History of Machine Learning in Finance and Economics

D Snow, published: 2021-01-08

Introduction

Finance and Economics have been slow to adopt modern machine learning techniques. Nevertheless, the researchers and practitioners in these respective domains have been essential in laying the bedrock of what we now refer to as machine learning. The use of mathematics in the service of social and economic analysis dates back to the 17th century. Then, mainly in German universities, a style of instruction emerged which dealt explicitly with the detailed presentation of data as it relates to public administration. Gottfried Achenwall lectured in this fashion in the mid-18th century, coining the term *statistics*. By the 19th century, we saw a flurry of economic statistics, some of which gave rise to statistical learning methods. Then at the start of the 20th century, French mathematician Louis Bachelier published his treatise *Theory of Speculation* that is considered the first scholarly work on mathematical finance.

In 1966 Joseph Gal in the Financial Analyst Journal wrote that ''It will soon be possible for portfolio managers and financial analysts to use a high-speed computer with the ease of the desk calculator''[1]. Today, machine learning code has been streamlined; in less than 10-lines of code, you can create a close to state-of-the-art machine learning option pricing model with free online computing power. This is reminiscent of the 1970s, where not long after the creation of the Chicago Board Options Exchange, Black-Scholes option values could be easily calculated on a handheld calculator. We are not there yet, but it is in within reach. This article seeks to understand the use and the development of what we now refer to as machine learning throughout the history of finance and economics.

In this paper, we will discover the development of linear regressions and correlation coefficients, and the use of least squared methods in astronomy for predicting orbital movements. Although the method of least squares had its start in astronomy, the discipline has since moved on to more fundamental equations that underpin planetary movements. Modern astronomers do not just take raw statistical readings from their telescope to throw into the hopper of a correlation machine as we now do in social sciences. Finance and economic practitioners have tried to model some of these fundamental equations with theoretical foundations, but so far, they produce lacklustre prediction performance. So far, the weight of evidence is that a hodgepodge of correlations is the best prediction machines in disciplines that have some human behavioural component.

For finance, there are broadly two hypotheses that explain why we will never uncover the true fundamental relationships in financial markets, being the efficiency market hypothesis, and adaptive markets hypothesis, neither of which is necessarily in conflict with the other. Professional quantitative traders have for many years called out public researchers for their unreliable prediction models. However they do appreciate new ideas that emerge from the field. One of the earliest critical studies was performed by a respected researcher in the area, Alfred Cowles, who published in Econometrica on the subject of stock market forecasting. In 1933 he had compared 24 financial publication with 24 random portfolios for the period January 1928 to June 1932 and found no evidence that the financial publications outperformed random portfolios except for one.

He came back in 1944 to acknowledge that one method did perform well, outperforming the market with around two percentage points for approximately 40 years, with mostly consistent performance

[1] https://www.cfainstitute.org/en/research/financial-analysts-journal/1966/man-machine-interactive-systems-and-their-application-to-financial-analysis

over all sub-periods. This small win provided some relief to active managers participating in this debate. Two decades later in 1966 in correspondence with colleagues, he revealed that he does indeed believe that knowledge of financial patterns could help to explain the witnessed performance, leaving the door open to demonstrate the performance of individual funds over the period.

The mid-to-late 1980s was the first-time advanced machine learning methods had been used in the industry. This movement started because of traders like Edward Thorp, and Richard Dennis showed remarkable success by combining technical trading methods with statistics. Soon enough, labs like the Morgan Stanley ATP group started with people like Nunzio Tartaglia at its head in 1986. A year later in 1987, Adams, Harding, and Leuck started Man AHL London. In 1987 two years after joining Morgan Stanley, David Shaw decided to start his own quantitative fund DE Shaw & Co. That same year James Simons renamed his Monemetrics to Renaissance Technologies to emphasise its new quantitative focus, and a few months after that Israel Englander launched Millennium Management.

This surge was no coincidence, on the back of these methods sitting in old papers, waiting to be used, we saw the proliferation of computers. In 1980, Sinclair in the U.K. sold a home computer for £99.95, that is less than £500 in today's money. In 1981, IBM announced its Personal Computer, which becomes the basis of most of the modern personal computer industry. This advance not only led to improved computation power but also improved storage capacity and data processing ability. It allowed for the collection of large datasets, which was in turn used by faster and more powerful computers to mathematically scour for patterns in financial data to find predictive trading signals. Throughout this period, much of what we now referred to as machine learning has already been well tried and tested by these firms. Neural networks, nonparametric kernel regressions, and dynamic programming methods were already experimented with by ex-IBM folk and notable academics at Renaissance Technologies in the late 1980s.

Market Patterns

In 1956 a prescient student at the University of Aberdeen, KI Gordon wrote that in the price of commodities there is a vast interdependent system that would require some form of complex modelling to calculate continuous space solutions. He prophesises that in the future with the use of Cybernetics, the prices and availability of ''various primary products and shares'' could be included in the input of models for the ''prediction of future fluctuations in supplies and prices'' and ''future booms, slumps, or periods of stability''.

He goes on to say that government action could then be taken to maintain conditions of stability and prevent slumps. Delighting himself in Keynesian future-speak, he predicts that aimed with these models there would be ''no more depressions, no more unemployment, no more inflation''. He quickly pulls back and says, maybe that view is too rosy because we are yet to consider the effects that such predictions might have on the market. Following along the modern discussion of the inverted yield curve, Gordon points out that the ''mere appointment of a Senate Investigating Committee to examine the stock market was enough to lower share prices by about 10%. In a similar manner, talk of an impending slump is liable to precipitate the very condition it is desired to avoid.''

Technical trends have been explicated by Joseph de la Vega as he analysed the Dutch financial markets in the 17th century. Homma Munehsia in the 18th century developed candlestick techniques. In the 19th century, Clement Juglar studied financial and economic time series and introduced the idea of observable ''cycles'' in business activity in 1862 after studying banking statistics like metallic reserves, banknote circulation, and deposits[2]. Early technical analysis exclusively features the analysis of charts

[2] *Des Crises Commerciales et de leur Retour Périodique en France, en Angleterre et aux Etats-Unis*

due to the lack of computing power available for statistical analysis, Charles Dow originated one of the first point and figure chart analysis. In fact, Dow theory from the collective writing of Dow Jones, inspired the development of modern technical analysis at the end of the 19th century. In the early 20th century, Richard Wycoff developed, among others, a measure of the dullness of the market, in essence, an oscillating wave with a decreasing amplitude that peters out[3]. In the 1920s and 30s Richard Schabacker published books such as Stock Market Theory and Practice and Technical Market Analysis.

The basic principles come down to the belief that prices move in trends and that history repeats itself. Harry Roberts, in 1959 writing in the Journal of Finance, discusses analyst's obsession with patterns that offer alleged 'clues' about the future. In so doing, he takes note of the fact that some draw inspiration from physical processes like tides and waves, and that some are purely empirical. Regardless of these criticisms, indirect evidence for technical analysis was provided by Tabell and Tabell in 1964[4], and more direct support by 1988 Priott and White[5]. In a 2007 review, Irwin and Park reported that more than half of around 100 modern studies produce positive results[6]. Moreover, traders like Richard Dennis, William Eckhardt, and Paul Tudor Jones have also amassed large fortunes using technical analysis, much of which can't be explained by mere luck.

The evidence for patterns is quite evident in both finance and economics, both in the cross-section, the time series, the panel, and the network. In 1929, Eugen Slutzky already discovered the potential of cumulative effects from independent random influences affecting a business cycle favourable or unfavourable[7]. Patterns exist in most human-domains, and pattern-recognition software is here to stay. Machine learning models are set out to spot the difference between Yules nonsense-correlations between time series and proper financial cycles[8]. In 1974 Russell Fogler started to identify patterns in business processes that can be used for forecasting. He essentially reclassifies observations into patterns to improve the accuracy of forecasts.

It is sometimes said that the ability to predict is the final test of science. Successful prediction requires two conditions: first knowledge of how the science behaves under different circumstance, and second, the successful recording and transmission of this behaviour to a prediction model. Failures of prediction are due to the lack of either of these two essential conditions. As a result, a good default assumption is that prediction models mostly fail and sometimes work, especially in social science. An example of a failure of prediction due to imperfect knowledge of facts is found in the case of the closure of the Indian mints to silver in 1893. It was expected that the value of the silver rupee would be maintained at 16 pence. But no account was taken of the large number of silver coins among the natives causing a failure in prediction. In this case, the failure of prediction at first was due, not to any defect in monetary science, but to ignorance of Indian history, something that is not easily written into a machine learning model.

<u>Programmability</u>

[3] http://libarch.nmu.org.ua/bitstream/handle/GenofondUA/16293/949111f86496672faafa6a518c329e91.pdf?sequence=1&isAllowed=y

[4] https://sci-hub.st/10.2469/faj.v20.n2.67

[5] https://sci-hub.st/https://doi.org/10.3905/jpm.1988.409149

[6] https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2007.00519.x

[7] https://sci-hub.st/10.2307/1907241

[8] https://www.jstor.org/stable/2341482?seq=1

Some tasks can be programmed, and some tasks cannot[9]. GPE Clarkson showed in 1962 how the portfolio selection decision made by bank trust investments could be automated using discriminator nets, i.e., a sequential branching computer program[10]. Later systems like that by Jesup & Lamont, a New York brokerage firm not only routinised investment decisions based on decision heuristics but also self-learned new patterns to further refinement.[11]

Soon enough, these decision-making systems were used to replace recruitment advisors in 1964[12], loan officers in 1964[13], treasurers in 1966[14], and sales analysts in 1967[15]. These systems didn't just perform well; their performance was often indistinguishable from professional investors, analysts, and advisors. This line of research has in those years been dubbed as behavioural theory and have taken particular update at Carnegie tech. These researchers have shown that the data can be fit very well with a few simple equations. It did not take too long for these advances to reach the popular news in a 1966 Business week article, ''The Shake-up of Conventional Economics''.

These and similar methods have slowly percolated through the industry, and the nonparametric statistical routines that Jesup & Lamont have used in the 1960s have now become commonplace. Skilled staff at JP Morgan Chase have recently suffered the fate of the machine. A new contract intelligence programme was established to interpret commercial-loan agreements that previously required 360k hours of legal work by loan officers and lawyers[16]. Other machine learning-enabled tasks include the automation of post-allocation requests at UBS[17] and the automation of policy pay-outs at Fukoku Mutual Life Insurance[18].

GPE's portfolio selection model was validated using the ''Turing's Test'' that compares man with model, the model's final portfolio bears an astonishing resemblance with that selected by man. The machine was programmed so that the output appears in the English language. A 1963 review says that ''many portions of a portfolio selection process are treated as data by Clarkson, rather than as objects of analysis'' like a list of the suitable assets considered for trust investment, a list of relevant industry groups for respective investment goals. It is precisely this reliance on data that allows the model to perform so well. The dissertation was six years after the Dartmouth workshop that to many was the official creation of the field of artificial intelligence and one year after Herbert Simon's Ford lecture and book, *Organisations*.

''In common with other heuristic programs, the process is iterative. Lists of industries and companies are searched for particular attributes; sublists are created, searched, and again divided. For example, to obtain a high-growth portfolio, the list of companies stored in memory is searched to obtain shares with the desired characteristics. Additional criteria are employed to narrow (or expand) this list. Further search and matching against desired criteria yield the specific selection of stocks to buy''[19].

[9] https://www.jstor.org/stable/2390654
[10] Portfolio Selection: A Simulation of Trust Investment
[11] The computer, new partner in investment management
[12] Project manager selection: the decision process
[13] https://books.google.com.au/books/about/Bank_Procedures_for_Analyzing_Business_L.html?id=hHQIHAAACAAJ&redir_esc=y
[14] A theory of the budgetary process
[15] Simulation and Sales Forecasting
[16] JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours
[17] Robots enter investment banks{'} trading floors
[18] This Japanese Company Is Replacing Its Staff With Artificial Intelligence
[19] https://sci-hub.st/https://doi.org/10.1111/j.1540-6261.1960.tb02764.x

At the time, various researchers rightly pointed to the fact that although specific tasks and subroutines can be automated by learning from humans and programming that into a machine, not everything is programmable. Baumol wrote in 1966, ''There is a rich body of information, much of it hardly tangible, which necessarily influences the recommendations of an investment counsellor, but which the computer is not able to take into account. Knowledge about the personality of a retiring company president, or about the implications of current Defense Department discussions for the saleability of a company's products is surely important, but there is no way in which it can be taken into account automatically and mechanically in accord with predeveloped mathematical formulae.''[20]

These concerns were for the time pushed aside, and researchers kept working on new ways to program heuristics to make decisions. The Harvard Business Review in 1966, wrote an essay ''Heuristic Programs for Decision Making'' where they discuss how strange words like cybernetics have been adopted into the vocabulary of the business world. Macrae already claimed in 1951 that the word cybernetics is familiar to most sociologists[21]. In the late 1950s, there was a lot of excitement for the potential of automation to help out accountants and bankers speed up their work[22].

Heuristic has also been given a new meaning with problem-solving connotations. A heuristic can be as easy as "buy when the price is low" for stock trading, "buy when you are down to your last two items" for inventory control, "first-in-first-out" for accounting, "first come first serve" for job scheduling. Heuristic programming, unlike linear programming, is not concerned with finding the optimal answer, but a satisfactory one, often the only solution when the problem is too large to be solved with analytical techniques. It is also valid when questions are ill-structured and involve intuition such as which you want to imitate a human problem-solving process. These heuristic models were one of finance and economics first inroads to *learn from data*.

In the inaugural issue of Institutional Investors in 1967, Gilbert Kaplan wrote about how computer applications could help money managers improve their portfolio performance. In the mid-1960s, NYSE firms were already spending $100mn annually to computerise their operations[23]. A 1962 Wall Street Journal article described how as far back as 1954, some brokers have already started to use data-processing equipment for market analysis to look at changes 92 different industry groups and 22 stocks broken down by size[24]. Another author in 1967 wrote that the availability of ''machine-readable data'' made it fashionable to be doing something with computers.

Heuristic programming is not unlike the 20-person team that was said to translate Ray Dalio's unique financial worldview into algorithms[25]. Employees have referred to the project as trying to turn Ray's brain into a computer[26]. Steven Cohen at Point72 is also testing models that mimic the trades of their portfolio managers[27]. And Paul Tudor has also assigned a team of coders to create ''Paul in a Box''[28].

[20] https://sci-hub.st/10.2469/faj.v22.n5.95

[21] https://sci-hub.st/10.2307/587385

[22] http://www.newlearner.com/courses/hts/bat4m/pdf/accounting_technology.pdf

[23] https://books.google.com.au/books/about/A_Financial_History_of_the_United_States.html?id=l8-ZAAAAIAAJ&redir_esc=y

[24] Stock Analysts Study Computers as Aid in Investing, See No Easy Road to Riches

[25] https://www.bloomberg.com/news/features/2017-08-10/bridgewater-s-ray-dalio-has-a-plan-to-outlive-himself

[26] https://www.wsj.com/articles/the-worlds-largest-hedge-fund-is-building-an-algorithmic-model-of-its-founders-brain-1482423694

[27] https://www.bloomberg.com/news/articles/2017-06-28/fund-manager-who-traded-for-cohen-trains-algos-to-copy-his-brain

[28] https://www.bloomberg.com/news/articles/2017-10-20/automation-starts-to-sweep-wall-street-with-tons-of-glitches

In Alchemy and Artificial Intelligence, 1965, Hubert L Dreyfus writes, ''…In problem-solving once the problem is structured and planned a machine could take over to work the details…as in the case of machine shop allocation or investment banking''. More and more so, financial institutions are coming to terms with the rising prospects of automation.

The rule-based systems that took off in the 1960s continued for the next couple of decades, but the terminology changed from heuristics to expert systems. Expert systems appeared in finance around the early 1980s. At that time, Dupont had built more than a hundred expert systems that helped save them and estimated $10mn a year[29]. In finance, the Protrader expert system built in 1989 by KC Chen was able to determine optimal investment strategies, execute transactions, and modify the knowledgebase through a learning mechanism[30]. Investment advisors also had their own version in the form of PlanPower made available in 1986, offering tailored financial plans to individuals with incomes about $75 thousand[31]. And a few years later Chase Lincoln First Bank provided an improved wealth management tool, selling reports for a meter $300[32]. An early indication of what was to follow in modern robo-advisors. This trend continued into the 1990s, by 1993 the U.S. Department of Treasury has also launched a system called FinCEN that was used to determine 400 potential incidents of money laundering over two years equalling $1 billion in value[33].

Expert Systems slowly disappeared as the expectations did not meet reality, and they were complicated to use. For example, G.E. developed a system called Commercial Loan Analysis Support System to assess commercial loans, they secured agreements with two large New York firms, but it fell apart. Even though it performed as intended, there was no established person to manage the system, and the project failed and died[34]. And where they were not too complicated, they changed name, maybe to Excel VBA aka macros. Expert systems did not disappear in its entirety, as some subcomponents still remain today.

In recent years, researchers have been leveraging machine learning methods like reinforcement learning strategies to develop end-to-end derivatives trading businesses[35]. These methods could allow you to model any portfolio of derivatives. The agent is exposed to an environment that consists of the market and other constraints faced by real-world agents and then asked to hedge a position with a set of available securities. We have therefore moved from a world where the rules are automatically learned from data. Moreover, if the business process can be simulated, data can be recorded from the real environment and be further used to automate an activity. By simply learning from data, machine learning has increased the subset of economically viable programmable processes.

The implication for such advances, such as to automation of derivative businesses could have immense consequences. Will these automated businesses exhibit improved hedging performance for lower fees and hence proliferate like exchange-traded funds? To this day analysts are forced to develop an

[29] https://www.sciencedirect.com/science/article/pii/B9780124438804500454

[30] http://www.ecrc.nsysu.edu.tw/liang/paper/3/PROTRADER%20An%20Expert%20System%20for%20Program%20Trading.pdf

[31] https://prism.ucalgary.ca/bitstream/handle/1880/48295/Nielson_Expert_Systems_1990_publishercopy.pdf;jsessionid=26F8564F2BA1536A94A96AD0753F8E7D?sequence=1

[32] https://prism.ucalgary.ca/bitstream/handle/1880/48295/Nielson_Expert_Systems_1990_publishercopy.pdf;jsessionid=26F8564F2BA1536A94A96AD0753F8E7D?sequence=1

[33] https://www.aaai.org/Papers/IAAI/1995/IAAI95-015.pdf

[34] https://www.sciencedirect.com/science/article/abs/pii/037722179290274D

[35] Deep hedging: hedging derivatives under generic market frictions using reinforcement learning

intuition for where their traditional models make false assumptions and have to build a litany of tools and workarounds that could in the future be made superfluous by well-executed reinforcement learning strategies.

It isn't just end-to-end reinforcement learning models that are automated; there is a movement towards automating all prediction problems. A new movement in automated machine learning called AutoML is taking foot. AutoML automates feature engineering, model selection, and hyperparameter optimisation for supervised learning. AutoML had its "Alpha Go" moment when a Google AutoML engine was pitted against hundreds of the world's best data scientists and came second. For now, it is costly to run and AutoML system, but Google is betting on an AutoML model that would become even smarter and cheaper. Kaggle has hosted challenges from the likes of Two Sigma challenge, Winton, and Jane Street, all of whom have readily programmable and automatable prediction problems.

Least Squares Method

Supervised learning takes its roots in Astronomy. In the pre-least-squares era, there were attempts to fit a straight line to data. This problem comes down to determining the best average and measure of variability, the solution of which depends on the distribution of the errors or residuals. This problem dates as far back as Galileo Galilei (1632) who after making orbital observations have noted that the errors needed ''to be corrected for us to get the best possible information from the observations''[36]. The body of statistical theory which looks at this and related problems are called the theory of errors.

John Graunt was 12 years old when Galileo published his ''Dialogue Concerning the Two Chief World Systems''. In a few decades time, he would be regarded as the founder of *demography* and even later referred to as the father of statistics[37]. In his career, Graunt, with others like Edmond Haley established the life table or table of mortality. Karl Pearson referred to him as the first to deal with vital statistics. It is unlikely that Graunt and Galilei know much if anything of each other's work, but in a few centuries, Galilei's mathematical theory of errors would intersect Graunt's statistics. Since time immemorial, we have fitted patterns to data using a range of techniques. However, the theory of errors is really the start of modern machine learning as we know it.

After Galileo, Leonhard Euler (1749) and Johann Tobias Mayer (1750) independently came up with a method of averages for fitting a linear equation to observed data. Since then, there was a range of innovations, such as that by Laplace (1786) who seek to minimise the maximum absolute deviation from the fitted straight line. In 1793 he improved this by saying that the sum of absolute deviations should be a minimum and the sum of deviations should be zero.

The specific method that brought the theory into common practice is the method of least squares. Legendre (1805) and Gauss (1809) discovered and proposed the least-squares method as the best method to fit a good linear line through a set of points. In 1805 Adrien Marie Legendre while not the first to use the method of least squares was the first to publish it. The first use of the method has been claimed by Carl Friedrich Gauss, alleging that he has used it for 11 years before Legendre's publication. In certain circles, these methods soon got particularly popular around the time, and new innovations soon followed like Gergonne' polynomial regression published in 1915 that sparked the start of non-linear regressions[38].

[36] https://sci-hub.st/10.2307/1403077

[37] http://www.cs.xu.edu/math/Sources/Graunt/graunt.html#:~:text=John%20Graunt%20was%20the%20true,the%20Royal%20Society%20of%20London.

[38] https://www.sciencedirect.com/science/article/pii/0315086074900330?via%3Dihub

Although not explicitly using the method of least squares, Quetelet was the first statistician that had the necessary skills to yield the theory of errors and apply it to social science. In 1823, he spent a sojourn at the Paris Observatory under Bouvard to acquire the knowledge to use the method. Quetelet believed statistics could be used as a unifying method, not only for astronomy, but for all sciences dependent on observation, including meteorology, biology, and the social sciences. His first statistical work (1826) utilised Belgian birth and mortality tables as the basis for the construction of insurance rates[39]. As far back as Lloyd's coffee house data and statistics have been a critical component to the insurance business because companies have to assess the premium as which they would, for example, accept a sea voyage.

Although Quetelet found himself in the perfect position to yield the least-squares method to social science, there is no evidence that he made use of the technique; he was generally more fixated on averages and the bell curve. When he looked at the analysis of conviction rates in 1835, there is mention of the method of ''least squares''[40] to obtain he precise numbers for conviction rates, but the source of the numbers is not clear, leading one to believe that he did not practically know how to apply the method to social science, which in itself is not an easy task.

Dr K Heym who is referred to as the 'creator of invalidity insurance science' and was working as part of a team of actuaries for the German railway pension fund, was the first to use the least-squares method in social science with an application to insurance in 1854. Heym used the method to calculate the probability of death using a least-squares method.[41] It seems that each time this technique was applied to social science, the quality of the implementation was hotly contested. The quality of said work has been debated and written about in 1860 by Dr Fischer, after which he proposes an adjusted least squared method; many more papers followed since[42].

In 1871, William Stanley Jevons presented a short schedule relating the price of corn to different harvest sizes, and although thought to have use least squares method, it was later proven wrong. Jevons was taught least-squares at UCL and was both an innovator in mathematical economics and a great statistician. If there was anyone to bring the techniques of the theory of errors into the broader social sciences, it was Jevons, but he did not, and just like Quetelet, he failed in this endeavour. However, it must be said that in his paper examining the condition of the coinage in 1868, although his analysis is not mathematically sophisticated, it was the informal equivalent of regression analysis, more specifically weighted least squares.

Jevons might have been the first to apply the theory of errors to the pricing or understanding of the pricing of assets. Still, another formidable statistician, F.Y Edgeworth was the first to use the least squared regression version rather successfully in his 1883 paper ''On the Method Ascertaining a Change in the Value of Gold''. Edgeworth was a leading economist and statistician at the time. In this paper, he makes a few critical remarks to the quality of Jevons past work. ''It follows that the geometric mean constructed by Professor Jevons has not much ground to stand upon''. The record shows that although being friends, Edgeworth and Jevons were critical of each other's work. In summary, demographers, insurers and economists, like Quetelet, Heym, Fischer, Jevons, and Edgeworth all had a role to play in bringing the method into social science. Their work was later instrumental in establishing statistical learning theory.

[39] https://sci-hub.st/10.1086/228630

[40] Quetelet, 1835, vol. 2, pp. 307 - 308; 1842, p. 105

[41] http://www.staff.uni-oldenburg.de/dietmar.pfeifer/Karup_Lebensversicherung.pdf

[42] https://www.jstor.org/stable/41135294?seq=1#metadata_info_tab_contents

Linear Regression

The popularisation of the least square method came later in the form of Galton, Edgeworth, Pearson, and Yule. Galton's concept of regression that can found in his 1869 book Hereditary Genius, was finally mathematically developed by Yule and Pearson, student-teacher collaborators. It was not just an adaption of linear least squares regression to a new set of problems; it also introduced new ways of looking at multivariate data and the mathematical formalisation of an association statistic called correlation. This unique formulation addressed the issues of heterogeneity found by Lexis in the 1800s when a bell-curve could not capture premature deaths by controlling for the effects of measured variables. Better yet, these techniques would return to astronomy and transform it on top of sociologic, genetic, and economics.

The concept of a measure of correlation had been introduced by Francis Galton in 1888[43], and after some input from F. Y. Edgeworth, Pearson[44] gave it definitive form in a series of articles in 1896[45]. The Galton adopted concepts of correlation and regression, were the most important ideas to come out of Pearson and Yule's work[46]. Looking at diminishing ancestorial effects, Galton also came across the concept of multiple regression after his 1877 address and published something in that tune a two decades later in the journal Nature[47]. Galton was never able to breed the mathematical formulae to capture this effect. Like with the concept of correlation[48], Galton laid the imaginative groundwork for Yule and Pearson to follow through with mathematical rigour. Pearson's 1896 series were one of the first examples of multivariate equations[49]. Multivariate equations also opened up the opportunity to include interaction terms between variables[50].

Apart from helping with the development of correlation research, which includes, among other things the coinage of ''correlation coefficient'', Edgeworth, kept asking interesting economic questions. In his *The Mathematical Theory of Banking* in 1888, he pioneered a study in inventory theory, showing how a record of past demands on the reserves of a bank could be used to predict the odds that a given level of reserves will be sufficient in the future, and most notably recognised the existence of serial correlation. He suggested that the behaviour might be similar to the moving average of independent random terms.

It is strange that before Galton re-popularised the least-squares method, it was sparsely used in social science and barely escaped the confined of astronomical and geodetic work, even though some early social scientistic like Adolphe Quetelet were also astronomers. Within the next few decades, Udny Yule and Karl Pearson brought this concept of regression into a more statistical context. Throughout the next few decades, Udny like Edgeworth took the applied economic lens and Pearson remained on the method side. Following on from his 1895[51] paper studying correlation

[43] Galton, 1888;

[44] https://sci-hub.st/https://doi.org/10.1098/rsta.1894.0003

[45] f "mathematical contributions to the theory of evolution" published in the Philosophical Transactions of the Royal Society

[46] http://www.economics.soton.ac.uk/staff/aldrich/hope%202010.pdf

[47] https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537

[48] Galton, 1888;

[49] https://sci-hub.st/https://doi.org/10.1098/rsta.1896.0007

[50] https://www.cambridge.org/core/journals/journal-of-agricultural-science/article/studies-in-crop-variation-ii-the-manurial-response-of-different-potato-varieties/326342EA3B72577737653A3D54741D46

[51] On the correlation of total pauperism with proportion of
out-relief, I: all ages. Economic journal 5: 603 -611.

patterns of pauperism, the first such application in economics, Yules 1899[52] establishes a multiple regression that investigates the causes of pauperism. This study has long remained the only multiple regression exercise for that era. It was the very first exposition and use of multiple regression in its least-squares form in the literature[53].

In 1901, Arthur Bowley in the first English language statistics textbook tried to capture this range of new techniques; the book was titled *Elements of Statistics,* which exactly a century later in 2001 would be succeeded by the *Elements of Statistical Learning*, the ultimate machine learning textbook by Friedman, Tibshirani, and Hastie.

In 1964 the first regression analysis called CAPM model took a similar form as a simple OLS regression[54].

$$\mathrm{R_{i,t}} - \mathrm{r_f} = \alpha_\mathrm{i} + \beta_\mathrm{i}\left(\mathrm{R_{m,t}} - \mathrm{r_f}\right) + \varepsilon_\mathrm{i,t}$$

To assess the alpha and the beta, you estimate the model by performing an OLS regression. An OLS regression is a supervised learning task where a target is being predicted based on independent variables. In 1973 there was an Arbitrage Theory of Capital Asset Pricing, leading to a multifactor model of asset pricing.[55] These early economists and sociologist were instrumental in not just using linear regression methods, but also by improving and popularising them for all other sciences.

Multivariate Analysis

Yule and Pearson's development of correlation and regression opened the way for multivariate analysis in the works of Fisher and Harold Hotelling. Fisher, a studious American PhD, had written to the greatest such as Edgeworth, Pearson, and Yule to invite them to the U.S. He was however slow to adopt their method of correlation. His first use was in *Purchasing Power of Money* in 1911. However, his colleague JP Norton used it in his thesis *Statistical Studies in the New York Money Market*, to which Yule responded that they study performed ''in a very able manner''[56].

Fisher was primarily responsible for the body of research called the analysis of variance, which includes hypothesis setting and significance tests, some of which was foreshadowed by Edgeworth[57]. H.L. Moore was another interesting character from the era, in 1917 he published *Forecasting the Yield and Price of Cotton,* his research later appeared in a review study Forecasting the Acreage, Yield, and Price of Cotton.[58] He, like others, sought the opinion of Pearson, who largely ignored him. However, Yule reviewed some of his work somewhat favourably. It is therefore quite remarkable that the statistical

[52] An investigation into the causes of changes in pauperism in England,
chiefly during the last two intercensal decades, I .
journal of the Royal Statistical
Society 62: 249 -
295.

[53] https://sci-hub.st/https://doi.org/10.1068/d291

[54] https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1964.tb02865.x

[55] https://www.top1000funds.com/wp-content/uploads/2014/05/The-Arbitrage-Theory-of-Capital-Asset-Pricing.pdf

[56] http://www.economics.soton.ac.uk/staff/aldrich/hope%202010.pdf

[57] "On Methods of Ascertaining Variations in the Rate of Births, Deaths, and Marriages," was read on September 12, 1885.

[58] https://drum.lib.umd.edu/bitstream/handle/1903/18044/DP70111.pdf?sequence=1

economics or econometrics approach entertained by Fisher, Norton, and Moore, did not find any footing in England except with Yule and Bowley[59].

The American branch started moving the field forward and innovating on multiple topics. Statistics and Economics were very much intermingled in the U.S. at this time. In 1928, 69% of American Statistical Association members were also a member of the American Economic Association[60]. Hotelling was probably the most prominent member; he was also to some extent the loudest bastion and follower of Fisher's ideas[61]. He worked at Stanford and then later went to Columbia to replace Moore as professor of economics.

In 1946, Gerhard Tinter, by essay to the American Statistical Association ''proposes to introduce the economic statistician to some of the newer methods of multivariate analysis.'' Here he introduces the use of PCA (principal component analysis), LDA (linear discriminant analysis), and CCA (canonical-correlation analysis) all remnants from Harold Hotelling who played a central role in the development of mathematical economics, initially inspired by Eric Temple Bell, and R.A. Fisher, who soon went on to inspire Milton Friedman, and Kenneth Arrow[62]. Harold Hotelling exemplifies the fact that state-of-the-art machine learning methods emerged out of the applied disciplines starting with his seminal 1933 paper ''Analysis of a complex of statistical variables into principal components.''[63] Not long after Tinter's introduction, Richard Stone took up the offer to publish a paper in 1947 on the interdependence of blocks of transactions[64].

Multiple regressions, in particular, was a boon in economics and finance literature at the time, with almost no problem left untouched. Even Harry Markowitz mean-variance portfolio weights introduced in 1952 can now use regression models to map returns directly into portfolio weights as a simple one-step procedure[65]. Hammond, in 1964 said that a multiple-regression technique could be used for inference in clinical studies. Ronen, in 1974 writes that Hammond's approach can be used by analysts to fit terms like P/E ratios, dividend yields, earnings to analyst's judgements to capture their personal weighting policy within the framework of a linear model[66].

A mere decade after Gerard Tinter's introduction to multivariate methods, it has reached financial markets with a Principal Component Analysis of stock market indices[67], and industry clustering research in the 1960s[68]. Multivariate analysis extended into the concept of dimensionality reduction practised today. It also includes the use of multivariate techniques for other purposes, for example,

[59] 1930a. British Statistics and Statisticians Today. Journal of the American Statistical Association 25:186–90.

[60] Biddle, J. 1999. Statistical Economics, 1900–1950. HOPE 31:607–51

[61] Bennett, J. H., ed. 1990. Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher. Oxford: Oxford University Press.

[62] https://sci-hub.st/10.2307/2280570

[63] https://psycnet.apa.org/record/1934-00645-001

[64] https://sci-hub.st/10.2307/2983570

[65] https://static1.squarespace.com/static/56def54a45bf21f27e160072/t/5a0d0673419202ef1b2259f2/1510803060244/The_Sampling_Error_in_Estimates_of_Mean-Variance_Efficient_Portfolio_Weights.pdf

[66] https://egrove.olemiss.edu/cgi/viewcontent.cgi?article=1715&context=dl_tr

[67] https://ideas.repec.org/p/cwl/cwldpp/175.html

[68] The only requirement to be called an unsupervised learning strategy is to learn a new feature space that captures the characteristics of the original space by maximizing some objective function or minimising some loss function.

Hotelling suggested the idea of Principal Component Regression in 1957[69]. Other innovations for added robustness followed like Partial Least Squares Regression in 1966[70]. To this day, it is an active and growing science.

Time Series Analysis

After researchers have moved from univariate least-squares methods, multivariate analysis, the next natural dimension to model is the time dimensions. As a result of being a well-researched and established science, time series analysis techniques are not commonly referred to as machine learning in modern times. However, many of the models are regression models that might, for example, use their own lagged values as inputs in the prediction model.

The first actual application of fitting autoregressive models to data can be brought back to the work of Yule and Walker in the 1920s. A few decades later, Herman Wold introduced to the removal of seasonal and other fluctuations by introducing a moving average in 1939, known as the AutoRegressive Moving Average (ARMA)[71]. The technique did not catch on until the 1970s due to difficulty in deriving a likelihood function to use maximum likelihood to estimate parameters. Box and Jenkins came around to solve the problem and published a popular book called *Time Series Analysis*[72]. The Box-Jenkins model that came from this book is one of the most popular time series models used extensively in finance and economics literature.

Many extensions have soon followed including Vector AutoRegressive (VAR), that are able to accept multivariate data. And Generalised and AutoRegressive Conditional Heteroscedasticity models (GARCH/ARCH) that are non-linear generalisations of the Box-Jenkins model that allow for the parametrisation and prediction of non-constant variance, making them very popular for financial time series leading to a Noble prize award for Granger and Engle in 2003. Additional innovations included models that possess regime-switching capabilities, like a Markov-switching GARCH model[73].

Logistic Regression

The logistic function was invented in the 19th century by one of Quetelet's students to describe the growth of populations[74]. Malthus (1789) said a population would increase in geometric progression, and Quetelet noted that this progression couldn't be sustained and there has to be some upper limit. He experimented with several adjustments and asked Verhulst his student to look into it. Verhulst published the first logistic function in 1838.

The roots of using the logistic function in a regression problem started with Fechner with research looking in human response to a stimulus (1860). It was later popularised by Gaddum (1933) and Bliss

[69] HOTELLING, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. Brit. J. Stat. Psychol., 10, 69-79

[70] https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=534296

[71] https://www.sv.uio.no/econ/english/research/networks/haavelmo-network/publications/files/TH1939g%20transl%20final.pdf

[72] https://books.google.co.uk/books/about/Time_Series_Analysis.html?id=5BVfnXaq03oC&redir_esc=y

[73] https://link.springer.com/article/10.1007/s001810100100

[74] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=360300

(1934). In 1935 Bliss published a brief note in science, introducing the term probit. An analysis that linked binary discrete response variables to discrete covariates became known as logistic regression.

Applications of the logistic regression appeared in economics and market research in 1954 when Farrel used a probit model to study car ownership of different vintages as a function of household income, and Adam in 1958 to fit log normal demand curves from surveys to the willingness to buy cigarettes at various prices. Computers at first did not have packaged routines for the maximum likelihood estimation (MLE) for logit and probit models, and in 1977 the first computer packages were released that offer this facility. MLE can be traced back to Edgeworth, and then Later Fisher. In 1969, Theil saw its potential as an approach to model shares. In 1973, McFadden linked the multinomial logit to the theory of discrete choice for which he earned the Nobel prize in 2000.

A further development was regularisation regression methods that allow one to deal with many potential covariates. Regularised regression was an essential technique for a world where not just the length of datasets was increasing, but also the width of datasets. Economists were slow to adopt most modern machine learning techniques, but they have shown some affinity for Ridge regression which, although discovered earlier, were fully developed in 1970[75]. The LASSO regression emerged later in 1996[76].

The Ridge regression was known initially as Tikhonov regularisation, by its namesake Andrey Tikhonov, and it is useful to mitigate the problem of multicollinearity in linear regression. The first recorded use was by Tikhonov himself in 1943 with ''on the stability of inverse problems'' where it was applied to integral equations[77]. Arthur Hoerl separately discovered it in 1959[78] in 1962 gave it its first statistical approach, from which point onwards it was referred to as ridge regression[79]. It became especially famous after his 1970s paper, ''Ridge Regression: Biased estimation for nonorthogonal problems.''

It has not been long since these methods were adopted into standard econometrics literature. In Phoebus Dhrymes, 1978 book, Introductory Economics, he introduces four remedies for dealing with multicollinearity two of which are Ridge regressions, and the use of principal components[80]. Even as early as 1985, it was hard to use these techniques in applied finance and economics, simply because no one wanted to make their code available. ''Although no major statistical computer packages perform ridge analysis, the author will gladly supply interested parties with either a Minitab macro or FORTRAN 77 source code''[81].

Instance-based Methods

Instance-based algorithms sometimes called memory-based algorithms, are all derived from the nearest neighbour classifier in 1967[82]. These algorithms learn training examples and then generalise to new instances based on some similarity metric. This method requires a large amount of memory,

[75] https://www.math.arizona.edu/~hzhang/math574m/Read/RidgeRegressionBiasedEstimationForNonorthogonalProblems.pdf
[76] https://www.jstor.org/stable/2346178?seq=1
[77] http://a-server.math.nsc.ru/IPP/BASE_WORK/tihon_en.html
[78] https://www.computerhistory.org/collections/catalog/102663238
[79] https://www.scienceopen.com/document?vid=77767d4d-bc39-4026-856b-9fdf2075fddb
[80] https://b-ok.cc/book/2127488/3b313c
[81] https://sci-hub.st/10.2307/2683926
[82] (Cover & Hart, 1967)

but can quickly adapt to new data, and the prediction time is swift. Each new point looks at the past labelled points that are closest to the new point, the nearest points are the neighbours, and they vote on what the new point's label should be.

Instance-based methods have extremely slow to percolate into economics, partly because of the discipline's distaste for nonparametric models, and partly because of small dataset problems. In 1990 researchers at the University of Akorn, studied which asset flow measures best classifies bond ratings. They used a linear discriminant analysis, nearest neighbour, and probit analysis[83]. In 1992 Miguel Delegado attempted to introduce the method in a paper, *Nonparametric and Semiparametric Methods for Economic Research*, trying to make the case that nonparametric models are useful even if only at an exploratory level because ''… nonparametric models make no precise assumptions about functional form. Instead, the data are allowed to speak for themselves''[84].

Instance-based learning can be used both for classification and regression. Instance-based learning has become more prevalent in recent years. It has been used to forecast U.S. interest rates in 1996[85], for forex prediction in 1997[86], and commodity forecasting in 1999[87]. Naturally, the nearest neighbours are not the only method-based instance method. Support Vector Machines have also been a popular model, and just three years after it was developed in 1995 Vladimir Vapkin, a practical stock prediction example appeared[88]. A few more methods are worth mentioning like Learning Vector Quantization, Self-Organizing Map, and Locally Weighted Learning.

Around this time, the nonparametric regression model became especially popular in quantitative hedge funds. In 1987, Rene Carmona in Renaissance Technologies started experimenting with them to model high-dimensional non-linear relationships. The methods took some time to reach financial academics, as the industry was keeping their cards close to their chest. In 2000 Andrew Lo published one of the first applications of kernel regression to finance. This is late considering that the runway for kernel regression has already been laid by both Nadaraya[89] and Watson[90] in 1964. However, by 1987 we have seen the application of kernel estimation in the context of econometric models[91].

Tree-based Methods

The simple age-gender cohort employed in actuarial tables of mortality offers a framework for thinking about decisions that date back several centuries. Manuel Lima's book, *The Book of Trees* traces the use of trees back millenia where it was used as a visualisation and mnemonic. The earliest reference to trees is most likely the reference to the *Tree of Knowledge* in the Book of Genesis.

Binomial option pricing models, for example, use decision tree analysis, decision trees are also used for real options analysis like expansion and abandonment option for capital budgeting[92]. Interest rate instruments can be priced with binomial trees. It has been used for decision making in finance and

[83] https://www.questia.com/library/journal/1G1-9363215/the-relationship-of-asset-flow-measures-to-bond-ratings
[84] https://sci-hub.st/https://doi.org/10.1111/j.1467-6419.1992.tb00151.x
[85] https://dlib.bc.edu/islandora/object/bc-ir:102901
[86] https://www.sciencedirect.com/science/article/abs/pii/S0169207097000010
[87] https://www.sciencedirect.com/science/article/abs/pii/S0165176599001561
[88] https://dspace.mit.edu/bitstream/handle/1721.1/9925/40121760-MIT.pdf?sequence=2
[89] http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=tvp&paperid=356&option_lang=eng
[90] https://pub.uni-bielefeld.de/record/2715511
[91] https://link.springer.com/article/10.1007/BF01972450
[92] https://www.jstor.org/stable/1926290?seq=1

organisations for a long time. A case study of the technique appears in a Harvard Business Review article in 1964[93].

Just like a logistic function is only a function without a classification problem, a decision tree is simply a decision tree if not hooked onto a regression or classification problem to supervise its learning. The first regression tree was invented in 1963, that utilised an impurity measure to split data in two subsets recursively[94]. A decision tree is a classifier that partitions data recursively into to form groups. The splitting of nodes is decided by algorithms like information gain, chi-square, Gini index.

The statistical antecedents of decision trees can be traced back to a 1959 article by William Belson, where he used a decision tree to match and predict population samples[95]. Since then, multiple supervised tree models have been formulated and developed using different algorithms and methods. In order, they have progressed through AID[96], ELISEE,[97] THAID[98], CHAID[99], and CART[100] models. In the last few decades, decision trees have been used for all sorts of reason, like measuring and predicting firm performance[101], and detecting fraud through financial statements[102], or predicting bankruptcy[103], or detecting concept drift for economic time series prediction[104].

Clustering Methods

Both clustering and dimension reduction techniques are multivariate methods[105]. Both methods establish a new feature space by maximising some criteria. Clustering looks at the commonality across features to reduce the number of samples to a smaller number of patterns, whereas dimensionality seeks to remove redundant columns from the data. They are known as multivariate methods because they both use multiple features to perform a transformation.

Clustering was originated in psychology by Zubin[106] in 1938 and Tryon[107] in 1939 and anthropology by Driver and Kroeber[108]. The approach Tyron used was the first attempt at an agglomerative clustering

[93] https://hbr.org/1964/07/decision-trees-for-decision-making

[94] https://washstat.org/presentations/20150604/loh_slides.pdf

[95] https://www.jstor.org/stable/2985543

[96] https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500855

[97] https://scholar.google.com/scholar_lookup?title=Le+programme+%C3%89lis%C3%A9e%2C+pr%C3%A9sentation+et+application&author=CELLARD+%28J.+C.%2C+LABB%C3%89+%28B.+et+SAVITSKY+%28G.&publication_year=1967

[98] https://agris.fao.org/agris-search/search.do?recordID=US201300497437

[99] https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2986296

[100] https://books.google.co.nz/books/about/Classification_and_Regression_Trees.html?id=JwQx-WOmSyQC&redir_esc=y

[101] https://www.sciencedirect.com/science/article/abs/pii/S0957417413000158

[102] https://link.springer.com/article/10.1186/s40064-016-1707-6

[103] https://ieeexplore.ieee.org/document/5952826

[104] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.5260&rep=rep1&type=pdf

[105] https://online.stat.psu.edu/stat555/node/15/

[106] ZUBIN, J. A. 1938 "A technique for measuring likemindedness. " 3ournal of Abnormal and Social Psychology 33 (October):508-516.

[107] TRYON, R. C. 1939 Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Ann Arbor: Edwards Brothers.

[108] DRIVER, H. E. AND KROEBER, A. L. 1932 "Quantitative expression of cultural relationships." University of California Publications in American Archaeology and Ethnology 31:211-256.

technique. These earlier innovations laid the path out for more experimentation, especially concerning the efficiency of various algorithms.

W.D. Fisher showed the application of clustering and aggregation in economics in 1969. In finance, industry clustering was one of the first inroads of unsupervised learning into machine learning as presented in a paper by B.F. King in 1966[109]. He applies factor analysis on the observed covariance matrix of a large set of monthly returns to explain the degree of cross-sectional interdependence and form clusters of similar companies[110].

The field grew tremendously with every researcher attempting to develop their clustering method. Cormack in 1971 critically reviewed the area and suggested that researchers start implementing comparative studies on cluster procedures instead of just formulating new ones[111]. In recent years, clustering has been applied to portfolio management with success[112].

Ensemble Methods

The principle of combining prediction has been of interests to scientists in many fields. It has particular relevance to finance, where multiple models are often used and weighted for asset valuation purposes, or multiple estimates are combined for earnings forecasts. Many methods exist, such as voting or linear combinations. The weak learners could also take various forms like linear regression models or decision trees. The ensemble revolution had a slow start only kicking off in 1990, with *boosting* models.

The algorithm underlying ensemble methods generally take on a boosting, or bagging element, or even a combination of both. Boosting trains small or weak models sequentially, with a new model trained at each round[113] of which Gradient Boosting is a particular variety with enhanced training capacity due to fitting the derivative of a cost function. Bagging trains each member of the sample from a different training dataset[114] of which Random Forest developed in 1995 is a variety that also randomly samples at the column level. Boosting and bagging can also be combined using models like XGBoost and LightGBM developed by Microsoft.

Model stacking also generally falls under the ensemble umbrella. With stacking a set of models are constructed, it could be a set of ensembles, of which their outputs are tested on a hold-out set to combine the set of outputs in such a way so as to improve the prediction quality, in effect creating a model weighting scheme or a meta-model[115]. Stacking has been successfully used on both supervised learning tasks and unsupervised learning[116] It has become the preferred modelling setup in data science competitions. Over the years, the innovation in ensemble models generally came from optimisation and robustness improvements. These days modern methods like XGBoost and LightGBM

[109] King, 1966

[110] The only requirement to be called an unsupervised learning strategy is to learn a new feature space that captures the characteristics of the original space by maximizing some objective function or minimising some loss function.

[111] https://www.jstor.org/stable/2344237

[112] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708678

[113] R. E. Schapire. The Strength of Weak Learnability. Machine Learning, 5:197–227, 1990

[114] L. Breiman. Bagging Predictors. Machine Learning, 24(2):123–140, 1996

[115] D. H. Wolpert Stacked Generalization. Neural Networks 5(2)., pages 241–259, 1992.

[116] https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265

do not need a lot of hyperparameter tuning. Ensemble models perform very well on tabular data, and conveniently so, finance and economic models have traditionally been built on the back of tabular datasets. Ensemble models have been used in financial event predictions like earnings surprise prediction, bankruptcy prediction, and restaurant facility closures[117]. They, to a large extent, give you the non-linear high-dimensional benefit of neural networks, without all the architectural fuss.

Neural Networks

Norbert Wiener's first reference to cybernetics was in a paper "Behaviour, Purpose, and Teleology published in 1943, with a fuller representation submitted in Cybernetics published in 1948. The immediate origin of cybernetics can be found in the development of calculating machines, guided missiles, and even the physiology of the brain. Wiener says that the problems of computing machinery for partial differential equations turned his mind to these channels in 1940. In the same year, W.R. Ashby published an article which tacitly assumed many of the ideas that later contributed to Cybernetics. Not long after, the perceptron was introduced by Frank Rosenblatt (1957), which is considered to be the first instantiation of the modern neural network[118].

Soon enough in 1975[119] and 1976,[120] Jerry Felson applied a perceptron pattern recognition technique that uses iterative probabilistic learning from previous decision-making experiences for investment selection and market forecasting purposes and shows that you can have above-average performance. He notably came to believe that when the decision parameters become large, e.g., exceeds four, human decision-making deteriorates, and therefore, the programming of an investment decision process is useful.

White (1992) attempted to popularise neural nets in economics in the early 1990s, but at the time, they did not lead to substantial performance improvements and did not become popular in economics[121]. Before White, Kimoto & Asakawa in 1990 after a long drought since Felson's papers in the 1970s, wrote an article to predict the stock movements on the Tokyo Stock Exchange[122]. A flurry of neural network papers followed as American researchers started to play catch-up with Japan[123]. Unlike some of the areas before, no real innovation came directly from finance. Instead, innovation in neural architecture occurred, leading to new applications in finance, some of which are described below.

Neural networks are especially beneficial for many financial applications. Recurrent neural networks help with financial time series forecasting, and it was first introduced by John Hopfield in 1982.[124] A little more than a decade later, a longer-memory solution was developed called LSTM[125]. This type of neural network is often compared to the performance of ARIMA and other mechanical time series models in finance and economics. Not a lot of faith has been placed in these models until they were

[117] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3481555
[118] http://homepages.math.uic.edu/~lreyzin/papers/rosenblatt58.pdf
[119] https://sci-hub.st/https://doi.org/10.1109/TSMC.1975.4309399
[120] https://sci-hub.st/10.1016/S0020-7373(76)80042-9
[121] https://www.amazon.com/Artificial-Neural-Networks-Approximation-Learning/dp/1557863296
[122] https://sci-hub.se/10.1109/ijcnn.1990.137535
[123] http://web.ist.utl.pt/~adriano.simoes/tese/referencias/Papers%20-%20Adriano/NN.pdf
[124] https://www.pnas.org/content/pnas/79/8/2554.full.pdf
[125] https://web.archive.org/web/20150526132154/http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

shown to beat traditional models in the M4 time series competition in 2018[126]. These models can also be easily adapted for classification tasks by adding a logistic or sigmoid output function.

The Neocognitron published in 1979 by Kunihiko Fukushima[127], eventually led to what we now refer to as convolutional neural networks. At first, these methods were ridiculed because they subjected time-series data to the format of an image. This has led to some researchers converting time series data to images using techniques like Gramian Field Encoding and others. Later research has shown that that is not necessary and that the raw data can simply be fed into the neural network architecture. Convolutional neural networks can also successfully be used for assessing visually useful information like technical indicators and limit order books, among others.

The Autoencoder was popularised by D.E. Rumelhart and Hinton in 1986[128]. It can be used for feature representation engineering and dimensionality reduction. The basic constructs of generative adversarial learning were developed by Schmidhuber in 1990, and later 'rediscovered' and popularised by others[129]. They can be used to generate new time series data; variational autoencoders can also be used to remove the noise from financial data as well as generate new time series data. Any of these neural network architectures can be used to approximate analytical functions to improve latency. Neural networks can be used to solve the partial derivative equations in analytical procedures. Neural networks can be used to calibrate financial models. Neural networks provide an excellent function approximator for complex reinforcement learning models. Neural networks are online models that can train with new datapoints, they can also entertain transfer learning, where the ensembles models cannot. This technology is still very new, and not much has been done in this domain. For example, see my post on *The Neural Landscape for Option* modelling for eight different ways in which neural networks can be used to model option derivatives.

Agent Learning

New inventions often inspire new ideas. Not long after computers were first conceived in the 1800s, inventors like Ada Lovelace philosophised about the potential of them becoming intelligent. In this section, we will look at agents that automatically take actions and learn from those actions.

Established figures like William J. Baumol in 1966 said that ''Portfolio decisions must remain an art, employing at its best a fine balance of experience, judgement, and intuition all of which only a good analyst can supply''. At the time, these discussions were often hypothecations due to the lack of authorities that having access to large parts of valuable data. However, since 1962, it was particularly fashionably for every financial institution to point out that they have a computer[130]. Competitive pressures would motivate these institutions to show their customers that they actively participate in the computer revolution.

A 1959 case study from Jesup & Lamont, a New York brokerage firm founded in 1877 showed how computers could be successfully applied to investment management[131]. Jesup collected data manually for 30 securities listed on the NYSE; the data was collected from published reports. By around 1960, their input programs could translate, interpret, and fill unformatted information received from news

[126] https://www.sciencedirect.com/science/article/abs/pii/S0169207018300785
[127] http://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf
[128] https://web.stanford.edu/class/psych209a/ReadingsByDate/02_06/PDPVolIChapter8.pdf
[129] http://people.idsia.ch/~juergen/FKI-126-90ocr.pdf
[130] The Wall Street Journal, April 23, 1962 p 4
[131] https://dspace.mit.edu/bitstream/handle/1721.1/47024/computernewpartn00amst.pdf?sequence=1

service wires. At about 1962, they were monitoring all common stocks traded on the NYSE, and it was useful in anticipating the market decline during April and May that year. The developers noted a few issues:

(1) The existing data sources could not be depended on, and it required frequent correction and manual review to reduce substantial errors.
(2) The output of the computers still far exceeded management's capacity to assimilate the information.
(3) Traditional financial performance measures were inadequate to capture market behaviour over extended periods of time.

''Until 1963 those interested in computer-aided investment analysis face an appalling data acquisition task. Fundamental data has to be manually culled from published reports, and technical data was also manually transcribed from published reports or encoded from wire service lines.'' In 1963 this changed when Standard Statistics Inc, a subsidiary of S&P's, released the *Compustat* tapes containing fundamental data compiled annually from 1947 for 900 companies, also in 1963 Scantlin Electronics, Utronics, and Bunker-Ramo began offering price, volume, earning, and dividend data for all listed securities.

This allowed Jesup to fix many of their problems by around 1964. Their new design included the ability learn so as to contribute to future refinements, heuristics were incorporated to evaluate alternative investment approaches, and an adaptive design allowed the procedures to change over time, leading to new models and processes that contributed to the systems future success. Therefore, unlike other rule-based methods around the time, there was self-learning that occurred here. It also provided more generally a source of data for manual investigation, and an approach to validate the firm's investment models.

The firm implemented parametric and nonparametric statistical routines to facilitate the analysis of fundamental and technical relationships. They realised the importance of ''Real-Time Data Acquisition'', that would give analysts the ability to evaluate relevant market conditions as they developed. Using the validated analytic procedure, i.e., backtests, the system would recommend actions, i.e., buy, hold, sell, sell short. The system was able to recreate conditions existing since 1959 and test the effectiveness of an alternative analytical approach to determine the performance of a particular decision or rule during a specified period.

This recursive feedback for optimal performance also received some attention in academia. In 1966, Merton Miller introduced a dynamic policy to model the demand for money by firms[132], three years later in 1969 Daellenbach, formulated the optimal cash balance for a bank as a dynamic programming problem to minimise the sum of the variable transaction costs and interest costs to obtain an optimal policy[133]. This formalisation of dynamic programming is the start of a long line of literature in reinforcement learning were recursive feedback is an essential part of the learning process.

Here we, therefore, advocate for agent learning to be any system that self-learns through recursion. Within this definition, a customised self-learning function fits the bill, as well as more mathematically sophisticated methods like dynamic programming. Dynamic programming had some early advocates in finance, for example, Leonard Baum of the acclaimed Baum-Welch algorithm used to find unknown parameters in a hidden Markov model (HMM), in 1979 he was hired into Monemetrics by Jim Simons

[132] https://sci-hub.st/10.2307/1880728
[133] https://sci-hub.st/https://doi.org/10.2307/2329701

now famous for Renaissance Technologies to try some of these mathematical techniques to the financial market.

Reinforcement learning as we know it today also greatly benefited from Christopher Watkins who in 1989 developed Q-learning. It also soon was shown possible to use neural networks in reinforcement learning with one of the first applications looking at controlling an inverted pendulum[134].

Present State

For 40 years from 1960 to 2000, quantitative financial methods have primarily been performed and improved behind the veil of corporate enterprise. The tide has slowly turned with a new funds launched in the late 90s such as AQR capital management that emphasised the publication of original research. In recent years, we have also seen the Winton Group and Two Sigma two relatively new fund started in 1997 and 2001 respectively, developing data science competitions not just as a recruiting tool, but also as a way of getting machine learning feedback from the crowd. In the 2000s, we have also experienced a larger push for open-source software for machine learning with software like Scikit-Learn and Torch. Funds came to realise that creative solutions can be found wherever data scientists have access to large datasets. It is true that small scale researchers cannot push efficiencies attributable solely to data processing capabilities, but they have ample skills in identifying patterns in large datasets.

It wasn’t long since the academic research and industry interest gained popular interest. In 2013, Tucker Balch, then at Georgia Tech now at JP. Morgan, published one of the first free online courses in computation finance on Coursera. In 2015 he published a course on Machine Learning for Trading via the Udacity platform. The field has been given a second breath in 2018 with industry insiders publishing more elaborate content like Marco Lopez de Prado’s Advances in Financial Machine Learning. Institutions like NYU, Cornell, and Columbia has also moved towards a more machine learning based curricula. The amount of educational and open source content is increasing exponentially, and research interest is back in full swing with ICAIF one of the world’s first conference on AI in Finance taking place last year. This is just a quick summary of the current state, I will at some point publish a three-four pager on the subject.

[134] Anderson, C. W. (1989). Learning to control an inverted pendulum using neural networks. Control Systems