Visualization for Data Science
CMPT 733

Steven Bergner
sbergner@cs.sfu.ca

Outline

- Visualization: What, Why, and How?
- Examples and goals
- Guidelines and Techniques

Recap: Data Science Pipeline

What	When	Who	Goal
Computer Science	1950-	Software Engineer	Write software to make computers work

Plan → Design → Develop → Test → Deploy → Maintain

What	When	Who	Goal
Data Science	2010-	Data Scientist	Extract insights from data to answer questions

Collect→ Clean → Integrate → Analyze → Visualize → Communicate

What role does Visualization play?

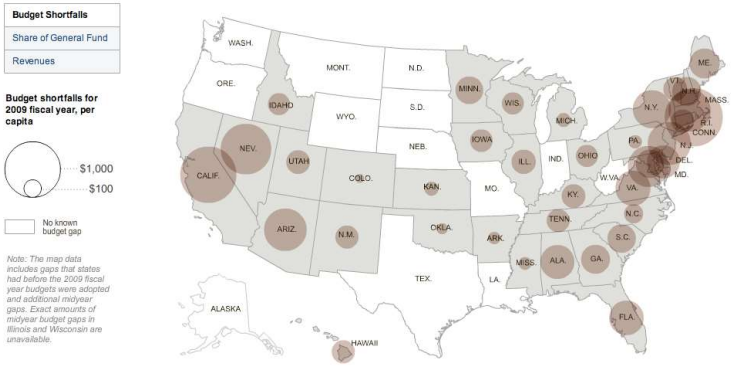
vi·su·al·ize

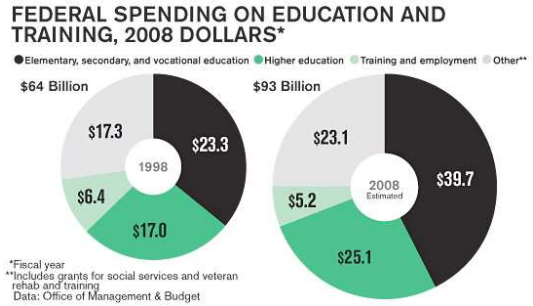
1. To form a mental image
2. To make visible

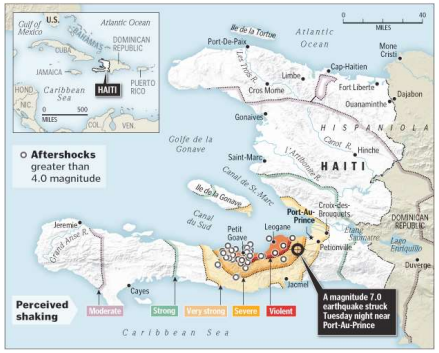
[The American Heritage Dictionary]

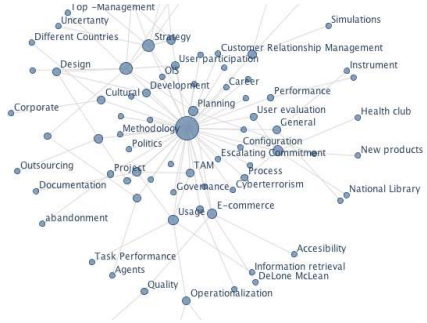
What is Data Visualization?

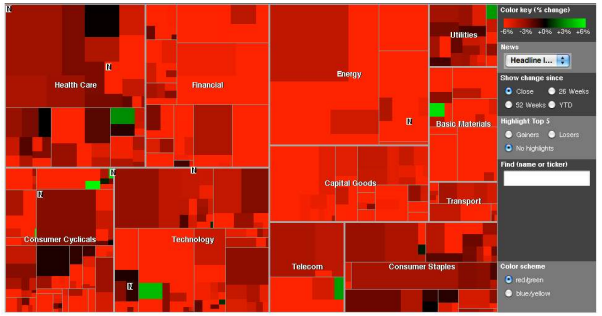
Visualization: To convey information through visual representations







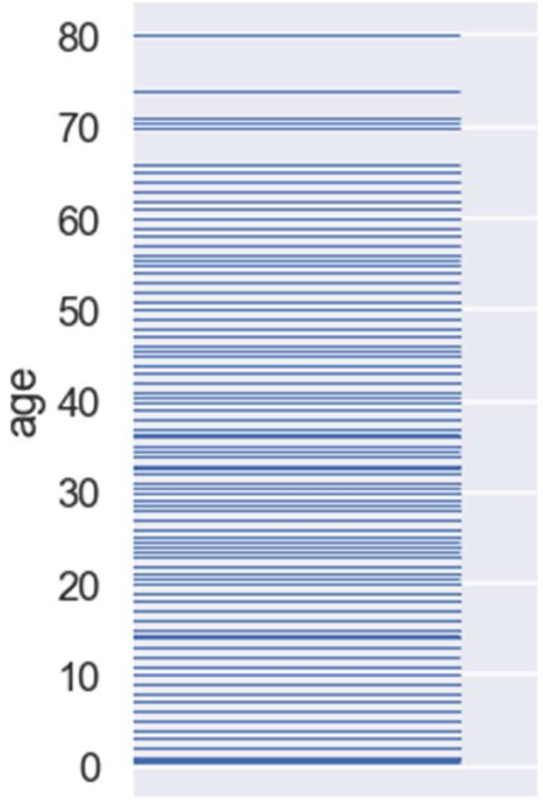




Computer Readable

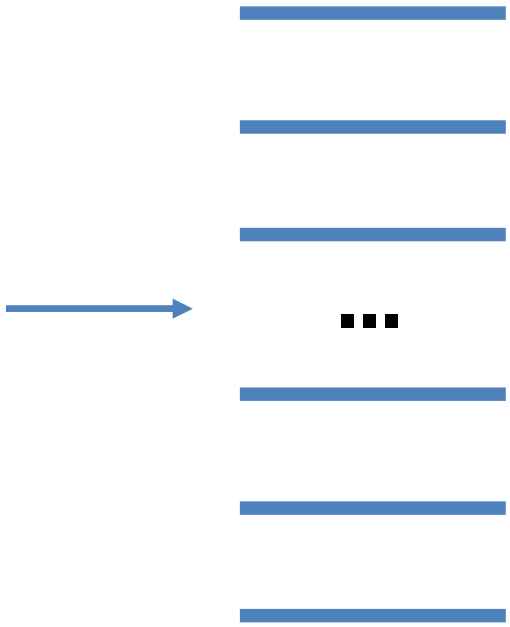
	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0

Human Readable

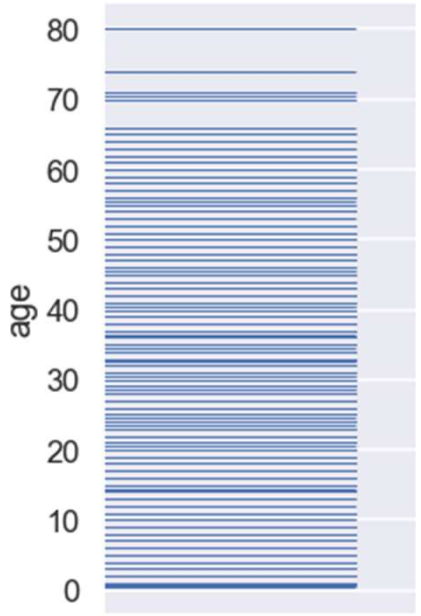


Data from https://www.kaggle.com/c/titanic

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0



10px
16px
11px
...
0px
11px
15px



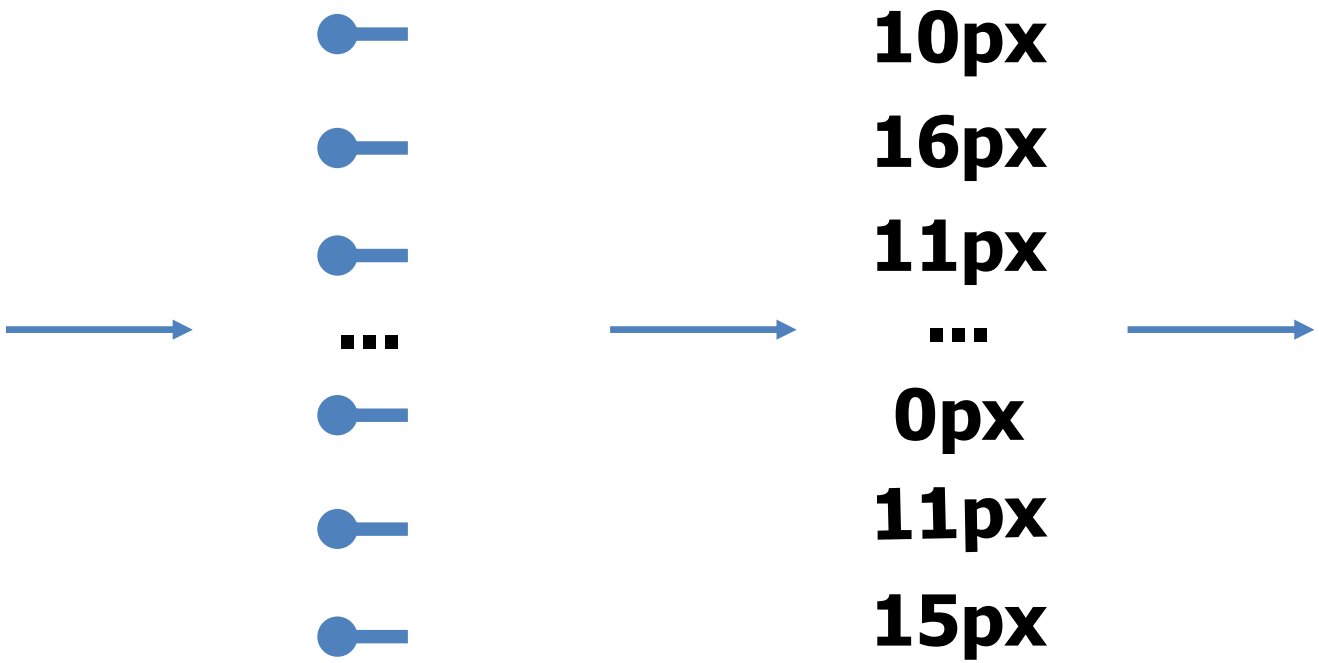
Mark

(Represents a datum)

Encoding

(Maps datum to visual position)

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0

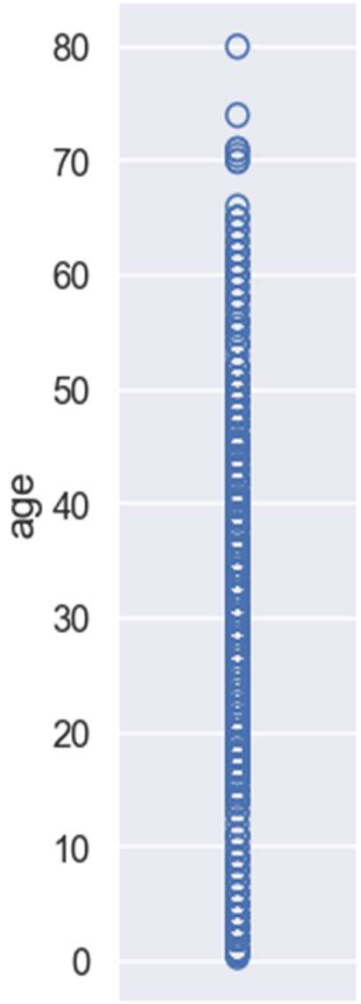


Mark

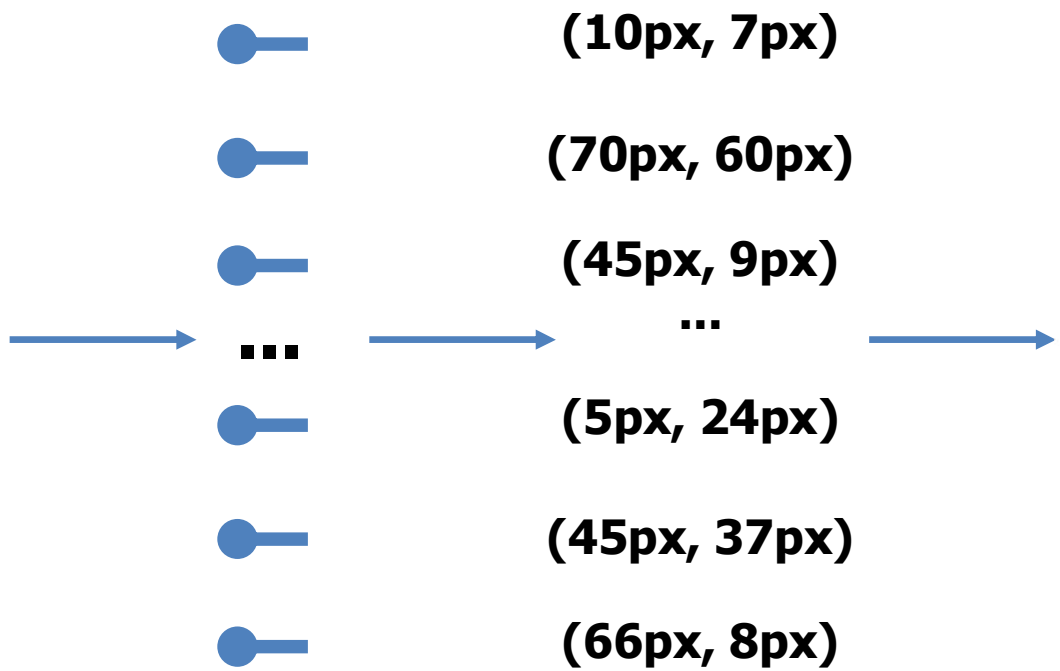
(Represents a datum)

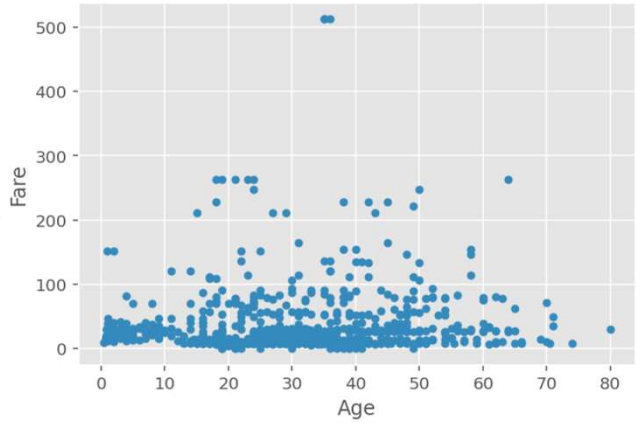
Encoding

(Maps datum to visual position)



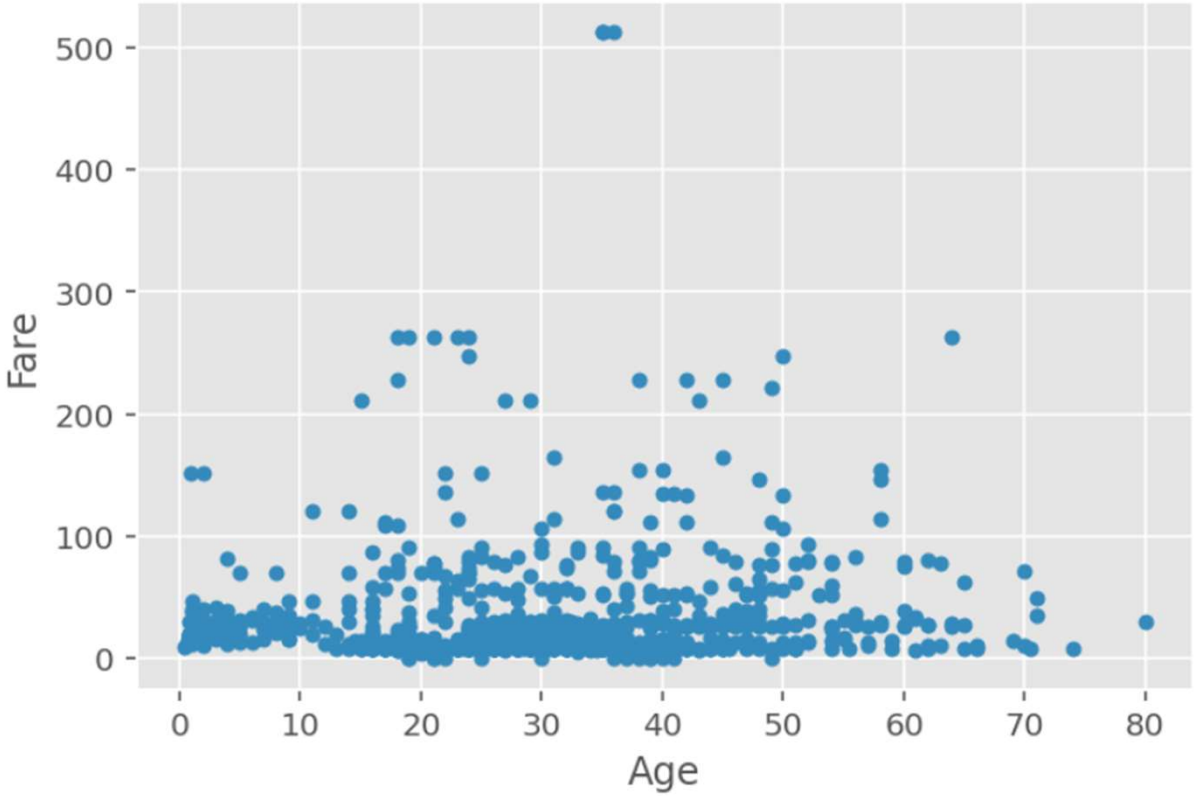
	age	fare
0	22.0	7.25
1	38.0	71.28
2	26.0	7.92
...	...	...
888	NaN	23.45
889	26.0	30.00
890	32.0	7.75





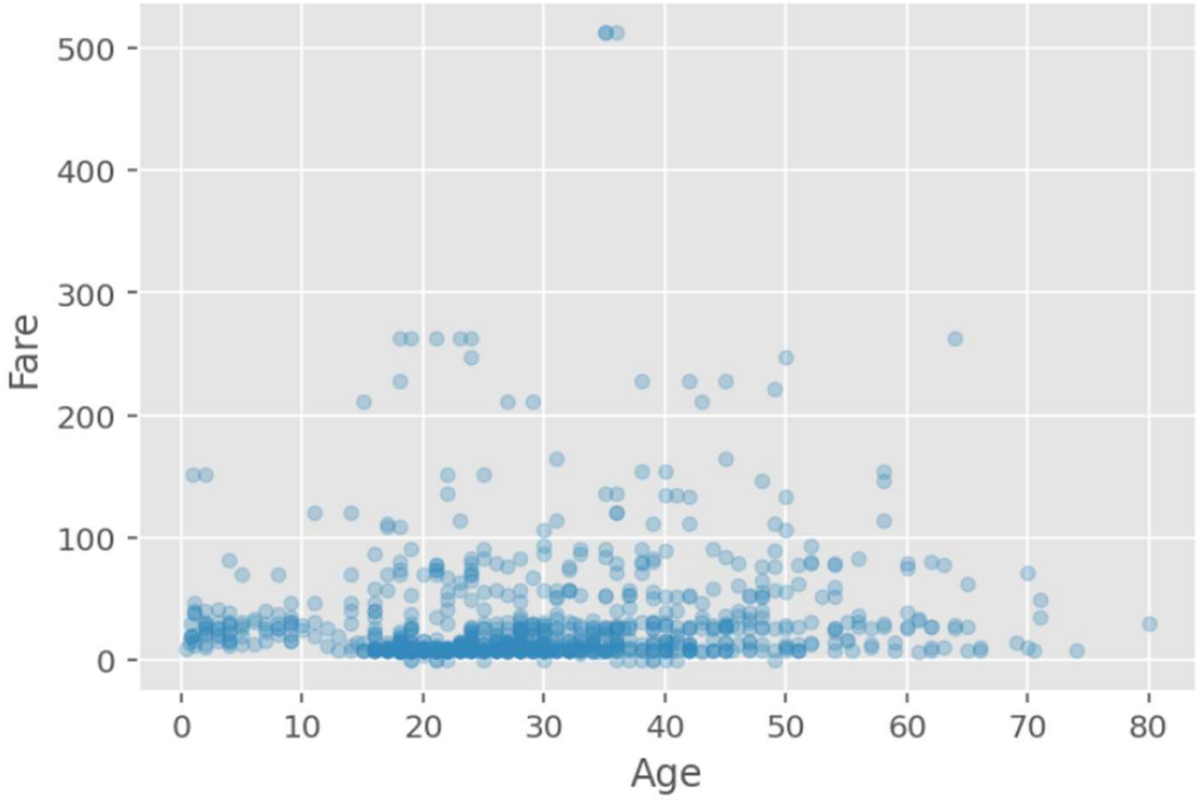
Visualizations are for Humans

"Looks like older people didn't spend more than younger people."



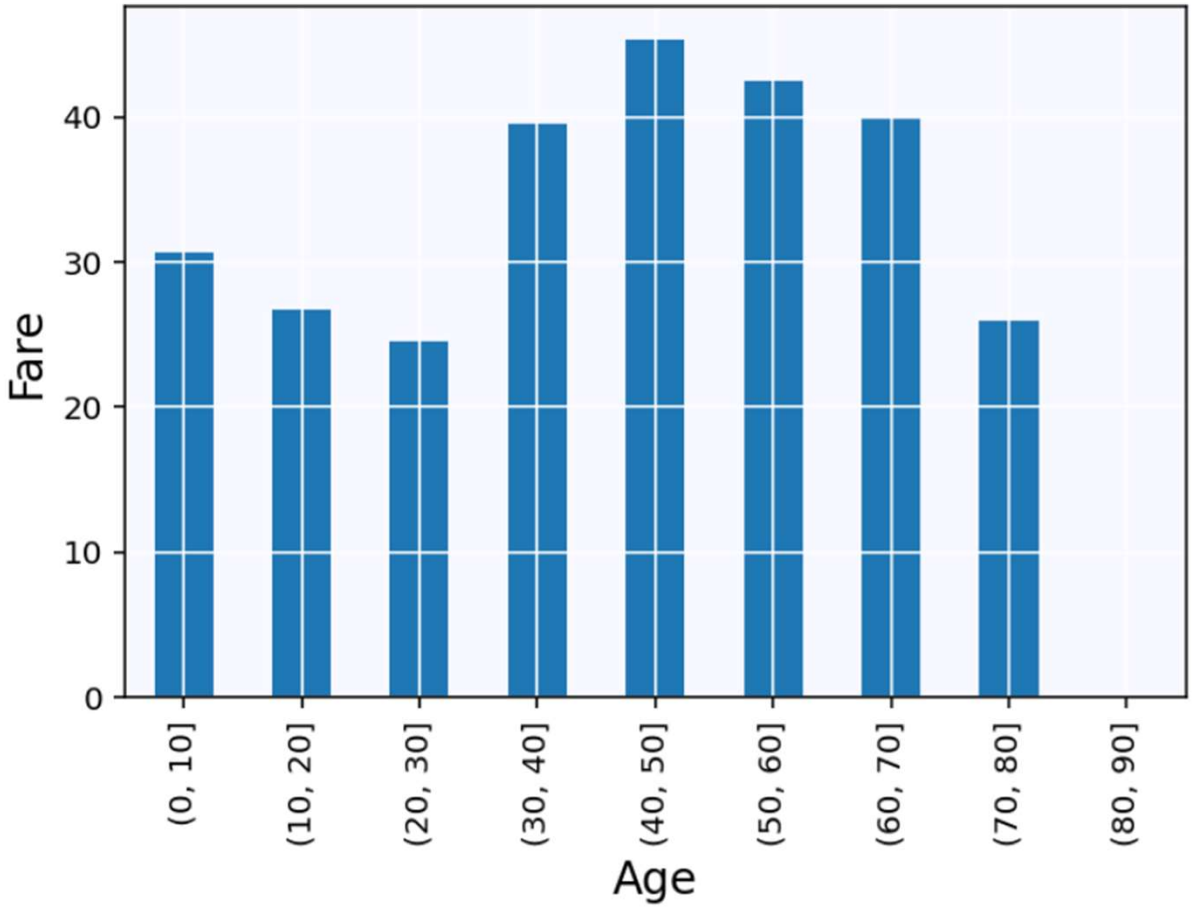
Visualizations are for Humans

"Looks like older people didn't spend more than younger people."



Visualizations are for Humans

"Looks like older people didn't spend more than younger people."



Visualizations are for Humans

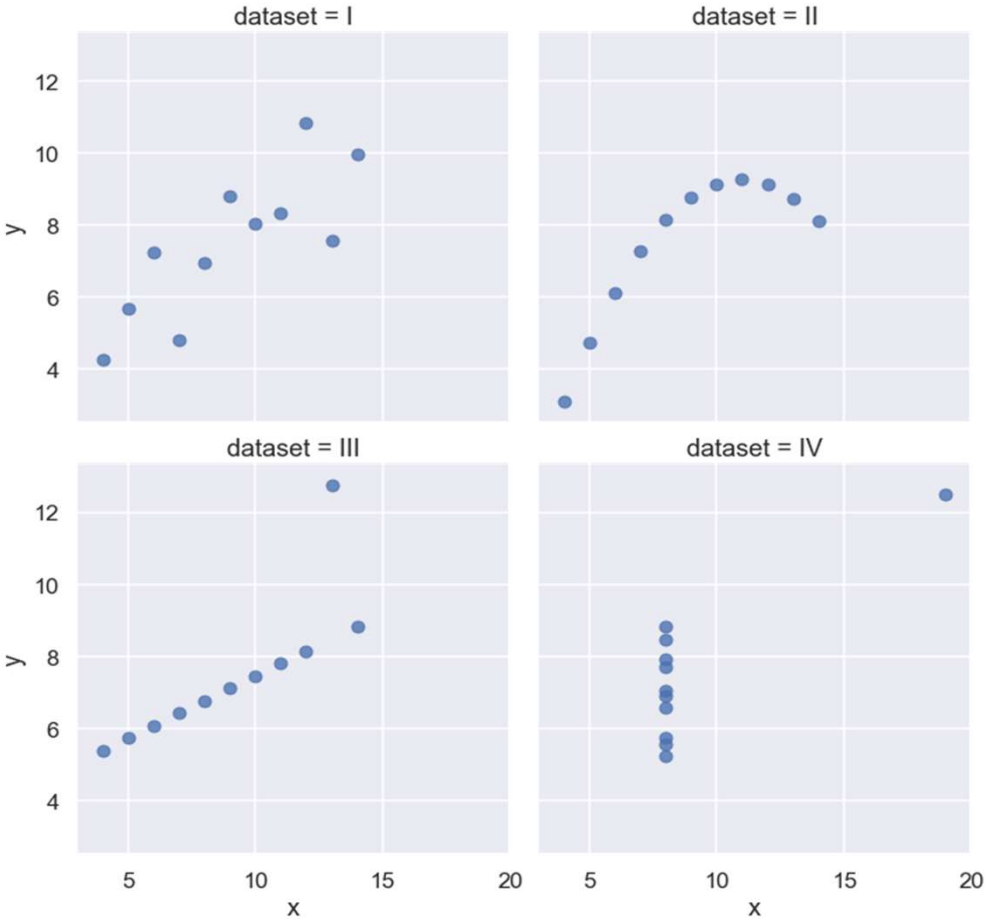
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

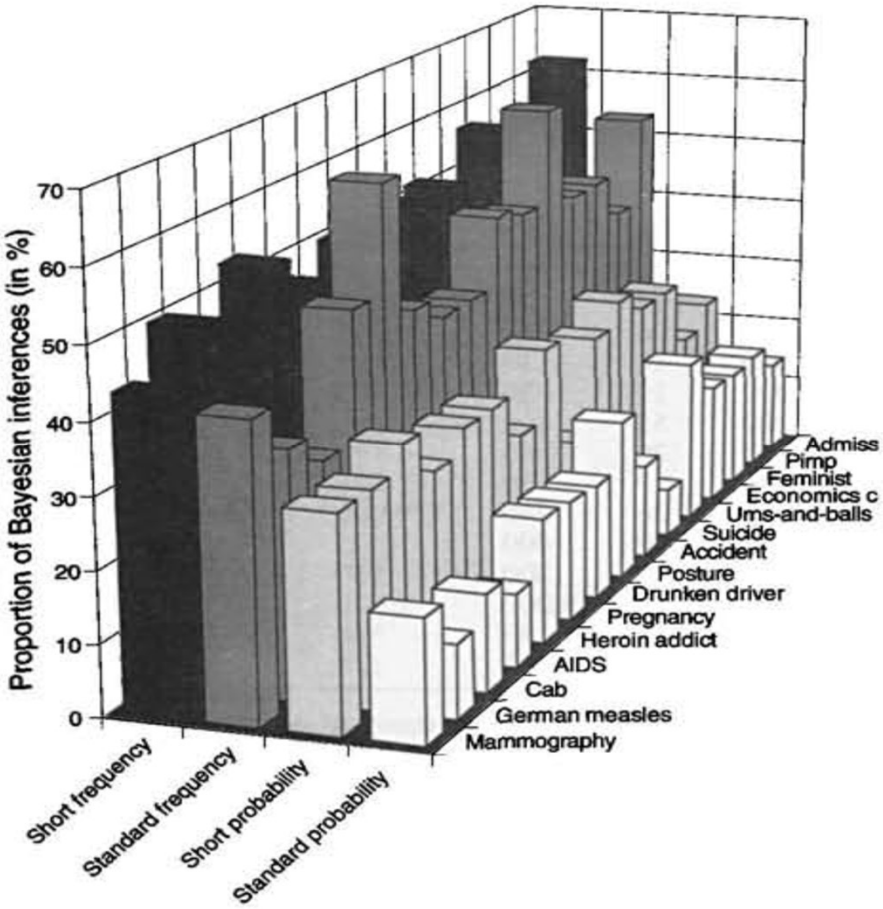
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

Visualizations are for Humans



Human eyes good at seeing visual patterns!

Visualizations are for Humans



Human eyes good at seeing visual patterns!...

Sometimes.

Why Data Visualization?

- One goal of data science is to inform human decisions
  - Excellent plots directly address this goal
  - Sometimes the most useful results from data analysis are the visualizations!

- Data viz is not as simple as calling plot()
  - Many plots possible, but only a few are useful
  - Every visualization has tradeoffs

Python example: seaborn

Demo/Tutorial: https://seaborn.pydata.org/tutorial.html

seaborn

Best used with tidy (aka long-form) data

- Seaborn will perform groupby automatically

Typical usage:

```
sns.someplot(x='...', y='...', data=...)
```

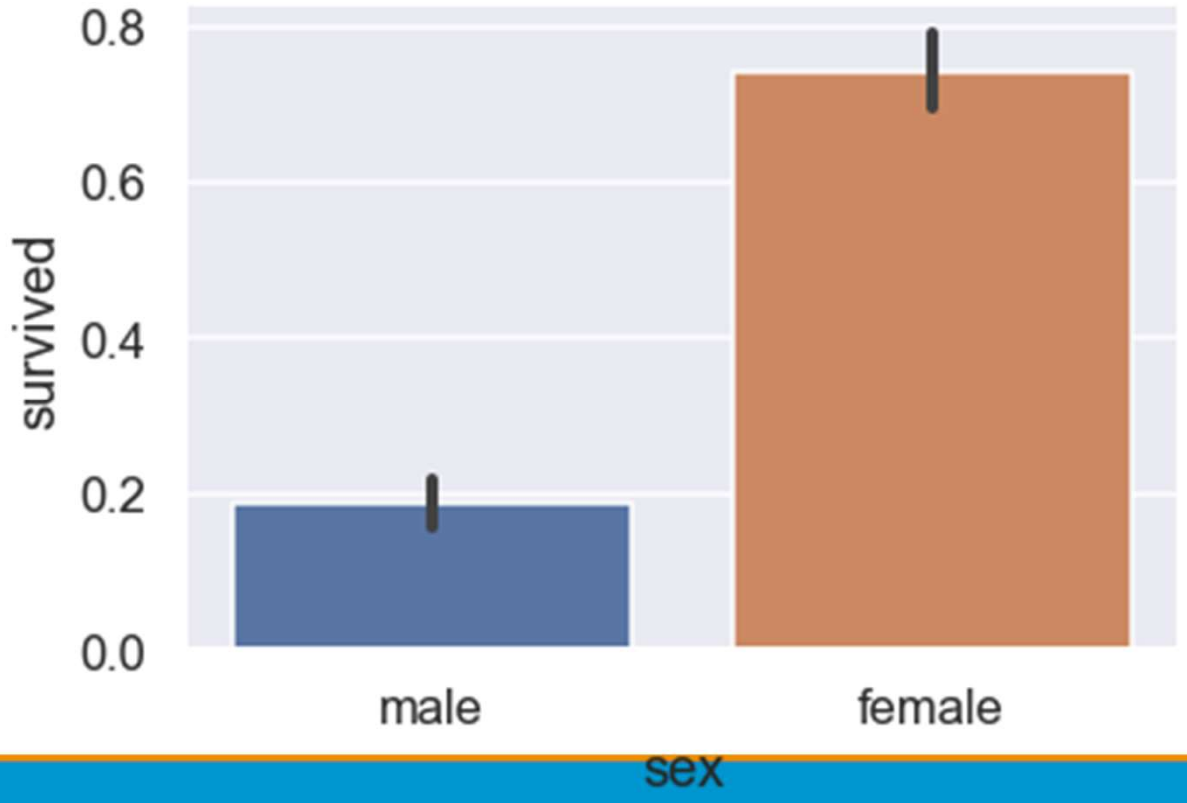
Plot types

- Dot plot, Rug plot
- Jitter plot
- Error bar plot
- Box plot
- Histogram
- Kernel density estimate
- Cumulative distribution function

seaborn

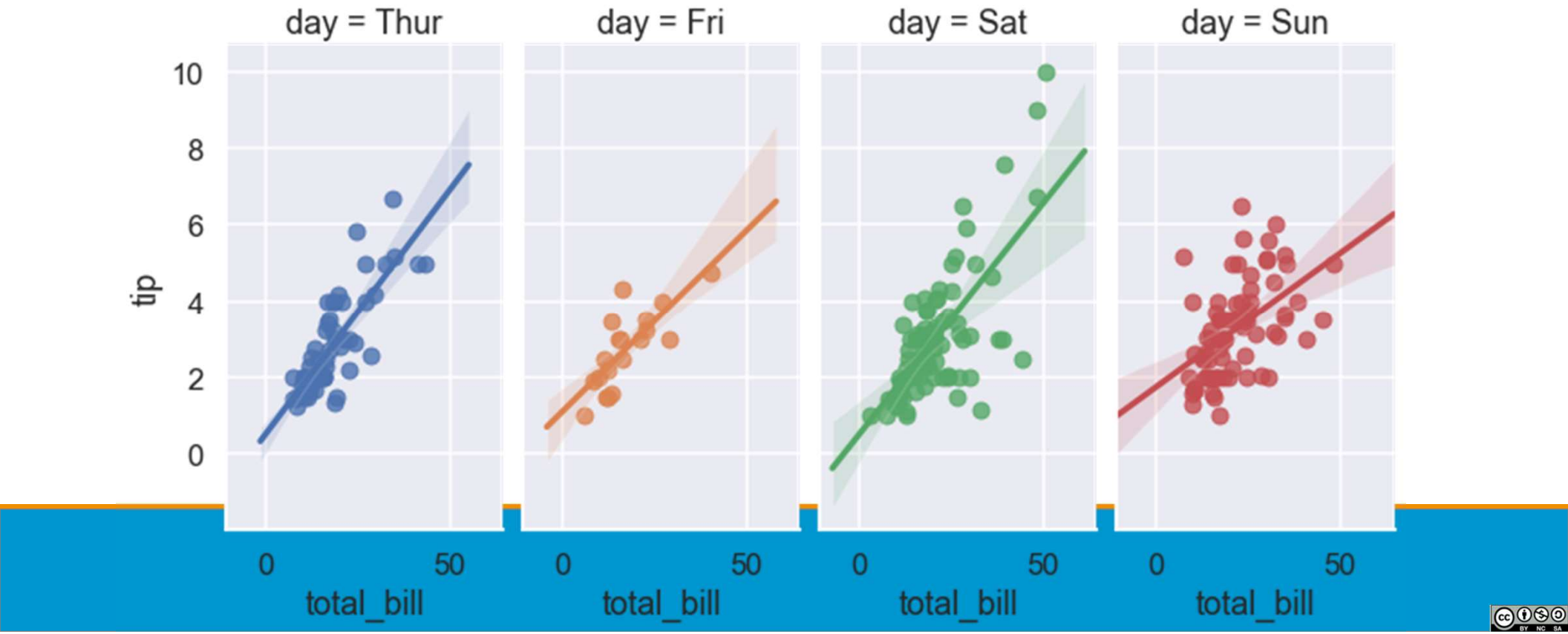
```
sns.barplot(x='sex', y='survived', data=ti)
```

	survived	class	sex	age	fare
0	0	Third	male	22.0	7.25
1	1	First	female	38.0	71.28
2	1	Third	female	26.0	7.92
...	...	...	...	...	...
888	0	Third	female	NaN	23.45
889	1	First	male	26.0	30.00
890	0	Third	male	32.0	7.75



seaborn

```
sns.lmplot(x="total_bill", y="tip",
           col="day", hue="day", data=tips)
```



Customizing Plots using matplotlib

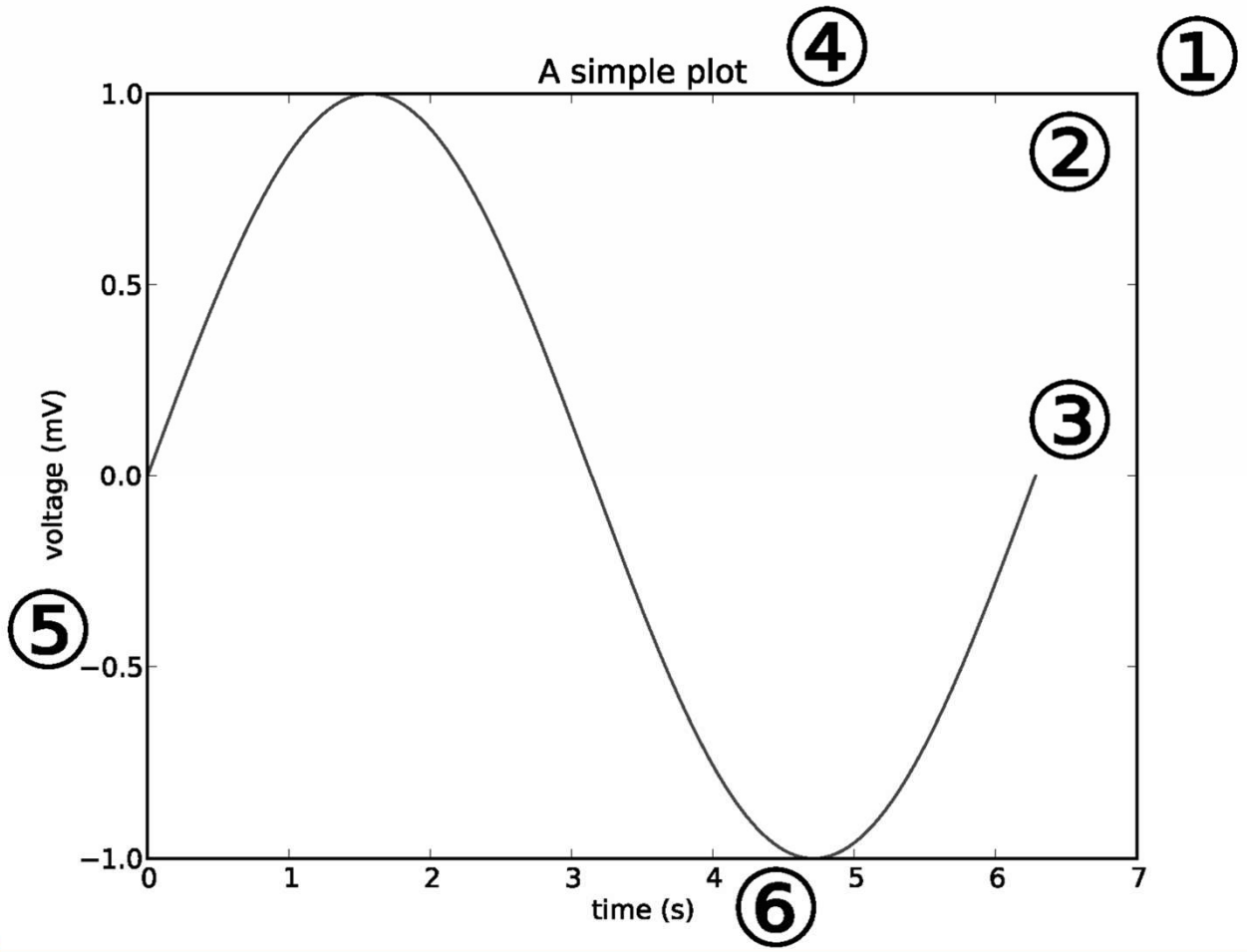
Demo/Tutorial: https://matplotlib.org/tutorials/introductory/pyplot.html

matplotlib

- Underlying library for seaborn, pandas, and most other Python plotting libraries
- A Figure contains several Axes. Each Axes contains a plot.
- When creating a plot, a new figure + axes is created if not already initialized.
  - Matplotlib remembers that axes for the duration of the cell (hidden state!)
- Note: Axes = one chart within a larger Figure
  - Axis = x or y-axis within a chart (sorry!)

matplotlib

1. Figure
2. Axes
3. Line
4. Title
5. YAxis
6. XAxis



Typical Workflow

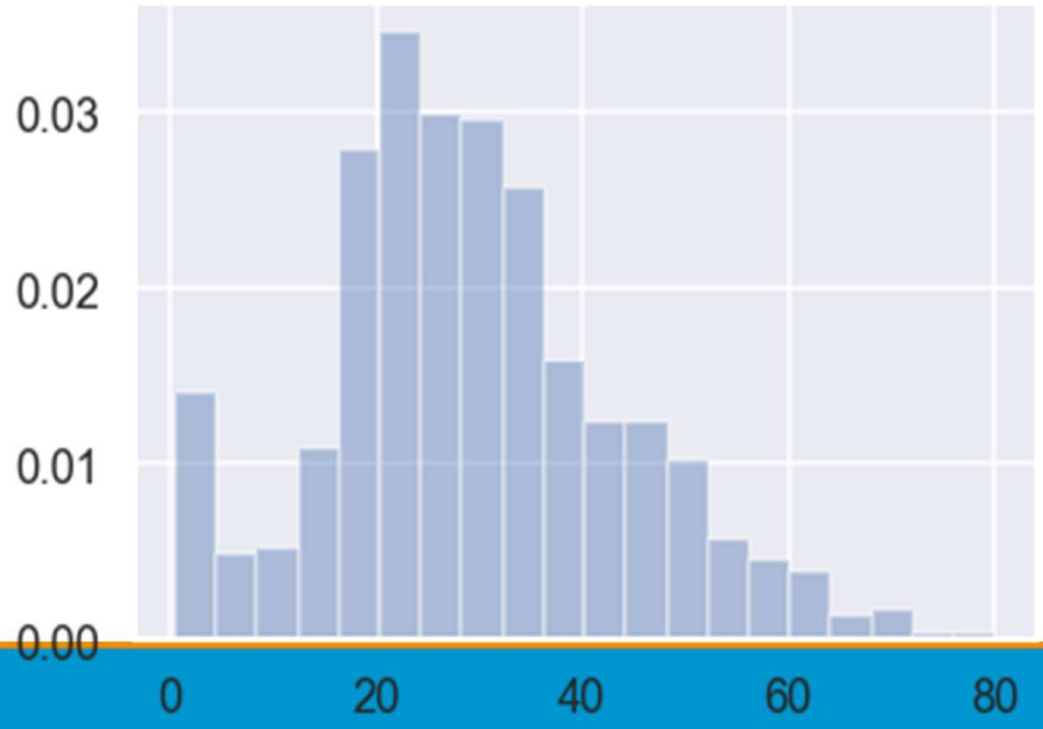
- Start with seaborn plot
  - Get as close to desired result as possible
- Fine-tune with matplotlib, e.g:
  - Changing title, axis labels
  - Annotating interesting points
- Publication-ready plots take lots of fine-tuning!

Common Visualizations for One Quantitative Variable

Histograms

Always have proportion per unit on y-axis

- Total area = 1
- Deciding on number of bins is hard! Trial-and-error process.



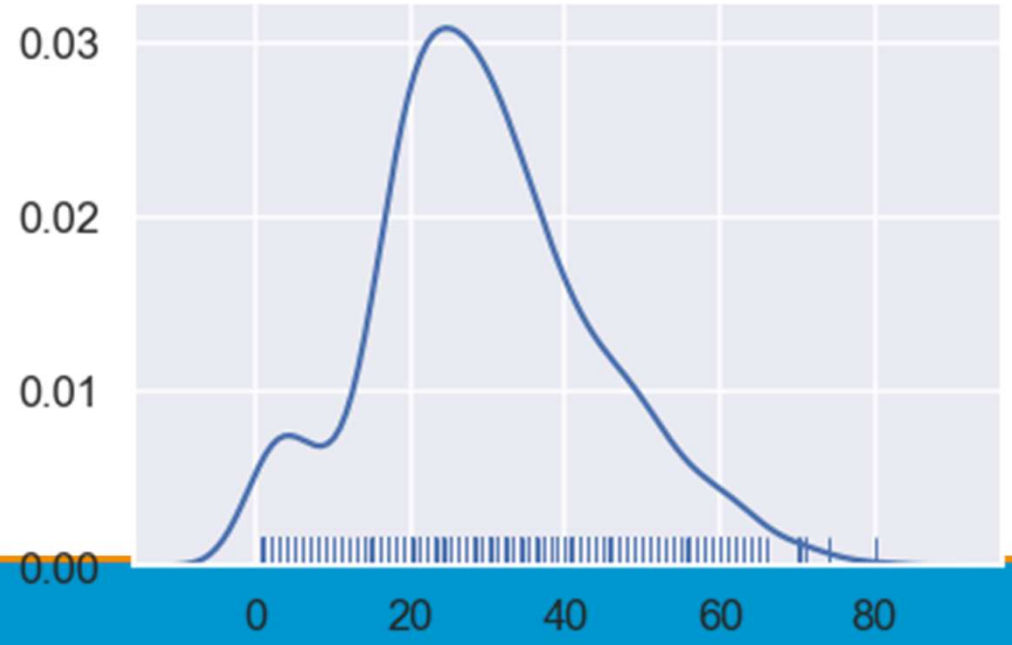
Density Plots

Density plots similar to a “smoothed” histogram

- More on smoothing tomorrow

Rug plots put a tick at each data point

- Used to show all points

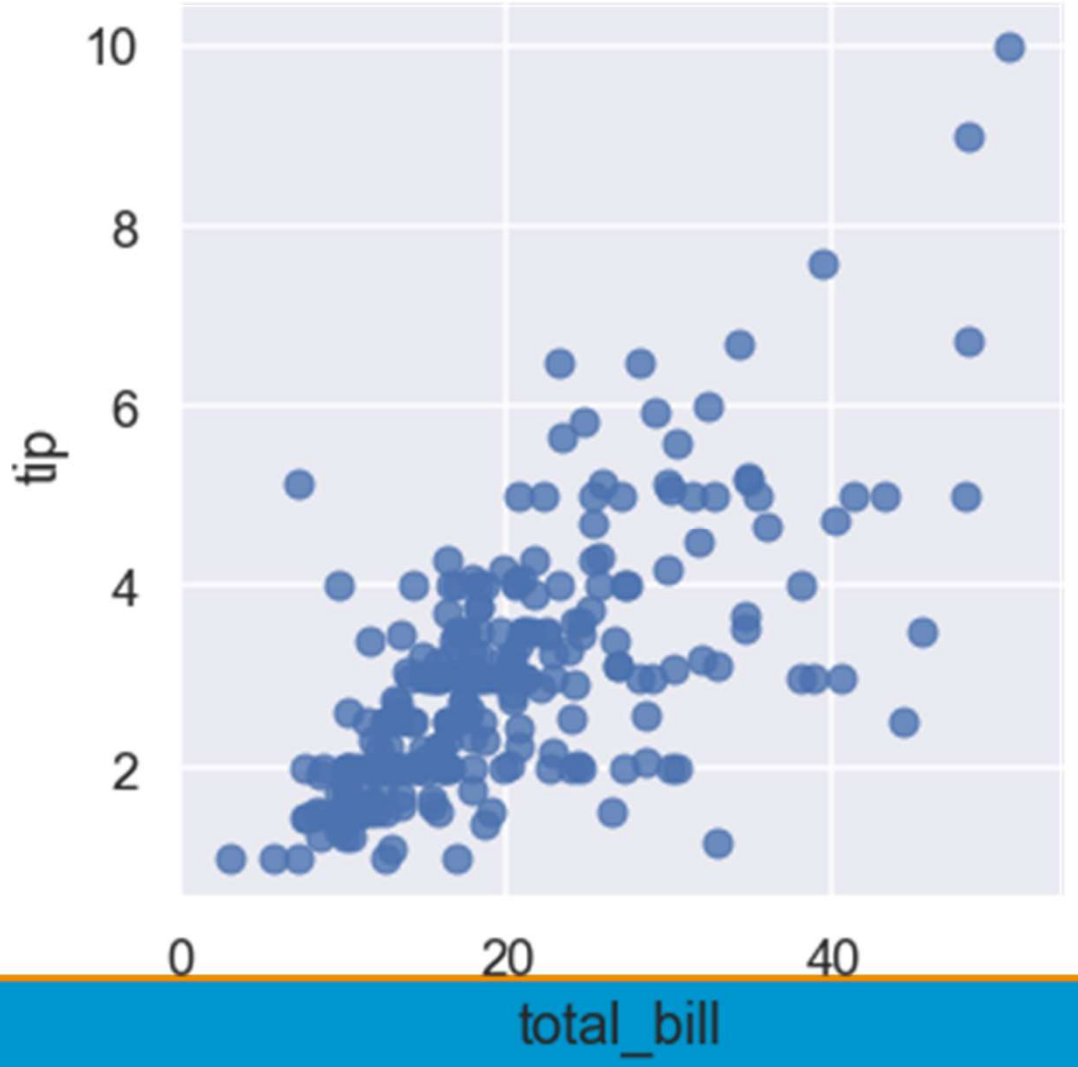


Common Visualizations for Two Quantitative Variables

Scatter Plots

Used to reveal relationships between pair of variables

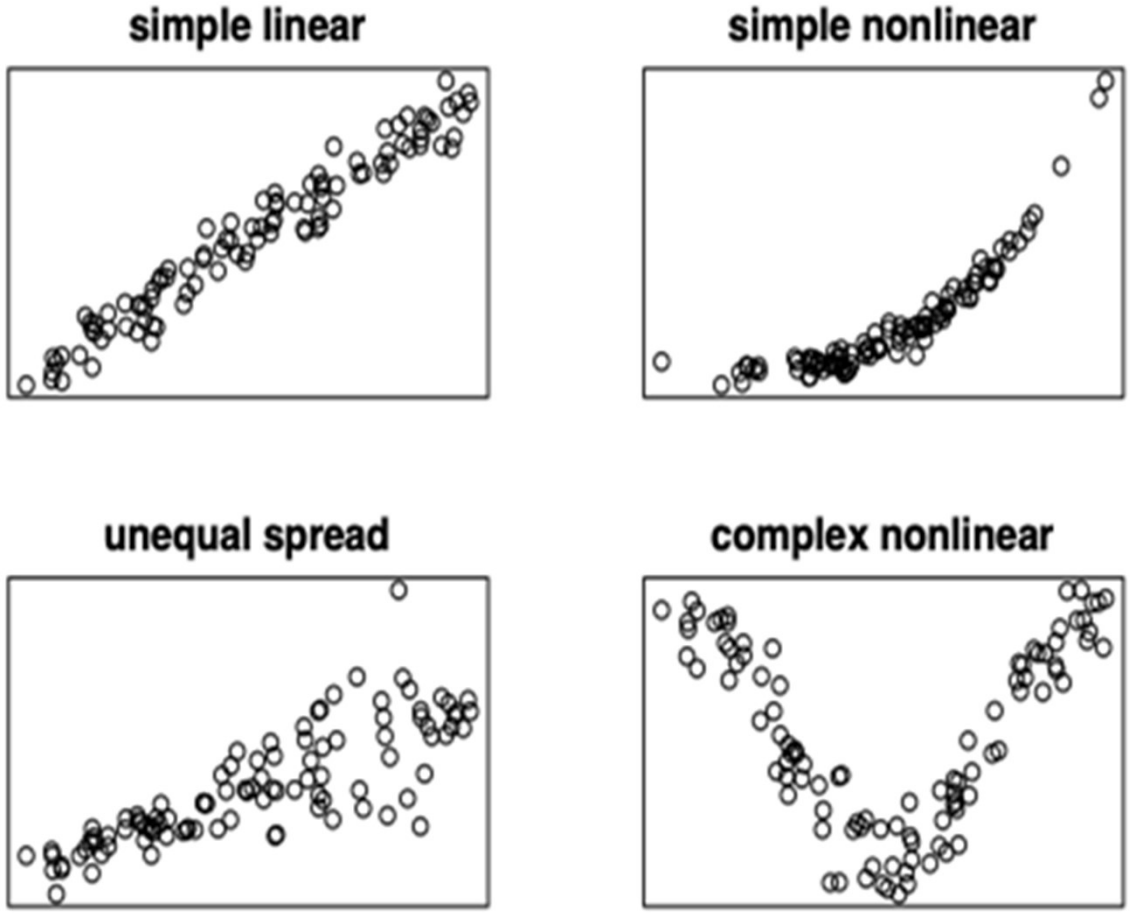
- Susceptible to overplotting
  - Points overlap!



Scatter Plots

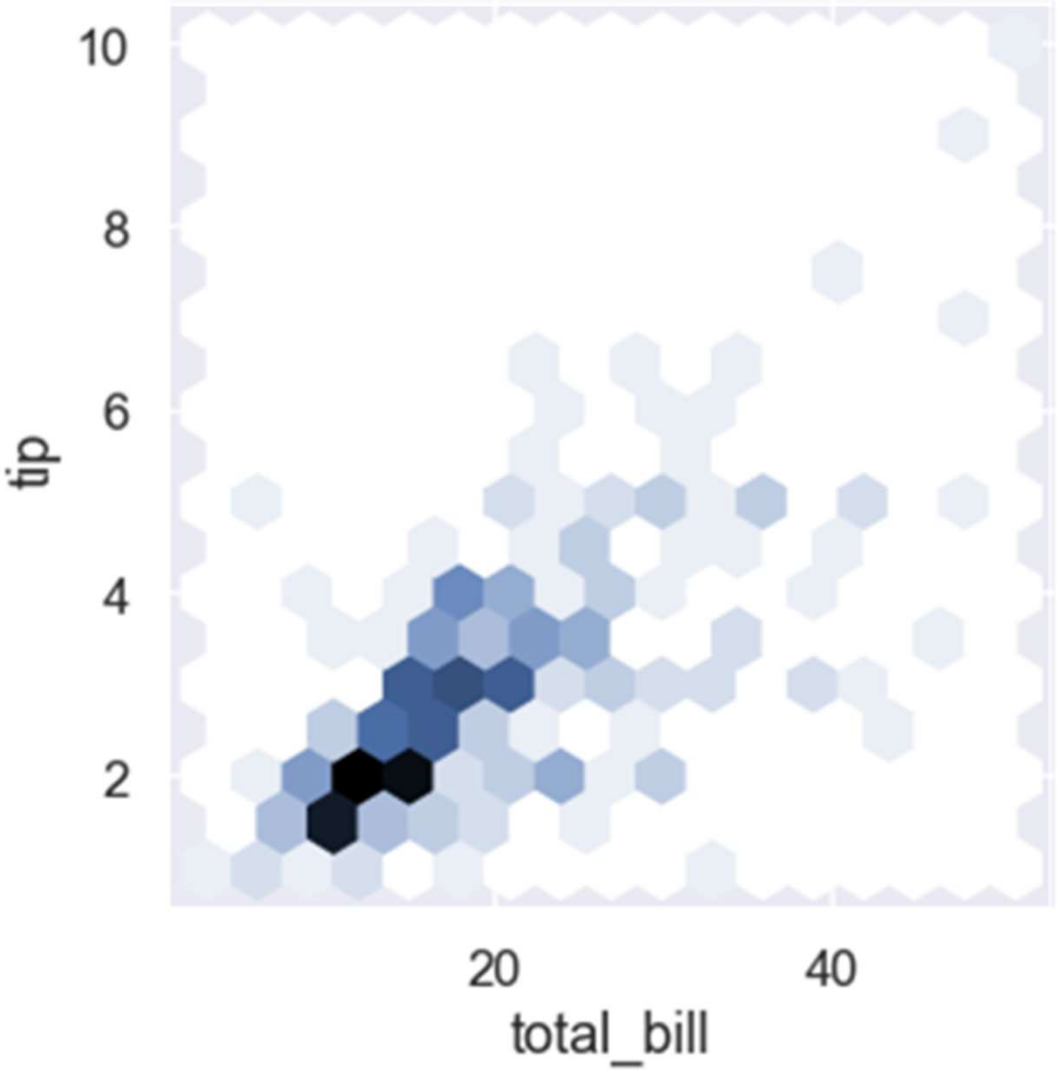
Used to inform model choices

- E.g. simple linear model requires linear trend and equal spread.



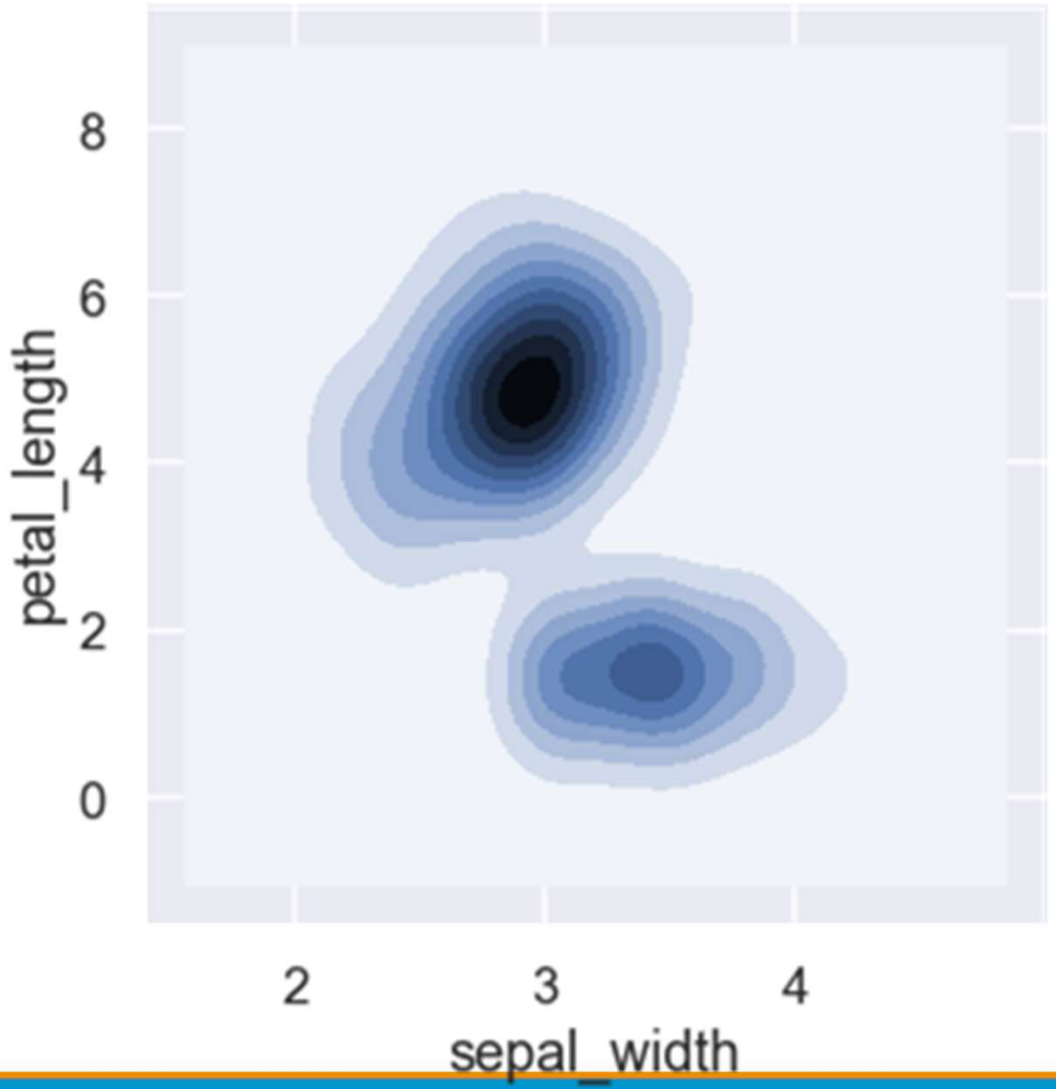
Hex Plots

- Equivalent of histogram in two dimensions
- Shaded hexagons usually correspond to more points



2D Density Plots

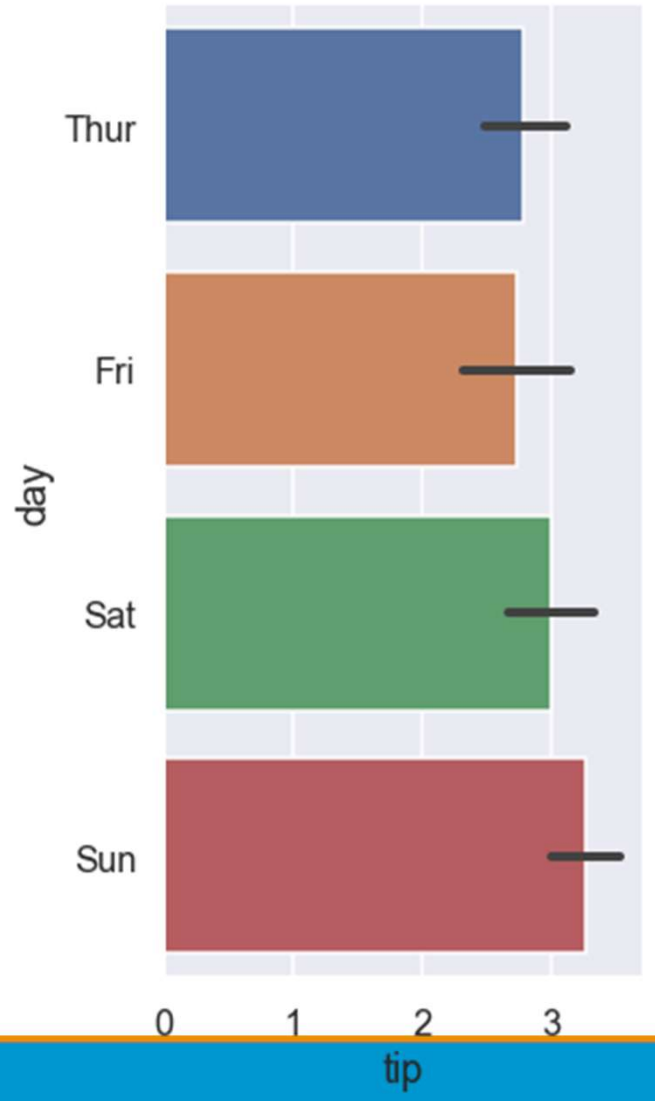
- Density plots also work in two dimensions!



Common Visualizations for Qualitative + Quantitative Variable

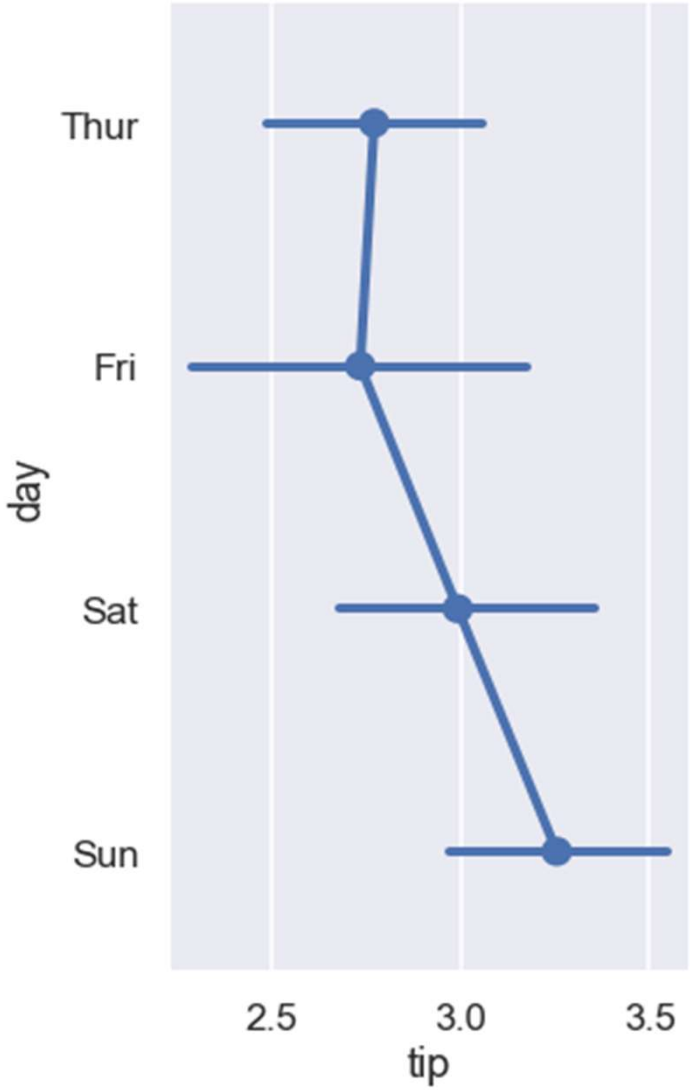
Bar Plots

- Typically use horizontal bars to avoid label overlap
- Can also plot confidence intervals on bars if appropriate



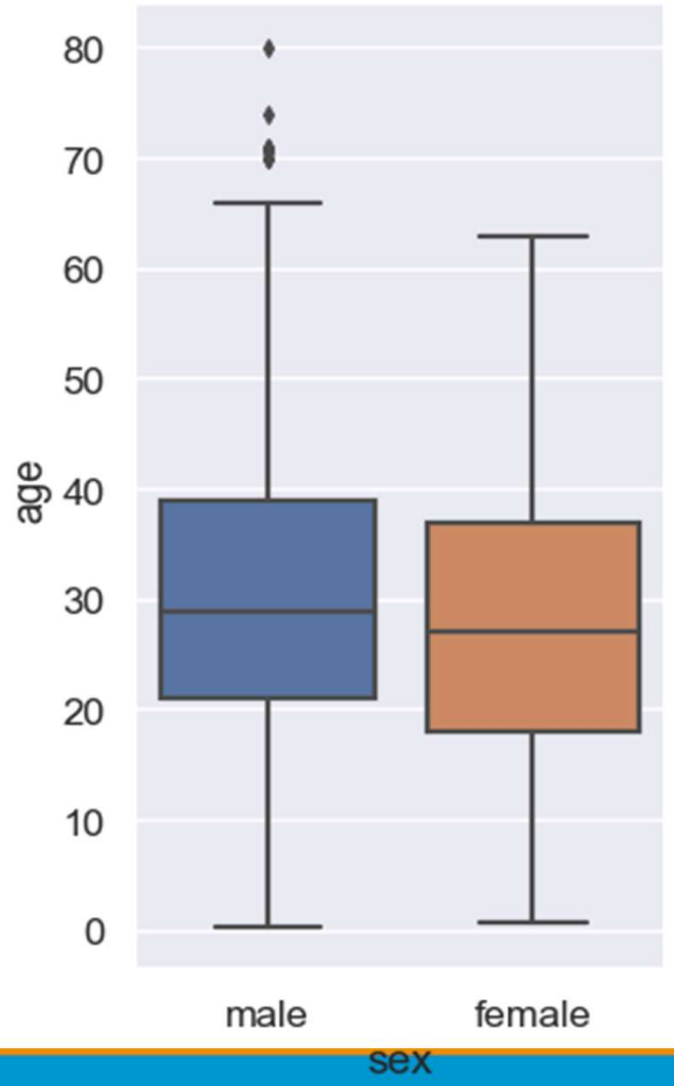
Point Plots / Dot Plots

- Minimal cousin of the bar plot
- Some prefer point plots since the bar widths in a bar plot have no meaning



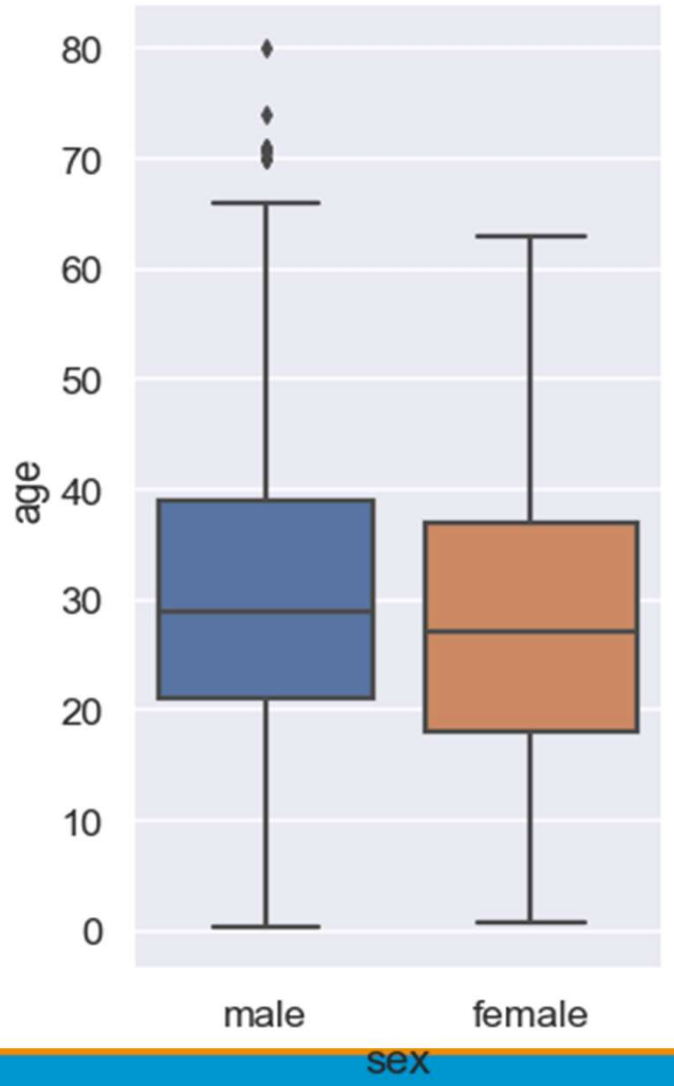
Box Plots

- Used to compare distributions
- Uses quartiles
  - Q1: 25th percentile
  - Q2 (median): 50th
  - Q3: 75th
- Middle line = median
- Box shows 1st and 3rd quartile
- Whiskers show rest of data
- Outliers = 1.5 * (Q3 - Q1) past Q1 or Q3



Box Plots

- Outliers plotted beyond whiskers
- Interquartile range IQR = Q3 - Q1
- Outliers are defined as:
  - 1.5 * IQR beyond Q1 or Q3
- Example for male ages:
  - Q1 = 21; Q2 = 29; Q3 = 39
  - IQR = 18; 1.5*IQR = 27
  - Outliers are:
    - Above Q3 + 1.5*IQR = 66
    - Below Q1 - 1.5*IQR = -6

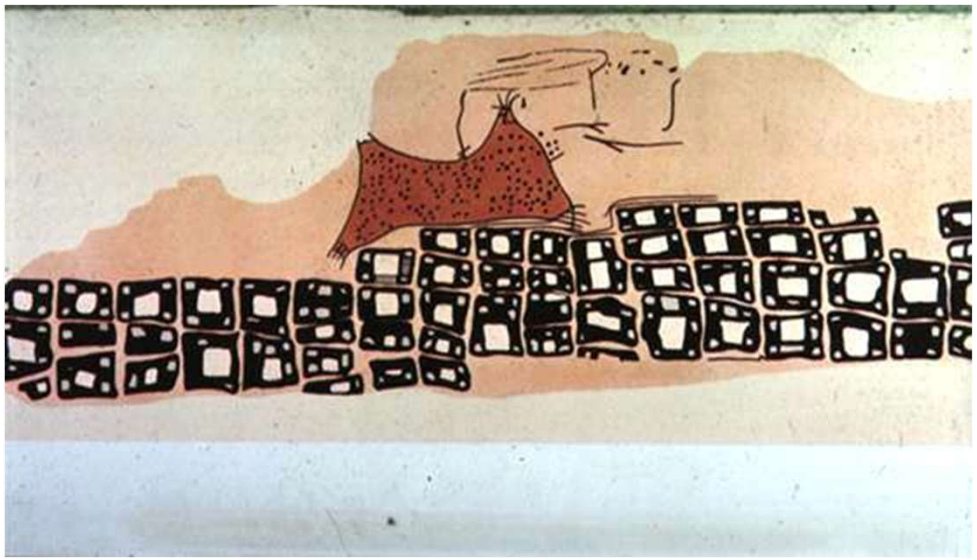


Summary

- Data visualization is underappreciated!
- Use seaborn + matplotlib
  - Pandas also has basic built-in plotting methods
- Types of variables constrain the charts you can make
  - Single quantitative: histogram, density plot
  - 2+ quantitative: scatter plot, 2D density plot
  - Quantitative + qualitative: bar plot, point plot, box plot

Visualization Goals

Map

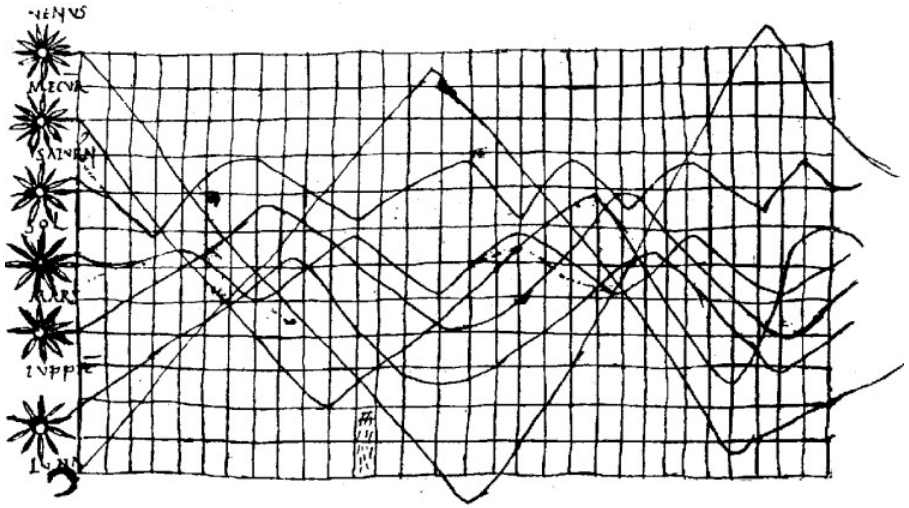


Konya town map, Turkey, c. 6200 BC

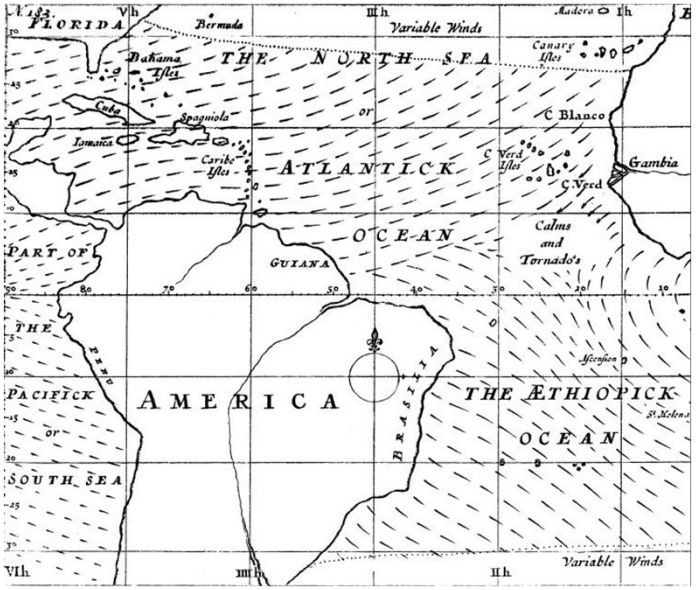
Anaximader's Map of the World

Anaximander of Miletus, c. 550 BC

Map

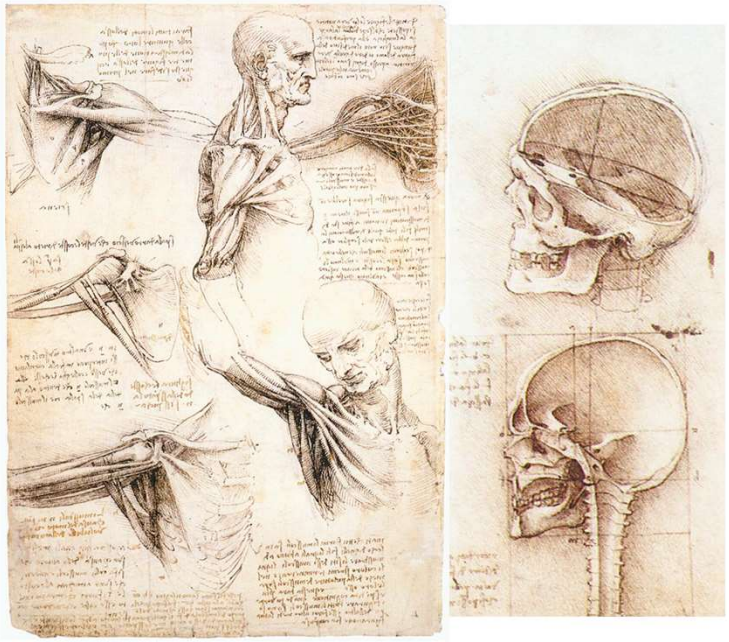


Planetary Movement Diagram, c. 950



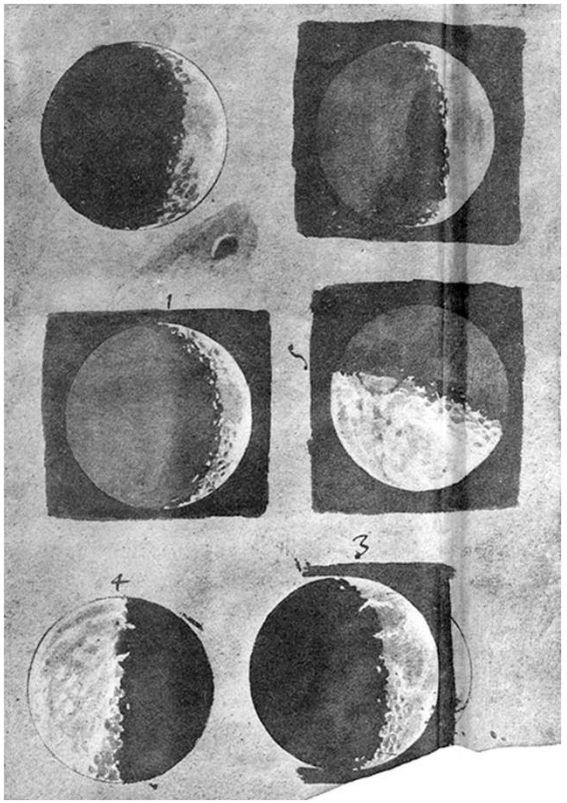
Halley's Wind Map, 1686

Record



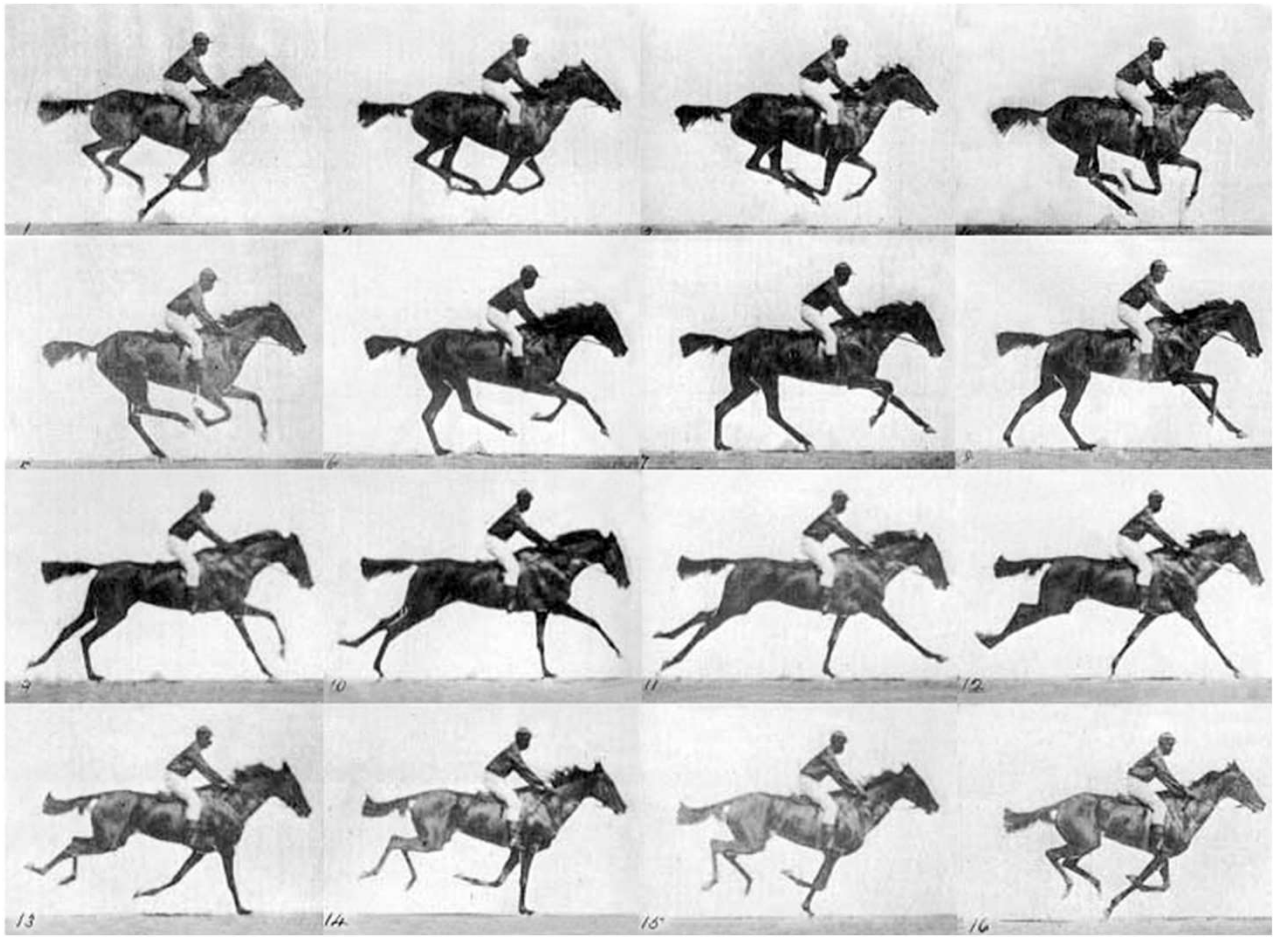
Leonardo Da Vinci, ca. 1500

William Curtis (1746-1799)



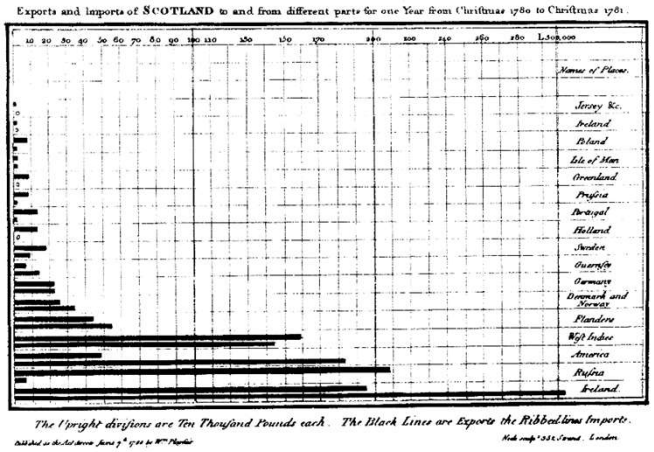
Galileo Galilei, 1616

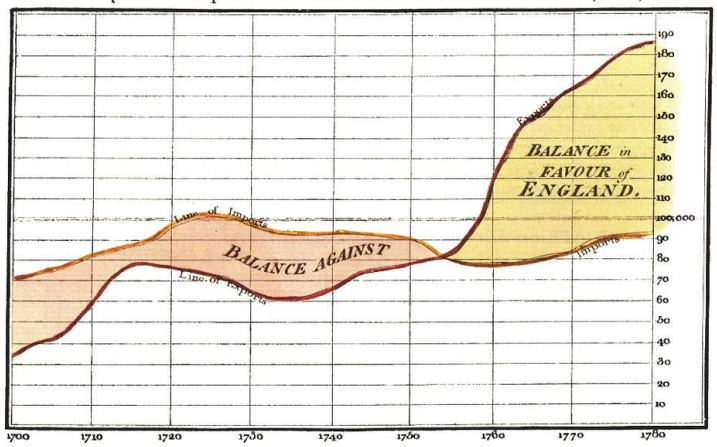
Record



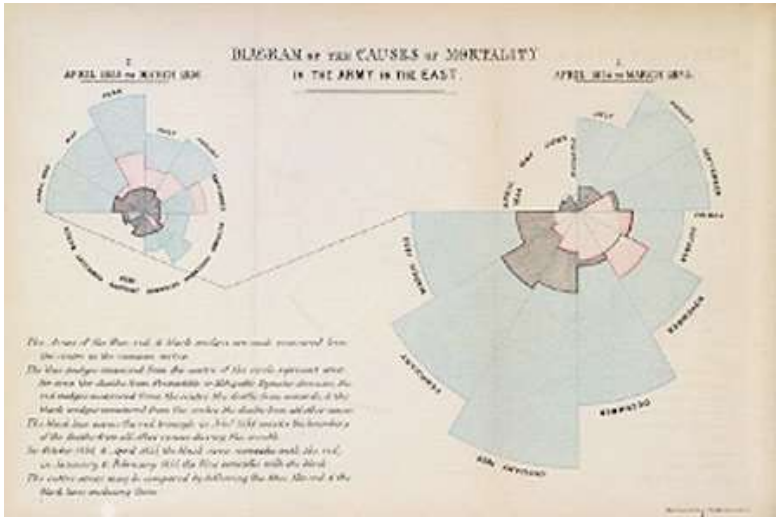
E. J. Muybridge, 1878

Abstract



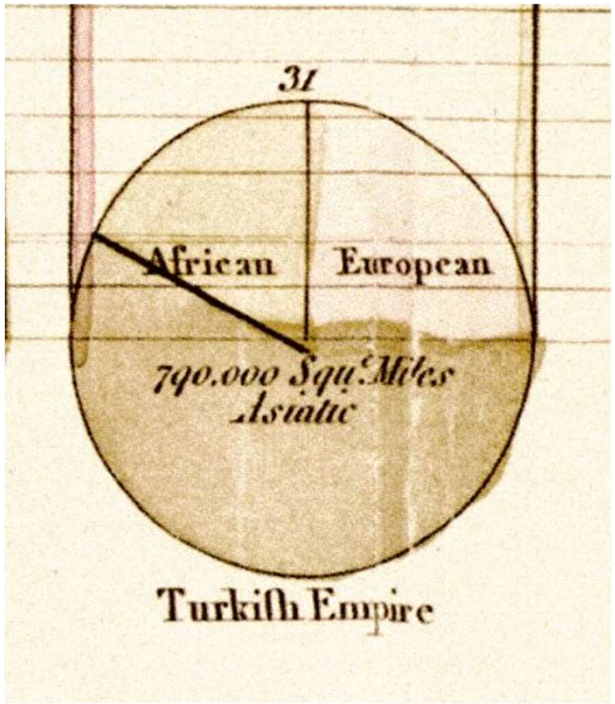


W. Playfair, 1786

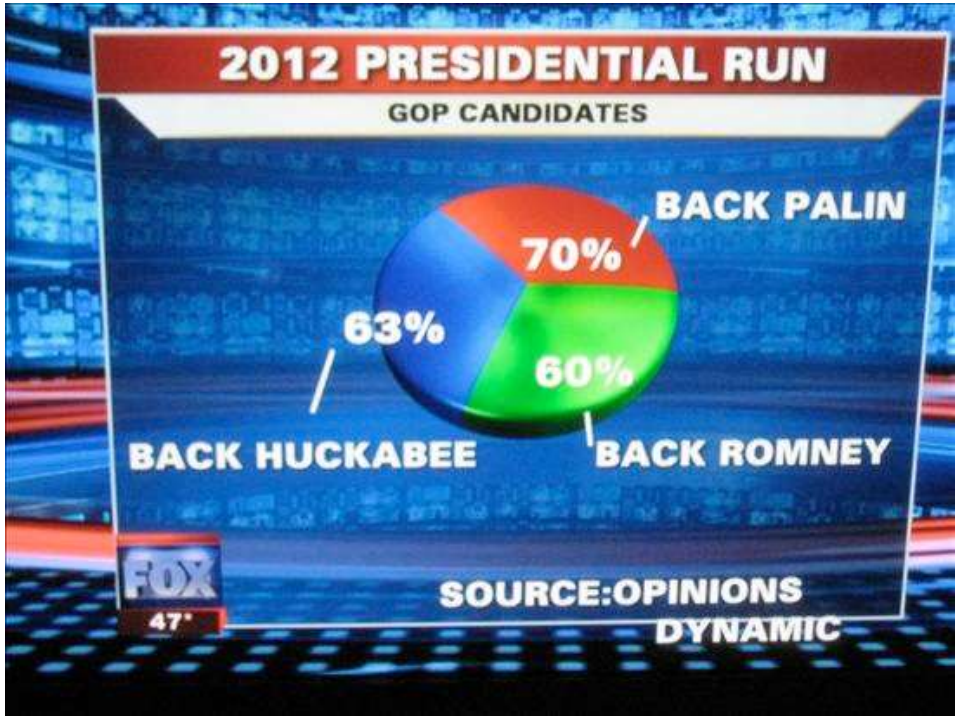


F. Nightingale, 1856

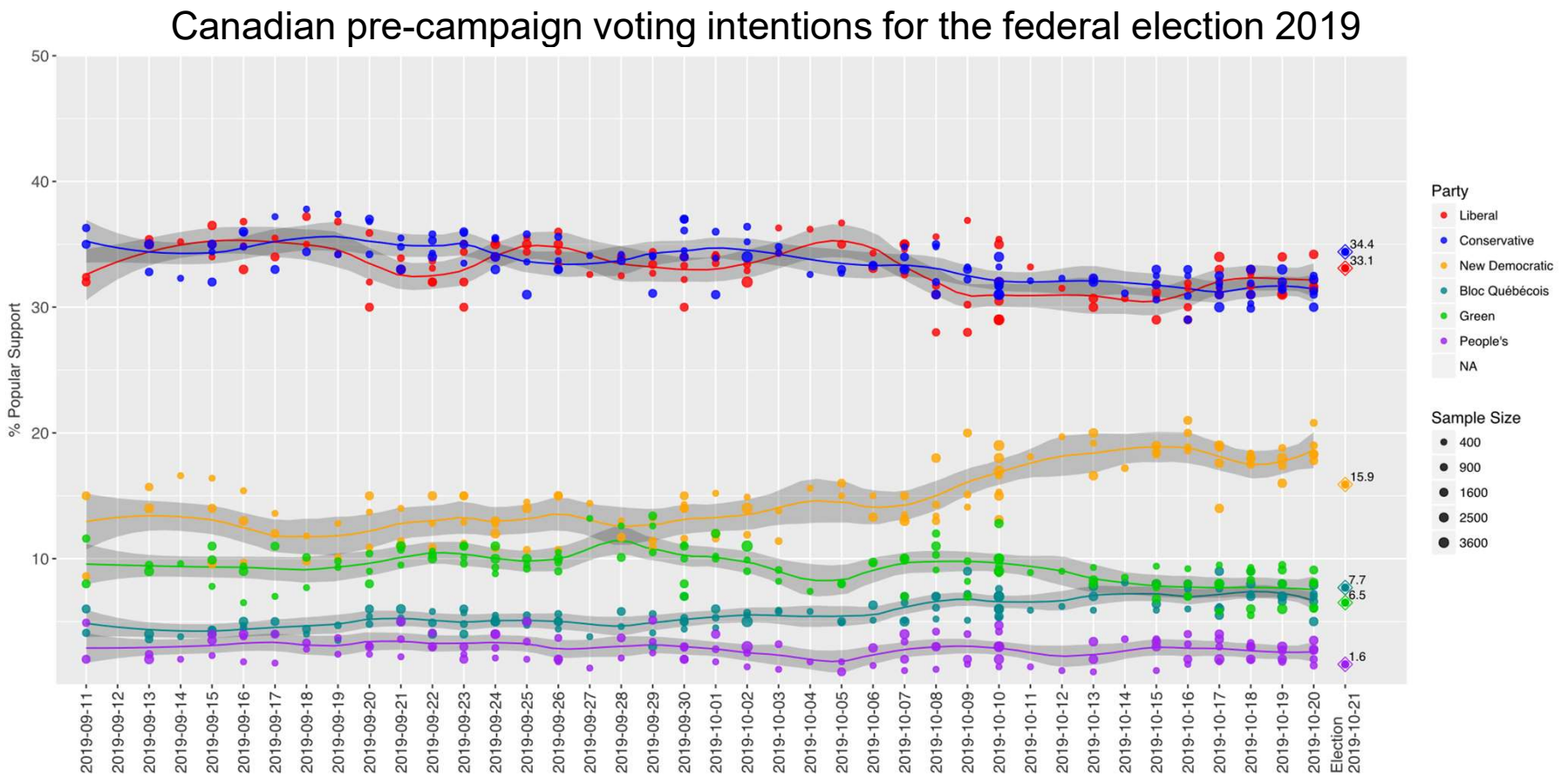
Abstract



W. Playfair, 1801

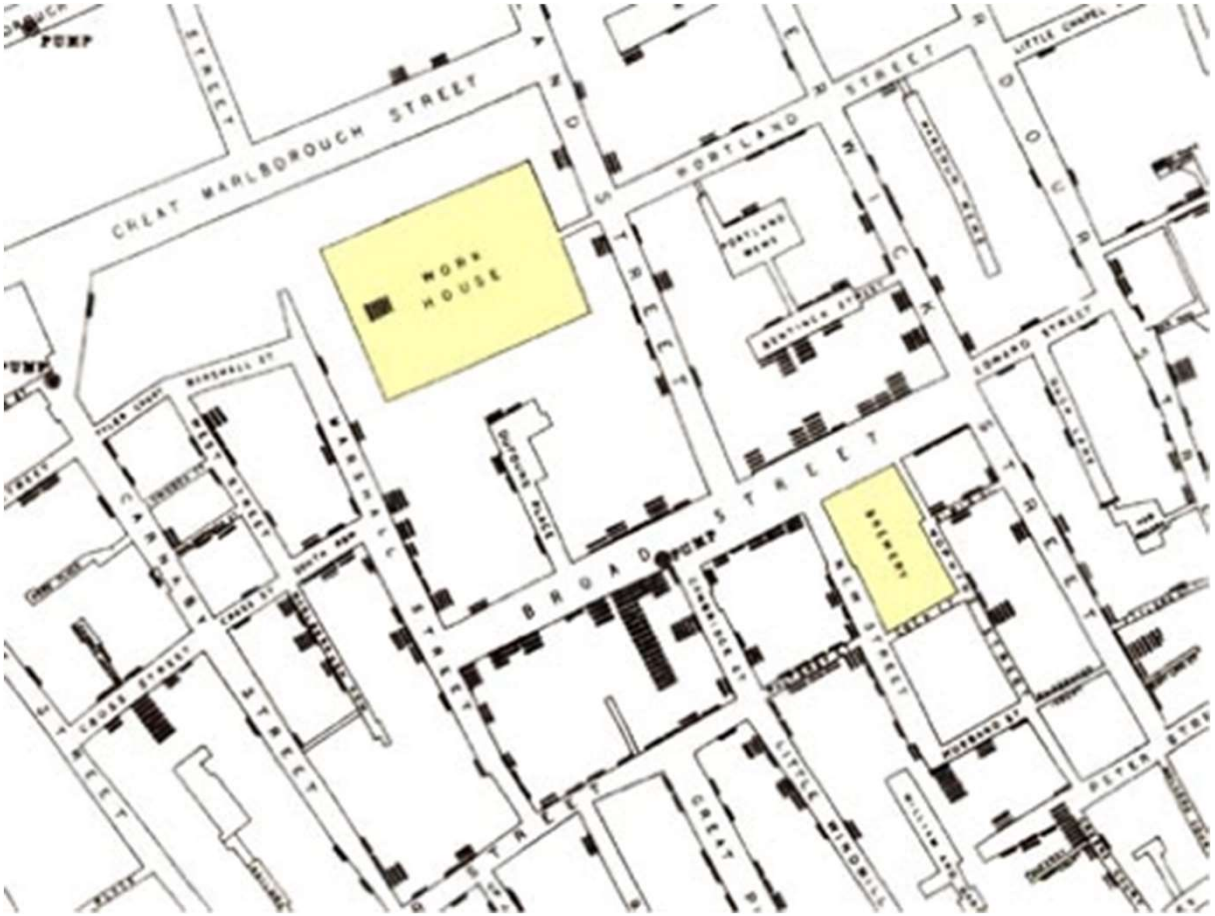


Abstract



Source wikipedia.org
Code available at https://github.com/tylerecouture/wikiplot/blob/master/canadian-federal-polls-pre43rd.R

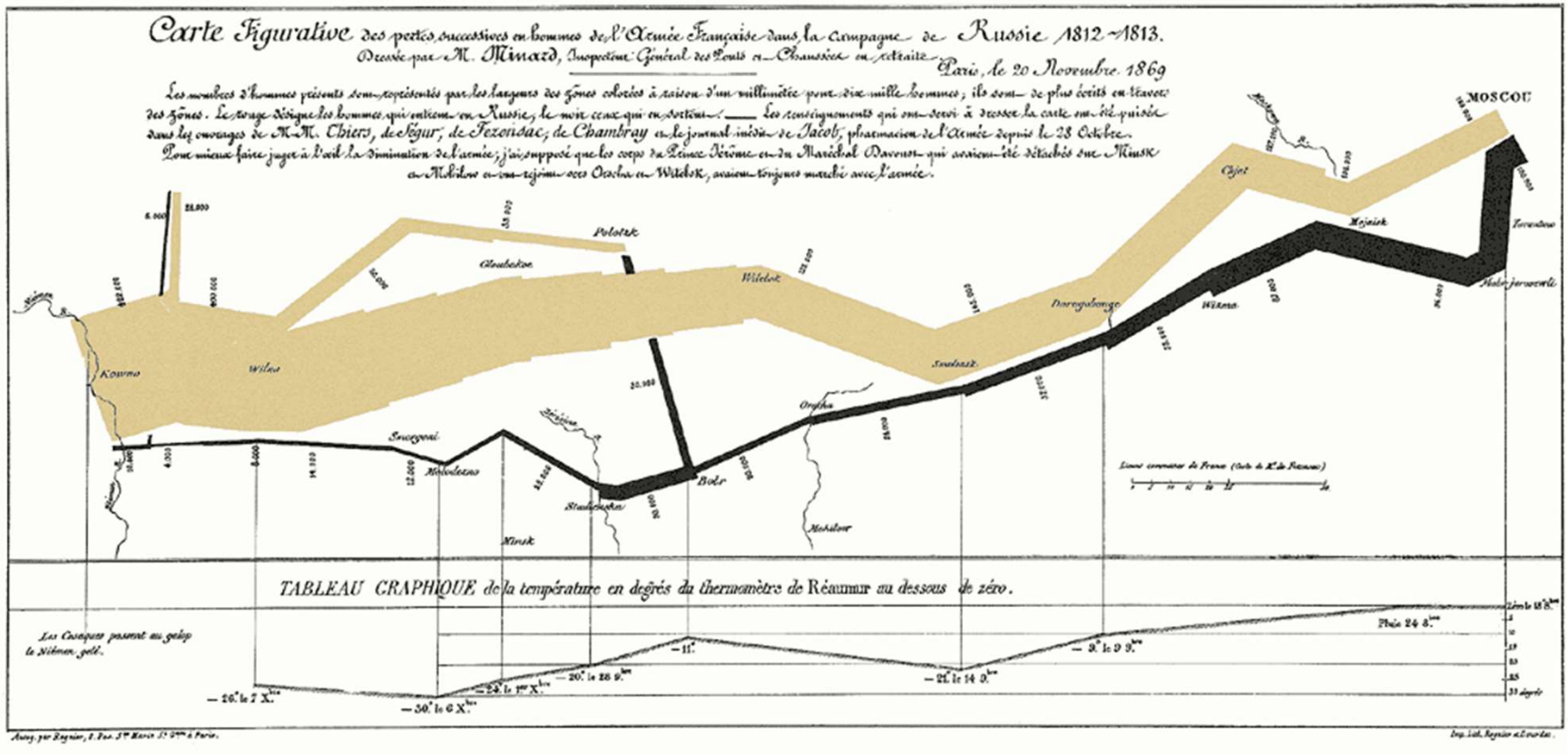
Discover



John Snow, 1854

E. Tufte, Visual Explanations, 1997

Discover

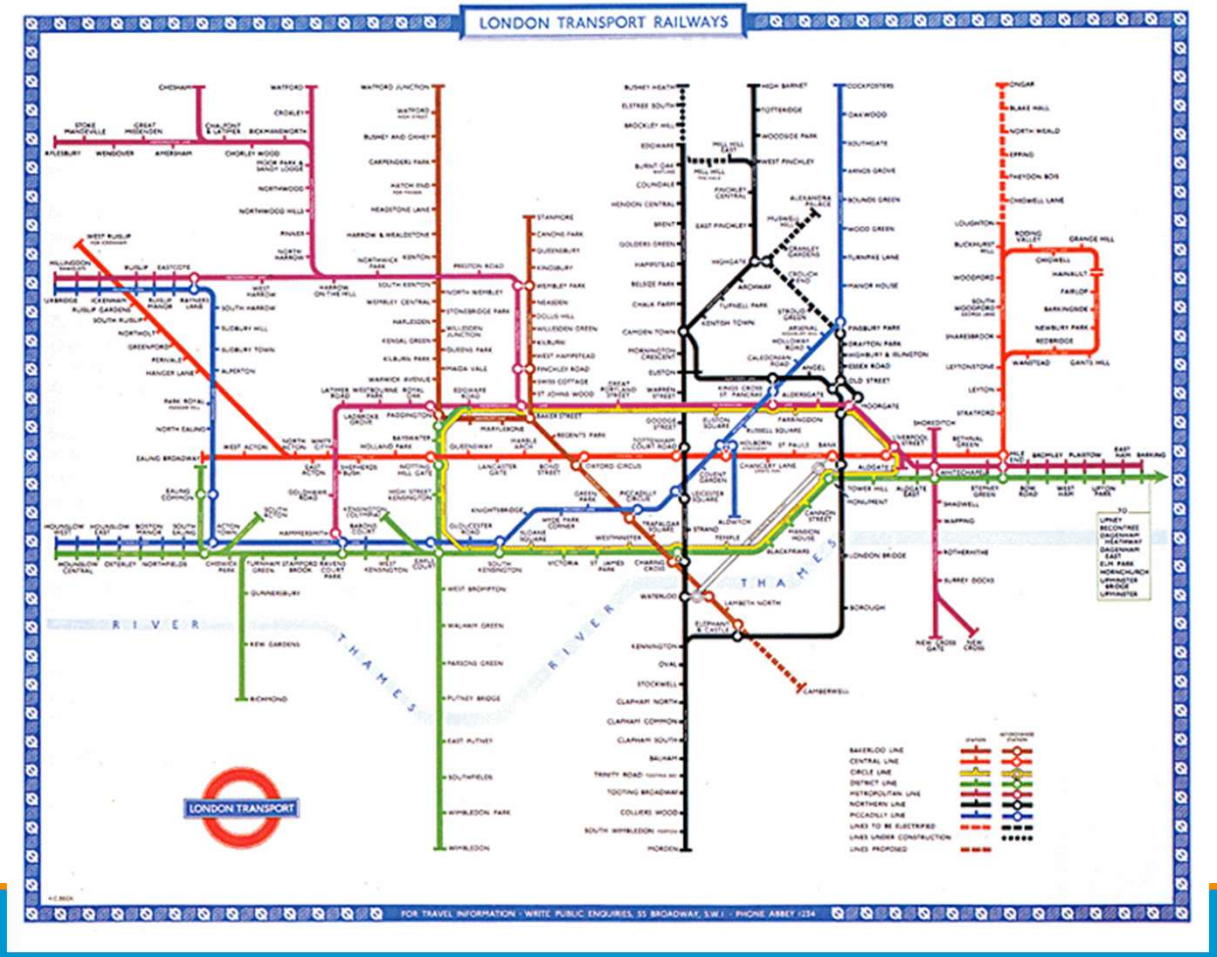


C.J. Minard, 1869

Clarify

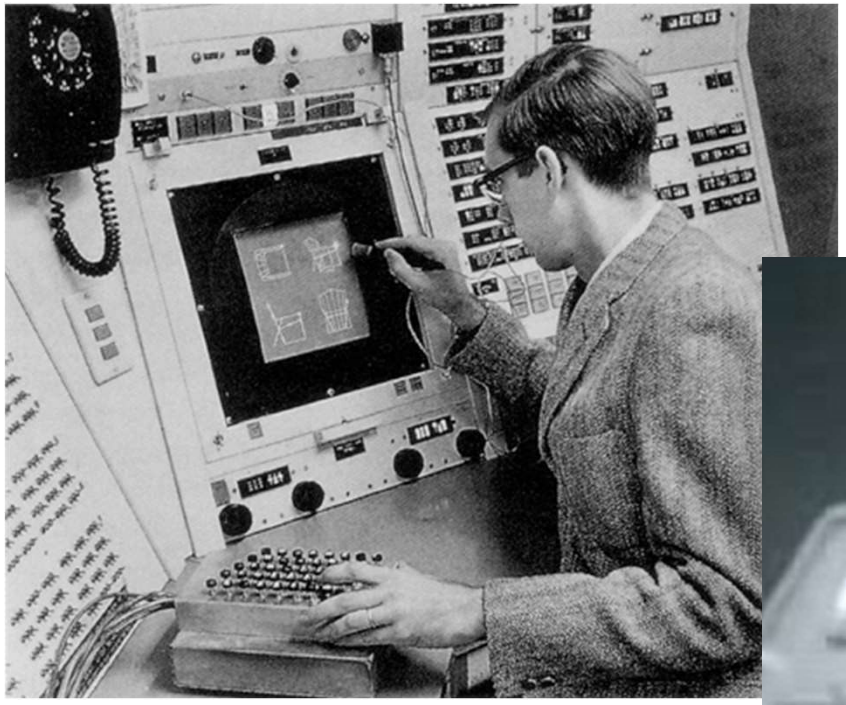
London Subway Map, 1927

Clarify



Harry Beck, 1933

Interact

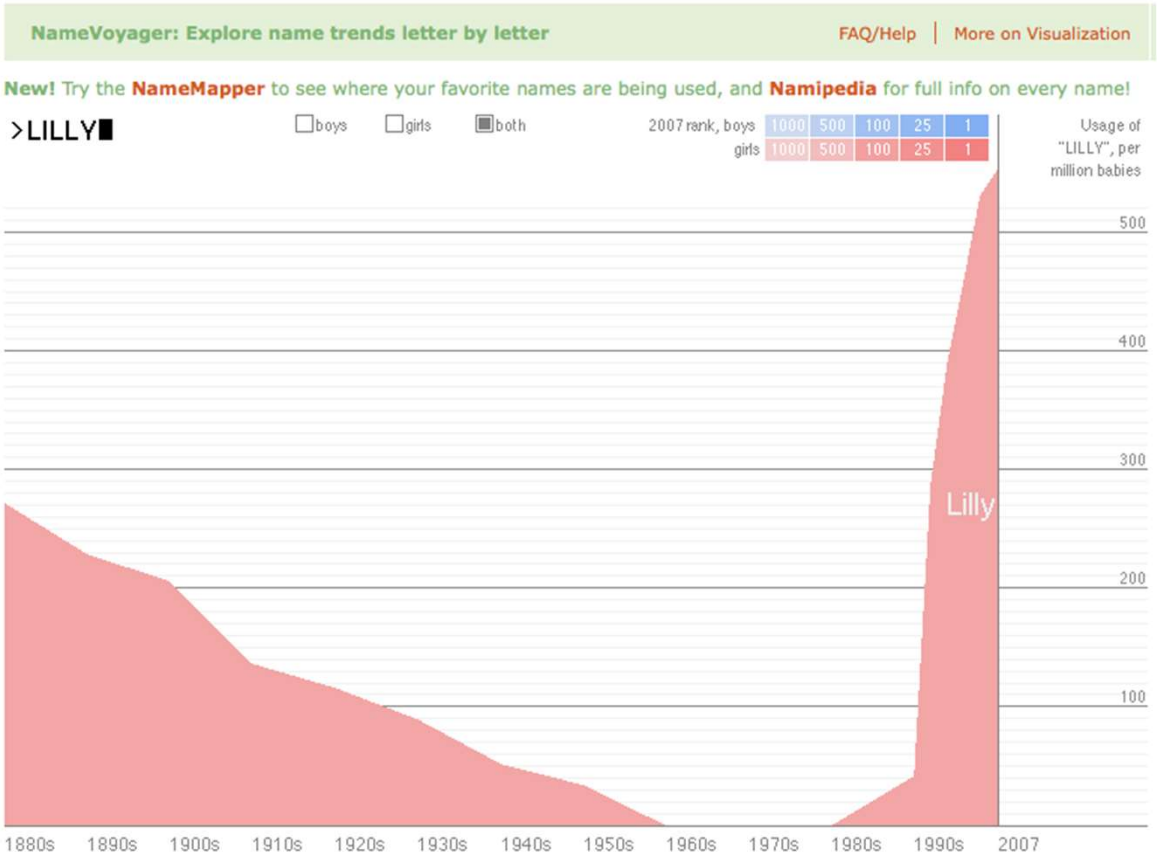


Ivan Sutherland, Sketchpad, 1963

Doug Engelbart, 1968

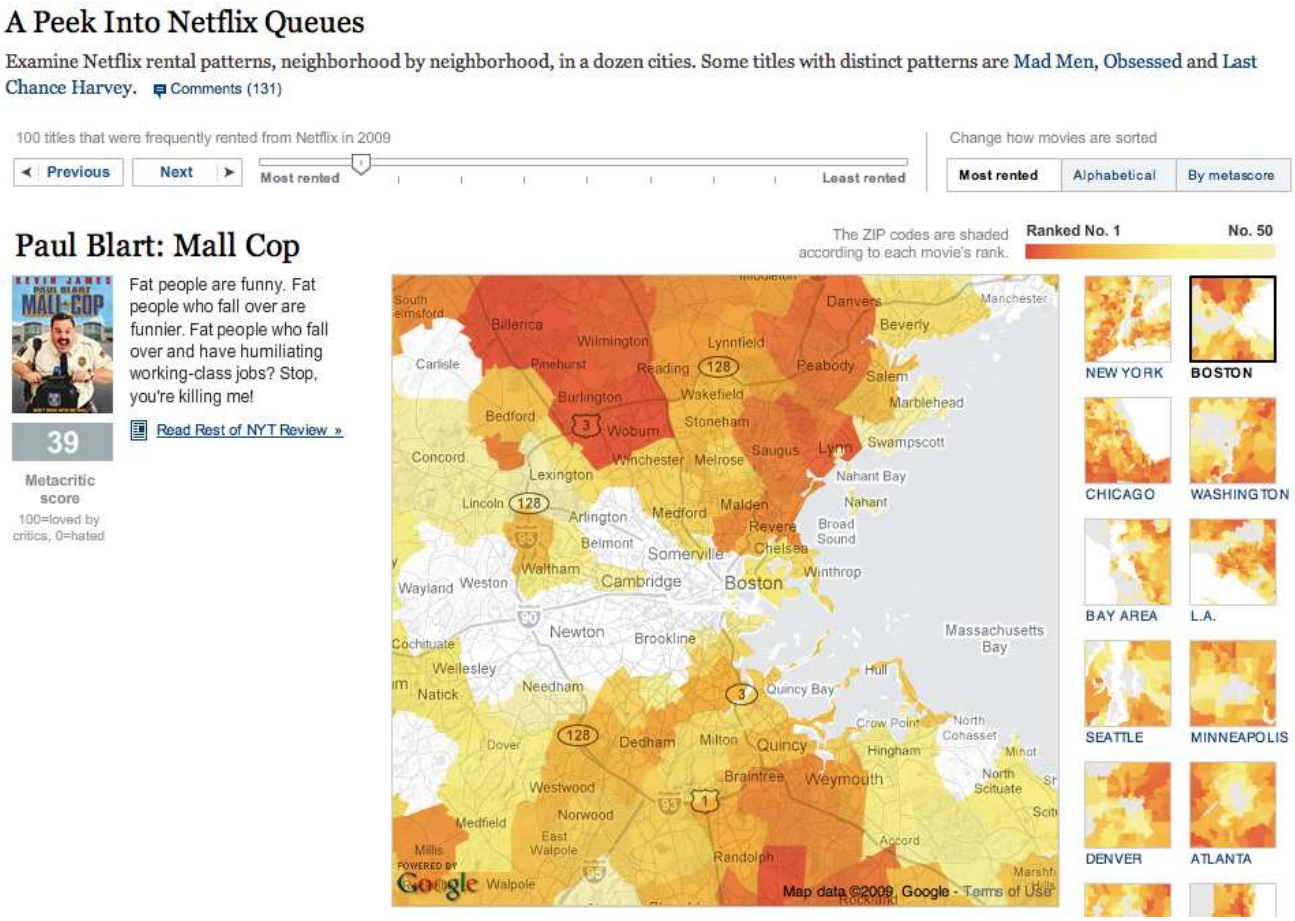
[play Engelbart.mov]

Interact



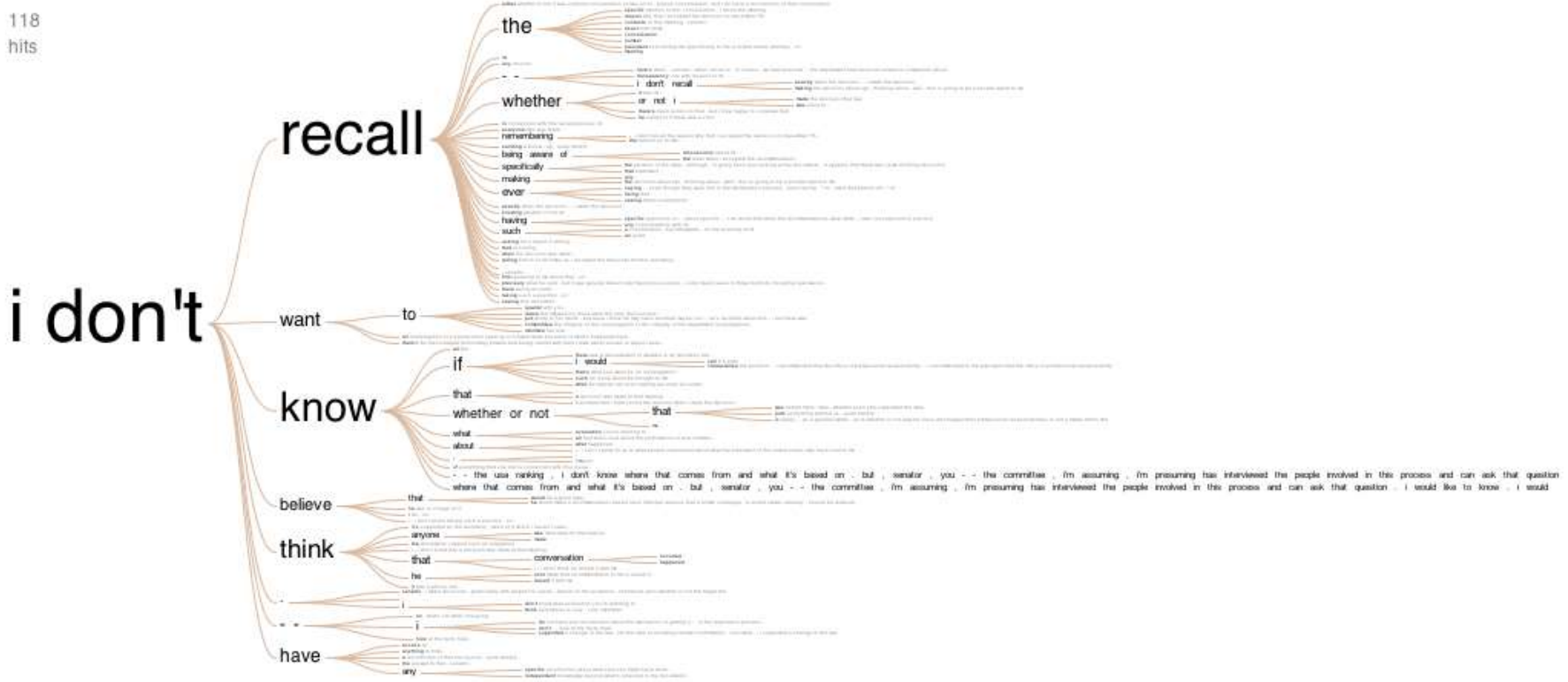
M. Wattenberg, 2005

Interact



NY Times

Communicate



"Many Eyes", M. Wattenberg 2007

Communicate

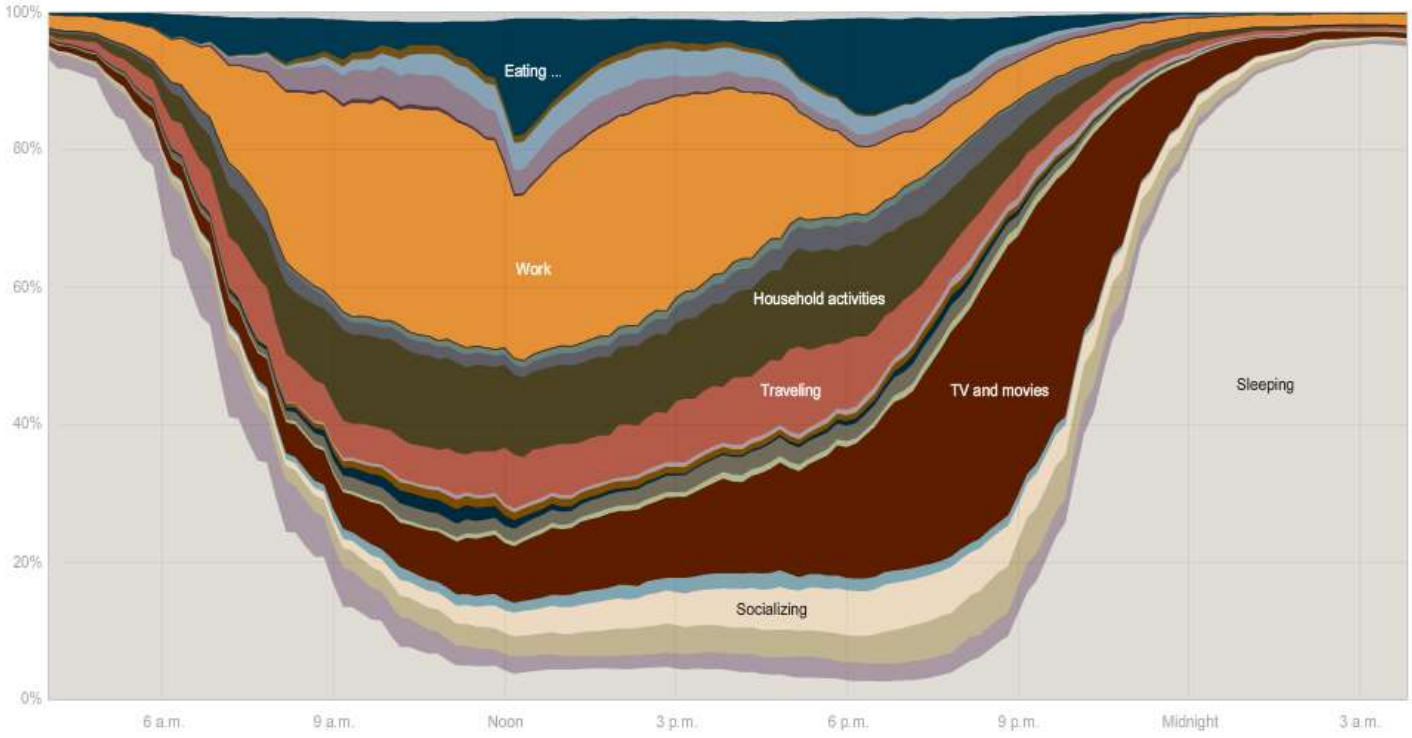
How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. Related article

Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab...	Hispanic	Age 65+	Advanced	Two+ children



NY Times


Formation of a Spiral Galaxy
DIRECTOR Takaaki Takeda
AFFILIATION 4D2U Project

Inspire / Tell a Story

Hans Rosling, TED 2006

Visualization

- To convey information through visual representations

  Map
  Record
  Abstract
  Discover
  Clarify
  Interact
  Communicate
  Inspire

Goals

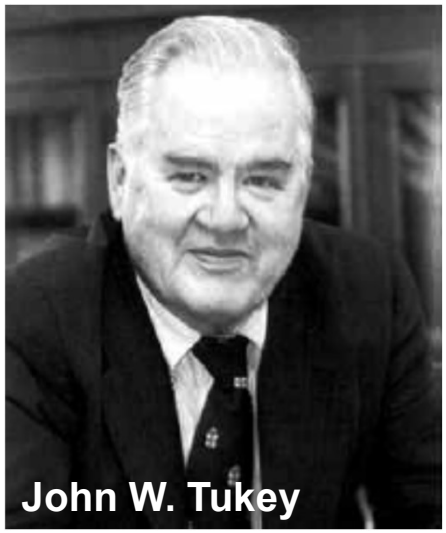
- Insight and analysis
  - Extract the information content
  - Make things and relationships visible
  - Analyze the data by means of the visual representation
- Communication
  - Allow the non–expert to understand
  - Guide the expert into the right direction
- Exploration
  - Interactive control
  - Use visual representation to understand the phenomena

- “The purpose of computing is insight not numbers” (Hamming 1962)

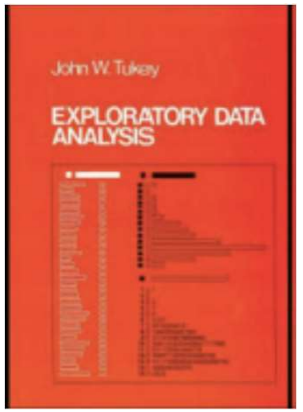
Exploratory Data Analysis (EDA)

EDA is the process of doing Descriptive Statistics

- Aim to understand the data
- Data summarization, visualization, etc.



- Professor at Princeton University
- Founding chairman of the Princeton statistics department in 1965
- Worked on EDA at Bell Labs since 60's
- Wrote a book entitled "Exploratory Data Analysis" in 1977



EDA is like detective work

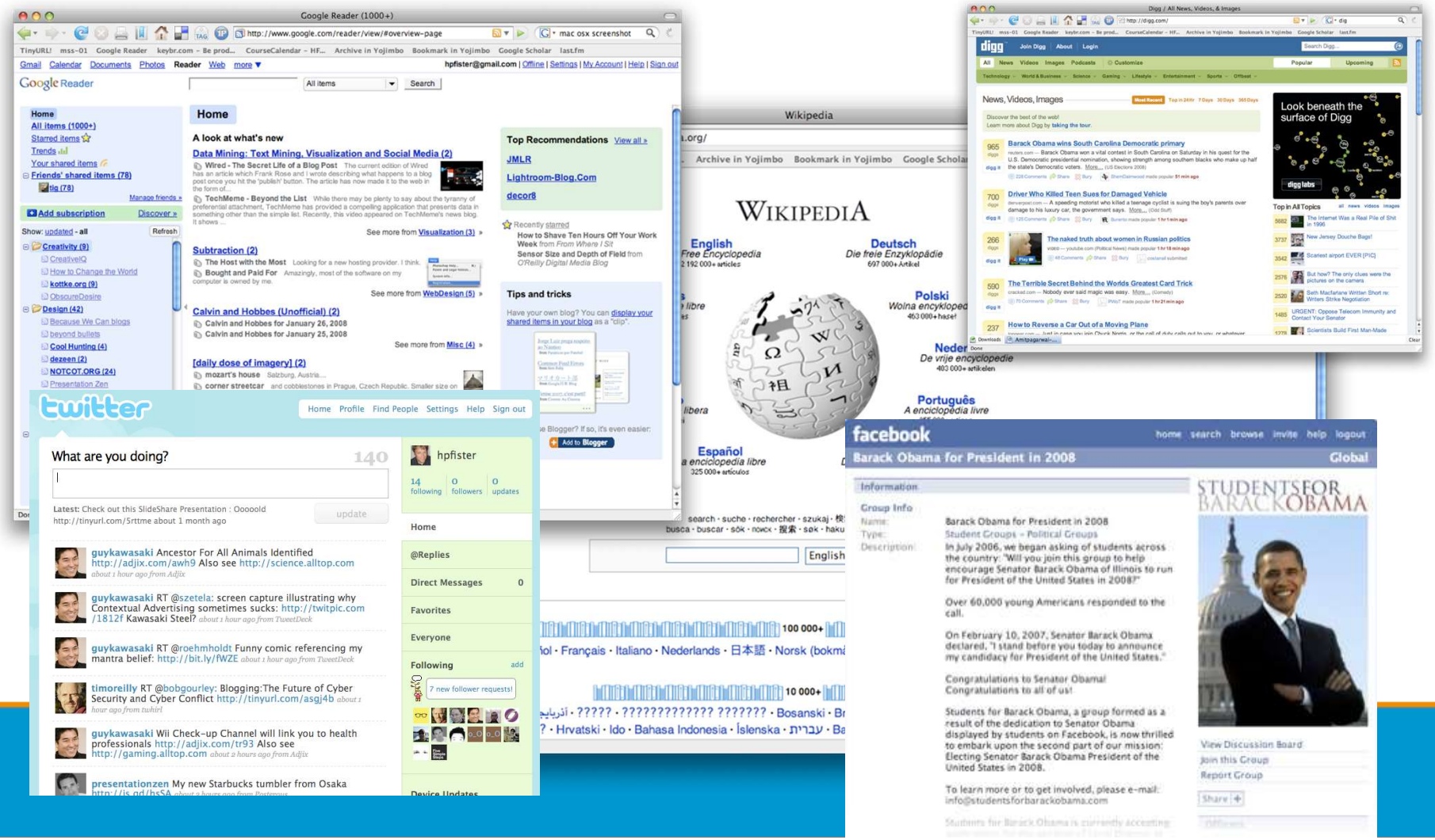
John Tukey:

"Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there."

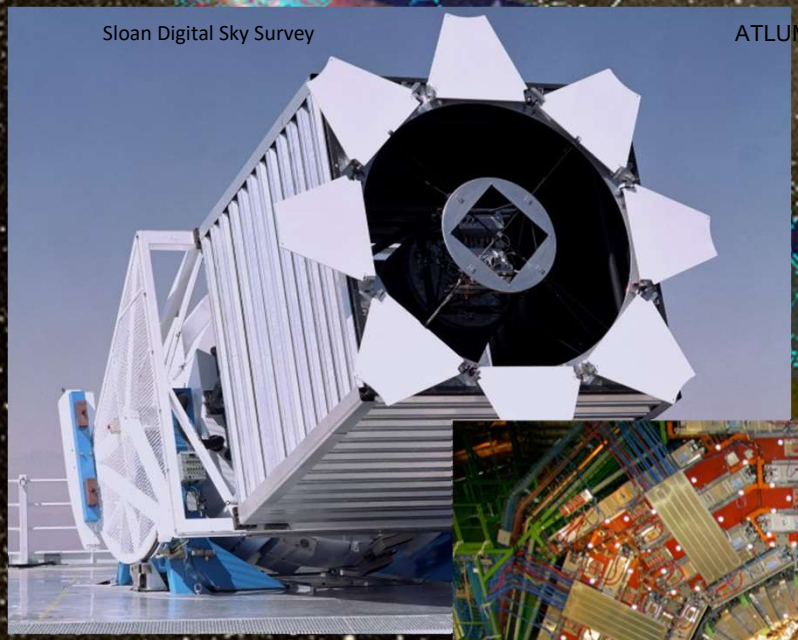
Why Data Visualization?

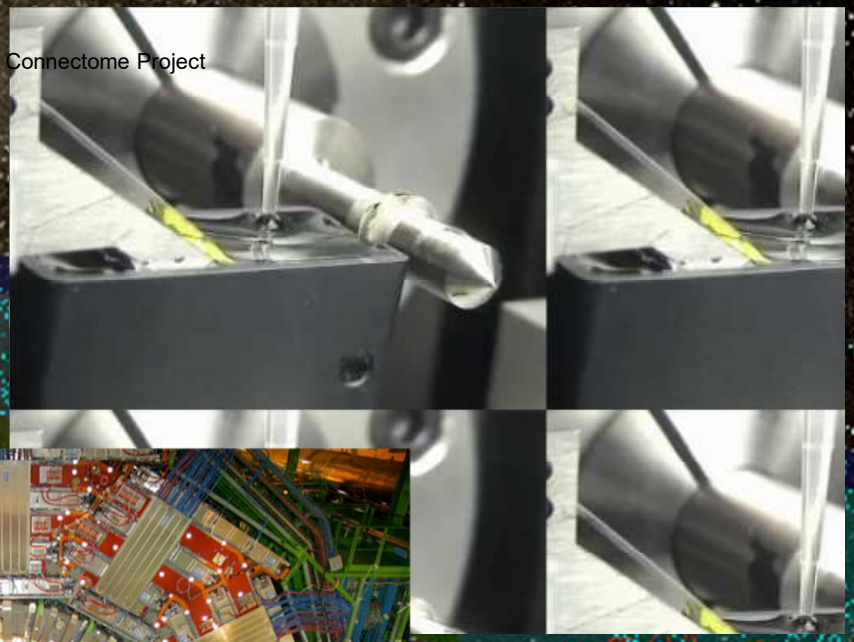
- What?
- Why?
- Who?
- How?

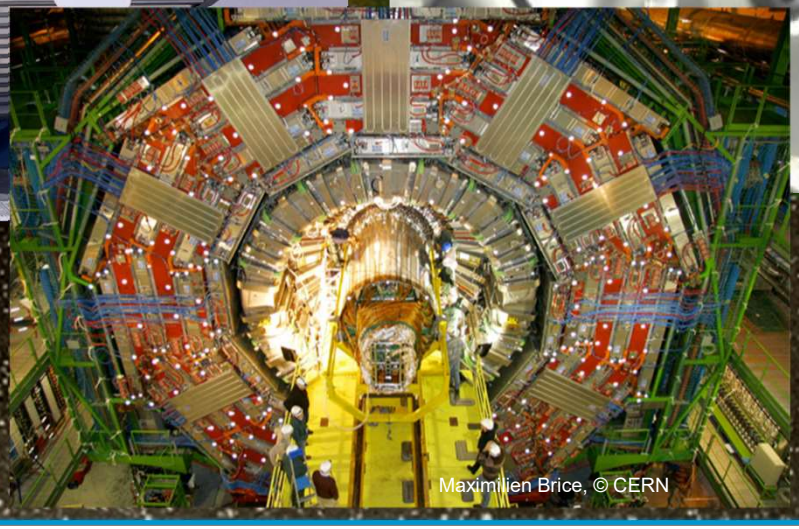
Information Explosion / Big Data



Instrument Data Explosion


Sloan Digital Sky Survey


ATLUM / Connectome Project


Maximilien Brice, © CERN

“The Industrial Revolution of Data”

Joe Hellerstein, UC Berkeley

Limits of Cognition

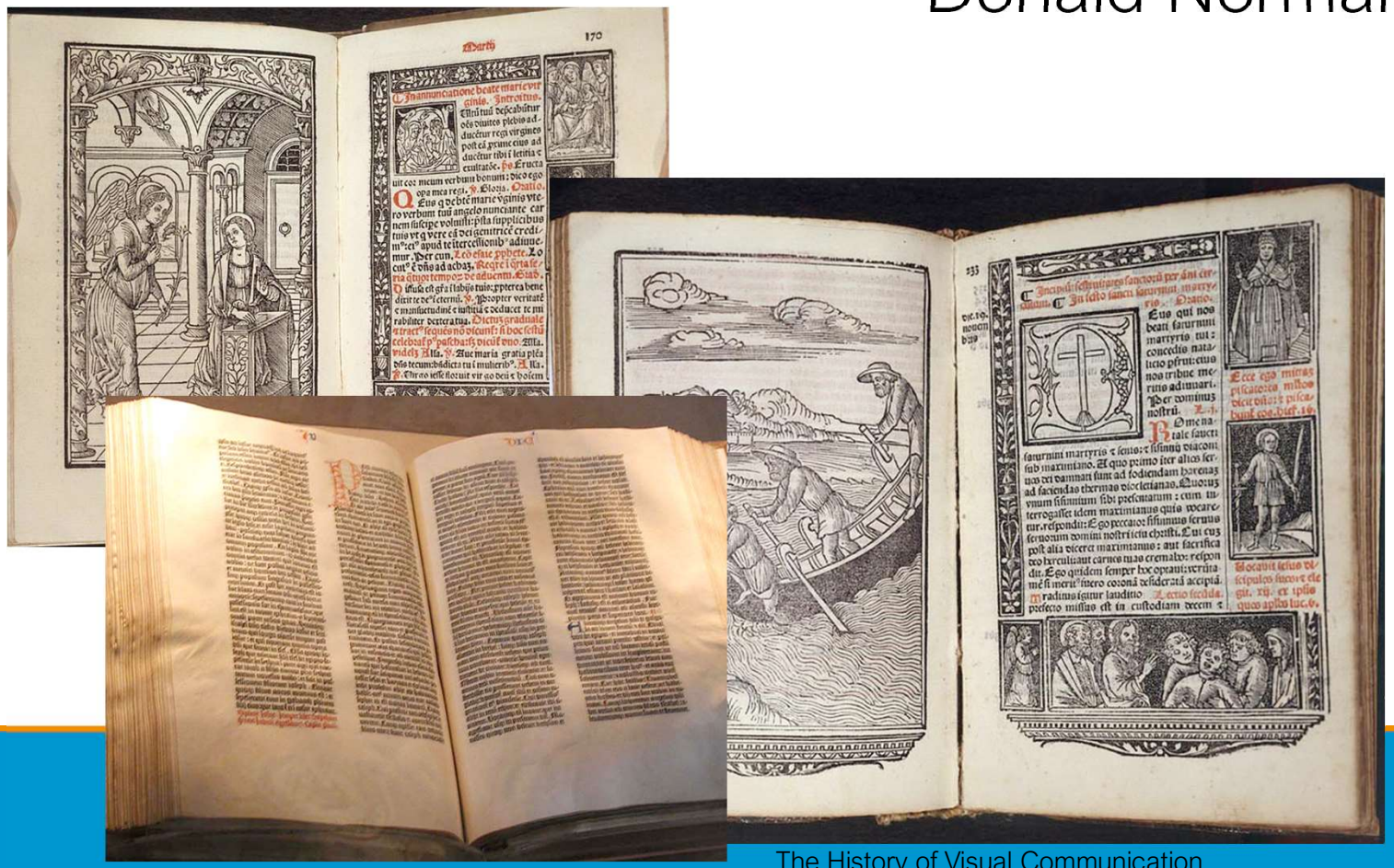
Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

“It is things that make us smart.”

Donald Norman

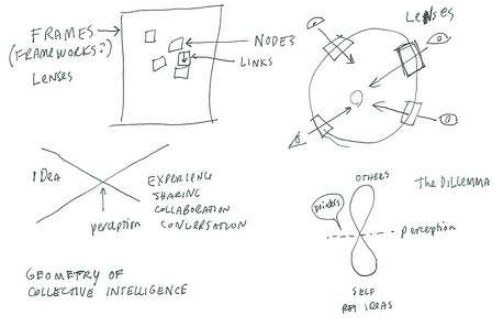
“It is things that make us smart.”

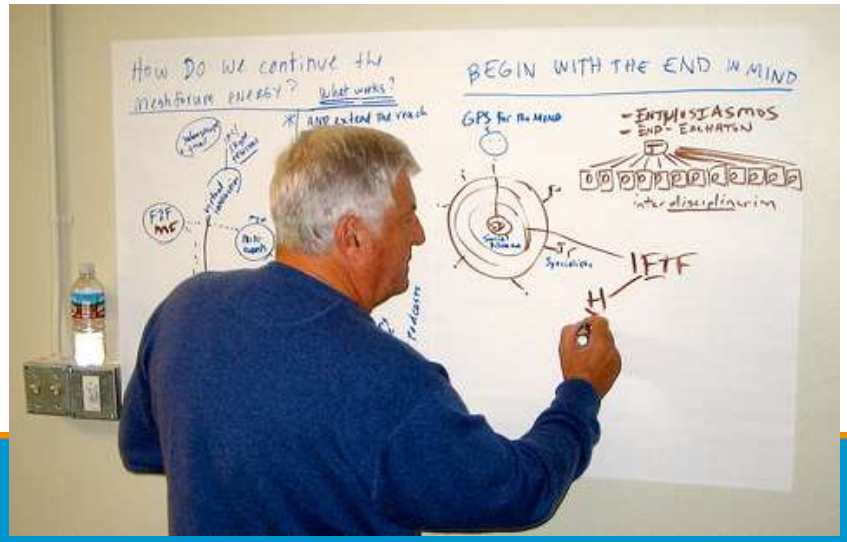
Donald Norman



“It is things that make us smart.”

Donald Norman





Visual Thinking Collection, Dave Grey

Idea Maps, by Jamie Nast

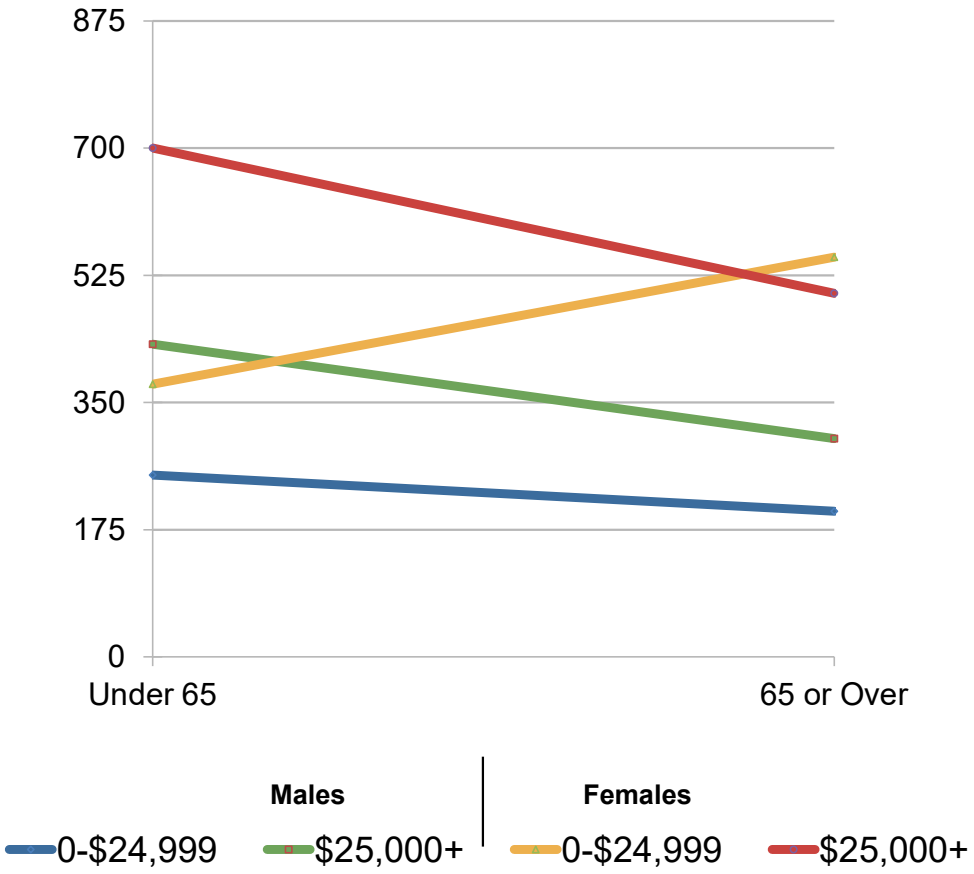
Mental Queries

Which gender or income level group shows different effects of age on triglyceride levels?

	Males		**Females**	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-$24,999	250	200	375	550
$25,000+	430	300	700	500

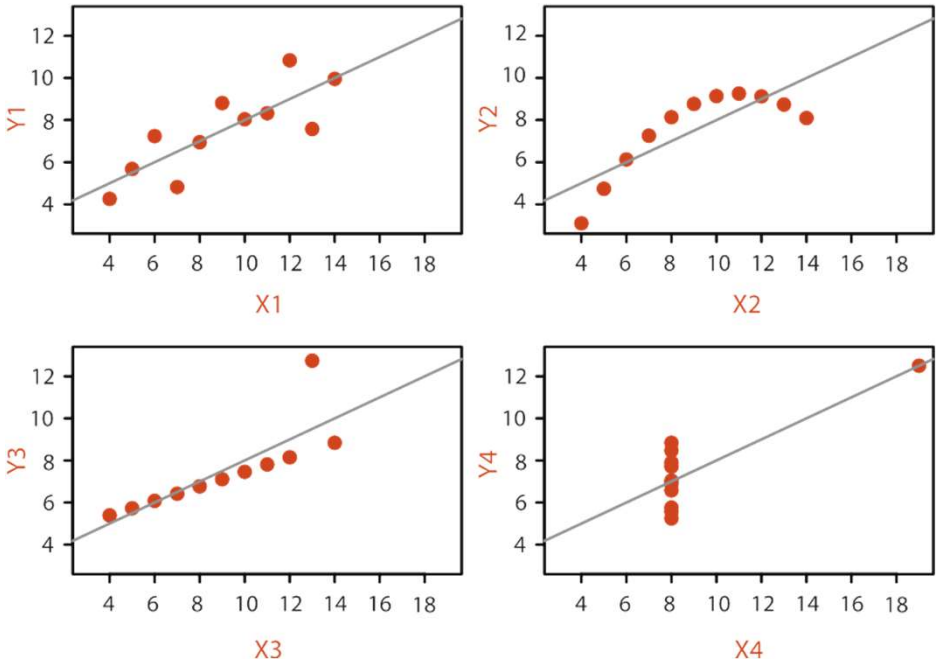
Slide after Stephen Kosslyn, Clear and to the Point

Visual Queries



Slide after Stephen Kosslyn, Clear and to the Point

Why use an external representation?

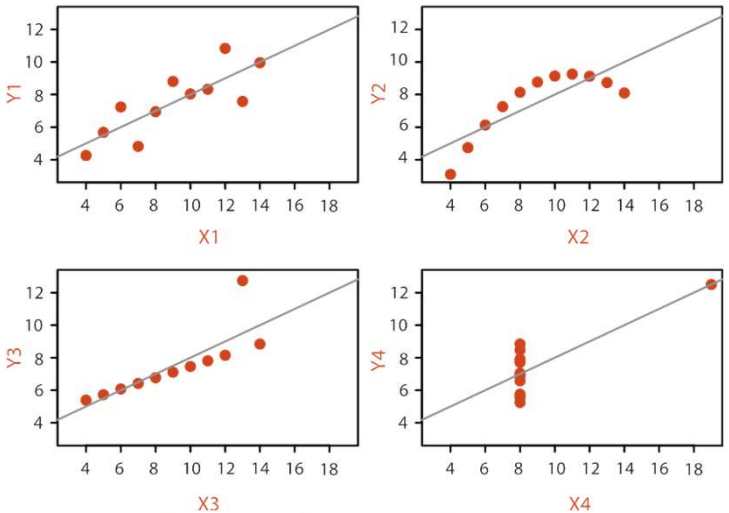


- Replace cognition with perception

	I		II		III		IV	
	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	9,13	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

[F. J. Anscombe, 1973]

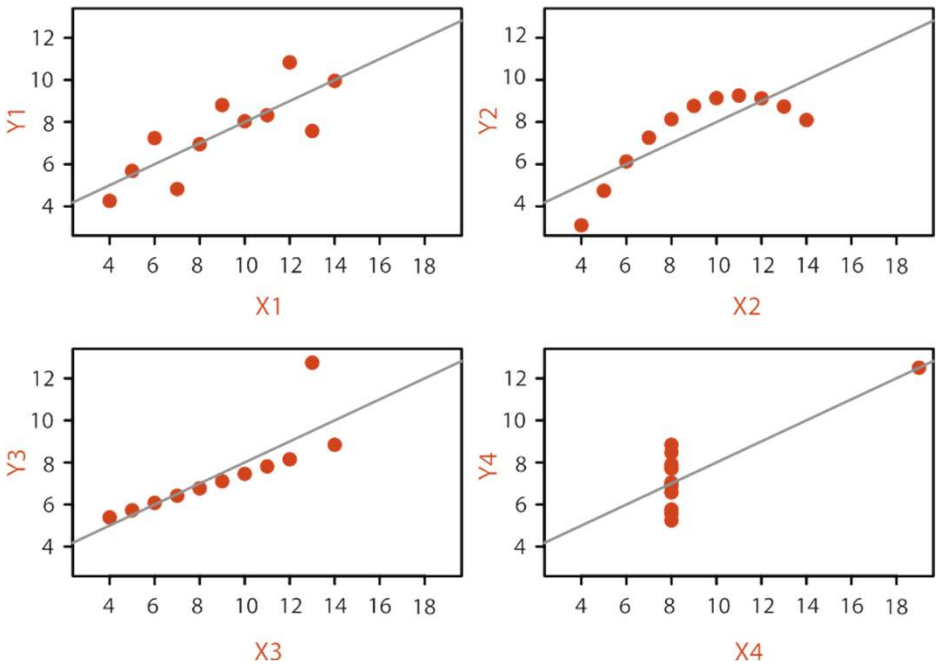
Why represent all the data?



I		II		III		IV	
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	9,13	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- Summaries lose information, details matter
  - Confirm expected and find unexpected patterns
  - Assess validity of statistical model

“Numerical calculations are exact, but Graphs are rough”



- Same relationship among each pair of variables?
- Identical statistics

X mean	9
X variance	10
Y mean	7.5
Y variance	3.75
<X,Y> correlation	0.816

[F. J. Anscombe, 1973]

Visualization

- Helps us think
- Reduces load on working memory
- Offloads cognition
- Uses the power of human perception

Defining Visualization (Vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

["Visualization Analysis and Design" by T. Munzner, 2014]

Why have a human in the loop?

- Not needed when automatic solution is trusted
- Good for ill-specified anlaysis problems
  - Common setting: "What questions can we ask?"

Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Munzner, T. (2014)

Long-term use

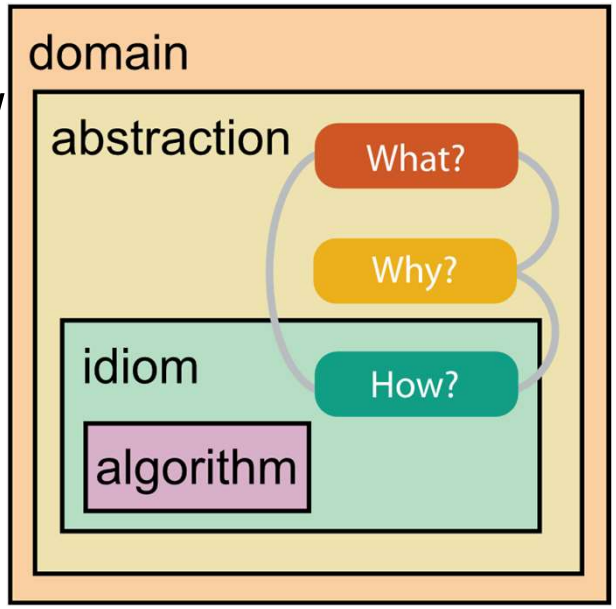
- Exploratory analysis of scientific data
- Presentation of known results

Short-term use

- For **developers** of automatic solutions:
  - Understand requirements for model development
  - Refine/debug and determine parameters
- For **end users** of automatic solutions: verify, build trust

Analysis framework: four *levels*

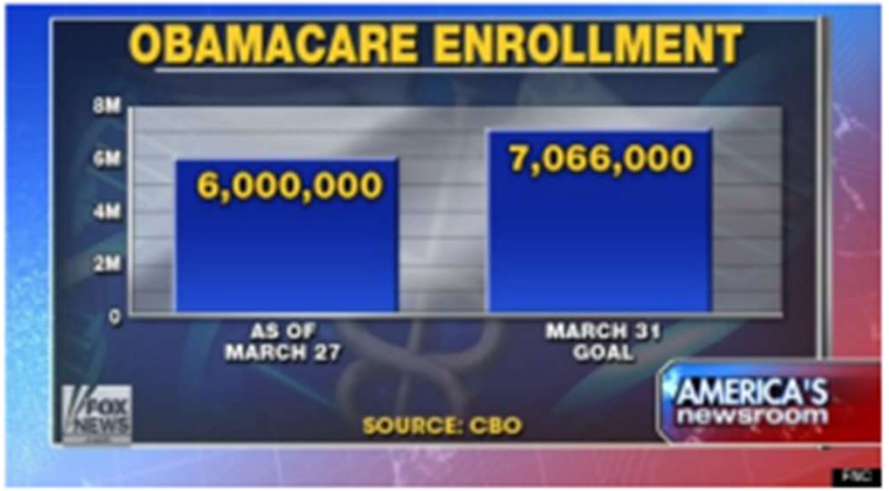
- ***Domain*** situation: Who are the target users?
- ***Abstraction***: Translate from specifics of domain to vocabulary of vis
- **What** is shown? *Data abstraction*
  - Don't just draw what you're given: transform to new
- **Why** is the user looking at it? *Task abstraction*
- **How** is it shown? ***Idiom***
  - Visual encoding idiom: How to draw
  - Interaction idiom: How to manipulate
- ***Algorithm***: efficient computation



[A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Pitfalls

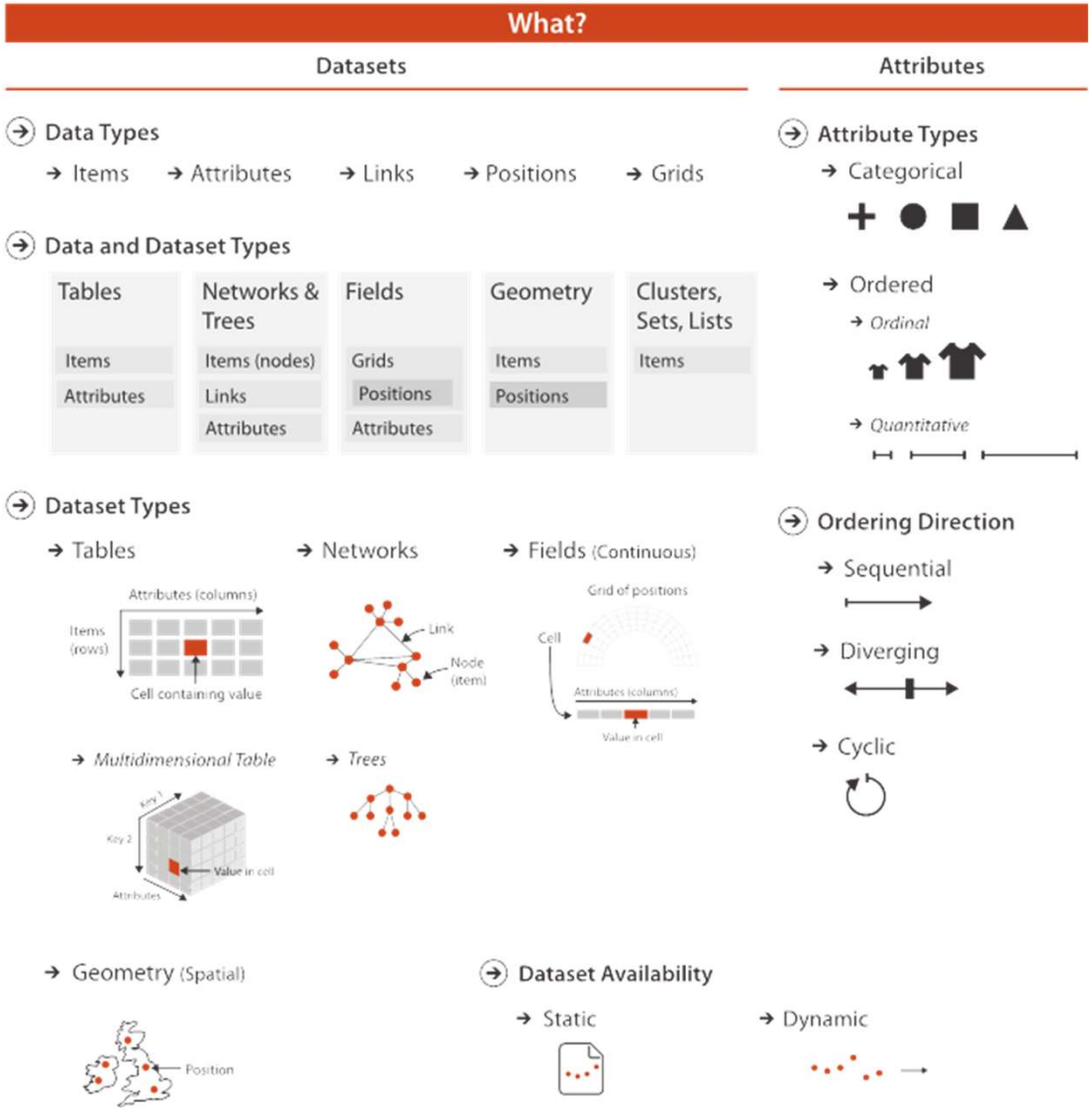
- WTF Visualizations (http://viz.wtf)
- Without **knowing the principles**, you might make a lot of mistakes like this!



Resource limitations

- **Computational** limits
  - Processing time and system memory
- **Human** limits
  - Human attention and memory
  - Understanding abstractions
- **Display** limits
  - Pixels are precious
  - Information density tradeoff: Info encoding vs unused whitespace

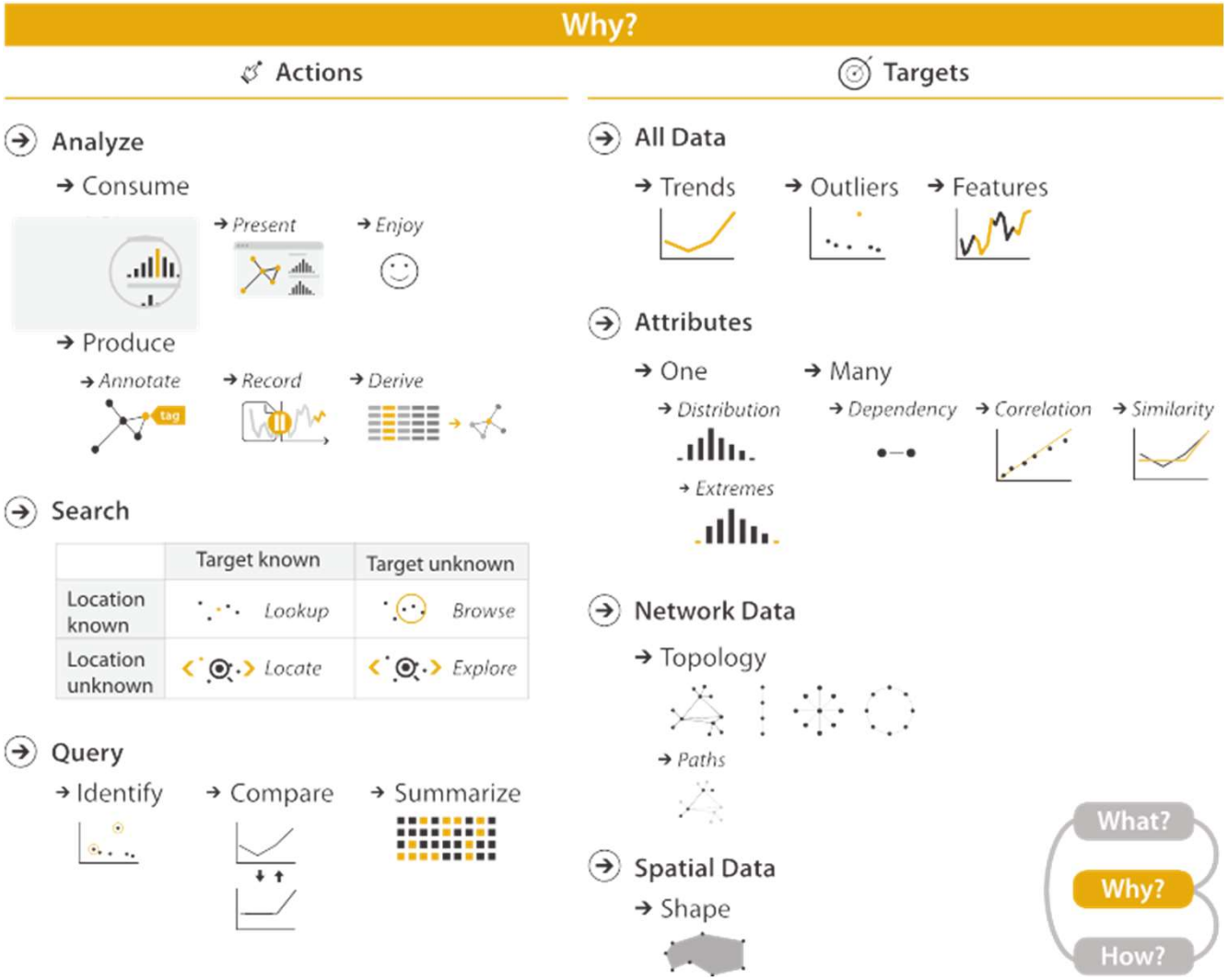
Understand
Data, Task, and Encoding



Data Types

- Items and attributes as rows and columns of tables
- Position and time are special attributes
- Spatial data on grids makes computation easier

[T. Munzner, 2014]



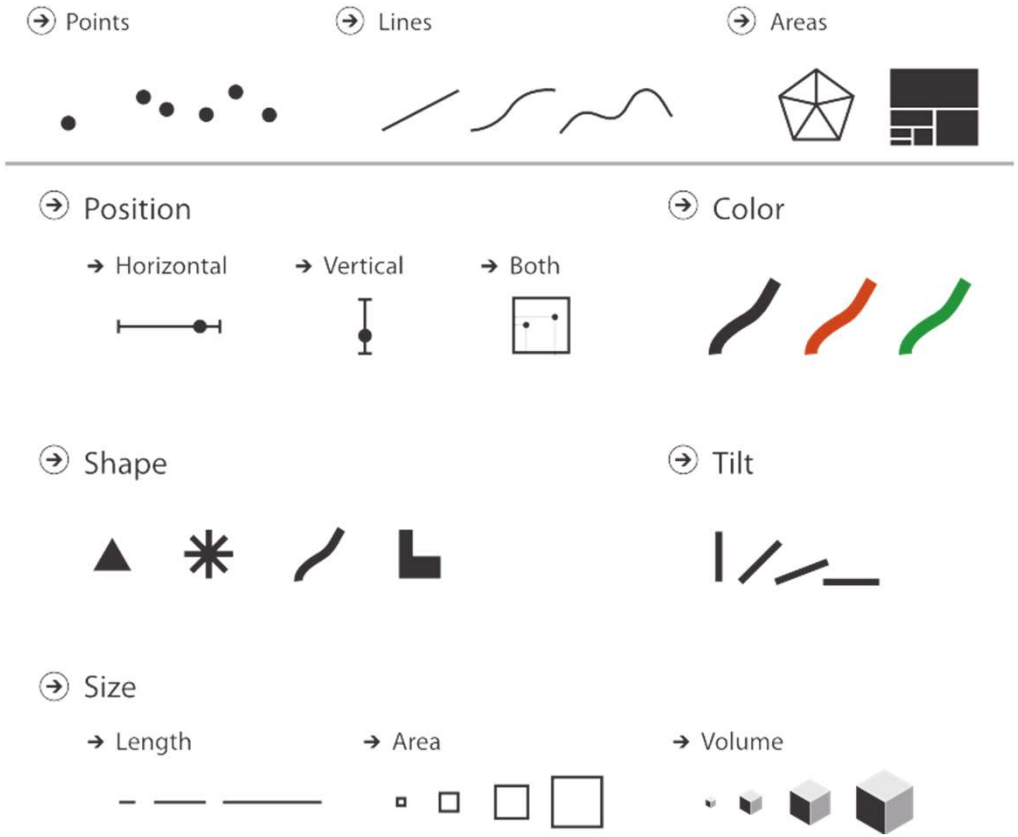
Tasks

- Actions
  - Analyze
  - Search
  - Query
- Targets
  - Item & Attributes
  - Topology & Shape

[T. Munzner, 2014]

Visual Encoding – How?

- Marks
  - Geometric primitives
- Channels
  - Appearance of marks
  - Redundant coding with multiple channels possible

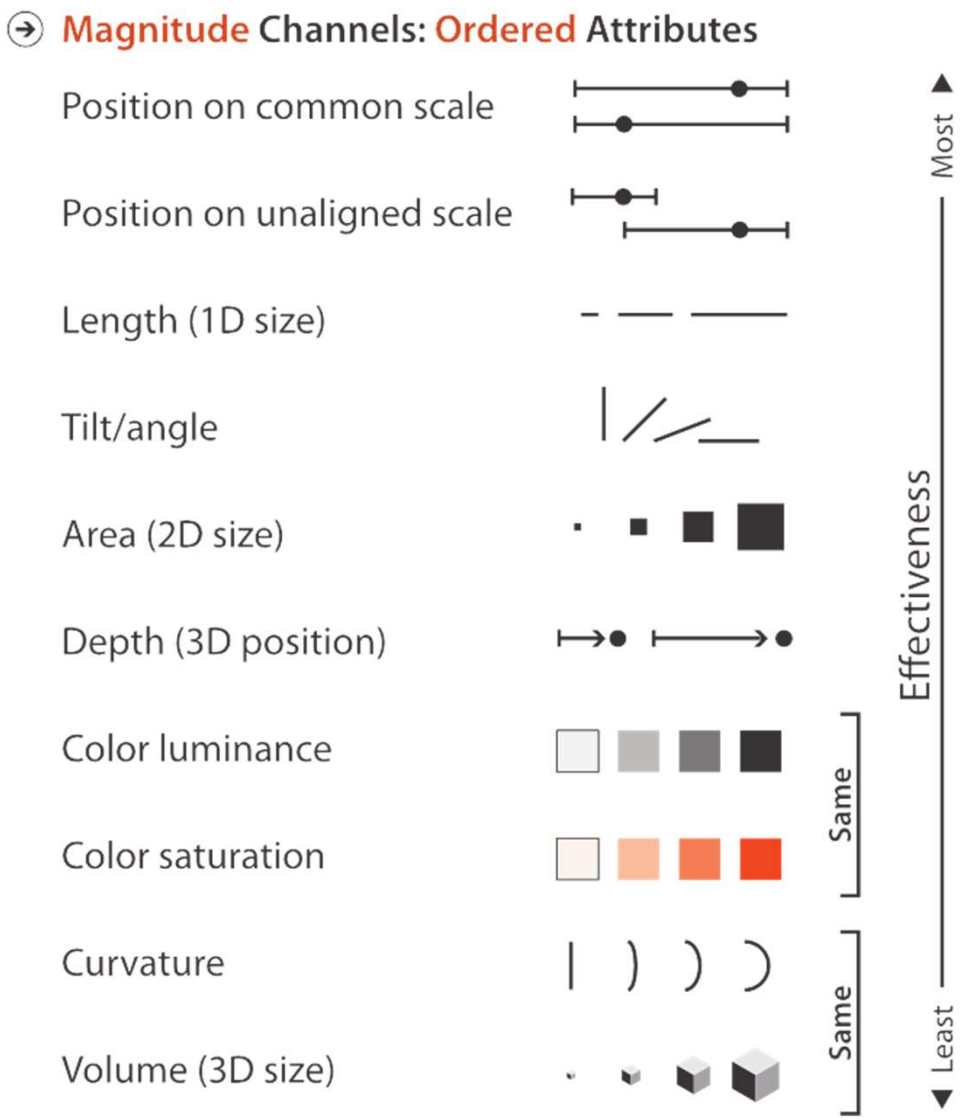


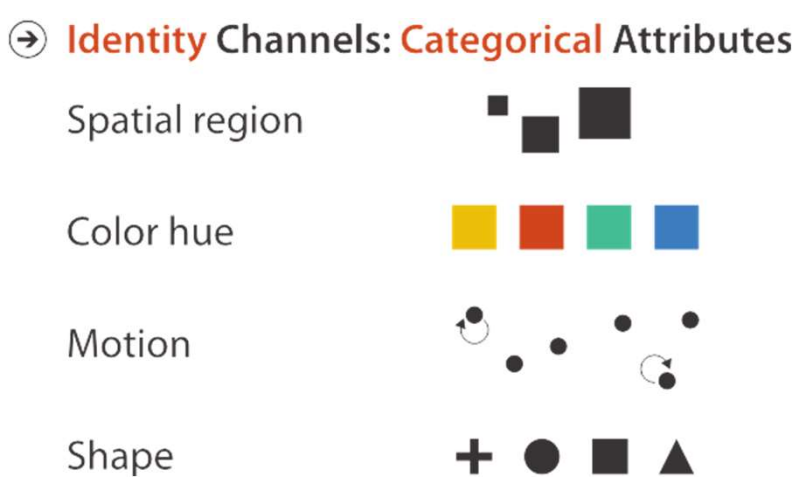
[T. Munzner, 2014]

Design Principles for Task Effective Visualization

Task and effectiveness

- Most idioms ineffective for particular task/data
  - Recast tasks from domain-specific vocabulary to abstract form
  - Systematic thinking about choices imposes structure on design space
  - Analyze existing as step to design new – iterate and compare
- What counts as effective?
  - Novel: enable entirely new kinds of analysis
  - Faster: speed up existing workflows





Expressiveness principle

- **Match channel and data characteristics**

Effectiveness principle

- **Encode important attributes with higher ranked channels**

[T. Munzner, 2014]

Chart Design: Simplifying

Example from Tim Bray

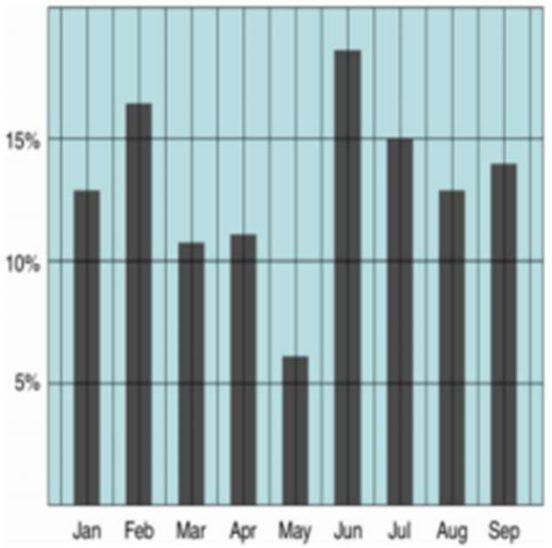


Chart Design: Simplifying

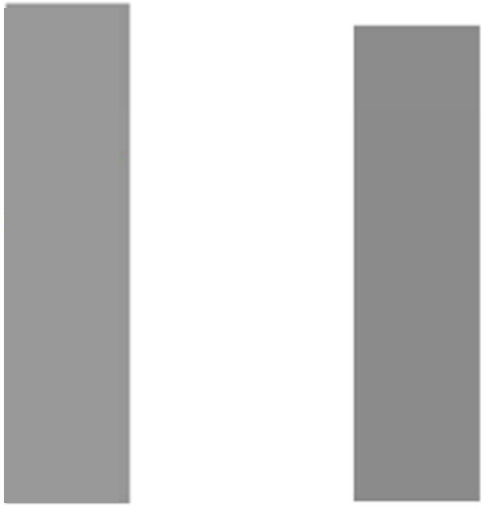
Example from Tim Bray

Principle 2: Understand Magnitudes

Which one is brighter?

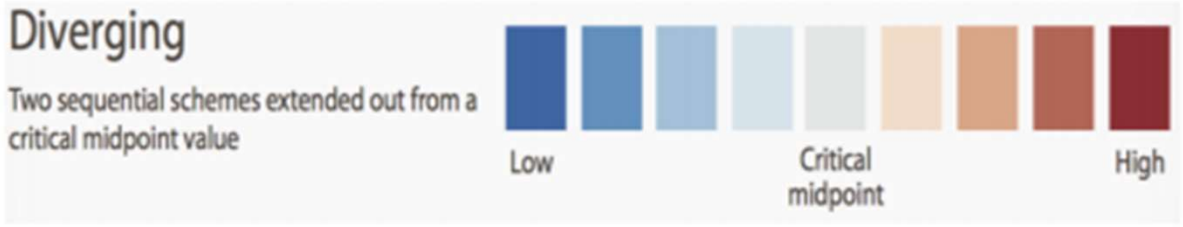
Principle 2: Understand Magnitudes

Which one is longer?



Principle 3: Use Color

- **Make your visualization look beautiful**
  - Colour Lovers: http://www.colourlovers.com
- **Work for different kinds of data**

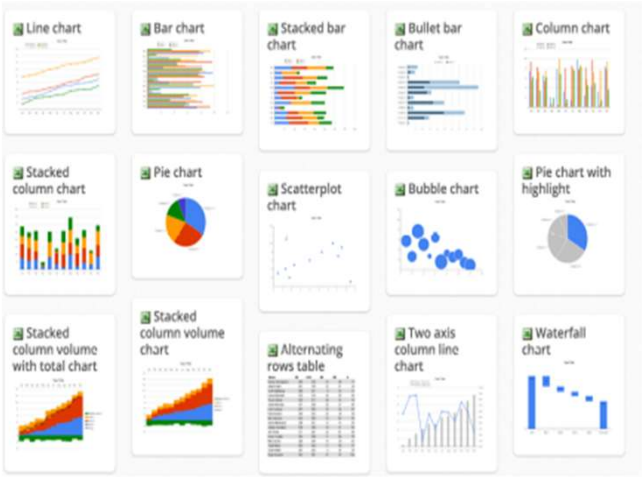


Colormaps

- Choose a color map appropriate for the data type
  - Categorical, ordinal, quantitative
  - Sequential, perceptually uniform
  - Diverging, cyclic, qualitative
- See https://matplotlib.org/3.1.0/tutorials/colors/colormaps.html

Principle 4: Use Structure

- Chart chooser: http://labs.juiceanalytics.com

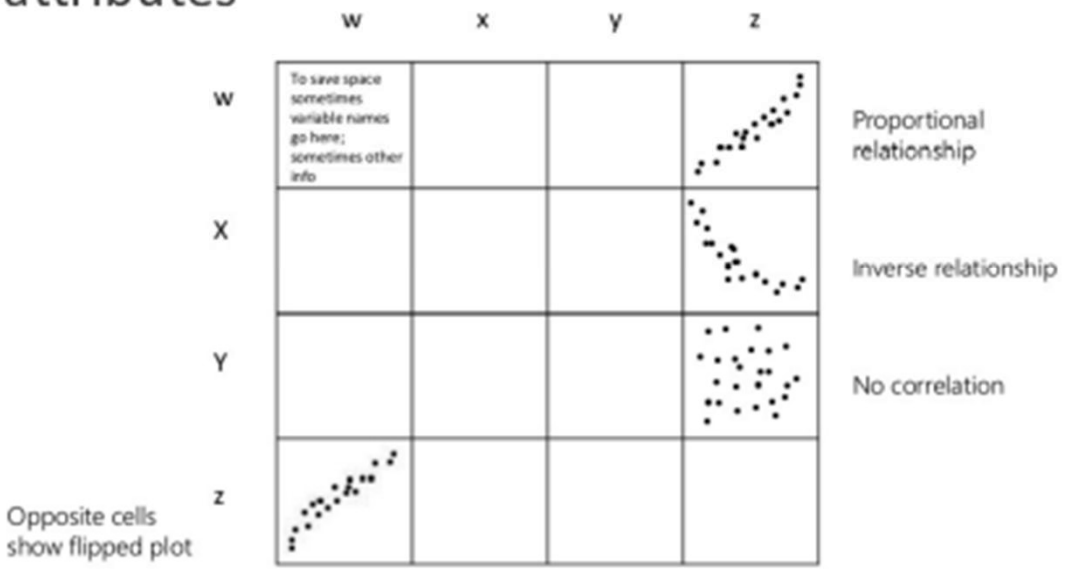


- Galleries (modify for your purposes)
  - https://**seaborn**.pydata.org/examples/index.html
  - https://observablehq.com/@**d3**/gallery
  - https://**altair-viz**.github.io/gallery/

Principle 4: Use Structure

Correlation Visualization

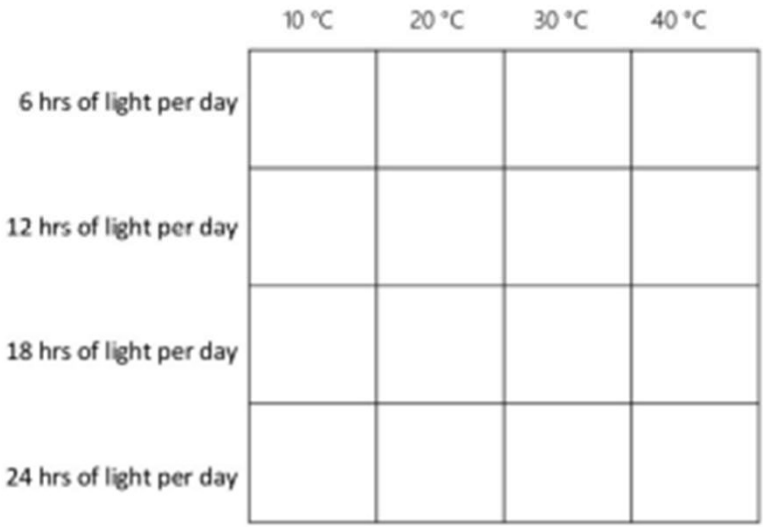
- Consider a table with n=4 attributes



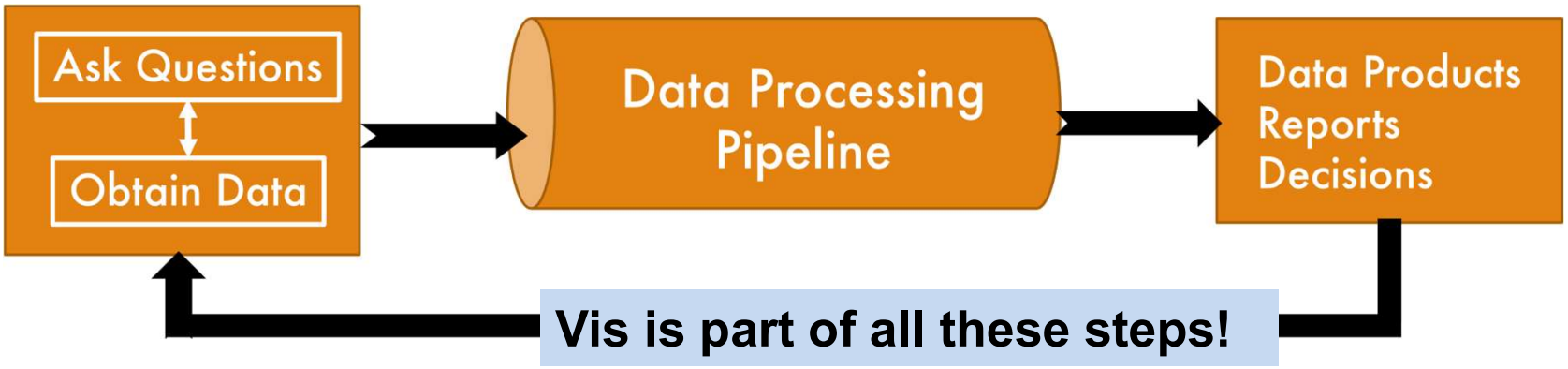
Principle 4: Use Structure

Correlation Visualization

- Conduct a deeper analysis on each pair of attributes



Recap: Data Lifecycle



Related Processes

Big Data *Journey*

- Business transformations as a company becomes more data-centric

Data Visualization *Process*

- Acquire, Parse, Filter, Mine, Represent, Refine, Interact [**Ben Fry '07**, Visualizing Data]

Data Visualization *Pipeline*

- Analyse (Wrangling), Filter, Map to visual properties, Render geometry

Sources

Books

- Tamara Munzner ”Visualization Analysis and Design”, 2014
- Lau, Gonzalez, Nolan: “Principles and Techniques of Data Science”

Slides

- Jiannan Wang’s CMPT 733 slides, Spring 2017
- Torsten Möller’s Visualization course, Spring 2018
- UC Berkley Data 100 (Lau, Nolan, Dudoit, Perez)

Visualization group exercise

- How would you visualize demographic data?
  - Total population per year
  - Broken down by gender
  - Differentiate by age
  - Include ...?
- Why would you visualize demographic data?