UC Berkeley's FODA Institute: Foundations of Data Analysis

Michael W. Mahoney

Michael Jordan
Richard Karp
Fernando Perez
Bin Yu
(A large halo of other people TBD going forward)

UC Berkeley
October 2017

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

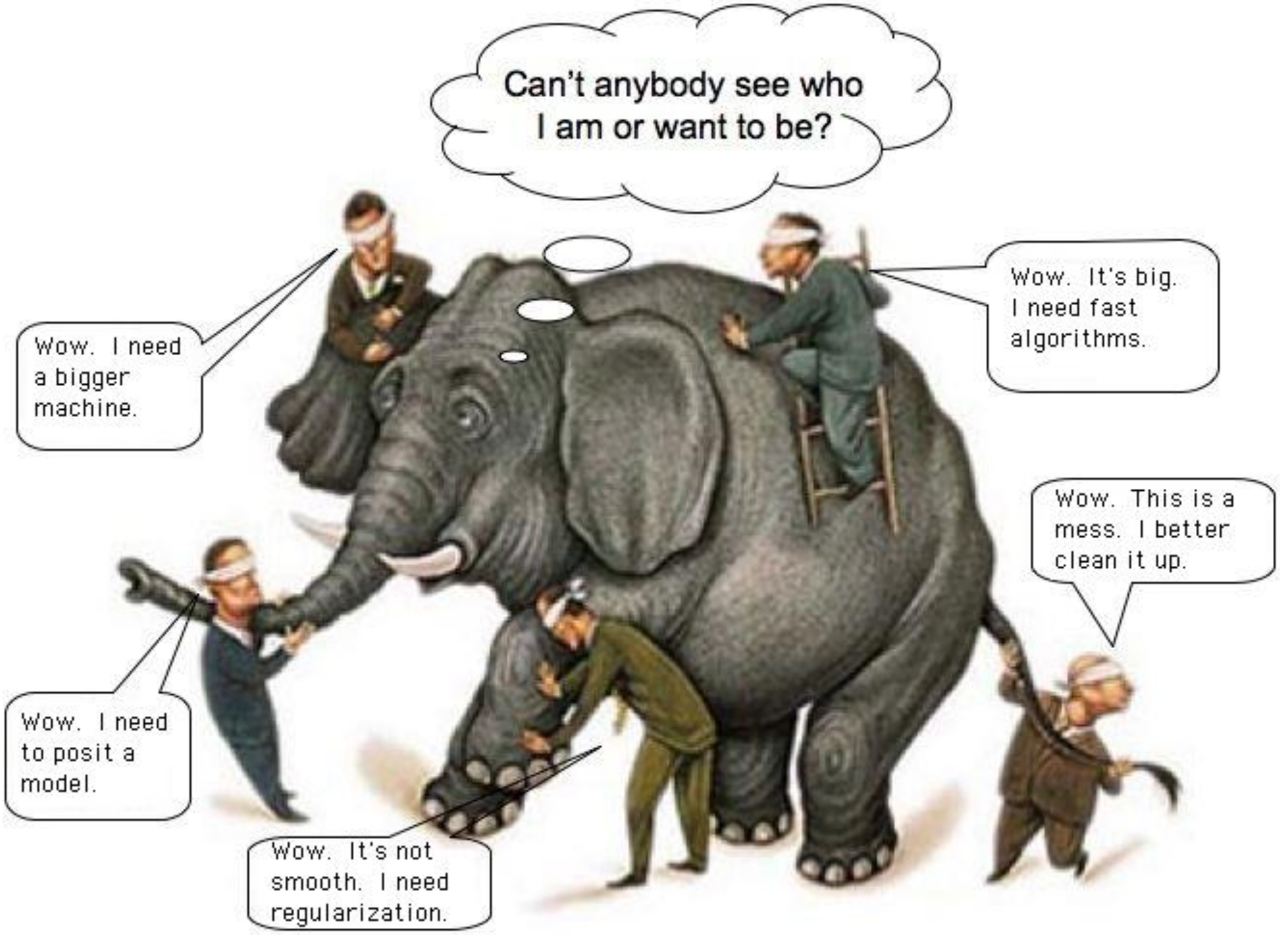
Also: **But algorithms/statistics are also a "microscope/telescope" to probe data.*

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?


Can't anybody see who I am or want to be?
Wow. I need a bigger machine.
Wow. It's big. I need fast algorithms.
Wow. This is a mess. I better clean it up.
Wow. I need to posit a model.
Wow. It's not smooth. I need regularization.

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.

Where is mathematics and the (applied) mathematicians?

UC Berkeley FODA PI and coPIs

- Michael Mahoney
- Michael Jordan
- Richard Karp
- Fernando Perez
- Bin Yu

Overview

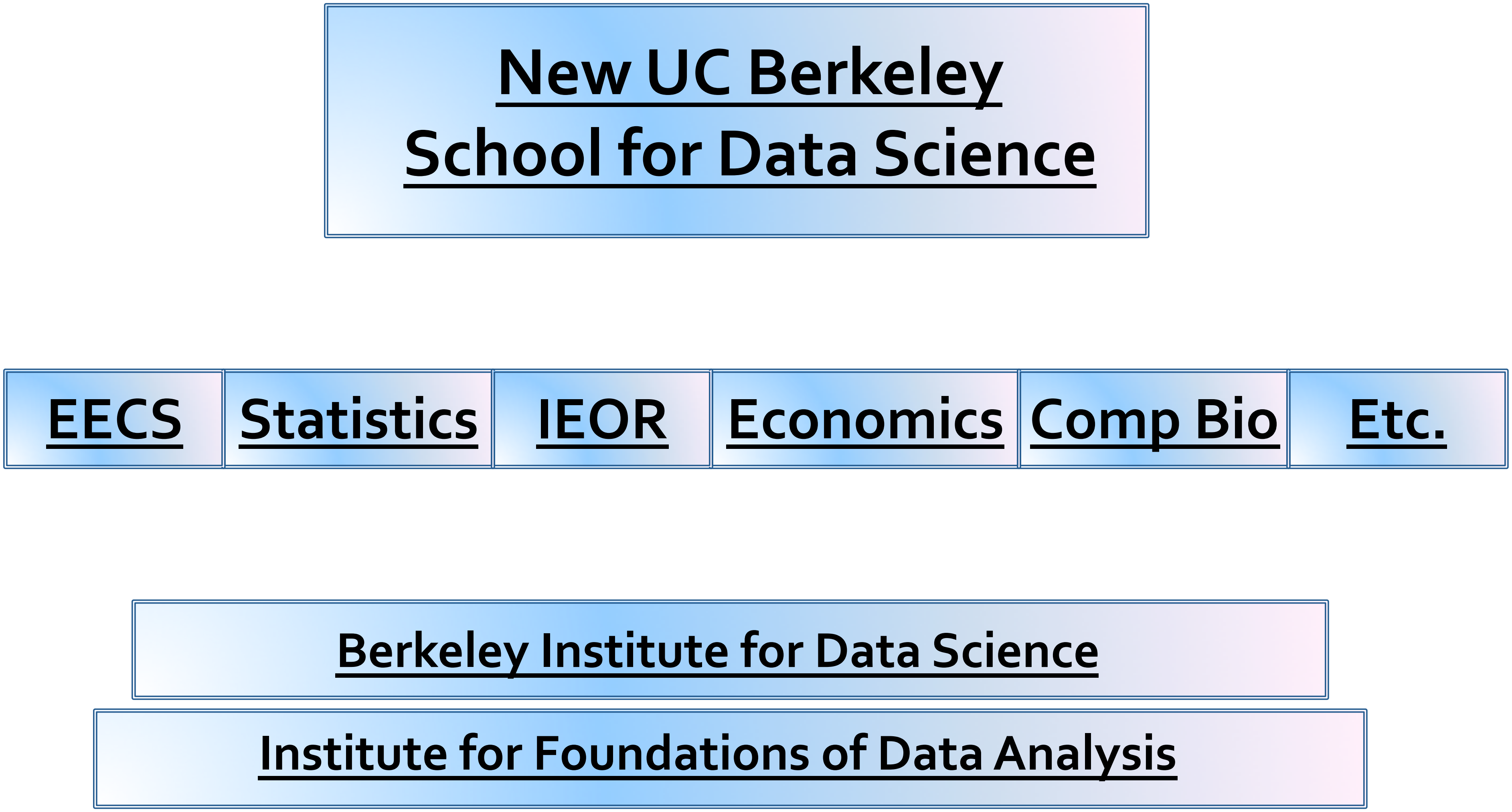
- Foundations of Data Analysis (FODA) Institute

- Development, teaching, and applications of foundational methods

- None of the many entities at Berkeley are devoted to addressing the interdisciplinary foundations of data

- New division-level, dean-led academic unit on data science (largest structural reorganization in decades)

- Will enable unified approach, woven directly into interdisciplinary scientific and educational innovation at Berkeley.

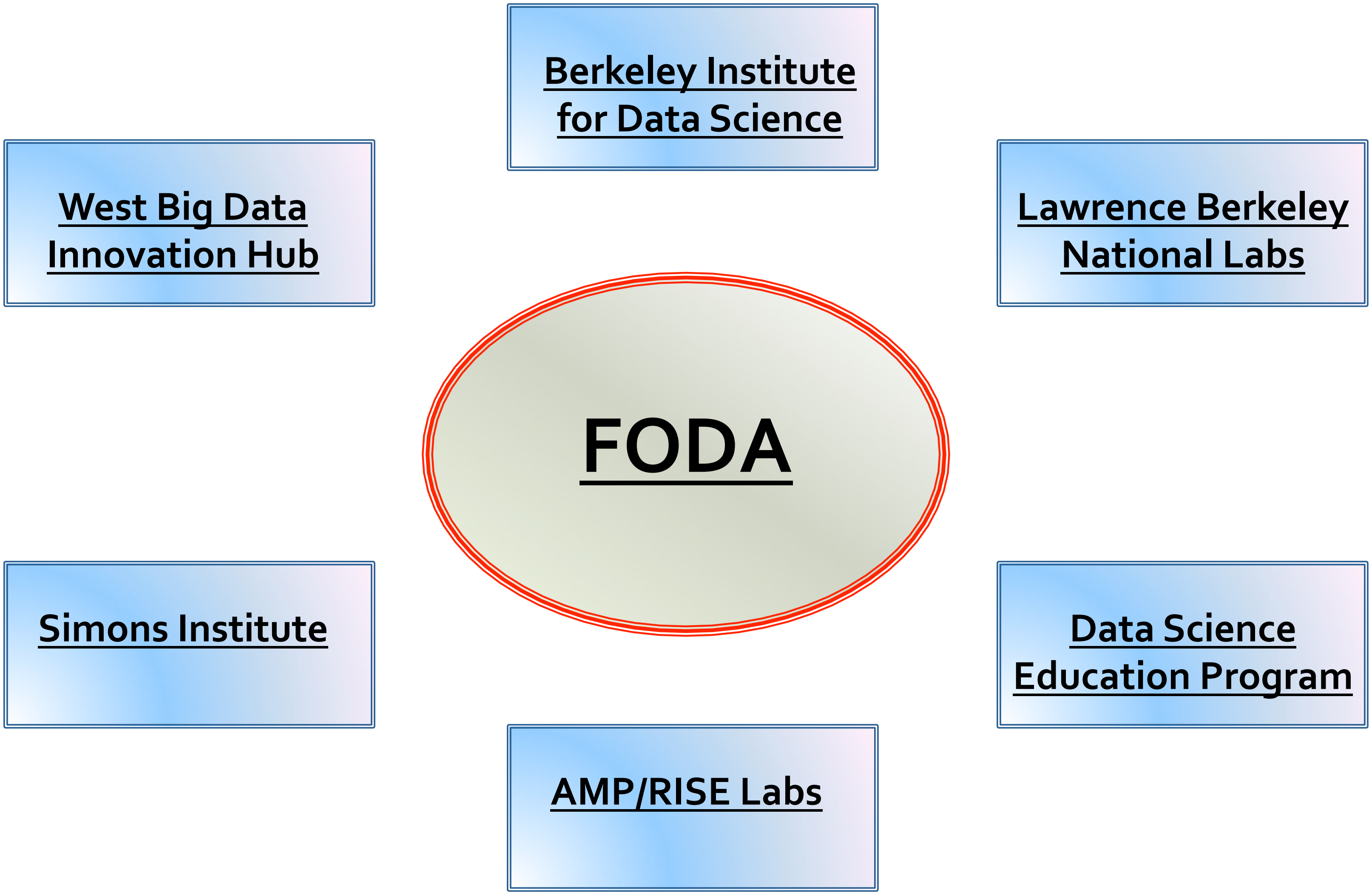


New UC Berkeley School for Data Science

EECS | **Statistics** | **IEOR** | **Economics** | **Comp Bio** | **Etc.**

Berkeley Institute for Data Science

Institute for Foundations of Data Analysis



Berkeley Institute for Data Science

West Big Data Innovation Hub

Lawrence Berkeley National Labs

FODA

Simons Institute

Data Science Education Program

AMP/RISE Labs

Intellectual Merit

As a start, four fundamental challenges:

- Complexity theory of inference: upper/lower bounds for inferential optimization problems
- Stability: A unifying computational-inferential principle
- Data-driven computational mathematics: Randomness as a statistical versus algorithmic versus mathematical resource
- Science-based models and data-driven models: Can they be combined

Education: from *researchers* to *graduate students* to *freshman*

Broader Impact

- Integrated into the undergraduate and graduate Data Science Education Program (DSEP) at Berkeley

- Collaborations with: BIDS; Simons Institute; RISE/AMP Labs; LBNL; and NSF Big Data Regional Innovation Hub

- Groundwork for interactions between theoretically-inclined data science researchers and researchers in diverse domains

- More principled extraction of domain insights from data across a wide range of domains

Example of a Grand Challenge Problem

When two parameters, e.g., model complexity and data size, diverge together:

- Theory people: fix one parameter and let the other get large
- Practical people: hope big data means enough for things to average out
- Actually: Factor of 10 more data means a factor of 10 different data

Informally:

- Data are not a meaningful perturbative approximation of a nice limit state*

*Great place to look for fundamental advances:
e.g., Markov Chain Monte Carlo

Some specific ideas/issues/challenges

- Formal Home
- Informal Ecosystem
- Physical Space
- Education and workforce development
- Goal-based finite-horizon research (initially four focal points … but maybe convex combination of NSF STC and Math Institutes)
- Collaborative working groups (maybe Miller-like mechanism?)
- Government lab connections (are they really interested?)
- Industrial connections (find the right "value proposition"?)
- Campus partners/investments (including BIDS, Simons, LBNL, AMP/RISE, NSF Big Data Regional Innovation Hub; etc.)

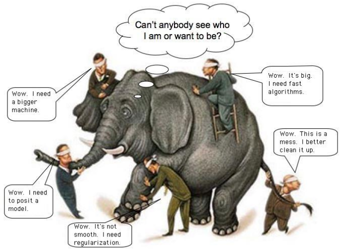
Simons Institute: F13: "Theoretical foundations of big data analysis"

O(DATA)

Main workshops:

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

BIG success!


Can't anybody see who I am or want to be?
Wow. I need a bigger machine.
Wow. It's big. I need fast algorithms.
Wow. This is a mess. I better clean it up.
Wow. I need to posit a model.
Wow. It's not smooth. I need regularization.

"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)

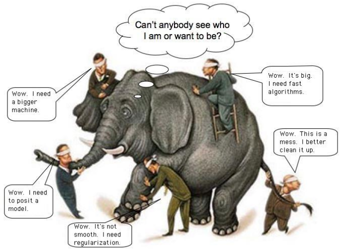
QUESTION: Why is that? What could/should we do about that?

Simons Institute: F18: “Foundations of data science”

Main workshops:

- Kickoff Boot Camp (8/27-31, 2018)
- Randomized Numerical Linear Algebra and Applications (9/24-28, 2018)
- Robust and High-Dimensional Statistics (10/29-11/2, 2018)
- Sublinear Algorithms and Nearest-Neighbor Search (11/27-30, 2018)

(We are planning on a) BIG success!


Can't anybody see who I am or want to be?
Wow. I need a bigger machine.
Wow. It's big. I need fast algorithms.
Wow. This is a mess. I better clean it up.
Wow. I need to posit a model.
Wow. It's not smooth. I need regularization.

“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)

GOAL: Have this less true in 2018 and even less true in 2023 than in 2013 or 2008.

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 year to year
- The organizers of PCMI 2016 are **John Duchi** (Stanford), **Anna Gilbert** (Michigan), and **Michael Mahoney**, (UC, Berkeley).

- *Edited volume of lectures (to teach from) available soon!*

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)

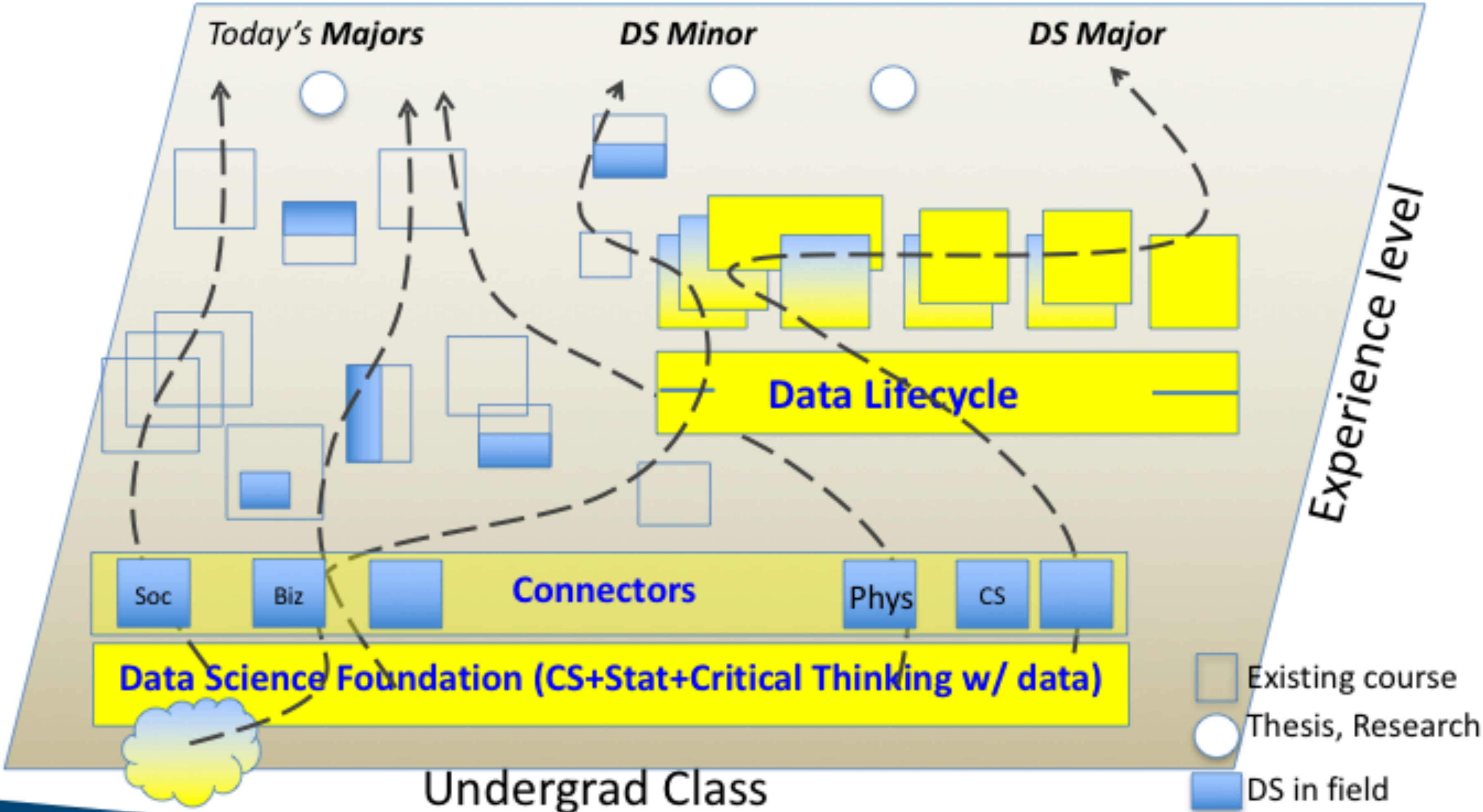
Mauro Maggioni (Geometric and graph methods for high-dimensional data)

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)

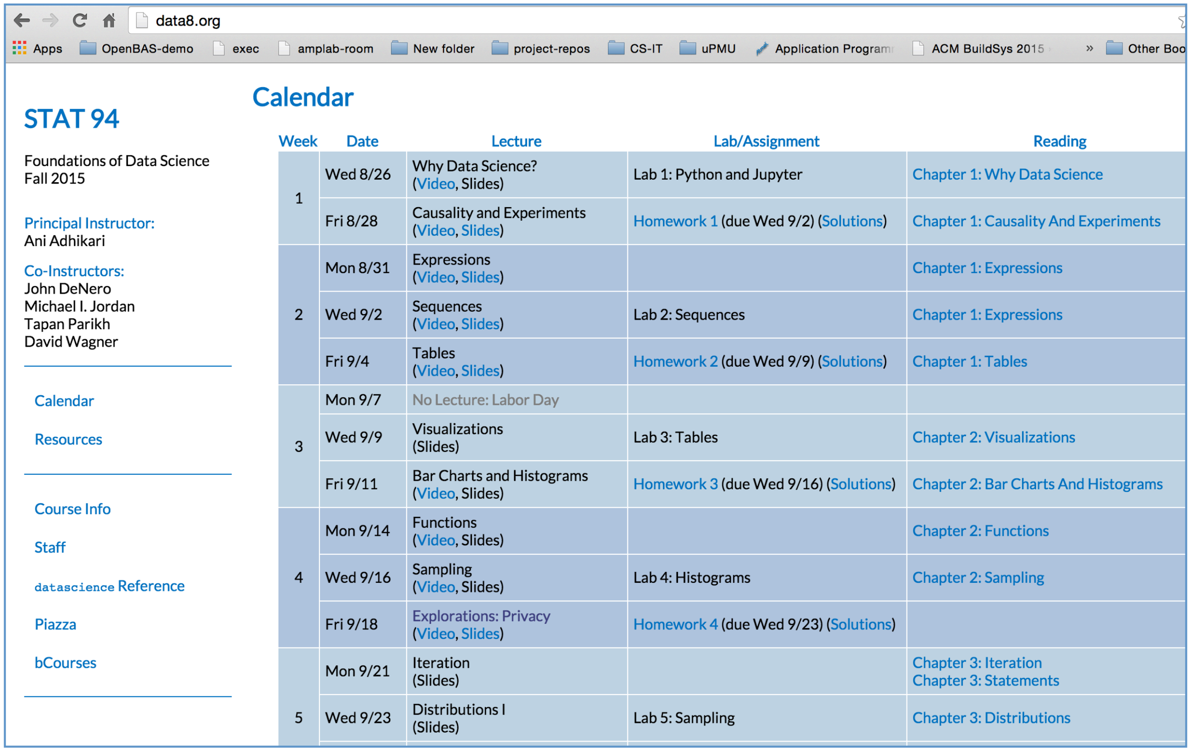
Undergrad experience transformed



Berkeley
UNIVERSITY OF CALIFORNIA

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



STAT 94

Foundations of Data Science
Fall 2015

Principal Instructor:
Ani Adhikari

Co-Instructors:
John DeNero
Michael I. Jordan
Tapan Parikh
David Wagner

Calendar

Resources

Course Info

Staff

`datascience` Reference

Piazza

bCourses

Calendar

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

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

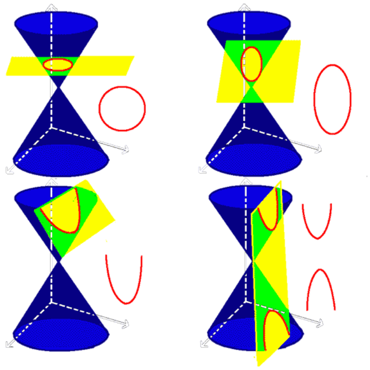
- Placing data science in a context of questions in the world
- Technical depth beneath the "foundation" experience
- Hybrids
- Fall: 6 pilots

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 (Math of Data): 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 can we best use NSF funding in the ecosystem?