Statistical Inference, Learning and Models for Big Data

Nancy Reid

University of Toronto

December 2, 2015

UC San Diego
School of Medicine

Division of Biostatistics & Bioinformatics
In the Department of Family Medicine and Public Health

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
Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander , Nancy Reid, Martin Wainwright

FEBRUARY 23 – 27, 2015
Workshop on Visualization for Big Data: Strategies and Principles
Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson

MARCH 23 – 27, 2015
Workshop on Big Data in Health Policy
Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis , Sharon-Lise Normand

APRIL 13 – 17, 2015
Workshop on Big Data for Social Policy
Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson

JUNE 13 – 14, 2015
Closing Conference
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University

GRADUATE COURSES

JANUARY TO APRIL 2015
Large Scale Machine Learning
Instructor: Ruslan Salakhutdinov (University of Toronto)

JANUARY TO APRIL 2015
Topics in Inference for Big Data
Instructors: Nancy Reid (University of Toronto), Mu Zhu (University of Waterloo)

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 sciences. It is expected that all activities will be webcast using the FieldsLive system to permit wide participation. Allied activities planned include workshops at PIMS in April and May and CRM in May and August.

ORGANIZING COMMITTEE

Yoshua Bengio (Montréal)
Hugh Chipman (Acadia)
Sallie Keller (Virginia Tech)
Lisa Lix (Manitoba)
Richard Lockhart (Simon Fraser)
Nancy Reid (Toronto)
Ruslan Salakhutdinov (Toronto)

INTERNATIONAL ADVISORY COMMITTEE

Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehelata Huzurbazar (Wyoming)
Nicolai Meinshausen (ETH Zurich)
Dale Schuurmans (Alberta)
Robert Tibshirani (Stanford)
Bin Yu (UC Berkeley)

For more information, allied activities off-site, and registration, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/bigdata

Image Credits: Sheelagh Carpendale & InnoVis

NSERC CRSNG

Ontario

CANSSI INCASS

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

Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander , Nancy Reid, Martin Wainwright

FEBRUARY 23 – 27, 2015

Workshop on Visualization for Big Data: Strategies and Principles

Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson

MARCH 23 – 27, 2015

Workshop on Big Data in Health Policy

Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis , Sharon-Lise Normand

APRIL 13 – 17, 2015

Workshop on Big Data for Social Policy

Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson

JUNE 13 – 14, 2015

Closing Conference

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University

Canadian Institute for Statistical Sciences

Fields Institute for Resesarch in the Mathematical Sciences

Pacific Institute for Mathematical Sciences

Centre de Recherches Mathématiques

Workshops

Opening Conference and Bootcamp	Jan 9 – 23
Statistical Machine Learning	Jan 26 – 30
Optimization and Matrix Methods	Feb 9 – 11
Visualization: Strategies and Principles	Feb 23 – 27
Big Data in Health Policy	Mar 23 – 27
Big Data for Social Policy	Apr 13 – 16

Networks, Web mining, and Cyber-security	May, CRM
Statistical Theory for Large-scale Data	April, PIMS
Challenges in Environmental Science	May, PIMS
Complex Spatio-temporal Data	April, Fields
Commercial and Retail Banking	May, Fields

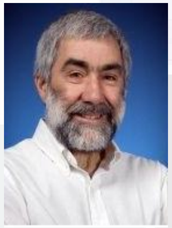
Closing Conference: Statistical and Computational Analytics
June 12 – 13, Halifax

Deep Learning Summer School	August 3 – 12

And more

Distinguished Lecture Series in Statistics

- Terry Speed, ANU, April 9 and 10
- Bin Yu, UC Berkeley, April 22 and 23




Coxeter Lecture Series

- Michael Jordan, UC Berkeley, April 7 – 9

Distinguished Public Lecture,

- Andrew Lo, MIT, March 25


Graduate Courses

- Statistical Machine Learning
- Topics in Big Data

Ruslan Salakhutdinov, Toronto


Mu Zhu, Waterloo

Industrial Problem Solving Workshop

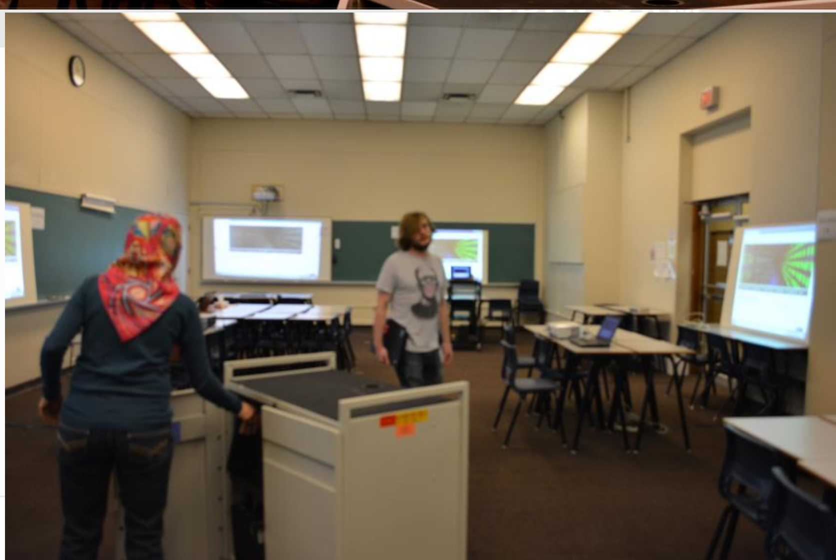
- May 25 – 29

Fields Summer Undergraduate Research Program

- May to August, 2015

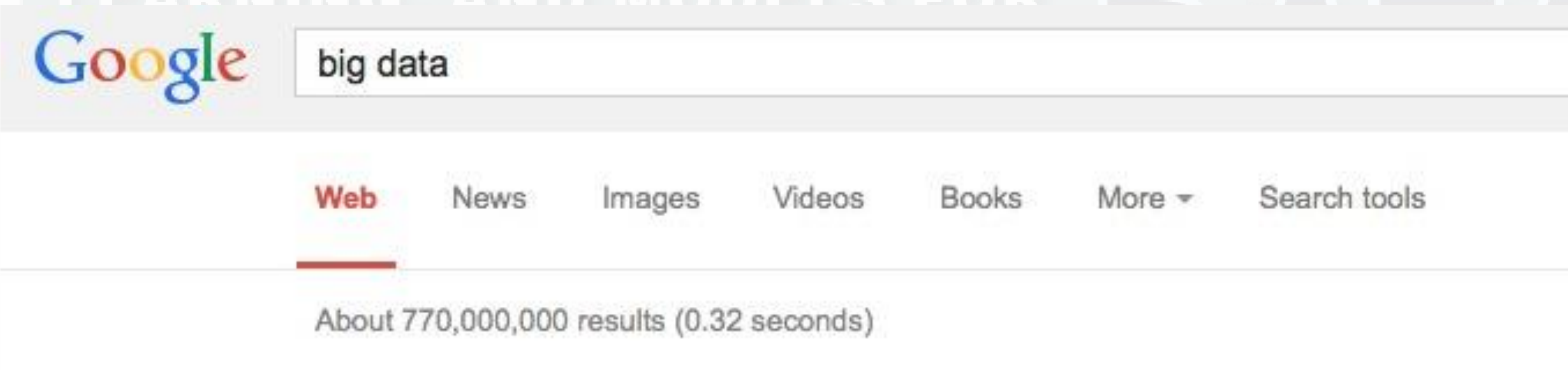
Watch events on Fields*Live*

MDM 12 – Einat Gil et al.



Big Data – Big Topic

- Where to start?
- Look up some references



- Likelihood 78 m
- Statistical inference 7m

MARCH 29, 2013

STEAMROLLED BY BIG DATA

BY GARY MARCUS

Five years ago, few people had heard the phrase "Big Data." Now, it's hard to go an hour without seeing it. In the past several months, the industry has been mentioned in dozens of New York *Times* stories, in every section from metro to business. (*Wired* has even already declared it passé: "STOP HYPING BIG DATA AND START PAYING ATTENTION TO 'LONG DATA'.") At least one corporation, the business-analytics firm SAS, has a Vice-President of Big Data. Meanwhile, nobody seems quite sure exactly what the phrase

Gartner Hype Cycle
July 2013

The Blogosphere

I view "Big Data" as just the latest manifestation of a cycle that has been rolling along for quite a long time

Steve Marron, June 2013

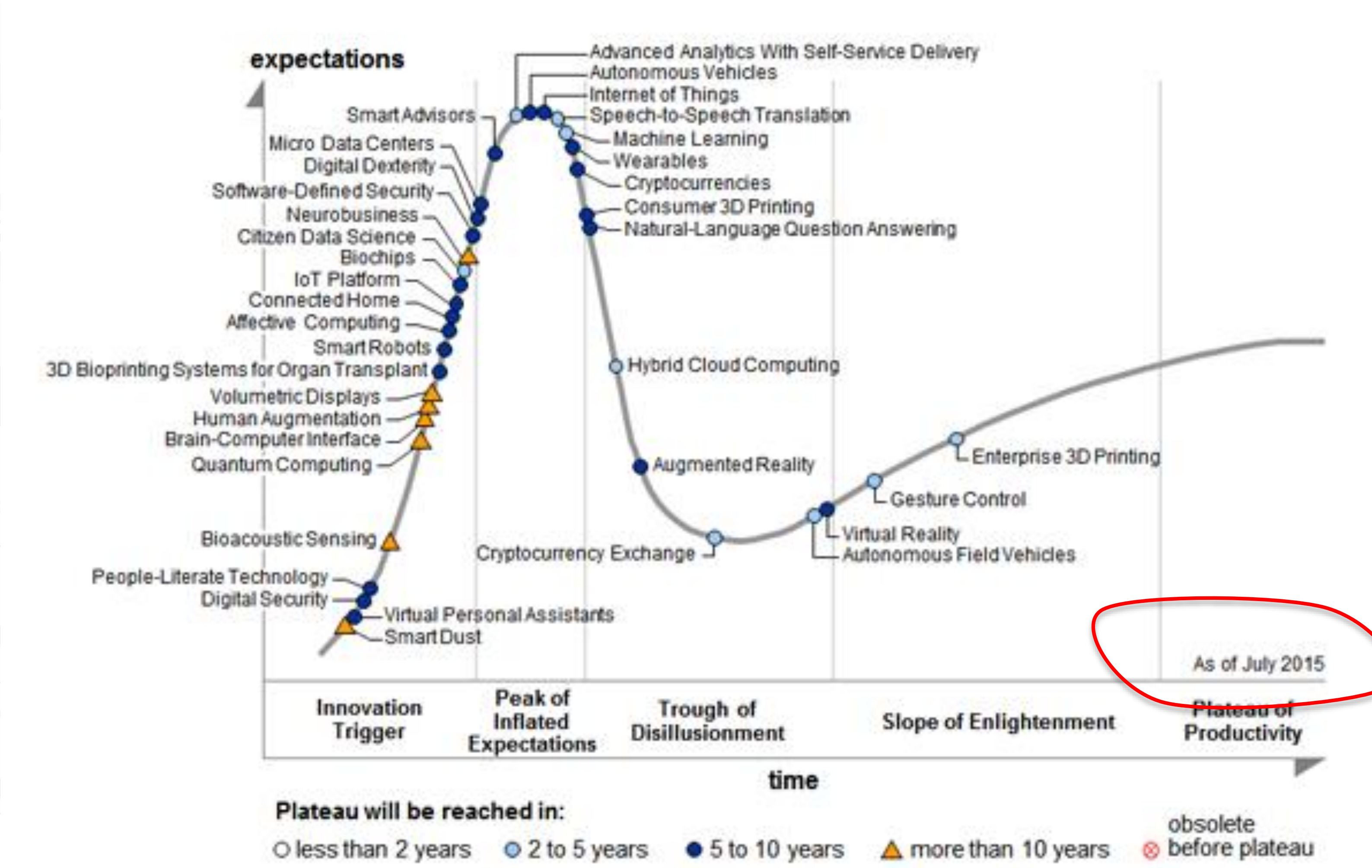
- Statistical Pattern Recognition
- Artificial Intelligence
- Neural Nets
- Data Mining
- Machine Learning

As each new field matured, there came a recognition that in fact much was to be gained by studying connections to statistics

Gartner Hype Cycle
July 2014

falling-99183_640 →

https://etechlib.wordpress.com/tag/hype-cycle/



Big Data Types

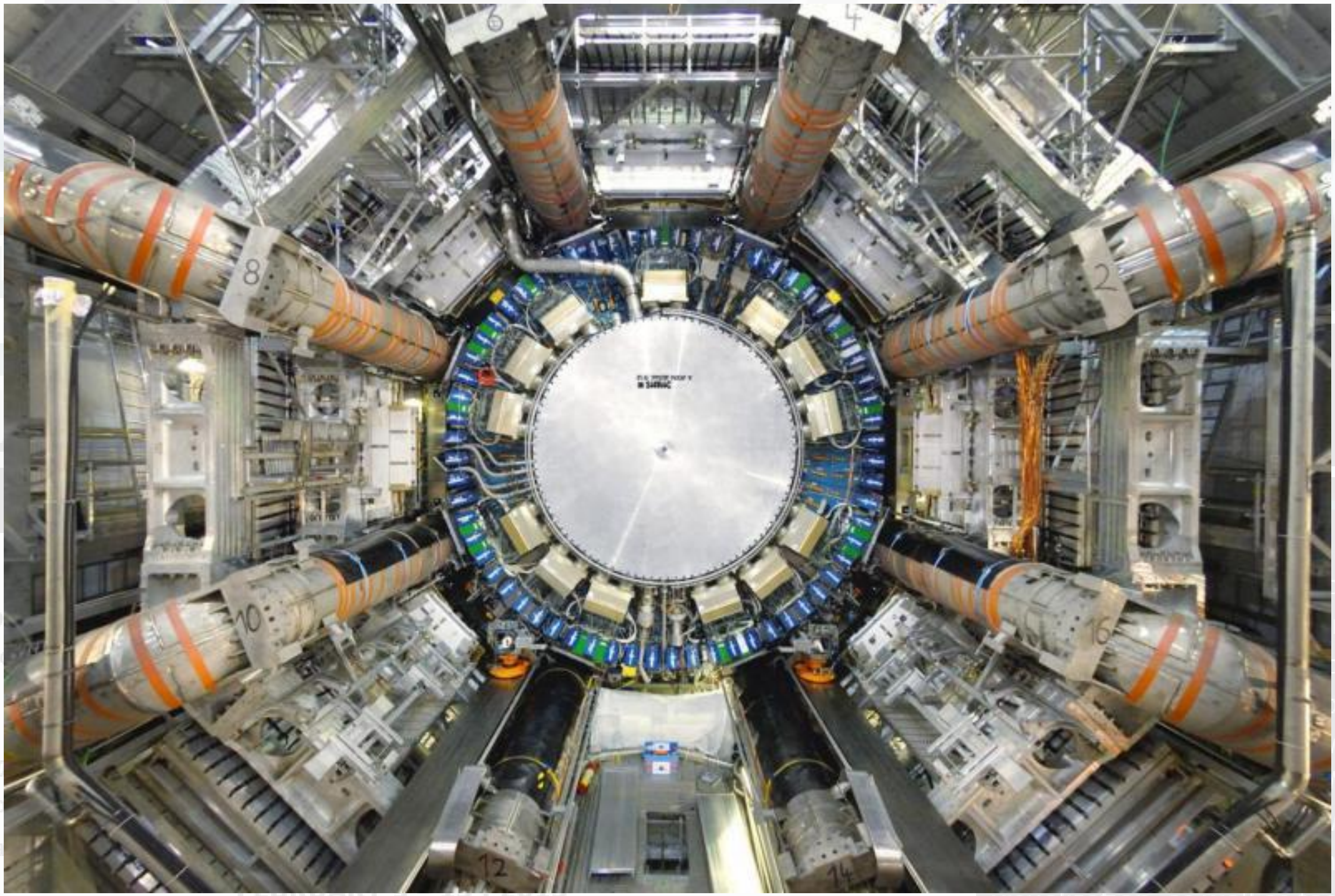
- Data to confirm scientific hypotheses
- Data to explore new science
- Data generated by social activity – shopping, driving, phoning, watching TV, browsing, banking, …

- Data generated by sensor networks – smart cities
- Financial transaction data
- Government data – surveys, tax records, welfare rolls, …
- Public health data – health records, clinical trials, public health surveys

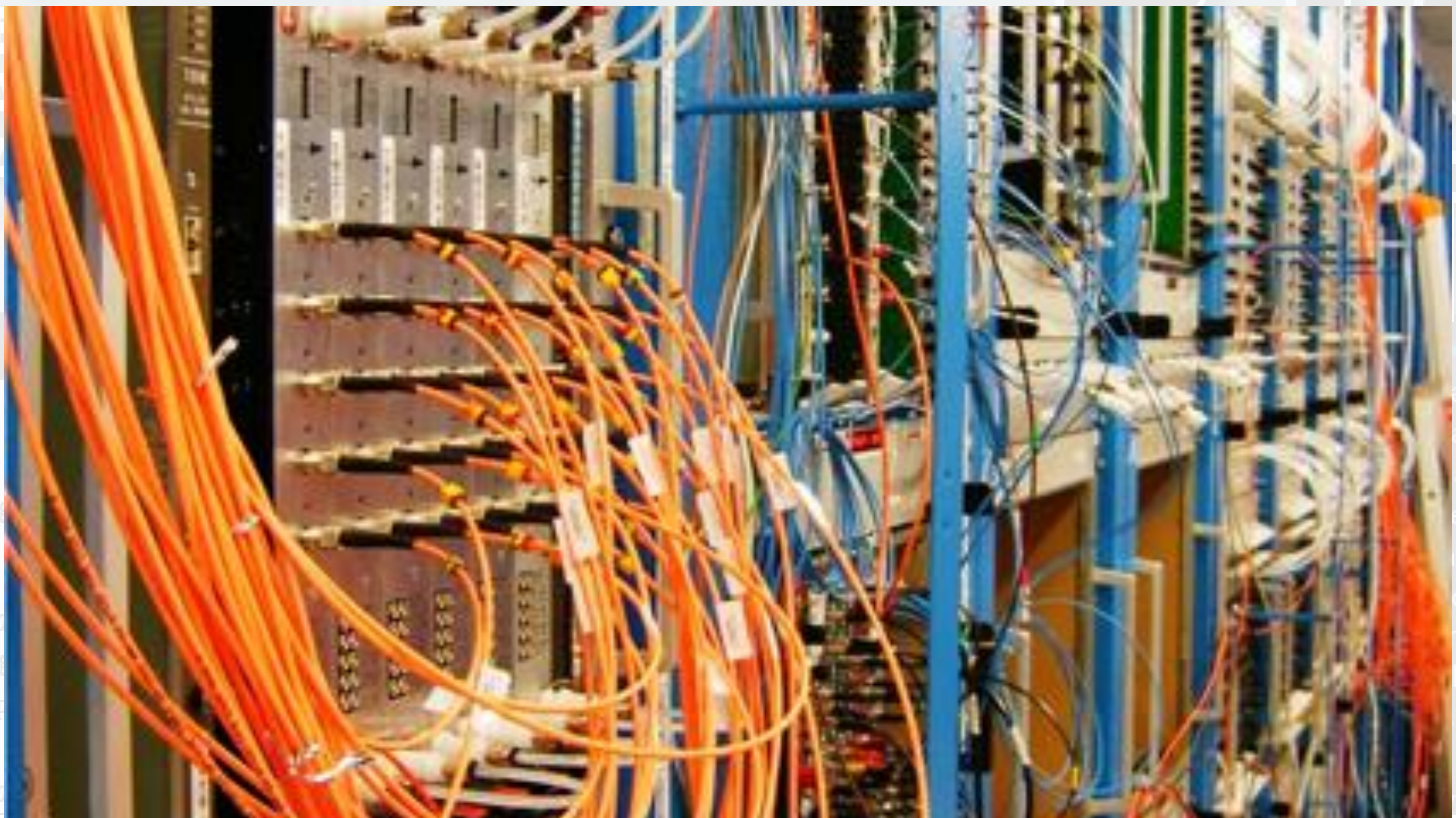
Jordan 06/2014

The Atlas experiment – CERN

http://atlas.ch/what_is_atlas.html#5



If all the data from ATLAS were recorded, this would fill 100,000 CDs per second. This would create a stack of CDs 450 feet high every second, which would reach to the moon and back twice each year. The data rate is also equivalent to 50 billion telephone calls at the same time. ATLAS actually only records a fraction of the data (those that may show signs of new physics) and that rate is equivalent to 27 CDs per minute. http://atlas.ch/what_is_atlas.html - 5

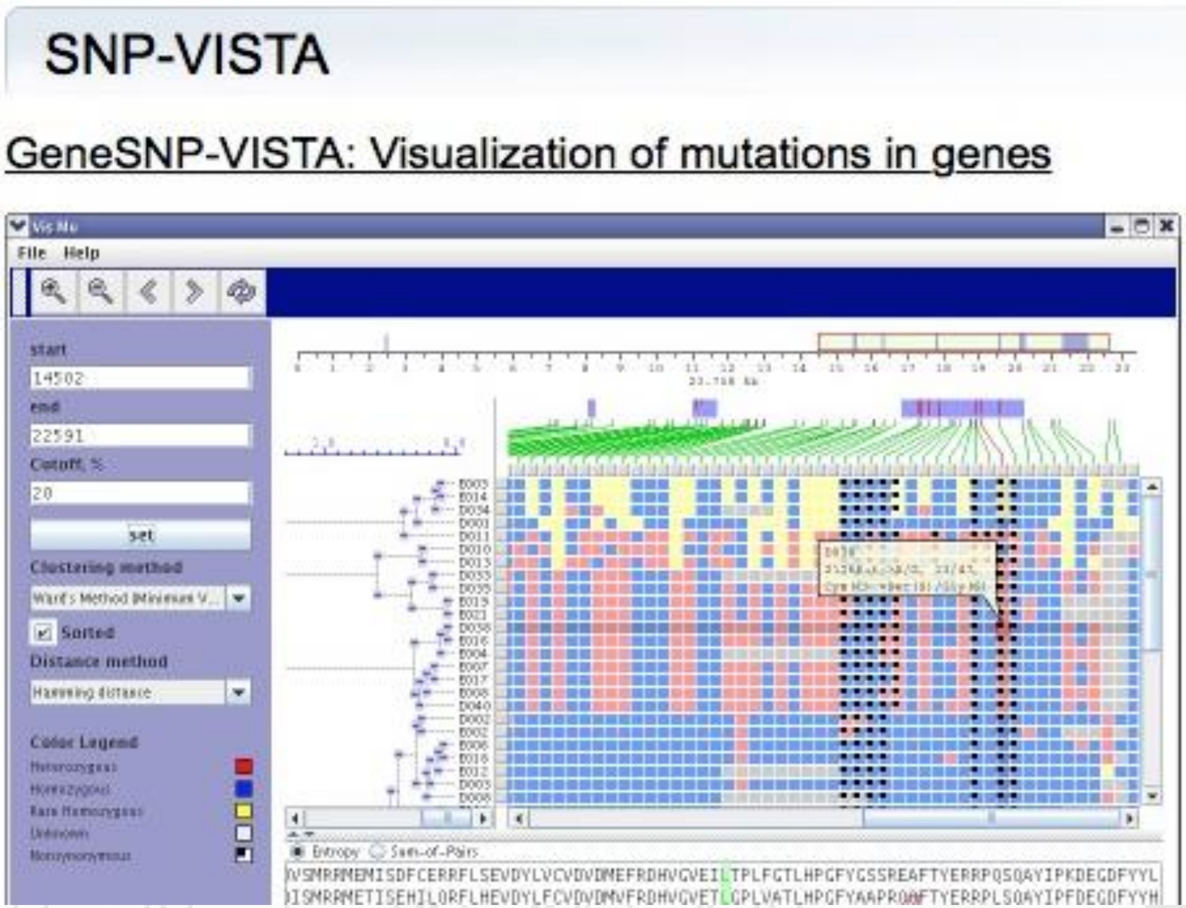


Exploration: the Square Km Array

https://www.skatelescope.org/location/

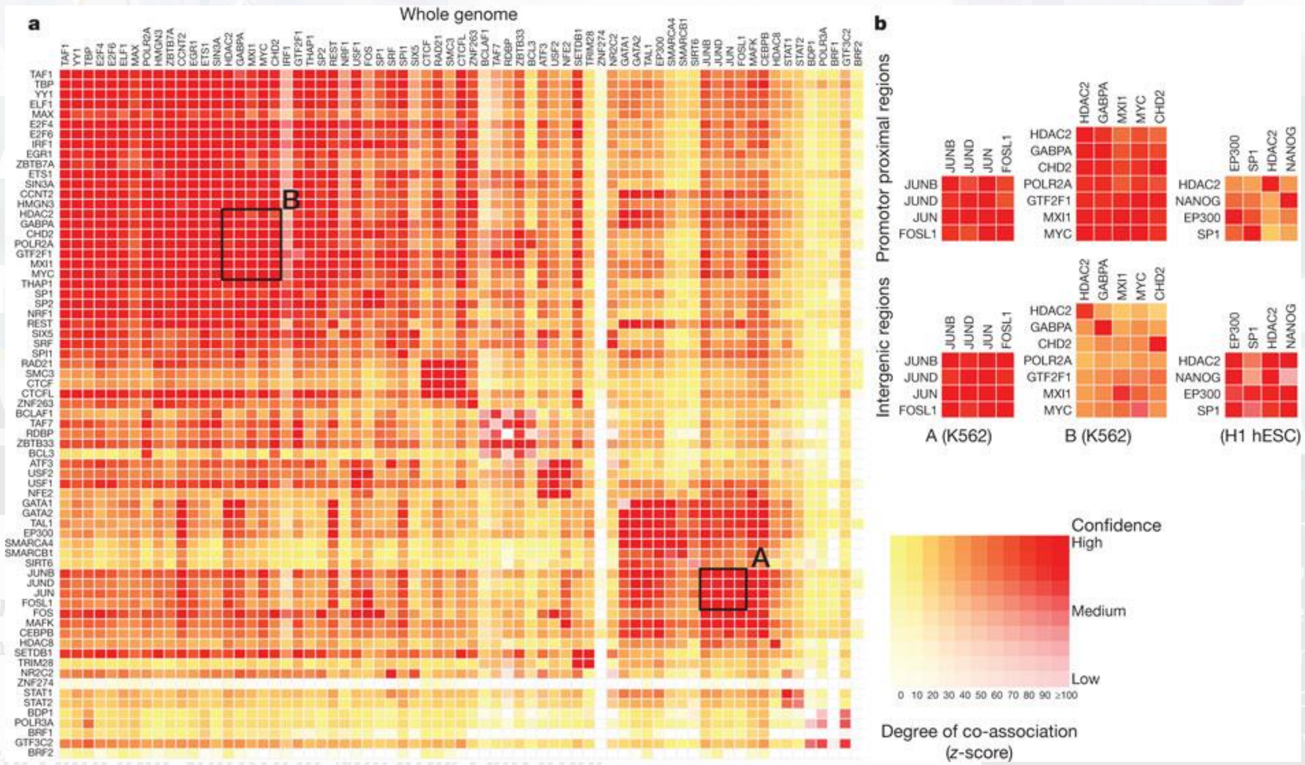
- The Square Kilometre Array (SKA) project is an international effort to build the world's largest radio telescope, with a square kilometre (one million square metres) of collecting area.
- World leading scientists and engineers designing and developing a system which will require supercomputers faster than any in existence in 2013, and network technology that will generate more data traffic than the entire Internet.

Exploration: genomics



Exploration: genomics

Co-association between transcription factors.



I Dunham *et al. Nature* **000**, 1-18 (2012) doi:10.1038/nature11247

Social Activity

Workshop on Optimization and Matrix Methods in Big Data

Big Data Structures

- Too much data: Large N
  - Bottleneck at processing
  - Computation
  - Estimates of precision
- Very complex data: small n, large p
- New types of data: networks, images, …
- "Found" data: credit scoring, government records, …

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

JANUARY 26 – 30, 2015

Workshop on Big Data and Statistical Machine Learning

FEBRUARY 9 – 13 , 2015

Workshop on Optimization and Matrix Methods in Big Data

serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops

cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life

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

Highlights from the workshops

Jan 9 – 23: Bootcamp

Jan 26 – 30: Statistical Machine Learning

Feb 9 – 11: Optimization and Matrix Methods

Feb 23 – 27: Visualization: Strategies and Principles

Mar 23 – 27: Health Policy

April 13 – 16: Social Policy

Opening Conference and Bootcamp

- Overview
  - Robert Bell, ATT: "Big Data: it's not the data"
  - Candes, Stanford: Reproducibility
  - Altman, Penn State: Generalizing PCA
  - …
- One day each: inference, environment, optimization, visualization, social policy, health policy, deep learning, networks
- Franke, Plante, et al. (2015): "A data analytic perspective on Big Data", http://arxiv.org/abs/1509.02900

Big Data and Statistical Machine Learning

- Mu Zhu – Towards deep learning
- Brendan Frey – The infinite genome project
- Samy Bengio – The battle against the long tail

Statistical Machine Learning

- Markov Random Field is essentially an exponential family model:

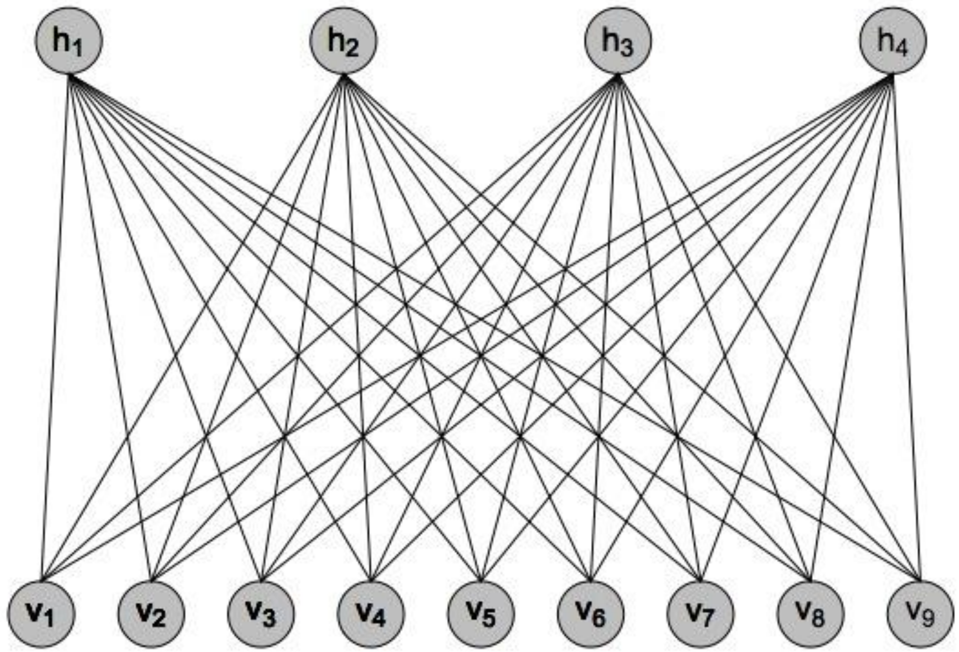
$$p(x;\eta) \propto \frac{1}{Z(\eta)} \exp\{\eta^T t(x)\}$$

- Restricted Boltzmann machine is a special case:

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

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

Restricted Boltzmann Machine



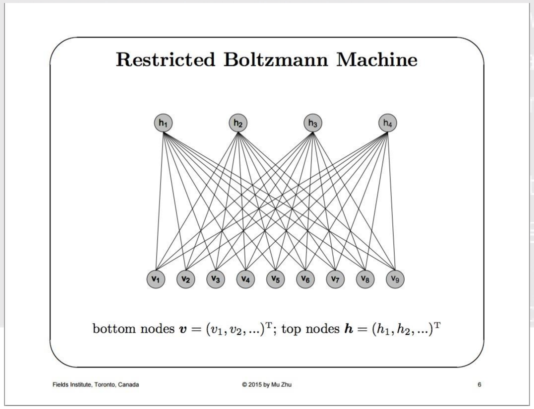
bottom nodes $\boldsymbol{v} = (v_1, v_2, ...)^{\mathrm{T}}$; top nodes $\boldsymbol{h} = (h_1, h_2, ...)^{\mathrm{T}}$

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

- if just one binary top node, model for $h \mid \underline{v}$ is a logistic regression

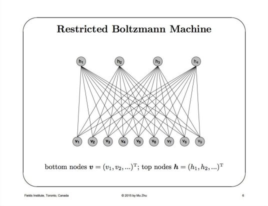
$$\log\{P(h=1 \mid v)/P(h=0 \mid v)\} = a + v^T w$$

- 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 $\underline{v}$



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

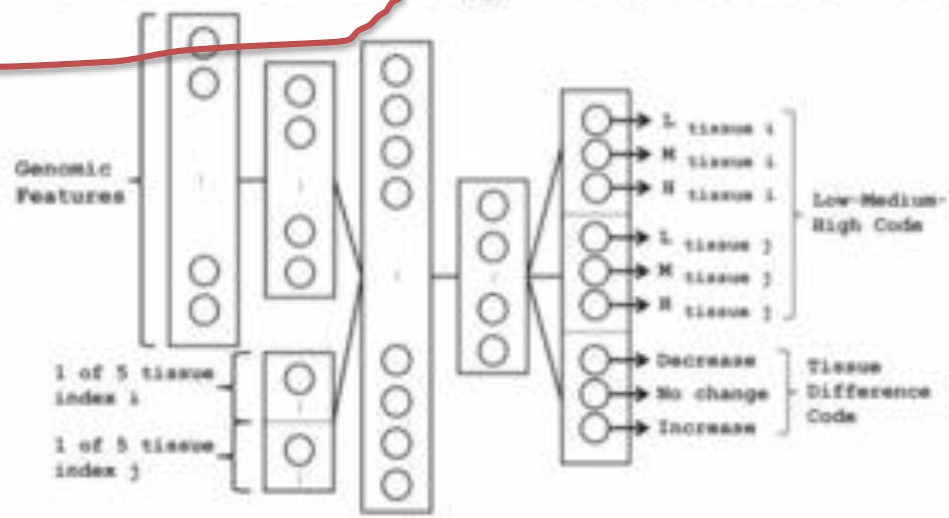
- top nodes $h_1, h_2, \dots$; bottom nodes $v_1, v_2, \dots$
- model for $h_t \mid \underline{v}, h_{-t}$ is logistic regression



- stack these in layers; top nodes for one layer become bottom nodes for the next layer
- some applications of deep learning have ~10 layers, with millions of units in each layer
- estimating parameters becomes an optimization and computational problem

Training

- ~160,000 training cases
  - 10,000 exons x 16 human tissues
- Target: Three Ψ levels (low, medium, high)
- Input: ~1400 features derived from genome
- Logistic regression, lasso, SVMs, …
- Bayesian neural network: Xiong et al, Bioinformatics 2011
- Deep neural network: Leung et al, Bioinformatics 2014



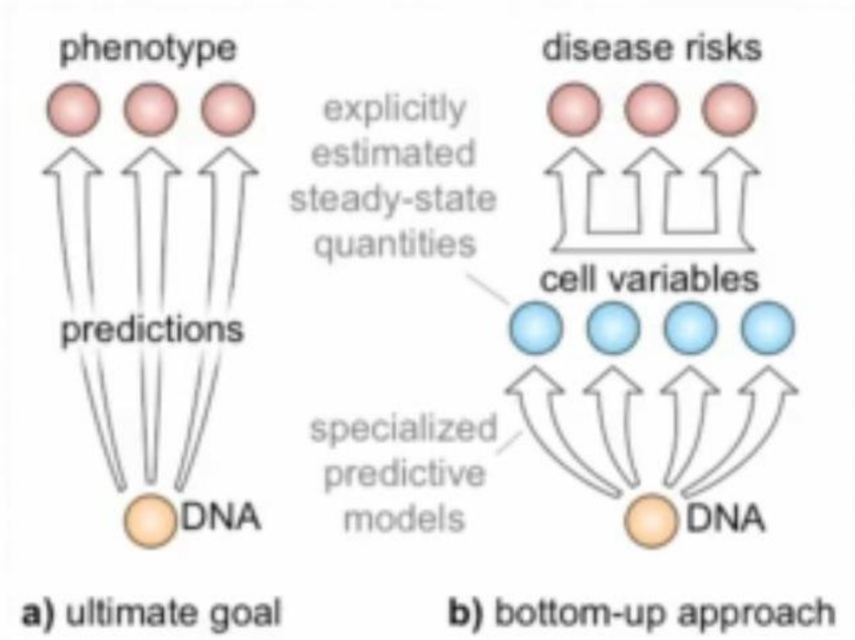
See also Barash et al, Nature 2010; Xiong et al, Science 2015

[Brendan Frey, *The Infinite Genomes Project*]

My group: The infinite genome project

Xiong et al, Science 2015; Barash et al, Nature 2010

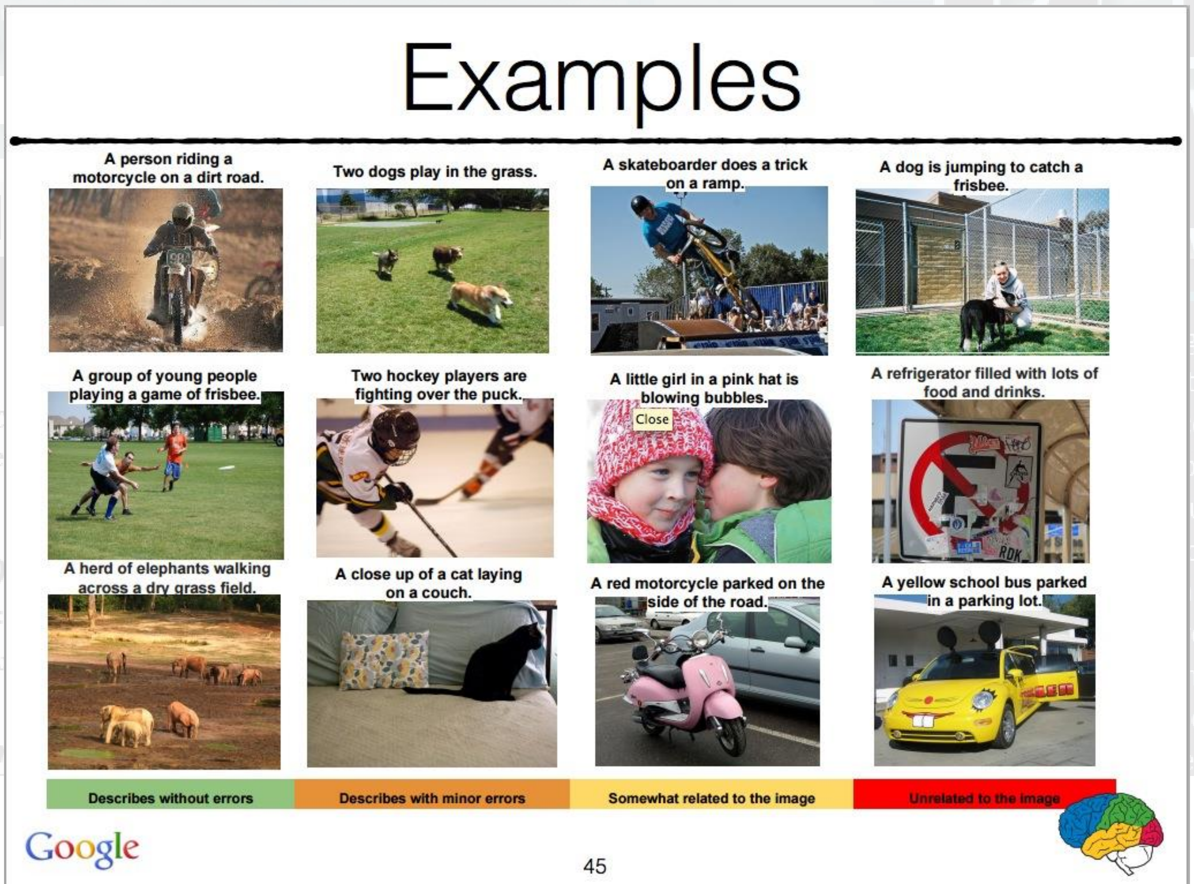
- Use statistical induction to infer a computational model that mimics crucial aspects of cell biology
- Use it to ascertain disease mutations



[Brendan Frey, *The Infinite Genomes Project*]

Statistical Machine Learning

- Bengio, S. (2015). The battle against the long tail. slides



Statistical Machine Learning

"The rise of the machines", *Economist*, May 9 2015

Optimization

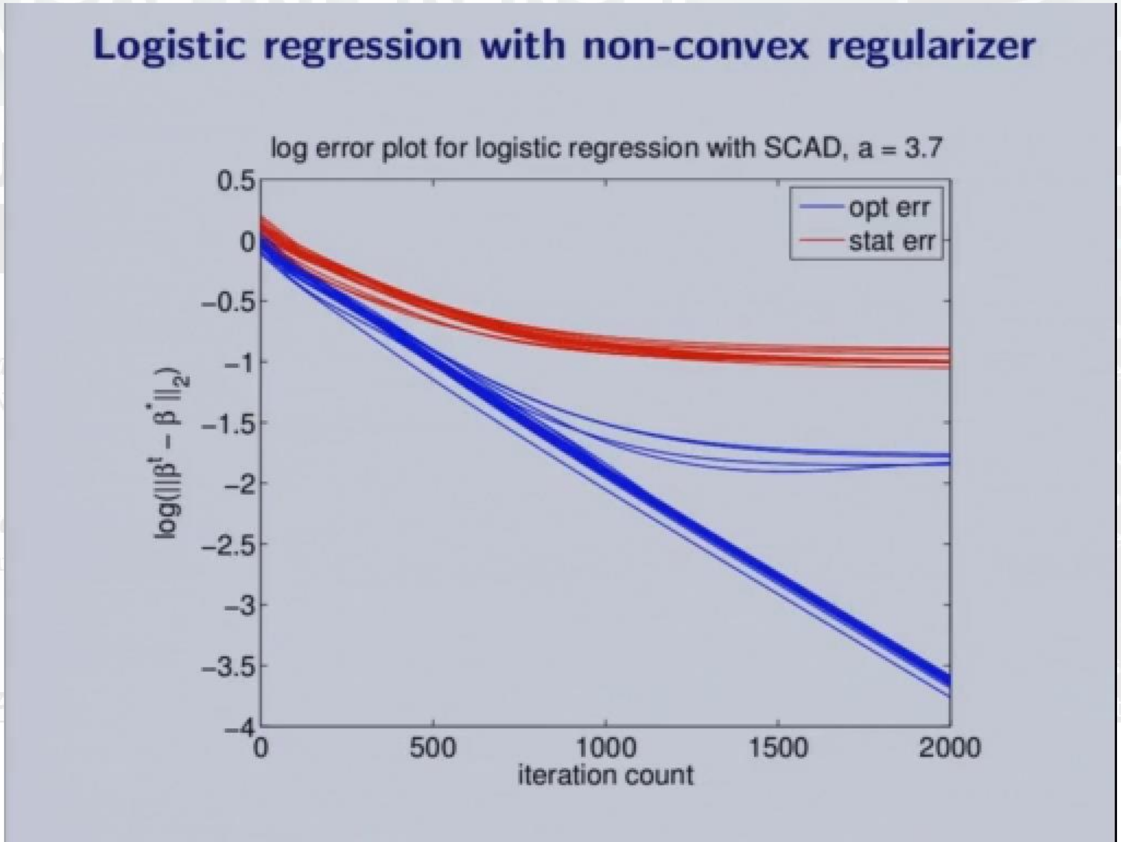
- Wainwright – non-convex optimization
- example: regularized maximum likelihood

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

- lasso penalty $||\theta||_1$ is convex relaxation of $||\theta||_0$
- many interesting penalties are non-convex
- optimization routines may not find global optimum

Wainwright and Loh

- distinction between statistical error $\hat{\theta} - \theta^*$
- and optimization error $\theta_t - \hat{\theta}$ (iterates)

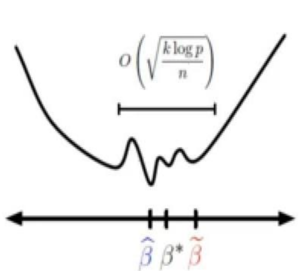


Wainwright and Loh

- a family of non-convex problems
- with constraints on the loss function (log-likelihood) and the regularizing function (penalty)
- conclusion: any local optimum will be close enough to the true value
- conclusion: can recover the true sparse vector under further conditions

Loh, P. and Wainwright, M. (2015). Regularized *M*-estimators w nonconvexity. *J Machine Learning Res.* 16, 559-616.

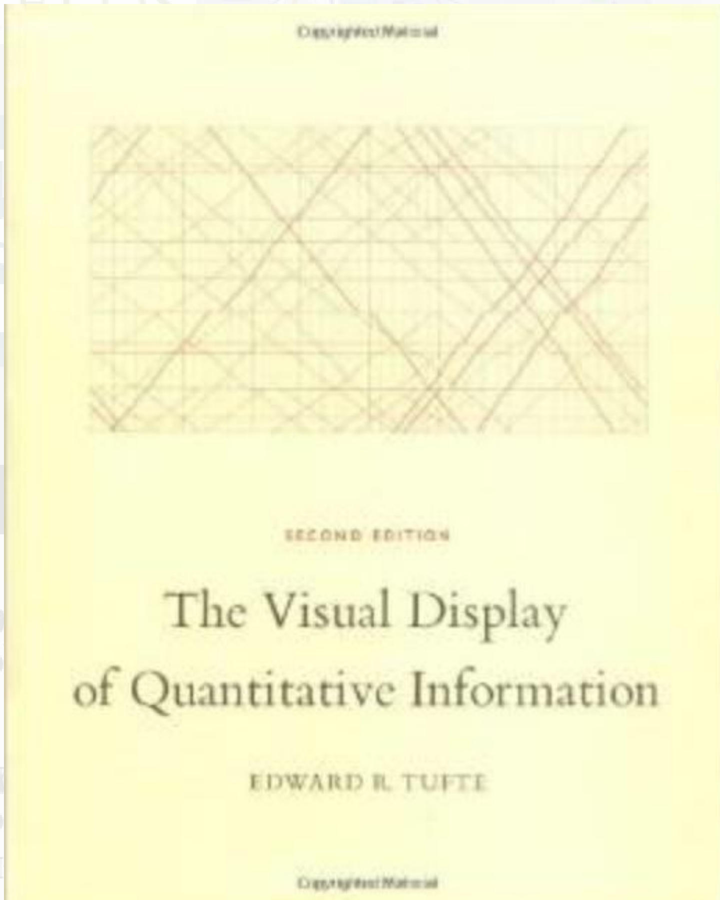
Loh, P. and Wainwright, M. (2014). Support recovery without incoherence. http://arxiv.org/abs/1412.5632



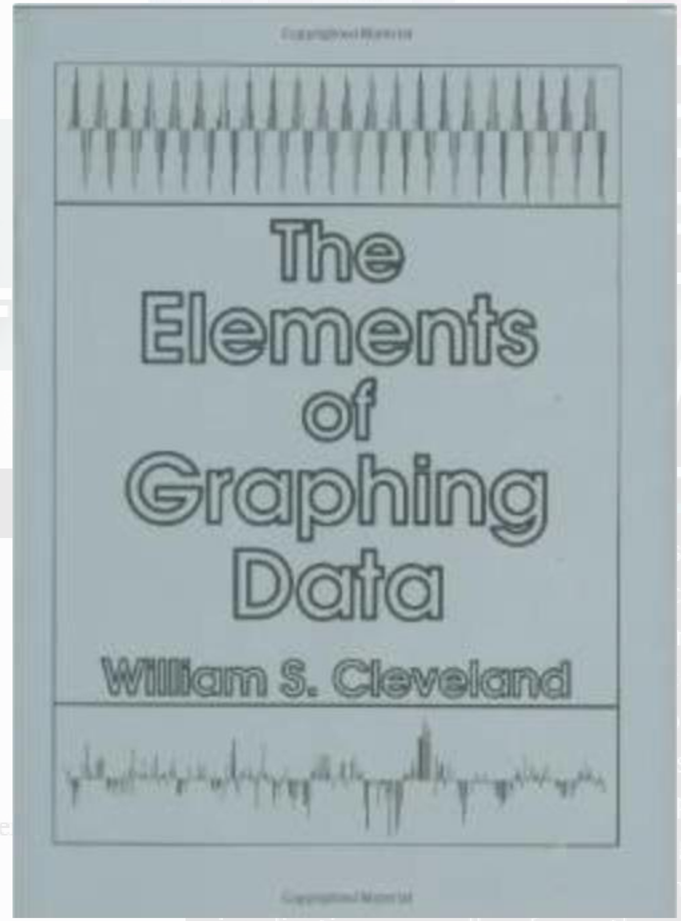
Visualization for Big Data
Strategies and Principles

- data representation
- data exploration via filtering, sampling and aggregation
- visualization and cognition
- information visualization
- statistical modeling and software
- cognitive science and design

Visualization for Big Data:
Strategies and Principles

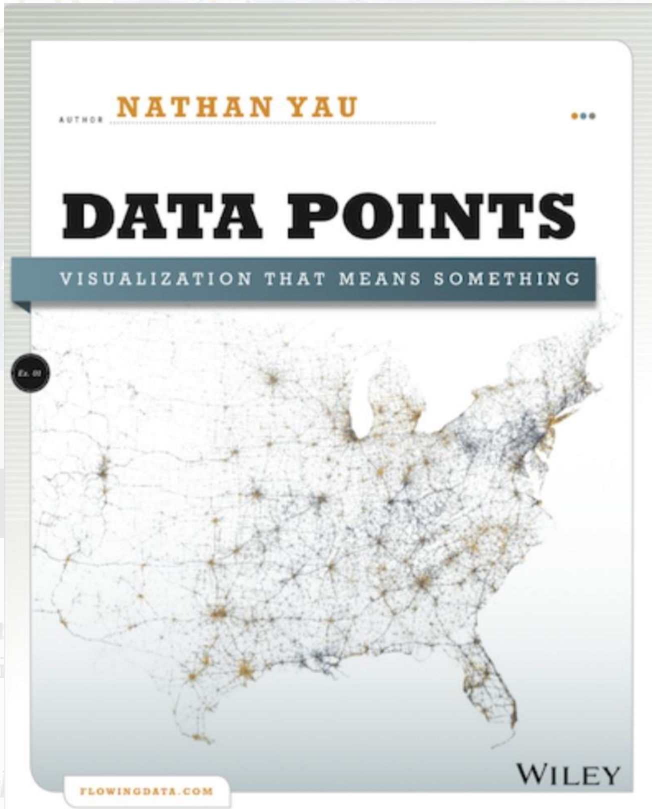


1983

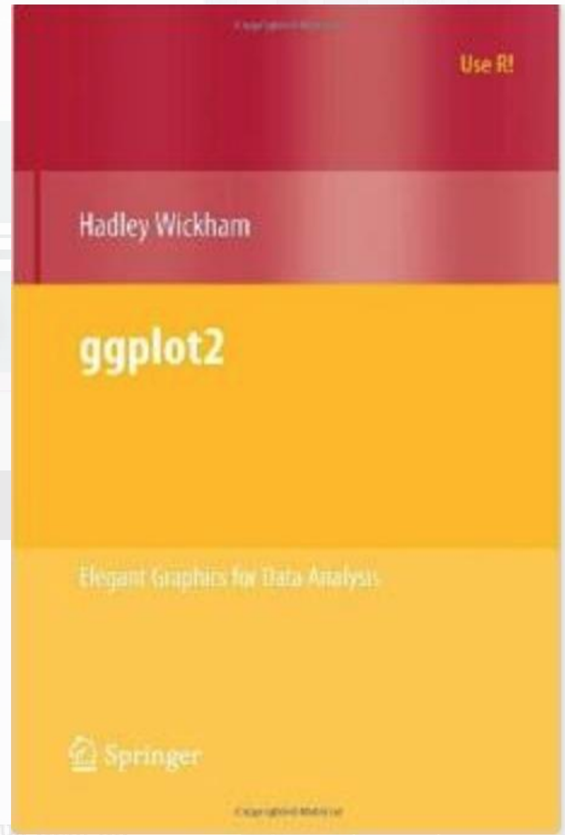


1985

Visualization for Big Data:
Strategies and Principles



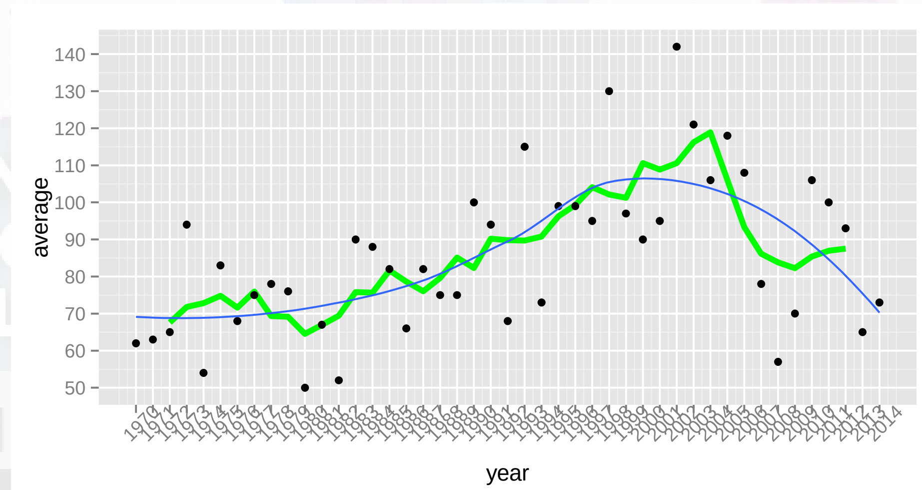
2013



2009

Statistical Graphics

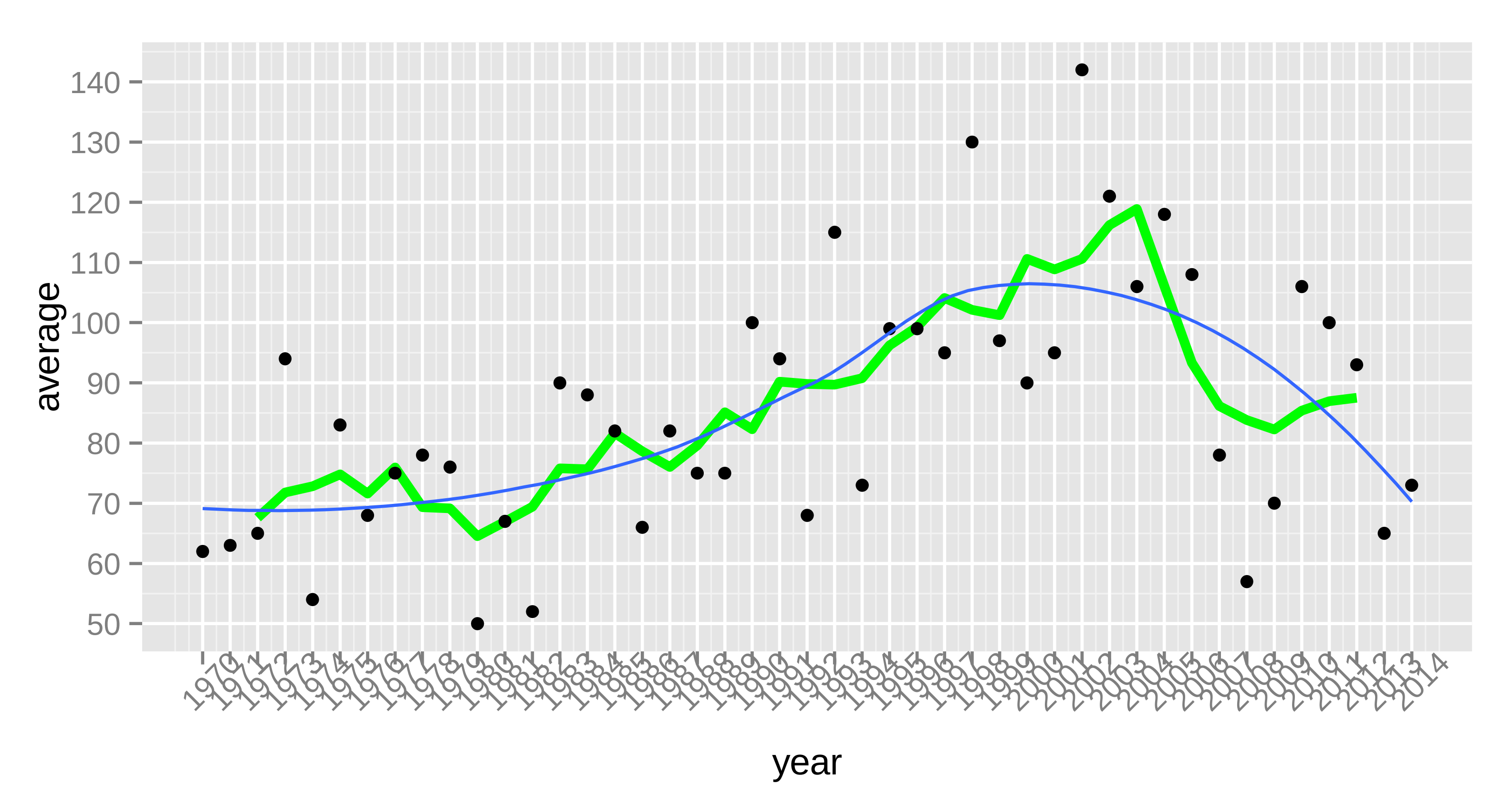
- convey the data clearly
- focus on key features
- easy to understand



- research in perception
- aspects of cognitive science

- must turn ‘big data’ into small data

- Rstudio, R Markdown
- `ggplot2, ggvis, dplyr, tidyr,`
- [cheatsheets]()




```
honeyplot +
    geom_line(aes(honey$year,honey$runmean),col = "green",size=1.5) +
    geom_point(aes(honey$year,honey$average),) +
    scale_x_continuous(breaks=1970:2014) +
    geom_smooth(method="loess",span=.75,se=F) +
    scale_y_continuous(breaks=seq(0,140,by=10)) +
    theme(axis.text.x = element_text(angle=45))
```

Big Data for Health Policy

- big data, causal inference, challenges
- graphical models and visualization
- data quality assessments of administrative data

  - pragmatic clinical trials
  - comparative effectiveness research
  - evidence mining research in e-record
  - health determinants
  - propensity score methods

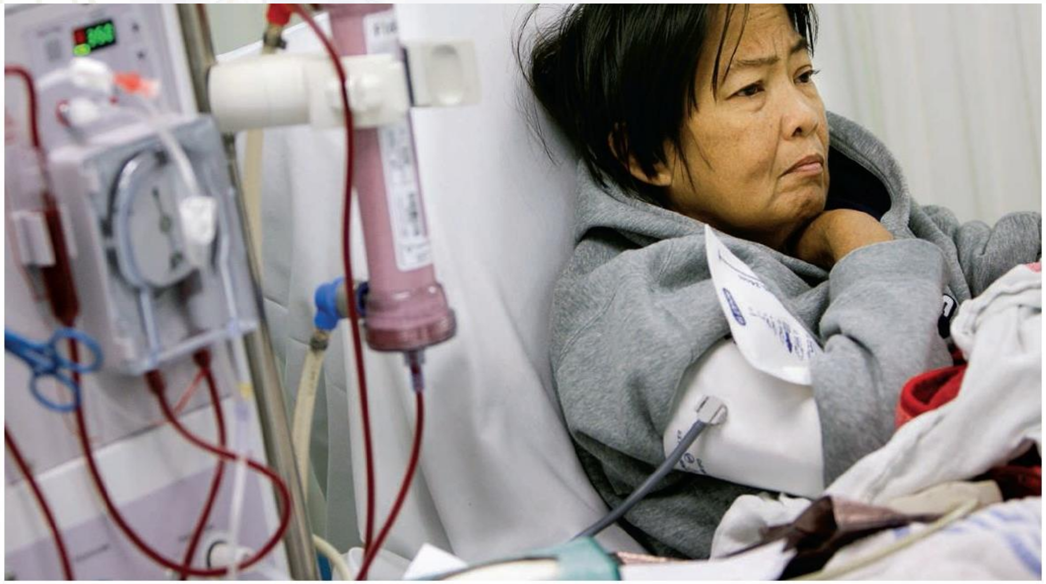
- data from multiple jurisdictions
- diagnostic test assessment
- marginal structural models
- dynamic periodicity and trend

  - big data and health policy research

Big Data for Health Policy

- Pragmatic clinical trials
  - Patrick Heagerty, Fred Hutchison
- Linking health and other social data-bases
  - Thérèse Stukel, ICES
- Privacy

Heagerty – Pragmatic Clinical Trials



One pragmatic clinical trial compares different approaches to dialysis. Studies like this will enroll a broader cohort, including more women and minorities.

MEDICAL RESEARCH

Clinical trials get practical

Many clinical trials don't help doctors make decisions. A new breed of studies aims to change that

By **Jennifer Couzin-Frankel**, *in Philadelphia, Pennsylvania*

trials will involve more women, more minorities, a range of incomes," says Monique Anderson, a cardiologist at Duke University

tend to focus on health behaviors or compare available treatments, not test experimental drugs, although that could change.

Nine Collaboratory trials are under way. One tests whether patients on dialysis are more likely to survive and stay healthier if the dialysis treatment itself lasts longer. The study is randomizing about 400 dialysis centers around the country to either continue with their usual routine—dialysis typically ranges from about 3 to 5 hours in the United States—or administer it for at least 4.25 hours. Patients receive information about the trial at their clinic and a toll-free number to call if they have questions for the research team or wish to opt out.

An opt-out model is an option only for some of the lowest risk clinical trials: U.S. regulations require active informed consent for studies of experimental drugs. Because current pragmatic trials are comparing approaches doctors already use routinely, even ethicists agree that enrolling everyone, unless someone objects, is often reasonable.

Other challenges come in figuring out the best way to design pragmatic studies.

Heagerty – Pragmatic Clinical Trials

Common Trial Designs

Parallel

Time 1
X
X
X
X
O
O
O
O

Crossover

Time 1	Time 2
X	O
X	O
X	O
X	O
O	X
O	X
O	X
O	X

Lisa Lix, U Manitoba

Heagerty – Pragmatic Clinical Trials

Stepped Wedge Design

Time

1	2	3	4	5
O	X	X	X	X
O	O	X	X	X
O	O	O	X	X
O	O	O	O	X

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

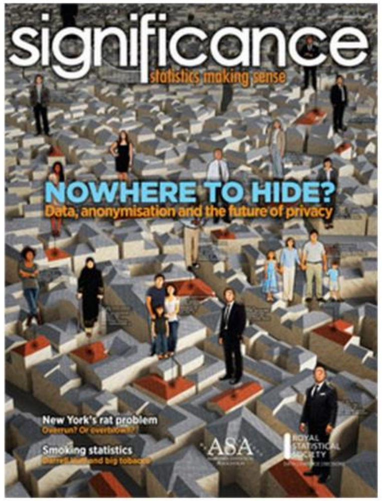
Thérèse Stukel, ICES

Big Data for Social Policy

Significance - October 2014 (Volume 11 Issue 4)

News, Interview and Editorial

Using Xbox polls to predict elections. The ISIS terror in numbers. Why South Koreans are heading for extinction. Tackling the reproducibility problem. How statistical models helped in the aftermath of the Boston Marathon bombings. And finally ... Fantasy author Jasper Fforde explains his theory of expectation-influenced probability.



Visualisation

Cultural movements

Mauro Martino on cognitive computing and mapping the migration of Western culture.

Special report: Data and privacy

Now you see me, now you don't

Does data anonymisation work? The answer depends on who you talk to. But finding a way to preserve privacy while sharing valuable data is crucial to the future of our information society.

Privacy

- anonymization/de-identification “HIPAA rules”
  - privacy commissioner of Ontario:
  - [“Big Data and Innovation, Setting the record straight: De-identification does work”]
  - Narayanan & Felten (July 2014) [“No silver bullet: De-identification still doesn’t work”]
- multi-party communication (Andrew Lo, MIT)
- statistical disclosure limitation
- differential privacy

  Slavkovic, A. -- Differentially Private Exponential Random Graph Models and Synthetic Networks

- Statistical Disclosure Limitation

- multi-party computation

- differential privacy

Current Issue: Volume 6, Issue 2 (2014)

Article

PDF [Face Recognition and Privacy in the Age of Augmented Reality]
Alessandro Acquisti, Ralph Gross, and Fred Stutzman

PDF [Top-Coding and Public Use Microdata Samples from the U.S. Census Bureau]
Nicole Crimi and William Eddy

PDF [Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record Linkage]
Frank Niedermeyer, Simone Steinmetzer, Martin Kroll, and Rainer Schnell

PDF [A Graph-based Approach to Key Variable Mapping]
duncan smith and Mark Elliot

Reader from: Nairobi, Nairobi Area, Kenya

Global Measures of Data Utility for Microdata Masked for Disclosure Limitation
Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, Alan F. Karr

What did we learn?

1. Big data means big models: complex, high-dimensional
   - regularization to induce sparsity
   - sparsity assumed or imposed
   - layered architecture complex graphical models
   - dimension reduction PCA, ICA, etc.
   - ensemble methods aggregation of predictions

2. Computational challenges include size and speed
   - ideas of statistical inference get lost in the machine

3. Data owners understand 2., but not 1.

4. Data science may be the best way to combine these

Gartner Hype Cyc
July 2014

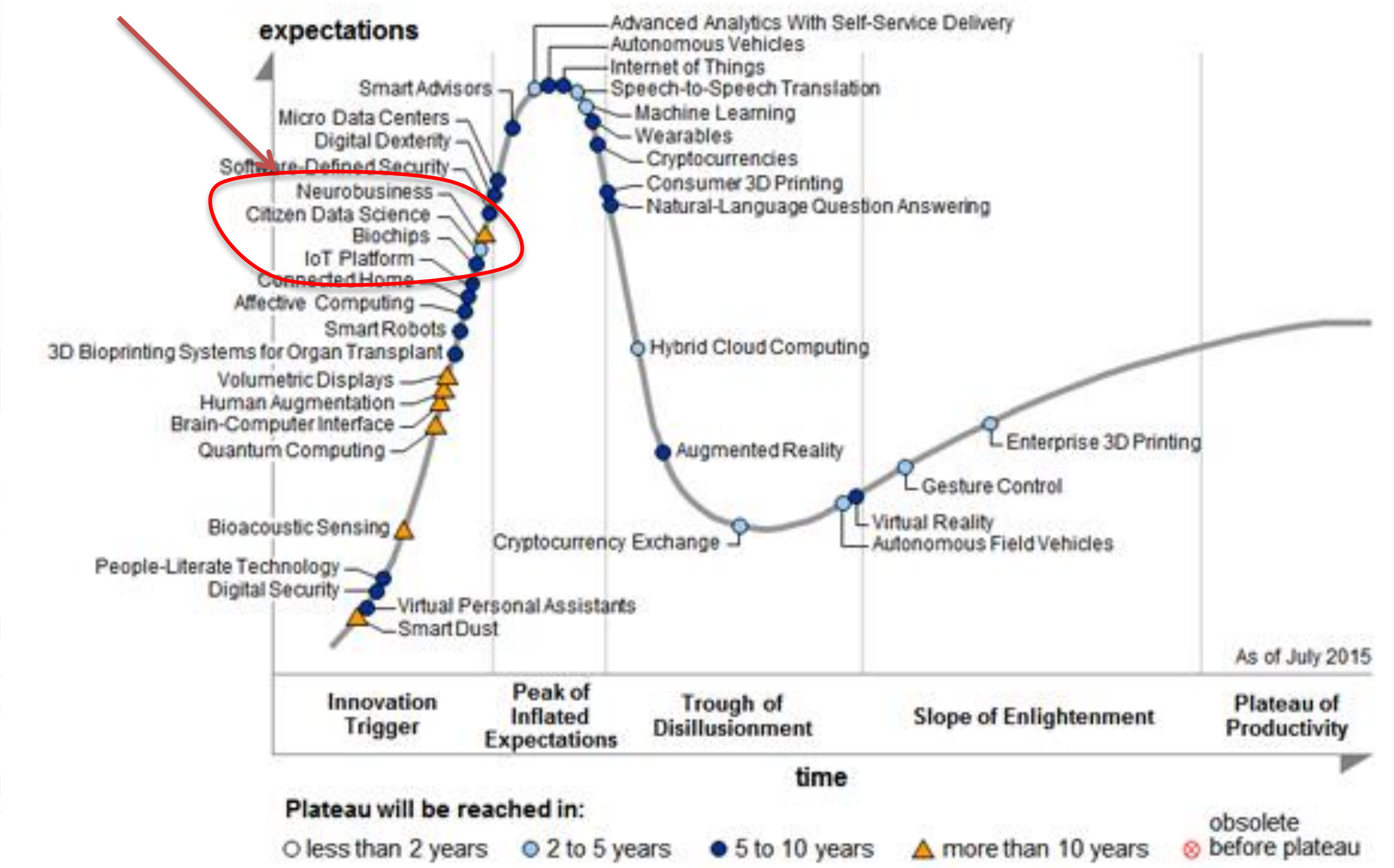
falling-99183_640 →

ttp://www.gartner.com/technology/research/hype-cycles/

https://etechlib.wordpress.com/tag/hype-cycle/

July 2015

"Citizen Data Science"



What did I learn?

- Big Data is real, and here to stay
- Big Data often quickly becomes small
  - by making models more and more complex
  - by looking for the very rare/extreme points
  - through visualization
- Big Insights build on old ideas
  - planning of studies, bias, variance, inference
- Big Data is a Big Opportunity

A few resources

- Franke, Plante et al. (2015). Statistical inference, learning and models in big data.
  - http://arxiv.org/abs/1509.02900
- Talks from the closing workshop for the Big Data program
- data science programs: U Michigan, Beijing, Johns Hopkins, UC Berkeley, Columbia, NYU, Dalhousie, UBC, U Toronto, ...