Ten Financial Applications of Machine Learning

Prof. Marcos López de Prado

Seeing Beyond The Hype

- Financial ML offers the opportunity to gain insight from data:
  - Modelling non-linear relationships in a high-dimensional space
  - Analyzing unstructured data (asynchronous, categorical)
  - Learning patterns with complex interactions (hierarchical, non-parametric)
  - Focusing on predictability over parametric adjudication
  - Controlling for overfitting (early-stopping, cross-validation)
- At the same time, **Finance is not a plug-and-play subject** as it relates to machine learning.
  - Modelling financial series is harder than driving cars or recognizing faces.
  - **A ML algorithm will always find a pattern, even if there is none!**
- In this presentation, we review a few important financial ML applications.

What is Machine Learning?

What Is Machine Learning?

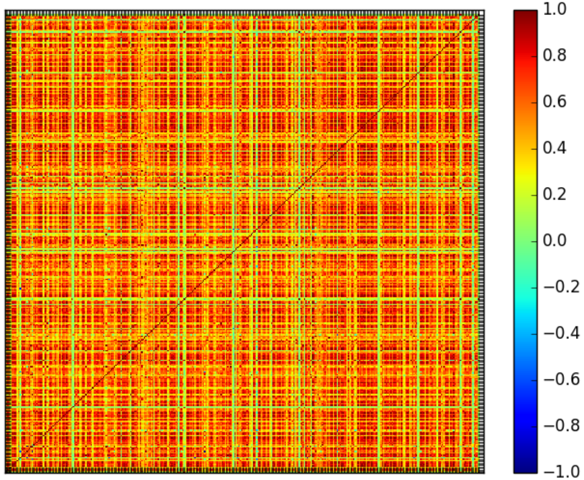
> **“An ML algorithm learns complex patterns in a high-dimensional space without being specifically directed.”**

Advances in Financial Machine Learning (2018, p.15)

Let’s break this statement into its components.

- **“learns … without being specifically directed”**: Unlike with other empirical tools, researchers do not impose a particular structure on the data. Instead, researchers let the data speak.
- **“learns complex patterns”**: The ML algorithm may find a pattern that cannot be easily represented with a finite set of equations.
- **“learns … in a high-dimensional space”**: Solutions often involve a large number of variables and the interactions between them.

Suppose that you have a 1000x1000 correlation matrix...



What Is Machine Learning?

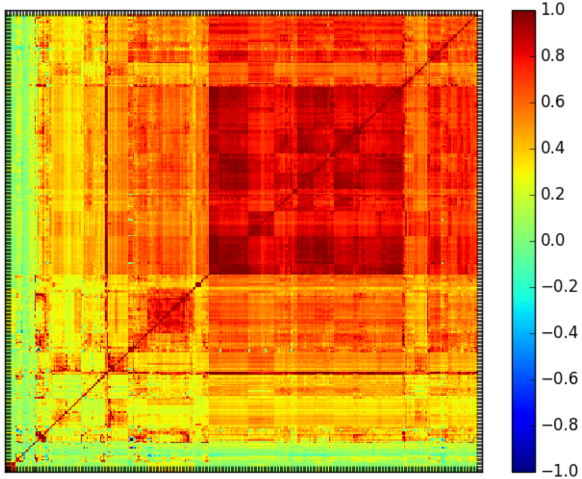
> **“An ML algorithm learns complex patterns in a high-dimensional space without being specifically directed.”**

Advances in Financial Machine Learning (2018, p.15)

Let’s break this statement into its components.

- **“learns … without being specifically directed”**: Unlike with other empirical tools, researchers do not impose a particular structure on the data. Instead, researchers let the data speak.
- **“learns complex patterns”**: The ML algorithm may find a pattern that cannot be easily represented with a finite set of equations.
- **“learns … in a high-dimensional space”**: Solutions often involve a large number of variables and the interactions between them.

Suppose that you have a 1000x1000 correlation matrix... A clustering algorithm finds that there are 3 blocks: Highly correlated, low correlated, uncorrelated.



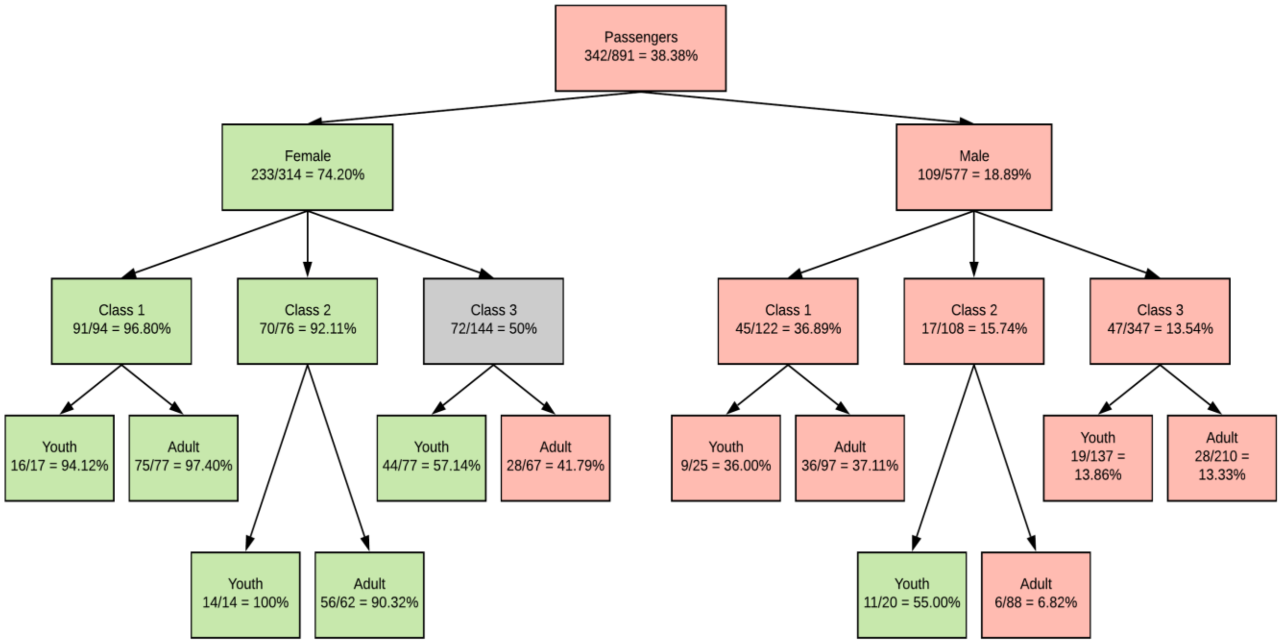
A Simple Example: The Survival Rate At The Titanic

How would you predict the probability that a particular passenger at the Titanic survived?

An ML algorithm will find that three variables are relevant: *Gender*, *ticket class*, and *age*.

The algorithm will also find that the there is a hierarchical structure in the data.

For example, for the purpose of surviving, being female is more important than being young or having a first class ticket.



Passengers 342/891 = 38.38%

Female 233/314 = 74.20%

Male 109/577 = 18.89%

Class 1 91/94 = 96.80%

Class 2 70/76 = 92.11%

Class 3 72/144 = 50%

Class 1 45/122 = 36.89%

Class 2 17/108 = 15.74%

Class 3 47/347 = 13.54%

Youth 16/17 = 94.12%

Adult 75/77 = 97.40%

Youth 44/77 = 57.14%

Adult 28/67 = 41.79%

Youth 9/25 = 36.00%

Adult 36/97 = 37.11%

Youth 19/137 = 13.86%

Adult 28/210 = 13.33%

Youth 14/14 = 100%

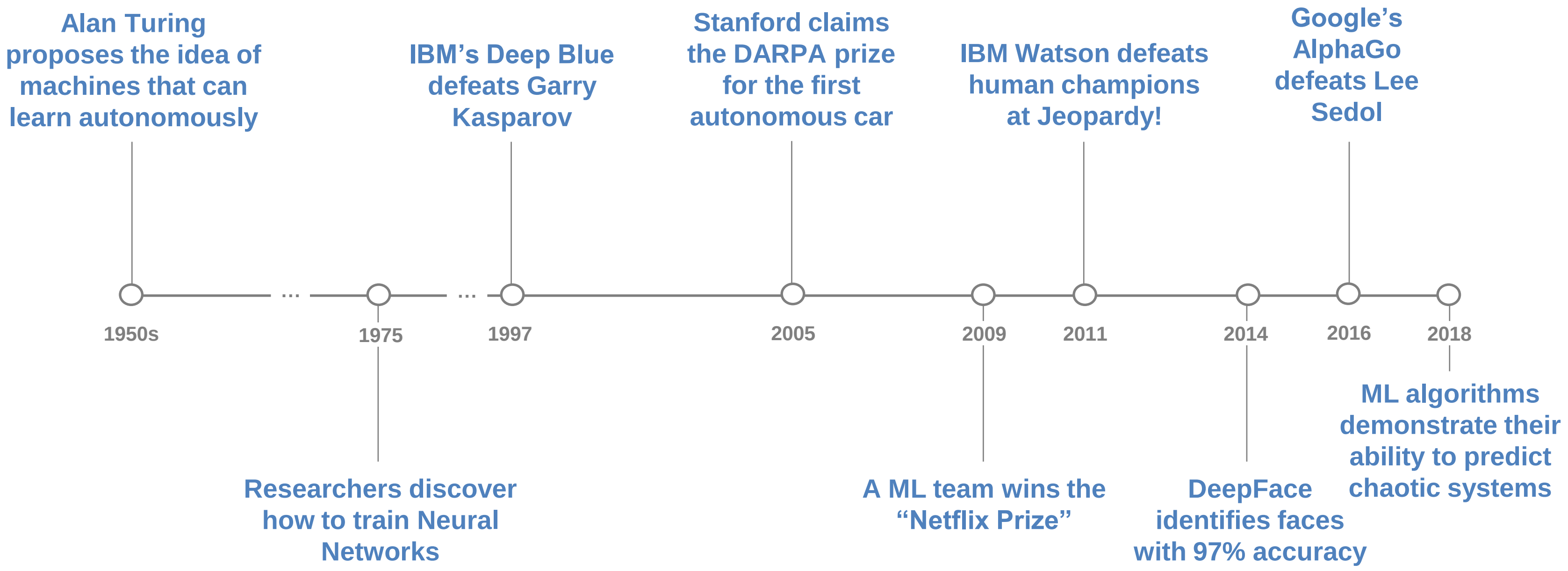
Adult 56/62 = 90.32%

Youth 11/20 = 55.00%

Adult 6/88 = 6.82%

Source: Lam, E. and C. Tang. For illustrative purposes only. Available at https://bit.ly/2UjVmOI

Timeline



1950s

Alan Turing proposes the idea of machines that can learn autonomously

1975

Researchers discover how to train Neural Networks

1997

IBM’s Deep Blue defeats Garry Kasparov

2005

Stanford claims the DARPA prize for the first autonomous car

2009

A ML team wins the “Netflix Prize”

2011

IBM Watson defeats human champions at Jeopardy!

2014

DeepFace identifies faces with 97% accuracy

2016

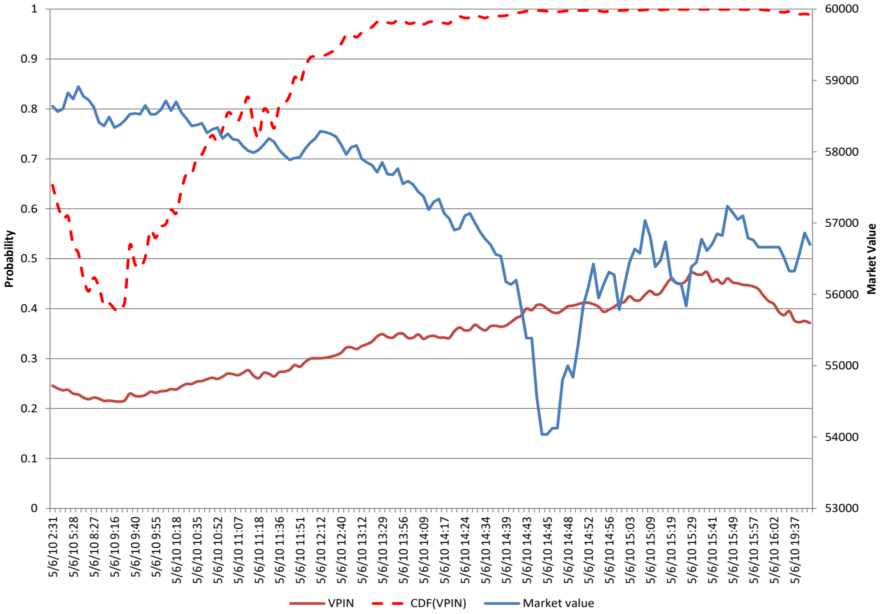
Google’s AlphaGo defeats Lee Sedol

2018

ML algorithms demonstrate their ability to predict chaotic systems

Can ML Predict Black Swans?

A **black swan** is an extreme event that has not been observed before. E.g., the “flash crash” of May 6 2010.



VPIN detected extreme levels of persistent order flow imbalance two hours prior to the crash

The official investigation into the flash crash found that the likely cause was an order to sell 75,000 E-mini S&P 500 futures contracts at a high participation rate.

That large order caused a persistent imbalance in the order flow, which triggered a cascade of stop-outs across market makers, until nobody stood on the bid.

Imbalanced order flow is the norm, with various degrees of persistency. The 10% sudden drop in prices was a black swan, but the causes were known to microstructure theory.

Conclusion: **Black swans can be predicted *by theory,* even if they cannot be predicted *by algorithms***.

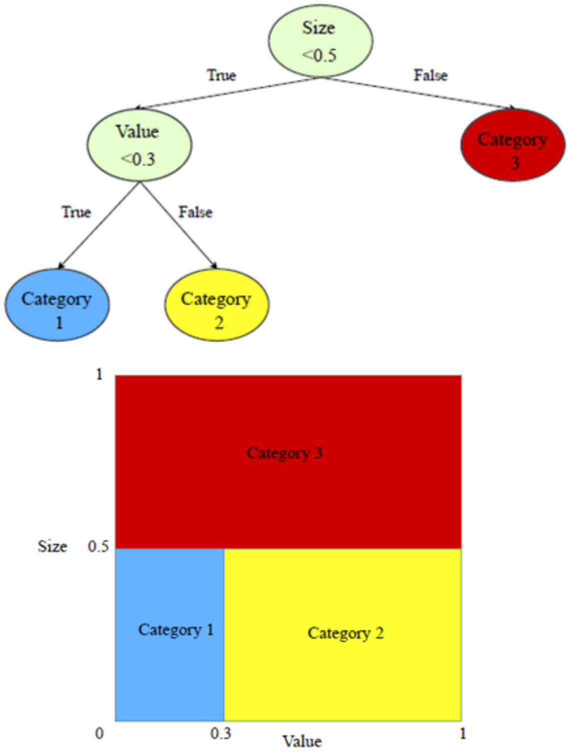
Corollary: Use ML for developing theories, and let the theories make the predictions (not the algorithms).

Simons on Machine Learning

Current Applications of Financial ML

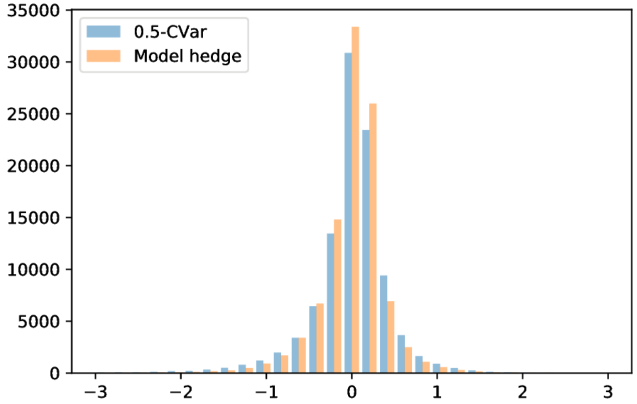
1. Price Prediction

- ML methods allow the modelling of complex relations among the 6 or 7 widely accepted economic factors, including
  - Non-linear relations
  - Threshold relations
  - Hierarchical relations
  - Categorical variables
  - Unknown specification
  - Interaction effects
  - Control variables
- Econometric methods fail to recognize complex relationships, hence leading to inferior results.

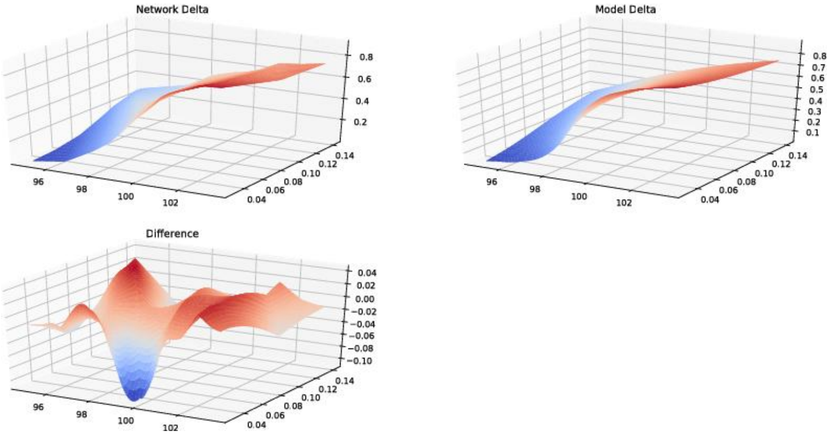


2. Hedging

- Analytical hedging is problematic in presence of market frictions, such as transaction costs, market impact, liquidity constraints, risk limits, etc.
- Reinforcement learning approaches are Greek-free and model free. They are purely empirical, with very few theoretical assumptions.
  - These models consider many more variables and data points when making hedging decisions, and can generate more accurate hedges at greater speeds.

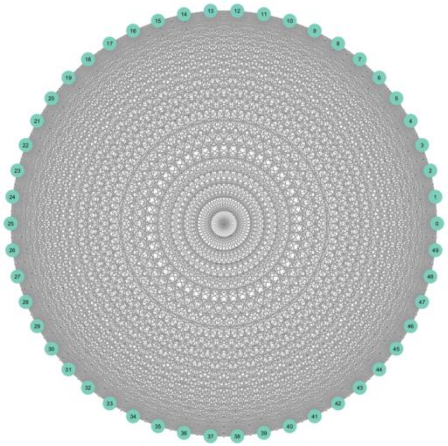


Comparison of model hedge and deep hedge associated to 50%-expected shortfall criterion



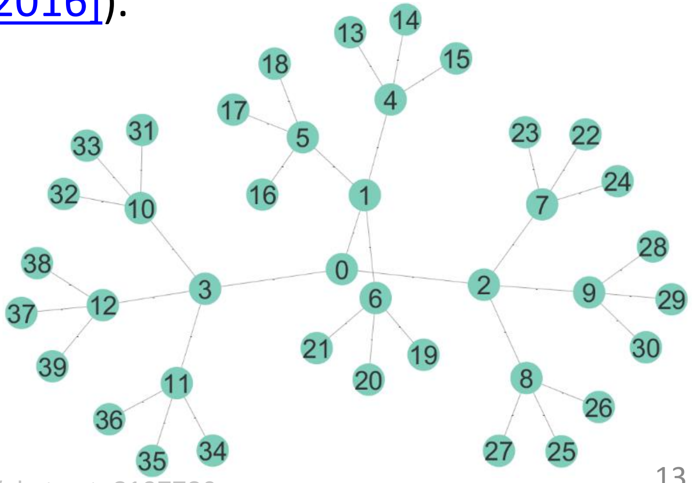
3. Portfolio Construction / Risk Analysis

- Most firms continue to allocate trillions of dollars using mean-variance portfolio optimization (MVO). *“The most expensive piece of beautiful math in history.”*
- It is widely known that MVO underperforms the naïve allocation out-of-sample (De Miguel et al. [2009]).
- In contrast, ML solutions outperform MVO (and 1/N) out-of-sample, with gains in Sharpe ratio that exceed 31% (López de Prado [2016]).



Covariance-based models require the independent estimation of $N(N+1)/2$ variables.

ML models need only $N-1$ *hierarchical* estimates, making them more robust and reliable.



4. Structural Breaks / Outlier Detection

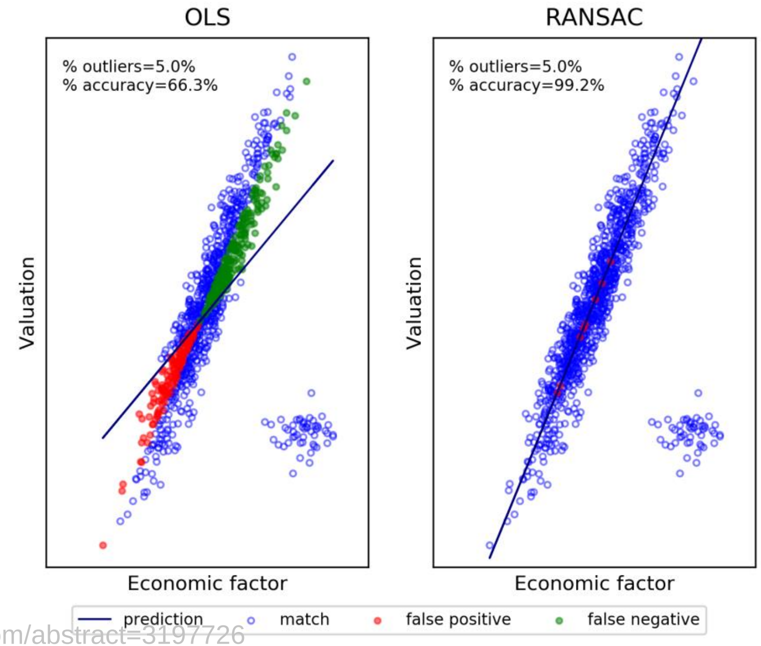
Cross-sectional studies are particularly sensitive to the presence of outliers. Even a small percentage of outliers can cause a very large percentage of wrong signals: Buys that should be sells (false positives), and sells that should be buys (false negatives).

In this plot we run a regression on a cross-section of securities, where a very small percentage (only 5%) are outliers:

- The **red dots** are securities that are expensive, but the regression wrongly classified as cheap.
- The **green dots** are securities that are cheap, but the regression wrongly classified as expensive.

With only 5% of outliers, the cross-sectional regression produced a 34% classification error. In contrast, RANSAC's classification error was 1%, involving borderline cases.

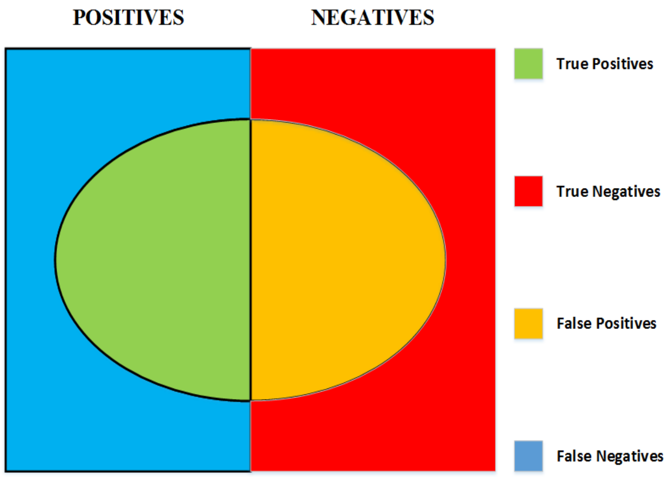
Whenever you suspect the presence of outliers in your data, consider applying RANSAC or similar ML methods.



4. Structural Breaks / Outlier Detection

5. Bet Sizing / Alpha Capture

- Suppose that you have a model for making a buy-or-sell decision:
  - You just need to learn the size of that bet, which includes the possibility of no bet at all (zero size).
  - This is a situation that practitioners face regularly. We often know whether we want to buy or sell a product, and the only remaining question is how much money we should risk in such bet.
  - Meta-labeling: Label the outcomes of the primary model as 1 (gain) or 0 (loss).



- Meta-labeling builds a secondary ML model that learns how to use a primary exogenous model.
- The secondary model does not learn the *side*. It only learns the *size*.
- We can maximize the F1-score:

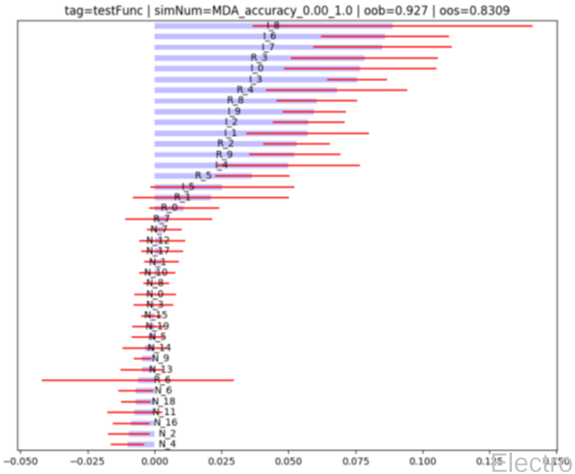
$$F1 = 2\frac{precision \cdot recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

6. Feature Importance

- ML algorithms identify patterns in a high dimensional space.
- These patterns associate features with outcomes.
- The nature of the relationship can be extremely complex, however we can always study what features are more important.
  - E.g., even if a ML algorithm may not derive an analytical formula for Newton’s Gravitational Law, it will tell us that *mass* and *distance* are the key features.



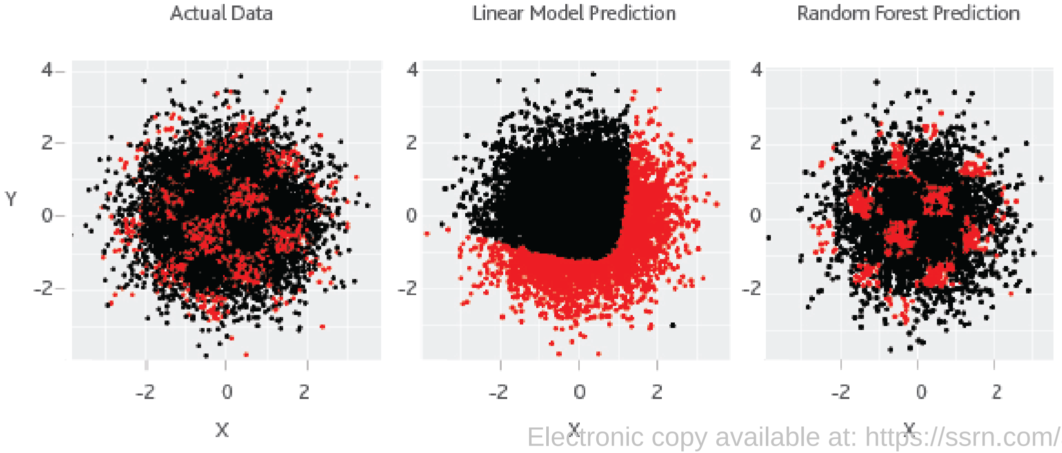
In traditional statistical analysis, key features are often missed as a result of the model’s misspecification.

In ML analysis, we give up closed-form specifications in exchange for identifying what variables are important for forecasting.

Once we know *what* are the factors at play, we can develop a theory of *how*.

7. Credit Ratings, Analyst Recommendations

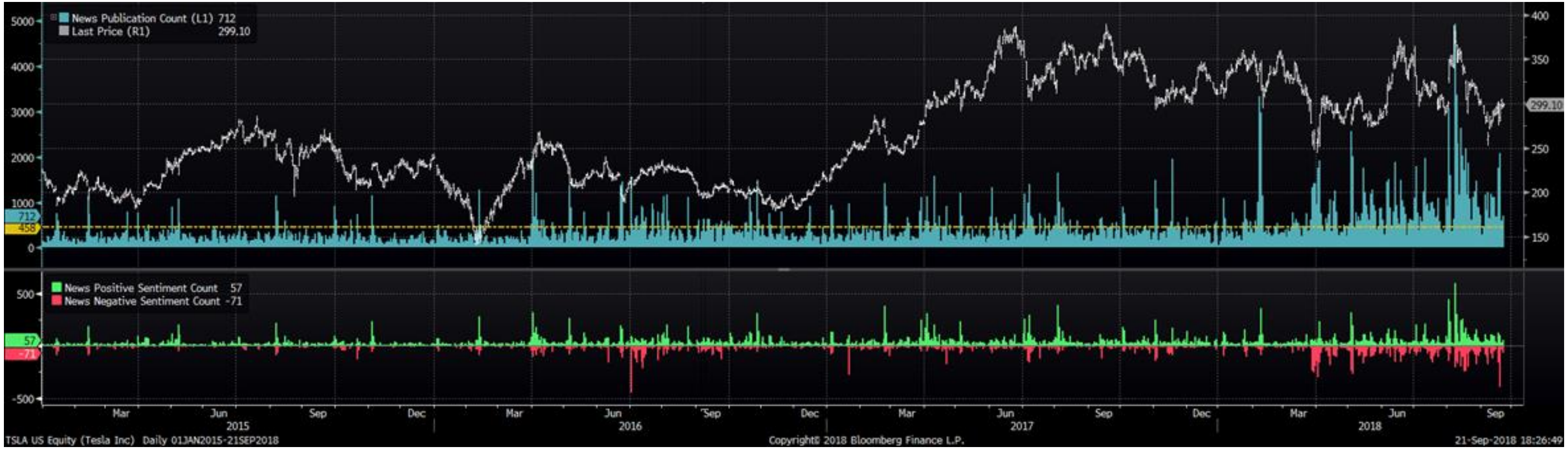
- Stock analysts apply a number of models and heuristics to produce credit and investment ratings.
- These decisions are not entirely arbitrary, and correspond to a complex logic that cannot be represented with a simple set of formulas or a well-defined procedure.
- Machine learning algorithms have been successful at replicating a large percentage of recommendations produced by bank analysts and credit rating agencies.



In this example by Moody's, the left figure shows a scatter plot of bonds as a function of two features (X,Y), where defaults are colored in red. The middle plot shows that traditional econometric methods fail at modelling this complex, non-linear relationship. The right plot shows that a very simple ML algorithm performs well.

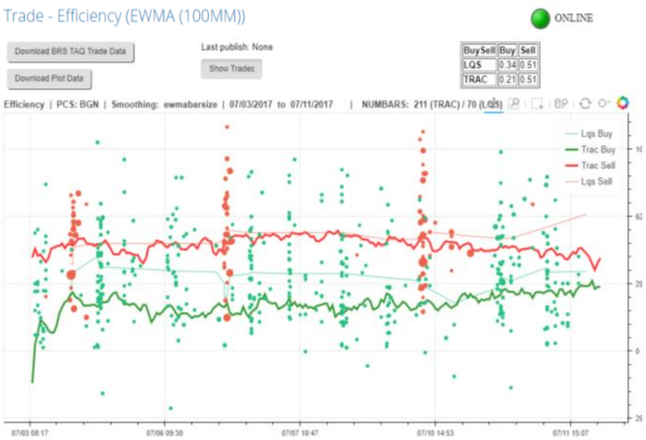
8. Unstructured Data

- In the plot below, an algorithm has identified news articles containing information relevant to Tesla (TSLA US Equity).
  - Blue bars: Daily count of the total number of articles. The average is 458 articles/day, with a maximum of ~5000.
  - Green bars: Daily count of articles expressing a positive sentiment.
  - Red bars: Daily count of articles expressing a negative sentiment.



9. Execution

- Credit instruments are
  - traded over-the-counter
  - relatively illiquid (they may not trade for days and weeks)
- Kernel-based methods identify “similar” trades based on their common features.
  - The set of common trades enables us to derive theoretical prices.
  - If we buy a bond at a price higher than subsequent “similar” bonds, we can bust the trade.



Trade - Efficiency (EWMA (100MM))

Last publish: None

Buy Sell	Buy	Sell
LQS	0.34	0.51
TRAC	0.21	0.51

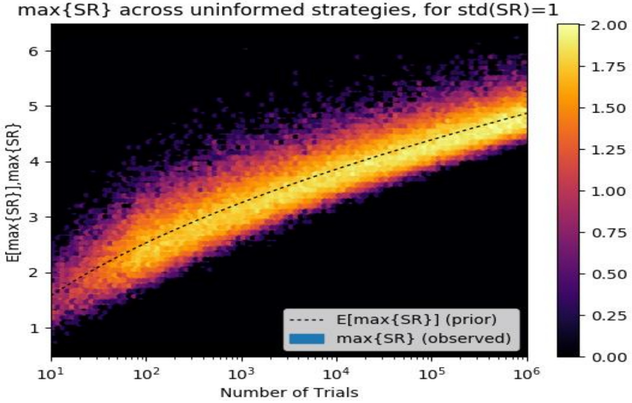
Efficiency | PCS: BGN | Smoothing: ewmabarsize | 07/03/2017 to 07/11/2017 | NUMBARS: 211 (TRAC) / 70 (LQS)

This plot shows the trade efficient of buys (green) and sales (red):

- A **buy** has efficiency 0 when it prints at the quoted offer, and it has efficiency 100 when it prints at the quoted bid.
- A **sale** has efficiency 0 when it prints at the quoted bid, and it has efficiency 100 when it prints at the quoted offer.
- Both have efficiency 50 at the mid.

In this example, the rebalancing of the portfolio has been profitable, as it has captured about 1/3 of the bid-ask spread (approx. 50 bps in price).

10. Detection of False Investment Strategies

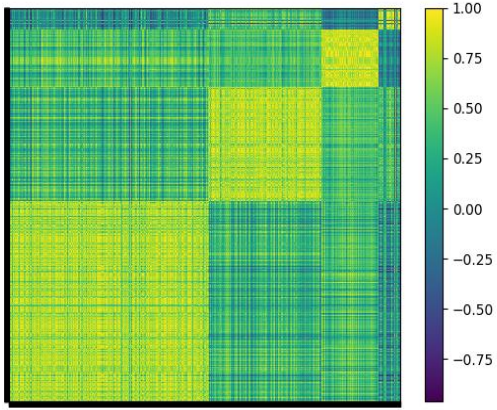


The y-axis displays the distribution of the maximum Sharpe ratios (max{SR}) for a given number of trials (x-axis). A lighter color indicates a higher probability of obtaining that result, and the dash-line indicates the expected value.

For example, after only 1,000 independent backtests, the expected maximum Sharpe ratio (E[max{SR}]) is 3.26, even if the true Sharpe ratio of the strategy is zero!

Most quantitative firms invest in false discoveries.

Solution: Deflate the Sharpe ratio by the number and variance of trials.



Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Strat Count	3265	1843	930	347
aSR	1.5733	1.4907	2.0275	1.0158
SR	0.0974	0.0923	0.1255	0.0629
Skew	-0.3333	-0.4520	-0.4194	0.8058
Kurt	11.2773	6.0953	7.4035	14.2807
T	2172	2168	2174	2172
StartDt	2010-01-04	2010-01-04	2010-01-04	2010-01-04
EndDt	2018-05-01	2018-04-25	2018-05-03	2018-05-01
Freq	261.0474	261.0821	261.1159	261.0474
sqrt(V[SR_k])	0.0257	0.0256	0.0256	0.0257
E[max SR_k]	0.0270	0.0270	0.0270	0.0270
DSR	0.9993	0.9985	**1.0000**	0.9558

The selected strategy belongs to Cluster 2. After taking into account the number and variance of trials involved in the discovery, the probability that $SR > 0$ is virtually 1. Hence, the backtest is unlikely to be overfit.

The Perils of Financial ML

The “spilled samples” problem (1/2)

- Most non-financial ML researchers can assume that observations are drawn from IID processes. For example, you can obtain blood samples from a large number of patients, and measure their cholesterol.
- Of course, various underlying common factors will shift the mean and standard deviation of the cholesterol distribution, but the samples are still independent: There is one observation per subject.
- Suppose you take those blood samples, and someone in your laboratory spills blood from each tube to the following 9 tubes to their right.
  – That is, tube 10 contains blood for patient 10, but also blood from patients 1 to 9. Tube 11 contains blood from patient 11, but also blood from patients 2 to 10, and so on.

The “spilled samples” problem (2/2)

- Now you need to determine the features predictive of high cholesterol (diet, exercise, age, etc.), without knowing for sure the cholesterol level of each patient.
- That is the equivalent challenge that we face in financial ML.
  - Labels are decided by outcomes.
  - Outcomes are decided over multiple observations.
  - Because labels overlap in time, we cannot be certain about what observed features caused an effect.

My friend Luna can recognize faces, like Google or FaceBook. She is not so good at investing, and Google’s ML would probably fail miserably if applied to financial markets.

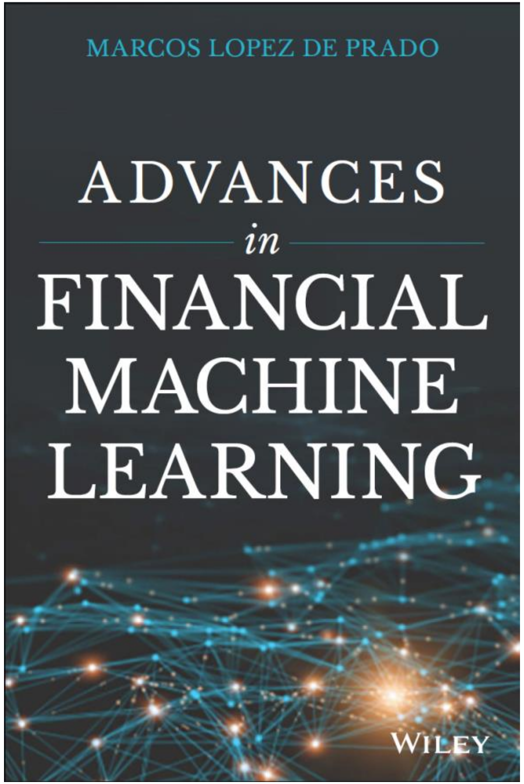
Finance is not a plug-and-play subject as it relates to ML. There are no “West Coast” solutions to “East Coast” problems.

Financial ML as a specific subject

- Financial series exhibit properties that are inconsistent with standard ML assumptions. **A ML algo will always find a pattern, even if there is none!**

PROBLEM	A SOLUTION
Non-stationarity with long memory	Fractional differentiation
Variable information arrival rate	Order imbalance bars
Outcomes span multiple observations	Triple barrier method, with uniqueness weighting
Regime switches	Structural-break methods
Dependence, serial and cross-sectional	K-fold CV with purging, embargoing
Single path for backtesting	Combinatorial cross-validation
Low signal/noise ratio. Backtest overfitting	Deflated Sharpe ratio by controlling for the number of trials

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 document are the authors’ and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-