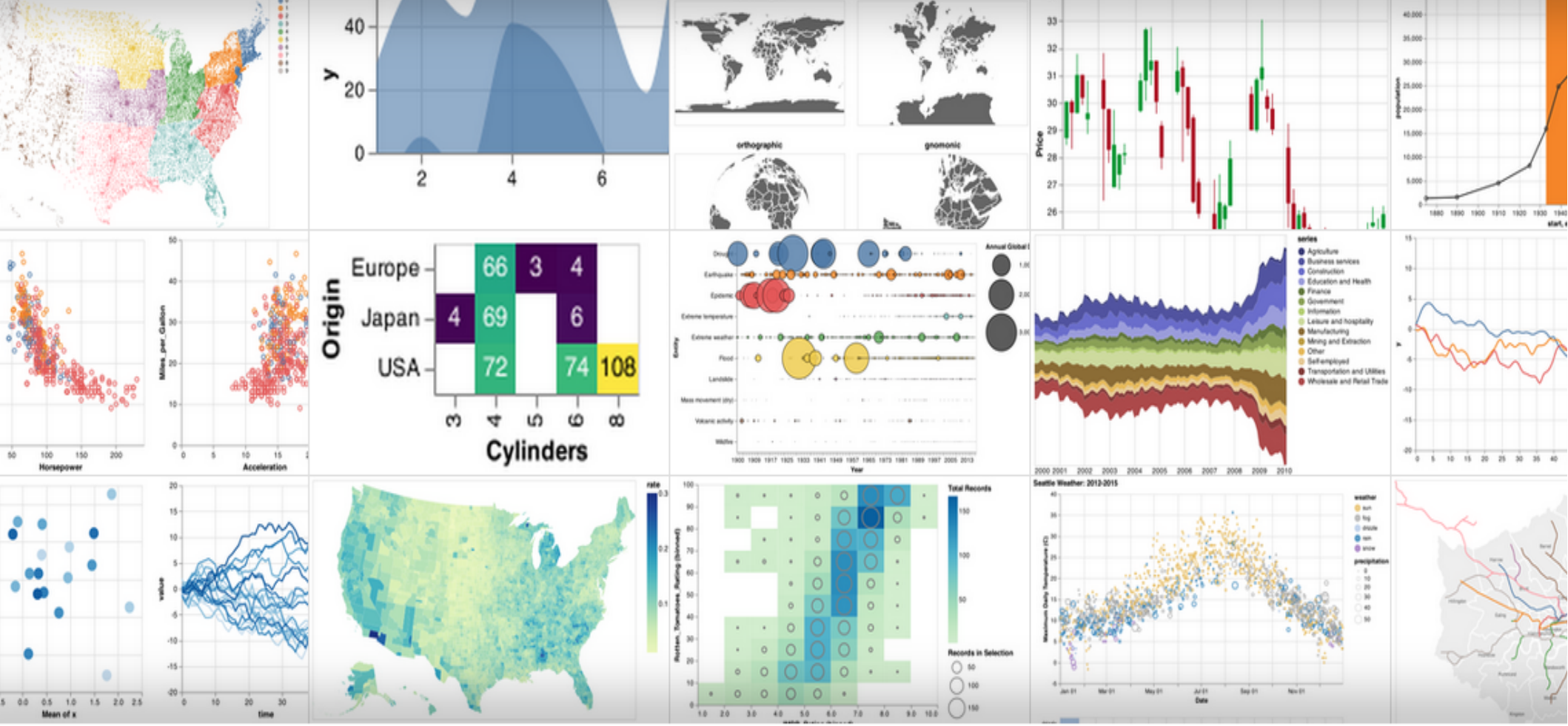
Data 227
Data Visualization & Communication
Autumn 2023



Data 227
Data Visualization & Communication
Autumn 2023

Just kidding.

Greetings

- 2023-09-26 Tuesday class
- I am Will Trimble
  - educated at University of Chicago and University of Washington (go Huskies)
- You are ?

Greetings

- I am Will Trimble
  - educated at Univeristy of Chicago and University of Washington (go Huskies)
- You are ?

  [Undergraduates, mostly 3rd and 4th year, mostly economics and social sciences]

- Have you seen python? R?
  -

Course Plan

- Plan for the course
  - Half slides/discussion, half interactive programming
  - Five homeworks, two projects, no exams
  - No particular technology required, but I can help you with python.
- This class:
  - Quick overview of topics
  - Software choices for visualization
  - jupyter notebook 10 minutes to pandas
  - HW 1 due Friday Oct 6 - census visualization+ data cleaning

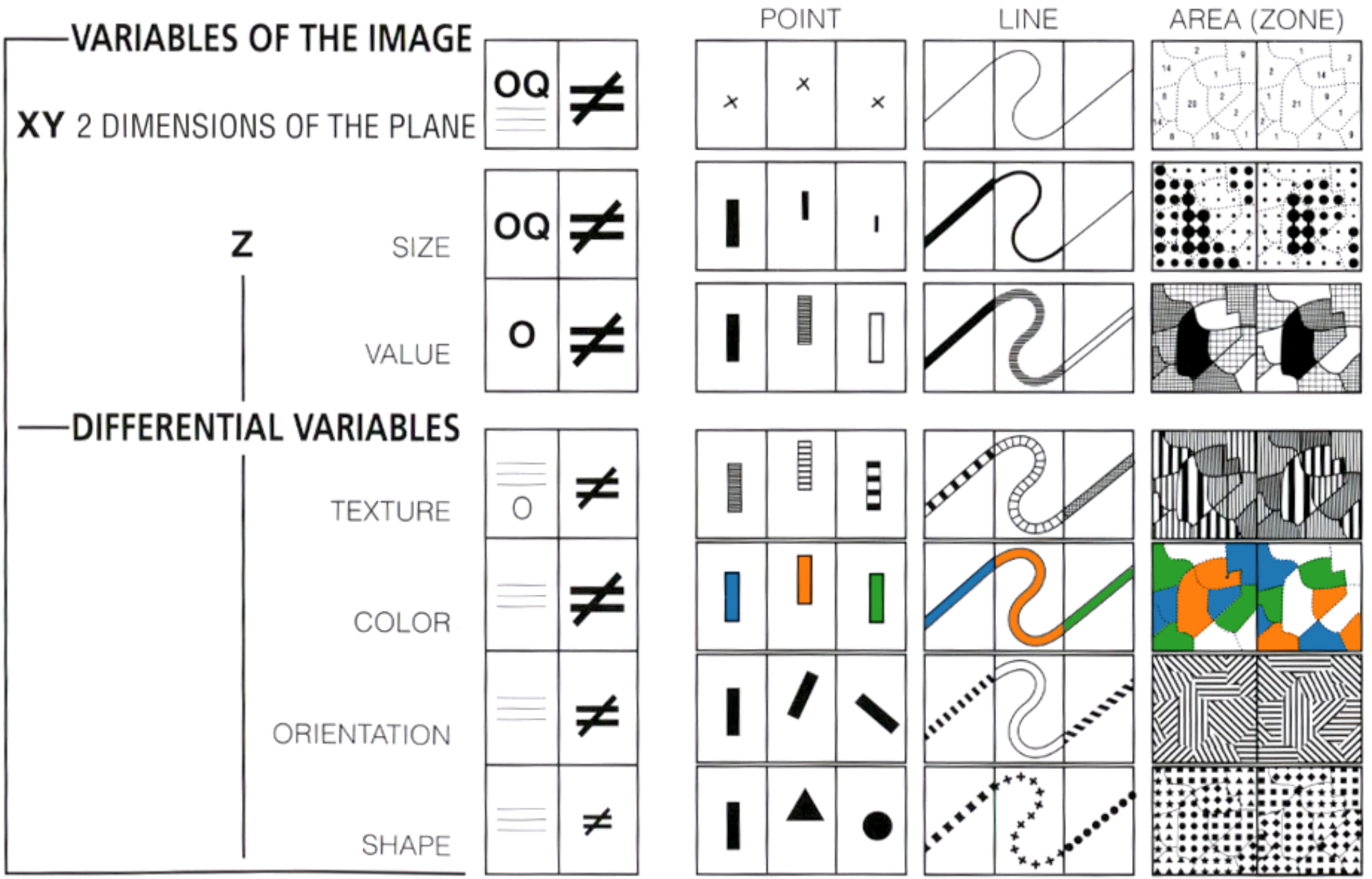
Week 1 readings

- Hadley Wickham. Tidy Data. Journal of Statistical Software 2014

- The Power of Representation, Chapter 3 in Things That Make Us Smart. Don Norman.

- Chapter 1: Graphical Excellence, In The Visual Display of Quantitative Information. Tufte.

Encoding channels



Bertin's taxonomy of visual encodings from Sémiologie Graphique, as adapted by Mike Bostock.

Data types

- integer : stored exactly, but with a limit, positive and negative
- floating point: stored in scientific notation, not perfectly exact, positive and negative
- dates: usually need to be handled separately from numbers and strings (can't add “2021-09-28” and 21:15 otherwise)
- complex: encountered usually in physics and engineering; each point has two real numbers with a slightly different algebra
- vectors and matrices: 1 x n or m x n objects, usually numbers

Shoutout to ISO-8601. There is a right way to write dates (for science). YYYY-MM-DD always. Why ?

More data types

- character: single letter, number, or punctuation mark, conventionally set aside by quotation marks: “A” “8” 'j' '^M'
- strings: sequences of ordered symbols (text)

  “This course introduces best practices for presenting and communicating quantitative data.”
- unicode: generalization of strings to every human alphabet and writing system “اَلسَّلَامُ عَلَيْكُمْ” “A tsunami (from from Japanese: 津波 , lit. 'harbour wave') is a series of waves”
- Boolean: stores True or False, 0 or 1, “binary” data flag

When you import your data, if your computing environment does not guess the data types correctly, you will not be able to use the data.

Attribute types

- Nominal : Labels or categories

  =, ≠ ex: nationality, species, “type”, postal code (can be grouped)

- Ordinal: Ordered

  =,≠ <> ex: Likert scale (agree strongly, somewhat agree..)
  Grades A/B/C/D/F

- Quantitative (interval):

  =,≠ <> +- ex: dates, location

  (ratios meaningless, differences meaningful)

- Quantitative (ratio): physical measurement w/ meaningful zero

  =,≠ <> + * / ex: mass, length, counts

Attribute types

- Nominal : Labels or categories

  =, ≠ ex: nationality, species, “type”, postal code (can be grouped)

- Ordinal: Ordered

  =,≠ <> ex: Likert scale (agree strongly, somewhat agree..) Grades A/B/C/D/F

- Quantitative (interval):

  =,≠ <> +- ex: dates, location

This less-familiar categorization helps us understand what visualization techniques are feasible with which data.

Chart types

What?

Datasets

Data Types

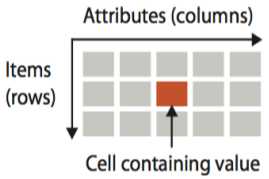
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

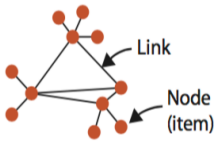
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

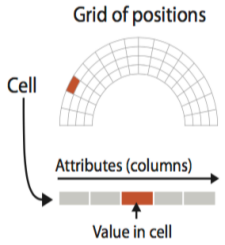
→ Tables



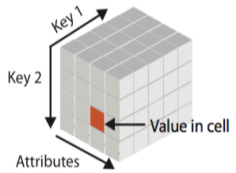
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

→ Geometry (Spatial)

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

→ *Quantitative*

Ordering Direction

→ Sequential

→ Diverging

→ Cyclic

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
  - One Variable per Item
    - Many Categories
      - Table or Table with Embedded Charts
    - Few Categories
      - Many Items
        - Bar Chart
      - Few Items
        - Column Chart
- Over Time
  - Many Periods
    - Cyclical Data
      - Circular Area Chart
    - Non-Cyclical Data
      - Line Chart
  - Few Periods
    - Single or Few Categories
      - Column Chart
    - Many Categories
      - Line Chart

Relationship

- Two Variables
  - Scatter Chart
- Three Variables
  - Bubble Chart

Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

Composition

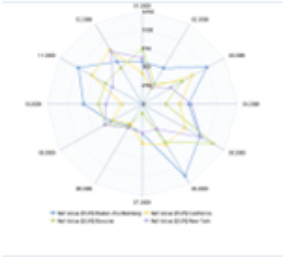
- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

Chart types

Area chart

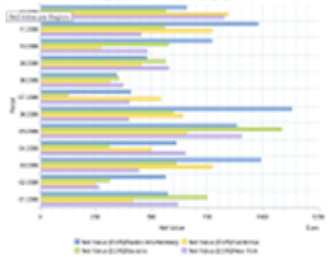
Pie chart

Radar chart

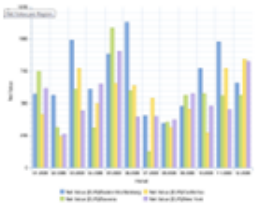


Stacked columns chart

Bar chart

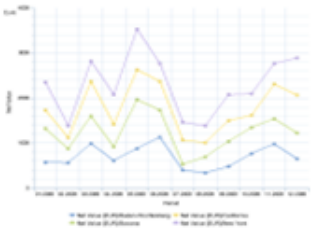


Pipeline chart

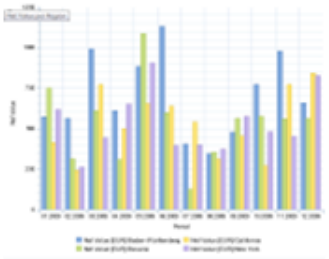


Speedometer chart

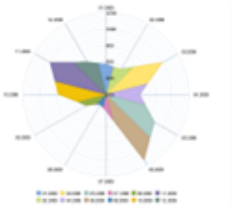
Stacked line chart



Vertical bar chart

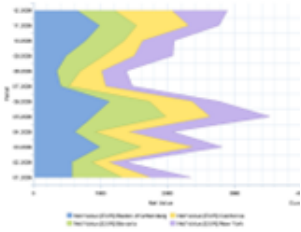


Polar chart



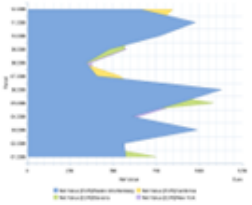
Split pie chart

Stacked profile area chart

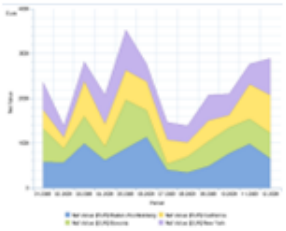


Doughnut chart

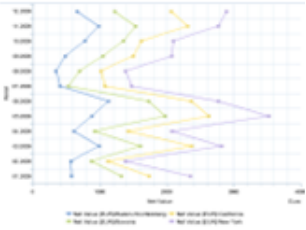
Profile area chart



Stacked area chart

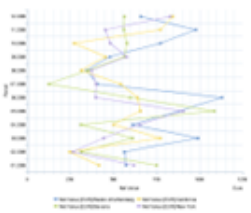


Stacked profile chart

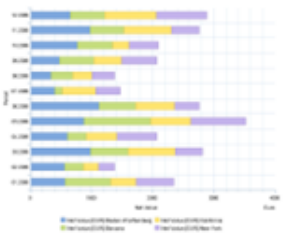


Line chart

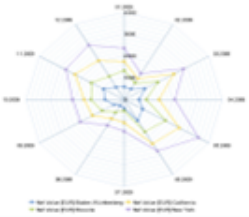
Profile chart



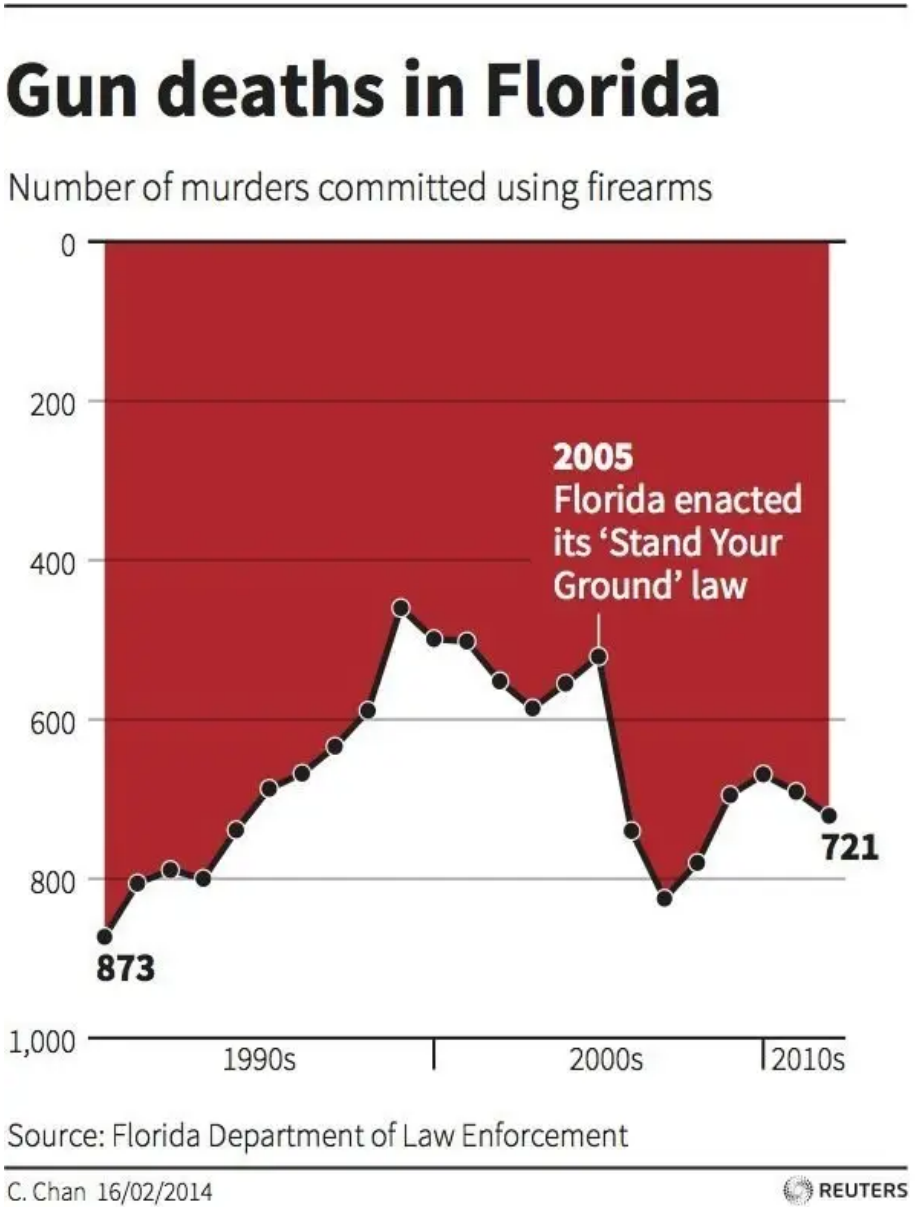
Stacked bar chart



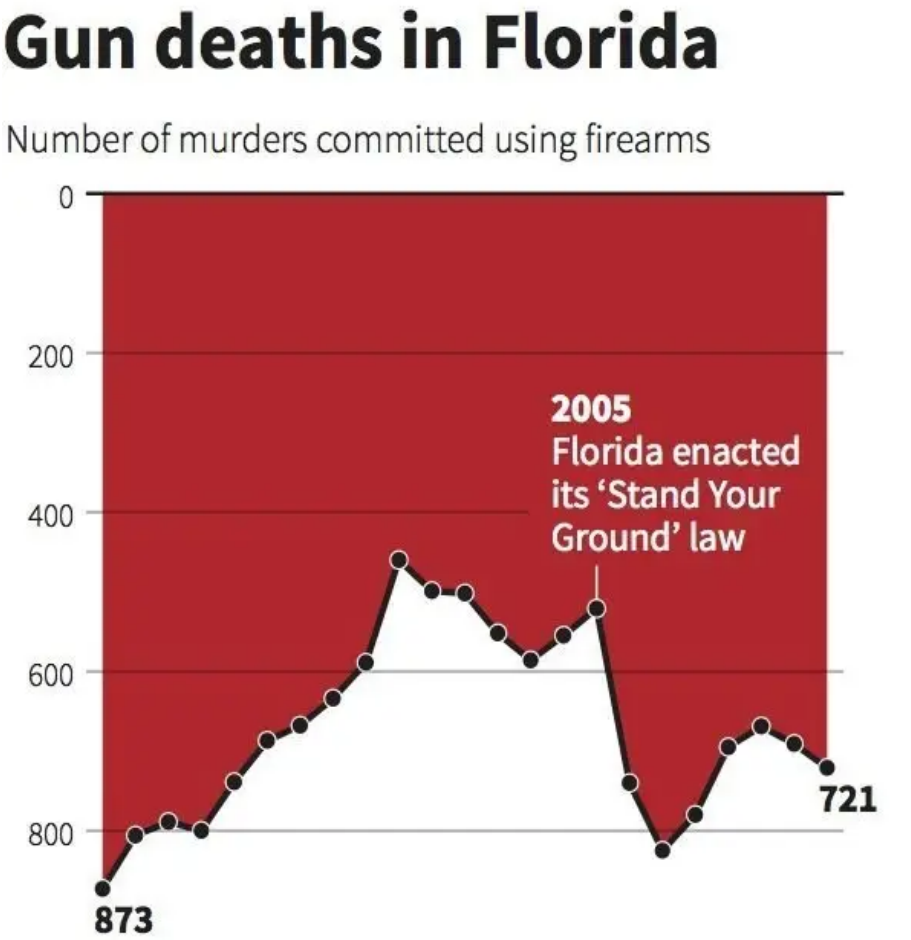
Stacked radar chart



Ethics

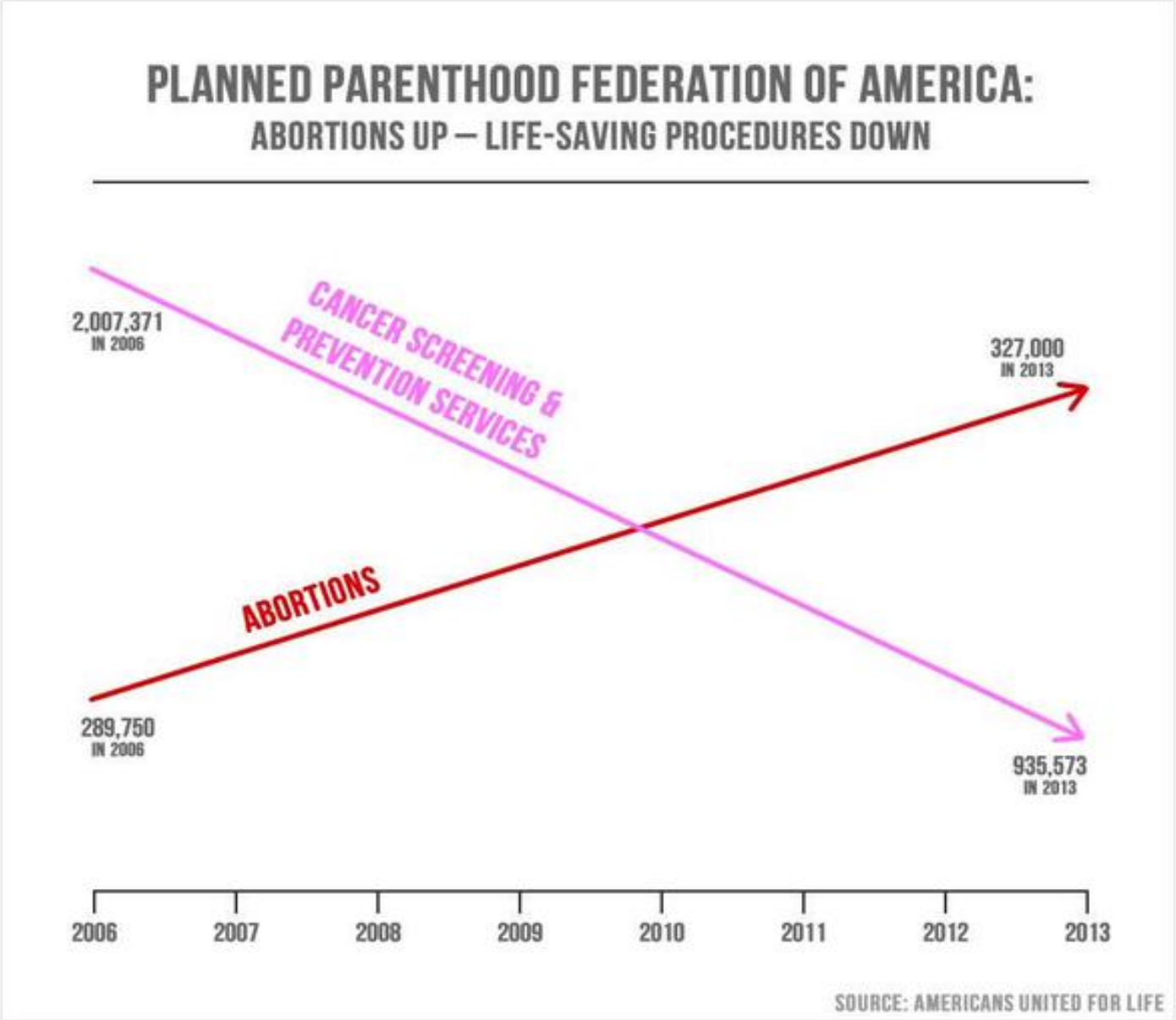

Gun deaths in Florida
Number of murders committed using firearms
0
200
400
600
800
1,000
2005
Florida enacted its 'Stand Your Ground' law
873
721
1990s
2000s
2010s
Source: Florida Department of Law Enforcement
C. Chan 16/02/2014
REUTERS

Ethics

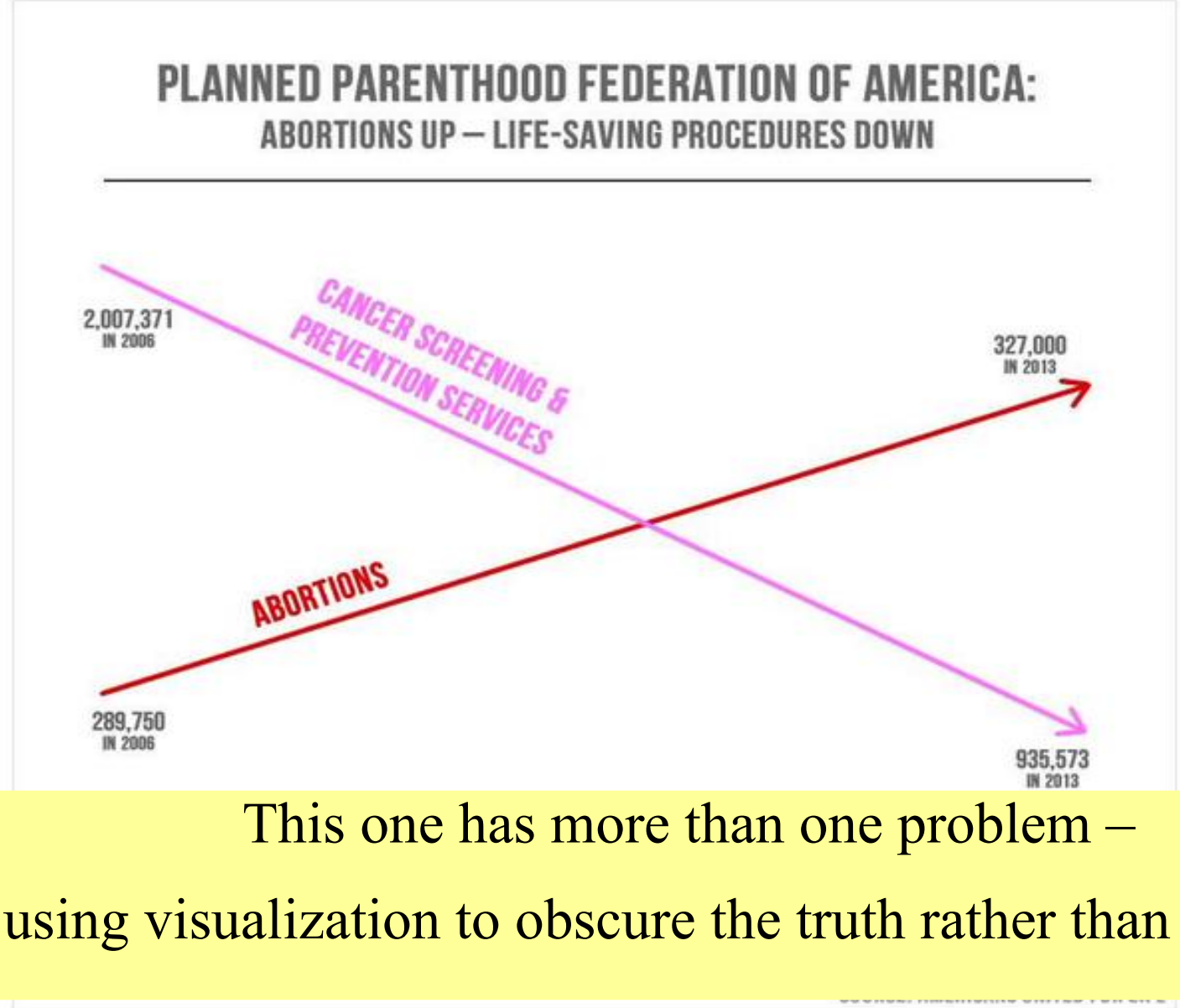


This only has one problem with it..

Ethics


PLANNED PARENTHOOD FEDERATION OF AMERICA:
ABORTIONS UP – LIFE-SAVING PROCEDURES DOWN
2,007,371
IN 2006
CANCER SCREENING & PREVENTION SERVICES
327,000
IN 2013
ABORTIONS
289,750
IN 2006
935,573
IN 2013
2006
2007
2008
2009
2010
2011
2012
2013
SOURCE: AMERICANS UNITED FOR LIFE

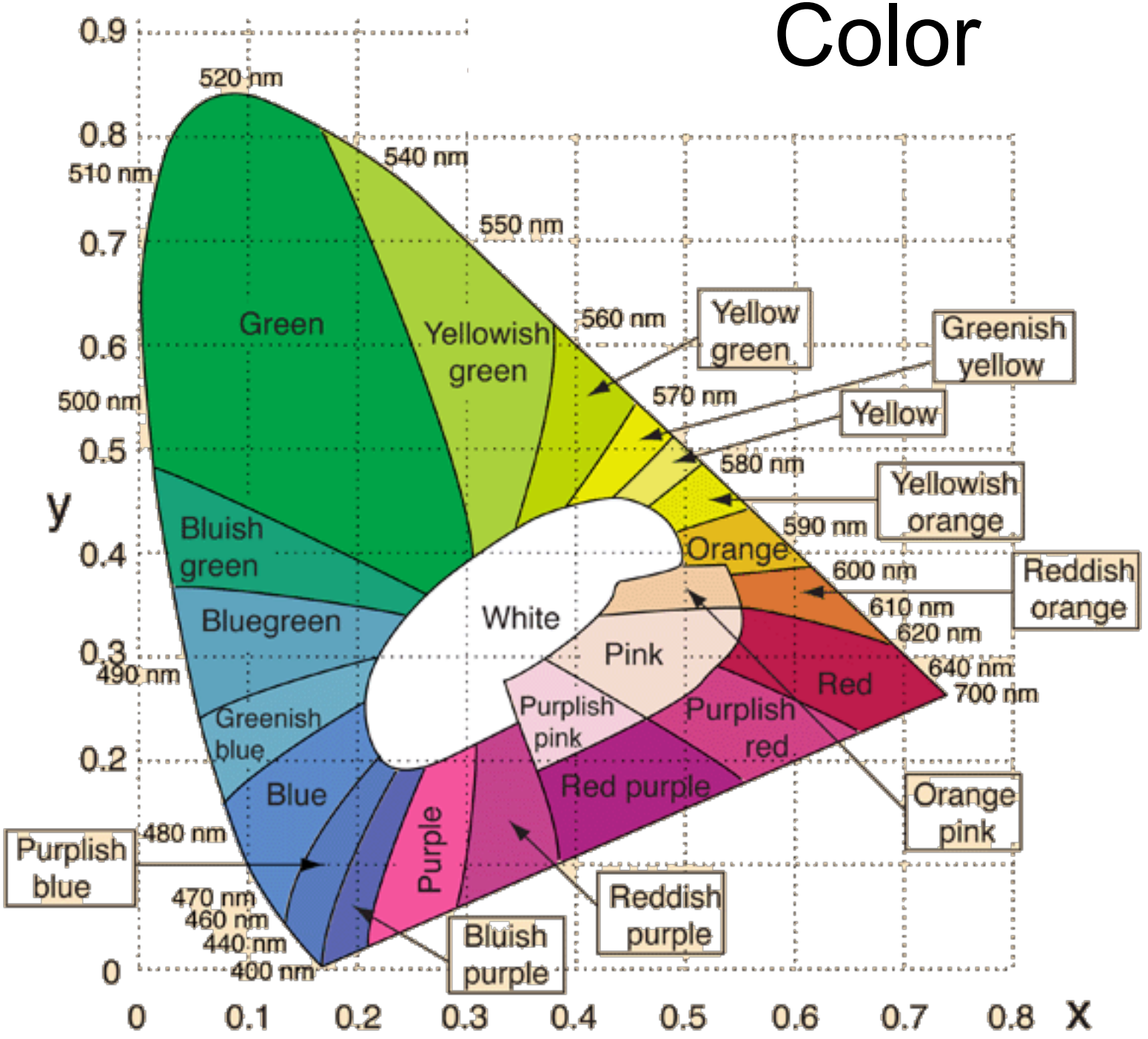
Ethics



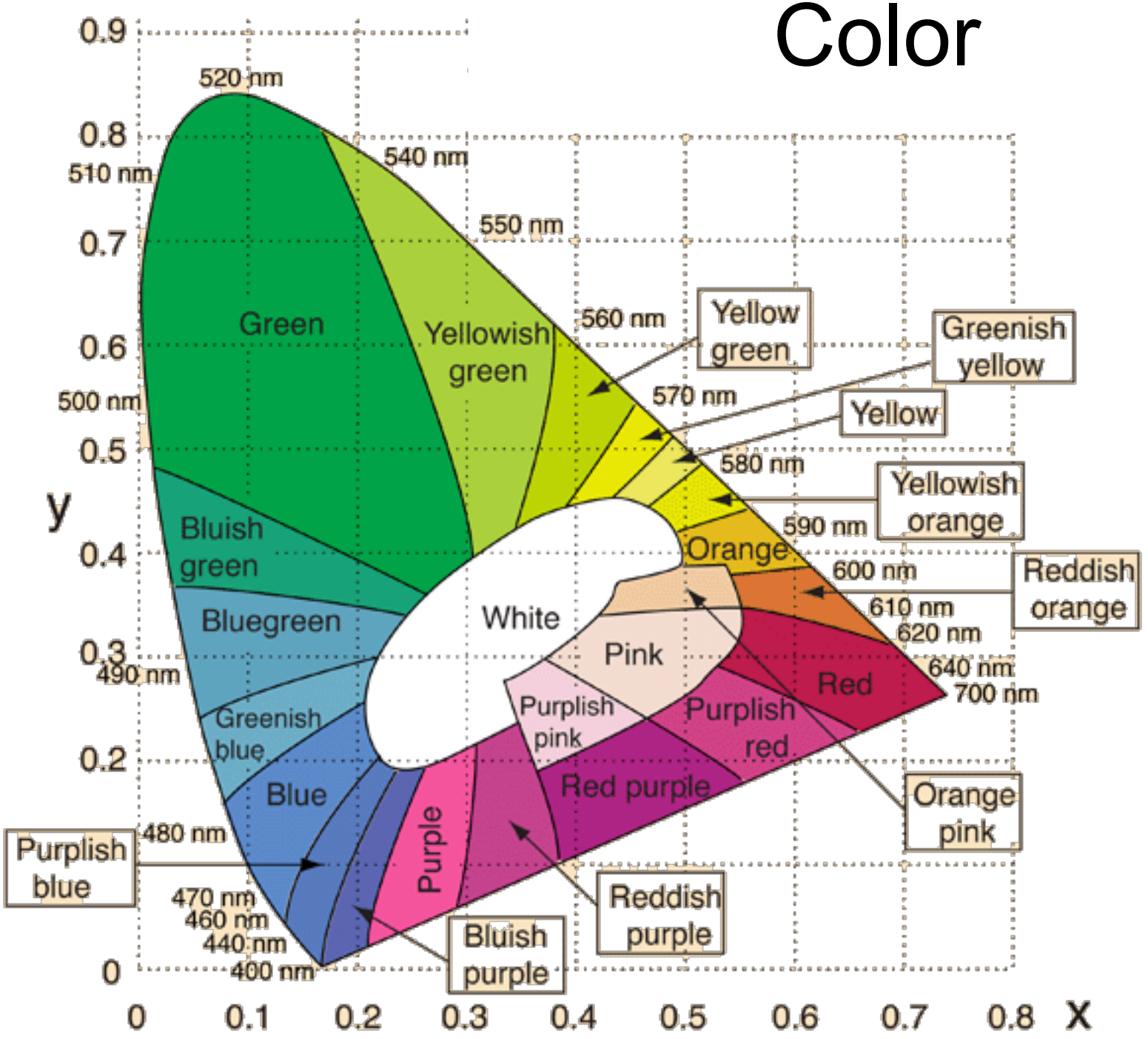
This one has more than one problem –

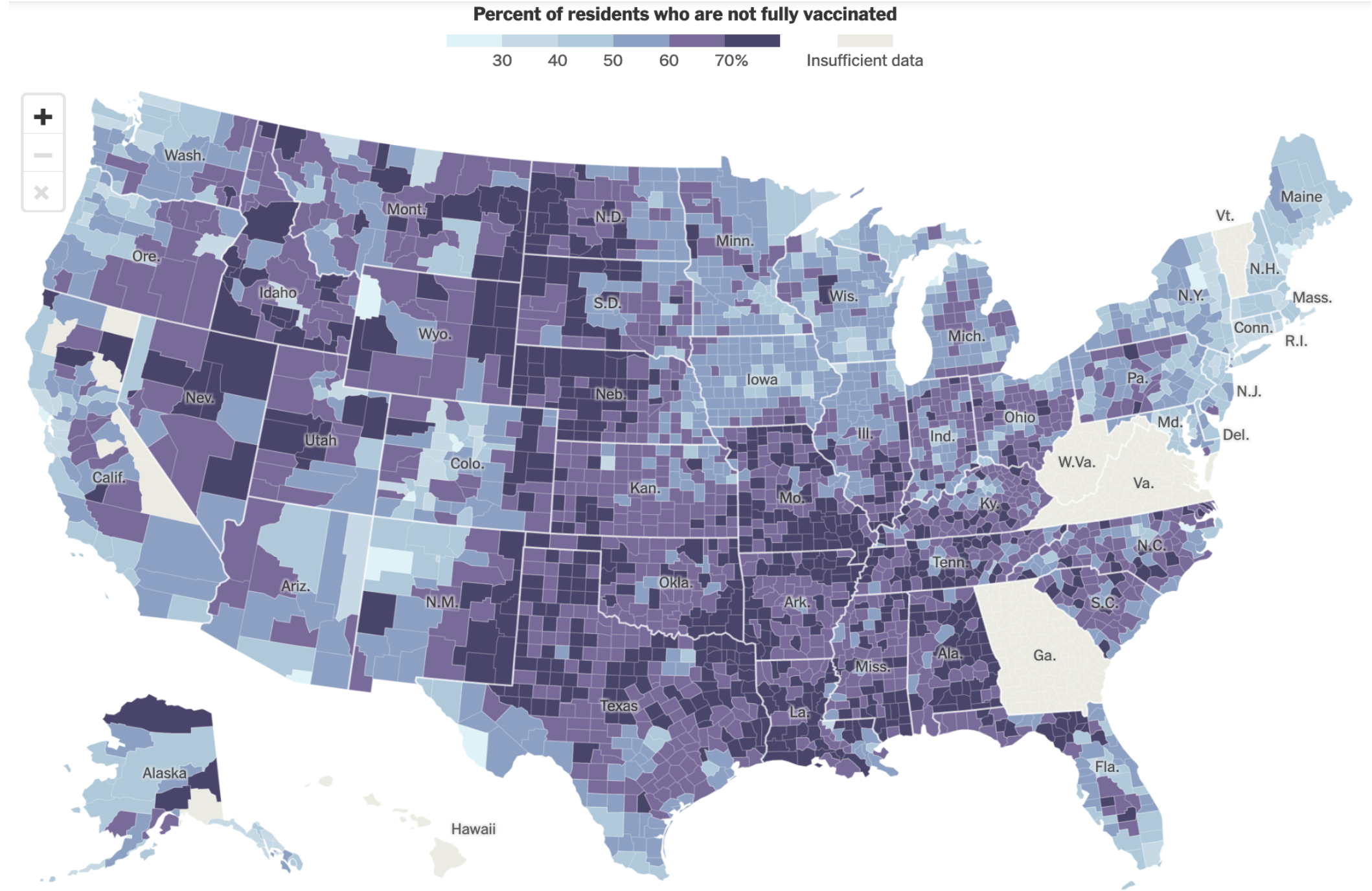
using visualization to obscure the truth rather than reveal it

Color


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
y
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
x
520 nm
540 nm
550 nm
560 nm
570 nm
580 nm
590 nm
600 nm
610 nm
620 nm
640 nm
700 nm
510 nm
500 nm
490 nm
480 nm
470 nm
460 nm
440 nm
400 nm
Green
Yellowish green
Yellow green
Greenish yellow
Yellow
Yellowish orange
Orange
Reddish orange
Red
Bluish green
Bluegreen
White
Pink
Purplish pink
Purplish red
Red purple
Greenish blue
Blue
Purple
Purplish blue
Orange pink
Reddish purple
Bluish purple

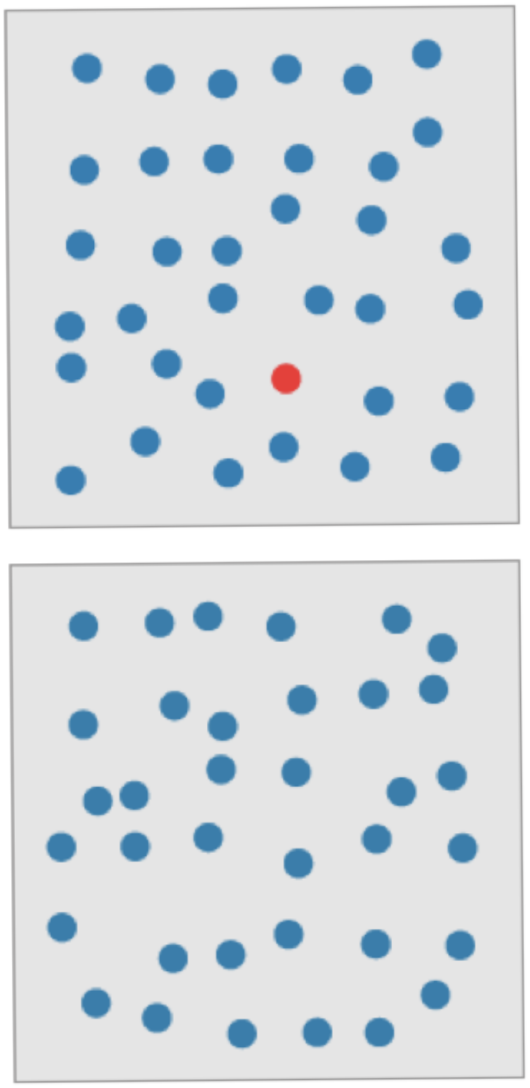
Color

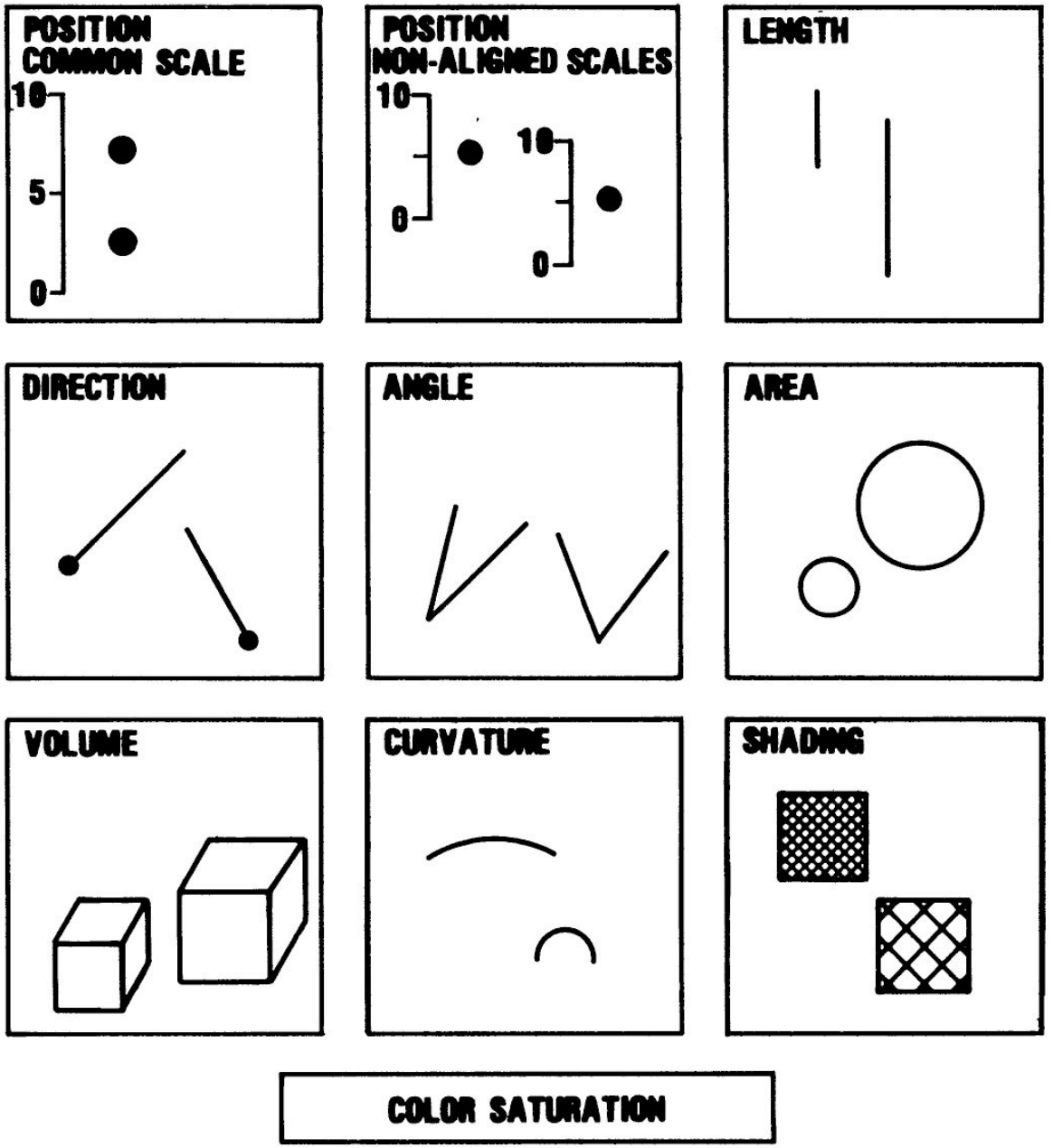

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
y
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
x
520 nm
540 nm
550 nm
560 nm
570 nm
580 nm
590 nm
600 nm
610 nm
620 nm
640 nm
700 nm
510 nm
500 nm
490 nm
480 nm
470 nm
460 nm
440 nm
400 nm
Green
Yellowish green
Yellow green
Greenish yellow
Yellow
Yellowish orange
Orange
Reddish orange
Red
Bluish green
Bluegreen
White
Pink
Purplish pink
Purplish red
Red purple
Greenish blue
Blue
Purple
Purplish blue
Orange pink
Reddish purple
Bluish purple



Note: No C.D.C. data available for Hawaii, Texas and some counties. Four additional states were excluded because more than a quarter of the data was missing. Data from Te and Colorado excludes shots given by most federal agencies. Data is as of July 29. • Sources: Centers for Disease Control and Prevention; Texas Department of State Healt Services; Colorado Department of Public Health & Environment; Massachusetts Department of Public Health; U.S. Census Bureau • By Ashley Wu and Albert Sun

Perception

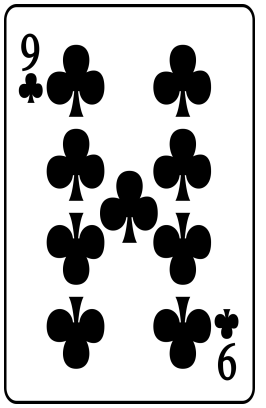



POSITION
COMMON SCALE
10
5
0
POSITION
NON-ALIGNED SCALES
10
0
10
0
LENGTH
DIRECTION
ANGLE
AREA
VOLUME
CURVATURE
SHADING
COLOR SATURATION

What is Visualization for?

- Document, record, preserve
- Analyze
  - Tool to help us understand
    - enhance memory
    - take advantage of perceptual preprocessing
  - Reveal peculiarities, limitations of data
- Communicate
  - Inform, persuade
  - Make decisions

Activity: Number Scrabble



- Two-player game played with cards A-9
- Players alternately choose one card
- First player to collect any three cards which sum to 15 wins
- If all cards are exhausted and no three cards held by one player sum to 15, game ends in a draw.

Herbert Simon, The Sciences of the Artificial, 1969

Number Scrabble example game

- Player A takes 8
- Player B takes 2
- Player A takes 4
- Player B takes 3
- Player A takes 5
- B to move?

Number Scrabble example game

- Player A takes 8
- Player B takes 2
- Player A takes 4
- Player B takes 3
- Player A takes 5
- B to move?

How much do you know about arithmetic?

Number Scrabble example game

- Player A takes 8
- Player B takes 2
- Player A takes 4
- Player B takes 3
- Player A takes 5
- B to move?

A: 8, 4, 5

B: 2, 3

Unclaimed: 1,6,7,9

Number Scrabble example game

A: 8, 4, 5

B: 2, 3

Unclaimed: 1,6,7,9

3 choose 2 = 3 pairs

A can try to play next

Number Scrabble example game

A: 8, 4, 5

B: 2, 3

Unclaimed: 1,6,7,9

8 + 4 + 3 = 15

8 + 5 + 2 = 15

4 + 5 + 6 = 15

Number Scrabble example game

A: 8, 4, 5

B: 2, 3

Unclaimed: 1,6,7,9

8 + 4 + 3 = 15
8 + 5 + 2 = 15
4 + 5 + 6 = 15

Number Scrabble example game

A: 8, 4, 5

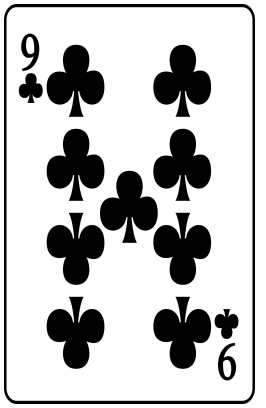
B: 2, 3

Unclaimed: 1,6,7,9

8 + 4 + 3 = 15
8 + 5 + 2 = 15
4 + 5 + 6 = 15

If I color them, B's only non-losing move is evident.

Activity: Number Scrabble



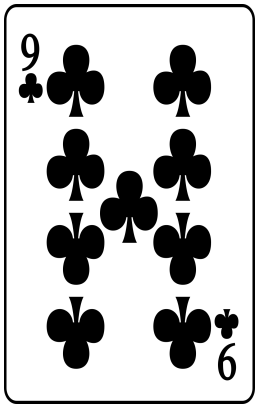
- There are eight triplets of distinct natural numbers between 1 and 9.that sum to 15.

1	5	9
1	6	8
2	4	9
2	5	8
2	6	7
3	4	8
3	5	7
4	5	6

- Four solutions involve 5, and four solutions are imbalanced around 5

Herbert Simon, The Sciences of the Artificial, 1969

Activity: Number Scrabble



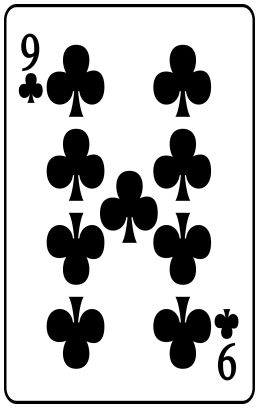
- There are eight triplets of distinct natural numbers between 1 and 9...

1	5	9
1	6	8
2	4	9
2	5	8
2	6	7
3	4	8
3	5	7
4	5	6

- 1, 3, 9, 7 have only two paths to victory
- 2, 4, 6, 8 have three each
- 5 is in four of the solutions

Herbert Simon, The Sciences of the Artificial, 1969

Activity: Number Scrabble



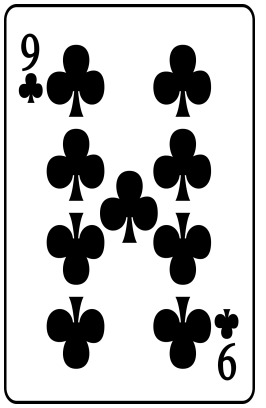
- There are eight triplets of distinct natural numbers between 1 and 9...

1	5	9
1	6	8
2	4	9
2	5	8
2	6	7
3	4	8
3	5	7
4	5	6

- 1, 3, 9, 7 have only two paths to victory
- 2, 4, 6, 8 have three each
- 5 is in four of the solutions

Herbert Simon, The Sciences of the Artificial, 1969

Activity: Number Scrabble



- 1, 3, 9, 7 have only two paths to victory
- 2, 4, 6, 8 have three each
- 5 is in four of the solutions

Herbert Simon, The Sciences of the Artificial, 1969

Number Scrabble example game

A: 8, 4, 5

B: 2, 3

Unclaimed: 1,6,7,9

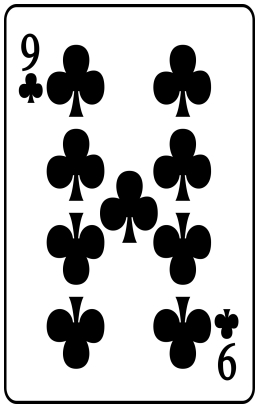
6	7	2
1	5	9
8	3	4

8 + 4 + 3 = 15

8 + 5 + 2 = 15

4 + 5 + 6 = 15

Activity: Number Scrabble



6	7	2
1	5	9
8	3	4

- Writing down the cards helped us see what was done, what was missing
- Coloring the numbers helped us see some group theory properties
- The tic-tac-toe representation turns the problem into a solved one.

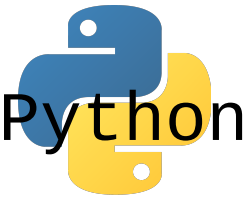
Visualization tools

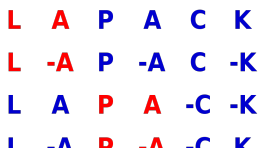
- Ubiquitous, limited to keypresses and mouse clicks
- Excellent for interacting with primates / getting data and judgements out of your head

- Commercial, free license for students, entirely graphical and purported to be easy to use.

- Browser-based free data wrangling power tool; find outliers, perform transforms on columns. Won't need for this class

The stack





Stratiographic layers -- lower layers are older, will likely last longer into the future.

The stack

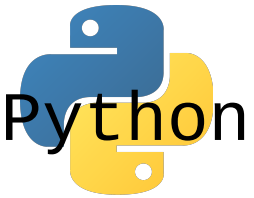

plotnine

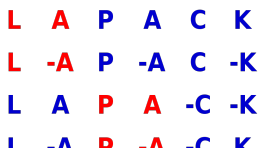

seaborn


matplotlib


pandas


NumPy


Python


L A P A C K
L -A P -A C -K
L A P A -C -K

The stack


plotnine


seaborn

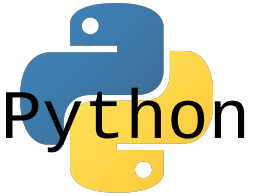

Vega-Lite

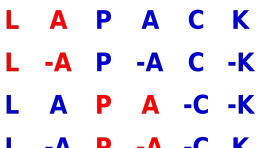

matplotlib


pandas


Vega


NumPy

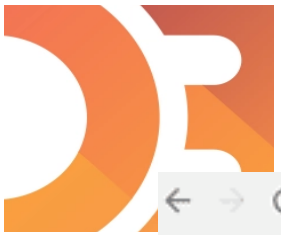

Python

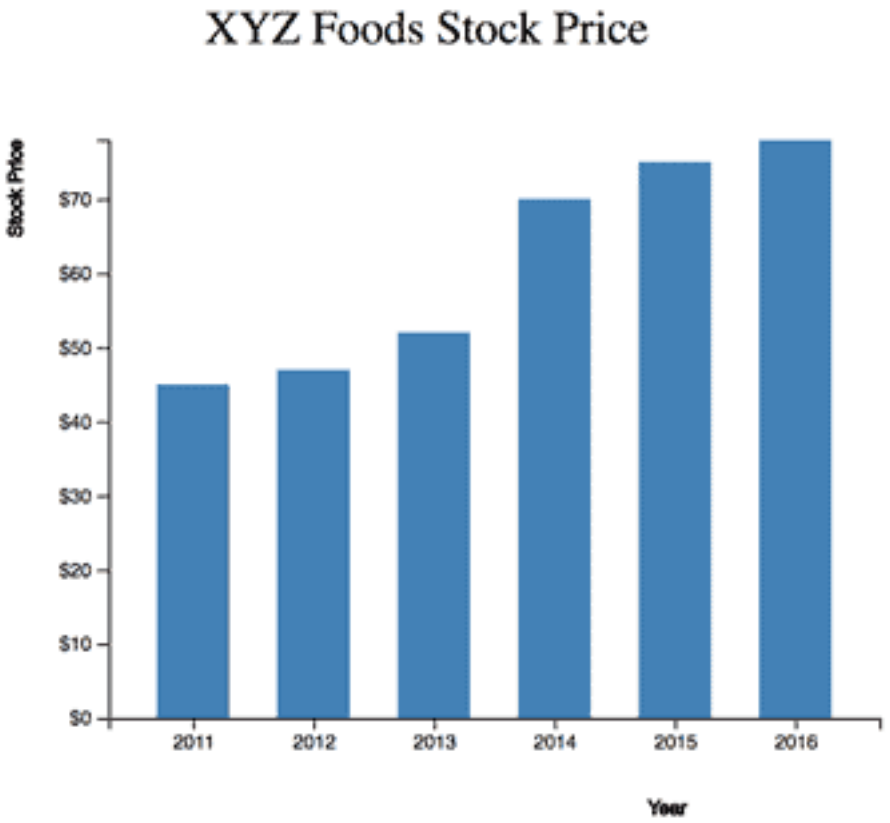

L A P A C K
L -A P -A C -K
L A P A -C -K


JS

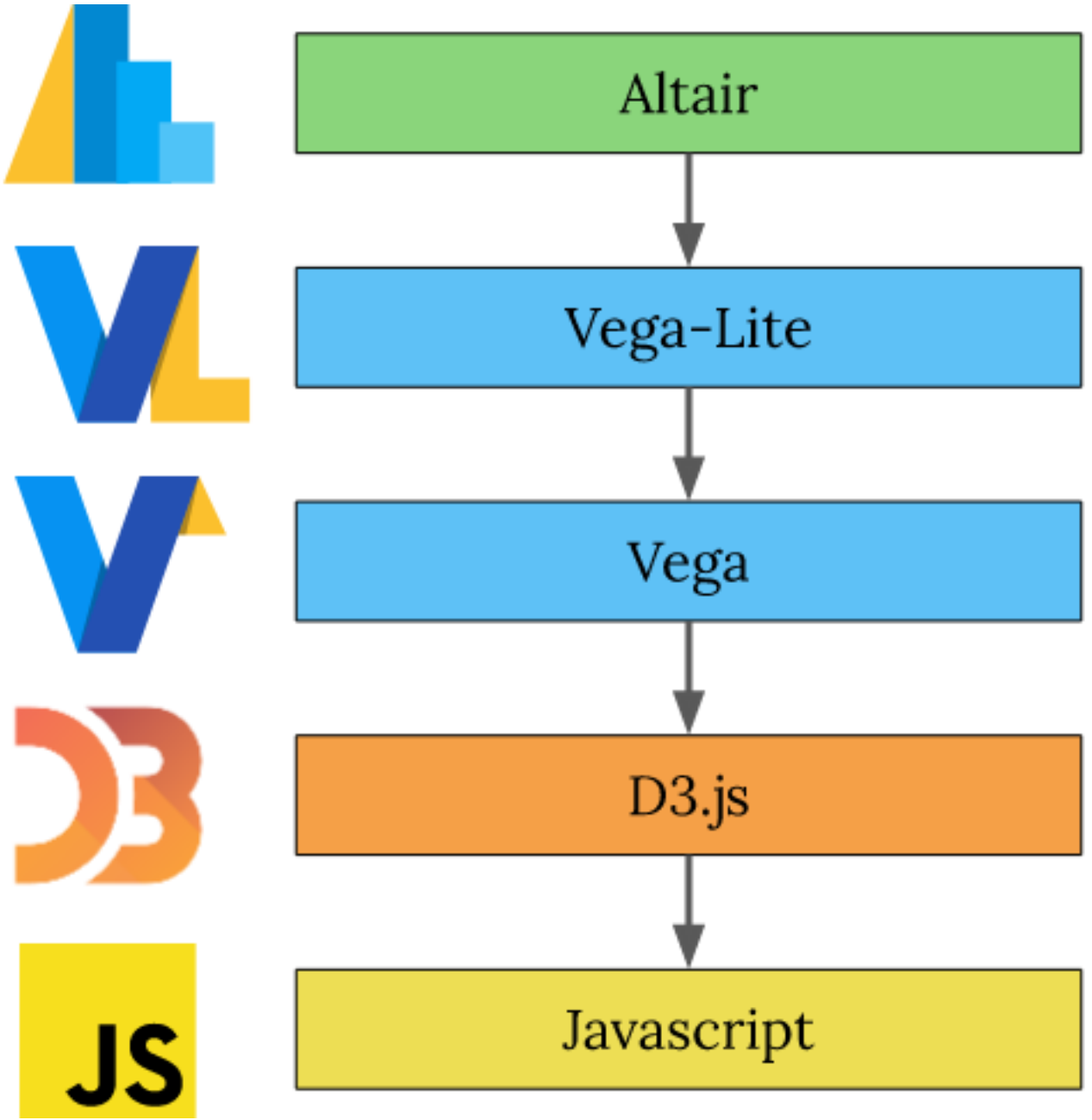
```
<!doctype html>
<html>
<head>
    <style>
        .bar {
            fill: steelblue;
        }
    </style>
    <script src="https://d3js.org/d3.v4.min.js"></script>
<body>
<svg width="600" height="500"></svg>
<script>
    var svg = d3.select("svg"),
        margin = 200,
        width = svg.attr("width") - margin,
        height = svg.attr("height") - margin
    svg.append("text")
       .attr("transform", "translate(100,0)")
       .attr("x", 50)
       .attr("y", 50)
       .attr("font-size", "24px")
       .text("XYZ Foods Stock Price")
    var xScale = d3.scaleBand().range([0,
width]).padding(0.4),
        yScale = d3.scaleLinear().range([height, 0]);
    var g = svg.append("g")
               .attr("transform", "translate(" + 100 + ","
+ 100 + ")");
    d3.csv("XYZ.csv", function(error, data) {
        if (error) {
            throw error;
        }
        xScale.domain(data.map(function(d) { return
d.year; }));
        yScale.domain([0, d3.max(data, function(d)
{ return d.value; })]);
        g.append("g")
         .attr("transform", "translate(0," + height + ")")
         .call(d3.axisBottom(xScale))
         .append("text")
         .attr("y", height - 250)
         .attr("x", width - 100)
         .attr("text-anchor", "end")
         .attr("stroke", "black")
         .text("Year");
        g.append("g")
         .call(d3.axisLeft(yScale).tickFormat(function(d){
             return "$" + d;
         })
         .ticks(10))
         .append("text")
         .attr
         .attr
         .attr
         .attr
         .attr
         .text
        g.selec
         .data(data)
         .enter().append("rect")
         .attr("class", "bar")
         .attr("x", function(d) { return
xScale(d.year); })
         .attr("y", function(d) { return
yScale(d.value); })
         .attr("width", xScale.bandwidth())
         .attr("height", function(d) { return height -
yScale(d.value); });
    });
</script>
</body>
</html>
```

This is the sample code for a bar graph





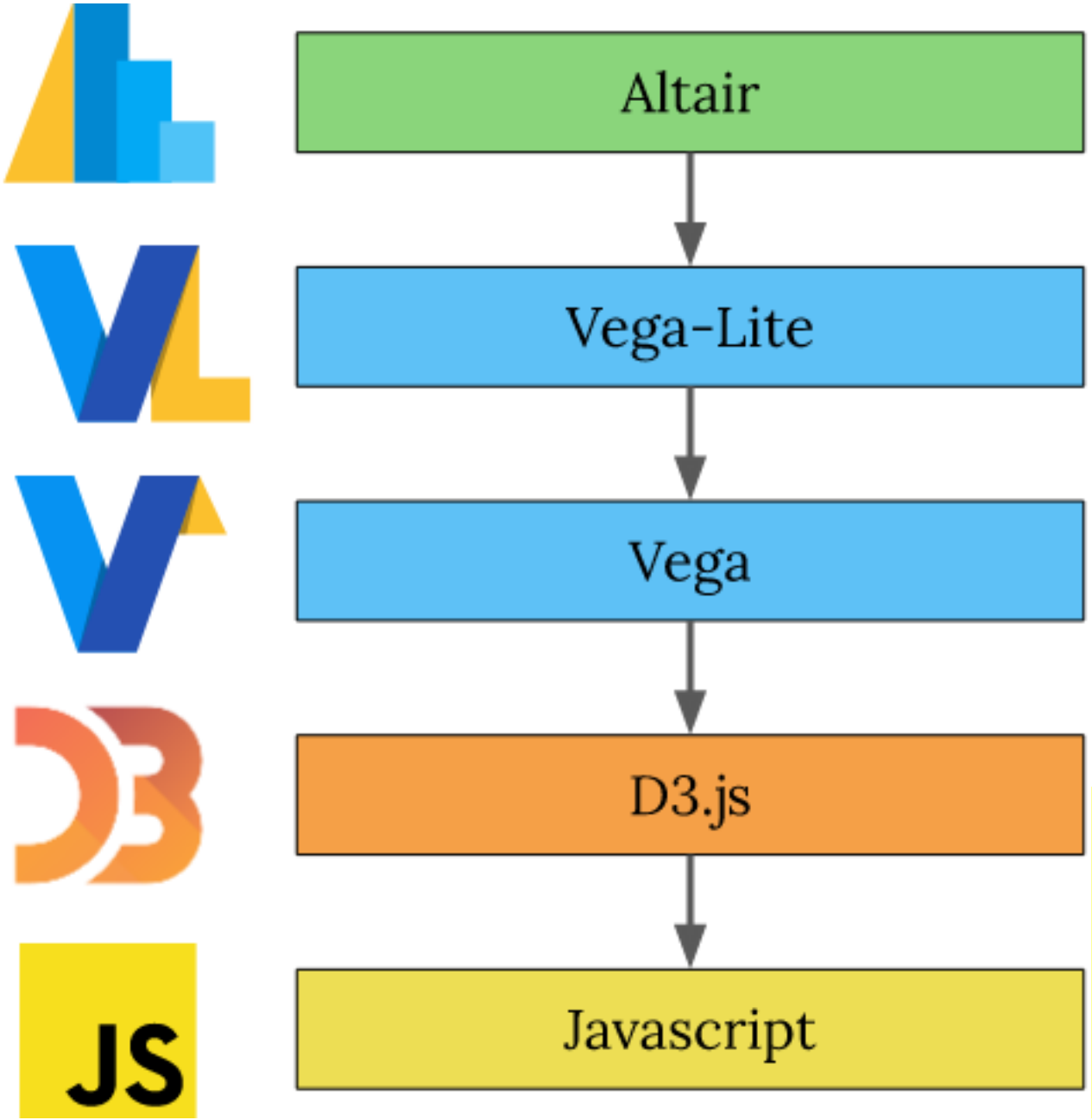

```
<!doctype ht
<html>
<head>
    <style>
        .bar

        }
    </style>
    <script
<body>
<svg width="
<script>
    var svg
        marg
        widt
        heig
    svg.appe
       .attr
       .attr
       .attr
       .attr
       .text
    var xSca
width]).padd
        ySca
    var g =

+ 100 + ")")
    d3.csv("
        if (error) {
            throw error;
        }
        xScale.domain(data.map(function(d) { return
d.year; }));
        yScale.domain([0, d3.max(data, function(d)
{ return d.value; })]);
        g.append("g")
                + height + ")")

mat(function(d){

            .attr("width", xScale.bandwidth())
            .attr("height", function(d) { return height -
yScale(d.value); });
    });
</script>
</body>
</html>
```




```
FacetedEncoding({
  x: X({
    field: 'Year',
    type: 'temporal'
  }),
  y: Y({
    aggregate: 'count',
    type: 'quantitative'
  })
})
```

This is a specification for a bar graph in Vega-lite

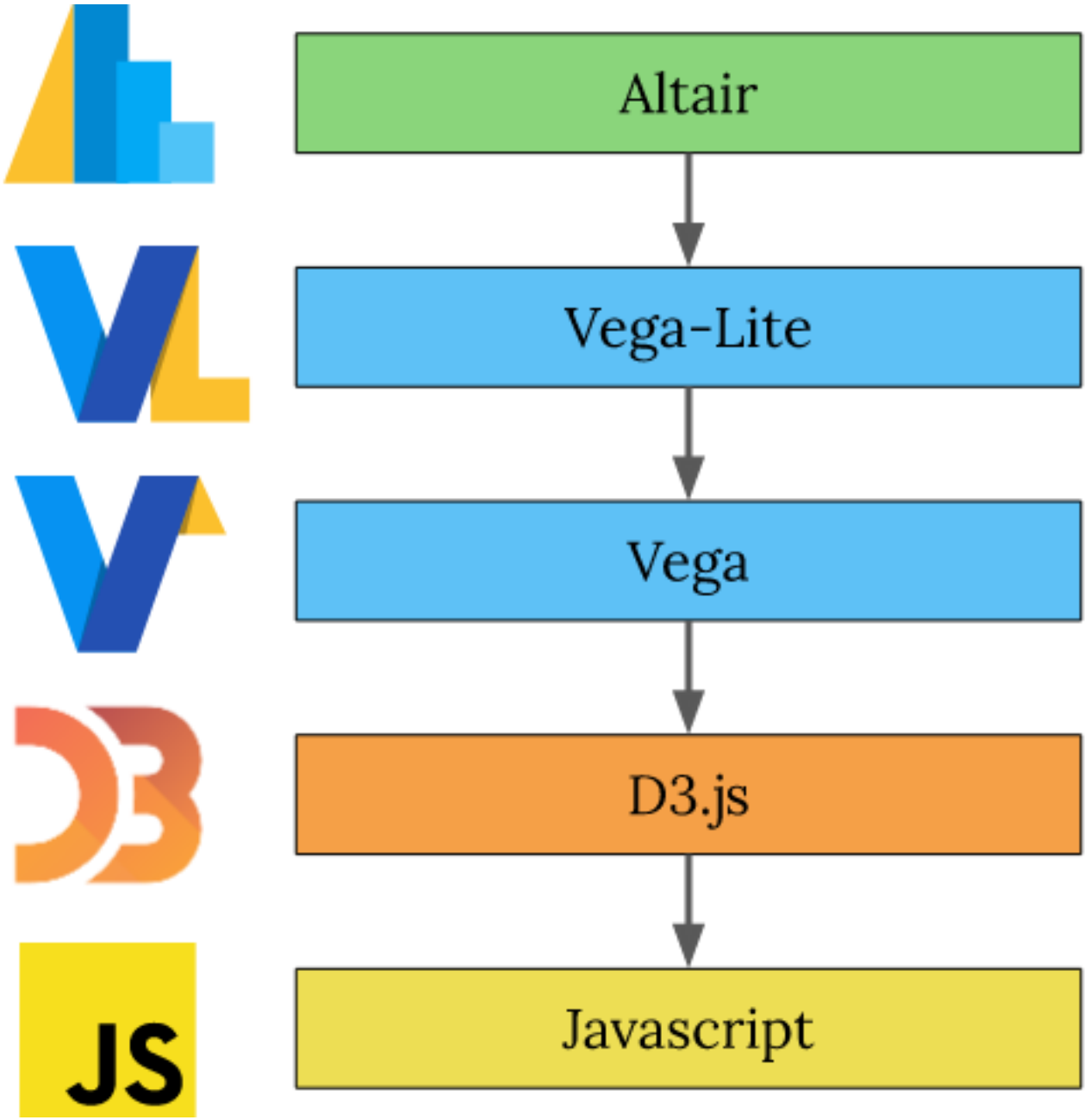
https://eitanlees.github.io/altair-stack/




```
alt.Chart(cars).mark_bar()
.encode(
    alt.X('Year'),
    alt.Y('count()'))
```

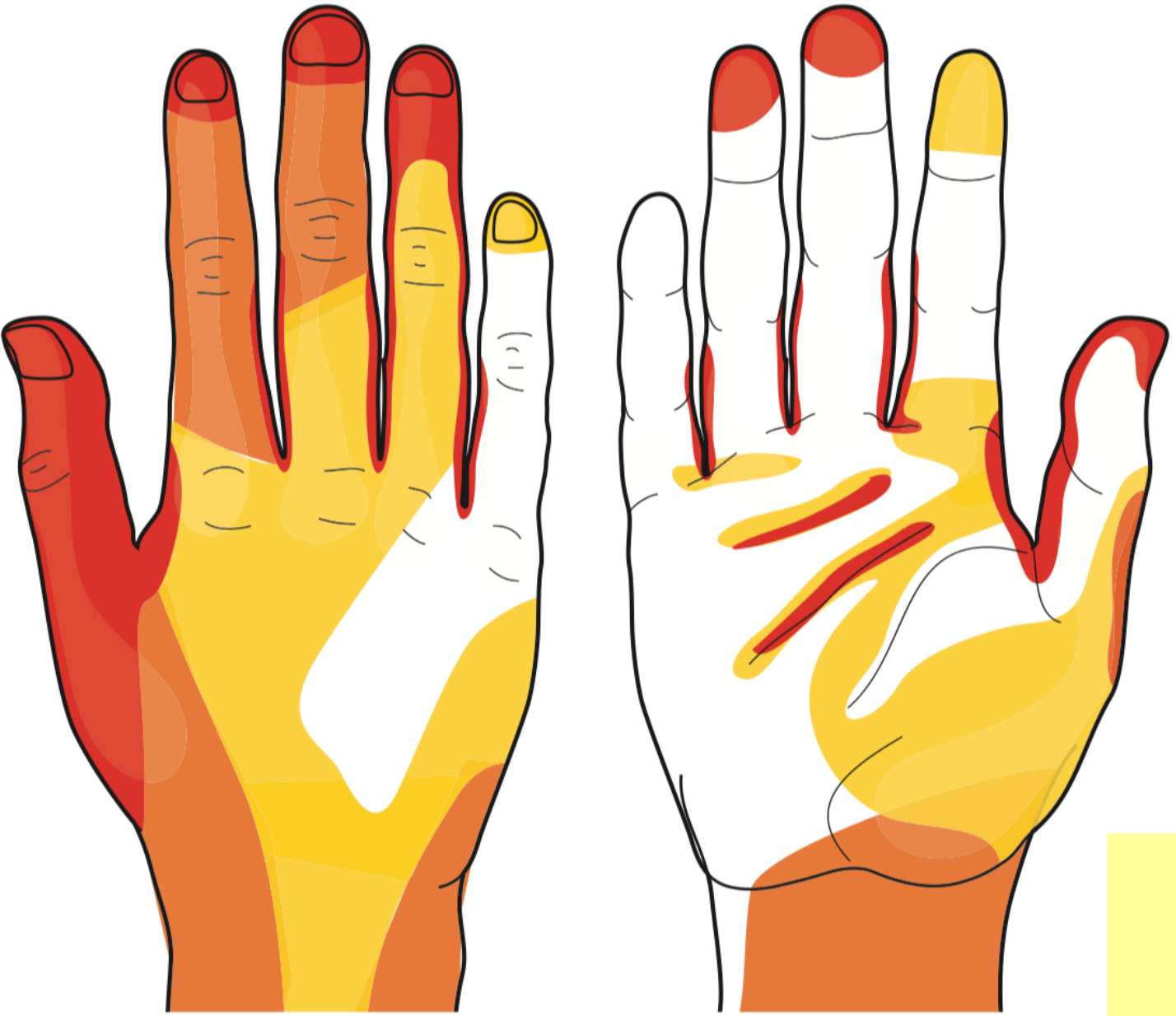
And here is the specification in altair (python wrapper for vega-lite)

https://eitanlees.github.io/altair-stack/



If I find out that altair/vega-lite really isn't working out by 2nd week we can move back to pure python.

https://eitanlees.github.io/altair-stack/



MOST OFTEN MISSED AREAS

OFTEN MISSED AREAS

LESS OFTEN MISSED AREAS

Reference: Taylor, L.J. An evaluation of handwashing techniques. *Nursing Times*. January 1978.

Beautiful, easy-to-understand graphic for these interesting times.

Administrivia

- Office hours
  - WT: 1-2:30 Wednesday,
  - WT: 11-12 Thursday, Friday; other times by appointment
  - AM:
  - AS:
- 7 homeworks, due on Friday 11:59 pm. First one friday of 2nd week.

Teaching Staff

Teaching Assistants

Anna Moise

Abby Star

Professor

Will Trimble
Instructor, Data Science
W 1-2:30
Th 11-212
F 11-12

This and that

TAs:

Abby Starr <abigailstarr@uchicago.edu>

Anna Moise <amoise@uchicago.edu>

- 10 minutes to pandas which takes way more than 10 minutes:
- https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html