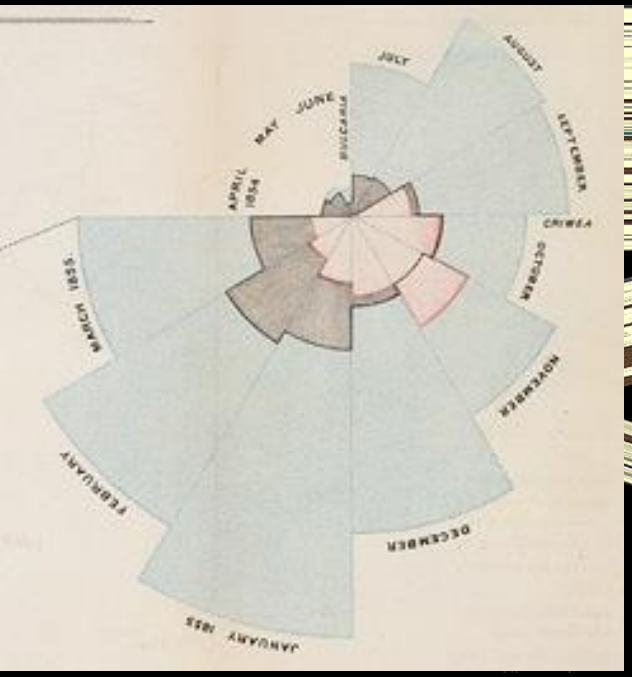
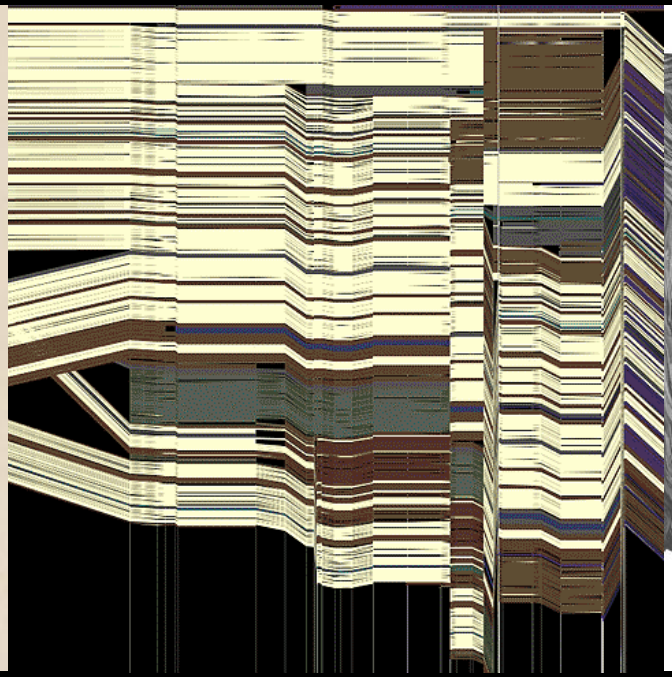
CSE 442 - Data Visualization

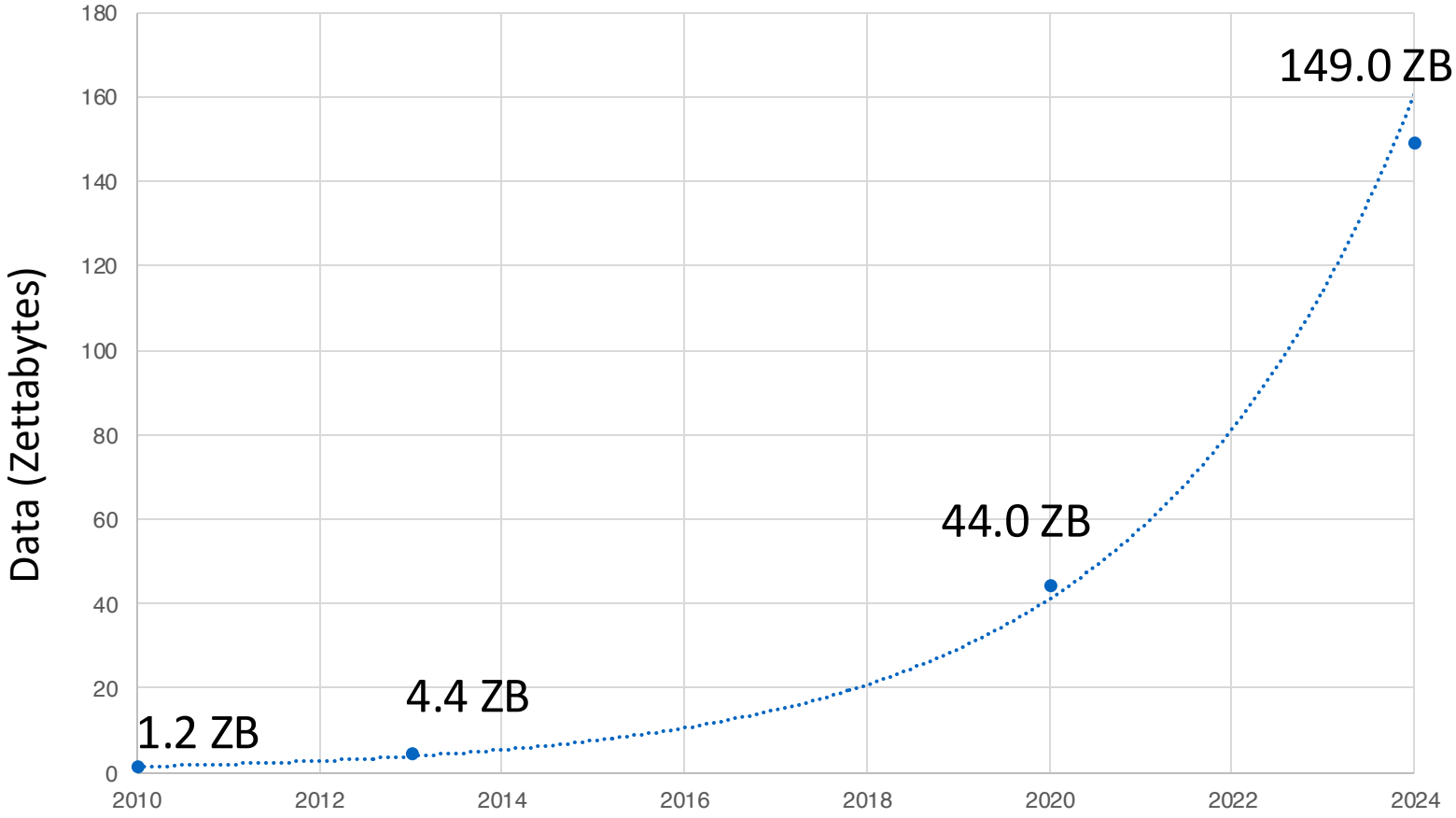
The Value of Visualization





Leilani Battle University of Washington

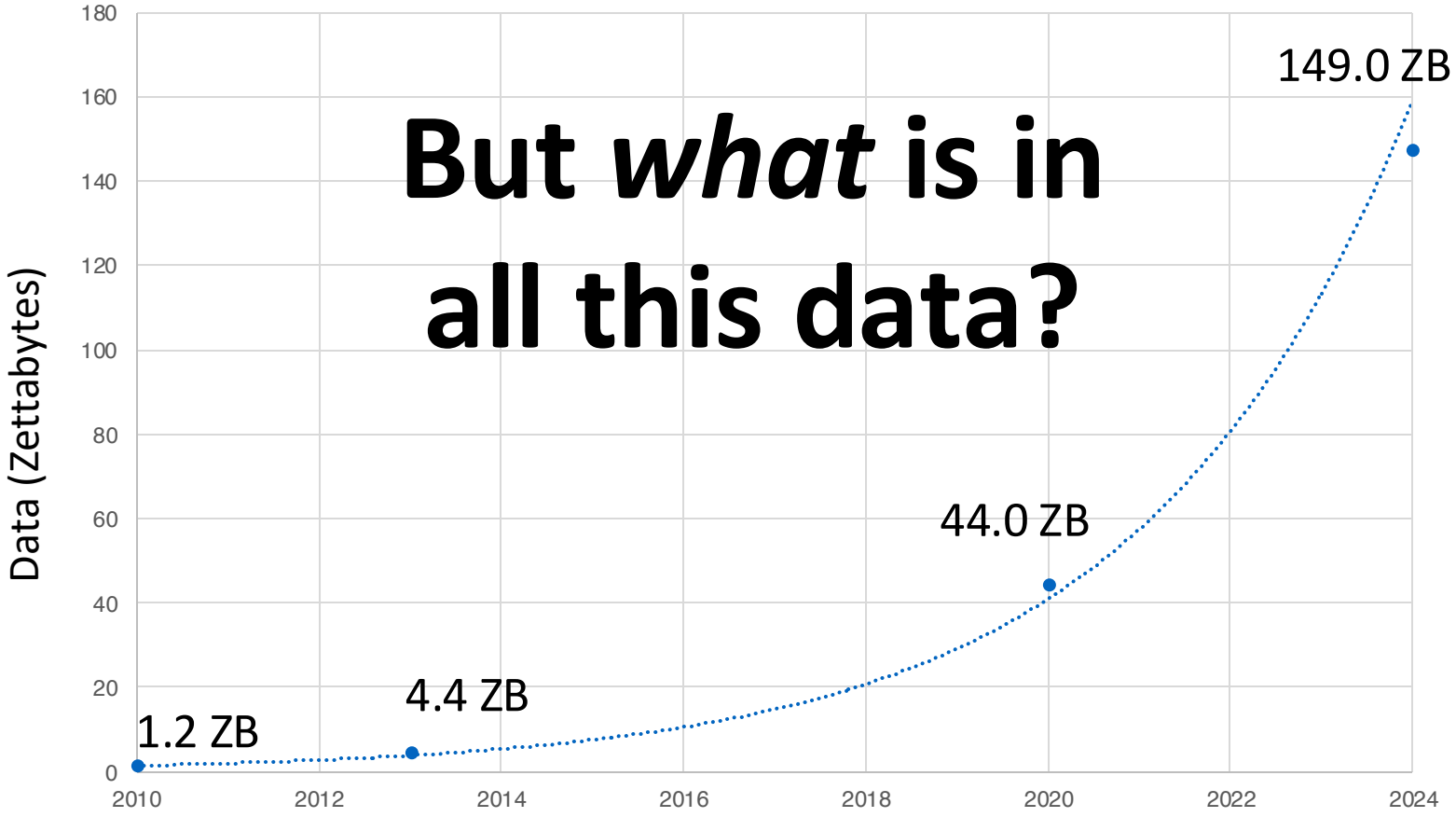
Data Doubles Roughly Every Year



~2x every 2 years

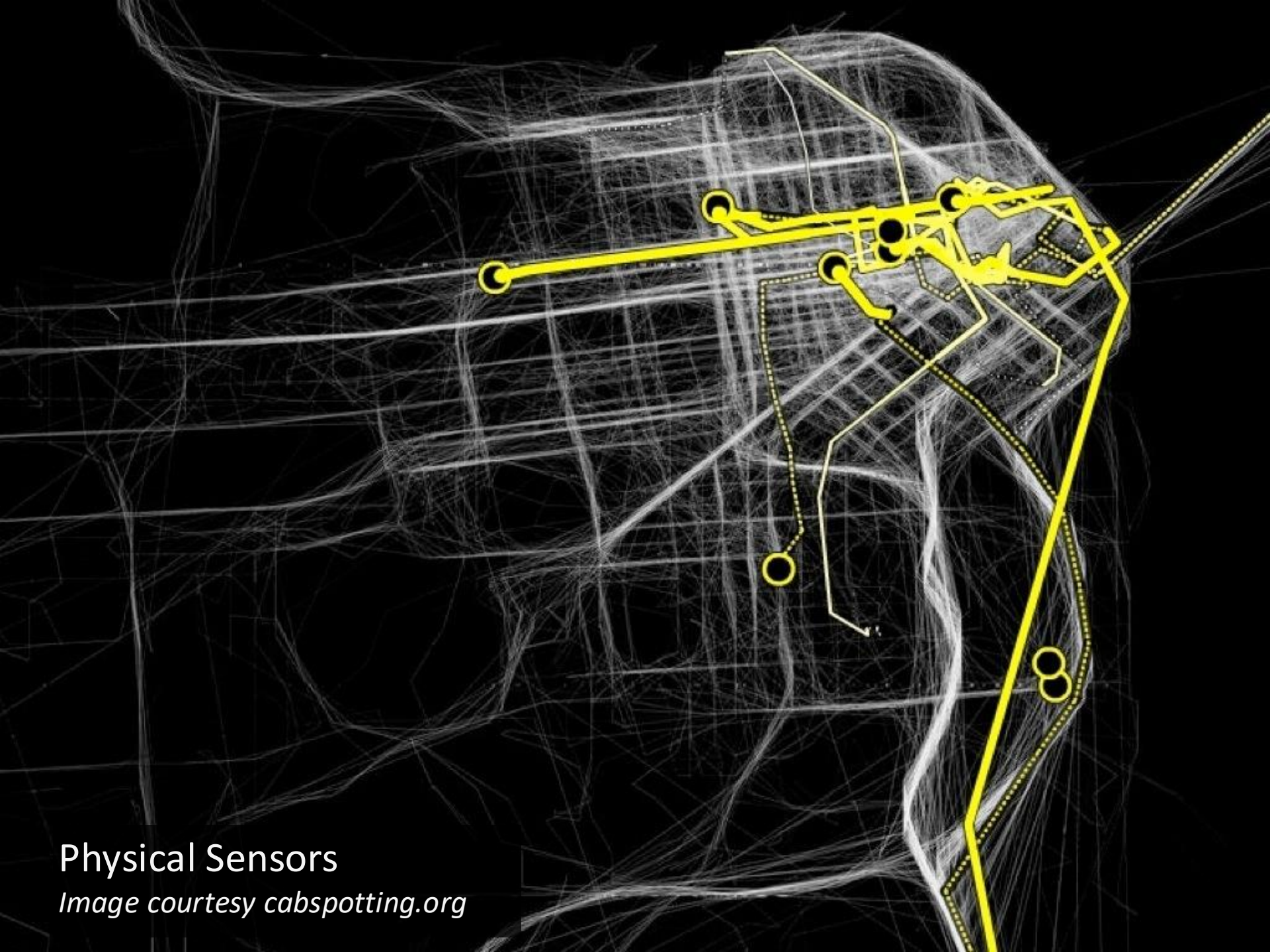
Data Created & Consumed
Source: IDC Digital Universe & Statista

But *what* is in all this data?



~2x every 2 years

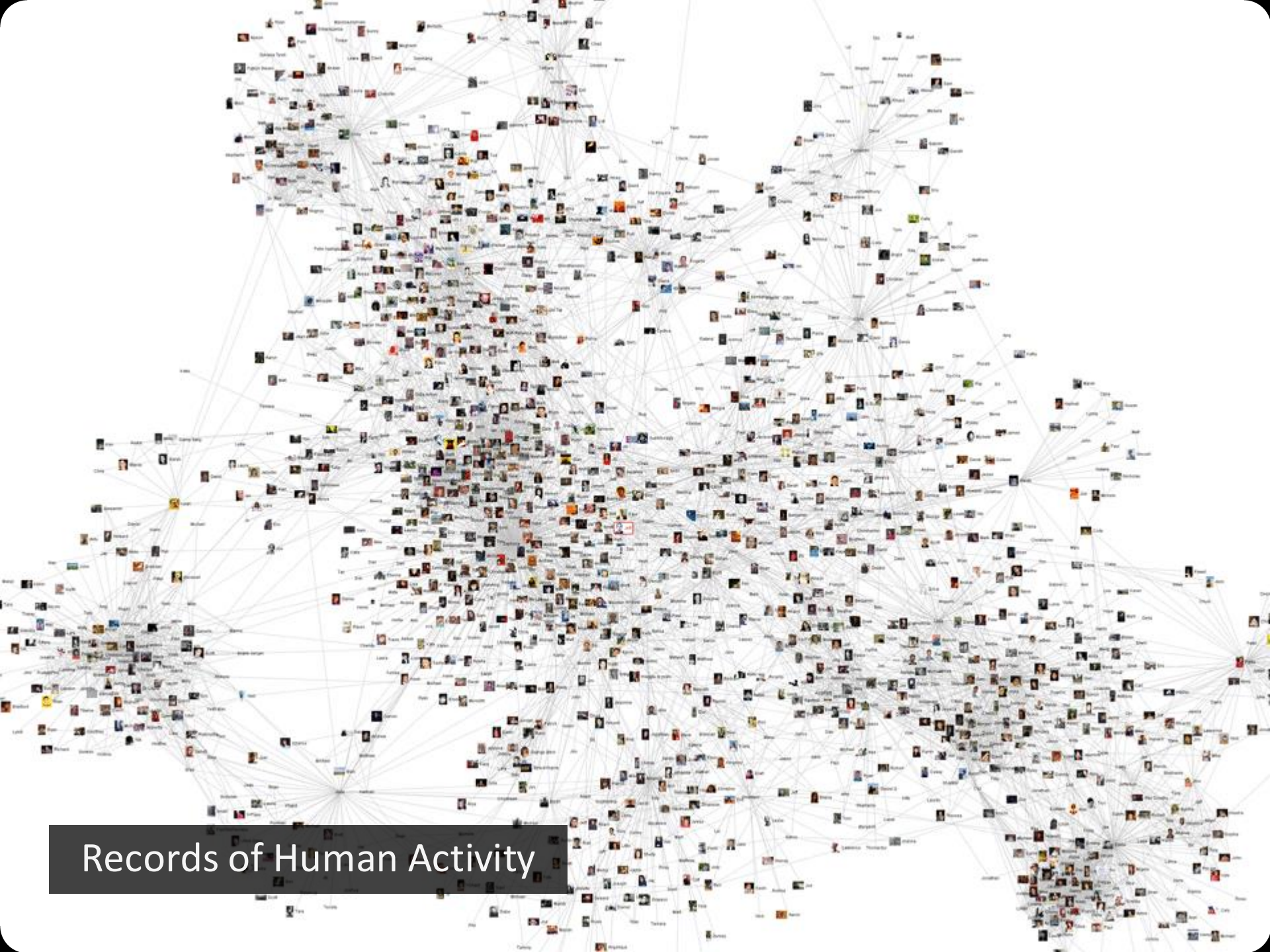
Data Created & Consumed
Source: IDC Digital Universe & Statista



Physical Sensors

Image courtesy cabspotting.org

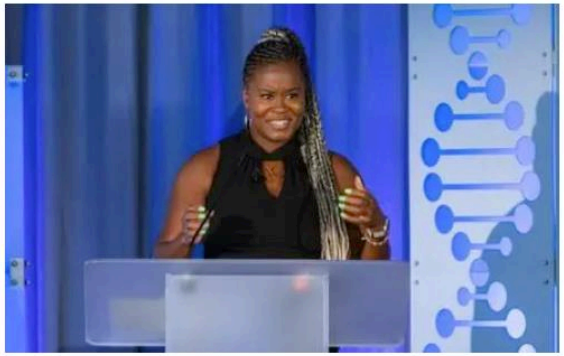
Health & Medicine


Records of Human Activity

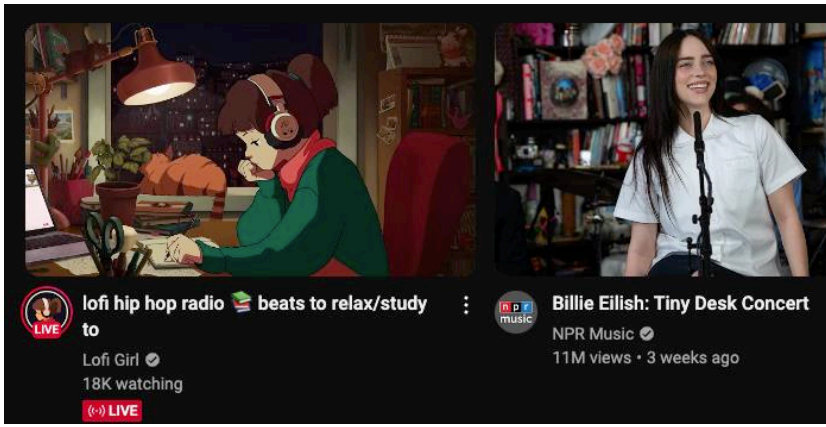
Advancing patient care with AI-driven tools

Meanwhile, Cherisse Marcou, Ph.D., co-director of the center's Digital Omics Program and co-director and vice chair of Information Technology and Bioinformatics in the Clinical Genomics Laboratory, is focused on bringing these advancements directly into patient care. Her team has created an artificial intelligence-driven tool that helps specialists simplify genomic test selection and accelerate genomic data analysis and interpretation.

Known as the Genetic Optimization and Appropriateness of Testing Tool, or GOATT, this innovation is used by laboratory experts to automate the summarization of important and relevant patient information across the electronic medical record into a concise format. Access to this comprehensive patient clinical summary helps ensure that the right test is



Cherisse Marcou, Ph.D., presents her research at the Individualizing Medicine Conference, Sept. 2024.



USAID
FROM THE AMERICAN PEOPLE

Data-Driven Agriculture: The Future of Smallholder Farmer Data Management

Home > Digital Agriculture > Documents > Data-Driven Agriculture: The Future of Smallholder Farmer Data Management

More than 500 million smallholder farms worldwide play a significant role in food production and the genetic diversity of the food supply. Until now, it has been difficult to get information to or from smallholder farmers, compounding basic infrastructural problems such as access to inputs, markets, financing, and training. The spread of mobile technology, remote-sensing data, and distributed computing and storage capabilities are opening new opportunities to integrate smallholder farmers into the broader agri-food system. The scale of these changes holds out the potential for another agricultural revolution.

As mobile technology use increases and improves in rural areas, the paradigm is also shifting for how smallholder farmers are profiled, how their needs are understood and met, how the impact of agricultural services is measured, how farmer data is shared, and how a global body of knowledge can be built by drawing on typically siloed expertise and data. To help describe

Home » Data and Reports

Environmental data

From daily weather forecasts, severe storm warnings, and climate monitoring to fisheries management, coastal restoration and supporting marine commerce, NOAA's products and services support economic vitality and affect more than one-third of America's gross domestic product.

NOAA's dedicated scientists use cutting-edge research and high-tech instrumentation to provide citizens, planners, emergency managers and other decision makers with reliable information they need when they need it.

Life-size cutouts of Facebook CEO Mark Zuckerberg are displayed by a progressive advocacy group on the lawn of the U.S. Capitol on Tuesday.

Carolyn Kaster / Reuters

My Facebook Was Breached by Cambridge Analytica. Was Yours?

How to find out if you are one of the 87 million victims

ROBINSON MEYER | APR 10, 2018 | TECHNOLOGY

Share Tweet ...

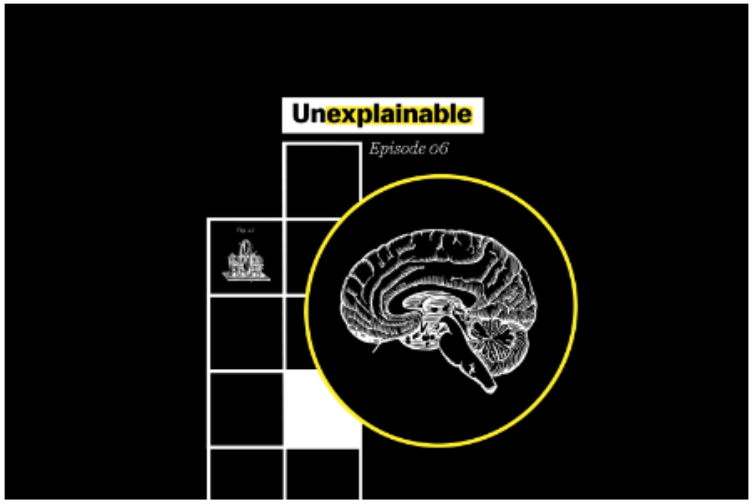
TEXT SIZE − +

SCIENCE

The replication crisis devastated psychology. This group is looking to rebuild it.

The Psychological Science Accelerator could be the future of the field around the globe — if they can sustain it.

by **Brian Resnick**
Updated Apr 7, 2021, 3:03 PM PDT



Vox

High potential for data abuse…

Inequality

Rise of the racist robots – how AI is learning all our worst impulses

There is a saying in computer science: garbage in, garbage out. When we feed machines data that reflects our prejudices, they mimic them – from antisemitic chatbots to racially biased software. Does a horrifying future await people forced to live at the mercy of algorithms?

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 28 August 2024

AI generates covertly racist decisions about people based on their dialect

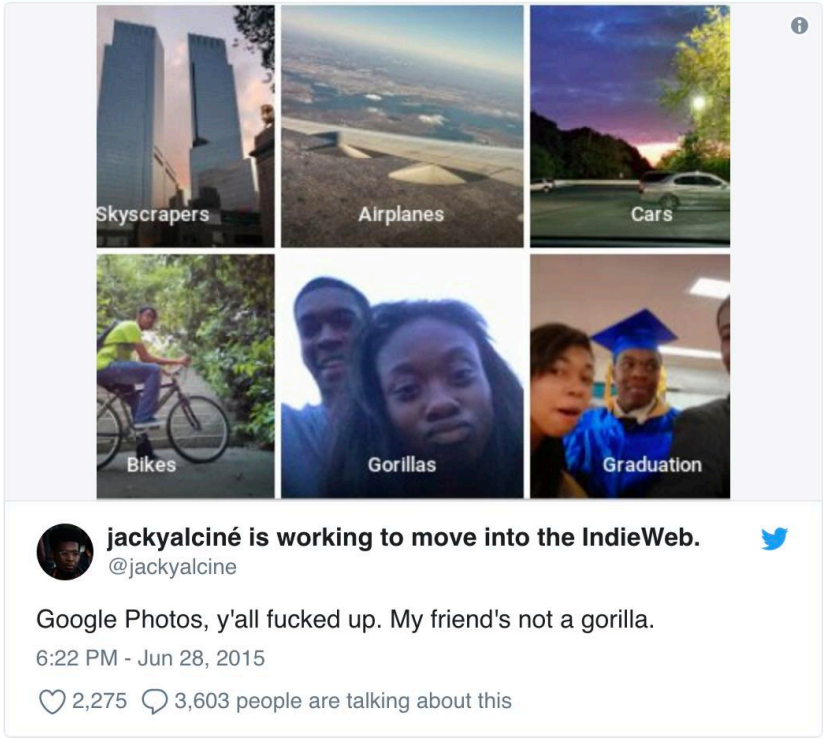
Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky & Sharese King

Nature **633**, 147–154 (2024) | Cite this article

78k Accesses | **5** Citations | **418** Altmetric | Metrics

Abstract

Hundreds of millions of people now interact with langu
ranging from help with writing[1,2] to informing hiring de
language models are known to perpetuate systematic r
their judgements biased in problematic ways about gro
Americans[4,5,6,7]. Although previous research has focus



…amplified by “big data” and ML systems.

We move from data to information to knowledge to wisdom, and separating one from the other, being able to distinguish among and between them that is, knowing the limitations and the danger of exercising one without the others while respecting each category of intelligence, is generally what serious education is about.

Toni Morrison, American Novelist

The Source of Self Regard

How might we use **visualization** to **empower understanding** of data and analysis processes?

What is Visualization?

"Transformation of the symbolic into the geometric"
[McCormick et al. 1987]

"... finding the artificial memory that best supports our natural means of perception." [Bertin 1967]

"The use of computer-generated, interactive, visual representations of data to amplify cognition."
[Card, Mackinlay, & Shneiderman 1999]

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.32$

$u_Y = 7.5$ $\sigma_Y = 2.03$

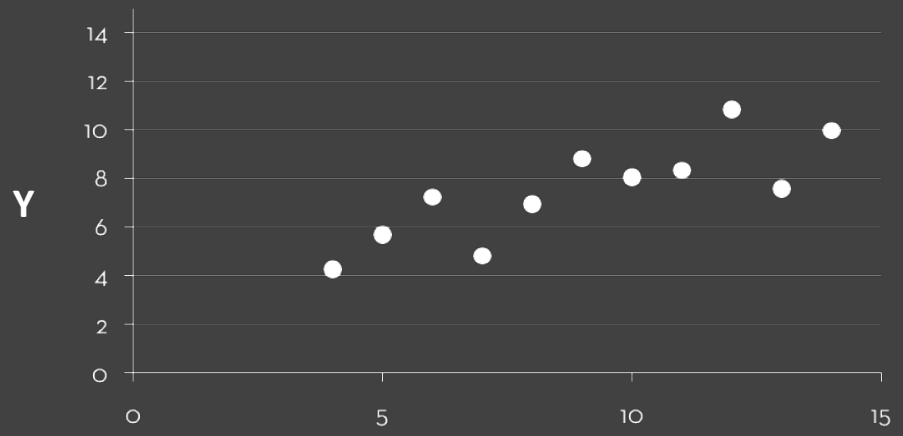
Linear Regression

$Y = 3 + 0.5\ X$

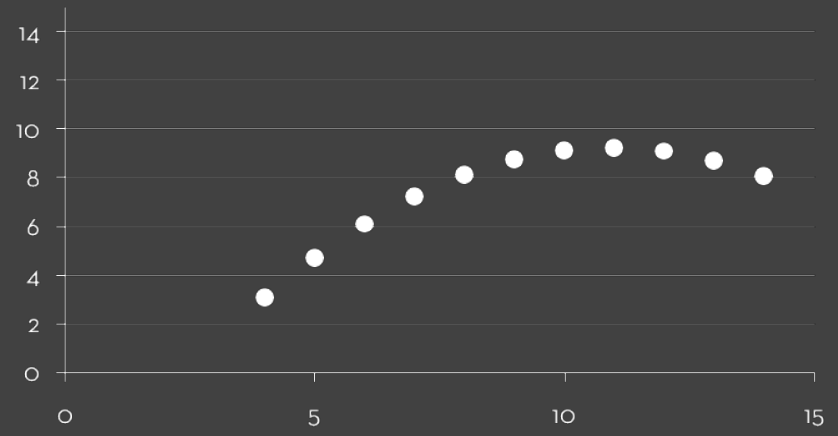
$R^2 = 0.67$

[Anscombe 1973]

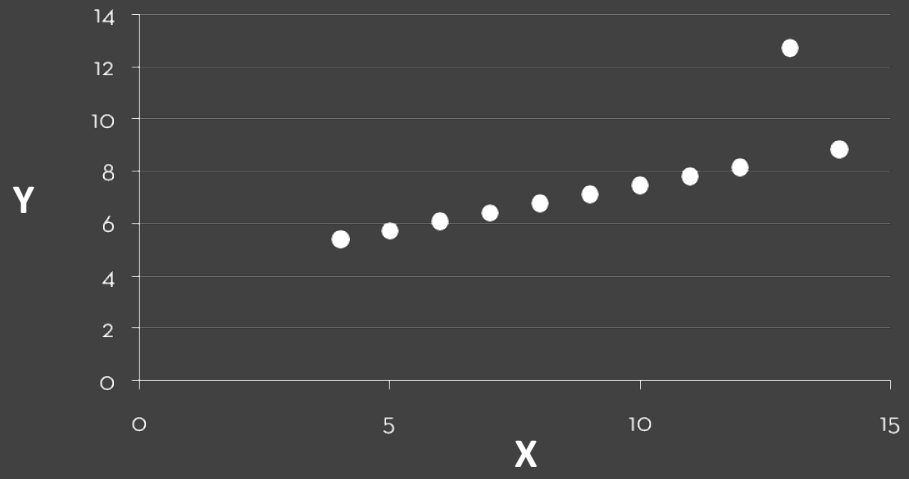
Set A



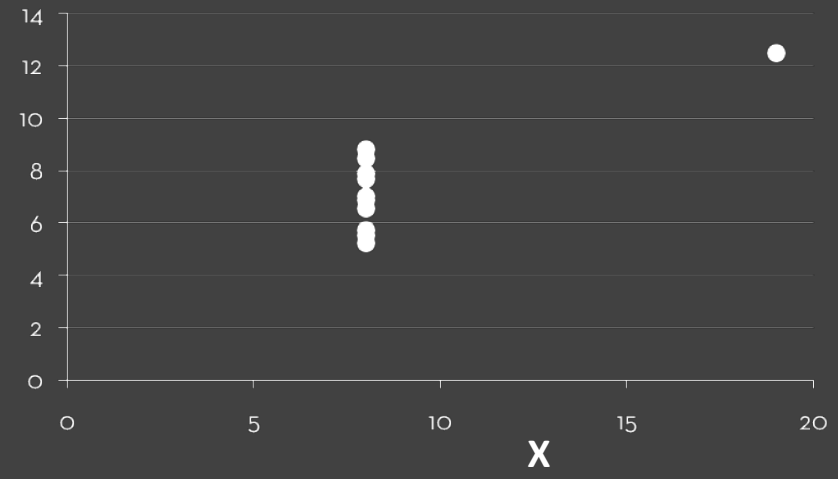
Set B



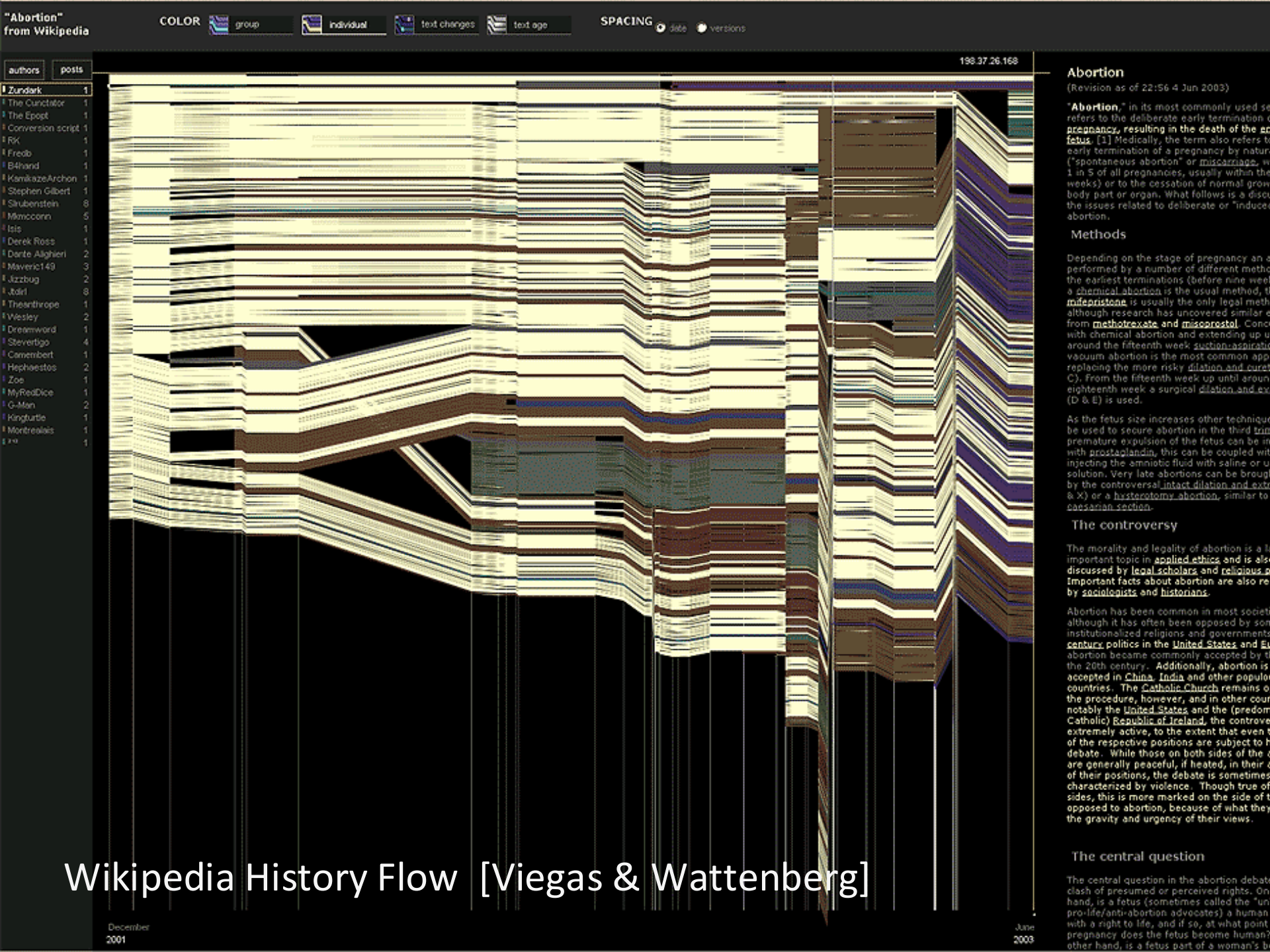
Set C



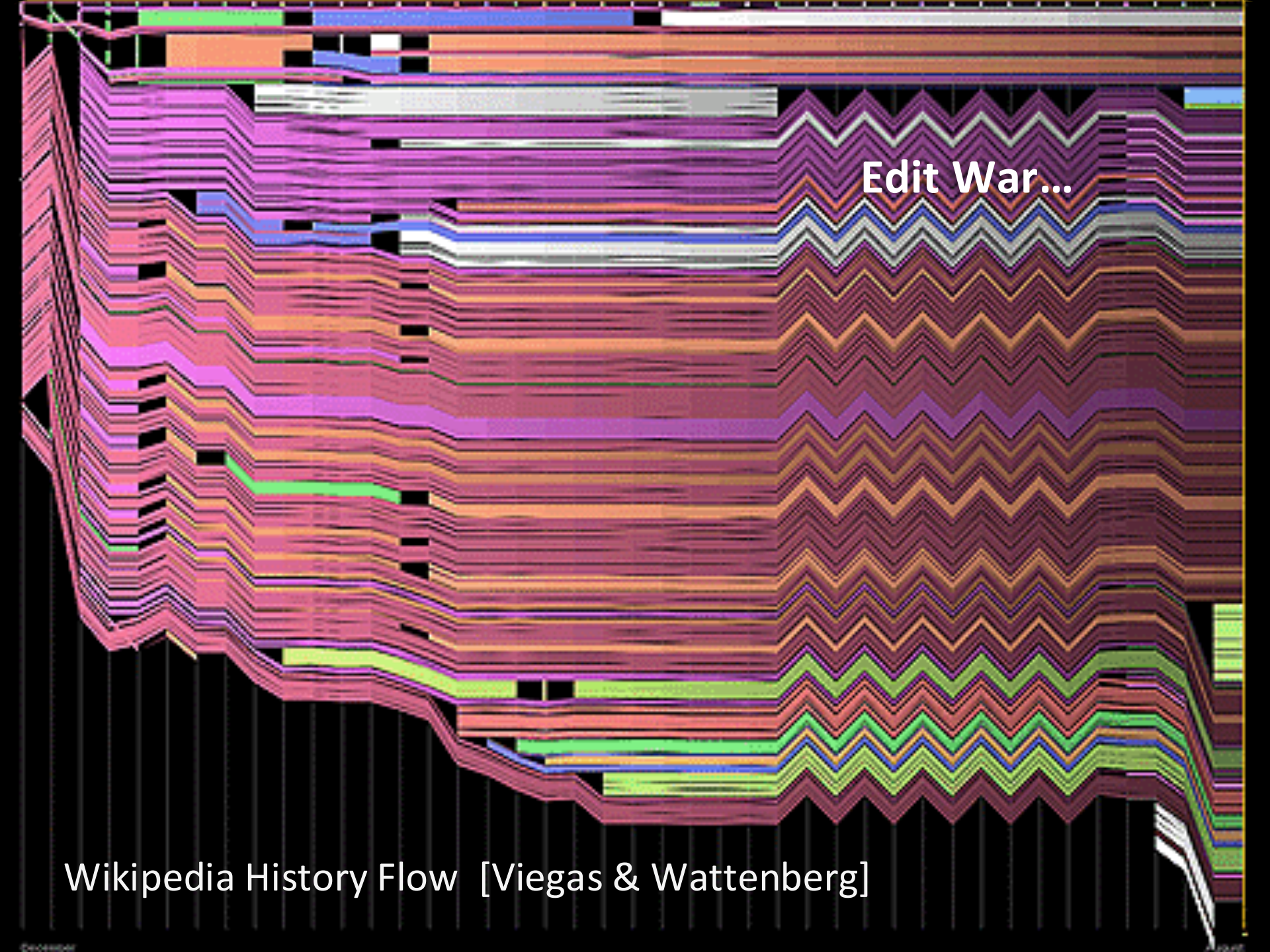
Set D



[Anscombe 1973]



Wikipedia History Flow [Viegas & Wattenberg]



Wikipedia History Flow [Viegas & Wattenberg]

Why Create Visualizations?

Why Create Visualizations?

Answer questions (or discover them)
Make decisions
See data in context
Expand memory
Support graphical calculation
Find patterns
Present argument or tell a story
Inspire

The Value of Visualization

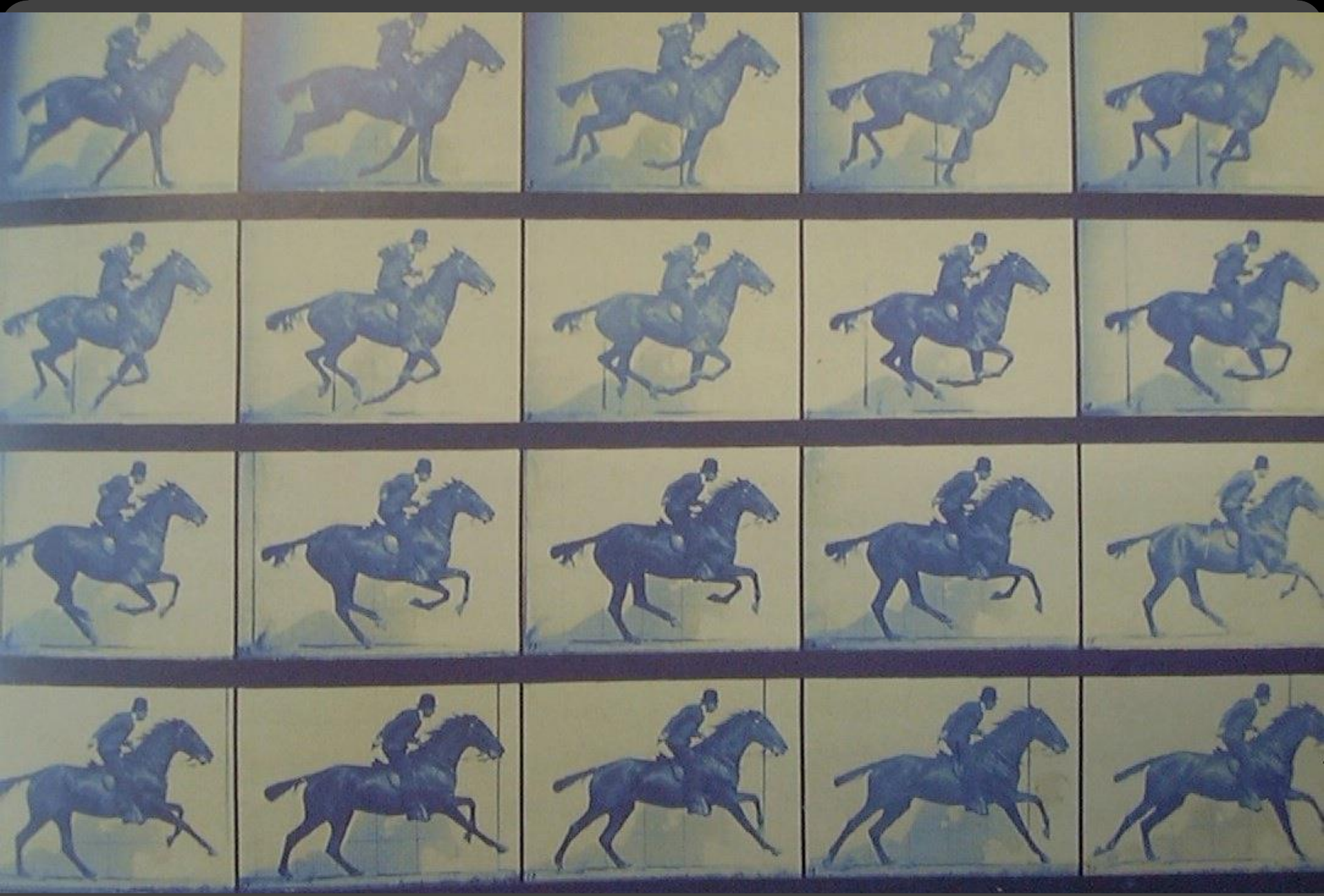
Record information

Analyze data to support reasoning

Convey information

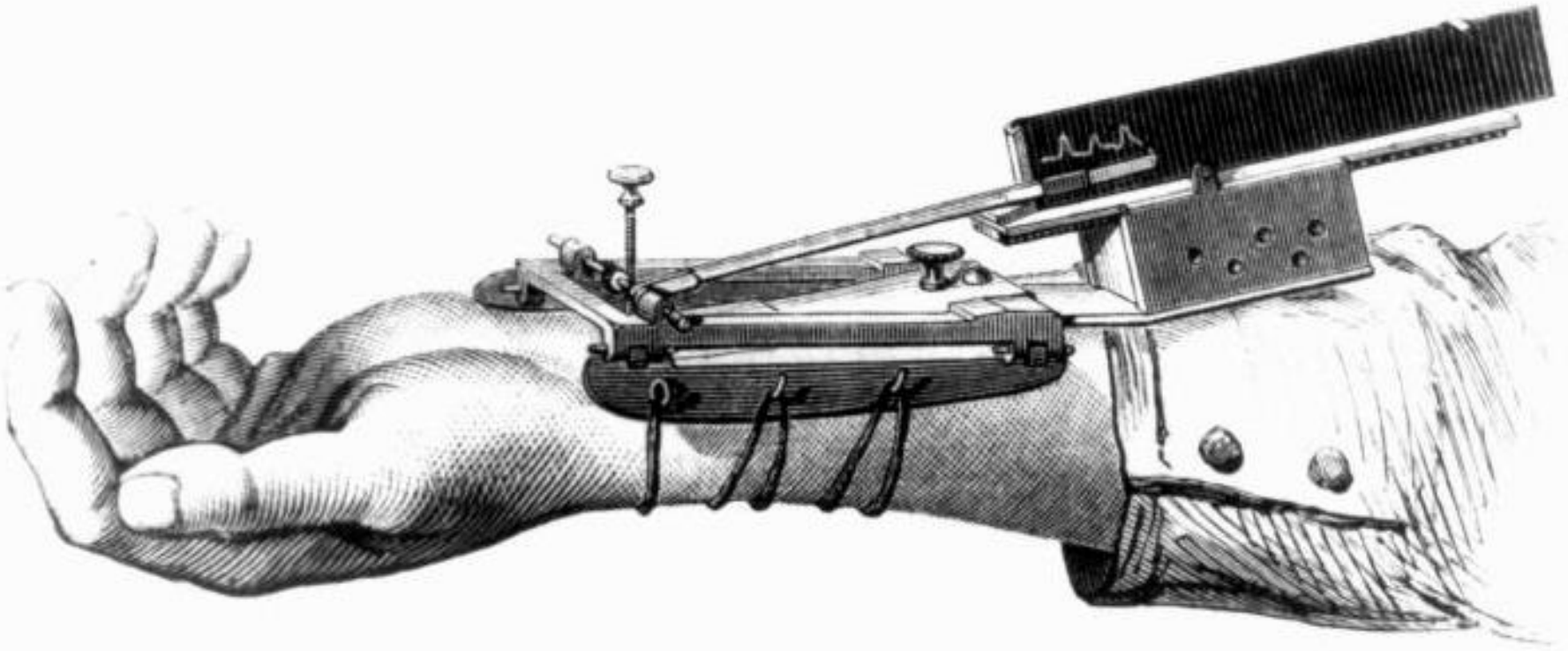
Record Information

https://www.flickr.com/photos/brittgow/4782591644



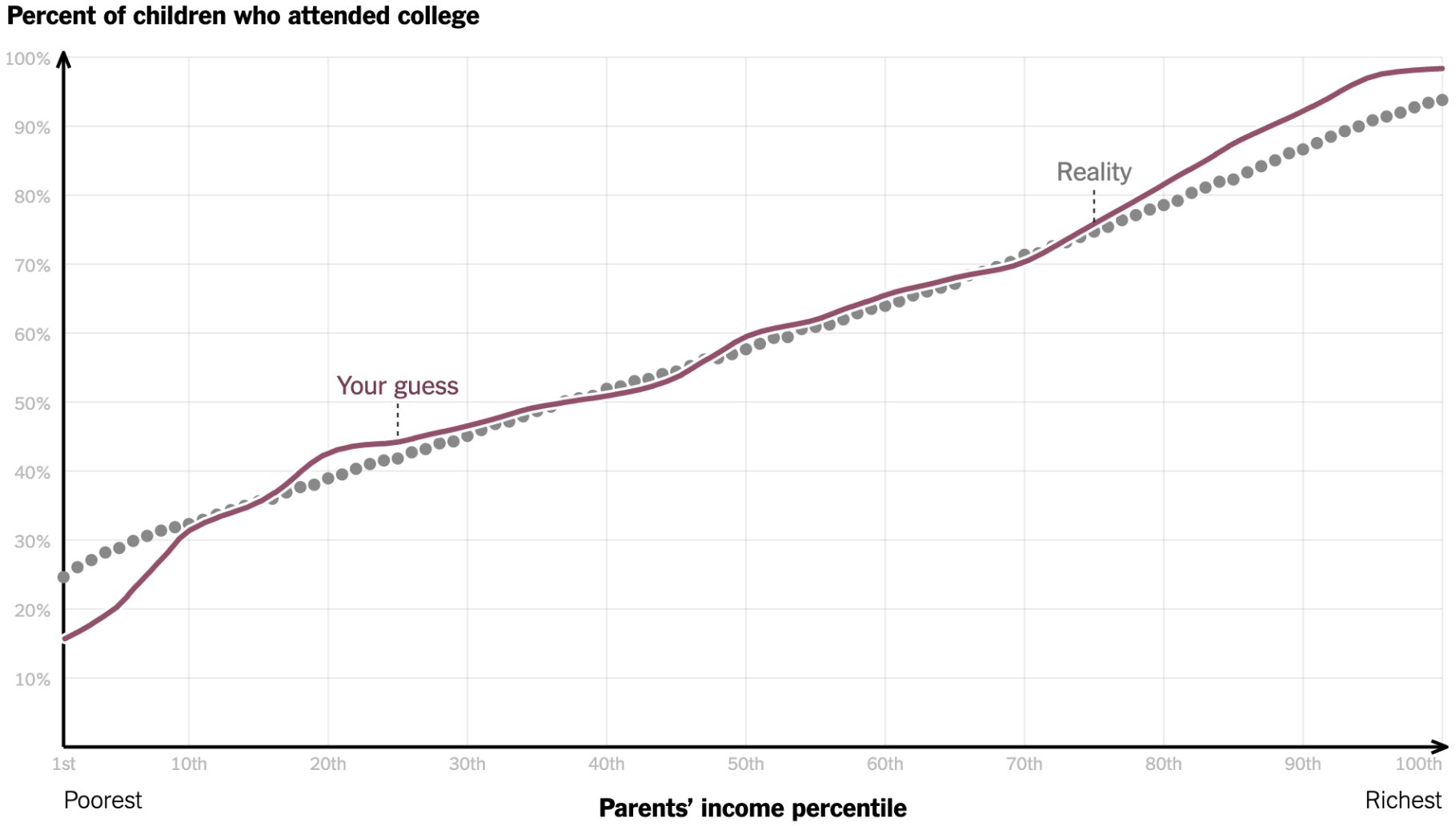
Gallop, Bay Horse “Daisy” [Muybridge]

Frederick Douglass. Photograph. Retrieved from the Library of Congress, <www.loc.gov/item/2017895330/>

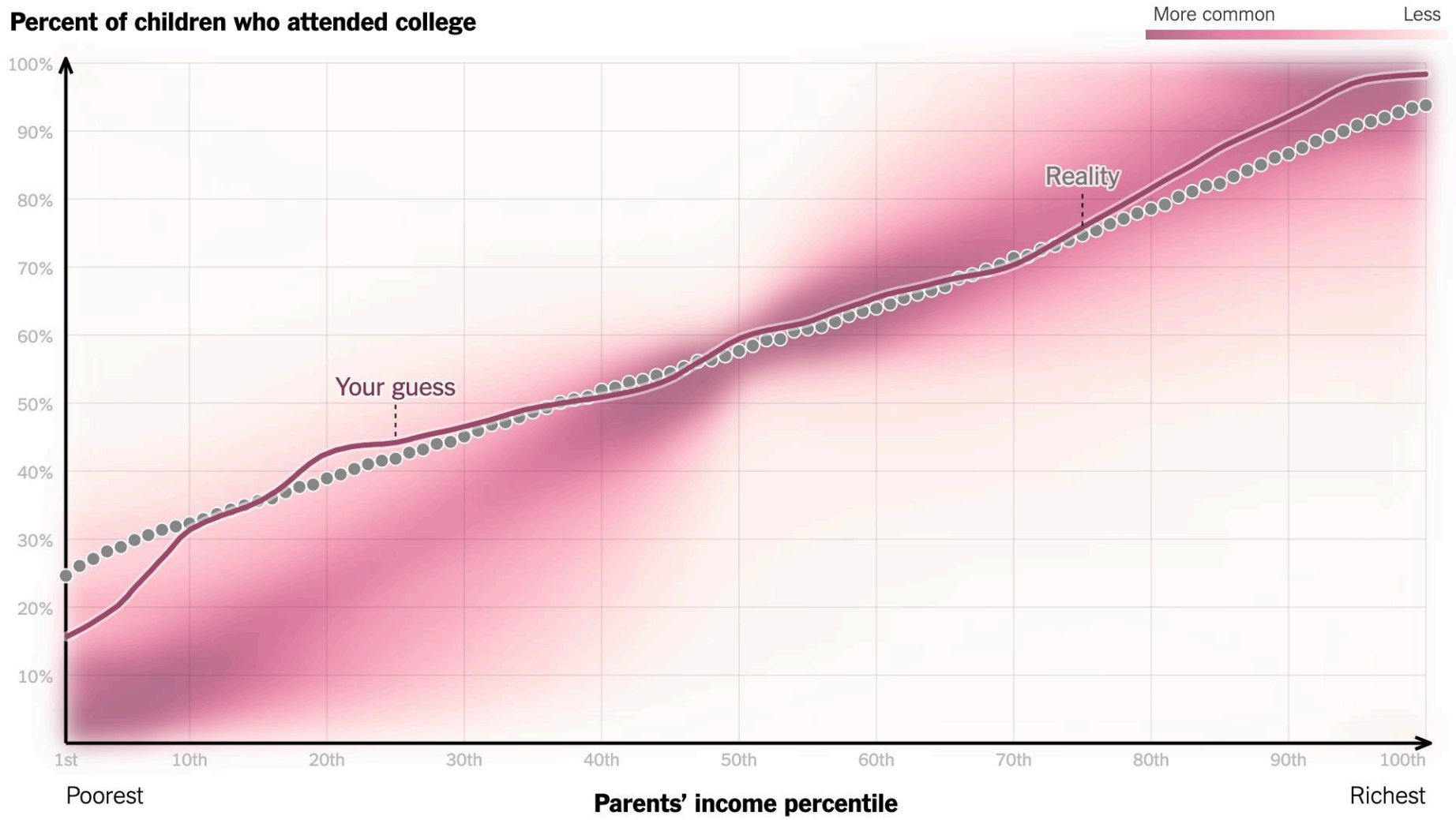


1.
Marey's sphygmograph in use, 1860. *La méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine.*

E.J. Marey's sphygmograph [from Braun 83]



You Draw It: How Family Income Predicts Children's College Chances
[New York Times, May 28, 2015]



You Draw It: How Family Income Predicts Children's College Chances
[New York Times, May 28, 2015]

Support Reasoning

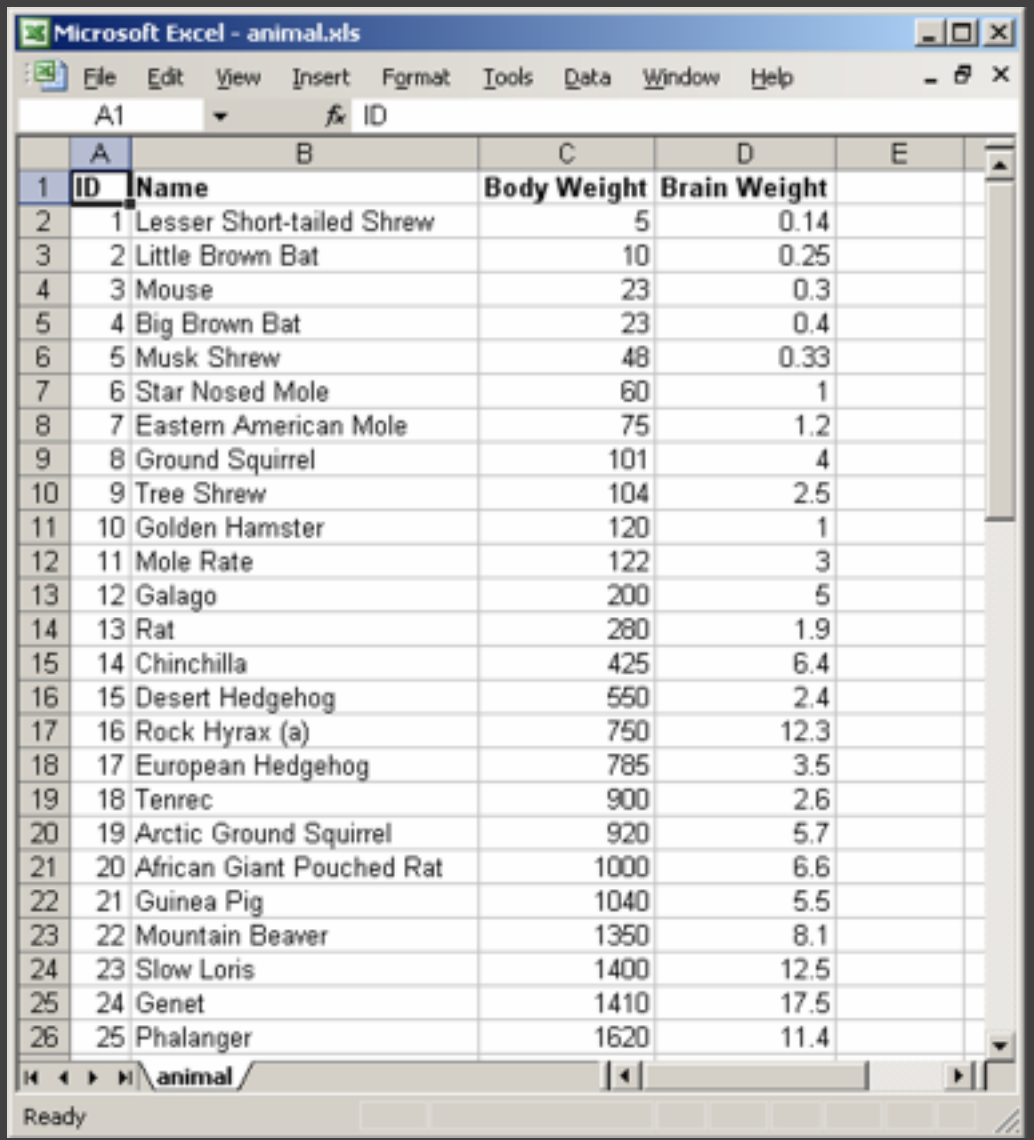
Data in Context: Cholera Outbreak

In 1854 John Snow plotted the position of each cholera case on a map. [from Tufte 83]

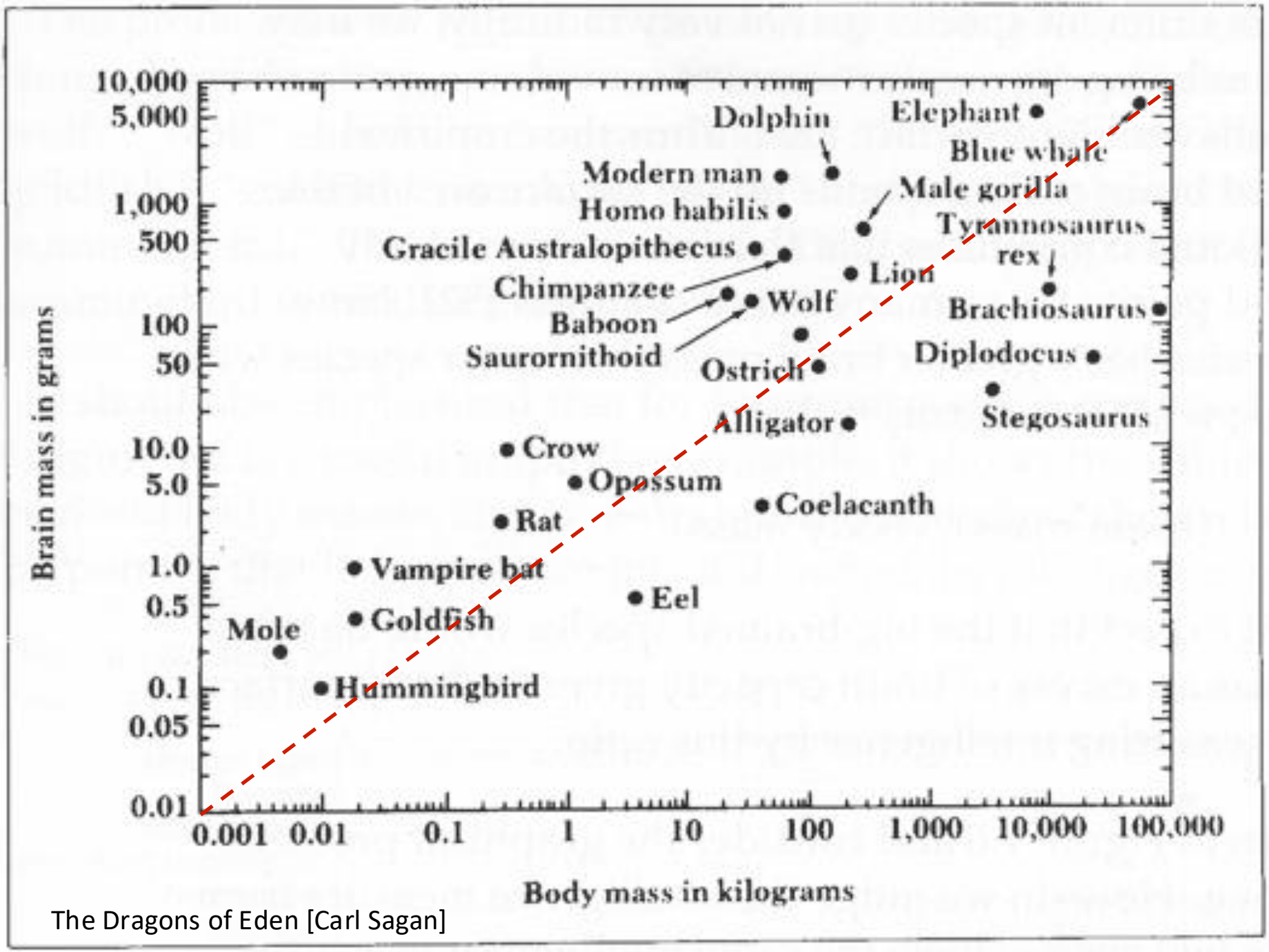
Data in Context: Cholera Outbreak

Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

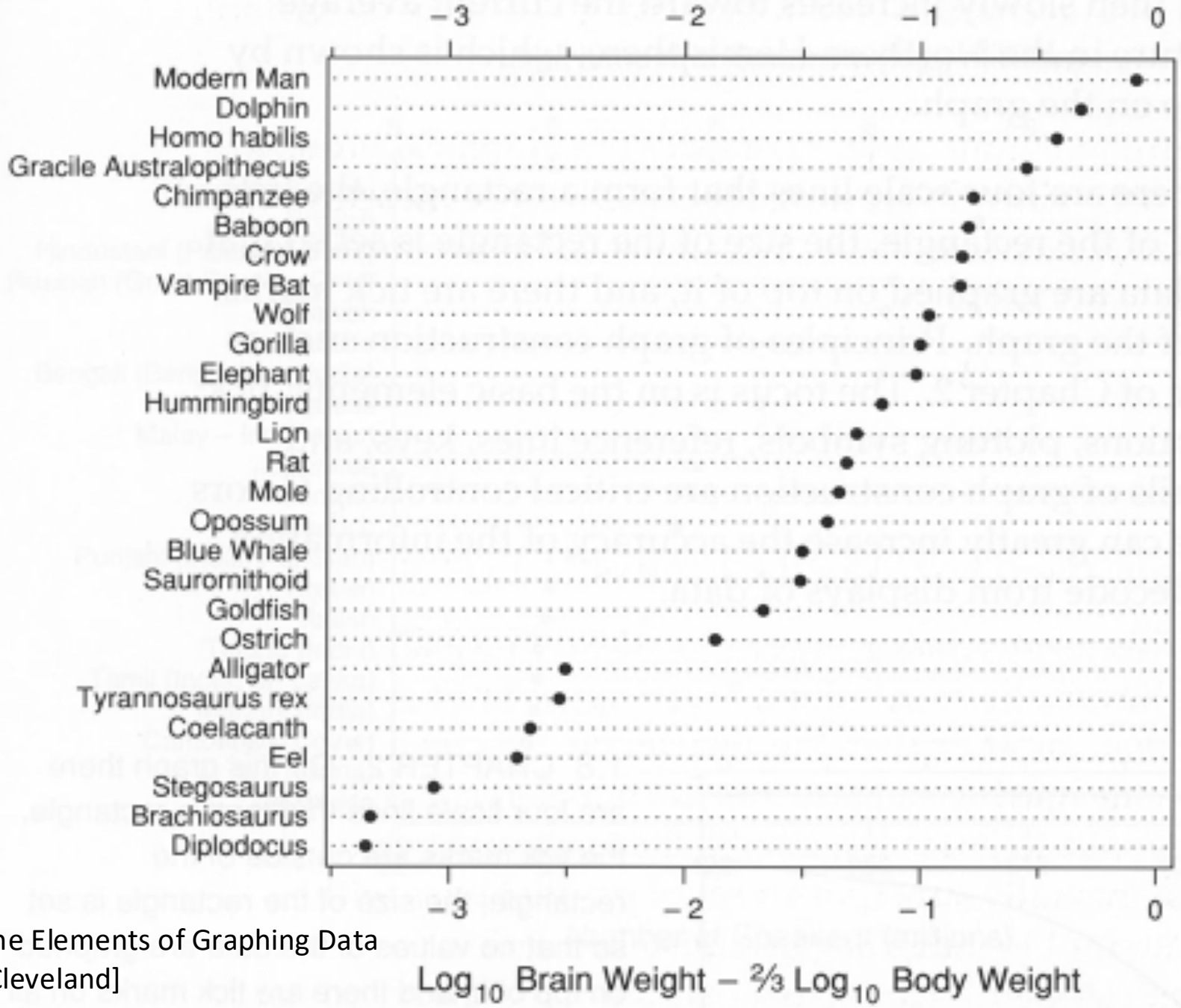
Answer Questions: Brain Power?



	A	B	C	D	E
1	ID	Name	Body Weight	Brain Weight	
2	1	Lesser Short-tailed Shrew	5	0.14	
3	2	Little Brown Bat	10	0.25	
4	3	Mouse	23	0.3	
5	4	Big Brown Bat	23	0.4	
6	5	Musk Shrew	48	0.33	
7	6	Star Nosed Mole	60	1	
8	7	Eastern American Mole	75	1.2	
9	8	Ground Squirrel	101	4	
10	9	Tree Shrew	104	2.5	
11	10	Golden Hamster	120	1	
12	11	Mole Rate	122	3	
13	12	Galago	200	5	
14	13	Rat	280	1.9	
15	14	Chinchilla	425	6.4	
16	15	Desert Hedgehog	550	2.4	
17	16	Rock Hyrax (a)	750	12.3	
18	17	European Hedgehog	785	3.5	
19	18	Tenrec	900	2.6	
20	19	Arctic Ground Squirrel	920	5.7	
21	20	African Giant Pouched Rat	1000	6.6	
22	21	Guinea Pig	1040	5.5	
23	22	Mountain Beaver	1350	8.1	
24	23	Slow Loris	1400	12.5	
25	24	Genet	1410	17.5	
26	25	Phalanger	1620	11.4	

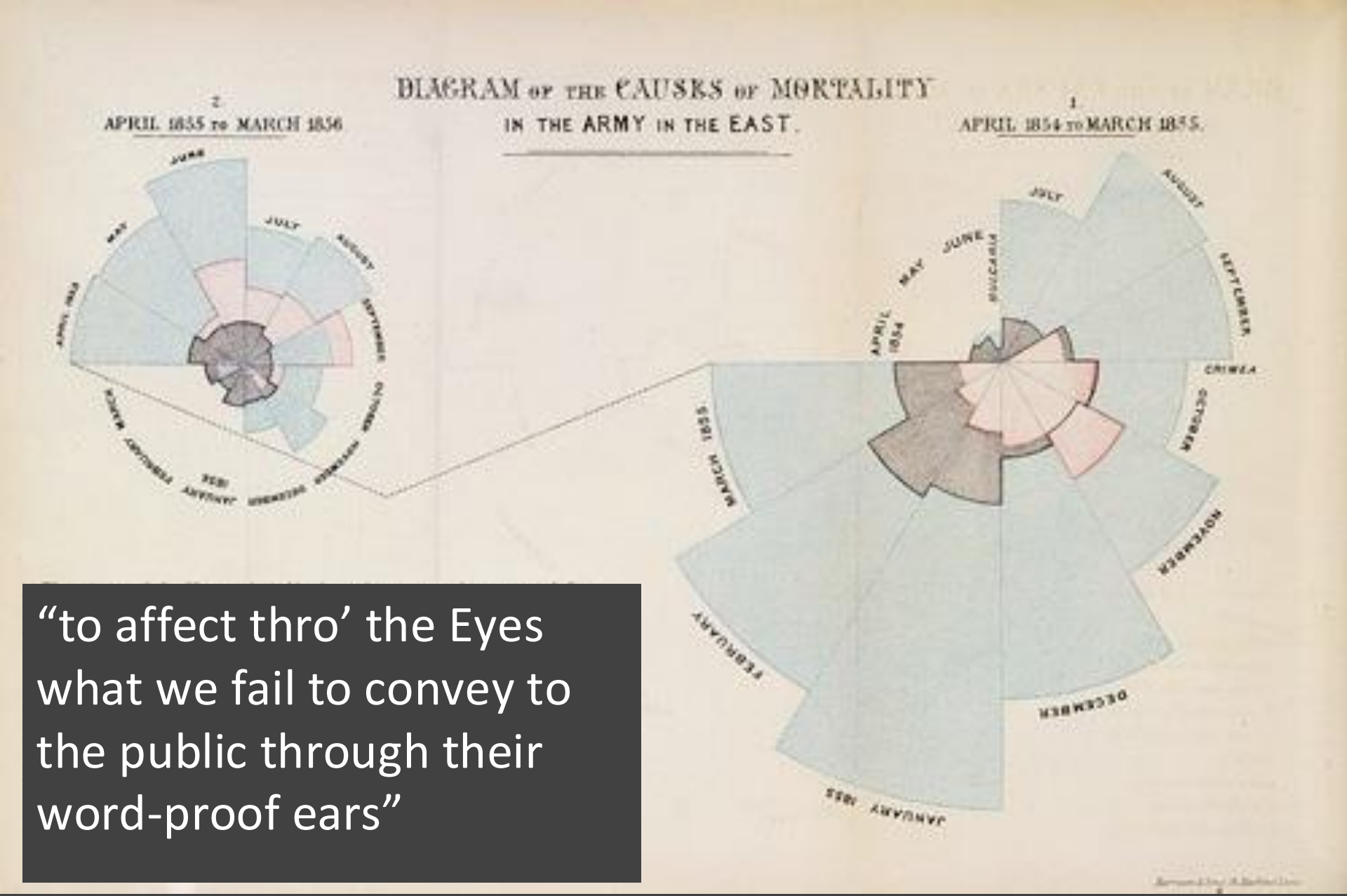


The Dragons of Eden [Carl Sagan]



The Elements of Graphing Data
[Cleveland]

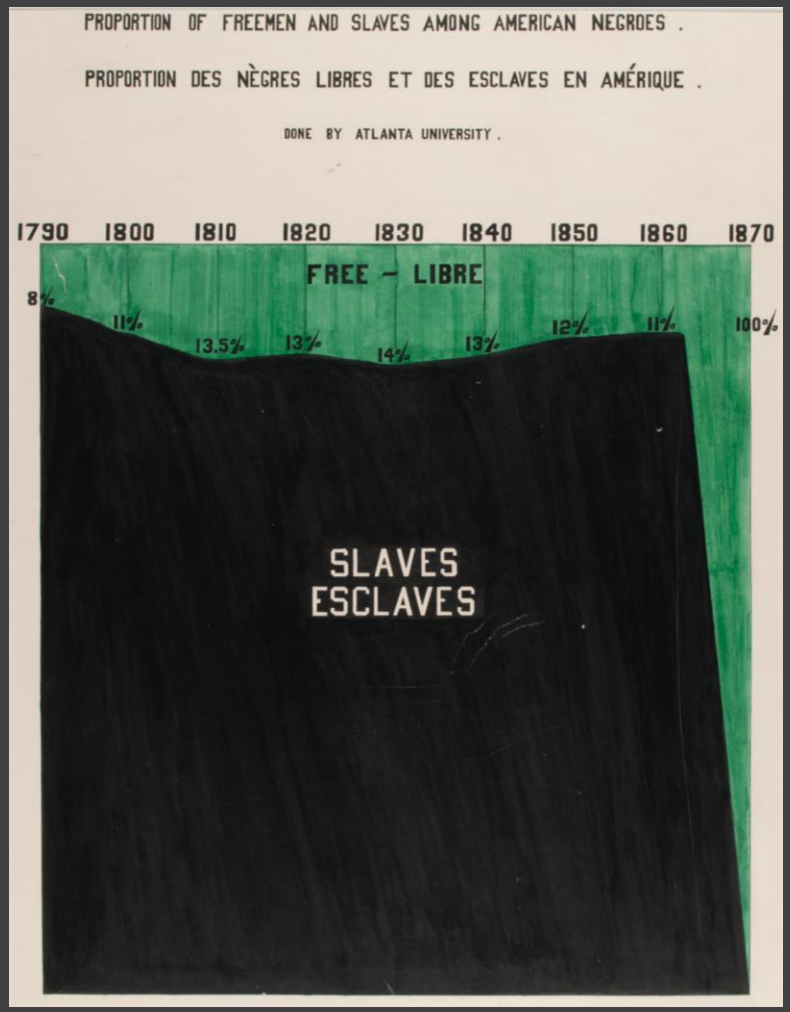
Convey Information



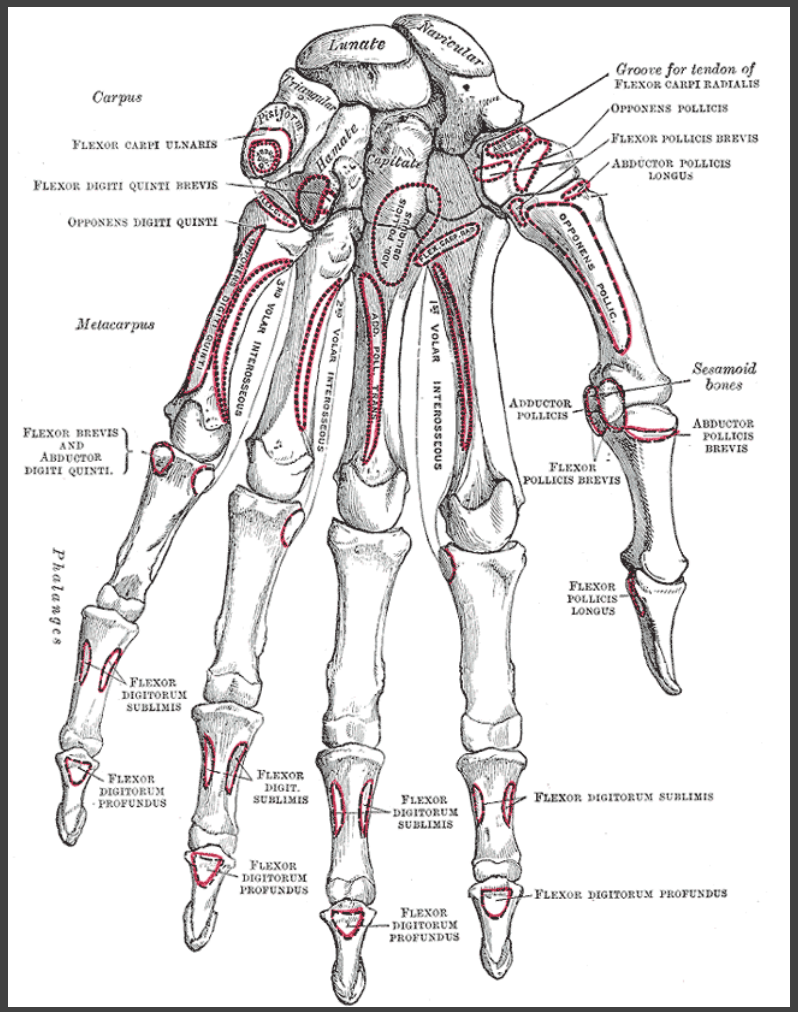
"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

Communicate, Inform, Inspire



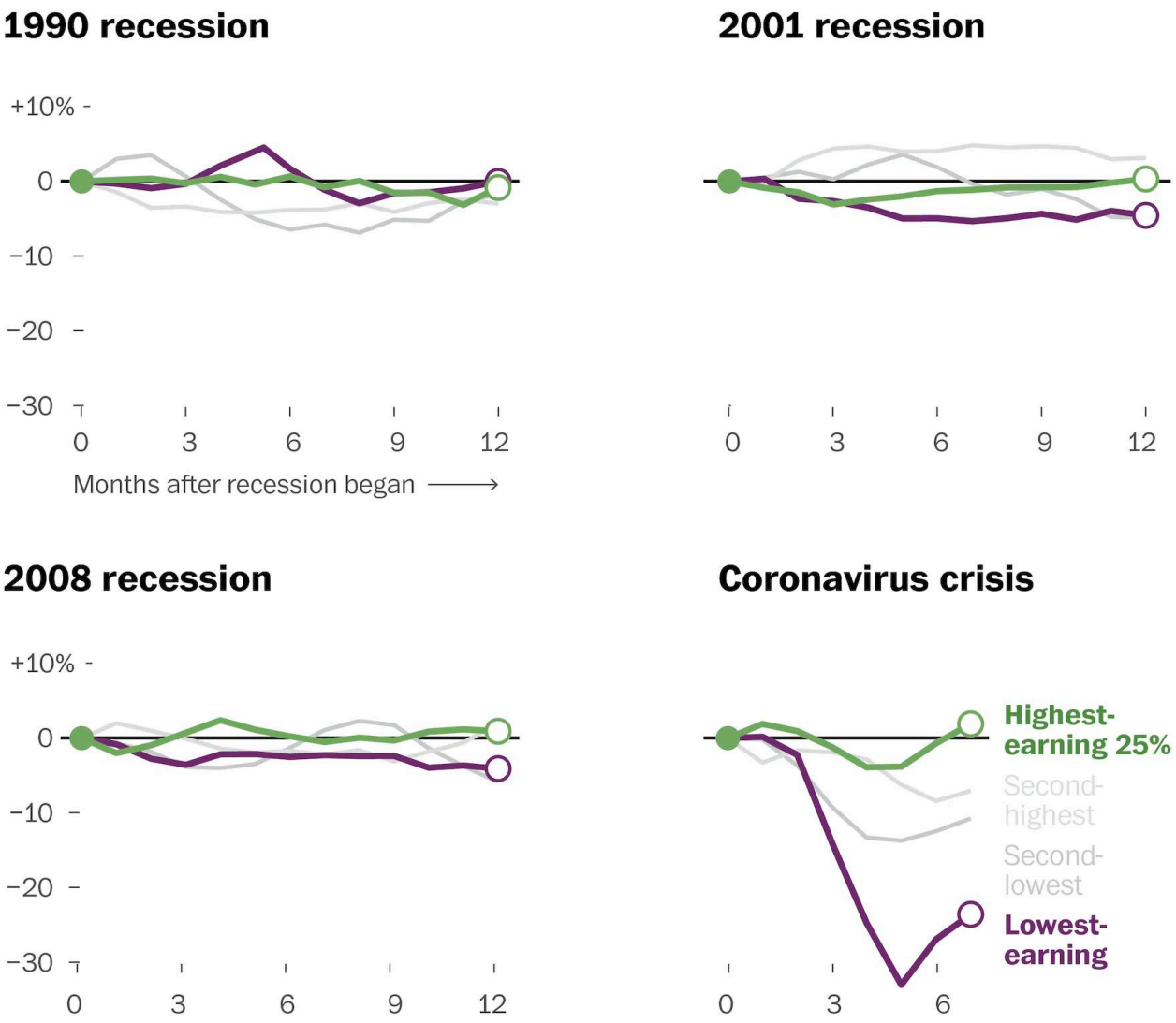
Visualizing Black America, Du Bois et al. 1900



Bones in hand, Gray's Anatomy 1918 ed.

The coronavirus crisis is different

Job growth (or loss) since each recession began, based on weekly earnings



Notes: Based on a three-month average to show the trend in volatile data.

Source: Labor Department via IPUMS, with methodology assistance from Ernie Tedeschi of Evercore ISI

THE WASHINGTON POST

The Covid Economy
Washington Post

The Value of Visualization

Record information

- Blueprints, photographs, seismographs, …

Analyze data to support reasoning

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Convey information

- Communicate, inform, inspire
- Collaborate and revise

Goals of Visualization Research

1 **Understand** how visualizations convey information

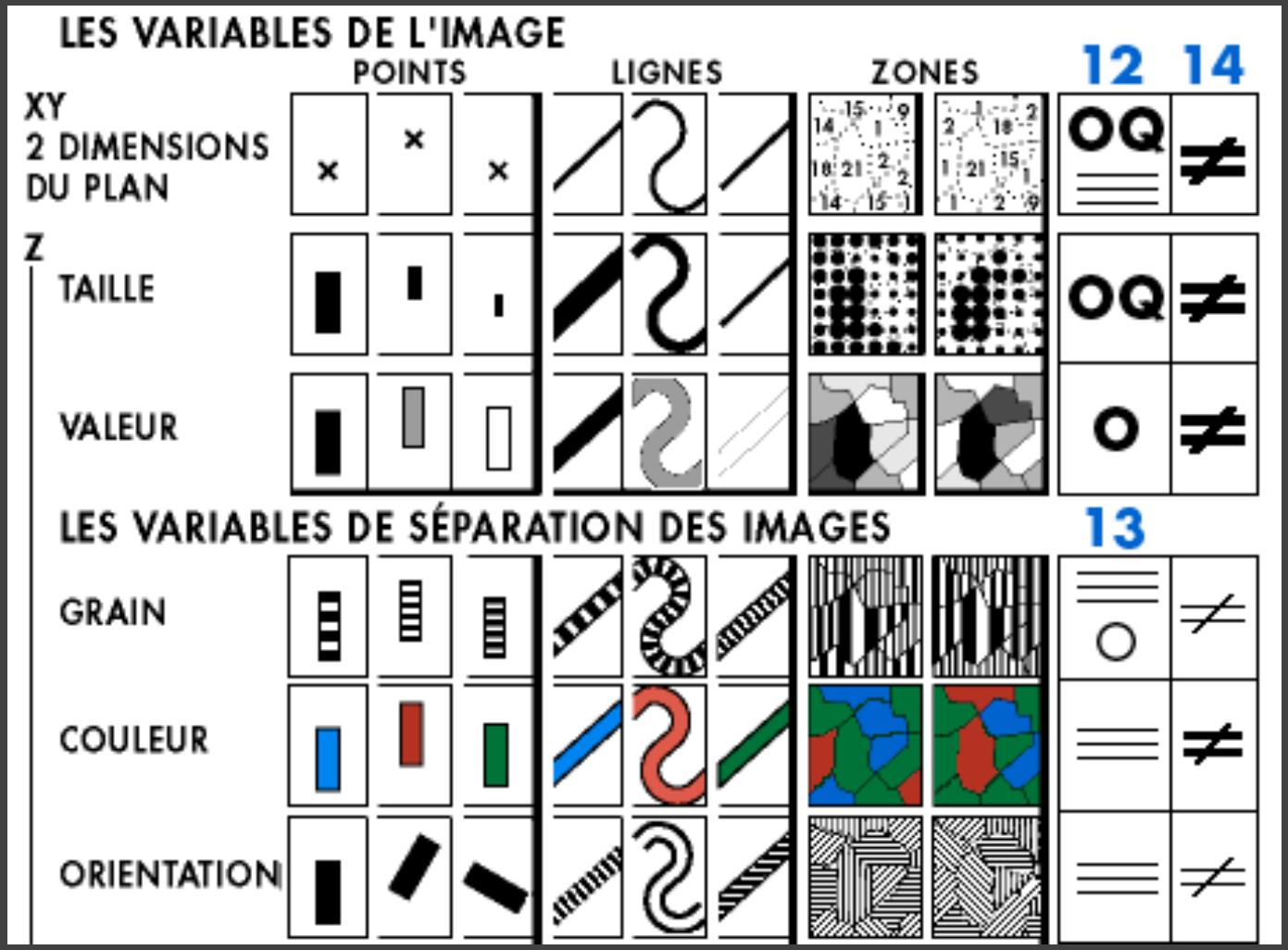
What do people perceive / comprehend?
How do visualizations inform mental models?

2 **Develop principles and techniques** for creating effective visualizations and supporting analysis

Leverage perception & augment cognition
Improve ties between visualization & mental model

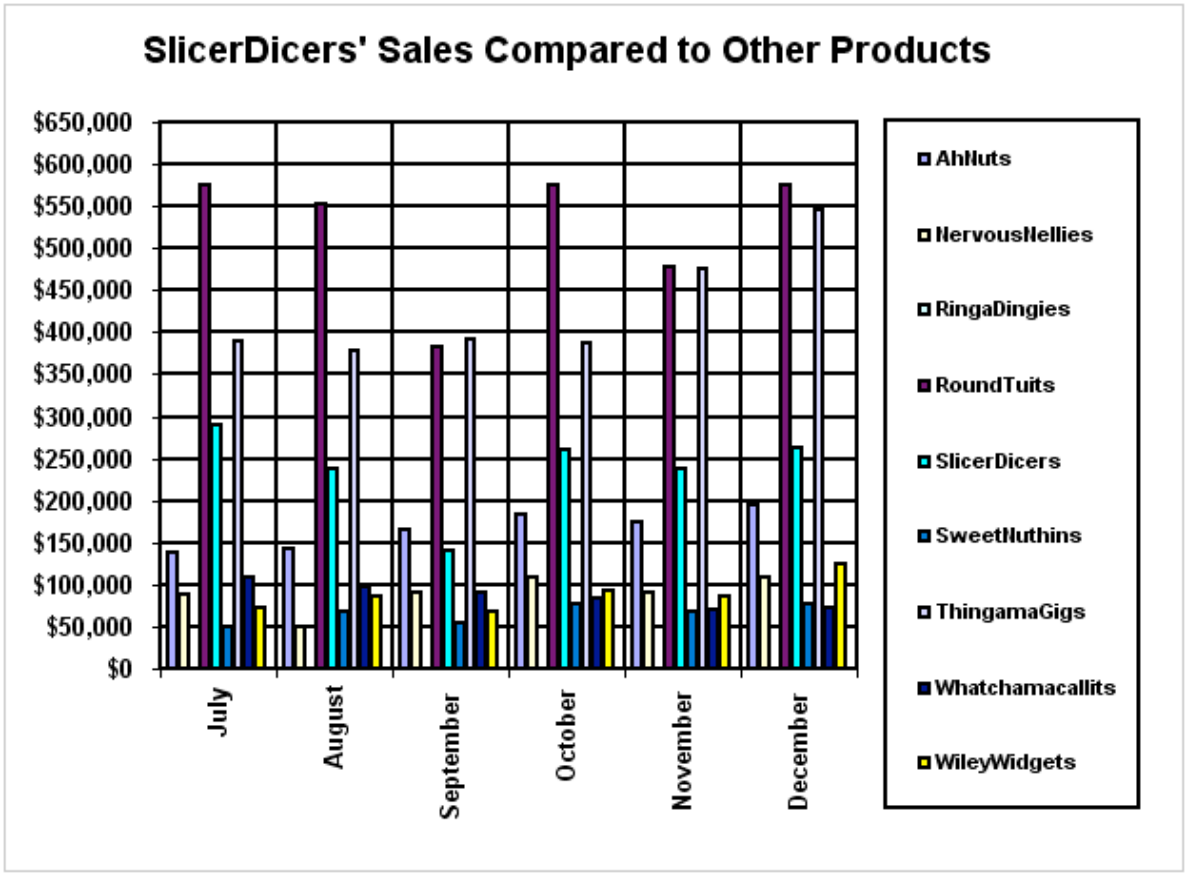
Course Topics

Data and Image Models

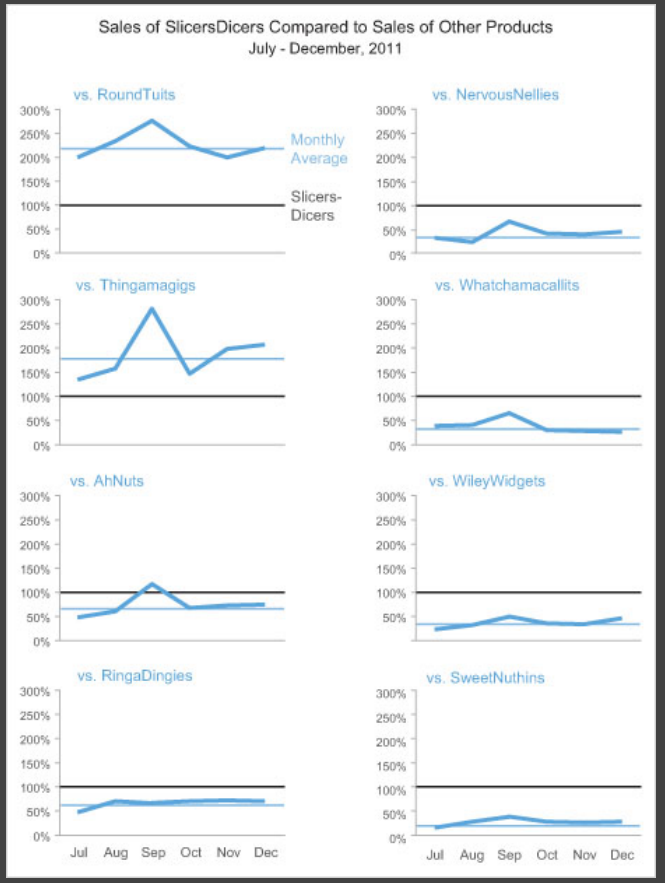


Sémiologie Graphique [Bertin 67]

Visualization Design

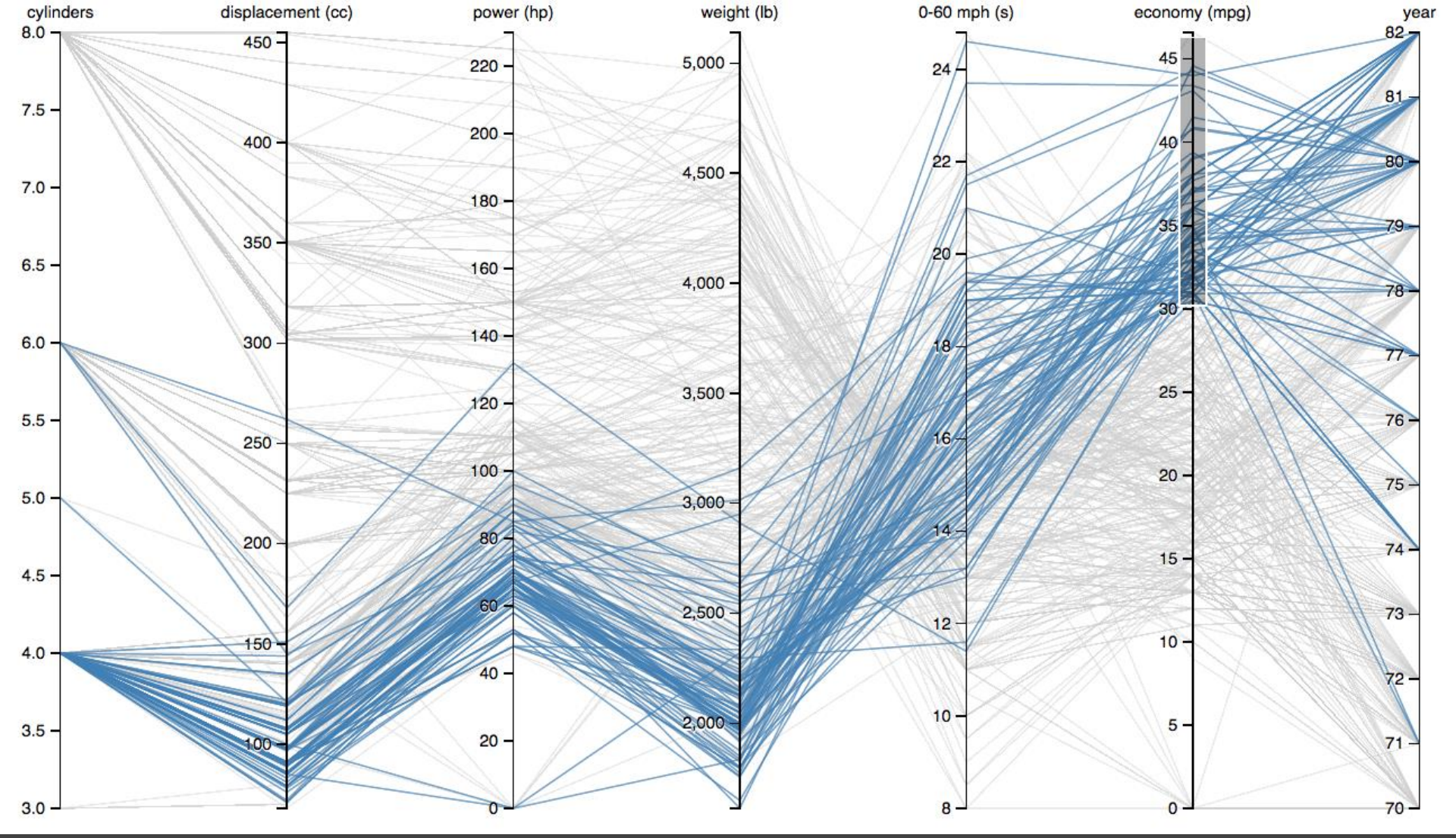


Problematic design

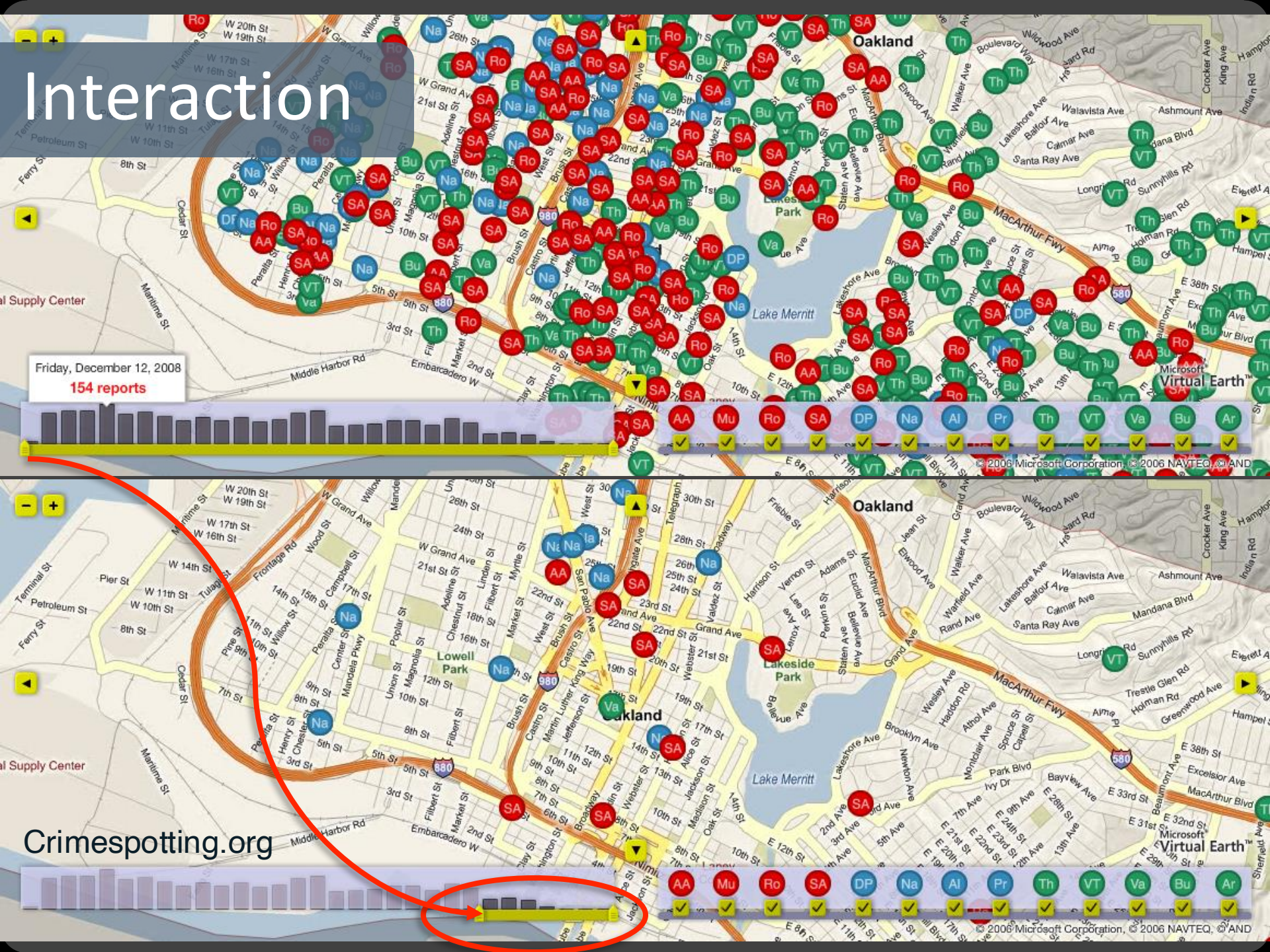


Redesign

Exploratory Data Analysis

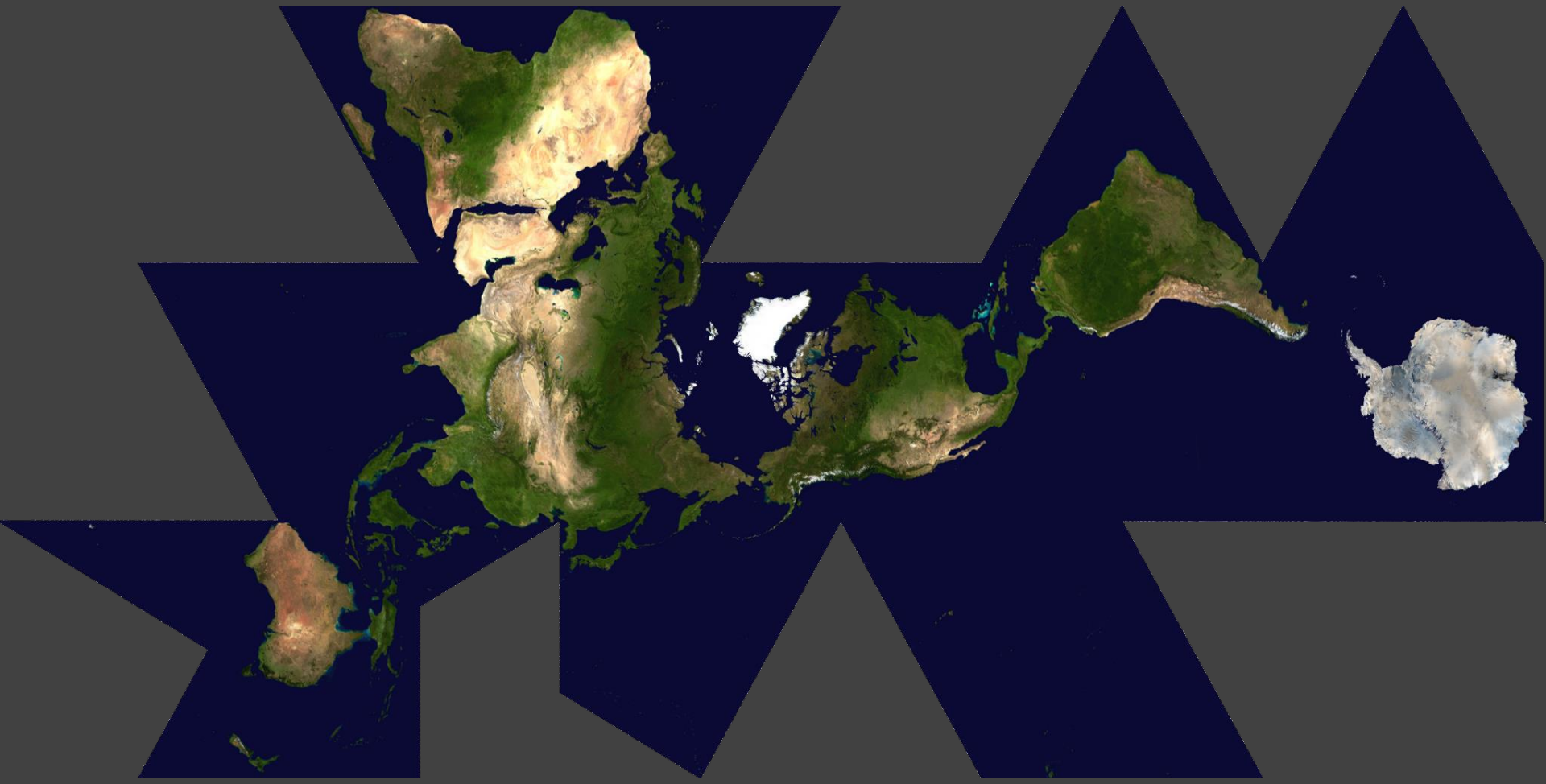


Interaction



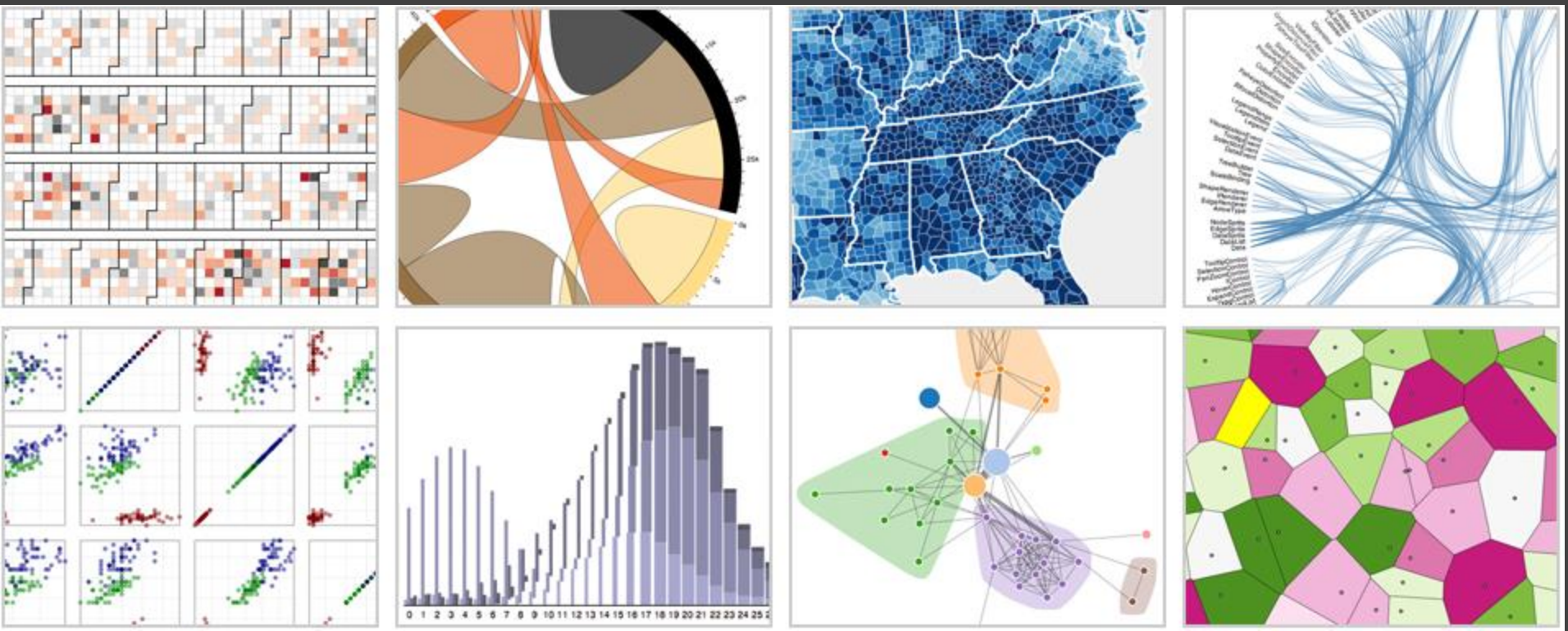
Crimespotting.org

Maps



Dymaxion Maps [Fuller 46]

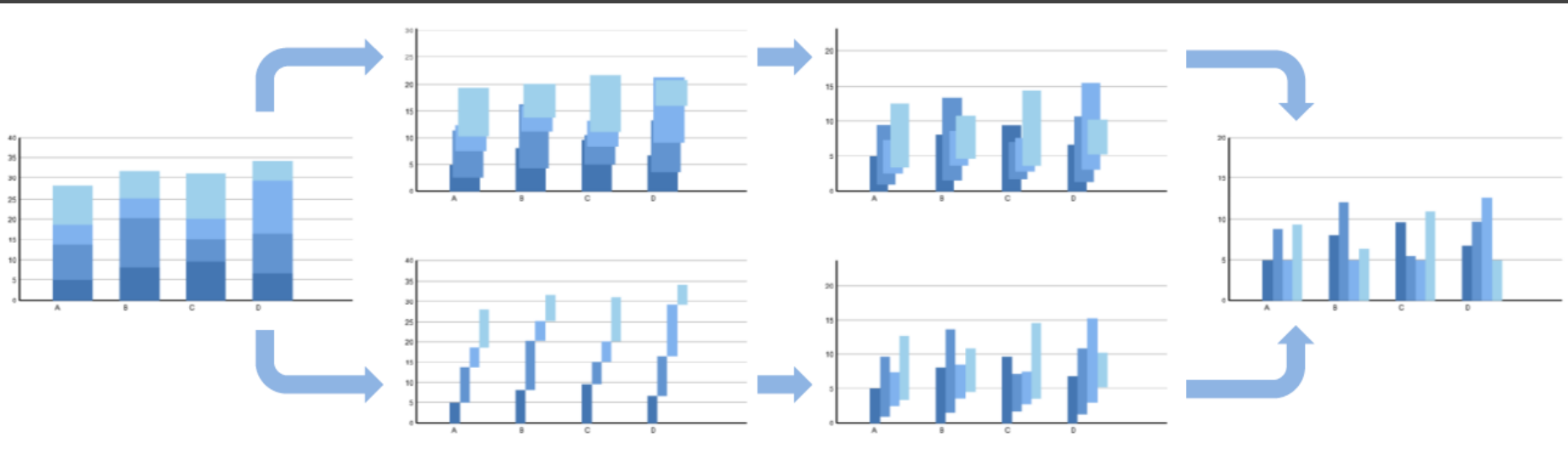
Visualization Software



D3: Data-Driven Documents

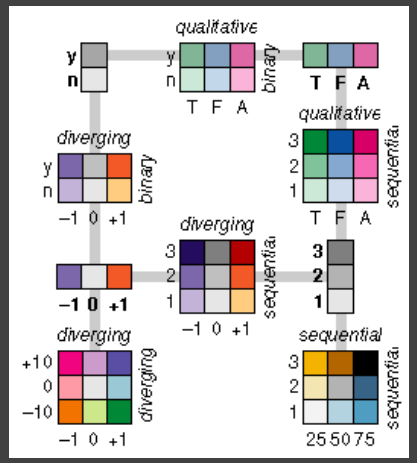
Vega-Lite / Altair

Animation

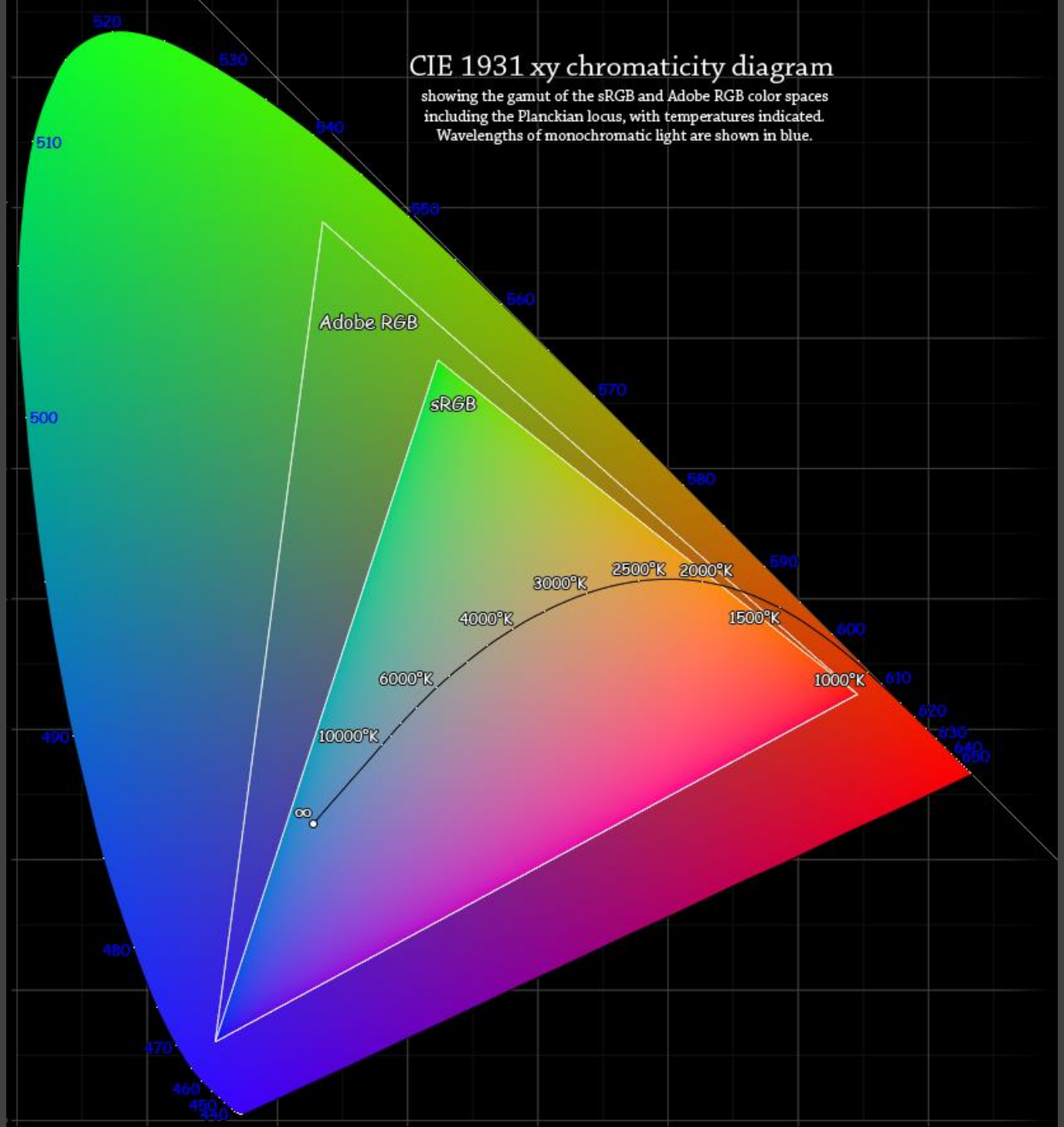


Animated transitions in statistical data graphics [Heer & Robertson 07]

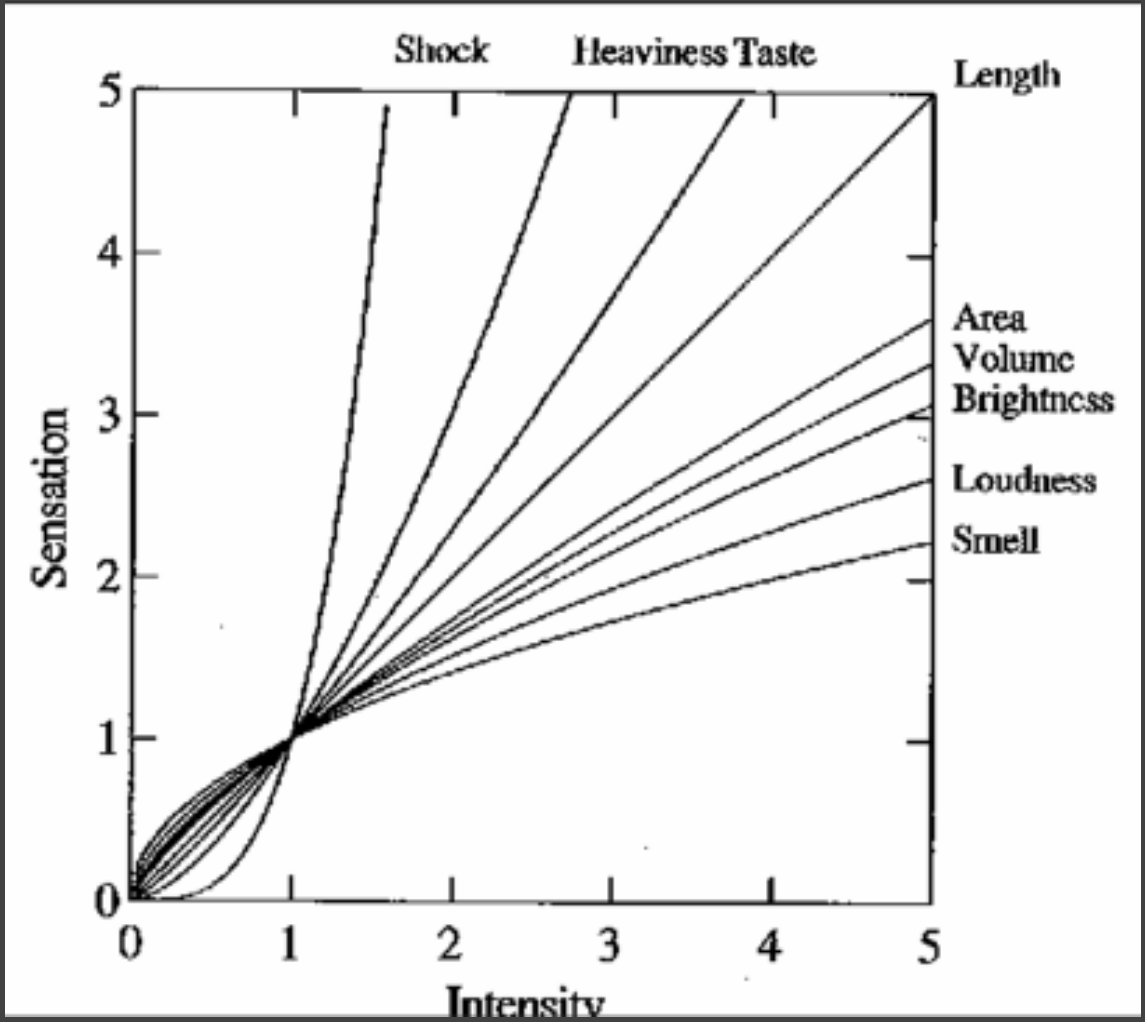
Color



Color Brewer

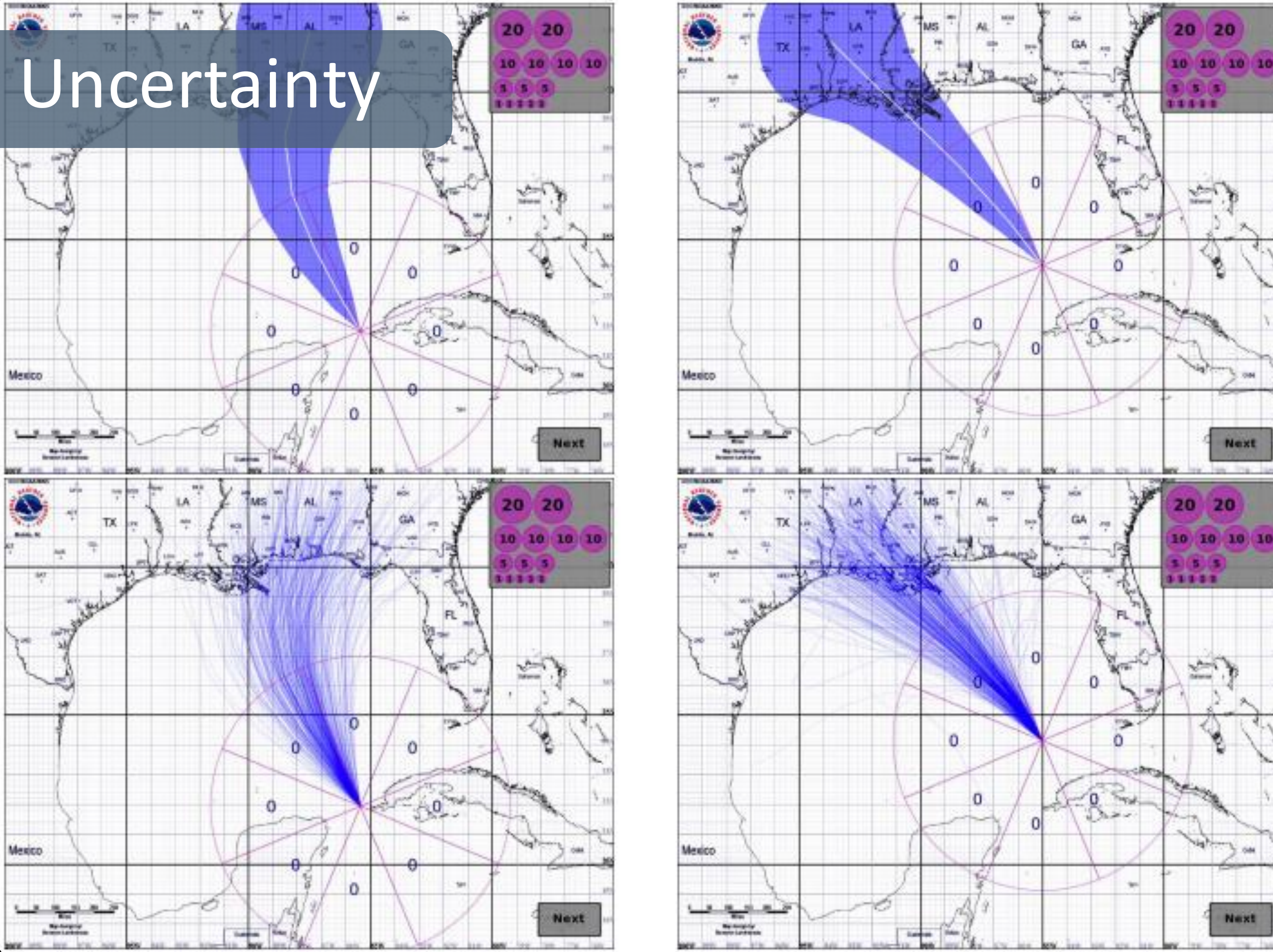


Graphical Perception

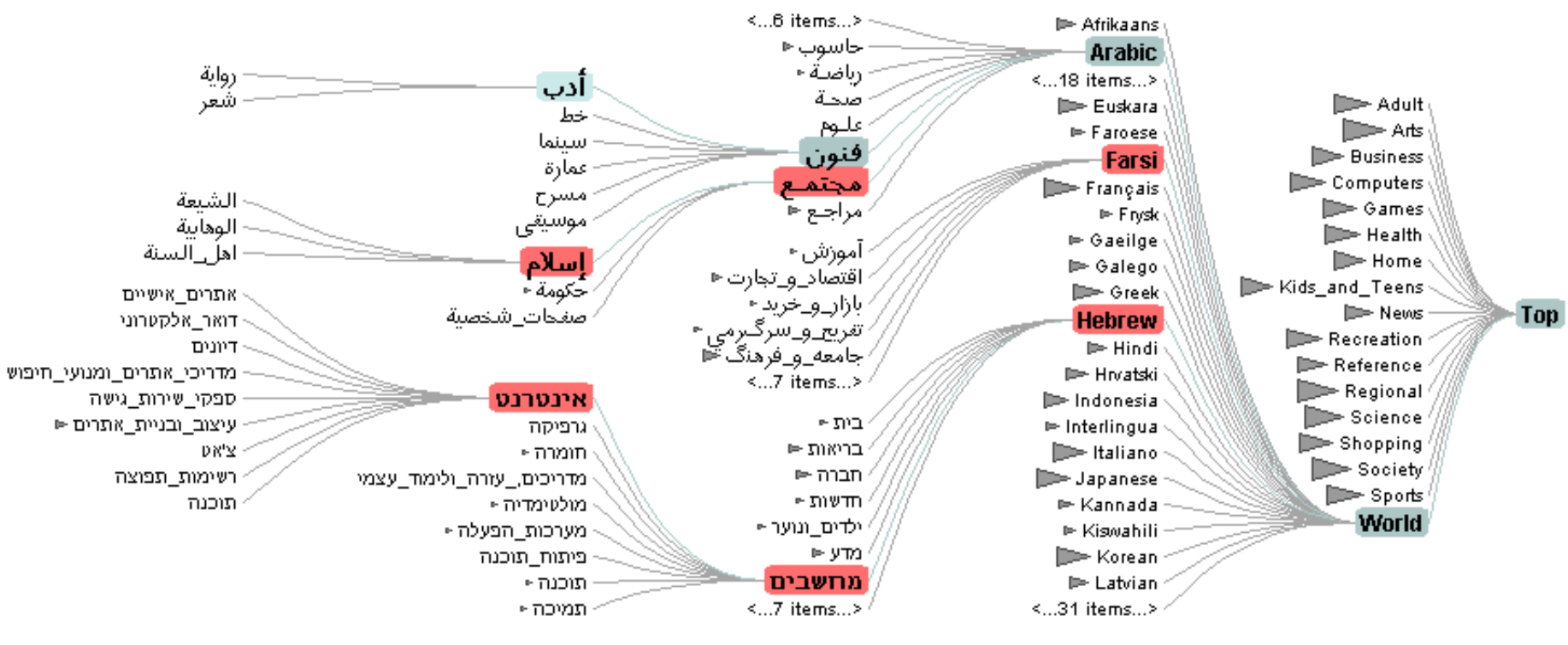


The psychophysics of sensory function [Stevens 61]

Uncertainty

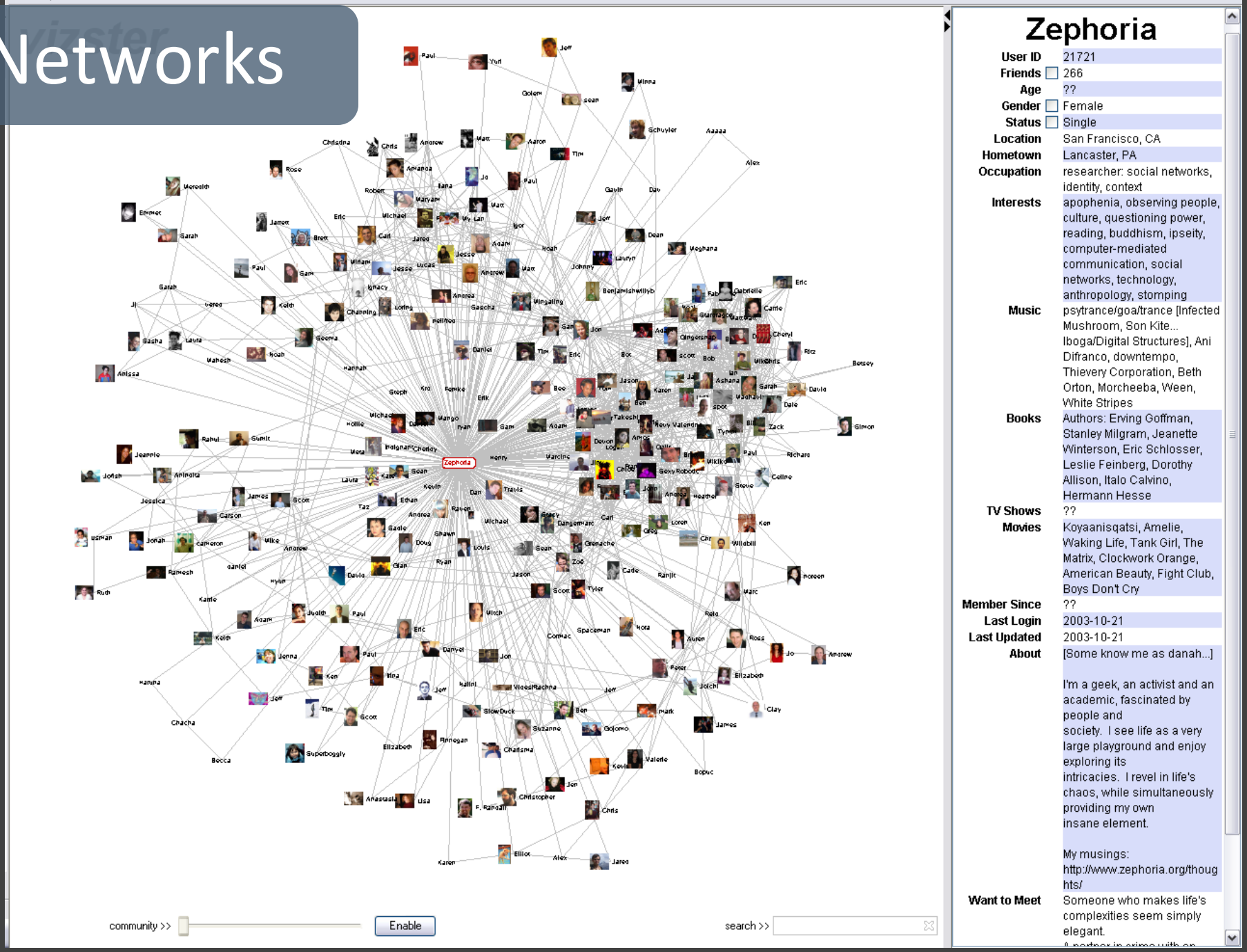


Hierarchies



Degree-Of-Interest Trees [Heer & Card 04]

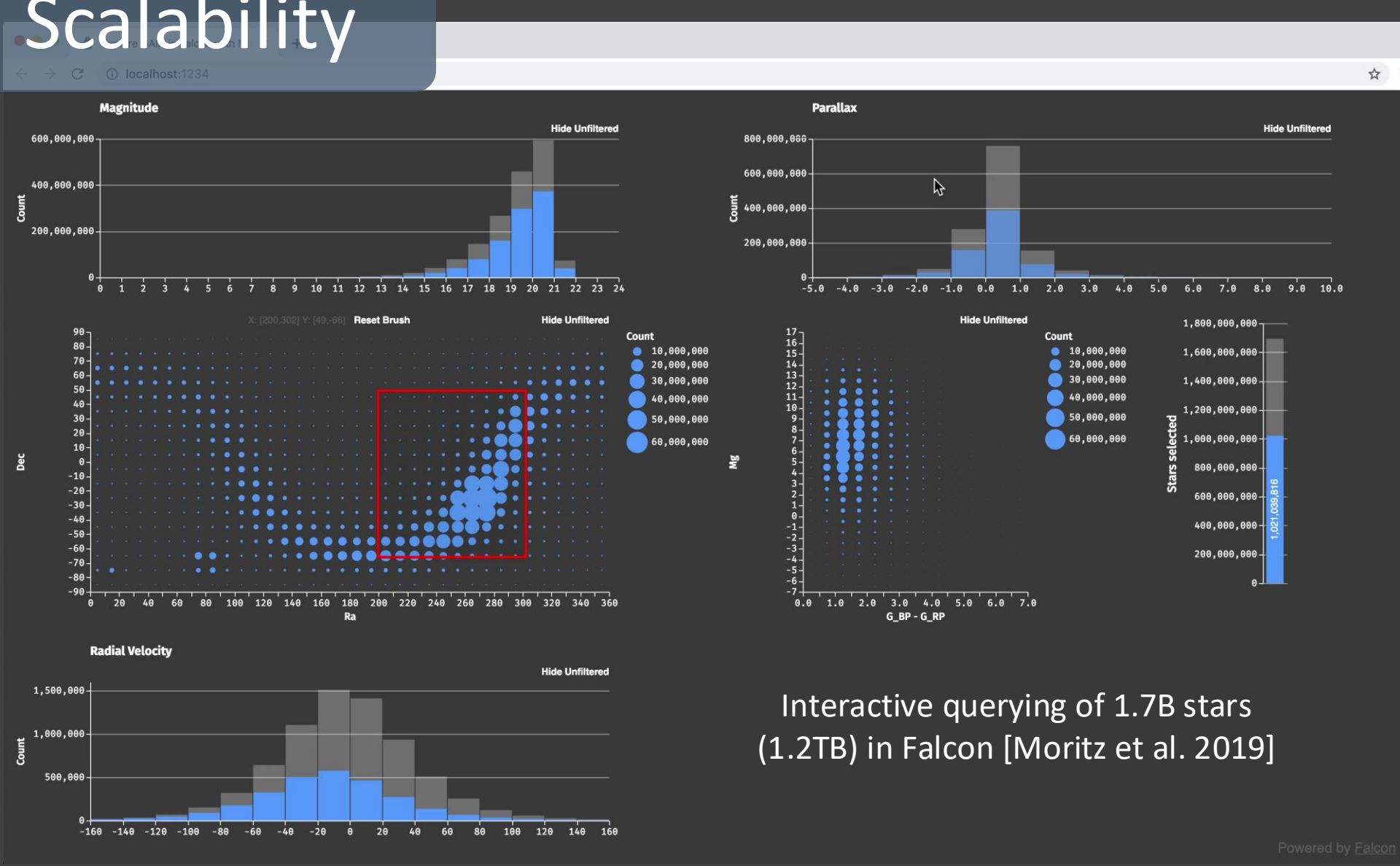
Networks



Zephoria

User ID	21721
Friends	266
Age	??
Gender	Female
Status	Single
Location	San Francisco, CA
Hometown	Lancaster, PA
Occupation	researcher: social networks, identity, context
Interests	apophenia, observing people, culture, questioning power, reading, buddhism, ipseity, computer-mediated communication, social networks, technology, anthropology, stomping
Music	psytrance/goa/trance [Infected Mushroom, Son Kite... Iboga/Digital Structures], Ani Difranco, downtempo, Thievery Corporation, Beth Orton, Morcheeba, Ween, White Stripes
Books	Authors: Erving Goffman, Stanley Milgram, Jeanette Winterson, Eric Schlosser, Leslie Feinberg, Dorothy Allison, Italo Calvino, Hermann Hesse
TV Shows	??
Movies	Koyaanisqatsi, Amelie, Waking Life, Tank Girl, The Matrix, Clockwork Orange, American Beauty, Fight Club, Boys Don't Cry
Member Since	??
Last Login	2003-10-21
Last Updated	2003-10-21
About	[Some know me as danah...] I'm a geek, an activist and an academic, fascinated by people and society. I see life as a very large playground and enjoy exploring its intricacies. I revel in life's chaos, while simultaneously providing my own insane element. My musings: http://www.zephoria.org/thoughts/
Want to Meet	Someone who makes life's complexities seem simply elegant.

Scalability



Interactive querying of 1.7B stars (1.2TB) in Falcon [Moritz et al. 2019]

Narrative

Recent elections have placed a heavy emphasis on "swing states" — Ohio, Florida and the other competitive states. Y[illegible]tween the Democratic and Republican parties. A look at how the states st[illegible] how they have shifted over past elections.

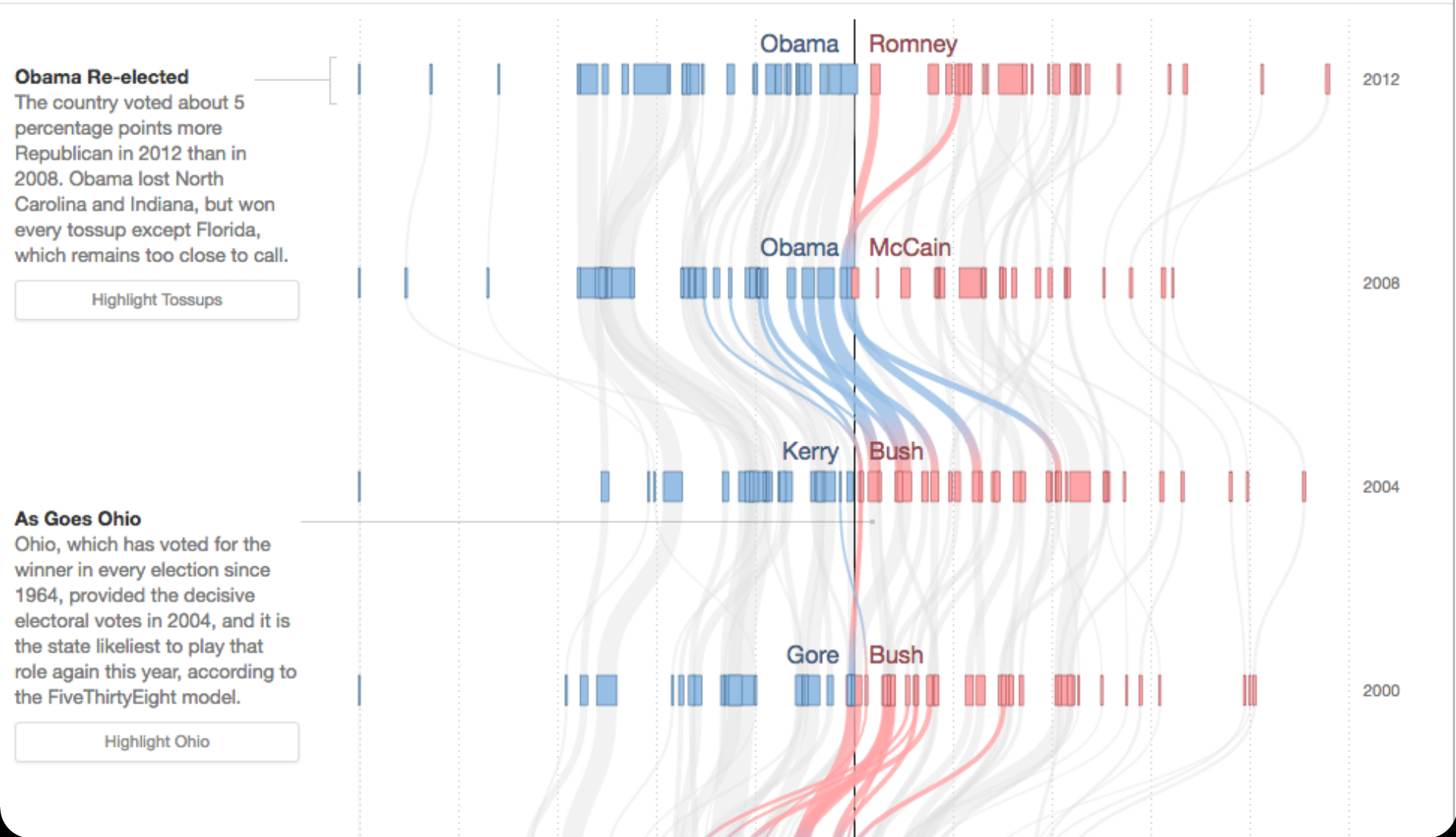
Each box represents a state sized by number of electoral votes.

Each curve shows how much it shifted left or right between elections

Chart **Size of Lead** | Chart **Electoral Votes**

← MORE DEMOCRATIC | MORE REPUBLICAN →

≥50% +40% +30% +20% +10% | +10% +20% +30% +40% ≥50%



Obama Re-elected
The country voted about 5 percentage points more Republican in 2012 than in 2008. Obama lost North Carolina and Indiana, but won every tossup except Florida, which remains too close to call.

Highlight Tossups

As Goes Ohio
Ohio, which has voted for the winner in every election since 1964, provided the decisive electoral votes in 2004, and it is the state likeliest to play that role again this year, according to the FiveThirtyEight model.

Highlight Ohio

Course Mechanics

You should expect to:

1. *Evaluate and critique* visualization designs
2. *Learn* visualization techniques & theory
3. *Implement* interactive data visualizations
4. *Develop* a substantial visualization project

Lectures & Office Hours

Watch the pre-recorded video before class on Thursdays!

Tues = Lectures. Thurs = in-class activities.

All Tues lectures will be in-person + recorded. We will use self assessments to gauge learning.

Please attend in person but **NOT** if you feel ill.

Office hours will be held in person or on Zoom.

Links are on Canvas for virtual office hours.

We strongly encourage using Ed to post questions and seek help!

Readings

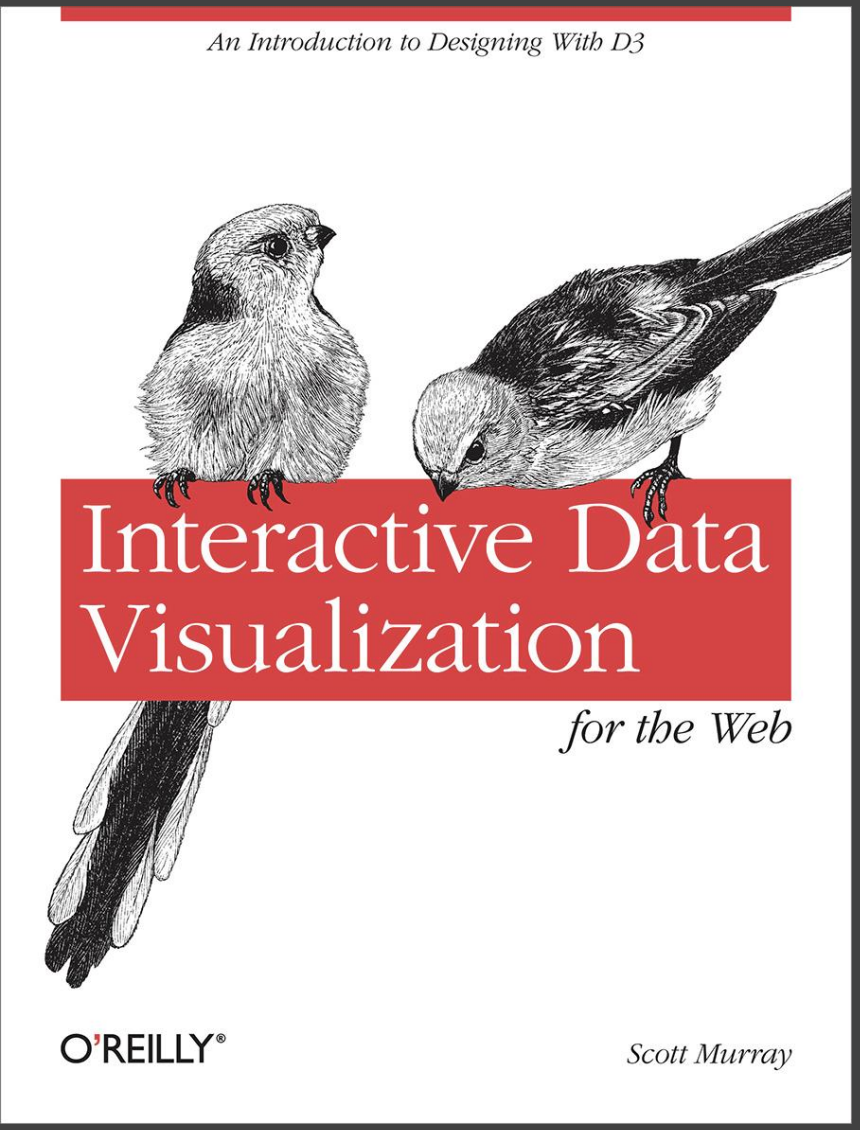
There is no one universal textbook on visualization!

So we will draw on books, notebooks, and linked articles.

Material in class will loosely follow readings.

Readings should be read by start of class.

Textbook

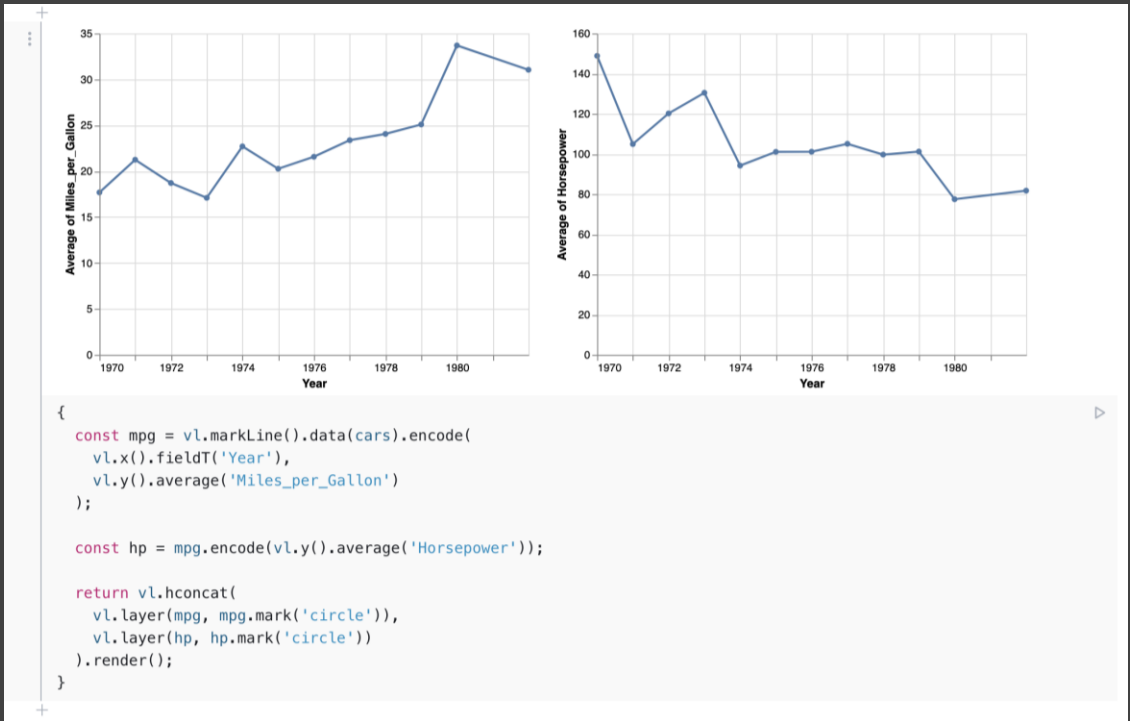


Interactive Data Visualization for the Web, 2nd Edition

For learning D3!
Book available online.
Code / examples on GitHub.

We will be using **D3 v7**.
https://d3js.org

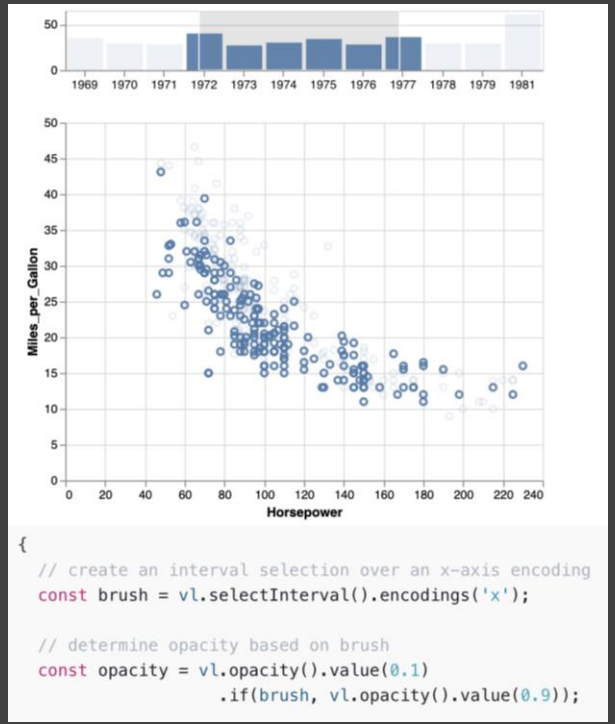
Interactive Vega-Lite Notebooks




```
{
  const mpg = vl.markLine().data(cars).encode(
    vl.x().fieldT('Year'),
    vl.y().average('Miles_per_Gallon')
  );

  const hp = mpg.encode(vl.y().average('Horsepower'));

  return vl.hconcat(
    vl.layer(mpg, mpg.mark('circle')),
    vl.layer(hp, hp.mark('circle'))
  ).render();
}
```




```
{
  // create an interval selection over an x-axis encoding
  const brush = vl.selectInterval().encodings('x');

  // determine opacity based on brush
  const opacity = vl.opacity().value(0.1)
                    .if(brush, vl.opacity().value(0.9));
```

Hands-on engagement with course concepts and tools using Observable (JavaScript) notebooks.

Assignments (No exams)

CP Class Participation (10%)

A1 Expository Visualization (10%) - *Due 1/17*

A2 Deceptive Visualization (15%) - *Due 1/29*

Peer Review (5%) - *Due 2/4 (deadline < 1 week)*

A3 Interactive Prototype (19%) – *Due 2/18*

Team Registration (1%) - *Due 2/7*

Peer Review (5%) - *Due 2/25 (deadline < 1 week)*

FP Final Project (35%)

Proposal - *Due 2/14*

Prototype - *Due 2/26*

Demonstration Video - *Due 3/11*

Final Prototype - *Due 3/18*

Grading Philosophy

A *great* submission gets a *great grade* (A- to A, 3.6 – 3.8), but an *exceptional grade* (A+, 3.9 – 4.0) requires *exceptional creativity/design (typically top 10%).*

Example: Typical A1 grades (out of 10 points).

Everyone starts with a high score (9/10).

Then, we *deduct* points for errors. We also *add* points for creativity and design above and beyond the assignment requirements.

The median score for A1 is typically 8.5 out of 10 (considered an A-).

Final Project

Produce an **explorable visual explanation**

Initial **prototype** and **design review**

Final deliverables and **video presentation**

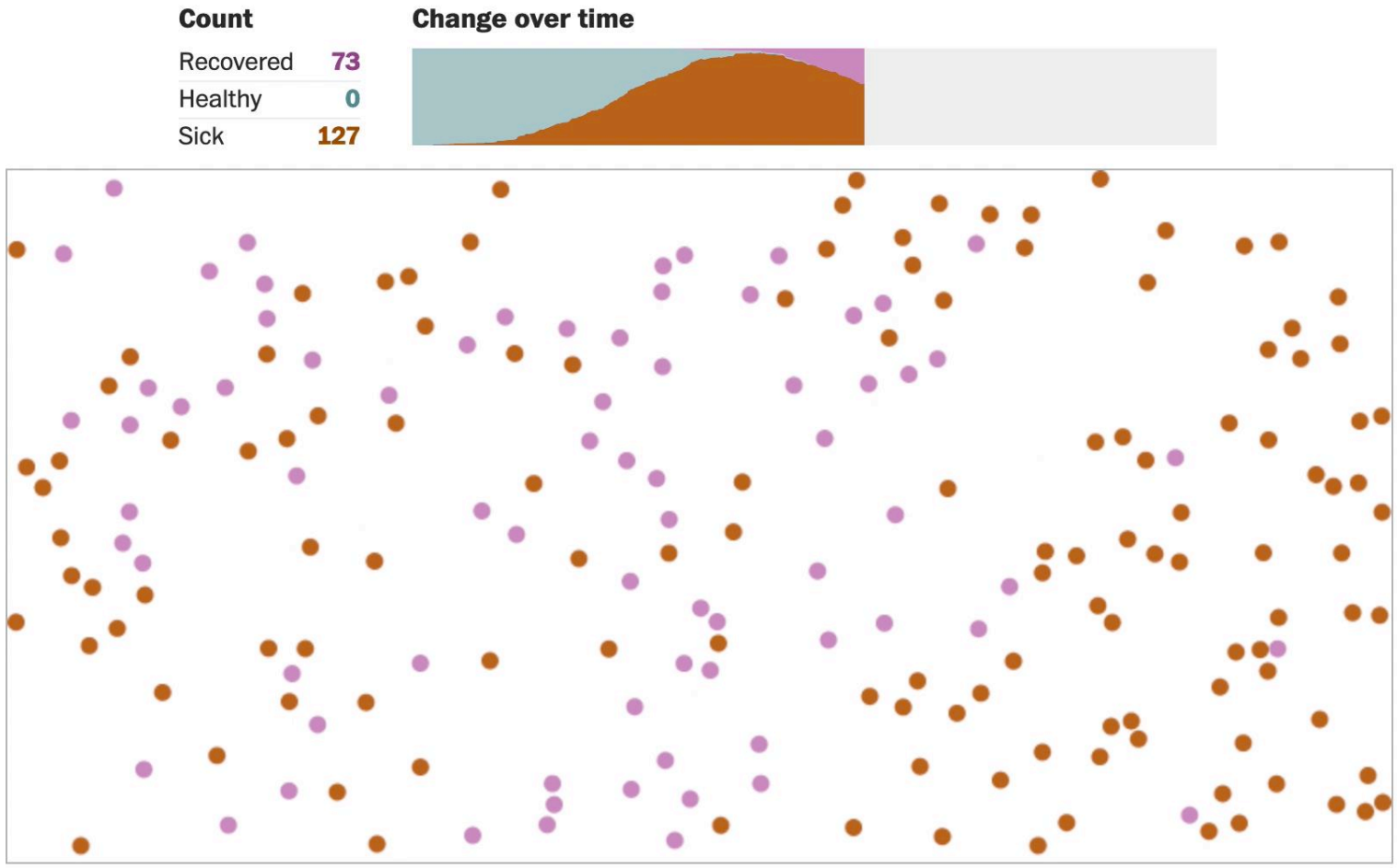
Submit and **publish online** (GitLab)

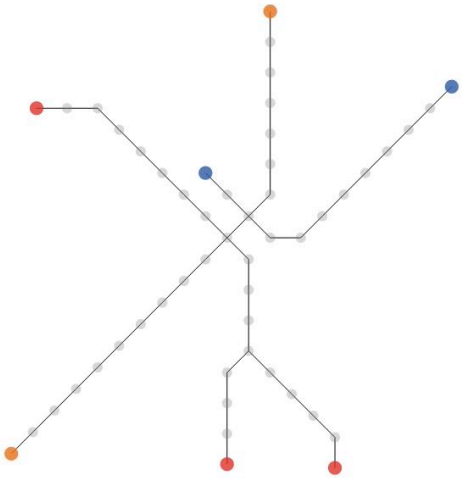
Projects from **previous classes** have been:

- Published as research papers
- Shared widely (some in the New York Times!)
- Released as successful open source projects

Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve"

Harry Stevens, Washington Post 2020

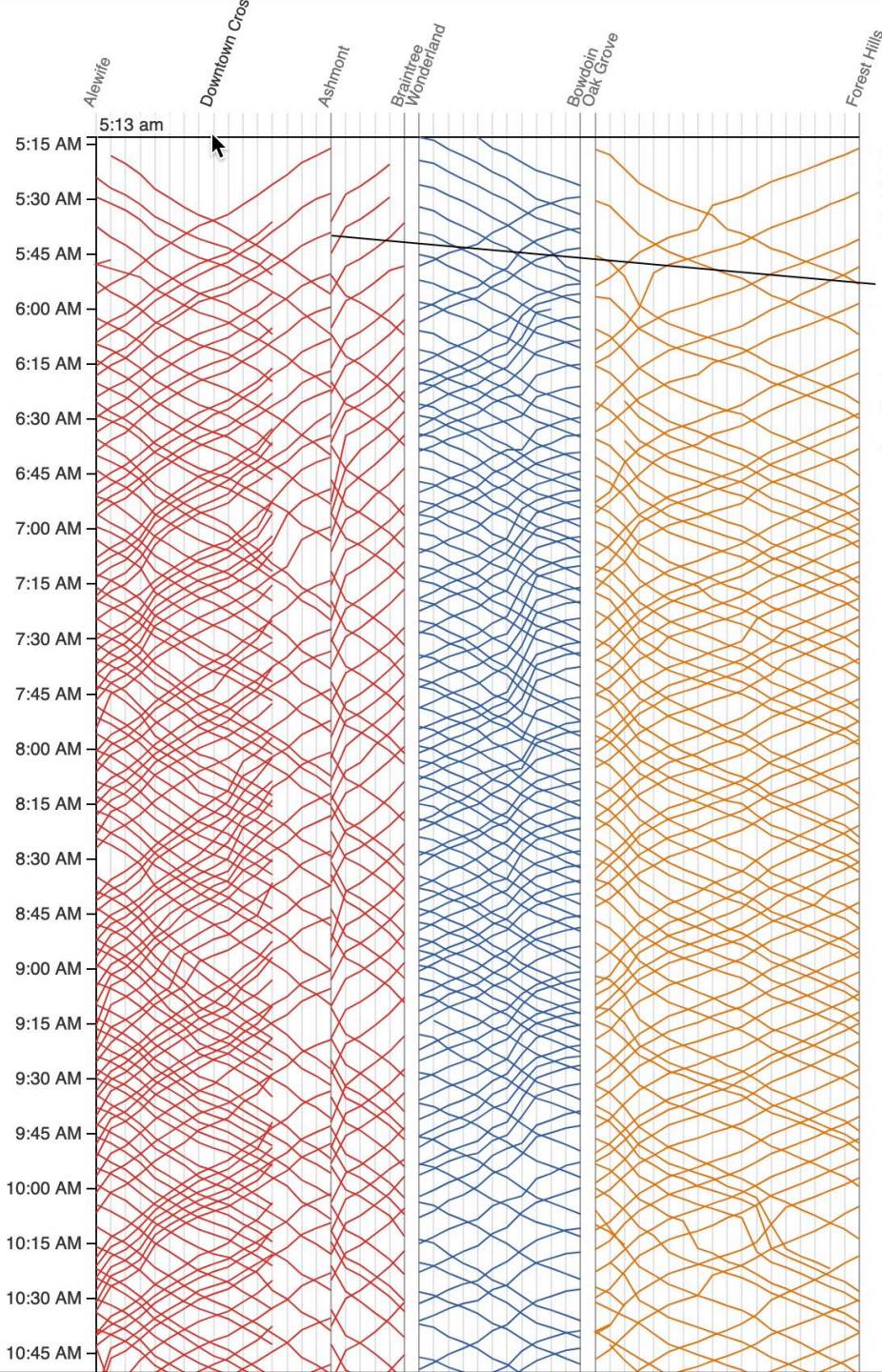




Locations of each train on the red, blue, and orange lines at 5:13 am. Hover over the diagram to the right to display trains at a different time.

Trains are on the right side of the track relative to the direction they are moving.

See the morning rush-hour, midday lull, afternoon rush-hour, and the evening lull.



Service starts at 5AM on Monday morning. Each line represents the path of one train. Time continues downward, so steeper lines indicate slower trains.

Since the red line splits, we show the Ashmont branch first then the Braintree branch. Trains on the Braintree branch "jump over" the Ashmont branch.

Train frequency increases around 6:30AM as morning rush hour begins.

MBTA Viz

Barry & Card

KEYBOARD WALKING

Passwords with a “keyboard walking” pattern start at an arbitrary key, then move in a direction (usually right or down) while continuing to hit keys. Sometimes this is combined with holding down the SHIFT key, so that some characters are uppercase or symbols to improve complexity.

While the generated password may seem to be random and unhackable, password crackers check for these keyboard patterns and guess them early on.

Many passwords in the leaked passwords dataset have a spatial pattern. Other than the numeric passwords like 123456, common keyboard walking offenders include qwerty and 1qaz@wsx.

Semantic Passwords

Vishal Devireddy (CSE 512, Spring ’21)

Course Participation

Thur In-Class Activities – in-person teams

Online self-assessments – virtual

No exams!

Online Self Assessments

We assign assessments to gauge student learning each week.

Assessments are due each Monday by 11:59pm, starting next week. If you participate, you will receive full credit.

Self assessments only count towards course participation.

Coming Up Soon!

Thur Jan 16: In-Class Activity

This Thursday (Jan 9) will be another lecture.

We will have our first in-class activity on Thursday next week!

You need to watch the pre-recorded lecture video before the activity. (We will post them soon!)

Observable + Data Tutorial

This Friday Jan 10, 4-5:30pm. Virtual.

Introduction to Observable notebooks, JavaScript basics, and data management and transformation, led by Tae.

Zoom link will be available on Canvas.
The tutorial will be recorded.

A1: Expository Visualization

Design a static visualization for a data set.

The climate of a place can have a tremendous impact on people's lived experience. You will examine average monthly climate measurements for six major U.S. cities, roughly covering the edges of the continental United States.

You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

A1: Expository Visualization

Pick a **guiding question**, use it to title your vis.
Design a **static visualization** for that question.
You are free to **use any tools** (inc. pen & paper).

Deliverables (upload on Gradescope; see A1 page)
Image of your visualization (PNG or JPG format)
Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm, Fri Jan 17**.

Seeking Help From Course Staff

The fastest way to reach us is through the Ed Discussion Board

Email us ASAP if you need access to edstem.org!

We also hold virtual and in-person office hours each week (schedule on the next slide).

We can also be reached over email at cse442@cs.washington.edu

Instructors

cse442@cs

Instructor

Leilani Battle OH: *Wed 5-6pm (virtual)*
Assistant Professor, CSE

Teaching Assistants

Lisa Elkin OH: *Mon 4:30pm (virtual)*

Tae Jones OH: *Online / Ed*

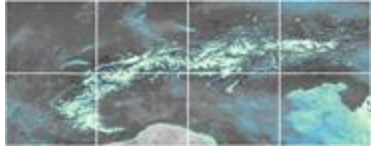
Heer Patel OH: *Online / Ed*

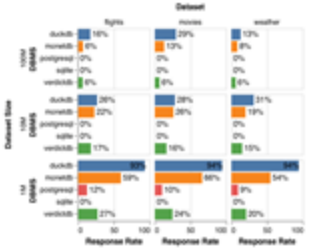
Han Zhang OH: *Thur 10:30am (virtual, TBD)*

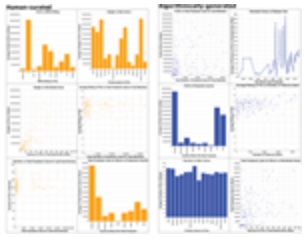
Jiawen Zhu OH: *Fri 10:45am (virtual, TBD)*

Leilani Battle (she/her)

Assistant Professor, UW CSE
Co-Director, CSE Interactive Data Lab
https://homes.cs.washington.edu/~leibatt/





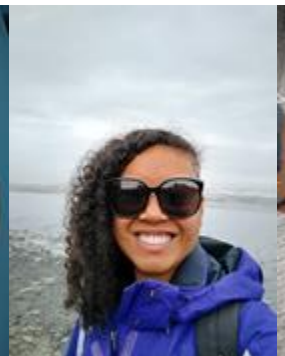


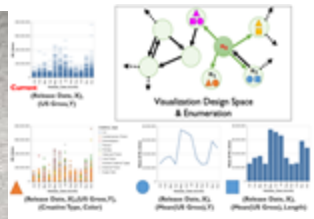
Visualization / HCI / Data management / Data Science

I model how people interact with data analysis systems.

I use these models to build ***behavior-driven*** optimizations, UI features, and performance benchmarks for interactive data analysis

Hobbies: disc golf, reading, cooking, board games, travel.





Lisa Elkin (She/Her)

Academic Background

BMath, C&O and Pure Math, University of Waterloo, 2012
MET, Entertainment Technology, Carnegie Mellon, 2015
MMath, Computer Science, University of Waterloo, 2018
PhD, UW CSE, 2018 - ???

TA Experience

HCI, Data Viz, Linear Algebra, Calculus, Intro CS, CS for non-majors

Industry Internships

MSR 2018, Apple 2021, Meta 2022, Meta 2023

Winston Elkin

Academic Background

- Doggy School Level 1*
- PhD, UW CSE, 2019 - ???
  - Dissertation: Evaluating the Impact of Tree Species on a Dog's Desire to Pee on it: an Autoethnography

Industry Internships

- Meta 2022, Meta 2023. Even has the badge to prove it.

* Indicates candidate did not complete all requirements but passed due to being very cute.

Tae Jones

4th Year PhD Student, CSE

Email: taejones@cs.washington.edu

Research Interests

- User/Patient Engagement, Physical-Mental Health Comorbidities, Mental Health & Wellbeing, Behavioural Change Support

Current Project

- Increasing patient engagement in long term collaborative interventions by understanding and reducing barriers

Current Hobbies

- Rhythm Skating on Quads, Gardening & Vermicomposting, Woodland Park Zoo Ambassador, Yoga

... and Miyazaki

Han Zhang (she/her)

5th year PhD student, CSE

micohan@uw.edu

Research interests

- Leveraging human-centered AI/ML to better understand human behaviors.
- Building responsible behavioral models that balance technical robustness and social impact.
- Designing user-centered AI systems to improve accessibility.

Other interests

- Badminton
- Hiking
- Traveling

Heer Patel

heerpate@cs.washington.edu

- 4th year BS/MS
- **Interests**
  - Data Science (Data Viz)
  - HCI
  - Business
- **Hobbies**
  - Squash (sport, not veggie)
  - Traveling to sunny locations :)
  - Henna

Jiawen Stefanie **Zhu**

stef-zjw.github.io
jiawenz2@uw.edu

UNIVERSITY OF WATERLOO — B.CS.

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING — Ph.D. in CSE

Research Interests

Human-Computer Interaction (HCI)
Human-AI Collaboration
Language, e.g. Multilingualism

Hobbies

Martial Arts
Hiking
Watercolour

Questions?