Big Data

= Big Machines

= Lots of Computing

= Complex Architectures

= Computer Science
Small data

equations and formulas

= mathematical modelling

= a little computing

= Statistical Science

\[ p(v, h; \eta) \propto \frac{1}{Z(\eta)} \exp\{a^T v + b^T h + v^T W h\}, \]
\[ \eta = (a, b, W) \]
Big Data

• Interesting
• Detailed
• Informative
• Fun
Small Data

So yesterday
Small Data
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.
Canadian Institute for Statistical Sciences

Fields Institute for Research in the Mathematical Sciences

Centre de Recherches Mathématiques

Pacific Institute for Mathematical Sciences
Workshops

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

Introduction to topics at following workshops
One day on each topic
Many speakers started by trying to define big data

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

But I know it when I see it …”

Justice Potter Stewart; Jacobellis v. Ohio 22 June 1964

Robert Bell, Google, Plenary Opening Lecture
Statistical Inference, Learning and Models in Big Data

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

\(^1\)University College London, London, UK
\(^2\)HEC Montréal, Montréal, Québec, Canada
\(^3\)Freie Universität, Berlin, Germany
\(^4\)University of Waterloo, Waterloo, Ontario, Canada
\(^5\)University of Toronto, Toronto, Ontario, Canada
\(^6\)Western University, London, Ontario, Canada
\(^7\)University of Witwatersrand, Johannesburg, South Africa
\(^8\)McMaster University, Hamilton, Ontario, Canada

Summary
Some highlights

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

- Statistical Machine Learning

Restricted Boltzmann Machine

bottom nodes $v = (v_1, v_2, \ldots)^T$; top nodes $h = (h_1, h_2, \ldots)^T$
Statistical Machine Learning

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

Restricted Boltzmann Machine

bottom nodes \( v = (v_1, v_2, \ldots)^T \); top nodes \( h = (h_1, h_2, \ldots)^T \)
Restricted Boltzmann machine

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

- natural gradient ascent

\[ \eta \leftarrow \eta + \epsilon \, i(\eta)^{-1} \nabla_\eta \ell(\eta; v, h) \]

- uses Fisher information as metric tensor

\[ \ell = \log f \]

\[ i = \mathbb{E}(-\ell'') \]

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

- Gaussian graphical model approximation to force sparse inverse

Grosse and Salakhutdinov (2016) 32\textsuperscript{nd} Internat. Conf. on Machine Learning
Restricted Boltzmann machine

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

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

Brendan Frey, Infinite Genomes Project

FieldsLive January 27 2015

Leung et al Bioinformatics 2014
Some highlights

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

- Optimization

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

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

- lasso penalty
  $$\mathcal{P}_\lambda(\theta) = \lambda \|\theta\|_1 = \lambda \sum |\theta_j|$$

- $\|\theta\|_1$ is convex relaxation of $\|\theta\|_0$

- many interesting penalties are non-convex

- optimization routines may not find global optimum
Optimization

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

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

Wainwright FieldsLive Jan 16 2015

Loh and Wainwright JMLR 2015
Some highlights

• Statistical Machine Learning

• Optimization

• Visualization

• Health Policy

• Social Policy
Some highlights

- Visualization

innovis.cpsc.ucalgary.ca
Visualization

• statistical graphics
  – data representation
  – data exploration
  – filtering, sampling aggregation

• information visualization

• scientific visualization

• cognitive science and design
Visualization

KPMG Data Observatory, IC
Visualization

KPMG Data Observatory, IC
Visualization

It’s all about the 538 Electoral College votes

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

The winding path to 270 electoral votes

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

How unpopular is Donald Trump?
An updating calculation of the president's approval rating, accounting for each poll's quality, recency, sample size and partisan lean. How this works »

52.4% Disapprove
41.3% Approve
Visualization

How Trump compares with past presidents
- Approval rating
- Disapproval rating
- Not approval

Barack Obama 2009-17

George W. Bush 2001-09

Bill Clinton 1993-2001

George H.W. Bush 1989-93

Ronald Reagan 1981-89

Jimmy Carter 1977-81

Gerald Ford 1974-77

Richard Nixon 1969-74

Lyndon B. Johnson 1963-69

fivethirtyeight.com

guns
Some highlights

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

- Health Policy

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

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

Thérèse Stukel, ICES
Some highlights

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

- Social Policy
ICES Data Repository is globally unique in scope and breadth

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

Thérèse Stukel, ICES
Privacy

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

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

• Statistical Disclosure Limitation
• Differential Privacy
• Multi-party Communication
Some highlights

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

- inference, environmental science, networks, genomics, finance, physical sciences, software infrastructure, …
What did we learn?

• Statistical models for big data are complex, high-dimensional
  – inference is well-studied, but difficult

• Computational challenges include size and speed
  – ideas of statistical inference get lost in the machine

• Data owners understand 2., but not 1.
• Data modellers understand 1., but not 2.

• Data science may be the best way to combine these
That was yesterday

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

Harvard launches data science initiative

Francesca Dominici and David Parkes named co-directors

• Berkeley, Hopkins, CMU, Washington, UBC, Toronto, …
What is data science?

- a course?
- a set of courses?
- a job?
- a technology?
- a new field of research?
- a collaboration?
Data Science Program(s)

- mathematical reasoning
- statistical theory
- statistical and machine learning methods
- programming and software development
- algorithms and data structure
- communication results and limitations
Good Enough Practices in Scientific Computing

Greg Wilson¹,‡*, Jennifer Bryan²,‡, Karen Cranston³,‡, Justin Kitzes⁴,‡, Lex Nederbragt⁵,‡, Tracy K. Teal⁶,‡
¹ Software Carpentry Foundation / gvwilson@software-carpentry.org
² University of British Columbia / jenny@stat.ubc.ca
³ Duke University / karen.cranston@duke.edu
⁴ University of California, Berkeley / jkitzes@berkeley.edu
⁵ University of Oslo / lex.nederbragt@ibv.uio.no
⁶ Data Carpentry / tkteal@datacarpentry.org

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

• Data Management – from raw to ‘analysable’

• Software – programming

• Collaboration

• Project Organization

• Keeping Track

• Writing
Data Science Research

• data collection and data quality

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

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

• ‘new’ types of data: networks, graphs, text, images, …
  – “alternative sources”
… Data Science Research

• collaboration and communication

• data wrangling, database development, record linkage, privacy

• replicability, reproducibility, new workflows

• visualization

• outside the ivory tower -- industry, government, media, public
Tripods (Transdisc Research in Princ…)

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

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

I. Data-Centric, Exploratory Methods

I. Efficient Algorithms

II. Graph Approaches

III. Model Fitting and Regularization

IV. Ensemble Methods

V. Causal Inference

VI. Targeted Learning
The push back

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

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

How big data threatens democracy and increases inequality
The push back

Big data in social sciences: a promise betrayed?

Posted on March 22, 2017

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

“Big data is neither easier nor faster nor cheaper”

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

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

“Before I write my name on the board, I’ll need to know how you’re planning to use that data.”
Facial recognition database used by FBI is out of control, House committee hears

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

March 27
The push back

RSS 2014 Significance Lecture - The Big Data trap
Big data: are we making a big mistake?

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

“Big data” has arrived, but big insights have not
“A range of other problems”

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

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

accurate answers quickly; meaningful error bars; merge various data sources; visualize and present conclusions; diagnostics; non-stationarity; targeted experiments within databases
Caution can be a good thing

"Digital Hippocratic Oath"
Caution can be a good thing

“…from data we will get the cure for cancer as well as better hospitals;
schools that adapt to children’s needs making them happier and smarter;
better policing and safer homes;
and of course jobs.”
Big Data

As of July 2014

Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity

Plateau will be reached in:
- less than 2 years
- 2 to 5 years
- 5 to 10 years
- more than 10 years
- obsolete before plateau
Thank You!