Lecture 20: Visualization I

COSC 480 Data Science, Spring 2017

Michael Hay

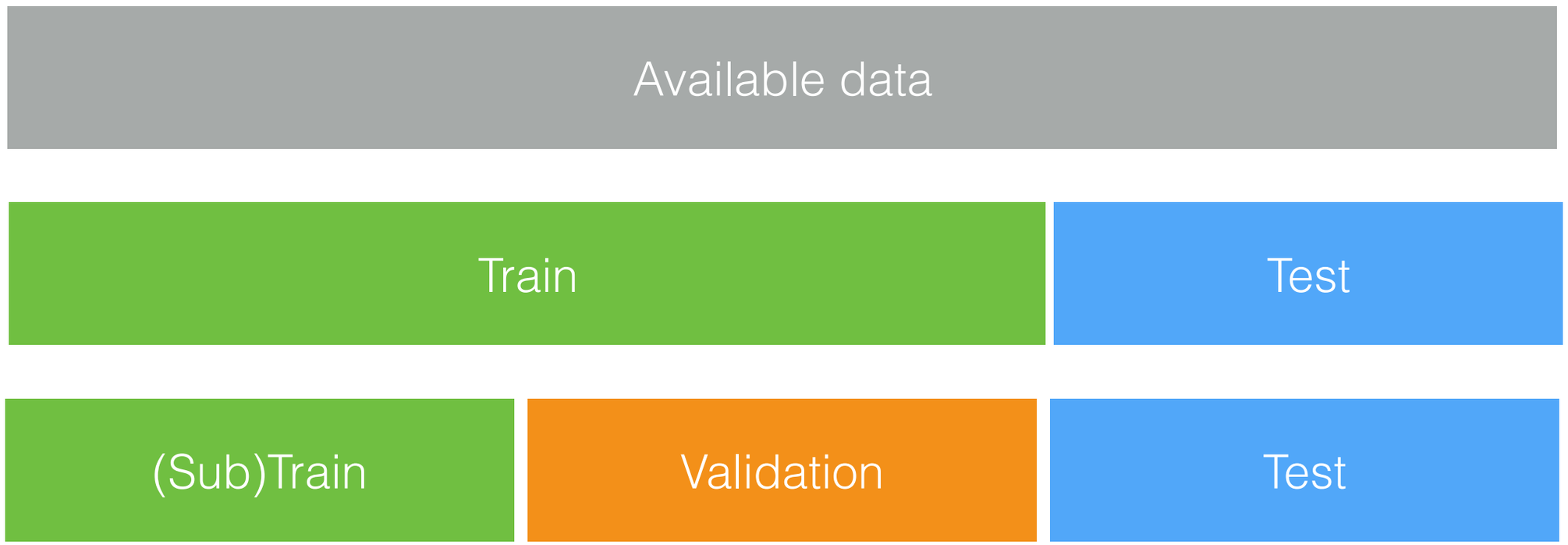
Logistics

- Quiz tomorrow
  - Included: Perceptron, Naive Bayes, Decision Trees
  - Not included: learning theory, overfitting

Today

- Finish up cross-validation
- Start on viz

Recap



- Train variety of models on **train**
- Pick best based on error on **validation**
- Evaluate your final choice on **test**

Drawbacks of splitting into train/validation/test

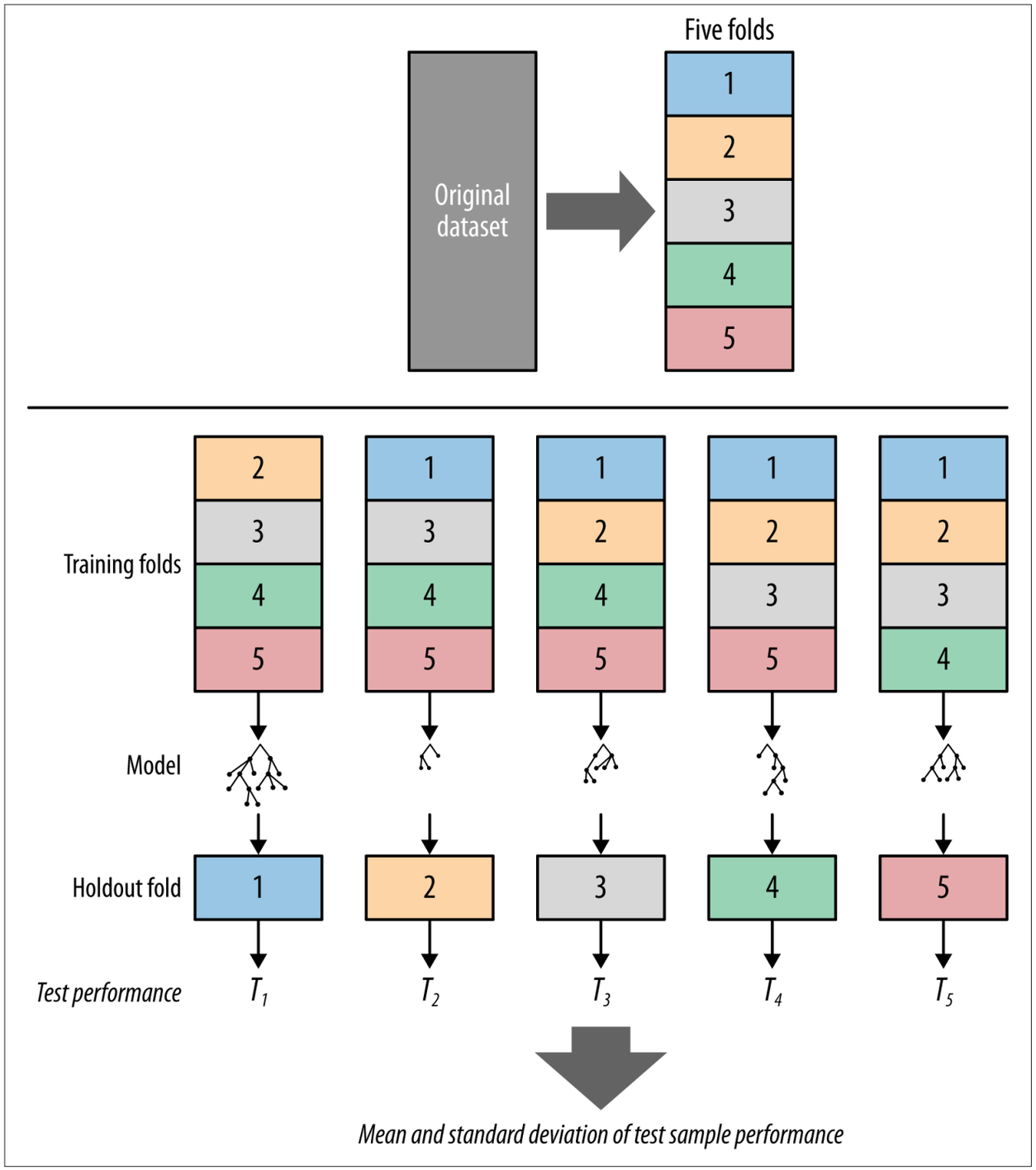
- A fixed amount of labeled examples are now divided into three subsets
- We want each subset to be *as big as possible*
  - Bigger (sub)train: more *information* for training
  - Bigger validation and test: more reliable estimates of "true" error

Cross validation

- Applicable for splitting a dataset into **two** parts:
  - Example: **train/test**
  - Example: (sub)train/validate
- Idea: use every example as a test example
- Extreme version: ***Leave one out cross validation***
  - Learn on n-1 examples, test on n^{th} example, record result
  - Repeat this n times! Average the results.

Credit: Data Science for Business, Provost and Fawcett

Cross validation



Credit: Data Science for Business, Provost and Fawcett

Question

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Suppose we have already decided that we are going to fit a max depth=5 decision tree. We want to (a) train it, and (b) evaluate its performance (i.e., test it).

If the outcome of the evaluation is good, we will deploy our decision tree in the "wild."

If we use 5 fold cross validation, how many decision trees do we build in total?

Question

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Suppose we have already decided that we are going to fit a max depth=5 decision tree. We want to (a) train it, and (b) evaluate its performance (i.e., test it).

If the outcome of the evaluation is good, we will deploy our decision tree in the "wild."

If we use 5 fold cross validation, how many decision trees do we build in total?

6 in total. 5 during cross validation (1 per "holdout" fold — see previous figure). If the results are good, we will train a final tree using all of the data and deploy that one.

Question

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Suppose we want to fit a decision tree but we are unsure of the max depth parameter. We want to identify the "best" setting of max depth.

We will try 6 settings of max-depth 1, 2, …, 6.

Rather than simply split our training data into (sub)train and validate, we will use 5 fold cross-validation.

How many decision trees do fit in total?

Question

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

Suppose we want to fit a decision tree but we are unsure of the max depth parameter. We want to identify the "best" setting of max depth.

We will try 6 settings of max-depth 1, 2, …, 6.

Rather than simply split our training data into (sub)train and validate, we will use 5 fold cross-validation.

How many decision trees

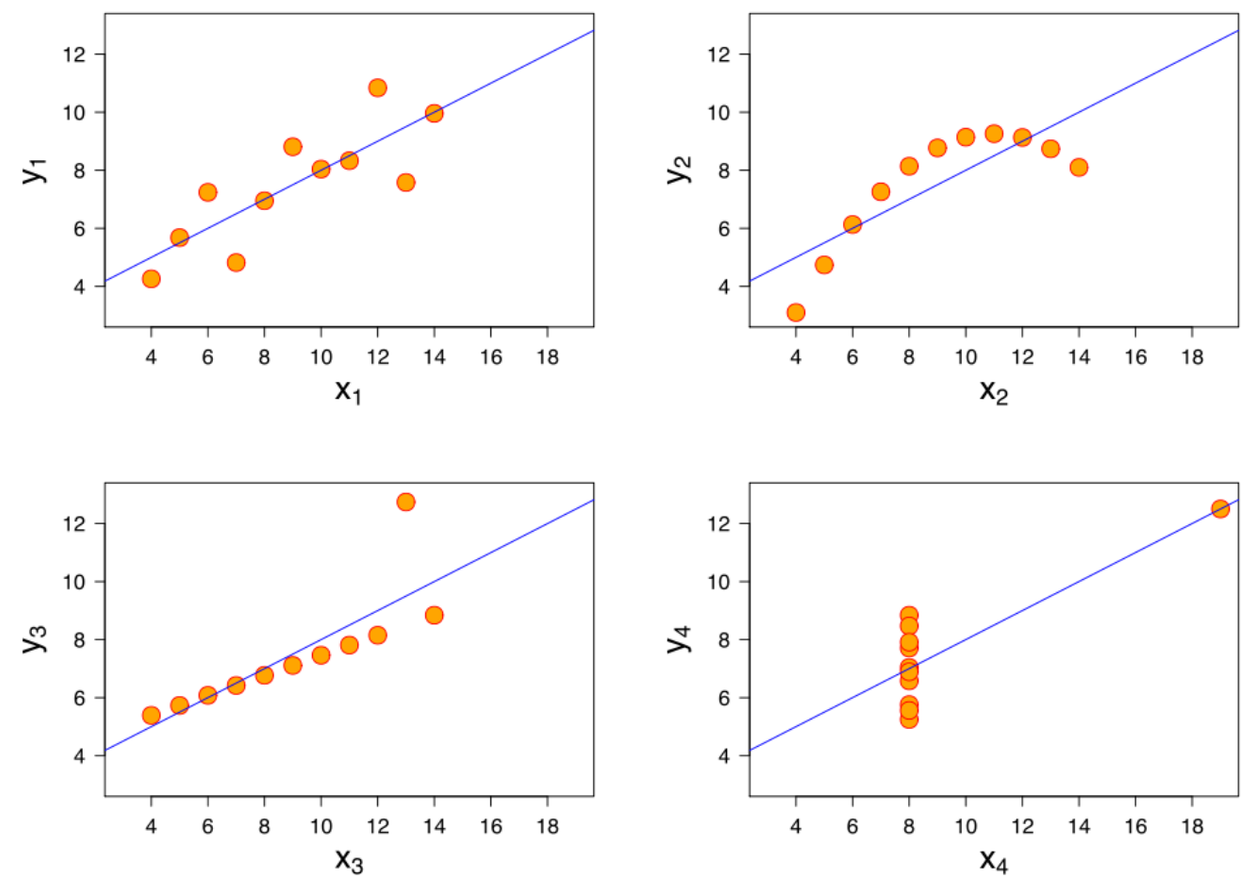
- (6 x 5) + 1 = 31 in total.
- 6 per "holdout" fold — see previous figure — for the sixth depths.
- This is repeated 5 times (5 fold cross val).
- The outcome is 5 "best" settings of max depth
- Retrain one last tree with max depth fixed

Why visualize?

Preserve complexity

Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



4 datasets, identical in terms of basic stats

Evaluate data quality

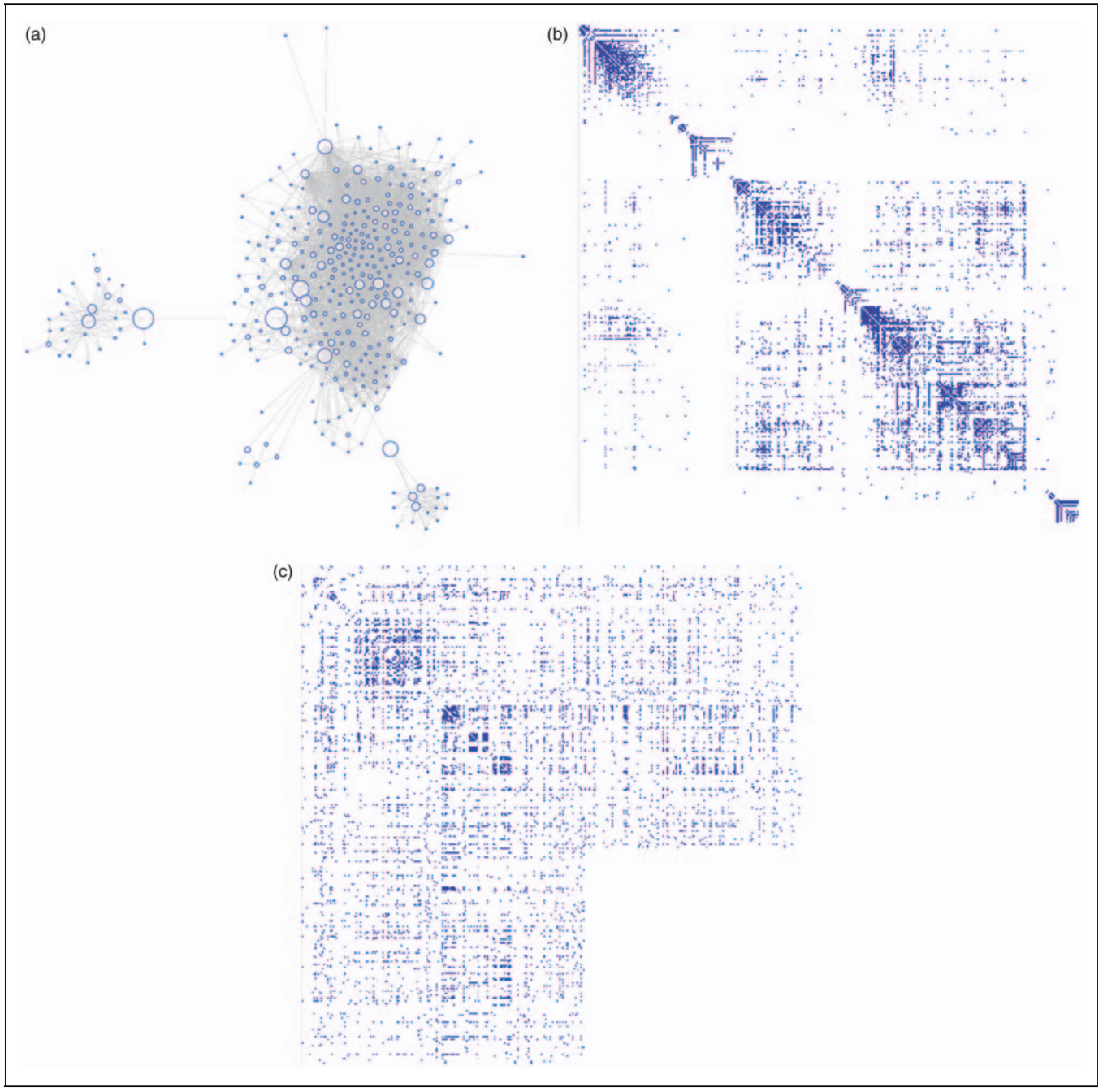


Figure 2. The choice of visual representation impacts the perception of data quality issues. (a) A node-link diagram of a social network does not reveal any irregularities. (b) A matrix view sorted to emphasize connectivity shows more sub-structure, but no errors pop out. (c) Sorting the matrix by raw data order reveals a significant segment of missing data.

Figure 2(c) is more revealing. The rows and columns are instead sorted in the order provided by the Facebook API. We now see a striking pattern: the bottom-right corner of the matrix is completely empty. Indeed, this is a missing data problem, as Facebook enforces a 5000-item result limit for a query. In this case, the maximum was reached, the query failed silently, and the mistake went unnoticed until visualized.

http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

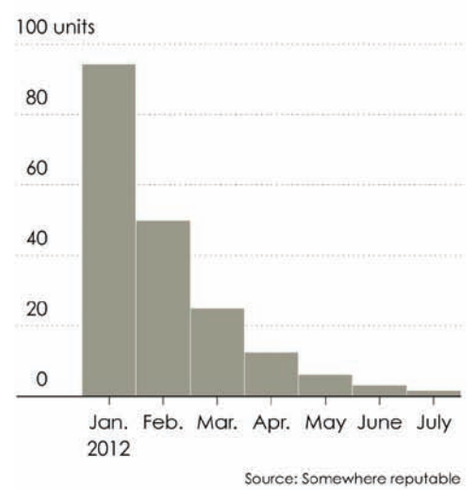
Tell a story

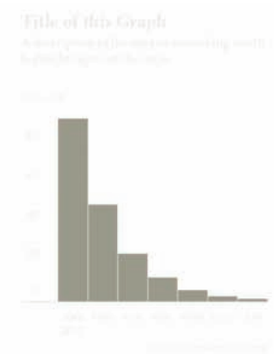
https://www.youtube.com/watch?v=OwII-dwh-bk

Visualization components

Working parts

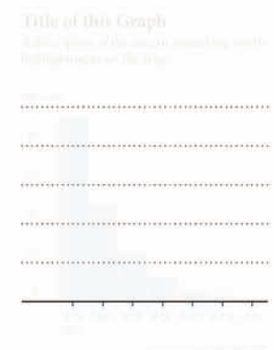
Several pieces work together to make a graph. Sometimes these are explicitly shown in the visualization and other times they form a visual in the background. They all depend on the data.





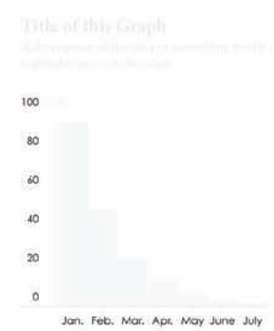
Visual Cues

Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.



Coordinate System

You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.



Scale

Increments that make sense can increase readability, as well as shift focus.

Context

If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.

Visual cues

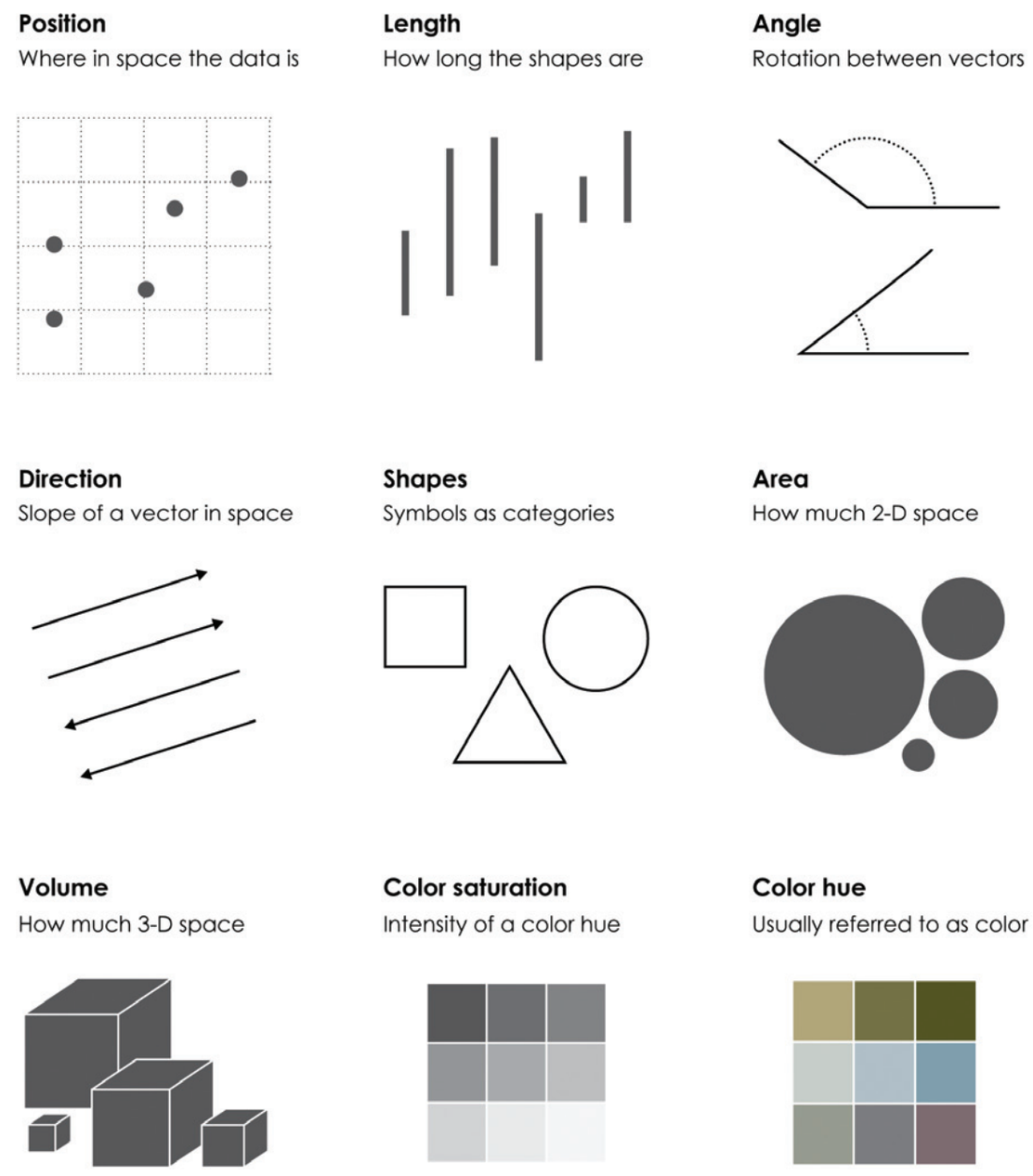
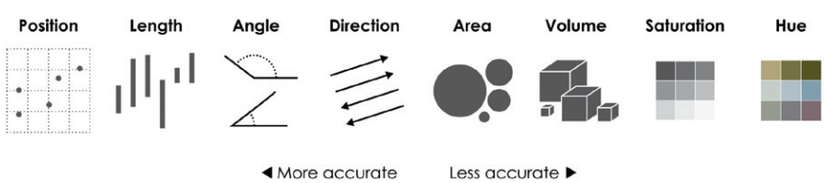


FIGURE 3-3 *Visual cues*

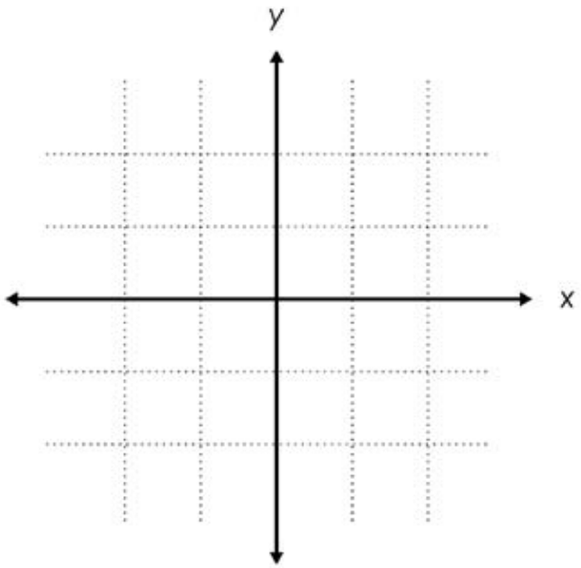
Human perception



Coordinate systems

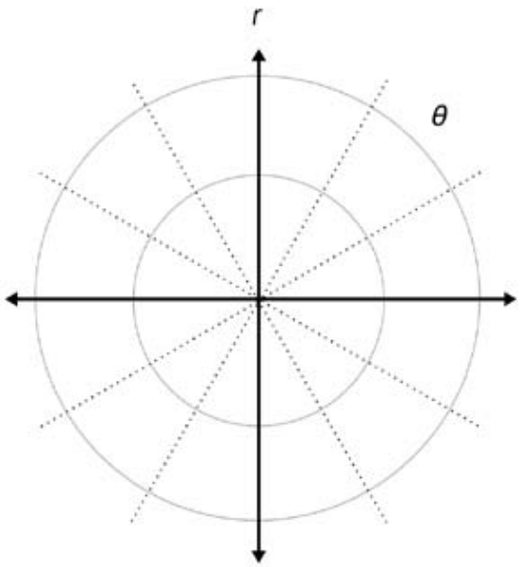
Cartesian

If you've ever made a graph, the x- and y-coordinate system will look familiar to you.



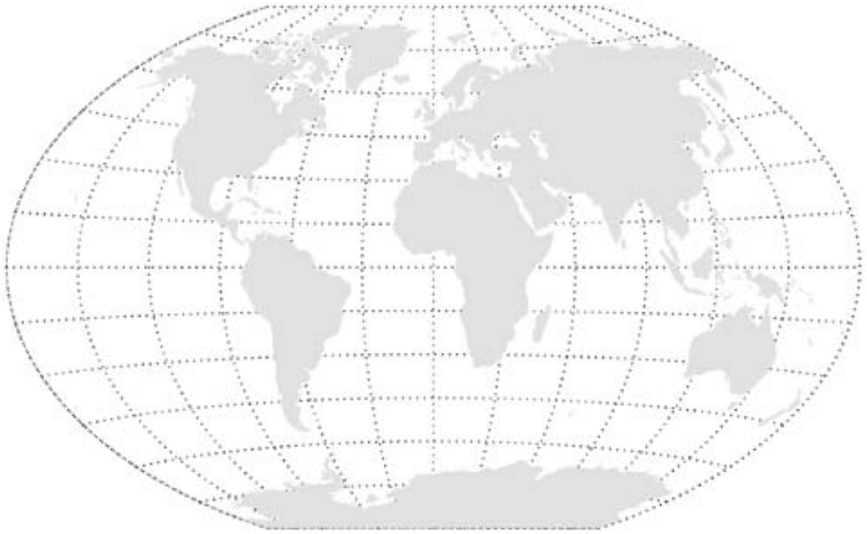
Polar

Pie charts use this system. Coordinates are placed based on radius *r* and angle *θ*.



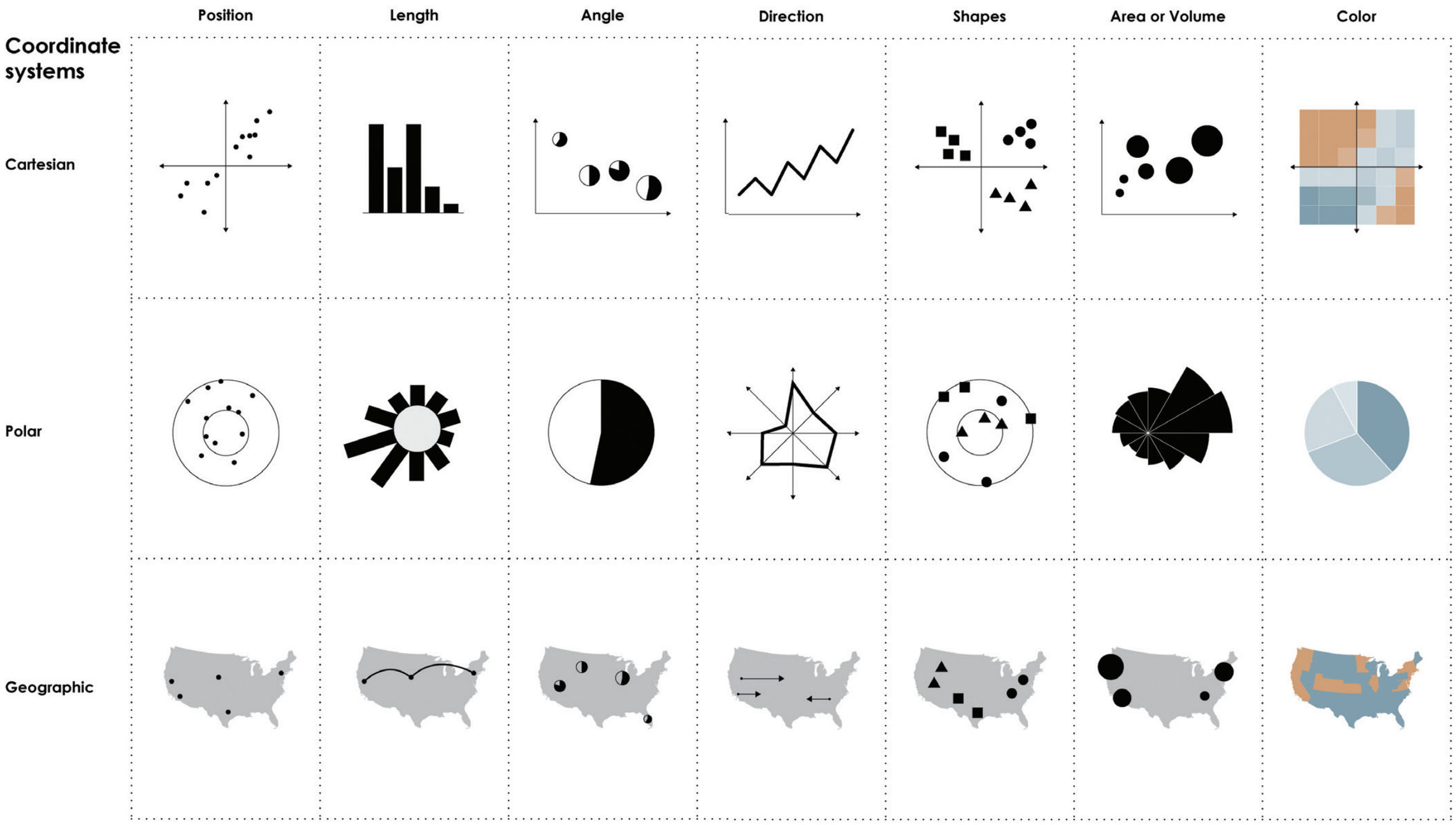
Geographic

Latitude and longitude are used to identify locations in the world. Because the planet is round, there are multiple projections to display geographic data in two dimensions. This one is the Winkel tripel.

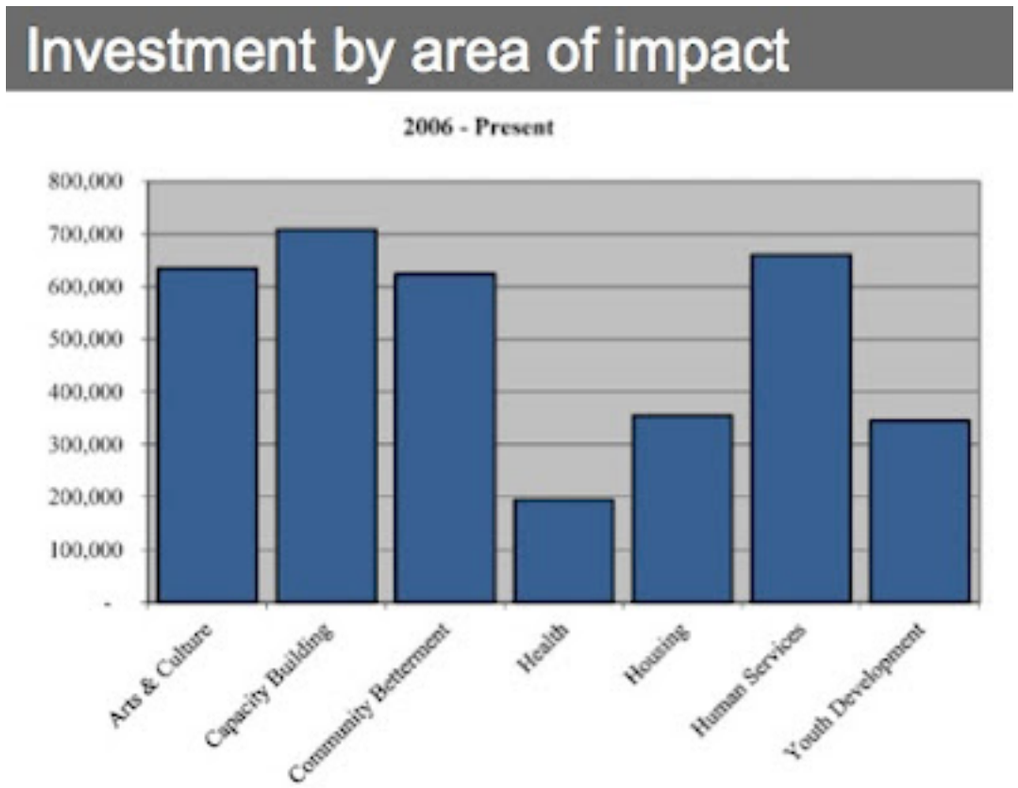


Scales

Time is special… why?


Position
Length
Angle
Direction
Shapes
Area or Volume
Color
Coordinate systems
Cartesian
Polar
Geographic

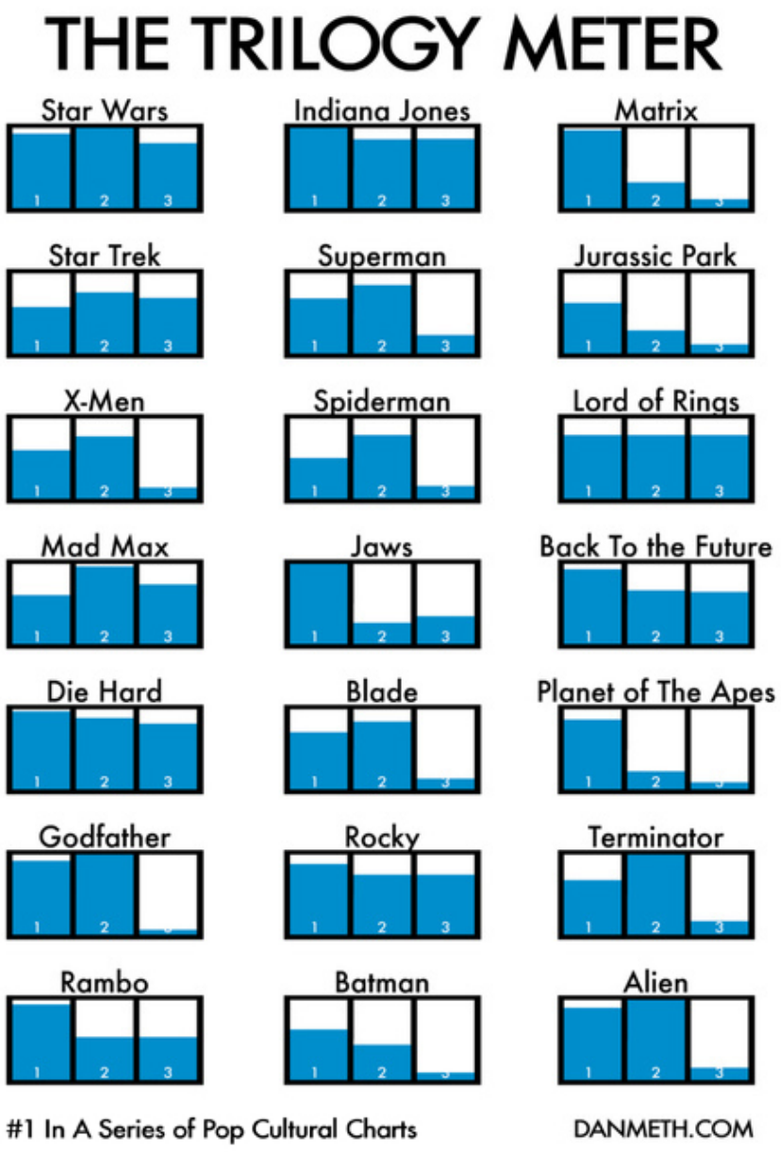
Context



- Frame the data visualization for the readers with titles, pre-amble text, annotations, etc.

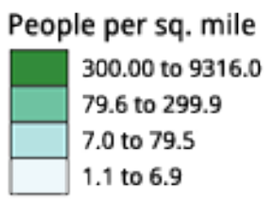
- Minimize clutter

Context: small multiples

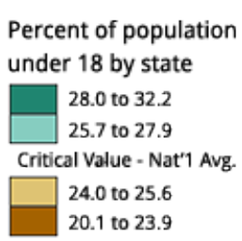


Color

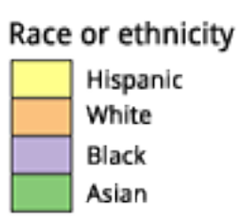
- Color is **difficult**
- People have trouble differentiating more than 5-7 colors/shades.
- Color schemes (see http://colorbrewer2.org/)



- Sequential: suited to ordered data that progress from low to high



- Diverging: suited for ordered data that diverges above and below some "middle" value



- Qualitative: distinguish between categorical data without implied order

http://colorbrewer2.org/

Scales for color

- How do you map *numerical* data into color?
- Linear: map color onto linear space (using color theory) then map data to linear space
- Quantile: sort data, break into groups of equal number of records
- Quantize: sort data, break max-min into equal width ranges

http://roadtolarissa.com/coloring-maps/

Question

For each graphic, identify the visual cues, coordinate system, and scale(s).

How many variables are depicted in each graphic? Link each variable to visual cue.

Critique the viz, including use of *context*.

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

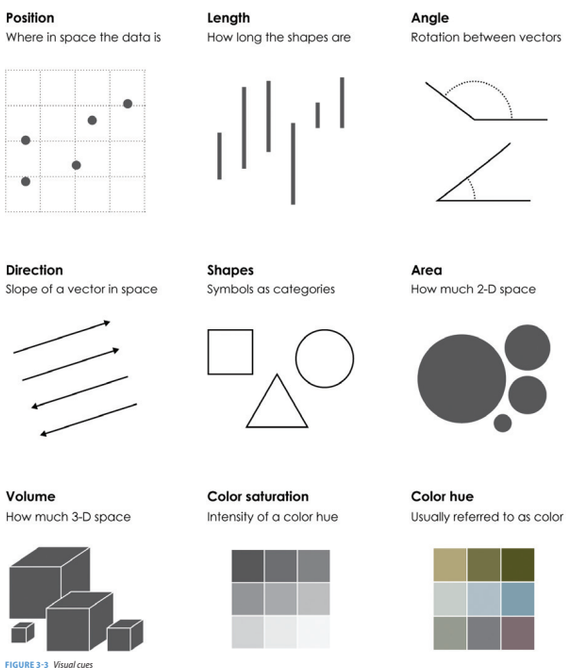
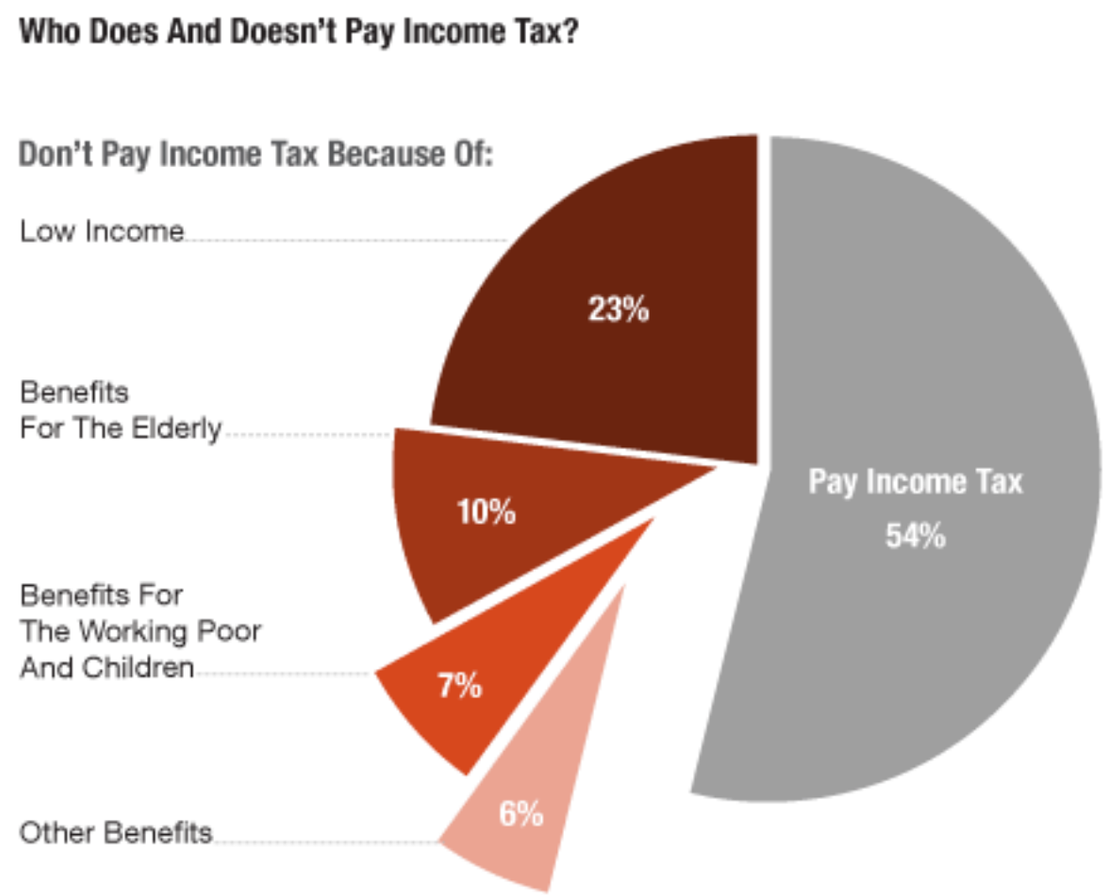


FIGURE 3-3 *Visual cues*



Coordinate systems:

- Cartesian (x,y)
- Polar
- Geographic

Scales:

- linear
- log
- categorical
- ordinal
- percent
- time

http://mdsr-book.github.io/exercises.html

Question

For each graphic, identify the visual cues, coordinate system, and scale(s).

How many variables are depicted in each graphic? Link each variable to visual cue.

Critique the viz, including use of *context*.

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

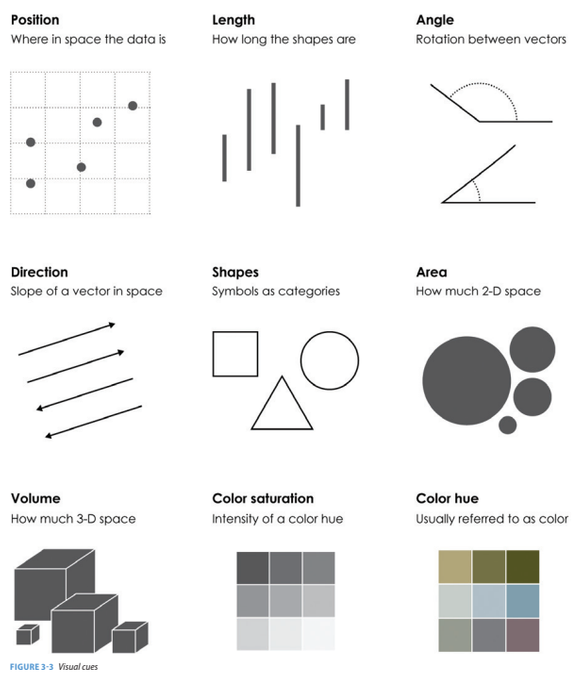
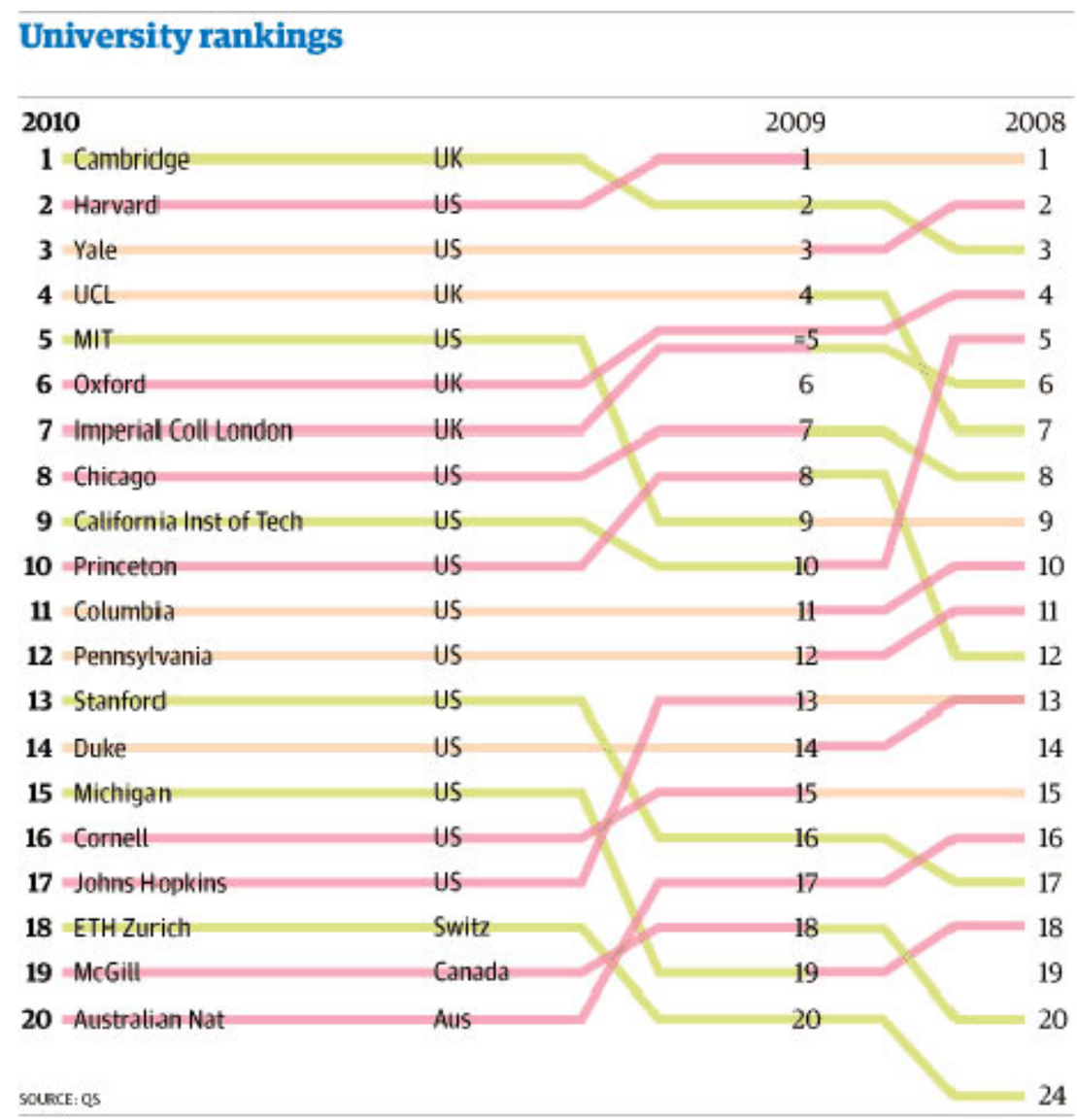


FIGURE 3-3 *Visual cues*



Coordinate systems:

- Cartesian (x,y)
- Polar
- Geographic

Scales:

- linear
- log
- categorical
- ordinal
- percent
- time

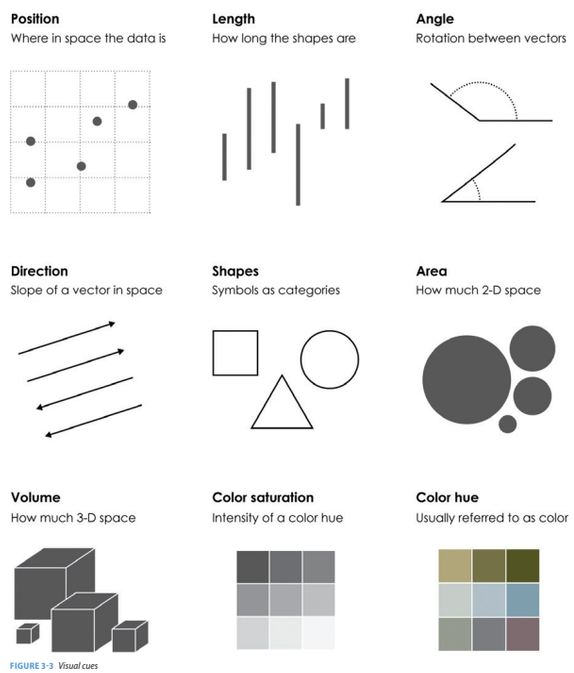
http://mdsr-book.github.io/exercises.html

Question

For each graphic, identify the visual cues, coordinate system, and scale(s).

How many variables are depicted in each graphic? Link each variable to visual cue.

Critique the viz, including use of *context*.



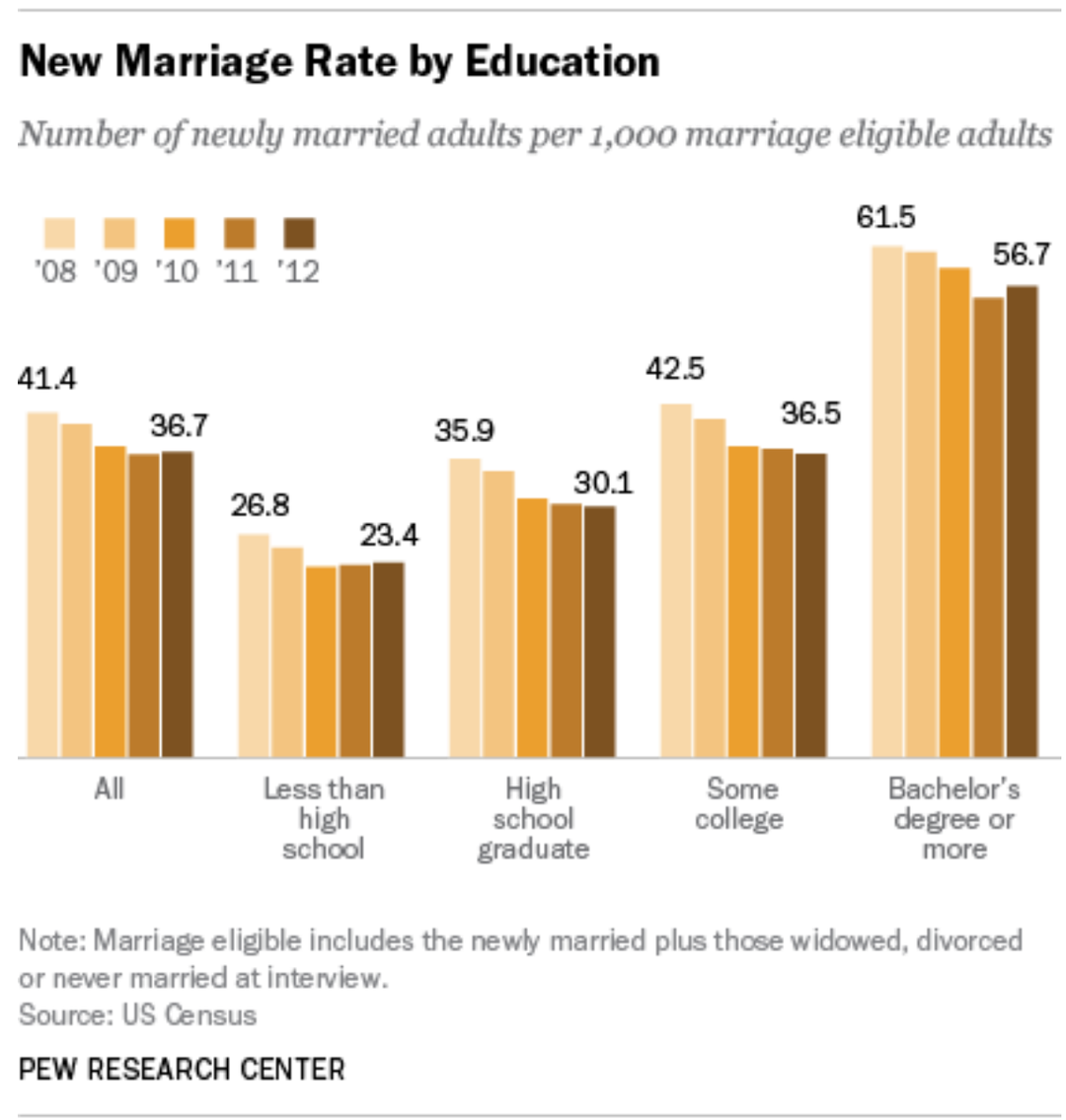
Coordinate systems:

- Cartesian (x,y)
- Polar
- Geographic

Scales:

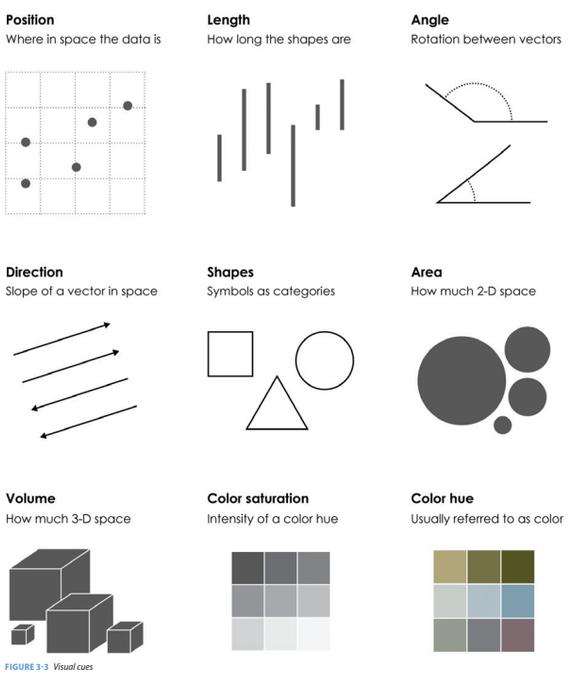
- linear
- log
- categorical
- ordinal
- percent
- time

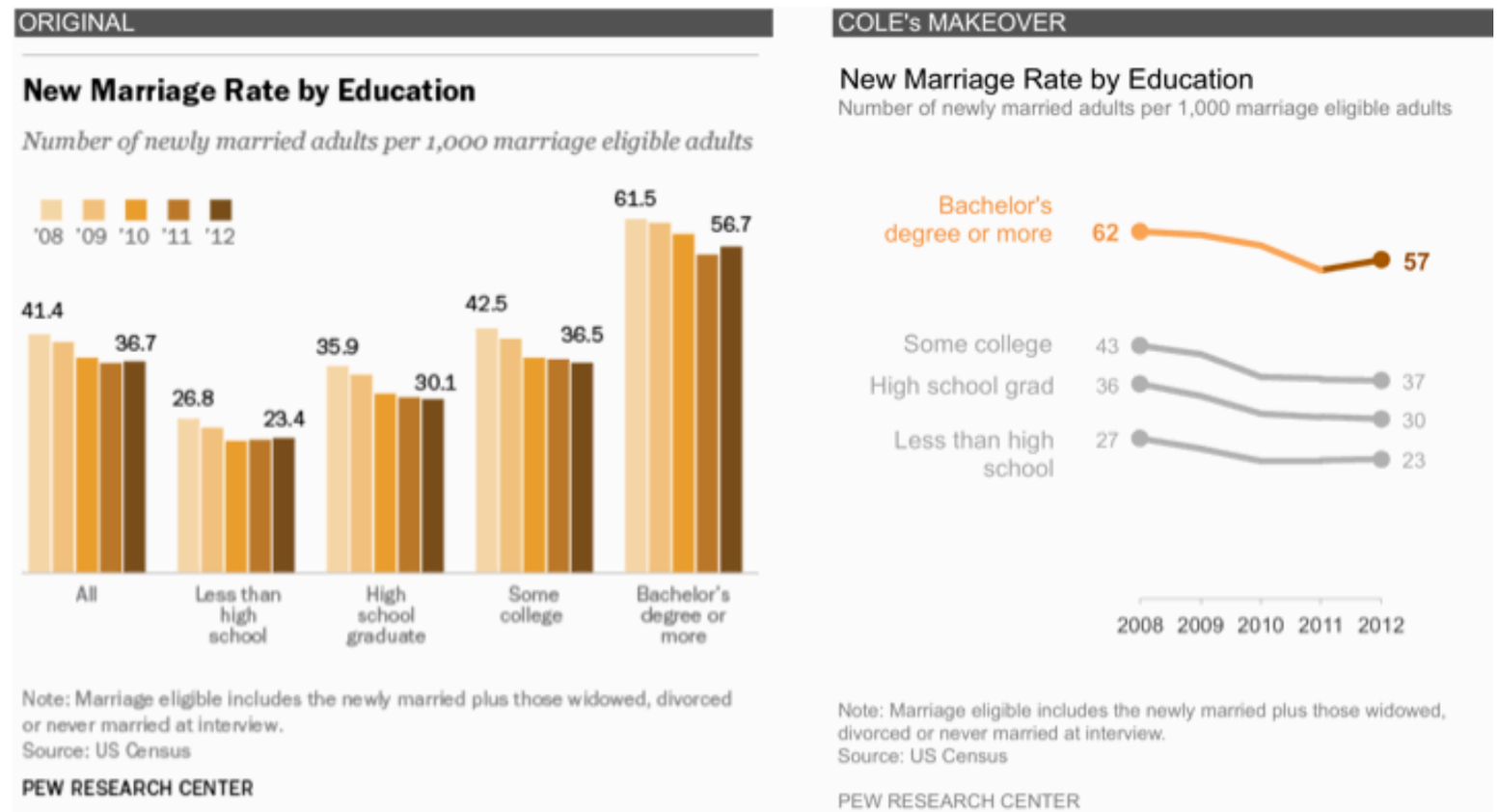
Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you



http://www.storytellingwithdata.com/blog/2014/02/more-americans-are-tying-knot

Visual Makeover





Coordinate systems:

- Cartesian (x,y)
- Polar
- Geographic

Scales:

- linear
- log
- categorical
- ordinal
- percent
- time

http://www.storytellingwithdata.com/blog/2014/02/more-americans-are-tying-knot

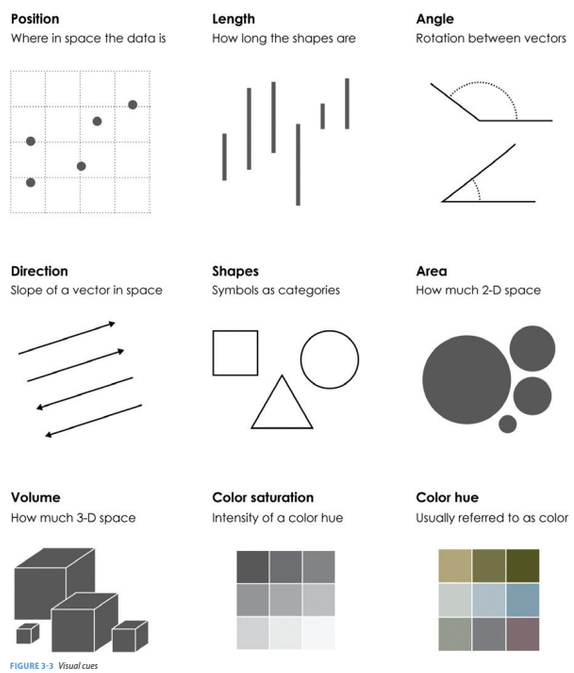
Question

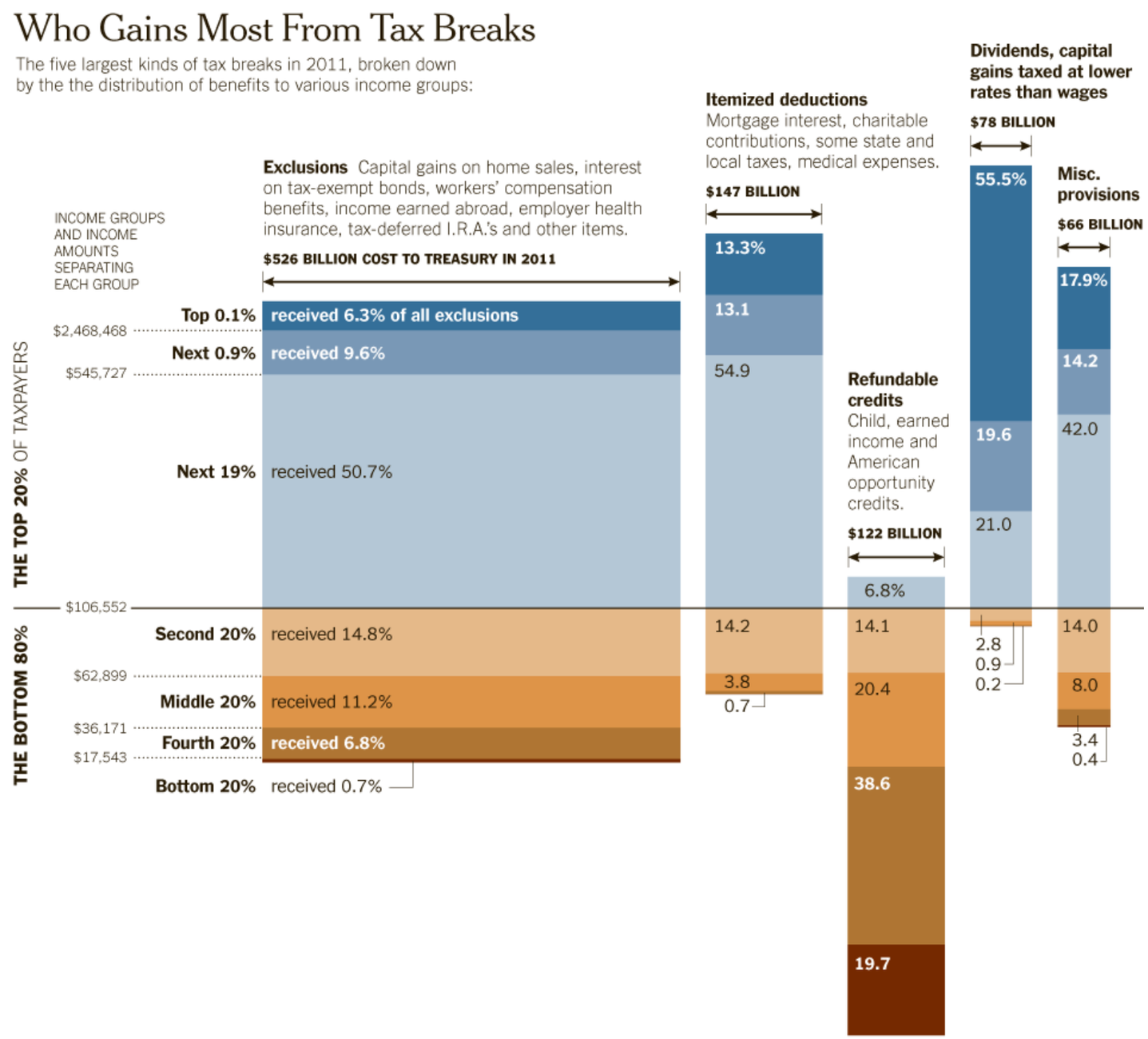
For each graphic, identify the visual cues, coordinate system, and scale(s).

How many variables are depicted in each graphic?
Link each variable to visual cue.

Critique the viz, including use of *context*.

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you





Coordinate systems:

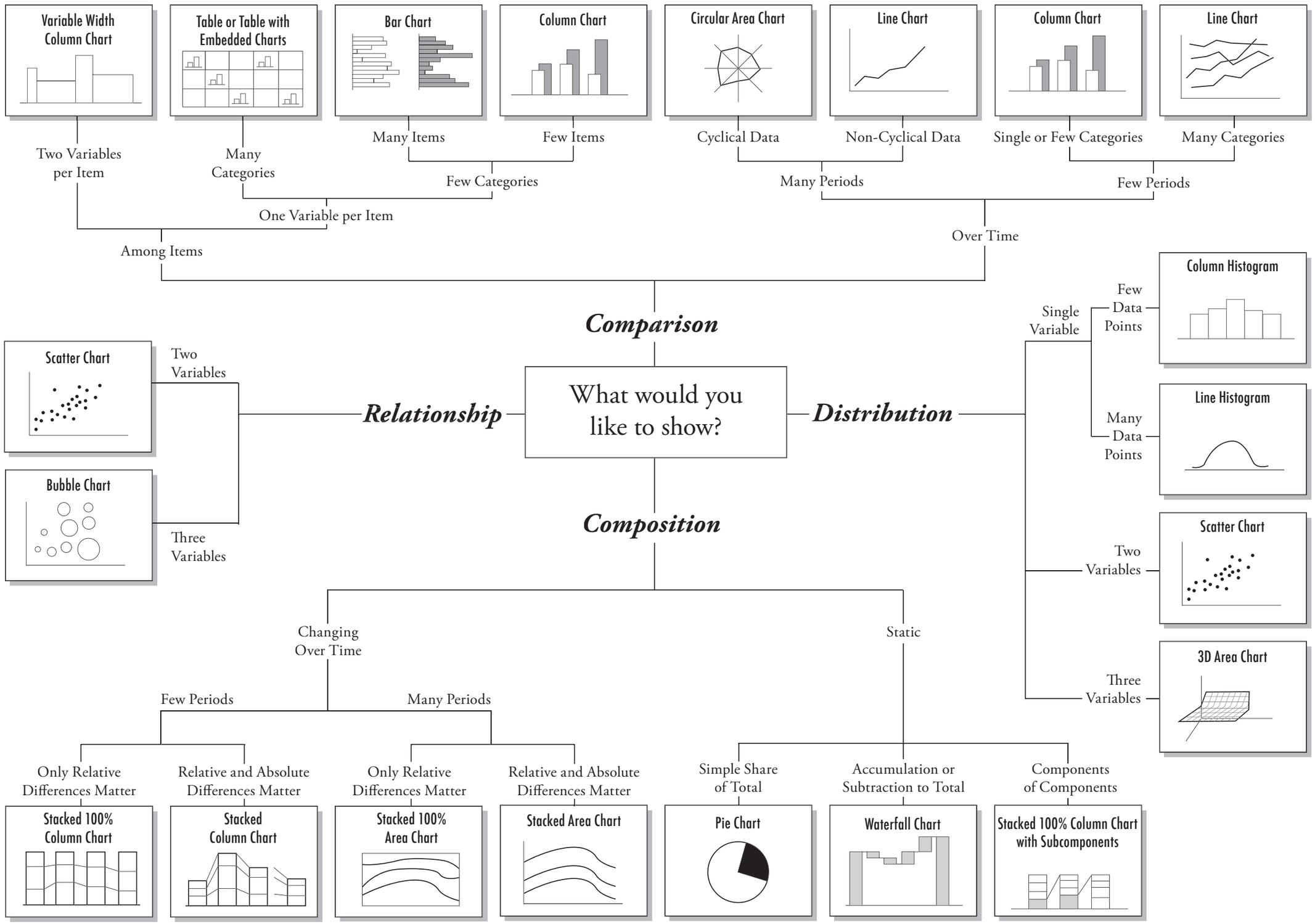
- Cartesian (x,y)
- Polar
- Geographic

Scales:

- linear
- log
- categorical
- ordinal
- percent
- time

http://mdsr-book.github.io/exercises.html

Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Useful resources

- Nathan Yau's book *Data Points* (available thru library)
- Makeovers, http://thewhyaxis.info/remakes/
- WTF Viz, http://wtfviz.net/
- Thumbs Up Viz, http://thumbsupviz.com/
- Help Me Viz, https://policyviz.com/helpmeviz/
- Top 10 do's and don'ts: http://guides.library.duke.edu/datavis/topten