Foundations of Data Science, Fall 2020

1. Introduction: Data Science

Prof. Dan Olteanu

Sept 15, 2020

https://lms.uzh.ch/url/RepositoryEntry/16830890400

https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09

Data Science: Origins

John Tukey: *The Future of Data Analysis*. The Annals of Math. Stats., 1962

”All in all I have come to feel that my central interest is in

data analysis,

which I take to include, among other things:

procedures for analyzing data,

techniques for interpreting the results of such procedures,

ways of planning the gathering of data to make its analysis easier, more precise or more accurate,

and all the machinery and results of (mathematical) **statistics which apply to analyzing data**”

Data Science: These Days

Coupling of scientific discovery and practice that involves

the collection, management, processing, analysis, visualisation, and interpretation

of **vast amounts of heterogeneous data**

associated with a diverse array of

scientific, translational, and interdisciplinary applications

Reactions from the Statistics and Computer Science Communities

Statistics = Science of collecting and analysing numerical data in large quantities

- *Aren't WE Data Science?*
- *A grand debate: Is Data Science just a 'rebranding' of statistics?*
- *Why Do We Need Data Science When We've had Statistics for Centuries?*

Reactions from the Statistics and Computer Science Communities

Statistics = Science of collecting and analysing numerical data in large quantities

- *Aren't WE Data Science?*
- *A grand debate: Is Data Science just a 'rebranding' of statistics?*
- *Why Do We Need Data Science When We've had Statistics for Centuries?*

Computer Science pragmatic view:

- Data science is concerned with **really big data**, which traditional computing resources could not accommodate
- Data science trainees have the skills needed to cope with such big datasets.

An account to these reactions and their legitimacy:

David Donoho: *50 years of Data Science*. 2015

The Two Cultures

Leo Breiman: *Statistical Modeling: The Two Cultures*. Statistical Science, 2001.

1. Generative Modeling
    - Develop stochastic models which fit the data
    - Make inferences about the data-generating mechanism based on model structure
    - Implicit assumption: There is a true model generating the data, and often a ‘best’ way to analyse the data.

    Proponents: Academic Statisticians

2. Predictive Modeling
    - Silent about the underlying mechanism generating the data
    - Allows for many different predictive algorithms
    - Interest: accuracy of prediction made by different algorithm on various datasets
    - Epicenter: Machine Learning; sitting within CS departments

    Proponents: Computer scientists and *industrial* statisticians.

The Two Cultures

”The statistical community has been committed to the almost exclusive use of **[generative] models**. *This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting problems.*

The Two Cultures

”The statistical community has been committed to the almost exclusive use of **[generative] models***. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting problems.*

[Predictive] modeling*, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets.*

If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on [generative] models”

The Predictive Culture's Secret Sauce: The Common Task Framework

(a) Publicly available training datasets with feature measurements and class label for each observation.

(b) Competitors whose *common task* is to infer a class prediction rule from the training data.

(c) A scoring referee, to which competitors can submit their prediction rule.

The referee runs the prediction rule against a testing dataset which is sequestered behind a Chinese wall.

The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule.

CTF applied by DARPA successfully in many problems, e.g.,:

- machine translation, speaker identification, fingerprint recognition,
- information retrieval, OCR, automatic target recognition.

General Experience with CTF

1. Error rates decline by some percentage each year, to an asymptote depending on task and data quality.
2. Progress usually comes from many small improvements
   A change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.

Those fields where machine learning has scored successes are essentially those fields where CTF has been applied systematically.

The Common Task Framework is the single idea from machine learning and data science that is most lacking attention in today's statistical training.

Driving Forces behind this new Science

1. The formal theories of statistics

   Statistics thus represents a fraction of data science

2. Accelerating developments in computers

   Faster hardware, better algorithms

3. The challenge, in many fields, of more and ever larger bodies of data

   Sciences and society become increasingly more digitalised

4. The emphasis on quantification in an ever wider variety of disciplines

   Extract compact knowledge out of a sea of data

   As science itself becomes a body of data that we can analyze and study, there are opportunities for improving the accuracy and validity of science, through the scientific study of data analysis.

We Currently Witness an Industrial Revolution of Data!

- Much cheaper to generate data

  Inexpensive sensors, smart devices, social software Web 2.0, multiplayer games, Internet of Things connecting homes, cars, appliances, RFIDs, GPS, software logs, audio & video

- Much cheaper to process data

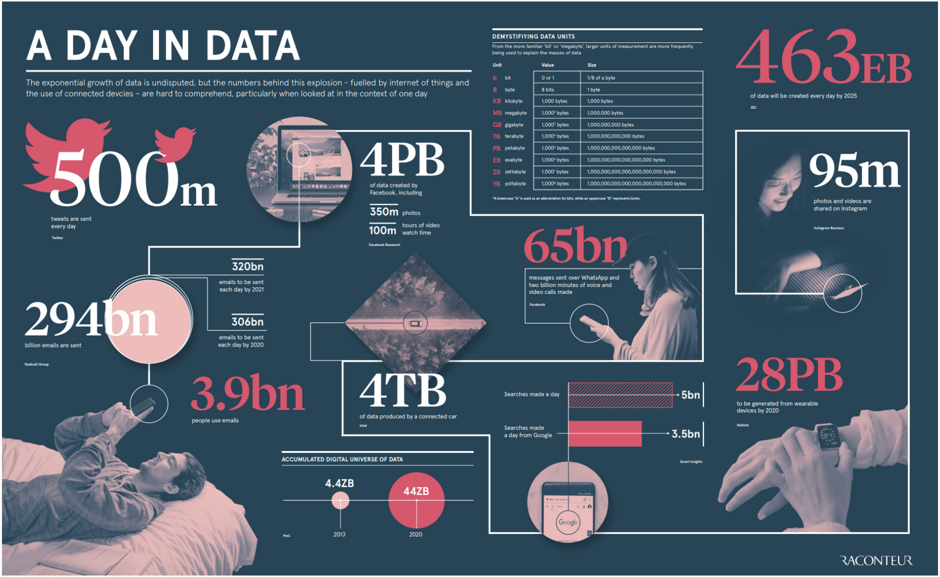
  Advances in multicore CPUs, inexpensive cloud computing, open source software, unlimited fibre power broadband

- Society has become increasingly more computational

  Many categories of people involved in generating, processing, and consuming data

http://www.flickr.com/photos/revsorg/1168346563/

How much Data is Generated each Day? (World Economic Forum, 2019)



Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

The ”Big Data” Buzz

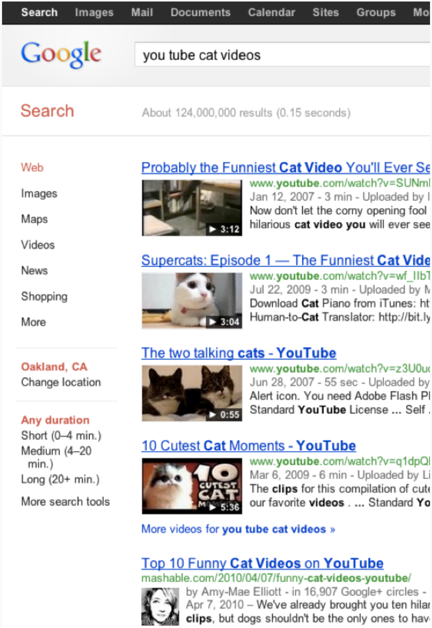
“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

Eric Schmidt, Google (and others)

The "Big Data" Buzz

"Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days."

Eric Schmidt, Google (and others)



Sciences Become Increasingly More Computational

http://www.economist.com/node/15579717

The Economist: *“Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn’t just more. More is different.“*

The Data Deluge Makes the Scientific Method Obsolete

- George Box, statistician (1970s):

  ”All models are wrong, but some are useful.“

- Peter Norvig, Google’s research director (2008):

  ”All models are wrong, and increasingly you can succeed without them.”

- Anand Rajaraman, academic/VC, and others (2012):

  “The new oil/oxygen of Google/Facebook/Twitter/. . . = Simple models + big data.” No need for a-priori sophisticated and inherently wrong models.

Full Scope of Data Science

1. Data Exploration and Preparation
2. Data Representation and Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualisation and Presentation
6. Science about Data Science

Each of the above facets of data science require special skills beyond those taught in e.g., Statistics and Computer Science when taken alone.

Data Exploration and Preparation

80% of the effort devoted to data science is diving into messy data

- Exploratory Data Analysis requires serious time and effort
  - to learn about the data and
  - to prepare it for further exploitation.
- Data cleaning to address anomalies
- Value recoding and reformatting
- Value grouping

Data Representation and Transformation

Central step: Implement an appropriate transformation restructuring the originally given data into a new and more revealing form.

- Modern data management and database skills
    - managing unstructured data (text)
    - spreadsheets
    - (no)SQL DB
    - distributed DB
- Maths representations
    - Fourier transform for acoustic data
    - wavelet transform for image and sensor data

Computing with Data

Data scientists need to keep current on new computing idioms

Programming languages for

- Data analysis and processing
- Text transformation and managing complex computational pipelines

Efficient centralised and distributed computing paradigms

- Distributive computation, algorithms, computational complexity
- Cloud computing to run massive number of jobs
- Documenting and abstracting commonly recurring pieces of software

Data Modeling

Data scientists use tools and viewpoints from Breiman's modelling cultures

- Generative modeling
  - Propose stochastic models that could have generated the data
  - Derives methods to infer properties of the underlying generative mechanism
- Predictive (algorithmic) modeling
  - Construct methods that predict well over some concrete dataset

Data Visualisation and Presentation

Crystallise understanding of a dataset by developing a new plot which codifies it

- Histograms, scatterplots, time series plots
- Dashboards for monitoring data processing pipelines that access streaming or widely distributed data
- Visualisations for presenting conclusions from a modelling exercise or CTF challenge

Science about Data Science

The true effectiveness of a tool: the probability of deployment times the probability of effective results once deployed

Identify commonly-occurring analysis/processing workflows

- Use data about their frequency of occurrence in scholarly/business domains
- Measure the effectiveness of standard workflows in terms of performance metrics: human time, computing resource, analysis validity
- Uncover emergent phenomena in data analysis, e.g.,
  - new patterns arising in data analysis workflows
  - disturbing artefacts in published analysis results

Scope of this Course:

Basics of Data Modelling

˜ Machine Learning˜