Big Data, Data Science, and Other Buzzwords that *Really Matter*

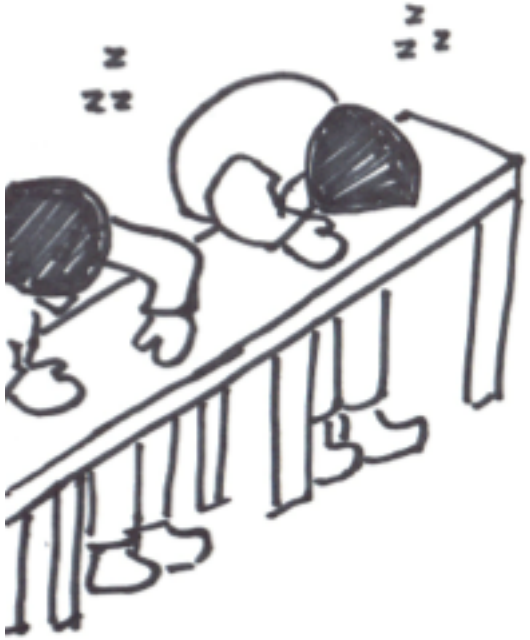
Michael Franklin
July 21, 2014

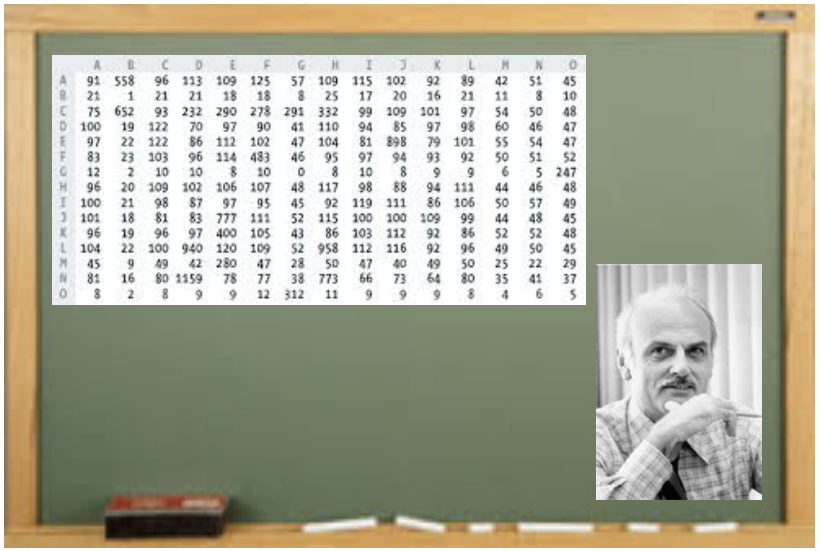
CRA Snowbird Conference

BERKELEY INSTITUTE FOR
DATA SCIENCE

MY DATA STORY

Teaching Databases (first 20 yrs)



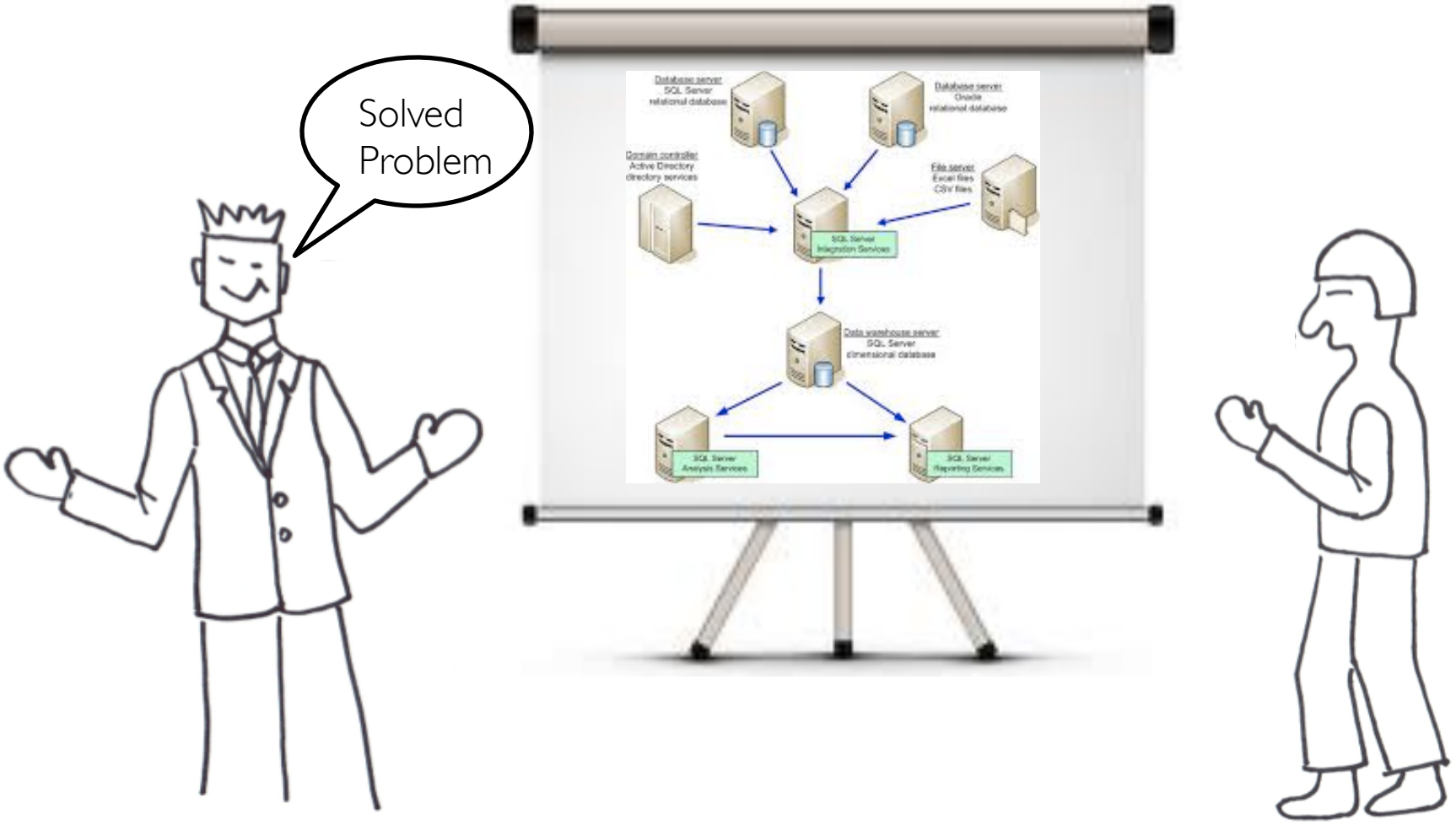


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	91	558	96	113	109	125	57	109	115	102	92	89	42	51	45
B	21	1	21	21	18	18	8	25	17	20	16	21	11	8	10
C	75	652	93	232	290	278	291	332	99	109	101	97	54	50	48
D	100	19	122	70	97	90	41	110	94	85	97	98	60	46	47
E	97	22	122	86	112	102	47	104	81	898	79	101	55	54	47
F	83	23	103	96	114	483	46	95	97	94	93	92	50	51	52
G	12	2	10	10	8	10	0	8	10	8	9	9	6	5	247
H	96	20	109	102	106	107	48	117	98	88	94	111	44	46	48
I	100	21	98	87	97	95	45	92	119	111	86	106	50	57	49
J	101	18	81	83	777	111	52	115	100	100	109	99	44	48	45
K	96	19	96	97	400	105	43	86	103	112	92	86	52	52	48
L	104	22	100	940	120	109	52	958	112	116	92	96	49	50	45
M	45	9	49	42	280	47	28	50	47	40	49	50	25	22	29
N	81	16	80	1159	78	77	38	773	66	73	64	80	35	41	37
O	8	2	8	9	9	12	312	11	9	9	9	8	4	6	5

Teaching DATA SCIENCE

Pitching Data Management

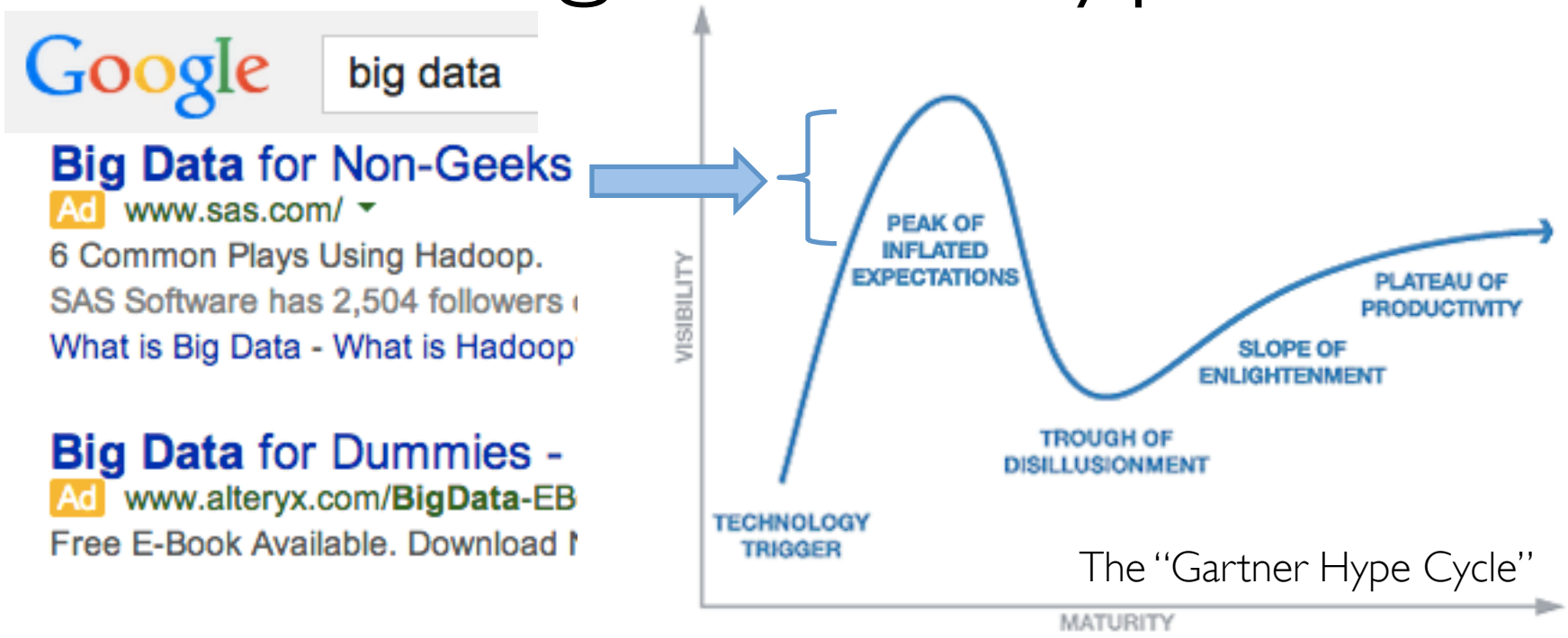
(to Funders, Investors, Deans…)



Pitching Big Data

(to Funders, Investors, Deans…)

"Big Data" Hype?



The "Gartner Hype Cycle"

Just because it's hyped, doesn't mean we can or should ignore it

WHAT'S THE BIG DEAL?

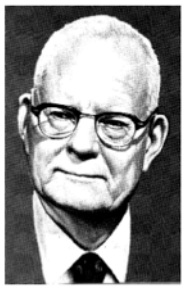
Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Demming



1958: "A Business Intelligence System"

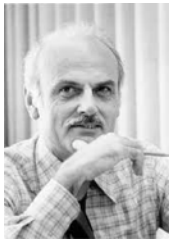
Peter Luhn

1997: "Machine Learning"

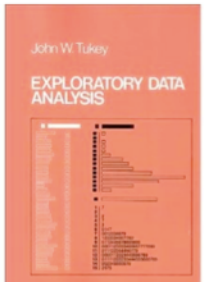
1977: "Exploratory Data Analysis

1989: "Business Intelligence"

1970: Relational Database



E.F. Codd



Howard Dresner

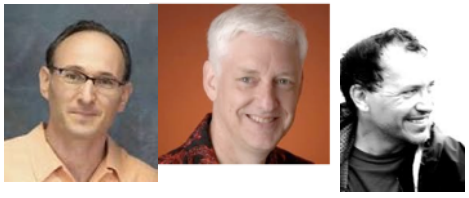
2010:"The Data Deluge"

2007:"The Fourth Paradigm"

2009:"The Unreasonable Effectiveness of Data"

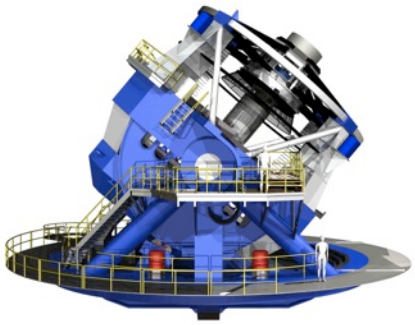
1996: Google

Google™



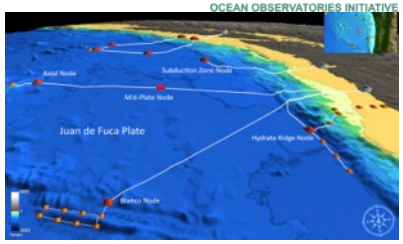
Abridged Version of Jeff Hammerbacher's timeline for CS 194, 2012

Nearly every field of endeavor is transitioning from "data poor" to "data rich"



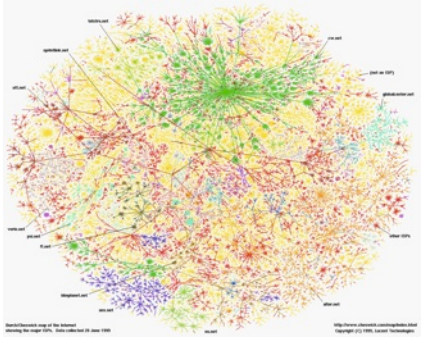
Astronomy: LSST

Physics: LHC



Oceanography: OOI

Neuroscience: EEG, fMRI



Sociology: The Web

Biology: Sequencing

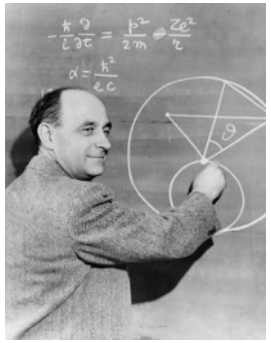
Economics: mobile, POS terminals

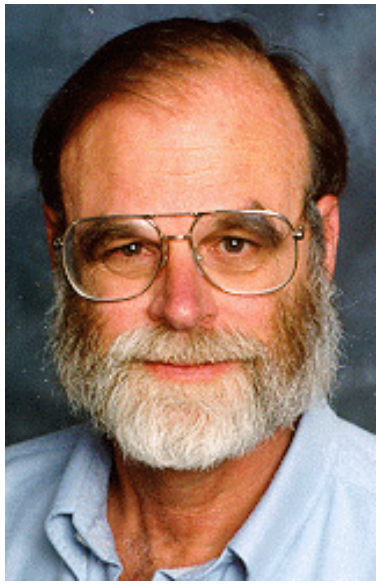
Data-Driven Medicine

Sports

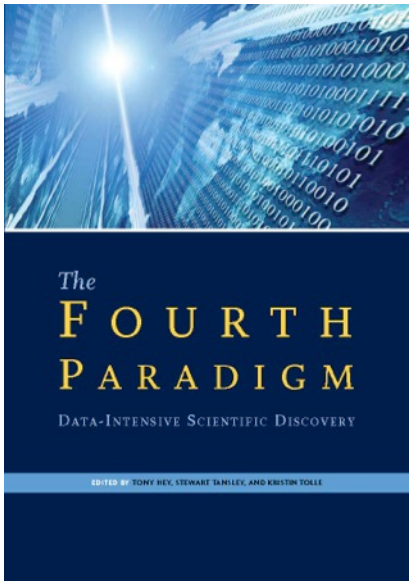
The Fourth Paradigm

1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive

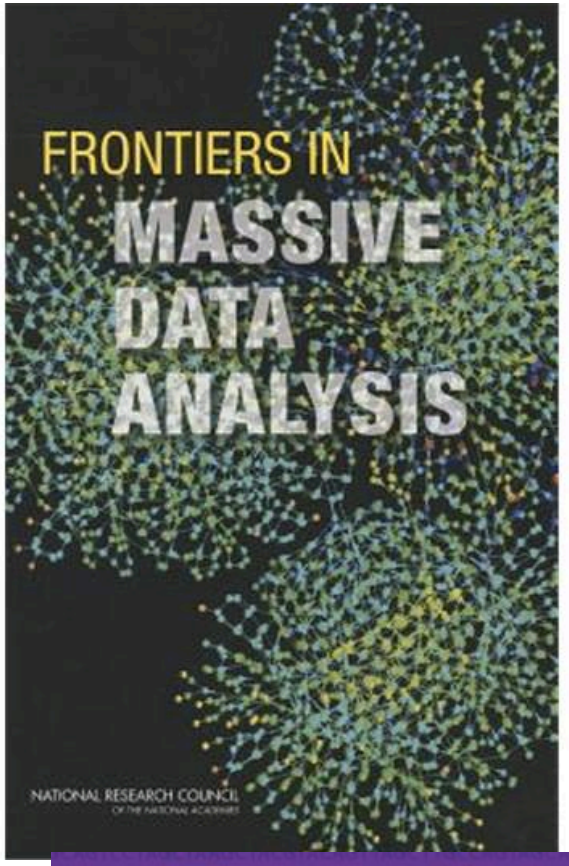




Jim Gray



It's National



Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

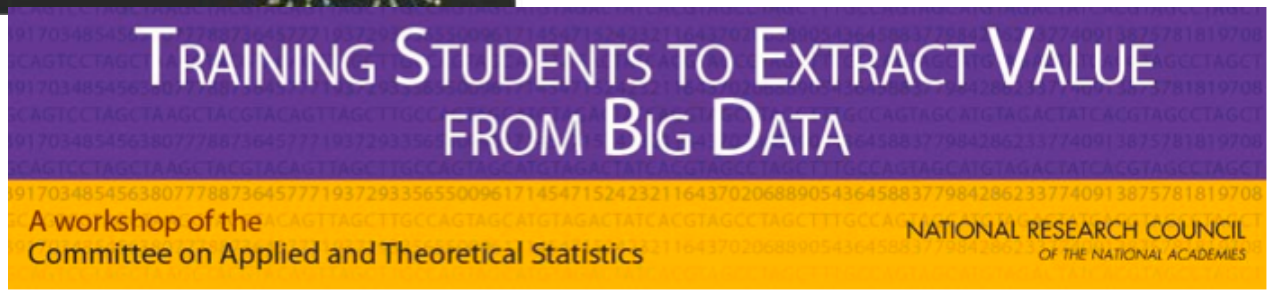
FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

NITRD

Big Data Senior Steering Group (BDSSG)



SOME DEFINITIONS
(DATA SCIENCE ≠ BIG DATA)

Data Science

"The Sexiest Job of the 21st Century"

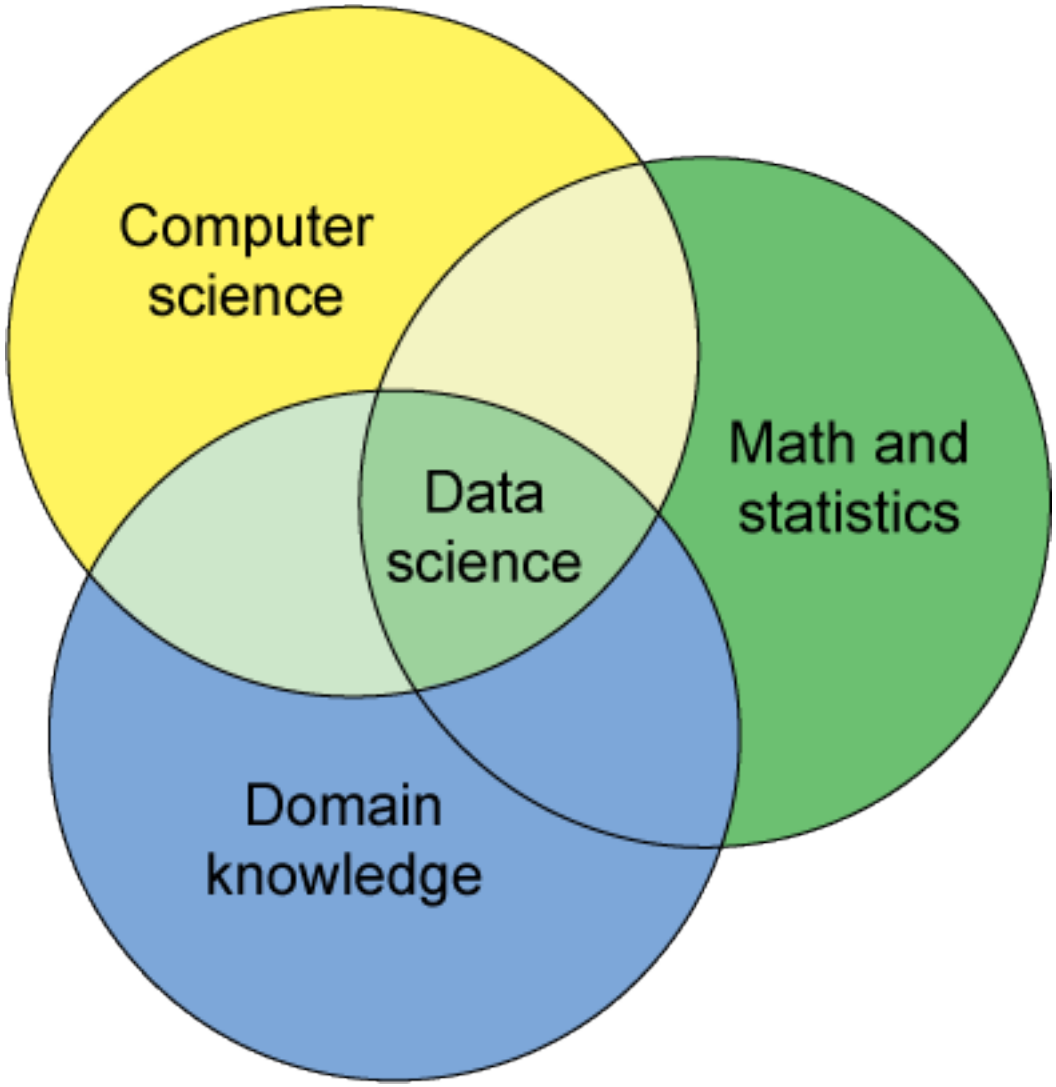
- Harvard Business Review (after Hal Varian)

"A Data Scientist is a Statistician who Lives in San Francisco." *- unknown*

"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician."

- Josh Wills, Cloudera

Data Science



based on Drew Conway, NYU

Question:

Any new intellectual content?

Data Science

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

- Hilary Mason, chief scientist at bit.ly

Big Data – A Bad Definition

Data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process.

- Dictionary.com

Big Data – The 3 V's

Volume – there are those billions and trillions

Velocity – ever more data coming at you

Variety – coming from all sorts of places

(this is an improvement, but still problem-based)

Big Data – A Better Definition?

"For a long time, we thought that Tamoxifen was roughly 80% effective for breast cancer patients. But now we know much more:

we know that it's 100% effective in 70% to 80% of the patients, and ineffective in the rest."

- Tim O'Reilly et al. "How Data Science is Transforming Health Care"

If you can get and analyze enough of the right data you can figure out who the treatment will work for.

Not just more rows, but more columns!

Big Data Technology

Regardless of your definition, one thing is clear:
The technology has fundamentally changed.

- Massively scalable storage
- Cheap, scalable processing
- Flexible schema on read vs. schema on write
- Easier integration of search, query and analysis
- Variety of interface/interaction languages
- Open source ecosystem driving innovation

AMPLAB: COLLABORATIVE BIG DATA RESEARCH

AMPLab: Integrating 3 Resources

Algorithms
- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines
- Clusters and Clouds
- Warehouse Scale Computing

People
- Crowdsourcing, Human Computation
- Data Scientists, Analysts

AMPLab Overview

amplab
UC BERKELEY

60+ Students, Postdocs, Faculty and Staff from:
Systems, Machine Learning, Databases, Security, and Networking

Franklin Jordan Stoica Culler Goldberg Joseph Katz Patterson Recht Shenker

Industry Sponsors + White House Big Data Program:
NSF CISE Expeditions in Computing and Darpa XData

Shared lab space – no offices, lots of meeting rooms, good coffee…

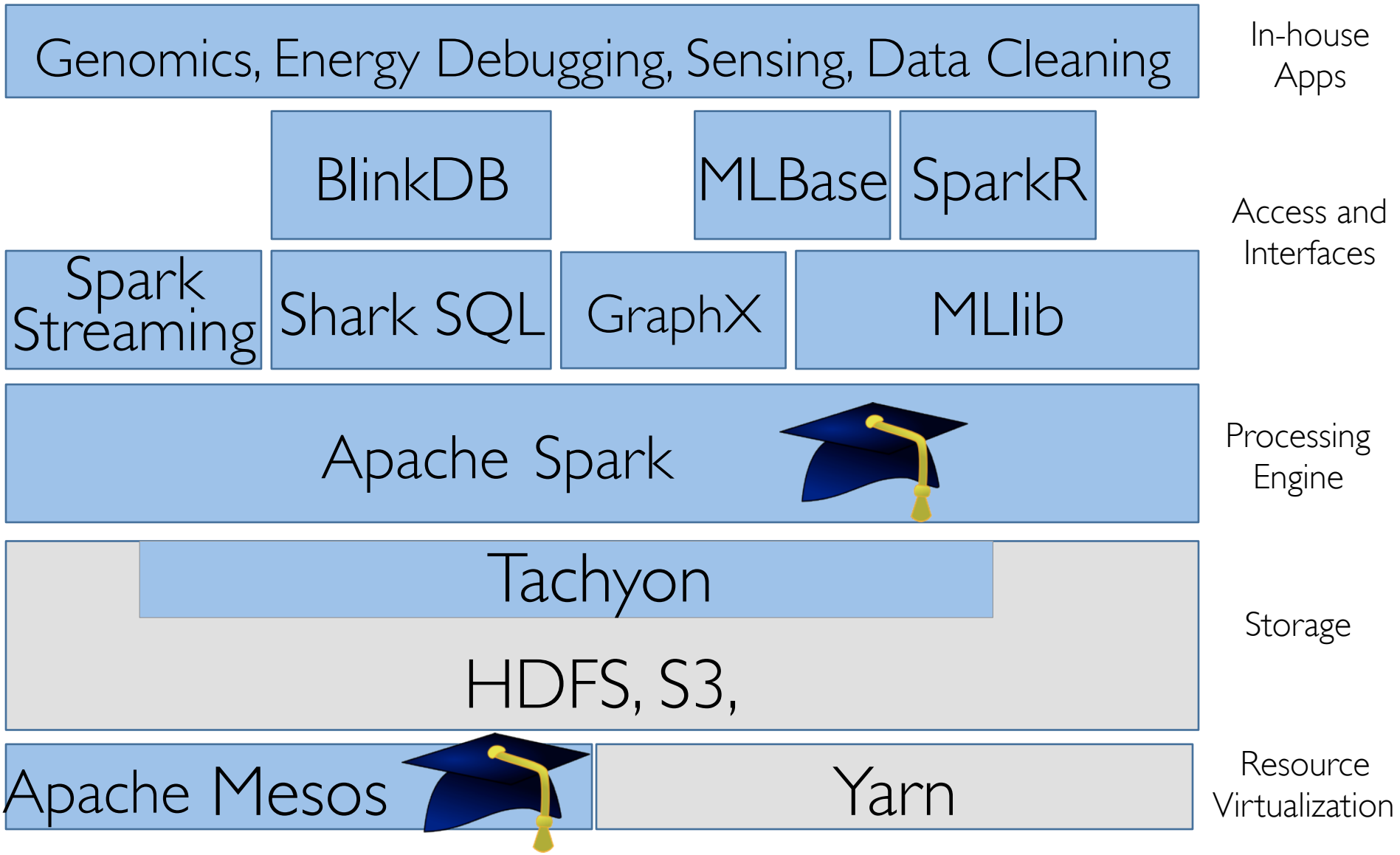
Launched: Jan 2011
Scheduled conclusion: Dec 2016

See Dave Patterson "How to Build a Bad Research Center", CACM March, 2014

A Nexus of Industrial Engagement

Regular interactions with top technologists at leading data-driven companies (e.g., twice-yearly 3-day offsite retreats; AMPCamp)

Berkeley Data Analytics Stack
(open source software)



Community Building

MeetUp on MLBase @Twitter (Aug 6, 2013)

Spark Meetup Groups

Groups	Members	Interested	Cities	Countries
17	5,711	455	16	6

Spark Summit SF (June 30, 2014)

big data / hadoop / open source

Welcome to Berkeley: Where Hadoop isn't nearly fast enough

by Derrick Harris APR. 17, 2013 - 4:19 PM PDT

WANdisco Announces Support for In-Memory Data Processing Technologies, Spark and Shark

Big Data Journal: Blog Post

Big Data: Enterprise Class Machine Learning with Spark and MLbase

New Infrastructure Aims at Economical, In-memory, Large Scale Machine Learning

big data / hadoop / open source

Spark is a really big deal for big data, and Cloudera gets it

Hortonworks®

This week we will also release a tech preview of Apache Spark for distributed in-memory processing in HDP 2.1

Vendors Align to Drive Apache Spark Adoption

Mike Vizard | IT UNMASKED

| | Share

DBMS2

December 13, 2012

Introduction to Spark, Shark, BDAS and AMPLab

DATA CENTER > CLOUD

Cloudera, MapR, IBM, and Intel bet on Spark as the new heart of Hadoop

Screeching sound heard amid Spark stampede as Cloudera changes opinion on Hive

By Jack Clark, 30 Jun 2014 Follow 5,004 followers

Spark: Open Source Superstar Rewrites Future of Big Data

BY CADE METZ 06.19.13 6:30 AM

WIRED

Strata

Making Data Work

Shark: Real-time queries and analytics for big data

Shark is 100X faster than Hive for SQL, and 100X faster than Hadoop for machine-learning

by Ben Lorica | @bigdata | Comment | November 27, 2012

Databricks announces "Certified on Spark" Program

BIG DATA // SOFTWARE PLATFORMS

NEWS

4/11/2014

11:33 AM

MapR Brings Spark In-Memory Analysis To Hadoop

GIGAOM EVENTS RESEARCH JOBS

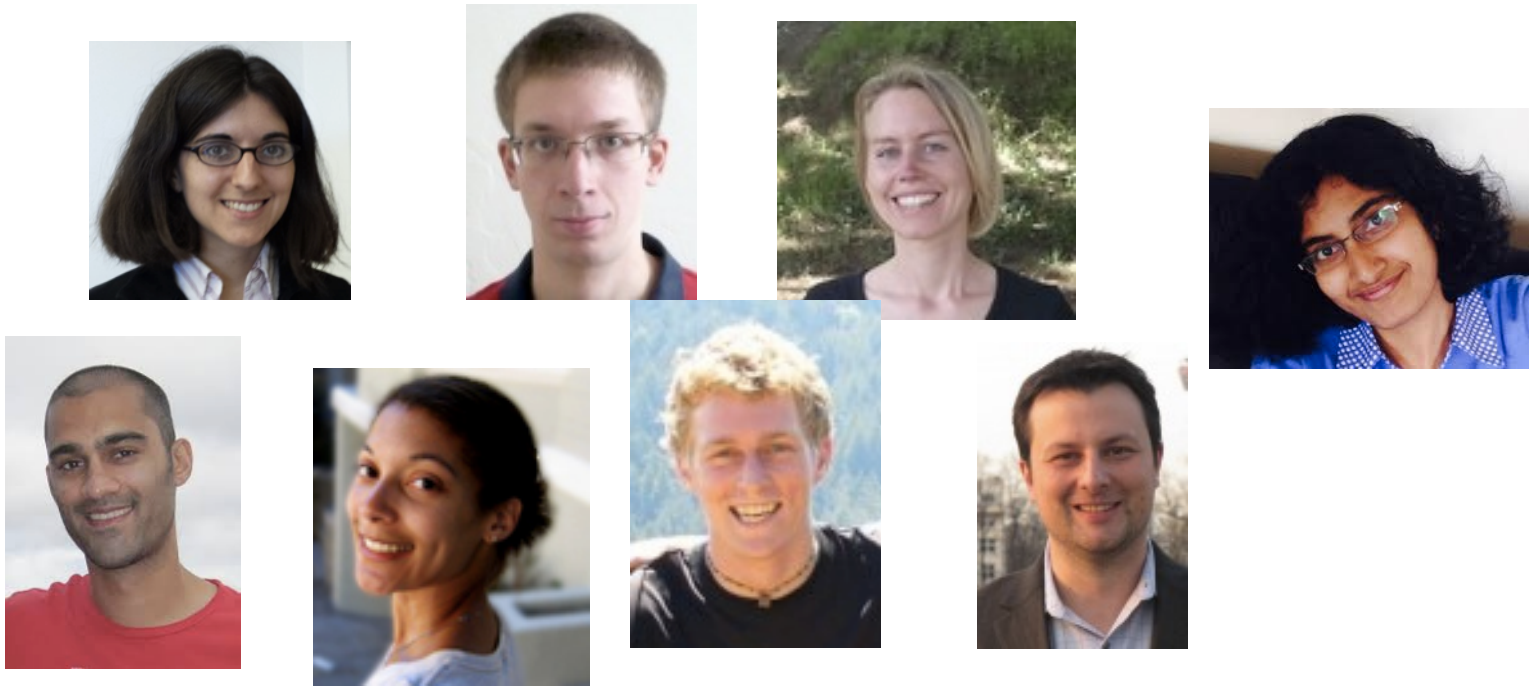
APPLE CLEANTECH CLOUD DATA EUROPE MEDIA MOBILE

big data / hadoop / open source

As MapReduce fades, Apache Spark is now a top-level project

Educational Impact

amplab UC BERKELEY



Our Students and Postdocs are in high demand in industry (obviously) and also in academia.
The alumni above have accepted faculty jobs at Brown, Harvey Mudd, MIT, Stanford, UCLA, UT Austin,…

SO WHAT DO DATA SCIENCE AND BIG DATA MEAN FOR CS?

WE COMPUTER SCIENCE PEOPLE HAVE THE PATENT ON THE BYTE AND THE ALGORITHM. WE HAVE THE PATENT ON INFORMATION.

Jim Gray, 2002

SO, OBVIOUSLY DATA SCIENCE IS ONE SUB-AREA OF COMPUTER SCIENCE – RIGHT?

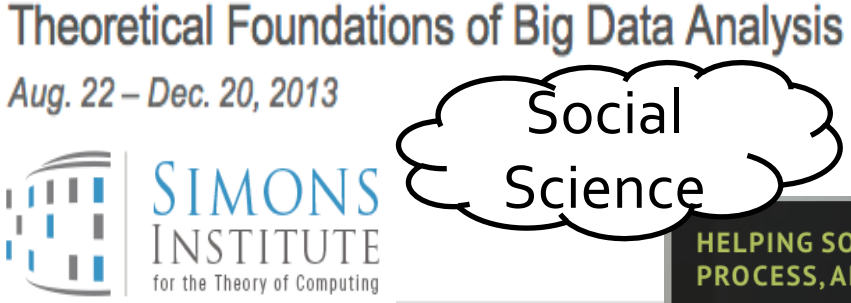
Big Data on Campus (e.g. Berkeley)

Astro

Stats

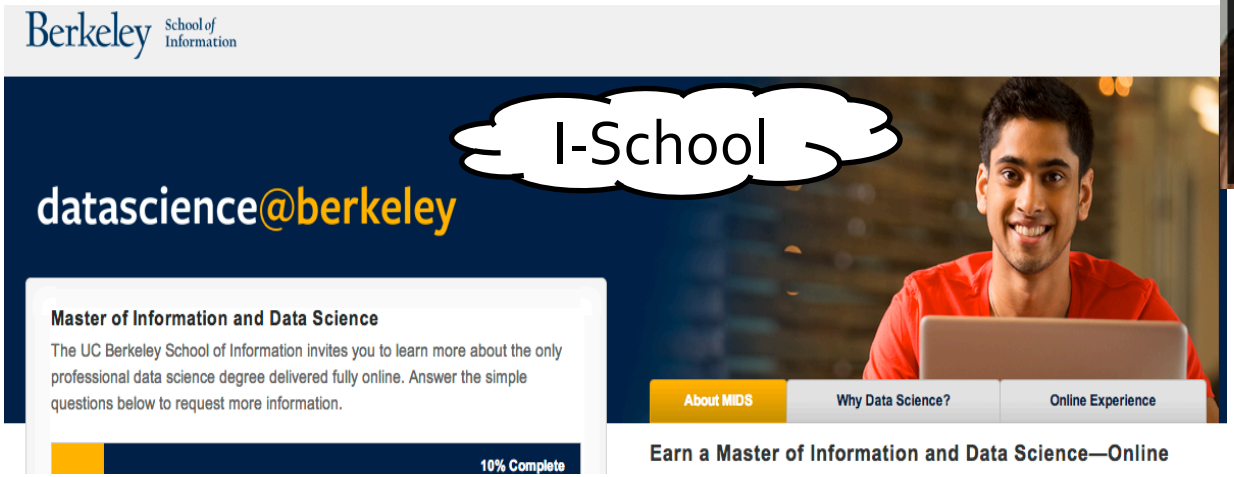
EECS

Simons



Social Science

Law



I-School

CITRIS

Cross & Multi-Campus

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

PI & Co-PI homes:

- Astronomy
- Bioengineering
- Comp Sci
- Libraries
- Math
- Neuroscience
- Physics
- Public Policy
- Social Sciences
- Statistics

An accelerator for data driven discovery
An agent of change in the modern university as Data Science takes hold
An incubator for the next generation of Data Science technology and practice

NEW YORK UNIVERSITY

Berkeley UNIVERSITY OF CALIFORNIA

UNIVERSITY of WASHINGTON

Masters Degrees are Everywhere

MS or MBA in [Business | Marketing | Predictive | “ “] Analytics

Information Systems Mgmt w/concentration in Business Intelligence and Data Analytics

MS in CS w/concentration in Machine Learning or Info Mgmt and Analytics

MEng w/Concentration in Data Science Systems

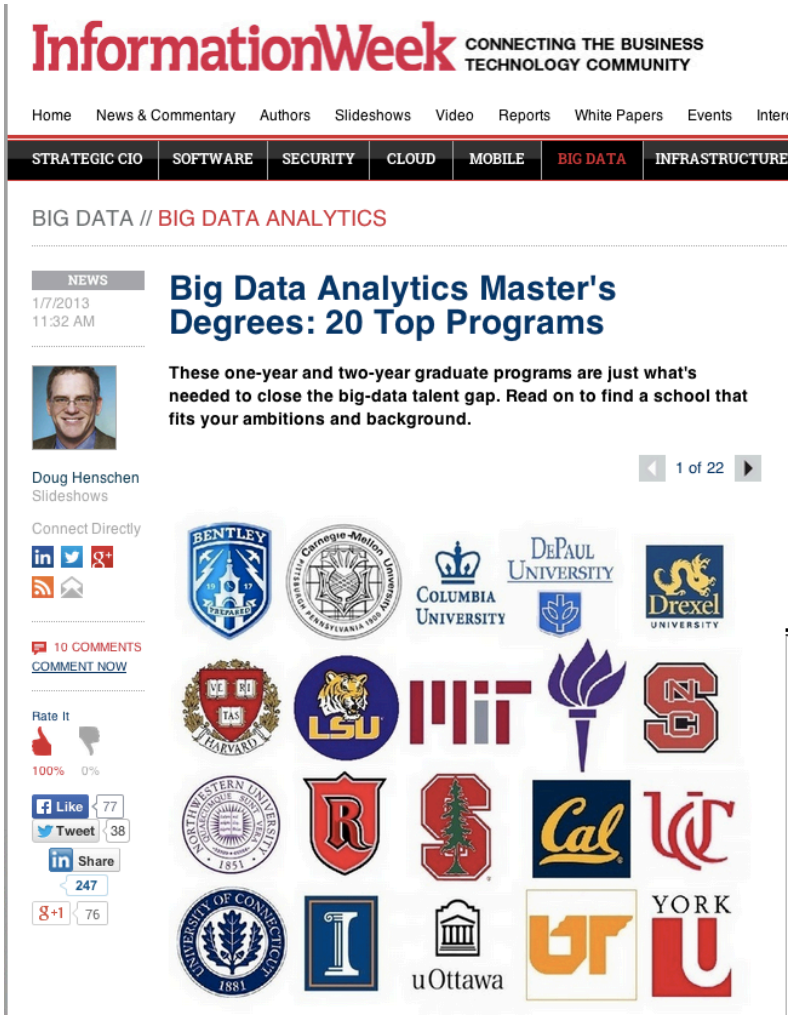
MS in Computational Science and Eng

M of Biz and Sci in OR and Biz Analytics

MS in Biz Analytics and Project Mgmt

MS in Statistics: Analytics Concentration

Masters of Data Science (in various ways)



WHY CARE?

1970's: EE + Math → Computer Science

2010's: CS + Stats + ?? → Data Science

Is something fundamental emerging here?

If so, what role should CS play?

Are new principles needed? Or just pick & choose?

What do we bring to the table? What are we missing?

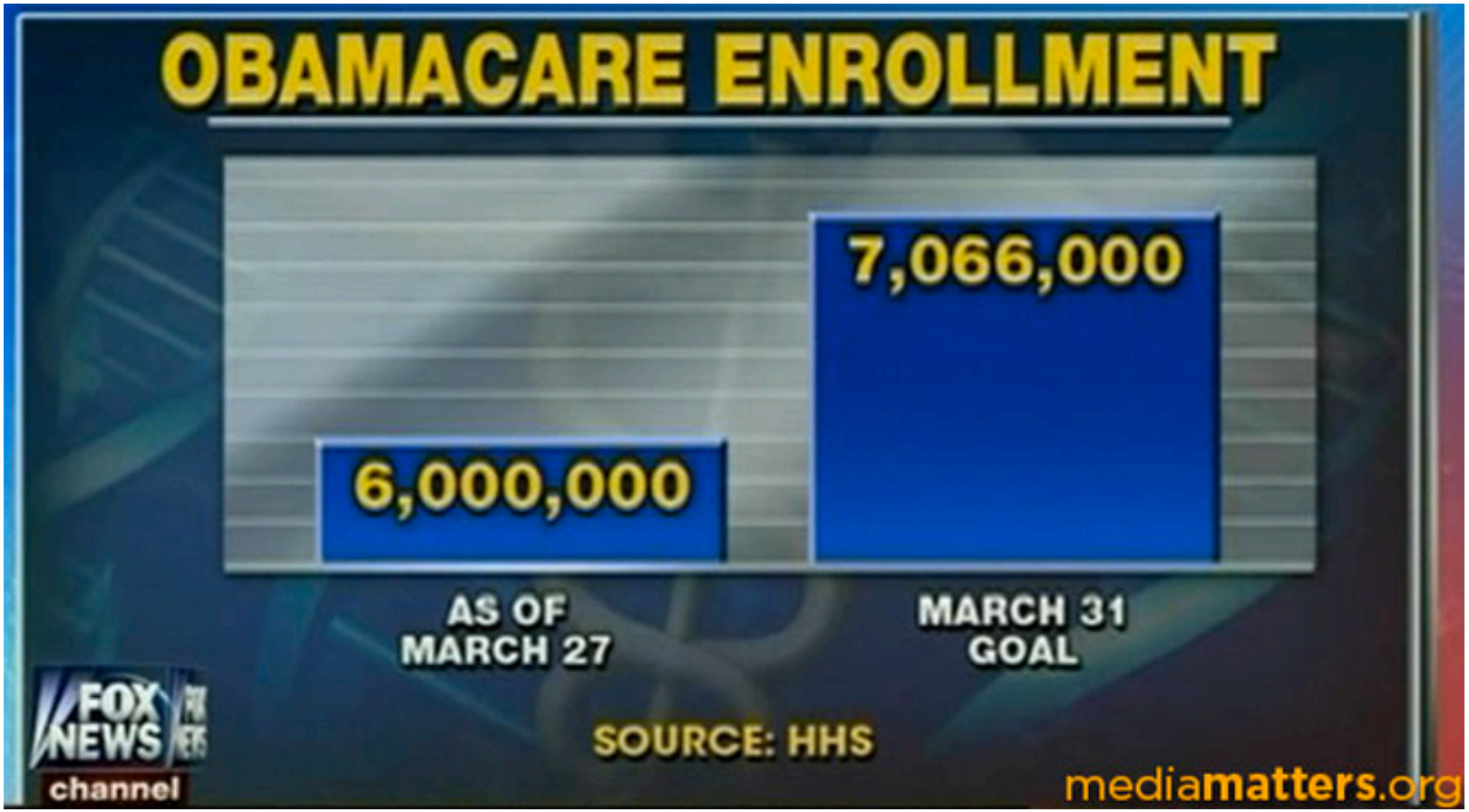
How, What and Whom to teach?

- Computational Thinking → Analytical Thinking?

How best to engage

- At Campus Level
- Nationally and beyond

Another Reason it Matters



Note: Fox News did broadcast a corrected version of this graph with a real y-axis the following day, acknowledging the graph as a mistake.

Data Science: Still Work to Do…