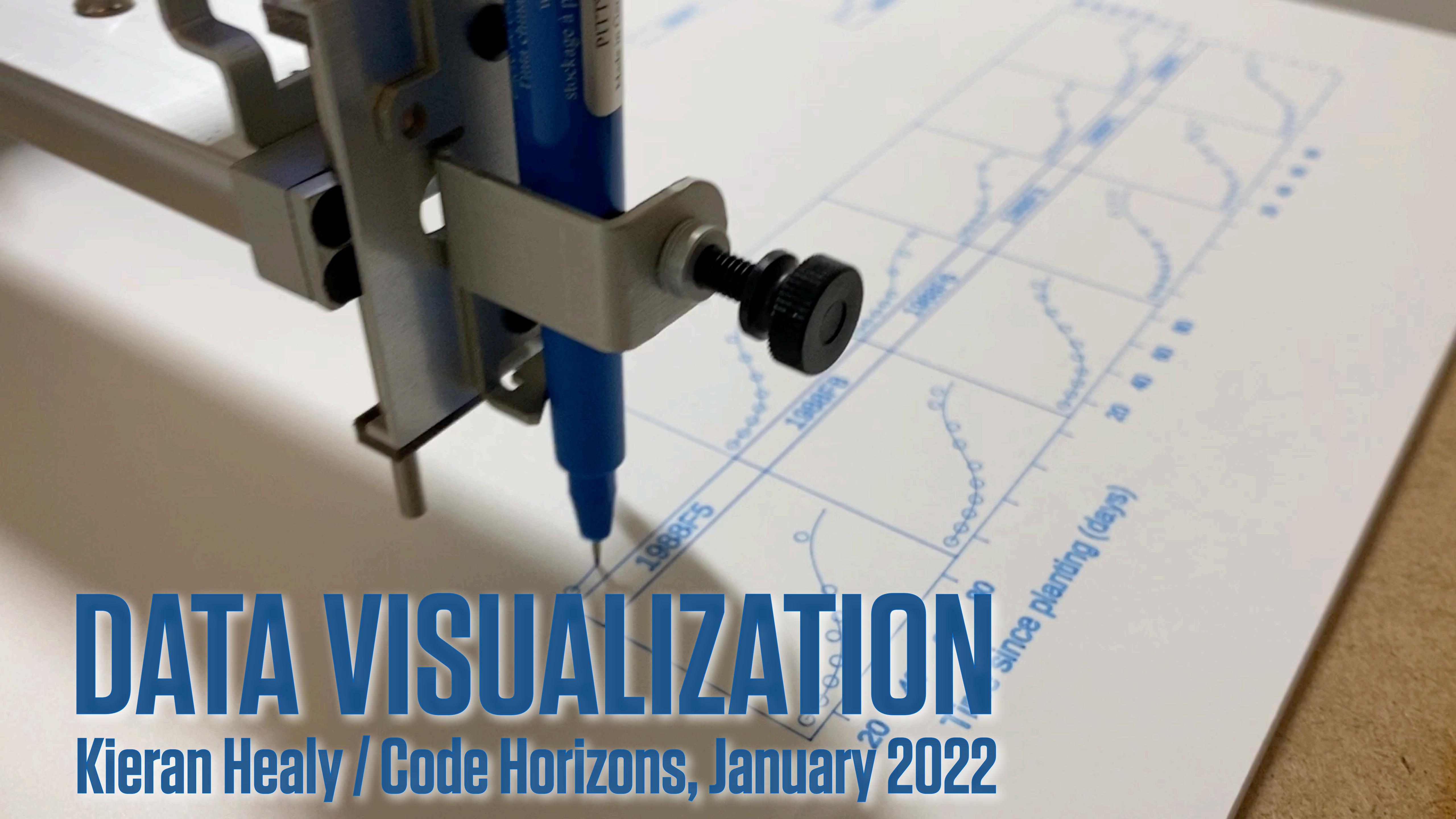


DATA VISUALIZATION

Kieran Healy / Code Horizons, January 2022



Lecture: 10am-2pm EST

Breaks on the hour, or thereabouts

Use Zoom's Chat to ask questions during class

Use Slack to discuss things between classes

Office hour: 4pm-5pm EST

Housekeeping



RStudio

Get Up & Running

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

stathorizons_0820 - master - RStudio

Console Jobs ~/Documents/courses/stathorizons_0820/

R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

knitr hook "anchor" is now available
Loading required package: testthat

Attaching package: 'testthat'

The following object is masked from 'package:devtools':

test_file

> |

RStudio

Files	Plots	Packages	Help	Viewer
New Folder	Delete	Rename	More	
Home > Documents > courses > stathorizons_0820				
Name	Size	Modified		
..				
.gitignore	40 B	Aug 3, 2020, 10:00 AM		
01_introduction.Rmd	4 KB	Aug 3, 2020, 11:50 AM		
02_get_started.Rmd	4 KB	Aug 3, 2020, 11:46 AM		
		Aug 3, 2020, 11:46 AM		
04_show_the_right_numbers.Rmd	5 KB	Aug 3, 2020, 11:46 AM		
05_tables_and_labels.Rmd	10.9 KB	Aug 3, 2020, 11:46 AM		
06_models.Rmd	14.7 KB	Aug 3, 2020, 11:46 AM		
07_maps.Rmd	12.5 KB	Aug 3, 2020, 11:46 AM		
08_refine_plots.Rmd	21.7 KB	Aug 3, 2020, 11:46 AM		
09_supplementary_material.Rmd	17 KB	Aug 3, 2020, 11:46 AM		
assets				
data				
figures				
keynote				
LICENSE.md	18.1 KB	Jul 21, 2020, 11:16 AM		
materials				
R				
README.md	5.4 KB	Aug 3, 2020, 11:46 AM		
slides				
stathorizons_0820.Rproj	205 B	Aug 3, 2020, 11:50 AM		

RStudio

The screenshot shows the RStudio interface with a red box highlighting the left pane. The left pane contains an R Markdown file named '01_introduction.Rmd'. The code includes metadata at the top, followed by sections for notes and a statement about the document being an R Markdown file. The right pane shows the project structure for 'stathorizons_0820', which includes an 'assets' folder, several RMD files (e.g., 01_introduction.Rmd, 02_get_started.Rmd, 03_make_a_plot.Rmd), and other files like .gitignore, .Rhistory, LICENSE.md, and README.md.

```
stathorizons_0820 - master - RStudio
```

```
01_introduction.Rmd x
```

```
1 ---  
2 title: "Data Visualization"  
3 author: "Kieran Healy"  
4 date: "10-January-2020"  
5 output: html_document  
6 ---  
7  
8 ## Data Visualization Notes  
9  
10 This is a starter RMarkdown project template to accompany courses taught with  
*Data Visualization*. You can use it to take notes, write your code, and  
produce a good-looking, reproducible document that records the work you have  
done. At the very top of the file is a section of *metadata*, or information  
about what the file is and what it does. The metadata is delimited by three  
dashes at the start and another three at the end. You should change the title,  
author, and date to the values that suit you. Keep the 'output' line as it is  
for now, however. Each line in the metadata has a structure. First the *key*  
("title", "author", etc), then a colon, and then the *value* associated with  
the key.  
11  
12 ## This Document is an RMarkdown File  
13  
14 Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word  
# Data visualization =
```

Console Jobs ~ /Documents/courses/stathorizons_0820/

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
knitr hook "anchor" is now available  
Loading required package: testthat  
  
Attaching package: 'testthat'  
  
The following object is masked from 'package:devtools':  
  
test_file  
>
```

```
Environment History Connections Git Tutorial
```

```
Insert Import Dataset
```

```
Global Environment
```

```
Functions
```

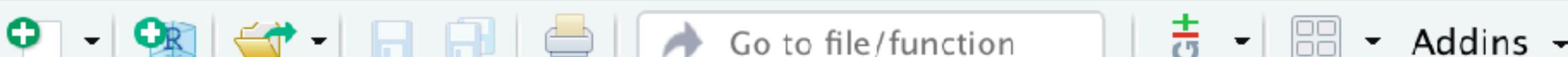
```
set function (name, value)
```

```
Files Plots Packages Help Viewer
```

```
New Folder Delete Rename More
```

```
Home Documents courses stathorizons_0820
```

Name	Size	Modified
..		
.gitignore	40 B	Aug 3, 2020, 10:00 AM
.Rhistory	579 B	Aug 3, 2020, 11:50 AM
01_introduction.Rmd	4 KB	Aug 3, 2020, 11:46 AM
02_get_started.Rmd	4 KB	Aug 3, 2020, 11:46 AM
03_make_a_plot.Rmd	5.8 KB	Aug 3, 2020, 11:46 AM
04_show_the_right_numbers.Rmd	5 KB	Aug 3, 2020, 11:46 AM
05_tables_and_labels.Rmd	10.9 KB	Aug 3, 2020, 11:46 AM
06_models.Rmd	14.7 KB	Aug 3, 2020, 11:46 AM
07_maps.Rmd	12.5 KB	Aug 3, 2020, 11:46 AM
08_refine_plots.Rmd	21.7 KB	Aug 3, 2020, 11:46 AM
09_supplementary_material.Rmd	17 KB	Aug 3, 2020, 11:46 AM
assets		
data		
figures		
keynote		
LICENSE.md	18.1 KB	Jul 21, 2020, 11:16 AM
materials		
R		
README.md	5.4 KB	Aug 3, 2020, 11:46 AM
slides		
stathorizons_0820.Rproj	205 B	Aug 3, 2020, 11:50 AM



01_introduction.Rmd



```
1 ---  
2 title: "Data Visualization"  
3 author: "Kieran Healy"  
4 date: "10-January-2020"  
5 output: html_document  
6 ---  
7  
8 ## Data Visualization Notes  
9  
10 This is a starter RMarkdown project template to accompany courses taught with  
*Data Visualization*. You can use it to take notes, write your code, and  
produce a good-looking, reproducible document that records the work you have  
done. At the very top of the file is a section of *metadata*, or information  
about what the file is and what it does. The metadata is delimited by three  
dashes at the start and another three at the end. You should change the title,  
author, and date to the values that suit you. Keep the 'output' line as it is  
for now, however. Each line in the metadata has a structure. First the *key*  
("title", "author", etc), then a colon, and then the *value* associated with  
the key.
```



Environment History

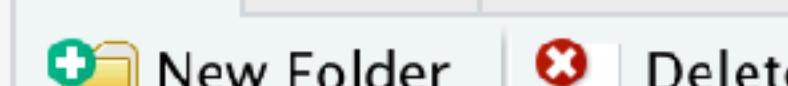


Global Environment

Functions

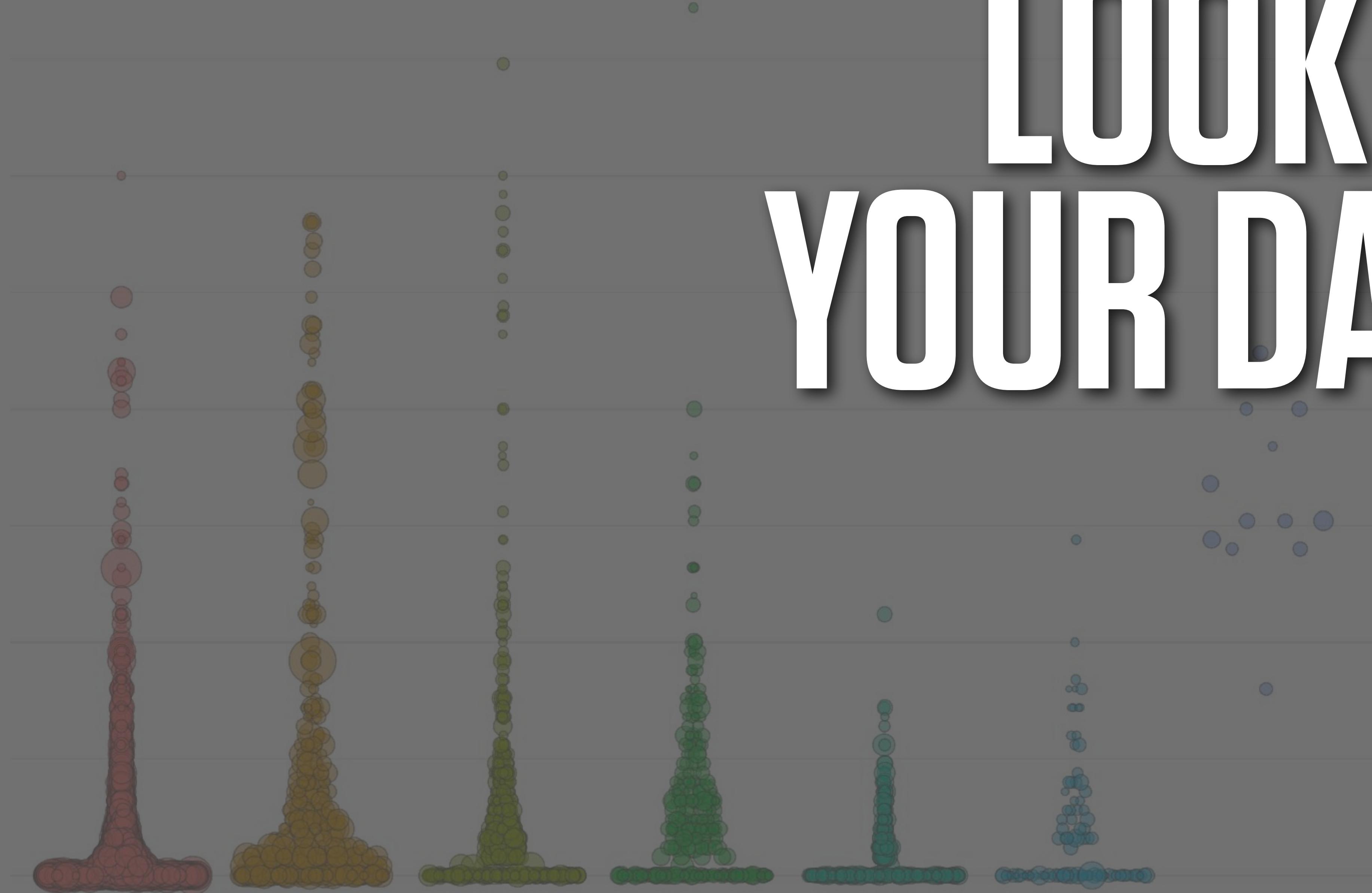
set

Files Plots Packages

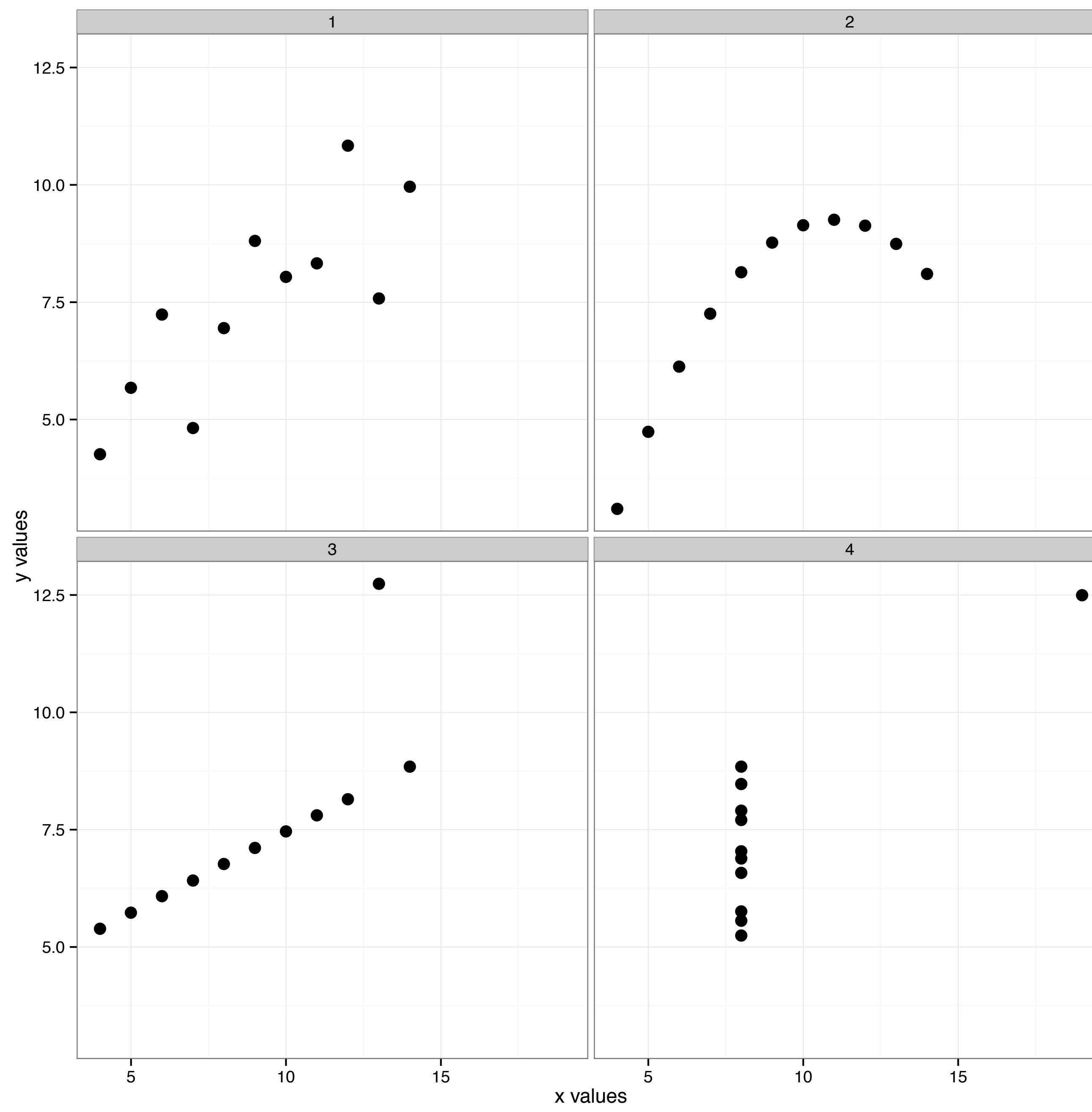


Home > Documents

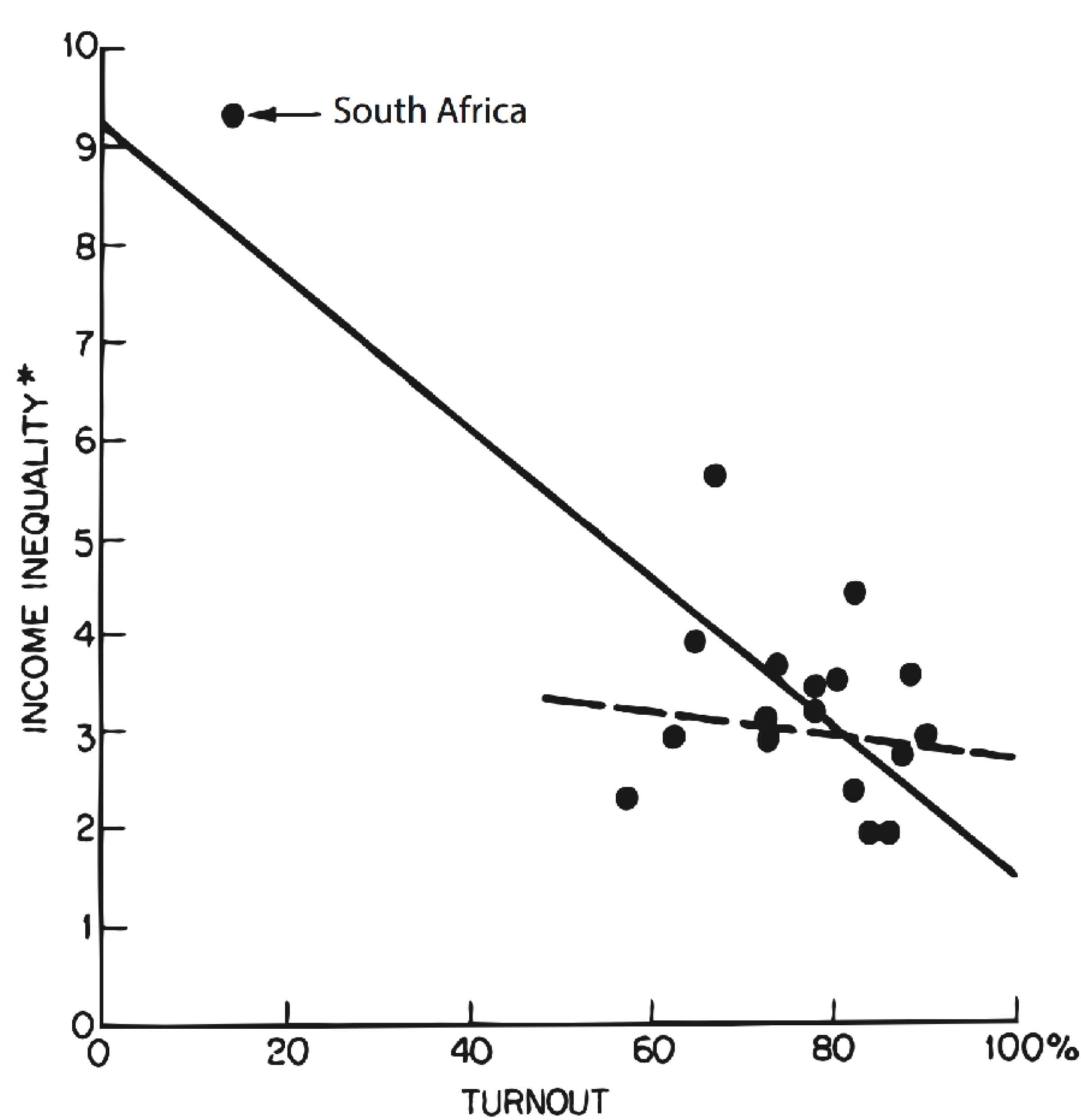
LOOK AT YOUR DATA



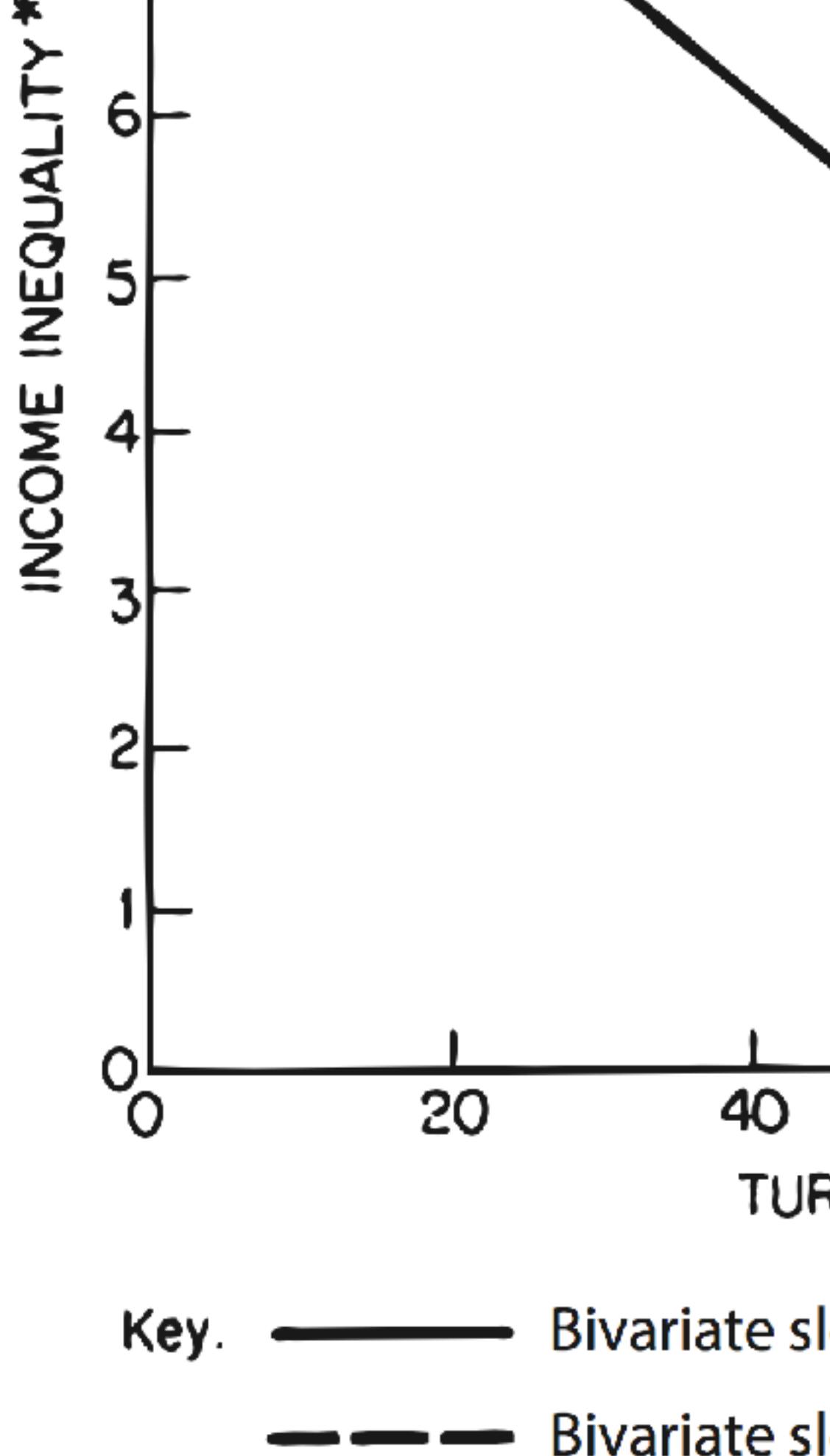
SEEING THINGS

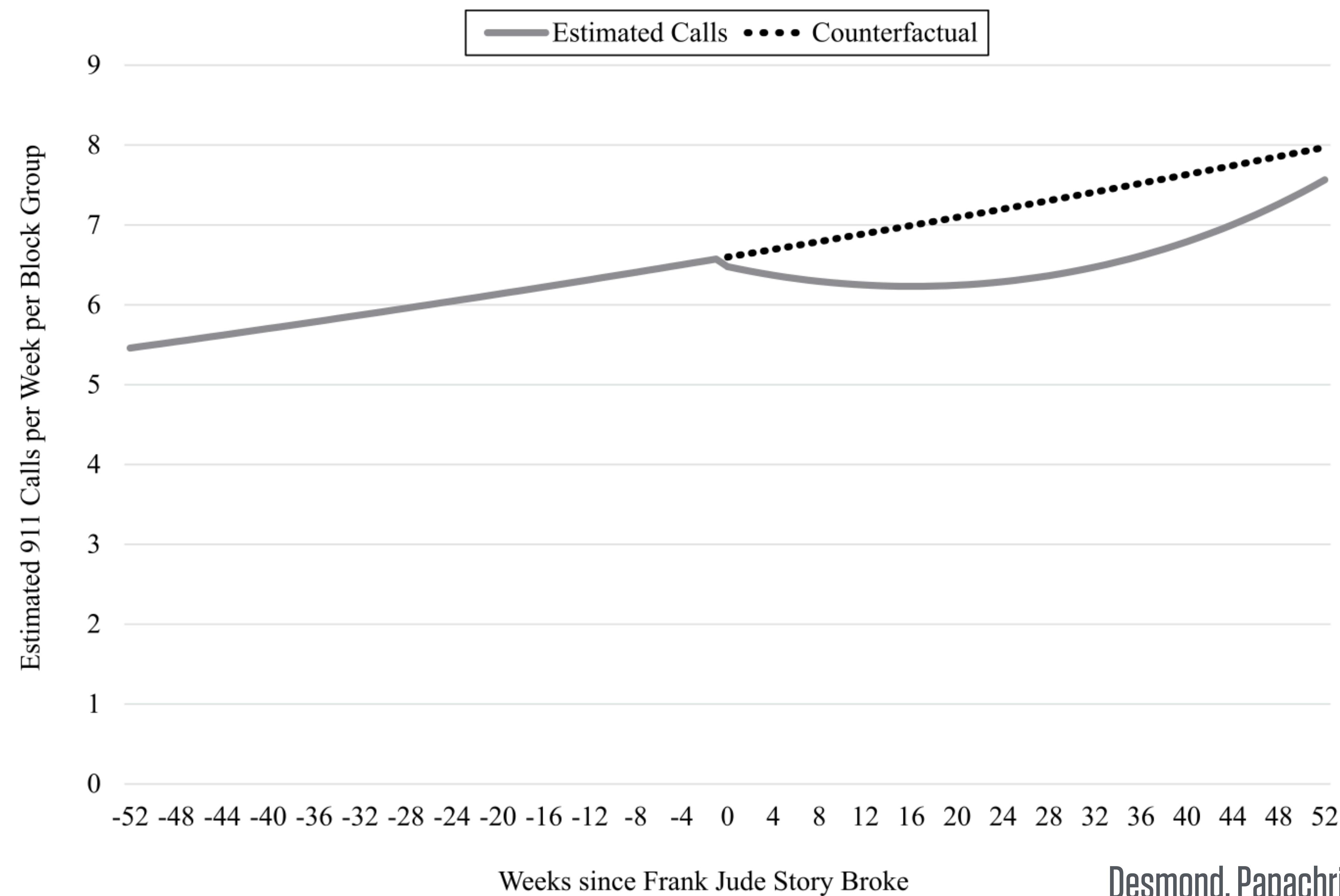


For all panels, N=11; Mean=7.5; Regression: $Y=3 + 0.5(X)$; SE of slope estimate: 0.118, t=4.24; Sum of Squares (X-X): 100; r=0.82.



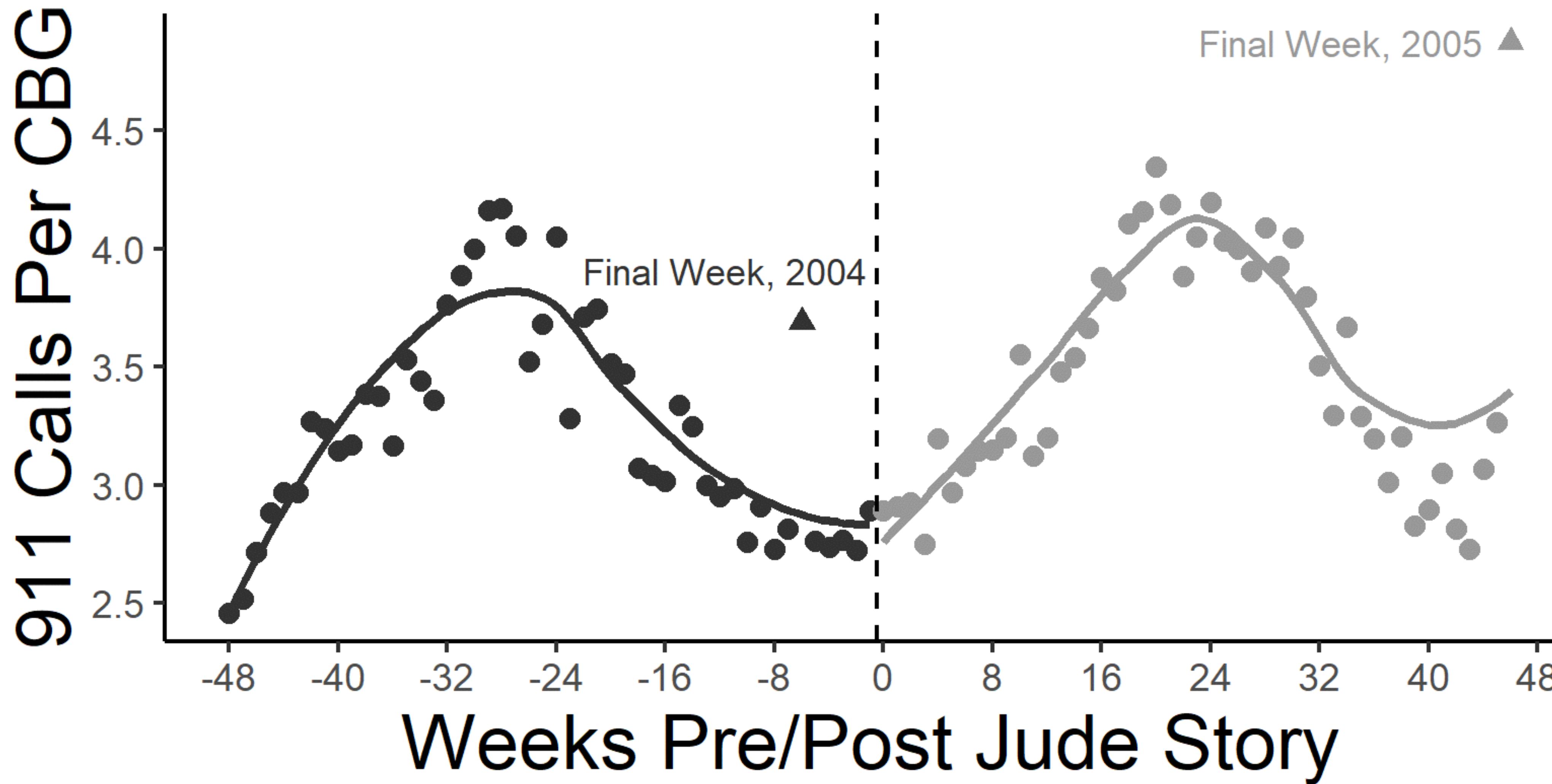
Key. — Bivariate slope including South Africa ($N = 18$)
— Bivariate slope excluding South Africa ($N = 17$)

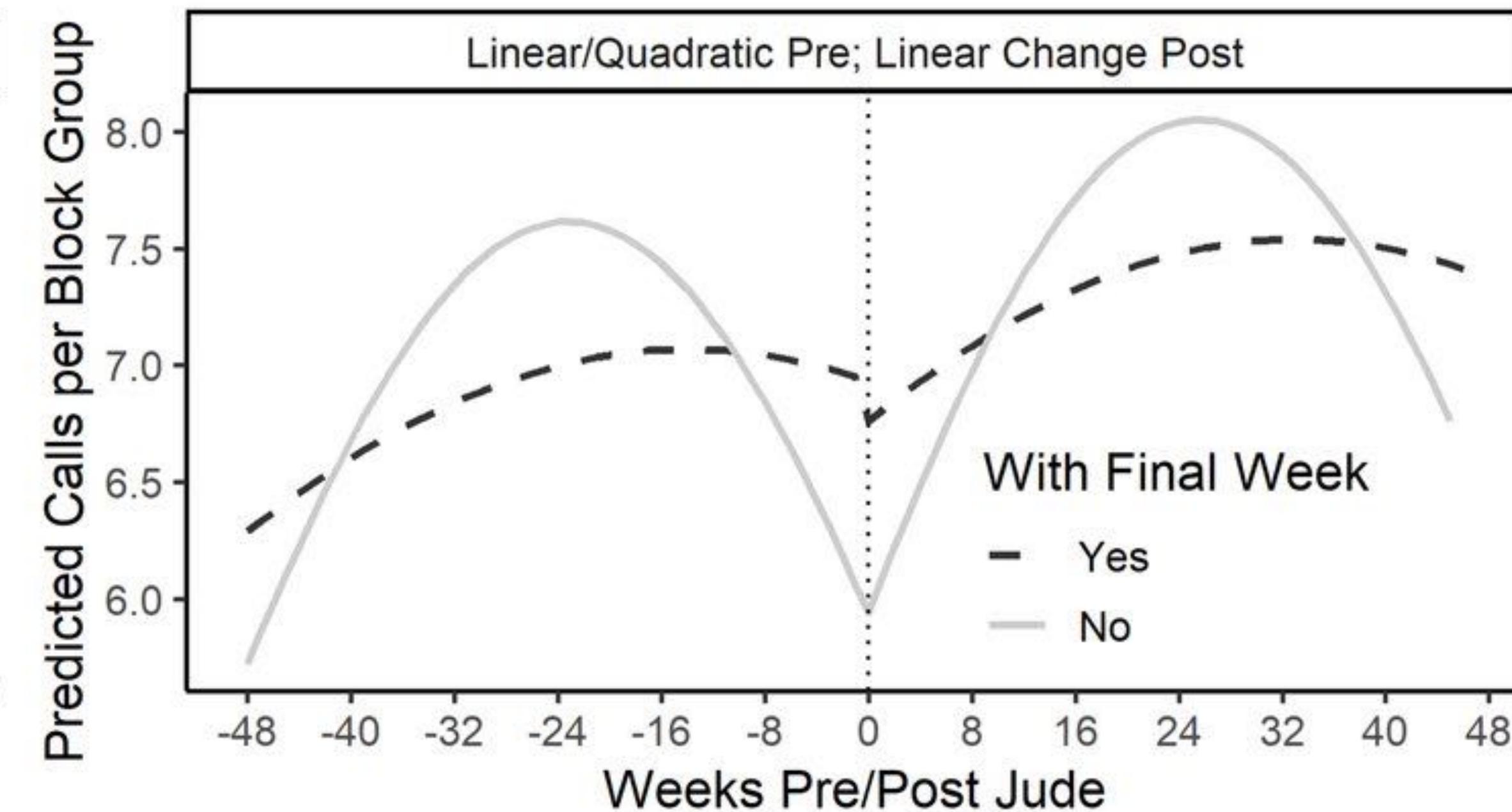
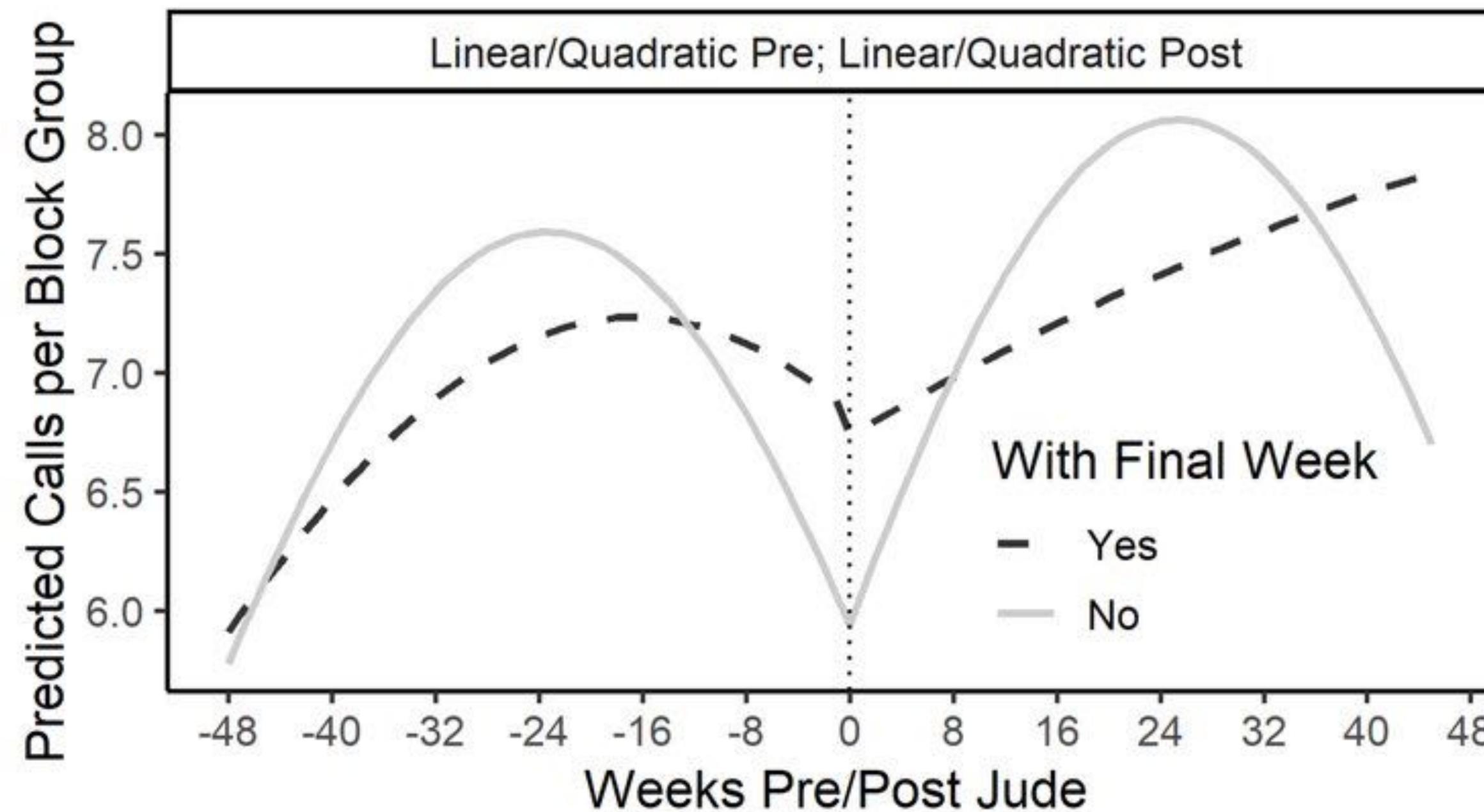
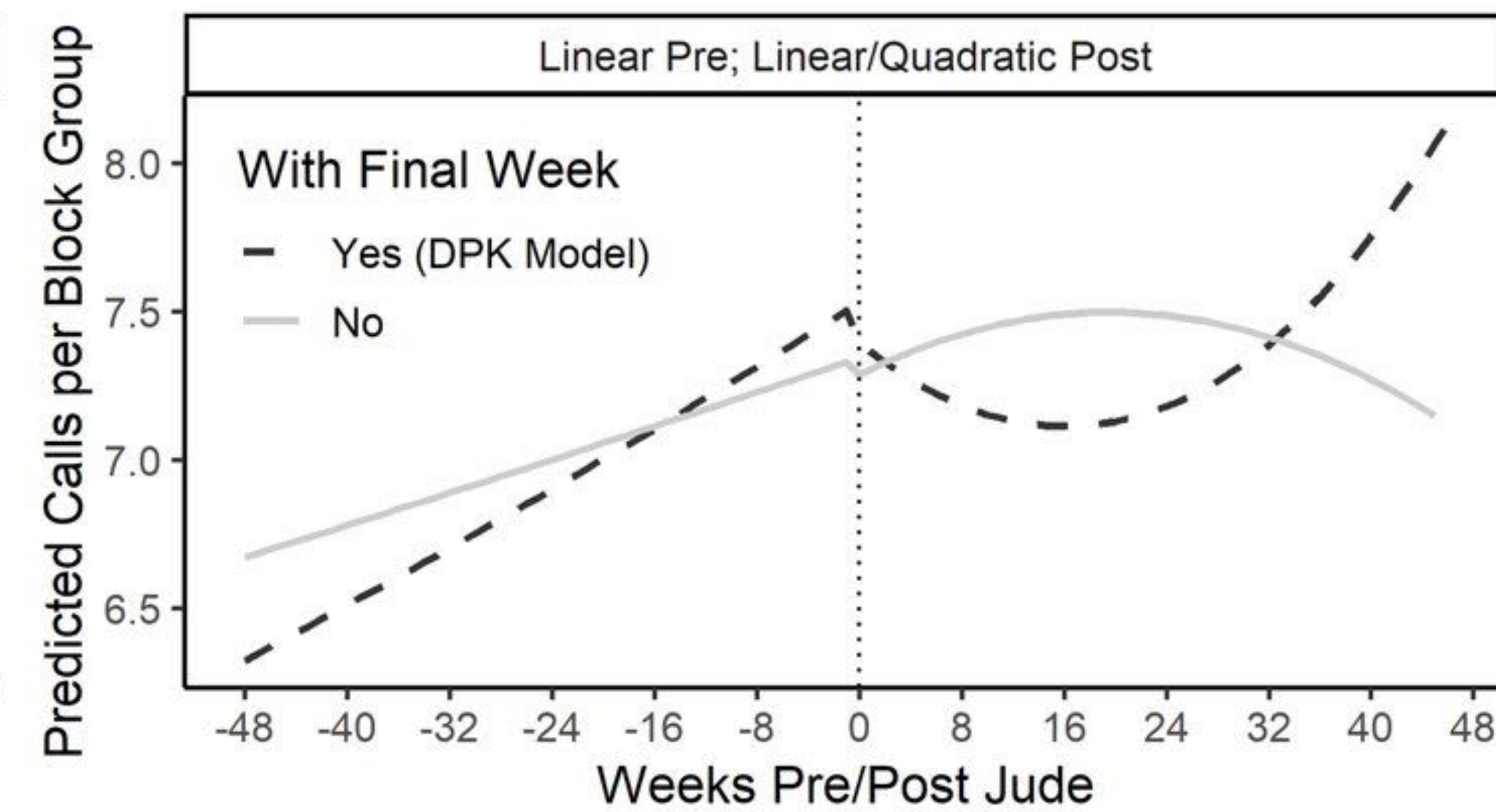
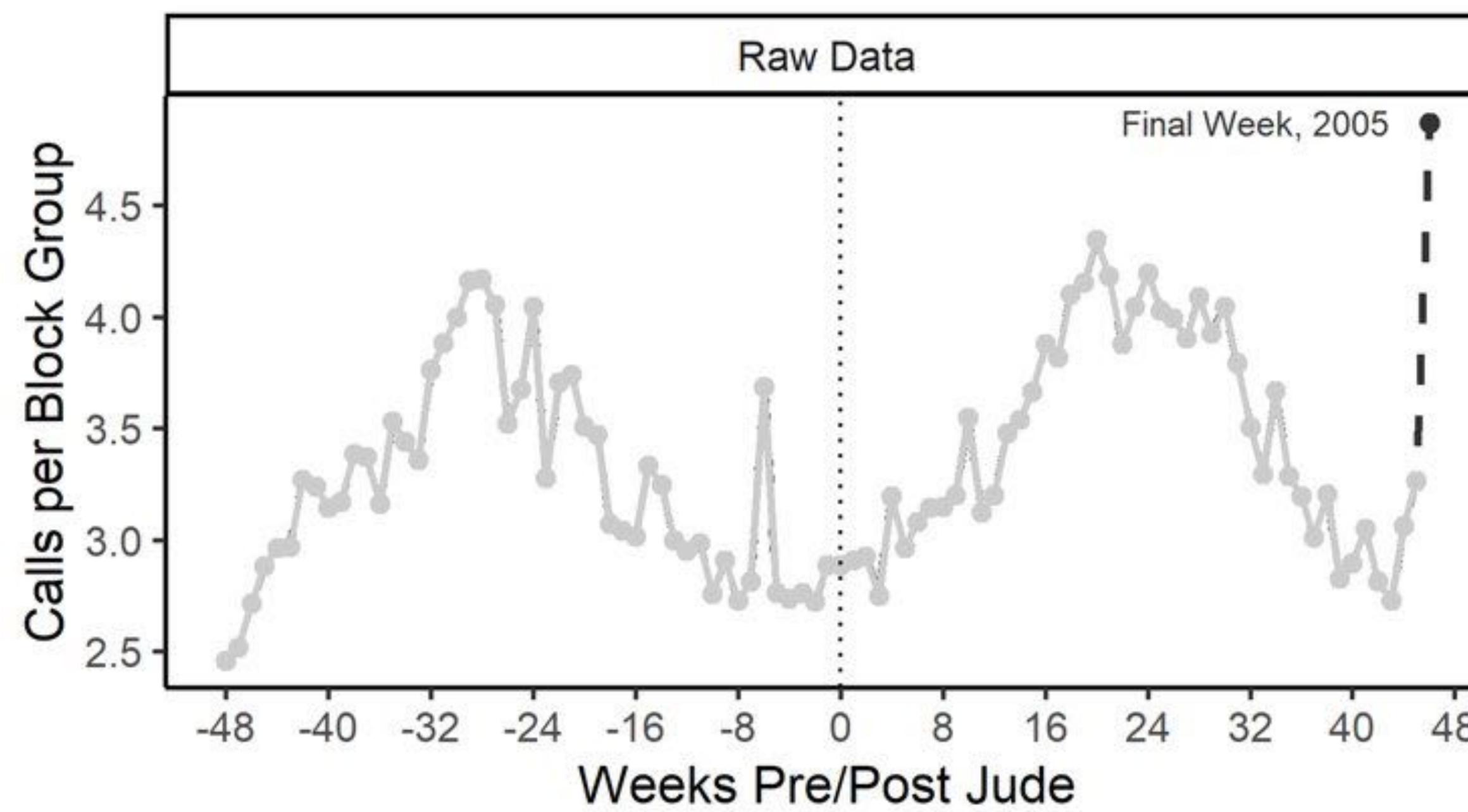




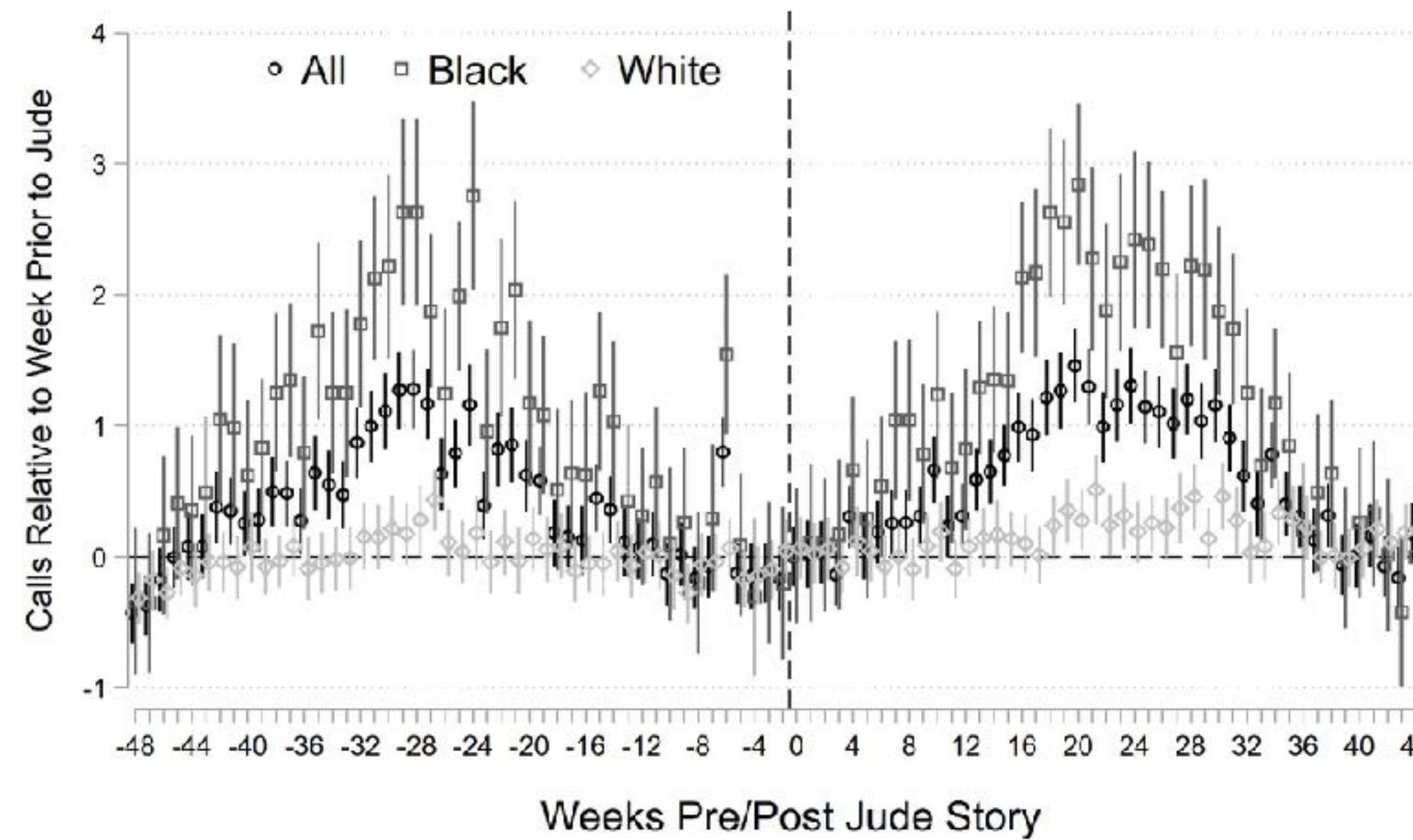
Desmond, Papachristos & Kirk (2016)

All Neighborhoods

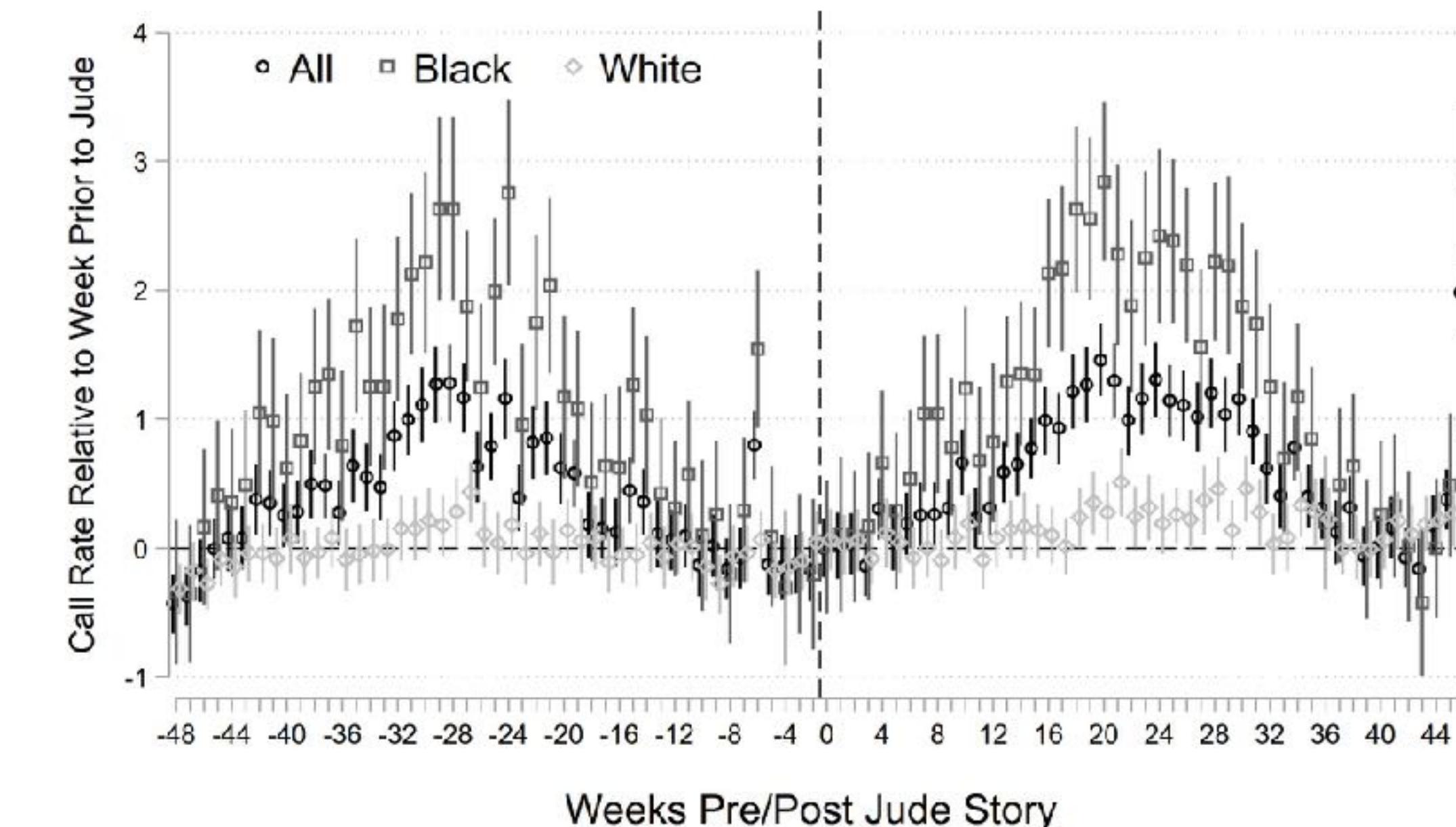




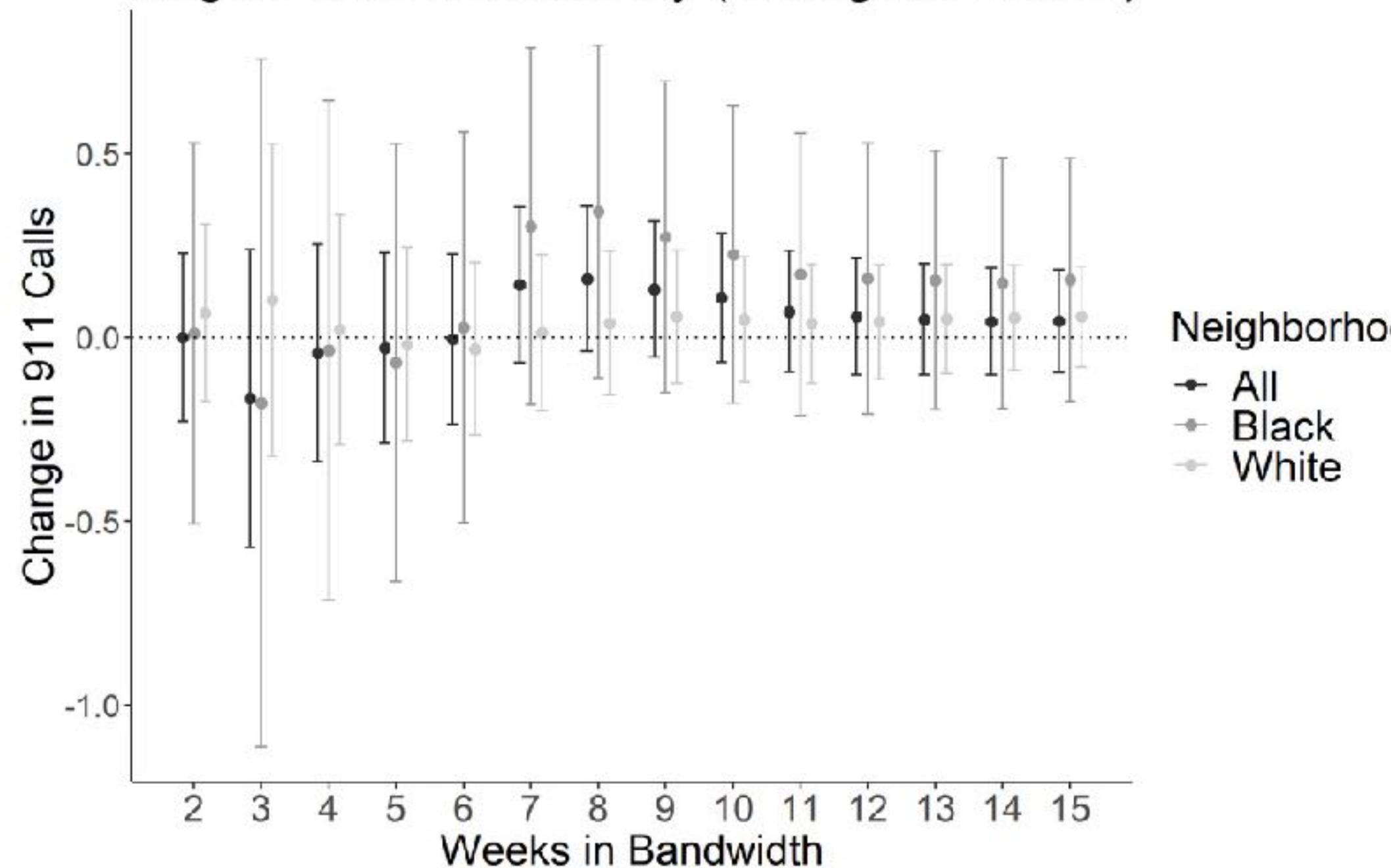
Event Study: Leads/Lags



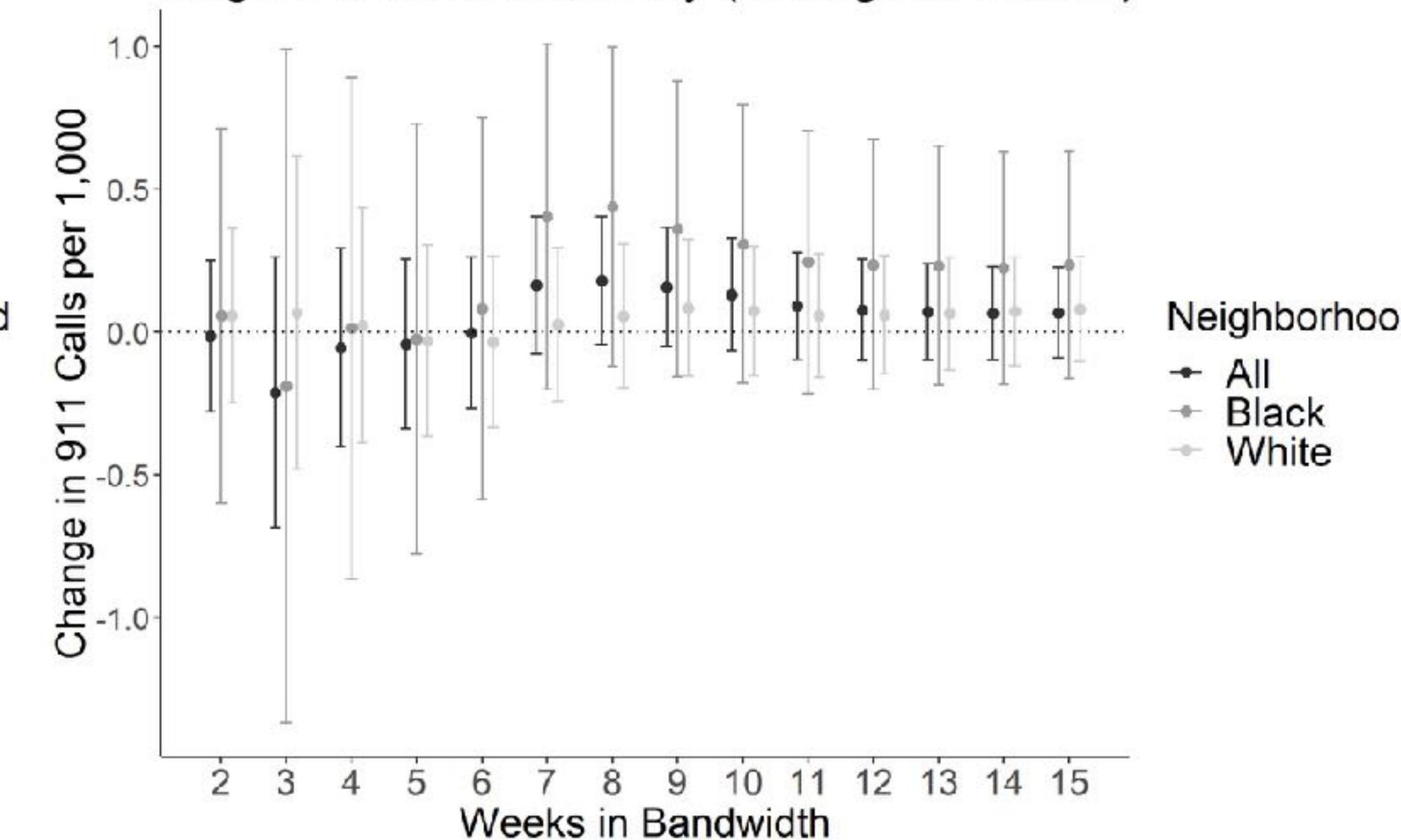
Event Study: Leads/Lags



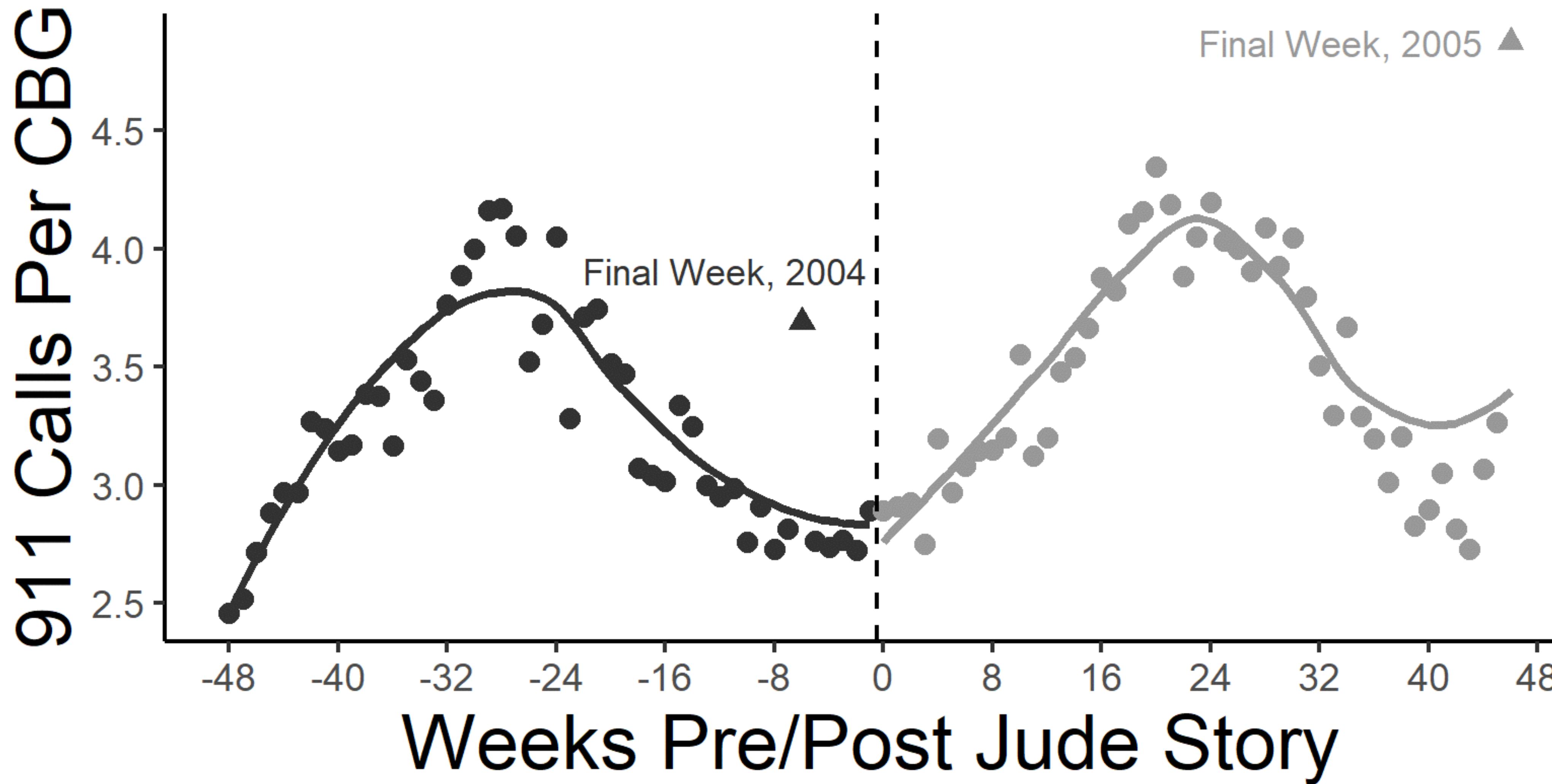
Regression Discontinuity (Triangular Kernel)

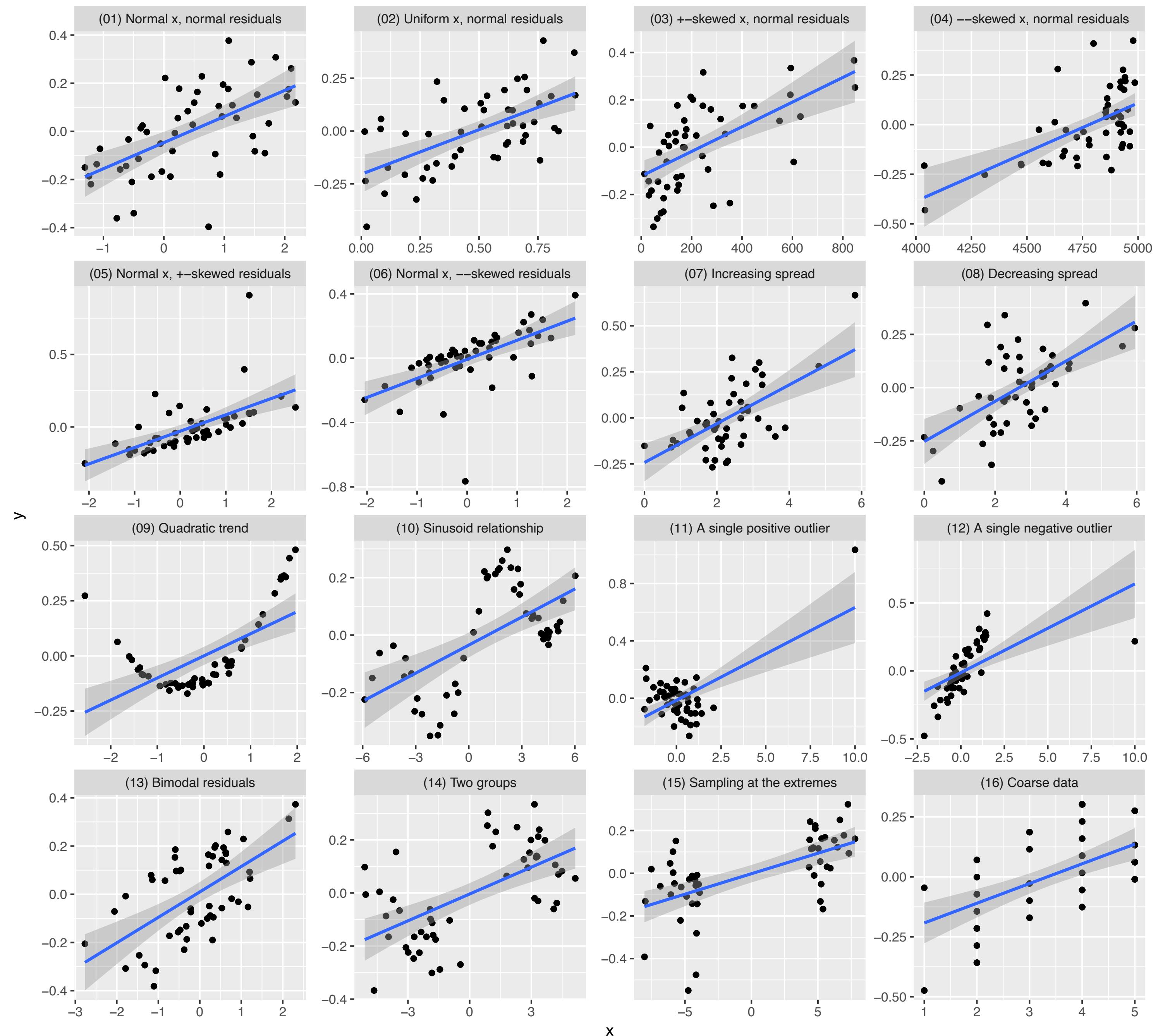


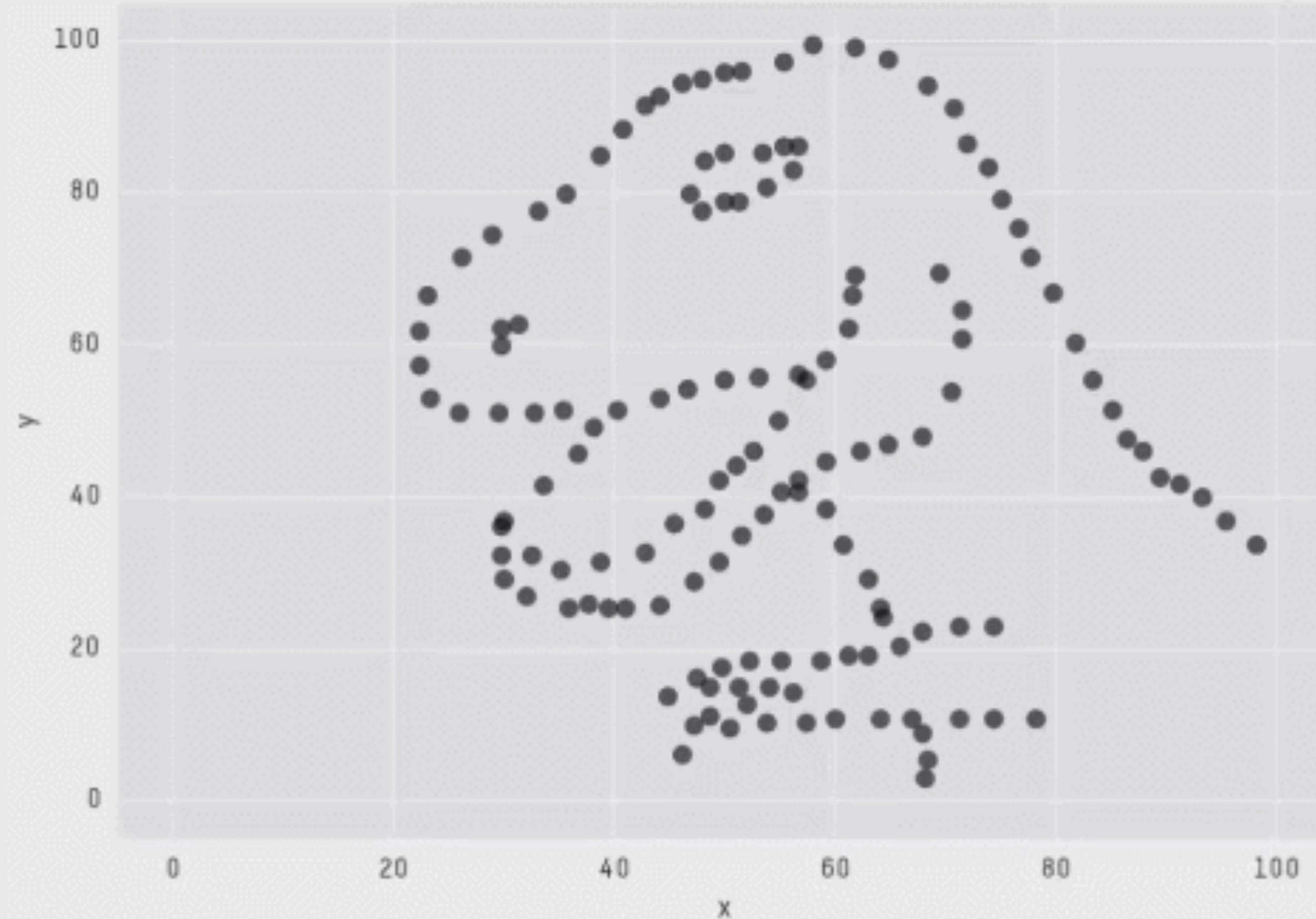
Regression Discontinuity (Triangular Kernel)



All Neighborhoods

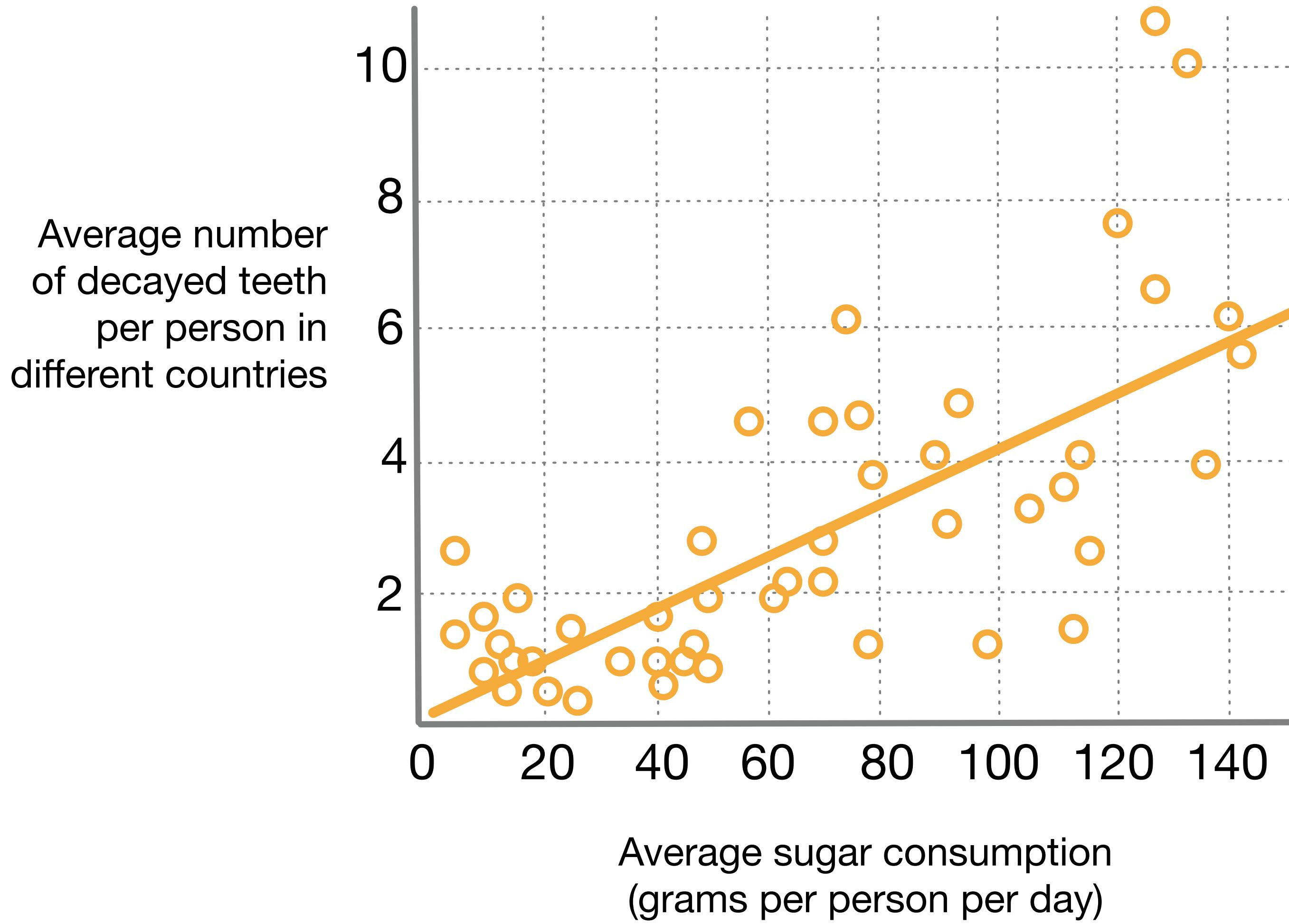




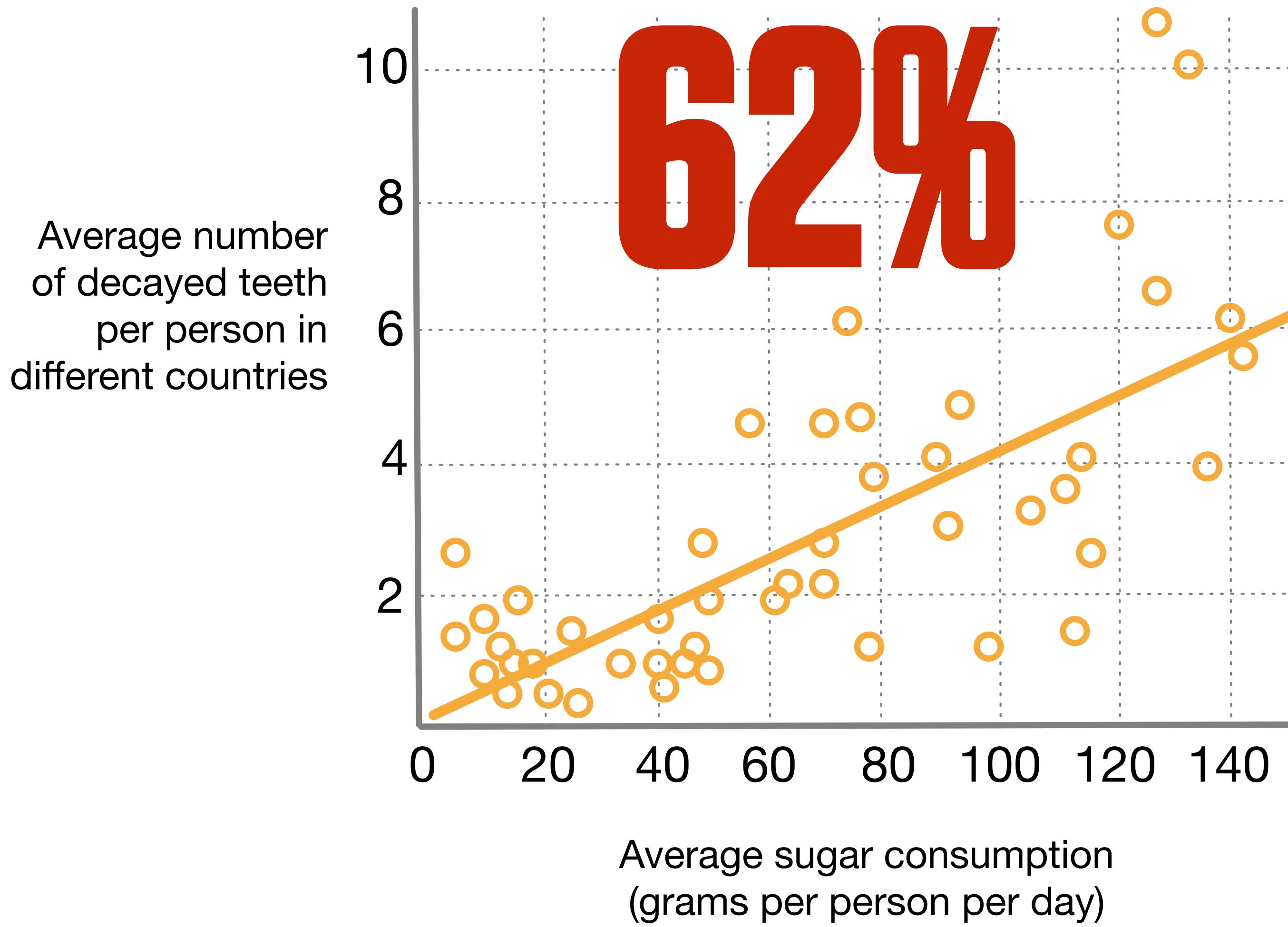


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Which of the following statements best describes the data in the graph below?



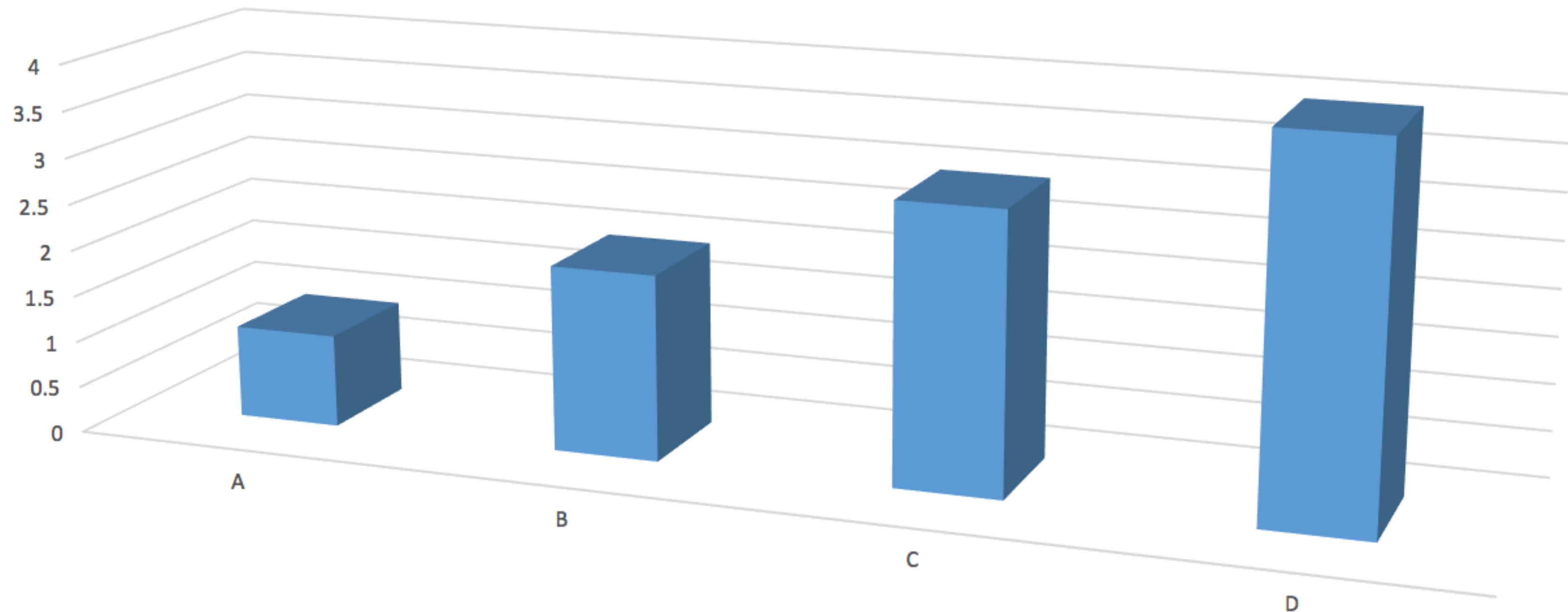
Which of the following statements best describes the data in the graph below?

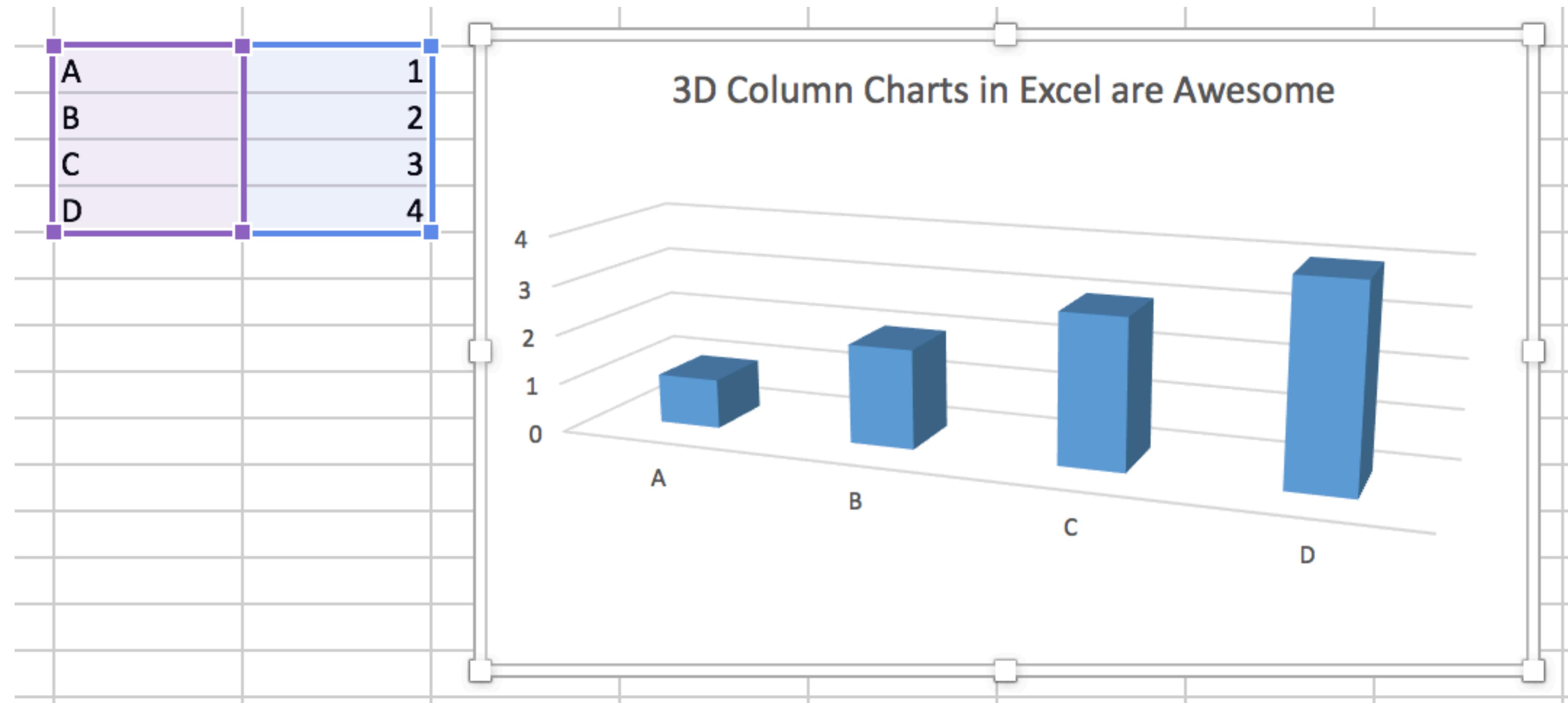


- A. In recent years, the rate of cavities has increased in many countries
- B. In some countries, people brush their teeth more frequently than in other countries
- C. The more sugar people eat, the more likely they are to get cavities**
- D. In recent years, the consumption of sugar has increased in many countries

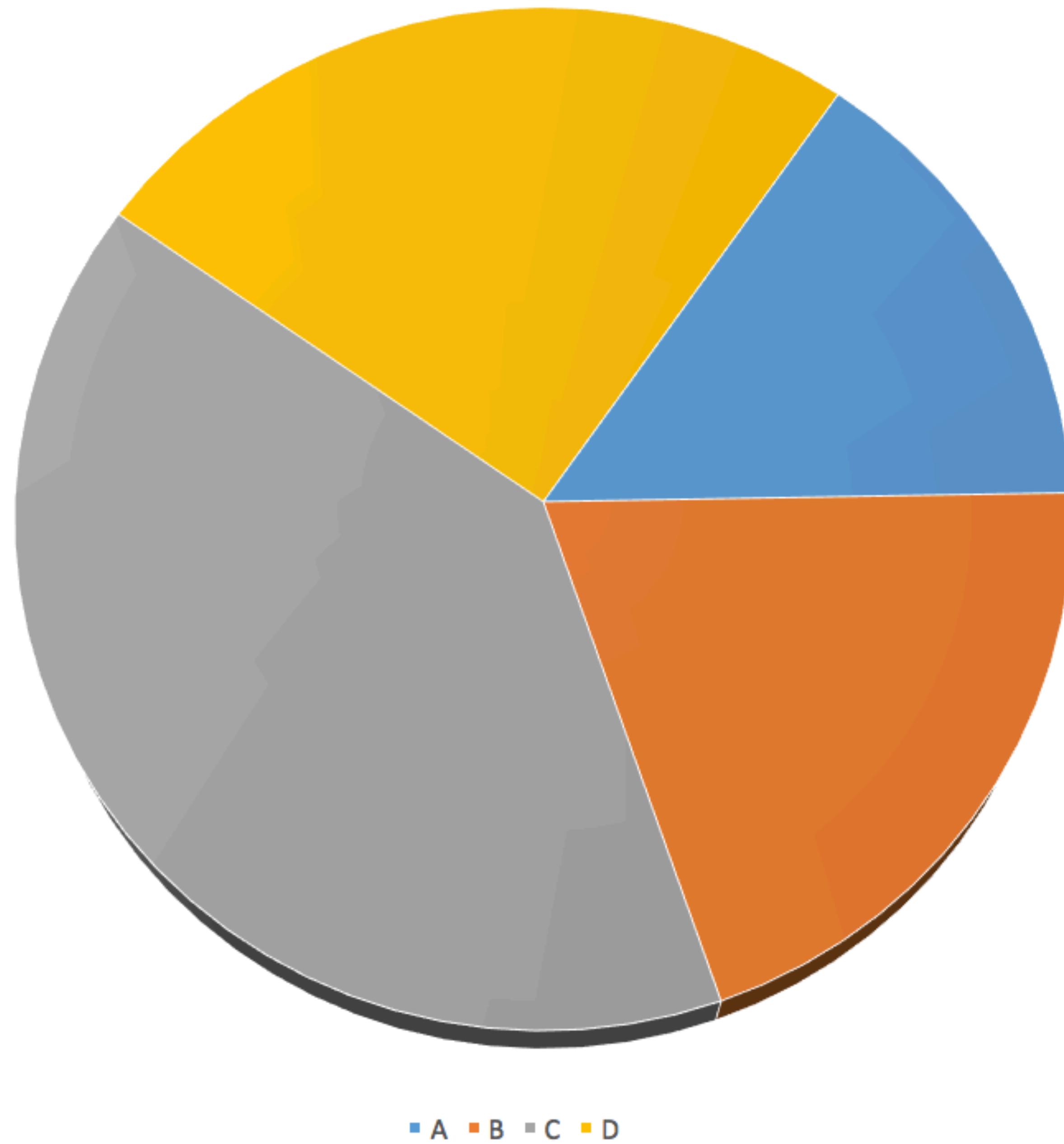
NOT
SEEING THINGS

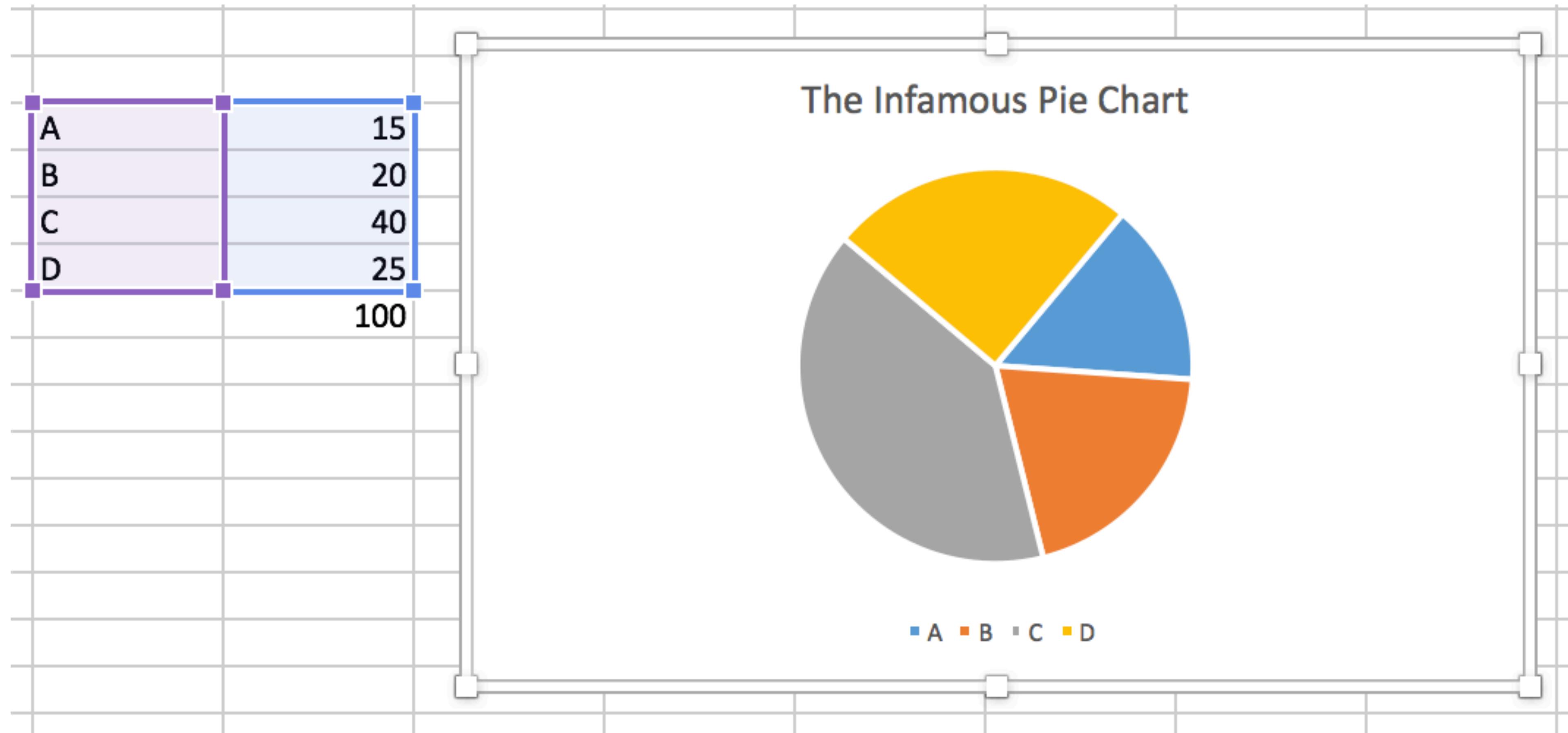
3D Column Charts in Excel are Awesome





The Infamous Pie Chart





BAD TASTE

BAD DATA

BAD PERCEPTION

TASTE

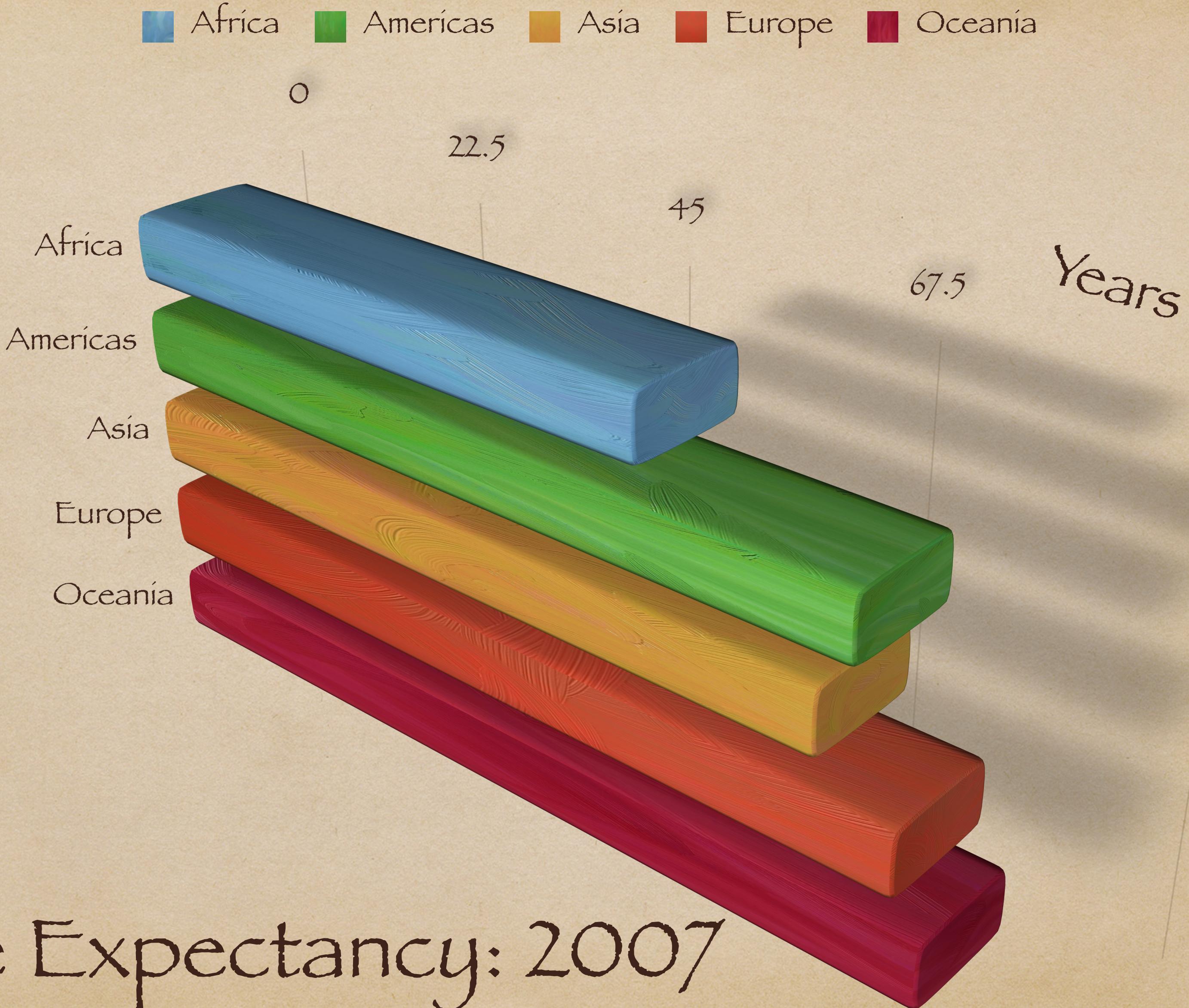
SIMPLIFY, SIMPLIFY?

The Visual Display
of Quantitative Information

EDWARD R. TUFTE

Life Expectancy: 2007

CONTINENT



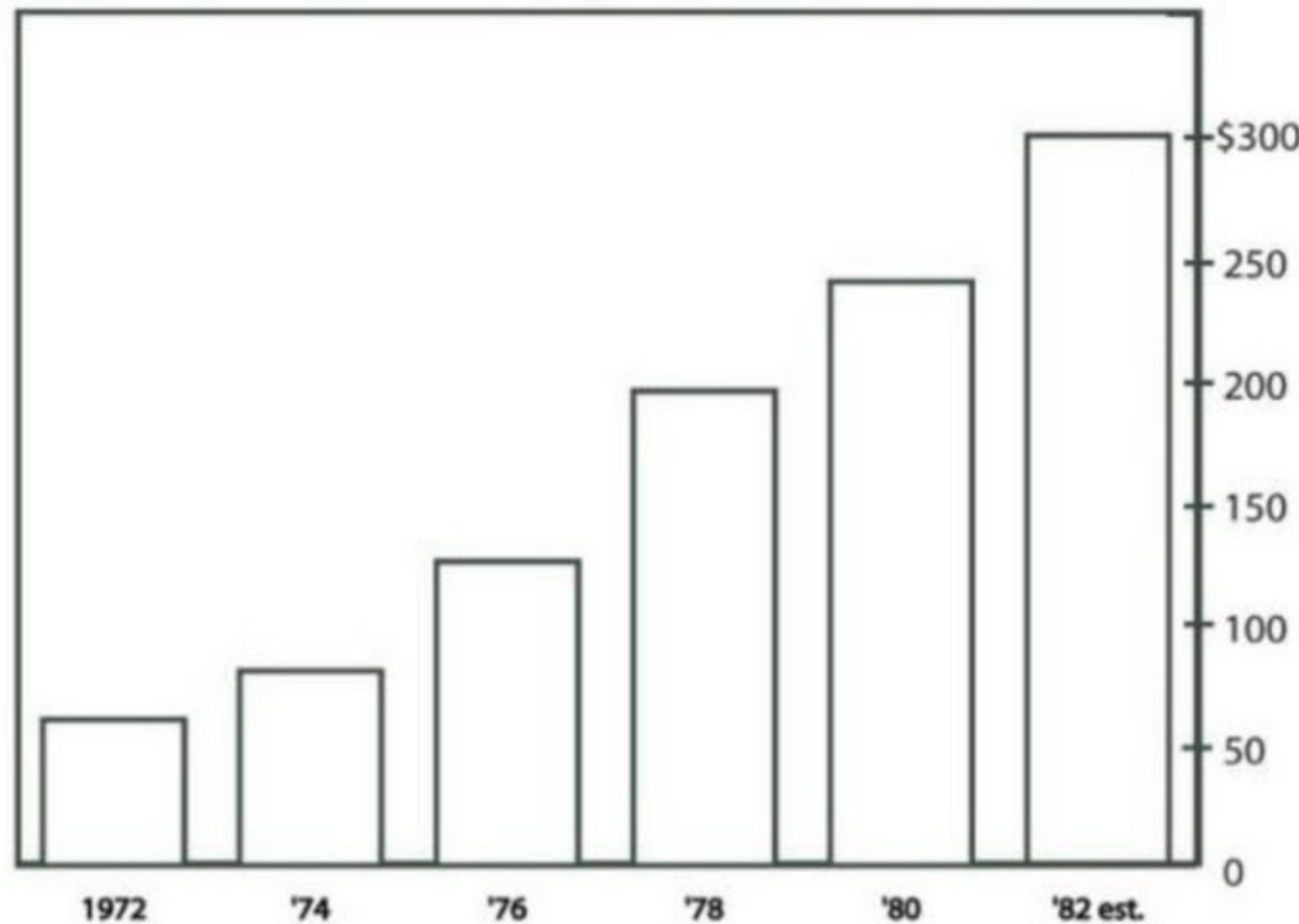
Years

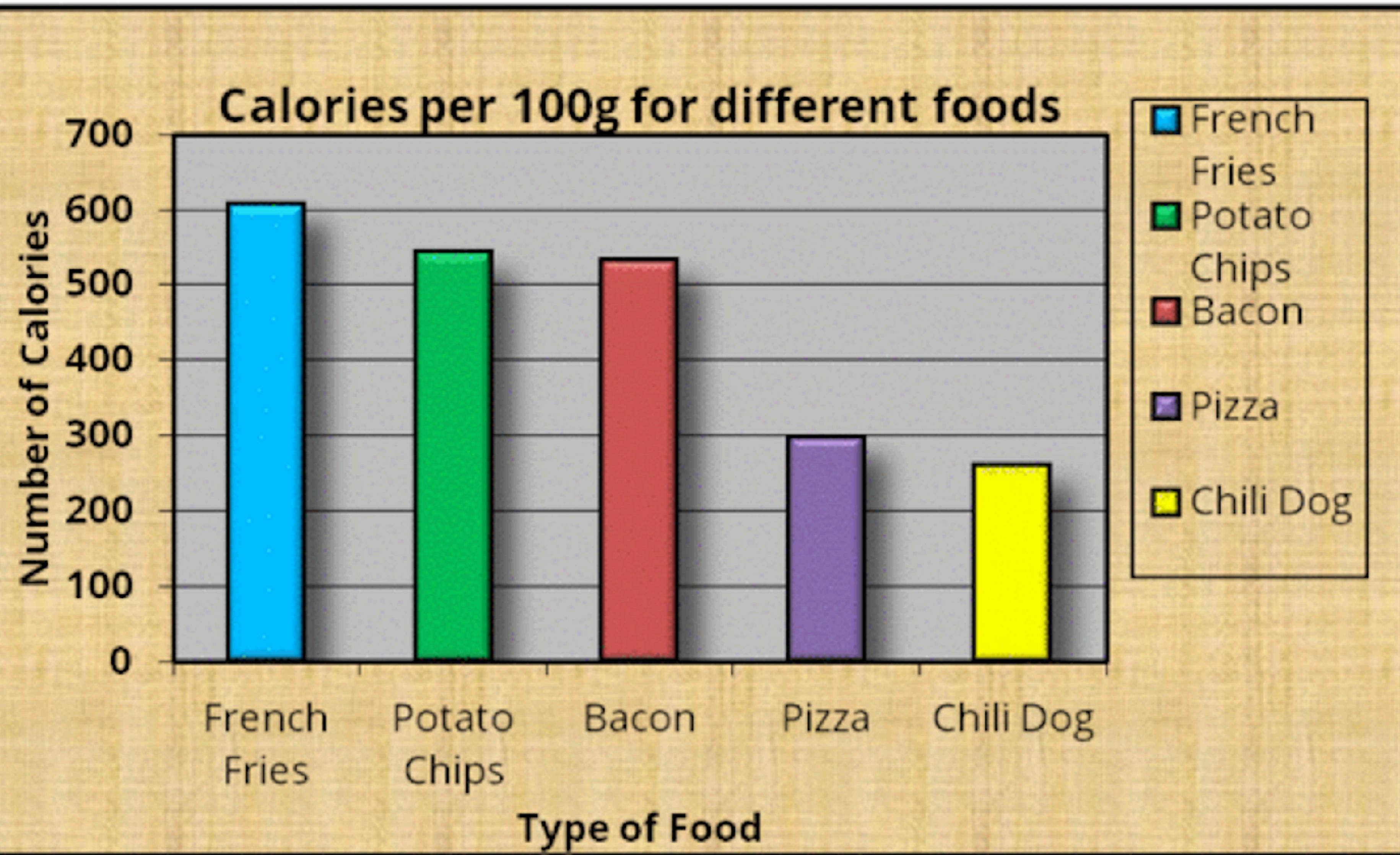
MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions

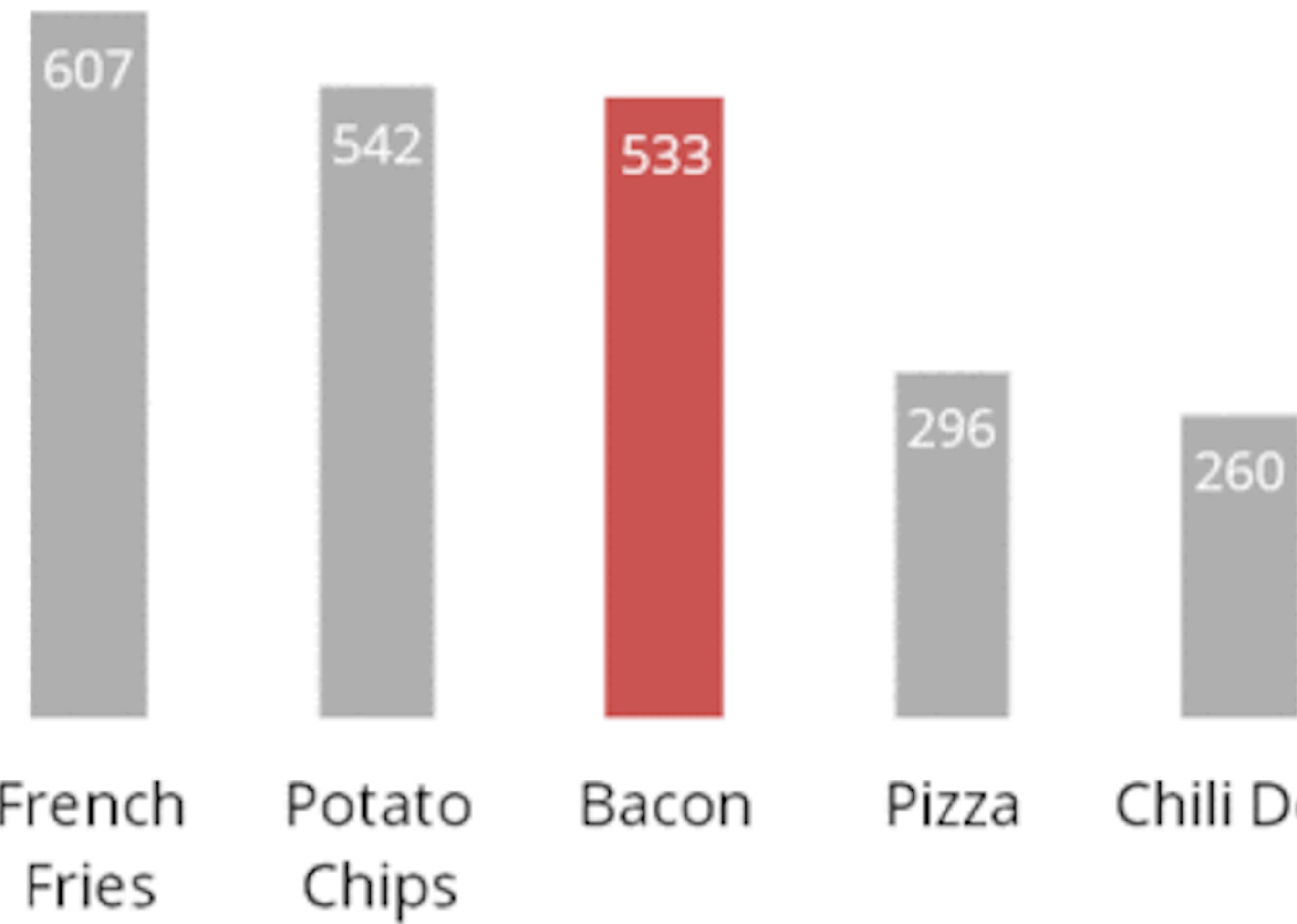


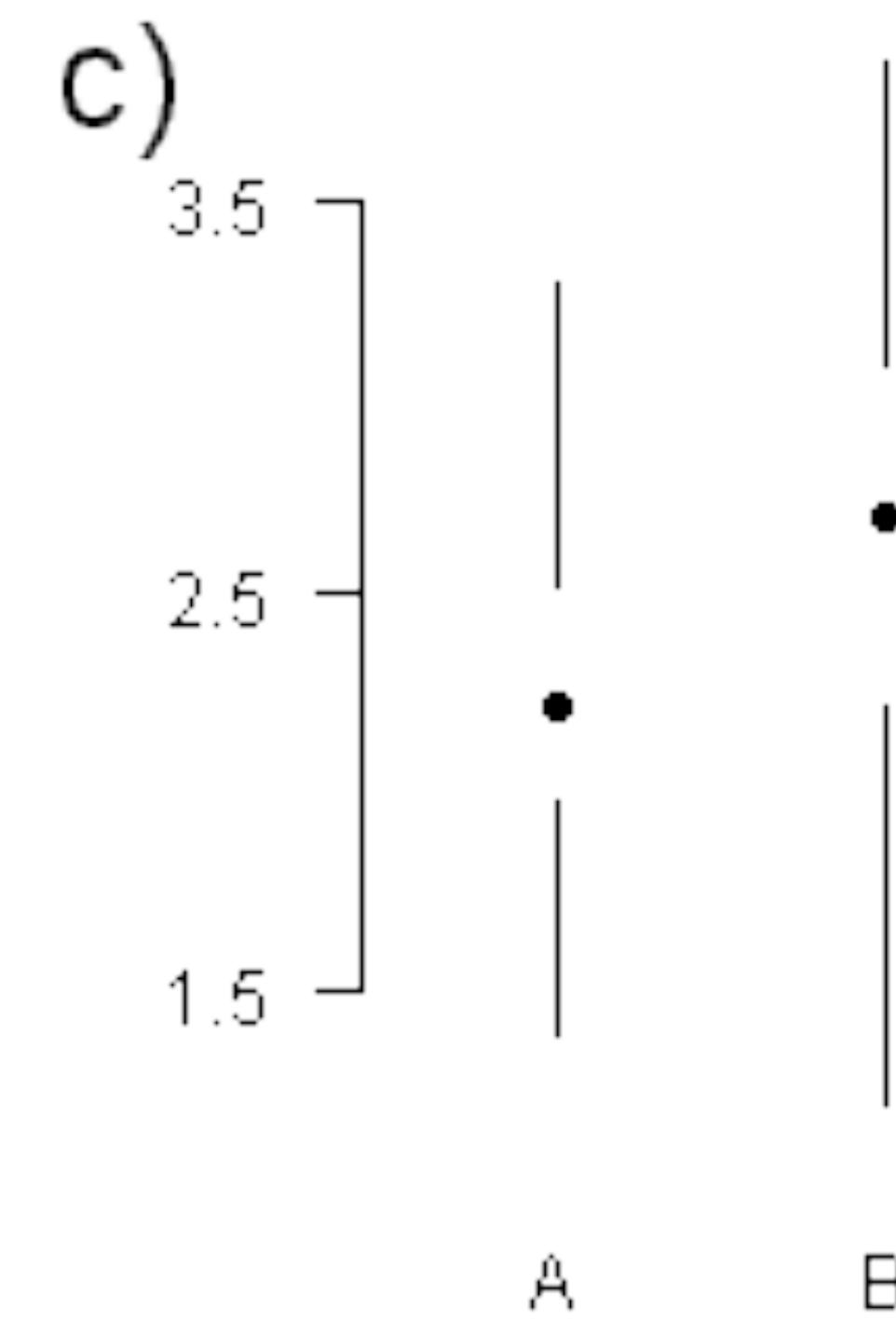
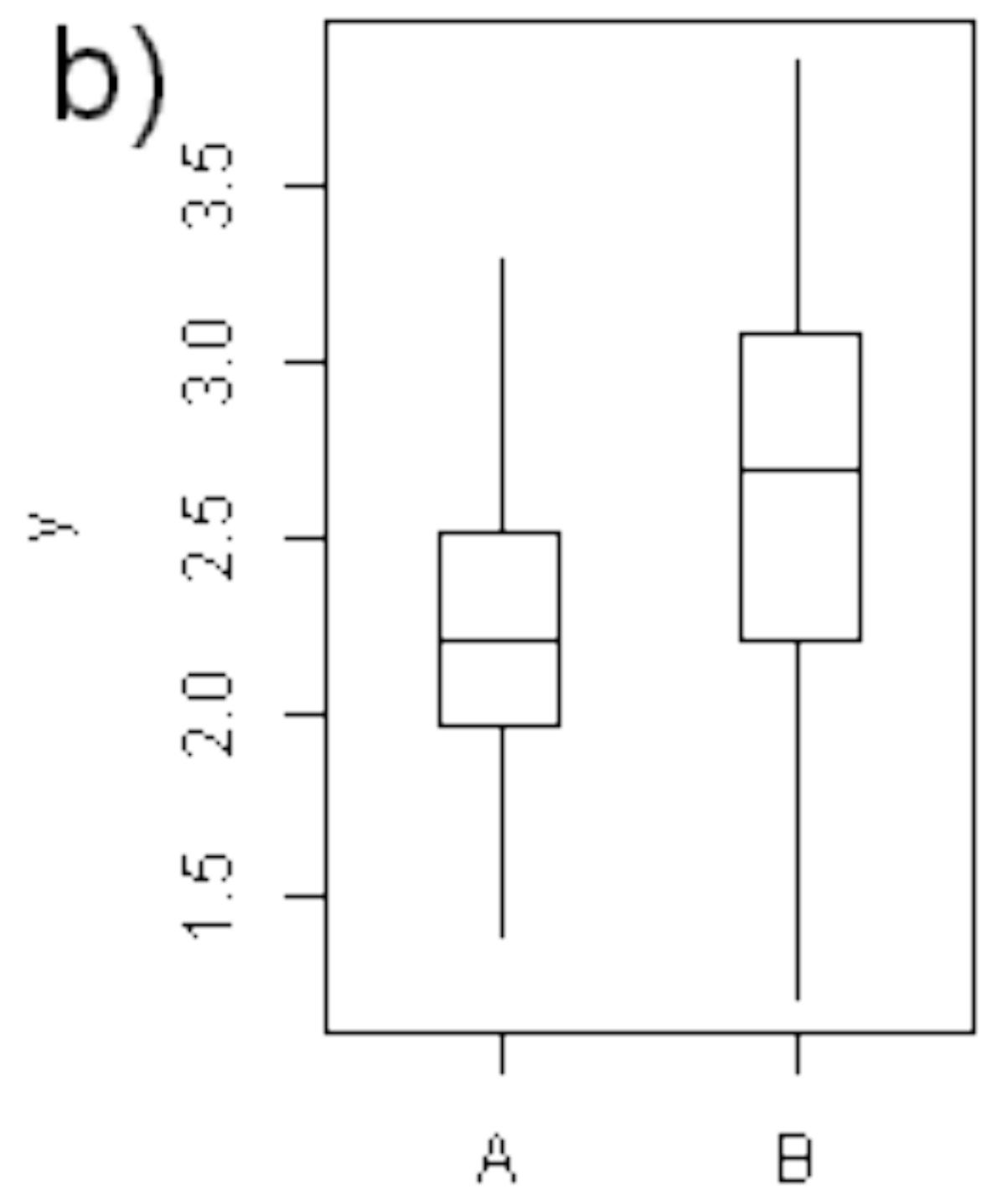
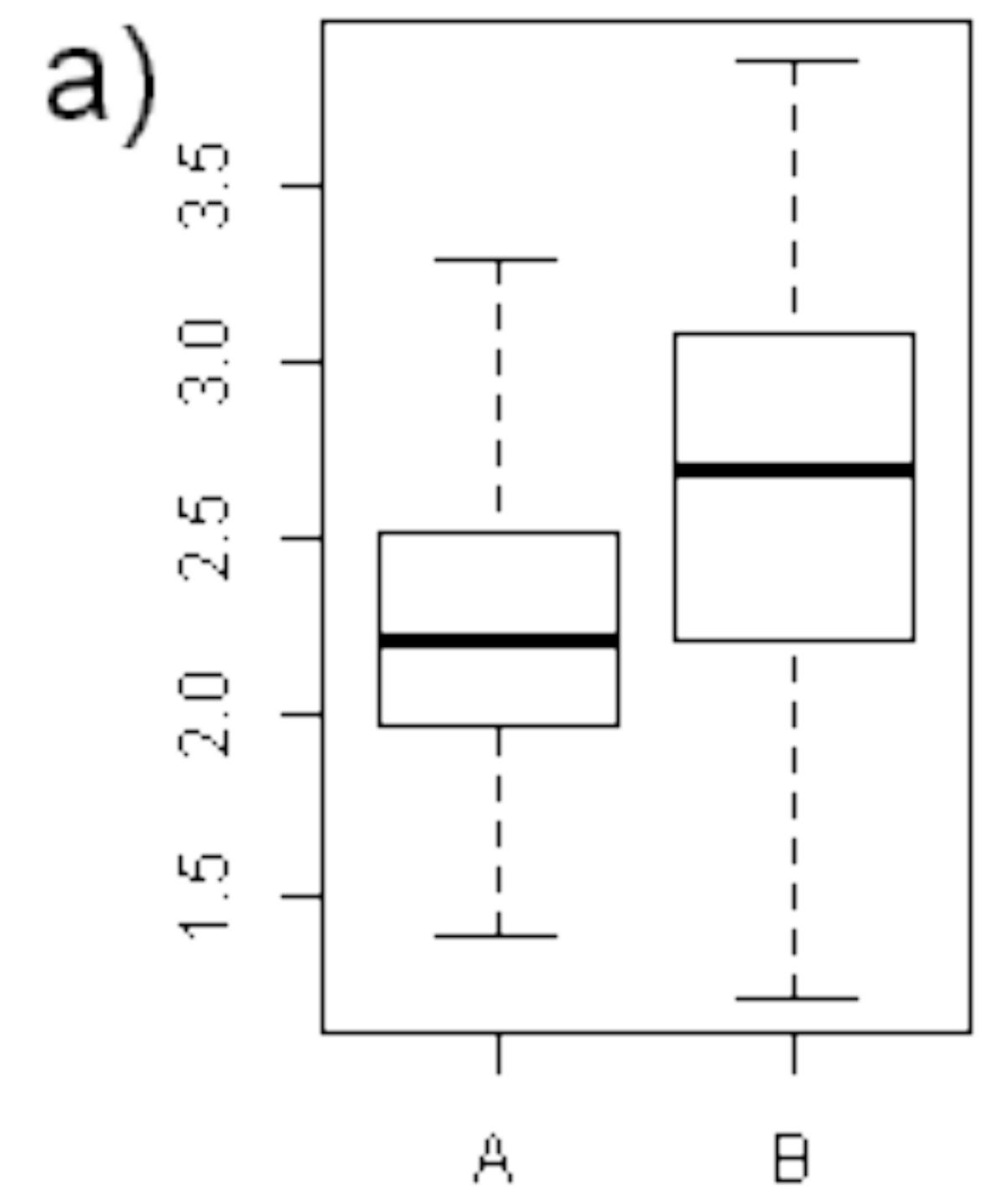
MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions

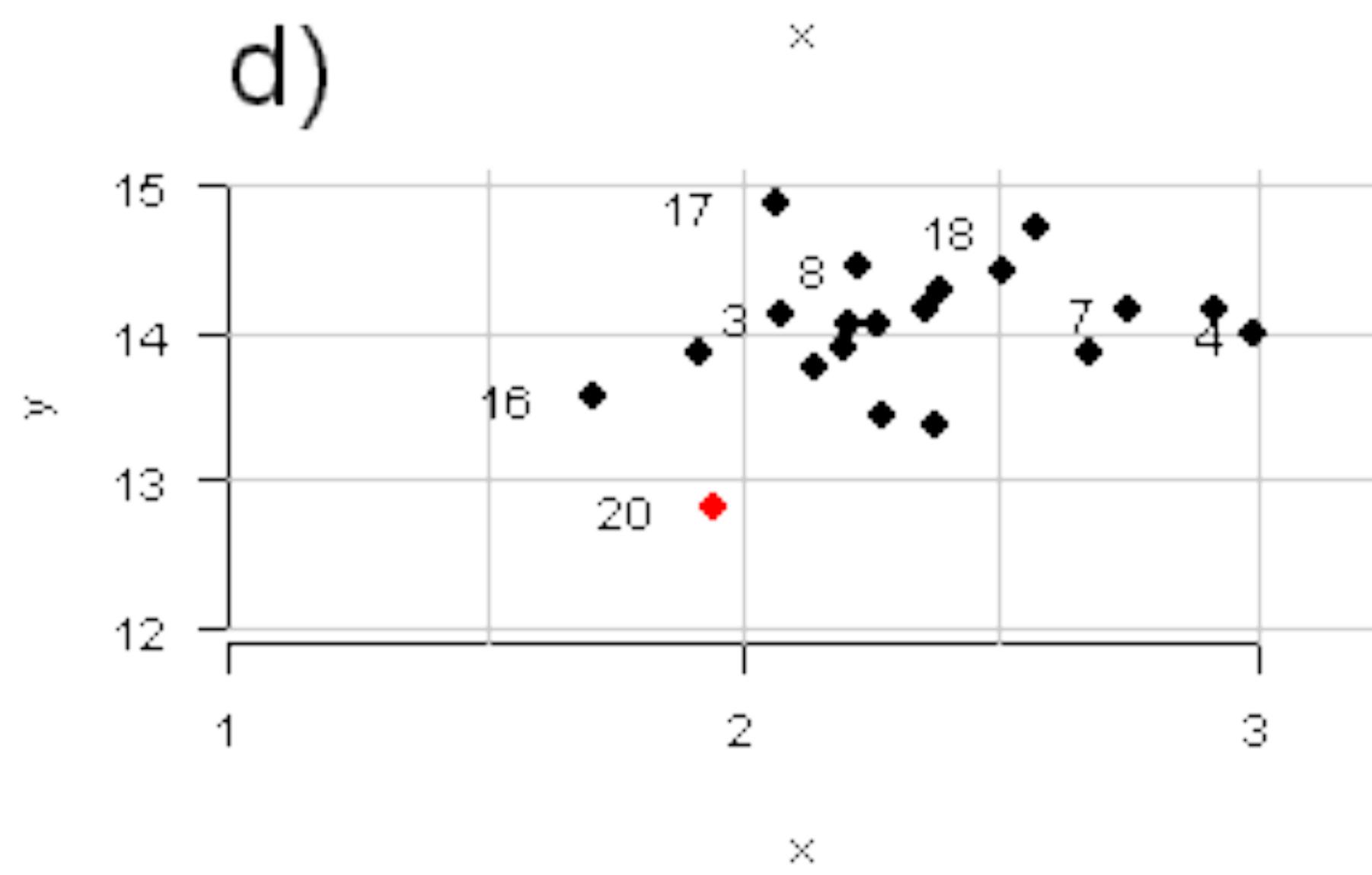
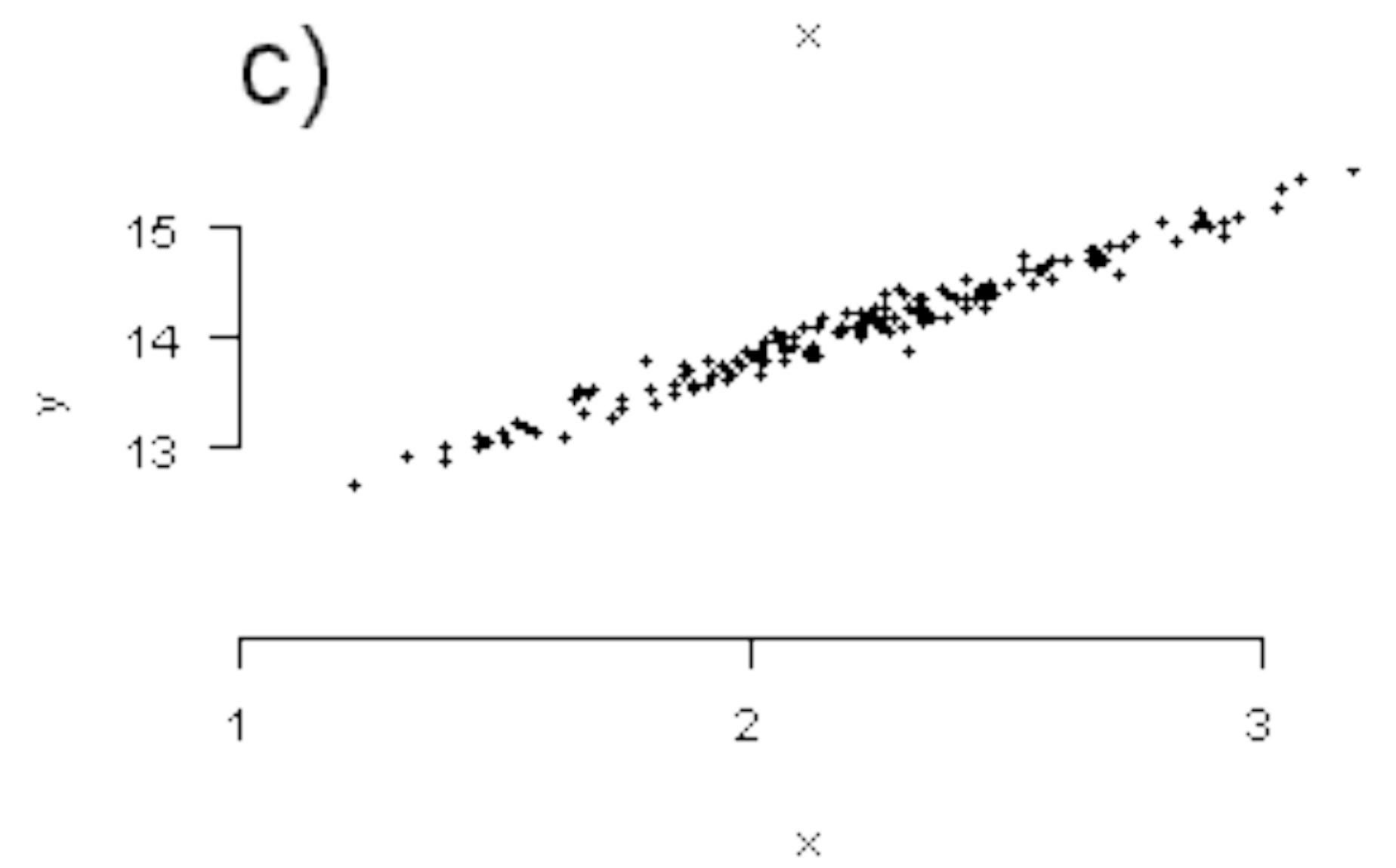
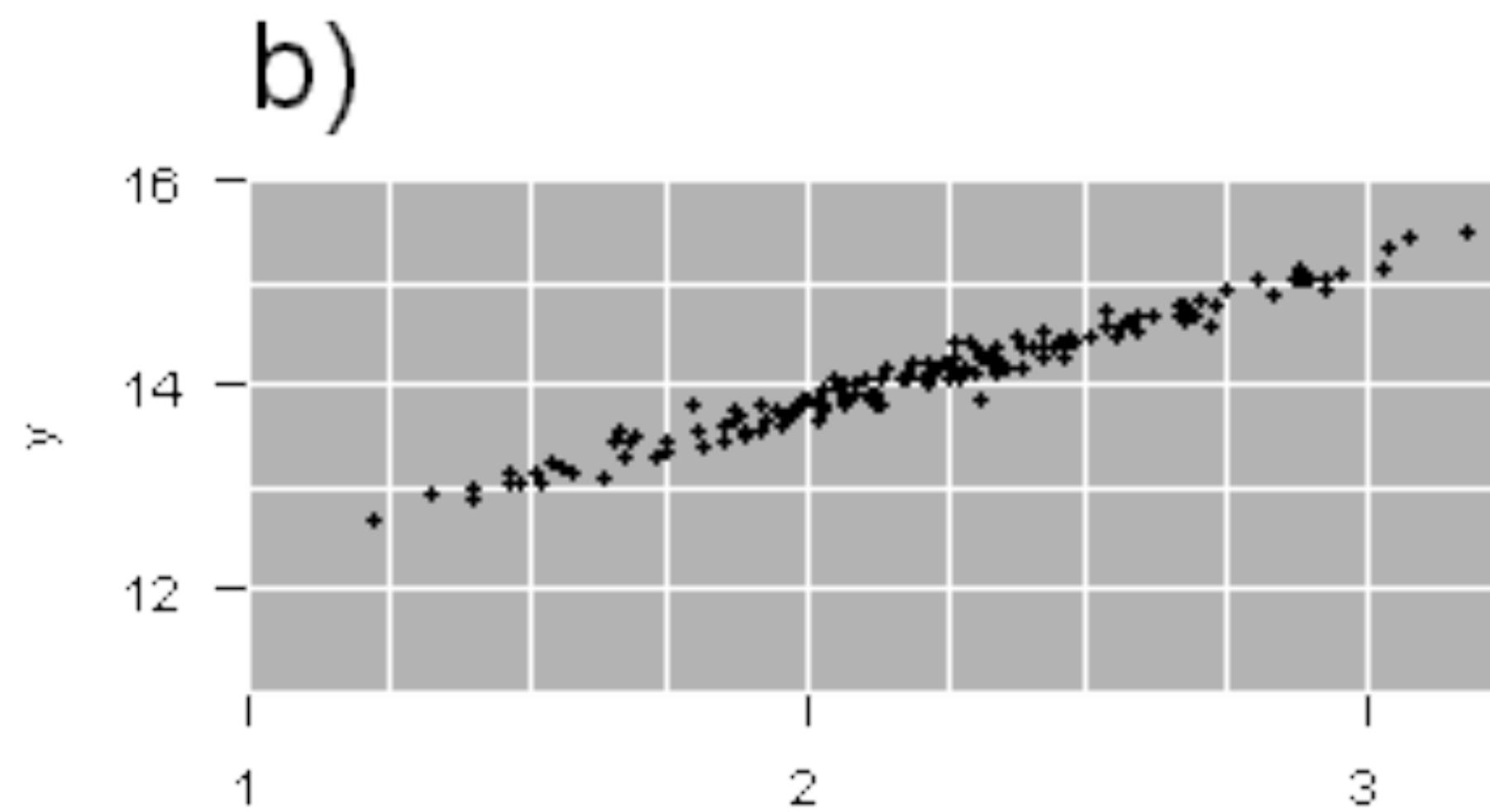
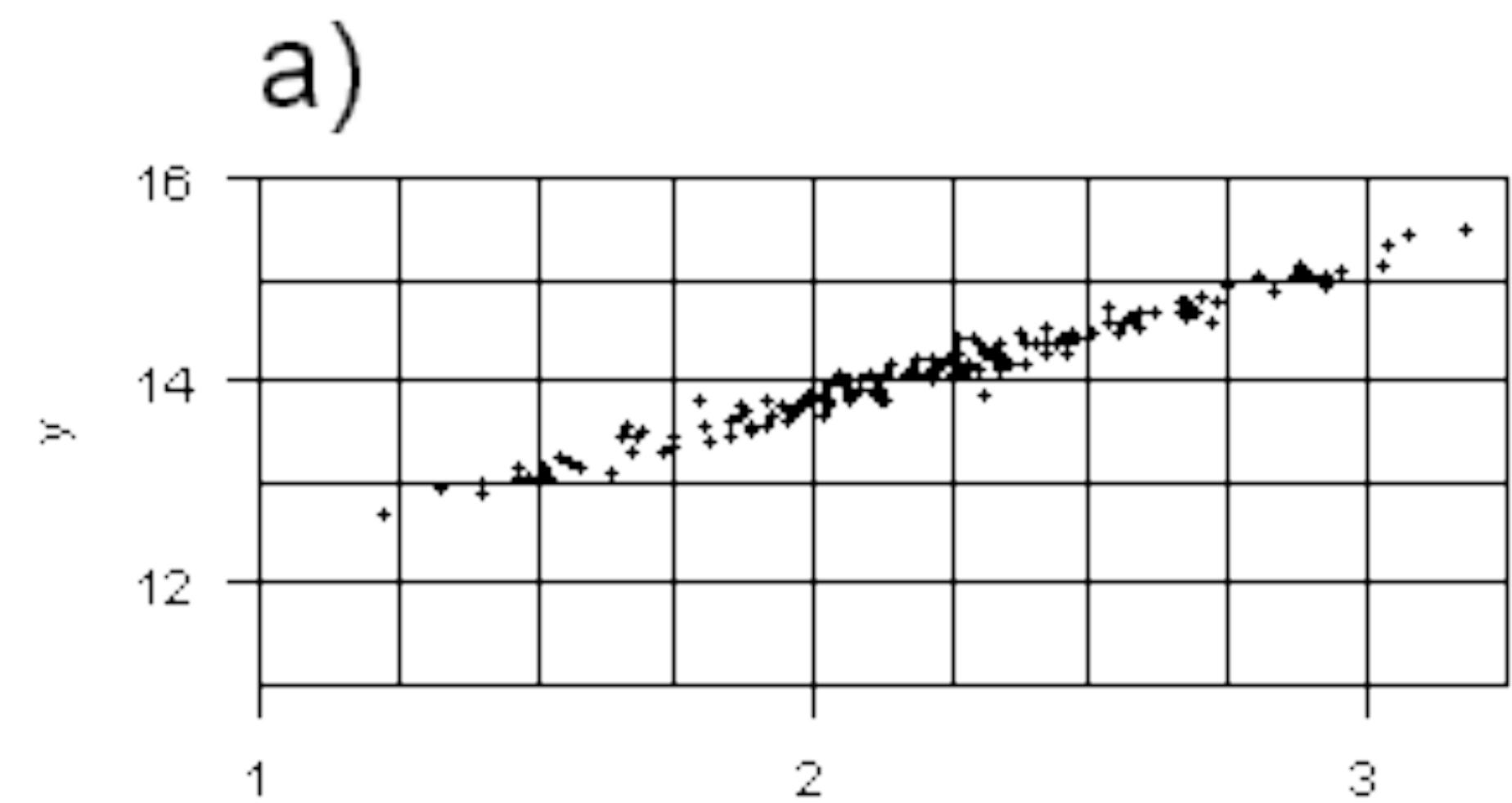




Calories per 100g



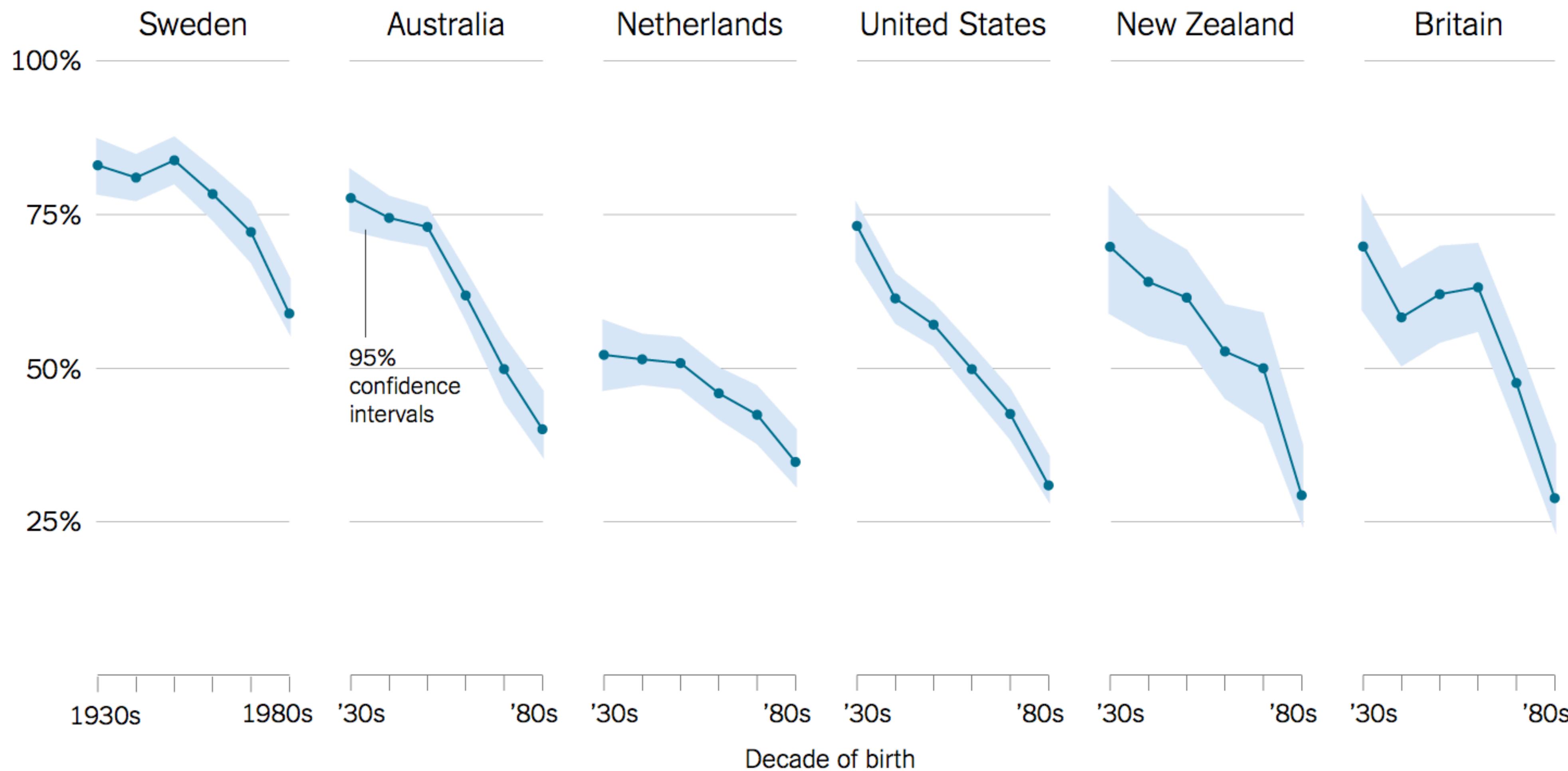




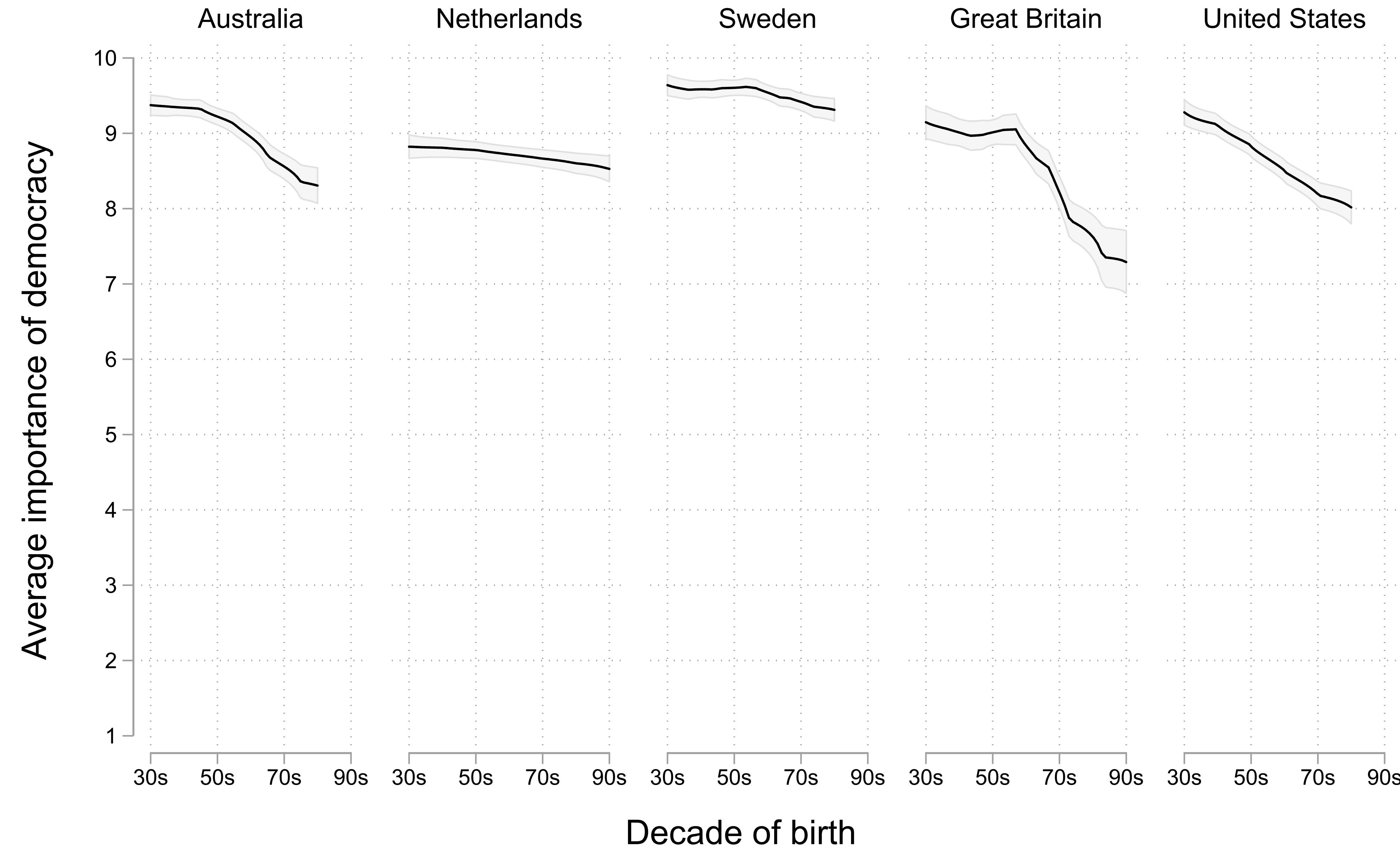
DATA

**JUNK-FREE
JUNK CHARTS**

Percentage of people who say it is “essential” to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, “The Signs of Democratic Deconsolidation,” Journal of Democracy | By The New York Times

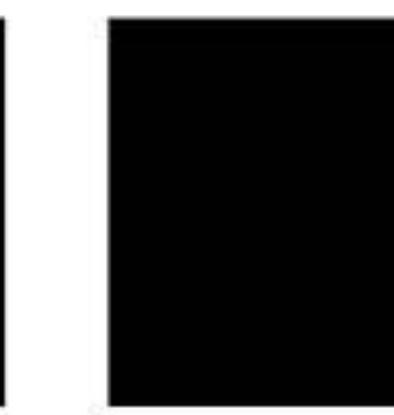
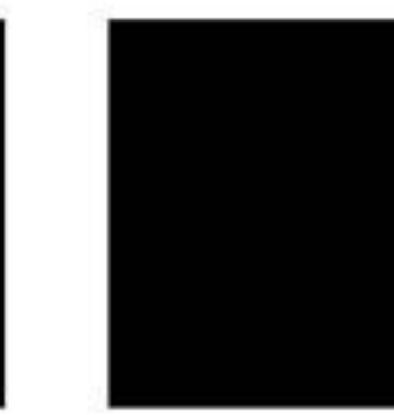


Graph by Erik Voeten, based on WVS 5

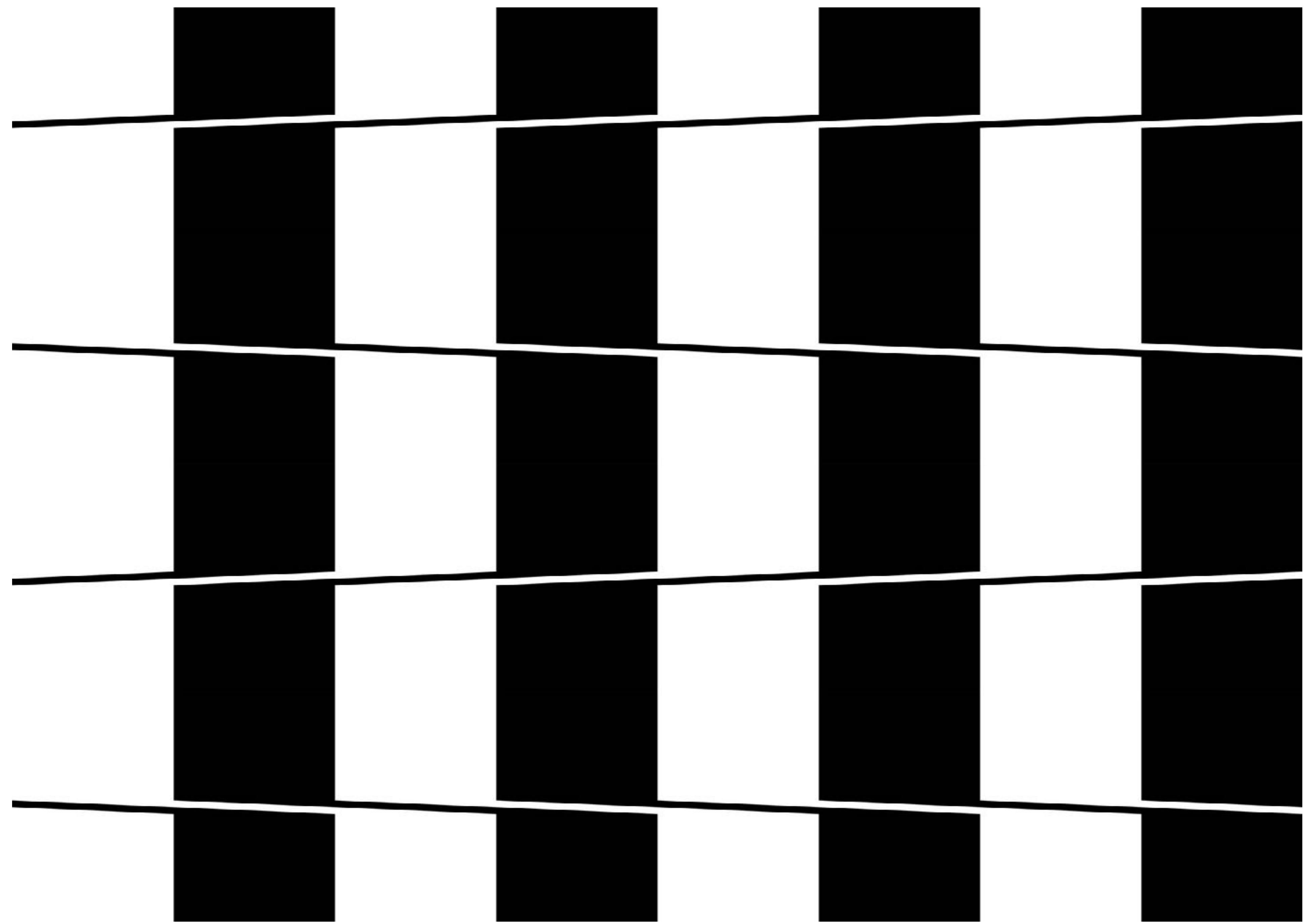
PERCEPTION

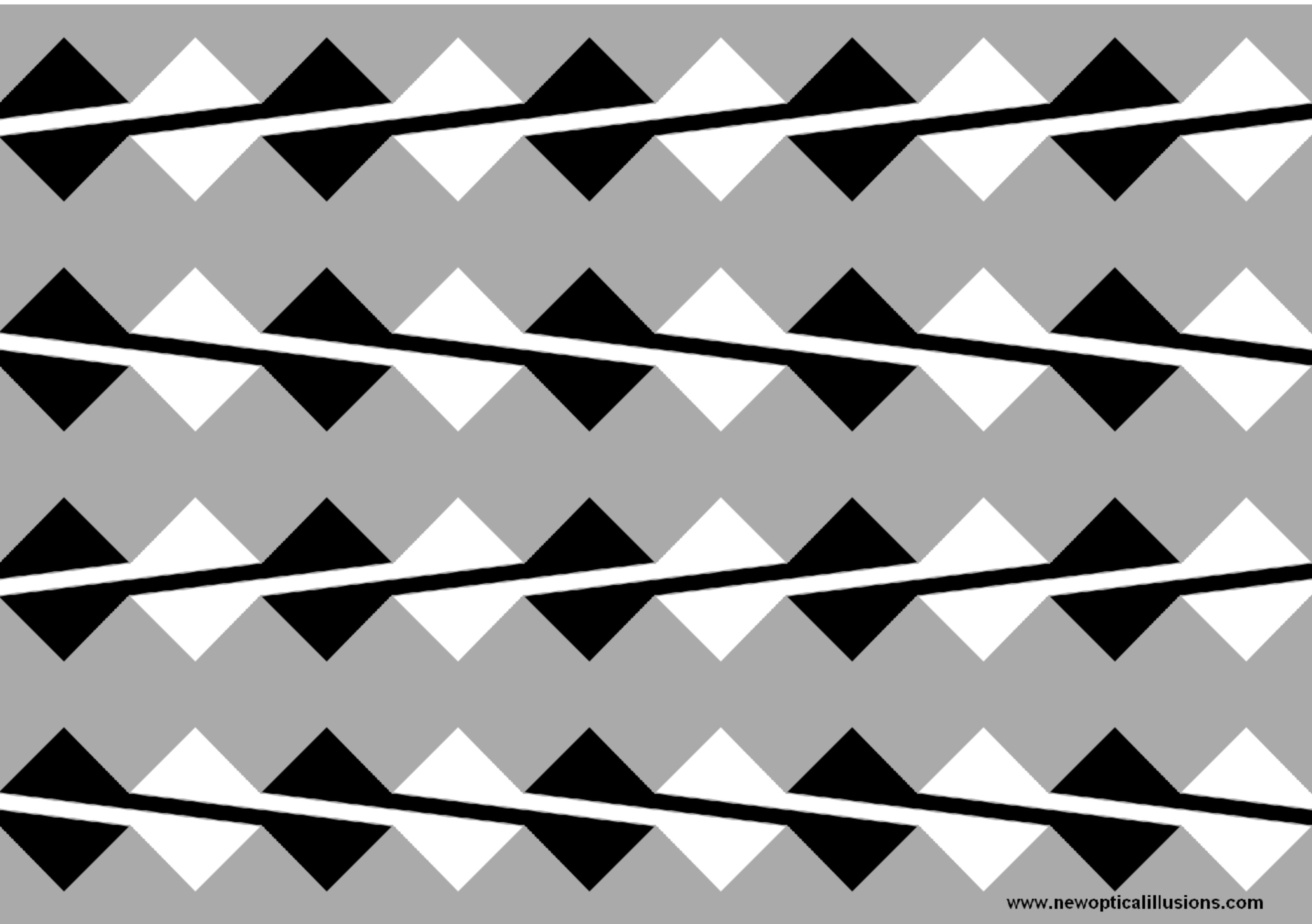
SIGHT AND INSIGHT

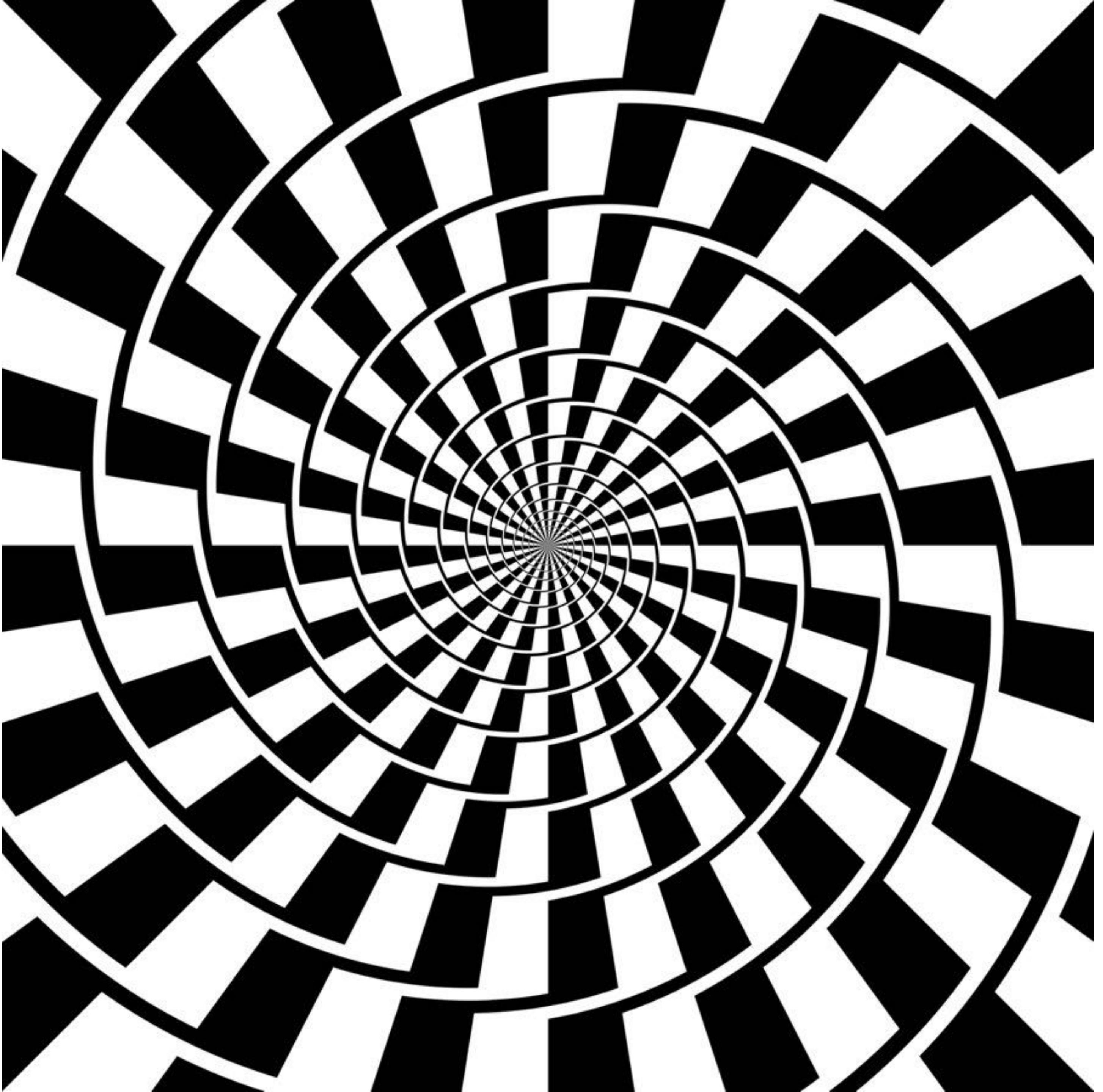
**Edges &
Contrasts**

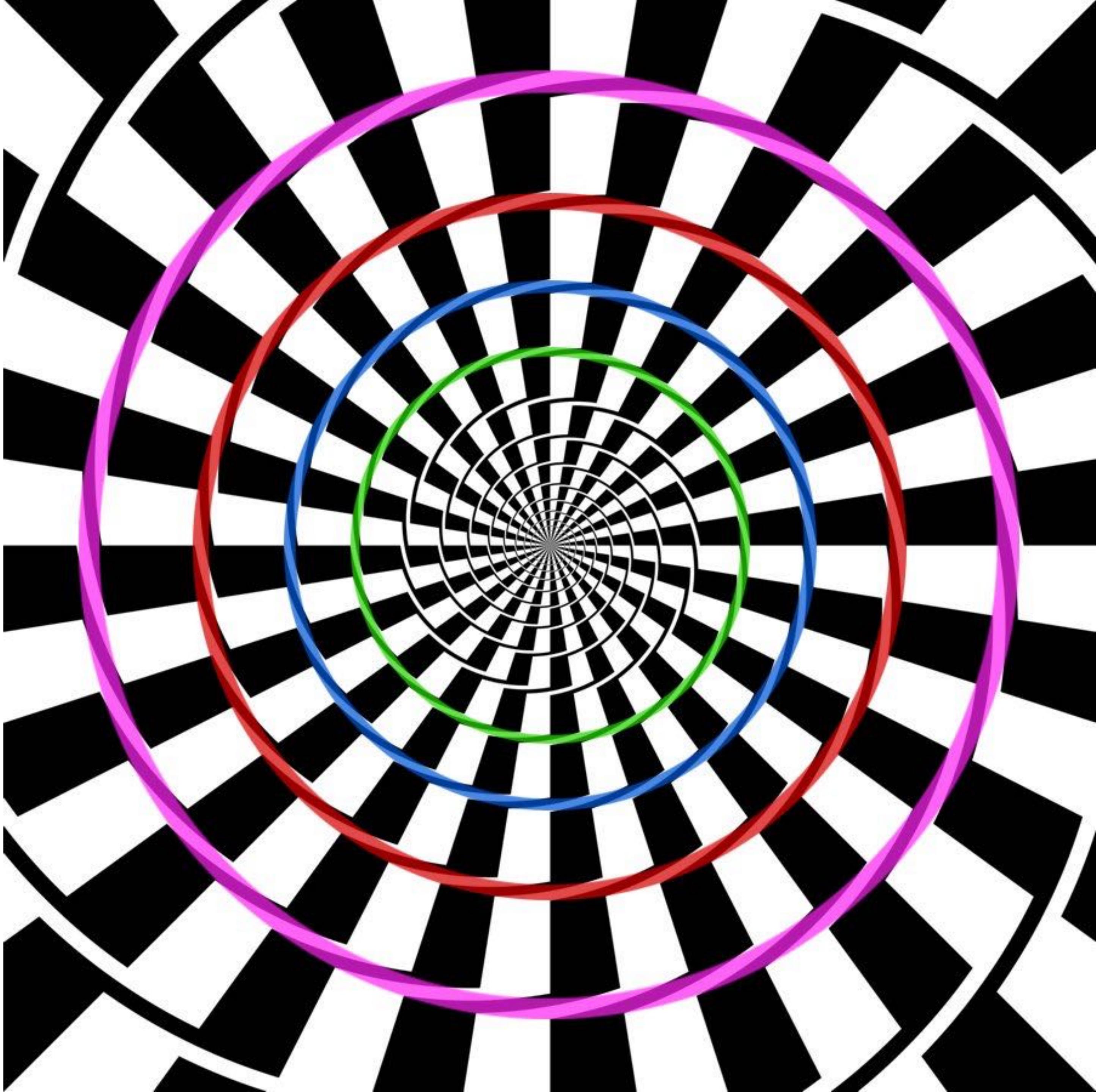


fin

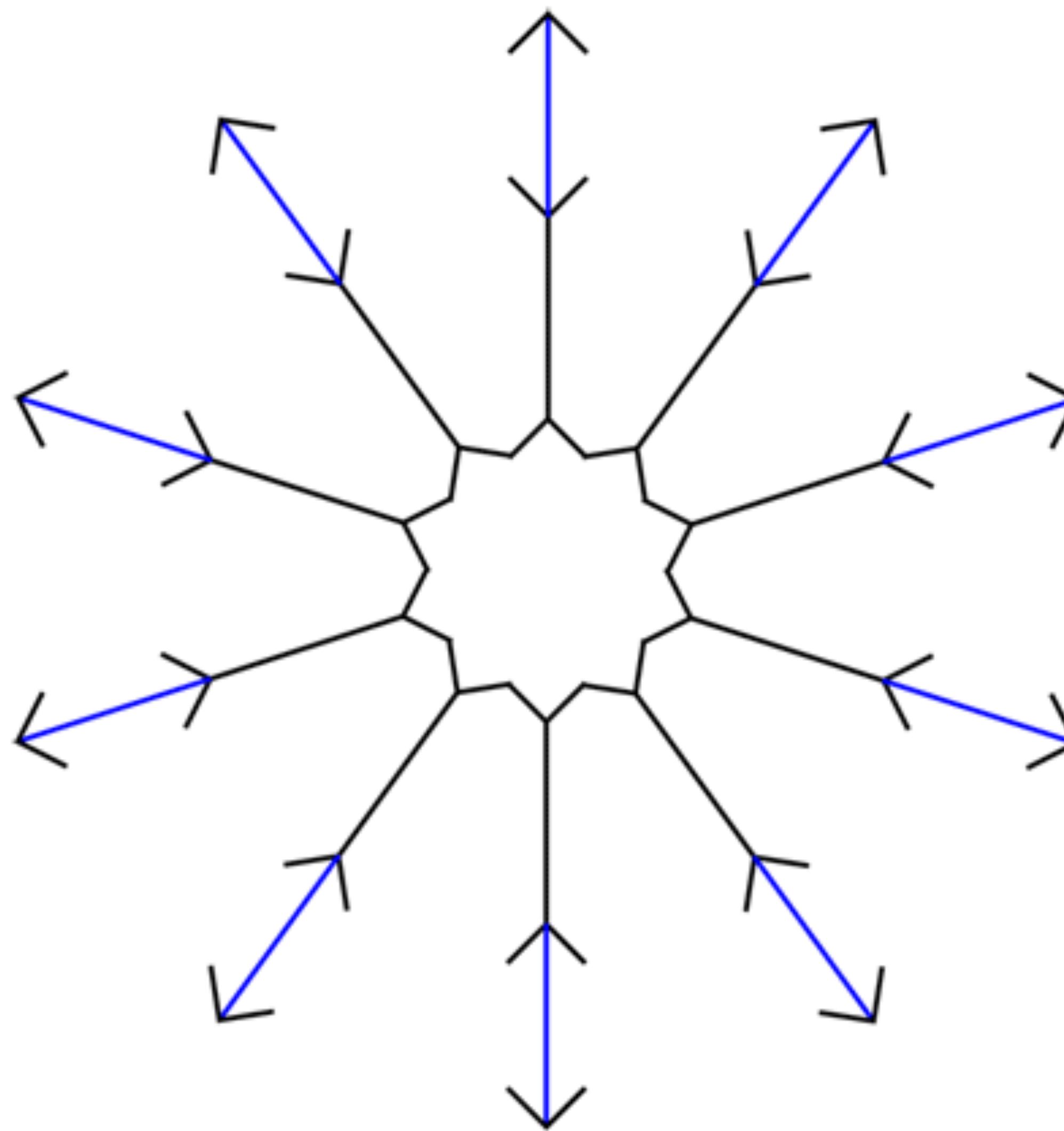






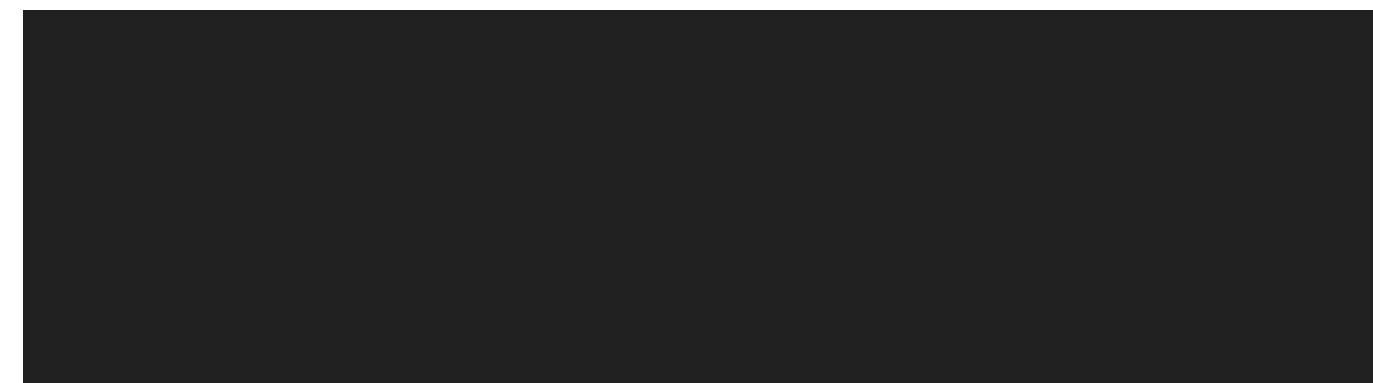


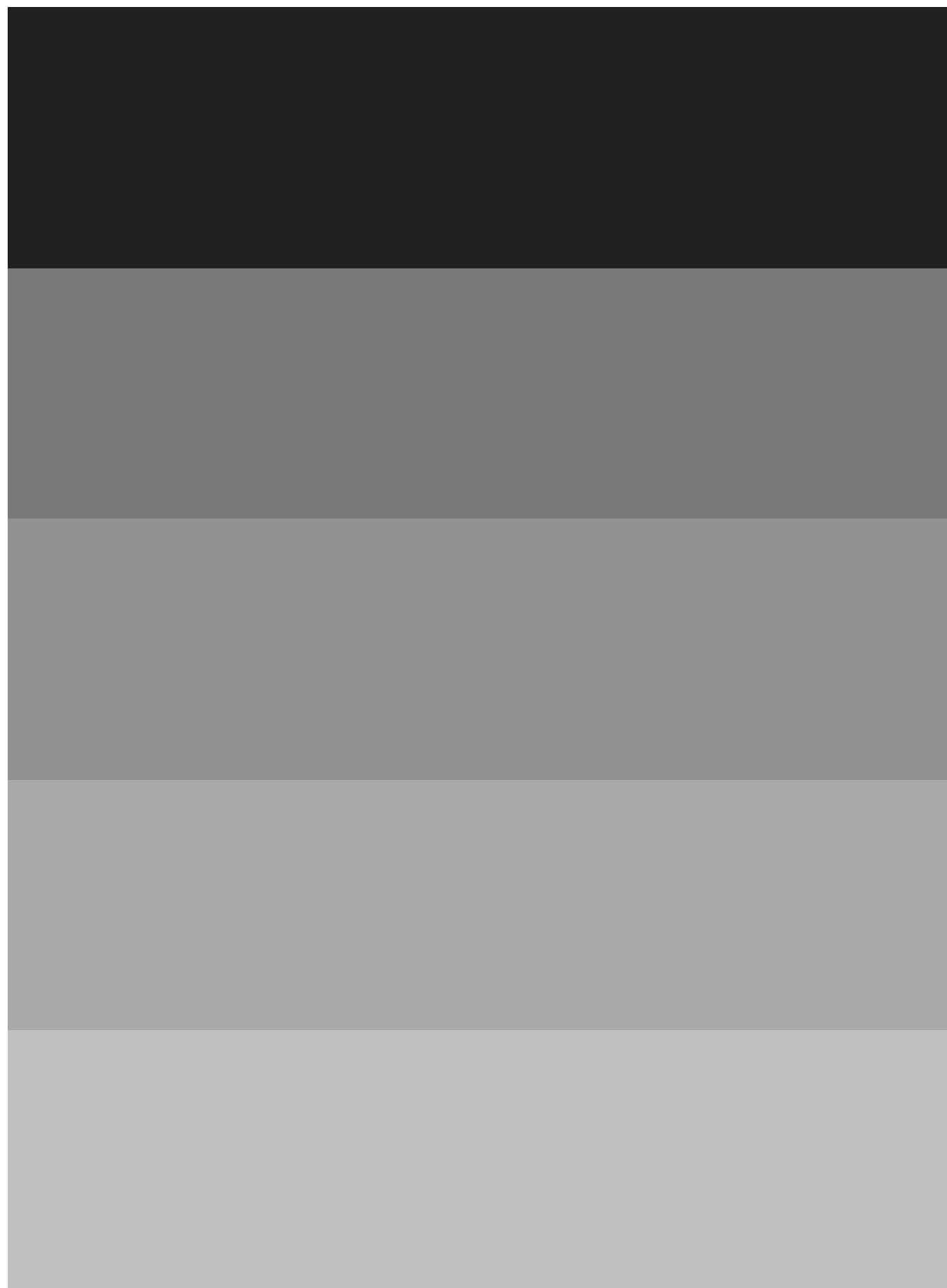
Sarcone's Dynamic Müller-Lyer Illusion



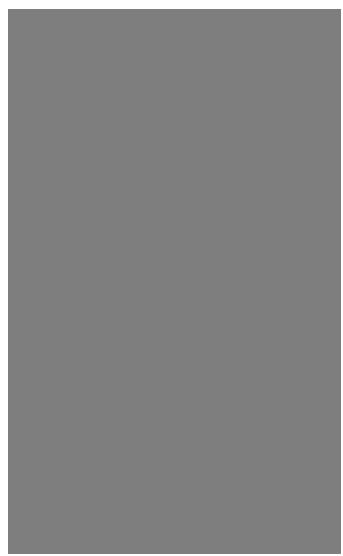
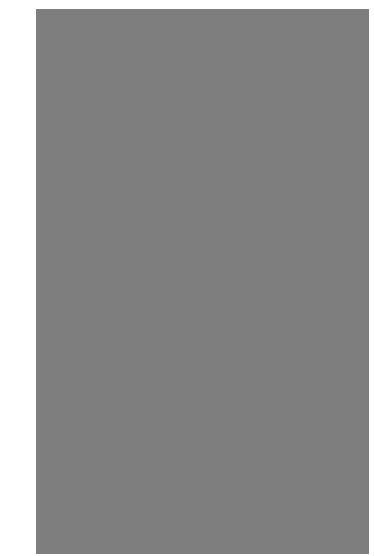


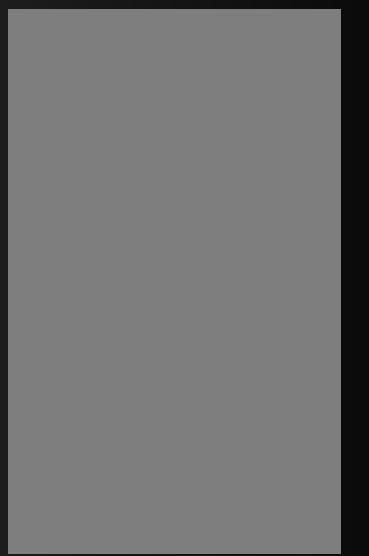
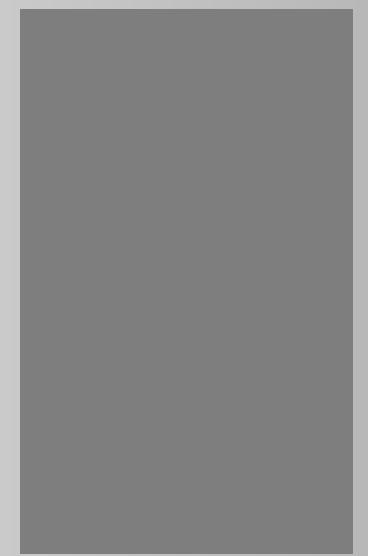
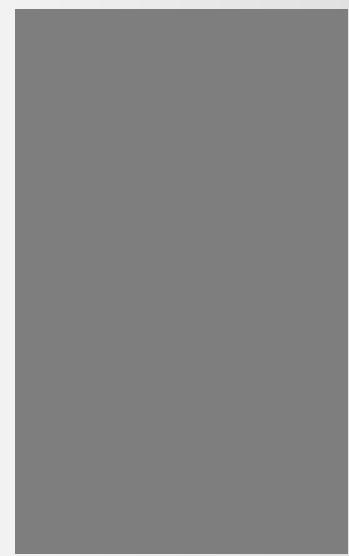
These are two perfectly
geometrical circles

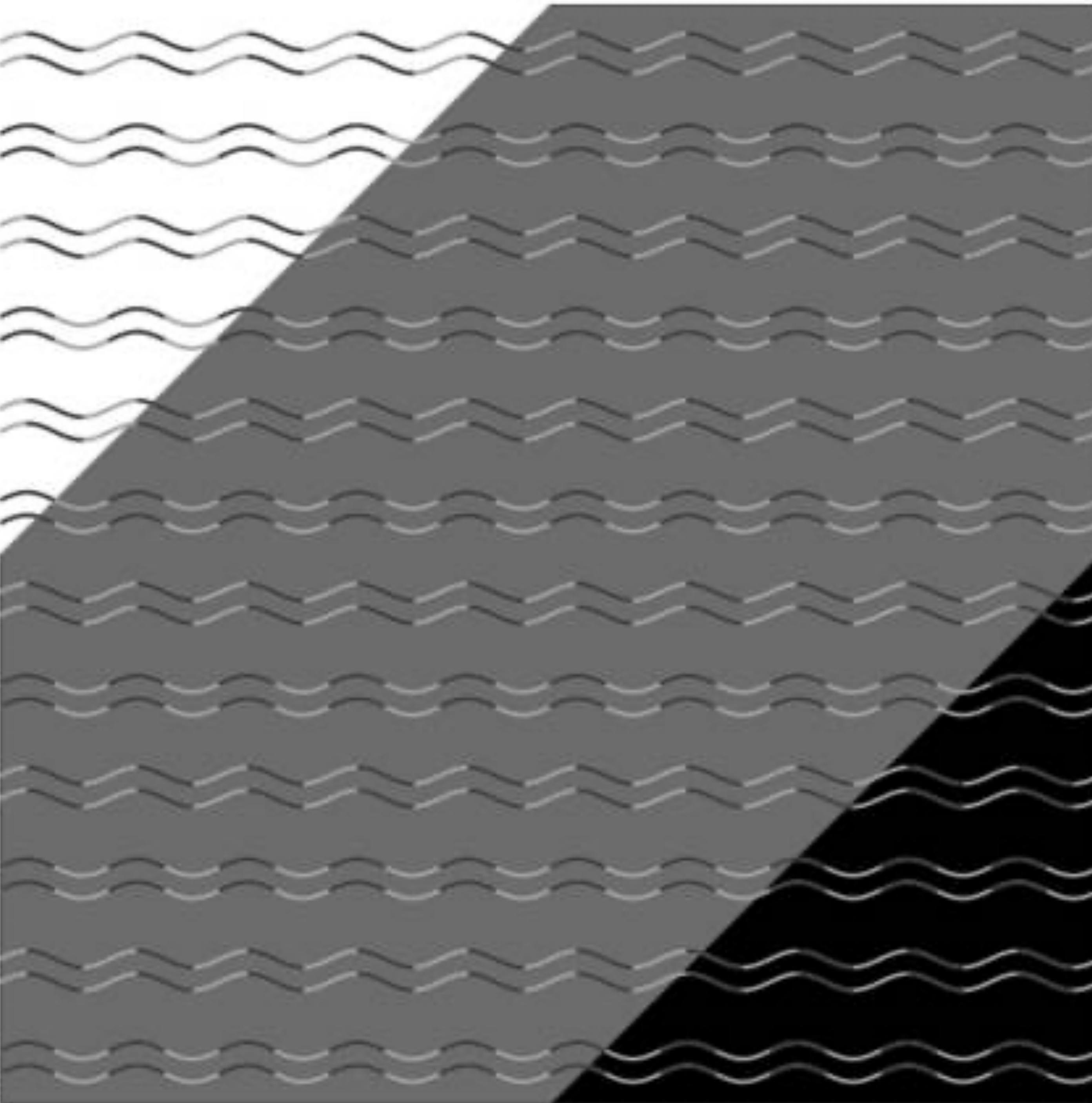


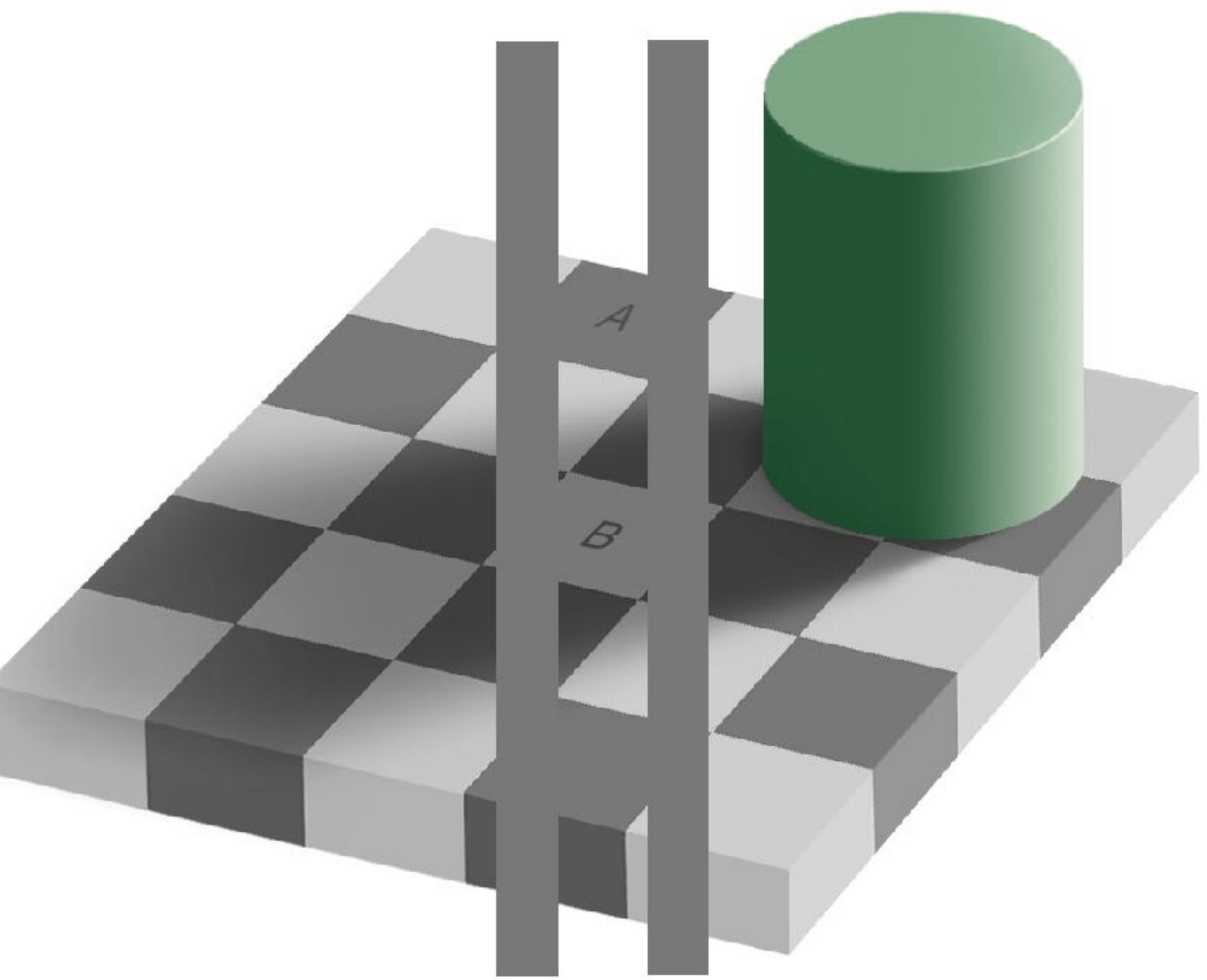
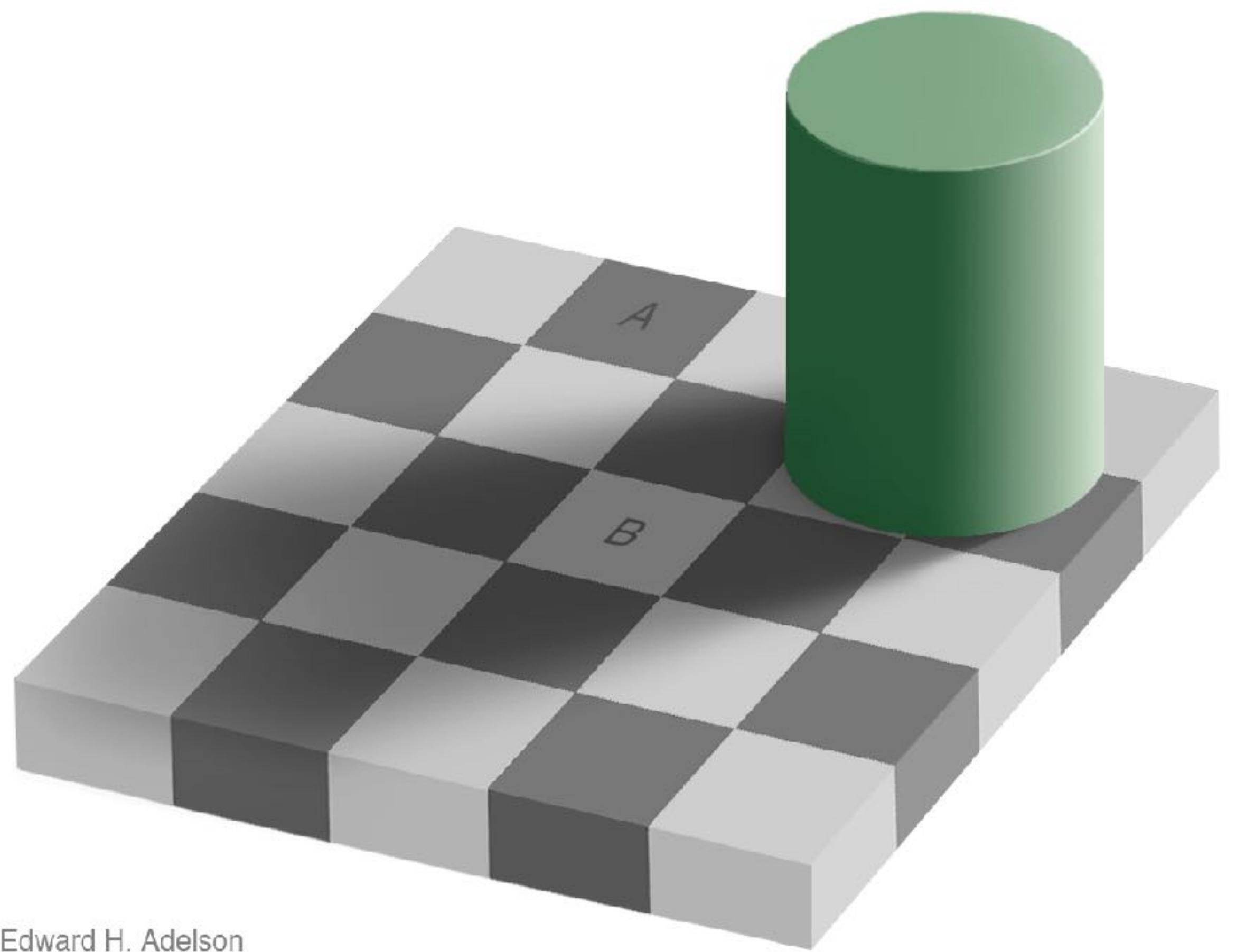




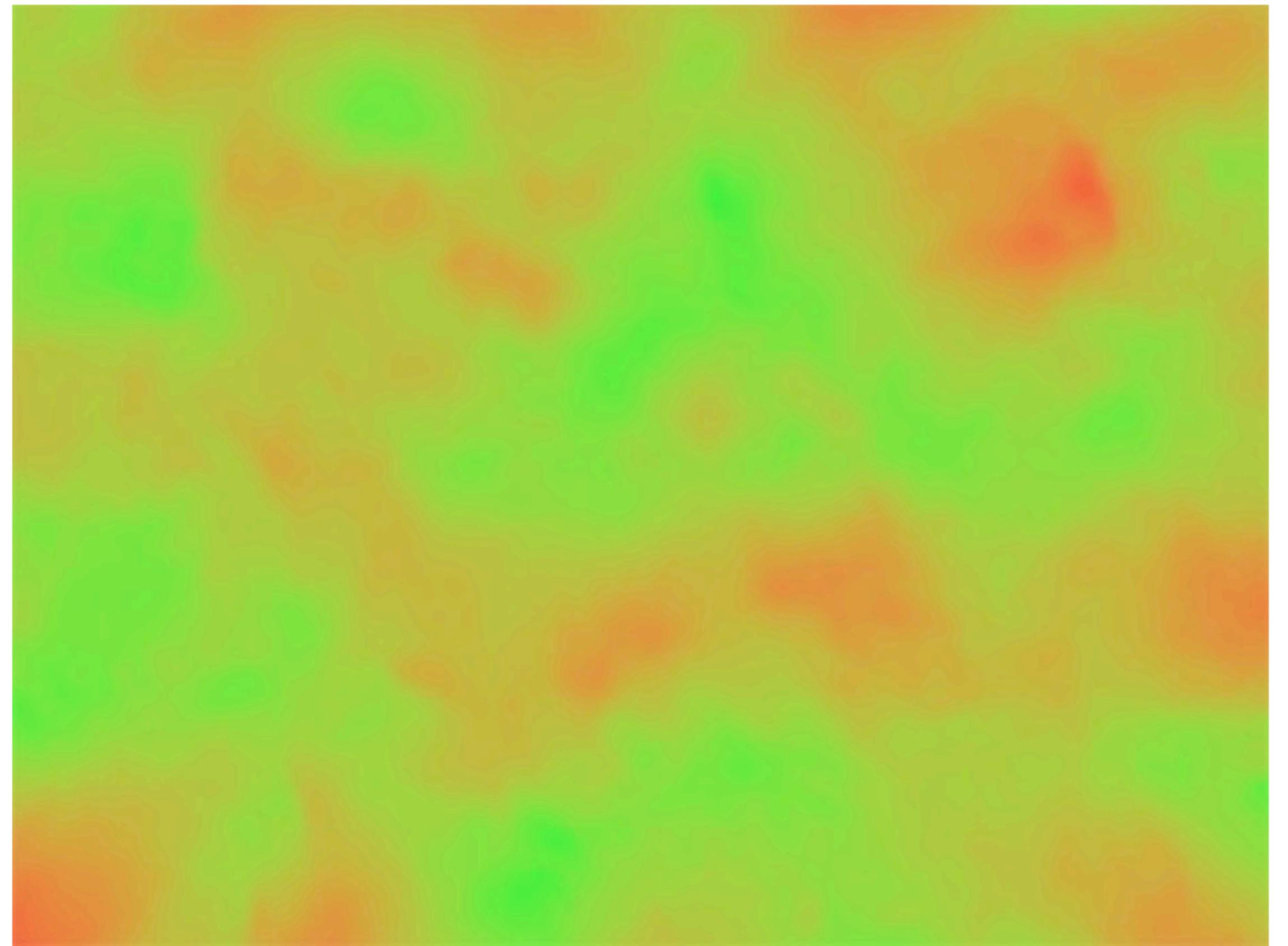
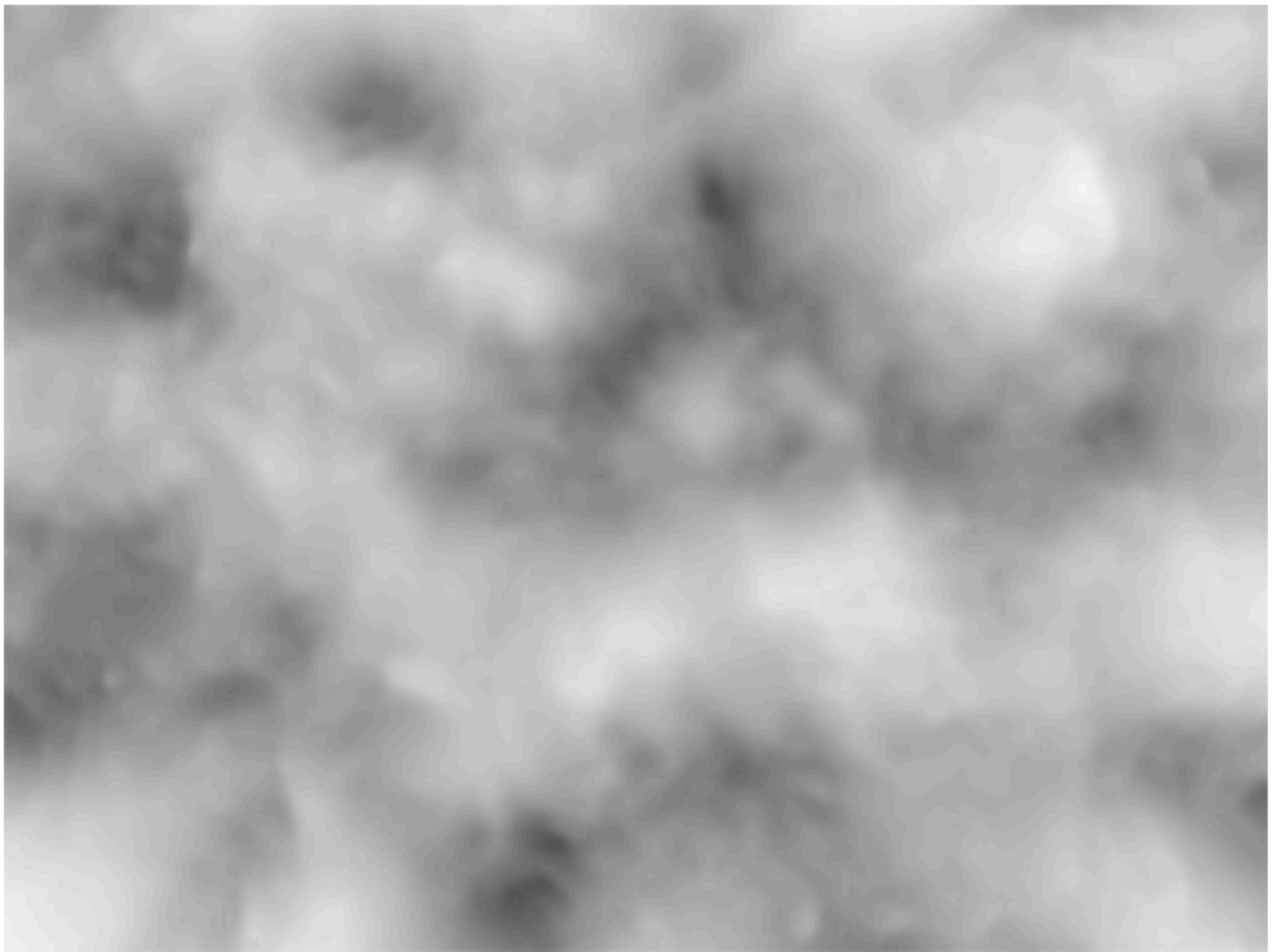


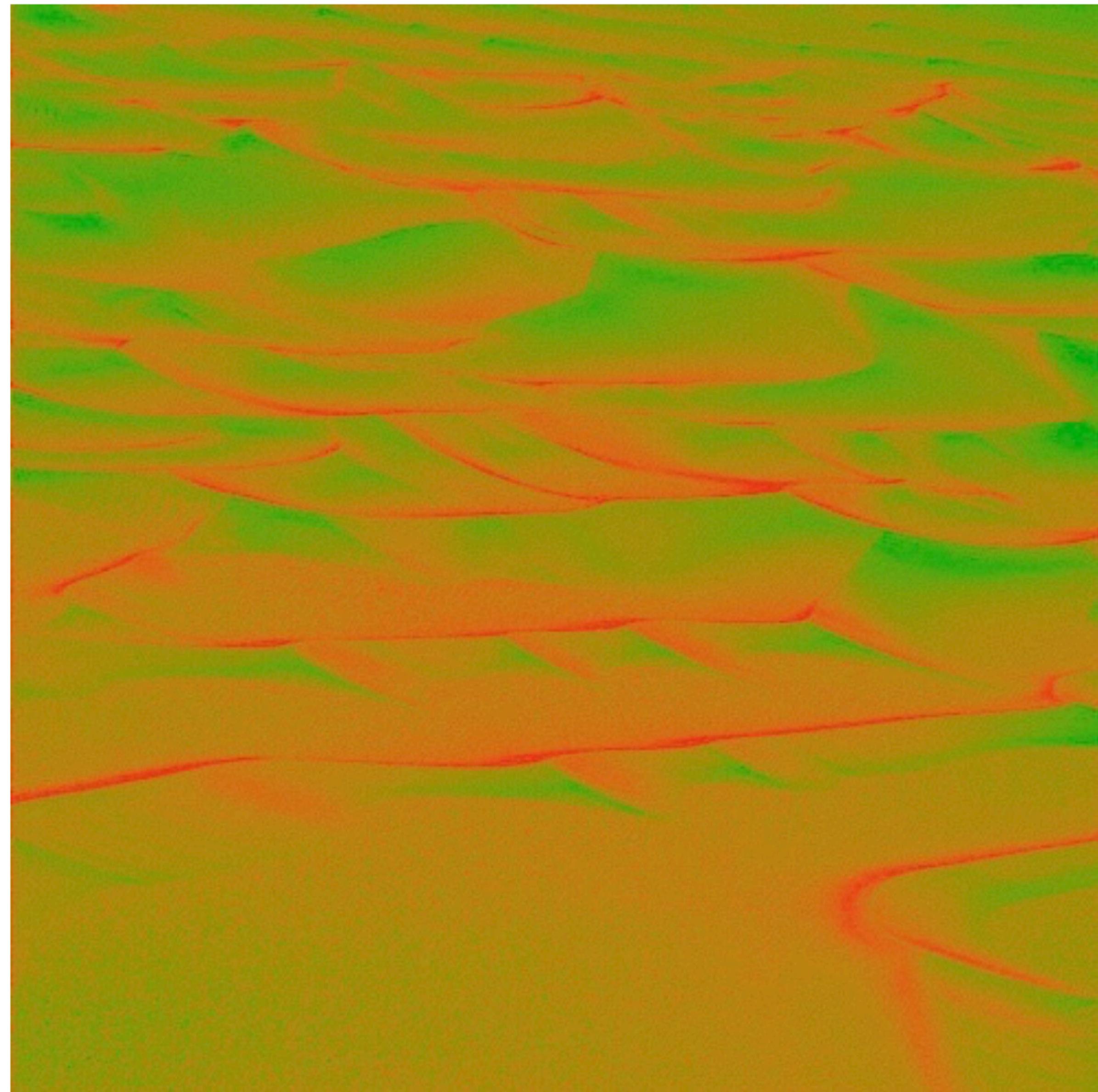


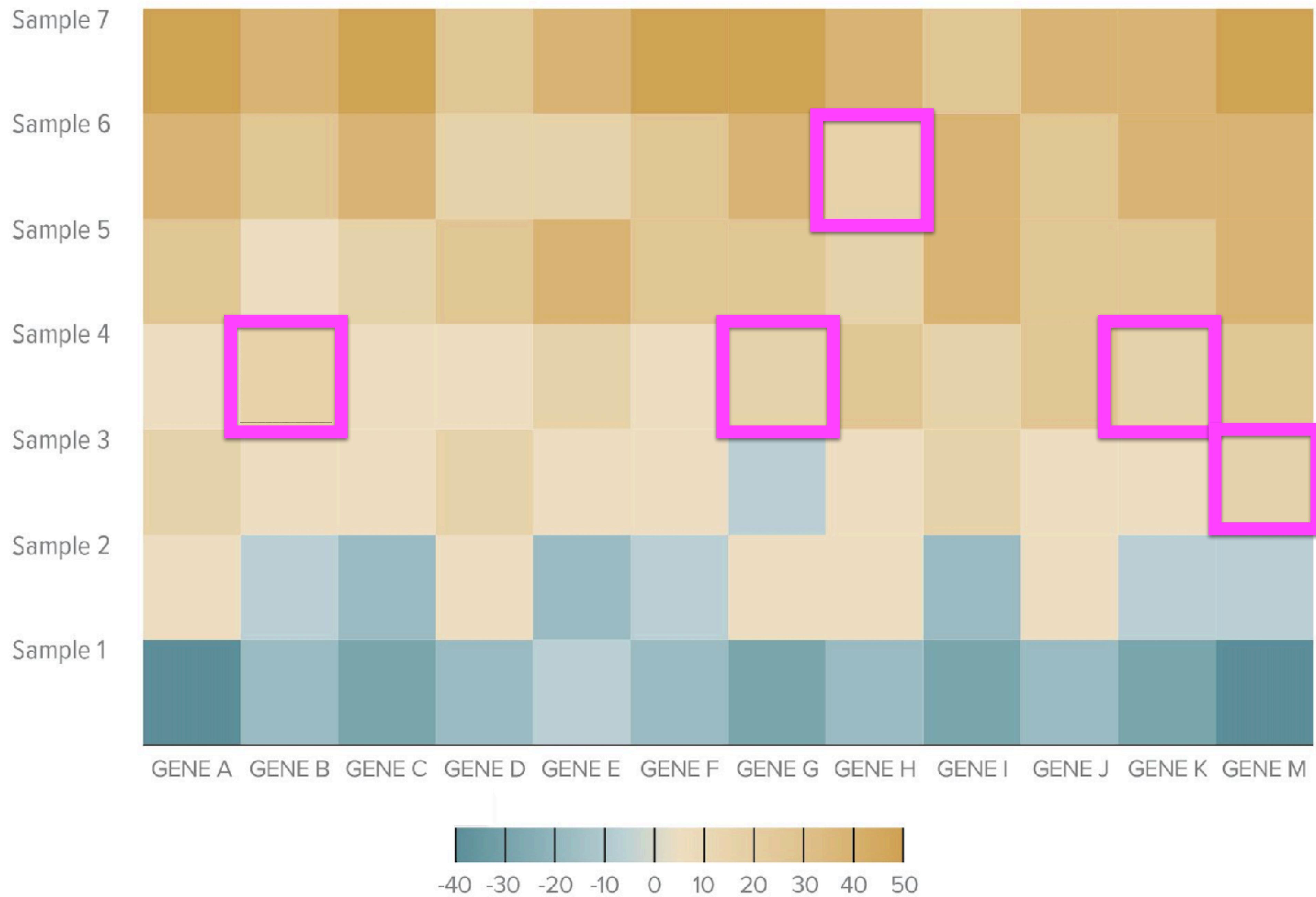


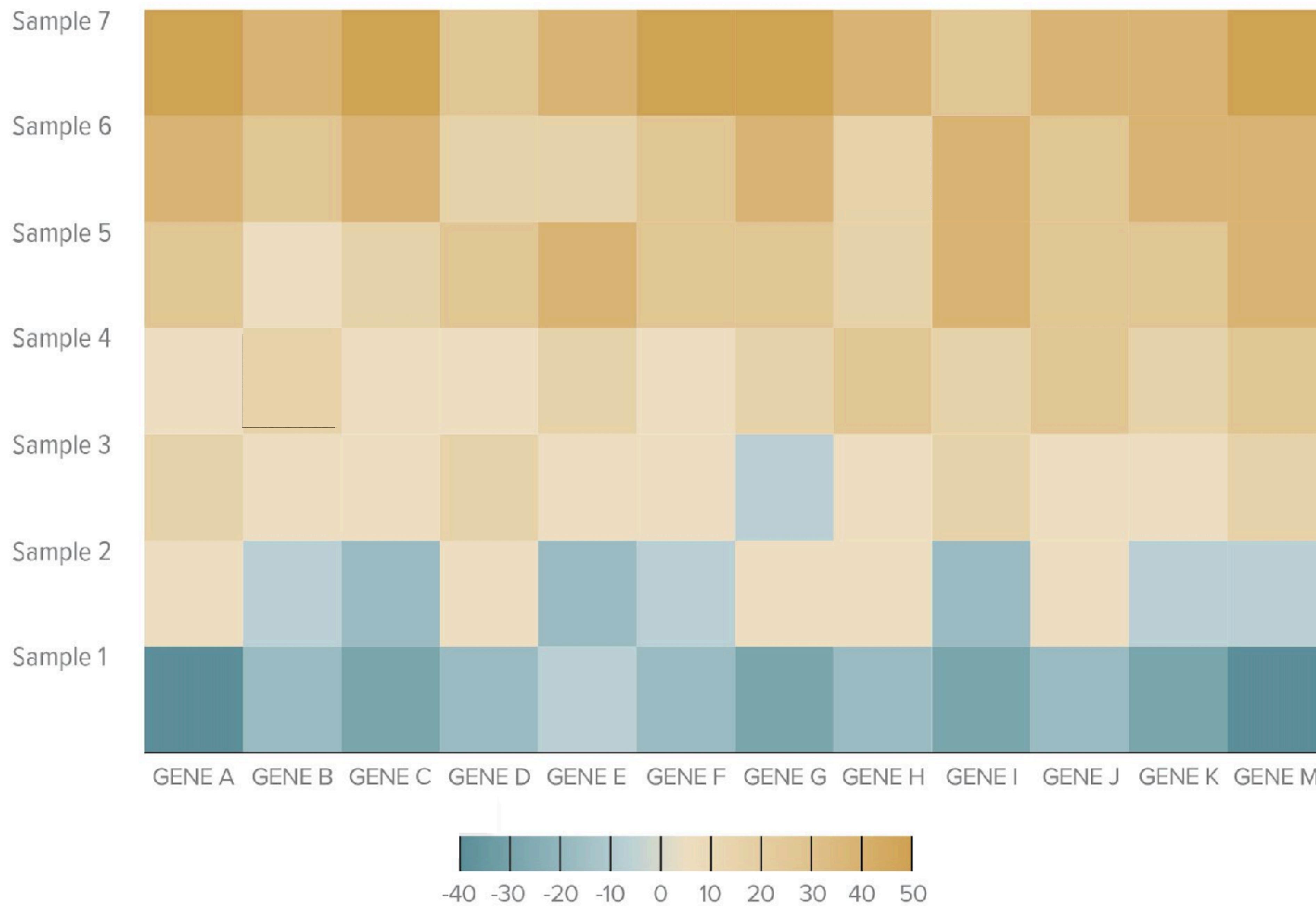


Edward H. Adelson





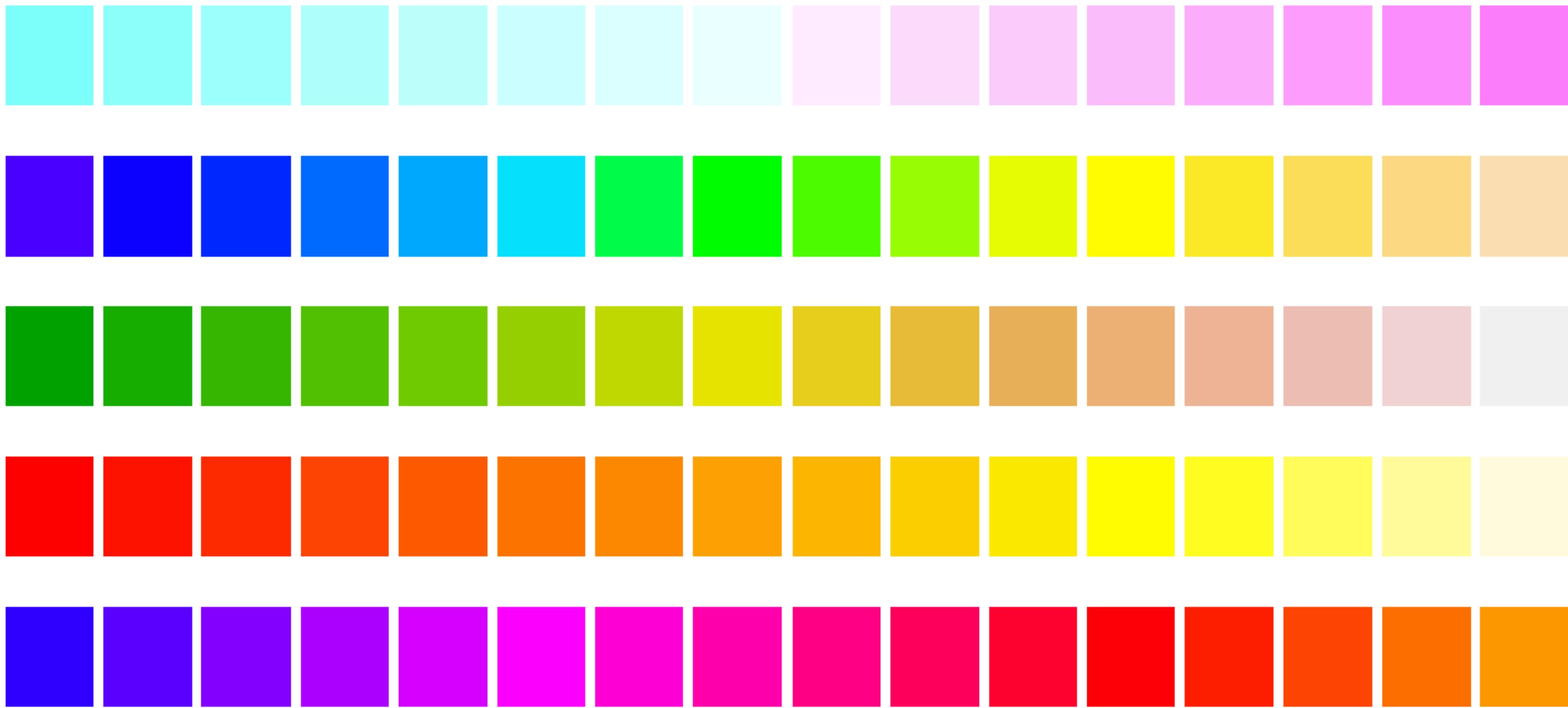




Color

+





viridis



magma



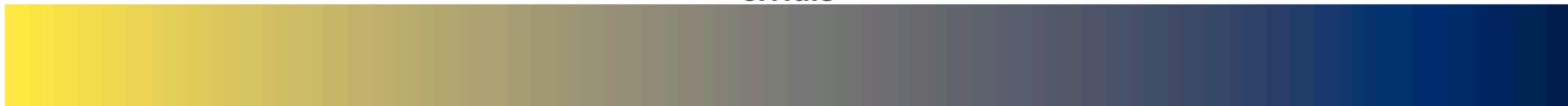
plasma

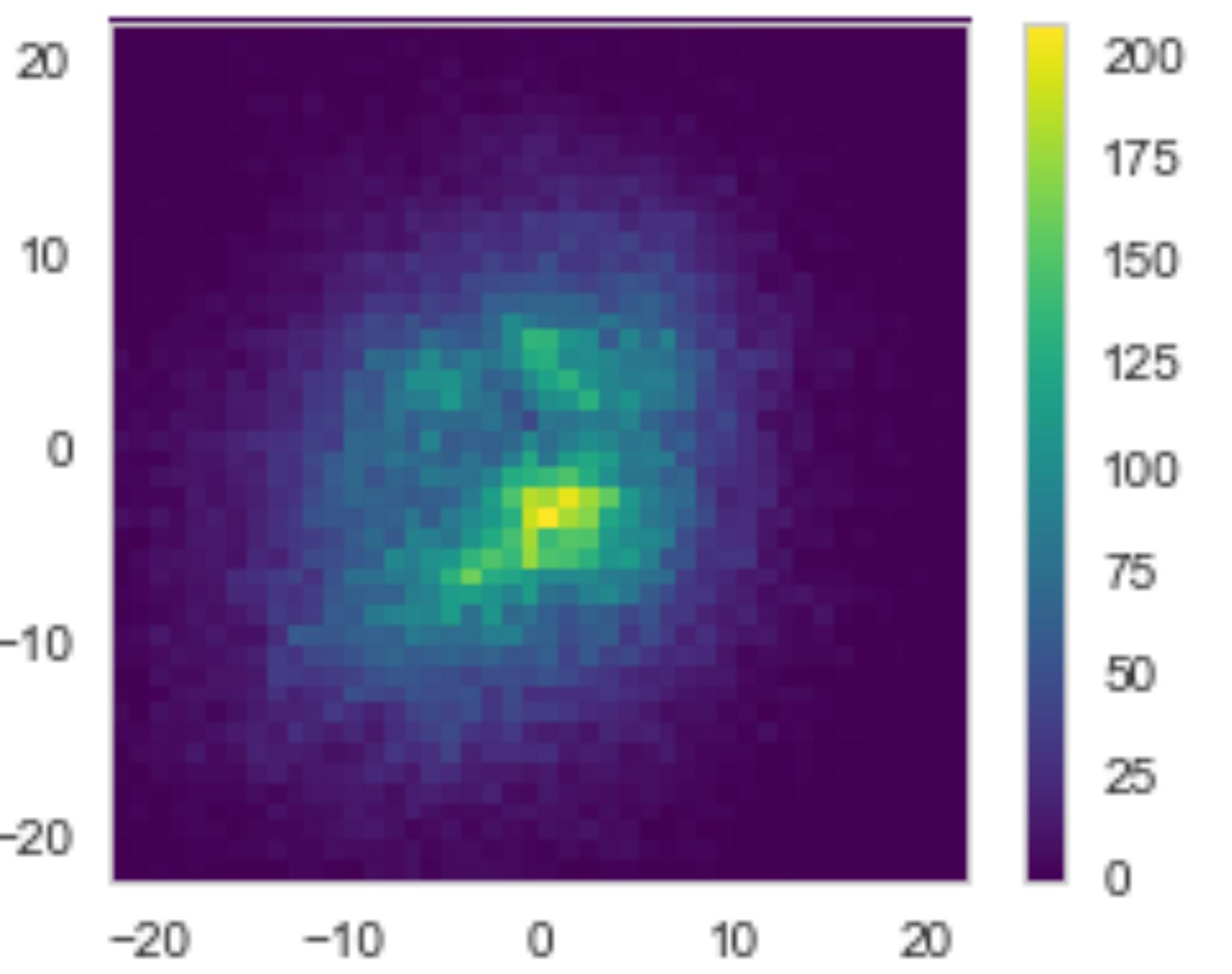
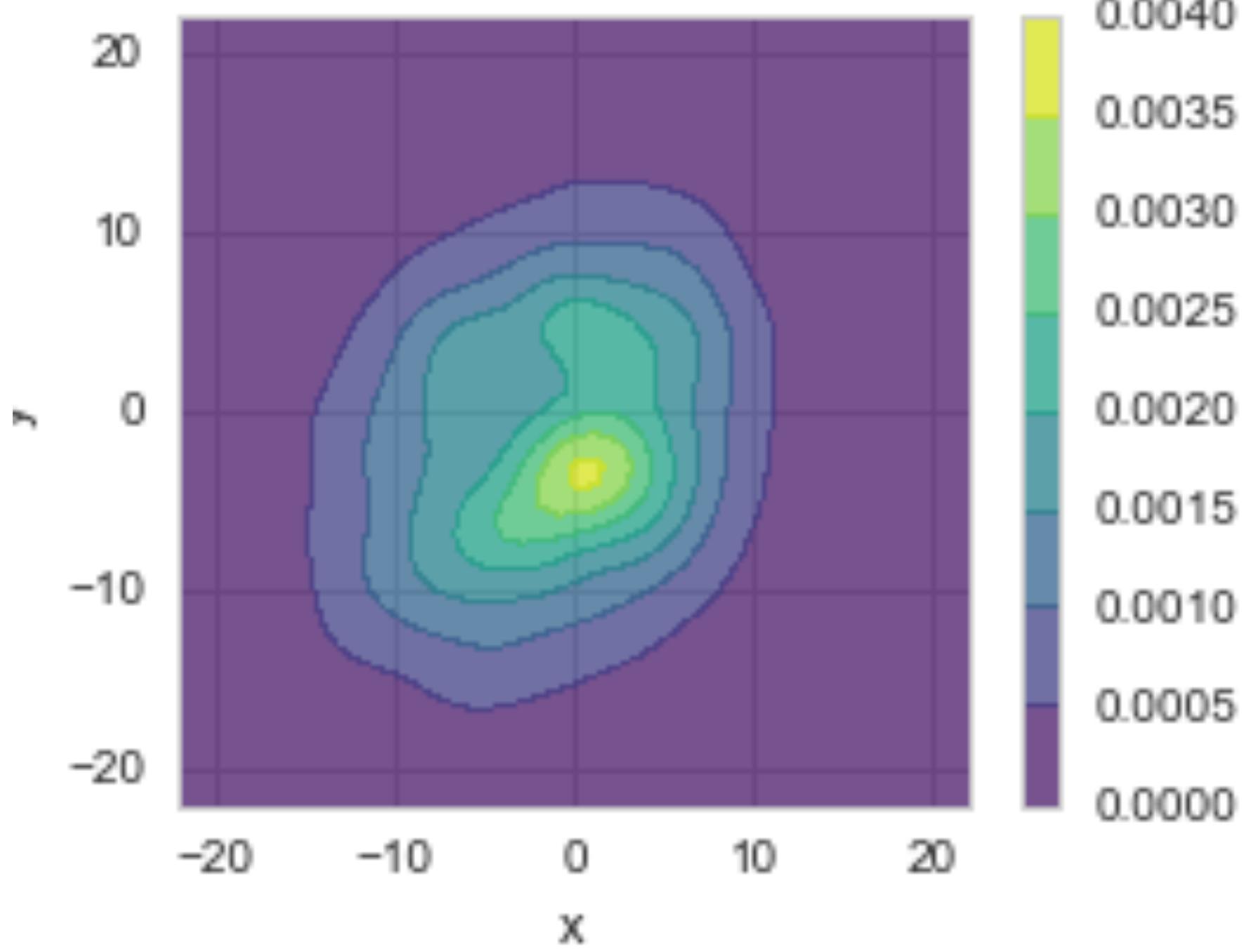
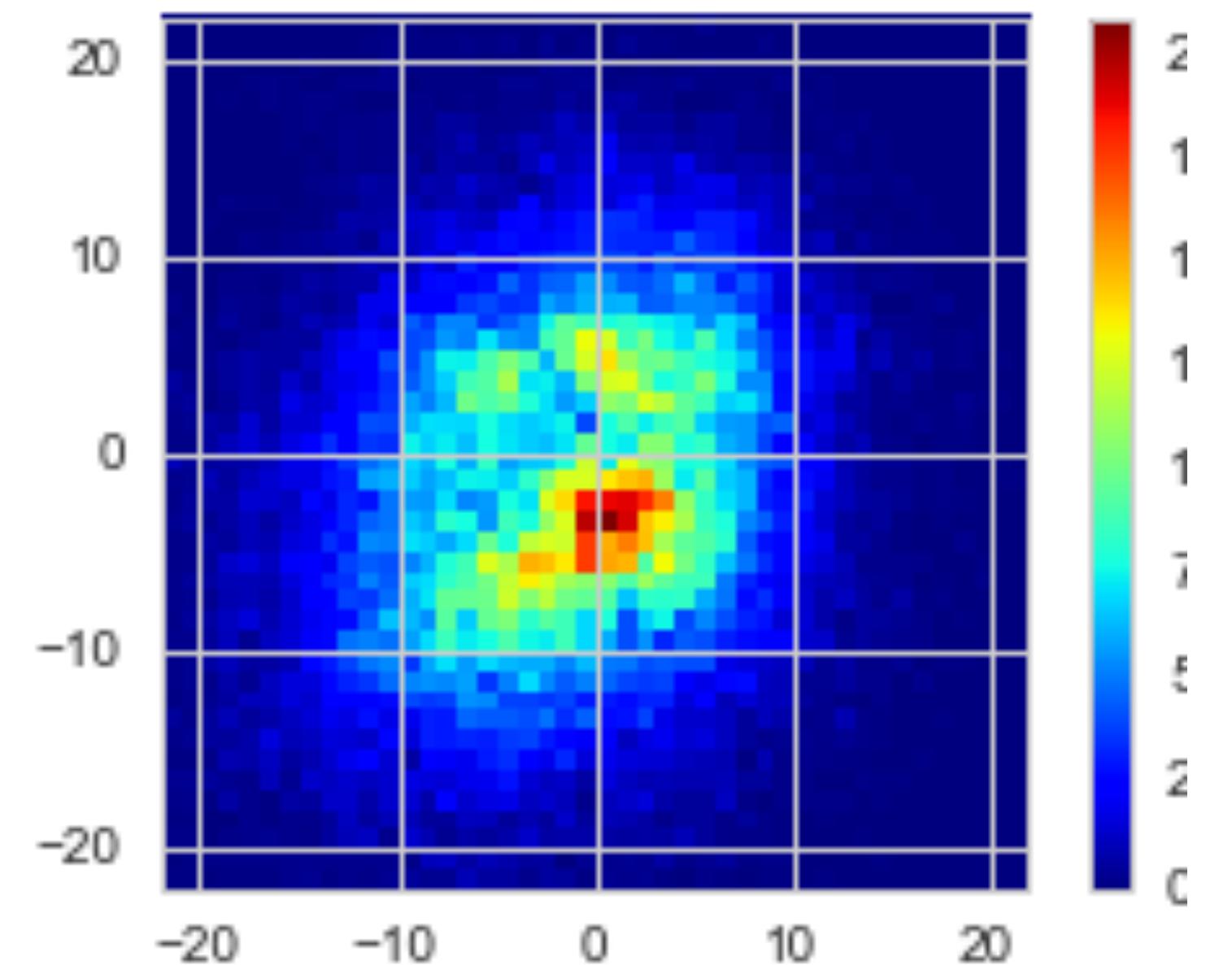
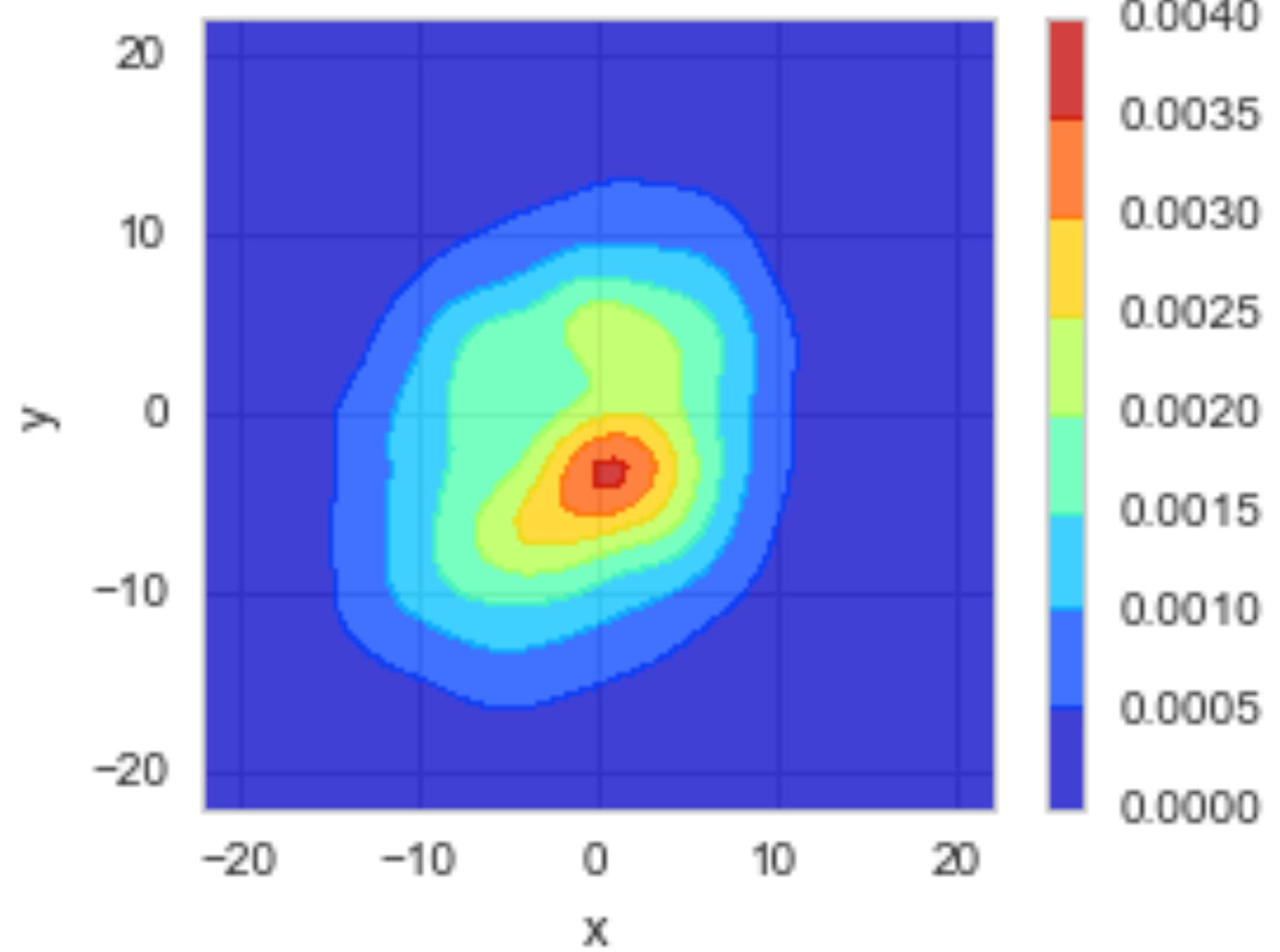


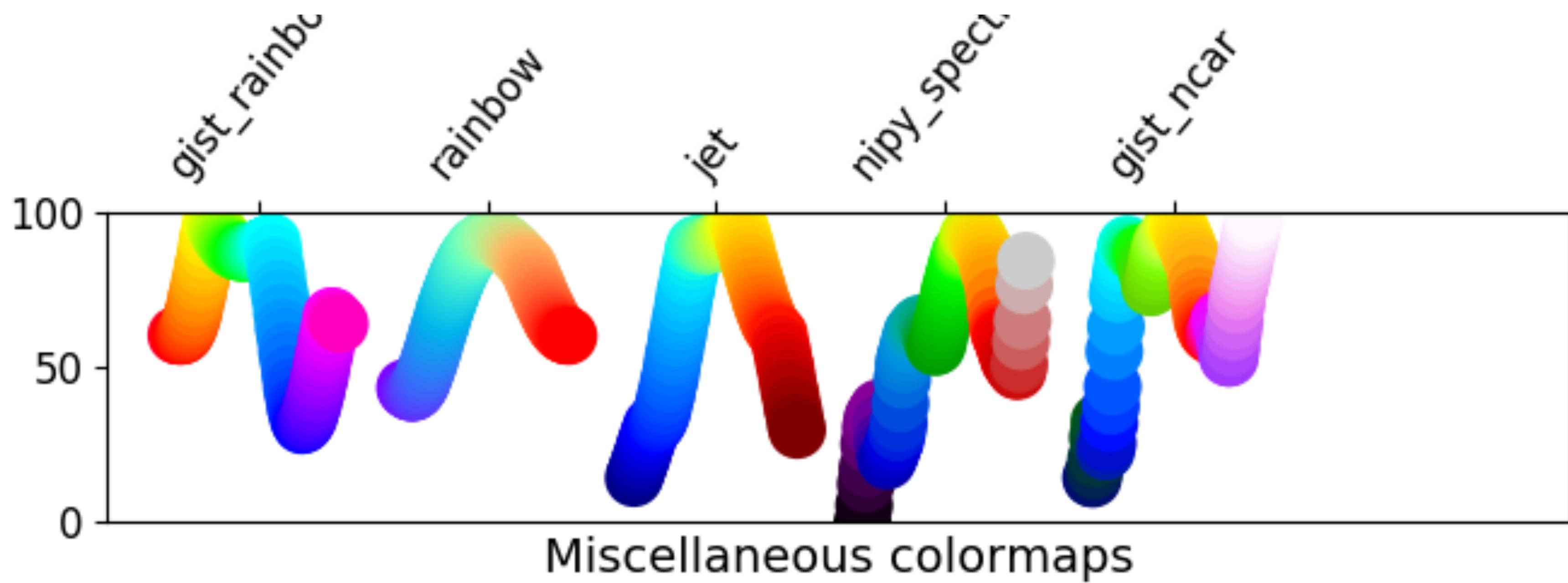
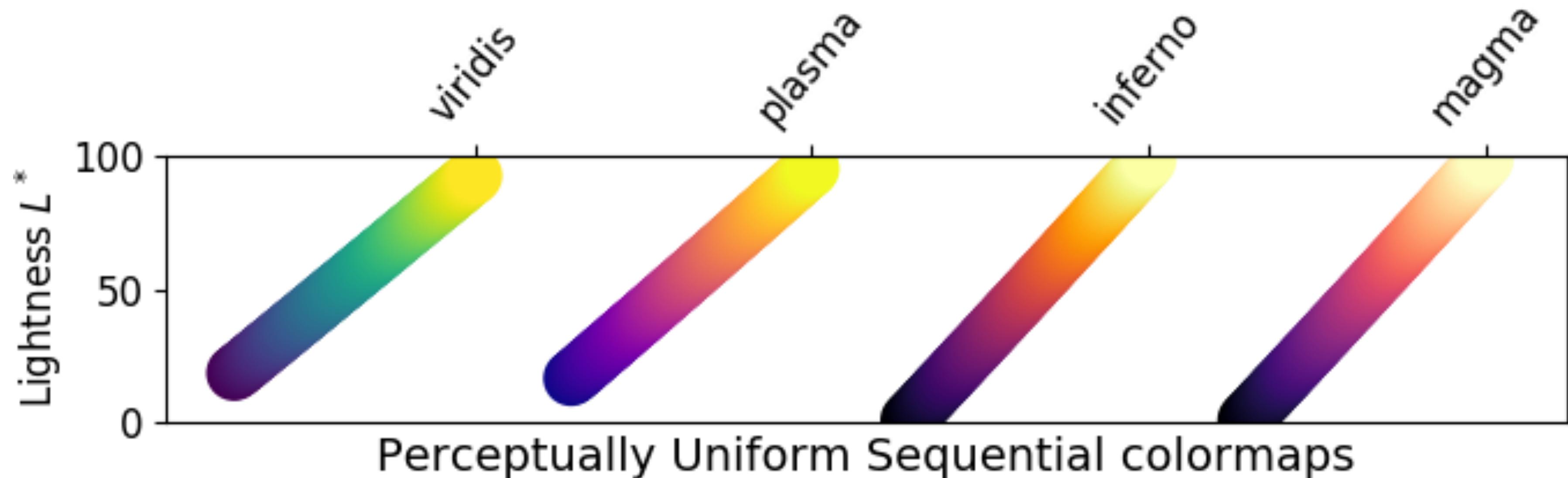
inferno

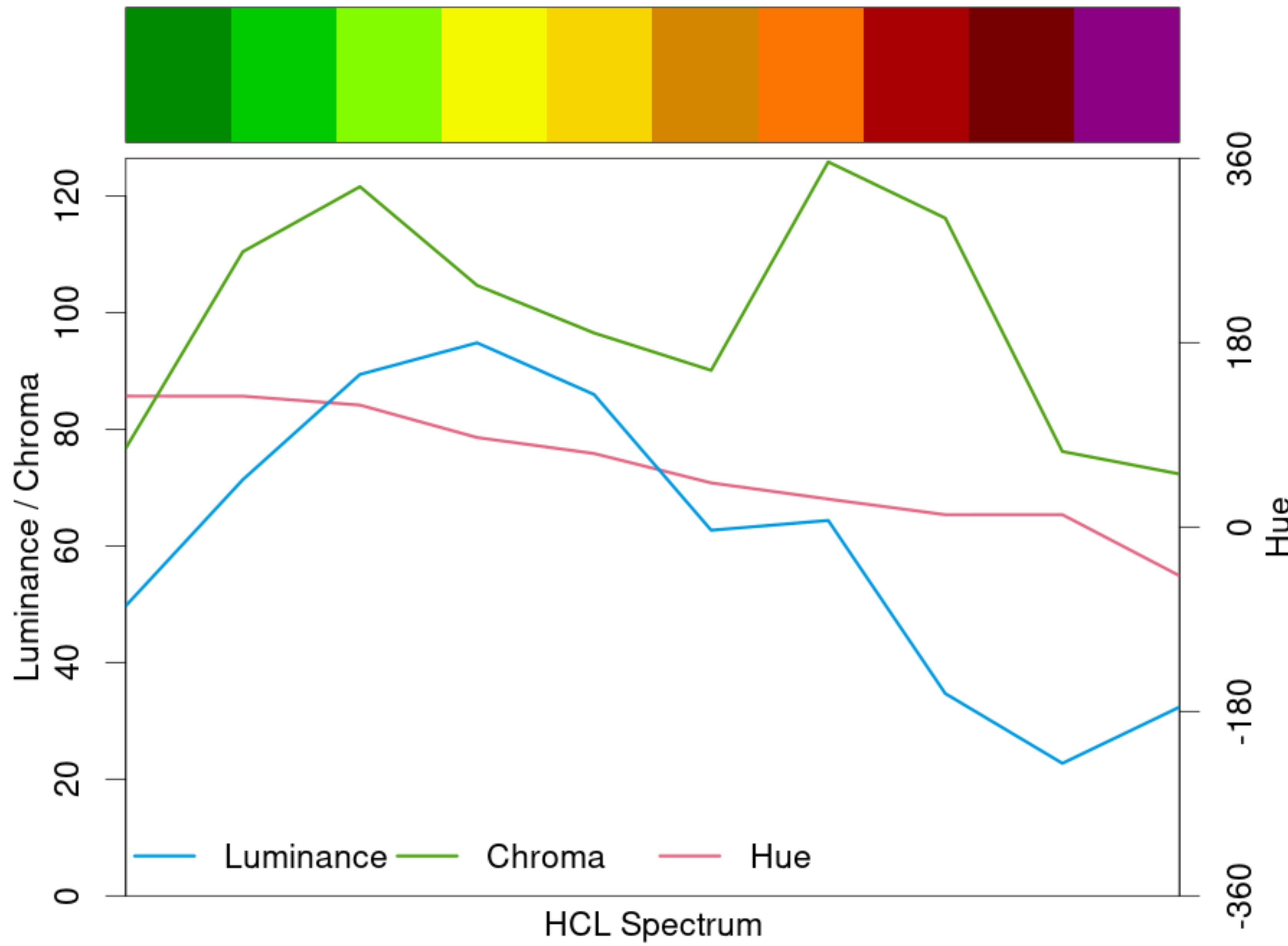


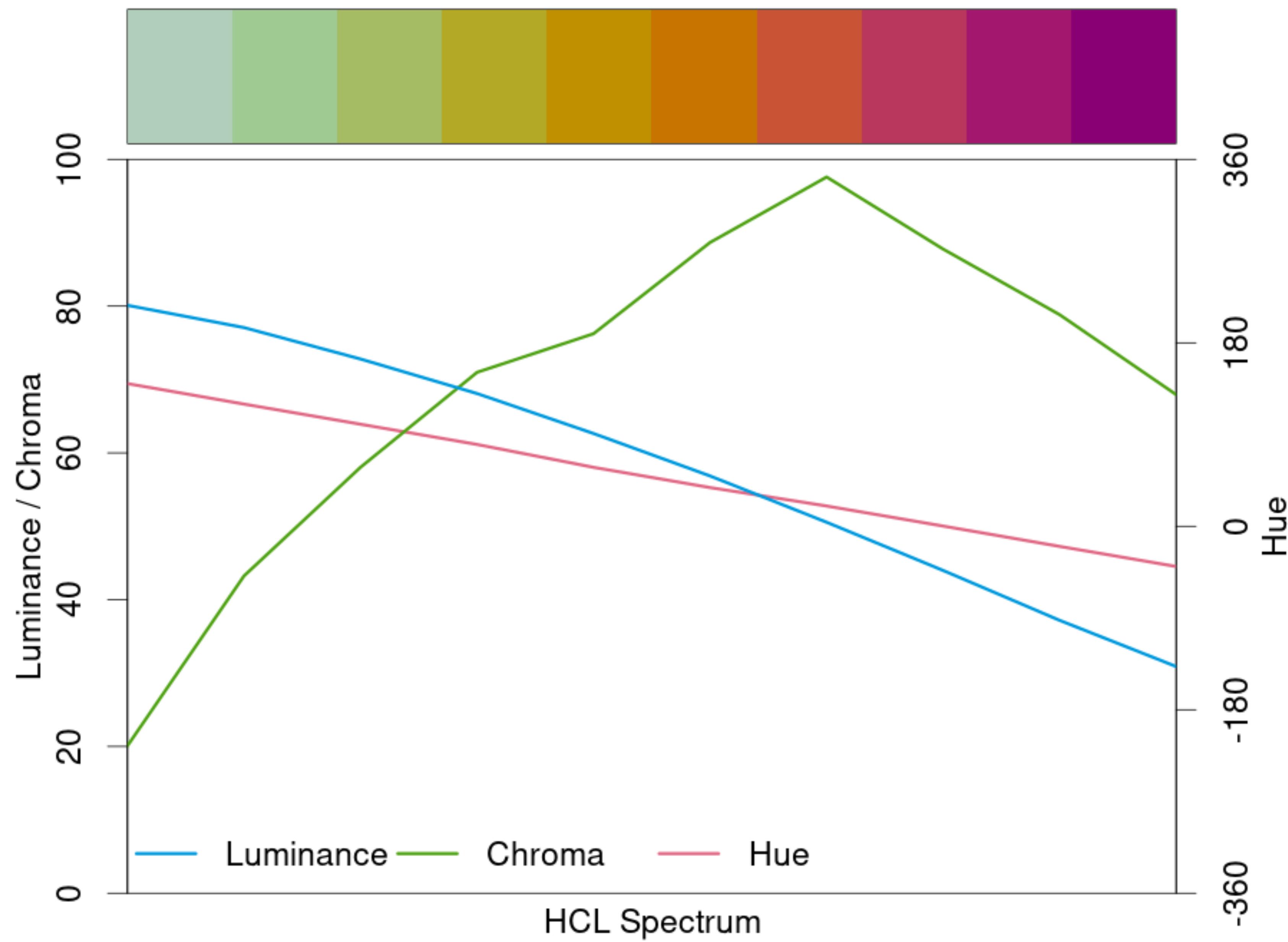
cividis

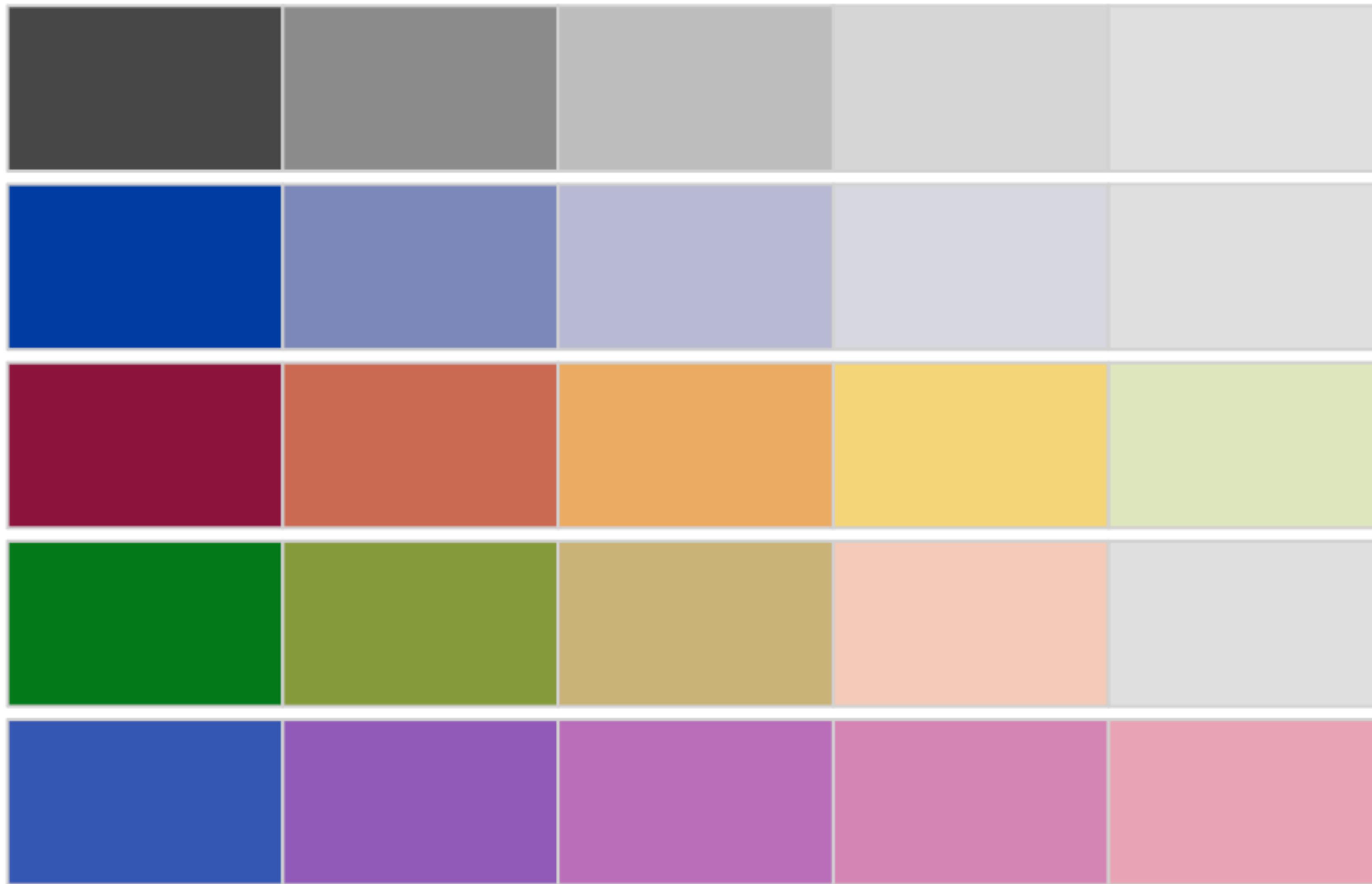


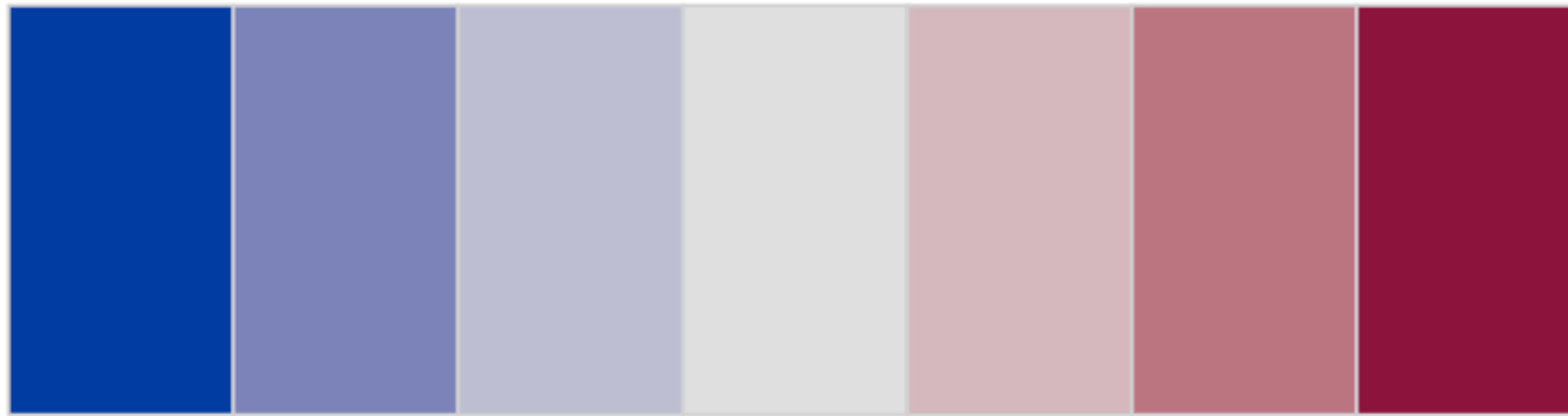








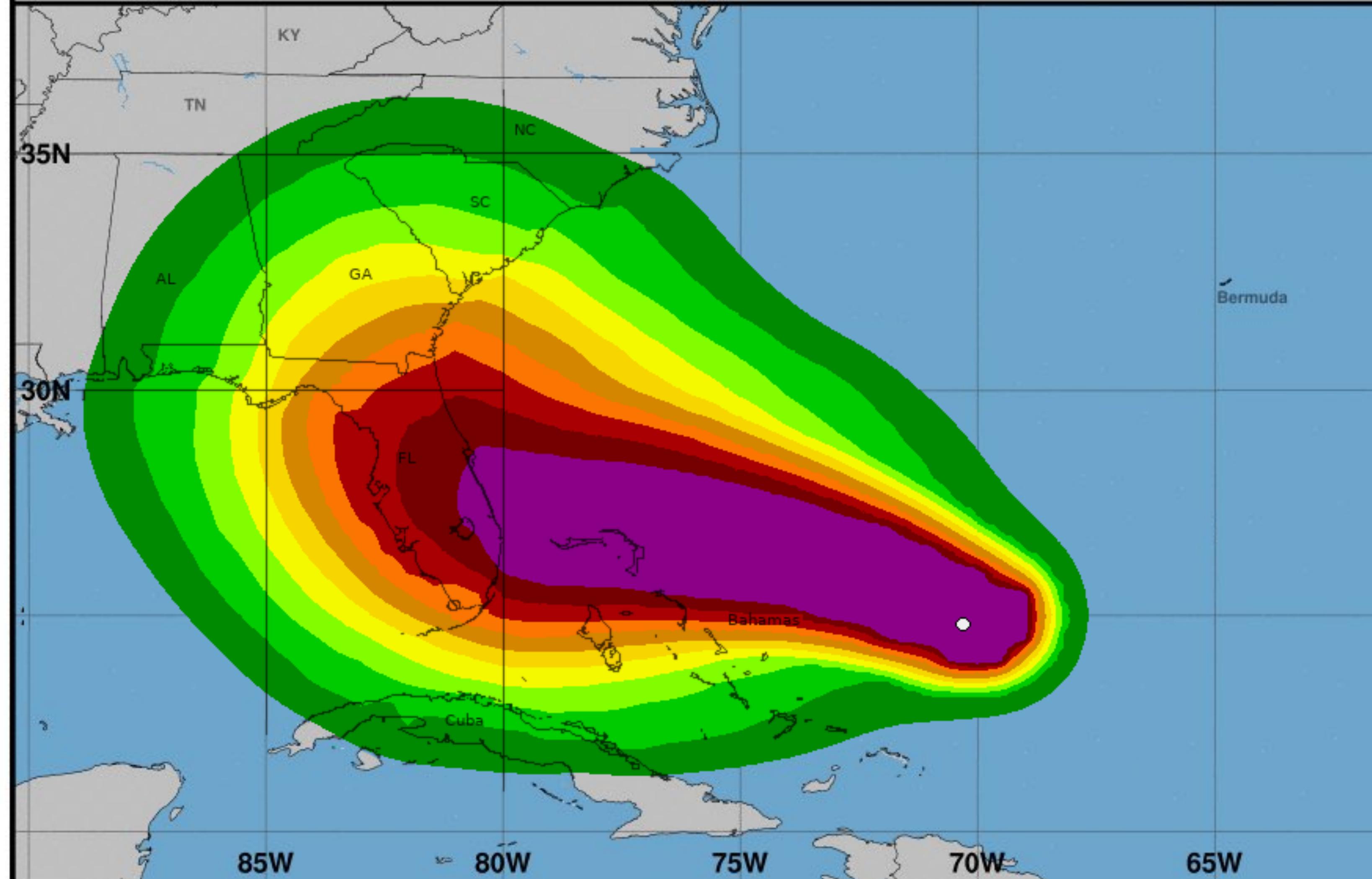






Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



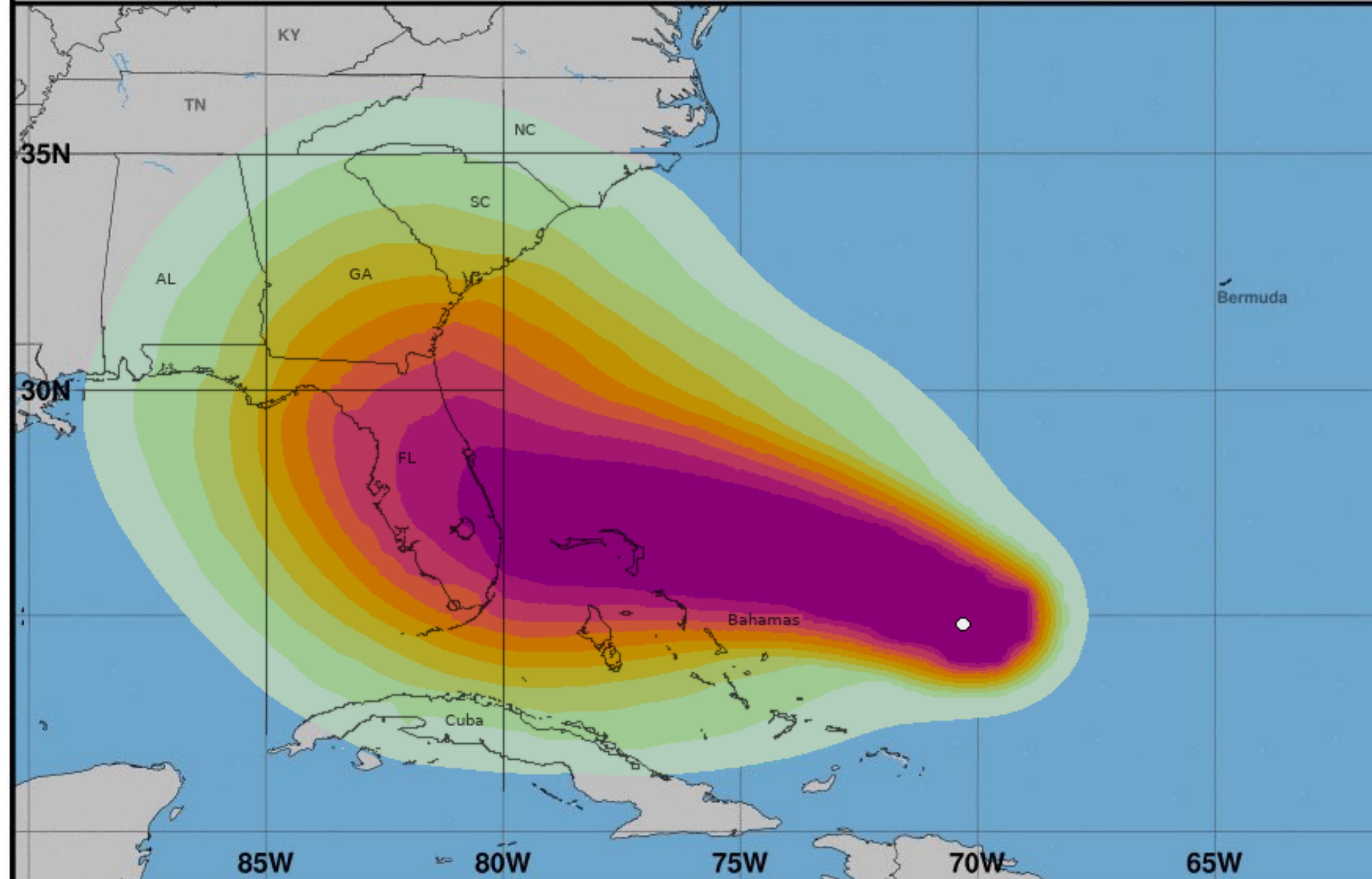
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
O indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)

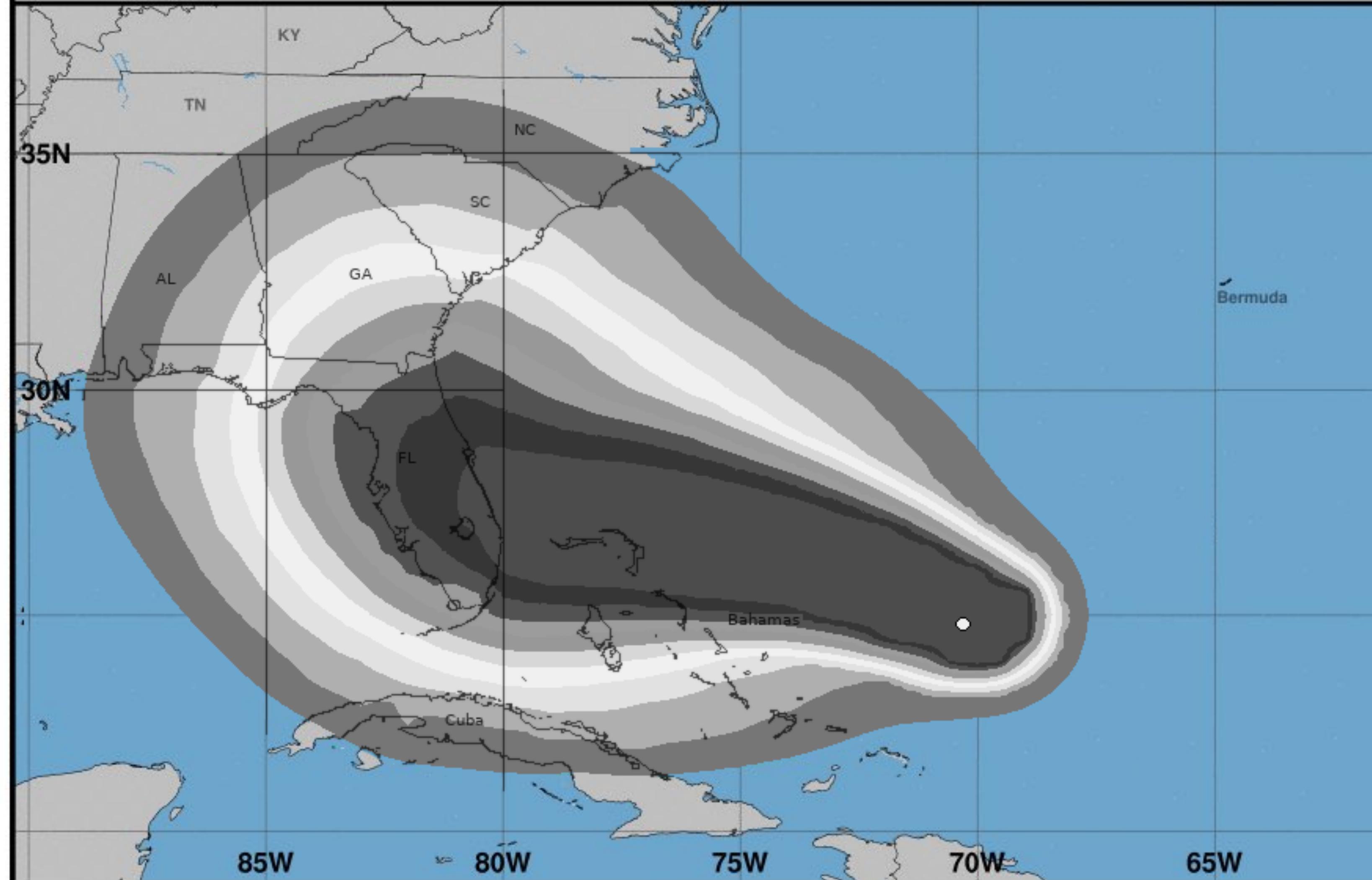


Achim Zeileis
<https://bit.ly/35zBIYL>



Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



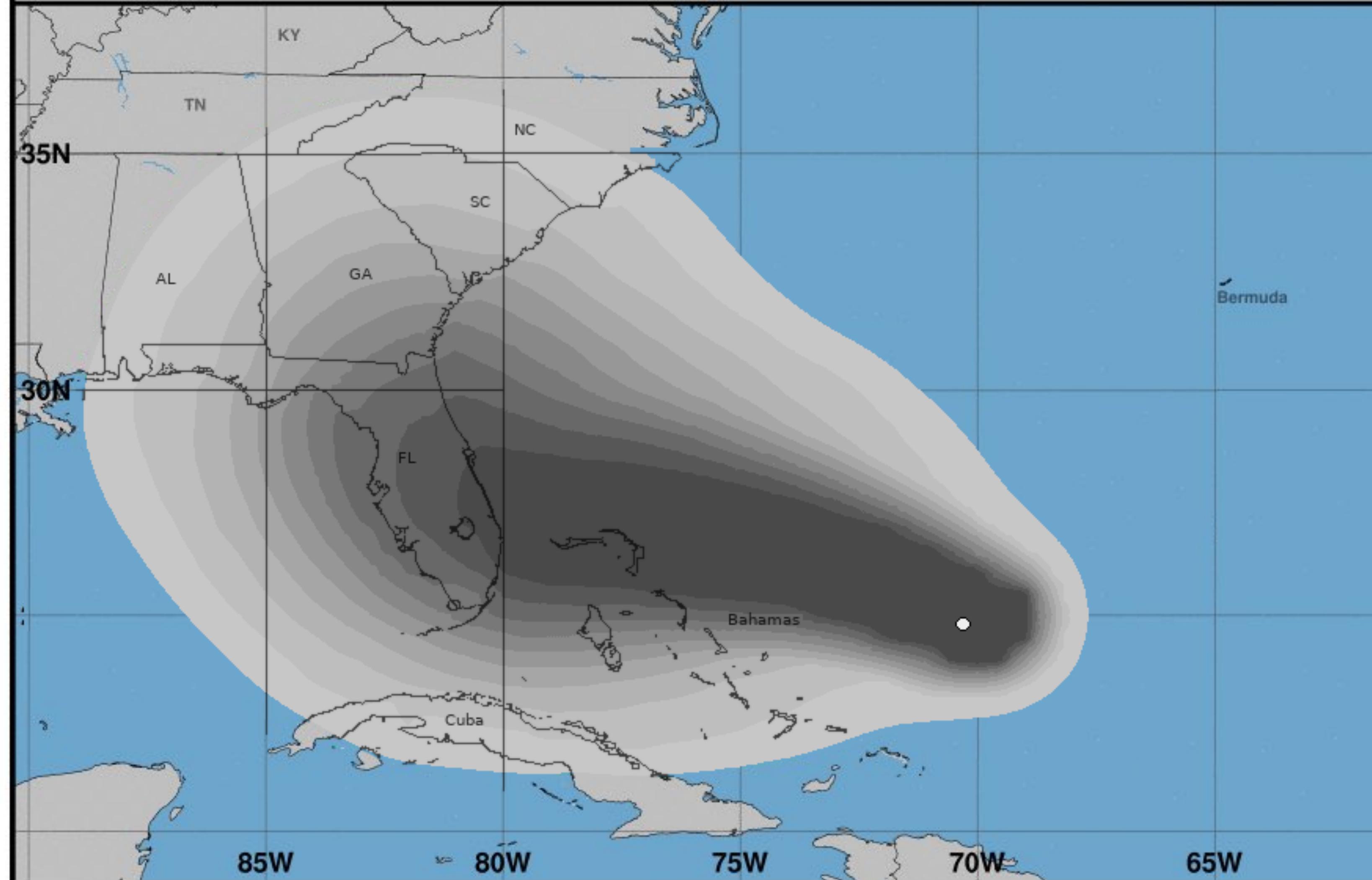
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



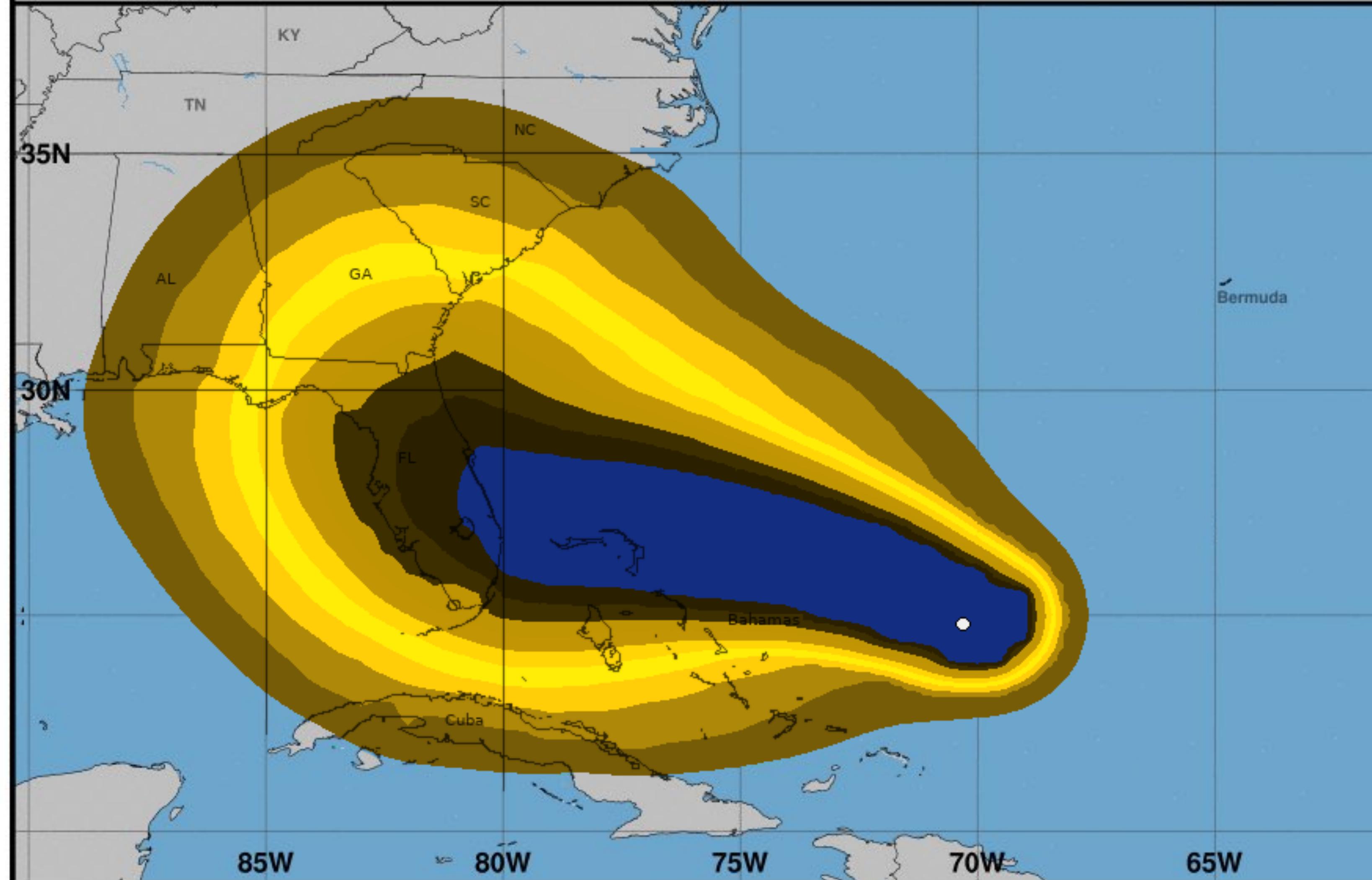
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



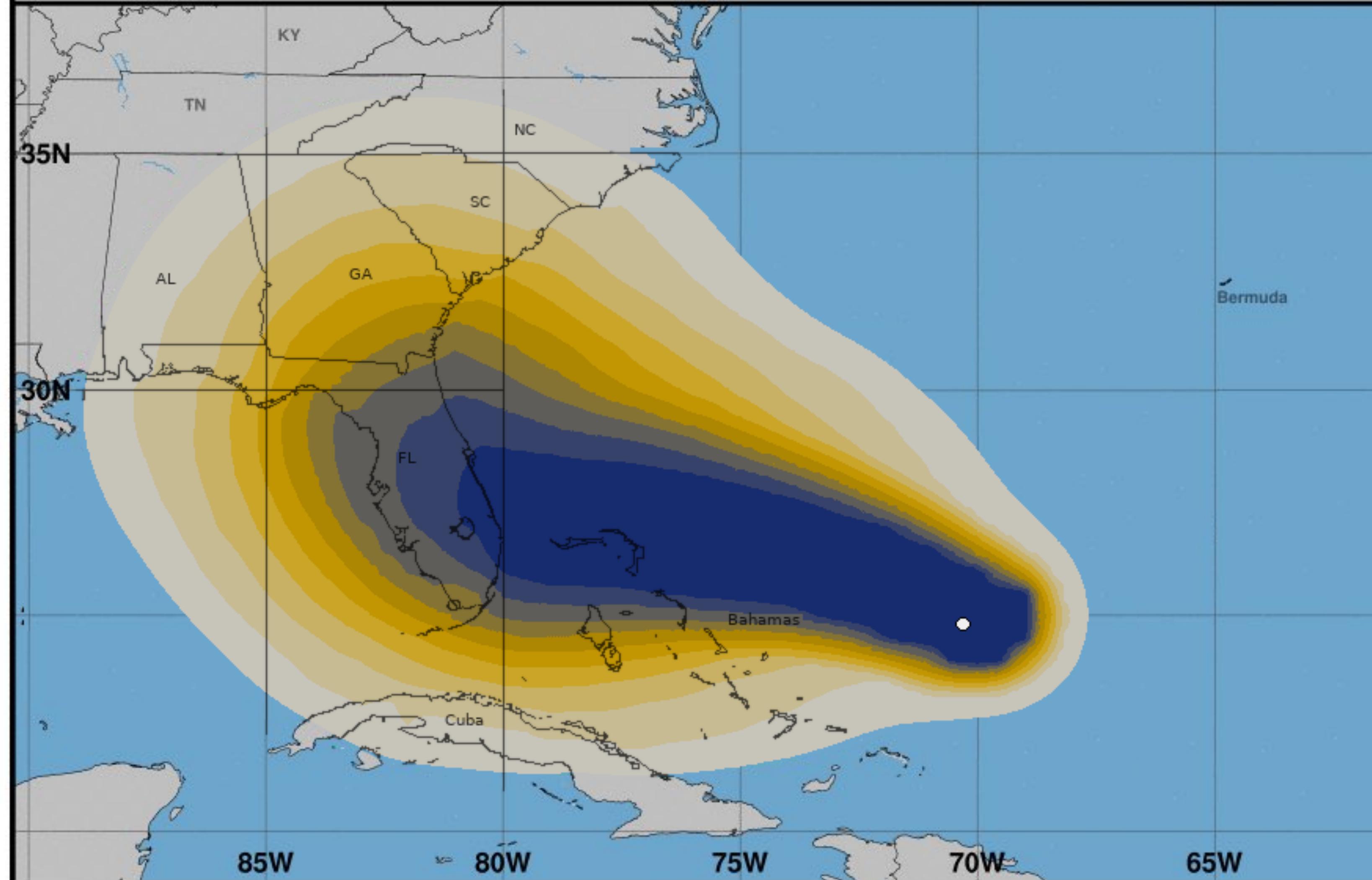
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





Tropical-Storm-Force Wind Speed Probabilities

For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04

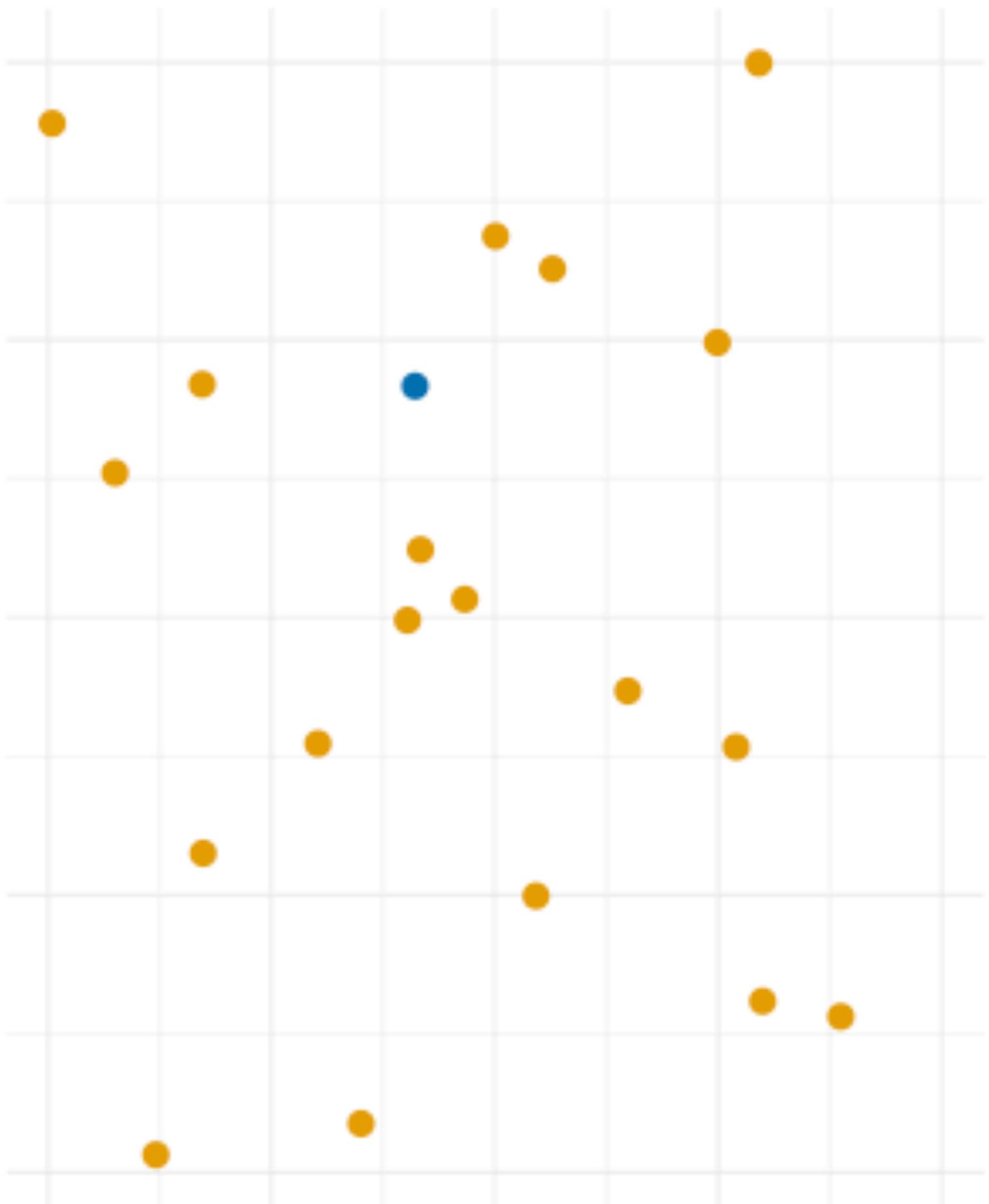


Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)

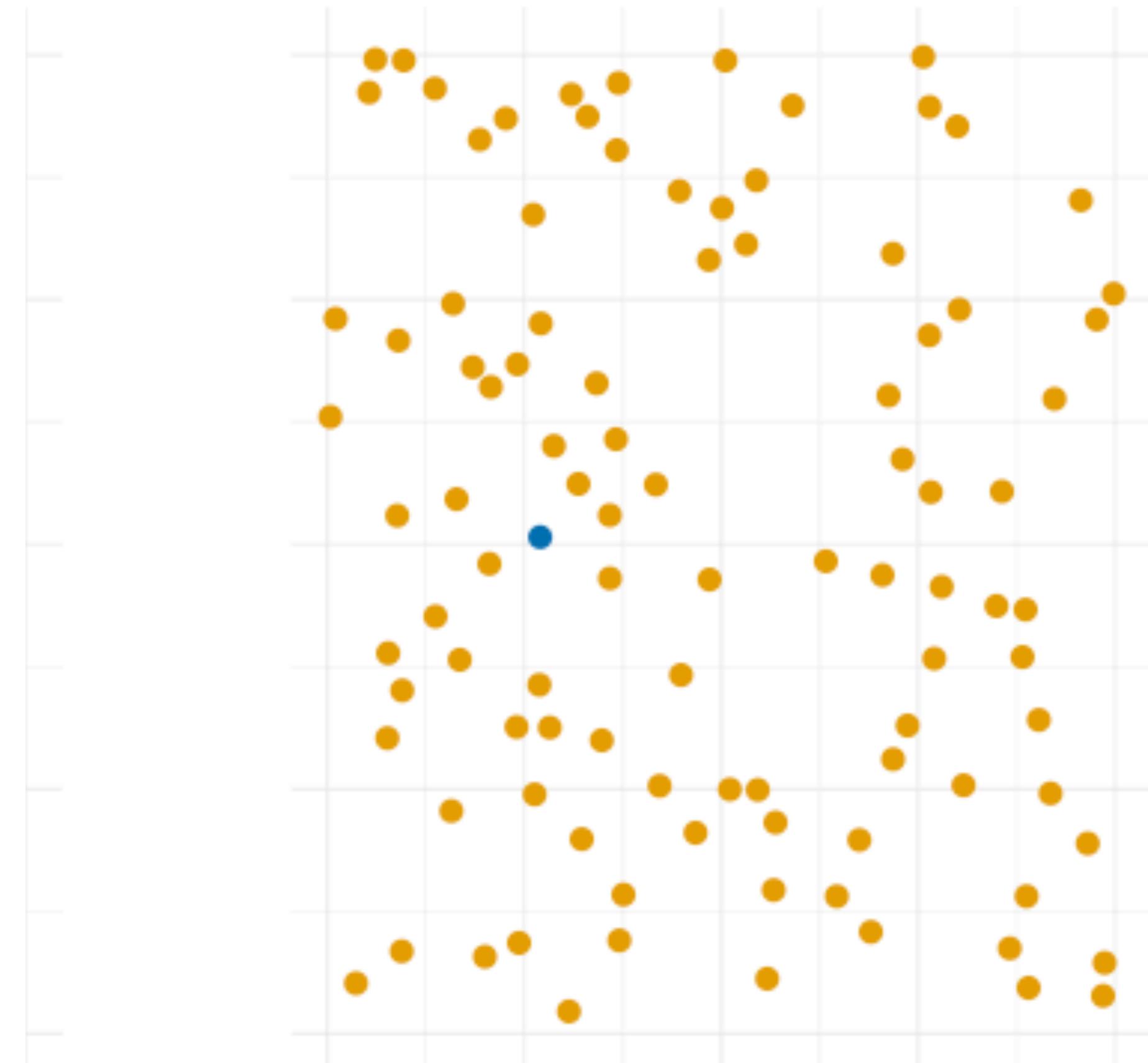


PRE-ATTENTIVE PROCESSING

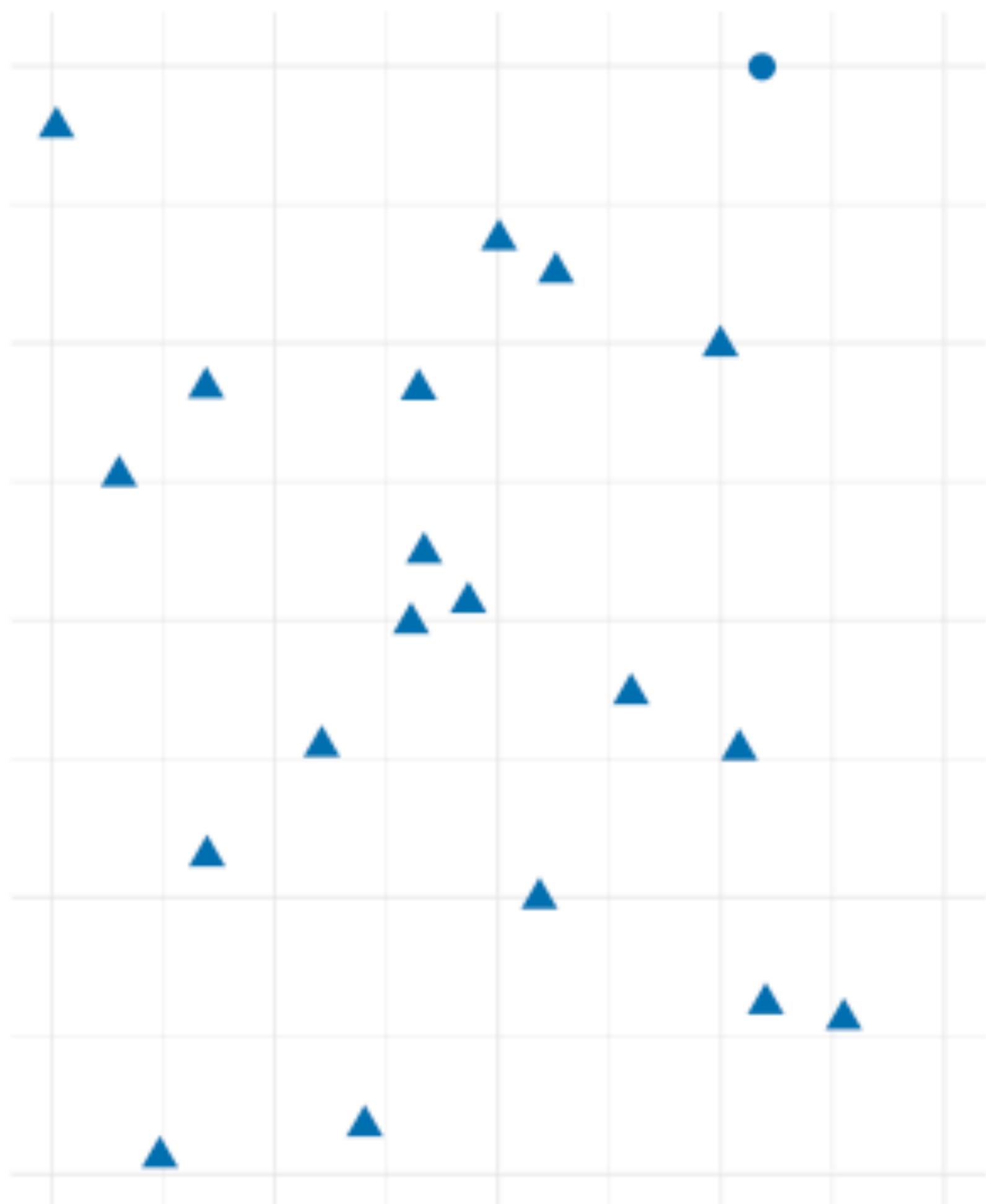
Color Only, N=20



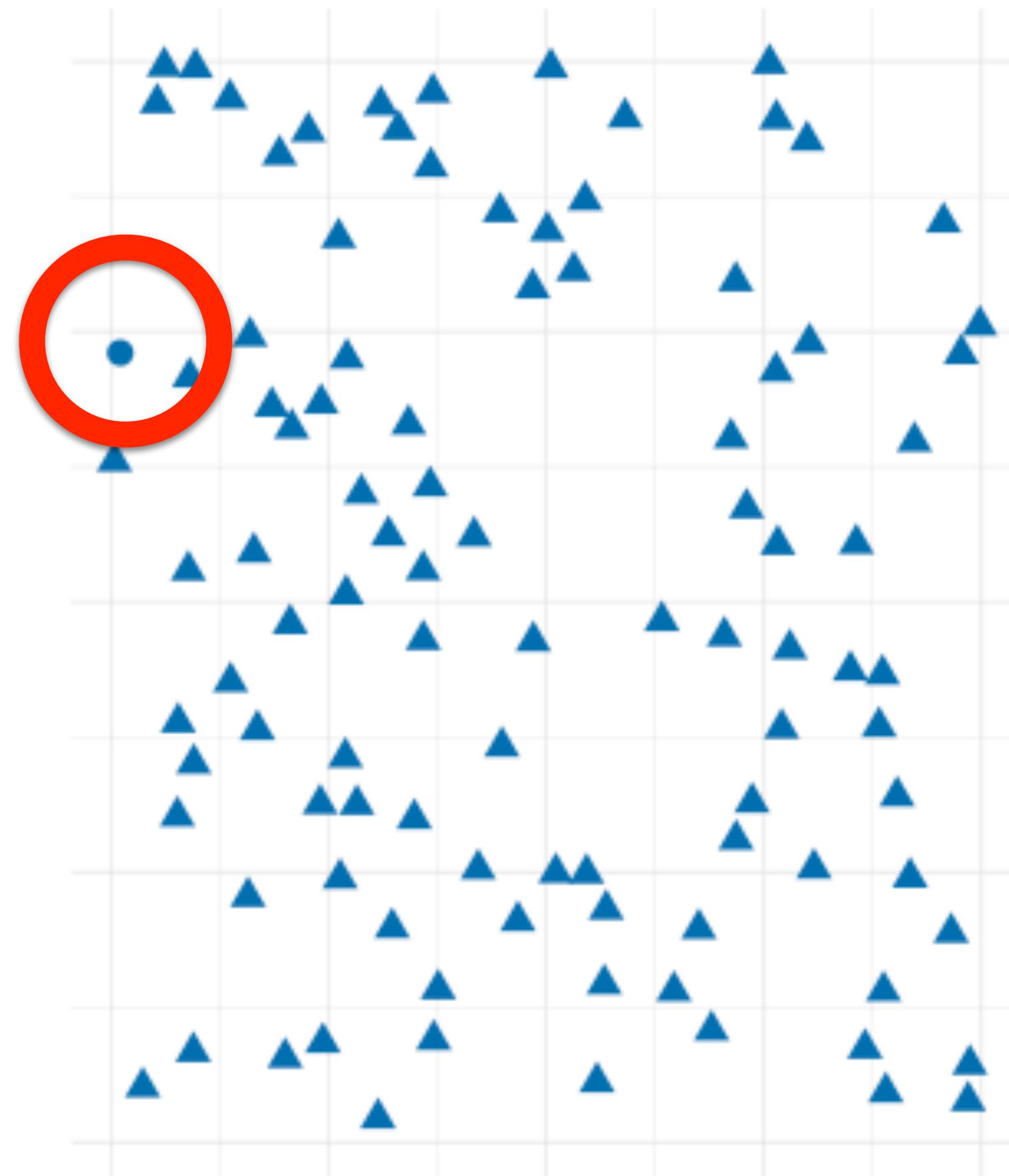
Color Only, N=100



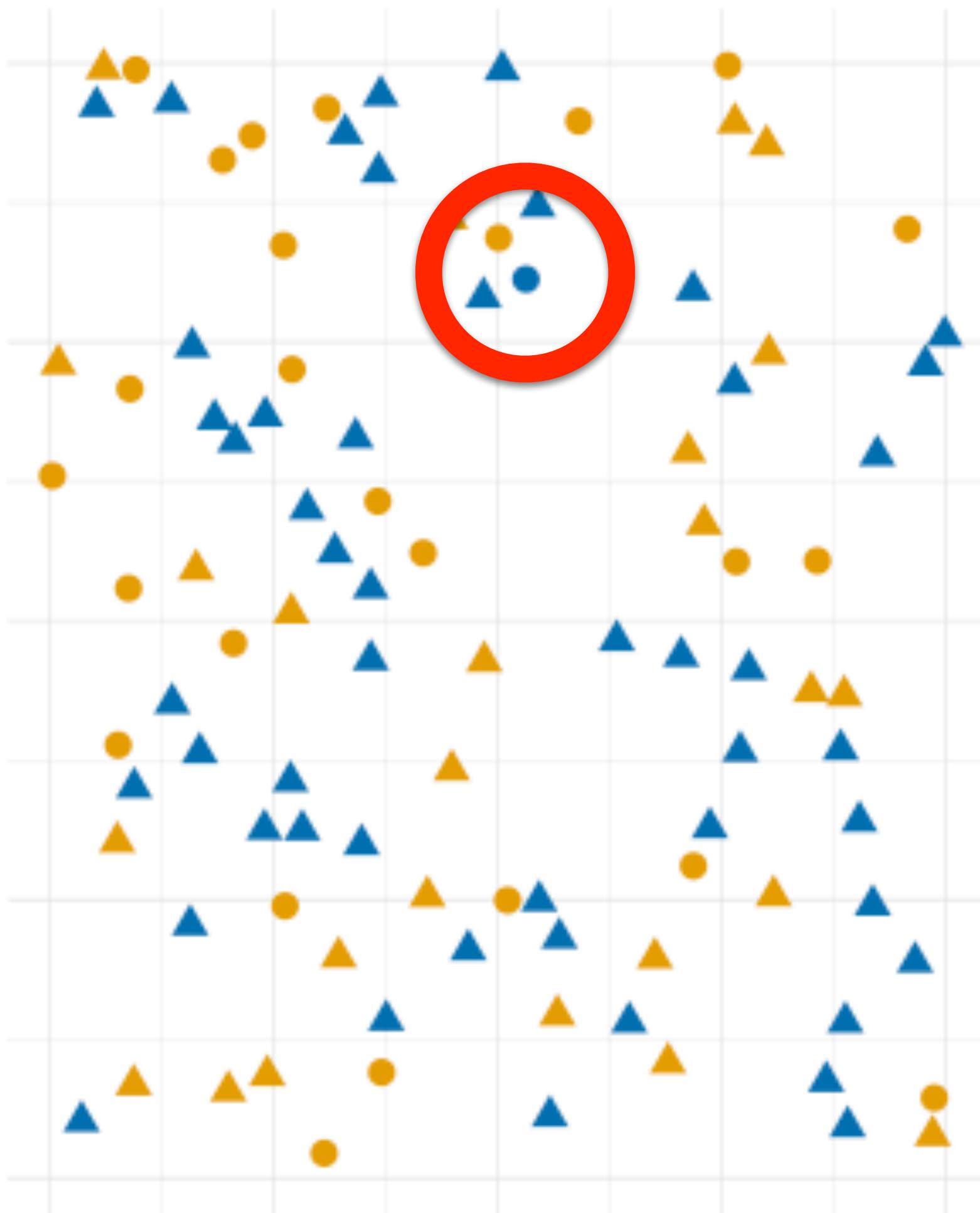
Shape Only, N=20

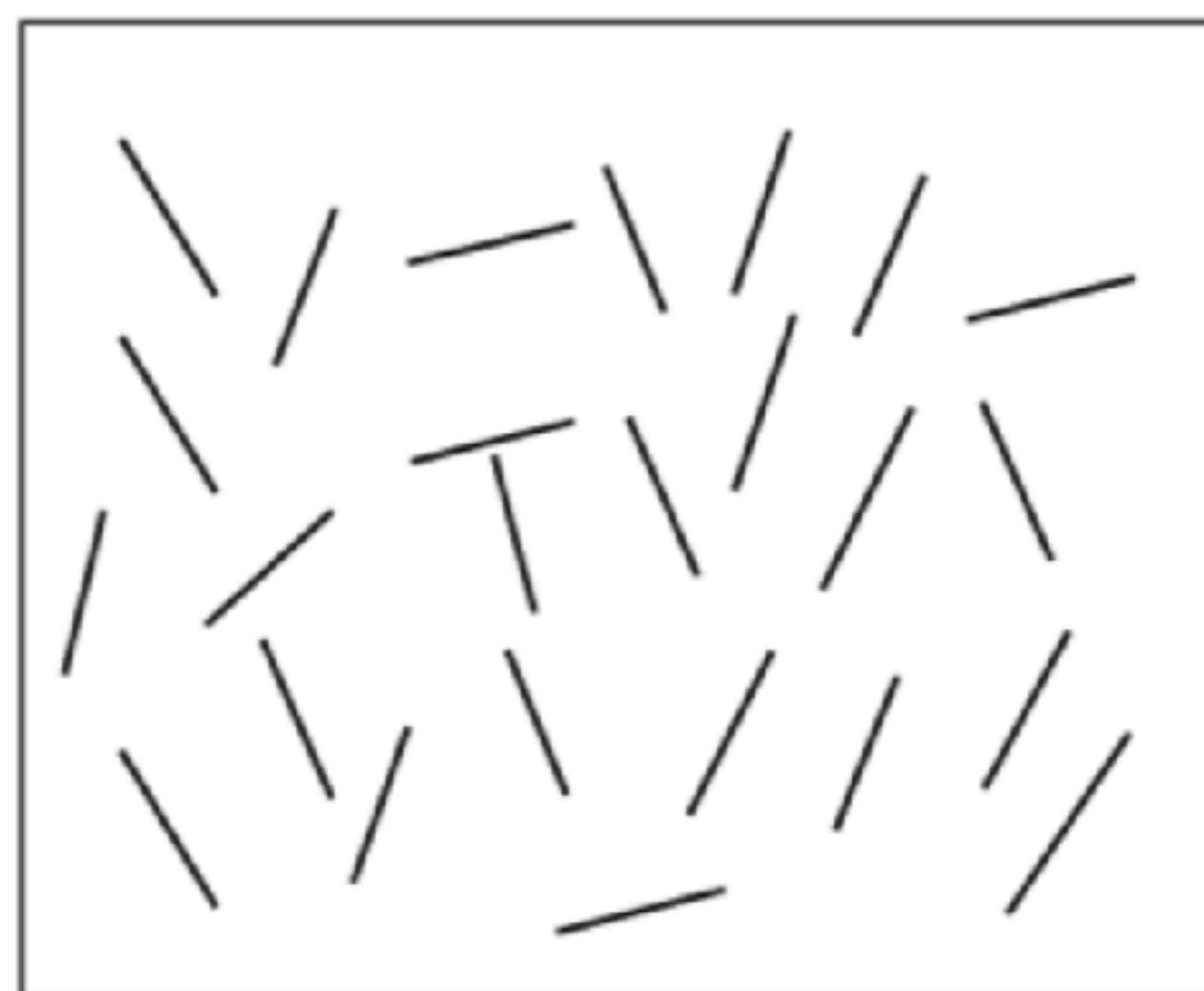
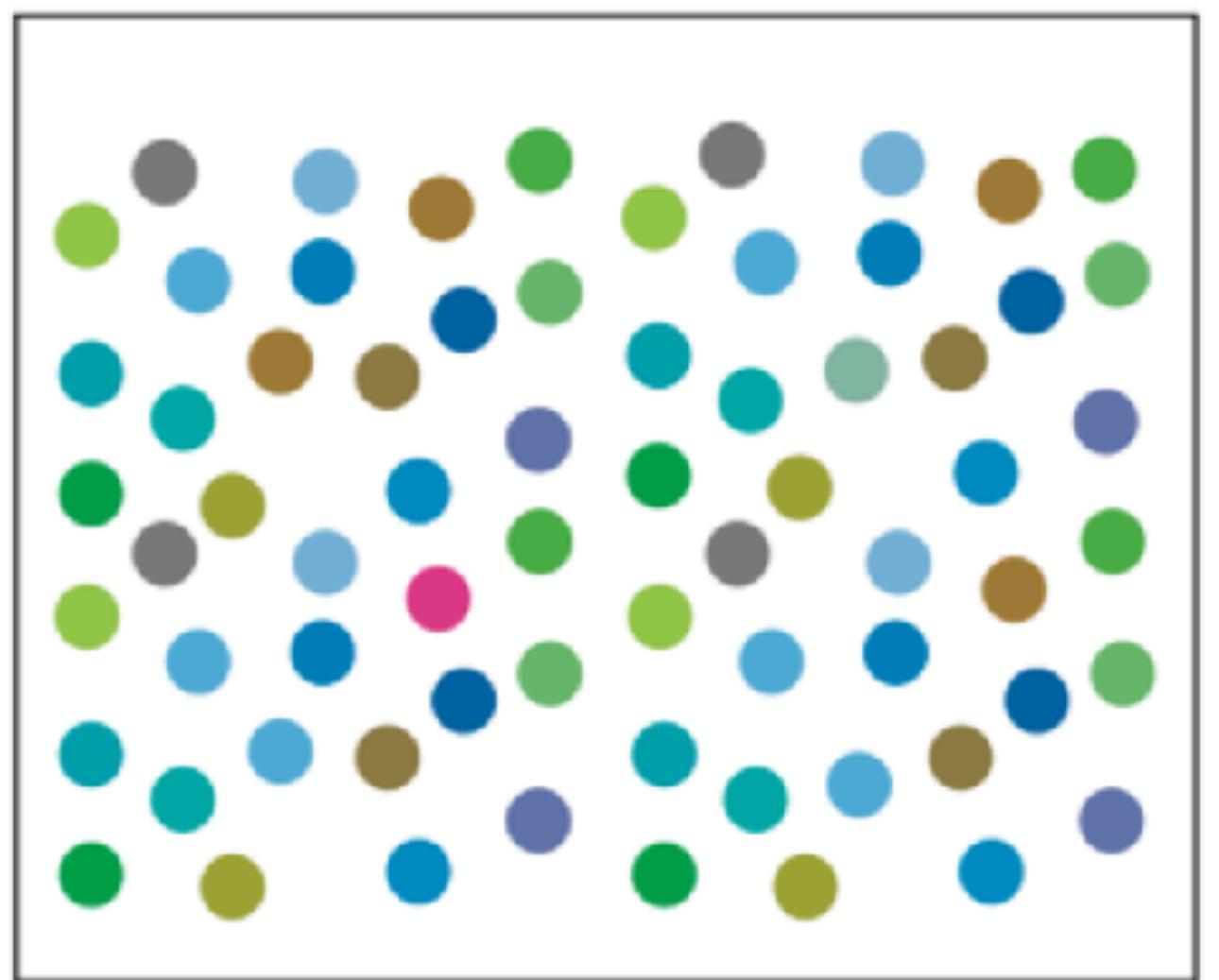
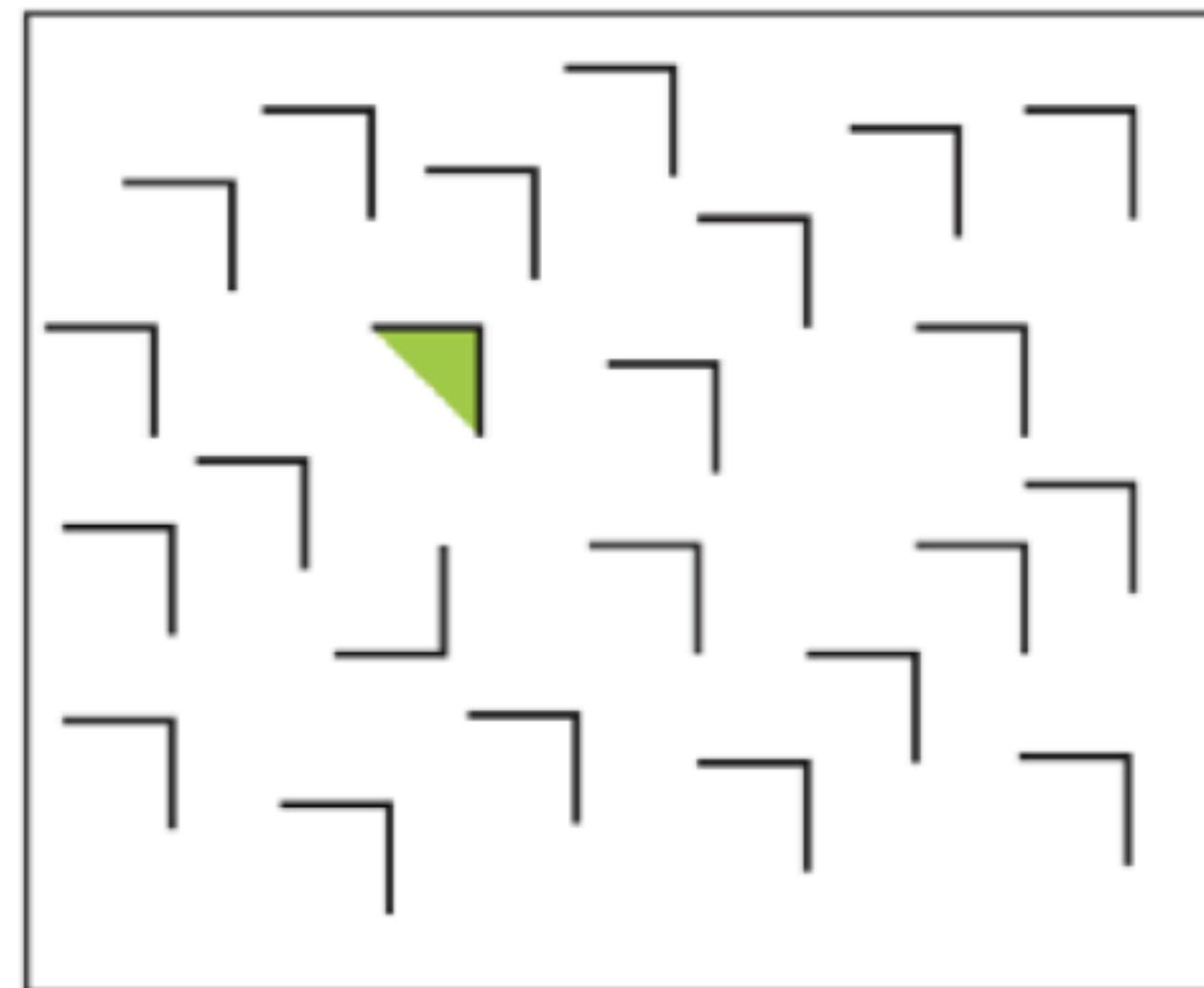
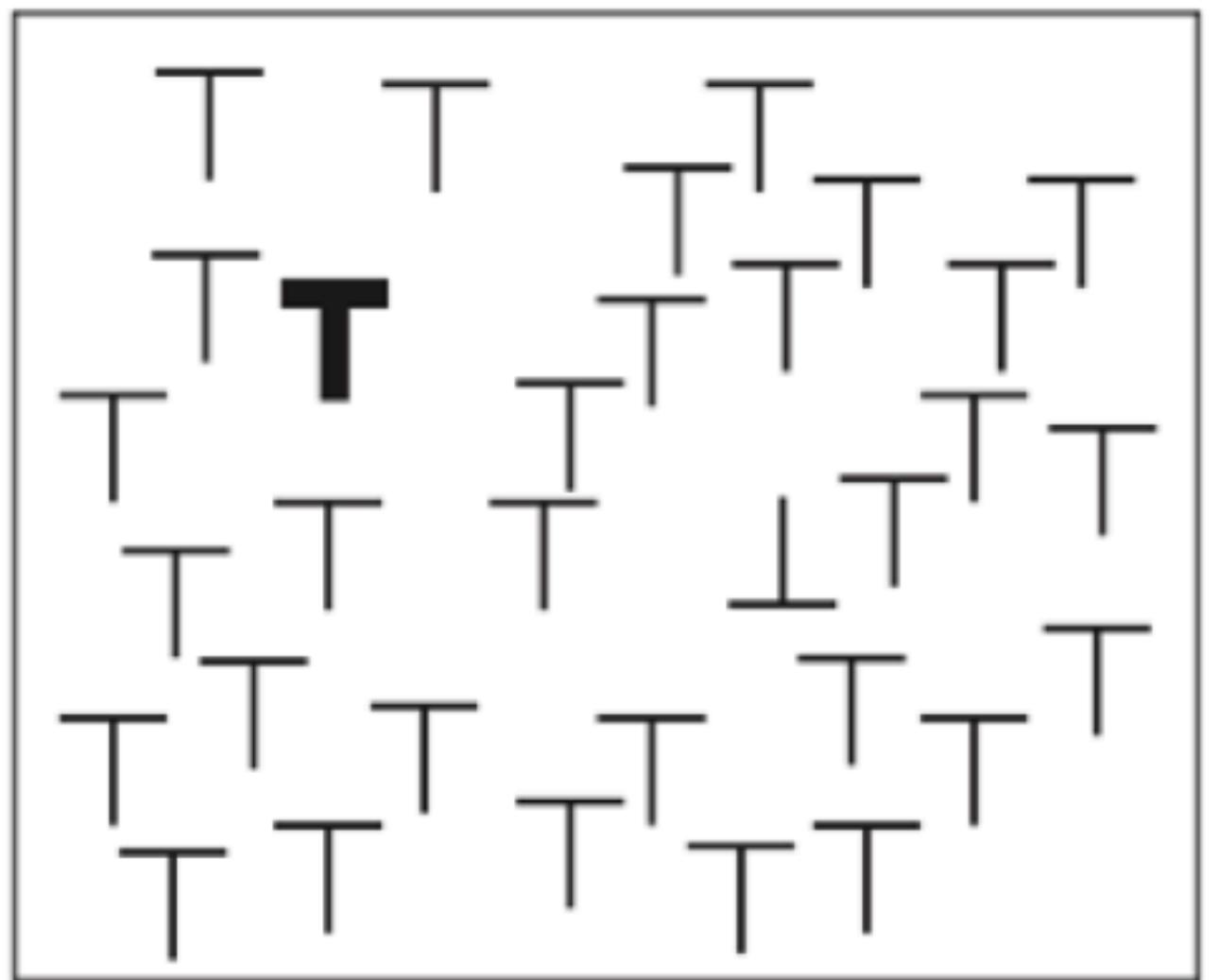


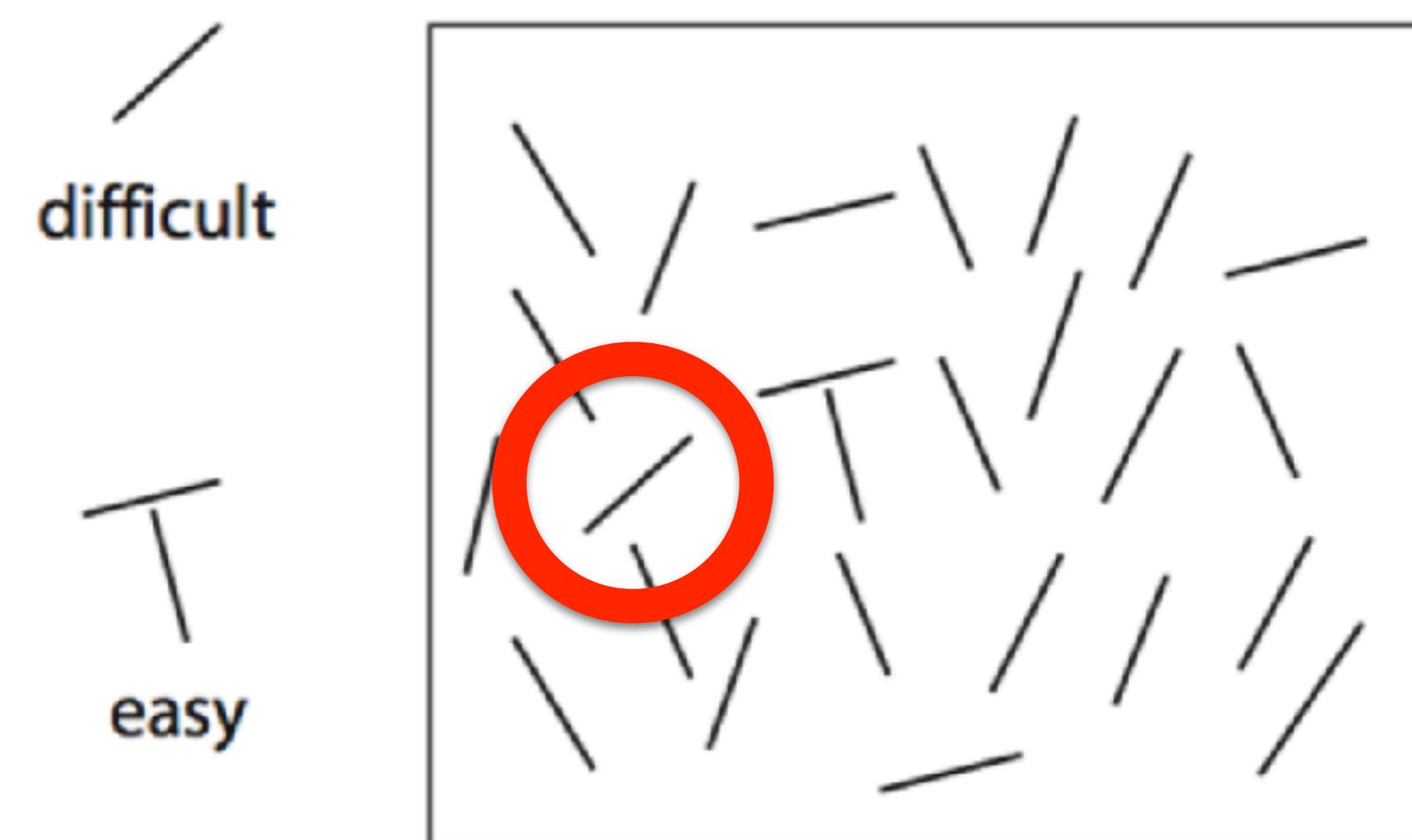
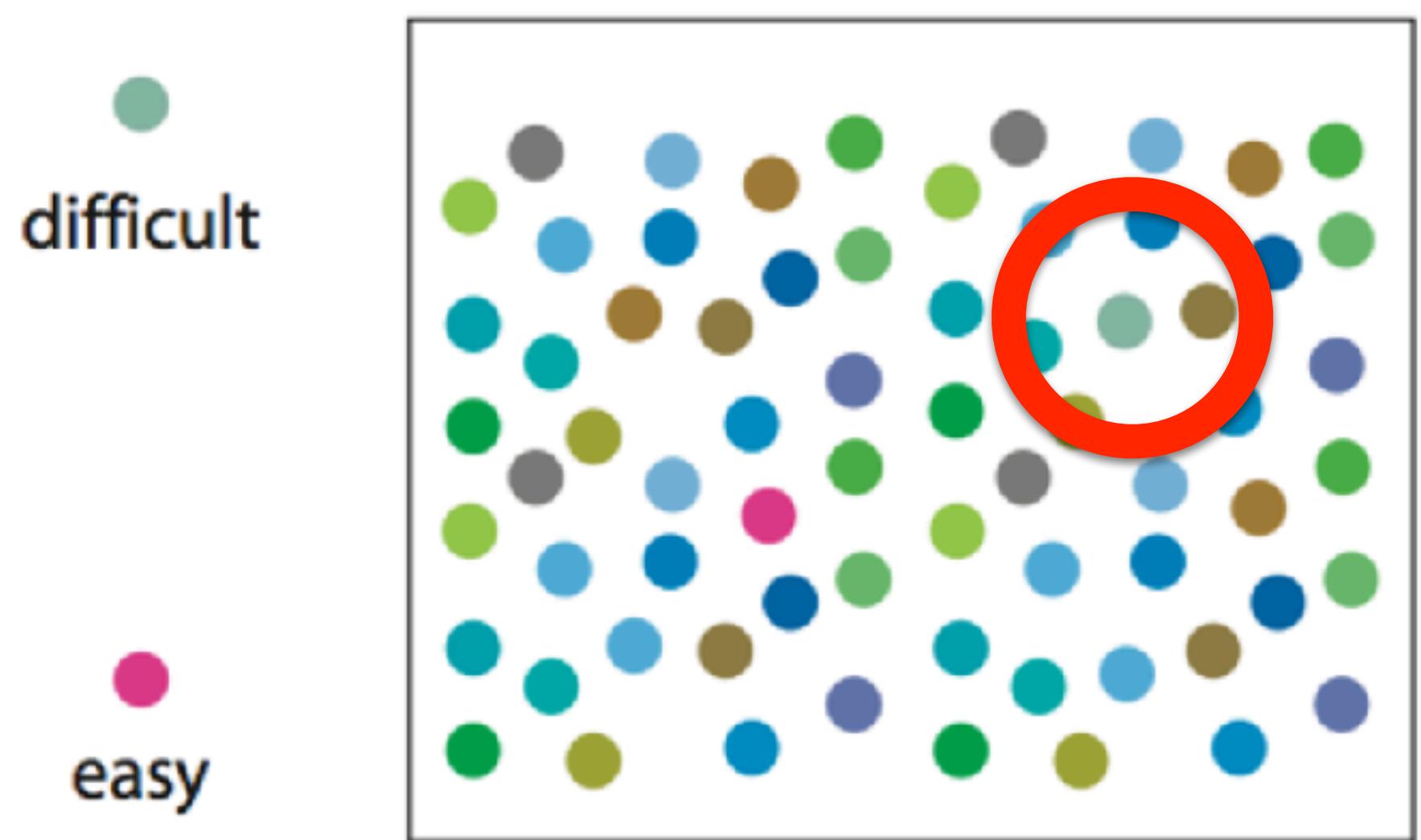
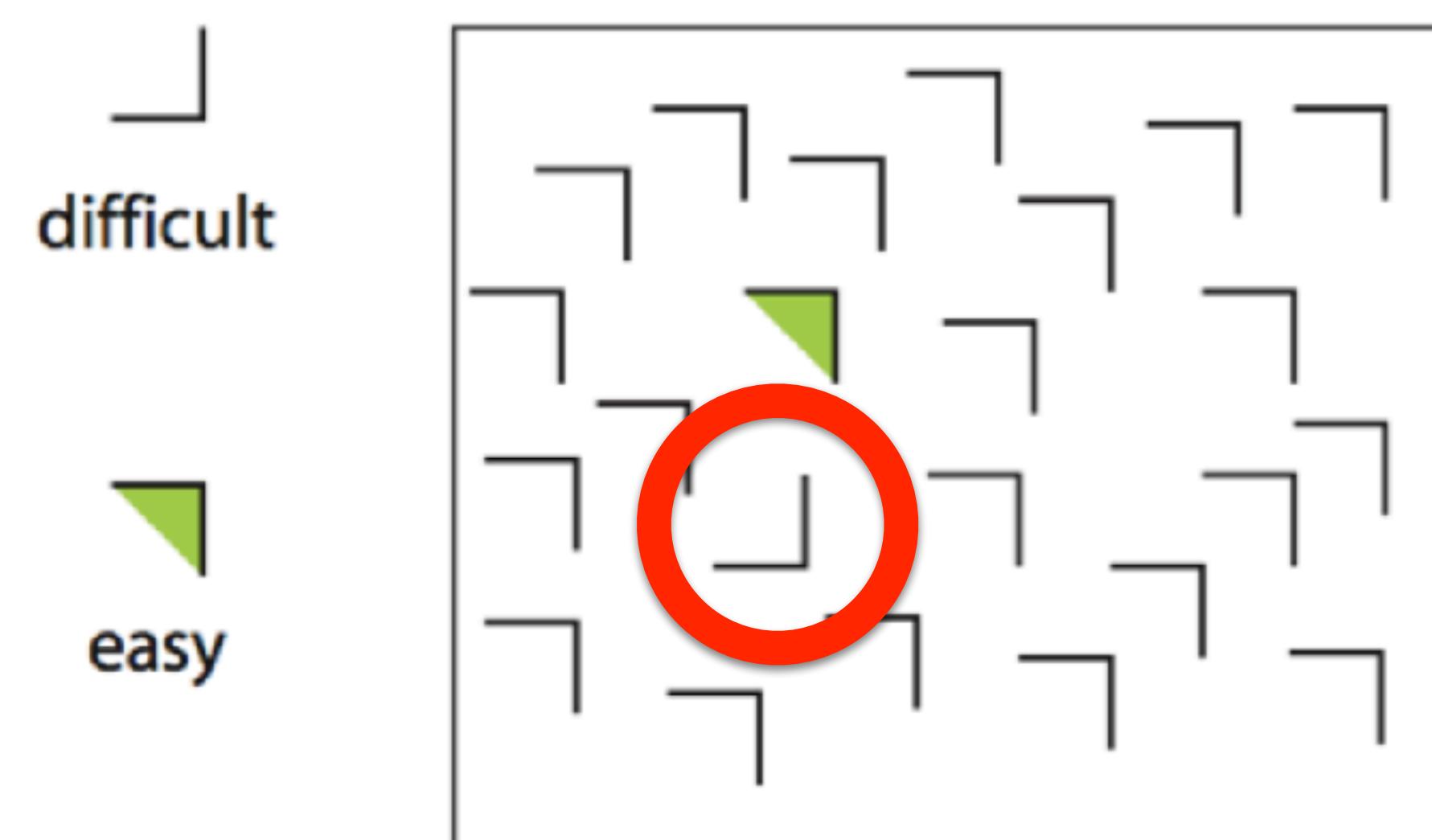
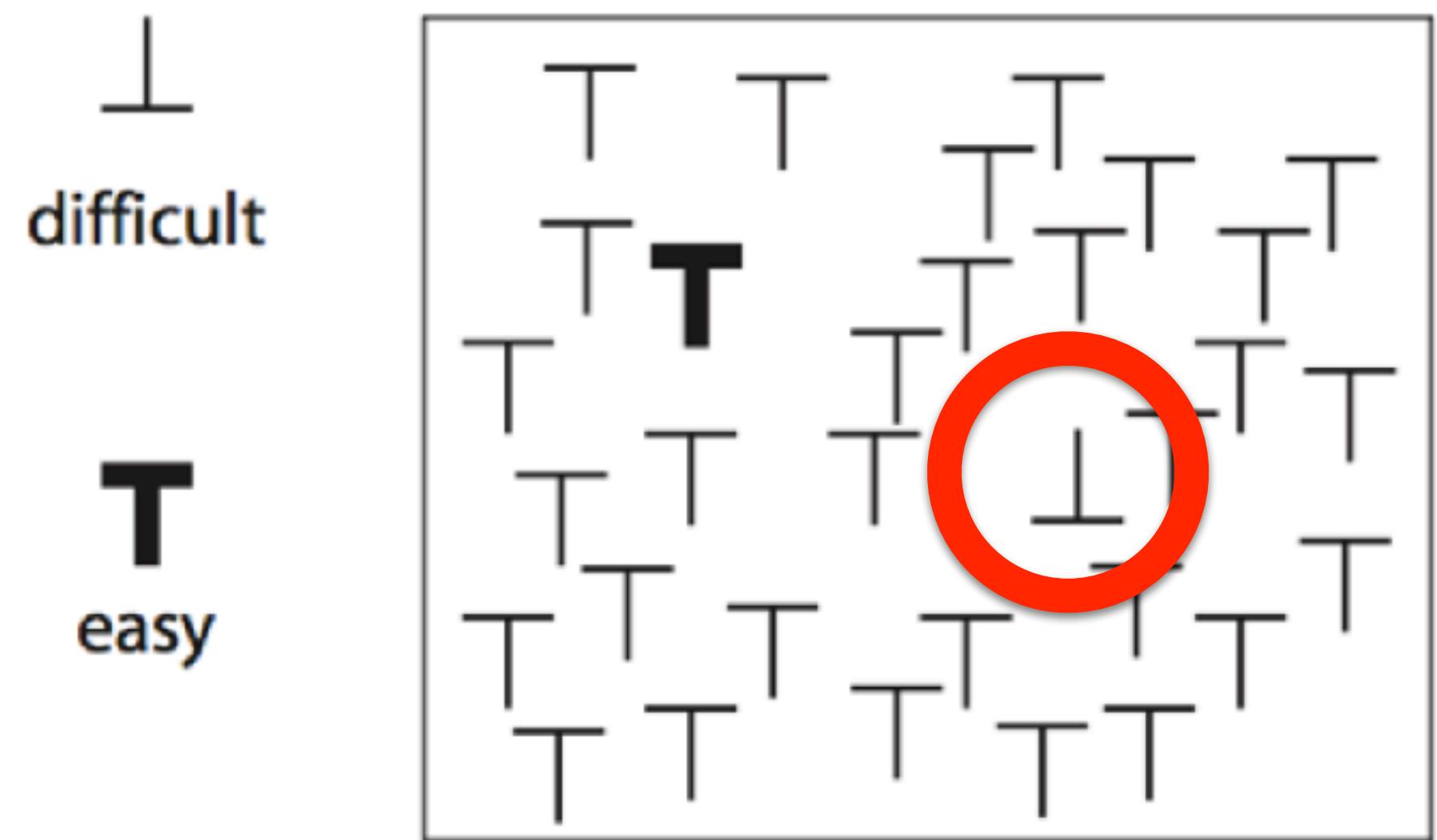
Shape Only, N=100

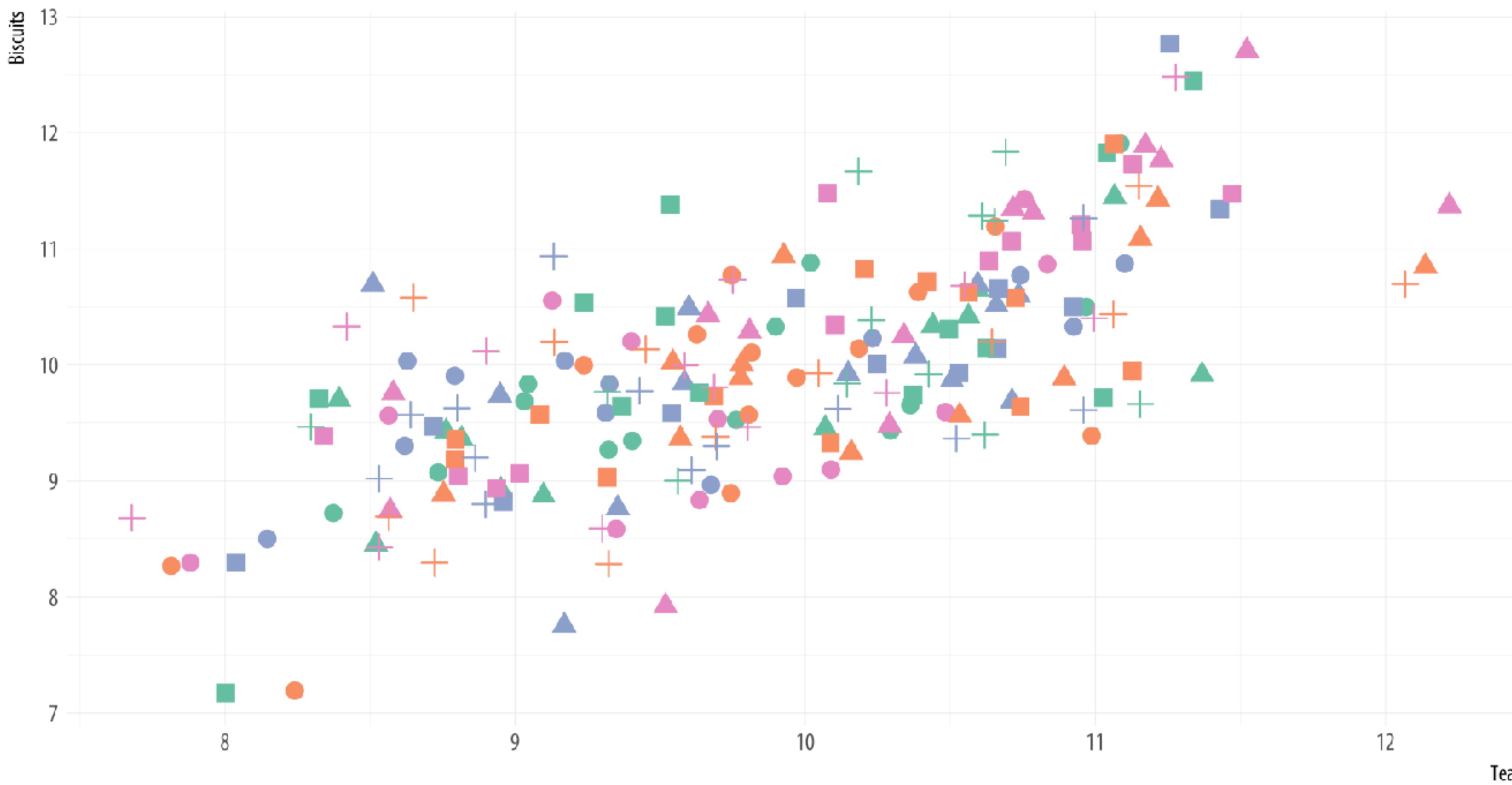


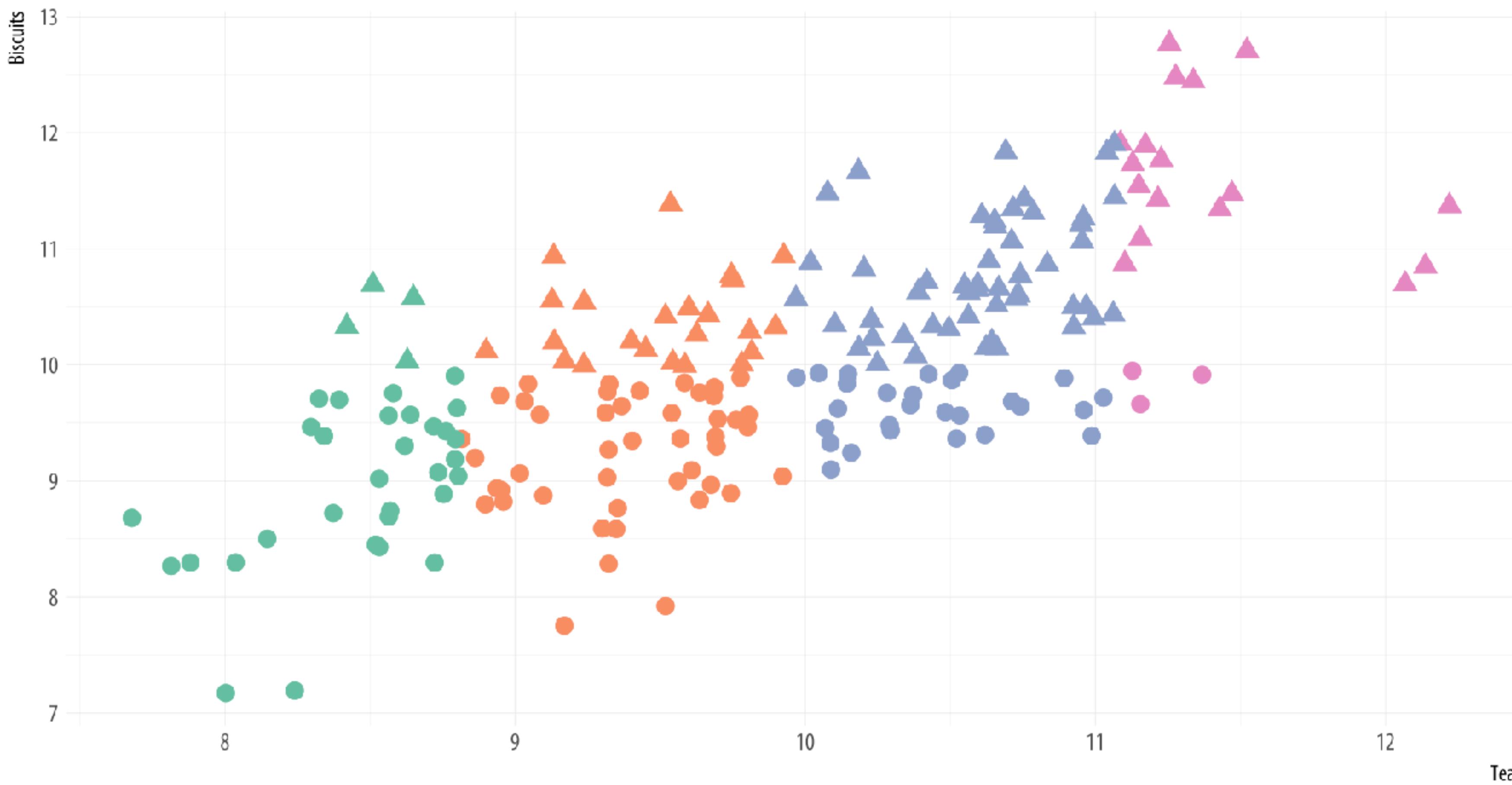
Color & Shape, N=100







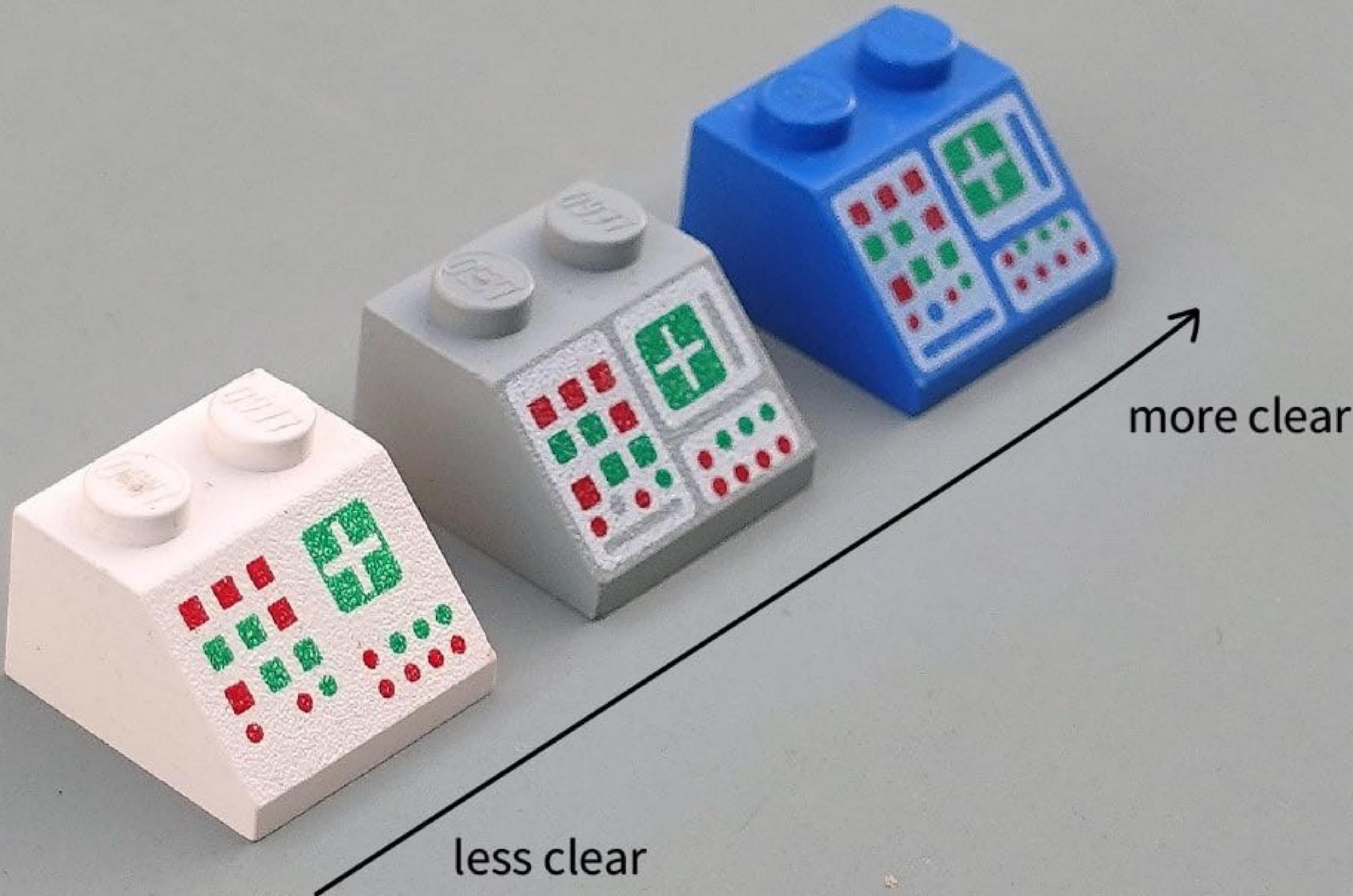




GESTALT INFERENCES



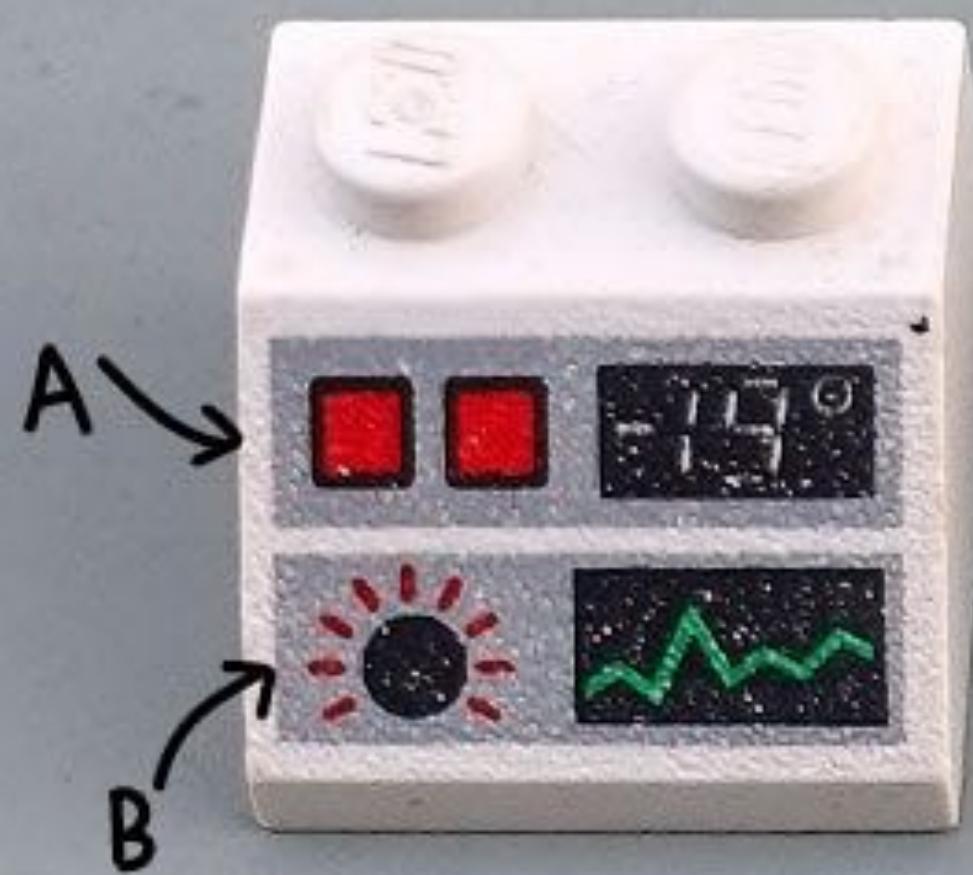
George Cave



less clear

more clear

by feature



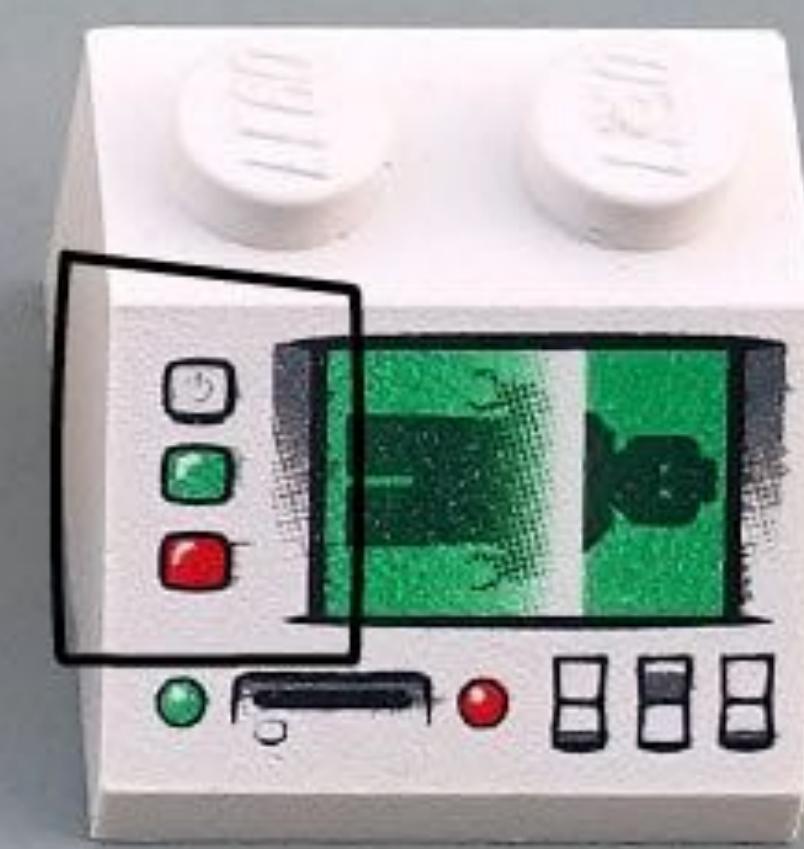
by operation

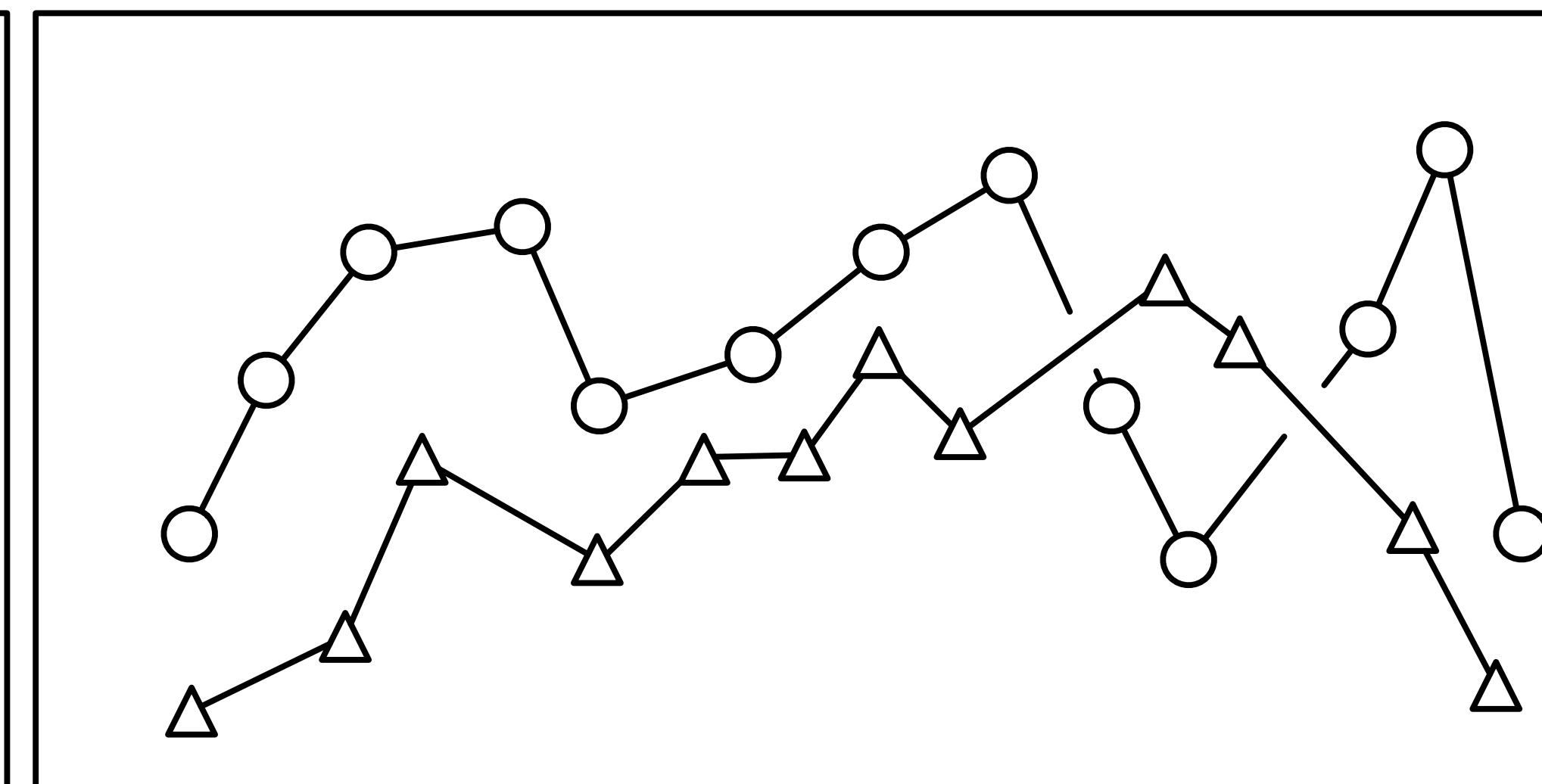
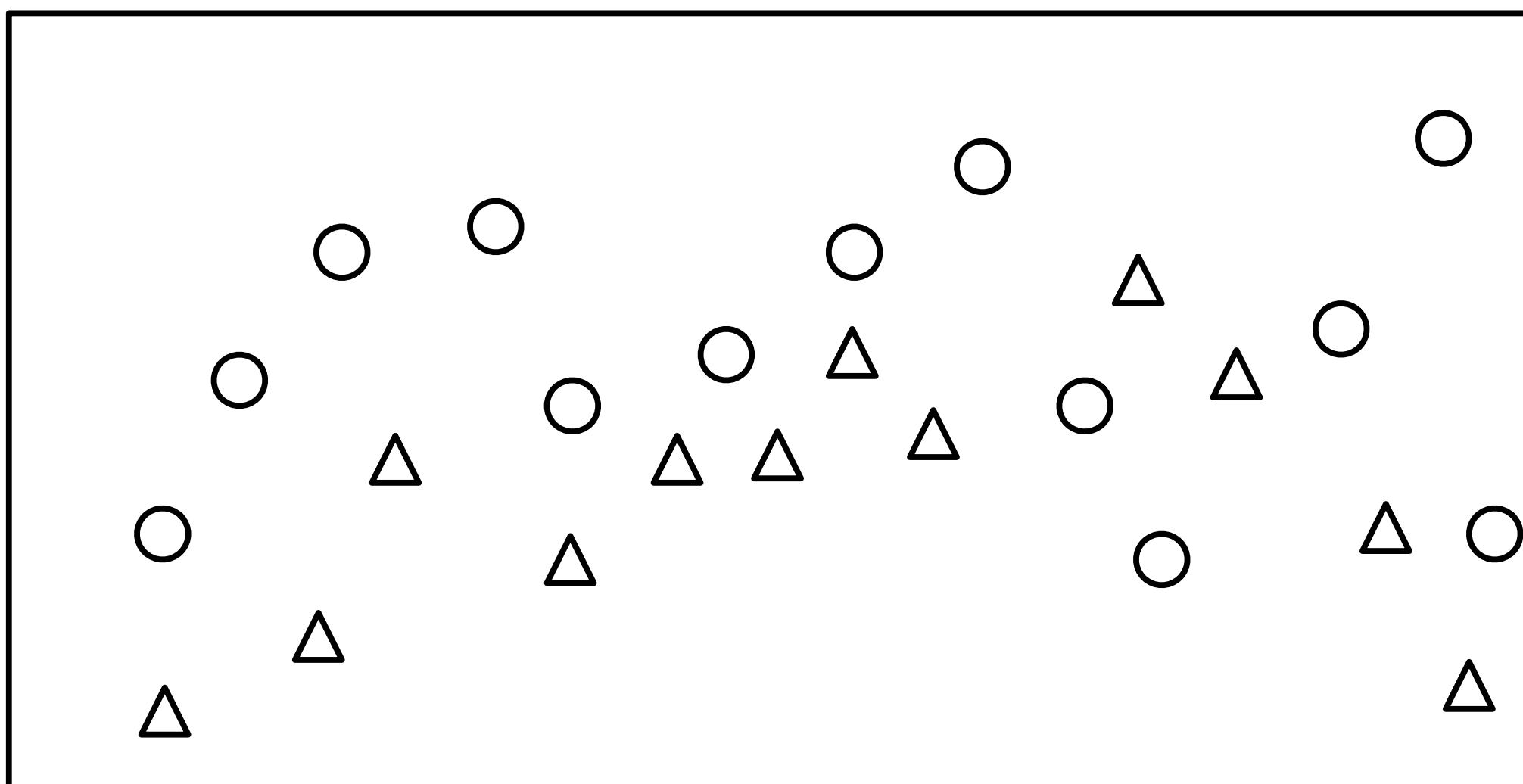
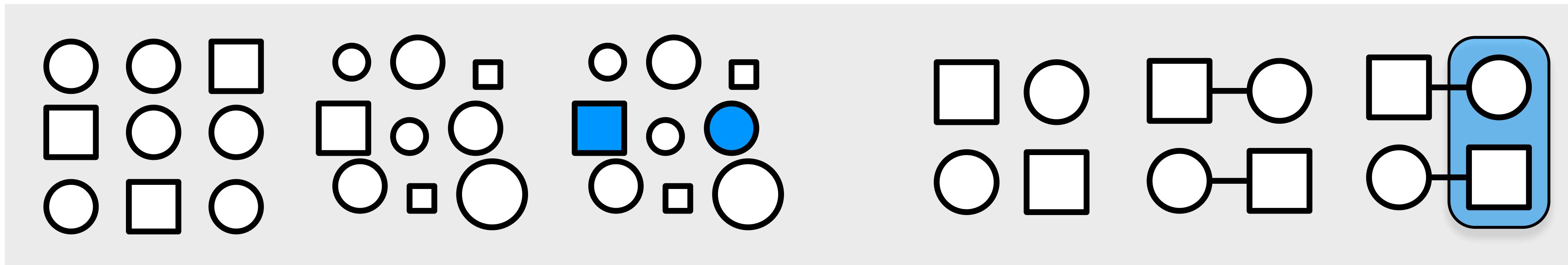
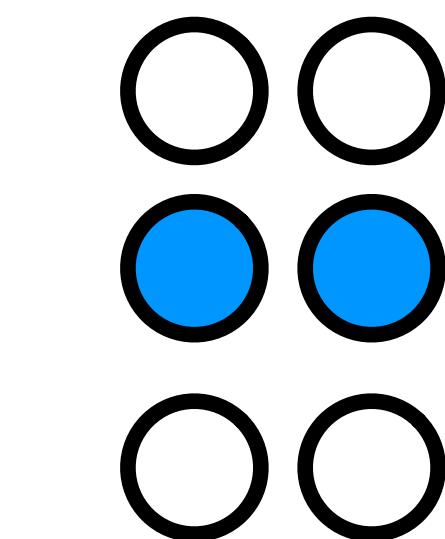
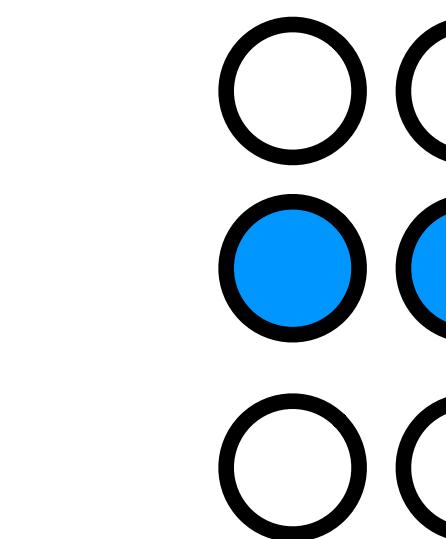
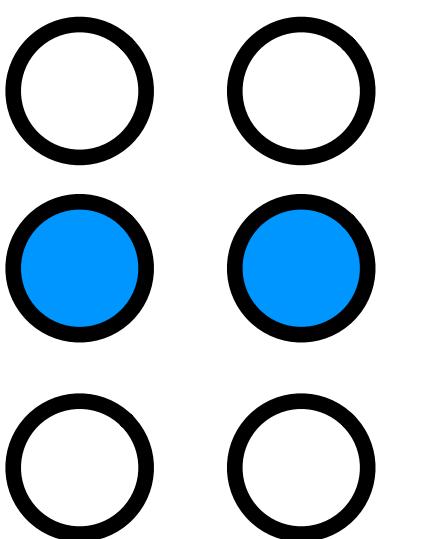
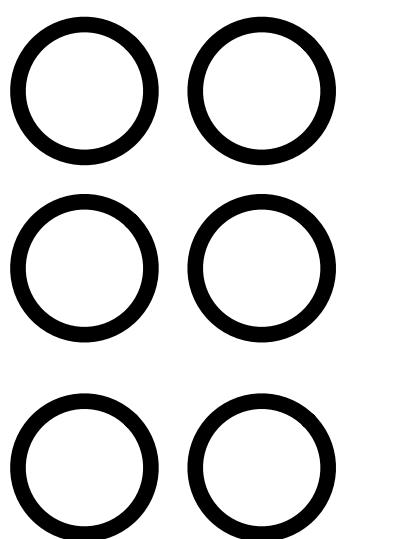
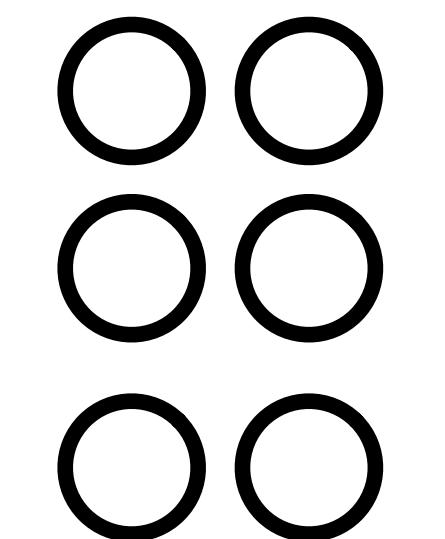
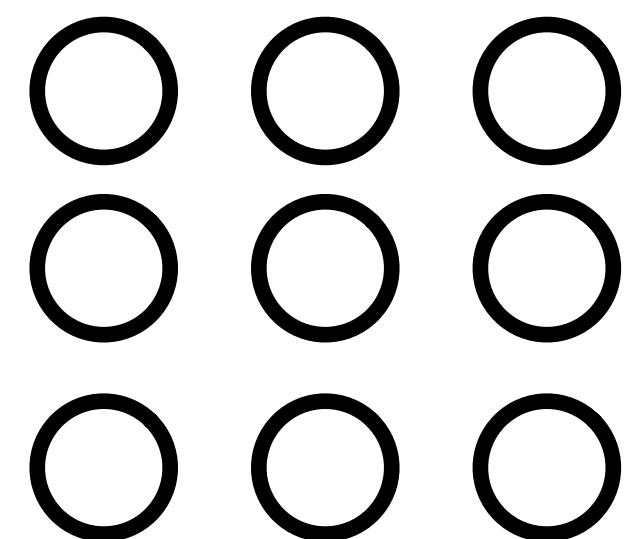


by technology

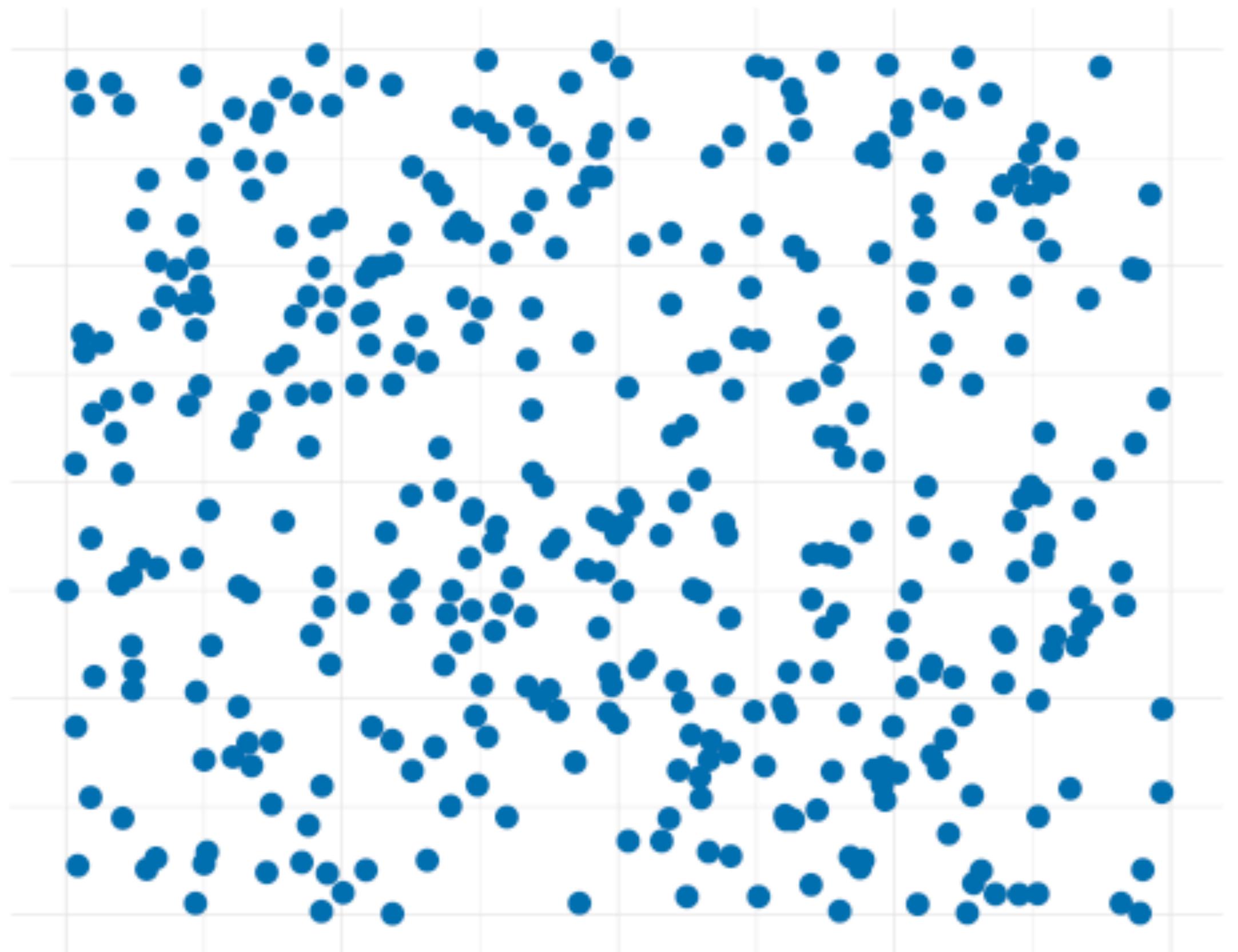


by use-case

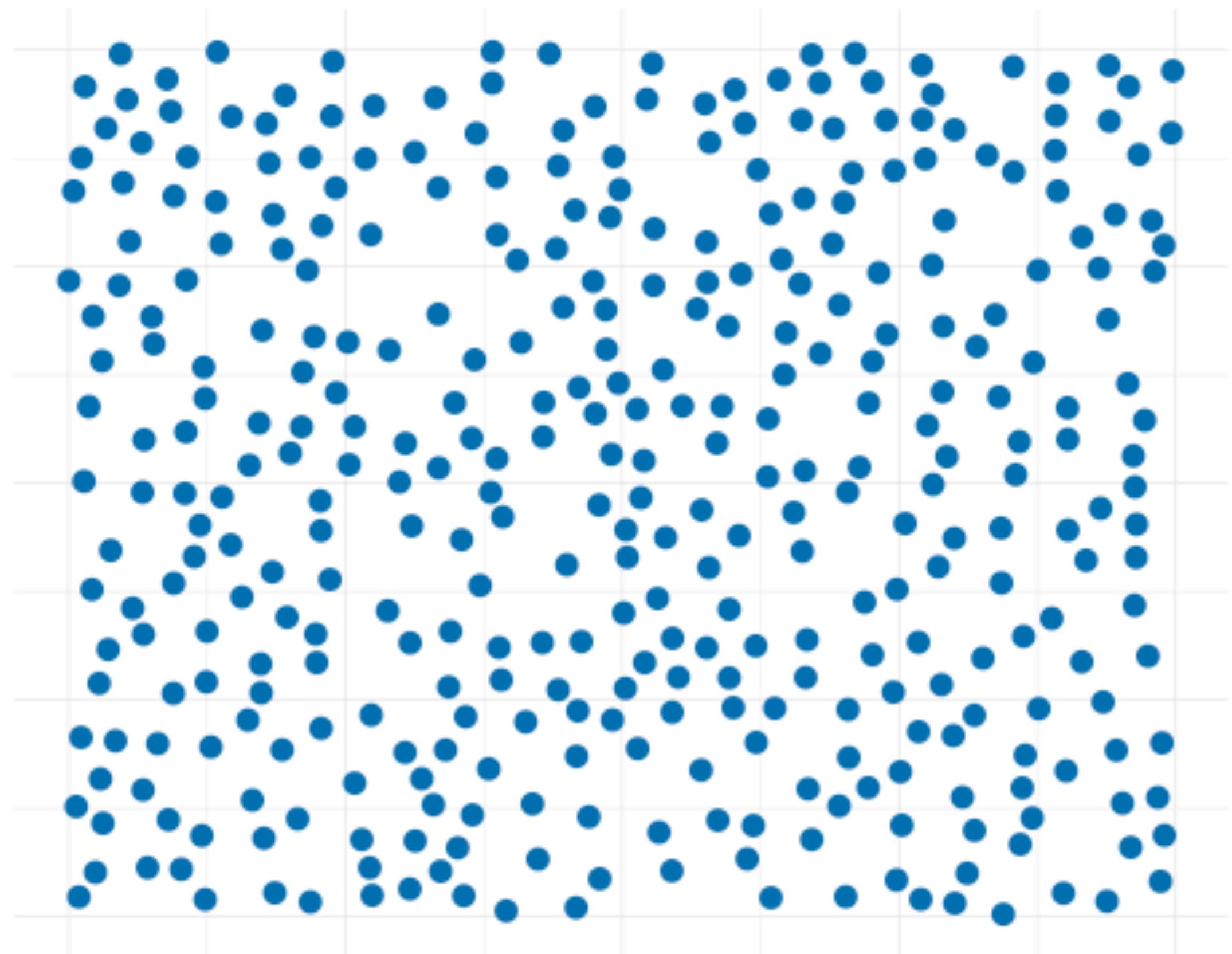




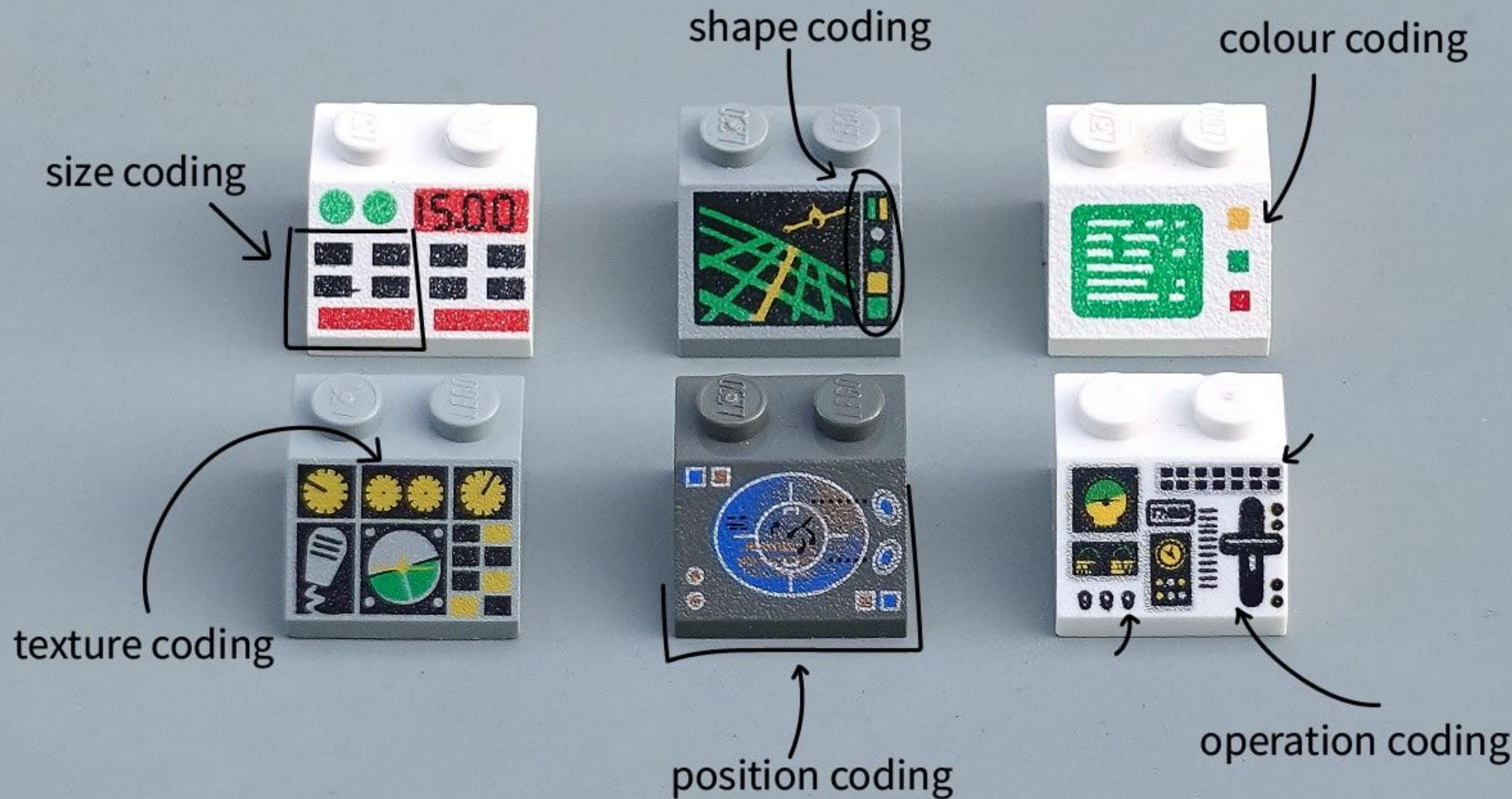
Poisson



Matérn

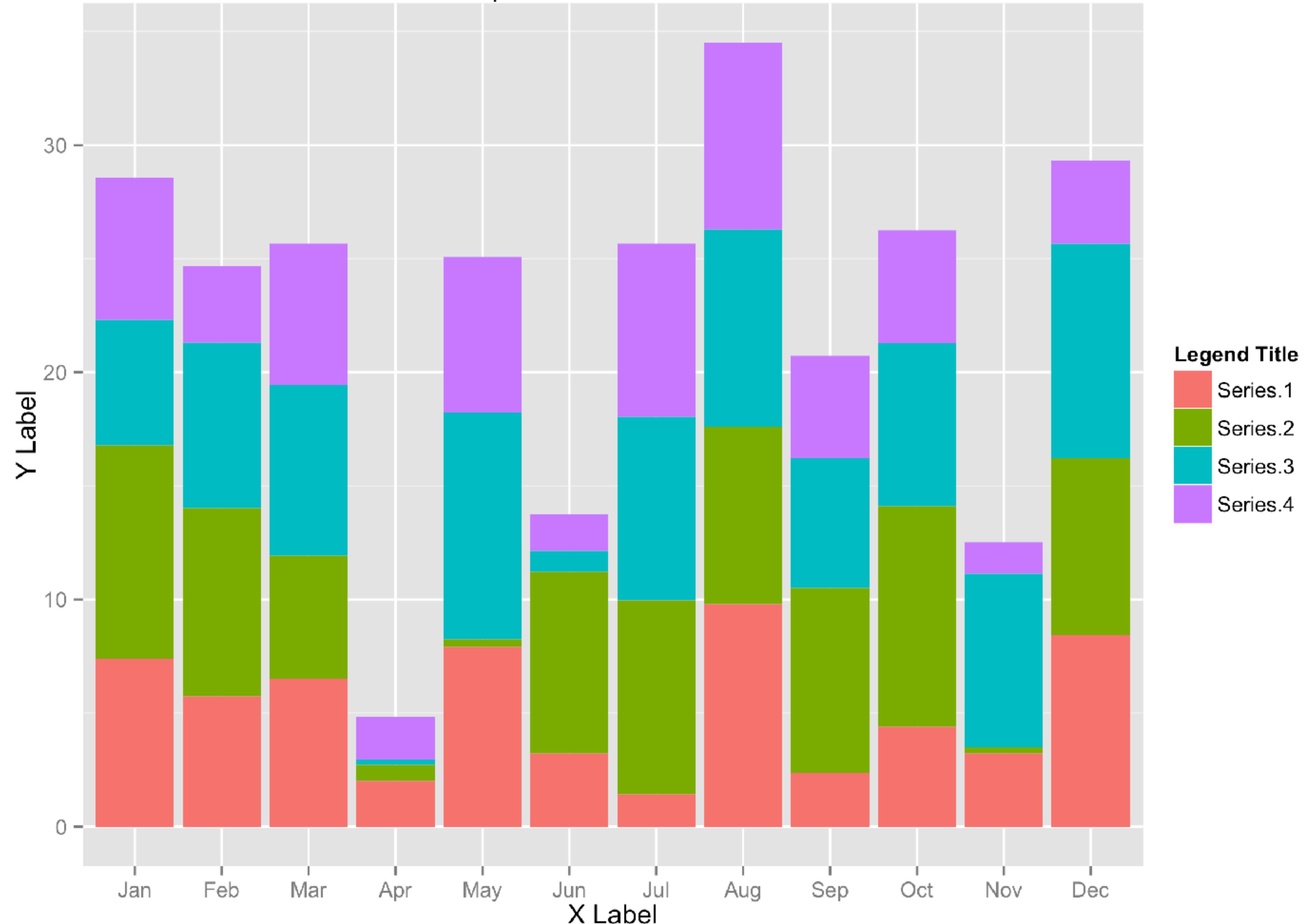


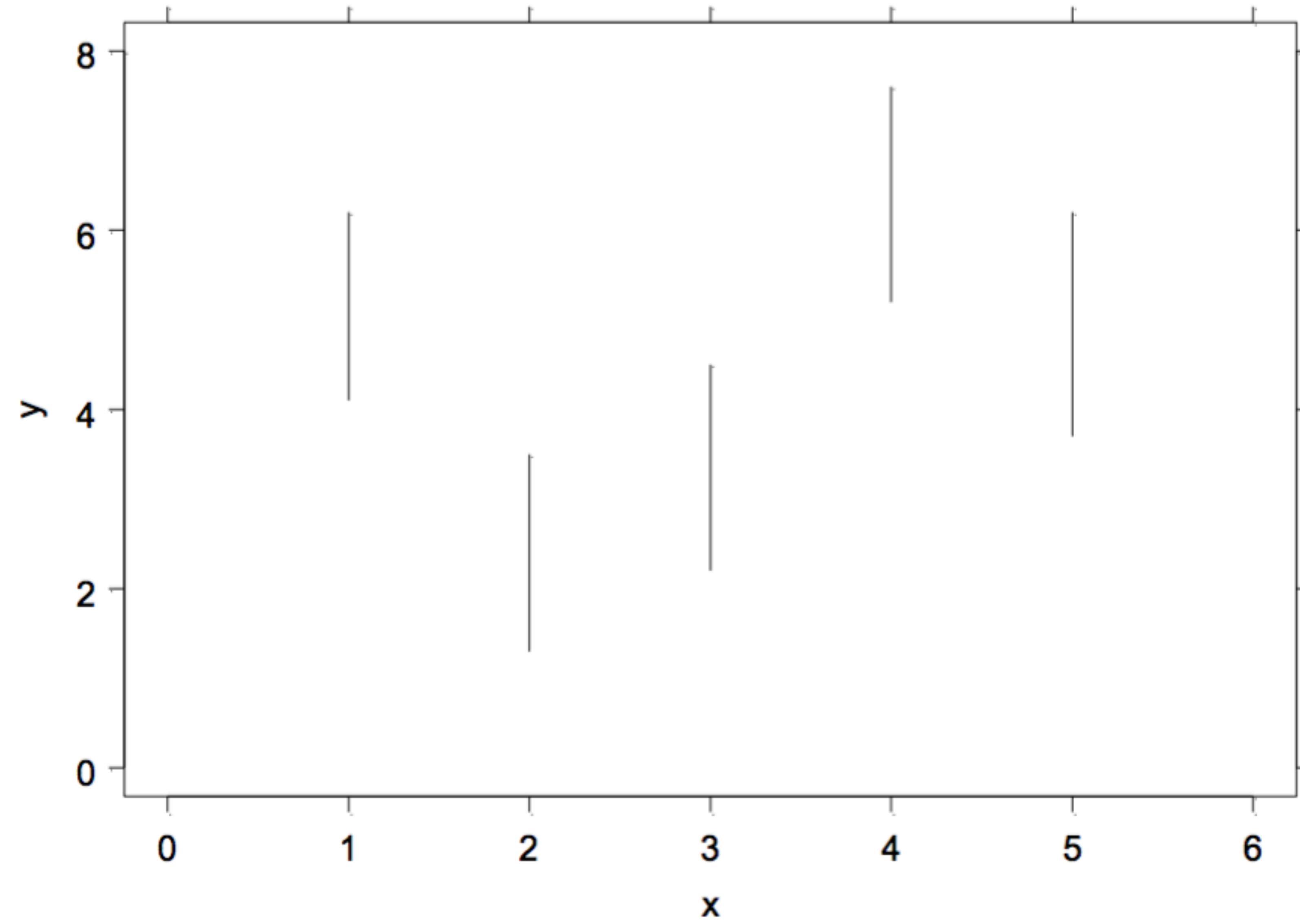
MAPPINGS FOR DATA

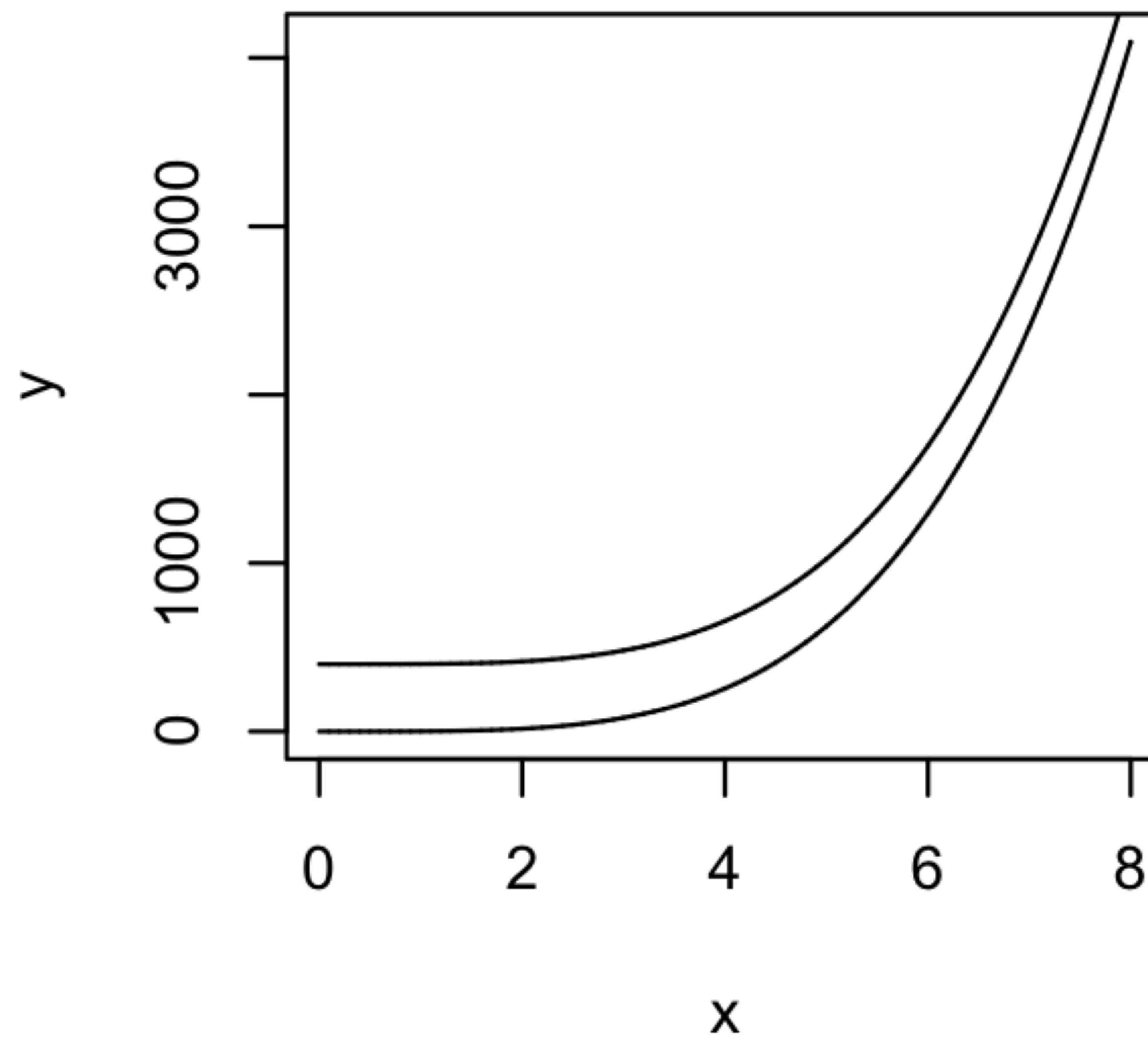


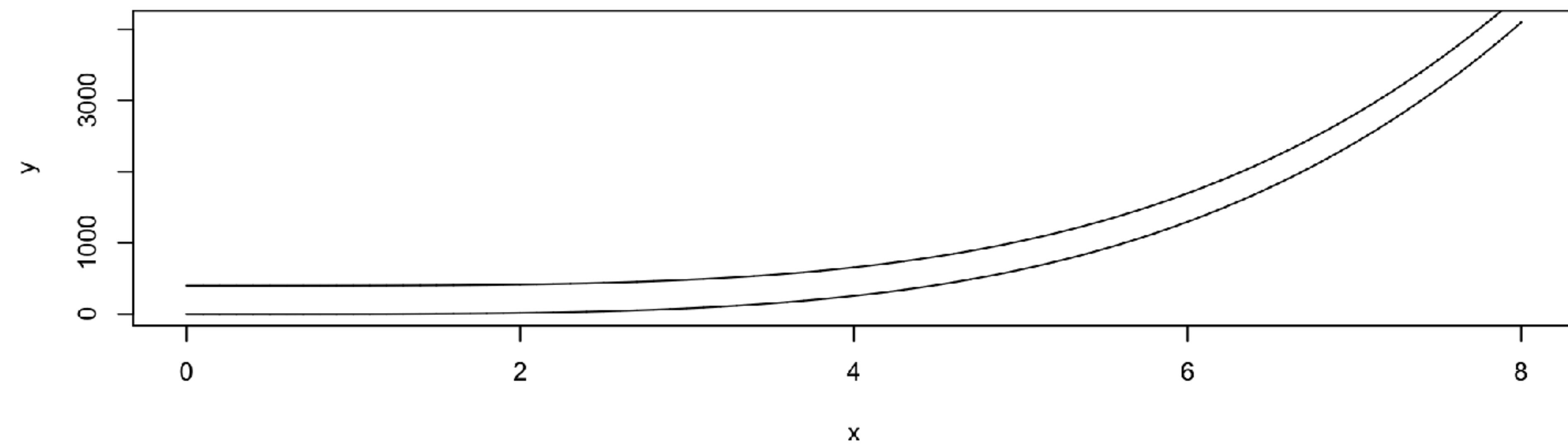
Visual Tasks in Decoding Graphs

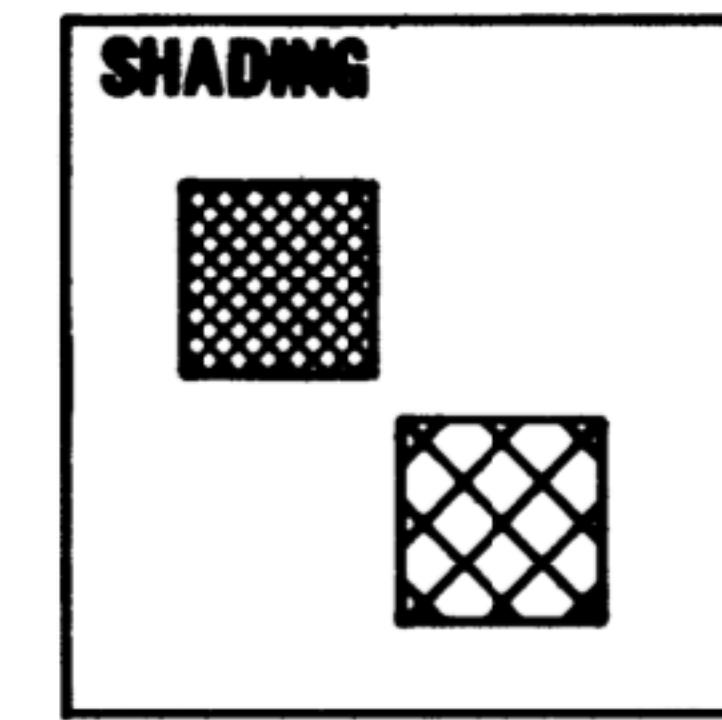
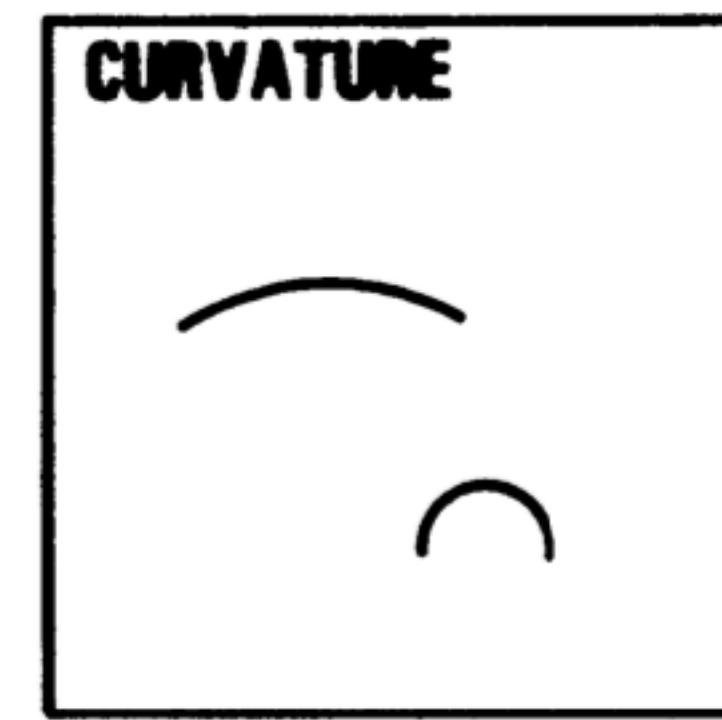
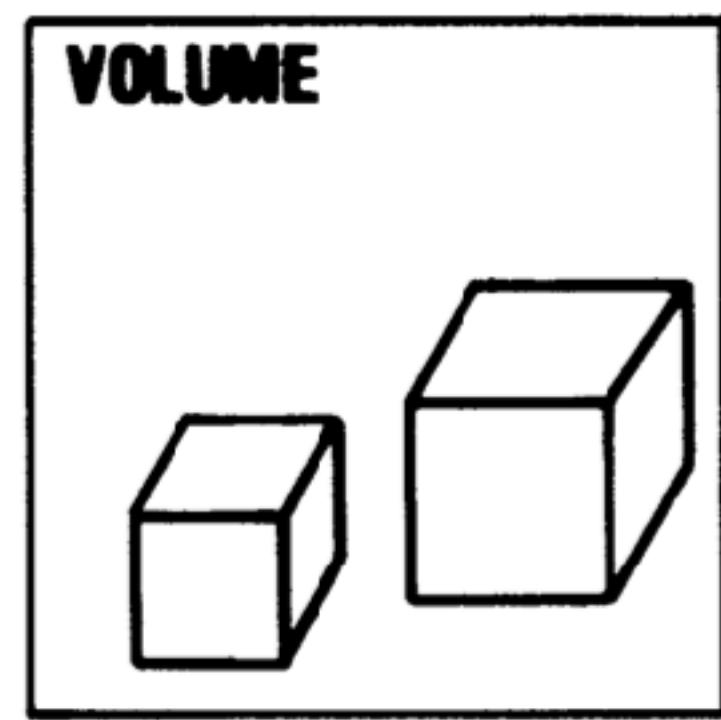
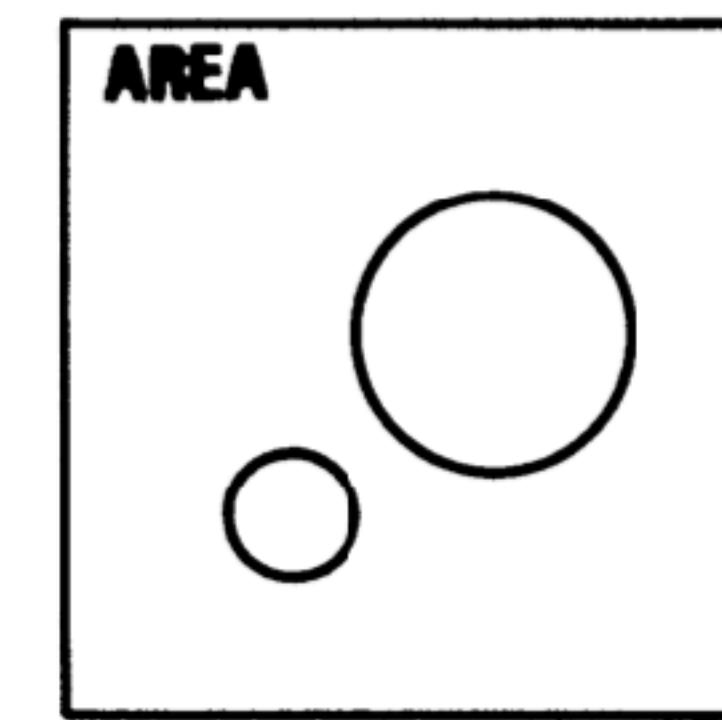
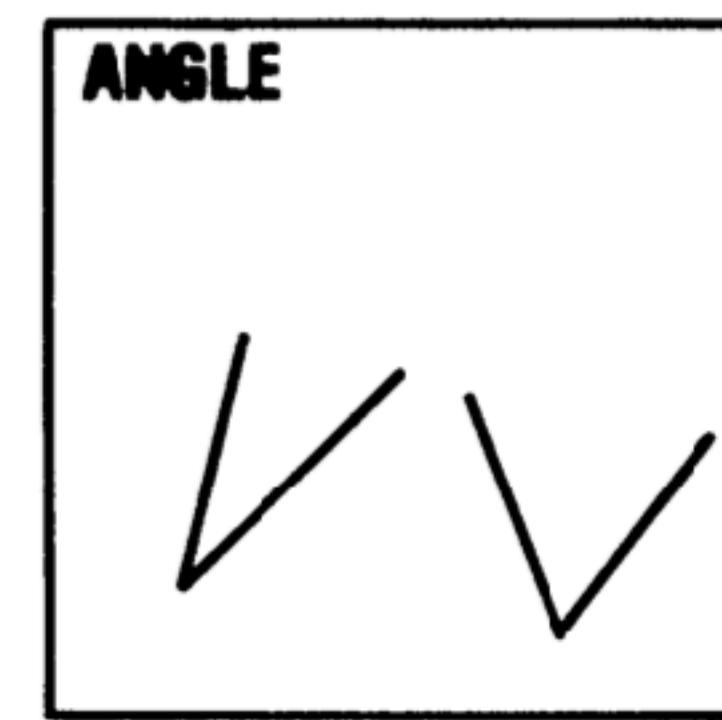
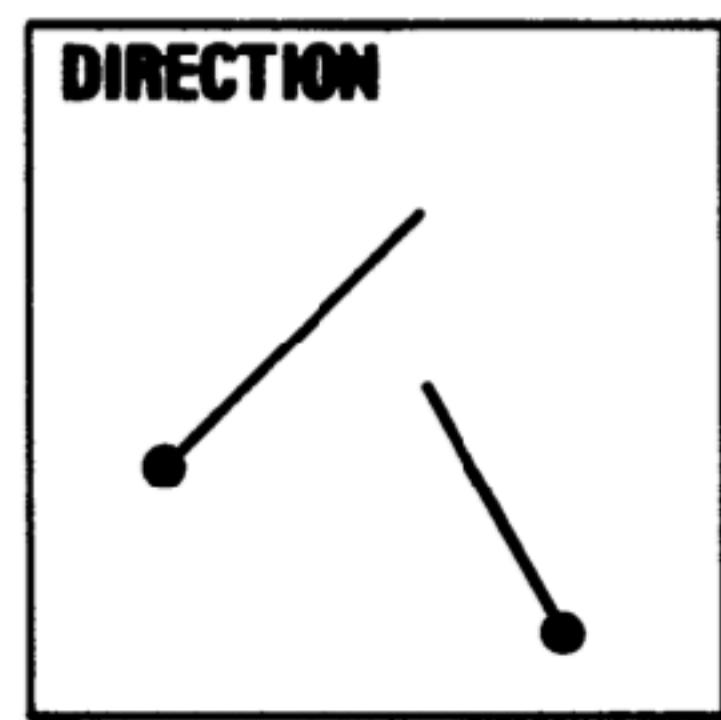
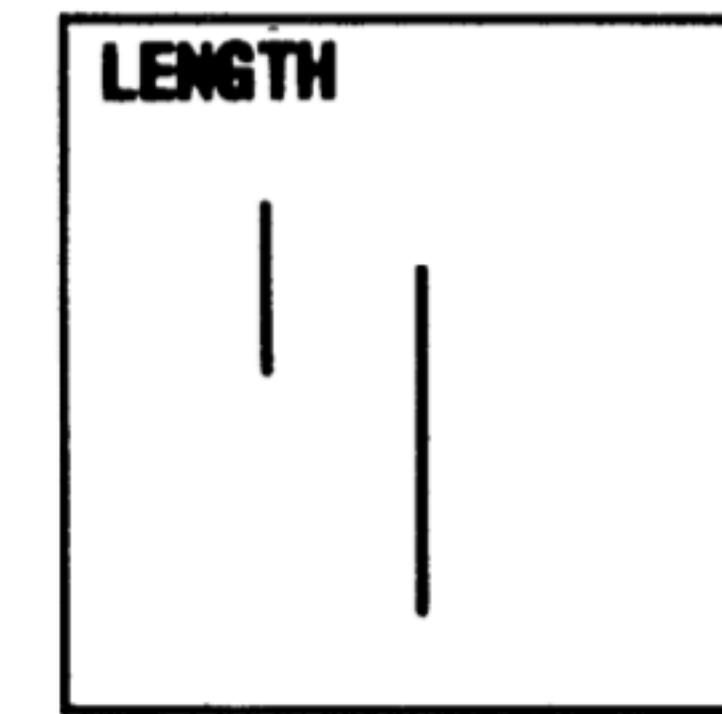
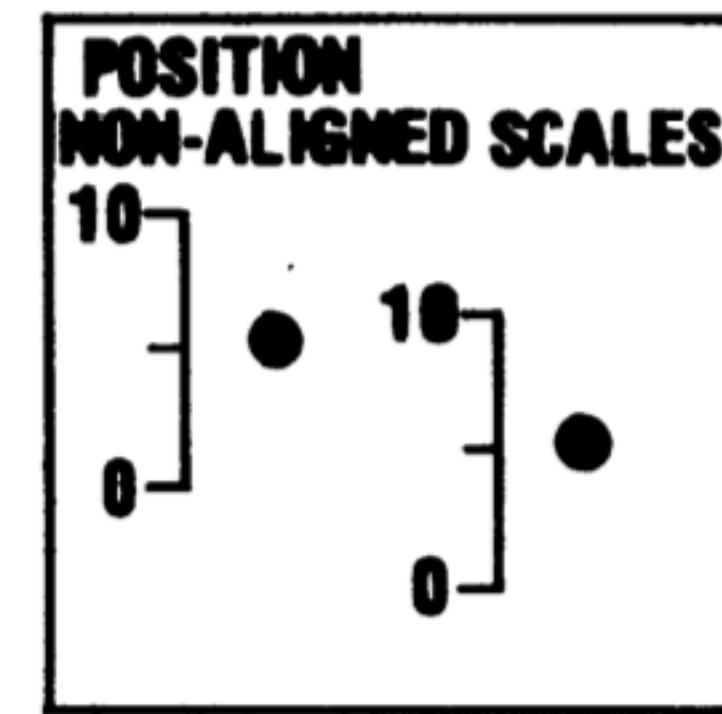
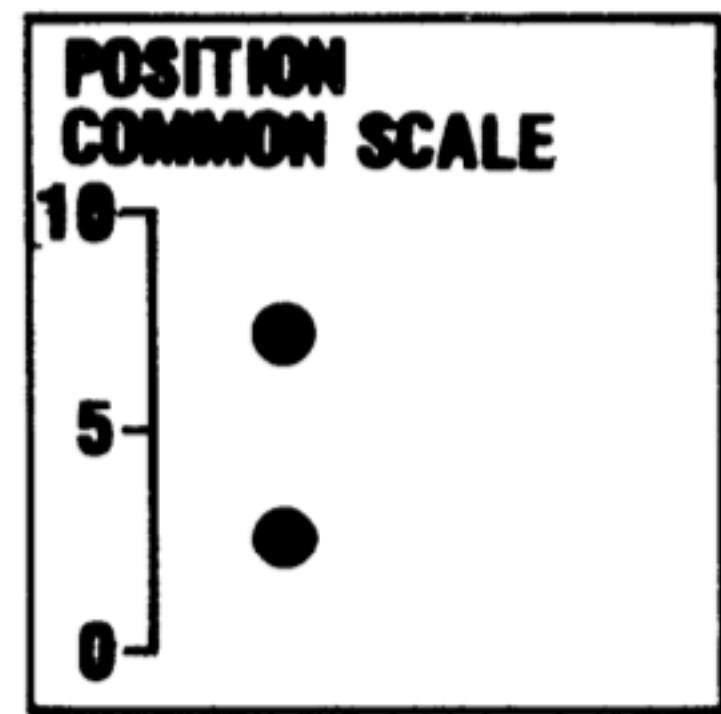
An Example Stacked Column Chart











COLOR SATURATION

Figure 1. Elementary perceptual tasks.

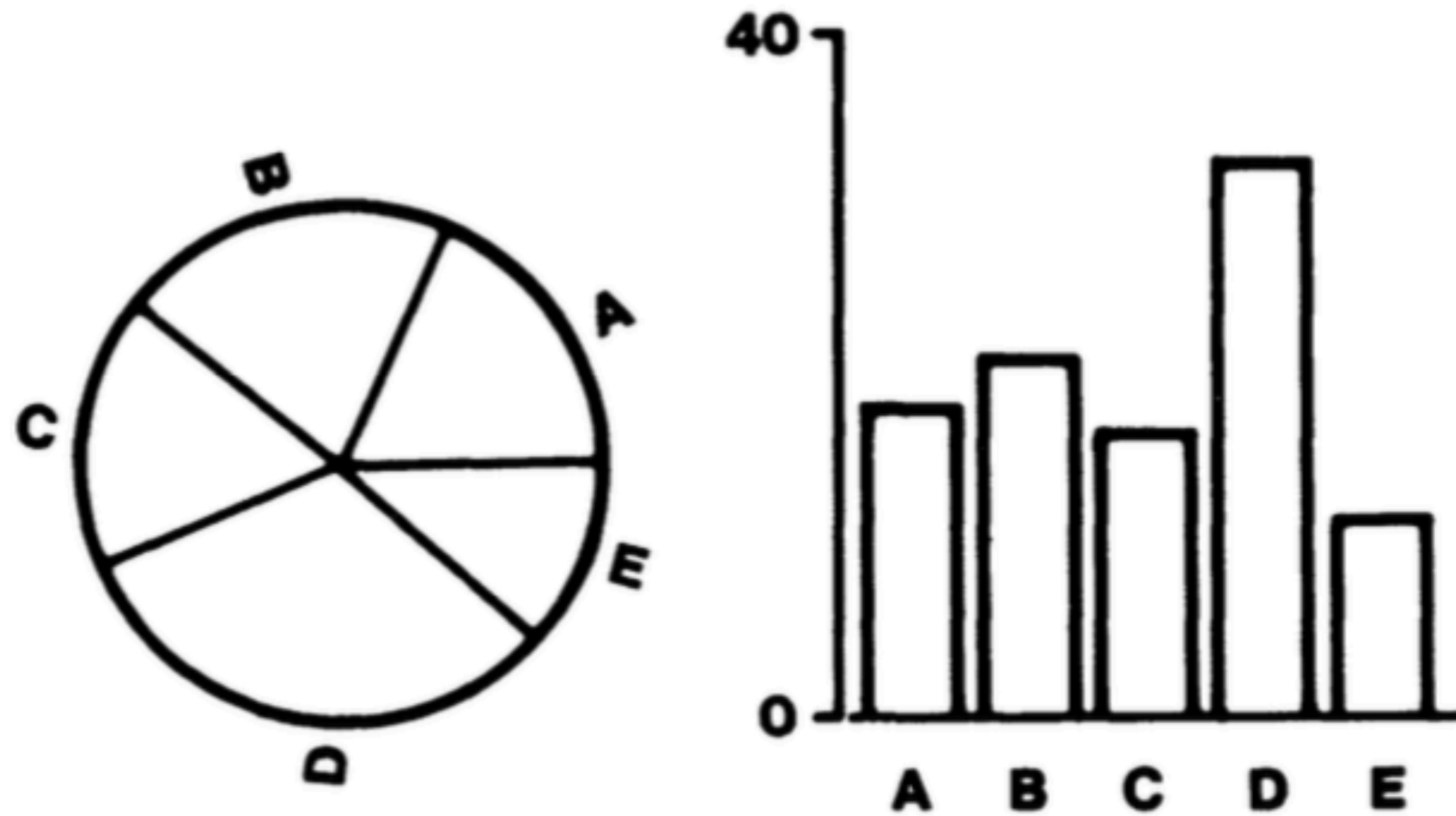


Figure 3. Graphs from position-angle experiment.

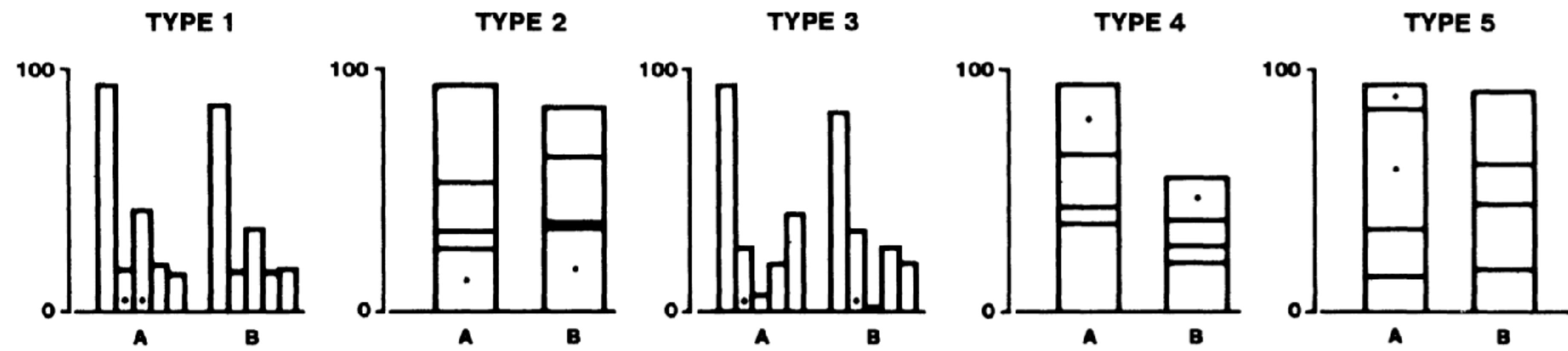
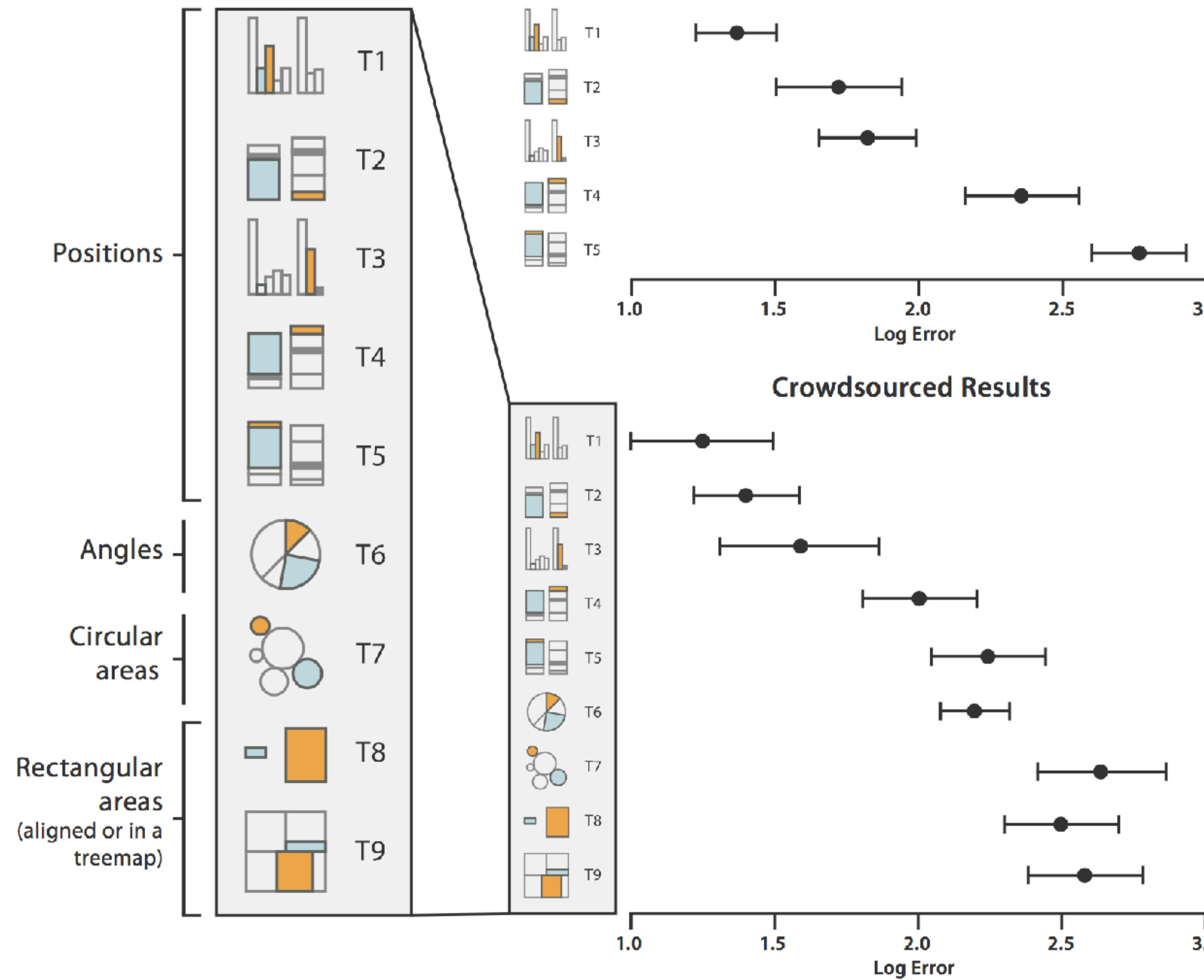


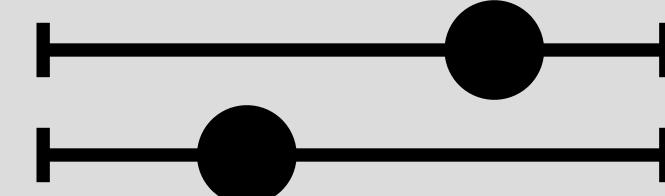
Figure 4. Graphs from position-length experiment.

Cleveland & McGill's Results



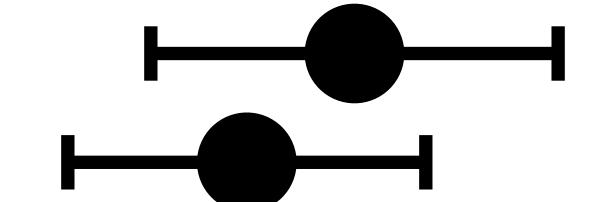
A Rough Hierarchy of Mappings

Position on a common scale



Better

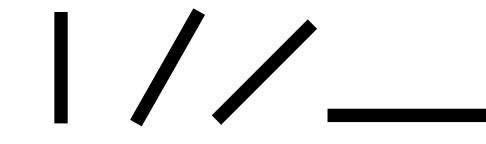
Position on unaligned scale



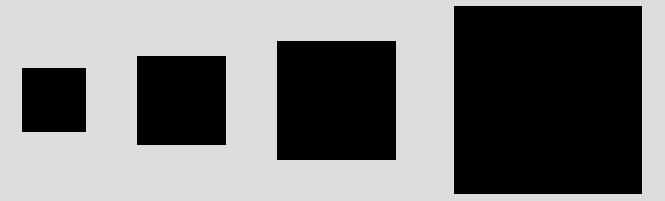
Length (1D as size)



Tilt or Angle



Area (2D as size)



Effectiveness

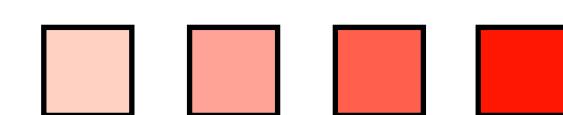
Depth (3D as Position)



Color luminance [brightness]

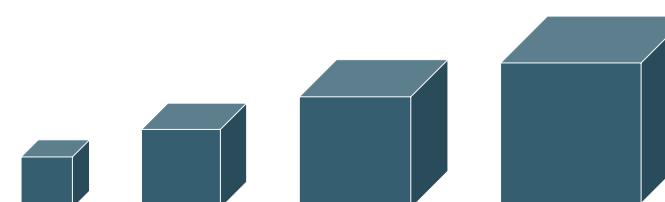


Color saturation [intensity]

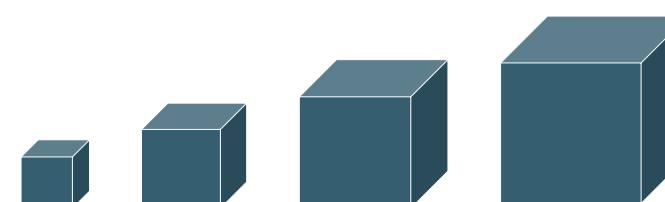


Worse

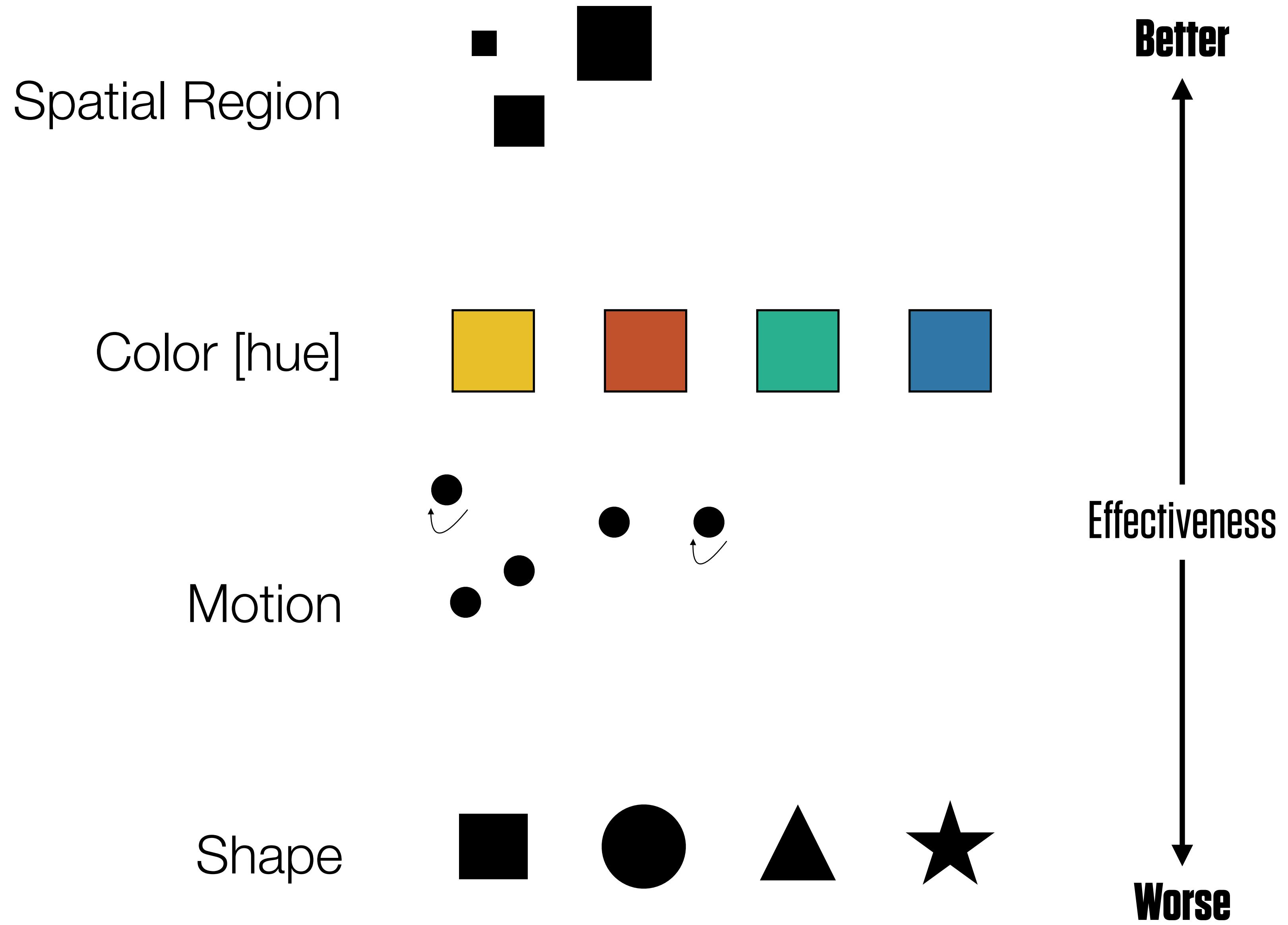
Curvature



Volume (3D as size)



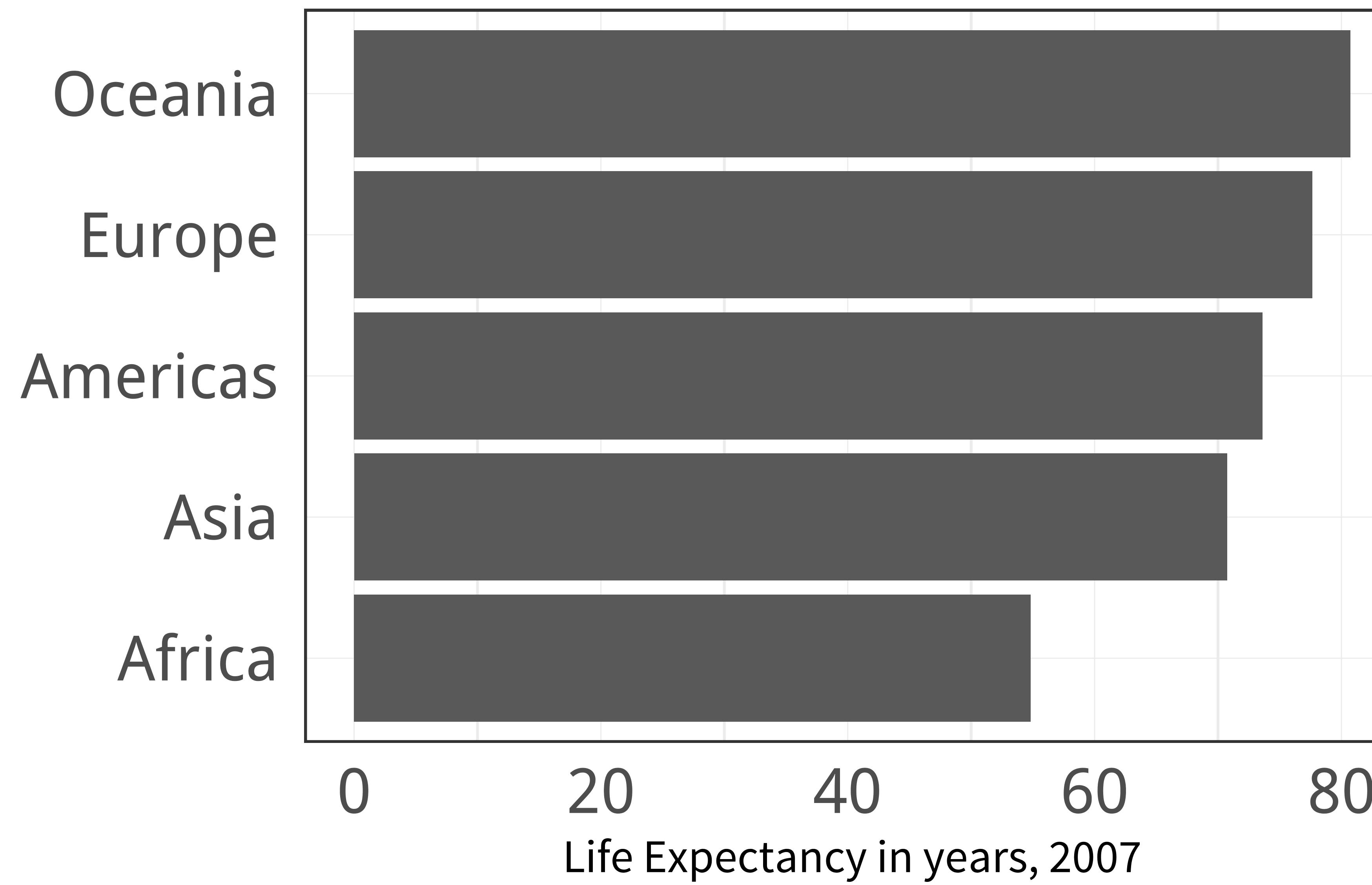
See e.g. Munzer (2014)

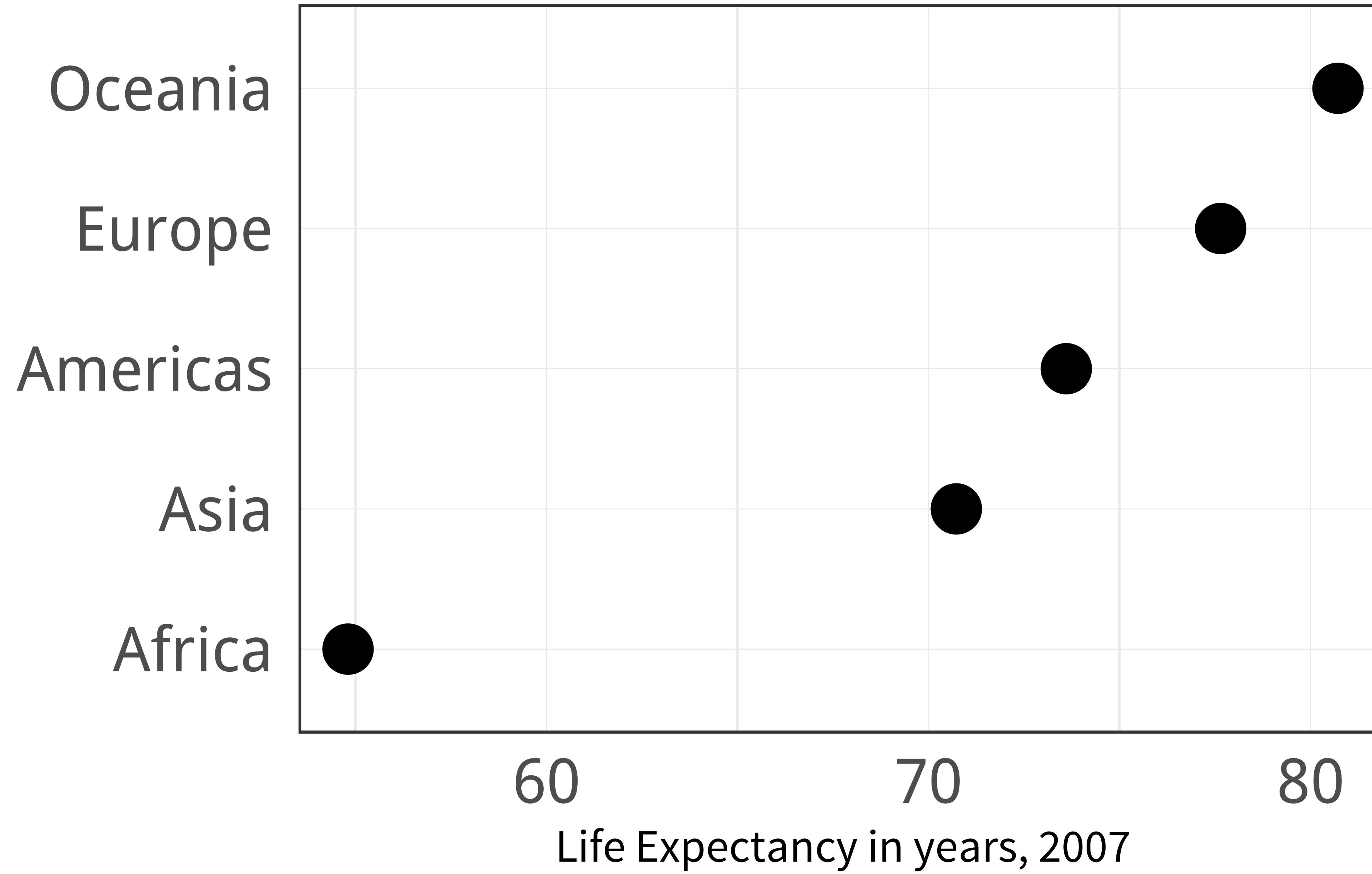


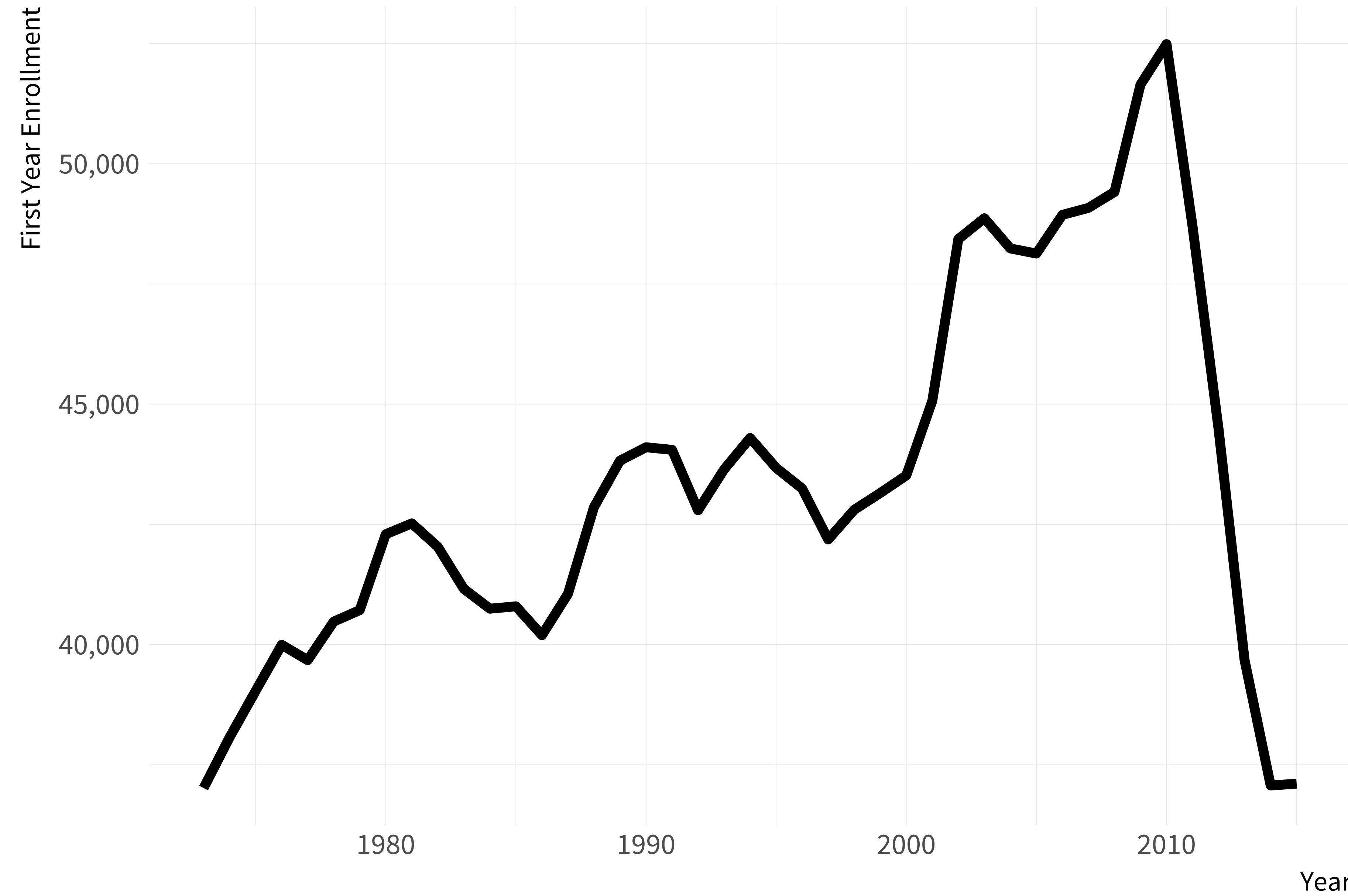
See e.g. Munzer (2014)

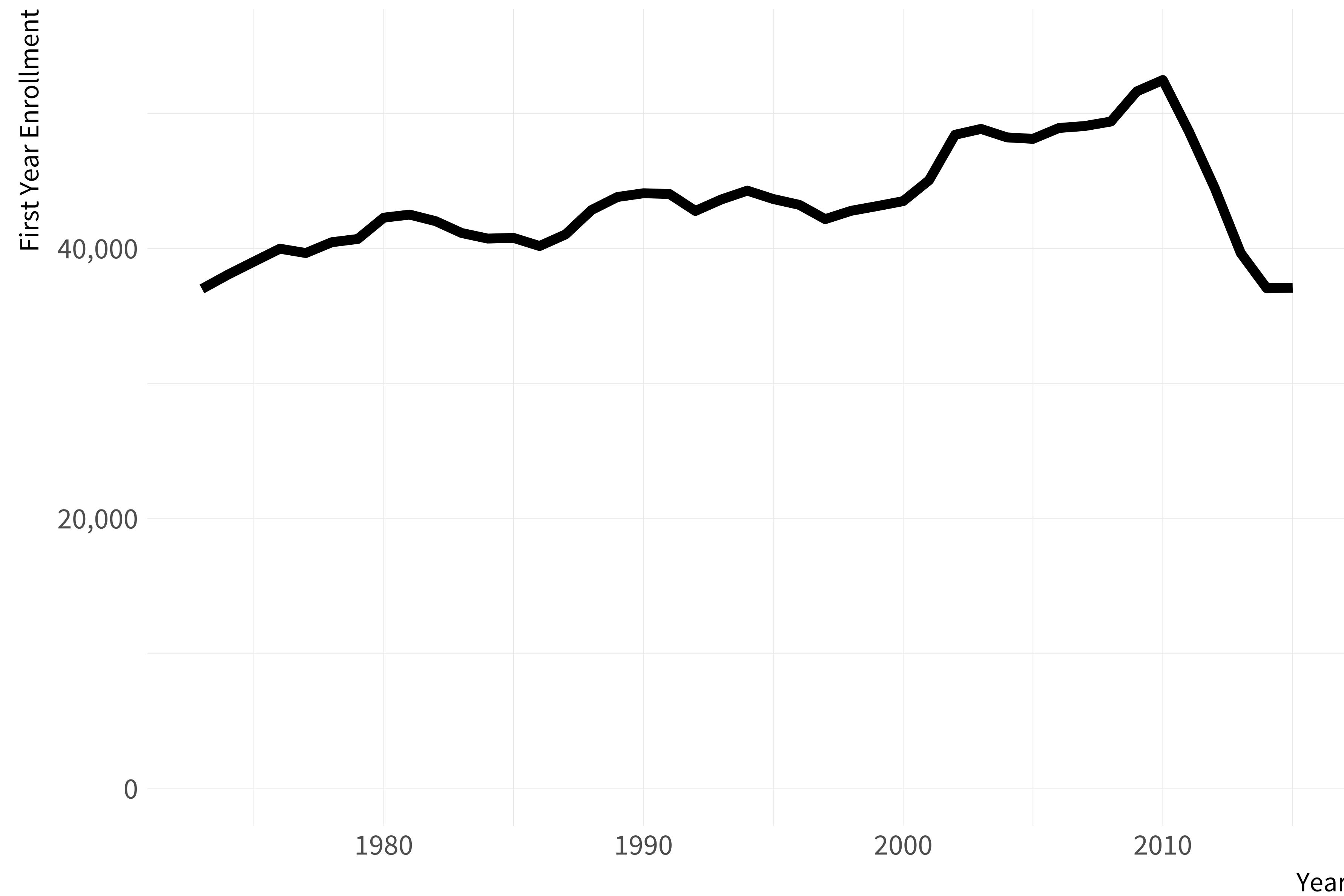
**HONESTY AND
GOOD JUDGMENT**

**Example:
Baselines**



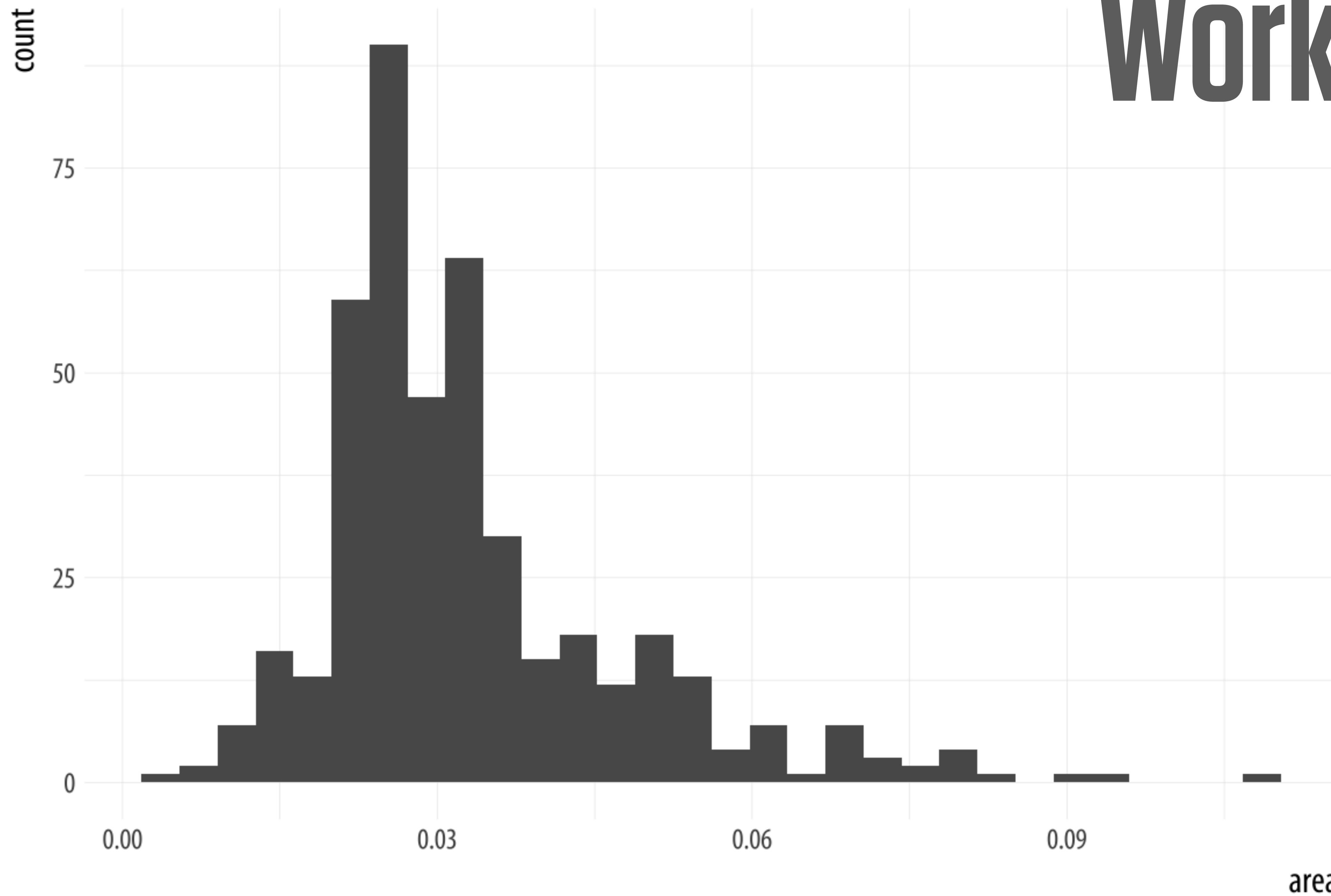






In Practice

Workhorses



lifeExp

80

60

40

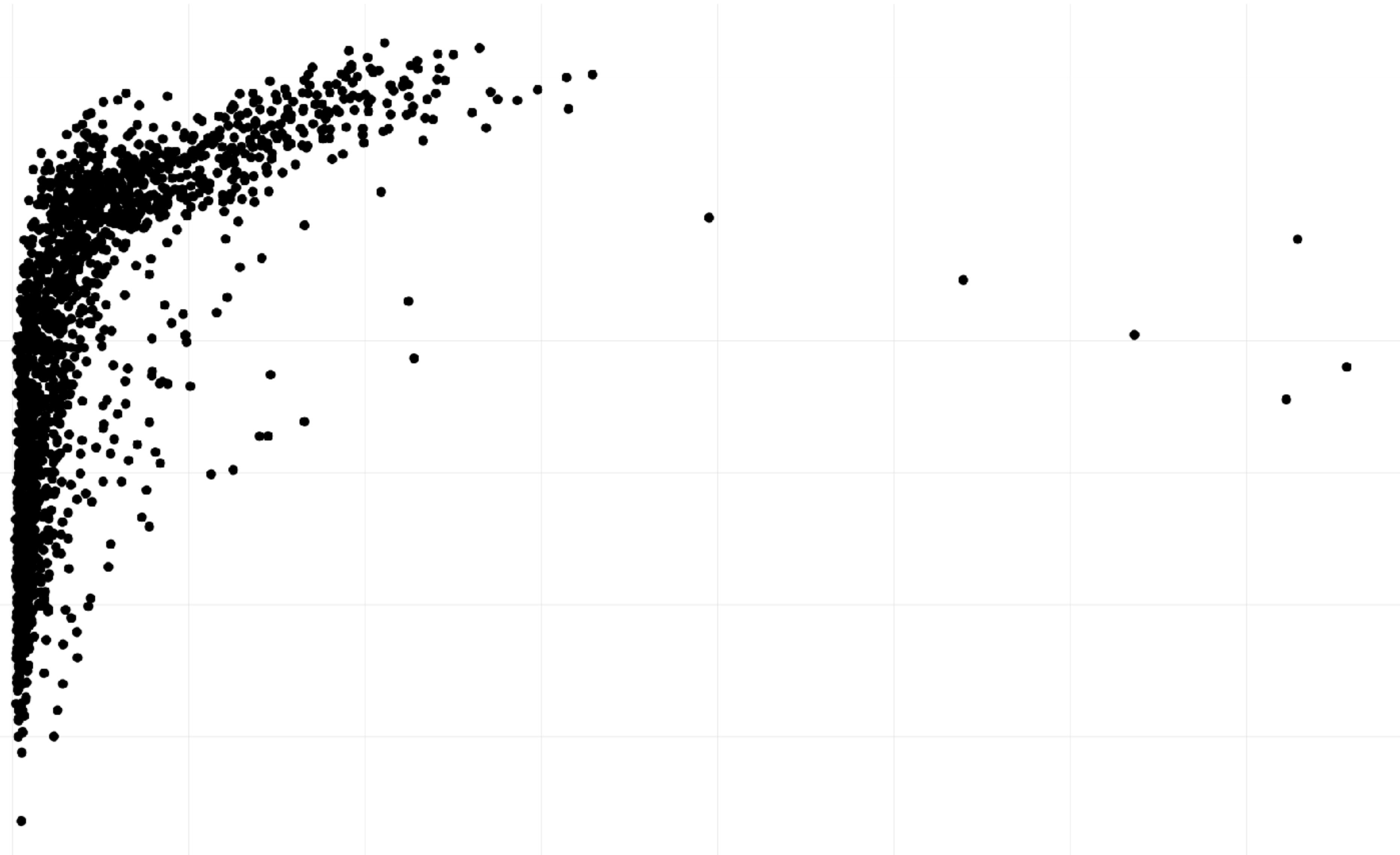
0

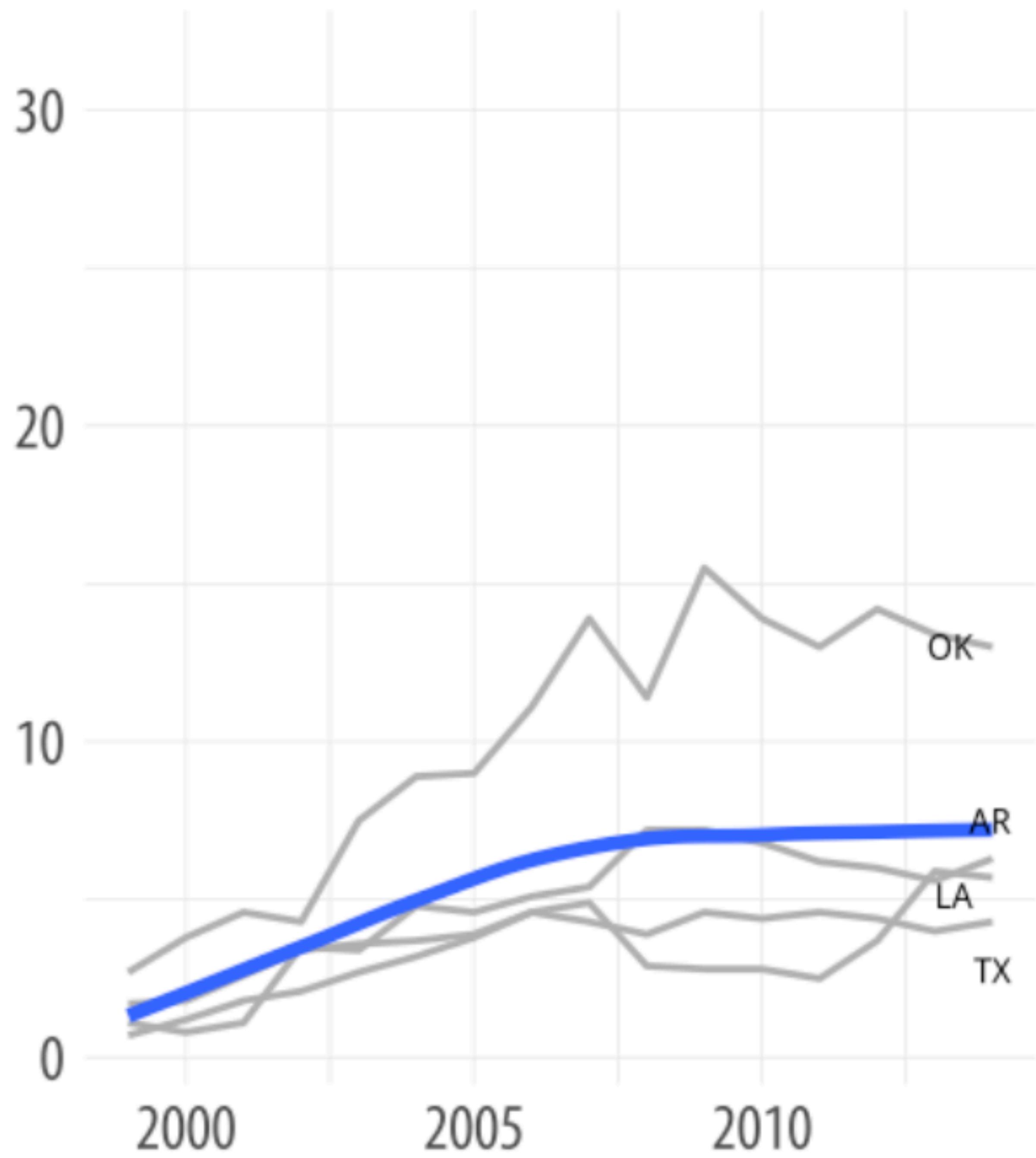
30000

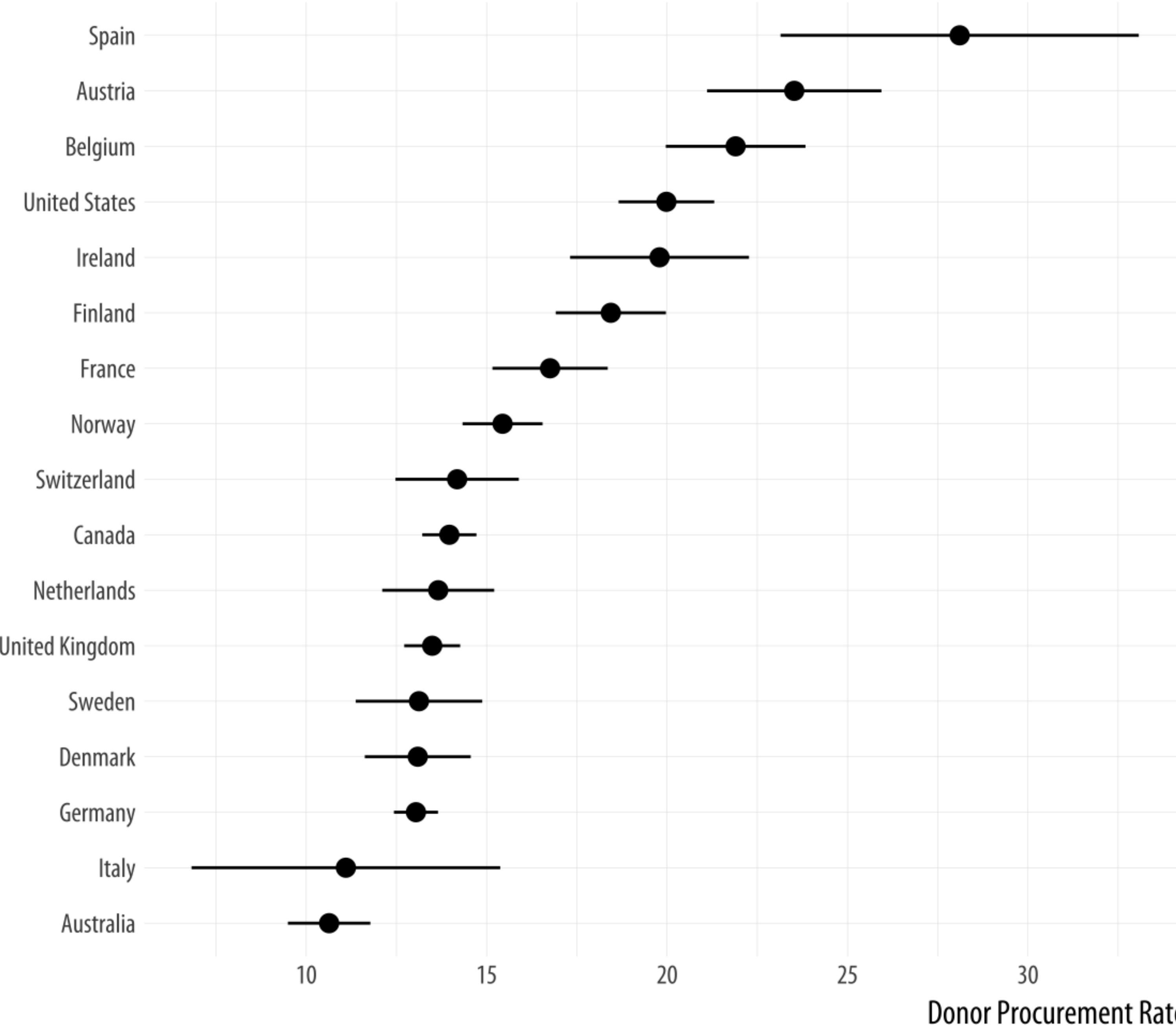
60000

90000

gdpPerCap

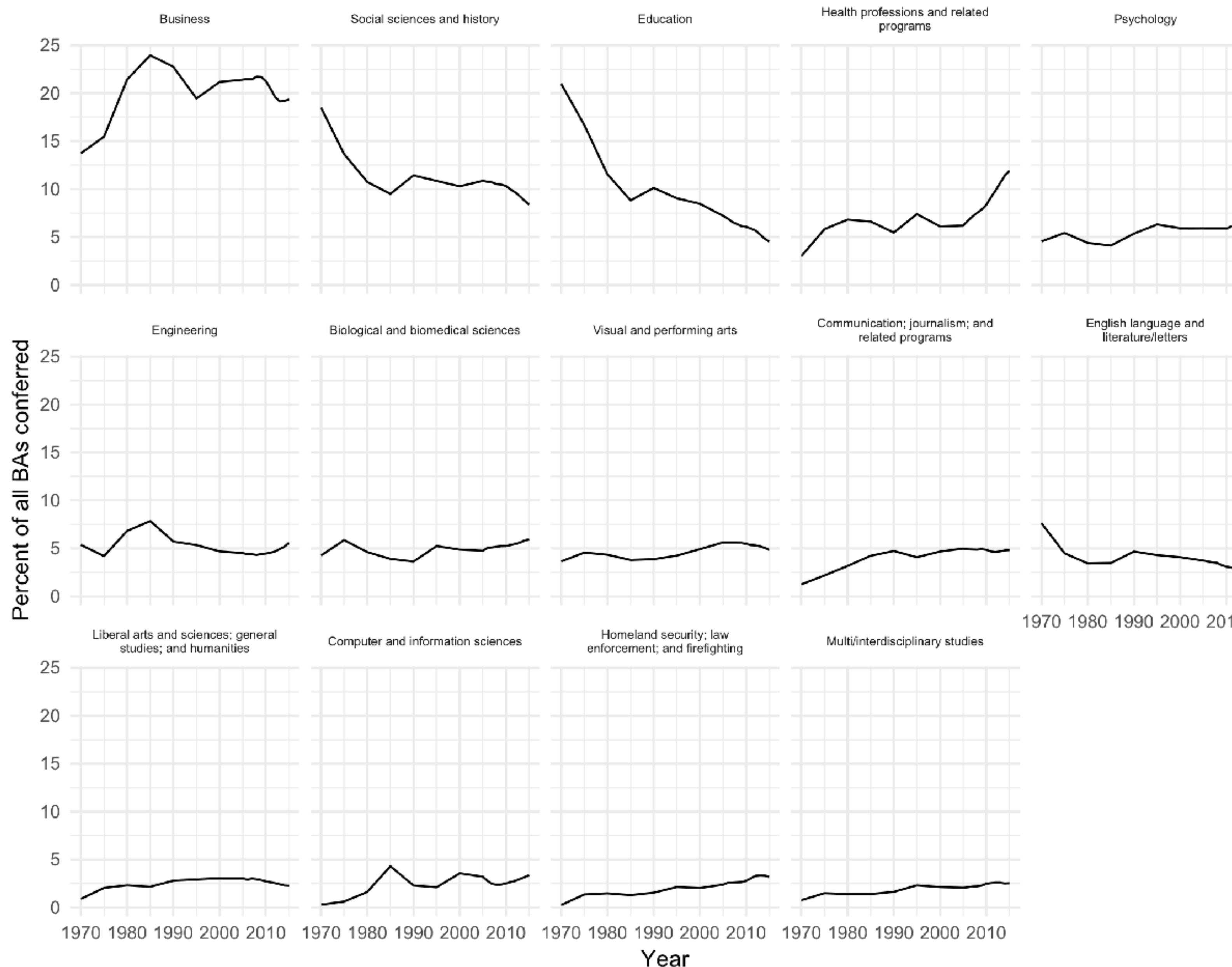






US Trends in Bachelor's Degrees Conferred, 1970-2015, for Areas averaging more than 2% of all degrees

Observations are every 5 years from 1970-1995, and annually thereafter



Data from NCES Digest 2017, Table 322.10.

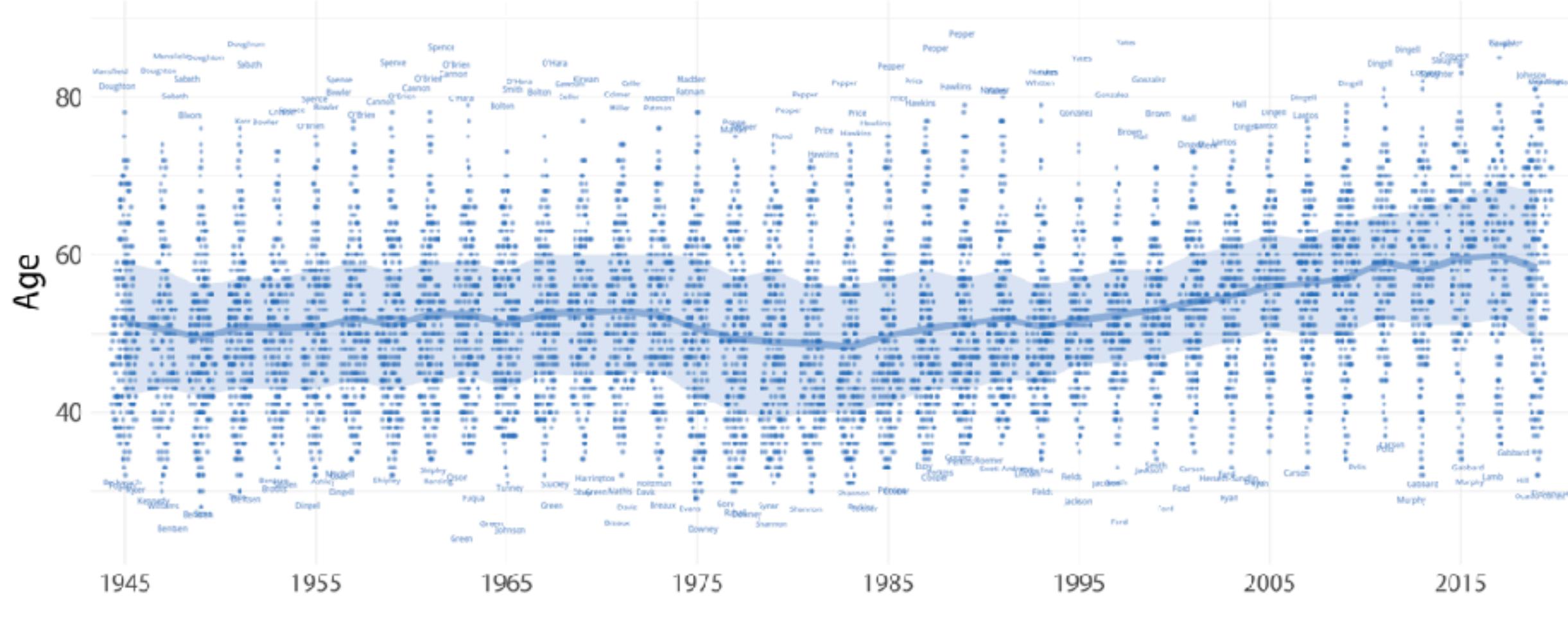
Show Ponies

Age Distribution of Congressional Representatives, 1945-2019

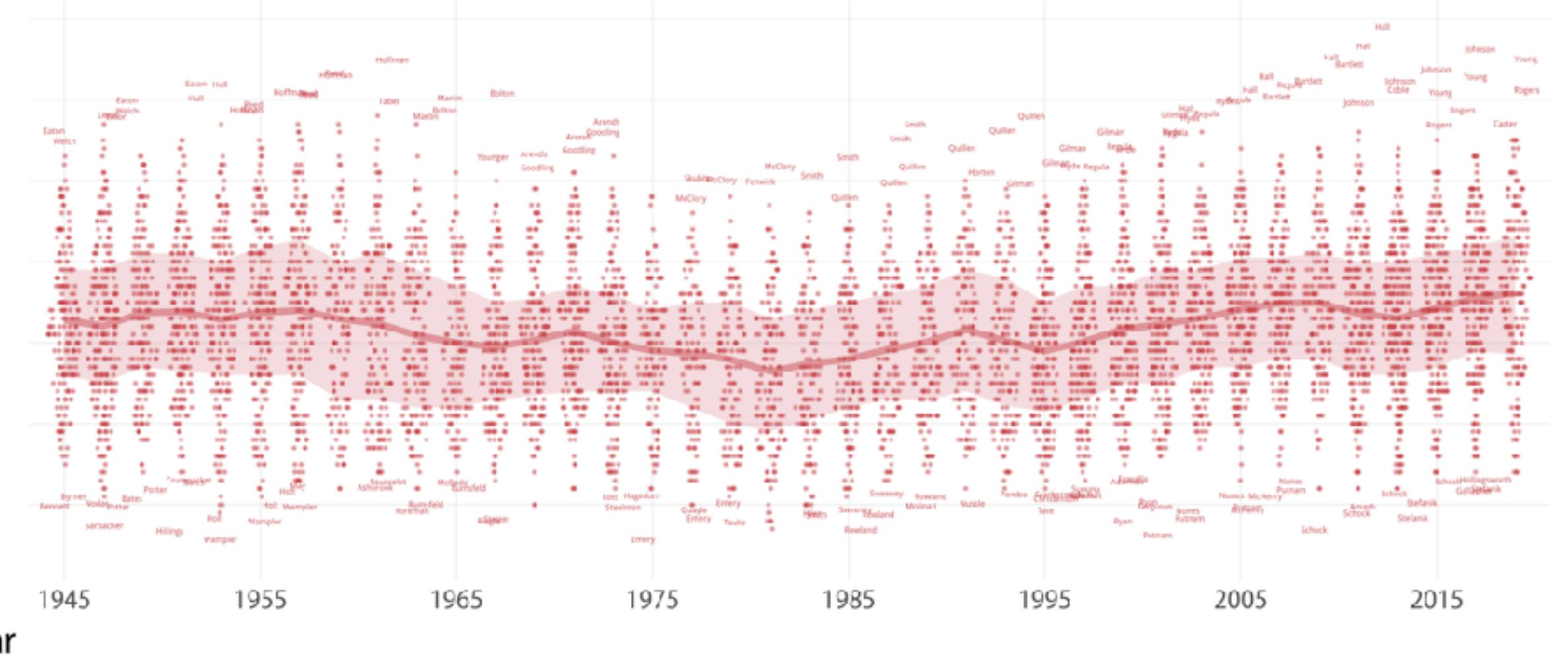
Trend line is mean age; bands are 25th and 75th percentiles of the range.

Youngest and oldest percentiles are named instead of being shown by points.

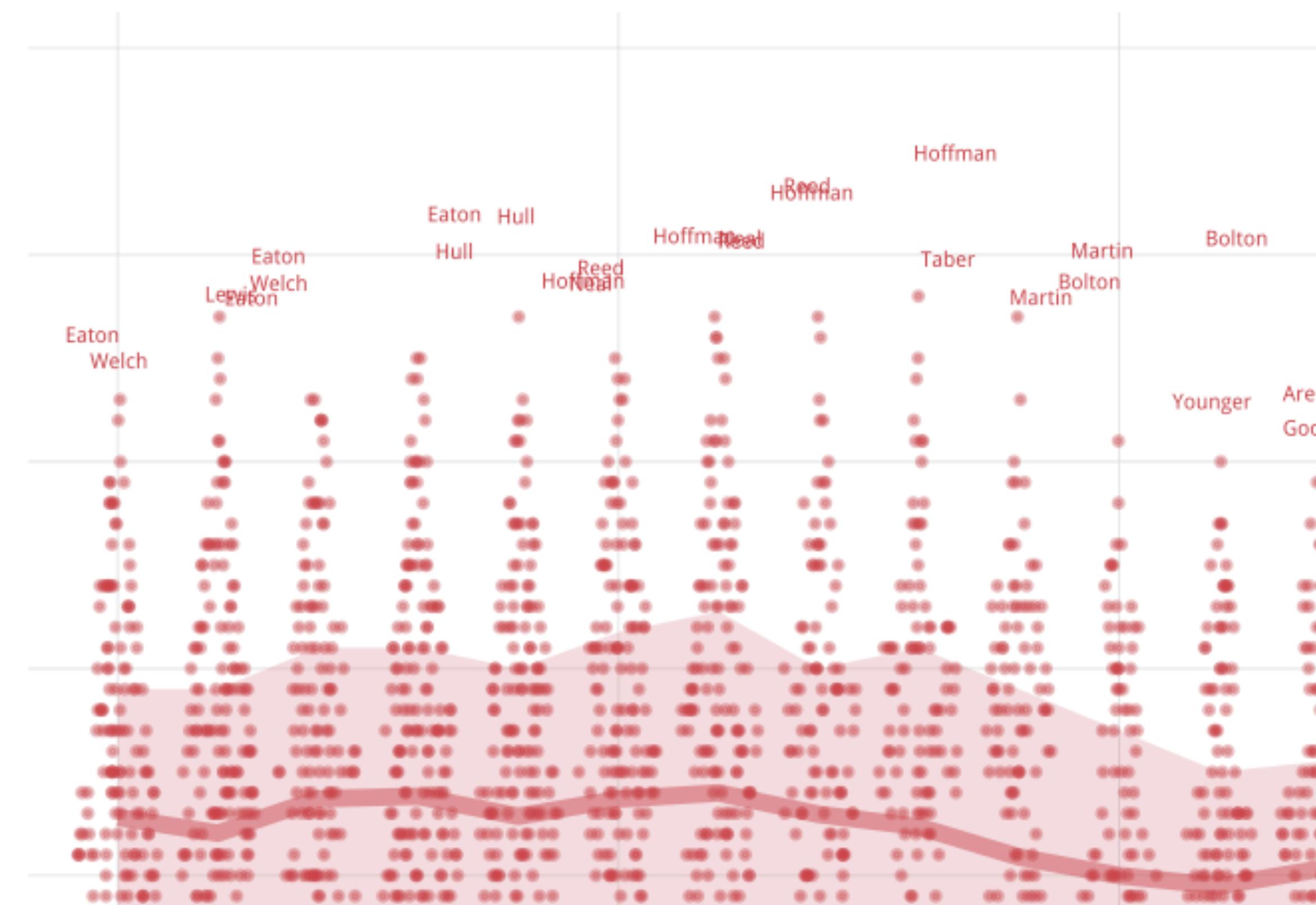
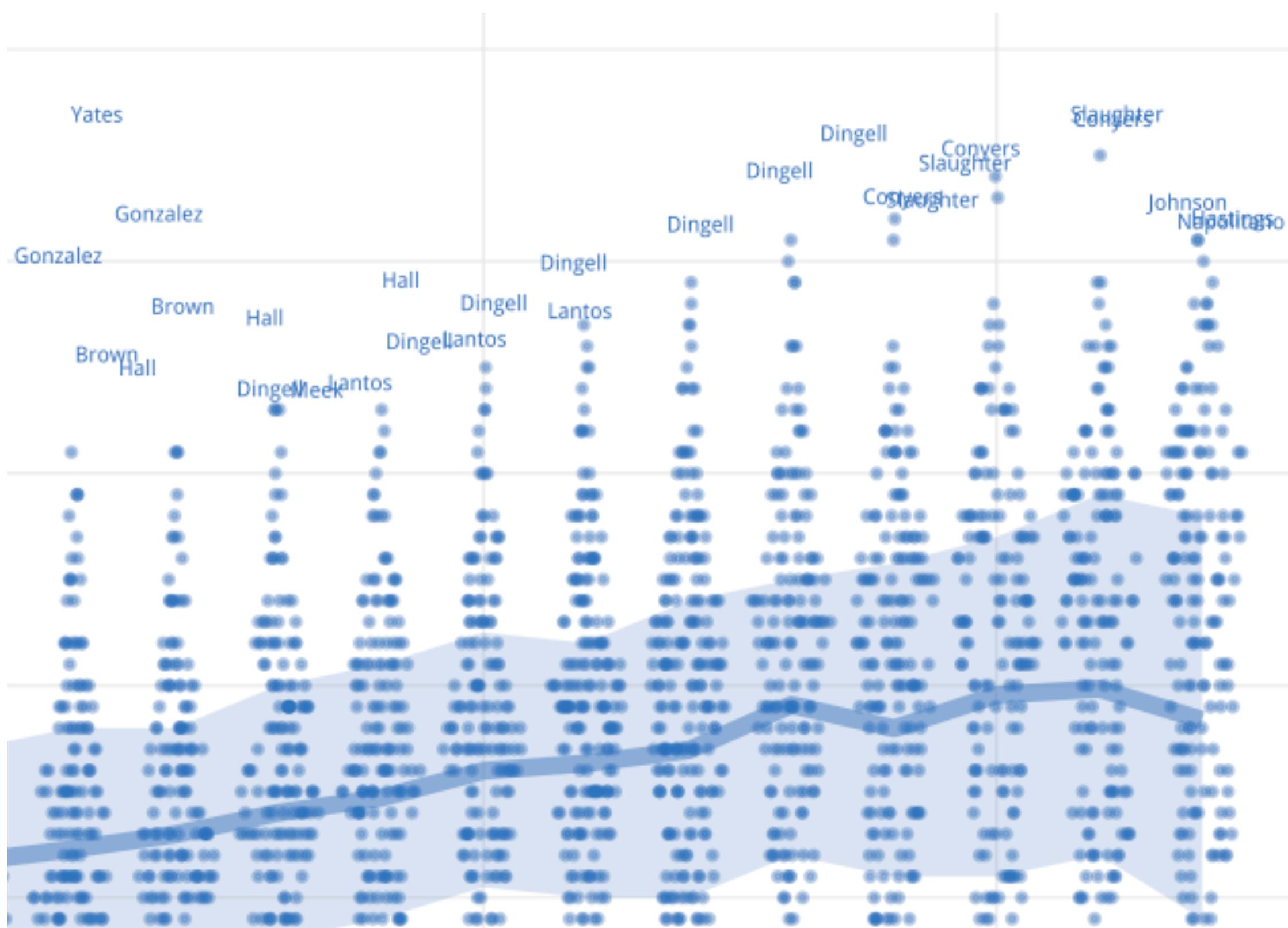
Democrats



Republicans



Show Ponies



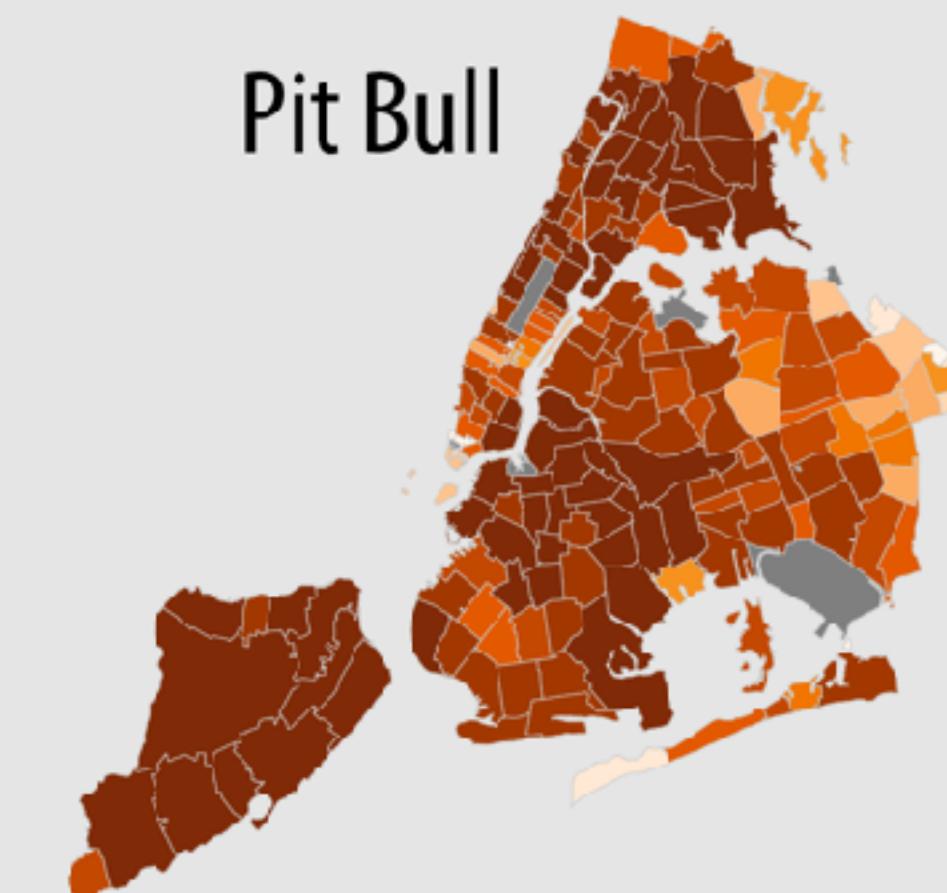
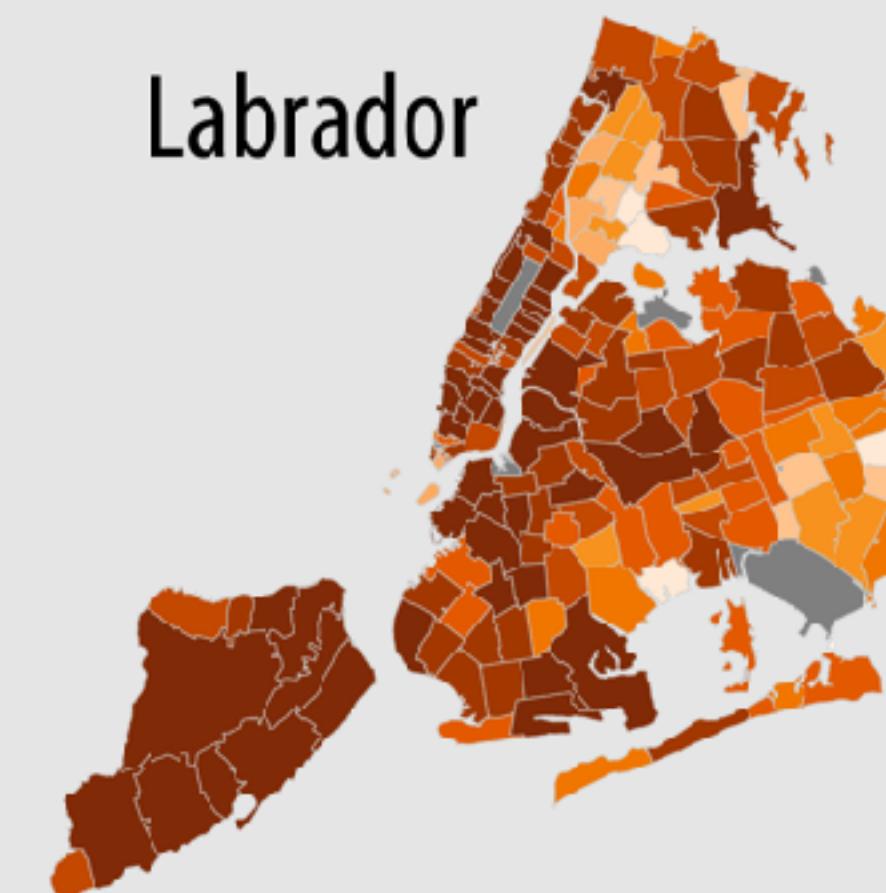
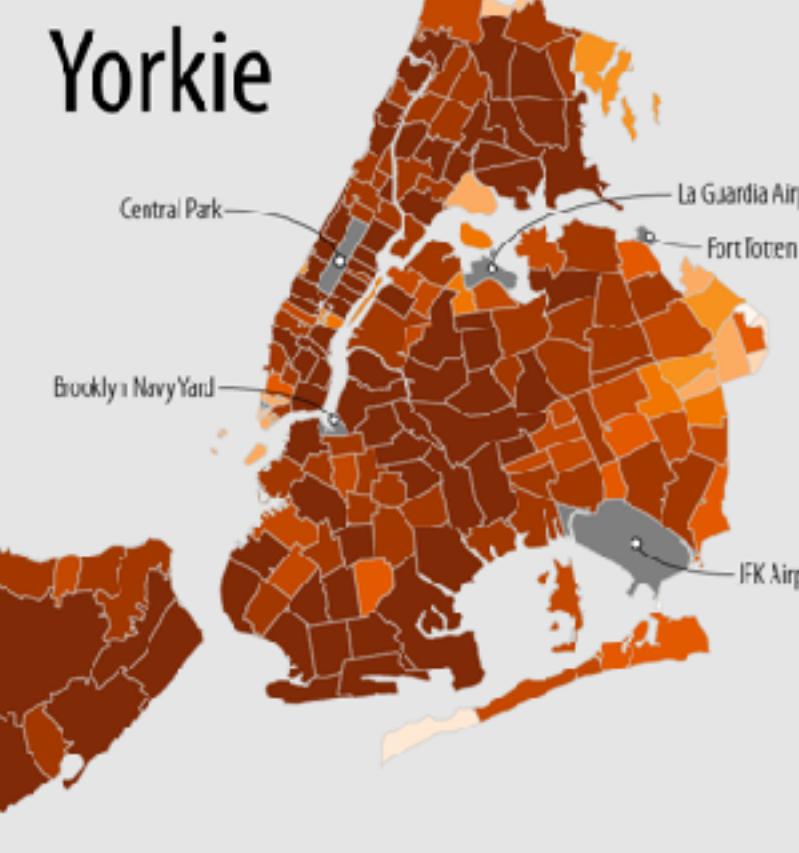
Relative prevalence of breed, from Most to Least



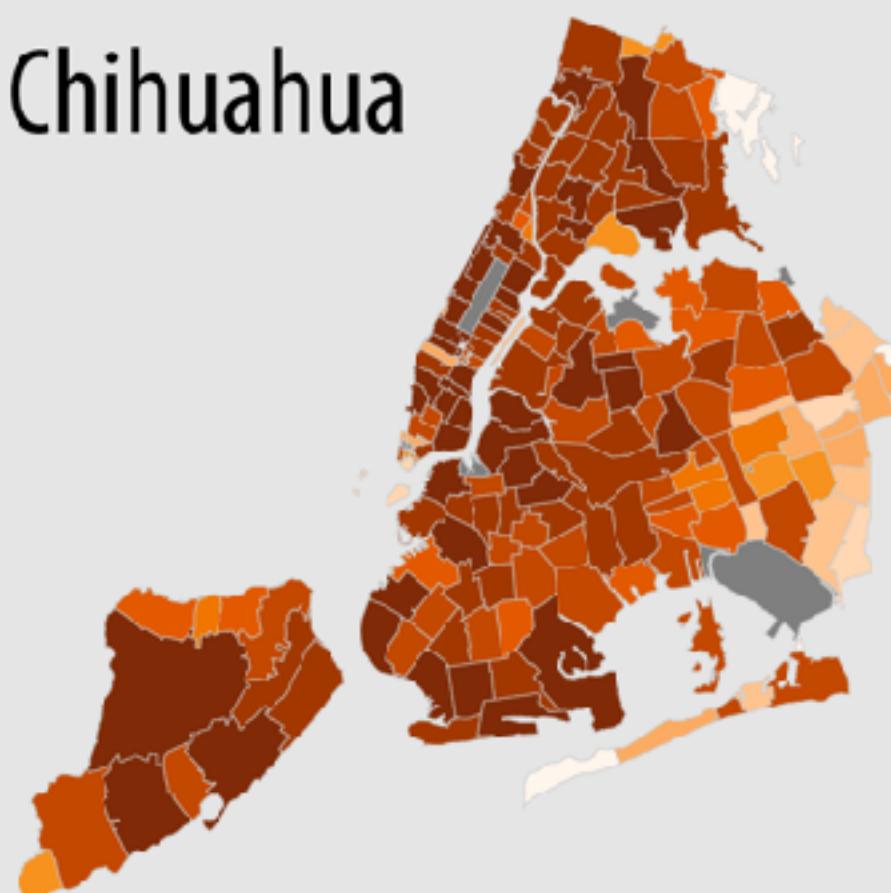
More Common

Less Common

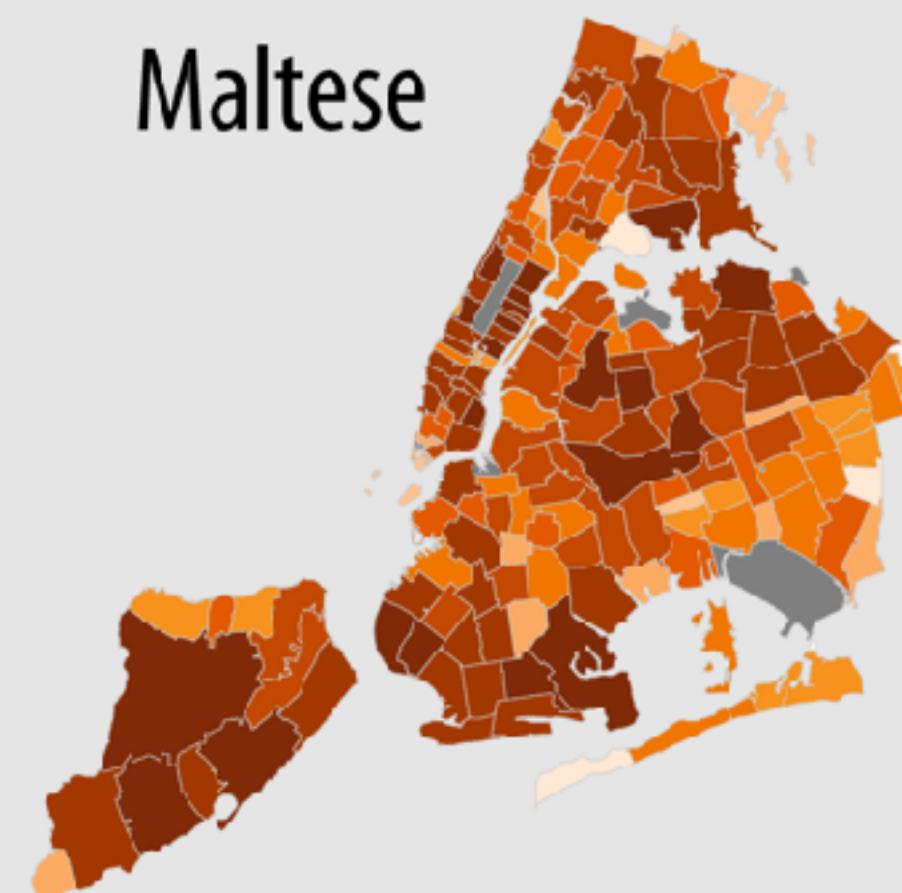
The most common
NYC breed is
“Mixed” or
“Unknown”



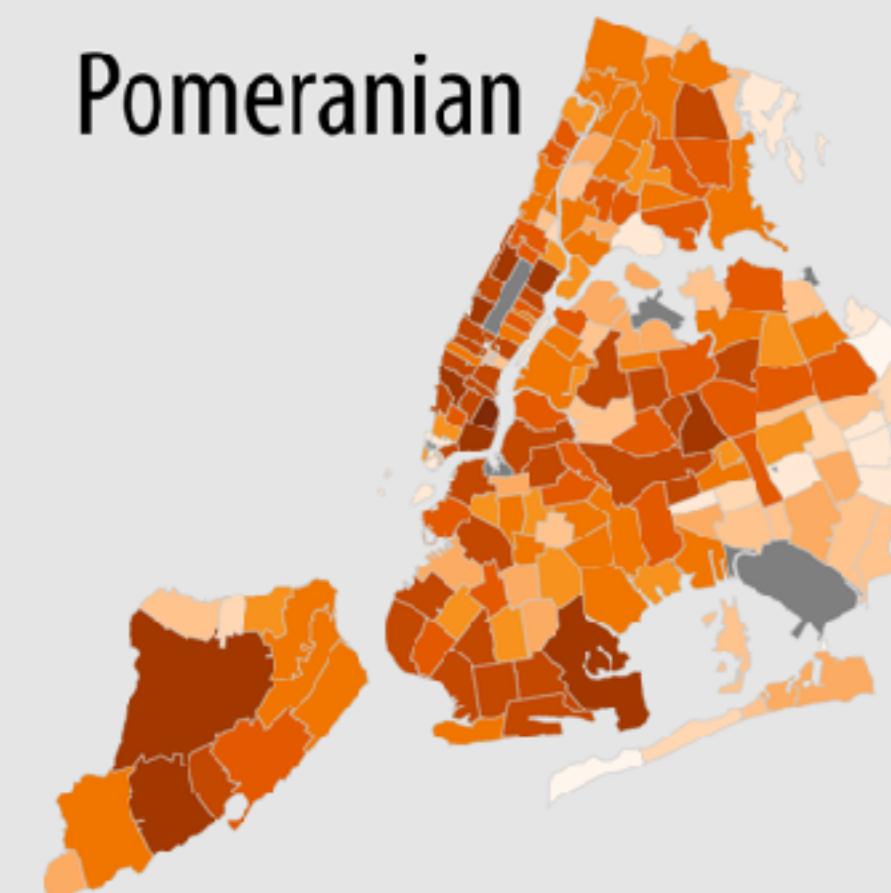
Chihuahua



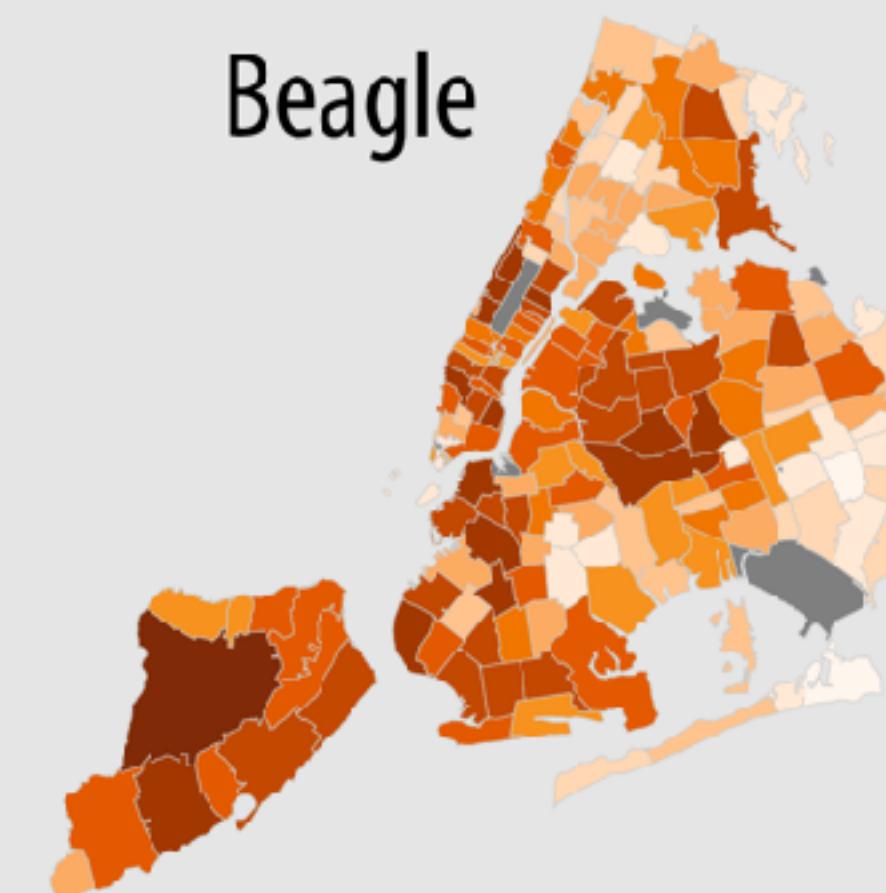
Maltese



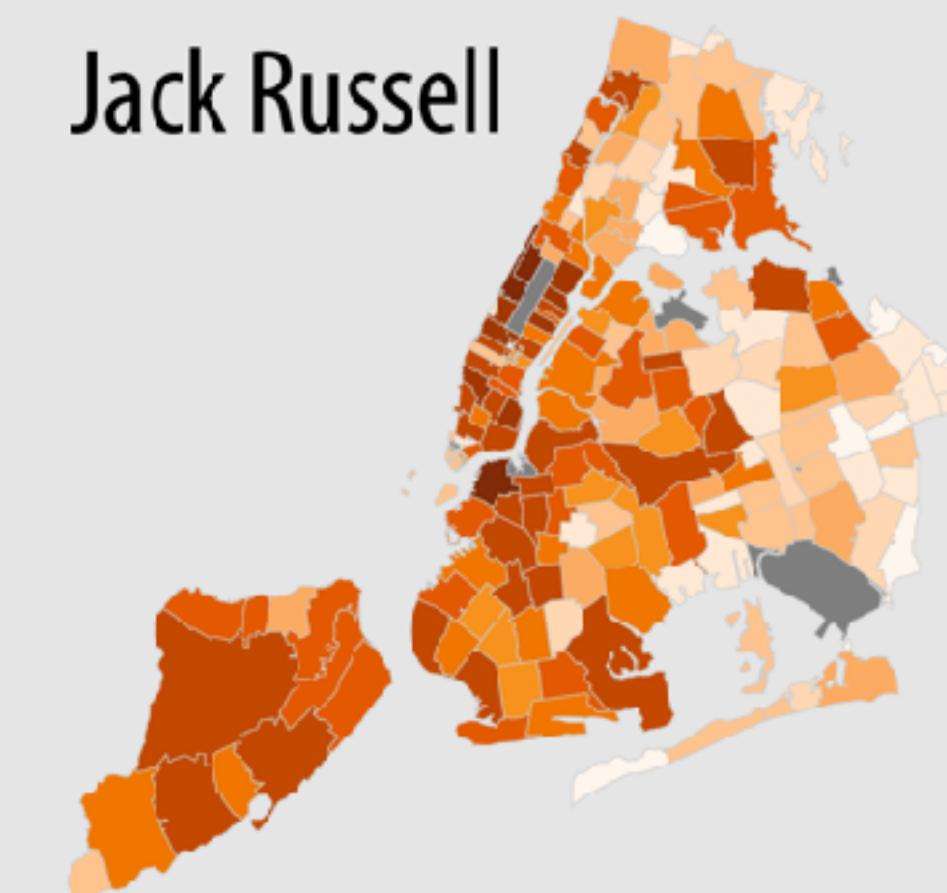
Pomeranian



Beagle

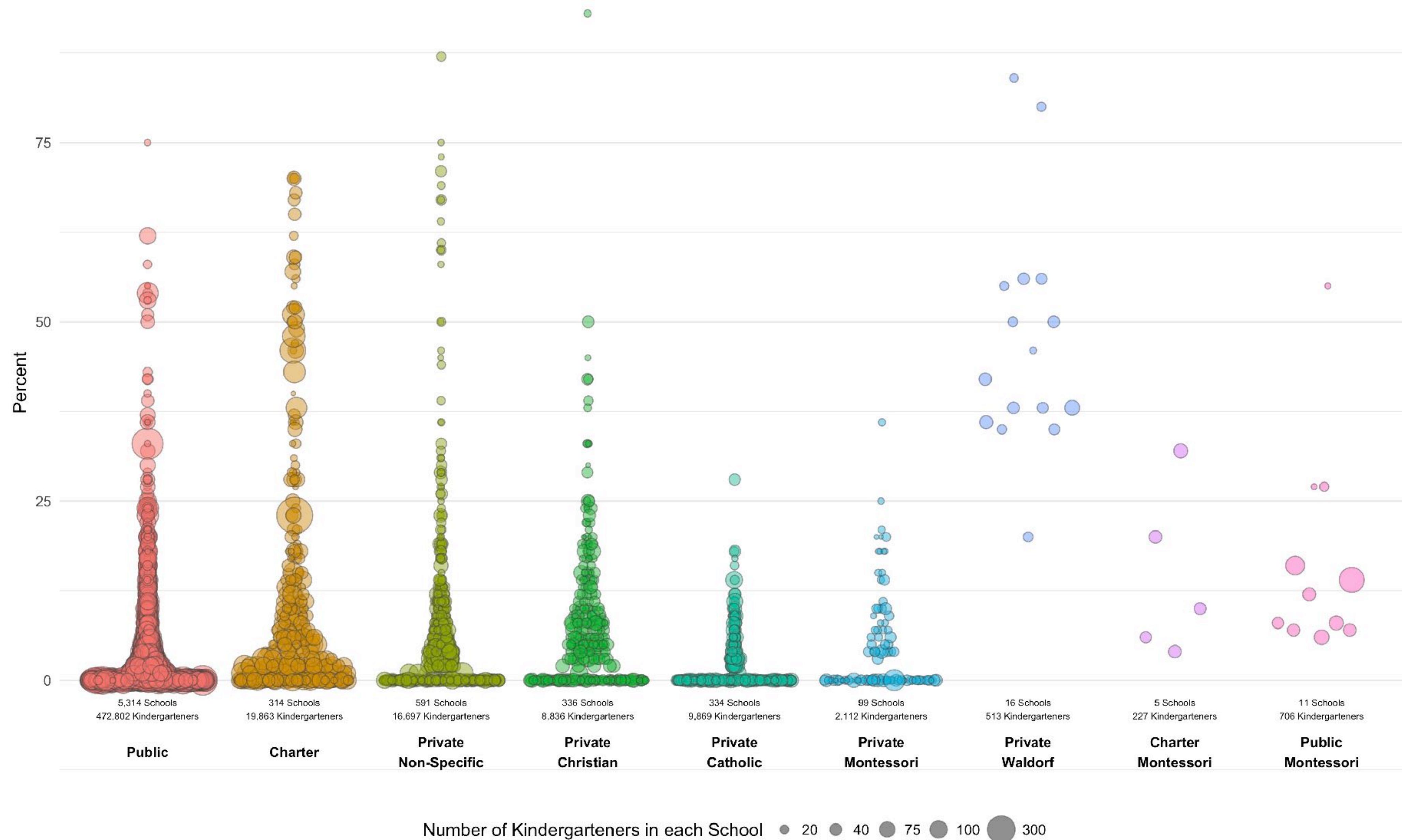


Jack Russell



Vaccination Exemption Rates in California Kindergartens

Percent of Kindergarteners with a Personal Belief Exemption, by Type and Size of School.



MORTALITY IN FRANCE 1816 – 2016

100 | Males

75

50

25

0

100 | Females

Age

75

50

25

0

1820

1845

1870

1895

1920

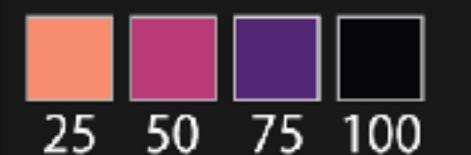
1945

1970

1995

2015

Death Rate Percentile



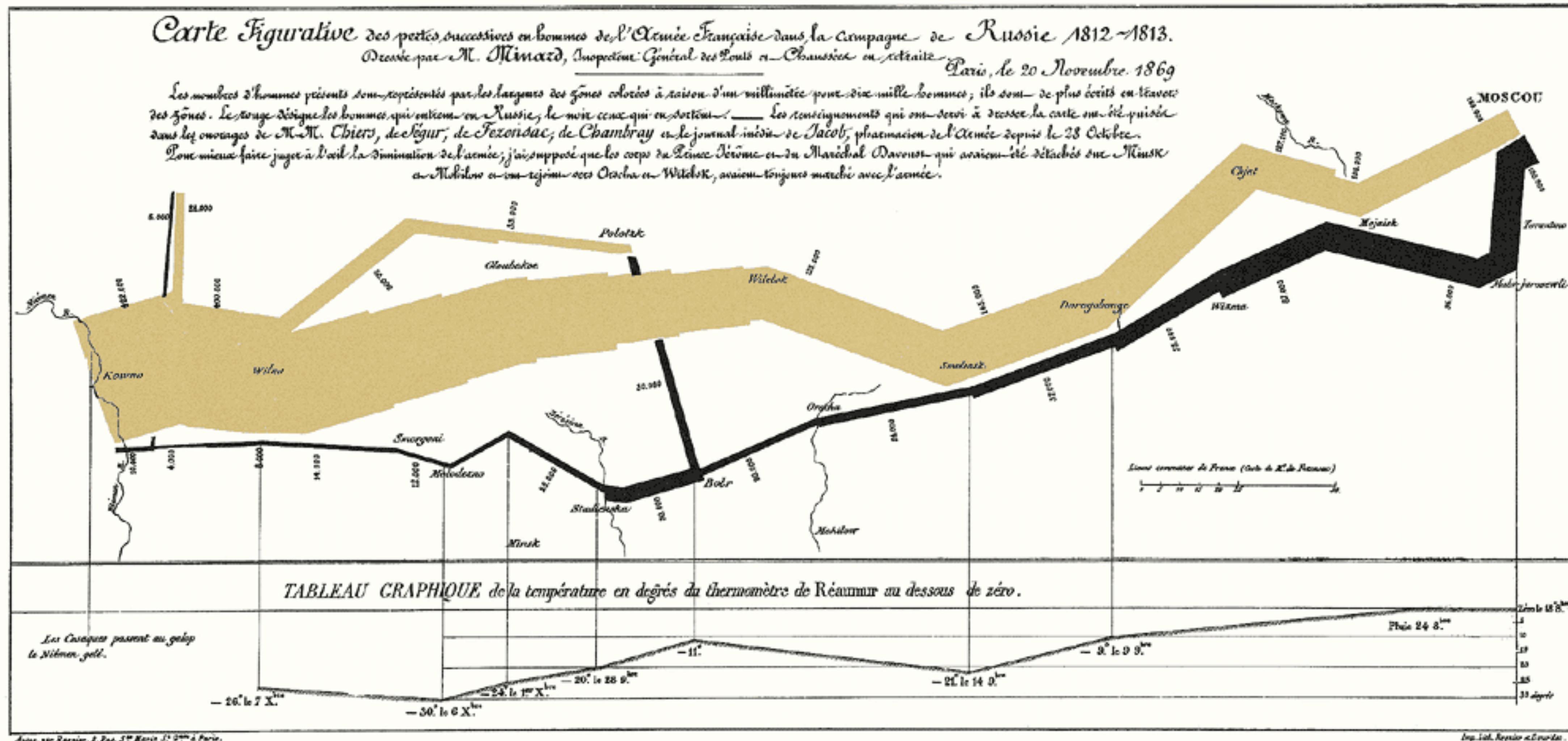
Mortality rates are calculated for each age in each year and binned by percentile. The darker the color at any particular point, the more people of that age die in that year. The lighter the color, the more people of that age survive in that year.

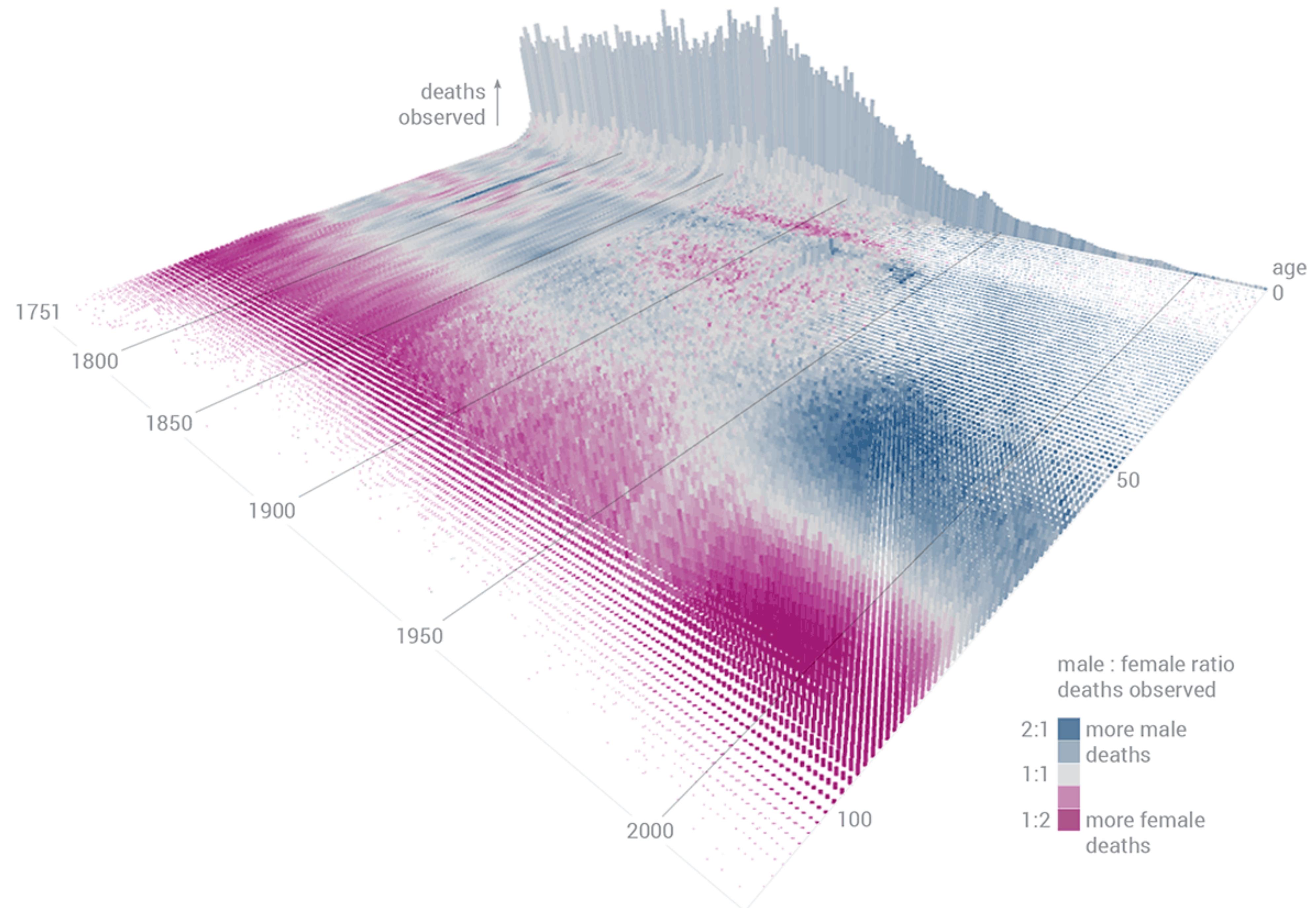
Historical trends are visible, such as the rapid decrease in infant mortality rates after World War II, as well as increased life expectancy overall. Specific events show up as vertical streaks in the graph. The death toll due to wars is evident for Males. Pandemics are also visible, most notably the 1918 Influenza pandemic and the death toll due to Smallpox outbreaks after the Franco-Prussian war of 1870.

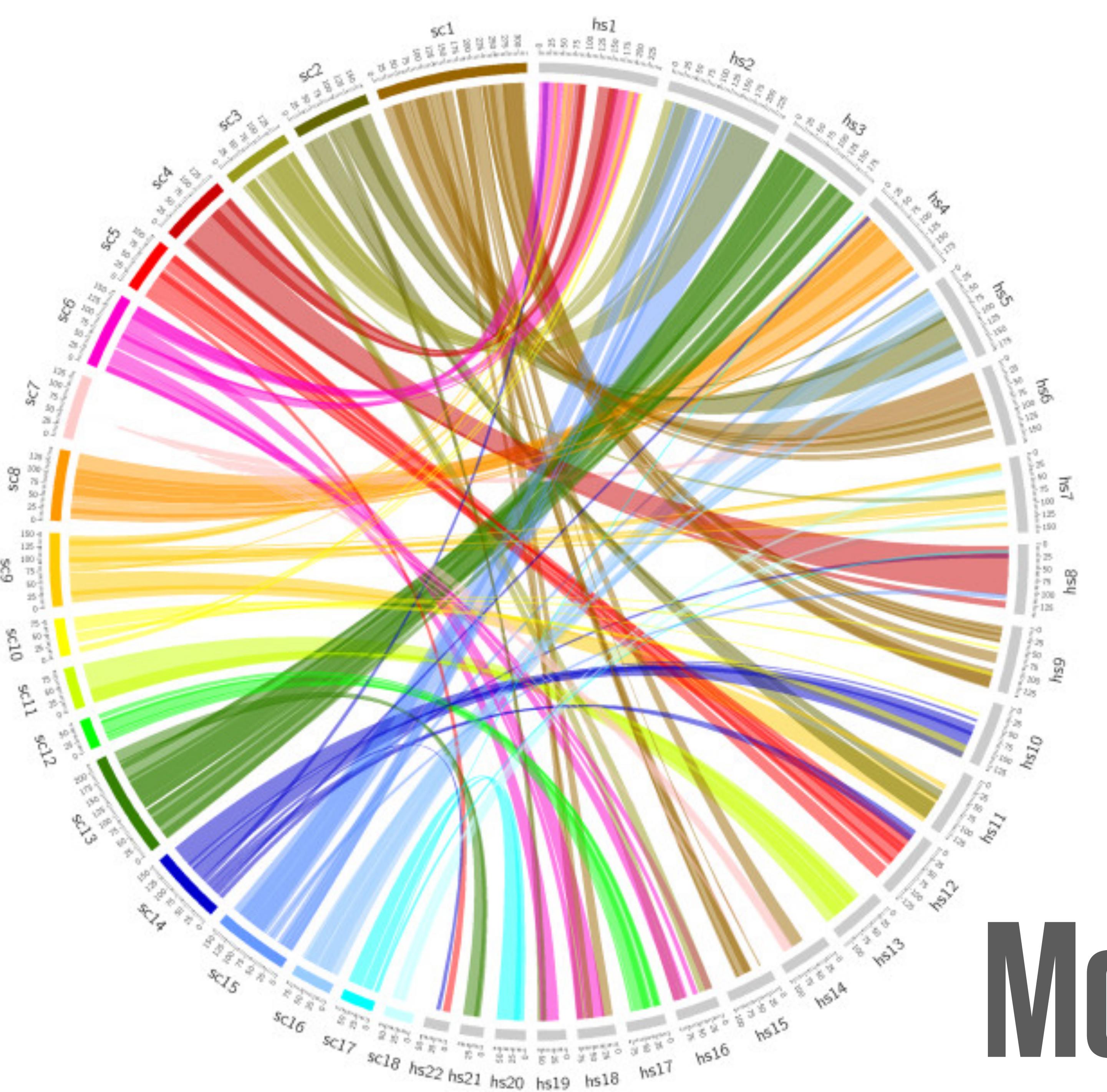
Diagonal streaks in the data are visible in some parts of the graph. These are artifacts due to the estimation of the mortality rate in some years. Single-year-of-age figures prior to 1900 are calculated from five-year age groups, as no single-year data is available in the original mortality tabulations from which the rates were derived.

Kieran Healy / [socrata.co](#)
Data Visualization: A Practical Introduction
is published by Princeton University Press

Unicorns ...







... or
Monsters