Big Data, Data Science, Statistics

Nancy Reid

31 March 2017

Queen's University
Department of Mathematics and Statistics

Big Data

= Big Machines

= Lots of Computing

= Complex Architectures

= Computer Science

Small data

= equations and formulas

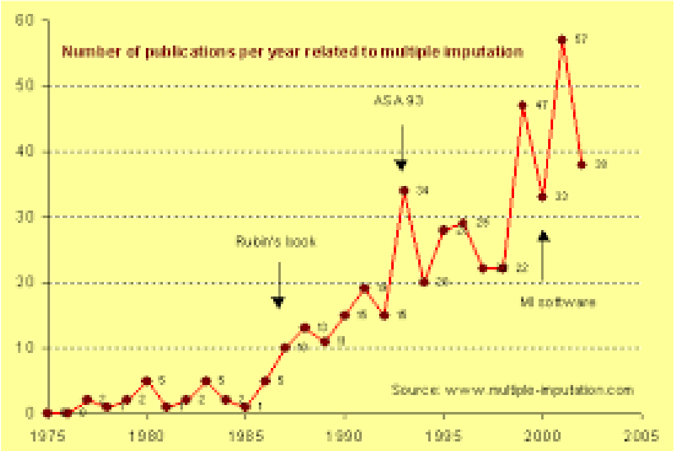
= mathematical modelling

= a little computing

= Statistical Science

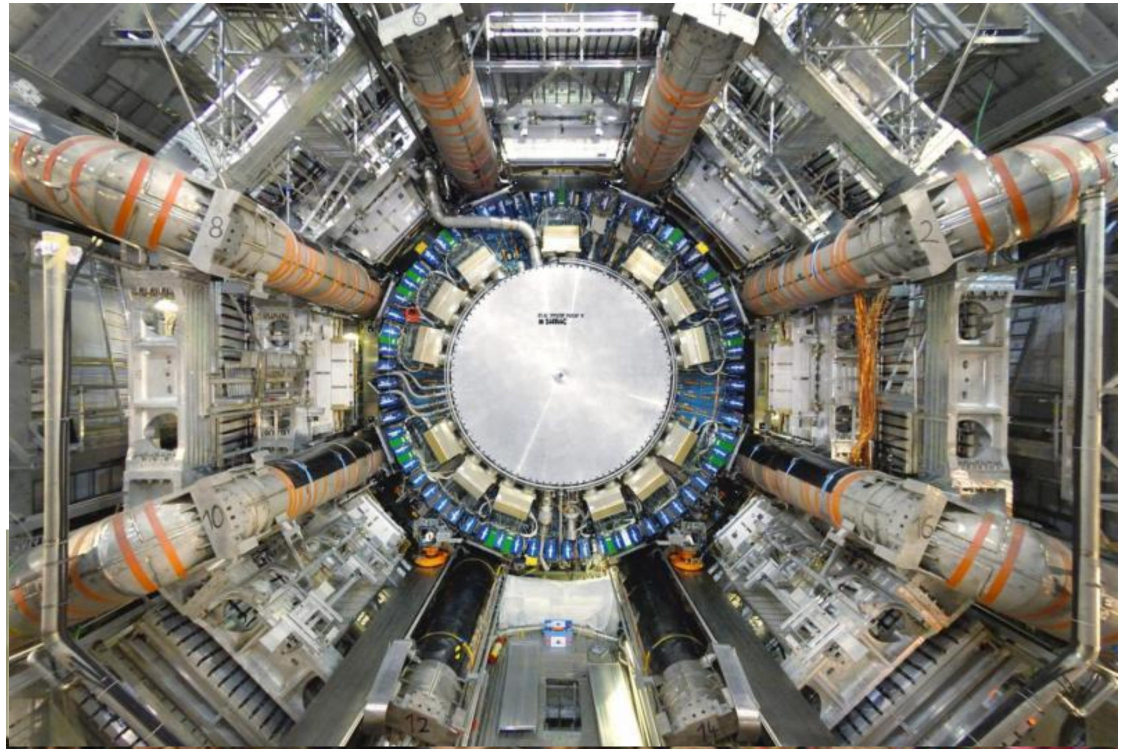
$$p(v,h;\eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\},$$

$$\eta = (a, b, W)$$



Big Data

- Interesting
- Detailed
- Informative
- Fun




Small Data

So yesterday

Small Data

FIELDS

THE FIELDS INSTITUTE

THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR BIG DATA

JANUARY - JUNE, 2015

PROGRAM

JANUARY 12 – 23, 2015

Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

JANUARY 26 – 30, 2015

Workshop on Big Data and Statistical Machine Learning
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

FEBRUARY 9 – 13 , 2015

Workshop on Optimization and Matrix Methods in Big Data

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops throughout the program will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life

Canadian Institute for Statistical Sciences

CANSSI INCASS

Fields Institute for Resesarch in the Mathematical Sciences

FIELDS

Pacific Institute for Mathematical Sciences

CRM

Centre de Recherches Mathématiques

NSERC CRSNG

Ontario

Workshops

- Opening Conference and Bootcamp
- Statistical Machine Learning
- Optimization and Matrix Methods
- Visualization: Strategies and Principles
- Big Data in Health Policy
- Big Data for Social Policy
- Networks, Web mining, and Cyber-security
- Statistical Theory for Large-scale Data
- Challenges in Environmental Science
- Complex Spatio-temporal Data
- Commercial and Retail Banking

CANSSI INCASS

The Fields Institute for Research in Mathematical Sciences

FIELDS

Fields*Live* Video Archive

CENTRE DE RECHERCHES MATHÉMATIQUES

Pacific Institute *for the* Mathematical Sciences

Opening Conference and Bootcamp

Introduction to topics at following workshops

One day on each topic

Many speakers started by trying to define big data

"I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so.

But I know it when I see it … "

Justice Potter Stewart; *Jacobellis v. Ohio* 22 June 1964

Robert Bell, Google, Plenary Opening Lecture

International Statistical Review

International Statistical Review (2016), 84, 3, 371–389 doi:10.1111/insr.12176

Statistical Inference, Learning and Models in Big Data

Beate Franke[1], Jean-François Plante[2], Ribana Roscher[3], En-Shiun Annie Lee[4], Cathal Smyth[5], Armin Hatefi[5], Fuqi Chen[6], Einat Gil[5], Alexander Schwing[5], Alessandro Selvitella[8], Michael M. Hoffman[5], Roger Grosse[5], Dieter Hendricks[7] and Nancy Reid[5]

[1]*University College London, London, UK*

[2]*HEC Montréal, Montréal, Québec, Canada*

[3]*Freie Universität, Berlin, Germany*

[4]*University of Waterloo, Waterloo, Ontario, Canada*

[5]*University of Toronto, Toronto, Ontario, Canada*
E-mail: reid@utstat.utoronto.ca

[6]*Western University, London, Ontario, Canada*

[7]*University of Witswatersrand, Johannesburg, South Africa*

[8]*McMaster University , Hamilton, Ontario, Canada*

Summary

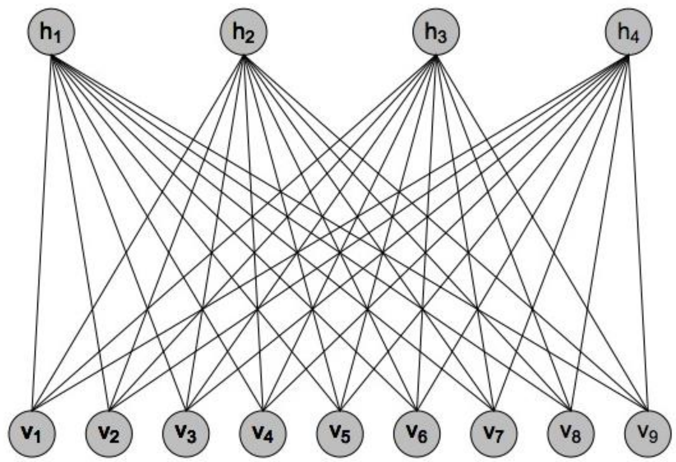
Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

- Statistical Machine Learning

Restricted Boltzmann Machine



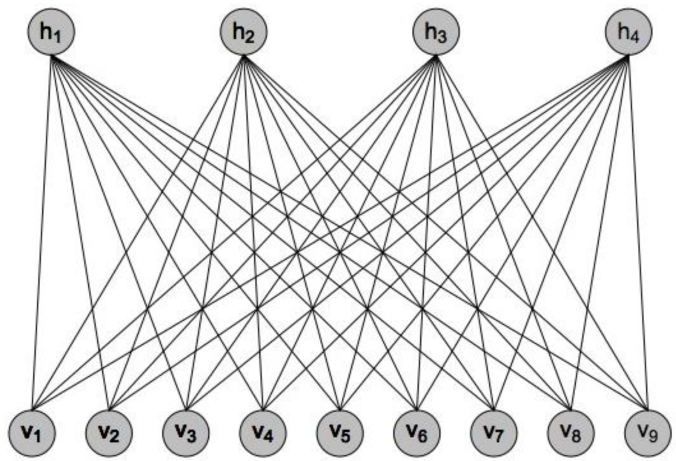
bottom nodes $\boldsymbol{v} = (v_1, v_2, \ldots)^{\mathrm{T}}$; top nodes $\boldsymbol{h} = (h_1, h_2, \ldots)^{\mathrm{T}}$

Fields Institute, Toronto, Canada © 2015 by Mu Zhu 6

Statistical Machine Learning

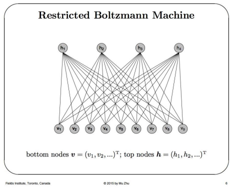
$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\} \qquad \eta = (a, b, W)$$

Restricted Boltzmann Machine



bottom nodes $\boldsymbol{v} = (v_1, v_2, ...)^{\mathrm{T}}$; top nodes $\boldsymbol{h} = (h_1, h_2, ...)^{\mathrm{T}}$

Restricted Boltzmann machine



$$f(v,h;\eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- natural gradient ascent

$$\eta \longleftarrow \eta + \epsilon\, i(\eta)^{-1} \nabla_\eta \ell(\eta; v, h)$$

$$\ell = \log f$$

$$i = \mathrm{E}(-\ell'')$$

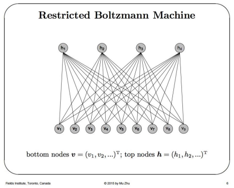
- uses Fisher information as metric tensor

Girolami and Calderhead (2011); Amari (1987); Rao (1945)

- Gaussian graphical model approximation to force sparse inverse

Grosse and Salakhutdinov (2016) 32nd Internat. Conf. on Machine Learning

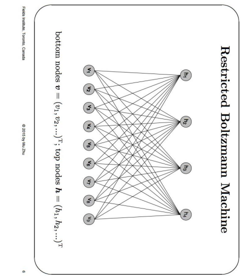
Restricted Boltzmann machine

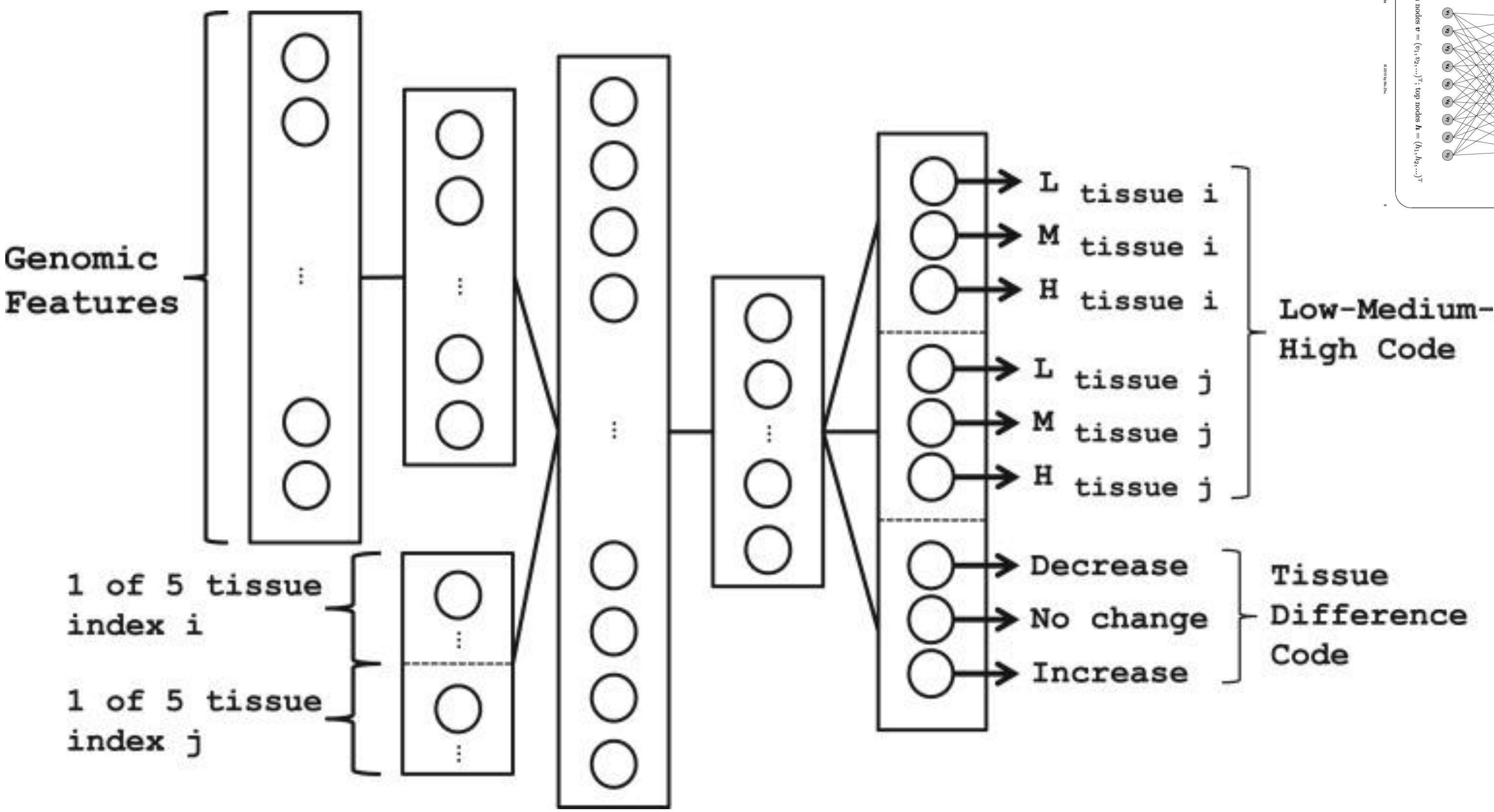


$$f(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}$$

- if just one binary top node, model for $h \mid \underline{v}$ is a logistic regression
- with several binary top nodes, model for $h_t \mid \underline{v}, h_{-t}$ is also a logistic regression, with odds ratio depending only on $\underline{v}$
- deep learning has ~10 layers, with millions of units in each layer
- estimating parameters is an optimization problem

Restricted Boltzmann machine





Brendan Frey, Infinite Genomes Project

Fields*Live* January 27 2015

Leung et al Bioinformatics 2014

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

- Optimization

$$\max_{\theta}\{\frac{1}{n}\sum_{i=1}^{n}\log f(y_i \mid x_i;\theta) - \mathcal{P}_{\lambda}(\theta)\}$$

Optimization

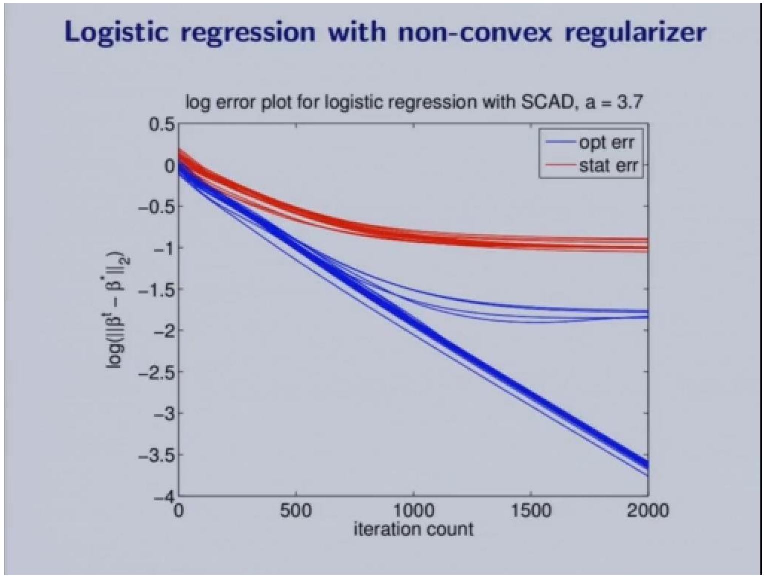
$$\max_{\theta}\{\frac{1}{n}\sum_{i=1}^{n}\log f(y_i \mid x_i;\theta) - \mathcal{P}_\lambda(\theta)\}$$

- lasso penalty $\mathcal{P}_\lambda(\theta) = \lambda||\theta||_1 = \lambda\Sigma|\theta_j|$
- $||\theta||_1$ is convex relaxation of $||\theta||_0$
- many interesting penalties are non-convex
- optimization routines may not find global optimum

Optimization

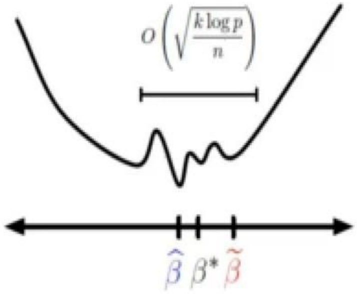
$$\max_{\theta}\{\frac{1}{n}\sum_{i=1}^{n}\log f(y_i \mid x_i;\theta) - \mathcal{P}_\lambda(\theta)\}$$

- statistical error $\hat{\theta} - \theta^*$ neighbourhood of true value
- approximation error $\theta_t - \hat{\theta}$ iterating over *t*



Wainwright Fields*Live* Jan 16 2015

Loh and Wainwright *JMLR* 2015

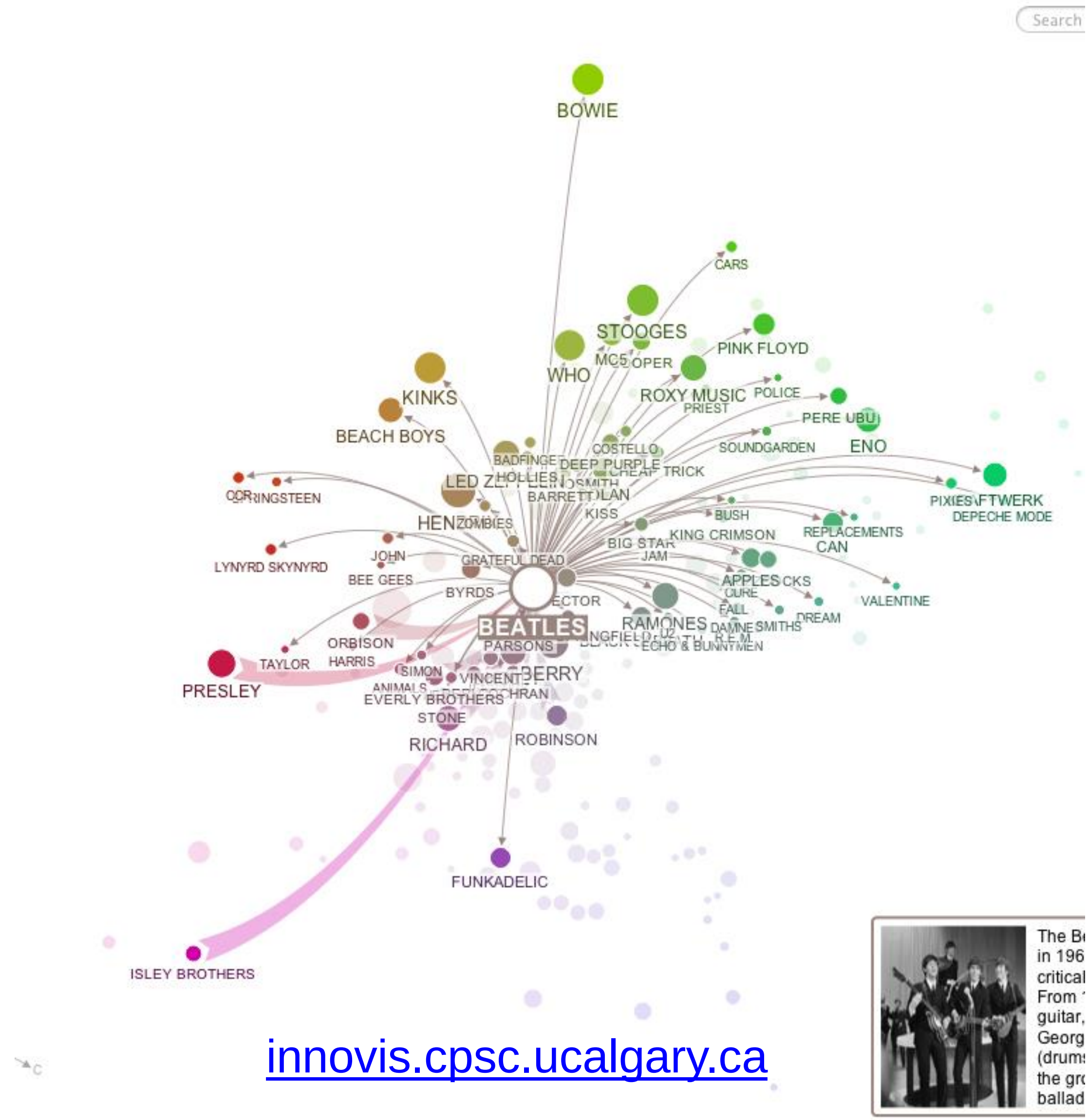


Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

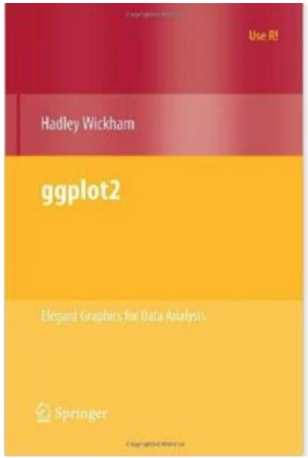
- Visualization

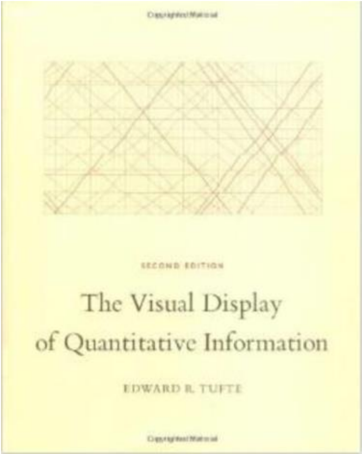


innovis.cpsc.ucalgary.ca

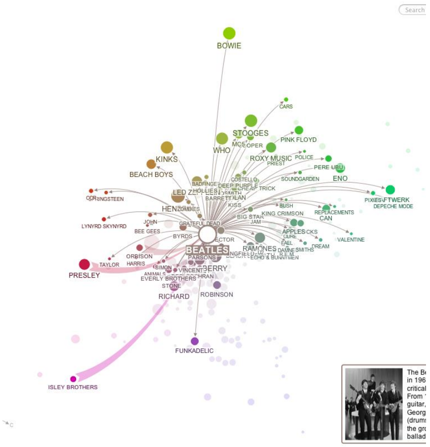
Visualization

- statistical graphics
  - data representation
  - data exploration
  - filtering, sampling aggregation



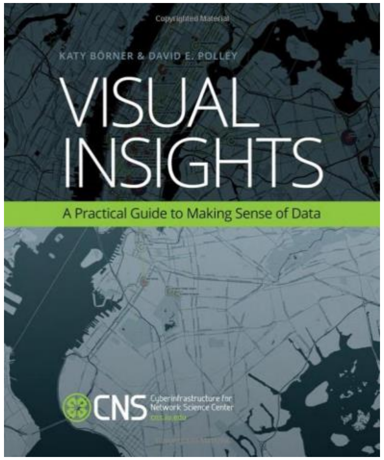


- information visualization



- scientific visualization

- cognitive science and design

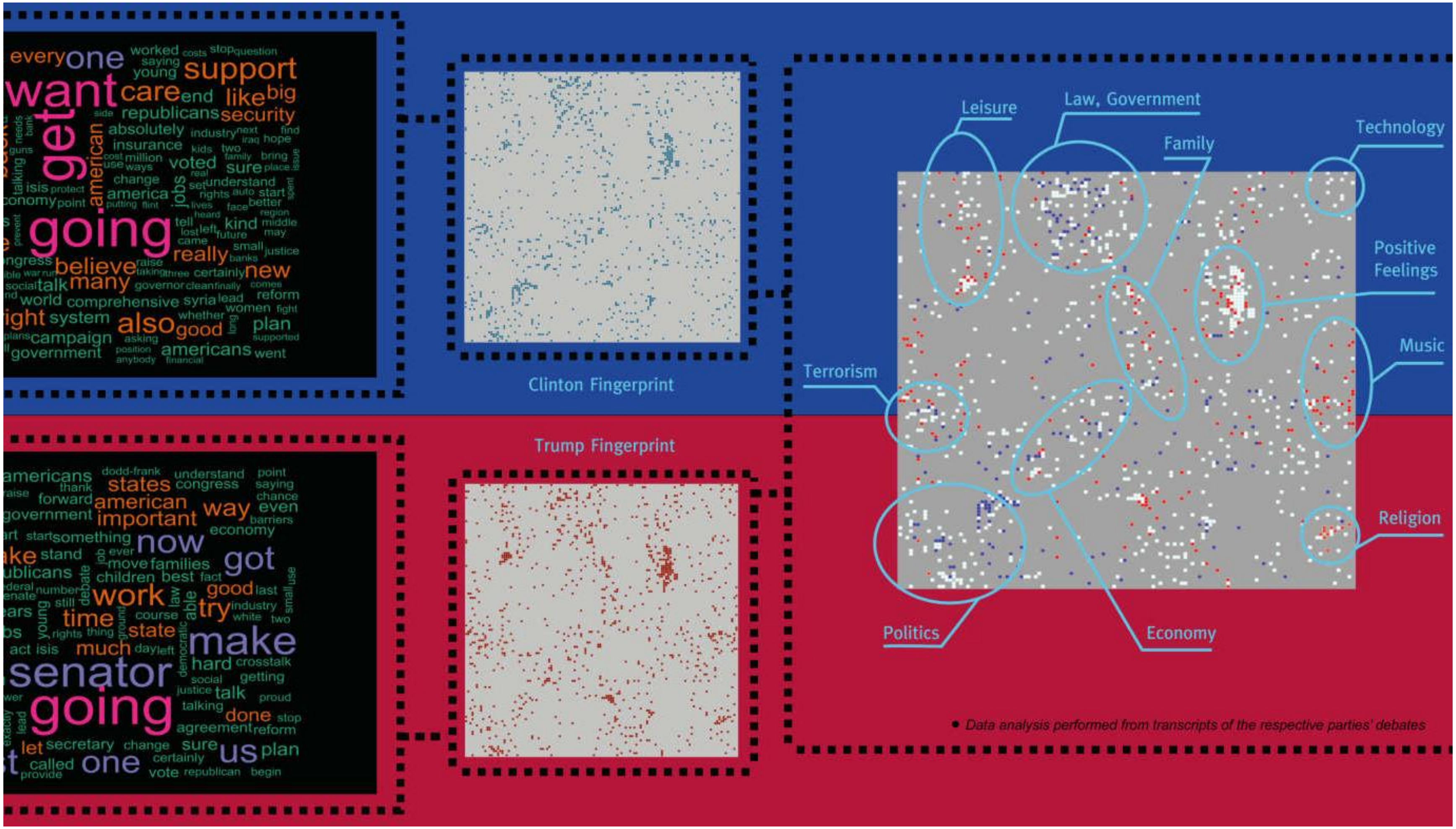


Visualization

[KPMG Data Observatory, IC]

Visualization

[KPMG Data Observatory, IC]

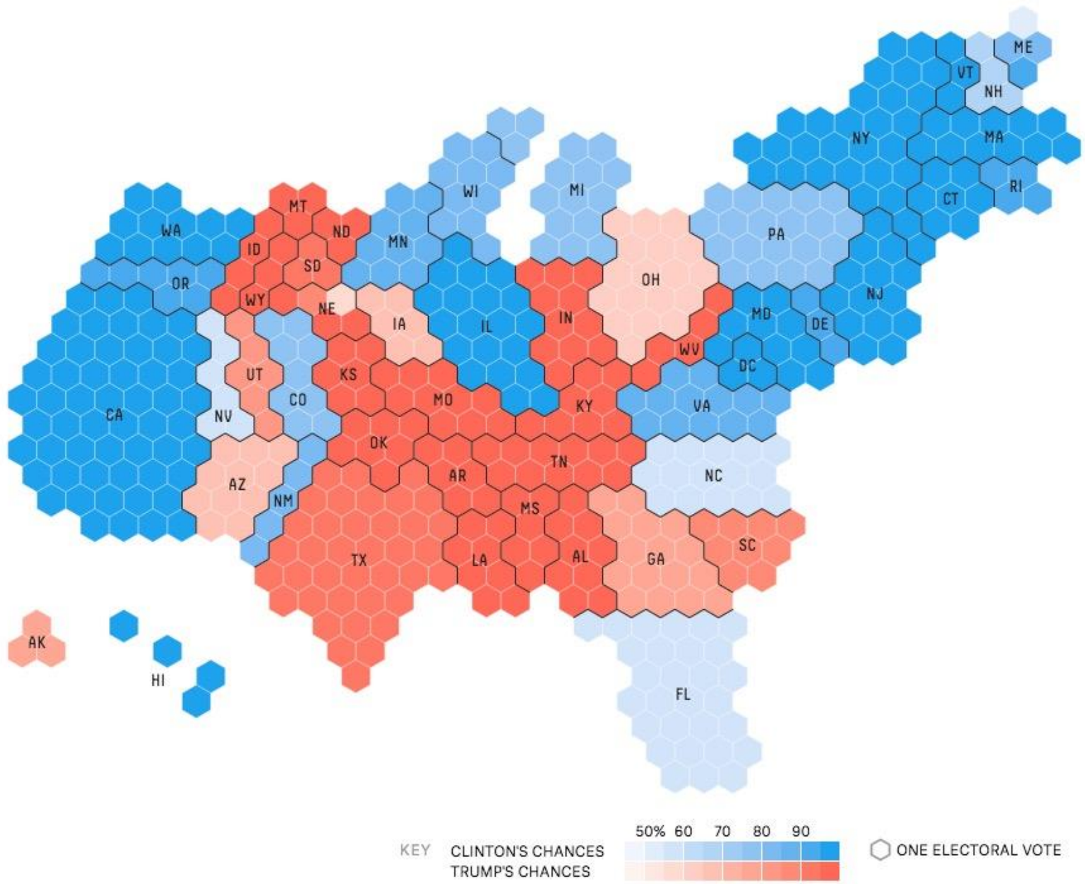


Visualization

fivethirtyeight.com

It's all about the 538 Electoral College votes

Here's a map of the country, with each state sized by its number of electoral votes and shaded by the leading candidate's chance of winning it.

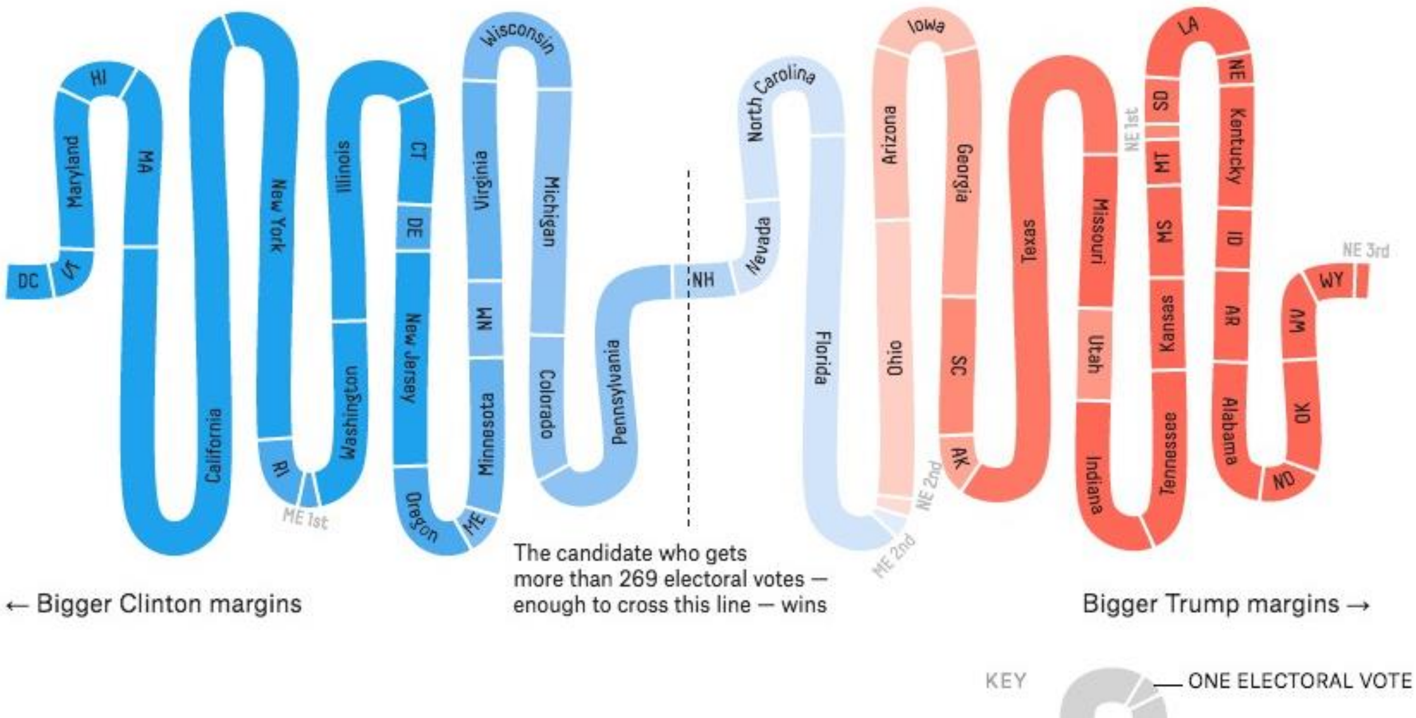


Visualization

fivethirtyeight.com

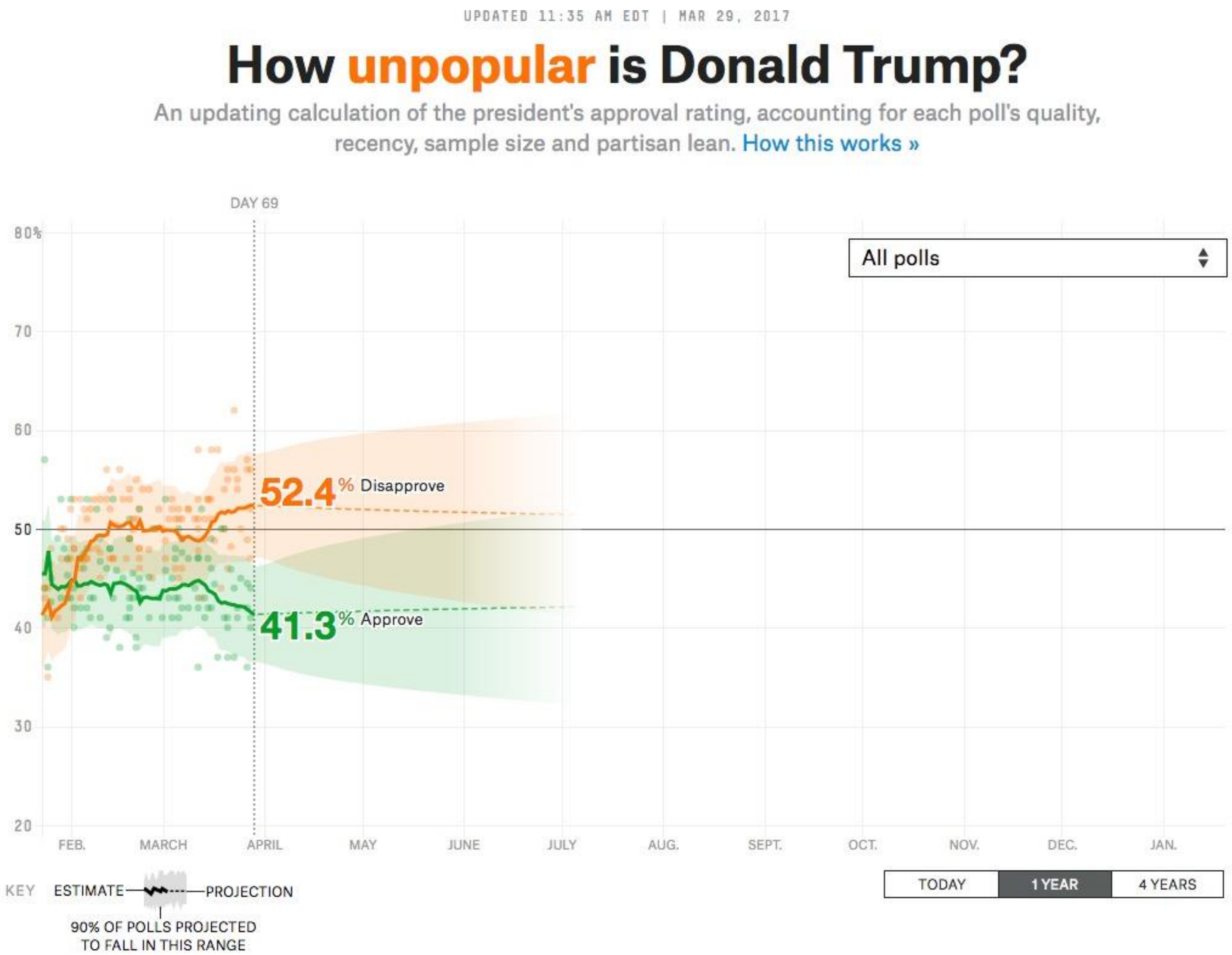
The winding path to 270 electoral votes

A candidate needs at least 270 electoral votes to clinch the White House. Here's where the race stands, with the states ordered by the projected margin between the candidates — Clinton's strongest states are farthest left, Trump's farthest right — and sized by the number of electoral votes they will award.



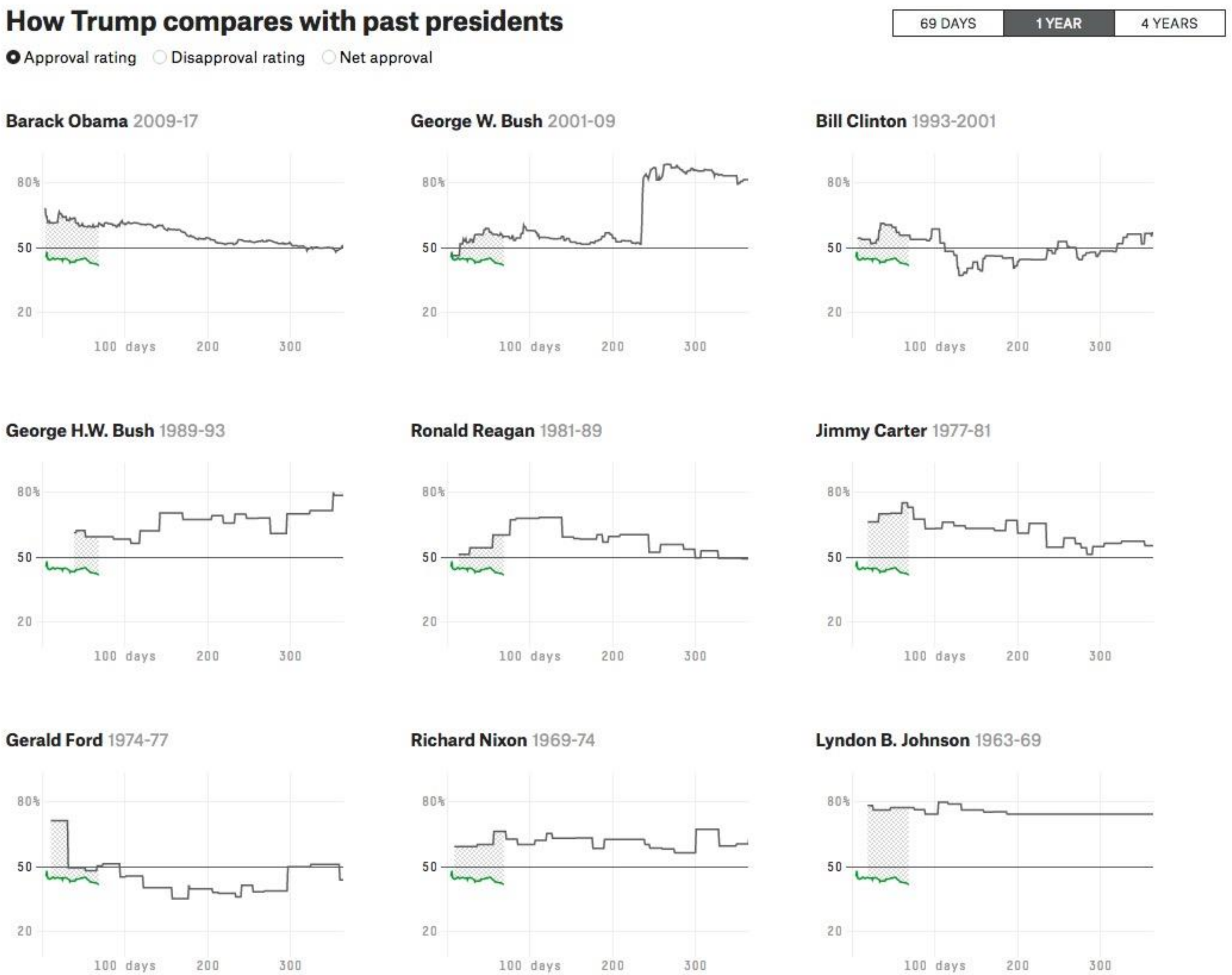
Visualization

fivethirtyeight.com



Visualization

fivethirtyeight.com



guns

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights

- Health Policy

The ICES Data Repository consists of record-level, coded and linkable healt
encompasses much of the publicly funded administrative health services rec
Ontario population eligible for universal health coverage since 1986 and is c
integrating research-specific data, registries and surveys. Currently, the repo
health service records for as many as 13 million people.

The ICES Data Repository consists of record-level, coded and linkable health data sets. It encompasses much of the publicly funded administrative health services records for the Ontario population eligible for universal health coverage since 1986 and is capable of integrating research-specific data, registries and surveys. Currently, the repository includes health service records for as many as 13 million people.

Institute for Clinical and Evaluative Sciences

Health Policy

Administrative Databases

The ICES Data Repository consists of record-level, coded and linkable health data sets. It encompasses much of the publicly funded administrative health services records for the Ontario population eligible for universal health coverage since 1986 and is capable of integrating research-specific data, registries and surveys. Currently, the repository includes health service records for as many as 13 million people.

I WANT TO...

-- Select --

DATA & PRIVACY

- **ICES Data**
  - Data Dictionary
  - Types of ICES Data
  - Working with ICES Data
  - Special Data Projects
- Privacy at ICES
  - Privacy FAQs
  - Questions or Complaints

Institute for Clinical and Evaluative Scienc

ICES Data Repository is globally unique in scope and breadth

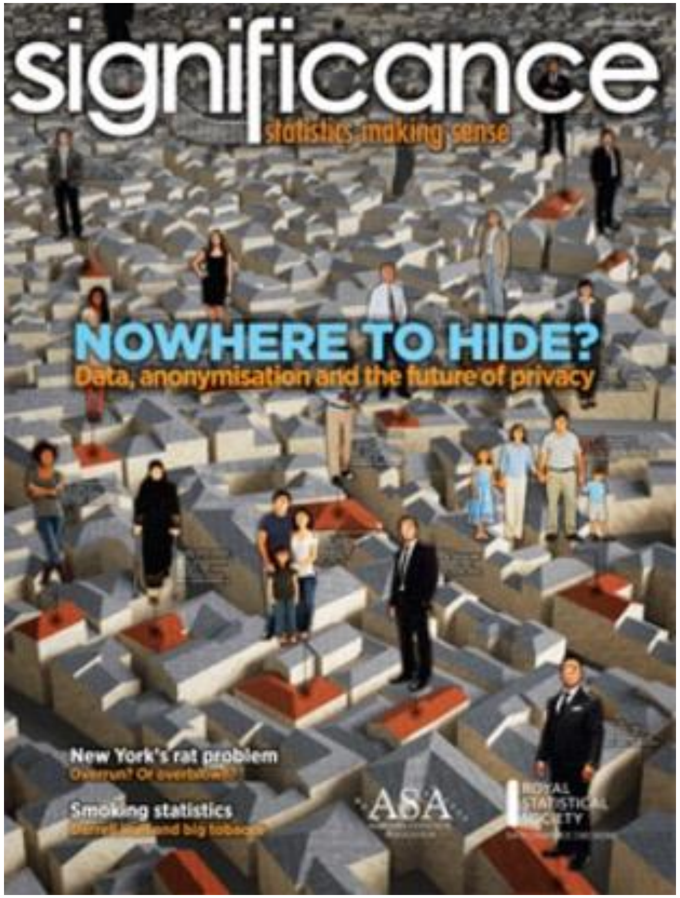
- Individual level: reflects people and their health care experiences
- Linkable: once linked, provide information about continuity of care
- Longitudinal: most health care records over time since 1991
- Population based: health records of 13M people in 2012; 4M Electronic Medical Records profiling 330,000 Ontarians
- Breadth of services: most publicly funded health services, some services outside health
- De-identified: unique ICES Key Number – encrypted health card number
- Secure and Privacy Protected: approved by Office of the Information and Privacy Commissioner

Thérèse Stukel, ICES

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy

Some highlights



- Social Policy

ICES Data Repository is globally unique in scope and breadth

- Individual level: reflects people and their health care experiences
- Linkable: once linked, provide information about continuity of care
- Longitudinal: most health care records over time since 1991
- Population based: health records of 13M people in 2012; 4M Electronic Medical Records profiling 330,000 Ontarians
- Breadth of services: most publicly funded health services, some services outside health
- De-identified: unique ICES Key Number – encrypted health card number
- Secure and Privacy Protected: approved by Office of the Information and Privacy Commissioner

Thérèse Stukel, ICES

Privacy

- “Big Data and Innovation, Setting the Record Straight: De-identification *Does* Work”
  [Privacy Commissioner of Ontario, July 2014]

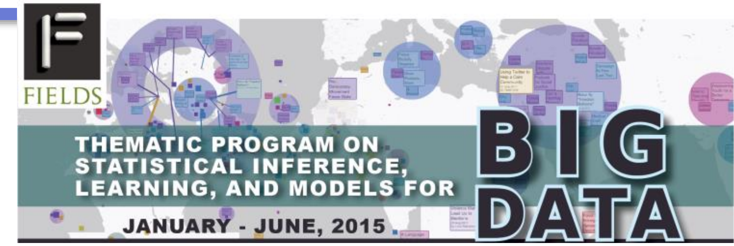
- “No silver bullet: De-identification still doesn’t work”
  [Narayan & Felten, July 2014]

- Statistical Disclosure Limitation
- Differential Privacy
- Multi-party Communication

Some highlights

- Statistical Machine Learning
- Optimization
- Visualization
- Health Policy
- Social Policy
- inference, environmental science, networks, genomics, finance, physical sciences, software infrastructure, …

What did we learn?



- Statistical models for big data are complex, high-dimensional
  - inference is well-studied, but difficult
- Computational challenges include size and speed
  - ideas of statistical inference get lost in the machine
- Data owners understand 2., but not 1.
- Data modellers understand 1., but not 2.
- Data science may be the best way to combine these

That was yesterday

- Data science programs “springing up like mushrooms after rain”

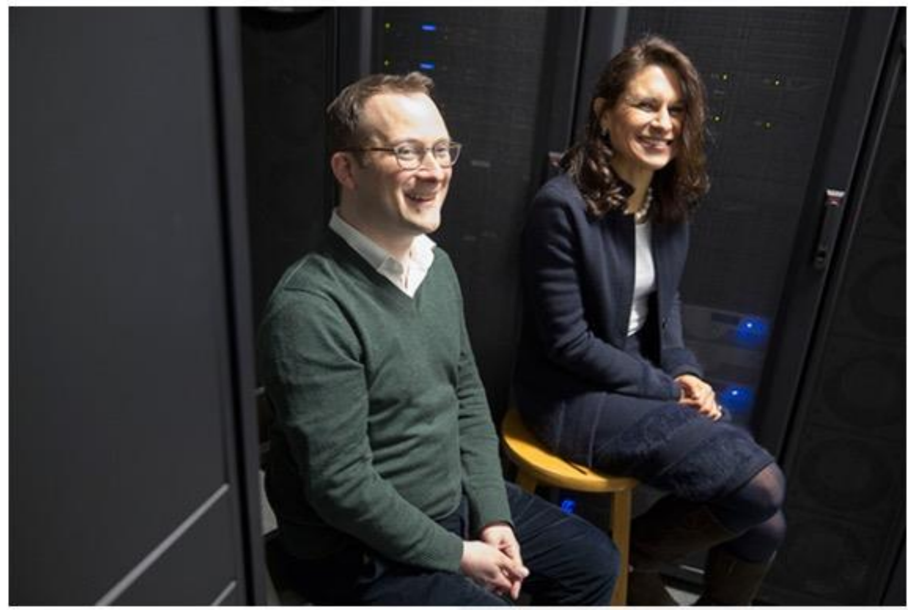
HARVARDgazette

SCIENCE & HEALTH > ENGINEERING & TECHNOLOGY

Harvard launches data science initiative

Francesca Dominici and David Parkes named co-directors

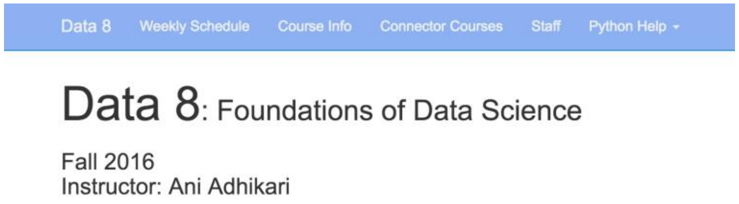
March 28, 2017 |



- Berkeley, Hopkins, CMU, Washington, UBC, Toronto, …

What is data science?

- a course?
- a set of courses?
- a job?
- a technology?
- a new field of research?
- a collaboration?



Data Science Program(s)

University of Toronto
New Undergraduate Program Proposal

'This template has been developed in line with the University of Toronto's Quality Assurance Process.)

- mathematical reasoning
- statistical theory
- statistical and machine learning methods
- programming and software development
- algorithms and data structure
- communication results and limitations

LOS | SUBMISSION

http://arxiv.org/abs/1609.00037v1

Good Enough Practices in Scientific Computing

Greg Wilson[1,‡*], Jennifer Bryan[2,‡], Karen Cranston[3,‡], Justin Kitzes[4,‡], Lex Nederbragt[5,‡], Tracy K. Teal[6,‡]
1 Software Carpentry Foundation / gvwilson@software-carpentry.org
2 University of British Columbia / jenny@stat.ubc.ca
3 Duke University / karen.cranston@duke.edu
4 University of California, Berkeley / jkitzes@berkeley.edu
5 University of Oslo / lex.nederbragt@ibv.uio.no
6 Data Carpentry / tkteal@datacarpentry.org

‡ These authors contributed equally to this work.
* E-mail: Corresponding gvwilson@software-carpentry.org

… Good Enough

- Data Management – from raw to ‘analysable’
- Software – programming
- Collaboration
- Project Organization
- Keeping Track
- Writing

knitr

tidyr

dplyr

ggplot2

ggvis

Github

Data Science Research

- data collection and data quality

- large N, small p
  - computational strategies, e.g. Spark, Hadoop
  - divide and conquer

- small n, large p
  - inferential and computational strategies
  - dimension reduction
  - post-selection inference
  - inference for extremes

- ‘new’ types of data: networks, graphs, text, images, …
  - “alternative sources”

… Data Science Research

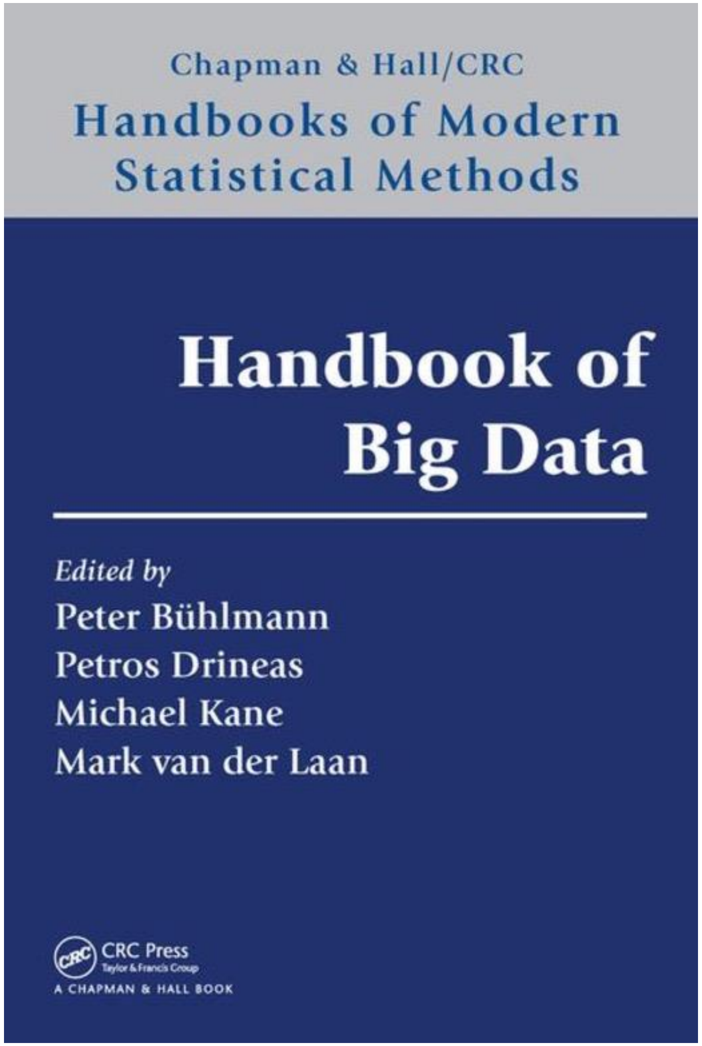
- collaboration and communication
- data wrangling, database development, record linkage, privacy
- replicability, reproducibility, new workflows
- visualization
- outside the ivory tower -- industry, government, media, public

Tripods (Transdisc Research in Princ…)

Fundamental research areas that may be a part of the focus of a transdisciplinary collaboration under this solicitation include, but are not limited to:

- Combinatorial inference on complex structures;
- Tradeoffs between computational costs and statistical efficiency;
- Randomized numerical linear algebra;
- Representation theory and non-commutative harmonic analysis;
- Topological data analysis (TDA) and homological algebra; and
- Multiple areas in machine learning including deep learning.

Published Feb 2016



I. General Perspectives
I. Data-Centric, Exploratory Methods
I. Efficient Algorithms
II. Graph Approaches
III. Model Fitting and Regularization
IV. Ensemble Methods
V. Causal Inference
VI. Targeted Learning

The push back

Big data
The Guardian's Science Weekly

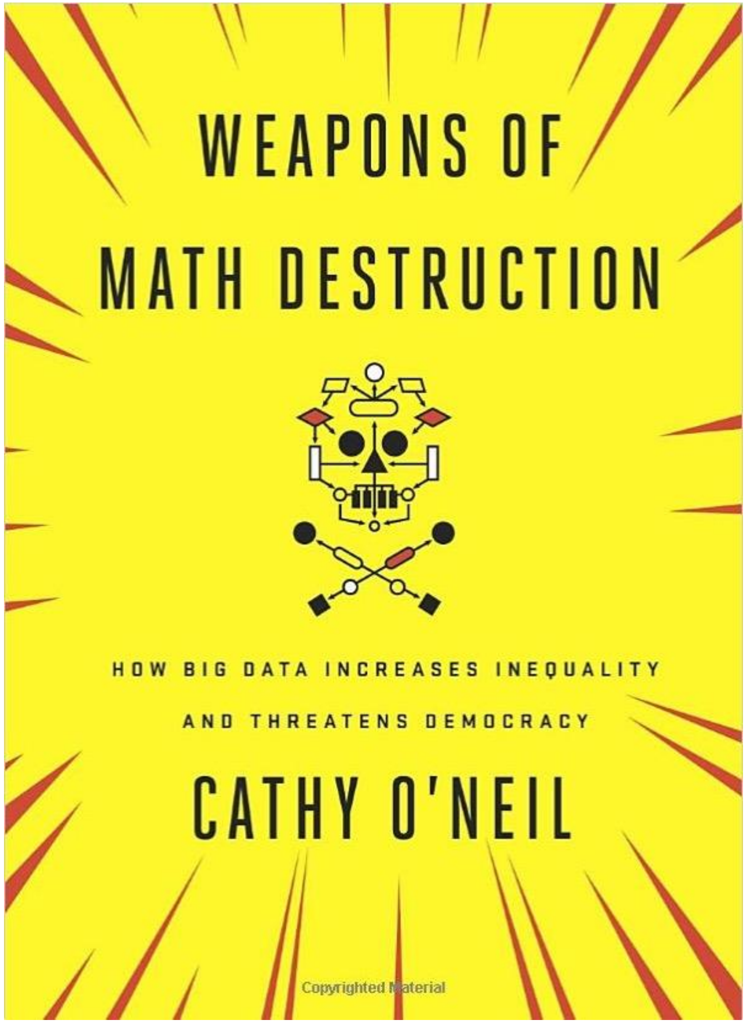
Weapons of math destruction: how big data and algorithms affect our lives - podcast

WS More or Less: Algorithms, Crime and Punishment

When maths can get you locked up.

Download MP3

Available now
9 minutes



"if the assessment never asks about race, how could the algorithm throw up racially biased results?"

"Credit scores are used by nearly half of American employers to screen potential employees"

How big data threatens democracy and increases inequality

The push back

Big data in social sciences: a promise betrayed ?

Posted on March 22, 2017

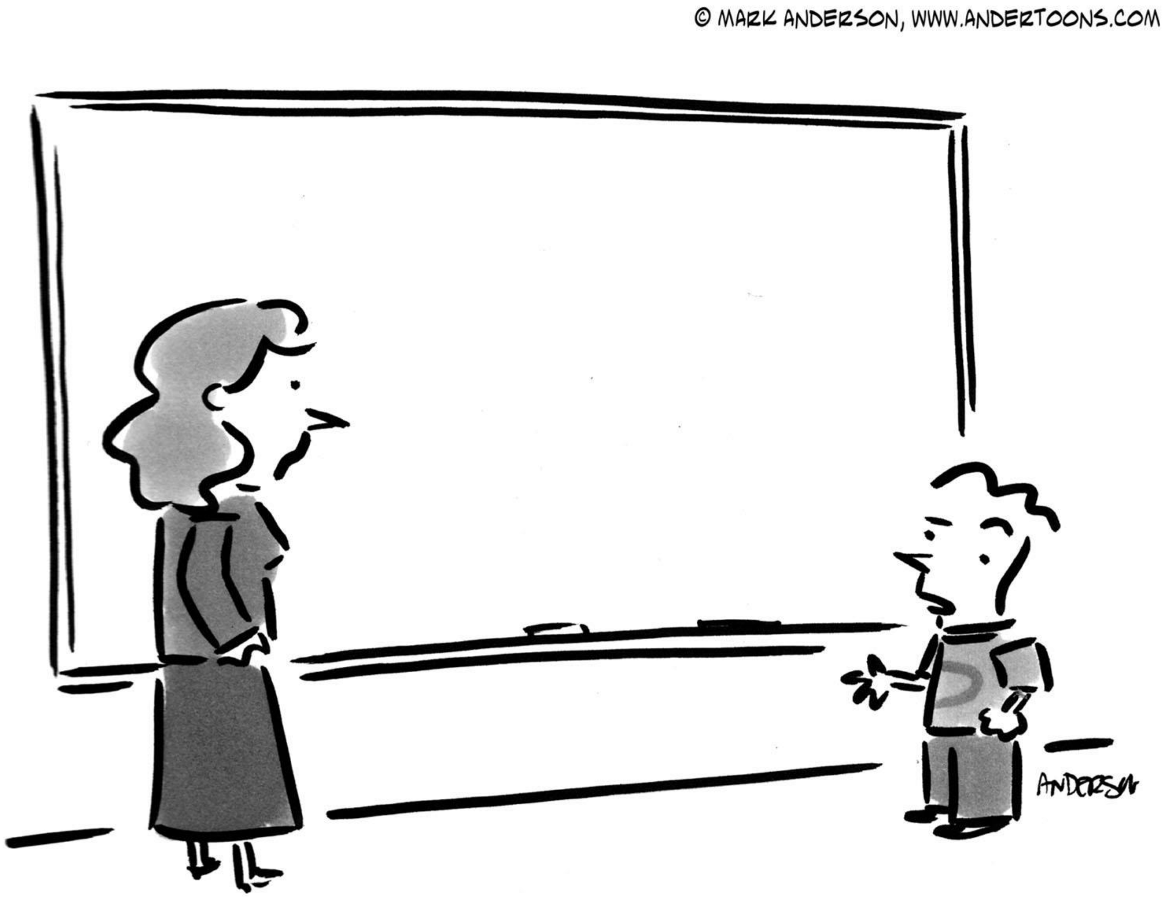
In just 5 years, the mood at conferences on social science and big data has shifted, at least in France. Back in the early 2010s, these venues were buzzing with exchanges about the characteristics of the “revolution” (the 4Vs) with participants marveling at the research insights afforded by the use of tweets, website ratings, Facebook likes, Ebay prices or

“Big data is neither easier nor faster nor cheaper”

“Building a database doesn’t create its own uses”

My impression was that there is a sense in which ML is to statistics what robotization is to society: a job threat demanding a compelling reexamination of what is left for human statisticians to do,

Privacy



"Before I write my name on the board, I'll need to know how you're planning to use that data."

Privacy


sign in | become a supporter | subscribe | search

jobs | dating | more | International edition

theguardian

UK | world | sport | football | opinion | culture | business | lifestyle | fashion | environment | tech | travel | all sections

home

Public Leaders Network
Global public leaders

How do you feel about the government sharing our personal data? - livechat

[March 29](#)

Facial recognition database used by FBI is out of control, House committee hears

Database contains photos of half of US adults without consent, and algorithm is wrong nearly 15% of time and is more likely to misidentify black people

[March 27](#)

The push back

The push back

Big data: are we making a big mistake?

Economist, journalist and broadcaster **Tim Harford** delivered the 2014 *Significance* lecture at the Royal Statistical Society International Conference. In this article, republished from the *Financial Times*, Harford warns us not to forget the statistical

"Big data" has arrived, but big insights have not

“A range of other problems”

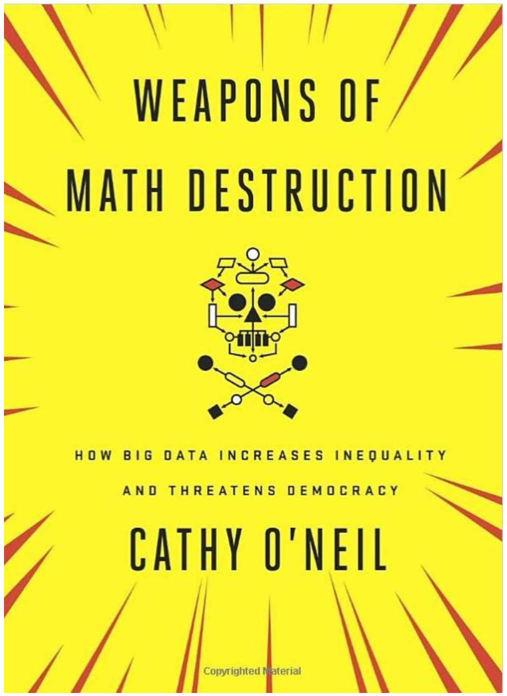
“while I do think of neural networks as one important tool in the toolbox, I find myself surprisingly rarely going to that tool when I’m consulting out in industry.

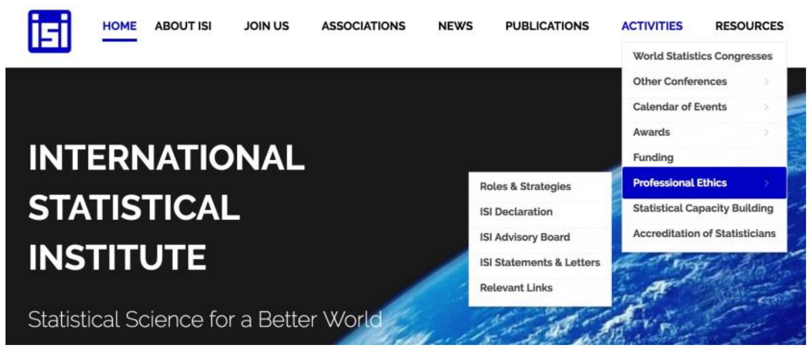
I find that industry people are often looking to solve a range of other problems, often not involving “pattern recognition” problems”

Michael Jordan, UC Berkeley

accurate answers quickly; meaningful error bars; merge various data sources; visualize and present conclusions; diagnostics; non-stationarity; targetted experiments within databases

Caution can be a good thing





"Digital Hippocratic Oath"

Caution can be a good thing

Guardian 2 July 2016

"…from data we will get the cure for cancer as well as better hospitals;

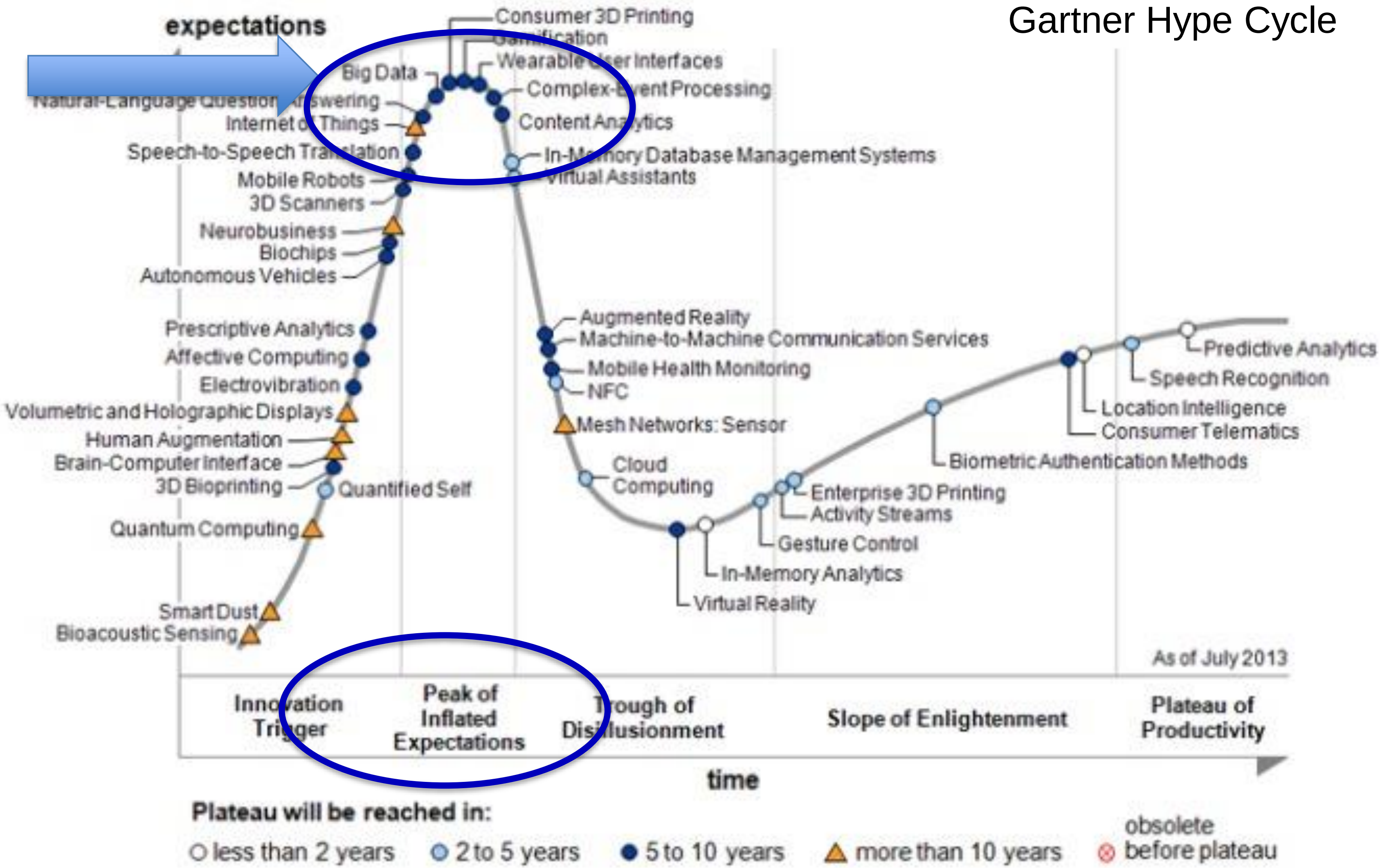
schools that adapt to children's needs making them happier and smarter;

better policing and safer homes;

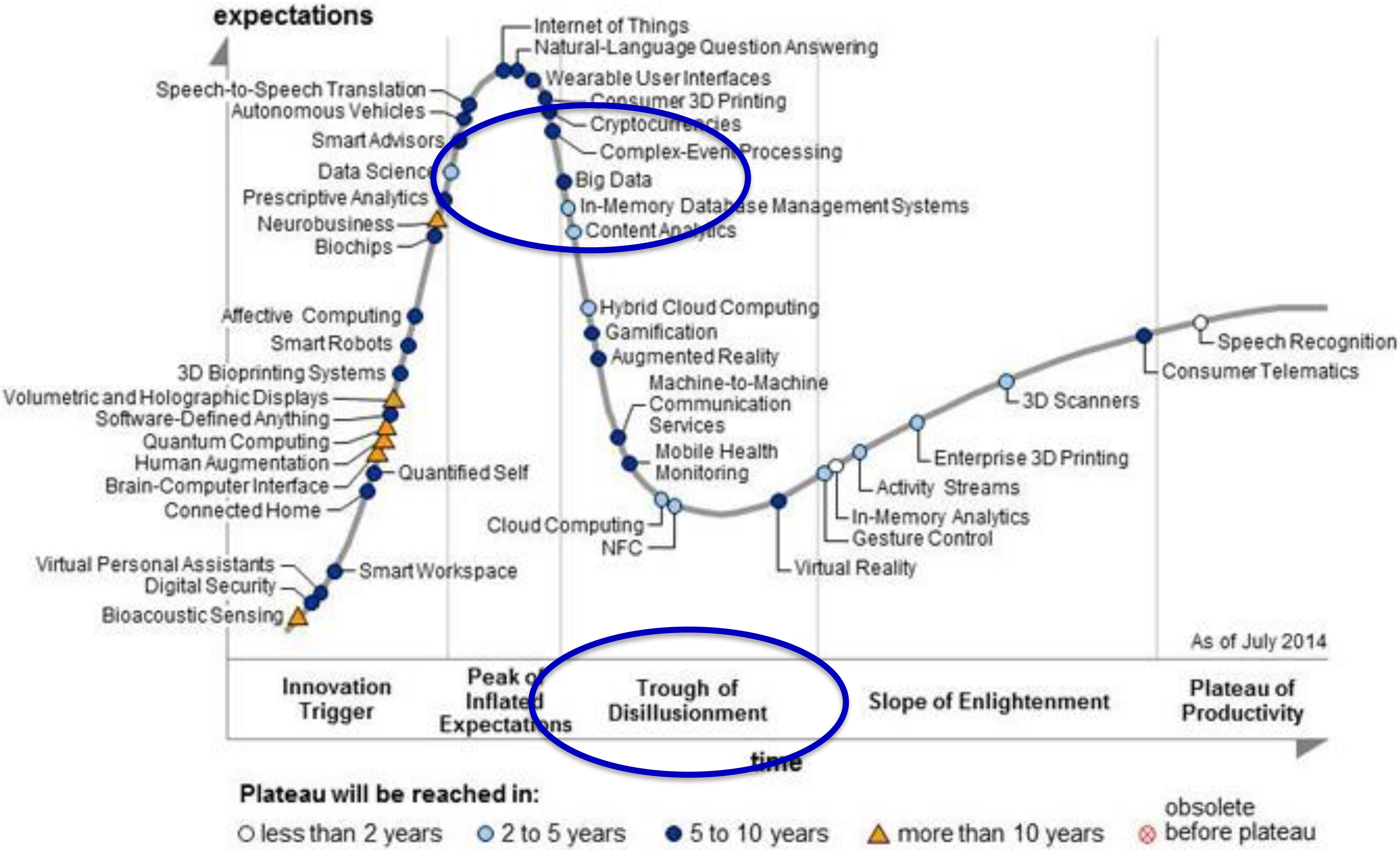
and of course jobs."

Big Data 2013

Gartner Hype Cycle

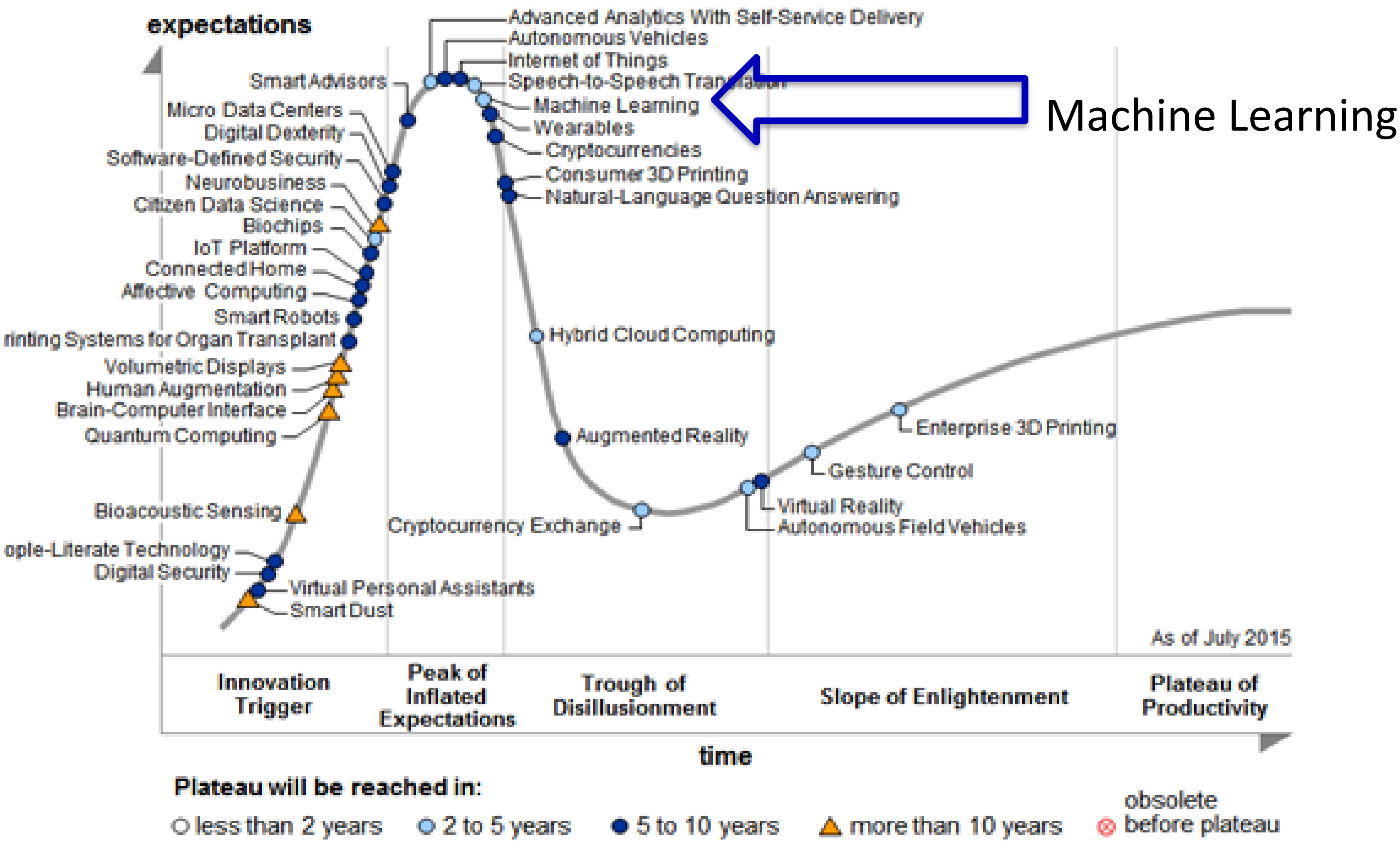


Big Data 2014

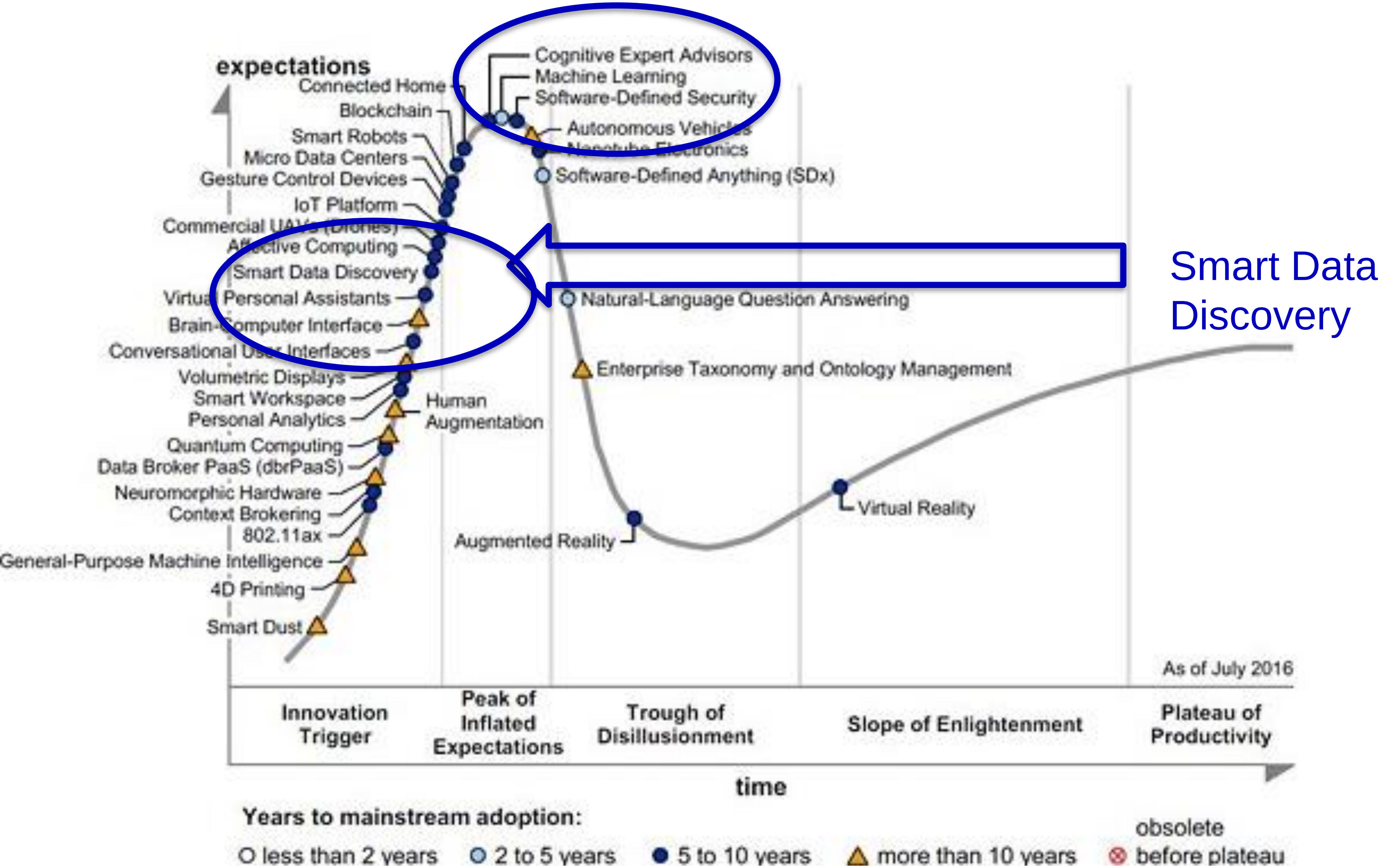


Big Data

2015



Gartner Hype Cycle 2016



Thank You!

Queen's University

Department of Mathematics and Statistics