

Statistical Inference, Learning and Models for Big Data

FIELDS

THE FIELDS INSTITUTE

THEMATIC PROGRAM ON
STATISTICAL INFERENCE,
LEARNING, AND MODELS FOR

Nancy Reid
University of Toronto

JANUARY - JUNE, 2015

BIG
DATA

[交通地图](#) | [光华地图](#) | [周边住宿](#) | [用户登录](#)



北京大学统计科学中心

Center for Statistical Science, Peking University

Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio,
Hugh Chipman, Bin Yu

August 12, 2015

FEBRUARY 9 - 13, 2015

Workshop on Optimization and Matrix Methods in Big Data

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



THE FIELDS INSTITUTE

FIELDS

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

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)

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)

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



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



Pacific
Institute for
Mathematical
Sciences



FIELDS



Centre de Recherches Mathématiques



**NSERC
CRSNG**



Ontario

Fields Institute
for Research in
the
Mathematical
Sciences

JANUARY 12 - 23, 2011
Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Li

JANUARY 26 - 30, 2011
Workshop on Big Data and Statistical Machine Learning
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio,
Hugh Chipman, Bin Yu

Workshop on Optimization

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 emerging themes, such as learning and inference, as well as focus themes for applications in the social, physical and life

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
- Deep Learning Summer School August 3 – 12, CRM

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

Industrial Problem Solving Workshop

May 25 – 29

Fields Summer Undergraduate Research Program

May to August, 2015



Ruslan Salakhutdinov, Toronto



Mu Zhu, Waterloo

Watch  events on **FieldsLive**



MDM 12 – Einat Gil et al.

THE FIELDS INSTITUTE

G
A



Bin Yu

mans, Y

Data



els
ll
n,
d

and
or
life

Big Data – Big Topic

- Where to start?
- Look up some references

Google

big data

Web

News

Images

Videos

Books

More ▾

Search tools

About 770,000,000 results (0.32 seconds)

- Likelihood 78 m

- Statistical inference 7m

FEBRUARY 9 - 13, 2015

Workshop on Optimization and Matrix Methods in Big Data

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

The Blogosphere



Gil Press
Contributor

TECH

9/03/2014 @ 8:01AM | 12,133 views

12 Big Data Definitions: What's Yours?

[+ Comment Now](#)

[+ Follow Comments](#)

PROGRAM

JANUARY 12 - 23, 2015

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Krawcheck

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

Forbes ▾

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

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



The Blogosphere



FIELDS

WIRED

OPINION

big data

Business

memes

Science

Stop Hypeing Big Data and Start Paying Attention to 'Long Data'

BY SAMUEL ARBESMAN 01.29.13 | 9:30 AM | PERMALINK

f Share

15

🐦 Tweet

6

g+1

156

in Share

26

Pin it



The Blogosphere

Kim Crawley, researcher, InfoSec Institute

December 22, 2014

The problem with Big Data

Share this article:



Big Data is a big buzzword these days. It's completely understandable why so many people in tech talk about it, even though few people completely understand it.

So, what is Big Data?

With the massive growth of data centers worldwide in the past thirty years or so, we're creating, transmitting, and storing more data than ever before.

We're well beyond terabytes, petabytes and now even exabytes. We're quickly zooming into zettabytes in global capacity and



The Blogosphere

datascience@berkeley
Blog

datascience@berkeley Blog

What Is Big Data?

September 3rd, 2014 by *Jenna Dutcher* 5 Comments

Filed Under: *Big Data*

datascience@berkeley



JANUARY - JUNE, 2015

PROGRAM

JANUARY 12 - 23, 2015

To settle the question once and for all, we asked more than 40 thought leaders in publishing, fashion, food, automobiles, medicine, marketing, and every industry in between how exactly they would define the phrase "big data." Their answers might surprise you! Take a look below to find out what big data is:

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

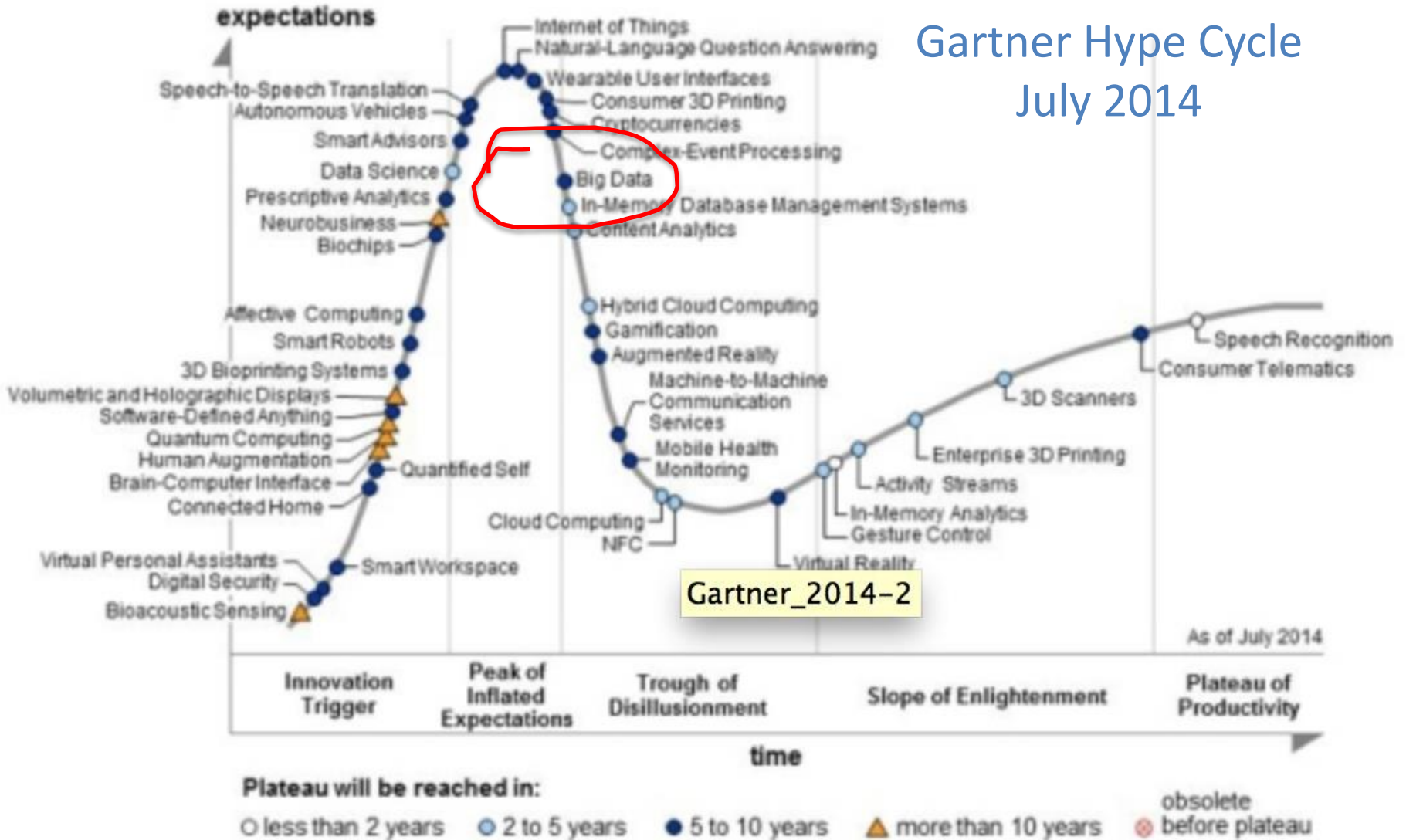
This thematic program emphasizes both applied and theoretical aspects of

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

Gartner Hype Cycle July 2013



Gartner Hype Cycle July 2014



falling-99183_640 →

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

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

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

Workshop on Big Data and Statistical Machine Learning

Organizing committee: Bin Yu (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Nancy Reid

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 concentrate on overview lectures and the program, concentrating on overview lectures and statistical inference. 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



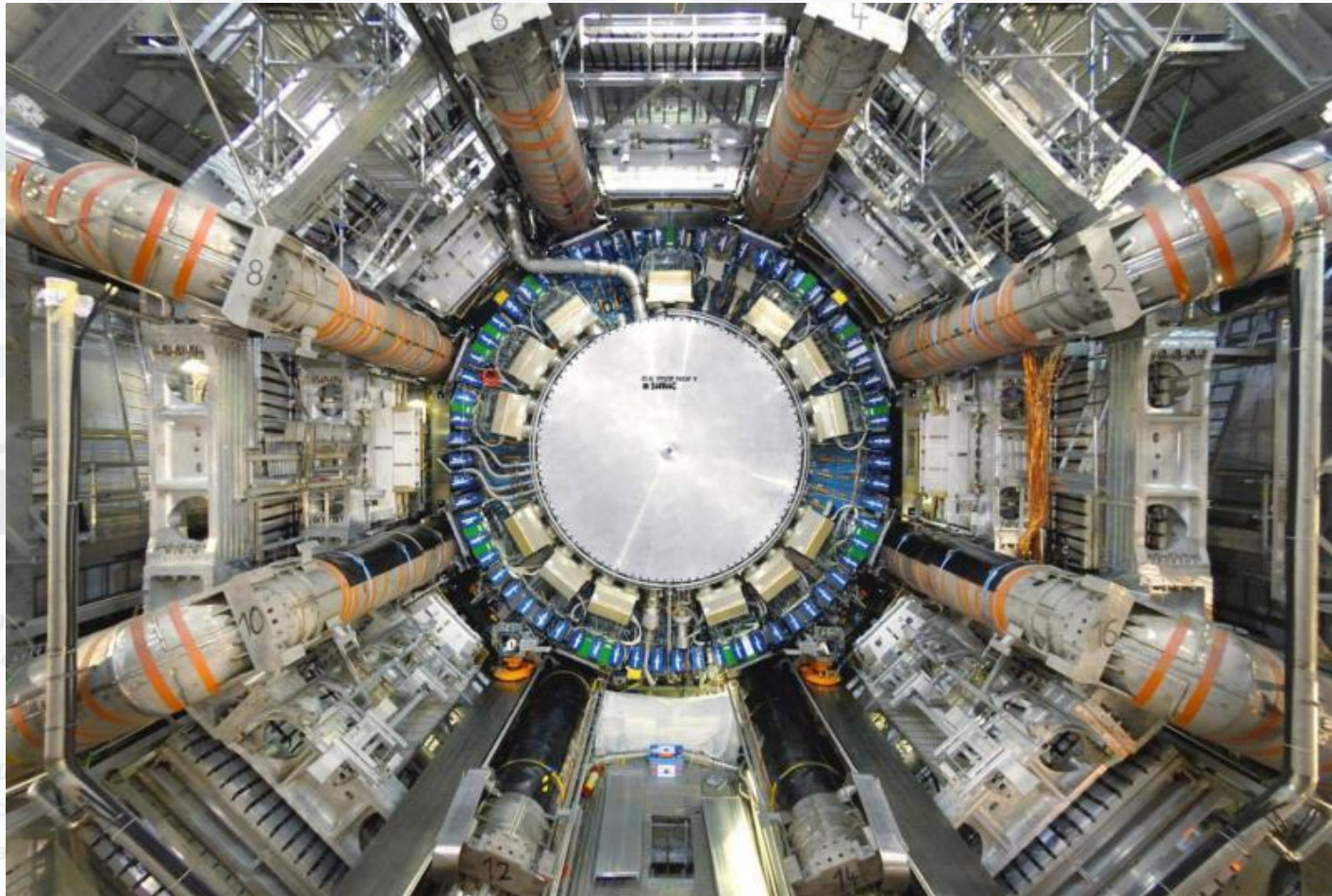
Big Data Types

- 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



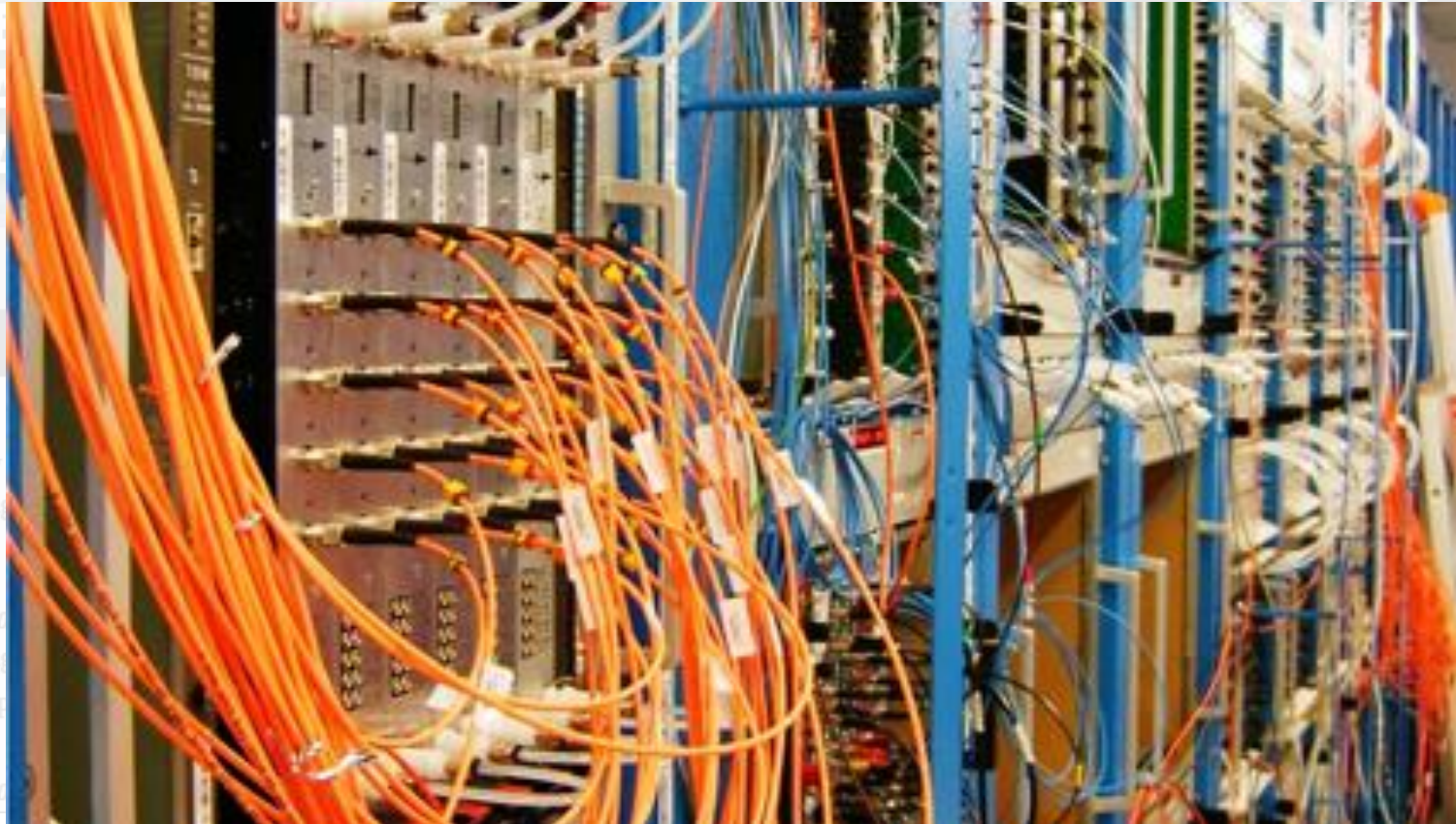
*Opening Co
Organizing Co*

*Workshop
Organizing co
Hugh Chipma*

Workshop on Optimization and Matrix Methods in Big Data

phases
al aspects of
ng and models
inference will
the program,
lectures and
Workshops
will highlight
as learning and
visualization, as well as focus themes for
applications in the social, physical and life

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



Opening
Organizing

Workshop
Organizing
Hugh Chip

Workshop

emphasizes
aspects of
and models
reference will
the program,
structures and
workshops
highlight
learning and
themes for
physical and life

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.



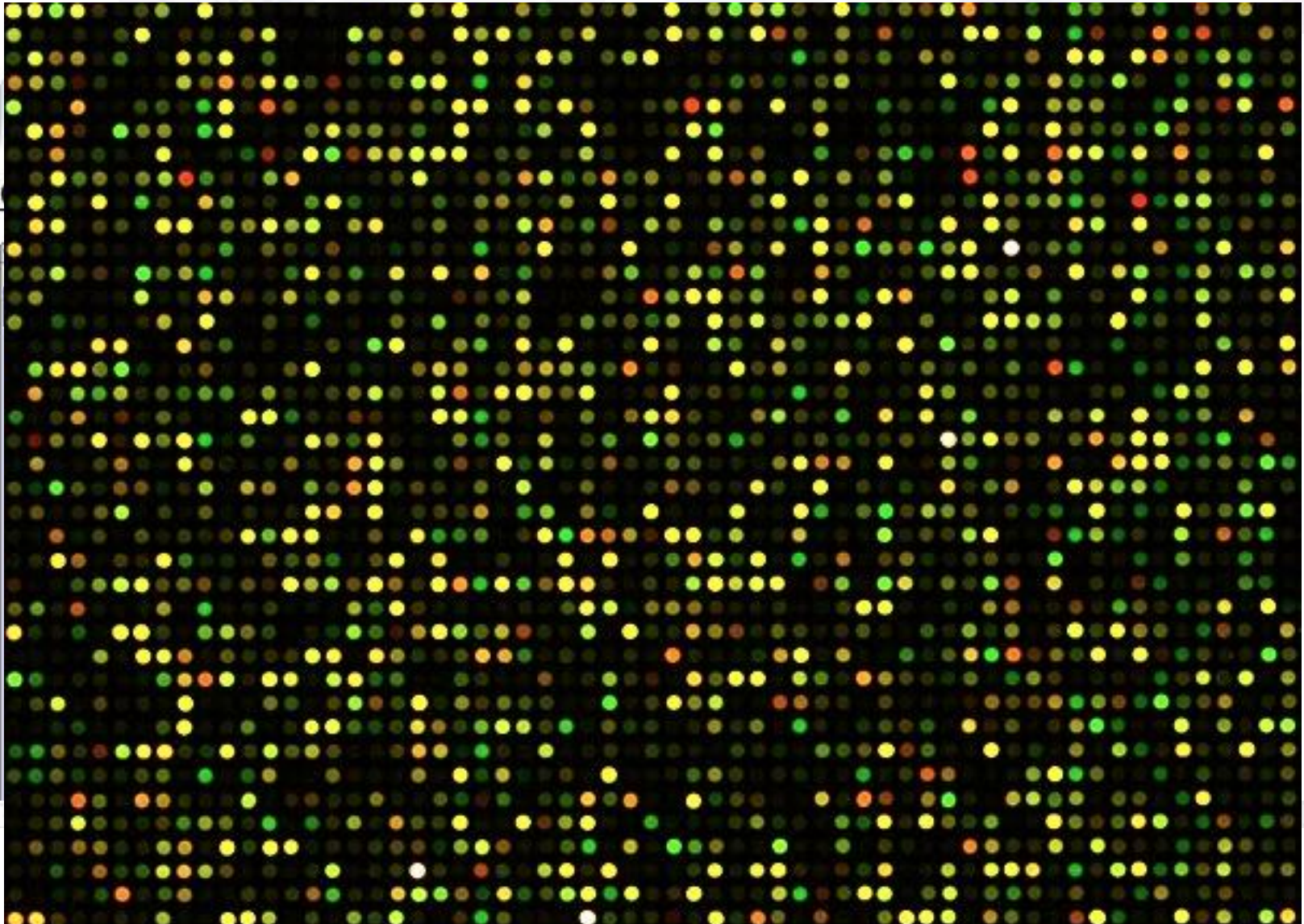
Opening Conference
Organizing Committee:

Workshop on Big Data
Organizing committee: R
Hugh Chipman, Bin Yu

Workshop on Optimiz

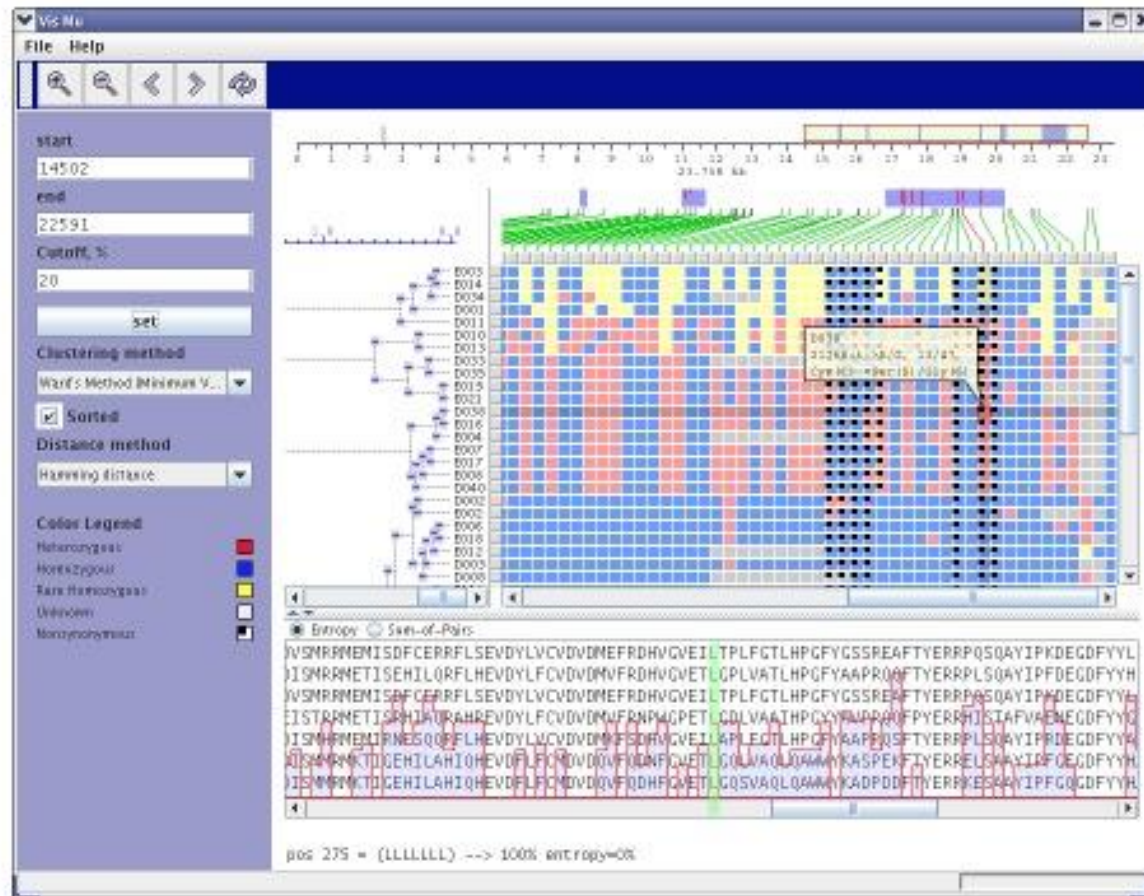
rogram emphasizes
d theoretical aspects of
nce, learning and models
opening conference will
duction to the program,
n overview lectures and
paration. Workshops
rogram will highlight
emes, such as learning and
well as focus themes for
applications in the social, physical and life

Exploration: microarray



SNP-VISTA

GeneSNP-VISTA: Visualization of mutations in genes



TECHNOLOGY

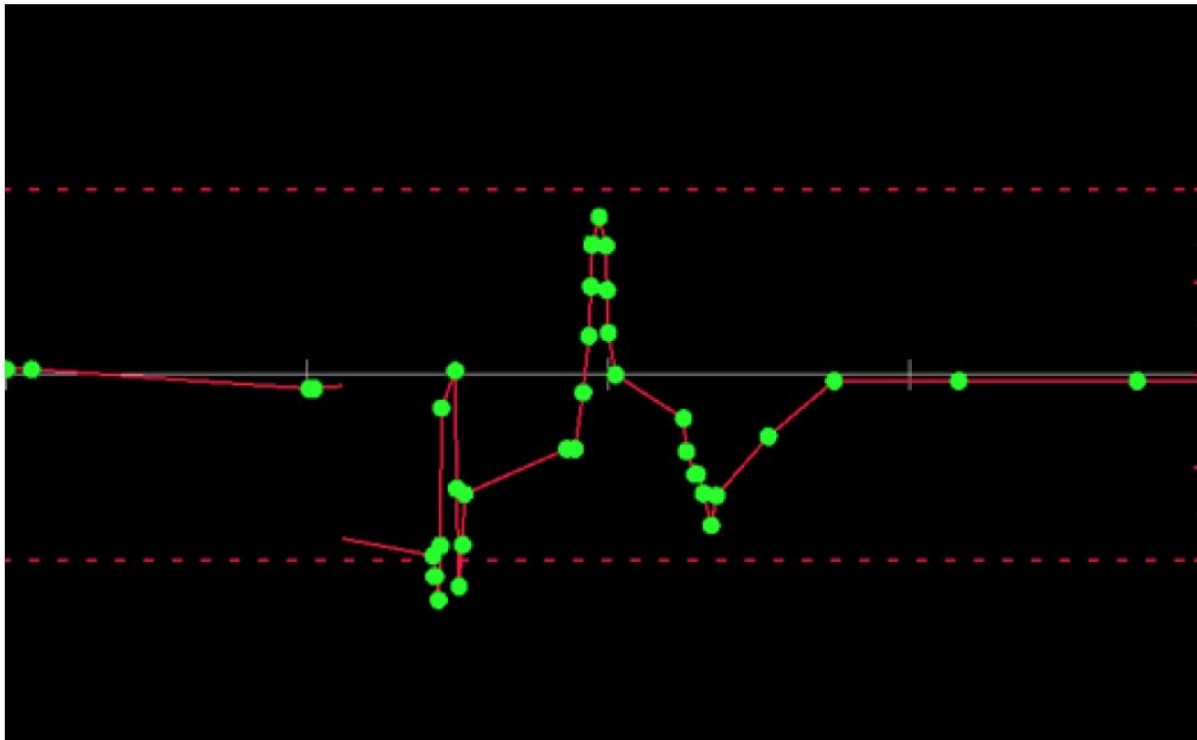
BOSTON'S 'STREET BUMP' APP TRIES TO AUTOMATICALLY MAP POTHOLES WITH ACCELEROMETERS AND GPS

By Clay Dillow Posted February 10, 2011





 247 Shares



**COUNT THE SW
MOMENTS, NO
THE CALORIES.**

STIR THINGS UP >

**STEVIA
IN THE
RAW**



JTF

of
odels
will
gram,
and
s
ght
ing and
es for
nd life

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

Organizing Committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yashua Bengio, Hugh Chipman, and Y

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 will highlight 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: Deep Learning
- Feb 9 – 11: Optimization
- Feb 23 – 27: Visualization
- Mar 23 – 27: Health Policy
- April 13 – 16: Social Policy

**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

Opening Conference and Bootcamp

- Overview

- Bell: “Big Data: it’s not the data”
- Candes: Reproducibility
- Altman: Generalizing PCA

- One day each: **inference**, environment, **optimization**, visualization, **social policy**, health policy, **deep learning**, networks

PROGRAM

JANUARY 12 – 23, 2015

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

Workshop on Big Data and Statistical Machine Learning

Organized by: Sallie Keller, Bin Yu, (Co-organizers) Sallie Keller, Bin Yu, 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 machine learning and models in big data. The opening conference will provide 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

Big Data and Statistical Machine Learning

- Roger Grosse – Scaling up natural gradient by factorizing Fisher information
- Samy Bengio – The battle against the long tail

PROGRAM

JANUARY 12 - 23, 2015

Opening Conference and Boot Camp

Organizing committee: Ruslan Salakhutdinov, Yoshua Bengio, Li Deng, D. Ermiya, and Li Fei-Fei

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 for 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

Statistical Machine Learning

- Grosse, R. (2015). Scaling up natural gradient by factorizing Fisher information.

to appear in *International Conference on Machine Learning*.

- Markov Random Field is essentially an exponential family model:

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

- Restricted Boltzmann machine is a special case:

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

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

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 introductory lectures and hands-on preparatory 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

Statistical Machine Learning

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

- natural gradient ascent

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

- uses Fisher information as metric tensor
- Gaussian graphical model approximation to force sparse inverse

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

Statistical Machine Learning

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

Examples

A person riding a motorcycle on a dirt road.



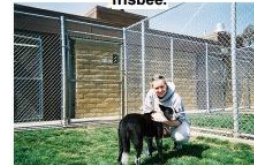
Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



Describes without errors

Describes with minor errors

Somewhat related to the image

Unrelated to the image



Statistical Machine Learning

Some you win, some you lose

Image-recognition software's analysis of what a picture represents



"A person riding a motorcycle on a dirt road"



"A yellow school bus parked in a car park"

Source: "Show and Tell: A Neural Image Caption Generator", Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan

FEBRUARY 9 - 13, 2015

Workshop on Optimization and Matrix Methods in Big Data

visualization, as well as focus themes for applications in the social, physical and life

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

Optimization

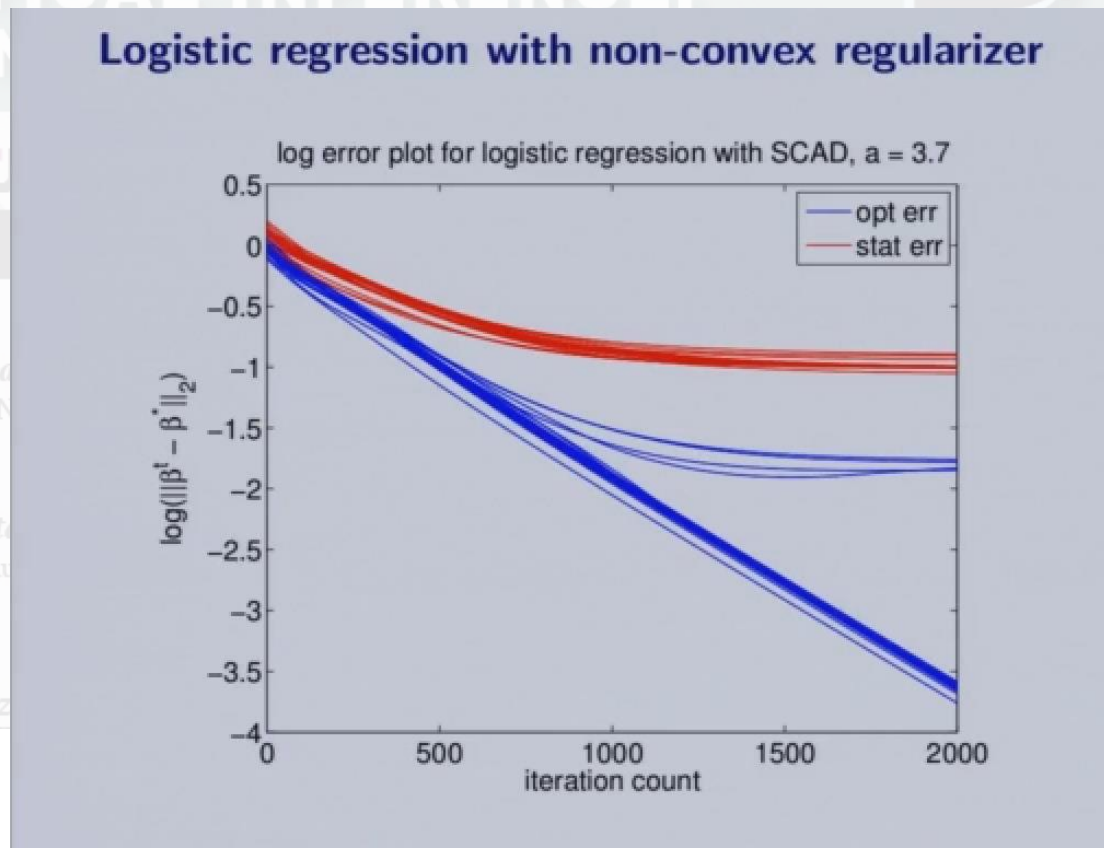
- Wainwright – non-convex optimization
- example: regularized maximum likelihood

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

- 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)



Opening Conference of
Organizing Committee: N

Workshop on Big Data
Organizing committee: Ru
Hugh Chipman, Bin Yu

Workshop on Optimiz

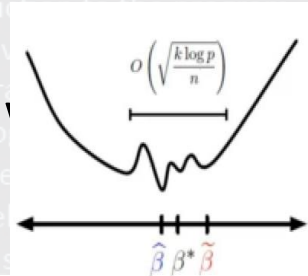
program emphasizes
d theoretical aspects of
nce, learning and models
opening conference will
roduction to the program,
n overview lectures and
paration. Workshops
program will highlight
emes, such as learning and
well as focus themes for
he social, physical and life

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 and nonconvexity. *J Machine Learning Res.* 16, 559-616.

Loh, P. and Wainwright, M. (2014). Support recovery without incoherence. Arxiv preprint

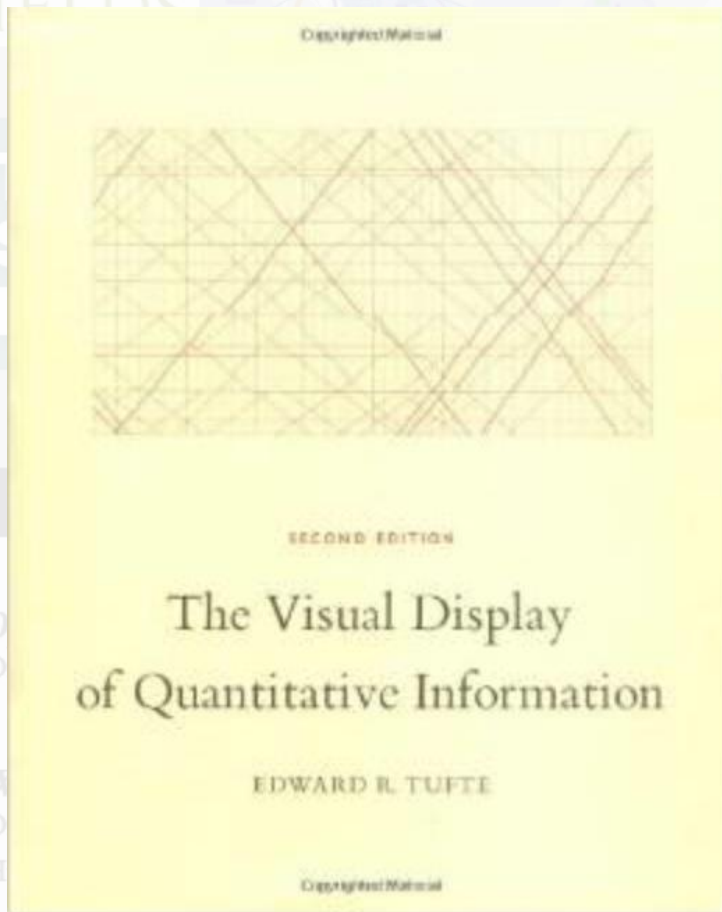


Visualization for Big Data Strategies and Principles

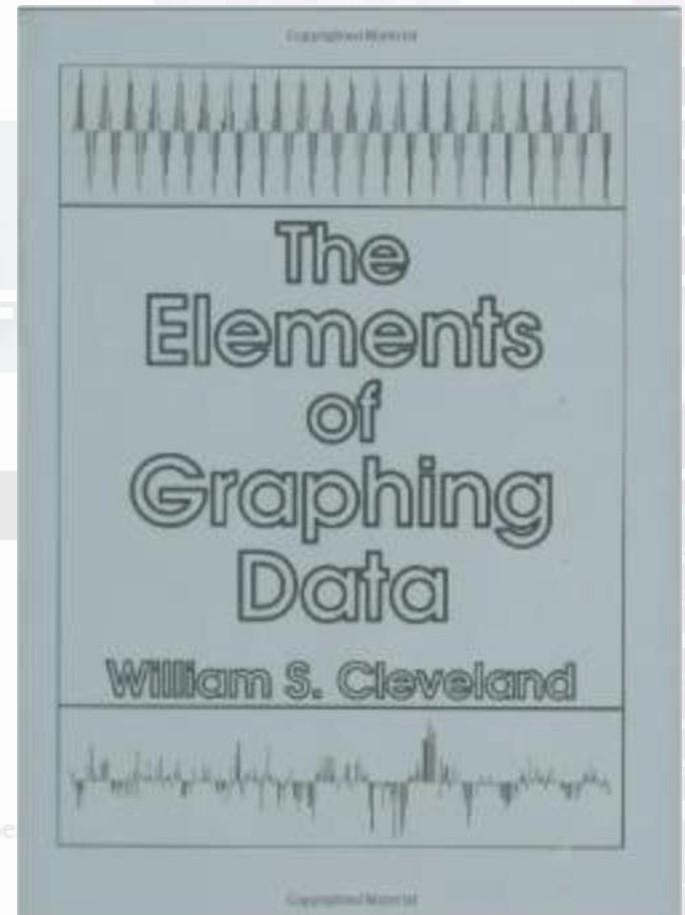
- data representation
- data exploration via filtering, sampling and aggregation
- visualization and cognition
- information visualization
- statistical modeling and software
- cognitive science and design

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models for 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

Visualization for Big Data: Strategies and Principles

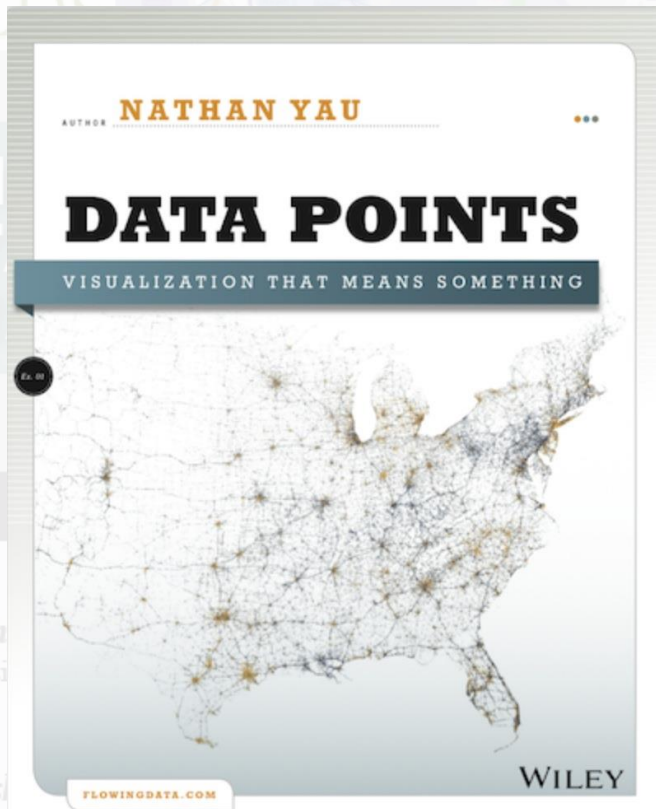


1983

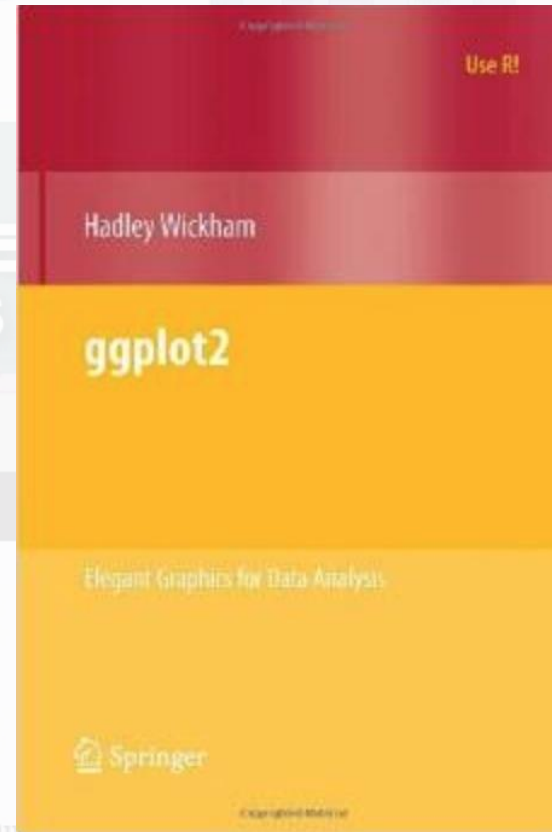


1985

Visualization for Big Data: Strategies and Principles



2013



2009

Openin
Organiz

Worksh

Organizing
Hugh Chipman, Bin Yu

Workshop on Optimization and Matrix Methods in Big Data

ON
ENCE
DELS
2015

Lix, Bin Yu

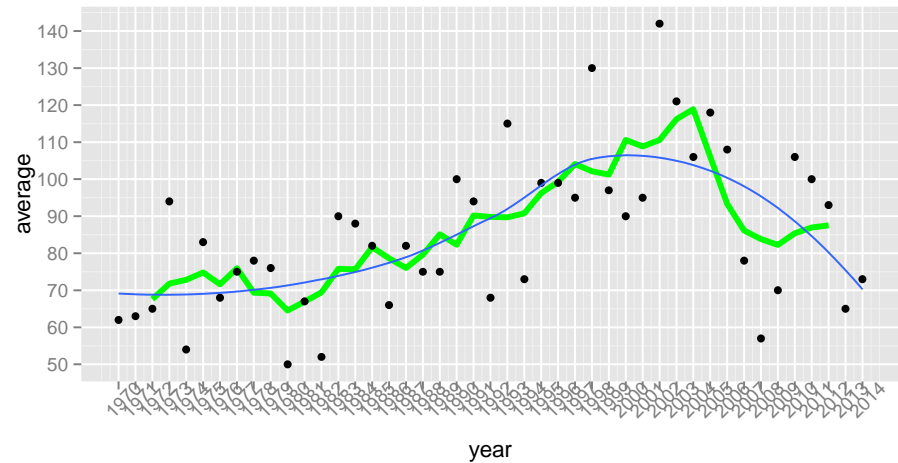
ning

huurmans, Yoshu

emphasizes
retical aspects of
arning and models
g conference will
n to the program,
view lectures and
n. Workshops
throughout the program will highlight
cross-cutting themes, such as learning and
ization, as well as focus themes for
applications in the social, physical and life

Statistical Graphics

- convey the data clearly
- focus on key features
- easy to understand
- research in perception
- aspects of cognitive science



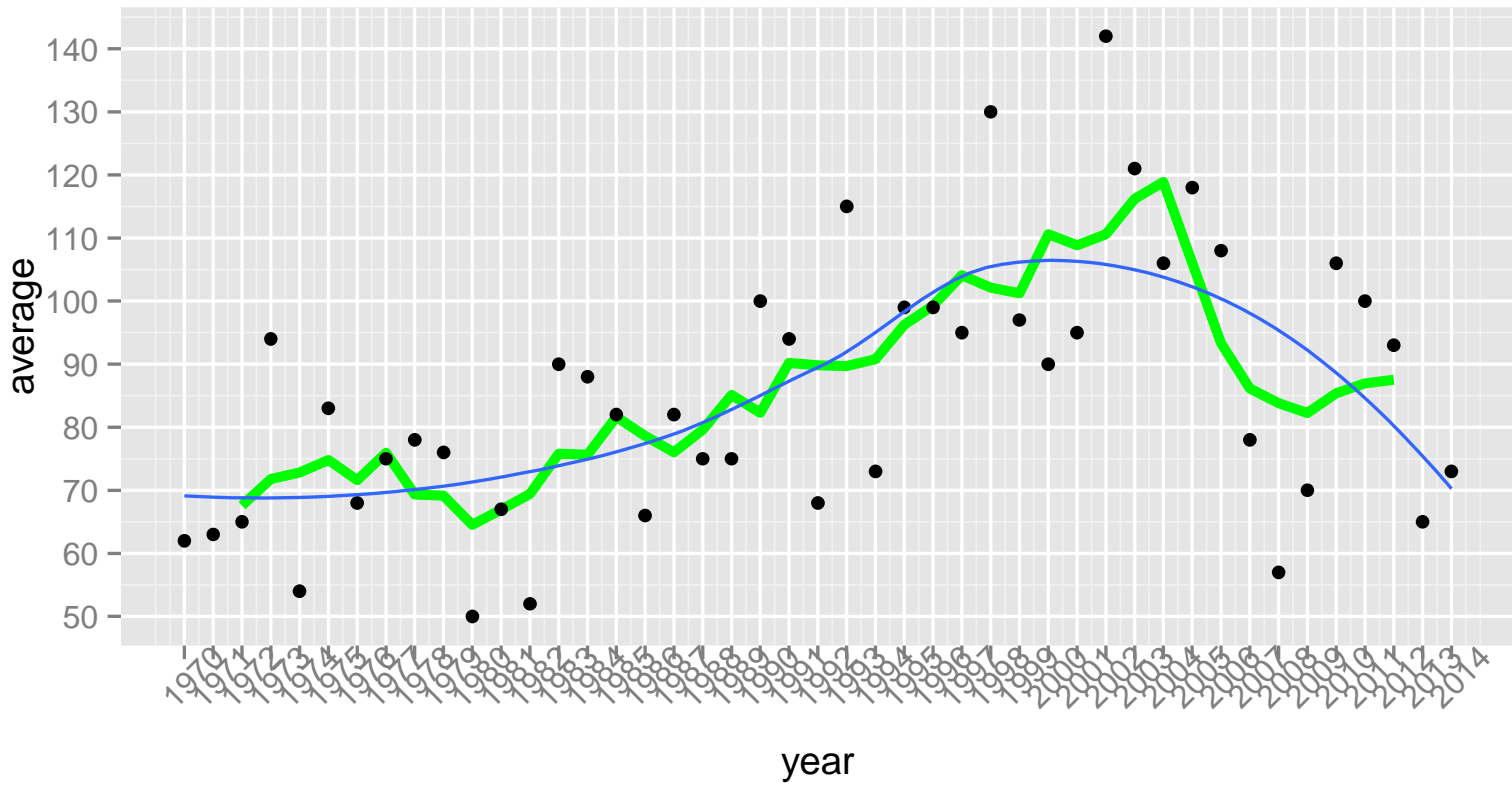
- must turn 'big data' into small data

- Rstudio, R Markdown

- `ggplot2`, `ggvis`, `dplyr`, `tidyr`,

- [cheatsheets](#)

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



Opening Conference and Boot Camp

Organized by Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

```

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))

```

both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as a starting point for 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

Information Visualization

- <http://www.infovis.org>
- a process of transforming information into visual form
- relies on the visual system to perceive and process the information
- <http://ieevis.org/>
- involves the design of visual data representations and interaction techniques

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 learning and visualization, as well as focus themes for applications in the social, physical and life



Highlights

- Sheelagh Carpendale: info-viz

<http://innovis.cpssc.ucalgary.ca/>

- representation
- presentation
- interaction

Example: [Edge Maps](#)

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



Highlights

- [Katy Borner](#): scientific visualization
- advances understanding or provides solutions for real-world problems
- impacts a particular application

PROGRAM

JANUARY 12 - 23, 2015

Opening Conference and Boot-Camp

Organizing committee: Katy Borner, Ken-ichi Kawarabayashi, Bin Yu

- <http://scimaps.org/>

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

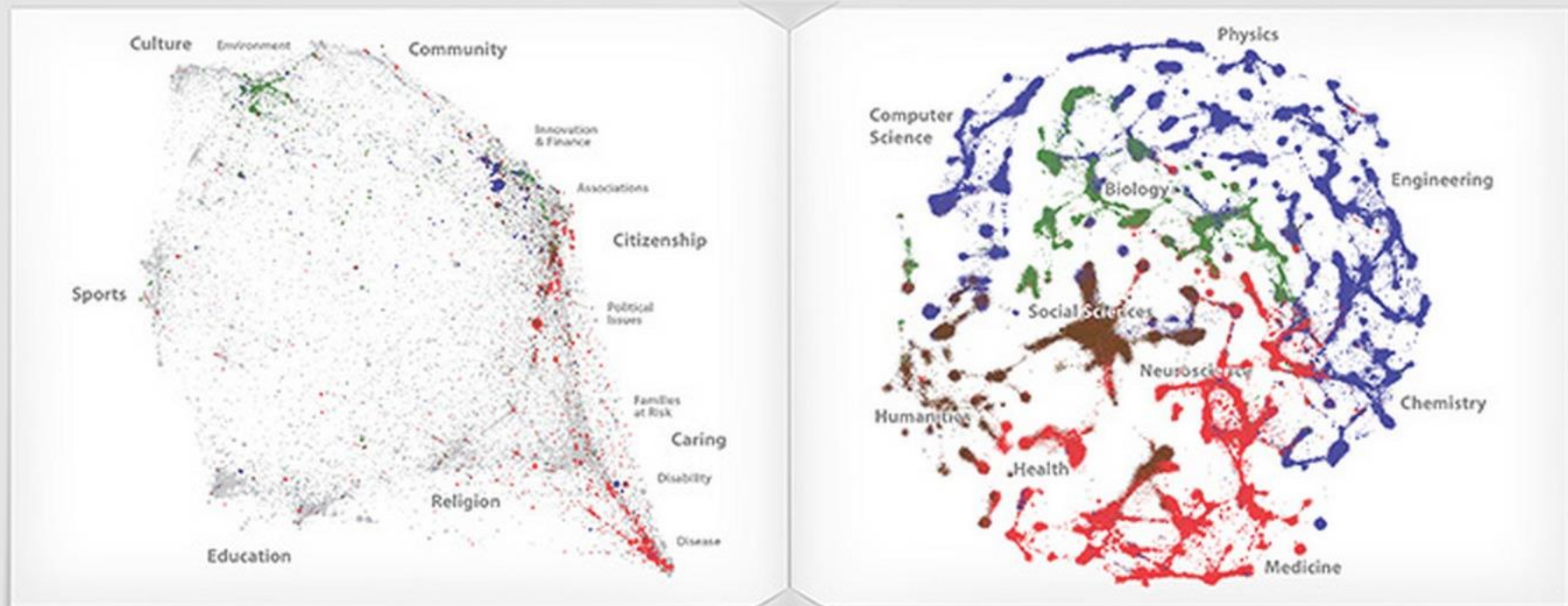
Workshop on Optimization and Matrix Methods in Big Data

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

Exploring the Relationships Between a Map of Altruism and a Map of Science

How is altruism related to science? Altruism is about individual selfless intentions. Science is about discovery and problem solving. On the surface these two facets of society may seem unrelated. In reality they may be strongly linked. Altruistic missions explain historical (and may predict future) patterns of scientific investments. The map of altruism (left) represents altruistic missions, and displays the relative positions of nearly 100,000 non-profit organizations (NPOs) in the United States based on mission-related text from their websites. This map of altruism reveals the issues that we care most about as a society: Culture, Sports, Education, Religion, Community, Citizenship, and Caring. The map of science (right) represents decades of funded research in the natural and medical sciences, engineering, technology, social sciences and humanities. It displays over 43,000,000 documents that are grouped together using a combination of citation and textual similarity.

These two maps are shown side-by-side to illustrate how the altruistic intentions of a society correlate with where we focus our discovery and problem solving efforts. The map of science has been divided into four major areas, shown in four different colors. NPOs whose National Taxonomy of Exempt Entities (NTEE) codes indicate that they explicitly fund scientific activities in these four areas are correspondingly colored in the map of altruism. Altruistic missions associated with these four areas are considered in more detail below, along with projections of how altruistic missions not currently associated with funding of scientific research might benefit from such funding in the future.



Citizenship is linked to Physics, Chemistry, Engineering and Computer Science. The specific aspect of Citizenship active here is the belief that funding should be provided for entrepreneurship and innovation so that the economy can flourish. The funding of science-based innovation from governments and NPOs is reasonably mature and is expected to remain high.



Caring is the basis for funding medical research. The aspects of Caring vary, and include curing of disease, providing opportunities for the disabled, and the treatment of mental health issues. A scientific understanding of these issues has been well funded by individuals, e.g. through donations to NPOs; and through government funding, e.g. the National Institutes of Health.



All Seven Aspects of Altruism are potentially important for childhood development. Scientific research related to this topic is currently focused on social issues, e.g. risk factors, and Education. The altruism map raises an interesting question: is this the right balance, or should more scientific attention be paid to childhood development in other areas, such as Culture, Community, Sports, and Citizenship? Time will tell.

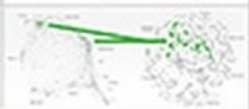
Historical

Citizenship is a major factor in the funding of the Social Sciences. The specific aspect of Citizenship active here is aligned with the belief that rational analysis and the scientific method can contribute to the resolution of political issues. "Think tanks" are examples of non-profit organizations that are funded from this altruistic motive.

Culture and Citizenship contribute to the funding of environmental research. Culture supports that aspect of environmental research that is more concerned with the preservation of our planet for the future enjoyment of our children. Citizenship supports the research focusing on innovative solutions and political tradeoffs which arise from the toxic consequences of current practices.

Future

Community is an important altruistic mission that represents a potential funding opportunity. We know very little about how different communities (geographical, professional, social, etc.) have evolved in terms of providing altruistic services to their members. There are lessons to be learned from how communities variously emphasize Culture, Sports, Education, Religion, Care, or Civic responsibility.





Highlights

- Alex Gonçalves: Visualization for the masses

- to build communion

- for social change

- powerful stories

- “duty of

beauty” <http://www.nytimes.com/newsgraphics/2014/02/14/fashion-week-editors-picks/>

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, and throughout the program will highlight learning and visualization, as well as focus themes for applications in the social, physical and life



Big Data for Health Policy

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

JANUARY 12 - 23, 2015

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

- Privacy

JANUARY 26 - 30, 2015

Workshop on Big Data and Statistical Machine Learning

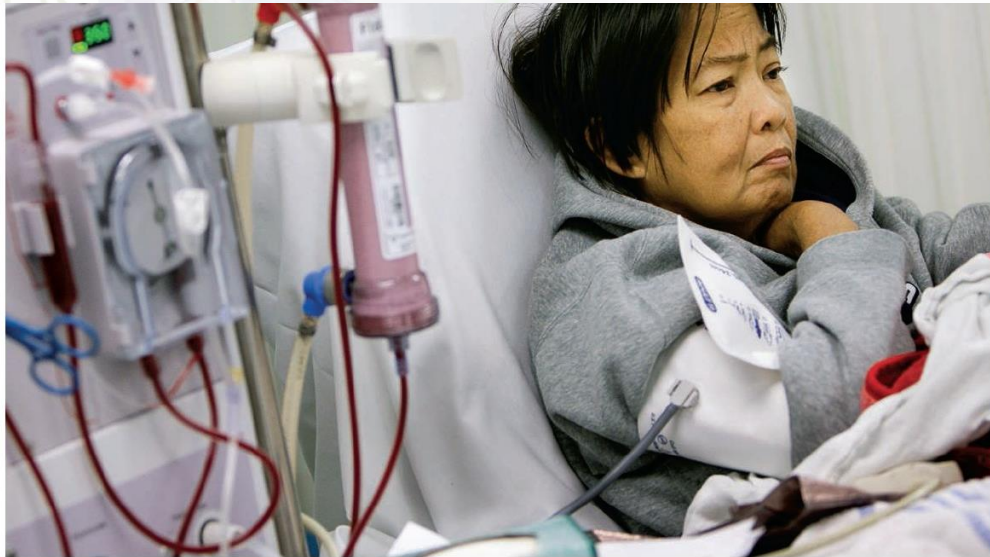
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

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

Heagerty – Pragmatic Clinical Trials



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

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

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

2

X

O

X

O

X

O

X

O

O

X

O

X

O

X

O

X

Heagerty – Pragmatic Clinical Trials

Stepped Wedge Design

					Time	
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>		
0	X	X	X	X		
0	0	X	X	X		
0	0	0	X	X		
0	0	0	0	X		

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

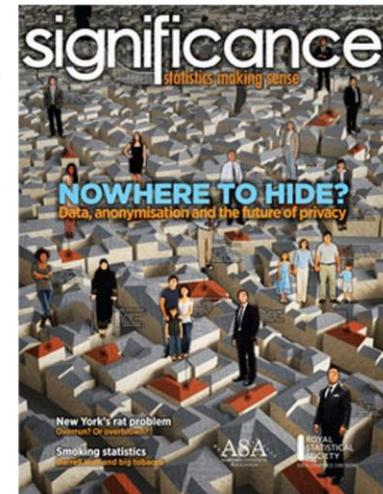
Cultural movements

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.



Worksh
Organizin
Hugh Chi
Worksh

**Carnegie
Mellon
University**

Journal of Privacy and Confidentiality

[Home](#) [About](#) [FAQ](#) [Policies](#) [My Account](#)

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”](#)

PROGRAM

- multi-party communication (Andrew Lo, MIT)
- statistical disclosure limitation and differential privacy
 - Slavkovic, A. -- Differentially Private Exponential Random Graph Models and Synthetic Networks

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

Workshop on Big Data and Statistical Machine Learning

Organizing Committee: Sallie Keller, Bin Yu

Hugh Chipman

FEBRUARY 9 - 13, 2015

Workshop on Optimization and Matrix Methods in Big Data

This 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, and the opening conference, lectures and workshops will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life

- Statistical Disclosure Limitation
 - released data is typically counts, or magnitudes, cross-classified by various characteristics – gender, age, region, ...
 - an item is sensitive if its publication allows estimation of another value of the entity too precisely
 - rules designed to prohibit release of data in cells at ‘too much’ risk, and prohibit release of data in other cells to prevent reconstruction of sensitive items – Cell Suppression

PROGRAM

- computer science -- privacy-preserving data-mining; multi-party computation, differential privacy
- theoretical work on differential privacy has yielded solutions for function approximation, statistical analysis, data-mining, and sanitized databases
- it remains to see how these theoretical results might influence the practices of government agencies and private enterprise

What did we learn?

1. Statistical models are 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 Cycle July 2014



falling-99183_640 →

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

JANUARY 12 - 23, 2015

Opening Conference and Boot Camp

Organizing Committee: Nance Reid (Chair), Sellic Keller, Lisa Lix, Bin Yu

- Big Insights build on old ideas

JANUARY 26 - 30, 2015

Workshop on Planning of Studies, Learning, and Inference

Organizing committee: Kuslan Suresh, Shantanu Das, Bin Yu, Yoshua Bengio,

Hugh Chipman, Bin Yu

- planning of studies, bias, variance, inference

FEBRUARY 9 - 13, 2015

Workshop on Optimization and Machine Learning Methods in Big Data

- Big Data is a Big Opportunity

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 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

A few resources

- Franke, Plante et al. (2015). Statistical inference, learning and models in big data. *preprint: account of the opening workshop*

- [Talks from the closing workshop](#)
for the Big Data program

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

- data science programs: Beijing, Johns

Hopkins, Berkeley, Columbia, NYU, Dalhousie

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, with an overview lectures and background preparation. Workshops will provide an opportunity for in-depth research, teaching and visualization, as well as focus themes for applications in the social, physical and life

A haphazard web walk

Science article (Khoury & Ioannidis) Big Data Meets Public Health

<http://www.sciencemag.org.myaccess.library.utoronto.ca/content/346/6213/1054.full>

Science Article (Ruths & Pfeffer) Social media for large studies of behaviour

<http://www.sciencemag.org.myaccess.library.utoronto.ca/content/346/6213/1063.full>

McGill Newsroom re above Science article:

<http://www.mcgill.ca/newsroom/channels/news/social-media-data-pose-pitfalls-studying-behaviour-240450>

A haphazard web walk

Economist Data Viz

<http://www.economist.com/blogs/graphicdetail/2014/12/new-data-visualisations?fsrc=scn%2Ftw%2Fdc%2F&%3Ffsrc%3Dscn%2F=tw%2Fdc> Interactive visualizations (dec 2)

Flowing Data — books <http://flowingdata.com/data-points/>

Best Viz 2014 http://flowingdata.com/2014/12/19/the-best-data-visualization-projects-of-2014-2/?utm_source=dlvr.it&utm_medium=twitter

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will cover the program, including 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

A haphazard web walk

Big data Music Industry

<http://venturebeat.com/2014/12/18/how-big-data-can-change-the-music-industry/>

The problem with big data <http://www.scmagazine.com/the-problem-with-big-data/article/388691/>

Open models <http://radar.oreilly.com/2014/11/we-need-open-models-not-just-open-data.html>

Workshop on Big Data and Statistical Machine Learning

Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

JANUARY 12 - 23, 2015

Workshop on Optimization and Matrix Methods in Big Data

This thematic program emphasizes both applied and theoretical aspects of machine learning and models. This four-day program and models 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