Big Data Applications for Finance: Overview

Daniele Bianchi[1]

whitesphd.com

[1] School of Economics and Finance
Queen Mary, University of London

Contents

Syllabus and objectives

Syllabus

Week 1: Introduction to the module an Matlab

Week 2: Introduction to Machine Learning

Week 3: Credit risk models I

Week 4: Penalised regressions

Week 5: Asset Management I

Week 6: Asset Management II

Week 7: Text as data and Sparse portfolios

Week 8: Unsupervised learning

Week 9: Large-scale learning

Objective

This is an introductory course on "Big Data" applications in finance.

Thus, the primary purpose of the course is to use *data science methods* in finance.

The goal is learning to interpret empirical estimates, and generate them.

Four-fold objective:

- Use of machine learning techniques for empirical analysis.
- Critically evaluate the informativeness of empirical estimates.
- Visualize complex information sets.
- Introduction to some applications in financial markets.

Reading material

"Big Data" is a field which is relatively new in financial markets.

- For this course we do not follow any particular textbook.
- Readings will be based on original research work.

Throughout the module I will refer to examples/insights from few textbooks:

- *"An introduction to Statistical Learning"*, by Gareth James, Daniele Witten, Trevor Hastie and Robert Tibshirani (2013), freely available at **https://statlearning.com**.
- *"Machine Learning for Factor Investing"*, by Guillaume Coqueret and Tony Guida (2020), freely available at **http://www.mlfactor.com**.
- *"Machine Learning: A Probabilistic Perspective"*, by Kevin P. Murphy (2012)

Occasionally, I might post some popular readings from the press on QMplus.

What is “Big Data”?

Introduction to Big Data

"There was five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the peace is increasing."
(*Eric Schmidt, former Google CEO, 2010*).

"I keep saying that the sexy job in the next years will be statisticians, and I'm not kidding"
Hal Varian, Chief Economist at Google.

"Torture the data, and it will confess to anything"
Ronal Coase, Nobel Laureate in Economics.

New datasets in Finance

New types of data are becoming more common in financial analysis:

Unstructured text data: Social media, web pages, news, court judgements, company annual reports.

Web: clicks, likes, searches, advertising clickthroughs, etc.

Geographic: postcodes/zipcodes and all associated information, as well as satellite imagery.

Electoral activity: electoral rolls, voting records.

Financial: company balance sheets, house prices, stock prices, FX rates, etc.

And many more...

The value in big data

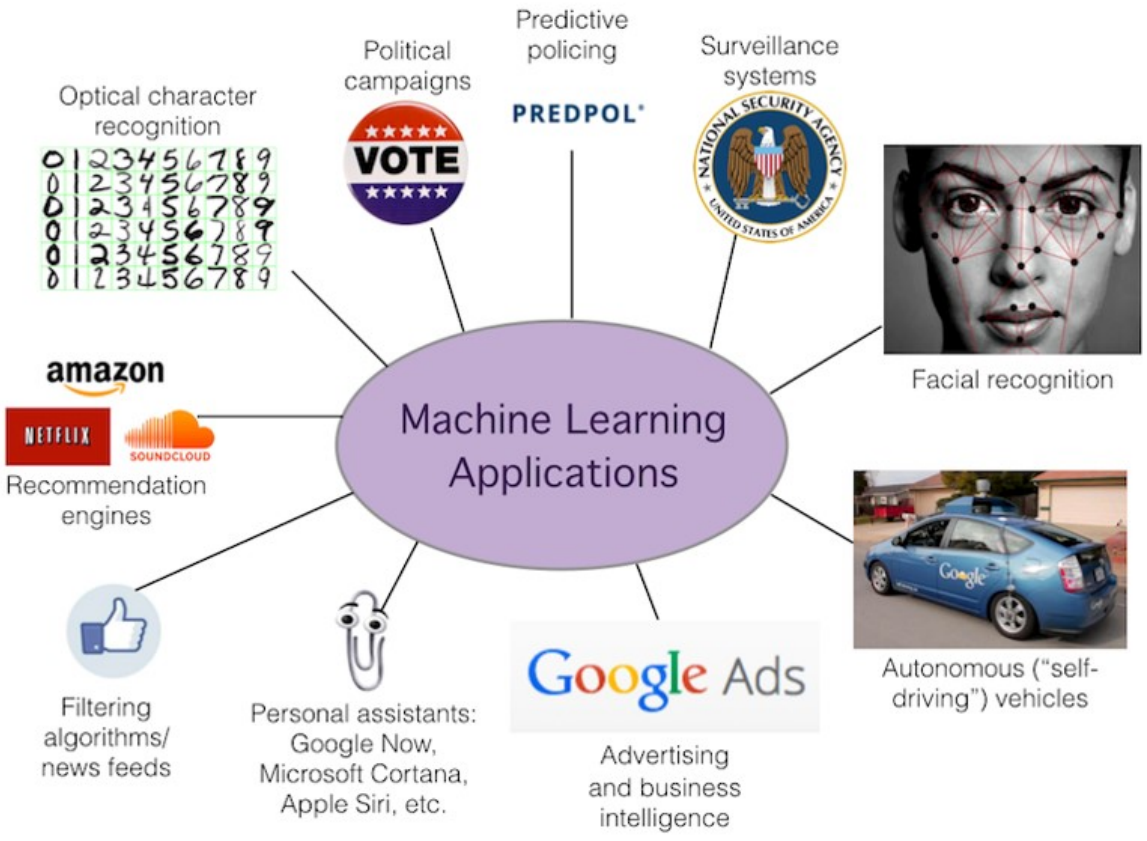
Today data are:

- Easy to come by; commoditized.
- Created from every human action. For e.g., your entrance in the Tube is available as data!

The value is not in the fact that this "data" exists. It is in the *analytics* that you can do with it.

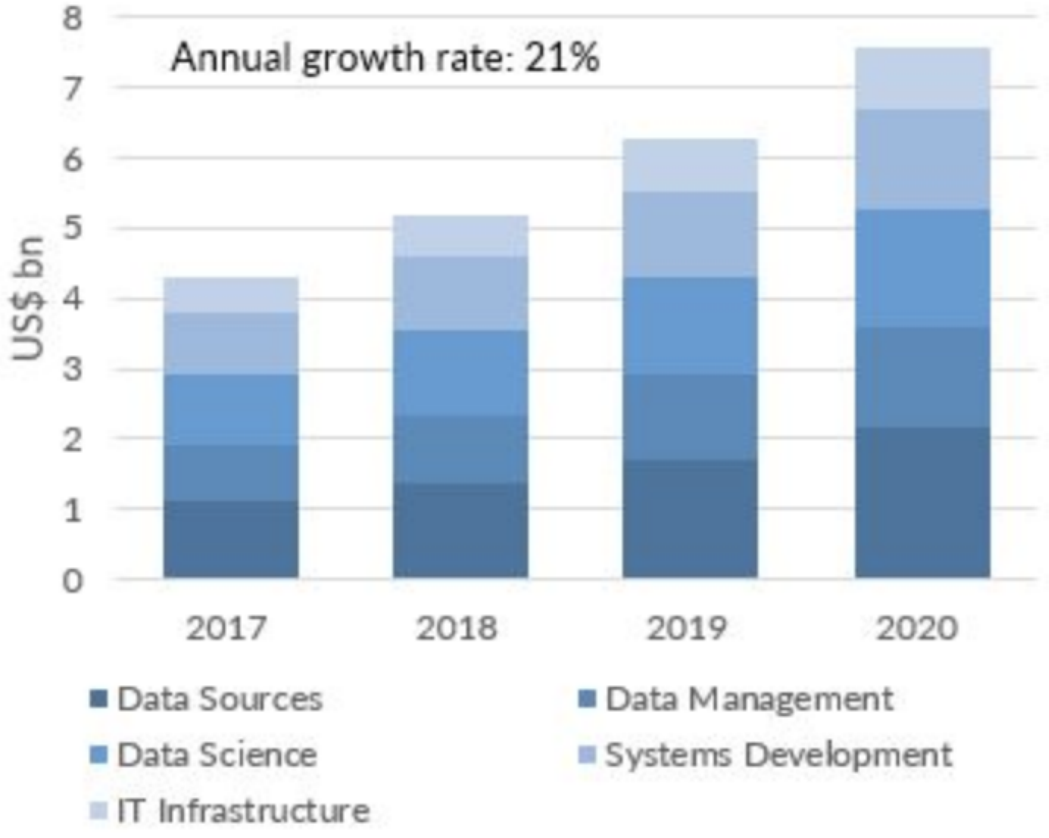
- A well designed algorithm is possibly more valuable than an expensive computer.
- The value in data analysis is the analyst!! (human capital).

Some non-finance examples



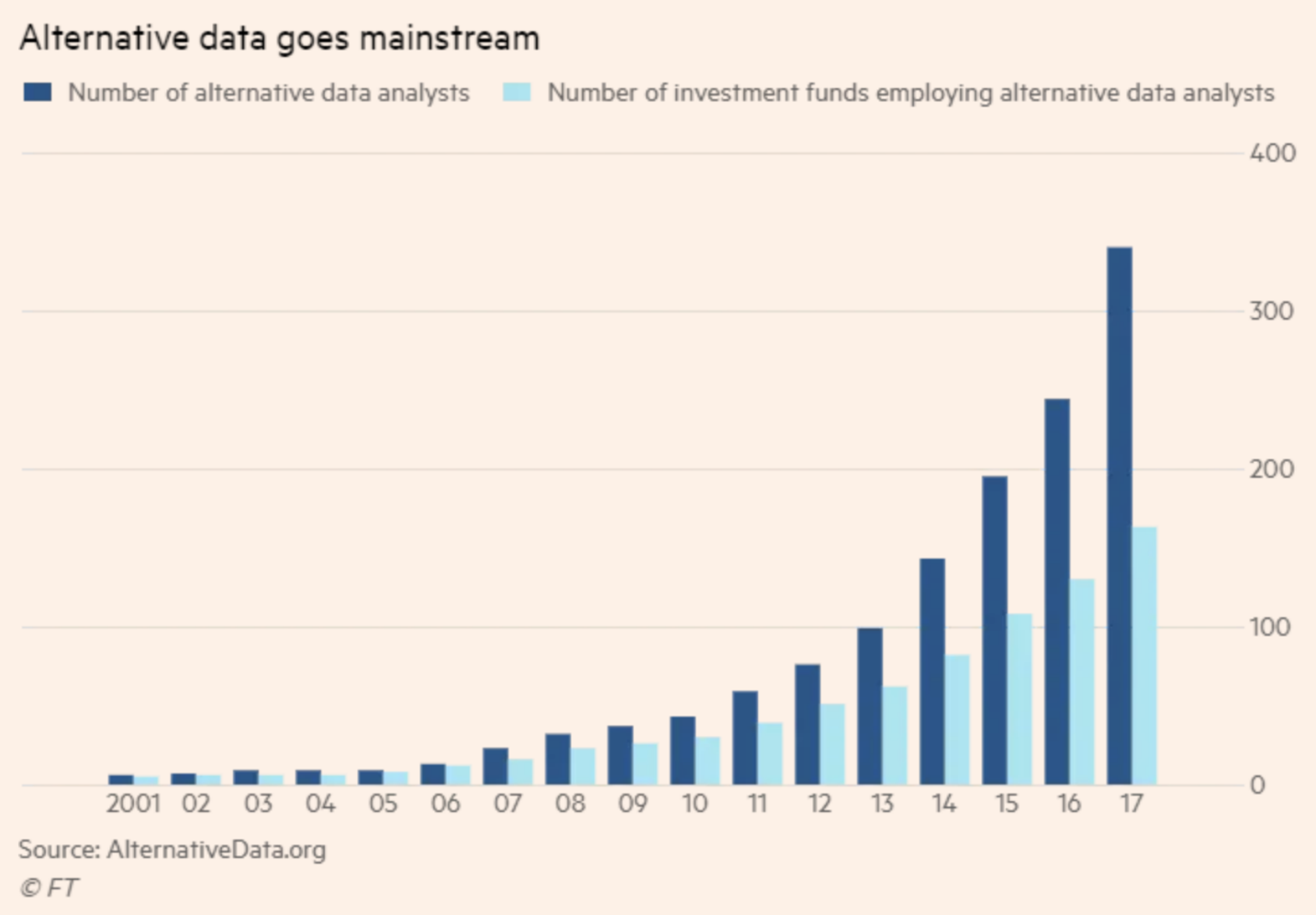
Mainstay in the finance industry

Spending on alternative data is growing rapidly, and will excede US$ 7 billion by 2020...



Source: Opimas Analysis

Mainstay in the finance industry



Thus, what are the building blocks?

Some of the building blocks in big data analytics:

- Basic econometrics: Linear regression (assumed as familiar to this audience).
- The use of statistical learning from vast data.
- Machine learning is one of the tools.
- "Big Data" is not really about the size of the data.
- It is about developing statistical tools to learn. One can learn with very small datasets too!

A bird's-eye view of statistical learning

A bird's-eye view of statistical learning

Suppose we observe Y_i and $X_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, p$.

And suppose that we believe that there is a relationship between Y and at least one of the X's

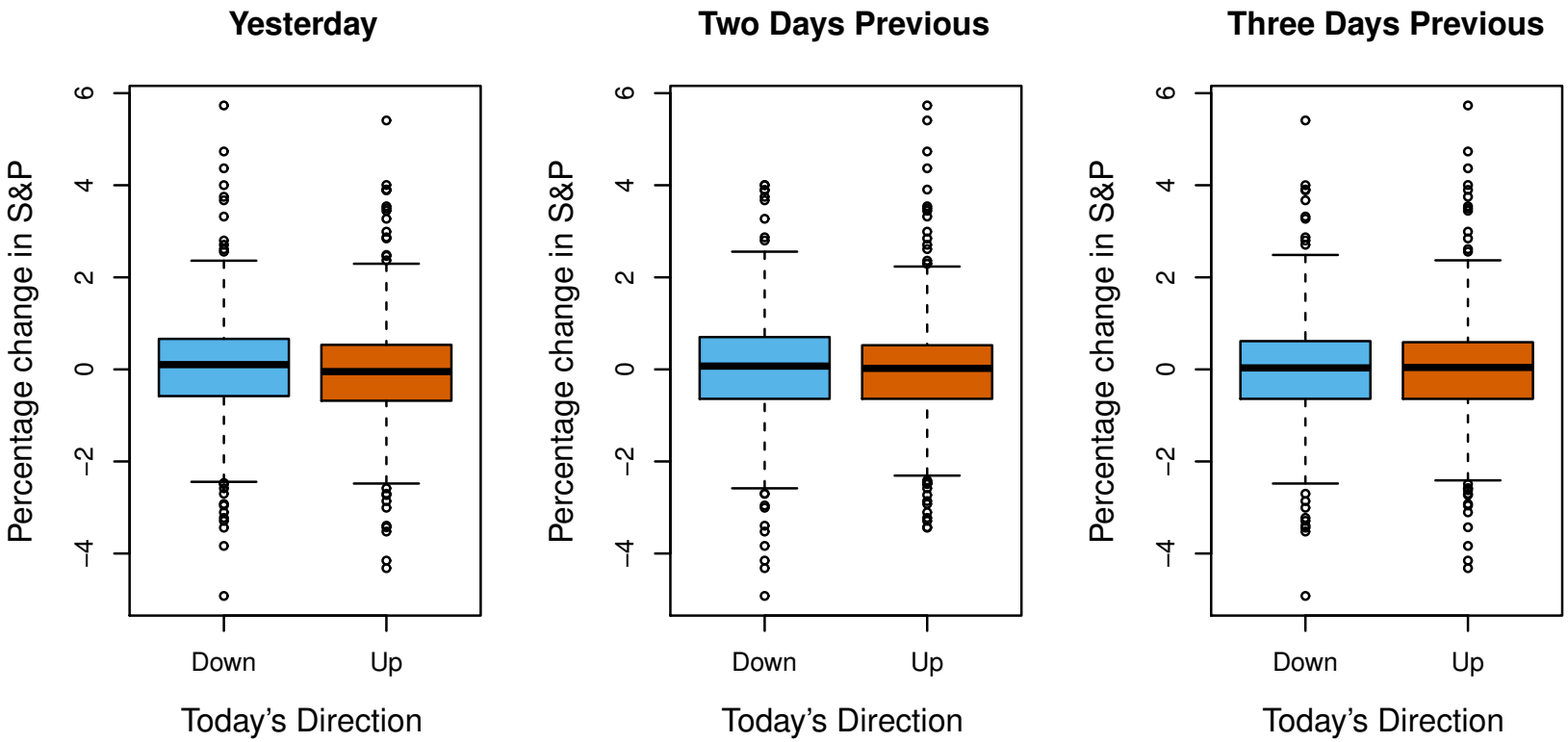
Then we can model the relationship as $Y_i = f(X_i) + \epsilon_i$, where $f(\cdot)$ is an unknown function and ϵ is a random error term with mean zero.

Statistical learning is the science (and art) of discovering $f(\cdot)$. N.B., it is called "learning", because we are using the data to "learn" $f(\cdot)$.

Learning $f(\cdot)$ will be useful for two main purposes:

- Prediction: What's going to happen (given a new X)?
- Inference: Which predictors are critical and why?

What are we trying to learn?



Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased. Rest: Same as left panel, but the percentage changes for 2 and 3 days.

Example of a prediction from a "model"

Quadratic discriminant analysis. Train a model using 2001-2004 data. Predict probability of a stock market decrease using 2005 data. Simple model, correctly predicts 60% of the time.