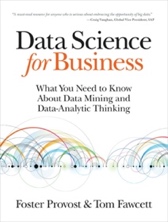
**Topic 1: Introduction to Data Science and Big Data**

* + 1. **Data analytic thinking (Ch. 1)**



1. **Discuss the ubiquity of data opportunities.**
2. Targeted marketing
3. Online advertising
4. Recommendations for cross-selling
5. Analyze customer behavior
6. Credit scoring
7. Trading
8. Fraud detection
9. Supply chain management
10. **Compare and contrast data science, engineering, and data-driven decision making.**

Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data.

“Data” engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state.

1. **Explain data and data science capability as a strategic asset.**

Data as Strategic Asset. Analytics is about having the right information and insight to create better business outcomes. It means the ability to can quickly access the right data points to evaluate key performance and revenue indicators in building successful growth strategies.

1. **Describe data analytic thinking.**

﻿When faced with a business problem, you should be able to assess whether and how data can improve performance.

﻿Firms in many traditional industries are exploiting new and existing data resources for competitive advantage.

1. **Compare data science and the work of the data scientist.**

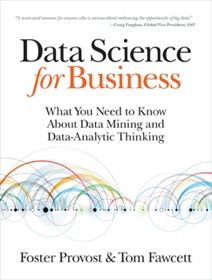
Data analystssift through data and seek to identify trends. What stories do the numbers tell? What business decisions can be made based on these insights? They may also create visual representations, such as charts and graphs to better showcase what the data reveals.

Data scientists are PROs at interpreting data, but also tend to have coding and mathematical modeling expertise.

Most data scientists hold an advanced degree, and many actually went from data analyst to data scientist. They can do the work of a data analyst, but are also hands-on in machine learning, skilled with advanced programming, and can create new processes for data modeling. They can work with algorithms, predictive models, and more.

A data scientist is someone who can predict the future based on past patterns whereas a data analyst is someone who merely curate meaningful insights from data.

* + 1. **Business problems and data science solutions (Ch. 2)**



1. **Describe how one transitions from business problems to data mining tasks.**
2. Classification. Scoring and probability estimations.
3. Regression. Predicting outcome.
4. Similarity matching.
5. Clustering
6. Co-occurrence grouping. Association rule. Onions and mustard for hotdogs. What items are purchased together.
7. Profiling
8. Link prediction. Link movies to people. Form recommendation system.
9. Data reduction
10. Casual modeling
11. **Compare supervised and unsupervised methods.**



Note: A teacher supervises.

Metaphorically, a teacher “supervises” the learner by carefully providing target information along with a set of examples. Labelled inputs.

An unsupervised learning task might involve the same set of examples but would not include the target information. The learner would be given no information about the purpose of the learning but would be left to form its own conclusions about what the examples have in common.

Supervised

Classification

Regression

Casual modeling

Both

Similarity matching. For recommendations.

Link prediction. Friend suggestions on Facebook.

Data reduction

Unsupervised

Clustering

Co-occurrence. Onions and mustard for hotdogs.

Unsupervised: “Do our customers necessarily fall into groups?”

Supervised: “Can we find groups who are likely to cancel their service soon after their contracts expire?”

1. **Describe the difference between data mining and using the results of data mining.**

Data mining is used on an existing dataset to find patterns. Historical data.

Using the results is for machine learning. The machine is trained on a 'training' data set, which teaches the computer how to make sense of data, and then to make predictions about new data sets.

1. **Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling and evaluation.**

A close up of a logo

Description automatically generated

Business understanding

Be creative. Think about the problem to be solved.

Data Understanding

Costs of the data will vary.

Data Preparation

Converting to tabular format. Removing spaces. Normalize data or scale it. Leakage can be a problem. This is when information is in historical data but not available when decision has to be made. Example: how many web pages were surfed during a session.

Data Modelling

Techniques and algorithms are applied to the data. Short section in book.

Data Evaluation

Assess the results. Look for patterns. Make sure they are valid.

* + 1. **Defining Big Data Guida**

A close up of a sign

Description automatically generated

* + - 1. **Discriminate between alternative data and big data**

Alternative Data

Alternative data refers to data used to obtain insight into the investment process. These data sets are often used by hedge fund managers and other institutional investment professionals within an investment company.

Big Data

Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

* + - 1. **Contrast drivers of adoption of alternative data with its challenges in the investment community.**

Drivers

﻿This is evidenced by the number of buy side firms actively seeking and researching alternative data sources. However, the adoption of alternative data is at the cusp of transitioning into an early majority phase as we observe a larger number of asset managers, hedge funds, pension funds, and sovereign wealth funds setting up alternative data research capabilities.

Challenges

﻿The task of evaluating, processing, ensuring legal compliance, and procuring a large number of datasets requires an overhaul of existing processes and can represent a significant organizational challenge.

﻿For example, at a large fund the procedure often requires (i) due diligence on the new data provider, (ii) signing legal agreements for (in most cases free) test data, and (iii) approval by compliance teams.

* + - 1. **Identify the largest categories of alternative data types in use today.**

A picture containing drawing

Description automatically generated

1. **Evaluate the usefulness of an alternative data set.**

The criteria:

1. Data history length
2. Data frequency
3. Universe coverage
4. Market Obscurity
5. Crowding factor
6. Uniqueness
7. Data quality
8. Annual Price

1. **Describe the likely attributes that differentiate alternative data sets in terms of COST.**

﻿To an asset manager, the value‐add of a dataset will be dependent on many factors, such as investment strategy, time horizon, universe size, and many other factors that will be unique to a fund manager strategy. The marginal alpha of a new alternative dataset could be too small if the new data source is highly correlated with datasets already used by an asset manager.

﻿

The first is that in many cases new providers' understanding and knowledge of peer or comparable data subscription pricings is non‐existent or very limited.

Second, data providers do not know how their data will be used by the buy side and how much value or alpha a dataset provides to an asset.

**A close up of a device

Description automatically generated**

1. **Discuss some of the most prominent alternative data trends.**

1. ﻿Alt data is that it is applicable to all asset classes and not just to listed equities,
2. ﻿In 2017 we saw a large increase in location, web, and app tracking sources. Forty per cent of all new commercially available sources
3. ﻿With regard to demand, the top categories enquired about were ESG, Transactional, Sentiment, and Economic data in the majority of months in 2017.
   * 1. **Implementing alternative data in an INVESTMENT process. (Ch. 4.1-4.6)**

A close up of a sign

Description automatically generated

1. **Describe the “Quant Quake” and how it motivated the search for alternative data.**

A multi-strategy firm suffered losses. They liquidated their equity books to cover margin calls. The space was crowded:

Crowded factors

* + - 1. Earnings Yield
      2. 12-month price momentum
      3. 5-day price reversal.

Note that the full table 4.1 is cut off in the ebook version. The full version of the table appears at the end of this section.

1. **Discuss reasons for “the chasm” in the alternative data adoption life cycle and reasons that the chasm had been difficult to cross for many fund managers.**



A close up of a map

Description automatically generated

The mainstream market wants a technology that is more proven before adopting.

1. **Discuss METHODS for improving the efficiency of evaluating data sets for finding Alpha.**
2. Research

Allocating increased research resources specifically to new datasets, setting a clear time horizon for evaluating each (say, 4–6 weeks), and then making a definitive decision about the presence or absence of added value from a dataset. This requires maintaining a pipeline of new datasets and holding to a schedule and

1. Backtesting

Building a turnkey back testing environment which can efficiently evaluate new alphas and determine their potential added value to the existing process. There will always be creativity involved in testing datasets, but the more mundane data processing, evaluation and reporting aspects can be automated to expedite the process in (1).

1. Experienced Sr. Analyst

Assigning an experienced quantitative analyst to be responsible for evaluating new datasets – someone who has seen a lot of alpha factors before and can think about how the current one might be similar or different. Alternative data evaluation should be viewed as a core competency of any systematic fund.

1. ﻿Data suppliers with Innovative Data

Increasing outreach to innovative data suppliers rather than what's available from the big data providers, whose products are harder to consider truly alternative.

1. ﻿Start with Easy Data

Giving priority to datasets which are relatively easy to test, in order to expedite one's exposure to alternative alpha. More complex, raw or unstructured datasets can indeed get one to more diversification and more unique implementations, but at the cost of sitting on one's existing factors for longer – so it may be best to start with some low hanging fruit if one is new to alternative data

1. ﻿Be OK with Limited History

Gaining more comfort with the limited history length that we often see with alternative datasets. With many new datasets, one is ‘making a call’ subject to a short history. One can't necessarily judge these datasets by the same criteria of 20‐year back tests as we can with more traditional factors, both because the older data simply isn't there and because the world 20 years ago has little bearing on the crowded quant space of today. But the alternative, not evaluating these datasets, can be considered riskier. Below, we address some

1. **Describe issues involved with selecting a DATA SOURCE for evaluation within the context of a quant equity process.**

* Finding the right Data Vendor with good Data
* Which data has investment value. Is it Backtested?

1. **Explain why and under what circumstances a FUNDAMENTAL prediction may be more appropriate than an asset price prediction when working with alternative data.**

 Note: Templeton’s value

When is Fundamental Analysis better than Alternative? Depends on the data.

1. Satellite imagery lacks breadth. Parking lots and oil containers are limited.
2. Twitter and blogs have too much NOISE. Hard to scale.
3. Hard to formulate at least one general hypothesis from dataset.
4. **Apply the fundamental Law of Active Management and describe how it applies to discretionary managers and how it applies to quant managers.**

Note:Ray Dalio

The **Fundamental Law of Active Management** is the observation that the information ratio of any trading strategy is proportional to the square root of the number of independent bets made per year. Used like the Sharpe Ratio.



**Information ratio** is useful in comparing a group of funds with similar management styles. William Sharpe extended his work. Difference between Active and the Benchmark.

Generally, the information ratio compares the returns of the manager's portfolio with those of a benchmark such as the yield on three-month [Treasury bills](https://en.wikipedia.org/wiki/United_States_Treasury_security#Treasury_bill) or an equity index such as the [S&P 500](https://en.wikipedia.org/wiki/S%26P_500).

﻿The information coefficient (IC) = Skill

N is the breadth.

It is the correlation between the manager's predictions and subsequent realized returns; a measure of skill.

1. **Describe the transition from fundamental analysis to “quantimental analysis.”**

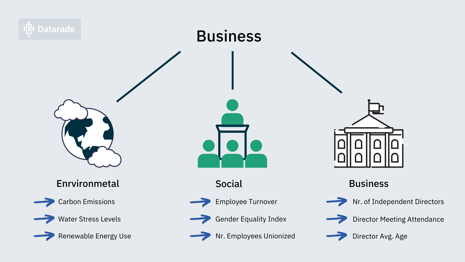
A mix of Both. Fundamental and Quant

Quants are all data driven.

Discretionary is everything else.

New quantitative and alternative data sources will have to be delivered to fundamental teams using the aforementioned workflow tools such as visualization, screening and email alerts, which will enable greater adoption, and the quantimental teams will need to shift some of their attention away from the typical tools in order to best leverage new datasets.

1. **Describe how alternative data can be used to generate a trading signal using examples including blogger sentiment, online consumer demand, transactional data, and environment, social, and governance (ESG) data.**



1. Blogger Sentiment

TipRanks and ExtractAlpha generate sentiment for site articles. Created the TRESS algorithm to rank.

1. Online Consumer Demand

Alpha-DNA performs web-search, websites, social media analysis. Created the DRS to rank.

1. Transactional Data

Credit card transactional data. ﻿Datasets are used to get an early view into an

industry's revenues ahead of corporate earnings announcements.

1. Environmental, Social and Governance (ESG) data.

Exposure and Management main factors.

﻿

﻿Companies with relatively more complaints face greater business risk, either because they have alienated their customers or because they are more likely to be subject to punitive regulatory actions.

We can see that the firms with the most complaints tend to be more volatile and more highly levered and that they have lower dividend yields.

* + 1. **Using alternative and big data to trade MACRO assets (Ch. 5.1-5.4)**

A close up of a sign

Description automatically generated

1. **Define general concepts and terms for the use of big data and alternative data, including “exhaust data.”**

A picture containing transport

Description automatically generated

Data exhaust or exhaust data is the **trail** of data left by the activities of an Internet or other computer system users during their online activity, behavior and transactions.

An enormous amount of often raw data is created, which can be in the form of cookies, temporary files, logfiles, storable choices, and more.

* 1. **Compare traditional model building approaches and machine learning.**

Specifically, machine learning revolves around the problem of *prediction*, while many economic applications revolve around parameter estimation.

Machine learning algorithms are now technically easy to use you can download convenient packages in R or Python

* 1. **Discuss how big data and alternative data can be used to improve economic forecasts and “nowcasts.”**

Nowcasting is the prediction of the present, the very near future and the very recent past in economics.

Predicting the present. Taking the pulse of the economy using real time indicators.

Google searches for economics correlated with present economic anxiety. Whereas,

the gross domestic product (GDP), is only determined after a long delay, and are even then subject to subsequent revisions.

* 1. **Describe how case studies show that alternative data is related to the following types of macro data: US Treasury yield, implied volatility in the foreign exchange market, and investor anxiety.** 
     1. **US Treasury yields**

A person wearing glasses and smiling at the camera

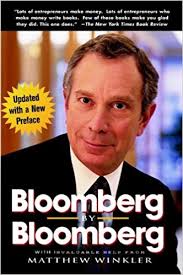
Description automatically generated FOMC Statements. TEXT.

﻿The raw input data consists of text extracted from Federal Reserve communications, which is of a relatively small size.

It is then structured into a tabular data format. This includes speeches, statements and minutes released by the Federal Reserve. Metadata is then derived from this text, such as sentiment scores, which are generated using natural language processing.

* + 1. **Implied volatility in the foreign exchange market (FX)**

We see that there is very often a statistically significant positive correlation between i**mplied volatility and news volume**. This suggests that we can use news volume as an input to model implied volatility.

 Note: More news means more FX volatility

* + 1. **Investor anxiety**



Google Trends: Key Word Search. “Face Masks”

﻿Investopedia is a financial education website. Can we glean anything from the topics that users search for on Investopedia?

The principle behind its Anxiety Index is to track search terms made by users, which results in Investopedia page views.