Topics in inference for big data

- Course web page
- Course information
- Topics covered
  - Adapted to weekly workshops as offered
- Today (NR)
  - background on the organization of the program
  - a short look ahead
Development of a Thematic Program

- Spring 2013 – Fields Institute approached CANSSI
- June, 2013 – CANSSI director established steering committee
- July 2013 – First conference call
- Sep 2013 – Letter of Intent submitted to Fields
- March 2014 – Full proposal submitted to Fields
- May 2014 – Funding approved
- June 2014 – Invitations, PDFs, participants
A National Program

• March 2014 – Workshop proposal to CRM

• Sep 2014 – Two workshop proposals to PIMS
Big Data – Big Topic

• Where to start?
• Look up some references

- Likelihood 78 m
- Statistical inference 7m
Big Data – Big Hype?

- Expectations
- Peak of Inflated Expectations
- Trough of Disillusionment
- Slope of Enlightenment
- Plateau of Productivity

- Innovation Trigger
- Plateau will be reached in:
  - less than 2 years
  - 2 to 5 years
  - 5 to 10 years
  - more than 10 years
  - obsolete before plateau
The Blogosphere

12 Big Data Definitions: What's Yours?

Gil Press
Contributor

+ Comment Now  + Follow Comments

PROGRAM

JANUARY 12 – 23, 2015
Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie K.

JANUARY 26 – 31, 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
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 WHAT IT CAN ACTUALLY DO.”)
The problem with Big Data

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 in processing. And that’s just the start. We’ve barely scratched the surface in real-time data and the analysis of it.
Stop Hyping Big Data and Start Paying Attention to ‘Long Data’

BY SAMUEL ARBESMAN  01.29.13  |  9:30 AM  |  PERMALINK
The Blogosphere

7 Tips to Succeed with Big Data

GET THE GUIDE

SEE A PREVIEW

Organizing Committee: Kailash Babu (Indian Institute of Science), Dhairya Das (University of California, Berkeley), Hugh Chipman, Bin Yu

February 9 – 13, 2015

Workshop on Optimization and Matrix Methods in Big Data
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.
Statistical Inference, Learning and Models for Big Data

- 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
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 – OHIP records, clinical trials, public health surveys

Jordan 06/2014
The Atlas experiment – CERN

http://atlas.ch/what_is_atlas.html#5
If all the data from ATLAS were recorded, this would fill 100,000 CDs per second. This would create a stack of CDs 450 feet high every second, which would reach to the moon and back twice each year. The data rate is also equivalent to 50 billion telephone calls at the same time. ATLAS actually only records a fraction of the data (those that may show signs of new physics) and that rate is equivalent to 27 CDs per minute. [http://atlas.ch/what_is_atlas.html](http://atlas.ch/what_is_atlas.html)
The Square Kilometre Array (SKA) project is an international effort to build the world’s largest radio telescope, with a square kilometre (one million square metres) of collecting area.

World leading scientists and engineers designing and developing a system which will require supercomputers faster than any in existence in 2013, and network technology that will generate more data traffic than the entire Internet.
Exploration: 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
About the Study

The Ontario Health Study is one of the largest long-term health studies in Canada. Since 2010 almost 225,000 Ontarians have taken a short online survey to help researchers better understand the causes of chronic diseases, such as diabetes, heart disease, and cancer.

Learn more about the study and how you can participate by visiting the website at https://www.ontariohealthstudy.ca/en/home.
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 Structures**

Volume: Scale of Data
- 40 Zettabytes (42 Trillion Gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
- 6 billion people have cell phones.
- Most companies in the U.S. have at least 100 terabytes (100,000 gigabytes) of data stored.
- World population: 7 billion.

Velocity: Analysis of Streaming Data
- The New York Stock Exchange captures 1 TB of trade information during each trading session.
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
- By 2016, it is projected there will be 18.9 billion network connections — almost 2.5 connections per person on earth.

Variety: Different Forms of Data
- It's estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day.
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes).
- By 2014, it's anticipated there will be 420 million wearable, wireless health monitors.
- 4 billion hours of video are watched on YouTube each month.
- 400 million tweets are sent per day by about 200 million monthly active users.

Veracity: Uncertainty of Data
- 1 in 3 business leaders don't trust the information they use to make decisions.
- 27% of respondents in one survey were unsure of how much of their data was inaccurate.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, NIDCD, GAO
A Warning about Big Data

Big data has been enjoying a lot of hype, with promises it will help deliver everything from increased corporate profits to better healthcare. While the potential is certainly there, Tim Harford asks if the...
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 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...
Warnings

- Google `flu trends
  - Google’s engineers weren’t trying to figure out what cause what
  - A theory-free analysis of mere correlations is inevitably fragile
- Paul et al (2014): Twitter Improves Influenza Forecasting
- According to the Pew Research Internet Project, in 2013, US-based Twitter users were disproportionately young, urban or suburban, and black
  - New, large, cheap data sets and powerful analytical tools will pay dividends – nobody doubts that.
  - “Big data” has arrived, but big insights have not.
... warnings

Is the Privacy of Network Data an Oxymoron?

Stephen E. Fienberg*

1 Introduction

While social networks are now a part of everyday life for the vast majority of people using computers, smartphones, and tablets, privacy is but an afterthought. Google+ has in excess of 100 million users a month while Facebook has topped 1 billion. Other more specialized networks such as Linked-in add to the fray. But from a privacy perspective the biggest concern for users should be the efforts to integrate the networking apps into all other forms of online activity as well as the constant effort to link additional data to network information, in addition to the network owners’ efforts to market that information to third party vendors. Further, Facebook and other networking sites have
... warnings

• **Did Big Data kill the statistician?**
  - ... why do we think Big Data is more than just a new name for a collection of old ideas, and why do we think that data science is forward looking and statistics is just dealing with the past? Why do we lend more credibility to rebranding than to historical fact?
  - A good statistician will understand that “not everything that counts can be counted, and not everything that can be counted counts”. A quote which is variously attributed to either Albert Einstein or William Bruce Cameron...
Statistical Inference, Learning and Models for Big Data

- 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

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.
Important but omitted

- Business analytics
- Financial data
- Physical sciences
- Statistical genetics
- Genomics, proteomics, ...
- Hardware and software
Important but often forgotten

- Design of experiments
- Principles of sampling
- Sources of variation
- Opportunities for bias
- Simplicity
- Modelling
- Understanding
- Clarity
## WEEK ONE: JANUARY 12 - 16

### Monday January 12: Introductory Lectures and Overview

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00</td>
<td>Coffee and Registration</td>
</tr>
<tr>
<td>8:45</td>
<td>Welcome</td>
</tr>
<tr>
<td>9:00-9:30</td>
<td>Nancy Reid, University of Toronto</td>
</tr>
<tr>
<td>9:30-10:30</td>
<td><strong>Keynote Lecture:</strong> Bob Bell, AT&amp;T Labs - Research</td>
</tr>
<tr>
<td></td>
<td><em>Big Data: It's Not the Data</em></td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Coffee</td>
</tr>
<tr>
<td>11:00-12:00</td>
<td>Adam Kalai, Microsoft</td>
</tr>
<tr>
<td></td>
<td><em>Machine learning and crowdsourcing</em></td>
</tr>
<tr>
<td>12:00-2:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>2:00-3:00</td>
<td>Hugh Chipman, Acadia University</td>
</tr>
<tr>
<td></td>
<td><em>An overview of Statistical Learning</em></td>
</tr>
<tr>
<td>3:00-3:30</td>
<td>Tea</td>
</tr>
<tr>
<td>3:30-4:30</td>
<td>Yulia Gel, University of Waterloo</td>
</tr>
<tr>
<td></td>
<td><em>The Role of Modern Social Media Data in Surveillance and Prediction of Infectious Diseases: from Time Series to Networks</em></td>
</tr>
<tr>
<td>4:30</td>
<td>Cash Bar Reception</td>
</tr>
</tbody>
</table>

### Tuesday January 13: Introductory Lectures and Overview

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30-10:30</td>
<td><strong>Keynote Lecture:</strong> Emmanuel Candes, Stanford University</td>
</tr>
<tr>
<td></td>
<td><em>Big Data and the Reproducibility of Scientific Research: What Can Statistics Offer</em></td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Coffee break</td>
</tr>
<tr>
<td>11:00-12:00</td>
<td>Steve Scott, Google Inc</td>
</tr>
<tr>
<td></td>
<td><em>Bayes and Big Data: The Consensus Monte Carlo Algorithm</em></td>
</tr>
<tr>
<td>12:00-1:30</td>
<td>Lunch break</td>
</tr>
</tbody>
</table>
A haphazard web walk

Quantathon from Waterfont International [http://quantathon.ca](http://quantathon.ca)


Science article (Khoury & Ioannidis) Big Data Meets Public Health [http://www.sciencemag.org.myaccess.library.utoronto.ca/content/346/6213/1054.full](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](http://www.sciencemag.org.myaccess.library.utoronto.ca/content/346/6213/1063.full)
A haphazard web walk


Freakonometrics blog http://freakonometrics.hypotheses.org

A haphazard web walk


A haphazard web walk

Big data Music Industry http://venturebeat.com/2014/12/18/how-big-data-can-change-the-music-industry/


Katy Borner’s exhibit http://scimaps.org

stats blogs http://blogs.ams.org/blogonmathblogs/2015/01/05/return-of-the-statistics-blogs/#sthash.7VVUU14i.9D4iOVSI.dpws

UCL Jan meeting http://www.ucl.ac.uk/bigdata-theory/schedule/