

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

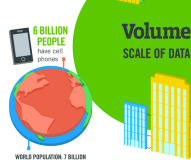
MORE IS DIFFERENT

- ▶ 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

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by
2020, an increase of 300
times from 2005



It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



Most companies in the
U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange captures

**1 TB OF TRADE
INFORMATION**
during each trading session



By 2016, it is projected
there will be
**18.9 BILLION
NETWORK
CONNECTIONS**
— almost 2.5 connections
per person on earth



Modern cars have close to
100 SENSORS
that monitor items such as
fuel level and tire pressure

Velocity ANALYSIS OF STREAMING DATA



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States.



As of 2011, the global size of
data in healthcare was
estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook
every month



Variety DIFFERENT FORMS OF DATA



By 2014, it's anticipated
there will be
**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**

**4 BILLION+
HOURS OF VIDEO**
are watched on
YouTube each month



400 MILLION TWEETS
are sent per day by about 200
million monthly active users



1 IN 3 BUSINESS LEADERS

don't trust the information
they use to make decisions



in one survey were unsure of
how much of their data was
inaccurate

Veracity UNCERTAINTY OF DATA

Poor data quality costs the US
economy around
\$3.1 TRILLION A YEAR



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTTEC, SAS

IBM

CLASSIC SCIENCE

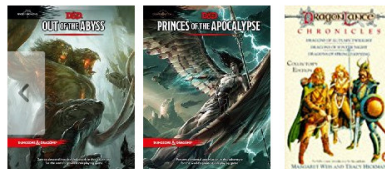
- ▶ The original data science field
- ▶ SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6×10^6 TB), (Zhang, Yanxia, and Yongheng Zhao. “Astronomy in the Big Data Era.” *Data Science Journal* 14 (2015).)¹
- ▶ Bioinformatics
- ▶ Medical science



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

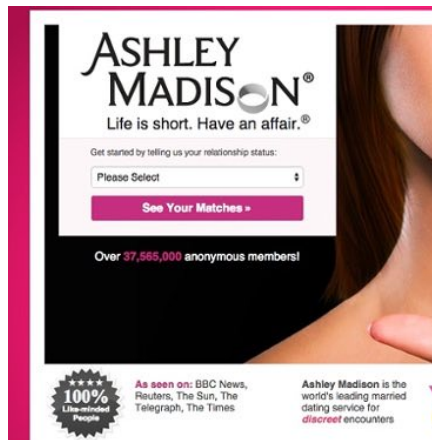


Digital Cameras best sellers [See more](#)



DATA JOURNALISM

- ▶ Wikileaks 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 trading²
- ▶ 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*) Blog³

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 nation⁴
- ▶ 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



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?

⁶http://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

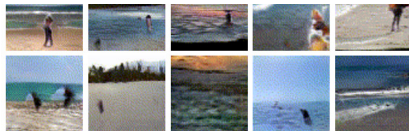
- ▶ Churn models
 - ▶ Why are my customers leaving?
- ▶ Customer segmentation
 - ▶ What kinds of customers do I have?
 - ▶ Is a specific customer of a certain kind?
- ▶ Product development
 - ▶ What is a successful product?

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



Train Station



⁶http://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

MACHINE INTELLIGENCE 3.0

ENTERPRISE INTELLIGENCE

VISUAL Ortelius Insight Planet darfon DEPRIVISION cortica iQvision SPACE KNOW iQvision netra deepomatic	AUDIO Gridspace TalkIQ nexidia CAPIO Expect Labs Clover Qurbulous payit archive	SENSOR PREDIX GYOT MAANA Sentinal PLANET OS UPTAKE thingwork KODAK Aluatum	INTERNAL DATA PREMIER DEWATSON Dypar Palantir ADAMO Alation Sapho Outlier Digital Reasoning	MARKET mattermark Quid Datafux PREMISE Bottlenuse enigma OTRACK predata
---	--	--	--	---

ENTERPRISE FUNCTIONS

CUSTOMER SUPPORT DigitalGenius Kasisto Eloquent T/S/Seo ACTIONIQ Zendesk Pinct CLARABRIDGE	SALES collective sense fuse/machines AVISO salesforce INSIDE Zennight Zennight clari COM	MARKETING MINTIGO Lattice RADIUS Lattice PERSADO brightfunnel retention COSMOCON AURA Cmapa	SECURITY CYCLANCE DARKTRACE ZIMPERIUM dependant Sentinel DEMISTO graphistry drawbridge SignalSense AppZen	RECRUITING textio envilo Wide & Wandy h univise SpringRole GIGSTER HireVue
---	--	---	---	---

AUTONOMOUS SYSTEMS

GROUND NAVIGATION drive.ai AdaraWorks ZOOX USER Google TIRSLA AutoRobotics	AERIAL SKYDIO SHIELD AI Airware DJI DroneDeploy plato SKYWATCH	INDUSTRIAL JAYBRIDGE OSARO CLEARPATH fetch KINQ3ED HANWIST nathis robotics	PERSONAL amazon alexa Cortana ALO facebook Siri Repika	AGENTS PROFESSIONAL butter.ai POPS SKIPFLAG clara x.ai slack tallia Zoom sudo
---	---	--	---	---

INDUSTRIES

AGRICULTURE BLUE DRIVER mavix tule TRACE P Dribbble AGRI-DATA iStockphoto	EDUCATION KNEWTON volley gradescope CTI COURSERA UUDACITY school	INVESTMENT Bloomberg iSENTIUM KENSHC alphaSense Dotominr Coudant	LEGAL blue J BEAGLE Everlaw RAVEL Seal ROSS LEGAL ROBOT	LOGISTICS MAUTO Acenta PRETECKT Routific clearmatel MARBLE PITSTOP
--	---	---	--	---

INDUSTRIES CONT'D

MATERIALS zymergen Citrine Eigen Innovations M nanoionics CALCULARIO	RETAIL FINANCE TALA finance Lendo earnest affirm MIRADOR wealthfront Betterment	PATIENT PULSE CareStore ZENITH HEALTH Oncoda SEPTENIA Atomwise Numerate	IMAGE BUTTERFLY 3SCAN ARTERYS enlitic BAYLABS imago Google DeepMind	BIOLOGICAL CarbonX color GRAIL deep genomics RECURSION UMINIST illuminate Atomwise verity
---	--	--	--	--

TECHNOLOGY STACK

AGENT ENABLERS
 OCTANE.AI howdy Maluba KITT.AI
 OpenAI Gym Kasisto AUTOMAT
 semanticmachines

DATA SCIENCE
 DOMINO SPARKBEYOND rapidminer
 kaggle DataRobot yhat AYASDI
 dataiku seldon yscope bigml

MACHINE LEARNING
 CognitiveScale GoogleML context relevant
 Dypar HyperScience n2o logic minds.ai H2o.ai
 SCALES INFERENCE sparkognition loop
 deepsense reactive skyminde bonsai

NATURAL LANGUAGE
 agolo FLYLIE LEXALYTICS
 Narrative Science spaCy LUMINOSO
 cortical.io MonkeyLearn

DEVELOPMENT
 SIGOPT HyperOpt fuzzy okite
 rainforest lobe Anodot
 Signifai LAYER6 bonsai

DATA CAPTURE
 CrowdFlower diffbot CrowdAI import
 Paxala DATASET amazon mechanical turk enigma
 WorkFusion DATALOGUE TRIFACTA parsehub

OPEN SOURCE LIBRARIES
 Keras Chainer CNTK TensorFlow Caffe
 H2O DEEPLARNING4J theano torch
 DSSNET scikit-learn AzureML neon
 MXNet DMTK Spark PaddlePaddle WEKA

HARDWARE
 KNUPHAT TENSTORRENT Cirascale
 NVIDIA nvidia nervana Movius
 terisilica GoogleTPU 10 Labs
 Cerebras Isosemi

RESEARCH
 OpenAI Inria Numenta ELEMENT vicarious
 KNOGIN Numenta Kmera Systems Cogitai

shivonzills.com/MACHINEINTELLIGENCE · Bloomberg BETA

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?