THE PURPOSE OF VISUALIZATION

CS 448B | Fall 2025

MANEESH AGRAWALA

1

How much data (bytes) did we produce in 2023?

2

2023: 123 zetabytes

Statistica 2024

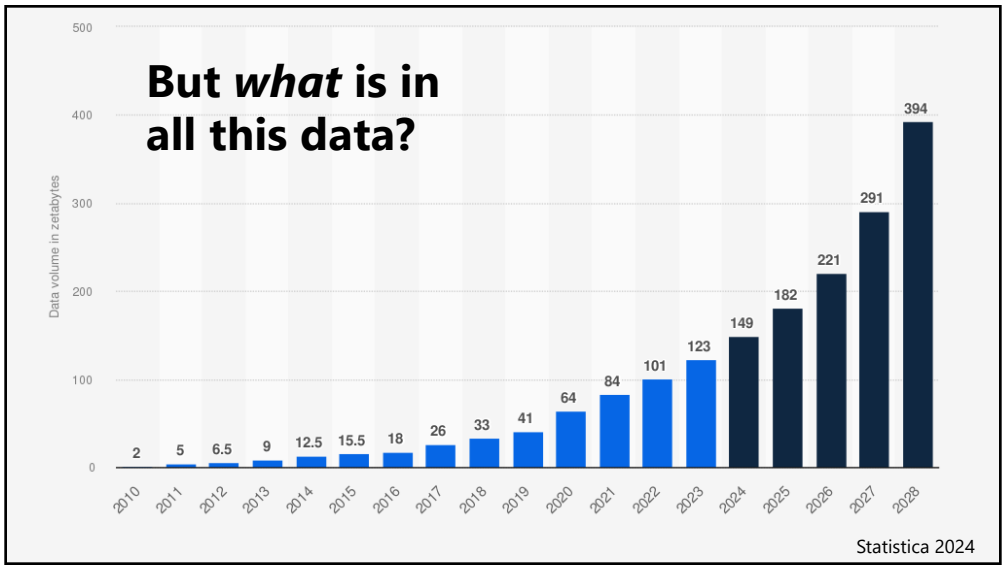
3

2023: 123 zetabytes

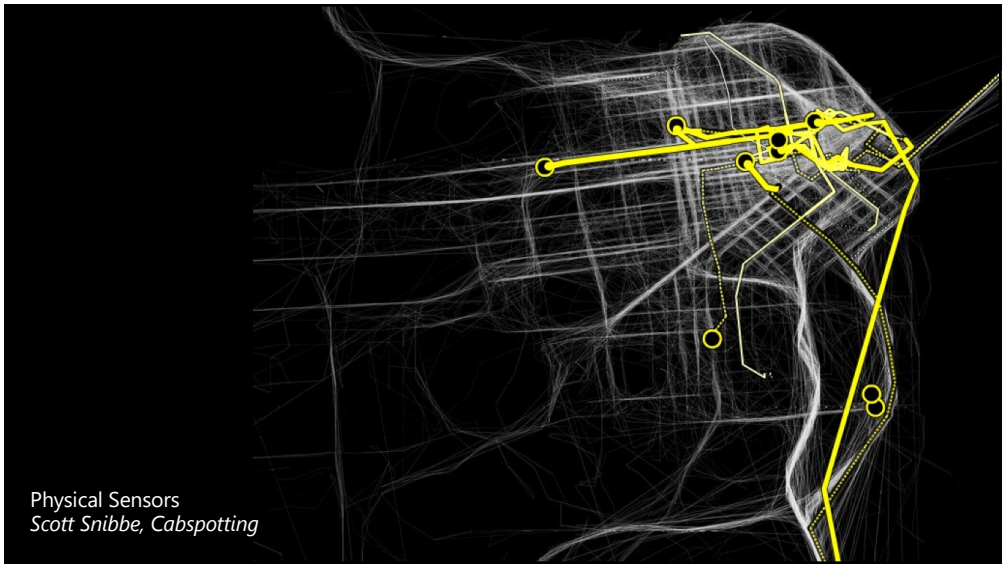
~2x increase every 2 years since 2010

Statistica 2024

4


But *what* is in all this data?
Data volume in zetabytes
500
400
300
200
100
0
2
5
6.5
9
12.5
15.5
18
26
33
41
64
84
101
123
149
182
221
291
394
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
Statistica 2024

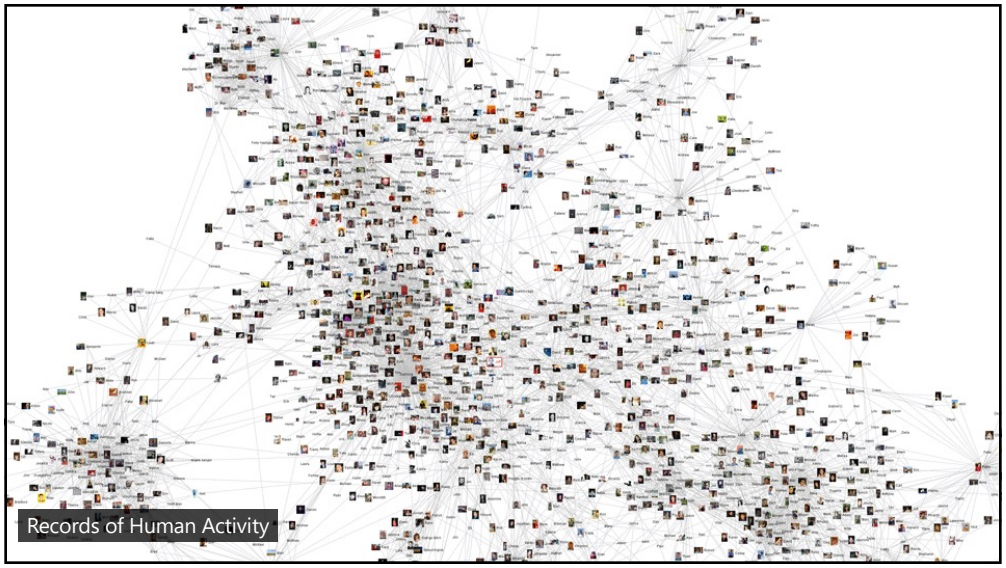
5


Physical Sensors
Scott Snibbe, Cabspotting

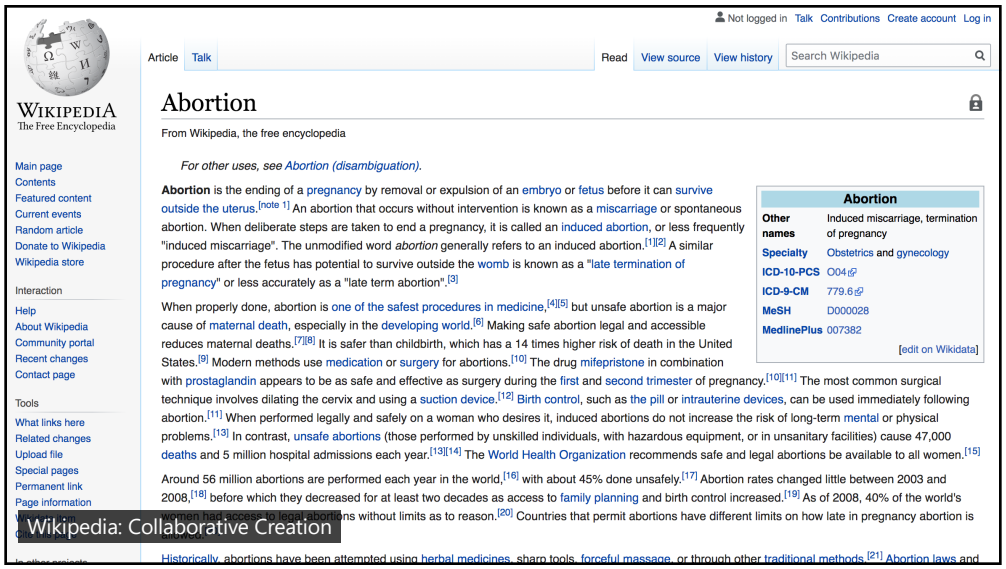
6


HR/ECG
1/min
97
Art
mmHg sys/dia
82/60
SpO2
%
99
RR/CO2
1/min
HR/ECG
1/min
79
Art
mmHg sys/dia
152/79
SpO2
%
95
RR/CO2
1/min
HR/ECG
1/min
64
Art
mmHg sys/dia
93/55
SpO2
%
99
RR/CO2
1/min
HR/ECG
Health & Medicine

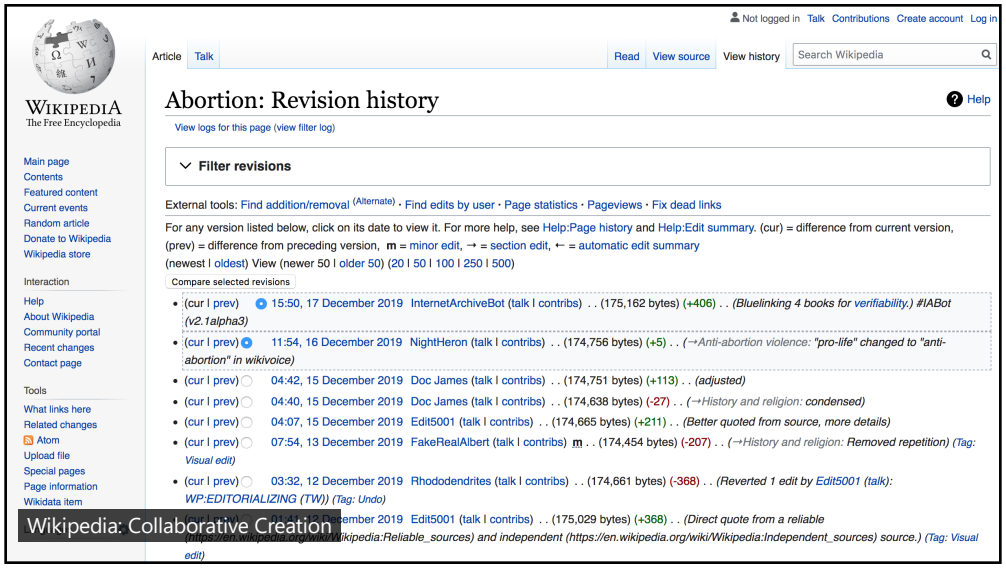
7


Records of Human Activity

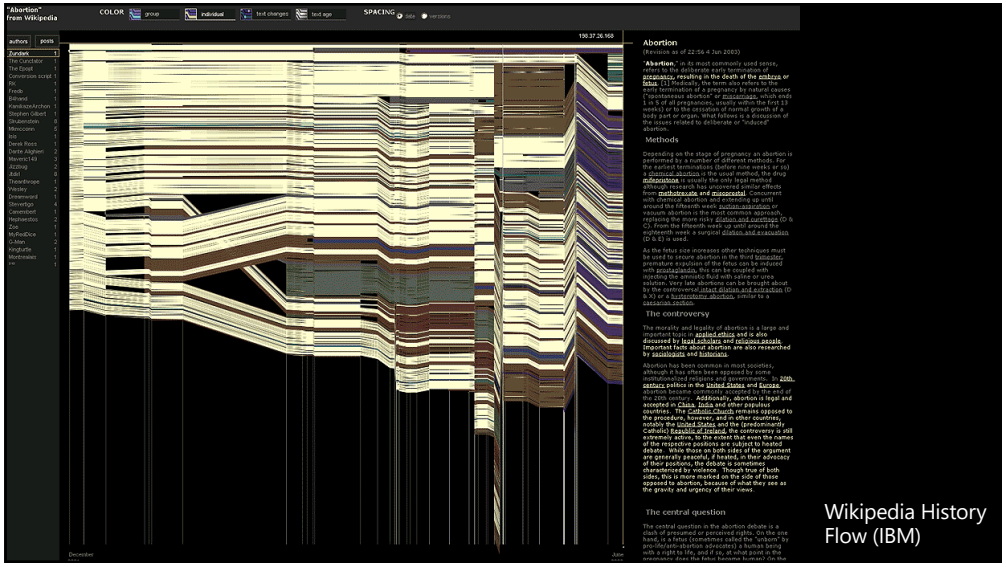
8


Not logged in Talk Contributions Create account Log in
Article Talk
Read View source View history
Search Wikipedia
WIKIPEDIA
The Free Encyclopedia
Abortion
From Wikipedia, the free encyclopedia
For other uses, see Abortion (disambiguation).
Abortion
Other names
Induced miscarriage, termination of pregnancy
Specialty
Obstetrics and gynecology
ICD-10-PCS
O04
ICD-9-CM
779.6
MeSH
D000028
MedlinePlus
007382
[edit on Wikidata]
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikipedia: Collaborative Creation

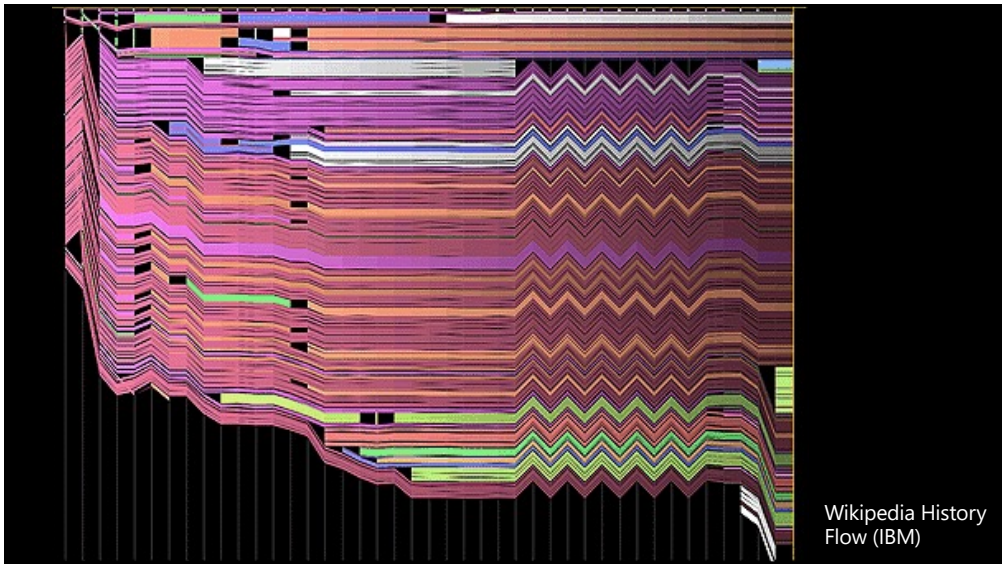
9


Not logged in Talk Contributions Create account Log in
Article Talk
Read View source View history
Search Wikipedia
WIKIPEDIA
The Free Encyclopedia
Abortion: Revision history
Help
View logs for this page (view filter log)
Filter revisions
External tools: Find addition/removal (Alternate) · Find edits by user · Page statistics · Pageviews · Fix dead links
Compare selected revisions
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Atom
Upload file
Special pages
Page information
Wikidata item
Wikipedia: Collaborative Creation

10

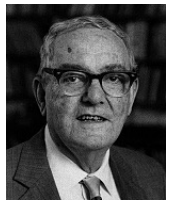

"Abortion" from Wikipedia
COLOR
SPACING
Abortion
Methods
The controversy
The central question
Wikipedia History Flow (IBM)

11


Wikipedia History Flow (IBM)

12

"What information consumes is rather obvious: **it consumes the attention of its recipients**. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."



Herb Simon
as quoted by Hal Varian
Scientific American
September 1995

13

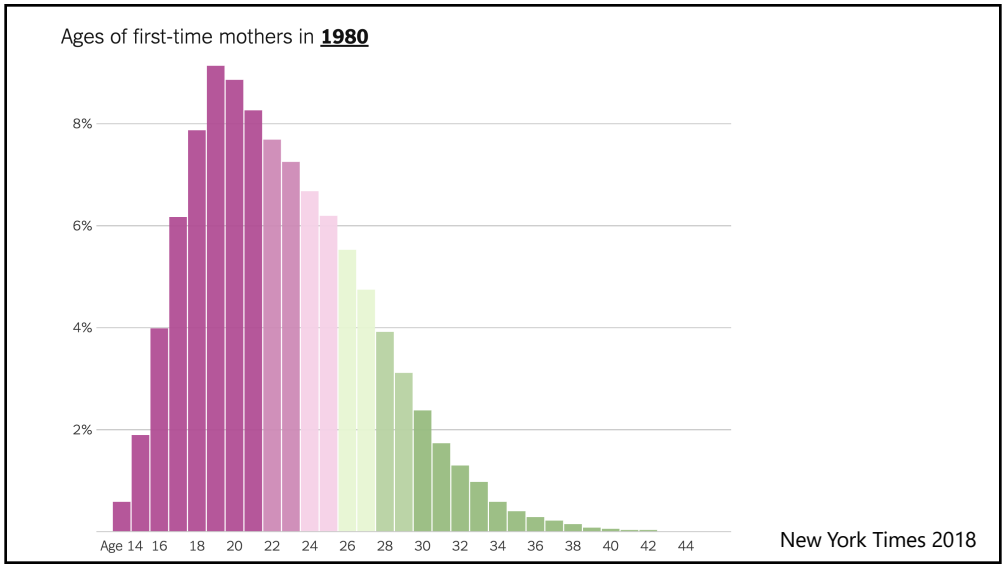
"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it."

Hal Varian
Google's Chief Economist
The McKinsey Quarterly
January 2009

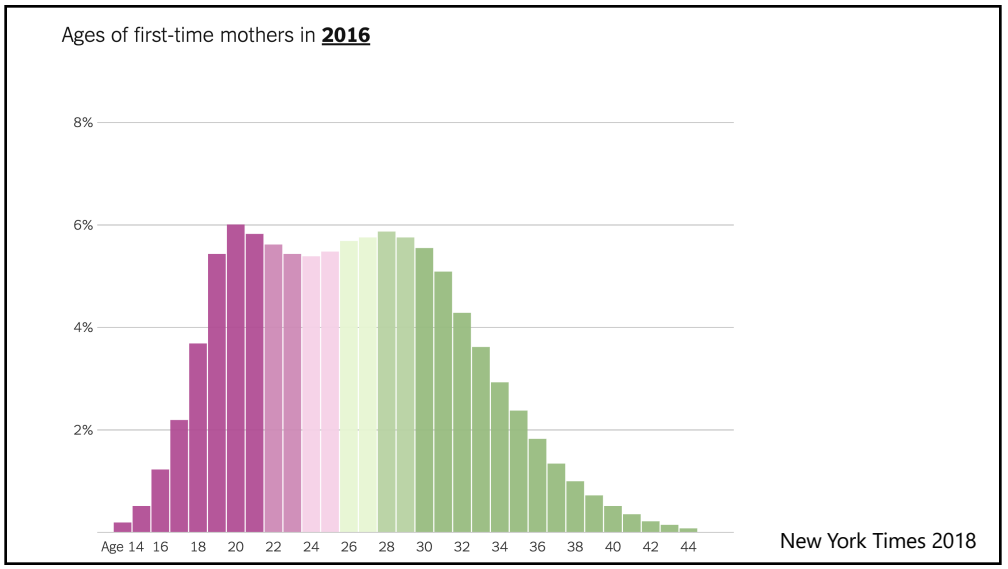
14

WHAT IS VISUALIZATION?

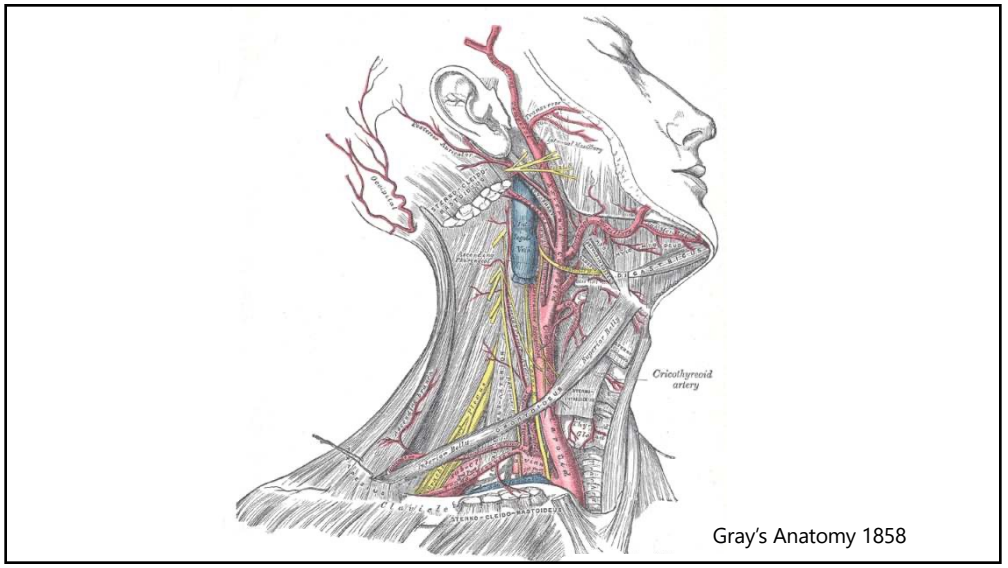
15



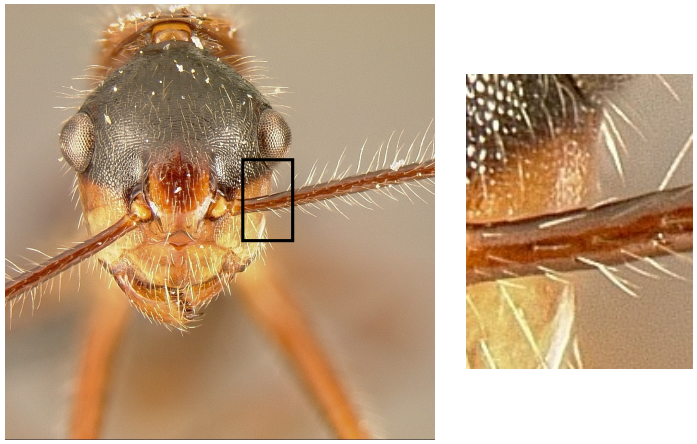
16


Ages of first-time mothers in **2016**
8%
6%
4%
2%
Age 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
New York Times 2018

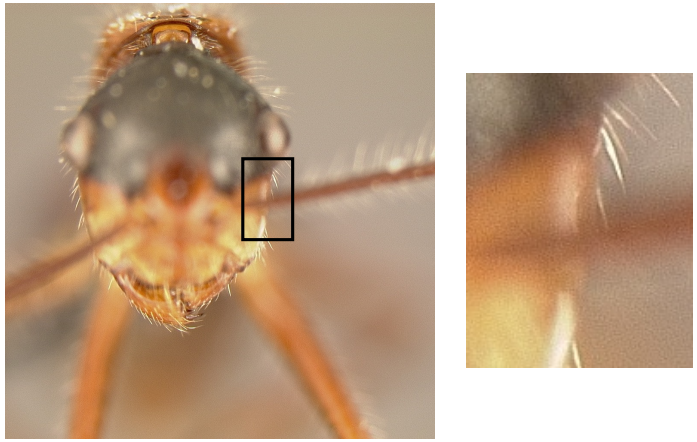
17


Cricothyroid artery
Gray's Anatomy 1858

18



19



20

Combination of in-focus regions

[Agarwala 2004]

21

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]

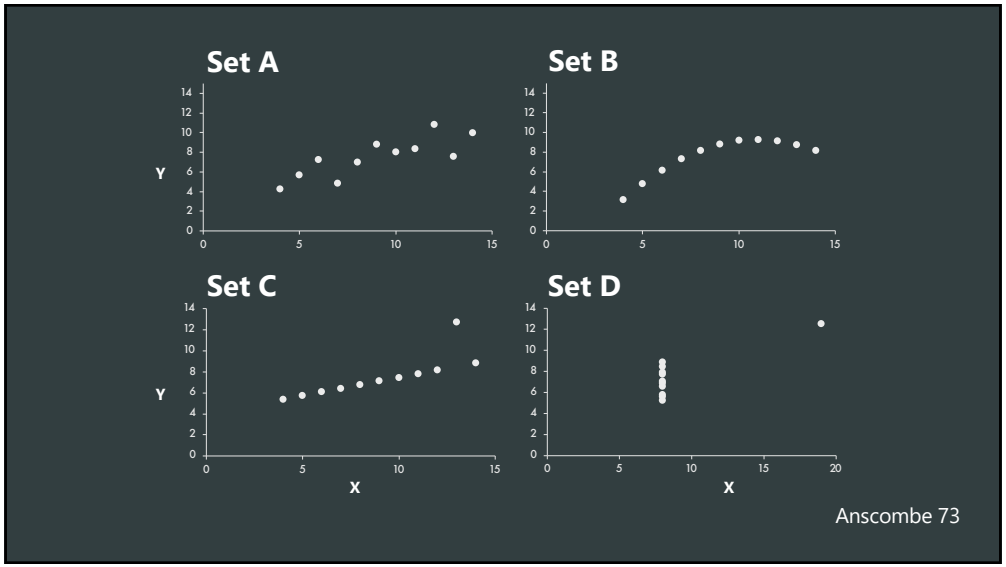
24

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
$Y = 3 + 0.5\ X$
$R^2 = 0.67$

25



26

WHY DO WE CREATE VISUALIZATIONS?

27

Why do we create visualizations?

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

28

THE PURPOSE OF VISUALIZATION

Record information
- Photographs, blueprints, …

Support reasoning about information (analyze)
- Process and calculate
- Reason about data
- Expand memory

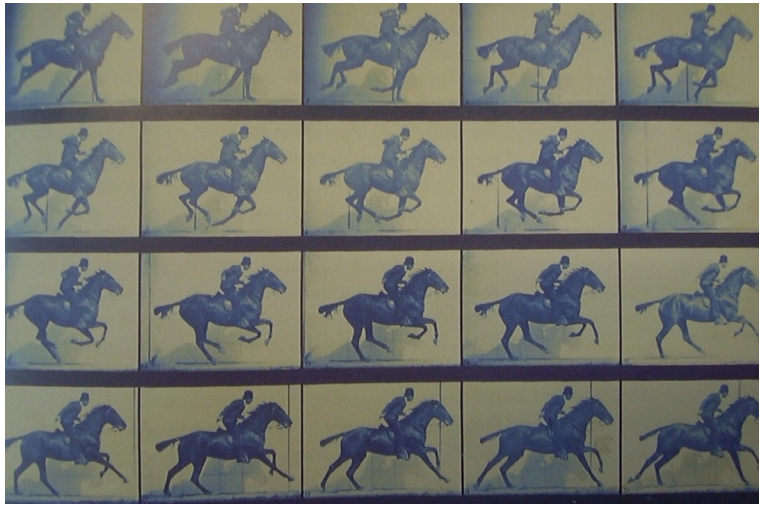
Communicate, inform, inspire (present)
- Share and persuade
- Emphasize important aspects of data

29

RECORD

30

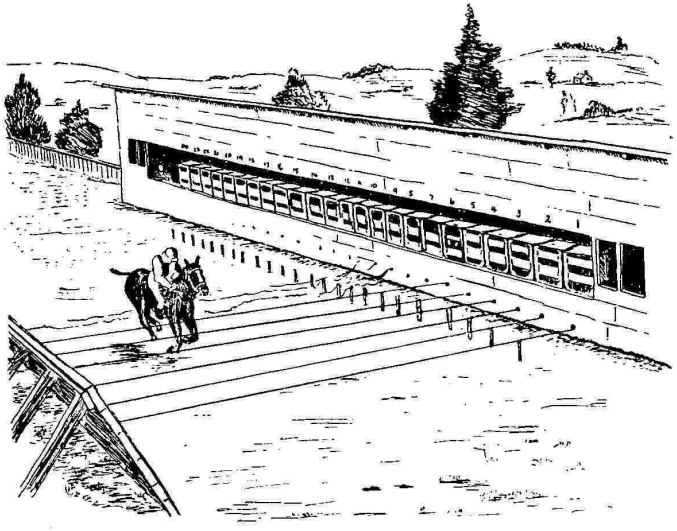
Answer question



Gallop, Bay Horse "Daisy" [Muybridge 1884-86]

31

Answer question



Muybridge's setup [1884-86]

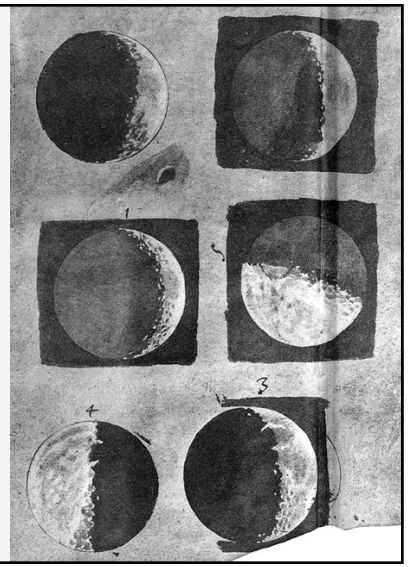
32

Photograph:
Phases of the moon

Yaorusheng / Getty Images

33

Drawing:
Phases of the moon



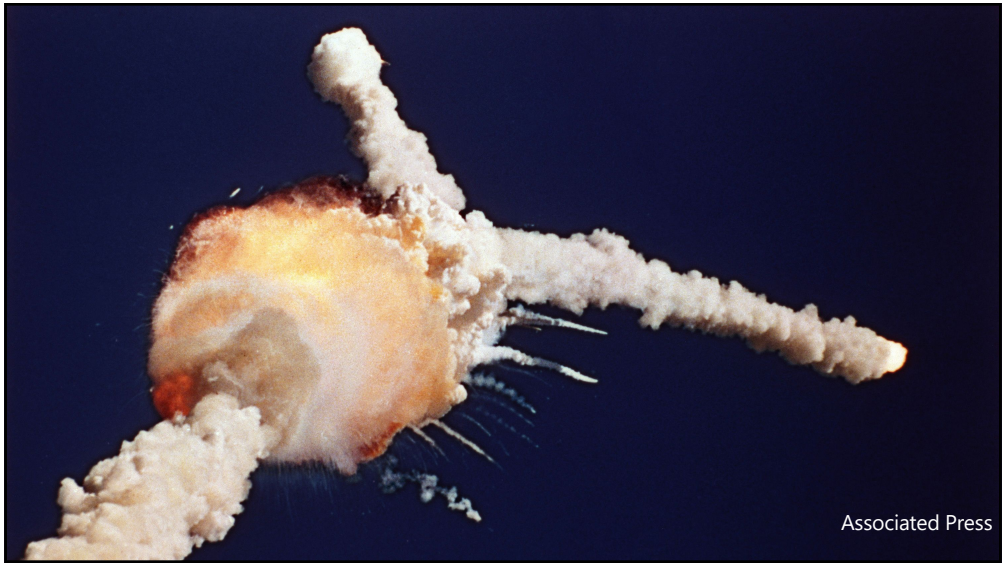
Galileo's drawings of the phases of the moon from 1616
http://galileo.rice.edu/sci/observations/moon.html

34

SUPPORT REASONING

37

38



39

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	350°-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 (90°), (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

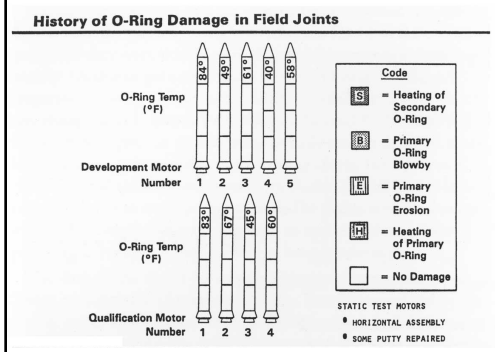
HISTORY OF O-RING TEMPERATURES (DEGREES - F)

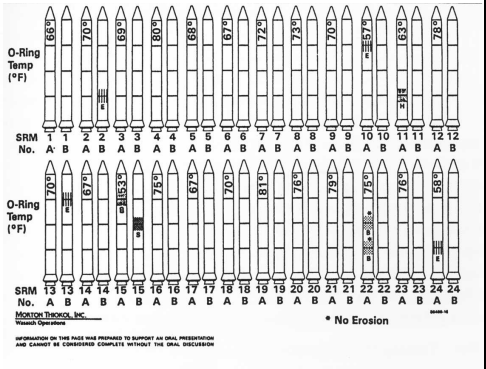
MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-1	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

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

40

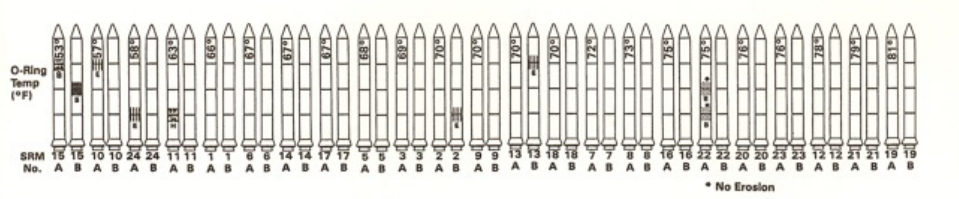
Make a decision: Challenger





41

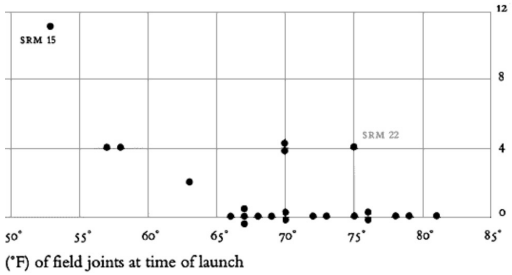
Make a decision: Challenger



Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

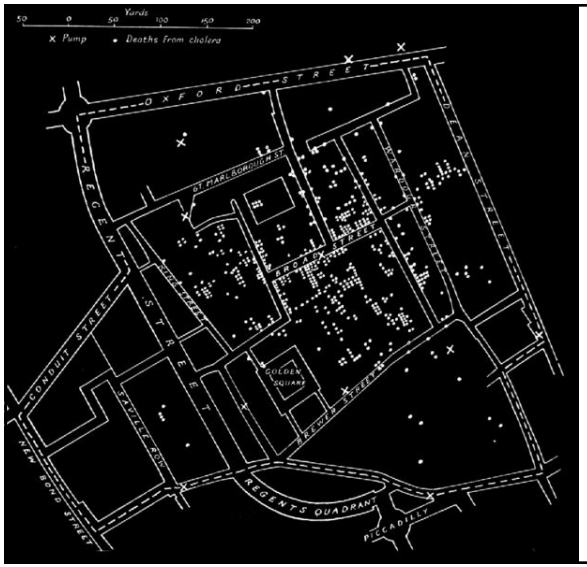
42

Make a decision: Challenger



Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

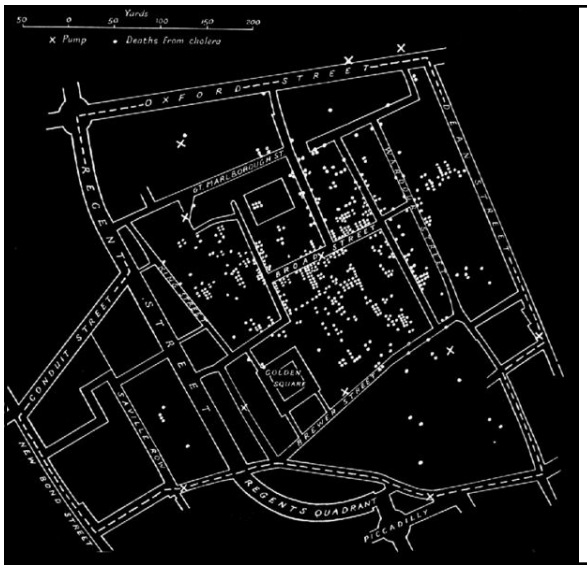
43



Data in context

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

44



Data in context

Used map to support hypothesis Broad St. pump was the cause. [from Tufte 83]

45

Expand memory: Multiplication

Class exercise

48

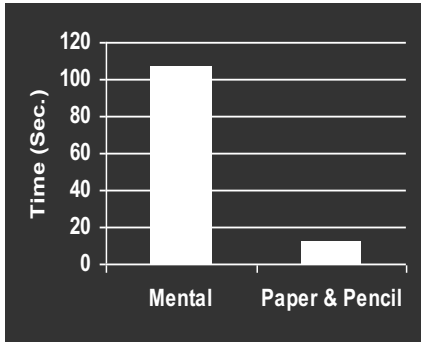
Expand memory: Multiplication

```
  34
x 87
```

49

Expand memory: Multiplication

```
  34
x 87
 238
2720
2958
```



50

Most powerful brain?

Microsoft Excel - animal.xls

File Edit View Insert Format Tools Data Window Help

A1 ID

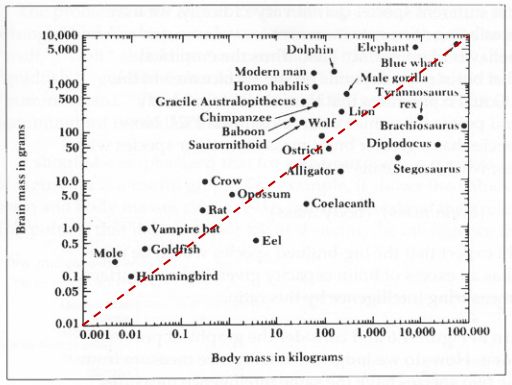
	A	B	C	D	E
1	ID	Name	Body Weight	Brain Weight	
2	1	Lesser Short-tailed Shrew	5	0.14	
3	2	Little Brown Bat	10	0.25	
4	3	Mouse	23	0.3	
5	4	Big Brown Bat	23	0.4	
6	5	Musk Shrew	48	0.33	
7	6	Star Nosed Mole	60	1	
8	7	Eastern American Mole	75	1.2	
9	8	Ground Squirrel	101	4	
10	9	Tree Shrew	104	2.5	
11	10	Golden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

54

Most powerful brain?



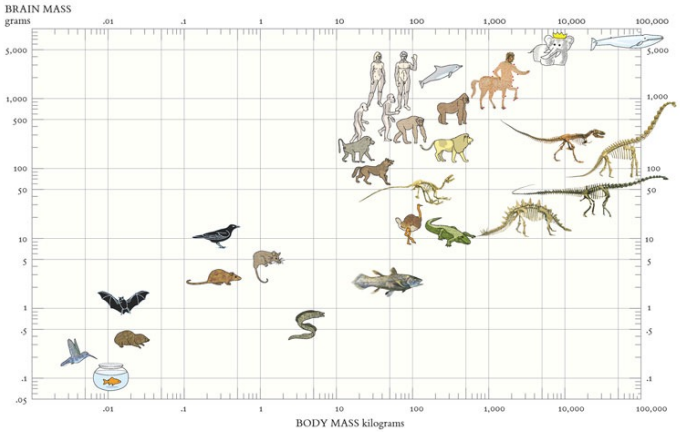
The Dragons of Eden [Carl Sagan 1977]

55

COMMUNICATE, INFORM, INSPIRE

57

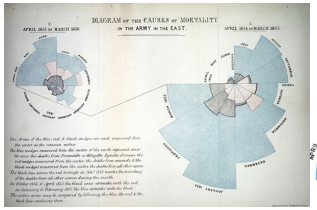
Tell a story

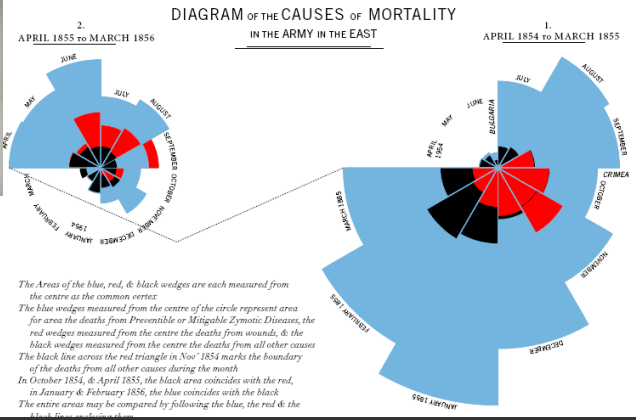


Beautiful Evidence [Tufte 2006]

58

Present an argument

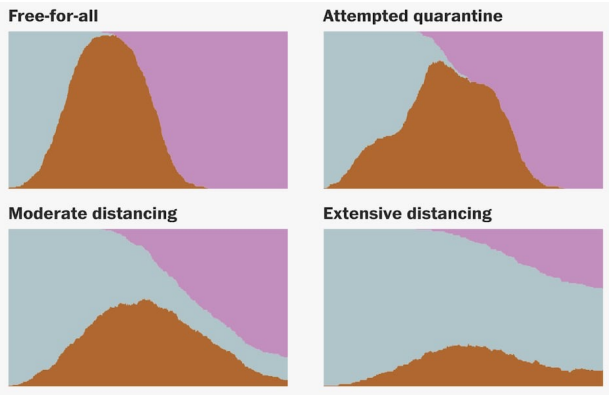




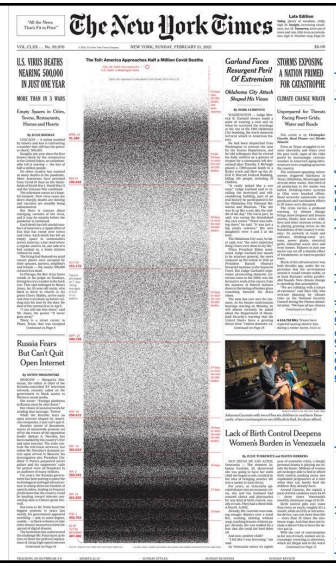
"to affect thro ' the eyes what we fail to convey to the public through their word-proof ears"

59

Inform

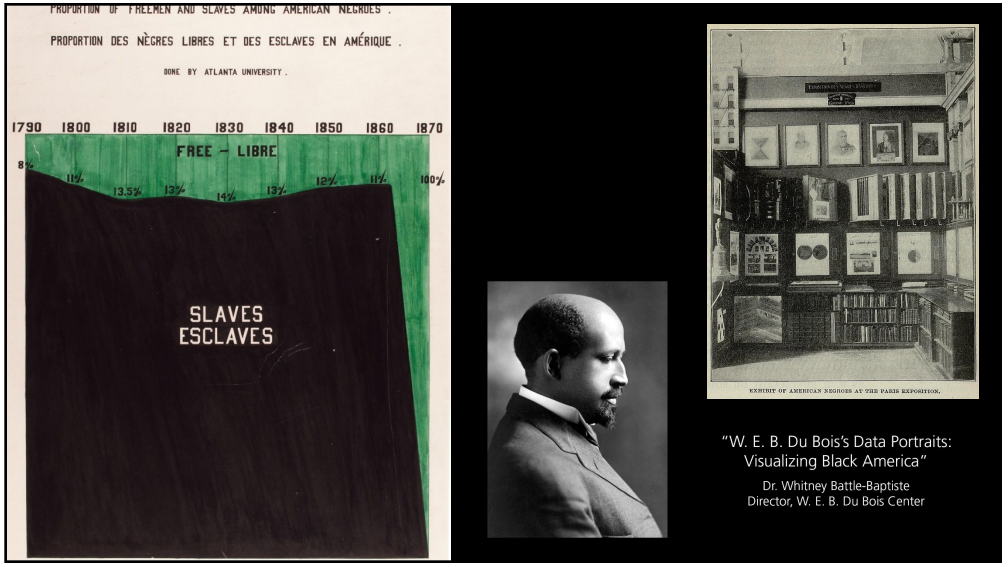


https://www.washingtonpost.com/graphics/2020/world/corona-simulator/



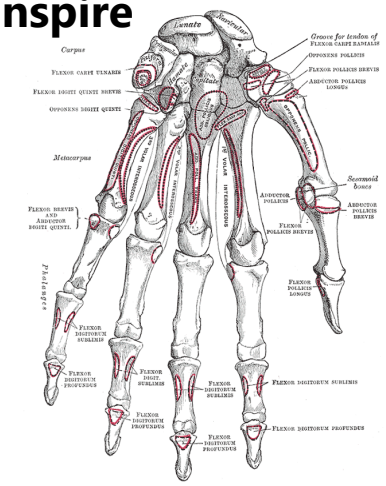
https://www.nytimes.com/2021/02/21/insider/covid-500k-front-page.html

61

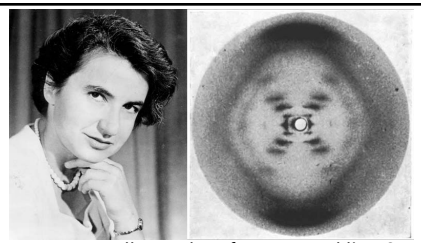


62

Inspire



Bones in hand [Gray's Anatomy 1918 ed.]



X-ray crystallography of DNA [Franklin 52]

Double helix model
[Watson and Crick 53]

63

IN THIS COURSE, …

67

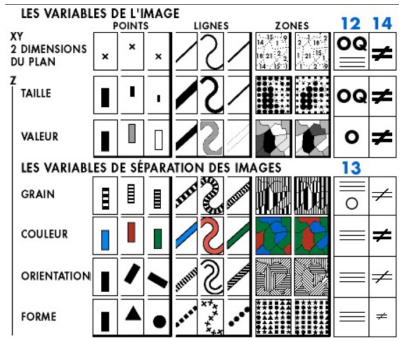
Mon: The Purpose of Visualization

Record information
- Photographs, blueprints, …

Support reasoning about information (analyze)
- Process and calculate
- Reason about data
- Expand memory

Communicate, inform, inspire (present)
- Share and persuade
- Emphasize important aspects of data

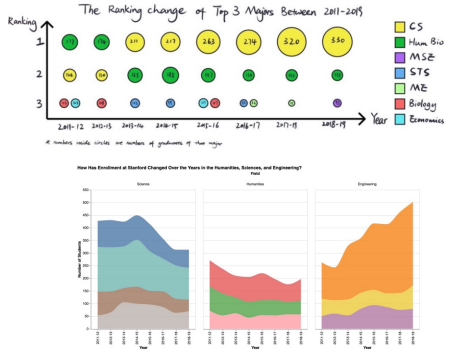
Wed: Data and Image Models



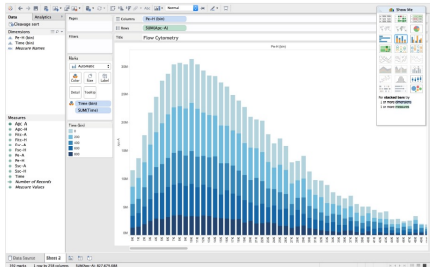
Week 1

68

Mon: Visualization Design



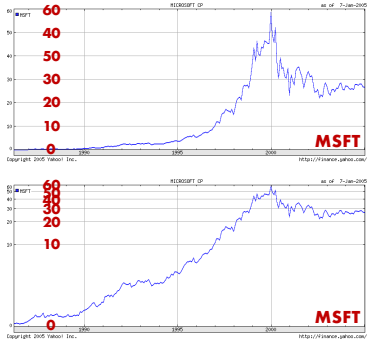
Wed: Exploratory Data Analysis



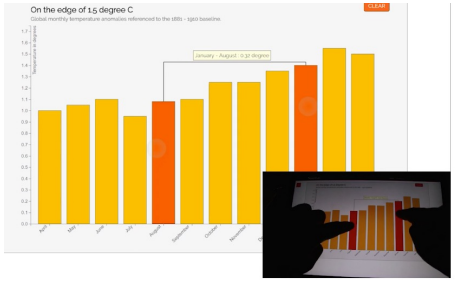
Week 2

69

Mon: Chart Design



Wed: Interaction



Week 3

70

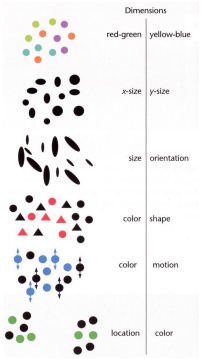
Mon: Introduction to D3 **Wed: D3 Tutorial**

Data-Driven Documents

Week 4

71

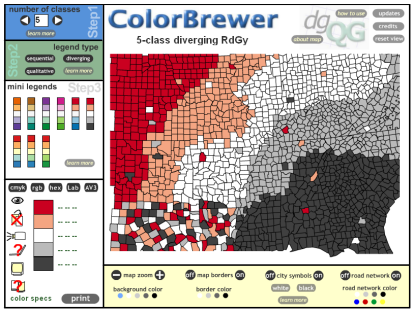
Mon: Perception **Wed: Visual Explainers**



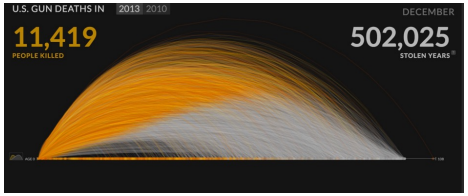
Week 5

72

Mon: Color



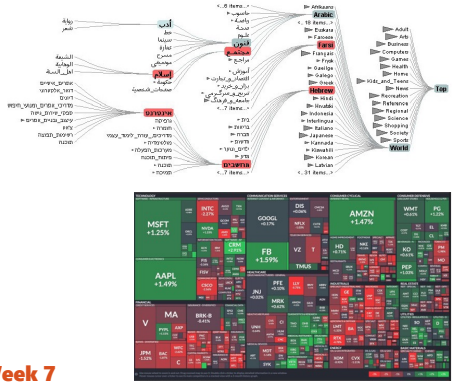
Wed: Animation



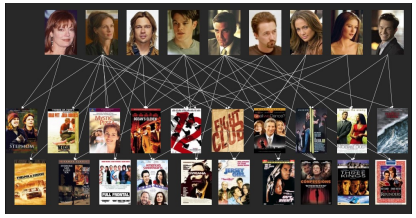
Week 6

73

Mon: Network Layout



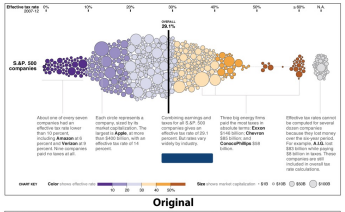
Wed: Network Analysis

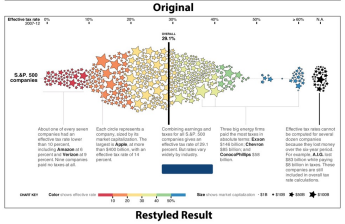


Week 7

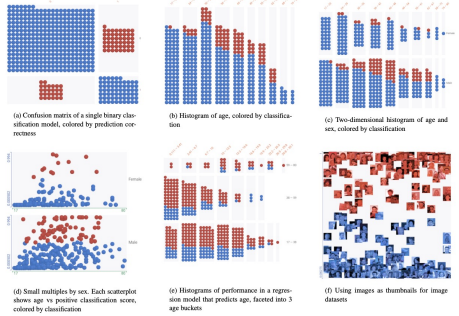
74

Mon: Deconstructing Visualizations





Wed: Visualization and AI



(a) Confusion matrix of a single binary classification model, colored by prediction correctness

(b) Histogram of age, colored by classification

(c) Two-dimensional histogram of age and sex, colored by classification

(d) Small multiples by sex. Each scatterplot shows age vs positive classification score, colored by classification

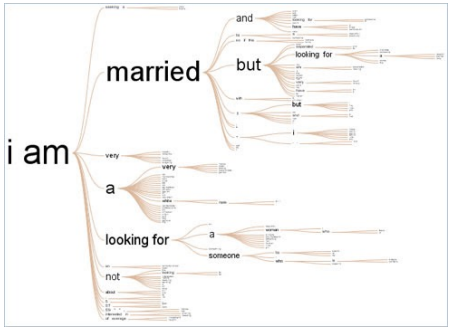
(e) Histograms of performance in a regression model that predicts age, faceted into 3 age buckets

(f) Using images as thumbnails for image datasets

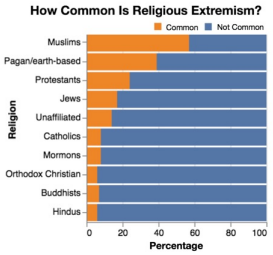
Week 8

75

Mon: Text Visualization



Wed: Visualization and NLP



Q1: *What is the percentage of response 'Common' for Catholics?*

A(Sempre): ***92***

A(Ours): ***8****. I looked up the length of the orange bar for 'Catholics'.*

Q2: *Which religion has the longest orange component?*

A(Sempre): ***Hindus***

A(Ours): ***Muslims****. I looked up 'Religion' of the longest orange bar.*

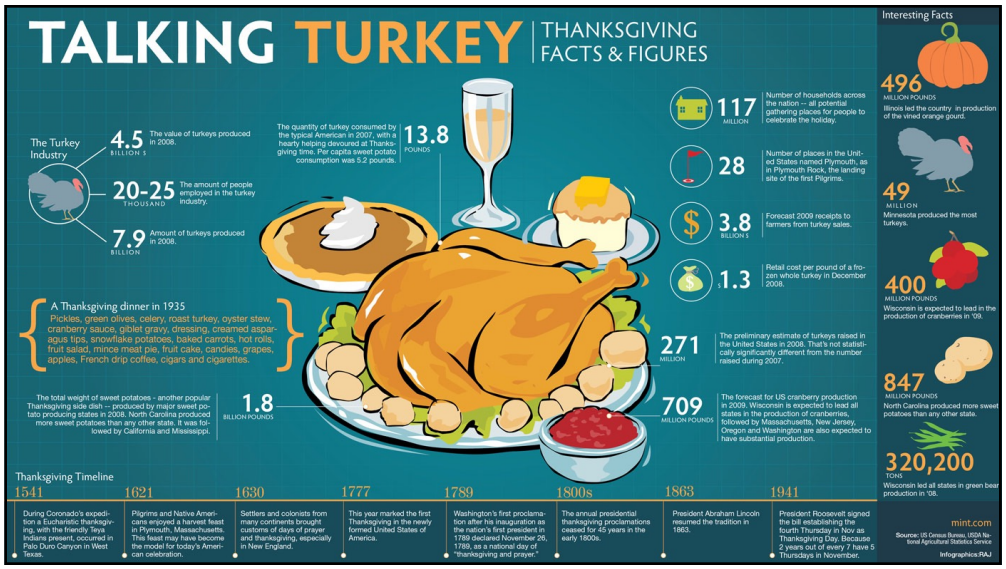
Q3: *What does the blue field represent?*

A(Sempre): ***24***

A(Ours): ***Not Common****. I looked up what blue represents by looking at the legend.*

Week 9

76



77

Final Project Review and Feedback

Week 10

78

LEARNING GOALS

- An understanding of key visualization techniques and theory, including data models, graphical perception and methods for visual encoding and interaction.
- Exposure to several common data domains and corresponding analysis tasks, including exploratory data analysis and network analysis.
- Practical experience building and evaluating visualization systems using Vega-Lite and D3.js.
- The ability to read and discuss research papers from the visualization literature.

79

YOU SHOULD EXPECT TO

- *Design, evaluate and critique* visualizations
- *Explore* data using existing visualization tools
- *Implement* interactive data visualizations
- *Develop* a substantial visualization project

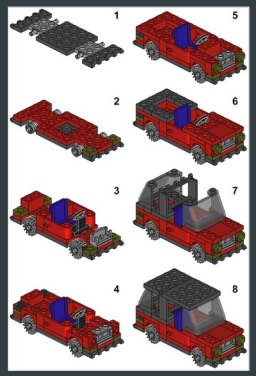
80

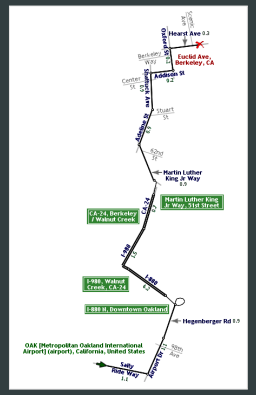
COURSE MECHANICS

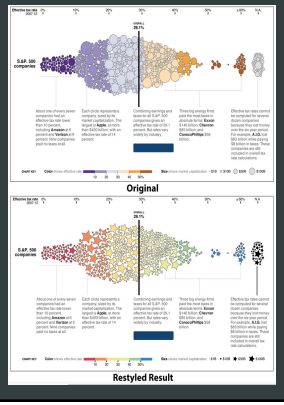
https://magrawala.github.io/cs448b-fa25/

82

Instructor: Maneesh Agrawala

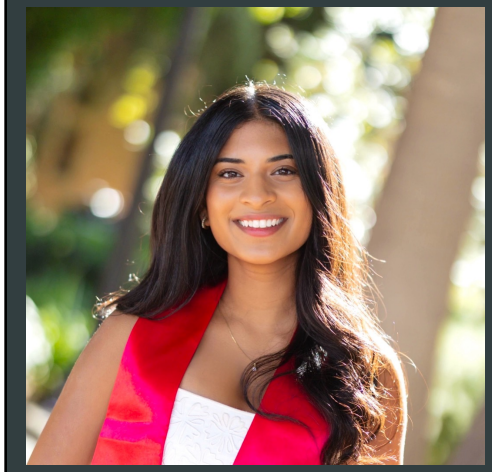






83

Course Assistant: Riya Karumanchi



Office Hours: 4:30-5:30pm Wednesdays
Location: CoDA 3rd Floor

84

Course Assistant: Shriya Reddy

Office Hours: 3:30-4:30 Mondays
Location: CoDA 3rd Floor

85

Office Hours

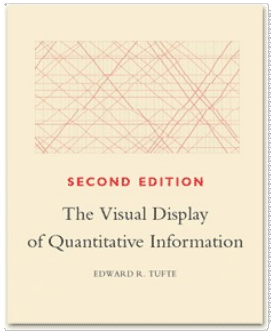
Maneesh: 2-3pm Thu, CoDA E362
Riya: 4:30-5:30pm Wed, CoDA 3rd Floor
Shriya: 3:30-4:30 Mon, CoDA 3rd Floor

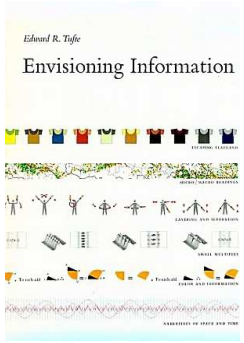
Happy to schedule other OH by appointment
Outside of OH use Slack to connect with us

https://canvas.stanford.edu/courses/215694/external_tools/11232

86

Textbooks

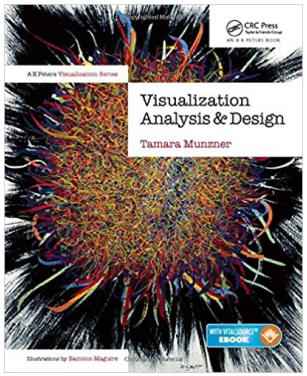




See also: www.edwardtufte.com

87

Optional Textbook



For additional theory and depth

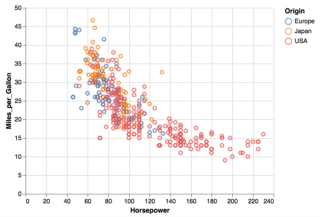
88

Interactive Notebooks

Interactivity

In addition to basic plotting and view composition, one of Vega-Lite's more exciting features is its support for interaction.

Starting with a scatter plot, we can add a basic (yet valuable!) form of interactivity – tooltips upon mouse hover – by including a `tooltip` encoding channel:




```
vl.markPoint().data(cars).encode(
  vl.x().fieldQ('Horsepower'),
  vl.y().fieldQ('Miles_per_Gallon'),
  vl.color().fieldN('Origin'),
  vl.tooltip(['Name', 'Origin']) // show the Name and Origin fields in a tooltip
).render()
```

Hands-on engagement with course concepts and modern visualization tools (Vega-Lite / D3), in JavaScript (Observable)

Extra: Fri 9/26

We will provide an intro to JavaScript/Observable on Zoom.

More details on course website.

89

Readings

From books, notebooks and linked articles

Many open to public, some may require SUNetID/Password

Material in class will be loosely based on readings

Readings should be read by start of class

Post comment (on reading, notebooks or lecture) in Canvas Discussion

One comment per week through week 9
Must post by ***end of the week*** (Sun at 8pm)
You have 1 *pass* for the quarter

Class home page
https://magrawala.github.io/cs448b-fa25/

90

Reading Responses

Good responses typically exhibit one or more

Critiques of arguments made in the papers/lectures
Analysis of implications or future directions for ideas in readings/lectures
Insightful questions about the readings/lectures

Responses should not be summaries

Should be substantive (1-2 paragraphs is typical)

91

In-Person Discussion

Discussion and critique are essential for effective design and evaluation of visualizations

- In-person discussion is more effective than online and benefits all attendees
- Attendance is **mandatory** for non-CGOE students
- Will be considered in grading for non-CGOE students

92

Assignments

Class participation (10%)

Assignment 1: Visualization Design (10%) due 9/29

Assignment 2: Exploratory Data Analysis (15%) due 10/13
Use Tableau or Vega-Lite

Assignment 3: Interactive Prototype (25%) due 10/27
Should be familiar with Javascript **(start now if you are not)**
Will cover basics of Vega-Lite and D3 in class

Final Project (40%) proposal due 11/3, design review 12/1, 12/3, final submission 12/7

93

Final project

Produce an interactive visual explainer

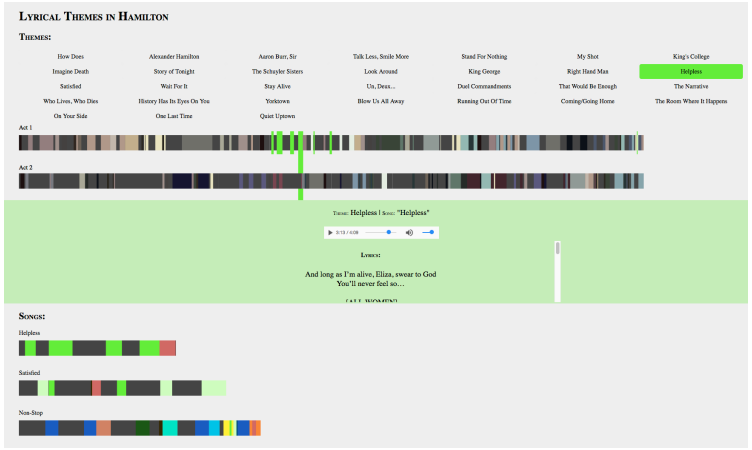
- Initial prototype and design review
- Final deliverables and video

Projects from previous classes have been

- Published as research papers
- Shared widely (e.g., gone viral on blogs)
- Released as successful open source projects

94

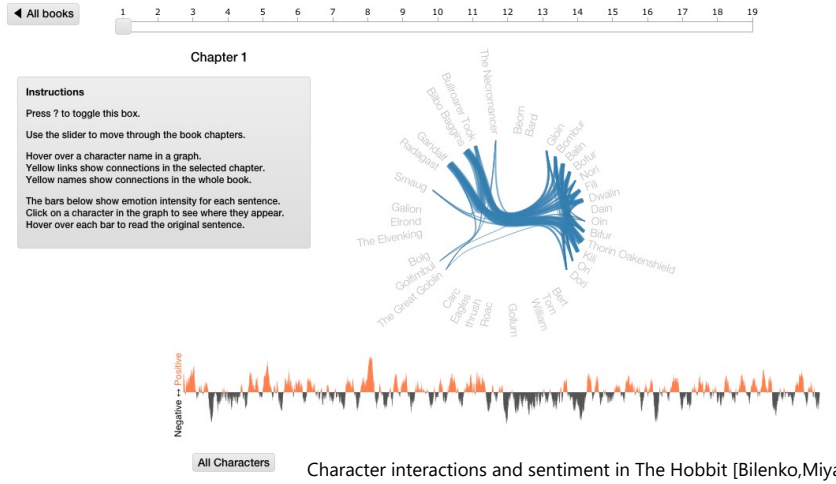
Structure of Musicals



Lyrical themes in Hamilton [Townley-Smith, Sterman, Cook 2016]

95

Visualization of Narrative Structure



Character interactions and sentiment in The Hobbit [Bilenko,Miyakawa 2013]

96

ASSIGNMENT 1: VISUALIZATION DESIGN

Due 9/29 6 AM

Design a static visualization for a data set

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

Data: Halloween Candy Class Ranking

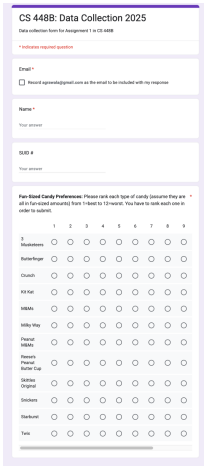
As part of the first lecture in this class you have been asked to fill out a form asking you to rank 12 types of candy commonly passed out on Halloween (fun-sized portions) from 1=best to 12=worst. We have aggregated and wrangled the data to produce a data table with the following information.

Number of records: 12
Data fields:

- **candy:** Name of candy.
- **chocolate:** Does it contain chocolate?
- **fruity:** Is it fruit flavored?
- **caramel:** Is there caramel in the candy?
- **peanutyalmondy:** Does it contain peanuts, peanut butter or almonds?
- **nougat:** Does it contain nougat?
- **crispedricewafer:** Does it contain crisped rice, wafers or a cookie component?
- **hard:** Is it hard candy?
- **bar:** Is it a candy bar?
- **pluribus:** Is it one of many candies in a bag or box?
- **sugarpercent:** The percentile of sugar it fall under within a larger data set of 85 candies.
- **pricepercent:** The unit price percentile compared with a larger data set of 85 candies.
- **classwinpercent:** The win percentage based on all the pairwise ranking match-ups in our class.

The data will be available in csv format (after we process and wrangle it) at: classHalloweenCandy2025.csv.

98


CS 448B: Data Collection 2025

Data collection form for Assignment 1 in CS 448B

* Indicates required question

Email *

Record agoneata@gmail.com as the email to be included with my response

Name *

Your answer

SUID #

Your answer

Fun-Sized Candy Preferences: Please rank each type of candy (assume they are all in fun-sized amounts) from 1=best to 12=worst. You have to rank each one in order to submit. *

	1	2	3	4	5	6	7	8	9
3 Musketeers	○	○	○	○	○	○	○	○	○
Butterfinger	○	○	○	○	○	○	○	○	○
Crunch	○	○	○	○	○	○	○	○	○
Kit Kat	○	○	○	○	○	○	○	○	○
MilkMs	○	○	○	○	○	○	○	○	○
Milky Way	○	○	○	○	○	○	○	○	○
Peanut MilkMs	○	○	○	○	○	○	○	○	○
Reese's Peanut Butter Cup	○	○	○	○	○	○	○	○	○
Skittles Original	○	○	○	○	○	○	○	○	○
Snickers	○	○	○	○	○	○	○	○	○
Starburst	○	○	○	○	○	○	○	○	○
Twix	○	○	○	○	○	○	○	○	○

https://forms.gle/Sd7ngHrSuCxLDb6C6

99

Assignment 1: Visualization Design

Due: Monday Sep 29, 2025 by 6am

In this assignment, you will design a visualization for a small data set and provide rigorous rationale for your design choices. You should in theory be ready to explain the contribution of every pixel in the display. You are free to use any graphics or charting tool you please–including drafting it by hand.

Data: Halloween Candy Class Ranking

As part of the first lecture in this class you have been asked to fill out a form asking you to rank 12 types of candy commonly passed out on Halloween (fun-sized portions) from 1=best to 12=worst. We have aggregated and wrangled the data to produce a data table with the following information.

Number of records: 12
Data fields:
- **candy:** Name of candy.
- **chocolate:** Does it contain chocolate?
- **fruity:** Is it fruit flavored?
- **caramel:** Is there caramel in the candy?
- **peanutyalmondy:** Does it contain peanuts, peanut butter or almonds?
- **nougat:** Does it contain nougat?
- **crispedricewafer:** Does it contain crisped rice, wafers or a cookie component?
- **hard:** Is it hard candy?
- **bar:** Is it a candy bar?
- **pluribus:** Is it one of many candies in a bag or box?
- **sugarpercent:** The percentile of sugar it fall under within a larger data set of 85 candies.
- **pricepercent:** The unit price percentile compared with a larger data set of 85 candies.
- **classwinpercent:** The win percentage based on all the pairwise ranking match-ups in our class.

The data will be available in csv format (after we process and wrangle it) at: classHalloweenCandy2025.csv.

Note that all the fields other than **classwinpercent** are from a subset of Walt Hickey's Halloween Candy Article as available from Kaggle.

Assignment

Your task is to download this data and design a static (i.e., single image) visualization that you believe effectively communicates one aspect of the data and provide a short write-up (no more than 4 paragraphs) describing your design choices. Start by choosing a question you'd like your visualization to answer. **Design your visualization to answer that question, and use the question as the title of your graphic.**

While you must use the data set given, you are free to filter, transform and augment the data as you see fit. Such transformations may include (but are not limited to) log transformation, computing percentages or averages, grouping elements into new categories, and/or removing data that are not relevant to your driving question. You are also free to incorporate external data as you see fit. Your chart should be interpretable without recourse to your short write-up. Do not forget to include title, axis labels or legends as needed! *Hint: Good design often requires omitting data when it is irrelevant to the question your visualization is designed to answer.*

As different visualizations can emphasize different aspects of a data set, your write-up should document what aspects of the data you are attempting to most effectively communicate. In short, what story are you trying to tell? Just as important, also note which aspects of the data might be obscured due to your visualization design.

100