Why the Big Deal About "Big Data"?

Ed Lazowska

Bill & Melinda Gates Chair in
Computer Science & Engineering

Founding Director, eScience Institute

University of Washington

Washington Innovation Summit

November 2012

http://lazowska.cs.washington.edu/bigdata.pdf

This morning …

- Overview presentation
  - Ed Lazowska
- Panel discussion + Q&A on "big data" and the transformation of businesses
  - Christian Chabot (Tableau Software), Mike Fridgen (Decide), Cameron Myhrvold (Ignition Partners), Ruben Ortega (Google)

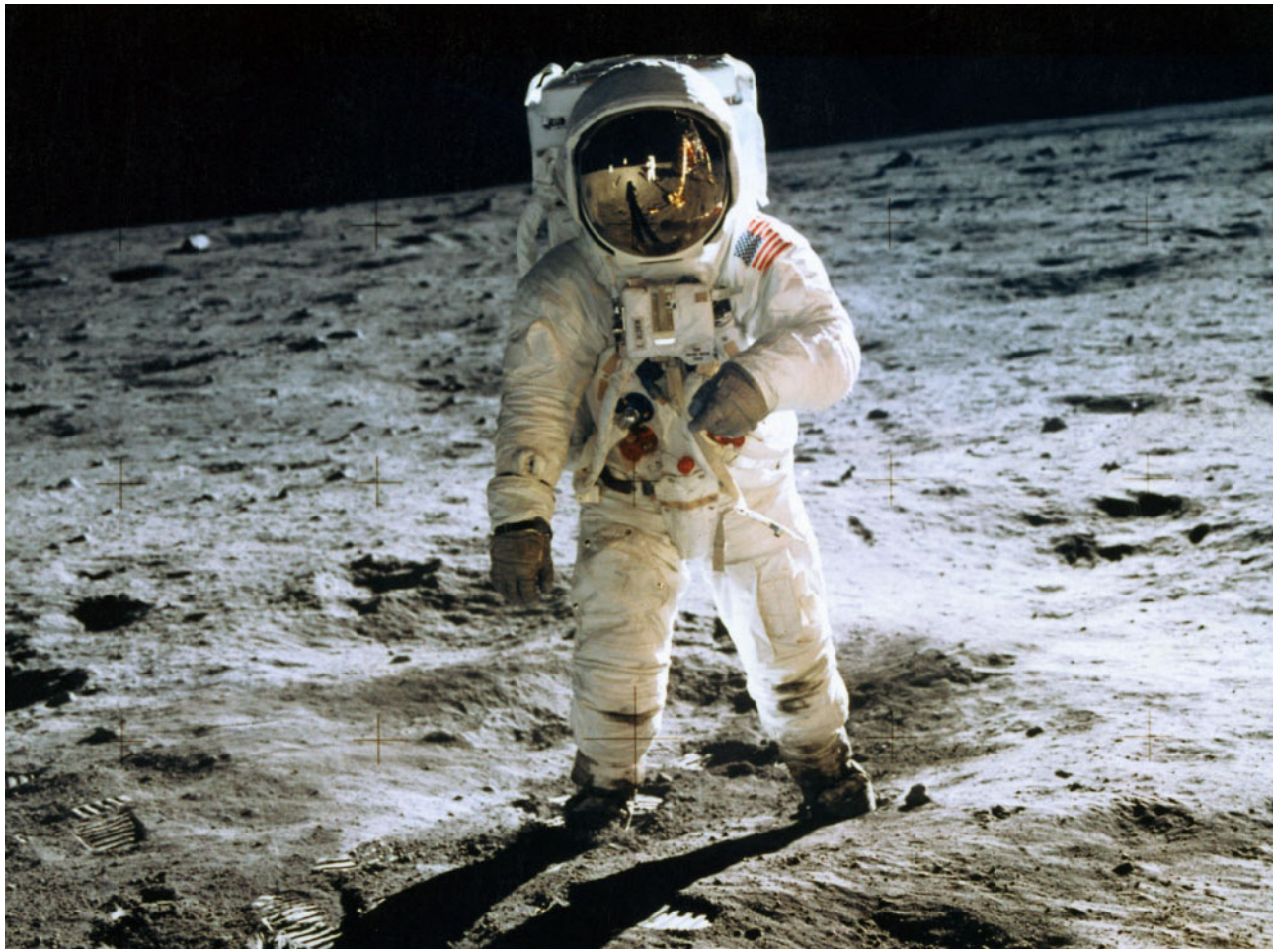
- Break

- Short presentations + Q&A on "big data" and national challenges
  - Shwetak Patel (UW: smart home), Carl Imhoff (PNNL: smart grid), Mike Kellen (Sage Bionetworks: smart discovery)

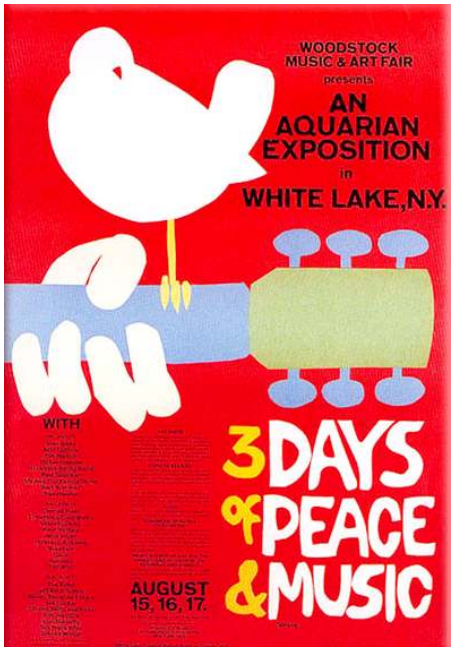
- Presentation on related workforce issues
  - Bill Richter (Isilon Storage Division, EMC)

Forty years ago …

Credit: Peter Lee, Microsoft Research

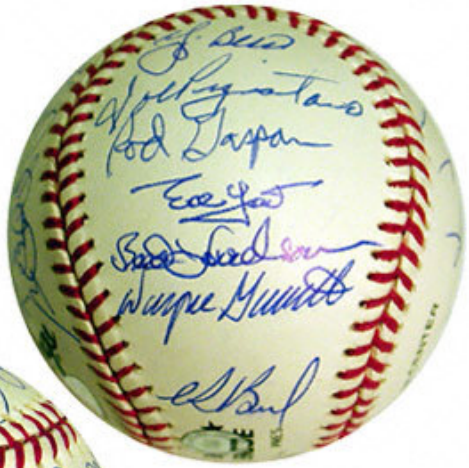



1969


WOODSTOCK
MUSIC & ART FAIR
presents
AN
AQUARIAN
EXPOSITION
in
WHITE LAKE, N.Y.
WITH
3 DAYS
of PEACE
& MUSIC
AUGUST
15, 16, 17.


1969

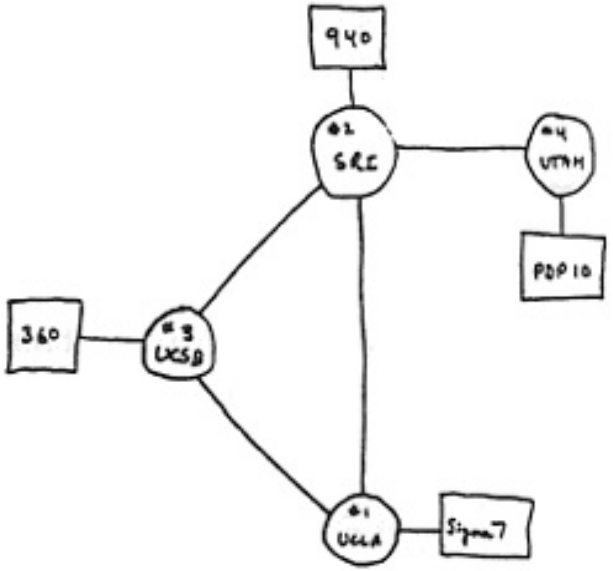

1969 WORLD SERIES
NEW YORK METS




WORLD
CHAMPIONS
NEW YORK METS
1969
RYAN
30


1969


Mets



THE ARPA NETWORK
DEC 1969
4 NODES

29OCT69	2100	LOADED OP. PROGRAM FOR BEN BARKER BBN	CSK
	22:30	Talked to SRI Host to Host	CSK
		Left IMP program running after sending a host dead message to IMP.	CSK

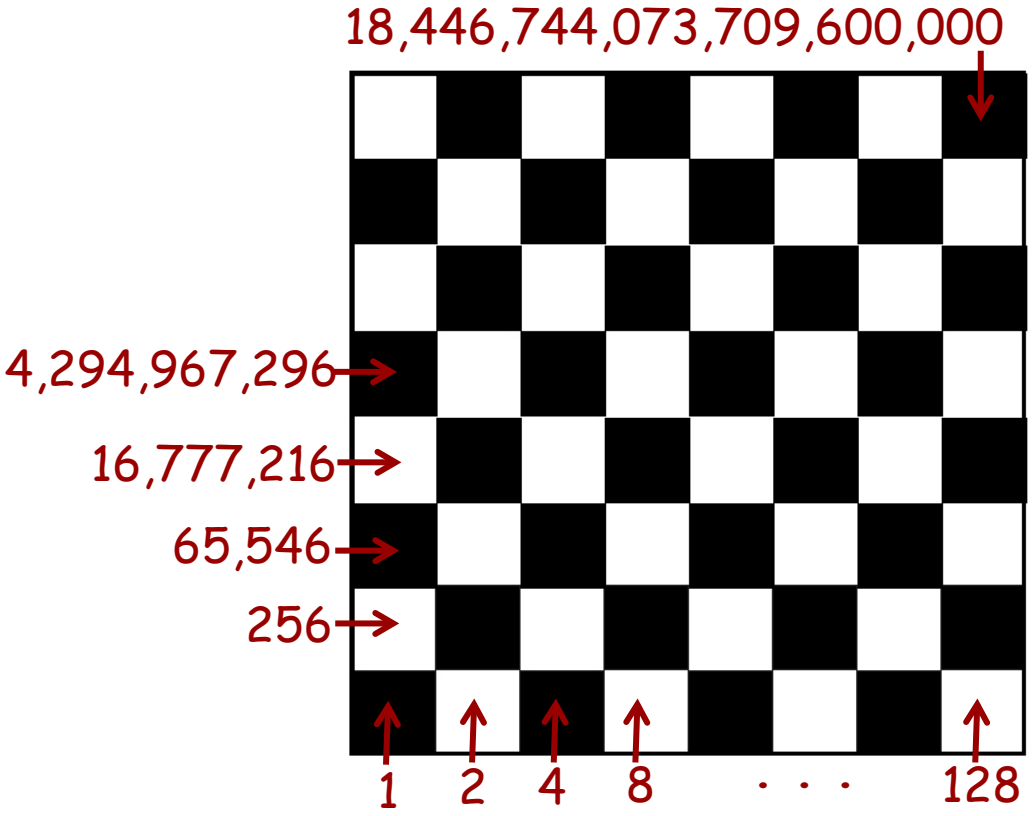
With forty years hindsight, which had the greatest impact?

- Unless you're big into Tang and Velcro (or sex and drugs, or fan futility), the answer is clear …

- And so is the reason …

EXPONENTIALS 'R' US

Exponentials are rare – we're not used to them, so they catch us unaware



Every aspect of computing has experienced exponential improvement

- Processing capacity
- Storage capacity
- Network bandwidth
- Sensors
- (Astonishingly, even algorithms in some cases!)

You can exploit these improvements in two ways

- Constant capability at exponentially decreasing cost
- Exponentially increasing capability at constant cost

Today, these exponential improvements are enabling the "big data" revolution

- A proliferation of sensors
  - Think about the sensors on your phone
- More generally, the creation of almost all information in digital form
  - It doesn't need to be transcribed in order to be processed
- Dramatic cost reductions in storage
  - You can afford to keep all the data
- Dramatic increases in network bandwidth
  - You can move the data to where it's needed

- Dramatic cost reductions and scalability improvements in computation
  - With Amazon Web Services, or Google App Engine, or Microsoft Azure, 1000 computers for 1 day cost the same as 1 computer for 1000 days!
- Dramatic algorithmic breakthroughs
  - Machine learning, data mining – fundamental advances in computer science and statistics

"Big data" is how we're putting the "smarts" in …

- Smart homes
- Smart cars
- Smart health
- Smart robots
- Smart science (confronting the data deluge)
- Smart crowds and human-computer systems
- Smart interaction (virtual and augmented reality)

xconomy
Business + Technology in the Exponential Economy

Computer Science & Engineering

UNIVERSITY of WASHINGTON

JEFF HEER

CARLOS GUESTRIN

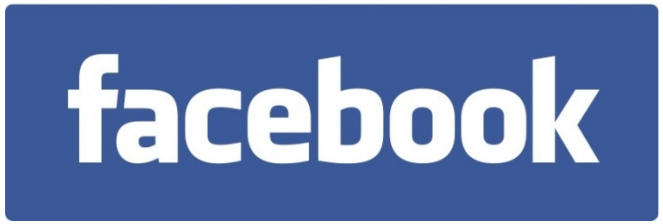
EMILY FOX

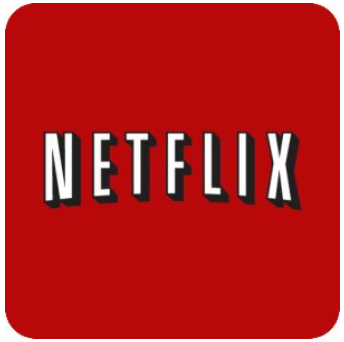
BEN TASKAR

Senior hires catapult the University of Washington in machine learning and "big data"

Some examples of "big data" in action

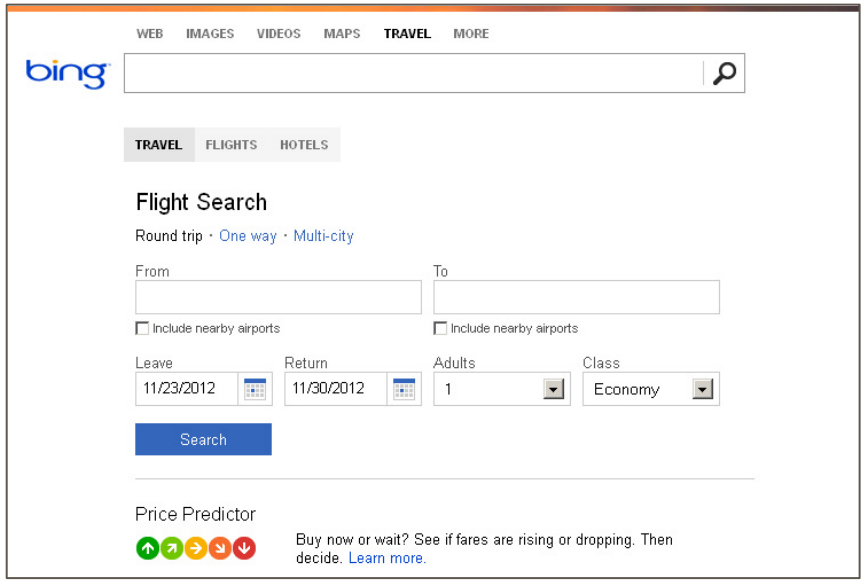
- Collaborative filtering





- Fraud detection

Price prediction



Hospital re-admission prediction

caradigm™

a Microsoft | GE Healthcare company

Products | Partners | About

AMALGA - REDUCE READMISSIONS

Avoid preventable readmissions, reduce costs and deliver higher quality care

Service Offerings

- Improve Quality of Care
- Reduce Readmissions
- Ease Care Transitions
- Manage Chronic Disease

Background

With payers implementing penalties for readmissions, it's critical to start addressing readmissions risks today. Solutions that help healthcare enterprises count last month's readmissions are no longer sufficient. Enterprises need to know which patients in their hospitals today are at risk for being readmitted within 30 days of discharge, so they can take action and address those risks before the patient walks out the door.

Amalga helps healthcare organizations proactively identify inpatient and Emergency Department (ED) patients at risk for readmissions and helps them take action to avoid preventable readmissions, reduce costs and deliver higher quality care – today and tomorrow.

Overview

By using predictive modeling technologies, Amalga can help reduce preventable readmissions by enabling healthcare delivery organizations to:

- Effectively define and monitor patient groups across the enterprise
- Use data collected in Amalga to predict readmission probability based on a given hospital's historical data
- Proactively manage at-risk patients throughout their stay and at discharge
- Access patterns in key indicators to identify and address root causes of readmissions

Features and Benefits

With Amalga, organizations can:

- Actively identify and track patient groups
- Integrate disparate systems and identify patient cohorts based on key characteristics
- Use predictive modeling technologies to help identify patients at risk for readmission
- Analyze readmission patterns and monitor 30-day inpatient and 72-hour ED readmissions
- View simplified reports that support identification and addressing of root causes sooner

- Travel time prediction under specific circumstances

Microsoft Research

Search Microsoft Research

Videos

Home | Our Research | Connections | Careers | Hub

Worldwide Labs | Research Areas | Research Groups

> Projects > Predictive Analytics for Traffic

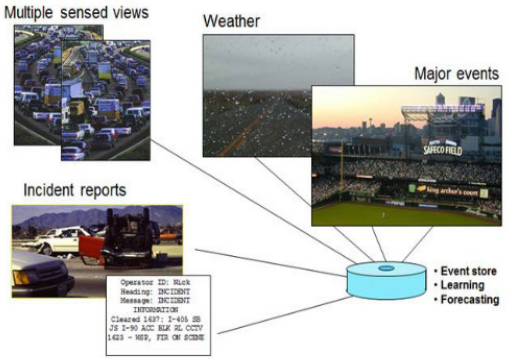
Predictive Analytics for Traffic

Machine learning and intelligence for sensing, inferring, and forecasting traffic flows

Machine learning and intelligence are being applied in multiple ways to addressing difficult challenges in multiple fields, including transportation, energy, and healthcare. Research scientists at Microsoft Research have been engaged in efforts in all of these areas. We focus on multiyear efforts at Microsoft Research to infer and forecast the flows of traffic. The work leverages machine learning to build services that make use of both live streams of sensed information and large amounts of heterogeneous historical data. This has led to multiple prototypes and real-world services such as traffic-sensitive directions in Bing Maps. Focused work in this realm also stimulated new efforts in related areas, such as privacy and routing.

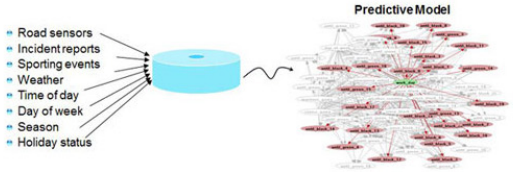
Predicting Traffic Jams and Flows: JamBayes

Machine learning methods have been applied to create methods that provide estimates of flows inferences about current and future traffic flows. Research on the JamBayes project, started in 2002, was framed by the frustrations encountered with navigating through Seattle traffic, a region that has seen great growth amidst slower changes to the highway infrastructure.



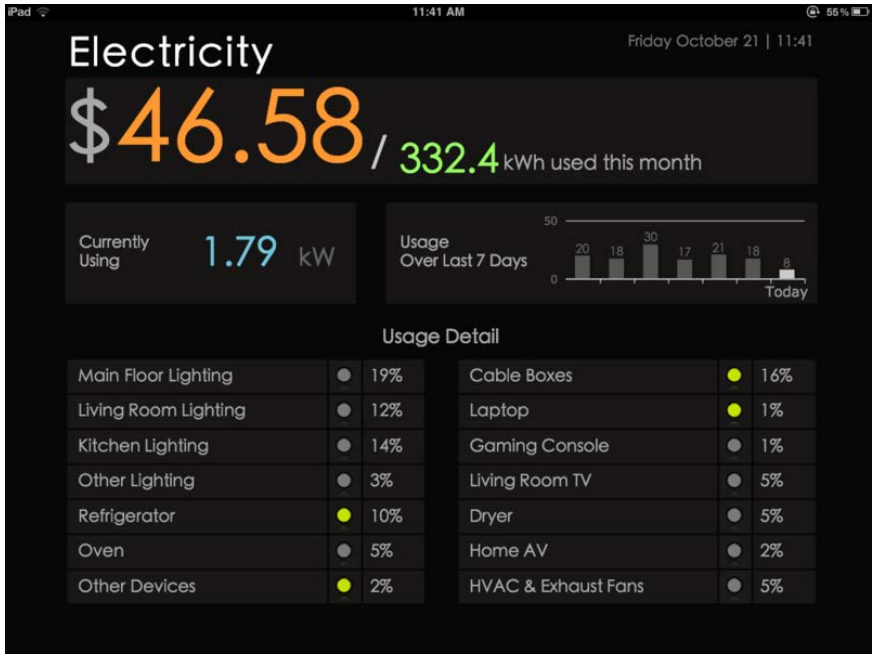
JamBayes focused on learning to forecast flows from multiple streams of information.

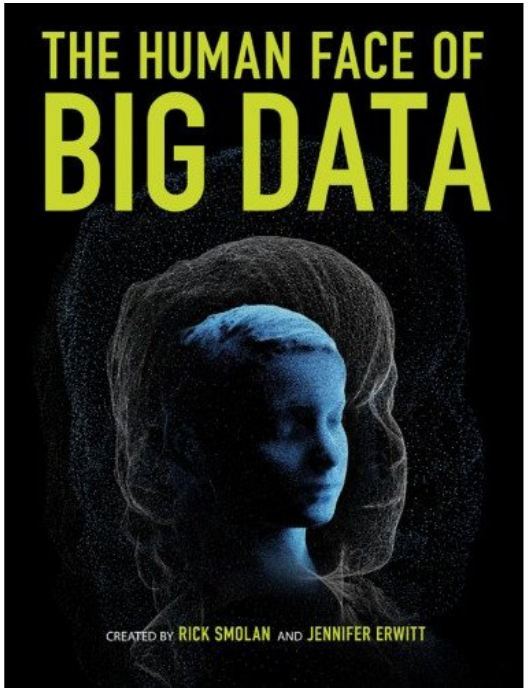
The challenge was to predict the future of traffic flow: How long would it be until a current traffic jam on the highway system of Seattle would melt? How long until open flows on different segments of the highway system of Seattle would become clogged? The idea was to combine heterogeneous streams and histories of information to make these predictions. These streams included multiple years of different types of data, including sensed highway data, reports of accidents throughout the highway system, weather, and major regional events such as Mariners and Supersonics games.

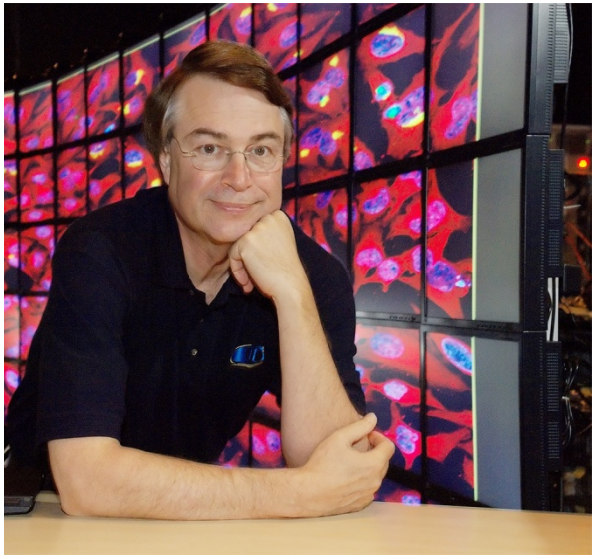


Sports

Home energy monitoring







Larry Smarr, UCSD

John Guttag & Collin Stultz, MIT

Google self-driving car

Gordon Bell, Microsoft Research

Speech recognition

Siri. Beta
Your wish is its command.

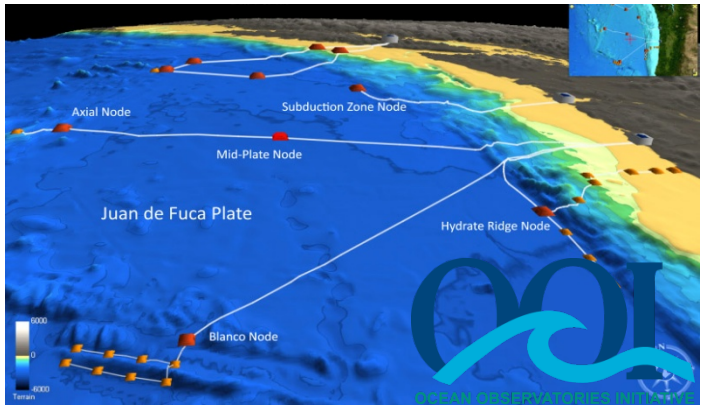
Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more.* Ask Siri to do things just by talking the way you talk. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

Machine translation

- Speech -> text
- Text -> text translation
- Text -> speech in speaker's voice

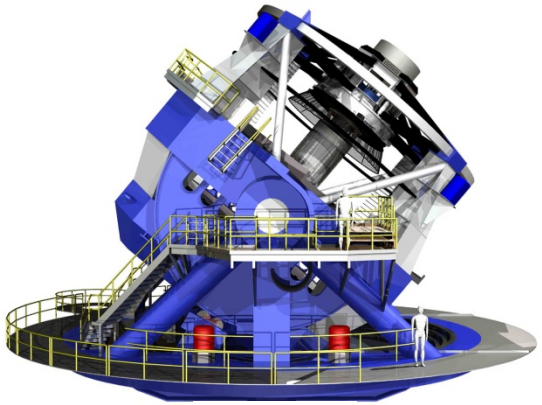
http://www.youtube.com/watch?v=Nu-nlQqFCKg&t=7m30s
7:30 – 8:40

Scientific discovery



Ocean Observatories Initiative


Gene Sequencing


Large Synoptic Survey Telescope


Large Hadron Collider

eScience Institute

Presidential campaigning

2012 ELECTION

Inside the Secret World of the Data Crunchers Who Helped Obama Win

By Michael Scherer | Nov. 07, 2012 | 268 Comments

Like 16k Tweet 6,662 +1 656 Share 1,479

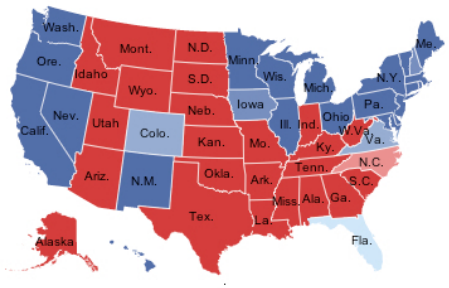
Electoral forecasting

DATA MINING

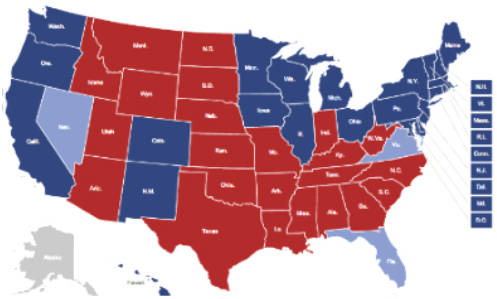
Nate Silver's Sweep Is a Huge Win for 'Big Data'

The data utopia awaits.

By Nitasha Tiku 11/07 11:10am

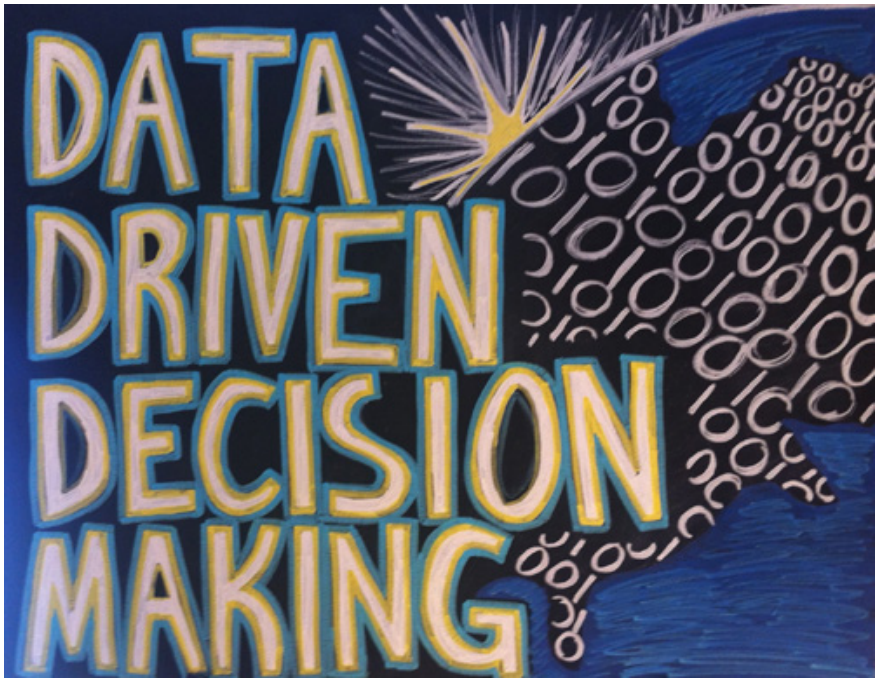


Nate Silver's Map



The Actual Map

- *Real* data-driven decision-making (vs. MBA baloney) for every sector!



Innovative opportunities abound

- Systems

cloudera

- Specialist tools

- Non-specialist tools

Infinite applications

- All tied to the cloud

Is this a great time, or what?!?!

http://lazowska.cs.washington.edu/bigdata.pdf