Statistical Inference, Learning and Models for Big Data

Nancy Reid
University of Toronto

P. R. Krishnaiah
1932-1987

P.R. Krishnaiah Memorial Lecture
2015 Rao Prize Conference
Penn State University
May 15, 2015

P. R. Krishnaiah
1932 – 1987

- 1960 – 1963: Remington Rand Univac
- 1963 – 1976: Wright-Patterson Air Force Base
- 1976 – 1988: University of Pittsburgh
  - founded Center for Multivariate Analysis
  - joint professor in the Graduate School of Business
  - 6 international symposia on Multivariate Analysis
  - founder and editor of *JMVA*
  - founder and editor of *Developments in Statistics*
  - founder and editor of *Handbook of Statistics*

M.M.Rao (1988) Journal of Multivariate Analysis

P. R. Krishnaiah

P. R. **Krishnaiah**
1932-1987

PUBLICATIONS

Books Edited

1. (1966). *Multivariate Analysis.* Academic Press, New York.
2. (1969). *Multivariate Analysis-II.* Academic Press, New York.
3. (1973). *Multivariate Analysis-III.* Academic Press, New York.
4. (1977). *Multivariate Analysis-IV.* North-Holland, Amsterdam.
5. (1979). *Multivariate Analysis-V.* North-Holland, Amsterdam.
6. (1985). *Multivariate Analysis-VI.* North-Holland, Amsterdam.
7. (1977). *Applications of Statistics.* North-Holland, Amsterdam.
8. (1978). *Developments in Statistics, Vol.* 1. Academic Press, New York.
9. (1979). *Developments in Statistics, Vol.* 2. Academic Press, New York.
10. (1980). *Developments in Statistics, Vol.* 3. Academic Press, New York.
11. (1983). *Developments in Statistics, Vol.* 4. Academic Press, New York.
12. (with KALLIANPUR, G., AND GHOSH, J. K.) (1981). *Statistics and Probability: Essays in Honor of C. R. Rao*, North-Holland, Amsterdam.
13. (1980). *Handbook of Statistics, Vol.* 1: *Analysis of Variance.* North-Holland, Amsterdam.
14. (with KANAL, L.) (1982). *Handbook of Statistics, Vol.* 2: *Classification, Pattern Recognition and Reduction of Dimensionality.* North-Holland, Amsterdam.
15. (with BRILLINGER, D. R.) (1983). *Handbook of Statistics, Vol.* 3: *Time Series in Frequency Domain.* North-Holland, Amsterdam.
16. (with SEN, P. K.) (1984). *Handbook of Statistics, Vol.* 4: *Nonparametric Methods.* North-Holland, Amsterdam.
17. (with HANNAN, E. J., AND RAO, M. M.) (1985). *Handbook of Statistics, Vol.* 5: *Time Series in the Time Domain.* North-Holland, Amsterdam.
18. (with RAO, C. R.) (1988). *Handbook of Statistics, Vol.* 6: *Sampling.* North-Holland, Amsterdam.
19. (with RAO, C. R.) (1988). *Handbook of Statistics, Vol.* 7: *Quality Control and Reliability.* North-Holland, Amsterdam.
20. (with SCHUURMANN, F. J.) (In press). *Computations of Complex Multivariate Distributions.* North-Holland, Amsterdam. (In press).
21. *Simultaneous Test Procedures.* Dekker, New York. (In press).

M.M.Rao (1988) Journal of Multivariate Analysis

P. R. Krishnaiah

P. R. Krishnaiah
1932-1987

1976

a. (with CHANG, T. C., AND LEE, J. C.) On the distribution of the likelihood ratio test statistic for compound symmetry. *South African Statist. J.* **10** 49–62.

b. Some recent developments on complex multivariate distributions. *J. Multivariate Anal.* **6** 1–30.

c. (with LEE, J. C., AND CHANG, T. C.) Approximations to the distributions of the likelihood ratio statistics for tests of certain covariance structures of complex muiltivariate normal populations. *Biometrika* **63** 543–549.

d. (with KIM, D. G., DUBRO, G. A., PHADIA, E., AND SCHUURMANN, F. J.) The measurement of flow velocity by transit time techniques. Unpublished.

1977

a. (with CHANG, T. C., AND LEE, J. C.) Approximations to the distributions of the likelihood ratio statistics for testing the hypotheses on covriance matrices and mean vectors simultaneously. In *Applications of Statistics* (P. R. Krishnaiah, Ed.), pp. 97–103. North-Holland, Amsterdam.

b. (with KHATRI, C. G. AND SEN, P. K.) A note on the joint distribution of correlated quadratic forms. *J. Statist. Planning Inference* **1** 299–307.

c. (with LEE, J. C.) Infeence on the eigenvalues of the covariance matrices of real and complex multivariate normal populations. In *Multivariate Analysis-IV* (P. R. Krishnaiah, Ed.), pp. 95–103. North-Holland, Amsterdam.

d. On generalized multivariate Gamma type distributions and their applications in reliability. In *Proceedings, Conference on the Theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods, Vol.* 1 (I. N. Shimi and C. P. Tsokos, Eds.), pp. 475–494. Academic Press, New York.

e. (with LEE, J. C., AND CHANG, T. C.) Approximations to the distributions of the likelihood ratio statistics for testing certain structures on the covariance matrices of real multivariate normal populations. In *Multivariate Analysis-IV* (P. R. Krishnaiah, Ed.), pp. 105–118. North-Holland, Amsterdam.

f. (with LEE, J. C., AND CHANG, J. C.) Approximations to the distributions of the determinants of real and complex multivariate Beta matrices. *South African Statist. J.* **11** 13–26.

M.M.Rao (1988) Journal of Multivariate Analysis

P. R. Krishnaiah

Technical Report No. 86–38, Center for Multivariate Analysis, University of Pittsburgh. *IEEE Trans.*, in press.

u. (with TAUXE, W. N., KLEIN, H. A., BAGCHI, A., KUNDU, D., AND TEPE, P.) *Clinical Evaluation of the Filtration Fraction: A Multivariate Statistical Analysis.* Technical Report No. 86–41, Center for Multivariate Analysis, University of Pittsburgh.

v. (with NISHI, R., AND BAI, Z. D.) *Strong Consistency of Certain Information Theoretic Criteria for Model Selection in Calibration, Discriminant Analysis and Canonical Correlation Analysis.* Technical Report No. 86–42, Center for Multivariate Analysis, University of Pittsburgh.

w. (with MIAO, B. Q., AND ZHAO, L. C.) *On Detection of Change Points using Mean Vector.* Technical Report No. 86–47, Center for Multivariate Analysis, University of Pittsburgh. In *Handbook of Statistics*, Vol. 7 (P. R. Krishnaiah and C. R. Rao, Eds.). Academic Press, New York.

x. (with CHEN, X. R., AND LIANG, W. Q.) *Estimation of Multivariate Binary Density Using Orthonormal Functions.* Technical Report No. 86–48. Center for Multivariate Analysis, University of Pittsburgh.

y. (with CHEN, X. R.) *Test of Linearity in General Regression Models.* Technical Report No. 86–49, Center for Multivariate Analysis, University of Pittsburgh. *J. Multivariate Anal.*, in press.

z. (with CHEN, X. R.) *Estimation and Testing in Truncated and Nontruncated Linear*

M.M.Rao (1988) Journal of Multivariate Analysis

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

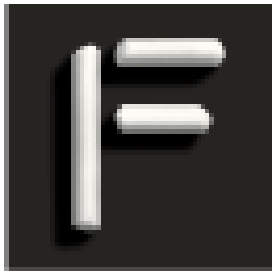
Image Credits: Sheelagh Carpendale & InnoVis

NSERC CRSNG

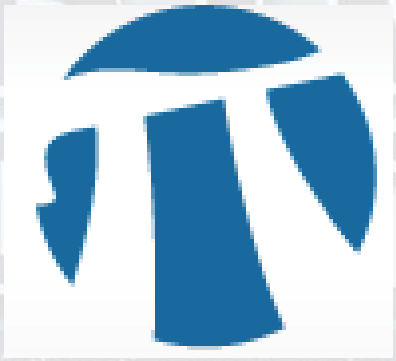
Ontario

CANSSI INCASS

Canadian Institute for Statistical Sciences



FIELDS



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

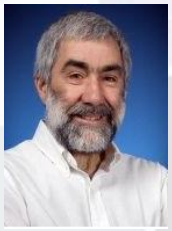
- Networks, Web mining, and Cyber-security — May, CRM
- Statistical Theory for Large-scale Data — April, PIMS
- Challenges in Environmental Science — May, PIMS
- Complex Spatio-temporal Data — April, Fields
- Commercial and Retail Banking — May, Fields

- Closing Conference: Statistical and Computational Analytics — June, SSC

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

MDM 12 – Einat Gil et al.

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

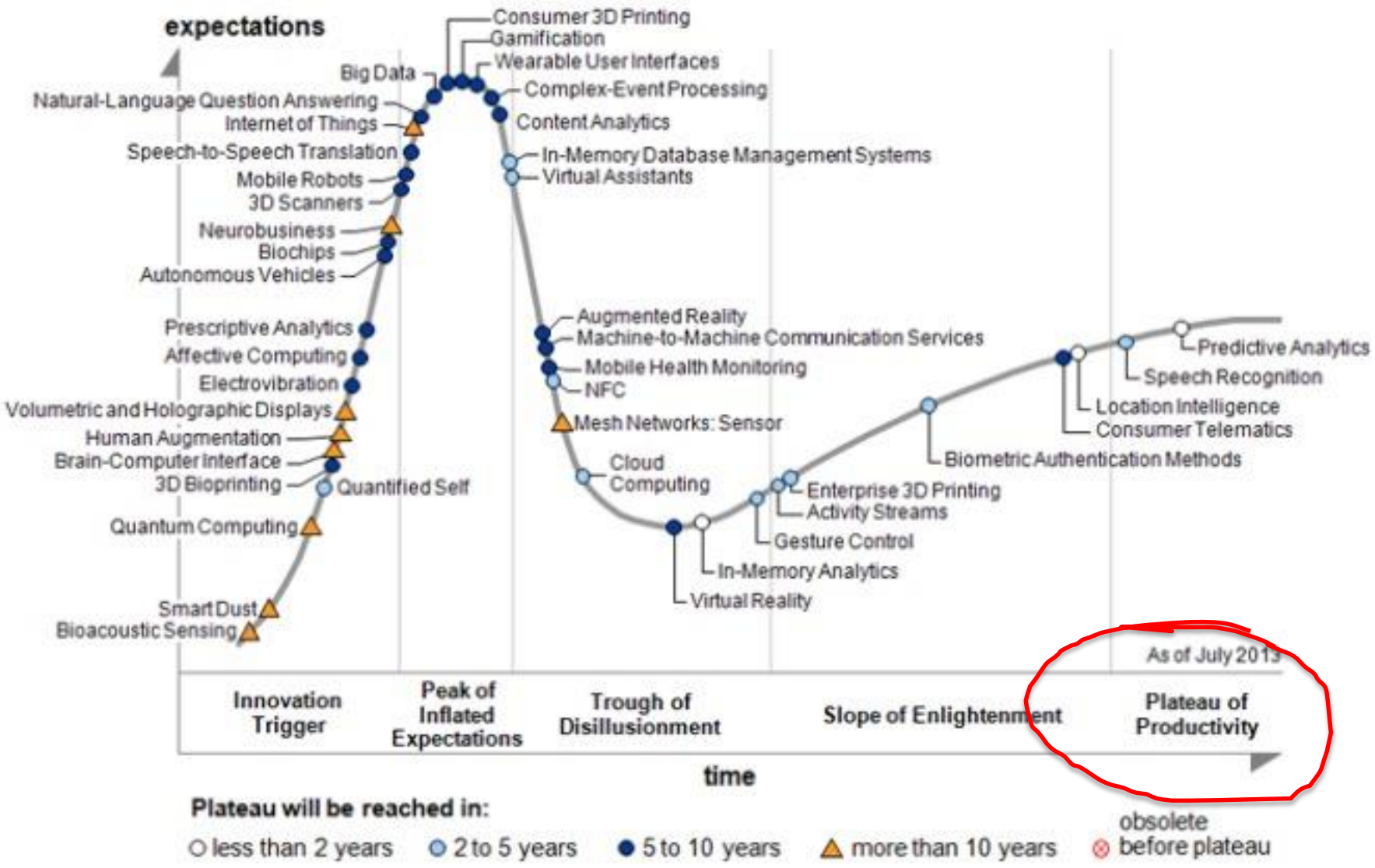
As each new field matured, there came a recognition that in fact much was to be gained by studying connections to statistics

Gartner Hype Cycle
July 2013

Gartner Hype Cyc
July 2014

falling-99183_640 →

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


expectations
Consumer 3D Printing
Gamification
Wearable User Interfaces
Big Data
Complex-Event Processing
Natural-Language Question Answering
Internet of Things
Content Analytics
Speech-to-Speech Translation
In-Memory Database Management Systems
Virtual Assistants
Mobile Robots
3D Scanners
Neurobusiness
Biochips
Autonomous Vehicles
Augmented Reality
Prescriptive Analytics
Machine-to-Machine Communication Services
Affective Computing
Mobile Health Monitoring
Electrovibration
NFC
Volumetric and Holographic Displays
Mesh Networks: Sensor
Human Augmentation
Brain-Computer Interface
Cloud Computing
3D Bioprinting
Quantified Self
Enterprise 3D Printing
Activity Streams
Quantum Computing
Gesture Control
In-Memory Analytics
Virtual Reality
Smart Dust
Bioacoustic Sensing
Predictive Analytics
Speech Recognition
Location Intelligence
Consumer Telematics
Biometric Authentication Methods
As of July 2013
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

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

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

"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

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

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
- Franke, Plante, et al. (2015): “A data analytic perspective on Big Data”, forthcoming

Big Data and Statistical Machine Learning

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

Statistical Machine Learning

- Grosse, R. (2015). Scaling up natural gradient by factorizing Fisher information. under review for *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:

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

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

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

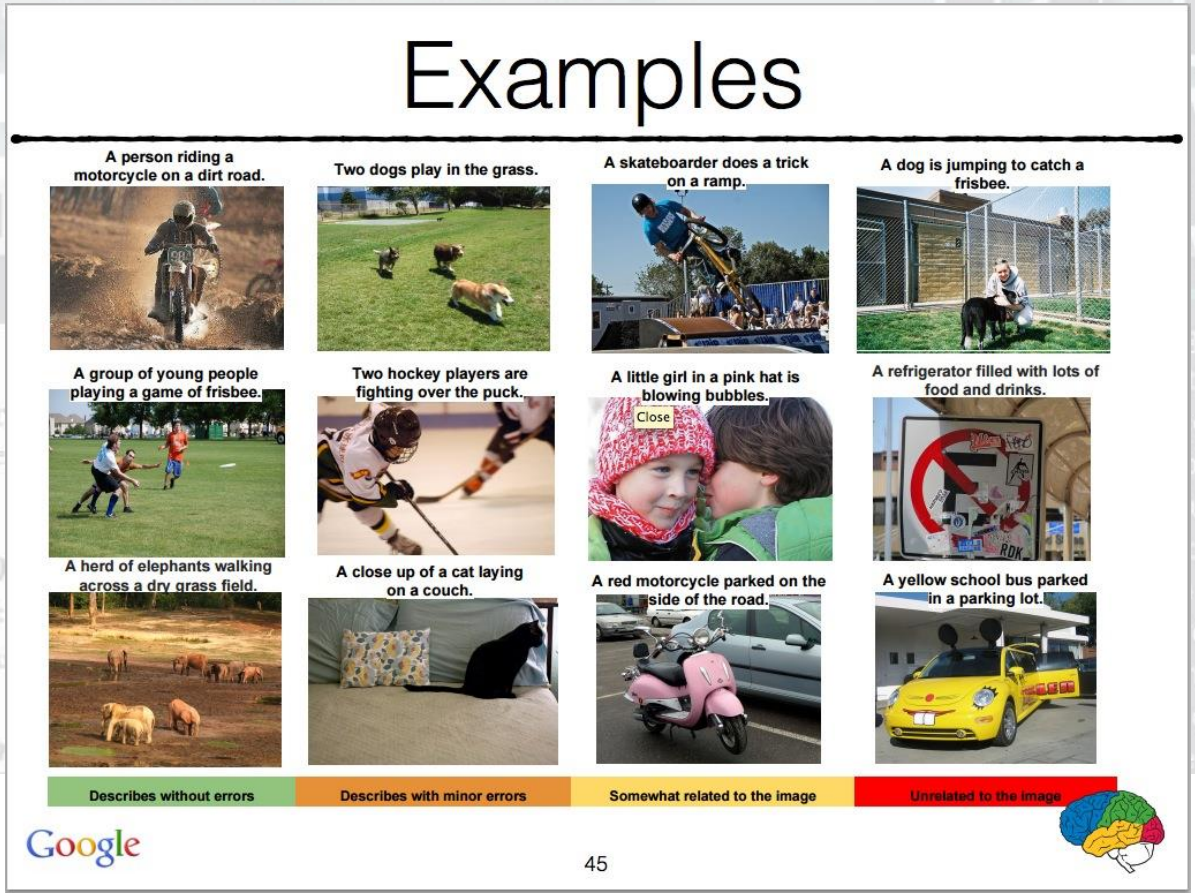
$$\eta \longleftarrow \eta + \epsilon \, i(\eta)^{-1} \nabla_\eta \ell(\eta; v, h)$$

- 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]



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

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

Optimization

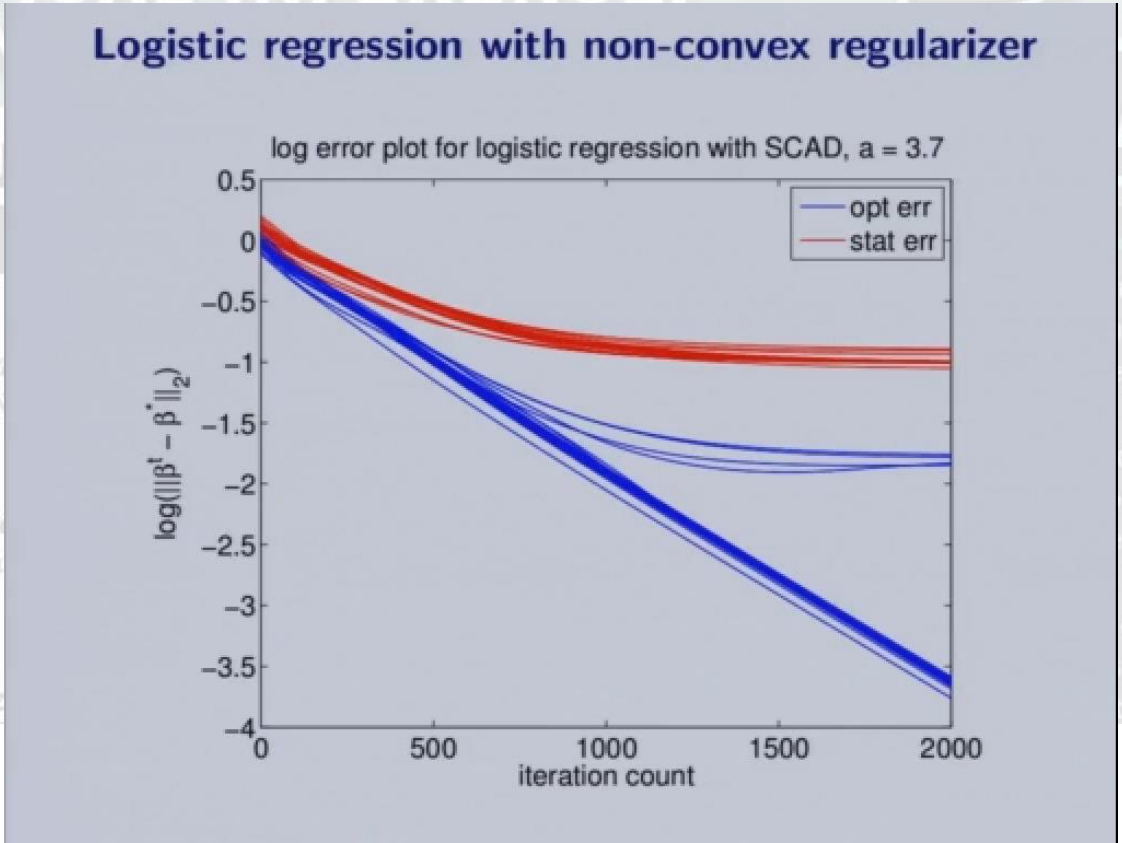
- Wainwright – non-convex optimization
- example: regularized maximum likelihood

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

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

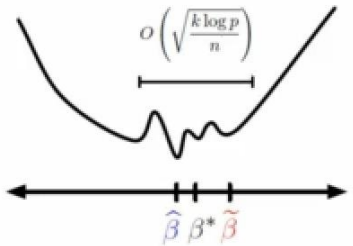
Wainwright and Loh

- 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*
nonconvexity. *J Machine Learning Res.* 16, 559-61

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

Visualization for Big Data Strategies and Principles

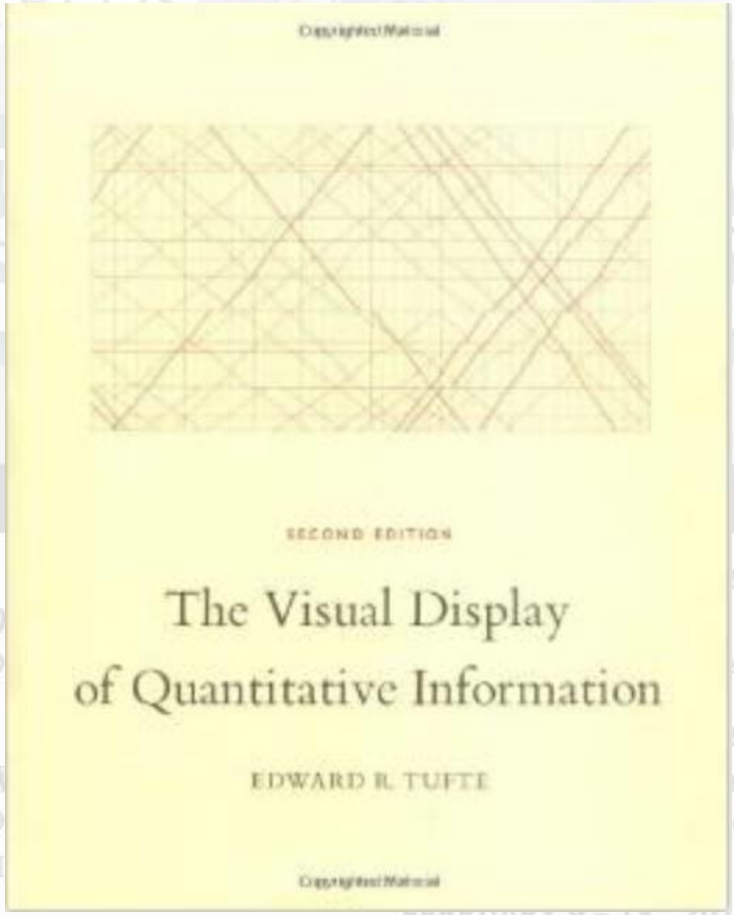
- data representation
- data exploration via filtering, sampling and aggregation
- visualization and cognition
- information visualization
- statistical modeling and software
- cognitive science and design

Visualization for Big Data:
Strategies and Principles

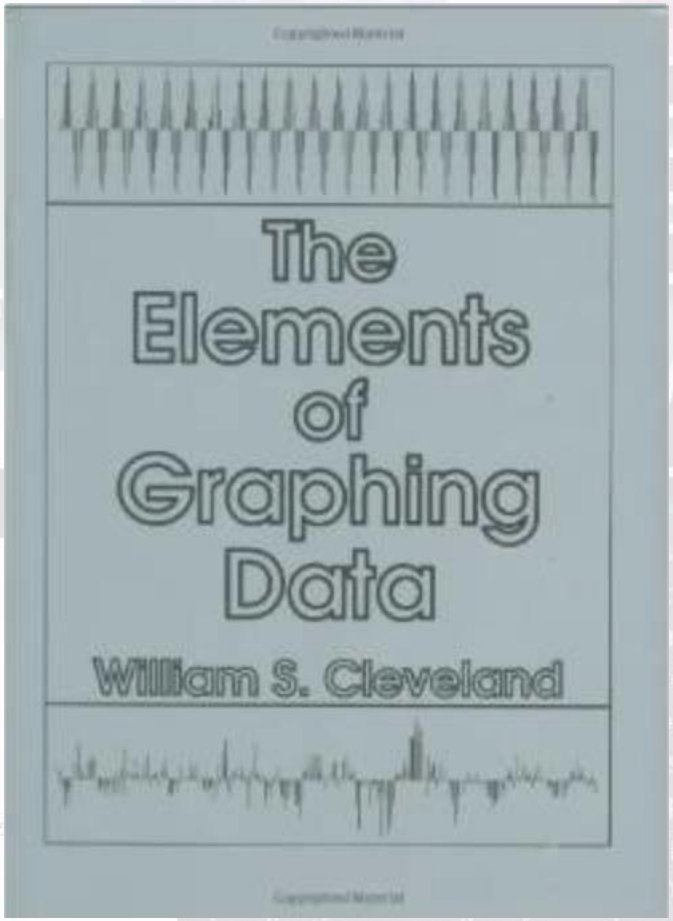
Visualization for Big Data:
Strategies and Principles

In addition to approaches such as search, query processing, and analysis, visualization techniques will also become critical across many stages of big data use--to obtain an initial assessment of data as well as through subsequent stages of scientific discovery.

Visualization for Big Data:
Strategies and Principles

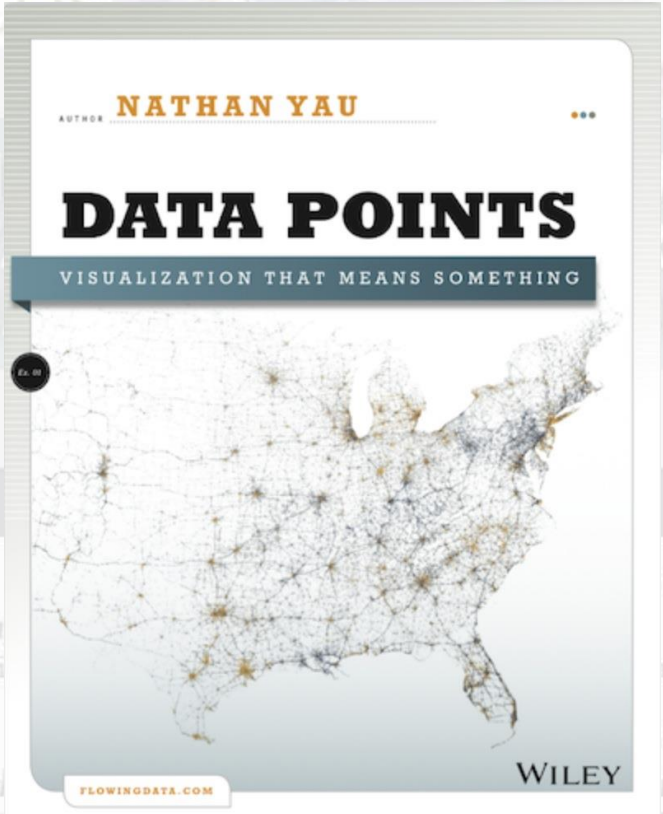


1983

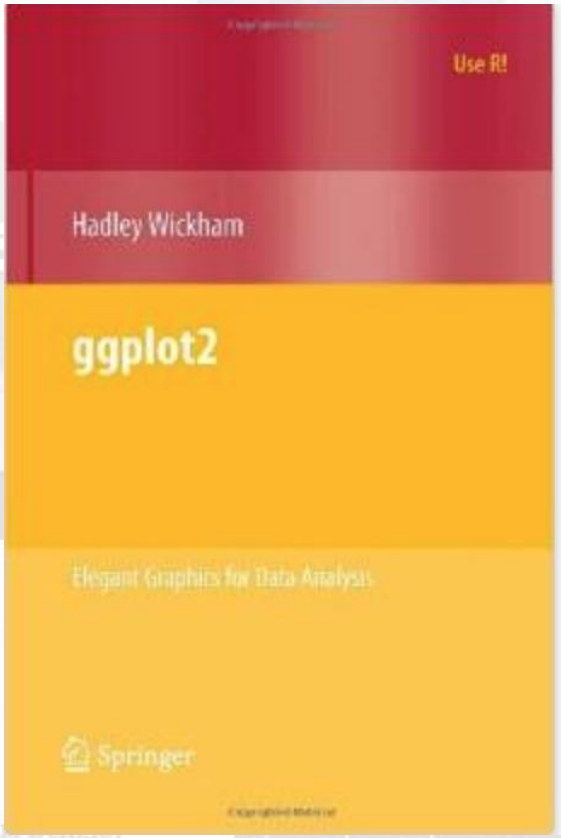


1985

Visualization for Big Data:
Strategies and Principles



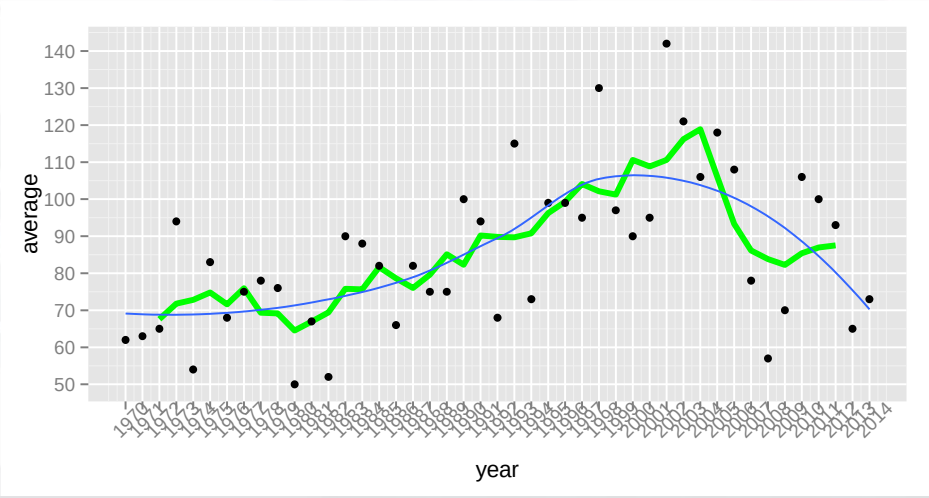
2013



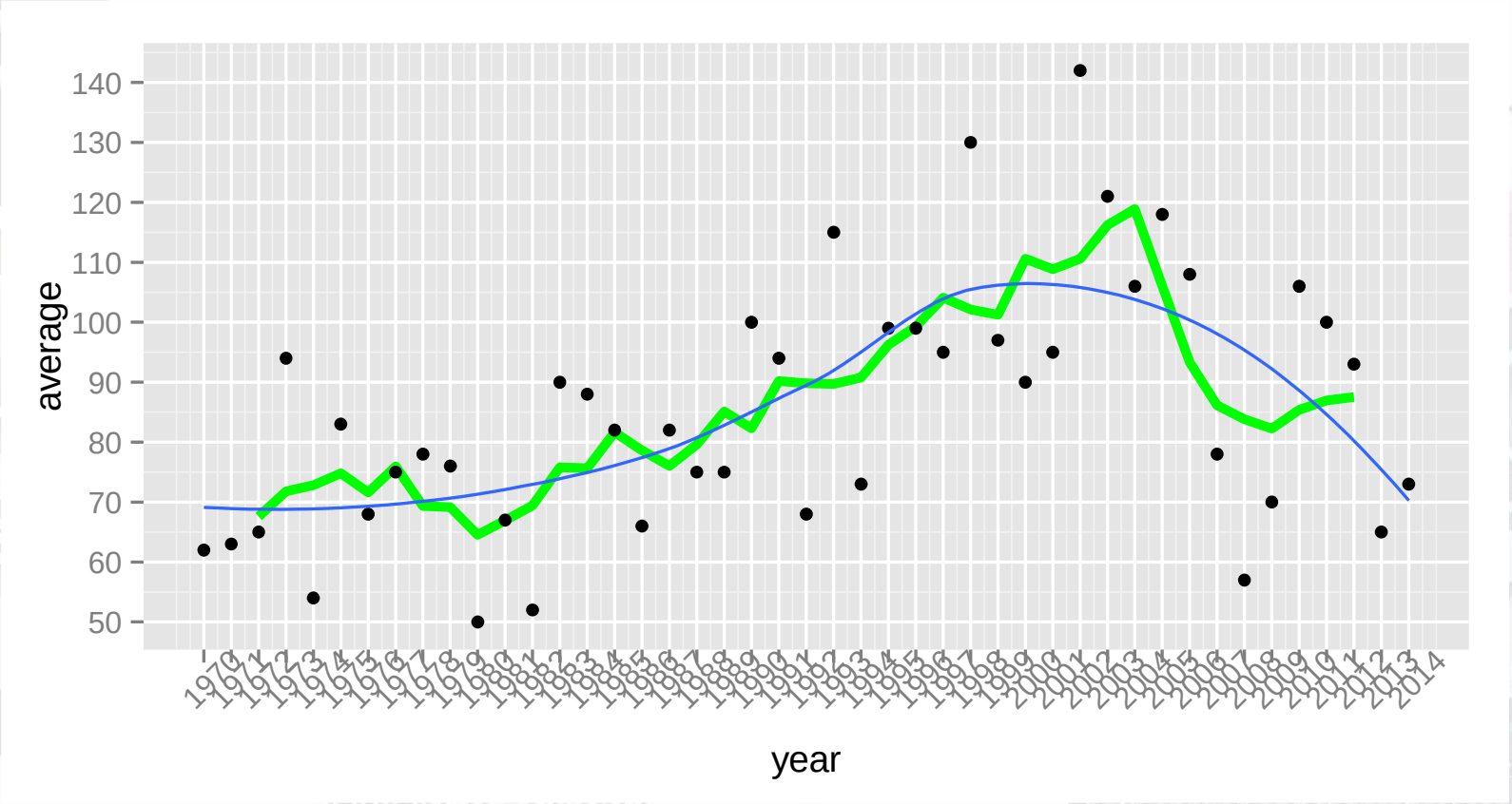
2009

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))
```

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://ieeevis.org/
- involves the design of visual data representations and interaction techniques

Highlights

- Sheelagh Carpendale: info-viz

  http://innovis.cpsc.ucalgary.ca/

- representation
- presentation
- interaction
- Example: Edge Maps web page

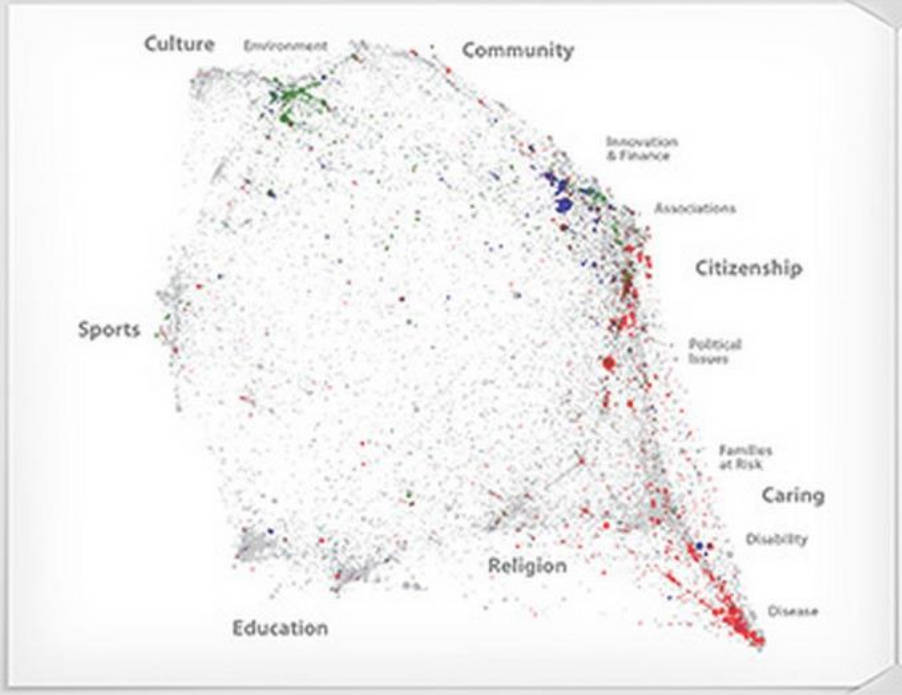
Highlights

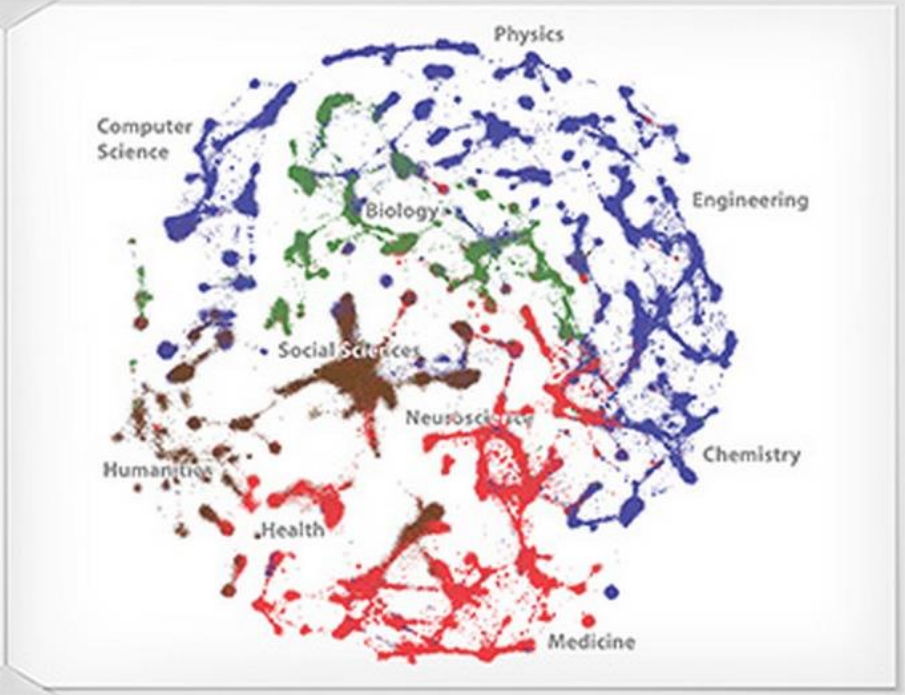
- Katy Borner: scientific visualization
- advances understanding or provides solutions for real-world problems
- impacts a particular application
- http://scimaps.org/

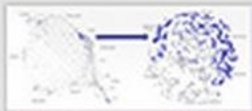
Exploring the Relationships Between
a Map of Altruism and a Map of Science

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.





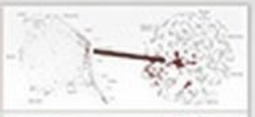


Citizenship is linked to *Physics, Chemistry, Engineering* and *Computer Science*. The specific aspect of *Citizenship* active here is the belief that funding should be provided for entrepreneurship and innovation so that the economy can flourish. The funding of science-based innovation from governments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

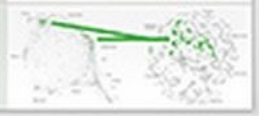
Caring is the basis for funding medical research. The aspects of *Caring* vary, and include curing of disease, providing opportunities for the disabled, and the treatment of mental health issues. A scientific understanding of these issues has been well funded by individuals, e.g., through donations to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

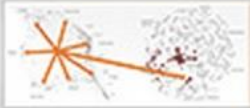
Historical

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.

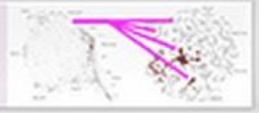




All Seven Aspects of Altruism are potentially important for childhood development. Scientific research related to this topic is currently focused on social issues, e.g., risk factors, and *Education*. The altruism map raises an interesting question: is this the right balance, or should more scientific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

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.



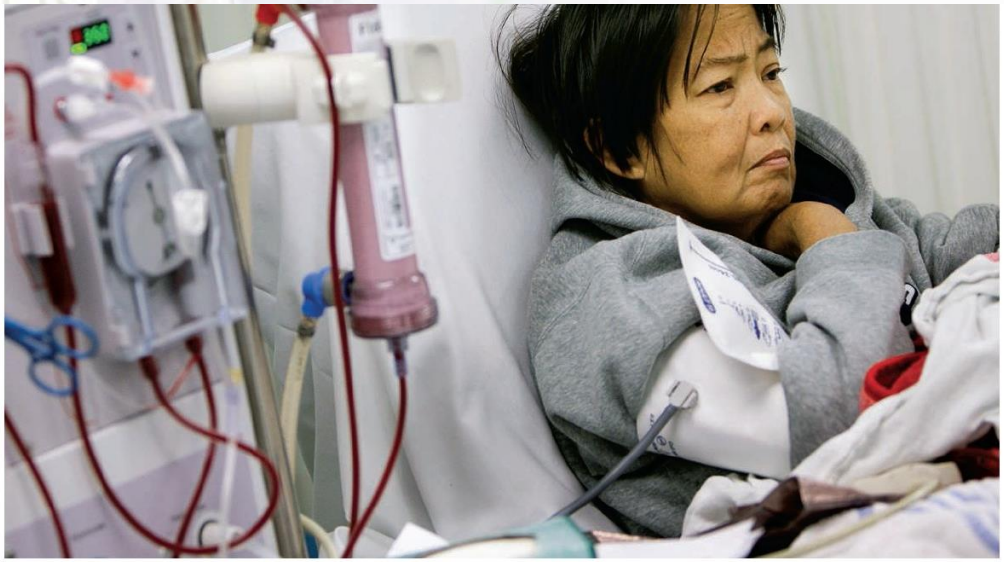
Highlights

- 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/

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.

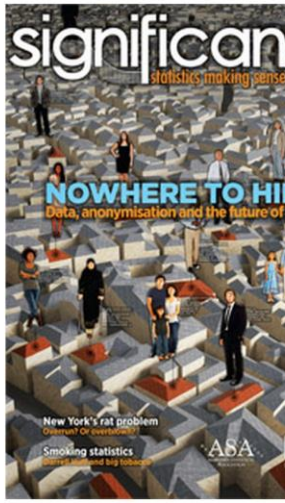
Other challenges come in figuring out the best way to design pragmatic studies.

Big Data for Social Policy

Significance - October 2014 (Volume 11 Issue

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 pre
privacy while sharing valuable data is crucial to the future of our information society.

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"
- multi-party communication (Andrew Lo, MIT)
- statistical disclosure limitation and differential privacy

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

- Statistical Disclosure Limitation
  - released data typically counts or magnitudes stratified by characteristics of the entities to which they apply
  - 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

- computer science -- privacy-preserving data-mining; secure 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 Cyc
July 2014

falling-99183_640 →

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

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

INTERNET ARCHIVE WayBackMachine

http://www.utstat.utoronto.ca/stats/srivasconf/confpage.html Go

1 captures
20 Jun 01 – 20 Jun 01

MAY JUN JUL
2000 20 2001 2002

Close

Help

Conference in Honor of Professor Muni Srivastava

July 11, 12 at the Fields Institute, Toronto

Professor Muni Srivastava will be retiring from the Department of Statistics at the University of Toronto in June, 2001. Professor Srivastava has made important contributions to several areas of statistics, including multivariate statistics, sequential analysis nonparametric inference and statistical quality control.

The Department of Statistics will be holding a conference in his honour on July 11 and 12, 2001 at the Fields Institute at the University of Toronto. The meeting will be held in Room 230 of the Fields Institute. There will be a banquet on the evening of July 11 at the Faculty Club and Professor C.R. Rao will present the after dinner speech.

The conference will feature a distinguished group of invited speakers, including Barry Arnold, Rudy Beran, Martin Bilodeau, Jerry Lawless, Hisao Nagao, Ashish Sen, David Tyler, Dietrich von Rosen, Yanhong Wu and Shelly Zacks. The schedule for the talks can be found on the Conference Schedule.

The conference is being organized by Professors Keith Knight and Nancy Reid

Professor C. R. Rao will present the af

Statisticians should seize the opportunity to lead on Big Data