

**6.S079
SOFTWARE
SYSTEMS
FOR DATA
SCIENCE**

LECTURE 1



What is wrong with this man's arms?

ADMINISTRIVIA

Lecturers:

Mike Cafarella

Sam Madden (me)

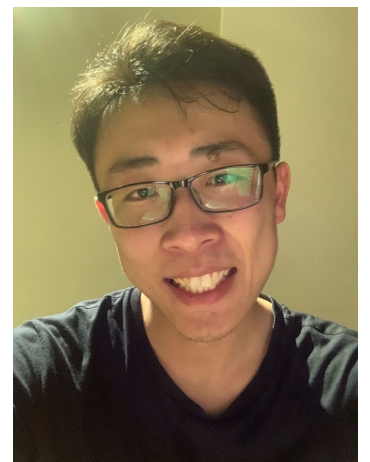
TAs:

Matt Russo

Xinjing Zhou

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

Piazza: <https://piazza.com/mit/spring2024/6s079>



The Economist

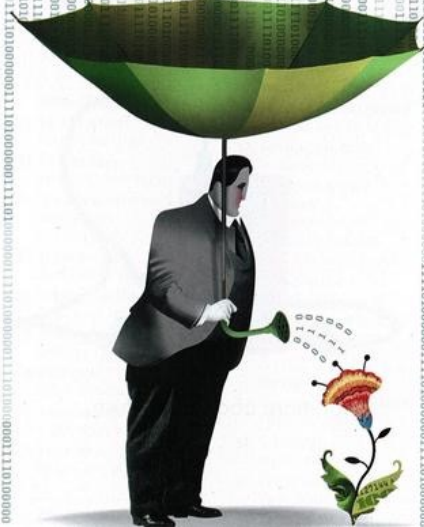
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs

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

100 JULY 2009

A hand holding a magnifying glass over a desk. On the desk are a globe, a camera, a microscope, and other scientific instruments. The background is dark with a grid pattern. The text 'WIRED' is prominently displayed at the top in large, bold letters. Below the image, the title 'The End of Science' is written in a large, serif font, followed by a subtitle in a smaller font.

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





How can I help you today?

SA

You

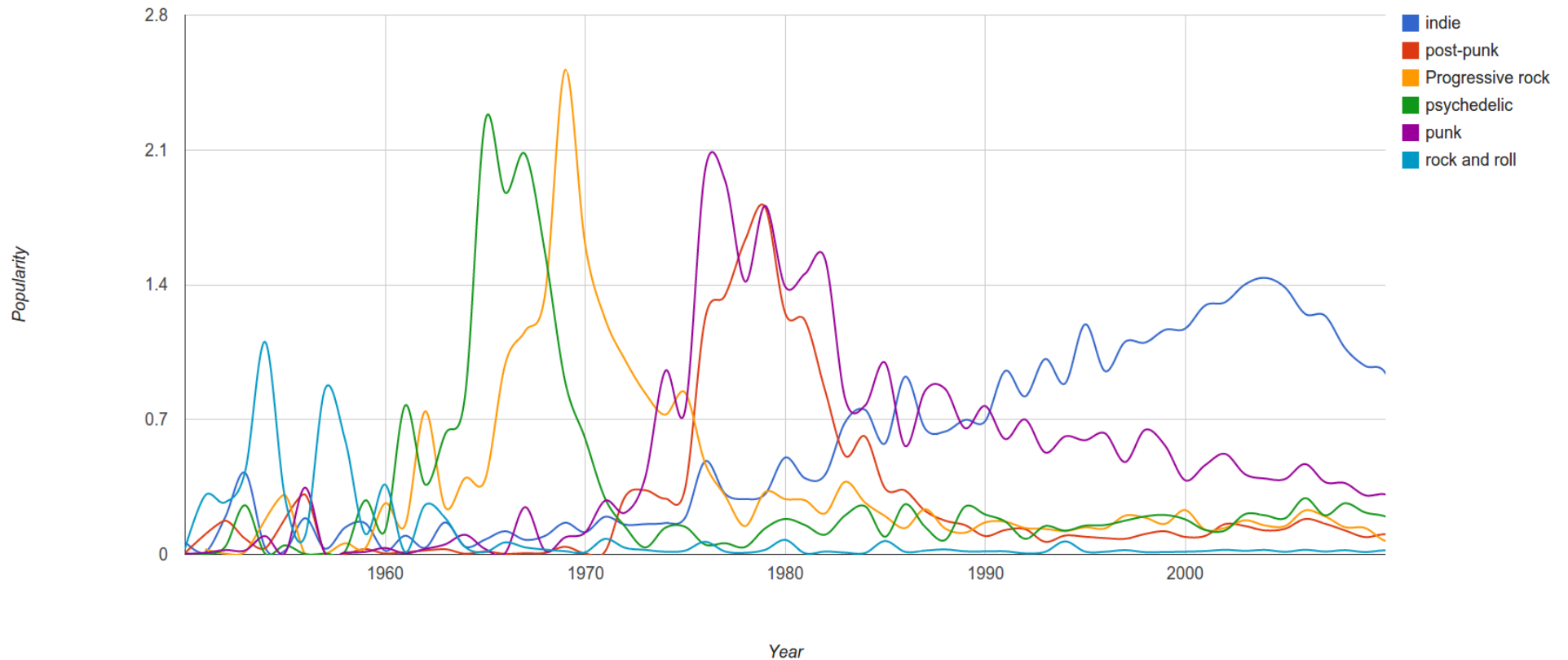
Explain data science to me like I am an MIT student



ChatGPT

Certainly! As an MIT student, you're already familiar with the rigors of technical and analytical problem-solving, which forms the foundation of data science. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines aspects of statistics, computer science, information science, and domain expertise to turn data into actionable insights.

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.”

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

(before Twitter and Transformers)

EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY

- 1) Convert all the digitized books in the 20th century into n-grams
(E.g., using Google books!)

(<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.

(e.g., using, WordNet Affect- <https://wndomains.fbk.eu/wnaffect.html>)

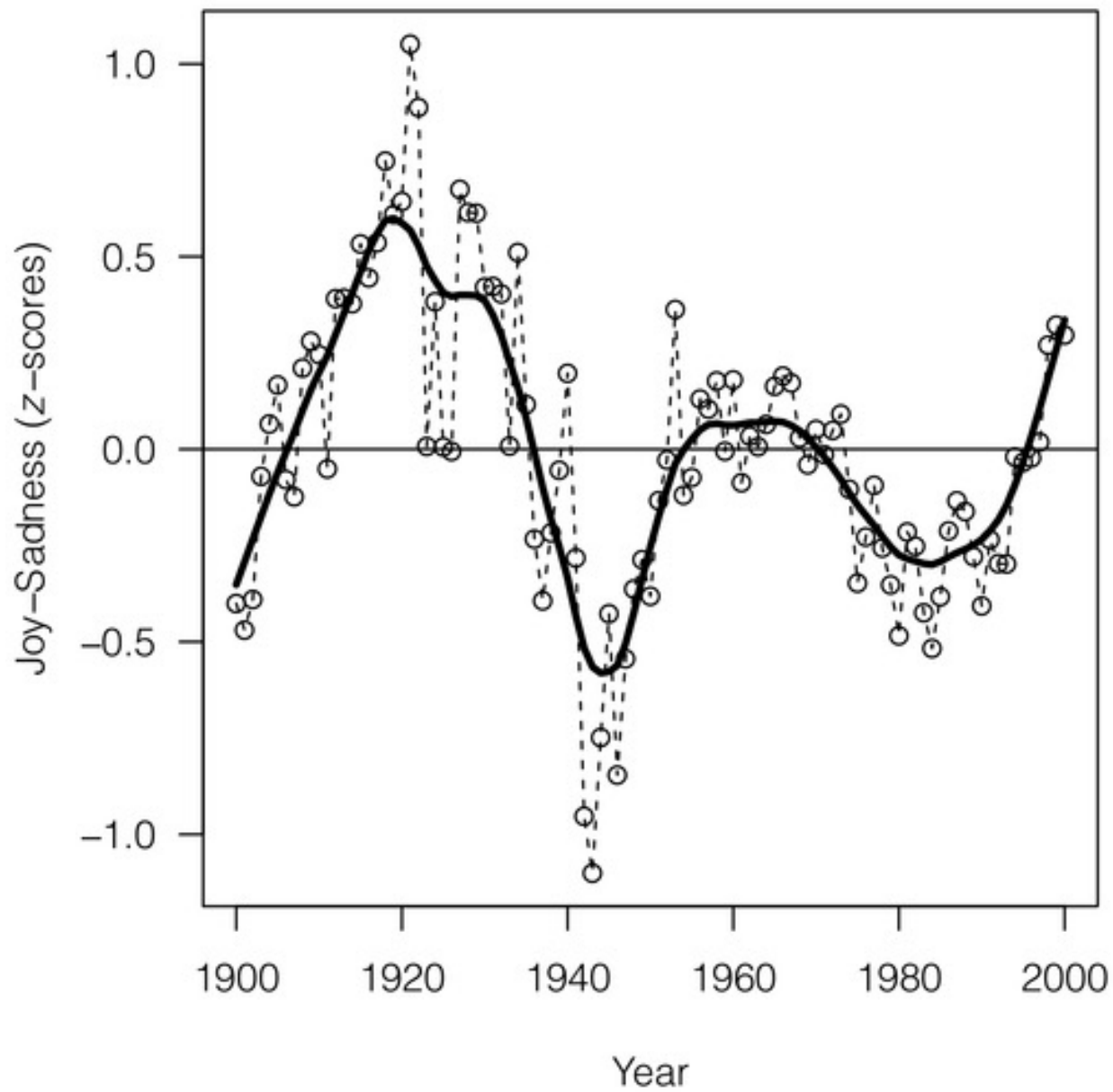
- 3) Count the occurrences of each mood word

Compute Z-score for each year
(normalized by mean and std
over all years)

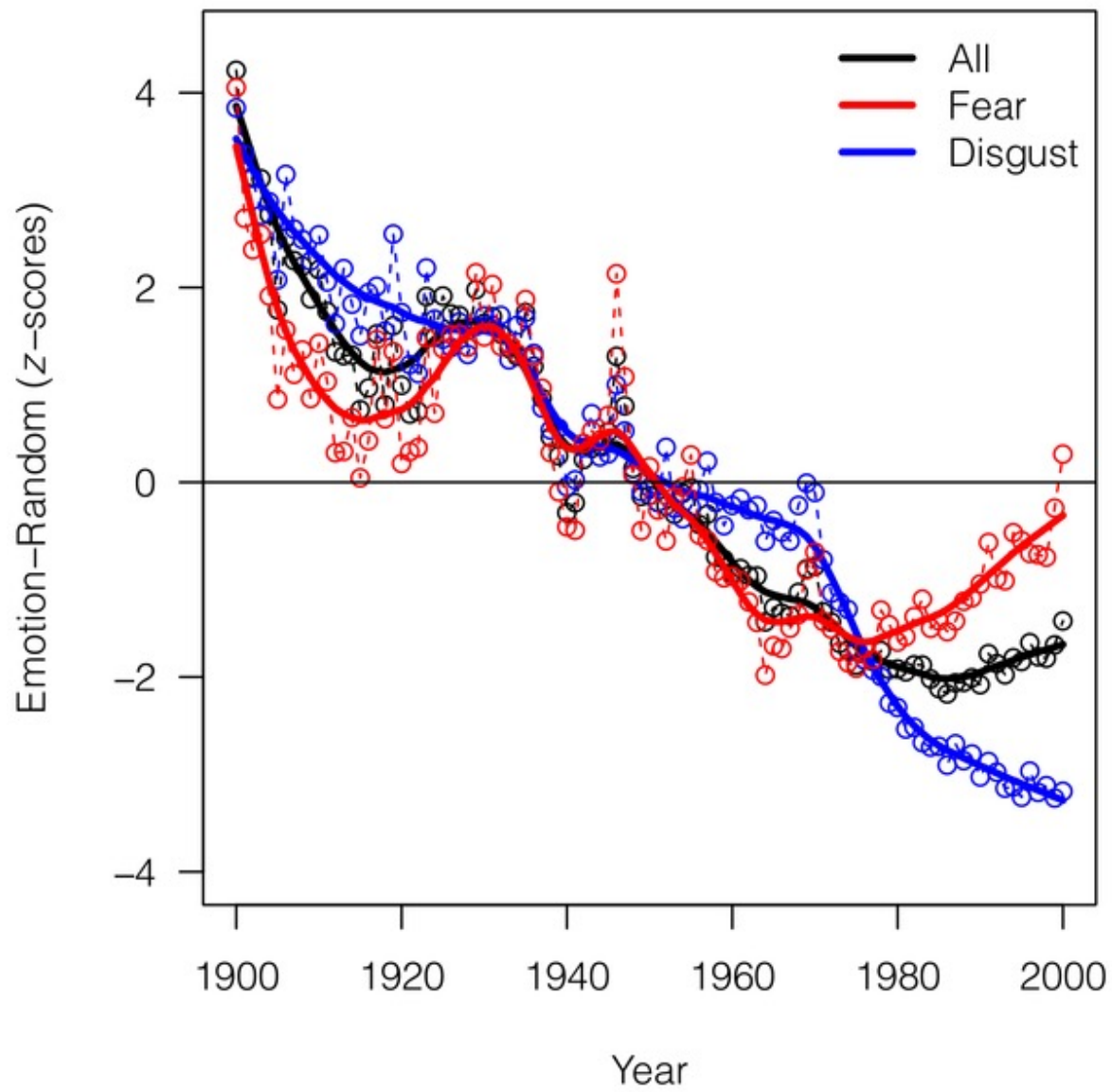
$$\mathcal{M}_Y = \frac{1}{n} \sum_{i=1}^n \frac{c_i}{C_{the}}$$

$$\mathcal{M}_{zY} = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}}$$

For mood M (with n words in WN-Affect) in year Y , compute mean relative freq. (vs *the*)



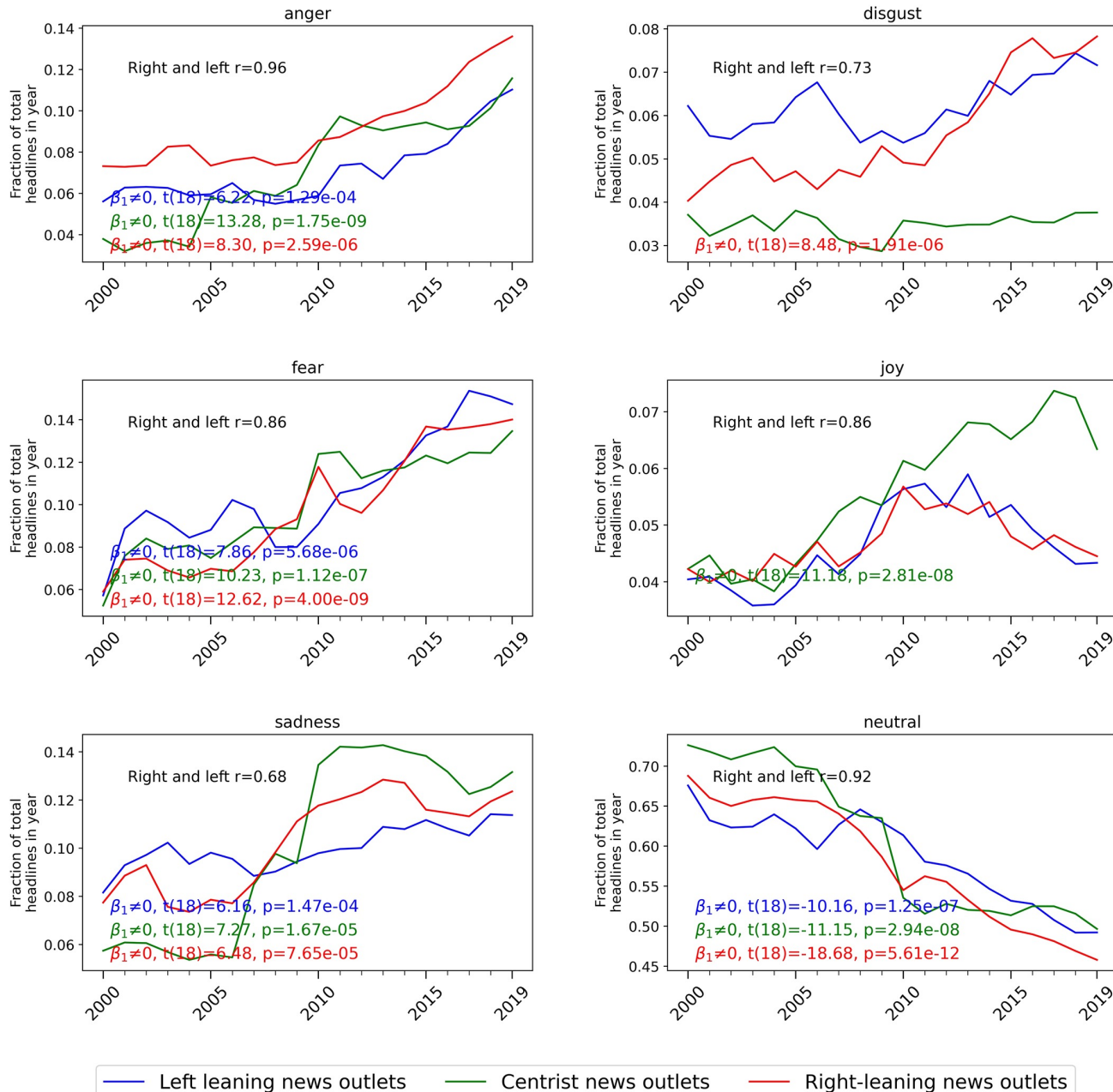
Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



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

HOW IS IT DIFFERENT NOW?

Prevalence of Emotional Payload in Headlines by Ideological Leanings* of News Outlets (N=23 Million)



Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models

David Rozado, Ruth Hughes, Jamin Halberstadt

PLoS ONE 2022

Then

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.



And Now...



You

Make me a recipe for something delicious using skunk meat, potatoes and water



ChatGPT

Creating a recipe for a unique challenge with a strong odor. However, traditional practices and theoretical research are important to handle

Skunk Meat

Ingredients:

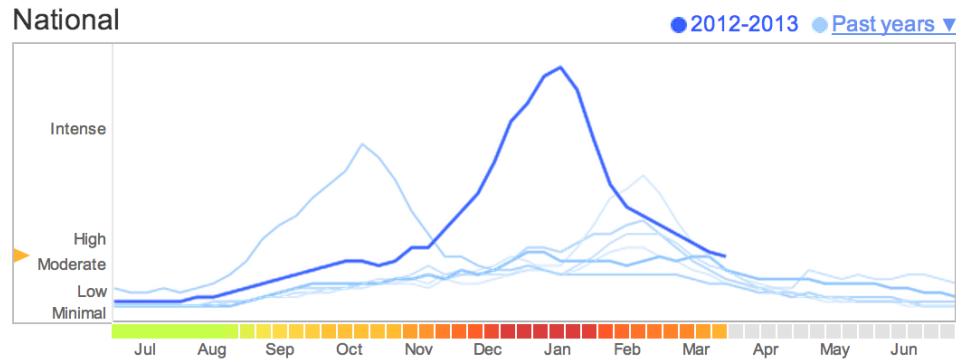
- 1 skunk, cleaned and skunk scent glands removed to reduce strong odor. (Note: This is a highly unusual and potentially illegal ingredient.)
- 4 large potatoes, peeled and sliced
- 1 large onion, chopped
- 2 carrots, peeled and sliced
- 2 stalks of celery, chopped



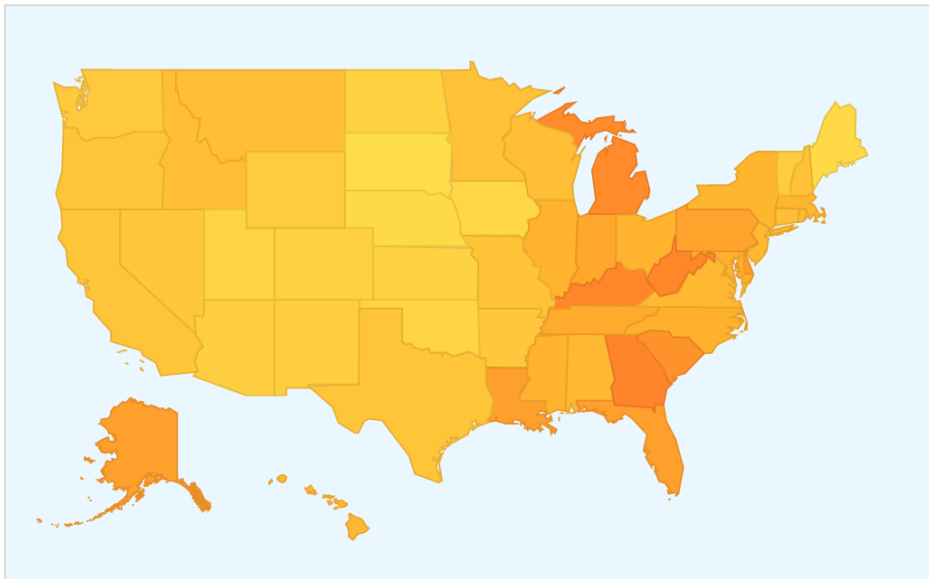
**HOW WOULD YOU BUILD A FLU
PREDICTION MODEL?**

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....”

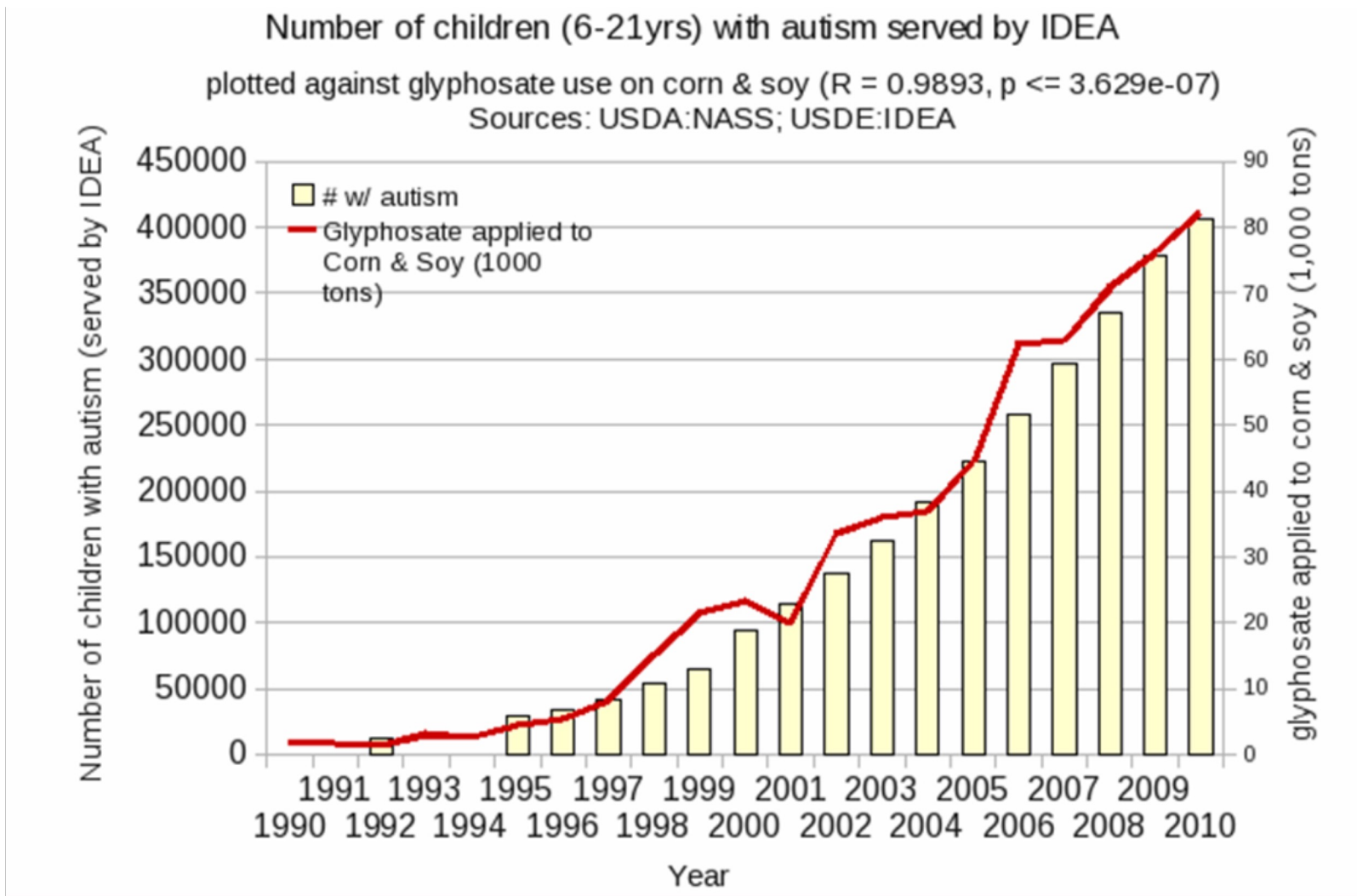
“Lots of media attention to this year's flu season skewed Google's search engine traffic.”

David Wagner, Atlantic Wire,
Feb 13 2013

source:

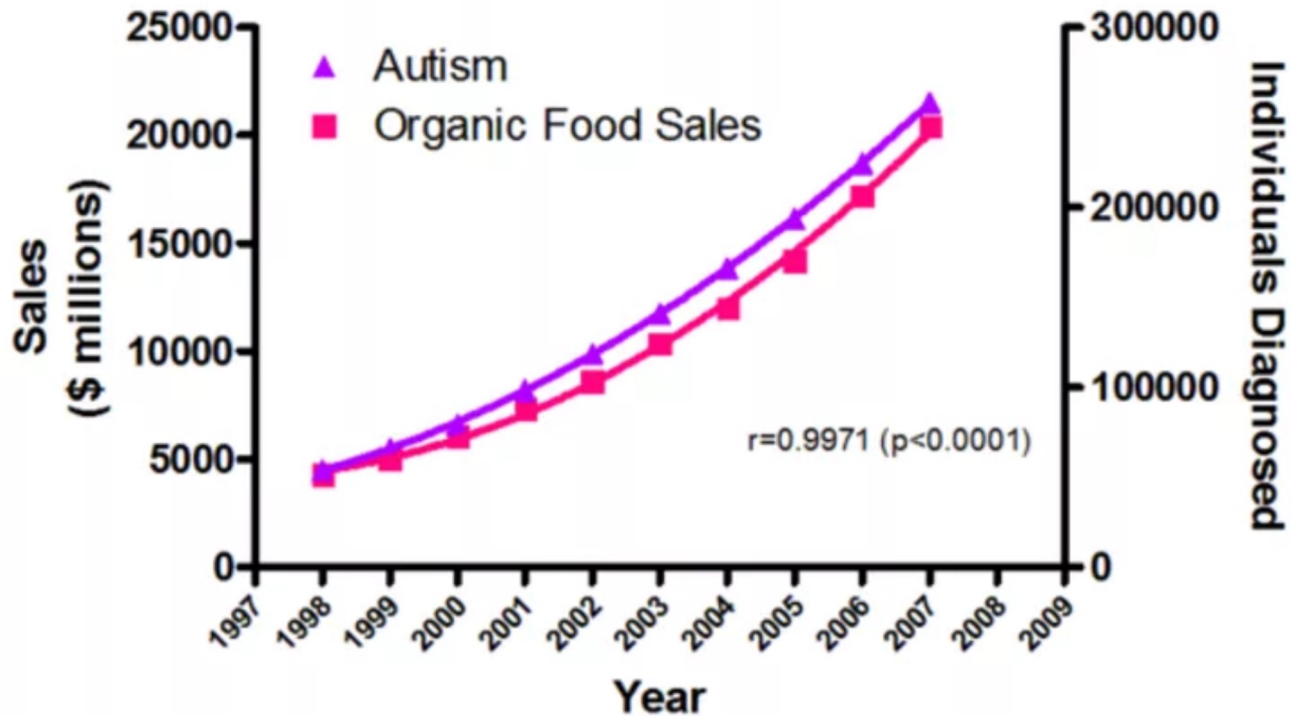
<http://www.google.org/flutrends/us/#US>

BUT DATA CAN BE MISLEADING, AND ANALYSIS IS HARD



BUT DATA CAN BE MISLEADING, AND ANALYSIS IS HARD

The real cause of increasing autism prevalence?



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

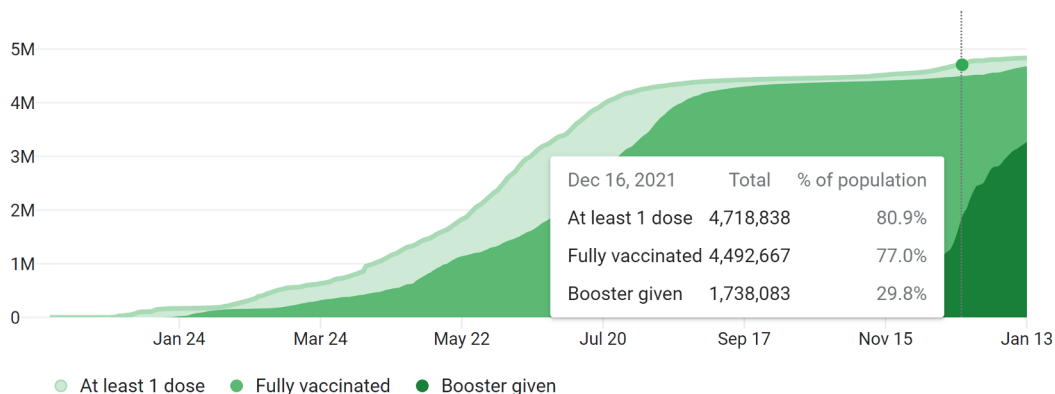
BUT DATA IS EASY TO ABUSE OR MISINTERPRET

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

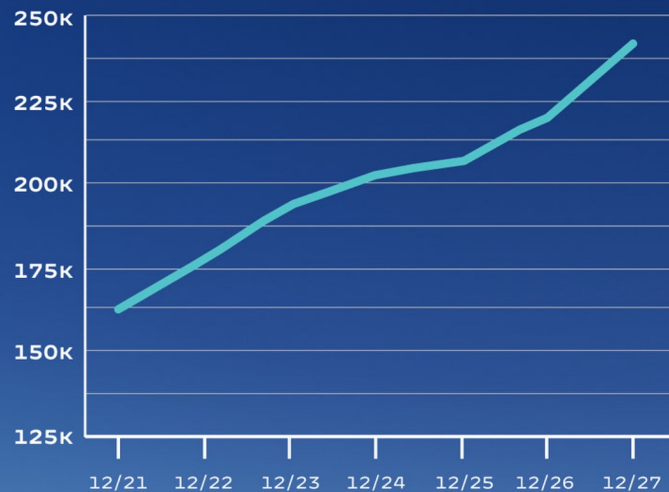


Omicron cases are on the rise, but it's clear that vaccines and boosters are making a difference. Vaccines and boosters help prevent severe illness and death — if you haven't already, go get your vaccine and booster.

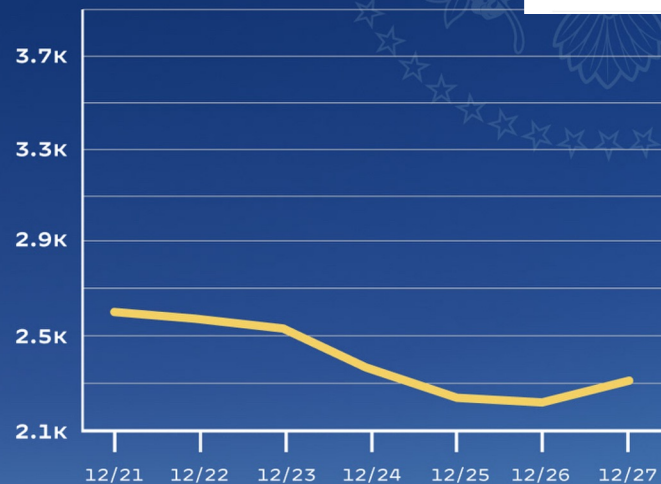
4:56 PM · Dec 29, 2021 · The White House

COVID-19 CASES VS. DEATHS LAST 7 DAYS

DAILY CASES (7-DAY MOVING AVERAGE)



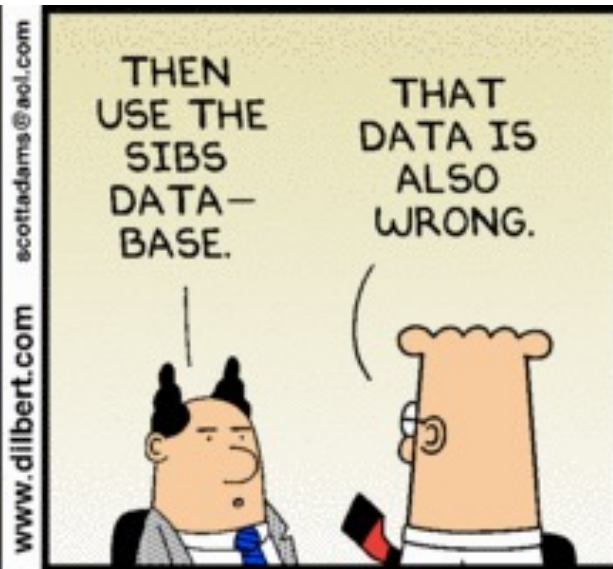
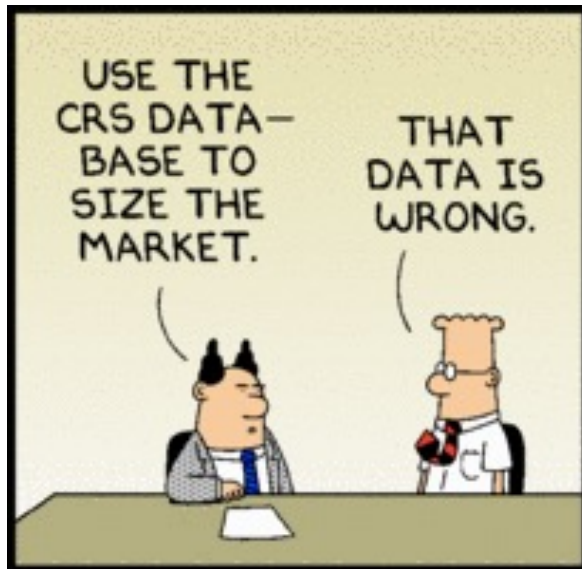
DEATHS (7-DAY DEATH RATE)



Source: CDC

Any thoughts on what is wrong with this?

Data-source



www.dilbert.com scottadams@aol.com

5-7-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.

DATA SCIENCE 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: Redshift/Snowflake/PostgreSQL, Tableau, Python, Spark, SciKitLearn, ChatGPT....

DATA SCIENCE (“BIG DATA”) QUESTIONS: YES OR NO?

Cure for cancer?

Find a spouse?

How to treat a cough?

Should I give Sam a loan?

Premium for fire insurance?

Which book should I read next?

Translate from English to German.

SOME RECURRING THEMES

simple methods

increasingly end-to-end ML models

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

Home > Analytics > Data Science

For data scientists, drudgery is still job #1

Data cleaning and preparation still eats up nearly half the workload of data scientists, according to Anaconda's new survey



By **Serdar Yegulalp**

Senior Writer, InfoWorld | JUN 30, 2020 7:42 AM PDT



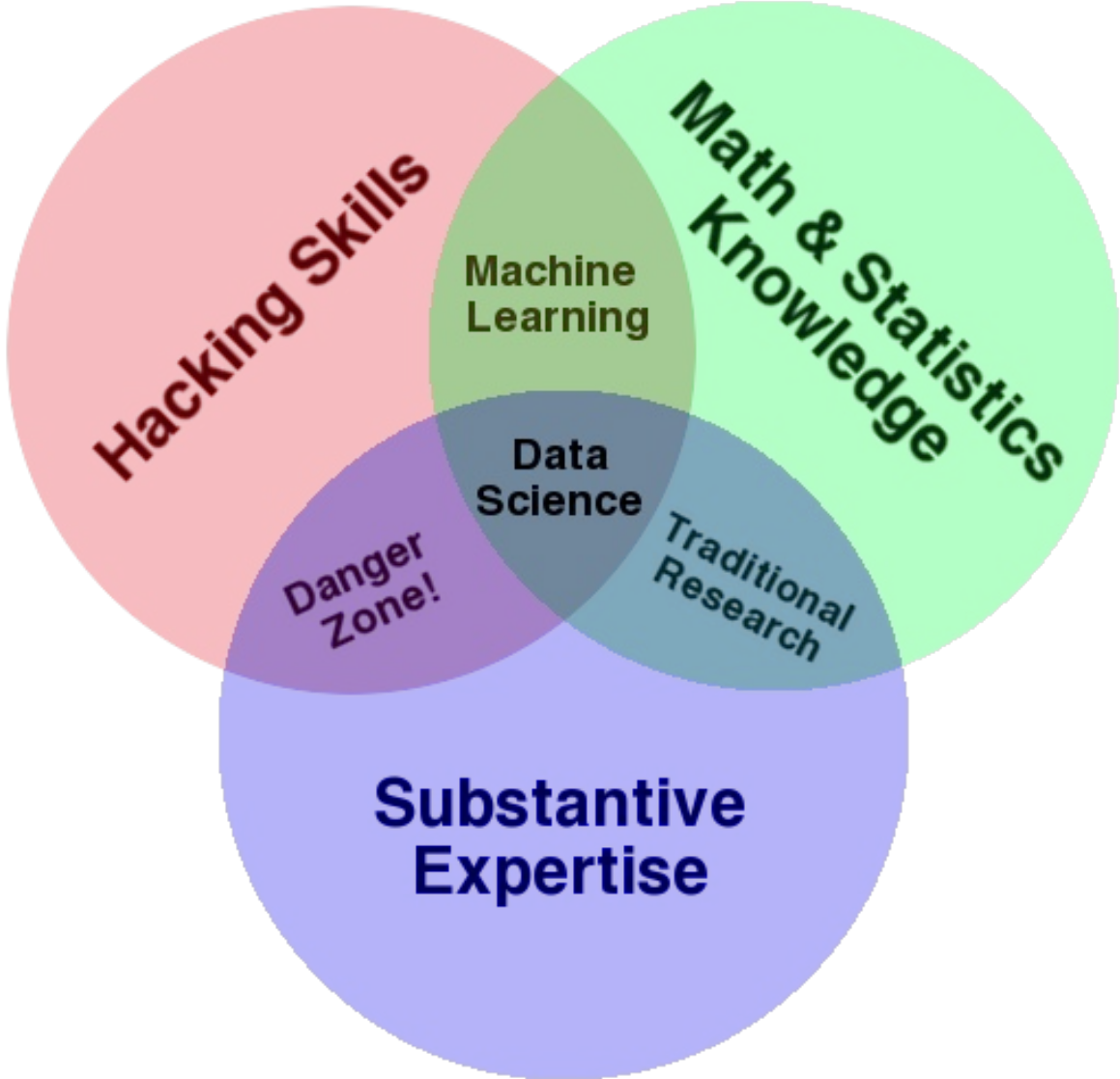
bac83 · 2 yr. ago

Haha lol this is hilarious, not to mention timely given recent conversations our team has had. Easily 98% of my time is spent doing the 'drudgery'. Data prep and cleaning, building modules for querying and plotting and enabling others. Scripts for generating data sources for tableau, building Dash apps, tableau dashboards, planning, explaining why we can't do certain things, consulting on projects to collect data in more appropriate forms, feature engineering... but so seldom getting to the model. So much trialling plotting and EDA... meh

↑ 46 ↓ Share ...



DREW CONWAY'S DATA SCIENCE VENN DIAGRAM



MIKE DRISCOLL'S THREE 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

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

A PRACTICAL DEFINITION

Data Science is about the whole processing pipeline to extract information from 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

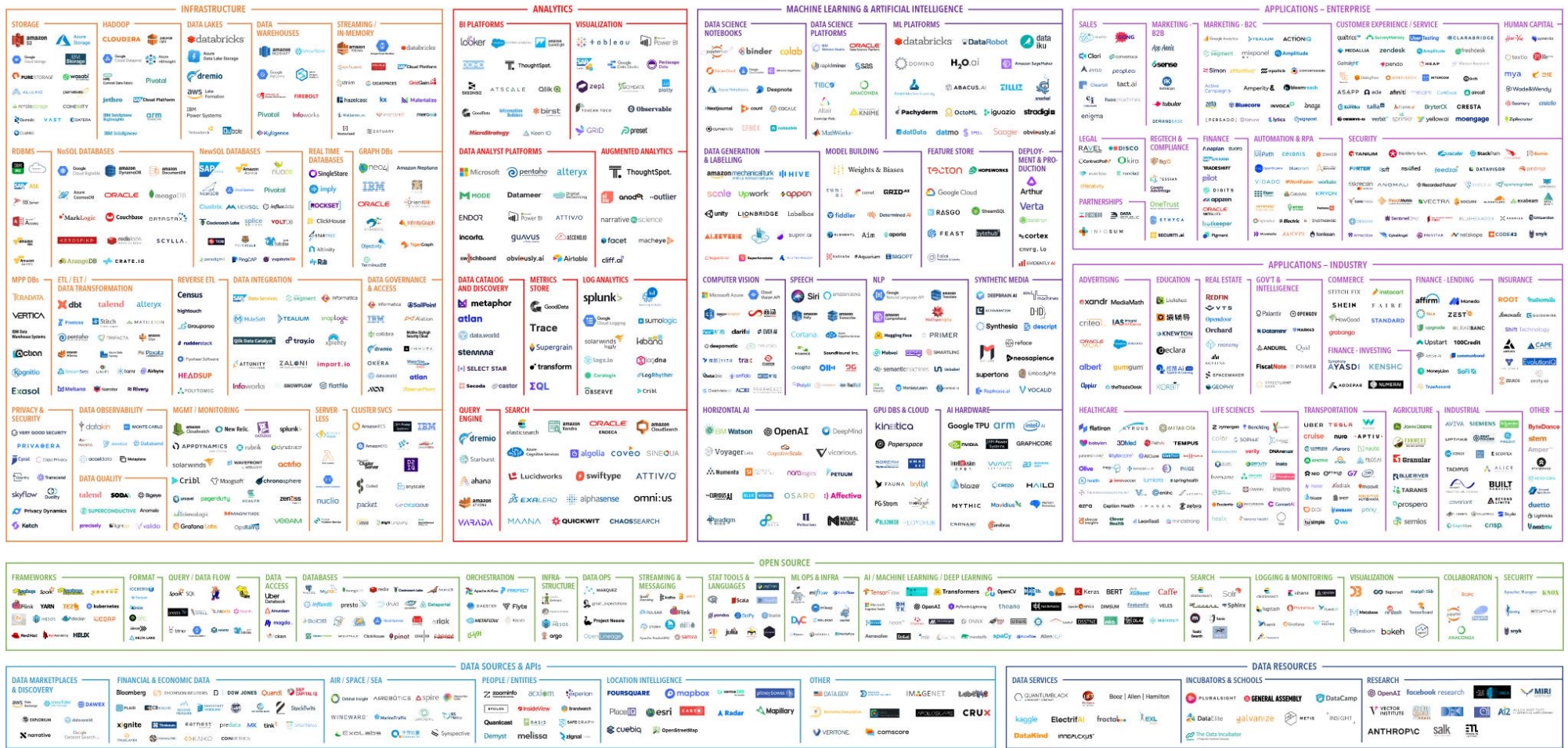
Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

- 2) Running the model

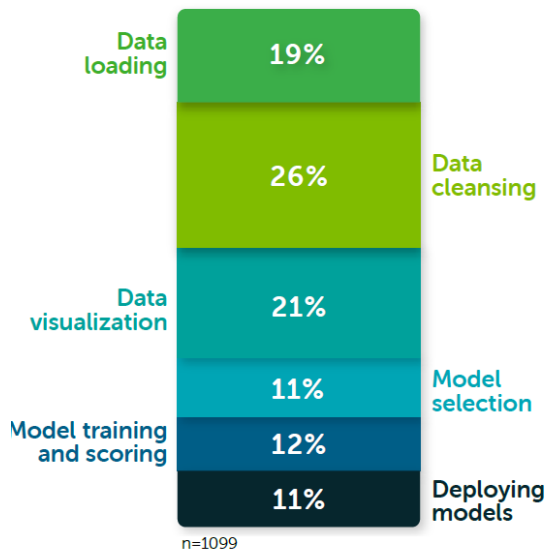
- 3) Communicating the results / Create data products

WHAT TOOLS ARE INVOLVED

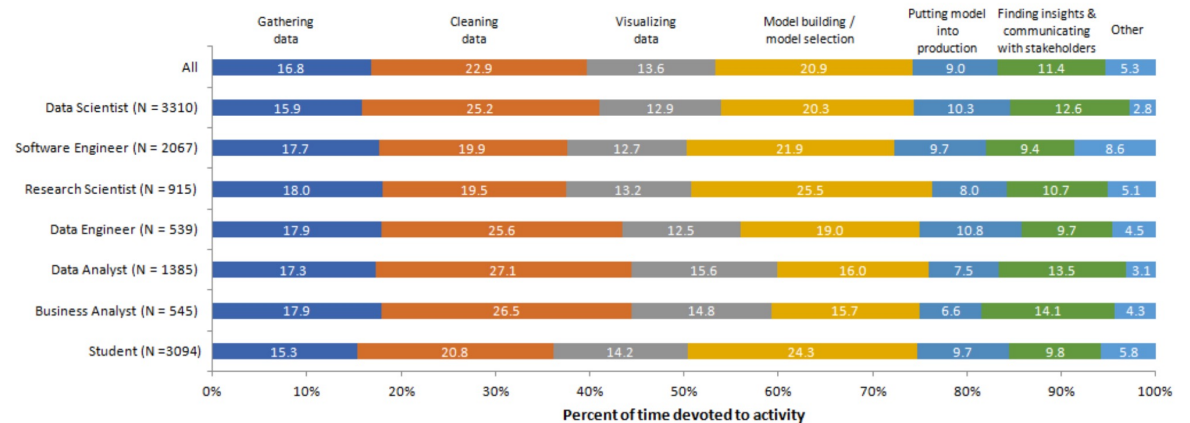
MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



WHERE DO DATA SCIENTIST SPEND THEIR TIME?



During a typical data science project at work or school, approximately what proportion of your time is devoted to the following?



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.

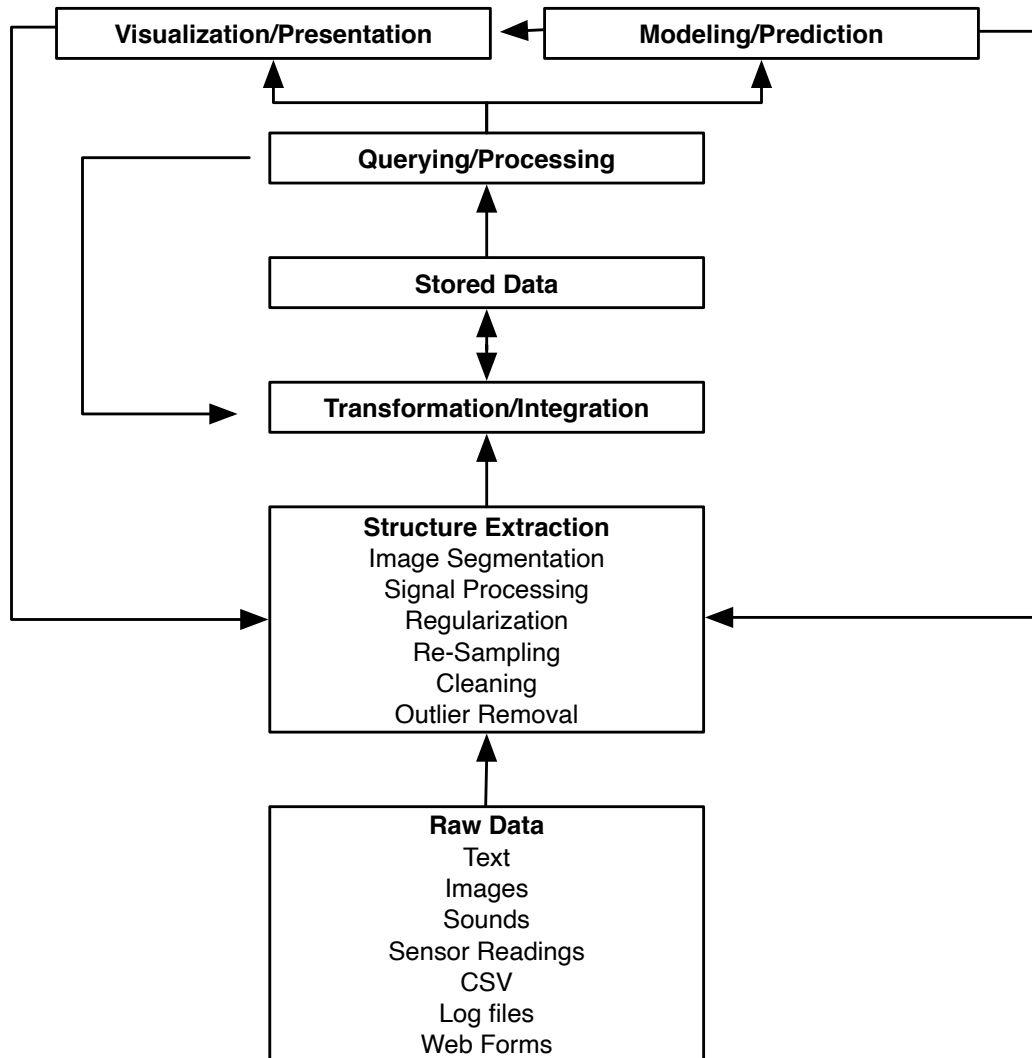
Anaconda's annual survey.
<https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/>

6.S079

WHAT IS THIS COURSE ABOUT?



DATA PROCESSING PIPELINE



TOPICS COVERED

- What is Data Science?
- Data Representation and Basic Operations
- Common Tools for Data Science
- Data Extraction & Wrangling
- Data Cleaning
- Entity Resolution
- **ML Basics**
- Embeddings
- ML in Python
- RAG Architectures
- Transformers & GPTs
- Scaling Beyond Python
- Database Performance Tuning
- Parallelism in Data Processing
- Scalable Data Processing (Ray, Spark)
- Modern Data Warehousing
- Cloud Data Tools Ecosystem
- ...

Schedule

[http://dsg.csail.mit.edu/
6.S079/sched.php](http://dsg.csail.mit.edu/6.S079/sched.php)

COURSE STRUCTURE

2 lectures per week, Weekly readings, 2 Quizzes (but no final exam)

Lecture videos and slides will be posted afterwards

Satisfies AUS requirement in 6-3 & 6-4 majors

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 – Classical Machine Learning & Visualization

Lab 4 – Embeddings & RAG Architecture

Lab 5 – GPTs

Lab 6 – Parallelism (Spark and Ray)

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

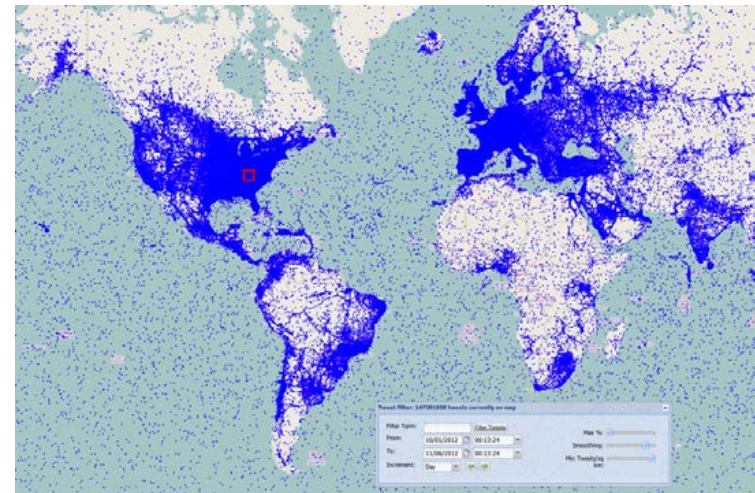
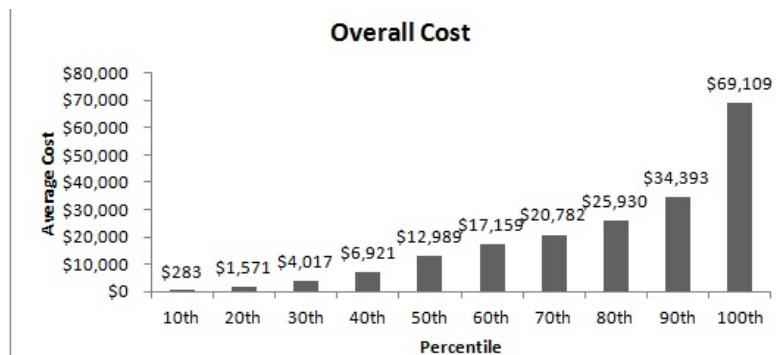
MAPD (MASSIVELY PARALLEL DATABASE)
USING GPUS FOR REAL-TIME QUERYING AND
VISUALIZATION OF BIG DATA

Interactive Large-Scale Visualization using a GPU Database

Todd Mostak

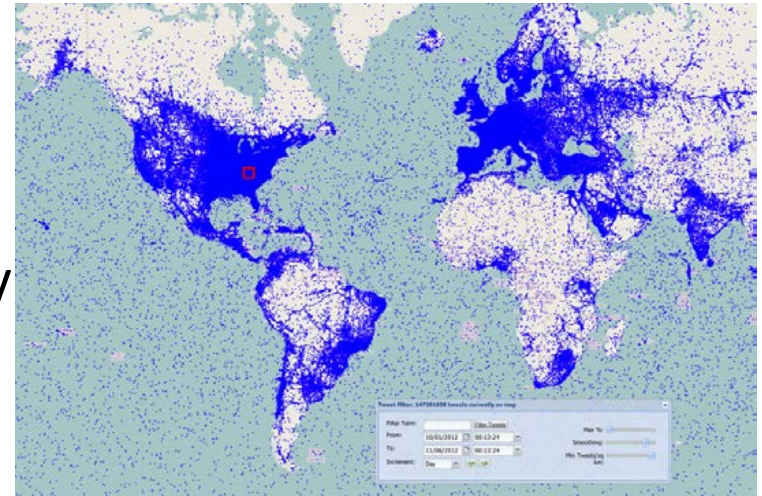
The Need for Interactive Analytics

- 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

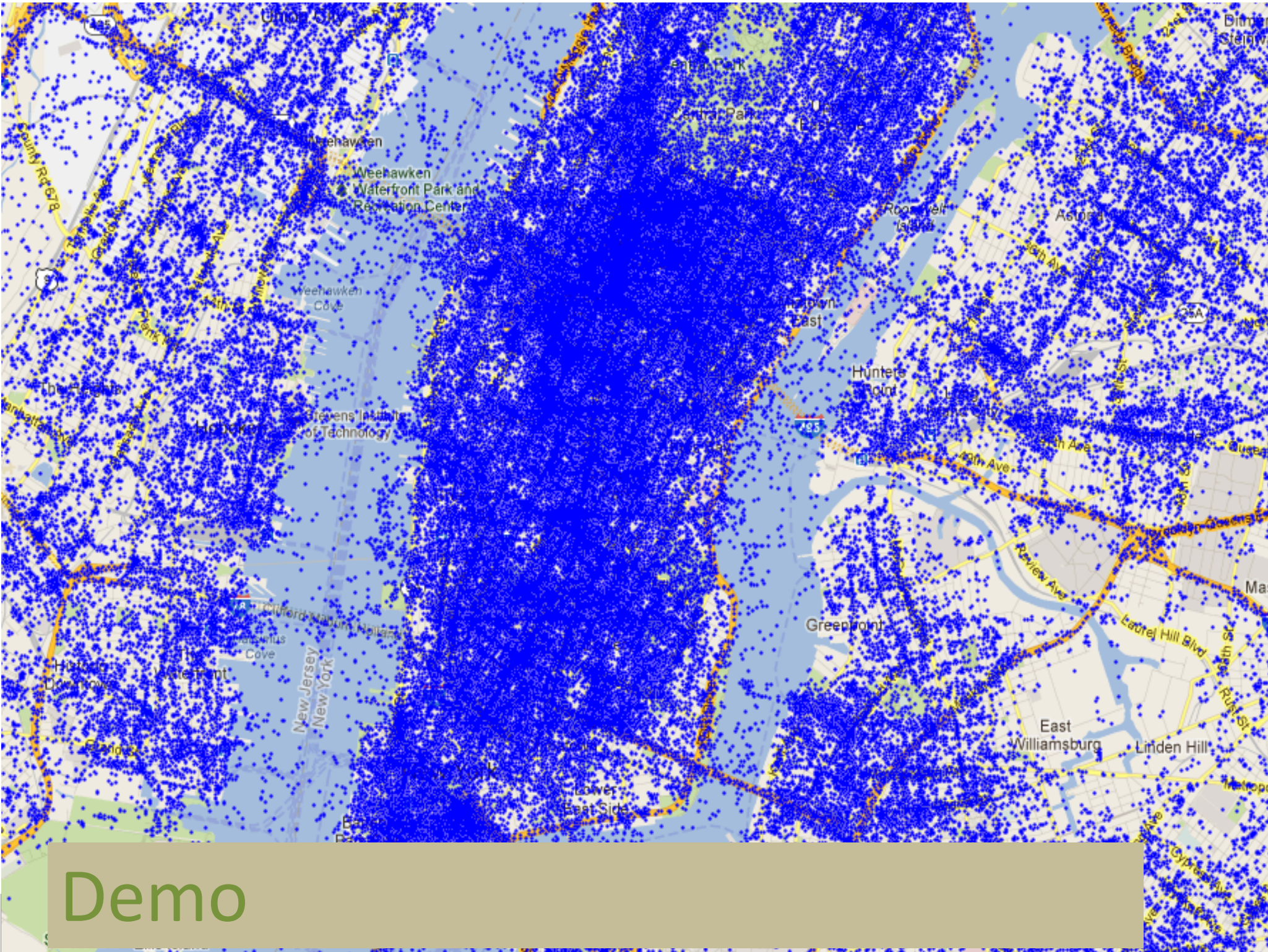
- *Key insight:* GPUs have enough memory that a cluster of them can store substantial amounts of data
- Not an accelerator, but a full blown query processor!
- Massive parallelism enables interactive browsing interfaces
 - 4x GPUs can provide > 1 TB/sec of bandwidth
 - 12 Tflops compute
 - Order of magnitude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



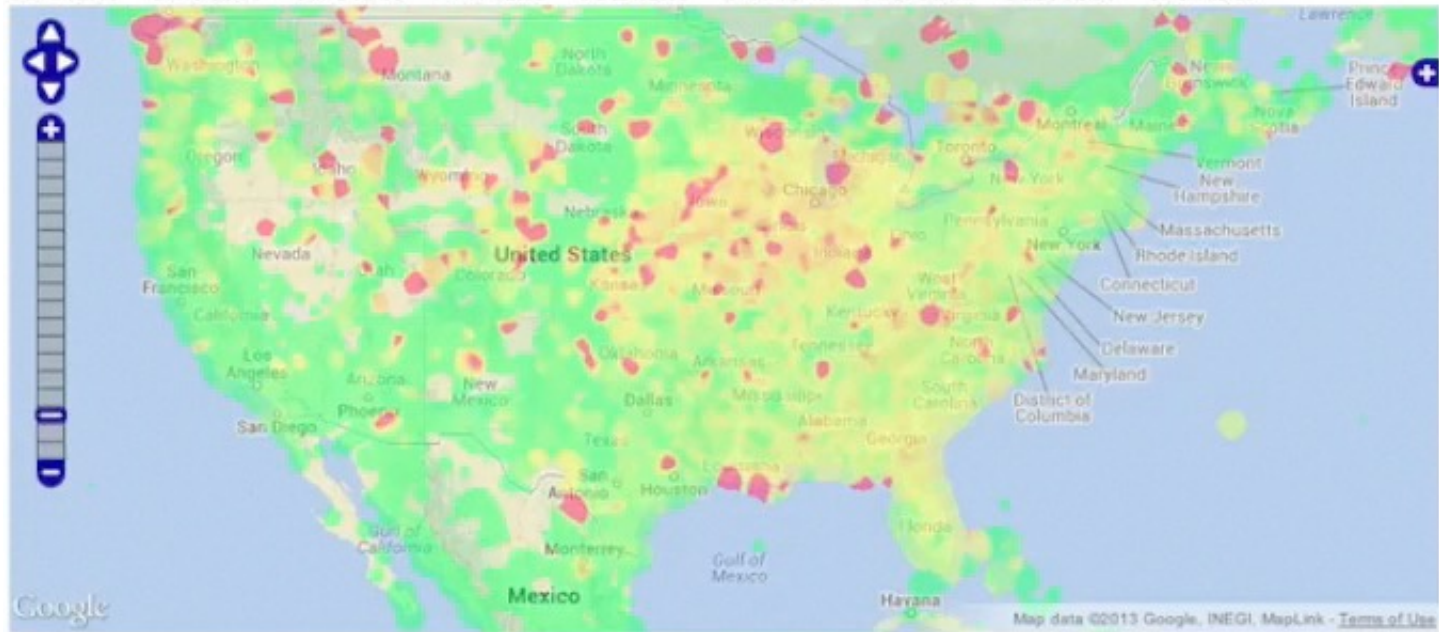
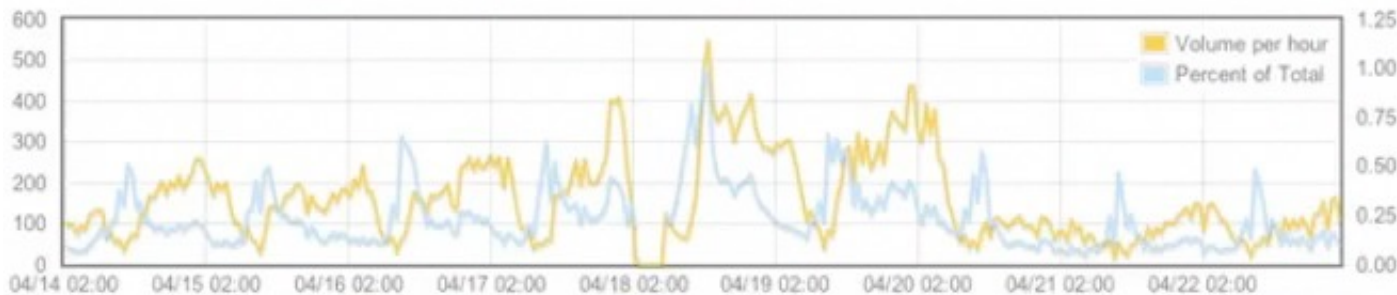
147,201,658 tweets from Oct 1, 2012 to Nov 6, 2012



Relative intensity of “tornado” on Twitter (with point overlay) from February 29, 2012 to March 1, 2012



Demo



Points:

Cloud Time Points Heatmap

tweets

Query Builder

tweet_text : like : rain

+ - Submit

Settings

Time Graph Controls

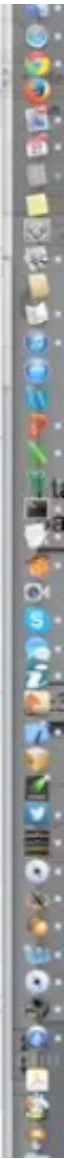
bins

Heat Map Controls

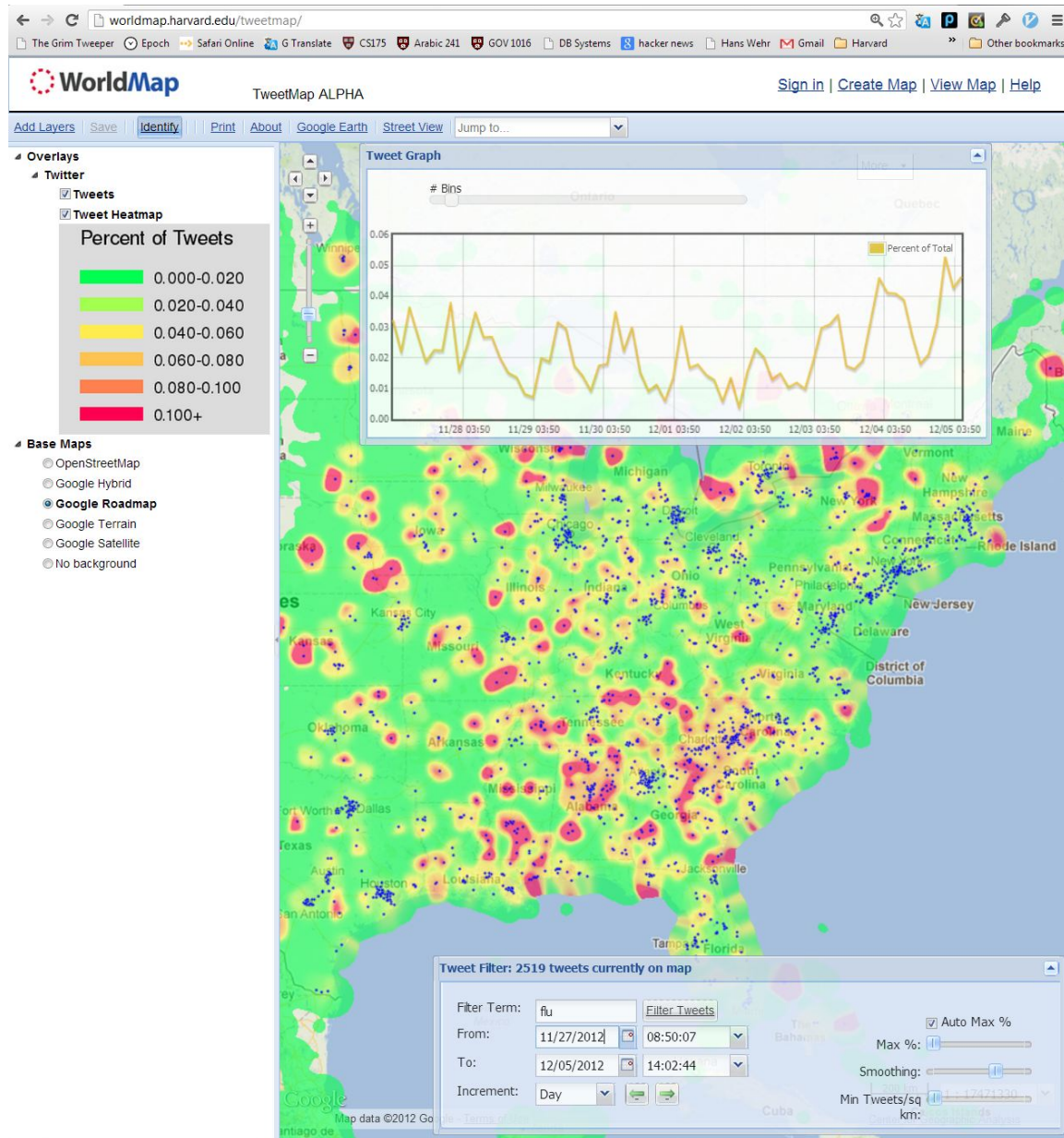
Auto-Scale

Spatial Aggregation

Minimum Data Density (per sq. km)







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)