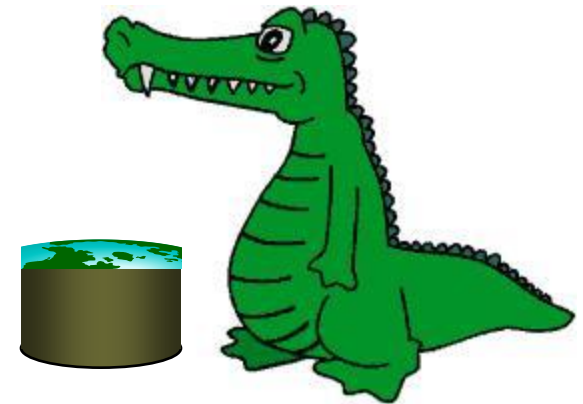


CAP4770/5771

Introduction to Data Science

Fall 2015

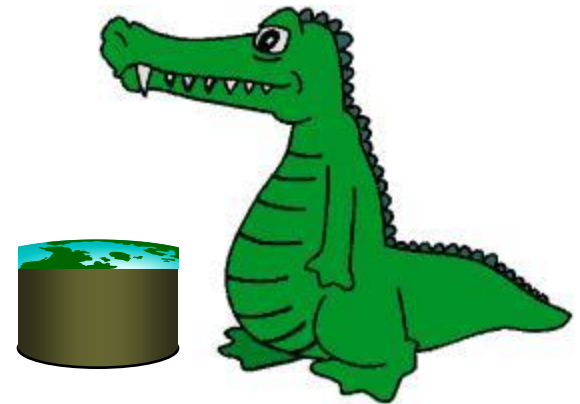
University of Florida, CISE Department
Prof. Daisy Zhe Wang



Based on notes from CS194 at UC Berkeley by Michael Franklin, John Canny, and Jeff Hammerbacher

Data Science Overview

Why, Where, What,
How, Who





Outline

- Data Science -- Why all the excitement?
 - history
 - examples
- Where does data come from?
- What is Data Science?
- How to do Data Science?
- Who are Data Scientists?



Data Science – Why all the excitement?



Data Analysis Has Been Around for a While...

1935: "The Design of Experiments"

R.A. Fisher



1939: "Quality Control"

W.E. Deming



1958: "A Business Intelligence System"

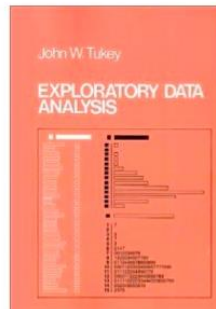


Peter Luhn

1997: "Machine Learning"



1977: "Exploratory Data Analysis"



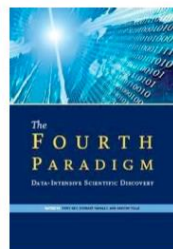
1989: "Business Intelligence"

Howard Dresner



2010: "The Data Deluge"

2007: "The Fourth Paradigm"



2009: "The Unreasonable Effectiveness of Data"



1996: Google





Data Science: Why all the Excitement?



Exciting new effective applications of data analytics

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Prediction model is built on
Various data sources,
types and analysis.



Why the all the Excitement?

elections2012

Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding

guardian.co.uk, Wednesday 7 November 2012 10.45 EST




Predicting political champagne and election Outcome:

*the signal and the
and the noise and
the noise and the
noise and the no
why most noise a
predictions fail to
but some don't n
and the noise and
the noise and the
nate silver noise
noise and the no*



Data and Election 2012 (cont.)

- ...that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, **helped save the president's candidacy**. In Chicago, the campaign recruited a team of behavioral scientists to build an **extraordinarily sophisticated database**
- ...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. **The power of this operation stunned Mr. Romney's aides on election night**, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.
-- New York Times, Wed Nov 7, 2012
- The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist, Feb. 18th 2015




A history of the (Business) Internet: 1997

BackRub Search: university

BackRub Query Results


BackRub's Highest Ranked Sites

University of Illinois at Urbana-Champaign

 <http://www.uiuc.edu/>


694.687 8460 backlinks 12k - 10/25/96 - 11/1/96

Stanford University Homepage

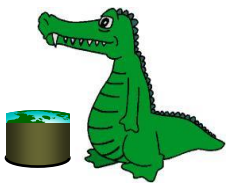
 <http://www.stanford.edu/>

609.303 8857 backlinks 4k - none - 11/1/96

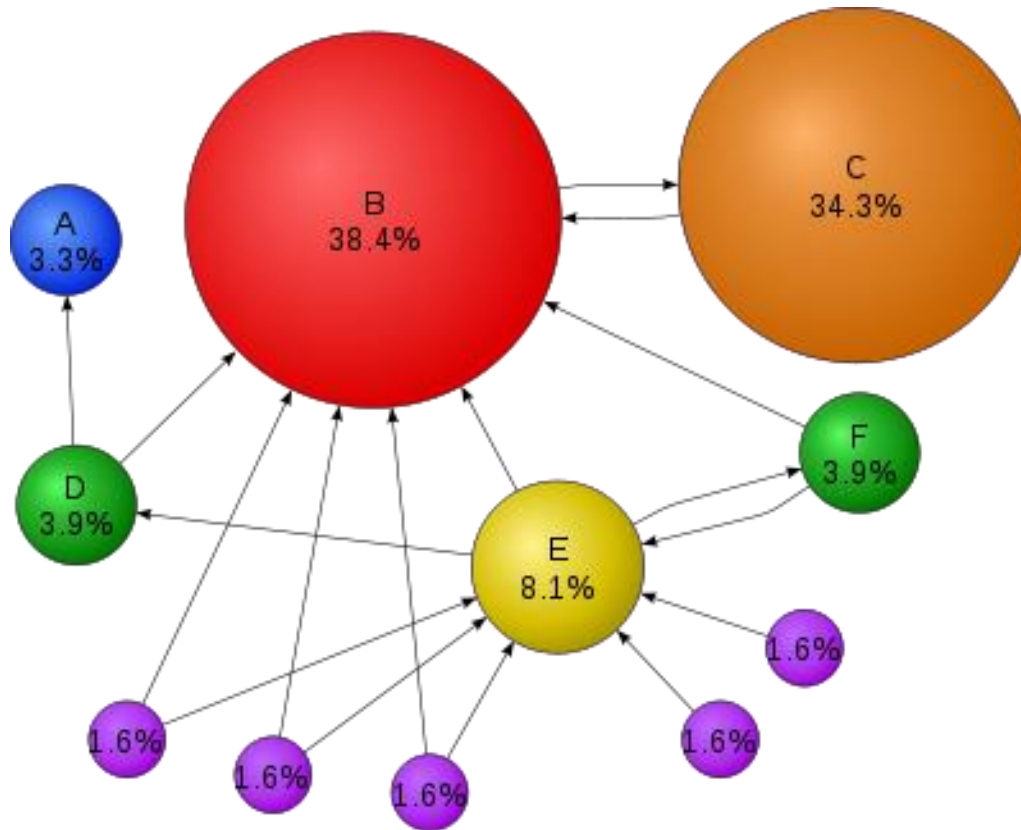
Stanford University: Portfolio Collection

 <http://www.stanford.edu/home/administration/portfolio.html>

167.919 34 backlinks



PageRank: The web as a behavioral dataset





Sponsored search

[Advanced Search](#)

Web [Show options...](#)

Results **1 - 10** of about **661,000** for **gatco towel bars**. (0.33 seconds)

Gatco Towel Bars

www.eFaucets.com/Gatco

Low Price Guarantee, Free Shipping. 110% Price Match, No Tax, Shop Now!



Gatco Towel Bars

www.AmericanHomePlus.com/Gatco

110% Low Price Guarantee. Buy Now. Free Shipping On Gatco Accessories.

Gatco Towel Bars

www.PlumberSurplus.com

Free S/H Available. Large Selection Gatco Towel Bars, ready to ship!



Gatco Towel bars, Gatco toilet paper holders, Gatco Bathroom ...

Gatco Towel bars, Gatco toilet paper holders, Gatco Bathroom Accessories, Gatco robe hooks, Gatco bathroom shelves, Gatco hotel shelves, Gatco double towel ...

www.kitchensnbath.com/gatco-bath-accessories.html - [Cached](#) - [Similar](#) -

Discount Towel Bars, Towel Bars, Towel Racks

Discount Towel Bars offers only the highest quality bath hardware. We are factory direct distributors for Moen, Baldwin, Dynasty Hardware and Gatco. ...

www.discounttowelbars.com/ - [Cached](#) - [Similar](#) -

Gatco at Lowe's: 24" Franciscan Chrome Double Towel Bar

24" Franciscan Chrome Double Towel Bar - 69656 5286.

www.lowes.com/lowes/lkn?action=productDetail... - [Cached](#) - [Similar](#) -

Shopping results for gatco towel bars



[Gatco Bleu 18 in. Towel Bar - Polished Chrome](#)

\$39.99 new - [Sears](#)

[Gatco 4240 24-Inch Latitude II Towel Bar. Chrome](#)

\$30.53 new - [Amazon.com](#)

[4621 Camden Towel Bar 18 Gatco Inc](#)

Sponsored Links

Sponsored Links



[Gatco Bleu 18 in. Towel Bar - Polished Chrome](#)

\$39.99 - [Sears](#)



[Gatco Spa Towel Rack, 3 Tier - Satin Nickel](#)

\$94.99 - [Sears](#)



[Gatco Chenille 18 in. Towel Bar - Vintage ...](#)

\$44.99 - [Sears](#)

Gatco Bath Accessories

Lowest prices.

All Gatco collections

www.TheHomeDecor.net

Towel Bars on Sale

Save 20%-50% Off List- New Styles

Lowest Prices + Free Shipping!

www.FixtureUniverse.com





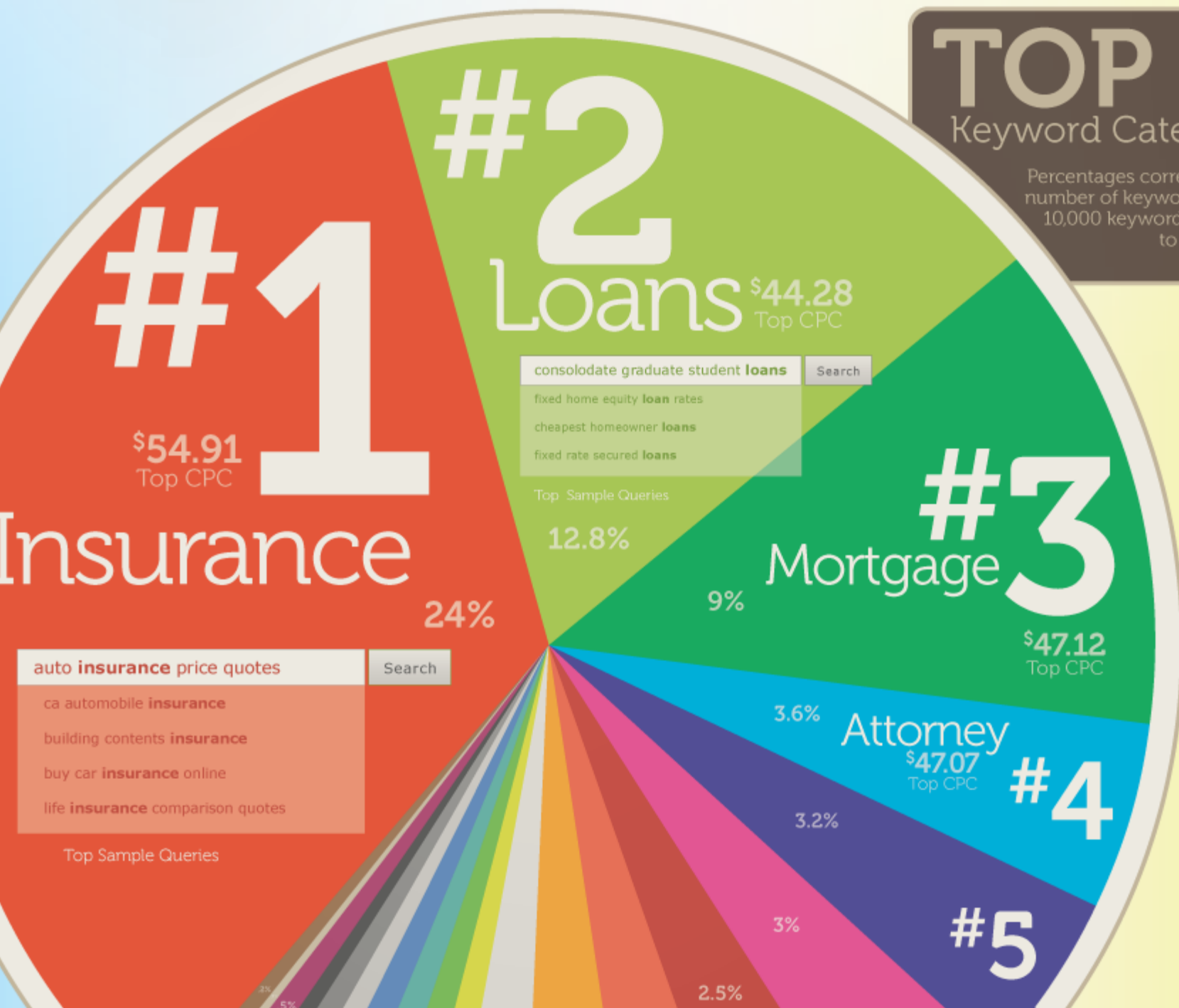
Sponsored search

- Google revenue around \$50 bn/year from marketing, 97% of the companies revenue.
- Sponsored search uses an auction – a pure competition for marketers trying to win access to consumers.
- In other words, a competition for **models** of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.
- There are around 30 billion search requests a month. Perhaps a **trillion events** of history between search providers.
- Google Adwords and Adsense

TOP 20

Keyword Categories

Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.





Other Data Science Applications

- Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)
- Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things
- Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery
- Software Log Data → Automatic Trouble Shooting (Splunk)
- Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine

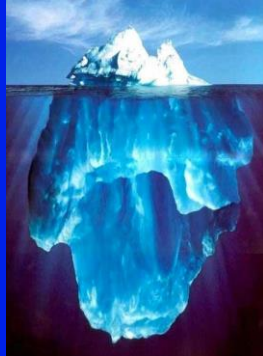


Where does data come from?



"Big Data" Sources

It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,...
Server request
Transaction
Network message
Fault

...

User Generated (Web & Mobile)



...

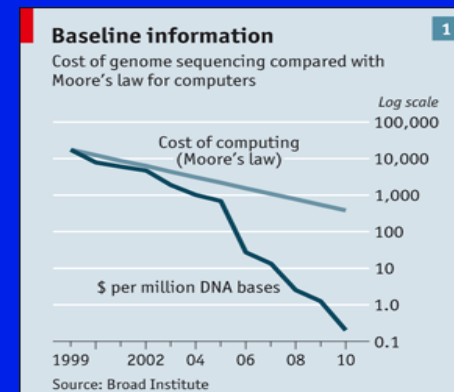
..



Internet of Things / M2M



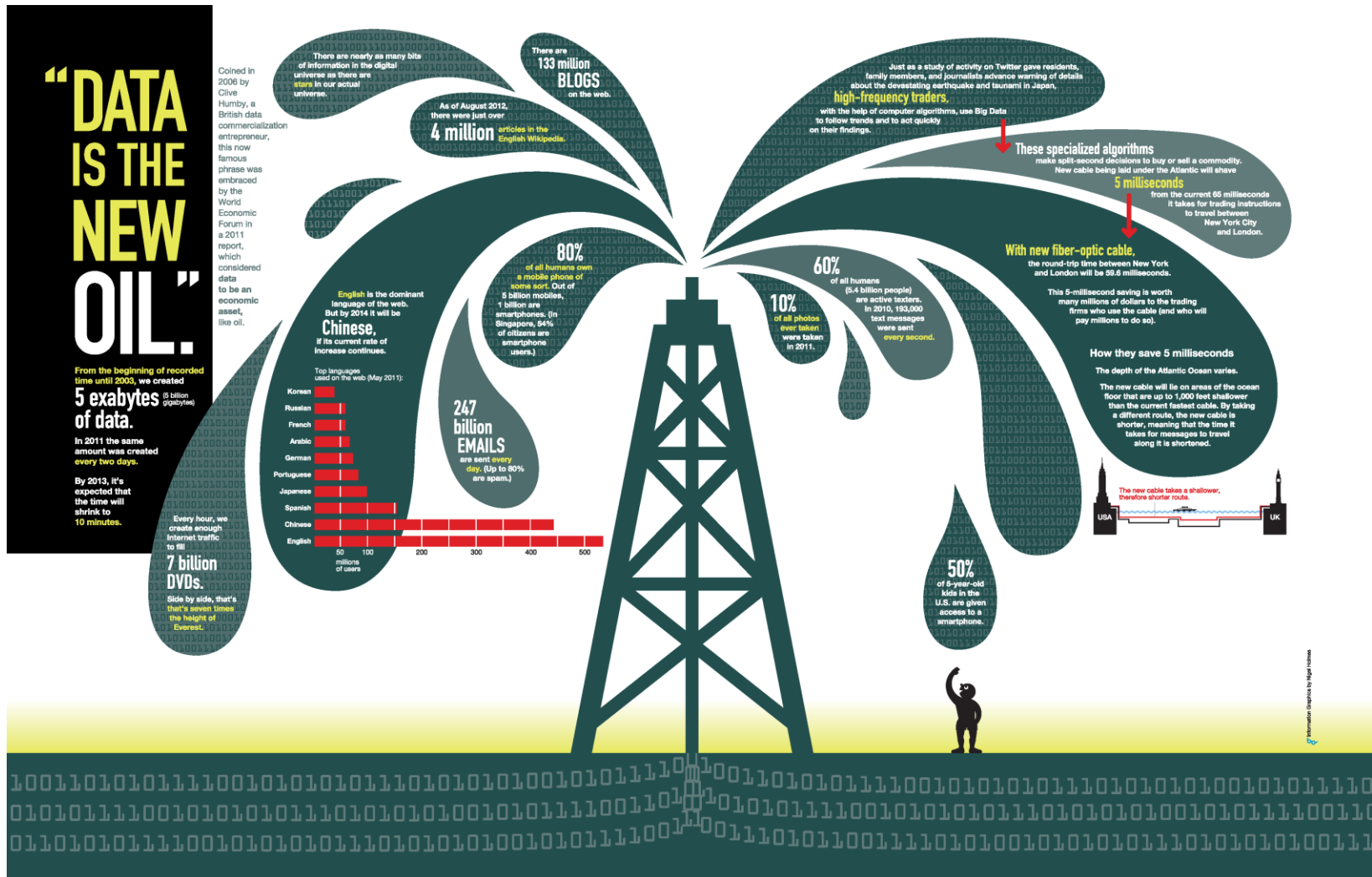
Health / Scientific Computing





"Data is the New Oil"

– World Economic Forum 2011





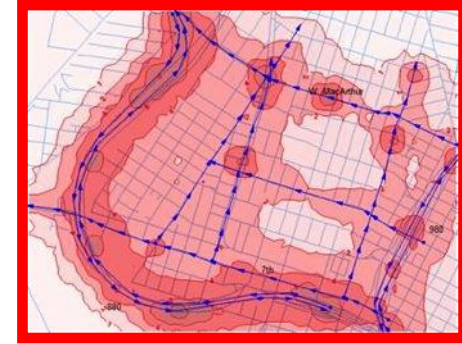
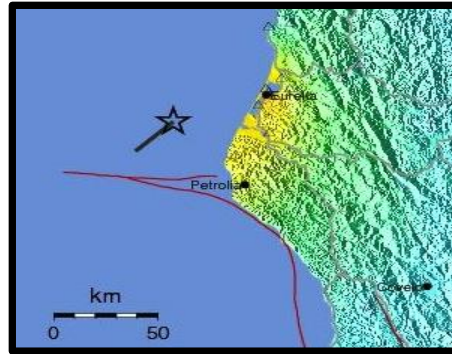
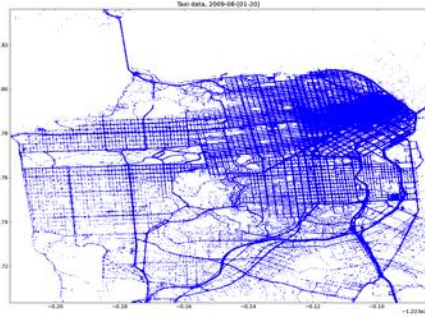
5 Vs of Big Data

- Raw Data: Volume
- Change over time: Velocity
- Data types: Variety
- Data Quality: Veracity
- Information for Decision Making: Value



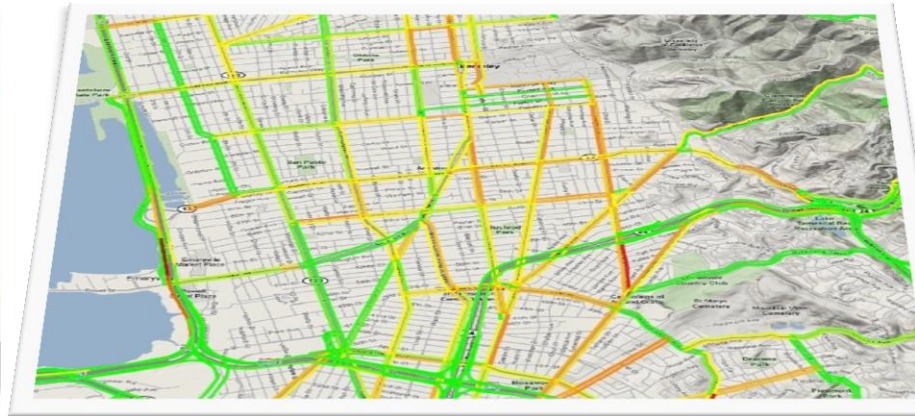
What can you do with the data?

Traffic Prediction and Earthquake Warning



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



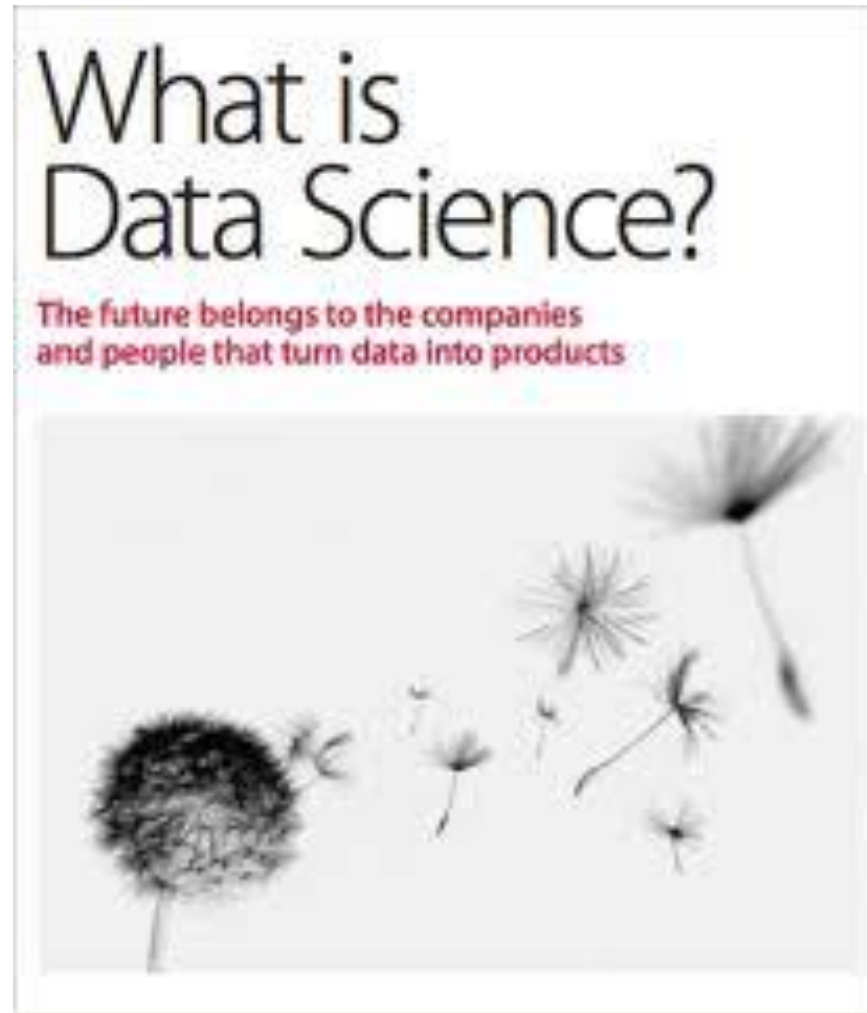
From Alex Bayen, UCB, Director, Institute for Transportation Studies



What is Data Science?



"Data Science" an Emerging Field



O'Reilly Radar report, 2011



Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to **collect, clean, integrate, analyze, visualize, interact** with **data** to **create data products**.















Goal of Data Science

Turn **data** into **data products**.



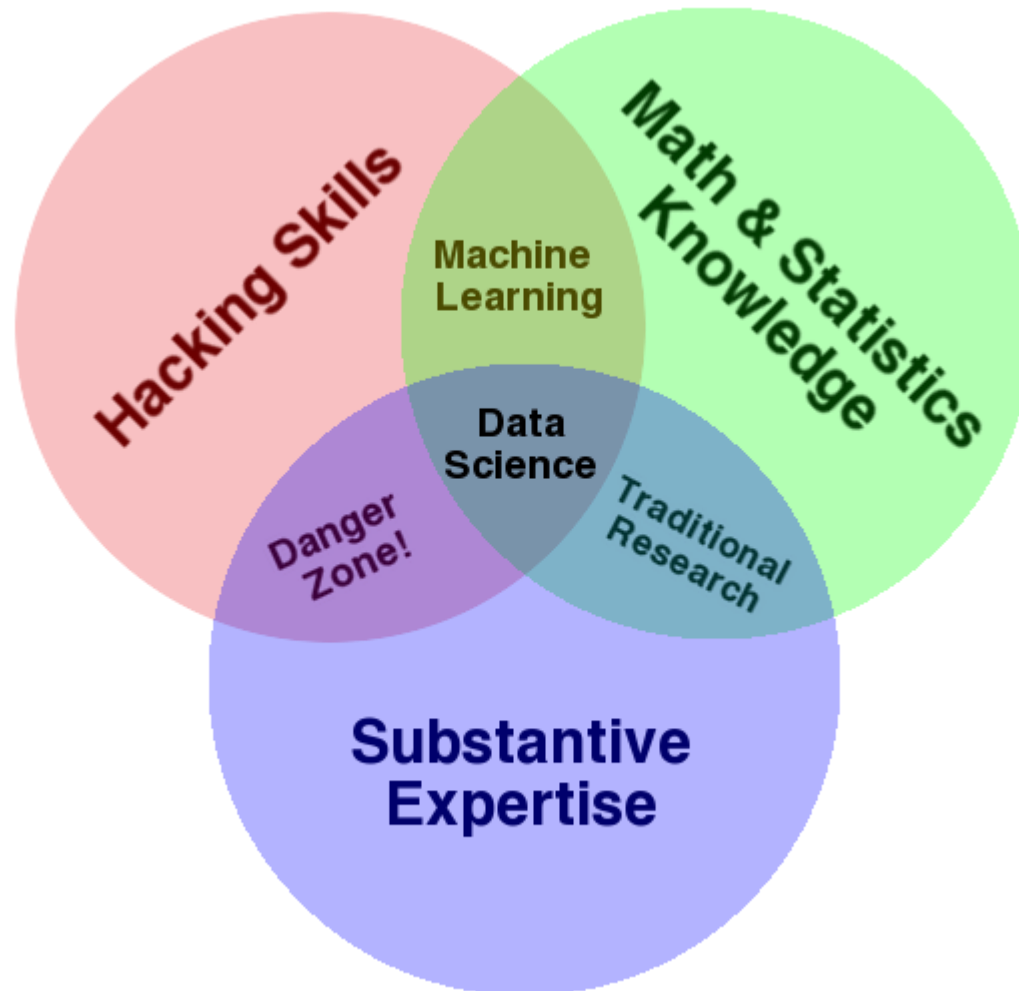
Some recent ML Competitions at <https://www.kaggle.com/>

NIST Pre-Pilot Data Science Evaluation – likely to be incorporated to be part of Labs/Final project

Active Competitions				kaggle	
		Flight Quest 2: Flight Optimization Final Phase of Flight Quest 2	33 days Coming soon \$220,000		
		Packing Santa's Sleigh He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams \$10,000		
		Flu Forecasting  Predict when, where and how strong the flu will be	41 days 37 teams		
		Galaxy Zoo - The Galaxy Challenge Classify the morphologies of distant galaxies in our Universe	2 months 160 teams \$16,000		
		Loan Default Prediction - Imperial College Lon... Constructing an optimal portfolio of loans	52 days 82 teams \$10,000		
		Dogs vs. Cats Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag		



Data Science – A Visual Definition





Contrast: Databases

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra, Riak, Memcached, Apache River, ...

ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition
Tolerance



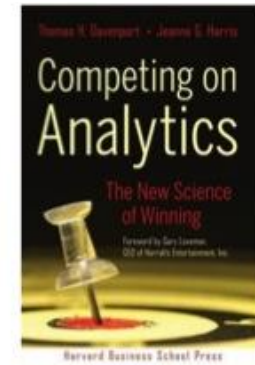
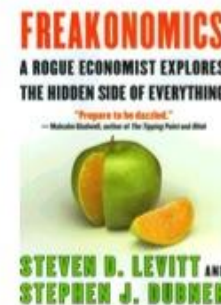
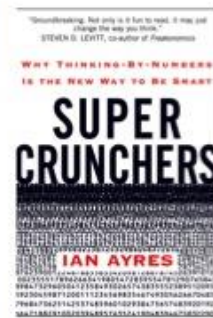
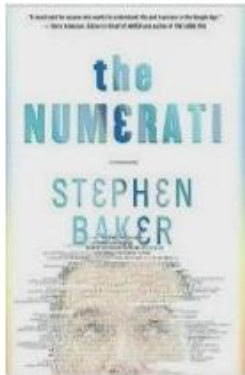
Contrast: Business Intelligence

**Business
Intelligence**

Querying the past

Data Science

Querying the past
present and future





Contrast: Machine Learning

Machine Learning

Develop new (individual) models

Prove mathematical properties of models

Improve/validate on a few, relatively clean, small datasets

Publish a paper

Data Science

Explore many models, build and tune hybrids

Understand empirical properties of models

Develop/use tools that can handle massive datasets

Take action!