Introduction

Spyros Samothrakis
Research Fellow, IADS
University of Essex

January 16, 2017

Course Structure

- 10 weeks
- Each week:
  - 2-hour lecture
  - 3-hour lab: a whole semester of Role-Playing!
- Assessment:
  - 2 assignments
    - Project description (more on this later) - 15%
    - Final application and report - 70%
  - Labs - 15% (1.5% each)
    - You **must** complete each weekly lab!
- This is the first and only non-technical lecture!
- *Feel free to interrupt me at any point with questions/comments*

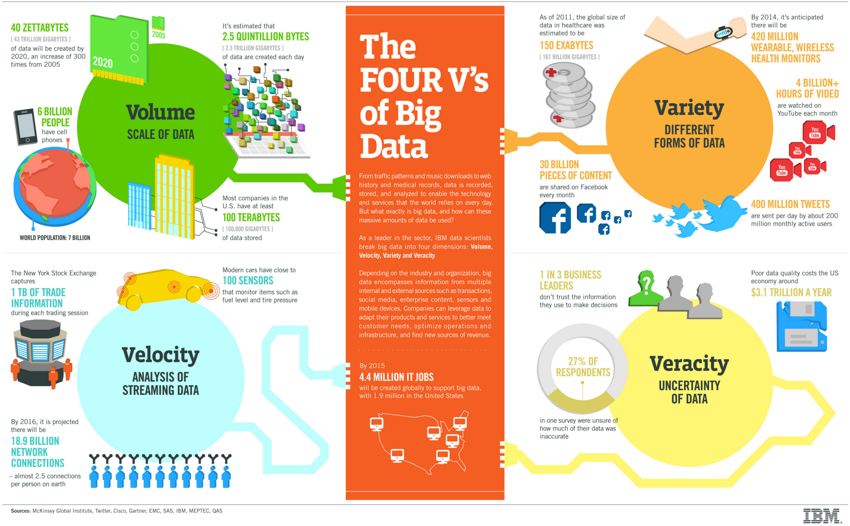
Better living through data

- The term “Data Science” was coined by Jim Grey
  - As the fourth “Science Paradigm”
- We are going to make sense of the world by using tons of data
- An umbrella term that could just mean a “Statistician of the 21st Century”
- Mixing statistics and computer science (databases, machine learning)

Mixing statistics, philosophy of science and machine learning

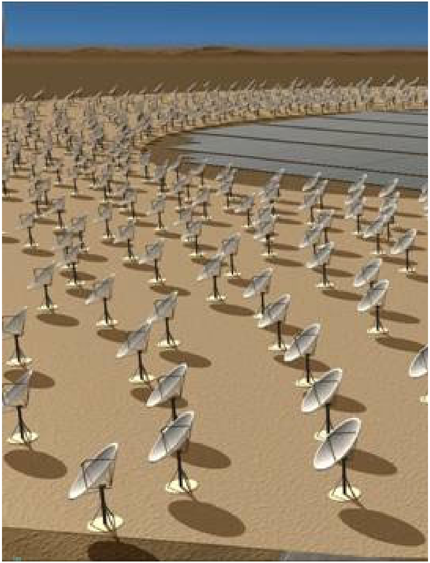
- Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical Science 16.3 (2001): 199-231.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
- Anderson, Philip W. "More is different." Science 177.4047 (1972): 393-396.
- Science is the epistemology of causation

IBM's Infographic



Classic science

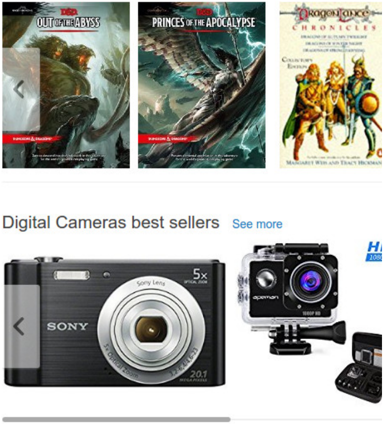
- The original data science field
- SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6e+6 TB), (Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the Big Data Era." Data Science Journal 14 (2015).)[1]
- Bioinformatics
- Medical science



[1] http://datascience.codata.org/article/10.5334/dsj-2015-011

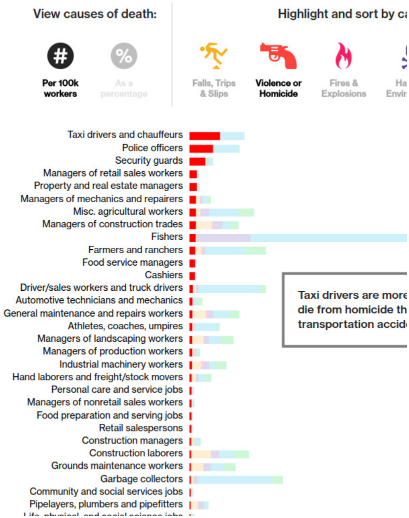
Recommender Systems

- One of the most popular applications of data science
- Propose products to customers based on past history
- Almost all online vendors do it
- Made popular by the Netflix prize



Data journalism

- One can report news from data dumped from public bodies
- e.g. The Deadliest Jobs in America[2]
- Searching and indexing datasets / leaks (think wikileaks)



[2] https://www.bloomberg.com/graphics/2015-dangerous-jobs/

Finance & Insurance

- Predict stock prices (Hedge Funds)
- Insurance models
- Credit score
- In fact, a lot of trading that currently happens is algorithmic trading[3]
- Sudden drops in share prices often caused by defective algorithms

[3] http://www.bbc.com/news/business-34264380

Politics (current)

"...This included a) integrating data from social media, online advertising, websites, apps, canvassing, direct mail, polls, online fundraising, activist feedback, and some new things we tried such as a new way to do polling (about which I will write another time) and b) having experts in physics and machine learning do proper data science in the way only they can – i.e. far beyond the normal skills applied in political campaigns..."

Dominic Cummings's (Head of *Vote Leave*) Blog[4]

[4] https://dominiccummings.wordpress.com/2016/10/29/on-the-referendum-20-the-campaign-physics-and-data-science-vote-leaves-voter-intention-collection-system-vics-now-available-for-all/

Politics (historical)

- New Yorker - THE PLANNING MACHINE: Project Cybersyn and the origins of the Big Data nation[5]
- Cybersyn / Chile during Alliente's rule, co-designed by Stafford Beer
- Plan was to use data fed directly from each industry to automate production

[5] http://www.newyorker.com/magazine/2014/10/13/planning-machine

Question answering

- e.g. Antol, Stanislaw, et al. "VQA: Visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.[6]
- Input can be videos, websites, et
- Think google

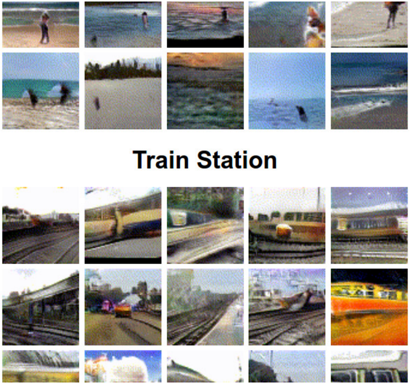
[6] http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

Digital Marketing

- Is a new product I just created well received by our customers?
- Is a new marketing campaign e-mail sent detrimental to our efforts?
- What is the content a chain of e-mails should have?
- Customer segmentation
- What adverts should I present to a user?

Creative artificial intelligence (recipes, music, art, text)

- e.g. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." Advances In Neural Information Processing Systems. 2016.[7]
- Generate an artefact
  - Generate videos
  - Generate text
  - Generate music



[6] http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

Game playing

- We recently have seen a resurgence of game playing machines
- A computer GO programme finally outperformed top humans (AlphaGO)
- No-limit heads up poker (matches still played as we speak!)
- New labs are opening from major game companies dealing with game AI
- Though directly related, game analytics

Artificial intelligence

- Everything we have seen so far are basically applications of Artificial Intelligence and Machine Learning
- Inductive reasoning from a limited amount of examples
    - Structured learning
    - One-shot models
- Deductive reasoning
    - From concepts to data
    - Platonic forms

Some sample data

- takes_off_road: owner takes the vehicle off road
- company_vehicle: it belongs to a business
- is_over_30: age of vehicle is over 30
- regular_service: is the vehicle serviced regularly?
- brake_down: will it break down within three months of our inspection date?

takes_off_road	company_vehicle	is_over_30	regular_service	brake_down
0	1	1	0	1
0	0	1	1	0
1	1	1	1	1
0	1	1	0	1
0	0	1	0	0
0	1	0	0	0
1	0	0	1	0
1	1	1	1	1
1	0	0	1	1
0	1	1	0	1
1	0	0	1	0
1	1	0	0	0
0	0	0	0	0

Predictions

- The most common data science operation
- Can you predict if a car will break down given the data, and if yes with what probability?
- Can you learn a model, that if provided with a tuple $< takes\_off\_road, company\_vehicle, is\_over\_30, regular\_service >$ predict *break_down*?
- The tuple represents a vehicle
- Columns are called *features*
- If we call the model M, can you learn $P(C|D;M)$
- You might have seen this as *supervised learning*
- You can also try to predict if a vehicle was taken off-road, given that it broke down

Clustering

- Another very common request
- Imagine there is some hidden property in the data, another feature that we have not observed
  - This feature groups together vehicles
  - Again we are looking for $P(C|D;M)$, but C is a fictional/latent variable
- Unsupervised learning
- The probabilistic intuition I provided is not unique

Inferring what-if scenarios from the data

- Say your vehicle broke down
- What would have happened if you have not driven if off-road?
- Have a look at the data - what can you say?
- Do you have enough data of the needed type?
- Causality from observational data
  - Super hard, but super important
  - Think of smoking!

Acquiring new data

- We can't really answer what would happen to vehicle from the data collected already
- We might need to set a controlled experiment where:
  - We find vehicles of similar characteristics
  - Drive them off-road
  - See if they break down
  - What is the optimal way of doing such a procedure?
- Causality from experimental data - mostly what science is all about
  - **Science is the epistemology of causality**

Anomaly Detection

- If we are given a new vehicle, can we say if it is “special” in a way?
- Maybe it’s the only vehicle with certain features
- Maybe it’s a unique vehicle
- Somehow we need to find bizarre samples that do not conform to expect norm
- Multiple formal definitions

Generate new data

- Can I generate fictional vehicles and their properties?
- Mathematically, learn P(D;M), a model of the data
- You can then use your plausible, but fictional vehicles for entertainment
- "Learning to draw before learning to see"
  - $P(D, C; M) = P(C|D)P(D)$
  - $P(D|C; M)$

Dimensionality reduction

- Maybe we only need some feature combination above
- Maybe some features only carry noise with them - they are irrelevant
- For example, how important the *car_colour* feature would be?
- What happens if we learn based on irrelevant features?
- Spurious correlations are everywhere
- Kicking out useless features might make the model more interpretable

Linking with other data/collecting labels

- What if the data we have is not enough?
- In our example, model make is not provided
- Can we inquire data providers to find that?
- How expensive would that be?
- How easy is to label the data?
  - Active learning
  - Labelled data often very expensive

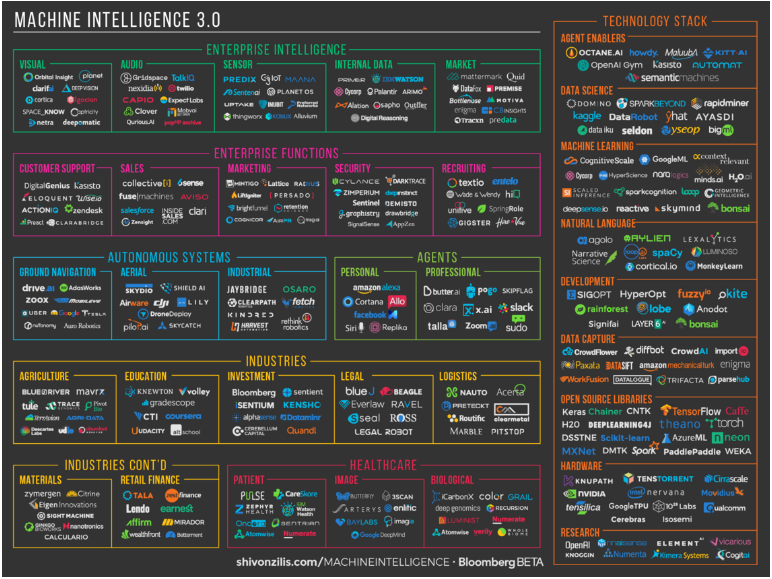
Making decisions from data

- Now that we have a model
- Let's say you know that a vehicle will break down after three months with a certain probability
  - How much do we charge for insurance on it?
  - Should we even sell insurance to the owner?
  - What is the risk of actually selling insurance?
- We are missing another model (that of the customer)
  - Do we actually need the model?
  - Do customer preferences change over time?
- Bandits, reinforcement learning

Some notes

- *"If you torture the data enough, nature will always confess."*
  - *Disputed*
- *"If you torture the data long enough, it will confess to anything."*
  - Huff, D. "How to lie with statistics (illust. I. Geis)." NY: Norton (1954).
- *Lies, damned lies, and statistics*
  - *Disputed*

Startup mayhem



The law

“We summarize the potential impact that the European Union’s new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which”significantly affect" users. The law will also effectively create a **right to explanation**, whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation"

Goodman, Bryce, and Seth Flaxman. “European Union regulations on algorithmic decision-making and a” right to explanation“.” arXiv preprint arXiv:1606.08813 (2016).

The social impact of AI/Machine Learning

"We examine how susceptible jobs are to computerisation. To assess this, we begin by implementing a novel methodology to estimate the probability of computerisation for 702 detailed occupations, using a Gaussian process classifier. Based on these estimates, we examine expected impacts of future computerisation on US labour market outcomes, with the primary objective of analysing the number of jobs at risk and the relationship between an occupation's probability of computerisation, wages and educational attainment. According to our estimates, about 47 percent of total US employment is at risk. We further provide evidence that wages and educational attainment exhibit a strong negative relationship with an occupation's probability of computerisation"

- Not sure I believe them, but read the article

Frey, Carl Benedikt, and Michael A. Osborne. "The future of employment: how susceptible are jobs to computerisation." Technological Forecasting and Social Change (2014).

Overall on Data and Society

- Think about how much of your life you spend online
  - Not just on a computer, but mobile phones, GPS signals etc., car sensors
  - Soon your fridge and coffee machine (IoT)
- Tons of data flying around
  - They are being used to make decisions on a micro level (i.e. about you)
- Regulations are set in place
- New El-Dorado?

Linux VM

- Download the VM for this module
- External link `https://docs.google.com/uc?id=0B_kDfEzMuWD6ZGJFU1VfeEY3TnM&export=download`
- The VM contains all (or most) of what you need if you are to create a successful python project
- Username/password is `mlvm/mlvm`
- You will have a USB stick were you should copy the VM folder (after you un-rar the archive)
- More about this on the labs

Python

- Python is the language of this module
- You are expected to be competent python programmers (or willing to put the extra effort)
- Python has evolved to be one of the two "data science" languages (the other is **R**)
- Python has/is:
  - An excellent list of features coming from functional programming
  - A huge number of related libraries
  - Easy to learn
  - Object oriented programming capabilities
  - Can be extended via *C* trivially
  - A massive amount of related libraries

Ipython/Jupiter

- A better command line interface to python
- Has something called a “notebook”
  - A notebook combines code + natural language
- See here for a very nice example

`https://github.com/rhiever/Data-Analysis-and-Machine-Learning-Projects/blob/master/example-data-science-notebook/Example%20Machine%20Learning%20Notebook.ipynb`

Numpy

- Numpy is possibly the most important library in Python for numerical computing
- Provides vector and matrix operations on top of *arrays*
- Almost every other library manipulates numpy arrays underneath

Scipy

- A scientific computing framework
- Linear Algebra
- Optimisation
- Statistics
- Clustering

Scikit-learn

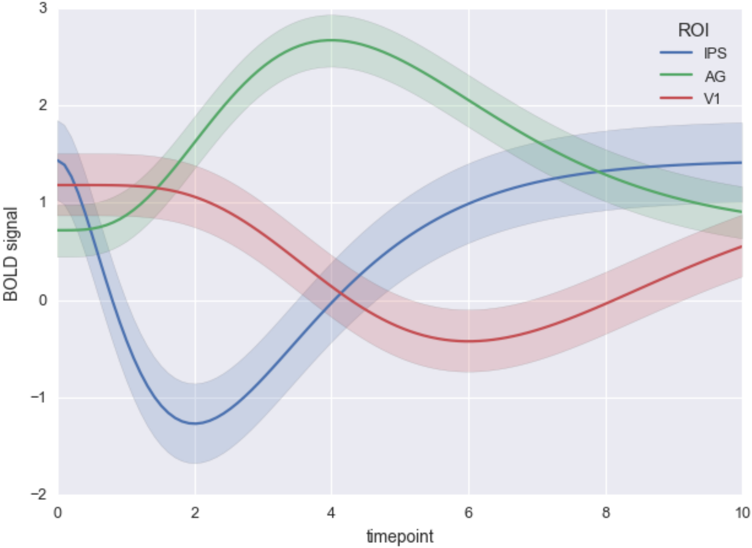
- A machine learning framework
- Includes almost everything, apart from neural networks
- We are going to use it extensively
- Super-fast trees
- Excellent documentation

Keras

- A neural networks framework
- Very popular
- Uses theano or tensorflow underneath
- We will use this as well
- Though notice this is not a module on neural networks
  - But you can delve into this if you want
  - Not trivial, but not super hard either
  - Again, a lot of examples and online tutorials

Matplotlib, seaborn

- Standard visualisation tools



Twitter API

- We are going to use this a lot
- Most of the labs will be about creating twitter bots!
- Twitter has a python API
  - You will use it to send tweets and collect tweets
  - Also `tweepy`
- If you haven't used twitter, create an account now and tweet a bit
  - Read other people's tweets
  - Get a feeling of the platform
- You will need to create both a gmail account and a twitter account for the labs

Pandas

- *R* had dataframes
  - Essentially, a very SQL-like table-like data structure
- "DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object"
- You can manipulate these, and it helps a lot with cleaning up and re-shaping your data
- This is a big part of data science!
  - Data munging/data wrangling

XGBoost

- The competition winner!
- Used a lot by kaggle participants
- (Kaggle) `https://www.kaggle.com/`
- Now runs on GPUs!
- We will deal with boosting at a later lecture

Apache Spark

- The clustering framework
- You need it when you have tons of data to process
- Has its own machine learning library (mlib), which we are not going to use
  - But it makes sense to use it if your data doesn't fit in memory
  - Can be used with 3rd party modules in conjuction with sk-learn
- Sits on top of HDFS (which we are going to install and use later on)

Github

- All your code for your project will need to be publicly available
- Create a github account if you don't have one
- Two directories (`/src`, `/pdf`)
  - One for the pdf of the project
  - One for the code
  - If you have an ipython ipnb it should go here
- Add a README.md as well!

Assignments

- Let's spend some time looking at possible projects

Final remarks

- This is a huge field
- We will not (and cannot) cover everything, so feel free to explore
- We have only scrapped the surface
- The aim of this module is to get you practical skills that will help you survive the data science arena
- Coding + ML + statistics!
- We will try to get as much of a unified view of the field as possible
- Next week: boostrapping!