Imperial College
London

MATH96007 - MATH97019 - MATH97097

Methods for Data Science

Prof Mauricio Barahona

What is data science?

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Wikipedia

Imperial College
London

"The ability to take data to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educaLonal level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data." Hal Varian

Imperial College
London

"The ability to take data to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educaLonal level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data." Hal Varian
Google's Chief Economist

Imperial College
London

There is a lot of hype

Truth despite the hype

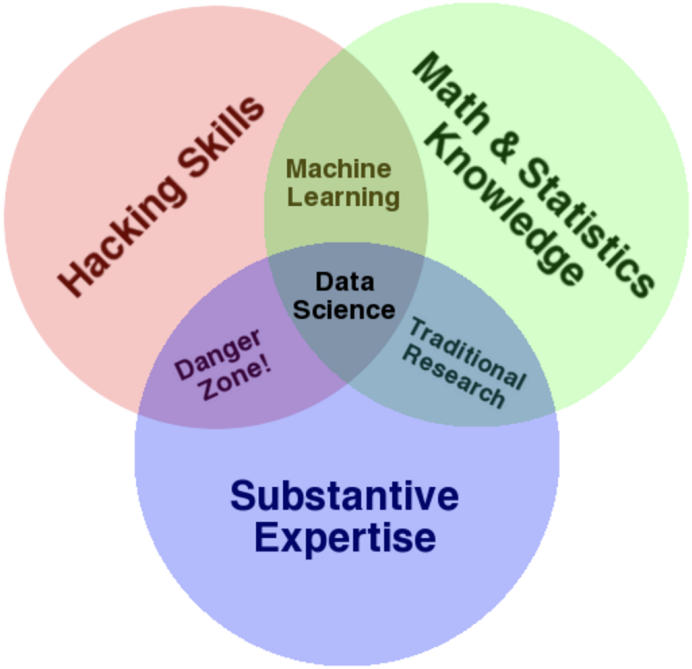
I think there *is* something new in "data science". It's not a science. It's more of a process; a merging of (1) skills in applied mathematics, computer science, statistics, visualisation and communication, with (2) rich datasets and domain knowledge.

In practice, it's usually done in teams. No one has deep knowledge of statistics, mathematics, computer science, visualisation together with detailed knowledge of the data source, industry, interesting questions.

New aspects include: massive amounts of data (even "big data"), combined with cheap computing power.

Imperial College
London

Venn diagram: where does data science fit



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The title "data scientist"

There's no academic field called "data science". After all, scientists have always used data. Statisticians theorise about data, and analyse it too. Very few academics are "professors of data science" .

Cathy O'Neil's book "Doing Data Science" says that an academic data scientist might be defined as: "a scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem."

O'Neil, Cathy; Schutt, Rachel. Doing Data Science: Straight Talk from the Frontline (Kindle Locations 480-482). O'Reilly Media.

Data science jobs

There are a lot of jobs as data scientists. What do these people do?

Fundamentally, a data scientist is someone who can extract meaning from data and communicate the results clearly.

Data science a company typically includes:

- data collection, storage and management
- who has access to what data
- how will the data be used, how does it add value
- privacy considerations
- * analysis of data
- * communicating the results

Case studies: successes of data science

Moneyball! Data in baseball outperforms standard predictors of who will be successful. There's a book, and a movie with Brad Pitt.

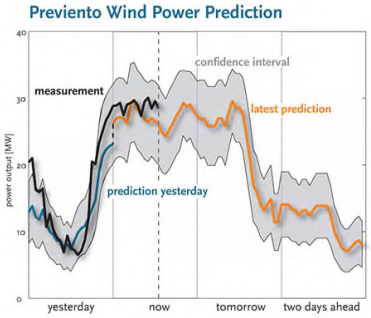
Other big success areas:

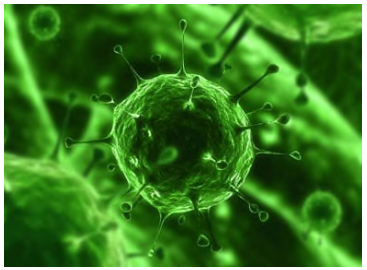
- Spam filters
- Fraud detection
- Election predictions (Nate Silver, `https://fivethirtyeight.com`)
- Predictions of crime hotspots
- Text analysis: twitter, blogs
- Targeted advertising
- Song, movie and product recommendations

Imperial College
London

Case studies: successes of data science

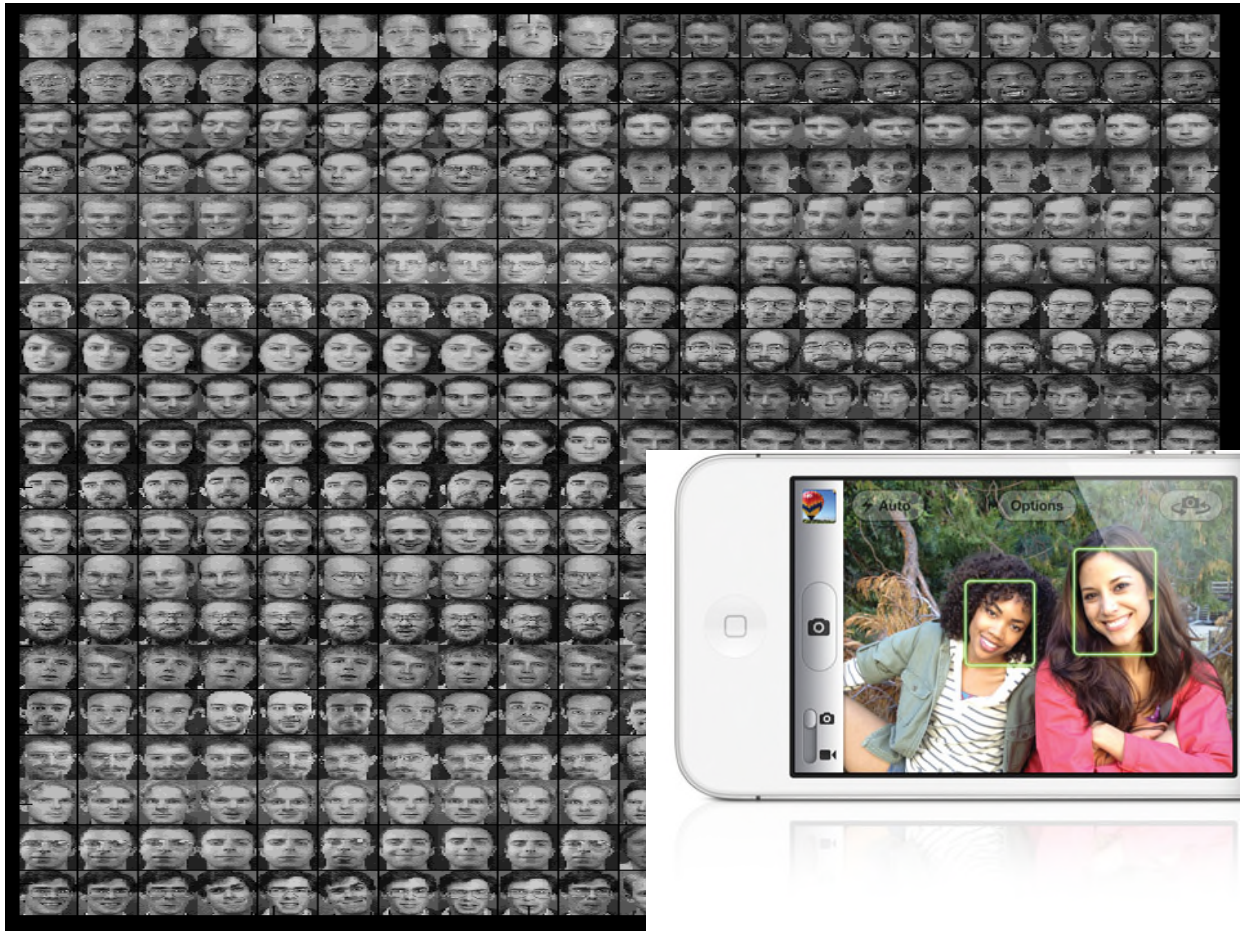
- Forecasting (*e.g.* *Energy demand prediction, finance*)
- Imputing missing data (*e.g.* *Netflix recommendations*)
- Detecting anomalies (*e.g.* *Security, fraud, virus mutations*)
- Classifying (*e.g.* *Credit risk assessment, cancer diagnosis*)
- Ranking (*e.g.* *Google search, personalization*)
- Summarizing (*e.g.* *News zeitgeist, social media sentiment*)
- Decision making (*e.g.* *AI, robotics, compiler tuning, trading*)





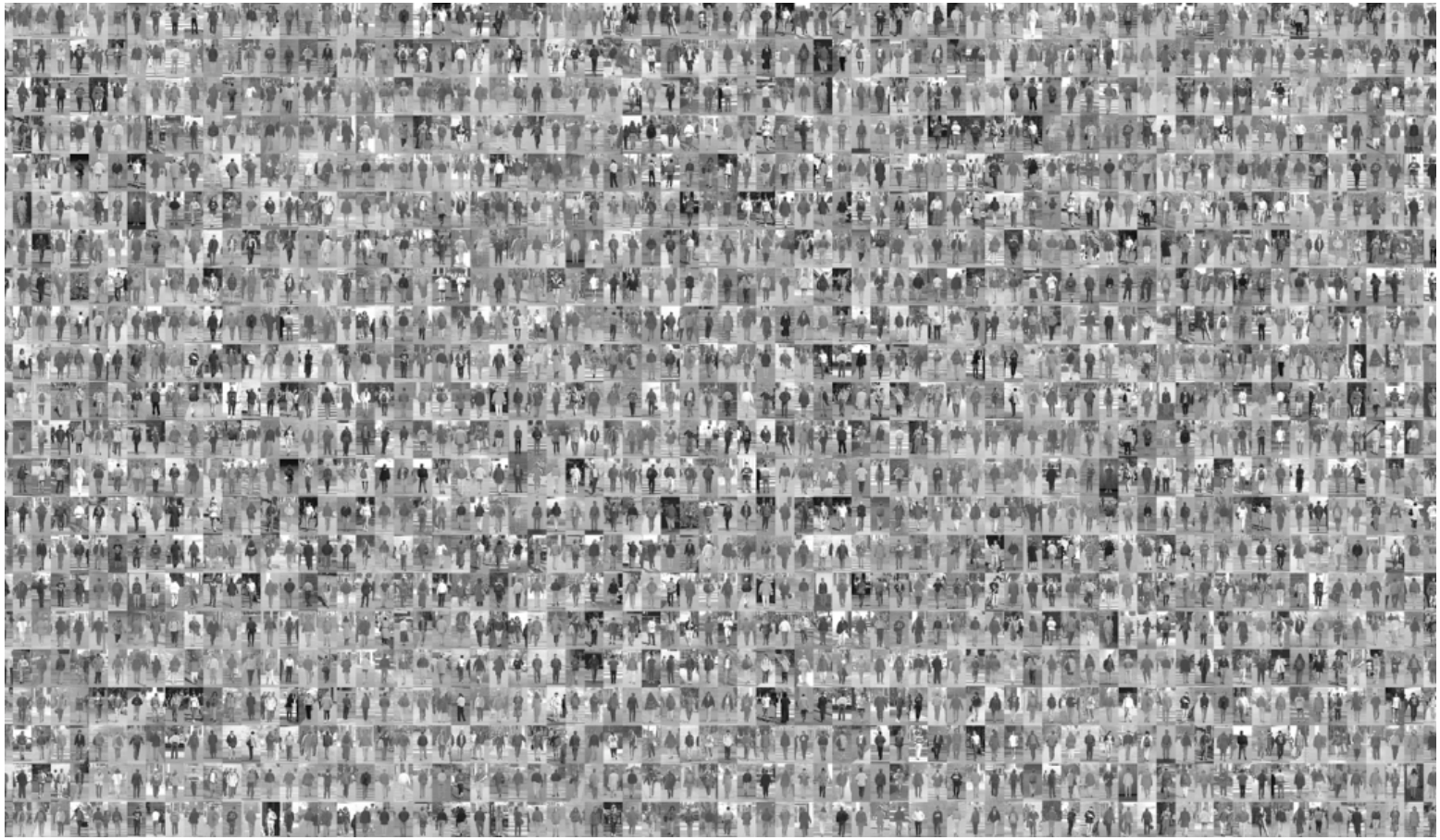
Imperial College
London

Case studies: successes of data science



Imperial College
London

Case studies: successes of data science



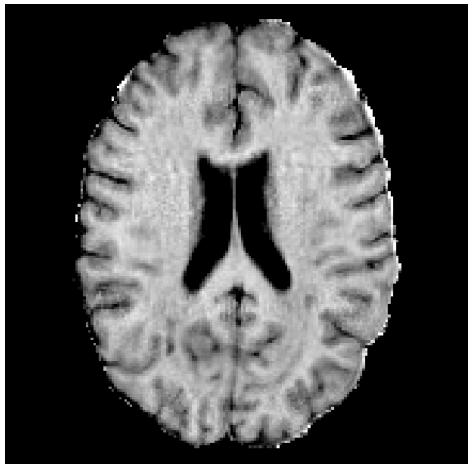
Millions of labeled examples are used to build real-world applications, such as pedestrian detection

Imperial College
London

Case studies: successes of data science

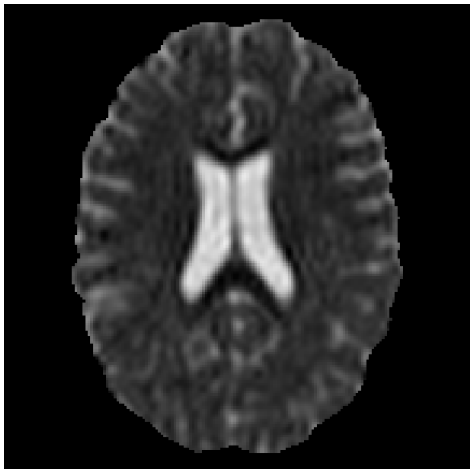
Imperial College
London

Case studies: successes of data science

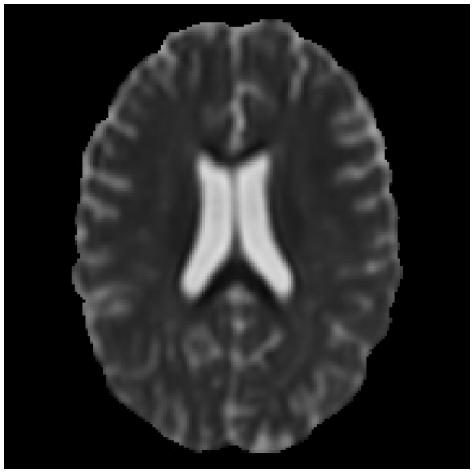


Real nMTR image

Synthetic nMTR image using L1 loss



Real RD image



Synthetic RD image

Imperial College London

Case studies: successes of data science

Natural Language Processing

Imperial College London

Case studies: successes of data science

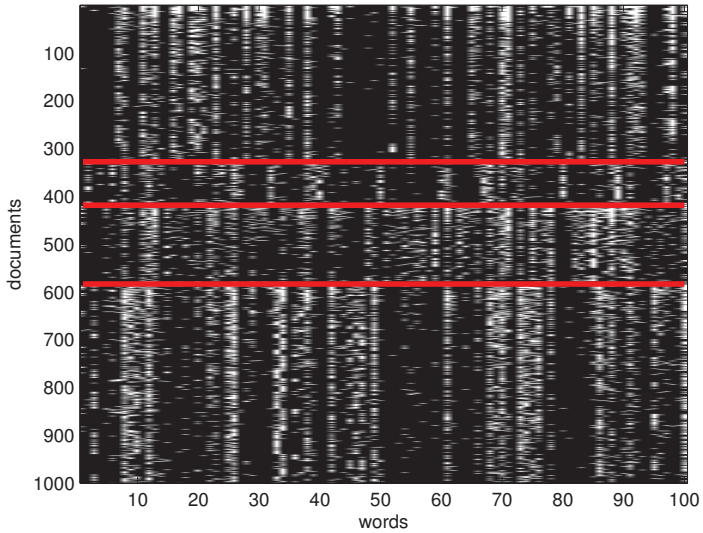


Figure 1.2 Subset of size 16242 x 100 of the 20-newsgroups data. We only show 1000 rows, for clarity. Each row is a document (represented as a bag-of-words bit vector), each column is a word. The red lines separate the 4 classes, which are (in descending order) comp, rec, sci, talk (these are the titles of USENET groups). We can see that there are subsets of words whose presence or absence is indicative of the class. The data is available from `http://cs.nyu.edu/~roweis/data.html`. Figure generated by `newsgroupsVisualize`.

Natural Language Processing

Imperial College
London

Case studies: successes of data science

Speech recognition - Translation

In all these examples, one of the key steps was the mathematical conceptualisation/formulation

Specific questions that you could ask

- Predict whether a heart attack patient will have a second heart attack, using demographic, diet and clinical data
- Predict the price of a stock in 6 months, using company performance measures and economic data. (Don't invest your money by your answers...)
- Identify the numbers in a handwritten ZIP code, from a digital image.
- Estimate the amount of glucose in the blood of a diabetic person from the infrared absorption spectrum of that persons blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic data
- Predict which flu strain will be successful next year based on its DNA sequence and relationships to other flu strains

The process of data science

- The science: what's a good question and what's the dataset you will use to answer it?
- What is the input and what are you trying to determine
- Data handling, cleaning and exploration: computer programming and knowing things
- THE MATH and the main analysis:
  - statistics, computing: what do you compute, why, and what does it mean?
  - machine learning: supervised, unsupervised..
  - networks: characteristics, analysis, comparisons
- Communication: words, visualisations, summaries to tell the story

Imperial College London

Aims and outcomes

- Learn mathematical concepts underpinning methods in learning from data
- The process of conceptualisation of the analysis
- The process of explanation of the analysis
- The mathematical justification of the analysis

- Getting a good exposure to current methods and their use in practice
- Challenges in dealing with (realistic) data

What the tools are, how they work mathematically, and how to analyse a dataset and clearly communicate the results

Imperial College
London

Lots of resources online.
Use them **but** make things your own
and understand the principles