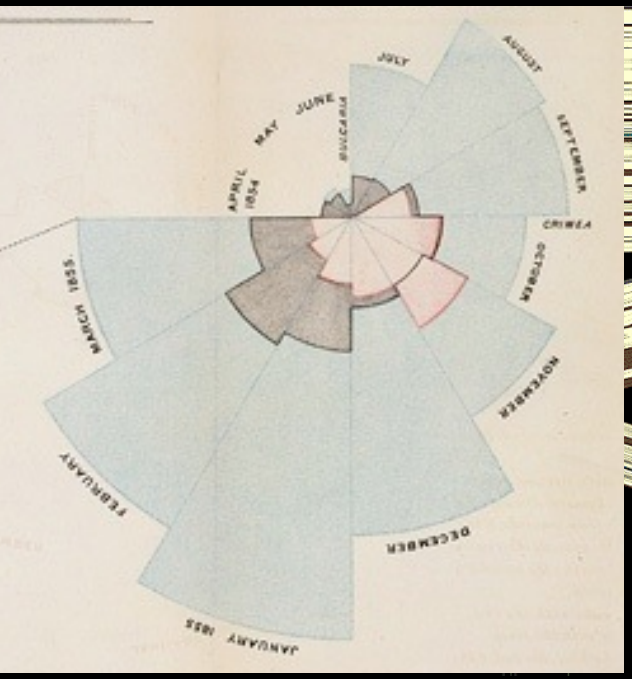
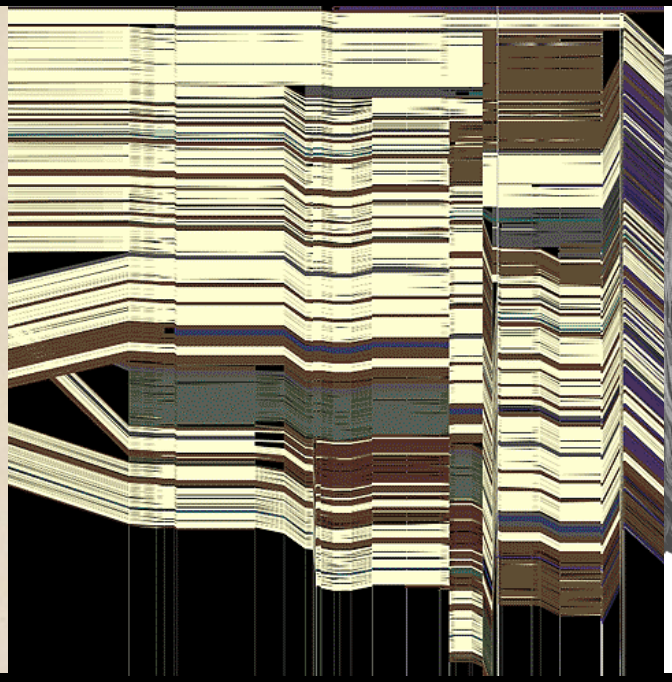
CSE 442 - Data Visualization

The Value of Visualization





Leilani Battle University of Washington

Testing poll everywhere setup

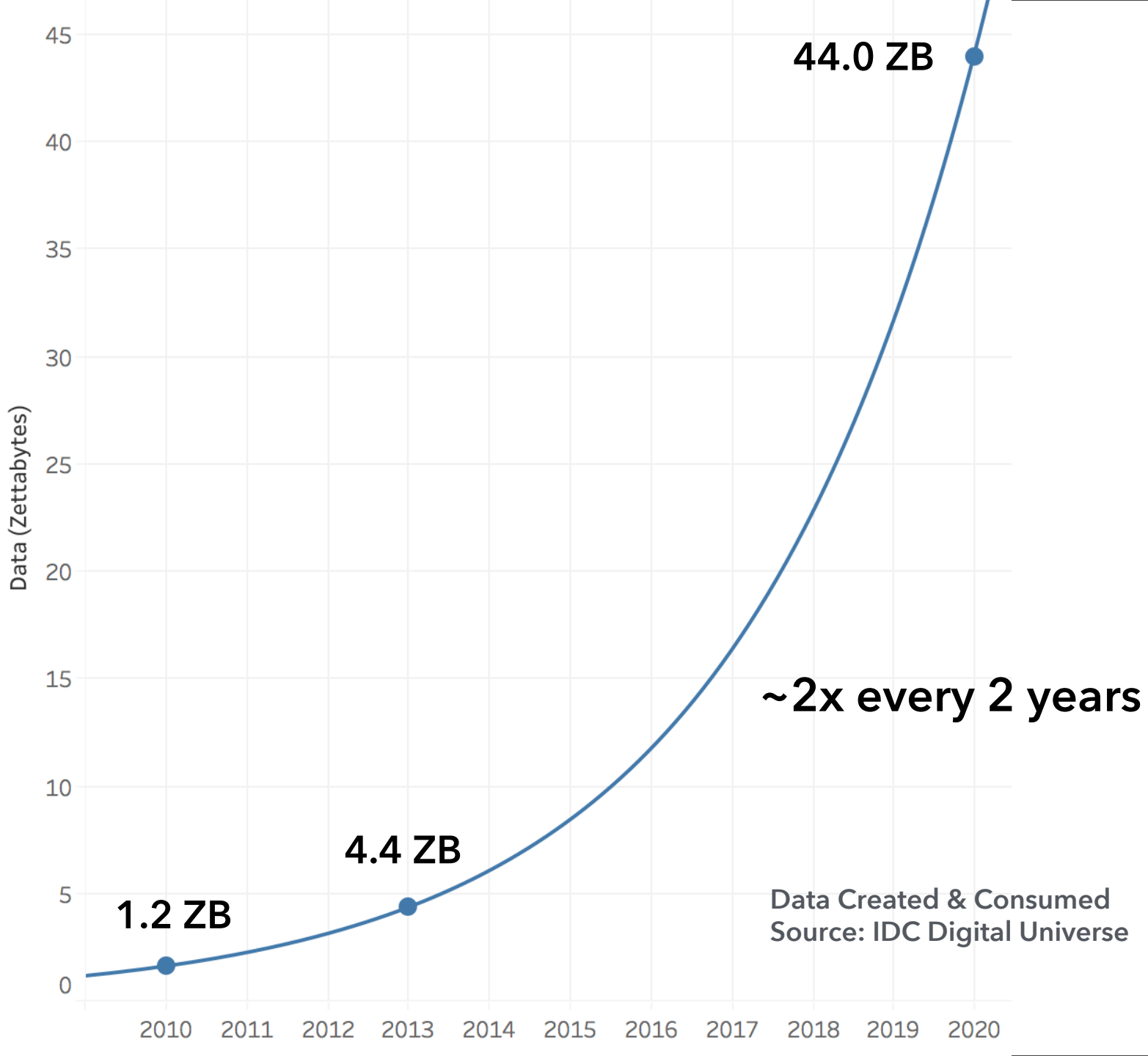
Respond here: pollev.com/leibatt

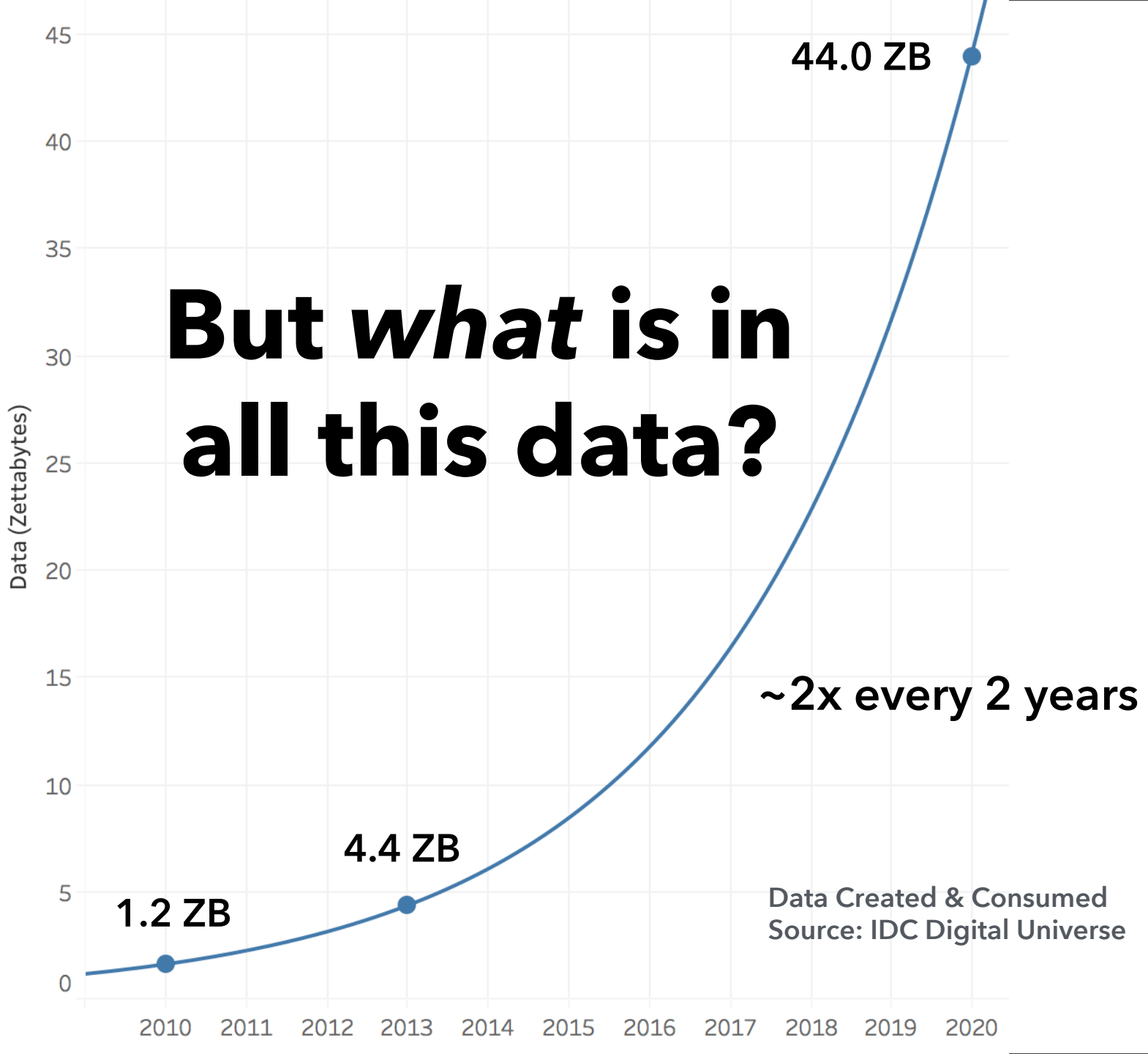
How much data (bytes) did we produce in 2010?

Respond here: pollev.com/leibatt

2010: 1,200 exabytes
and exponential growth…

Gantz et al., 2008, 2010

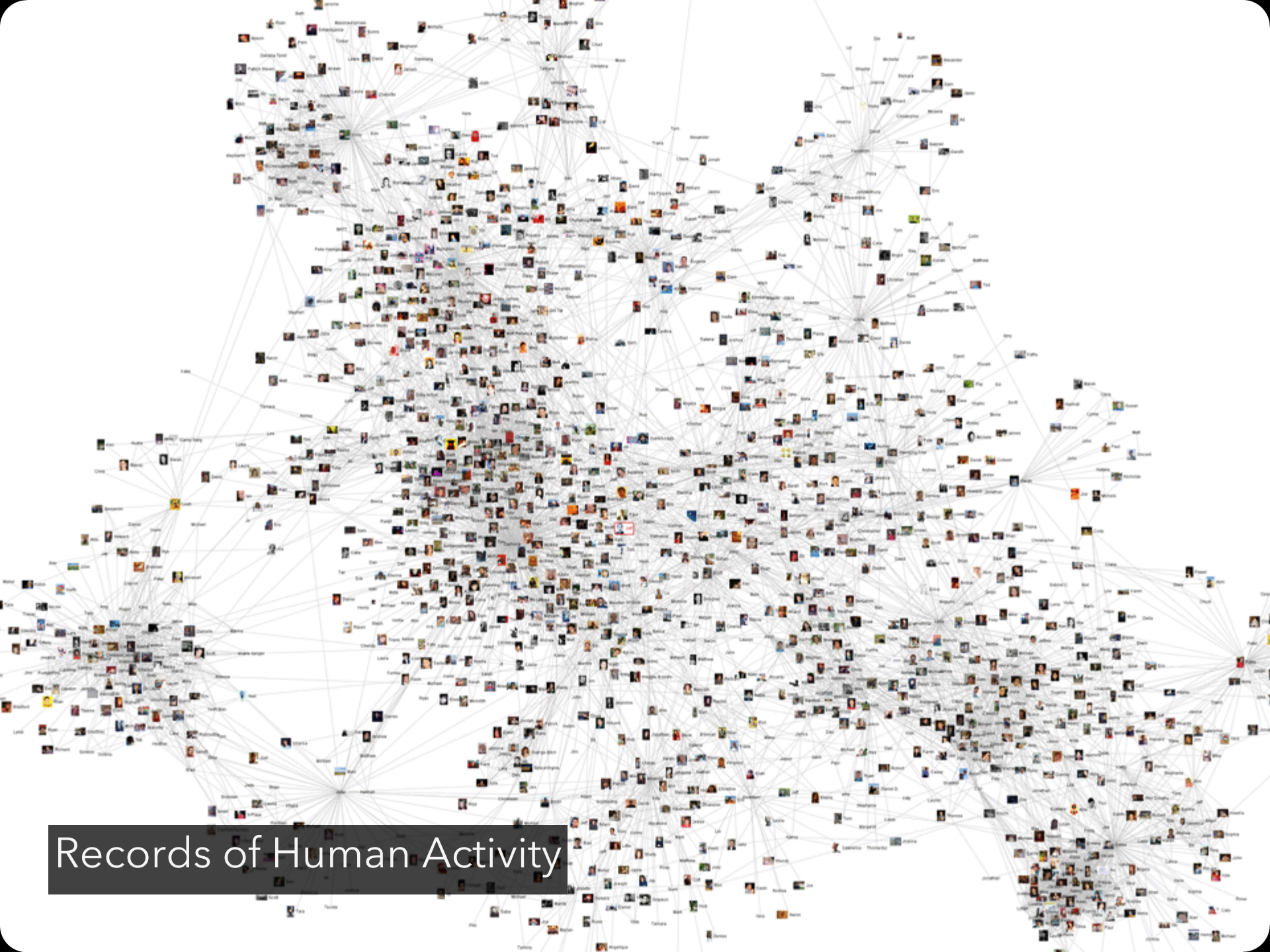

Data (Zettabytes)
45
40
35
30
25
20
15
10
5
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
44.0 ZB
4.4 ZB
1.2 ZB
~2x every 2 years
Data Created & Consumed
Source: IDC Digital Universe


But *what* is in all this data?
Data (Zettabytes)
45
40
35
30
25
20
15
10
5
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
1.2 ZB
4.4 ZB
44.0 ZB
~2x every 2 years
Data Created & Consumed
Source: IDC Digital Universe

Physical Sensors

Image courtesy cabspotting.org

Health & Medicine


Records of Human Activity

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 in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

But wait!

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 in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it.

"free" to whom?

"ubiquitous" about whom?

…to whose benefit?

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

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 − +

tdwi
CHICAGO
MAY 6–11
LEARN
Machine Learning & Advanced Analytics

Psychology's Replication Crisis Can't Be Wished Away

It has a real and heartbreaking cost.

ED YONG | MAR 4, 2016 | SCIENCE

Share Tweet ...

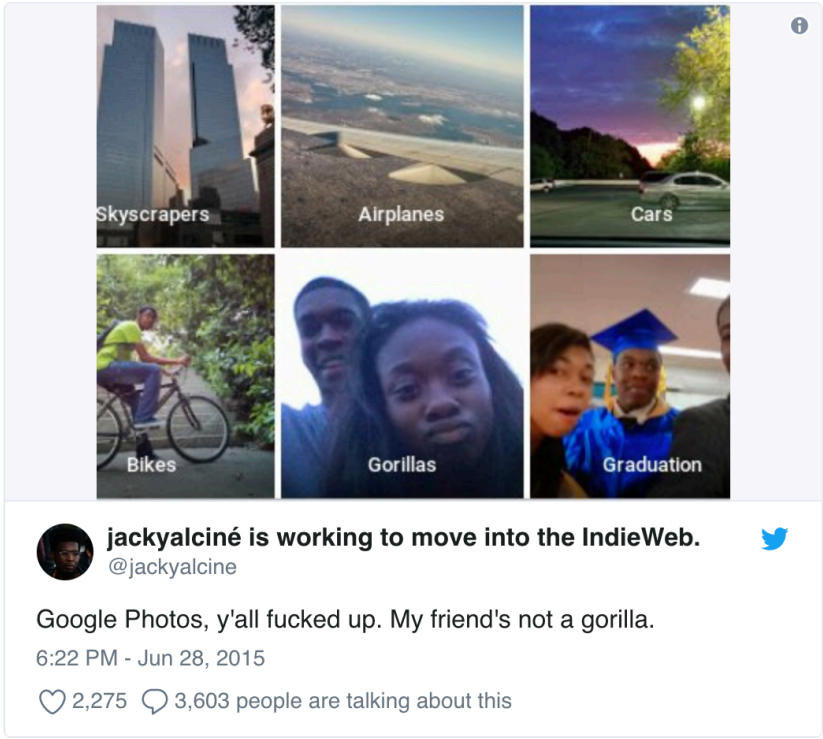
TEXT SIZE − +

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?



…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

$u_X = 9.0$ $\sigma_X = 3.32$

$u_Y = 7.5$ $\sigma_Y = 2.03$

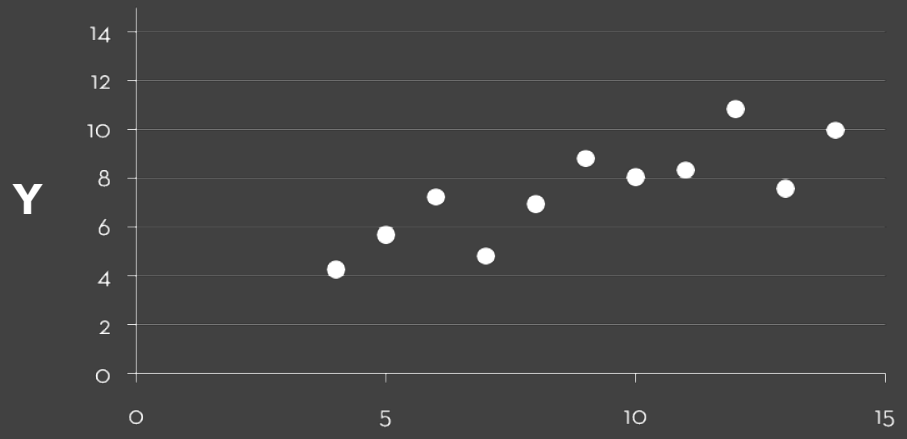
Linear Regression

$Y = 3 + 0.5\ X$

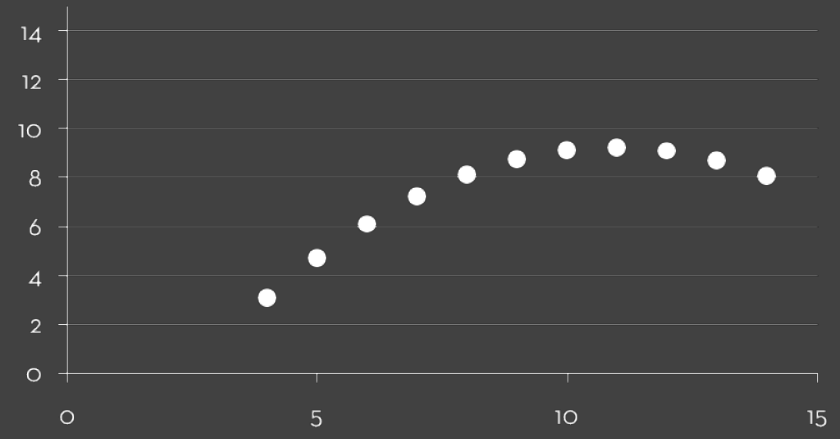
$R^2 = 0.67$

[Anscombe 1973]

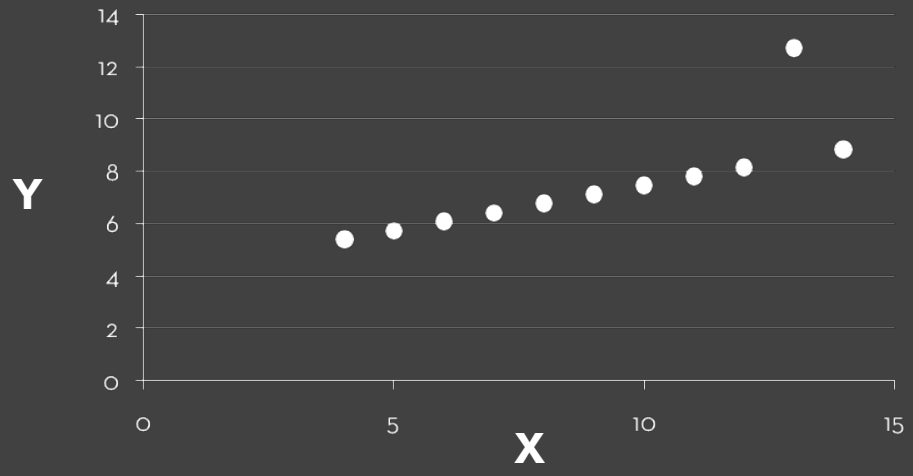
Set A



Set B

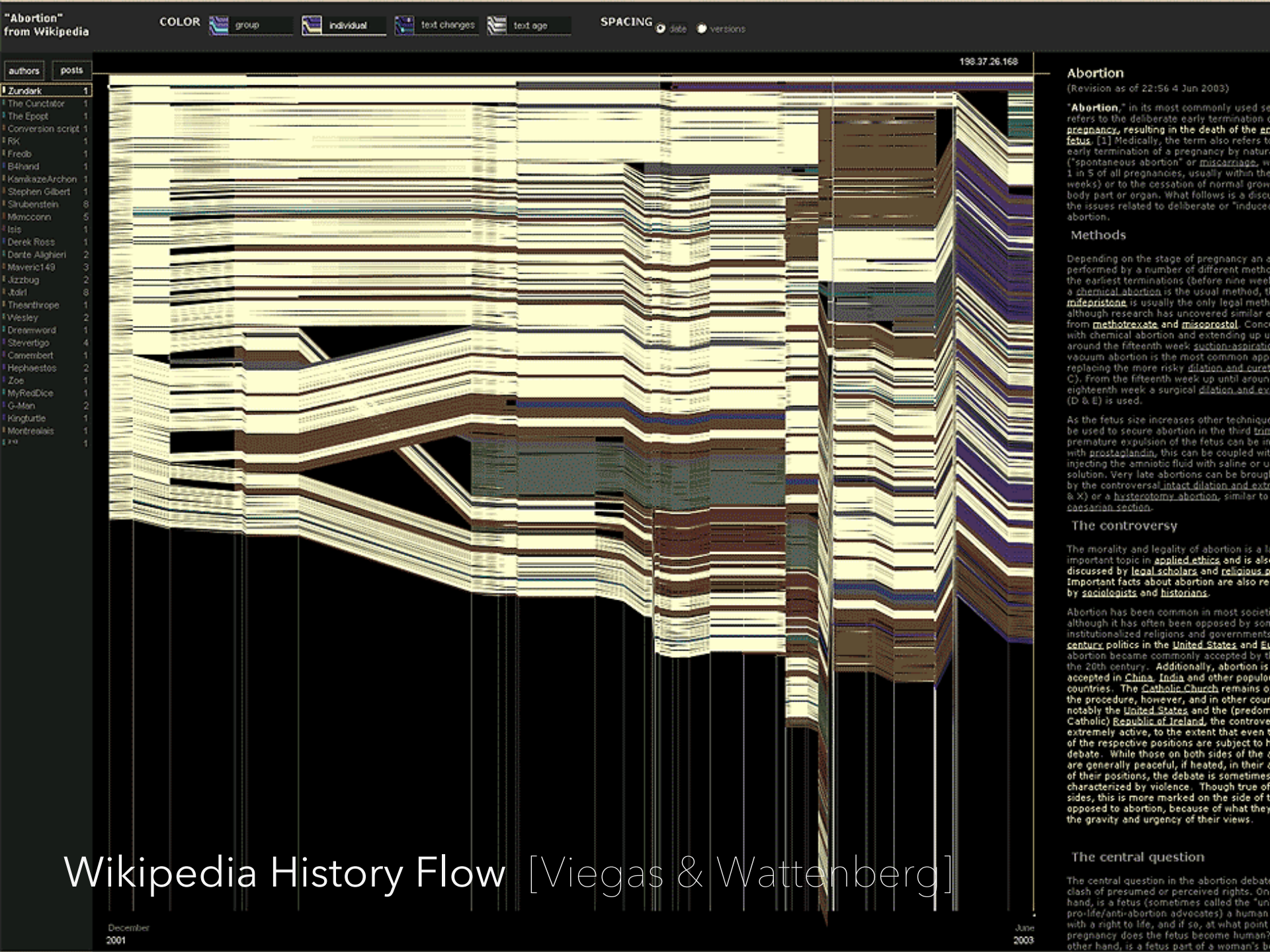


Set C



Set D

[Anscombe 1973]



Wikipedia History Flow [Viegas & Wattenberg]

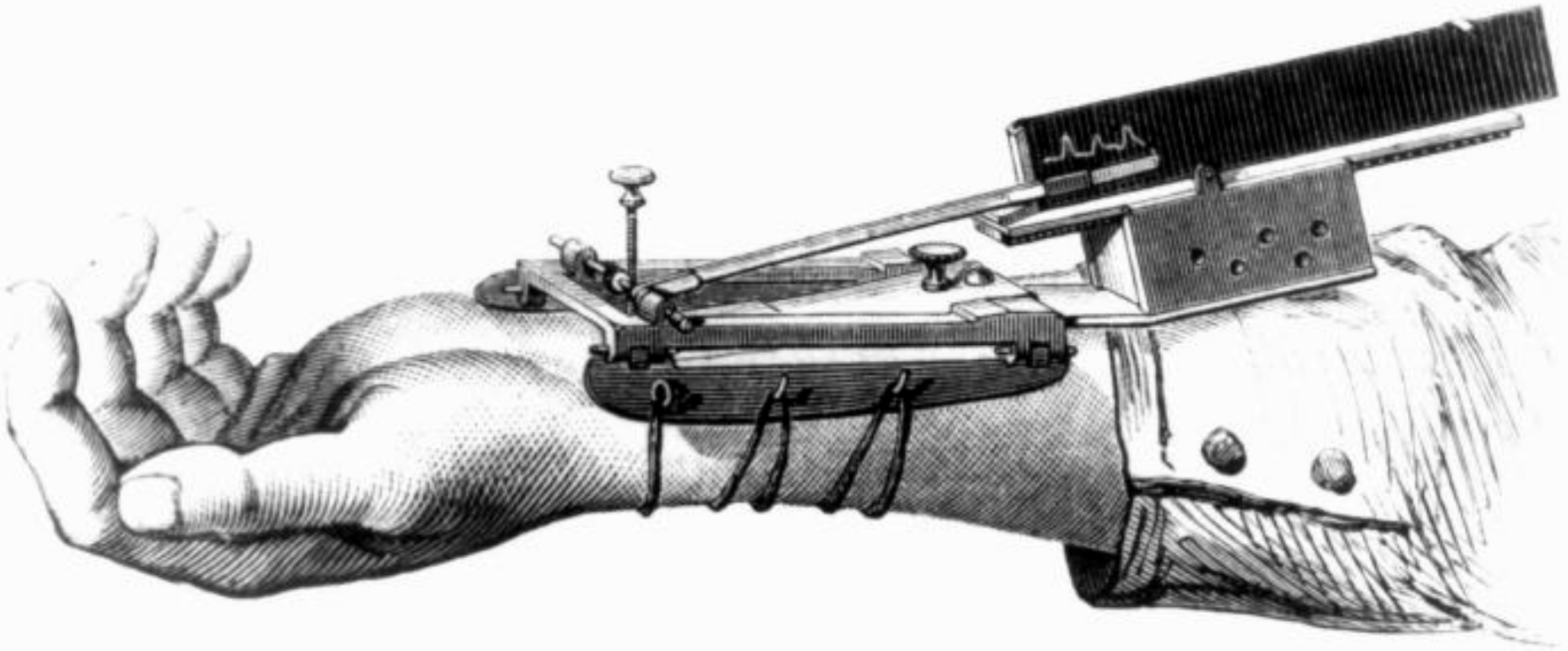
Wikipedia History Flow [Viegas & Wattenberg]

Why Create Visualizations?

Why Create Visualizations?

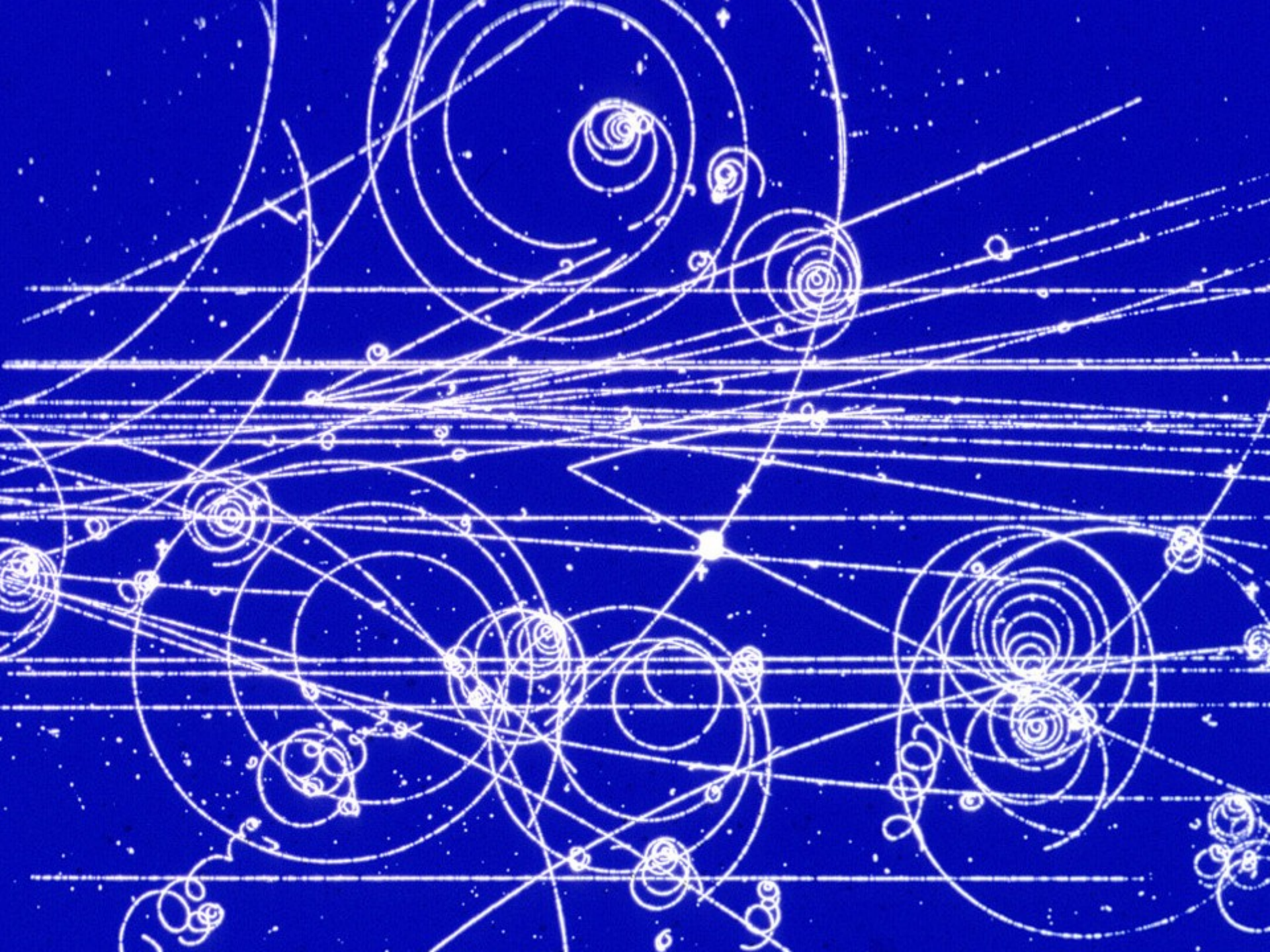
Answer questions (or discover them)
Make decisions
See data in context
Expand memory
Support graphical calculation
Find patterns
Present argument or tell a story
Inspire

Record Information

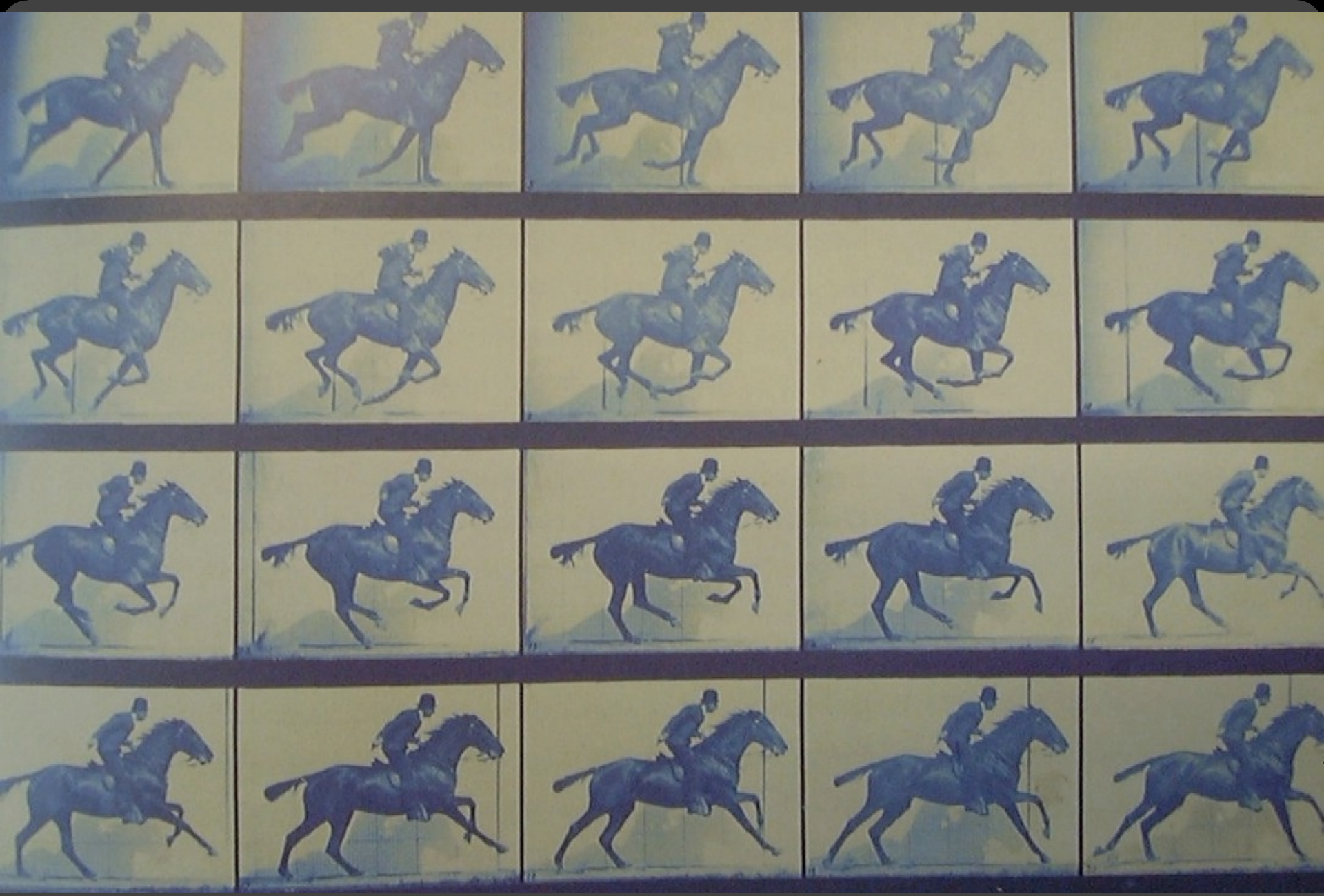


1.
Marey's sphygmograph in use, 1860. *La méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine.*

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

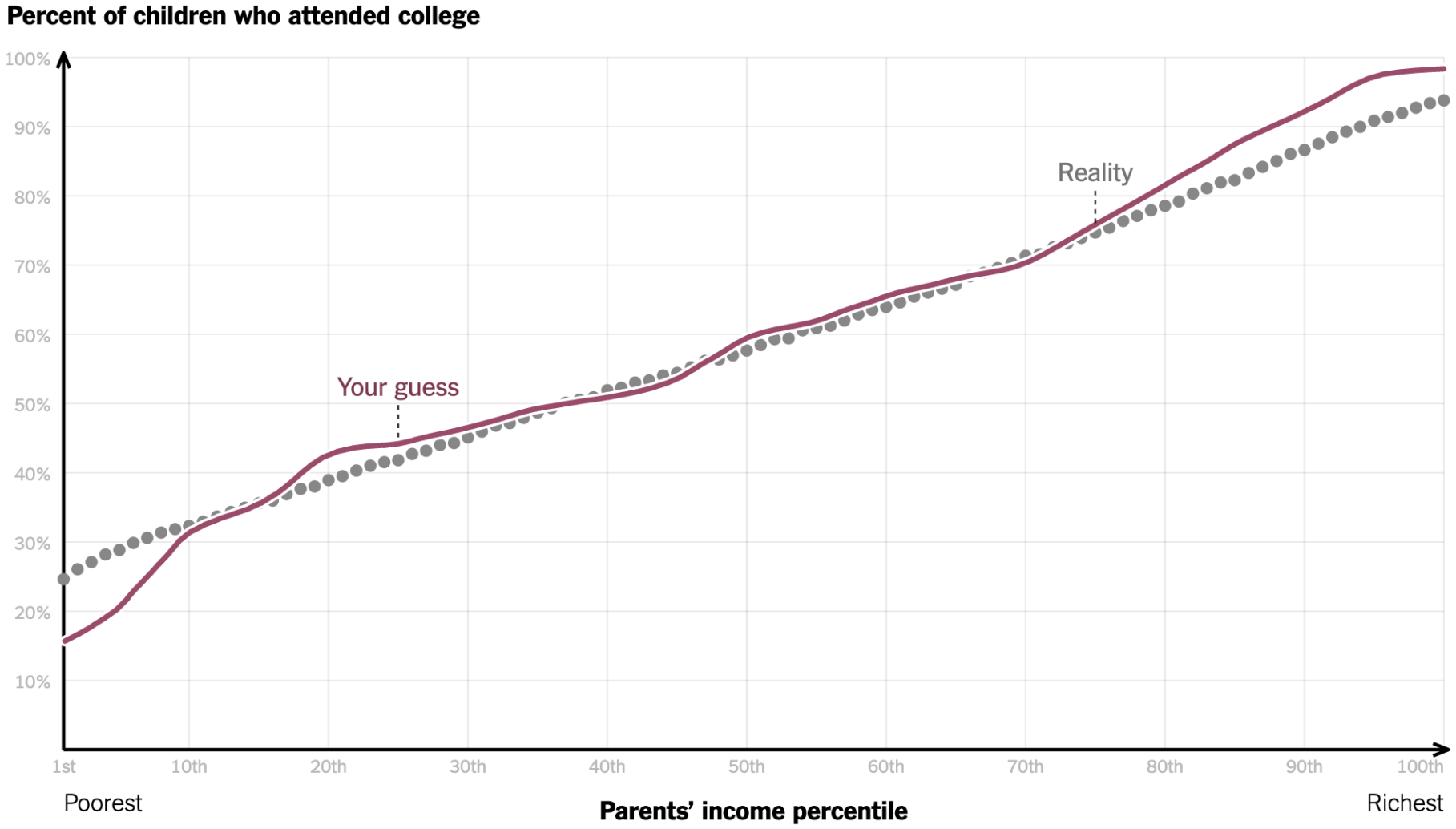


https://www.flickr.com/photos/brittgow/4782591644



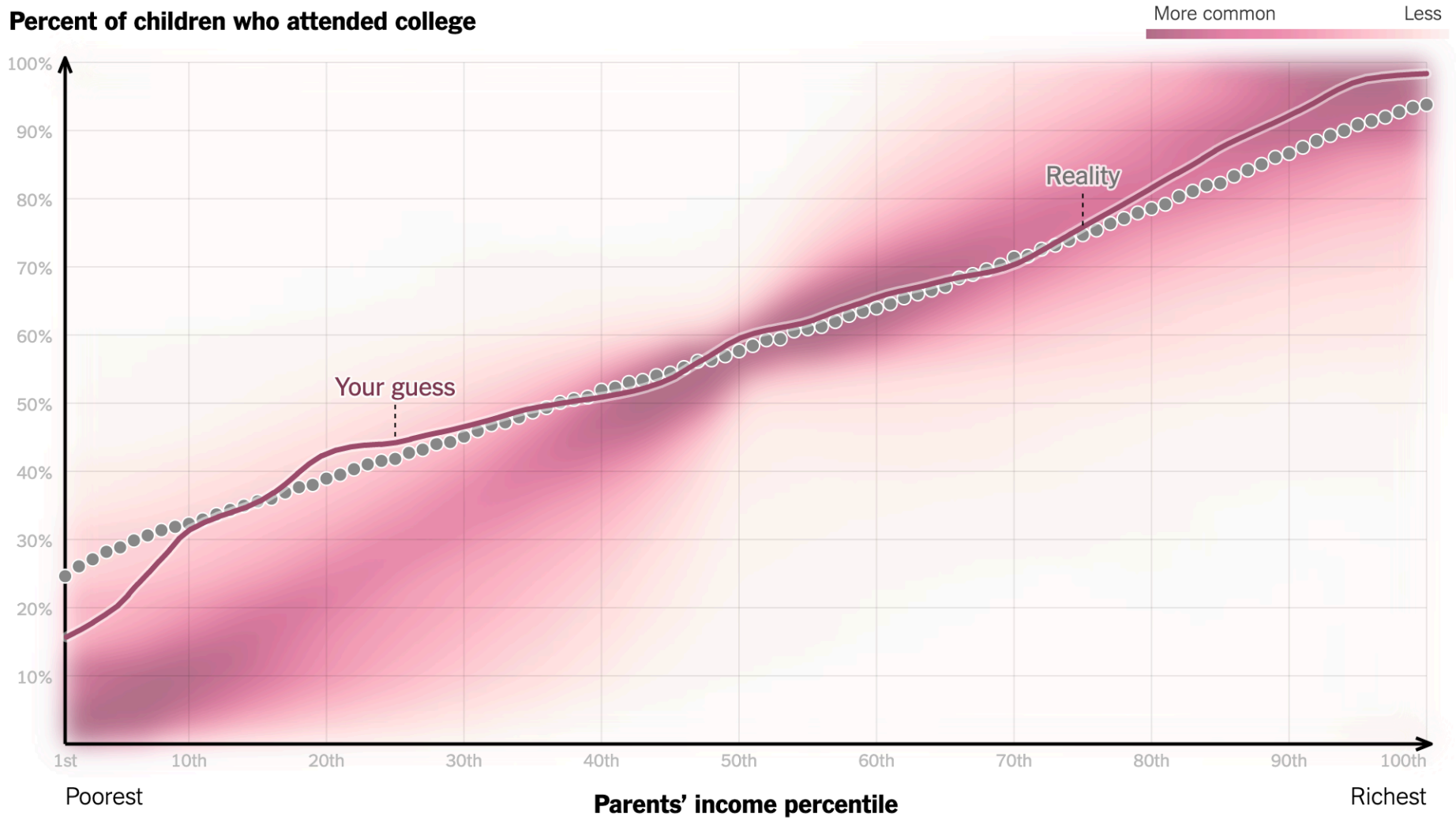
Gallop, Bay Horse “Daisy” [Muybridge]

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



You Draw It: How Family Income Predicts Children's College Chances

[New York Times, May 28, 2015]

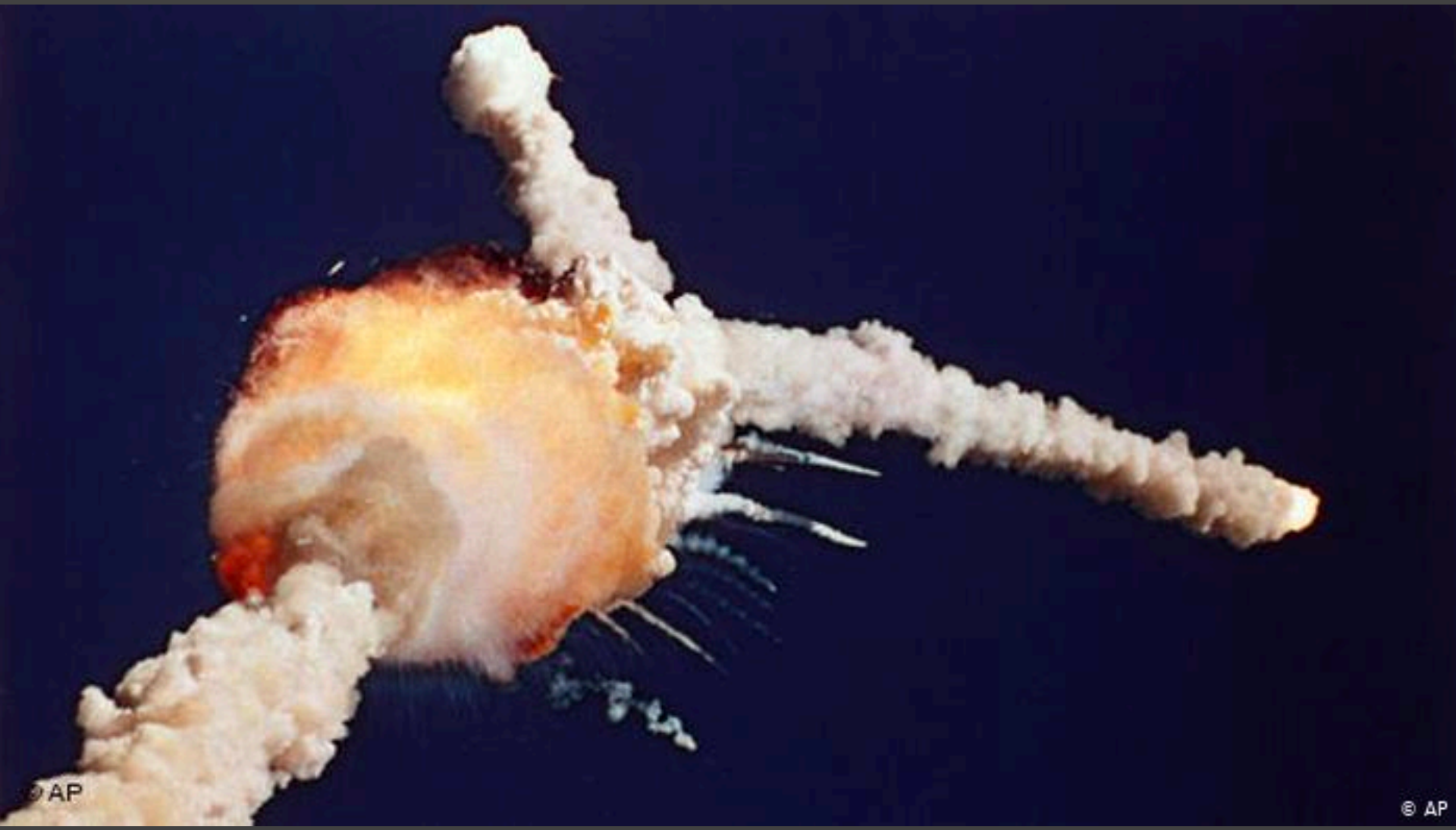


You Draw It: How Family Income Predicts Children's College Chances

[New York Times, May 28, 2015]

Support Reasoning

© AP
© AP


© AP
© AP

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36° --66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

BLOW BY HISTORY

SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

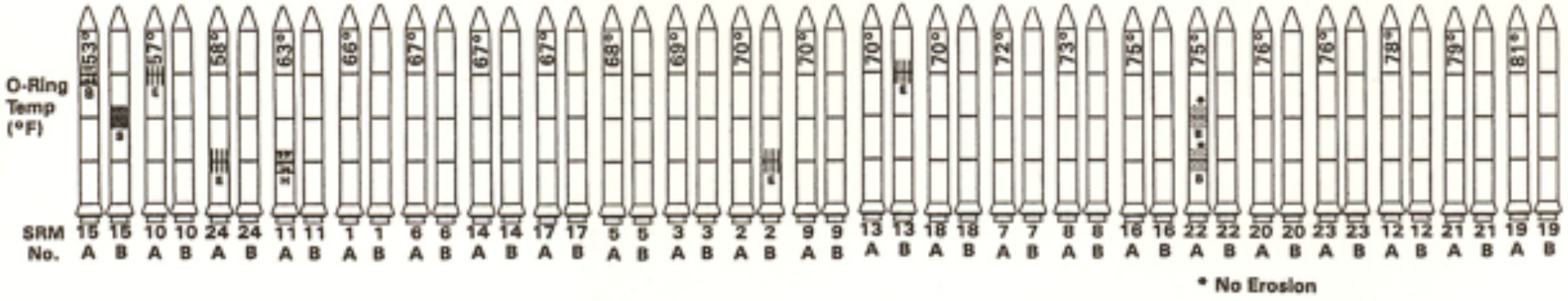
SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

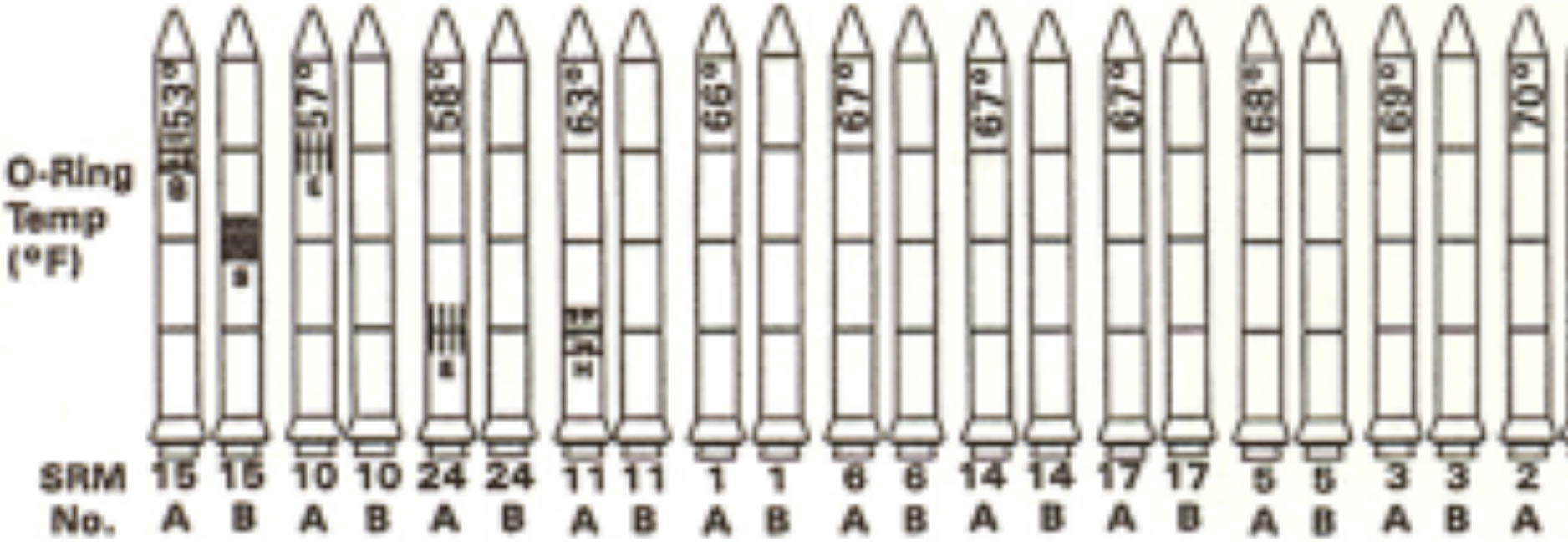
HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

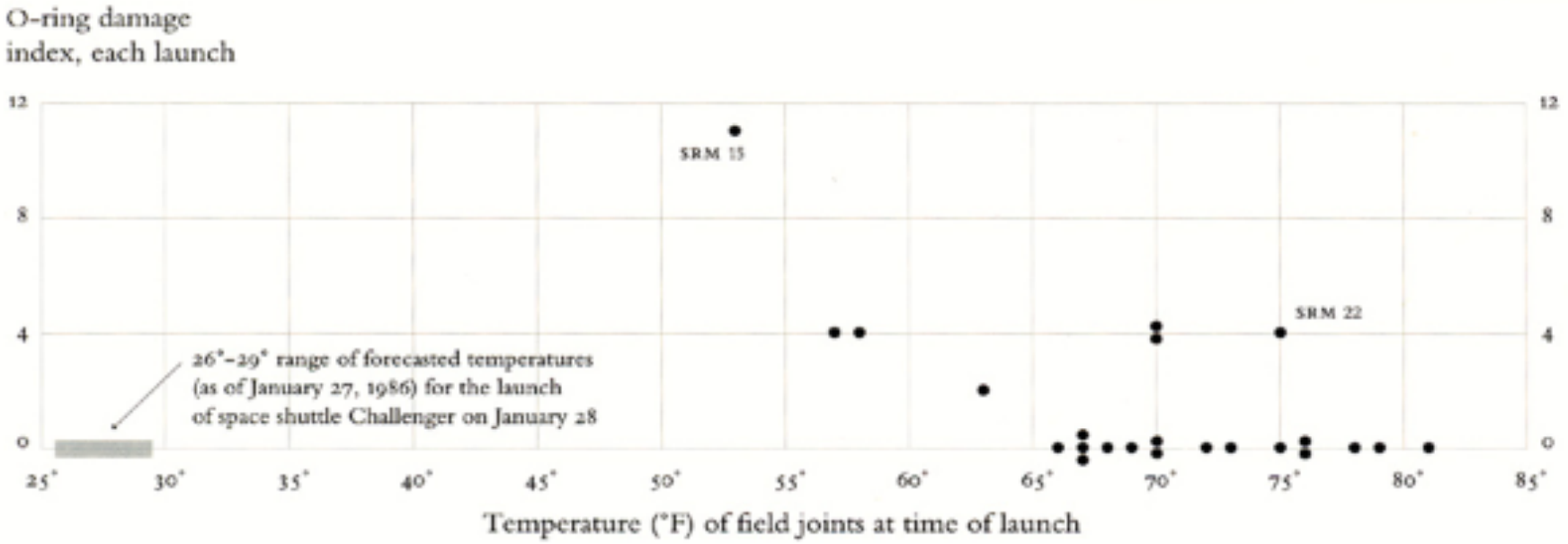
2 of 13 pages of material faxed to NASA by Morton Thiokol [from Tufte 1997]

Make Decisions: Challenger





Make Decisions: Challenger



But wait! What is an appropriate "damage index"?
Which temperatures, O-ring or outside air?

Chart of temperatures vs. O-ring damage [Tufte 97]

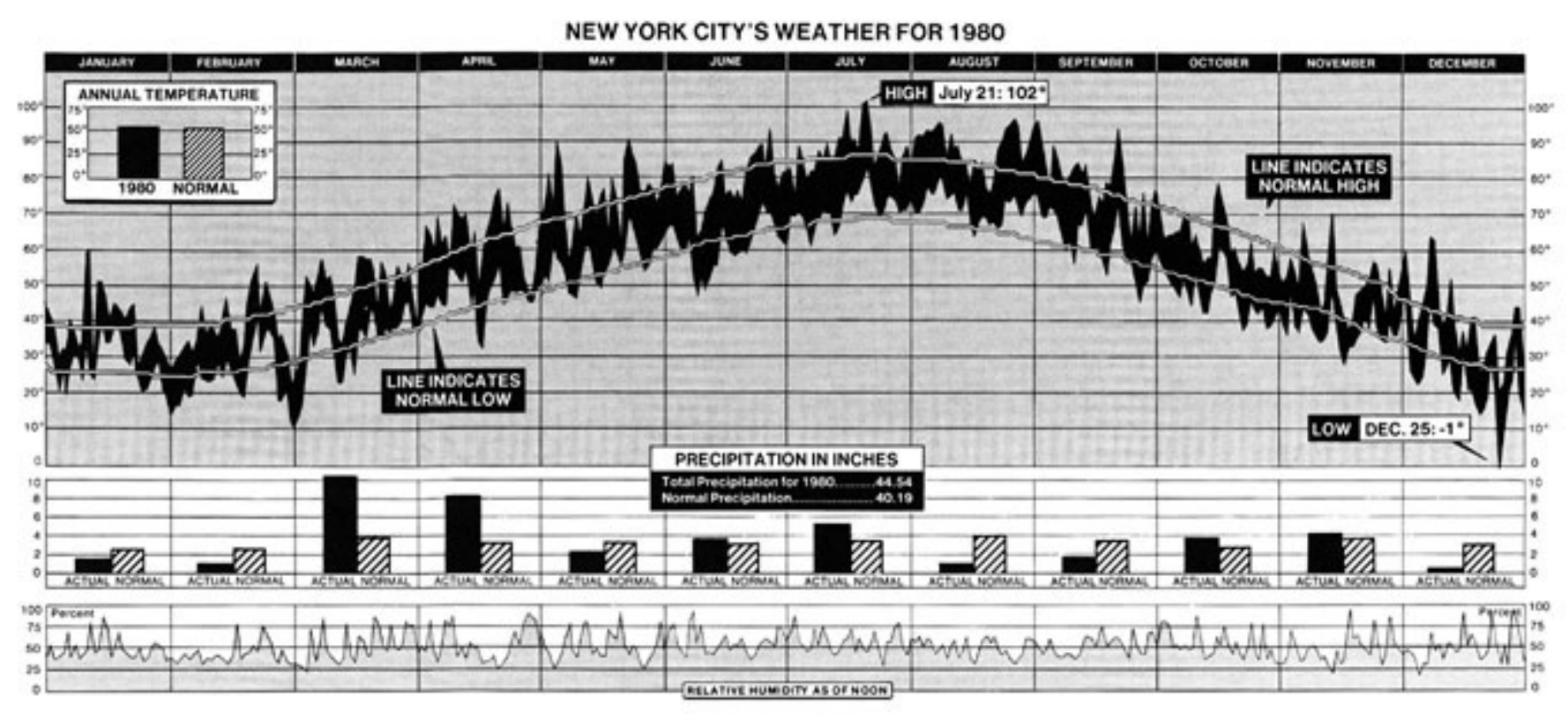
Data in Context: Cholera Outbreak

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

Data in Context: Cholera Outbreak

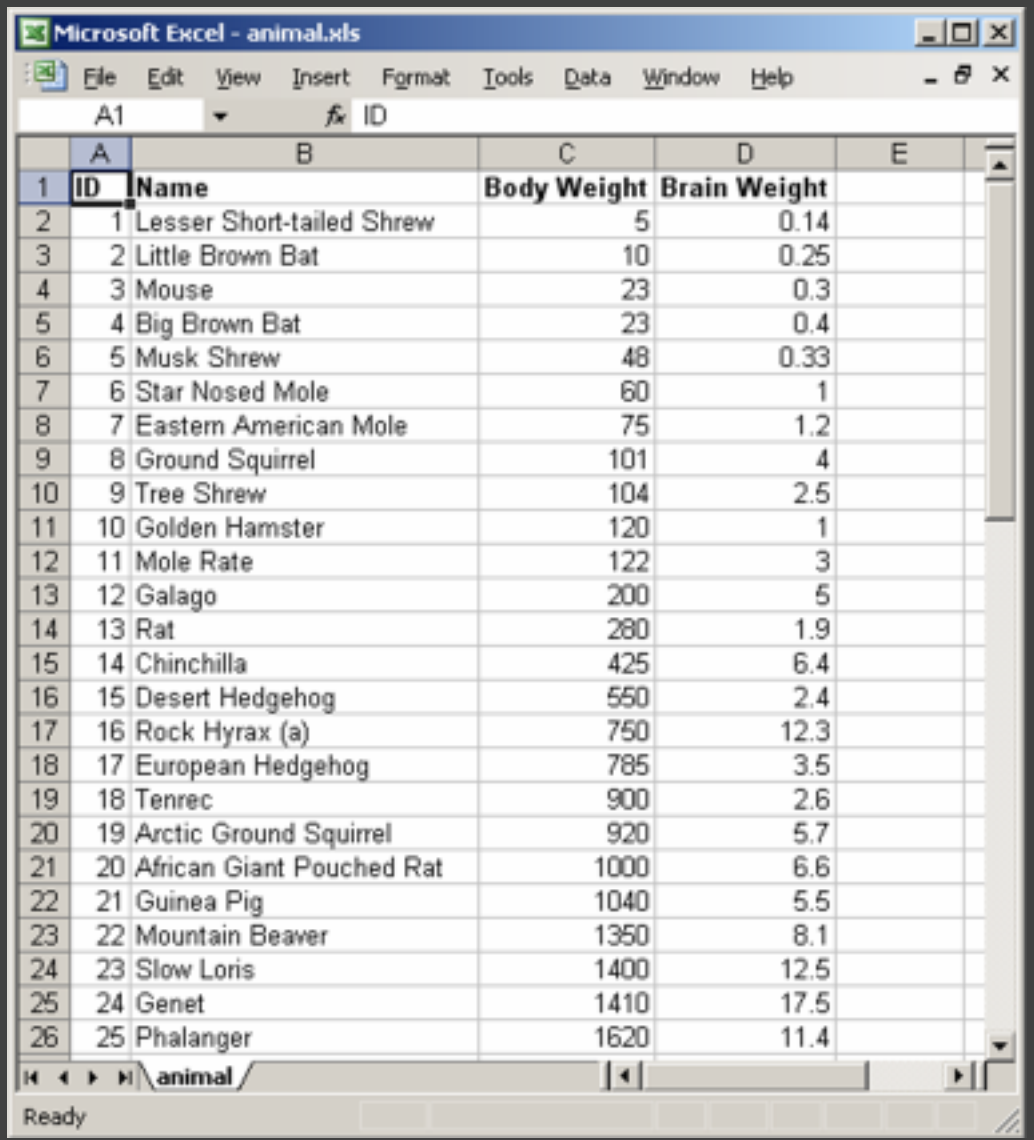
Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

Find Patterns: NYC Weather



[New York Times 1981]

Answer Questions: Brain Power?


Microsoft Excel - animal.xls

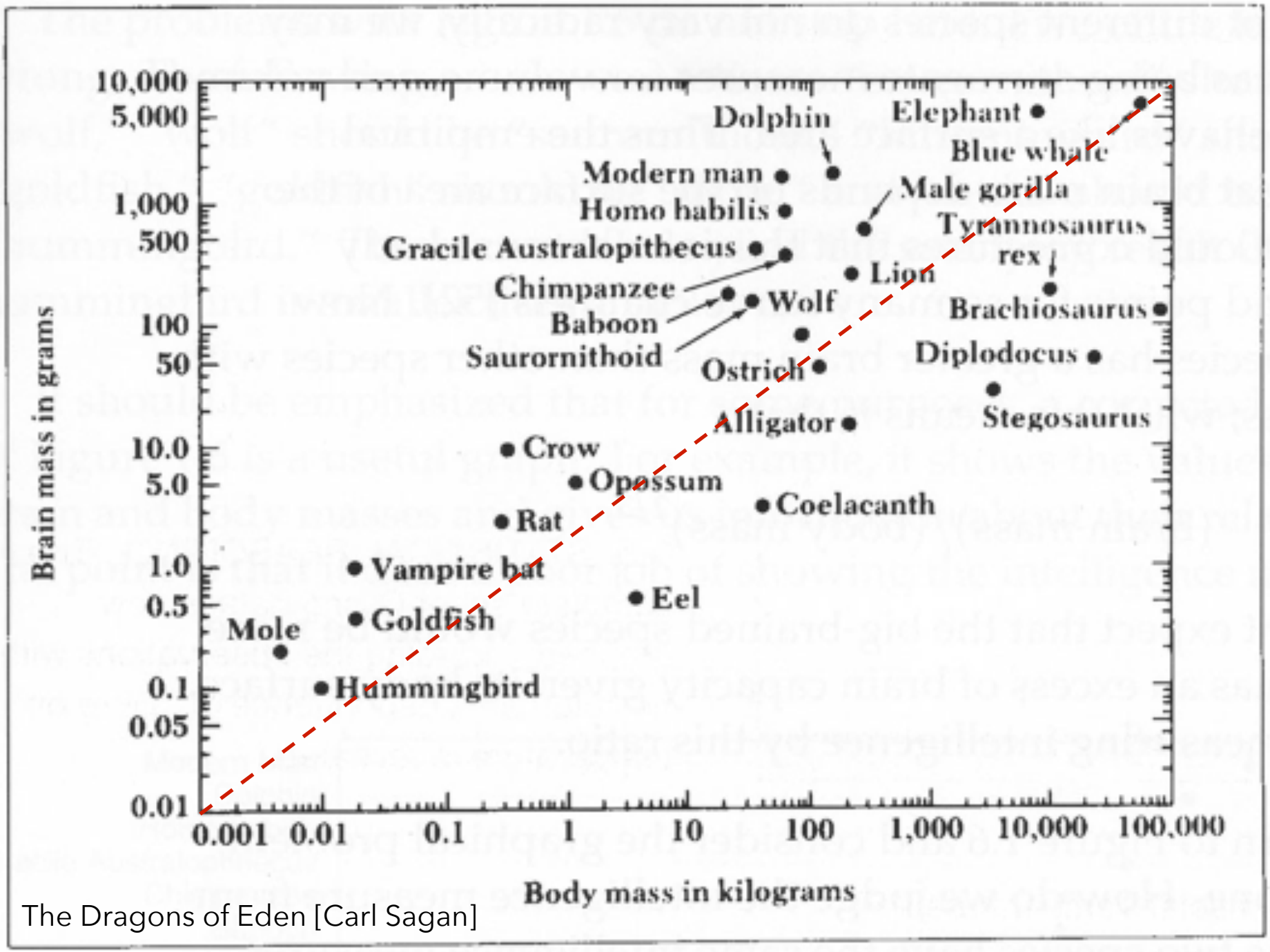
File Edit View Insert Format Tools Data Window Help

A1 ID

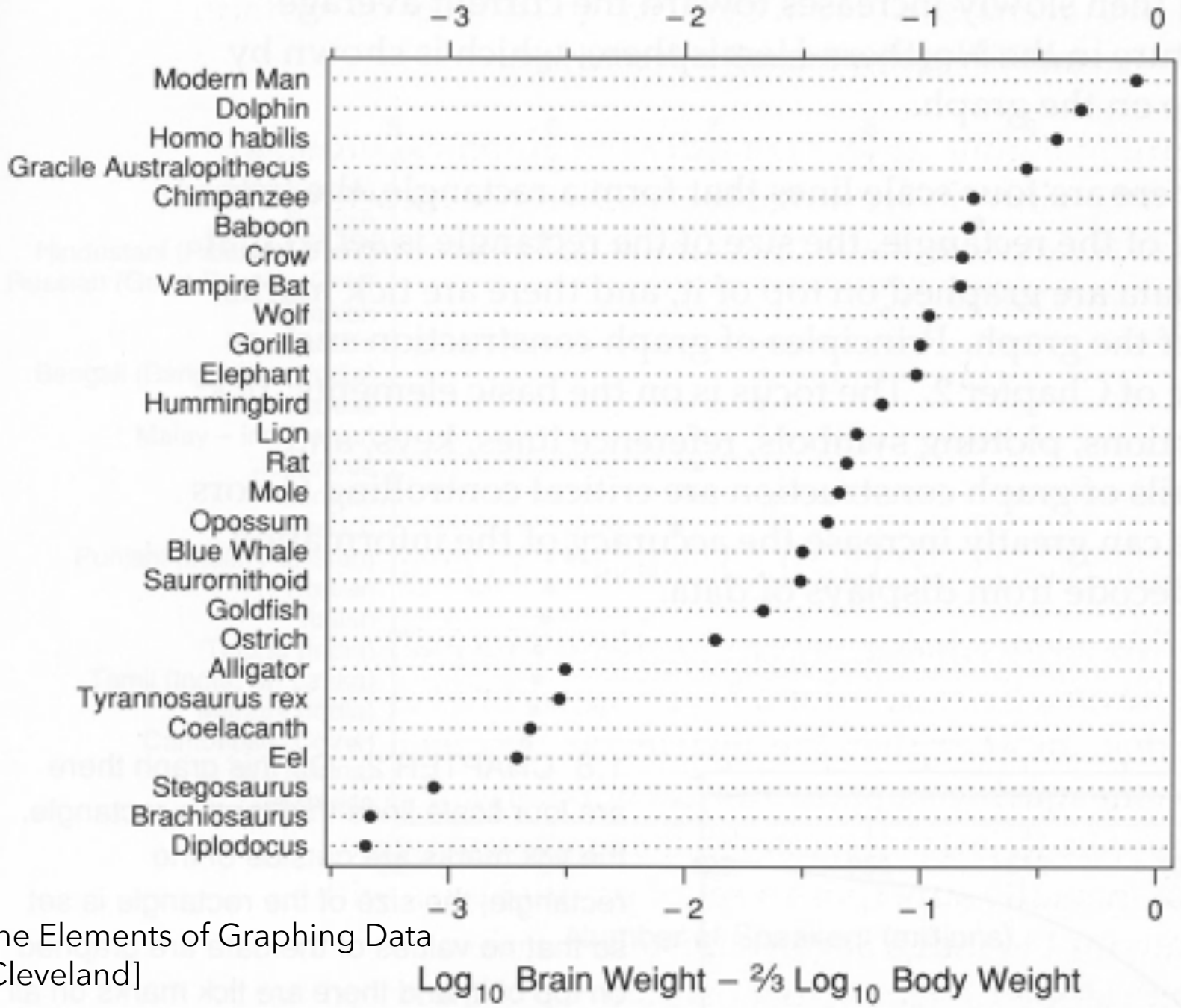
	A	B	C	D	E
1	ID	Name	Body Weight	Brain Weight	
2	1	Lesser Short-tailed Shrew	5	0.14	
3	2	Little Brown Bat	10	0.25	
4	3	Mouse	23	0.3	
5	4	Big Brown Bat	23	0.4	
6	5	Musk Shrew	48	0.33	
7	6	Star Nosed Mole	60	1	
8	7	Eastern American Mole	75	1.2	
9	8	Ground Squirrel	101	4	
10	9	Tree Shrew	104	2.5	
11	10	Golden 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	

animal

Ready

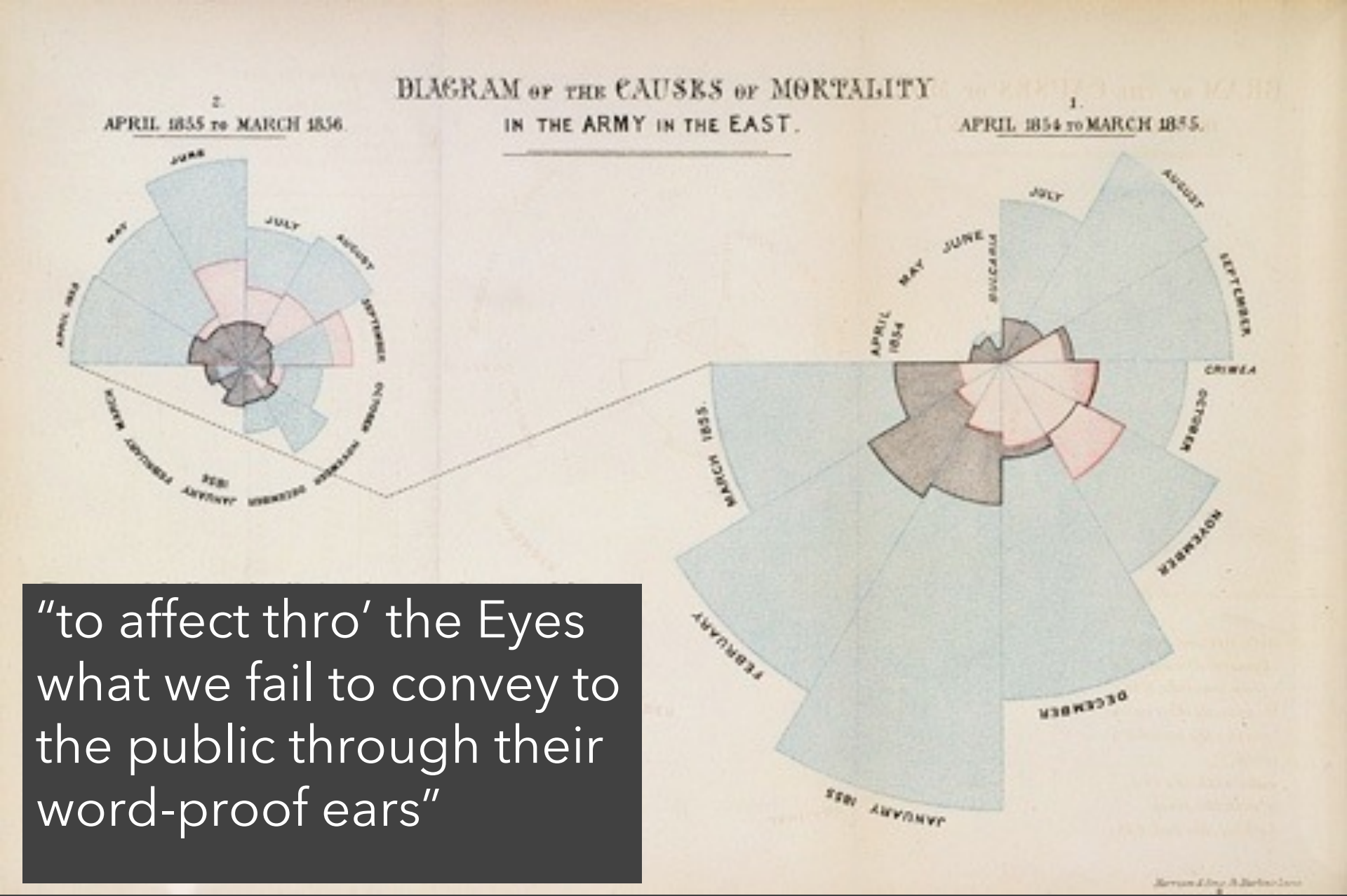


The Dragons of Eden [Carl Sagan]



The Elements of Graphing Data
[Cleveland]

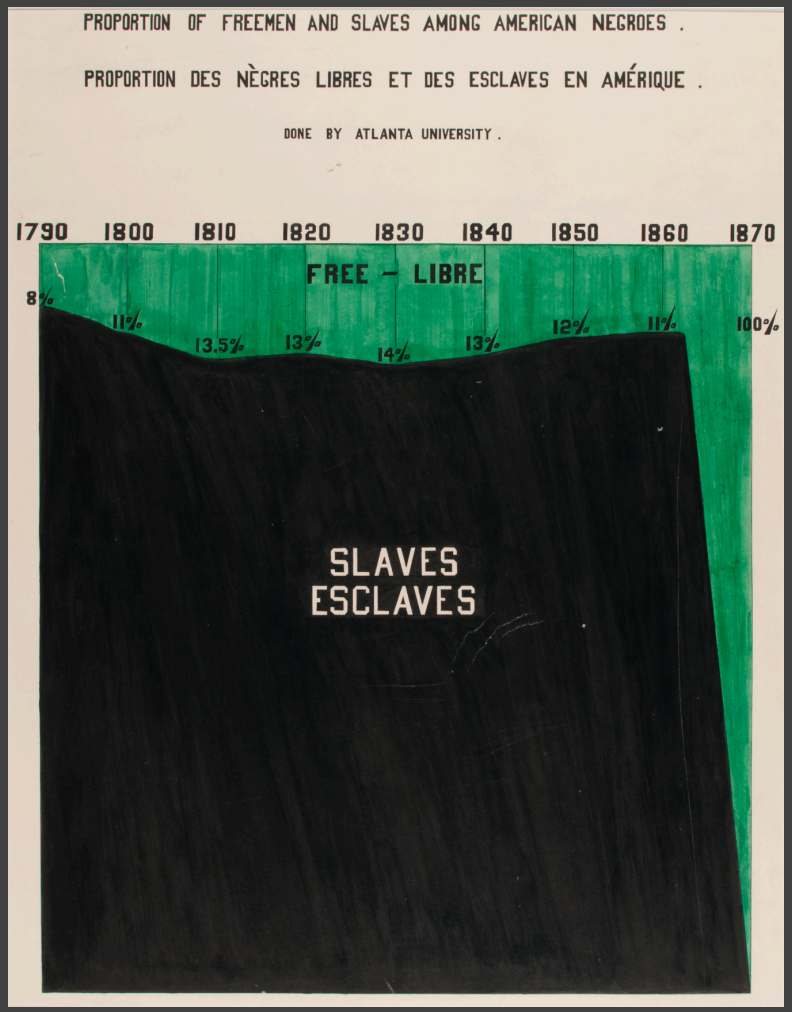
Convey Information



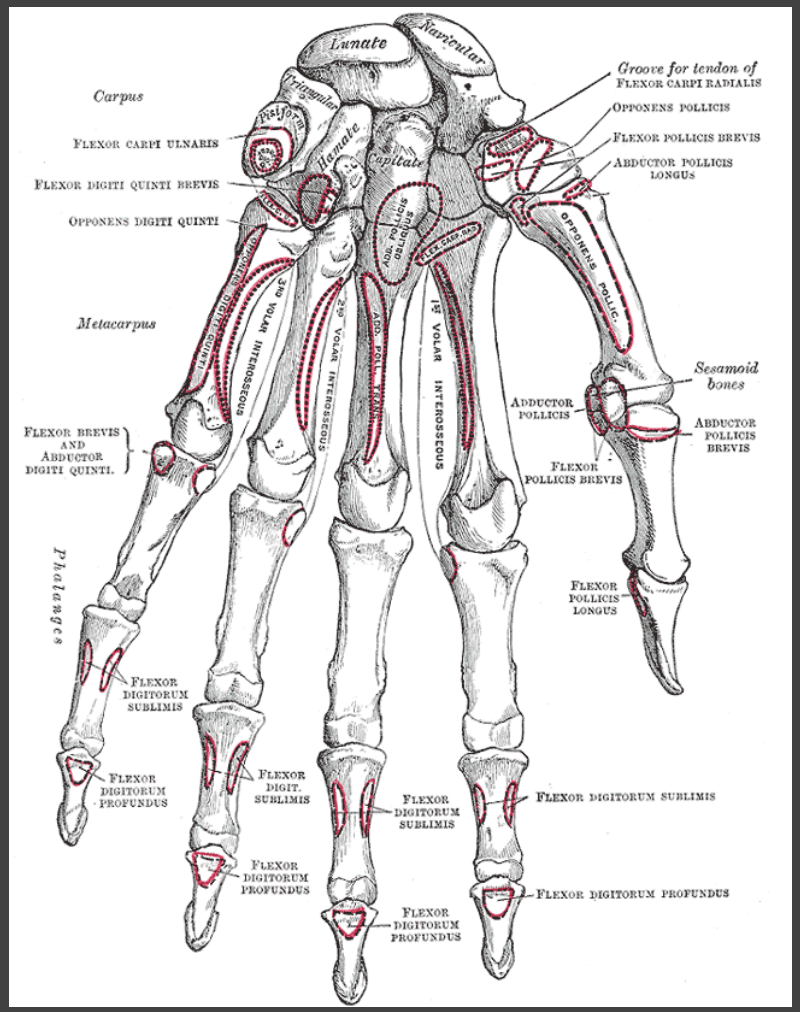
"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

Communicate, Inform, Inspire



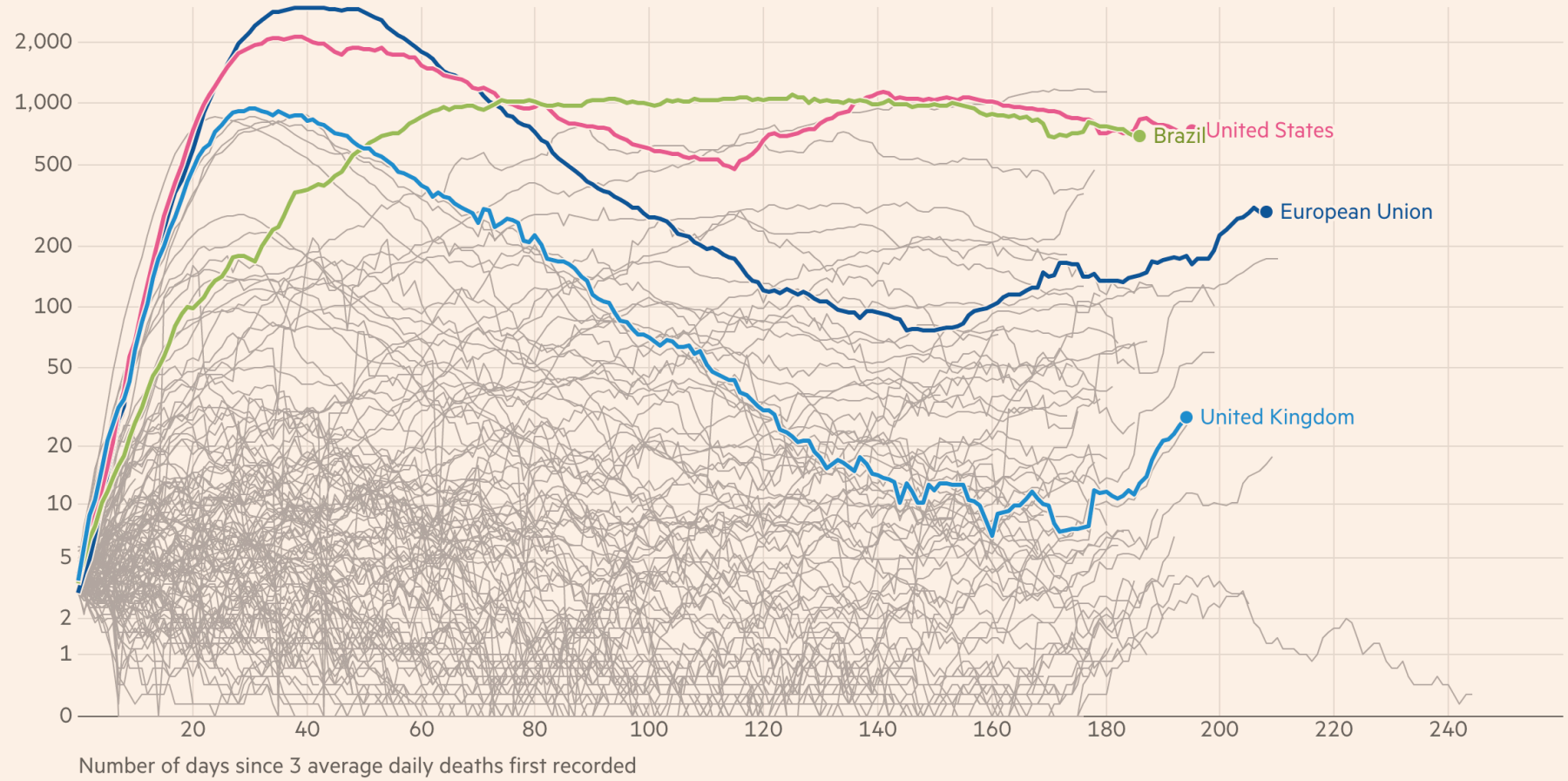
Visualizing Black America, Du Bois et al. 1900



Bones in hand, Gray's Anatomy 1918 ed.

New deaths attributed to Covid-19 in European Union, United States, Brazil and United Kingdom

Seven-day rolling average of new deaths, by number of days since 3 average daily deaths first recorded



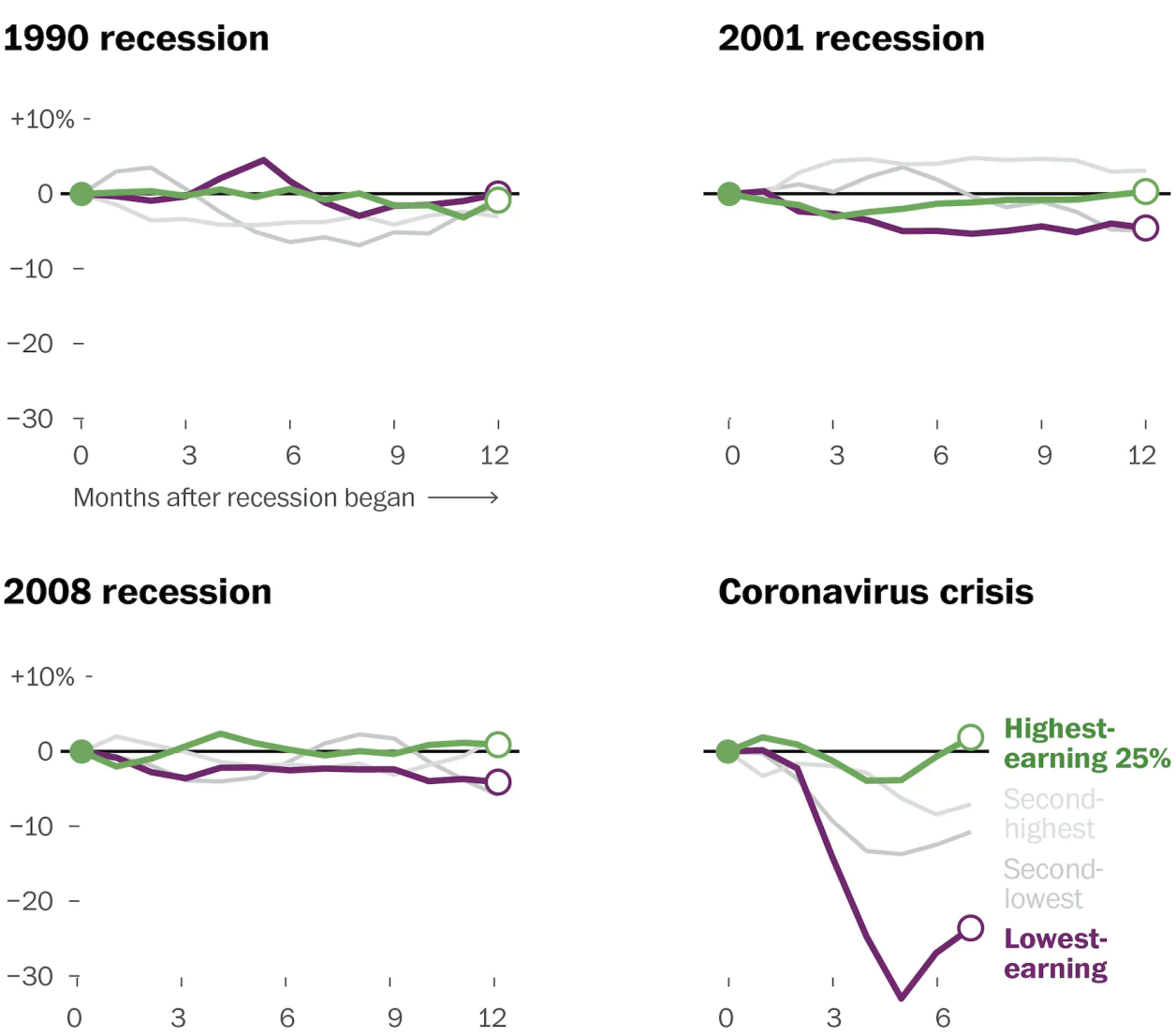
Source: Financial Times analysis of data from the European Centre for Disease Prevention and Control, the Covid Tracking Project, the UK Dept of Health & Social Care and the Spanish Ministry of Health.
Data updated September 25 2020 12.46pm BST. Interactive version: ft.com/covid19

FINANCIAL TIMES

Coronavirus Tracked John Burn-Murdoch & Financial Times

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

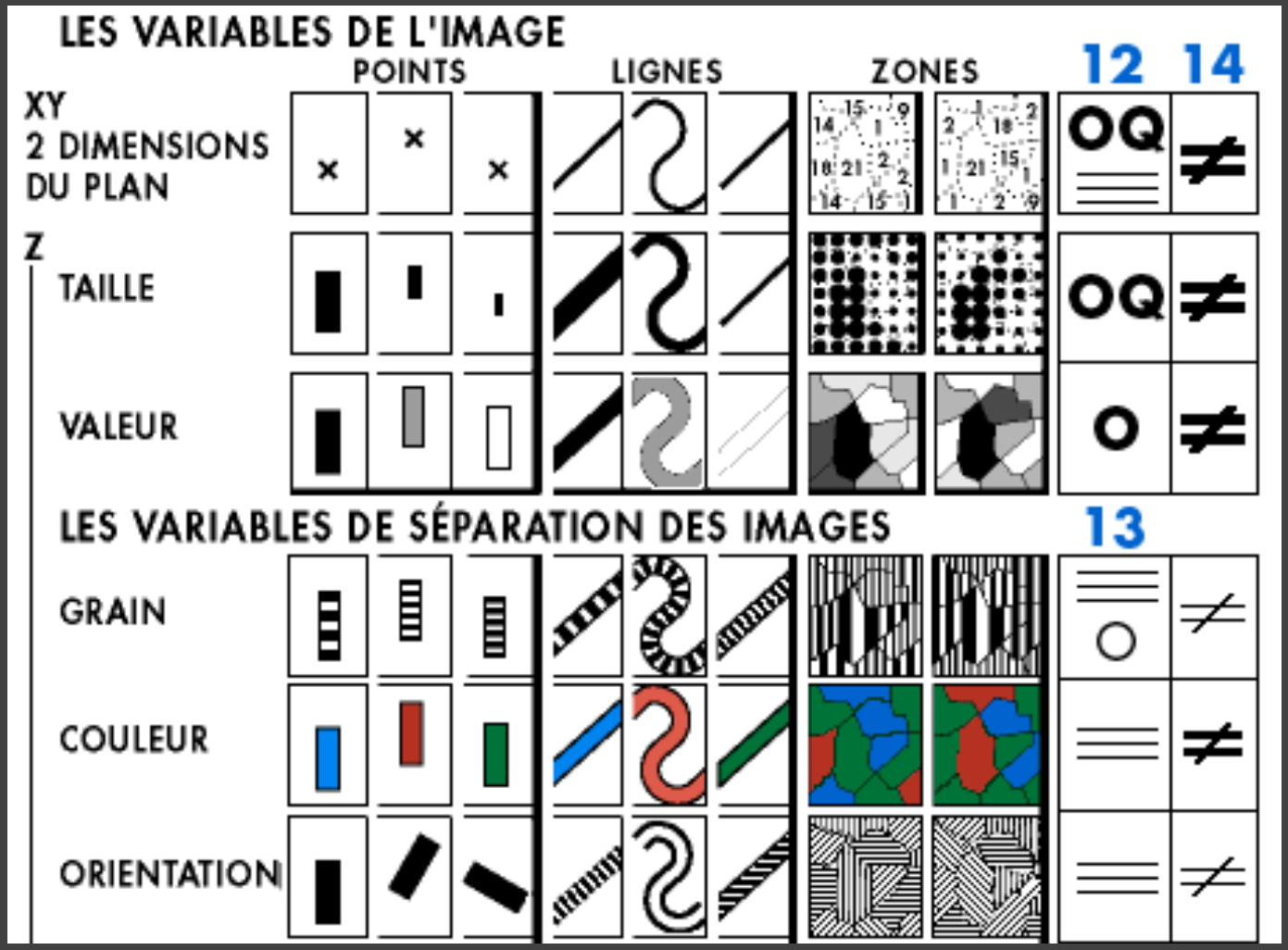
   What do people perceive / comprehend?
   How do visualizations inform mental models?

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

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

Course Topics

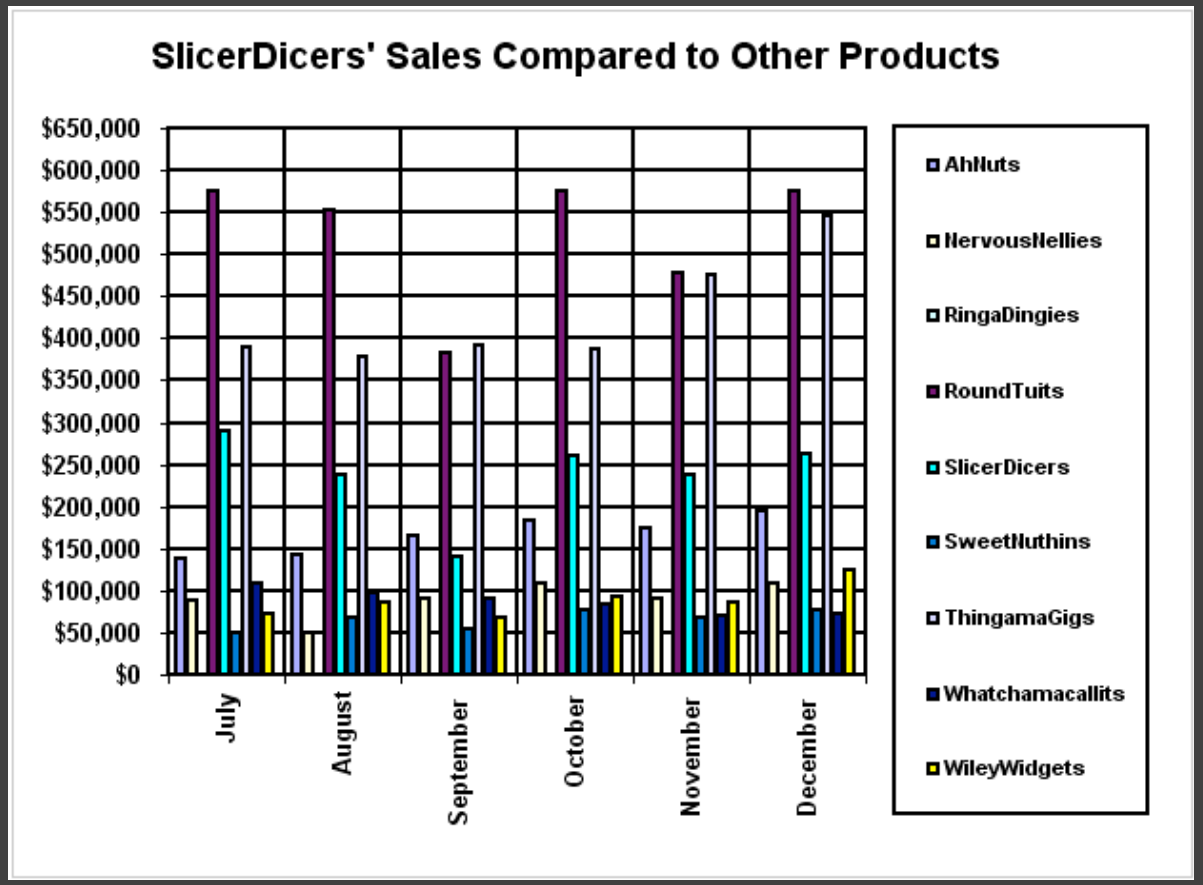
Data and Image Models



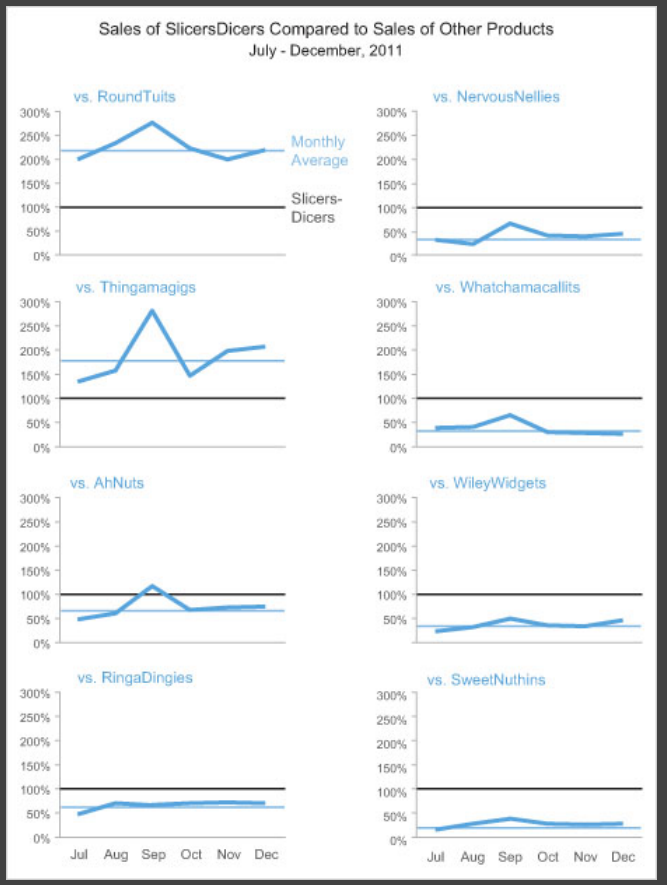
Sémiologie Graphique [Bertin 67]

Visualization Design

Respond here: pollev.com/leibatt

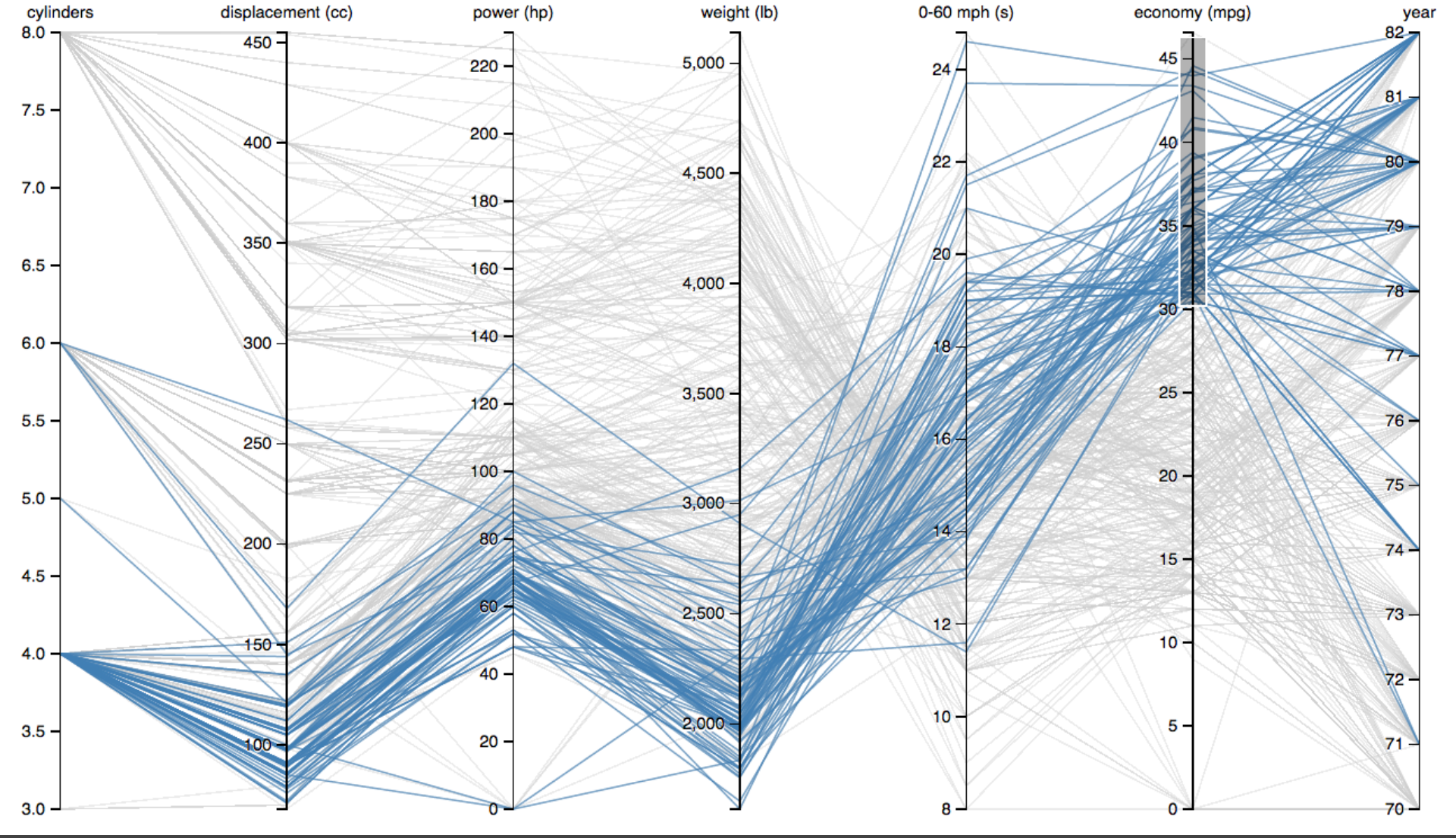


Problematic design

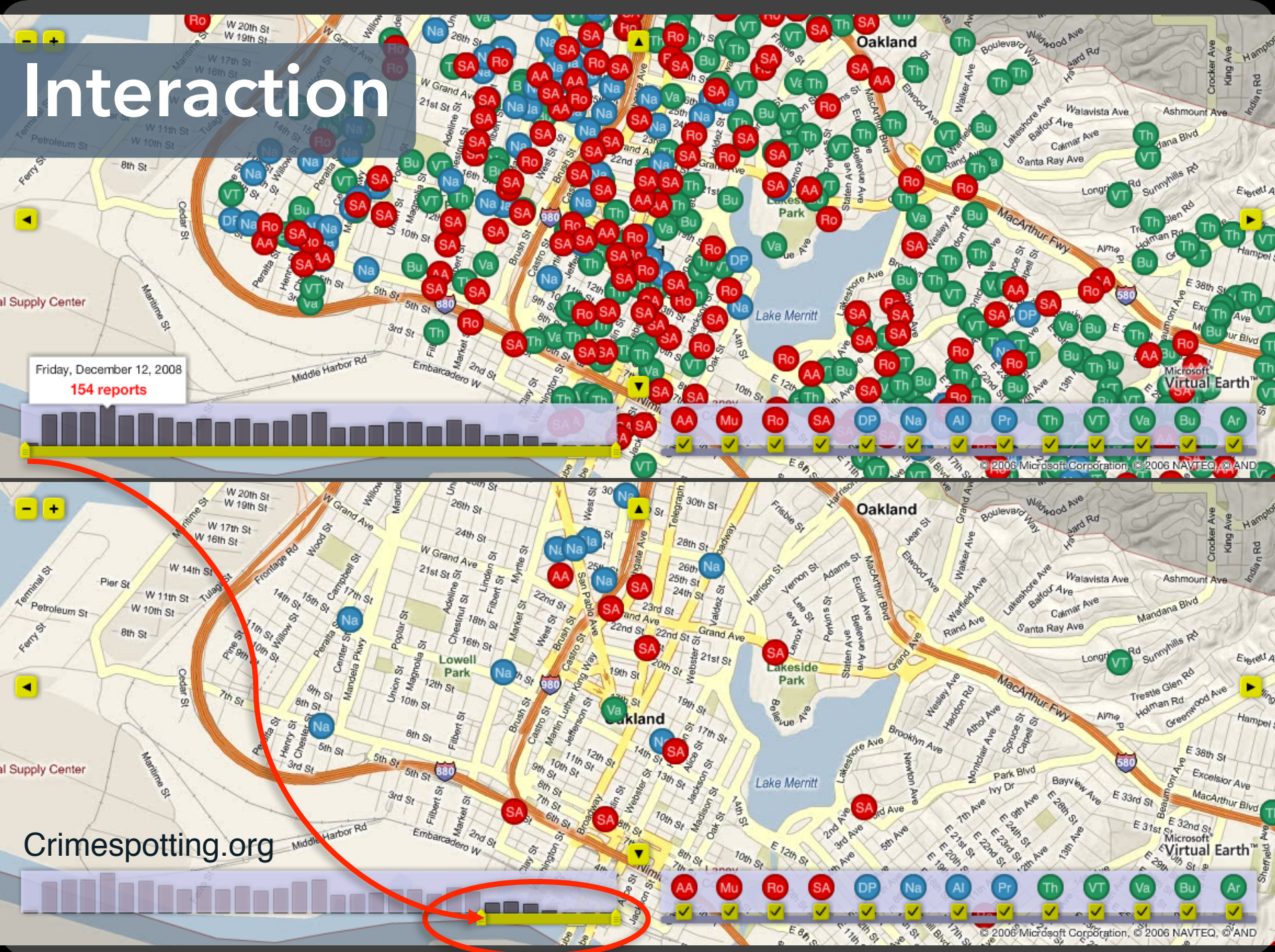


Redesign

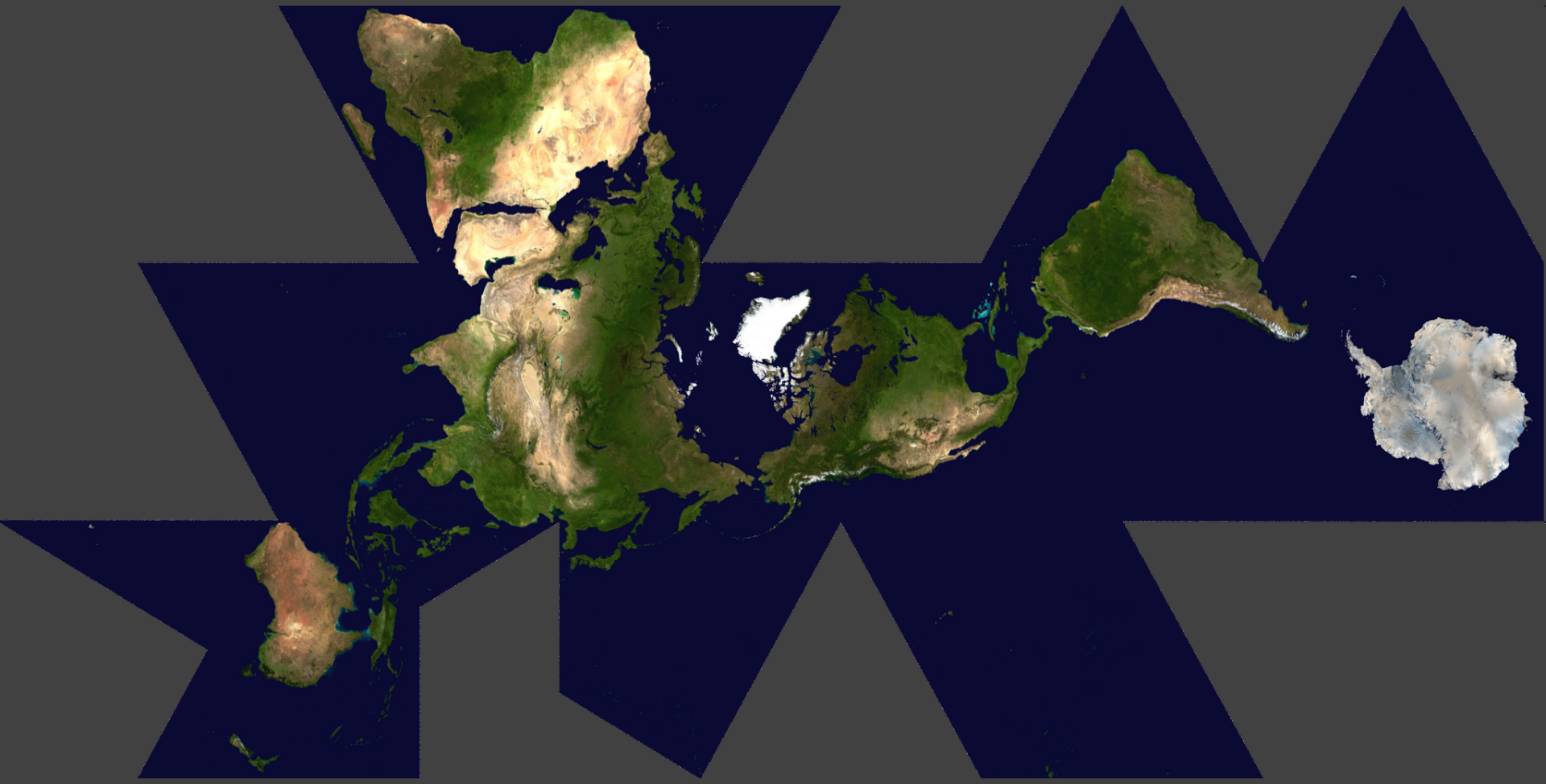
Exploratory Data Analysis



Interaction

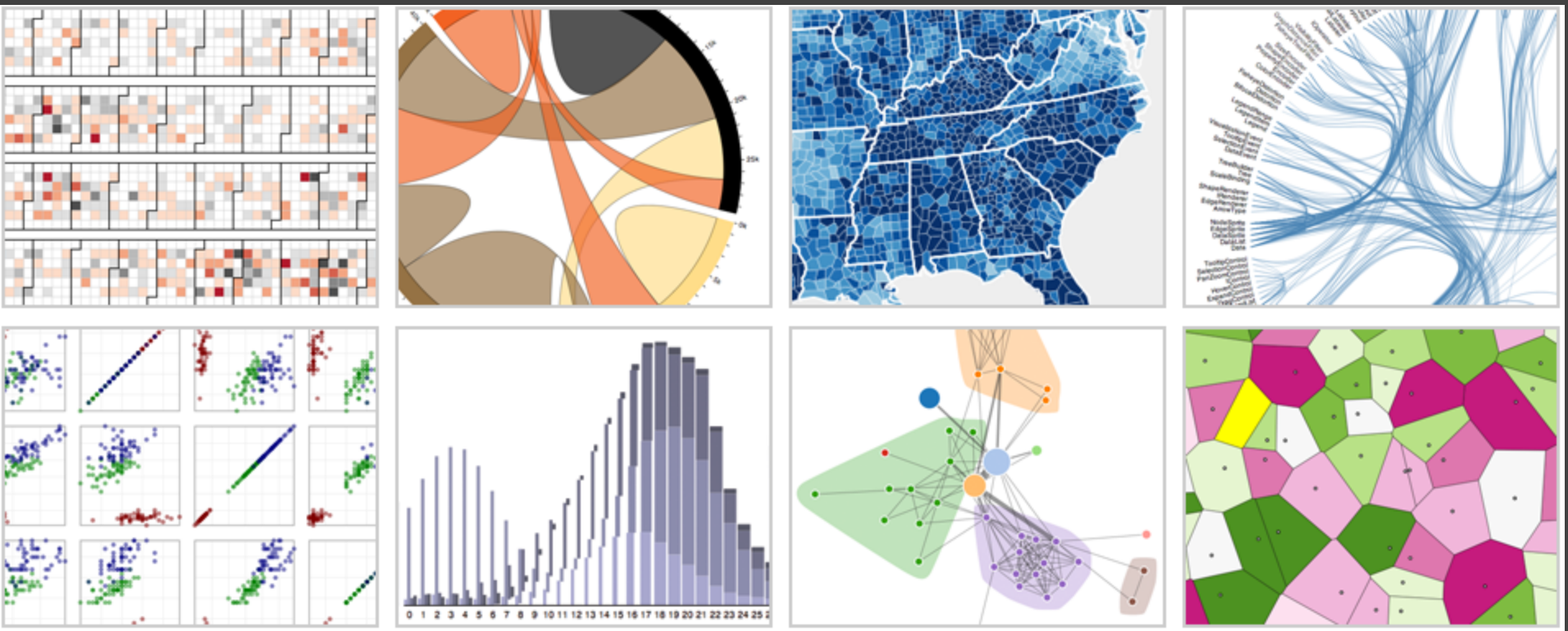


Maps



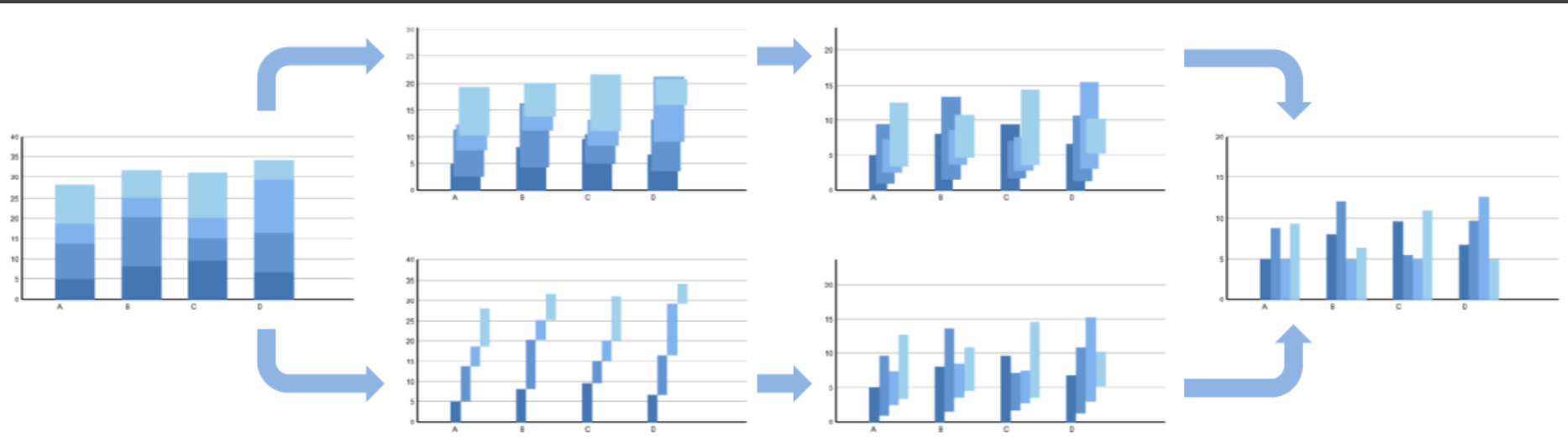
Dymaxion Maps [Fuller 46]

Visualization Software



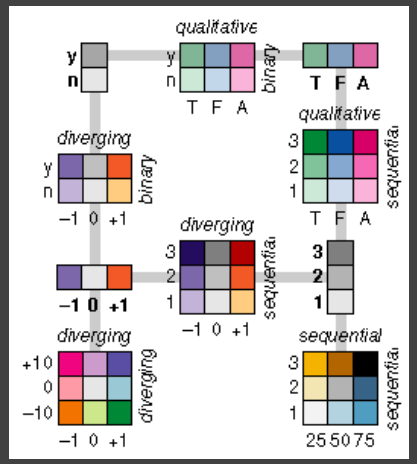
D3: Data-Driven Documents
Vega-Lite / Altair

Animation

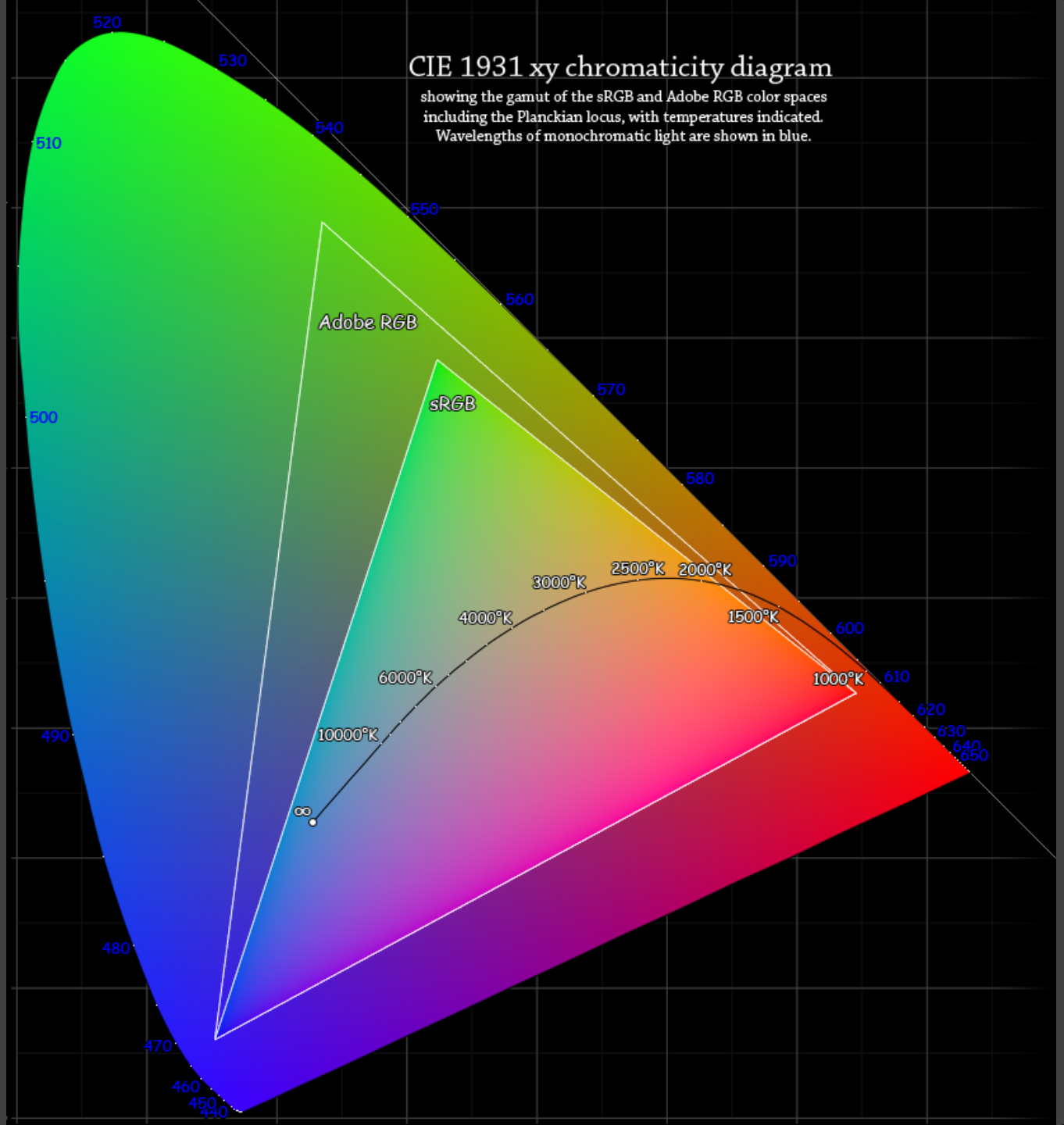


Animated transitions in statistical data graphics [Heer & Robertson 07]

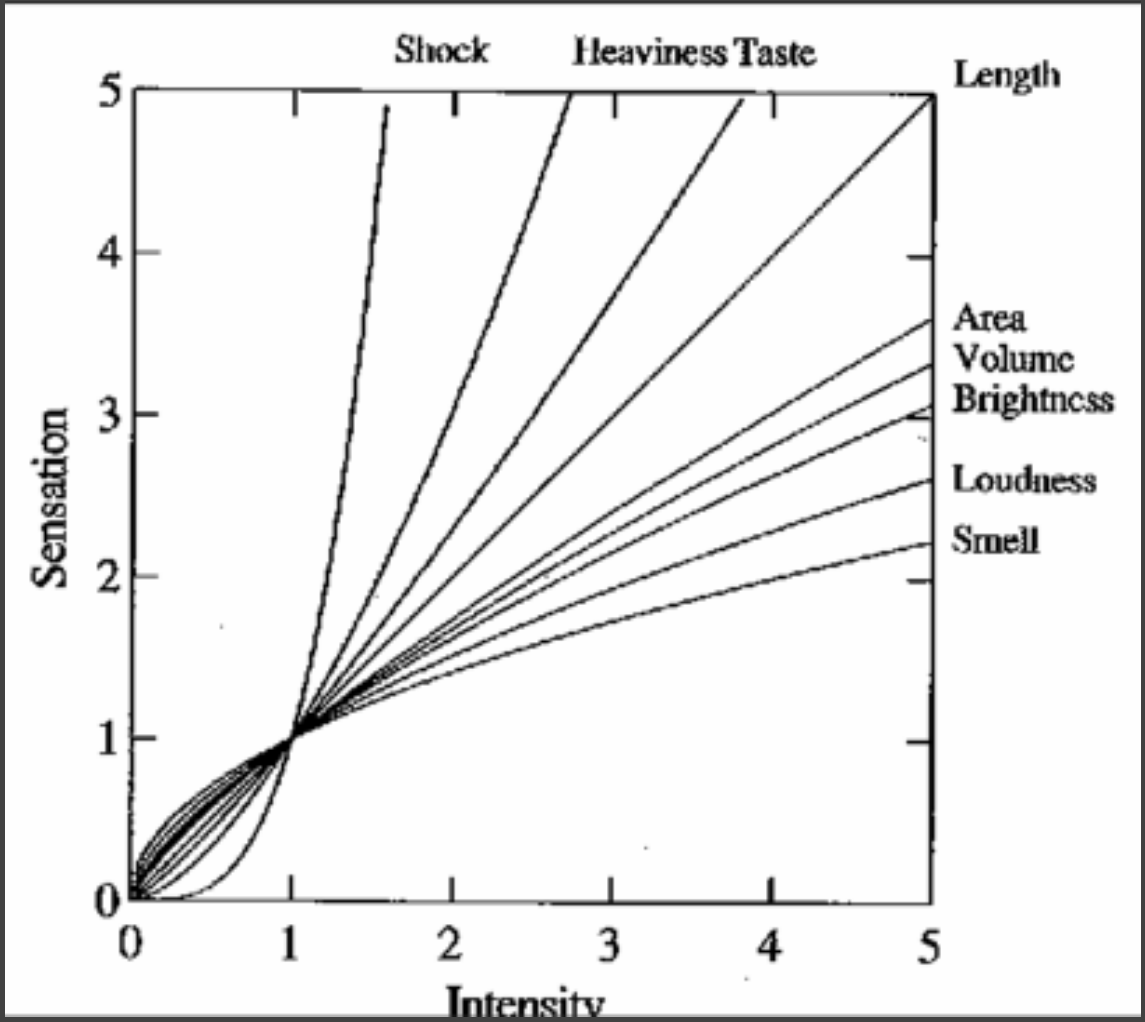
Color



Color Brewer

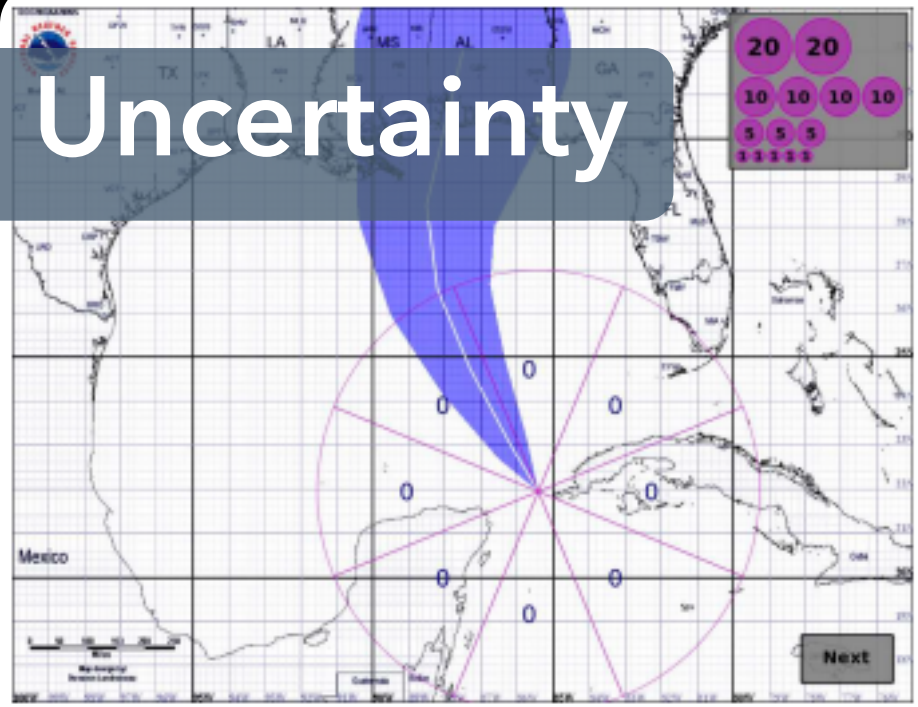


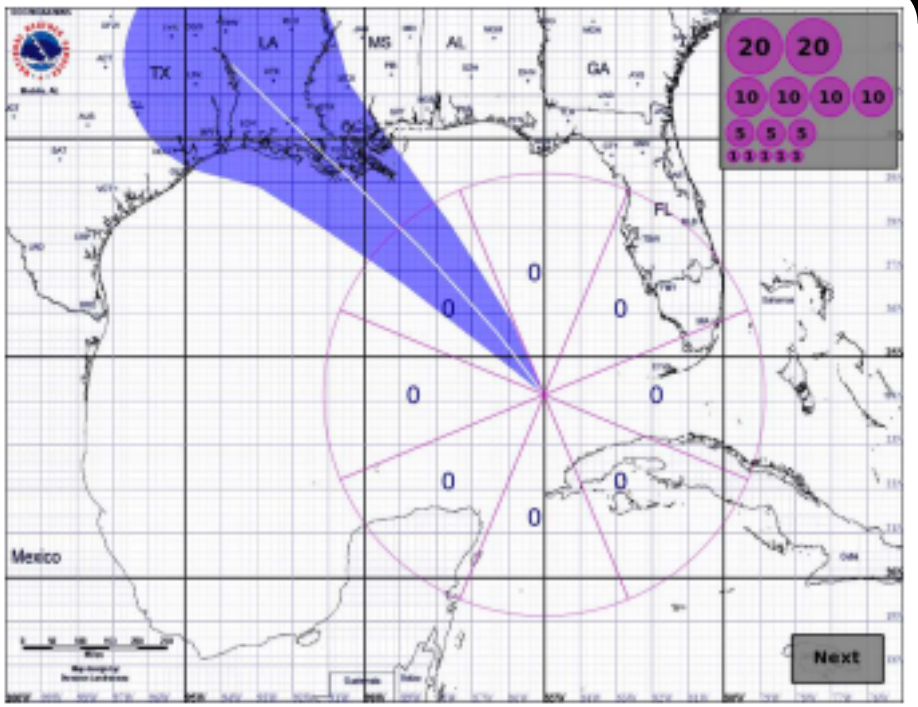
Graphical Perception

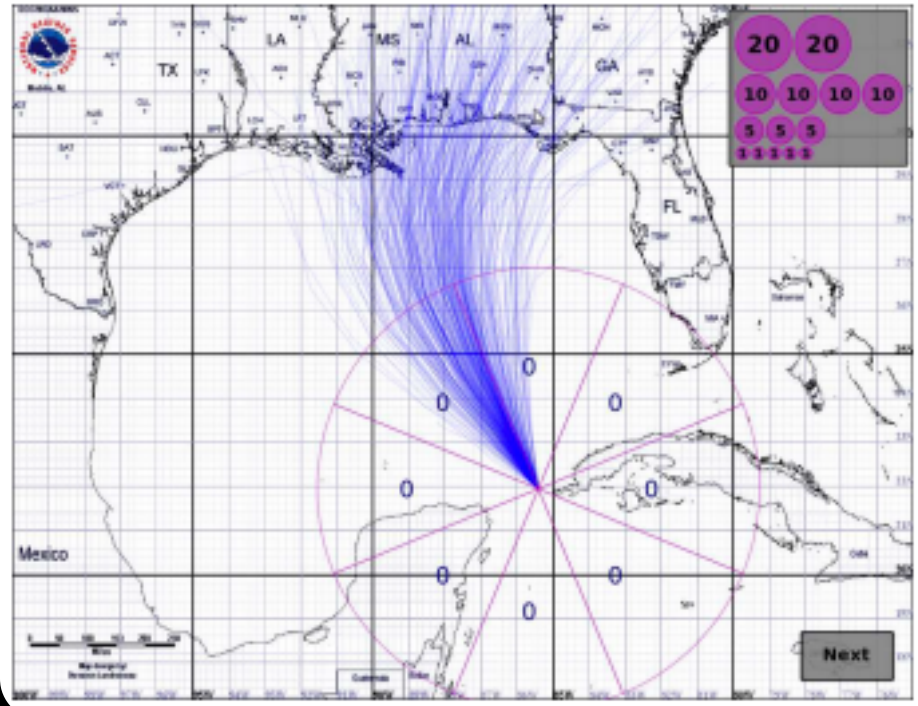


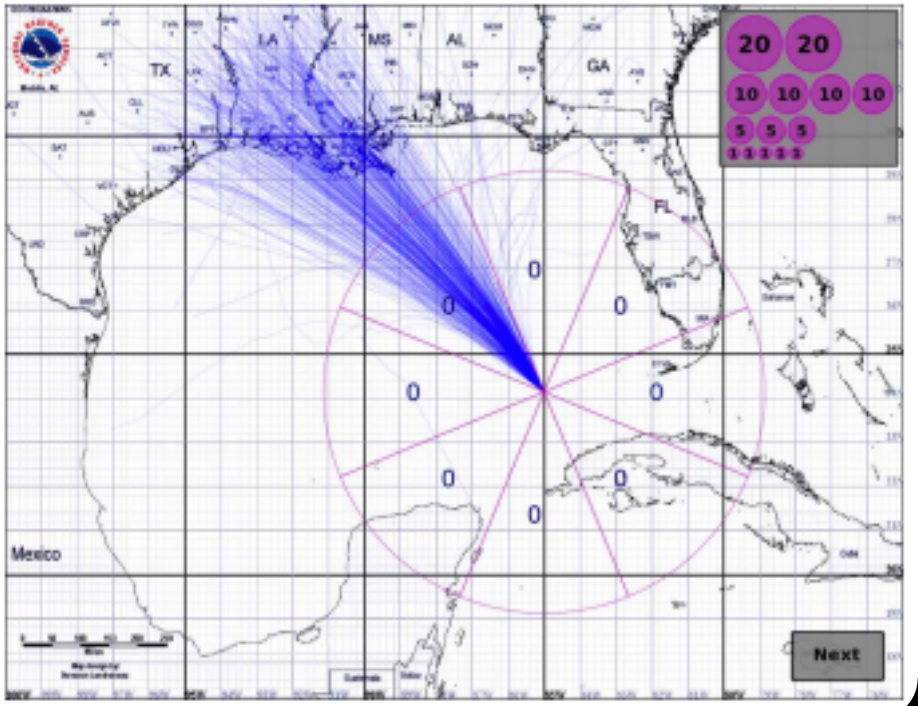
The psychophysics of sensory function [Stevens 61]

Uncertainty

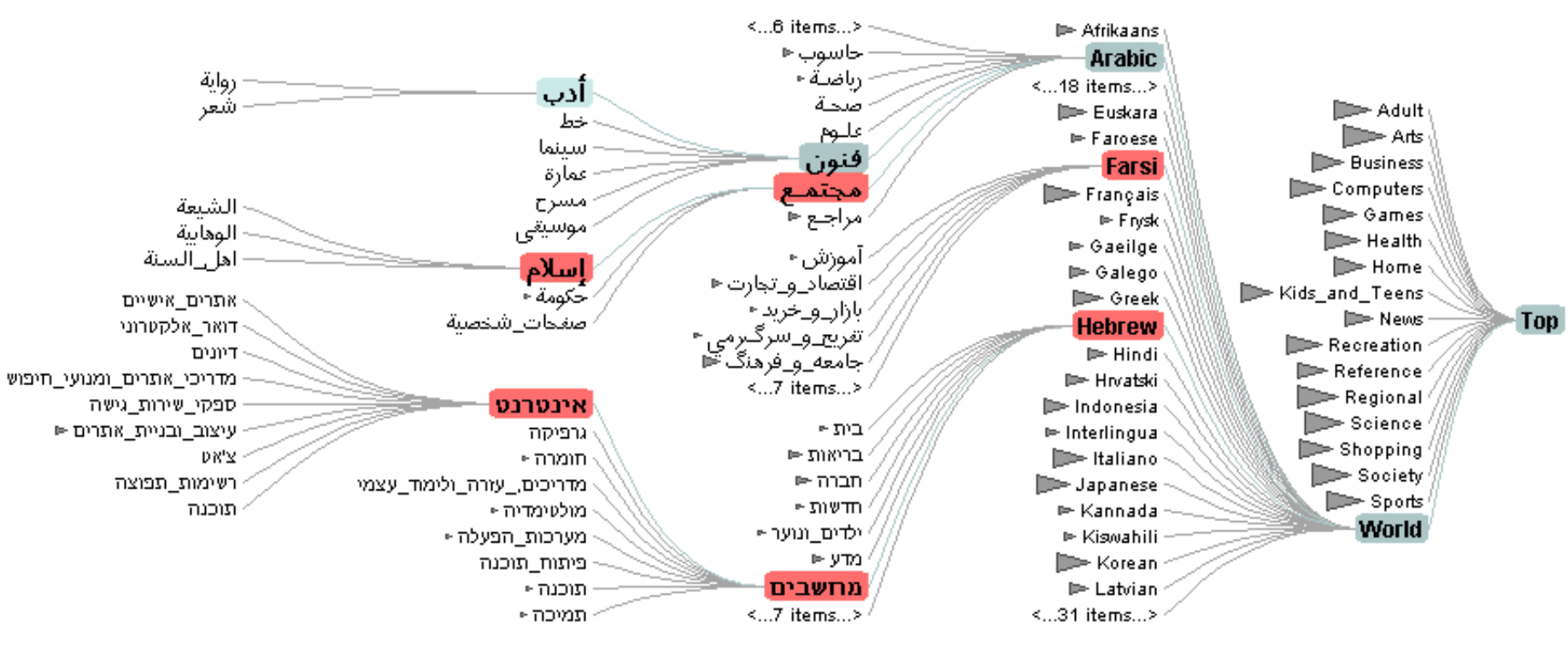






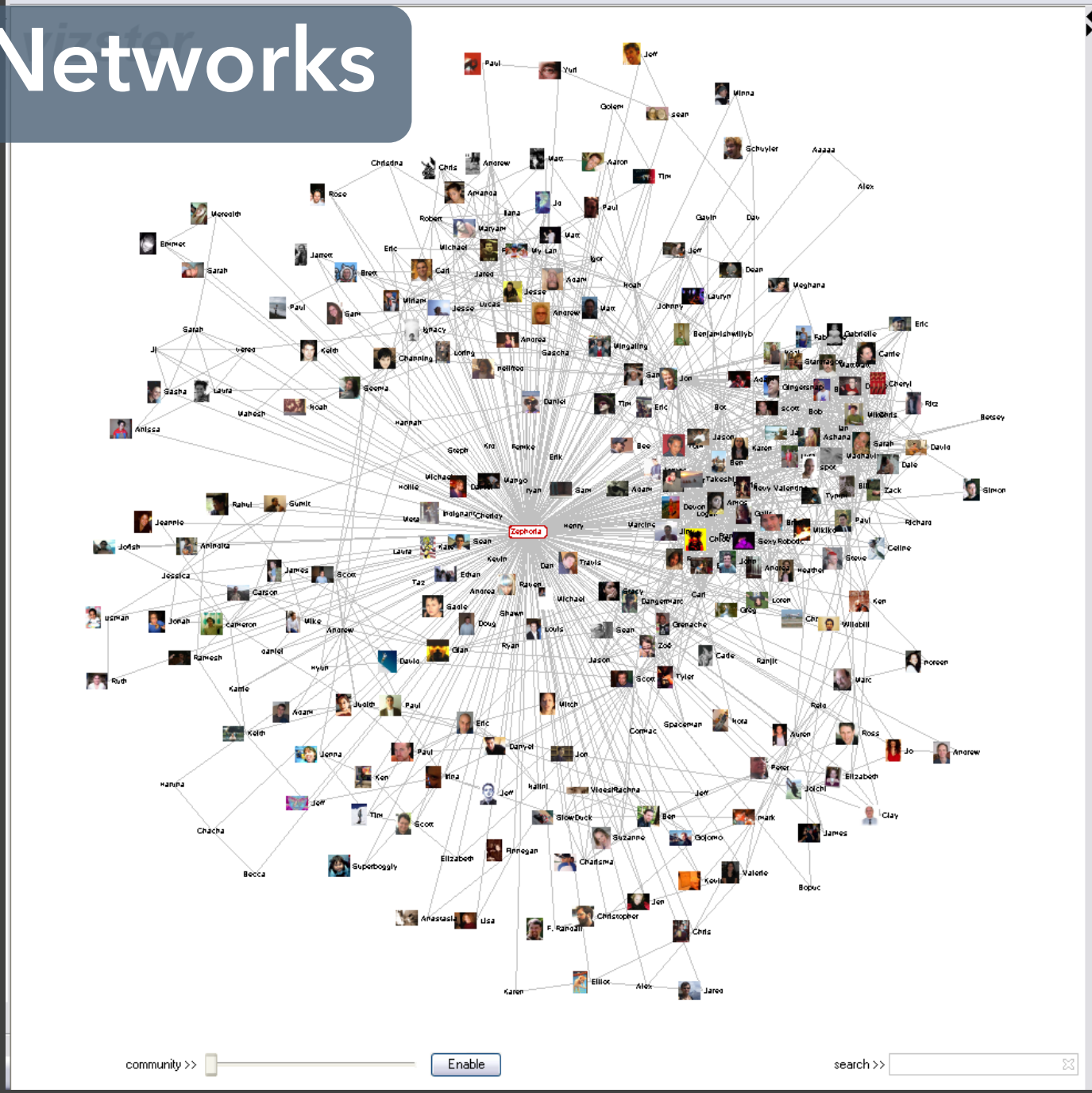


Hierarchies



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

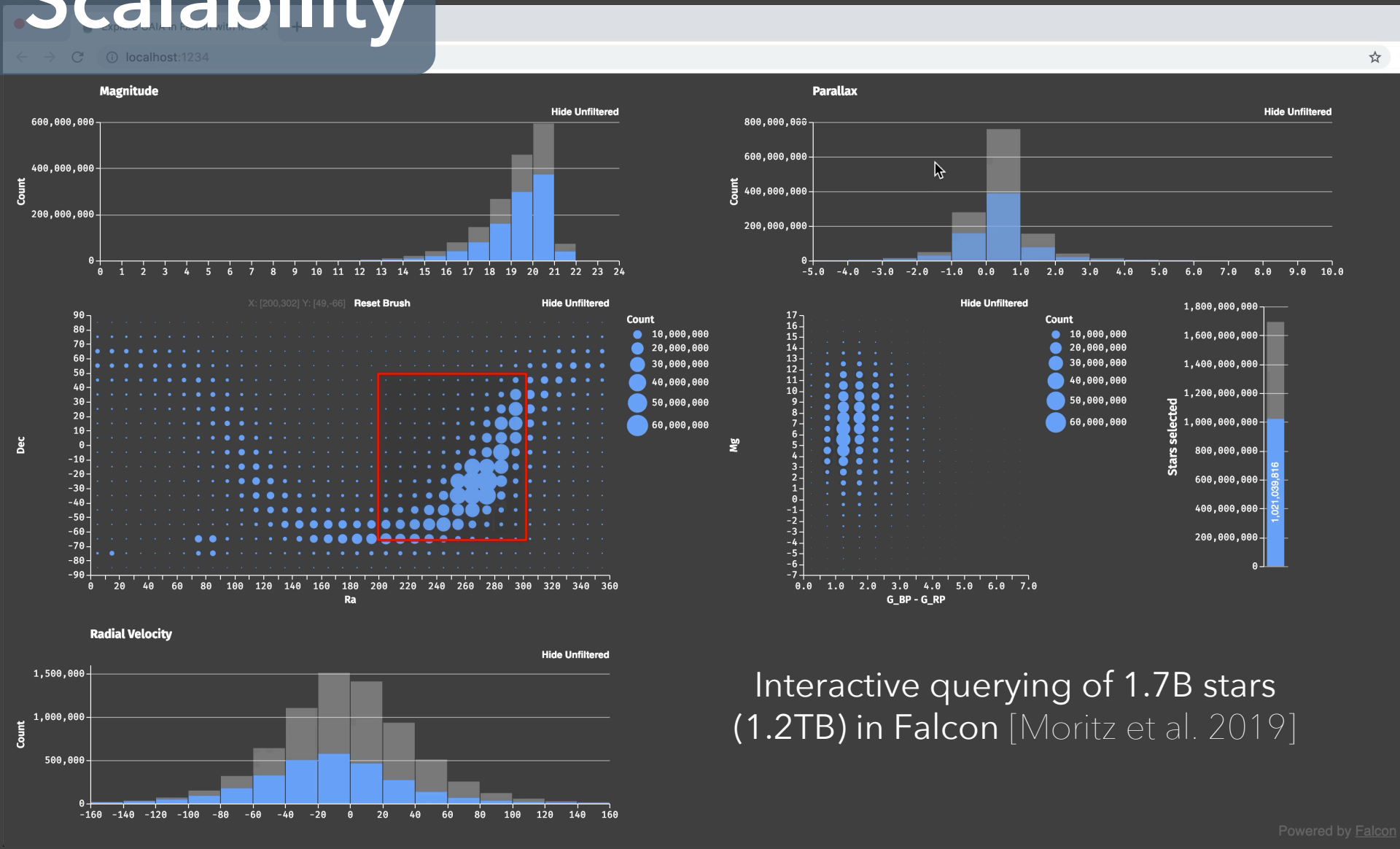
Networks



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]

Powered by Falcon

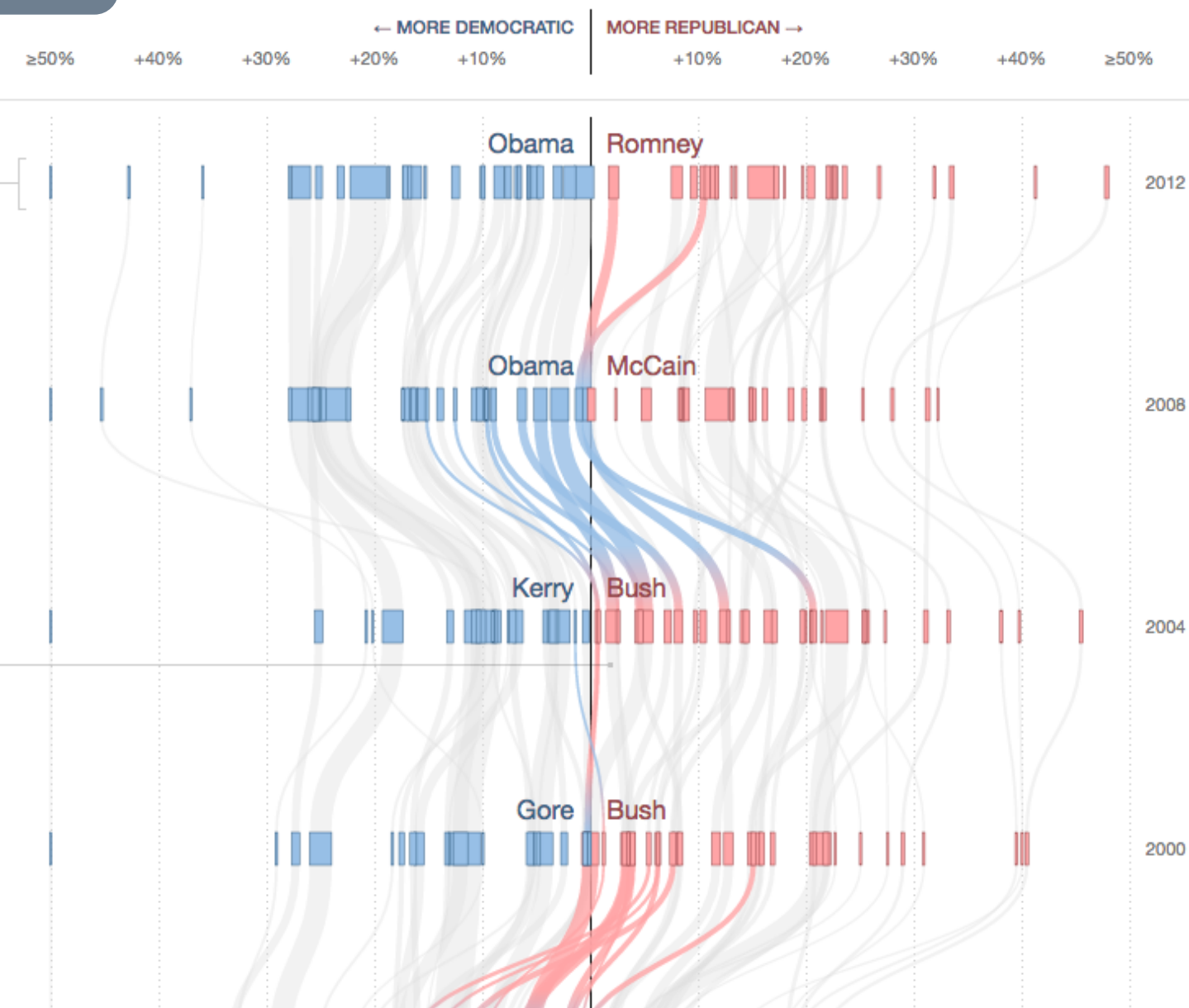
Narrative

Recent elections have placed a heavy emphasis on "swing states" — Ohio, Florida and the other competitive states. Y[illegible] between the Democratic and Republican parties. A look at how the states stacked up in the [illegible] 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**



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 Data and Image Models lecture videos before class on Thursday!

Tues = Lectures. Thurs = in-class activities.

All Tues lectures will be in-person + recorded. We will use PollEV to track engagement.

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

From books, notebooks, and linked articles.

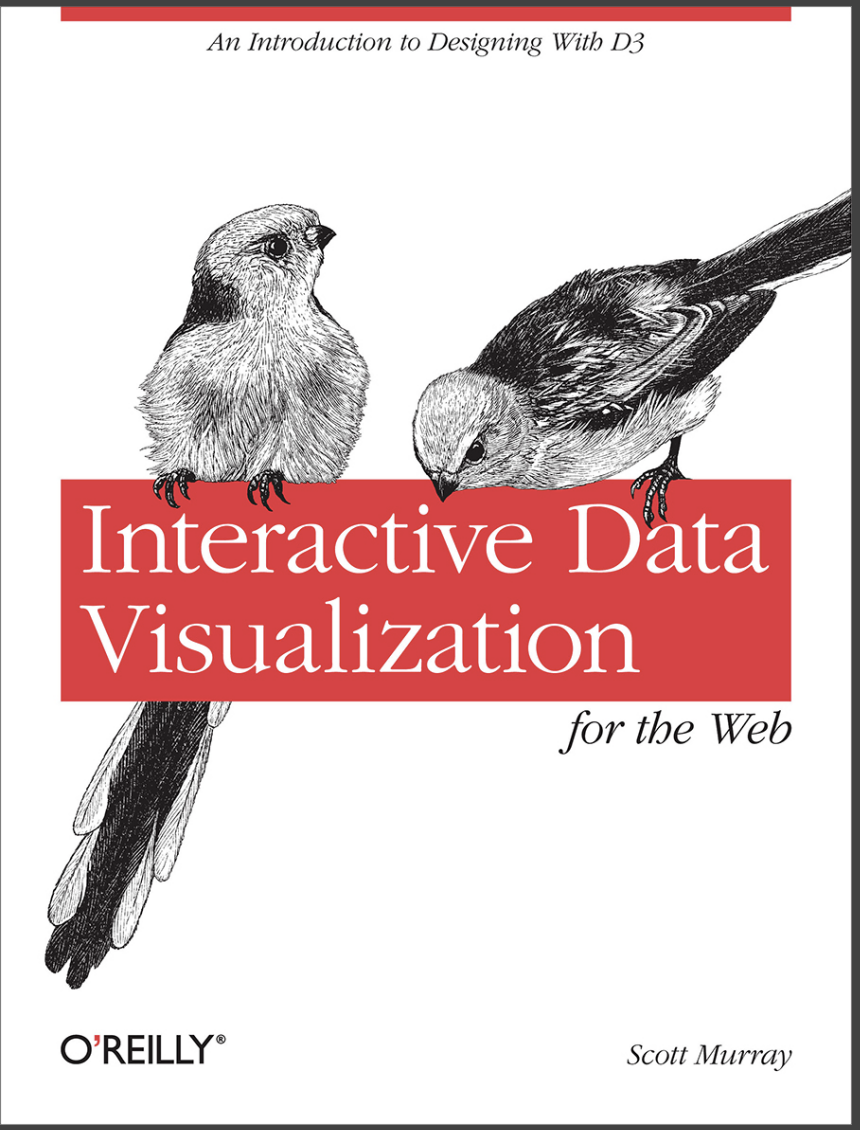
Material in class will loosely follow readings.

Readings should be read by start of class.

Post comments & quizzes on class forum.

- One comment per week (up through week 8).
- Post comments by Friday 11:59pm.
- You have 1 “pass” for the quarter.

Textbook

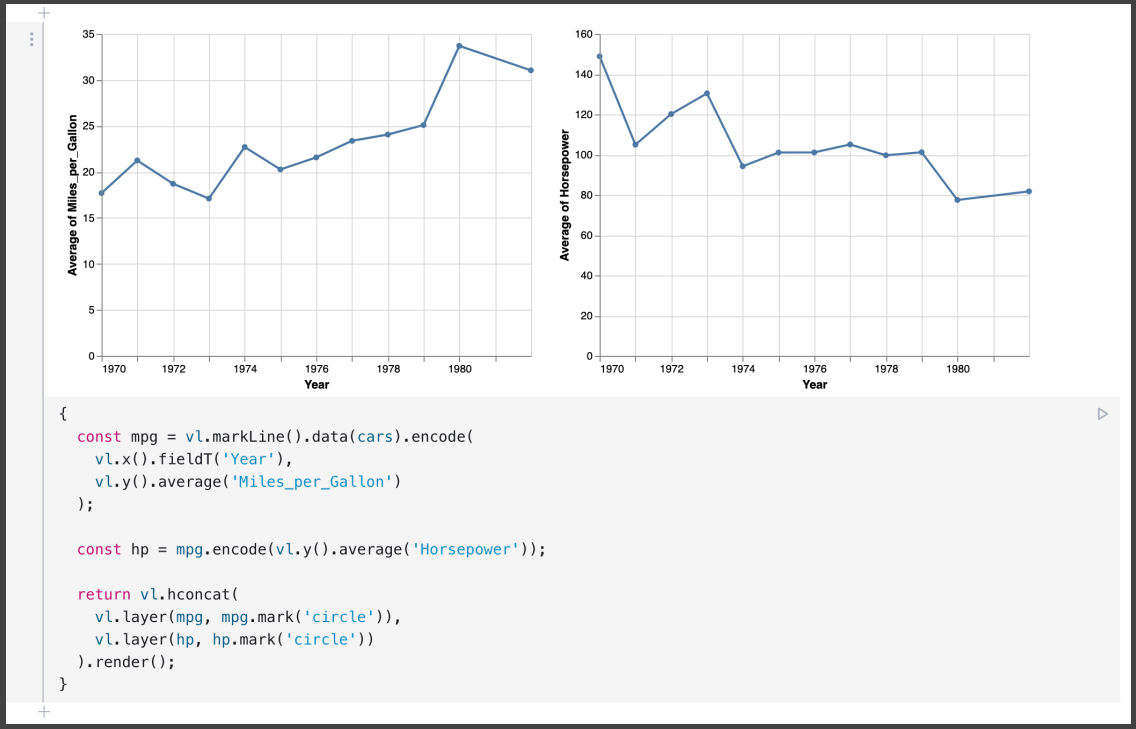


Interactive Data Visualization for the Web, 2nd Edition

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

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

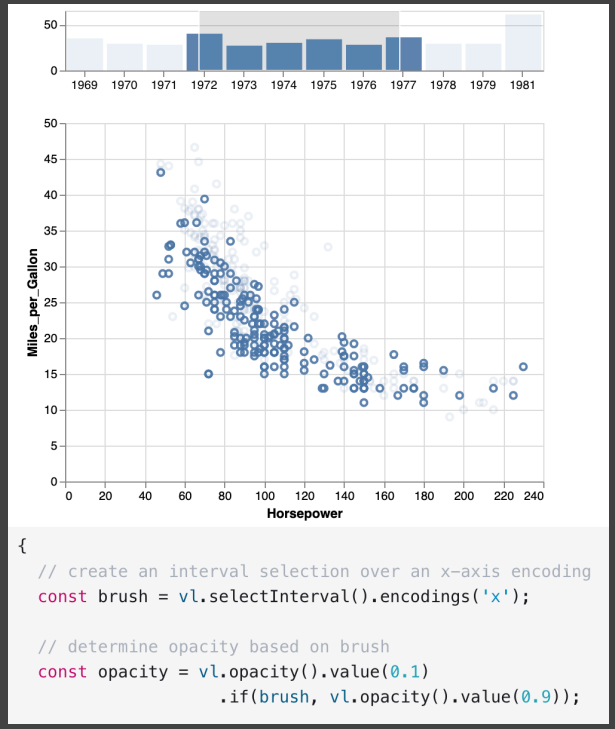
Interactive Vega-Lite Notebooks




```
{
  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

CP Class Participation (10%)

A1 Expository Visualization (10%) - *Due 01/11*

A2 Deceptive Visualization (15%) - *Due 01/25*

Peer Review (5%) - *Due 02/01*

A3 Interactive Prototype (20%) - *Due 02/13*

Peer Review (5%) - *Due 02/21*

FP Final Project (35%)

Proposal - *Due 02/14*

Prototype - *Due 02/27*

Demonstration Video - *Due 03/7*

Final Prototype - *Due 03/13*

Grading Philosophy

A *good* submission generally gets a *good grade* (A- to A, 3.6 – 3.8), but a *great grade* (A+, 3.9 – 4.0) requires *more effort*.

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 going 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**

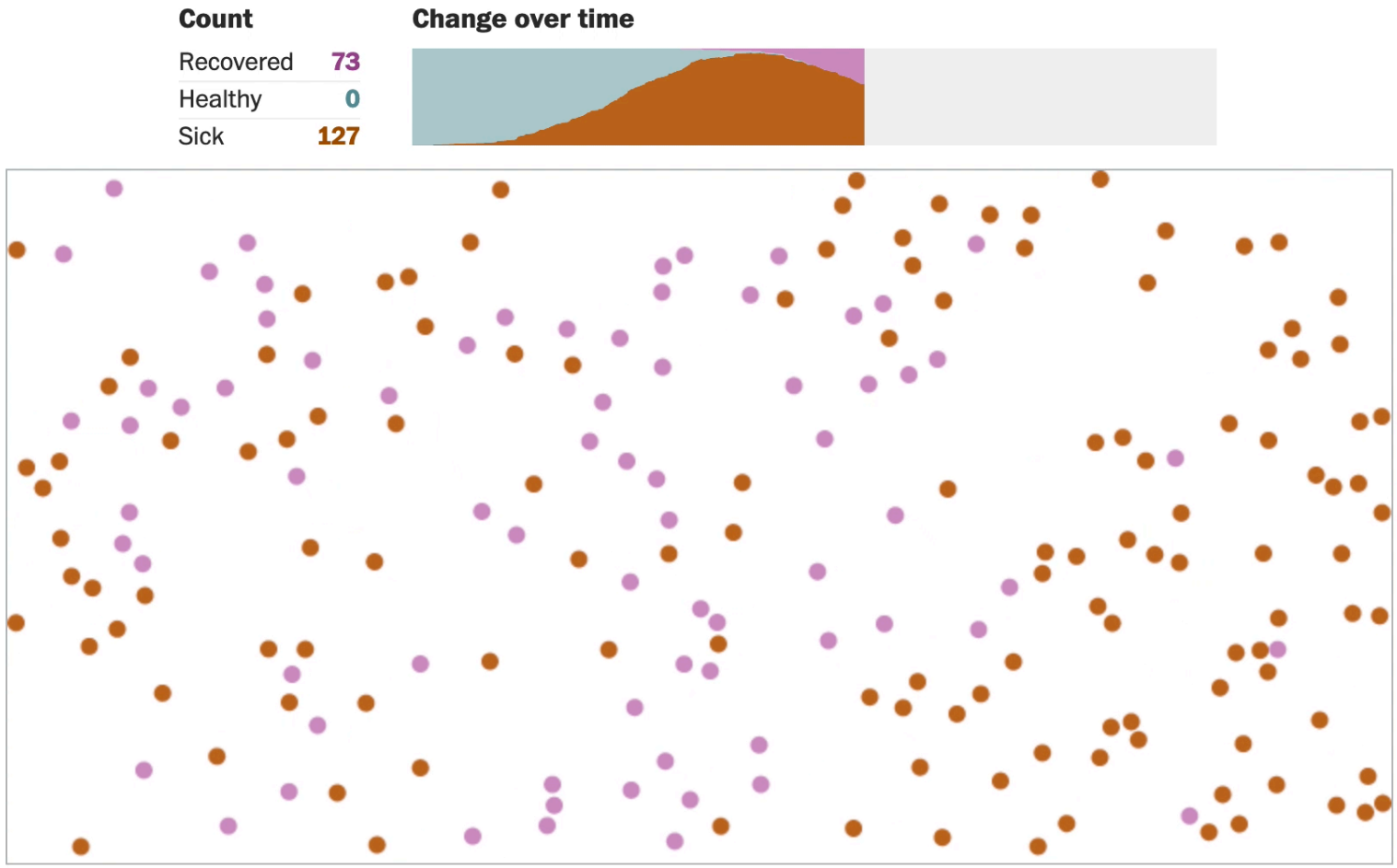
Submit and **publish online** (GitLab)

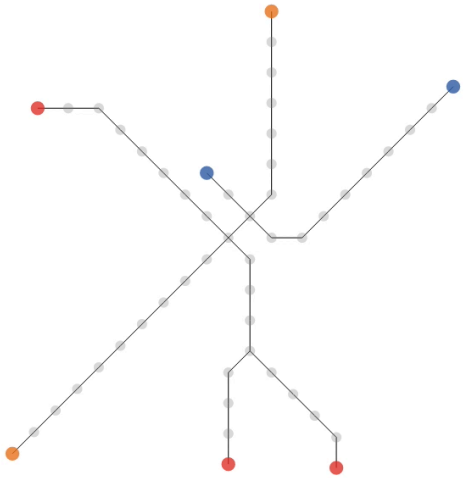
Projects from **previous classes** have been:

- 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





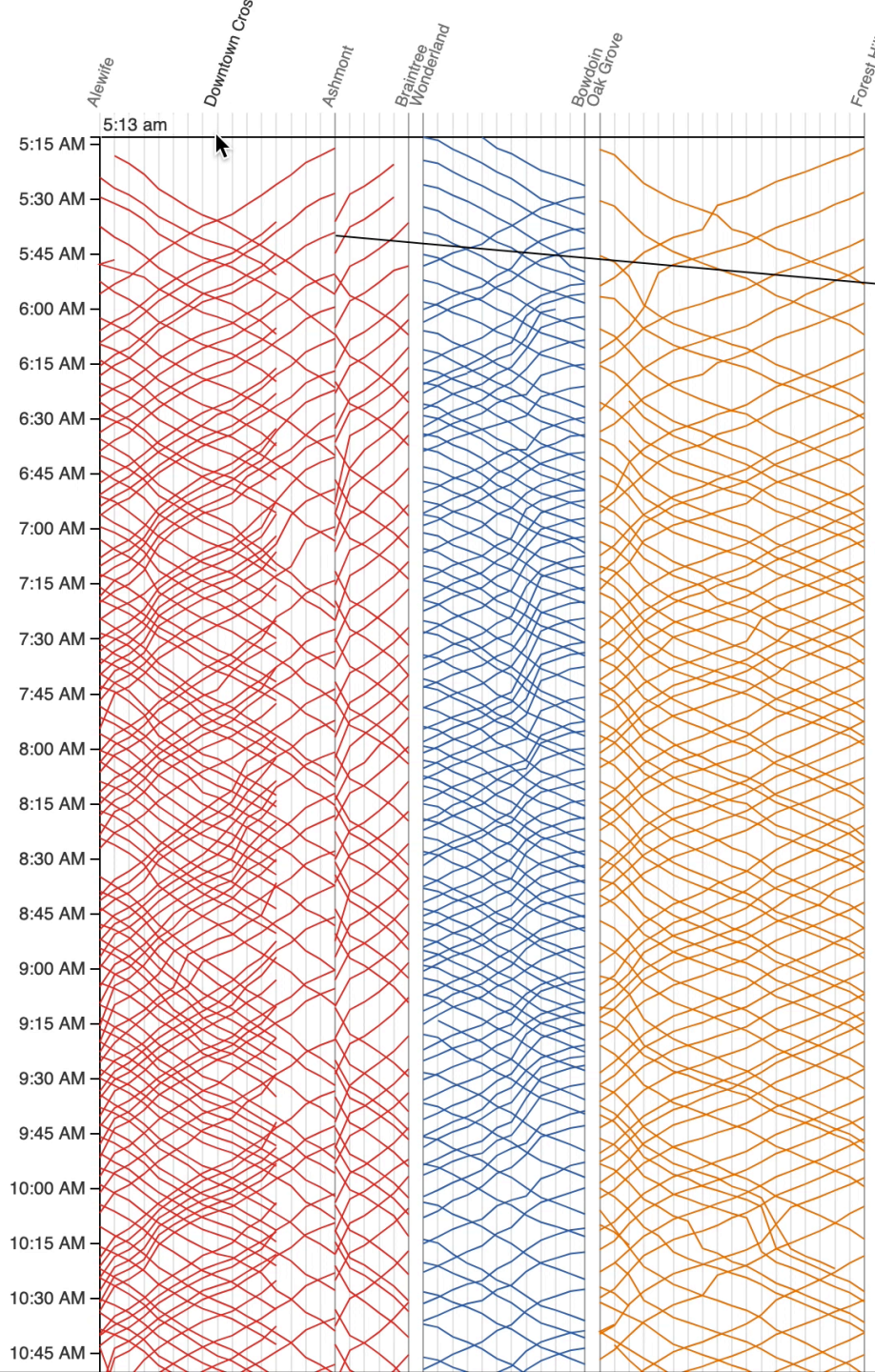
Locations of each train on the red, blue, and orange lines at 5:13 am. Hover over the diagram to the right to display trains at a different time.

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.

MBTA Viz

Barry & Card



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.

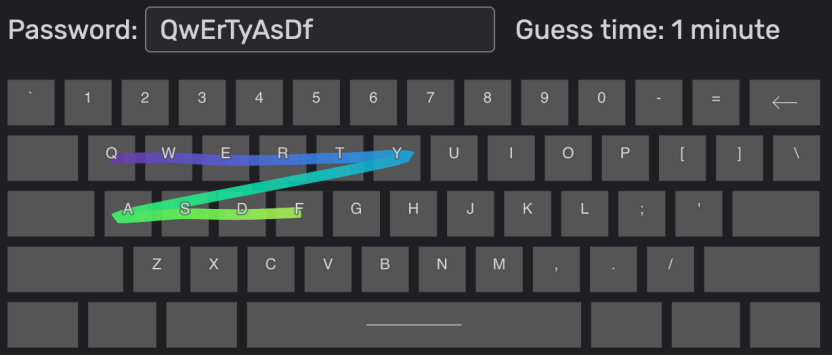
Train frequency increases around 6:30AM as morning rush hour begins.

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)

Coming Up Soon!

Thursday Jan 05: In-Class Activity

We will try designing our own data visualizations in class!

You need to watch the pre-recorded Data and Image Models lecture videos beforehand.

We will have a quick poll on PollEV before diving into the activity.

Observable + Data Tutorial

Friday Jan. 6, 4:30-6pm

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

Zoom link is 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, Wed Jan 11**.

Online Course Participation

Quiz & discussion comments on class forum (Ed).
Both are due by Friday, 11:59pm.

Quizzes start this week.
One comment per week.
You have 1 “pass” (quiz + comment) for the quarter.

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

Instructors

cse442@cs

Instructor

Leilani Battle OH: *Wed 2-3pm (virtual)*
Assistant Professor, CSE

Teaching Assistants

Katherine Juarez OH: *TBD*

Sonia Saitawdekar OH: *TBD*

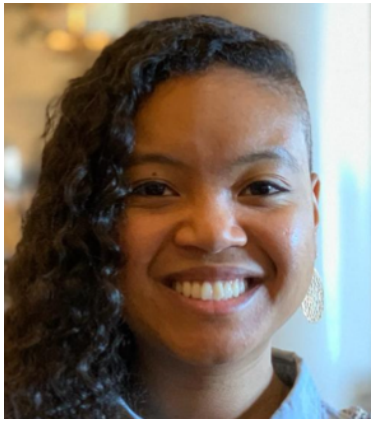
Aakash Srazali OH: *TBD*

WeiJun Tan OH: *TBD*

Yuanjie 'Tukey' Tu OH: *TBD*

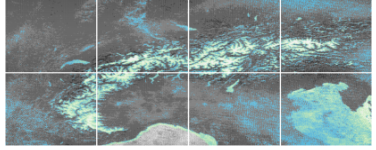
Erin Wilson OH: *TBD*

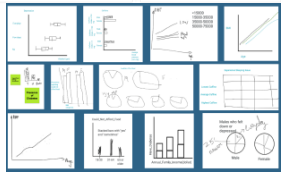
Yu Xin OH: *TBD*

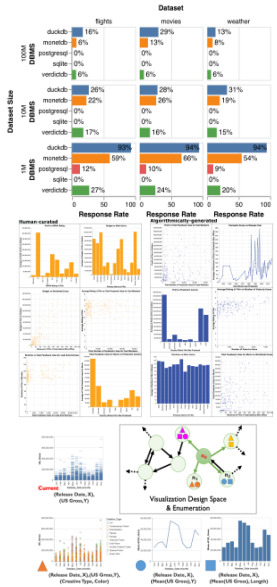


Leilani Battle

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





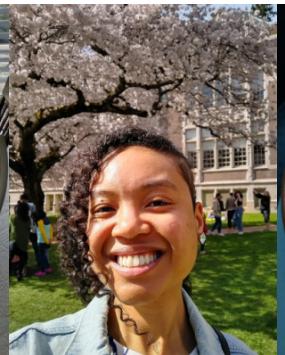


Visualization / HCI / Data management / Data Science

I model relationships between analysts' ***intents***, i.e., analysis goals, and ***behaviors***, i.e., patterns of interaction 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, travel, board games, etc.



Katherine Juarez

kajuarez@cs.washington.edu

- Second Year PhD Student
- Research Area: Human-Computer Interaction, Cross-Culture Studies
- Hobbies:
  - Kayaking
  - Reality TV
  - Trying out new restaurants

Sonia Saitawdekar

sonias02@cs.washington.edu

- Senior
- CS major, Business minor
- Interests: front-end development, data science
- Hobbies: dancing, listening to audiobooks, trying new foods

Aakash Shameer Bin **Srazali**

- Kuala Lumpur, Malaysia
- Senior in Computer Science
- 5th time being a TA - **CSE 442, CSE 333, CSE 351**
- Hobbies: Playing football(soccer) / Collecting shoes
- Contact: aaksra@cs.washington.edu
- Website: aakashshameer.com

WeiJun Tan
wj428@cs.washington.edu

- From Selangor, Malaysia
- Junior - CS / Stat
- 2nd time TA CSE 442
- Interest - systems / data science
- Hobbies: chess / table tennis / badminton / competitive programming

Yuanjie (Tukey) Tu

yuanjt2@cs.washington.edu

- From Jiangxi, China
- PhD student – Civil Engineering
- Research: Self-driving vehicles
- Hobbies: Hiking, swimming, traveling

Erin Wilson
ewilson6@cs.washington.edu

- PhD student in Computer Science
- From Northern California
- Research interests: computational biology, data science, sustainability
- Goofiest data visualization: analyzed social dynamics in the fantasy novel *Mistborn* by Brandon Sanderson based on how often each character “raised an eyebrow” at other characters.

Yu Xin

- 5th year master (BS/MS) student
- Area of Interest
  - Systems (cloud computing, distributed systems)
  - Machine Learning (CV, NLP)
- Hobby
  - Swimming / badminton
  - Solve puzzles
  - Video games

Questions?