Stats5 Seminar: Introduction to Data Science

Winter 2018

Professor Padhraic Smyth
Departments of Computer Science and Statistics
University of California, Irvine

Outline

- Class organization and topics
- History of data analysis
- Data science and real-world applications
- The Data Science Major
- Limitations of what we can do with data

Class Organization

- Meet weekly for 40 minute seminar with 5-10 minute discussion
- 8 topics (with guest speakers), weeks 2 through 9
  - You are encouraged to ask questions during and after the talks
- Intro and wrap-up talks in weeks 1 and 10
- Class Web site is at www.ics.uci.edu/~smyth/courses/stats5
  - Slides and related materials will be posted during the quarter

Schedule of Lectures

Date	Speaker	Department Or Organization	Topic
Jan 9	Padhraic Smyth	Computer Science	Introduction to Data Science
Jan 16	Padhraic Smyth	Computer Science	Machine Learning
Jan 23	Michael Carey	Computer Science	Databases and Data Management
Jan 30	Sameer Singh	Computer Science	Statistical Natural Language Processing
Feb 6	Zhaoxia Yu	Statistics	An Introduction to Cluster Analysis
Feb 13	Erik Sudderth	Computer Science	Computer Vision and Machine Learning
Feb 20	John Brock	Cylance, Inc	Data Science and CyberSecurity
Feb 27	Video Lecture (Kate Crawford)	Microsoft Research and NYU	Bias in Machine Learning
Mar 6	Matt Harding	Economics	Data Science in Economics and Finance
Mar 13	Padhraic Smyth	Computer Science	Review: Past and Future of Data Science

Submission of Review Forms (Weeks 2 to 10)

- Submit Review forms for Lectures 2 through 10

- Review forms will be available online at the start of each class
  - A few relatively short questions based on the lecture that day
  - Needs to be submitted to EEE by noon for each lecture
  - Bring your laptop or other device

- Requirements to pass the class
  - Attend and submit review form for least 8 lectures for weeks 2 through 10

    (allowed to miss one if you need to for some reason)

- No final exam: pass/fail based on attendance and review forms

Academic Integrity

- The review form you submit each week must be
  (a) written by you, and
  (b) written during the lecture that week

- Failure to adhere to this policy may result in failing the class

- It is the responsibility of each student to be familiar with UCI's Academic Integrity Policies and UCI's definitions and examples of academic misconduct. See the class Web site for additional info.

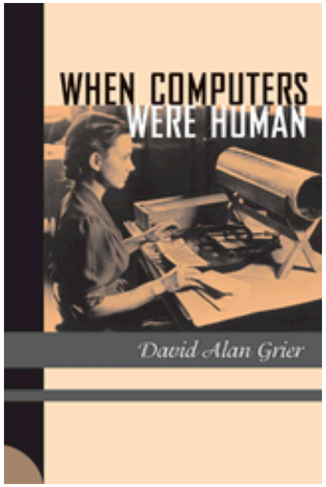
A BRIEF HISTORY OF DATA ANALYSIS AND COMPUTING

Computers and Data

The historical meaning of the term "computer":

"one who computes" (i.e., a person)

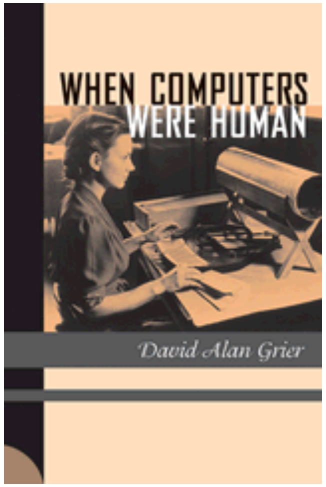
Since the 1700's, statisticians have been using "computers" to analyze data – so its not a new idea



Computers and Data

The historical meaning of the term "computer":

"one who computes" (i.e., a person)



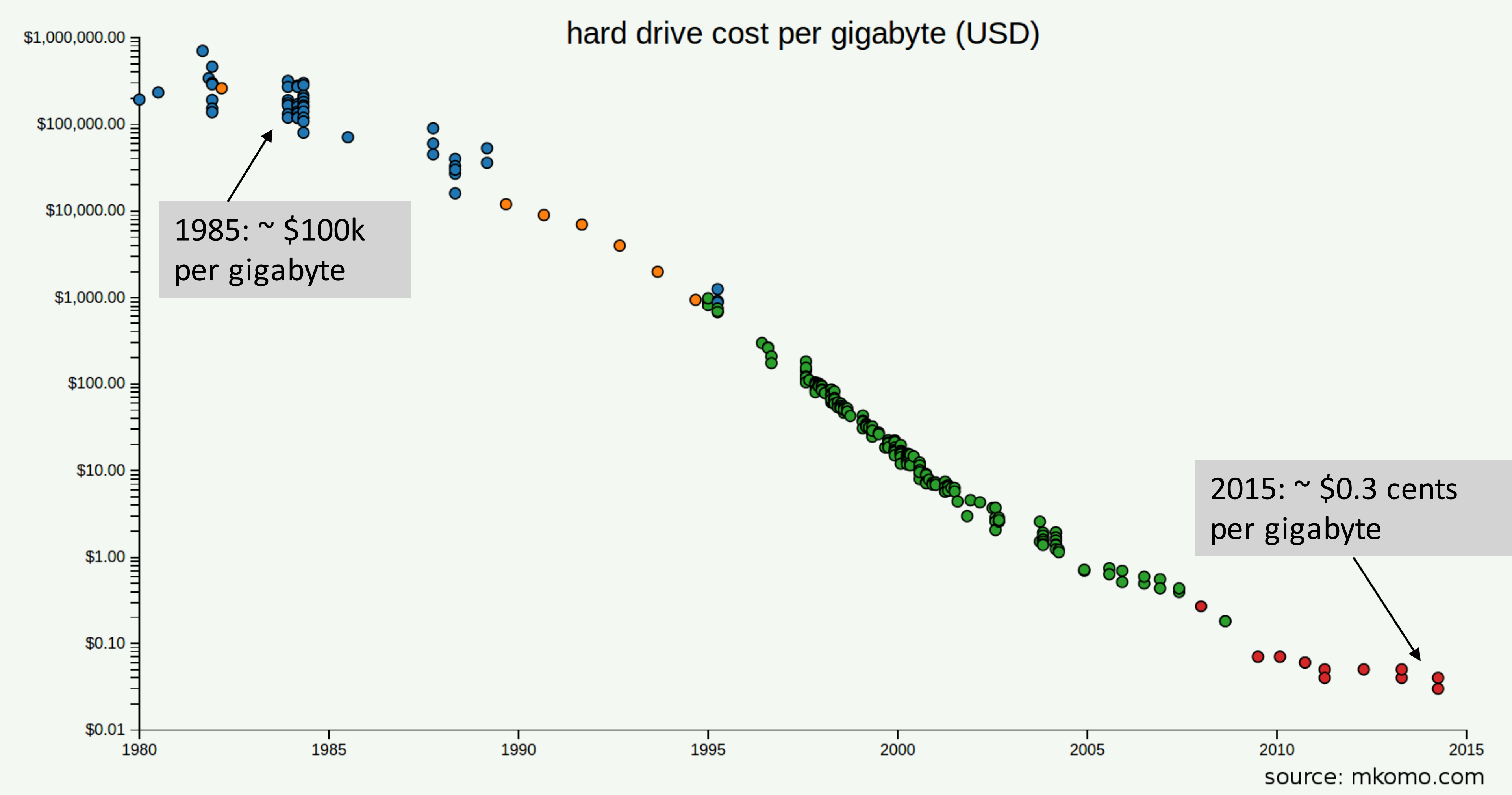
Since the 1700's, statisticians have been using "computers" to analyze data – so its not a new idea

For example, Karl Pearson, one of the founders of statistics, directed a team of "computers" in his lab in London around the early 1900's

…..but for many years, "computers" could only work on relatively small problems

Statistics and Modern Computing

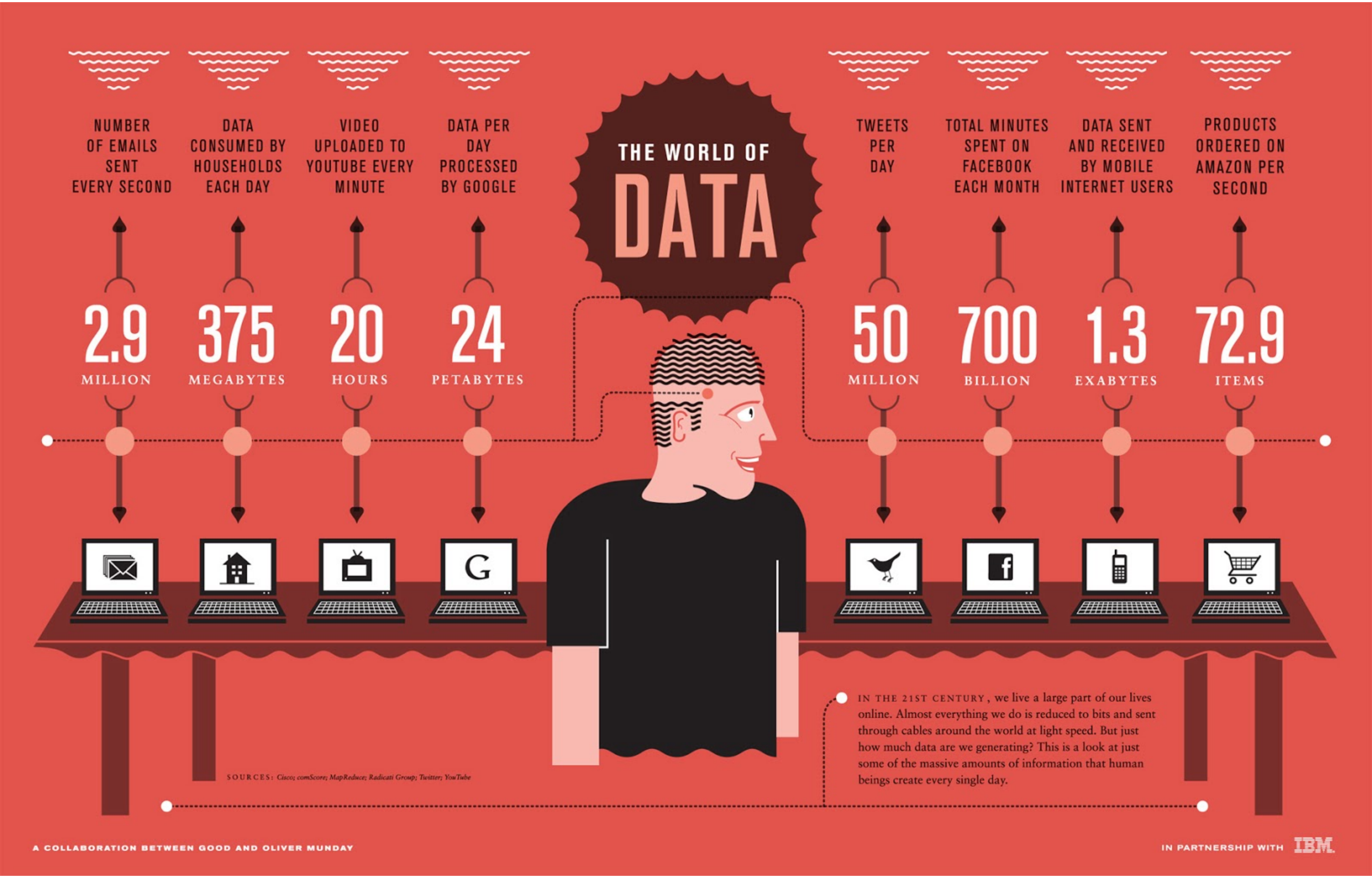
- Post World War II
  - Increasing use of computing to solve algorithmic aspects of statistical analyses
- 1960’s
  - Development of statistical computing and exploratory data analysis
- 1980’s
  - Computing allowed statisticians to explore more flexible models
  - Increase in use of “non-parametric” techniques and simulation methods
- 1990’s
  - Development of “machine learning” – very flexible predictive modeling techniques developed in computer science
- Today
  - Data science = computing + statistics + applicatinos



1985: ~ $100k per gigabyte

2015: ~ $0.3 cents per gigabyte

source: mkomo.com

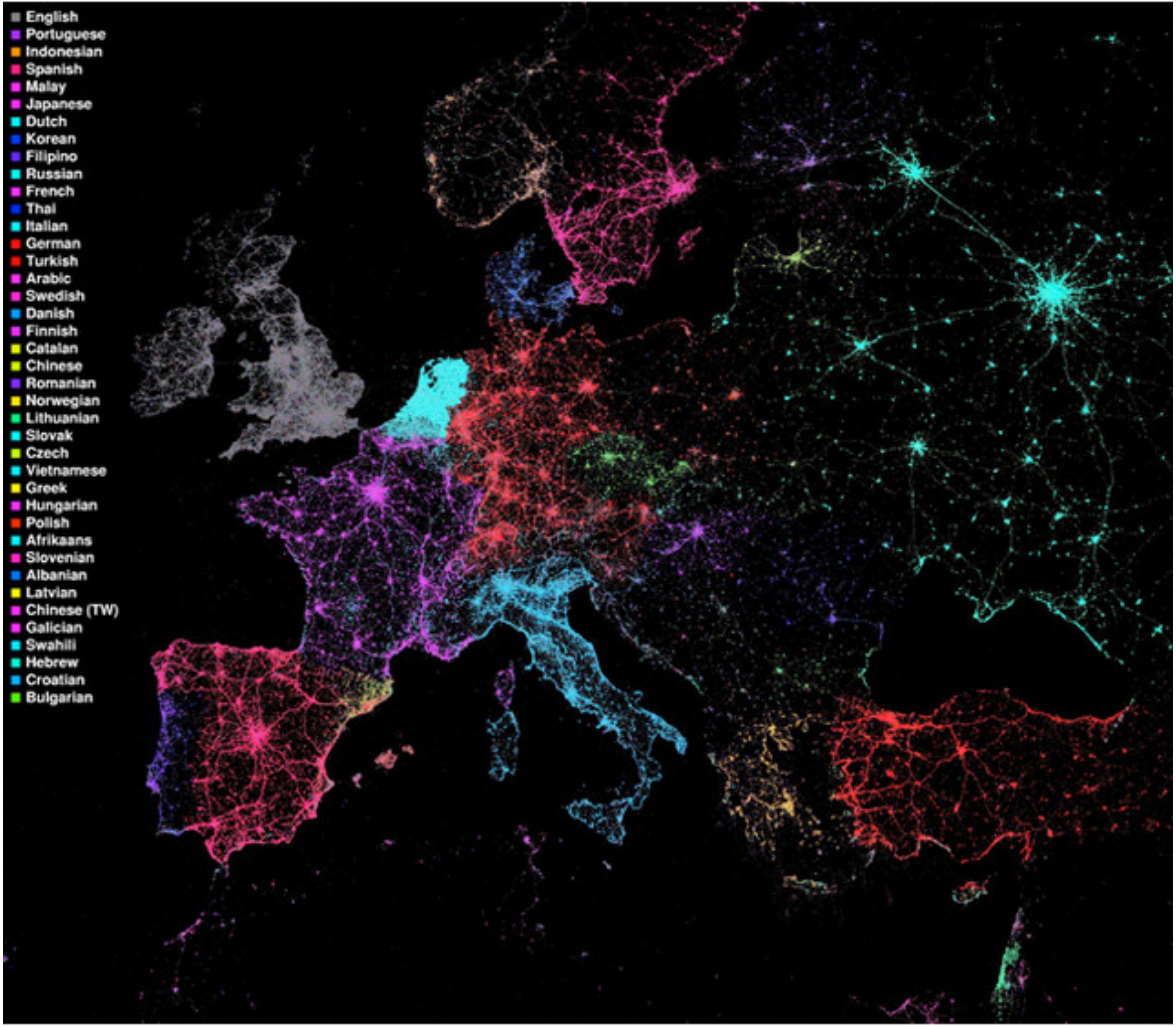


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
f
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

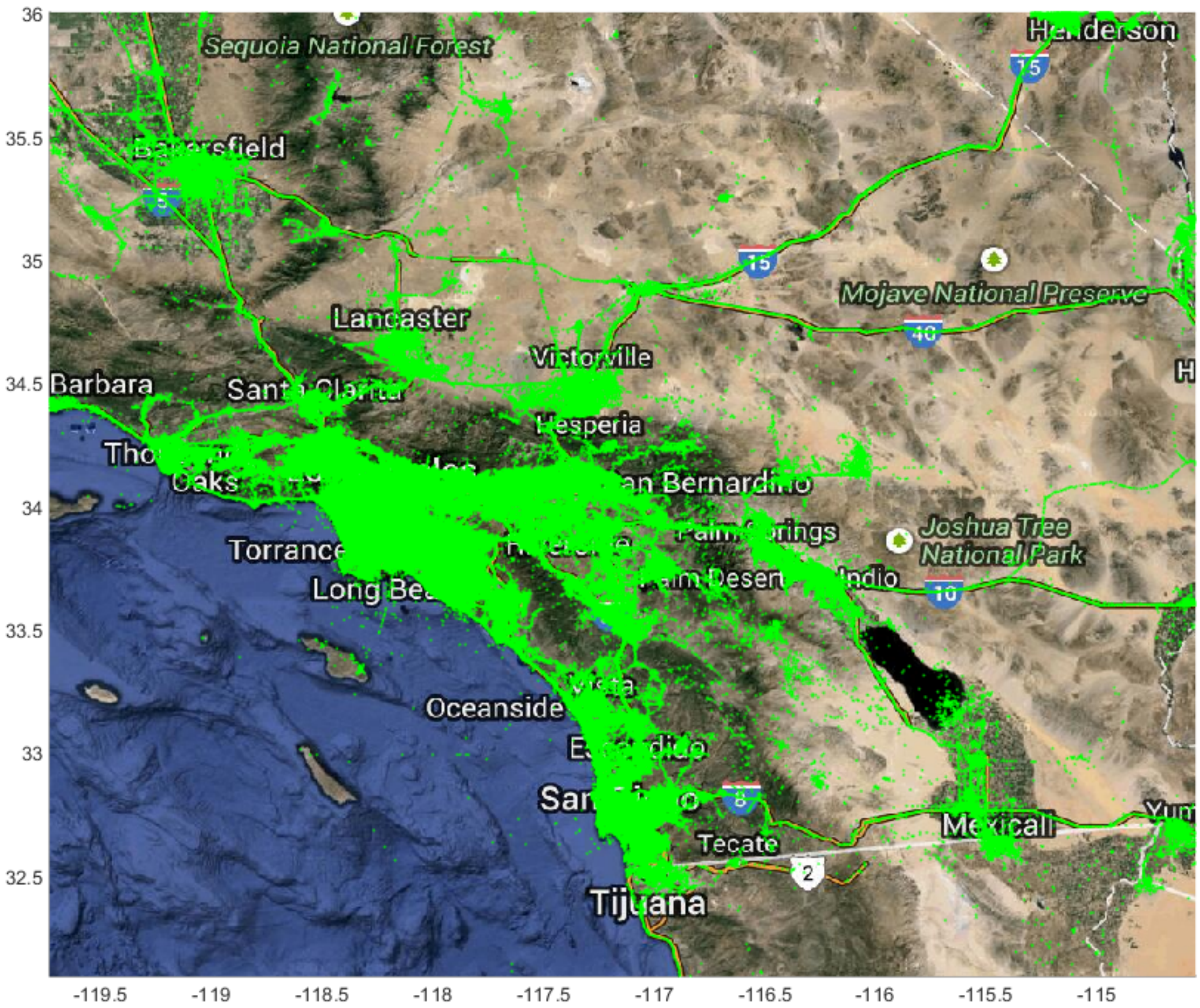
From http://exploringbigdata.blogspot.com/



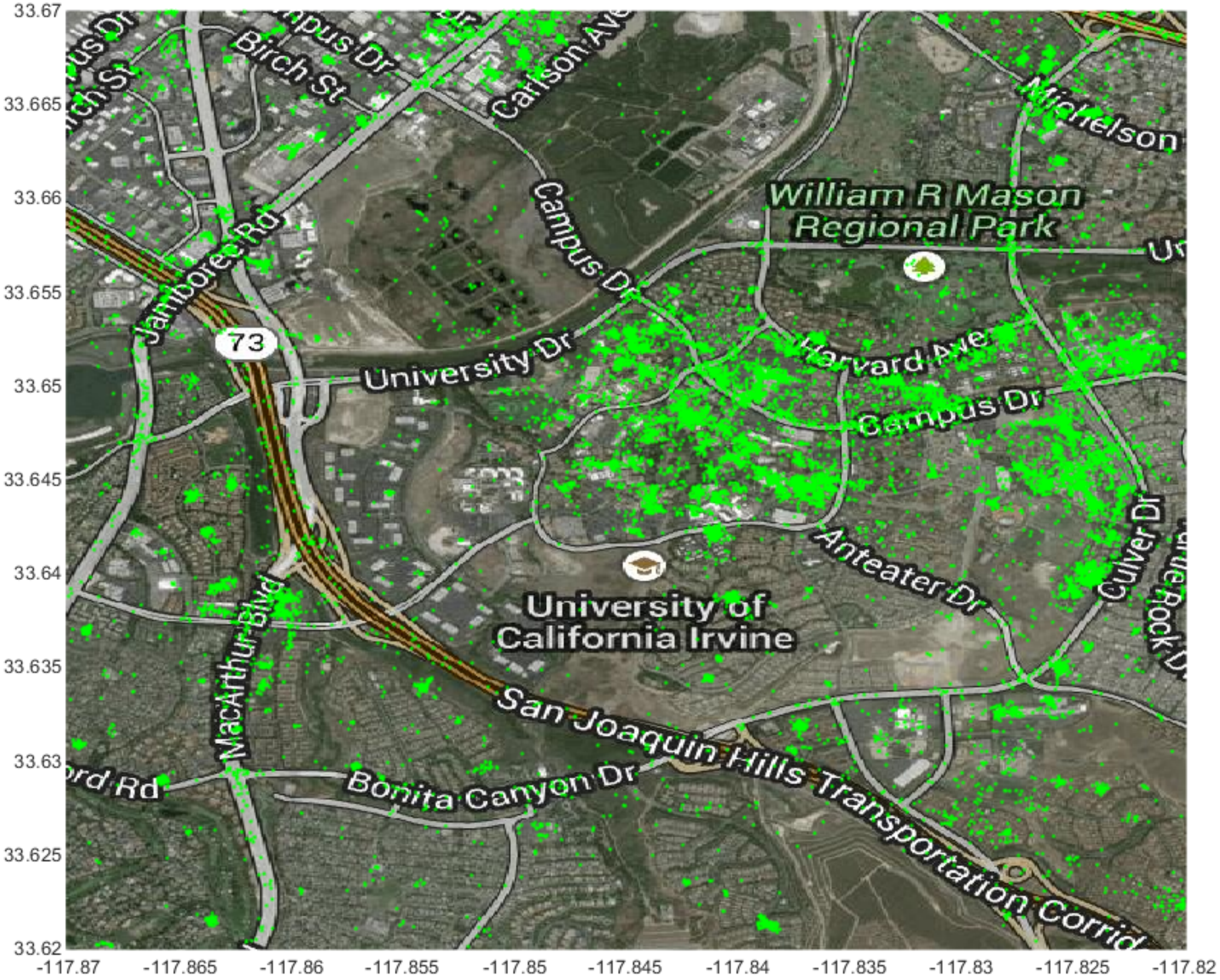
http://www.mapsandthecity.com/wp-content/uploads/2011/11/Schermafbeelding-2011-11-09-om-10.13

Modeling Human Behavior using Social Media

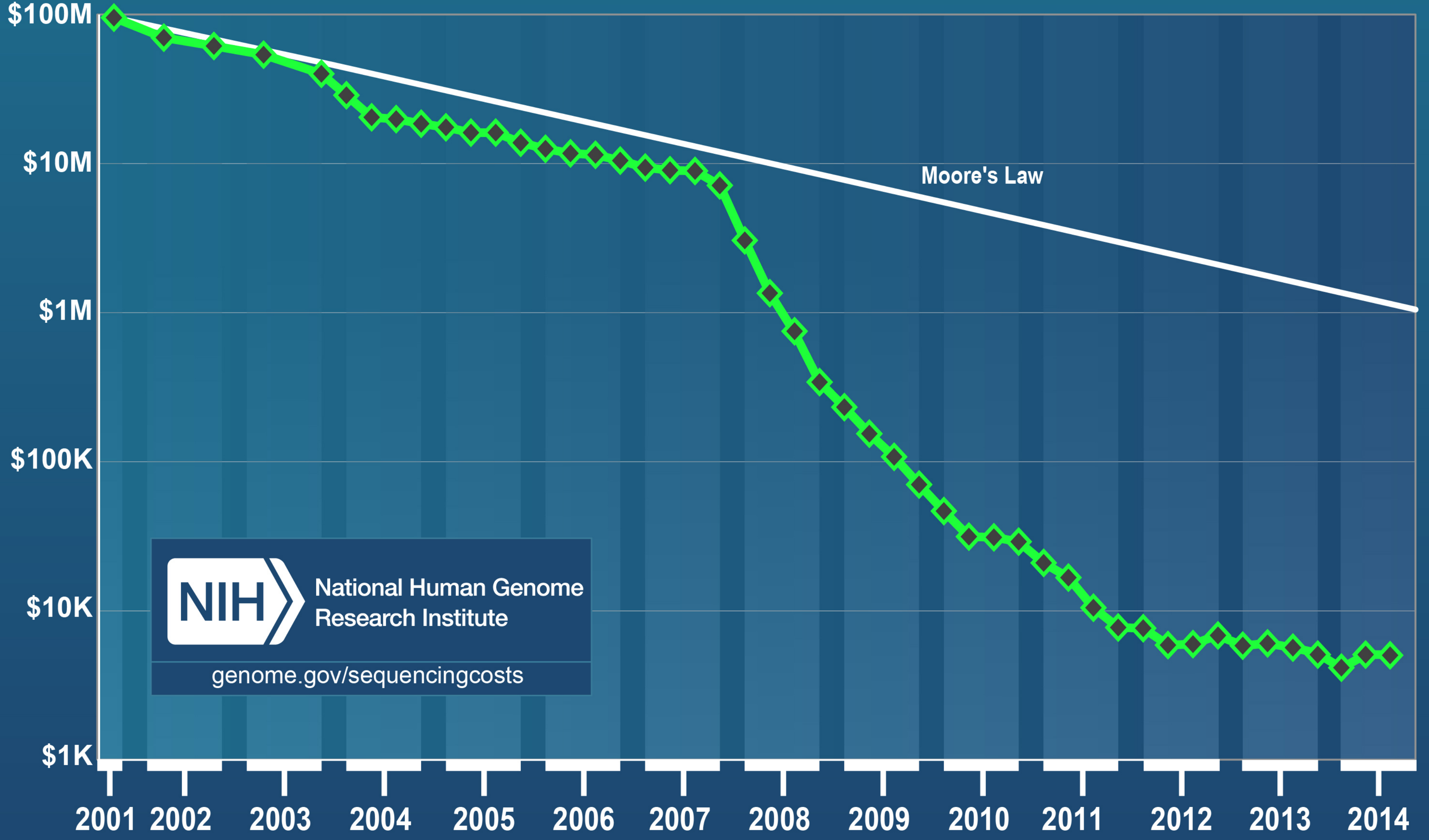
From Lichman and Smyth, ACM SIGKDD 2014



Geolocated Tweets around UC Irvine



Cost per Genome



From: healthitananalytics.com

Scientific Data: Large Hadron Collider at CERN

60 Terabytes/day
20 Petabytes/year

1 Terabyte = 10^{12} bytes
1 Petabyte = 10^{15} bytes

A Paradigm Shift in Data Analysis

- Technological drivers
  - Sensors (cheap and ubiquitous, e.g., GPS on your phone)
  - Data storage (we are all “data owners”)
  - Computational power
  - Data analysis methods (statistics and machine learning)
  - Internet and wireless communication (can collect and share data)

- Convergence…..tremendous demand for data analysis
  - In business, in sciences, in medicine, in engineering, and more……

- In the past, this demand was met by statistics
  - Does not scale up – there are not nearly enough statisticians
  - Need more tools than just statistics….need databases, algorithms, machine learning,….

DATA SCIENCE IN THE REAL WORLD

What is Data Science?

Data science involves the full lifecycle of data:

from real-world unstructured data…..to predictions and decisions

Data science is broader than just databases, statistics, ML, algorithms

…..but these are all critical components

Key aspects of data science include

- Domain knowledge and problem definition
- Data preparation/organization/management
- Understanding of uncertainty (statistics)
- Computing, algorithms, fitting models, machine learning
- Iterative exploration and experimentation
- Human judgement and interpretation

How is Data Science used in each of these Organizations?

How is Data Science used in each of these Organizations?

Organizations

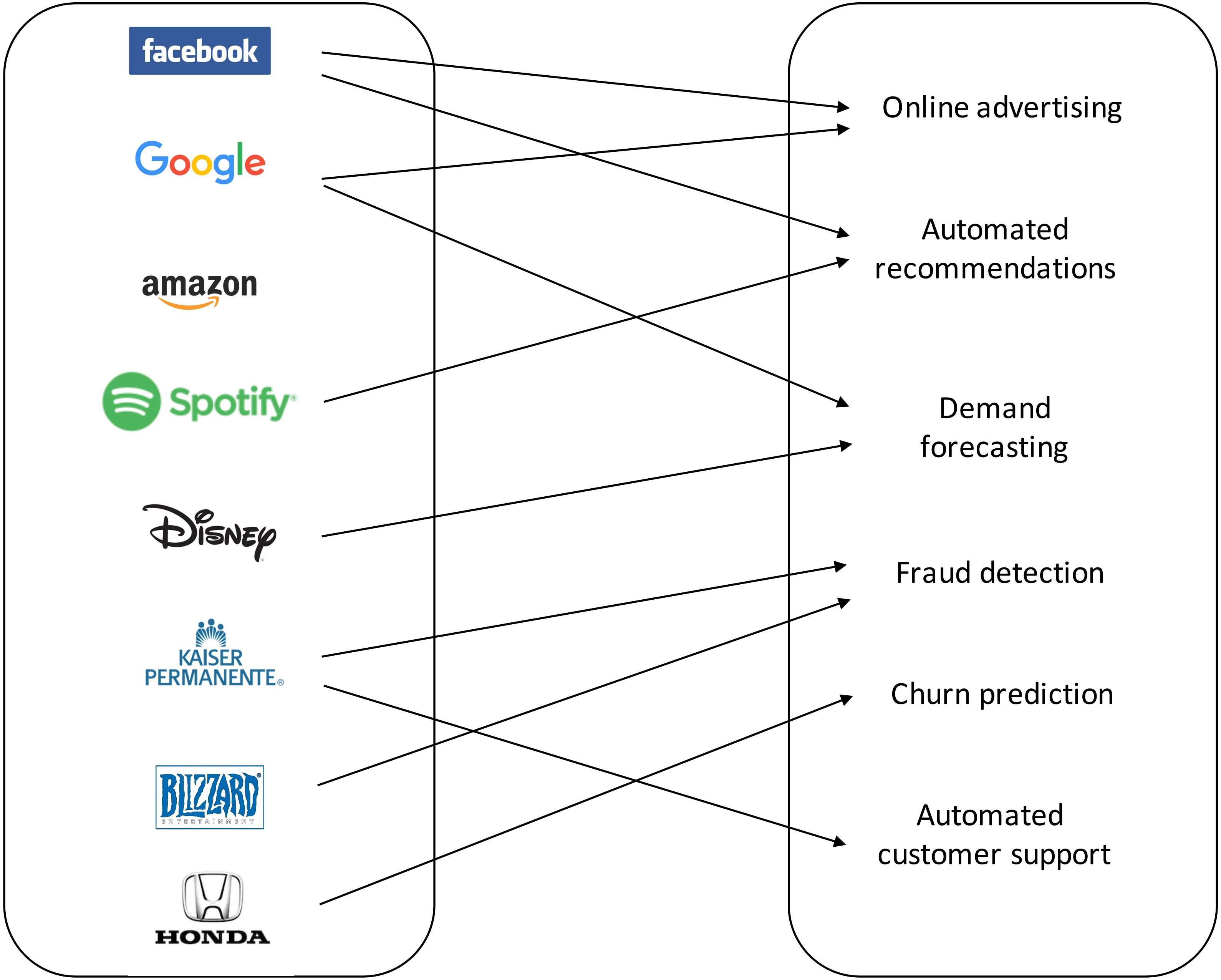
- facebook
- Google
- amazon
- Spotify
- Disney
- KAISER PERMANENTE
- BLIZZARD ENTERTAINMENT
- HONDA

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support

Organizations

Data Science Applications



Online advertising

Automated recommendations

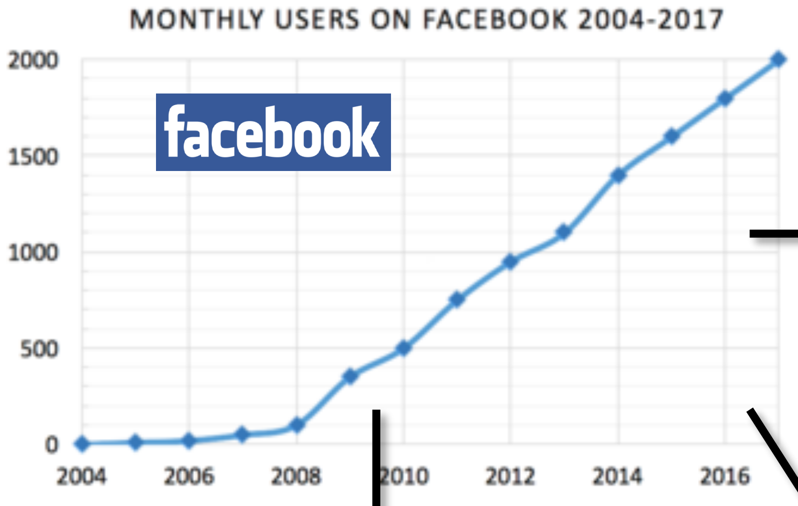
Demand forecasting

Fraud detection

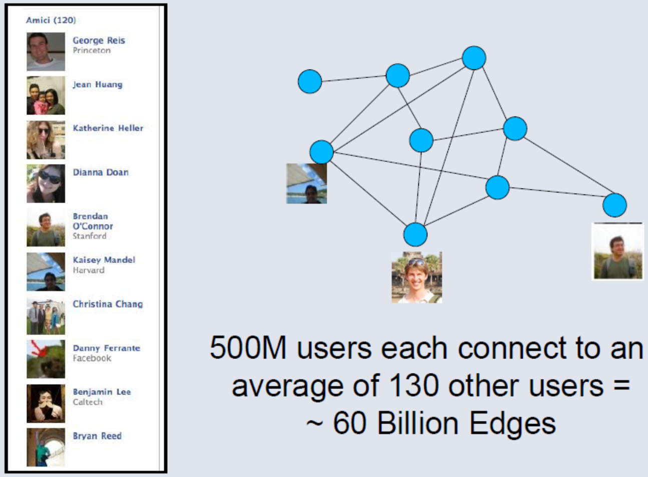
Churn prediction

Automated customer support

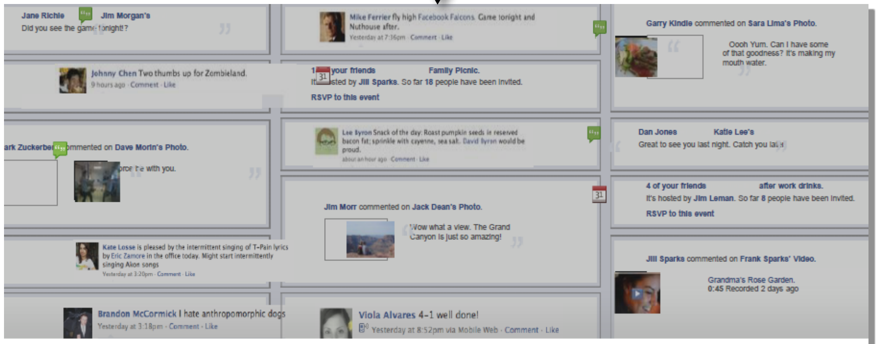
How does Facebook predict what content to show you?


MONTHLY USERS ON FACEBOOK 2004-2017

The Friendship graph



500M users each connect to an average of 130 other users = ~ 60 Billion Edges


Jane Richie Jim Morgan's
Did you see the game tonight!?

Mike Ferrier fly high Facebook Falcons. Game tonight and Nuthouse after.
Yesterday at 7:36pm · Comment · Like

Garry Kindle commented on Sara Lima's Photo.
Oooh Yum. Can I have some of that goodness? It's making my mouth water.

Johnny Chen Two thumbs up for Zombieland.
9 hours ago · Comment · Like

1 your friends Family Picnic.
It sted by Jill Sparks. So far 18 people have been invited.
RSVP to this event

ark Zuckerber mmented on Dave Morin's Photo.
orce e with you.

Lee Byron Snack of the day: Roast pumpkin seeds in reserved bacon fat; sprinkle with cayenne, sea salt. David Byron would be proud.
about an hour ago · Comment · Like

Dan Jones Katie Lee's
Great to see you last night. Catch you lat

4 of your friends after work drinks.
It's hosted by Jim Leman. So far 8 people have been invited.
RSVP to this event

Jim Morr commented on Jack Dean's Photo.
Wow what a view. The Grand Canyon is just so amazing!

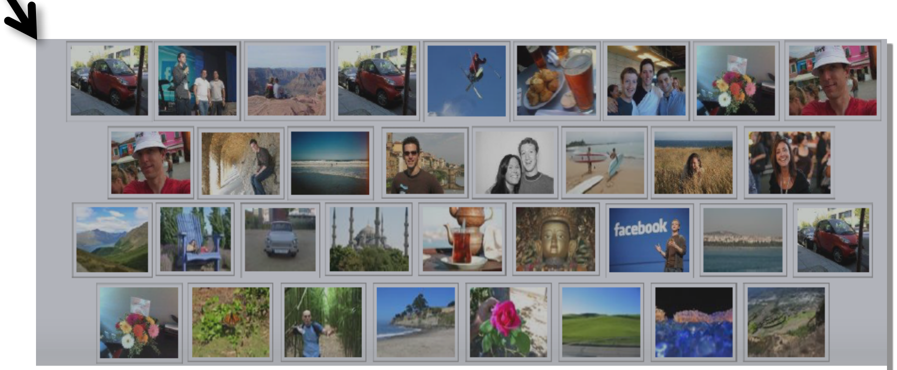
Kate Losse is pleased by the intermittent singing of T-Pain lyrics by Eric Zamore in the office today. Might start intermittently singing Akon songs
Yesterday at 3:20pm · Comment · Like

Jill Sparks commented on Frank Sparks' Video.
Grandma's Rose Garden.
0:45 Recorded 2 days ago

Brandon McCormick I hate anthropomorphic dogs
Yesterday at 3:18pm · Comment · Like

Viola Alvares 4–1 well done!
Yesterday at 8:52pm via Mobile Web · Comment · Like

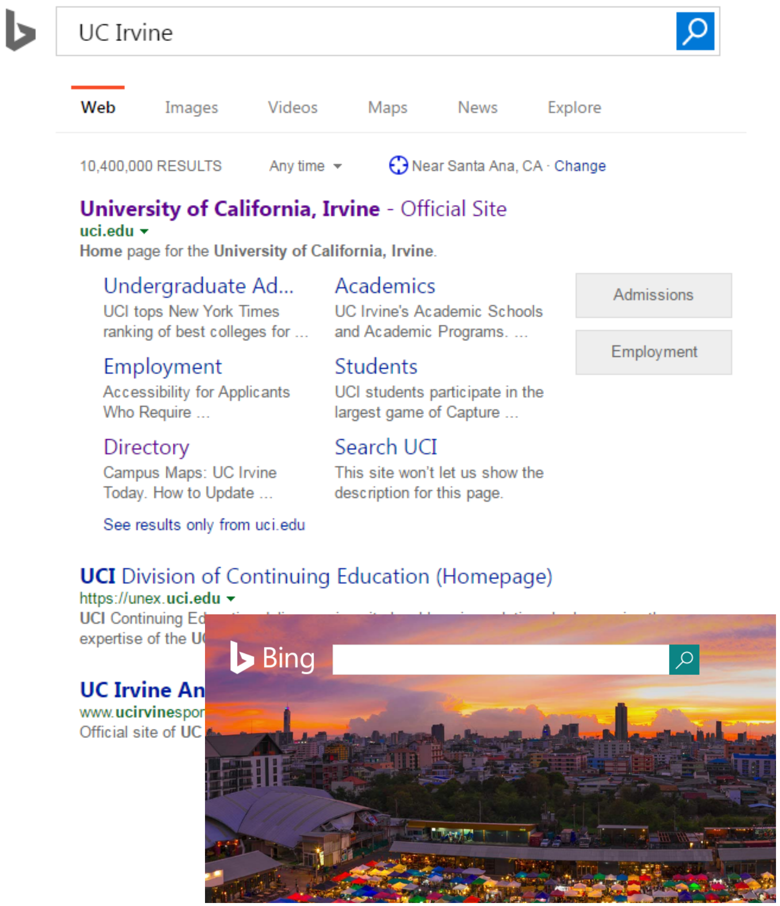
Over 30 billion pieces of content shared every month



Over **3 billion** photos uploaded each month

Graphics from Lars Backstrom, ESWC 2011

Web Search: How do search engines rank search results?



UC Irvine

10,400,000 RESULTS Any time Near Santa Ana, CA · Change

University of California, Irvine - Official Site
uci.edu
Home page for the **University of California, Irvine**.

Undergraduate Ad...
UCI tops New York Times ranking of best colleges for ...

Academics
UC Irvine's Academic Schools and Academic Programs. ...

Employment
Accessibility for Applicants Who Require ...

Students
UCI students participate in the largest game of Capture ...

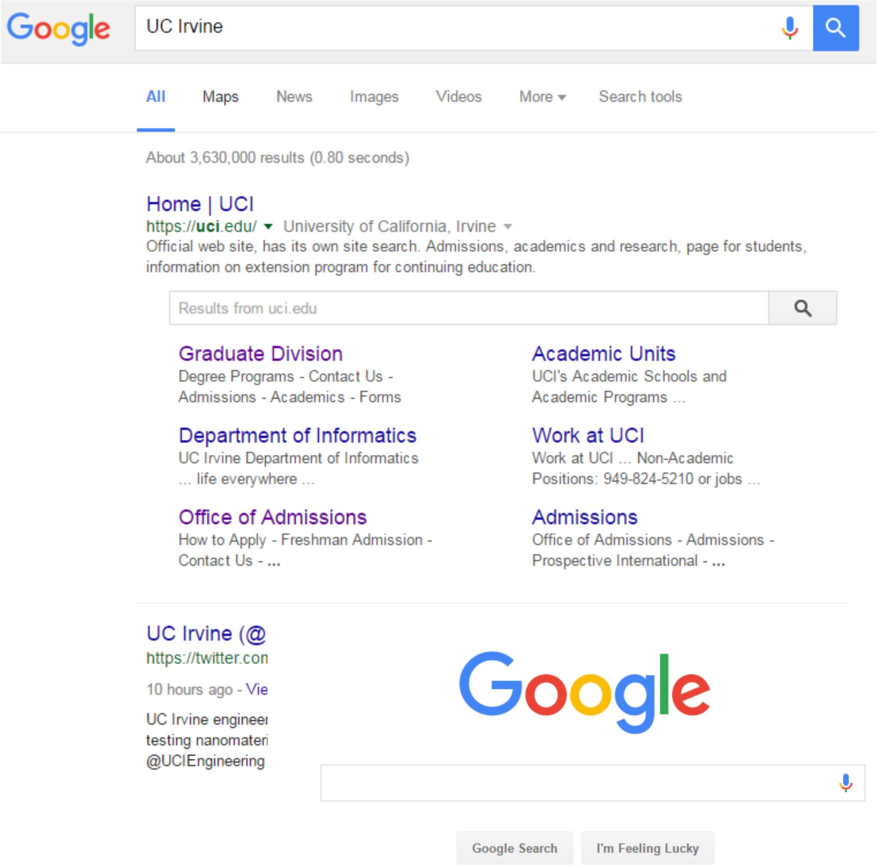
Directory
Campus Maps: UC Irvine Today. How to Update ...

Search UCI
This site won't let us show the description for this page.

See results only from uci.edu

UCI Division of Continuing Education (Homepage)
https://unex.**uci.edu**
UCI Continuing Ed
expertise of the **U**

UC Irvine An
www.**ucirvinespor**
Official site of UC



UC Irvine

About 3,630,000 results (0.80 seconds)

Home | UCI
https://**uci**.edu/ University of California, Irvine
Official web site, has its own site search. Admissions, academics and research, page for students, information on extension program for continuing education.

Results from uci.edu

Graduate Division
Degree Programs - Contact Us - Admissions - Academics - Forms

Academic Units
UCI's Academic Schools and Academic Programs ...

Department of Informatics
UC Irvine Department of Informatics ... life everywhere ...

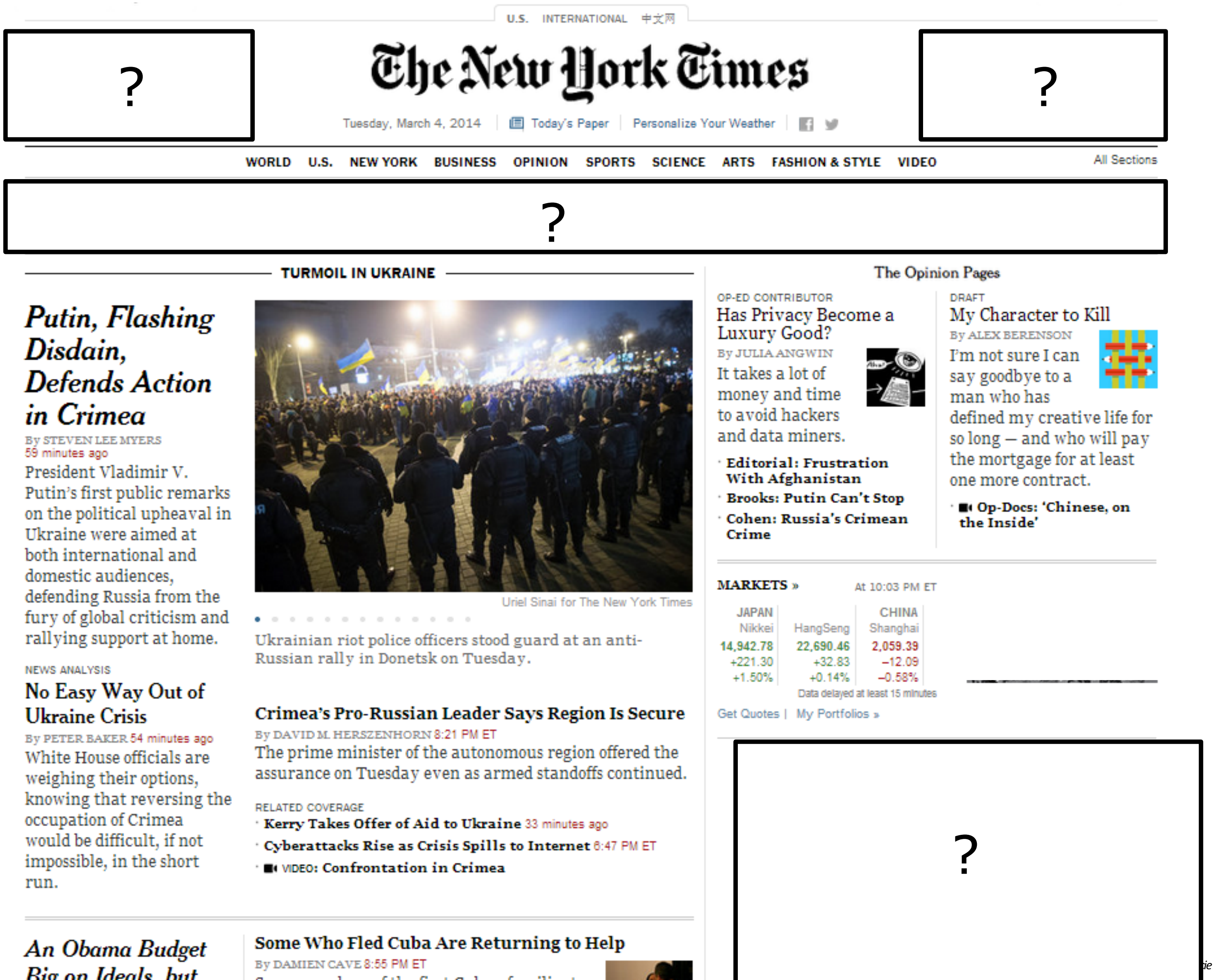
Work at UCI
Work at UCI ... Non-Academic Positions: 949-824-5210 or jobs ...

Office of Admissions
How to Apply - Freshman Admission - Contact Us - **...**

Admissions
Office of Admissions - Admissions - Prospective International - **...**

UC Irvine (@
https://twitter.con
10 hours ago - Vie
UC Irvine engineer
testing nanomateri
@UCIEngineering

How do ad companies decide what online ads to show you?



U.S. INTERNATIONAL 中文网

?

The New York Times

?

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather |

WORLD U.S. NEW YORK BUSINESS OPINION SPORTS SCIENCE ARTS FASHION & STYLE VIDEO All Sections

?

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS 59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 6:47 PM ET
- VIDEO: **Confrontation in Crimea**

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- **Op-Docs: 'Chinese, on the Inside'**

MARKETS » At 10:03 PM ET

JAPAN Nikkei	HangSeng	CHINA Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

Get Quotes | My Portfolios »

?

An Obama Budget Big on Ideals, but

Some Who Fled Cuba Are Returning to Help

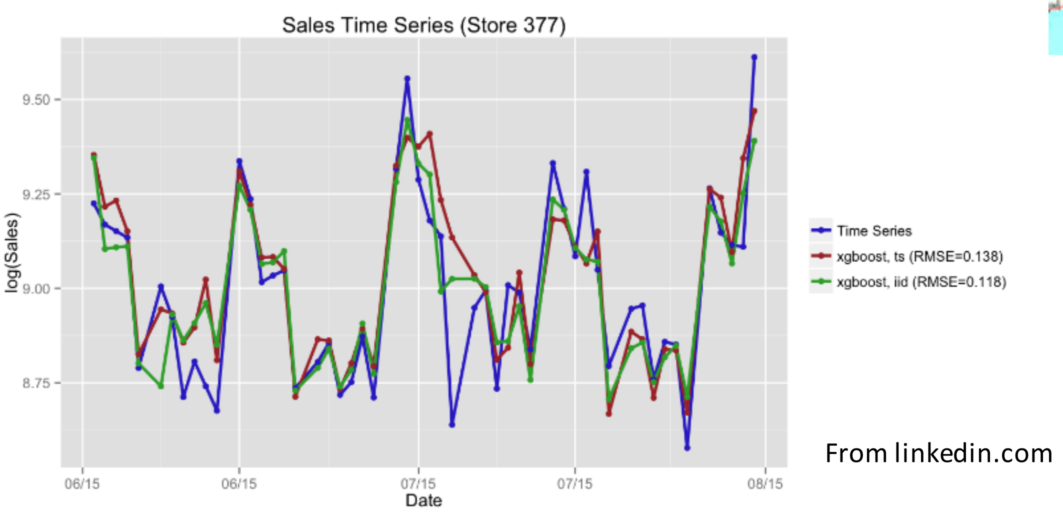
By DAMIEN CAVE 8:55 PM ET

How does Amazon forecast how many items for its warehouses?

From dailymail.co.uk

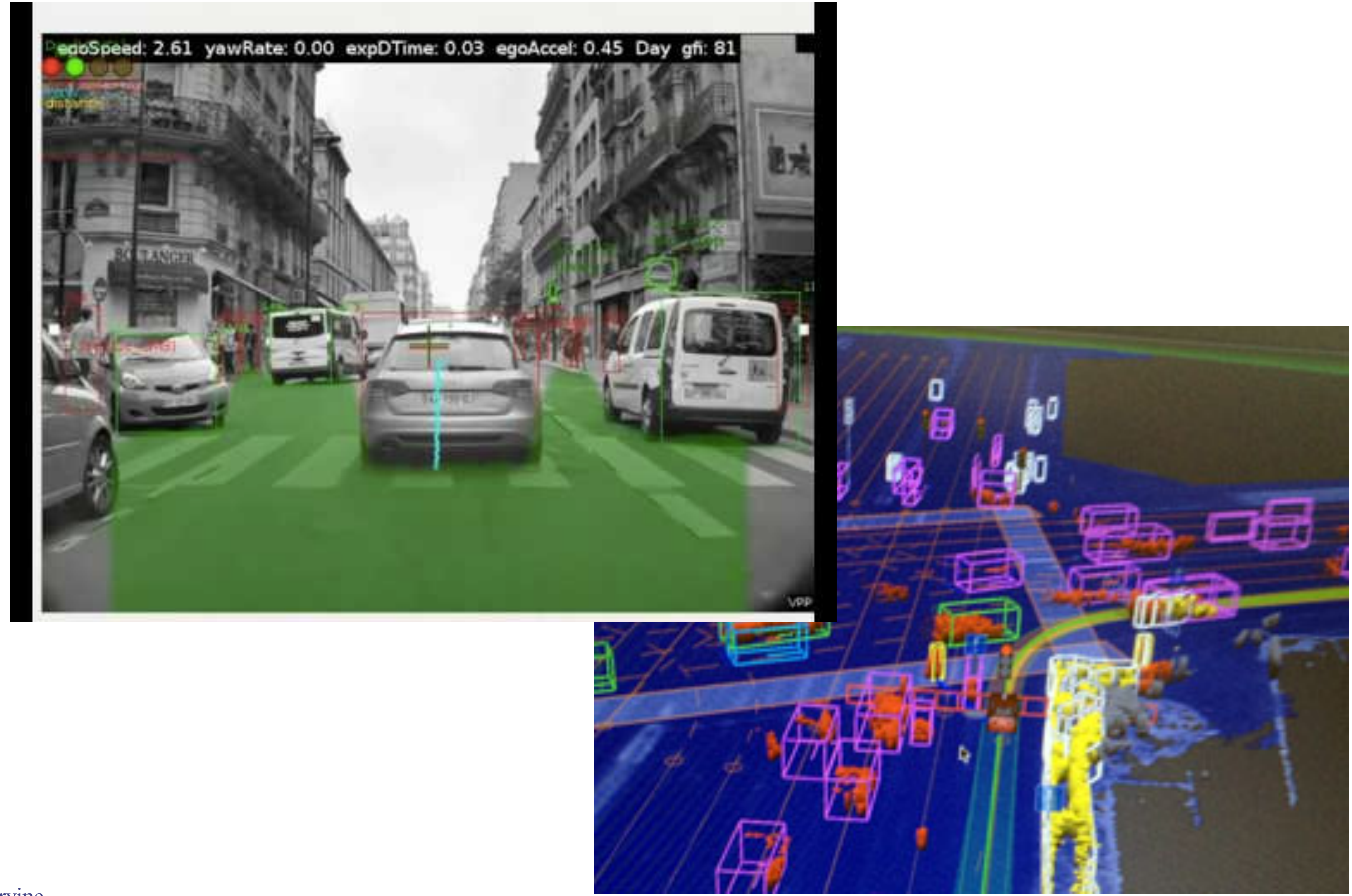
CANADA
ATLANTIC OCEAN
PACIFIC OCEAN

From www.formaspace.com

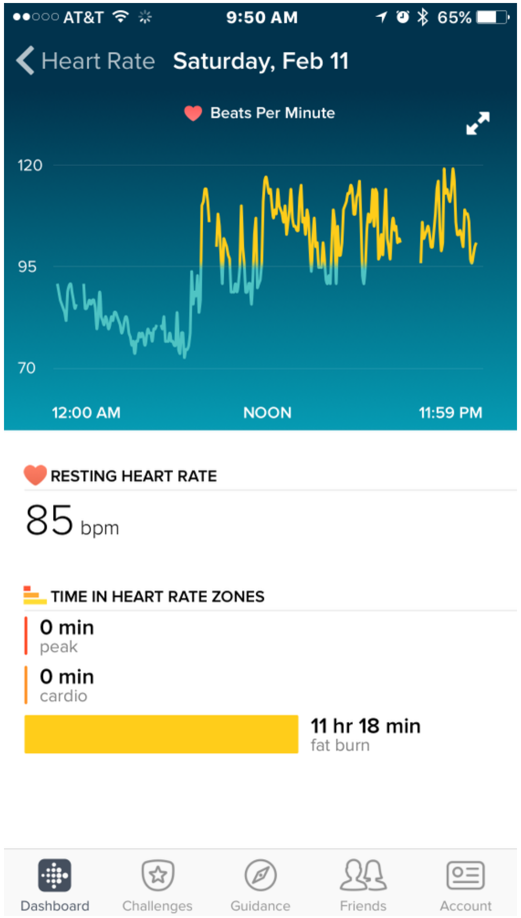


From linkedin.com

How do autonomous cars recognize objects in image data?


egoSpeed: 2.61 yawRate: 0.00 expDTime: 0.03 egoAccel: 0.45 Day gfi: 81
VPP

How can we use wearable data to improve our health?


●●○○○ AT&T 9:50 AM 65%

Heart Rate Saturday, Feb 11

Beats Per Minute

RESTING HEART RATE

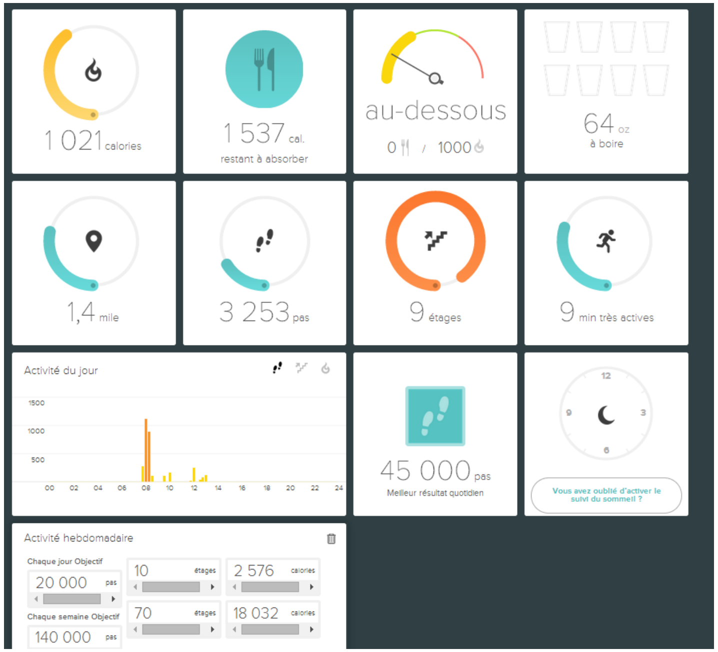
85 bpm

TIME IN HEART RATE ZONES

0 min
peak

0 min
cardio

11 hr 18 min
fat burn


1 021 calories

1 537 cal.
restant à absorber

au-dessous
0 / 1000

64 oz
à boire

1,4 mile

3 253 pas

9 étages

9 min très actives

Activité du jour

45 000 pas
Meilleur résultat quotidien

Vous avez oublié d'activer le suivi du sommeil ?

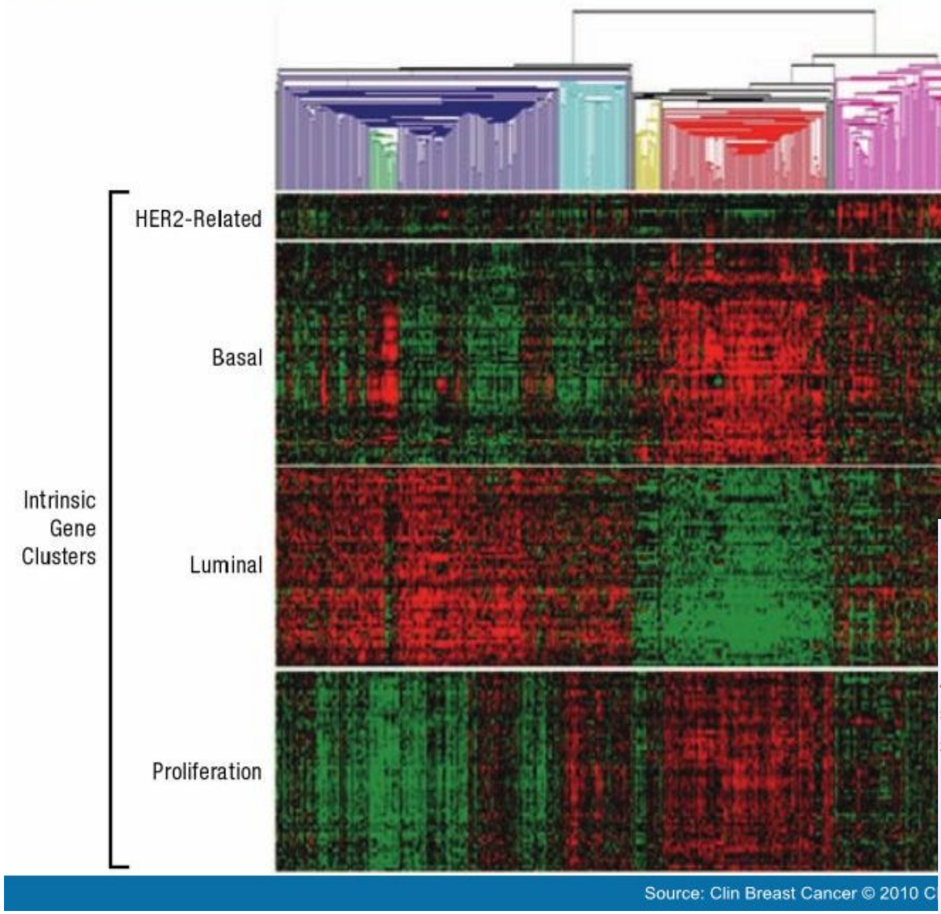
Activité hebdomadaire

Chaque jour Objectif
20 000 pas | 10 étages | 2 576 calories

Chaque semaine Objectif
140 000 pas | 70 étages | 18 032 calories

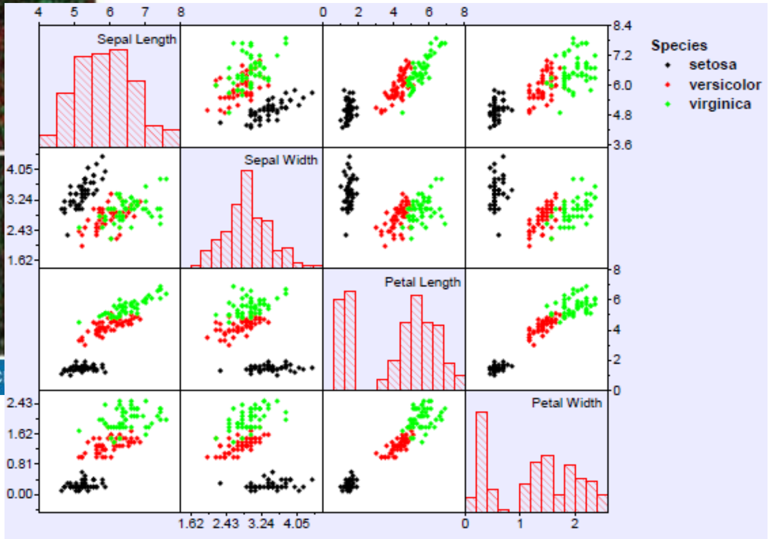
Images from community.fitbit.com

How can we make personalized recommendations in medicine?



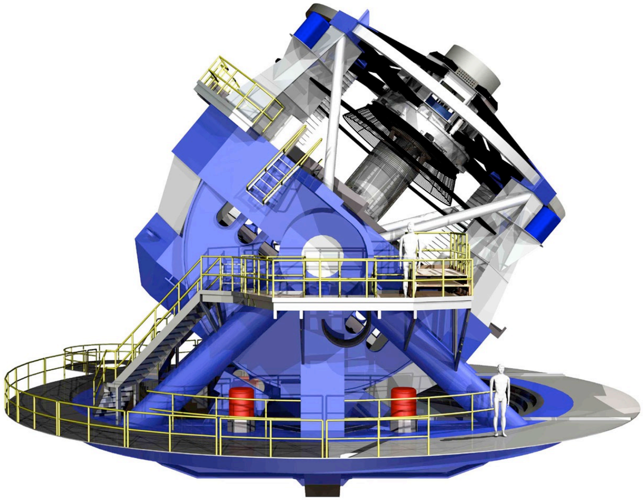
Source: Clin Breast Cancer © 2010 C

Data Matrix:
Rows = genes
Columns = patients



From www.originlab.com

Astronomy: How can we process terabytes/day of telescope data?

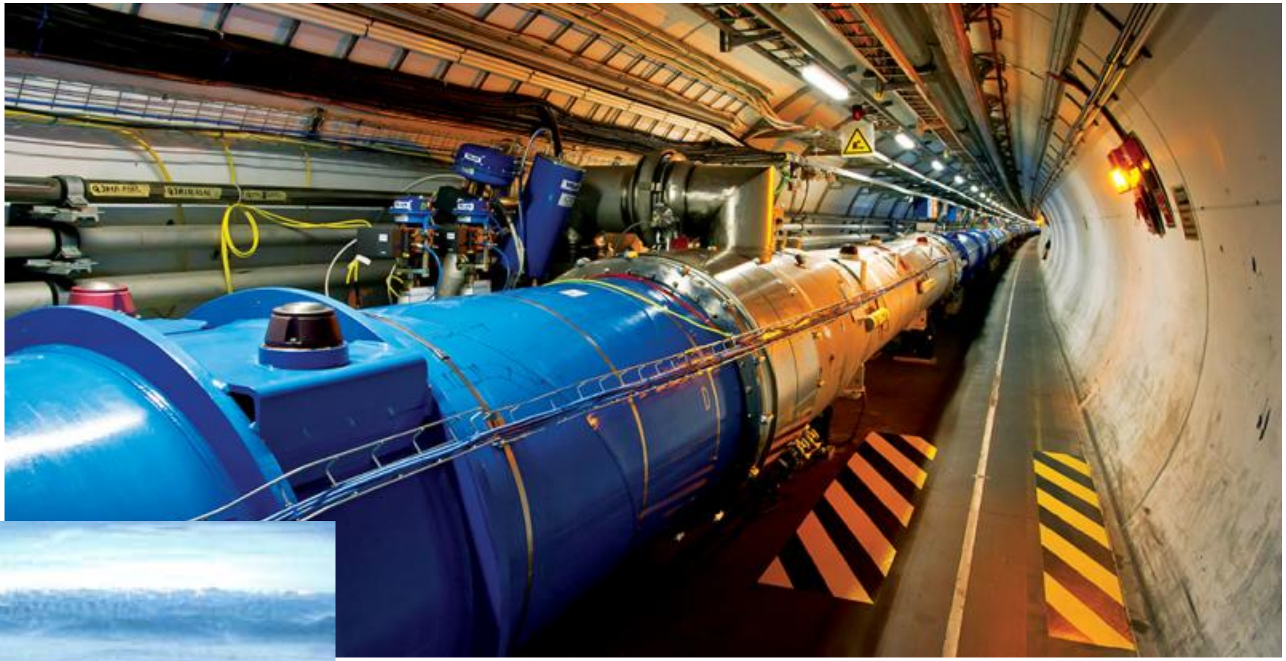


Large Synoptic Telescope (LST)
15 Terabytes/day
100+ Petabytes in 10 years

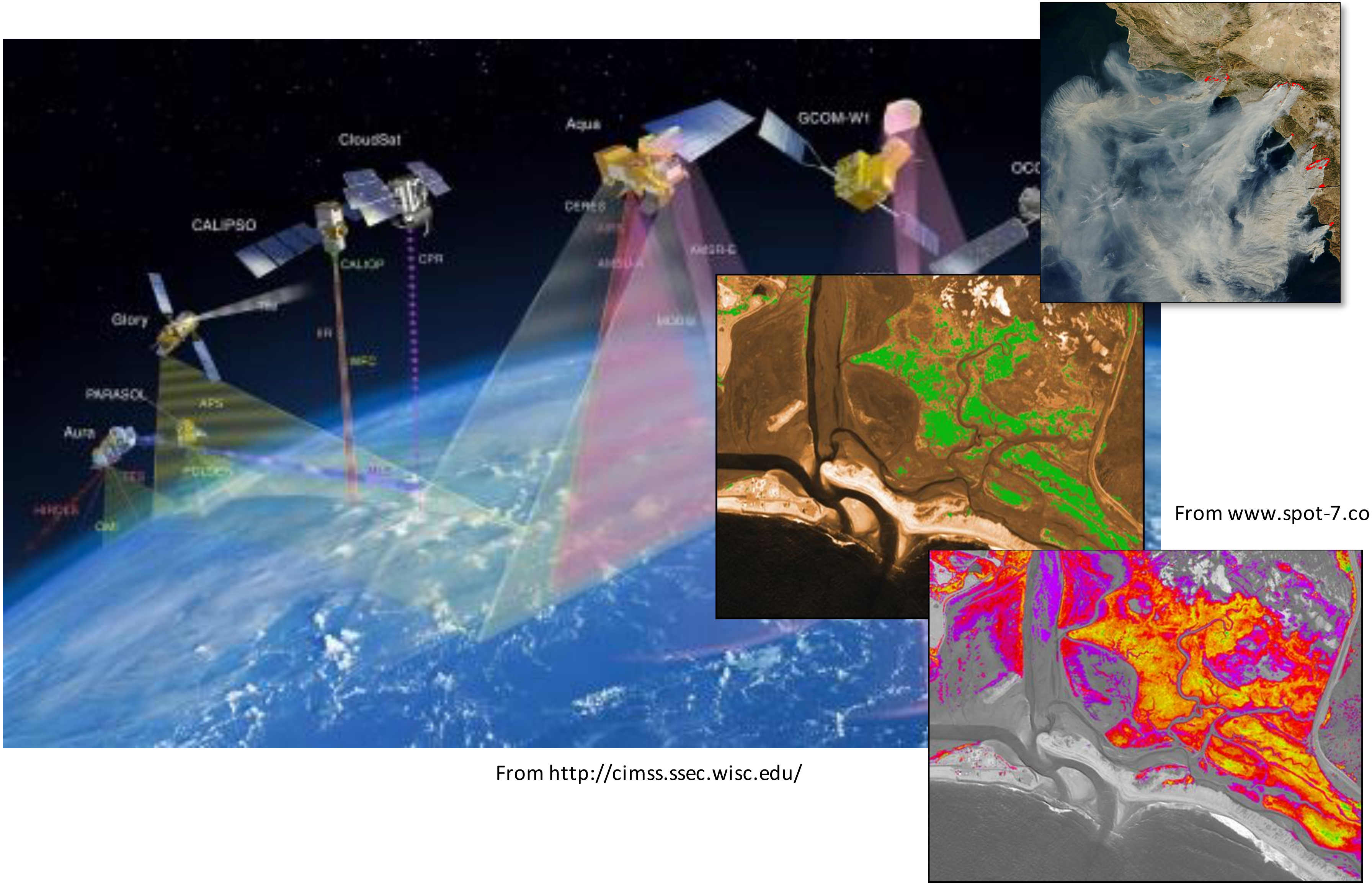
From Raddick et al, Astronomy Education Review, 2009

Physics: What is required to search for new physics particles?

Large Hadron Collider:
700 Mbytes/second
60 Terabytes/day
20 Petabytes/year



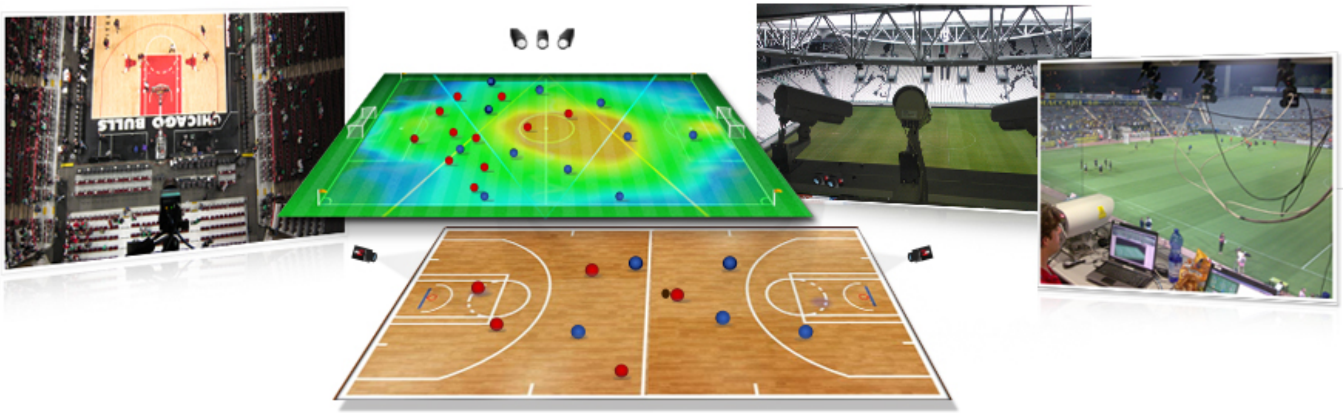
How can we detect land changes in NASA satellite images?



From www.spot-7.com

From http://cimss.ssec.wisc.edu/

How can algorithms interpret and summarize sports data?



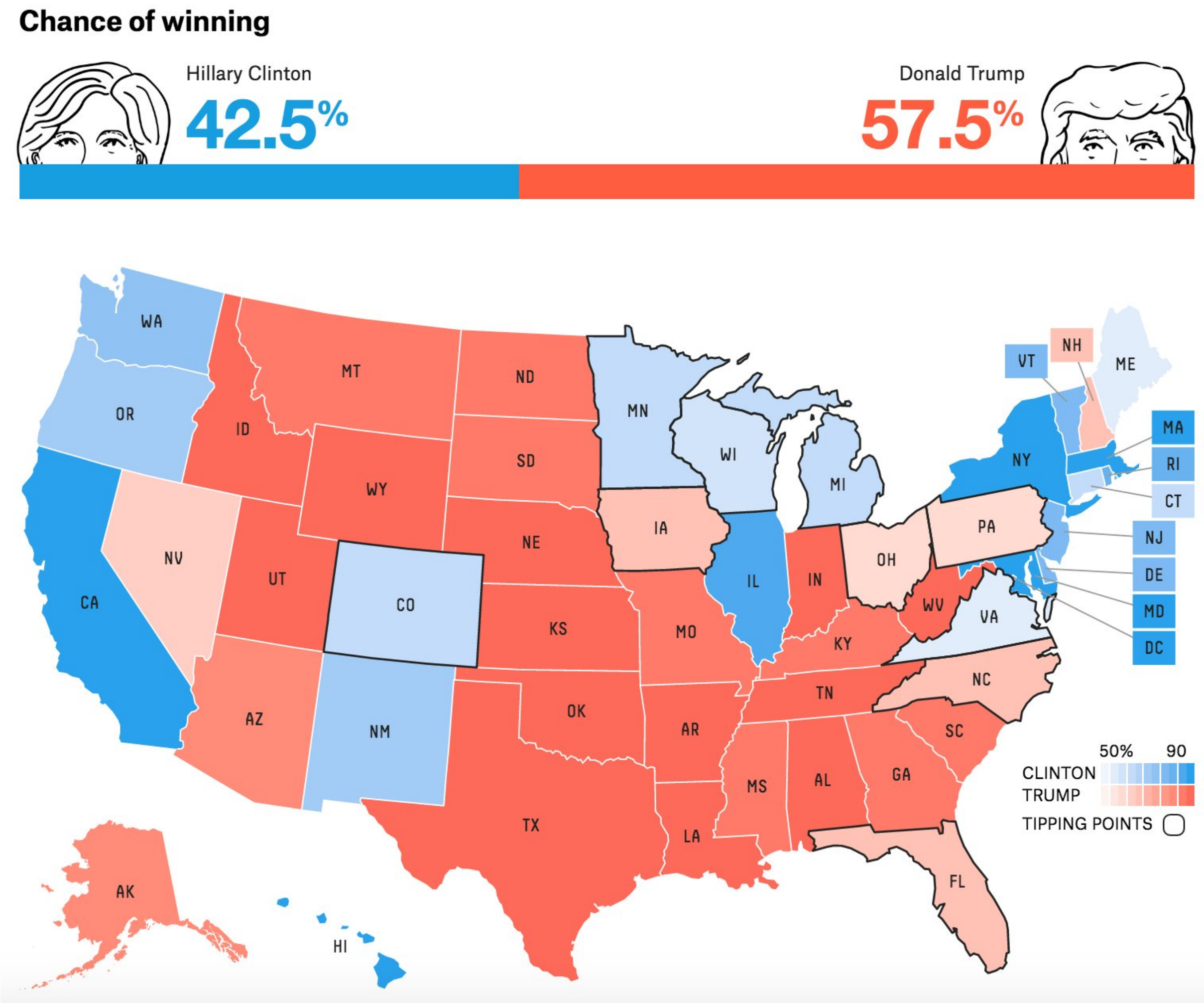
OPEN SHOT BREAKDOWN

LOS ANGELES LAKERS		NEW YORK KNICKS
12	FGM	16
25	FGA	50
48%	FG%	32%
6	3PM	5
12	3PA	20
50%	3P%	25%

DISTANCE 3,036 m
SPEED 6.6 km/h
SPRINTS 2
FOX SPORTS 3 HD LIVE
GALAXY Tab PLAYER STATISTICS

Graphics from www.stats.com/sportvu/sportvu.asp

Politics: How can we reliably predict events like elections?



Chance of winning

Hillary Clinton **42.5%**

Donald Trump **57.5%**

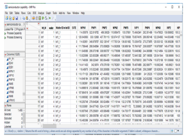
“Nowcast” forecast: Downloaded on July 25th 2016,
from http://projects.fivethirtyeight.com/2016-election-forecast/

Data Pipelines

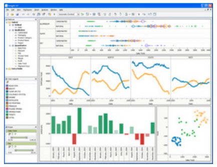
Unstructured Data

Extracted Data

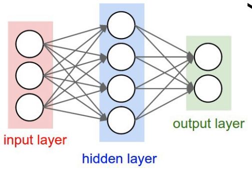
Transformed Data



Data for Modeling



Predictive Model



Predictions/ Decisions

Hidden Technical Debt in Machine Learning Systems

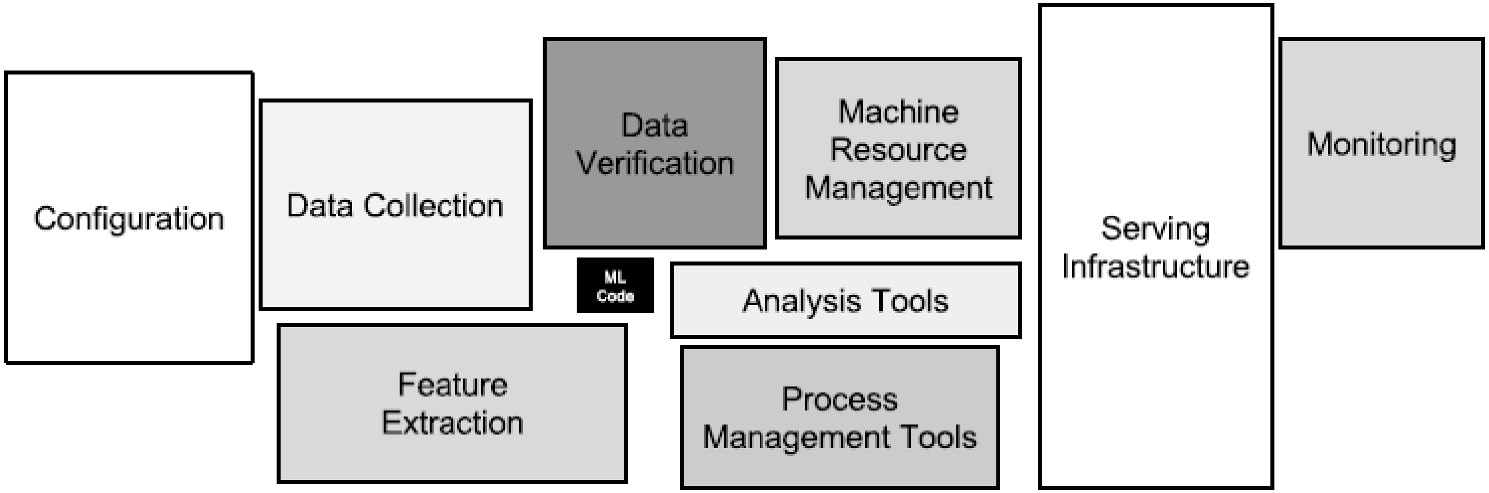


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

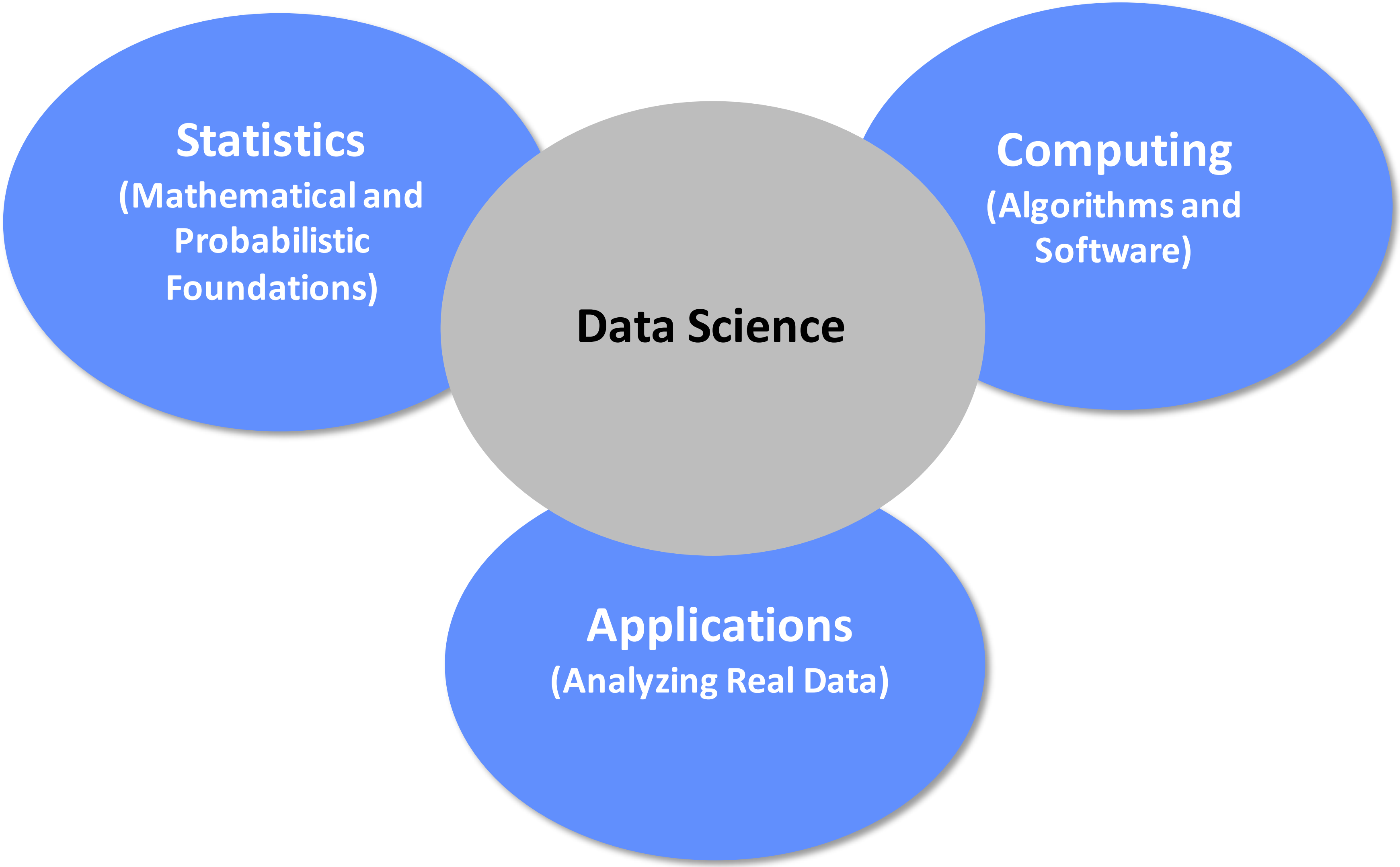
Scullley et al, NIPS 2015 Conference

THE DATA SCIENCE MAJOR

All of the applications we discussed are built on ideas from…

- Database systems
- Algorithms
- Software engineering
- Machine learning
- Probabilistic and statistical models
- Quantification of uncertainty
- Data visualization
- and more…

Components of Data Science



Statistics

(Mathematical and Probabilistic Foundations)

Computing

(Algorithms and Software)

Data Science

Applications

(Analyzing Real Data)

What Classes will you take in the DS Major?

Statistics

Stats 120 ABC: Intro to Prob and Stats
Stats 68: Exploratory Data Analysis
Stats 110-112: Statistical Methods
CS 178: Machine Learning
(Stats 140: Multivariate Statistics)

Computing

ICS 46: Data Structures
IFMTX 43: Intro to Software Engineering
CS 122A: Intro to Data Management
CS 161: Design and Analysis of Algorithms
(CS 131: Parallel and Distributed Computing)
(CS 172: Neural Networks/Deep Learning)

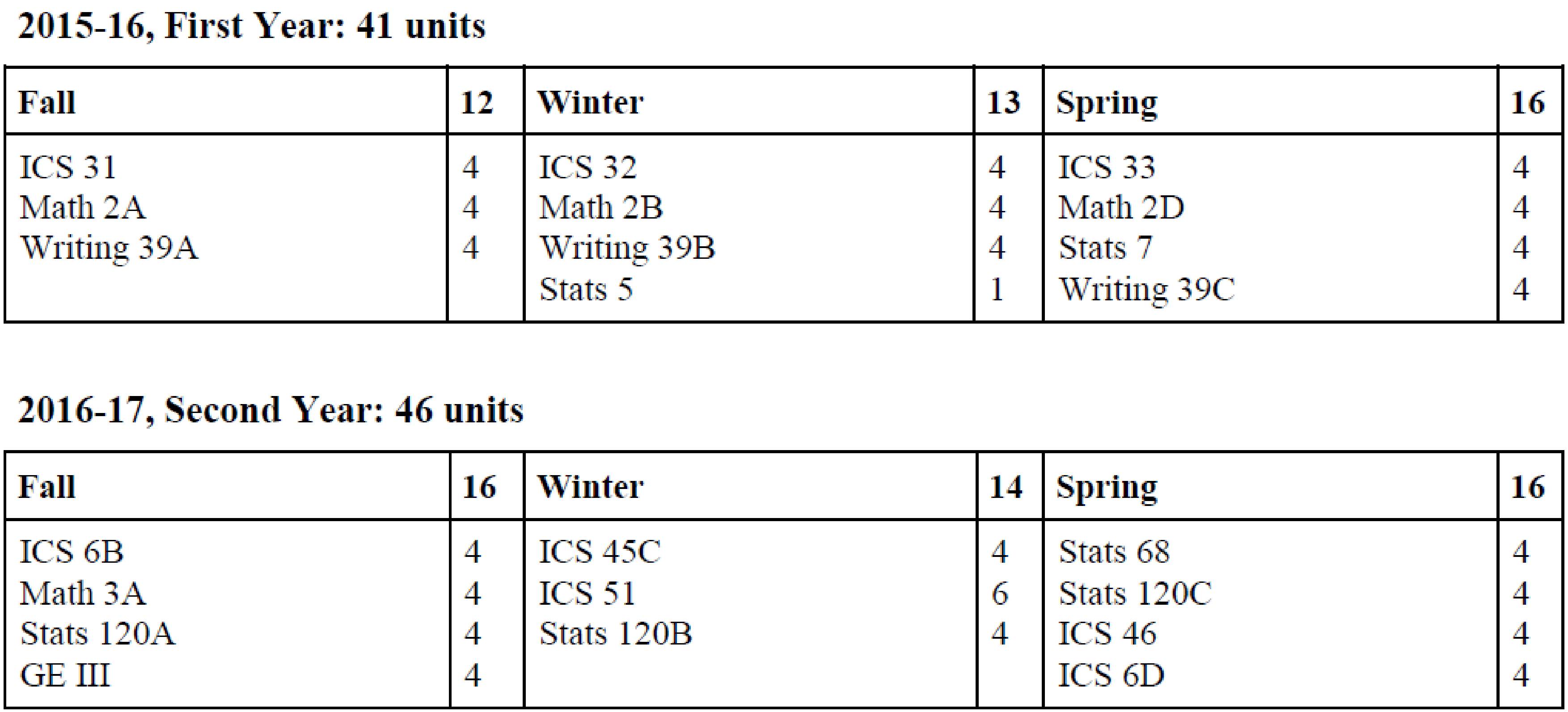
Applications

Stats 170AB: Data Science Capstone Project
INF 143: Information Visualization
(INF 131: Human Computer Interaction)
(CS 121: Information Retrieval)
(CS 122B: Project in Databases/Web Applications)
(Summer intermships, e.g., junior year)

(Sample electives shown in parentheses)

Sample Course of Study in the Major

Years 1 and 2: foundational courses in computer science, mathematics, statistics, including statistical computing



2015-16, First Year: 41 units

Fall	12	Winter	13	Spring	16
ICS 31	4	ICS 32	4	ICS 33	4
Math 2A	4	Math 2B	4	Math 2D	4
Writing 39A	4	Writing 39B	4	Stats 7	4
		Stats 5	1	Writing 39C	4

2016-17, Second Year: 46 units

Fall	16	Winter	14	Spring	16
ICS 6B	4	ICS 45C	4	Stats 68	4
Math 3A	4	ICS 51	6	Stats 120C	4
Stats 120A	4	Stats 120B	4	ICS 46	4
GE III	4			ICS 6D	4

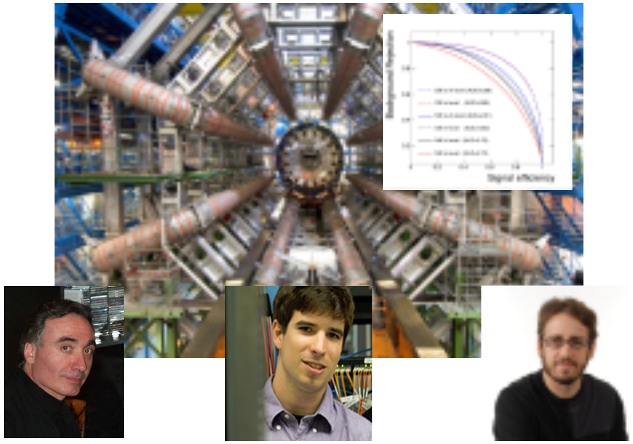
Years 3 and 4: more emphasis and specialization in data science topics such as machine learning, databases, visualization, advanced statistics

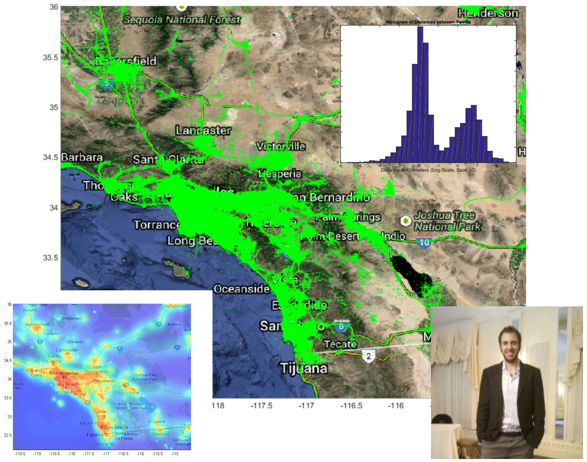
Year 3: sample program

Fall		Winter		Spring
Stats 110, Statistical Methods for Data Analysis I CS 161, Design and Analysis of Algorithms In4matx 43, Introduction to Software Engineering GE IV/VIII,		Stats 111, Statistical Methods for Data Analysis II CS 178, Machine Learning and Data-Mining ICS 139W, Critical Writing on Information Technology GE III/VII,		Stats 112, Statistical Methods for Data Analysis III CS 122A, Introduction to Data Management In4matx 143, Information Visualization GE VI,

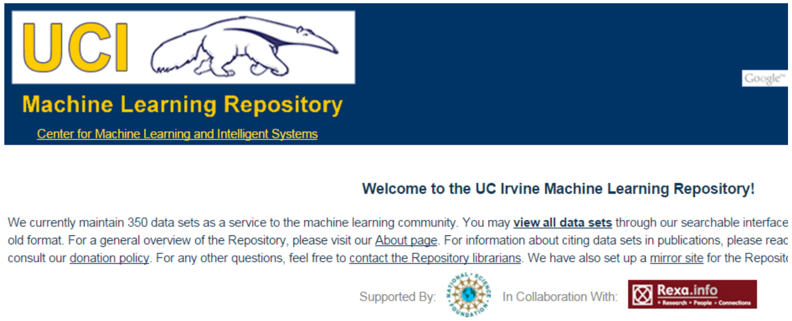
Year 4: two-quarter capstone "data-intensive" project, + statistics and CS electives

Research at UC Irvine in Data Science





UCI Data Science Initiative



UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 350 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please reac consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Reposit

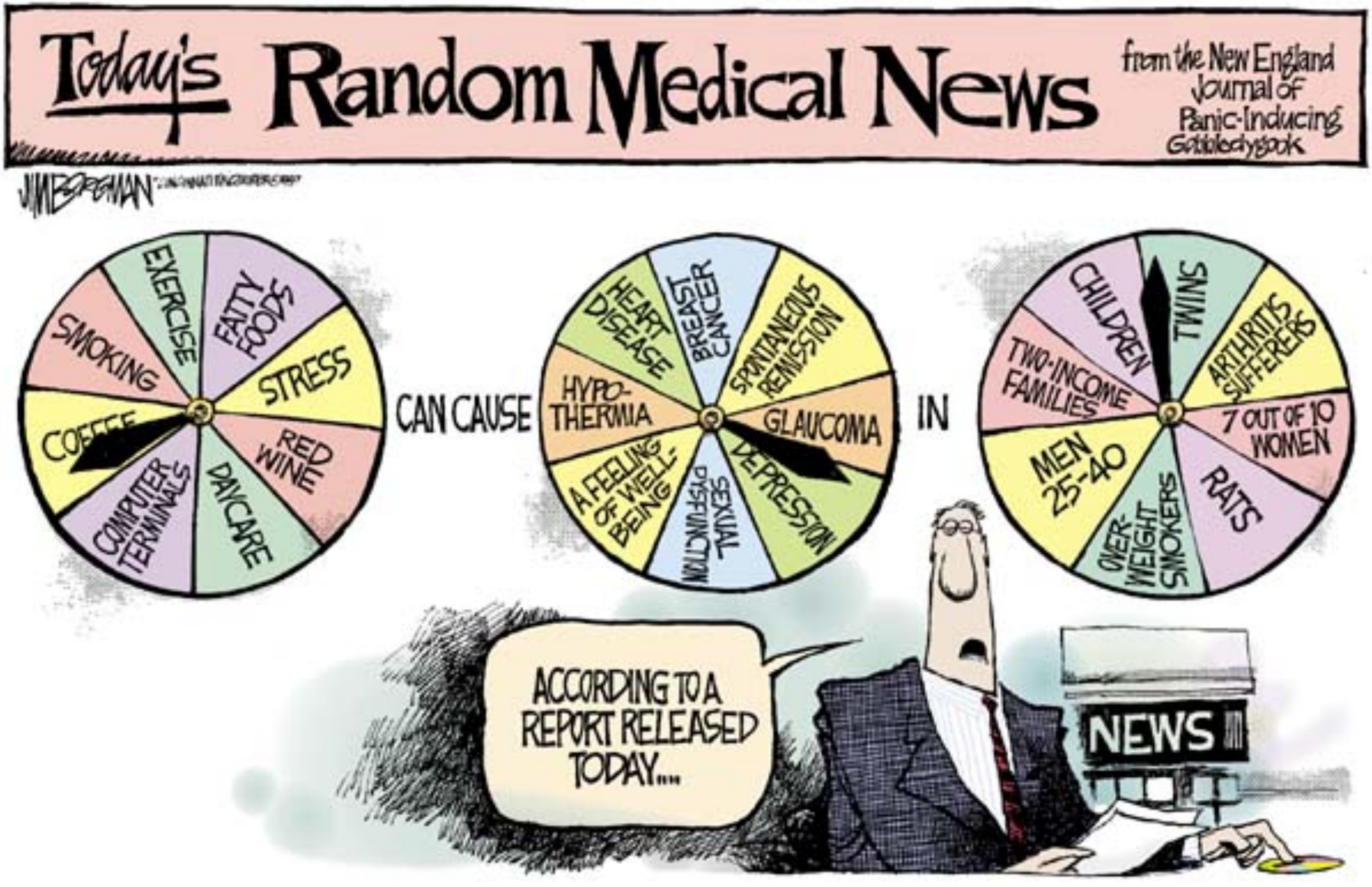
Supported By: In Collaboration With: Rexa.info • Research • People • Connections

Center for Machine Learning and Intelligent Systems

Bren School of Information and Computer Science

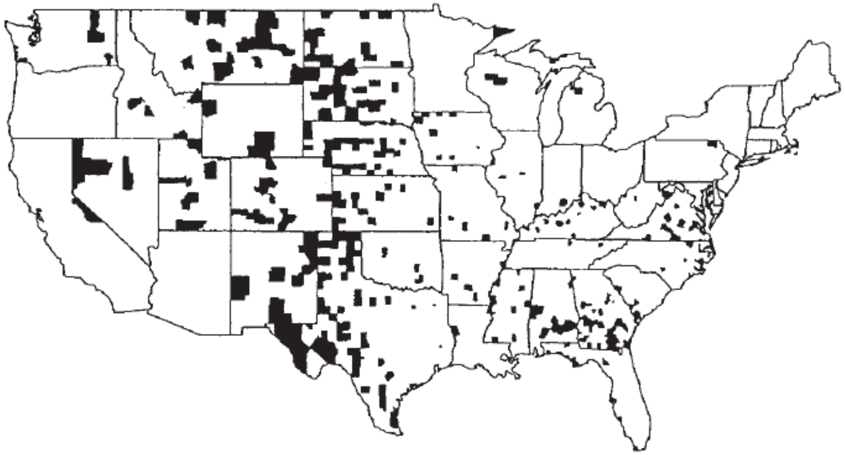
University of California, Irvine

LIMITATIONS OF WHAT WE CAN LEARN FROM DATA

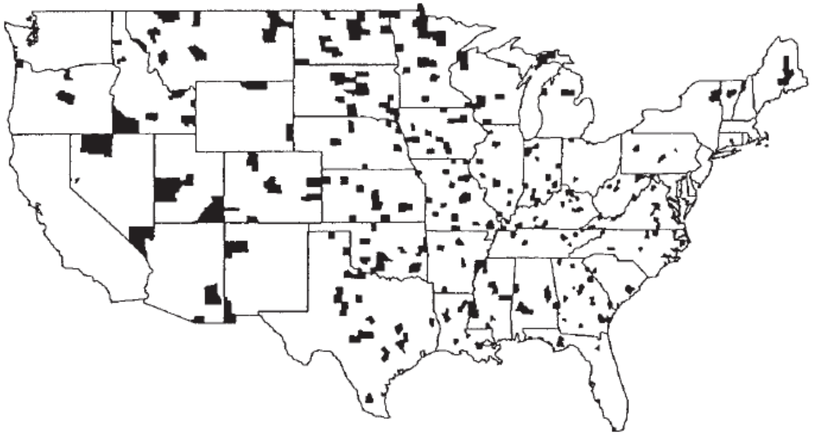

Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
RED WINE
DAYCARE
COMPUTER TERMINALS
COFFEE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
GLAUCOMA
DEPRESSION
SEXUAL DYSFUNCTION
A FEELING OF WELL-BEING
HYPO-THERMIA
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
7 OUT OF 10 WOMEN
RATS
OVER-WEIGHT SMOKERS
MEN 25-40
TWO-INCOME FAMILIES
ACCORDING TO A REPORT RELEASED TODAY....
NEWS

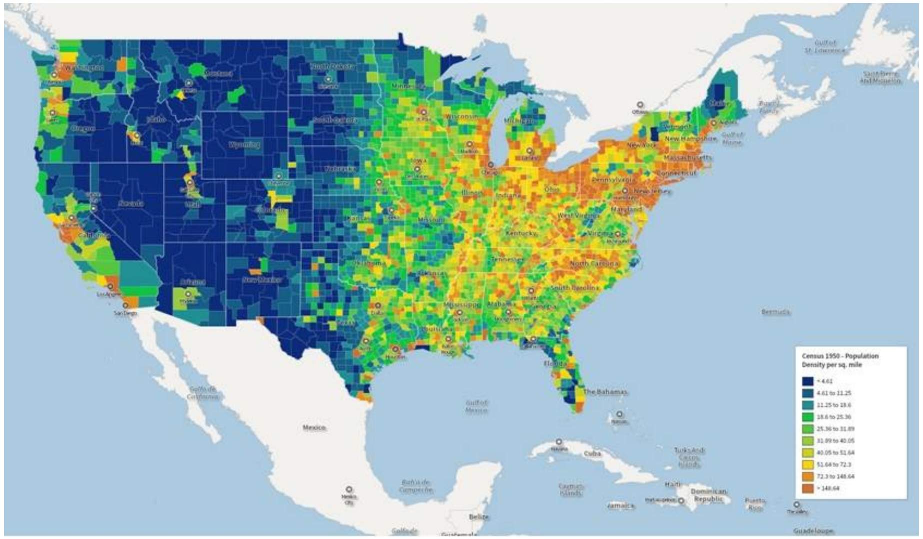
Kidney Cancer Death Rates by County in the US

Lowest Rates



Highest Rates





From A. Gelman and D. Nolan
Oxford University Press, 2002

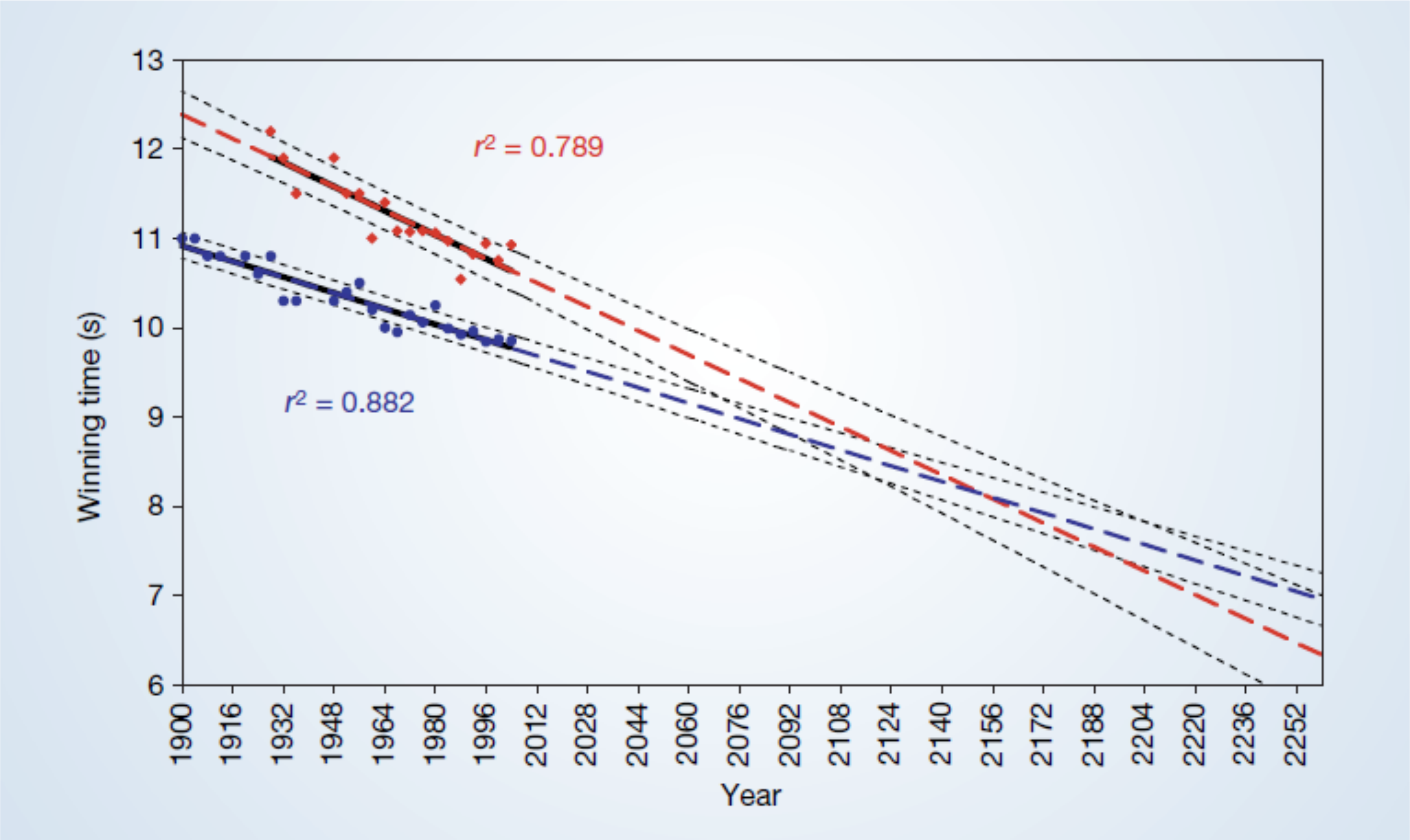


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

From Tatem et al., Nature 2004.
(see also response letters at http://faculty.washington.edu/kenrice/natureletter.pdf)

How Much Climate Data Do We Actually Have?

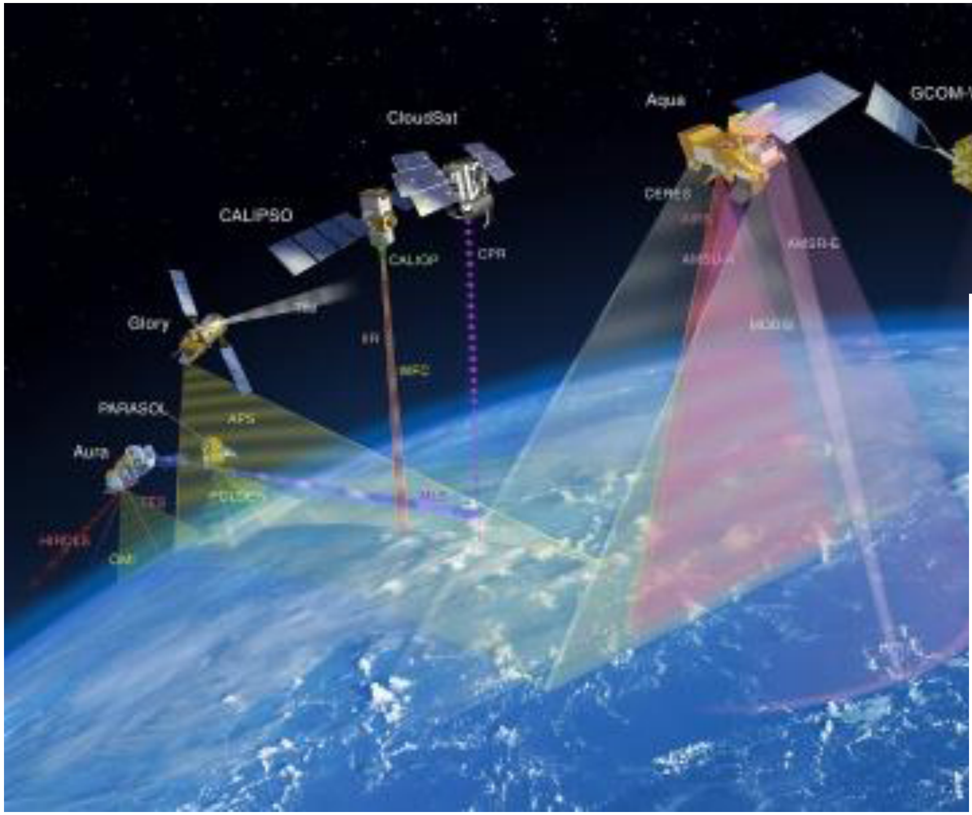


Image from http://cimss.ssec.wisc.edu/

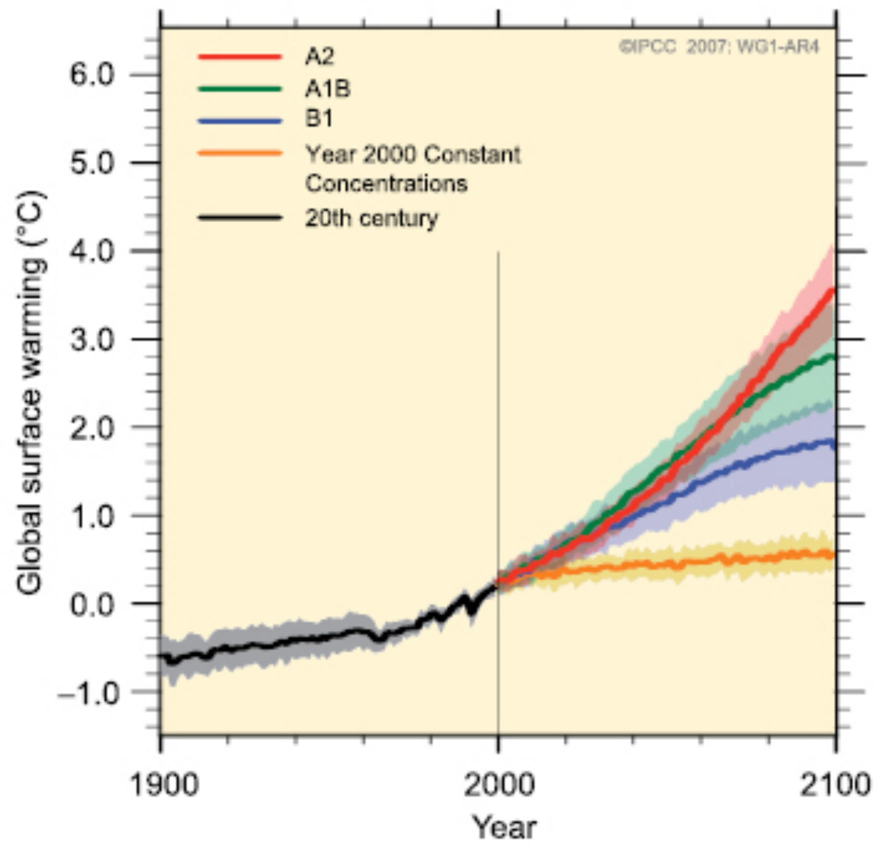


Image from ipcc.ch

Schedule of Lectures

Date	Speaker	Department Or Organization	Topic
Jan 9	Padhraic Smyth	Computer Science	Introduction to Data Science
Jan 16	Padhraic Smyth	Computer Science	Classification Algorithms in Machine Learning
Jan 23	Michael Carey	Computer Science	Databases and Data Management
Jan 30	Sameer Singh	Computer Science	Statistical Natural Language Processing
Feb 6	Zhaoxia Yu	Statistics	An Introduction to Cluster Analysis
Feb 13	Erik Sudderth	Computer Science	Computer Vision and Machine Learning
Feb 20	John Brock	Cylance, Inc	Data Science and CyberSecurity
Feb 27	Video Lecture (Kate Crawford)	Microsoft Research and NYU	Bias in Machine Learning
Mar 6	Matt Harding	Economics	Data Science in Economics and Finance
Mar 13	Padhraic Smyth	Computer Science	Review: Past and Future of Data Science