6.S079 SOFTWARE SYSTEMS FOR DATA SCIENCE

ADMINISTRIVIA

Lecturers:

Tim Kraska

Sam Madden

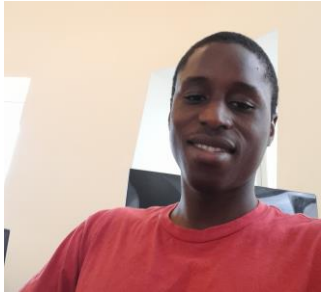
TAs:

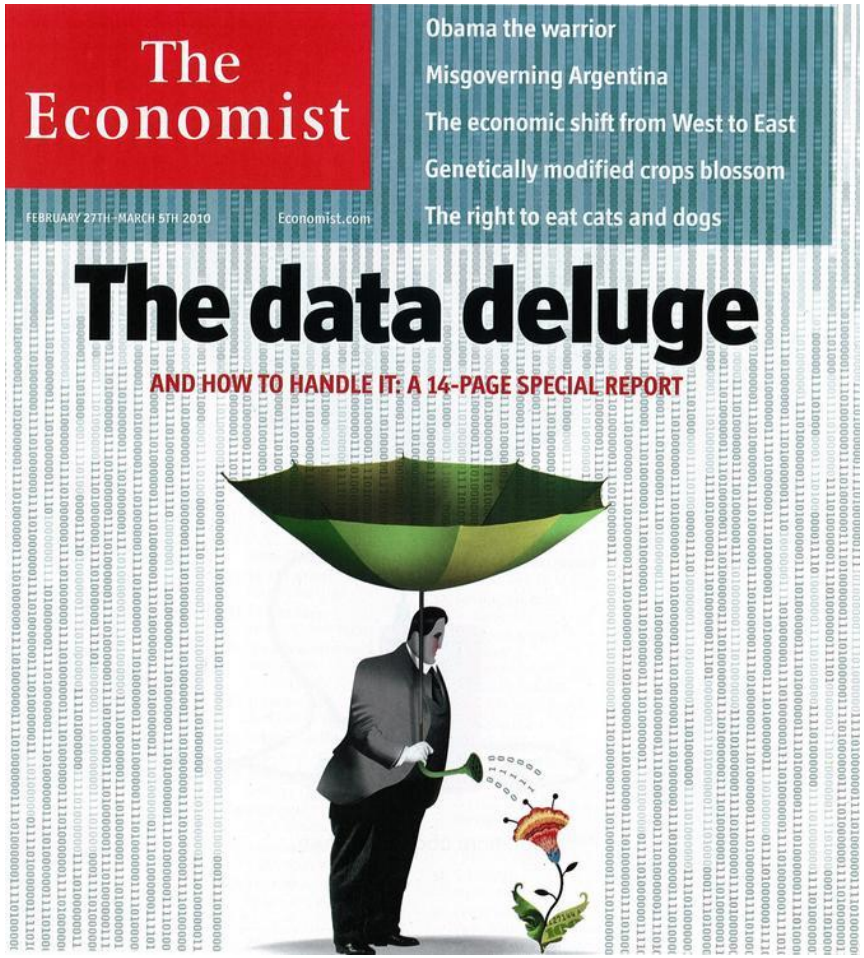
Markos Markakis

Amadou Ngom

Website: http://dsg.csail.mit.edu/6.S079/

Piazza: http://piazza.com/mit/fall2022/6s079




The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
The End of Science
The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.

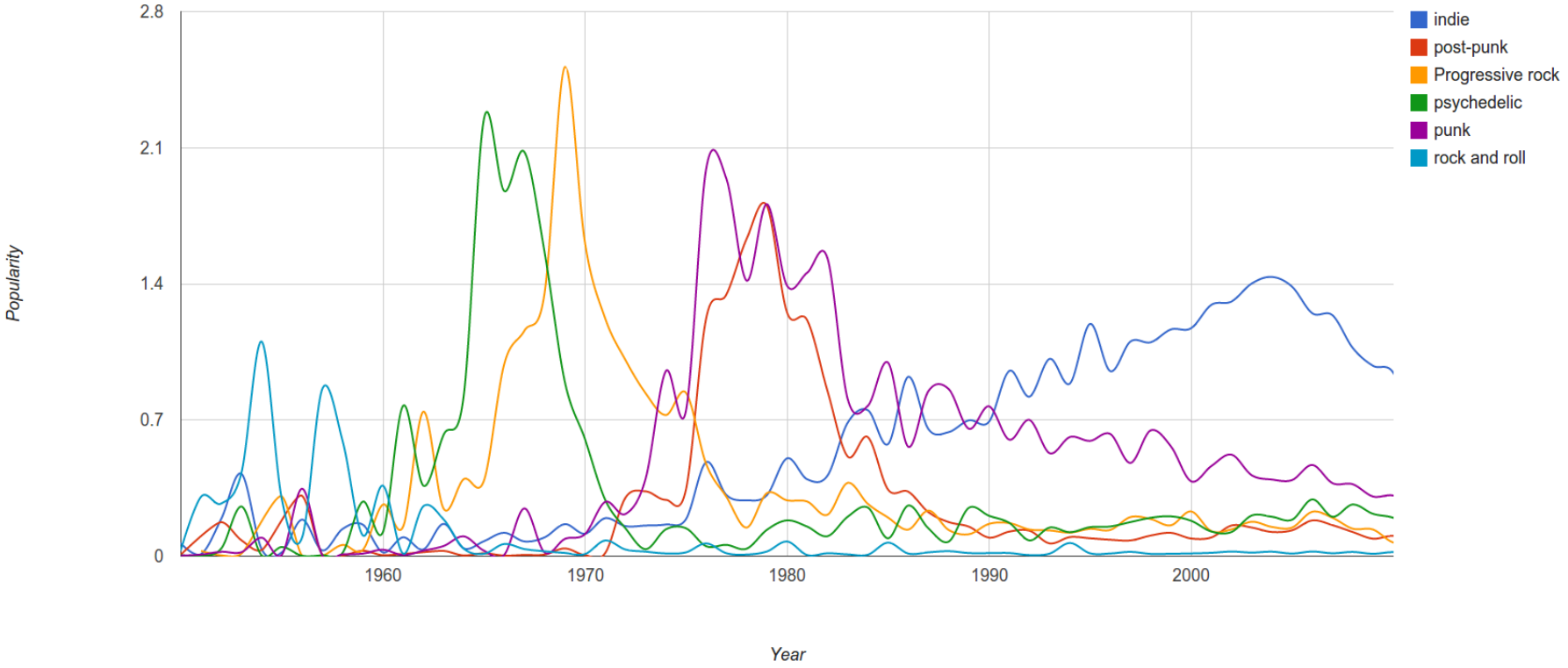
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google


NETFLIX
MENU
tv


Google

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

HOW WOULD YOU ESTIMATE THE EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY?

(read: before Twitter)

EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY

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"

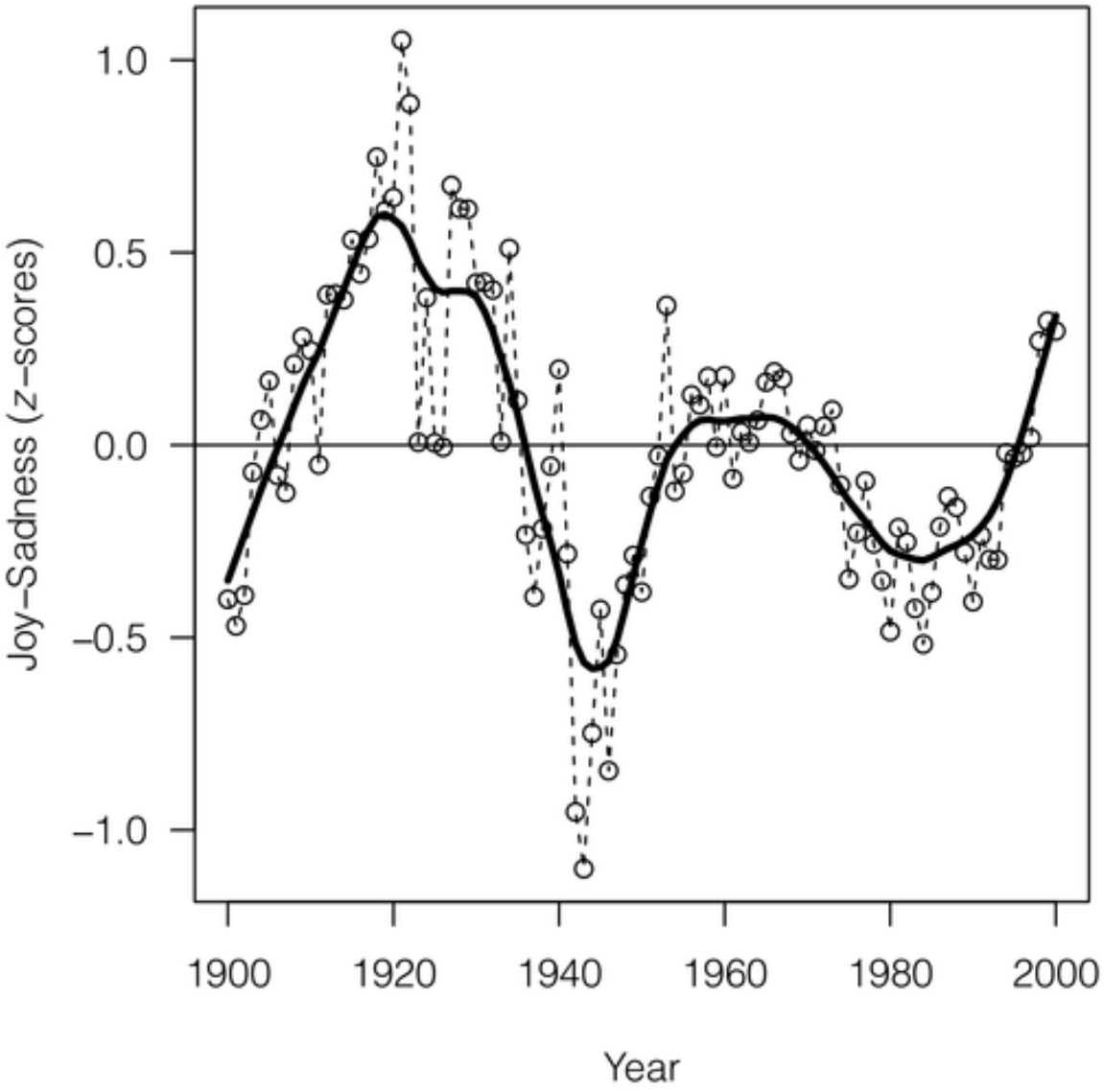
2) Label each 1-gram (word) with a mood score.
(Thanks, WordNet Affect)

3) Count the occurrences of each mood word

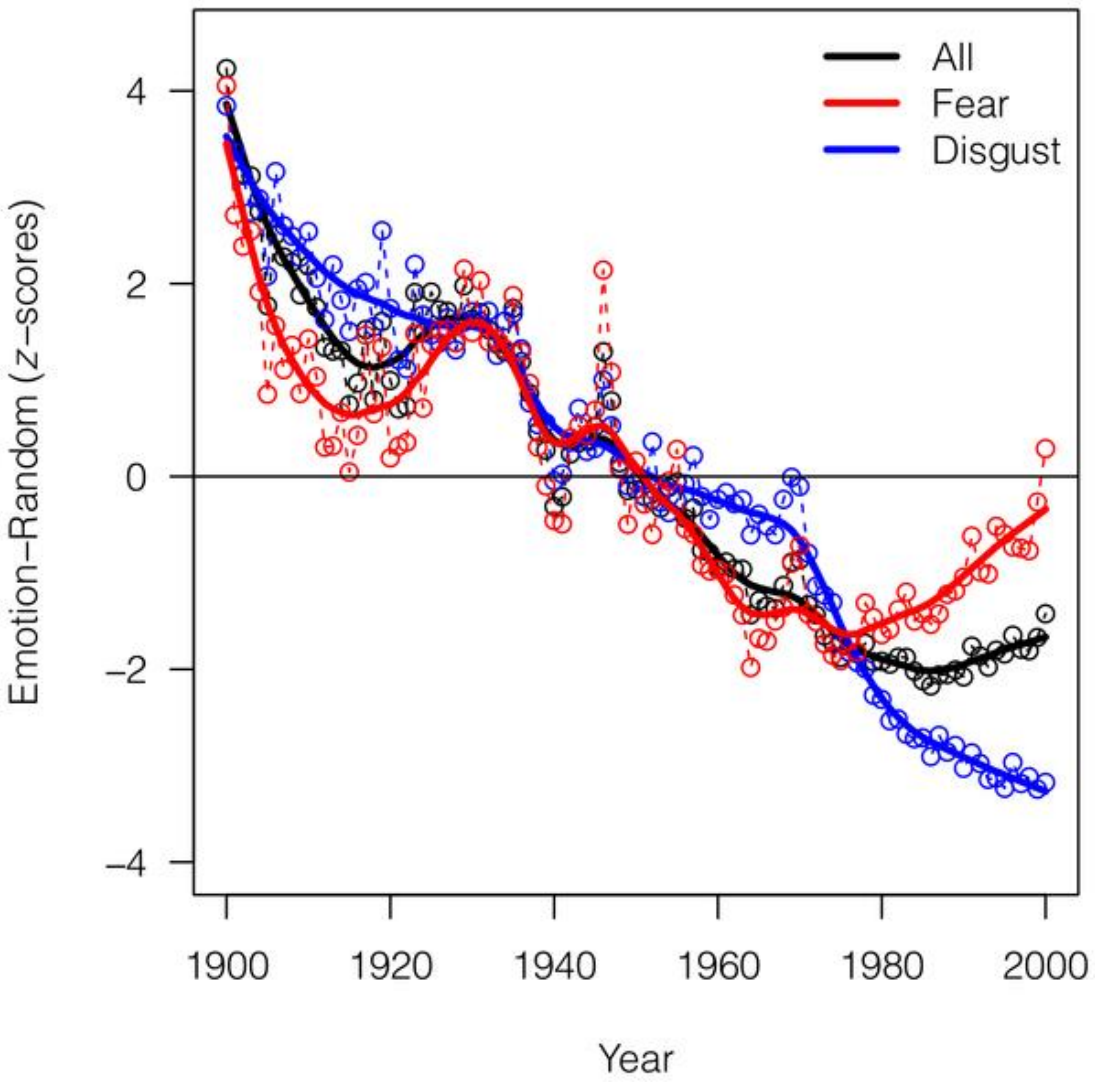
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

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

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

PAPERS CITED BY THEM

…

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

…

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

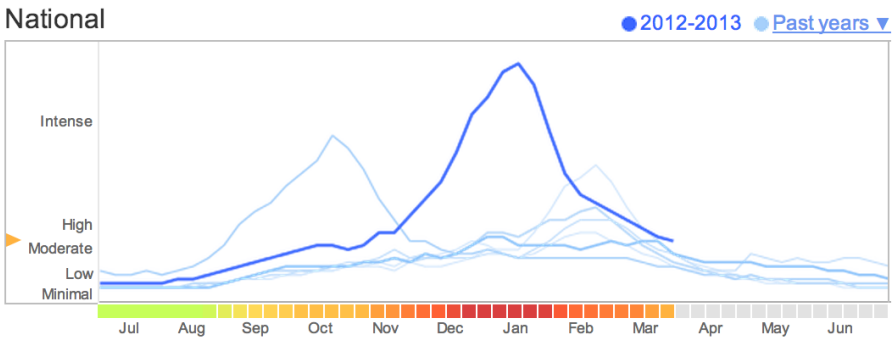
Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

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

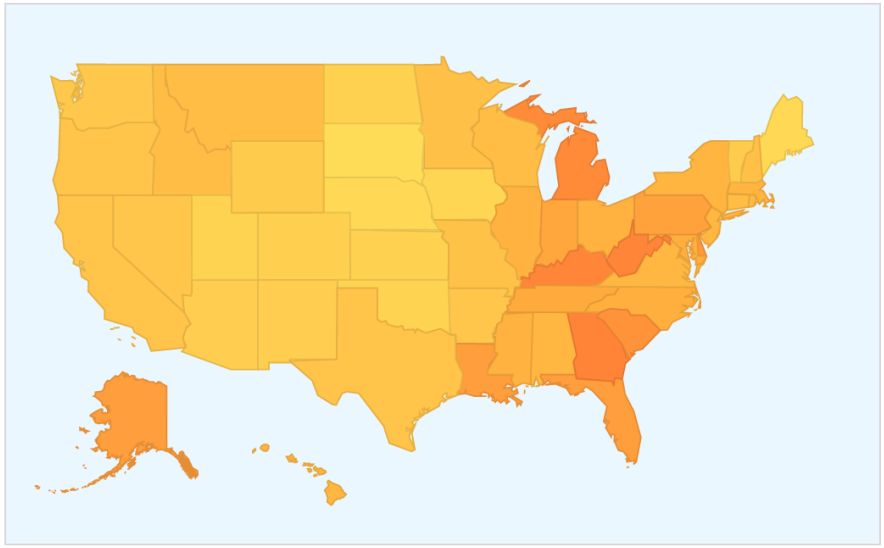
HOW WOULD YOU BUILD A FLU PREDICTION MODEL?

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.

flu risk

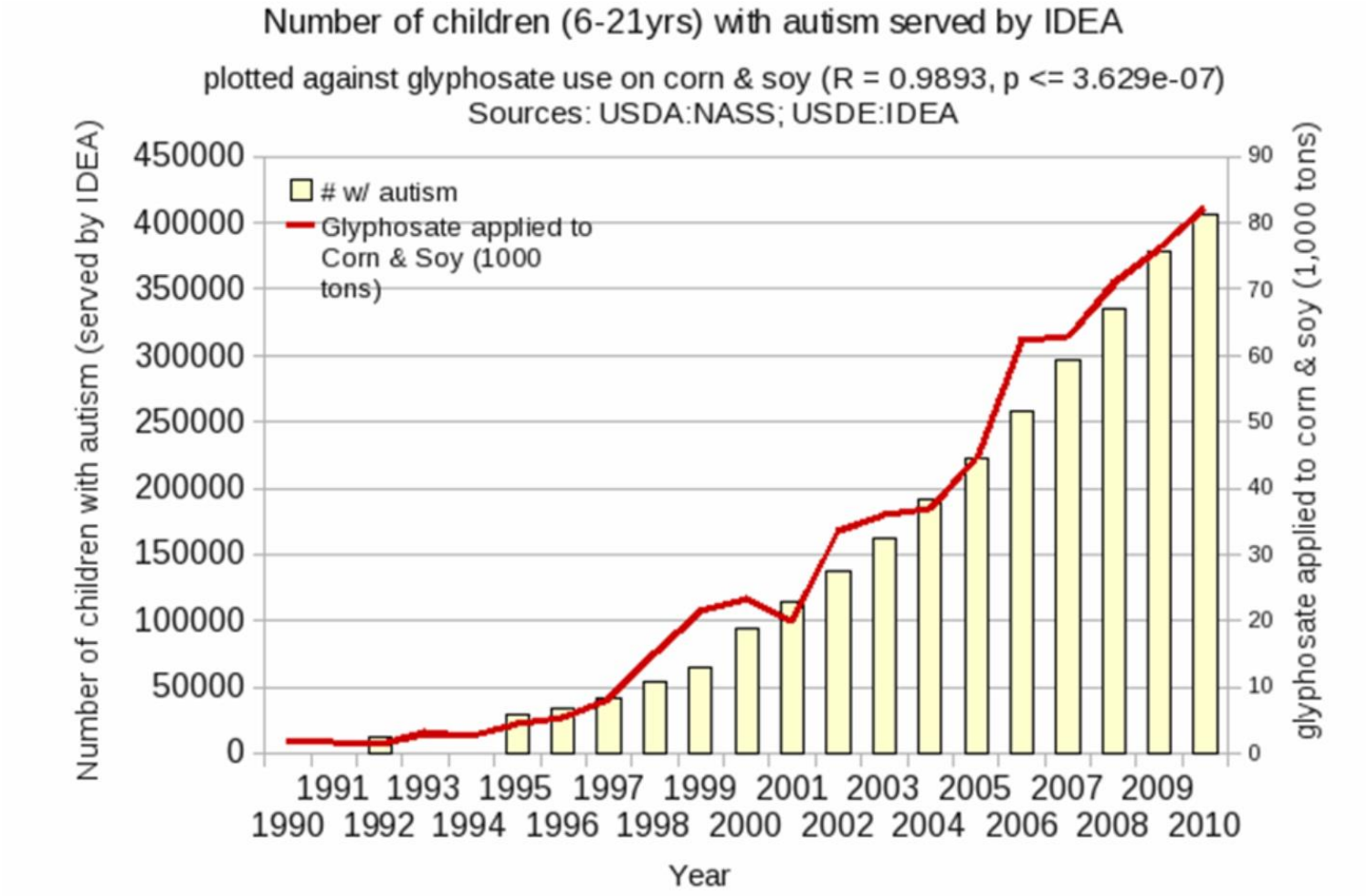
"Scientific hindsight shows that Google Flu Trends far overstated this year's flu season…."

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

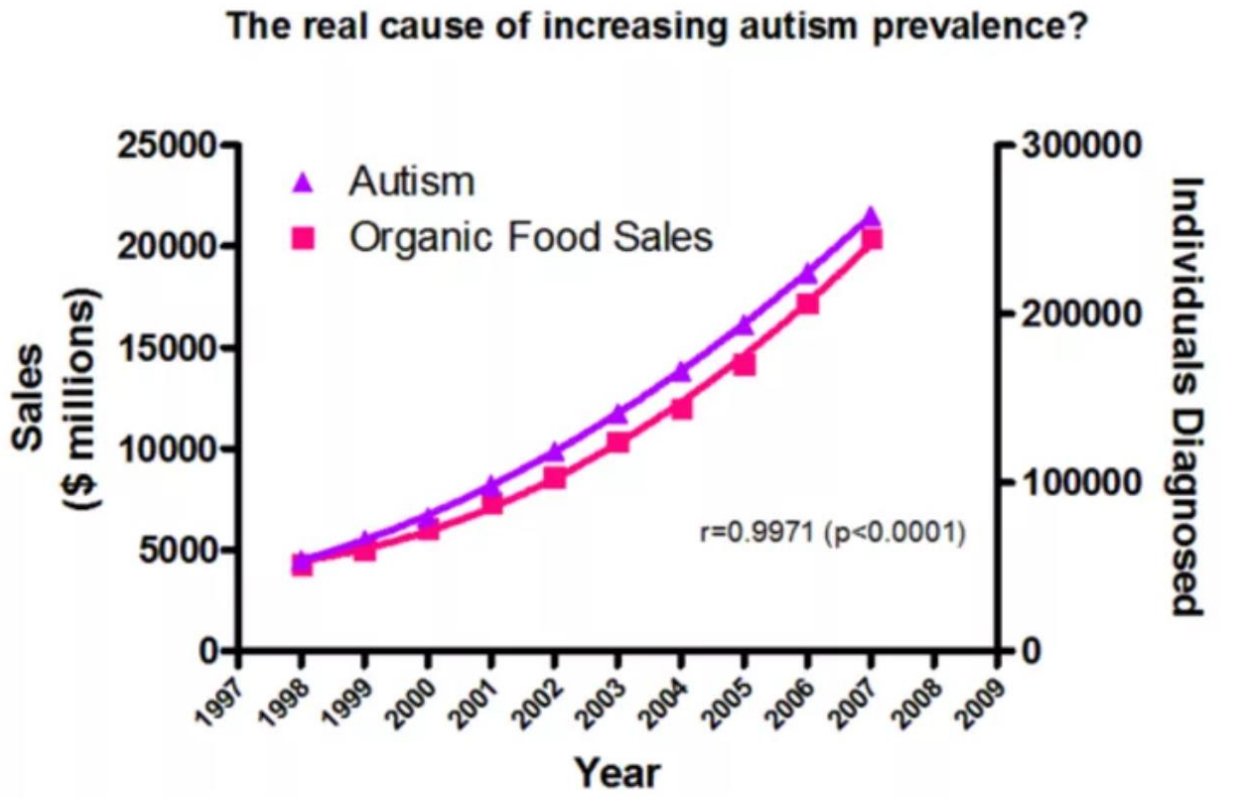
David Wagner, Atlantic Wire, Feb 13 2013

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

BUT DATA CAN BE MISLEADING, AND ANALYSIS IS HARD



BUT DATA CAN BE MISLEADING, AND ANALYSIS IS HARD (SOURCE)



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

BUT DATA IS EASY TO ABUSE OR MISINTERPRET

Table 4. Vaccination status for individuals ≥12 years infected with Omicron compared to other variants, data included in the table are from 22 November to 16 December 2021

Tabel 4. Vaccinationsstatus for personer ≥12 år med omikron-infektion sammenlignet med andre varianter i perioden fra og med 22. november 2021 til og med 16. december 2021

Vaccination status (12+ year olds)	Other variants (No. of cases)	Other variants (%)	Omicron (No. of cases)	Omicron (%)
Booster vaccinated	8,866	8.6	1,851	10.8
Fully vaccinated	67,034	65.3	13,546	79.0
Not vaccinated	23,492	22.9	1,454	8.5
Received first dose	3,216	3.1	304	1.8
Total	102,608	99.9	17,155	100.0

Individuals aged 5-11 years have recently been invited for COVID-19 vaccination, hence the vaccination coverage is relatively low in this age group and not included in Table 4.

William Makis MD
@MakisMD

Stunning numbers from Denmark:

56% double vaccinated are catching 65% of "other variants" & 79% of Omicron

Most vulnerable group to Omicron BY FAR

25% boosted still catch 10% of Omicron cases, while unvaccinated catch 8.5%

This is worse than vaccine failure. This is damage.

1:08 AM · Dec 20, 2021 · Twitter Web App

805 Retweets **136** Quote Tweets

1,146 Likes

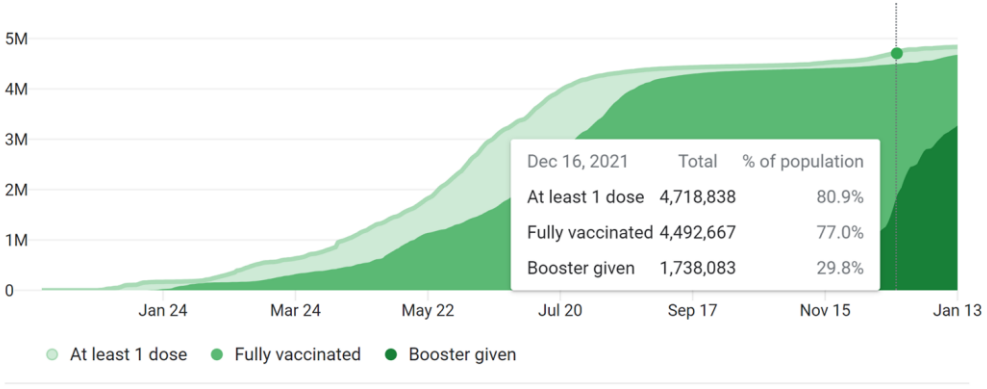
BUT DATA IS EASY TO ABUSE OR MISINTERPRET

Table 4. Vaccination status for individuals ≥12 years infected with Omicron compared to other variants, data included in the table are from 22 November to 16 December 2021

Tabel 4. Vaccinationsstatus for personer ≥12 år med omikron-infektion sammenlignet med andre varianter i perioden fra og med 22. november 2021 til og med 16. december 2021

Vaccination status (12+ year olds)	Other variants (No. of cases)	Other variants (%)	Omicron (No. of cases)	Omicron (%)
Booster vaccinated	8,866	8.6	1,851	10.8
Fully vaccinated	67,034	65.3	13,546	79.0
Not vaccinated	23,492	22.9	1,454	8.5
Received first dose	3,216	3.1	304	1.8
Total	102,608	99.9	17,155	100.0

Individuals aged 5-11 years have recently been invited for COVID-19 vaccination, hence the vaccination coverage is relatively low in this age group and not included in Table 4.



William Makis MD
@MakisMD

Stunning numbers from Denmark:

56% double vaccinated are catching 65% of "other variants" & 79% of Omicron

Most vulnerable group to Omicron BY FAR

25% boosted still catch 10% of Omicron cases, while unvaccinated catch 8.5%

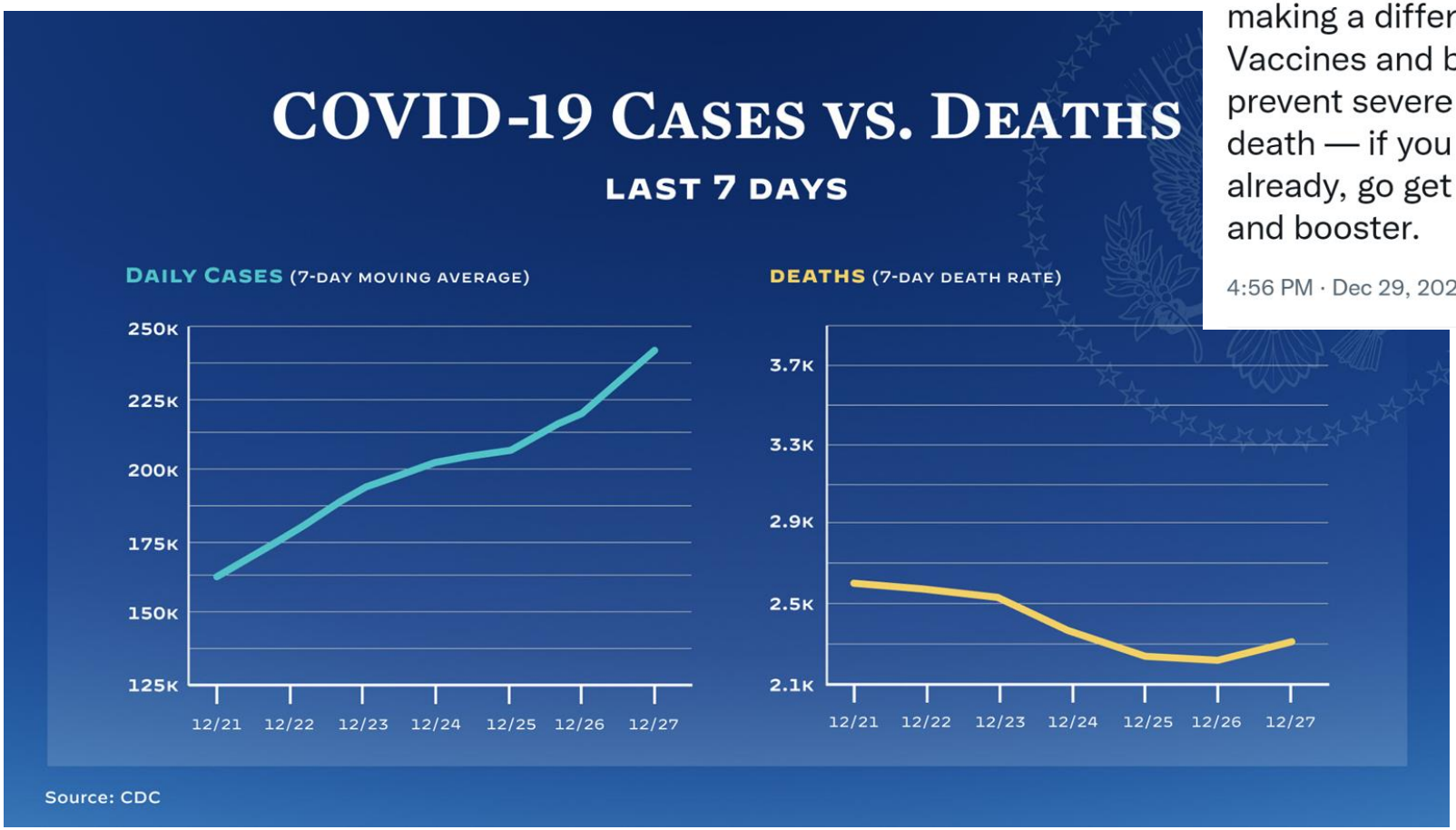
This is worse than vaccine failure. This is damage.

1:08 AM · Dec 20, 2021 · Twitter Web App

805 Retweets **136** Quote Tweets

1,146 Likes

BUT DATA IS EASY TO ABUSE OR MISINTERPRET



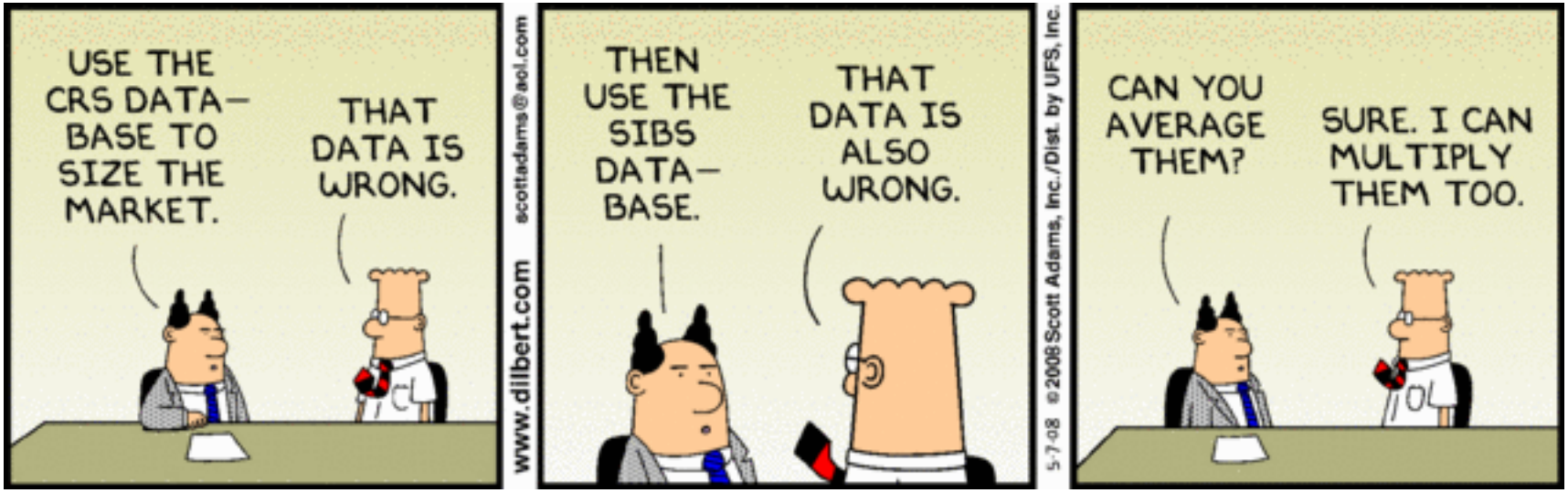
The White House
@WhiteHouse

Omicron cases are on the rise, but it's clear that vaccines and boosters are making a difference. Vaccines and boosters help prevent severe illness and death — if you haven't already, go get your vaccine and booster.

4:56 PM · Dec 29, 2021 · The White House

Any thoughts on what is wrong with this?

Data-source


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
www.dilbert.com scottadams@aol.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5-7-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

SIMPLE TRUTHS

"Power of data"

- the more data the merrier (GB -> TB)
- data comes from everywhere in all shapes
- value of data often discovered later

Services turn data into $

- the more services the merrier
- need to adapt quickly

E.g.: Google, Amadeus, Disney, Walmart, BMW, ...

Tools: Oracle/Vertica/PostgreSQL, Tableau, Python, Spark, SciKitLearn,....

BIG DATA QUESTION: YES OR NO?

Cure for cancer?

Find a spouse?

How to treat a cough?

Should I give Tim a loan?

Premium for fire insurance?

Which book should I read next?

Translate from English to German.

SOME RECURRING THEMES

simple methods

repurposing data

communication matters

Other themes

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

WHAT IS DATA SCIENCE?

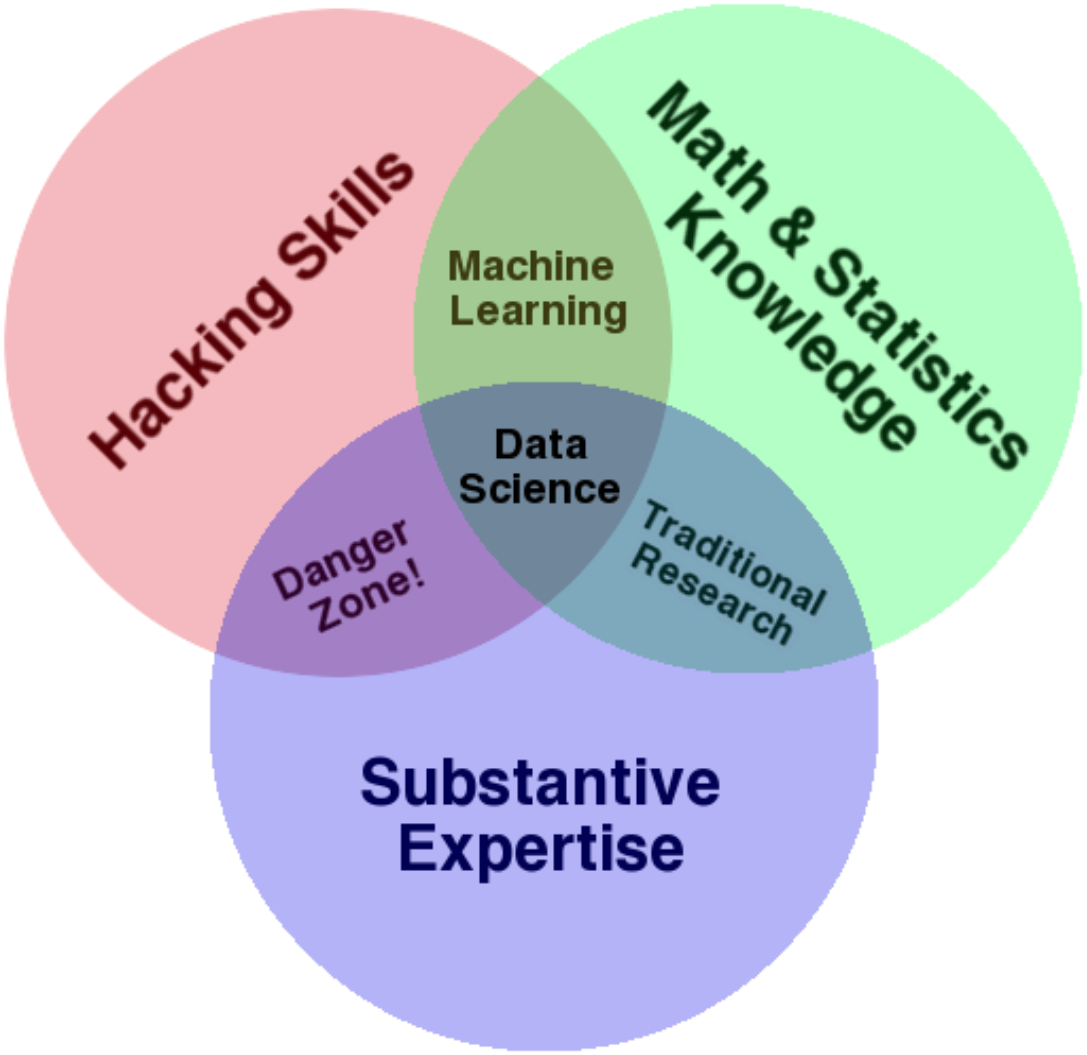
"Data Scientists:
The Definition of Sexy"

Forbes, 2012

"Data Scientist: The Sexiest
Job of the 21st Century"

Harvard Business Review, 2012

DREW CONWAY'S DATA SCIENCE VENN DIAGRAM



MIKE DRISCOLL'S THREE SEXY SKILLS OF DATA GEEKS

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

Data Wrangling

- parsing, scraping, and formatting data

Statistics

- traditional analysis

Visualization

- graphs, tools, etc.

DOING DATA SCIENCE

PETER HUBER:

1. **Inspection**
2. **Error checking**
3. **Modification**
4. **Comparison**
5. **Modeling and model fitting**
6. **Simulation**
7. **What-if analyses**
8. **Interpretation**
9. **Presentation of conclusions**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

DOING DATA SCIENCE

BEN FRY:

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

COLIN MALLOWS:

1. Identify data to collect and its relevance to your problem
2. Statistical specification of the problem
3. Method selection
4. Analysis of method
5. Interpret results for non-statisticians

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

A PRACTICAL DEFINITION

Data Science is about the whole processing pipeline to extract information out of data

Data Scientist understand and care about the whole data pipeline and produce data products

A data pipeline consists of 3 steps:

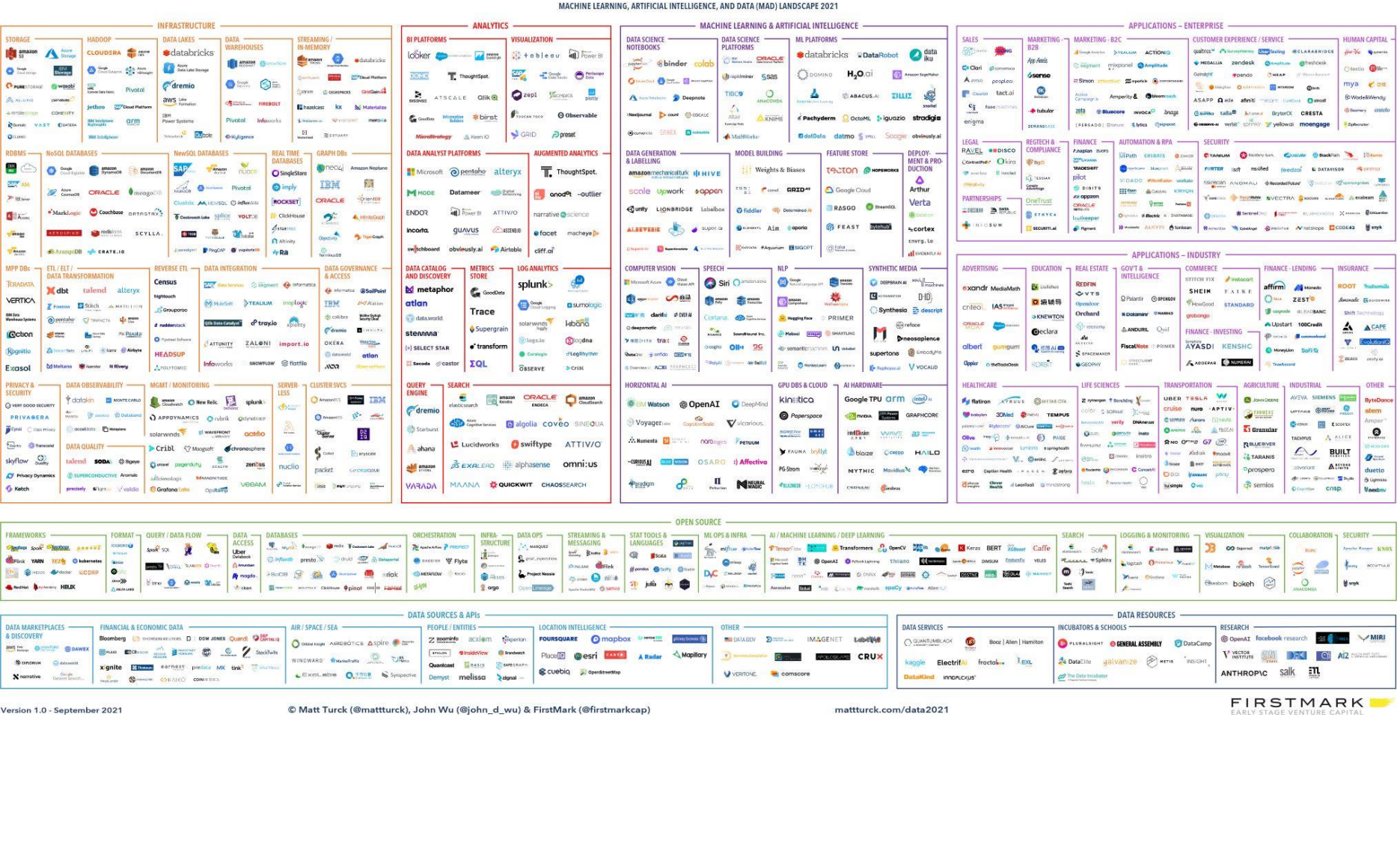
1) **Preparing to run a model**

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

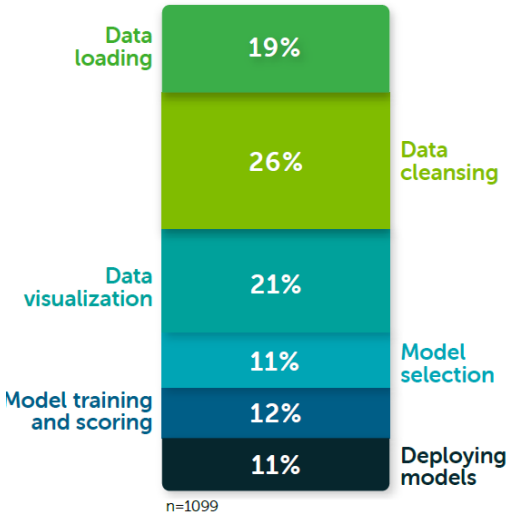
2) **Running the model**

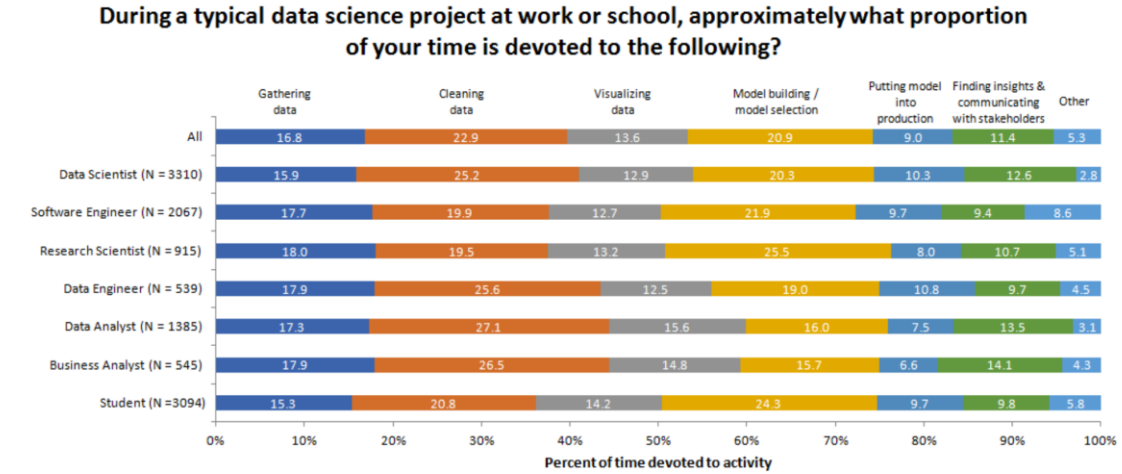
3) **Communicating the results / Create data products**

WHAT TOOLS ARE INVOLVED



WHERE DO DATA SCIENTIST SPEND THEIR TIME?



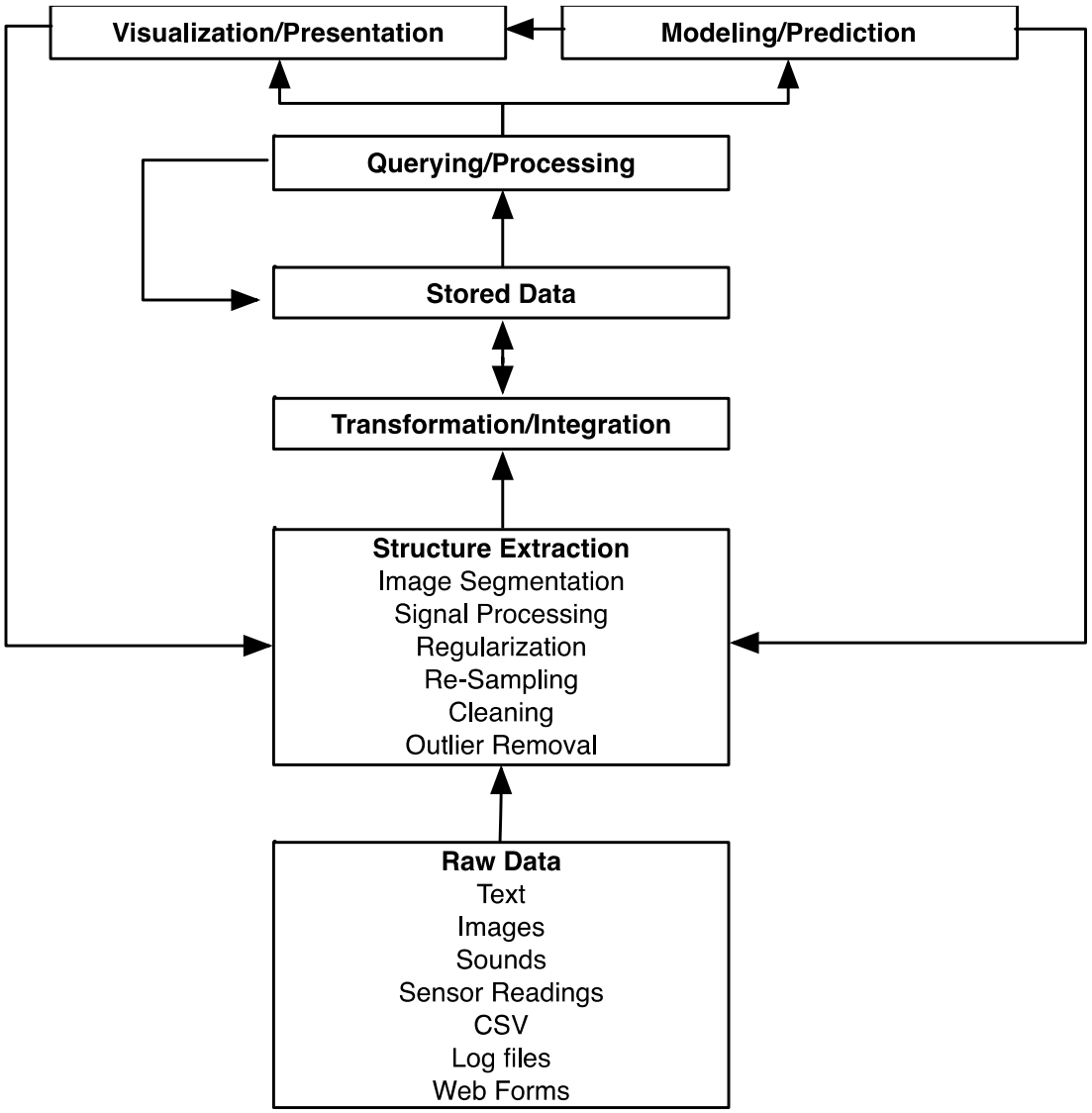


Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 15937 respondents who provided an answer to this question. Only selected job titles are presented.

Anaconda's annual survey.
https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

6.S079
WHAT IS THIS COURSE ABOUT?

DATA PROCESSING PIPELINE



Course will investigate all of these topics

Structured data
- SQL Databases / Pandas

Data Cleaning & Wrangling

Entity Resolution

ML & Embeddings
- Not a deep treatment

Visualization

Scaling & Performance

Cloud Data

COURSE STRUCTURE

2 lectures per week

Weekly readings

2 Quizzes (but no final exam)

Grading Rubric:

Final Project: 35%

- Proposal: 2.5%
- Check-in 1: 2.5%
- Check-in 2: 5%
- Poster & presentation: 10%
- Report: 15%

Labs: 35% (weighted equally)

Quizzes: 25%

- Quiz 1: 12.5%
- Quiz 2: 12.5%

Participation: 5% (Piazza, class, …)

LABS

Lab 0 – Setting up the environment

Lab 1 – Working with Data (SQL, Dataframes)

Lab 2 – Data Cleaning and Wrangling

Lab 3 – Using Machine Learning

Lab 4 – Embeddings

Lab 5 – Visualization

Lab 6 – Parallelism

PROJECT

Two options:

1. **Build a system for working with data**

2. **Choose a data set and do some end to end modeling on it**

EXAMPLE "SYSTEMS"

Given a twitter keyword, analyze the distribution of sentiment in tweets about it

Build a tool to extract structured data from a particular type of document, i.e., go from scanned PDF ➔ tabular data

Build a high performance visualization system for some data set

MAPD (MASSIVELY PARALLEL DATABASE)

USING GPUS FOR REAL-TIME QUERYING AND VISUALIZATION OF BIG DATA

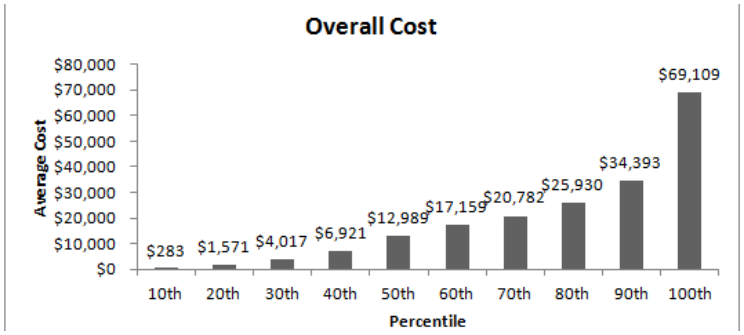
Interactive Large-Scale Visualization using a GPU Database

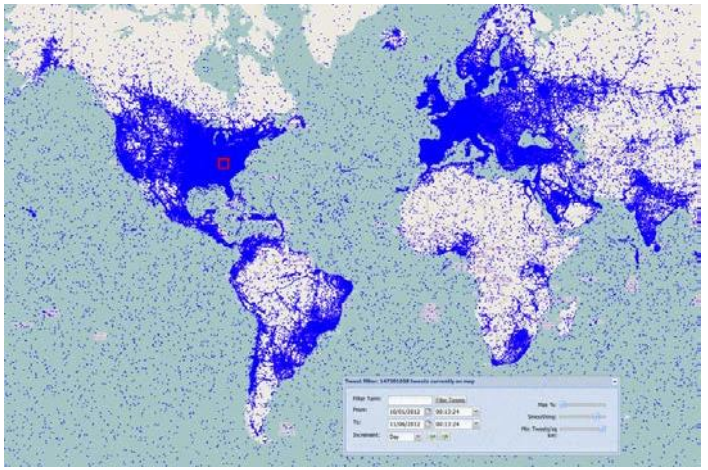
Todd Mostak

The Need for Interactive Analytics

- Idea: often need to browse massive data sets
- Browsing is best supported through visualization

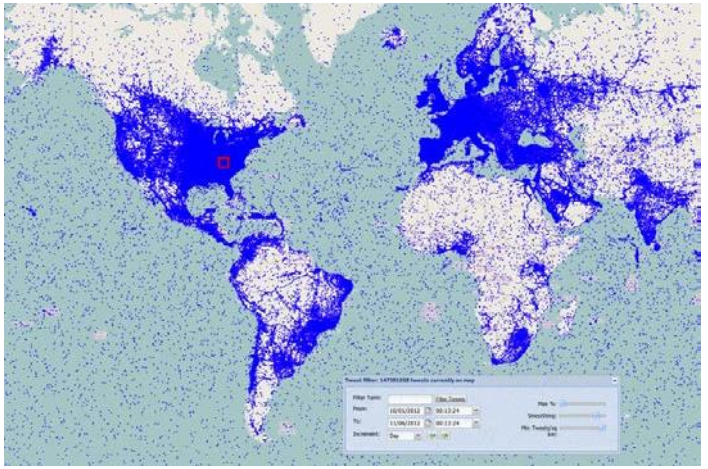
➔ ad-hoc analytics, with millisecond response times





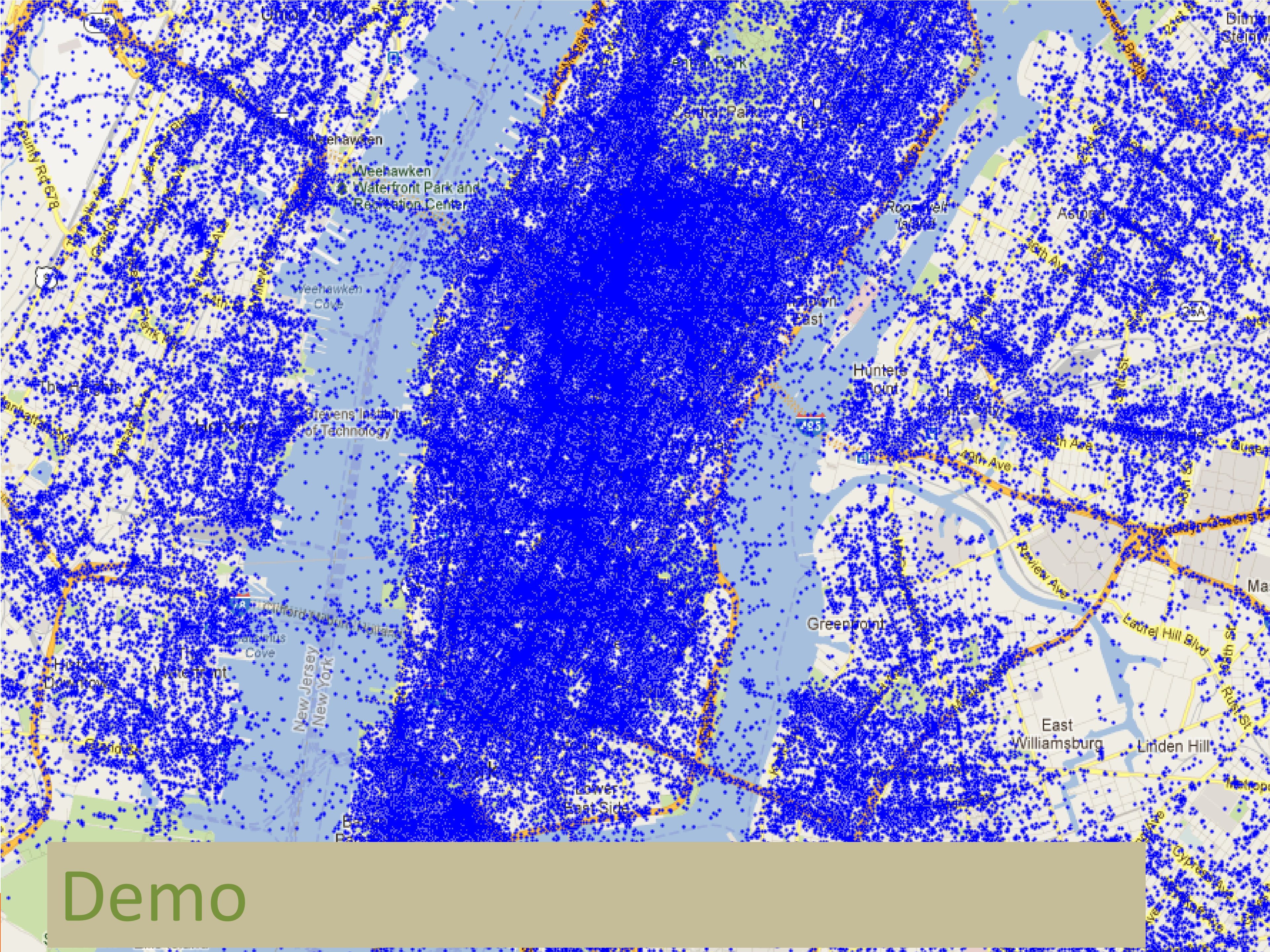
MapD: GPU Accelerated SQL Database

- *Key insight:* GPUs have enough memory that a cluster of them can store substantial amounts of data
- Not an accelerator, but a full blown query processor!
- Massive parallelism enables interactive browsing interfaces
  - 4x GPUs can provide > 1 TB/sec of bandwidth
  - 12 Tflops compute
  - Order of magnitude speedups over CPUs, when data is on GPU
- "Shared nothing" arrangement



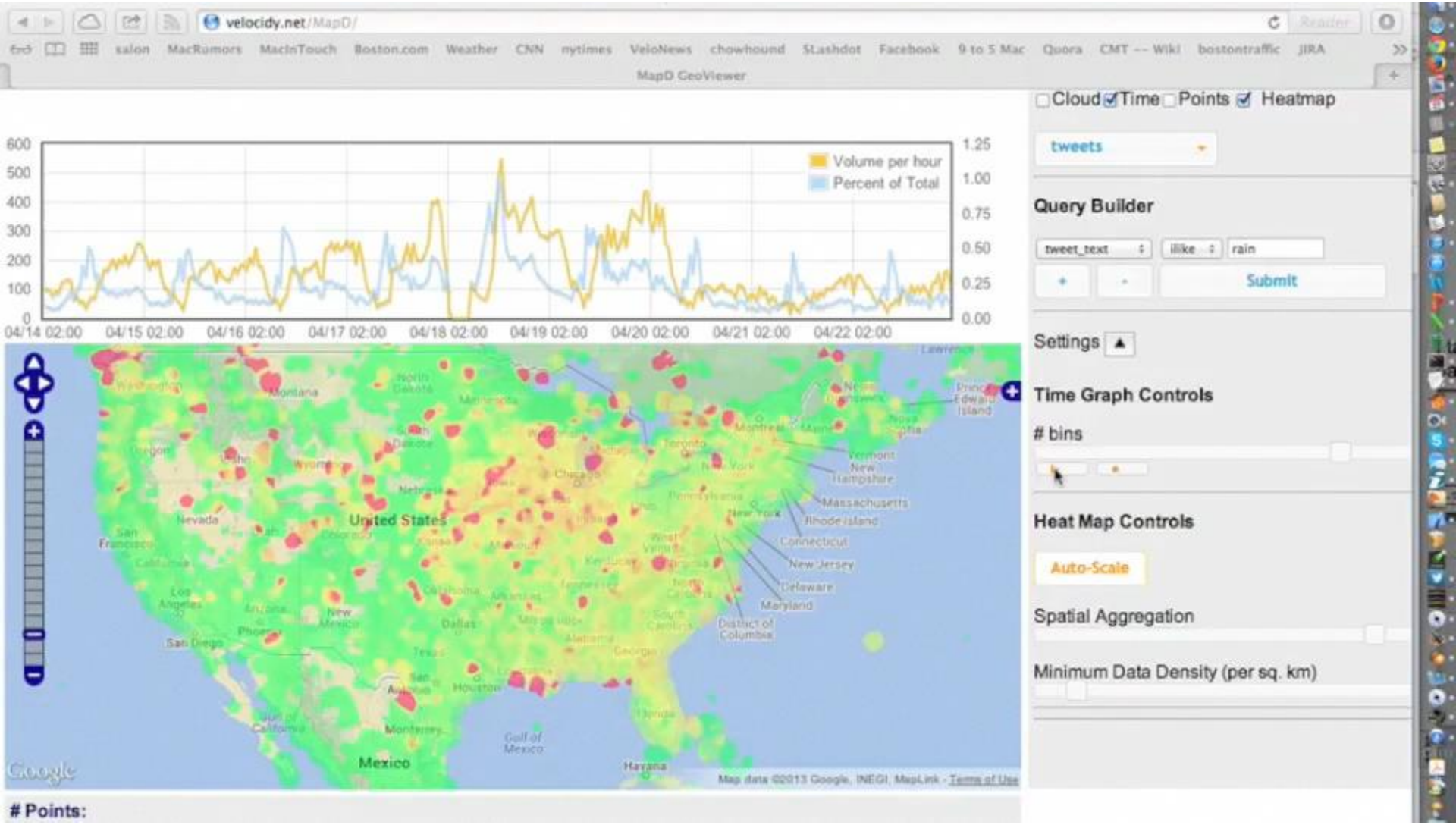
147,201,658 tweets from Oct 1, 2012 to Nov 6, 2012

Relative intensity of "tornado" on Twitter (with point overlay) from Febuary 29, 2012 to March 1, 2012

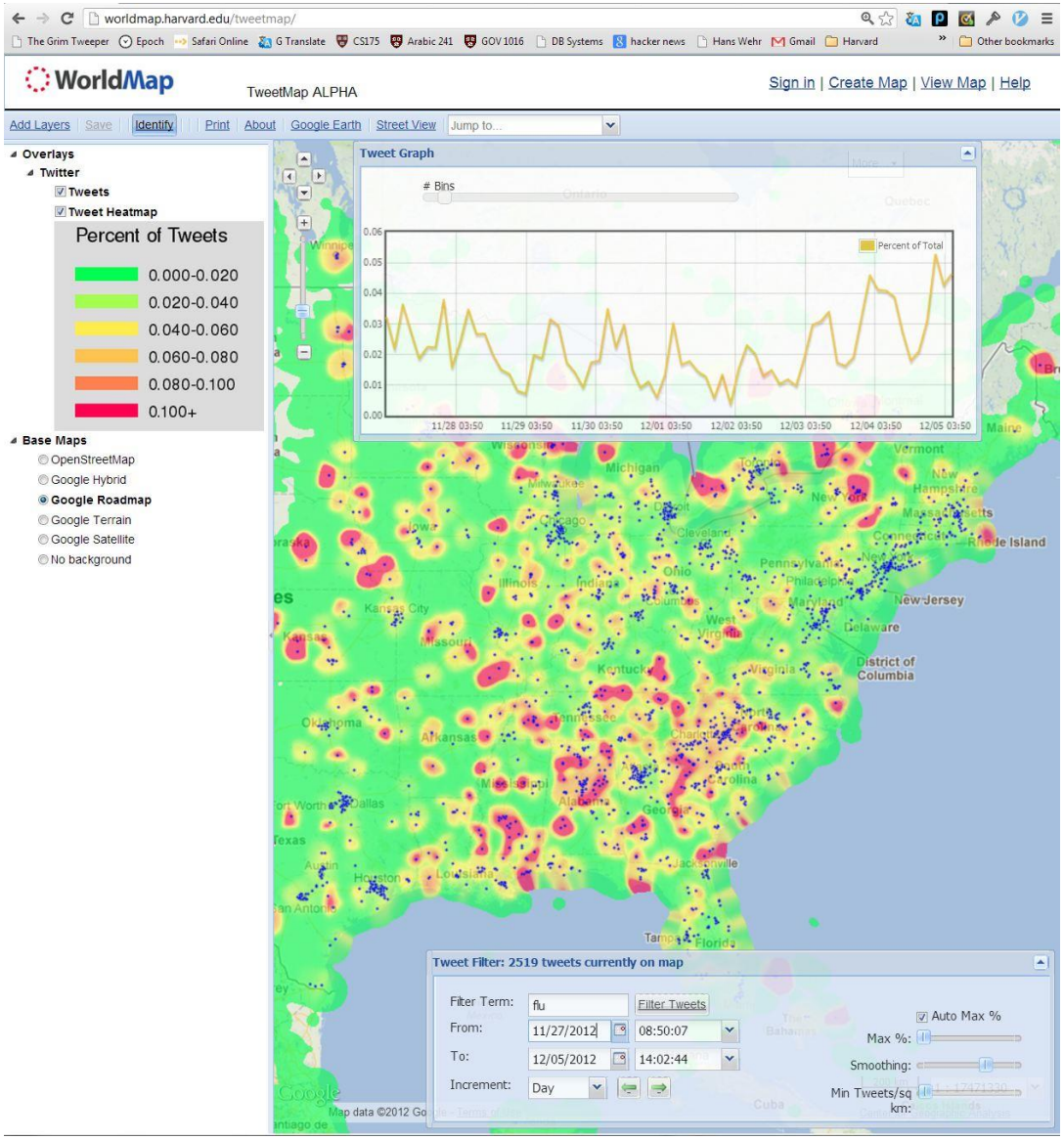


Weehawken
Waterfront Park and
Recreation Center
Weehawken
Cove
Stevens Institute
of Technology
New Jersey
New York
Hunters
Point
Greenpoint
East
Williamsburg
Linden Hill
Laurel Hill Blvd
Review Ave
Lower
East Side

Demo


velocidy.net/MapD/
MapD GeoViewer
Cloud
Time
Points
Heatmap
tweets
Query Builder
tweet_text
ilike
rain
+
-
Submit
Settings
Time Graph Controls
bins
Heat Map Controls
Auto-Scale
Spatial Aggregation
Minimum Data Density (per sq. km)
Volume per hour
Percent of Total
04/14 02:00
04/15 02:00
04/16 02:00
04/17 02:00
04/18 02:00
04/19 02:00
04/20 02:00
04/21 02:00
04/22 02:00
United States
Mexico
Google
Points:


Tweet Filter: 2082 tweets currently on map
Filter Term:
central park
Filter Tweets
From:
10/14/2012
20:17:29
To:
11/04/2012
20:17:29
Increment:
Day
Max %:
Smoothing:
Min Tweets/sq km:
Central Park
Manhattan
Upper West Side
Upper East Side
Midtown East
Theater District
Times Square
Hell's Kitchen
Chelsea
Greenwich Village
West Village
Union Square
Flatiron Distrct
Gramercy Park
Kips Bay
East Harlem
Marcus Garvey
Riverside Park
Port Morris
Randalls Island
Ditmars Steinway
LaGuardia Airport
East Elmhurst
Jackson Heights
Astoria
Maspeth
Greenpoint
Rego Park
Middle Village
Linden Hill
Chinatown
New York
Guttenberg
West New York
Bulls Ferry & Jacob's Ferry



Search for “flu” showing outbreak over Southeastern U.S.

EXAMPLE DATA SETS

- Election data (donations, fundrasing)
- Sports data, at fine granularity (e.g., individual shots / passes)
- Education data (admissions, jobs, costs, loans)
- Medical data (medicare, billing, etc.)
- Federal funding (defense, nsf, etc)
- Real estate (transactions, property prices, restaurants, etc)