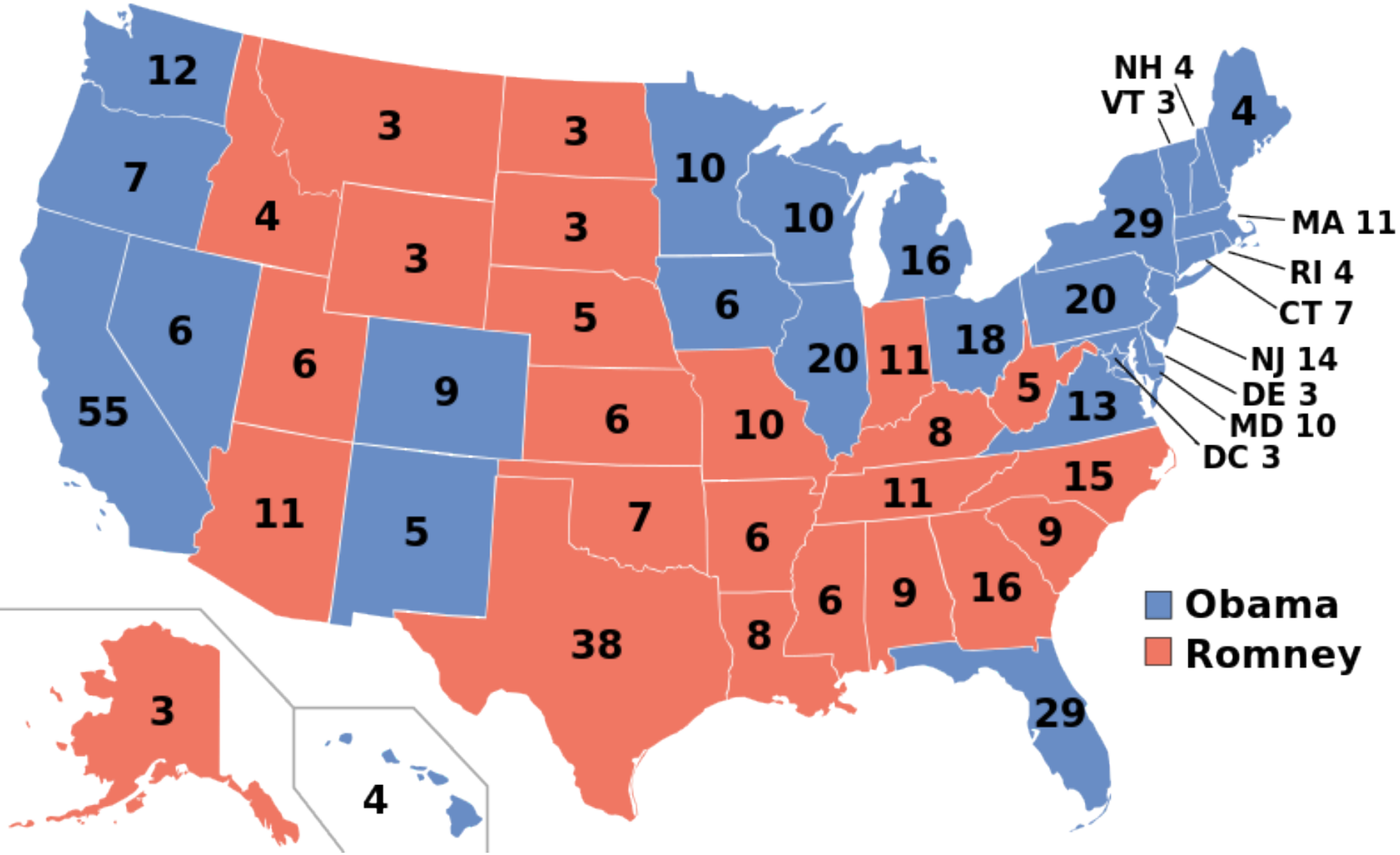
INTRODUCTION TO DATA SCIENCE

BILL HOWE, PHD

DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS

UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Nate Silver

source: randy stewart

"The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes."

Nate Silver, Oct. 26, 2012

fivethirtyeight.com

"…the argument we're making is exceedingly simple. Here it is: Obama's ahead in Ohio."

Nate Silver, Nov 2, 2012

fivethirtyeight.com

"The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign."

Nate Silver, Nov. 10, 2012

DailyBeast

OBAMA CAMPAIGN'S DATA-DRIVEN GROUND GAME

"In the 21st century, the candidate with [the] best data, merged with the best messages dictated by that data, wins."

Andrew Rasiej, Personal Democracy Forum

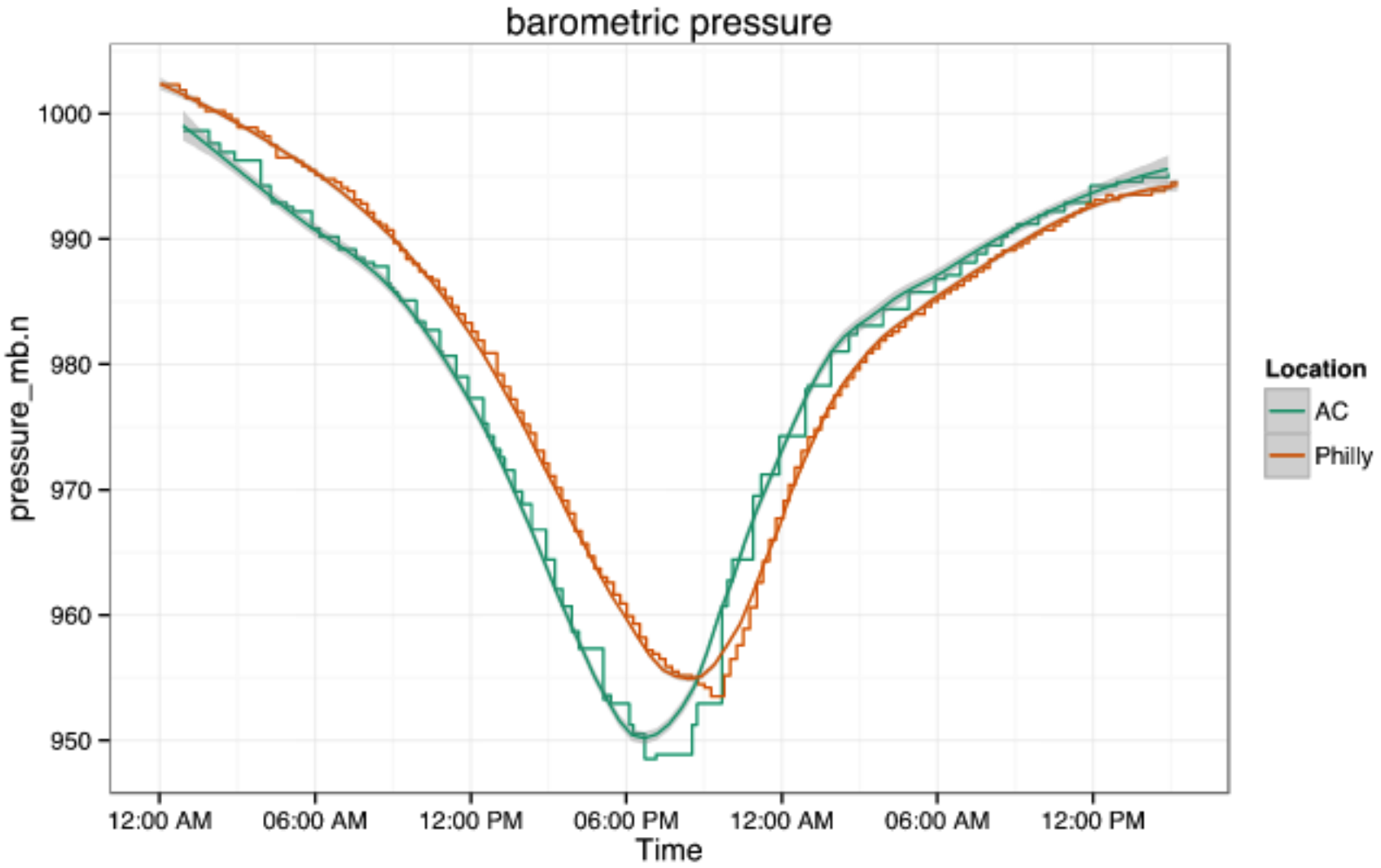
"…the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

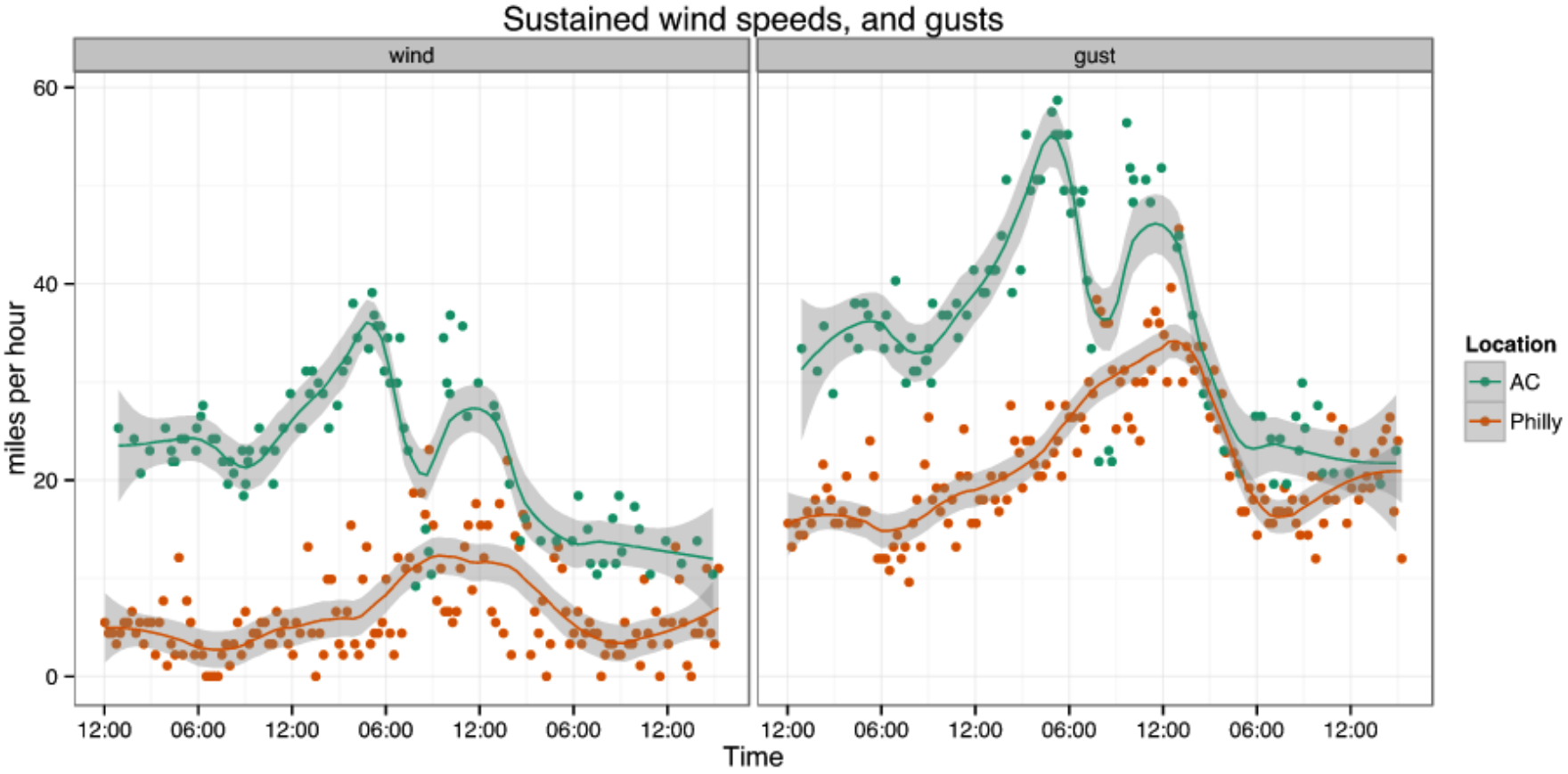
HURRICANE SANDY



http://rpubs.com/JoFrhwld/sandy

Josef Fruehwald

HURRICANE SANDY



http://rpubs.com/JoFrhwld/sandy

Josef Fruehwald

1) Convert all the digitized books in the 20th century into n-grams (Thanks, Google!)

(http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

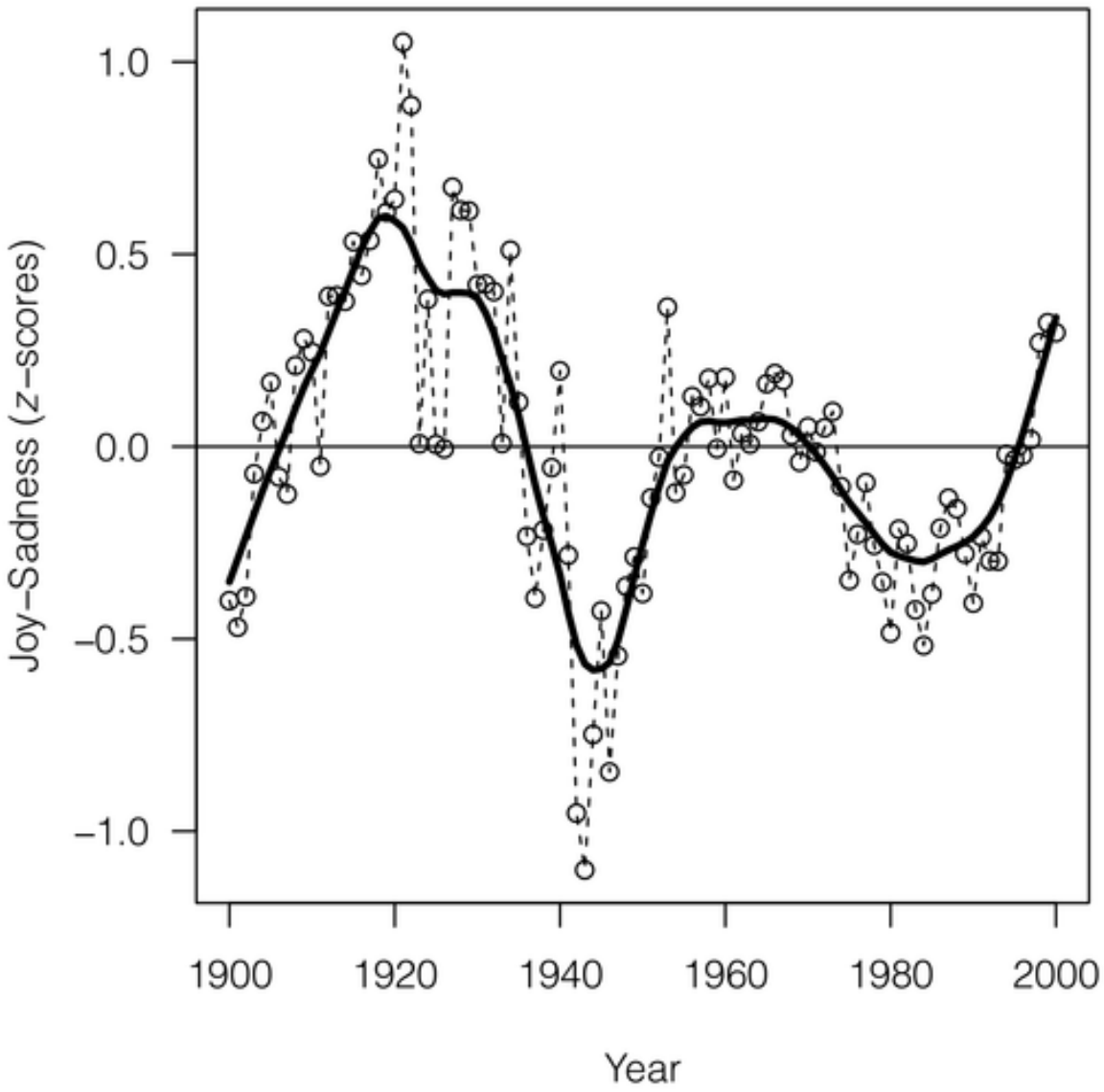
2) Label each 1-gram (word) with a mood score. (Thanks, WordNet!)

3) Count the occurences of each mood word

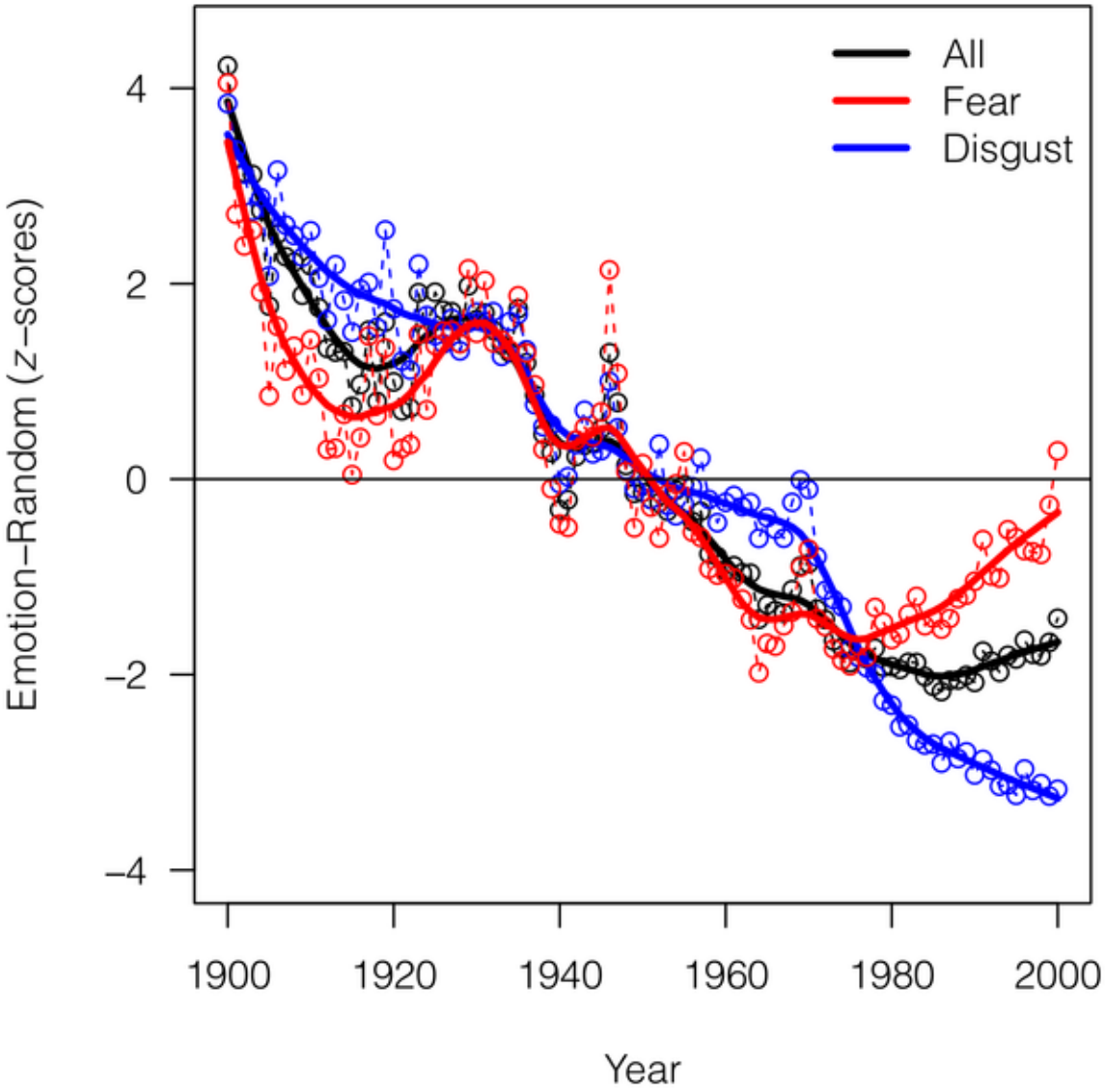
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

…
2. Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) ***Quantitative analysis of culture using millions of digitized books***. Science 331: 176–182. doi: 10.1126/science.1199644. Find this article online
3. Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) ***Quantifying the evolutionary dynamics of language***. Nature 449: 713–716. doi: 10.1038/nature06137. Find this article online
4. Pagel M, Atkinson QD, Meade A (2007) ***Frequency of word-use predicts rates of lexical evolution throughout Indo-European history***. Nature 449: 717–720. doi: 10.1038/nature06176. Find this article online
…
6. DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) ***Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics***. Psychology of Aesthetics, Creativity and the Arts 5: 200–207. doi: 10.1037/a0023195. Find this article online
…

Related: Obama campaign's data-driven ground game

"In the 21st century, the candidate with [the] best data, merged with the best messages dictated by that data, wins."

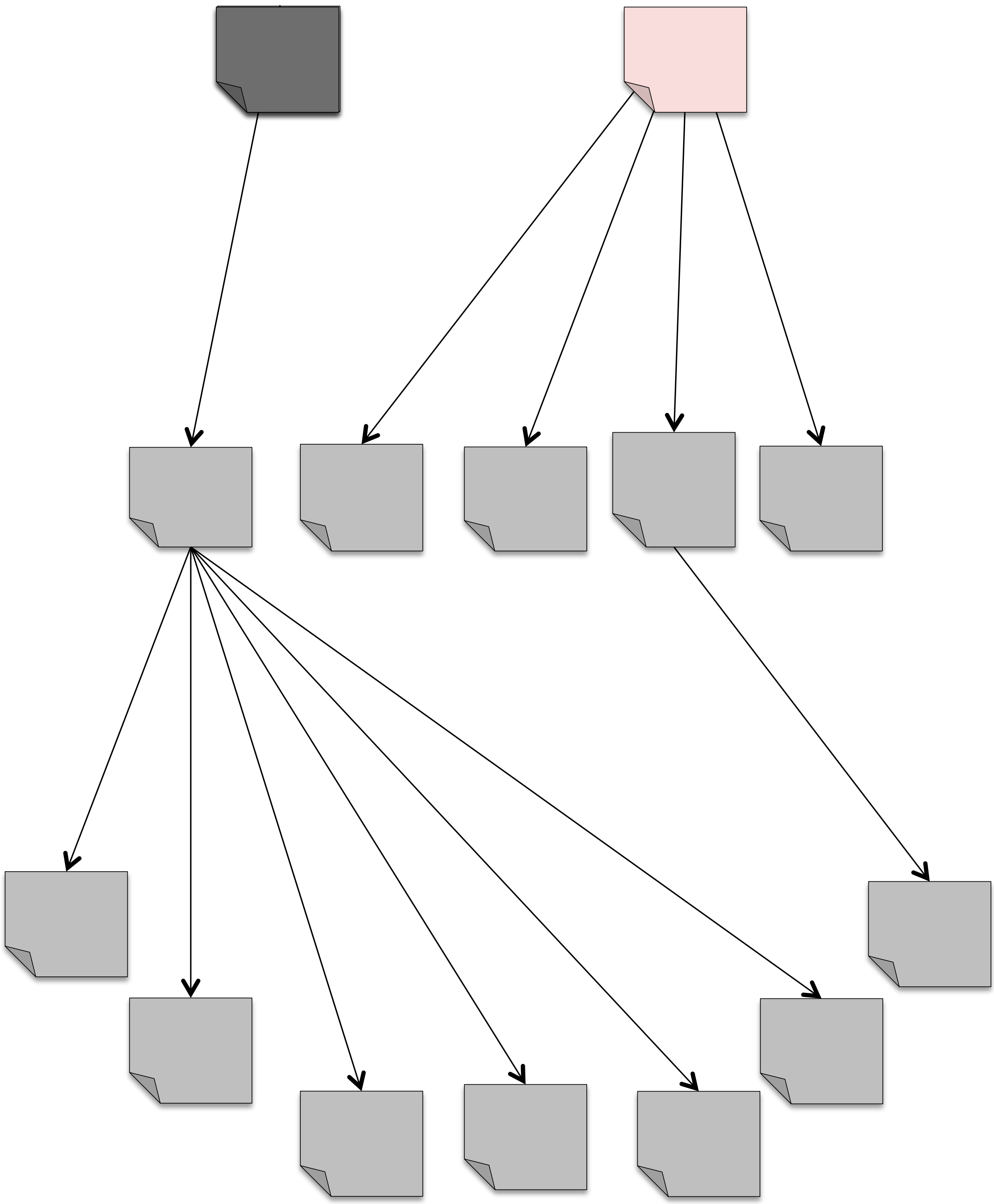
Andrew Rasiej, Personal Democracy Forum

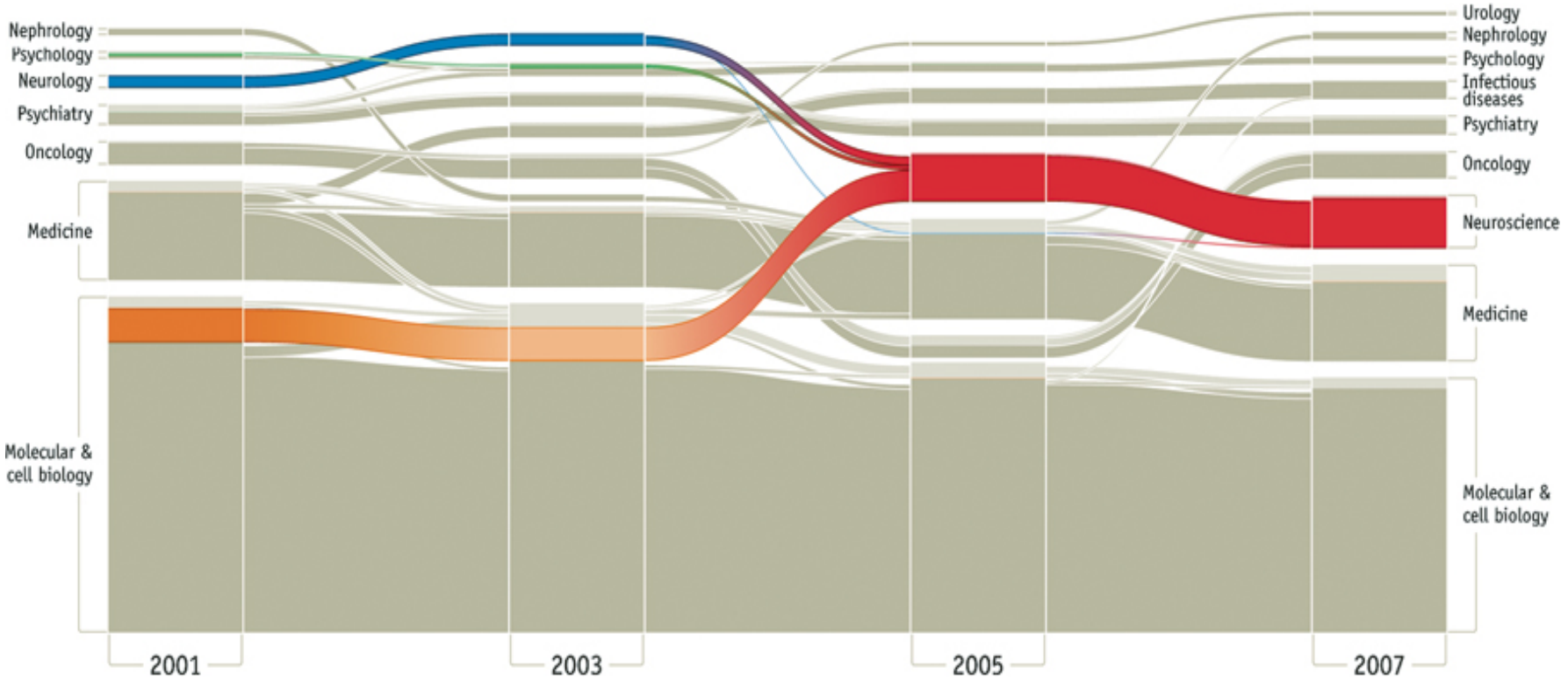
"…the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies





Carl Bergstrom, Martin Rosvall, 2011

WHAT ARE THE LIMITS?

Are subjective topics such as food and music preferences out of reach for quantitative analysis?

Flavor network and the principles of food pairing

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási

Affiliations | **Contributions** | **Corresponding authors**

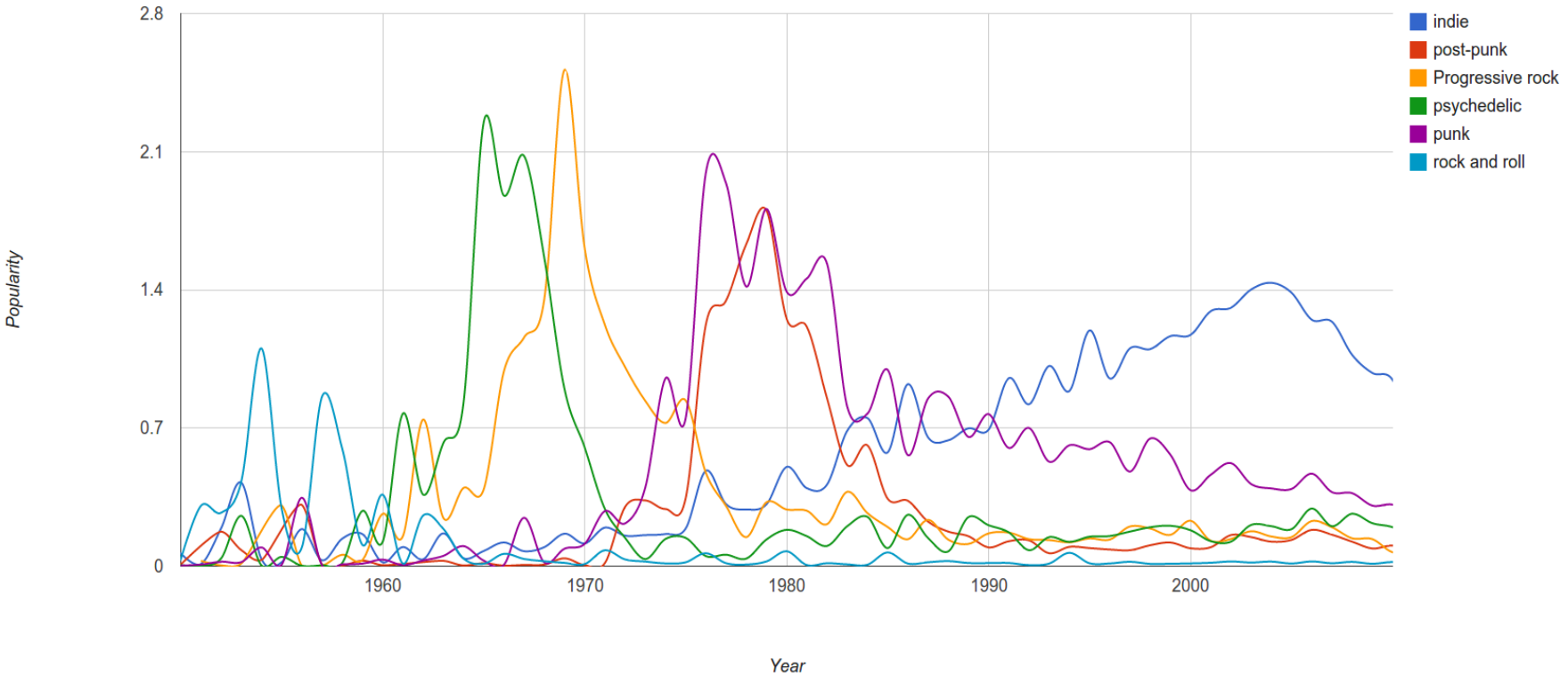
Scientific Reports **1**, Article number: 196 | doi:10.1038/srep00196

Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.

MUSIC POPULARITY

Last.FM



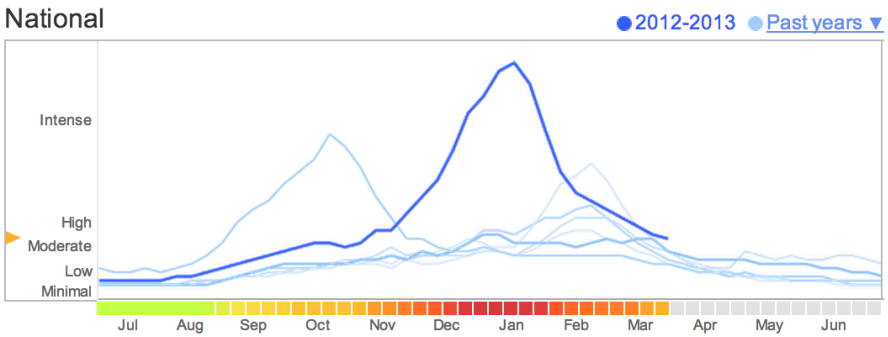
"Since we have a massive amount of user tag data available we can easily correlate tags and years and measure "popularity" of a genre by counting the number of artists formed in a specific year."

Janni Kovacs, Last.FM

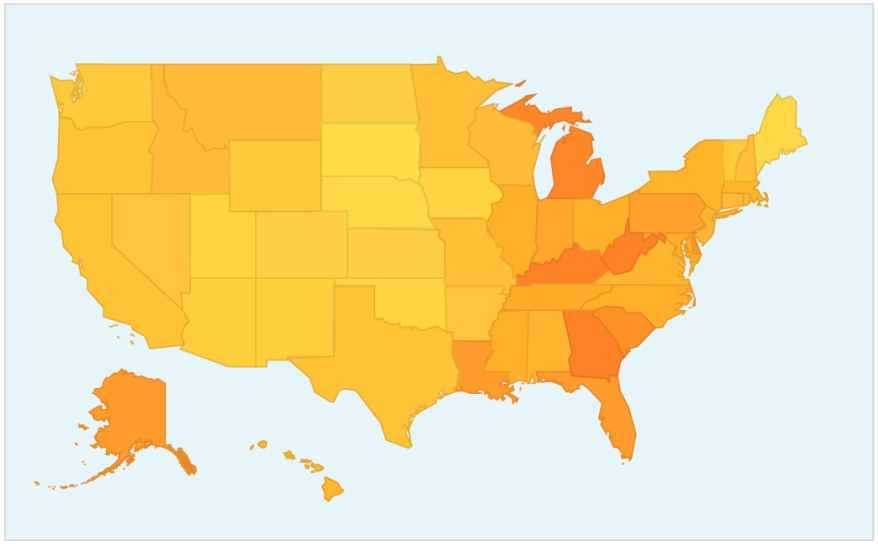
FLU TRENDS

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



States | Cities (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.

http://www.google.com/permissions/using-product-graphics.html

source:
http://www.google.org/flutrends/us/#US

flu risk

"Scientific hindsight shows that Google Flu Trends far overstated this year's flu season…."

"Lots of media attention to this year's flu season skewed Google's search engine traffic."

David Wagner, Atlantic Wire, Feb 13 2013

PREDICTIONS

October 22, 2012

Six Italian seismologists convicted of manslaughter for failing to predict magnitude 6.3 earthquake in April 2009.

Locals were concerned about seismic activity; researchers deemed "too reassuring" in the verdict.

SOME RECURRING THEMES

simple methods

repurposing data

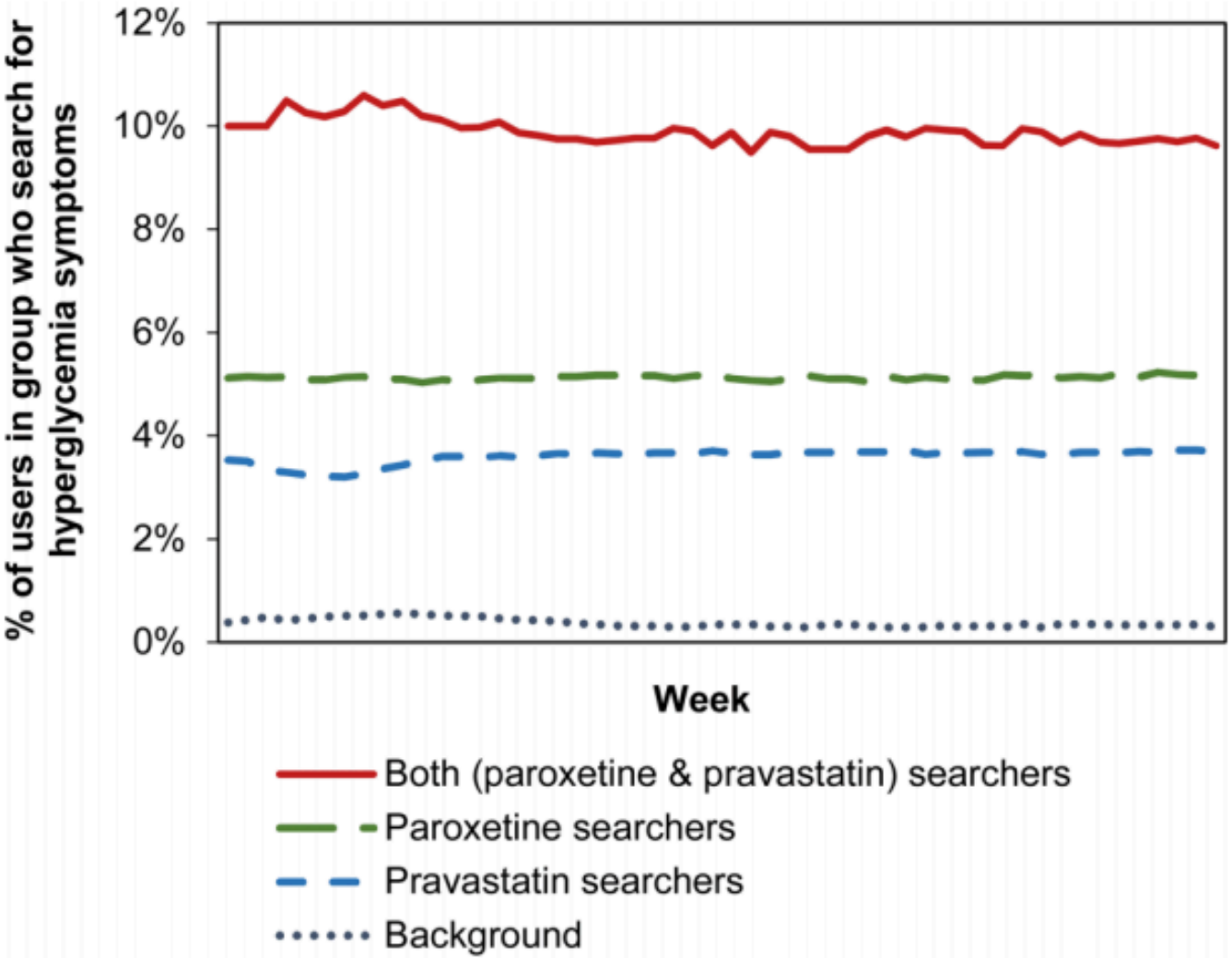
communication matters

Graph analytics
Databases
Visualization
Large datasets

Other themes

- "Data products" – not just answers
- "Speed of thought" analysis

SIGNALS FROM THE CROWD



Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, Eric Horvitz, ***Web-scale pharmacovigilance: listening to signals from the crowd****, J Am Med Inform Assoc, March 2013, doi:10.1136/amiajnl-2012-001482*

INTRODUCTION TO DATA SCIENCE

BILL HOWE, PHD

DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS

UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE

WHAT IS DATA SCIENCE?

Fortune

- "Hot New Gig in Tech"

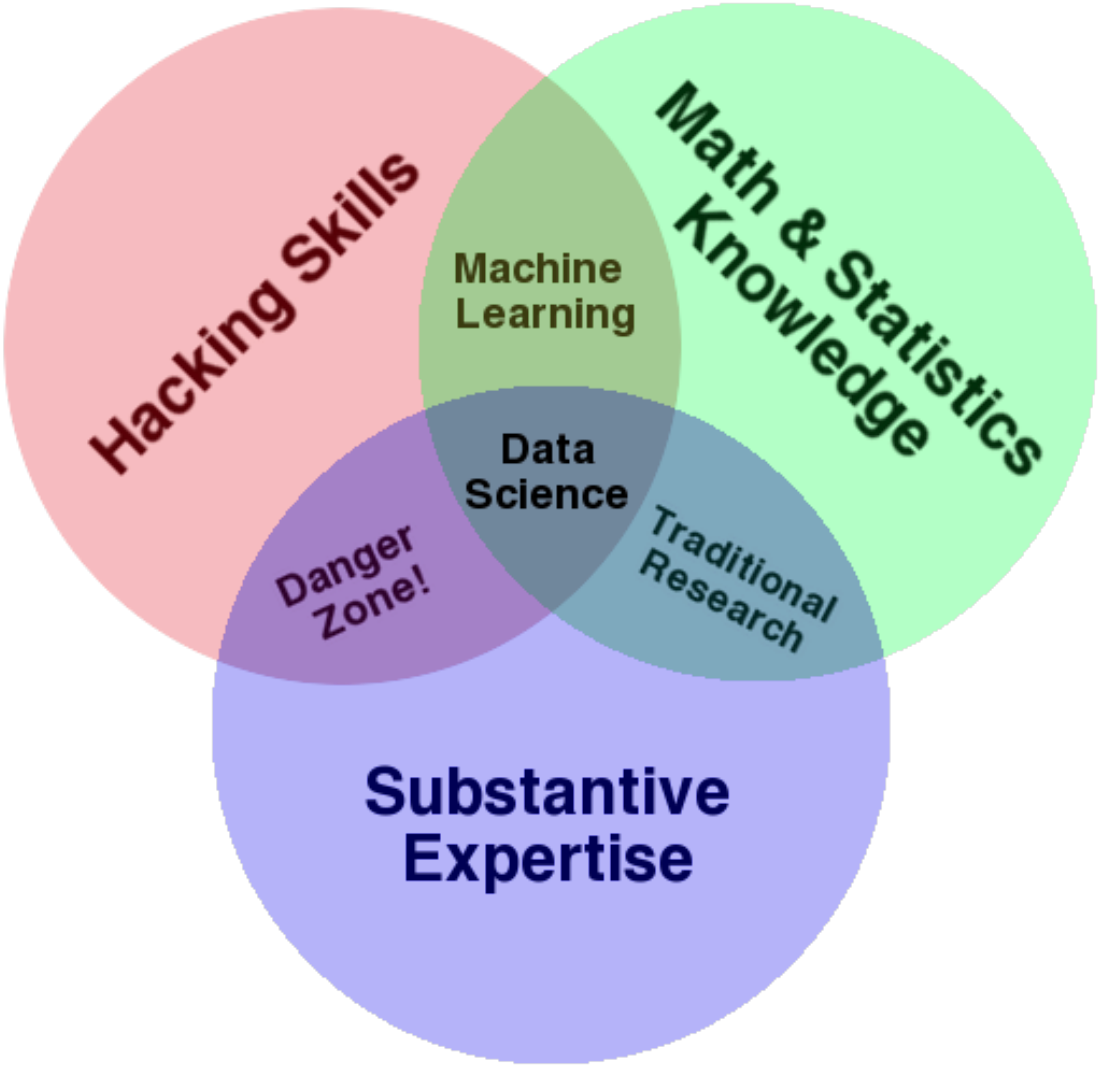
Hal Varian, Google's Chief Economist, NYT, 2009:

- "The next sexy job"
- "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."

Mike Driscoll, CEO of metamarkets:

- "Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics."
- "Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

DREW CONWAY'S DATA SCIENCE VENN DIAGRAM



WHAT DO DATA SCIENTISTS DO?

"They need to find nuggets of truth in data and then explain it to the business leaders"

- Richard Snee, EMC

Data scientists "tend to be "hard scientists", particularly **physicists**, rather than computer science majors. Physicists have a strong **mathematical background, computing skills**, and come from a discipline in which survival depends on **getting the most from the data.** They have to think about the big picture, the big problem."

- DJ Patil, Chief Scientist at LinkedIn

MIKE DRISCOLL'S THREE SEXY SKILLS OF DATA GEEKS

Statistics

- traditional analysis

Data Munging

- parsing, scraping, and formatting data

Visualization

- graphs, tools, etc.

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information."

- Jeffrey Stanton
Syracuse University School of Information Studies
An Introduction to 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

"data wrangling"
"data jujitsu"
"data munging"

THREE TYPES OF TASKS:

1) Preparing to run a model

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

2) Running the model

3) Communicating the results

DATA SCIENCE IS ABOUT *DATA PRODUCTS*

"Data-driven apps"

- Spellchecker
- Machine Translator

(Mike Loukides)

Interactive visualizations

- Google flu application
- Global Burden of Disease

Online Databases

- Enterprise data warehouse
- Sloan Digital Sky Survey

Data science is about building data products, not just answering questions

Data products empower others to use the data.

May help communicate your results (e.g., Nate Silver's maps)

May empower others to do their own analysis (e.g., Global Burden of Disease)

DISTINGUISHING DATA SCIENCE FROM…

Business Intelligence

Statistics

Data(base) Management

Visualization

Machine Learning

Huge number of relevant courses, new and existing.

- Concepts in Computing with Data, Berkeley
- Practical Machine Learning, Berkeley
- Artificial Intelligence, Berkeley
- Visualization, Berkeley
- Data Mining and Analytics in Intelligent Business Services, Berkeley
- Data Science and Analytics: Thought Leaders, Berkeley
- Scalable Machine Learning, Berkeley
- Analyzing Big Data with Twitter, Berkeley
- Machine Learning, Stanford
- Paradigms for Computing with Data, Stanford
- Mining Massive Data Sets, Stanford
- Data Visualization, Stanford
- Algorithms for Massive Data Set Analysis, Stanford
- Research Topics in Interactive Data Analysis, Stanford
- Data Mining, Stanford
- Machine Learning, CMU
- Statistical Computing, CMU
- Machine Learning with Large Datasets, CMU
- Machine Learning, MIT
- Data Mining, MIT
- Statistical Learning Theory and Applications, MIT
- Data Literacy, MIT
- Introduction to Data Mining, UIUC
- Learning from Data, Caltech
- Introduction to Statistics, Harvard
- Data-Intensive Information Processing Applications, University of Maryland
- Statistical Inference, UPenn
- Introduction to Data Science, Columbia
- Dealing with Massive Data, Columbia
- Data-Driven Modeling, Columbia
- Introduction to Data Mining and Analysis, Georgia Tech
- Computational Data Analysis: Foundations of Machine Learning and Data Mining, Georgia Tech
- Applied Statistical Computing, Iowa State
- Data Visualization, Rice
- Data Warehousing and Data Mining, NYU
- Data Mining in Engineering, Toronto
- Machine Learning and Data Mining, UC Irvine
- Knowledge Discovery from Data, Cal Poly
- Large Scale Learning, University of Chicago
- Data Science: Large-scale Advanced Data Analysis, University of Florida
- Strategies for Statistical Data Analysis, Universität Leipzig
- Data Analysis, Johns Hopkins (via Coursera)
- Computing for Data Analysis, Johns Hopkins (via Coursera)

"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

- Aaron Kimball, CTO Wibidata

WHAT "DATA SCIENCE" TELLS ME:

If you're a DBA, you need to learn to deal with unstructured data

If you're a statistician, you need to learn to deal with data that does not fit in memory

If you're a software engineer, you need to learn statistical modeling and how to communicate results.

If you're a business analyst, you need to learn about algorithms and tradeoffs at scale

BREADTH

TOOLS

ABSTRACTIONS

Hadoop	MapRedu=ce
PostgreSQL	Relational Algebra
glm(…) in R	Logistic Regression
Tableau	InfoVis

DEPTH

STRUCTURES

STATISTICS

Management

Relational Algebra

Standards

Analysis

Linear Algebra

ad hoc files

SCALE

DESKTOP

CLOUD

main memory

R

local files

distributed

Hadoop

S3, Azure Storage

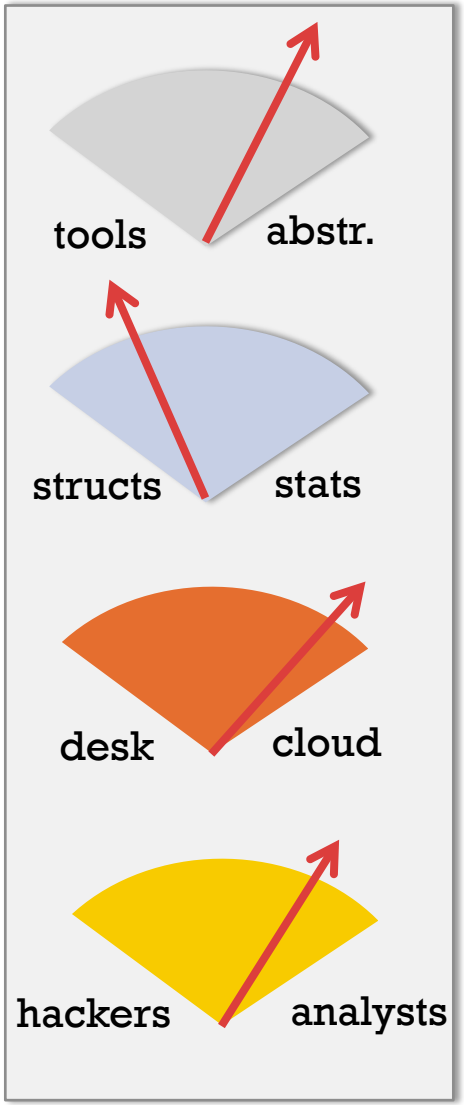
TARGET

HACKERS

ANALYSTS

Assume proficiency in Python, Java, R

Assume little or no programming


tools
abstr.
structs
stats
desk
cloud
hackers
analysts

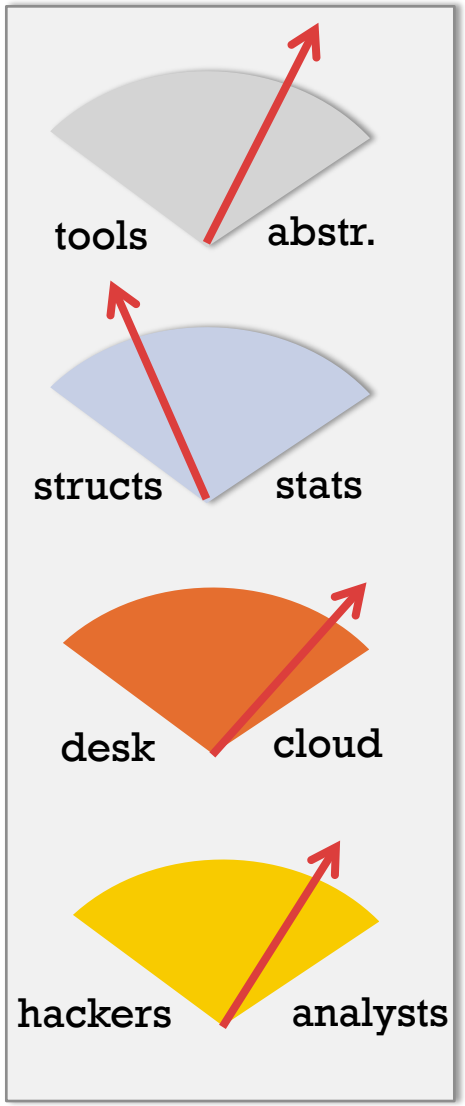
DIMENSIONS

BILL HOWE, PHD

DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS

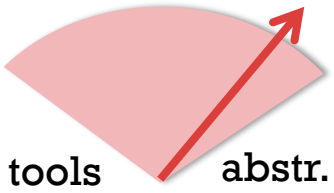
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE

THIS COURSE



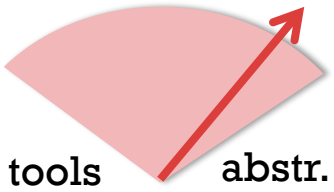
WHAT GOES AROUND COMES AROUND

- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop
- 2009 Hbase: Indexing for Hadoop
- 2010 Dietrich et al. Schemas and Indexing for Hadoop
- 2012 Transactions in HBase (plus VoltDB, other NewSQL systems)



- But also some permanent contributions:
  - Fault tolerance
  - Schema-on-Read
  - User-defined functions that don't suck

ABSTRACTIONS OF DATA SCIENCE

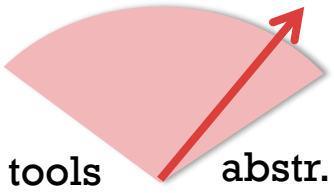


"Data Jujitsu"
"Data Wrangling"
"Data Munging"

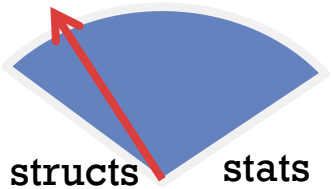
Translation: "We have no idea what this is all about"

ABSTRACTIONS OF DATA SCIENCE

matrices and linear algebra?
relations and relational algebra?
objects and methods?
files and scripts?
data frames and functions?



STRUCTURES



"80% of analytics is sums and averages"

- Aaron Kimball, wibidata

God created the integers; all else is the work of man

Codd created relations; all else is the work of man

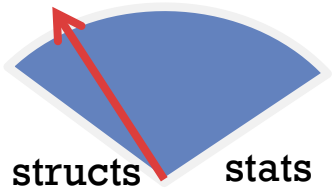
THREE TYPES OF TASKS:

1) Preparing to run a model

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

"80% of the work"

- Aaron Kimball

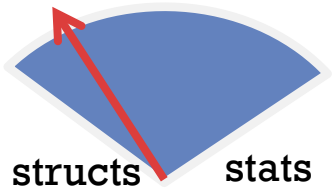


2) Running the model

3) Interpreting the results

DEALING WITH DATA

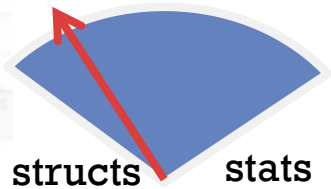
How much time do you spend "handling data" as opposed to "doing science"?



Mode answer: "90%"

STATISTICAL TOOLS

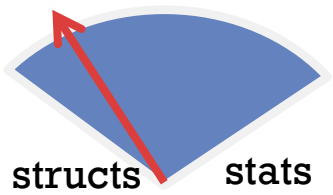
- Databases and Statistical Packages
  - Many analysts download data to use in Excel/SAS/ Matlab/R or their favorite programming language?
    FORTRAN??
  - Use matrix/vector operations
  - Most of these stat packages require data to fit in RAM
    - Taking samples from the full data to fit into ram results in loss of precision
  - External toolkits may also lack parallelism



src: Christian Grant, MADSkills

SPARSE MATRIX MULTIPLY IN SQL

```
SELECT A.row_number, B.column_number, SUM(A.value * B.value)
FROM A, B
WHERE A.column_number = B.row_number
GROUP BY A.row_number, B.column_number
```



src: Christian Grant, MADSkills

ASIDE: SCHEMA-ON-WRITE VS. SCHEMA-ON-READ

A schema* is a shared consensus about some universe of discourse

At the frontier of research, this shared consensus does not exist, *by definition*

Any schema that does emerge will change frequently, *by definition*

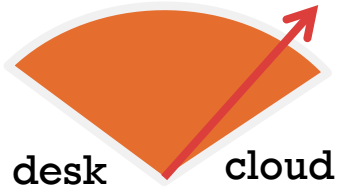
Data found "in the wild" will typically not conform to any schema, *by definition*

But this doesn't mean we have to live with ad hoc scripts and files

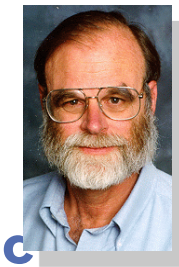
My answer: Schema-later, "lazy schemification"

* ontology/metadata standard/controlled vocabulary/etc.

DATA ACCESS HITTING A WALL



Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow



You can GREP 1 MB in a second

You can GREP 1 GB in a minute

You can GREP 1 TB in 2 days

You can GREP 1 PB in 3 years.

You can FTP 1 MB in 1 sec

You can FTP 1 GB / min (~1$)

… 2 days and 1K$

… 3 years and 1M$

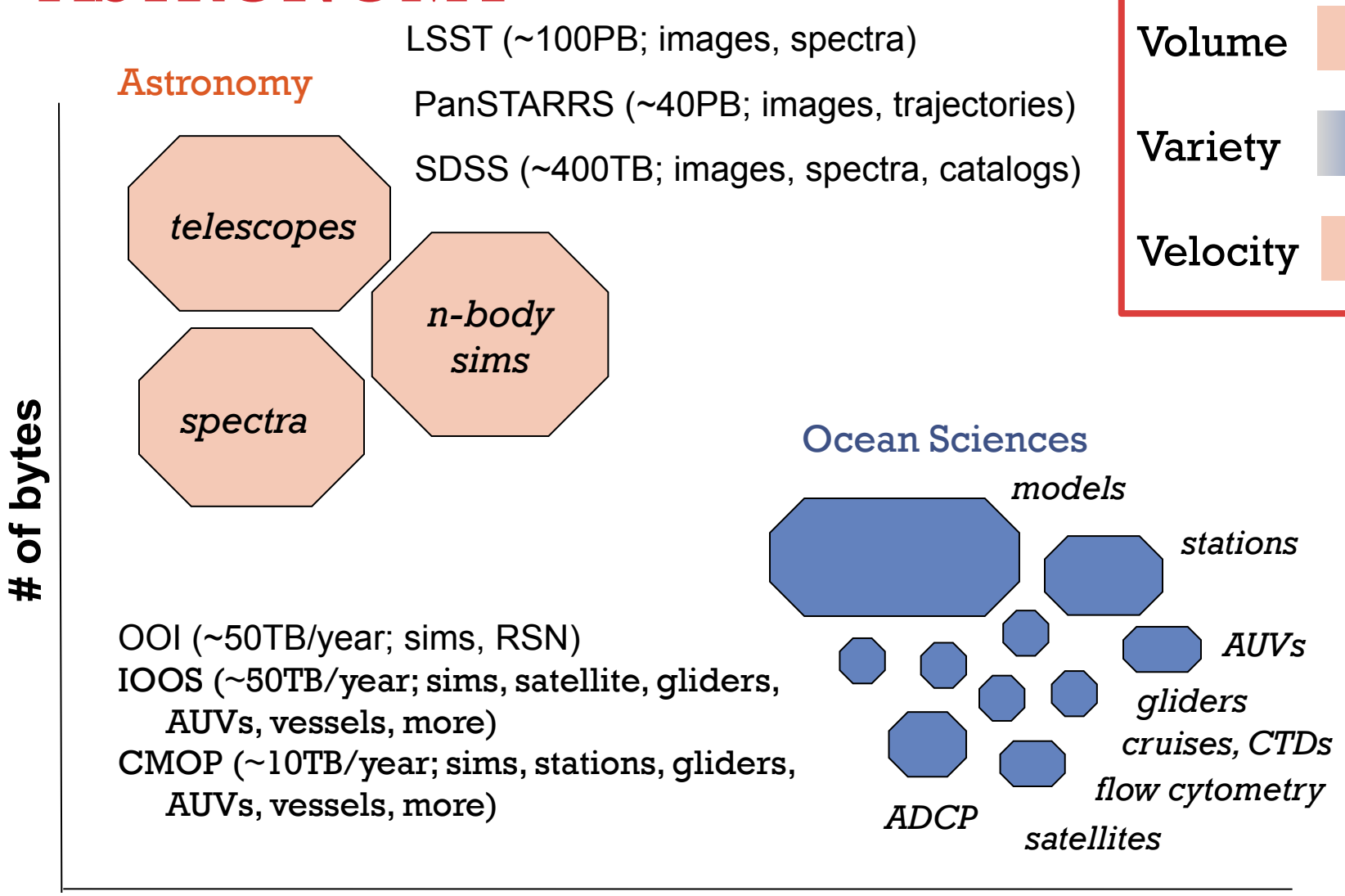
Oh!, and 1PB ~5,000 disks

At some point you need
indices to limit search
parallel data search and analysis

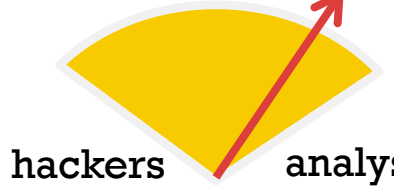
This is where databases can help

[slide src: Jim Gray]

ASTRONOMY


3 V's of Big Data
Volume
Variety
Velocity
Astronomy
LSST (~100PB; images, spectra)
PanSTARRS (~40PB; images, trajectories)
SDSS (~400TB; images, spectra, catalogs)
telescopes
n-body sims
spectra
Ocean Sciences
models
stations
AUVs
gliders
cruises, CTDs
flow cytometry
satellites
ADCP
OOI (~50TB/year; sims, RSN)
IOOS (~50TB/year; sims, satellite, gliders, AUVs, vessels, more)
CMOP (~10TB/year; sims, stations, gliders, AUVs, vessels, more)
# of bytes
# of data sources

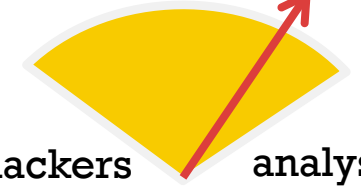
SHORTAGE OF ANALYSTS



US faces shortage of 140,000 to 190,000 people "with deep analytical skills, as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

--Mckinsey Global Institute

BIOLOGY



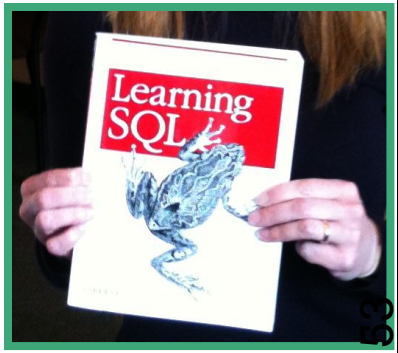
Biologists are beginning to write very complex queries (rather than relying on staff programmers)

Example: Computing the overlaps of two sets of blast results

```
SELECT x.strain, x.chr, x.region as snp_region, x.start_bp as snp_start_bp
    , x.end_bp as snp_end_bp, w.start_bp as nc_start_bp, w.end_bp as nc_end_bp
    , w.category as nc_category
    , CASE WHEN (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
  THEN x.end_bp - x.start_bp + 1
  WHEN (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
  THEN x.end_bp - w.start_bp + 1
  WHEN (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
  THEN w.end_bp - x.start_bp + 1
END  AS len_overlap

FROM [koesterj@washington.edu].[hotspots_deserts.tab] x
INNER JOIN [koesterj@washington.edu].[table_noncoding_positions.tab] w
ON x.chr = w.chr
WHERE (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
OR (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
OR (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
ORDER BY x.strain, x.chr ASC, x.start_bp ASC
```

We see thousands of queries written by non-programmers



DATA SCIENCE IN SCIENCE

BILL HOWE, PHD

DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS

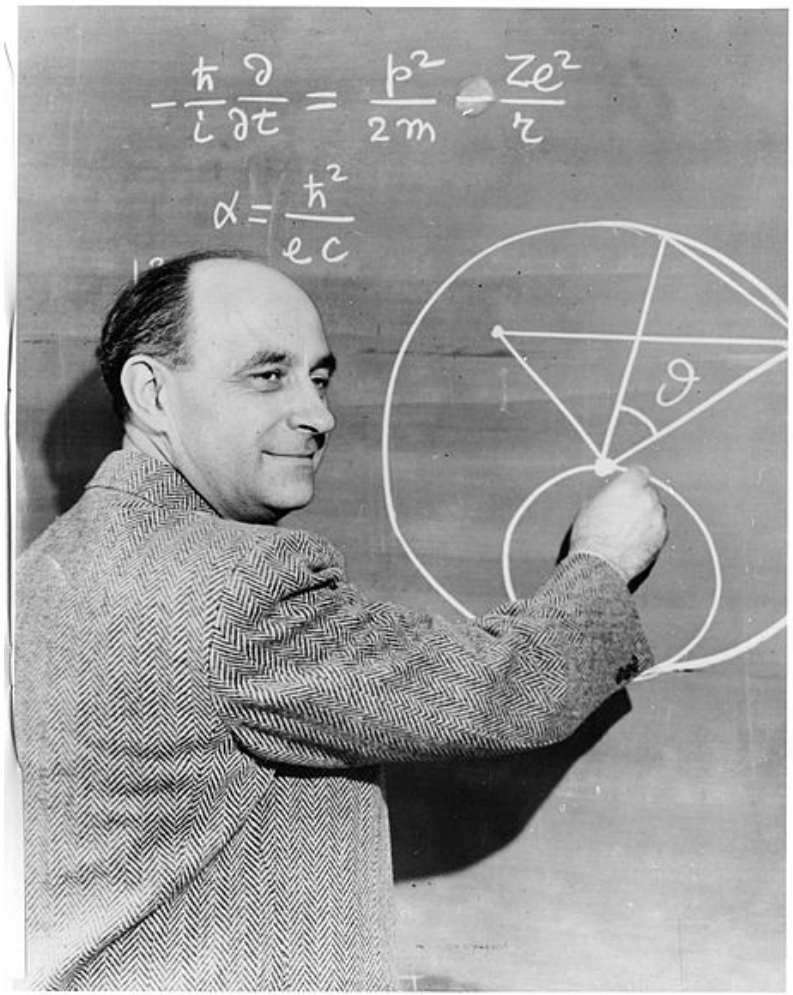
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE

“ESCIENCE” = “DATA SCIENCE”

EMPIRICAL THEORETICAL COMPUTATIONAL

public domain

EMPIRICAL THEORETICAL COMPUTATIONAL



EMPIRICAL THEORETICAL **COMPUTATIONAL**

EMPIRICAL THEORETICAL COMPUTATIONAL ESCIENCE

SLOAN DIGITAL SKY SURVEY

SCIENCE IS ABOUT ASKING QUESTIONS

Traditionally: "Query the world"

Data acquisition activities coupled to a specific hypothesis

eScience: "Download the world"

Data acquired en masse in support of many hypotheses

The cost of data acquisition has dropped precipitously thanks to advances in technology

- **Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)**
- **Life Sciences: lab automation, high-throughput sequencing,**
- **Oceanography: high-resolution models, cheap sensors, satellites**

The cost of finding, integrating, and analyzing data, then communicating results, is the new bottleneck

ESCIENCE IS DRIVEN BY *DATA* MORE THAN BY COMPUTATION

Massive volumes of data from sensors and networks of sensors

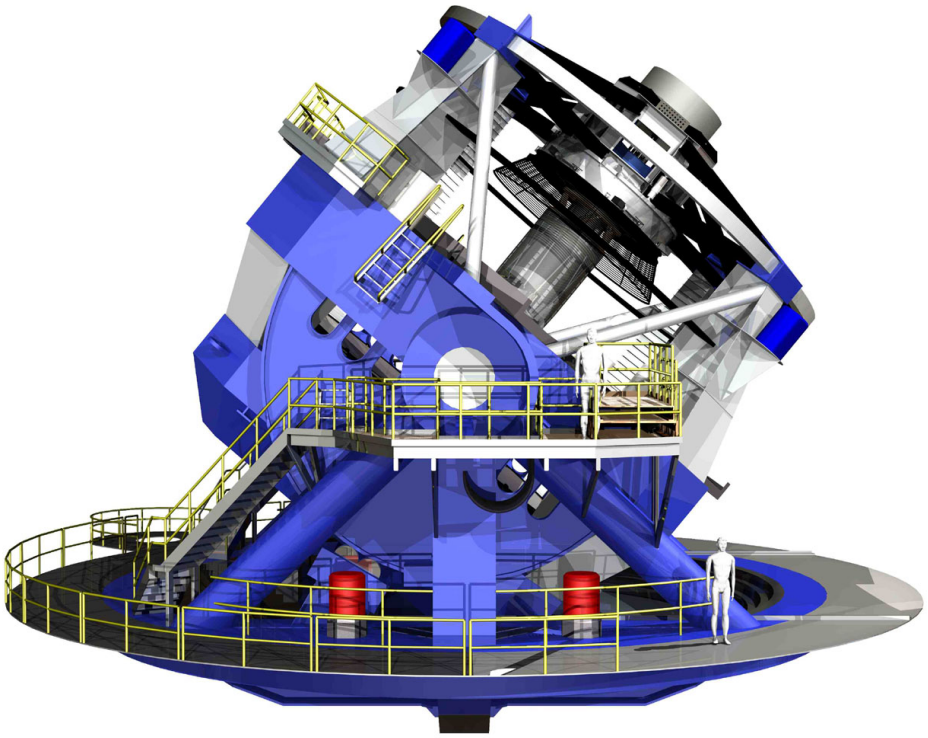
Apache Point telescope, SDSS

80TB of raw image data (80,000,000,000,000 bytes) over a 7 year period

Large Synoptic Survey Telescope (LSST)

40TB/day (an SDSS every two days), 100+PB in its 10-year lifetime

400mbps sustained data rate between Chile and NCSA



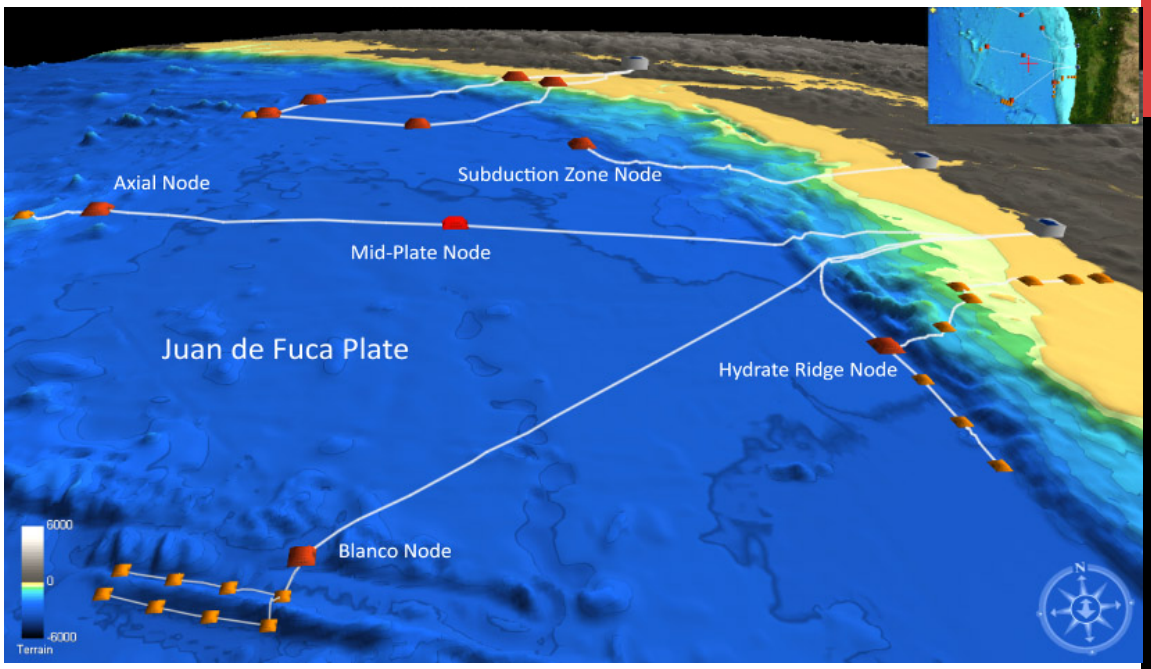
Illumina
HiSeq 2000 Sequencer

~1TB/day

Major labs have 25-100 of these machines

Regional Scale Nodes of the NSF Ocean Observatories Initiative

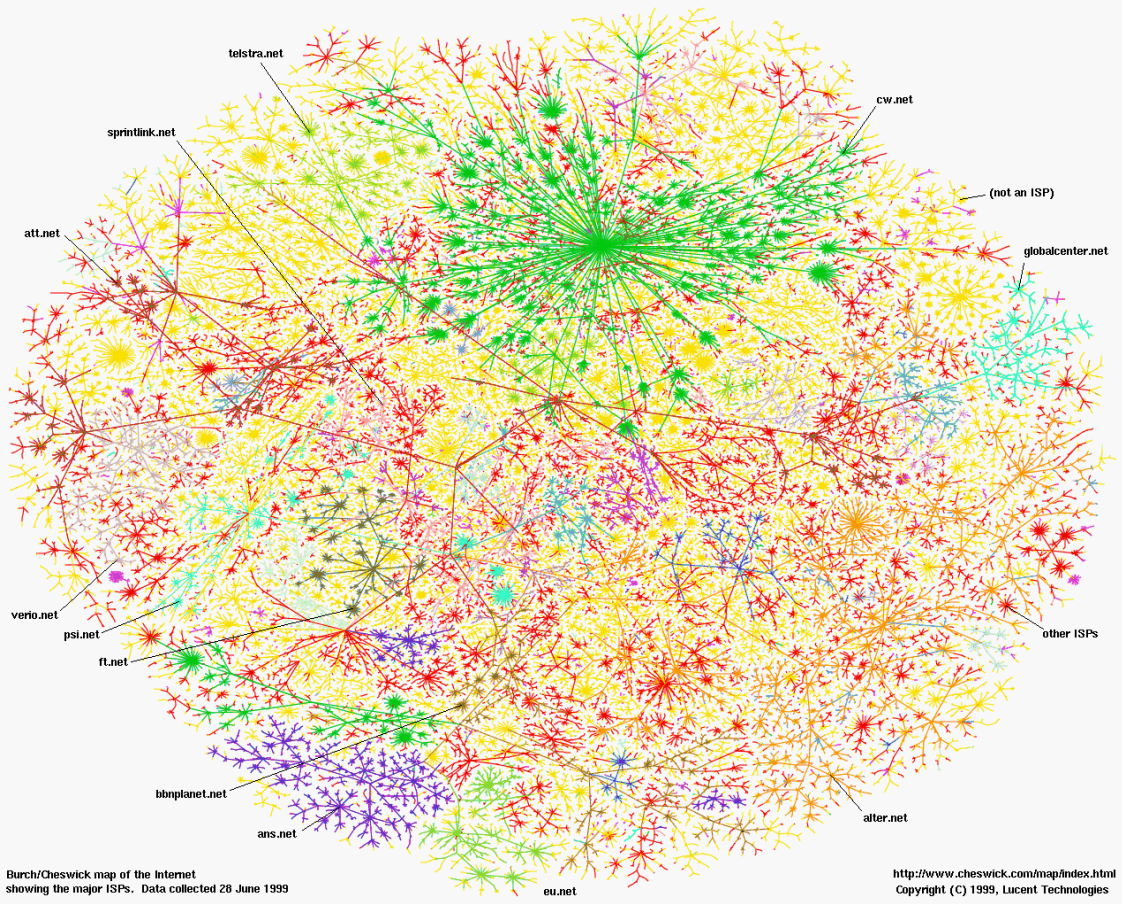
1000 km of fiber optic cable on the seafloor, connecting thousands of chemical, physical, and biological sensors



The Web

20+ billion web pages x 20KB = 400+TB

One computer can read 30-35 MB/sec from one disk => 4 months just to read the web



ESCIENCE IS ABOUT THE *ANALYSIS* OF DATA

The automated or semi-automated extraction of knowledge from massive volumes of data

- There's simply too much of it to look at
- But it's not just a matter of volume

The Three V's of Big Data:

- Volume: number of rows / objects / bytes
- Variety: number of columns / dimensions / sources
- Velocity: number of rows / bytes per unit time

More V's:

- *Veracity: Can we trust this data?*

SUMMARY

Science is in the midst of a generational shift from a data-poor enterprise to a data-rich enterprise

Data analysis has replaced data acquisition as the new bottleneck to discovery

What does this have to do with business?

Business is beginning to look a lot like science

- Acquire data aggressively and keep it around
- Hire data scientists
- Make empirical decisions

ASIDE ON "BIG DATA"

BILL HOWE, PHD
DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE

BIG DATA: THREE CHALLENGES

Volume

- the size of the data

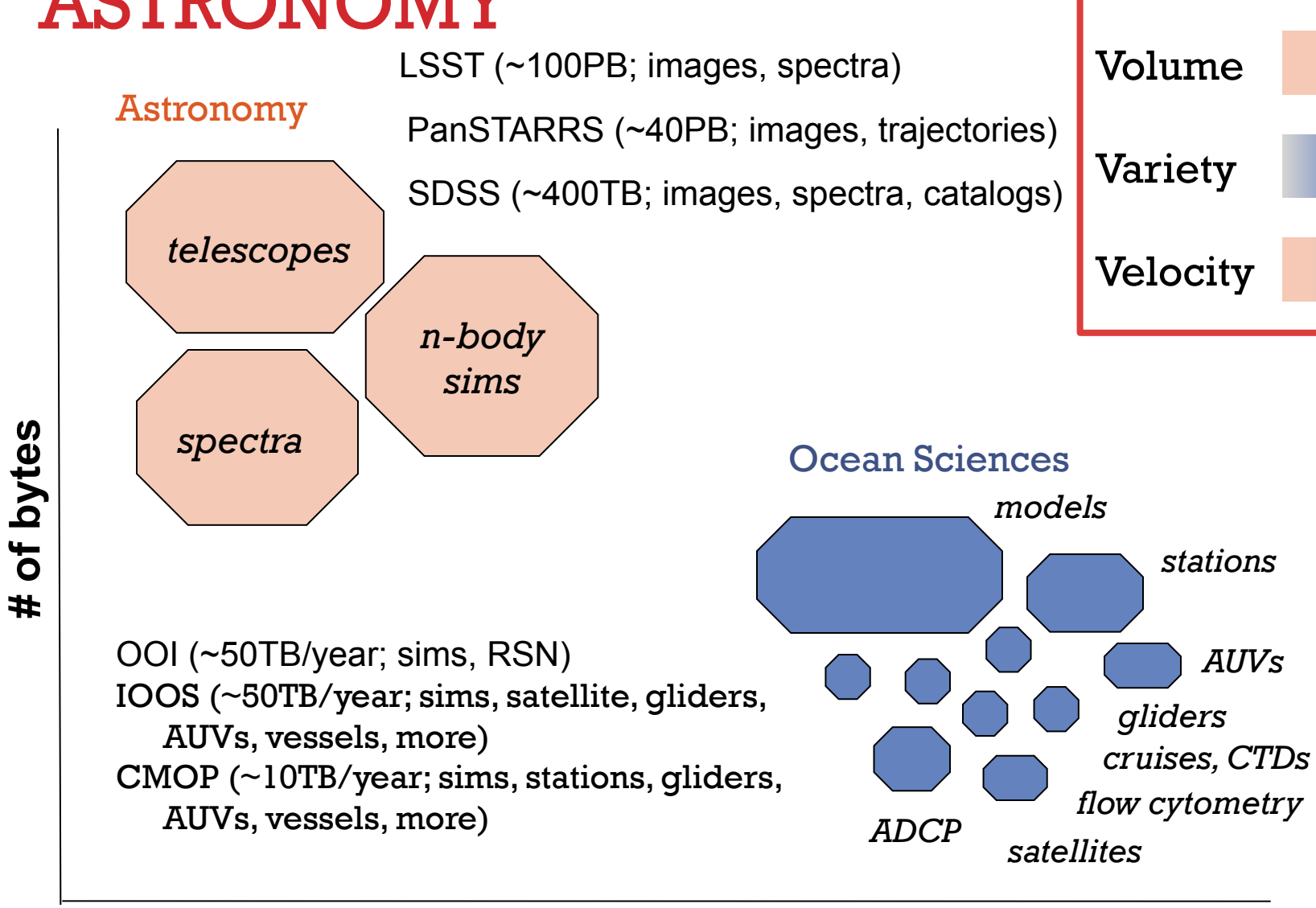
Velocity

- the latency of data processing relative to the growing demand for interactivity

Variety

- the diversity of sources, formats, quality, structures

ASTRONOMY


3 V's of Big Data
Volume
Variety
Velocity
LSST (~100PB; images, spectra)
Astronomy
PanSTARRS (~40PB; images, trajectories)
SDSS (~400TB; images, spectra, catalogs)
telescopes
n-body sims
spectra
Ocean Sciences
models
stations
AUVs
gliders
cruises, CTDs
flow cytometry
ADCP
satellites
OOI (~50TB/year; sims, RSN)
IOOS (~50TB/year; sims, satellite, gliders, AUVs, vessels, more)
CMOP (~10TB/year; sims, stations, gliders, AUVs, vessels, more)
# of bytes
# of data sources

BIG DATA

"Big Data is any data that is expensive to manage and hard to extract value from."

Michael Franklin
Thomas M. Siebel Professor of Computer Science
Director of the Algorithms, Machines and People Lab
University of Berkeley

Key idea: "Big" is relative! "Difficult Data" is perhaps more apt!

HISTORY OF THE TERM "BIG DATA"

Erik Larson, 1989, Harper's magazine

"The keepers of big data say they do it for the consumer's benefit. But data have a way of being used for purposes other than originally intended."

Takeaway: private data is becoming commoditized

Predates the rise of the Internet, but foreshadows emerging topics in data science: ethics, validation, privacy

BIG DATA HISTORY

"E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety."

On Volume:
"The lower cost of e-channels enables and enterprise to offer its goods or services to more individuals or trading partners, and up to 10x the quantity of data about an individual transaction may be collected—thereby increasing the overall volume of data to be managed."

On Velocity:
"E-commerce has also increased point-of-interaction (POI) speed, and consequently the pace data used to support interactions and generated by interactions"

On Variety:
"Through 2003/04, no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics."

- Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", Gartner, 2001

HISTORY OF THE TERM "BIG DATA"

"Big Data … and the Next Wave of InfraStress"

-- John R. Mashey, former Chief Scientist, SGI

Takeaway: Disk capacities growing incredibly fast, disk latencies not keeping pace: trouble ahead!

A technology-oriented view of Big Data

BIG DATA NOW

"…the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences – arguably the key scientific theme of our times."

- Francis X. Diebold
  Paul F. and Warren S. Miller Professor of Economics
  School of Arts and Sciences
  University of Pennsylvania

WHERE DOES BIG DATA COME FROM?

“data exhaust” from customers

new and pervasive sensors

the ability to “keep everything”

EXAMPLE

Car black boxes: Privacy nightmare or a safety measure?

February 15, 2013 | By Ronald D. White

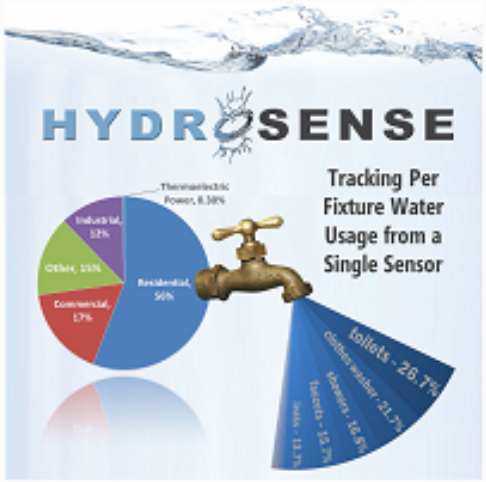
Email Share +1 1 Tweet 0 Recommend 271

What if the black box in your new car becomes a tool to invade your privacy? What if, on the other hand, it winds up saving your life after an accident?

Those are some of the questions being raised this week over black box data event recorders in cars. Privacy advocates worried on Thursday that the data could be misused. Safety advocates argued on Friday that a watered-down version of the recorders would slow safety innovations.

photo in public domain

EXAMPLE

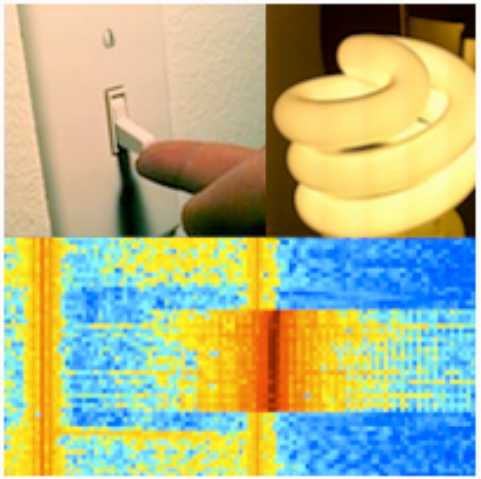


HydroSense

Water Fixture Usage with a Single Sensor

HydroSense is a pressure-based sensor that automatically determines water usage activity and flow down to the source (e.g., dishwasher, laundry, shower) from a single non-intrusive installation point.

Lead Researchers: Jon Froehlich, Eric Larson, Shwetak Patel



ElectriSense

Electrical Device Energy Usage with a Single Sensor

ElectriSense is a single plug-in sensor that provides whole home device level usage data. That is, using a single sensor plugged in anywhere in the home, ElectriSense can infer which electrical appliances are on and which off. This data could be used for numerous applications, for example, for providing home owners with itemized electrical bill that not only shows the total energy consumption but breaks the total on a per appliance basis (TV consumed 20 KWh, Lighting consumes 18 KWh and so on).

Lead Researchers: Sidhant Gupta, Shwetak Patel

INTRODUCTION TO DATA SCIENCE: LOGISTICS

BILL HOWE, PHD
DIRECTOR OF RESEARCH, SCALABLE DATA ANALYTICS
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE

HOW THIS COURSE IS ORGANIZED

- **A "guided tour" of important trends and technologies**
- **A "deep dive" into selected must-know algorithms, techniques, and technologies**
- **A set of hands-on assignments to deliver specific skills and experiences**

PREQUISITES

We assume

- some prior programming experience in some language
- "muscle memory" with basic college statistics
- some exposure to databases and database concepts

One assignment will require writing SQL

Two assignments will require writing Python

One (optional) assignment will involve processing ~1TB of data using Amazon Web Services

- You will pay for these resources, should you choose to complete the assignement

One assignment will involve solving a prediction problem on kaggle.com using whatever tools you wish.

Some understanding of distributed systems will be helpful, but not required

LEARNING OBJECTIVES

- **The ability to describe the landscape of data science concepts, tools, algorithms, and technologies**
- **Hands-on experience in data manipulation, analysis and prediction**
- **You will be an "advanced beginner" in a variety of data science topics**

COURSE PHILOSOPHY

The skills needed by a data scientist span a variety of different areas

- statistics, programming, databases, systems, visualization

The traditional organization of topics is not ideal

- It is difficult to acquire introductory-level knowledge in all areas
- Cross-cutting concepts and abstractions are obscured

Our goal: Expose and simplify the underlying commonalities between these areas

NON-GOALS FOR THIS COURSE

You will not emerge an expert in statistics

- Though you will apply basic statistical methods

You will not emerge an expert in machine learning

- Though you will be familiar with some important concepts and will have the chance to exercise them

You will not emerge an expert in databases and NoSQL

- Though you will understand the concepts they share and know how to apply them

You will not emerge an expert in R, Python, MapReduce, or SQL

- Though you will use all of these in assignments

SLIDES CAN BE FOUND AT:

TEACHINGDATASCIENCE.ORG