Designing Visualizations

INFO 201

Joel Ross

Winter 2017

https://slides.com/joelross/info201w17-visualization/live

Final Project Description

Today's Objectives

By the end of class, you should be able to

- Identify the purpose and usage of data visualizations
- Assess the effectiveness of different **visual encodings**
- Organize data in the proper **shape**
- Plot geographic data on **maps**

London, 1854

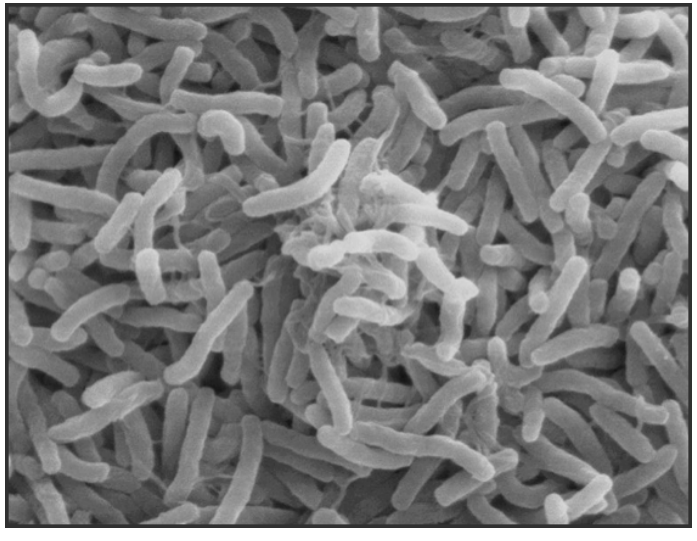
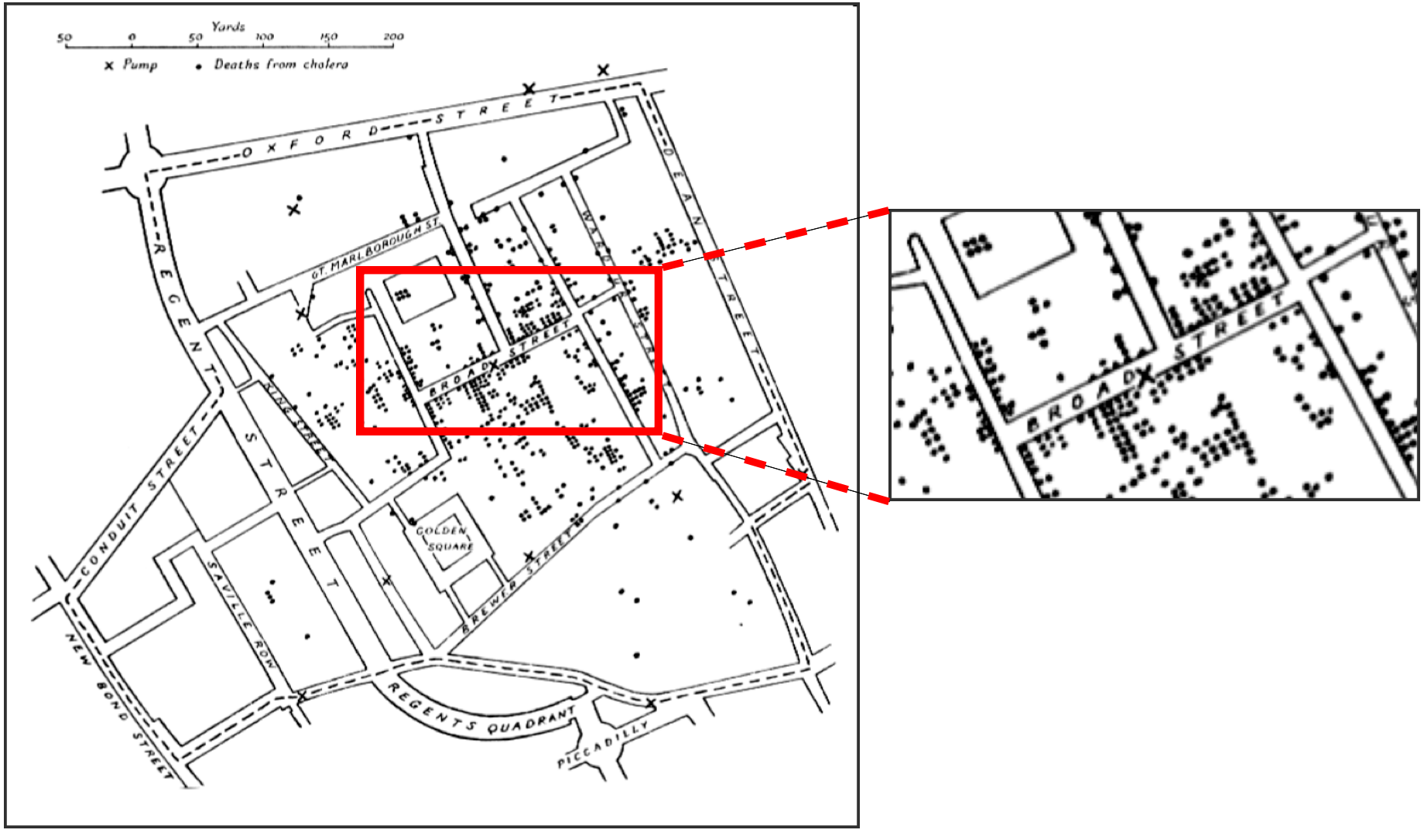


TABLE I

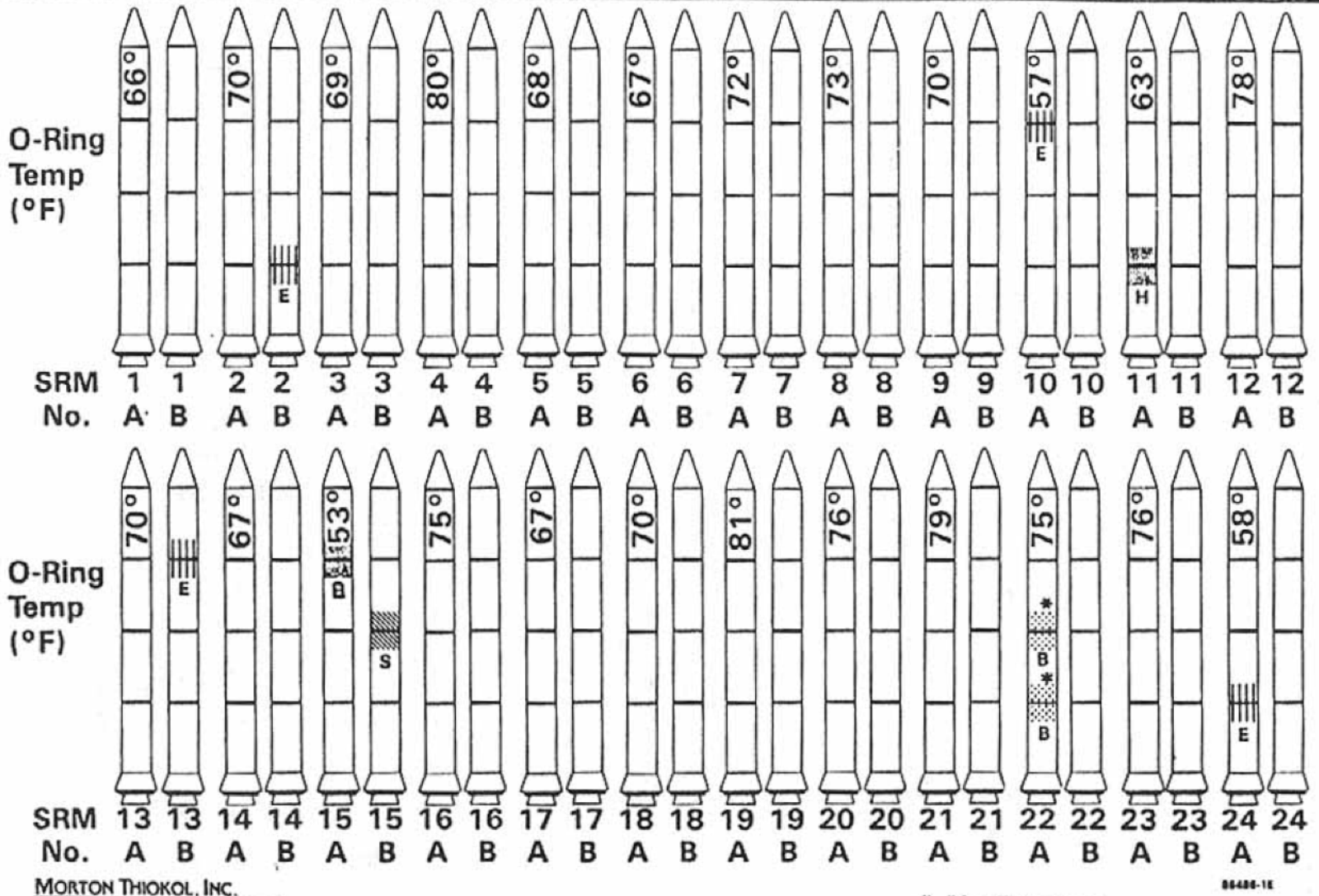
Date.		No. of Fatal Attacks.	Deaths
August	19	1	1
"	20	1	0
"	21	1	2
"	22	0	0
"	23	1	0
"	24	1	2
"	25	0	0
"	26	1	0
"	27	1	1
"	28	1	0
"	29	1	1
"	30	8	2
"	31	56	3
Sept.	1	143	70
"	2	116	127
"	3	54	76
"	4	46	71
"	5	36	45
"	6	20	37
"	7	28	32

John Snow


50 0 50 100 150 200 Yards
× Pump • Deaths from cholera
OXFORD STREET
REGENT STREET
GT. MARLBOROUGH ST.
BROAD STREET
GOLDEN SQUARE
BREWER STREET
CONDUIT STREET
SAVILLE ROW
NEW BOND STREET
REGENTS QUADRANT
PICCADILLY
DEAN STREET

Challenger Explosion (1986)

History of O-Ring Damage in Field Joints (Cont)



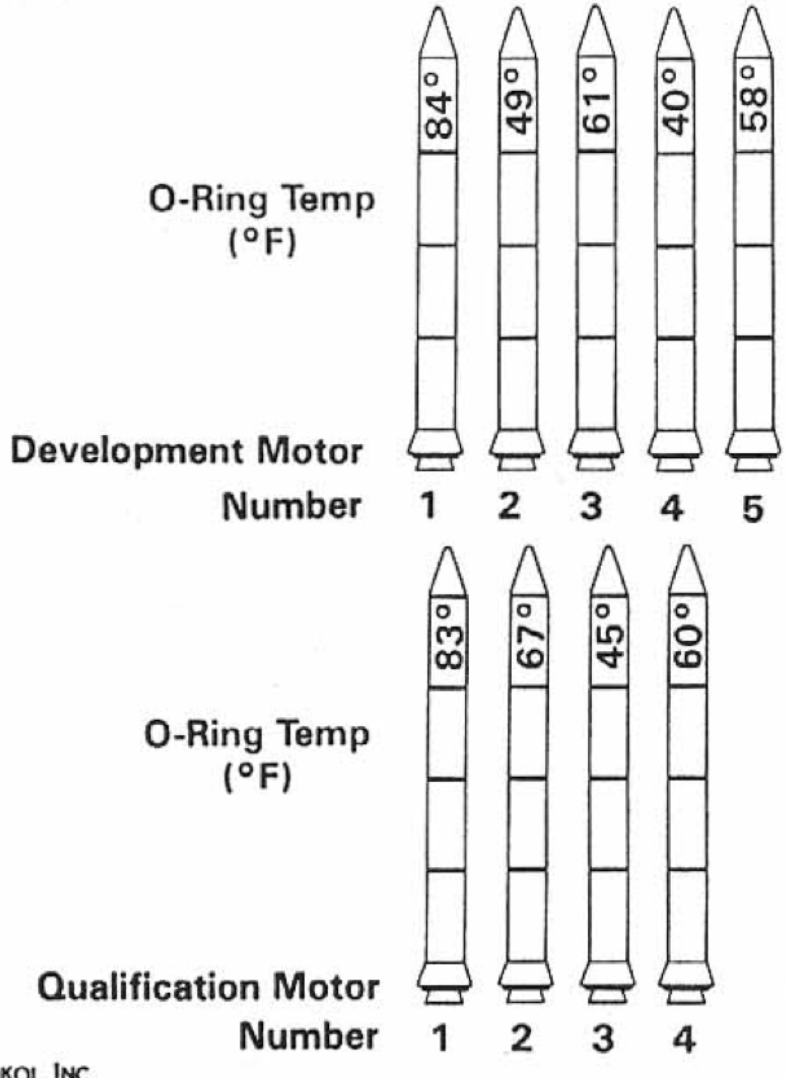
MORTON THIOKOL, INC.
Wasatch Operations

* No Erosion

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

[Ref. 2/26-2 2 of 3]

History of O-Ring Damage in Field Joints



Code

- S = Heating of Secondary O-Ring
- B = Primary O-Ring Blowby
- E = Primary O-Ring Erosion
- H = Heating of Primary O-Ring
- = No Damage

STATIC TEST MOTORS

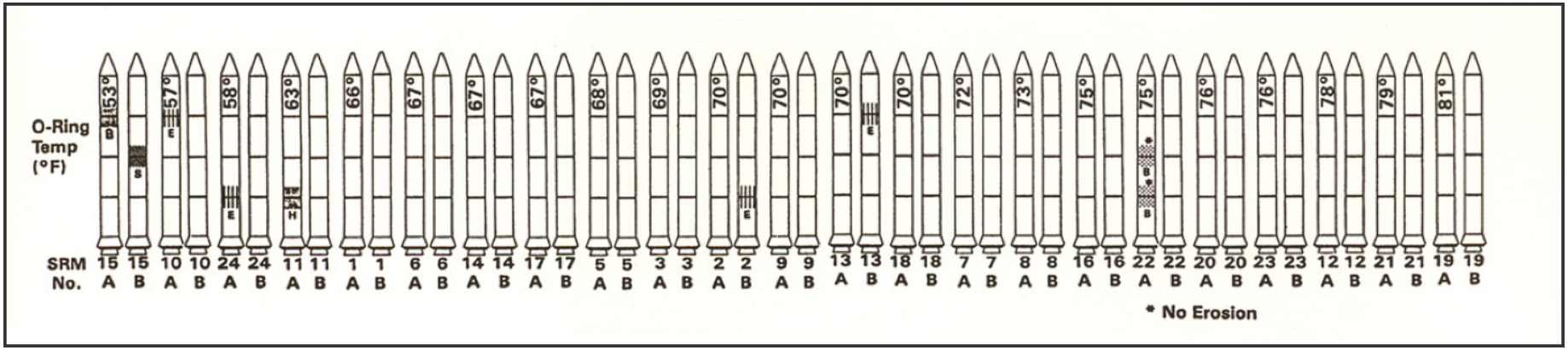
- HORIZONTAL ASSEMBLY
- SOME PUTTY REPAIRED

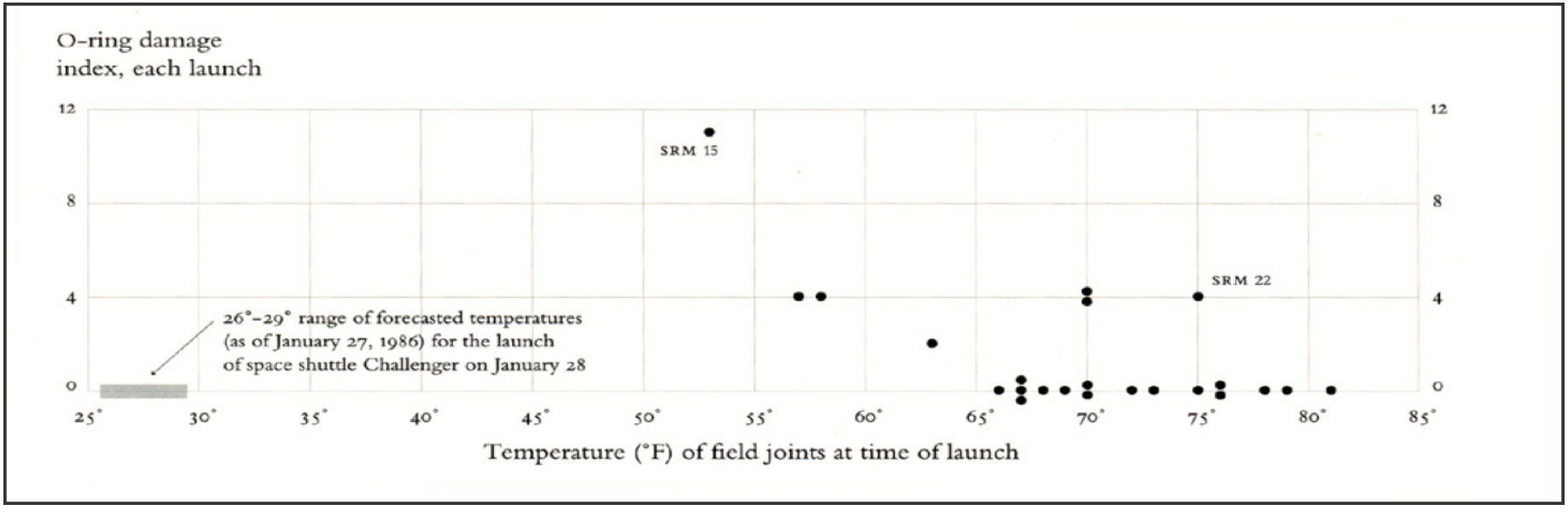
MORTON THIOKOL, INC.
Wasatch Operations

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

[Ref. 2/26-2 1 of 3]

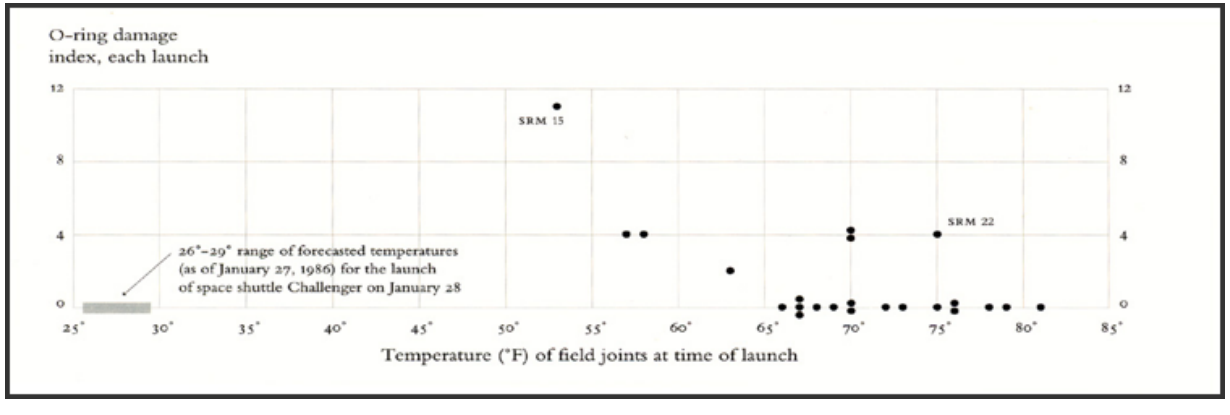
Tufte's Revision





Reasoning (Exploration)

Communication (Explanation)



Impact

Example: how segregated are U.S. cities?

http://vallandingham.me/racial_divide/

Visualizations **encode** data to represent it visually

Rene Magritte, *The Treachery of Images*

Importance of Representations

Let's play a game!

- Have 9 digits: **1 2 3 4 5 6 7 8 9**
- Two players take turns selecting a digit
  - Each digit can be selected only once
- First player whose digits sum to **15** wins!

A takes **8**, B takes **2**, A takes **4**, B takes **3**, A takes **5**

What should B do?!

Importance of Representations

4	3	8
9	5	1
2	7	6

4	3	8
9	5	1
2	7	6

X	O	X
	X	
O		

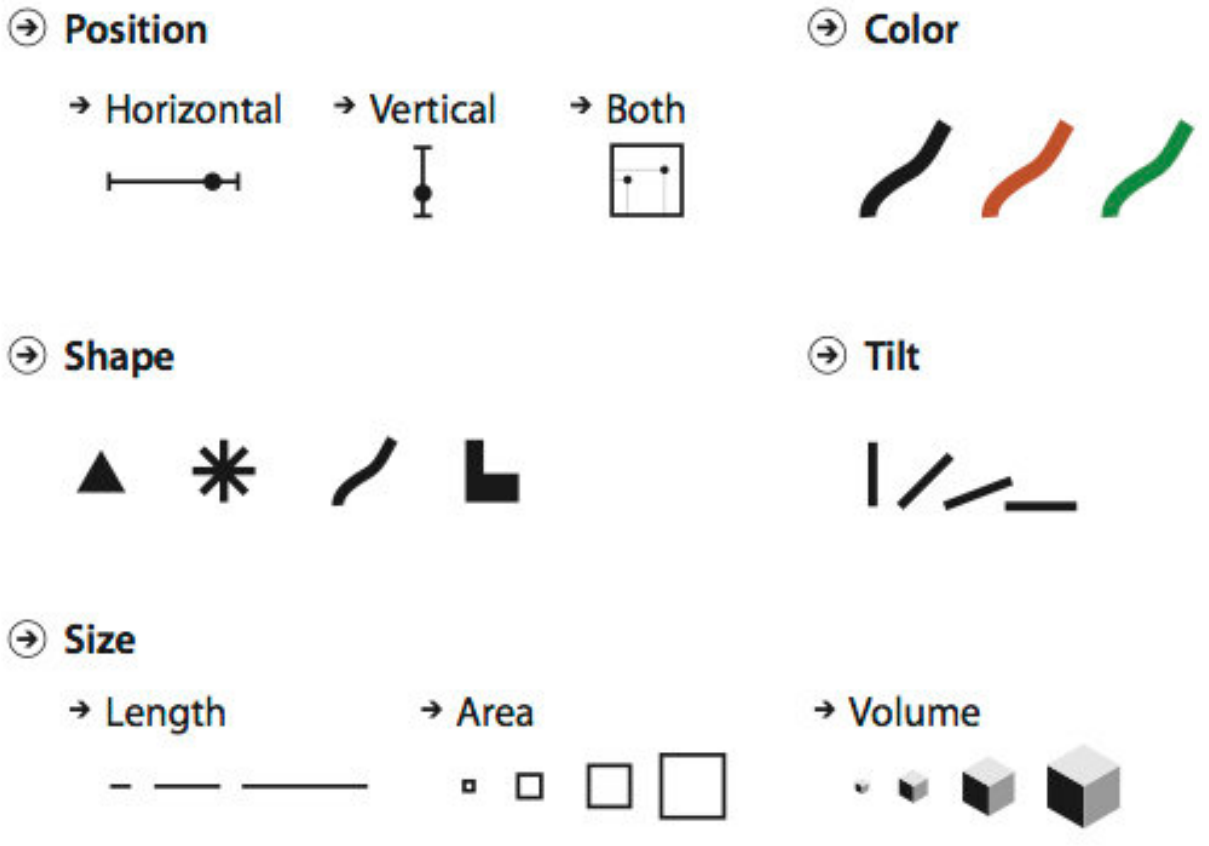
What should B do?!

"A good representation captures the essential elements of the event, deliberately leaving out the rest"

- Donald Norman

Visual Channels

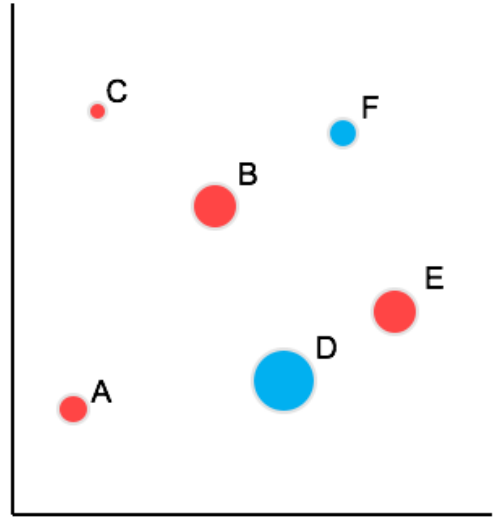
A visual property that can communicate meaning. An **aesthetic** that data can map to.



Aesthetic Mappings

Dimensions

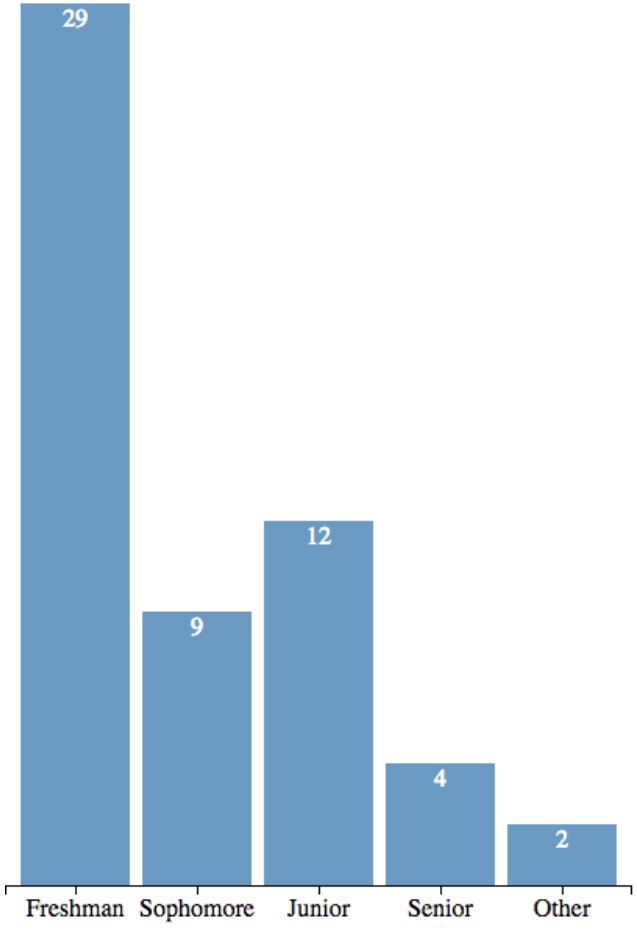
Petal length	Sepal length	Plant height	Species
1	2	3	Setosa
4	2.5	10	Mimosa
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…

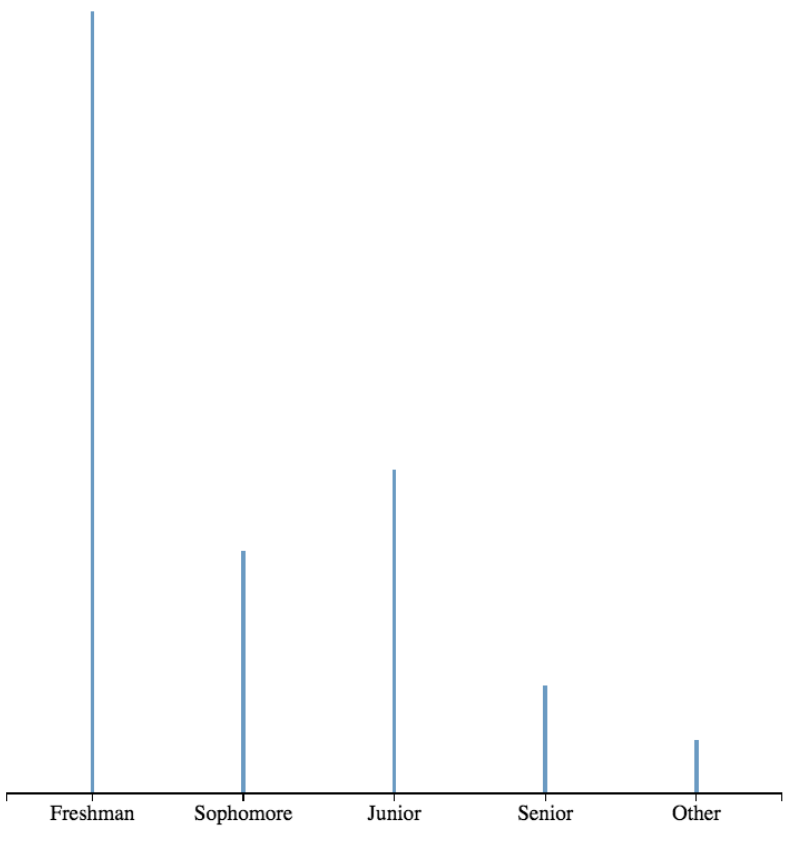


x-position y-position size color

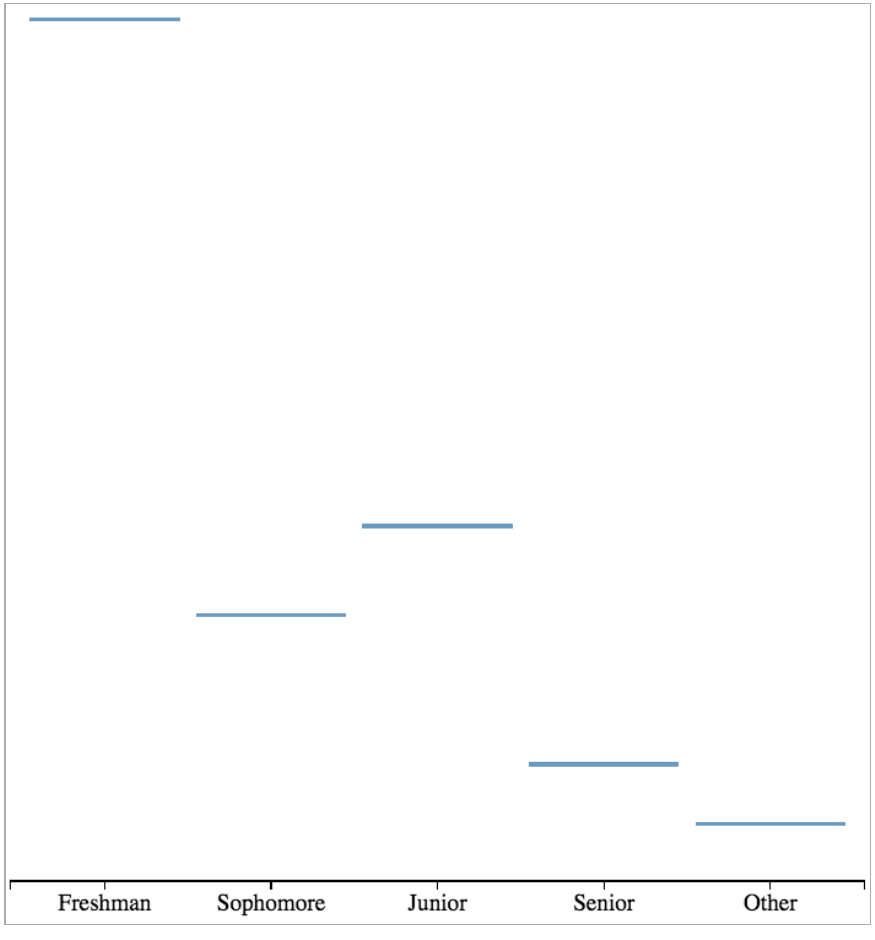
Visual Channels

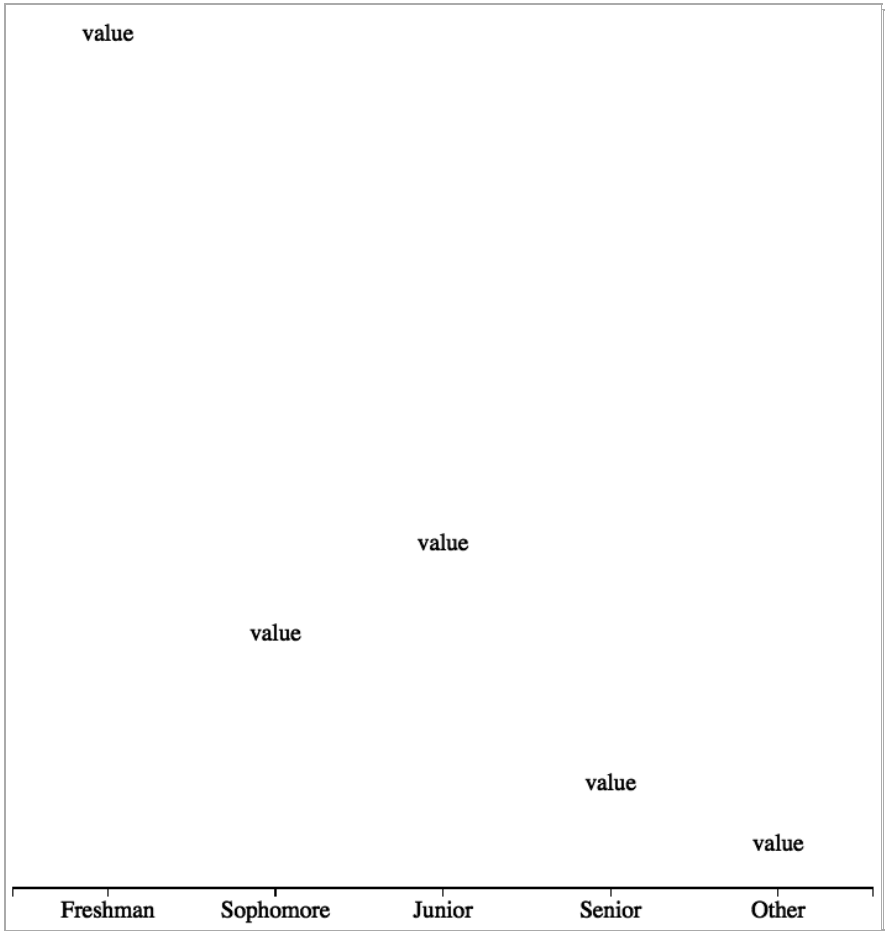
Different Encodings

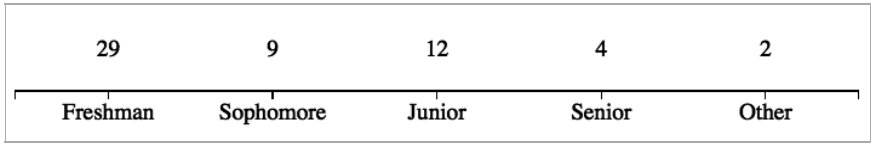




Different Encodings







What makes a good encoding?

Mackinlay's Expressiveness Criteria

A set of facts is **expressible** in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Unexpressive Visualizations

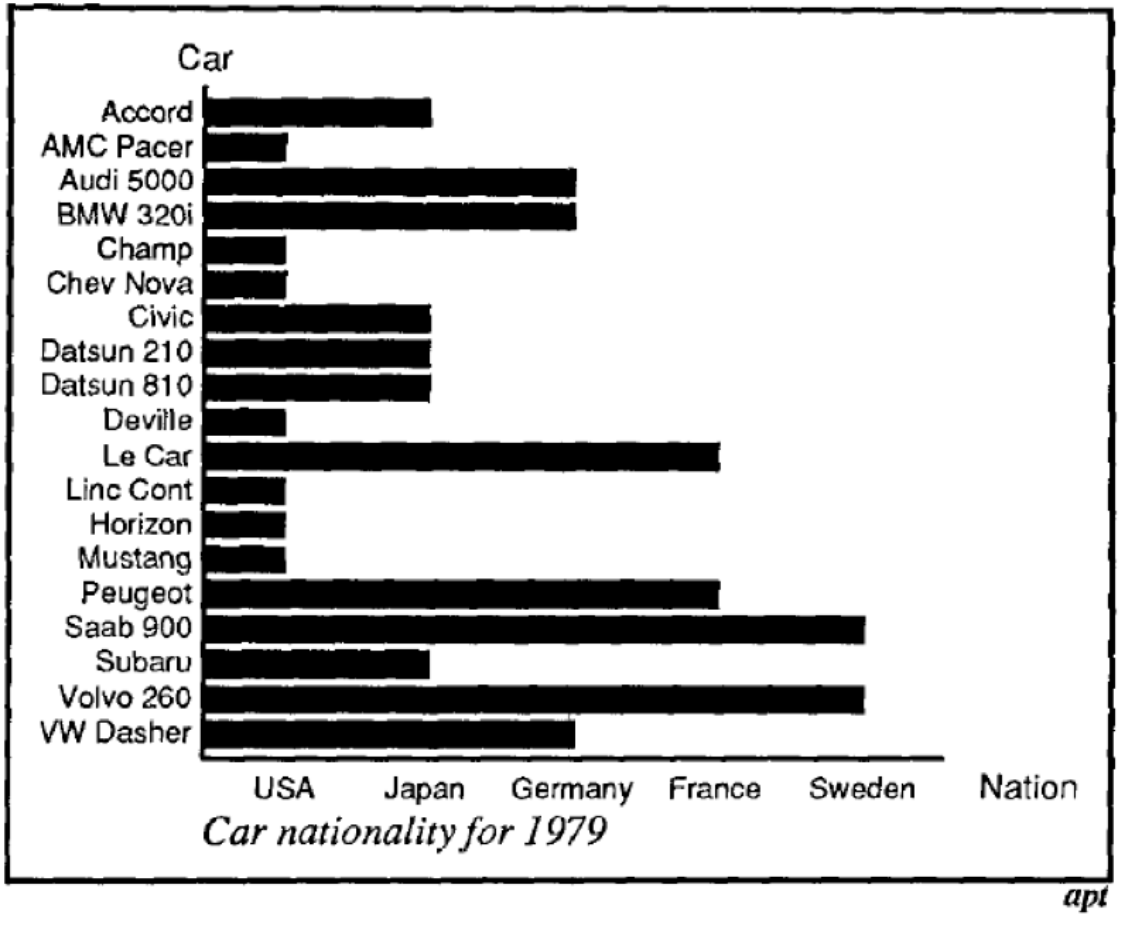


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

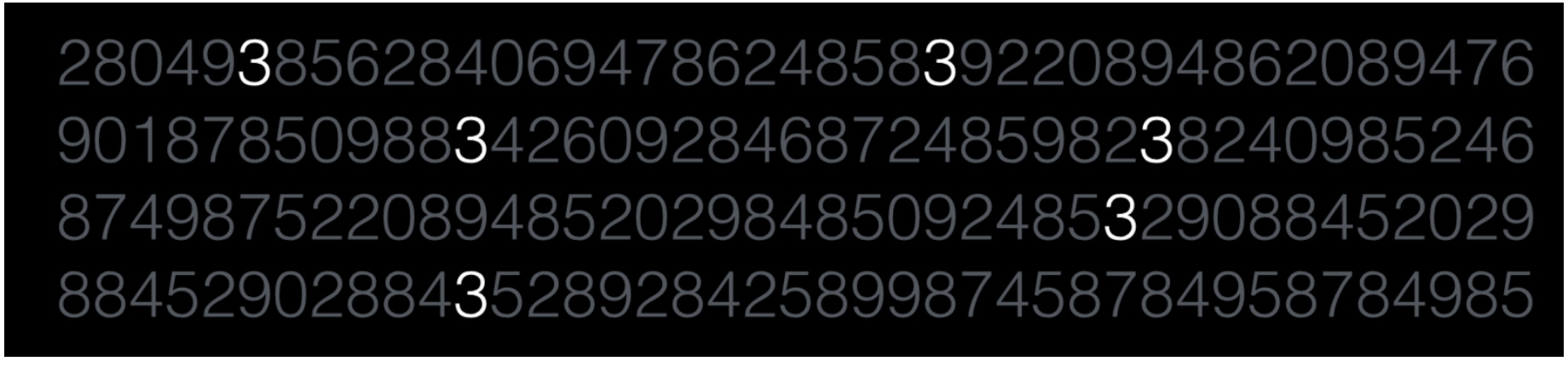
(Mackinlay, 1986)

Mackinlay's Effectiveness Criteria

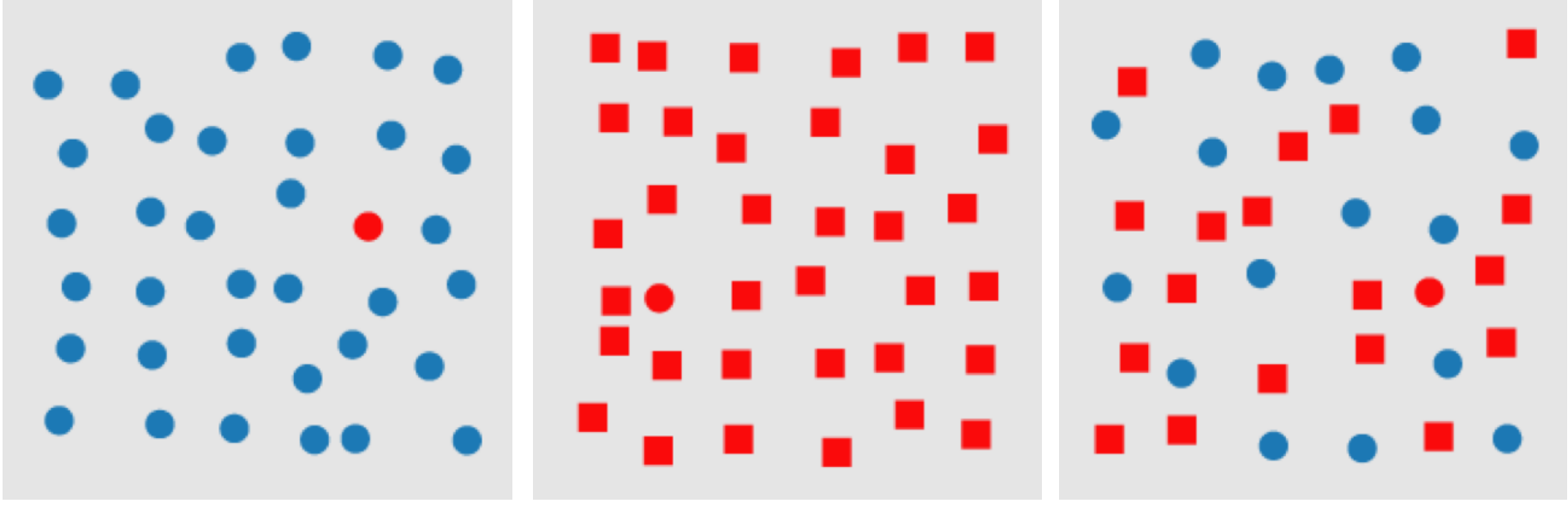
A visualization is more **effective** than another visualization if the information conveyed by one visualization is more readily **perceived** than the information in the other visualization.

```
28049385628406947862485839220894862089476
90187850988342609284687248598238240985246
87498752208948520298485092485329088452029
88452902884352892842589987458784958784985
```

How many 3's are there?

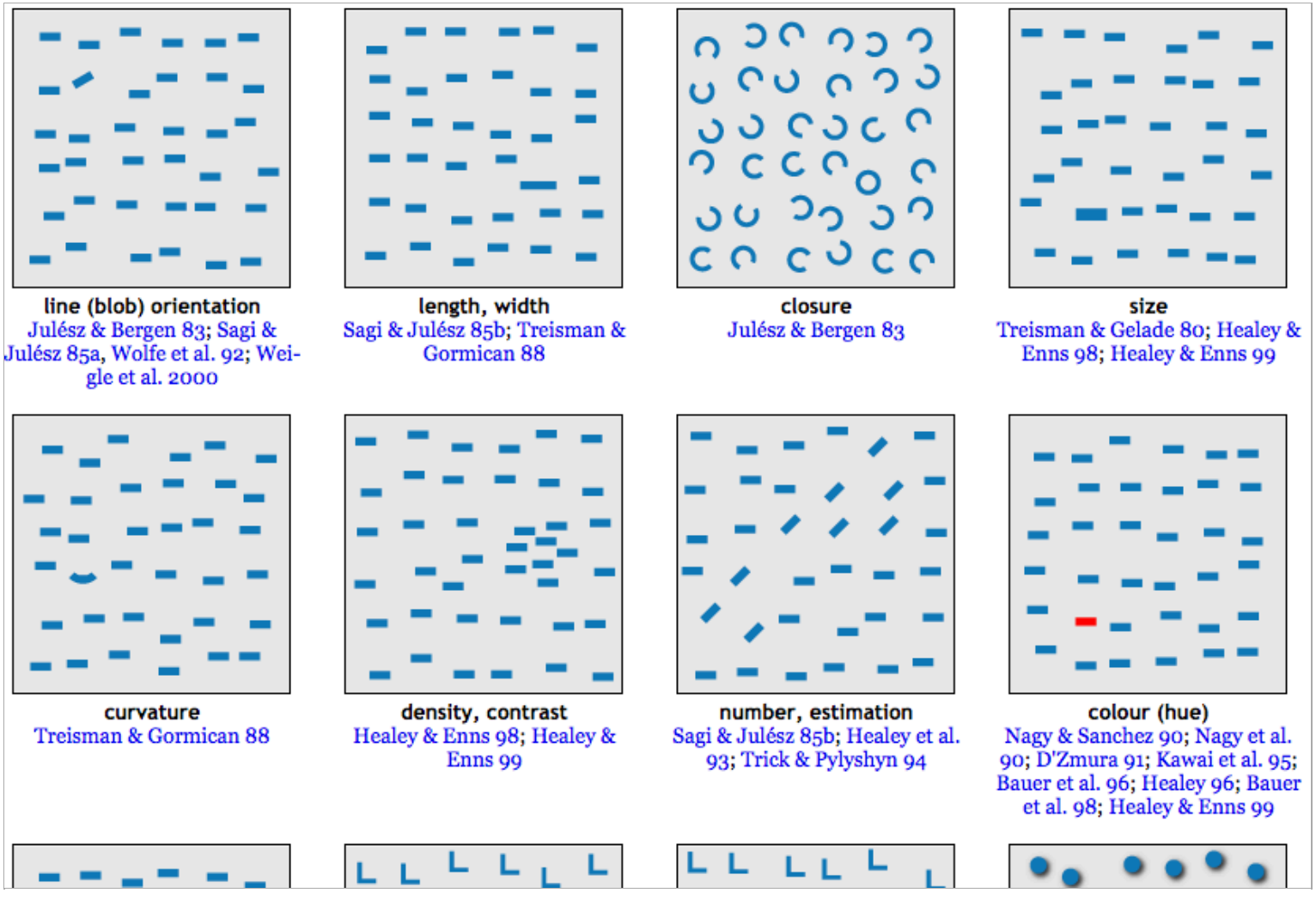


How many 3's are there?



Find the red dot

Pre-attentive Visual Perception



http://www.csc.ncsu.edu/faculty/healey/PP/

How do we choose a visual channel to map to?

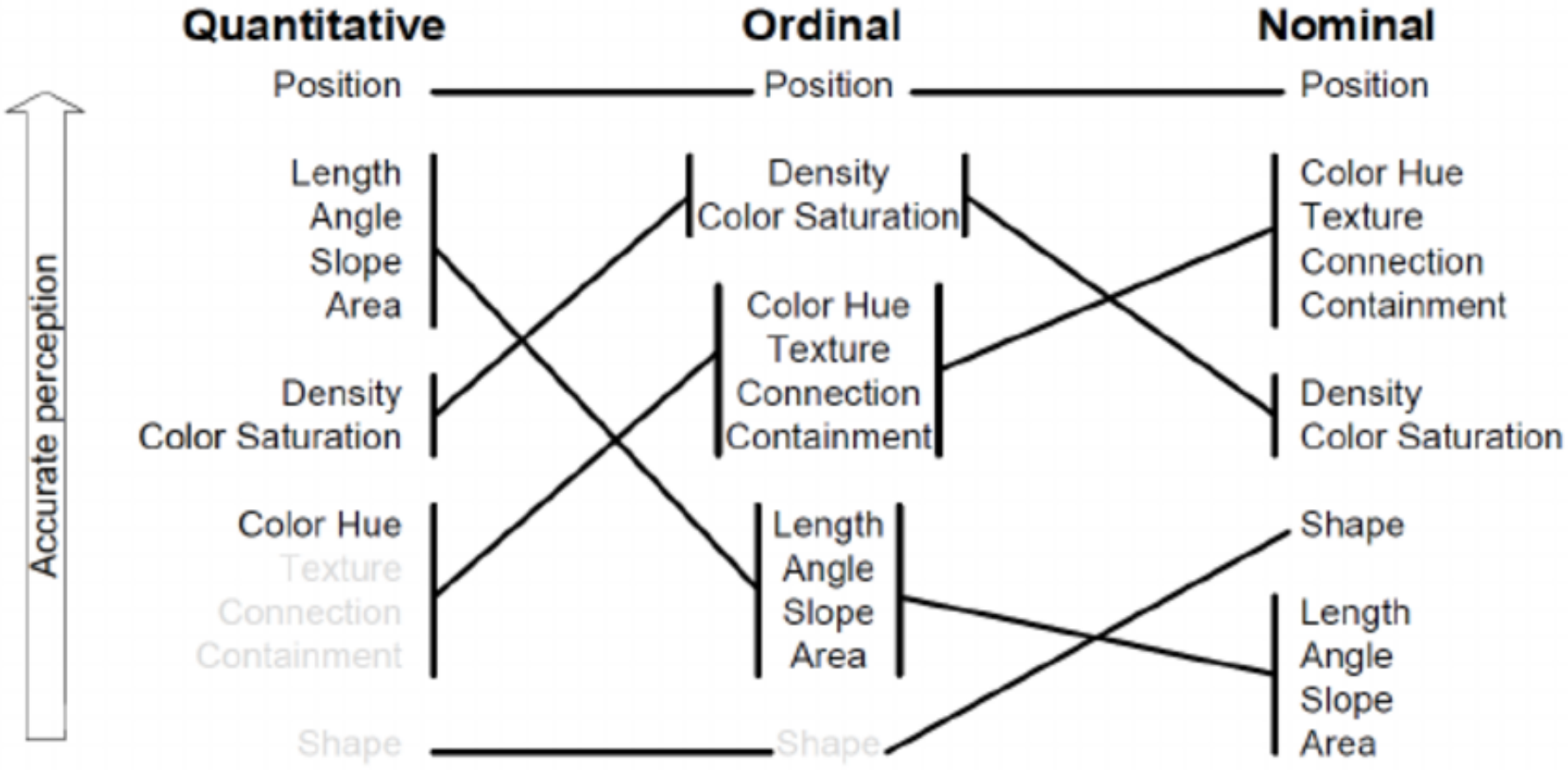
Level of Measurement

RECALL

A way of classifying the nature of data values. Applies to all data analysis, distinct from the R "data type".

Level	Example	Operations
★ **Nominal** unordered used for classification	*Fruits:* apples, bananas, oranges, etc.	**== !=** "same or different"
★ **Ordinal** ordered can comparison	*Grade of meat*: Grade A, Grade AA, Grade AAA, etc.	**== != < >** "bigger or smaller"
★ **Interval (Quantitative)** ordered, no set "zero" can find difference	*Dates:* 05/15/2012, 04/17/2015, etc.	**== != < >** **+ –** "3 units bigger"
Ratio (Quantitative) ordered, fixed "zero" can find magnitude	*Lengths:* 1 inch, 1.5 inches, 2 inches, etc.	**== != < >** **+ – * /** "twice as big"

Visual Channel Effectiveness



(Mackinlay, 1986)

Position

The most accurate visual channel for all data types.

Resemblance (nominal)

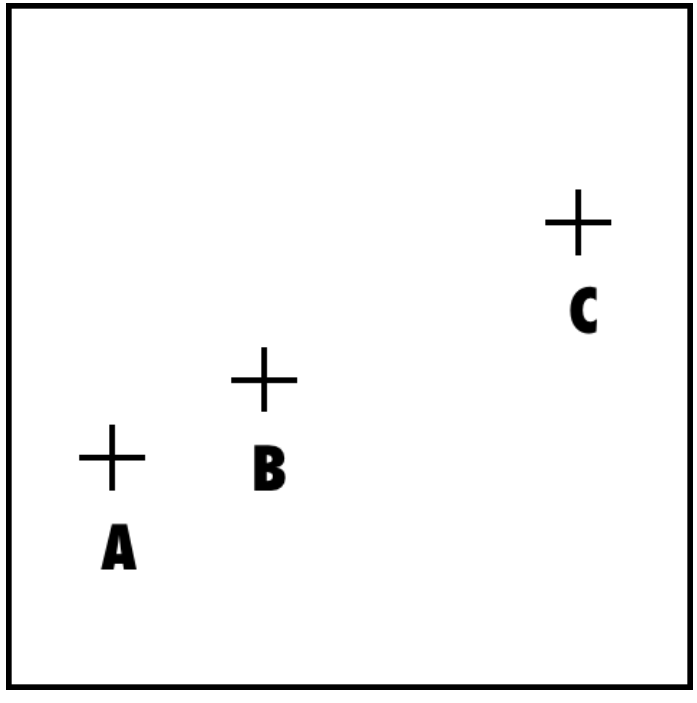
(A != B != C)

Order (ordinal)

(B is between A and C)

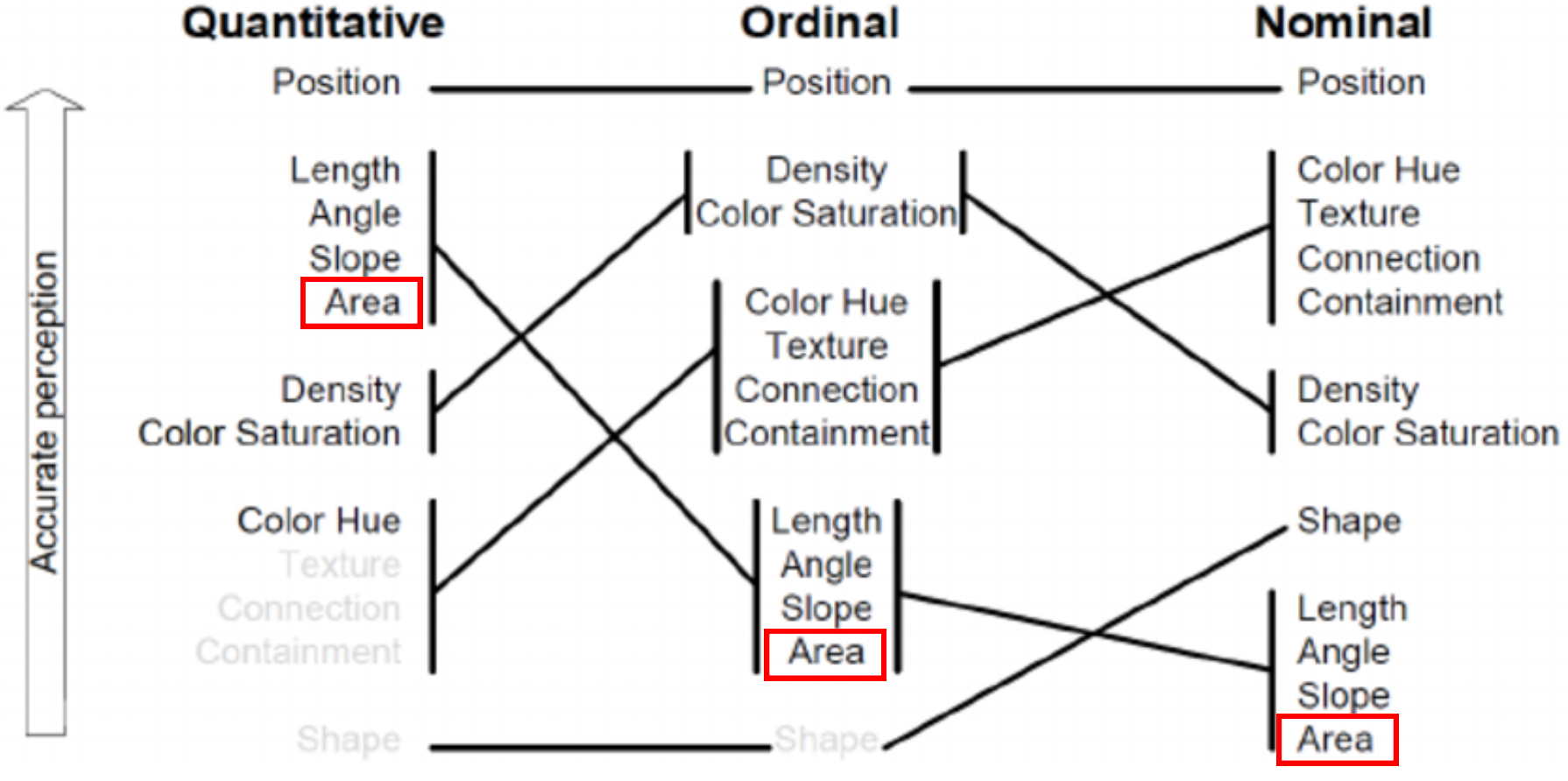
Proportion (quantitative)

(BC is 2x long as AB)



Size

Less accurate as data is less quantitative



Color: Hue, Brightness, Saturation

Hue

(red, yellow, green, etc)

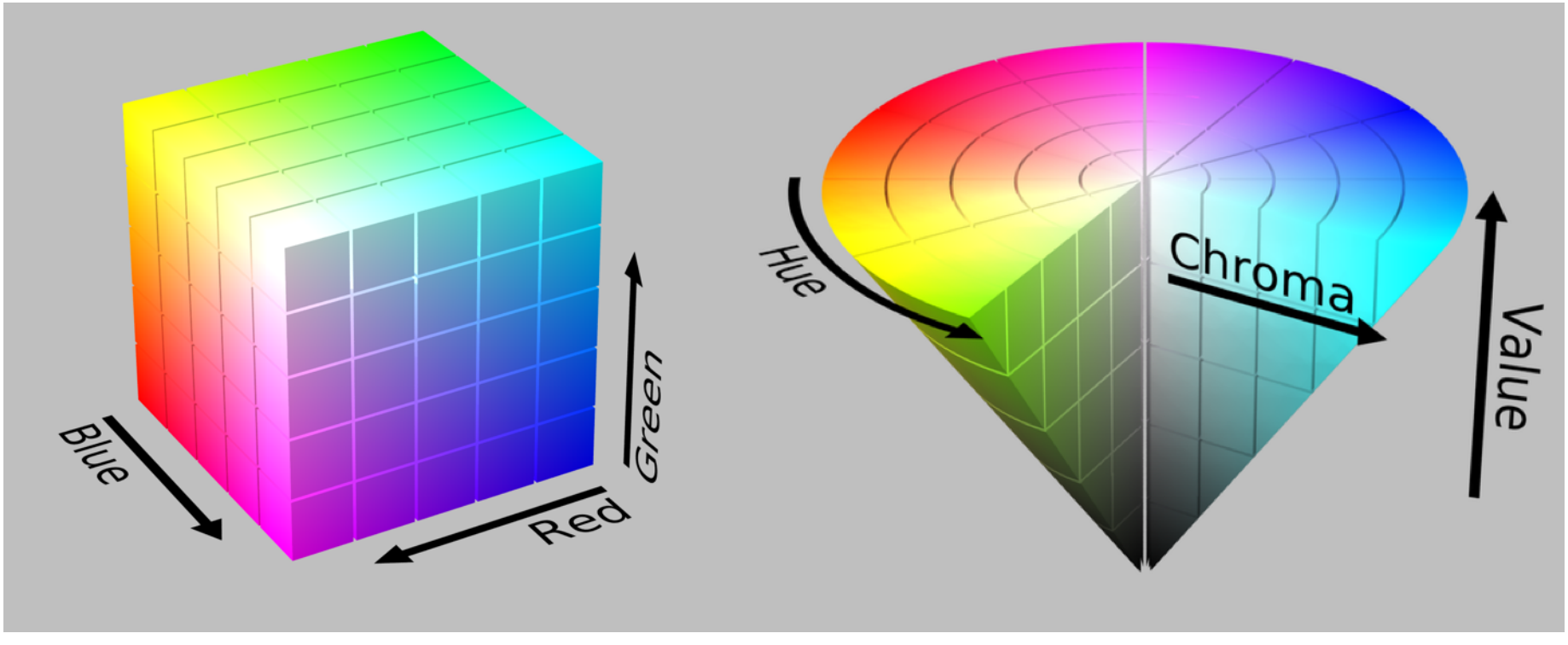
Saturation

(vivid vs. washed out)

Brightness or **Value**

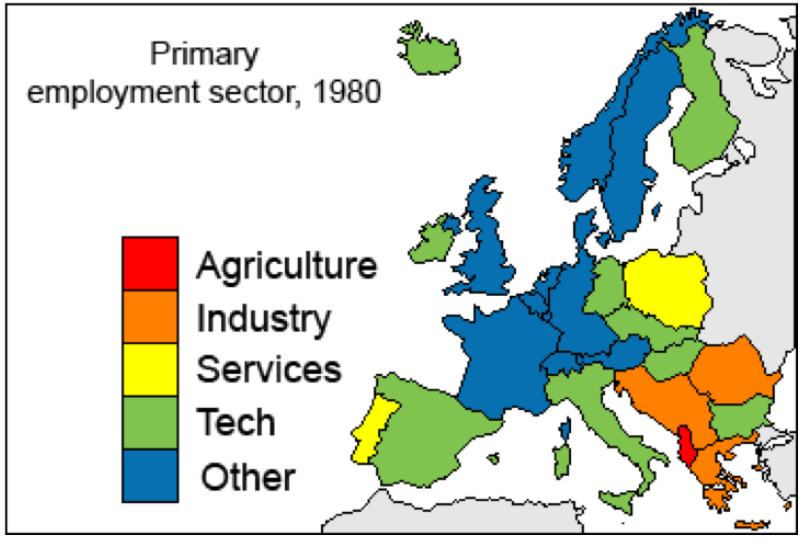
(luminance)

Color: RGB vs. HSB

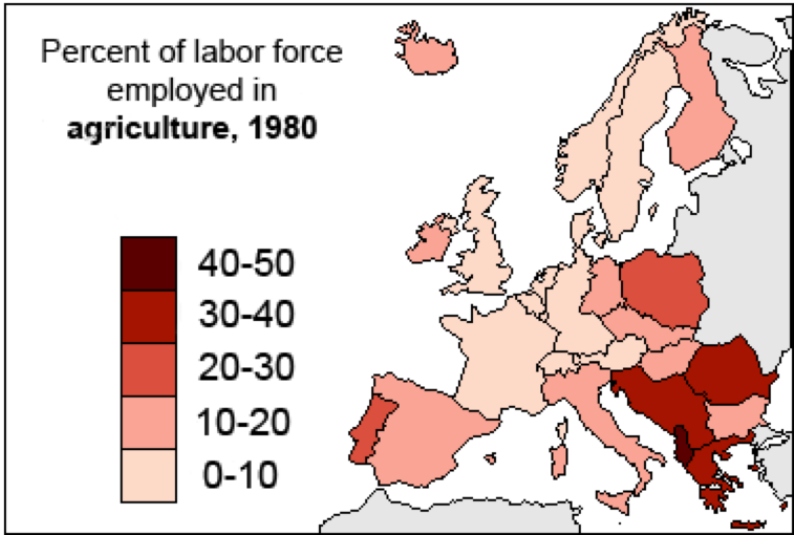


Color

Hue is good for *nominal* variables,
Saturation is good for *ordinal* variables

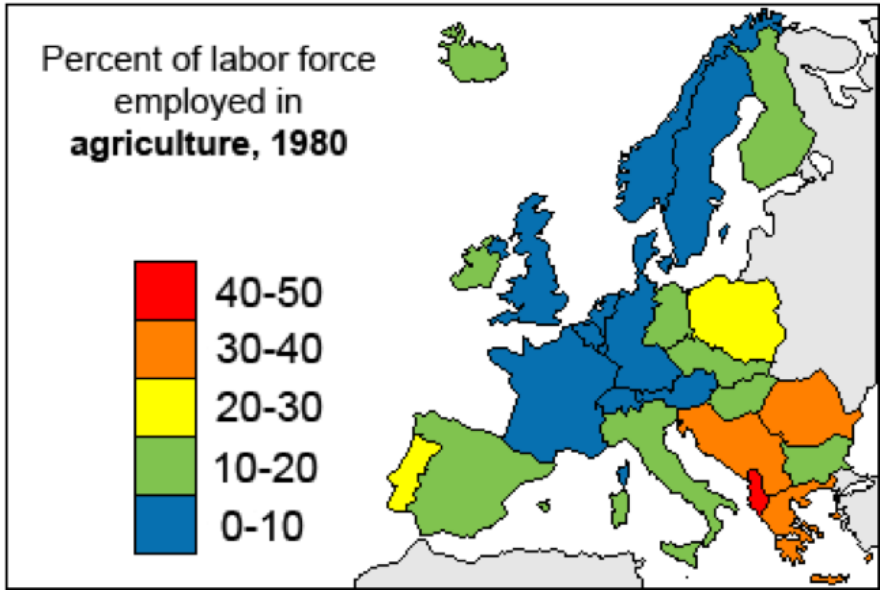


Hue (nominal)

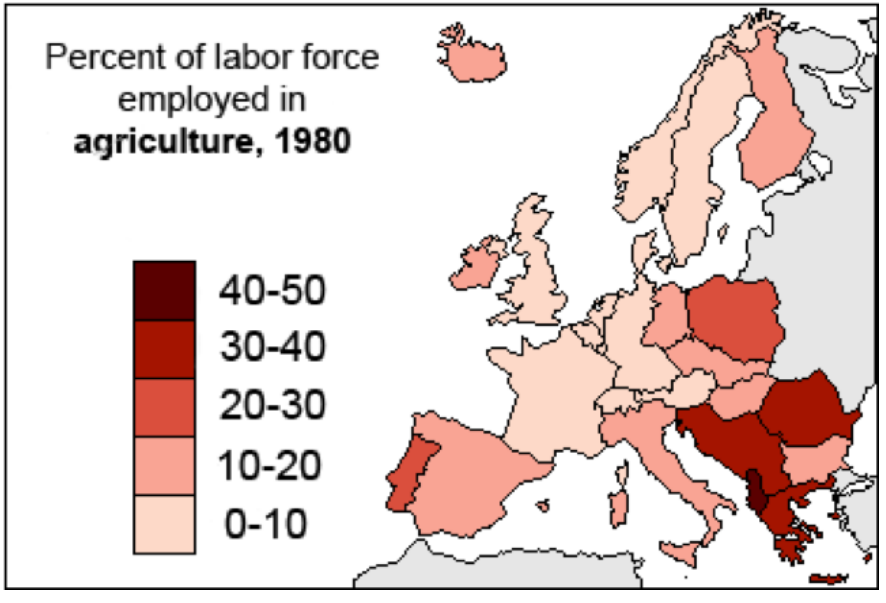


Saturation (ordinal)

Which is more effective?

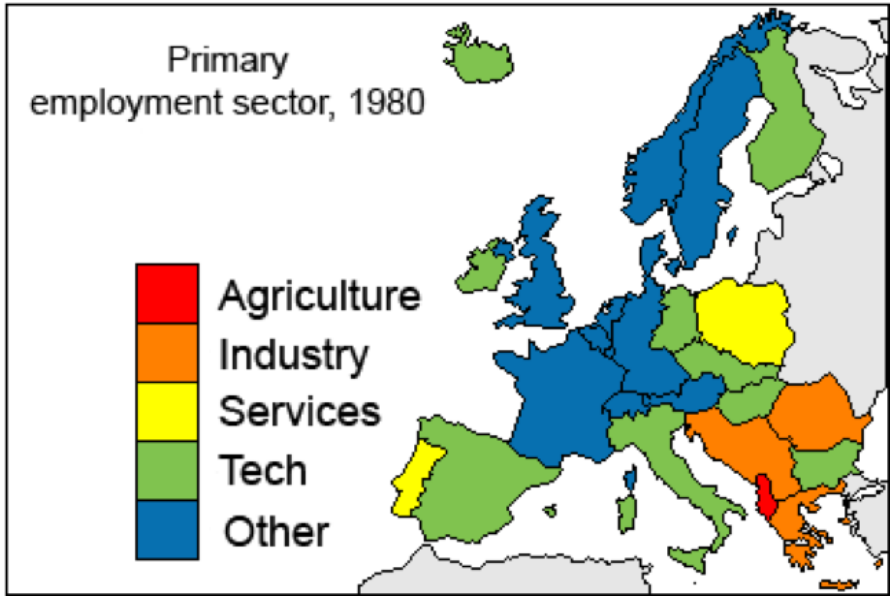


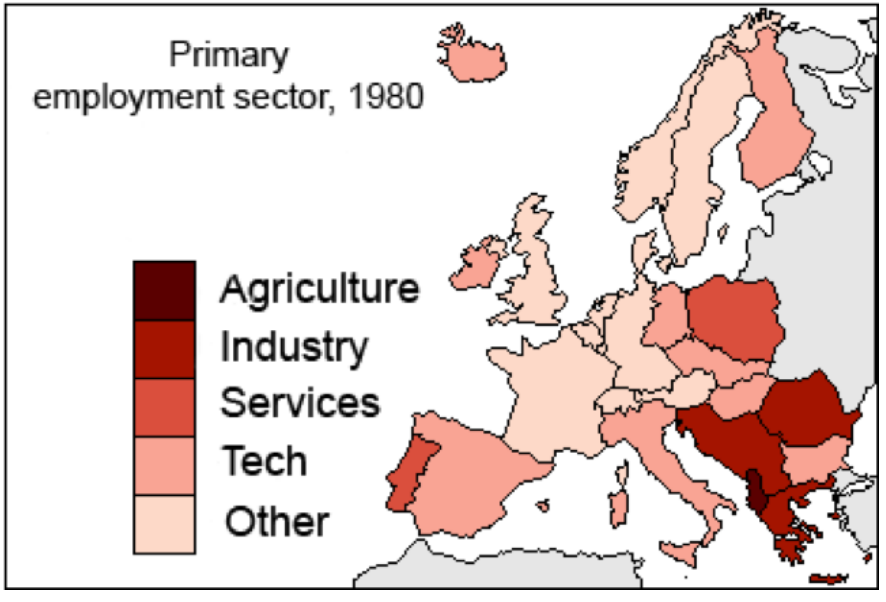
A



B

Which is more effective?





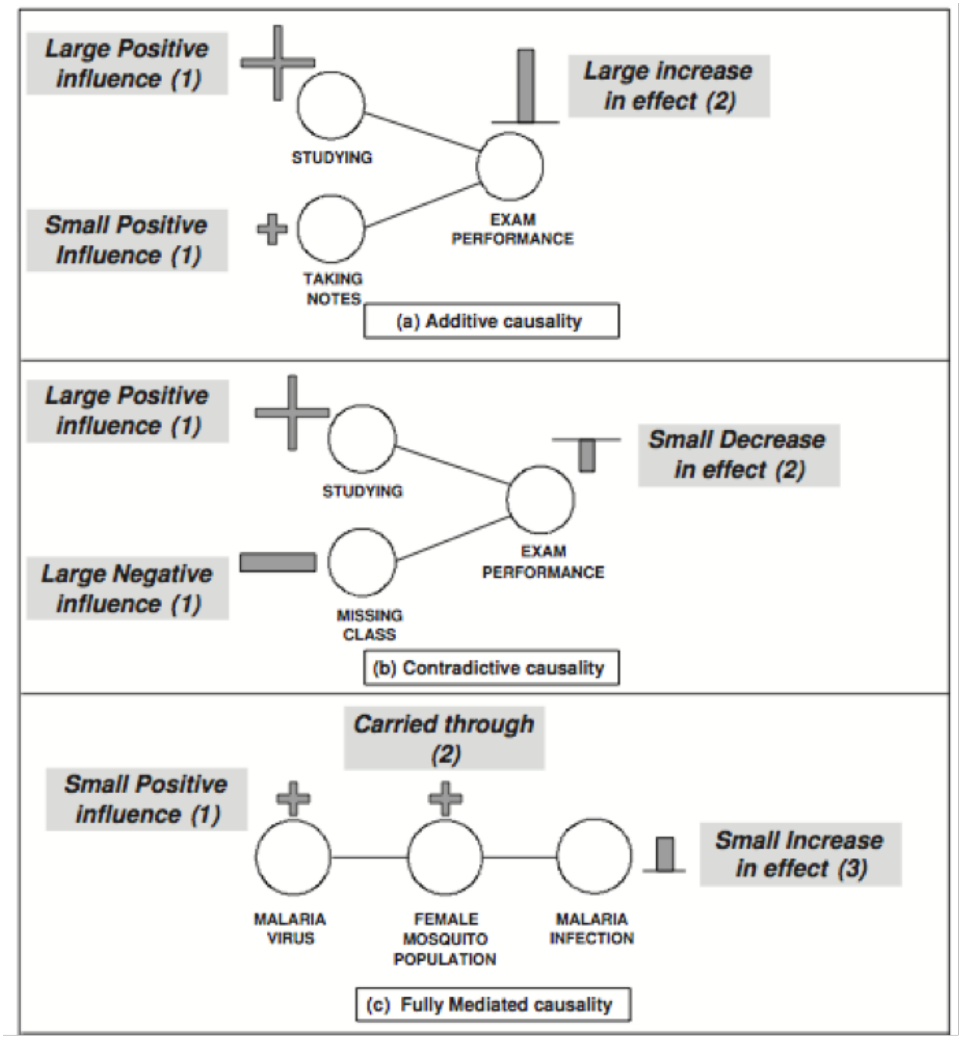
A

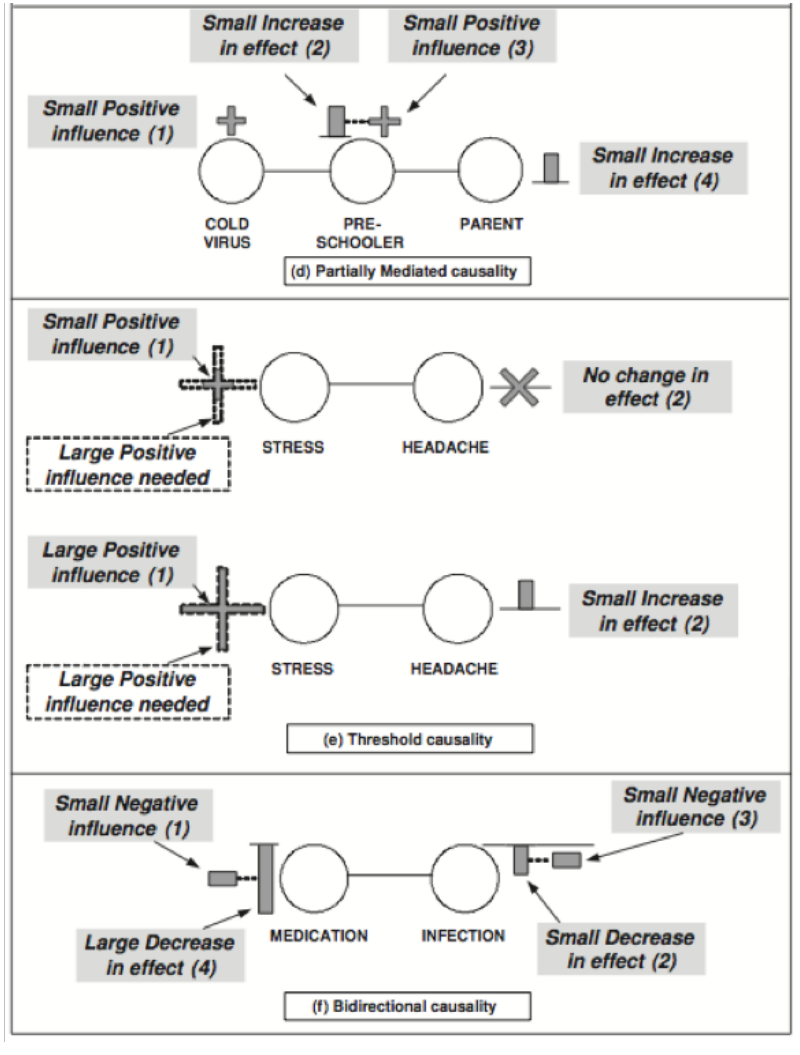
B

Other Channels: Orientation

http://hint.fm/wind/

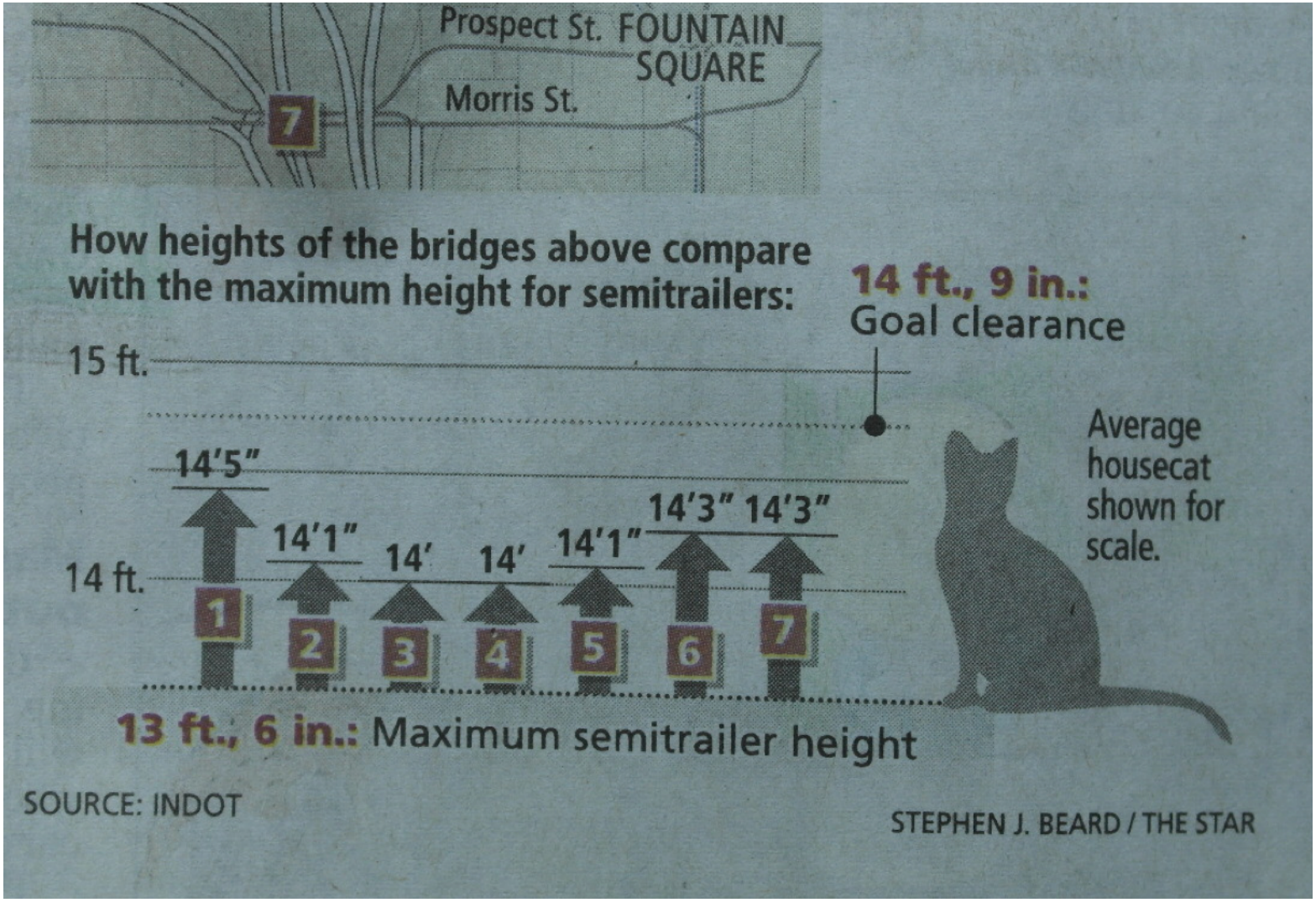
Labels & Glyphs



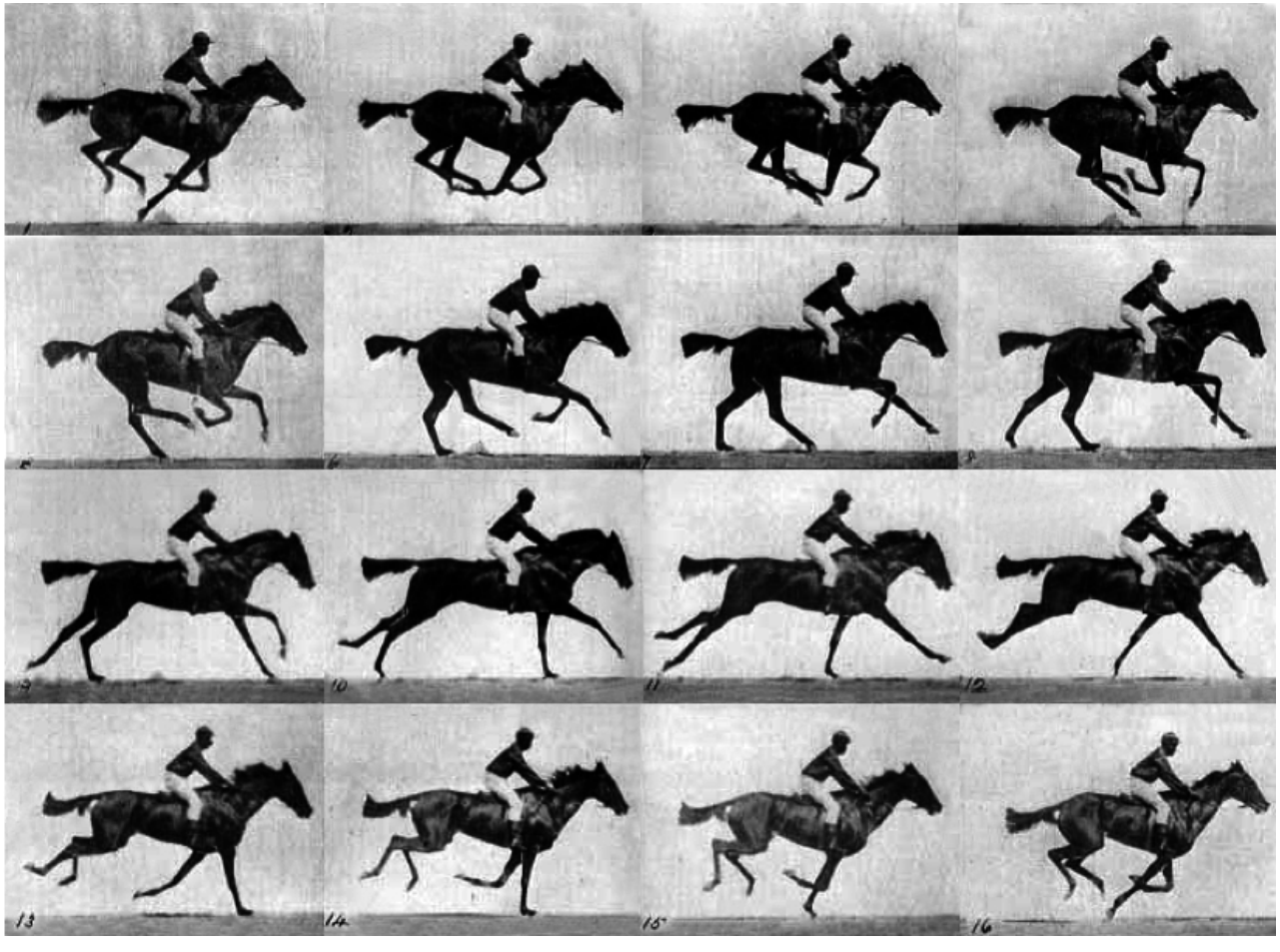


(Kabada et al. 2007)

Poor Labeling



Animation



E.g.: Michotte's Launching Effect (visualizing causality)
http://cogweb.ucla.edu/Discourse/Narrative/michotte-demo.swf

Utilize the channels that are most effective for the data that is most important

But you need to know what data is important to your audience!

Back to programming....

Grammar of Graphics

Words used to describe the visual components and aspects of a graphic.

- **Data** shown in the plot
- **Geometric objects (geoms)** that appear on the plot
- **Aesthetic mappings** from the data to the **geoms**
- **Statistical transformation** used to calculate the data
- **Scales** (range of values) for each **aesthetic**
- **Coordinate system** to organize the **geoms**
- **Facets** or groups of data shown in different plots

ggplot2

Use the `ggplot()` function to draw a plot, specifying plot elements via the grammar.

```
# plot the `mpg` data set, with highway milage
# on the x axis and engine displacement (power)
# on the y axis:

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

data to plot

add geometry

geometric objects (points)

aesthetic mappings

Facets

Break a plot into parts with **facets** (similar to `group_by()` in `dplyr`). Each facet acts like a "level" in a factor, with a plot for each level.

```
# a plot with facets based on vehicle type.
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~class)
```

A **formula** , read as

"as a function of"

Consider a data set...

	name	section	math_exam1	math_exam2	spanish_exam1	spanish_exam2
1	Mason	a	91	88	79	99
2	Tabi	a	82	79	88	92
3	Bryce	a	93	77	92	92
4	Ada	b	100	99	83	82
5	Bob	b	78	88	87	85
6	Filipe	b	91	93	77	95

What if we want to facet by exam?

Data Shape

Wide Data

	name	section	math_exam1	math_exam2	spanish_exam1	spanish_exam2
1	Mason	a	91	88	79	99
2	Tabi	a	82	79	88	92
3	Bryce	a	93	77	92	92
4	Ada	b	100	99	83	82
5	Bob	b	78	88	87	85
6	Filipe	b	91	93	77	95

6 rows x 4 cols
= 24 scores

Long Data

	name	section	exam	score
1	Mason	a	math_exam1	91
2	Tabi	a	math_exam1	82
3	Bryce	a	math_exam1	93
4	Ada	b	math_exam1	100
5	Bob	b	math_exam1	78
6	Filipe	b	math_exam1	91
7	Mason	a	math_exam2	88
8	Tabi	a	math_exam2	79
9	Bryce	a	math_exam2	77
10	Ada	b	math_exam2	99
11	Bob	b	math_exam2	88
12	Filipe	b	math_exam2	93
13	Mason	a	spanish_exam1	79
14	Tabi	a	spanish_exam1	88

24 rows x 1 col
= 24 scores

Data Shape

We can convert between **wide** and **long** data (and vice versa) using the `tidyr` package.

```
# Alternatively, install "tidyverse"
install.packages("tidyr")  # once per machine
library("tidyr")



# Make a data.frame (example)
students <- data.frame(
  name = c('Mason', 'Tabi', 'Bryce', 'Ada', 'Bob','Filipe'),
  section = c('a','a','a','b','b','b'),
  math_exam1 = c(91, 82, 93, 100, 78, 91),
  math_exam2 = c(88, 79, 77, 99, 88, 93),
  spanish_exam1 = c(79, 88, 92, 83, 87, 77),
  spanish_exam2 = c(99, 92, 92, 82, 85, 95)
)
```

http://tidyr.tidyverse.org/

http://tidyverse.org/

Data Shape

Convert from **wide** to **long** using `gather()`. The **key** is a new column containing *gathered colnames*, and **value** is a new column with their values.

```
students.long <- gather(students.wide,
                        key = exam,
                        value = score,
                        math_exam1, math_exam2,
                        spanish_exam1, spanish_exam2
                       )
```

names for new columns

col data to populate with

Convert from **long** to **wide** using `spread()`. The **key** is where to get the *new colnames*, and **value** is where to get the values

```
# spread by column "exam"
stu.wide <- spread(students.long, key = exam, value = score)

# spread by column "name"
stu.wide.name <-
         spread(students.long, key = name, value = score)
```

Map Visualizations

Polygons

Use **geom_polygon** to draw shapes.

```
rect <- data.frame(x.coords = c(3, 5, 5, 3),
                   y.coords = c(4, 4, 2, 2))

ggplot(data = rect) +
  geom_polygon(aes(x = x.coords, y = y.coords))
```

each row is a corner point

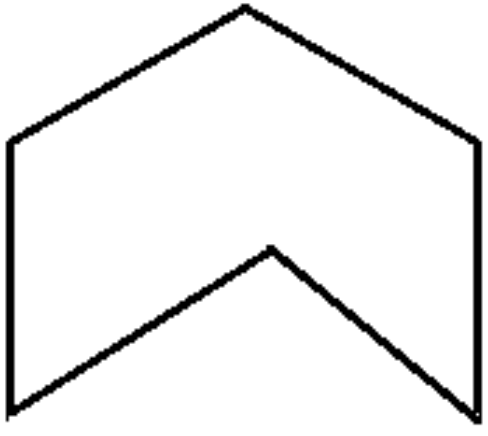
Draw multiple polygons by **grouping** points.

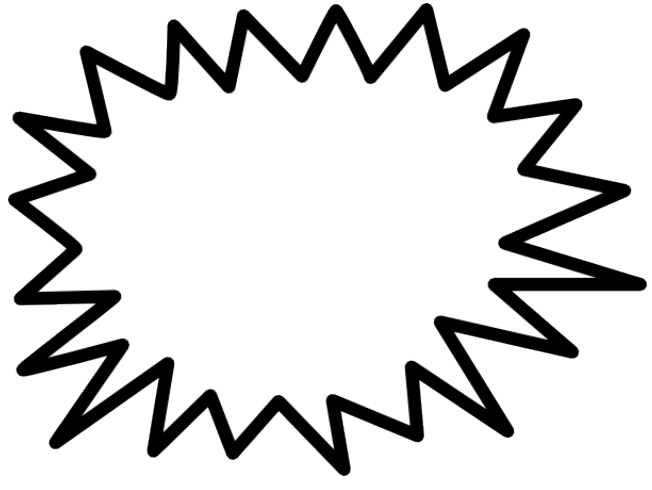
```
double.rect <- data.frame(x.coords = c(1,2,2,1, 3,4,4,3),
                          y.coords = c(2,2,1,1, 2,2,1,1),
                          rect.num = c(1,1,1,1, 2,2,2,2))

ggplot(data = double.rect) +
  geom_polygon(aes(x=x.coords, y=y.coords, group = rect.num))
```

which rect the point goes with

Polygons





Map Data

ggplot2 provides a set of data frames (from the `maps` library) which include polygon definitions for different geographic maps.

Access these data frames with the `map_data()` function.

```
# load the data
usa.states <- map_data("state")

# plot the polygons
ggplot(data = usa.states) +
  geom_polygon(aes(x = long, y = lat, group = group)) +
  coord_quickmap()  # map coordinate system!
```

Module 13 exercise-3

Clone a fresh copy, or pull changes!

```
# In Terminal (in your repo directory)

# Add a new "bookmark" to remote called "upstream" to the repo
git remote add upstream https://github.com/info201-w17/module13-ggplot2

# download changes from the upstream repo
git pull upstream master
```

Action Items!

- Be comfortable with **module 13**
- Assignment 5 due ***Thursday before class***
  - (Assignment 6 online soon)

Thursday: What makes a good visualization?
Also maps.