Data Science

Spyros Samothrakis
Research Fellow, IADS
University of Essex

February 6, 2017

Better science through data

Hey, Tony, Stewart Tansley, and Kristin M. Tolle. "Jim Gray on eScience: a transformed scientific method." (2009).

- Thousand years ago: empirical branch
  - You observed stuff and you wrote down about it
- Last few hundred years: theoretical branch
  - Equations of gravity, equations of electromagnetism
- Last few decades: computational branch
  - Modelling at the micro level, observing at the macro level
- Today: data exploration
  - Let machines create models using vast amounts of data

Mixing statistics, philosophy of science and machine learning

- Wu, C. F. J. "Statistics= data science." (1997).
- Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical Science 16.3 (2001): 199-231.
- Science is the epistemology of causation
- Data science is basically science on arbitrary data
  - But quite often we only care about predictions
- Possibly a re-branding of data mining, machine learning, artificial intelligence, statistics

Better business through data

- There was a report by Mckinsey

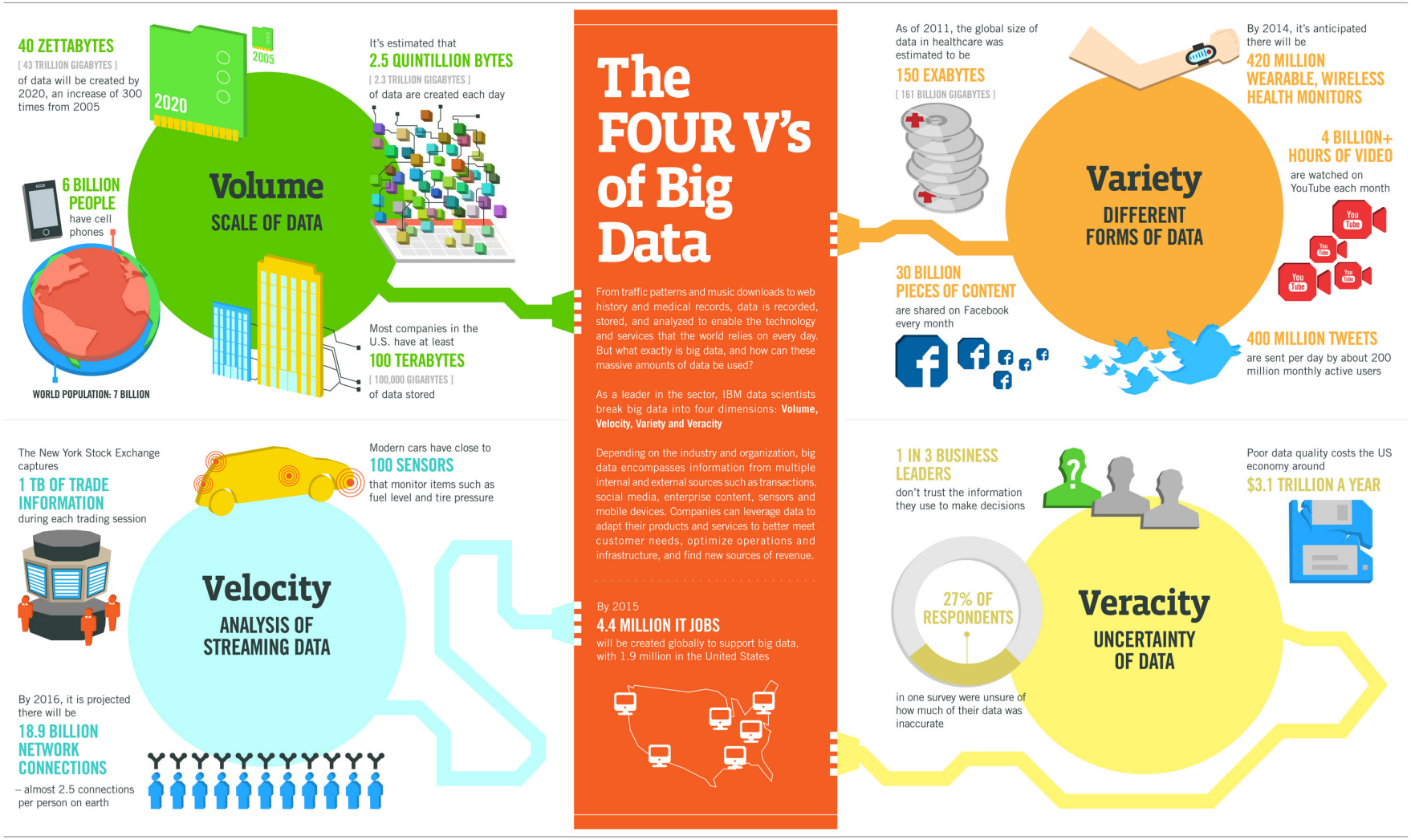
Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).

- Urges everyone to monetise "Big Data"
- Use the data provided within your organisation to gain insights
- Has some numbers as to how much this is worth
- Proposes a number of methods, most of them associated with machine learning and databases

More is different

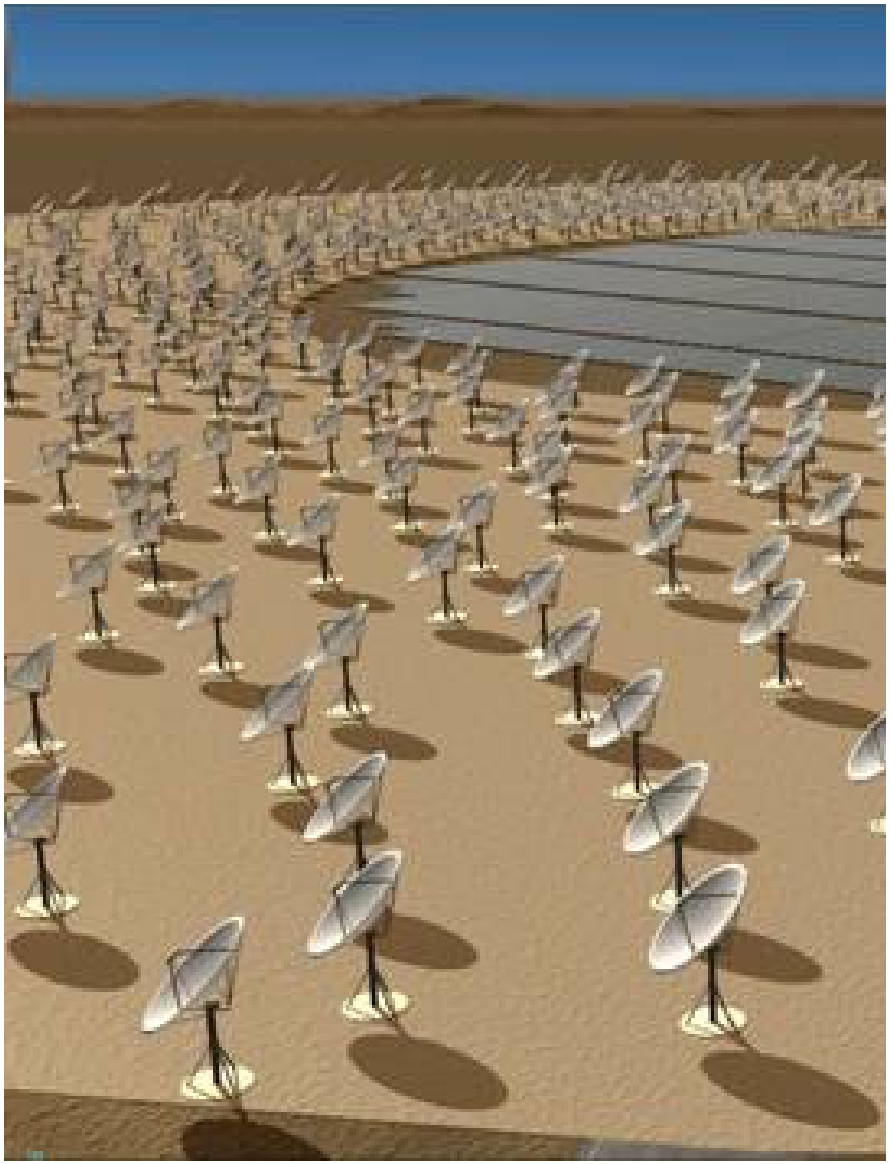
- Anderson, Philip W. "More is different." Science 177.4047 (1972): 393-396.
- The idea of emergence
- You put stuff together, you go from physics to chemistry
- . . . from chemistry to biology
- . . . from biology to psychology and zoology
- . . . from psychology to sociology
- "quantity changes into quality"

IBM's Infographic



Classic science

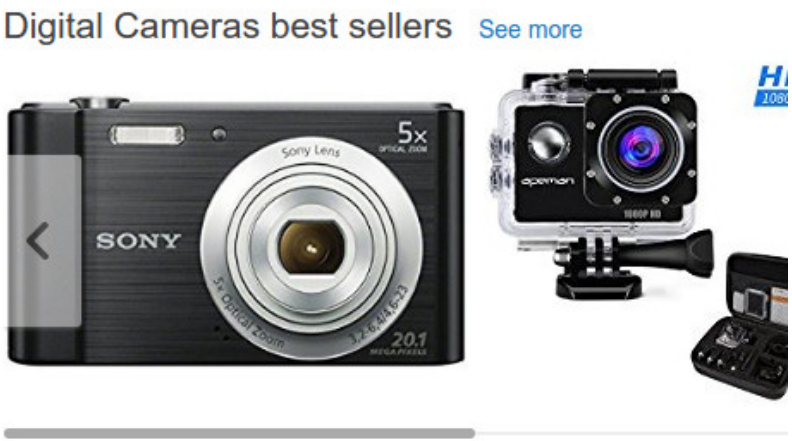
- The original data science field
- SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6e+6 TB), (Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the Big Data Era." Data Science Journal 14 (2015).)[1]
- Bioinformatics
- Medical science



[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

- Wikileak style data dumps are everywhere
- The Ashley-Madison Affair, 2015
- "Just three in every 10,000 female accounts on infidelity website are real"
- "The website claims 5.5 million of its 37 million accounts are 'female' "

[2] http://www.independent.co.uk/life-style/gadgets-and-tech/news/ashley-madison-hack-just-three-in-every-10000-female-accounts-on-infidelity-website-are-real-10475310.html

Finance & Insurance

- Predict stock prices (Hedge funds)
- Insurance models
- Credit score
- In fact, a lot of trading that currently happens is algorithmic trading[2]
- 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[3]

[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[4]
- 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.[5]
- 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

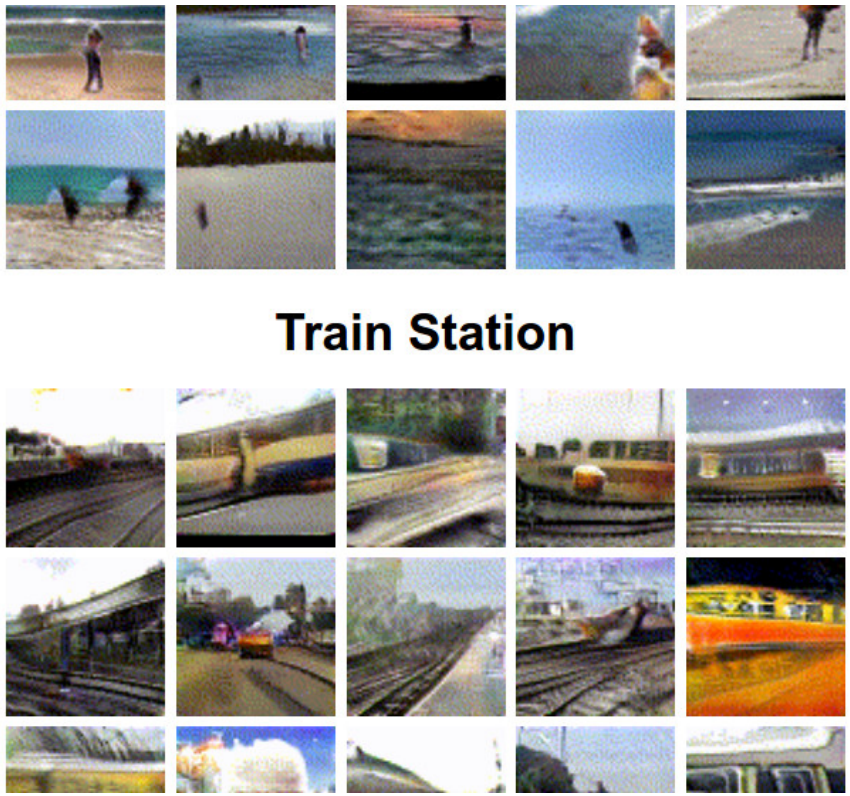
- Is a new product I just created well received by our customers?
- Is a new marketing campaign e-mail sent detrimental to our efforts?
- What is the content a chain of e-mails should have?
- What adverts should I present to a user?

Business Analytics

- Churn models
  - Why are my customers leaving?
- Customer segmentation
  - What kinds of customers do I have?
  - Is a specific customer of a certain kind?
- Product development
  - What is a successful product?

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

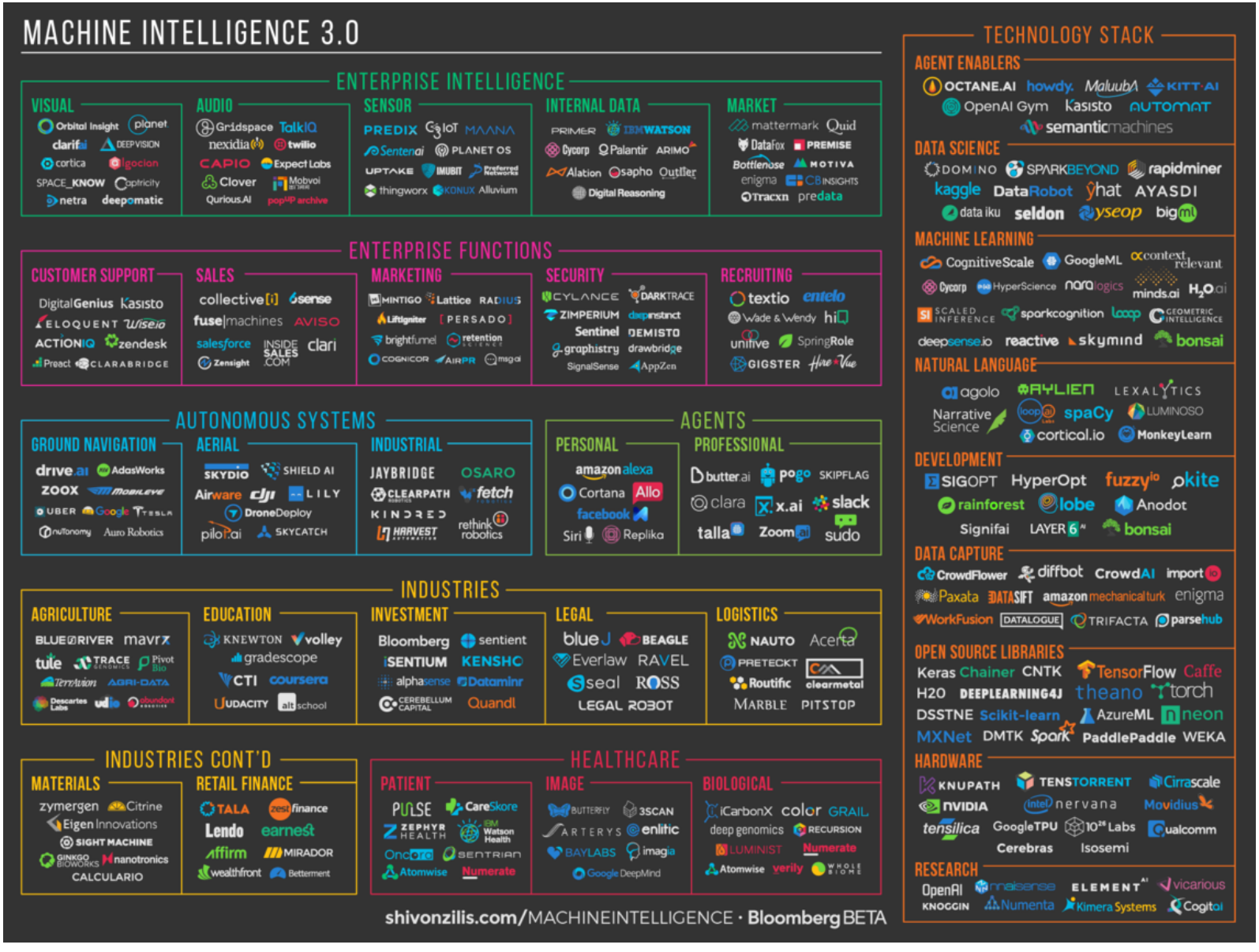


[6] http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

Game playing

- 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

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?

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*

Final remarks

- This is a huge field
- Question almost everything you read about statistics
- Startups are being taken over left and right
- Big business is investing mega-dollars/pounds/etc
- Small businesses?