

What Data Scientists Can Learn from History

Aaron Lai

Abstract We argue that technological advances and globalization are driving a paradigm shift in data analysis. Data scientists add value by properly formulating a problem. A deep understanding of the context of a problem is necessary because our incomplete answer will be worse than incorrect—it is misleading. Therefore, we propose three innovative analytical tools that define the problem in a solvable way: institution, data, and strategy. Afterward, we use three historical examples to illustrate this point and ask “What would a ‘typical’ data scientist do?” Finally, we present the actual solutions and their business implications, as well as data mining techniques we could have used to tackle those problems.

1 Introduction

Benjamin Disraeli said “What we anticipate seldom occurs; what we least expected generally happens.” As the volume of data grows exponentially, quantitative analysis, statistical modeling, and data mining are becoming more important. Predictive modeling is the use of statistical or mathematical techniques to predict the future behavior of a target group. It is different from forecasting in that forecasting uses time-series data to forecast the future. Predictive models are independent of time¹ so it will only be affected by random factors. Predictive modeling assumes that people, as a group, will behave in the same way given the same situation. The variations or errors are caused by an individual’s unobserved characteristics.

Part of the material of this article is based on my presentation titled “Predictive Innovation or Innovative Prediction?” for the Predictive Analytics Summit held in San Francisco in November 2010. Only the Powerpoint version was distributed to the participants. This paper has not been submitted to any other places. All opinions are my own personal views only and do not necessarily reflect those of my employer or my affiliation.

¹ In technical terms, they are called stationary.

A. Lai (✉)
Market Analytics, Blue Shield of California,
San Francisco, CA, USA
e-mail: aaron.lai@st-hughs.oxon.org

Of course, prediction is not the only thing a data scientist will do. Data science, a new yet undefined term, is to make sense out of data. It could be statistical analysis, algorithmic modeling, or data visualization. High volumes of data, which is commonly known as Big Data, require a new approach in problem solving. To succeed, we need an innovative approach to data analysis.

In this article, we argue that model building processes will be changed due to technological advances and globalization of talents. We analysts add value by a creative adaptation of modeling and an innovative use of modeling. It is the survival of the fittest and not the survival of the strongest!

As Louis Pasteur said centuries ago, “Chance favors prepared minds.” Predictive methods, when used properly and innovatively, could result in sparkling outcomes. Competitive pressure will make it just too important to leave it to non-professionals. It is very common for a half-knowing analyst to jump into the labyrinth of modern tools without thinking. It is thus essential to be innovative.

We look at the model building process from three angles: Institution, Data, and Strategy. We will use three historical examples to illustrate this point by asking, “What would a ‘typical’ data scientist do?” It is not uncommon for an inexperienced analyst to blindly apply what he or she learned from the textbooks irrespective of the root cause. We will contrast our “default” answers to the ingenious historical solutions. In describing the aftermath of the Long-term Capital Management fiasco, Niall Ferguson wrote “To put it bluntly, the Nobel prize winners had known plenty of mathematics, but not enough history. They had understood the beautiful theory of Planet Finance, but overlooked the messy past of Planet Earth. And that, put very simply, was why Long-Term Capital Management ended up being Short-Term Capital Mismanagement”. [4, p. 329] Andrew Lo of MIT used another “P envy” [16] that echoed Ferguson’s comment as it was titled “WARNING: Physics Envy May Be Hazardous To Your Wealth!”

The model development cycle is being compressed at an unprecedented speed. This is due to three factors: technological advance, outsourcing, and innovation diffusion. The latest statistical or data-mining software can easily replace a team of analysts. For example, SAS has a fully automated forecasting system that can create and fit a series of ARIMA models; Tableau analyzes data and suggests what type of chart would be most appropriate. Since we live in a global village, if I can formulate the problem in an equation or write down a specification, I can recruit an expert across the world to solve it. There are many sites or companies that allow people to pose questions and source answers. Crowdsourcing makes geographical limitation irrelevant. The last factor is an escalating pace of innovation as news travel fast—the latest techniques could be instantly imitated.

2 Case I: The War Chest

This was 1694 England. The Crown was under severe financial pressure and no easy answer was in sight.

2.1 Background

The main source of income of William the Conqueror since 1066 was the possession of royal properties (Royal Demesne) and the feudal system of land tenure (Feudal Aids). Feudal Aids was the right for the King to levy a tax for his ransom should he be taken prisoner by an enemy (thus we have the term the King's Ransom). This land tax system had been abused by the Crown so much that the nobles needed to create the Magna Carta to protect the lender's right. Customs were invented in 1643 after adopting the Holland system of excise taxes. The first record of currency debasement in England (decreasing the amount of precious metals and thus lowering the value of the coins) is from the reign of Edward I in 1300. There were many subsequent debasements. The metal content of the same coin dropped to only one-seventh from the beginning to the end of the reign of Henry VIII!

Henry III had the first recorded debt. Since interest payment was forbidden (usury), the Crown only needed to pay back the principal in those early days. During the Hundred Years War (1337–1453), Henry V had incurred so much debt that he would need to secure his debts by securities such as tax and jewels in 1421. In the twentieth century, those securities were called revenue bonds and asset-backed securities. Henry VIII defaulted on his loans several times by releasing himself from repaying those borrowed monies while Elizabeth I had excellent credit (could borrow at 10 % interest from Antwerp) and she finally paid all her loans.[5, p. 61, 67, 70, 72–74]

The financial situation was indeed very challenging in 1690s. William of Orange arrived in England in 1688 and England was at war with France in 1689 for the Nine Years War (1689–1697). The credit of the Crown remained weak until the Glorious Revolution of 1688 institutionalized the financial supremacy of the Parliament. The Parliament controlled new taxes and limited the power of the King. The whole system changed from the King to the King in Parliament and thus it established the financial superiority of the Parliament. One of the financial revolutions was to make notes transferable [19]. The governmental expenditure increased from £ 0.5 million in 1618 to £ 6.2 million in 1695 while debt increased from £ 0.8 million in 1618 to £ 8.4 million in 1695 [19]!

2.2 Problem Statement

Governmental debt was increasing at an astonishingly high rate. Even after some costly wars in continental Europe, there were no signs that any kind of peace would come soon. The King and Country needed a lot of money to finance military build-up and prepare for the next war.² The Crown had recovered his credit standing and thus was able to borrow more. In 1693, there was a large long-term loan (£ 1 million) secured by new taxes but it was almost immediately exhausted by 1694 [19].

² In fact the War of the Spanish Succession (1702–1713) was just around the corner.

The creditors were growing uneasy about the debt level and they demanded interest rate as high as 14 % in 1693 and 1694 [19]. Since those debts were “asset-backed securities”³, the HM Treasury officials had already used up high quality assets to do credit-enhancement.

2.3 *What if We Were There?*

Government revenue comes from two sources: tax and borrowing. Following a standard modeling approach, we could create an econometric model to investigate the elasticity of taxation. We could also use a segmentation model to put citizens/institutions into buckets, since they all had different coefficient of elasticity. A tax maximization policy would tax the most tax inelastic groups, subject to their ability to pay. It would be a typical constrained optimization exercise.

On the borrowing side, we would have to estimate the borrowing capability for our sovereign debts. We might run some macroeconomic models to assess our financial strength so as to present a credible plan to convince the market of our credit worthiness. There are only three ways a country can handle her debt: grow out of it, inflate over it, or default on it. Of course the investors hate the last two options. Thus it is the job of the Chancellor of Exchequer to make a convincing case.⁴ This is also why the central banks need to be considered as independent so that their will to fight inflation is strong.

2.4 *The Endgame*

Two important innovations helped drive down the borrowing cost and increase the borrowing capability. The first was the invention of fractional reserve by goldsmith-bankers and the second one was the incorporation of the Bank of England. During the medieval time, people stored gold and other valuables in the vault protected by the goldsmiths. The depositor received a certificate that could be redeemed on demand. Since only the goldsmiths knew the exact amount in a vault, they found that they could lend money (by issuing certificates, just like the Certificates of Deposit we have now) without doing anything [1]. The goldsmiths could then lend a substantial amount of money to both the Crown and the public. They also used reserve ratio and loan diversification to manage risk; operation risk for the former one and credit risk for the latter one. In the case of Sir Francis Child, he maintained 50–60 % reserve-to-asset ratio and diversified his lending to the general public and various Crown debts backed by different revenue stream such as Customs, Excise, East India Goods, Wine

³ They were backed by additional excise and duties on imports respectively.

⁴ For an explanation on the history of the Bank of England could help understanding the eurozone crisis, refer to [13].

and Vinegar etc. The increasing use of discounting (delay payments in exchange of a fee) by bankers like Sir Francis facilitated the circulation and liquidity of long-term debts. Discounting also allowed them to shorten the term structure of their liabilities [21].

Given the insights of using fractional reserve to increase the loan (i.e. money) supply and using high quality assets to enhance investment attractiveness, we could reformulate this problem into a portfolio optimization exercise. Following the standard mean-variance approach pioneered by Markowitz, we could create efficient portfolio of assets based on risk and return, as well as the inter-asset covariance. Many optimization algorithms could help solve this problem and a classical solution is quadratic programming. An alternative approach to optimization is econometrics modeling. We could use discrete choice analysis to find out who is going to buy what type of asset. In addition, Monte Carlo simulation and Agent-based Modeling (ABM) could also be employed. This kind of approach would allow us to model the dynamic interactions and inter-agent interactions in various consumption and preference trade-offs.

In modeling a solution, we need to be aware of the principal-agency problem as perceived by the investors. The HM Treasury served at the pleasure of the King and it was not there to serve the investing public. Therefore, any solution needed to be a credible solution from the point-of-view of the investors; they needed to be reassured that the government was determined to repay her debt. People said, "It is not about the return of money; it is about the return of my money."

The subscribers of government debts were invited to incorporate as the Bank of England in 1694. The Bank was responsible for handling the loans and the promised distributions. One of the most important characteristics was that the Bank could not lend the Crown money or purchase any Crown lands without the explicit consent of the Parliament [19]. To further lower the risk of the lenders, the government created a separate fund to make up deficiencies in the event that the revenue earmarked for specific loans was insufficient to cover the required distribution [19].

Government needs money and wars need a lot of money. The ability to borrow a large amount of long-term money cheaply was the reason that Britain beat France and emerged as a major power of the world [19]. Finance was so important that the Prime Minister was also the Chancellor of the Exchequer until the eighteenth century. The modern Chancellor of the Exchequer is always the Second Lord of the Treasury (No. 11 Downing Street) while the Prime Minister is still the First Lord of the Treasury. The official sign is still nailed to the front door of No. 10 Downing Street. These two innovations fundamentally changed the financing ability of Britain and that led to centuries of British Empire, especially for the funding of an expensive Royal Navy. The Bank of England became so prominent that it even had a nickname "The Old Lady" since 1797. Institution arrangement is very important to economic development, and Douglass North received his Nobel Prize because of his contribution to this area [25, p. 21] (Fig. 1).

Given the incomplete nature of old data, it would be difficult for us to assess the situation via quantitative analysis. However, researchers have [25] built a VAR (Vector Autoregressive) model to study the dynamics of the determination of interest rate on government debt from 1690 to 1790. They found that industrial revolution,

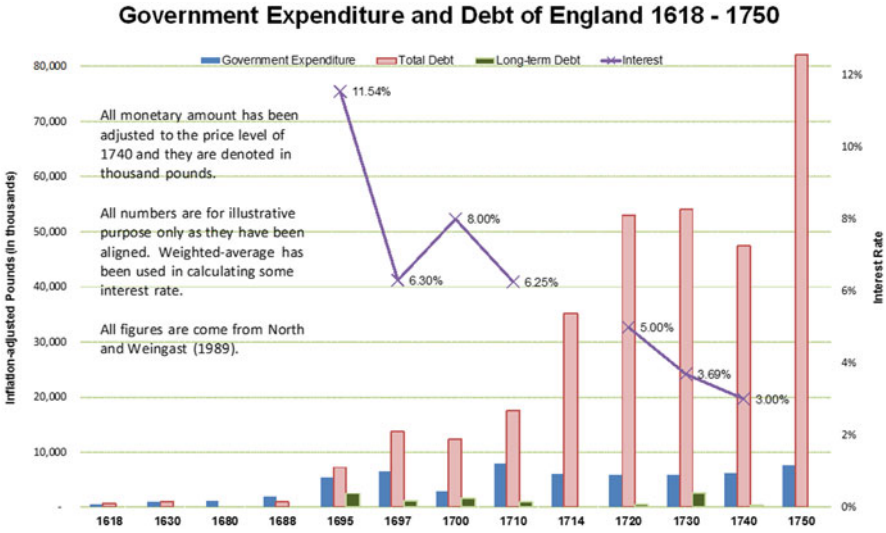


Fig. 1 Debt of England based on the figures listed in [19, pp. 820, 822, and 824]

military victories, and institutional reforms contributed a lot, especially the flight of capital from Napoleon’s reign.

2.5 Business Implication

GMAC was an example of an institutional innovation. It was originally created as a wholly owned subsidiary of General Motors to provide financing support to GM dealers. With this new institution, GM could offer incentive car loans to customers or dealers with very low interest rates. The increased sales further lowered the production cost (average fixed cost from the economies of scale) of a car. This kind of institutional arrangement has become a standard practice in the automotive industry. Now all major car manufacturers have subsidiaries to do automobile financing. The same idea has been extended to private label credit cards and other manufacturer financing. Data mining could help determine the optimal asset allocations for both the parent and the spin-off. Financial engineering can also decide the best capital structure and borrowing level.

3 Case II: London Outbreak

This was an ordinary August day (24th) in 1854. Mrs. Lewis of 400 Broad Street was washing her baby’s diaper in water, and she subsequently emptied the water into

a cesspool in front of the house. Little did she know that this simple action would cause 700 deaths within a 250-yard radius of a nearby water pump since her baby was infested with cholera [18].

3.1 Background

England was in a state of panic as there were over 20,000 deaths in England and Wales in 1853–1854. Asiatic cholera reached Great Britain in October 1831 and the first death occurring in that month was at Sunderland [7]. Cholera was first found in 1817. It caused 10,000 deaths out of a population of 440,000 in St. Petersburg in August 1831.⁵ Even though it had been researched extensively in a previous India outbreak⁶, no one really knew much about the disease and the Russians had even offered a prize for the best essay on *cholera morbus*. Miasma (spread via air) was the prevailing theory of transmission for the greater part of the nineteenth century. The irony was that even though sanitarians' casual theory was incorrect, they were able to demonstrate how and where to conduct the search for causes in terms of the clustering of morbidity and mortality. Jakob Henle argued in 1840 that cholera was caused by minute organism, and John Snow's works in 1849 to 1854 were consistent with this theory. Unfortunately, nothing until Louis Pasteur's experiment in 1865 could the establishment accept infectious disease epidemiology [24]. Snow questioned the quality of water, and after performing some microscopic works, he was not able to find the cholera micro-organisms [9, p. 99].

3.2 What if We Were There?

Snow was a very analytical person and is one of the pioneers of analytical epidemiology. William Farr, an established epidemiologist at that time, realized that the "Bills of Mortality" would be much more amenable to analysis when they contained variables in addition to names and parishes. His reports published in mid-1840s counted deaths not only by 27 different types of disease, but also by parish, age, and occupation. Snow used Farr's data to investigate the correlations among them.

If we were there, we could develop some logistic models with all variables to see if we could support or refute the prevalent theories⁷. However, we would have difficulties in developing a comprehensive model because we could not directly test both the contagion and the miasmatic hypotheses. And according to sanitarians, organic matters were not the direct causes of disease themselves, but as raw materials

⁵ p. 1, 16 [8].

⁶ Just the Madras volume ran to over 700 pages, p. 30 and 31 [8].

⁷ In fact, a paper used a logistic model on the Farr data and it rejected the Farr theory that cholera was caused by elevation [2].

District	Deaths from cholera in 1849 per 10,000 inhabitants	Elevation above high water (feet)	Annual deaths from all causes 1838 - 1844 per 10,000 inhabitants	Persons per acre	Persons per inhabited house	Average annual value of house (£)	Annual value of house per person (£)	Poor rate precept per pound of house value	Water supply ^a
Newington	144	-2	232	101	5.8	22	3.788	0.075	1
Rotherhithe	205	0	277	19	5.8	23	4.238	0.143	1
Bermondsey	161	0	264	66	6.2	18	3.077	0.134	1
St George	164	0	267	181	7.0	22	3.318	0.089	1

Fig. 2 Eight possible explanatory variables [2, p. 389]

Explanatory variable	Low 95% CL	Odds ratio	High 95% CL	P
Constant	6.006×10^{-4}	0.002626	0.01149	-
Water from Thames between Battersea Bridge and Waterloo Bridge ^a		1.00		<0.001
Water from New River and Rivers Lea and Ravensbourne	0.44	0.59	0.79	
Water from Thames between Kew and Hammersmith	0.22	0.40	0.72	
Increase in elevation above high water (10 feet)	0.85	0.91	0.98	<0.01
Decrease in poor rate (£/100)	0.87	0.91	0.96	<0.001
Average annual death rate 1838 - 1844	1.00	1.00	1.01	0.48
Persons per inhabited house	0.89	1.03	1.19	0.71
Persons per acre	1.00	1.00	1.00	0.67
Average house value per person (£)	1.00	1.00	1.00	0.35
Average house value within district (£)	1.00	1.00	1.00	0.79

^a Baseline.

Fig. 3 Logistic regression results [2, p. 392]

to be operated upon by disease “ferments” presented in the atmosphere during epidemics [20]. The significance results from miasmatic research at that time could be caused by the spurious correlation problem. Spurious correlation is the appearance of correlation caused by unseen factors.

Figures 2, 3 and 4 provide some tables and results from [2]. This model shows that poverty is the most significant factor!

3.3 The Endgame

Dr. Snow marked each death on the map as an individual event⁸ rather than a location of death. He did find that all deaths were within a short walking distance from the pump. Secondly, he made another map to show that those deaths were indeed closer to the Broad Street pump than the others [10]. Thirdly, he obtained water samples from several pumps in the area but the Broad Street water looked cleanest. Furthermore, he had two “negative data” points that supported his case: no deaths in the Lion Brewery (workers drank the beer) and the workhouse (which had its own well) [18].

⁸ Many people, including Edward Tufte and the CDC, took E.W. Gilbert’s version of map (with dots instead of bars) as John Snow’s original maps.

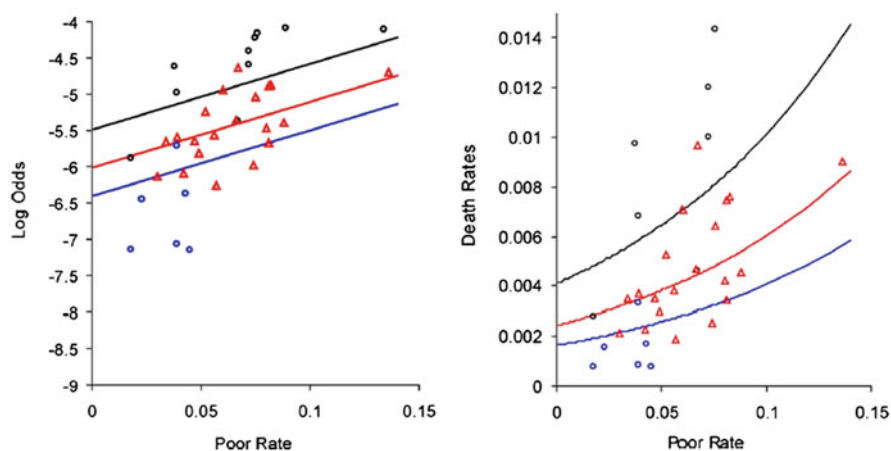


Fig. 4 Odds diagrams [2, p. 392]

The success of Snow's hypothesis rested on its narrow focus, while the Board of Health had a general hypothesis only. Snow's predictions were so specific that only a few observations were contradictory. Snow personally investigated on-site for those contradictory observations (e.g. brewery and workhouse) until he was satisfied with them. His hypothesis was also consistent with clinical observation. Snow insisted that the disease was gastrointestinal and all symptoms could be explained by fluid loss from the gastrointestinal tract. This led him to conclude that the infecting agent was oral and not respiratory [18, 20].

This is an important point for data scientists because our results or conclusions have to be consistent with all other information, both within and outside our model. The results need to be not only statistically and logically sound, but also must be consistent with observation. If you find something that contradicts to common sense, it is more likely for you to have made a mistake than to have discovered a new world.

Henry Whitehead did a survey to try to refute the conclusion of Snow. However, his results were in fact confirmed Snow's analysis. For those who drank water from the Broad Street pump, 58 % developed cholera compared with only 7 % of those who did not. Snow found that the mortality was related to the number of people who drunk from the pump during the infested period (from the date of washing the infested diaper to the removal of the pump handle). Another engineering survey⁹ concluded that there had been a consistent leak from the cesspool to the pump shaft [18]. For a more detailed discussion on the contribution of John Snow to analytical epidemiology, see [14].

⁹ The Board opened up the brick shaft but it seemed perfectly in order.

3.4 Business Implications

Google Maps has opened many possibilities of marrying data and geographical information. People create map-based websites ranging from restaurant guides to Haiti disaster relief.¹⁰ It is impossible to underestimate the impact of seeing information displayed on a map! This is the power of Data Innovation—collecting, using, and displaying data in innovative ways.

A recent BBC report¹¹ showed that Google, Microsoft, and Apple were all eyeing the rapidly growing spatial information market. We predict that spatial analysis and data visualization will gain lots of momentum when our infrastructure could support collecting, storing, and analyzing vast amount of data everywhere anytime.

Nevertheless, data visualization provides hints to the solution but cannot be the solution itself. Given almost identical information (even similar maps), the Board and Snow arrived at completely different conclusions. Why? It was because the Board analyzed the situation through a conventional len. They were all distinguished scholars or practitioners, and they fitted the facts into the model rather than retrofitting the model for the facts. The success of Snow rested on his particular attention to anomalous cases [10]. It is very common for us to downplay the importance of outliers rather than drilling down to the root cause of those “unfitted” observations. We tend to blame the customers for not behaving as our model predicted and not acknowledging it as a limitation of the model. The same rationale can be extended to financial model development as well [15].

When we perform spatial analysis and data visualization, we need to be careful that we are convincing rather confusing our audience. As explained in a New York Times article, an Army platoon leader in the Iraq war could spend most of his time making PowerPoint slides [3].

4 Case III: A Tale of Two Navies

This was 1904. Russia under the Tsar was an established European power with high self-image while Japan was a rising industrial power in Asia after victory in the Sino-Japanese War (1894–1895).

4.1 Background

Russians did not think highly of the Japanese navy because 50 years earlier Japan had no fleet at all. The Russian Foreign Minister, when asked about the possibility of

¹⁰ Dan Mascia wrote in January 14, 2010 for Fast Company titled, Haiti Earthquake Disaster: Google Earth, Online-Map Makers, Texts “Absolutely Crucial” <http://www.fastcompany.com/blog/dan-macsai/popwise/haiti-earthquake-google-maps-web-tech>.

¹¹ “Tech giants compete over mapping” from BBC Click, August 10, 2012.

<http://www.springer.com/978-3-319-07811-3>

Real World Data Mining Applications

Abou-Nasr, M.; Lessmann, S.; Stahlbock, R.; Weiss, G.M.
(Eds.)

2015, XVI, 418 p. 144 illus., 96 illus. in color., Softcover

ISBN: 978-3-319-07811-3