

COMP-5360 / MATH-4100

Introduction to Data Science

Visualization

Alexander Lex
alex@sci.utah.edu



visualization

pictures

The purpose of computing is insight, not numbers.

- Richard Wesley Hamming

- Card, Mackinlay, Shneiderman

Banana

M. acuminata

Date

P. dactylifera

Cress

Arabidopsis thaliana

Rice

Oryza sativa

Sorghum

Sorghum bicolor

Brome

Brachypodium distachyon

Visualization Definition

**Visualization is the process that transforms
(abstract) data into
interactive graphical representations for the purpose of
exploration, confirmation, or presentation.**

Good Data Visualization

- ... makes data accessible
- ... combines strengths of humans and computers
- ... enables insight
- ... communicates

Why Visualize?

To inform humans: Communication

How is ahead in the election polls?

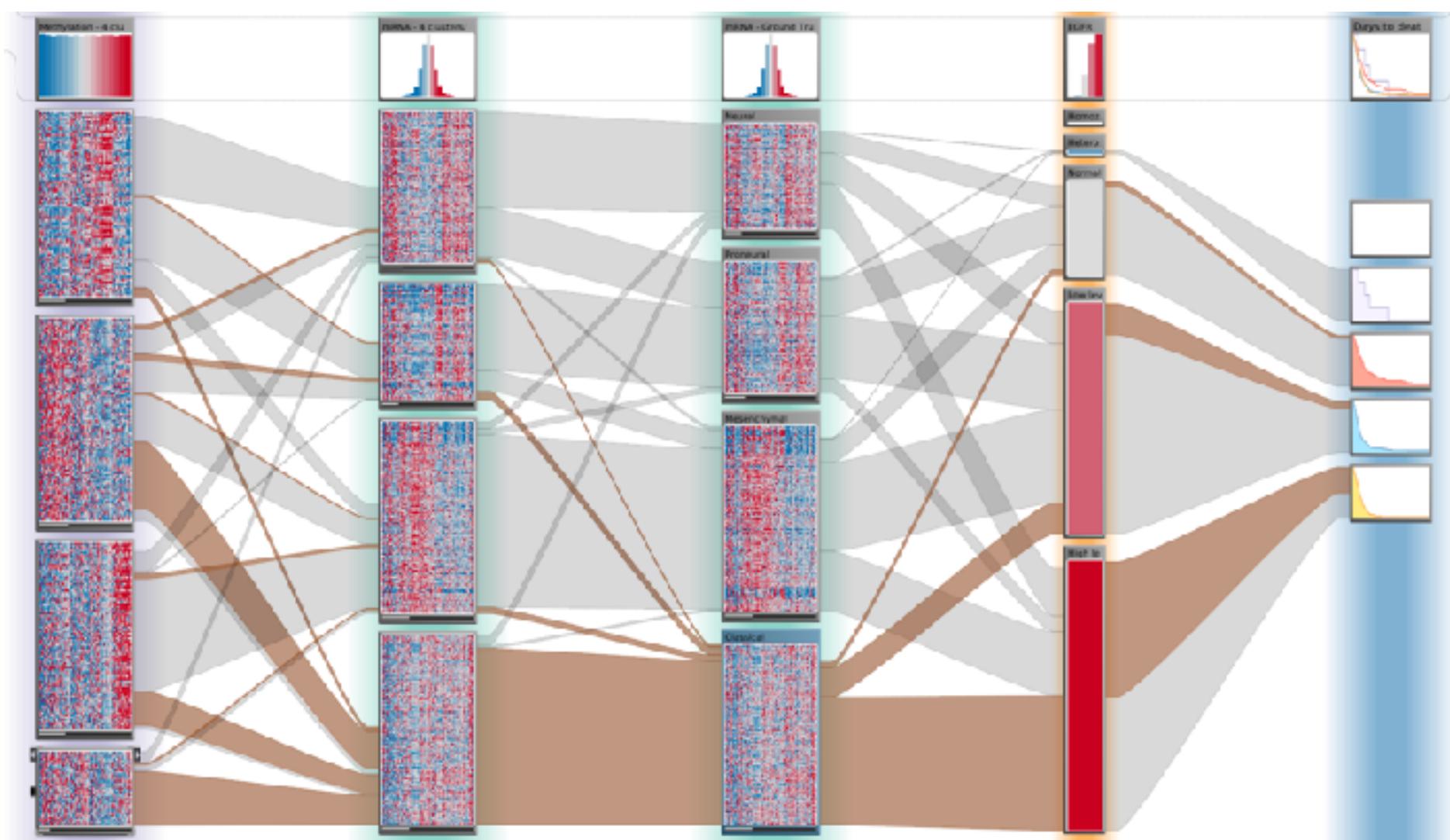
When questions are not well
defined: Exploration

What is the structure of a terrorist network?

Which drug can help patient X?

Purpose of Visualization

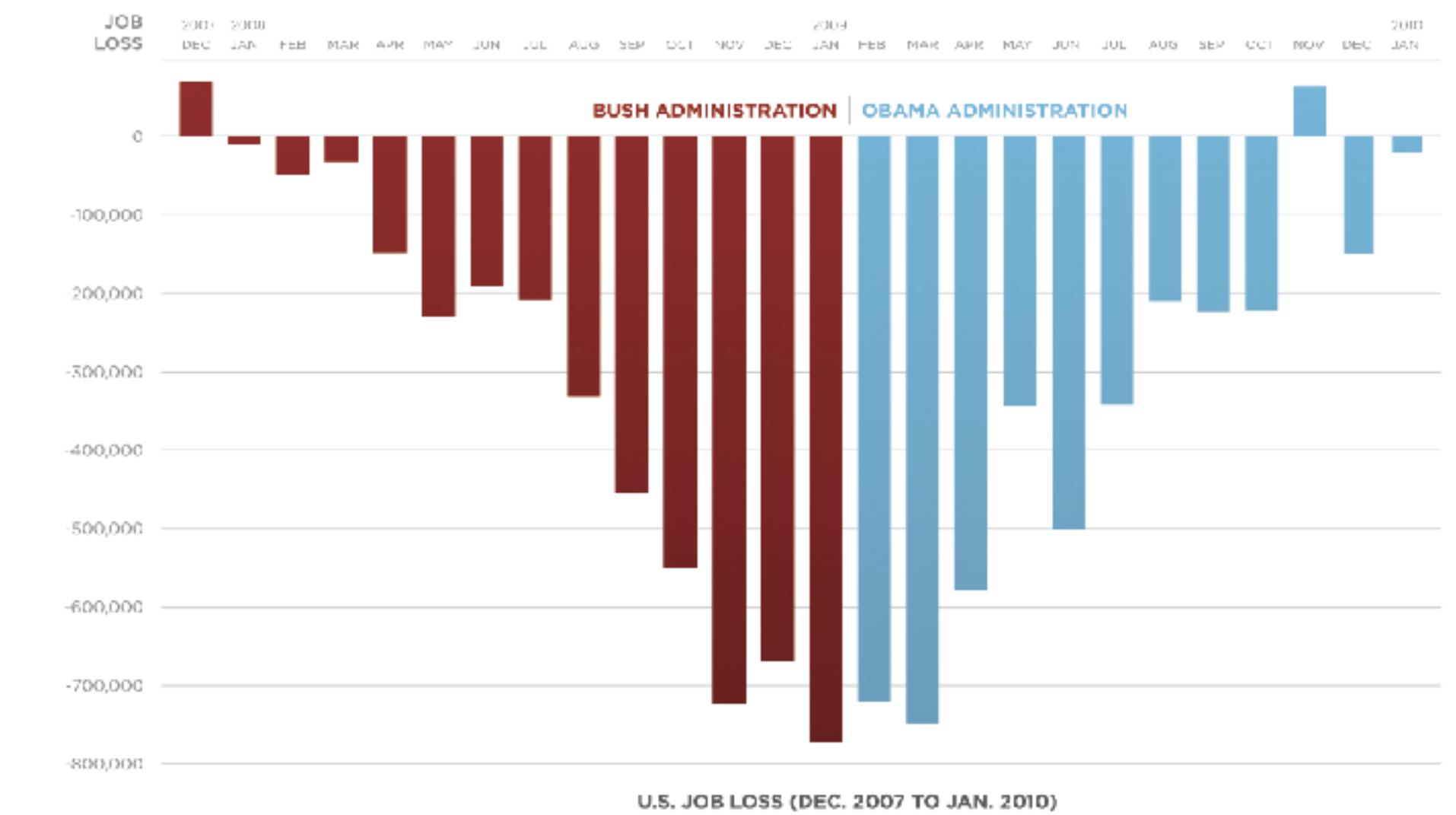
[Obama Administration]



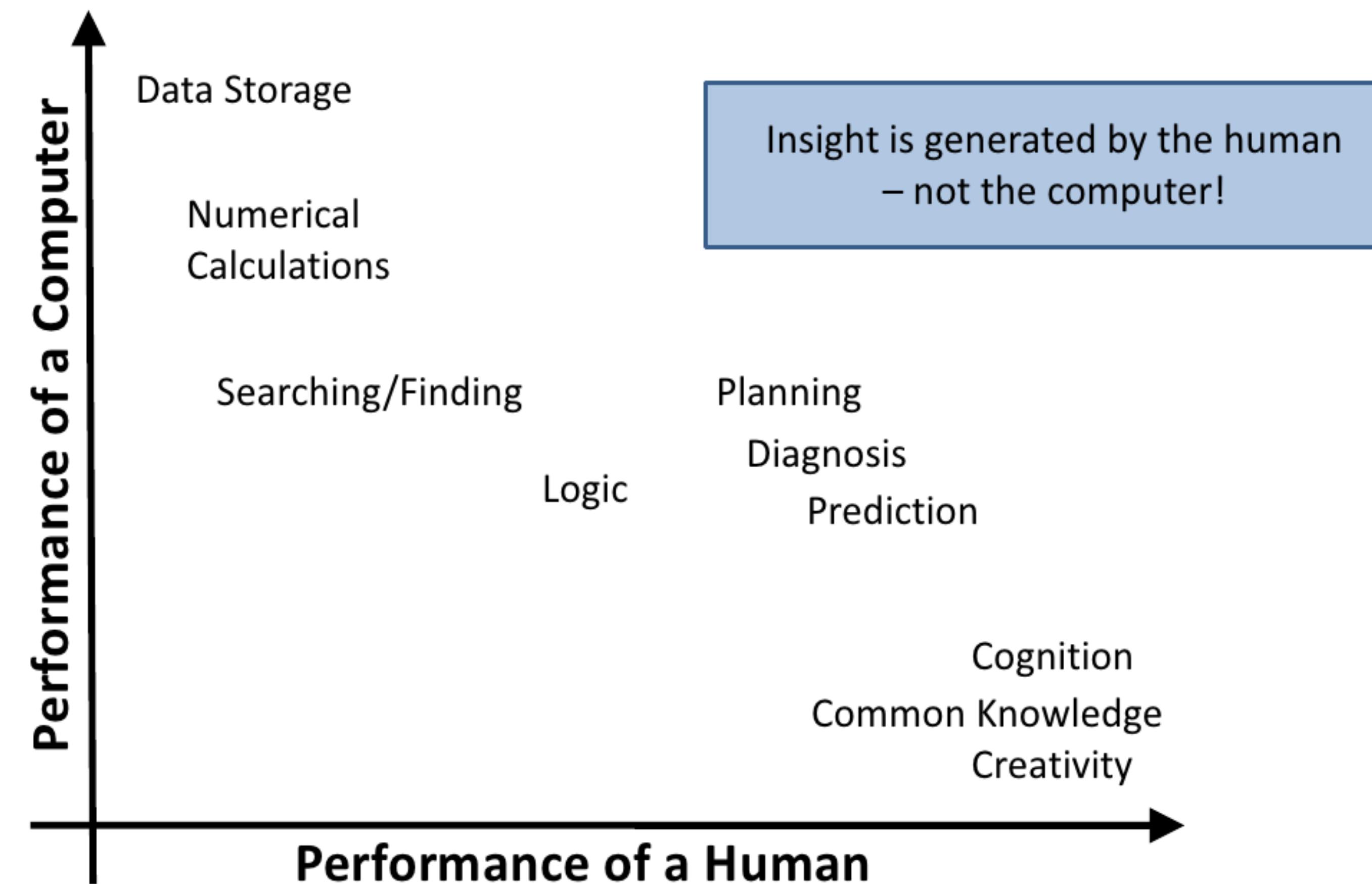
Open Exploration

Confirmation

Communication



The Ability Matrix



Why not just use Statistics?

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.1	8	7.9
5	5.5						6.8

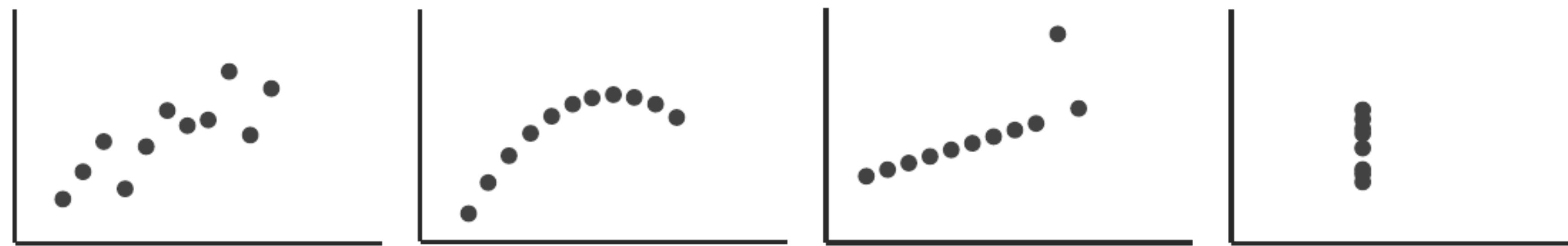
Mean x: 9 y: 7.50

Variance x: 11 y: 4.122

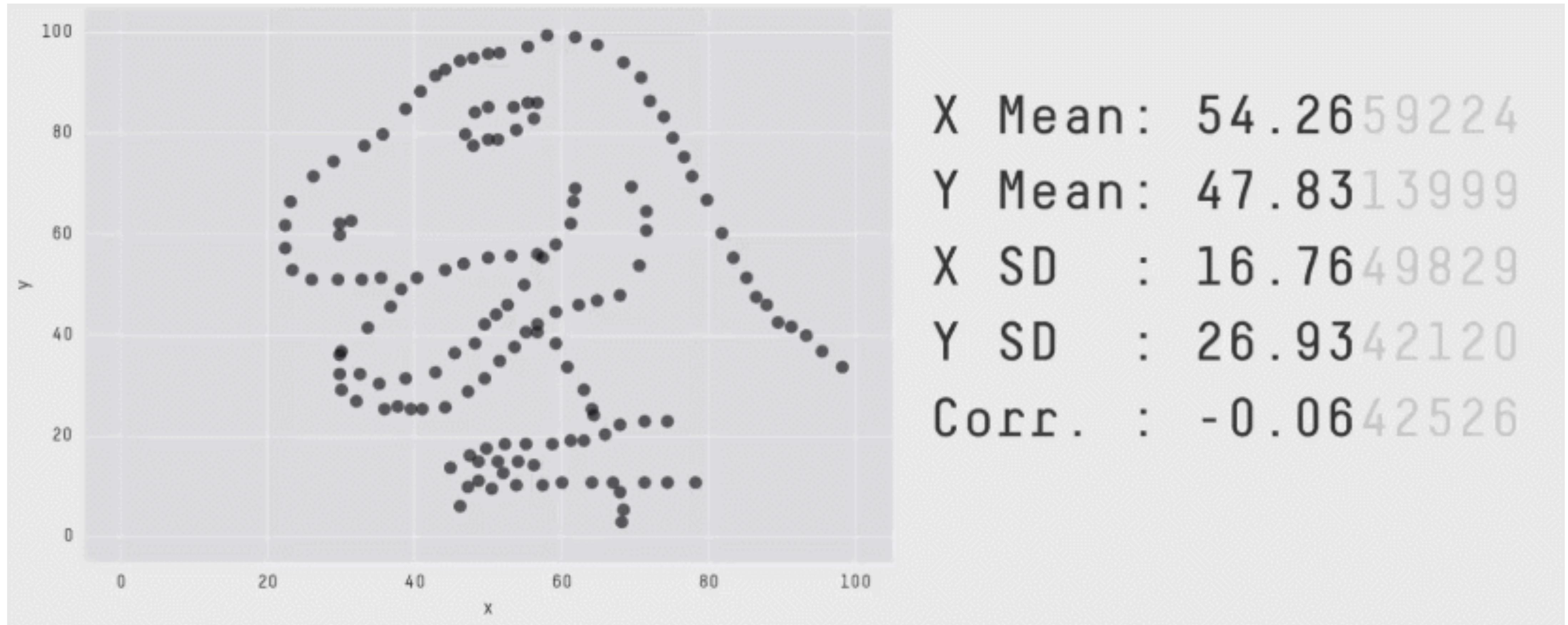
Correlation x - y: 0.816

Linear regression: $y = 3.00 + 0.500x$

Anscombe's Quartett



Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x - y: 0.816
Linear regression: $y = 3.00 + 0.500x$



Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, CHI 2017, Justin Matejka, George Fitzmaurice

Visualization =

Human Data Interaction

Visualization in the Data Science Process

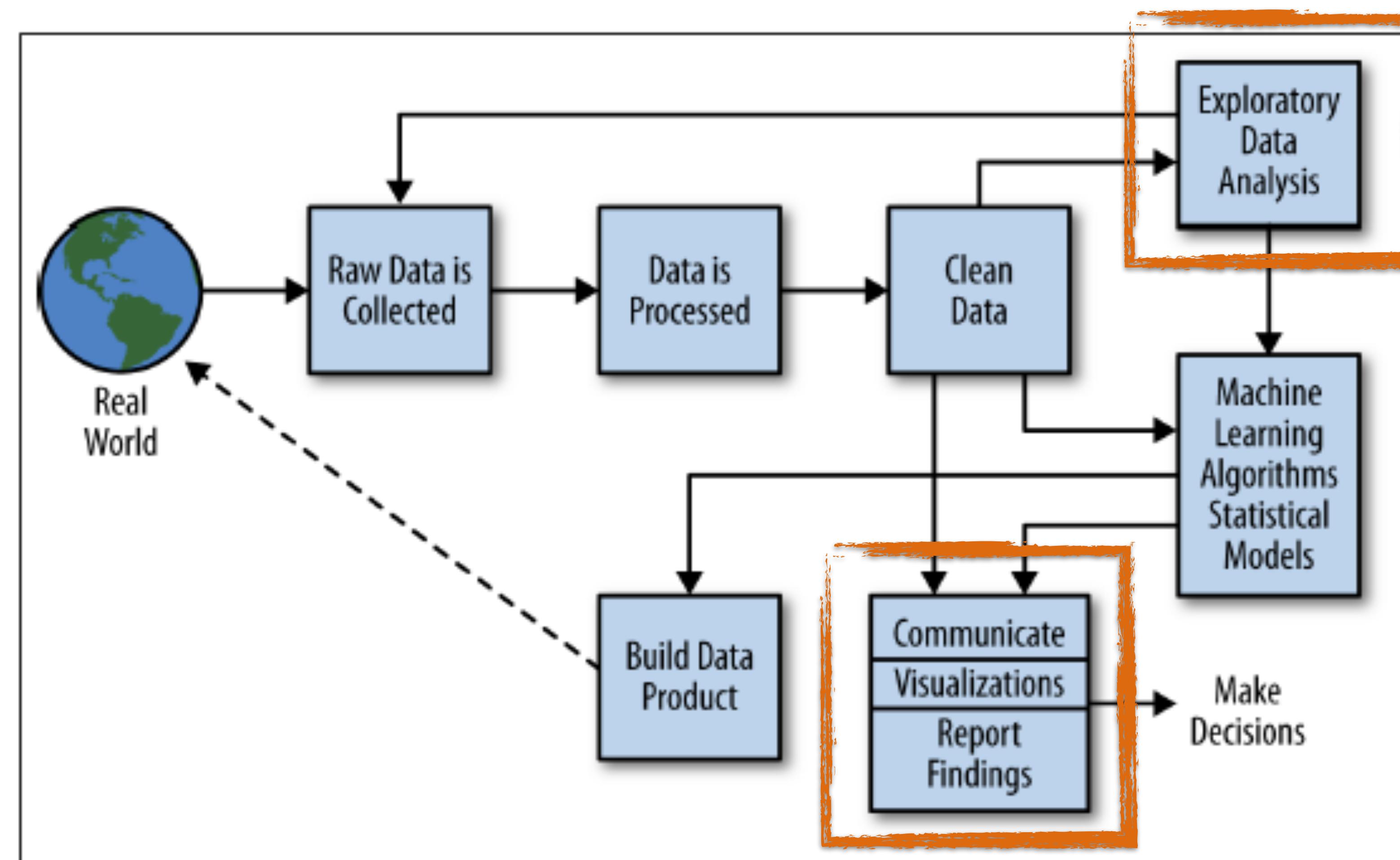


Figure 2-2. The data science process

Why Humans?

Leveraging human capabilities

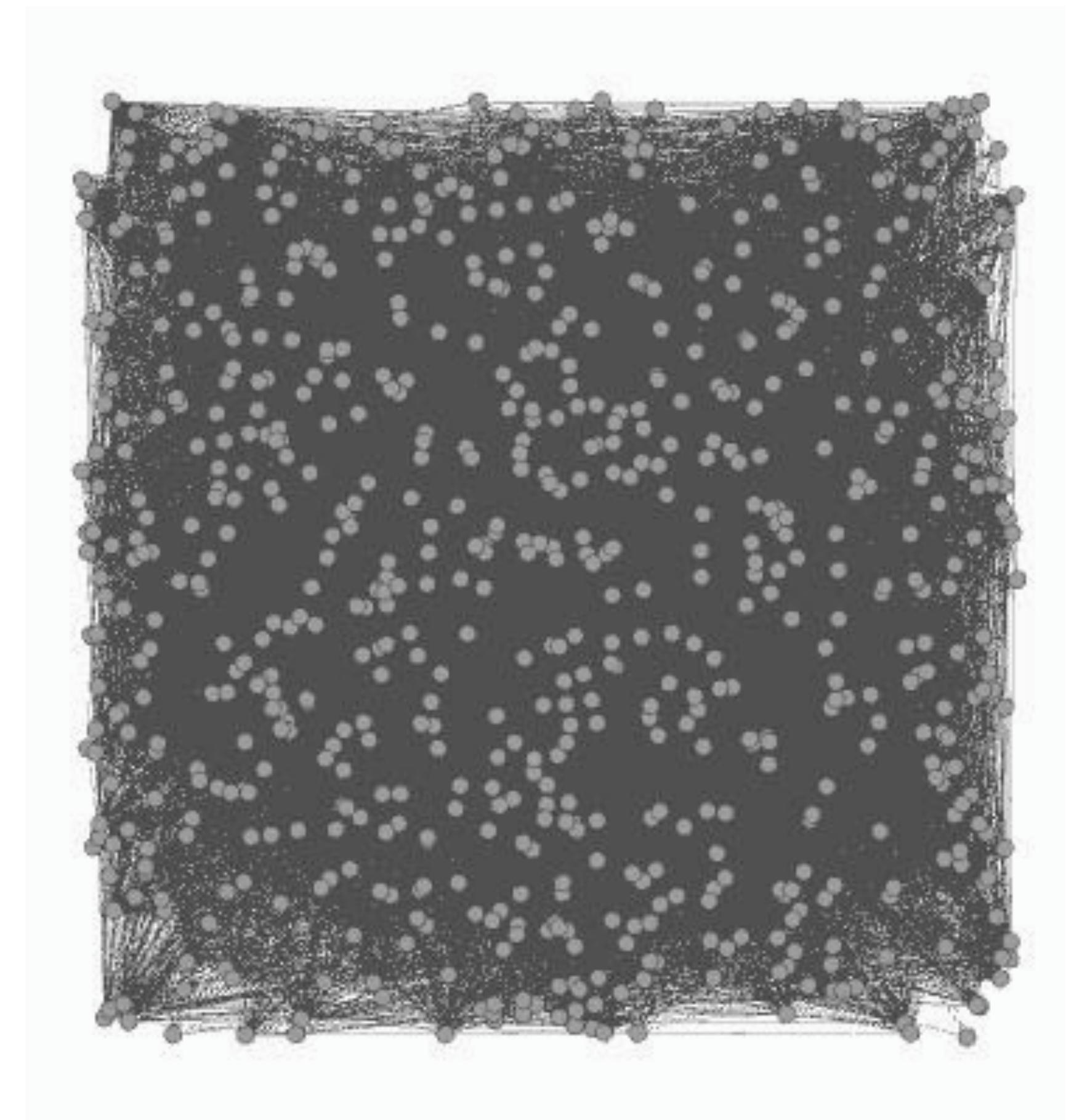
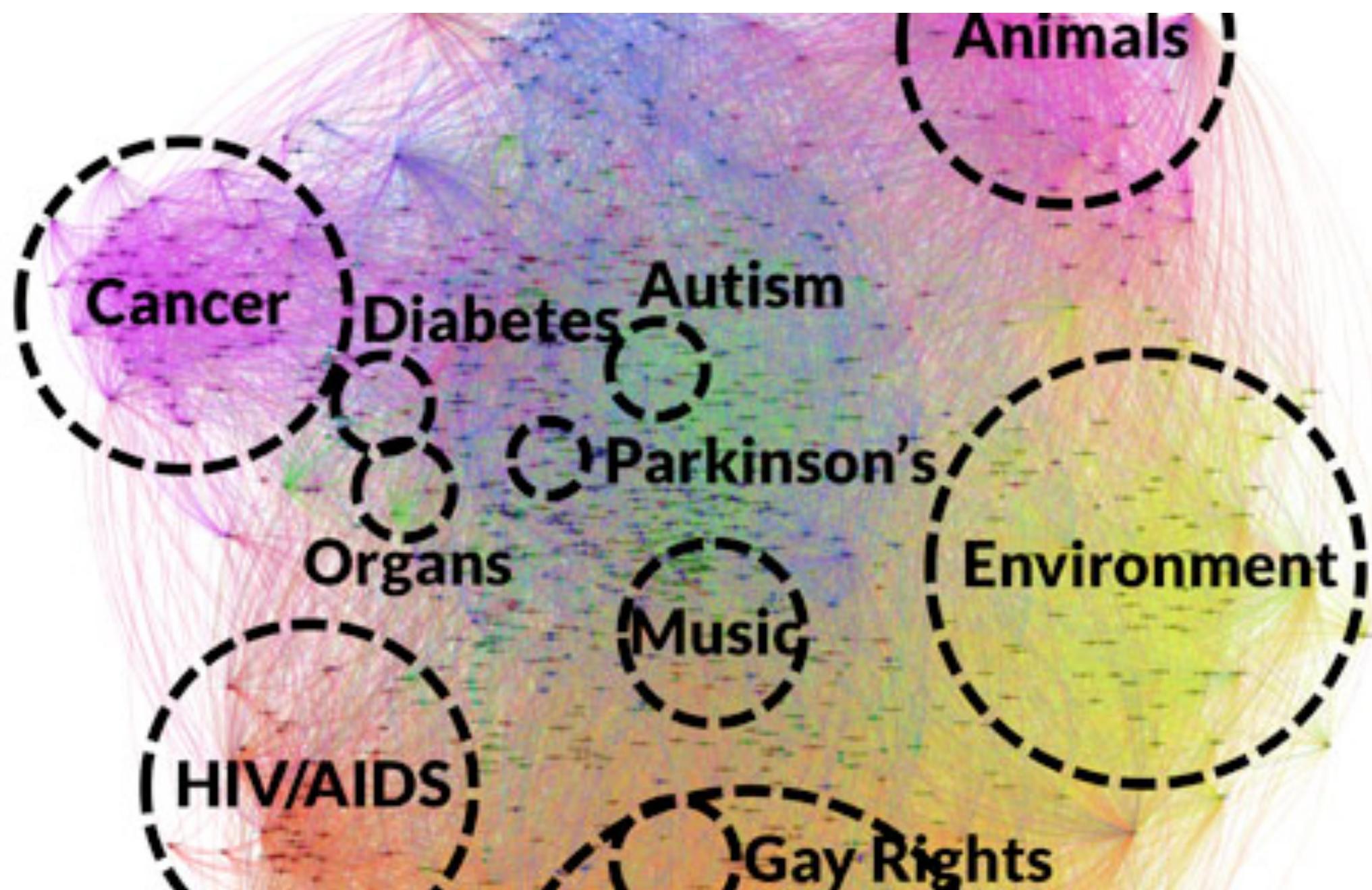
Pattern Discovery: clusters, outliers, trends

Contextual Knowledge: expectations for dataset, explanations for patterns

Action: humans learn and take action

But: we also have to **design for Humans and their limitations**

Not everything that can be drawn can be read!

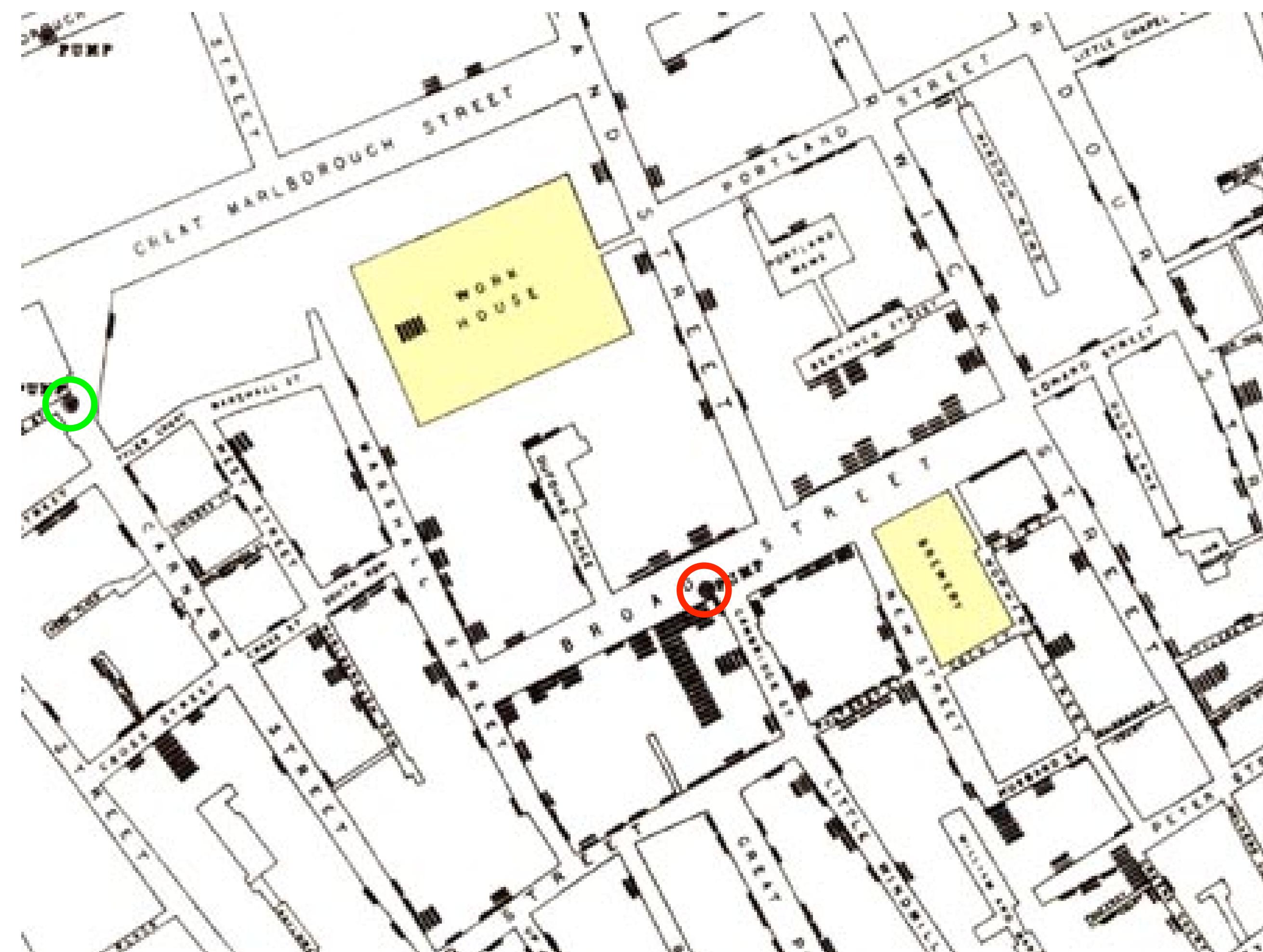


Limits of Cognition



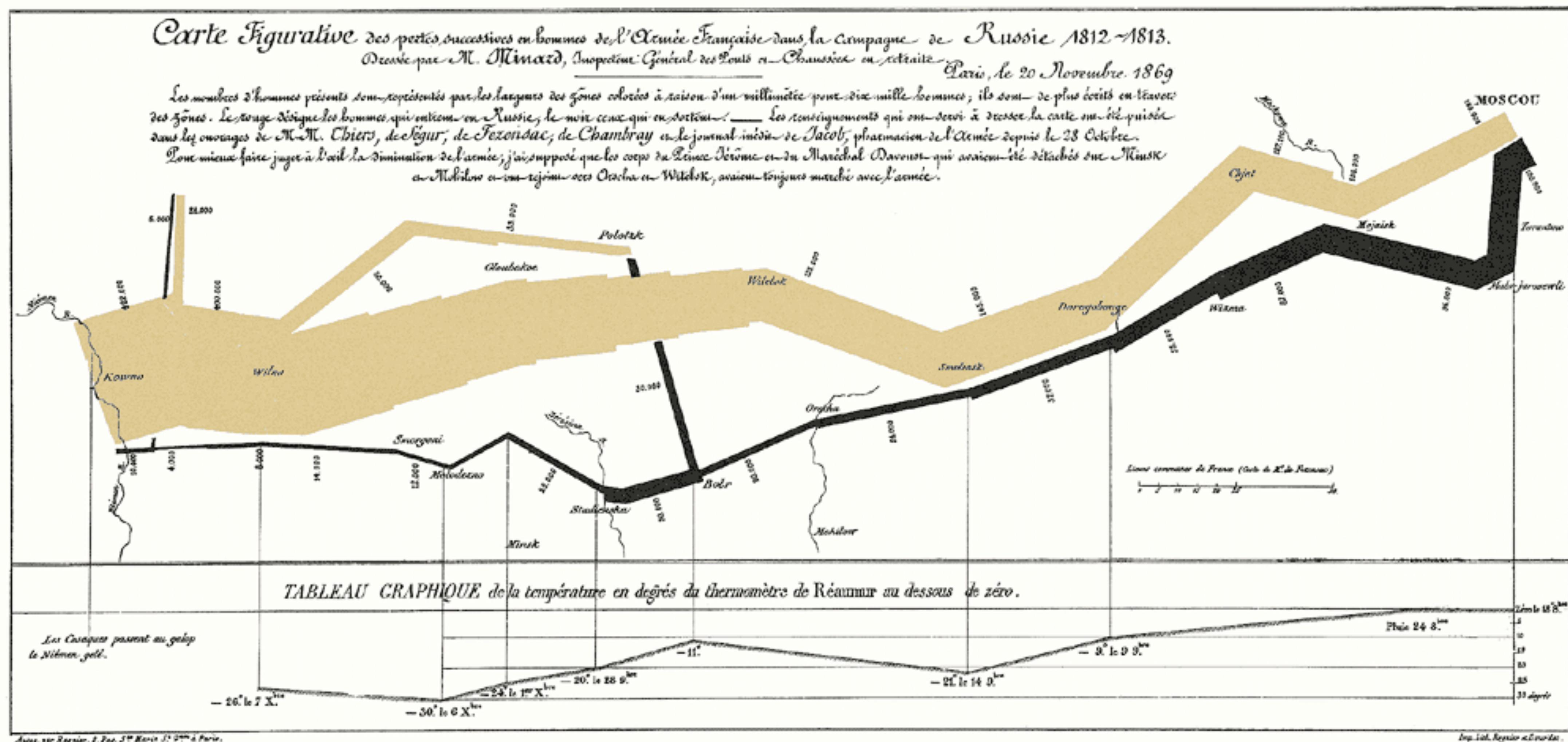
Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

History: John Snow



John Snow, 1854

Minard: Napoleon's Russian Campaign



Communicate



Hans Rosling, TED 2006

Marks & Channels

How can I visually represent two numbers, e.g.,
4 and 8

Marks & Channels

Marks: represent **items** or **links**

Channels: change **appearance** based on **attribute**

Channel = Visual Variable

Marks for Items

Basic geometric elements

→ Points



0D

→ Lines



1D

→ Areas

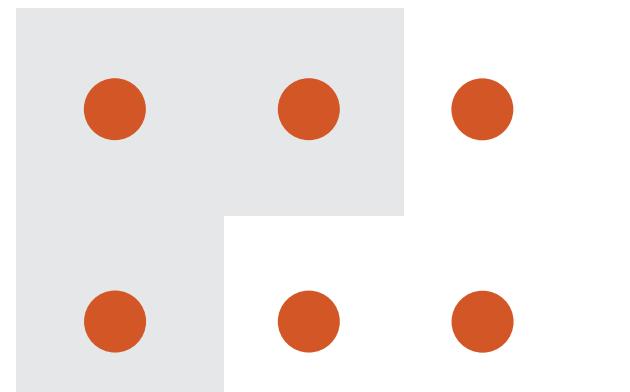


2D

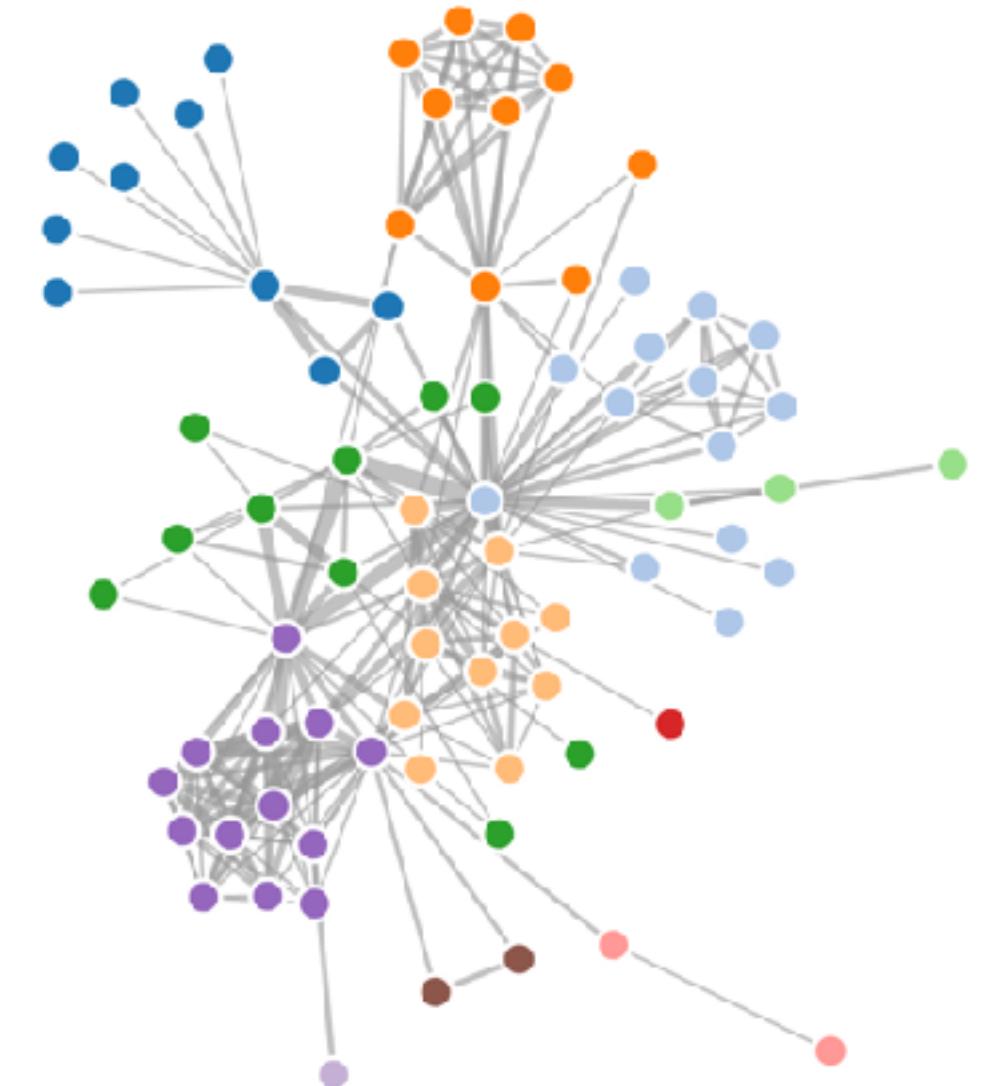
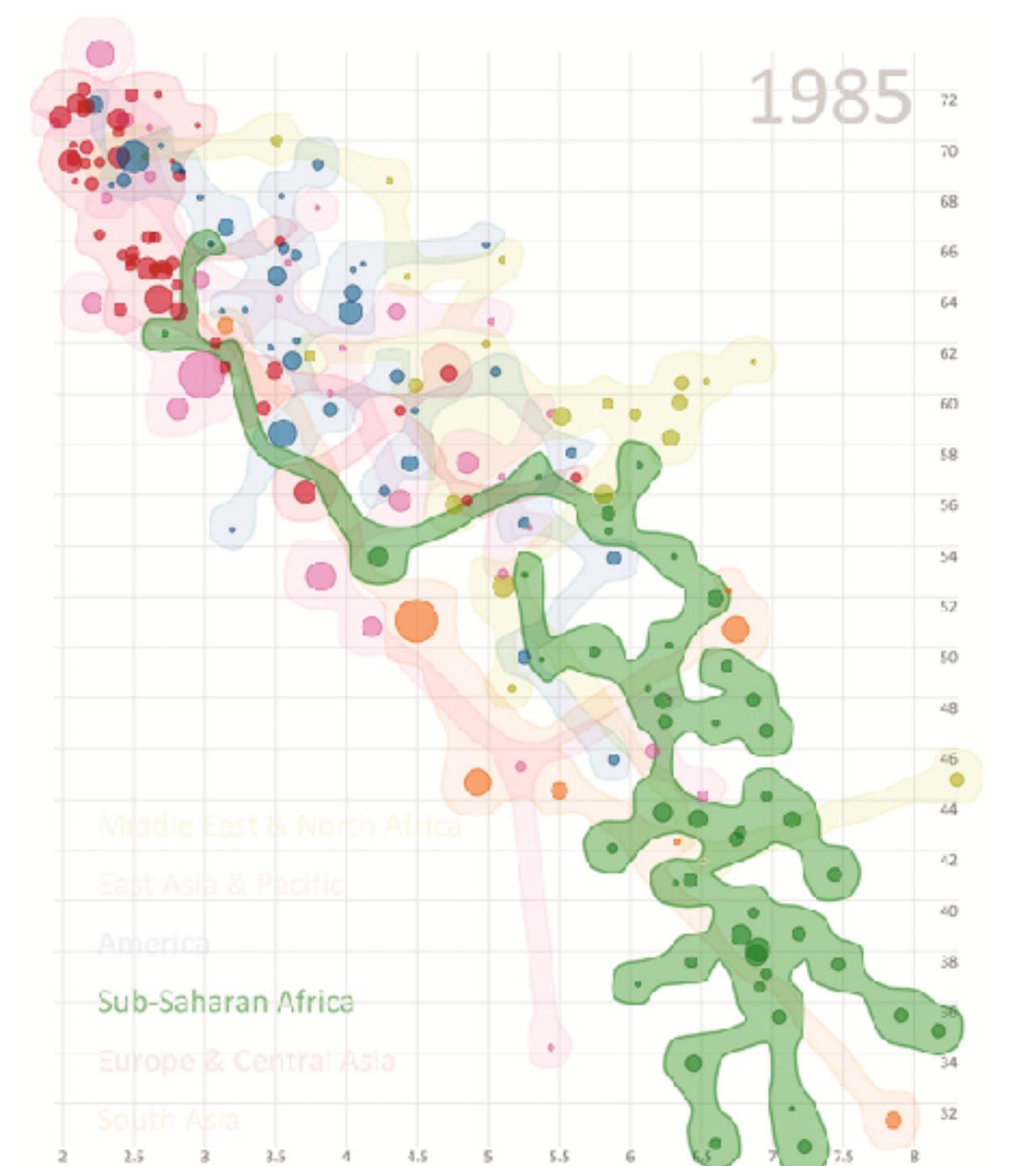
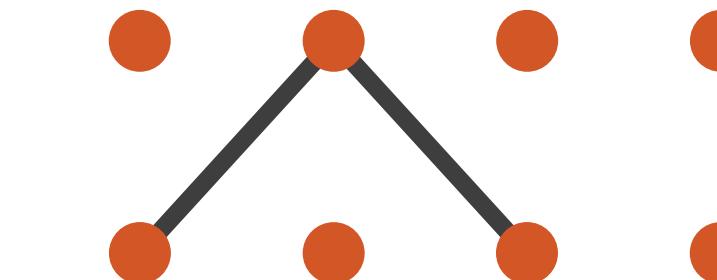
3D mark: Volume, but rarely used

Marks for Links

→ Containment



→ Connection



Channels (aka Visual Variables)

Control appearance
proportional to or
based on attributes

→ Position

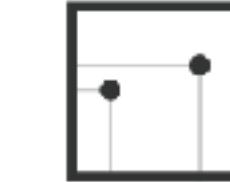
→ Horizontal



→ Vertical



→ Both



→ Color



→ Shape



→ Tilt



→ Size

→ Length



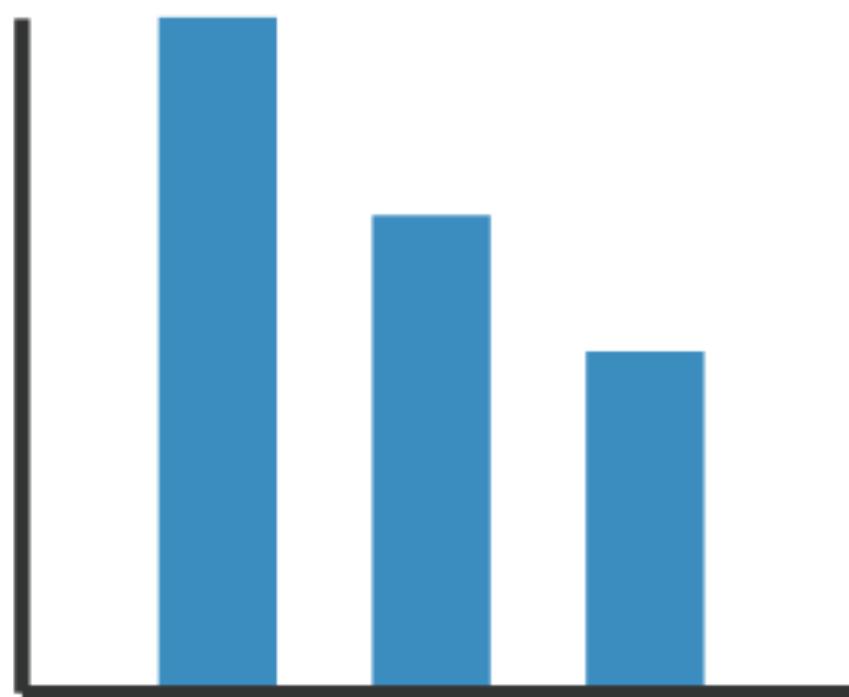
→ Area



→ Volume



Using Marks and Channels

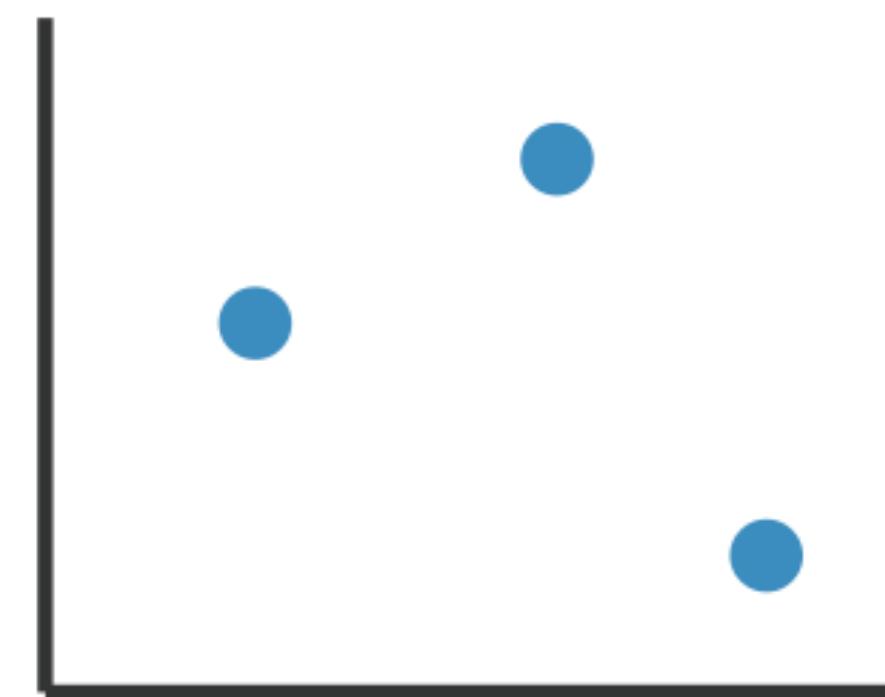


Mark: Line

Channel: Length/Position

1 quantitative attribute

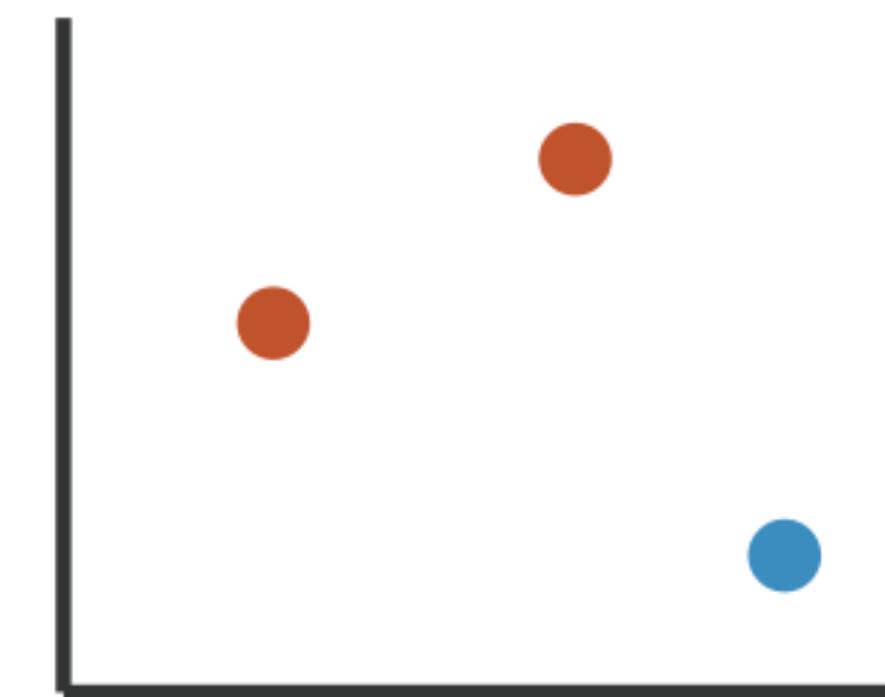
1 categorical attribute



Mark: Point

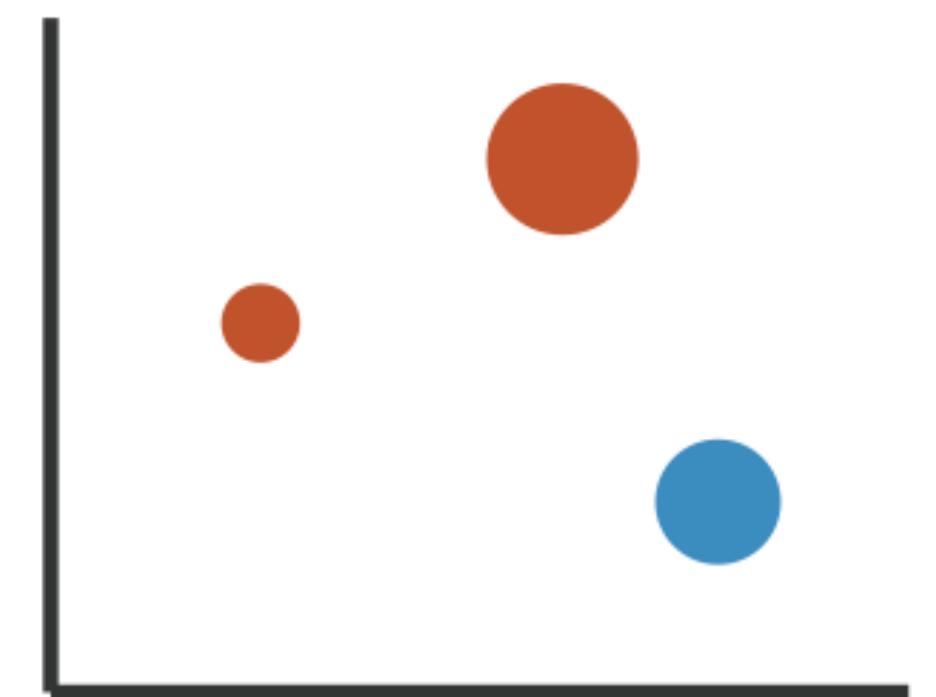
Channel: Position

2 quantitative attr.



Adding Hue

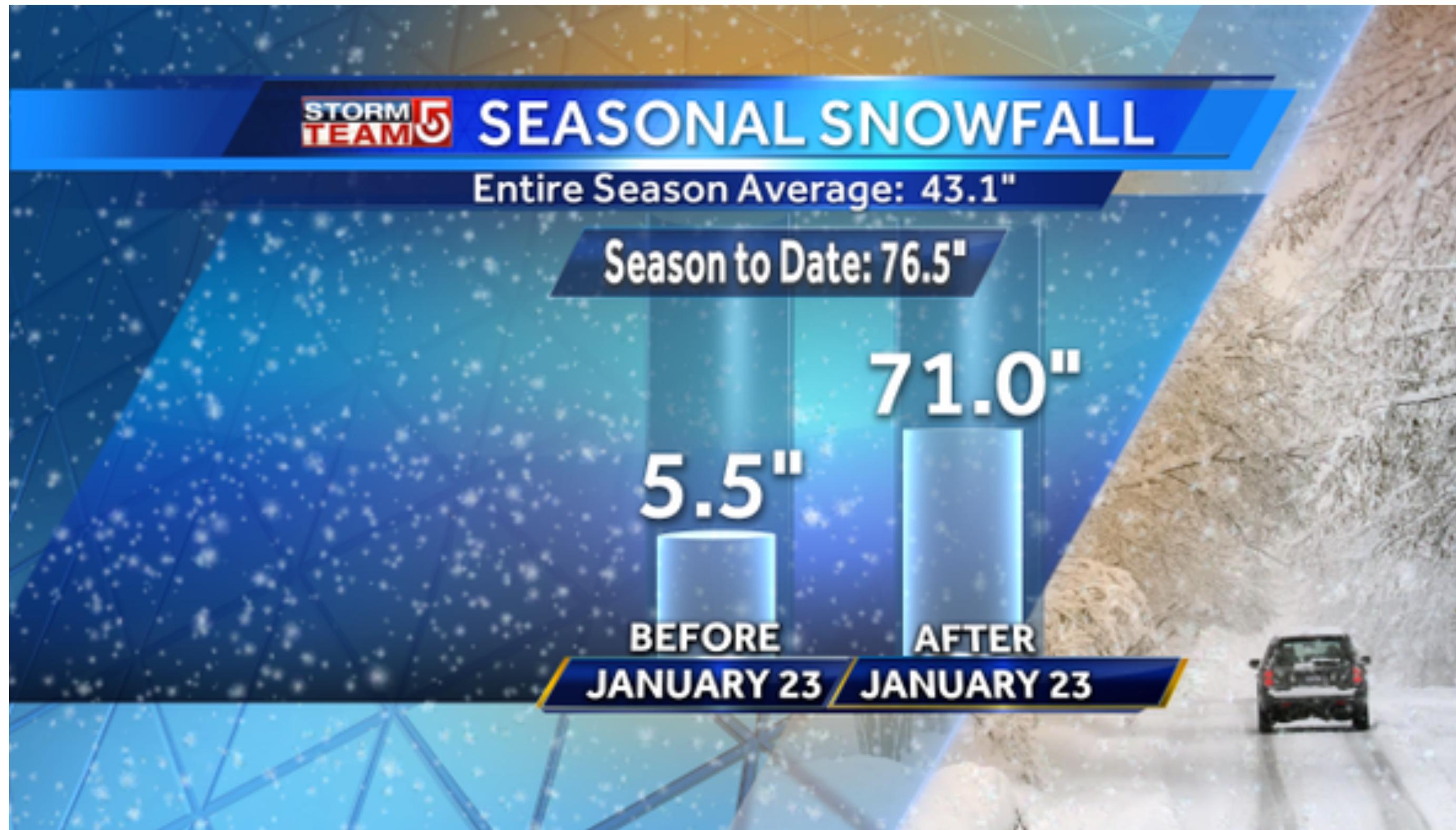
+1 categorical attr.



Adding Size

+1 quantitative attr.

Good bar chart?



Rule: Use channel proportional to data!

Types of Channels

Magnitude Channels

How much?

Position

Length

Saturation ...

Identity Channels

What? Where?

Shape

Color (hue)

Spatial region ...

Ordinal & Quantitative Data

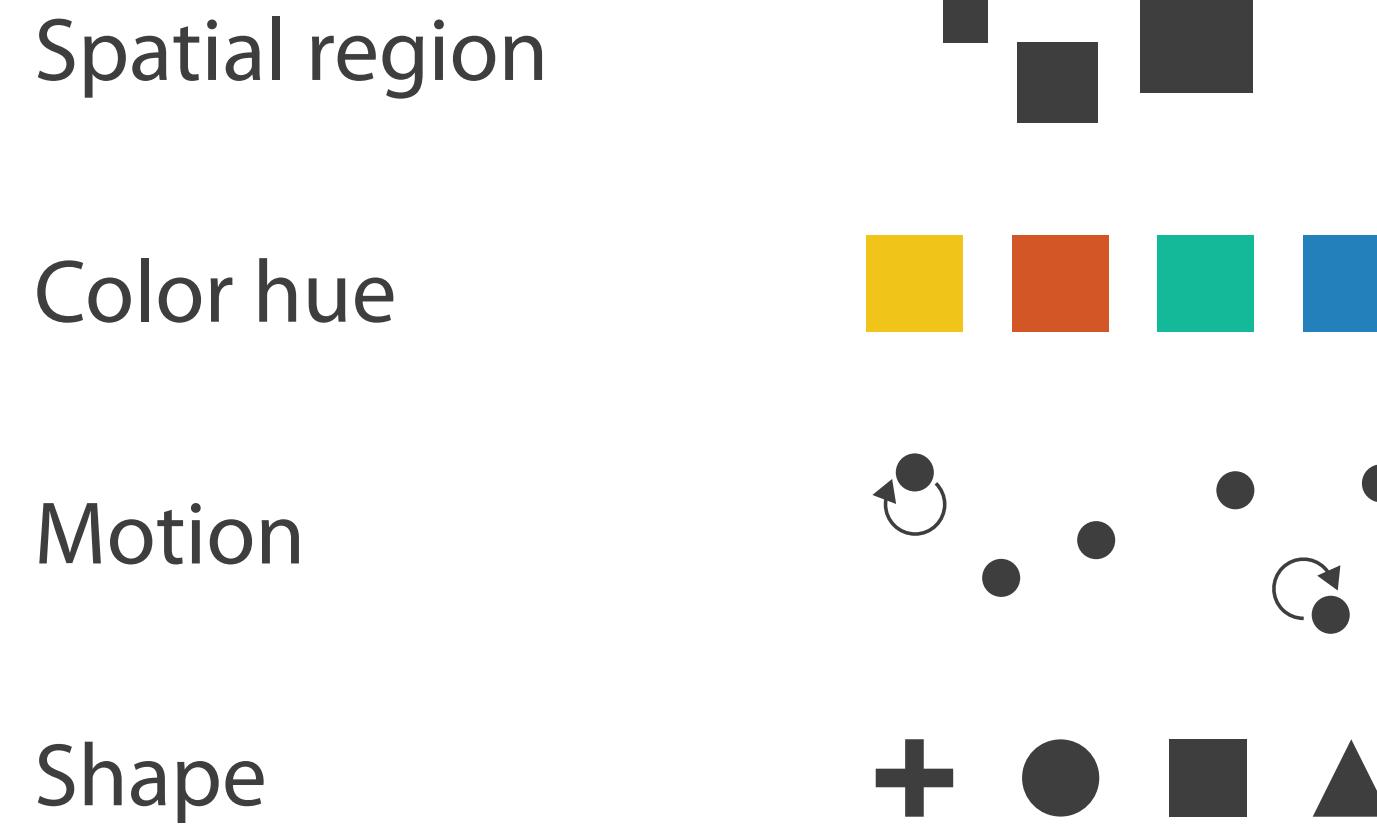
Categorical Data

Channels: Expressiveness Types and Effectiveness Ranks

→ Magnitude Channels: Ordered Attributes



→ Identity Channels: Categorical Attributes



▲ Most
Effectiveness
Least ▼

[Same]

Position

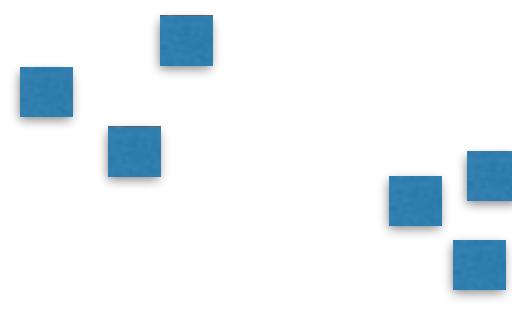
Strongest visual variable

Suitable for all data types

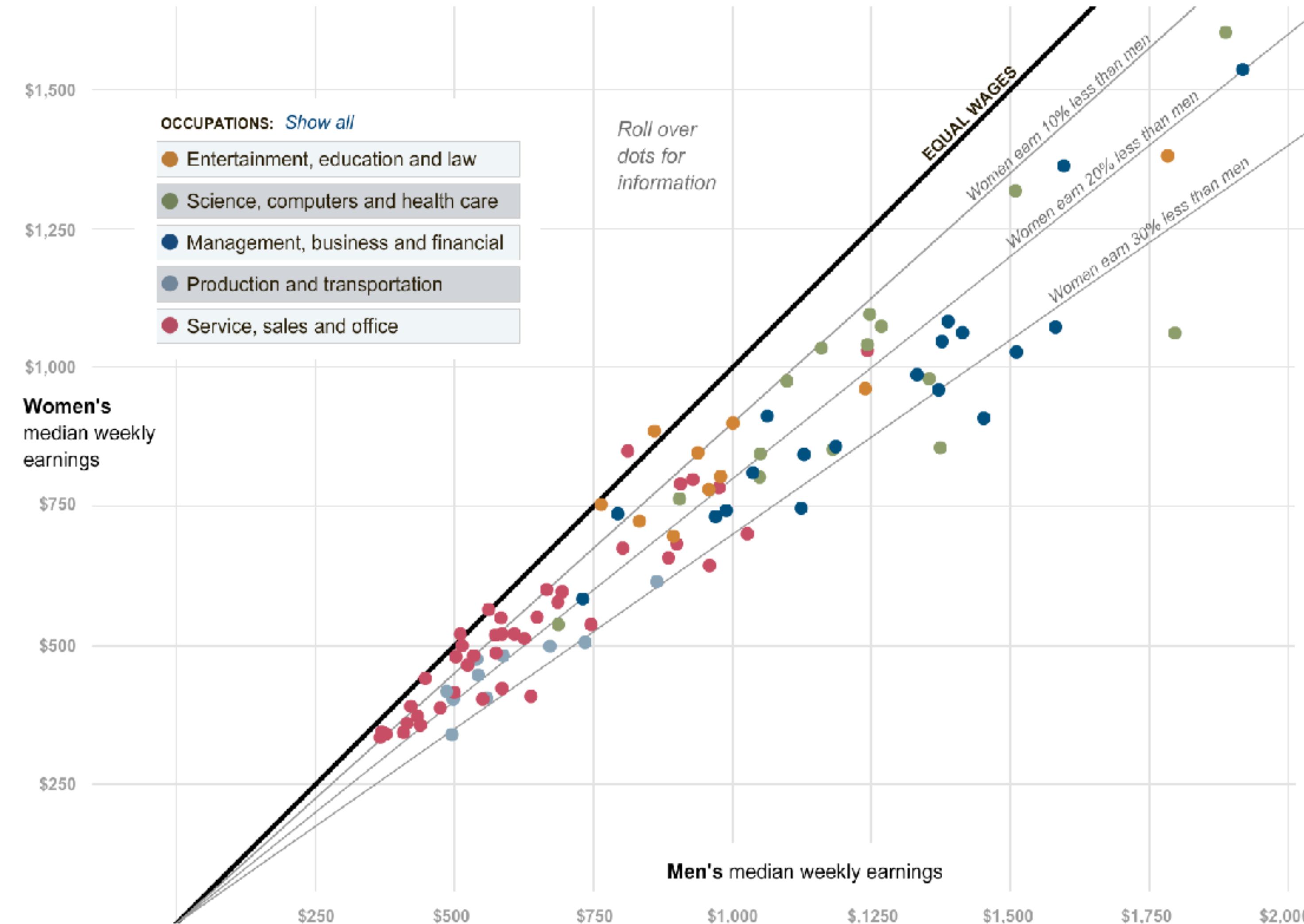
Problems:

Sometimes not available
(spatial data)

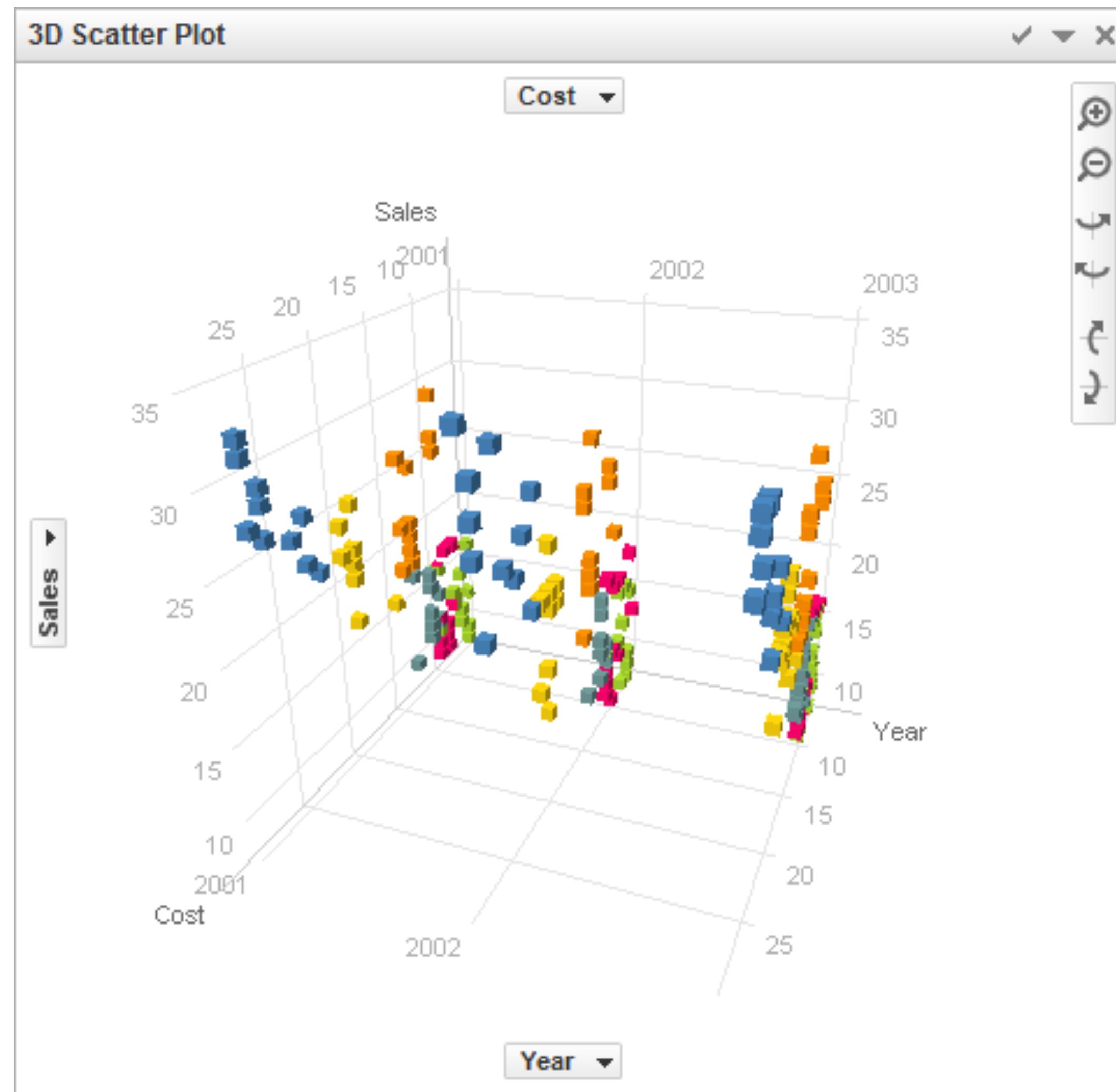
Cluttering



Example: Scatterplot



Position in 3D?



Length & Size

Good for 1D, OK for 2D, Bad for 3D

Easy to see whether one is bigger

Aligned bars use position redundantly



Example 2D Size: Bubbles

Four Ways to Slice Obama's 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

All Spending

Types of Spending

Changes

Department Totals

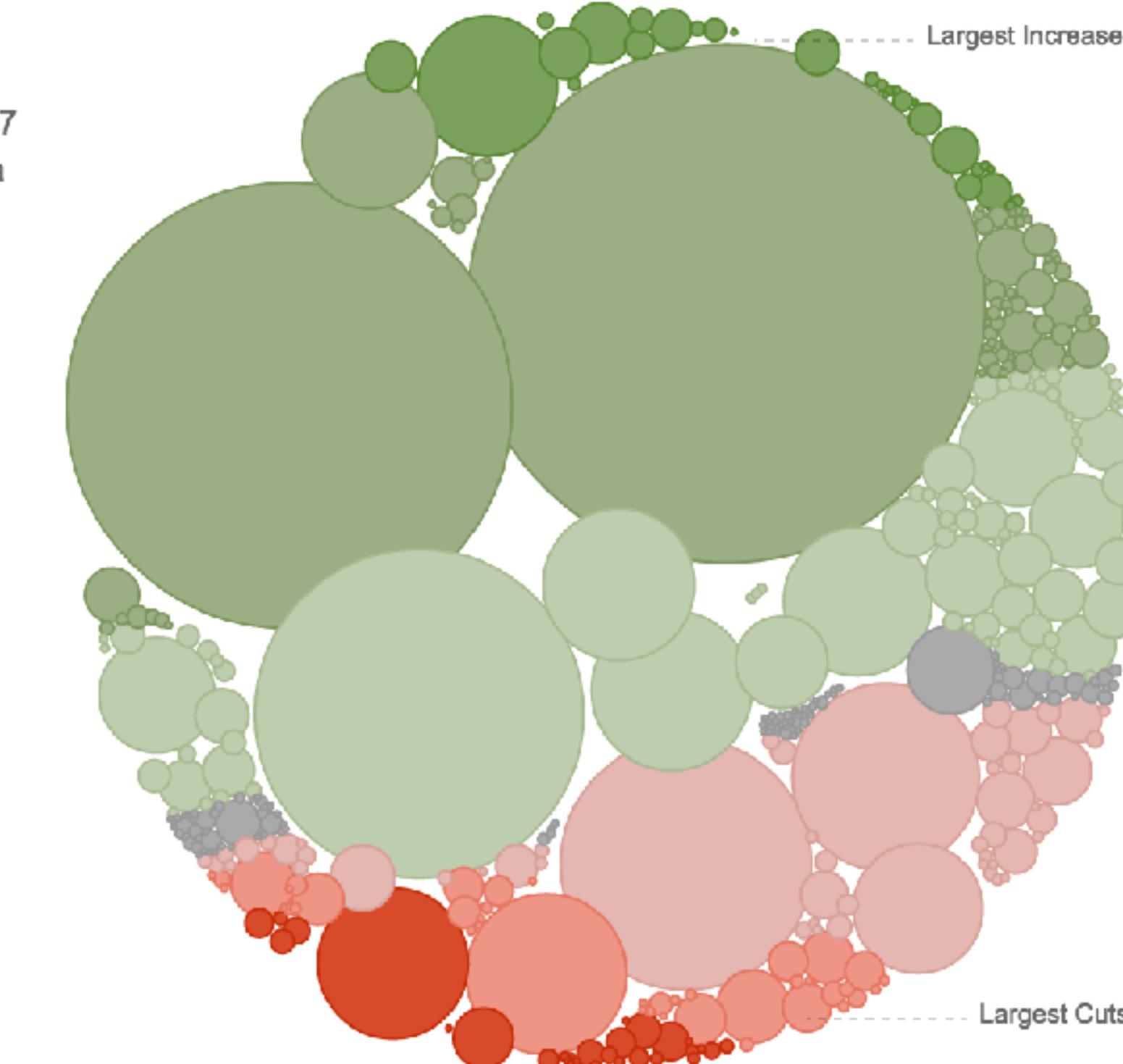
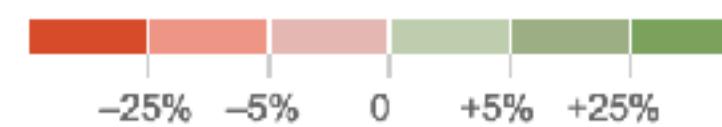
How \$3.7 Trillion Is Spent

Mr. Obama's budget proposal includes \$3.7 trillion in spending in 2013, and forecasts a \$901 billion deficit.

Circles are sized according to the proposed spending.



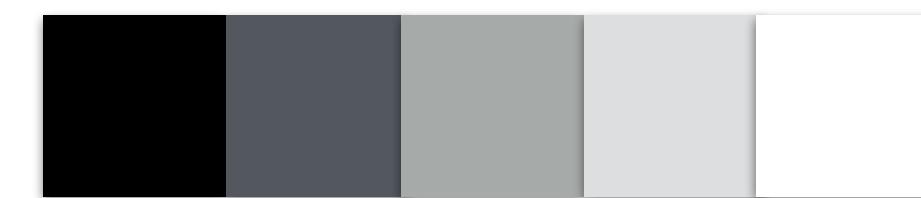
Color shows amount of cut or increase from 2012.



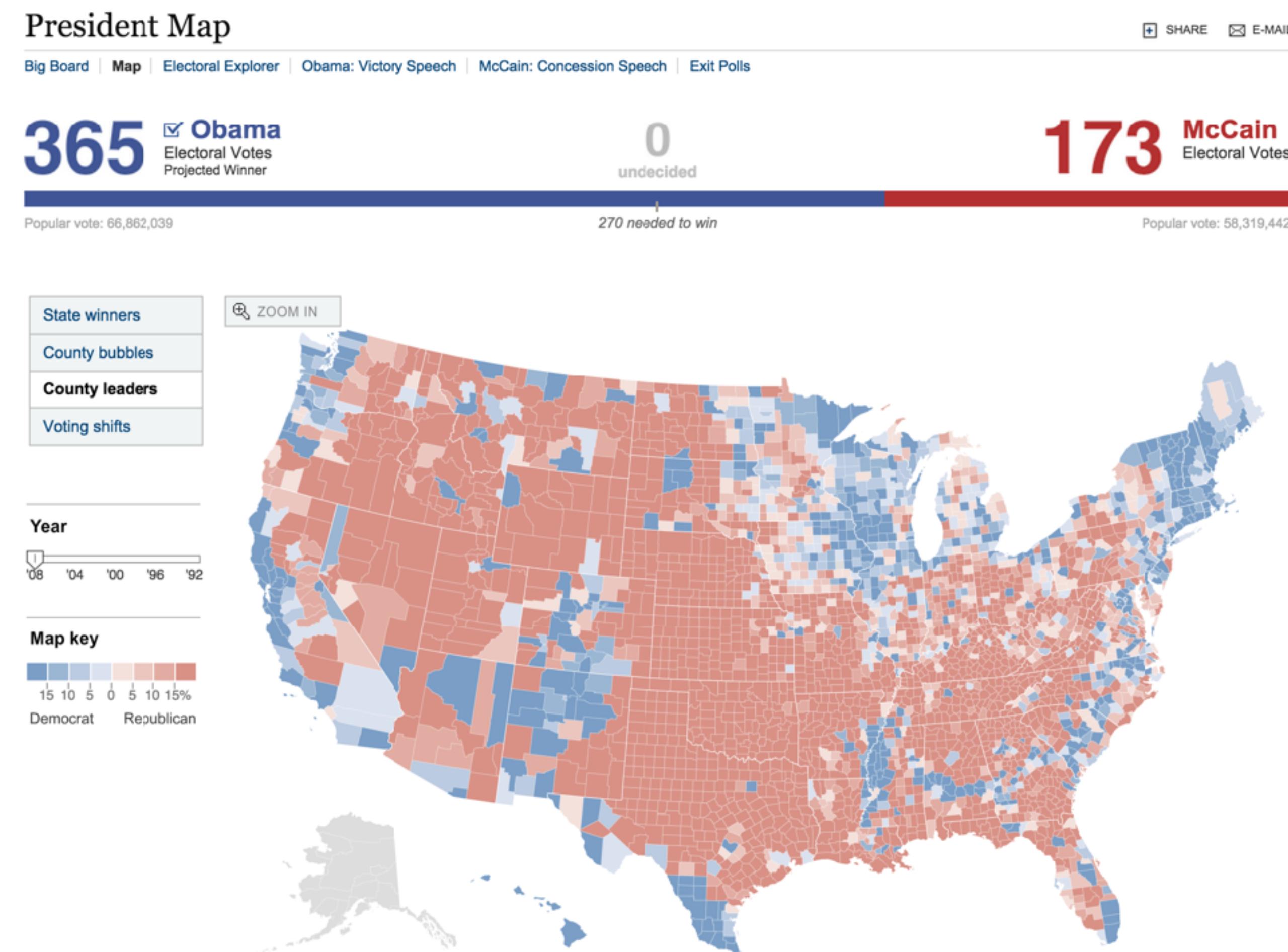
Value/Luminance/Saturation

OK for quantitative data when length & size are used.

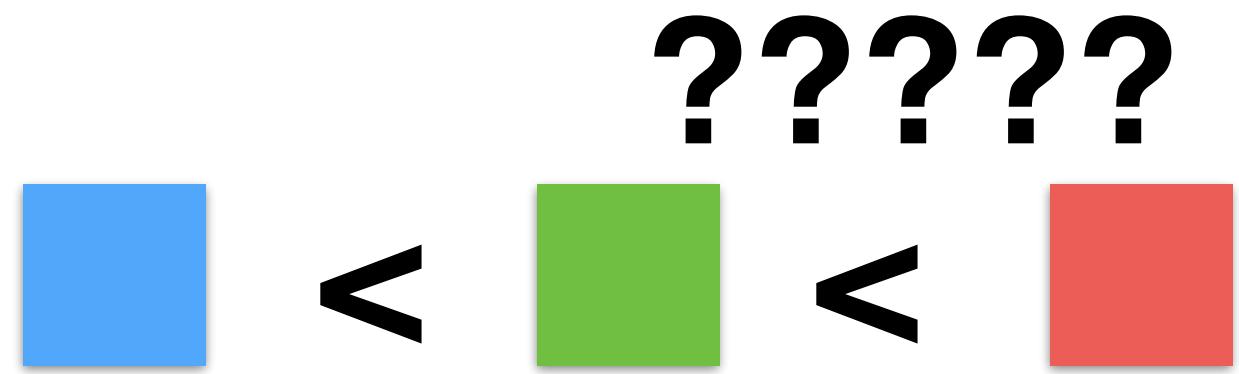
Not very many shades recognizable



Example: Diverging Value-Scale



Color



Good for qualitative data (identity channel)

Limited number of classes/length (~7-10!)

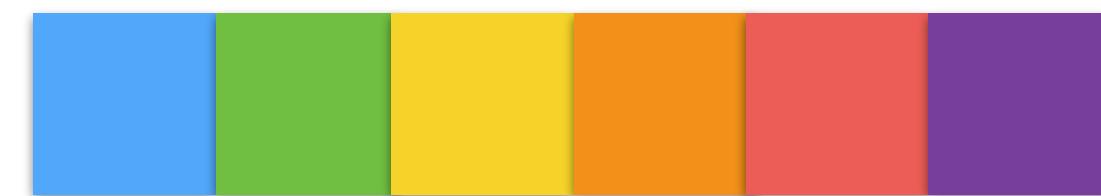
Does not work for quantitative data!

Lots of pitfalls! Be careful!

My rule:

minimize color use for encoding data

use for brushing



What is a colormap?

[0,8]



specifies a mapping between color and values

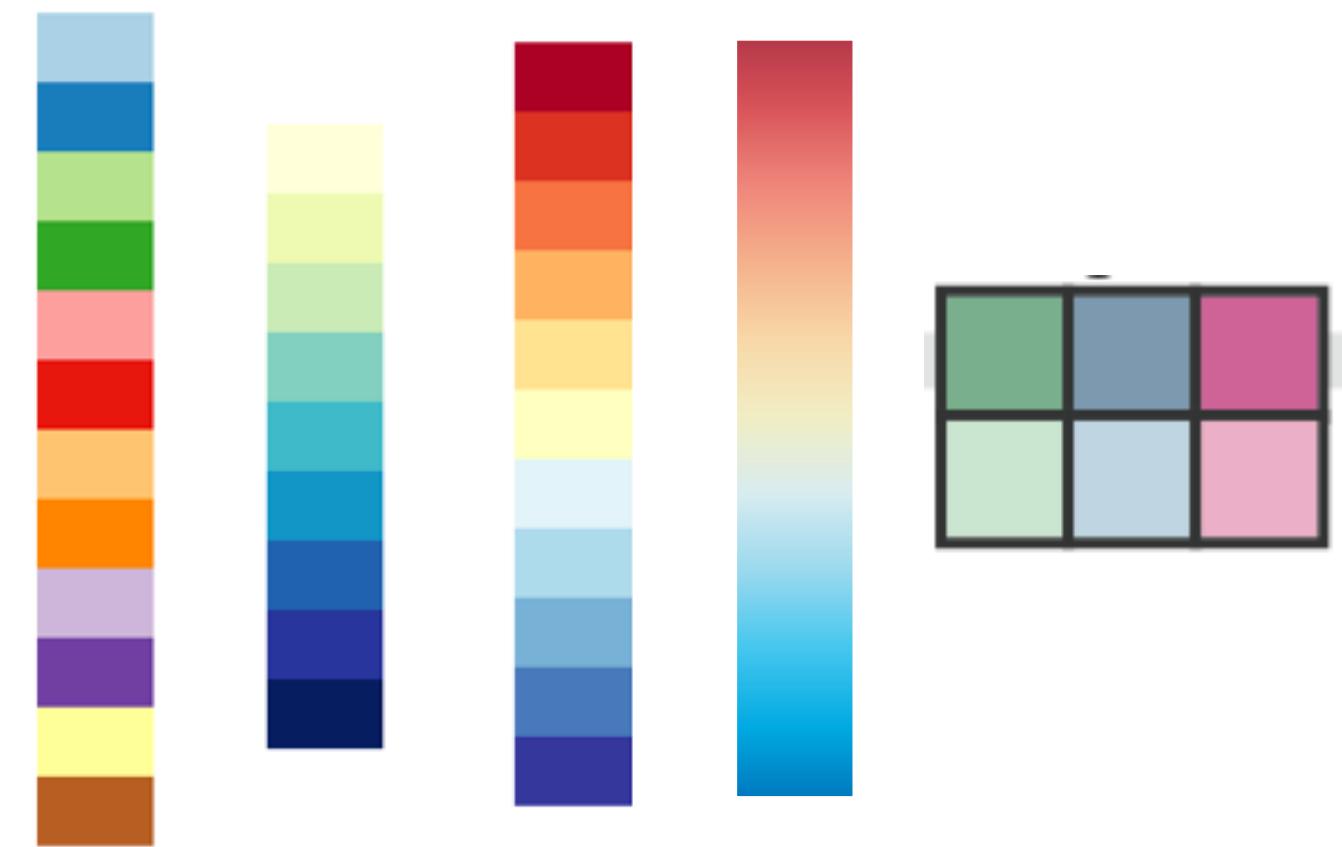
categorical vs ordered

sequential vs diverging

segmented vs continuous

univariate vs bivariate

expressiveness: match colormap to attribute characteristics!

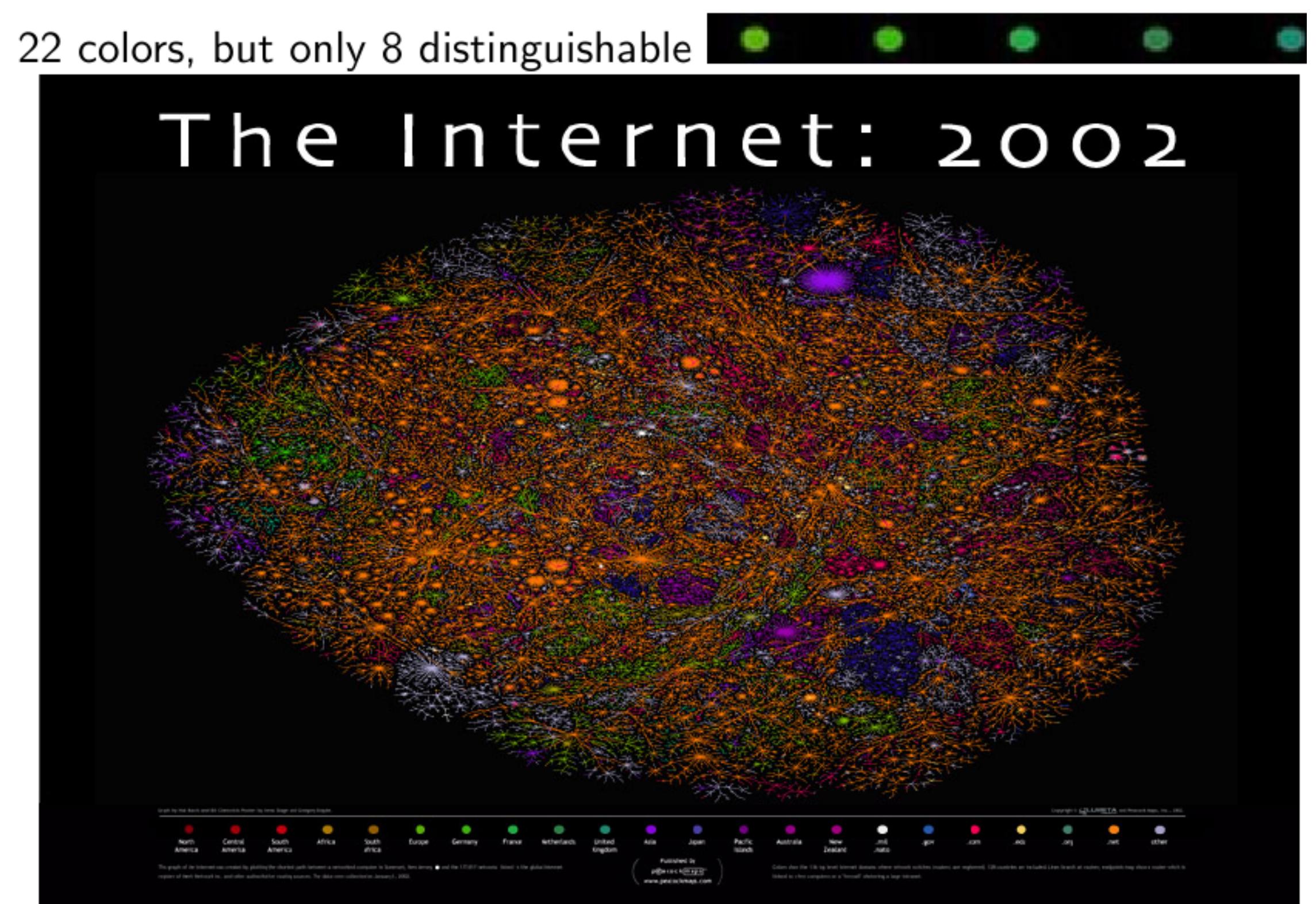


Qualitative Data Vis

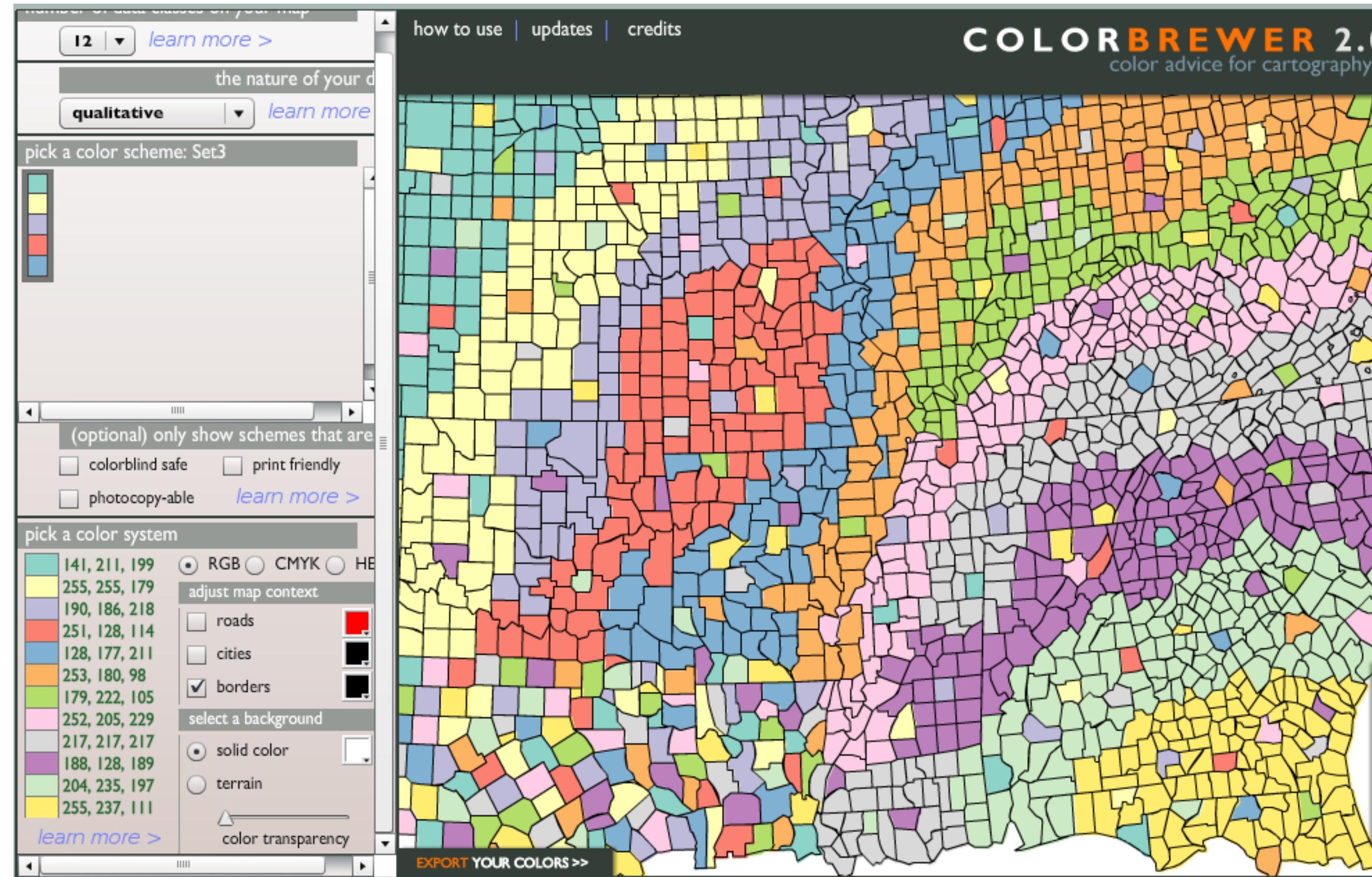
Color labeling (nominal information coding)

recommended: about 6, no more than 10

22 colors, but only 8 distinguishable



Application: Labeling



Carefully designed color scheme for 12 colors [colorbrewer]

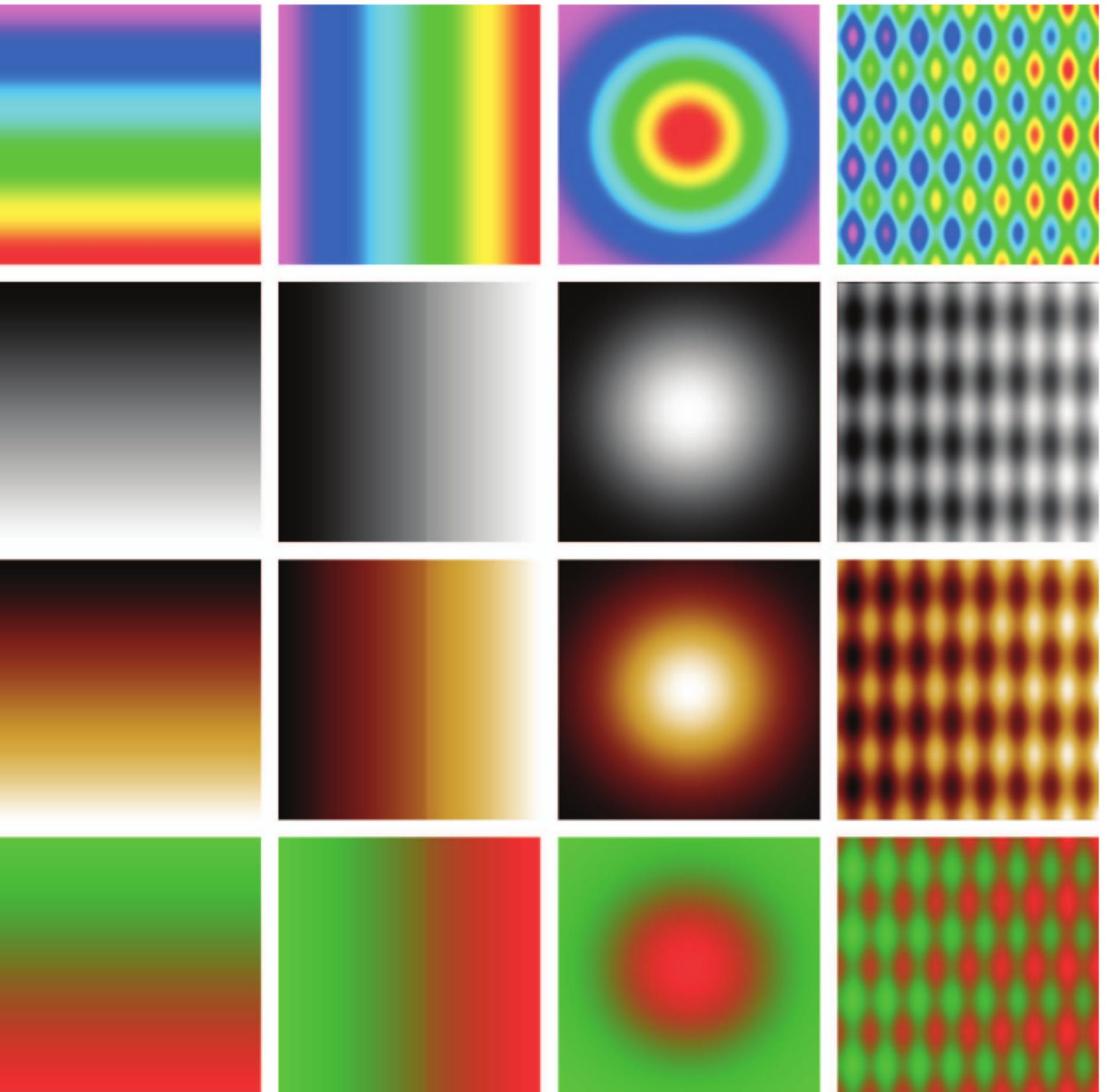
Quantitative Data Vis

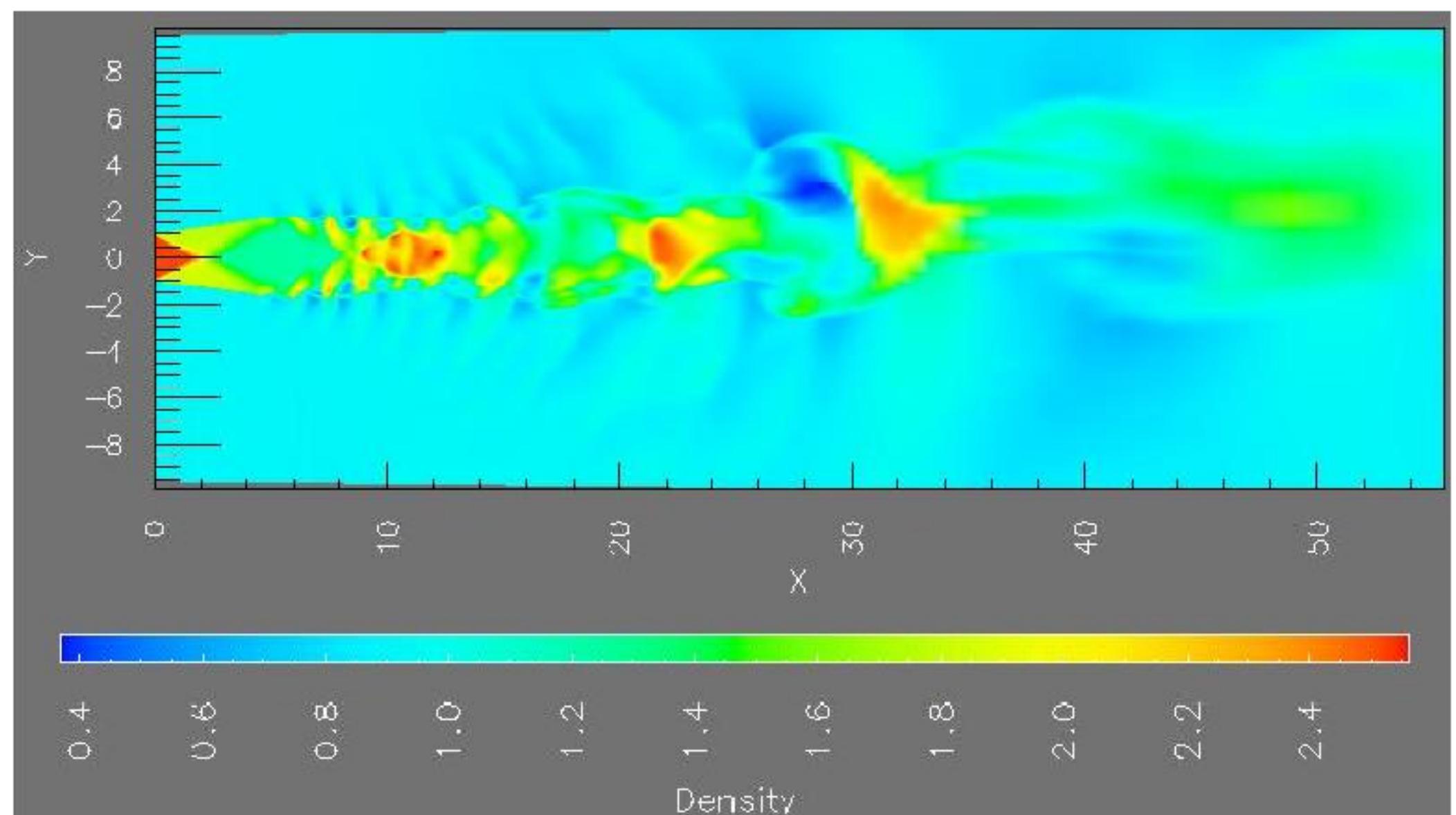
use value

saturation works but not as good

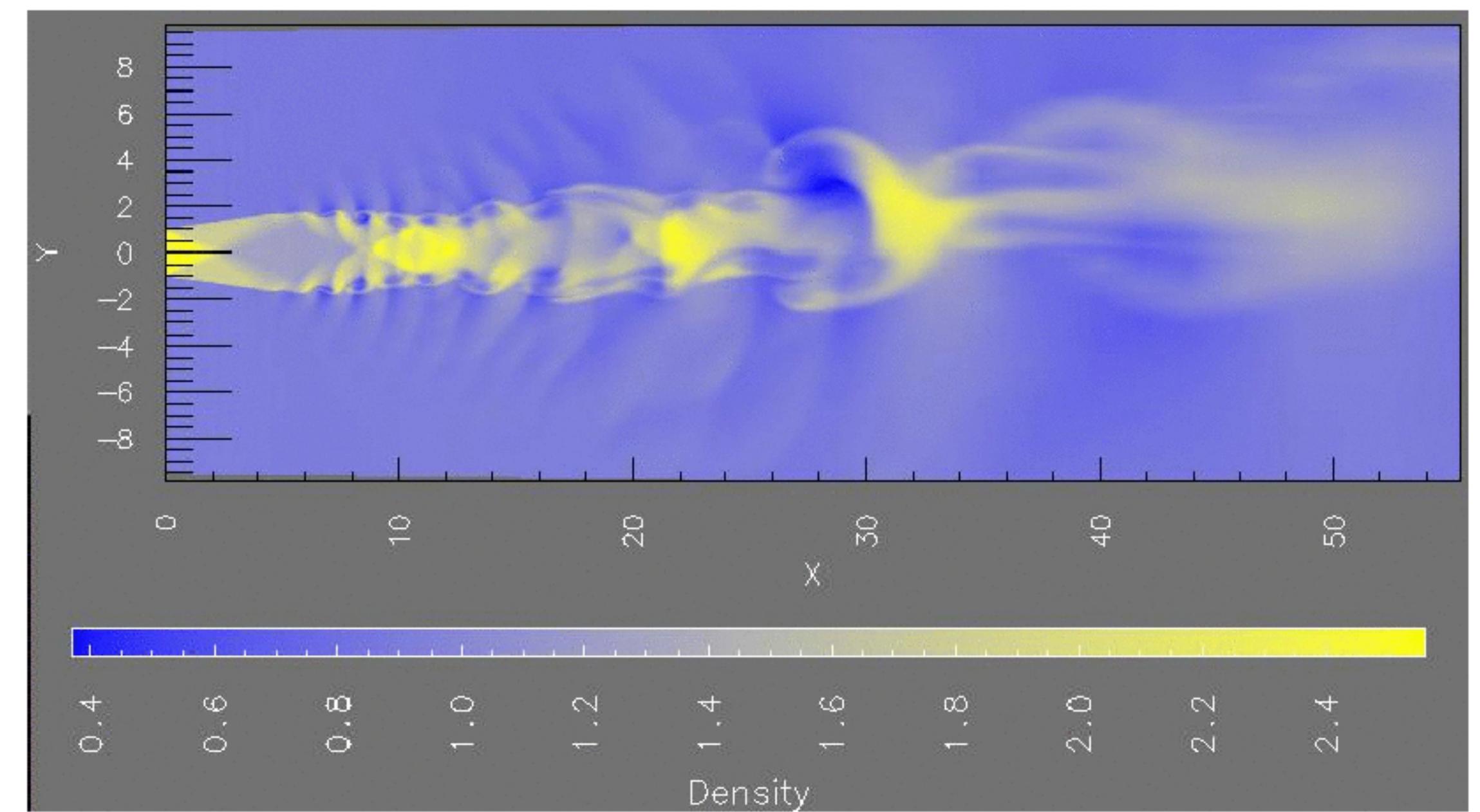
don't use hue!

Danger: rainbow color map



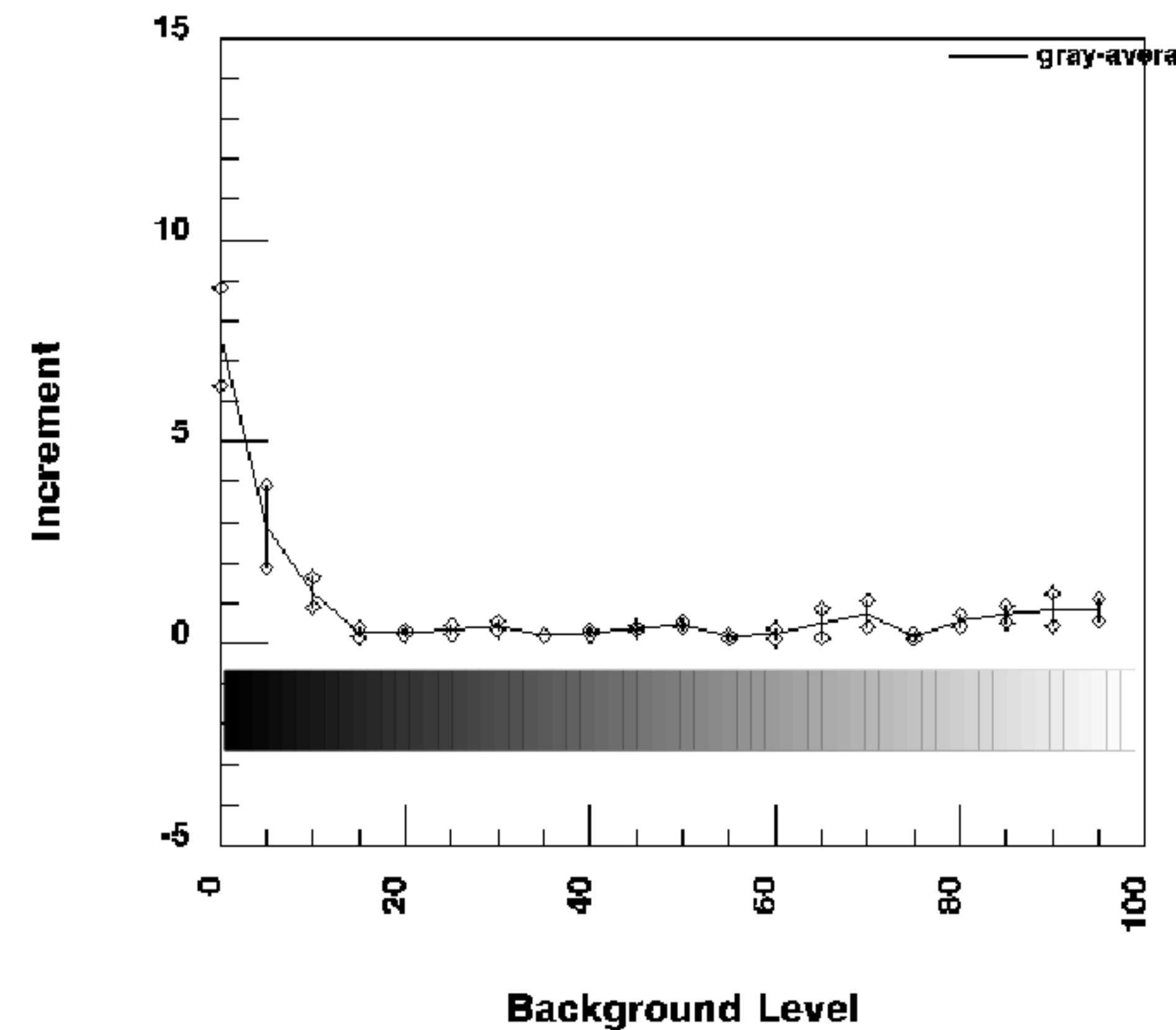
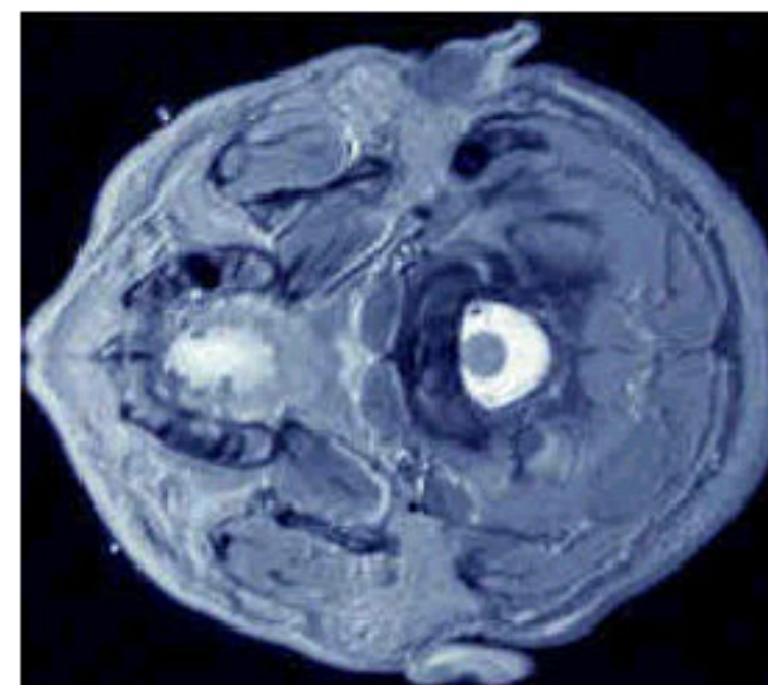


[Rogowitz and Treinish, Why Should Engineers and Scientists Be Worried About Color? <http://www.research.ibm.com/people/l/lloydt/color/color.HTM>]

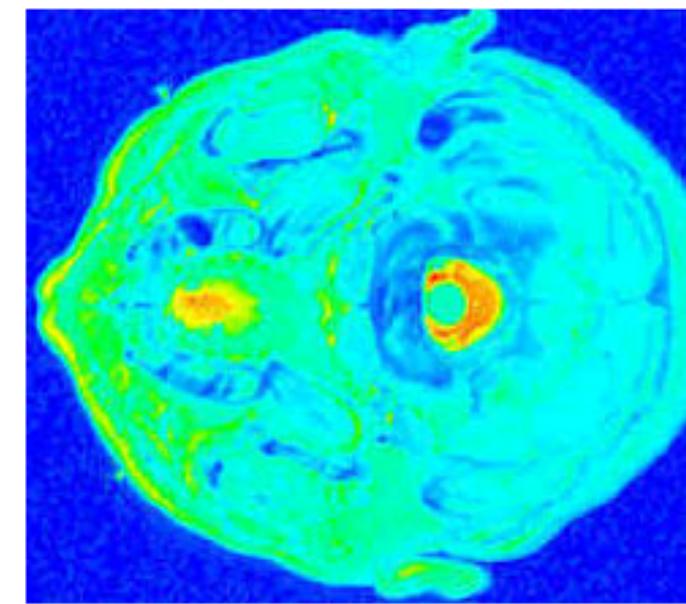


[Rogowitz and Treinish, How NOT to Lie with Visualization, www.research.ibm.com/dx/proceedings/pravda/truevis.htm]

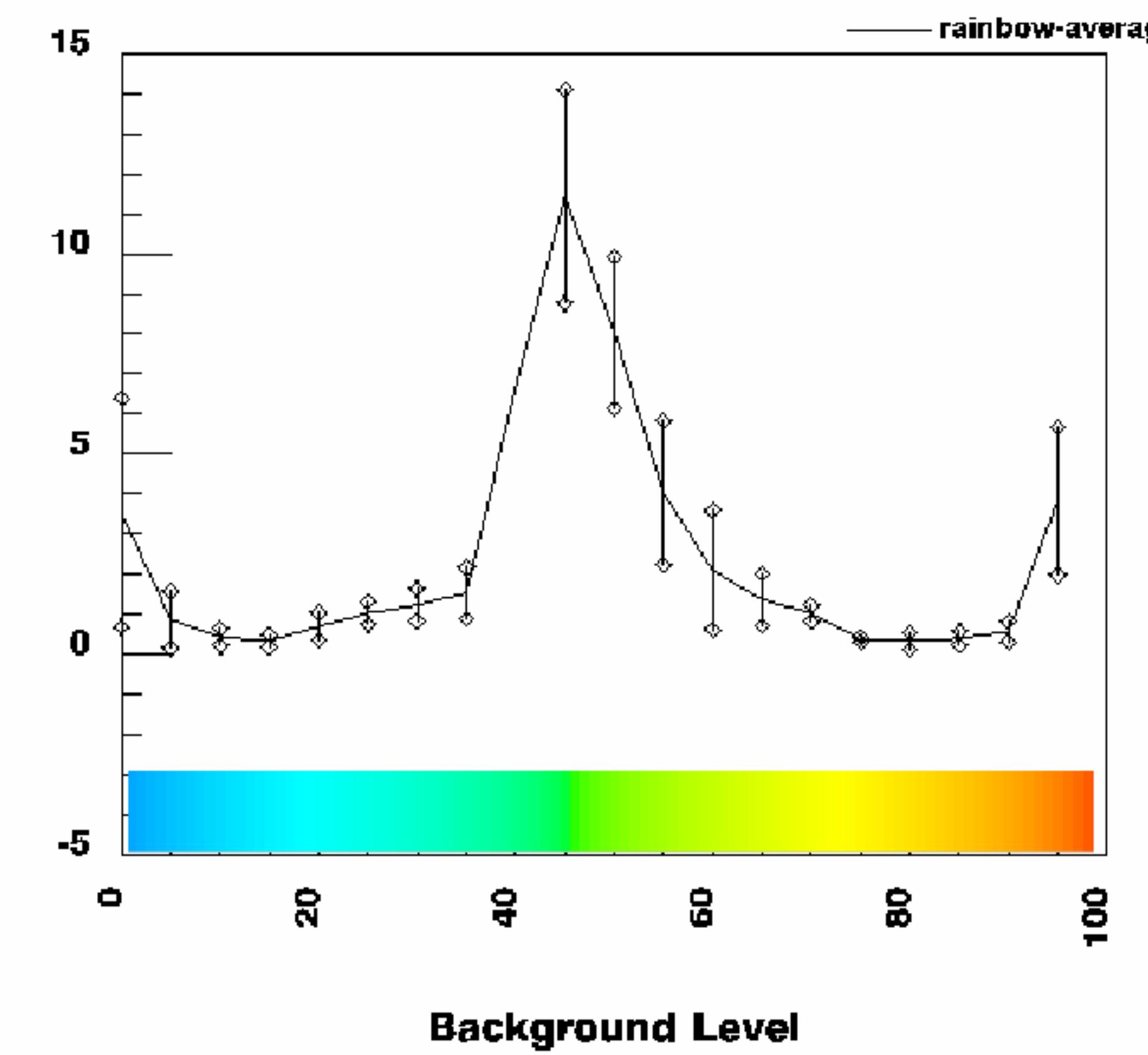
Increment Threshold Experiment: Luminance



Increment Threshold Experiment: Rainbow Color Scale



Increment



Color Blindness

10% of males, 1% of females (probably due to X-chromosomal recessive inheritance)

Most common: red-green weakness / blindness

Reason: lack of medium or long wavelength receptors, or altered spectral sensitivity (most common: green shift)

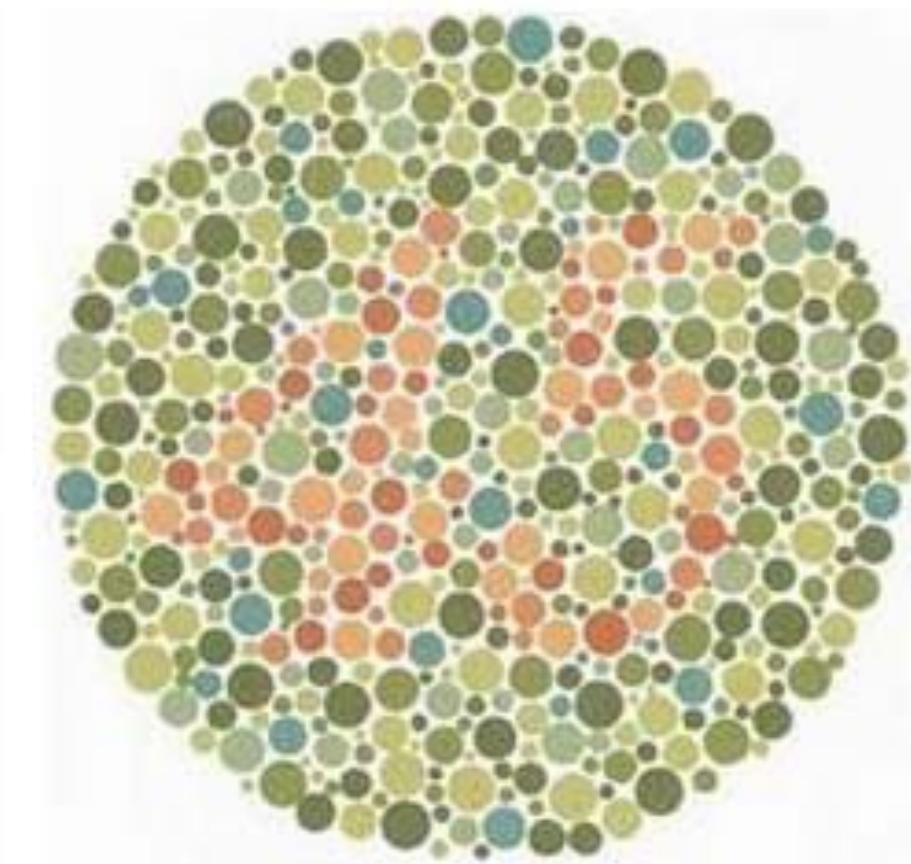
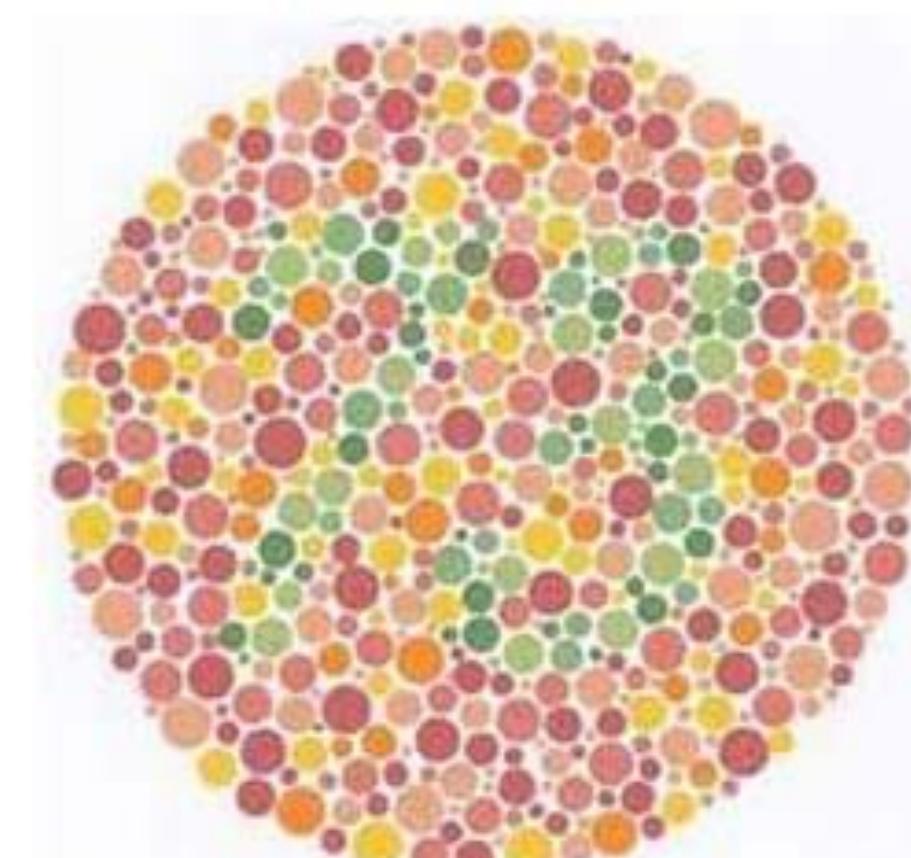
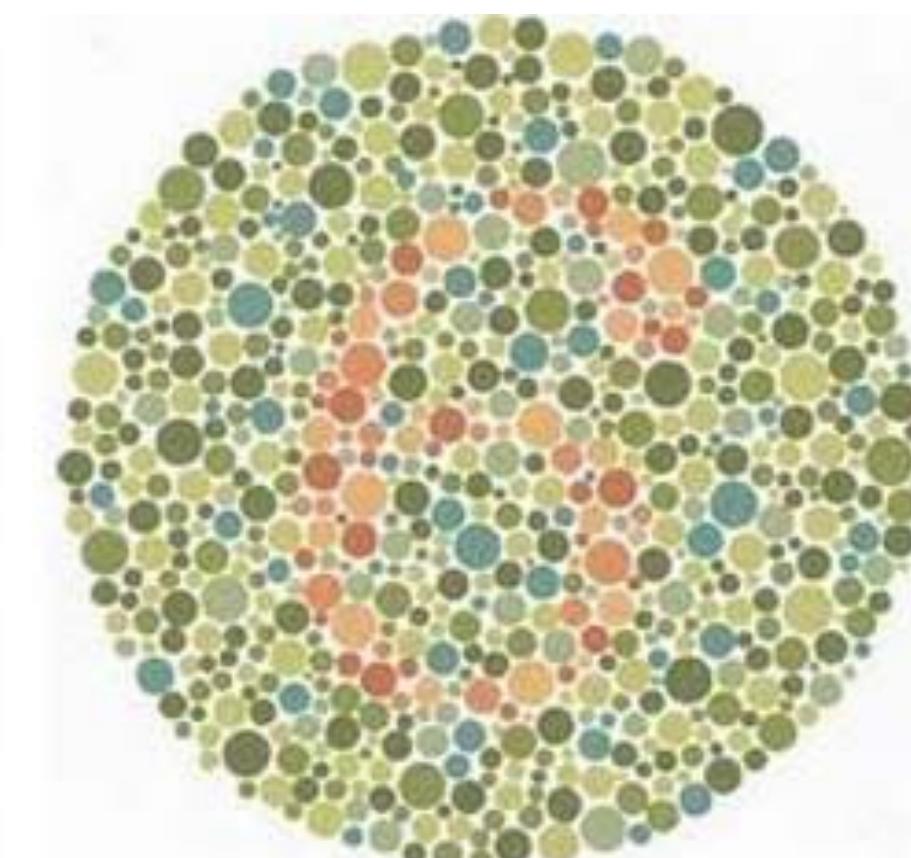
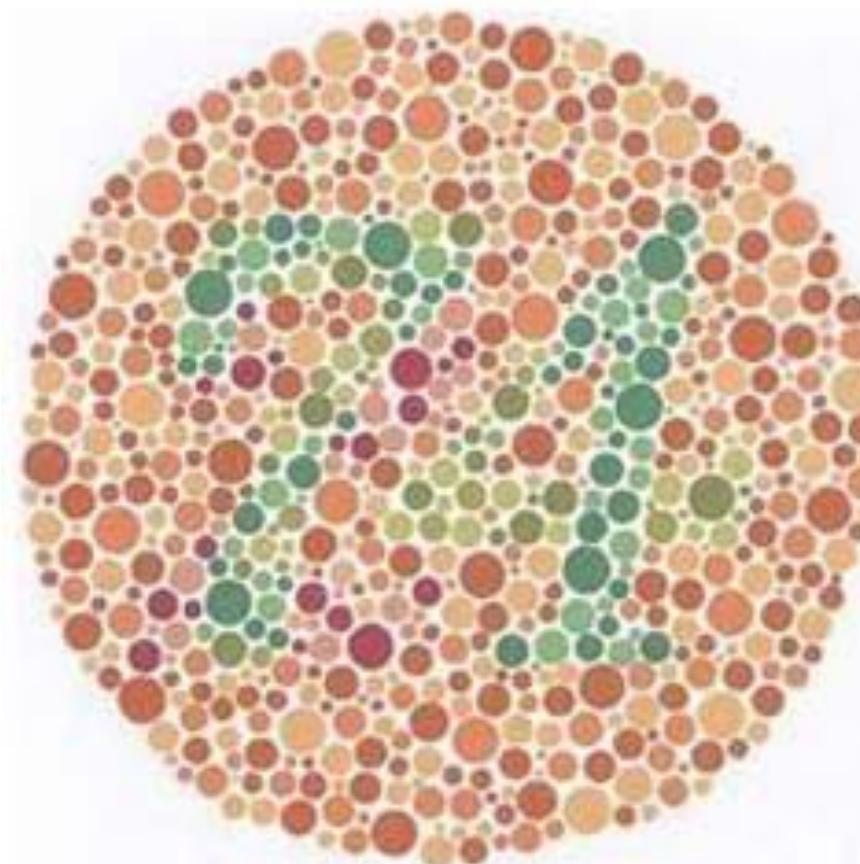
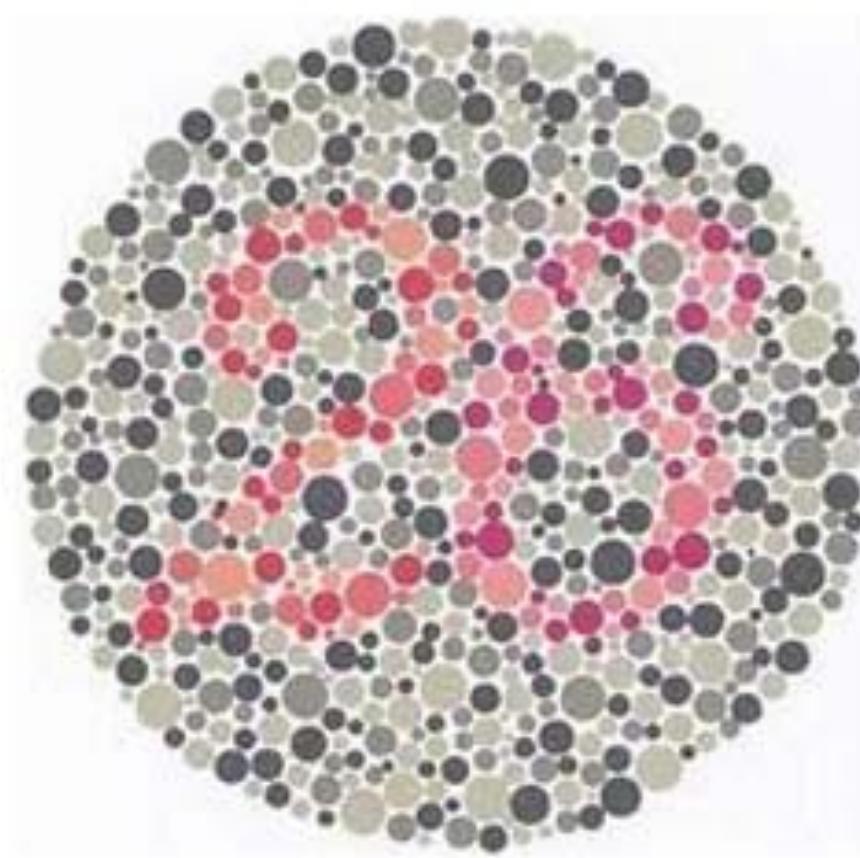
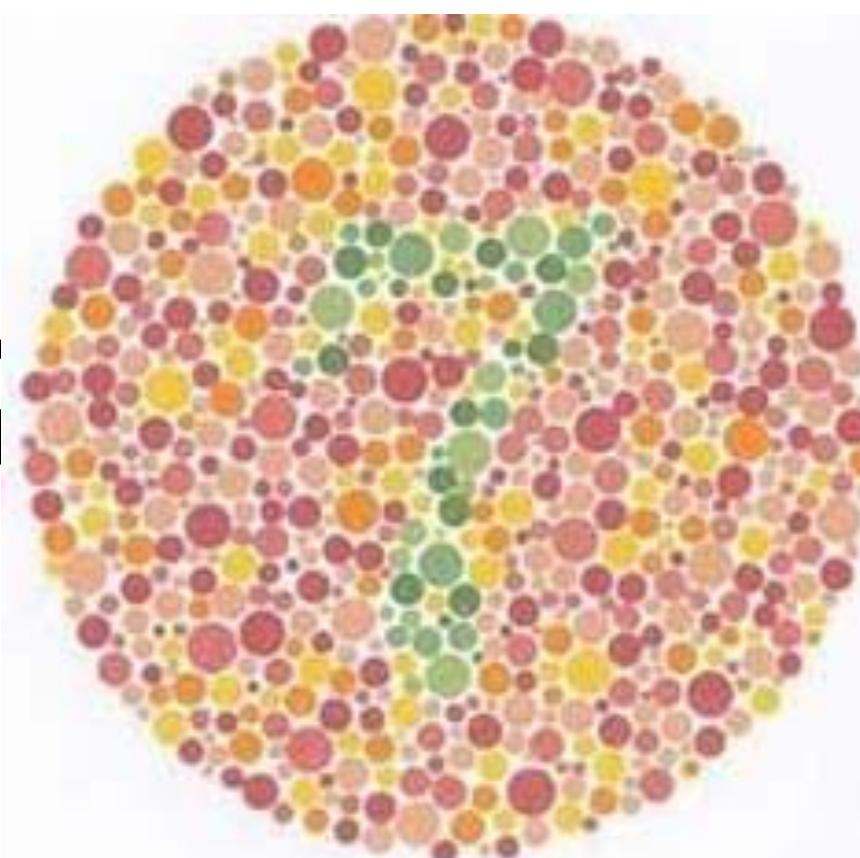
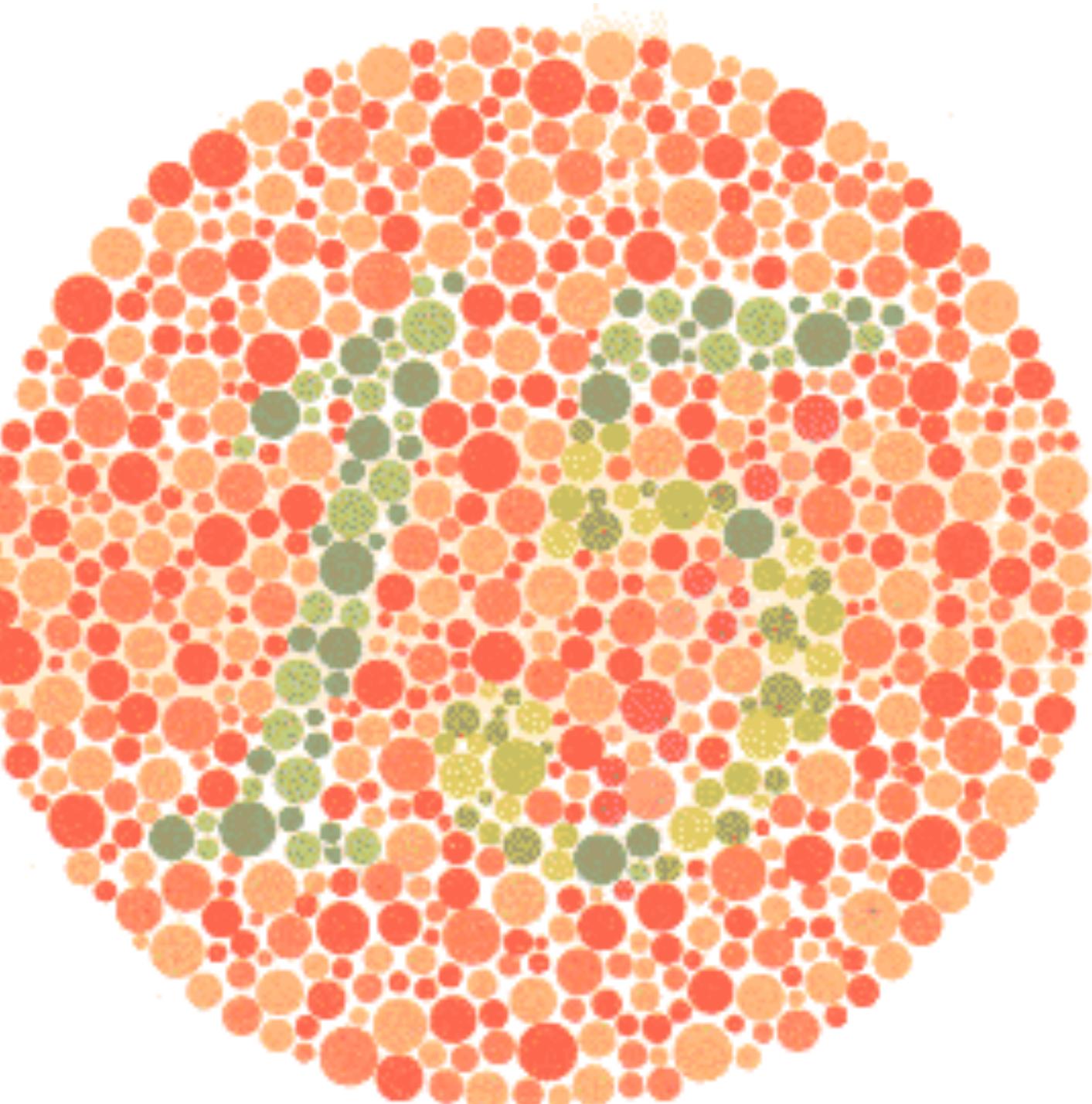
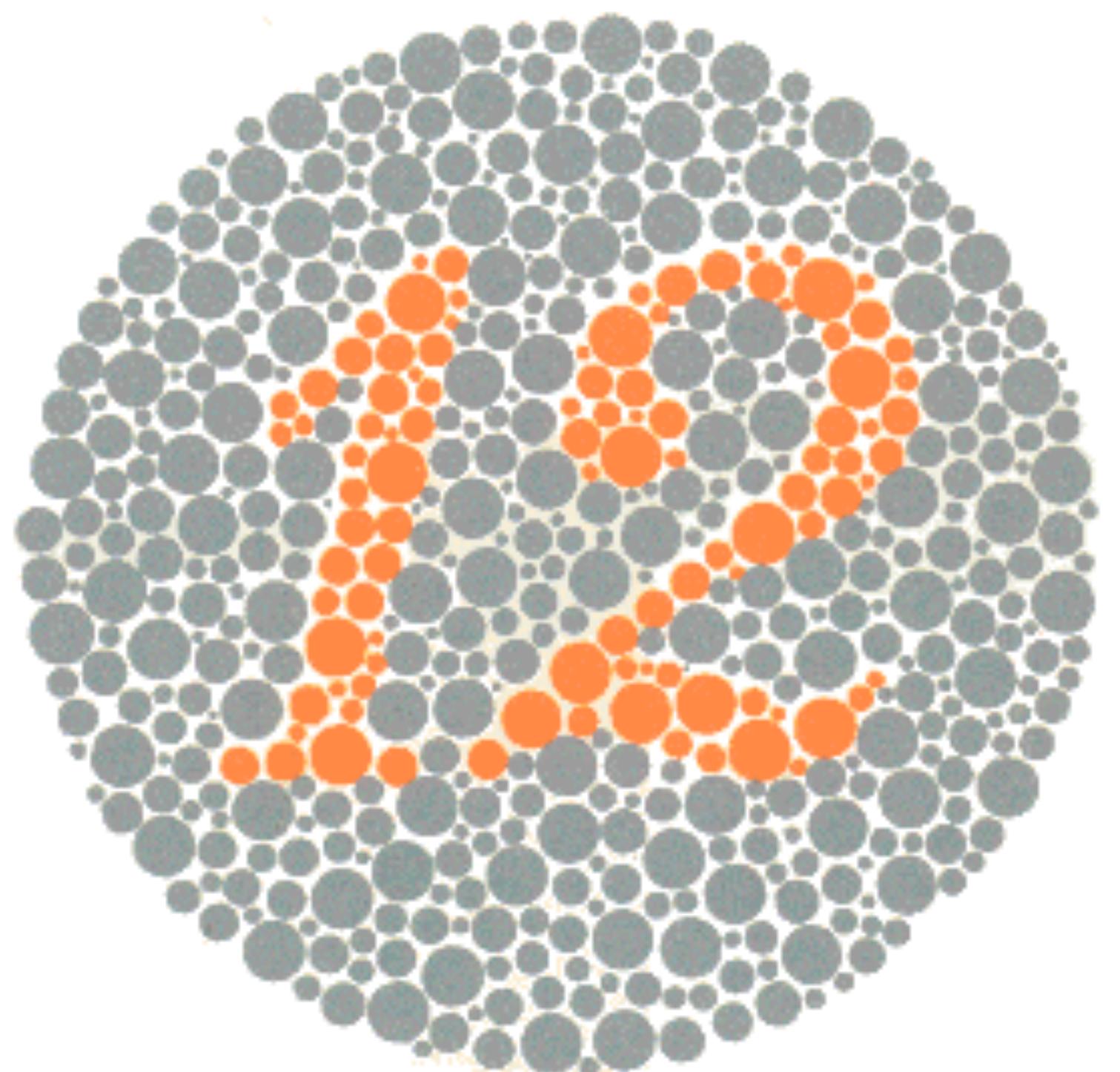


Normal Color Perception

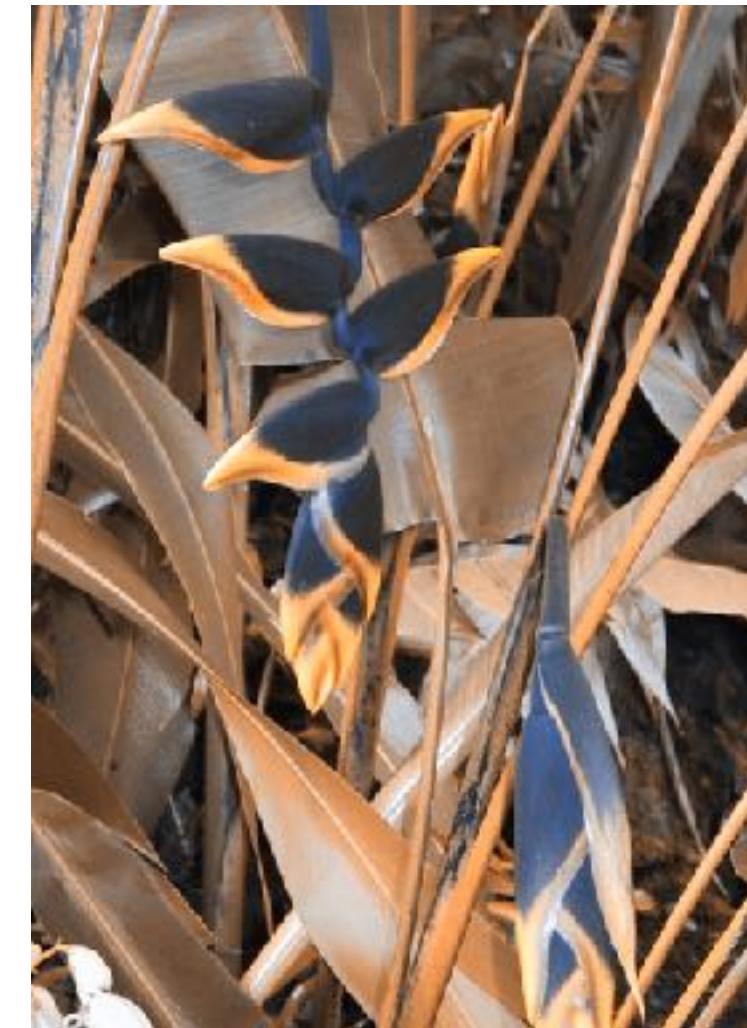
Deuteranopia (no green receptors)

Protanopia (no red receptors)

Color Blindness Tests



Color Blindness



Protanope

No L cones

Red / green
deficiencies

Deuteranope

No M cones

Tritanope

No S cones

Blue / Yellow
deficiency

All Spending Types of Spending Changes Department Totals

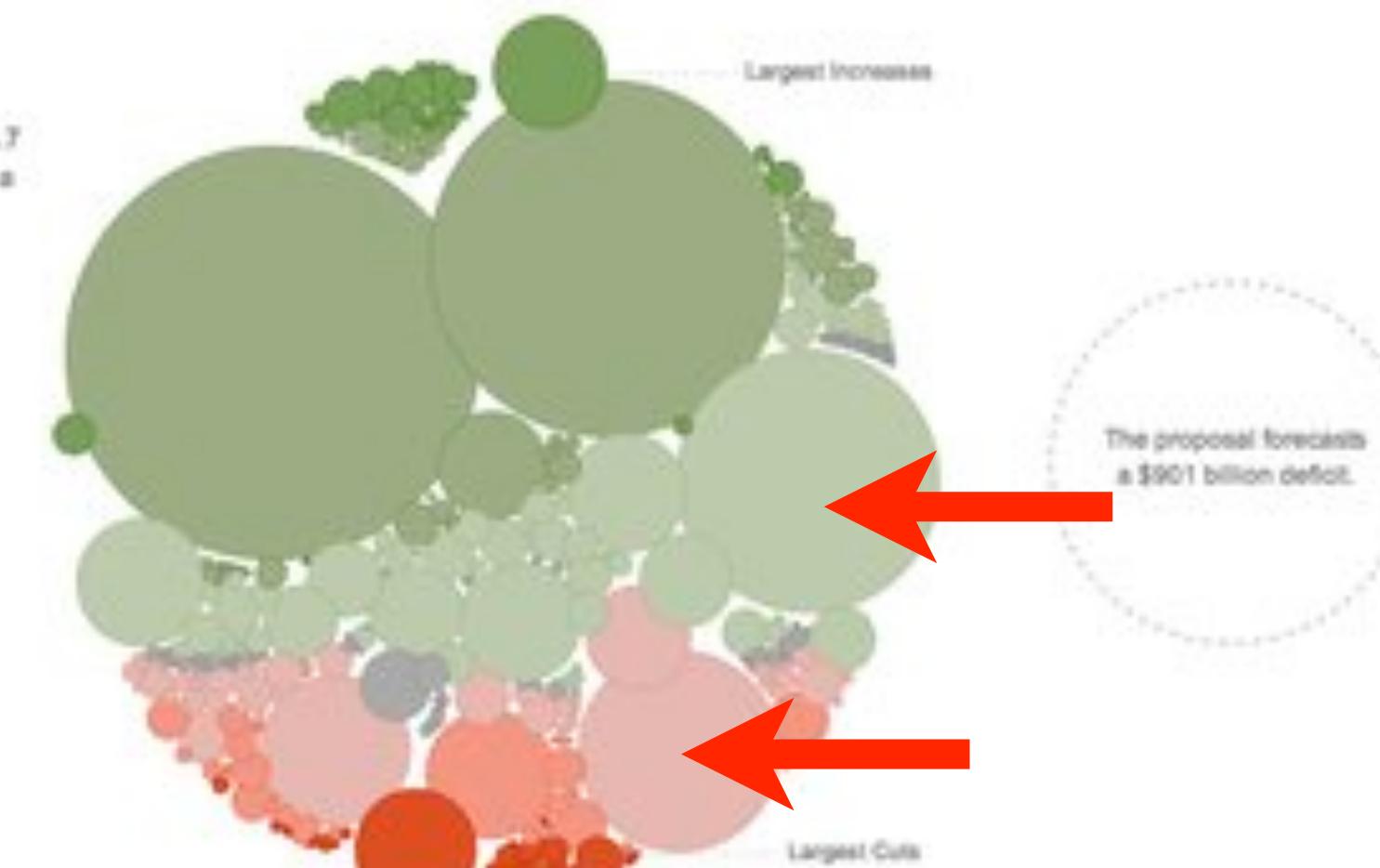
How \$3.7 Trillion Is Spent

Mr. Obama's budget proposal includes \$3.7 trillion in spending in 2013, and forecasts a \$901 billion deficit.

Circles are sized according to the proposed spending.



Color shows amount of cut or increase from 2012.



All Spending Types of Spending Changes Department Totals

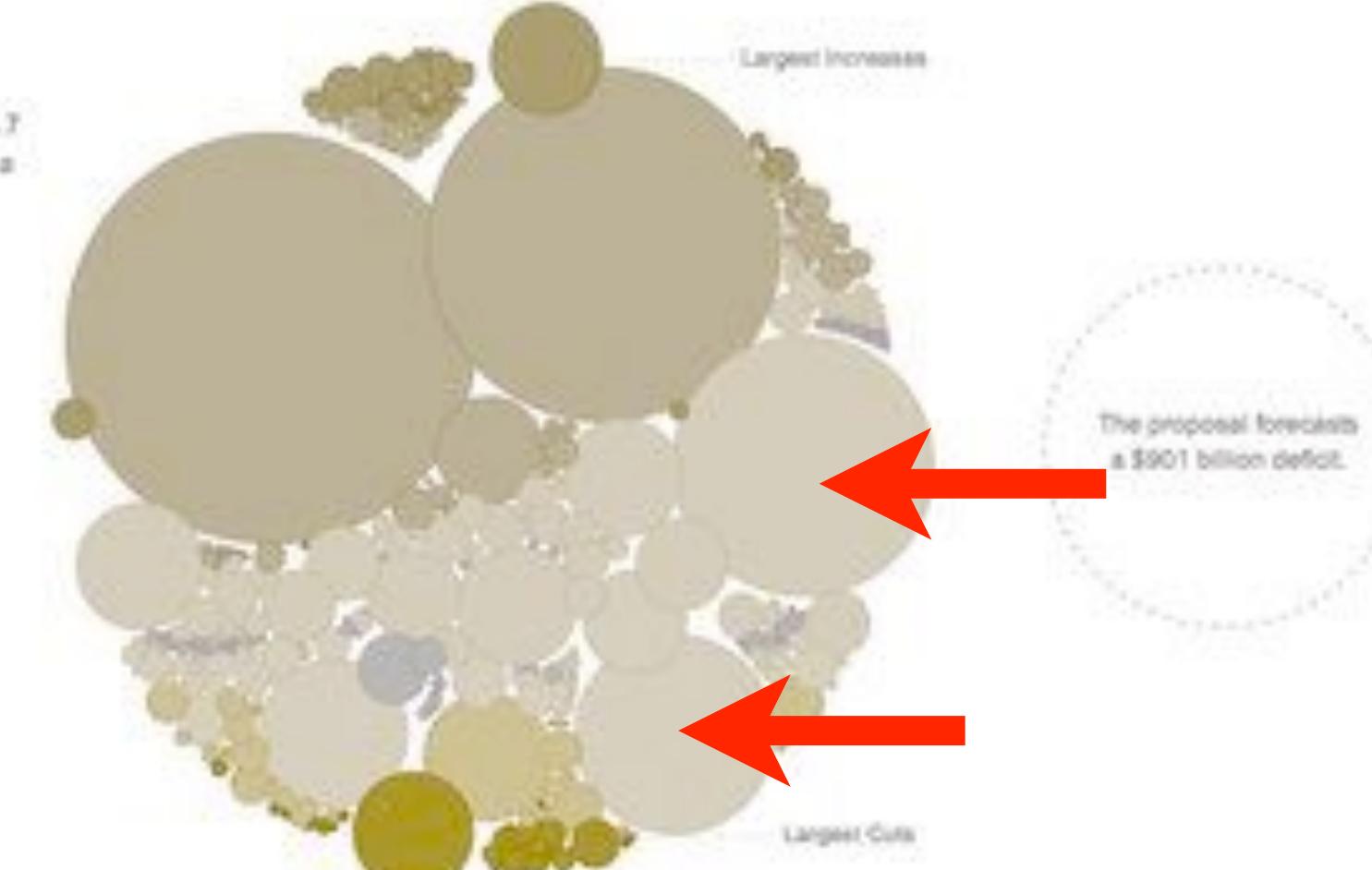
How \$3.7 Trillion Is Spent

Mr. Obama's budget proposal includes \$3.7 trillion in spending in 2013, and forecasts a \$901 billion deficit.

Circles are sized according to the proposed spending.



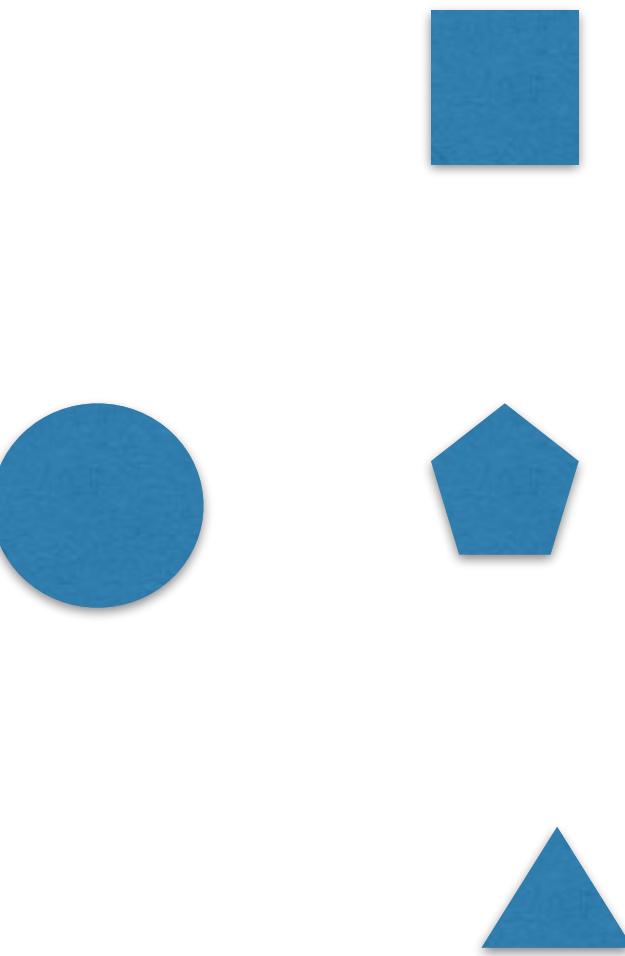
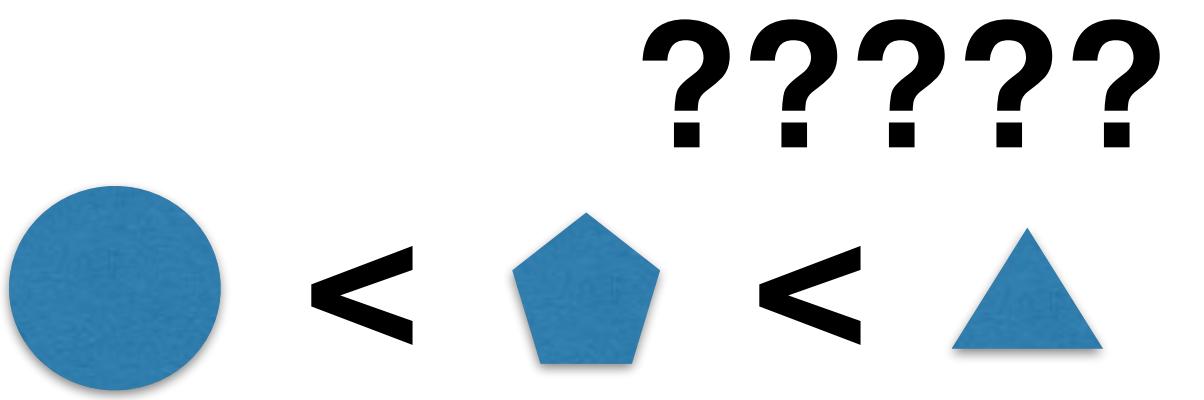
Color shows amount of cut or increase from 2012.



Shape

Great to recognize many classes.

No grouping, ordering.





ASTON MARTIN



Audi

MITSUBISHI
MOTORS

BENTLEY

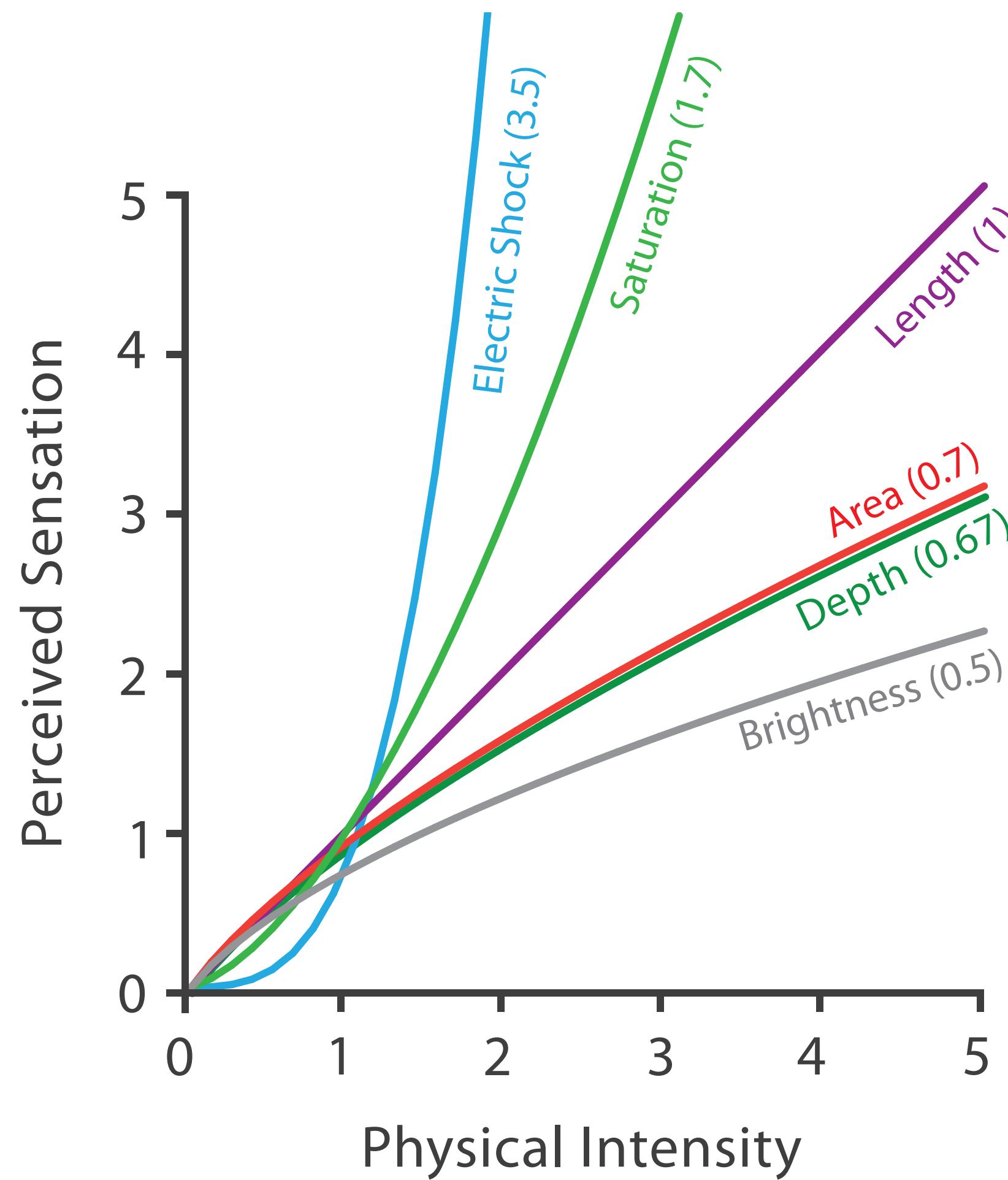


FOR SAFETY You must know German trafficsigns																



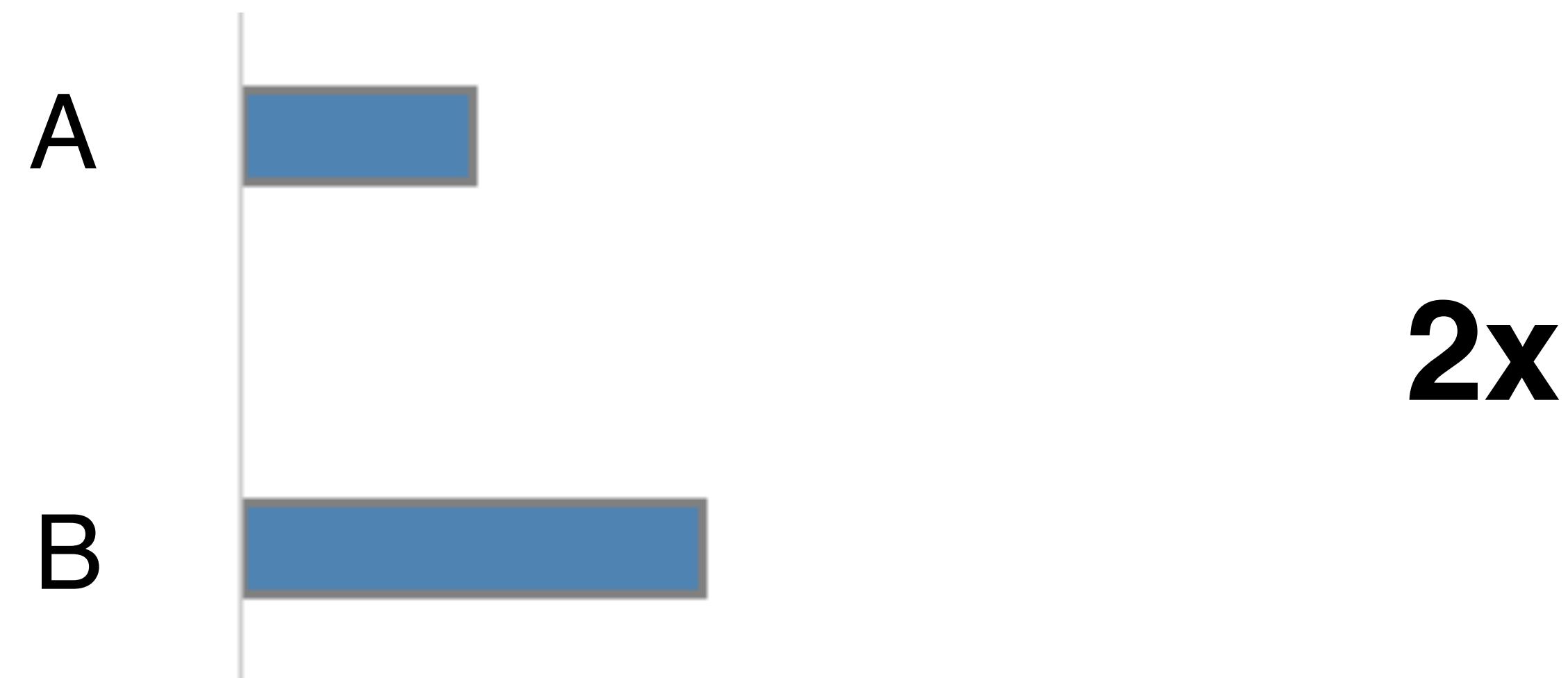
Why are quantitative channels different?

Steven's Psychophysical Power Law: $S = I^N$



S = sensation
 I = intensity

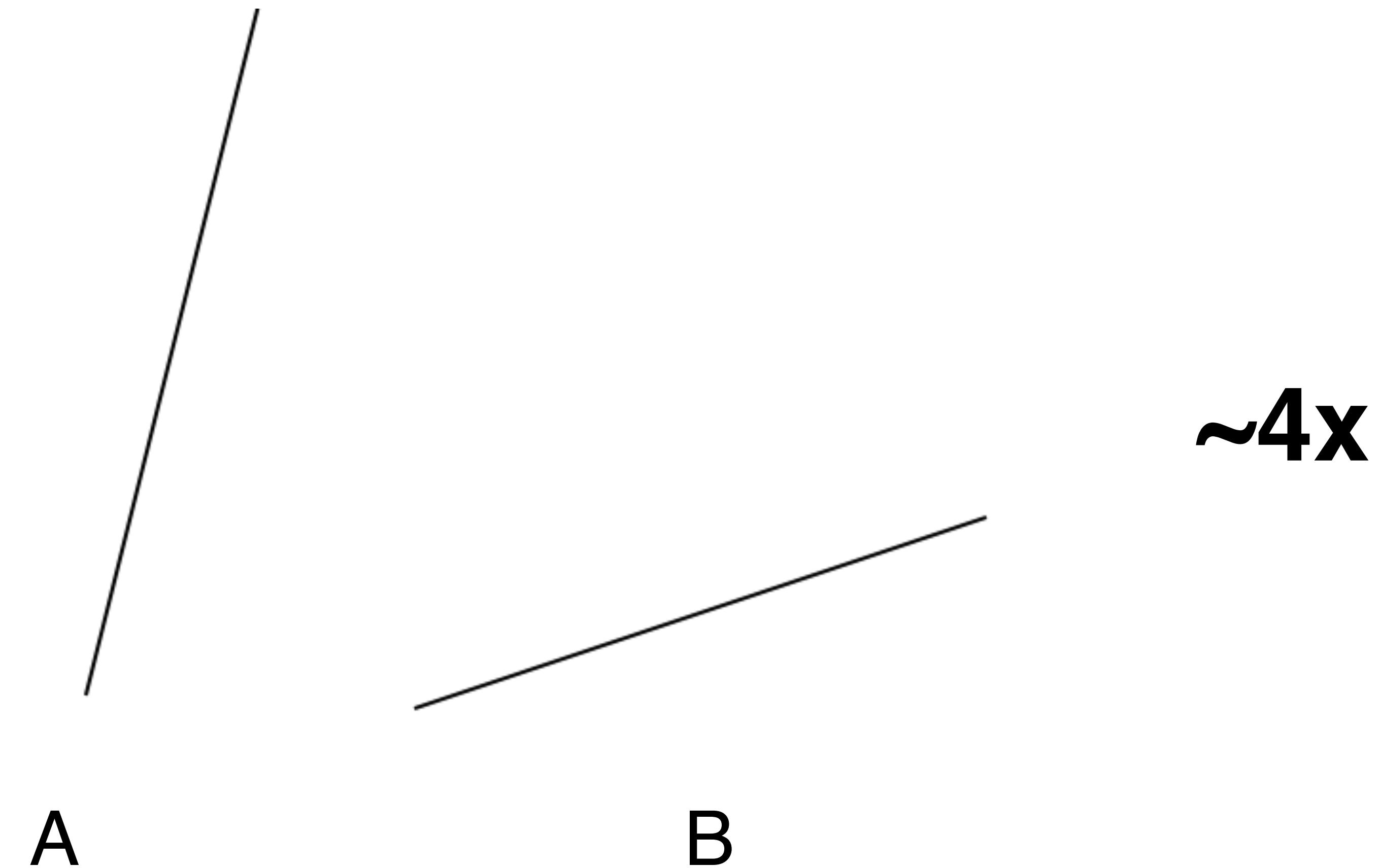
How much longer?



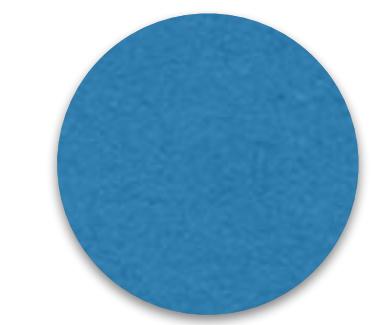
How much longer?



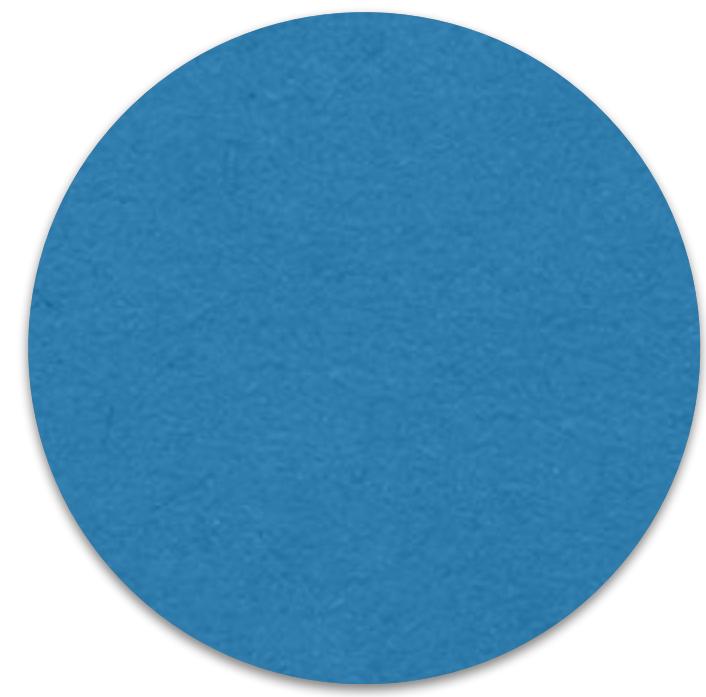
How much steeper?



How much larger?



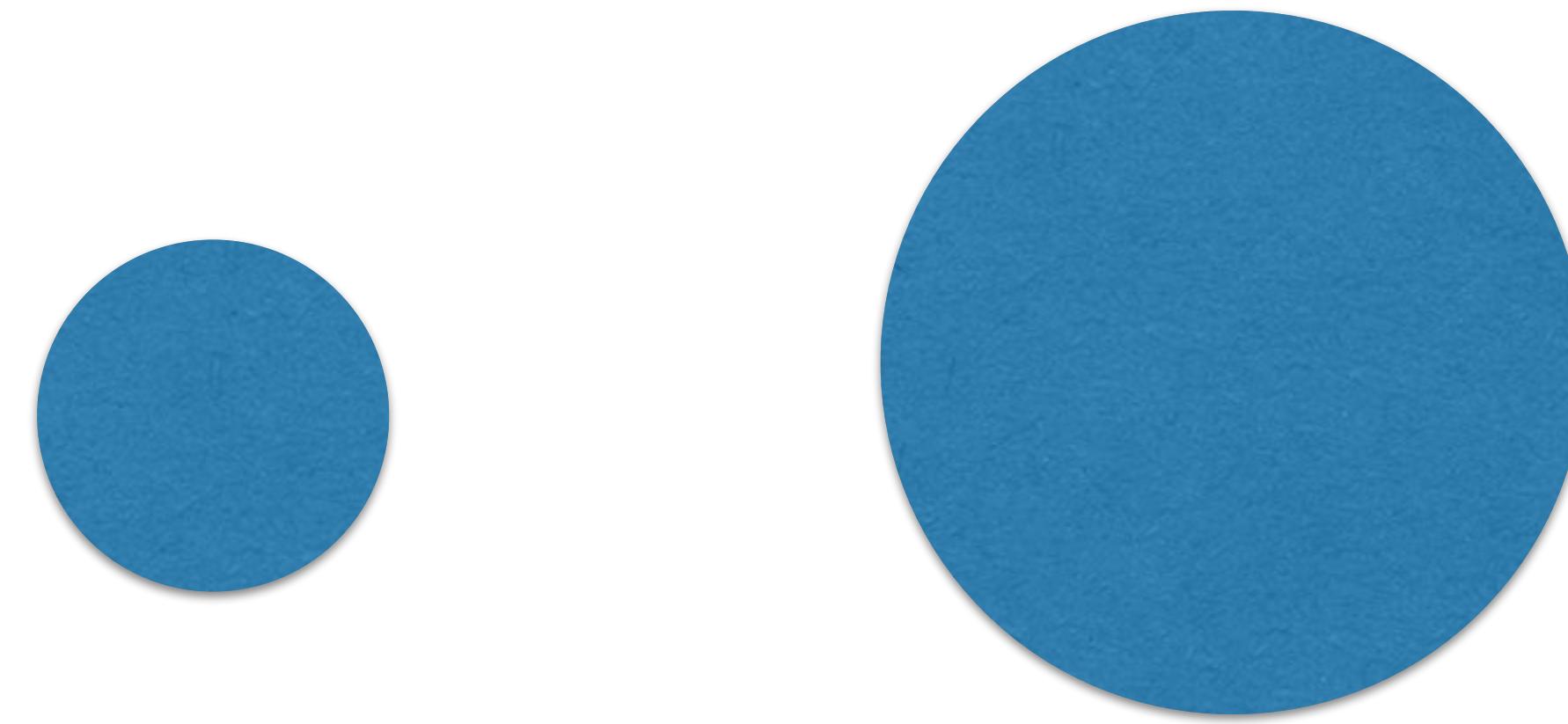
A



B

5x

How much larger?



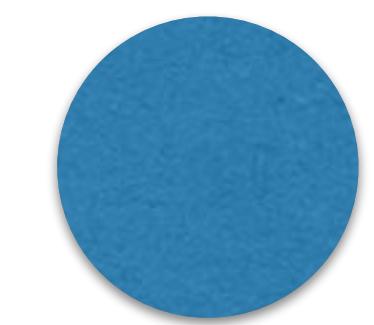
A

B

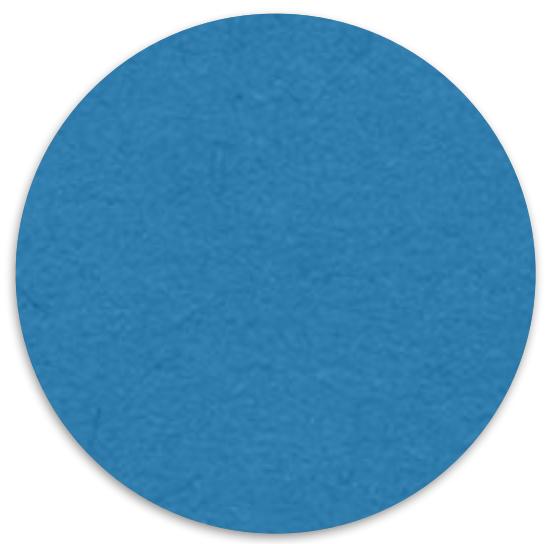
**2x
diameter
4x area**

area is proportional to
diameter squared

How much larger (area)?



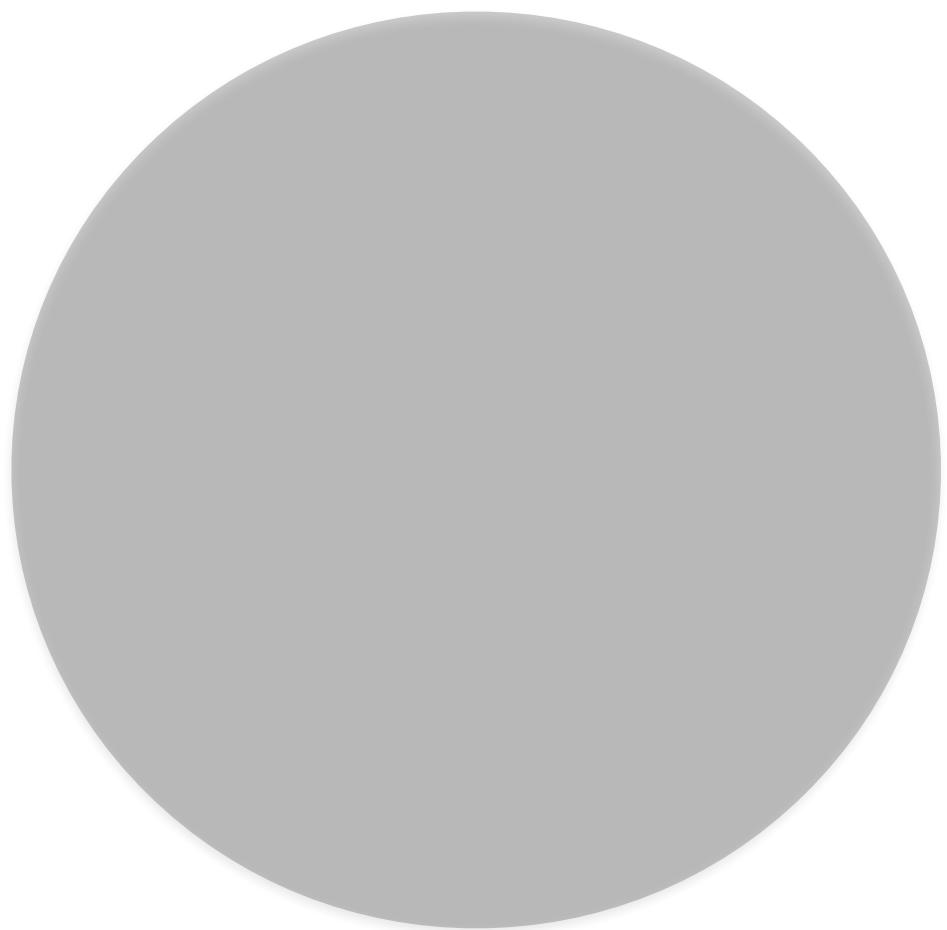
A



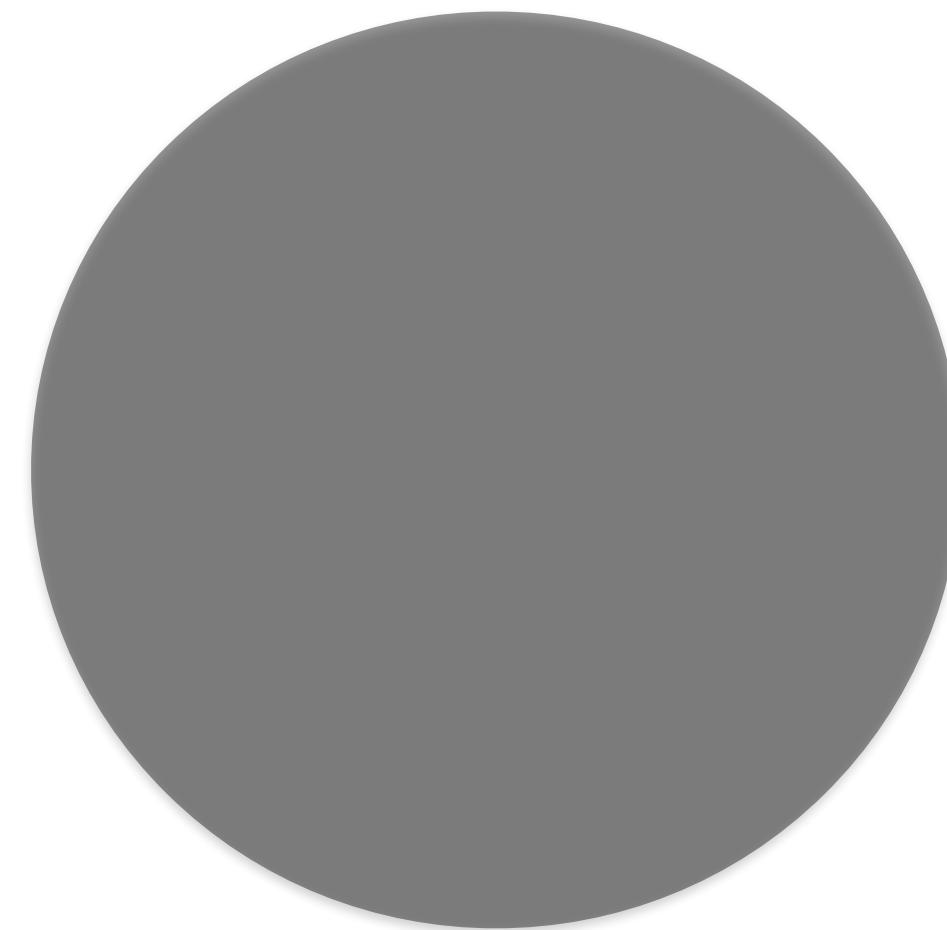
B

3x

How much darker?



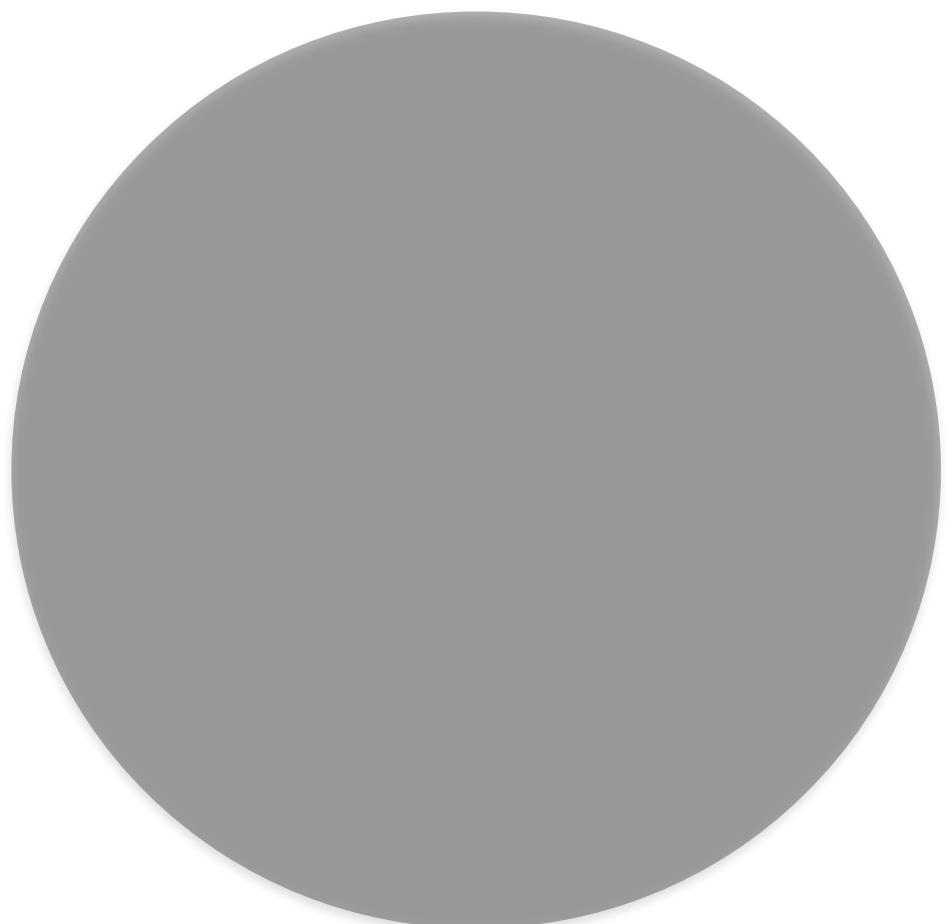
A



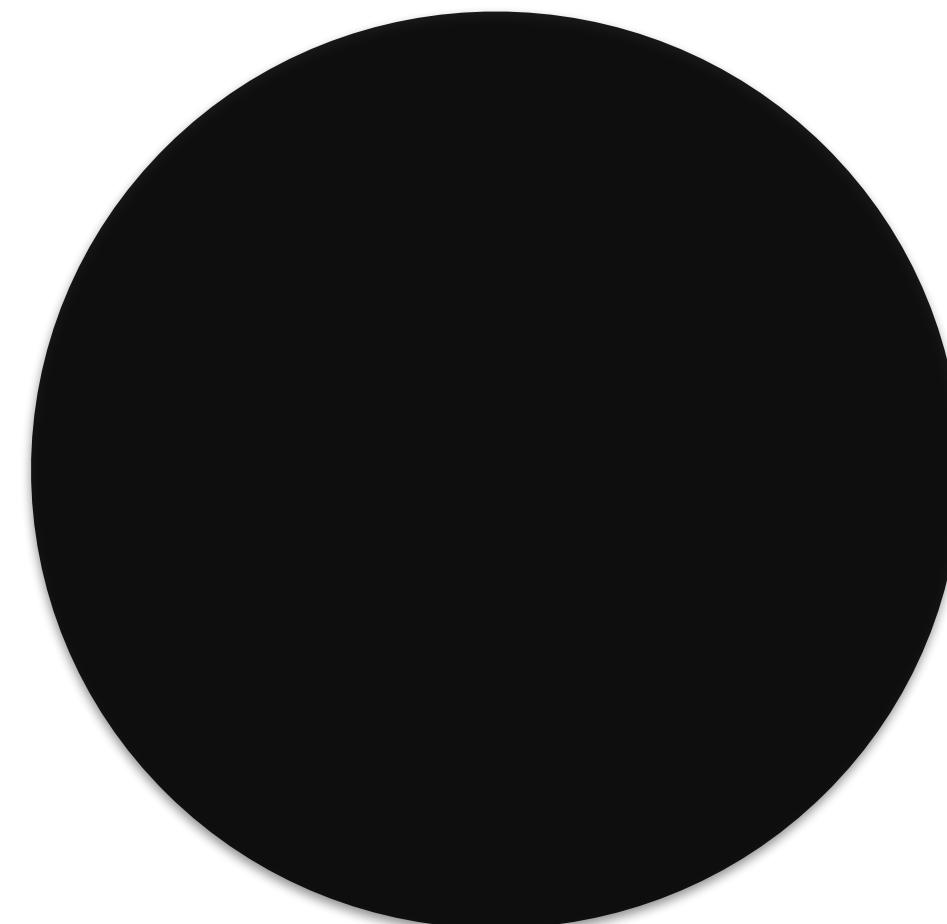
B

2x

How much darker?



A

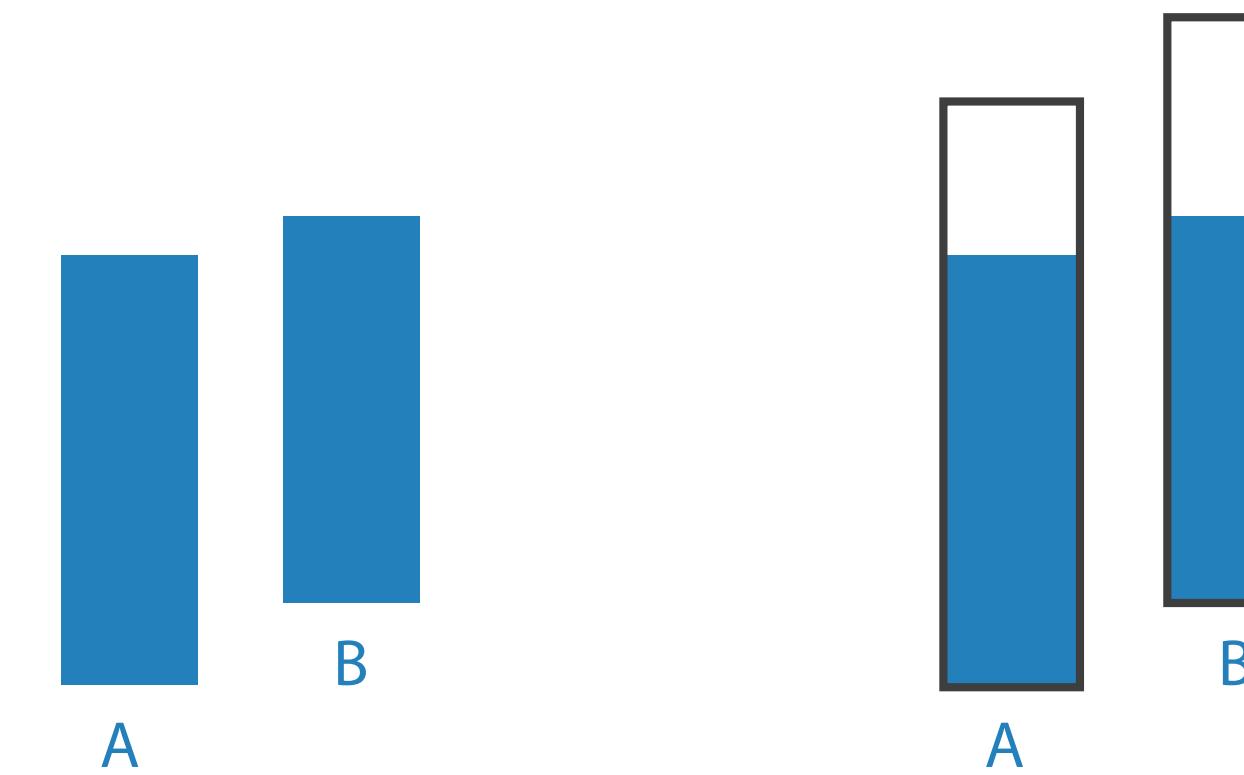


B

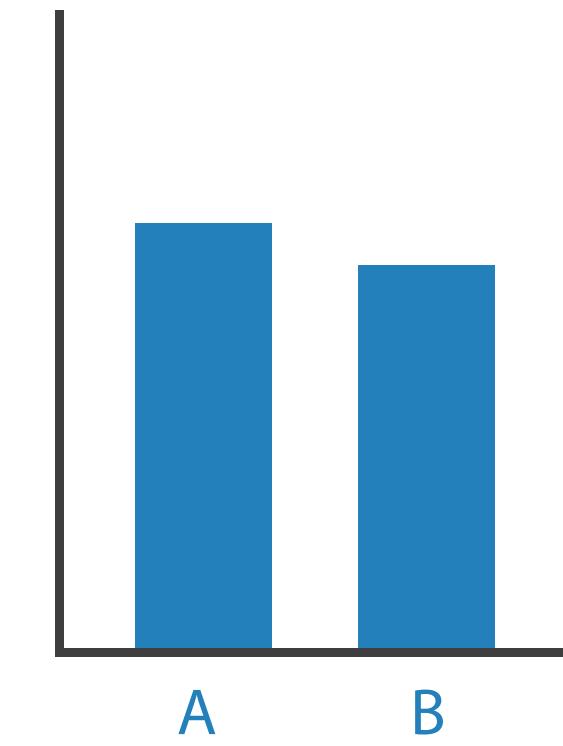
3x

Other Factors Affecting Accuracy

Alignment



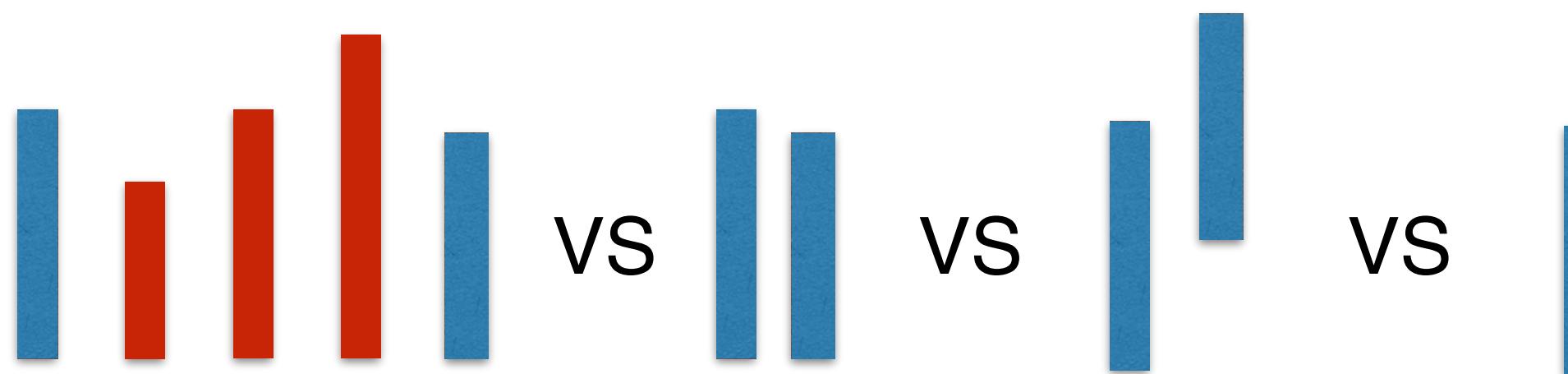
Distractors



Distance

Common scale

...



Channels: Expressiveness Types and Effectiveness Ranks

→ Magnitude Channels: Ordered Attributes

Position on common scale



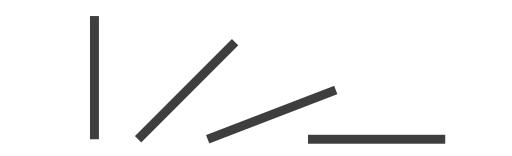
Position on unaligned scale



Length (1D size)



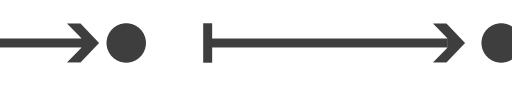
Tilt angle



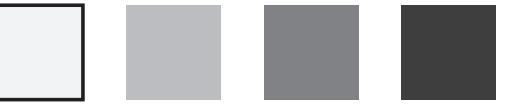
Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



Effectiveness ↑
Same
Least ↓

→ Identity Channels: Categorical Attributes

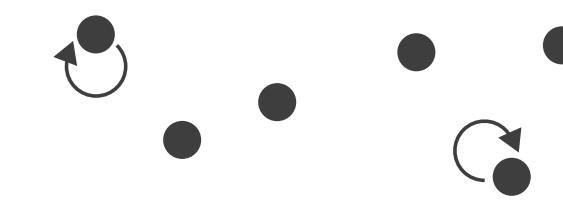
Spatial region



Color hue



Motion



Shape



Design Guidelines

Rule #1: Use the Best Visual
Channel Available
for the Most Important
Aspect of your Data

Effectiveness Principle

Rule #2: The visualization
should show all of the data,
and only the data

Expressiveness Principle

Tufte's Integrity Principles

Show **data variation**, not design variation

Clear, detailed, and thorough **labeling and appropriate scales**

Size of the **graphic effect** should be **directly proportional to the numerical quantities** (“lie factor”)

The Lie Factor

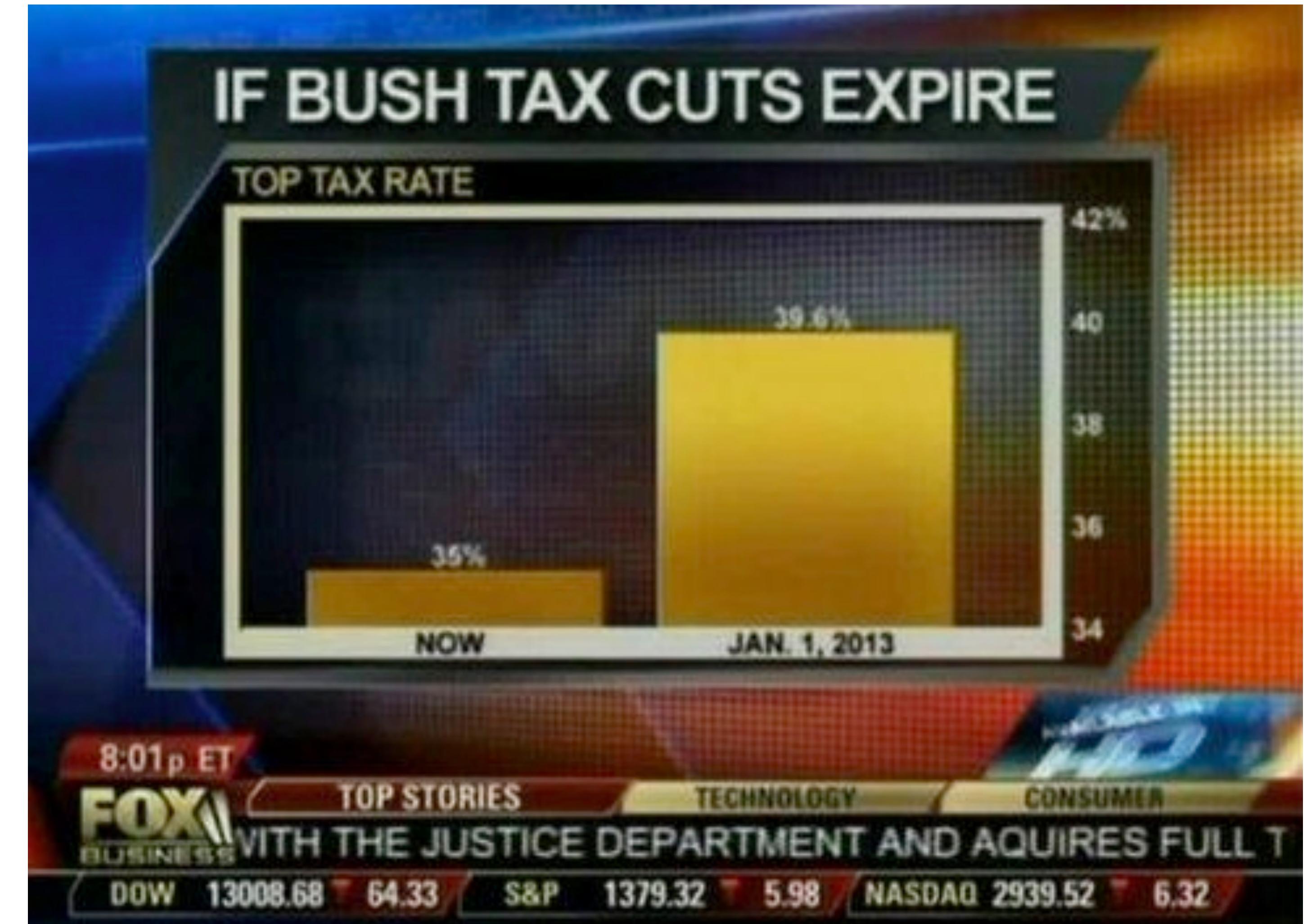
Size of effect shown in graphic

Size of effect in data

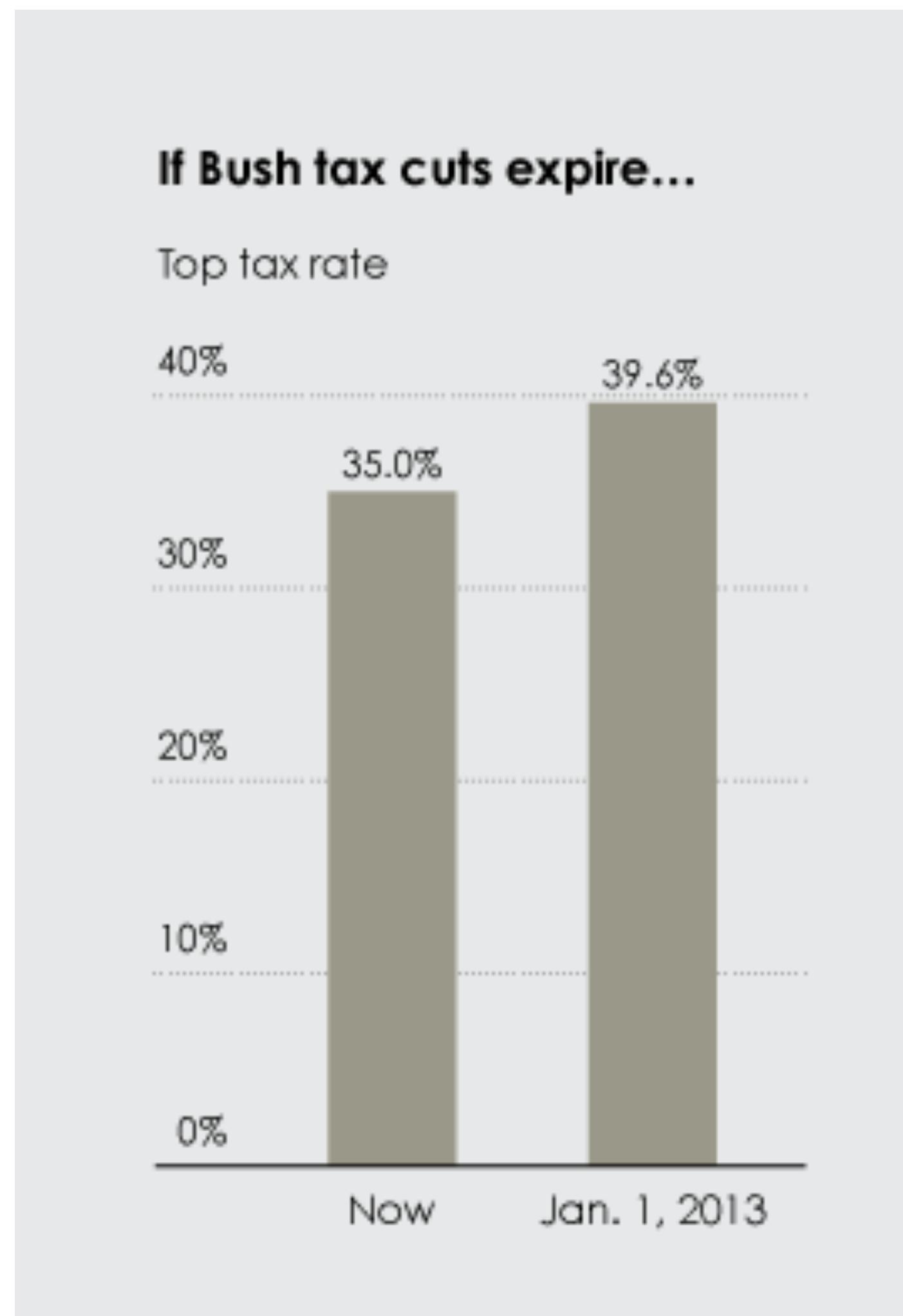
Lie Factor - Graphical Integrity

Magnitude in data
must correspond to
magnitude of mark

Effect in Data: factor 1.14
Effect in Graphic: factor 5
Lie Factor: $5/1.14 = 4.38$



Scale Distortions



What's wrong?



Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

What's wrong?



Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um



Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

What's wrong?

Grafik der Kronenzeitung



Viele Bezieher mit "ungeklärter Staatsbürgerschaft"
Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

Grafik
in echt



Viele Bezieher mit "ungeklärter Staatsbürgerschaft"
Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

OBAMACARE ENROLLMENT

7,100,000

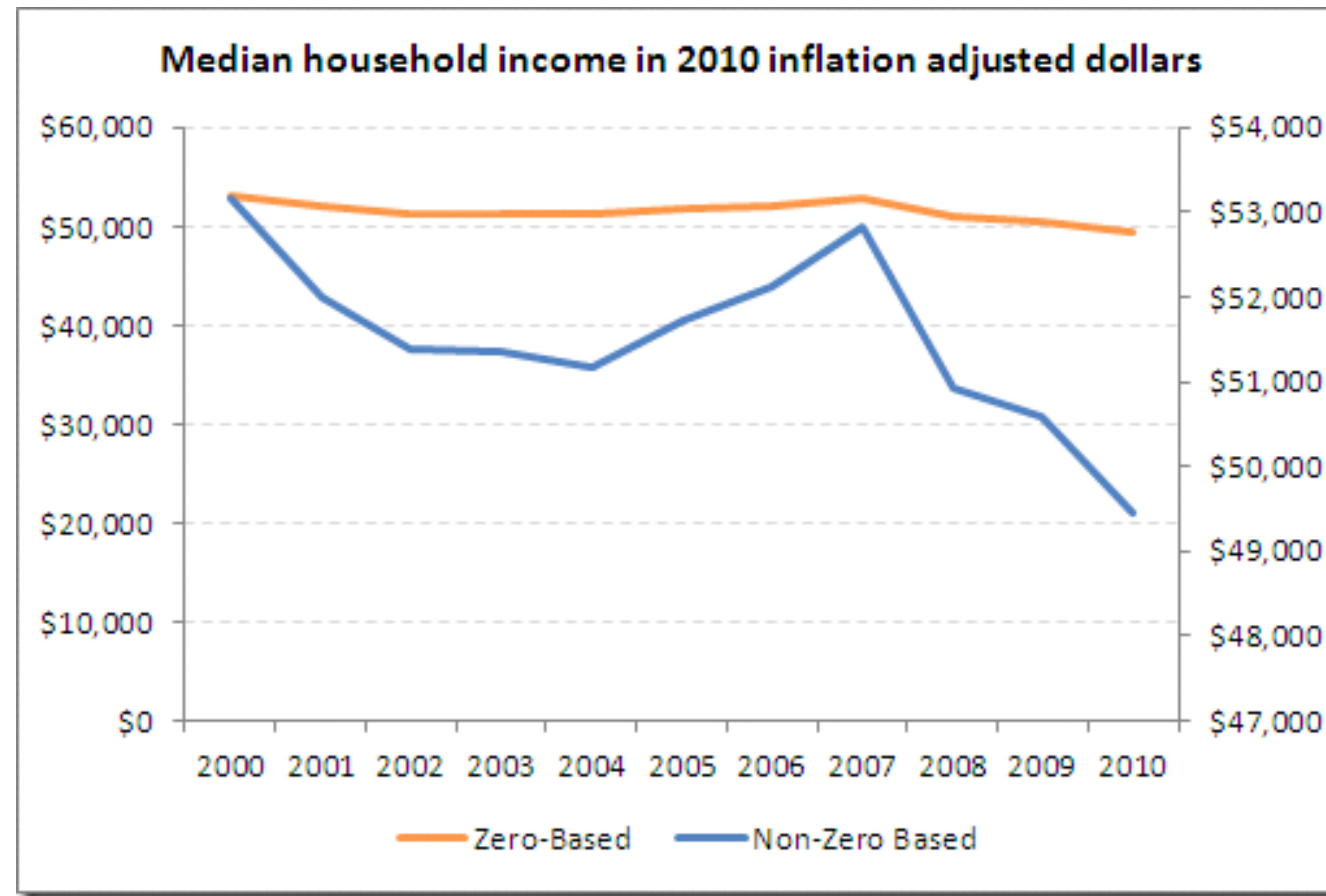
ACTUAL
ENROLLMENT

7,000,000

GOAL



Start Scales at 0?

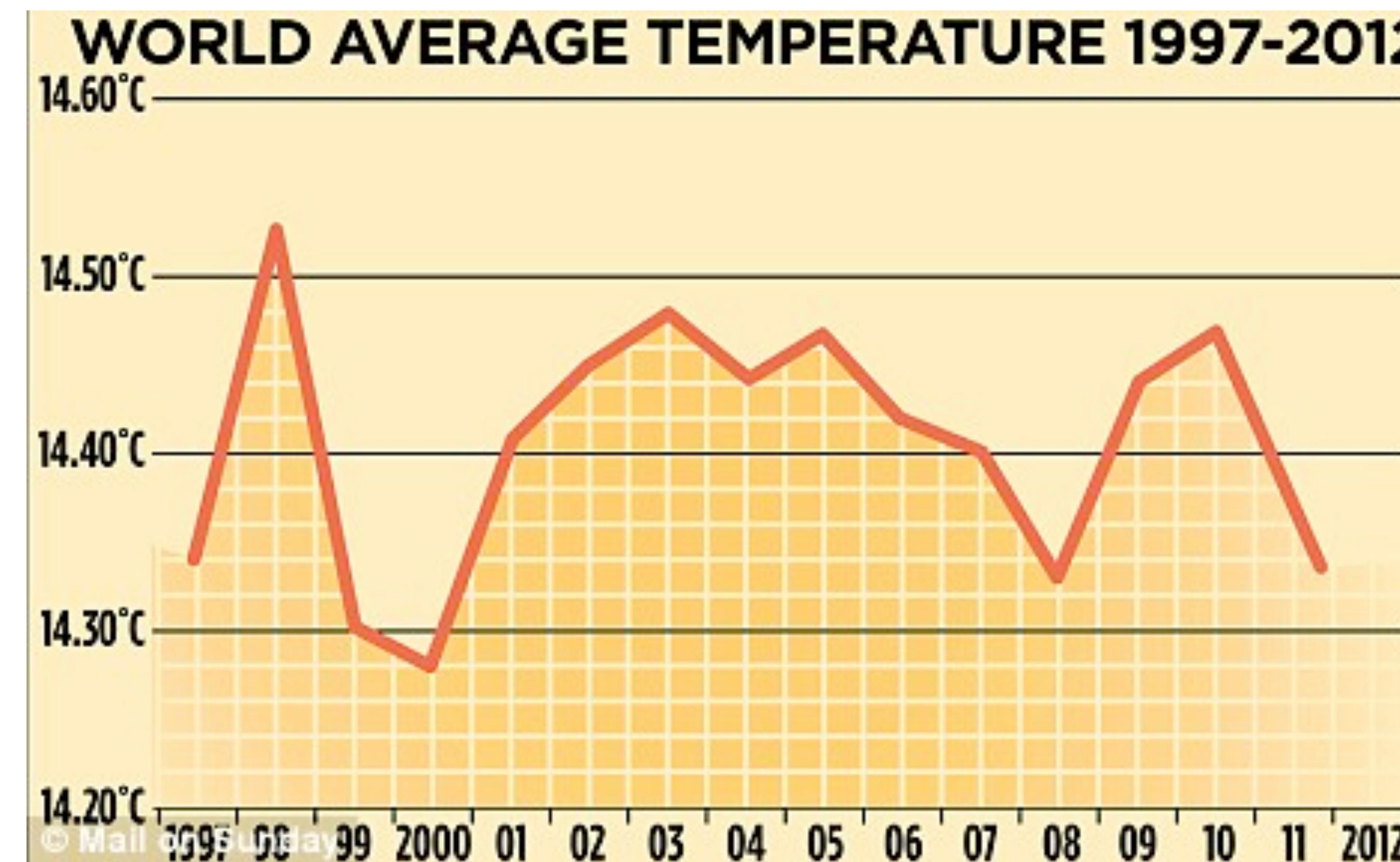


Scales at 0

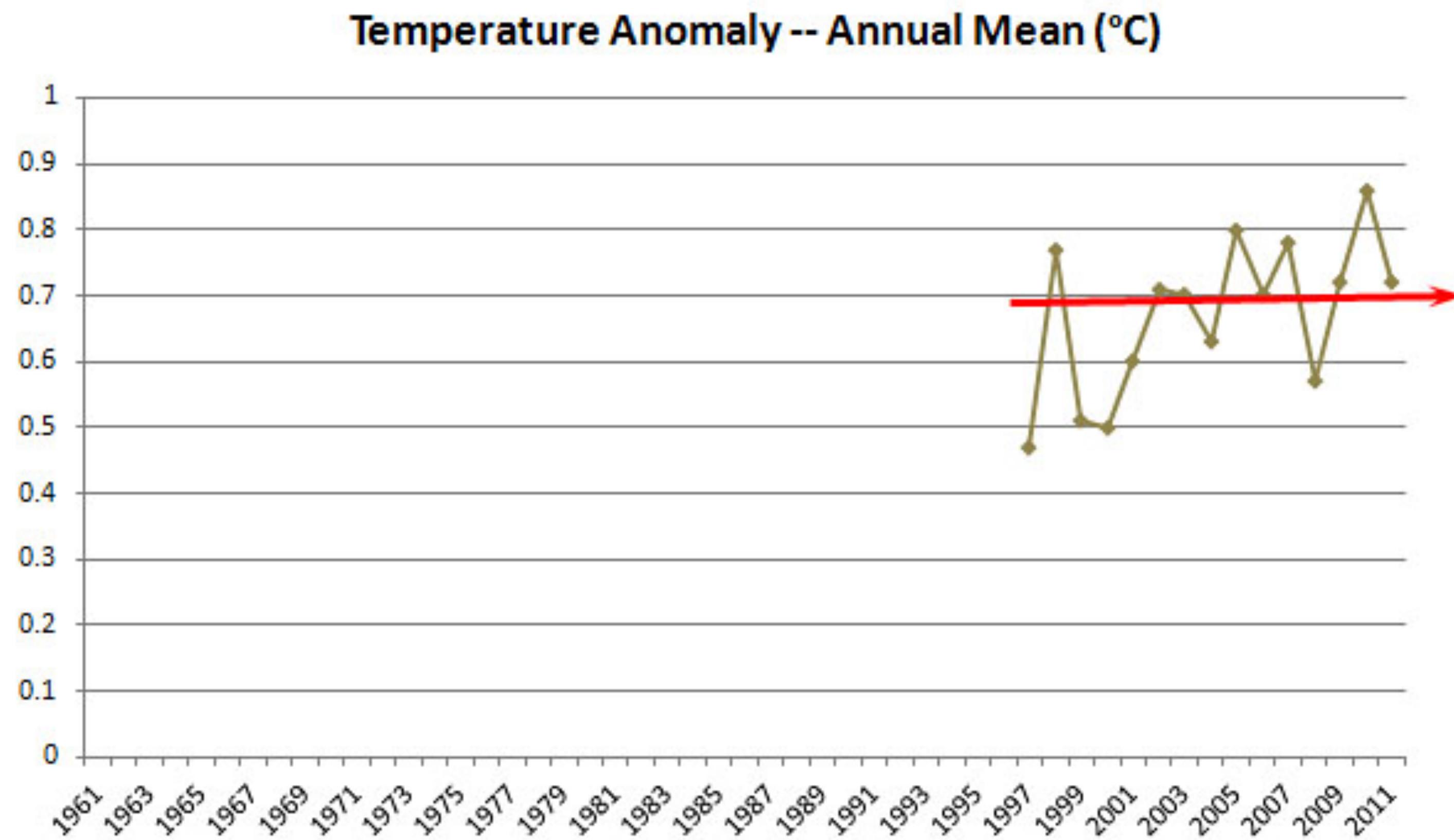


**Use a baseline that shows the
data, not the zero-point.**

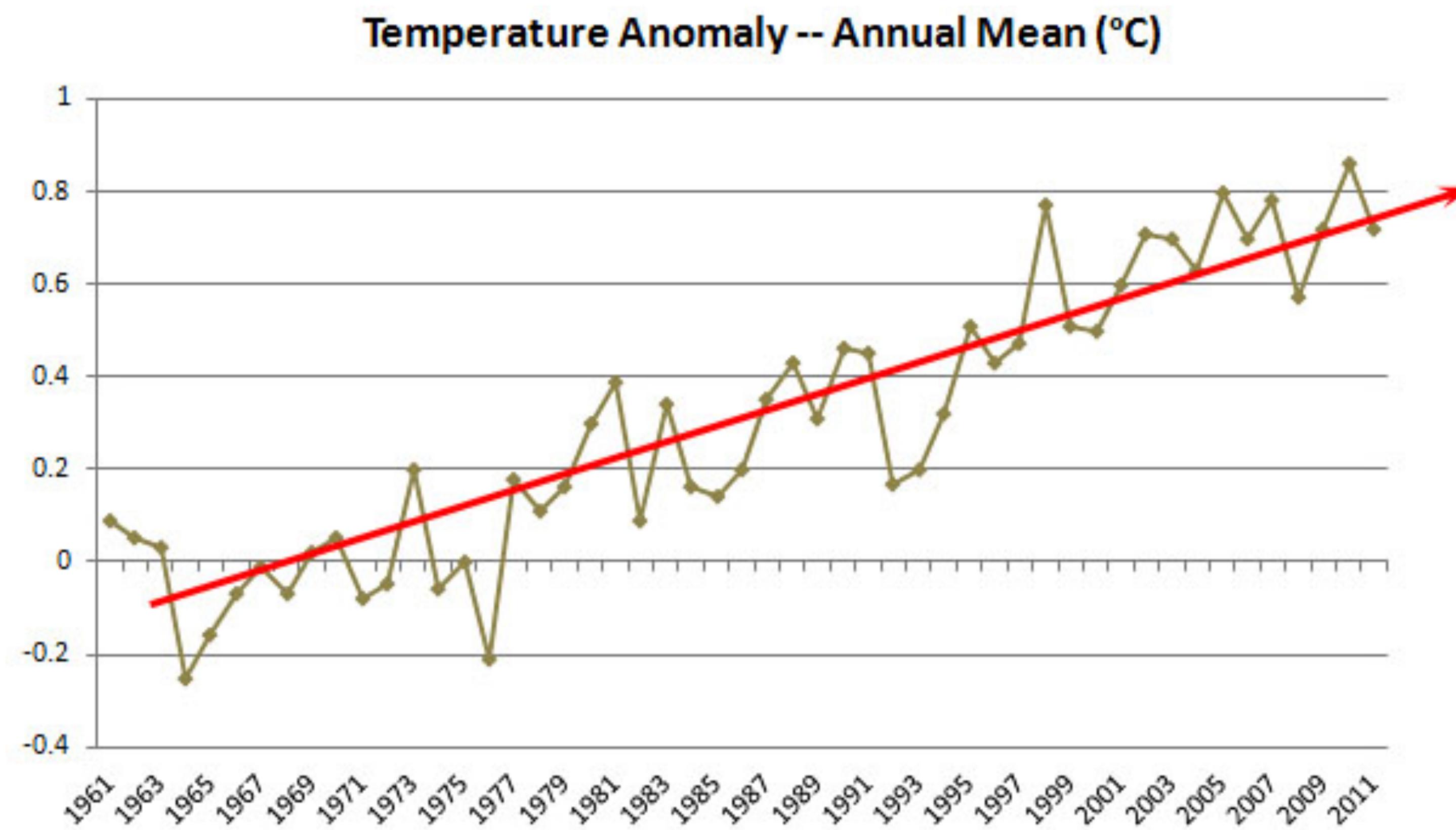
Global Warming?



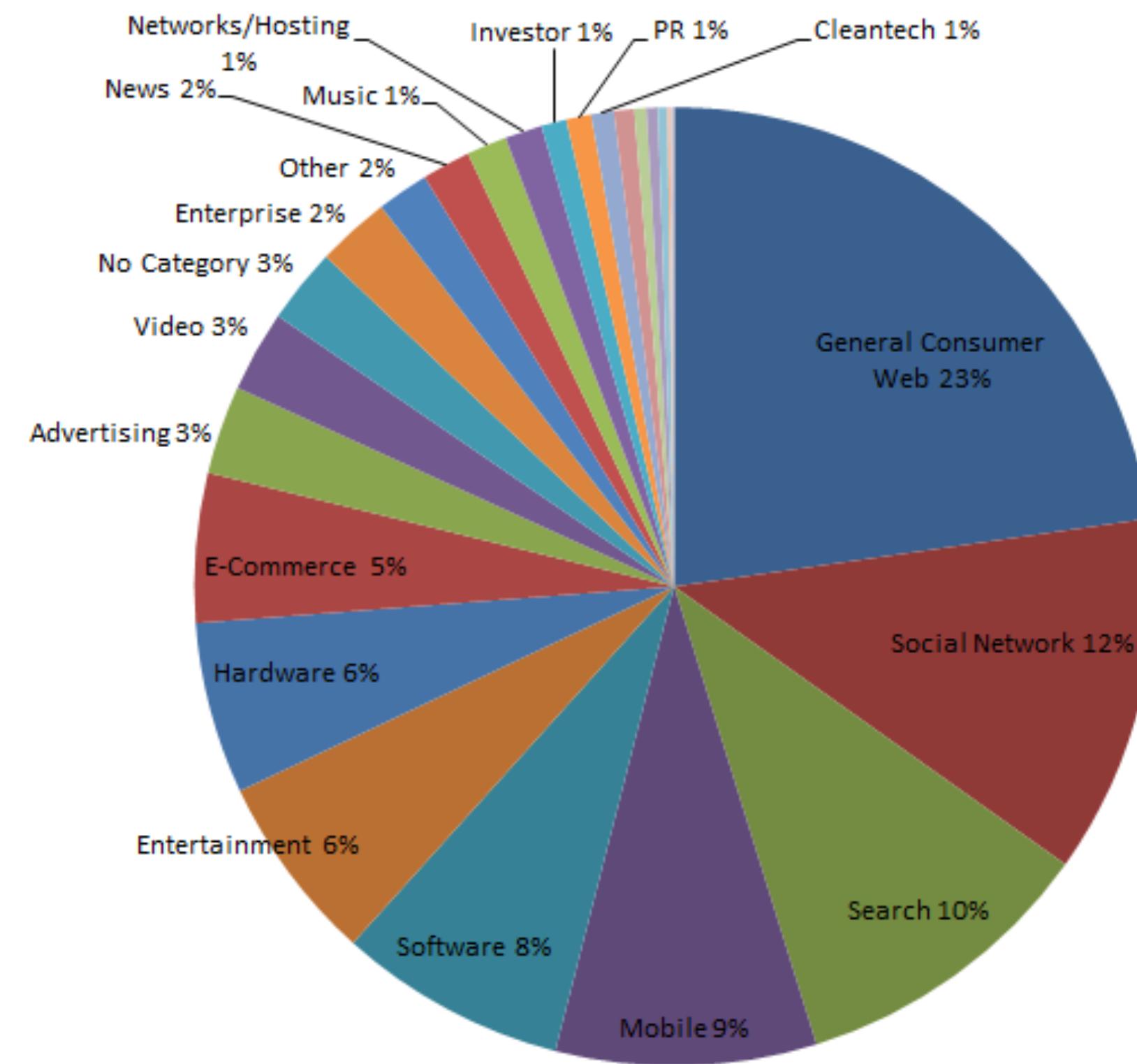
Global Warming?



Global Warming - Frame the Data



Death to Pie Charts

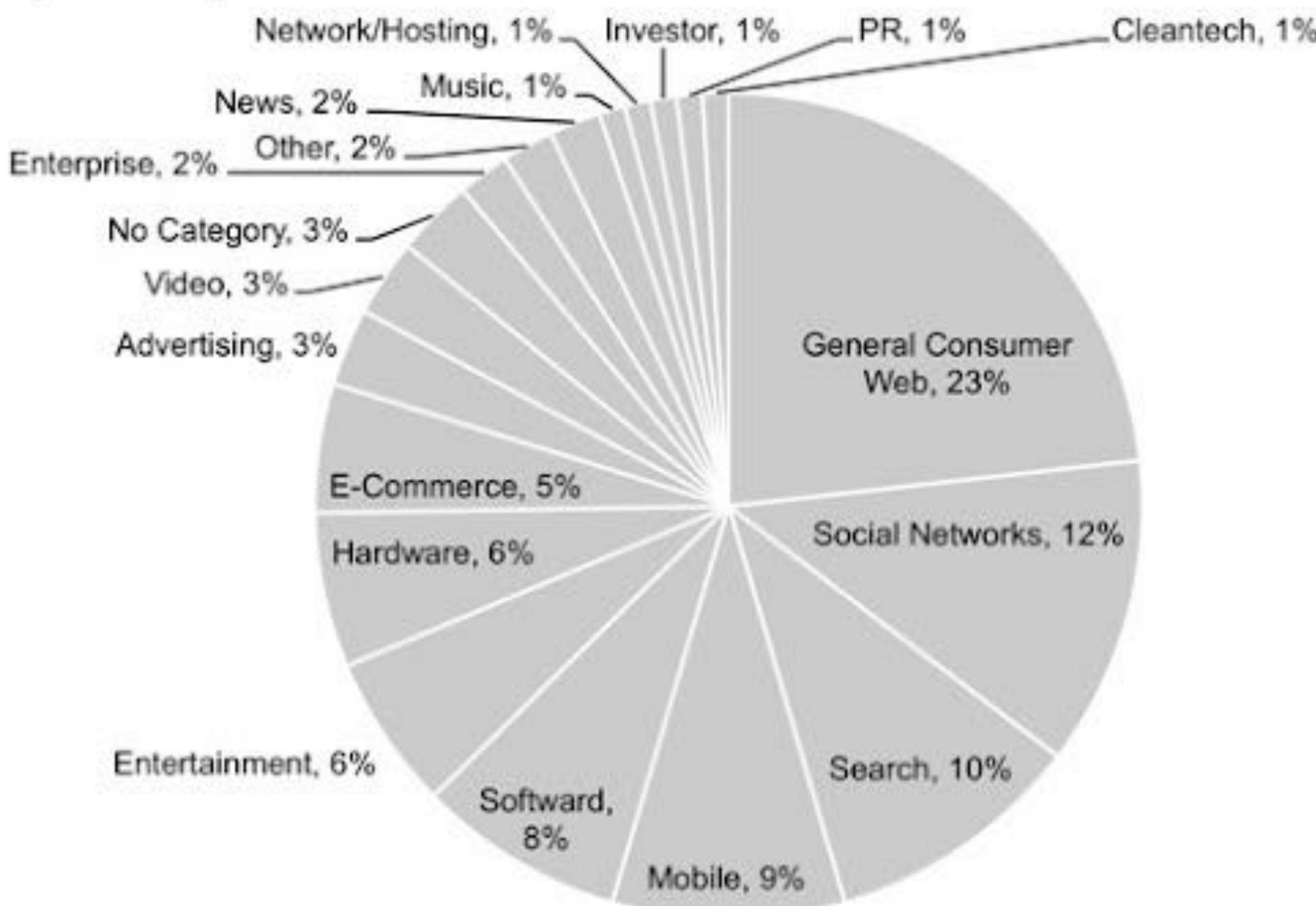


Share of coverage
on TechCrunch

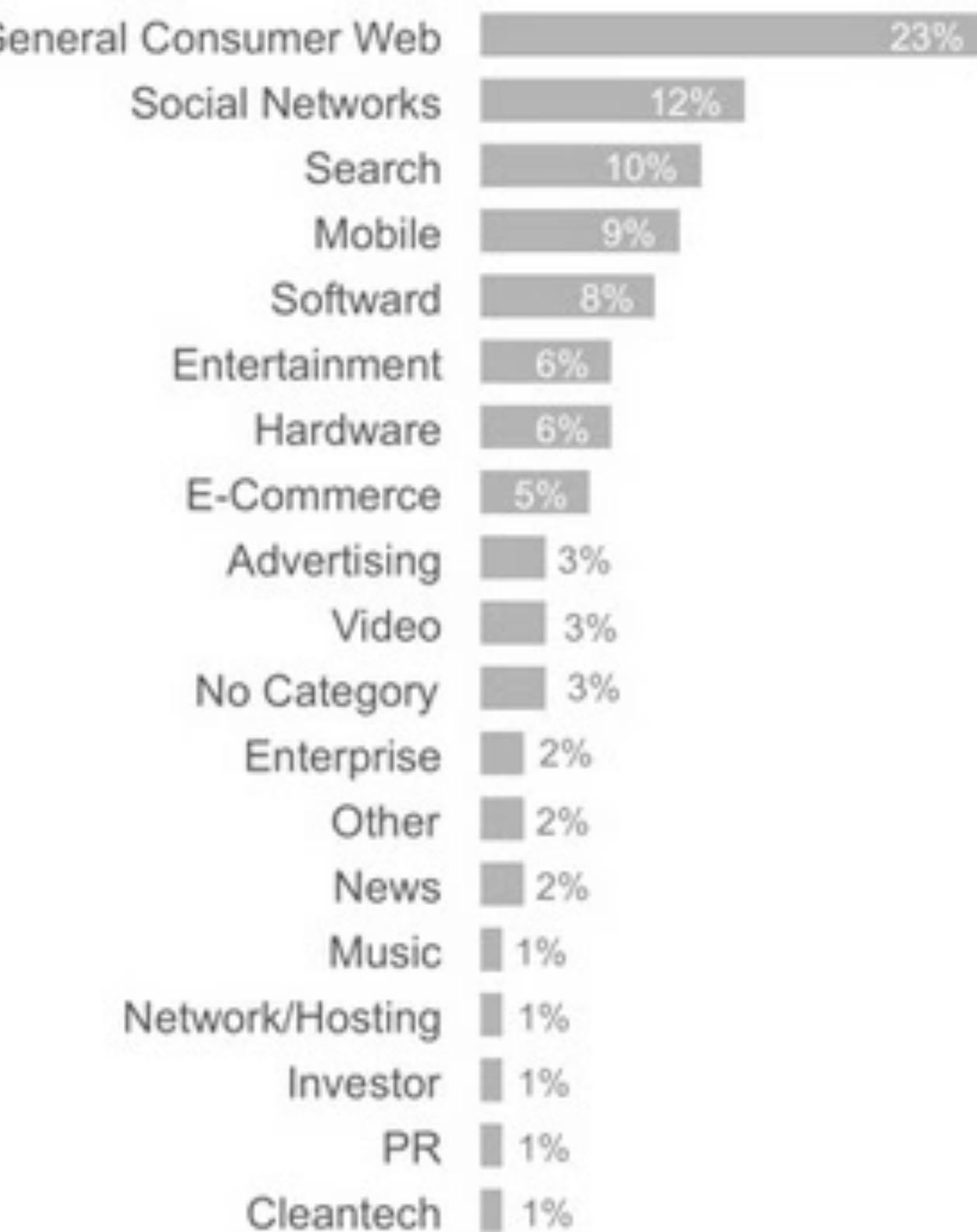
“I hate pie charts.
I mean, really hate them.”

Redesign

TechCrunch Coverage: 2005 - 2011
A slightly better pie?

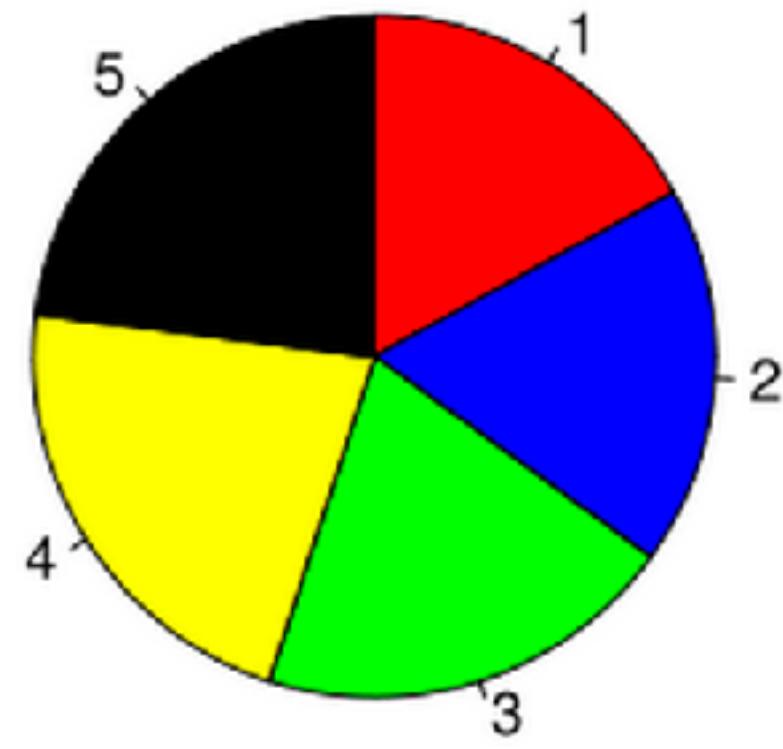


TechCrunch Coverage: 2005 - 2011
Bars are best!

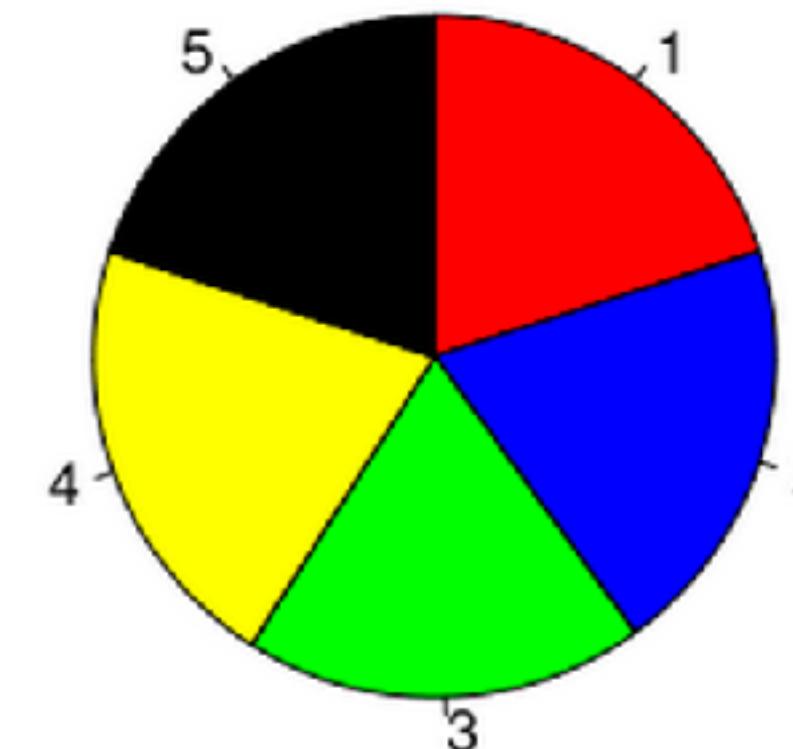


Can you spot the differences?

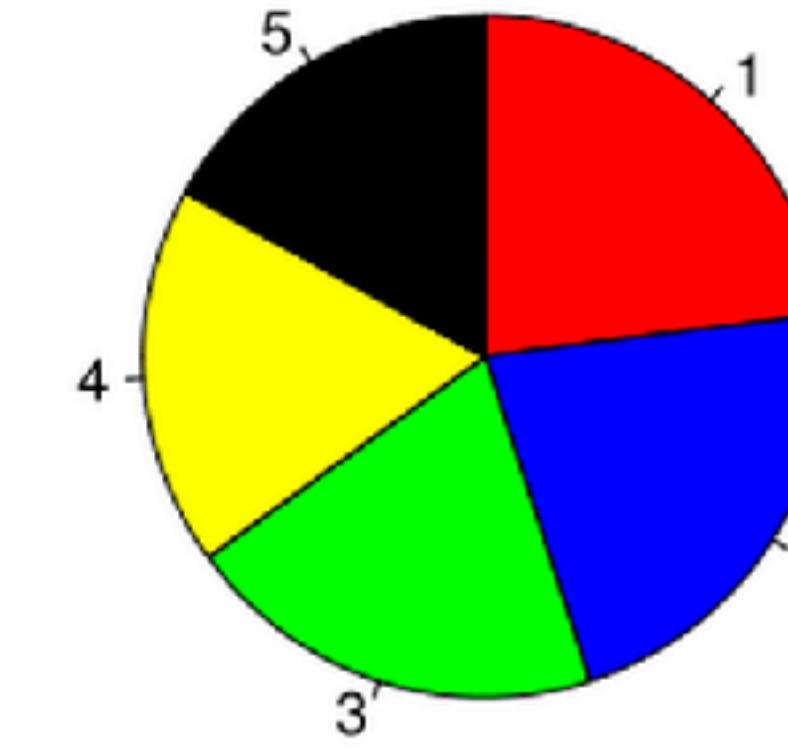
A



B

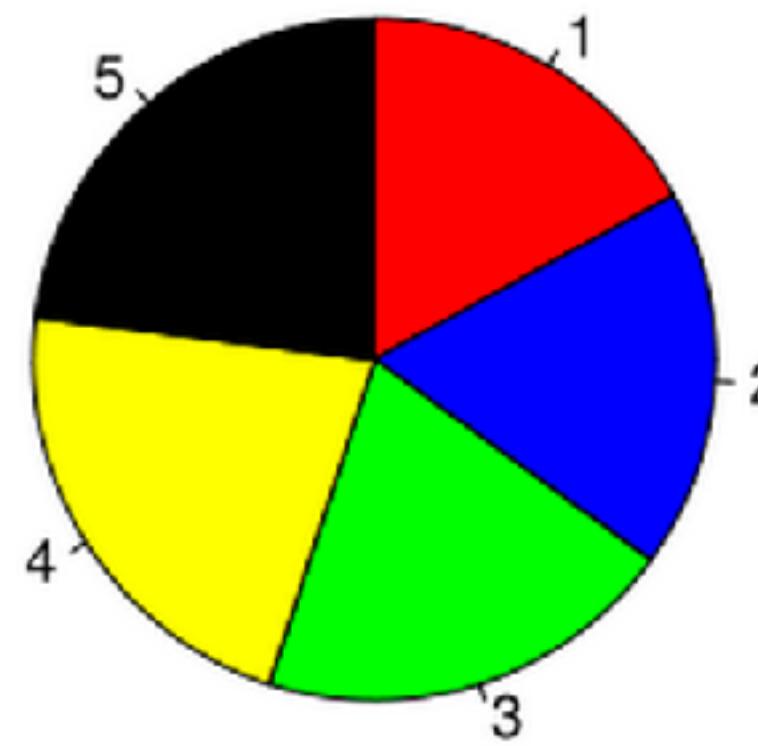


C

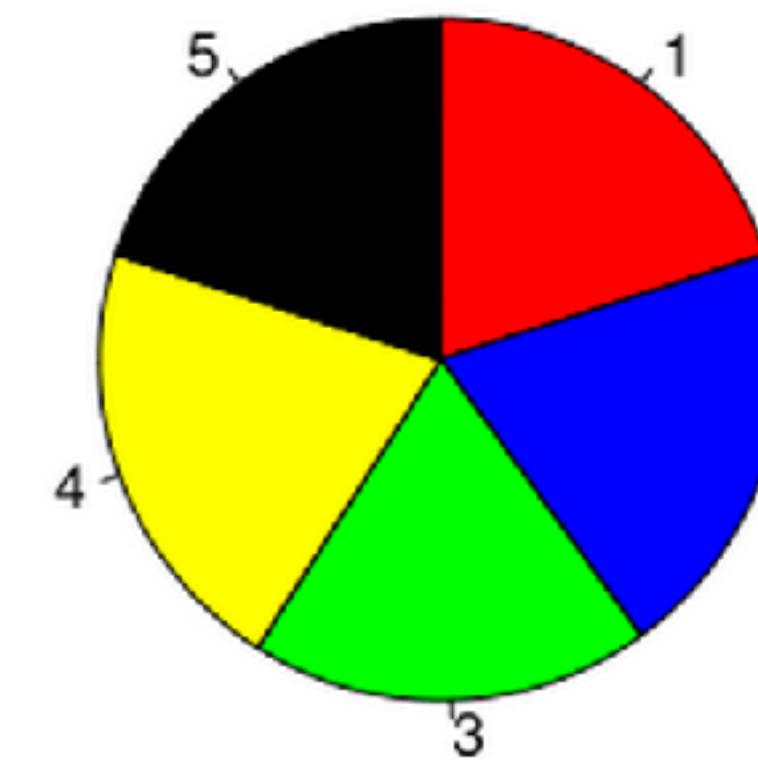


Can you spot the differences?

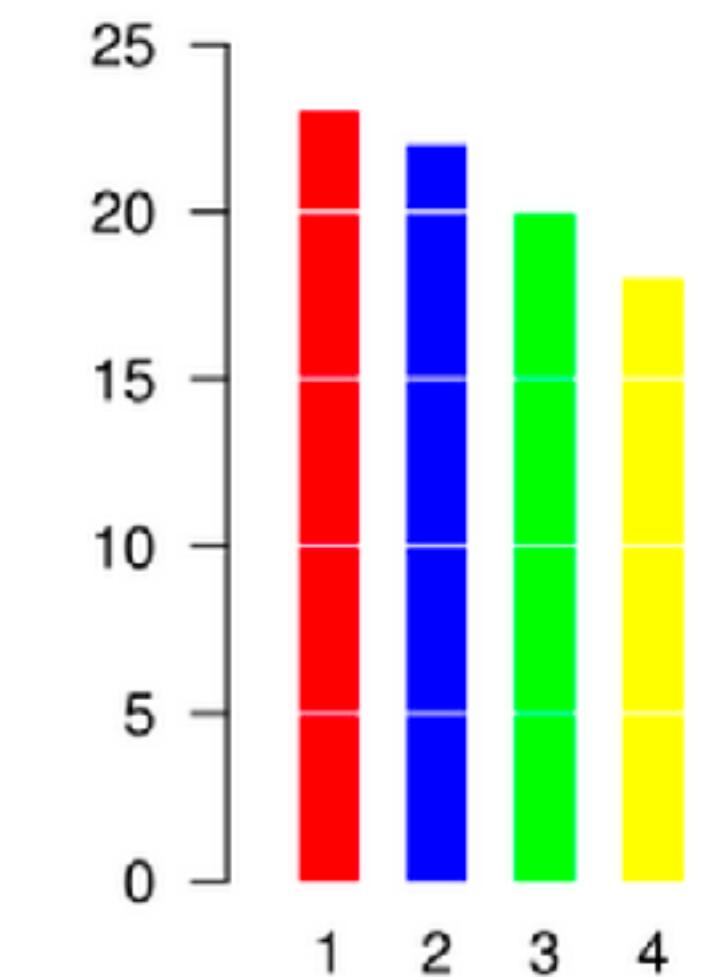
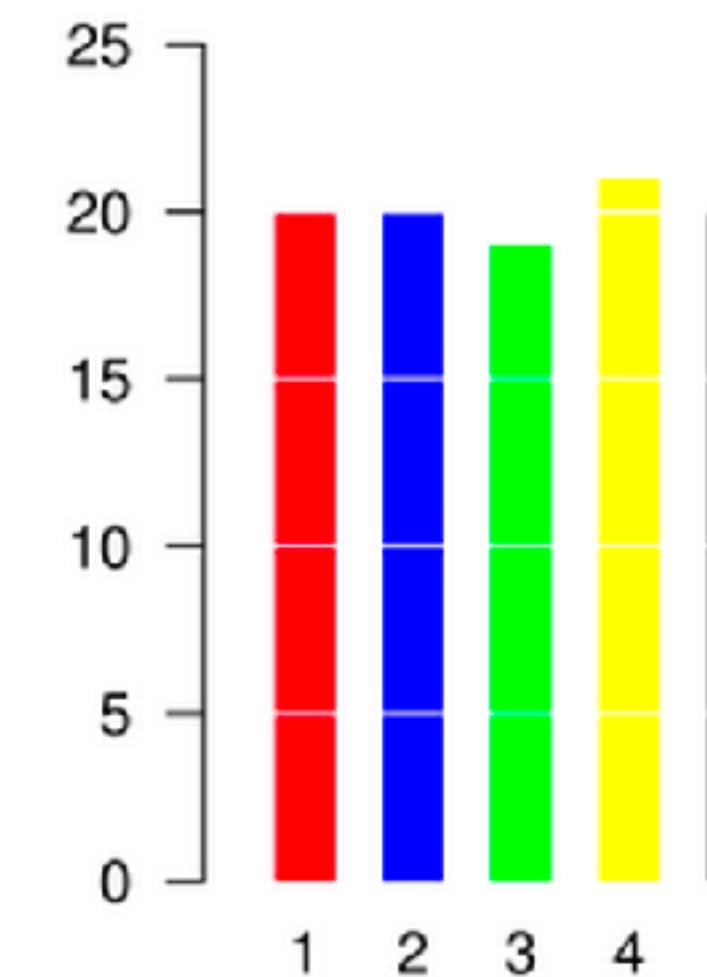
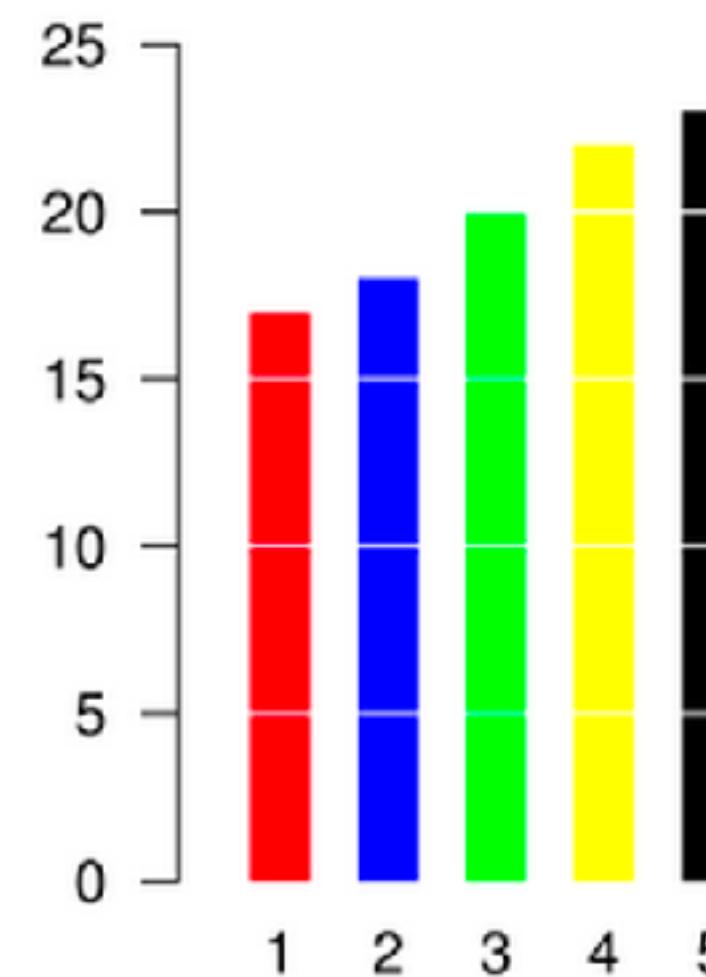
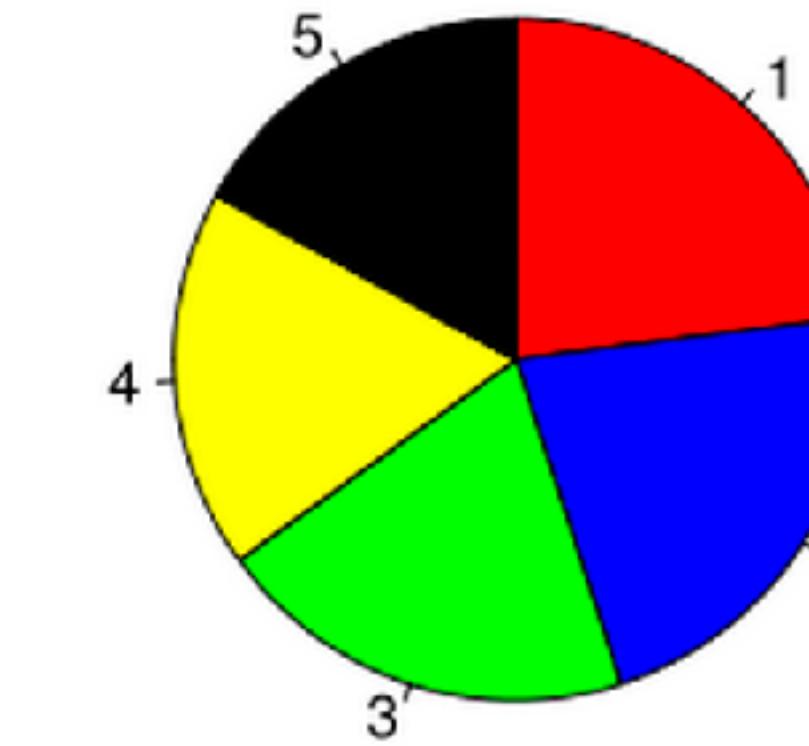
A



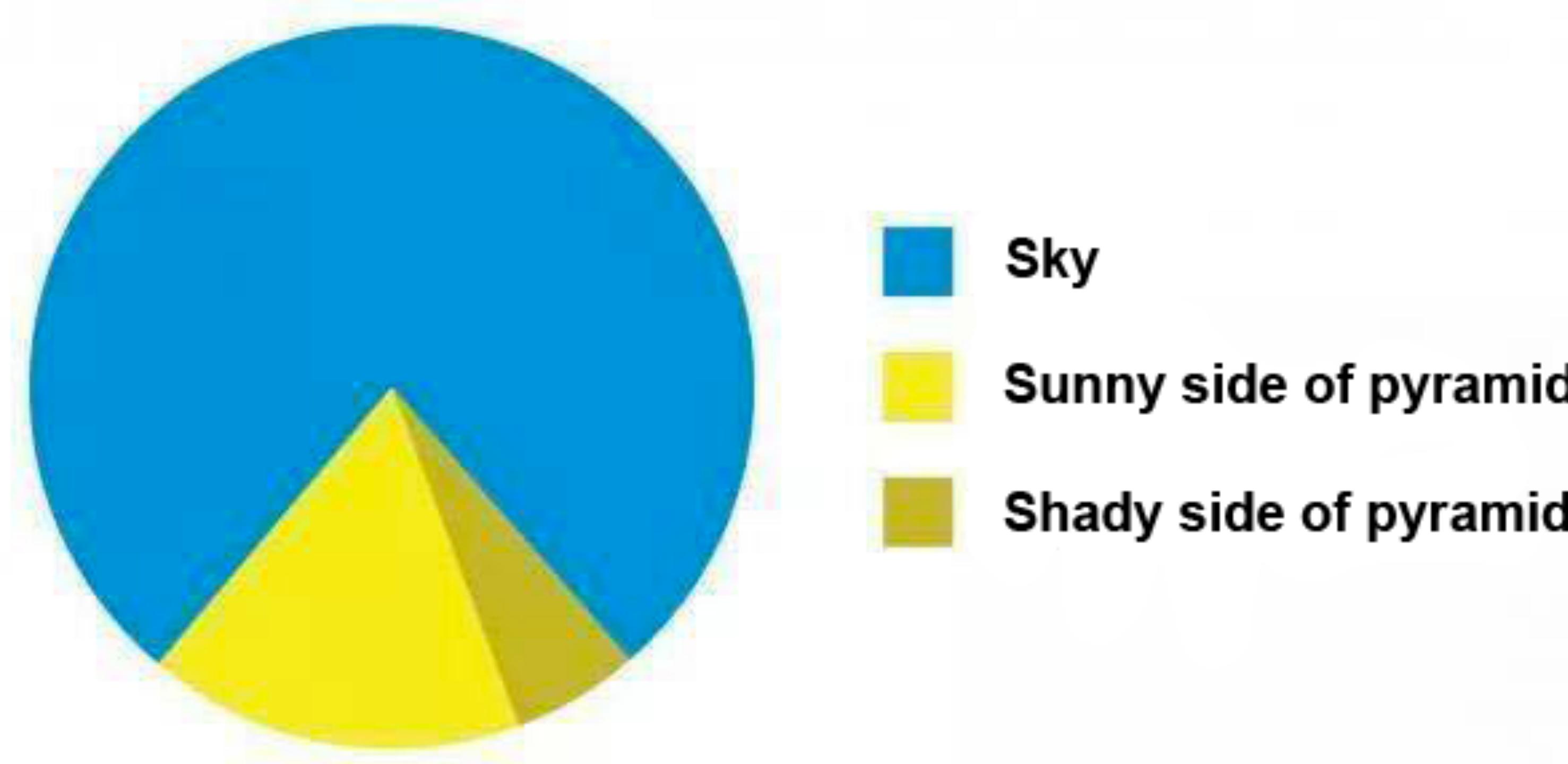
B



C



My favorite pie chart



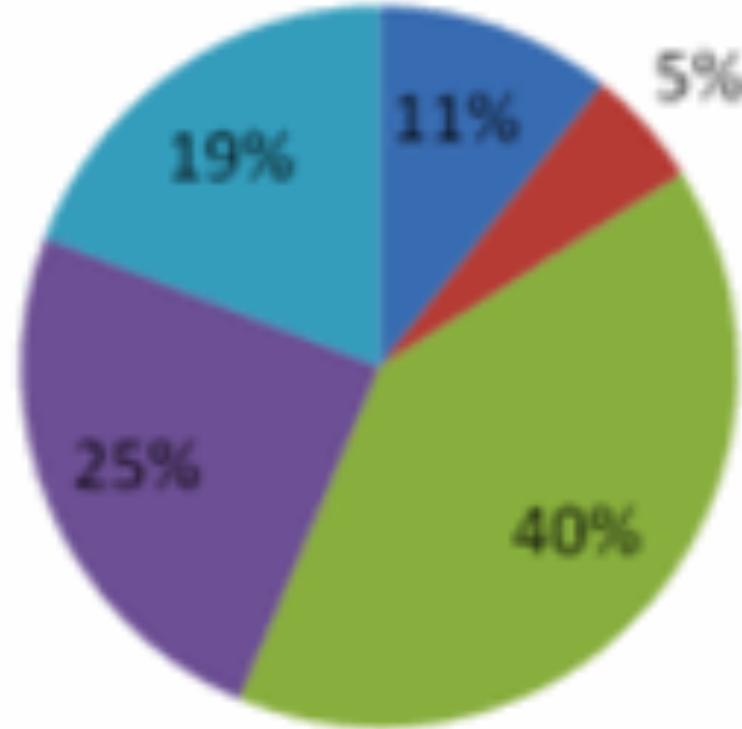
My second favorite pie chart



So, what to use instead?

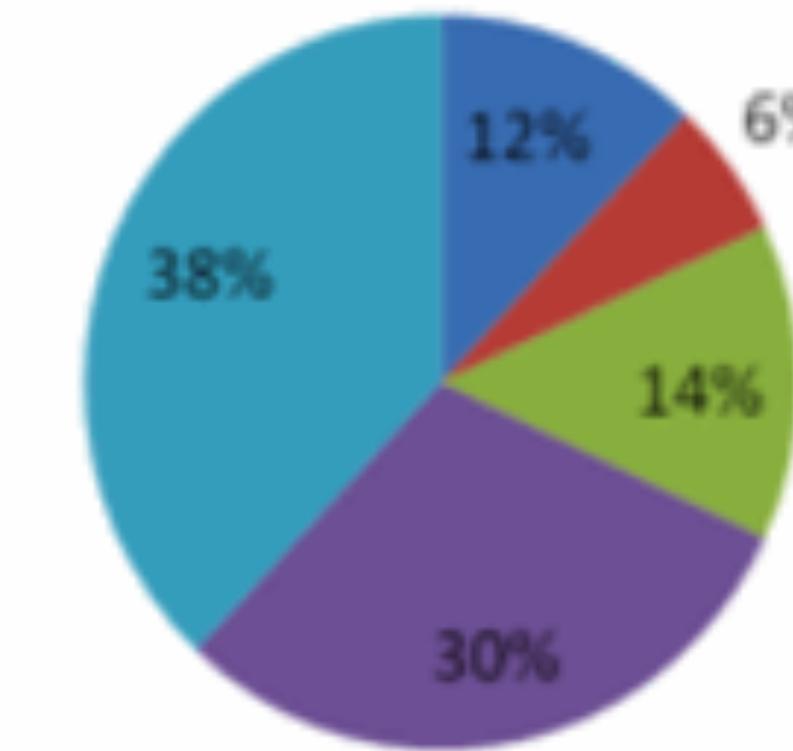
PRE: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



POST: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



imagine you just completed a pilot summer learning program on science aimed at improving perceptions of the field among 2nd and 3rd grade elementary children

Alternative #1: Show the Number(s) Directly

After the pilot program,

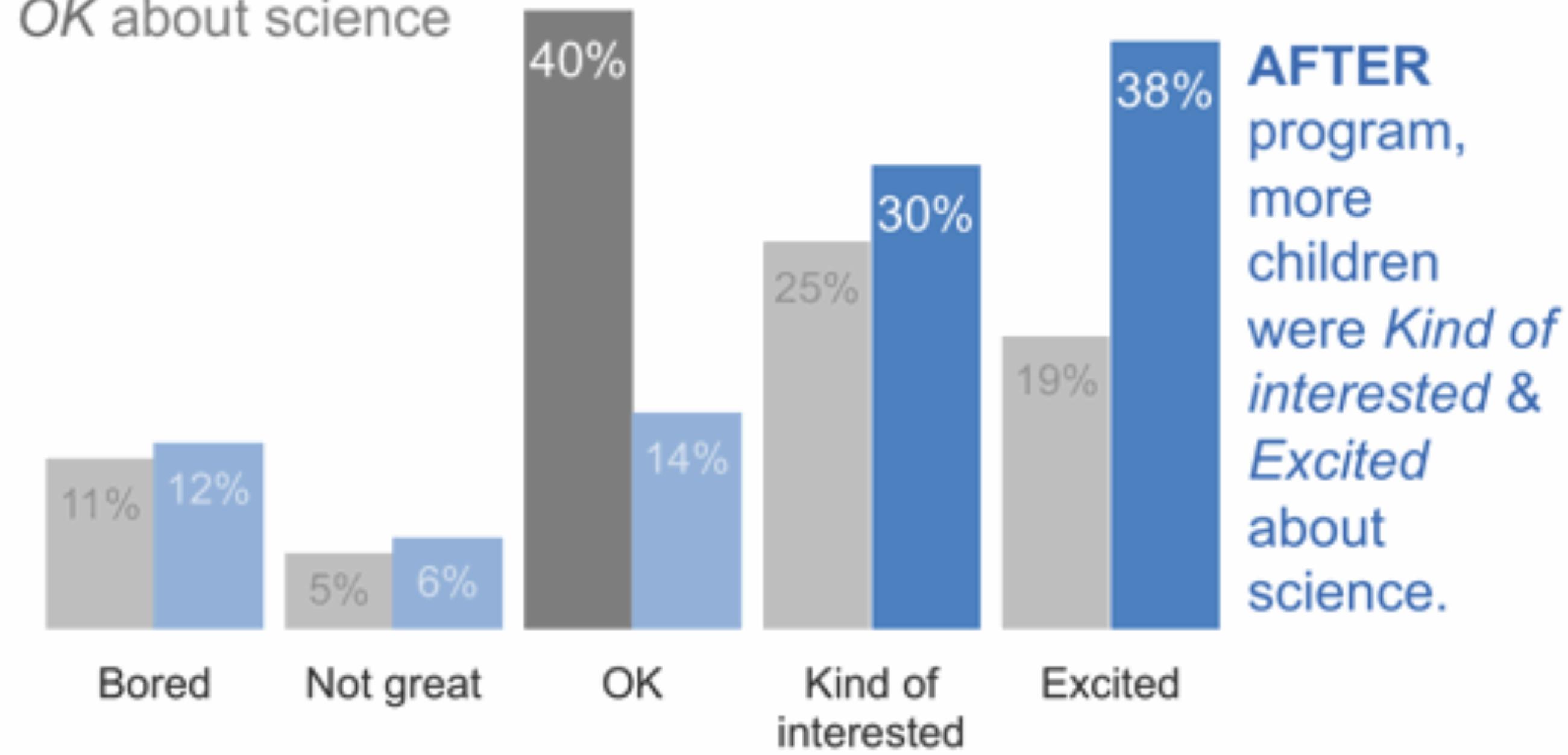
68%

of kids expressed interest towards science,
compared to 44% going into the program.

Alternative #2: Simple Bar Graph

How do you feel about science?

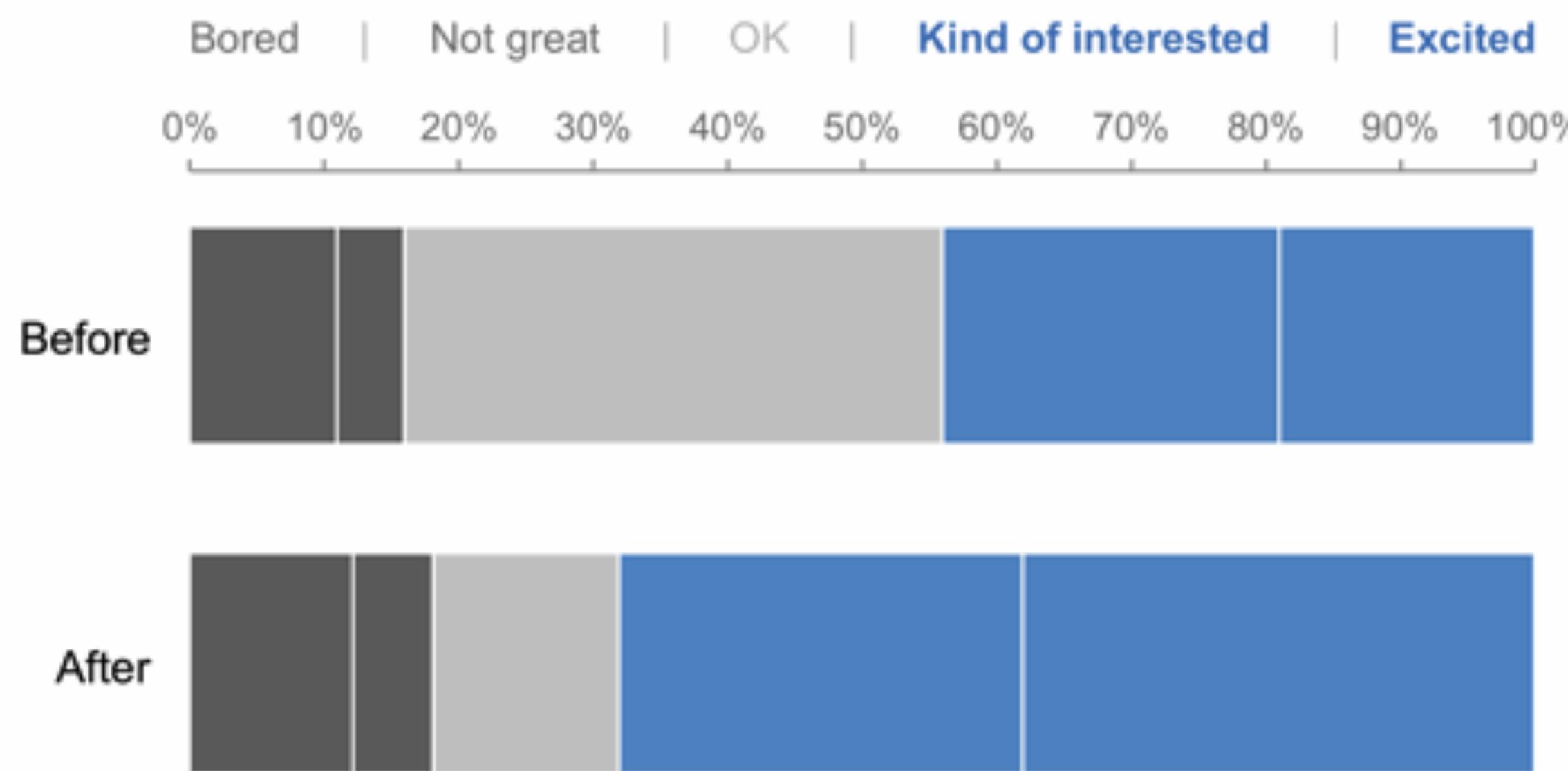
BEFORE program, the majority of children felt just OK about science



AFTER program,
more
children
were *Kind of
interested &
Excited*
about
science.

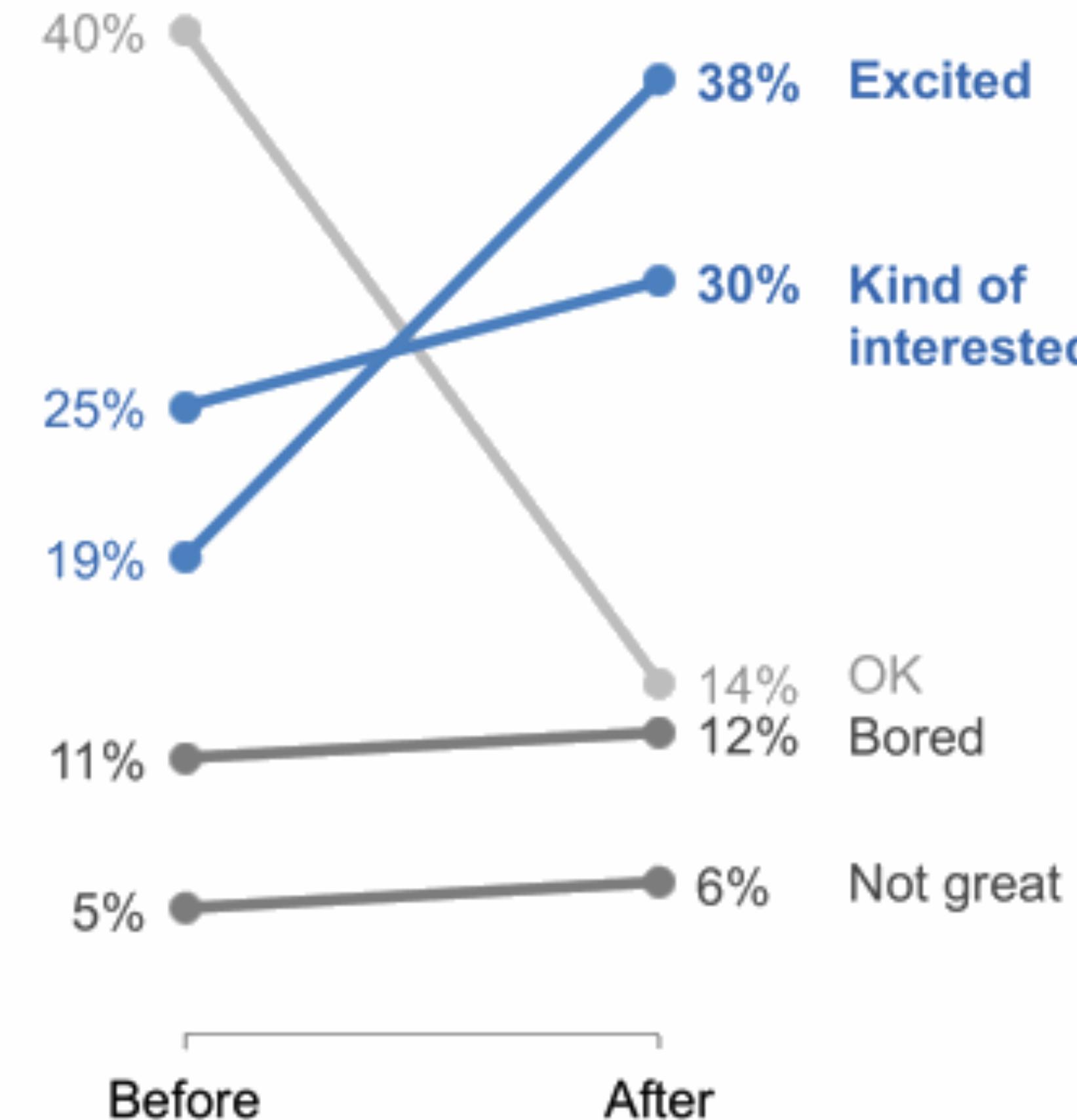
Alternative #3: 100% Stacked Horizontal Bar Graph

How do you feel about science?



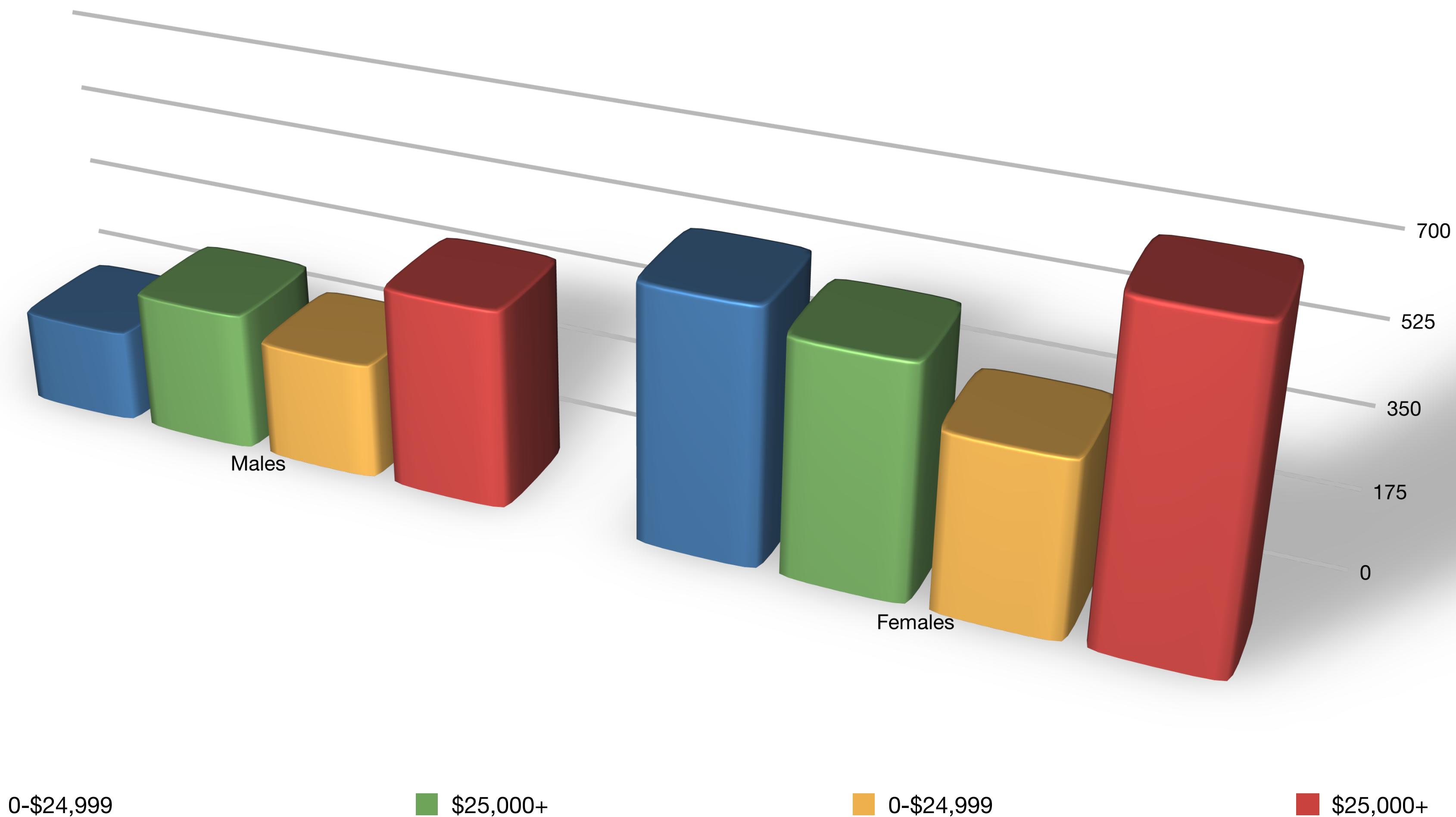
Alternative #4: Slopegraph

How do you feel about science?

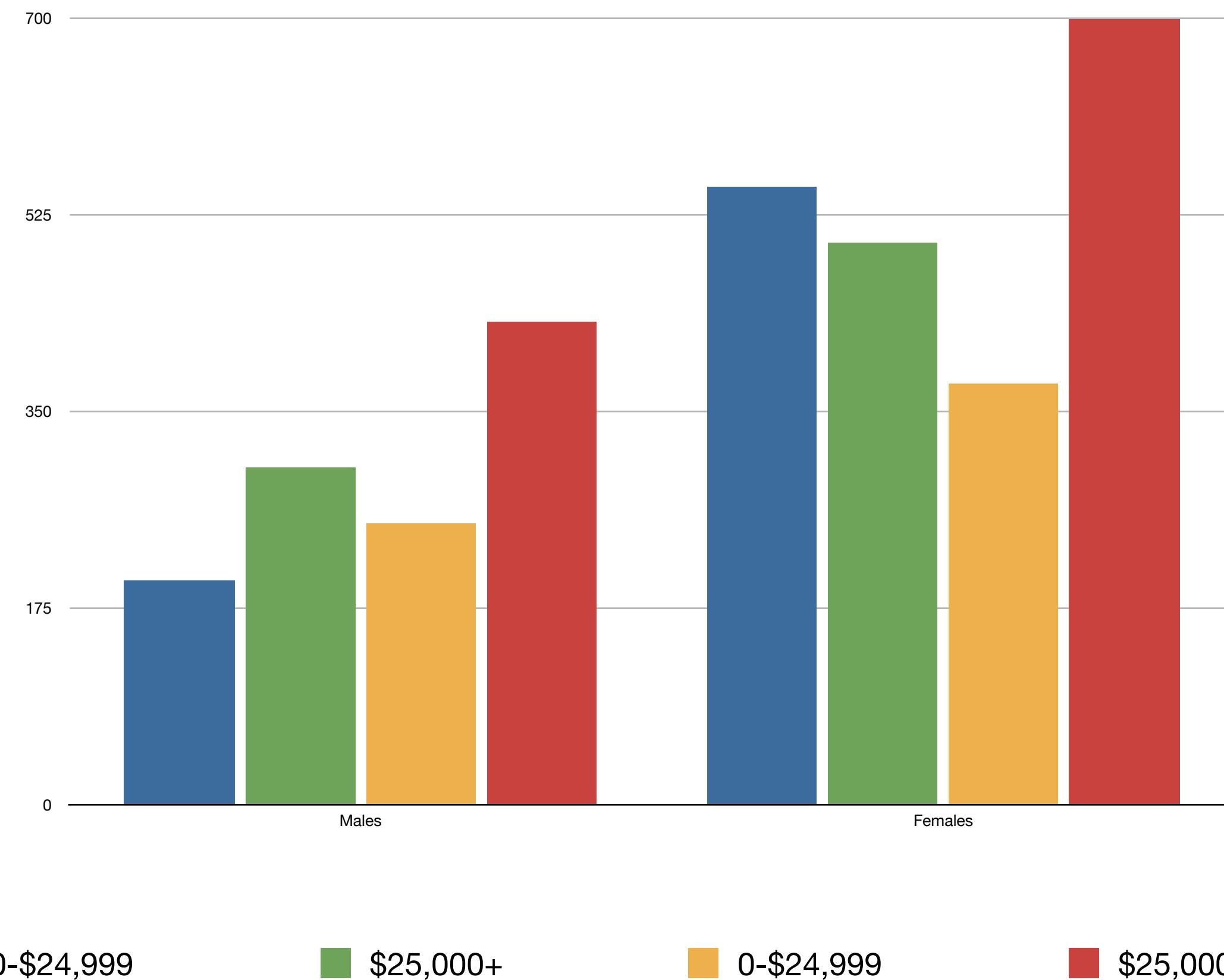


Visualization Design Principles

Maximize Data-Ink Ratio

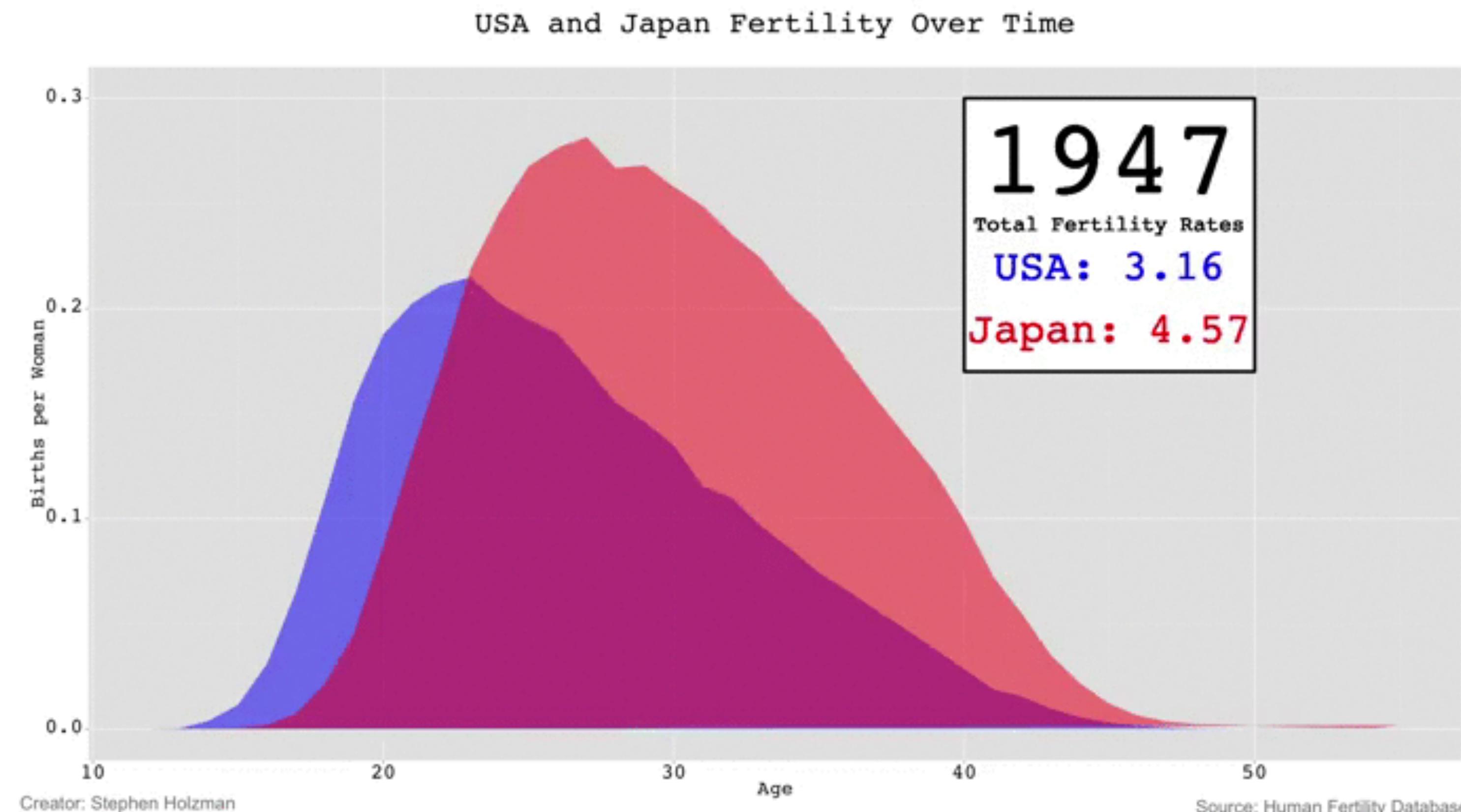


Maximize Data-Ink Ratio

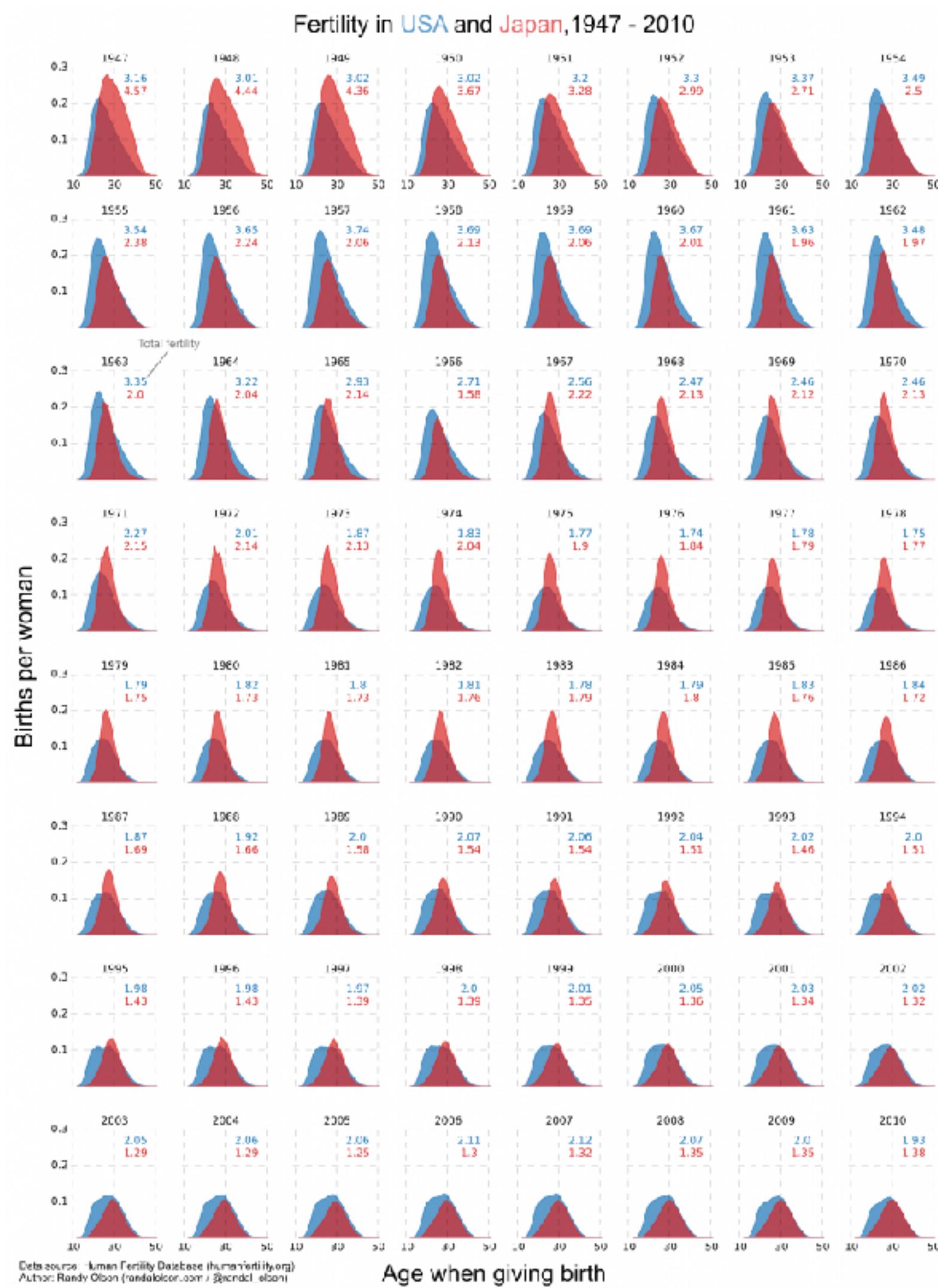


Eyes Beat Memory

Don't make people memorize: Show them

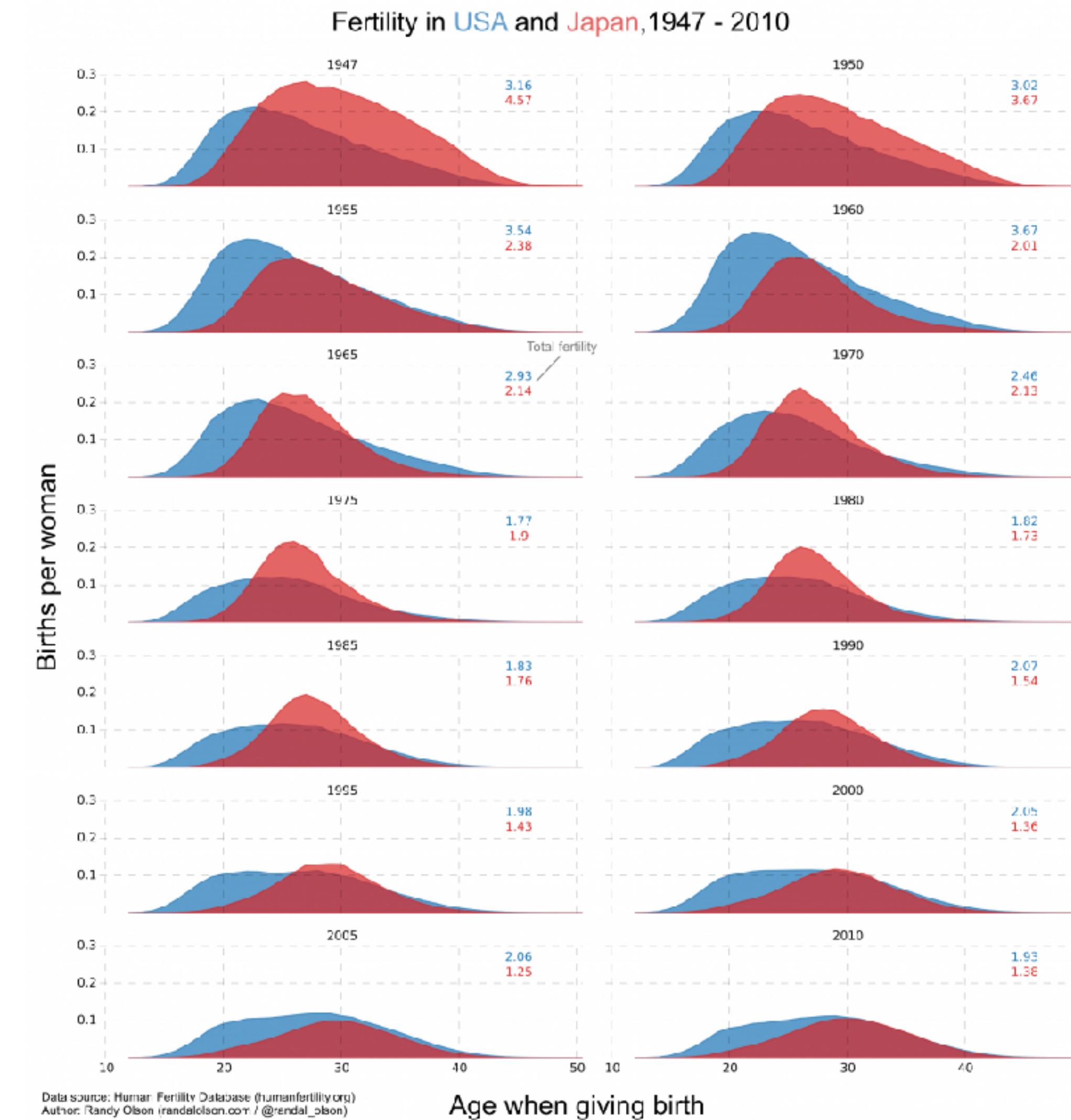


Eyes Beat Memory: Small Multiples



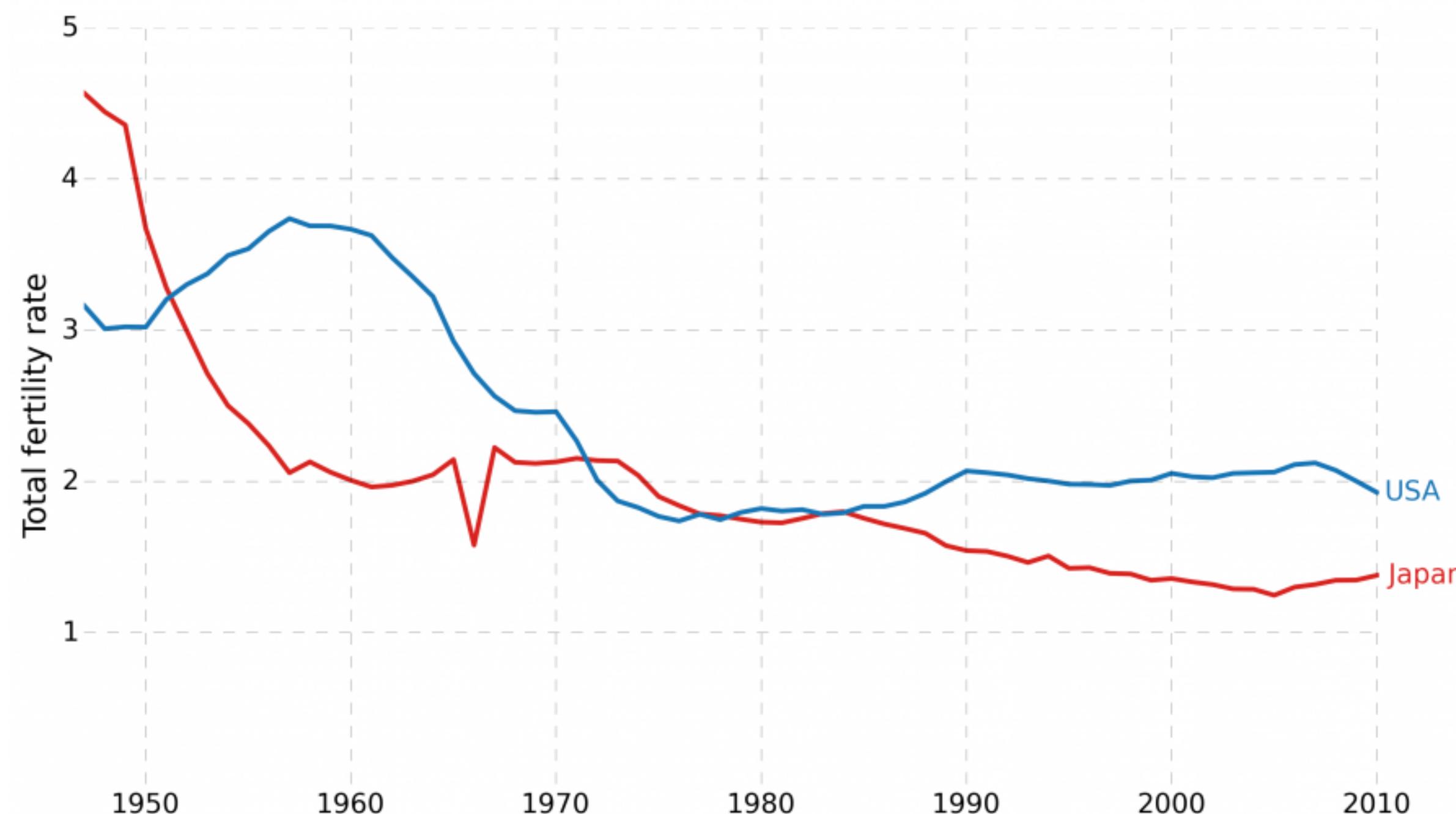
A lot of charts
Do we need all of them?

Eyes Beat Memory: Small Multiples



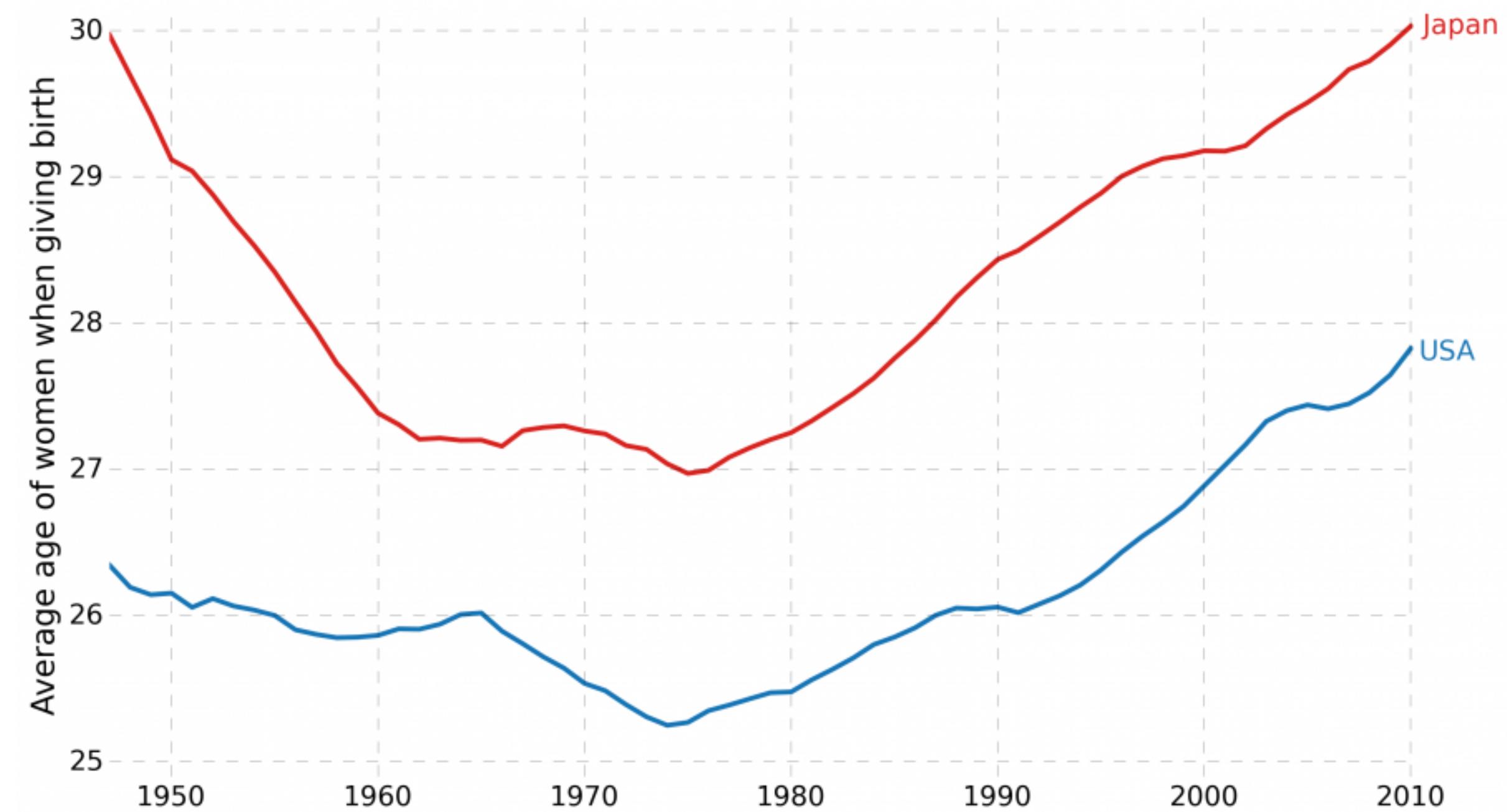
Simplify!

Total fertility rate in USA and Japan, 1947 - 2010



Data source: Human Fertility Database (humanfertility.org)
Author: Randy Olson (randalolson.com / [@randal_olson](https://twitter.com/randal_olson))

Average age when giving birth in USA and Japan, 1947 - 2010



Data source: Human Fertility Database (humanfertility.org)
Author: Randy Olson (randalolson.com / [@randal_olson](https://twitter.com/randal_olson))

Tabular Data

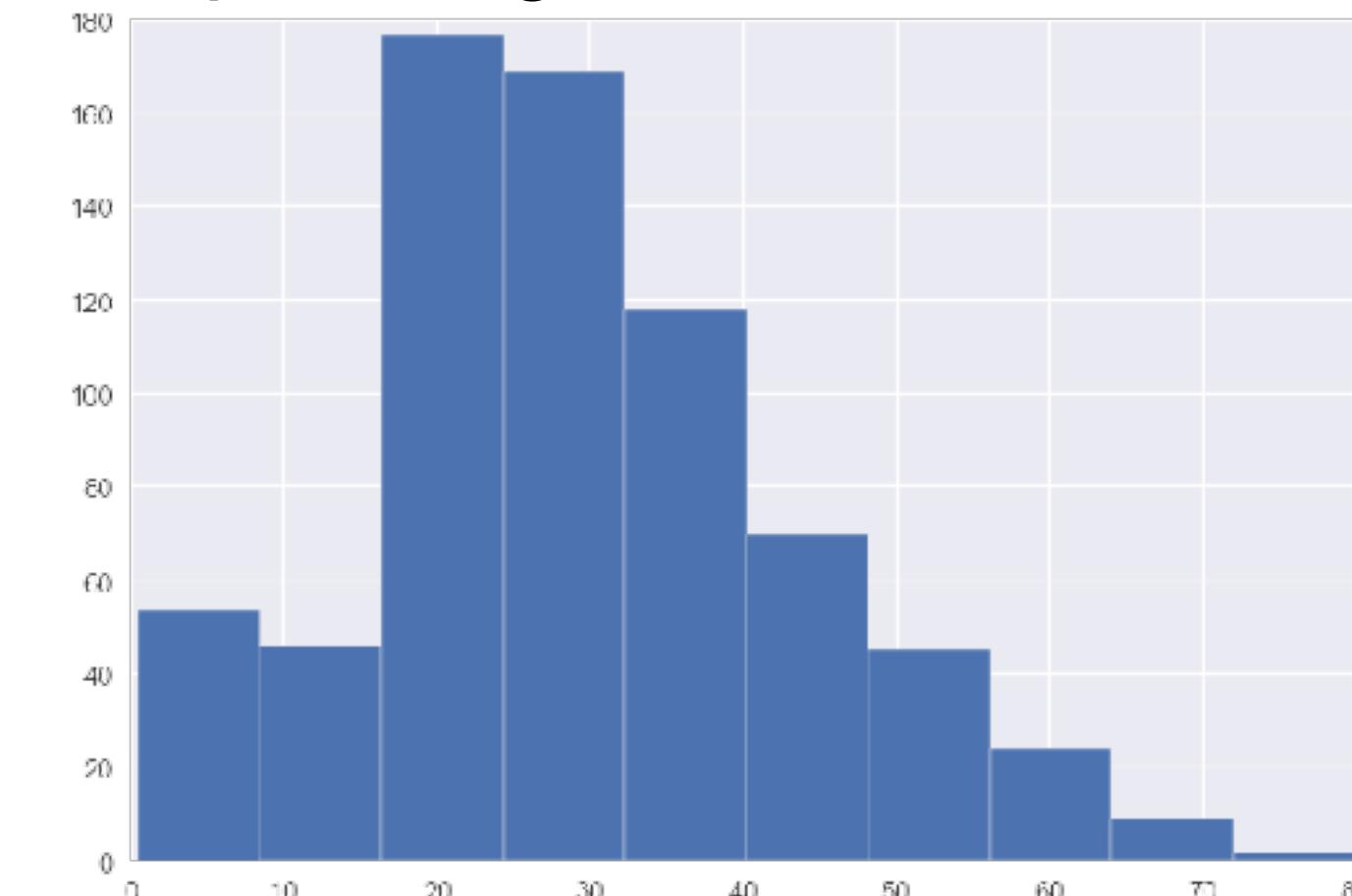
Histogram

Good #bins hard to predict
make interactive!
rules of thumb:

$$\# \text{bins} = \sqrt{n}$$

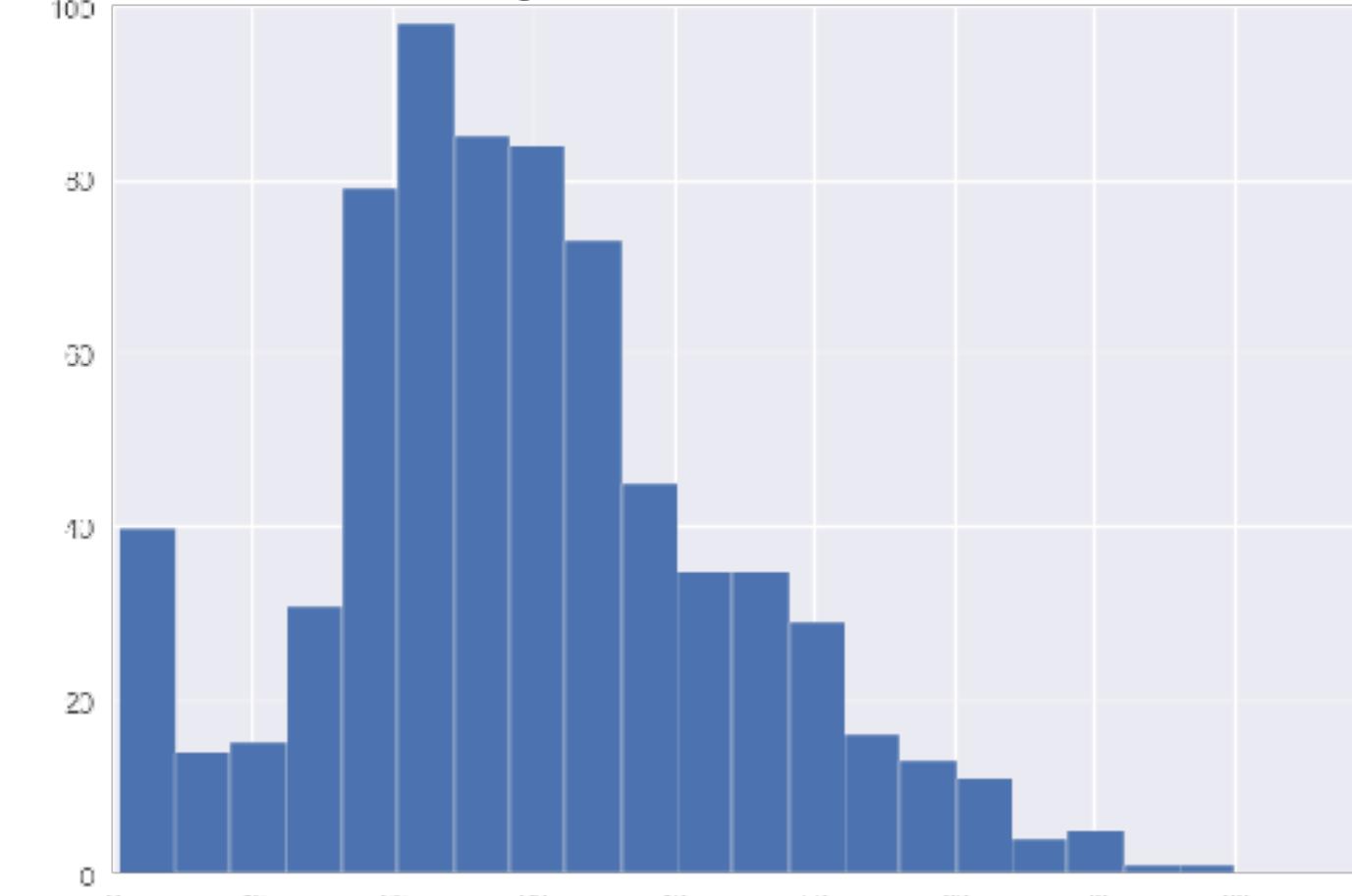
$$\# \text{bins} = \log_2(n) + 1$$

passengers



10 Bins

passengers



20 Bins

age

age

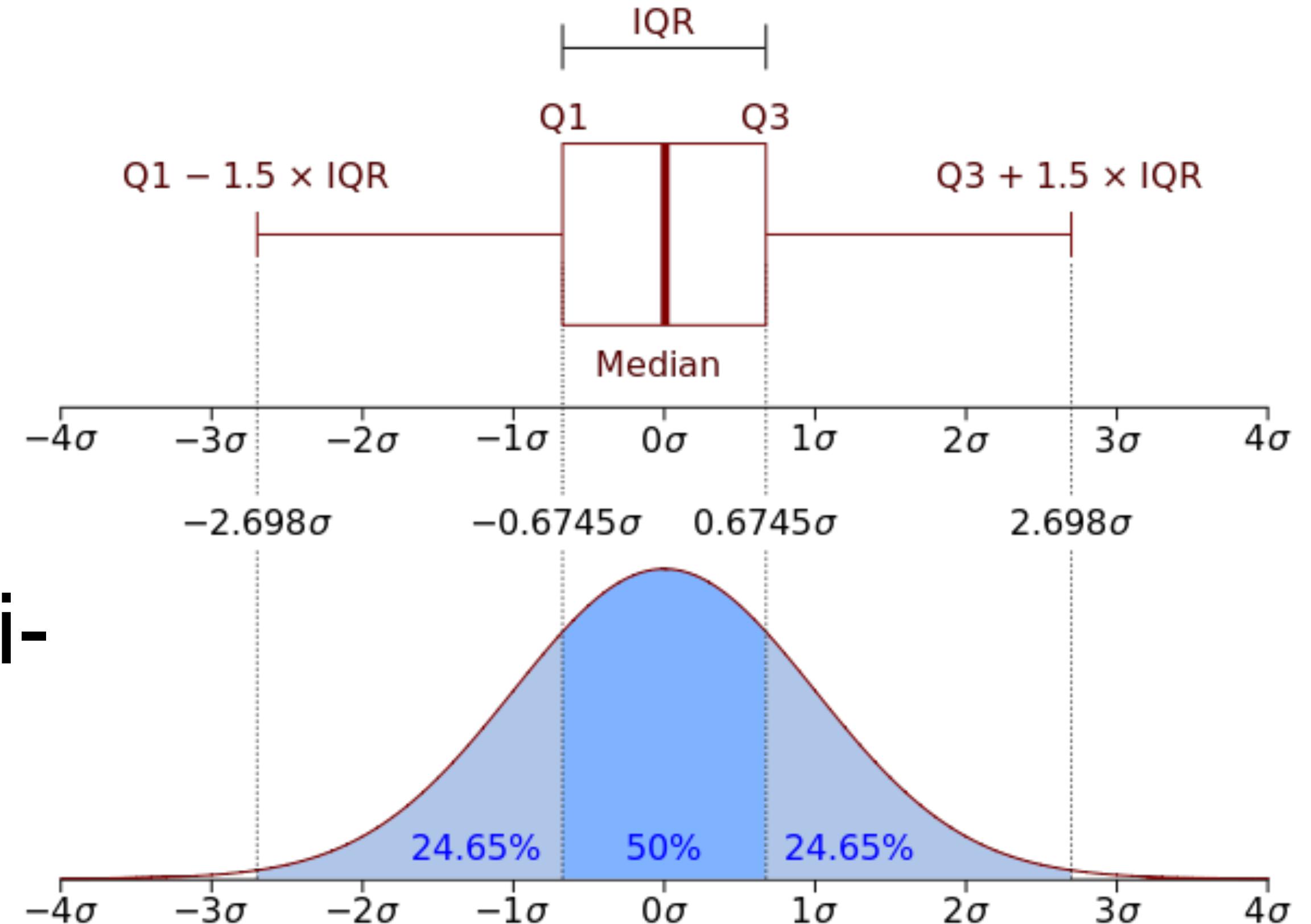
Box Plots

aka Box-and-Whisker Plot

Show outliers as points!

Not so great for non-normal distributed data

Especially bad for bi- or multi-modal distributions



One Boxplot, Four Distributions

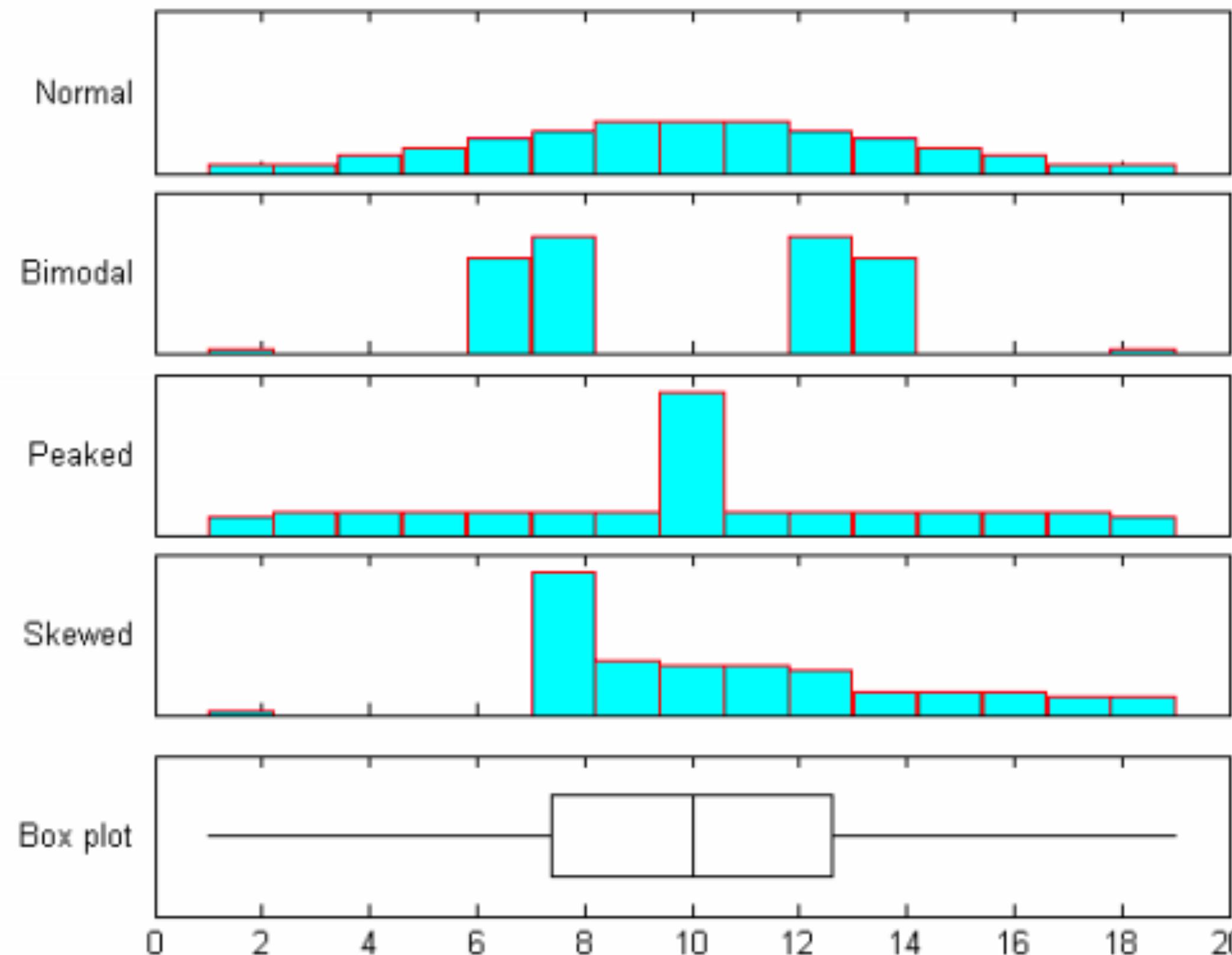
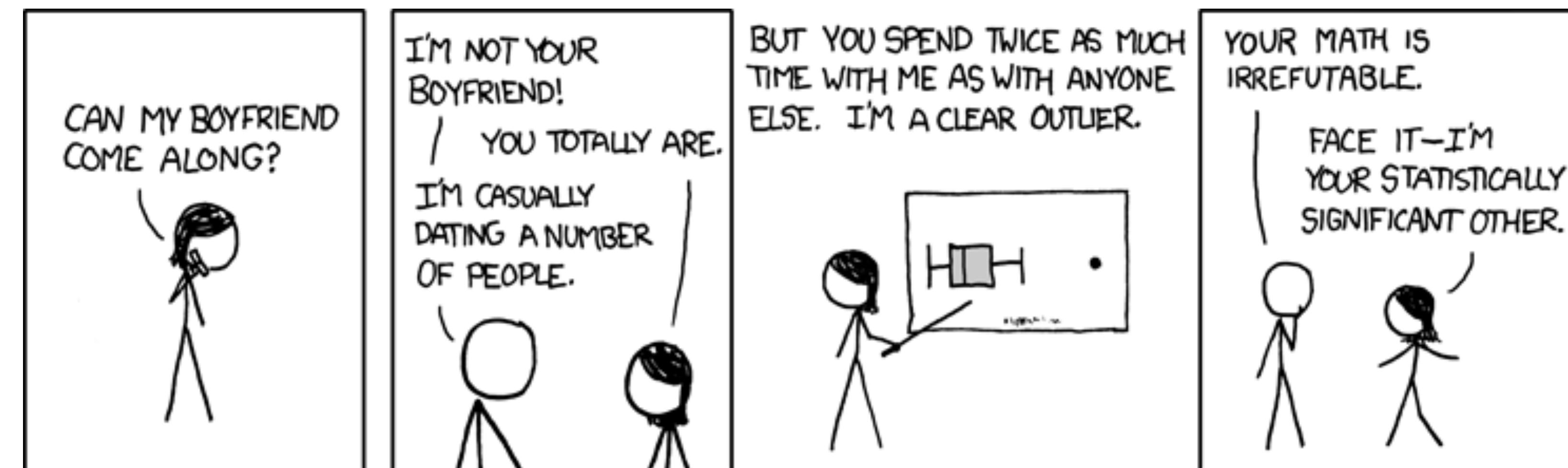
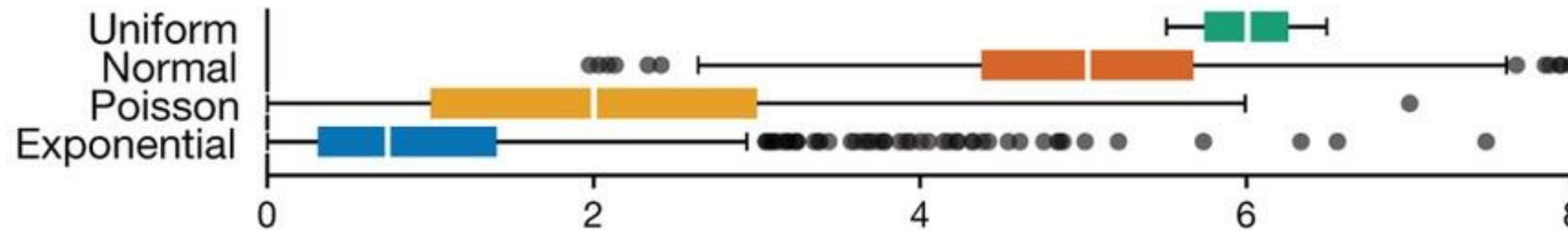


Figure 1: Histograms and box plot: four samples each of size 100

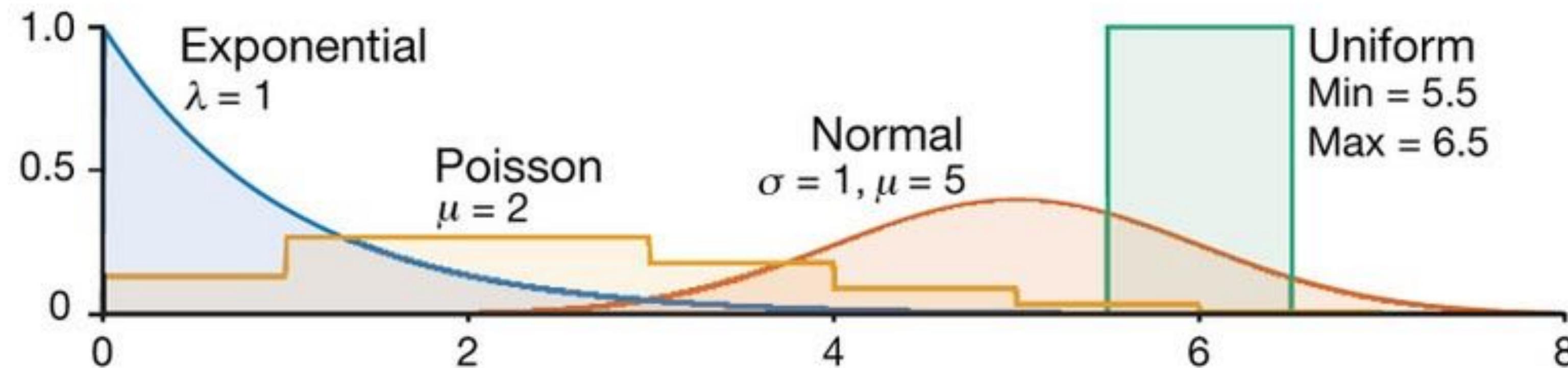
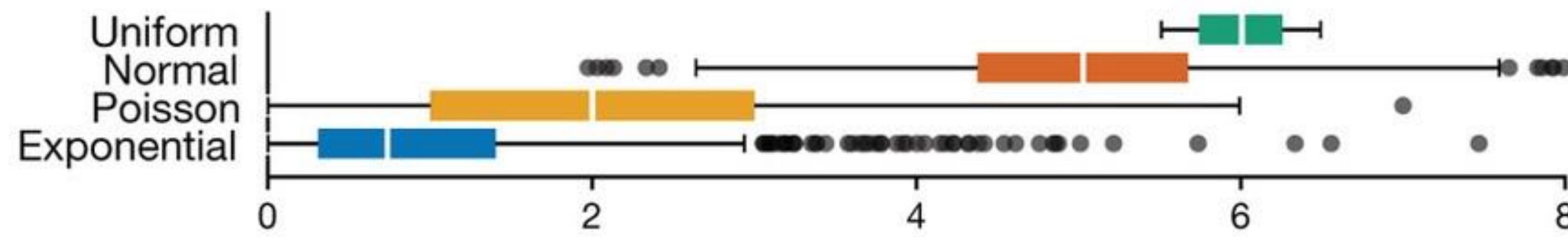
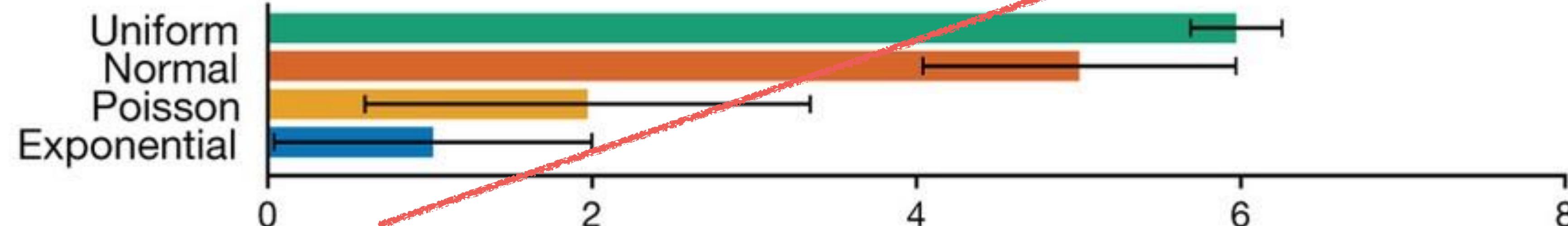
Box(and Whisker) Plots



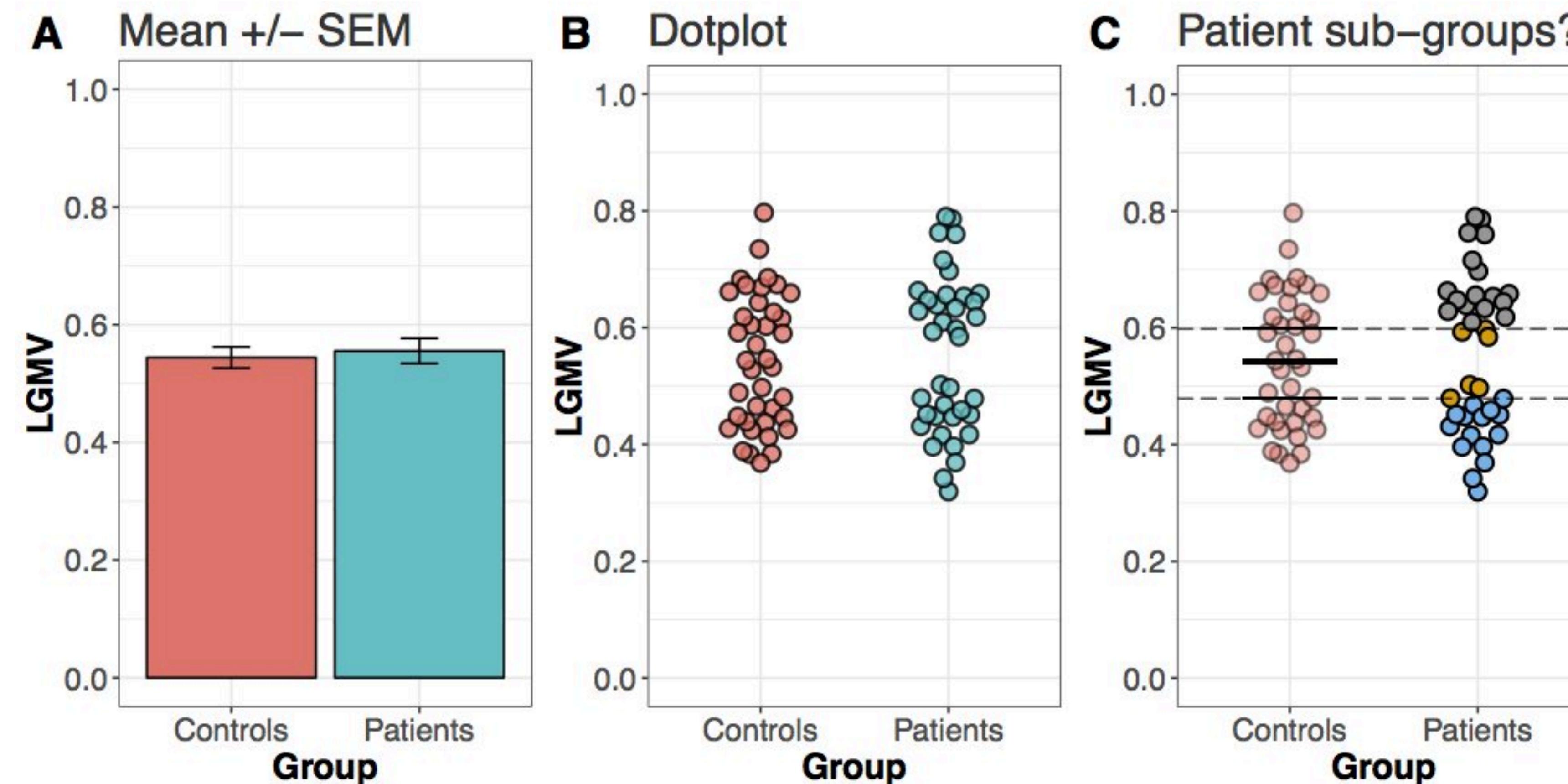
<http://xkcd.com/539/>



Comparison

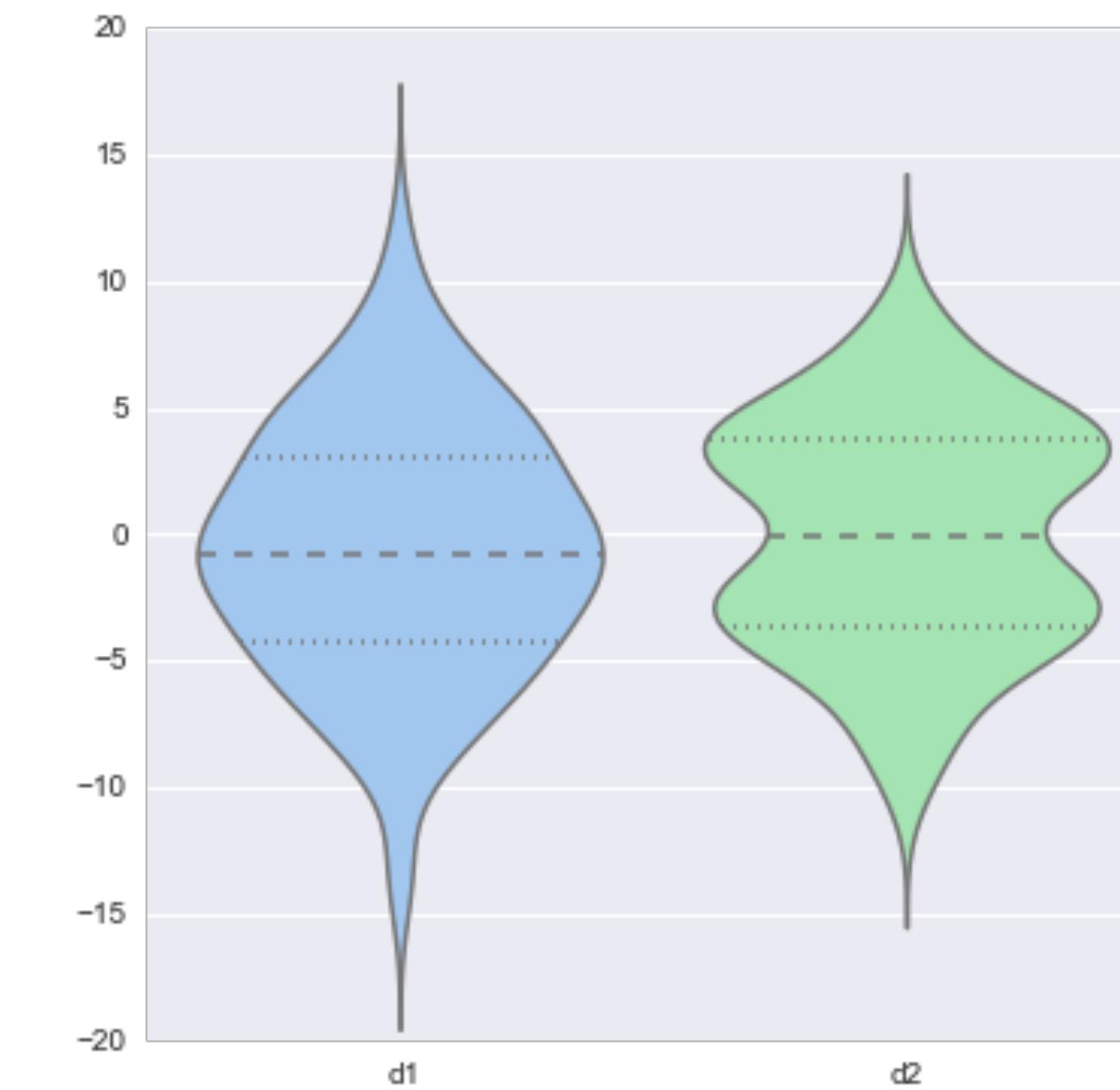
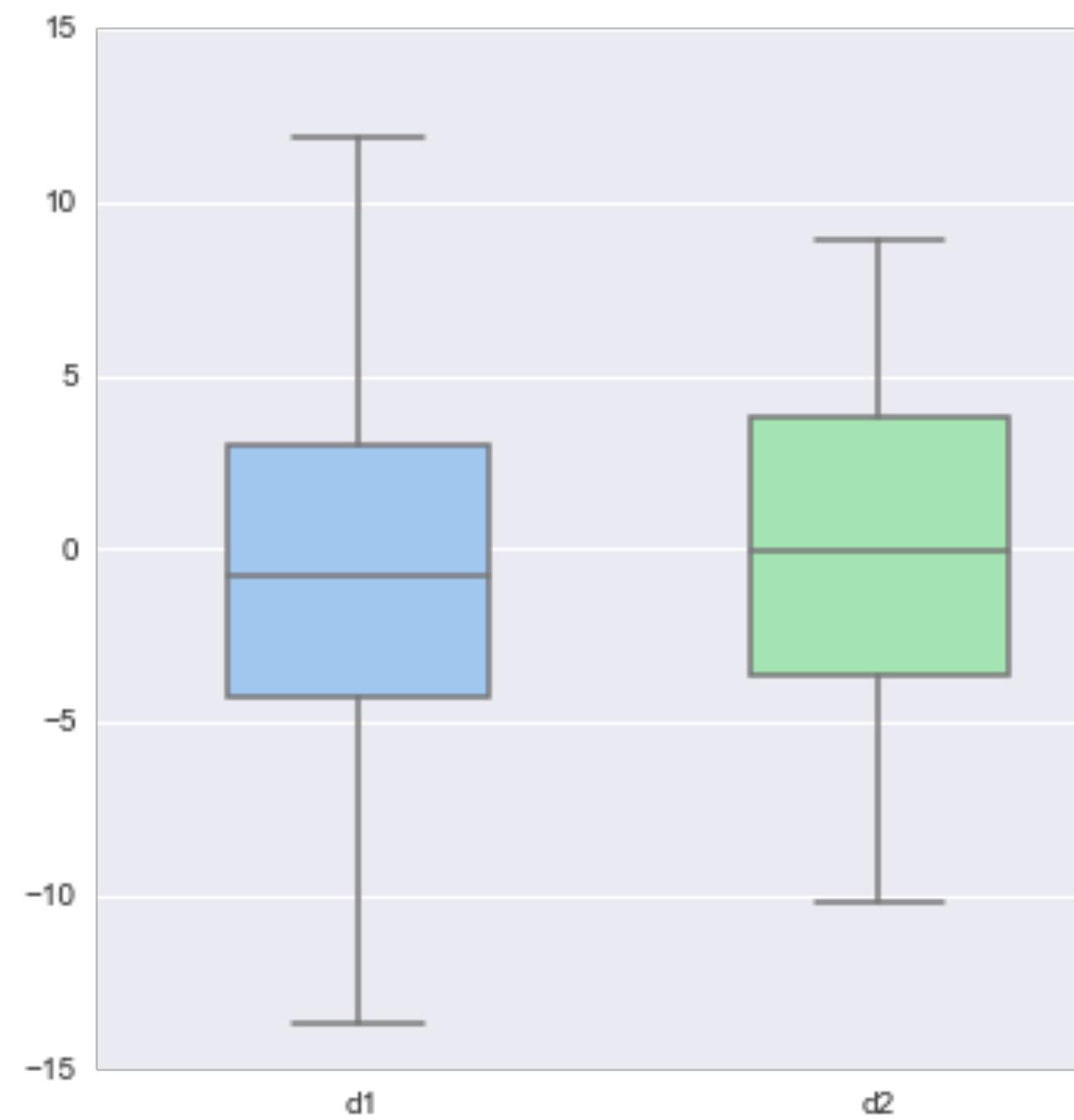


Bar Charts vs Dot Plots



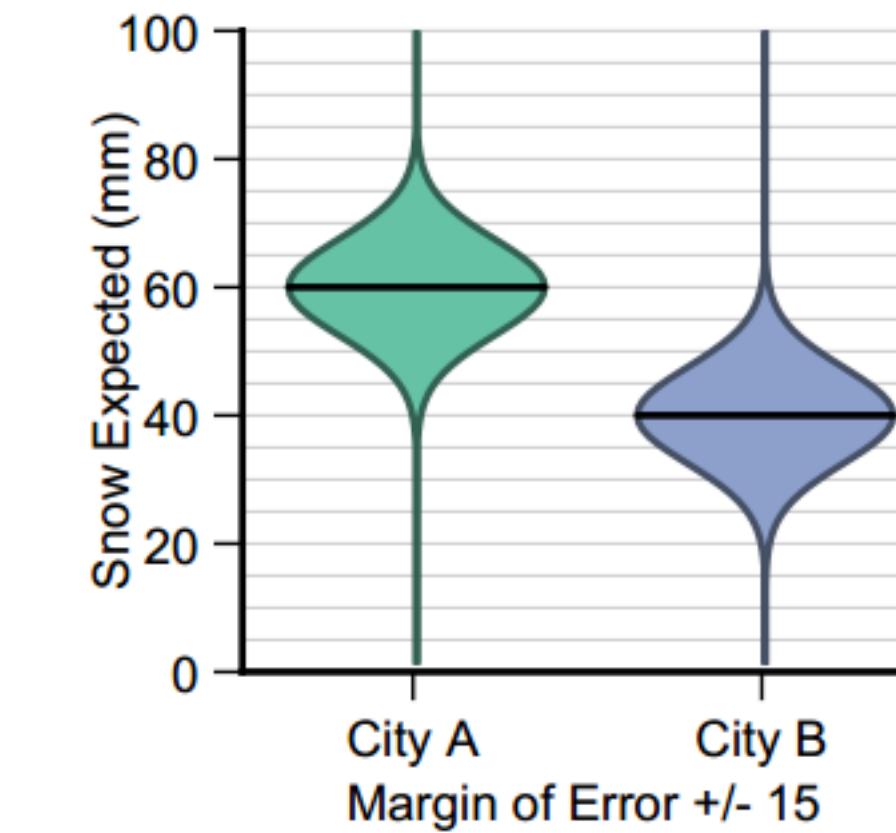
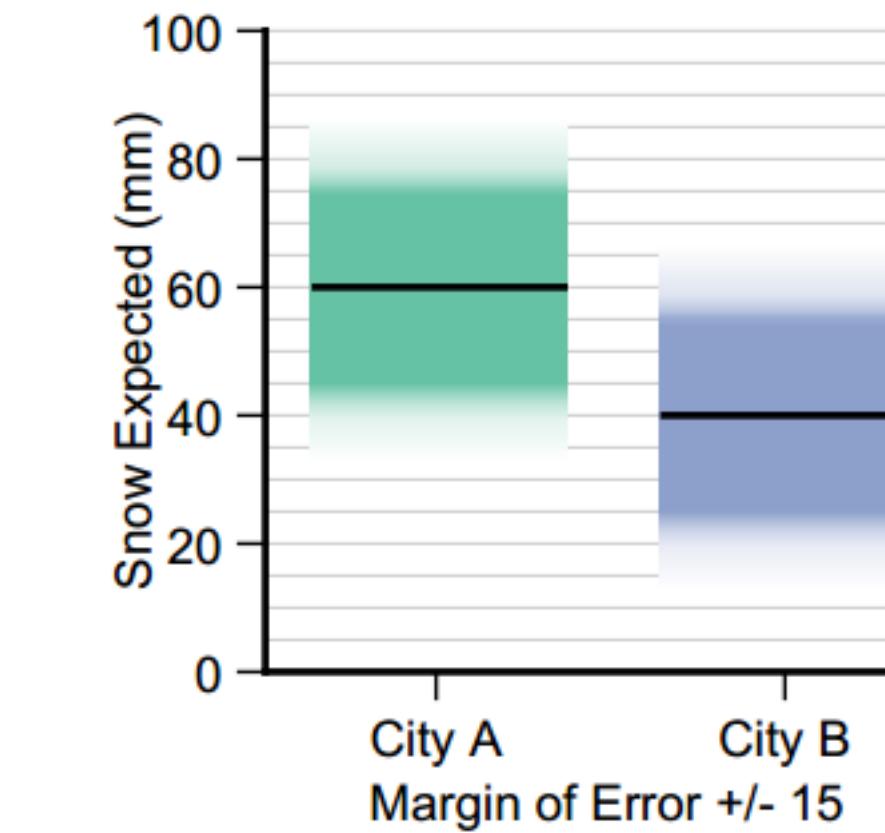
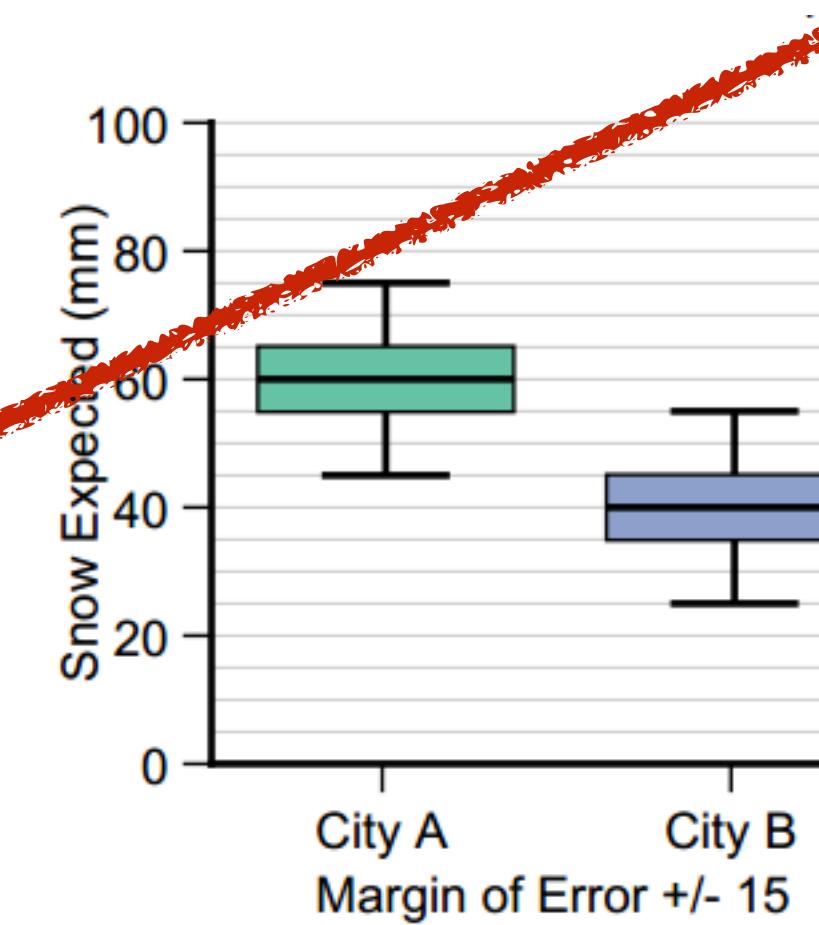
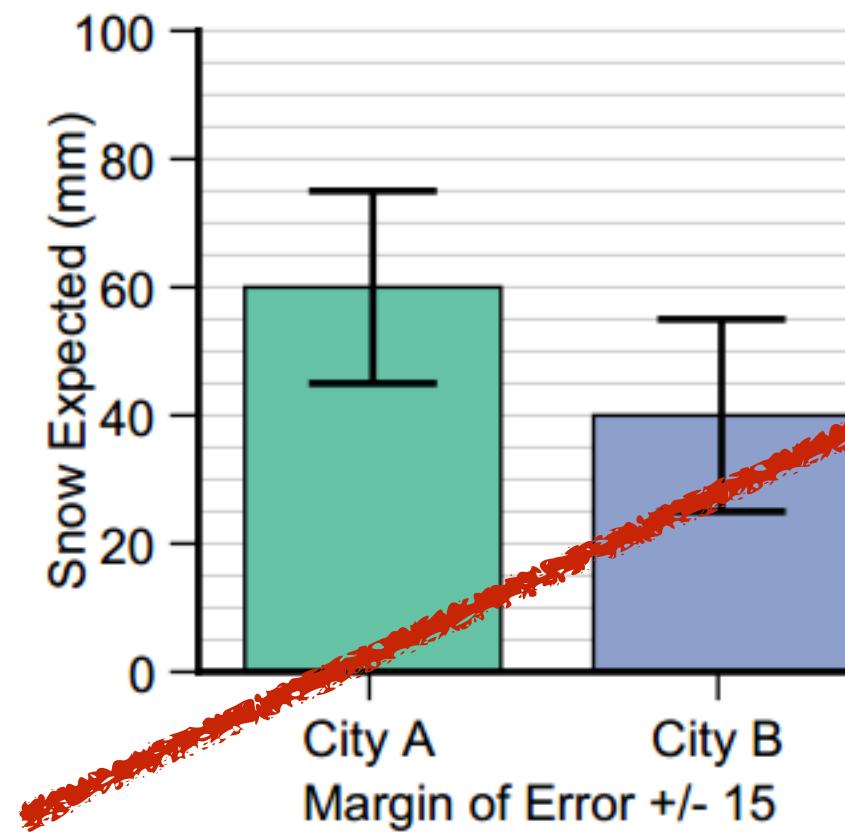
Violin Plot

= Box Plot + Probability Density Function



Showing Expected Values & Uncertainty

NOT a distribution!

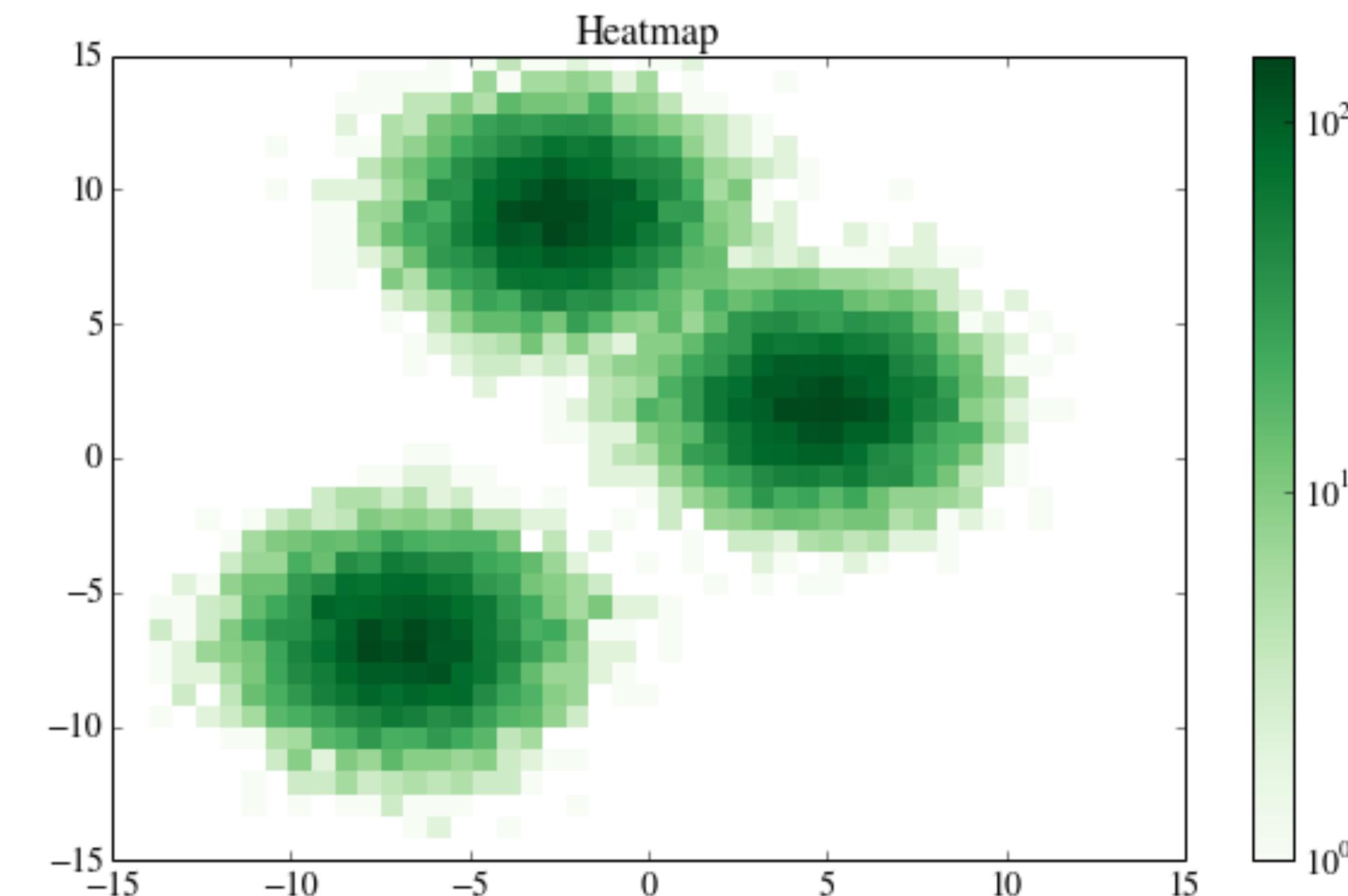
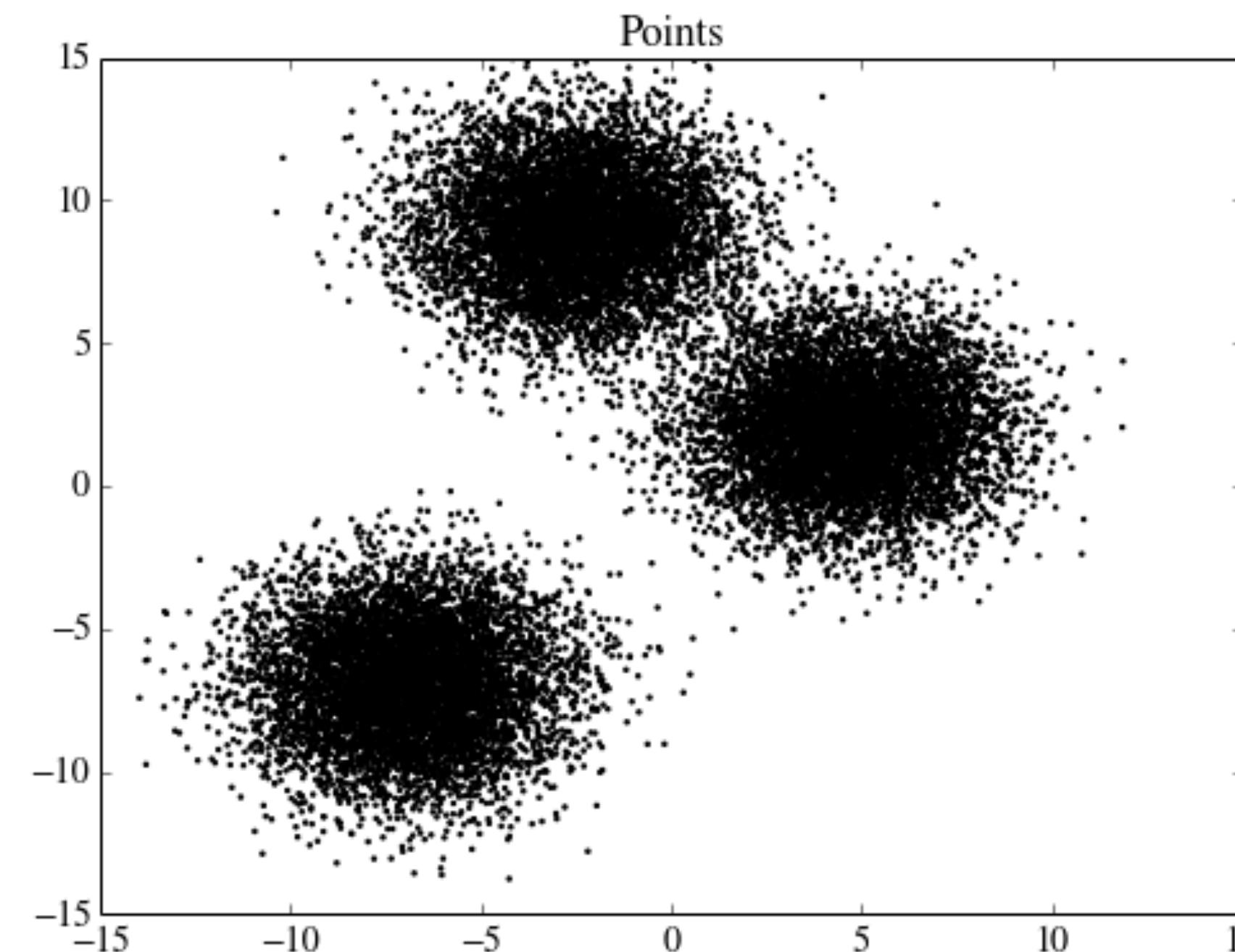


Error Bars Considered Harmful:
Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher

Heat Maps

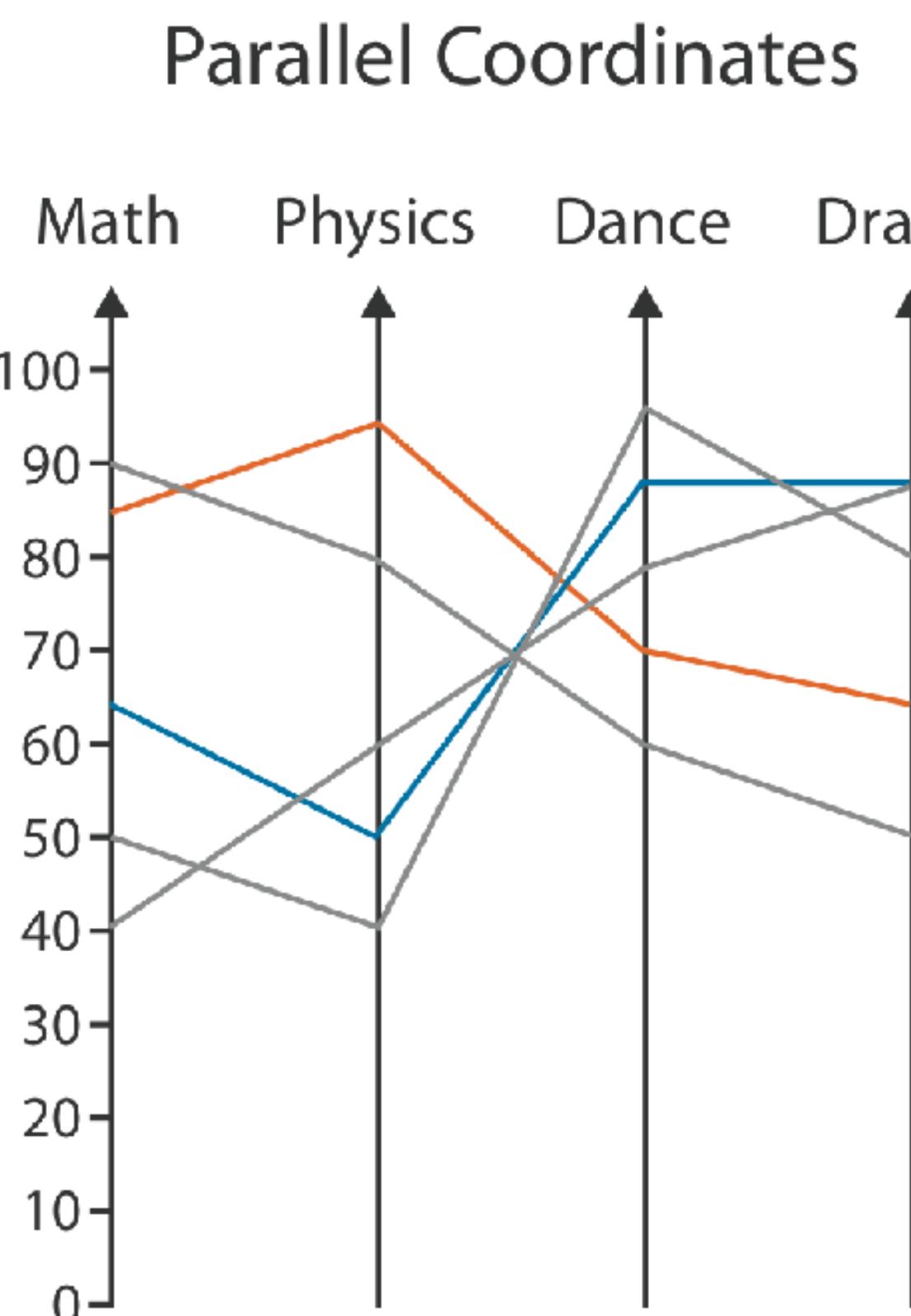
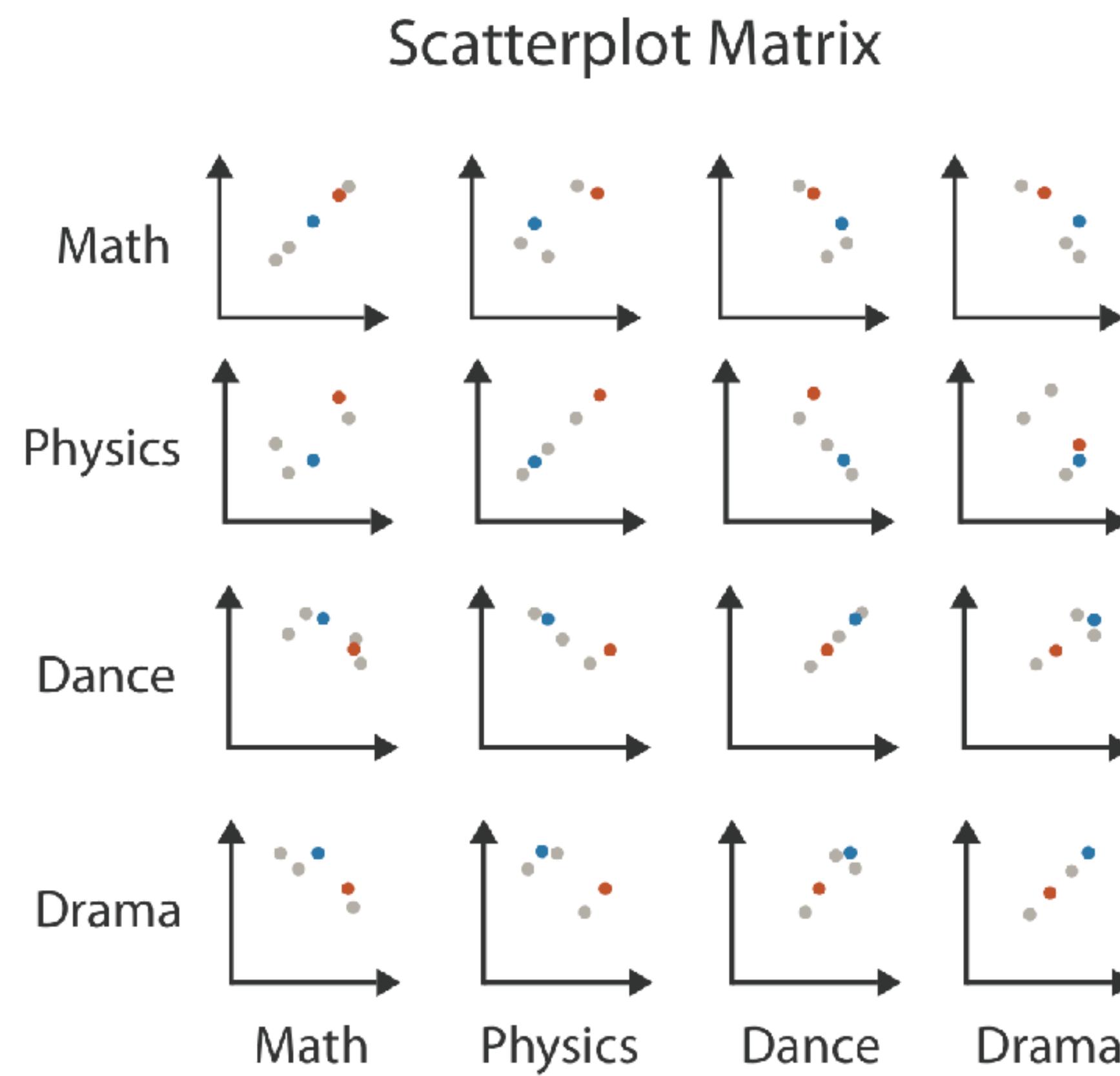
binning of scatterplots

instead of drawing every point, calculate grid and intensities



2D Density Plots

Scatterplot Matrix & PCs



Table

	Math	Physics	Dance	Drama
Math	85	95	70	65
Physics	90	80	60	50
Dance	65	50	90	90
Drama	50	40	95	80
	40	60	80	90

Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

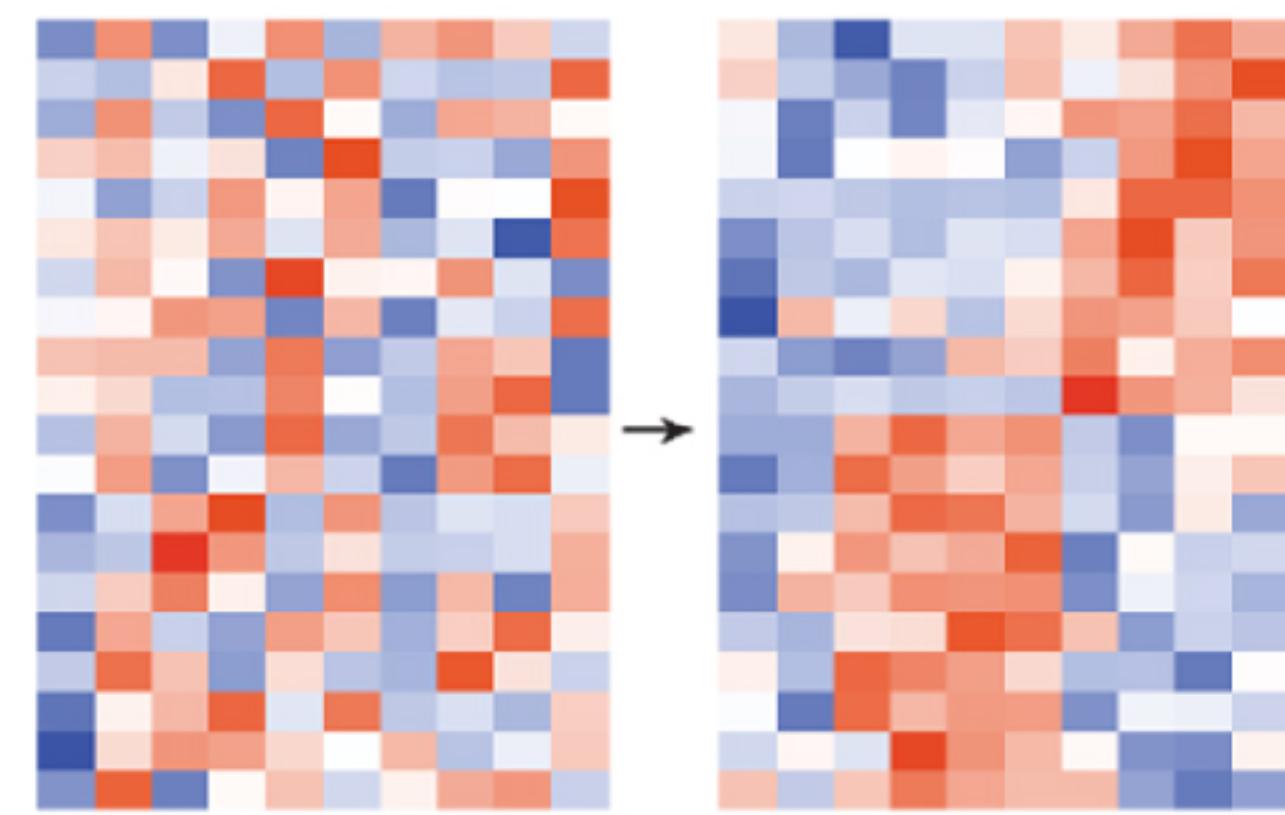
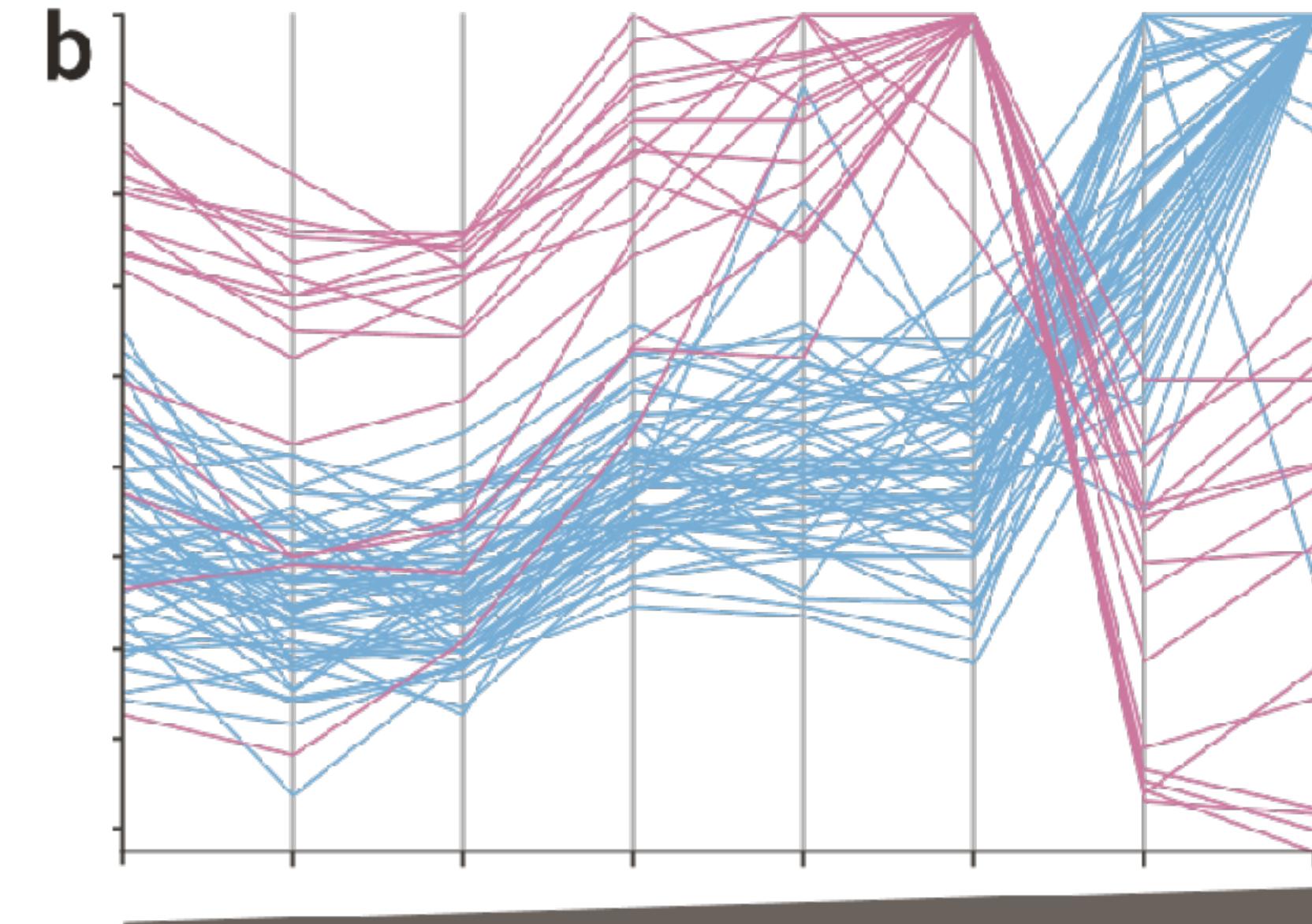
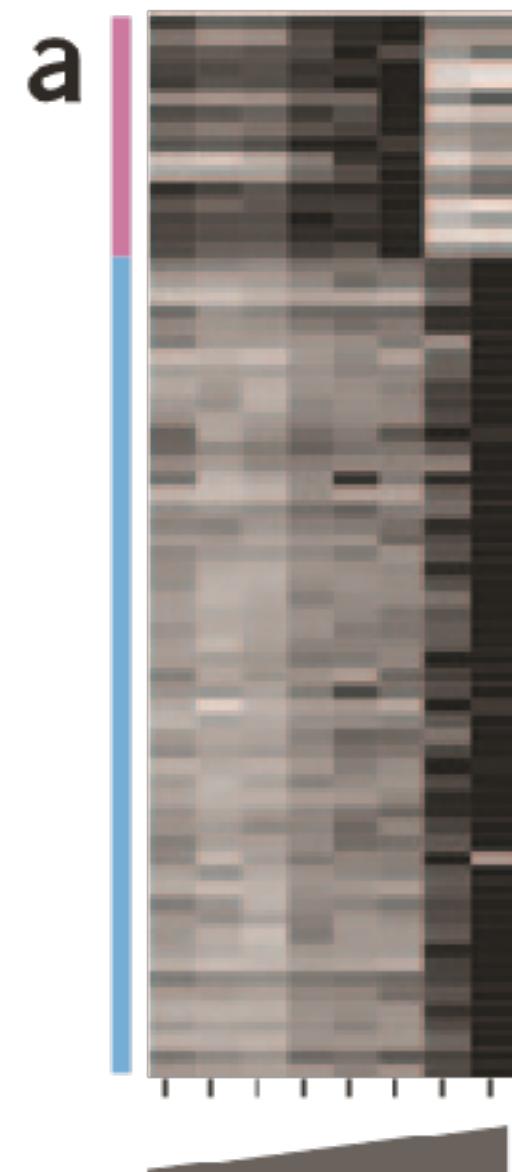
Ordering critical for interpretation

If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

same scale & type



Heterogeneous Data?

