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

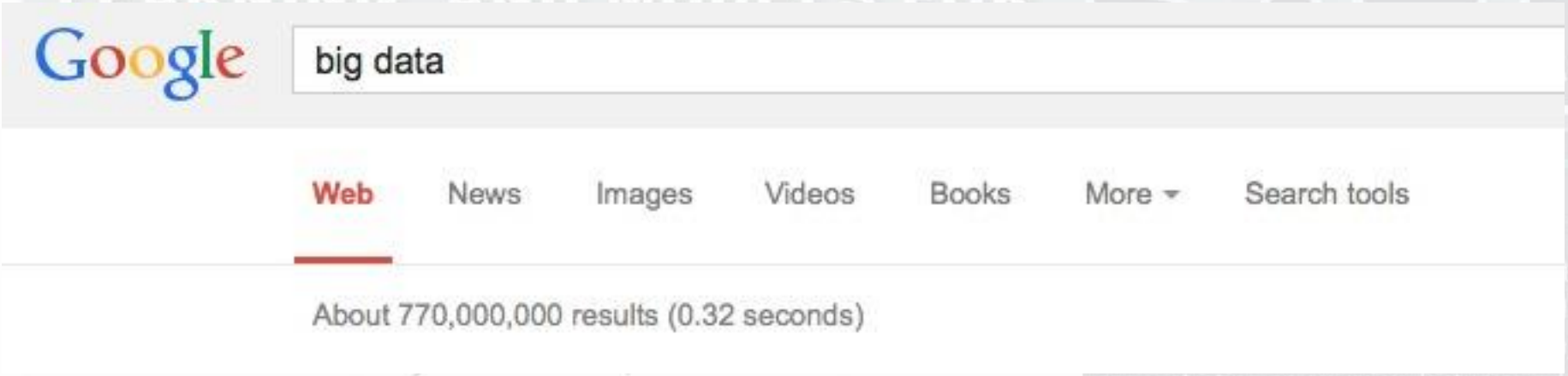
CENTRE DE RECHERCHES MATHÉMATIQUES

- Sep 2014 – Two workshop proposals to PIMS

PIMS

Big Data – Big Topic

- Where to start?
- Look up some references



- Likelihood 78 m
- Statistical inference 7m


expectations
Consumer 3D Printing
Gamification
Wearable User Interfaces
Big Data
Complex-Event Processing
Natural-Language Question Answering
Content Analytics
Internet of Things
Speech-to-Speech Translation
In-Memory Database Management Systems
Virtual Assistants
Mobile Robots
3D Scanners
Neurobusiness
Biochips
Autonomous Vehicles
Prescriptive Analytics
Augmented Reality
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
Quantum Computing
Smart Dust
Bioacoustic Sensing
Predictive Analytics
Speech Recognition
Location Intelligence
Consumer Telematics
Biometric Authentication Methods
Enterprise 3D Printing
Activity Streams
Gesture Control
In-Memory Analytics
Virtual Reality
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

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

Forbes

PROGRAM

JANUARY 12 – 23, 2015

Opening Conference and Boot Camp

Organizing Committee: Nancy Reid (Chair), Sallie K

JANUARY 26 – 3

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

The Blogosphere

MARCH 29, 2013

STEAMROLLED BY BIG DATA

BY GARY MARCUS

THE NEW YORKER

Share Tweet g+1

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

The Blogosphere

Kim Crawley, researcher, InfoSec Institute

December 22, 2014

The problem with Big Data

Share this article:

SC MAGAZINE
FOR IT SECURITY PROFESSIONALS

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

WIRED

OPINION | big data | Business | memes | Science

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

BY SAMUEL ARBESMAN 01.29.13 | 9:30 AM | PERMALINK

Share 15 Tweet 6 g+1 156 in Share 26 Pin it

The Blogosphere

Gartner_2014-2

By MIKE WHEATLEY | Published August 18, 2014 | Full size is 540 × 337 pixels

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

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

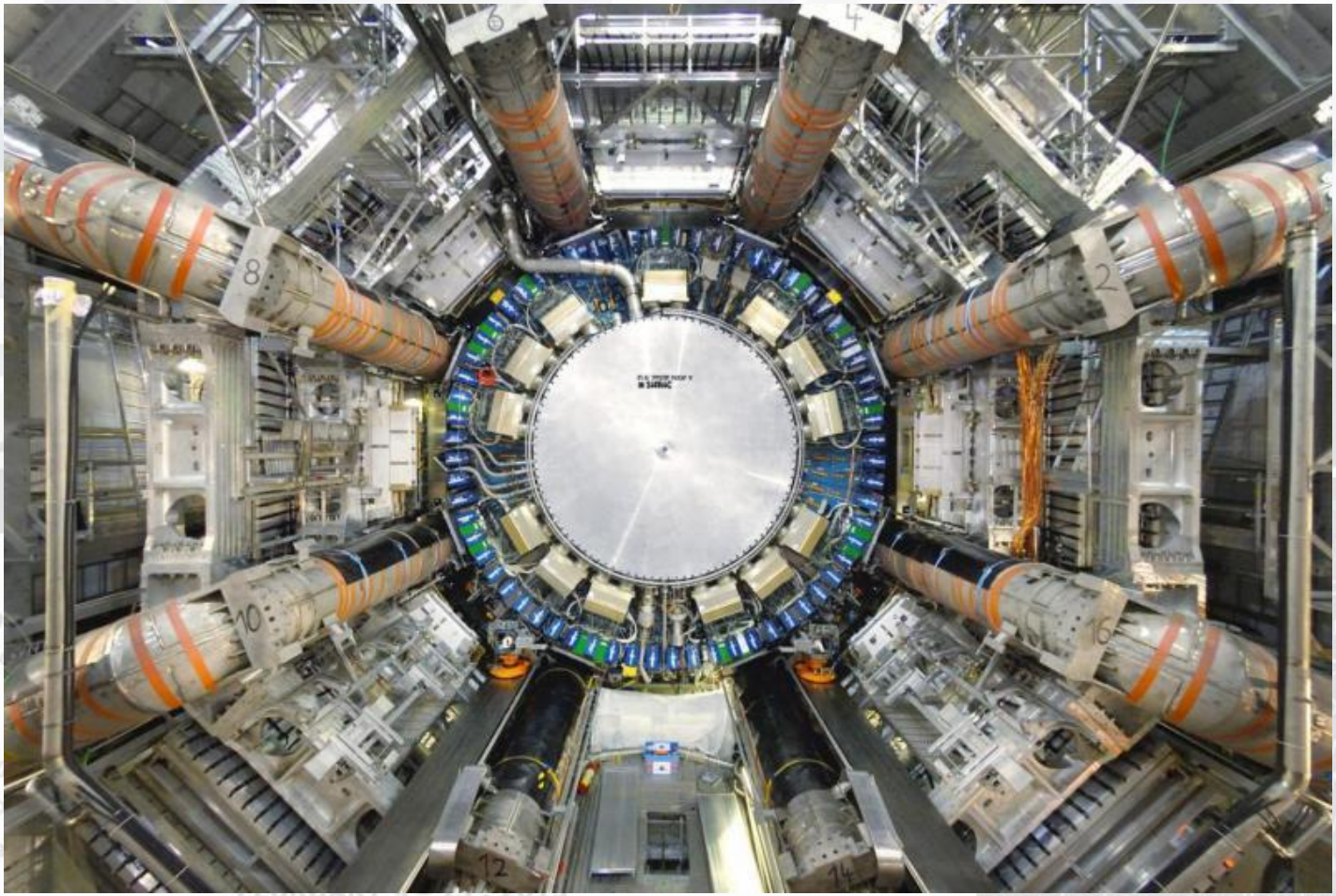
Big Data Types

- 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

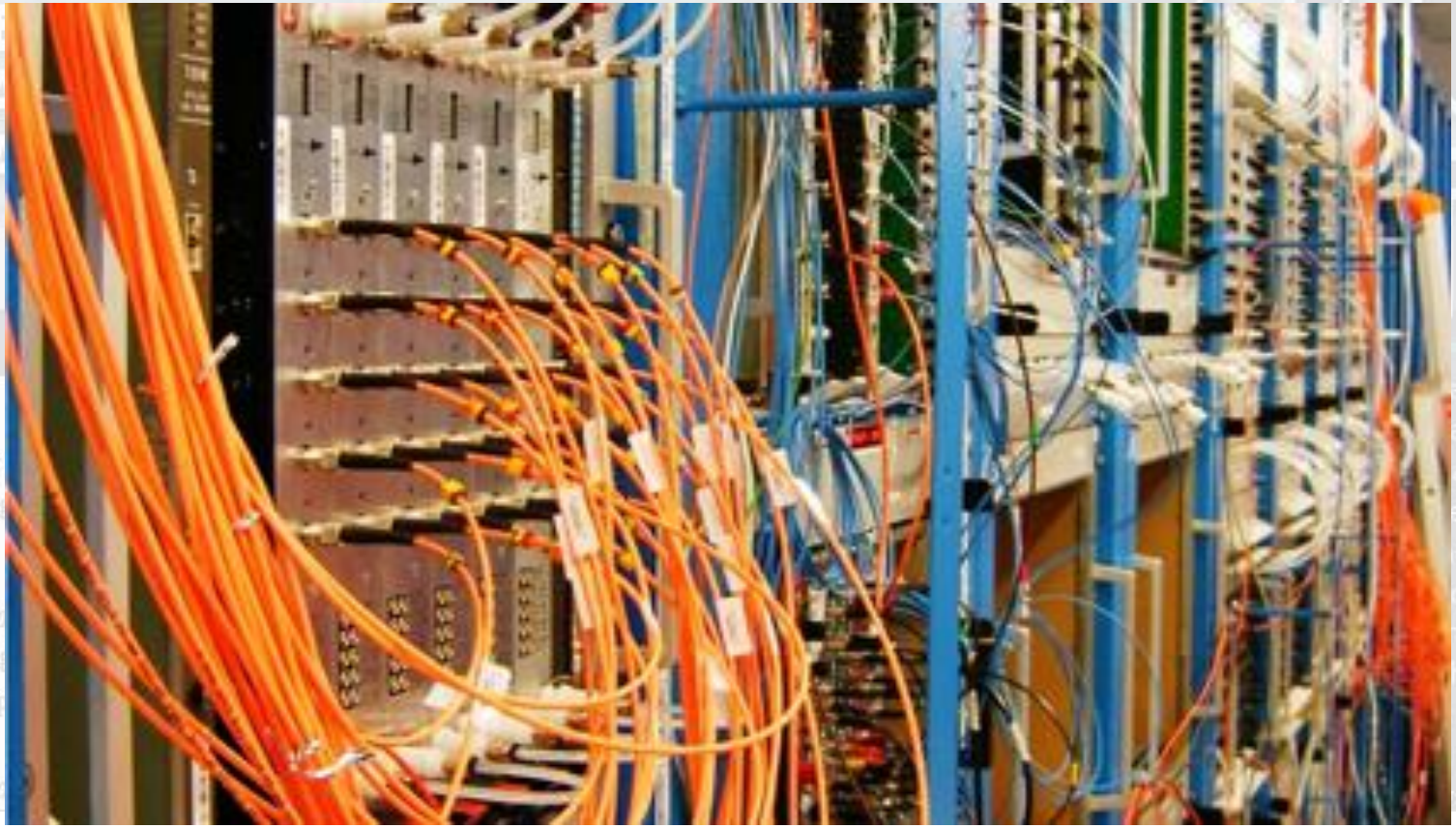
Jordan 06/2014

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 - 5

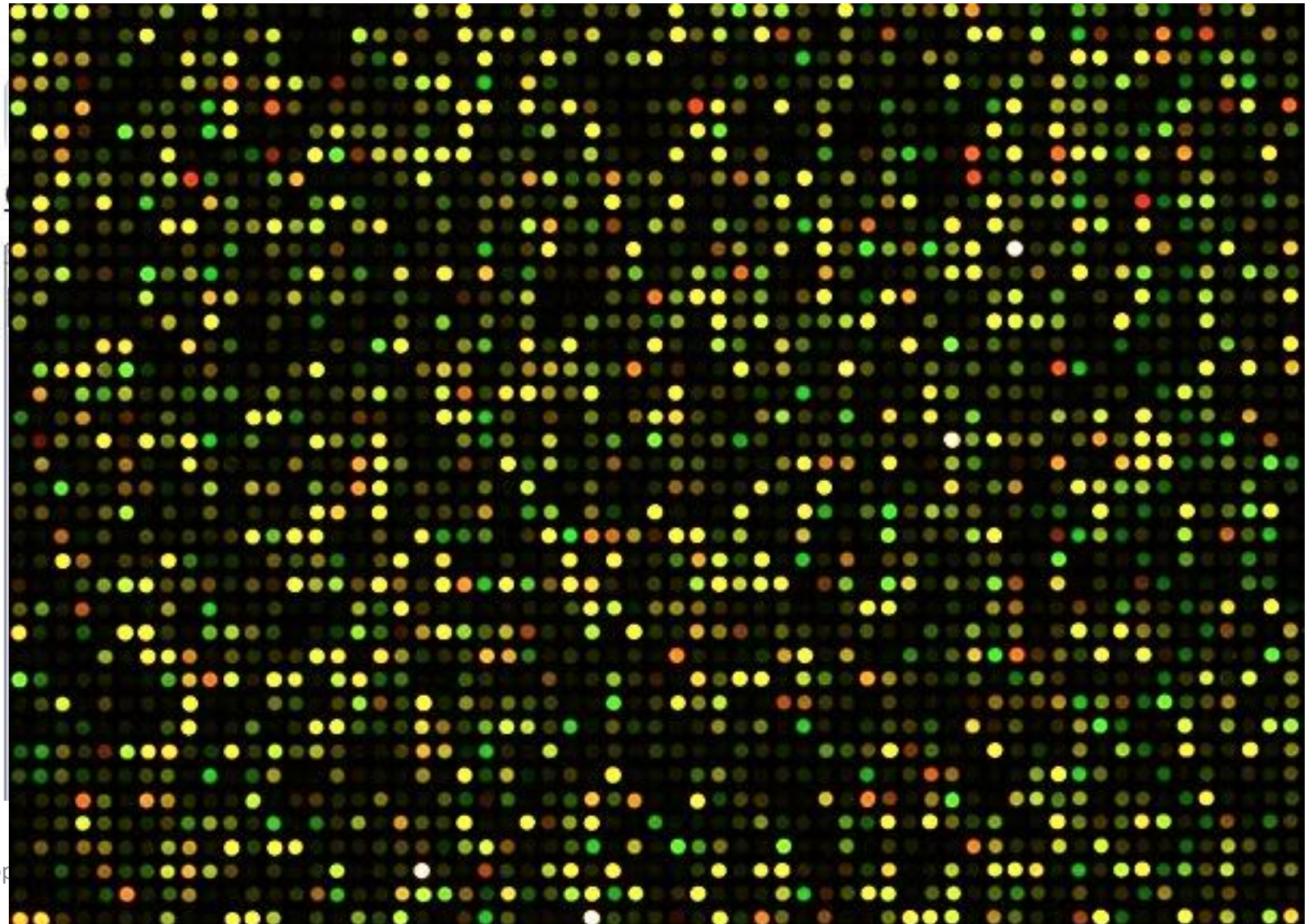


Exploration: the Square Km Array

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

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

Exploration: microarray

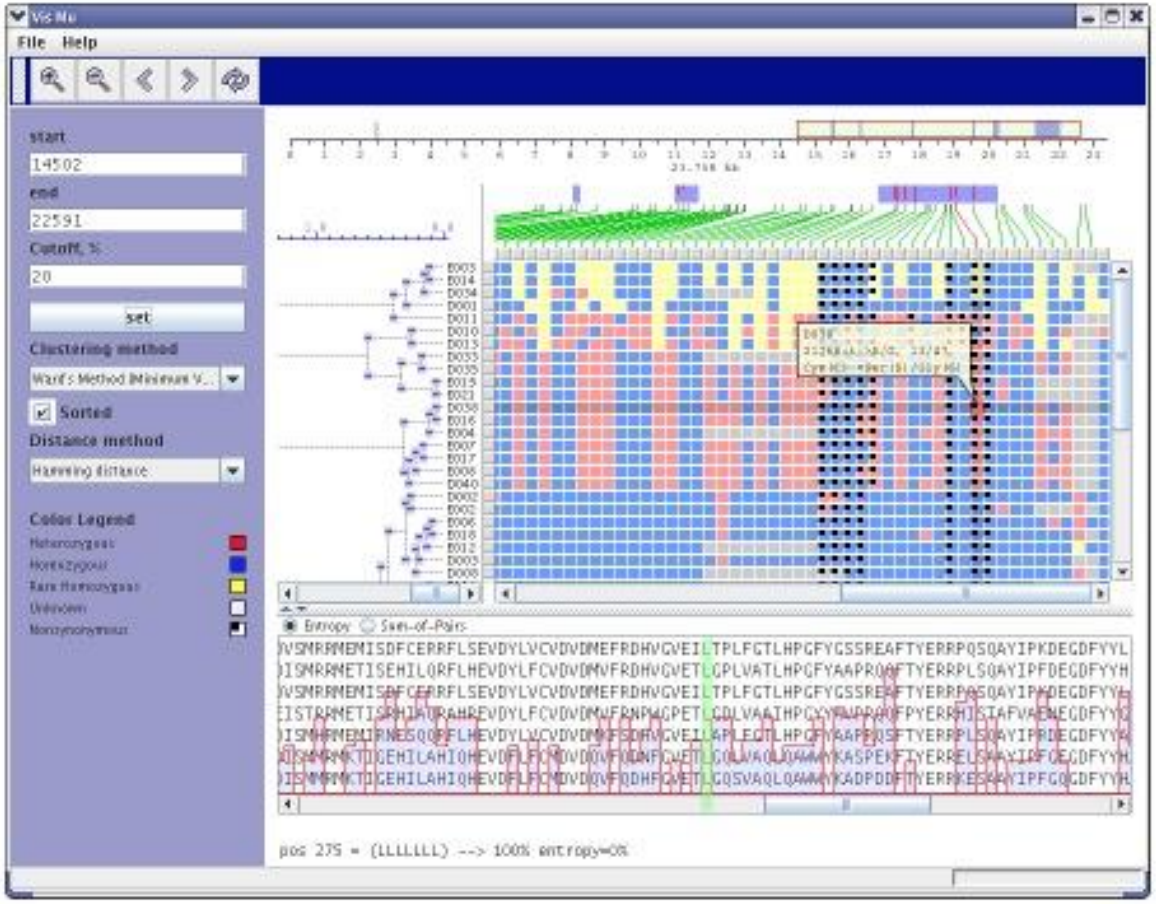


VISTA Tools for Comparative Genomics

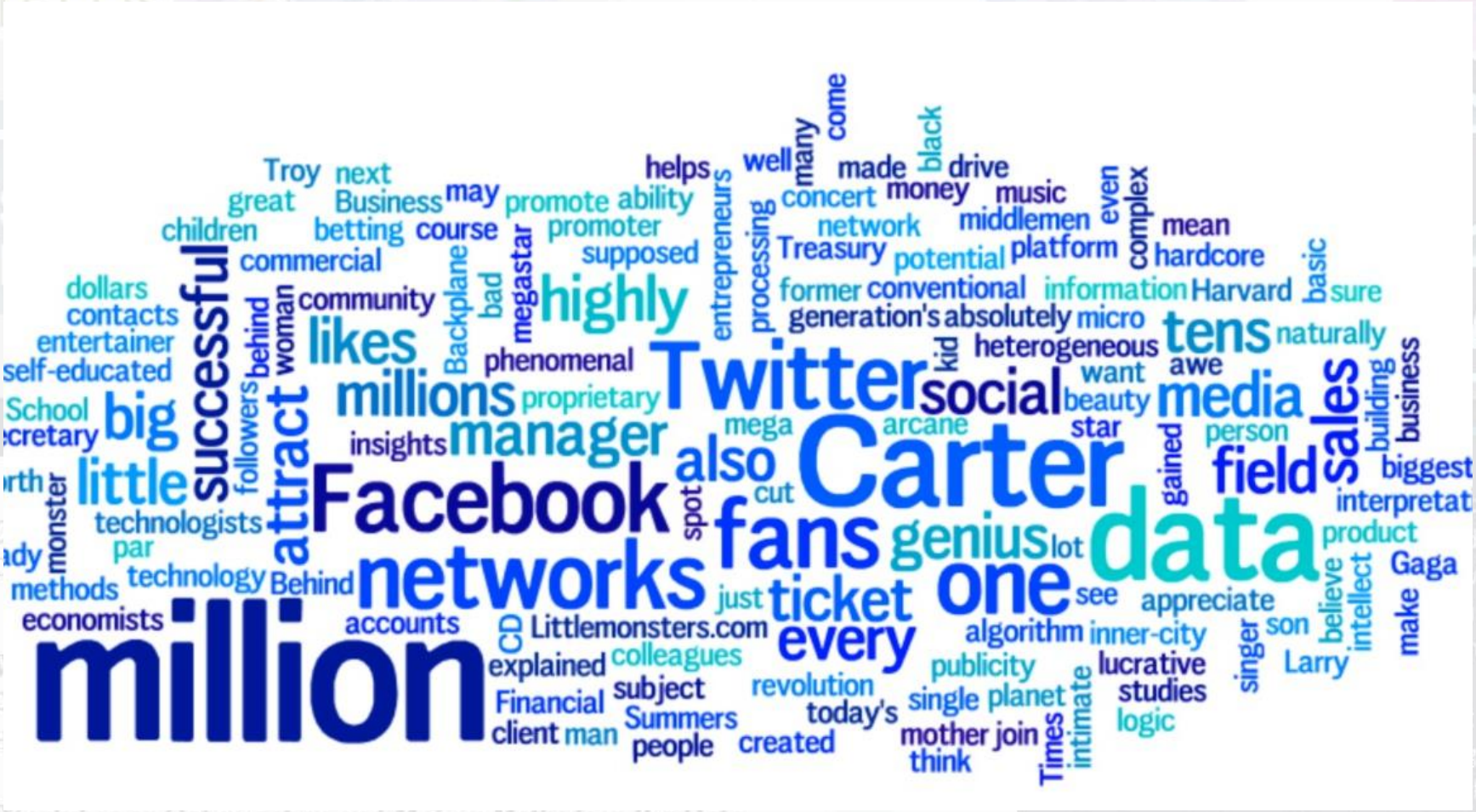
VISTA Home | Custom Alignment | Browser | Enhancer DB | Downloads | Publications

SNP-VISTA

GeneSNP-VISTA: Visualization of mutations in genes



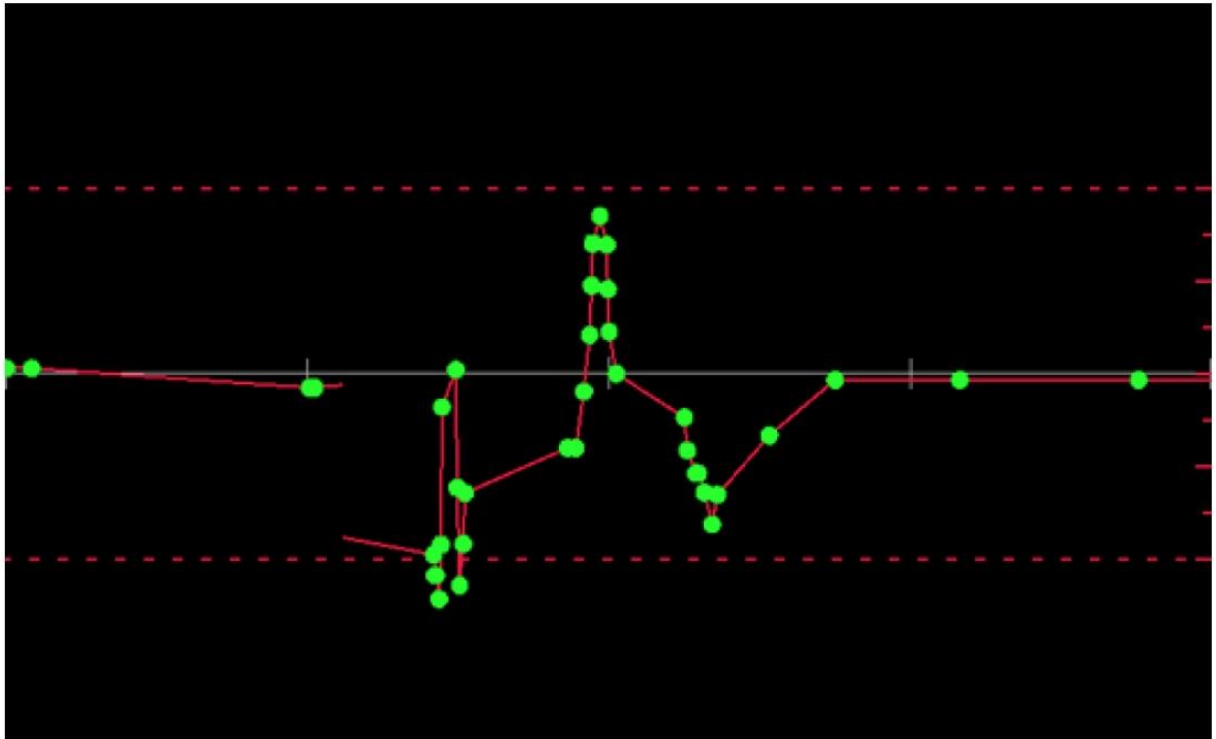
Social Activity



TECHNOLOGY

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

By Clay Dillow Posted February 10, 2011



https://www.ontariohealthstudy.ca/en/home

English Français

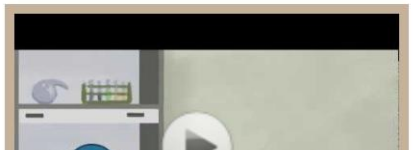
Ontario Health Study
Étude sur la santé Ontario

Already Registered? Log In Here

Home | About The Study | Privacy | Taking Part in the Study | Community

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



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, …

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

2005
2020

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

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

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

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

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

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, NEPTEC, QAS

IBM

http://www.bbc.co.uk/programmes/p028cm6q

More or Less Home Episodes Join us on Facebook Follow us on Twitter

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…

Available now

10 minutes

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

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

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

Journal of Privacy and Confidentiality (2012) 4, Number 2, 1–5

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

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

A look ahead

WEEK ONE: JANUARY 12 - 16

Monday January 12: Introductory Lectures and Overview	
8:00	Coffee and Registration
8:45	Welcome
9:00-9:30	**Nancy Reid**, University of Toronto
9:30-10:30	**Keynote Lecture: Bob Bell**, AT&T Labs - Research *Big Data: It's Not the Data*
10:30-11:00	Coffee
11:00-12:00	**Adam Kalai**, Microsoft *Machine learning and crowdsourcing*
12:00-2:00	Lunch
2:00-3:00	**Hugh Chipman,** Acadia University *An overview of Statistical Learning*
3:00-3:30	Tea
3:30-4:30	**Yulia Gel,** University of Waterloo *The Role of Modern Social Media Data in Surveillance and Prediction of Infectious Diseases: from Time Series to Networks*
4:30	Cash Bar Reception
Tuesday January 13: Introductory Lectures and Overview	
9:30-10:30	**Keynote Lecture: Emmanuel Candes**, Stanford University *Big Data and the Reproducibility of Scientific Research: What Can Statistics Offer*
10:30-11:00	Coffee break
11:00-12:00	**Steve Scott**, Google Inc *Bayes and Big Data: The Consensus Monte Carlo Algorithm*
12:00-1:30	Lunch break

A haphazard web walk

Quantathon from Waterfont International http://quantathon.ca

Forbes article “give me the data!” http://www.forbes.com/sites/lutzfinger/2014/11/21/give-me-data-why-data-ideas-fail/

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

A haphazard web walk

McGill Newsroom re above Science article: http://www.mcgill.ca/newsroom/channels/news/social-media-data-pose-pitfalls-studying-behaviour-240450

Forbes "12 big data definitions" http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/

Freakonometrics blog http://freakonometrics.hypotheses.org

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)

A haphazard web walk

Social media in ROB: http://www.theglobeandmail.com/report-on-business/industry-news/marketing/study-highlights-gap-between-social-medias-likers-and-lurkers-for-brands/article21996548/

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

Start-ups etc. - http://www.nytimes.com/2014/12/15/technology/in-big-data-shepherding-comes-first-.html?smid=tw-share

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

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

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.dpbs

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