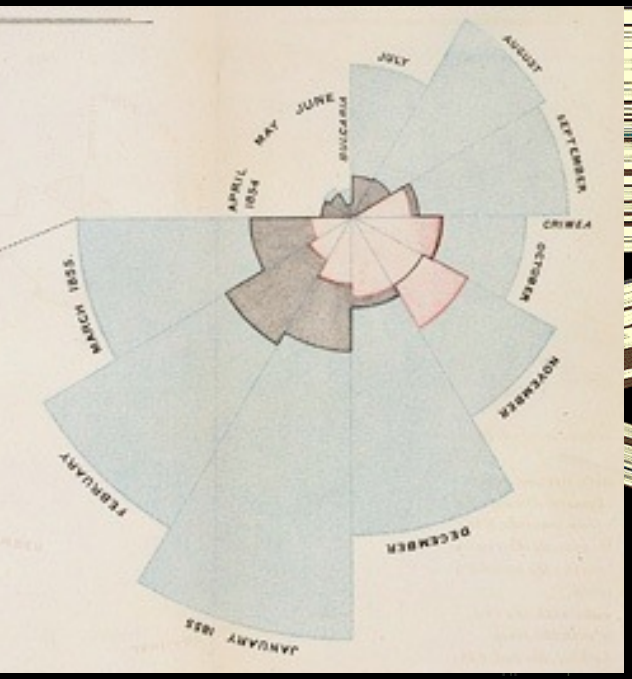
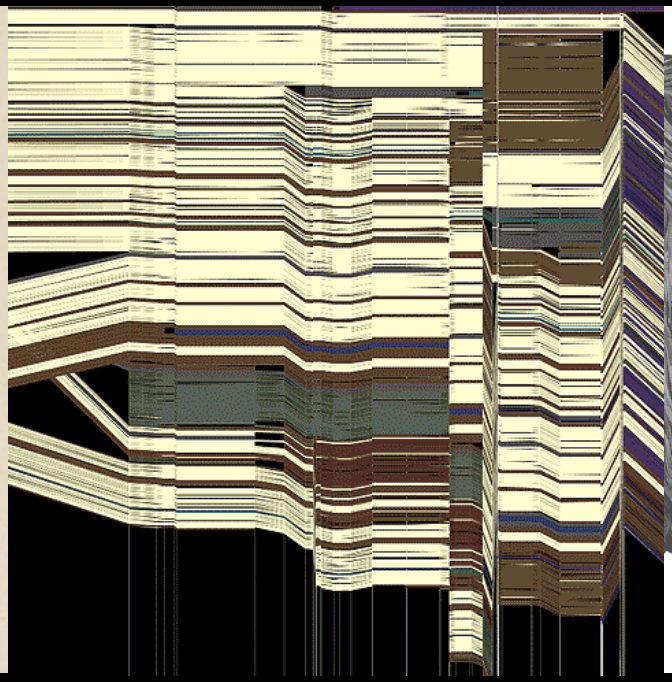
CSE 412 - Intro to Data Visualization

The Value of Visualization





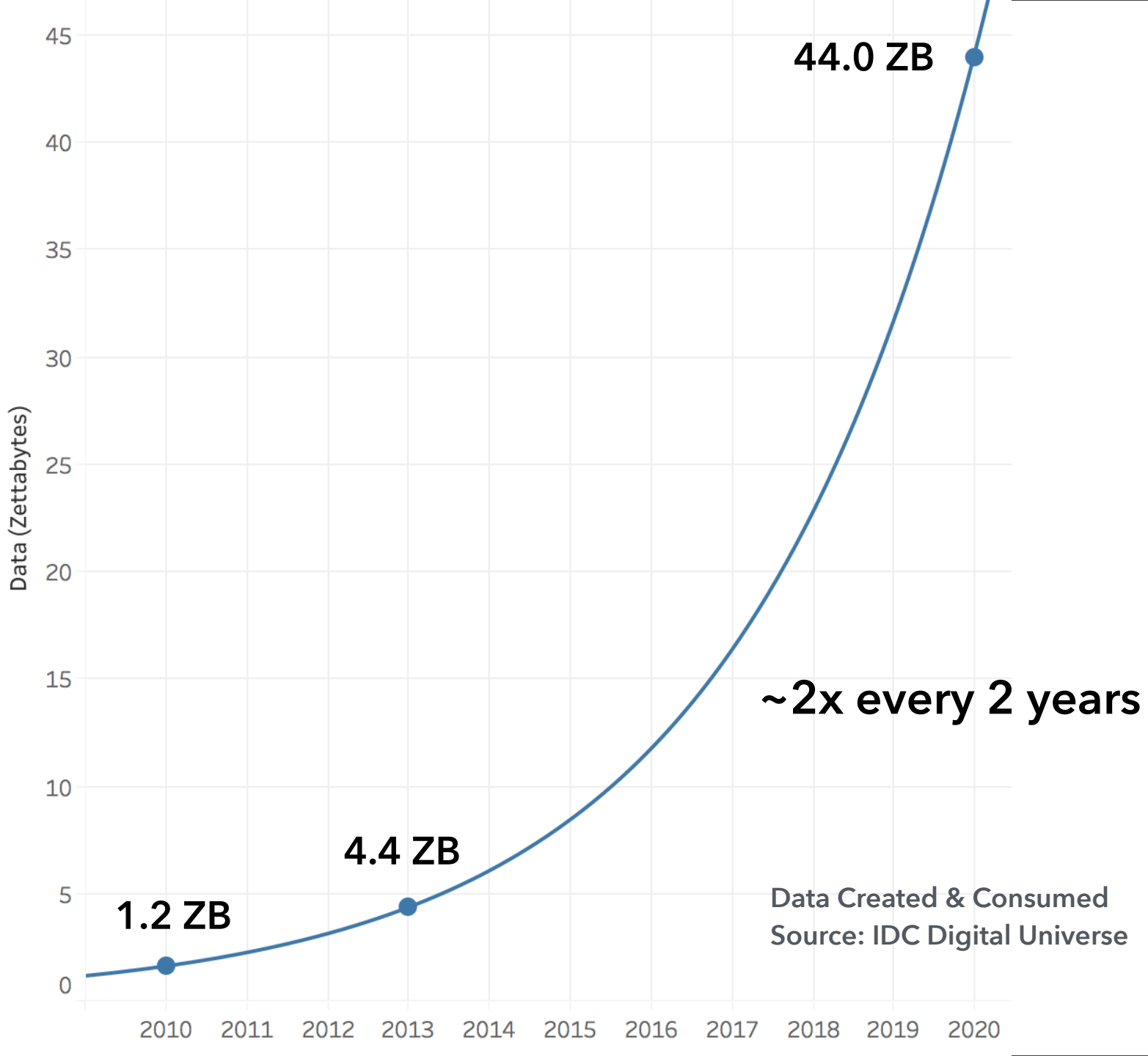
Jane Hoffswell University of Washington

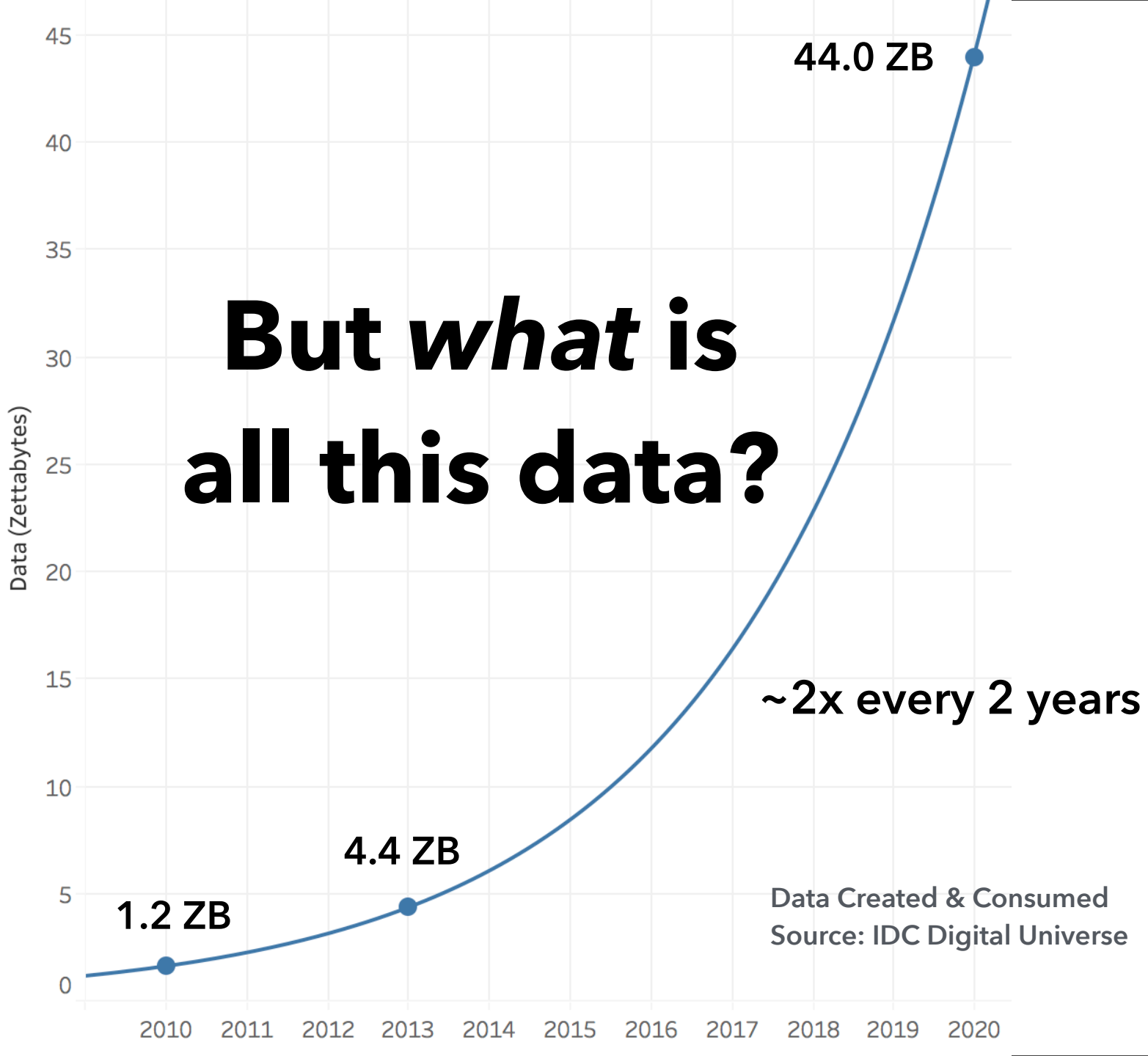
How much data (bytes) did we produce in 2010?

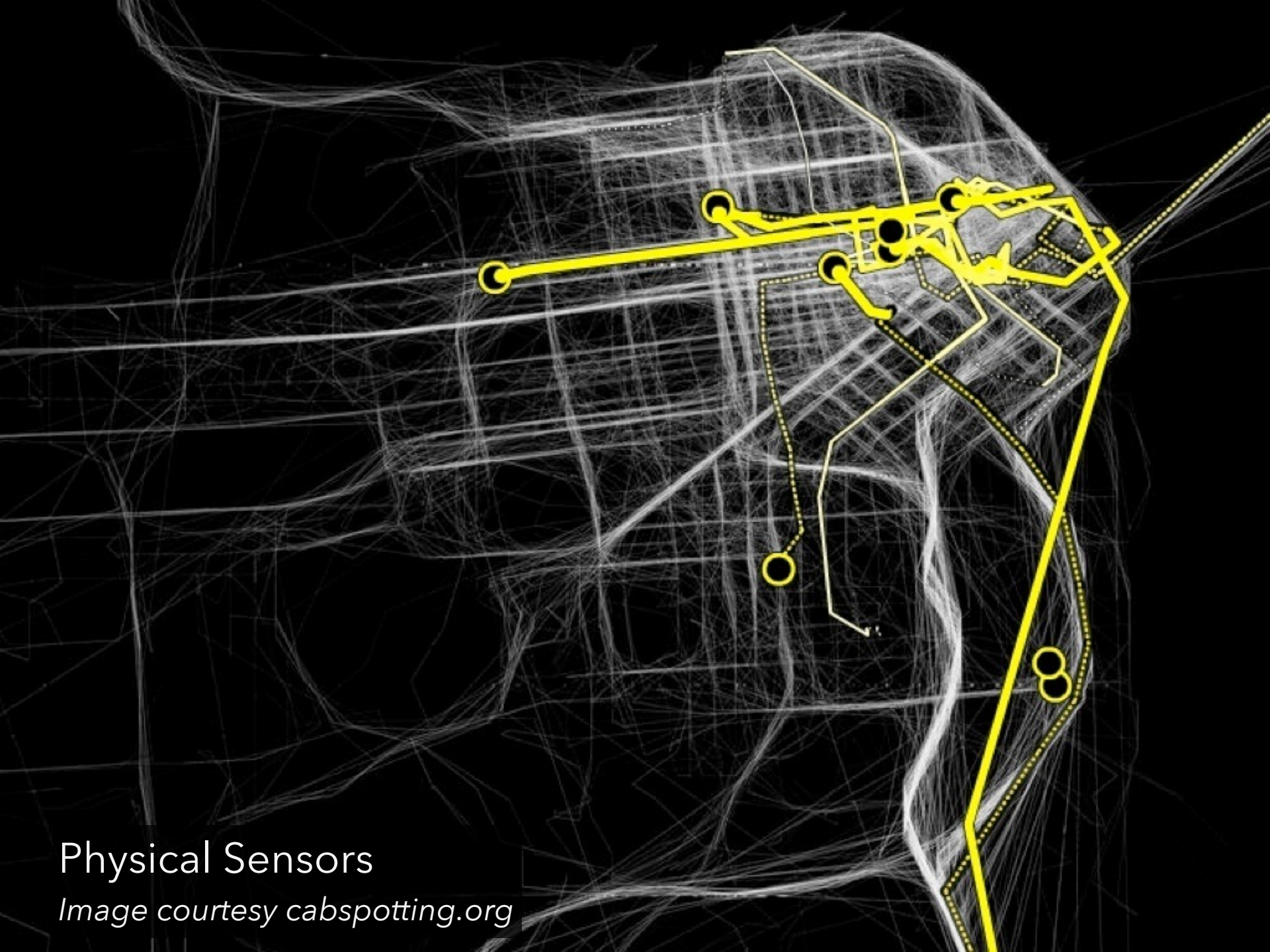
2010: 1,200 EB (1.2 ZB)

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

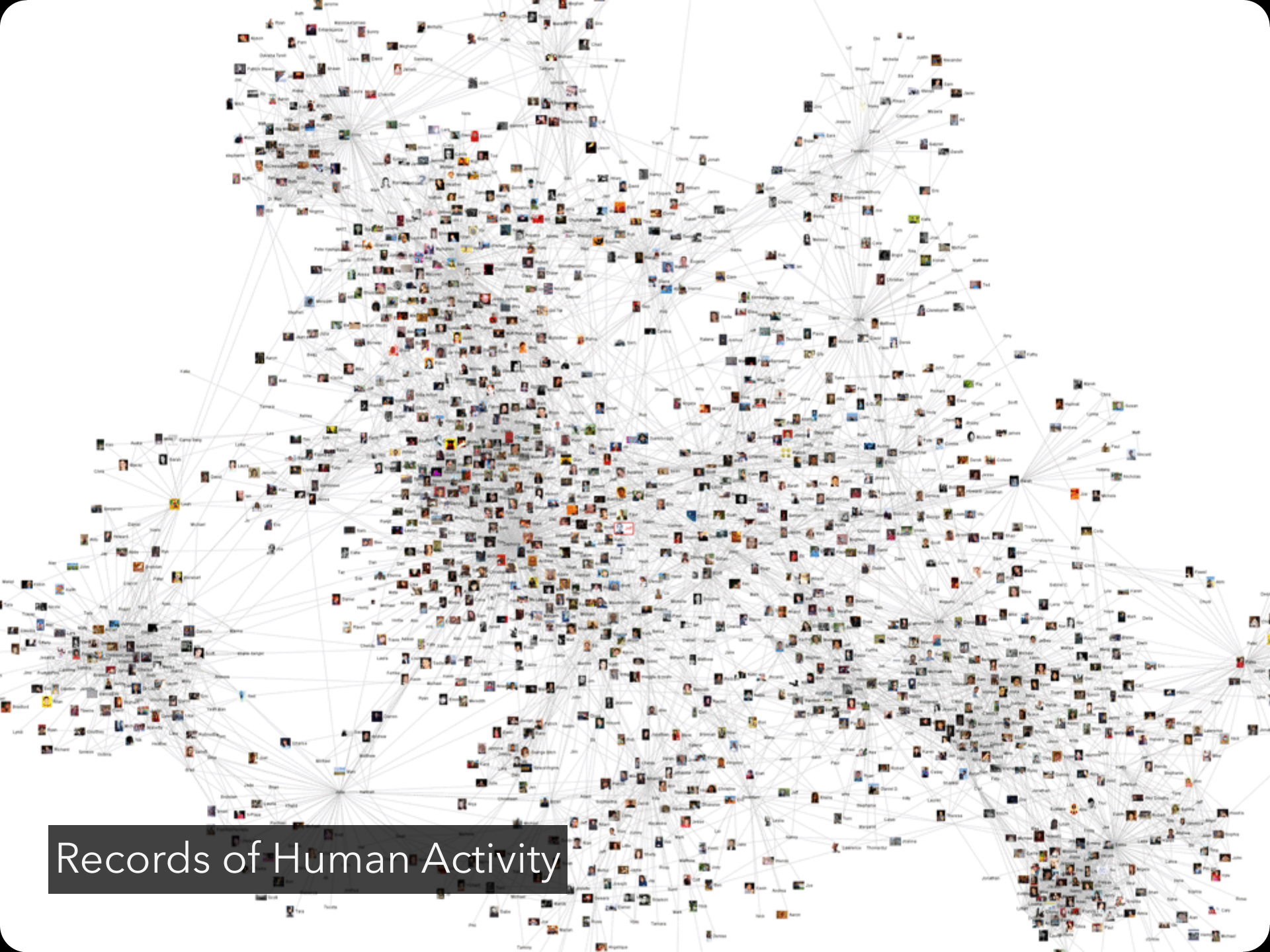

But *what* is all this data?
Data (Zettabytes)
45
40
35
30
25
20
15
10
5
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
1.2 ZB
4.4 ZB
44.0 ZB
~2x every 2 years
Data Created & Consumed
Source: IDC Digital Universe



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 − +

Psychology's Replication Crisis Can't Be Wished Away

It has a real and heartbreaking cost.

ED YONG | MAR 4, 2016 | SCIENCE

Share Tweet ...

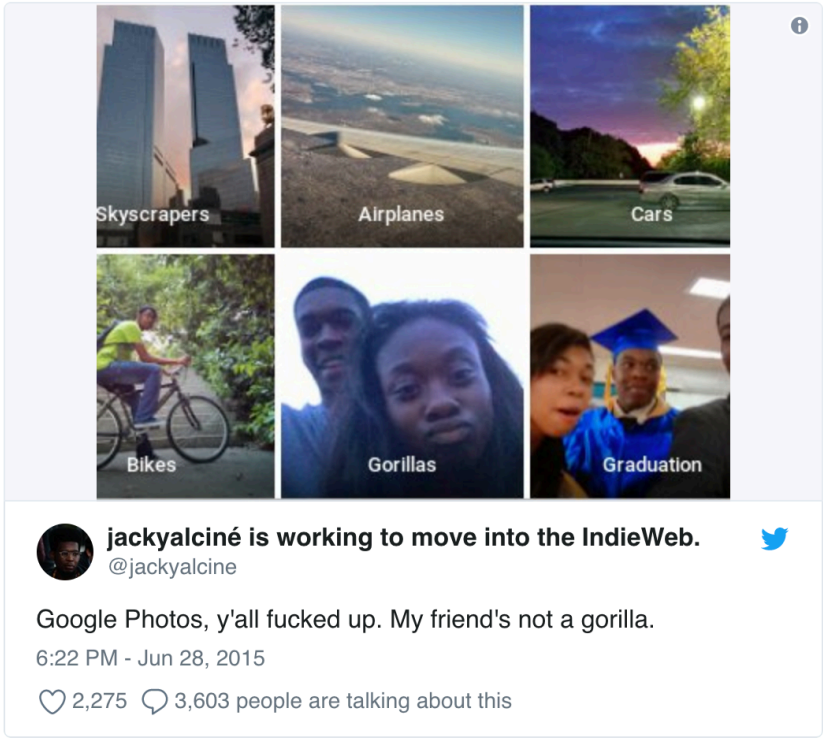
TEXT SIZE − +

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.

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

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

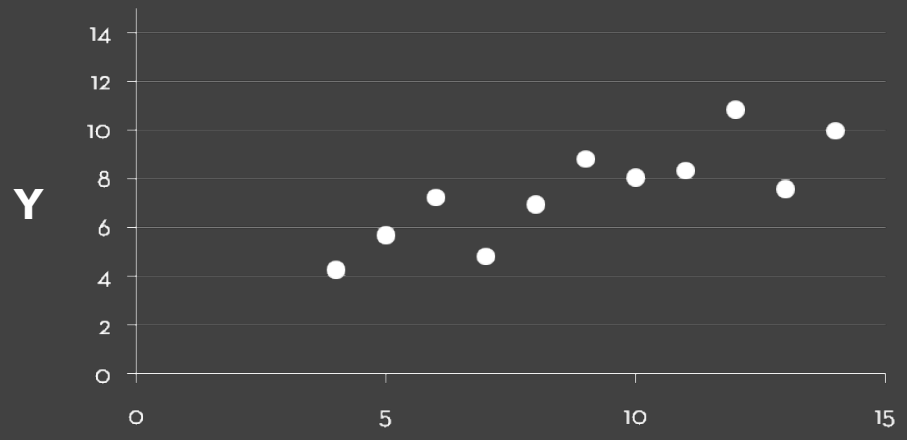
Linear Regression

$Y = 3 + 0.5\ X$

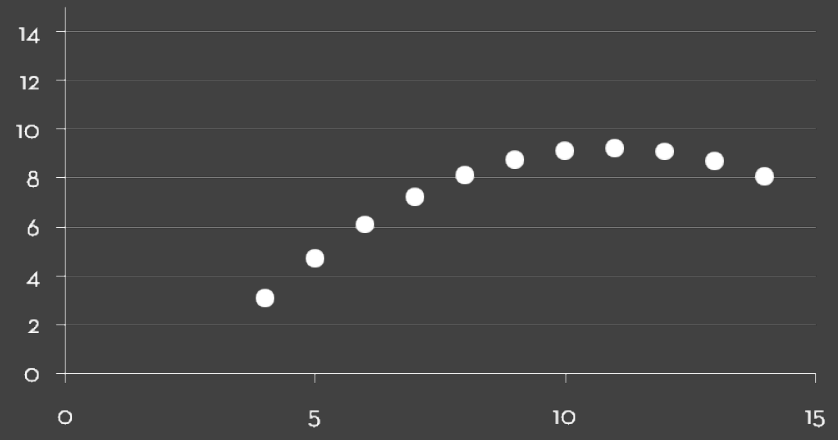
$R^2 = 0.67$

[Anscombe 1973]

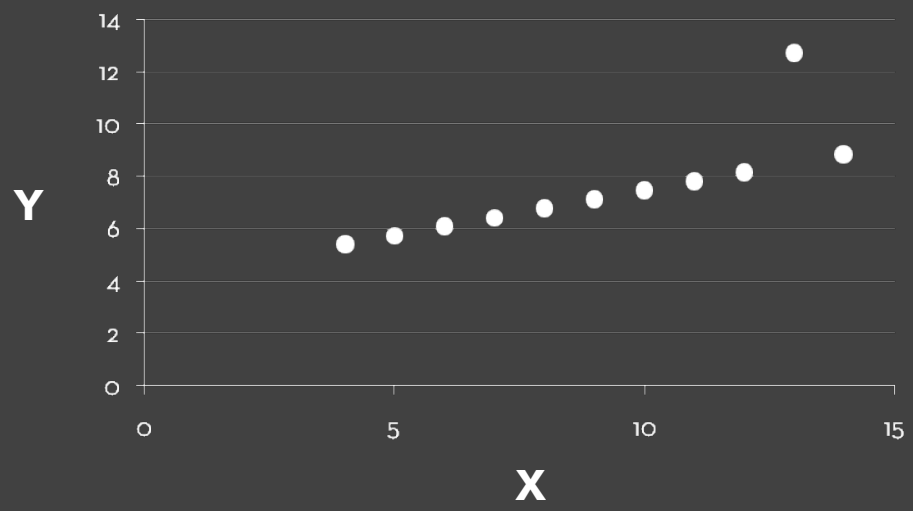
Set A



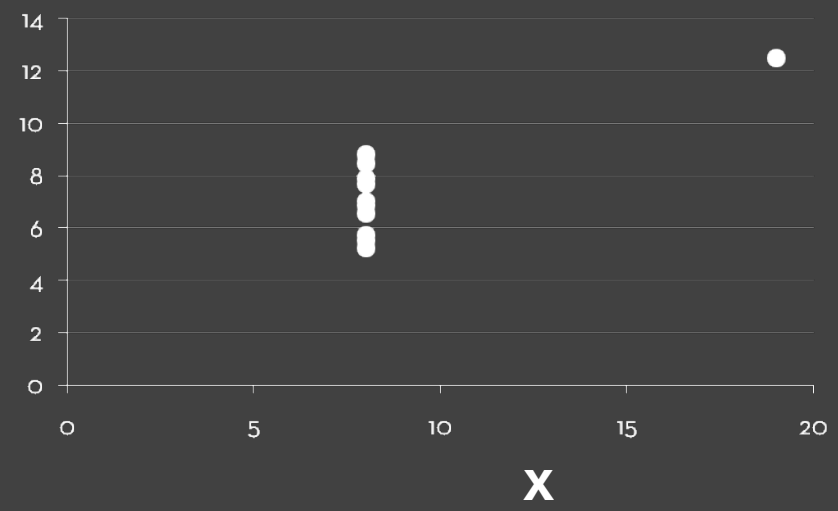
Set B

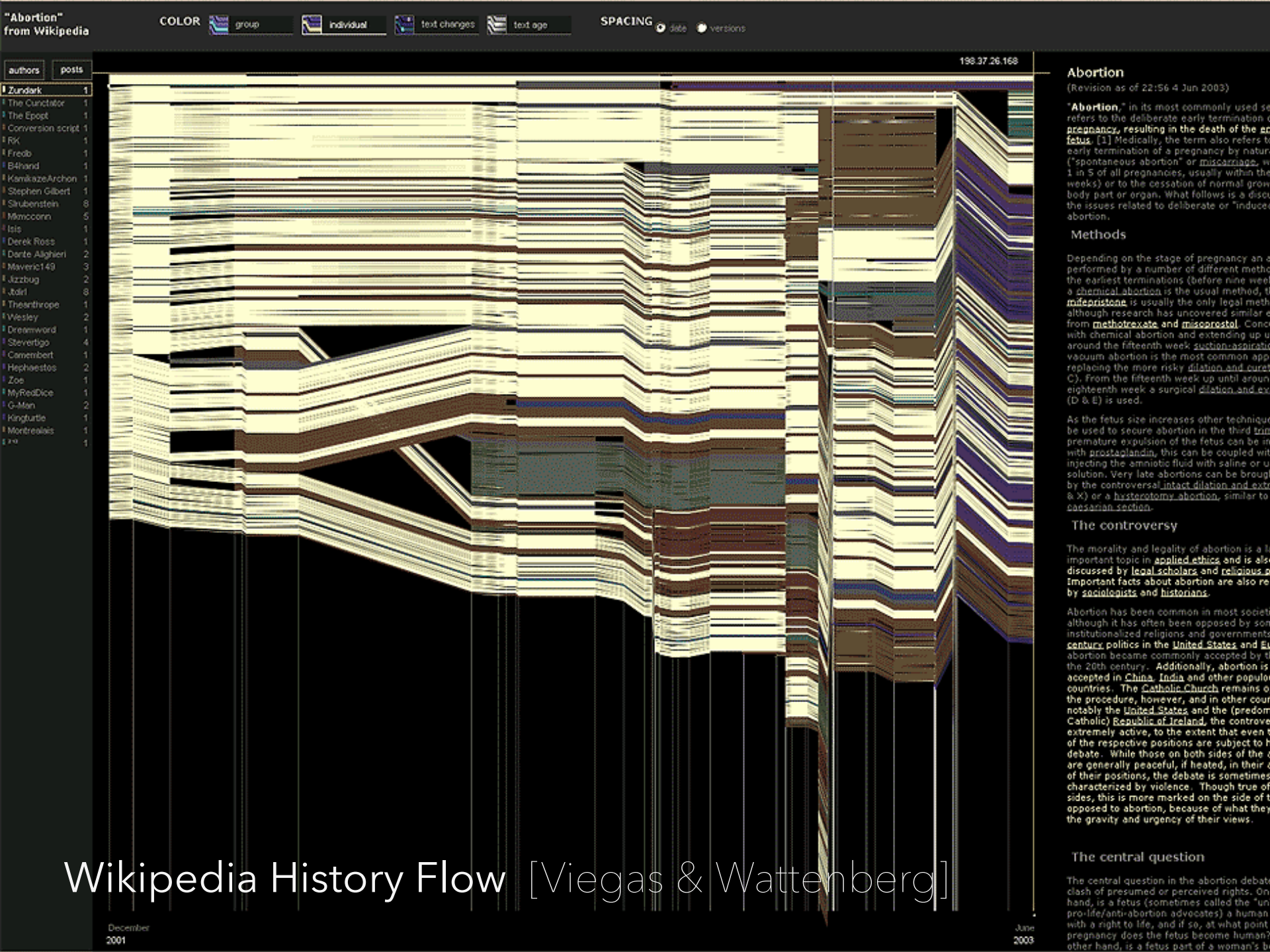


Set C



Set D





Wikipedia History Flow [Viegas & Wattenberg]

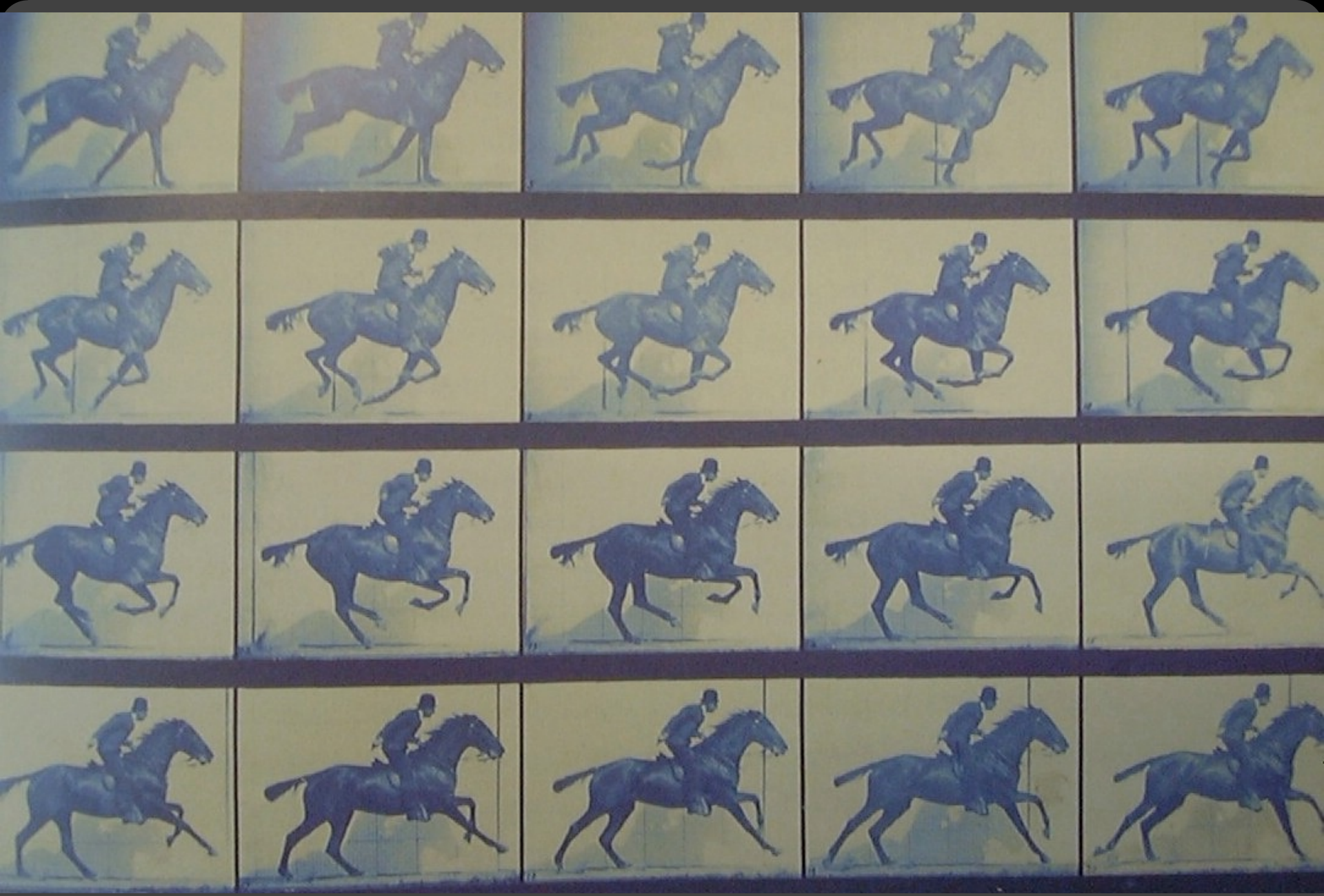
Wikipedia History Flow [Viegas & Wattenberg]

Why Create Visualizations?

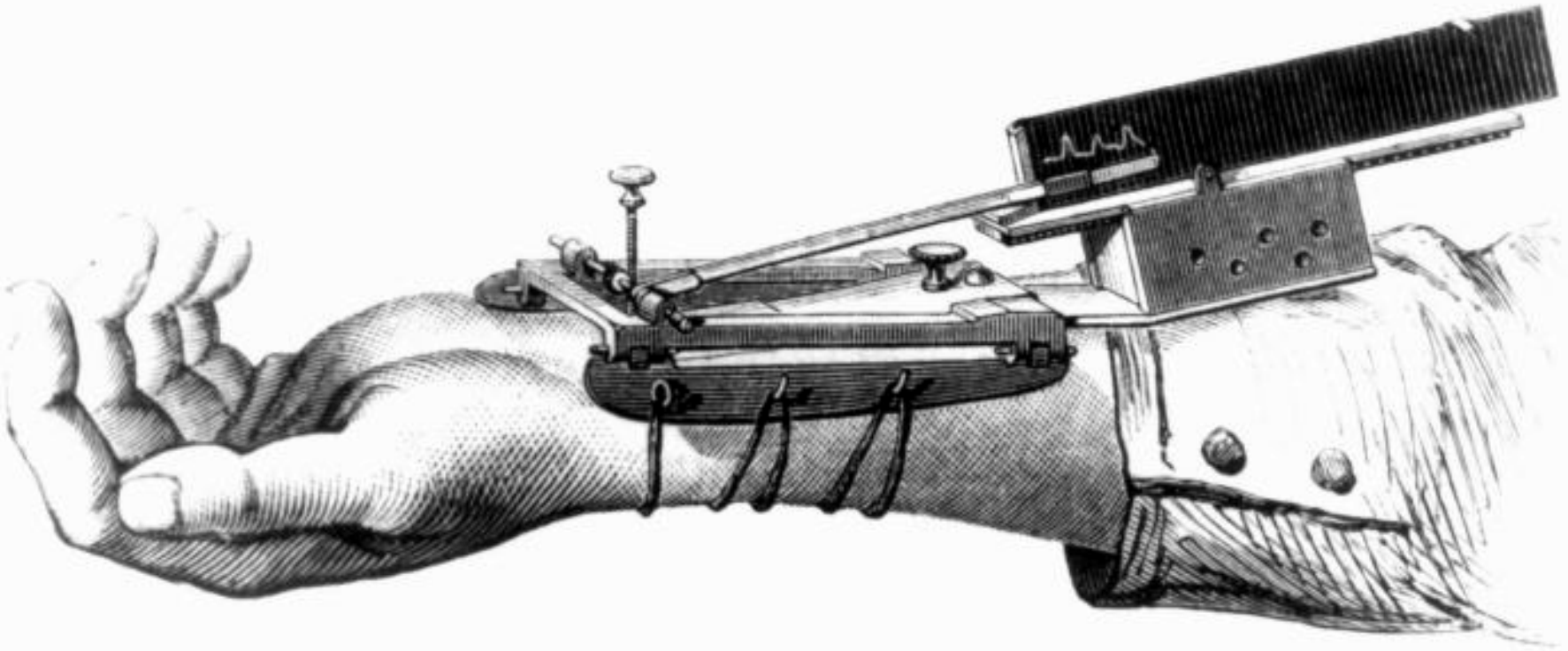
Why Create Visualizations?

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

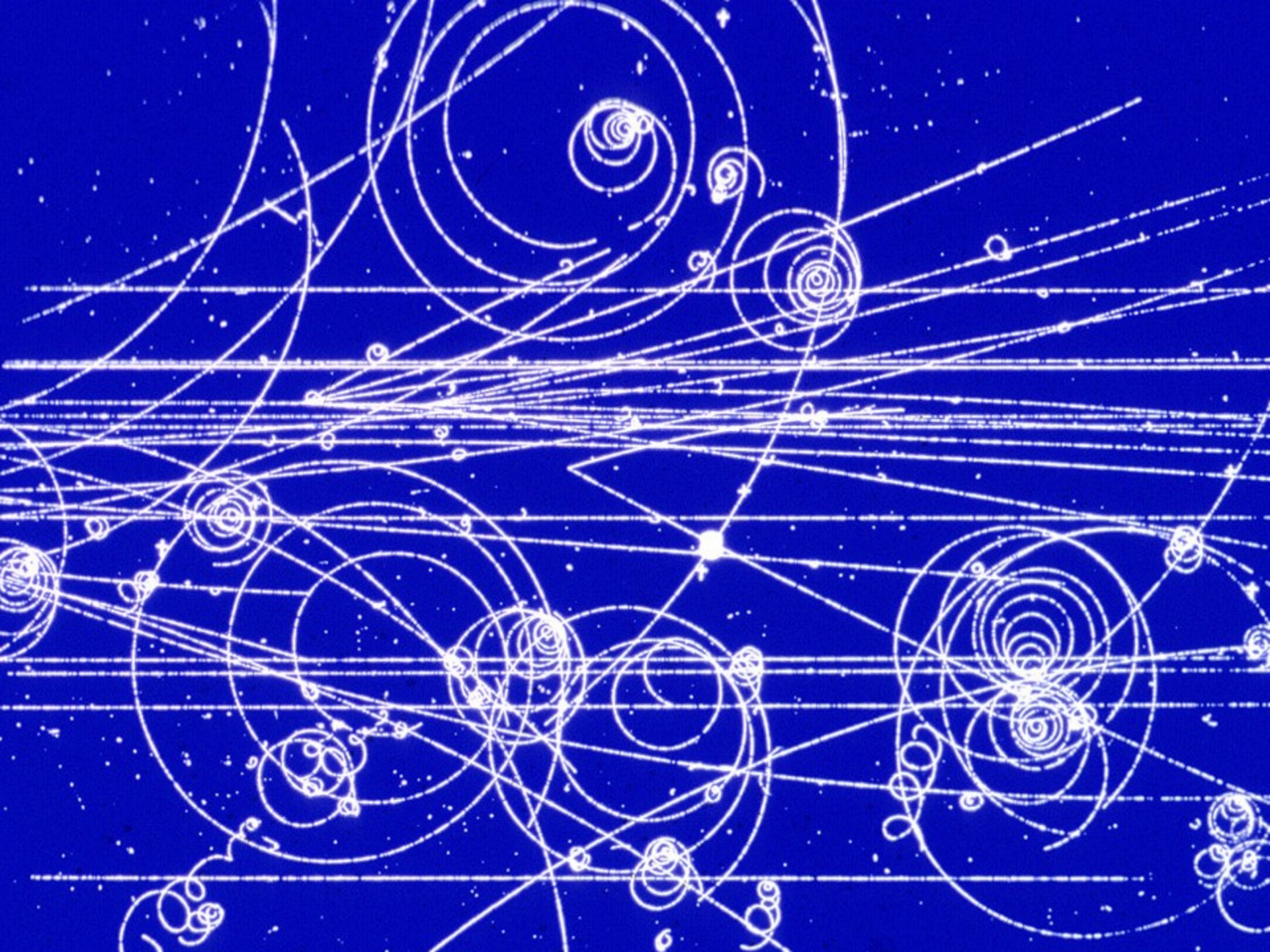


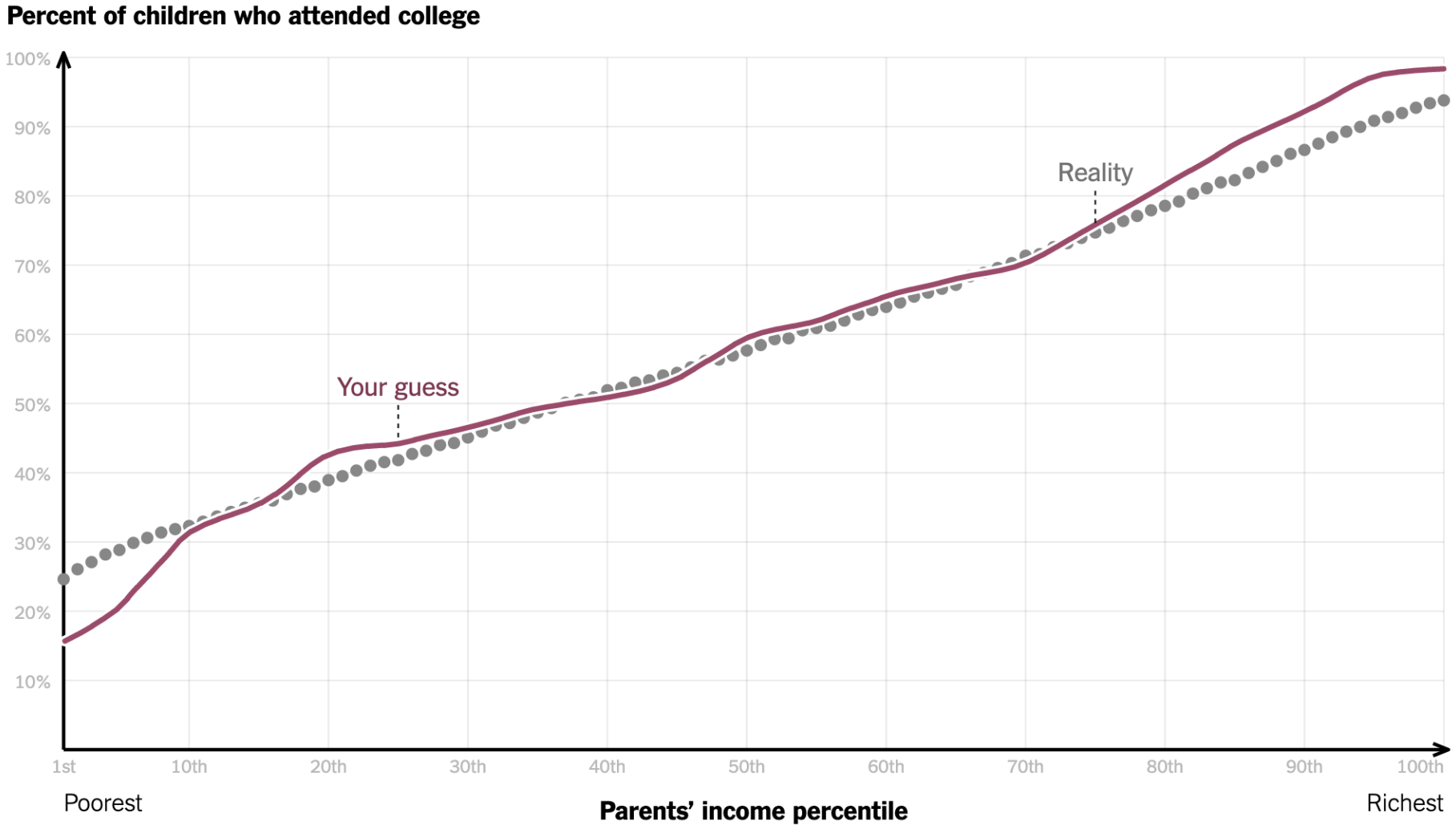
Gallop, Bay Horse “Daisy” [Muybridge 1884-86]



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

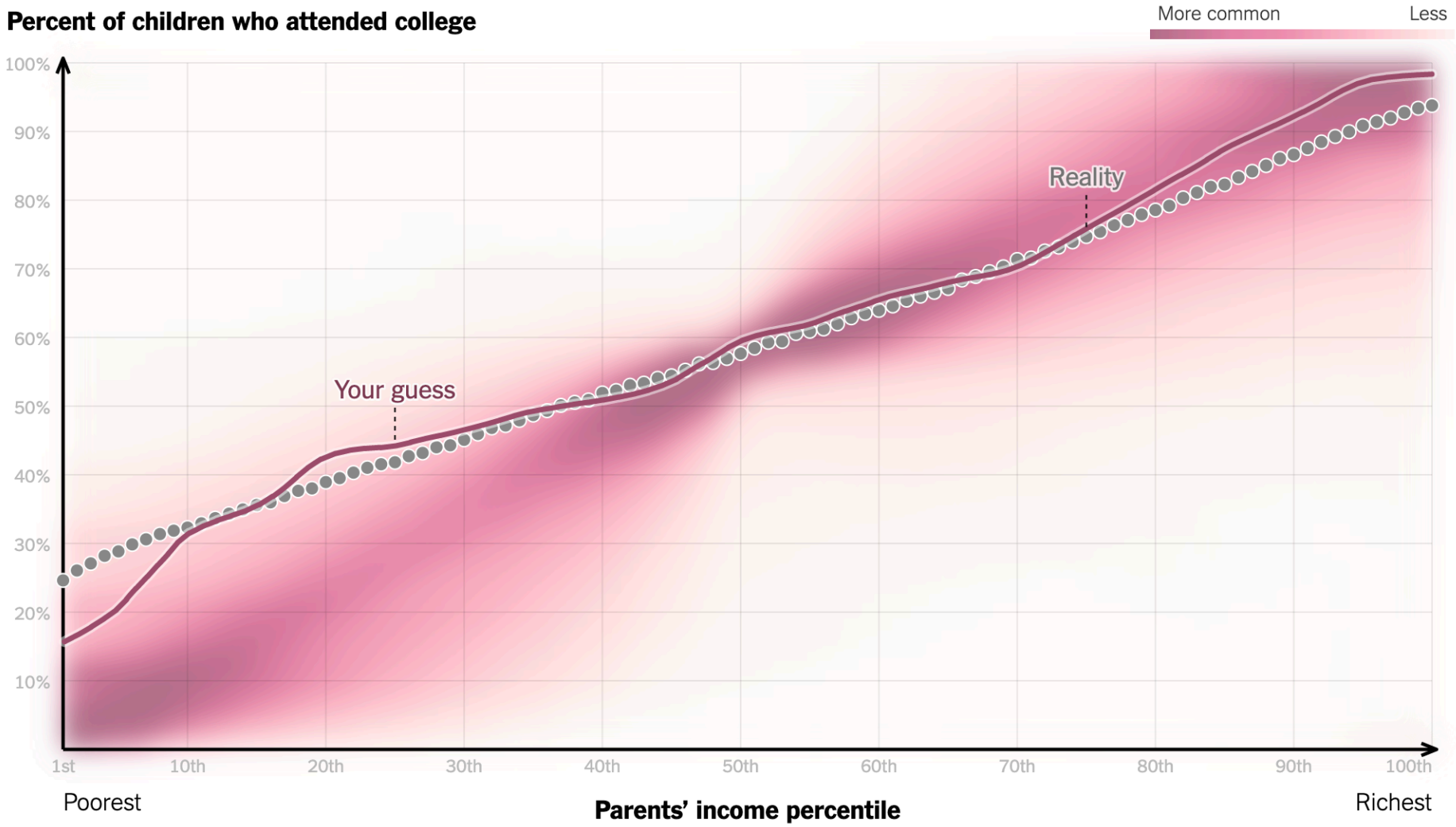
E.J. Marey's sphygmograph [from Braun 83]





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

[New York Times, May 28, 2015]

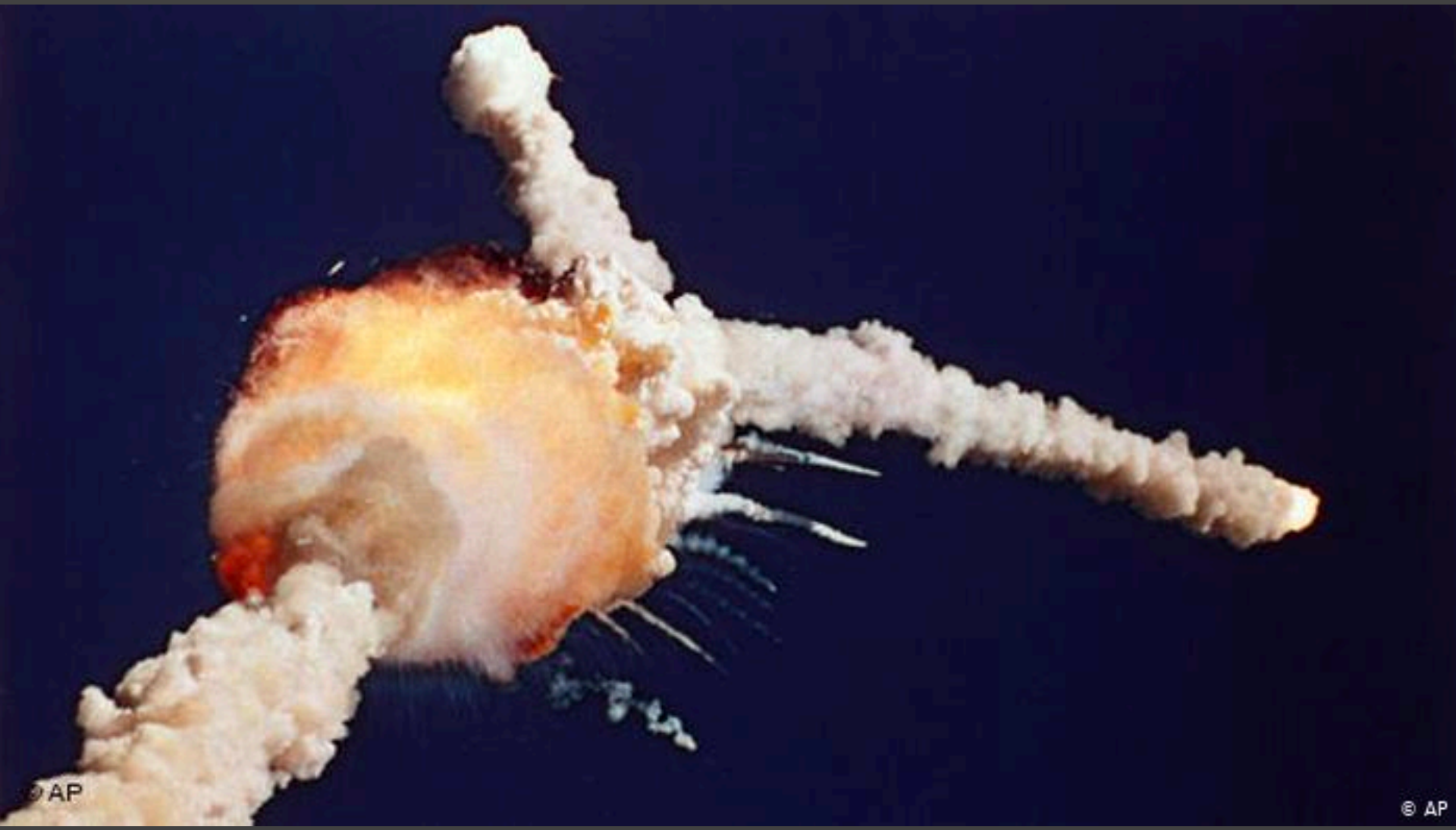


You Draw It: How Family Income Predicts Children's College Chances

[New York Times, May 28, 2015]

Support Reasoning


© 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°)

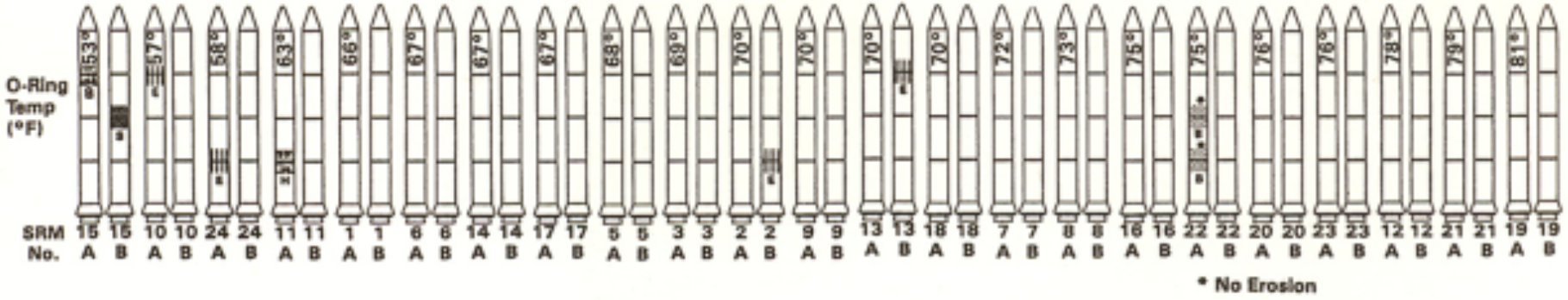
SRM-13A, 15, 16A, 18, 23A 24A
- nozzle Blow-By

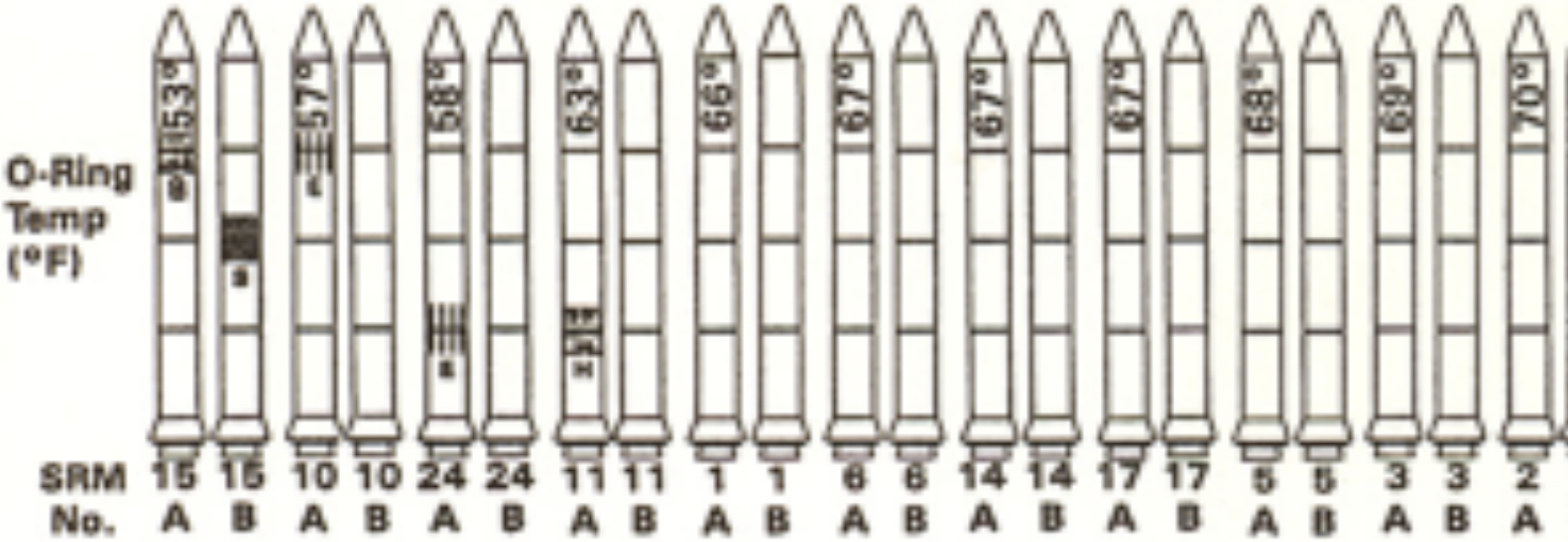
History of O-ring Temperatures (Degrees - F)

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

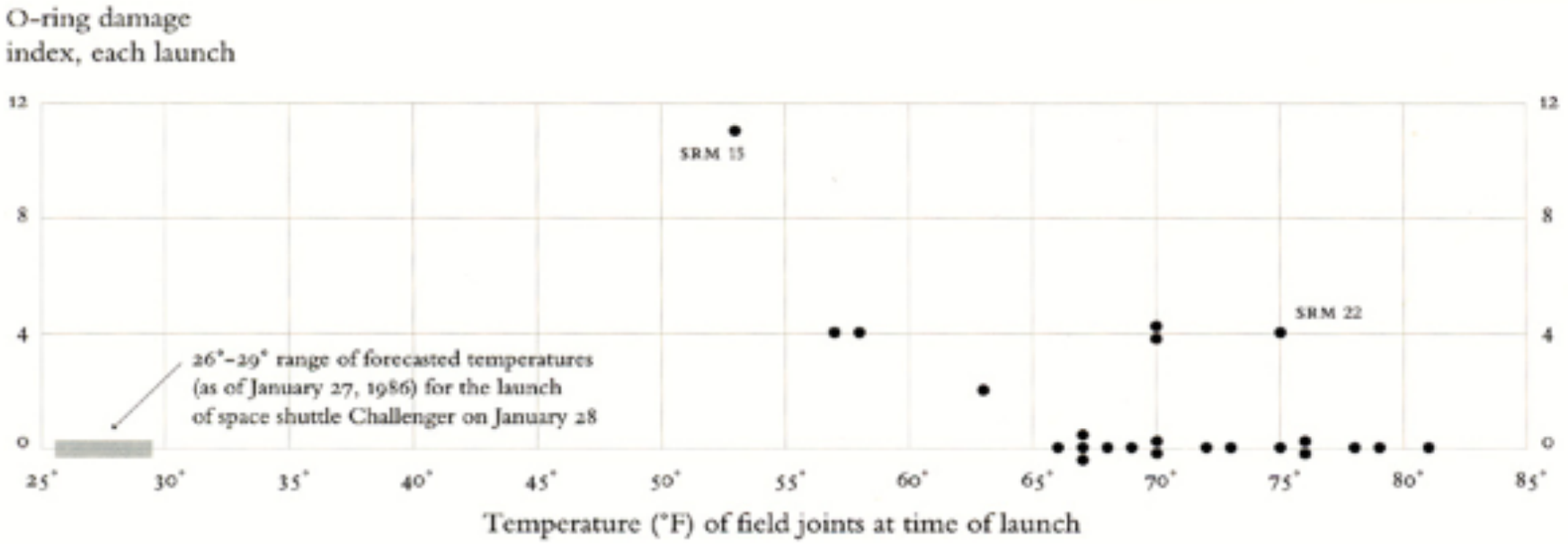
2 of 13 pages of material faxed to NASA by Morton Thiokol [from Tufte 1997]

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]

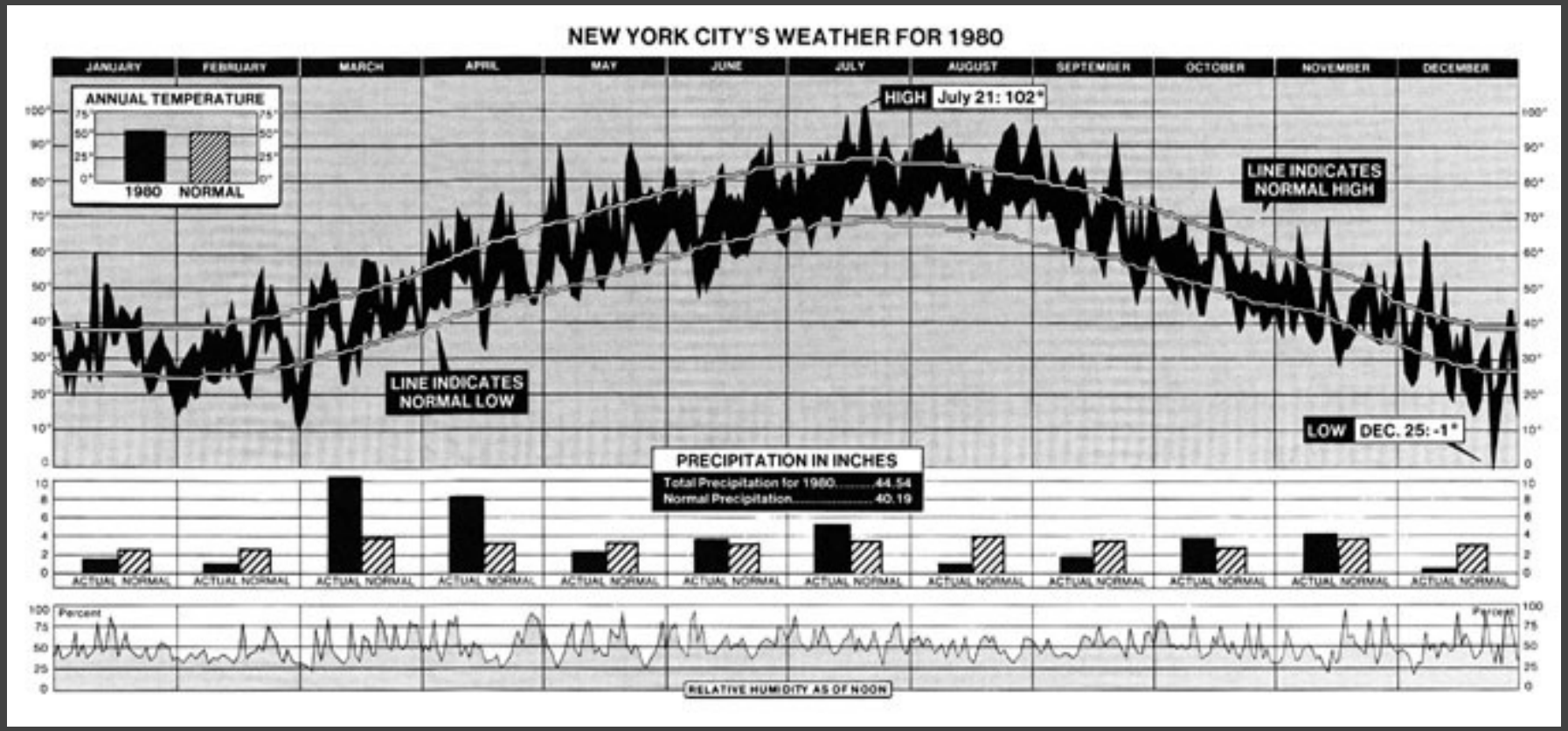
Data in Context: Cholera Outbreak

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

Data in Context: Cholera Outbreak

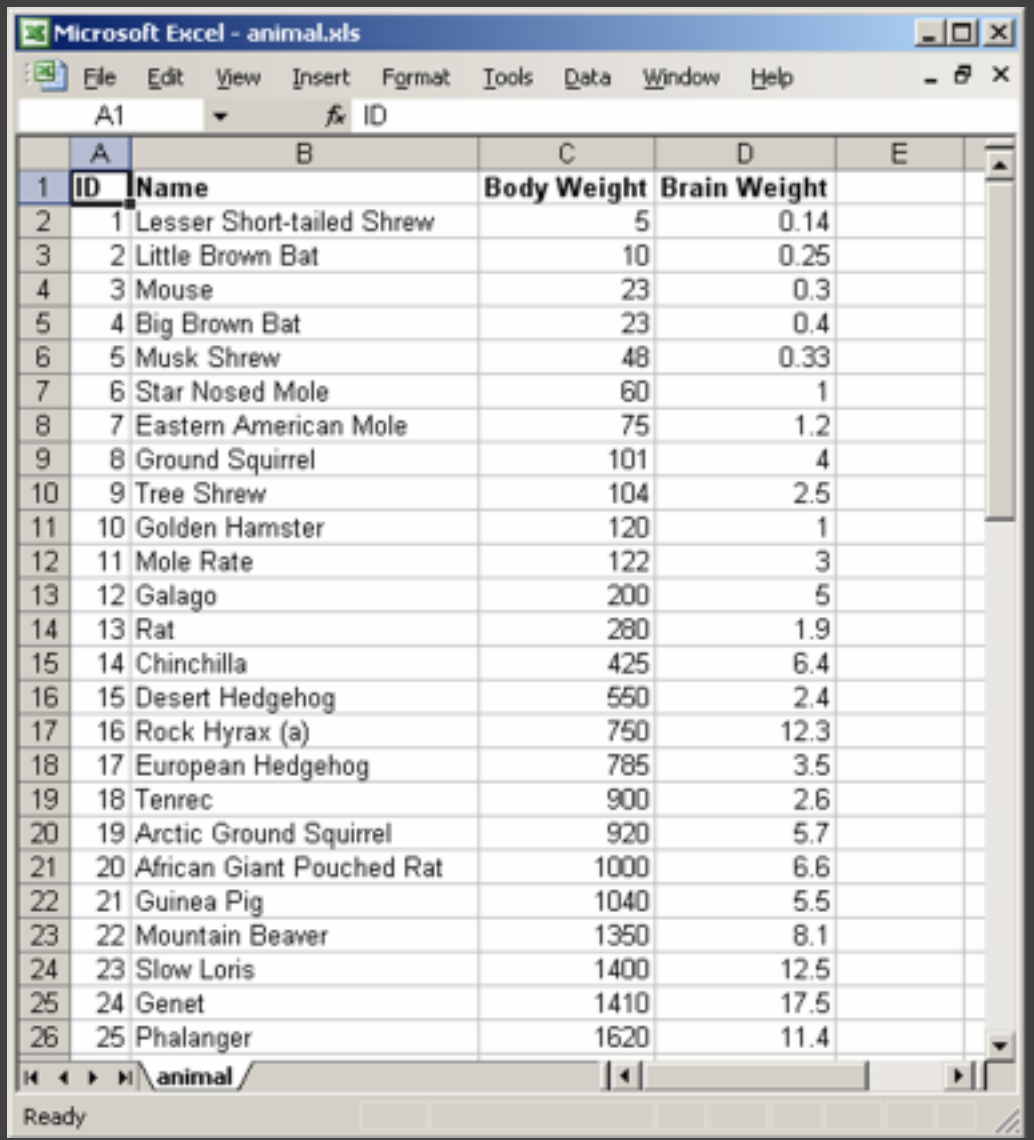
Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

Find Patterns: NYC Weather

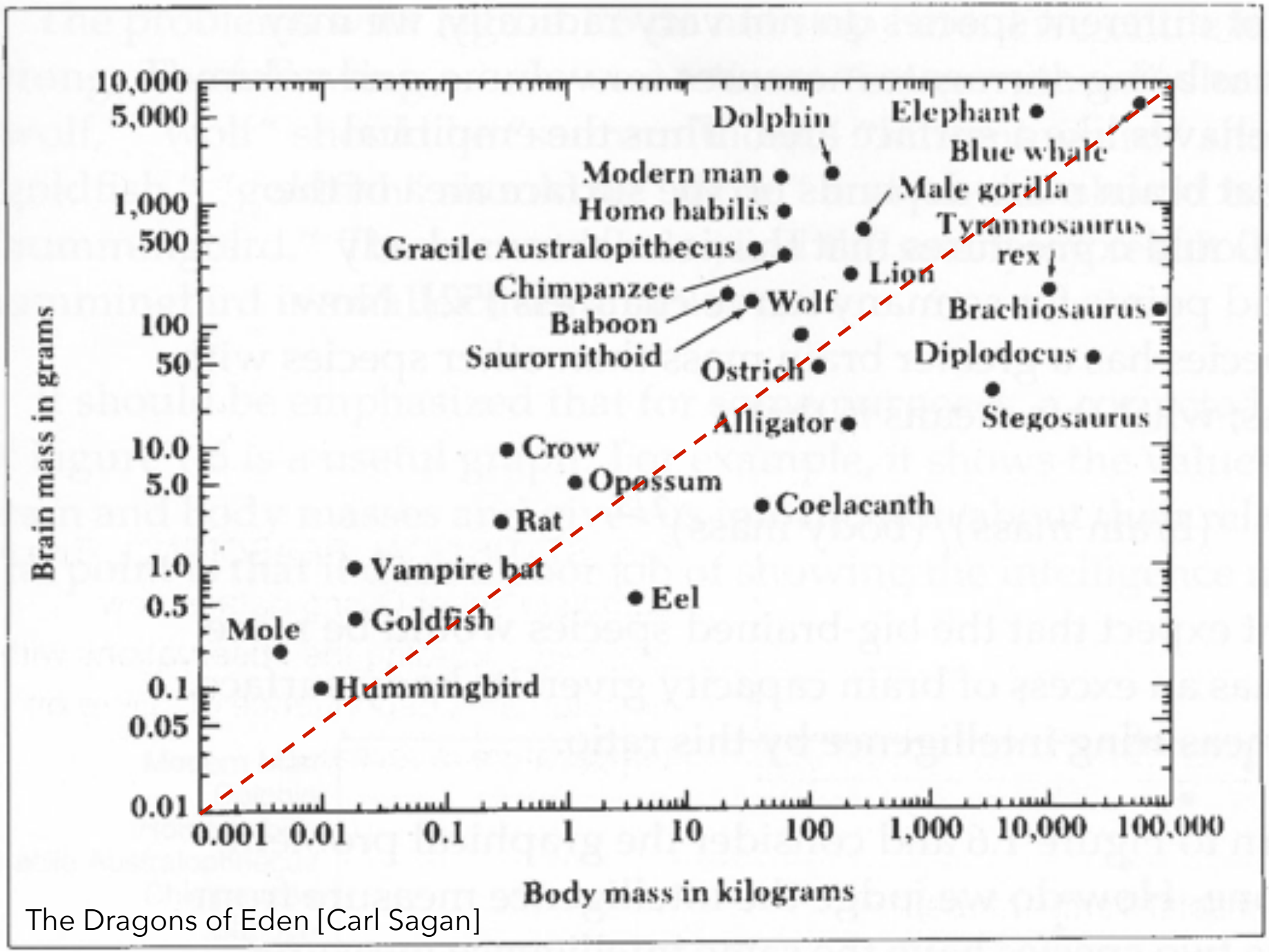


[New York Times 1981]

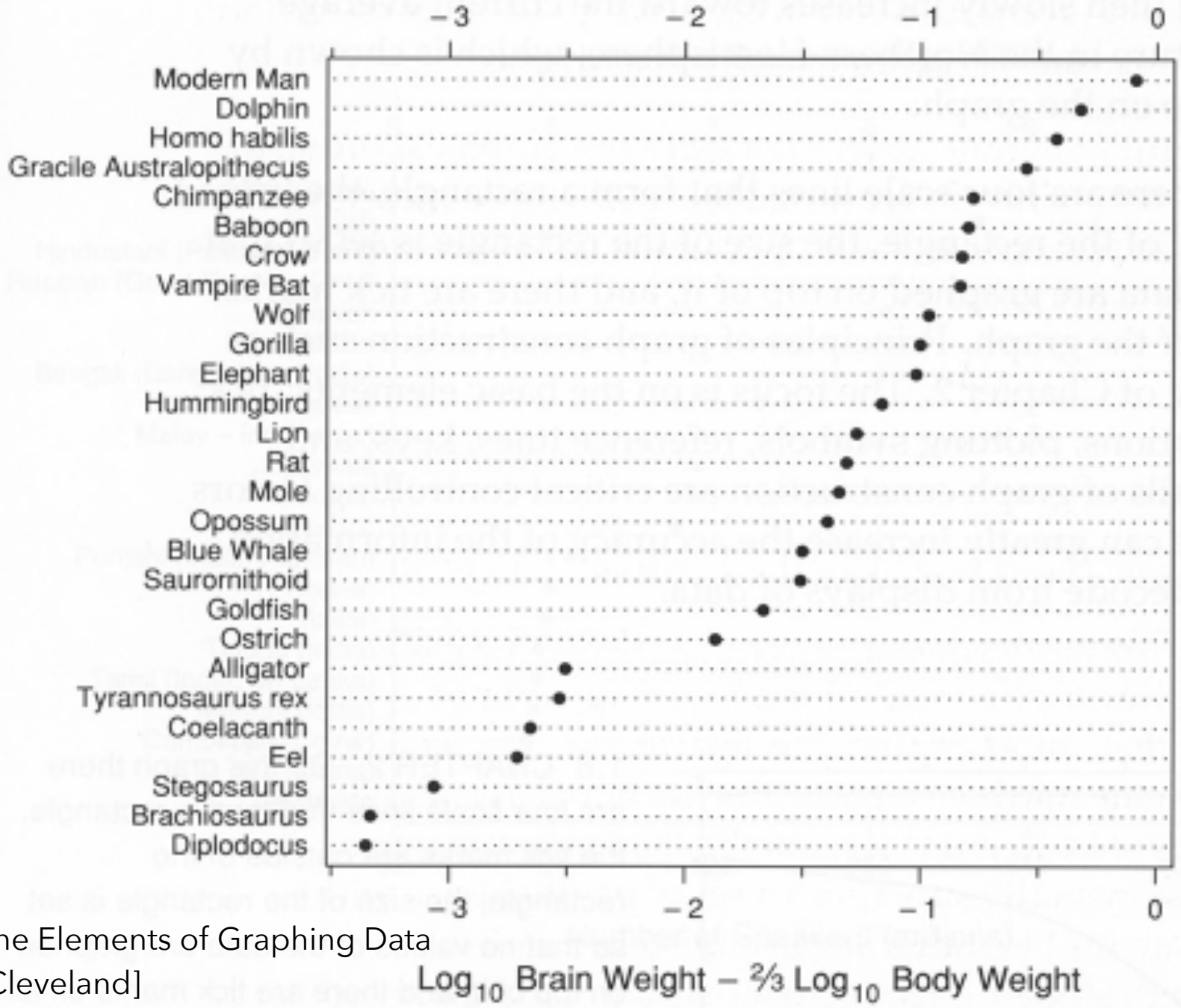
Answer Questions: Brain Power?



	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	

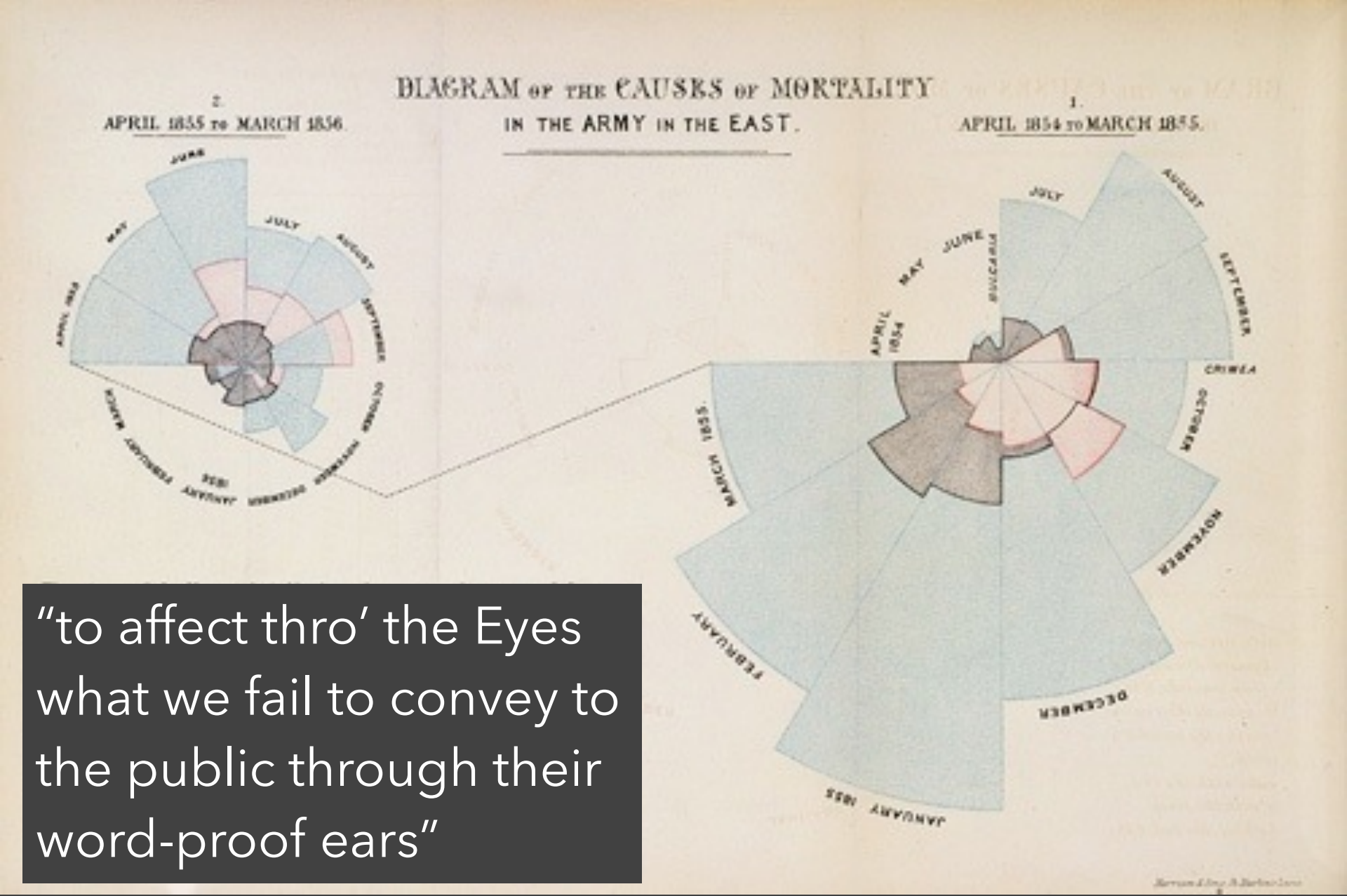


The Dragons of Eden [Carl Sagan]



The Elements of Graphing Data
[Cleveland]

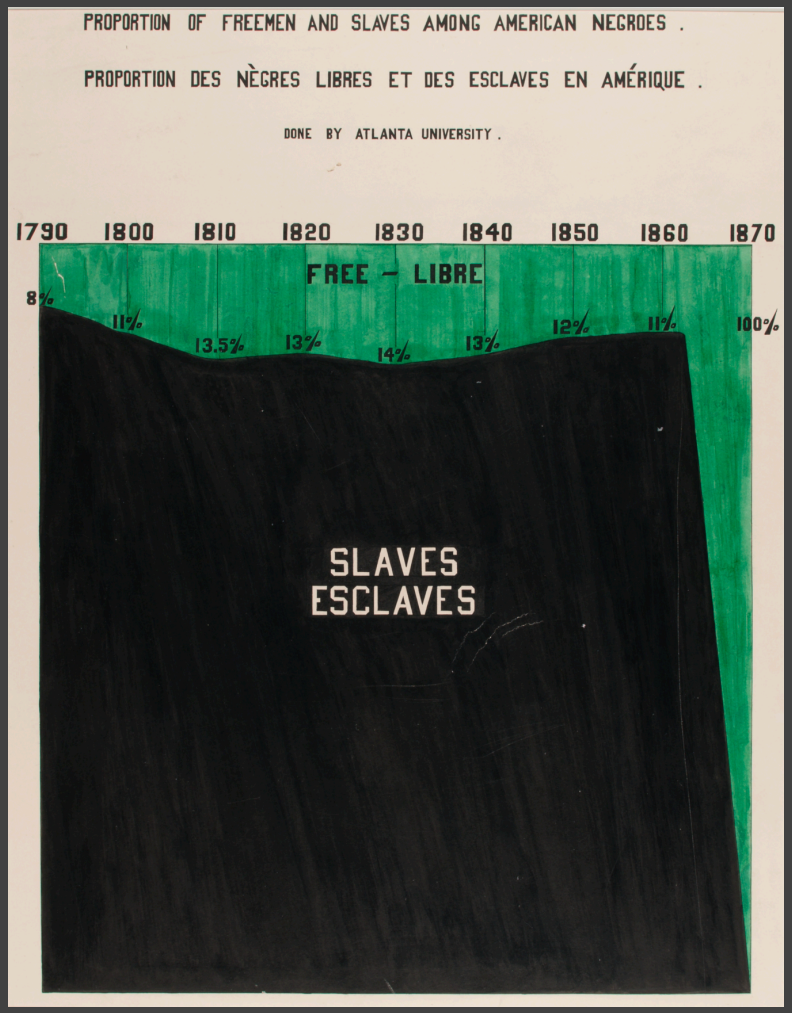
Convey Information



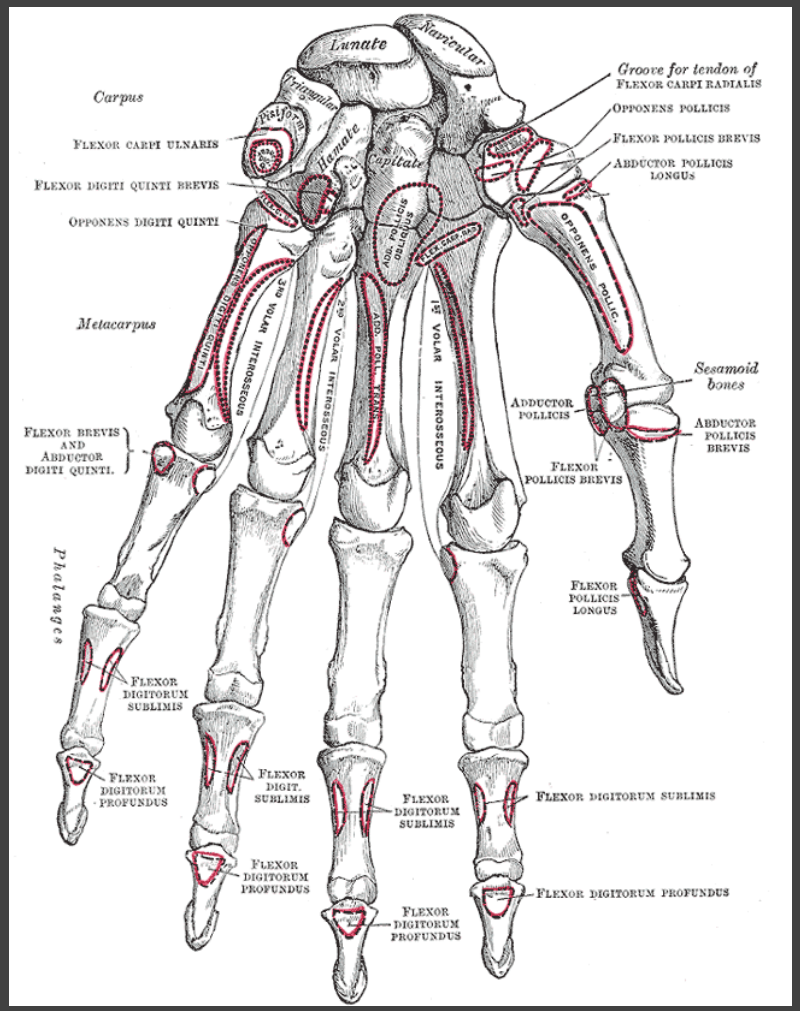
"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

Communicate, Inform, Inspire



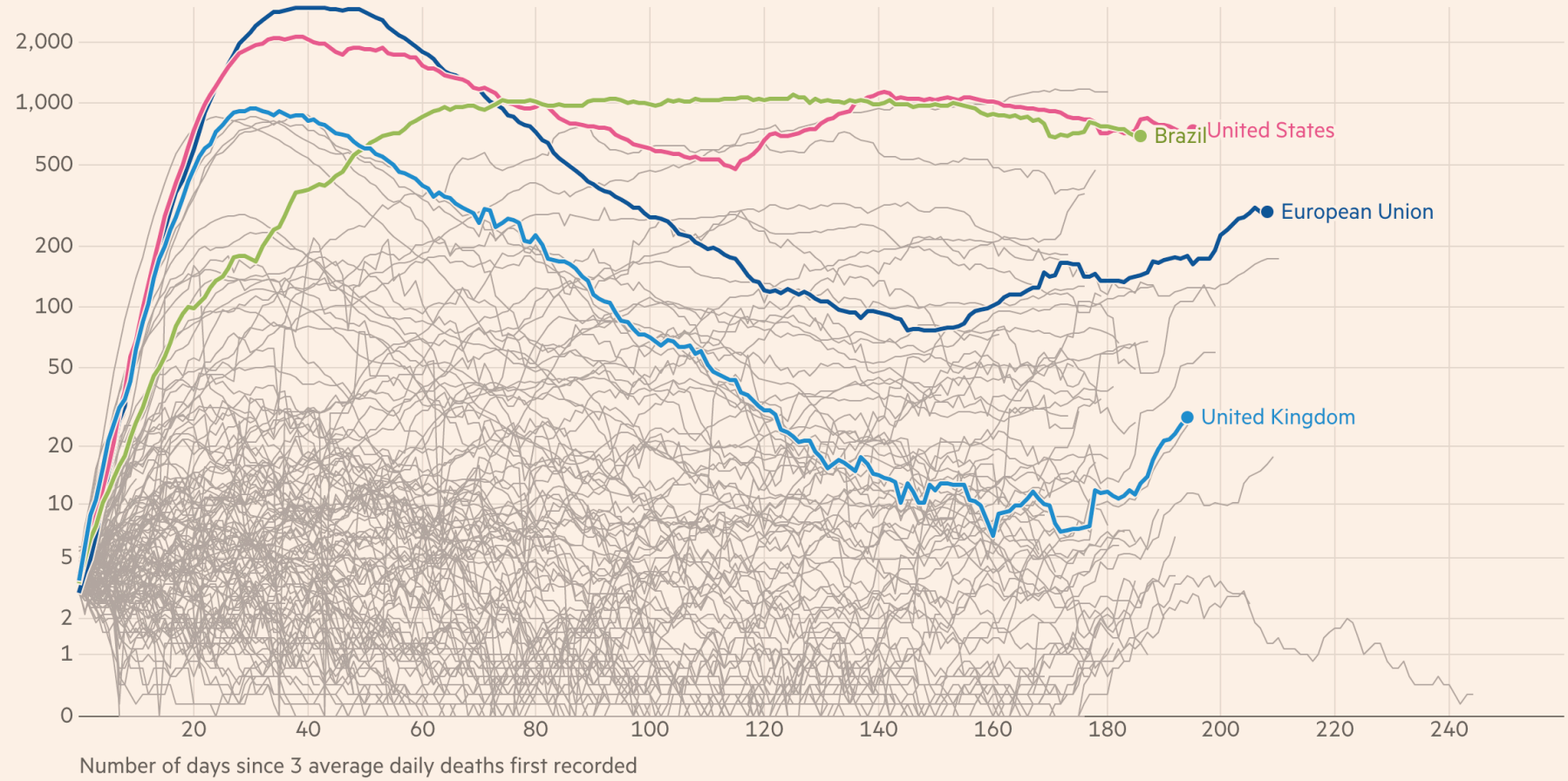
Visualizing Black America, Du Bois et al. 1900



Bones in hand, Gray's Anatomy 1918 ed.

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 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 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

Instructors

cse412-staff@cs

Jane Hoffswell OH: *Wed After Lecture*

Teaching Assistants

Aayush Chhabra OH: *By Appointment*

Dalton Hildreth OH: *Mon 12-1pm*

Kalyani Marathe OH: *Thur 11:30-12:30pm*

Yue Zhang OH: *TBD*

Yueqian Zhang OH: *Tue 5:30-6:30pm, and BA*

Aayush Chhabra

OH: By Appointment

Hi. I'm a senior graduating in June 2021.

Major: CS
Minor: Applied Math

My interests include :

- Data Viz
- Artificial Intelligence
- Machine Learning
- Robotics
- and all kinds of cool stuff with data.

Most recently, I interned as a Quantitative Researcher at a hedge fund (Citadel).

I love startups and I'm joining one after graduation. Talk to me if startups fascinate you.

Looking forward to learning and working with all of you :)

Dalton Hildreth

OH: Monday, 12pm - 1pm

2nd year Ph.D. student in GRAIL

I love computer graphics! (geometry, simulation) With a seasoning of ML

For fun, I also enjoy drawing, board gaming, game jams, and bouldering

I’m excited to learn with all of you!

Kalyani Marathe

OH : Thursday 11:30 am - 12:30 pm

- **Background**: Ph.D. student at GRAIL Lab, interested in Computer Vision, Machine Learning, Data Science, and Data Visualization
- **Work Experience** : 2 years as a Software Engineer at IDeaS, A SAS Company
- **Technical Experience**: Python, Java, Groovy, ScikitLearn, NumPy, SciPy, ScikitImage, OpenCV
- **Fun:** Quilling art and playing music

Yue Zhang

she/her
Office hours: Friday 6-7pm

6th year (😨) PhD student in computational biology

Research areas: interactive data analysis and machine learning for single-cell genomics, natural language processing on scientific literature

Fun: books, games, honestly just surviving right now

Yueqian Zhang

OH: Tuesday 5:30 pm - 6:30 pm & By Appointment

I am a junior **majoring in CS** and **minoring in Education**.

I am interested in **HCI**.

Outside of CS I enjoy **anime** and **drawing**.

Come talk to me! I'd love to meet everyone :)

Some of my artworks:

Attending Lectures via Zoom

If you are comfortable doing so, please turn on your cameras. Otherwise, consider adding a profile picture to Zoom!

Important Links

Course Website:

https://courses.cs.washington.edu/courses/cse412/21sp/

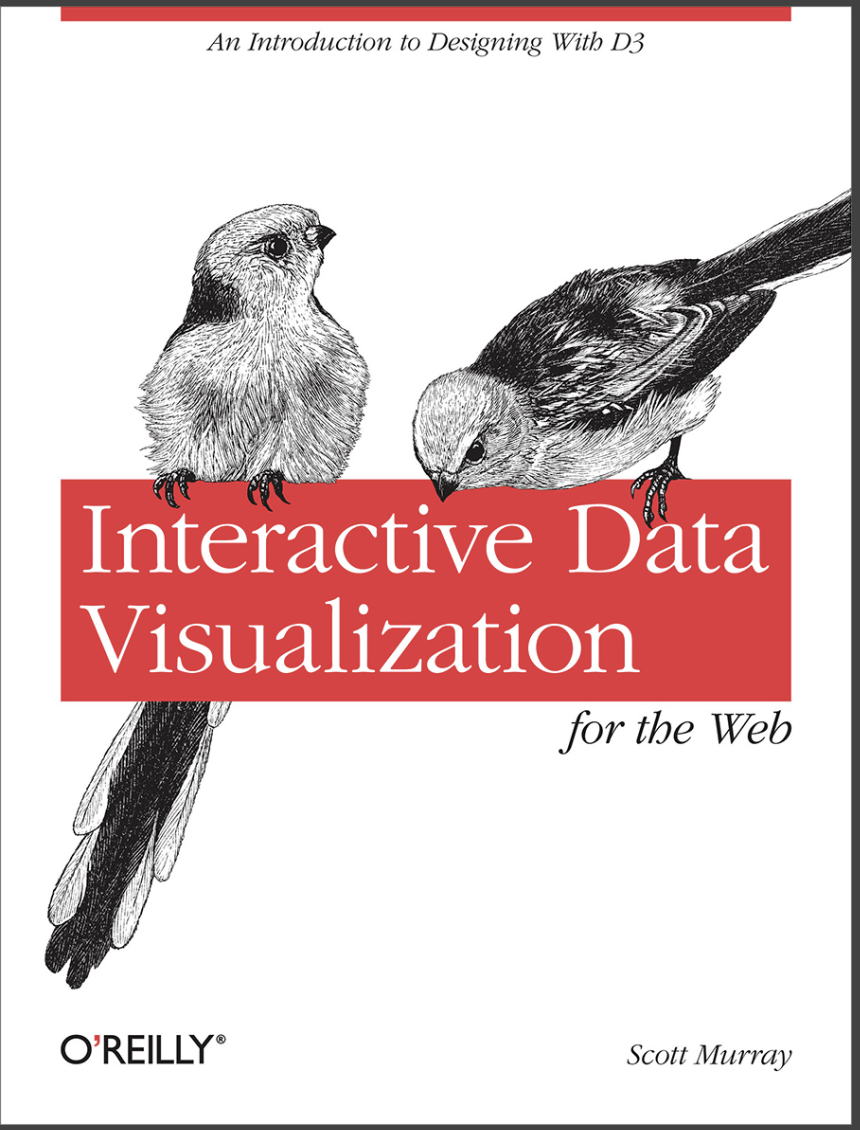
Canvas:

https://canvas.uw.edu/courses/1475556

Ed Discussion:

https://edstem.org/us/courses/4910/discussion/

Textbook

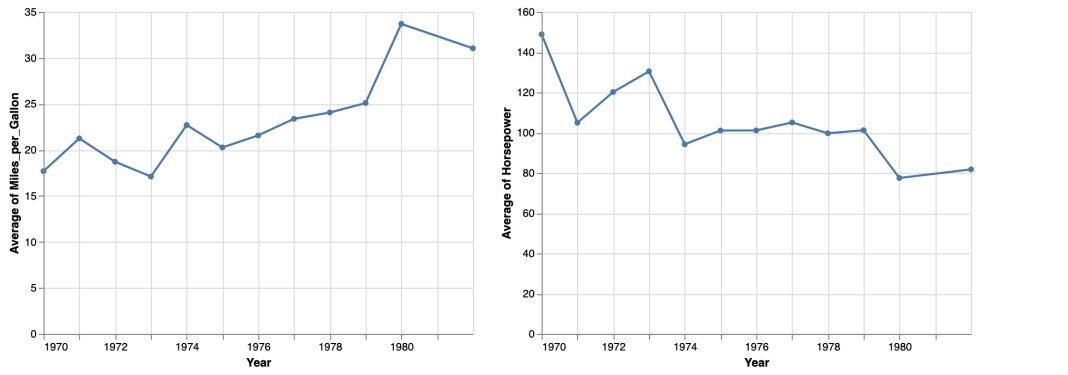


Interactive Data Visualization for the Web, 2nd Edition

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

We will be using **D3 v6**.
https://d3js.org

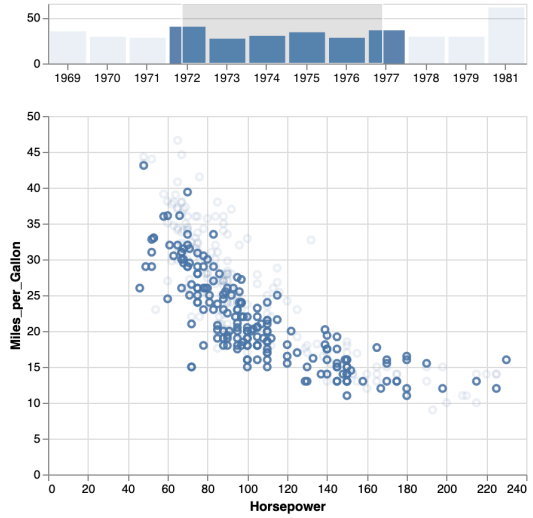
Interactive Vega-Lite Notebooks




```
{
  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.

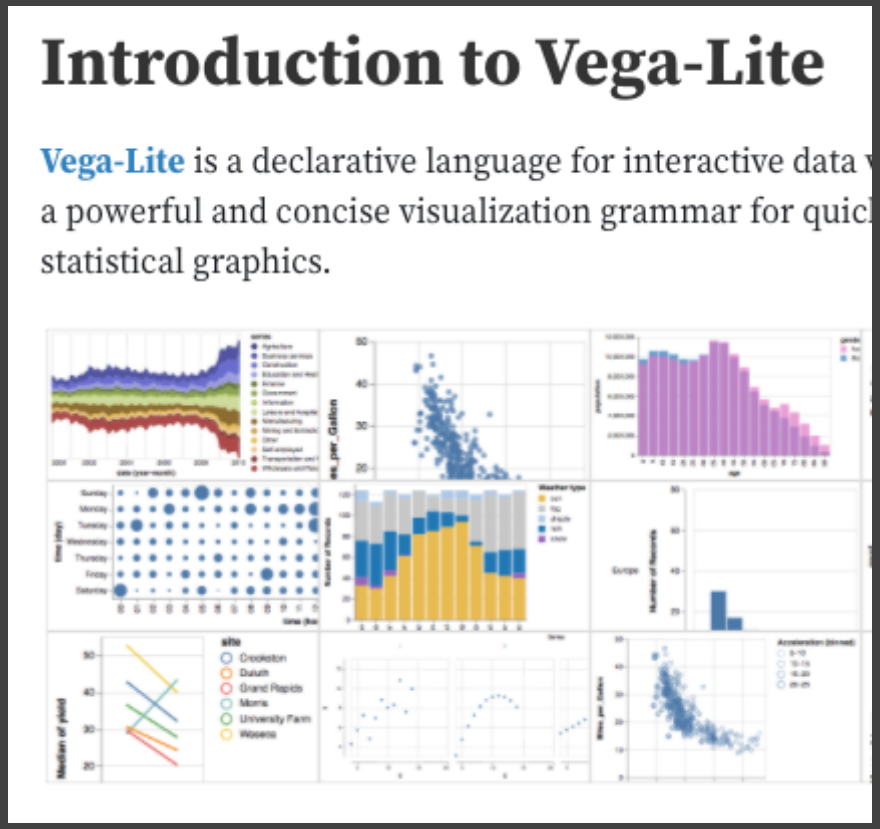
Readings

From books, notebooks, and linked articles.

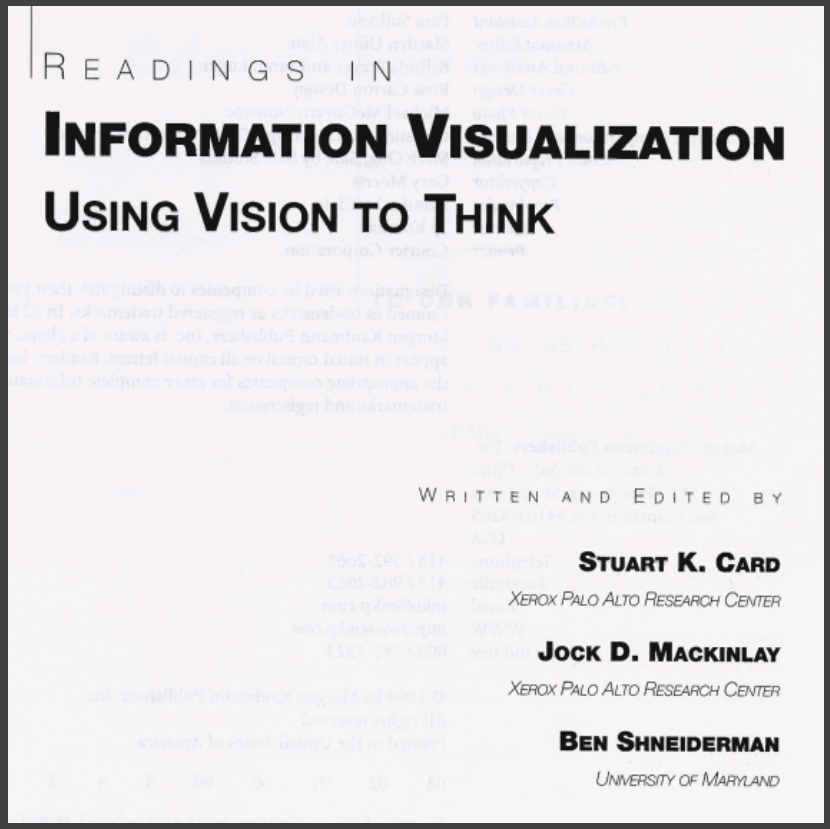
Material in class will loosely follow readings.

Readings should be read by start of class.

Required Readings for 3/29 (today)



Notebook: Introduction to Vega-Lite.



Chapter 1: Information Visualization.
Stuart Card, Jock Mackinlay, and Ben Shneiderman.
Readings in Information Visualization. 1999.

Required Readings for Wed 3/31

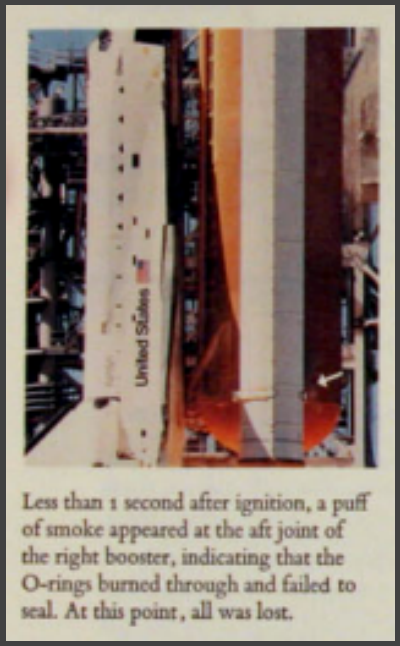
Data Types, Graphical Marks, and Visual Encoding Channels

A visualization represents data using a collection of *graphical marks* such as bars, lines, and point symbols. The attributes of a mark — such as its position, shape, size, or color — serve as *channels* in which we can encode underlying data values.

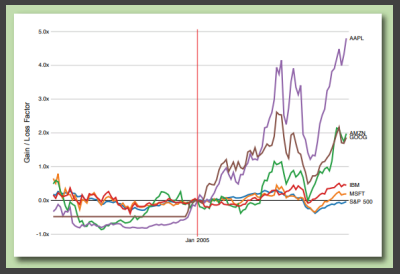
With a basic framework of *data types*, *marks*, and *encoding channels*, we can concisely create a wide variety of visualizations. In this notebook, we explore each of these elements and show how to use them to create custom statistical graphics.

Notebook: Data Types, Graphical Marks, and Visual Encoding Channels.

Optional Readings for Week 1



MON Decision to Launch the Challenger.



FRI A Tour through the Visualization Zoo.

CS-TR-3665 July 1996
ISR-TR-96-66

The Eyes Have It:
A Task by Data Type Taxonomy
for Information Visualizations

Ben Shneiderman
Department of Computer Science
Human-Computer Interaction Laboratory,
and Institute for Systems Research
University of Maryland, College Park, Maryland 20742 USA
ben@cs.umd.edu, http://www/cs.umd.edu/projects/hcil/

WED The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.

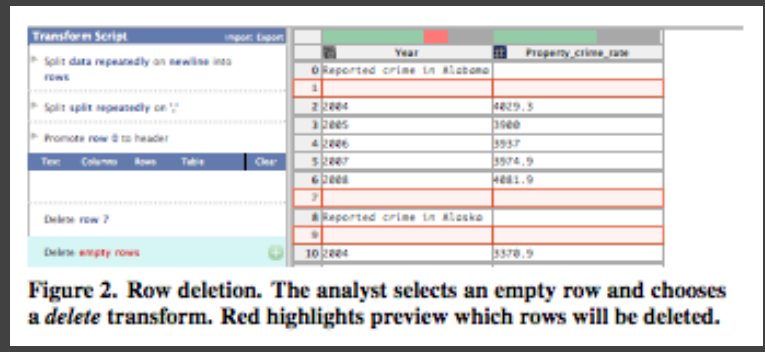


Figure 2. Row deletion. The analyst selects an empty row and chooses a *delete* transform. Red highlights preview which rows will be deleted.

THUR Wrangler: Interactive Visual Specification of Data Transformation Scripts

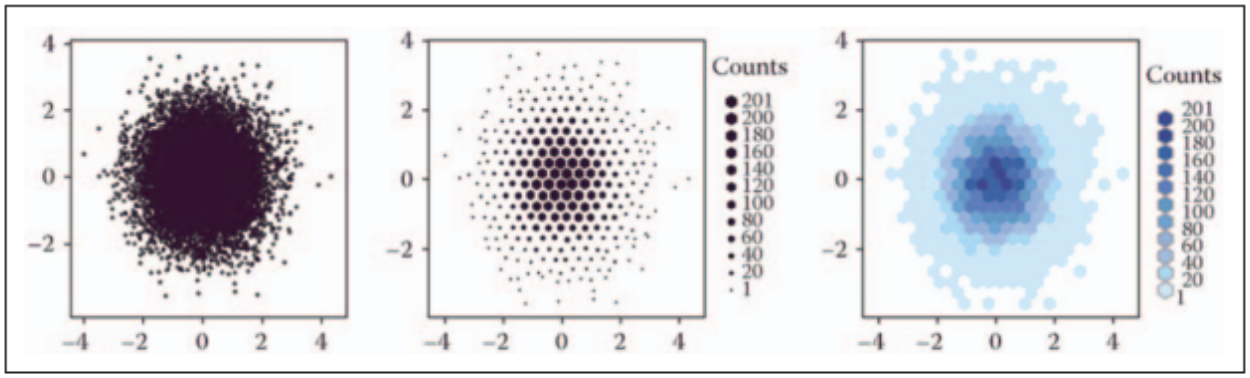


Figure 3. Visualizing 'raw' data at scale, taken from Carr et al.[16] (a) A traditional scatter plot. (b) A binned plot using a size encoding. (c) A binned plot using a color encoding. Note the discontinuity in color between 0 and 1, making cells with a single element readily apparent.

THUR Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data.

Assignments

CP Class Participation (10%)

A1 Visualization Design (10%) - *Due 4/5 (next Monday)*

A2 Exploratory Data Analysis (15%) - *Due 4/19*

A3 Ethical & Deceptive Visualization (20%) - *Due 5/3*

Peer Evaluation (5%) - *Due 5/10*

FP Final Project (40%)

Proposal - *Due 5/7*

Milestone Prototype - *Due 5/21*

Demonstration Video - *Due 5/31*

Final Prototype - *Due 6/7*

Weekly Discussion Post and Quiz

First discussion thread and quiz will be posted on Ed later this afternoon, due **Mon 4/5, 11:59pm**.

Short quizzes reinforce course concepts.
Quizzes are graded on participation not score.

Share substantive discussion post on Ed about course readings or lecture material.

Weekly Discussion Post and Quiz

First discussion thread and quiz will be posted on Ed later this afternoon, due **Mon 4/5, 11:59pm**.

Short quizzes reinforce course concepts.
Quizzes are graded on participation not score.

Share substantive discussion post on Ed about course readings or lecture material.

- Critiques of arguments made in the papers
- Analysis of implications or future directions
- Clarification of some point or detail presented in the class
- Insightful questions about the readings or answers to other's questions
- Links and commentary for relevant web resources or examples

Weekly Discussion Post and Quiz

First discussion thread and quiz will be posted on Ed later this afternoon, due **Mon 4/5, 11:59pm**.

Short quizzes reinforce course concepts.
Quizzes are graded on participation not score.

Share substantive discussion post on Ed about course readings or lecture material.

Posts are due each Monday by 11:59pm up through week 8. You have 1 free “pass” for the quarter.

Final Project

Produce **interactive web-based visualizations**

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

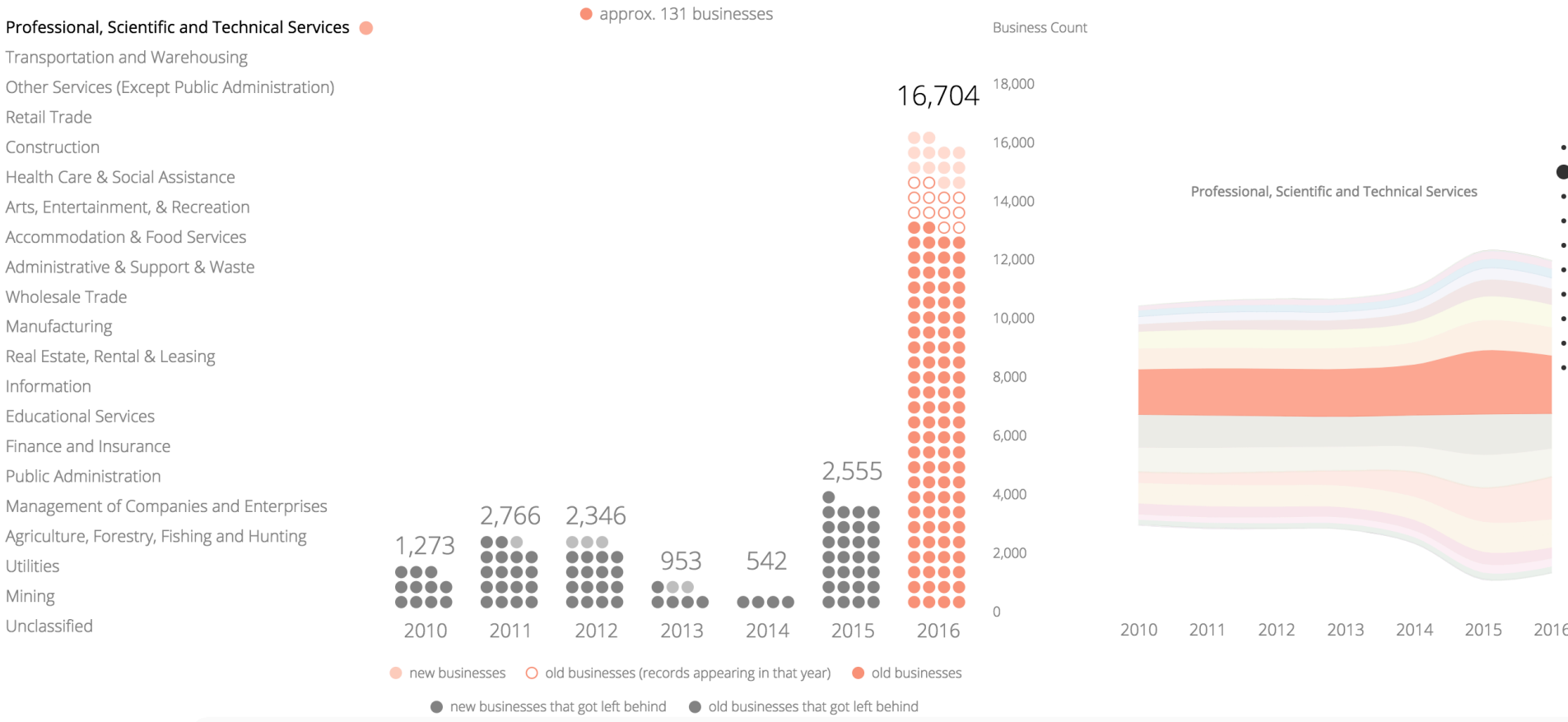
Final deliverables and **video presentation**

Submit and **publish online** (GitHub)

Projects from **previous classes** have been:

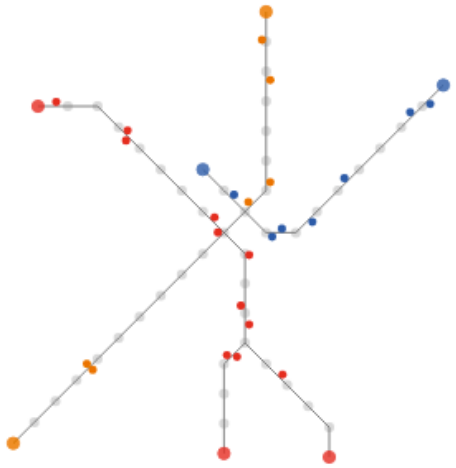
- Published as research papers
- Shared widely (some in the New York Times!)
- Released as successful open source projects

Inspiration…



Change In Times (CSE 442, Spring '17)

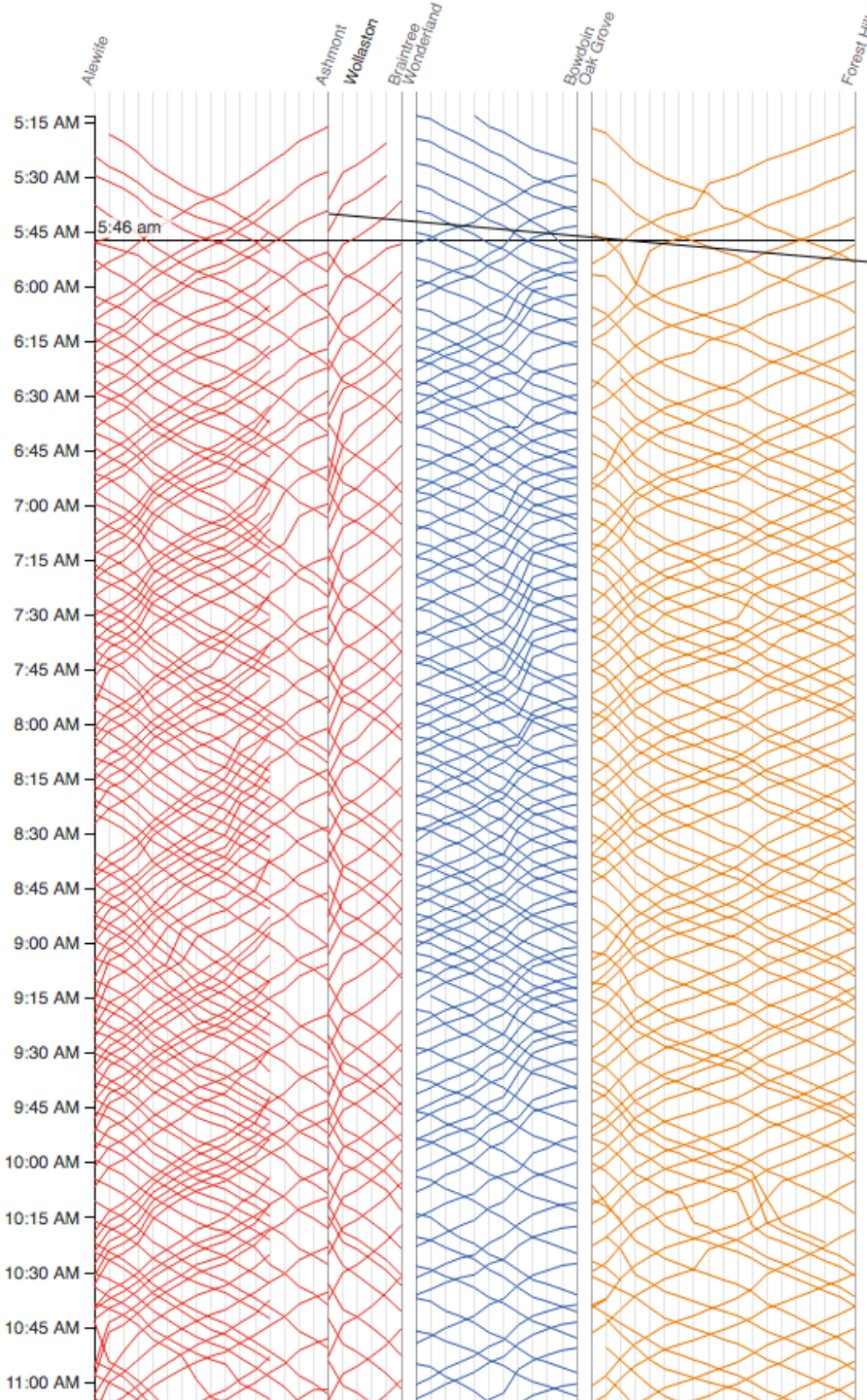
Gunnar Olson, Halden Lin, Lilian Liang, and Shobhit Hathi



Locations of each train on the red, blue, and orange lines at 5:46 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

Questions?

A1: Visualization Design

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: Visualization Design

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 via Canvas; see A1 page)
Image of your visualization (PNG or JPG format)
Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm PT, next Monday April 5th**.