ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

Lecture 4 and 5
Data Visualization 1/3

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)
AI-Experience-Lab (reflect9.github.io/ael)

Things to Learn

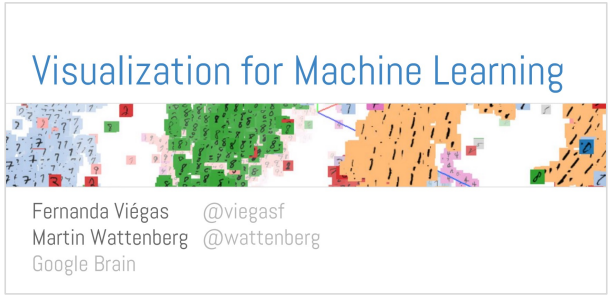
(Data Visualization 1)

1. What is Data Visualization
2. Historical and Modern Data Visualization
3. Principles and Theories

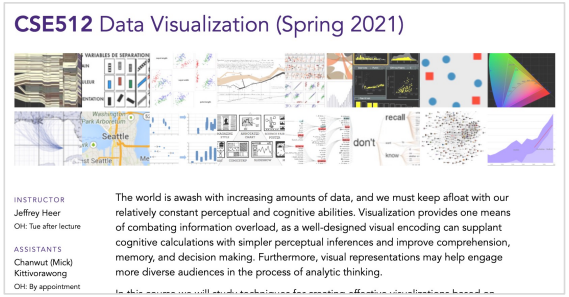
Week 2

Acknowledgements

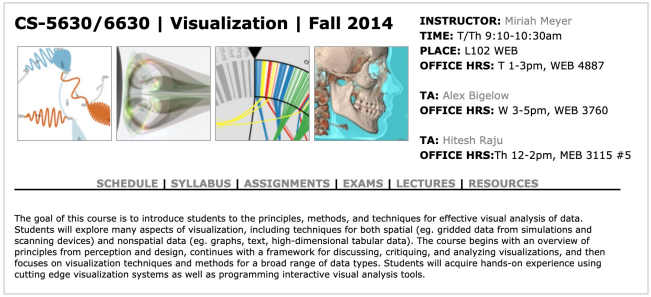
Content and structure of this slide deck is inspired by (or copied-and-pasted from) the following lectures



Fernanda Viégas and Martin Wattenberg are pioneers of data visualization. While this talk focuses on ML researchers, it covers the introductory concepts of data visualization.
https://media.neurips.cc/Conferences/NIPS2018/Slides/Visualization_for_ML.pdf



Jeffrey Heer is best known for his work on information visualization and interactive data analysis. At University of Washington, he's been teaching data visualization over a decade.
https://courses.cs.washington.edu/courses/cse512/21sp/



Miriah Meyer's *Visualization* classes provide comprehensive summary of data visualization
https://miriah.github.io/teaching/cs6630/

What is Data Visualization

How many ‘A’ are in the text below?

MTHIVLWYADCEQGHKILKASKDFAKMSD
AKSDFDKLFKLDKLFAKSLDFKLASKLDFK
KSDFLSKLDFKDKLFKLKLFKLSDKFERT
DKLFASDFKLSLDKFKSDKLFKLASDKLF

How many ‘A’ are in the text below?

MTHIVLWYADCEQGHKILKASKDFAKMSD
AKSDFDKLFKLDKLFAKSLDFKLASKLDFK
KSDFLSKLDFKDKLFKLKLFKLSDKFERT
DKLFASDFKLSLDKFKSDKLFKLASDKLF

Can you see the difference between sets?

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

$u_X = 9.0$ $\sigma_X = 3.32$

$u_Y = 7.5$ $\sigma_Y = 2.03$

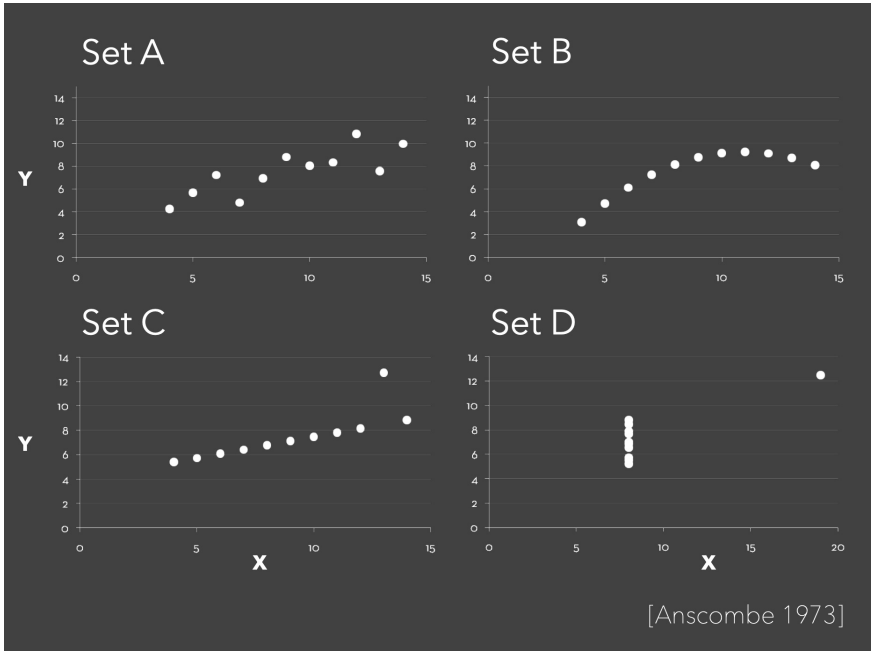
Linear Regression

$Y = 3 + 0.5\,X$

$R^2 = 0.67$

[Anscombe 1973]

All the sets have the same summary statistics



But they look different when visualized

Data Visualization is…

“Transforms data into visual marks” [Viégas and Wattenberg 2018]

“Transformation of the symbolic into the geometric” [McCormic 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, and Shneiderman 1999]

Four Goals of Datavis

- **Record** information
- **Analyze** data to support reasoning
- **Confirm** hypotheses
- **Communicate** ideas to others

Symbolic Data

Cognition

Working Memory

Visual Marks

Perception

Visual Bandwidth

DATA

Exploratory Data Analytics **(EDA)**

INSIGHTS & VISUALIZATIONS

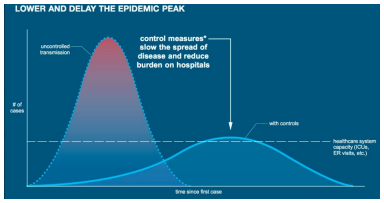
Design Storytelling Communication

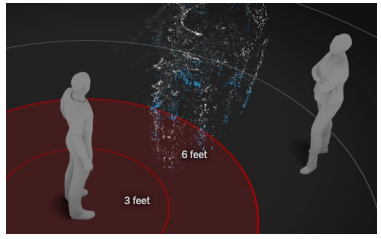
Infographics

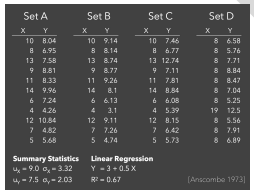
Dashboards

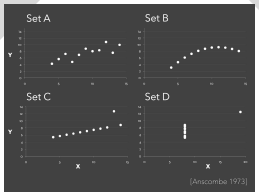
Products/Services

Data Videos







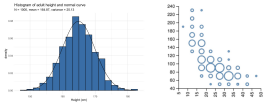


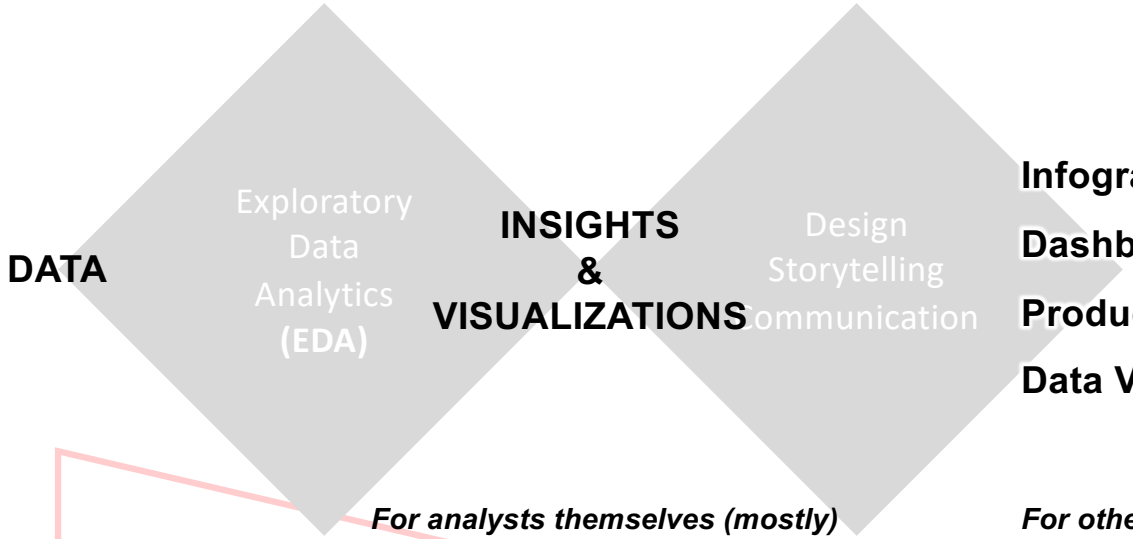
Order ID	Order Date	Order Priority
3	10/14/06	5-Low
6	2/21/08	4-Not Specified
32	7/16/07	2-High
32	7/16/07	2-High
32	7/16/07	2-High
32	7/16/07	2-High
35	10/23/07	4-Not Specified
35	10/23/07	4-Not Specified

Data Cleaning and Transformation;

Making and Testing hypotheses;

Trial-and-errors of drawing charts





DIKW

For analysts themselves (mostly)

- See data in context

- Discover insights

- Make decisions

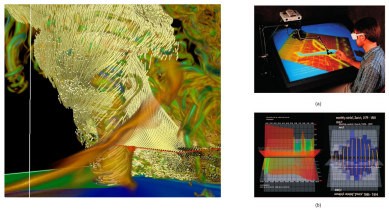
For other people

- *Guide the viewer's attention*
- *Communicate data and insights to the viewer*
- *Enable the viewer to get their own insights*

Data Visualization and its siblings

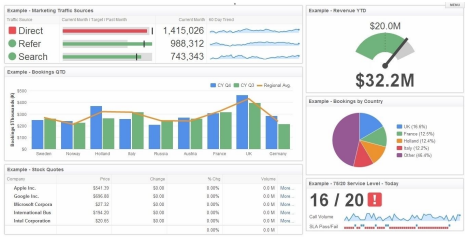
Scientific Visualization

- Visualizing continuous data on in a multi-dimensional space
- There is an inherent spatial position (e.g. 3D position)
- Relationship between variables are well understood
- Studied in the computer graphics domain



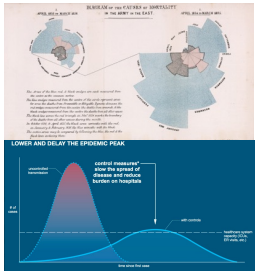
Information or Data Visualization

- Visualizing discrete and abstract variables
- There is no inherent way to position entities
- Relationship between variables are not well understood
- Studied in the Human-Computer Interaction domain



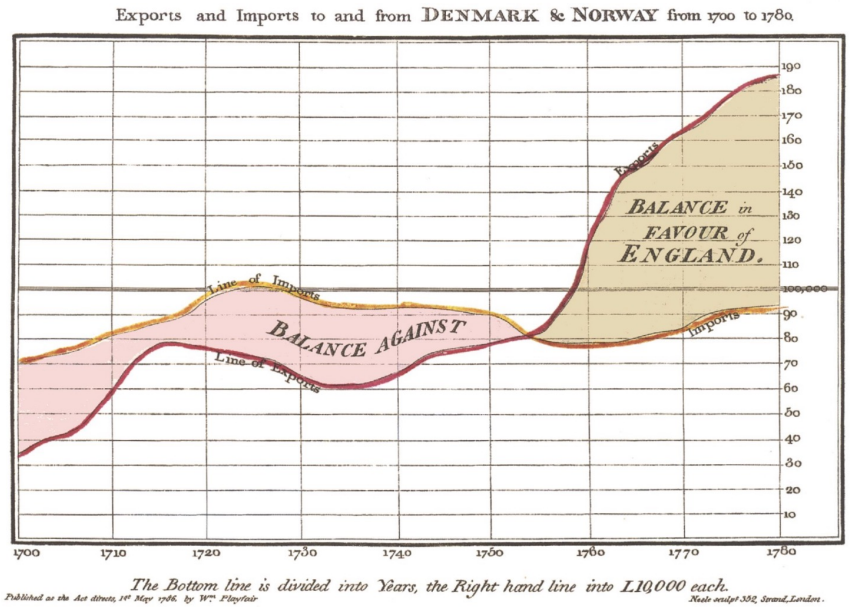
Infographics

- Communicating messages for a specific purposes (e.g. persuasion, education)
- Focusing on effective visual communication
- Studied in the visual communication and graphic design domains



Long long time ago…

William Playfair (1786)

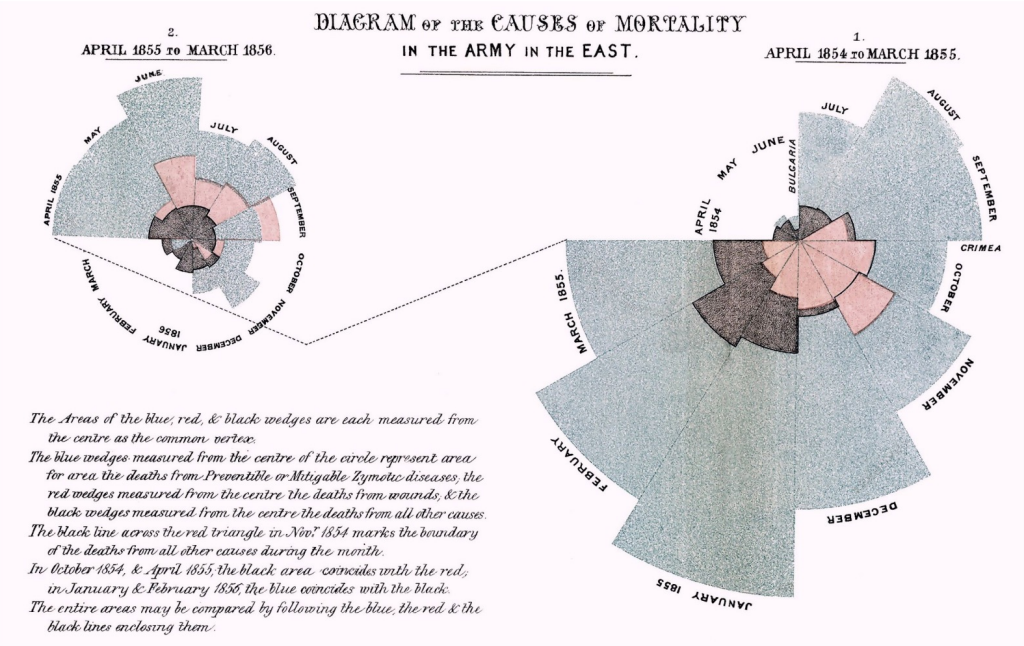


Line, bar, pie charts were all invented by the same person!

Aside from revolutionizing graphics, Playfair was an economist, engineer, and even a secret agent.

(Image: Wikipedia)

Florence Nightingale (1858)

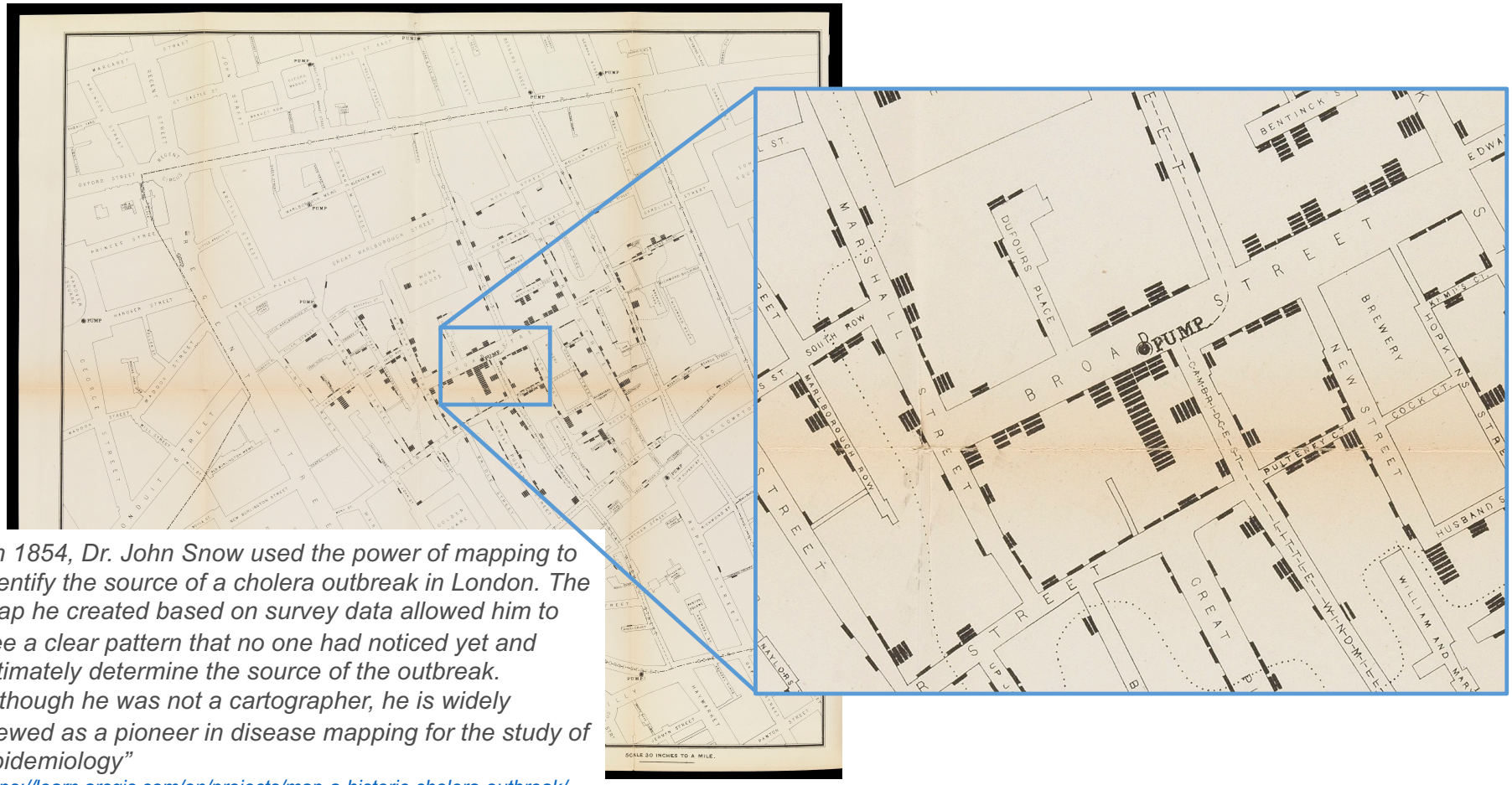


These charts led to the adoption of better hygiene / sanitary practices in military medicine, saving millions of lives.

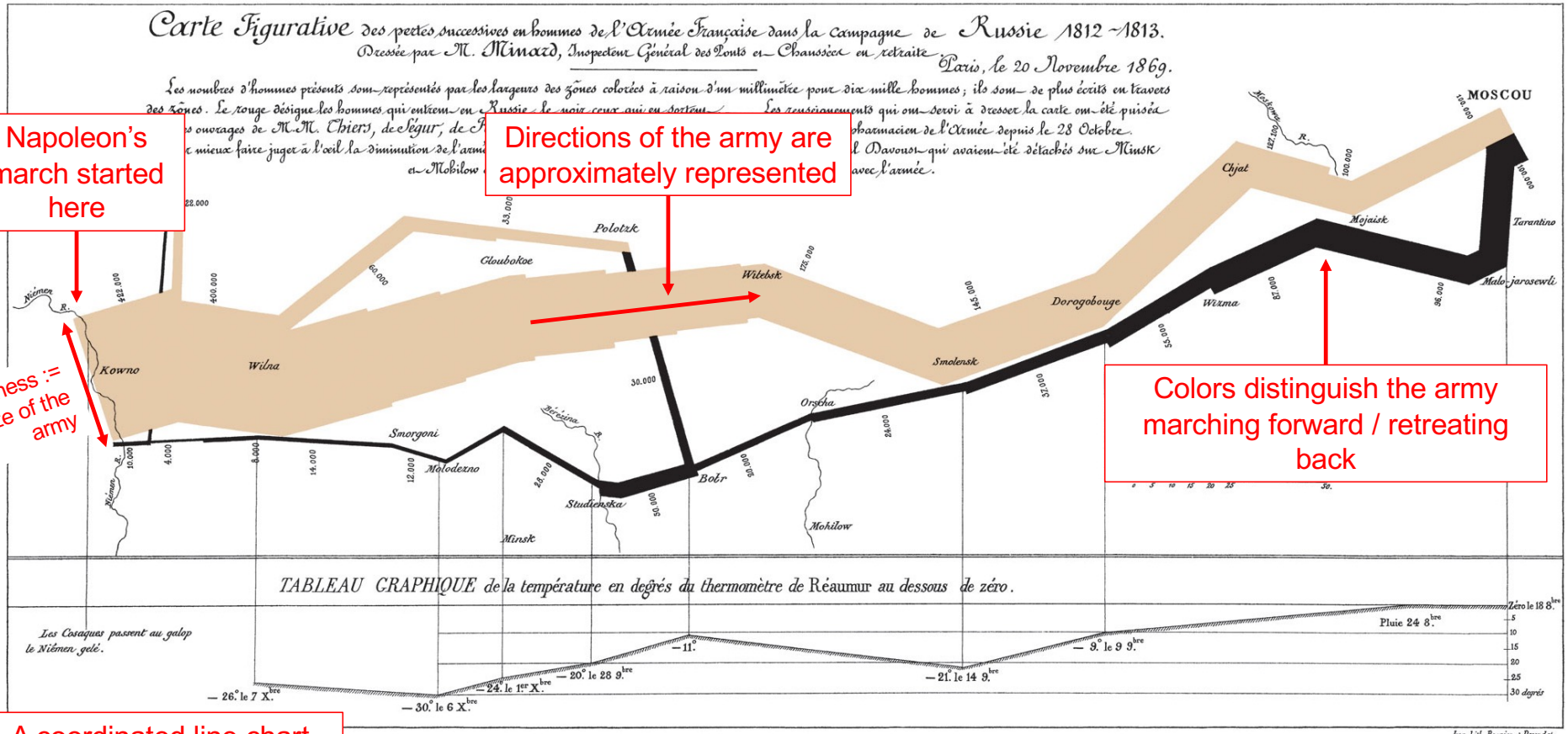
Arguably the most effective visualization ever!

This particular visualization technique would be frowned on today. Lesson: technique is less important than having the right data and right message.

(Image: Wikipedia)



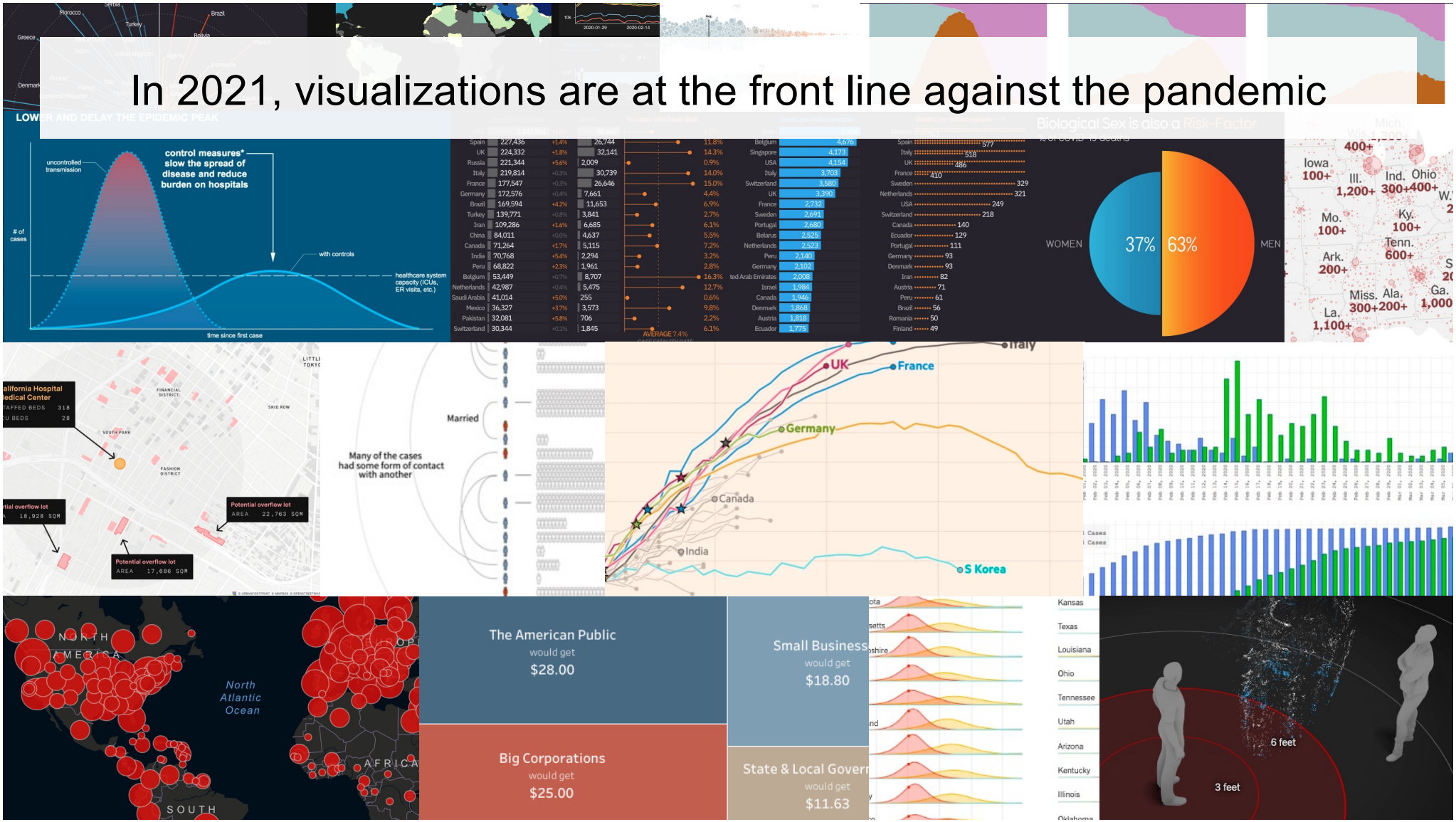
"In 1854, Dr. John Snow used the power of mapping to identify the source of a cholera outbreak in London. The map he created based on survey data allowed him to see a clear pattern that no one had noticed yet and ultimately determine the source of the outbreak. Although he was not a cartographer, he is widely viewed as a pioneer in disease mapping for the study of epidemiology"
https://learn.arcgis.com/en/projects/map-a-historic-cholera-outbreak/



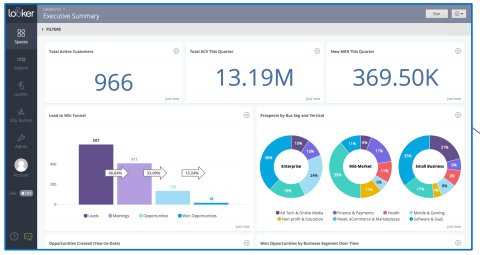
A coordinated line chart shows how low the temperature while retreating

Charles Minard's Flow Map of Napoleon's Russian Campaign of 1812

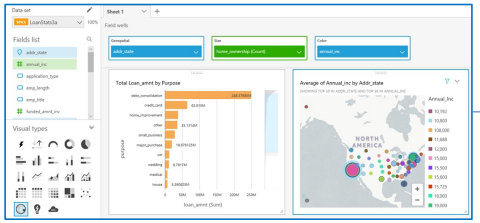
In 2021, visualizations are at the front line against the pandemic



Data Visualization is the core of Business Intelligence

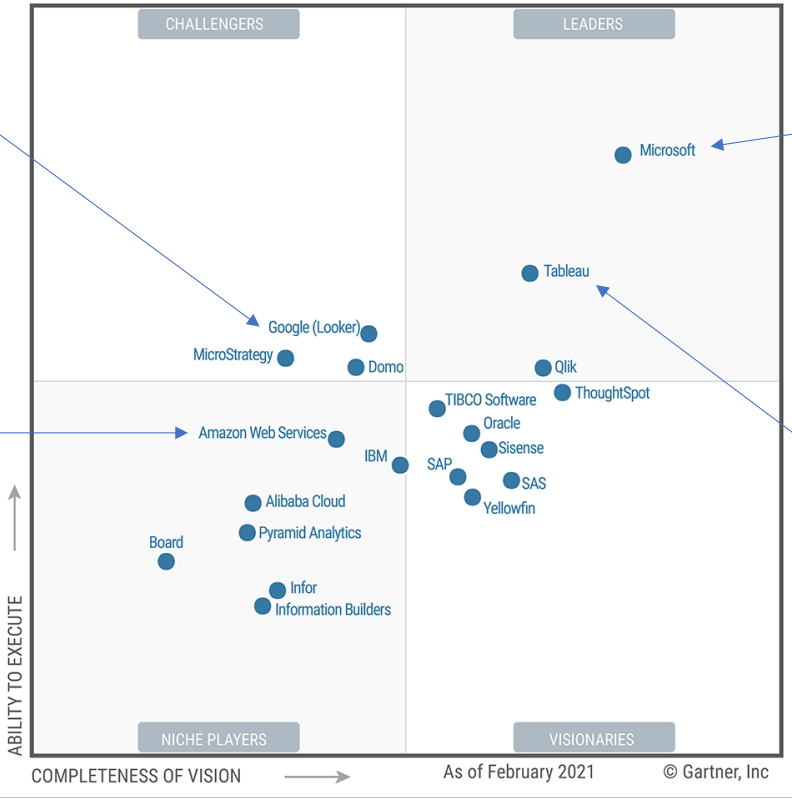


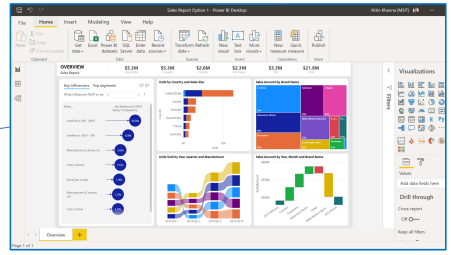
Google Looker



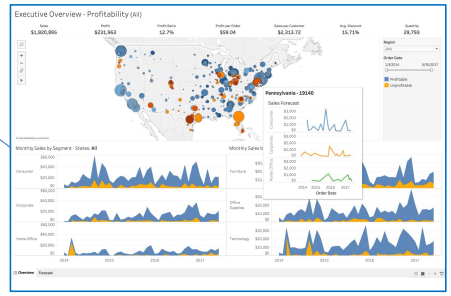
AWS Quicksight

Gartner's Magic Quadrants of Analytics and Business Intelligence



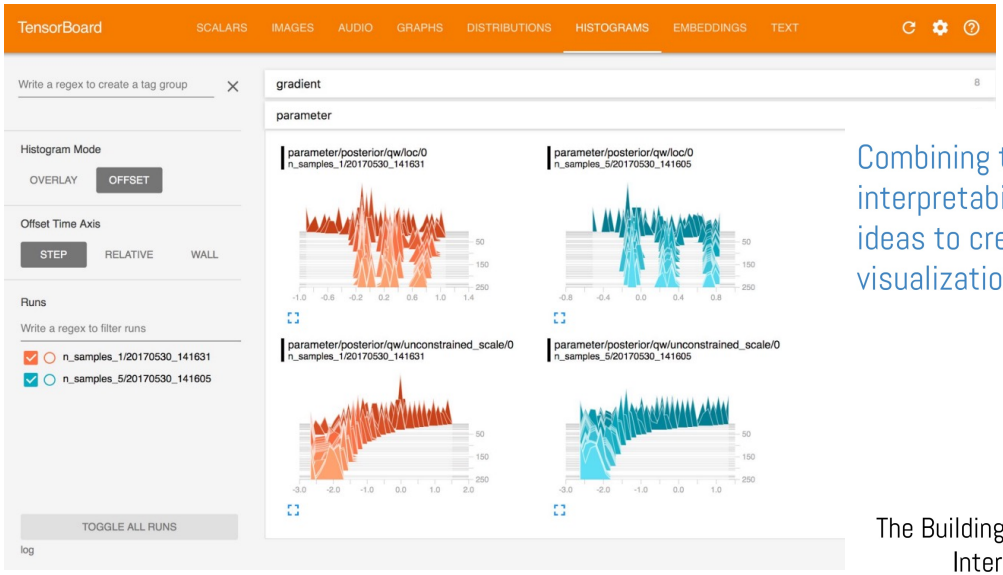


Microsoft Power BI

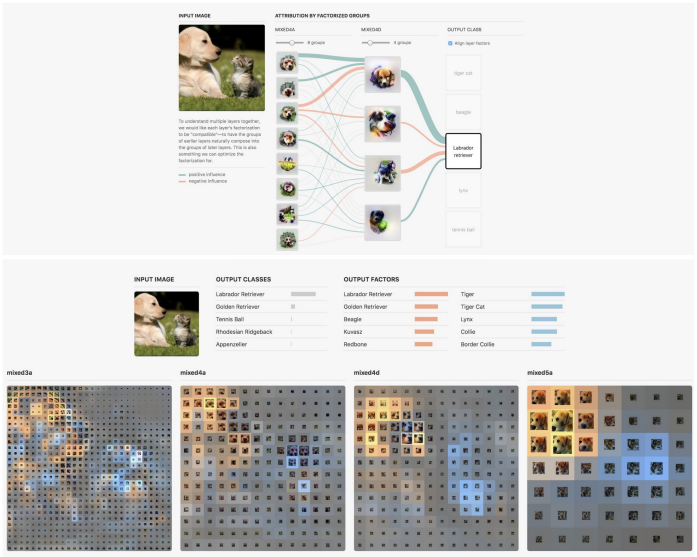


Tableau

Data Visualizations Make AI Interpretable



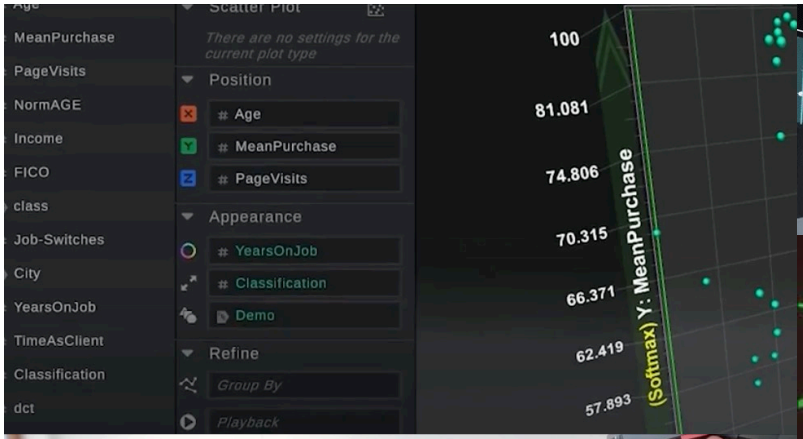
Combining these interpretability ideas to create new visualizations

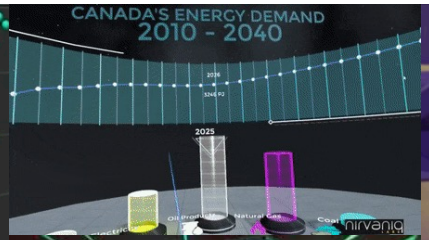


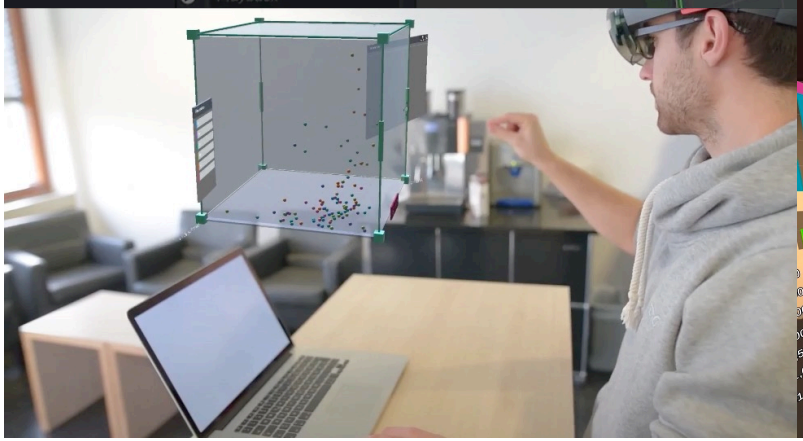
The Building Blocks of Interpretability

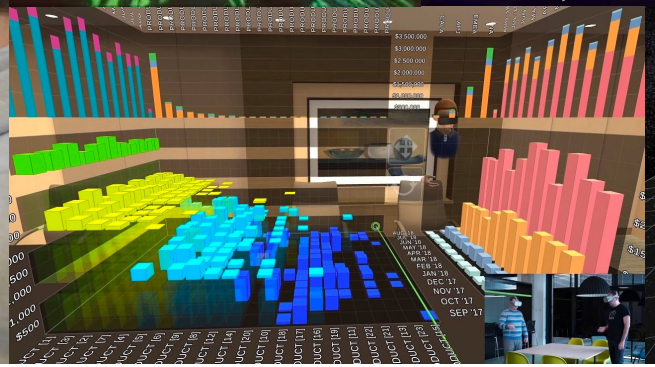
Olah, Satyanarayan, Johnson, Carter, Schubert, Ye, Mordvintsev

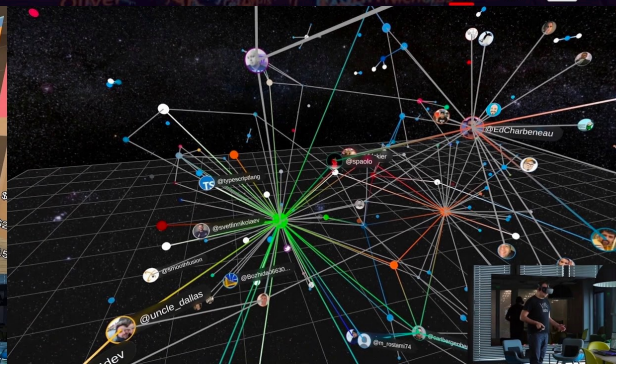
Diverse Channels > Immersive Visual Analytics











Diverse Channels > Immersive Data Storytelling

We are scratching the surface of datavis

For three weeks we will focus

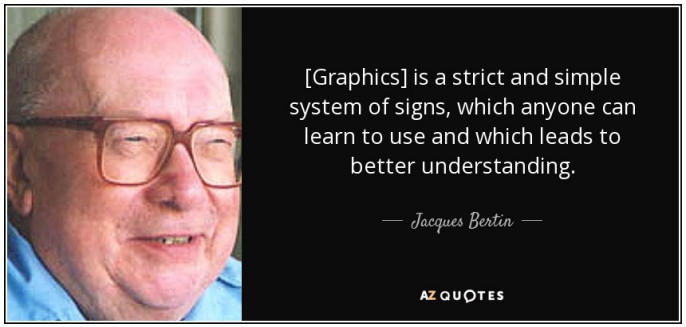
1. Basic principles and theories of visualization
2. Hands-on practices of common datavis toolkits
3. Effectively communicating with others via simple reports and dashboards

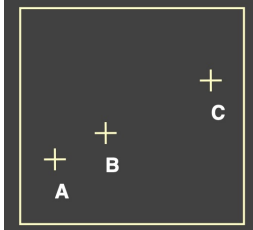
However, we are NOT covering

- How to design and develop infographics or data stories
- Designing consumer products or services embedding data visualization
- Building interactive tools for data analysis

Principles and Theories

Visual Language is a Sign System

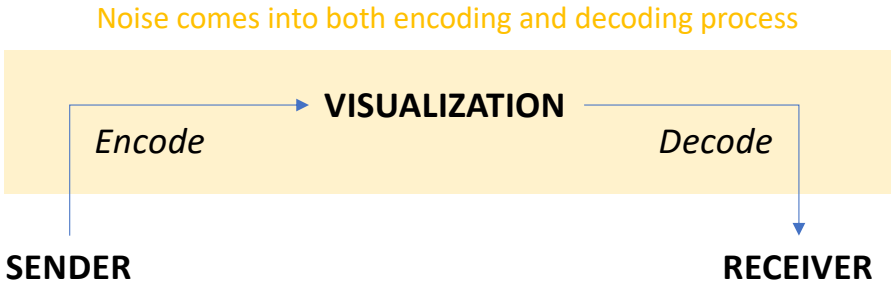


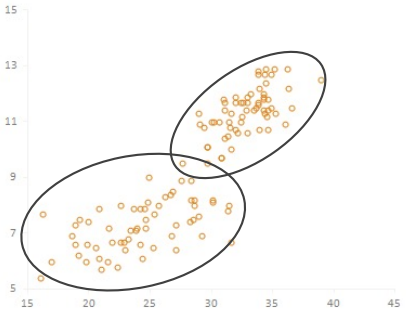


1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables

"Resemblance, order and proportional are the three signfields in graphics." - Bertin



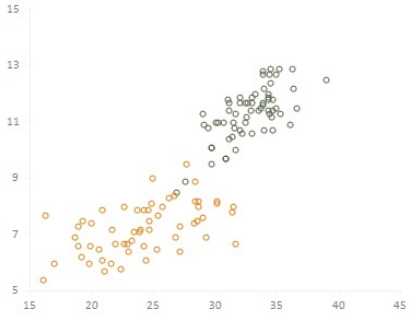


PROXIMITY

When objects placed together, the eye perceives them as a group.

SIMILARITY

When objects look similar to one another, the eye perceives them as a group or pattern.

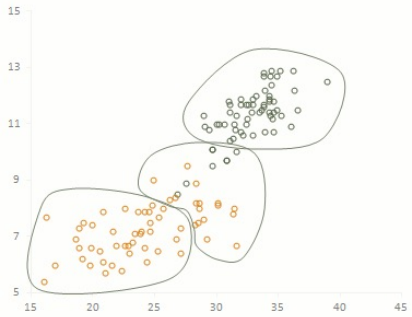


CONTINUANCE

The eye is compelled to move from one object through another.

CLOSURE

When an object is incomplete or not completely enclosed.



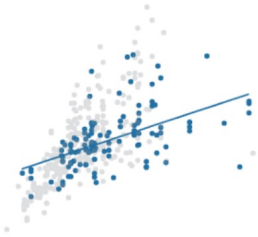
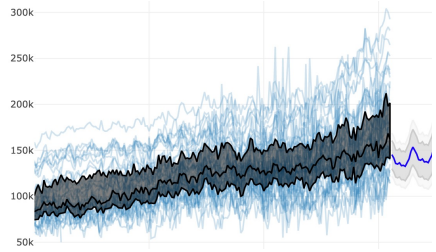


FIGURE & GROUND

When the eye differentiates an object from its surrounding area.

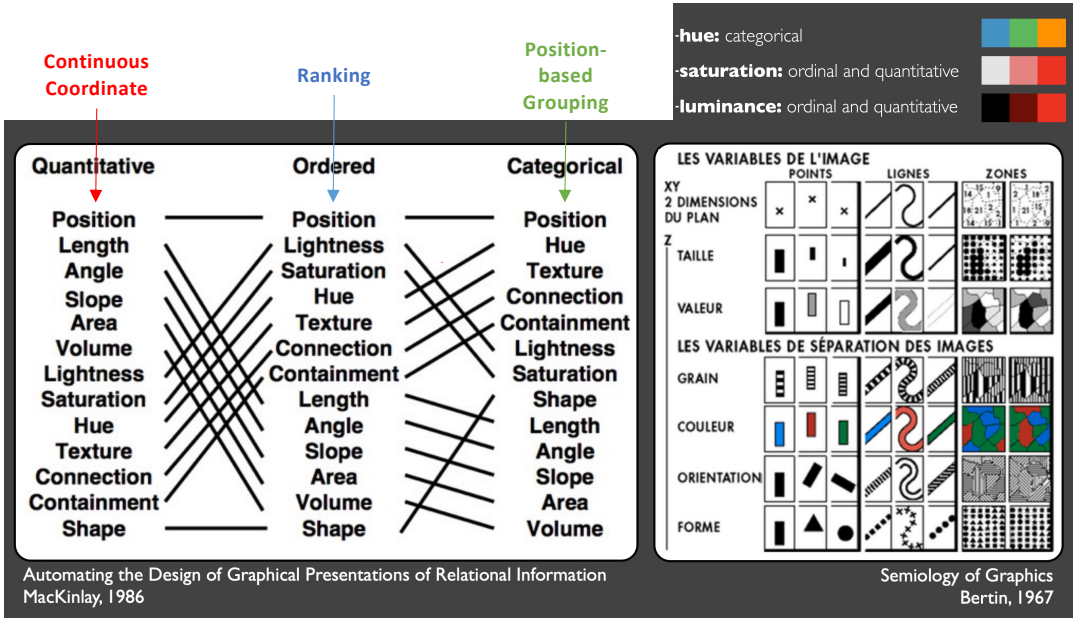


https://design4users.com/gestalt-theory-for-ux-designers-principle-of-similarity/

Cheat Sheet: Data Type → Visual Encoding

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

quantitative
ordinal
categorical



- **Position** (scatterplot) is the best for any level of measurement.
- **Length** (bar) makes viewers assume the axis is quantitative
- **Lightness** and **Saturation** guides viewer's attention order
- **Hue** is the best way to distinguish different categories (better than **shape**)

Cheat Sheet: Data Type → Visual Encoding

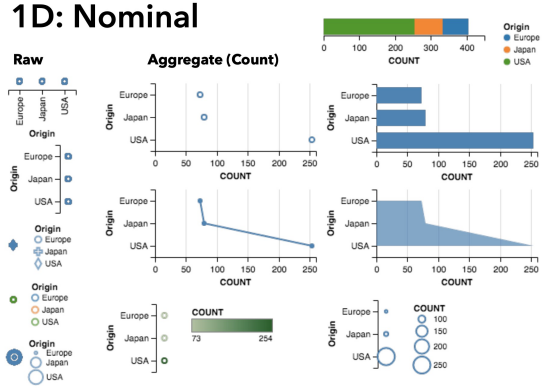
Mapping Data to Visual Variables

Assign **data fields** (e.g., with *N, O, Q* types) to **visual channels** (*x, y, color, shape, size, …*) for a chosen **graphical mark** type (*point, bar, line, …*).

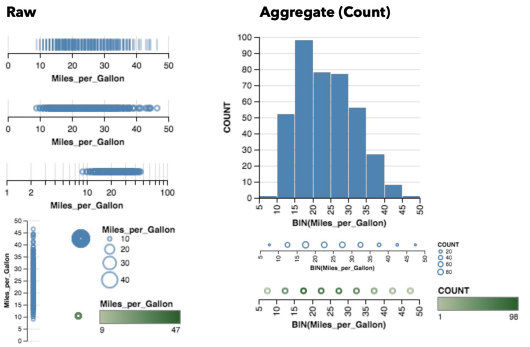
Additional concerns include choosing appropriate **encoding parameters** (*log scale, sorting, …*) and **data transformations** (*bin, group, aggregate, …*).

These options define a large combinatorial space, containing both useful and questionable charts!

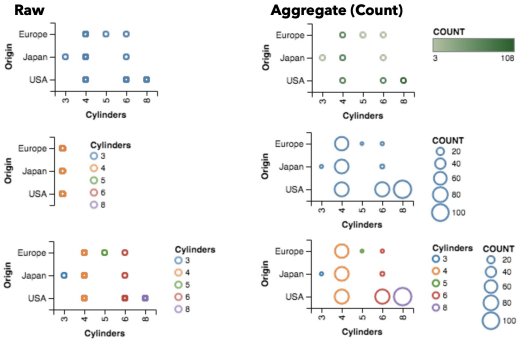
1D: Nominal



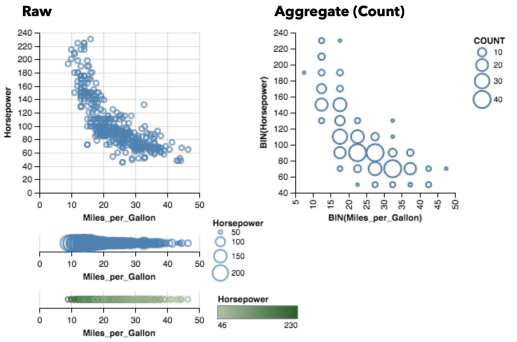
1D: Quantitative



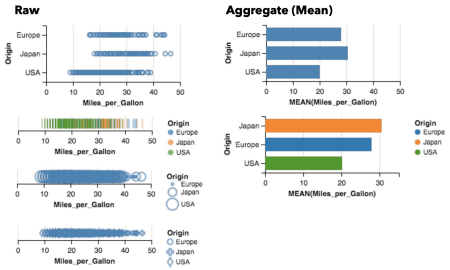
2D: Nominal x Nominal



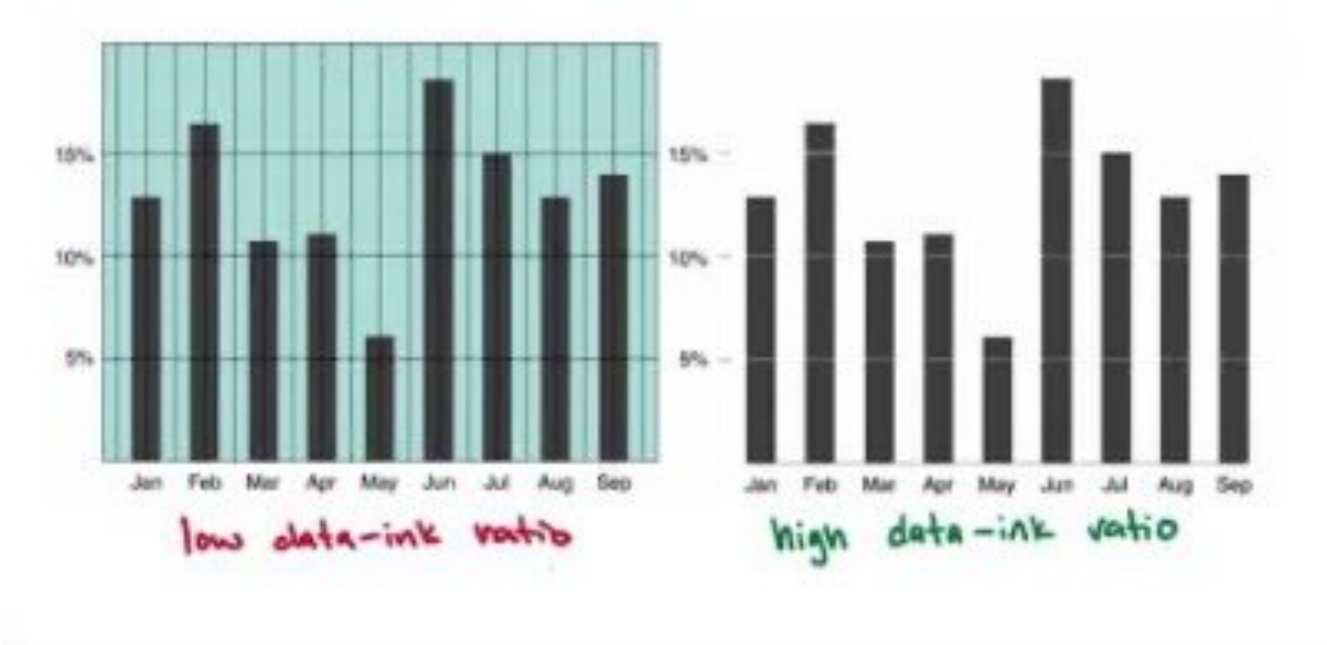
2D: Quantitative x Quantitative



2D: Nominal x Quantitative

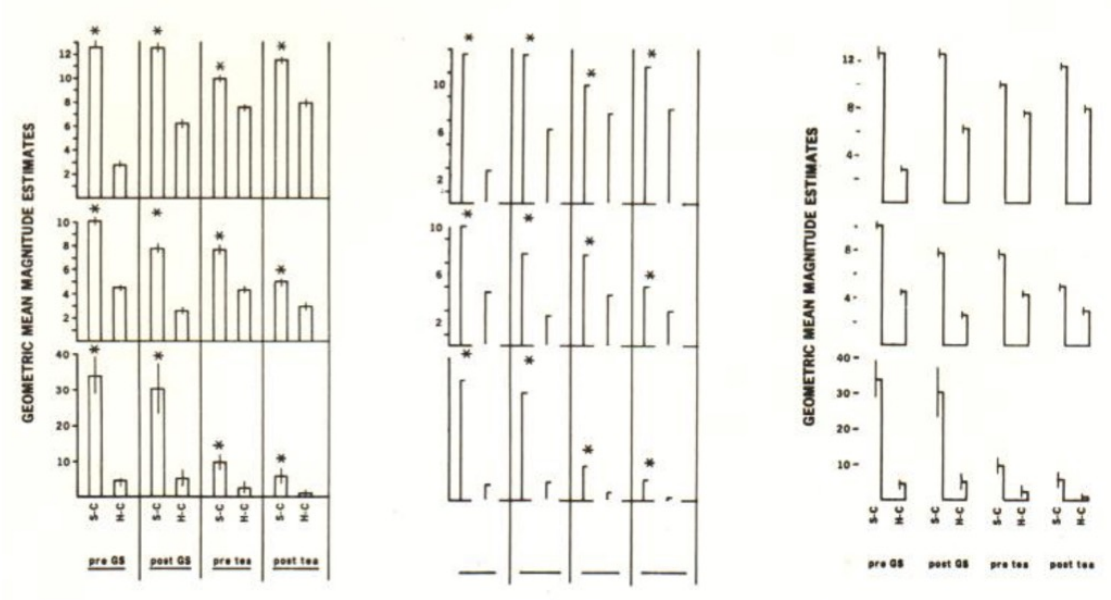


Edward Tufte > Maximize the data-ink ratio



https://www.youtube.com/watch?v=JIMUzJzqaA8

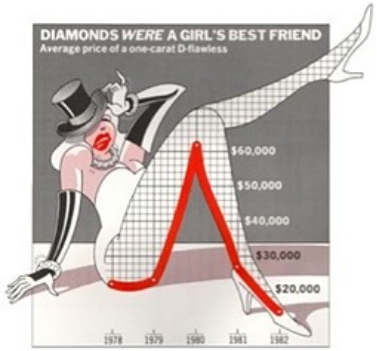
Edward Tufte > Maximize the data-ink ratio

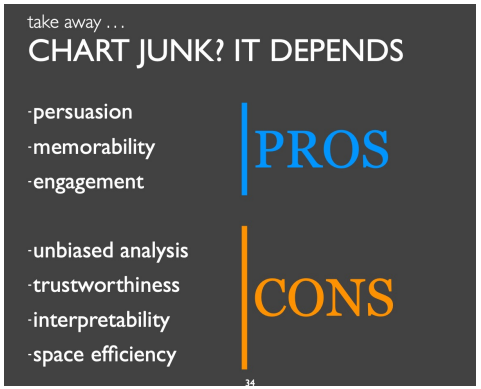


Note. There are justified criticisms on charts with extremely high data-ink ratios.
- Charts of extremely high data-ink ratios are often difficult to understand without training
- People tend to say “more the better” until they get tired of being cognitively overloaded
- My heuristics are (1) consider the viewer’s persona (2) don’t go too extreme (3) use subtle colors when you are unsure about complete removal

Edward Tufte > Avoid Chart Junk

"Nearly all those who produce graphics for mass publication are trained exclusively in the fine arts and have had little experience with the analysis of data. Such experiences are essential for achieving precision and grace in the presence of statistics... Those who get ahead are those who ***beautified data****, never mind* ***statistical integrity****."* Tufte, Edward R. (1983). The Visual Display of Quantitative Information





From Jeffery Heer's Lecture Note

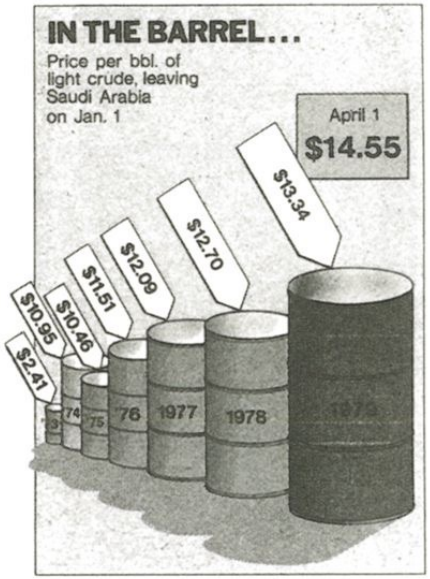
- *Data-ink ratio* and *Chart Junk* are largely overlapping concepts, since most junk charts waste lots of ink (i.e. viewer's cognitive workload)
- There are criticisms on this principle, "*chart junk increases long-term memorability of the chart*" (Christopher et al. 2010)
- Edward Tufte is a rational idealist who loved minimalism and statistical integrity (i.e. "Never lie with visualizations"). His principles are noteworthy, but following them won't guarantee success in real world.

Edward Tufte > Integrity

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity."

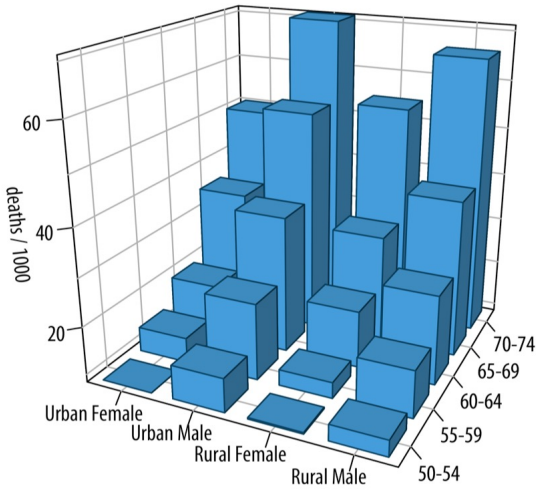
"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented."

"Show data variation, not design variation."



Lie Factor 9.5... by volume 59.4

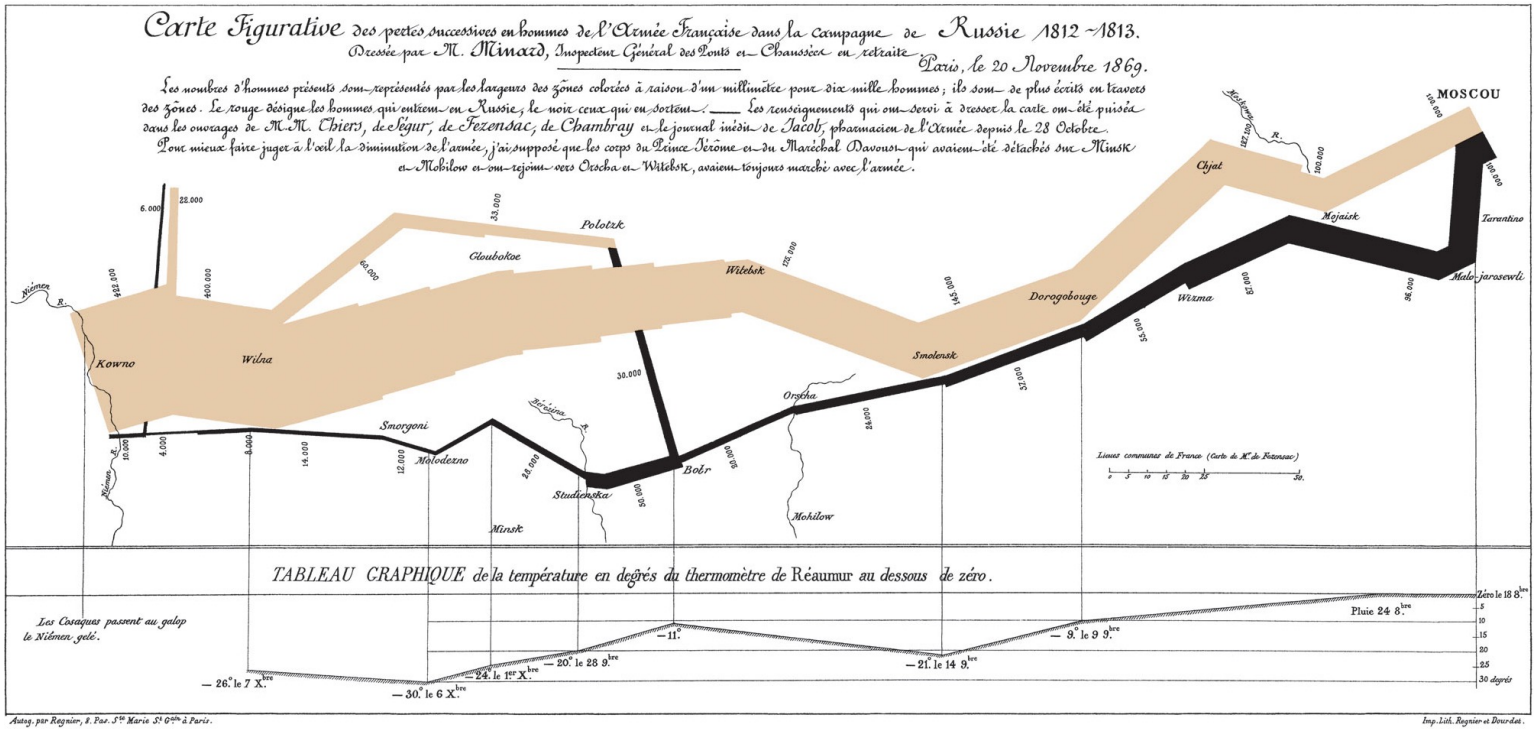
New York Times, January 27, 1981, p. D-1



3D perspectives make the bars behind look smaller than actual values.

Tufte (2001) The visual display of quantitative information, p. 70-71

Edward Tufte > Multifunctioning, Layered information



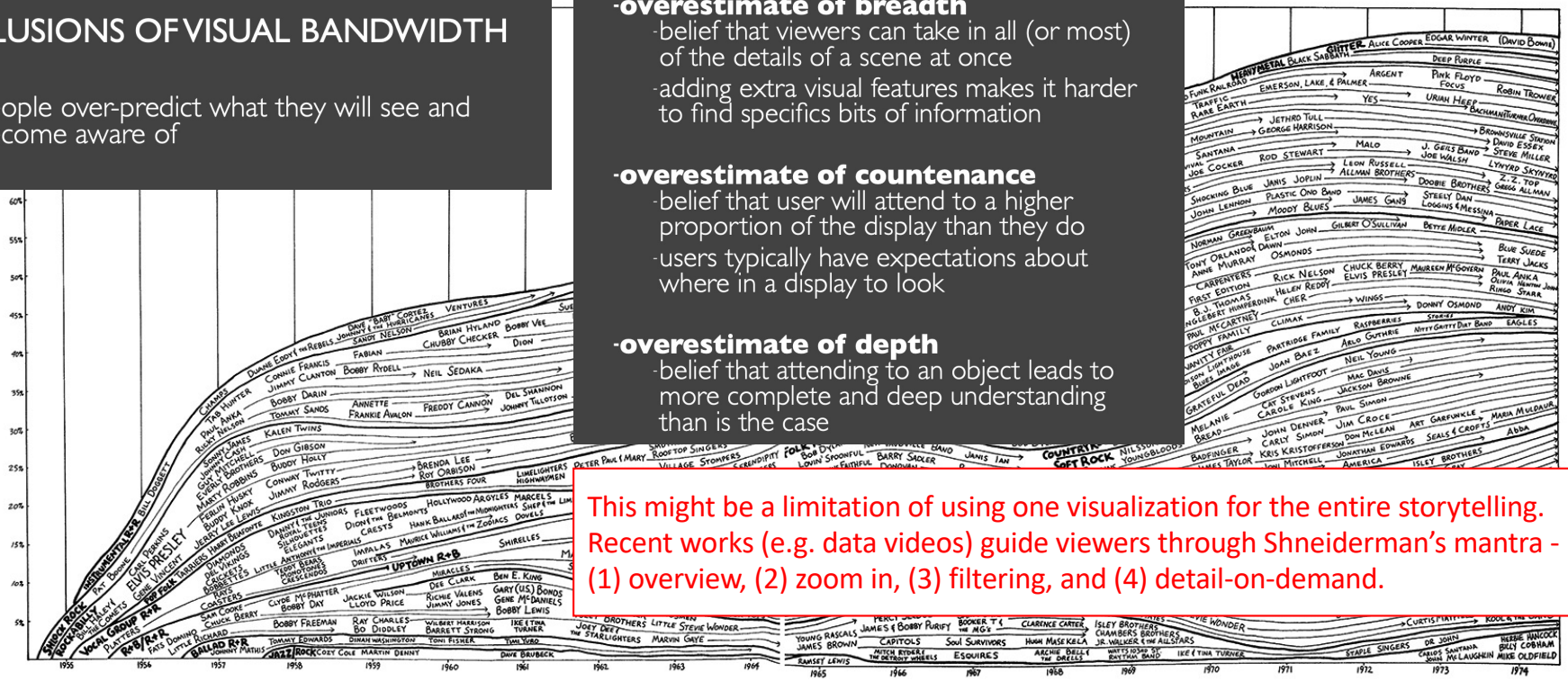
No wonder why Tufte loved Minard's work

Edward Tufte > Maximize Amount of Data Shown

ILLUSIONS OF VISUAL BANDWIDTH

people over-predict what they will see and become aware of

- **overestimate of breadth**
  - belief that viewers can take in all (or most) of the details of a scene at once
  - adding extra visual features makes it harder to find specifics bits of information
- **overestimate of countenance**
  - belief that user will attend to a higher proportion of the display than they do
  - users typically have expectations about where in a display to look
- **overestimate of depth**
  - belief that attending to an object leads to more complete and deep understanding than is the case



The Genealogy of Rock / Pop Music by Reebee Garofalo

This might be a limitation of using one visualization for the entire storytelling. Recent works (e.g. data videos) guide viewers through Shneiderman's mantra - (1) overview, (2) zoom in, (3) filtering, and (4) detail-on-demand.

Next Week

- Advanced Visualizations
  - Network
  - Hierarchy
  - High-dimensional space (ML model interpretation)
  - Trellis / Small Multiples
  - Scatterplot matrix
  - Multiple-coordinated views
  - Parallel Coordinates
  - Radial / Star chart
- Interaction
  - Brush-and-Highlighting
  - Shneiderman’s mantra
- Narratives
- Animation
- Latest research topics
  - AI-supported dataviz
  - Immersive Data Storytelling