Table of Contents

Copyrighted material

Copyrighted material

List of Illustrations

Copyrighted material

Copyrighted material

Copyrighted material

"Doloc's book is a masterfully written and essential handbook for anyone involved in utilizing data to gain insights into their respective industries. With intellectual honesty, Doloc separates hype from reality, skillfully and intricately weaving a framework to harness the advances and recent developments in quantitative and computational finance. He challenges readers to adopt the best approaches for their applications, knowing the potential but also the limitations, and wisely problem solve. The author may have expertly designed this book for the trading community, but the takeaways are industry agnostic. A must-read for any academic or practitioner in data science, machine learning, and AI fields."

—Rob Friesen, president & COO, Bright Trading, LLC; CEO & Director of Education, StockOdds, Inc.

"Cris Doloc's book is a great introduction to a fascinating field of Computational Intelligence and its applications to quantitative finance. Through examples and case studies covering a wide range of problems arising in quantitative finance from market making to derivative valuation and portfolio management the author demonstrates how to apply complex theoretical frameworks to solving practical problems. Using a sequence of case studies, Doloc shows quantitative researchers and practitioners the power of emerging Computational Intelligence and machine learning technologies to build intelligent solutions for quantitative finance."

—Yuri Burlakov, Ph.D., head of Proprietary Research, Volant Trading

"Cris Doloc has created a valuable guide to Computational Intelligence and the application of these technologies to real-world problems. This book establishes a firm foundation to update the Financial Mathematics program curriculum and practitioners in this domain by presenting a systematic, contemporary development of data-intensive computation applied to financial market trading and investing. Using a sequence of case studies, Doloc shows quantitative researchers and practitioners the power of emerging Computational Intelligence and machine learning technologies to build intelligent solutions for quantitative finance."

—Jeff Blaschak, Ph.D., data scientist and co-founder, Social Media Analytics, Inc.

"Cris Doloc has written a book that is more than just a solid introduction to the current state of the art in AI for quants; it is a solid introduction in how to *think about* AI for quants. In a field that is changing daily, the focus on application of techniques and critical thinking about the strengths and weaknesses of different approaches rather than on details of the latest tools makes time spent with this book a good investment in the future. The case studies in particular help ground the material in the real world of quantitative finance and provide powerful examples of the informed application of AI to finance."

—John Ashley, Ph.D., director of Global Professional Services, Nvidia

Copyrighted material

“Doloc's book masterfully distills the complex world of quantitative trading into a clear guide that's an ideal starting point for new, would-be quants. It provides so many fresh insights into the space that even more seasoned practitioners can learn from it.”

—James L. Koutoulas, Esq., CEO, Typhon Capital Management

“Through a series of case studies, Doloc illustrates a number of examples of real-world problems designed to prepare the reader to work in the contemporary world of quantitative finance. I recommend this book to students of financial engineering and quantitative finance, and to all quantitatively oriented participants in all areas of finance.”

—Ilya Talman, president, Roy Talman & Associates, Inc.

Copyrighted material

Applications of Computational Intelligence in Data-Driven Trading

CRIS DOLOC

WILEY

Copyrighted material

Copyright © 2020 by Cris Doloc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993, or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com/go/permissions.

Library of Congress Cataloging-in-Publication Data is Available:

ISBN 978-1-119-55050-1 (hardback)

ISBN 978-1-119-55052-5 (ePDF)

ISBN 978-1-119-55051-8 (ePub)

Cover Design: Wiley

Cover Images: © metamorworks/iStock.com, © BiksuTong/Shutterstock, © AVIcon/Shutterstock, © Classica2/Shutterstock

Copyrighted material

Dedicated to the memory of my father, Emil

Copyrighted material

About the Author

Cris Doloc holds a PhD in Computational Physics and worked for more than two decades at the intersection of Quantitative and Computational Finance. He is an accomplished technology leader, who designed and led the implementation of several firm-wide trading, valuation, and risk systems. Cris's expertise extends from enterprise software architecture to High Performance Computing and Quantitative Trading.

Cris is currently teaching at the University of Chicago in the Financial Mathematics program, and is the founder of FintelligeX, a technology platform designed to promote quantitatively data-driven education. He is very passionate about the opportunities that recent developments in Cognitive Computing and Computational Intelligence could bring to the field of Quantitative and Computational Finance.

Copyrighted material

Acknowledgments

The metamorphosis of my ideas into the format of a book would not have been possible without the participation and help of many people. Unfortunately, I will not be able to name all of them, but I would like to start by thanking Bill Falloon, my editor at Wiley, who was at the origin of this project. Bill believed in this project from the beginning and helped me tremendously to navigate through the very complex and time-consuming process of writing a book. I would also like to thank Michael Henton, Beula Jaculin and Elisha Benjamin from Wiley for all the editorial help they have provided me with throughout this process.

This project could not have been completed without the constant understanding and support of my beloved wife, Lida, and of my precious daughter, Marie-Louise.

I am extremely grateful to my reviewers for their time and invaluable feedback. I would like to thank Professor Dan Nicolae, the Chair of the Statistics Department at the University of Chicago, Professor Roger Lee, the director of the Financial-Mathematics program at the University of Chicago, and Linda Kreitzman, the Executive Director of the MFE program at UC Berkeley, for their guidance and suggestions throughout the review process. I feel privileged to have had among the reviewers of this book some very influential names from the practitioner's realm like:

- Robert Friesen, the President and COO of Bright Trading and the CEO and Director of Education at StockOdds in Vancouver, Canada.
- Dr. Gerald Hanweck, a pioneer in the field of GPU applications to finance, the CEO and founder of *Hanweck Associates, LLC, New York*, a leading provider of real-time risk analytics for global derivatives markets.
- James Koutoulas, Esq., the CEO of Typhon Capital Management, in Chicago.
- Dr. John Ashley, the Director of Global Professional Services at NVidia.
- Dr. Jeff Blaschak, data scientist and the co-founder of Social Media Analytics.
- Dr. Yuri Burlakov, head of the proprietary research group at Volant Trading, New York.
- Ilya Talman, the president of Roy Talman and Associates in Chicago.

I am profoundly grateful to a large group of people that helped me to grow in my career, both as a physicist, and as a quant-technologist. I am deeply grateful to my high school physics teacher, Constantin Vasile, and to my PhD thesis adviser, Dr. Gilles Martin, for instilling in me the love for physics and problem solving. I am also very thankful to many amazing entrepreneurs and business leaders who entrusted me with important projects throughout my career. I would like to acknowledge many of my colleagues and mentors who helped me to shape my current views on how to apply the latest technology to solve the most important problems at hand.

Finally, I would like to thank my students for helping me to understand the

Copyrighted material

importance of promoting problem-solving skills over content acquisition or tools management. The complexity of modern financial markets demands a continuous assimilation of the newest technology available, and a new breed of quant workforce will have to emerge. The quant of the twenty-first century will have to combine classical quant skills with deep knowledge of computer science and hands-on knowledge of modern HPC technologies. My message to them is this: *There is no magic tool other than our own intelligence! Neither AI, nor any other "intelligence"-containing idiom could be a substitute for human intelligence! Our duty as educators is to kindle your interest in innovating, to nurture your problem-solving skills, and to guide your professional development*. My earnest hope is that this book will be a useful device for reaching this purpose!

Cris Doloc, PhD
August 2019, Chicago

Copyrighted material

About the Website

Additional materials for the book can be found at: https://www.fintelligex.com/book

The website includes the following materials: book's cover and description, table of contents and weblinks to coding resources.

Copyrighted material

Introduction

"Life on earth is filled with many mysteries, but perhaps the most challenging of these is the nature of Intelligence."

–Terrence J. Sejnowski, computational neurobiologist

Two decades of participation in the digital transformation of the trading industry as a system architect, quant, and trader, coupled with the experience of teaching in the Financial-Mathematics program at the University of Chicago, provided me with a unique perspective that I will convey to the reader throughout this book. As both a practitioner and an educator, I wrote this book to assert the fact that the trading industry was, and continues to be, a very fertile ground for the adoption of cutting-edge technologies.

The central message of this book is that the development of *problem-solving* skills is much more important for the career advancement of a quantitative practitioner than the accretion and mastering of an ever-increasing set of new tools that are flooding both the technical literature and the higher education curricula. While the majority of these tools become obsolete soon after their release into the public domain, acquiring an adequate level of problem-solving expertise will endow the learner with a long-lasting *know-how* that will transcend ephemeral paradigms and cultural trends.

If the use of an exhaustive tool set is providing the solution architect with *horizontal* scalability, mastering the expertise of what tools should be used for any given problem will grant the user with the *vertical* scalability that is absolutely necessary for implementing *intelligent* solutions. While the majority of books about the application of *machine intelligence* to practical problem domains are focused on how to use tools and techniques, this book is built around six different types of problems that are relevant for the quantitative trading practitioner. The tools and techniques used to solve these problem types are described here in the context of the case studies presented, and not the other way around.

MOTIVATION

The impetus to write this book was triggered by the desire to introduce to my students the most recent scientific and technological developments related to the use of *computationally intelligent* techniques in quantitative finance. Given the strong interest of my students in topics related to the use of Machine Learning in finance, I decided to write a companion textbook for the course that I teach in the Financial-Mathematics program, titled Case Studies in Computing for Finance.

Soon after I started working on the book, I realized that this project could also

Copyrighted material

benefit a much larger category of readers, the quantitative trading practitioners. An important motivation for writing this book was to create awareness about the promises as well as the formidable challenges that the era of data-driven decision-making and Machine Learning (ML) are bringing forth, and about how these new developments may influence the future of the financial industry. The subject of *Financial Machine Learning* has attracted a lot of interest recently, specifically because it represents one of the most challenging *problem spaces* for the applicability of Machine Learning.

I want to reiterate that the central objective of this book is to promote the primacy of developing problem-solving skills and to recommend solutions for evading the traps of keeping up with the relentless wave of new tools that are flooding the markets. Consequently the main purpose of this book is pedagogical in nature, and it is specifically aimed at defining an adequate level of engineering and scientific clarity when it comes to the usage of the term *artificial intelligence*, especially as it relates to the financial industry.

The term AI has become the mantra of our time, as this label is used more and more frequently as an *intellectual wildcard* by academicians and technologists alike. The AI label is particularly abused by media pundits, domain *analysts*, and venture capitalists. The excessive use of terms like AI disruption or AI revolution is the manifestation of a systemic failure to understand the technical complexity of this topic. The hype surrounding the so-called *artificial intelligence revolution* is nothing but the most noticeable representation of a data point on Gartner's *hype curve of inflated expectations*.

This hype could be explained eventually by a mercantile impulse of using any opportunity to promote products and services that could benefit from the use of the AI label. It is rather common that a certain level of misunderstanding surrounds novel technology concepts when they are leaving the research labs and are crossing into the public domain. The idea that we are living in an era where the emergence of *in silico* intelligence could compete with human intelligence could very well qualify as “*intellectual dishonesty*”, as Professor Michael Jordan from Berkeley said on several occasions. Consequently, one of the main goals of this book is to clarify the terminology and to adjust the expectations of the reader in regard to the use of the term AI in quantitative finance.

Another very important driver behind this book is my own opinion about the necessity of updating the Financial-Mathematics curriculum on two contemporary topics: data-driven decision-making (trading and investing) and Computational Intelligence. As a result, the first half of this book is dedicated to the introduction of two modern topics:

- *Data-driven trading*, as a contemporary trading paradigm and a byproduct of the fourth scientific paradigm of data-intensive computation.
- *Computational Intelligence*, as an umbrella of computational methods that

Copyrighted material

could be successfully applied to the new paradigm of data-driven trading.

The general confusion created by the proliferation of the term AI is at the same time enthralling and frightening. While *mass fascination* comes from the failure to grasp the complexity of applying *machine intelligence* techniques to practical problems, the fear of an AI-world taking over humanity is misleading, distracting, and therefore counterproductive. Whether or not Science will be able any time soon to understand and properly model the concept of *Intelligence*, enrolling both computers and humans into the fight to enhance human life is a major challenge ahead.

While solving the challenge of understanding *general intelligence* will be quintessential to the development of *Artificial Intelligence* it may also represent the foundation of a new branch of engineering. I will venture to call this new discipline *Quantitative and Computational Engineering* (Q&CE). Like many other classic engineering disciplines that have emerged in the past (e.g. Civil, Electrical, or Chemical), this new engineering discipline is going to be built on already mature concepts (i.e. *information*, *data*, *algorithm*, *computing*, and *optimization*). Many people call this new discipline *Data Science*. No matter the label employed, this new field will be focused on leveraging large amounts of data to enhance human life, so its development will require perspectives from a variety of other disciplines: from quantitative sciences like Mathematics and Statistics to Computational, Business, and Social sciences. One of the main goals of writing this book is to acknowledge the advent and to promote the development of this new engineering discipline that I label *Quantitative and Computational Engineering*.

The intended purpose of this book is to be a practical guide for both graduate students and quantitative practitioners alike. If the majority of books and papers published on the topic of *Financial Machine Learning* are structured around the different types and families of tools, I decided to center this book on practical problems, or *Case Studies*. I took on the big challenge to bridge the perceived gap between the academic literature on quantitative finance, which is sometimes seen as *divorced from the practical reality*, and the world of practitioners that is sometimes labeled as being *short on scientific rigor*. As a result I dedicated the second half of the book to the presentation of a set of Case Studies that are contemporarily relevant to the needs of the financial industry and at the same time representative of the problems that practitioners have to deal with. For this purpose I will consider categories of problems such as trade execution optimization, price dynamics forecast, portfolio management, market making, derivatives valuation, risk, and compliance. By reviewing dozens of recently peer-reviewed publications, I selected what I believed to be the most practical, yet scientifically sound studies that could illustrate the current state-of-the-art in Financial Machine Learning. I earnestly hope that this review of recently published information will be useful and engaging for both Financial-Mathematics students as well as practitioners in quantitative finance who have high hopes for the applicability of Machine Learning, or more generally Computational Intelligence

Copyrighted material

techniques in their fields of endeavor.

Last but not least I hope that other industries and sectors of the digital economy could use the financial industry's adoption model to further their business goals in two main directions: automation and innovation. Therefore, another important motivation in writing this book was to share with decision-makers from other industries (e.g. Healthcare and Education) valuable lessons learned by the financial industry during its digital revolution.

The message that I want to convey in this book is one of confidence in the possibilities offered by this new era of *data-intensive computation*. This message is not grounded on the current hype surrounding the latest technologies, but on a deep analysis of their effectiveness and also on my two decades of professional experience as a technologist, quant, and academic. Throughout my career I was driven by the passion to adopt cutting-edge technologies for as long as they could be useful in solving real-world problems. I wanted to convey this philosophy to my students as well as to the readers of this book. This book is an attempt to introduce the reader to the great potential offered by the new paradigm of Data-Intensive Computing, or to what is called the fourth paradigm of scientific discovery to a variety of industries. Throughout this book I am going to promote the concept of Computational Intelligence as an umbrella of new technologies aimed at augmenting human performance (through automation) and engendering intelligence (via innovation and discovery) with examples from the emerging field of data-driven trading. The use of computer systems to analyze and interpret data, coupled with the profound desire to learn from them and to reason without constant human involvement, is what Computational Intelligence is all about. As a means to convey the message I chose to introduce the reader to the realm of Computational Intelligence by presenting a series of Case Studies that are actionable and relevant in today's markets, as well as modern in their data-driven approach.

TARGET AUDIENCE

This book is primarily intended for students and graduate students who contemplate becoming practitioners in the field of Financial Machine Learning and Computational Intelligence as well as for more-seasoned trading practitioners who are interested in the new paradigm of data-driven trading by using *machine intelligence* methodologies.

Another possible target audience is represented by technologists and decision-makers from other sectors of the economy that currently undergo structural digital transformations and could have a major societal impact, like Education and Healthcare. This very large potential audience could learn extremely useful lessons

Copyrighted material

from the digital revolution that shaped the financial industry in the last 10 to 15 years and could apply similar approaches for the successful early adoption of the newest technology available.

As mentioned before, the main goal of this book is to promote and advocate for the use of Computational Intelligence framework in the field of data-driven trading.

Since this is a quite novel and technically advanced topic, I choose to embed this message into a more readable narrative, one that will not exclude readers who may not be very fluent in the language of quantitative and computational sciences. By embedding the main message into a more readable narrative, I hope it will make it more appealing to nontechnical people.

BOOK STRUCTURE

The first part of the book is dedicated to introducing the two main topics of the book: Data-Driven Decision-Making and Computational Intelligence. As such:

- Chapter 1 describes the historical evolution of trading paradigms and the impact that technological progress had on them. A good portion of this chapter is spent on describing the new paradigm of data-driven trading.
- Chapter 2 introduces the reader to the role that data is playing in trading and investing, especially in light of the new data-driven paradigm. This chapter will guide the reader through a fascinating journey from Data to Intelligence.
- Chapter 3 endeavors to *de-noise* the AI hype by introducing an adequate level of scientific clarity for the usage of the term *Artificial Intelligence*, especially as it relates to the financial industry.
- Chapter 4 introduces the framework of Computational Intelligence, as a more realistic and practical framework compared to the AI narrative. Novel approaches to the *solvability* problem are presented and the Probably Approximately Correct framework is introduced.
- Chapter 5 exemplifies the use of Computational Intelligence in Quantitative Finance. It starts with assessing the viability of this methodology in the context of financial data and it presents a brief introduction to Reinforcement Learning as one of the most promising methods used in the next chapters on case studies.

The second part of the book introduces the reader to a series of Case Studies that are representative of the needs of today's financial industry. All the Case Studies presented are structured as follows: an introduction to the problem, a brief presentation on the state-of-the-art in that specific area, a description of the implementation methodology employed, and a presentation of empirical results and conclusions.

- Chapter 6, Case Study 1: *Optimizing trade execution*. This chapter gives a

Copyrighted material

short introduction to the Market Microstructure topic, specifically as it relates to Limit Order Book dynamics in a high-frequency trading context, and then it describes a series of methods for optimizing the Market *impact* problem.

- Chapter 7, Case Study 2 – *Price dynamics forecast*. Several practical examples that use Reinforcement Learning and a variety of Deep Neural Networks are presented.
- Chapter 8, Case Study 3 – *Portfolio management*. This chapter compares the more traditional methods for portfolio construction and optimization with the more modern approaches like Reinforcement Learning and Deep Learning.
- Chapter 9, Case Study 4 – *Market making*. Reinforcement Learning and Recurrent Neural Network algorithms are applied to the problem of liquidity provisioning and several practical examples are presented.
- Chapter 10, Case Study 5 – *Valuation of derivatives*. This chapter introduces the reader to a fascinating new set of applications of ML. Well-established valuation models like Black-Scholes are becoming outdated by the use of Deep Neural Networks and Reinforcement Learning.
- Chapter 11, Case Study 6 – *Financial risk management*. This last chapter dedicated to Case Studies exemplifies understanding and controlling credit, market, operational, and regulatory risk with the help of ML techniques.

The book concludes with Chapter 12, a summary of the three main goals of this book, namely to:

- Describe the new paradigm of Data-Driven Trading and the application of Computational Intelligence techniques to implement it.
- Present from both a scientific and an engineering perspective a critical opinion on the use of the term *Artificial Intelligence* attempting to *de-noise* it.
- Draw the blueprint of a new engineering discipline that in my opinion will be absolutely quintessential to furthering the progress of Computational Intelligence and its applications in Finance and other sectors of the digital economy.

Copyrighted material

CHAPTER 1
The Evolution of Trading Paradigms

"You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete."

– Buckminster Fuller, inventor, system theorist

1.1 INFRASTRUCTURE-RELATED PARADIGMS IN TRADING

Since the beginning of human civilization, commerce has been the main engine of progress. The Cambridge dictionary defines commerce as "the business of buying and selling products and services." The exchange of valuables has been the main driver of progress in any type of economy throughout history, and it was primarily accomplished through trading. The mechanism of trading is considered to have been the main instrument that linked different peoples and acted as the main channel of communication for cultural and intellectual exchange. The primal forms of trade appeared when prehistoric peoples started exchanging valuables for food, shelter, and clothing. The concept of exchange for sustenance became a reality in a physical space known as the *marketplace*. The concept of a *marketplace* as an area designated for the exchange of goods or services became associated with a set of rules to operate within it. As human civilization progressed and the sophistication of trading practices advanced, the need for more modern avenues to trade have become prevalent, and the world of financial instruments was created. Pioneering markets, like the Dojima Rice Market or the Amsterdam Stock Exchange, were the early promoters of modern trading, transacting products such as equities, futures contracts, and debt instruments.

The long history of trading (Spicer 2015) as the main vehicle to exchange valuables and information could be studied by considering the evolution of different trading paradigms. Since a paradigm is a conceptual representation for looking at, classifying, and organizing a specific human endeavor, one can look at trading paradigms from two different perspectives:

- The *infrastructure* required to establish a marketplace, and
- The *methods* required to support and generate trading decisions.

Since the dawn of the financial markets, trading was strongly associated with the technological progress of the time, by heavily employing the most recent breakthroughs. This section is meant to be a very brief history of the *love affair* between trading and technology.

1.1.1 Open Outcry Trading

In its earliest manifestation, trading took place in a setting called the *open outcry* system. This mechanism of transacting involved the matching of buyers and sellers

Copyrighted material

characteristic to successful traders and therefore has an important survivorship bias. The third attribute, the ability to rapidly adapt to changing market conditions, is what one calls nowadays data-driven trading.

As a relatively new field of research and as a new paradigm, data-driven trading draws inspiration from a vast repository of trading knowledge as well as from the multidisciplinary field of Data Science (see Figure 1.2). The *wisdom* of markets claims that past performance is not indicative of future performance – or in other words, one should not care much about past performance, but instead try to forecast how well a strategy is going to perform in the future. On the other hand, Data Science teaches that historical data is the only vehicle that one could utilize to train the learning models on. Equally important is the fact that not all aspects of past data are likely to occur in the future.

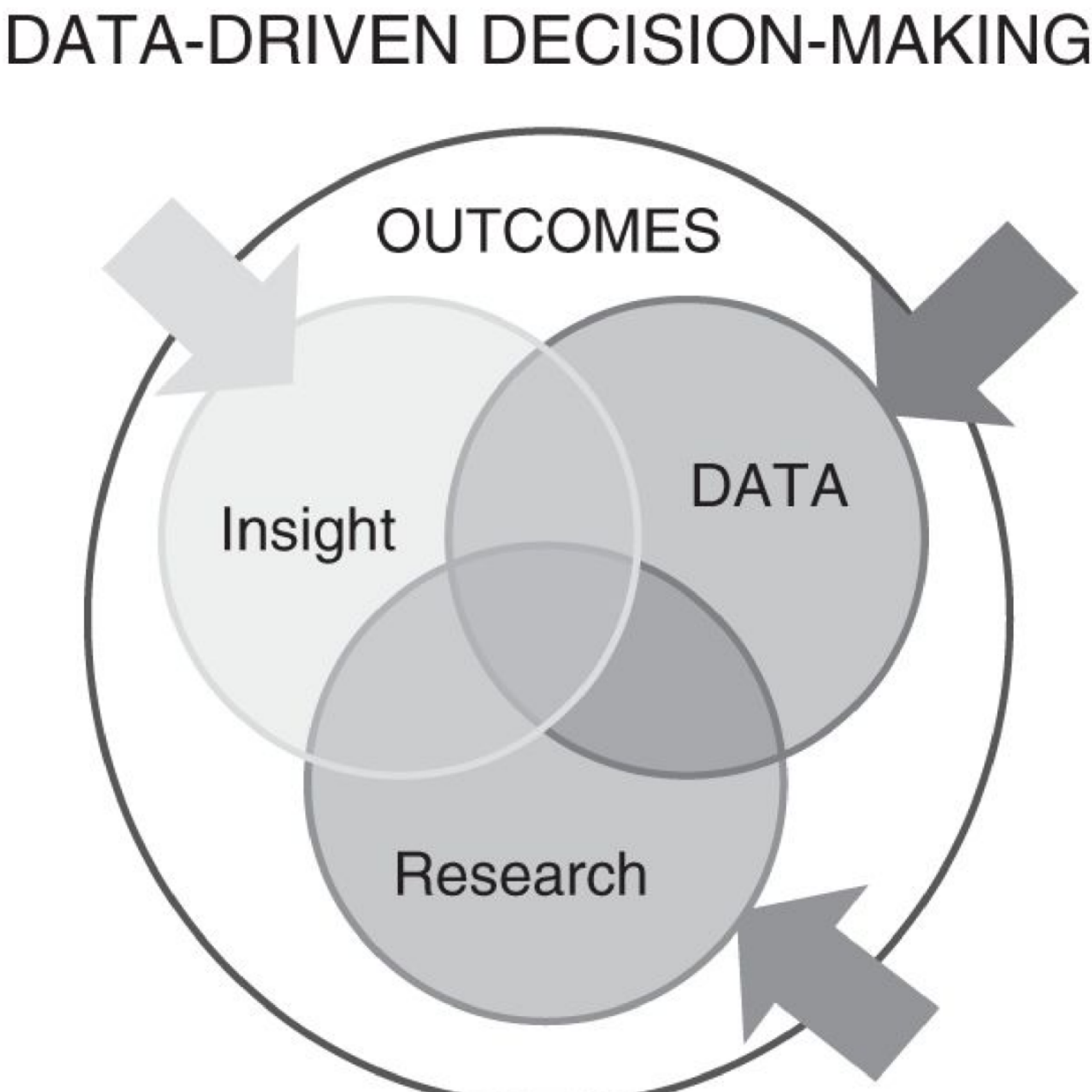

FIGURE 1.2 The data-driven trading concept.

Adapting to current market conditions is not a new endeavor. What is really new is

Copyrighted material

Unfortunately there was a general lack of consensus on how to define the process of Learning. For a long period of time Learning was not regarded as a field that would qualify as a scientific discipline. Until relatively recently Learning was considered as an extension of Education Psychology and other related non-quantitative disciplines. In spite of the fact that Learning is a very reproducible process, a rigorous quantitative modeling is still not available.

Last decade has seen a resurgence of interest in modeling the process of Learning. Pioneering work in this area was published in 2012 by Koedinger and collaborators at Carnegie Mellon (Koedinger, Corbett, and Perfetti 2012). They introduced the *Knowledge-Learning-Instruction* framework that was used to identify a broad range of learning events and influencing factors such as memory, induction, understanding, and sense-making processes. Other researchers believe that Learning must be modeled from both a statistical and a computational perspective. Since Learning is a central component to Computational Intelligence, this section is dedicated to understanding the relationship between Learning, Evolution, and Intelligence.

Dr. Leslie Valiant is one of the most prestigious computational theorists and computer scientists of our time. Professor Valiant was the recipient of the Turing award in 2010 for his "transformative contributions to the theory of computation, including the theory of probably approximately correct (PAC) learning, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing." He is the T. Jefferson Coolidge Professor of Computer Science and Applied Mathematics at Harvard and also the author of the *Probably Approximately Correct* (PAC) learning model (Valiant 1984, 2013) that I am going to use in this section as the framework upon which to build the case for the use of CI in Quantitative Finance. Professor Valiant is well known for avoiding the use of the term AI. At the beginning of his scientific career, while talking to the famous Edsger Dijkstra (one of the most influential computer scientists of the twentieth century, inventor of Dijkstra's algorithm), he was asked about the subject of research that he worked on at that time. After proudly responding AI, Dijkstra said: "Why don't you work first on the 'Intelligence' part?" That was a WOW moment for Dr. Valiant that prompted him to dedicate most of his scientific career to studying the mechanisms of Learning and Intelligence.

This section will introduce the PAC framework, which is a revolutionary way of studying and emulating Intelligence. The PAC theory was introduced by Leslie Valiant in 1984 and for this contribution he was given the Turing award in 2010. Prof. Valiant is one of the pioneers of formalizing the fundamental equivalence between the capabilities of brains and computers. Several decades of research in this field allowed him to come up with the *Probably Approximately Correct* model, which defines mathematically the conditions under which a mechanistic system could be said to learn information.

4.2.1 The Probably Approximately Correct Framework

Algorithms are defined in classical computing as step-by-step instructions needed to achieve an expected outcome, similar to recipes in cooking. The designer of the recipe has full knowledge and is in full control of the environment utilized for

Copyrighted material

achieving the desired goal.

PAC theory introduced a novel algorithmic concept called the ecorithm. The ecorithms are a special category of algorithms. Unlike classical algorithms, they run in environments that are initially unknown to the designer. The ecorithms learn new information that was not available at the design time by interacting with the environment without being programmed to do so. After sufficient interaction with the unknown environment, the ecorithms will gain new knowledge that was not provided by the designer but extracted from the environment instead.

The Probably Approximately Correct model provides a mathematical framework by which algorithmic designers could evaluate the expertise achieved through the Learning process (the Representation step) and devise the Cost function associated with it (the Estimation step). The performance of an ecorithm is evaluated against input information collected from a rather uncontrolled and unpredictable environment and its goal is to perform well enough to ensure survivability.

Another novelty introduced by the PAC theory is that ecorithms are not merely characteristics of computers. Dr. Valiant generalizes the ecorithms as computational concepts that could be used to explore fundamental mysteries related to the evolution of life on Earth. Evolutionary Biology explains how evolution was shaped by living organisms interacting with and adapting to their environments. According to PAC the combination of bio-inheritance and the new knowledge acquired through learning from the environment are the major factors on determining the dynamics of a system or bio entity. The PAC theory suggests a unified way of studying the mechanisms of learning, evolution, and intelligence using computer science methods.

If algorithms are currently implemented in *in silico* systems, ecorithms could be applied to a much broader category of systems, from simple organisms to entire biological species. The PAC framework illustrates a computational equivalence between the way that individuals learn and the way that entire biological systems could evolve. For both cases, ecorithms are describing adaptive behavior in a mechanistic way. PAC's declared goal is to find “mathematical definitions of learning and evolution which can address all ways in which information can get into systems.” (Valiant 1984, 2013) A possible outcome will result in integrating life and computer sciences in novel ways never attempted before. The notions of Learning and Intelligence could then be expanded to include non-biological entities.

Chapter 2 illustrates in great detail the transformative journey of data from Information into Knowledge and eventually Intelligence. The PAC theory defines as theoryful the mathematical rules for predicting the process of transforming Information into Knowledge. Everything else is termed as theoryless. Theoryless processes, including evolution in biological systems or decision-making in cognitive systems, are considered as innovative applications of the ecorithm concept.

The computational features of the Learning process as modeled by the PAC theory have the following important properties:

- The learning process should take place in a *relatively limited number of steps*

Copyrighted material

(in polynomial time).

- The *number of interactions* with the environment from which the entity is learning should be *limited*.
- The *probability of making errors* in applying the knowledge acquired by learning should be *sufficiently small*.

One of the main assumptions of AI has been that one could eventually emulate in software the computations that our brains are performing by identifying their algorithms. This claim asserts that Artificial Intelligence and General Intelligence are practically one and the same. The biggest failure of AI so far lies in its inability to precisely determine what these computations should be and what the algorithms responsible are. Fortunately Machine Learning has been proven to be a quite effective mechanism of bypassing this deadlock. One of the biggest challenges for achieving the AI dream is the ability to implement computations that model evolutionary behavior.

A typical example of this problem is the toddler learning to walk problem. What is the process by which a small child is learning to walk by crawling, touching, and sensing the surrounding environment? This is a process of learning that involves acquiring knowledge that is not described in a *user manual*, could not be coded in the classical sense of hard computing, and is very domain and individual specific.

Although this problem is clearly not of a hard computing type, a certain level of computational activity is performed by the learner while the Learning process is unfolding. Until very recently the general assumption was that Learning could take place exclusively in biological systems. The novelty of the PAC theory is that Intelligence is made up of tangible, mechanical, and ultimately understandable processes. According to Professor Valiant, "We will understand the intelligence we put into machines in the same way we understand the physics of explosives – that is, well enough to be able to render their behavior predictable enough that in general they don't cause unintended damage." (Valiant 2013)

4.2.2 Why AI Is a Very Lofty Goal to Achieve

The central idea of PAC's theory is that the successful Learning of any concept of unknown nature should involve the determination of a high-probability hypothesis that represents a good approximation of it. This will assert that most decisions, either conscious or evolutionary, could be represented in terms of PAC learning.

The field of Machine Learning has convincingly demonstrated that the notion of Learning is central to Intelligence. Unfortunately the Learning algorithms responsible for human intelligence are yet to be identified. Although the current ML algorithms are detecting regularities or patterns that are learned from data, trying to understand and emulate Intelligence, especially human Intelligence, requires a lot more than that. The mere fact that one could detect regularities in data does not make a problem simpler to solve. It has been experimentally proven that even the most complex and theoryless data could exhibit predictable patterns. Functional MRI analysis has shown that predictable patterns of blood flow in the brain could be detected when the experimental subject is reading a text. This problem pertains undoubtedly to the theoryless realm since the understanding of

Copyrighted material

how knowledge is represented in the brain is almost nonexistent.

How is computer science currently dealing with problems when using the hard computing paradigm? For any given problem, a computer:

- Could be programmed to solve it via a predefined algorithm, or
- Could be instructed to learn how to solve it by giving it access to lots of data, or
- Could use a combination of these two approaches.

Because the process of learning is statistical in nature, it cannot be made error-free. If developing a flawless program is an achievable goal, the learning alternative will always be exposed to the risk of not being sufficiently accurate. For problems where the desired outcome could be explicitly specified, programming would be the best solution, assuming that one is able to do so. For a variety of reasons there are situations where one may not be able to program a solution. As a consequence, the learning solution becomes vital whenever one cannot specify explicitly the outcome or one cannot get direct programming access to the system. When the learner is a human agent, all these conditions may apply, and there is no alternative to learning. When the learner is a computer system, some or all of these conditions could be present and then the learning solution is the only possibility.

The inability to explicitly specify outcomes is the most common use of general-purpose ML applications. E-mail spam detection is a typical example. As new sources of spam are rapidly developing, the task of manually incorporating ways of detecting it into e-mail systems would be prohibitive. Instead ML algorithms learn specific patterns from e-mail data that enable them to distinguish between e-mails that users label as spam from e-mail that they do not.

The success of Machine Learning is due in large part to the effectiveness of several learning algorithms, such as boosting or ensemble learners. One of the most remarkable innovations in ML is the boosting methodology, which is used currently as a generic technique for improving the performance of almost any basic learning algorithm. The building block is represented by a weak learner that is using a learning style in which the hypothesis employed predicts just marginally better than random guessing. Then the Boosting algorithm will translate the weak learning algorithm into a strong learning algorithm. The idea is to use the weak learning method several times to get a succession of hypotheses and keep the focus on the examples that previous hypotheses found difficult to classify. Because weak learning works for any distribution, modifying the distribution at each stage will enable the learner to achieve better results by this repeated refocusing. Boosting has proved to be a very robust method for improving the predictive accuracy of a wide variety of simple learning methods.

Besides the choice of what learning algorithm to use, the selection of features to represent the problem is another important aspect in Machine Learning. It was empirically proven that good choices yield to more accurate predictions. But how can one gauge what a better choice of features is? The process of feature engineering is computational in nature but not quite well understood. Biological systems for example are using *high-level* features. These features are acquired

Copyrighted material

most likely through the evolution process and were passed from generation to generation in a genetically encoded fashion. Learning these high level features from scratch individually every time one needs to use them would make the evolution a lot less efficient. ML has benefited greatly not only from the development of better algorithms and from the access to large volumes of data but also from the development of hardware accelerators like GPUs or FPGAs. The success of ML on a broad variety of problems is powerful evidence for the effectiveness of learning in areas related to human information processing.

Undoubtedly Machine Learning is the most successful branch of AI to date. One of the most sobering questions about the use of this methodology is how can one predict for which problems one expects ML to succeed, and for which to fail. This question is especially important in Quantitative Finance. A strong requirement would be that the distribution on which the system needs to perform well must be identifiable. The distribution does not need to be described explicitly, but just unambiguously. This means that one should have the assurance that algorithms trained on labeled data sets of typical examples will perform at least as well when used out in the wild on unseen data of the *same kind*. But even when this data distribution consistency is present, the ML approaches may fail because the patterns that one searches are either inherently hard to learn or because the information in the data set is not sufficient for the task at hand.

What really makes AI so difficult to achieve?

The short answer is the inability of machines to handle common sense knowledge in a way that is similar to humans. According to Professor Valiant, the human cognitive system is the outcome of a very long period of evolution coupled with a lifetime of learnable target pursuit since birth. Alan Turing's dream was to educate a computer as one would educate a child. That would entail endowing the computer with human-like cognitive capabilities. But because human cognitive abilities are the result of complex evolutionary ecorithms, and because of our very limited understanding of how our brain is hardwired, this process is absolutely theoryless. There have been attempts to describe the algorithms of evolution, but the experimental results (the data) are no doubt, theoryless.

Let's suppose that one may have access to a powerful super-theory that will allow the computation of the atomic features of the human nervous system at birth. This encoded information could be used to educate a computer in a similar way humans are educating their offspring. But what to do about the knowledge encoded in the human DNA by the evolutionary processes? The only alternative will be to start from the beginning of life on Earth and simulate all evolution stages. Quite an unfeasible task, not just computationally but also because the inputs and parameters that accompanied evolution are just impossible to determine. According to Professor Valiant, the most important barrier to meaningful advances in AI is related to the lack of understanding of how humans acquire knowledge through learning after birth. There are several means of communication through language, vision, smell, taste, and touch. But encoding this knowledge into a computer-ready format is beyond our current capabilities. One empirical observation is that the more common sense the knowledge is, the more difficult is to encode it into a computer-ready format. And this is just a reflection of our

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material