Data Science

Spyros Samothrakis
Deputy Director, IADS
University of Essex

February 12, 2019

Data Science

Analytics

Examples

Prescriptive Analytics

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 often science on observational 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 that 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

Overview - Analytics

- Descriptive
  - "I'll create high level, compressed views of your data"
  - What is sometimes termed analytics, unsupervised learning
- Predictive
  - "I'll try to predict a possible version of your future".
  - Show me some good customers in your database, I'll try getting you more
  - Supervised learning (mostly)
- Prescriptive
  - "What should I do to achieve certain results?"
  - **Reinforcement Learning, bandits and causality**
    - e.g. Send the right sequence of e-mails to engage an audience

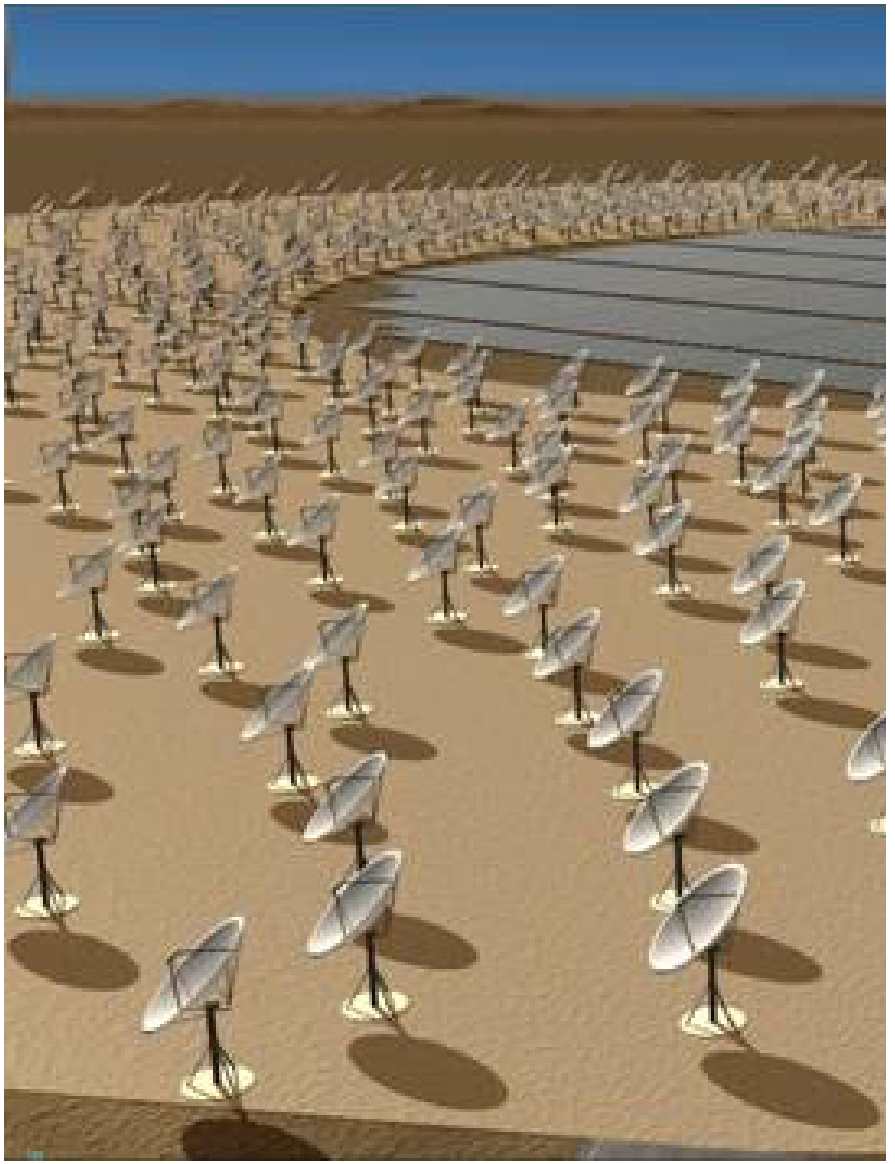
[1] Chris Wiggins. "Lectures delivered Aug 8-9, 2016 at MLSS.cc (Arequipa, Peru).
https://www.slideshare.net/chrishwiggins/machine-learning-summer-school-2016/75

Waaaaaait...

- "We've been doing this thing for years"
  - econometrics
  - statistics
  - you-name-it
- Mostly true, in fact most of these ideas have been developed in parallel in many communities
  - Deep insights, larger scale

Classic science

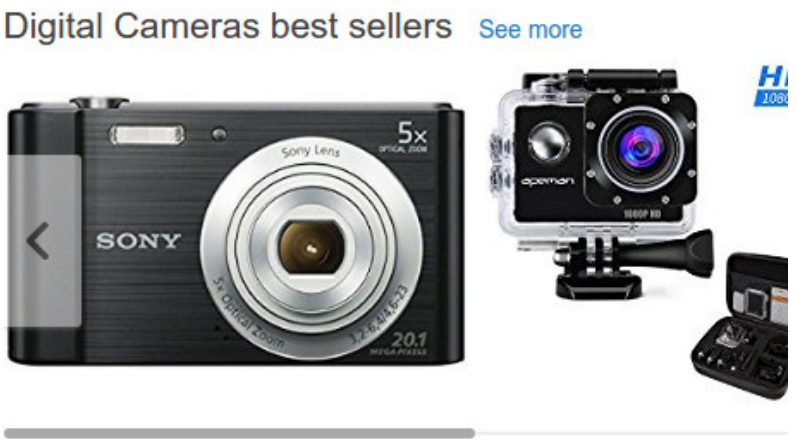
- The original data science field
- SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6e+6 TB), (Zhang, Yanxia, and Yongheng Zhao. "Astronomy in the Big Data Era." Data Science Journal 14 (2015).)[1]
- Bioinformatics
- Medical science



[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
- Insurance models
- Credit scores
- Sudden drops in share prices sometimes 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[2]

[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/

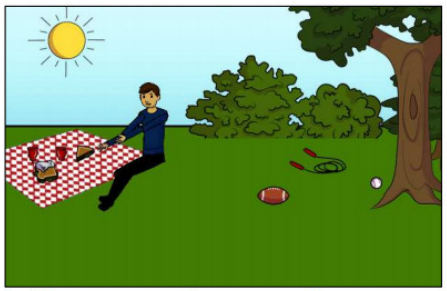
Politics (historical)

- New Yorker - THE PLANNING MACHINE: Project Cybersyn and the origins of the Big Data nation[3]
- 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.[4]
- 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

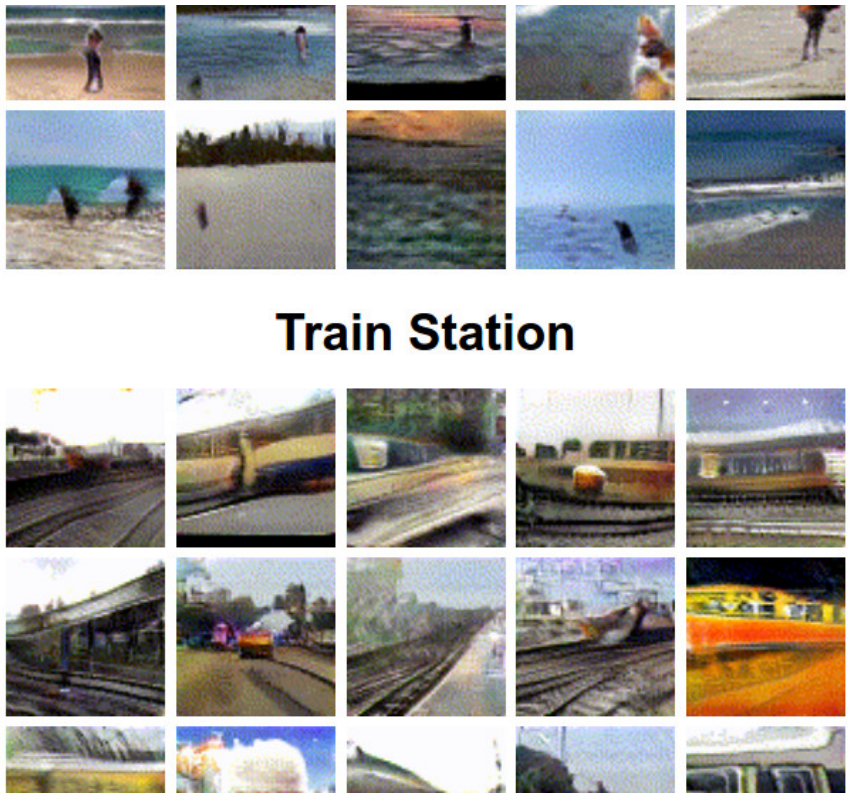
- Is a new product I just created well received by our customers?
- Is a new marketing campaign e-mail sent detrimental to our efforts?
- What is the content a chain of e-mails should have?
- What adverts should I present to a user?

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)

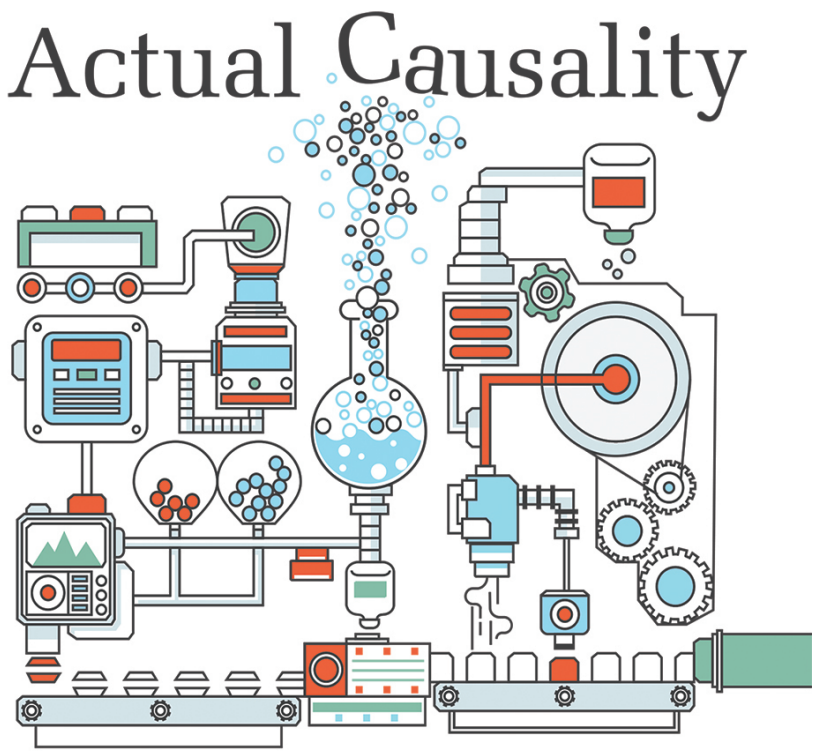
- e.g. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." Advances In Neural Information Processing Systems. 2016.[5]
- Generate an artefact
  - Generate videos
  - Generate text
  - Generate music



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

Causality / off-policy reinforcement learning

- Forward Causality
  - "The effects of causes" - ML Causality
  - "How much would my sales increase if I send more aggressive e-mails to my customers?"
- Backward Causality
  - "The causes of effects" - ML interpretability on causally consistent models
  - "What was the main cause of the drop in sales?"



Joseph Y. Halpern

Reinforcement learning / bandits

- What if we had access to a simulator?
  - Or close - we might be able to build one using our data
- We can search the space of available policies using the simulator and find the best one
- . . . then deploy in the real world

Our Capacity

- BLG Data Research Center
- IADS
- iads@essex.ac.uk, ssamot@essex.ac.uk