CE888: Data Science and Decision Making
Lecture 1: Introduction

Ana Matran-Fernandez
University of Essex

January 14, 2019

Course Structure

- 10 weeks
- Each week:
  - 2-hour lecture
  - 3-hour lab
- Assessment:
  - No exam!
  - 10 Labs - 15% (1.5% each)
    - You **must** complete each weekly lab!
    - When you finish, call one of the tutors to show it to us
  - 2 assignments
    - Project description (more on this later) - 15%
    - Final application and report - 70%
    - 10-page IEEE journal format report (No more, no less!)
- This is the first and only non-technical lecture
- *Feel free to interrupt me at any point with questions/comments*

Course Lecturers

- Module supervisor: Spyros Samothrakis (ssamot@essex.ac.uk)
- Lectures 1–5: Ana Matran-Fernandez (amatra@essex.ac.uk)
- Lectures 6–10: Haider Raza (h.raza@essex.ac.uk)
- Teaching assistant: Lina Barakat
- We're all here to help you. Don't hesitate to get in touch when you need to!

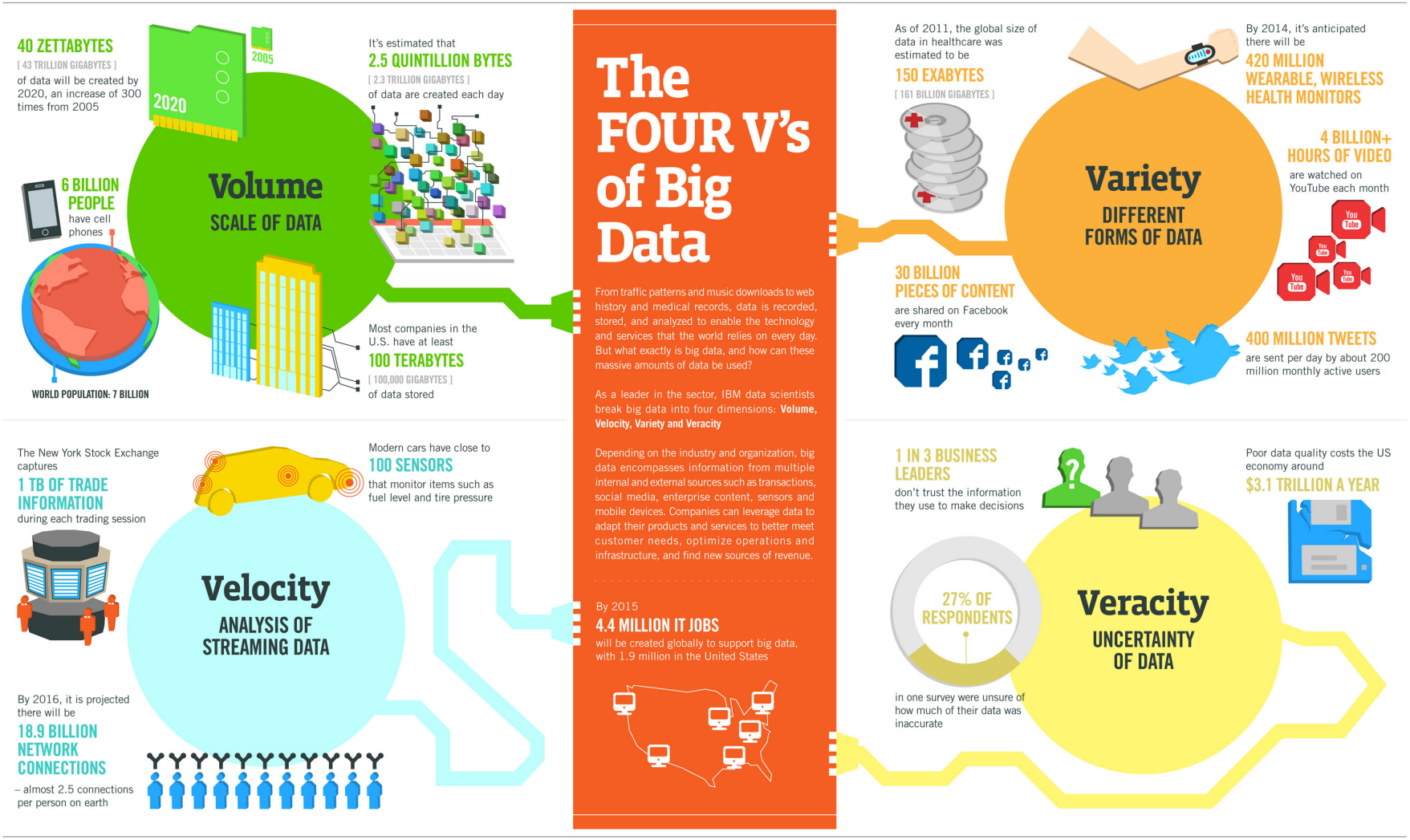
Better living through data

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

Mixing statistics, philosophy of science and machine learning

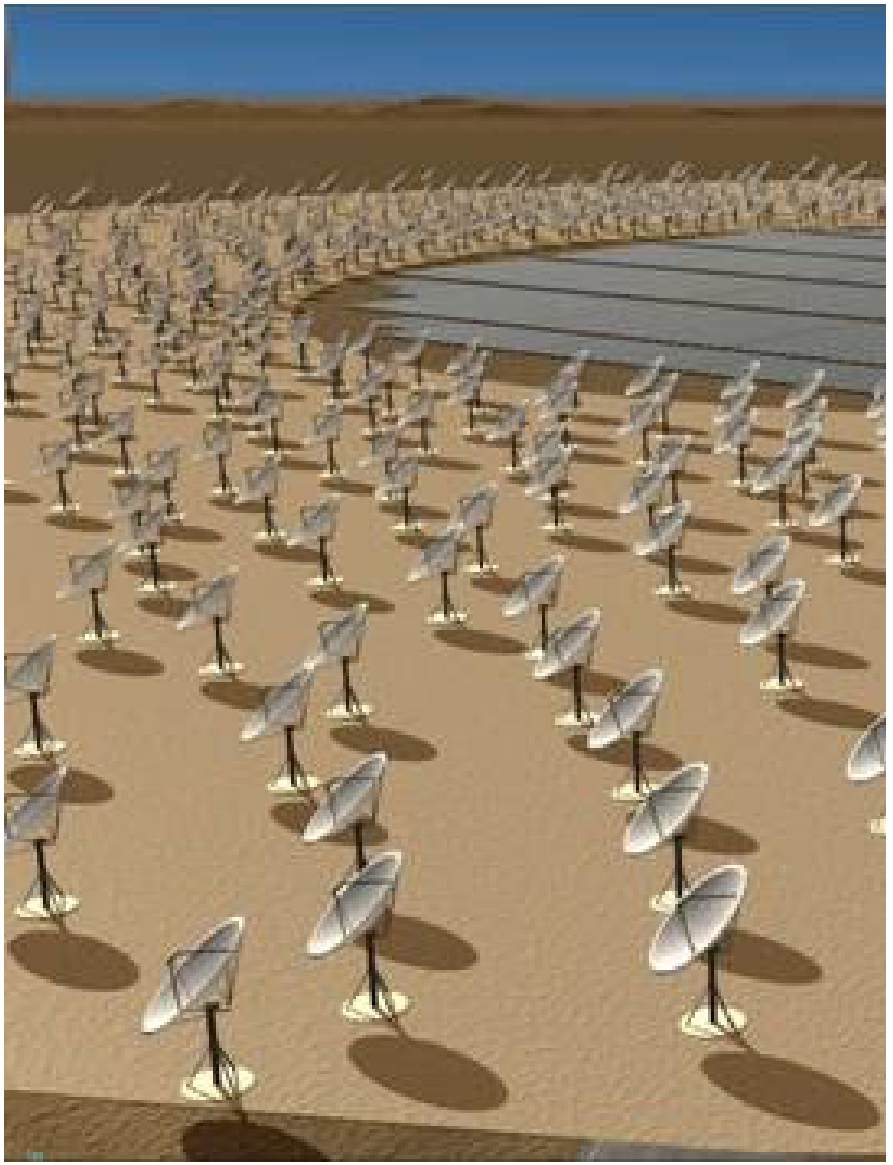
- Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical Science 16.3 (2001): 199-231.
- Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. "The elements of statistical learning." Vol. 1. Springer, Berlin: Springer series in statistics,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

IBM's Infographic



Classic science

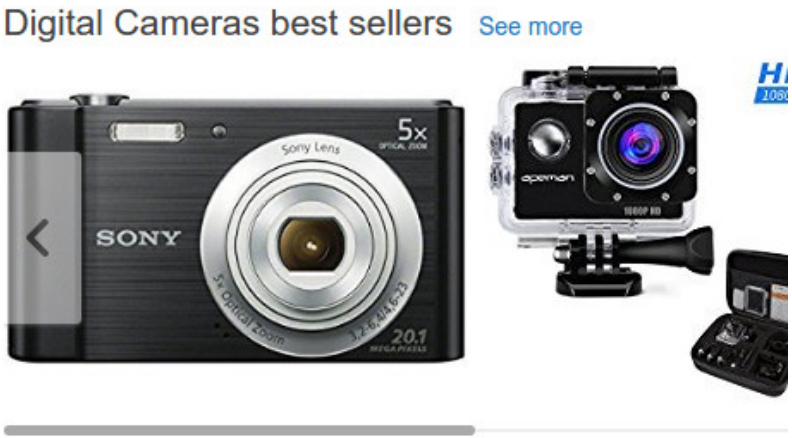
- The original data science field
- SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6e+6 TB), (Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the Big Data Era." Data Science Journal 14 (2015).)[1]
- Bioinformatics
- Medical science



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

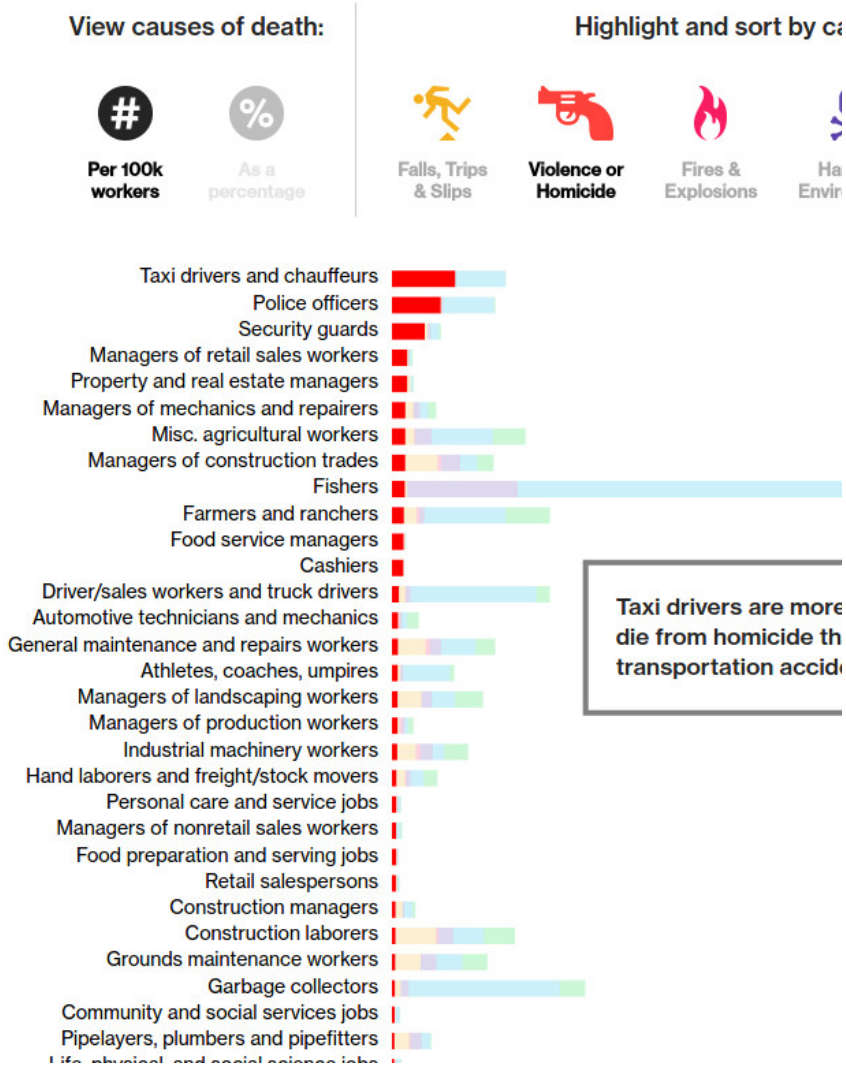
Recommender Systems

- 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

"...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 can88 – i.e. far beyond the normal skills applied in political campaigns..."

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

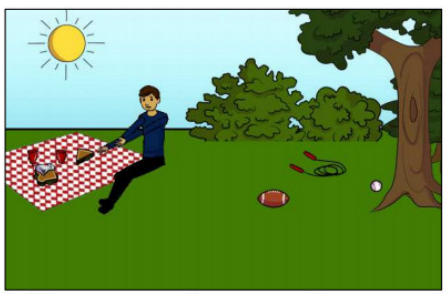
[4] `https://dominiccummings.wordpress.com/2016/10/29/on-the-referendum-20-the-campaign-physics-and-data-science-vote-leaves-voter-intention-collection-system-vics-now-available-for-all/`

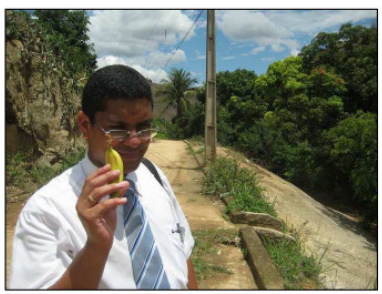
Question answering

- e.g., Antol, Stanislaw, et al. “VQA: Visual question answering.” Proceedings of the IEEE International Conference on Computer Vision. 2015.[5]
- Input can be videos, websites, etc.
- Think Google


What color are her eyes?
What is the mustache made of?


How many slices of pizza are there?
Is this a vegetarian pizza?


Is this person expecting company?
What is just under the tree?


Does it appear to be rainy?
Does this person have 20/20 vision?

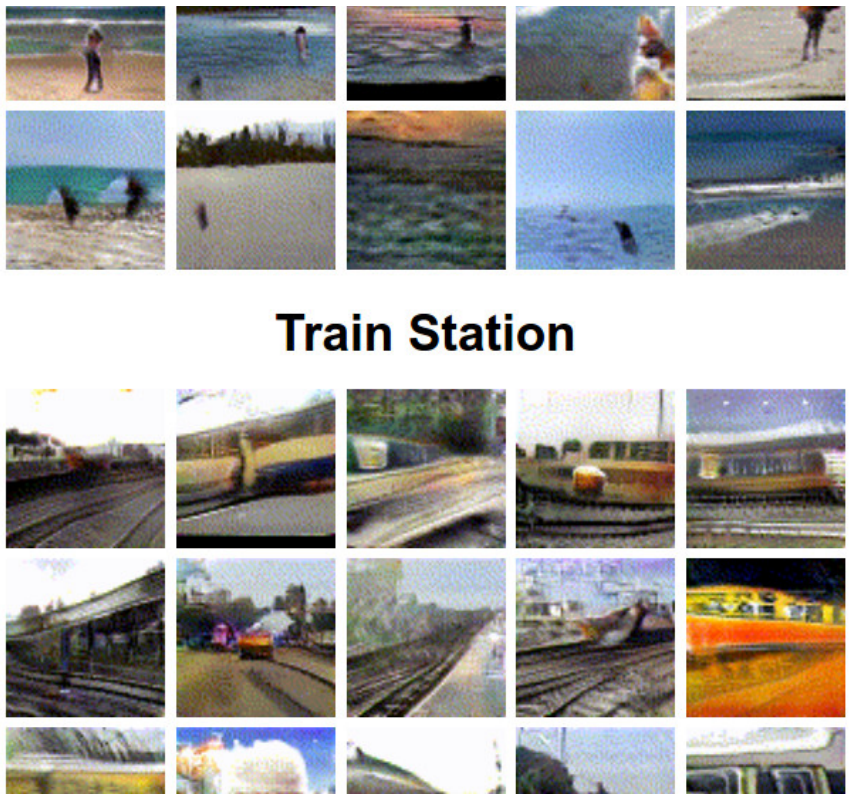
[6] http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

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)

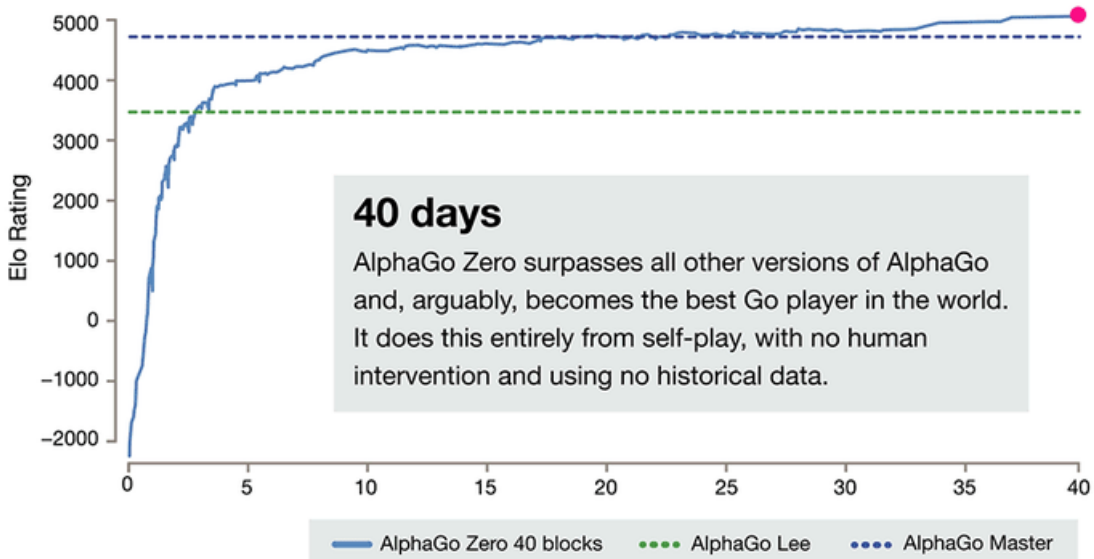
- e.g. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." Advances In Neural Information Processing Systems. 2016.[6]
- Generate an artefact
  - Generate videos
  - Generate text
  - Generate music



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

Game playing

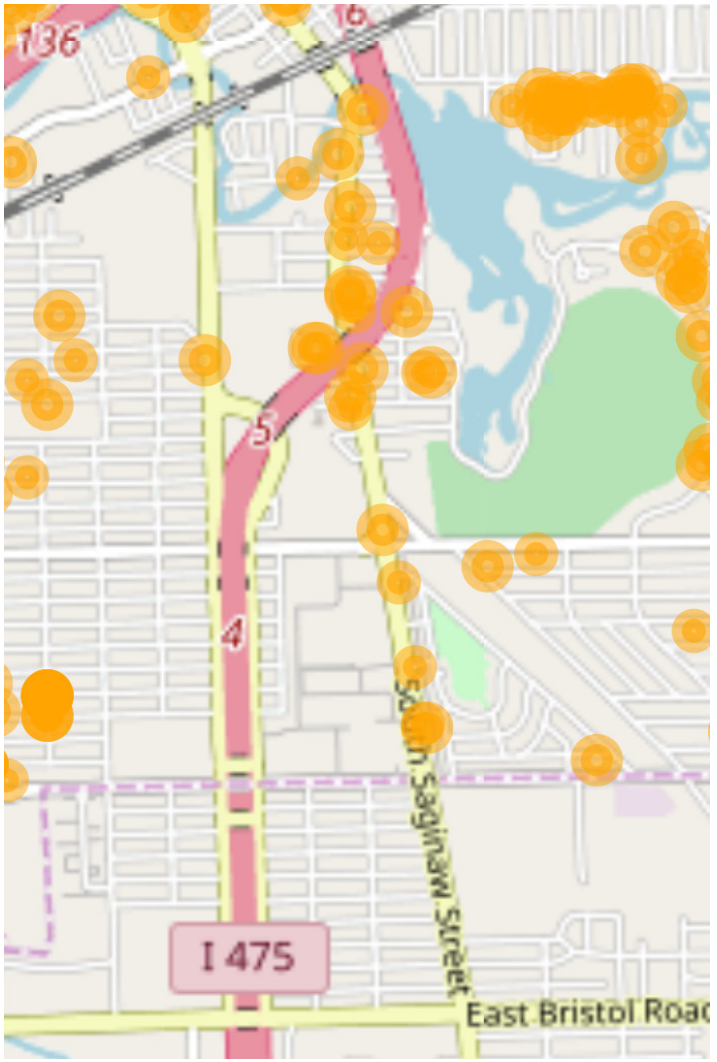
- Go, Chess machines are superhuman with no embedded human knowledge
- Heads-up limit/no limit poker - same deal, different kind of game



[7] "DeepMind AlphaGo Zero learns on its own without meatbag intervention" http://www.zdnet.com/article/deepmind-alphago-zero-learns-on-its-own-without-meatbag-intervention/

Public Health

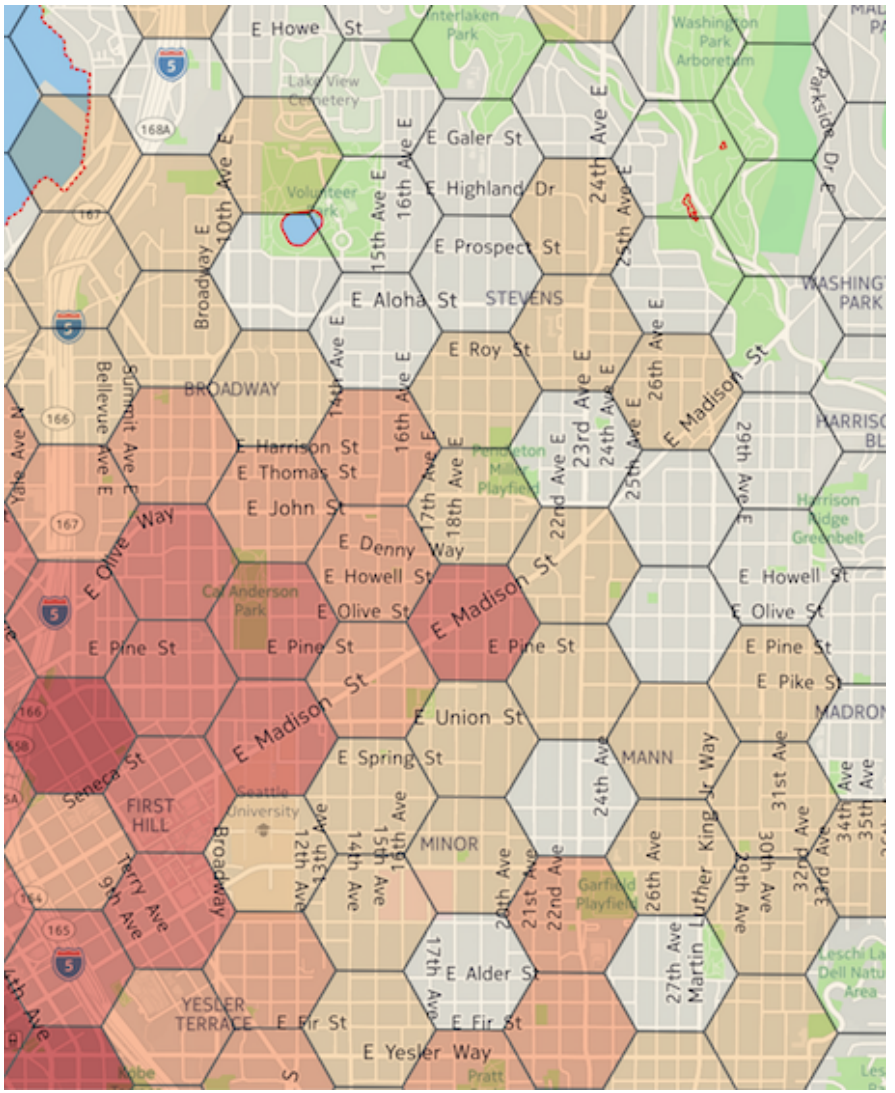
- University of Michigan, Flint Water Crisis
- "There is lead in Flint's water. Where it is? Which homes are most at risk? When will the lead levels decrease?"
- "We have data for over 8,000 properties, but there are over 50,000 parcels in Flint. Which of the not-yet-tested properties are at risk?"



[8] The Michigan Data Science Team (MDST) Work on the Flint Water Crisis" http://web.eecs.umich.edu/~jabernet/FlintWater/data_dive_summary.html

Predictive Firefighting

- FireCast
  - Risk of each building catching fire
  - Collects about 60 features per building
  - (V.3) Input for each building of about 8K features!
- Image from Seattle
- Act on Risk



[9] New York City Fights Fire with Data
http://www.govtech.com/public-safety/New-York-City-Fights-Fire-with-Data.html

Interventions

- "The collection of delinquent fines is a massive public administrative challenge. In the United Kingdom for instance, unpaid court fines amounted to more than £600 million in 2011"
- Send personalized text messages/emails, tailored to individuals needs

[2] Assessing the Effectiveness of Alternative Text Messages to Improve Collection of Delinquent Fines in the United Kingdom`https://www.povertyactionlab.org/evaluation/assessing-effectiveness-alternative-text-messages-improve-collection-delinquent-fines`

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?
- break_down: will it break down within three months of our inspection date?

takes_off_road	company_vehicle	is_over_30	regular_service	break_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 so, with what probability?
- Can you learn a model, that if provided with a tuple <takes_off_road, company_vehicle, is_over_30, regular_service> predicts *break_down*?
- The tuple represents a vehicle
- Columns are called *features*
- This is equivalent to: given a model M, can we 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*

Inferring what-if scenarios from the data

- Say your vehicle broke down
- What would have happened if you have not driven it 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

Acquiring new data

- We can't really answer what would happen to the vehicle from the data collected already
- We might need to set a controlled experiment where:
  - We find vehicles with 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

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

Generate new data

- Can I generate fictional vehicles and their properties?
- You can then use your plausible, but fictional vehicles for entertainment

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

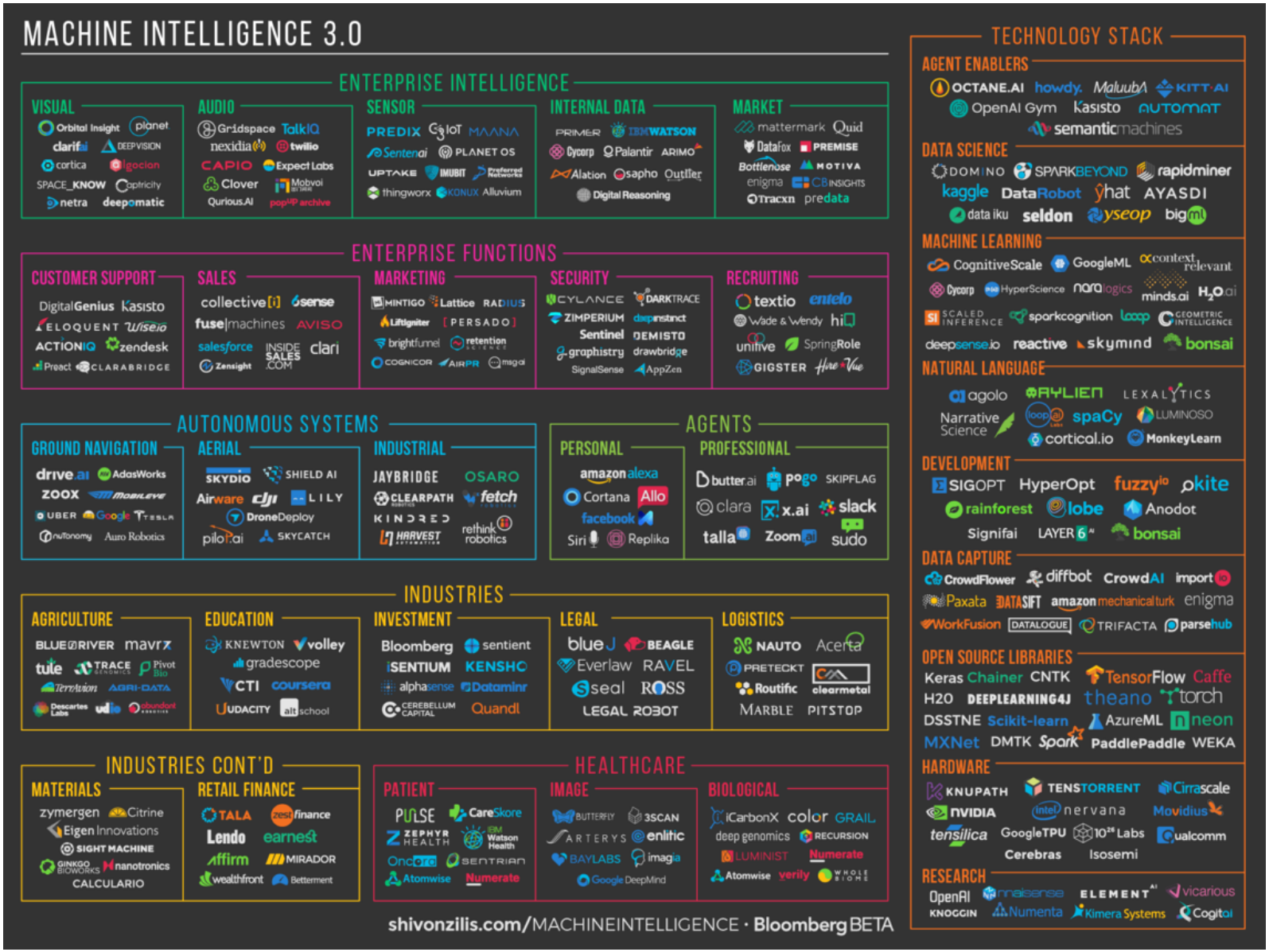
Linking with other data/collecting labels

- What if the data we have are 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 it to label the data?
  - Active learning
  - Labelled data often very expensive

Making decisions from data

- Now 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

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% 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, car sensors, etc.
  - Soon (or already!) 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

Linux Virtual Machine

- Download the VM for this module **here**
- The VM contains all (or most) of what you need to create a successful Python project
- You need to have a USB stick where you should copy the VM folder (after you un-zip the archive)
- More about this in future labs

Python

- Python is the programming language for 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/Jupyter

- A "better" command line interpreter for Python
- Has something called a "notebook"
  - A notebook combines code + natural language
- See **here** for a very nice example:

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
- The focus of today's lab

Scipy

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

Scikit-learn

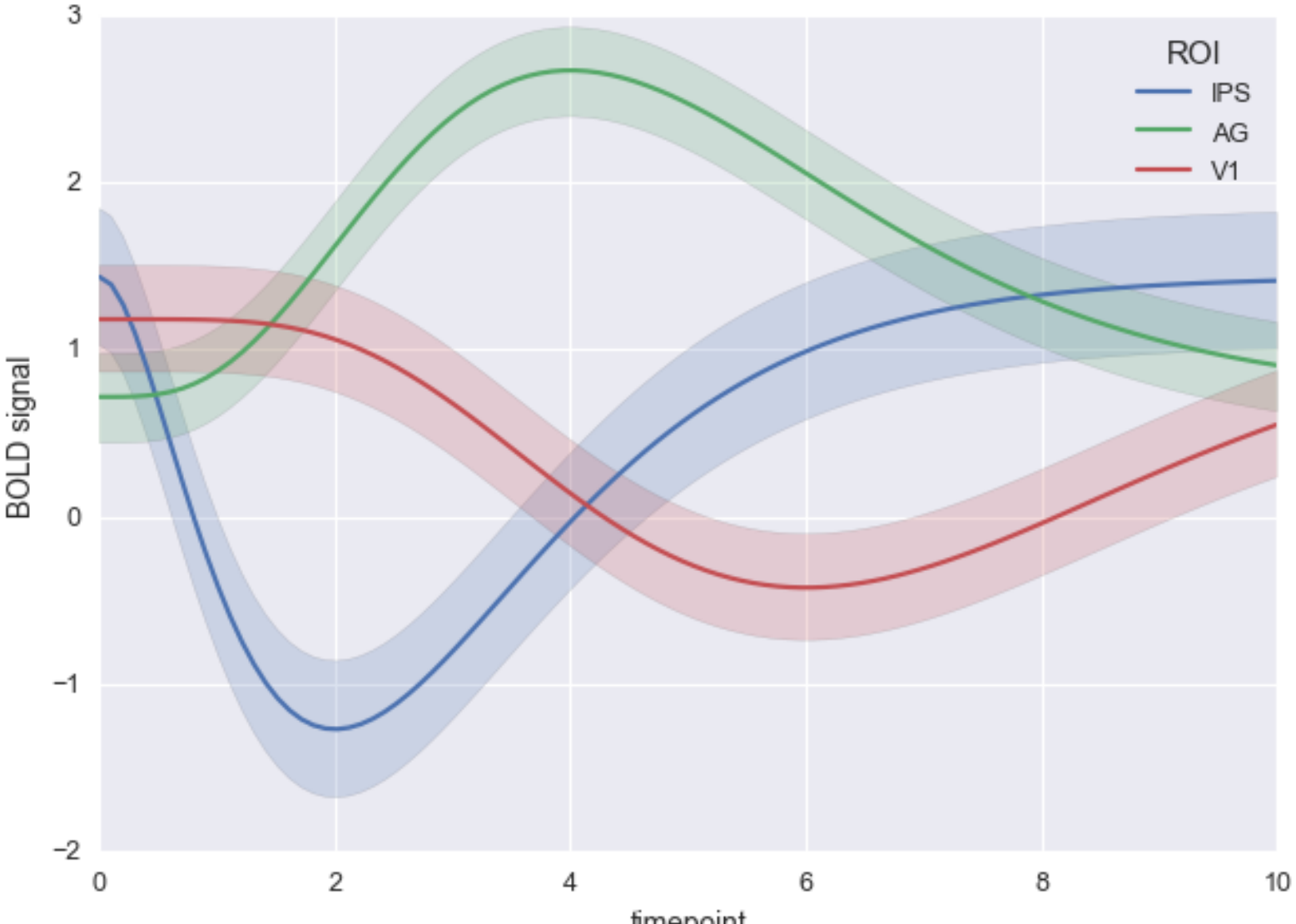
- A machine learning framework
- Includes almost everything, apart from neural networks
- We are going to use it extensively
- Super-fast trees
- Excellent documentation

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



Pandas

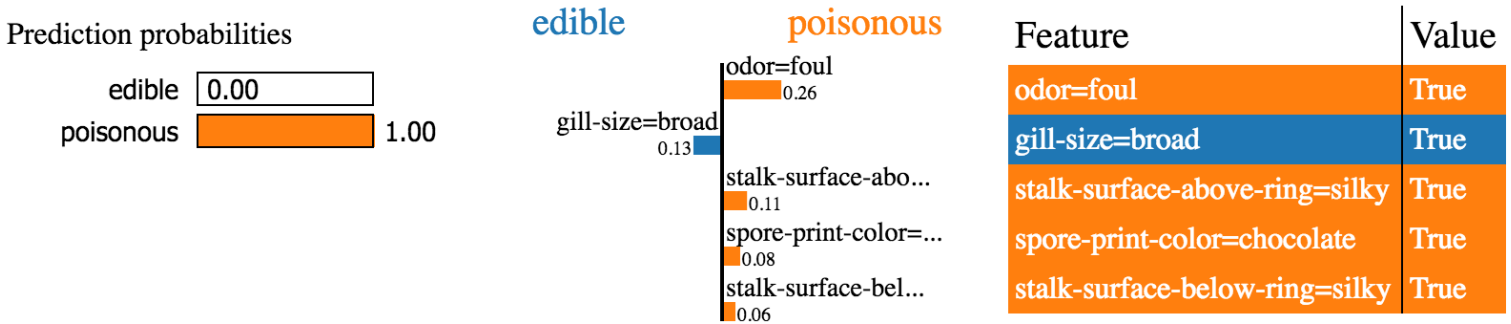
- *R* had dataframes
  - Essentially, a very SQL-like/table-like data structure
- "A Pandas DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dictionary of Series objects. It is generally the most 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

LIME

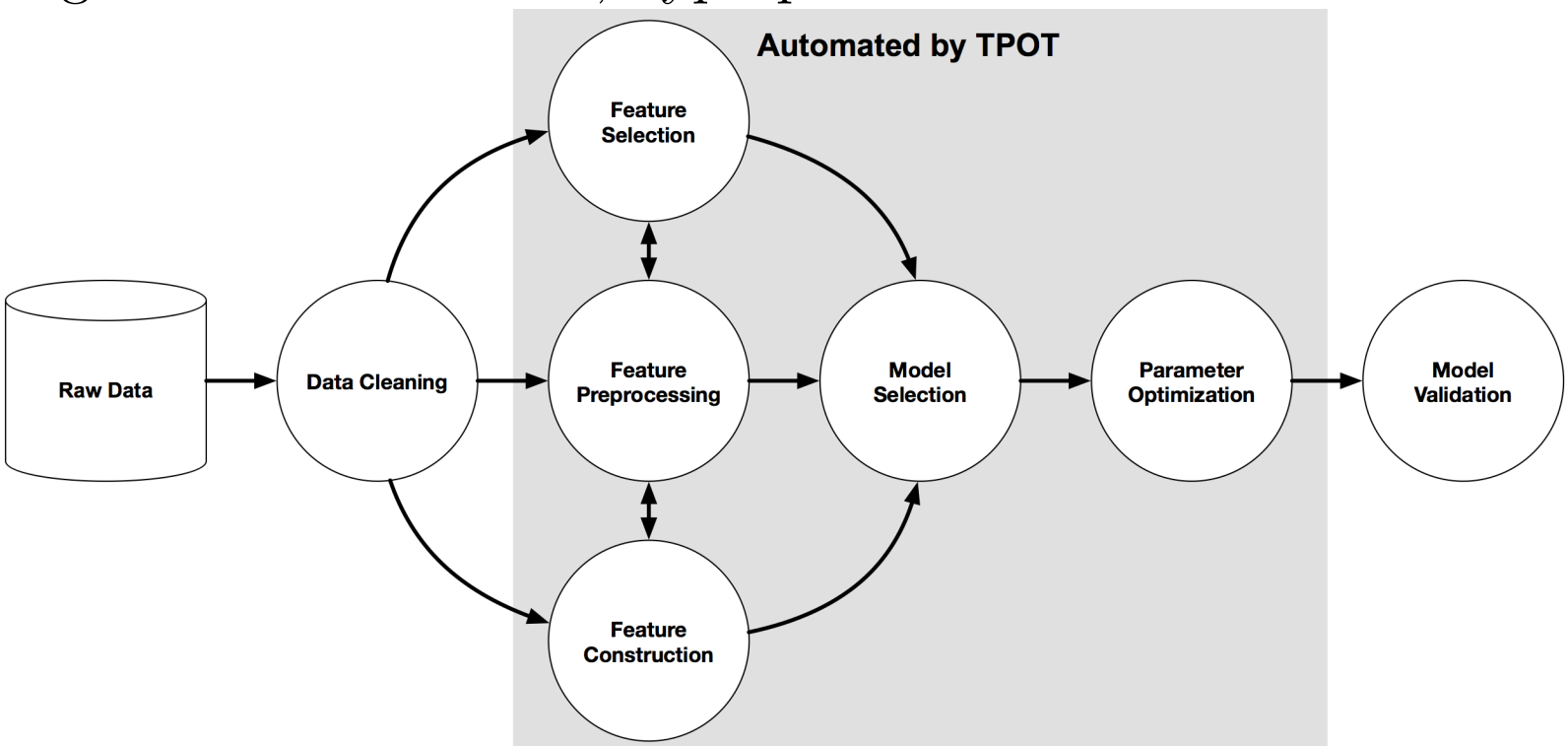
- `https://github.com/marcotcr/lime`
- LIME is a project about explaining what machine learning classifiers (or models) are doing.



- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier
- With GDPR, it is now a legal requirement to be able to explain your models

T-POT

- The ML pipeline is getting more and more complicated
- A number of tools has been developed to automate algorithm design
- E.g. feature extraction, hyperparameter choices etc.



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 scikit-learn
- Sits on top of HDFS (which we are going to install and use later on)

Github

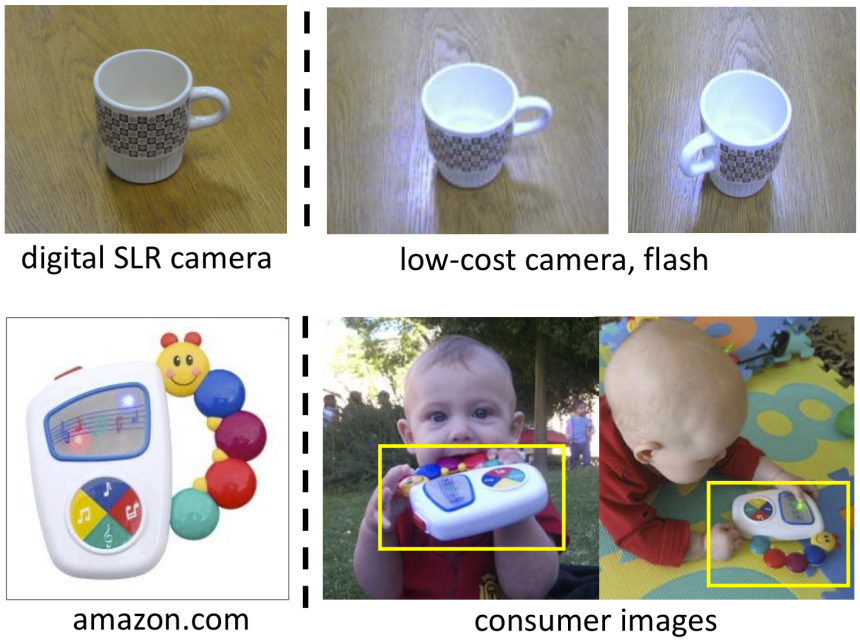
- All your code for your project will need to be publicly available
- Create a github account if you don't have one
- Two directories:
  - `/pdf` for the pdf of the project (the 10-page report)
  - `/src` for the code
  - If you have an IPython notebook (.ipynb) it should go in `/src`
- Add a README.md as well!
- And a `/labs` folder for your work in the labs

Assignments

- 7 Projects
  - You will be assigned to one of these projects randomly
  - Individual assignments
  - You are encouraged to discuss within the group
  - **But it's still your own project**
- Work on your own
- DO NOT WAIT UNTIL THE VERY LAST MINUTE, EXPERIMENTS TAKE TIME
- Project Proposal/Initial Report deadline: **21-Feb-2019 11:59:59**
- Final Project Deadline: **25-Apr-2019 11:59:59**

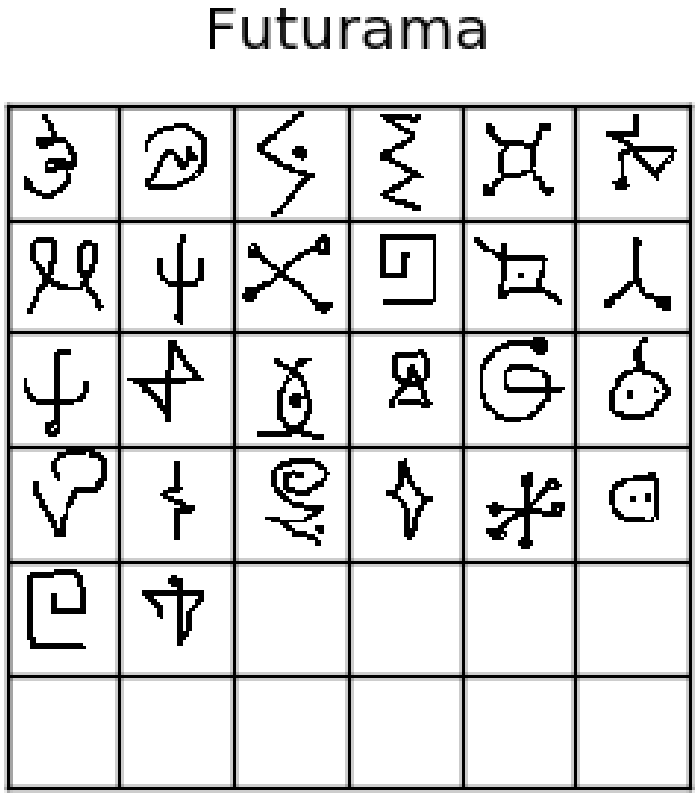
Domain Adaptation

- The usual assumption is that the training and test set come from the same distribution
- This is not always the case - in face almost never
- What can we do about this?
- Auto-ML for domain adaptation



One-Shot learning

- Most machine learning algorithms require a huge amount of data
  - This is not always possible
- Humans tend to generalise nicely using very few data samples
- Massive datasets not always available
- Auto-ML and Metric Learning



Reinforcement Learning and Interpretability

- We need to be able to explain the models
- This is a a *legal* requirement
- We will use LIME to try and interpret some game-playing agents

Continual Learning

- One of the hottest problems in ML right now
- *Humans forget, Machines tend to forget catastrophically*
- Most ML algorithms cannot learn without forgetting all past experience

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