Centre for
Computational Finance
and Economic Agents

The Thesis Committee for Jorge Martins Faleiro Junior

Certifies that this is the approved version of the thesis:

SUPPORTING LARGE SCALE COLLABORATION AND CROWD-BASED INVESTIGATION IN ECONOMICS: A COMPUTATIONAL REPRESENTATION FOR DESCRIPTION AND SIMULATION OF FINANCIAL MODELS

Supervisors

Professor Edward Tsang

Dr. Carmine Ventre

Research Committee

Dr. Nigel Newton

Dr. Michael Fairbanks

Dr. Steve Phelps

Viva Examiners

Professor Kuo-Ming Chao

Dr. Daniel Karapetyan

Dr. Florentina Hettinga

SUPPORTING LARGE SCALE COLLABORATION AND CROWD-BASED INVESTIGATION IN ECONOMICS: A COMPUTATIONAL REPRESENTATION FOR DESCRIPTION AND SIMULATION OF FINANCIAL MODELS

by

Jorge Martins Faleiro Junior, MSc

Doctoral Thesis

Presented to the Faculty of the Centre of Computational Finance and Economic Agents of

University of Essex

in Partial Fulfillment

of the Requirements

for the Degree of

PhD in Computational Finance

University of Essex

February of 2018

DEDICATION

This thesis is dedicated to the memory of Mom and Dad, “Dona Juliêt e Seu Jorge”.

For your example of wisdom, dedication, and unconditional love.

Essa dissertação de doutorado é dedicada aos meus pais, “Dona Juliêt e Seu Jorge”.

Pelo seu exemplo de sabedoria, dedicação e amor incondicional.

Saudades.

ACKNOWLEDGEMENTS

This thesis would not be possible without the help, support, and guidance of several exceptional individuals that, in one way or the other, contributed to the preparation and successful completion of this research.

First and foremost, I want to thank Nanda, for her loving support and encouragement. You were always keen to know what I was doing, and how I was doing it. And here is the result, one more momentous milestone complete.

My very special gratitude goes to all the members of my research committee, past and present, for the quality of the timely interaction and relevant feedback: Professor Edward Tsang, Dr. Carmine Ventre, Dr. Nigel Newton, Dr. Michael Fairbank, and Dr. Steve Phelps.

I would like to especially thank the supervisor of this work, Professor Edward Tsang, for his quality feedback, insights, guidance, and for allowing the exact degree of latitude required for such innovative and sophisticated work of research.

Thanks to all family, friends, and co-workers for encouragements on the beginnings and your support along the way, and for your help with reviews and comments. You coped with my absence and showed the utmost patience and understanding with my divagations and utterances about a line of research that became a new passion of mine.

I would also like to thank and acknowledge the importance of an anonymous crowd of enthusiasts, unsuspecting participants on this research, for all insights given by answering to my open-ended posts, and for providing suggestions on my published open source models and puzzles. Unknowingly you were putting to the test the very idea of this research: collaborative crowds.

And finally, last but by no means least, thanks to the University of Essex support staff for the unyielding assistance on several topics related to the nitty-and-gritty execution of this research.

ABSTRACT

SUPPORTING LARGE SCALE COLLABORATION AND CROWD-BASED INVESTIGATION IN ECONOMICS: A COMPUTATIONAL REPRESENTATION FOR DESCRIPTION AND SIMULATION OF FINANCIAL MODELS

Jorge Martins Faleiro Junior, PhD

University of Essex, 2018

Supervisors: Edward Tsang

Carmine Ventre

Finance should be studied as a hard science, where scientific methods apply. When a trading strategy is proposed, the underlying model should be transparent and defined robustly to allow other researchers to understand and examine it thoroughly. Any reports on experimental results must allow other researchers to trace back to the original data and models that produced them.

Like any hard sciences, results must be repeatable to allow researchers to collaborate and build upon each other's results. Large-scale collaboration, when applying the steps of scientific investigation, is an efficient way to leverage *crowd science* to accelerate research in finance.

Unfortunately, the current reality is far from that. Evidence shows that current methods of investigation in finance in most cases do not allow for reproducible and falsifiable procedures of scientific investigation. As a consequence, the majority of financial decisions at all levels, from personal investment choices to overreaching global economic policies, rely on some variation of try-and-error and are mostly non-scientific by definition. We lack transparency for procedures and evidence, proper explanation of market events, predictability on effects, or identification of causes. There is no clear demarcation of what is inherently scientific, and as a consequence, the line between fake and genuine is blurred.

In this research, we advocate the use of a next-generation investigative approach leveraging forces of human diversity, micro-specialized crowds, and proper computer-assisted control methods associated with accessibility, reproducibility, communication, and collaboration.

This thesis defines a set of very specific cognitive and non-cognitive *enablers* for crowd-based scientific investigation: *methods of proof*, *large-scale collaboration*, and a domain-specific *computational representation* for the field of economics. These enablers allow the application of procedures of structured scientific investigation powered by crowds, a collective brain in which neurons are human collaborators.

TABLE OF CONTENTS

LIST OF EQUATIONS

LIST OF FIGURES

LIST OF CONTRIBUTIONS

RESEARCH PUBLICATIONS

Papers, Reports, Book Chapters

Faleiro Jr, Jorge M, and Edward P. K. Tsang. *Supporting Crowd-Powered Science in Economics: FRACTI, A Conceptual Framework for Large-Scale Collaboration and Transparent Investigation in Financial Markets. 14th Simulation and Analytics Seminar*. Helsinki: Bank of Finland, 2016. (Faleiro Jr e Tsang 2016a)

Faleiro Jr, Jorge M, and Edward P. K. Tsang. "Black Magic Investigation Made Simple: Monte Carlo Simulations and Historical Back Testing of Momentum Cross-Over Strategies Using FRACTI Patterns." Submitted to *Algorithms: Special Issue Algorithms in Computational Finance*. MDPI AG, 2019. (Faleiro Jr e Tsang 2016)

Faleiro Jr, Jorge M. *Automating Truth: The Case for Crowd-Powered Scientific Investigation in Economics*. Report, Centre of Computational Finance and Economic Agents, University of Essex, Colchester: University of Essex, 2016a. (J. M. Faleiro Jr 2016a)

Faleiro Jr, Jorge M. *A Language for Large Scale Collaboration in Economics: A Streamlined Computational Representation of Financial Models*. Report, Centre of Computational Representation and Economic Agents, University of Essex, Colchester: University of Essex, 2017. (J. M. Faleiro Jr 2017)

Faleiro Jr, Jorge M, and Edward P. K. Tsang. "Crowd-Powered Monitoring in Large Scale: A Collaborative Environment for Early Detection and Investigation of Systemic Failures in Financial Markets." Submitted to *Handbook of Global Financial Markets: Transformations, Dependence, and Risk Spillovers*. World Scientific Publishing, 2019. (Faleiro Jr e Tsang 2019)

Open Source

—. *QuantLET: an open source, event-driven framework for real-time analytics*. 08 2008. http://quantlet.net (J. M. Faleiro Jr 2008)

CHAPTER 1. INTRODUCTION

"The duty of a man who investigates the writings of scientists, if learning the truth is his goal, is to make himself an enemy of all he reads… attack it from every side. He should suspect himself as he performs his critical examination so to avoid falling into either prejudice or leniency"- Ibn al-Haytham (Abdelhamid 2003)

1.1. MOTIVATION AND BACKGROUND

The long path of this research started just a few years after the great recession of 2008. At that time the consequences of harmful economic policies that led to the crisis were still fresh, and pieces and bits of the global havoc were still ricocheting in distant corners. Regardless of where you looked, the apocalyptical chaos, the confusion, the financial and emotional losses served as clear signs that wrong incentives and bad economic policies carry the same destruction potential of war arsenals.

Individuals, municipalities, and sovereign states will forever carry the losses and scars of destruction. History will record the consequences for posterity in the hopes that we can avoid the same mistakes in the future, but only if we can clearly understand the causes, and the complex process that led to the disaster.

Over the next several years the financial community took to the board, looking for causes and answers, trying to explain why and how we had fallen into such a trap. Especially at a time when we had access to latest and greatest technologies, and we were going over the peak of our intellectual enlightenment, amassed over centuries of the application of the scientific method. The same scientific method we chose to ignore.

The majority of the answers, in essence, pointed to a justification along the lines of "we didn't know better". This research, however, points to a slightly different answer: *we thought, and pretended, we knew better*. And that's when the inquiry that led to the idea of this research started.

There is extensive evidence (Freedman 2011) (Cassidy 2013) (Reinhart and Rogoff 2010) (Olsen and Cookson 2009) (Lehrer 2011), unlike other subjects we commonly associate to hard sciences[1], that the discipline of economics[2] has been distancing itself from the guidelines of the scientific method[3]. This research was born out of this observation and can be condensed into one simple question:

> *Why don't scientific procedures carry in economics the same weight they carry in other hard sciences, like engineering, physics or biomedicine?*

This is indeed a simple question, and this research will show that, as it is usually the case with most simple questions, the answer is multifaceted and of relative complexity.

The search for an answer requires practitioners to approach the problem and the complex issues surrounding it with candor, ingenuity, and transparency. In the words of Ibn al-Haythan, in the opening quote of this chapter, "we should suspect ourselves, and attack our pre-conceptions and biases from every side, and avoid falling into prejudice or leniency". We shall seek the truth.

One could argue that our inability to accurately answer this scientific question is indeed the enabler of all our economic maladies. Several of the policies that lead to

[1] What we call on this paragraph "science" could be more accurately referred to as a "scientific method of investigation", described in details in Section 3.3.3

[2] In the context of this document the terms "financial sciences" and "economics" have interchangeable connotations.

[3] This observation discounts the harmful consequences of the rebirth of the anti-science movement (Holton 1993) so fashionable at the time of this writing.

the great recession of 2008 were born out of inexistent or defective experimentation, grew organically for many years, feeding on our collective and individual biases, and ultimately were left unattended, compounding themselves to catastrophic end results.

Fast-forwarding to the time of the writing of this thesis, when we are closing this phase of this research, several years after the crisis started, many of the consequences are still lingering around.

By now it is clear that decisions that affected billions of lives were made with no scientific insight into hard evidence and away from proven investigative procedures. The same procedures we use for the definition of what we call *hard science*[45]. The broad consensus at this time is that the cost of driving economic policies by loose investigation procedures in economics is just too high. Society can no longer afford such a cost.

So, referring back to our original question: if we know the tools and understand the harsh consequences of treating economics differently than other hard sciences, why do we insist in repeating the same mistakes over and over again? If we can safely fly in vehicles made of metal, perform unmanned pinpoint landings in dashing meteorites, and quickly continue to solve the mysteries of life hiding deep into our chromosomes, why can't we achieve similar results from our explorations in economics?

---

[4] We use the term "hard sciences" as it was coined by Nobel Prize winner in Economics in 1978 Herbert Simon, "for his pioneering research into the decision-making process within economic organizations", on his words: "The social sciences, I thought, needed the same kind of rigor and the same mathematical underpinnings that had made the *'hard' sciences* so brilliantly successful" (Simon 1978)

[5] We incorporate the colloquial definition of "hard" and "soft" sciences to respectively discern between natural sciences (e.g., biology, chemistry, and physics) and social sciences (e.g., economics, psychology, sociology) based on "evidence of a hierarchy of sciences" (Fanelli and Glanzel 2013).

Searching for answers, this research had to look at different sides of the problem, but before all, we had to recognize that, despite the simple initial observation, we are indeed dealing with systems of extreme complexity. To model this level of complexity, we[6] advocate leveraging a powerful machine. A machine we just now start to understand and tap into its full potential: the "collective brain" of human crowds (Muthukrishna and Henrich 2016).

Over the following chapters, we define a conceptual framework to research finance and economics as strictly scientific disciplines, uncovering what we call *enablers* for crowd-based investigation procedures, powered by a collective brain in which neurons are human collaborators (Nielsen 2012, 18).

[6] This thesis is the product of an individual work, strongly influenced by intellectual mentors, literally thousands of different references and supporting ideas. Therefore, throughout this thesis "we" is used instead of "I" for fairness, accounting for all those contributions.

1.2. OBJECTIVES

Objective 1 *Identify what is required for the adequate use of crowds in structured, scientific investigation*

Economics and finance are particular domains of knowledge in which related systems and agents – markets, humans and their relationships – are tough, if not impossible, to model. Adequate financial models must deal with the intrinsic complexity of economic agents (Foster 2004) (Arthur 2013) (Freedman 2011). On this research, we advocate the use of crowds for the resolution of complex problems in general and in economics in particular, an approach we are calling *crowd-based investigation*. There is empirical evidence for the suitability of crowds for investigation and the resolution of complex problems, but the mechanisms that should be in place to allow that to happen are not clear. Hence, the first objective of this research is to identify what properties are required for the adequate use of crowds in structured, scientific investigation.

Criteria for Success: Identify a set of specific cognitive and non-cognitive (computational) requirements for the adequate use of crowds in structured, scientific investigation. We are calling these requirements *enablers* of crowd-based investigation. Enablers for crowd-based scientific investigation are described in Chapter 3.

Objective 2 *Define a specialized computational representation to allow proper controls and collaboration in large scale in the field of economics.*

There is evidence describing the widespread increase over time in the misuse of procedures and computational resources in science, intentionally or not (Nuzzo 2014) in which the primary symptom is the lack of reproducibility and falsifiability. Such increase can be correlated to the exponential increase over time in available computational power (Goecks, Nekrutenko and Taylor 2010) (Amdahl 1967). In such correlation, causality cannot be inferred, but a natural relationship can. Computational power can be seen as the "glue" for collaboration and interaction in large scale, but with the condition that it has to be leveraged carefully. Computational power without proper control breads chaotic data and methods, therefore impeding reproducibility. This research advocates that (a) proper control and collaboration in large scale can only be achieved if an adequate computation representation is used; and (b) computational representation is bound to the specific domain of knowledge it tries to model. Hence, the second objective of this research is a specialized computational representation that allows proper controls and collaboration in large scale in the field of economics.

Criteria for Success: Given enablers outlined as a result of Objective 1, define a computational representation to support collaboration in large-scale for the field of economics. This computational representation is described in Chapter 4.

Objective 3 *Select a non-trivial problem to demonstrate how a real inquiry in economics can be studied using the scientific method, tools, and procedures defined as a result of Objective 1 and Objective 2.*

The third and last objective of this research is to select a non-trivial problem to demonstrate how a real inquiry in economics can be studied using the scientific methods, tools, and procedures described in Chapter 3 and Chapter 4. The problem should demonstrate a relevant inquiry in economics, be widely known as to encourage specialist involvement and argumentation, and yet still allow limited complexity to a level in which it facilitates discussions amongst a broader range of economic participants in diverse roles.

Criteria for Success: Define an end-to-end investigation exercise in which we measure the actual efficiency of technical analysis using formal methods and historical trading data, conducted step by step. The complete investigation exercise is demonstrated in Chapter 5.

1.3. DOCUMENT ORGANIZATION

This thesis is organized into four major parts: an introduction in Chapter 1; a literature survey in Chapter 2; research topics in Chapters 3 to 5; and a conclusion in Chapter 6.

The introduction in this chapter describes the motivation and background for this research in addition to objectives listed in Section 1.2. Each objective highlights a number of criteria for success and a reference to a research chapter in which that objective was attained.

In Chapter 2 is listed the main pieces of literature and previous knowledge relevant to this research across specific topics of research: philosophy of science, scientific learning, scientific support systems, large-scale collaboration, and computational representation.

The research is described in chapters 3 to 5. In Chapter 3 is described the cognitive and non-cognitive requirements, or enablers, for crowd-based investigation: methods of quantitative proof; large-scale collaboration; and computational representation. In Chapter 4 is described the computational representation for the field of economics through a representation system and cases of use. Chapter 5 describes a non-trivial problem to demonstrate how a real inquiry in economics can be studied using the scientific methods, tools, and procedures defined in previous chapters 3 and 4. The end of each of the research chapters brings a synopsis section, describing a relationship between the content of that chapter and the specific objective listed in the introduction

In Chapter 6 is listed a summary of this work, assumptions, contributions, known limitations, and a list of suggestions for future work.

CHAPTER 2. LITERATURE SURVEY

"If history is any indication, all truths will eventually turn out to be false" – Dean Kamen (Kamen 2014).

2.1. TOPICS OF RESEARCH

The scientific question stated in Section 1.1 and the assumptions for crowd-based investigation stated in Section 3.1 drove our study into five different topics of literature: scientific learning and economics; scientific support systems; large-scale collaboration; computational models; and a non-trivial study in economics.

Over the following sections we will synthesize the main references on current and established literature on each of these topics, highlight equivalencies between this research and other related studies, as well as differences in approach and points of view.

2.2. SCIENTIFIC LEARNING AND ECONOMICS

The scientific question stated in Section 1.1 requires a fresh look into why we learn and prove things in economics differently than other hard sciences. This question is not asked for the first time on this present research. A limited number of previous works have been inquiring the same and getting to different answers (Chen 2005) (Tsang 2010) (Simon 1978).

In order to get to a specific answer relevant to this research we had to find a commonly accepted definition for a modern scientific method (Munir 2010) and apply these same general concepts to crowds (Franzoni and Sauermann 2014). The explanation of how humans form ideas and understand knowledge, an integral and hidden part of the scientific method, uses abstract concepts that are explained by a

subject usually called philosophy of science (Oberdan 2016) (Godfrey-Smith 2003) (Mulder and van de Velde-Schlick 1978a).

The subject of philosophy of science is complex, multidisciplinary (Camerer, Lowenstein and Prelec 2005) (Camerer and Loewenstein 2002) (Madhavan 2000), mostly abstract, and under constant review and debate (Nola and Irzik 2006) (Hawthorne 2017). To avoid extending the scope of this research beyond what would be practical for a doctoral dissertation, our investigation of this subject was indeed the delineation of a thin line we had to walk very carefully. As a consequence we had to take a pragmatic approach to pinpoint specific literature related to our two very specific concerns:

- Delineate a "pipeline of proof" that could be leveraged in crowd-based investigation (Popper 2005) (Peirce 1883) (Hawthorne 2017) (Kapitan 1992) (Rodrigues 2011).

- Identify a quantitative method to infer conclusions - "quantitative inference" - that could be used as a gold standard for validation of evidence and exchanged across crowds of participants (Kass 2011) (Lenhard 2006).

The first concern, a "pipeline of proof", is an original idea proposed in this research, adjusted from similar pre-existing insights. It defines an algorithmic sequence of steps of investigation, played by a crowd and orchestrated by computers, in which each step of the pipeline is expected to generate a byproduct useful for the process of investigation. These products must be commonly understood and accepted by the crowd of participants (Gauch 2003) (Horgan, (b) 1993) (Horgan, (a) 2016).

The more general idea of a method of investigation is as old as the "method of hypotheses" of Plato (Nola and Irzik 2006). Later Francis Bacon proposed a more

specific step-by-step, methodical approach for investigation, and a general "record of observations" in *Novum Organum* in 1620 (Alkhateeb 2017) (Jürgen 2016) (Cintas 2003). More recently similar ideas are used for automation of experimentation (Soldatova, et al. 2016). Similarly still, the approach of arranging a step-by-step process as a pipeline is proposed for biomedical research and is referred to as a "statistical pipeline" (Frazee, et al. 2014) (Ochs 2010). Although similar in an overreaching purpose, the standardization of the understanding of what should be considered true or false, and the scope of what the pipeline would entail is different than what this research proposes. Their scope is limited to software patterns and a computational platform. Their metrics and records are concentrated on specific quantitative metrics.

Our second concern comes as a requirement for the proper definition of a proof pipeline: the need of standard quantitative metrics to get to conclusions that are commonly agreeable by all participants in a crowd (Fetzer 2017). For the scope of this research conclusions are attained based on quantifiable metrics by looking at statistical characteristics of data, and using probability alone (Lindley 2000) (Apolloni, Malchiodi and Gaito 2006). This process is by definition called statistical or quantitative inference. Once more, as it was the case previously while researching literature related to the philosophy of science, the foundations for the process of quantitative inference are abstract and subject to multiple interpretations and debate (R. A. Fisher 1922) (Jeffreys 1933) (Neyman 1934) (Savage 1954) (Efron 1978). In the scope of this research we acknowledge that all those differences are important, and instead of concentrating on the contentiousness of the debate, we emphasize how these differences are complementary. This approach is called statistical pragmatism (Kass 2011).

The final part of the literature review of scientific learning applied to economics concentrates in collecting evidences pointing to the misuse of scientific procedures (Ioannidis, Allison, et al. 2009) (Camerer, Dreber, et al. 2016) (Open Science Collaboration 2015) and the main causes for this deviation (Colquhoun 2016) (Jha 2012) (Cokol, et al. 2007) (Ioannidis 2005) (Martinson, Anderson and De Vries 2005). We were especially interested when the case under study was related to economics or financial sciences (Reinhart and Rogoff 2010) (Herndon, Ash and Pollin 2013) (Cassidy 2013) (Olsen and Cookson 2009) (Lehrer 2011) (Tsang 2014) (Nuzzo 2014).

Literature related to scientific learning and economics, expressing their respective limitations, can be compared mainly over the following topics:

- The underlying subject of philosophy of science is complex and far from a commonly understood consensus (Camerer, Lowenstein and Prelec 2005) (Camerer and Loewenstein 2002) (Madhavan 2000) (Nola and Irzik 2006) (Hawthorne 2017). To avoid extending the scope of this research beyond what would be practical for a doctoral dissertation, we had to investigate a pragmatic approach applicable to crowds (Kass 2011) (Lenhard 2006).

- A “pipeline of proof” is an original idea proposed in this research, adjusted from similar pre-existing insights investigation (Popper 2005) (Peirce 1883) (Hawthorne 2017) (Kapitan 1992) (Rodrigues 2011). The approach of arranging a step-by-step process as a pipeline is proposed for biomedical research and is referred to as a “statistical pipeline” (Frazee, et al. 2014) (Ochs 2010). Although similar in an overreaching purpose, the standardization of the understanding of what should be considered true or

false, and the scope of what the pipeline would entail is different than what this research proposes

- The foundations for the process of quantitative inference are abstract and subject to multiple interpretations and debate (R. A. Fisher 1922) (Jeffreys 1933) (Neyman 1934) (Savage 1954) (Efron 1978). In the scope of this research we acknowledge that all those differences are important, and instead of concentrating on the contentiousness of the debate, we emphasize how these differences are complementary. This approach is called statistical pragmatism (Kass 2011).

- There is widespread evidence in the literature pointing to the misuse of scientific procedures (Ioannidis, Allison, et al. 2009) (Camerer, Dreber, et al. 2016) (Open Science Collaboration 2015) and the main causes for this deviation (Colquhoun 2016) (Jha 2012) (Cokol, et al. 2007) (Ioannidis 2005) (Martinson, Anderson and De Vries 2005). We were especially interested when the case under study was related to economics or financial sciences (Reinhart and Rogoff 2010) (Herndon, Ash and Pollin 2013) (Cassidy 2013) (Olsen and Cookson 2009) (Lehrer 2011) (Tsang 2014) (Nuzzo 2014).

2.3. SCIENTIFIC SUPPORT SYSTEMS

Scientific support systems are computer systems utilized to control and automate procedures and workflow related to scientific investigation. There are examples in the literature of the use of computers specifically in data and workflow systems in scientific procedures (Goecks, Nekrutenko and Taylor 2010) and in some cases the function is conflated with typical workflow management systems (Curcin and Ghanem 2008).

Our research relies on a pre-existing open source project (J. M. Faleiro Jr 2008) that provides a dialect for a scientific support system. This dialect has been providing important insights to this research, and on the other way, the research also fed back to the dialect several ideas that were materialized as concrete extensions. This dialect relies on many underlying computational resources (Pérez and Granger 2007) (Jones, Oliphant and Peterson 2001) (McKinney 2010) (Hunter 2007) and we should expect more to come, as the chain of direct and indirect dependencies is fluid and is always changing.

Literature related to scientific support systems, expressing their respective limitations, can be compared mainly over the following topics:

- There are examples in the literature of the use of computers for the function of scientific support system, specifically limited as data and workflow systems (Goecks, Nekrutenko and Taylor 2010).

- The function of a scientific support system is conflated with typical workflow management systems (Curcin and Ghanem 2008)

2.4. LARGE-SCALE COLLABORATION

The implicit assumption of the use of crowds in scientific research calls for collaboration in large-scale. The research of the topic of collaboration in large-scale was done in terms of requirements (Nielsen 2012) (Udell 2002), complexity (Gowers and Nielsen 2009) (Gowers 2009a) (Gowers 2009b) (Polymath 2012) (Nielsen 2011) (Beall 2008) (Shen and Björk 2015), and an original perspective that offers a justification for large-scale collaboration as a consequence of the historical evolution of scientific participants and society (Krauss 2012).

2.5. COMPUTATIONAL MODELS

Some of the novelties of this research are related to the introduction of computational representations, the assumption that computational representations are closely tied to a domain of knowledge, and the assumption that a proper computational representation is a requirement for an effective crowd-based investigation, where machines orchestrate the interaction of participants in a crowd. The introduction of a computational representation as a domain-specific representation requires a number of references in terms of visualization and understanding of computer models in general (Tufte 2006) (Schwab, Karrenbach and Claerbout 2000) (Gentleman 2005).

Our research also opted for models in which the knowledge representation system in use is role-based (Davis, Shrobe and Szolovits 1993) (Rayo 2007) what brings as a consequence the irrelevancy of comparison between different models (Hayes 1978) and other important assumptions and consequences, described in Section 3.4.5.

For the definition of the computational representation for the field of economics were used a number of relevant empirical exercises (Faleiro Jr and Tsang 2016) (J. M. Faleiro Jr 2015) (J. M. Faleiro Jr 2014) (J. M. Faleiro Jr 2014) (J. M. Faleiro Jr 2008) (J. M. Faleiro Jr 2013), exemplifying very characteristic cases of use in finance. This research used a number of references to define properties for each of the facets: streaming (Stephens 1997) (McIlroy 1964), reactive (Bainomugisha, et al. 2013) (Harel and Pnueli 1985), distribution (Dean and Ghemawat 2004) (Hohpe and Woolf 2012), and simulation (Von Ronne 2012) (Robinson 2004).

2.6. INVESTIGATION EXERCISE

For the investigation exercise we selected an exploration of profitability of a momentum trading strategy based on a moving average cross over model (Schoeffel 2011) (Chestnutt 1955). To mitigate biases and provide a balanced analysis the investigation, the exercise checked for comparable literature and previous research testifying in favor (Patterson 2007) (Brown, Goetzmann and Kumar 1998) or against (Marshall, Cahan and Cahan 2010) the efficiency of technical analysis.

On final notes and evidences of profitability there are a number of surveys pointing for an explanation of our results based on biases (Park and Irwin 2007) (Young and Karr 2011) (Shermer 2014) (Marshall, Cahan and Cahan 2010), noise (Black 1986), past memory or private information (Fama 1965), and market inefficiencies (Marshall, Cahan and Cahan 2010) (Chaudhuri and Wu 2003). We consider each of these possibilities for an explanation of the final results in Section 5.5.

CHAPTER 3. ENABLERS OF CROWD-BASED INVESTIGATION

"The fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone"- Albert Einstein (Einstein and Infeld 1938)

This research advocates the use of crowds for the investigation and resolution of complex problems in general and in economics in particular. In this chapter, we identify what is required for the adequate use of crowds for scientific investigation, an approach we are calling *crowd-based investigation*. These requirements are called *enablers* for crowd-based investigation and are classified as either cognitive or non-cognitive.

Cognitive enablers relate to non-computational features related to mechanisms of human understanding, and as the name implies, related to cognition. Cognition is defined as "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" (Oxford English Dictionary 2011), and refer precisely to the subjective human process by which we build knowledge, and the underlying social fabrics of large-scale collaboration. Regardless of how evolved and subjective they are, cognition mechanisms are the same, regardless of the domain of knowledge in which they are applied. As a consequence, cognitive enablers are not domain specific, and should be the same regardless of the domain of knowledge under consideration.

Non-cognitive enablers, on the other hand, relate to features that can be directly and purely mapped to a computational description. Unlike cognitive enablers, non-cognitive enablers are domain specific. Different domains of knowledge require a different, specially tailored, computational description. This specifically tailored

computational description is called in the context of this research a *computational representation*.

Over the next sections, we describe assumptions and the rationale for the definition of enablers, and their importance on the definition of the subject of this research: *a crowd-based framework for investigation in economics*.

3.1. ASSUMPTIONS

Economics and finance are particular domains of knowledge in which related systems and agents – markets, humans and their relationships – are hard, if not impossible, to model. For the appropriate investigation, adequate financial models must be able to deal with this intrinsic complexity of economic systems and agents (Foster 2004) (Arthur 2013) (Freedman 2011).

This research advocates the use of crowds for the investigation and resolution of complex problems in general and in economics in particular, an approach we are calling *crowd-based investigation*.

At this point, available literature suggests that the suitability of crowds for the resolution of complex problems can only be confirmed by empirical evidence, as explained in the upcoming Section 3.3.2. In this research, we consider that the definition of mechanisms that should be in place to allow the use of crowds for the resolution of complex problems to be mostly axiomatic, based on three specific assumptions:

- The process by which we acquire objective knowledge must follow the rules dictated by the modern scientific method. As a consequence, the proof of observations as being real or false must be driven by a widely known set of quantifiable standards, as explained in Section 3.2.1.

- Human collaboration in large scale is an adequate method to investigate and resolve complex problems. Collaboration in large scale is enabled by providing the correct set of incentives to crowd participants, as explained in Section 3.3.1.

- Computers should fulfill the role of a tool to support discovery and should not serve as a replacement for the application of reproducible and falsifiable procedures of the scientific method.

Additional assumptions related to the overall research are listed in Section 6.3. The applications of these assumptions bring two immediate consequences concerning a method for resolution of complex problems in general:

- The need for an investigation method that applies to large groups of individuals, or crowds, for the resolution of complex problems, or problems of difficult representation through models. We are calling methods of investigation that apply to collaborative investigation methods of *crowd-based investigation*.

- The scientific effectiveness of an investigation based on crowds is related to the existence of specific environmental requisites that must be in place to allow, but not necessarily guarantee, the application of a proper scientific method by crowds of individuals[7]. We are calling these enabling requisites *enablers of a crowd-based investigation*.

The association between a crowd-based method of investigation and its cognitive and non-cognitive enablers is shown in Figure 1. Those enablers are classified as either cognitive or non-cognitive.

[7] The implicational relationship of the statement is that the existence of enablers in an environment is a necessary but not sufficient condition for the proper support of crowd-based investigation.

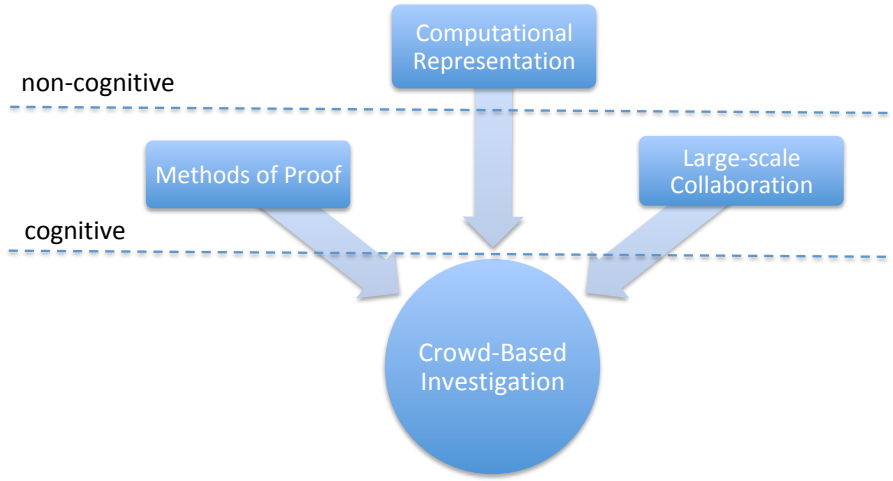


Figure 1. Enablers of Crowd-Based Methods of Scientific Investigation

Cognitive and non-cognitive requisites, or enablers, of the next wave of investigation methods based on crowds: methods of proof; large-scale collaboration; and a computational representation.

Cognitive enablers relate to non-computational features associated with the subjective mechanisms of human understanding of what to consider knowledge and the underlying fabrics of large-scale collaboration. Cognitive enablers are not domain specific, and as a consequence should be the same regardless of the domain of knowledge under consideration. Cognitive enablers are methods of proof and large-scale collaboration and are described respectively in the upcoming Sections 3.2 and 3.3.

Non-cognitive enablers, on the other hand, relate to features that can be directly and purely mapped to a computational description. This research refers to this description as a *computational representation*, explained in details in the

upcoming Section 3.4. Unlike cognitive enablers, non-cognitive enablers are domain specific, and as a consequence, each domain of knowledge must be supported by a different, specially tailored, computational representation. The generic process for a definition of computational representations for any domain of knowledge is called a *representational process* and is described in the upcoming Section 3.4.1.

3.2. METHODS OF PROOF

Humans learn new things through investigation. Careful investigation is what establishes if an observed phenomenon is real, or it should be deemed just a result of random forces of nature at play. The primary target of any investigation is to establish facts, as accurately as possible, by proving observations to be either true or false. That is how humankind has been accumulating objective knowledge for as long as we walk this earth, and this is why defining precise methods of proof is crucial.

The mental process we follow as individuals to investigate and learn about things is not straightforward. Even at this present date, science is still not able to unequivocally explain the process by which we learn and assess things. If this is true when we produce our thoughts on our own, we should expect an even more elaborate process to be at play when we introduce procedures of investigation that are performed by multiple individuals, organized in seemingly chaotic crowds. These adequate procedures of investigation are indeed the very nature of this research: the proposition of a conceptual platform in which objective knowledge is produced by collaboration and interaction of crowds of individual participants.

Given the number of participants and the nature of the interaction – formal scientific investigation - we can safely expect as a consequence a large number of hypotheses being generated and tested. On this scenario, ideas must be defined, exchanged, discussed, and tested in a sequence of steps, arranged like a pipeline.

Procedures in each step of the pipeline should potentially generate massive amounts of data, and each piece of data should be unquestionably tested as true or false. As a consequence, each of the steps must abide by standards and validation metrics that must be well understood and accepted by all participants.

In this section, we define requisites and the composition of this process and steps involved, what we call a *proof pipeline* in Section 3.2.1. The foundation of a method of proof also involves a correlation of the scientific method to economics and the use of computers in automation of tasks in the scientific method, discussed later, respectively in Section 3.2.2 and 3.2.3.

3.2.1. PROOF PIPELINE

A proof pipeline for scientific investigation is proposed as a process, composed of individual tasks, each task producing standard outputs that can be used to prove or reject an observation.

A simplistic description of such a pipeline would be a tube, where its input, taken on the head of the pipe, is a problem, or a set of ideas under investigation, and other intangible aspects such as the experience of the individual performing the inquiry or the investigation. On the tail of the tube, the result of the investigation, as either true or false. Over the extension of the tube, there are small holes, from where the process produces pre-defined, controlled evidence. For the sake of understanding, a diagram of a proof pipeline is shown later in this thesis, in Figure 2, on page 45.

The idea of arranging a sequence of pre-defined steps to assert a result of an investigation as true or false is not new. There are references in the literature to a step-by-step process in biomedical research, specifically for statistical measurements, referred to as a "statistical pipeline" (Frazee, et al. 2014) (Ochs 2010). Although

similar in its overreaching purpose and the intended standardization of the understanding of what is true or false, the scope of what that pipeline would entail is different than what this research proposes. Their scope is also limited specifically to software patterns and a computational platform. In the field of economics, there are proposals in the literature with a minor overlapping with the idea of proof pipelines, arranging economic models as testable pieces of engineering, not necessarily as pipelines, referred to as "economic wind tunnels" (Chen 2005).

As described earlier, a proof pipeline is a process, and as it is usually the case with processes, each step or part is composed of smaller mechanisms, smaller gears. Some gears are familiar and well understood, others not so much. One of those gears, required to establish a proof pipeline, is the underlying mechanism by which we get to conclusions based on premises taken from specific outcomes of an investigation. This process of getting to conclusions based on premises is called *inference*[8]. An inference is a mechanism we use to evaluate, learn, and create. This intricate mechanism is responsible for some of the most fundamental structures of the scientific thought.

The mechanism of inference is a complex and abstract subject, difficult to explain. Human ingenuity is attracted to things that can't be easily explained, so scientists have been looking at the general subject of inference for a long time, trying to understand and explain the specifics through studies in philosophy, biomedicine, and even religion. This lengthy inquiry is far from over. Formalizations of the exact

[8] In this thesis the term *inference*, used without qualifications, refers specifically to *human inference* and is defined as "the act of passing from one's proposition, statement, or judgment, considered as true, to another whose truth is believed to follow from that of the former" (Merriam-Webster 2018). Inference performed by humans and machines are related to different mechanisms and should not be used interchangeably (Gellatly 1989). It also differs for the term *statistical inference*, or *quantitative inference*, also used in this thesis, defined as "the act of passing from statistical sample data to generalizations usually with calculated degrees of certainty" (Merriam-Webster 2018).

mechanisms at play are mostly abstract, and as it is usually the case with philosophical subjects, surrounded by controversy (Lindley 2000) (Wang 1993).

For this reason, in this research, we want to carefully, and intentionally, stay away from the argument. While we understand the importance of the debate and study of general concepts around the topic called "philosophy of science" (Godfrey-Smith 2003), each of the small topics under the subject could lend a lengthy separate doctorate dissertation in itself. Would be impractical and redundant to explore in this thesis all the open controversies, different viewpoints, intricate details, and differences between methods.

Hence, it is essential at this point to carefully define our scope of interest when it comes to the general topic of inference, namely four specific topics, in line with the assumptions outlined in Section 3.1:

- **Support for falsifiable and testable inquiry**: the demarcation of what should be considered scientific is given by investigation propositions formalized by statements that can be tested and falsified. A scientific statement should be capable of conflicting with possible or conceivable observations, in line with the principle of falsifiability, "statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable observations" (Popper 1962, 39)

- **Step-wise, algorithmic nature**: methods of inference should fit a general algorithmic structure and a step-by-step, procedural description, mimicking the sequential arrangement of a pipeline.

- **Participation and collaboration in large-scale**: investigation should incentivize collaboration and interaction of a large number of participants.

Features or metrics of inference should be well understood and serve as a quantifiable standard for what to be considered true or false.

- **Computer augmented**: computers should serve as control points for collaboration and interaction of human participants, and not as agents of scientific inquiry themselves.

The most commonly accepted model of inference describing the scientific inquiry based on testing and falsifiability is the *Hypothetico-Deductive model*, or *H-D model* (Nola and Irzik 2006). The H-D model is a composition of all known modes of reasoning (Kapitan 1992) (Rodrigues 2011) (Peirce 1883). In this sense, reasoning is defined as the act of associating premises to conclusions and is described through three distinct modes of reasoning: deductive, inductive, and abductive.

In deductive reasoning, conclusions are derived from premises known or assumed to be true. The connections between premises and conclusions are made by terms. Deductive reasoning is a top-down logic, in a sense that, if all premises are true, and all terms are unambiguous, then the conclusion reached is necessarily true (Shin and Hammer 2016).

$$P \rightarrow Q: All\ men\ are\ mortal$$
$$P: Socrates\ is\ a\ man$$
$$Q: Socrates\ is\ mortal$$

In statements P and Q, if P is true, then that would cause Q to be true. Since P is true, we should expect Q to be also true.

In inductive reasoning, premises are seen as probable evidence[9], at various degrees, for the truth of the conclusion. Since evidence is uncertain, conclusions of an inductive argument are probable, depending on the evidence given. Induction reasoning is bottom-up, in a sense that a conclusion is reached by extrapolating the original, specific premises, to more general rules (Hawthorne 2017). Given for example two statements:

P: All biological life forms that we know depend on liquid water
Q: If we discover a new biological life form it will probably depend on liquid water to exist.

In these statements, following inductive reasoning, a conclusion Q is probable, based on the uncertainty of the premise P. The inductive reasoning asserts that in the future it is possible that a newly discovered biological life form does not depend on liquid water.

[9] The available body of facts or information indicating whether a belief or proposition is true or valid (Oxford University 2010)

In abductive reasoning a conclusion is a simplest or most likely explanation for a set of premises. In other words, abductive reasoning is the inference to the best explanation. For example, given a valid conclusion and a rule, abductive reasoning attempts to select premises that, when asserted, can support the conclusion. Given for example:

$$P \rightarrow Q: \textit{When it rains, the grass gets wet}$$
$$Q: \textit{The grass is wet}$$
$$P: \textit{It might have rained}$$

This abductive reasoning allows for investigation through the definition of various hypotheses, which can be further tested and falsified by the definition of additional statements, or more importantly – evidence, or data. Additionally, the abductive mode of reasoning is considered the core of Karl Popper's falsifiability and testability argument of any scientific hypothesis (Popper 2005).

Popper's surprisingly simple theory proposes discovery to occur in two steps. On the first step – *conjecture*[10] – a scientist offers a hypothesis that might explain some natural phenomena. The second step – *refutation* – the hypothesis is tested in order to show that the hypothesis is false (Popper 1962). If we succeed to show that the original conjecture is false, we go back to the first step, build a new conjecture, and follow the two-step process again. If in the second step we fail to test the hypothesis as false we should assume that the original conjecture is – for the moment, and as far as we could not prove otherwise – correct (Godfrey-Smith 2003).

[10] A conjecture is not materialized as a specific contribution. As a consequence, multiple experts can express the same semantic, and therefore one conjecture can possibly be reflected in different models.

Popper's theory is fundamental to the definition of the proof pipeline proposed in this section through a variation of the H-D model, and application of all modes of reasoning. The exact formulation of the H-D model vary, but in most cases, it is a combination of Karl Popper's view of falsifiability and testing, and a less skeptical view about confirmation[11] (Godfrey-Smith 2003, 69). A less skeptical view, in this case, means that our reliance on the notion that evidence can affect the credibility of hypothesis is necessarily fallible[12] (Crupi 2016).

The essence of the idea behind the hypothetic-deductivism in science is old, with its origins in Plato's dialogues, referred to in that work as "the method of hypotheses" (Nola and Irzik 2006). In a broader sense, the H-D model relies on a proposition of a hypothesis in a way that it can be falsified by a test of this proposition against observations, or evidence. The H-D model represents a formalization of the scientific method through a set of a simple sequence of four steps: observe; form a conjecture; deduce predictions from a conjecture; and test the predictions (Godfrey-Smith 2003).

Additionally, the H-D model formalizes a process of investigation through individual, sequential steps. The formalization of a process of investigation through a pre-defined sequence of steps defines the process of discovery as inherently algorithmic (Alkhateeb 2017). The idea of algorithmic procedures of investigation is not new. A precursor of the modern scientific method, Francis Bacon, arguably a predecessor of Karl Popper in respect of the method of falsification (Jürgen 2016) had foreseen two critical interconnected insights that are relevant to this research:

---

[11] Confirmation refers to "the problem of understanding when observations can confirm a scientific theory", and what is required in order to have an "observation evidence for the theory". This is a complex philosophical problem, often referred to as "the mother of all problems" (Godfrey-Smith 2003, 39).

[12] Observations cannot confirm theories or conclusions, i.e., "even with extensive and truthful evidence available, drawing a mistaken conclusion is more than a mere possibility", and as a consequence "under usual circumstances, reasoning from evidence is necessarily fallible" (Crupi 2016)

- The step-by-step, methodical approach to investigation, where Bacon used the word "machine" to describe his method in *Novum Organum* in 1620 (Alkhateeb 2017) (Jürgen 2016) (Cintas 2003).

- Bacon's method intended to leverage a "community of observers to collect vast amounts of information" and tabulate it into a central repository accessible to all (Alkhateeb 2017).

Following through on Bacon's hint, if discovery is algorithmic, then we can safely assume that machines could perform it. Alternatively, better yet, as this research advocates, discovery can be performed in large scale, having machines orchestrate the steps and rules of the collaboration of human crowds.

The application of this H-D model as an algorithm to a framework supporting crowd-based investigation can be described through a set of specific steps (Godfrey-Smith 2003, 236):

- **Observe**. The observer should use personal experience to understand and appreciate the problem under study. Gather previous contributions[13] relevant to the case of use at hand.

- **Form a conjecture or hypothesis[14]**. Form a supposition, or a proposed explanation for the phenomena under observation, based on whatever limited evidence has been currently gathered as a starting point for further investigation. State an explanation of your hypothesis. Materialize that conjecture as a model. Share that model.

[13] For now, the term "contribution" here is used in the same sense as when authors "contribute" to Wikipedia. A more detailed formalization of contributions is done later in this thesis in Section 3.4.3.

[14] Literature refers to the specific step in the H-D model where a supposition or specific explanation is made as a *conjecture*, which is equivalent to the most commonly known term *hypothesis*. For fairness, and accuracy, this thesis refers to both terms interchangeably.

- **Deduce predictions from your conjecture**. Formalize predictions, stating what should be expected if the conjecture is true. Incorporate those predictions as part of the model.

- **Test**. Experiment with the model, looking for evidence (observations) that might disprove your predictions. Record all evidence as contributions and share those contributions. If predictions are disproved, so is the hypothesis: go back to step 2 and repeat.

This sequence of steps in a pipeline of proof is depicted in Figure 2. Each of the four steps in this process is detailed respectively in the following sections 3.2.1.1, 3.2.1.2, 3.2.1.3, and 3.2.1.4.

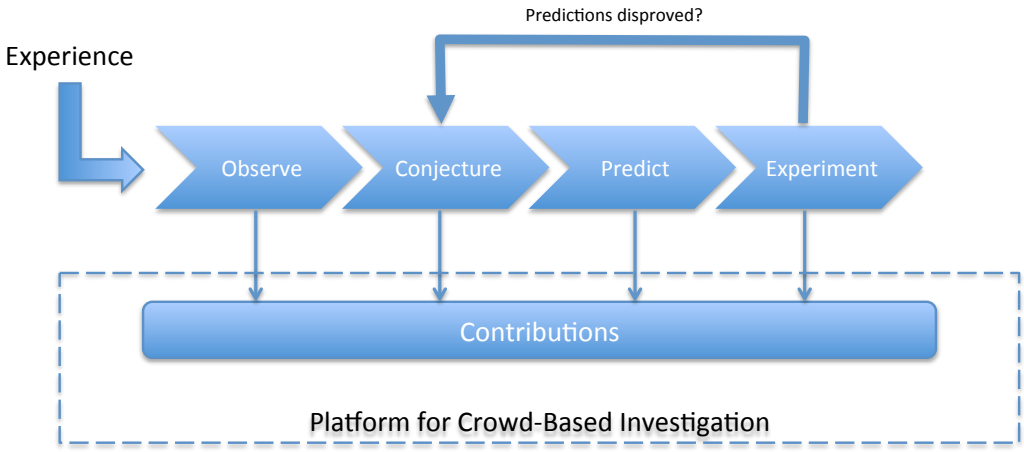


Figure 2. Method of Proof in Crowd-Based Investigation

A method of proof for collaboration in large-scale applying variations of the Hypothetico-Deductive Model to handle shared procedures of investigation in each of the phases: observe, conjecture, predict, and experiment.

This research advocates the use of this process, created from features of the H-D model, as a baseline for a proof pipeline for a crowd-based investigation and validation through the exchange of shared evidence

An end-to-end application of this algorithmic method is given during our investigation exercise of the profitability of momentum strategies, presented later in this thesis in Chapter 5.

3.2.1.1. Observation

The first step of a proof pipeline deals with human *observations*. Observations are organoleptic and by definition are subject to abstract human interpretation. Given its subjective nature, it would be hard, if not impossible, to use machines to automate the process by which we generate high quality, reliable observation records. On the other hand, machines should be ideal to establish a platform of collaboration in large scale where observations can be recorded and shared.

The idea of collecting observations in large scale is not new. In *Novum Organum,* in 1620, Francis Bacon proposed a method intended to leverage a "community of observers to collect vast amounts of information" and tabulate them into a central repository accessible to all (Alkhateeb 2017). Now technology allows that vision of a central registry of contributions that can be shared and evaluated by a community of observers.

An example of the observation step in practice is given later in this thesis, in Section 5.1 of the investigation exercise in Chapter 5, where we describe the overall problem, the mechanics of the mathematical models and algorithms involved, and observations that led to the original inquiry.

3.2.1.2. Conjecture

The second step of the proof pipeline is the generation of *conjectures*, or *hypotheses*, to attempt to explain the cause of phenomena under observation. Hypotheses must be falsifiable, and at the same time, by definition cannot be entirely and irrefutably confirmed. It should always be assumed that improved research methods should disprove a hypothesis at a later date.

Assuming an algorithmic nature of the discovery process, in what is commonly called *automated hypothesis generation*, hypotheses have a potential to be entirely generated by computers[15] (Spangler, et al. 2014). Automated hypothesis generation is still in initial stages, but research has produced a significant number of exceptional use cases.

Starting in the 1980's some experiments were able to hypothesize links between cause and effect, in two initially unrelated fields of study, without specialized knowledge in any of the subjects of study, and without conducting any experiments. The links were established by merely following algorithmic steps while connecting scientific papers with no citation overlaps (Swanson 1986). More recent research allows for limited automated hypothesis generation based on large-scale text mining of academic publications, natural language processing, mathematical modeling, and graph theory. Some equally relevant and related features, taken out of the techniques in use in hypothesis generation, include the prediction of a successful academic career based on the writing style of scientists on entry-level positions and quantifiable metrics of efficiency in scientific discovery (Rzhetsky, et al. 2015) (Grauwin, et al. 2017) (Sinatra, et al. 2016) (Gupta and Manning 2011) (Spangler, et al. 2014).

However, despite the evidences of potential and the slow advancement, science still lacks a complete theory for fully automated hypothesis generation. Additionally, these techniques currently rely on volumes of quality data associated with scientific publications, a scarce resource now that major scientific journals have placed severe restrictions on text mining of their content (Jha 2012).

[15] There are notable exceptions to the belief that discovery can be algorithmic. "Karl Popper insists there is no recipe for coming up with interesting conjectures" (Godfrey-Smith 2003)

Instead of a fully automated hypothesis generation, this research advocates the use of highly interconnected crowds orchestrated by computers. In such an environment, individual participants in a crowd would rely on computers to perform specialized discovery tasks, communication, and to produce metrics of quality on shareable contributions. In this scenario, hypotheses are generated by participants in a crowd, in an environment enhanced by computers, and not solely performed by machines.

An example of the conjecture step in practice is given later in this thesis, on the generation of hypotheses in Section 5.4.1. On that example, when we are assessing the profitability of momentum strategies, a conjecture is given by a well-defined, simple, falsifiable hypothesis, i.e.:

> *There are scenarios under which momentum strategies are consistently profitable.*

The consequences of a falsifiable conjecture are predictions. Predictions for this conjecture are expressed over the investigation exercise in Chapter 5, on page 177.

3.2.1.3. Prediction

The third step of the proof pipeline relates to *prediction*, where a researcher generates anticipations of probable outcomes of experimentations assuming that initial conjectures produced in the previous step, in Section 3.2.1.2, are true.

The mechanisms used to generate valuable, and high-quality predictions are similar to mechanisms we use to anticipate and track patterns in experience (Godfrey-Smith 2003). These mechanisms are subject to the complex rules that govern the

connection of experiences, or the rules of science itself[16] (Mulder and van de Velde-Schlick 1978a) (Oberdan 2016). These complex rules are bound to human traits of creativity and experience and, as of the time of this writing, there are no instances of efficient implementation in machines.

In the same manner, as defined for the first step previously described in Section 3.2.1.1, probable outcomes can be defined as different shocks of executions. Shocks are by definition an iteration of a simulation, as described in Section 4.3.4. The results of the execution of individual shocks are recorded in datasets as shareable evidence, allowing other participants to understand the expectations of a model better and assess predictions against actual outcomes.

An example of this prediction step in practice is given later in this thesis, in Section 5.4.1, when we define two specific predictions for a single conjecture, as part of the investigation exercise in Chapter 5.

3.2.1.4. Test

The last step of the proof pipeline is *testing*, where experiments are designed based on predictions produced during the third step, described in Section 3.2.1.3. Those experiments are performed in order to support or refute predictions, and the outcome of a test would either validate or falsify the original conjecture, or hypothesis.

In some fields of study reliant on intensive and controlled testing, procedures related to experimentation are widely automated. Scientists can submit a description of their experiments online and have that description subsequently converted to

[16] The core objective of science is to understand how experience shapes discovery, on the words of Moritz Schlick "what every scientist seeks (…) are the rules which govern the connection of experiences, and by which alone they can be predicted" (Mulder and van de Velde-Schlick 1979b)

specialized instructions and fed into robotic platforms to execute a battery of repeatable experiments (Soldatova, et al. 2016) (Alkhateeb 2017).

If we are to consider the assumption of standardized, quantifiable, and normalized results, as described in Section 3.1, it is important to introduce at this point the notion that experimentation on a complete method should also incorporate probabilities. In this case, a prediction should be expected to hold true $N\%$ of the time, in which case experimentation should be repeated to substantiate the probability N (Fetzer 2017).

On this sense, achieving unambiguous conclusions about a problem then becomes a numerical exercise, in which statistical inference[17] is the process of getting to conclusions about a specific problem by looking at statistical characteristics of data, and by using probability alone (Lindley 2000) (Apolloni, Malchiodi and Gaito 2006).

There is a widespread agreement that statistics depend on probability, but concomitantly there are disagreements as to what exactly is probability, and how probability is connected to statistics[18] (Savage 1954). Over the last several decades Ronald Fisher (R. A. Fisher 1922), Harold Jeffreys (Jeffreys 1933), Jerzy Neyman (Neyman 1934), Leonard Savage (Savage 1954), and many of their followers have defined several paradigms and have engaged in a number of debates that gave birth to controversies that were key to its formative properties (Efron 1978). A positive and

[17] The term *statistical inference*, or alternatively *quantitative inference*, is defined in this thesis as "the act of passing from statistical sample data to generalizations usually with calculated degrees of certainty" (Merriam-Webster 2018).

[18] The definition of probability is at the root of the division on the understanding of what is physical and evidential probability, as "it is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis." (Savage 1954, 2)

possibly unintended consequence of the debate is the multitude of statistical tools and the rich set of options available to the scientific community to conduct quantitative inference (Kass 2011) (Lenhard 2006).

On this research, we acknowledge that these differences are essential, but we assume that even more important is to leverage this toolset to concentrate on relationships between data and model, or how representations mapping measurements in the real to the theoretical world are made. This shift in paradigm, in which statistical models take a back seat to the understanding of the relationships between data and methods to infer conclusions, is called *statistical pragmatism* (Kass 2011). In statistical pragmatism numerical methods are seen as an eclectic practice, emphasizing mechanisms by which observed data is connected to statistical procedures, as described in Figure 3 (Kass 2011).

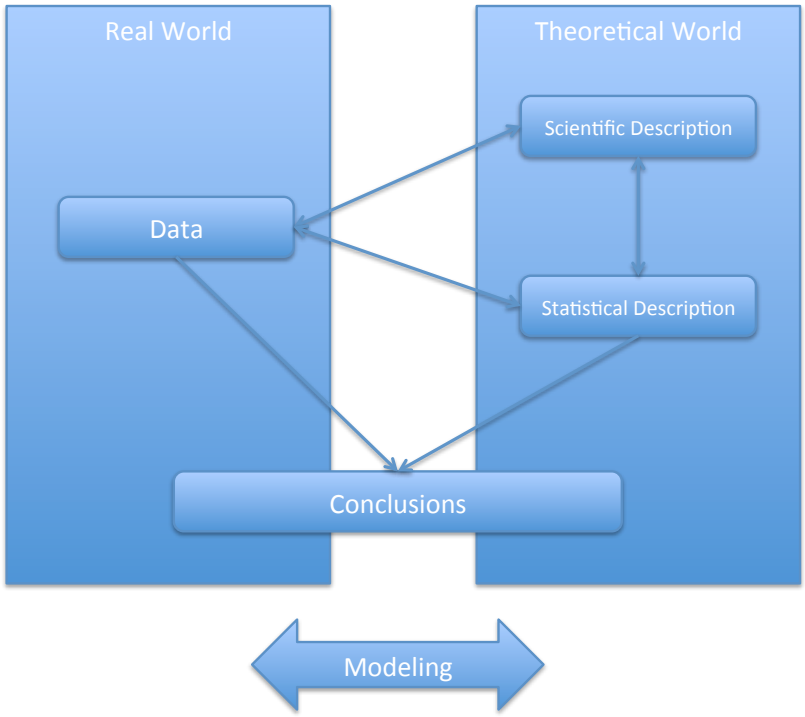


Figure 3. Pragmatic Statistics and the Mapping Between Data and Methods

Pragmatic statistics are defined through abstract mathematics constructs used to quantify and explain observable phenomena (Kass 2011).

In essence, pragmatic statistics is a vehicle to aggregate real data and theoretical descriptions into quantifiable results, and as such can be seen as a model to reach a set of conclusions based on real and theoretical constraints.

An example of the testing step and the partial application of pragmatic statistics is given later in this thesis, as part of the investigation exercise in Chapter 5. First on Section 5.4.2, for a Monte-Carlo simulation using stochastic generators on variations of arguments, and second, on Section 5.4.3, for the backtesting against constituents of the S&P 500 index (McGraw Hill Financial 2015a).

Evidence of testing results for the Monte-Carlo simulation using stochastic generators can be seen in Contribution 12 on page 182, and in Contribution 14 on page 186. Similar evidence for backtesting against constituents of the S&P 500 index is shown in Contribution 23 on page 196.

3.2.2. SCIENTIFIC LEARNING AND ECONOMICS

The scientific question posted in Section 1.1 brings in its core the observation that the process of investigation in economics should not be different from the process of learning we have in place in other hard sciences. As we have shown in Section 3.2, the process of learning and understanding followed by humans is abstract, fluid and subject to multiple definitions of what one might consider knowledge, assumptions, and beliefs. Given this abstract nature, it is essential to identify what is objective knowledge, or what is known, from what is not. Alternatively, in other words, demarcate the difference between what is considered science from what is not.

The clear demarcation of what is considered science, and what is not, is part of a controversial issue usually referred to as a *demarcation problem* (Hansson 2017). In a scenario where the intent is to produce new knowledge from a pre-existent foundation of knowledge by using large crowds of participants in scientific investigation, it becomes even more critical establish clear criteria for demarcation and for separation of what is known from everything else. This process of building new knowledge from a pre-existing foundation of what is considered to be true is called *scientific learning*. Scientific learning occurs as a result of two specific requirements.

The first requirement is that by definition scientific learning occurs by application of the principles of the scientific method. These principles show that, despite its power, science is indeed a simple tool. In science, we rule out things considered false based on hard evidence. What is true (truthfulness) is then inferred by exclusion[19] (Gauch 2003) (Horgan, (b) 1993) (Horgan, (a) 2016). We refer to this process of inference as the *modern scientific method*, outlined by a set of six principles (Munir 2010):

- The goal of scientific investigation should be to gain objective knowledge (Faleiro Jr and Tsang 2016a).

- Scientific knowledge is obtained through tests, experiments and observations. Tentative assumptions about a particular phenomenon may, however, be deduced from pre-existing knowledge.

- A hypothesis must be verifiable by some experimental or observational method.

- Experiments must be reproducible and must have controls

- The integrity of the data must be appropriately safeguarded.

In the modern scientific method, "each principle helps to increase the reliability and accuracy of knowledge resulting from scientific research" (Munir 2010). By that definition, these principles naturally address the requirements for objective scientific learning in economics described previously.

[19] Discounting the recent resurgence of the truth-conducive controversy, in which "it is fashionable among (…) some philosophers to say there are no principles of rationality that are truth-conducive (…) since there are no standards of rationality, there is no logic or method to science" (Gauch 2003).

The second requirement is that scientific learning is dependent on peculiar characteristics in a field of study. In our case, the field of economics, the process by which knowledge is acquired is defined, and dependent, on three specific peculiarities of the subject of study:

- **Complexity**: modern economics deals with a unique subject of study - a shared, intertwined, complex market – that cannot be rewound. Time like life moves towards one direction (Tsang 2010). Given the usually large number of inputs to such a complex system and the apparent independence between these input variables, once an event occurs, we cannot derive different futures from what the present currently describes (Faleiro Jr and Tsang 2016a).

- **Lack of proper theoretical models**: when taken from a recent historical perspective modern economics has been associated with compartmented classical fields like psychology, statistics, sociology, and computer sciences. Most of the assumptions in classical and theoretical sciences are inherently oversimplified and flawed when trying to predict or understand the behavior of a systemic market[20] (Tsang 2010) (Faleiro Jr and Tsang 2016a).

- **Multidisciplinary fields of study**: modern economics is, in essence, a multidisciplinary subject. Efforts to understand the market considering its most fundamental structures tend to rely on somewhat orthogonal fields of study like neuroeconomics (Camerer, Lowenstein and Prelec 2005), behavioral sciences (Camerer and Loewenstein 2002), and analysis of market micro-events (Madhavan 2000), among others. The interdependency of subjects in economics to bioengineering, neurosciences, social sciences, psychology, data and computer sciences, and related fields is diffuse and

[20] While we consider important to highlight this peculiarity, evaluating reasons for such limitations, or trying to entirely refute or confirm them is beyond the scope of this research.

difficult to correlate and at the same time, critical for scientific learning (Faleiro Jr and Tsang 2016a).

As a consequence of these peculiarities of our field of study (i.e., systemic complexity, lack of proper theoretical models, and novelty of correlated fields of study) modern research in economics becomes strictly dependent on high-performance computers[21], requiring the implementation of elaborate simulation-based techniques. Similar to what is used in other hard-sciences, such as physics, engineering, and biophysics (Chen 2005) (Kay-Yut 2006) (Faleiro Jr and Tsang 2016a). This dependency on high-performance computing has driven research in economics to favor specialized techniques for storage and processing speed. The field has been shaped so that the sheer generation of data and obscure ways to represent computational procedures is prioritized over proper control.

The process of scientific learning in economics is highly dependent on these limitations. The rigor of the study in economics as a hard science would require to store, share, and replicate results and methods of experiments across a vast community of participants. Additionally, the complexity of financial use cases is increasing and as a consequence requiring more storage and computing power. It is now usual to have the results of everyday experiments to grow into massive datasets and the description of concepts that are inherently related to the domain of economics to be only adequately decipherable by computer scientists.

This observation is a consequence of the fact that, in order to properly leverage more computational resources, the description of financial models had to

[21] A correlated consequence is that an ever-increasing dependency on high performance computers for scientific investigation makes it difficult to discern subjects that are specific to economics or computational finance. In other words, there is an incentive and a justification for economics and computational finance to have a significant overlap.

become more cryptic, thus requiring specialized computational skills that often the economics researcher does not possess.

The consequences of the absence of proper computational representation, and specifically its impact on the process of scientific learning are well documented, and it is beyond the scope of this research to list them all in details. This research identifies this and several other observable manifestations of disruption of current scientific procedures later in this thesis, in Section 3.3.3. These manifestations are an indication of a required change on current conventions for scientific investigation, making proper representation to become a priority for investigation and regulation in economics, other sciences, and even outside of academia in public and private institutions (Faleiro Jr and Tsang 2016a).

One of the cornerstones of this research is the realization that computing power is required during an investigation in economics to generate and store massive amounts of data and perform intensive computations, but if computing power is important in this scale, proper representation is indispensable. The limitation factor for scientific learning, and the consequential scientific advancement, has clearly shifted from availability of storage and computational resources to proper representation of investigative procedures (Ioannidis, Allison, et al. 2009) (Camerer, Dreber, et al. 2016) (Nuzzo 2014) (Reinhart and Rogoff 2010) (Herndon, Ash and Pollin 2013) (Cassidy 2013).

Simplification and streamlining of the representation of financial models is imperative to allow unquestionable transparency in the way raw data is obtained and stored, and corresponding results are modeled, used and calculated (Faleiro Jr and Tsang 2016a). This streamlined representation, called in the scope of this research a computational representation, will be described in details in Section 3.4 and Chapter 4 of this thesis.

3.2.3. SCIENTIFIC SUPPORT SYSTEMS

The modern scientific investigation cannot be performed without computers, especially when the agents of the investigation are participants in a crowd. In this section, we define a *scientific support system*[22], based on concepts presented previously in Section 3.2.1 and Section 3.2.2. A scientific support system is a specialized form of a workflow and data management system designed specifically to compose and execute a series of computational or data manipulation steps compatible with the scientific principle (Goecks, Nekrutenko and Taylor 2010) (NIH 2000) (Faleiro Jr and Tsang 2016a).

The modern investigation requires the handling of large volumes of data and the analysis of complex relationships between different financial models, each representing a relative simplification of real-world phenomena through experimentations (Press 2013). Each of these experiments may potentially yield a volume of data that would be impossible to understand and analyze by hand, hence the need for a specialized system to automate the steps of the investigation.

This scenario makes computational resources indispensable and virtually never enough considering the increasing complexity of the studies at hand. This research identifies this and several other observable manifestations of disruption of current scientific procedures, analyzed later in this thesis in Section 3.3.3. These are examples of the hard consequences of the lack of proper control. They provide the realization that, contrary to the good intent, computing power without proper controls is detrimental to science. This status quo establishes the idea of this research that, if

[22] Some literature refers to this specific class of systems as scientific workflow systems (Curcin and Ghanem 2008). In the scope of this research, to avoid confusion with those systems, dedicated to generic workflow management, we use a specific denomination of *scientific support system* (J. M. Faleiro Jr 2013a).

computing resources are indispensable, even more so is the idea of *computational controls*.

Computational controls in a financial investigation are enforced in scientific control systems by *control methods*. We define four specific required control methods: accessibility, reproducibility, communication, and collaboration.

- **Accessibility**: a computational representation must be *accessible*. The use of high-performance systems demands computer literacy from scientists, what is not always possible. There is a need for an accessible computational representation that shields the inherent complexities of modern computer systems from collaborators in any scientific field, focusing on simplicity, and therefore not requiring specialized computer literacy (Faleiro Jr and Tsang 2016a);

- **Reproducibility**[23]: The term *reproducibility* relates, in the scope of this thesis, to *reproducible research* and the assumption that reproducible scientific procedures apply "not to corroboration, but to transparency" (Goodman, Fanelli and Ioannidis 2016). The term reproducible procedures in modern literature is associated with a software platform and procedures that allow a researcher to understand a processing trail of a scientific product, from raw data, to figures, text, and tables (Claerbout and Karrenbach 1992). The same assumptions apply to reproducibility of non-stationary systems in general and economics in particular (Baiocchi 2007) (Hamermesh 2007) (Koenker and Zeileis 2009). This research, however, recognizes the potentially disruptive aspects of technology and computational power when

[23] Reproducibility, by definition, "refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, the second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results (…). Reproducibility is a minimum necessary condition for a finding to be believable and informative" (Bollen, et al. 2015)

these are used without proper controls. An abundance of computational power requires a potentially obfuscated representation of ways to transform information, yielding massive amounts of data. The paradoxical condition in which modern investigative procedures are entangled requires more computing power. More computer power enables to transform more data, and as a consequence, a higher risk of uncontrolled models and massive amounts of untraceable data that will, in turn, require more elaborate techniques and computing resources to trace and decipher that data (Goecks, Nekrutenko and Taylor 2010). Reproducibility of scientific procedures cannot be achieved without proper controls around provenance, and versioning of data and models (Faleiro Jr and Tsang 2016a);

- **Communication**: the massive amounts of computational inputs and outputs, as well as ways to transform the latter into the former, have to be adequately represented. Ways to *communicate* results through proper visualization and computational representation is crucial. The amount of data generated as input and output cannot be represented to humans the same way as they are to computers (Tufte 2006). To make research truly useful, we need human-friendly ways to visualize evidence. At the same time, communicating methods and procedures, regarded as of greater importance than explanatory texts and figures as experimental outputs (Schwab, Karrenbach and Claerbout 2000) (Gentleman 2005) cannot be addressed differently than other items that require human visualization (Faleiro Jr and Tsang 2016a);

- **Collaboration**: *collaboration* is allowing results from one experiment to be seamlessly utilized by other experiments, allowing extensions on models and data to fit additional scenarios, at the same time tracking ownership of each revision or improvement. Collaboration can occur only by exchange of

artifacts[24] that are traceable. Ways in which artifacts are produced and utilized have to be transparent to the overall community.

By definition or for the sake of purpose, scientific support systems have to be designed based on principles of the scientific method, described earlier in Section 3.2.2. By drafting a parallel between the control methods defined above, and the scientific principles in Section 3.2.2, we propose a framework for investigation conceived around the following drivers (Faleiro Jr and Tsang 2016a):

- Providing an easy-to-use environment for individuals themselves to create and test their hypotheses (models);
- Providing interactive tools for individuals, enabling them to execute their hypotheses and tests (scenarios) and view their results in real-time;
- Simplifying the process of sharing and reusing contributions (shareable evidence of investigation procedures) among individuals;
- Enabling individuals to track the provenance[25] of results and use the record of provenance to reproduce those results.

Given these drivers, a platform must provide adequate controls in terms of definition and testing of a hypothesis, as well as transparently tracing and safeguarding of the underlying data as scientific evidence. From that, the principles driving the definition of scientific support systems specifically:

- Allow the definition of a theoretically driven hypothesis;

[24] Artifacts are not synonym of evidence. Artifacts relate specifically to "observations in a scientific investigation or experiment that is not naturally present but occurs as a result of the preparative or investigative procedure" (Oxford University 2010)

[25] Chronology of the ownership, custody or location of historical entities (Oxford University 2010)

- Allow a hypothesis to be tested;
- Allow a hypothesis to be reproduced and verified by independent parties;
- Allow assumptions about a hypothesis to be deduced from historical data;
- Safeguard historical data.

We are calling the class of platform supporting these requirements a *scientific support system*. These drivers are the foundation to define the computational representation described later in this thesis in Section 3.4. Throughout the entirety of Chapter 4, these drivers are addressed again for the definition of computational representation specific for the field of economics.

3.3. LARGE-SCALE COLLABORATION

As we have explained in Section 3.1, this research advocates for a crowd-based method of scientific investigation based on the existence of cognitive and non-cognitive enablers. In this section, we outline requirements and justification for the second cognitive enabler for crowd-based investigation: *large-scale collaboration*.

The use of crowds for resolution of problems follows one of two distinct approaches. The first approach, named "wise crowds" (Surowiecki 2004) relies on empirical observations (K. Wallis 2014) (Galton 1907) and assumes the existence of some invisible, unquantifiable mechanism, somehow providing a certain level of knowledge to crowds, therefore allowing them to make wise decisions. The "wise crowd" approach relies on the assumption of complete independence and decentralization between participants of a crowd. The second approach, named *collaborative crowds*, assumes that knowledge is produced as a result of structured collaboration between participants of a crowd.

This research subscribes to the second approach, collaborative crowds, where large-scale collaboration occurs by the existence of particular requirements of collaboration, as a natural evolutionary response to the environment in which investigation takes place.

Over the following sections, we will list well-accepted requirements for large-scale collaboration, discuss evidence on the use of large crowds for the resolution of complex problems, provide a historical perspective corroborating this our view that crowd-based investigation is a natural evolutionary response to changes on culture, environment, and available technology.

3.3.1. REQUIREMENTS

The use of crowds as agents of investigation requires an organization of large number of individuals, in different roles and at different levels of technical understanding, to continually collaborate for the resolution of complex problems. However, as we can readily ascertain by observation, collaboration does not come out of thin air. We need something to drive effective collaboration, and in this section, we concentrate on explaining requirements for effective collaboration to take place.

Collaboration is what builds "some sort of a collective brain with the people in the group playing the role of neurons" (Nielsen 2012, 18) (Surowiecki 2004) and ultimately amplifies the intelligence of a group of people. Collaboration is facilitated as a result of four *requirements*: expert attention, proper cultural and intellectual development, manufactured serendipity, and human diversity.

- **Expert Attention:** Maximizing collaboration is primarily a problem of restructuring expert attention by designing the correct incentives that would encourage any single participant in a crowd to play the role of an expert at

times, whenever it is required. Given the over-specialized nature knowledge and the narrow window of expertize, these experts in crowds are called micro-experts. Being able to call "the attention of the right expert at the right time" is critical to problem resolution by crowds of individuals. "Expert attention is to creative problem solving what water is to life: it's the fundamental scarce resource" (Nielsen 2012).

- **Proper Cultural and Intellectual Development:** Collaboration must rely on participants in a proper stage of cultural and intellectual development. In the upcoming Section 3.3.3, we describe the historical evolution of the scientific process: from individual macro-experts to institutionalized science to the anticipated, next phase of crowd-based investigation. As part of this evolution, we start to notice evidence of disruption in the current discovery process based on hierarchies and institutions, and the transition to a new phase in which discovery is driven by crowds and micro-experts. This disruption and transition is discussed in details in the upcoming Section 3.3.3.

- **Manufactured Serendipity:** Collaboration requires the right participant with the ideal amount of micro expertise to help in the resolution of a problem. This phenomenon called manufactured serendipity, allow for fortunate discoveries of possible opportunities for collaboration by quasi-accident. Serendipitous connections between individuals are known to be essential in any creative or investigative work. The thinking behind "manufactured serendipitous connections" (Udell 2002) assumes connections between individuals cannot be fabricated, but the conditions by which they occur can be stimulated on purpose. In other words, "you can't automate accidental discoveries, but you can manufacture the conditions in which such events are more likely to occur" (Udell 2002).

- **Cognitive Diversity:** The last requirement calls for cognitive diversity. Cognitive diversity is the "extent to which a group of people reflects differences in knowledge, including abstract constructs like beliefs, preferences, and perspectives" (Miller, Burke and Glick 1998). Collaboration groups must be cognitively diverse, so to maximize instances of micro-expertise amongst its participants (Mitchell and Nicholas 2006). Putting it differently, to maximize collaboration, participants need a wide range of non-overlapping expertise. The minimum amount of shared knowledge must be the level that would allow participants to communicate effectively (Surowiecki 2004) (Nielsen 2012).

These requirements reflect the need for collaboration in investigative procedures that often are cross-disciplinary. Existing literature has identified by empirical methods a different set of requirements, but in a perspective that seems influenced by thinking on that specific field of study (e.g., in social sciences these same requirements for collaboration are outlined as process, understanding, utility, and knowledge integration (Jeffrey 2003)). As a common limitation, other instances in the literature lack quantitative metrics to show evidence of the importance of each prospective feature in collaboration. We call *collaboration metrics* the quantification of requirements for large-scale collaboration, and provide a roadmap for future research of this subject later in this thesis, in Section 6.6.

3.3.2. Collaborative Resolution of Complex Problems

As described in Section 3.1, this research considered particular assumptions as a starting point to the definition of the specific enablers of a crowd-based investigation. One of these assumptions is that organized human collaboration is well suited for the investigation and resolution of complex problems. In reality, it would

be impossible to infer absolute suitability of large-scale collaboration for the resolution of complex problems. Alternatively, we can enumerate results from empirical exercises, and their specific details, as evidence of resolution of complex problems by crowds.

The first example, the Polymath Project (Gowers and Nielsen 2009) is a brainchild of Fields Medal winner Timothy Gowers, a mathematician at Cambridge University. The Polymath project started from a pair of simple posts on his blog. The first post inquiring on the possibility of the use of crowds in the resolution of complex mathematics problems (Gowers 2009a), and then shortly after that a second post where Gowers proposed a particular problem to be resolved using massively collaborative investigation (Gowers 2009b).

Over the next 37 days, 27 people from around the globe – from mathematics enthusiasts to high school math teachers, and other Fields Medal winner Terence Tao – wrote 800 comments and more than 170,000 words on erratic movements of discovery (Nielsen 2012) following an open path of investigative try-and-error. After those 37 days, Gowers announced that the crowd had solved not only the original problem but also a harder, more generic problem that had the initially proposed problem as a special case (Polymath 2012)[26].

Considering the requirements for large-scale collaboration introduced in Section 3.3.1 the original problem was not proposed on the most appropriate platform for collaboration – basically, a sequence of textual comments on Gower's online blog – and the specifics of the methods of incentive for micro-specialists was not clear. Despite that, over the following months around a dozen of unresolved problems were proposed and resolved by a crowd of mathematics investigators, and the platform was

[26] The published author of the paper "D. H. J. Polymath" is a reference to the proposed problem, a new proof of the Density Hales-Jewett theorem, and to the crowd that took part in the resolution during the Polymath project

moved from an online blog to a wiki (Nielsen 2011). Despite lacking an adequate computational representation for investigation in mathematics, the Polymath Project is a successful example of investigation and resolution of specialized, very complex problems by crowds.

The second example is the control of predatory publishing in academia. Predatory publishing is a term popularized by Jeffrey Beall to refer to journals that charge huge fees to submit papers without proper peer review. Predatory publishing damage the scientific process by cheapening intellectual work and misleading scholars, especially early career researchers.

Beall created the list in 2008 (Beall 2008), and from 2010 to 2014 alone the size of the list increased ten-fold, growing to include thousands of journals and publishers (Shen and Björk 2015). Inclusion on the list was based on a metric derived from a 52-point criterion that Beall created himself (Beall 2015).

The list was controversial, mostly due to Beall's biases and previous positions against the open-access movement he described as "anti-corporatist, oppressive and negative" (Beall 2013), or strong statements in the lines of "predatory publishing damages science more than anything else" (Beall 2016). Despite the controversy, evidence points to the fact that Beall's list highlighted recognized problems in academia, and set to worsen (Shen and Björk 2015) (Bohannon 2013). Other studies point to the additional fact that the issue is strongly regional, and expected to worsen even further, as scientific research turns into a global endeavor (Seethapathy, Kumarand and Hareesha 2016) (Omobowale, et al. 2014) (Shen and Björk 2015).

On January 15th of 2017, Beall took his site and the list down, due to "threats and politics" (Straumsheim 2017).

Private initiatives swiftly took on to seize the opportunity and fill the void (Anderson 2017) through centrally managed "black" and "white" lists. As it is usually the case with centrally managed initiatives, it ignores "local knowledge"[27] (Hayek 1945) and fails to address the causes of the negative phenomena[28].

At that point, a community-based initiative, called "Stop Predatory Journals", ran by an anonymous community, took on the maintenance of the original list and extended it. The initiative mostly gathers contributions made through a simple configuration management platform and keeps a publicly available list of predatory journals, predatory publishers, hijacked journals[29], and misleading or fake metrics.

Despite a positive impact, this community-based initiative is still open to criticism, but more objectively, a crowd-based initiative has to look primarily at market incentives in large scale. In this sense, predatory publishing can be seen as a market-oriented, rational response to two factors:

- A poor system of incentives currently in place in academia (Crotty 2017);

- Bad funding models. There should be more than just 'author pays' or 'reader pays' models. The actual cost of publication is a fraction of what used to be when these systems were designed. Additionally, other financial costs like peer review are very relative in an environment that relies on a system of incentives for virtual collaboration (Schroter and Tite 2006).

[27] The "local knowledge problem" in economics is often used to explain why the central control of distributed resources (including centrally planned economies) do not work (Hayek 1945)

[28] Even if unintended, there is symbiotic relationship in place - the very existence of a centrally managed, subscription based list is justified by the existence of the damaging practice - that serves as a reverse incentive to ending the practice of predatory publishing altogether.

[29] A hijacked journal is a journal that had either their websites or branding co-opted by a predatory journal or publisher.

In closing, the current status of crowd-based surveillance of predatory journals is positive but fail to address the root causes of the phenomena. The overall solution lacks adequate computational representation for academic content and an aligned system incentives for collaboration, considering the requirements for large-scale collaboration introduced in Section 3.3.1.

In general, empirical evidence shows that collaborative crowds are more appropriate for the resolution of complex problems than conventional methods. Current research, however, is not able to pinpoint the exact reasons, or characteristics, of problems that would be more suitable for resolution by collaborative crowds (Tausczik, Kittur and Kraut 2014) (Brabham 2008). With a few exceptions, current literature lacks a quantitative analysis of the suitability of crowds for the resolution of complex problems (Guazzini, et al. 2015).

As a final observation on the topic, current research in the literature relies on experiments that fail to address the requirements outlined previously in Section 3.3.1 and further explored over the upcoming Section 3.4, namely the lack of a proper system of incentives and the absence of a dedicated, domain-specific computational representation. This realization seems to confirm the novelty of this research.

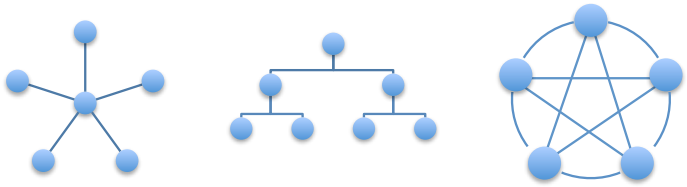
3.3.3. Scientific Evolution and Breakthrough

It should come without surprise to most people that the ability to build objective knowledge through a scientific method is what drove humans out of caves and shot our race towards the stars. As we have previously explained in Section 3.2.2, the scientific method can be seen as a cumulative process in a sense that, over time, we build new knowledge based on previous knowledge considered to be true. Truth, or at least what we perceive to be true, is not constant (Kamen 2014). Given our

history of understanding of the world around us, previous knowledge will almost certainly be ruled as false at some time in the future. On this erratic pathway a "tapestry" of "knowns" slowly evolves to take the infinite space of "unknowns" based on ever-changing knowledge foundations (Krauss 2012).

This dynamic and seemly chaotic method of acquiring an understanding of the world around us has been happening for as long as our ancestors started to grasp with inquiries and guesses about cause and effect of observable phenomena. Even if it was unintended, and we were not entirely aware of its exact mechanisms, this organic adaptation has been happening, constantly. It is so ingenious and so ancient that it has been organically adjusting itself to an ever-changing knowledge base, resources, and culture available at different points in history and time.

This adjustment occurs in evolutionary stages in response to available technology, the individual performing the scientific investigation, drivers, collaboration, creativity, and control in three distinct phases, as described in Figure 4.



	1st Phase	2nd Phase	3rd Phase
Topology	Hub and spoke	Tree	Mesh
Drivers	Macro-experts	Institutions	Crowds of micro-experts
Collaboration	Chance	Planning	Serendipitous design
Creativity	Macro-expert's trait	Parent's trait	Crowd's trait
Control	Macro-expert's trait	Strong	Weak

Figure 4. Phases of Collaborative Scientific Investigation

Characteristics of the evolution of collaborative scientific investigation from macro-experts, to institutions, and to crowds, depending on five factors: the individual performing the investigation, drivers, collaboration, creativity, and control

This chart summarizes each of the phases according to the individual performing the research, topology, drivers, collaboration, creativity, and control. The individual performing the investigation evolved from macro-experts, or "natural philosophers", to groups arranged hierarchically organized in institutions, to an upcoming phase in which micro-experts are arranged in mesh-like crowds, subject to the requirements listed in Section 3.3.1. Each of the topologies represents the communication paths between participants in each phase. The drivers for discovery, or the entity in charge for conducting the investigation and inquiry, in each phase, are macro-experts, institutions, and crowds of micro-experts. Collaboration occurs in each of the phases by chance, central planning, and by serendipitous design, as

explained previously in Section 3.3.1. Creativity in each phase is associated with the individual performing the investigation. Finally, control is either trait depending on the macro-expert, strong, or weak, depending on the topology in place, respectively hub-spoke, tree, or mesh.

The first phase of scientific investigation relied uniquely on "natural philosophers", bright individuals who were able to drive discovery based on their personal traits and occasional interaction with other "natural philosophers".

On this first phase, the domain of investigation was related to natural observations and research conducted by individuals almost in isolation. Given the relative simplicity of subjects under study, a few very bright individuals could still build on previous knowledge with little or no interaction with other individuals. The collaboration was done on an ad hoc basis, and opportunities for interaction were left to chance and rare social exchanges. The proximity with domains of study would allow for self-funding, and the management of resources is de-centralized and done in almost complete isolation. Ultimately, the expertise held by any single individual would determine the effectiveness of one's research – this is the golden age of polymaths or *macro-experts*.

Even with all of these intrinsic limitations, objective learning occurred, and the accumulation of knowledge led to increased complexity and a higher demand for resources to record, store and share knowledge. The ever-increasing body of knowledge demanded a more significant interaction with other researchers that would not necessarily share the immediate surroundings where research was taking place. Large institutions came along to fulfill the demand and manage the vast amount of resources needed for more complex methods, and more information.

That triggered to the second phase, in which institutions took over the task of organization of participants and the management of resources required for investigation and collaboration. As more and more resources were needed as multidisciplinary subjects increased in complexity, this second wave, the phase of institutionalized science, came to life.

On this second phase, participants were organized in hierarchical institutions across diverse kinds of institutions, interested or dependent on scientific advancement: academia, governments, or private corporations. Most scientific procedures evolved to match the hierarchical organization of these institutions, and so the production of objective knowledge followed.

As we move along through the second historical phase, institutionalized science started to shape research methods to fit into a more cumbersome, hierarchical communication, relying on larger, less efficient group sizes. Large institutions also brought along the unintended consequence of heavy top-down hierarchical communication and stronger controls. The immediate consequence overall was that complexity of research domains started to increase exponentially.

This increased complexity has been producing two major changing forces that are shaping the resurgence of a next historical phase:

- **Multidisciplinary collaboration**: multidisciplinary collaboration became mandatory. We cannot perform an objective investigation, on any field, without an understanding of orthogonal fields of knowledge. Unlike the first phase, no single participant, regardless of how bright, detains enough expertise to provide a full, overreaching solution to a modern-day problem. The natural limitation of individual participants of the scientific process in

dealing with an ever-growing knowledge body marks the beginning of the demise of the age of macro-experts.

- **Complexity requires control**: technology is an amplifier of features present in any environment, regardless of how we perceive the results of these features as positive or negative. This amplifying side effect of technology is observable everywhere: in politics, personal and business relationships, financial markets, and especially in our topic of interest: scientific procedures applied to economics. Technology is getting to a level of complexity and sophistication that its scientific use without control plays a role of a double-edged sword: it can cause more harm to the development of objective knowledge than good. One should expect the same scientific method that brought significant technological advancements would naturally improve the tools available for investigation, specifically computational tools. It did so to a certain extent. It is true we have advanced technology and methods available in the scientific investigation, but it is also true that we have abundant evidence of misuse of computational resources and methods in the scientific investigation leading to wrong or corrupt data and as a consequence defective research.

These two forces are bringing several disruptive manifestations as signs of an upcoming wave of transformation. These manifestations, listed over the next paragraphs, are evidence that this second historical phase of the scientific investigation is presenting signs of inadequacy with current status of technology, historical, and cultural developments:

- **Misaligned Academic Incentives:** despite an organized and commendable effort by scientific institutions to contain this harmful practice, evidence

shows that current academic incentives are fostering a culture of fraud. Based on pools and questionnaires, research finds an astonishing number of scientists engaging in a range of behaviors extending far beyond falsification, fabrication, and plagiarism (Martinson, Anderson and De Vries 2005). The issue is so prevalent that quantitative models can reliably predict and estimate the number of articles that should be retracted over time (Cokol, et al. 2007).

- **Hierarchies Stifles Creativity:** scientific research is mostly driven by creativity. As explained earlier in this section, in the current phase of scientific investigation, work is often performed in hierarchical structures. While hierarchies work reasonably well for control and decisions, there is evidence that hierarchies are detrimental to creativity (Burkus 2013) (Burkus 2012). There is also evidence that while hierarchical institutions usually verbally request for innovation, their top-down structures unintentionally reject them (Staw 1995) (Mueller, Melwani and Goncalo 2011) (Mueller 2014). The ideal structure to foster creativity is closer to a peer-to-peer association, the one presented by human crowds, than a top-down hierarchical structure (Muller, Wakslak and Krishnan 2014).

- **Human Limitation on Information Processing:** there are hard limitations on how much information humans can process. The limitation on how much information we can consume and understand also limits the throughput of quality research scientists can produce, review and reproduce (Tenopir, et al. 2015). There is evidence that scientists have already reached a plateau on how much information they can efficiently absorb, handle, and produce (Van Noorden 2014).

- **Unavailability of Quality Academic Content:** major journals have recently placed restrictions on mining and use of scientific data in large scale (Jha 2012). Similar limitations apply to the refusal of providing details on landmark research findings for "reasons of confidentiality" (Yong 2017) (Prinz, Schlange and Asadullah 2011) (Begley and Elis 2012). These restrictions undermine both the automation of hypothesis generation reliant on vast amounts of quality data and crowd collaboration of micro-experts dependent on access to peer-reviewed, quality academic content.

- **Lack of Means to Record and Share Reliable Data:** lack of appropriate means to record and share reliable data has been indicated as one of the limiting factors in modern investigation procedures. An additional limiting factor is the lack of a central authority to validate observations, and a central repository evidence-based knowledge (Wallis, Rolando and Borgman 2013). Other evidence in the field of economics describe examples of global economic policies defined based on flawed data stored in plain excel spreadsheets (Reinhart and Rogoff 2010) (Herndon, Ash and Pollin 2013) (Cassidy 2013).

- **Science Hacking:** evidence collected in a correlated field show a considerable rate of complex biotech experiments published in prominent journals, heavily reliant on advanced computational resources, just cannot be appropriately reproduced (Ioannidis, Allison, et al. 2009). Additionally, evidence shows that reproducibility is negatively correlated with the relative computational complexity of the experiment. In the field of economics, in particular, we have similar evidence (Camerer, Dreber, et al. 2016) measuring that only 61% of the articles in a major journal of economics can be successfully reproduced. Similar results have been found in psychology, in which only 38% of the

studies can be successfully reproduced (Open Science Collaboration 2015), or biotechnology where only 6 out of 53 "landmark cancer studies", i.e., 11%, could be properly reproduced (Begley and Elis 2012). Quantitative metrics also show examples of "statistics used wrong", proliferating the belief that p-values alone can determine findings to the true when in reality they are false (Colquhoun 2016) (Ioannidis 2005).

Modern science and the associated scientific method have taken a critical role in human societies. We have learned to blindly trust lives and outcomes of global reaching economic policies to findings that should be blinded from scrutiny by merely labeling them 'scientific'. If these manifestations listed above sound alarmist, the feeling is rooted in plausible reasons.

Now, here comes time for the third phase of collaborative scientific investigation based on multidisciplinary, diverse collaboration in large-scale through crowds. This research advocates for a crowd-based investigation to serve as an attenuator of the changing forces disrupting institutionalized science.

As described previously in Section 3.1, a crowd-based investigation occurs as a result of the existence of three requirements given as two cognitive enablers and one non-cognitive enabler. The cognitive enablers of crowd-based investigation are methods of proof, described previously in Section 3.2, and large-scale collaboration, described in this section.

Over the next section, we describe the third and last requirement: the non-cognitive enabler referred to as a computational representation.

3.4. COMPUTATIONAL REPRESENTATION

Languages are more than a vehicle for communication. They are often one's windows to reality. A language shapes how a person thinks, what can be achieved, and how can be achieved. Some languages often come from and facilitate the representation of concepts in a specific domain of knowledge, and if used outside of that specific domain, could make the representation of those same concepts more obscure. A person might, for example, use the German language for philosophy, or French for poetry. Using them the other way around might make a person write more, or being forcibly more verbose, or even lose clarity. In extreme cases using the wrong language for a domain of knowledge can impede the expression of the exact ideas one might intend.

Language defines reality (Lupyan and Ward 2013) (Rodriguez-Esteban and Rzhetsky 2008). Human observations that lead to the scientific inquiry, and drive our process of discovery are shaped by our method of questioning, and limited by the language we possess (Heisenberg 1958)[30].

This research explores literature and evidence showing that this is not only the case with natural languages but also with a layer of abstraction that is required to define domain-specific concepts in computers, in a way that these concepts can be shared in a crowd. We are calling this conceptual abstraction a *computational representation*.

Similarly to natural languages, a computational representation often grows from needs of a specialized domain, and therefore is better suited for use cases relevant to that specific domain. In some domains of knowledge, like architectural

[30] "We have to remember that what we observe is not nature itself but nature exposed to our method of questioning. Our scientific work (…) consists in asking questions about nature in the language that we possess and trying to get an answer from experiments by the means that are at our disposal" – Werner Heisenberg (Heisenberg 1958)

sciences, one would be more concerned about spaces, shapes, volumes or colors, and their relationships with a three-dimensional environment and the effect of the interaction of those concepts with humans. In legal sciences, one would be more concerned about possible associations between real-world entities, and rules defining their behavior and constraints for interaction. In some other domains, like bioinformatics, the ability to represent interconnected shapes and strings could be more relevant. In biophysics, it is essential to keep track of genotypical and phenotypical traits, and their relationships with encoded protein sequences with a vast number of possible combinations. In economics, our subject of concern, a researcher would be more interested in the way changes in quantitative measurements, over a time series, would affect the valuation.

A computational representation must mimic the inherently free flow of thoughts of the human mind and the speed of modern vehicles of collaboration and therefore, by similarity, a computational representation must be fluid. In contradiction, computational artifacts, like programming languages and databases, are born out of strictly technical aspects of a problem and bred outside of concerns relevant to specific domains of knowledge. Only after definition, they are forcedly introduced for use and therefore not able to follow the free-flow of the evolution of ideas. Computational artifacts remain frozen to domain-specific requirements of that specific point in time when the introduction occurred. When requirements on that domain evolve to follow the increasing complexity of the problems at hand, those artifacts would no longer fit, or in a best case require an additional verbosity, sacrificing the proper semantics of communication.

In opposition to computational artifacts, a computational representation must be dynamic, able to adapt and evolve to solve new classes of problems and organize increasingly complex and powerful computing environments. These new classes of

problems are different from problems we had to deal with just a few years back. They require the collaboration of multi-disciplinary specialists exchanging different types of artifacts that must be adequately described and tracked (Faleiro Jr and Tsang 2016a). An investigator must have adequate tools and methods to approach new problems correctly. On this sense, an adequate computational representation allows for the proper description and control of those tools and methods, allowing them to change in the face of new demands and be able to address new problems (Faleiro Jr and Tsang 2016).

Unfortunately, as previously described in Section 3.3.1, the status quo in exploratory research in economics defines a different reality. The lack of adequate representation and an abundance of computational power allow, and unintentionally require, a potentially obfuscated representation of ways to transform and store data, yielding massive amounts of convoluted and dissociated information. This paradoxical condition entangling modern investigative procedures define a vicious circle. Uncontrolled methods require more computing power, which enables to transform more data, which as a consequence bring an incentive for uncontrolled models and increasing amounts of untraceable data. These, in turn, require more opaque techniques and computing resources to trace and decipher that data, the "informatics crisis" (Goecks, Nekrutenko and Taylor 2010).

The way in which an investigator describes to an increasingly complex and powerful binary being a method to resolve a problem plays a fundamental role in communication and collaboration, and as a consequence in the traceability of the process of investigation and discovery. The amount of data generated in modern investigative procedures as input and output cannot be represented to humans the same way as they are to computers (Tufte 2006). To make research truly useful, we need human-friendly ways to visualize, track, store and understand the evidence. In

addition to representing evidence, communicating methods and procedures - regarded as of greater importance than explanatory texts and figures as experimental outputs (Schwab, Karrenbach and Claerbout 2000) (Gentleman 2005) - cannot be addressed differently than other items that require human visualization and interpretation.

3.4.1. REPRESENTATIONAL PROCESS

One of the cornerstones of this research is the assumption that a computational representation is defined by, and tightly coupled to, a specific domain of knowledge.

Given the intrinsic association of a computational representation to a domain of knowledge, it would be natural to expect that a computational representation could be derived from a domain of knowledge, given a set of well-defined inputs and general procedures. We are calling this organization of inputs and general procedures to produce a computational representation a *representational process*.

The outline of a representational process to define a computational representation for any domain of knowledge is described in Figure 5.

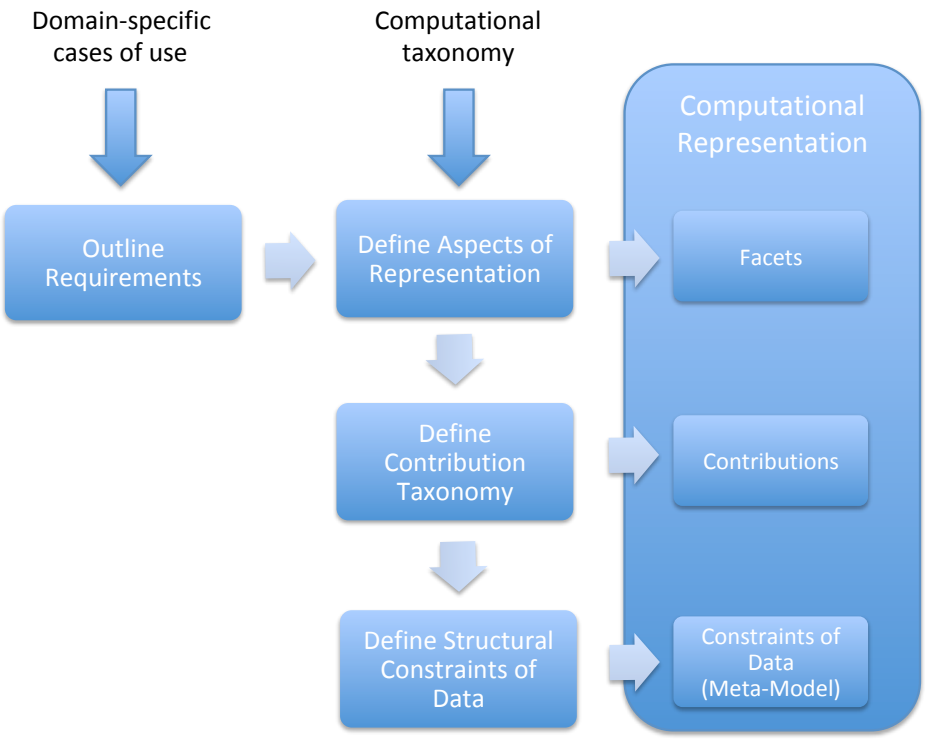


Figure 5. Representational Process

An outline of a generic process to define a computational representation for any domain of knowledge on four steps and two inputs: a list of exercises, or domain-specific cases of use, and a computational taxonomy.

The outline of the representational process defined in Figure 5 shows a set of two inputs and four distinct steps that are necessary to generate a computational representation composed of facets, contributions, and constraints of data. Arrows define the flow of data, and as a consequence, dependencies for the execution of a given step. An example of an actual application of the representational process, to produce a computational representation for the field of economics, is given later in this thesis in Chapter 4.

The two inputs for a representational process, represented on the top of the diagram, are the entry points for the representational process. The inputs are a set of domain-specific cases of use and a computational taxonomy.

Domain-specific cases of use are a collection of exercises reflecting specific characteristics of concern in that domain of knowledge. The selection of cases of use should represent an overreaching and diverse sample of the main activities relevant to that domain of knowledge. Each case of use defines the domain-specific knowledge necessary for that specific scenario to be understood and executed. We assume that the set of cases of use selected is representative enough to cover most of the scenarios of investigation relevant for that specific domain. A practical example of domain-specific cases of use is given in Section 4.2, while we produce an outline of requirements for the field of economics.

A computational taxonomy is an inventory of computer technologies available and relevant for the implementation of cases of use at that moment in time. Examples of items in a computational taxonomy are technologies to store, retrieve, analyze, and visualize data and computational methods. A computational taxonomy is fluid, in a sense that the exact definition of what is relevant is affected by qualities of the individual using this process, such as experience, and personal biases. A discussion on the non-deterministic nature of the process, concerning a computational taxonomy, is given in Section 3.4.5. An example of a computational taxonomy for the field of economics is given later on this thesis, in sections 4.3.1.1, 4.3.2.1, 4.3.3, and 4.3.4.

The four distinct steps of the representational process defined in Figure 5 are shown in individual solid boxes: outline requirements, define aspects of representation, define contribution taxonomy, and define structural constraints of data. Incoming arrows in each box define dependencies, and outgoing arrows define products, or results, of the execution of that specific step.

The first step outlines requirements that are relevant for the definition of a computational representation for a domain of knowledge. The outline of requirements is produced from a list of domain-specific cases of use, defined based on relevancy. Relevancy is given by, as we have mentioned before, the assumption that the set of cases of use is representative enough for most of the scenarios of investigation in that domain of knowledge. If the assumption is valid, we can infer as a consequence that any investigation exercise on that domain should depend, at least in a substantial part, with a combination of one or more of those requirements. For reasons of completeness, a proper computational representation for that domain of knowledge must address all these requirements. As a result, by definition, what we call a *proper computational representation* for a domain of knowledge should intend to represent all cases of use in the scope defined by the original list of cases of use[31].

The second step define aspects of representation based on the computational taxonomy and the domain-specific requirements produced as a result of the first step. The result of the second step is the set of facets of a computational representation. Facets and the specifics entailing their definition are described in details in the upcoming Section 3.4.2. An example of an execution of the step to define aspects of representation is given in the upcoming Section 4.3 when we describe facets and the process of their definition for a computational representation in economics.

The third step defines a contribution taxonomy based on facets produced as a result of the second step. The results of the third step are contributions of a computational representation. Contributions and specifics entailing their definition are described in details in the upcoming Section 3.4.3. An example of the step to define contributions is given in the upcoming Section 4.4 when we describe

[31] An example of an execution of the outline requirements step is given in the upcoming Section 4.2 when we define requirements for a computational representation for the field of economics.

contributions and the process of their definition for a computational representation in economics.

The fourth step defines structural constraints of data based on facets and contributions produced as the result of the second and third step. The results of the fourth step are constraints of data, or meta-model, of a computational representation. Constraints of data and specifics entailing their definition are described in details in the upcoming Section 3.4.4. An example of the step to define contributions is given in the upcoming Section 4.5 when we describe constraints of data and the process of their definition for a computational representation in economics.

The final result of a representational process, as shown in Figure 5 by a larger solid box on the right side, is a computational representation given by facets produced in step two, contributions produced in step three, and constraints of data produced in step four. Each of the components is depicted in Figure 5 as smaller boxes inside the computational representation. Facets, contributions, and constraints of data are detailed respectively over the upcoming sections 3.4.2, 3.4.3, and 3.4.4.

3.4.2. FACETS

A facet, in the context of this research, is defined as "one of the definable aspects that make up a subject or an object; denomination of things that are similar or related, but yet distinct things" (Merriam-Webster Online Dictionary 2016).

A more intuitive definition of what exactly is a facet is done by example and would come from a domain in which concepts are more tangible and organoleptic than in economics. Intuitiveness, as it is always the case, is achieved by representing concepts that are keen to one or more of traditional human senses.

Taking the domain of architecture, or civil engineering sciences, for example. The representation of ideas is done through the placement of volumetric shapes considering restrictions like light, gravity, and the mutually exclusive placement of objects in space. One example of a typical representation of that domain is shown in Figure 6.

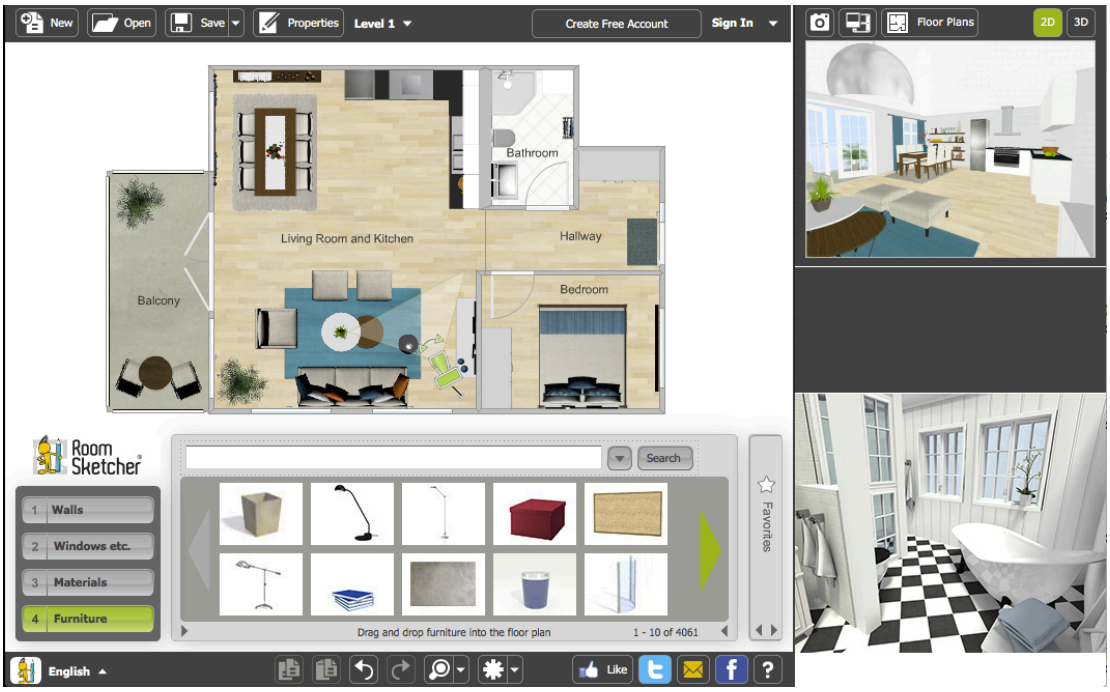


Figure 6. Example of the application of facets to a domain of knowledge

Aspects like volumetric shapes and specific coordinates in a 3D environment (facets) are used to describe a layout relevant to a specific domain of knowledge (architecture)

Three-dimensional shapes, textures, colors, and measurements can be combined to define concepts like pieces of furniture, rooms, ambiance, and then extended to derive in computers notions that can only be asserted at naked eye, anticipating the effect of the interaction of these concepts with individuals.

These primary, fundamental elements that can be combined to generate core concepts on the domain of knowledge are called *facets*.

In this example shapes, texture, colors, and measurements are facets, representable in computers, which make up a subject or an object relevant to that specific domain of knowledge: architecture.

At this point, it is important to emphasize one of the core assumptions of this research: the exact definition of what constitutes a facet in a specific domain of knowledge is empirical. In some cases, e.g., our example related to architectural sciences, the proximity to visual and spatial concepts makes the establishment of what is indeed a facet - shapes, textures, color, and measurements - somewhat intuitive, and as a consequence more natural to derive[32].

3.4.3. CONTRIBUTIONS

In the scope of this research, we call contributions the set of shareable and formal evidence[33] of an objective investigation. As shareable evidence they can be exchanged, reused and traced through something called a record of provenance[34], therefore becoming a vehicle for effective collaboration.

To be qualified as contributions in a crowd-based investigation scenario, any evidence has to carry specific traits: evidential properties, intrinsicality, and characteristics of communication and interaction.

---

[32] For the domain of knowledge of concern for this research, the selection of facets for a computational representation for the field of economics and their formalization is given in Chapter 4, specifically in Section 4.3.

[33] The available body of facts or information indicating whether a belief or proposition is true or valid (Oxford University 2010)

[34] Chronology of the ownership, custody or location of historical entities (Merriam-Webster Online Dictionary 2016)

3.4.3.1. Evidential Properties

To be defined, shared, reused and traced contributions must carry particular mandatory traits we call evidential properties: classification, identification, a record of provenance, and ownership and security (Faleiro Jr and Tsang 2016a).

- **Classification**: Contributions must follow a classification system of shareable entities, specific to the domain of knowledge under consideration, and referred to as taxonomy of contributions. This classification system is an organization of shareable artifacts, organized based on relevant features[35].

- **Identification**: Contributions should be appropriately identified following common standards for shared identification in a way to allow reference, sharing, and ownership (Berners-Lee, Fielding and Masinter 2005).

- **Provenance**: Contributions should carry a record of the chronology of ownership, custody or location of contributions, as well as the history of associations of contributions to entities or participants. We call this chronological description of custody and location a record of provenance.

- **Ownership and security**: Given the sensitive nature of contributions, contributions should ensure ownership and access only after proper authorization and authentication. For that reason, contributions must carry a record of ownership and authorization.

[35] The classification of contributions for the domain of economics is depicted in Figure 19.

3.4.3.2. Intrinsicality

Contribution properties are defined as either intrinsic or extrinsic. Intrinsic properties of contributions[36] are not explicitly described in the representation and are enforced by an implementation of the computational platform. They can be assumed to be in place based on physical aspects of the contribution, regardless of specific indications on the representation. On the other hand, extrinsic properties are explicitly represented.

For example, ownership of each revision or improvement in a contribution occurs without an explicit description in a representation. As a consequence tracking the ownership of contributions occurs by the natural exchange of artifacts that are inherently traceable. In that way artifacts are traced when they are produced and utilized, making the record or provenance transparent (Faleiro Jr and Tsang 2016a)

3.4.3.3. Characteristics of Communication and Interaction

Contributions must carry characteristics to allow collaboration to take place. Collaboration is a direct result of how well contributions foster communication and interaction. A contribution must support three characteristics of communication and interaction to support large-scale collaboration: analytical description, granularity, and simplicity.

- **Analytical description**: Problems must be proposed in a way that allows for an analytical description, following a top-to-bottom structure. Splitting the description of problems into sub-tasks allows micro-expertise to be harnessed

[36] Intrinsic elements of a representation are elements enforced by an implementation of the representation. An intrinsic element can be assumed to be in place regardless of any specific expressions on the representation itself.

more directly and contributions to be naturally generated and associated with solutions.

- **Granularity**: A computational representation should encourage short, small contributions. Small contributions would make simpler and more straightforward for experts to review incoming collaboration and assess if they are relevant to their investigation.

- **Simplicity**: Representation of contributions should be simple and straightforward. A streamlined representation would make it easier to refer to foundational knowledge, as well as making it easier for participants to communicate and describe contributions.

These properties of analytical description, granularity and simplicity allow input and results from one experiment to be seamlessly utilized by other experiments, easing extensions on models and data to fit additional scenarios by short and specialized description.

The contribution taxonomy for the field of economics, listing the relevant properties for that specific domain of knowledge, is defined in Section 4.4.

3.4.4. CONSTRAINTS OF DATA

Most domains express real entities and relationships using structural constraints of data. Those constraints define rules of associations that establish what is feasible in that domain, in the real world.

These rules of associations define structural constraints of data in place for a specific domain of knowledge. Those structural constraints use an abstract layer of

data to define restrictions on a separate layer of abstractions, based themselves on data, hence the term meta-data[37]. The set of structural constraints in a specific domain of knowledge is called meta-model.

Depending on the complexity of the domain of knowledge, and what should be represented, meta-data in a specific domain can follow a classification. A specific meta-model for the field of economics is described in Section 4.5.

3.4.5. DISCUSSION ON ASSUMPTIONS AND CONSEQUENCES

The conceptual layout of a computational representation is, in essence, a proposal to represent knowledge in a given specialized field through abstractions commonly called *models*.

The representation of knowledge through models is not something new. There is a long history of academic work attempting similar tasks in a variety of domains (Hayes 1978) (Davis and Shrobe 1983). However, most works concentrate on a comparative analysis, evaluating properties of specific representations against others.

Alternatively, this research assumes a role-based definition of knowledge representation. In a role-based definition, a description of a knowledge system is defined in terms of five core roles a specific representation plays (Davis, Shrobe and Szolovits 1993) (Faleiro Jr and Tsang 2016a).

- **Models are surrogates**: a surrogate is by definition a substitute for the target idea itself, and as a result, a measurement of how far or how close this surrogate is from calculations it intends to represent is secondary or irrelevant.

[37] Data that provides information about other data (Merriam-Webster Online Dictionary 2016)

- **Models define human expressions**: models should define measurements and concepts understood by humans in a language that is adequate for human consumption, even if not directly natural.

- **Models are a medium for efficient computation**: models are a medium for pragmatic efficient computation, or in other words, models should be able to be replicated in computers given appropriate technology and sufficient resources.

- **Models establish ontological commitments:** models define ontological commitments for a representation by defining "a set of decisions about how and what to see in the world" (Bricker 2016) (Rayo 2007). Models are approximations of reality, and as we define them, we make decisions of what to consider and what to ignore. These decisions are ontological commitments and are "not an incidental side effect but they are of essence in our representation" (Davis, Shrobe and Szolovits 1993).

- **Models define a theory of intelligent reasoning**: Models define a "fragmentary theory of intelligent reasoning" represented in terms of concepts and inferences, sanctioned and recommended. Models represent "some insight indicating how people reason intelligently" about a problem or investigation (Davis, Shrobe and Szolovits 1993).

The use of a role-based definition and these core roles bring important consequences when defining a computational representation for any domain of knowledge:

The first and most important consequence is that computational representations are non-discriminatory. Representations should not be measured on

how efficiently they represent a target idea, and therefore should not be compared to one another. Representations are abstract surrogates for a target idea and as such, they are just a set of decisions of what to see in a subject, and therefore bound to limitations and biases of an observer.

Second, a computational representation and associated models are fluid and not final. Representations are not set in stone and are expected to change whenever noticeable changes in technology bring new methods and tools, or a new case of use becomes relevant for that specific domain of knowledge.

These assumptions and consequences are critical when assessing and understanding features and limitations of any computational representation defined from the representational process defined in Section 3.4.1. These same assumptions and consequences should be expected in any representation, and more importantly, in the case of this research, a computational representation for the field of economics.

Over the following chapters, we apply this process to the domain of knowledge of our interest and will formalize requirements, facets, contributions, and constraints of data to define a specific computational representation for the field of economics in Chapter 4, respectively detailed in Sections 4.3, 4.4, and 4.5.

3.5. CHAPTER SYNOPSIS

This chapter fulfills the criteria of success described in Objective 1, on page 19 of this thesis, by identifying what is required for the adequate use of crowds in structured, scientific investigation.

We use the assumptions listed in Section 3.1 to define cognitive and non-cognitive requirements, or enablers, for crowd-based investigation: methods of quantitative proof, collaboration in large-scale, and a computational representation.

These enablers are formalized in details in three separate sections: methods of quantitative proof in Section 3.2; collaboration in large-scale in Section 3.3; and a computational representation in Section 3.4.

The first cognitive enabler, *methods of quantitative proof*, is required to provide a common framework for a step-wise, algorithmic investigation and standard control points based on statistical methods that apply to a crowd-based investigation. The general framework is defined as something we call a *proof pipeline*, detailed in Section 3.2.1. The discussion around proper methods of proof is extended in Section 3.2.2 and Section 3.2.3 respectively to define the scope of scientific learning in economics, and highlight requirements for a specific class of systems called scientific support systems.

The second cognitive enabler, *collaboration in large-scale*, establishes the mechanics and incentives that must be in place to leverage collaborative crowds. The mechanics and incentives for collaboration are listed in Section 3.3.1. The suitability of collaborative crowds for the resolution of complex problems is discussed in Section 3.3.2. In Section 3.3.3 we lay out the historical and evolutionary perspective to determine the use of collaborative crowds as the foundation for upcoming methods of scientific investigation.

The third enabler, *computational representation*, is a crowd-friendly representation system to define concepts related to a specific domain of knowledge. A computational representation is defined based on facets, contributions, and constraints of data. Facets are definable aspects that make up a subject or an object of a domain of knowledge and are defined in Section 3.4.2. Contributions are shareable and formal evidence of a crowd-based scientific investigation, defined in Section 3.4.3. Constraints of data define entities and rules of associations between entities in a specific domain of knowledge and are defined in Section 3.4.4. A generic process to

define a computational representation for any domain of knowledge is given in Section 3.4.1. This generic process is called a *representational process*.

This chapter brings particular novelty contributions to this research, specifically:

- Definition of specific cognitive and non-cognitive enablers to support structured scientific investigation based on large collaborative crowds.

- The conceptual definition of a computational representation, in terms of its core concepts and components, namely facets, contributions, and constraints of data.

- Definition of a representational process, a reproducible set of steps and procedures to generate computational representations from domain-specific requirements.

- Lining out of a historical perspective of scientific investigation, defining the drivers signaling the upcoming phase of crowd-based scientific investigation as a natural evolutionary process.

The representational process defined in this chapter is used to define a computational representation for the field of economics in Chapter 4. The non-cognitive enablers are used to define the method of proof and steps for investigation in Chapter 5.

CHAPTER 4. CONCEPTUAL FRAMEWORK FOR COLLABORATION AND TRANSPARENT INVESTIGATION IN ECONOMICS

"A good notation has a subtlety and suggestiveness which at times makes it almost seem like a live teacher." Bertrand Russell (Newman 1956)

This chapter defines a specialized computational representation for crowd-based scientific investigation for the field of economics. A computational representation, as previously defined in Section 3.4.1, is a layer of abstraction that is required to define domain-specific concepts in a way that these concepts can be shared with a crowd to allow collaborative investigation.

The computational representation for the field of economics defined throughout this chapter is built based on the representational process defined in the previous chapter, specifically in Section 3.4.1. As explained earlier in Section 3.4, a computational representation is defined by three complementary components called facets, contributions, and constraints of data. In this chapter, we will apply the guidelines of the representational process to define facets, contributions, and constraints of data for a computational representation for the field of economics.

4.1. A COMPUTATIONAL REPRESENTATION FOR ECONOMICS

A computational representation, as defined previously in Section 3.4, is a representation system based on facets, contributions, and constraints of data and used to define concepts related to a specific domain of knowledge, in a way these concepts can be shared with a crowd to allow controlled investigation in large-scale.

A computational representation can be defined for any domain of knowledge by following the steps of the representational process defined in Section 3.4.1.

According to the representational process, a computational representation can be generated for any domain of knowledge given a set of domain-specific requirements and a computational taxonomy. The set of domain-specific requirements for the field of economics is defined over the following Section 4.2, and the computational taxonomy is presented as we describe each facet, on sections 4.3.1.1, 4.3.2.1, 4.3.3, and 4.3.4.

4.2. DOMAIN-SPECIFIC REQUIREMENTS FOR ECONOMICS

According to the representational process defined in Section 3.4.1, a computational representation is built based on a set of domain-specific requirements selected by careful examination of specific features of a number of domain-specific cases of use.

Each case of use defines the knowledge necessary for that specific scenario to be understood and executed. For the definition of domain-specific requirements that will be used for the definition of a computational representation for economics, each case of use is a separate empirical exercise, listed as follows:

- Assessing the performance of momentum cross-over strategies using Monte Carlo simulations and historical backtesting[38] (Faleiro Jr and Tsang 2016)

- Simulation of the performance of real-time strategies through backtesting (J. M. Faleiro Jr 2015)

- Profitability of different moving average cross-over strategies (J. M. Faleiro Jr 2014)

- Real-time valuation of an equities portfolio (J. M. Faleiro Jr 2014)

[38] This exercise is the baseline for the exercise described over the upcoming Chapter 5.

- Assessment of profitability of strategies holding long positions on fixed-length intervals (J. M. Faleiro Jr 2013).

- Agent-based simulation of a central-limit order book (Foata, Vidhamali and Abergel 2011) (Panayi and Peters 2015) (Cont, Stoikov and Talreja 2010) (Murat, et al. 2009) (Chakraborti, et al. 2011) (Chan and Shelton 2001)

Some of these exercises are extensive and relate to novelty research subjects. Each one of those exercises expresses specific behaviors, later translated to an outline of features for proper representation of financial models, and as a consequence, requirements for a computational representation for the domain of economics. With that, the requirements for a computational representation for the field of economics as listed as follows:

- **Simplicity of communication**: a financial model is seldom defined and interpreted by one single group of users. The notation for its description should be simple enough to allow communication across a diverse community of users;

- **Predictability**: financial models are often defined with the intent of anticipating behavior or critical events;

- **Complexity of the domain of knowledge**: financial sciences deal with subjects that are inheritably complex and challenging to model;

- **Large volume of data**: virtually infinite history associated to a record of time: The record of the memory of financial models is associated with either datasets or streams of data that are virtually infinite.

- **Sliding window computations**: a sequence of fragments of data has to be evaluated so that adjacent members in the sequence, fitting a constant sliding time window, are relevant for the computation of a result[39].

- **Low latency**: responsiveness in near real-time. Given an event, or stimuli, some cases of use most respond as quickly as possible to avoid penalizing accuracy of measurements and profitability of the model itself;

- **Event-driven**: actions respond to events, originating from external and unpredictable sources;

- **Time-based**: tightly coupled with notions of value variations (e.g., prices, ratios) over discrete time series;

- **Graph-oriented**: financial models strongly rely on real-world entities and their ad-hoc relationships. Entities are associated with nodes and relationships to edges in graph-oriented representations. The sequence of transformations and steps to operate on real-world entities, either sequentially or not, is also graph oriented

We assume that the set of use cases is representative enough for most of the scenarios of investigation in economics. If the assumption is valid, we can infer as a consequence that any financial model should depend, at least in a substantial part, with a combination of one or more of those requirements. For reasons of completeness, a proper computational representation for the field of economics must address all these requirements. In this sense, by definition, a proper computational representational should intend to represent all cases of use in the scope of economics.

[39] Examples are a sequence of prices, in which a specific algorithm tracks features of price variations over different time windows, e.g., during the last hour, a day, a week. Different windows can be compared with adjacent or non-adjacent windows for identification of useful patterns.

As we have mentioned previously in Section 3.4.1, a computational representation is defined in terms of components called facets, contributions, and constraints of data. The definition of the procedures for the generation of each component was given previously in sections 3.4.2, 3.4.3, and 3.4.4. Over the following sections, we apply those procedures to define facets, contributions, and constraints of data for the computational representation for the field of economics.

4.3. FACETS

A facet, previously described in Section 3.4.2, is a definable aspect that makes up a subject or an object in a domain of knowledge. Ideas in a domain of knowledge are expressed using a composition of facets, and together with a taxonomy of contributions and constraints of data, define a computational representation for that domain of knowledge.

The definition of a facet can be understood intuitively. On the example we provided previously in page 86 in that same Section 3.3.3, we show that ideas and concepts in architectural and building sciences are tangible enough to allow for an almost immediate definition of facets relevant for that domain of knowledge. The proximity of ideas on architectural and building sciences to human senses make the definition of facets more intuitive.

On the other hand, the definition of ideas and concepts in financial sciences is mostly non-spatial, and as a consequence, the designation of facets on our specific case not as intuitive. In financial sciences, a researcher would be more interested, for example, in the way changes in quantitative measurements, over discrete time, would affect the price. These are abstract concepts, and as a consequence, it is hard to describe them through concrete, tangible similarities.

According to the representational process defined in Section 3.4.1 facets are defined based on two inputs: intrinsic requirements of a domain of knowledge and a computational taxonomy. The requirements for a computational representation for economics were previously defined in Section 4.2. A computational taxonomy, as previously defined in Section 3.4.2, is an inventory of computer technologies available and relevant for the implementation of domain-specific cases. The specific computational taxonomy in use is explored during the definition of each facet when we explore technology alternatives.

Over the next sections, we detail the exercise to find out the relevant set of facets for our domain of knowledge: economics. For that, we formalize the four facets required for the definition of aspects that make up subjects and ideas in the field of economics: streaming, reactives, distribution, and simulation.

4.3.1. STREAMING

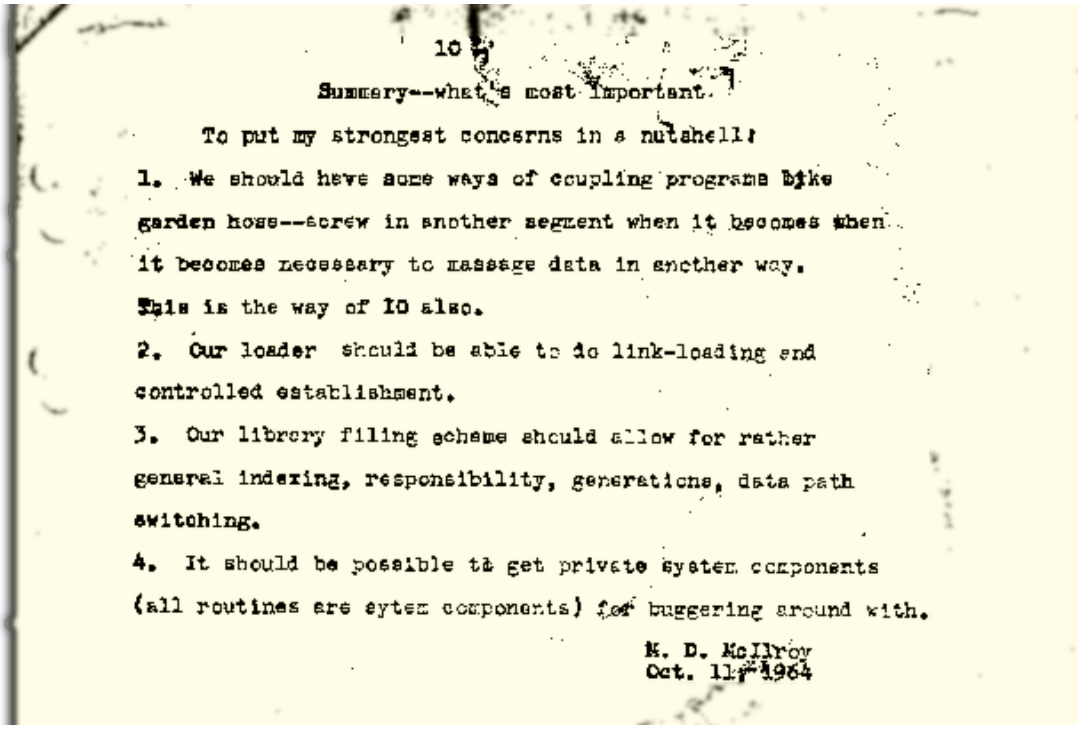
The original idea of streams, put merely, starts with a vision of a graph in which nodes are processors and edges are communication paths. Each node holds incoming and outgoing communication paths to other nodes in the graph. The basic idea of streaming relates to continuous sequences of data fragments traveling over communication paths, in which each node executes specific tasks upon arrival of fragments of data.

Streams are traditionally used in domains where concurrency and speed of processing is a core requirement. Some of those domains include micro-hardware control, image processing, graphics, sound processing, compression, networking, encryption, and digital signal filtering (Gordon 2010) (Thies 2009). Given similar requirements around performance, time series, sliding time windows computations,

and the graph-oriented nature of financial models, listed previously in Section 4.2, streaming is selected as the first facet in a computational representation for the field of economics.

4.3.1.1. Models of Computation

Streams have been used as a notation for representation of computational elements in domains of knowledge outside of economics for a long time (Stephens 1997). The first reference to an equivalent paradigm was on bullet notes given by Douglas McIlroy (McIlroy 1964) on October 11th of 1964.



10

Summary--what's most important.

To put my strongest concerns in a nutshell:

1. We should have some ways of coupling programs like garden hose--screw in another segment when it becomes when it becomes necessary to massage data in another way. This is the way of IO also.

2. Our loader should be able to do link-loading and controlled establishment.

3. Our library filing scheme should allow for rather general indexing, responsibility, generations, data path switching.

4. It should be possible to get private system components (all routines are sytem components) for buggering around with.

M. D. McIlroy
Oct. 11, 1964

Figure 7. The First Known Reference to Streams

A bullet summary by Douglas McIlroy on "what's most important", suggesting a function to "have some ways of coupling programs like garden hoses", what was referred by subsequent literature as *streams*

The original insight of "digital hoses", coined by Douglas McIlroy (McIlroy 1964) evolved through different milestones to consolidate the idea of streaming systems (Stephens 1997) (Thies 2009) (Gordon, Thies, et al. 2002). Each milestone of the evolution of what were initially *digital hoses* refers to a specific model of computation, as shown in Figure 8.

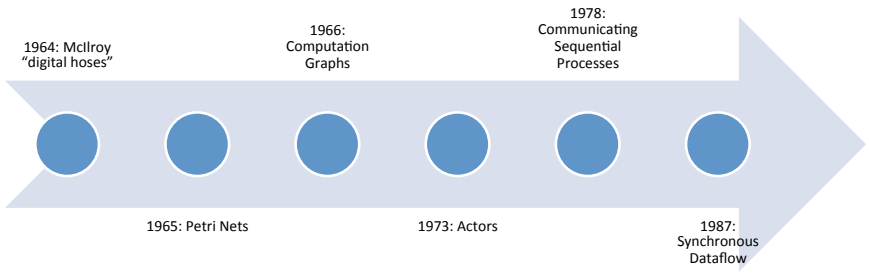


Figure 8. Timeline of Evolution: Models of Computation for Streams

Over time the original idea of "digital hoses" evolved on variations called "models of computation" from Petri Nets, Computation Graphs, Communicating Sequential Processes, and Synchronous Dataflow.

Each of these milestones, or models of computation, present different features and define a computational taxonomy of streaming systems (Thies 2009):

- **Petri Nets**[40]: a directed bipartite graph, where nodes either represent transitions or conditions. A directed edge specify which pre or post conditions

[40] Despite of Petri's original thesis of 1962 (Petri, Kommunikation mit Automaten (Communication with Automata) 1962) the formalization of Petri Nets as they are currently known only came a bit

are a requirement for a transition (Petri 1962) (Petri 1967) (Brauer and Reisig 2006) (Murata 1989);

- **Computation Graphs**: a graph-theoretic model for the description and analysis of parallel computations where computation steps correspond to nodes of a graph, while branches represent a dependency between computation-steps. Each branch is associated with independent queues of data (Karp and Miller 1966);

- **Kahn Process Networks**: a distributed model of computation where deterministic sequential processes are nodes, and FIFO channels are the edges of a graph network (Kahn 1974);

- **Actors**: a graph-based model of concurrent computation in which nodes are actors, and upon receipt of messages an actor can send new messages or create new actors. In this sense, edges can be created on demand and indicate a communication by message-passing (Hewitt, Bishop and Steiger 1973) (Greif 1975) (Clinger 1981) (Agha 1985) ;

- **Communicating Sequential Processes**: a textual and formal language for describing concurrent interaction based on primitive processes and events. Primitive processes are fundamental behaviors, and events represent indivisible and instantaneous interactions (Hoare 1978) (Hoare 2015);

- **Synchronous Dataflow**: a particular case of data flow in which each node represents a function, and each arc represents a signal path. It is a

later, in a 1965 colloquium (Petri, Grundsätzliches zur Beschreibung diskreter Prozesse 1967), published in 1967 (Brauer and Reisig 2006).

simplification of Kahn Process Network by limiting the number of messages each node consume and produce per signal (Lee and Messerschmitt 1987).

Despite lacking a standard nomenclature, topology, or modes of communication, all models of computation of streaming systems can be represented on a higher level by nodes and edges, arranged as graphs. In fact, the specific features of the models of computation can be normalized over three specific features (Stephens 1997) (Thies 2009) (Gordon, Thies, et al. 2002): topology, determinism, and dynamicity.

- **Topology**: defines the way in which nodes are arranged in a network;

- **Determinism**: establishes if the final results of execution are always the same, given the same set of inputs;

- **Dynamicity**: establishes if execution parameters (i.e., amount of buffering and communication patterns) can be decided and arranged statically and dynamically (i.e., at compilation time or runtime) (Gordon 2010) (Gordon, Thies, et al. 2002) (Thies 2009) (Thies, Karczmarek and Amarasinghe 2002)

The streaming facet translates these normalized features of models of computation – topology, determinism, and dynamicity – into three specific properties of financial models: synchronicity, connectivity, and plasticity. These properties are explained over the next topic when we explain the mechanics to define financial models using streams.

4.3.1.2. Defining Financial Models as Streams

The streaming facet defines a graph-oriented domain-specific language (van Deursen and Klint 2002) (Hohpe and Woolf 2012) to define financial models as a route of fragments of meta-data x through a chain of reusable and exchangeable processors P_i. The chain of processors P_i is arranged as function composition, as described in Equation 1 (J. M. Faleiro Jr 2007) (Spinellis 2001) (van Deursen, Klint and Visser 2000).

$$(P_1 \; o \; P_2 \; o \; \ldots \; o \; P_n)(x)$$

Equation 1. Function Composition

In the specific representation for the domain of economics, processors are chained together by a synchronicity operator δ giving a composition of processors the form shown in Equation 2.

$$x \rightarrow P_1 \; \delta_1 \; P_2 \; \delta_2 \; \ldots \; \delta_{n-1} \; P_n$$

Equation 2. Composition by Synchronicity Operator

We call this chain of processors P_1 ... P_n connected by δ a stream. Equation 3 gives an equivalent graph representation, based on edges and vertices, of the same stream.

$$\phi = (P_i, \delta_i)$$

Equation 3. Graph-Oriented Representation of a Stream

In Equation 3, ϕ is a directed sub-graph $\phi(V, E)$ in which V, the set of vertices P_i, are processors, and E, the set of edges δ_i, are synchronicity operators. The same graph ϕ can be visualized as a connected directional graph, as shown in Figure 9.

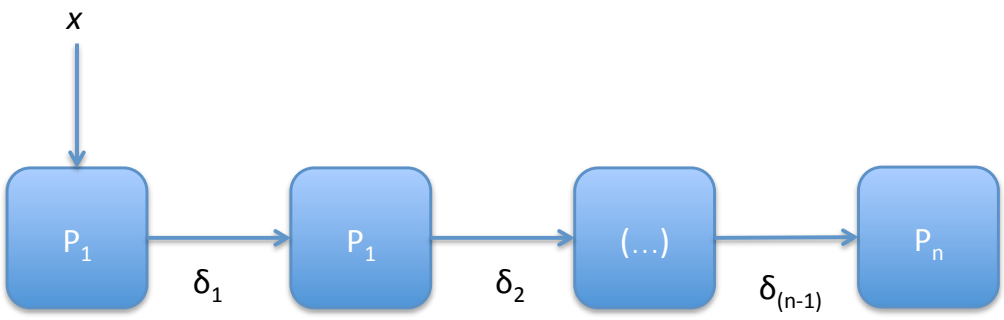


Figure 9. Streams as a directed graph

A stream can be visualized as a connected graph $\phi(V, E)$, in which edges are given by synchronicity operators $\eth_i$ and vertices, or nodes, by processors $\boldsymbol{P_i}$

We assume that a financial model, to be defined in terms of requirements listed in Section 4.2, must carry three fundamental properties:

- **Synchronicity**: financial models must operate on data fragments x synchronously or asynchronously, where x is defined in Equation 1;

- **Connectivity**: financial models are created by the composition of smaller, modular pieces that can often be recursively leveraged as smaller, reusable models;

- **Plasticity**: the composition and the behavior of a financial model can change, in real-time, upon arrival of new data fragments x, as defined in Equation 1.

Each of these three fundamental properties – synchronicity, connectivity, and plasticity - is formalized as stream elements, or interchangeably[41] as graph properties. Over the next sections, we formalize the representation of financial models over a stream-oriented language based on these three fundamental properties.

Synchronicity

In a stream ϕ, as described in Equation 3, processors P_i spawn tasks $t_{s,i}$ in pools of tasks T_i to handle data fragments x as they arrive. Each pool holds a variable number of tasks s, where the exact value of s, is irrelevant and associated with scheduling configuration details.

The synchronicity operator δ_i indicates how a fragment x is "handed over" from tasks in pools T_i in processors P_i, to P_{i+1}, where each δ_i can indicate two distinct modes: synchronous and asynchronous.

In synchronous mode, T_i depends on the completion of T_{i+1}, and therefore T_i can only proceed, and consume the next fragment x, after termination of task T_{i+1}.

[41] As we have shown in Section 4.3.1.1 through the different models of computation of streams, graph or stream representations are functionally interchangeable (Stephens 1997) (Thies 2009) (Gordon, Thies, et al. 2002).

Alternatively, in asynchronous mode, T_i does not depend on the completion of T_{i+1}, and therefore T_i can consume the next fragment x regardless of the result and termination of T_{i+1}.

Connectivity

Financial models are represented as directed graphs composed of a limited set of directed sub-graphs ϕ, as described in Equation 3, bound together by connectors. For all purposes, a connector C is a specialization type of processor P, as defined in Equation 1 as $P_{1..n}$, so that $C \cong P$.

As a specialized processor, a connector carries additional properties to allow the connection of multiple streaming sub-graphs $\phi = (P_i, \delta_i)$ into larger, interconnected networks of streams.

The composition of more elaborate, interconnected networks of streams allows the support of more complex functions. These functions include the plasticity property, described through a graph modification connector in the next section, reactive behaviors described in Section 4.3.2, the distribution facet described in Section 4.3.3, and enabling of distribution spaces described in Section 4.3.3.2. A complete outline of possible connectivity functions is provided later in this thesis, when we describe in details the processor contribution, in Section 4.4.2.

Plasticity

A graph representing a financial model should be able to modify itself upon arrival of relevant data fragments, depending on specific requirements of the model under study.

In the scope of this research, the ability to modify a graph Φ on demand is referred to as plasticity and is given by a special modification connector C_p. The connector C_p is given by function f of a predicate P on data fragment X, and a sub-graph template $\bar{\phi}$, formalized by Equation 4.

$$\bar{\phi} = (\bar{P}_i, \bar{\delta}_i)$$
$$P: X \rightarrow \{true, false\}$$
$$C_p = f(P, \Phi, \bar{\phi})$$

Equation 4. Definition of Plasticity Function

Plasticity occurs upon arrival of data fragment X. In case P resolves as a $true$ for X, a sub-graph ϕ based on template $\bar{\phi}$ is appended to graph Φ.

In short, this model in Equation 4 allows a graph to modify itself, if an arrival of a data fragment X causes the predicate P to resolve as $true$.

The best way to explain the plasticity property is through an example, and preferably in finance, through a common use case. In Figure 10, we depict the example of a definition of a new pricing route for a stock that should be set up upon arrival of a new symbol of that stock on a sequential feed of price ticks.

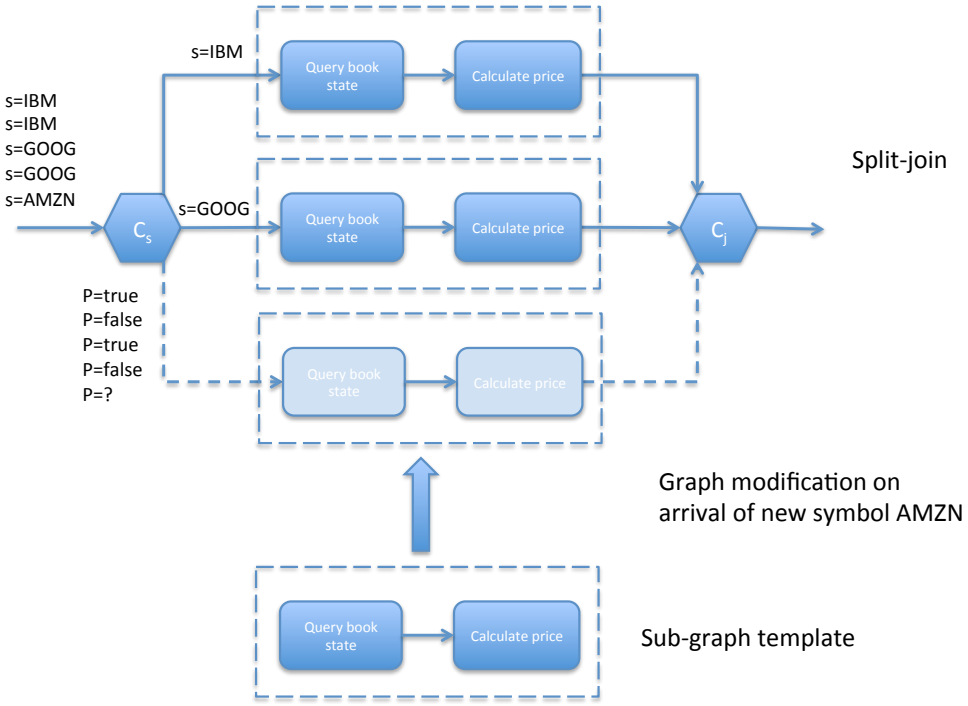


Figure 10. Graph Modification Connector Example

An example of plasticity through the application of a graph modification connector, when new symbols arrive and predicate P fires $true$, a new path on the graph is added based on the sub-graph template. In this example, the arrival of symbol AMZN will create a new branch and the modification of the overall graph.

On this use case it is assumed high-frequency requirements, so to maximize throughput every branch on the graph is dedicated to one symbol. The complete set of symbols is not known in advance. Therefore, a new branch must be created for every new incoming symbol, the first time an instance of this symbol is received. On this exercise, each branch executes the following steps for any given price tick:

- Query current state of the order book for current bid and ask prices of stock s;

- Calculate the price of a stock based on current mid-price, spread, and exponentially weighted moving average of the mid-price[42].

Every incoming fragment x, arriving at time t, carries a tuple (s, p) where s is symbol and p is the price. In this exercise, for the sake of simplification, since our concern is specifically to exemplify the property of plasticity, we are only interested in symbols. The sequence of incoming symbols is given on this example by the sequence in Equation 5.

$$s: (IBM, IBM, GOOG, GOOG, AMZN, \dots)$$

Equation 5. Sequence of Incoming Symbols

In response to each item in the sequence, predicate P turns to $true$ if this is the first arrival of that symbol on the sequence. In response to the sequence of symbols in Equation 5, P yields a correspondent sequence of results given by Equation 6.

$$P: (true, false, true, false, true, \dots)$$

Equation 6. Sequence of Predicate Results

In the previous Figure 10, the configuration of the graph Φ is shown as a snapshot in time right after the arrival of the AMZN symbol, as result of the use of a modification connector C_p. Before that snapshot, the symbol IBM had arrived,

[42] Signal attenuation functions are explored in details in the upcoming Chapter 5. The exponentially weighted moving average specifically is described in Section 5.2.2.4.

producing sub-graph $s = IBM$, followed by symbol GOOG, what produced the sub-graph $s = GOOG$.

On arrival of symbol AMZM, the predicate P yields $true$ and the third branch is created and associated with the newly arrived symbol AMZN. From that point on, arrivals of new symbols AMZN will be routed through the newly created branch.

The plasticity property allows for on-demand modifications on connections of a graph representing a financial model. This example is an important and common use case on models related to trading of financial instruments.

4.3.2. REACTIVES

Financial models must maintain continuous interaction with an ever-changing state that varies over time and is external to the financial model at hand, as stated on requirements defined in Section 4.2. Rules on the financial model have to trigger specific actions based on external events that can occur at unpredictable times.

These unexpected, unpredictable, external changes are hard to represent in conventional, sequential programming. External changes are associated to events, and require a number of non-sequential[43] properties for representation in financial models: inverted control, abstraction of time management, and abstraction of synchronicity details (Bainomugisha, et al. 2013):

- **Inverted Control**: Financial models keep a continuous and persistent interaction with their execution environment, executing actions based on

[43] Some literature considers reactives an extension of stream processing (Stephens 1997). Given the nature and requirements of financial models we opted to differentiate between sequential (streams) and non-sequential (reactives) as two separate and yet complementary facets.

events triggered by external sources. External sources then drive the order of execution, and as a consequence, in many particular cases, the rules and control flow of a financial model is inverted (Bainomugisha, et al. 2013).

- **Abstraction of Time Management**: Financial models often require a notion of a discrete time series, in which the modification of the event associated with each time t_i in the time series is performed by behaviors (Harel and Pnueli 1985) (Bainomugisha, et al. 2013). The event is associated to either a lifecycle change (e.g., corporate actions in a stock, roll-over operations in derivative instruments) or variations of value (e.g., the price of an asset, ratio of risk exposure) over time. After a relationship between reactive entities is set, computation dependencies and handling of events over time are automatic and the representation of time is intrinsic to every event (Nilsson, Courtney and Peterson 2002).

- **Abstraction of Synchronicity Details:** Financial models require abstraction of synchronicity details in the event-driven communication. Financial entities are often defined in terms of relationships with other entities. In a representation suitable for financial models, associations are established declaratively, similar to the way in which cells in a spreadsheet are defined and associated with a formula[44]. The declarative association through formula provides automatic management of associations between data dependencies. The event-driven communication synchronizing the state of those entities is intrinsic to the representation of the association and therefore transparent.

Functionally, these properties – inverted control, abstraction of time management and details of event-driven communication - are related in computer

[44] The designation of a formula is equivalent to the concept of a formula in an electronic spreadsheet

science to what is commonly called reactive programming (Harel and Pnueli 1985), and referred in the scope of this research as a reactive[45] facet.

The *reactive facet* is a declarative paradigm that allows the definition of what has to be done through reactive relationships, and let the computational representation automatically take care of when to do it, and who gets affected. A similar and more intuitive model is exemplified by a number of cells in an electronic spreadsheet representing a formula. Similarly, reactives allow for an intuitive representation of primitives and formula, in which composition of formula from primitives and other formula is defined declaratively (Harel and Pnueli 1988) (Bainomugisha, et al. 2013).

To describe declarative associations of reactive variables, we take for example the simple formula in Equation 7.

$$A = B + C$$

Equation 7. Reactive Formula Example

In a sequential representation, variables B and C would have to be set first, so that only then the computation of A could occur. Alternatively, in a reactive representation, the formula is declared first, setting a graph of reactive dependencies. In Figure 11 we show the graph of dependencies for the formula in Equation 7.

[45] Functionally equivalent patterns like observers, event-driven programming and asynchronous callbacks were also considered as possible alternatives to reactives, but unfortunately they carry their own impeding limitations. The coordination of individual callbacks, over a shared state, across numerous code fragments, in which the order of execution cannot be predicted, is an error prone, cryptic, daunting programming task (Bainomugisha, et al. 2013). Additionally, since callbacks do not produce a return value, these alternative programming patterns must perform side effects in order to affect the application state (Cooper and Krishnamurthi 2006).

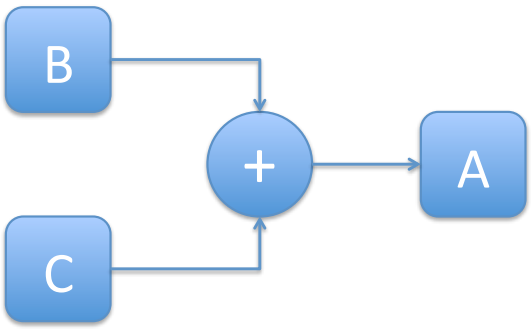


Figure 11. Graph of Reactive Dependencies

The reactive graph, representing a simple formula $A = B + C$. A formula of functions, operators, or other reactives is set as a graph of communication between reactives.

The graph in Figure 11 represents that, in case of a change on the value of either B or C, the executing environment abstracts the notion of a discrete time change and event-driven communication by propagating the modification across all dependencies in the graph. The exact way a value propagates over time through a graph of dependencies, in this case from B or C to A, can occur in more than one way and is abstracted from the representation itself.

The reactive paradigm is a broad concept, subject to a specific classification in terms of basic features, evaluation model, lifting, and directionality. These classes are related to special considerations for the use of the reactive paradigm is used to represent financial models, according to requirements defined in Section 4.2.

4.3.2.1. Basic Features

Two basic features define the reactive programming paradigm: behaviors and events. They are often referred to as *duals* because one can be used to represent the other. The behavior feature refers to mutable, time-varying values. The event feature refers to potentially infinite, immutable modifications that occur at discrete points in time.

In the computational representation for the field of economics, behaviors are associated with specialized processors R so that $R \cong P$. The dual of R is a virtually infinite sequence of events x, as previously discussed in Section 4.3.1.

Given for example two disjoint sub-graphs ϕ' and ϕ'', a generic synchronicity operator δ, and behaviors associated to reactive processors R' and R'', as described in Equation 8.

$$\begin{matrix} \phi' \ \delta \ R' \\ \phi'' \ \delta \ R'' \end{matrix}$$

Equation 8. Composition of Streams and Reactives

A reactive function ${f_r}'(R',\ R'')$ is evaluated on the arrival of either x' or x'', in each of the streams defined by sub-graphs ϕ' and ϕ'', as shown in Figure 12.

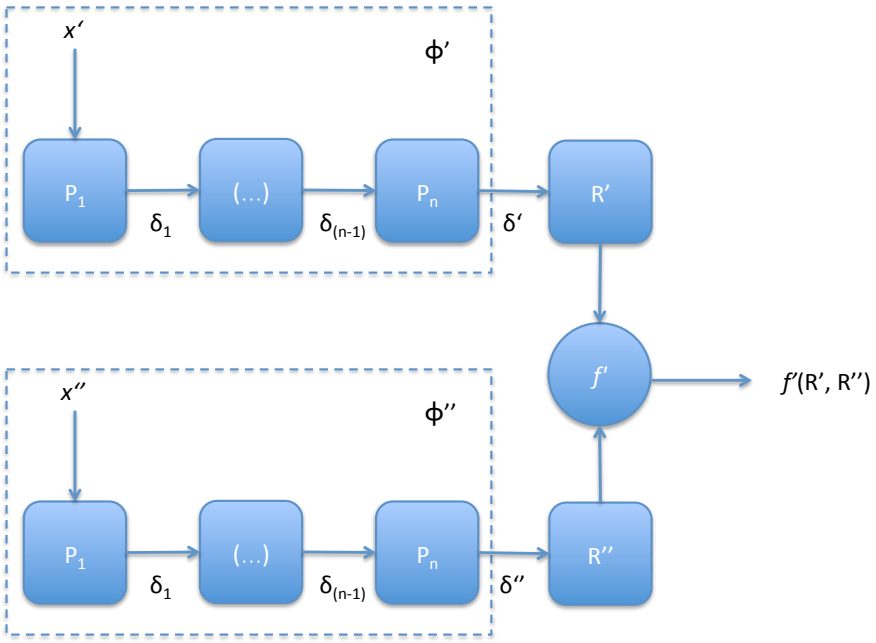


Figure 12. Composition of Streams and Reactives

Reactives are specialized processors in a stream, either as reactive steps in linear streams, like R' and R'', or as connectors for disjoint sub-graphs, like the reactive function $f_r{}'$.

As represented in Figure 12, reactives are specialized processors in a stream. In the case of R' and R'', they have to behave like regular, sequential processors for sub-graphs ϕ' and ϕ'', and at the same time act as reactives for the reactive function $f_r{}'(R', R'')$.

Reactive dependencies and respective flow of execution are defined declaratively. As a consequence, after the reactive graph is defined, calculations are not affected by the sequence of initialization of R' and R''.

4.3.2.2. Evaluation Model

The evaluation model of the reactive facet defines how a change x in stream ϕ propagates through a dependency graph of values and computations.

In a pull-based evaluation model, a value is calculated on demand, or in other words, a value has to be "pulled" from the source whenever required.

On the other way, in a push-based evaluation model, every change in value has to be sent to dependent computations. The push-based propagation is called data-driven since it occurs by the availability of new data.

The evaluation model has no direct implications on the representation, but it does, however, have implications on the distribution facet. Those implications are discussed in detail in the upcoming Section 4.3.3.

4.3.2.3. Lifting

We call lifting the transformation of a generic function $f(x)$ applied to x to a lifted function $f'(R < x >)$, where $R < x >$ is a reactive, or behavior, type of x (Bainomugisha, et al. 2013) given in Equation 9.

$$lift: f(x) \rightarrow f'(R < x >)$$

Equation 9. Reactive Lift Function

When looking at time step i, the resolution of a lifted function f' on value x_i yields the original function f, as shown in Equation 10. Mathematical operators (e.g.,

$+, -, *$) and user-defined functions, respectively, are functionally equivalent to lifted operators and functions.

$$f'(R < x_i >) \rightarrow f(x_i)$$

Equation 10. Original Lifted Function

The representation of the lifting transformation is classified further in terms of how much additional context is needed whenever an operator or a function has to be lifted to a reactive representation. This classification defines a lifting transformation as manual, explicit, or implicit.

- **Manual**: on manual lifting, a representation does not provide transparent lifting. A time-varying value has to be manually extracted and applied to operators, functions, or dependent variables.

- **Explicit**: on explicit lifting, the representation defines a number of unique operators that can be used to lift a function f to f'.

- **Implicit**: on implicit lifting, all operators and functions of a representation applied to x, user-defined or not, are transparently lifted to a reactive item $R < x >$.

For simplicity of communication, in the computational representation for the field of economics, all reactive transformations are implicitly lifted. This requirement will impose additional constraints on candidate implementations, but as a consequence gives a higher level of abstraction to the representation.

4.3.2.4. Directionality

A reactive representation may allow reactive propagation of changes to occur in one direction – unidirectional – or in either direction – multidirectional. In requirements listed in Section 4.2, there were no specific cases in which multidirectional propagation was an absolute requirement. As a consequence, for simplicity, for a computational representation for the field of economics, only a unidirectional propagation is required.

4.3.3. DISTRIBUTION

In Section 4.2 we listed some requirements explicitly related to the operation of financial models in large scale, both in terms of computational power and storage.

Those requirements - virtually infinite historical records, and responsiveness – require the use of distributed resources (Dean and Ghemawat 2004) to be able to scale to more than a single processing or storage unit. The *distribution facet* gives the computational representation the ability to communicate functions related to scaling up the workload of a financial model across multiple processors.

A distribution facet is, in essence, a particular application of connectors, as described in Section 4.3.1.2. A connector C is a specialized type of processor P so that $C \cong P$. That specialization means that in addition to the behavior of processors, a connector carries additional properties to allow the composition of streaming sub-graphs $\phi = (P_i, \delta_i)$ into larger, interconnected networks of streams.

On its more generic form, any connector C' allows for bridging of a number n of incoming sub-graphs ϕ' and a number m of outgoing sub-graphs ϕ'', as described in Figure 13:

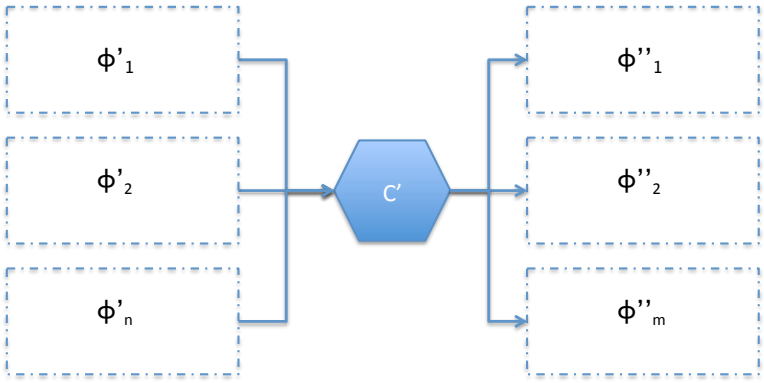


Figure 13. Connectors and Incoming and Outgoing Streams

Connectors in the distribution facet are used to compose streams by bridging a number n of incoming sub-graphs ϕ' and m outgoing sub-graphs to build more elaborate graphs.

The generic description of the distribution facet as connectors for a $n{:}m$ association of incoming to outgoing sub-graphs of streams has two significant consequences: improved expressiveness of the streaming notation, and leveraging of distributed and parallel processing in large scale.

4.3.3.1. Improved Expressiveness

The first consequence, improved expressiveness, is related to the possibility of laying out streams and connectors in different combinations to define more elaborated patterns (Hohpe and Woolf 2012). The placement of connectors in different locations of the streaming graph can define structures like split-joins and feedbacks (Gordon 2010) (Gordon, Thies, et al. 2002) (Thies 2009) (Thies, Karczmarek and Amarasinghe 2002) (Stephens 1997) (Kamburugamuve and Fox 2013). A visual representation of split-join and feedback loop is given in Figure 14.

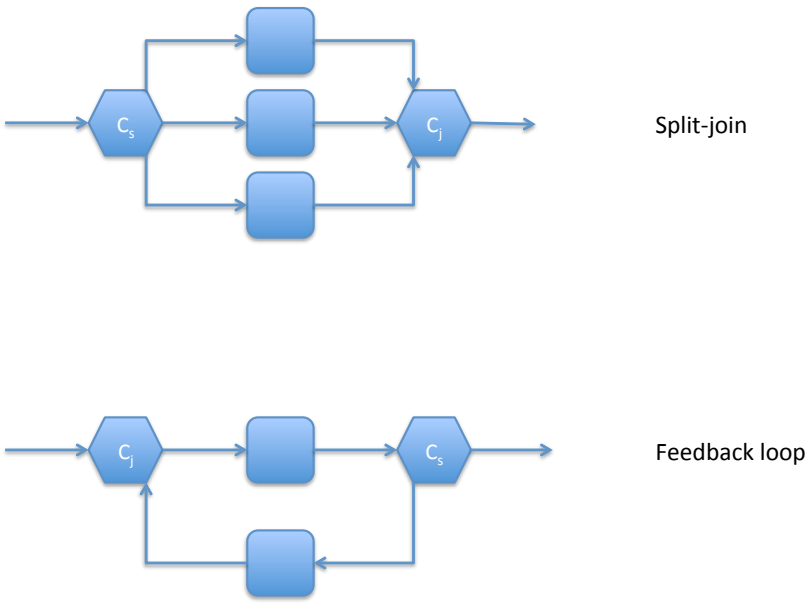


Figure 14. Use of Connectors for the Composition of Communication Patterns

Combination of connectors used to create more elaborate composite patterns like split-joins and feedback loops.

A split-join pattern is given by a pair of connectors, C_s and C_j, positioned around a set of processors. The connector C_s is placed on the splitting, inbound edge of the set of processors, while the connector C_j is placed on the joiner, outbound edge of the set of the processor.

Edges for communication in and out of the processor stack between C_s and C_j are given by the asynchronous operator δ. As a consequence, fragments leaving C_s might hit all processors concurrently, and the order of execution of these processors cannot be guaranteed. Unless specialized processors are inserted in the flow before the join of C_j, with the ability of re-establishing the original order of execution, the overall execution is non-deterministic.

A variation of the split-join pattern, the feedback loop, given in the lower part of Figure 14, is a re-arrangement of C_s and C_j to represent a loopback of data fragments. Edges for communication out of C_j and into C_s are asynchronous, i.e., the synchronicity operator δ is of type asynchronous.

Since there is a requirement of asynchronous communication on edges from and to C_s and C_j, results of the execution of the overall stream in a feedback loop are non-deterministic.

4.3.3.2. Parallelism and Distribution Spaces

The second consequence is the possibility of describing an execution flow spanning multiple computational environments and locations concurrently. Each of those disjoint computational environments is called *space*. A space by definition can be associated with different processors, in different locations, as required.

For example, given a connector C' and disjoint sub-graphs ϕ', ϕ'' and ϕ''' on a specific composition, as described in Equation 11:

$$\phi' \; \delta \; C'$$
$$C' \; \delta \; \phi''$$
$$C' \; \delta \; \phi'''$$

Equation 11. Connectors and Distribution Spaces

Following the definition of the connectivity property of financial models, described in Section 4.3.1.2, a larger graph Φ is defined as a result of connector C' applied on sub-graphs ϕ', ϕ'' and ϕ''' as shown in Figure 15.

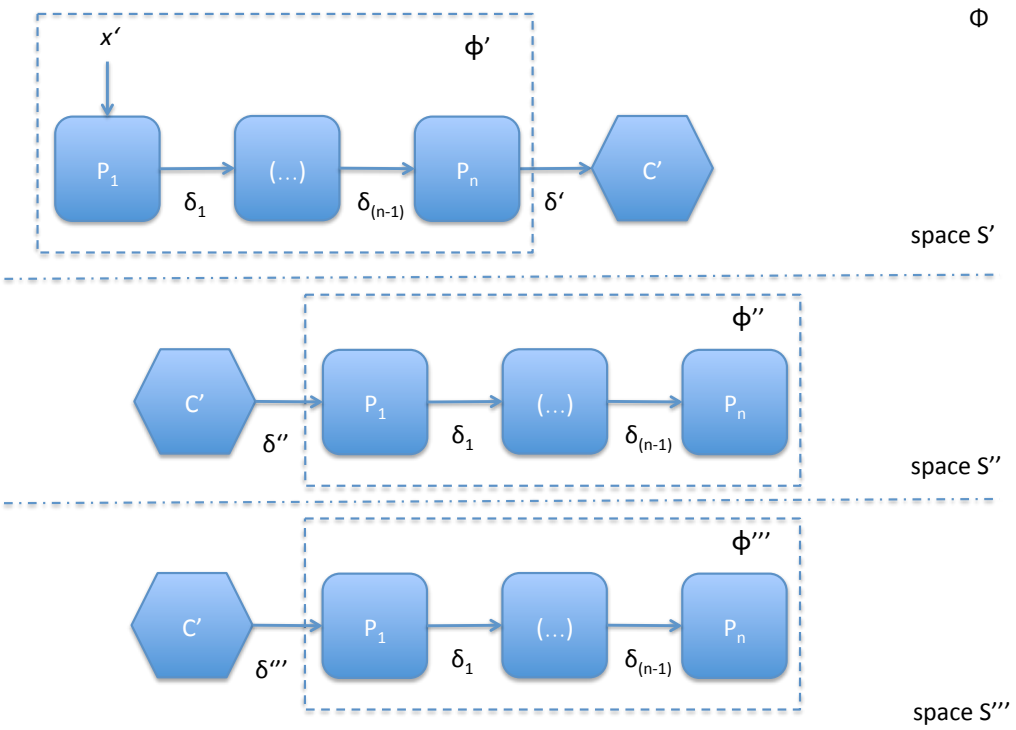


Figure 15. Graph Composition by Connectors and Spaces

The application of connectors to sub-graphs defines a generic graph Φ and multiple spaces S', S'' and S''', which can be optionally associated to computational resources spanning multiple locations.

Each space is a fragment of a complete graph Φ, in which boundaries of any space are given by incoming or outgoing edges of a connector. Each space abstracts details related to distribution or concurrency aspects of a financial model and can be at a later time associated to computational resources spanning multiple locations, without affecting the high-level representation of the financial model itself.

A complete graph Φ in Figure 15 shows the use of a connector C' to define multiple spaces S', S'' and S''', each associated to sub-graphs ϕ', ϕ'' and ϕ'''. Each space and sub-graph can be associated to different distribution contexts, without affecting the intuitive description of a financial model.

In essence, connectors allow scaling of a financial model to handle a virtually infinite load and volume of data by adding the notion of locality and distribution transparently through the use of spaces.

This notion is intrinsic to the representation in a sense that it is not defined in the financial model described, and decisions relative to performance, storage, and processing power can be made at a later time, with no modifications to the financial model itself.

4.3.4. SIMULATION

Financial models are a representation of complex systems in which the intent of defining one, in many cases, is to allow prediction of outcomes, through the application of disciplined, objective research methods.

Simulations are a fundamental technique for research of complex problems in many disciplines, especially in financial sciences, through the application of specialized algorithms (Von Ronne 2012) to define, search and test possible viable solutions. The exact placement of simulations in the proof pipeline proposed in this research is given in Section 3.2.1. According to that definition, simulations are the imitation of a system (Robinson 2004). Financial models are, in essence, an imitation – a controlled simplification to the right scale – of large, complex systems.

The facet simulation is responsible for representing methods allowing the anticipation of possible outcomes in financial models. As described in Section 3.2.1 the general topic of simulations is extensive and under active research. To that, the primary challenge when defining a simulation facet in a domain of knowledge is to establish the exact level of simplification that can be applied to models on that

representation, without affecting the quality of insights into the central problem under simulation.

According to the representational process previously defined in Section 3.4.1, any facet, and in this case precisely the simulation facet, must be selected based on domain-specific requirements and a computational taxonomy. Domain-specific requirements were previously defined in Section 4.2, and given the importance of the subject of simulation, should be augmented by the proof pipeline defined in Section 3.2.1. A computational taxonomy is given by various alternatives for classification simulation techniques (Leemis and Park 2006) (Robinson 2004) (Sulistio, Yeo and Buyya 2004) (Von Ronne 2012), shown in Figure 16. According to that representational process, the combination of requirements and techniques are enough to select and adjust relevant properties of simulation for financial models.

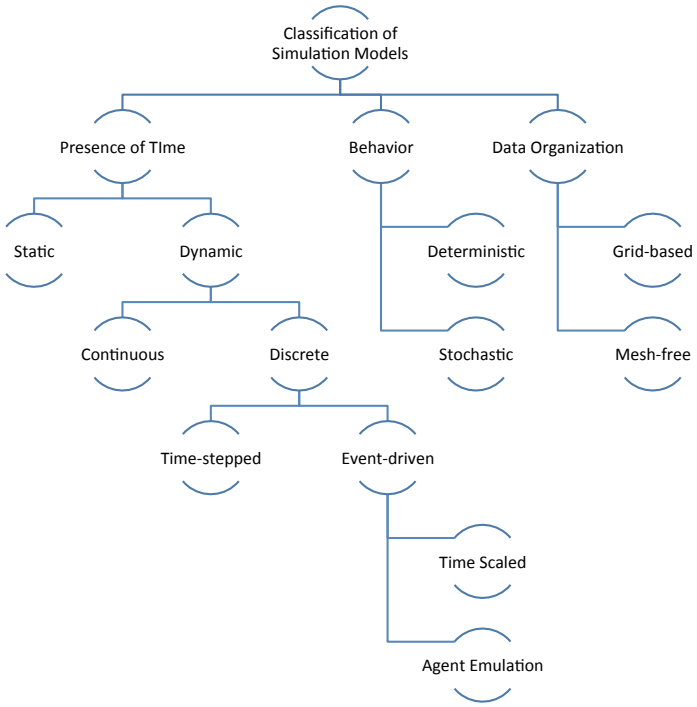


Figure 16. Simulation Taxonomy and Relevance to Economics

Simulation facet classified according to a generic taxonomy based on the presence of time, behavior and data organization. Leafs marked in a dotted like do not represent relevance to the field of economics.

The taxonomy of a simulation facet organizes all possible representations in three dimensions according to the presence of time, behavior and data organization. All three dimensions of classification are complete and not mutually exclusive, in a sense that a model requires a concomitant classification in each of the dimensions.

For example, a model that anticipates the effect of corporate actions over the price of an asset is static, deterministic and grid-based, while a model that uses random shocks to determine the influence of multiple features in the profitability of a portfolio in closing prices is dynamic time-stepped, stochastic and grid-based.

4.3.4.1. Presence of Time

The first classification takes into consideration the presence of time (Sulistio, Yeo and Buyya 2004), or time of change (Von Ronne 2012). As the name implies, this classification considers if time is a significant variable in defining the outcome of a simulation (Leemis and Park 2006).

Under this classification, system models can be classified as static or dynamic, respectively by observing if the system can be adequately modeled without or with a variable associated with time.

In our cases of use, previously defined in Section 4.2, it is clear that the absolute majority of financial models are dynamic. However, we cannot ignore that some critical exceptions do not require the presence of time. An example of a static system relevant to the field of financial sciences would be the influence of corporate actions on the valuation of equity assets on the day that a specific action took place.

Dynamic systems are further classified as either continuous or discrete (Sulistio, Yeo and Buyya 2004). Continuous dynamic systems consider that variables or features into consideration evolve continuously and are usually subject to modeling through differential equations, representing continuous modifications of a system. Some examples outside of economics are often related to classic mechanics like particles moving in gravitational fields, or an oscillating pendulum (Leemis and Park 2006). All observable phenomena are by nature continuous, but since by definition models are surrogates of real events, the use of discrete dynamic systems allow a significant simplification by considering that all variables of the system are piecewise constant functions of time, only possessing one of many values within a finite range.

Dynamic discrete systems can be classified even further depending on the irregularity of the time interval as time-stepped or event-driven systems.

On time-stepped systems, time intervals are constant or derived from fractions of time in which periodicity can be clearly ascertained. Examples of a dynamic discrete system in finance are models associated with changes in discrete values (e.g., prices, returns, risk ratios) over a time-series. Dynamic discrete time-stepped systems account for the majority of the cases of use in finance.

On event-driven systems[46], time interruptions occur in irregular intervals, driven by external sources of the model itself. In finance, such systems are not as usual as systems based on constant time steps. Event driven-systems should, however, be considered at least as necessary, and an adequate tool when investigating sophisticated use cases. Some examples are related to cases of agent-based simulation of a central limit order book. In these simulations, software agents play the role of market participants and are used to gauge the influence of real-world economic agents to study the effect of a pre-defined behavior in variations of the price of financial instruments (Foata, Vidhamali and Abergel 2011) (Panayi and Peters 2015) (Cont, Stoikov and Talreja 2010) (Murat, et al. 2009) (Chakraborti, et al. 2011) (Chan and Shelton 2001).

Dynamic discrete systems are represented in a computational representation for the field of economics by adjusting generic streams $\phi(P, \delta)$ in two specific points to represent either time-stepped systems or event-driven systems.

Time-stepped systems are replicated by replacing the generic endpoint X of fragments x in a stream $\phi(P, \delta)$ by a time-paced endpoint $f(T, X)$ so that the period

[46] We assume agent-based modeling is an extension of tools commonly used to simulate event-driven dynamic systems (Bandini, Manzoni and Vizzari 2009) (Borshchev and Filippov 2004).

T between events x_t can be adjusted, as shown in Figure 17.

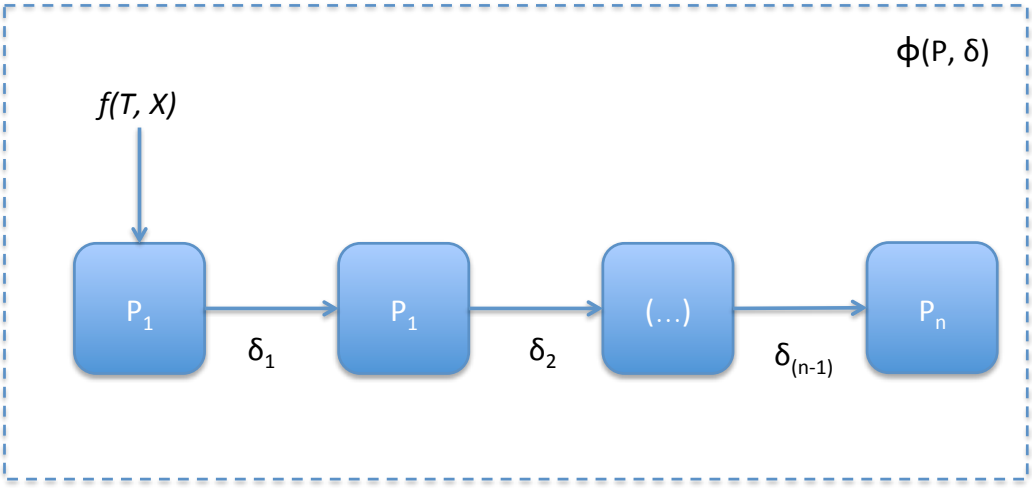


Figure 17. Simulation of Time-Stepped System

A time-stepped system is simulated by replacing the generic endpoint X of fragments x by a periodic function $f(T, X)$ in which the period T is a fraction of the time-step present in the original system under simulation.

Adjusting T to minimal amounts allows for the simulation in milliseconds of market behaviors that otherwise could only be observed over long periods, achieving for all practical purposes a time-squeezing effect. The simulation can then be replayed as many times as required, with different values for all relevant features (Faleiro Jr and Tsang 2016). An example of a time-stepped simulation is given in Chapter 5 when the hypothesis of profitability of momentum crossover strategies is tested against historical values of an index in Section 5.4.3.

Event-driven systems are represented in a computational representation for the field of economics by two different variations: time scaling and agent emulation.

In the time scaling variation, a generic endpoint X of fragments x in a stream $\phi(P, \delta)$ is replaced by an endpoint $f(k, X)$, allowing event-driven systems to be

replicated by replaying events x_t from X in a different scale of time. The same time-squeezing effect observed in time-stepped simulations is achieved by adjusting the time of occurrence of each event x_t to a shorter scale k as $x_{t'}$ as described in Equation 12.

$$t' = t_0 + \frac{1}{k}t$$

Equation 12. Time Scaling

Where t_0 is an arbitrary time assigned to the beginning of the simulation, and k is the time compression scale.

In the agent emulation variation, streams $\phi(P, \delta)$ play the role of individual software agents, similar to what some literary references call a process-oriented paradigm (Matloff 2008), process-modeling (Pedgen 2010) or process-interaction (Vangheluwe 2014).

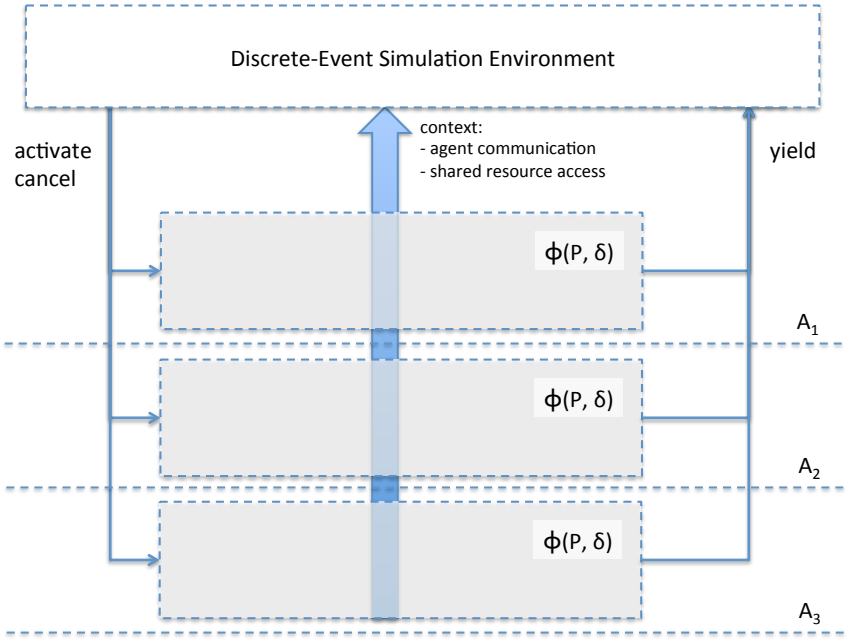


Figure 18. Discrete-Event Simulation Environment

Each stream $\phi(P, \delta)$ plays the role of software agents A_i, and execution is done by specific signals activate, cancel or yield. Agent communication and competing access for shared resources is done through the agent context.

Each software agent A_i, represented by stream $\phi(P, \delta)$, is used to model real-world entities that hold state and evolve in time. Agents interact through a shared context, either by direct communication or by modification of state in a shared resource.

A component called a Discrete-Event Simulation Environment is responsible for proper scheduling and coordination of an agent A_i by issuing and capturing variations of events of type activate, cancel, or yield.

In a higher level, an activate signal marks an agent as eligible for execution, while a cancel event forces an agent to release resources and yield execution. Agents notify the discrete-event simulation environment of specific changes in a task status

by issuing different types of yield signals. A yield signal tells the discrete-event simulation environment that the agent can be de-scheduled and action can go to an eligible agent if such an agent is available (Matloff 2008) (Scherfke 2014).

4.3.4.2. Behavior

The second classification of simulation models takes into consideration the randomness of results of the execution of a model given a constant set of inputs.

In deterministic models, the result of a model execution depends only on the input given to the model, what means that repeating a simulation several times will yield the same results (Sulistio, Yeo and Buyya 2004). On the other hand, in stochastic models, the result of a simulation varies randomly[47] (Perros 2009).

Both deterministic and stochastic behaviors are required for financial models, and the exact nature of a model is defined by behaviors of processors and topology of the underlying graph ϕ describing the model, as explained previously in Sections 4.3.1, 4.3.2, and 4.3.3.

4.3.4.3. Data Organization

The third classification of simulation models arranges simulations as grid-based or mesh-free, depending on the data organization scheme (Von Ronne 2012).

In the mesh-free organization (Gingold and Monaghan 1977), data is associated with individual and disconnected (i.e., mesh-free) nodes called particles.

[47] Pseudo-random models, in which a random outcome is emulated by a pre-defined sequence of random values, based on a number called a *seed*, are indeed a special case of a deterministic model. If following this definition, the simulation of a random walk in Section 5.4.2 should be more accurately referred to as a pseudo-random walk.

Updates to a particle are not bound to neighboring or connected relationships between particles, but instead are related to interactions to all particles considered relevant. The mesh-free organization enables the simulation of complex systems, at the expense of computing power and programming complexity. Use of mesh-free simulation in finance usually applies to overspecialized cases of use (Kim, Bae and Koo 2013) (Duffy 2006) (Lopez 2012) (Fasshauer 2006) and as a result was considered out of scope and left out of the list of cases of use in Section 4.2.

Alternatively, in the grid-based organization, the state of a simulation is arranged in discrete cells at particular locations in a grid. Updates occur to each cell based on previous state and those of its neighbors, or to cells to which it is connected. The absolute majority of the financial models are grid-based.

In this proposed computational representation for the field of economics some of the fundamental constructions - reactive primitives, functions, and operators - play the role of cells in a grid-based organization while connections reactive primitives constructions are arranged in the same way as cell dependencies.

4.4. CONTRIBUTIONS

As defined by the representational process introduced in Section 3.4.1, the second component of a computational representation is referred to as contributions. As introduced in Section 3.4.3, contributions are defined as shareable and formal evidence of a scientific crowd-based investigation.

According to the representational process in Section 3.4.1 contributions are a taxonomy of shareable evidence that is relevant to cases of use on the domain of knowledge under study, in this specific case, economics. The cases of use were described previously in Section 4.2. According to the evidential properties discussed

in Section 3.4.3, all contributions must follow a classification system, called taxonomy of contributions. Contributions for a computational representation for the field of economics must cover a broad range of models, methods, and results relevant to financial sciences (Herndon, Ash and Pollin 2013). Some examples include datasets in small, medium or large scale; time series in low, medium or high frequency; calculation processors and visualization plots; and results related to historical and real-time execution, simulation and backtesting. The taxonomy of contributions for the field of economics is shown in Figure 19.

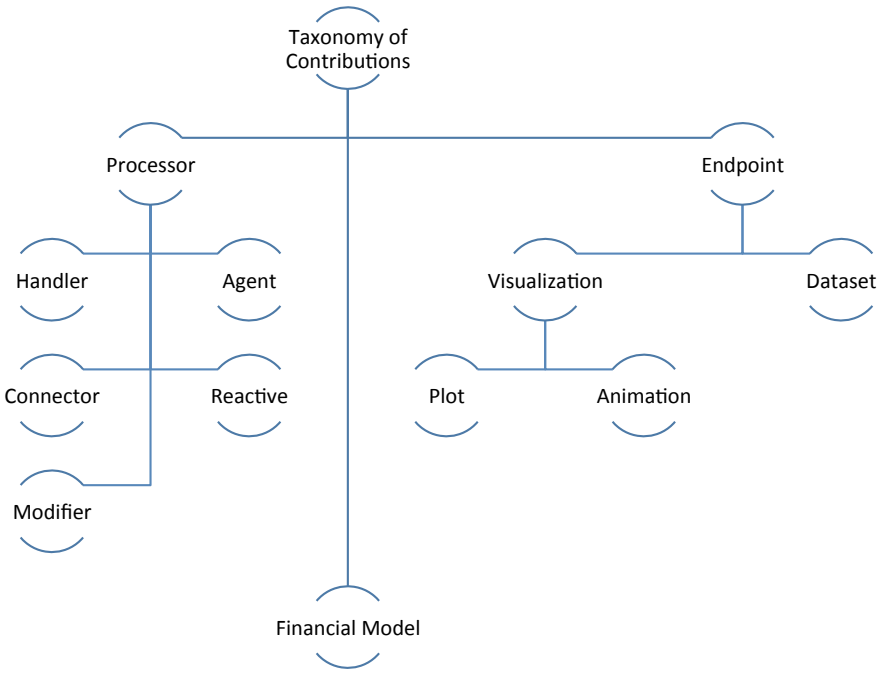


Figure 19. Taxonomy of Contributions

Contributions are classified as financial models, processors, or endpoints. Endpoints are either visualization (i.e., plots, animations) or datasets.

All contributions carry the evidential properties defined in Section 3.4.3, and therefore, in addition to falling in precisely one classification of the tree in Figure 19,

all contributions are also uniquely identified, carry a detailed record of provenance and hold enforceable ownership and access information. On a higher-level, contributions are classified as financial models, processors or endpoints.

4.4.1. FINANCIAL MODEL

The first type of contribution for a computational representation for the field of economics is a financial model. Financial Models are by definition a description of observable phenomena in the field of economics, simplified to the right scale, and adjusted for use in the process of a crowd-based investigation described in Section 3.2. An extensive outline of requirements of financial models is listed in Section 4.2.

Since financial models are contributions, they follow what we call evidential properties of contributions, explained in Section 3.4.3. As such, financial models can be defined and reused by different users.

From a representational perspective, financial models are built based on streams of processors and endpoints, arranged as graphs. The fundamentals for the description of financial models as streams are described in Section 4.3.1.2 on page 106. An example of a generic financial model is depicted as graph Φ in Figure 15 on page 126. Processors and endpoints as contributions are explained over the following sections 4.4.2 and 4.4.3 respectively.

4.4.2. PROCESSORS

The second type of contribution for a computational representation for the field of economics is a processor. Processors are steps on the execution stream and

are placed to perform specific computations on fragments of data x, as already explained in details in Section 4.3.1.2.

Since processors are contributions, they carry evidential properties of contributions, explained in Section 3.4.3. As such, processors can be defined and reused across different financial models (i.e., execution streams, as explained in Section 4.3.1.2), by different users, whenever that same specific function is required. Processors are further classified as handlers, connectors, modifiers, reactives, or agents.

- **Handler**: the simplest type of a processor is a handler. A handler applies transformations to a fragment of meta-data x as defined in Section 4.3.1.2;

- **Connector**: the composition of larger, more complex financial models from multiple smaller sub-graphs is possible by using a specialized type of processor called connectors. Connectors were explained in details in Section 4.3.1.2 on page 109 when the connectivity property of financial models as streams is explained;

- **Modifiers**: modifiers support the plasticity property of financial models, as extensively explained in Section 4.3.1.2 on page 109. The plasticity function was formalized in Equation 4. A practical example of plasticity applied to finance was described in Figure 10;

- **Reactives**: reactives allow the representation of declarative, non-sequential properties in financial models: inverted control, abstraction of time management, and abstraction of synchronicity details. The formalization of reactive processors was given previously in Section 4.3.2 on page 113.

- **Agents**: this specialization of a processor is used to support a sub-classification of event-driven simulation model called agent emulation. Agent emulation is described in details in Section 4.3.4.1 on page 130. On that section, an agent processor is described as software agent A_i, represented by stream $\phi(P, \delta)$, in Figure 18.

4.4.3. ENDPOINTS

The third type of contributions in a computational representation for the field of economics is called an endpoint. Endpoints can be used as either a source or a destination of data fragments x in the execution stream $\phi(P_i, \delta_i)$, as previously represented in Figure 9 on page 107.

Since endpoints are contributions, they follow what we call evidential properties of contributions, explained in Section 3.4.3. As such, endpoints can be defined and reused across different financial models (execution streams, as explained in Section 4.3.1.2), by different users, whenever that same endpoint, or state of data, is required. Depending on the intended use of the data, endpoints can be further classified as visualizations or datasets.

Visualizations can be static or dynamic. Static visualizations, called plots, show a complete and immutable representation of samples $(x_i \ldots x_j)$ in which the window associated to the interval $[i, j]$ is constant. On the other hand, dynamic visualizations - also referred to as animations - represent mutable windows, or samples, of data. Dynamic visualizations adjust a geometric representation in real time, depending on the arrival of new data.

The second type of endpoints is called a dataset. Datasets are a repository of transformed data fragments, as previously described in Figure 9, as either an entry point of virtually infinite streams of data fragments x, or a destination of the execution of stream $\phi(P_i, \delta_i)$.

Datasets can serve as intermediary entry and exit points of multiple sub-graphs or streams $\phi(P_i, \delta_i)$. In that sense, the resulting dataset of one execution stream can be a source dataset in a second, different, execution stream.

4.5. CONSTRAINTS OF DATA

According to the representational process defined in Section 3.4.1, the third component of a computational representation is called constraints of data. Constraints of data are explained in details in Section 3.4.4. Constraints of data define rules of association that establish what is feasible in a domain of knowledge. Those structural constraints use an abstract layer of data to define restrictions on a separate abstraction, based itself on data, hence the term meta-data. The set of structural constraints in a specific domain of knowledge is called meta-model.

For our domain of interest, financial sciences, structural constraints for associations between contributions and facets are defined in three of different groups of meta-models, based on its particular use: configuration, execution or simulation meta-model.

- **Configuration meta-model**: represents a versioned snapshot of a configuration of facets, arranged in a graph, over time. In other words, a structural description of all graphs defining the execution steps of a specific financial model. Since execution flows, or graphs, can change over time, a versioned configuration meta-model allows the exact definition and

reproducibility of execution flow, at any given moment in the past. Instances of this meta-model will determine a reproducible sequence of execution, versions and provenance tracking of all data used to generate any specific result set.

- **Execution meta-model**: represent fragments of hierarchical data that flow through one or more compatible steps of a model. Instances of an execution meta-model are related to one specific configuration meta-model. In a sense is a description, in structured data, of concepts inherent to financial sciences: entities, contracts, instruments, or relationships. An example of an execution meta-model is given later in the investigation exercise in Chapter 5, specifically on references to 'fitness', 'Ticker', and 'Adj. Close' features, in Contribution 22 on page 195.

- **Simulation meta-model**: supports the registration of experiments, results, and methods required to support an investigation. The registration is a permanent ternary association between financial models, shocks, and benchmarks. A financial model is a contribution describing the problem domain, the hypothesis under test, and the method under verification. The background for the definition of the hypothesis under tests and methods are part of the proof pipeline, previously described in Section 3.2.1. Shocks describe each of the executions of a financial model, used for recording utilized data, and results of each individual execution. Benchmarks describe the final comparison of results, of different shocks, and outline conclusions. An example of shocks and benchmarks on a Monte Carlo simulation is given later in the exercise in Chapter 5, in Contribution 9, on page 179.

It is important to note that meta-models are defined and dependent on a finance case of use, or exercise, and should be defined in an ad hoc fashion, as required. To define an extensive set of meta-models that could be used in a large number of financial use cases is not practical, and would yield no additional insights to justify the increased complexity.

Additionally, some cases of use might require a partial set of meta-models. For example, for a real-time stock pricing financial model, given a strict dependency on mid-prices and a static price calculation function, a simulation and a configuration meta-model would not be necessary. A financial model for this specific case of use can rely exclusively on an execution meta-model. A more extensive example is given in the investigation exercise in Chapter 5.

4.6. INVESTIGATION CASES OF USE

At this point, we have defined a computational representation for the field of economics following the representational process outlined previously in Section 3.4.1 given by facets, contributions, and constraints of data, respectively defined in this chapter under sections 4.3, 4.4, and 4.5. This representation enables a platform for an objective application of the scientific method that can be utilized efficiently in three specific scenarios, exemplified next: application of the scientific method, large-scale collaboration across a heterogeneous community of users, and support for simulation in economics.

On this section we present these three scenarios as examples of each of these cases of use, picturing a hypothetical exercise for the selection of fast learning methods for neural networks (Castillo, et al. 2006).

4.6.1. APPLICATION OF THE SCIENTIFIC METHOD

Researcher A is studying a new method for fast learning of neural networks based on sensitivity analysis. The study intends to find a fast learning method while predicting a Dow-Jones index for a given day using a historical dataset from 1994 to 1996. For that, Researcher A records in the platform the following contributions:

- Hypothesis H_1 and null hypothesis H_0;
- Six pluggable implementations, one for each of the possible problem resolution scenarios: standard algorithms, linear least square, second order, adaptive step size, appropriate weights, and rescaling.
- Pluggable implementation of a one-layer neural network
- Input datasets for the time series of price variations of the index, and output dataset registering learning time and fitting

Researcher A executes and collects the resulting data once for each of the six scenarios. For each cycle of execution and data collection, it is only necessary to switch the pluggable implementation of the resolution scenario.

Researcher A collects output datasets for each execution and contributes the results to the platform. He also records the final findings of his experiment, stating that adaptive step size is the fastest method and provides the best fitting overall.

From this point, their findings will be unquestionably and transparently bound to the method he followed (model), input data and findings, so any other participant in the community can leverage, inspect or challenge his findings.

4.6.2. LARGE SCALE COLLABORATION

Researcher B works in the same field as Researcher A. He develops a new approach called Sensitivity-Based Linear Learning Method (SBLLM) and wants to compare the performance of that approach to previous methods bound to findings of Researcher A.

Researcher B contributes a new pluggable implementation he calls SBLLM. Researcher B adds this single implementation to scenarios already contributed by Researcher A and re-runs what are now seven scenarios.

Researcher B records the new revision of the original financial model, what now brings an additional scenario, and records the results of the experiment: SBLLM is now the fastest learning method and provides the best fitting. From that point on any researcher in the world can participate in this scientific search for better learning methods.

4.6.3. SIMULATIONS

Researcher C is also studying the same subject as Researcher A and Researcher B. He believes previous findings are somewhat flawed because they failed to take into consideration a considerable number of independent variables.

Researcher C defines a new hypothesis stating that performance of learning methods are indeed affected by at least two independent variables:

- Number of layers N of the subject network;

- Generic input as a random walk of drift D and variance σ

Researcher C maps each independent variable to a parameter, defining parameters N, D, and σ, and contributes a new pluggable implementation of a neural network of N layers, as well as a random walk generator for drift D and variance σ. He also contributes a new revision of the model created by Researcher A, later augmented by Researcher B. This revised model accounts now for the random walk generator and a variable layer neural network.

Researcher C executes one run of all seven scenarios for every permutation of N, D, and σ as well as historical data of Dow-Jones from 1994-1996 as a baseline.

Researcher C contributes back to the platform his output datasets as well as an explanation of his final findings: SBLLM is a faster implementation for $N = 1$ for reasonable values of D and σ. For $N > 1$, other methods were found to perform better. Researcher C starts working on research explaining possible causes.

4.7. FRACTI

As previously explained in Chapter 3, enablers for *crowd-based scientific investigation* define what is required, or should be in place, to enable an investigation to occur across a large number of participants in a crowd (Franzoni and Sauermann 2014). There are two types of enablers: cognitive and non-cognitive enablers.

As we have explained in Chapter 3, cognitive enablers are domain independent, or in other words, should be the same regardless of the characteristics of the domain of knowledge under consideration. This research considers two cognitive

enablers, we call *methods of proof* and *collaboration in large-scale*, defined respectively in sections 3.2 and 3.3.

The non-cognitive enabler, on the other hand, is strongly dependent on the specifics of a domain of knowledge. The only non-cognitive enabler for crowd-based scientific investigation is called a computational representation. A computational representation is a representation system based on facets, contributions, and constraints of data used to define concepts related to a specific domain of knowledge. A computational representation is defined for a specific domain of knowledge based on the *representational process* described in Section 3.4.1. In this chapter, we applied this representational process to define a computational representation for the field of economics.

The computational representation defined in this chapter is the non-cognitive enabler that would allow a crowd-based scientific investigation to occur in economics. The use of this computational representation and other enablers for crowd-based investigation defined in Chapter 3, for support for scientific investigation and collaboration in large scale, brings tangible benefits:

- **Transparent Collaboration**: support for the transparent description of structure and contents of large datasets following a standard method of representation and visualization. Shared items can be examined in details and re-executed against different scenarios by different groups of users. Collaborators can easily define and backtest their hypotheses, support sharing**,** tracking and provenance of contributions, ensure that results are replicable and in this sense, playing a role of a scientific support system (J. M. Faleiro Jr 2013a) (Goecks, Nekrutenko and Taylor 2010).

- **Reproducibility**: a scientific approach to analytical research, in which reproducibility becomes a hard scientific requirement: models and scenarios have to be reproducible by anyone. Large sets of data and models can be re-executed, allowing different organizations and individuals to replicate results easily

- **Accessibility**: end users do not have to be proficient in computer science to be able to use, collaborate or visualize models or scenarios in the framework

- **Openness**: Instances of the meta-model representing a specific configuration, execution or simulation can be exchanged across environments or different implementations. Data and method of an investigation can be traced regardless of ownership, origin or location of a contribution.

From now on and throughout this thesis, we will refer to the set of cognitive and non-cognitive enablers of crowd-based investigation in the field of economics as *FRACTI* (Faleiro Jr and Tsang 2016a).

FRACTI stands for *FRAmework for Collaboration and Transparent Investigation in economics*, and it is an abstraction designating the three enablers of crowd-based investigation for the field of economics: methods of proof, requisites for large-scale collaboration, and the computational representation defined in this chapter. It is essential to clarify that FRACTI is a conceptual abstraction, and **is not a software implementation, or a programming language**.

4.8. CHAPTER SYNOPSIS

This chapter fulfills the criteria of success described in Objective 2 on page 20 of this thesis by defining a computational representation to support investigation and collaboration in large-scale for the field of economics. As previously defined in

Section 3.4, a computational representation is a representation system based on facets, contributions, and constraints of data, and used to define concepts related to a specific domain of knowledge for crowd-based investigation.

The computational representation for the field of economics was defined in this chapter by application of the generic representational process described in the previous chapter, specifically in Section 3.4.1. According to that representational process, a computational representation is closely tied to a domain of knowledge and defined based on domain-specific requirements and a computational taxonomy. Requirements for the field of economics are defined from financial sciences cases of use, described in Section 4.2. The computational taxonomy is described during definition of each of the facets in sections 4.3.1, 4.3.2, 4.3.3, and 4.3.4. As previously introduced in Section 3.4, a computational representation is given by three complementary components called faces, contributions, and constraints of data.

Facets are the definable aspects that make up a subject or an object and were previously introduced in Section 3.4.2. The facets of the field of economics are *streaming*, *reactives*, *distribution* and *simulation*, and are defined respectively in sections 4.3.1, 4.3.2, 4.3.3, and 4.3.4.

Contributions are shareable and formal evidence of an objective crowd-based investigation and were previously introduced in Section 3.4.3. Three specific classes of contributions define the taxonomy of contributions for the field of economics, namely *financial models*, *processors*, and *endpoints*. Each of these classes of contributions for the field of economics is defined respectively in sections 4.4.1, 4.4.2, and 4.4.3.

Constraints of data are structural constraints defining domain-specific rules of association between entities and relationships and were introduced in Section 3.4.4.

For the field of economics are proposed three types of meta-model: *configuration*, *execution*, and *simulation*, defined in Section 4.5.

This chapter brings several novelty contributions of this research, specifically:

- Definition of a computational representation based on the formal steps of the representational process previously defined in Section 3.4.1;

- Definition and formalization of a computational representation for the field of economics through the definition of facets, contributions, and constraints of data, respectively defined in sections 4.3, 4.4, and 4.5;

- Formalization of financial models through models of computation of graph-based streams, or agents, in sections 4.3.1.1 and 4.3.1.2.

The computational representation for the field of economics defined in this chapter and the cognitive enablers defined in Chapter 3 are jointly referred to as FRACTI, a FRAmework for Collaboration and Transparent Investigation in Economics, defined in Section 4.7. FRACTI is used in the upcoming Chapter 5 to formalize, investigate and resolve a non-trivial problem in economics.

CHAPTER 5. INVESTIGATING THE PROFITABILITY OF MOMENTUM TRADING STRATEGIES

"In God we trust; all others must bring data." – Edwin R. Fisher (E. Fisher 1978)[48]

This research chapter defines an end-to-end investigation exercise to measure the actual efficiency of a momentum strategy in technical analysis using historical trading data and the formal methods developed in previous chapters of this thesis. These methods were introduced in Chapter 3 when we define enablers for crowd-based investigation, and in Chapter 4, when we present a computational representation for the field of economics.

5.1. THE PROBLEM

Now has finally come the time to gather all the methods and tools defined in previous research chapters in this thesis and apply them to resolve a concrete problem in financial sciences. In this chapter, we present a real end-to-end example of a non-trivial investigation of the performance of a financial trading strategy using FRACTI, as defined previously in Section 4.7.

As previously explained in Chapter 3, enablers for crowd-based scientific investigation define what is required, or should be in place, to enable an investigation to occur across a large number of participants in a crowd, what we call *crowd-based investigation* (Franzoni and Sauermann 2014). There are two specific types of cognitive enablers, and by definition cognitive enablers are not dependent on specifics of a domain of knowledge. As a consequence, cognitive enablers should be the same across different domains of knowledge: psychology, architecture, physics,

[48] This quote is often erroneously attributed to William E. Deming (O'Toole 2017) (Walton 1986)

mathematics, etc., all should rely on the same characteristics of cognitive enablers if they wish to leverage crowds for investigation. These cognitive enablers are *methods of proof* and *large-scale collaboration* and were defined previously in sections 3.2 and 3.3.

Non-cognitive enablers, on the other hand, are strongly dependent on the specifics of a domain of knowledge and should be different depending on the domain of knowledge under consideration. In other words, we should not expect a non-cognitive enabler defined specifically for the field of economics to be directly applied to some other unrelated area, like psychology or engineering.

As explained before in Section 3.4, the non-cognitive enabler for crowd-based investigation is called a computational representation. A computational representation is a representation system based on facets, contributions, and constraints of data, and used to define concepts related to a specific domain of knowledge for crowd-based investigation.

A generic process to produce a computational representation is called a representational process, as previously defined in Section 3.4.1. A representational process produces computational representations described as a set of facets, contributions, and constraints of data, based on domain-specific cases of use and a computational taxonomy. In Chapter 4 we used the computational process to define a computational representation specific to the field of economics.

As explained in Section 4.7, we are referring to the cognitive and non-cognitive enablers for the field of economics, already defined in Chapter 3 and Chapter 4 as FRACTI.

In this chapter, we use FRACTI to perform an end-to-end exercise of investigation in financial sciences: the profitability of a specific trading strategy. We

consider this chapter an important formalization that will help consolidate the vision of this thesis, by providing concrete examples of the application of enablers for crowd-based investigation, as defined in Chapter 3.

We have been emphasizing that FRACTI **is not a software implementation\ or a programming language.** However, for the specific set of examples outlined in this chapter we have to rely on one particular implementation, or dialect, of FRACTI. This dialect is an open source platform called QuantLET (J. M. Faleiro Jr 2008). This dialect has evolved to serve as an illustration of concepts in FRACTI and is not intended to be a full-fledged implementation of all the concepts of FRACTI at this point. Throughout this work, the dialect has been providing valuable insights to this research, and vice-versa, this research fed back to the dialect several ideas that were materialized as concrete extensions. This dialect relies on many underlying computational resources (Pérez and Granger 2007) (Jones, Oliphant and Peterson 2001) (McKinney 2010) (Hunter 2007). Since the chain of direct and indirect dependencies is fluid and always changing, this list is not final.

We will be using financial models, as described in sections 4.4.1 and 4.3.1.2, to investigate if one common trading strategy prevalent in technical analysis is indeed profitable, and if so under which circumstances. Even for a somewhat complex and extensive exercise, we will show that financial models in FRACTI are a succinct and straightforward way to describe and communicate details and features of the trading strategy.

Throughout this investigation, you will learn how financial models are built as streams, by chaining processors, and endpoints, as described in sections 4.3.1.2, 4.4.2, and 4.4.3. As we had previously described in Section 4.4, financial models, processors and endpoints are contributions. As we have described in Section 4.4, contributions are shareable and formal evidence of an objective, crowd-based

investigation, and as evidence they can be exchanged, reused and traced through something called a *record of provenance,* therefore, becoming a vehicle for collaborative scientific investigation. For the convenience of the reader, shareable evidence of this exercise are captioned as *Contributions*, as per definition provided in Section 3.4.3, and a complete list of all contributions is supplied at the beginning of this thesis, on page xiii.

This exercise will also demonstrate how easy it is to adjust a financial model to perform various functions in the investigation. The example of this exercise will illustrate how slight adjustments can make the same model serve for tasks of visualization, benchmarking, simulation and even real-time trading.

This chapter is broken down into four specific parts, outlining the process of an investigation described in Section 3.2.

The first part, in Section 5.2 defines the fundaments of the scenario under investigation, a common technical trading strategy called *breakthrough crossover momentum strategy*. This exercise will dissect this technical strategy in the search for scenarios of profitability on two steps: first against random walks, and second against real historical data from components of a liquid index, the S&P 500 index.

The second part in Section 5.3 describes each of the components of a breakthrough crossover momentum strategy, presented in the first part, as FRACTI contributions. FRACTI was previously introduced and defined in Chapter 4.

The third part, in Section 5.4 describes the end-to-end exercise of investigation using the proof pipeline defined in Section 3.2.1: formulation of a hypothesis; outlining of limitations, simplifications, and expected outcomes; development of a model and components; definition of shocks and simulations; execution of simulations; and finally formulation of conclusions. During this exercise

we generate a number of contributions – as extensively mentioned before, contributions are a shareable record of scientific evidence – and a step-by-step analysis of that evidence. Throughout those steps, we will demonstrate how easy it is to formalize financial models and adjust them to cover a broad spectrum of functions, from visualization to benchmarking, to historical simulation. The definition and execution of financial models generate a number of relevant shareable and traceable contributions.

The last and fourth part of this chapter in Section 5.5 is an overall analysis of the investigation, in which by the end of the process, we list a number of possible explanations for our findings, given evidences produced.

5.2. MOMENTUM TRADING STRATEGY UNDER THE MICROSCOPE

The concrete example described in this chapter investigates a common trading strategy, a variation of a Moving Average Cross-Over (MAC-O) momentum strategy (Schoeffel 2011). This type of hybrid strategy uses *breakthrough* signals to identify *momentum* of a pseudo-random movement. A momentum indicates a tendency of a price movement to remain moving one way, up or down. We call this variation of MAC-O the Breakthrough Cross Over Momentum, or BCOM, strategy.

Breakthrough strategies are the most commonly used strategies in technical analysis and electronic trading and are strongly based on the assumption of momentum. Momentum strategies (Chestnutt 1955) identify profit opportunities by assuming that, unlike a purely random movement of a price (random walk), a price movement might carry some "inertia" and tend to gain on an already higher price, and in the assumption that in "many more times than not … the strong get stronger and the weak get weaker" (Chestnutt 1955).

Technical analysis is usually seen as an alternative to two other types of analysis called Fundamental and Quantitative analysis. Despite its continuous use in modern days, technical analysis has its origins on the Dow Theory and the works of Cowles (Cowles 1933), Dow and Hamilton (Hamilton and Dow 1922) (Hamilton 1903-1929) from early twentieth century. Since that early time, the effectiveness of technical analysis when compared to fundamental and quantitative analysis has been a matter of controversy, with references testifying for the efficacy of technical analysis (Patterson 2007) (Brown, Goetzmann and Kumar 1998) and against it (Marshall, Cahan and Cahan 2010). Different researchers test variations of these strategies using different data, which makes it very difficult to compare and contrast different tests and results used in each of these analyses.

On this exercise, we will be using FRACTI, as defined in Chapter 4, and financial models as previously described in sections 4.3.1.2 and 4.4.1 to determine the efficiency of the BCOM strategy as an example of the performance of one method of technical analysis. Throughout this investigation, we will move our biases and our pre-conceptions aside, and expect readers to do the same. We may not be able to offer a definitive conclusion for this investigation vis-à-vis evidences collected. The primary intent of this exercise is to serve as a showcase of the methods of crowd-based investigation proposed in this research.

Financial models that rely on breakthrough strategies intend to predict the behavior of a random – or pseudo-random - walk based on fluctuations of this walk crossing (or breaking through) the line given by attenuations of its past values. To establish momentum and the breakthrough, this type of financial models rely on four specific components: random walks, moving averages, rules for generation of buy and sell signals and portfolio management. Over the next sections, we will formalize each of these components.

5.2.1. RANDOM WALKS

The first components of BCOM strategies when we intend to evaluate its effectiveness when a series of prices on the underlying is ideally efficient are random walks. A typical random walk is a path of successive values following some efficient random generation pattern. The random walk theory, in essence, states that the path of prices over time is *efficient* when past prices do not influence future prices (Fama 1965).

Although random walks can be specified on multiple dimensions, a random price movement can be properly modeled as a one-dimension random walk. One-dimensional random walks are used extensively for simulation of stochastic movement of asset prices over time. On this exercise, we will specifically use a Brownian motion (R. Brown 1827), or Wiener process (Mandrekar 1995), characterized by the following properties:

$$W_0 = 0$$
$$W_t \sim N(0, t-s) + W_s$$

Equation 13: Brownian Motion

Where $t \rightarrow W_t$ is continuous, and $N(\mu, \sigma^2)$ is the normal distribution with expected value μ and variance σ^2. For any t, s, and u, $W_t - W_s$ and W_u are independent for $u \leq s < t$.

In FRACTI parlance, we can illustrate a random walk using a straightforward financial model of just one stream, as described in Section 4.3.1.2, and in one sentence, as shown in Contribution 1.

```
ts('2013-01-01', '2013-12-31') \
>> brownian(seed=42, s0=37) \
>> plot
```

Contribution 1. Random Walk Over a Time Series

A contribution of a financial model describing a random walk and a time series as one stream, showing two endpoints ts and $plot$, and one processor $brownian$

On this dialect, we represent one stream connecting a daily time-series from beginning to end of 2013, a Brownian motion random walk[49] for $W_0 = 10$, as defined in Equation 13. This financial model - like any other financial model in FRACTI, as explained in Section 4.3.1.2 - is a stream. As any financial models in FRACTI, it has precisely the same format of a stream formalized in Equation 2 in page 106, with the synchronicity operator δ in this specific dialect carrying a connotation $\gg$ to indicate synchronous communication.

The resulting contribution – as a reader might recall from definitions set forth previously in Section 4.4.3, plots are also an endpoint contribution – is a bi-dimensional plot shown in Contribution 2.

[49] On all instances of this exercise "seed" arguments are provided in order to achieve repeatability of results of a random series across different executions of the exercise (what in reality makes samples on this exercise pseudo-random and deterministic).

Contribution 2. Single One-Dimension Brownian Motion

A one-dimensional Brownian motion representing a pseudo-random price movement of an asset in a daily time series from 01/Jan/2013 to 31/Dec/2013

Additionally, for simulation purposes, the original financial model given in Contribution 1 can be extended to take into account multiple Brownian movements over a time-series. In Contribution 3 we generate in one stream a series of prices mimicking random walks for closing prices of three symbols: GOOG, IBM, and B:

```
q.seed(42)
ts('2013-01-01', '2013-12-31') \
>> brownian(s0=37, output='GOOG') \
>> brownian(s0=21, output='IBM') \
>> brownian(s0=42, output='B') \
>> plot
```

Contribution 3. Multiple Random Walks Over a Time Series

One financial model as a stream, showing one endpoint ts, defining a time series from 01/Jan/2013 to 31/Dec/2013, three instances of the same processor $brownian$, one for each symbol, and a final endpoint $plot$ for visualization of the combined set of random walks

This financial model defines a stream connecting from a daily time series for the entire year 2013 to three separate random walks with different W_0 for GOOG, IBM, and B. Like before, all results aggregated over a bi-dimensional plot, shown in Contribution 4.

Contribution 4. Multiple One-Dimension Brownian Motions

One endpoint contribution, a visualization, showing three different random walks simulating three symbols B, GOOG, and IBM

As we had previously described in Section 4.3.1.2, financial models are described as streams connecting processors using synchronicity operators denoted by δ. As it is shown in Contribution 1 and Contribution 3, this dialect represents the synchronous communication between processors by the operator $\gg$.

5.2.2. Signal Attenuation

The second major component in evaluating a BCOM strategy is the generation of an attenuation signal. The attenuation of a signal removes, or filters, variations of values around an intermediate point between high and low values. There are multiple ways to attenuate a signal, each providing different characteristics to the

filtered curve. On this specific exercise, filtering, or dampening, of signals is achieved by running or rolling averages[50] of past prices in a time series updated on each new tick of the price.

Moving averages in their basic form are computing intensive. For every new price, all past prices would have to be traversed and a new average calculated. To avoid this potentially inefficient computation, we skip the calculation of all values by acting on values on the tail of the series, and some residual representing past values.

We call this a *recursive representation*, and this exercise aims to represent calculations in a recursive form. In a recursive representation, iterations over past values of a series, in each calculation, are not necessary.

The most commonly used moving averages are cumulative, rolling, weighted, and exponentially weighted moving averages (Croarkin and Tobias 2012).

[50] Arithmetic mean

5.2.2.1. Cumulative Moving Average

The simplest moving average is an arithmetic average of the previous n values in a series. The non-recursive and recursive calculation are given by Equation 14:

$$CMA_n = \frac{1}{n}\sum_{i=0}^{n} x_i \quad (non\text{-}recursive)$$
$$CMA_n = \frac{1}{n}\left(x_n + (n-1)\,CMA_{n-1}\right) \quad (recursive)$$

Equation 14. Cumulative Moving Averages

A cumulative moving average is a particular case of moving average where there is no sampling window, so all data is considered equally in the calculation of the moving average.

5.2.2.2. Rolling Moving Average

The "rolling" average is an un-weighted mean of previous samples, considering a sampling window of size m'. A more precise definition accounts for n where $n < m'$, and the window is given by $m = \min(m', n)$. Equation 15 gives the non-recursive and recursive forms.

$$RMA_n = \frac{1}{m}\sum_{i=(n-m+1)}^{n} x_i \qquad (non\text{-}recursive)$$
$$RMA_n = RMA_{n-1} + \frac{1}{m}(x_n - x_{n-m+1}) \qquad (recursive)$$

Equation 15. Rolling Moving Averages

5.2.2.3. Weighted Moving Average

Weighted averages have a dampening factor (m) assigning different weights to values at different positions in a moving sample window of size m'. A more precise definition of the dampening factor, considering n samples, where $n < m'$, the dampening factor is given by $m = \min(m', n)$.

In this moving average m past factors are adjusted by a decreasing linear factor $(m - n + 1)$, so that more recent observations have a larger influence on the filtered signal. Its form is given by:

$$WMA_n = \frac{m.x_n + (m-1).x_{n-1} + \cdots + x_{n-m+1}}{m + (m-1) + \cdots + 1}$$

Where the denominator $\sum_{i=0}^{m-1}(m-i)$ is a triangular number (Weistein 2015) that can be reduced to $\frac{m\,(m+1)}{2}$, giving the model a final form shown in Equation 16. This model is partially recursive from 0 to the value of m for each iteration n.

$$WMA_n = \frac{2}{m\,(m+1)} \sum_{i=1}^{m} i\ .x_{n-m+i}$$

Equation 16. Weighted Moving Averages

5.2.2.4. Exponentially Weighted Moving Average

In the exponentially weighted moving average, a dampening factor α is used to decay older terms of the series. Its recursive form is given by:

$$EWMA_n = \alpha\, x_n + (1 - \alpha)\, EWMA_{n-1}$$

Equation 17. Exponentially Weighted Moving Averages

Where $0 < \alpha < 1$. Higher dampening factors (α) would give lower weight to older terms of the series, yielding slower filters. This model is recursive by definition and therefore easily adapted to computational forms.

5.2.2.5. Representation of Moving Averages

For the representation of any dampening effect in a time series, we can bind together multiple filters to a random walk and a visualization endpoint on the same stream using one single line, shown in Contribution 5.

```
ts('2014-1-1', '2014-12-31') \
>> brownian(seed=42, s0=37) \
>> ewma(alpha=0.2) \
>> rma(m=20) \
>> cma \
>> plot
```

Contribution 5. Stream for Visualization of a Random Walk

A financial model as a stream representing one time-series of random values, and three attenuation processors for exponentially weighted moving, rolling, and cumulative moving averages. The resulting visualization is sent to a plot endpoint.

In Contribution 5 is defined a financial model for a random walk with $W_0 = 37$ over a time series from 1/Jan/2014 to 31/Dec/2014[51]. Over this random signal is added three different filters: $ewma$ (exponentially moving average, in Section 5.2.2.4) with $\alpha = 0.2$, rma (rolling moving average, in Section 5.2.2.2) with $m = 20$ and cma (cumulative moving average, in Section 5.2.2.1). All results are chained to a plot endpoint for visualization, shown in Contribution 6.

[51] In Contribution 5, the parameter W_0 is represented on the dialect as $s0$.

Contribution 6. Multiple Filters Over a Random Walk

A visualization of three different attenuation filters of a random walk, given by signal $brownian$. The attenuation filters $ewma$, cma, and rma provide different levels of attenuation for the original random signal.

Momentum strategies tend to rely on filters that provide some control over how much decay should be applied to older terms of the series. Additionally, as explained before, on page 162, best results are achieved from filters that can be represented recursively. Given these characteristics, we will be using an exponentially weighted moving average as a filter for the BCOM strategy on this investigation.

5.2.3. DERIVATION OF MARKET SIGNALS

The third major component for evaluating a BCOM strategy is associated with mathematical rules for generation of market signals. Market signals are by definition an automated instruction to the market of either a buy or sell signal. In a BCOM strategy, any buy or sell signals S_i are generated depending on how a *fast* curve of prices F_i crosses over a *slow*, or dampened, curve D_i. The dampening effect is achieved by applying one of the filters described in Section 5.2.2.

In essence, a buy signal is generated whenever the fast curve crosses the slow curve upward. On the other hand, a sell signal is generated whenever the fast curve crosses the slow curve downward, as formalized in Equation 18.

$$S_{t+1} \rightarrow SELL \quad if\ D_{t-1} > F_t\ and\ D_{t-2} \leq F_{t-1}$$
$$S_{t+1} \rightarrow BUY \quad if\ D_{t-1} < F_t\ and\ D_{t-2} \geq F_{t-1}$$

Equation 18. Model for Derivation of Market Signals Using Cross-Overs

Where $(t-2, t-1, t)$ is the relevant tuple of three steps in time, shown in a sequence that is relevant to determine if the signal S in the immediately subsequent step $t+1$ will be either a sell or a buy signal.

A clear illustration of this behavior is given later in this investigation exercise, in Contribution 8 on page 174. On that example, red triangles denote $SELL$ signals, while green triangles denote BUY signals.

In Contribution 7, on page 172, we illustrate the application of this full strategy as the introduction of a $maco$ (moving average cross over) processor, which takes a time series as input and outputs a set of buy and sell signals.

On this model, there are two possible variations of rules for the derivation of signals, depending on what makes the faster and slower curves. These variations are called double and single crossover rules.

In a double crossover variation, both the slower and faster curves are themselves filtered curves of a spot price curve. The slower curve has a higher dampening factor (α) than the slower curve. For example, if we were to use moving averages, as defined in Section 5.2.2.1, as dampening filters, the fast curve MA''_i and slow curve MA'_i would differ by how far back in the past spot prices would be used in the average calculation. In other words, slow-moving averages track longer periods than fast moving averages does, i.e.:

$$F_i \approx MA''_i$$
$$D_i \approx MA'_i$$

In a single crossover variation, the faster curve is given by the spot price over time F_i, what could be for example the random walk W_i. The slower curve is given by the dampening of the spot price curve through some filter, usually a moving average MA_i, i.e.:

$$F_i \approx W_i$$
$$D_i \approx MA_i$$

On this investigation exercise, we will be using a single crossover variation, in which the model takes the final form in Equation 19.

$$S_{t+1} \rightarrow SELL \quad if\ MA_{t-1} > W_t\ and\ MA_{t-2} \leq W_{t-1}$$
$$S_{t+1} \rightarrow BUY \quad if\ MA_{t-1} < W_t\ and\ MA_{t-2} \geq W_{t-1}$$

Equation 19. Model for Derivation of Market Signals Using a Single Cross-Over

Where W_i, MA_i, and S_i are respectively values of the stochastic random walk, moving average filter, and a buy or sell signal at time $t = i$.

5.2.4. PORTFOLIO MANAGEMENT

The fourth and final major component of a BCOM strategy should account for portfolio management. Portfolio management is a comprehensive term that, in the scope of this exercise, refers to the act of keeping track of overall gains and losses of the application of any given trading strategy. On this model we will be tracking profitability by taking into consideration four quantities, as shown in Equation 20: the cash flow resulting from buys and sell signs, transaction costs, a cash balance, and price variations of the underlying (Markowitz 1952) (Markowitz 1959).

$$(P, B, \alpha)_t = pm(K_t, S_t, L_t, W_t)$$

Equation 20. Model for Portfolio Management

The function pm is the portfolio management function where the arguments K_i, S_i, L_i, and W_i are respectively the cash balance, signal (Buy or Sell), load (transaction costs) and the value of the stochastic random walk at time $t = i$. This function pm takes buy and sell signals generated by the crossover momentum strategy as input. Based on the cash balance (K), the function pm outputs a 3-tuple

$(P, B, \alpha)_t$ of the model at time t. This specific version of portfolio management is referred in the dialect as a processor called $cash\_stock$, so for all purposes of this exercise they are equivalent, i.e., $pm \approx cash\_stock$.

In this resulting tuple $(P, B, \alpha)_t$ on time t, the value P is the final signal to be sent to the marketplace (Buy, Sell or Nothing). The full portfolio balance B accounts for the summation of cash and non-cash positions. The rate α is the overall profitability of the model at t.

5.3. REPRESENTATION IN FRACTI

At this point, we have formalized the BCOM strategy using its core components and equations in Section 5.2, and have explained FRACTI in Section 4.7. In this section, we will combine these two pieces of knowledge to represent the required models for this investigation exercise using FRACTI facets and contributions.

As explained in Section 4.3.1.2, financial models in FRACTI are represented by a particular type of facet: a stream. A financial model as a stream is described in using steps, associated to either processors or endpoints, detailed respectively in previous sections 4.4.2 and 4.4.3.

Using that approach any financial model, regardless of its complexity, is described in terms of steps, performed by processors. Despite the relative complexity of the BCOM strategy, explained in Section 5.2, a financial model to benchmark the strategy's performance can be outlined by a specific set of steps:

- Generate price ticks, either from random Brownian generators or historical data, as described in Section 5.2.1;

- Generate moving average, preferably allowing the plugging-in of different types of moving averages to enable different comparisons and benchmarks, as described in Section 5.2.2;

- Generate signals buy or sell, based on breaks of the spot price curve through the attenuated signal, as described in Section 5.2.3;

- Decide whether to send an order to the market or not, based on current portfolio and liquidity (balance of available cash), as described in Section 5.2.4;

- Processing the resulting test data (e.g., plotting, storing).

The financial model associated with a benchmark of the BCOM strategy will connect all components defined in Section 5.2, in one line, as shown in Contribution 7.

```
ts('2013-1-1', '2014-12-31') \
>> brownian(seed=42, s0=37) \
>> ewma(alpha=0.05) \
>> maco \
>> cash_stock(initial_cash=10000, load=7.5) \
>> plot
```

Contribution 7. Breakthrough Momentum Strategy Model

A financial model for the use of a BCOM strategy for a time series from 1/Jan/2013 to 31/Dec/2014, for values in a random walk, attenuation signal given by an exponentially weighted moving average, a BCOM implementation given by a moving average crossover, and a portfolio management function based on cash and stock value.

This brief representation of this contribution is a financial model as a one-line stream. Starting from a first step defining a time-series for years 2013 and 2014 (ts), following to a simulation of a stochastic random walk of closing prices ($brownian$), as described in Section 5.2.1, with $W_0 = 37$. The result is sent to the third step of an exponentially weighted filter ($ewma$), as described in Section 5.2.2, with $\alpha = 0.05$. The result on a next step is a moving average crossover ($maco$) for cross-over signals, as described in Section 5.2.3; and a portfolio management function ($cash\_stock$), as described in Section 5.2.4, with an initial cash balance $K_0 = \$10{,}000$ and load per transaction $L_i = \$7.5$.

The resulting contribution of the execution of this stream is a visualization plot giving the first quick and structured glimpse into what to expect from an execution of a BCOM strategy, shown in Contribution 8.

Contribution 8. BCOM Performance of Random Prices

The visualization of a financial model representing the performance of the BCOM strategy, in which alpha indicates performance, or profitability if prices are to be considered efficient by following a pseudo-random walk. Green triangles indicate a buy signal, and red triangles indicate a sell signal over time.

In blue the random walk simulating closing prices of an underlying instrument, and in green the attenuating EWMA filter. The green and red triangles show when buy and sell signals are sent to the market. Finally in red, on the right vertical axis, in percentage points, is the profitability over time of the overall strategy.

A profitable strategy would show $alpha$ (red line) greater than one, so our definition of a profitable strategy would be:

$$\alpha > 1.0$$

This first run simulates a random walk, with the specific parameters of the initial price of the underlying $W_0 = 37$, dampening factor $EWMA_\alpha = 0.05$, initial cash balance of $K_0 = \$10{,}000$, and transaction costs, or load, of $L_i = \$7.5$. In this instance, a BCOM strategy is shown as not profitable, losing about ~6.5% overall in the period, two years.

As we have extensively discussed in Section 3.2, models are in essence a simulated simplification of a real world phenomenon. This BCOM model is no exception. To allow proper simulation and study, this FRACTI representation of BCOM will assume a few important simplifications:

- **Single-symbol order book**: support for one single order book, in other words, one symbol of the underlying asset. The same conclusions of this investigation should be assumed for additional symbols following the same price behavior;

- **Infinite market liquidity**: the marketplace guarantees market orders to be fully executed over the cycle of the next price tick;

- **No price lagging**: the marketplace guarantees market orders to be executed on the last price tick received.

These simplifications help with the understanding and investigation of the problem, but should not substantially impact the generality of the experiment under study.

5.4. THE INVESTIGATION EXERCISE

By now we finally have all we need to start the investigation exercise, and will be following the methods of proof outlined in Section 3.2.1. We have explained the foundations of FRACTI in Chapter 3 and Chapter 4, and the BCOM strategy is understood and defined using FRACTI concepts, each respectively done in Section 5.2 and Section 5.3. We have built a good idea of what we intend to measure and, as a consequence, given proper evidence – or contributions - we should be able to prove or disprove. We will now move forward with the investigation per se. Steps of the procedure described here are available as a FRACTI scratchpad (Pérez and Granger 2007) and can be inspected online (J. M. Faleiro Jr 2015).

Given first brief results from Contribution 8 there are a few immediate inquiries that need to be addressed and will form the basis for the remainder of this exercise:

- Are breakthrough momentum strategies money losers? Or are they ever profitable?

- If they are profitable, what features, if any, do we need to fine tune to make them consistently profitable?

Considering these preliminary inquiries, the first step on the scientific investigation method is to state our hypothesis. We will look into our hypotheses, test, and falsifiability criteria over the next section.

5.4.1. HYPOTHESIS

Based on the proof pipeline introduced in Section 3.2.1, we specify conjectures and predictions for this investigation exercise. Given the conjecture step of the proof pipeline, defined in Section 3.2.1.2, we define one specific, falsifiable hypothesis:

> *There are scenarios under which momentum strategies are consistently profitable.*

From this conjecture, according to terms previously defined in Section 3.2.1.3, we define two specific predictions:

- For some combinations of W_0, $EWMA_\alpha$, K_0 and L_i, as defined in Section 5.2, we expect the momentum strategy to be consistently profitable.

- If profitable against a random walk, we expect the strategy to be profitable against a representative sample of financial instruments that follow a quasi-stochastic price movement path.

We will test these predictions on two primary cases. First, on Section 5.4.2, a Monte-Carlo simulation using stochastic generators on variations of arguments of W_0, $EWMA_\alpha$, K_0, and L_i; and second, on Section 5.4.3, back testing against constituents of a well-known index: the S&P 500 index (McGraw Hill Financial 2015a).

5.4.2. MONTE CARLO SIMULATION OF BROWNIAN VARIATIONS

The first part of this exercise is an attempt to examine the first portion of our hypothesis: finding out for which combinations of parameters of a random walk W_0,

$EWMA_\alpha$, and L_i we should expect the momentum strategy to be consistently profitable.

We will answer that by defining a model that executes on different shocks. As previously described in Section 4.5, the term *shock* in FRACTI nomenclature denotes *one single iteration* of a recurring simulation. In other words, each shock carries one permutation of values of features relevant for a specific simulation. In this exercise specifically, each shock carries one possible variation of values of three particular features:

- An initial value of the random walk given by $shock.s0$, representing the value of W_0 in Equation 13 on page 157 when we model and describe random walks;

- The dampness factor of the filter given by $shock.alpha$, representing the value of α in Equation 17 on page 165, when we described the model for an exponentially weighted moving averages for dampening. On this context, α on that equation is referred to as $EWMA_\alpha$[52];

- Transaction costs, given by $shock.load$, representing the value of L_i in Equation 20 on page 170, when we model and describe portfolio management as a function of cash and stock balance.

The specific financial model for this simulation, with these three features, is shown in Contribution 9. The attentive reader will notice that this Contribution 9 is very similar to the original financial model for a BCOM strategy previously defined

[52] To avoid confusion with the fitness of the BCOM strategy, denoted by α, and defined in page 163.

in Contribution 7 on page 172. This realization should not come as a surprise since essentially they both describe the same financial model.

```
def momentum_simulation(shock):
    return ts('2013-1-1', '2014-12-31') \
    >> brownian(s0=shock.s0) \
    >> ewma(alpha=shock.alpha) \
    >> maco \
    >> cash_stock(initial_cash=10000, load=shock.load)
```

Contribution 9. Simulation Model

This contribution performs a simulation of the performance of a BCOM strategy against a random walk by permutations of possible values on features of the financial model. Each permutation is given by a $shock$ instance, carrying specific features $shock.s0$, $shock.alpha$, and $shock.load$.

The similarity between in Contribution 7 on page 172 and this Contribution 9 is an example of a practical consequence of the use of streams for the representation of financial models, in which one can come from definition and execution, and from there, to a repetitive simulation of a financial model with minimum descriptive changes. One could use the same stream to switch from visualization to a Monte Carlo simulation on the same model immediately, just changing from constants to arguments in a shock.

After defining the composition of a shock, the next step is to perform a simulation. In this exercise, for illustration purposes, we decided to use uniform samples to represent variations on each of the arguments.

In a uniform distribution, if $rs(n)$ is a random sample of size n, the continuous uniform distribution in the range $[a, b)$, denoted by $Unif[a, b)$ for $b > a$, is given by Equation 21.

$$Unif(a,b) = (b-a) * rs(n) + a$$

Equation 21. Uniform Distribution

In this dialect, a continuous uniform distribution using the same arguments as in Equation 21 is denoted by $unif(a,b,n)$.

The execution environment of this dialect is responsible for generating all shocks for the simulation, covering all possible variations of features W_0, $EWMA_\alpha$, and L_i. The simulation consists of the generation of an exhaustive variation of all these features, trying to identify for which combination of features, the performance of the BCOM strategy is higher. On this financial model, the performance of a BCOM strategy is given by $fitness$, a feature denoted by α, which indicates the overall profitability of the model.

We *shock* the model defined in Contribution 9 with values in $shock.s0$, $shock.alpha$, and $shock.load$ as permutations of values in uniform distributions. The entire definition of the procedure is done in one line, in Contribution 10.

```
b = montecarlo(momentum_simulation, \
                    s0=unif(5.0, 10.0, 3),\
                    alpha=unif(0.01, 0.8, 5),\
                    load=unif(1.0, 3.0, 3))
```

Contribution 10. Shocks of Permutations of Uniform Distributions

A Monte Carlo simulation of $momentum\_simulation$, defined in Contribution 9, by generating permutations of each of the arguments as uniform distributions as defined in Equation 21.

In Contribution 10, we specify how each of the permutations of each of the features on the simulation in Contribution 9 will be generated. In this case, a Monte Carlo simulation, each of the features will take permutations of random values taken from a uniform distribution $unif(a, b, n)$ as explained in Equation 21.

In this exercise, for illustration purposes, we intend to investigate correlations using linear regressions and scatter plot matrices[53] (Friendly and Denis 2005). For the first try of the investigation, we generate the scatter plot matrix by streaming the Monte Carlo simulation using one statement, as shown in Contribution 11.

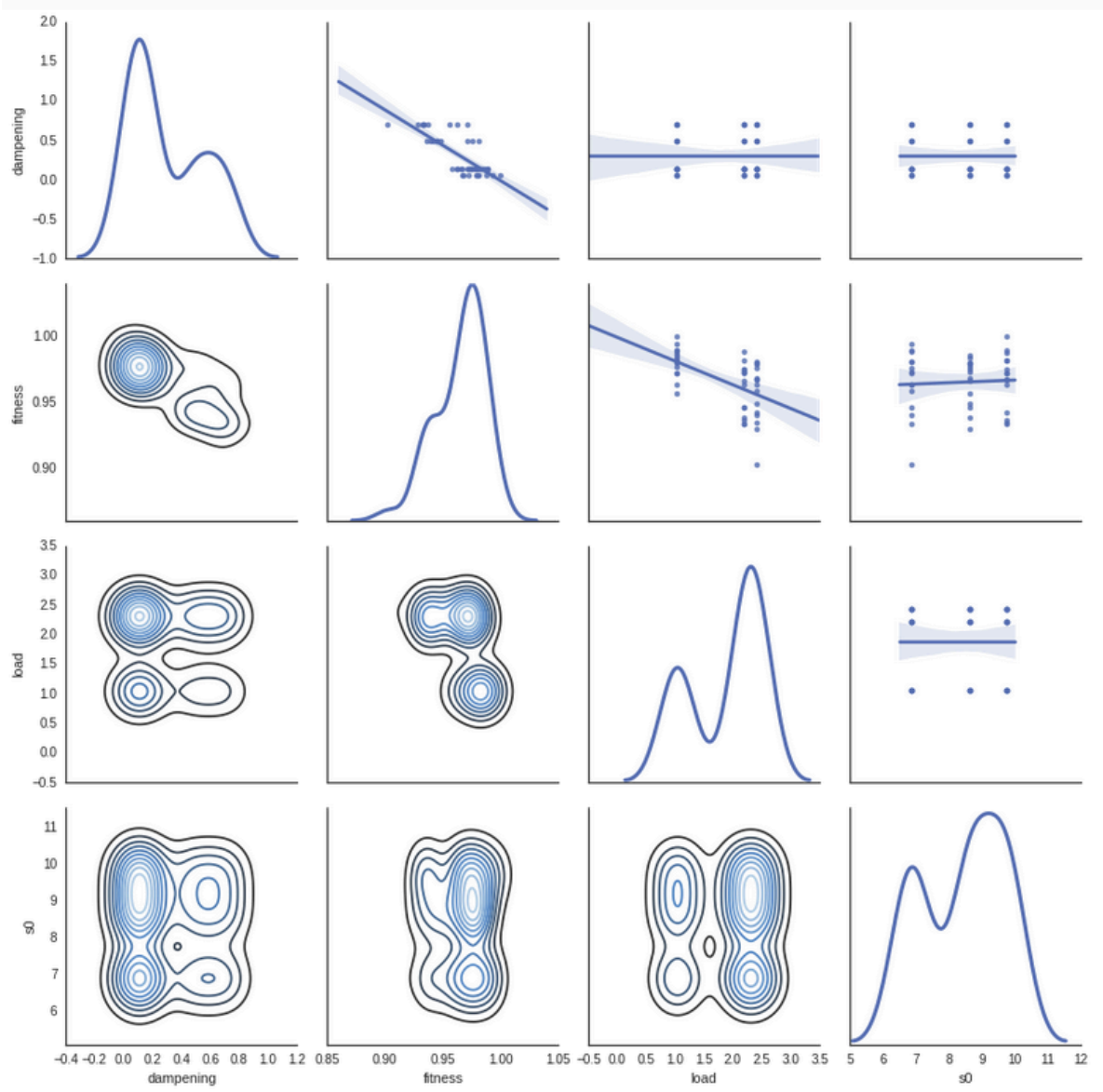
```
b >> scatter_matrix(index=None)
```

Contribution 11. Streaming a Scatter Matrix

Generation of a scatter matrix from a BCOM model previously defined in Contribution 9 to a Monte Carlo simulation previously defined in Contribution 10.

The result of the simulation defined in Contribution 9 in a visualization should show all features displayed on both x and y-axis. The resulting plot, shown in Contribution 12, allows for a visual description of the correlation between each feature.

[53] Usually real-world investigation scenarios will rely on a much higher number of features, which would make the use of visual methods impractical. Numerical methods on that case would be more appropriate.



Contribution 12. First Monte Carlo Simulation, Scatter Plot Matrix

The resulting plot of a first Monte Carlo simulation shows each feature arranged in rows and columns in a scatter plot matrix. On a scatter plot, features are arranged so that correlations between each feature can be visually inspected, and in the diagonal a normal interpolation gives a sense of mean and distribution of each feature.

The diagonal of the scatter plot brings a normal interpolation of each of the features, from where we can visually get a sense of mean and distribution for each of the features. The upper right half of the matrix brings a scatter plot of each pair of features, along with a linear interpolation of the pair. This visual arrangement gives

us a sense of mutual correlation. The lower left half of the matrix shows a cluster plot where we can observe any patterns of clustering on each of the pairs.

From top to bottom, and left to right, a scatter plot matrix presents the same sequence of features in horizontal and vertical. In this case, these are in order $EWMA_\alpha$, α, L_i, and W_0 – respectively representing the dampening factor of the EWMA, fitness (or profitability), load (or transaction costs) and the initial price of the instrument under simulation.

This first Monte Carlo simulation and the analysis of the correlation between features in the scatter plot matrix brings specific and essential insights into the BCOM strategy under investigation:

- This strategy is never profitable against random walks using the range of uniformly distributed arguments for this simulation. No shocks were able to bring $\alpha > 1$, our definition of *profitable* as explained in Contribution 8.

- The lower the dampening of the filter, the less money an investor will lose. The relationship between dampening and profitability is shown by the (dampening x fitness) scatterplot on the upper half right side of the scatter matrix, row 1 and column 2, with a negative line-of-fit. In other words, an investor will lose less money using slow filters; or using yet another phrasing, $EWMA_\alpha$ (dampening) is negatively correlated to α (fitness).

- As one could expect, this simulation shows the obvious negative correlation between transaction costs and profitability. The lower the transaction costs (load), the less money you lose. The lower the feature L_i (transaction costs, or load), the higher α (profitability, or fitness).

- As one could intuitively expect, the initial price of a stock does not influence the profitability of this model. The relationship between initial stock price and profitability is shown by the (s0 x fitness) scatterplot on the upper half right side of the scatter matrix, with a virtually neutral line-of-fit. In other words, the feature W_0 (initial price) does not correlate to α (profitability).

We can see that one of the features is irrelevant for what we are investigating. The quick inspection described above showed W_0 (initial price of a stock) does not influence α (profitability or fitness) and should be removed. On that note, we will adjust our model to remove W_0 and add a new feature W_σ, namely the variance of the Brownian random walk, as explained in Equation 13 through the parameter σ on that model.

```
def momentum_simulation_modified(shock):
    return ts('2013-1-1', '2014-12-31') \
    >> brownian(sigma=shock.sigma) \
    >> ewma(alpha=shock.alpha) \
    >> maco \
    >> cash_stock(initial_cash=10000, load=shock.load)
```

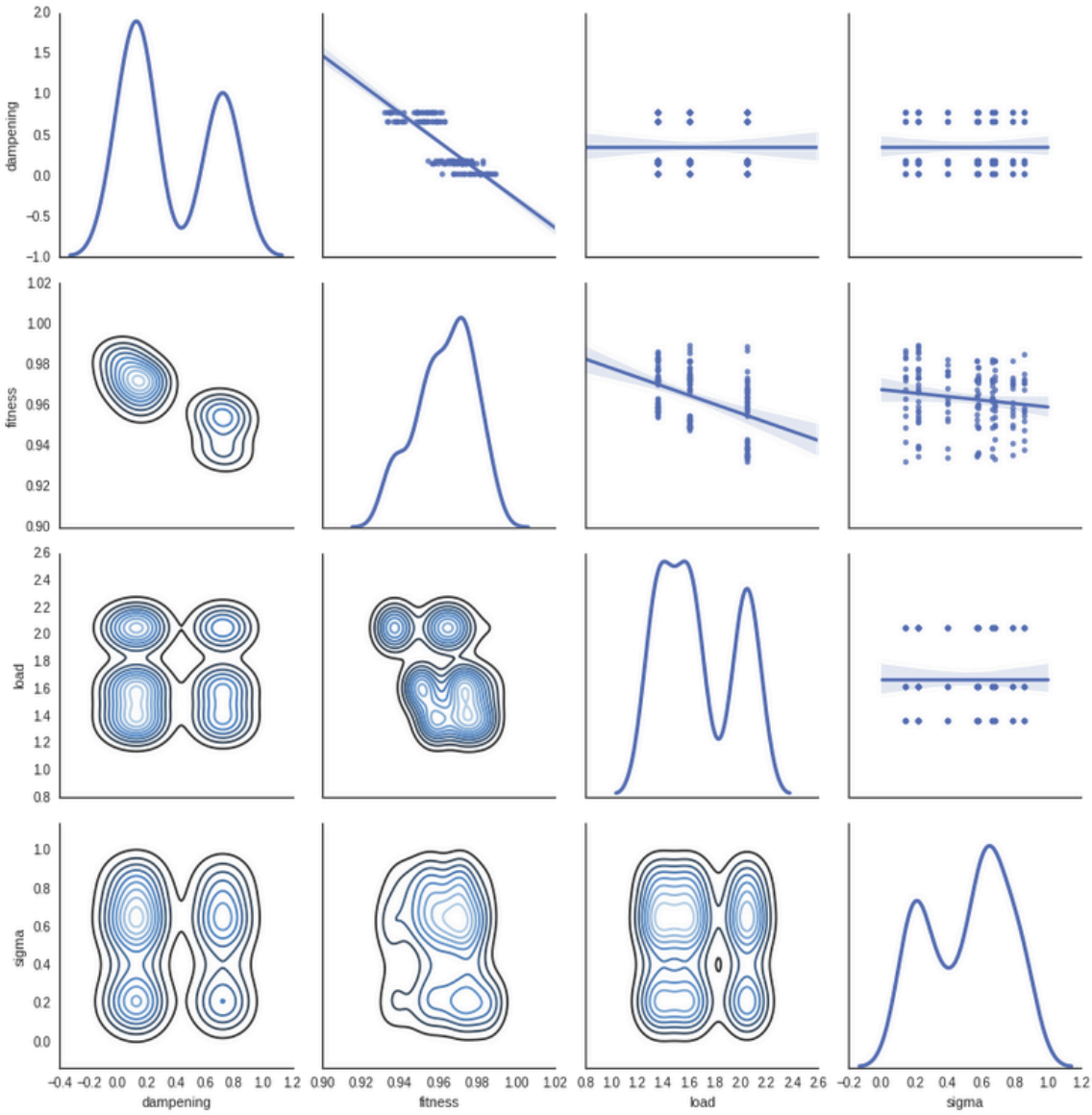
Contribution 13. Modified Simulation Model

This contribution performs a modified simulation of the performance of a BCOM strategy, by permutations of possible values of a different set of features in a financial model. Each permutation is given by a *shock* instance, carrying specific features $shock.sigma$, $shock.alpha$, and $shock.load$. For illustration purposes, compare this modified simulation to the original one, given in Contribution 9.

Our new model, in essence, represents the same as Contribution 9. However, it functionally does something substantially different: this new model investigates if

the variance of a random walk (W_σ) affects profitability (α) and if so under what circumstances.

The new shock features top to bottom and left to right are, in order, $EWMA_\alpha$, α, L_i and W_σ – respectively representing dampening of the EWMA, fitness (or profitability), load (or transaction costs) and variance of the Brownian motion. A scatter plot of this simulation is given in Contribution 14.



Contribution 14. Second Monte Carlo Simulation, Scatter Plot Matrix

The resulting plot of a second Monte Carlo simulation shows each feature into consideration. On this second simulation, we detect a slight negative correlation between the variance of the random walk and profitability of a BCOM strategy.

In Contribution 14 we see that all findings on the first try still hold true for the second try, with an extra insight. Now we get an additional conclusion about the correlation between W_σ and α: the (sigma x fitness) cell on the right side of the scatter matrix shows a scatterplot with a negative line-of-fit. This correlation between

the variance of a random walk and profitability indicates that the variance of a random walk (sigma, or W_σ) is negatively correlated to profitability (fitness, or α) or, in other words, we should expect to lose slightly less money when a random walk presents lower volatility.

The second try also confirms the bottom line of this simulation: against a random walk this model is never profitable.

5.4.3. BACK-TESTING AGAINST THE S&P 500 INDEX

For the second part of our hypotheses stated in Section 5.4.1, we will back-test the BCOM model for profitability against historical price data. For the sake of transparency, we selected to use constituents of the S&P 500 index. The S&P Index, created in 1957, was the first market-cap-weighted stock market and tracks US stocks with at least USD 5.3 billion of market cap (McGraw Hill Financial 2015b).

As we have explained in Section 4.4.3, one type of contributions in FRACTI is called an endpoint. As described in that section, endpoints are further classified into something called a dataset. Datasets are a repository of transformed data fragments, generated at one point in time, and reused, or consumed later.

Historical data is a dataset, and as such, it is also a contribution in our framework. Since historical data is a contribution, it can be leveraged as part of streams and be bound to other contributions, as explained in Section 4.4. As an example, we show in Contribution 15 how the dataset $historical$, representing historical time series of adjusted closing prices of an AAPL stock, can be used in a simple financial model to extract and plot a historical time series.

```
historical('AAPL', '2014-01-01', '2014-12-31', columns=['Adj. Close']) \
>> plot(index='Date')
```

Contribution 15. Model for Historical Adjusted Close Prices for APPL

A simple financial model shows how a dataset can be used as an endpoint to plot historical adjusted closing prices of stock of symbol AAPL for the year of 2014. The resulting stream is sent to another endpoint, a visualization plot.

This financial model represents an operation to get and show price points as a simple stream of two steps: from an endpoint dataset to an endpoint, as a static visualization explained in Section 4.4.3. On the first step, the first line on Contribution 15, the endpoint *historical* produces all adjusted closing prices[54] for Apple Computers for the year 2014. The second step simply directs the stream of prices to a static visualization, shown in Contribution 16.

[54] A stock's closing price on any given day of trading that has been amended to account for distributions and corporate actions that occurred at any time prior to that day's closing (Norton 2010).

Contribution 16. Adjusted Closing Prices for AAPL in 2014

A contribution of a static visualization showing all adjusted closing prices of Apple Computers Inc. for the year of 2014.

The dataset *historical* is a contribution. As we have described in Section 4.4.3, we can leverage historical data on any financial model as we would do with any other contribution. If we want to use historical data, instead of the random walk in our BCOM financial model shown in Contribution 7 in page 172, all we have to do is switch the first and second steps, related to a *ts* endpoint and a *brownian* processor, with one step: *historical*. The slightly modified financial model for historical data is shown in Contribution 17.

```
historical('AAPL', '2014-01-01', '2014-12-31', columns=['Adj. Close']) \
>> ewma \
>> maco \
>> cash_stock(initial_cash=10000, load=7.5) \
>> plot
```

Contribution 17. BCOM Applied Over APPL Historical Adjusted Closed Prices

A financial model used to back-check the profitability of a BCOM strategy on historical prices of Apple Inc. stock AAPL for the year of 2014, for an initial cash balance of $10,000 and a transaction cost (load) of $7.5 per transaction

This financial model in Contribution 17 compared to the original financial model in Contribution 7 in page 172 shows that they are descriptively very similar. This similarity is intentional, and desirable, since they essentially represent the same underlying attempt: to measure the profitability of a BCOM strategy. This slight modification, however, brings a relevant conceptual difference: the financial model in Contribution 7 gauges the performance of a BCOM strategy against a random walk, while the newly introduced financial model in Contribution 17 gauges the performance of that same strategy against an actual symbol, and its historical prices. The result of the static visualization, or plot, of the financial model in Contribution 17 is shown in Contribution 18.

Contribution 18. Historical BCOM Performance of AAPL

The visualization of a financial model representing the performance of the BCOM strategy, in the hypothetical case the strategy was used to trade AAPL stock during the year of 2014. Alpha indicates profitability, green triangles indicate buy signals, and red triangles indicate sell signals over time.

If we compare the BCOM performance over a random walk, shown in Contribution 8 back in page 174, to the newly assessed results of the same strategy applied over historical prices of AAPL, as shown in Contribution 18, it seems like the trend of poor performance repeats itself.

What this new Contribution 18 tells us is that, if one were to start using the BCOM strategy to trade APPL stock at the beginning of 2014, one should expect to lose approximately 3% of the original cash balance. In other words, of the initial $10,000 investment in APPL stocks on 1/Jan/2014, only about $9,700 would remain by 31/Dec/2014.

Since now we know how easy is to change features of our investigation, we can extend our inquiry. Would this poor performance be something specific to Apple Computers? How would a different stock perform given the same strategy? For example, how would Google Inc.[55] (GOOG), behave in the same period using the same BCOM strategy and features?

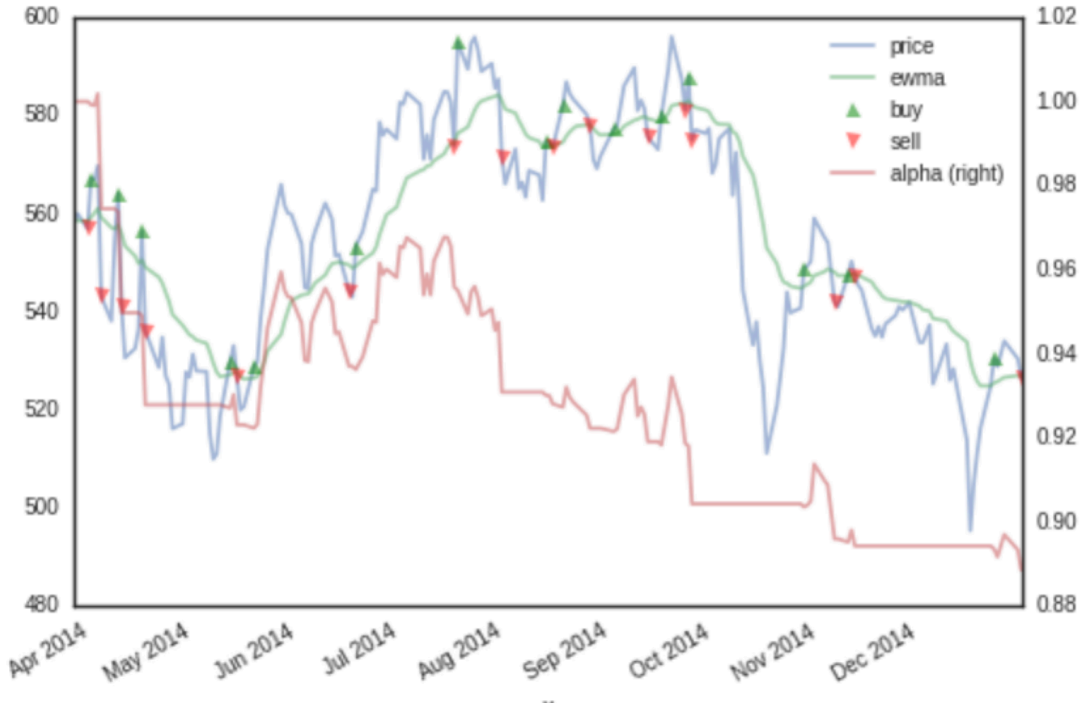
```
historical('GOOG', '2014-01-01', '2014-12-31', columns=['Adj. Close']) \
>> ewma \
>> maco \
>> cash_stock(initial_cash=10000, load=7.5) \
>> plot(out='goog_momentum.png')
```

Contribution 19. BCOM Applied Over GOOG Historical Adjusted Closed Prices

A financial model used to back-check the profitability of a BCOM strategy on historical prices of Google Inc. stock GOOGL for the year of 2014, for an initial cash balance of $10,000 and a transaction cost (load) of $7.5 per transaction

To answer these questions, we create a slightly different financial model, shown in Contribution 19, bringing only one modification: a change of a constant from 'APPL' to 'GOOG'. This small modification is all it is required to visualize the historical performance of this new instrument, shown in Contribution 20.

[55] Throughout this research the name of the company carrying the symbol GOOG changed from Google Inc. to Alphabet Inc. We kept the original designation, counting on the fact that readers would be more familiar with the original name.



Contribution 20. Historical BCOM Performance of GOOG

The visualization of a financial model representing the performance of the BCOM strategy, in the hypothetical case the strategy was used to trade GOOGL stock during the year of 2014. Alpha indicates profitability, green triangles indicate buy signals, and red triangles indicate sell signals over time.

Considering this new security on the simulation, GOOG, we lost even more money: as it is shown by the $alpha$ line Contribution 20, as much as 12% of the original investment would be lost over the same period.

Still, by now we have run the financial model twice, on two separate stocks, and neither was profitable. One could argue that the sample is not representative of expected behavior, both stocks are in the same sector, and both are large-cap stocks. In exact terms, the bad performance we have observed in both exercises, shown in Contribution 17 and Contribution 20, could be related to poor data selection and bias in our part.

To investigate further, looking to falsify these scenarios, we could back-test this model against a more significant, representative sample of stocks, on multiple sectors, and reasonably large caps. For consistency, and to minimize unintended biases, instead of trying to select a sample of stocks ourselves, we picked a set of stocks already present and tracked in a market index.

Therefore, in this next experiment, we will investigate the performance of a BCOM strategy if that strategy were to be applied to all stocks constituents of the S&P 500 index.

As we did on the first exercise, described in Contribution 9, the first step of a simulation exercise is to define the simulation model, specifying all features in a shock. The simulation model for this investigation is given in Contribution 21.

```
def snp500_model(shock):
    stream = historical(shock.symbol, '2014-01-01', '2014-12-31', columns=[shock.column]) \
    >> ewma \
    >> maco \
    >> cash_stock(initial_cash=10000, load=7.5)
```

Contribution 21. Simulation Model for a Generic Stock

This contribution performs a simulation of the performance of a BCOM strategy in the year of 2014, for a specific stock by permutations of possible values on features of the financial model. Each permutation is given by a $shock$ instance, carrying specific features $shock.symbol$, and $shock.column$

Under this dialect of FRACTI, a significant conceptual change in the simulation only requires a slight modification of the original model given in Contribution 9. In this new Contribution 21, we modify the original model to simulate different symbols in an index, instead of different characteristics of a

random walk. This specific new simulation case requires a smaller set of features, especially:

- The symbol of the stock, given by $shock.symbol$;
- The feature, or column, on the historical dataset, given by $shock.column$.

The simulation exercise consists of retrieving all constituents of the S&P 500 index and execute the benchmark of all adjusted close prices ('Adj. Close') for all symbols, using the simulation model given in Contribution 21, as shown in Contribution 22.

```
tickers = index('SP500') >> select(lambda x: x['Ticker'])
benchmark(snp500_model, symbol=tickers, column=['Adj. Close']) \
>> select(lambda x: x.fitness) \
>> hist
```

Contribution 22. Benchmark of All Constituents of S&P 500

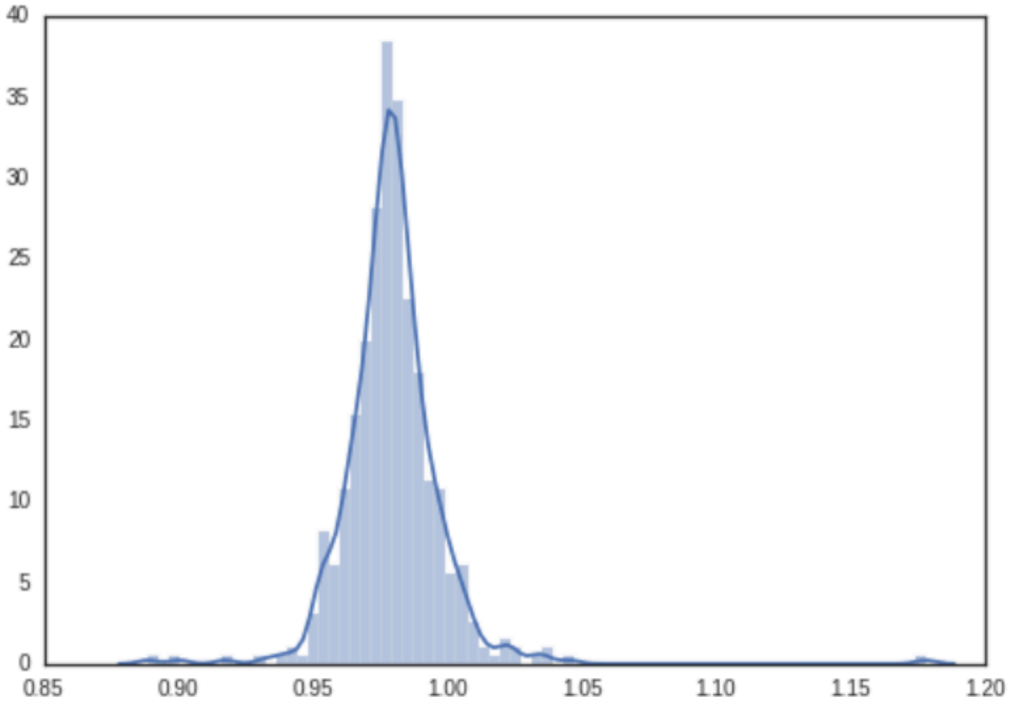
A benchmark of all symbols of the S&P Index, present in dataset $index$, uses the simulation model $snp500\_model$, previously defined in Contribution 21, and a specific column of the dataset associated to adjusted closing prices ('Adj. Close'). The $select$ statement specifies which feature of the simulation the histogram is generated against, $fitness$, or profitability of a specific symbol. The resulting plot will show a distribution of profitability of all stocks on the S&P.

This benchmark of the S&P Index generates the individual profitability of each symbol constituent of the index in two steps:

- The $index$ dataset, labeled 'SP500', contains all constituents of the S&P 500 index. For every entry of that dataset is extracted the symbol ('Ticker') of all constituents of the index;

- The simulation model snp_500, previously defined in Contribution 21, is executed once for each symbol in $tickers$, defined in the previous step. This step-wise sequential execution is called a benchmark. The benchmark executes individual shocks, in which each shock has a symbol and a column, in this case, the column associated with adjusted closing prices ('Adj. Close'). We collect the fitness of the model for this specific symbol and plot a histogram with the distributions of results.

The final result of the benchmark is given by a histogram of the profitability of individual stocks of the index, fitted by a normal approximation, as shown in Contribution 23.



Contribution 23. Distribution of Results, Simulation S&P 500

The histogram shows the distribution of profitability of the BCOM strategy, in the hypothetical scenario where the strategy would be used to trade on every single individual stock of the S&P 500 index during the entire year of 2014.

This resulting Contribution 23 presents the results of the benchmark, showing the profitability of the BCOM strategy in the hypothetical scenario where the strategy would be used to trade on individual stocks of the S&P 500 index for the year of 2014. In this hypothetical scenario, an investor would start with a hypothetical initial cash balance of $10,000 and would incur a cost of $7.50 per transaction, as described on the underlying simulation model in Contribution 21. The results closely fit a normal distribution, and the bottom line of this part of the exercise is clearly demonstrated:

- Evidence collected here indicate that this strategy is not profitable for stocks in the S&P 500 index. The mean stays at around 0.98, or 98% of the distribution.

- There is one single outlier in which we can achieve a return of approximately 18%. The right end tail of the distribution has a single outliner on 1.18, or 118% of fitness, or 18% profitability.

The conclusion, based on the evidence described by these contributions, is that on average, a stock in the S&P 500 index picked randomly and traded with a BCOM strategy would be losing around 2% for the year 2014. This scenario assumes an initial cash balance available for the trading of $10,000 and a transaction cost of $7.50.

Therefore, given the parameters of this exercise, the BCOM strategy for a set of diversified stocks in the S&P 500 is not profitable.

5.4.4. PROVENANCE OF CONTRIBUTIONS

Everything produced and stored in FRACTI is a contribution. By definition, as explained in details in Section 3.4.3, a contribution is a shareable and formal evidence of a crowd-based investigation. To be qualified as evidence in a crowd-based investigation scenario, collaborations have to carry a number of evidential properties such as a record of provenance, as previously described in Section 3.4.3.

To illustrate that, we investigate in this section the record of provenance of Contribution 20, a performance plot generated previously in this exercise, on page 193.

```
provenance('goog_momentum.png')
```

Contribution 24. Extracting the Provence of a Plot Contribution

Generating a record of provenance of a specific contribution generated previously. The original contribution is a plot, generated by $plot(out = \text{goog\_momentum.png})$ given in Contribution 19.

In this Contribution 24 is shown the statement for generation of a record of provenance for a specific contribution named 'goog_momentum.png'. The original plot is generated as a result of Contribution 19, on the endpoint $plot(out = \text{goog\_momentum.png})$. As described in Section 4.4.3, endpoints are contributions, and as such they must carry a record of provenance. A static visualization of the endpoint is shown in Contribution 20.

The statement in Contribution 24 extracts the record of provenance of Contribution 20, short-named on generation as 'goog_momentum.png'. The record of provenance shows all details related to ownership, time stamps, versioning, source,

and transformation steps to get to the generation of the plot. Additionally, the record of provenance should be complete enough to allow the re-creation of any specific contribution, if needed. The record of provenance is shown in Contribution 25.

```
**** PROVENANCE hdf://quantlet/jfaleiro/goog_momentum.png ****

quandl.get('https://www.quandl.com/data/WIKI/GOOG')
`-- v N/A jfaleiro @ 8/15/2015 14:35:12 EST

historical.cache('hdf://quantlet/pub/quandl/WIKI/GOOG.h5')
`-- v 0.0.1 jfaleiro @ 8/15/2015 14:35:12 EST MD5:d9d914b9bdcd9fccd913d4ab77909dbf

stream
    quantlet_analytics.dataset
    historical('GOOG', '2014-01-01', '2014-12-31', columns=['Adj. Close'])
    `--- v 0.0.1 jfaleiro @ 1/1/2015 8:34:02 EST MD5:29bc31efd050df78976a2b85250c2e54
    quantlet.filter
    ewma
    `--- v 0.0.1 jfaleiro @ 1/1/2015 8:34:02 EST MD5:9ed435ab2b00f1afdf4ff79cc74e4424
    quantlet.strats.momentum
    maco
    `--- v 0.0.1 jfaleiro @ 1/1/2015 8:34:02 EST MD5:9ed435ab2b00f1afdf4ff79cc74e4424
    quantlet.strats.portfolio
    cash_stock(initial_cash=10000, load=7.5)
    `--- v 0.0.1 jfaleiro @ 1/1/2015 8:34:02 EST MD5:9ed435ab2b00f1afdf4ff79cc74e4424
    quantlet_analytics.plot
    plot(out='goog_momentum.png')
    `--- v 0.0.1 jfaleiro @ 9/1/2015 8:15:35 EST MD5:ecfc5d3f498074e79cbb0b309dc1b786

file('hdf://quantlet/jfaleiro/goog_momentum.png')
`-- v N/A jfaleiro @ 9/17/2015 21:05:02 EST MD5:72802afc4283d8077da1cb368aa6c2cd

**** end of hdf://quantlet/jfaleiro/goog_momentum.png ****
```

Contribution 25. Record of Provenance of GOOG Plot

The record of provenance of an endpoint associated to a static visualization, short named 'goog_momentum.png'. The record of provenance shows all details related to ownership, time stamps, versioning, source, and transformation steps to get to the generation of the plot. Additionally, the record of provenance should be complete enough to allow the re-creation of any specific contribution

There are a number of important details we can infer from a record of provenance shown in Contribution 25:

- The source of all data and steps followed for generation are public and transparent. Ownership of each contribution (user jfaleiro), the timestamp of each contribution (i.e., date and timestamp at which the contribution was generated), an MD5 checksum, and source (indicated by a URI) are shown explicitly.

- From the record of provenance, we can observe that the data used to create this plot was obtained on-line (Quandl 2015) on a specific date shown on the record, and was passed through a number of stages for caching and transformation, according to a specific stream. The stages themselves are either endpoints or processors, are also contributions, as explained in Section 4.4.2 and Section 4.4.3. The stream is also a contribution as explained in Section 4.4.1. The generation of this version of the plot is also stated on the record of provenance.

- There is a URI associated with any contribution. In the case of this plot contribution, the URI of the version 0 of the plot `goog_momentum.png` is given by `hdf://quantlet/jfaleiro/goog_momentum.png:0` and indicates a universal location of the contribution and version. This URI allows this contribution to be shared with any collaborator with knowledge of this URI, as long as the recipient carry proper credentials.

- Contributions are created, read, updated, and deleted by respective operations referencing the URI of the contribution. A creation is always associated to version 0 of a contribution, and further operations of update increase the

version number. A removal is not definitive. An update on a removed contribution brings the contribution to a created state.

- All contributions are signed, safe-stored, and check-summed.

Although not the only evidential property of contributions, described in Section 3.4.3, provenance tracking is one of the fundamental features of FRACTI that allows for contributions to be treated as shareable evidence in a crowd-based investigation. Support for traceable sharing and collaboration among heterogeneous parties is one of the core features by which we achieve reproducibility in large-scale scientific research (Faleiro Jr and Tsang 2016a).

5.5. FINAL NOTES ON EVIDENCE OF PROFITABILITY

There are two opposing views regarding the efficiency of technical analysis similar to BCOM strategies. In one end are a number of skeptical studies, derisively equating technical strategies to "tea leaves reading" (Browning 2006), "black magic" (Samuelson 1965), or "financial astrology" (Huebscher 2009) (Carolan 1998). In the other opposing end of the spectrum are several claims of consistent return in the range of double-digit percentage points year after year. (Park and Irwin 2004)

Mimicking those same opposing views, but avoiding extreme stances, peer-reviewed scientific publications are apparently divided between supporting and repudiating claims of efficiencies of technical strategies (Park and Irwin 2007) (Balsara, Chen and Zheng 2007) (Hoffmann and Shefrin 2014). A survey of past studies indicated that from "95 modern studies, 56 concluded that technical analysis had positive results", and pointed to the difficulty in getting to conclusive findings due to "data-snooping bias and other problems" (Park and Irwin 2007) and "noise in trading price" (Black 1986).

Evidence from this investigation exercise corroborates these previous findings. The first simulation used in this chapter observed that entirely random walks were consistently unprofitable, retaining 96% of the initial cash investment in a year. On the second simulation, using real historical data shows that this model was only profitable in about 8% of stocks constituents of the S&P 500 Index, and in average retaining 98% of the initial cash investment during 2014.

We were not able to achieve claimed results indicating a consistently profitable behavior using either random walks or real historical data. Evidence produced in this exercise suggest that the strategy is not consistently profitable. As a consequence, the hypothesis we had outlined at the beginning of this exercise, in Section 5.4.1, is false.

These findings show strong discrepancies from claims from several investment resources. Given the lack of reliable evidence from previous studies the specific causality is difficult to assess, but taking into consideration our evidence, and indications on other studies, we can list a number of possible explanations:

- Data snooping[56] (Young and Karr 2011) and survivorship biases[57] (Shermer 2014). Evidence in this experiment might be getting the same results as previous studies did, in which technical "rules are profitable when considered in isolation, but these profits are not statistically significant after adjustment for data snooping and survivorship bias" (Marshall, Cahan and Cahan 2010).

- Our data samples are too efficient. The very notion of a pure random walk contradicts the assumptions followed by chartists that price movement carries

[56] Data snooping, also known as data fishing, data dredging, equation fitting or p-hacking, is the intentional or unintentional use of data inference techniques the researcher decides to perform *after* looking at the data (University of Texas 2011) (Simth and Ebrahim 2002)

[57] Survivorship bias is the unintentional error of concentrating on data items that have "survived" some process and overlooking those that have perished (Schemer 2014)

some “past memory” or private information (Fama 1965). In that sense, random walks and components of the S&P 500 Index might be just too efficient for momentum strategies to perform. “There is some evidence that technical trading rules perform better in emerging markets” due to their inefficiencies (Marshall, Cahan and Cahan 2010), and technical strategies tend to perform better in inefficient markets (Chaudhuri and Wu 2003).

- Traders might be using variations of this momentum strategy that are actually profitable, and not disclosing its details. Reproducibility, in this case, is obviously not possible.

- Either data or algorithms we relied on for these calculations are wrong or have bugs. Despite care, multiple reviews and regressions over this model, inaccuracies of this kind are unfortunately commonplace in scientific research (Bisig, et al. 2012) (NPR 2013) (Reinhart and Rogoff 2010) (Olsen and Cookson 2009) (Lehrer 2011) (Tsang 2014). In this case, all contributions in this exercise are available, traceable and verifiable by any interested parties if needed. This transparency of evidence that can be shared and investigated by crowds is indeed the primary motivation behind FRACTI (Faleiro Jr and Tsang 2016a).

To derive causality from computational artifacts that seem correlated at first sight is a hard task, especially when scenarios are not exhaustive. As we have discussed in previous publications (Faleiro Jr and Tsang 2016a), we should expect this determination to become increasingly more difficult as we have to deal with higher volumes of data, and more complex representations. As previously discussed in Section 3.3.3, evidence in the literature associates this phenomenon as a

consequence of the "informatics crisis" (Goecks, Nekrutenko and Taylor 2010) and the noise present in scientific investigation (Faleiro Jr and Tsang 2016a).

In closing, as we stated at the beginning of this exercise, the determination of the exact cause for a phenomenon in our investigation is secondary. The primary objective of this chapter is to demonstrate how simple it is to adjust and modify financial models and simulations in order to allow following fluid ideas, and how the FRACTI conceptual framework enforces collaboration through shared evidence (FRACTI contributions). At this point, we cannot determine the cause of the discrepancy between research claiming for or against the profitability of technical analysis in financial trading, and the answer will remain debatable and possibly the subject of future research. Our core argument on this overall research is that for increasingly complex cases of use, collaborative crowd-based scientific research is becoming the only way to achieve unequivocal answers.

5.6. CHAPTER SYNOPSIS

This chapter fulfills the criteria for success described in Objective 3, on page 21 of this thesis, the definition of an end-to-end investigation exercise to measure the actual efficiency of technical analysis using formal methods and historical trading.

The core argument of this research is that finance should be studied like any hard science, strictly following the procedures of the modern scientific method, leveraging computational controls and crowds, as previously explained in Chapter 3 and Chapter 4.

In this chapter, we have demonstrated how a trading strategy commonly used in technical analysis could be studied scientifically, for example, with control experiments (Brownian motion), generalization (testing on multiple assets) and

statistical analysis. By specifying the experiments under the proposed FRACTI framework, described in Chapter 4, researchers can communicate without ambiguity what experiments they have conducted. Other researchers can repeat the experiments to verify the results. This way, FRACTI can support crowd science, which is an efficient way to accelerate research in complex subjects of knowledge like finance.

We presented a concrete case in which we use FRACTI facets and contributions to outline a hypothesis, produce, record and collect evidence and get to an objective conclusion about a financial phenomenon under research, and as a consequence prove or disprove it.

In this exercise, we intentionally selected a model that is simple enough for a broad community of finance users to understand and test. An additional incentive is the exposure of this and similar strategies that are of constant public debate, in the media and academia, of opposing views of technical, fundamental and quantitative approaches to investing. The investigation exercise is assembled based on the steps and premises previously defined in Section 3.2.1.

A reasonably sophisticated trading strategy, called a *BCOM strategy*, is defined and formalized in Section 5.2 based on its components: *random walks*, *signal attenuation*, *derivation of market signals*, and portfolio management. Despite its complexity, this strategy can be described by FRACTI models in clear and accessible text as shown in Section 5.3, for example, Contribution 7 and Contribution 9. As described in a previous chapter, in Section 3.4, and demonstrated here, by using other researchers' contributions, a researcher does not have to be a computer specialist to understand, communicate and improve new contributions.

The hypotheses of this investigation exercise are defined in Section 5.4.1. The investigation exercise, described in Section 5.4, defines two steps: a Monte Carlo

simulation of Brownian variations in Section 5.4.2, and backtesting against stocks constituents of the S&P Index in Section 5.4.3.

Every single one of the shareable evidence generated in this exercise is called a *contribution*, defined as a *shareable and formal evidence of an objective crowd-based investigation*, previously described in Section 3.4.3. As explained in that session, contributions inherently carry a number of *traits* and *evidential properties*, such as ownership, provenance and access restrictions. These properties are produced and maintained transparently at the same time the contribution is generated, and as explained in Section 3.4.3, are crucial for scientific process of crowd-based investigation. An example of evidential properties of a contribution is given in Section 5.4.4.

Conclusions of the investigation exercise are given in Section 5.5. The first simulation in Section 5.4.2 observed that entirely random walks were consistently unprofitable, retaining 96% of the initial cash investment in a year. On the second simulation in Section 5.4.3, using real historical data shows that this model was only profitable in about 8% of stocks constituents of the S&P 500 Index, and in average retaining 98% of the initial cash investment during 2014.

This chapter brings a number of novelty contributions to this research, specifically:

- The study, formalization, and investigation of profitability of a BCOM strategy regarding Monte Carlo simulation of a random walk of prices, and historical prices of constituents of an S&P Index, respectively in Section 5.4.2 and Section 5.4.3;

- Representation of an end-to-end investigation of a non-trivial problem in economics using FRACTI concepts for crowd-based investigation, in Section 5.3;

- Representation of evidential properties of contributions, in Section 5.4.4.

Code excerpts provided in this chapter are examples of one possible dialect called QuantLET (J. M. Faleiro Jr 2008) are provided for illustration purposes only and are not the core subject under research. Despite that, these features align themselves with the vision of the long-term research and could be looked at as an illustration of the overall roadmap for FRACTI, as described in Section 4.7.

CHAPTER 6. CONCLUSION

"There is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole of life, from the moment you are born to the moment you die, is a process of learning." (Krishnamurti 2014)

6.1. SUMMARY

Exploration exercises usually start by trying to answer essential questions coming from the observation of a phenomenon. The exploration entailing this research is no exception. Why is it so difficult to predict outcomes in financial sciences, even when we know the social cost of an error is so high? Why are scientific methods in economics conducted so differently from investigations performed in other sciences? What are the core differences between the discipline of economics and other disciplines keener to hard sciences, where predictions and answers are trusted, often dealing with more complex subjects?

During the regular process of this research we had a chance to present our ideas to knowledgeable, specialized audiences in the field of economics, and use that opportunity to slip in these same inquiries. The most common answer we received expresses a belief that economics is a unique domain of knowledge, in which its principal subjects of study – humans – are difficult, if not impossible, to model. This answer seems to imply that human behavior is fundamentally non-deterministic and therefore adequate modeling is impossible. The absence of proper models makes prediction and measurement impractical; hence economics has to be handled like a soft science, similar to psychology, political sciences or sociology, and not like engineering, physics or astrophysics.

This research approached this answer as incomplete by cheer observation. If we can safely fly in vehicles made of metal, perform unmanned pinpoint landings in dashing meteorites, and quickly continue to solve the mysteries of life hiding deep into our chromosomes, why can't we achieve similar results from our explorations in economics?

Throughout this long research developed over the last several years, this study grew both in breadth and scope responding to new questions, the introduction of new ideas, additional fields of research, and contributions. From the beginning we assumed that the complexity or unpredictability of human behavior should not be used a justification for the lack of scientific methods in economics. The answer advocating a soft science approach for economics was assumed to be incomplete, and weak. We assumed the need for something else to enable structured investigation in economics, even if in the beginning we were not exactly sure about what this "something" would be. We knew, however, and literature seemed to confirm, that the answer had to be centered in human characteristics, and geared towards modeling of complex and dynamic, ever-changing, systems.

We began by concentrating solely on the computational aspect of an answer, maybe following our experience, or intuition. Despite its richness, and without a doubt required in the overall solution, the technological aspect showed itself limiting. The more we looked into a solely computational answer, and the more we surveyed existing literature, the research led to the realization that computers alone are more part of the problem than the solution. Computational power without proper control should be considered problematic. As it is the case with any advanced technology, computer power amplifies errors, risk, and unfitness of models of prediction to possibly disastrous consequences. We borrow the existing term "informatics crisis", coined on a somewhat related context, to describe the computer power paradox.

6.2. CONTRIBUTIONS

Contributions of this work are anticipated by the objectives defined in Section 1.2, and organized in specific research chapters of this thesis, described in Chapter 3, Chapter 4, and Chapter 5, and summarized respectively over the next paragraphs.

- **Enablers for Crowd-Based Investigation**: the computer power paradox brings indeed the first contribution of this research: *the definition of cognitive and non-cognitive enablers for crowd-based scientific investigation*, defined in Chapter 3. This first contribution fulfills the criteria of success described in Objective 1, on page 19 of this thesis, by identifying requirements for the adequate use of crowds in structured, scientific investigation. As far as this research could find, despite few references in which individual facilitators are listed, there is no other similar work outlining what is required for efficient, scientific, crowd-based investigation. A detailed bullet list of specific novelty contributions is given on the chapter synopsis, in Section 3.5, on page 93.

- **Specialized Computational Representation for Economics**: this initial insight allowed us to elaborate on one crucial aspect, the importance of a proper computational representation for structured scientific investigation in general, and specifically for economics. In Chapter 4 we describe the second contribution of this work: *a specialized computational representation for the field of economics*, given by a conceptual framework called FRACTI. This second contribution fulfills the criteria of success described on Objective 2 on page 20 of this thesis by defining a computational representation to support investigation and collaboration in large-scale for the field of economics. This conceptual framework is not a computer language or an implementation, but a representation system based on a set of fundamental building blocks to assemble and describe financial models, at a conceptual level. As far as this

research could find, there are no similar approach or idea for representation of financial models. A detailed bullet list of specific novelty contributions is given on the chapter synopsis, in Section 4.8, on page 148.

- **Non-Trivial Investigation Exercise**: the third and final contribution of this research is given on Chapter 5 by answering an age-old question in financial sciences: *how profitable is technical analysis?* We make sure we follow a strictly scientific approach to answering that question, by using the contributions we previously advocated on previous chapters: enablers of crowd-based investigation given in Chapter 3, and a computational representation for the field of economics given in Chapter 4. This third and last contribution fulfills the criteria for success described in Objective 3, on page 21 of this thesis, the definition of an end-to-end investigation exercise to measure the actual efficiency of technical analysis using formal methods and historical trading. This exercise is a practical end-to-end example of the use of concepts on this research based on reverse engineering, formalization, and data analysis to answer this traditional question in a way that no previous research has done, using a structured approach for crowd collaboration and structured, computational representation. A detailed bullet list of specific novelty contributions is given on the chapter synopsis, in Section 5.6, on page 204.

6.3. ASSUMPTIONS

All scientific research is axiomatic to a certain extent. Research deemed scientific has to build on top of pre-existing knowledge, considered certain and originating in previous works of science, and from there, consider particular

assumptions to build up further knowledge. While some assumptions might eventually be testable and falsifiable, others are considered approximations.

As much as would like to avoid it, this research was no exception. We considered a number of specific assumptions for the definition of the ideas in this research, listed in the following paragraphs.

- The intrinsic behavior of economic agents can be modeled, and the quest for such models is worth pursuing. The science of economics should be seen and treated as a hard science like physics, mathematics or engineering, and approached with objectivity, with similar methods. The argument of inherent complexity or unpredictability of economic systems as a justification for treating economics as a soft science – like psychology, social sciences, or political sciences – is assumed as incomplete and is refuted for the specific purposes of this research. The rationale for this assumption is described in Section 1.1 and Section 6.1.

- The ever-increasing reliance on high-powered computing resources for complex investigations in economics makes the subjects of economics and computational finance more and more intertwined. Given the nature and the scope of this research, there is no practical distinction between these two fields, as discussed in the peculiarities of our field of study in Section 3.2.2, specifically on page 55.

- The process by which we acquire objective knowledge must follow the rules dictated by the scientific method, as discussed in Section 3.2. The proof of observations as being real or false must be driven by a widely known and accepted collection of pragmatic and quantifiable standards, as described in Section 3.2.1.

- Human collaboration in large scale is assumed to be an adequate method to investigate and resolve complex problems. As explained in Section 3.3.2, this assumption is inferred empirically.

- Collaboration in large-scale is enabled by providing the correct set of incentives to crowd participants, as explained in Section 3.3.1.

- Computers should fulfill the role of a tool to support discovery and should not serve as a replacement for the application of reproducible and falsifiable procedures of the scientific method. In other words, computers should serve as control points for collaboration and interaction of human participants, and not as agents of scientific inquiry themselves, as explained in Section 3.1.

- A computational representation is defined by and tightly coupled to, a specific domain of knowledge. The representational process defined in Section 3.4.1 explains the concept of a computational representation and the inter-dependency between representation and a domain of knowledge.

- The exact definition of what constitutes a facet in a specific domain of knowledge is empirical, as explained in Section 3.4.2. In some cases, like architectural sciences, the proximity to visual and spatial concepts makes the establishment of what is indeed a facet - shapes, color, and measurements - somewhat intuitive, and as a consequence easier to derive. The same Section 3.4.2 provides the rationale for a facet.

- This research assumes a role-based definition of knowledge representation, as explained in Section 3.4.5. In a role-based definition, a description of a knowledge system is established in terms of five core roles a specific representation plays (Davis, Shrobe and Szolovits 1993). The explanation of a

role-based representation and significant consequences of this assumption are discussed in Section 3.4.5.

- Different models of computation of streams (i.e., topology, determinism, and dynamicity) are fully translated by variations of three specific properties of financial models, i.e., synchronicity, connectivity, and plasticity. Models of computation of streams are described in Section 4.3.1.1. Basic properties of financial models and the translation to models of computation of streams are described in Section 4.3.1.2.

- A financial model must carry three basic properties to fulfill the requirements for the domain of knowledge of economics, as defined in Section 4.2. The three basic properties are synchronicity, connectivity, and plasticity, all explained in Section 4.3.1.2. These properties are crucial for the representation of financial models as streams, explained on that same section.

- It would only make sense to entertain the investigation exercise in Chapter 5 with an assumption of a presence of price momentum on breakthrough strategies. The assumption of momentum and consequences are explained in Section 5.2.

More details about the context in which these assumptions were made, restrictions, and consequences, can be found on specific sections and references listed in each of bullets above.

6.4. LIMITATIONS

In this section we provide a few reservations and pre-empt a common questions and observations we received from reviewers of our peer-reviewed publications, lectures, readers, and followers of this research.

- The computational representation presented in Chapter 4 is not a programming language or a software platform. The computational representation we call FRACTI is a conceptual framework that relies on facets, contributions, and constraints of data, as explained in Section 3.4, to describe and assemble financial models. These financial models are abstract descriptions that be exchanged by participants in a crowd that are not necessarily specialized computer engineers.

- As explained in Section 3.4.5, a computational representation is a role-based knowledge system. As a consequence, comparisons of fitness of our computational representation to any other representations, in the most generic sense, or even to target ideas, and neither relevant nor appropriate. We assume that financial models on our computational representation are *surrogates* (Davis, Shrobe and Szolovits 1993), in a sense that these models are by definition a substitute for the target idea itself. As a consequence, any measurement of how far or how close this surrogate is from calculations it intends to represent is secondary or irrelevant.

- As we have described in Chapter 3, we use the term *enablers* to refer to requirements that enable, but do not guarantee, a crowd-based scientific investigation to occur in a given environment. In other terms, using implicational relationships enablers are a *necessary but not sufficient condition* for the proper support of a crowd-based investigation.

- The qualification given in Chapter 3 of either cognitive or non-cognitive refers to enablers, or requirements, which are purely computational or not. A non-cognitive enabler is purely computational, and a computational representation is by definition a non-cognitive enabler. On the other hand,

cognitive enablers are non-computational by nature, and as the name implies, related to cognition. Cognition, "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" (Oxford English Dictionary 2011), applies precisely to the context in which the two non-cognitive enablers of methods of proof and large-scale collaboration are used.

- Reproducibility of procedures is limited by current data and evidence, and cannot account for upcoming data and technology unknown at the time of this writing. Limitations of the scope of the general term *reproducibility* are given on page 59.

The use of structured computational methods to power crowd investigation is a relatively new subject and this research does not intend to exhaust all the possibilities on this upcoming field. The complexity of relevant problems and the damaging social impact of investigative mistakes in economics are only increasing over time and require more and more the participation of the right specialists at the right time.

We expect this research to serve as a foundation over which new concepts can be built, and to allow the transparent and objective collaboration of a multitude of diverse specialists, in different levels, in the search for right solutions. In the end, we are all companions and collaborators on the quest for the "scientific truth" (Ellerton 2012) that will serve as a positive agent of change in our communities, in local and global levels.

6.5. SIMILAR WORK

This thesis relates to the use of technology for improvement of methods of investigation. This topic covers a broad variety of related academic work, defined by two ends of a spectrum, as shown in Figure 20.

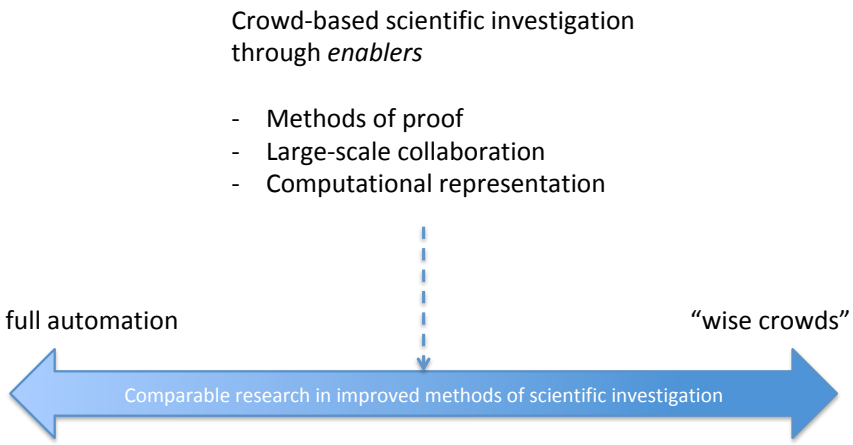


Figure 20. Comparable Research in Methods of Scientific Investigation

The spectrum defining the topic of improved methods of scientific investigation covers a broad range. On the right end of the spectrum are methods that rely on the full automation of methods of investigation. On the opposite end, methods relying on "wise crowds". This research, proposing the use of specific enablers of a crowd-based investigation, sits somewhere in-between the two opposing ends.

One end of the spectrum is defined by research that intends the full automation of the process of investigation. This approach relates to the assignment of computers to execute tasks that are usually performed by scientists. Some of those

tasks are associated to the registration of observations (Alkhateeb 2017), automated hypothesis generation (Spangler, et al. 2014) and contextual gaps (Swanson 1986), and automated testing based on robotics (Soldatova, et al. 2016). Fully automated research is not feasible given current technology, as previously discussed in the definition of the proof pipeline, in Section 3.2.1. Research on fully automated methods of investigation usually brings the same consequences and criticisms associated to data-driven research (Shih and Chai 2016), in which correlations of data are detected first, and only then, hypotheses are produced.

On the other end of the spectrum are solutions that rely on the existence of "wise-crowds", based on empirical evidence collected over the years (Surowiecki 2004) (K. Wallis 2014) (Galton 1907). The wise-crowds approach assumes the existence of some invisible, unquantifiable mechanism that makes crowds wise, and relies on the assumption of complete independence and decentralization. Paradoxically, the assumption of independence would diminish the value of structured collaboration in crowd investigation. Evidence on the existence of some mechanism enabling wise-crowds to occur is often empirical (Kelley and Tetlock 2013) and the subject some of criticism (Mannes 2009).

This research sits somewhere in the middle of this spectrum. We advocate the use of a crowd-based investigation through enablers, methods of proof, large-scale collaboration, and a computational representation. This research emphasizes the use of computers for mechanical and repetitive tasks, like the orchestration of scientific interaction in crowds and the record of scientific evidence as contributions, as previously described in Section 3.4.3. Additionally, this research also advocates for the importance of intangible human factors related to experience and creative thinking in science, and a hypothesis-driven process of discovery, as described in Section 3.2.1.

6.6. FUTURE WORK

This research offers a solution for large-scale crowd collaboration in finance investigation to potentially circumvent the inherent problems related to the information crisis we currently observe in scientific investigation, as described previously in Section 3.3.2. The foundations provided in this research are a significant improvement over the current methods of investigation in finance, as explained in Section 3.3.3, and opportunities for further research on the same subject are far from exhausted.

The subject of modern methods of investigation relies mostly on mechanisms of extreme complexity, as explained in Section 3.2, and an approach to find for a resolution of a complex problem is to search a prospective solution through organic and incremental iterations (Visser 2006) (T. Brown 2008) (Dorst 2015). This incremental, iterative approach is based on the classical principle of "design thinking" (Archer 1965). The main concerns of iterations of future research would be related to (a) a working minimum collaborative environment for crowd-based investigation; (b) versioning of complex run-time graphs; and (c) quantification of features of collaboration.

The initial step in an iterative approach would be the definition of a minimum environment in which the original ideas on this research could be exposed to an initially restricted group of participants and tested. A platform to support a controlled interaction, even if for a small and controlled number of research participants, will help explore algorithms and methods to record contributions and their evidential properties, as explained in Section 3.4.3. The use of evidential properties would require, amongst other things, the support of a record of provenance, as shown in Contribution 25, on page 199. The record of provenance would require storing and

versioning of the representation of the financial model as a graph associated with the stream of the execution, as explained in Section 4.3.1.2.

Versioning of complex graphs by a simple full copy of the entire graph on change of any attributes of the graph will trigger a space explosion. Some alternatives have been recently proposed for a particular type of graph called property graphs (Vijitbenjaronk, et al. 2017). Property graphs are oversimplified, compared to a FRACTI graph that intends to represent a flow of execution that must be materialized on different execution spaces, and as a consequence can be materialized on distinct, distributed processors, as explained in Section 4.3.3.2. An alternative to in-place versioning would be to apply changes to graphs by separate clone and merge operations, in a distributed approach (Torvalds and Hamano 2010). The drawback of a clone-and-merge approach is that, since processors and endpoints are in nature pervasive[58] as described in Section 4.4.2, a clone of a single graph can force cloning of others. The feasibility of a clone and a subsequent merge operation is limited by how large the affected graph might be.

For example, if both financial models ϕ_1 and ϕ_2 use a processor P, then a clone of ϕ_1 will force clone of ϕ_2, and the same for the merge operation. Due to other processors that can be possibly involved, the cloning chain can potentially involve other graphs. As a consequence, for real-world investigations, the final full size of the final chain can become an impediment for clone and merge operations.

Even if limited, in case not all of these questions are finally and adequately addressed, a minimum platform allowing a rudimentary interaction amongst participants of a scientific investigation is still useful for the investigation and additional financial models. The set of cases of use, defined in Section 4.2, was

[58] Given the intent of reusability, explained in Section 4.4.2, the same processors and endpoints should potentially repeat themselves in a large number of financial models.

enough for the definition of a functional computational representational for the field of economics introduced in Chapter 4, but in no way, they should be considered complete. Additional individual financial models should not be enough for a doctorate thesis in isolation but should bring two direct consequences. First, since a minimum platform would allow for reproducibility and traceability of evidence through the use of contributions, it should serve as an incentive for research to be conducted on the platform. Second, new financial models, or cases of use, would show the need for additional facets and possibly even contributions, extending the current computational representation.

The use of crowds for resolution of problems follows one of two distinct approaches. The first approach, named "wise crowds" (Surowiecki 2004) relies on empirical observations (K. Wallis 2014) (Galton 1907) and assumes the existence of some invisible, unquantifiable mechanism, somehow providing a certain level of knowledge to crowds, therefore allowing them to make wise decisions. The "wise crowd" approach relies on the assumption of complete independence and decentralization between participants of a crowd. The second approach, named *collaborative crowds*, assumes that knowledge is produced as a result of structured collaboration between participants of a crowd.

This research subscribes to the second approach, collaborative crowds, where collaboration in large scale occurs by the existence of particular requirements of collaboration, listed previously in Section 3.3.1. It is necessary to produce metrics and quantify the requirements that must be in place for large-scale collaboration, but proper quantification can only happen in a real collaboration environment, and proper participants are engaged. The quantification of requirements for large-scale collaboration, what we call *collaboration metrics*, would allow to measure the potential efficacy of a disjoint group of researchers and compare the performance of

different investigation exercises on items that are specifically related to how well collaboration takes place.

Quantifiable requirements would affect one or more *features of collaboration*. Features of collaboration would give, as a group, indications on the efficiency of the scientific investigation performed by the crowd, for a given specific investigation. Not all requirements for collaboration, previously listed previously in Section 3.3.1, are subject to quantification. Proposed quantifiable requirements for future research are the level of micro-expertise attention and cognitive diversity.

The first quantifiable requirement, the level of micro-expertise attention, is crucial to collaboration, and as a consequence, also to collaborative crowds[59]. The level of micro-expertise attention would measure specific features of collaboration: the relevancy of a participant in a crowd, per investigation subject; the quality of contributions produced per participant; and the influence of a participant.

Consider for illustration purposes a directed graph $\phi = (C, E)$ of vertices C and edges E. In $\phi = (C, E)$, C is a set of contributions $(c_0, \ldots, c_n)$, and E is a set of edges $(e_0, \ldots, e_n)$, indicating dependencies between contributions. Since contributions are produced as a result of a financial model[60], a contribution c' is considered a dependency of c'' if c'' is upstream to c' (in the sense that a financial model is a stream). A specific financial model might be tagged as belonging to zero or more subjects of investigation[61]. Given the streaming nature of a financial model,

[59] As already emphasized in Section 3.3.1, "expert attention is to creative problem solving what water is to life: it's the fundamental scarce resource" (Nielsen 2012)

[60] A financial model is also a directed graph, as previously explained in Section 4.3.1.2

[61] For example, in the exercise for investigation of performance of momentum strategies, previously described in Chapter 5, the financial model in Contribution 18 for example, could be tagged with "momentum", "trading strategy", "APPL", or "stocks"

this determination is straightforward. Features of collaboration are taken from the application of specific graph algorithms to the graph $\phi = (C, E)$:

- The in-degree centrality[62] of the contribution c_i would quantify the quality of that specific contribution;

- The relevancy of a participant in a crowd, per subject of investigation, is given by the summation, of the quality of all contributions C created by that participant;

- The level of influence of a participant is the summation of the relevancy of that participant, across all subjects of investigation.

The second quantifiable requirement is cognitive diversity. The literature points to metrics for cognitive diversity based on cognitive distance (Castner 2014) (Griffiths and Joshua 2009) (Tanenbaum and Griffiths 2000). Future research should also assume a possible reliance on the interaction of participants, or "texts of utterances of a collective's member" (Castner 2014), for measurement of a crowd's cognitive diversity. Another possibility, given recent literature just published, would be the use of a more sophisticated model of sentiment and emotion analysis using unstructured data (Rout, et al. 2018). This research could not find in the literature a proposal for quantification of cognitive diversity for scientific investigation.

This list of future research opportunities is obviously not final. The availability of a platform for crowd-based investigation would allow the collection of evidence to provide new insights into metrics and algorithms for additional items of research.

[62] The number of inbound edges to a vertex in a directed graph

6.7. FINAL NOTES

Finally, I would like to add a few personal notes related to the fascinating endeavor of this research. This research extended over several years, consuming thousands of hours of hard work, the reading of hundreds of books and articles of all kinds, and what seemed like an infinite number of paragraphs written on notes, peer-reviewed publications, and finally, in this thesis.

This research is primarily about one approach in many in how to improve the way we currently search for objective knowledge. Knowledge, as it seems, is the result of a recipe by which we mix the contents of a number of different buckets into where we keep distinct perceptions of the world around us: experiences, creative thoughts, hunches, beliefs, biases, and the systematic training. All in some measure is important, and they all have a role in producing objective knowledge. However, the exact recipe, or mechanisms, we follow in mixing the contents of those buckets hides deep in the way our collective minds work and is still an incognita. For that mysterious recipe, we advocate for the use of crowds to perform scientific investigation in complex subjects, specifically in our case, in the field of economics.

The use of crowds in scientific investigations is indeed a complex subject in itself, involving a multitude of subjects outside of our field of concern: philosophy, computer sciences, psychology, sociology, data sciences, and biomedicine. The many powerful ideas and insights from the ones that came before me are humbling, and make crystal clear the meaning of the term "standing on the shoulders of giants".

We are lucky to be alive these days, when the full power of knowledge of millennia, as well as just yesterday's, is available at the click of a mouse. Scientists in all these subjects are in one way or another looking at the same type of problems, dealing with their consequences, and coming up with their particular answers.

The evidence of breakage in traditional scientific research is plenty, and everywhere we look. The cause, one might argue, is the amplifying effects of technology combined with misaligned academic incentives currently at play in institutionalized science. The solution, we propose, is a crowd-based scientific investigation based on a combination of the adequate methods of proof, collaboration in large scale, and a domain-specific computational representation. These are the enablers we propose, and the ideal combination to harness the power of crowds using computers in a controlled, structured way.

Throughout this research, in addition to the use of conventional and formal methods of peer-review publications and interactions, we did as we say and also leveraged anonymous crowds in our investigation. Several of the topics were discussed in online question-and-answer sites. Platforms and financial models were open-sourced to the public for use and feedback. The involvement of crowds in this research should not stop as this thesis ends. Many of the ideas on this research can and should be extended.

> *"You want to explore in depth all the points that you have omitted, you want to chase all the tangential ideas that struck you but that you eliminated for brevity, you want to read other books, and you want to write essays. This is the sign that the thesis has activated your intellectual metabolism, and that it has been a positive experience. It is the sign that you are the victim of a compulsion to research, somewhat like Charlie Chaplin's character in Modern Times, a factory worker who keeps tightening screws even after a long day of work. Like Chaplin, you will have to make an effort to restrain yourself." (Eco 1977, 252)*

Indeed, as I write these final lines, I keep seeing screws that must be tightened everywhere. This "intellectual compulsion" is fascinating, but like everything else, this thesis has a scope, and must now come to an end (the thesis, not the compulsion).

I truly hope this tiny little speckle of several pages of organized thoughts will challenge similar curious minds into new inquiries and discoveries. These thoughts might ultimately help others to build on new ideas and solutions for the complex problems affecting humanity.

We have plenty to worry about during these challenging times, and thankfully, the ingenuity of the human scientific mind is standing by, ready to tackle and fix it all.

GLOSSARY

Ad hoc	Designates a non-generalizable solution or idea that is not intended to be adapted to other purposes. Applicable to the particular end or case at hand without consideration of wider application (Merriam-Webster 2018)
BCOM Strategy	Breakthrough Cross Over Momentum strategy, a variation of a MAC-O strategy that used break-through signals to identify momentum of a pseudo-random movement.
Benchmark	Benchmarks describe the final comparison of results, of different shocks, and outline of conclusions.
Characteristics of Communication and Interaction	One of the traits defining a contribution as a formal evidence for the purposes of crowd-based investigation. Collaboration in large-scale is a direct result of how well contributions foster communication and interaction, therefore a contribution must support three basic characteristics of communication and interaction: analytical description, granularity and simplicity

Computational Representation[63]	A representation system based on facets, contributions, and constraints of data and used to define concepts related to a specific domain of knowledge. A layer of abstraction that is required in order to define domain-specific concepts in computers, in a way these concepts can be shared in a crowd for the purposes of controlled investigation in large-scale.
Computational Taxonomy	An inventory of computer technologies available and relevant for the implementation of domain-specific cases of use. An exact composition of a computational taxonomy is non-deterministic and heavily dependent on biases and experience of the individual performing the selection, as well as his personal assessment of the relevancy of the technology for the case of use at hand
Constraints of Data	Structural constraints defining domain-specific rules of association between entities and relationships. These rules and associations describe for computers what is feasible for a domain of knowledge, in real world. Those structural constraints use an abstract layer of data to define restrictions on a separate layer of abstractions, based themself on data, hence the term meta-data to refer to constraints of data

[63] To avoid a common misunderstanding, it is important to note that a computational representation is **not** a programming language, or a software implementation.

Contribution	Shareable and formal evidences of an objective crowd-based investigation. To be qualified as shareable and formal evidences for a crowd-based investigation scenario, collaborations have to carry specific traits: evidential properties, representational perspectives, and characteristics of communication and interaction. As shareable evidences they can be exchanged, reused and traced therefore becoming a vehicle for collaborative scientific investigation.
Evidence	The available body of facts or information indicating whether a belief or proposition is true or valid (Oxford University 2010)
Evidential Properties	One of the traits defining a contribution as a formal evidence for the purposes of crowd-based investigation. Evidential properties are classification, identification, record of provenance, and ownership and security
Facet	A definable aspect that make up a subject or an object; denomination of things that are similar or related, but yet distinct things (Merriam-Webster Online Dictionary 2016).

Financial Model Financial Model is a type of a contribution, defined as a component of a computational representation for the field of economics. A financial model represents observable phenomena in economics, simplified to the right scale, and adjusted to use in a process of crowd-based investigation

FRACTI FRACTI is an acronym for a FRAmework for Collaboration and Transparent Investigation in economics. FRACTI is an abstraction designating the three cognitive and non-cognitive enablers of crowd-based investigation for the field of economics: methods of proof, large-scale collaboration, and the computational representation for the field of economics. FRACTI is a conceptual abstraction, and **is not a software implementation, or a programming language.**

Intrinsic Element Intrinsic elements of a representation are elements that do not have to be explicitly described, and are enforced by an eventual computer implementation of the representation. An intrinsic element can be assumed to be in place regardless of any specific expressions on the representation itself.

MAC-O Strategy Moving Average Cross-Over strategies are special types of BCOM strategies in which the break-through is the movement of price movement up or down its moving average.

Meta-model	The set of structural constraints of data in a specific domain of knowledge.
Momentum	A tendency of a movement to remain moving one way, up or down. Momentum strategies identify profit opportunities by assuming that, unlike a purely random movement of a price (random walk), a price movement might carry some "inertia" and tend to gain on an already higher price, "many more times than not … the strong get stronger and the weak get weaker" (Chestnutt 1955).
Record of Provenance	Chronology of the ownership, custody or location of contributions.
Representational Perspectives	One of the traits defining a contribution as a formal evidence for the purposes of crowd-based investigation. Representational perspectives are defined are either intrinsic or extrinsic.
Representational Process	Set of generic procedures to derive an ad hoc computational representation, for any given domain of knowledge.
Shock	Shocks describe each of the executions of a financial model, recording utilized data and results of each individual execution.

BIBLIOGRAPHY

09 22, 2016. https://www.merriam-webster.com.

Abdelhamid, Sabra. "Ibn al-Haytham - Brief life of an Arab mathematician: died circa 1040." *Harvard Magazine*, Sep-Oct 2003.

Agha, Gul. "Actors: A Model of Concurrent Computation in Distributed Systems." Technical Report, Artificial Intelligence Laboratory, MIT, Cambridge, 1985.

Alkhateeb, Ahmed. "Science has outgrown the human mind and its limited capacities." *Aeon*. April 24, 2017. https://aeon.co/ideas/science-has-outgrown-the-human-mind-and-its-limited-capacities (accessed Nov 18, 2017).

Amdahl, Gene. "Validity of the single processor approach to achieving large scale computing capabilities." *AFIPS spring joint computer conference*, 1967.

Anderson, Rick. "Cabell's New Predatory Journal Blacklist: A Review." *The Scholarly Kitchen*. Jul 25, 2017. https://scholarlykitchen.sspnet.org/2017/07/25/cabells-new-predatory-journal-blacklist-review/ (accessed Oct 12, 2017).

Apolloni, Bruno, Dario Malchiodi, and Sabrina Gaito. *Algorithmic Inference in Machine Learning*. Adelaide: Advanced Knowledge International, 2006.

Archer, Leonard Bruce. *Systematic method for designers*. London: Council of Industrial Design, 1965.

Arthur, Brian. *Complexity Economics: A Different Framework for Economic Thought*. Report, Santa Fe Institute, Santa Fe Institute, 2013.

Bainomugisha, Engineer, Andoni Lombide Carreton, Tom Van Cutsem, Stijn Mostinckx, and Wolfgang De Meuter. "A Survey on Reactive Programming." *ACM Computing Surveys* 45, no. 4 (8 2013).

Baiocchi, Giovanni. "Reproducible research in computational economics: guidelines, integrated approaches, and open source software." *Computational Economics* (Springer) 30, no. 1 (2007): 19-40.

Balsara, Nauzer J, Gary Chen, and Lin Zheng. "The Chinese Stock Market: An Examination of the Random Walk Model and Technical Trading Rules." *The Quarterly Journal of Business and Economics*, Spring 2007.

Beall, Jeffrey. "Criteria for Determining Predatory Open - Access Publishers." *Scholarly Open Access (Web Archive)*. Jan 1, 2015. https://web.archive.org/web/20161130184313/https://scholarlyoa.files.wordpress.com/2015/01/criteria-2015.pdf (accessed Oct 1, 2017).

Beall, Jeffrey. "Dangerous Predatory Publishers Threaten Medical Research." *Journal of Korean Medical Science* 31, no. 10 (Oct 2016): 1511-1513.

—. *Scholarly Open Access*. 12 10, 2008. https://scholarlyoa.com/.

Beall, Jeffrey. "The Open-Access Movement is Not Really about Open Access." *Triple C* (University of Westminster) 11, no. 2 (2013): 589-597.

Begley, Glenn, and Lee Elis. "Drug development: Raise standards for preclinical cancer research." *Nature* (Nature International Journal of Science), Mar 2012: 531-533.

Berners-Lee, T, R Fielding, and L Masinter. "Uniform Resource Identifier (URI): Generic Syntax." RFC, Network Working Group, The Internet Engineering Task Force, 2005.

Bisig, T, A Dupuis, V Impagliazzo, and R Olsen. "The scale of market quakes." *Quantitative Finance* 12, no. 4 (2012): 501-508.

Black, Fischer. "Noise." *The Journal of Finance* (The American Finance Association) 41, no. 3 (July 1986): 529-543.

Bohannon, John. "Who's Afraid of Peer Review?" *Science Magazine* (Science) 342, no. 6154 (Oct 2013): 60-65.

Bollen, Kenneth, John Cacioppo, Robert Kaplan, Jon Krosnick, and James Olds. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Subcommitte Report, Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Science, National Science Foundation, National Science Foundation, 2015.

Brabham, Daren. "Crowdsourcing as a Model for Problem Solving." *Convergence: The International Journal of Research into New Media Technologies* , Feb 2008: 75-90.

Brauer, Wilfried, and Wolfgang Reisig. "Carl Adam Petri und die "Petrinetze" (Carl Adam Petri and "Petri Nets")." *Informatik-Spektrum* (Springer Verlag) 29, no. 5 (Oct 2006): 369-374.

Bricker, Phillip. "Ontological Commitment." In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2016.

Brown, Robert. *A Brief Account of Microscopical Observations Made in the Months of June, July and August 1827, on the Particles Contained in the Pollen of Plants; and on the General Existence of Active Molecules in Organic and Inorganic Bodies*. Manuscript. London: Unpublished, 1827.

Brown, Stephen J, William N Goetzmann, and Alok Kumar. "The Dow Theory; William Peter Hamilton's Track Record Reconsidered." *SSRN*. Jan 23, 1998. http://ssrn.com/abstract=5869.

Brown, Tim. "Design Thinking." *Harvard Business Review*, Jun 2008.

Browning, E S. "Reading the Market's Tea Leaves." *Wall Street Journal*, Jan 2006.

Burkus, David. "How Hierarchies Kill Creativity." *The Creativity Post*. Jun 26, 2012. http://www.creativitypost.com/business/how_hierarchies_kill_creativity (accessed Dec 2, 2017).

—. "Why Great Ideas Get Rejected." *99u*. 12 2, 2013. http://99u.com/articles/7207/Why-Great-Ideas-Get-Rejected (accessed 12 2, 2017).

Camerer, Colin F, and George Loewenstein. "Behavioral Economics: Past, Present, Future." Department of Social and Decision Sciences, Carnegie-Mellon University, 2002.

Camerer, Colin, Anna Dreber, Eskil Forsell, Teck-Hua Ho, and Johannesson Magnus. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, Mar 2016.

Camerer, Colin, George Lowenstein, and Drazen Prelec. "Neuroeconomics: How Neuroscience Can Inform Economics." *Journal of Economic Literature* XLIII (03 2005): 9-64.

Campbell, W.G. *Form and Style in Thesis Writing, a Manual of Style*. Chicago: The University of Chicago Press, 1990.

Carolan, Christopher. "Autumn Panics: A Calendar Phenomenon." *Dow Award* (Market Technicians Association), 1998.

Cassidy, John. "The Reinhart and Rogoff Controversy: A Summing Up." *The New Yorker*, April 26, 2013.

Castillo, Enrique, Bertha Guijarro-Berdiñas, Oscar Fontenla-Romero, and Amparo Alonso-Betanzos. "A Very Fast Learning Method for Neural Networks Based on Sensitivity Analysis." *Journal of Machine Learning Research* 7 (07 2006): 1159–1182.

Castner, Johannes. "Measures of Cognitive Distance and Diversity." *SSRN*. Jul 31, 2014. https://ssrn.com/abstract=2477484.

Chakraborti, Anirban, Ioane Muni Toke, Marco Patriarca, and Frédéric Aberrgel. "Econophysics: Empirical facts and agent-based models." *Quantitative Finance*, no. 11 (Jan 2011): 1013-1041.

Chan, Nicholas Tung, and Christian Shelton. "An Electronic Market Maker." *DSpace@MIT*. 04 17, 2001. http://hdl.handle.net/1721.1/7220 (accessed 11 2, 2013).

Chaudhuri, Kausik, and Yangru Wu. "Random Walk vs Vreaking Trend in Stock Prices: Evidence from Emerging Markets." *Journal of Banking and Finance* 27 (2003): 575-592.

Chen, Kay-Yut. "An Economics Wind Tunnel: The Science of Business Engineering." *Experimental and Behavioral Economics - Advances in Applied Microeconomics* (John Morgan, Elsevier Press) 13 (2005).

Chestnutt, George A. *Stock Market Analysis: Facts and Principles*. American Investors Service, 1955.

Cintas, Pedro. "Francis Bacon: An Alchemical Odissey Through the Novum Organum." *Bulletin of the History of Chemistry* (American Chemical Society) 28, no. 2 (2003): 65-75.

Claerbout, Jon, and Martin Karrenbach. "Electronic documents give reproducible research a new meaning." *Society of Exploration Geophysicists*, 1992: 601-604.

Clinger, William Douglas. *Foundations of Actor Semantics*. Doctoral Dissertation, Electric Engineering and Computer Science, MIT, Cambridge: MIT, 1981.

Cokol, Murat, Ivan Iossifov, Raul Rodriguez-Esteban, and Andrey Rzhetsky. "How many scientific papers should be retracted?" *EMBO Reports* (European Molecular Biology Organization) 8, no. 5 (May 2007): 422-423.

Colquhoun, David. *The Problem With P-Values*. Oct 11, 2016. https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant (accessed 11 01, 2017).

Cont, Rama, Sasha Stoikov, and Rishi Talreja. "A Stochastic Model for Order Book Dynamics." *Operations Research* 58, no. 3 (May 2010): 549-563.

Cowles, Alfred. "Can Stock Market Forecasters Forecast?" *Econometrica* 1, no. 3 (Jul 1933): 309-324.

Croarkin, Carroll, and Paul Tobias. *e-Handbook of Statistical Methods*. Apr 2012. http://www.itl.nist.gov/div898/handbook// (accessed Oct 2015).

Crotty, David. "Predatory Publishing as a Rational Response to Poorly Governed Academic Incentives." *The Scholarly Kitchen*. Feb 27, 2017. https://scholarlykitchen.sspnet.org/2017/02/28/predatory-publishing-rational-response-poorly-governed-academic-incentives/ (accessed Oct 11, 2017).

Crupi, Vincenzo. *Confirmation*. Edited by Edward Zalta. 2016. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=confirmation.

Curcin, V, and M Ghanem. "Scientific workflow systems - can one size fit all?" Proceedings of the 2008 IEEE, CIBEC'08, Department of Computing, Imperial College London, London, 2008.

Davis, R, and H Shrobe. "Representing Structure and Behavior of Digital Hardware." *IEEE Computer*, Oct 1983: 75-82.

Davis, Randall, Howard Shrobe, and Peter Szolovits. "What Is a Knowledge Representation?" *AI Magazine*, Spring 1993.

Dean, J., and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." *OSDI '04*. San Francisco: Google Inc., 2004.

Dorst, Kees. *Frame Innovation*. Cambridge, Massachussets: The MIT Press, 2015.

Duffy, Daniel. "The Meshless (Meshfree) Method in Financial Engineering." In *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach*, by Daniel Duffy, edited by John Wiley & Sons. Oxford: John Wiley & Sons, 2006.

Eco, Humberto. *How to Write a Thesis*. Cambridge, Massachusetts: The MIT Press, 1977.

Efron, Bradley. "Controversies in the Foundations of Statistics." *The American Mathematical Monthly* (Mathematical Association of America) 85, no. 4 (1978): 231-246.

Einstein, Albert, and Leopold Infeld. *The Evolution of Physics*. New York: Simon & Schuster, 1938.

Ellerton, Peter. "What Do We Mean by Scientific Truth?" *Real Clear Science*. May 17, 2012. http://www.realclearscience.com/articles/2012/05/17/what_do_we_mean_by_scientific_truth_106273.html (accessed Jan 5, 2017).

Faleiro Jr, Jorge M. *QuantLET Example: Backtesting Momentum Strategies using Streams and Monte Carlo Simulations*. Jul 2015. http://goo.gl/EWGqyO.

—. *QuantLET Example: Moving Average Cross Over*. Sep 28, 2014. https://goo.gl/jEo6Mt (accessed Mar 13, 2015).

—. *QuantLET Example: Infinite Spreadsheets*. Sep 28, 2014. https://goo.gl/e5AzKU (accessed Apr 2015).

Faleiro Jr, Jorge M. *A Language for Large Scale Collaboration in Economics: A Streamlined Computational Representation of Financial Models*. Report, Centre of Computational Representation and Economic Agents, University of Essex, Colchester: University of Essex, 2017.

Faleiro Jr, Jorge M. "A Scientific Workflow System Applied to Finance." *Presentation slides for supervisory research meeting*. New York City, 2013a.

Faleiro Jr, Jorge M. *Automating Truth: The Case for Crowd-Powered Scientific Investigation in Economics*. Report, Centre of Computational Finance and Economic Agents, University of Essex, Colchester: University of Essex, 2016a.

Faleiro Jr, Jorge M. *Full Research Proposal*. Full Research Proposal, Centre of Computational Finance and Economic Agents, Colchester: University of Essex, 2014a.

Faleiro Jr, Jorge M. "Full Research Proposal." Progress Report, Centre of Computational Finance and Economical Agents, University of Essex, Colchester, 2015.

Faleiro Jr, Jorge M. *Outline Research Proposal*. Outline Research Proposal, Centre of Computational Finance and Economic Agents, Colchester: University of Essex, 2014.

—. *QuantLET: an open source, event-driven framework for real-time analytics*. 08 2008. http://quantlet.net.

Faleiro Jr, Jorge M. "Reference Model: Architecture, Concepts and Fundaments." *Presentation slides for supervisory research meeting*. New York City, 04 13, 2013.

Faleiro Jr, Jorge M. "Short Research Proposal." Research Proposal, 2012.

Faleiro Jr, Jorge M, and Edward P. K. Tsang. "Crowd-Powered Monitoring in Large Scale: A Collaborative Environment for Early Detection and Investigation of Systemic Failures in Financial Markets." In *Submitted: Handbook of Global Financial Markets: Transformations, Dependence, and Risk Spillovers*. World Scientific Publishing, 2019.

Faleiro Jr, Jorge M, and Edward P. K. Tsang. "Black Magic Investigation Made Simple: Monte Carlo Simulations and Historical Back Testing of Momentum Cross-Over Strategies Using FRACTI Patterns." Working Paper, Centre of Computational Finance and Economic Agents, University of Essex, Colchester, 2016.

—. "Supporting Crowd-Powered Science in Economics: FRACTI, A Conceptual Framework for Large-Scale Collaboration and Transparent Investigation in Financial Markets." *14th Simulation and Analytics Seminar*. Helsinki: Bank of Finland, 2016a.

—. "Supporting Crowd-Powered Science in Economics: FRACTI, A Conceptual Framework for Large-Scale Collaboration and Transparent Investigation in Financial Markets." *14th Simulation and Analytics Seminar*. Helsinki: Bank of Finland, 2016a.

Faleiro Jr, Jorge Martins. 07 01, 2007. http://goo.gl/w7ouS8 (accessed 07 01, 2007).

Faleiro Jr, Jorge Martins. "Doctoral Research Pathway Milestone M3: Full Research Proposal." Research Proposal, Centre of Computational Finance and Economic Agents, University of Essex, Colchester, 2014.

Faleiro Jr, Jorge Martins. *Implementation of a "Cyclical Long Position Strategy" in QuantLET*. Presentation. 05 2013.

Faleiro Jr, Jorge Martins. "Short Research Proposal." 2012.

Fama, Eugene F. "Random Walks in Stock Market Prices." *Financial Analysis Journal*, Sep-Oct 1965: 55-59.

Fanelli, Daniele, and Wolfgang Glanzel. "Bibliometric Evidence for a Hierarchy of the Sciences." *PLOS ONE* (US National Library of Medicine), Jun 2013.

Fasshauer, Greg. "Meshfree Methods." In *Handbook of Theoretical and Computational Nanotechnology*, edited by M. Rieth and W. Schommers, 33-97. American Scientific Publishers, 2006.

Fetzer, James. "Carl Hempel." In *The Stanford Encyclopedia of Phylosophy*, edited by Edward N Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2017.

Fisher, Edwin. *Effect of Smoking on Nonsmokers*. Hearing, House of Representatives, Washington D.C.: U.S. Government Printing Office, 1978, 2-5.

Fisher, Ronald A. "On the mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* (Royal Society) 222 (1922): 309-368.

Foata, Laurent, Michael Vidhamali, and Frédéric Abergel. "Multi-Agent Order Book Simulation: Mono- and Multi-Asset High-Frequency Market Making Strategies." In *Econophysics of Order-Driven Markets*. Springer, 2011.

Foster, John. *From Simplistic to Complex Systems in Economics*. report, School of Economics, University of Queensland, St Lucia: University of Queensland, 2004.

Franzoni, Chiara, and Henry Sauermann. "Crowd science: The organization of scientific research in open collaborative projects." Edited by Elsevier. *Research Policy* 43 (2014): 1-20.

Frazee, Ayssa, Sarven Sabunciyan, Kasper Hansen, and Rafael Irizarry. "Differential expression analysis of RNA-seq data at single-base resolution." *Biostatistics* 15, no. 3 (Jul 2014): 413-426.

Freedman, David. "Why Economic Models Are Always Wrong." *Scientific American*, Oct 2011.

Friendly, Michael, and Daniel Denis. "The Early Origins and Development of the Scatterplot." *Journal of the History of Behavioral Sciences* 41 (Spring 2005): 103-130.

Galton, Francis. "Vox Populi." *Nature* 75 (Mar 1907): 450-451.

Gauch, Hugh. *Scientific Method in Practice*. 1st. Cambridge: Cambridge University Press, 2003.

Gellatly, Angus. "Human Inference." In *Human and Machine Problem Solving*, 233-264. Boston, MA: Plenum Press, 1989.

Gentleman, Robert. "Reproducible Research: A Bioinformatics Case Study." *Statistical Applications in Genetics and Molecular Biology* 4, no. 1 (2005): Article 2.

Gingold, R A, and J J Monaghan. "Smoothed particle hydrodynamics: theory and application to non-spherical stars." *Monthly Notices of the Royal Astronomical Society* (Royal Astronomical Society) 181, no. 3 (Dec 1977): 375–389.

Godfrey-Smith, Peter. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: The University of Chicago Press, 2003.

Goecks, Jeremy, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biology*, 11 2010.

Goodman, Steven, Daniele Fanelli, and John Ioannidis. "What does research reproducibility mean?" *Science Translational Medicine* 8, no. 341 (Jun 2016): 1-6.

Gordon, Michael. *Compiler Techniques for Scalable Performance of Stream Programs and Multicore Architectures*. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge: Massachusetts Institute of Technology, 2010.

Gordon, Michael, et al. "A Stream Compiler for Communication-Exposed Architectures." *International Conference on Architectural Support for Programming Languages and Operating Systems* , 08 2002.

Gowers, Timothy. "A combinatorial approach to density Hales-Jewett." *Gower's Blog*. Feb 1, 2009b. https://gowers.wordpress.com/2009/02/01/a-combinatorial-approach-to-density-hales-jewett/ (accessed Feb 10, 2010).

—. "Is massively collaborative mathematics possible?" *Gowers;s Weblog*. Jan 27, 2009a. https://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible/ (accessed Feb 15, 2010).

Gowers, Timothy, and Michael Nielsen. "Massively collaborative mathematics." *Nature*, Oct 2009: 879-881.

Grauwin, Sebastian, et al. "Identifying and modeling the structural discontinuities of human interactions." *Scientific Reports* (Nature), Apr 2017.

Greif, Irene. "Semantics of Communicating Parallel Processes." Doctoral Dissertation, Electric Engineering and Computer Science, MIT, Cambridge, 1975.

Griffiths, Thomas, and Tenenbaum Joshua. "Theory-Based Causal Induction." *Psychological Review* (American Psychological Association) 116, no. 4 (2009): 661-716.

Guazzini, Andrea, Daniele Vilone, Camilo Donati, Annalisa Nardi, and Zoran Levnajić. "Modeling crowdsourcing as collective problem solving." *Nature* (Nature Research), Nov 2015.

Gómez-Torrente, Mario. "Logical Truth." In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2017.

Gupta, Sonal, and Christopher D Manning. "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers." *Proceedings of the 5th International Joint Conference on Natural Language Processing* . Chiang Mai: AFNLP, 2011. 1-9.

Hamermesh, Daniel. "Viewpoint: Replication in economics." *Canadian Journal of Economics* (Wiley), Jul 2007.

Hamilton, William P. "William Peter Hamilton's Editorials in the Wall Street Jornal." 1903-1929. http://goo.gl/MoklCh (accessed 10 5, 2015).

Hamilton, William P, and Charles H Dow. *The Stock Market Barometer*. New York: Harper & Brothers, 1922.

Hansson, Sven Ove. "Science and Pseudo-Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2017.

Harel, David, and Amir Pnueli. "On the Development of Reactive Systems." *Logics and Models of Concurrent Systems*, 1985: 477-498.

—. "Statemate: a working environment for the development of complex reactive systems." *ICSE '88 Proceedings of the 10th international conference on Software engineering*. Los Alamitos: IEEE Computer Society Press, 1988.

Hawthorne, James. "Inductive Logic." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2017.

Hayek, Friedrich A. "The Use of Knowledge in Society." *American Economic Review* (American Economic Association), Sep 1945: 519-530.

Hayes, Patrick J. *Naive Phisics I: Ontology for Liquids*. Working Paper, Institut Pour Les Etudes Semantiques et Cognitives, Universite de Geneve, Geneve: Fondazione Dalle Molle, 1978.

Heisenberg, Werner. *Physics and Philosophy: The Revolution in Modern Science*. 1st. Torchbooks, 1958.

Herndon, Thomas, Michael Ash, and Robert Pollin. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogof." Working Paper, Political Economy Research Institute, University of Massachusetts Amherst, Amherst, 2013.

Hewitt, Carl, Peter Bishop, and Richard Steiger. "A Universal Modular ACTOR Formalism for Artifical Intelligence." *IJCAI'73 Proceedings of the 3rd international joint conference on Artificial intelligence* (Morgan Kaufmann Publishers Inc.), Aug 1973: 235-245.

Hoare, Charles Anthony Richard. "Communicating Sequential Processes." *Communications of the ACM* (ACM) 21, no. 8 (Aug 1978): 666-677.

—. *Communicating Sequential Processes*. Oxford: Oxford University Computing Laboratory, 2015.

Hoffmann , Arvid I, and Hersh Shefrin . "Technical Analysis and Individual Investors." *Journal of Economic Behavior and Organization* 107, no. November (Feb 2014): 487-511.

Hohpe, G.,, and B. Woolf. *Enterprise Integration Patterns*. Boston: Addison-Wesley, 2012.

Holton, Gerald. *Science and Anti-Science*. Cambridge: Harvard University Press, 1993.

Horgan, John. "The Death of Proof." *Scientific American*, 1993.

Horgan, John. "Was Philosopher Paul Feyerabend Really Science's "Worst Enemy"?" *Scientific American*, Oct 2016.

Huebscher, Robert. *Burton Malkiel Talks the Random Walk*. Jul 7, 2009. http://goo.gl/1CPgpr.

Hunter, John. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, 2007: 90-95.

Ioannidis, John. "Why Most Published Research Findings Are False." *PLoS Med* 2, no. 8 (Aug 2005): 0696-0701.

Ioannidis, John, et al. "Repeatability of Published Microarray Gene Expression Analyses." *Nature Genetics* 41 (2009): 149-155.

Jürgen, Klein. "Francis Bacon." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2016.

Jeffrey, Paul. " Smoothing the Waters: Observations on the Process of Cross-Disciplinary Research Collaboration ." *Social Studies of Science* (SAGE Journals) 33, no. 4 (Aug 2003): 539-562.

Jeffreys, Harold. "Probability, Statistics, and the Theory of Errors ." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* (Royal Society) 140, no. 842 (Jun 1933): 523-535.

Jha, Alok. "Text mining: what do publishers have against this hi-tech research tool?" *The Guardian*, May 23, 2012.

Johnson, Steven. *Where Good Ideas Come From: The Natural History of Innovation*. New York: Riverhead Books, 2010.

Jones, Eric, Travis Oliphant, and Pearu Peterson. 2001. http://www.scipy.org/ (accessed 07 02, 2015).

Kahn, Giles. "The Semantics of a Simple Language for Parallel Programming." Edited by IRIA-Laboria and Commissariat a l'Energie Atomique. *Information Processing* (North-Holland Publishing Company) 74 (1974): 471-475.

Kamburugamuve, Supun, and Geoffrey Fox. *Survey of Distributed Stream Processing*. School of Informatics and Computing, Indiana University, Bloomington: Indiana University, 2013.

Slingshot, Documentary. Directed by Paul Lazarus. Performed by Dean Kamen. 2014.

Kapitan, Thomas. "Peirce and the autonomy of abductive reasoning." *Erkenntnis* (Kluwer Academic Publishers) 37, no. 1 (1992).

Karp, Richard, and Rayamond Miller. "Properties for a Model for Parallel Computations: Determinacy, Termination, Queuing." *SIAM Journal of Applied Mathematics*, Jan 1966: 1390–1411.

Kass, Robert E. "Statistical Inference: The Big Picture." *Statistical Science* (Institute of Mathematical Statistics) 26, no. 1 (2011): 1-9.

Kay-Yut, Chen. "Playing Games for Better Business: Using Economics Experiment to Test Business Policies." White Paper, Hewlett-Packard Laboratories, 2006.

Kelley, Eric, and Paul Tetlock. "How Wise Are Crowds? Insights from Retail Orders and Stock Returns." *The Journal of Finance* (The American Finance Association) 68, no. 3 (Feb 2013): 1229-1265.

Kim, Yongsik, Hyeong-Ohk Bae, and Hyeng Keun Koo. "Option pricing and Greeks via a moving least square meshfree method." *Quantitative Finance* (Taylor & Francis) 14, no. 10 (Nov 2013): 1753-1764.

Koenker, Roger, and Achim Zeileis. "On Reproducible Econometric Research." *Journal of Applied Econometrics* , 2009: 833-847.

Krauss, Lawrence. *A Universe From Nothing*. 1st. New York: Free Press, 2012.

Krishnamurti, Jiddu. *J. Krishnamurti Online*. 6 1, 2014. http://www.jkrishnamurti.org (accessed 6 3, 2017).

Lee, Edward Ashford, and David G Messerschmitt. "Static Scheduling of Synchronous Data Flow Programs for Digital Signal Processing." *IEEE Transactions on Computers* (IEEE) C-36, no. 1 (Jan 1987): 24-35.

Leemis, Lawrence, and Stephen Park. *Discrete-Event Simulation: A First Course*. Williamsburg, VA: Pearson, 2006.

Lehrer, Jonah. "Can We Prevent the Next Bubble?" *Wired*, 06 2011.

Lenhard, Johannes. "Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson." *British Journal for the Philosophy of Science* (Oxford University), 2006: 69-91.

Lindley, Dennis V. "The Philosophy of Statistics." *Journal of the Royal Statistical Society. Series D (The Statistician)* (Blackwell Publishing) 49, no. 3 (2000): 293-337.

Lopez, Alexander Guarin. "Meshfree methods in financial engineering." Thesis, University of Essex, Colchester, 2012.

Lupyan, Gary, and Emily Ward. "Language can boost otherwise unseen objects into visual awareness." *Proceedings of the National Academy of Sciences of the United States of America*. Stanford: PNAS, 2013. 14196-14201.

Madhavan, Ananth. "Market microstructure: A survey." *Journal of Financial Markets* (Elsevier) 3 (3 2000): 205-258.

Mandrekar, V. "Mathematical Work of Norbert Wiener." *Notices of American Mathematical Society* 42, no. 6 (Juliy 1995): 664-669.

Mannes, Albert. "Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision." *Management Science* (INFORMS), Jun 2009: 1267-1279.

Markowitz, Harry. "Portfolio Selection." *The Journal of Finance* 7, no. 1 (Mar 1952): 77-91.

—. *Portfolio Selection: Efficient Diversification of Investments*. Edited by Cowles Foundation for Research in Economics. New York: John Wiley & Sons, 1959.

Marshall, Ben R, Rochester H Cahan, and Jared M Cahan. "Techinical Analysis Around the World." Aug 1, 2010. http://ssrn.com/abstract=1181367 .

Martinson, Brian C, Melissa S Anderson, and Raymond De Vries. "Scientists Behaving Badly." *Nature* 435, no. 9 (June 2005).

Matloff, Norm. *Introducton to Discrete-Event Simulation and the SimPy Language*. Edited by University of California in Davis. University of California in Davis, 2008.

McGraw Hill Financial. "S&P 500." *S&P Dow Jones Indices*. 10 15, 2015a. http://us.spindices.com/indices/equity/sp-500 (accessed 10 15, 2015).

—. "S&P 500." *S&P 500 Fact Sheet*. Oct 15, 2015b. http://goo.gl/Ib5AXX (accessed Oct 15, 2015).

McIlroy, Douglas. *The Origin of Unix Pipes*. 10 11, 1964. http://doc.cat-v.org/unix/pipes/ (accessed 08 6, 2015).

McKinney, Wes. "Pandas: Data Structures for Statistical Computing in Python." *9th Python in Science Conference*. 2010. 51-56.

Merriam-Webster. *Ad hoc*. 2018. https://www.merriam-webster.com/dictionary/ad hoc. (accessed Feb 2, 2018).

—. *inference*. Jul 15, 2018. https://www.merriam-webster.com/dictionary/inference (accessed Jul 17, 2018).

Miller, Chet, Linda Burke, and William Glick. "Cognitive Diversity among Upper-Echelon Executives: Implications for Strategic Decision Processes." *Strategic Management Journal* (Wiley) 19, no. 1 (Jan 1998): 39-58.

Mitchell, Rebecca, and Stephen Nicholas. " Knowledge Creation in Groups: The Value of Cognitive Diversity, Transactive Memory and Open-mindedness Norms ." *The Electronic Journal of Knowledge Management* 4, no. 1 (2006): 67-74.

Mueller, Jennifer. "Managers Reject Ideas Customers Want." *Harvard Business Review*, Jul-Aug 2014.

Mueller, Jennifer, Shimul Melwani, and Jack Goncalo. "The Bias Against Creativity: Why People Desire But Reject Creative Ideas." *Cornell University, ILR School*. Aug 2011. http://digitalcommons.ilr.cornell.edu/articles/450/ (accessed Dec 2015).

Mulder, Henk, and Barbara van de Velde-Schlick. *Moritz Schlick Philosophical Papers: Volume I*. Translated by Peter Heath. Boston: Springer, 1978a.

Muller, Jennifer, Cheryl Wakslak, and Viswanathan Krishnan. "Construing creativity: The how and why of recognizing creative ideas." *Journal of Experimental Social Psychology*, 3 2014: 81-87.

Munir, Raji Al. *A Very Short Introduction to The Modern Scientific Method and The Nature of Modern Science*. Kindle. Edited by Amazon Books. rm@munir.info, 2010.

Murat, Ahmed, Anwei Chai, Xiaowei Ding, Yunjiang Jiang, and Yunting Sun. "Statistical Arbitrage in High Frequency Trading Based on Limit Order Book Dynamics." Stanford University, 2009.

Murata, Tadao. "Petri Nets: Properties, Analysis and Applications." *Proceedings of the IEEE*. IEEE, 1989.

Muthukrishna, Michael, and Joseph Henrich. "Innovation in the collective brain." *Phylosophical Transactions of the Royal Society* (The Royal Society) 371, no. 1690 (Feb 2016).

Newman, James. *The World of Mathematics*. New York: Simon and Schuster, 1956.

Neyman, Jerzy. "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." *Journal of the Royal Statistical Society* (Royal Statistical Society) 97, no. 4 (1934): 558-625.

Nielsen, Michael. *PolyMath Wiki*. Dec 11, 2011. michaelnielsen.org/polymath1 (accessed Jan 10, 2016).

—. *Reinventing Discovery: The New Era of Networked Science*. 1st. Princeton, New Jersey: Princeton University Press, 2012.

NIH. "Working Definition of Bioinformatics and Computational Biology." Bioinformatics Definition Committee, 2000.

Nilsson, Henrik, Antony Courtney, and John Peterson. "Functional Reactive Programming, Continued." *Proceedings of the 2002 ACM SIGPLAN*. Pittsburg: ACM Press, 2002. 51-64.

Hypothetico-Deductivism as a Methodology in Science. Vol. 28, in *Philosophy, Science, Education and Culture. Science & Technology Education Library*, by Robert Nola and Gurol Irzik. Dordrecht: Springer, 2006.

Norton, Vic. "Adjusted Closing Prices." Department of Mathematics and Statistics, Bowling Green State University, 2010.

NPR. 04 19, 2013. http://www.npr.org/blogs/money/2013/04/19/177999020/episode-357-how-much-should-we-trust-economics (accessed 10 01, 2014).

Nuzzo, Regina. "Scientific method: Statistical errors." *Nature*, Feb 2014.

Oberdan, Thomas. "Moritz Schlick." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2016.

Ochs, Michael F. "Genomics Data Analysis Pipelines." *Biomedical Informatics for Cancer Research* (Springer US), 2010: 117-137.

Olsen, Richard, and Clive Cookson. "How Science Can Prevent the Next Buble." *Financial Times*, 02 2009.

Omobowale, Ayokunle Olumuyiwa, Olayinka Akanle, Adebusuyi Isaac Adeniran, and Kamorudeen Adegboyega. "Peripheral scholarship and the context of foreign paid publishing in Nigeria." *Current Sociology* (SAGE Journals) 62, no. 5 (Sep 2014): 666-684.

Open Science Collaboration. "Estimating the reproducibility of psychological science." *Science* 349, no. 6251 (Aug 2015): 943-951.

O'Toole, Garson. "In God We Trust; Others Must Provide Data." *Quote Investigator*. Dec 29, 2017. https://quoteinvestigator.com/2017/12/29/god-data/ (accessed Feb 10, 2018).

Oxford English Dictionary. *cognition*. 12 01, 2011. https://en.oxforddictionaries.com/definition/cognition (accessed 03 20, 2017).

Oxford University. *Oxford Dictionary of English*. 3. Edited by Angus Stevenson. Oxford University Press, 2010.

Palard, Julien. *Pipes: A Infix Notation Library*. 8 29, 2012. https://github.com/JulienPalard/Pipe (accessed 04 15, 2013).

Panayi, Efstathios, and Gareth Peters. "Stochastic simulation framework for the limit order book using liquidity-motivated agents Read More: http://www.worldscientific.com/doi/abs/10.1142/S2424786315500139?journalCode =ijfe ." *International Journal of Financial Engineering* (World Scientific), Jul 2015.

Park, Cheol-Ho, and Scott H Irwin. "The Profitability of Technical Analysis: A Review." Research Report, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, Urbana-Champaign, 2004, 106.

Park, Cheol-Ho, and Scott H Irwin. "What Do We Know About the Profitability of Technical Analysis?" *Journal of Economic Surveys* 21, no. 4 (Jul 2007): 786-826.

Patterson, Scott. "Technically, a Challenge for Blue Chips." *The Wall Street Journal*, Nov 2007.

Pedgen, Dennis. "Advanced tutorial: Overview of simulation world views." *Proceedings of the 2010 Winter Simulation Conference*. Baltimore, 2010. 5-8.

Peirce, Charles Sanders. *Studies in Logic*. Edited by Members of the Johns Hopkins University. Boston, MA: Little, Brown, and Company, 1883.

Penman, W. "Using the Thesis and Dissertation Templates." *Thesis and Dissertations Templates*. The University of Texas at Austin Graduate School. January 2011. http://www.utexas.edu/ogs/pdn/downloads (accessed 01 18, 2014).

Pérez, Fernando, and Brian E. Granger. "IPython: A System for Interactive Scientific Computing." *Computing in Science and Engineering* (IEEE Computer Society) 9, no. 3 (May/June 2007).

Perros, Harry. *Computer Simulation Techniques*. Edited by North Carolina State University. Raleigh, NC: Computer Science Department, 2009.

Petri, Carl. "Grundsätzliches zur Beschreibung diskreter Prozesse." *Kolloquium über Automatentheorie* (Birkhäuser-Verlag), 1967: 121-140.

Petri, Carl. "Kommunikation mit Automaten (Communication with Automata)." Ph.D. Thesis, University of Bonn, Bonn, 1962.

Polymath, D H J. "A new proof of the density Hales-Jewett theorem." *Annals of Mathematics* (Princeton University & Institute of Advanced Study) 175, no. 3 (May 2012).

Popper, Karl. *Conjectures and Refutations. The Growth of Scientific Knowledge*. New York: Basic Books, 1962.

—. *The Logic of Scientific Discovery*. 2005, New York: Taylor & Francis, 2005.

Press, Gil. "A Very Short History Of Big Data." *Forbes*, 5 09, 2013.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. "Believe it or not: how much can we rely on published data on potential drug targets?" *Nature Reviews* 10, no. 712 (Aug 2011).

Quandl. 07 2015. http://www.quandl.com (accessed 07 01, 2015).

Rayo, Agustin. "Ontological Commitment." Massachusetts Institute of Technology, Cambridge, 2007.

Reinhart, Carmem, and Kenneth Rogoff. "Growth in a Time of Debt." *American Economic Review* 100, no. 2 (05 2010).

Robinson, Stewart. *Simulation: The Practice of Model Development and Use*. West Sussex: John Wiley & Sons, 2004.

Rodrigues, Cassiano Terra. "The Method of Scientific Discovery in Peirce's Philosophy: Deduction, Induction, and Abduction." *Logica Universalis* (SP Birkhäuser Verlag Basel) 5, no. 1 (2011): 127–164.

Rodriguez-Esteban, Raul, and Andrey Rzhetsky. "Six senses in the literature: The bleak sensory landscape of biomedical texts." *EMBO Reports* (European Molecular Biology Organization) 9, no. 3 (Mar 2008): 212-215.

Rout, Jitendra, Kim-Kwang Choo, Amiya Dash, Sambit Bakshi, and Sanjay Jena. "A model for sentiment and emotion analysis of unstructured social media text." *Electronic Commerce Research* (Springer) 18, no. 1 (Mar 2018): 181–199.

Rzhetsky, Andrey, G Jacob Foster, Ian T Foster, and A James Evans. "Choosing experiments to accelerate collective discovery." *PNAS*, Nov 2015.

Samuelson, Paul A. "Proof that Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review*, Spring 1965: 41-49.

Savage, Leonard. *The Foundations of Statistics*. New York: John Wiley & Sons, 1954.

Schemer, Michael. "How the Survivor Bias Distorts Reality." *Scientific American*, 9 2014.

Scherfke, Stefan. *Discrete-event simulation with SimPy*. Jul 25, 2014. https://stefan.sofa-rockers.org/downloads/simpy-ep14.pdf (accessed Ago 10, 2015).

Schoeffel, M. "The Three Main Trading Strategies and Their Variations." In *Algorithmic Trading Strategies*, by Fudancy Research Group, edited by Fudancy Technology. Nyon: Fudancy Research Group, 2011.

Schroter, Sara, and Leanne Tite. "Open Access Publishing and Author-Pays Business Models: A Survey of Authors' Knowledge and Perceptions." *Journal of the Royal Society of Medicine* (The Royal Society of Medicine Journals) 99, no. 3 (Mar 2006): 141-148.

Schwab, Matthias, Martin Karrenbach, and Jon Claerbout. "Making Scientific Computations Reproducible." *Computing Sci Eng*, no. 2 (2000): 61-67.

Seethapathy, G S, J U Kumarand, and A S Hareesha. "India's scientific publication in predatory journals: need for regulating quality of Indian science and education ." *Current Science* 111, no. 11 (Dec 2016): 1759-1764.

Shen, Cenyu, and Bo-Christer Björk. "'Predatory' open access: a longitudinal study of article volumes and market characteristics." *BMC Medicine* 13, no. 230 (Oct 2015): 1-15.

Shermer, Michael. "How the Survivorship Bias Distorts Reality." *Scientific American* 322, no. 3 (Aug 2014).

Shih, Willy, and Sen Chai. "Data-Driven vs. Hypothesis-Driven Research: Making sense of big data." *Academy of Management Proceedings*. Academy of Management Journal, 2016.

Shin, Sun-Joo, and Eric Hammer. "Peirce's Deductive Logic." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2016.

Simon, Herbert A. "Herbert A. Simon - Biographical." *The Official Web Site of the Nobel Prize*. Nobel Media AB. Dec 10, 1978. https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/simon-bio.html (accessed Feb 22, 2018).

Simth, George, and Shah Ebrahim. "Data Dredging, Bias or Confounding - They Can All Get You Into the BMJ and the Friday Papers." *British Medical Journal* 325, no. 7378 (12 2002).

Sinatra, Roberta, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. "Quantifying the evolution of individual scientific impact." *Science* 354, no. 6312 (Nov 2016).

Soldatova, Larissa N, Amanda Clare, Andrew Sparkes, and Ross D King. "An ontology for a Robot Scientist." *Bioinformatics* (International Society for Computational Biology) 22, no. 14 (Jul 2016): e464–e471.

Spangler, Scott, et al. "Automated Hypothesis Generation Based on Mining Scientific Literature." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM, 2014. 1877-1886.

Spinellis, Diomidis. "Notable Design Patterns for Domain-Specific Languages." *Journal of Systems and Software*, February 2001: 91–99.

Staw, Barry. "Why No One Really Wants Creativity." In *Creative Action in Organizations*, by Barry Staw, edited by Cameron Ford and Dennis Gioia, 161-166. California: SAGE Publications, 1995.

Stephens, Robert. "A Survey of Stream Processing." *Acta Informatica* 34, no. 7 (Jul 1997): 491–541 .

Straumsheim, Carl. " No More 'Beall's List' ." *Inside Higher Ed*. Jan 18, 2017. https://www.insidehighered.com/news/2017/01/18/librarians-list-predatory-journals-reportedly-removed-due-threats-and-politics (accessed Oct 10, 2017).

Sulistio, Anthony, Chee Shin Yeo, and Rajkumar Buyya. "A taxonomy of computer-based simulations and its mapping to parallel and distributed systems simulation tools." *Software - Experience and Practice* (John Wiley & Sons, Ltd.) 34 (Apr 2004): 653-673.

Surowiecki , James. *The Wisdom of Crowds*. 1st. New York: Random House, 2004.

Swanson, Don. "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge." *Perspectives in Biology and Medicine* (Johns Hopkins University Press) 30, no. 1 (Autumn 1986): 7-18.

Tanenbaum, Joshua, and Thomas Griffiths. "Structure Learning in Human Causal Induction." *Advances in Neural Information Processing Systems*. NIPS, 2000.

Tausczik, Yla, Aniket Kittur, and Robert Kraut. "Collaborative problem solving: a study of MathOverflow." *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* . Baltimore: ACM, 2014.

Tenopir, Carol, Donald King, Lisa Christian, and Rachel Volentine. "Scholarly article seeking, reading, and use: A continuing evolution from print to electronic in the sciences and social sciences." *Learned Publishing* 28, no. 2 (Apr 2015).

Thies, William. "Language and Compiler Support for Stream Programs." Thesis, Electrical Engineering and C omputer Science, Massachusetts Institute of Technology, Cambridge, 2009.

Thies, William, Michal Karczmarek, and Saman Amarasinghe. "StreamIt: A Language for Streaming Applications." *International Conference on Compiler Construction (CC 2002)*, 08 2002.

Torvalds, Linus, and Junio Hamano. *Git: Fast Version Control*. Feb 12, 2010. http://git-scm.com (accessed Mar 2, 2015).

Tsang, Edward. *High-frequency Finance Research Platform, A Wiki-style Global Project*. 2014. http://www.bracil.net/finance/HFF/platform.html (accessed 10 01, 2014).

Tsang, Edward. "New ways to understand financial markets." Working Paper WP046-10, Centre for Computational Finance and Economic Agents (CCFEA), University of Essex, Colchester, UK, 2010.

Tufte, Edward. *Beatiful Evidence*. Second Printing. Cheshire, CT: Graphics Press, 2006.

Turabian, K. L. *A Manual for Writers of Term Papers, Theses and Dissertations*. 5th ed. Chicago: The University of Chicago Press, 1987.

Udell, Jon. Mar 4, 2002. http://radio-weblogs.com/0100887/2002/03/04.html#a103 (accessed Dec 1, 2015).

University of Essex. *Policy on Thesis Submission, Deposit and Retention*. Policy, PGRE Team, Colchester: University of Essex, 2016.

University of Texas. *Common Mistakes in Using Statistics*. 11 3, 2011. https://www.ma.utexas.edu/users/mks/statmistakes/datasnooping.html (accessed 5 15, 2016).

van Deursen, Arie, and Paul Klint. "Domain-Specific Language Design Requires Feature Descriptions." *Journal of Computing and Information Technology* 1 (10 2002): 1-17.

van Deursen, Arie, Paul Klint, and Joost Visser. "Domain-Specific Languages: An Annotated Bibliography." *ACM SIGPLAN Notices* (ACM) 35, no. 6 (June 2000): 26-36.

Van Noorden, Richard. "Scientists May be Reaching a Peak in Reading Habits." *Nature*, Feb 2014.

Vangheluwe, Hans. "Discrete Event Modelling and Simulation." Article, McGill University, Quebec, 2014.

Vijitbenjaronk, Warut, Jinho Lee, Toyotaro Suzumura, and Gabriel Tanase. "Scalable time-versioning support for property graph databases." *IEEE International Conference on Big Data*. Boston, Massachussets: IEEE, 2017.

Visser, Willemien. *The Cognitive Artifacts of Designing*. 1st. Hillsdale, NJ: CRC Press, 2006.

Von Ronne, Jeffery. "Simulation: Overview and Taxonomy." *Department of Computer Science, Carnegie Mellon University*. Apr 16, 2012. http://www.cs.cmu.edu/~tcortina/15110sp12/Unit12PtB.pdf (accessed May 12, 2017).

Wallis, Jillian, Elizabeth Rolando, and Christine Borgman. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology." *PLoS One*, Jul 2013.

Wallis, Kenneth. "Revisiting Francis Galton's Forecasting Competition." *Statistical Science* (Institute of Mathematical Statistics) 29, no. 3 (2014): 420-424.

Walton, Mary. *The Deming Management Method*. New York: A Perigee Book: The Berkley Publishing Group, A Division of Penguin Group, 1986.

Wang, Chamont. *Sense and Nonsense of Statistical Inference*. New York: Marcel Dekker, 1993.

Weistein, Eric W. *Triangular Number*. Oct 10, 2015. http://mathworld.wolfram.com/TriangularNumber.html (accessed Oct 28, 2015).

Yong, Ed. "How Reliable Are Cancer Studies?" *The Atlantic*. Jan 18, 2017. https://www.theatlantic.com/science/archive/2017/01/what-proportion-of-cancer-studies-are-reliable/513485/ (accessed Jan 15, 2018).

Young, Stanley, and Alan Karr. "Deming, Data and Observational Studies." *Significance* (The Royal Statistical Society), Sep 2011: 116-120.

VITA

The author, Jorge M. Faleiro Jr, started it all in public schools in the beautiful city of Rio de Janeiro, Brazil. After a short tenure as an officer in the military and an honorable discharge as a Lieutenant, Jorge went on to receive an Electronic Engineering degree in Digital Systems and an MSc degree in Systems Engineering, Information Theory, and Informatics from the Instituto Militar de Engenharia (IME) in Rio de Janeiro, Brazil. Jorge holds specialization degrees in advanced financial strategies at the New York Institute of Finance and the Courant Institute of Mathematical Sciences at NYU, as well as a Data Sciences degree at the Johns Hopkins University. Jorge is a former lecturer at the MBA specialization at the NCE/COPPE at UFRJ, the Army War College (ESG), and at the Staff Specialization Center (CEP) of the Defense Ministry. Currently, Jorge lives in the fantastic City of New York, USA where, after holding senior positions in reputable financial institutions, he pursues his doctorate degree at the CCFEA, researching models for crowd-based scientific investigation and the application of deep learning and large-scale data analytics to computational trading. He is also the initiator of several open source projects and volunteers as a mentor in many important initiatives that are close to his heart. Jorge is a curious and avid learner of almost about anything, who believes that good example, hard work, formal education, and science done right will eventually fix everything. On his free time, Jorge enjoys challenging outdoor activities, the company of family and friends, snowboarding, climbing very tall mountains, sailing through steady seas, and scuba diving in exotic places.

Permanent e-mails: j <_at_> falei.ro, jorgemfaleirojr <_at_> gmail.com

Template for the format of this thesis obtained from (Penman 2011);
publishing guidelines from (Campbell 1990), (Turabian 1987);
submission guidelines from (University of Essex 2016)

The author typed this thesis.

A4; double-sided; left and right margin 3.2cm; top and bottom margin 2.5cm