Escaping The Sisyphean Trap:
How Quants Can Achieve Their Full Potential

Marcos López de Prado

ORIE 5256

Summary

- Investing can be characterized as a set of data science problems
  - Some of the most challenging problems in data science involve financial time series
  - Traditional statistics (econometrics) is ill-equipped to tackle the complexity of financial markets
  - This opens the door to a new breed of quantitative researchers
- While firms have attracted STEM talent, they have done a poor job at developing it
  - Firms hire specialists, but entice them to become generalists (e.g., portfolio managers)
  - Under the ubiquitous silo/platform structure, quants succumb to the Sisyphean trap
- A research lab structure offers a unique environment for developing scientists, by means of
  - co-specialization, working in a highly cooperative environment
  - tackling well-defined open investment problems
  - applying the scientific method

The Evolution of Investment Credentials

Towards Evidence-based Investing

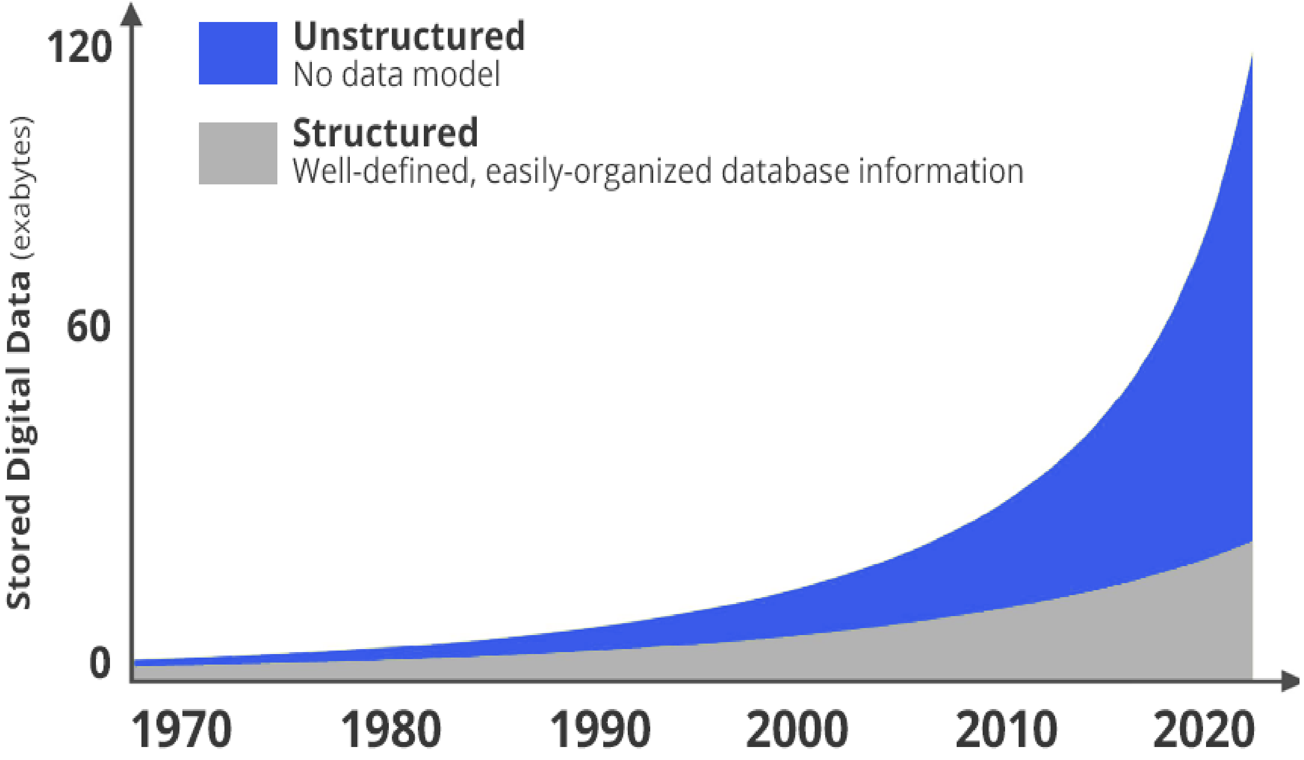
		1950	2000	2015
Conditions	Investing is largely driven by speculation, due to • no financial theories • no computers • simplistic math models • limited data	Academics formalize the concepts of diversification, risk-premia and valuation • MPT, CAPM, APT, risk factors, Black-Scholes, market microstructure	Technological advances in • data storage • supercomputing • networking Math models become more sophisticated	The explosion in alternative data shifts research objectives, from valuation and forecasting, towards nowcasting (direct estimation)
	Technical analysis	Fundamental analysis, econometrics	Market microstructure, HFT	Machine learning, causal inference
Rationale	Investment decisions are motivated by stories or themes	Investment decisions rely on financial analysis of specific opportunities	Investment decisions increasingly rely on research and technology. Some of the best performing funds are quant-driven	
Training	Investment professionals do not receive a formal education	During this period, CFA is the gold standard of financial accreditation	Firms hire STEM grads for research and software engineering roles	High demand for data scientists and automation experts

What Has Driven This Evolution?

- In search for opportunities, investors analyze **data**
  - In the past, data was limited
  - Today, data is plentiful, however complex
- **Machine learning** (ML) methods allow us to extract insight from complex data, at the cost of heavy calculations
- Technical advances in **automation** and **high-performance computing** (HPC) have led to productivity gains
- Corollary: Scientific backgrounds are in high demand

The amount of data has exploded in just a few decades. To put it in perspective, digital systems store today $44 \cdot 10^{21}$ bytes of data. In comparison, there are only $7.5 \cdot 10^{18}$ grains of sand in the entire world. Apollo 11's guidance computer operated with $32 \cdot 10^{3}$ bytes, the equivalent to $\sim$5 grains of sand.

Data Size & Data Complexity



According to IDC, about 90% of all data recorded throughout history has only been collected over the past 2 years.

About 80% of that data is unstructured, requiring the use of ML and HPC methods. Much of this data has investment applications.

Finance today offers some of the most challenging problems for data scientists.

What Makes Finance an Exceptional Field of Research

Curse #1: Barriers to Experimentation

- Association does not imply causation
  - X is associated with Y iff $P[Y,X] \neq P[X]P[Y]$
  - X causes Y iff $P\big[Y|do[X]\big] > P[Y]$
- Scientists propose causal theories to explain phenomena (observed associations)
  - A theory that cannot be falsified empirically, is not scientific
    - Note: Not scientific does not mean false
- In the natural sciences, causal theories are falsified through randomized controlled experiments
- Controlled financial experiments are rarely possible
  - E.g., we cannot repeat the Flash Crash of 2010 controlling for some market participants
  - Researchers engage in uncontrolled multiple testing
  - Causal inference offers an alternative

$$PS_1 \rightarrow TT_1 \rightarrow EE_1 \rightarrow PS_2$$

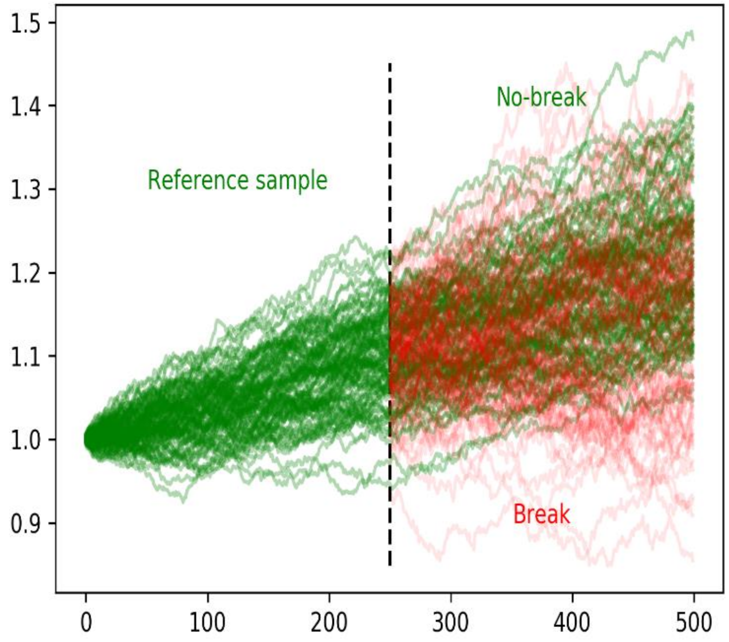
Popper's formula for explaining scientific progress

In response to a problem situation ($\boldsymbol{PS_1}$, e.g., an observed anomaly), researchers propose competing tentative theories ($\boldsymbol{TT_1}$), which are systematically subjected to falsification attempts. The error elimination ($\boldsymbol{EE_1}$) process purges false theories.

The simplest surviving theory is not truer, however it is better "fit" (in an evolutionary sense) to tackle more difficult problems ($\boldsymbol{PS_2}$) posed by the theory ($\boldsymbol{TT_1}$).

Curse #2: Non-Stationarity

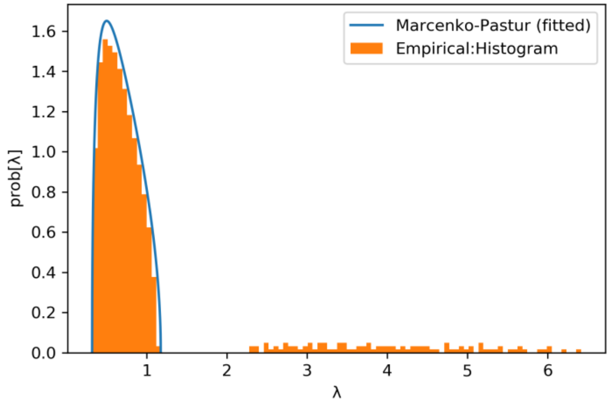
- Time series analysis relies on two key assumptions
  - **Ergodicity**: Observations span the entire support for the stochastic process
  - **Stationarity**: The unconditional joint probability distribution is time invariant
- Ergodicity can be recovered by combining non-ergodic processes
  - E.g., investing in diversified indices, rather than in an individual company that may go bankrupt
- Financial systems are non-stationary due to
  - structural breaks (e.g., changes in regulations)
  - drift in parameters that regulate the data-generating process (e.g., competition)

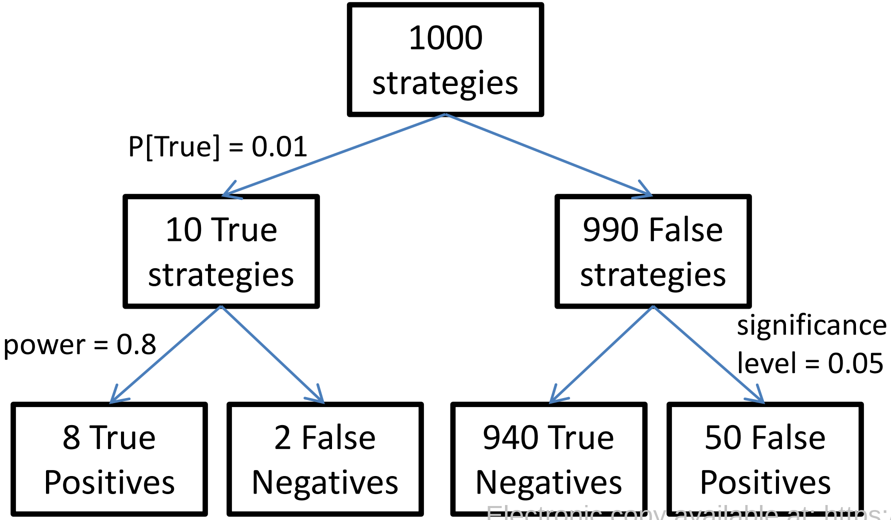


After 250 observations, a structural break takes place with probability 0.5. The problem is to detect the break as soon as possible, with high precision and high recall.

Curse #3: Strict Competition (Zero-Sum Game)

Strict competition among investment managers lowers the signal-to-noise ratio. For example, almost all the information contained in a factor covariance matrix (e.g., Barra) can be associated with noise rather than signal. As a result, the probability of finding a profitable investment strategy is extremely low, and alpha decays quickly.





Suppose that the probability of a backtested strategy being profitable is 1%.
Then, at the standard thresholds of 5% significance and (optimistic) 80% power, researchers are expected to make 58 discoveries out of 1000 trials, where 8 are true positives and 50 are false positives.
Under these circumstances, **a *p*-value of 5% implies that at least 86% of the discoveries are false!**

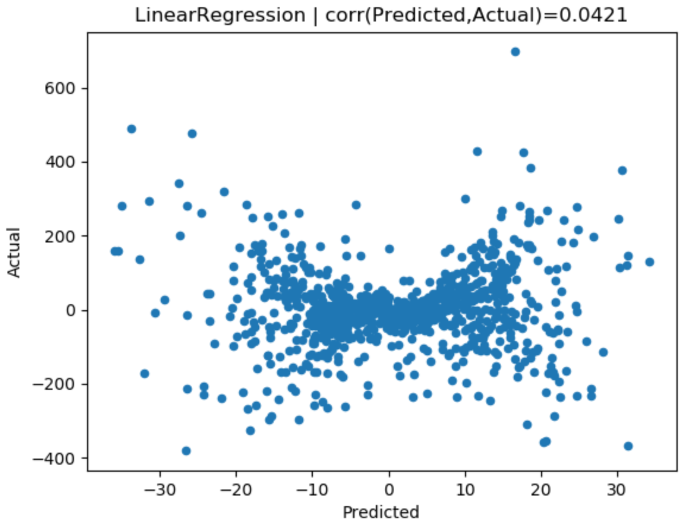
Curse #4: Systemic Complexity

- The Econometric canon was developed to model relatively straightforward processes

$$y_t = \alpha + \sum_{i=1}^{I} \beta_i X_{t,i} + \sum_{j=1}^{J} \gamma_j Z_{t,j} + \varepsilon_t$$

where ε_t is a (very well-behaved) random variable

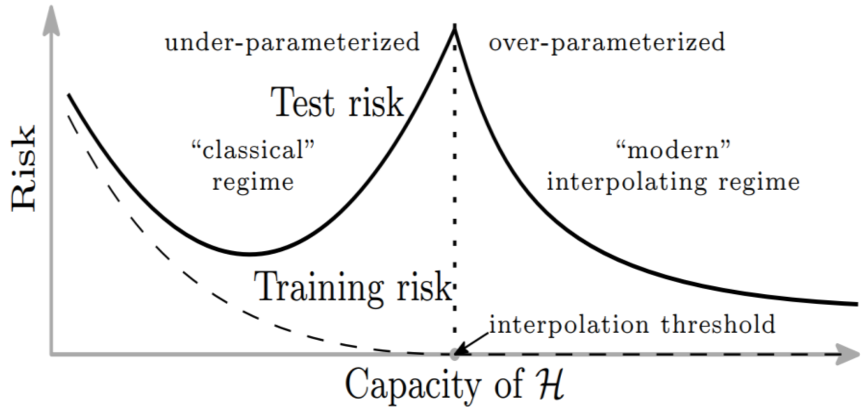
- This presumes that the researcher knows
  a) the predictive variables
  b) the functional form, including all interaction effects
- With the wrong (b),
  – we may reject the true (a) → false negative
  – we may accept the false (a) → false positive
- It is unreasonable to assume knowledge of (b)



In this example, the researcher has captured all the variables involved in a phenomenon, and used the correct specification, however he missed a single interaction effect. The result is that the researcher will incorrectly reject the true variables (a false negative). ML can help us decouple the variable search from the specification search.

Curse #5: Small Samples

- Financial datasets often exhibit
  - Short time series, few variables
  - Highly cross-dependent variables
  - Highly serially-dependent variables
- Financial researchers often must produce inference from small samples
  - Strong train-set overfitting
  - Strong test-set overfitting
- At standard confidence levels, statistical tests on financial datasets have low power
  - High proportion of false negatives
  - High false discovery rate



Belkin et al. [2019] propose the *possibility* of a "double descent" by fitting a model with thousands (possibly millions) of random Fourier features (RFF), in an attempt to achieve a lower mean squared error out-of-sample. Even if successful, the resulting model is a black-box: uninterpretable, and vulnerable to regime shifts, structural breaks and black-swans.

While massive overfitting may have commercial applications (e.g., marketing), it has limited value for institutional investors, due to scientific and legal reasons (e.g., fiduciary duty).

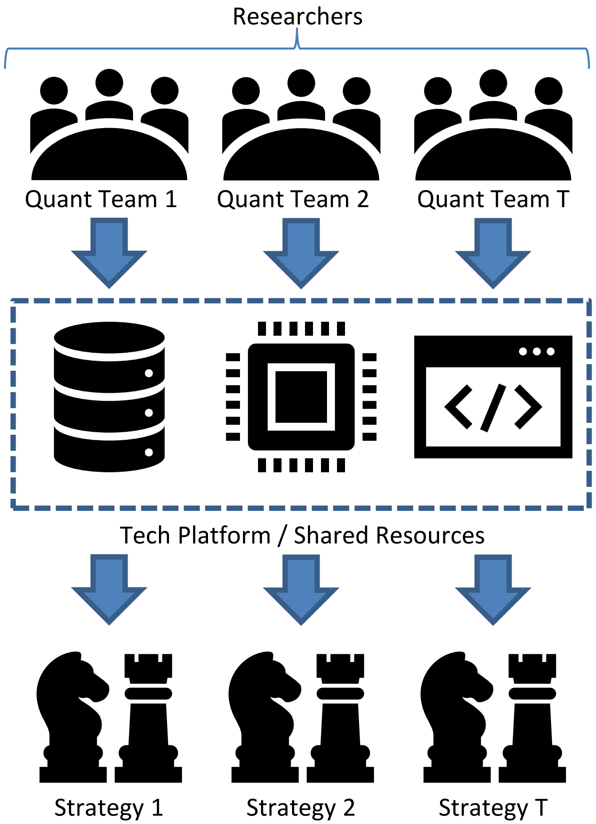
The Five Curses of Financial Research

Curse	Effect
Barriers to experimentation	• Cause-effect mechanisms cannot be easily isolated • Unfalsifiable (non-refutable) claims
Non-stationarity	• Predictions are unreliable • Risk of unforeseeable structural breaks, black swan events
Strict competition (zero-sum game)	• The publication of a discovery impacts its validity going forward • Low signal-to-noise ratio • Low probability of a positive, high false discovery rate • Limited capacity, alpha decay
Systemic complexity	• Basic forecasts may require substantial computational resources • Unknowable, hierarchical, high-dimensional, or possibly non-algebraic specification, with numerous interaction effects • Missing a single interaction effect may lead to false conclusions
Small samples	• Inference with low power • High risk of overfitting (train-set *and* test-set)

Overcoming the Five Curses that Afflict Finance

Silos/Platforms Curtail STEM Potential

- Traditional investment firms hire (small) self-sufficient teams, who will *not* collaborate
  - In fact, the teams compete with each other for allocations
- The firm's objective is to achieve diversification through competition
- However, this paradigm suffers from
  - Limited depth (teams of generalists)
  - Limited scalability (one team per strategy)
  - Incentivizes macroscopic alpha (due to limited resources)
  - Incentivizes false positives (due to competition)
  - No reusable knowledge (IP is owned by team)
  - No self-correcting mechanisms (due to lack of collaboration)
- Not best suited for scientific talent



The Sisyphean Trap

- Silo structures do not provide the co-specialization needed to overcome Finance's 5 curses
- Without division of labor, researchers cannot apply the checks and balances embedded in the scientific method
  - Peer review
  - Empirical falsification
- Silo structures make business sense, however they
  - transform specialists into generalists
  - do not offer the best value for investors
- Like Sisyphus, Silo quants end up performing recurrent futile tasks (e.g., backtest overfitting), that result in an endless cycle of false discoveries

In Greek mythology, the gods punished Sisyphus with an endless and futile task. Similarly, Silos stack the odds against quants, by asking them to overcome the 5 curses of Finance without a proper (scientific) setup.

Big Science

"No individual is alone responsible for a single stepping-stone along the path of progress... The day when the scientist, no matter how devoted, may make significant progress alone and without material help is past."

Ernest O. Lawrence (1940)

National Labs Work Like Assembly Lines

	Silos / Platforms	Assembly Line
Independent validation	X	✓
Reusable knowledge	X	✓
Microscopic alpha	X	✓
Scalability	X	✓

Research Assembly Lines strike a balance between creativity and planification, fostering breakthroughs at a predictable rate.

Prof. Ernest Lawrence founded Berkeley Lab in 1931 as a research factory, where interdisciplinary teams of scientists would tackle problems that universities could not.

Prof. Lawrence's idea led to the Manhattan Project, Big Science, the current network of DOE National Laboratories in the U.S., and a record 115+ Nobel Prizes (**~27% of all Physics + Chemistry awards**).

"Scientific progress requires a favorable environment."

Ernest O. Lawrence
February 29, 1940
Nobel prize acceptance speech

Structure of an Investment Assembly Line

Function	RESEARCH			TESTING	PRODUCTION		
Role	Data Curation	Feature Analysis	Strategy Development	Testing	Deployment	Portfolio Oversight	Portfolio Implementation
Question	Can we structure this dataset as (X, y)?	Has X predicted y in the past?	Why should X predict y going forward?	What's the false positive probability?	What's the most efficient deployment of the prototype?	Does the strategy perform as expected?	How can we achieve the target positions with minimal shortfall?
Popperian Cycle	Problem Situation (PS_n)		Tentative Theory (TT_n)	Error Elimination (EE_n)			

Strategy metadata ($n \leftarrow n + 1$) (meta-labeling)

For a description of the assembly line stations, read https://ssrn.com/abstract_id=3104847

Assembly Lines Develop STEM Talent

- In the Assembly Line paradigm, individuals
  - master a technical skill, not a strategy or asset class
  - join an interdisciplinary squad (co-specialization)
- Assembly Lines achieve competitive advantage by
  - solving open investment problems (reusable knowledge)
  - high throughput of diversified microscopic alpha
  - constant strategy renewal (flow)
- Scientific protocols ensure a clear separation between research and peer-review
  - No backtest overfitting: the testing team independently (a) validates the hypothesis, and (b) backtests the strategy
- Flat structure, with numerous checks and balances
  - There is no Portfolio Manager role
  - Self-correcting mechanisms at each station

The **Research team** formulates a strategy based on a falsifiable investment hypothesis

The **Testing team** independently backtests the strategy, *and* tests the validity of the economic rationale proposed by Research

The **Production team** produces code, calculates strategy allocations, monitors performance, and runs daily portfolio operations

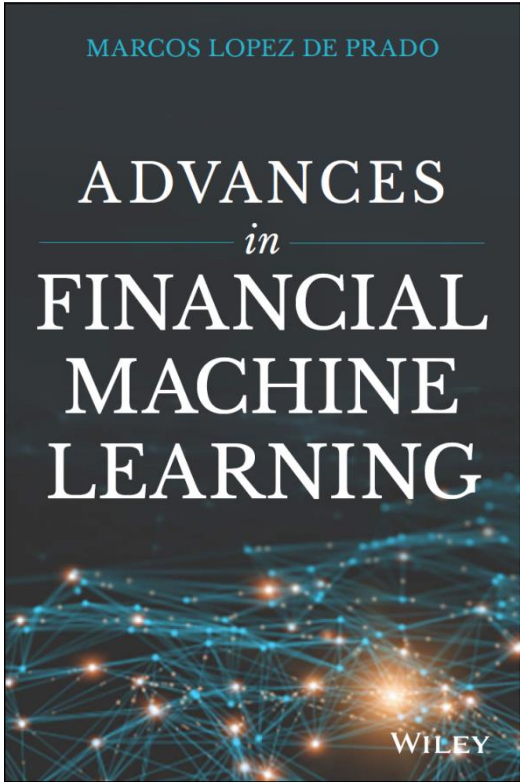
The Assembly Line In Action

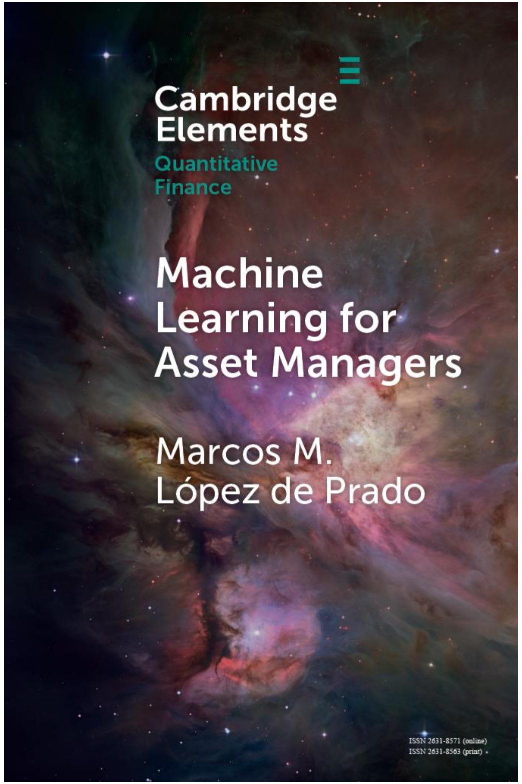
If you are passionate about research, and wish to prove yourself cracking some of the hardest open problems in data science, join ADIA's Research Lab.

To learn more, visit ADIA's job portal at

https://jobs.adia.ae/

For Additional Details





The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this presentation are my own, and do not necessarily reflect the views of Cornell University or Abu Dhabi Investment Authority
- No investment decision or particular course of action is recommended by this presentation
-

www.QuantResearch.org