

Foundations of Data Science

Michael W. Mahoney

ICSI and Dept of Statistics, UC Berkeley

April 2016

(For more info, see: <http://www.stat.berkeley.edu/~mmahoney>)

BIG data??? MASSIVE data????

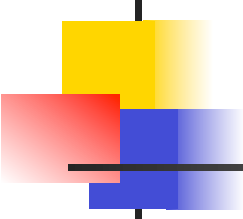


NYT, Feb 11, 2012: “The Age of Big Data”

- “What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. ...”

Why are big data big?

- Generate data at different places/times and different resolutions
- Factor of 10 more data is not just more data, but different data



BIG data??? MASSIVE data????

MASSIVE data:

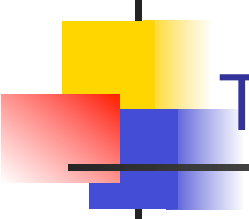
- Internet, Customer Transactions, Astronomy/HEP = “Petascale”
- One Petabyte = watching 20 years of movies (HD) = listening to 20,000 years of MP₃ (128 kbits/sec) = way too much to browse or comprehend

massive data:

- 10^5 people typed at 10^6 DNA SNPs; 10^6 or 10^9 node social network; etc.

In either case, main issues:

- Memory management issues, e.g., push computation to the data
- Hard to answer even basic questions about what data “looks like”



Thinking about large-scale data



Data generation is **modern version of microscope/telescope***:

- **See things couldn't see before**: e.g., fine-scale movement of people, fine-scale clicks and interests; fine-scale tracking of packages; fine-scale measurements of temperature, chemicals, etc.
- Those inventions ushered new scientific eras and **new understanding of the world** and new technologies to do stuff

**But algorithms are also a “microscope/telescope” to probe data.*

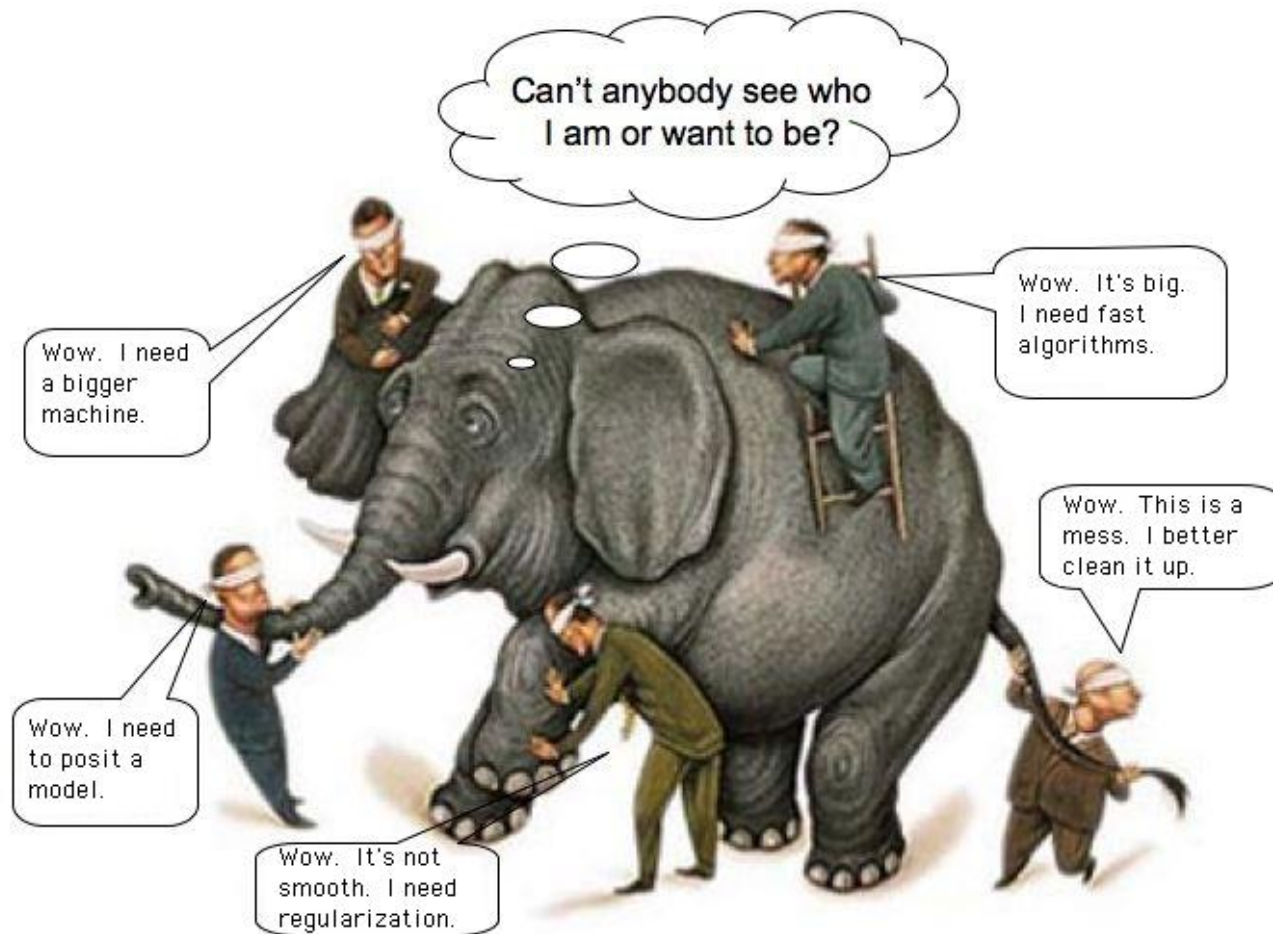
Easy things become hard and **hard things become easy**:

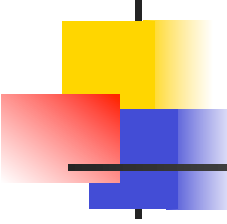
- Easier to see the other side of universe than bottom of ocean
- Means, sums, medians, correlations is easy with small data

Our ability to generate data far exceeds our ability to extract insight from data.



How do we view BIG data?





Algorithmic & Statistical Perspectives ...

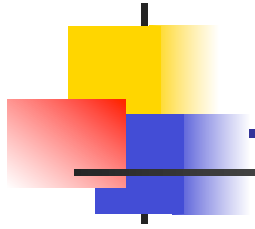
Lambert (2000), Mahoney (2010)

Computer Scientists

- *Data*: are a **record of everything** that happened.
- *Goal*: process the data to **find interesting patterns** and associations.
- *Methodology*: Develop approximation algorithms under different models of data access since the goal is typically **computationally hard**.

Statisticians (and Natural Scientists, etc)

- *Data*: are a **particular random instantiation** of an underlying process describing unobserved patterns in the world.
- *Goal*: is to **extract information** about the world from noisy data.
- *Methodology*: Make inferences (perhaps about unseen events) by **positing a model** that describes the random variability of the data around the deterministic model.



... are VERY different paradigms

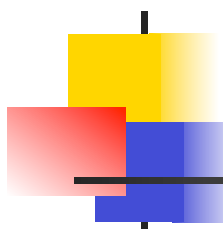
Statistics, natural sciences, scientific computing, etc:

- Problems often involve computation, but the study of *computation per se is secondary*
- Only makes sense to develop algorithms for *well-posed* problems*
- First, write down a model, and think about computation later

Computer science:

- Easier to study *computation per se in discrete settings*, e.g., Turing machines, logic, complexity classes
- Theory of algorithms *divorces computation from data*
- First, run a fast algorithm, and ask what it means later

*Solution exists, is unique, and varies continuously with input data



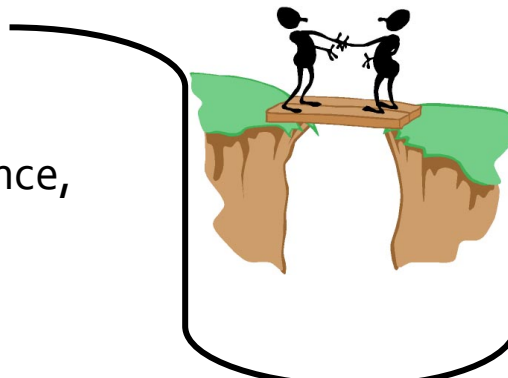
Anecdote 1: Randomized Matrix Algorithms

Mahoney “Algorithmic and Statistical Perspectives on Large-Scale Data Analysis” (2010)

Mahoney “Randomized Algorithms for Matrices and Data” (2011)

Theoretical origins

- theoretical computer science, convex analysis, etc.
- Johnson-Lindenstrauss
- Additive-error algs
- Good worst-case analysis
- No statistical analysis
- No implementations



Practical applications

- NLA, ML, statistics, data analysis, genetics, etc
- Fast JL transform
- Relative-error algs
- Numerically-stable algs
- Good statistical properties
- *Beats LAPACK & parallel-distributed implementations on terabytes of data*

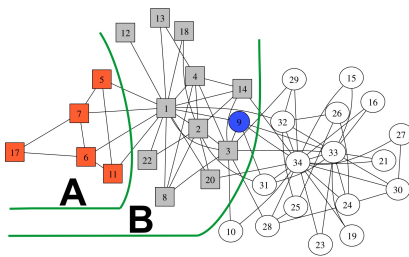
How to “bridge the gap”?

- *decouple (implicitly or explicitly) randomization from linear algebra*
- importance of *statistical leverage* scores!

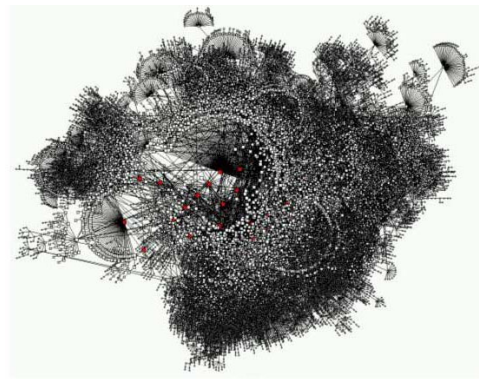
Anecdote 2: Communities in large informatics graphs

Mahoney “Algorithmic and Statistical Perspectives on Large-Scale Data Analysis” (2010)
Leskovec, Lang, Dasgupta, & Mahoney “Community Structure in Large Networks ...” (2009)

People imagine social networks to look like:



Real social networks actually look like:

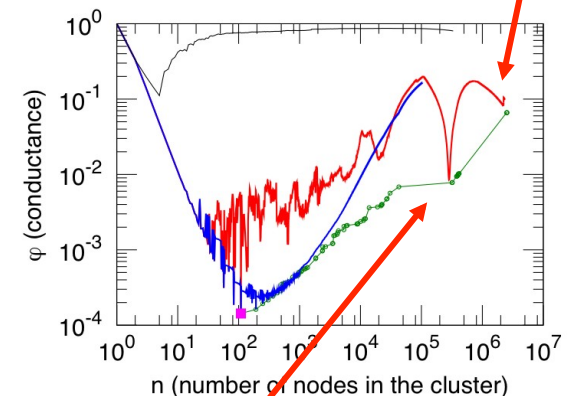


How do we know this plot is “correct”?

- (since computing conductance is intractable)
- Lower Bound Result; Structural Result; Modeling Result; Etc.
- Algorithmic Result (ensemble of sets returned by different approximation algorithms are very different)
- *Statistical Result* (Spectral provides more meaningful communities than flow)

Data are expander-like at large size scales !!!

Size-resolved conductance (degree-weighted expansion) plot looks like:



There do not exist good large clusters in these graphs !!!

Anecdote 3:

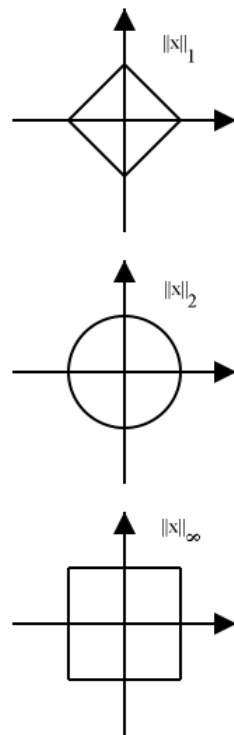
Approx. comp. and implicit regularization

Mahoney “Approximate Computation and Implicit Regularization for Very Large-scale Data Analysis” (2012)

Explicitly-imposed regularization

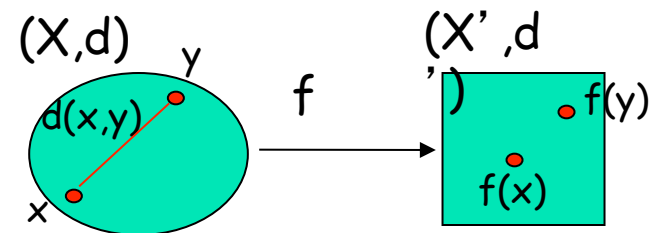
- Traditionally, regularization uses *explicit* norm constraint to make sure solution vector is “small” and not-too-complex

$$\min ||f|| + \lambda ||g(x)||$$



Implicitly-imposed regularization

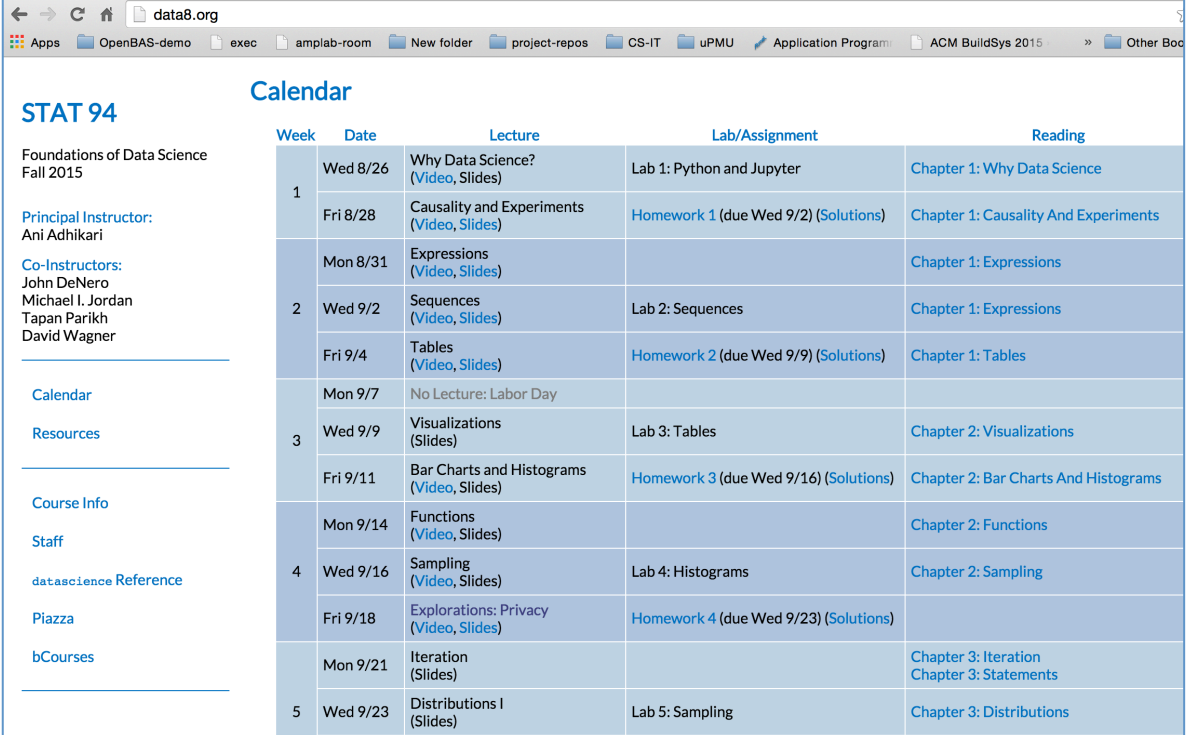
- Binning, pruning, early stopping, etc.
- Design decisions engineers make
- Approximation algorithms *implicitly* embed data in a “nice” metric/geometric place and then round the solution.



Big question: Can we formalize the notion that/when approximate computation in and of itself can *implicitly* lead to “better” or “more regular” solutions than exact computation? (Short answer: yes!)

Anecdote 4: Teaching Freshmen: Data8: Foundations of Data Science

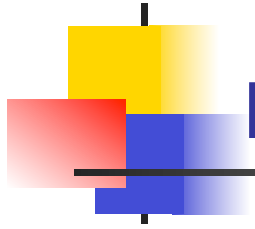
- <http://data8.org>
- Fundamental co-mingling of CS & Stat concepts on real data
- Learn computing concepts by doing interesting things on data
- Learn statistical concepts by observing what's interesting
- Codify understanding of concepts symbolically
- Explorations
- Connectors: $4+2=6$



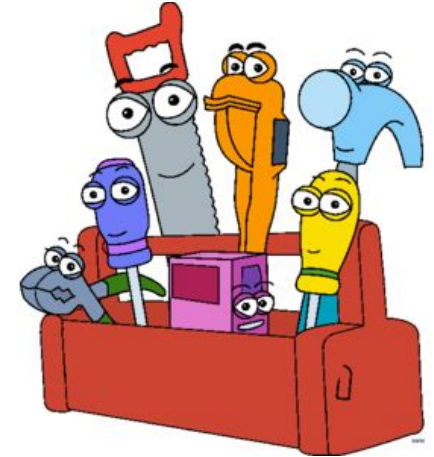
The screenshot shows the data8.org website with the course information for STAT 94, Foundations of Data Science, Fall 2015. The page includes a sidebar with links to Calendar, Resources, Course Info, Staff, datascience Reference, Piazza, and bCourses. The main content is a calendar table with columns for Week, Date, Lecture, Lab/Assignment, and Reading.

Week	Date	Lecture	Lab/Assignment	Reading
1	Wed 8/26	Why Data Science? (Video, Slides)	Lab 1: Python and Jupyter	Chapter 1: Why Data Science
	Fri 8/28	Causality and Experiments (Video, Slides)	Homework 1 (due Wed 9/2) (Solutions)	Chapter 1: Causality And Experiments
2	Mon 8/31	Expressions (Video, Slides)		Chapter 1: Expressions
	Wed 9/2	Sequences (Video, Slides)	Lab 2: Sequences	Chapter 1: Expressions
	Fri 9/4	Tables (Video, Slides)	Homework 2 (due Wed 9/9) (Solutions)	Chapter 1: Tables
3	Mon 9/7	No Lecture: Labor Day		
	Wed 9/9	Visualizations (Slides)	Lab 3: Tables	Chapter 2: Visualizations
	Fri 9/11	Bar Charts and Histograms (Video, Slides)	Homework 3 (due Wed 9/16) (Solutions)	Chapter 2: Bar Charts And Histograms
4	Mon 9/14	Functions (Video, Slides)		Chapter 2: Functions
	Wed 9/16	Sampling (Video, Slides)	Lab 4: Histograms	Chapter 2: Sampling
	Fri 9/18	Explorations: Privacy (Video, Slides)	Homework 4 (due Wed 9/23) (Solutions)	
5	Mon 9/21	Iteration (Slides)		Chapter 3: Iteration Chapter 3: Statements
	Wed 9/23	Distributions I (Slides)	Lab 5: Sampling	Chapter 3: Distributions

10/6/2015; Carson/Culler; DS experience @ UCB



Lessons from the anecdotes



We are being forced to **engineer a union between two very different worldviews** on what are fruitful ways to view the data

- in spite of our best efforts *not* to

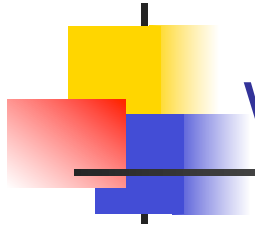
The forcing function (generation of lots of valuable data) is forcing us to **revisit old methods in a new light**

- often reinventing, but the forcing function makes that acceptable

Given existing forcing functions and disciplinary lines, **many methods and approaches are “undervalued”** for what non-foundational people want

- and it would be good not to lose them

QUESTION: How can we bridge the gap between these two worldviews?



What people say



“We are awash in a flood of data today”

“Big data does not arise out of a vacuum”

“Many scientific disciplines have become data driven”

“The era of big data is underway”

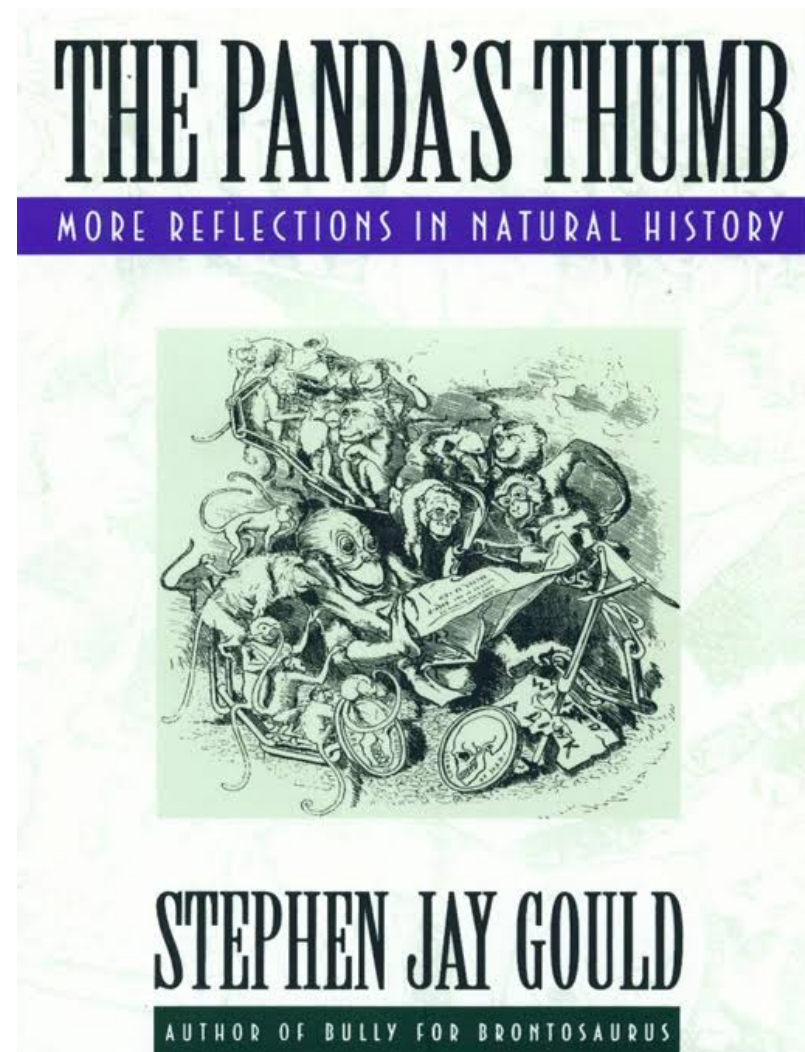
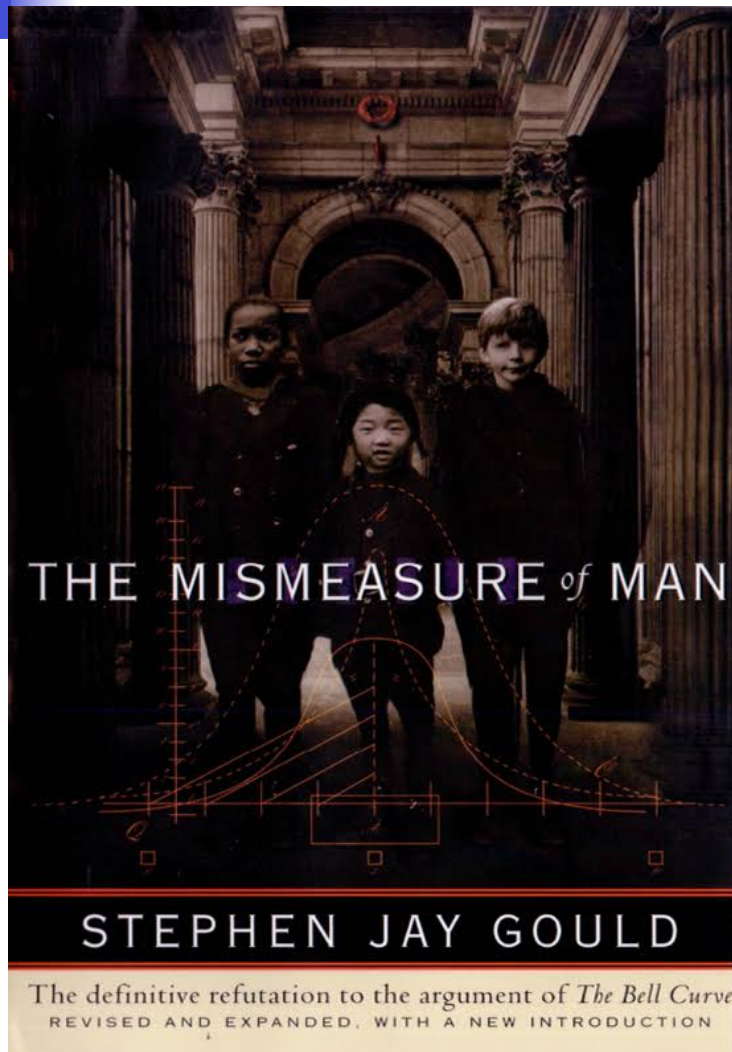
“There is growing enthusiasm for the notion of ‘big data’”

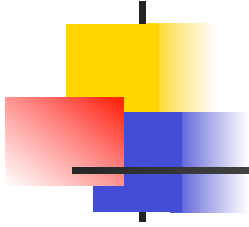
“We want to enable data-driven discovery and decision making”

...

QUESTION: What does all this mean?

Thinking big and thinking small





MMDS Workshop on “Algorithms for Modern Massive Data Sets”

(<http://mmds-data.org>)

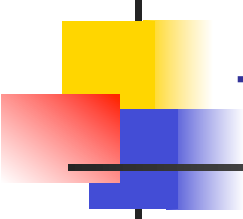
at UC Berkeley, June 21-24, 2016

Objectives:

- Address algorithmic, statistical, and mathematical challenges in modern statistical data analysis.
- Explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured data.
- Bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote cross-fertilization of ideas.

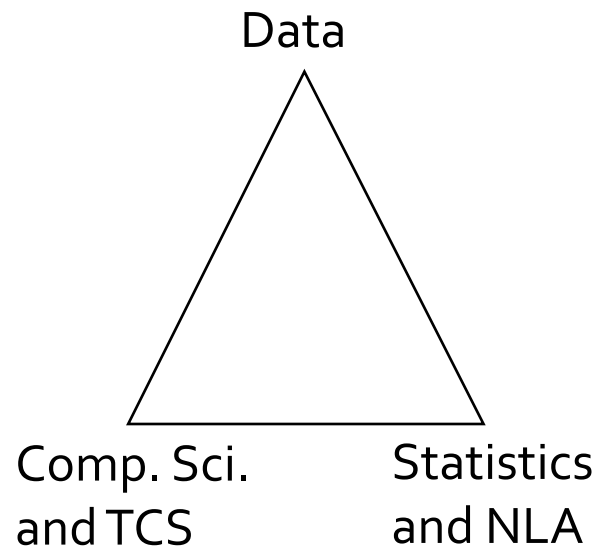
Organizers: M. W. Mahoney, A. Shkolnik, and P. Drineas,

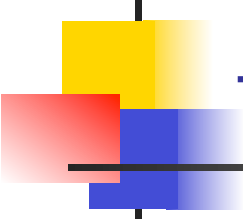
Registration is available now!



MMDS 2006:

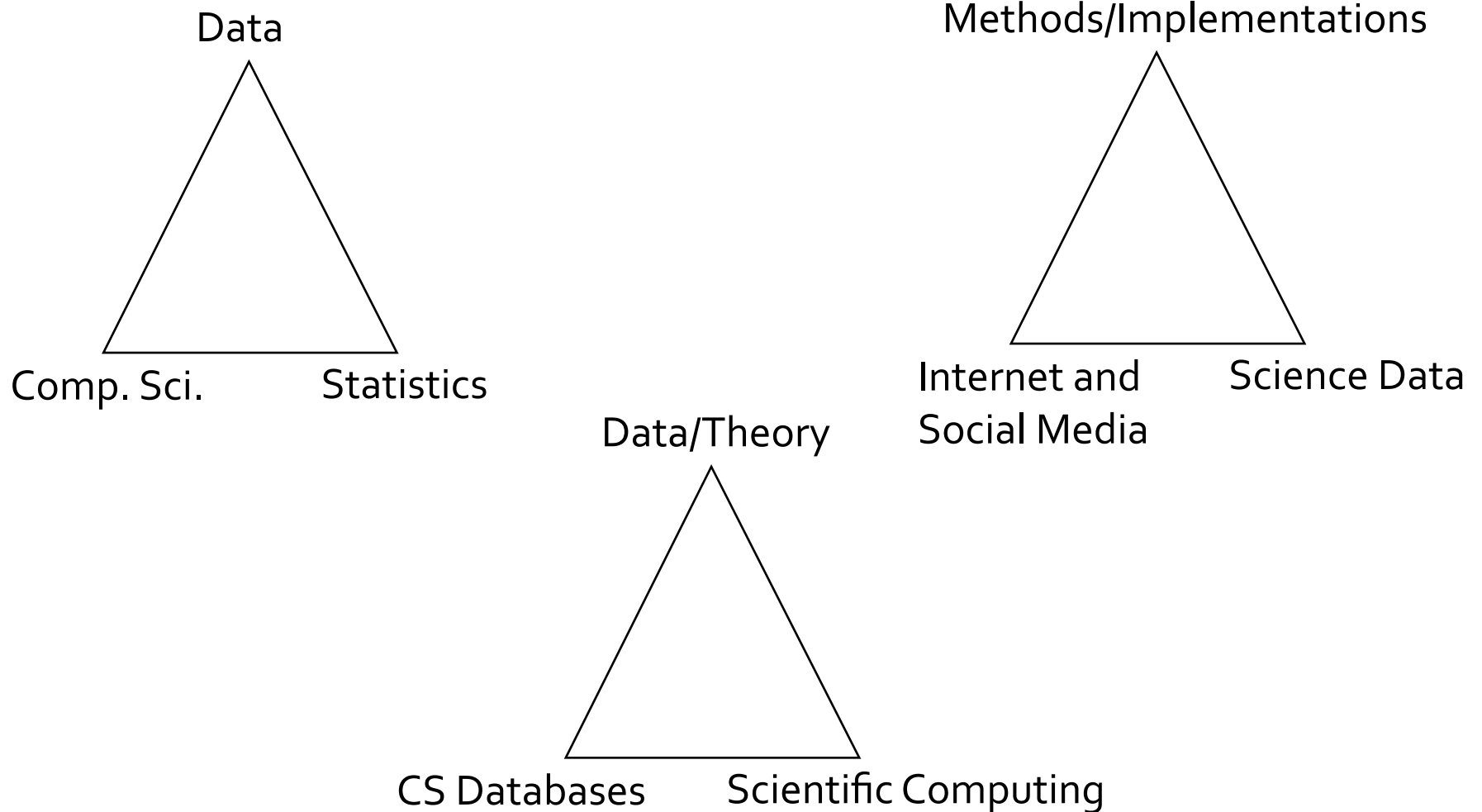
Things we worried about with MMDS





MMDS post-2006:

Things we worried about with MMDS



Pure and Applied Math

(matrix concentration inequalities, random matrix perturbation theory, multi-linear algebra, connections to fundamental conjectures, etc.)

Computer Science

(fast solvers for systems of linear equations, communication avoiding algorithms, optimization, graph-based and matrix-based data analysis algorithms, etc.)

RandNLA

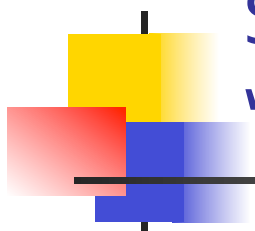
Randomization: a computational resource in designing and analyzing improved algorithms for large scale Linear Algebra problems

Statistics & Machine Learning

(implicit regularization and numerical robustness, optimization for complex statistical models, task-oriented sketching, interpretable analytics, etc.)

Implementations & Applications

(RandNLA in the field, petascale matrix computations, integration in numerical linear algebra and machine learning libraries, etc.)



Simons Institute: Fall 2013: “Theoretical foundations of big data analysis”

Three main workshops:

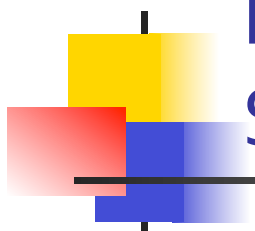
- Succinct Data Representations and Applications
- Parallel and Distributed Algorithms for Inference and Optimization
- Unifying Theory and Experiment for Large-Scale Networks

BIG success!

“These guys do a lot more theory *for* big data than the people at the theory *of* big data semester last term.”

- (Comment at the Computation-Intensive Probabilistic and Statistical Methods for Large-Scale Population Genomics workshop, at Simons Institute, Feb. 2014)

QUESTION: Why is that? What could/should we do about that?



Park City Mathematics Institute: Summer 2016: “Mathematics of Data”

- A 3-week mathematics program held each summer near Park City, UT
- An outreach program of the Institute for Advanced Study in Princeton, NJ
- Topics change from year to year
- The organizers of PCMI 2016 are **John Duchi** (Stanford), **Anna Gilbert** (Michigan), and **Michael Mahoney**, (UC, Berkeley).

Petros Drineas (RandNLA: Randomization in Numerical Linear Algebra)

John Duchi (Online, stochastic, and optimal methods in large-scale data analysis)

Cynthia Dwork (Differential privacy)

Robert Ghrist (Topological data analysis)

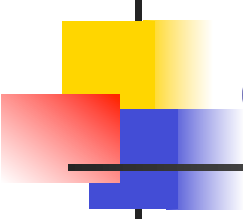
Piotr Indyk (Recent Developments in the Sparse Fourier Transform)

Mauro Maggioni (Geometric and graph methods for high-dimensional data)

Gunnar Martinsson (Randomized algorithms for matrix computations and analysis of high dimensional data)

Roman Vershynin (Random matrices, random graphs and high-dimensional inference)

Stephen Wright (Optimization Techniques for Data Analysis)



Teaching Freshmen: Connectors: (4+2=6)

Modular approach to accommodate the huge diversity of student interests, backgrounds, and perspectives

- Do justice to “critical thinking with data”
- Address applications and ethics in context of use

Spectrum of models

- Placing data science in a context of questions in the world
- Technical depth beneath the “foundation” experience
- Hybrids
- Fall: 6 pilots

Spring 16 connectors:

- INFO 88: Data and Ethics
- COGSCI 88: Data Science and the Mind
- ESPM 88A: Exploring Geospatial Data
- STAT 88: Probability and Mathematical Statistics in Data Science
- ESPM 88: Data Sciences in Ecology and the Environment
- HIST 88: How Does History Count?
- STAT 89A: Matrices and Graphs
- CS 88: Computational Structures in Data Science
- L&S 88-1: Health, Human Behavior and Data
- CE 88: Data Science for Smart Cities
- L&S 88-2: Literature and Data

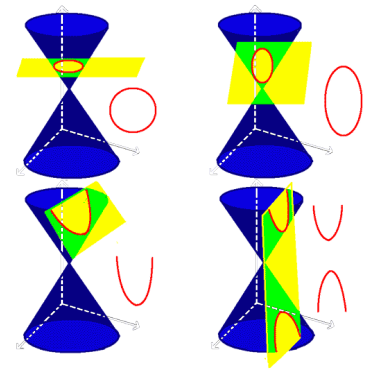
Teaching Freshmen: Stat89a “applied math” connector

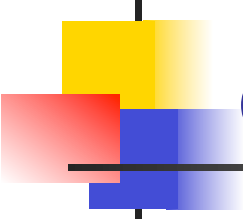
Pencil-and-paper homeworks

- Basic linear algebra, probability
- PCA/Least-squares/PageRank/diffusions via quadratic forms

Computational homeworks in python with ipynb

- Compute ratio of “ball to enclosing box” by throwing darts in 2D, to compute π
- Return mean and variance estimates for different number of throws
- Redo for 5D, 10D, 20D, etc.
- Compute distance to the origin and distributions of pair-wise distances, angles, etc. for points uniform in 2D sphere
- Redo for 5D, 10D, 20D, etc.
- What does this have to do with low versus high dimensional classification?
- What does this have to do with Chernoff-style tail bounds?
- Exploratory data analysis with PCA; multivariate regression with LS





Questions



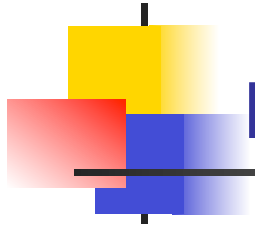
Why do computer science departments even exist?

What is statistics (inside academic departments versus outside academic departments)?

Where is applied mathematics in this big data area?

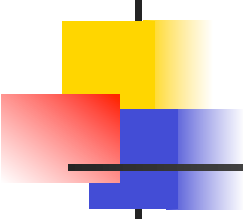
How does NSF fit into the funding ecosystem?





Bioinformatics: a cautionary tale?

- How did/does bioinformatics relate to computer science, statistics, and applied mathematics, “technically” and “sociologically”?
- How did NIH choose to fund graduate students and postdocs in the budget expansion of the 90s?
- What effect did this have on the number of American/foreign going into biomedical research?
- How will the pay structure of biomedical researchers effect which cs/stats “data scientists” engage you in your efforts?
- What effect does med schools deciding not to do joint faculty hires with cs departments have on bioinformatics and big biomedical data?
- How is this Big Biomedical Data phenomenon similar to and different than the Bioinformatics experience?



Big changes in the past ... and future

Consider the creation of:

- Modern Physics
- Computer Science
- Molecular Biology
- OR and Management Science
- Transistors and Microelectronics
- Biotechnology

These were driven by *new measurement techniques* and *technological advances*, but they led to:

- big new (academic and applied) questions
- new perspectives on the world
- lots of downstream applications

We are in the middle of a similarly big shift!

- Fundamental *foundational* challenges to statistics and computer science
- What will all this lead to ...

