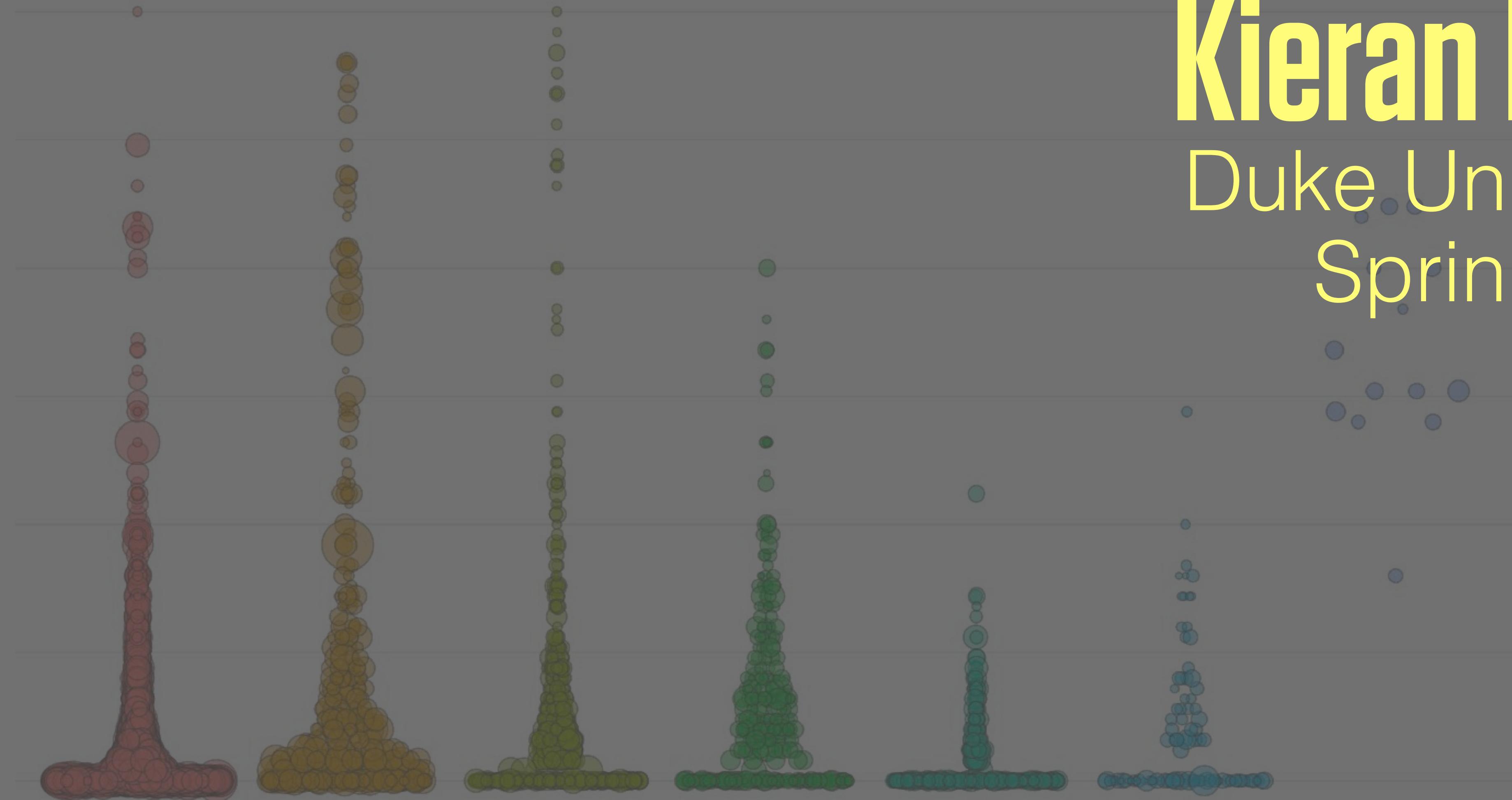


DATA VISUALIZATION

Kieran Healy
Duke University
Spring 2020





RStudio

Get Up & Running

Console

~/Desktop/

> |

library(tidyverse)

library(socviz)

Choose one, or pick a folder yourself

setup_course_notes("~/Desktop/")

setup_course_notes("~/Documents/")

Option 1



```
library(usethis)
```

```
# Read the options for where to put it
```

```
use_course("http://kjhealy.co/dv/dataviz_course_notes.zip")
```

Option 2

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

~/Desktop/dataviz_course_notes - RStudio

01_introduction.Rmd x Addins x

```
1 ---  
2 title: "Data Visualization"  
3 author: "Kieran Healy"  
4 date: "01-December-2018"  
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.
```

10:461 # Data Visualization Notes R Markdown

Console Terminal Jobs

~/Desktop/dataviz_course_notes/ ↵

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

> |

Environment History Files Build

Import Dataset Global Environment

Environment is empty

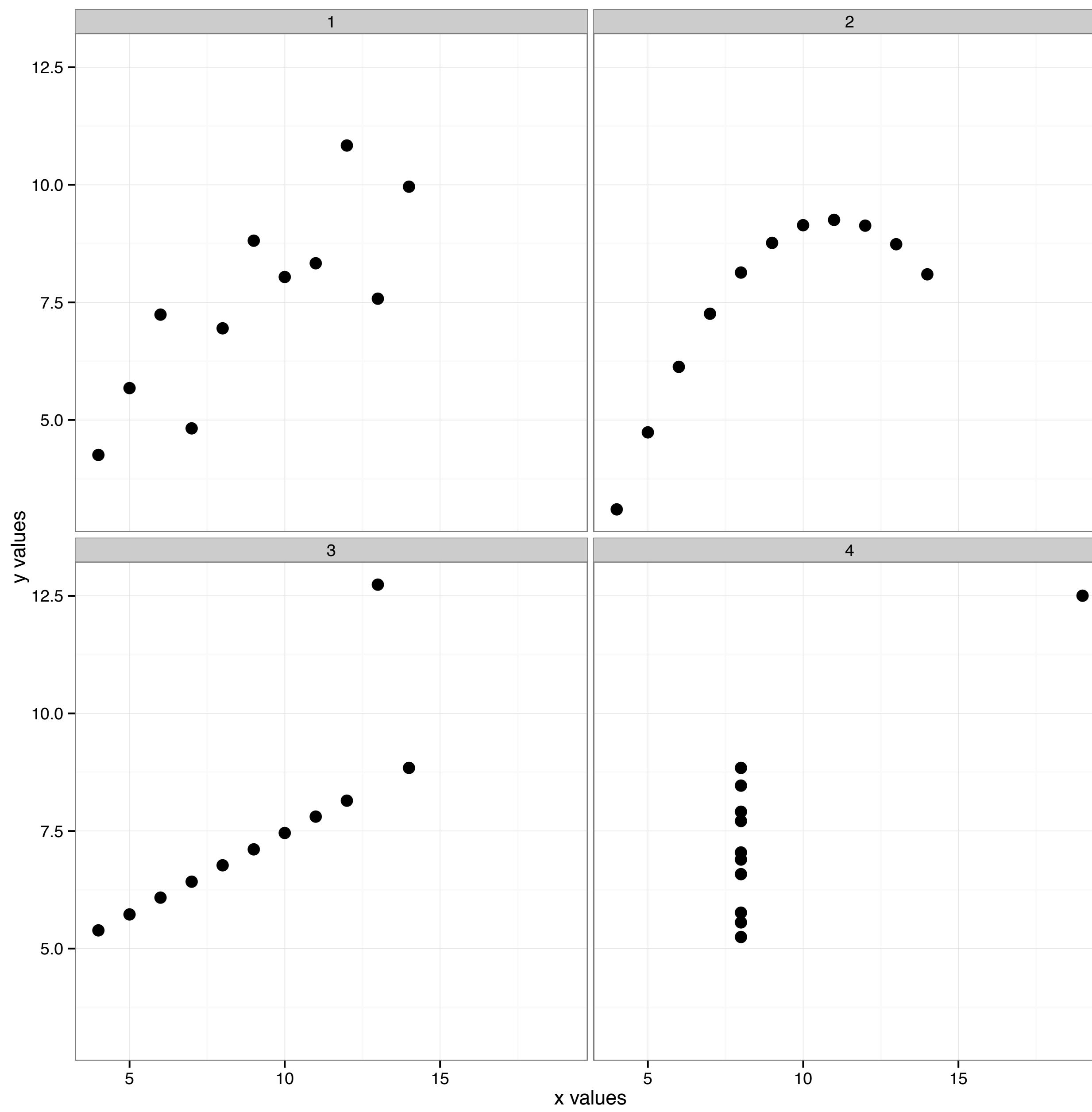
Plots Connections Packages Help Viewer

Zoom Export

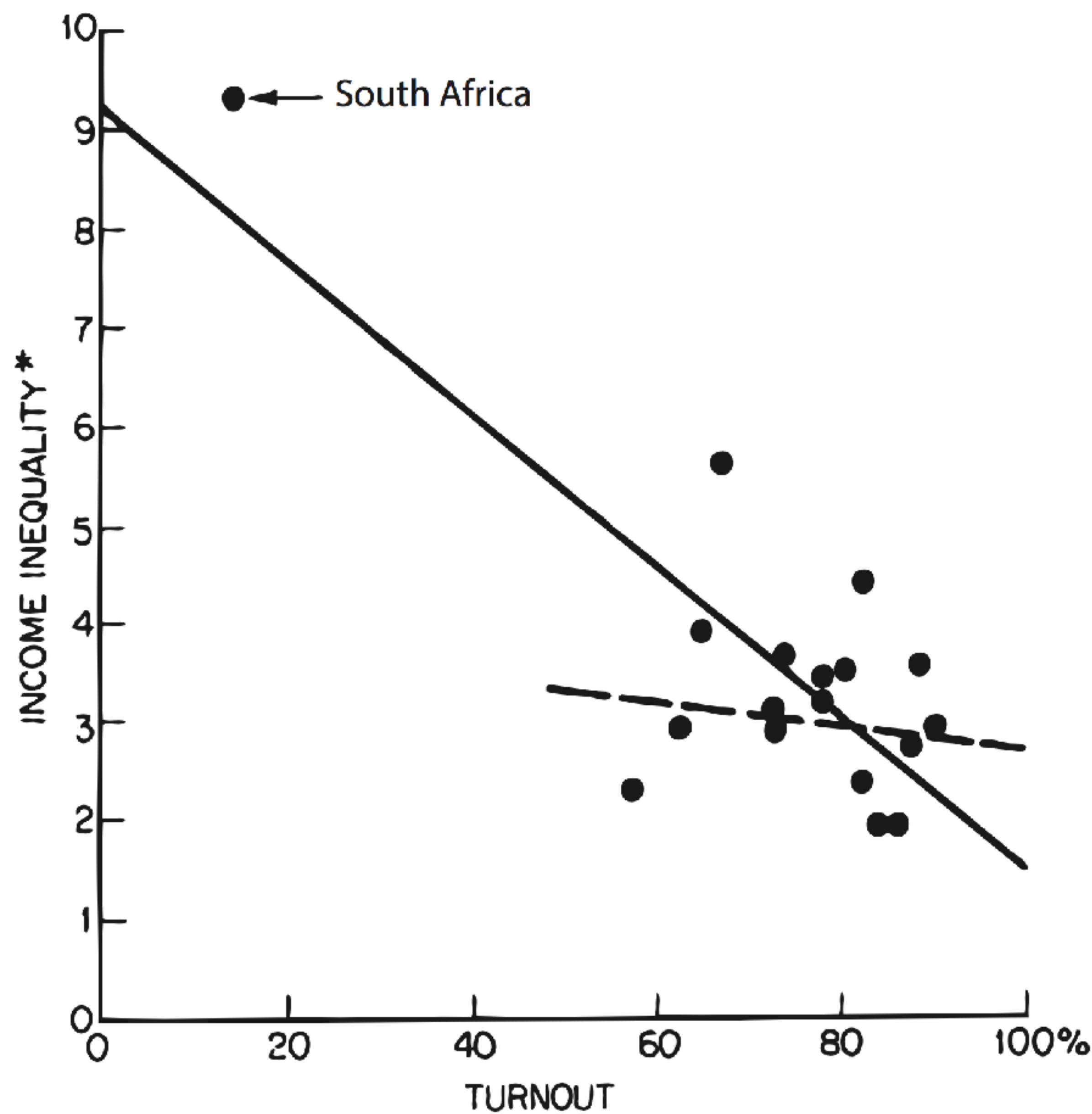
LOOK AT YOUR DATA



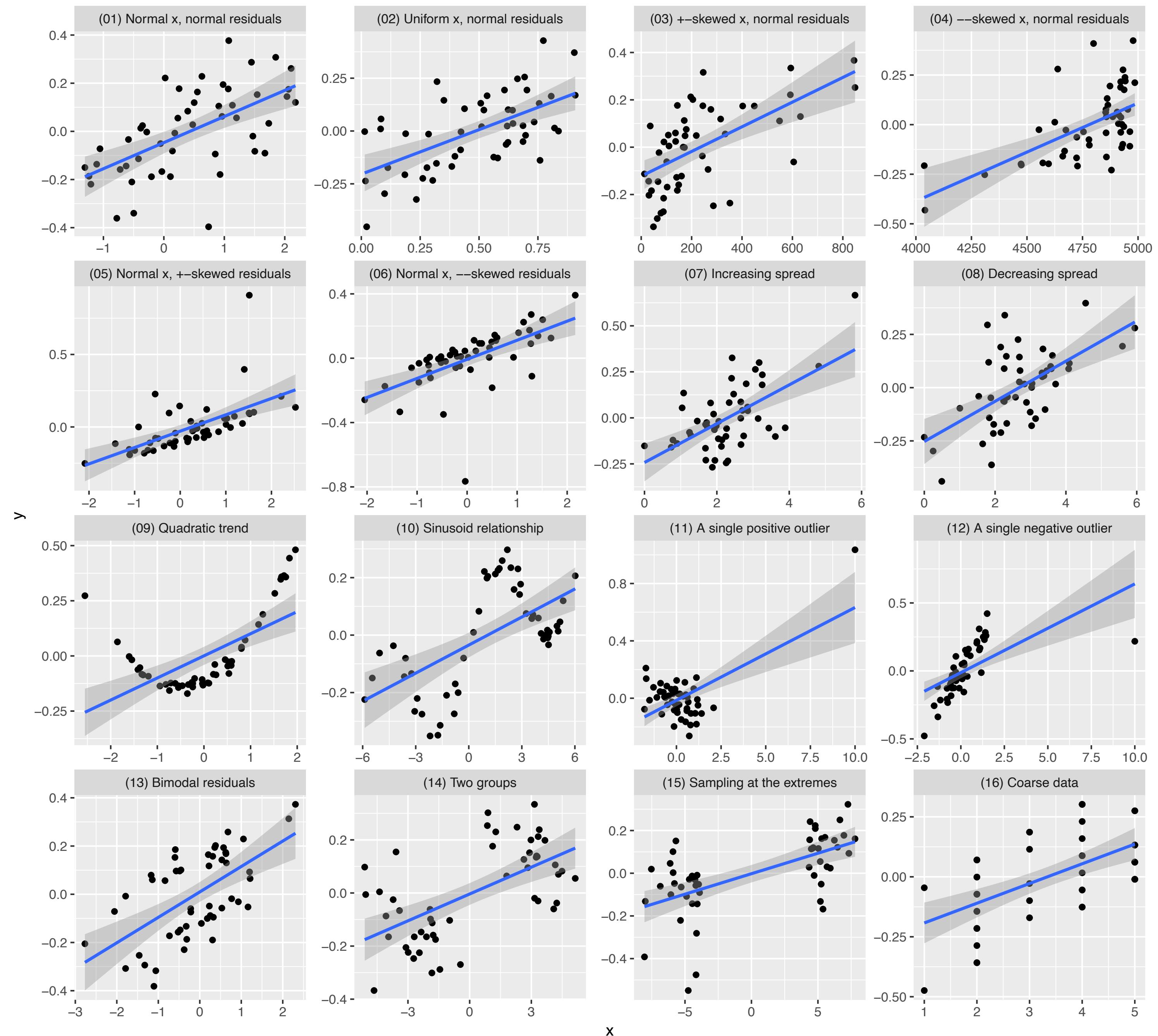
SEEING THINGS

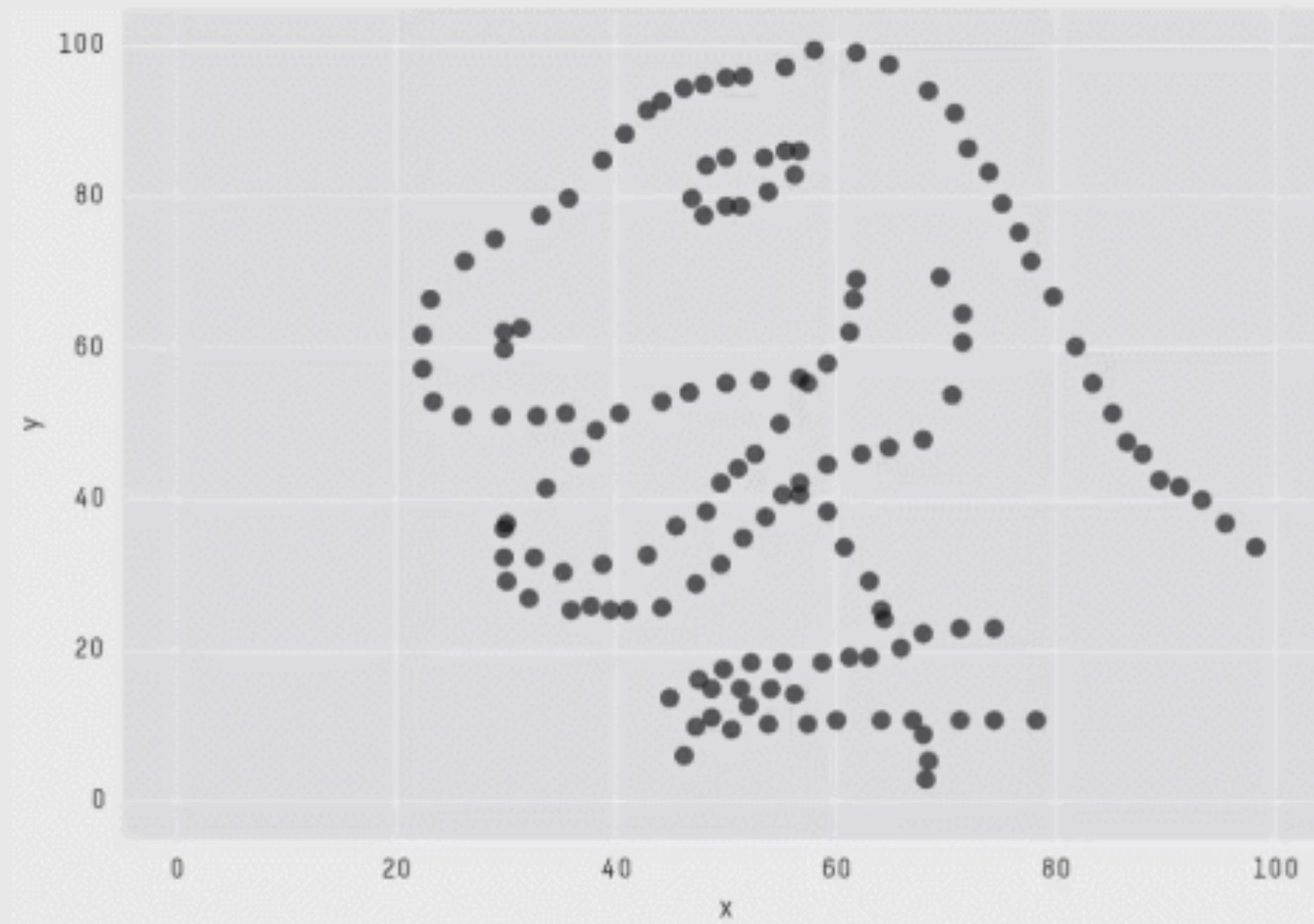


For all panels, N=11; Mean=7.5; Regression: $Y=3 + 0.5(X)$; SE of slope estimate: 0.118, t=4.24; Sum of Squares (X-X): 100; r=0.82.



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

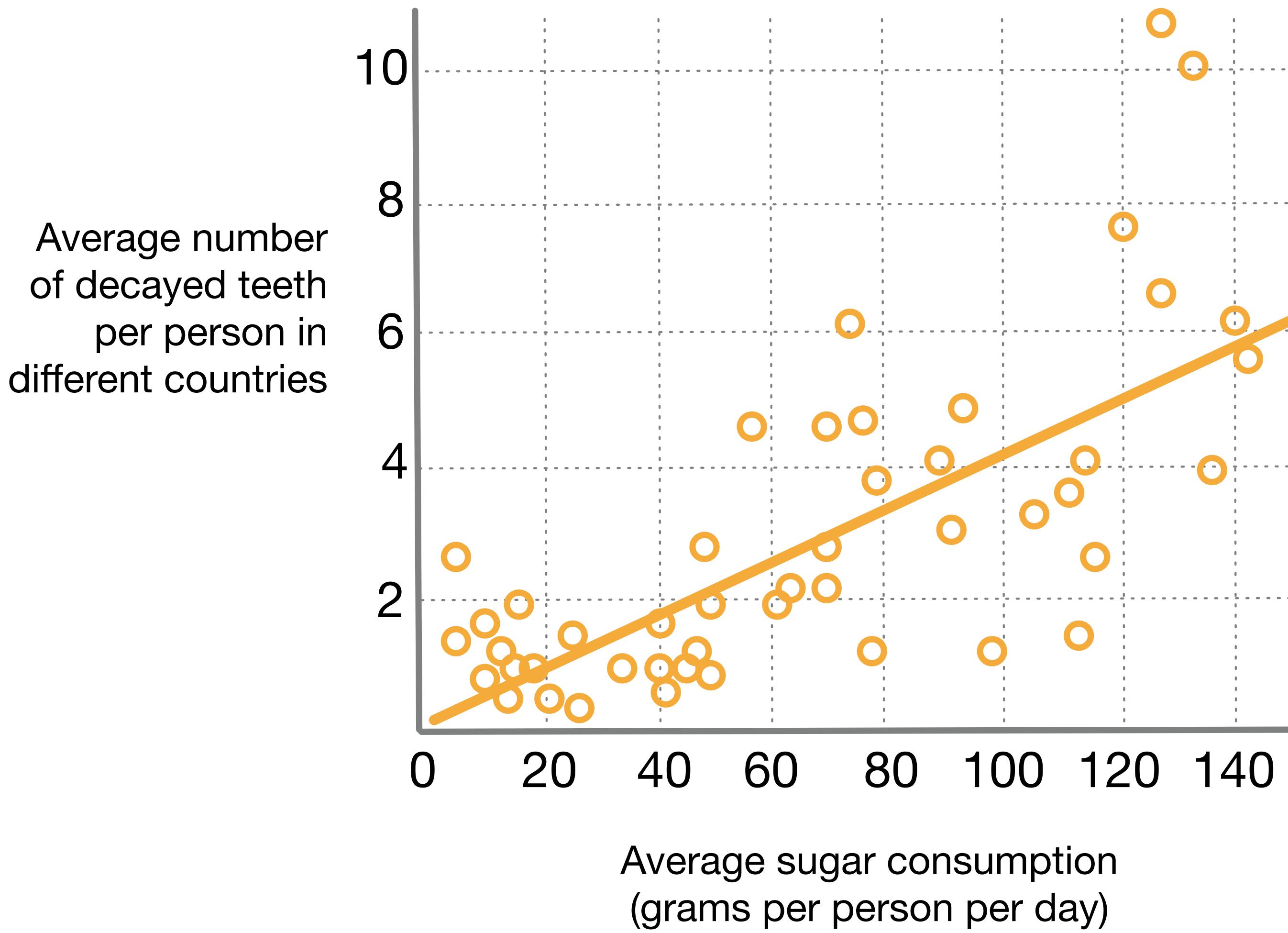




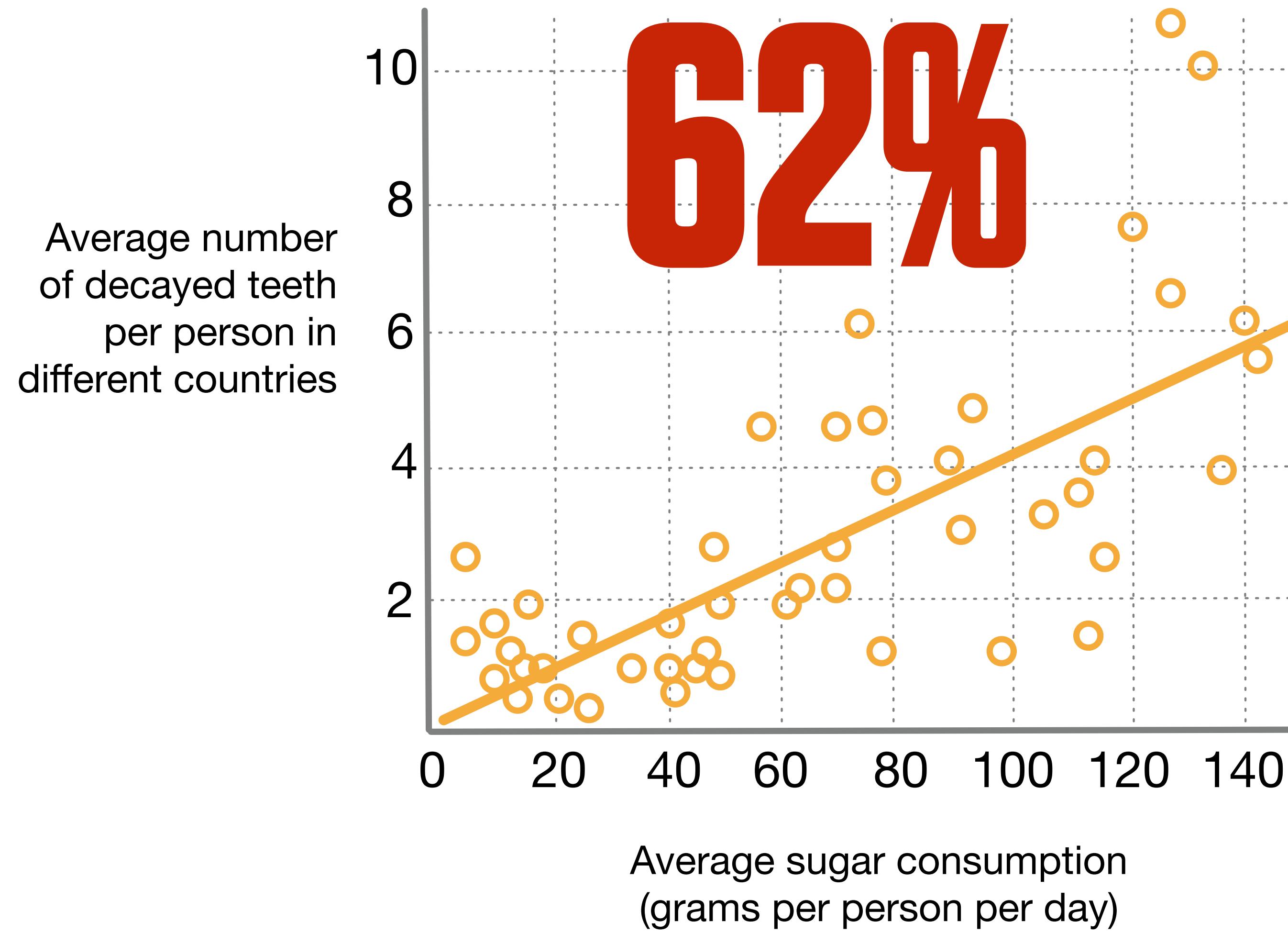
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Cairo; Matejka & Fitzmaurice

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



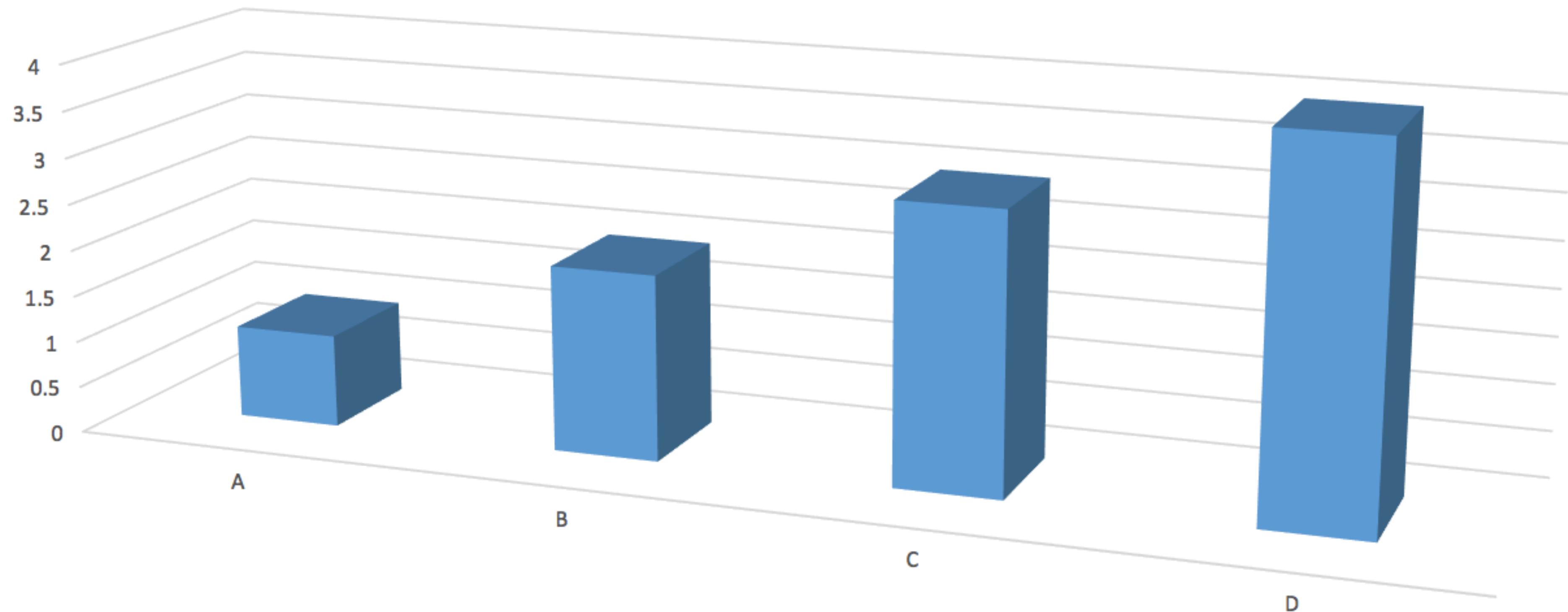
Which of the following statements best describes the data in the graph below?

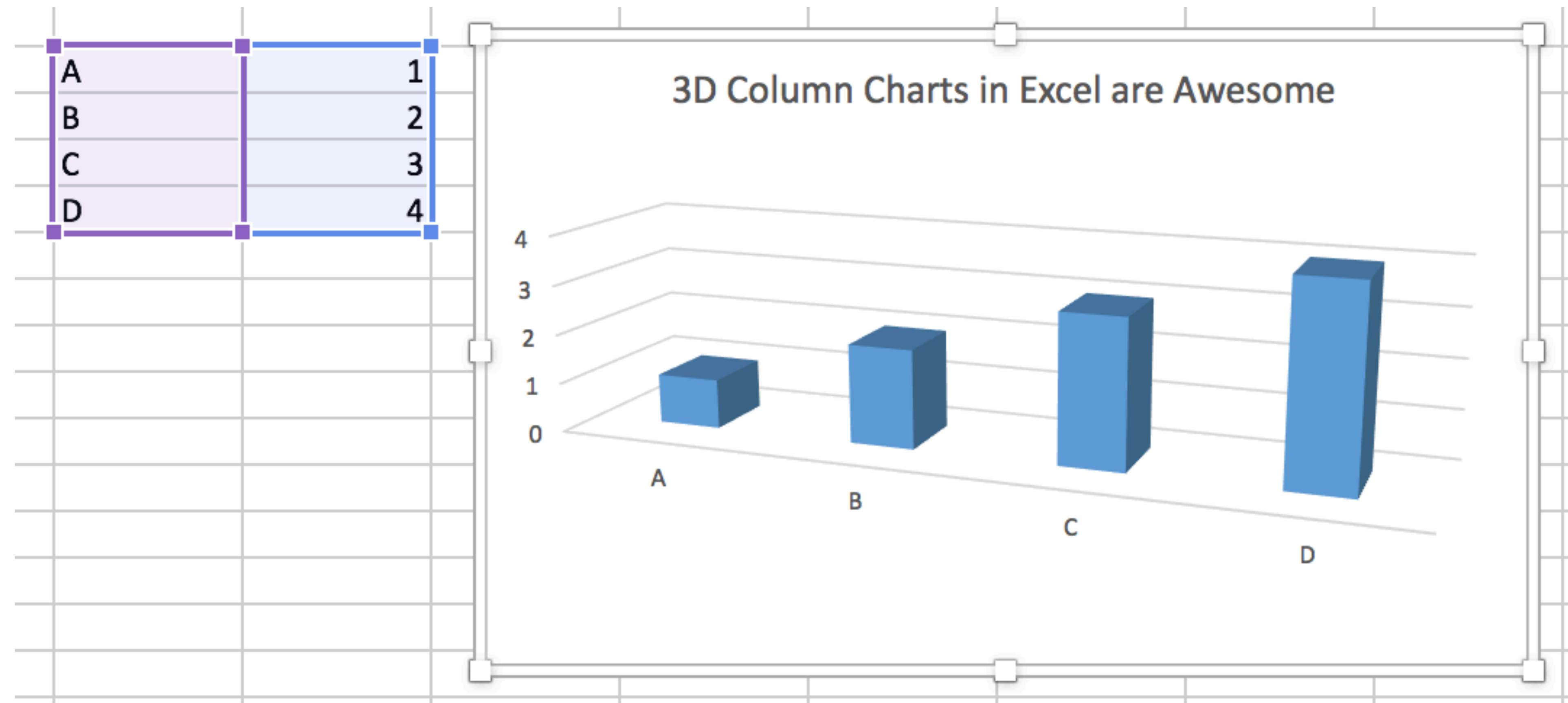


- A. In recent years, the rate of cavities has increased in many countries
- B. In some countries, people brush their teeth more frequently than in other countries
- C. The more sugar people eat, the more likely they are to get cavities**
- D. In recent years, the consumption of sugar has increased in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

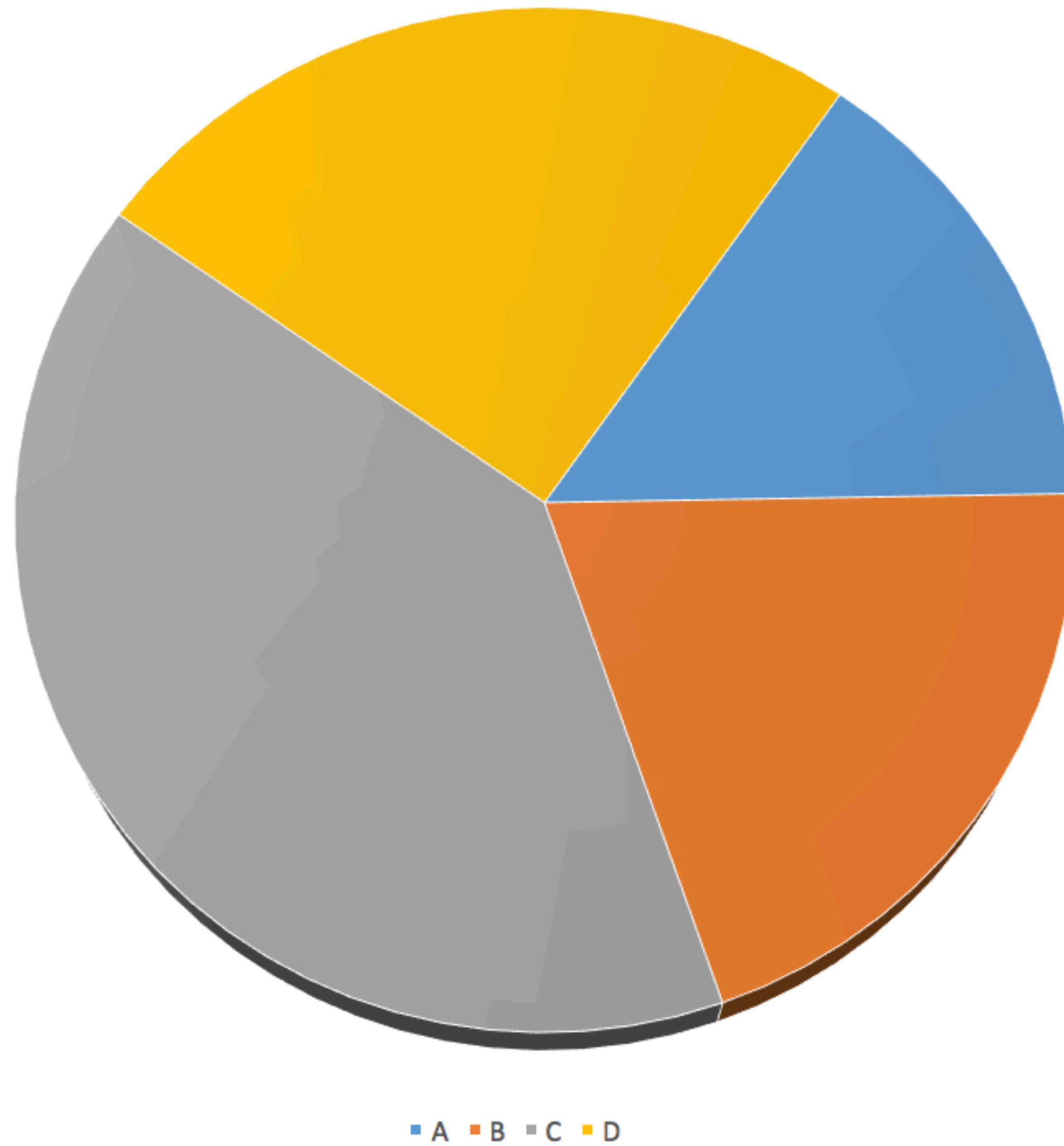
NOT
SEEING THINGS

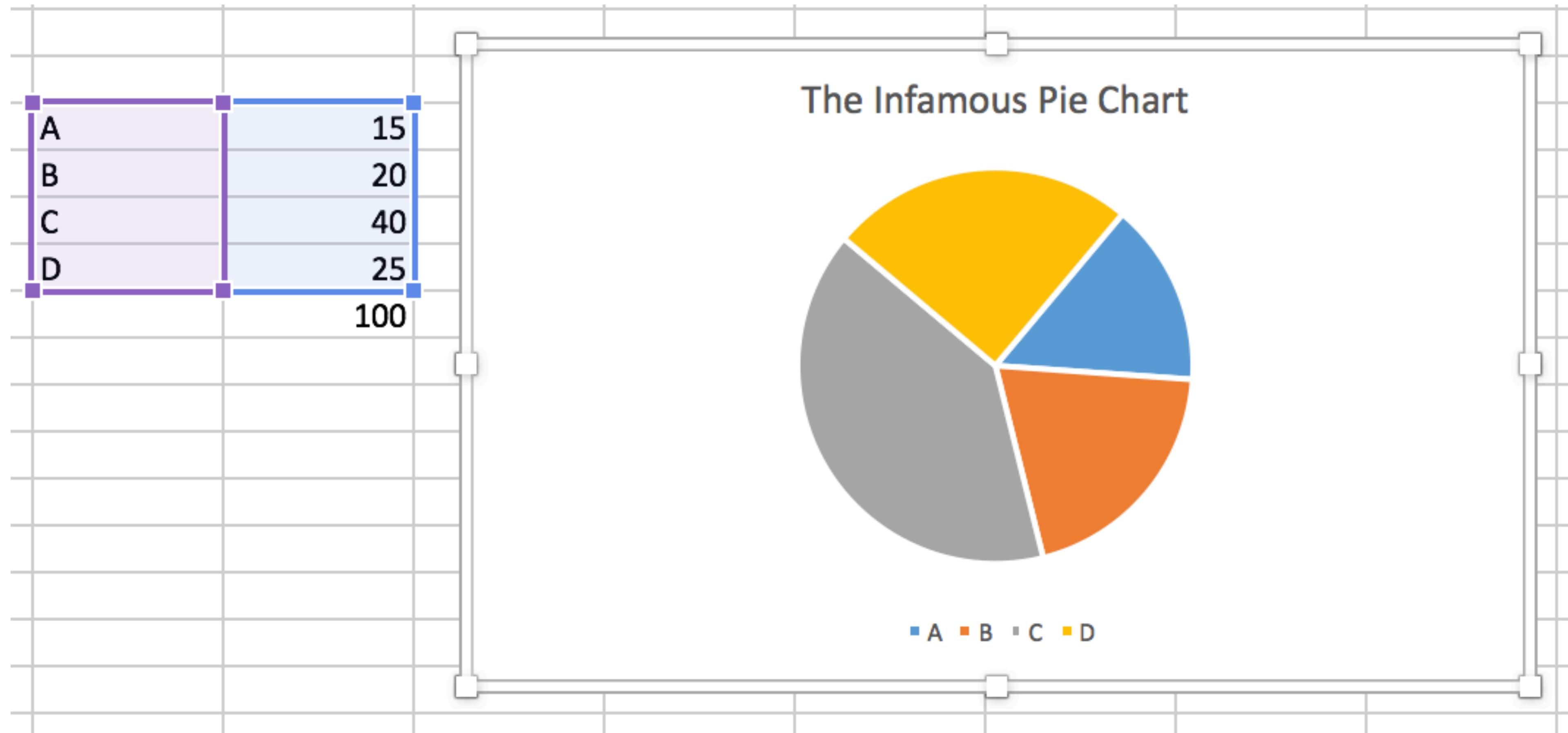
3D Column Charts in Excel are Awesome





The Infamous Pie Chart





BAD TASTE

BAD DATA

BAD PERCEPTION

TASTE

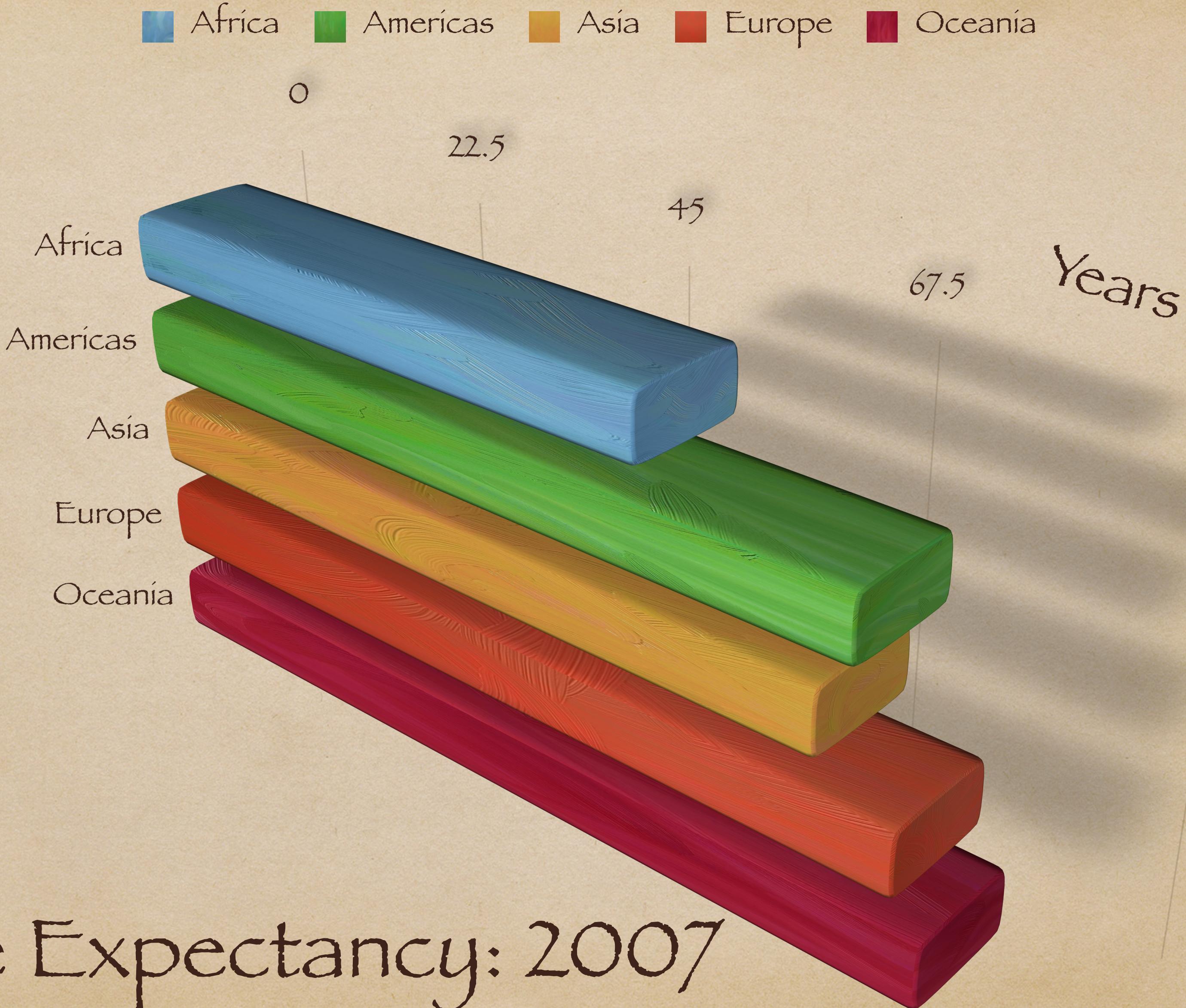
SIMPLIFY, SIMPLIFY?

The Visual Display
of Quantitative Information

EDWARD R. TUFTE

Life Expectancy: 2007

CONTINENT



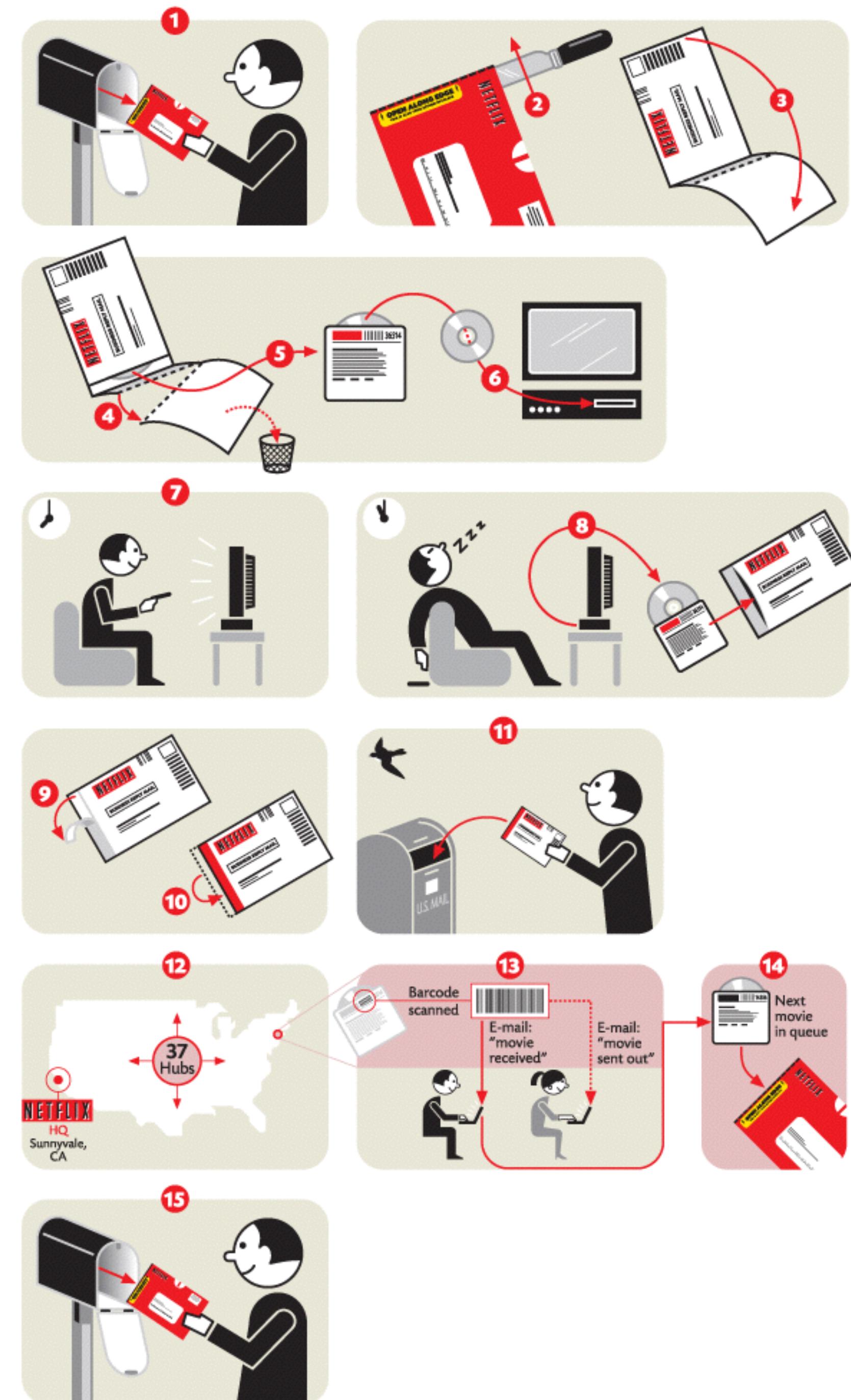
Years

MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions



HOW NETFLIX WORKS

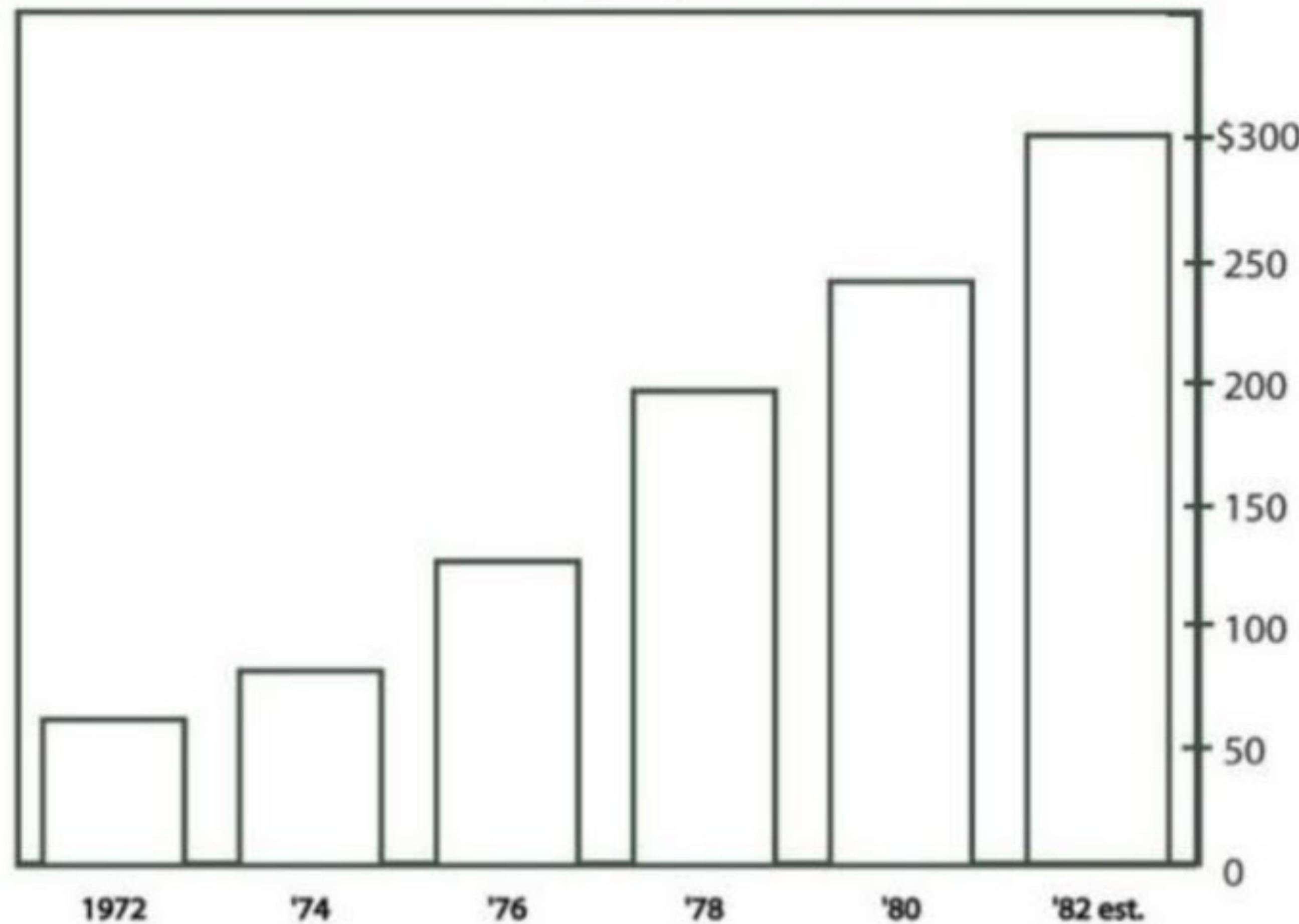


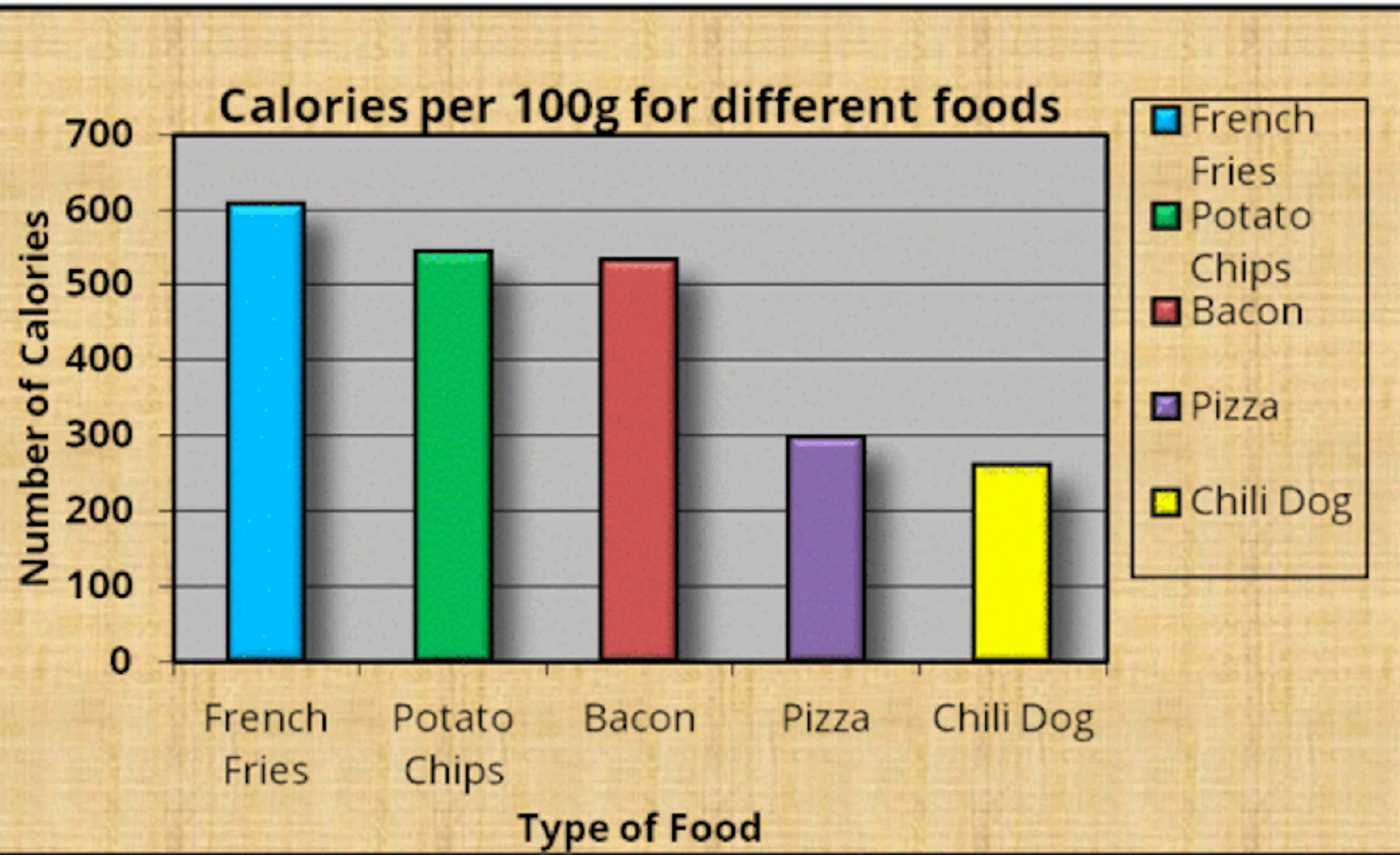
MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions

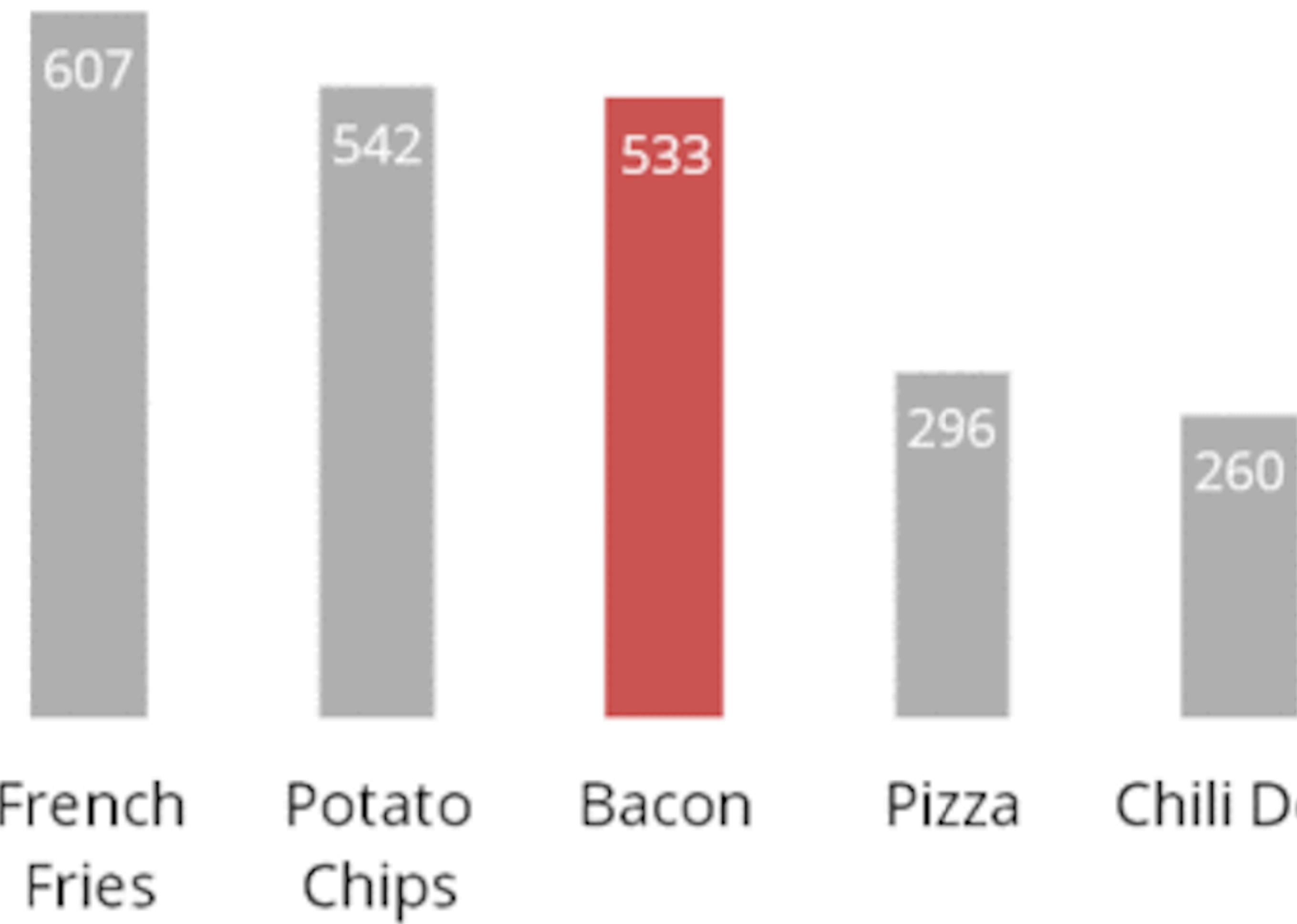


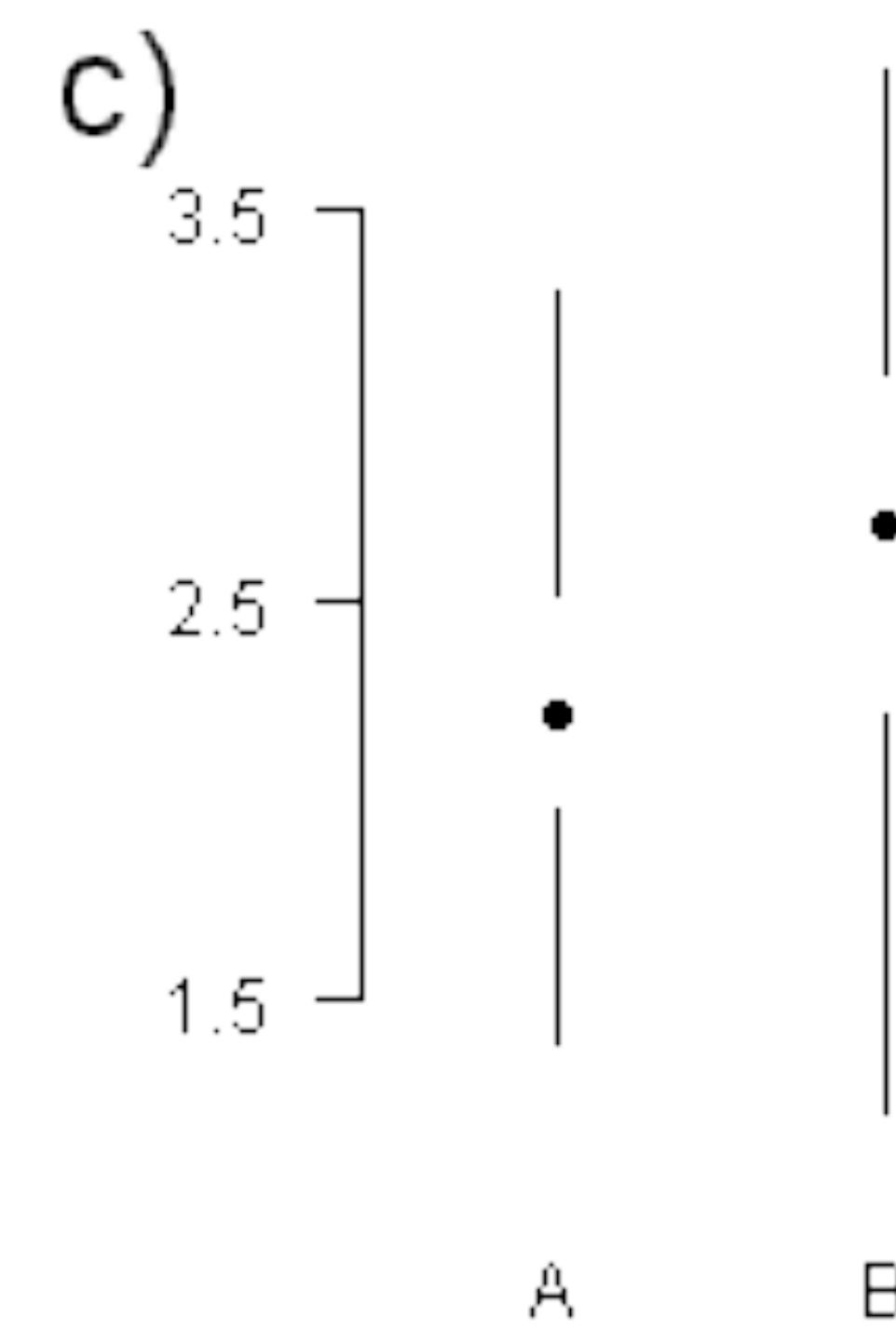
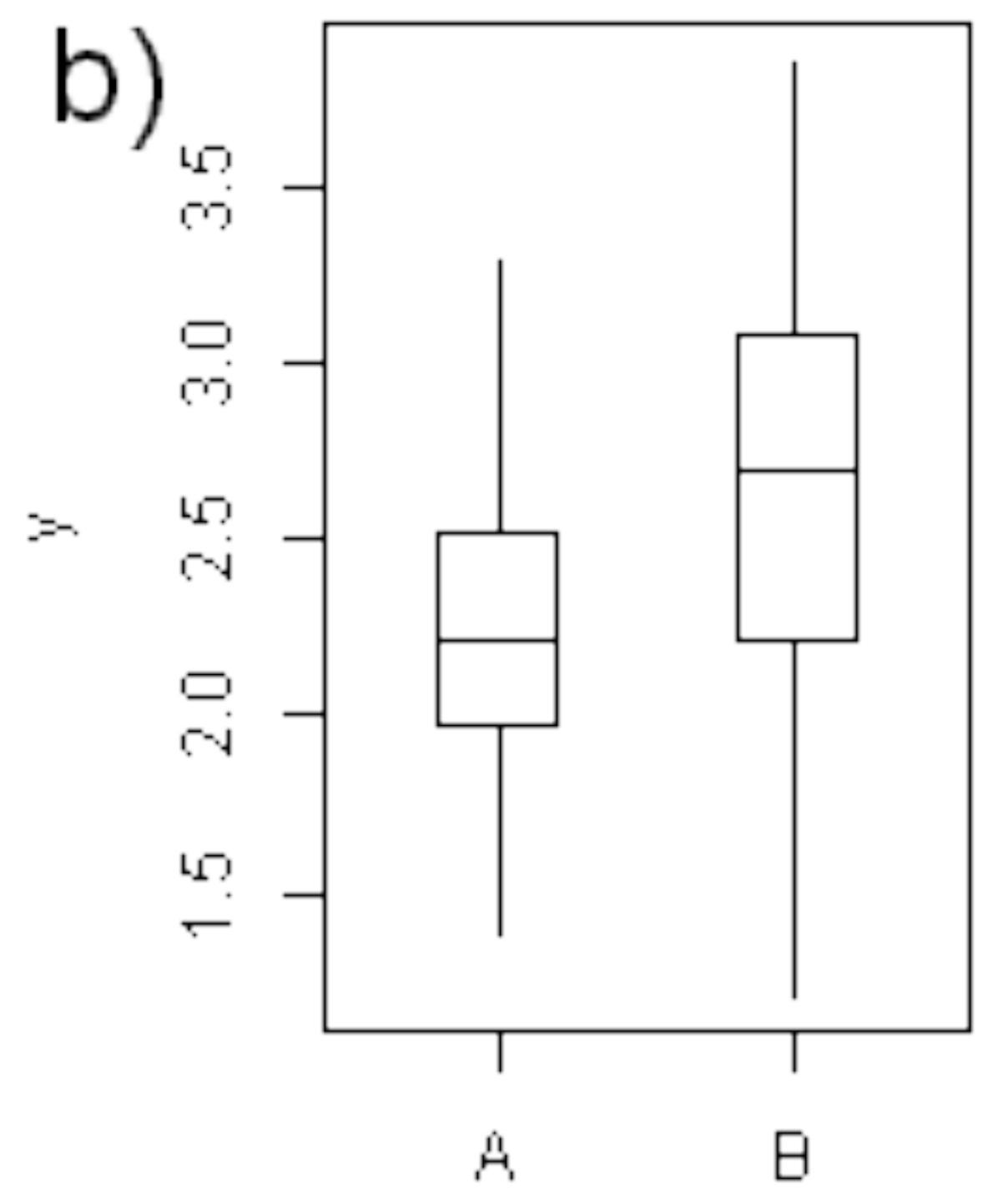
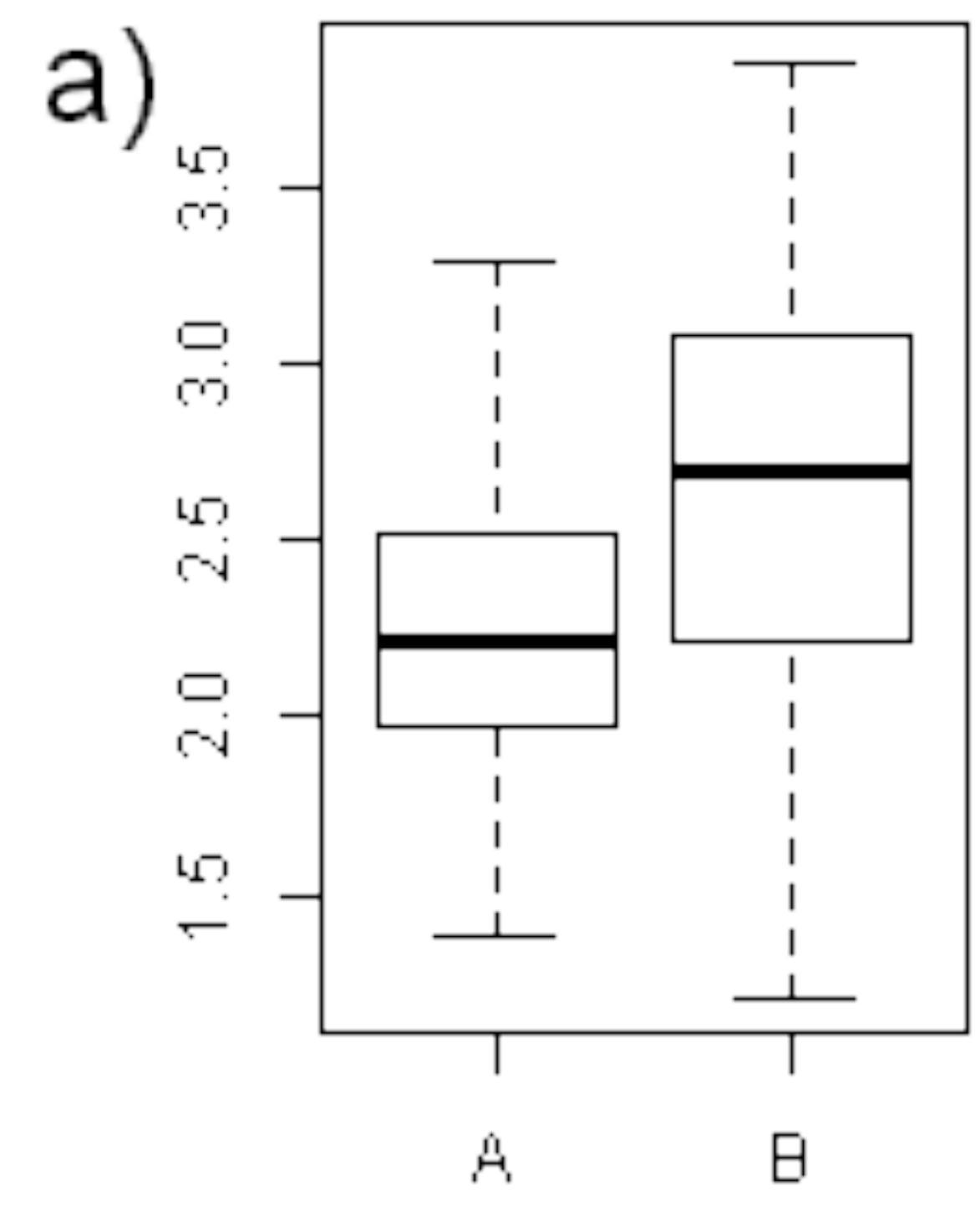
MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions

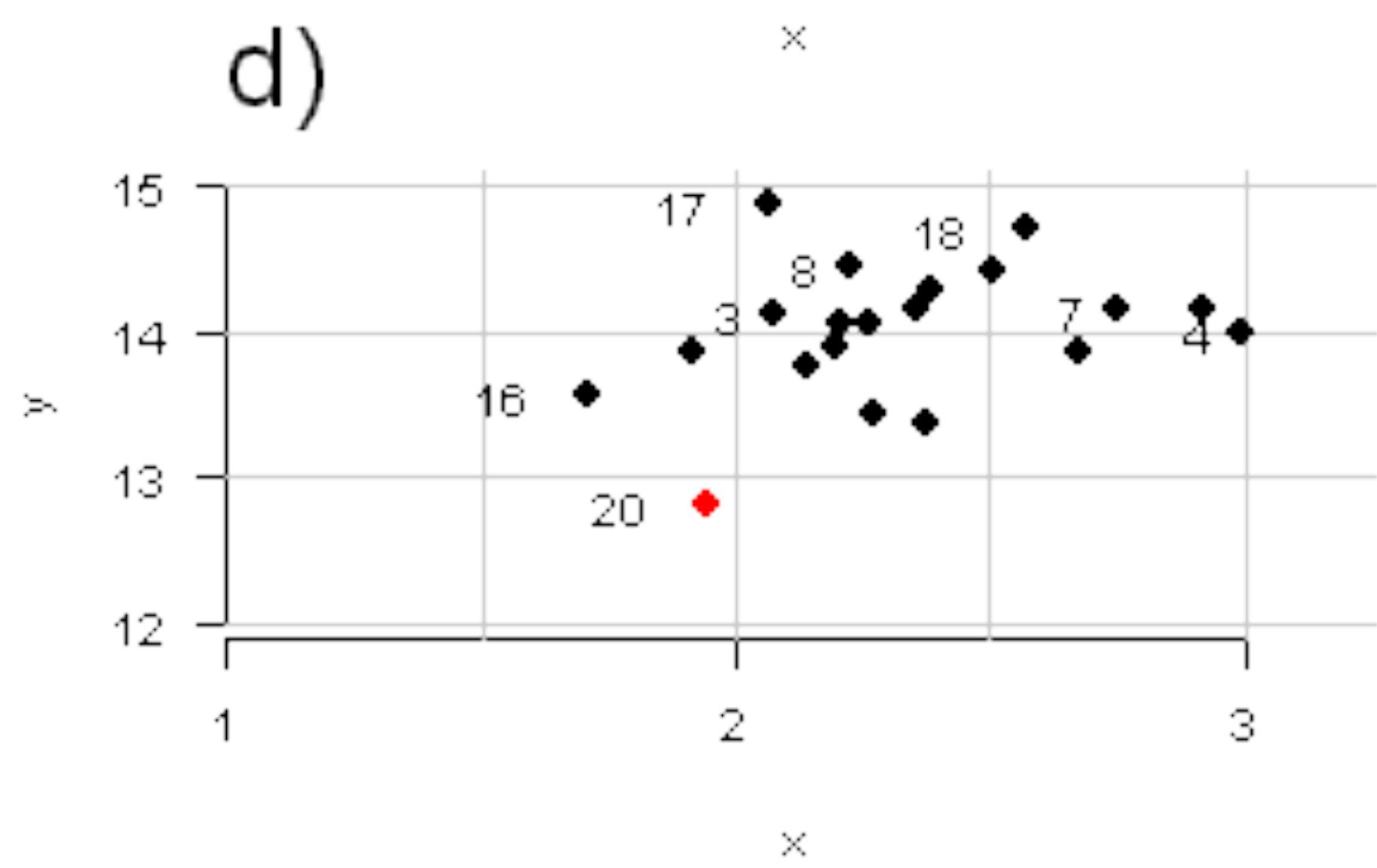
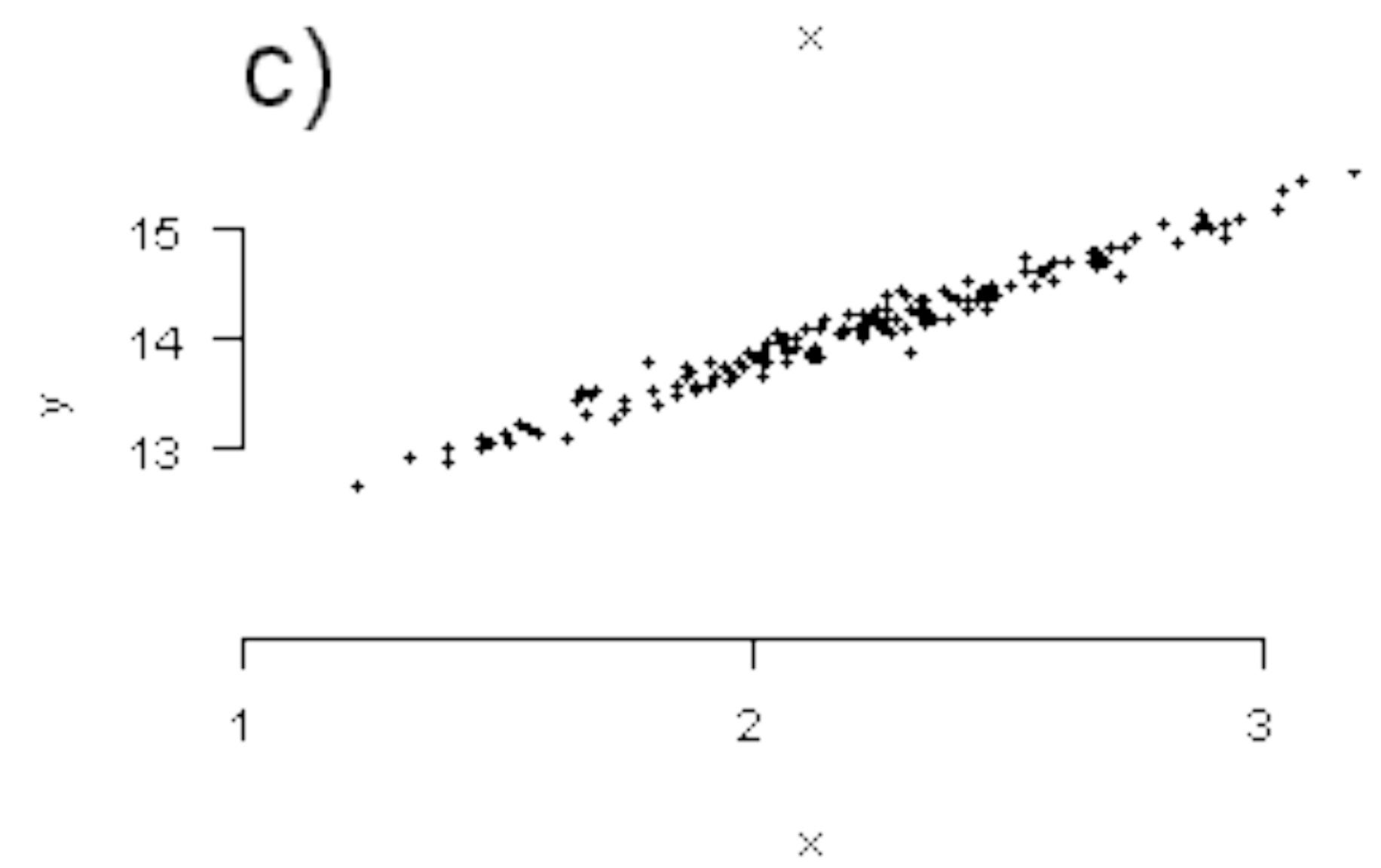
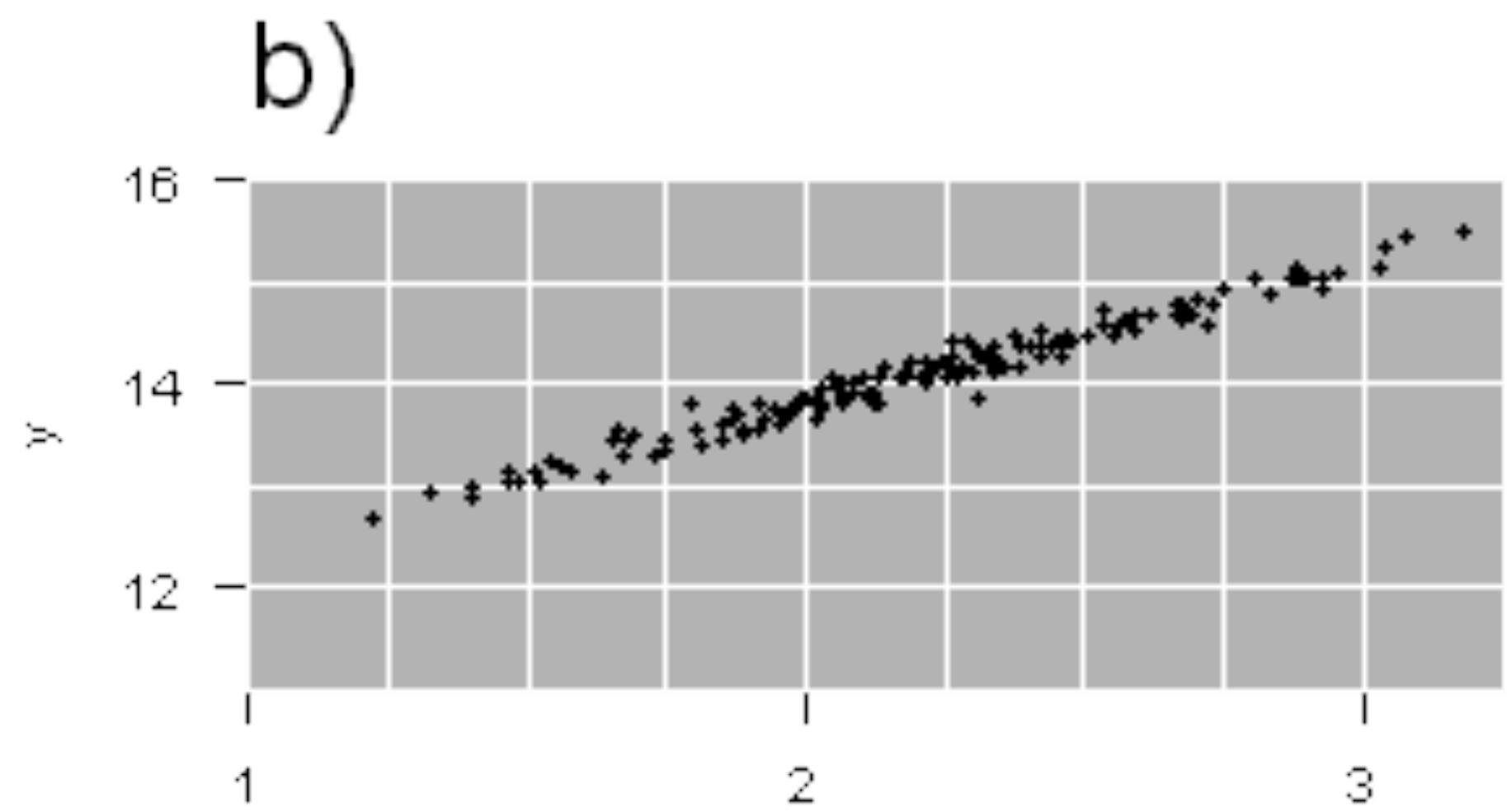
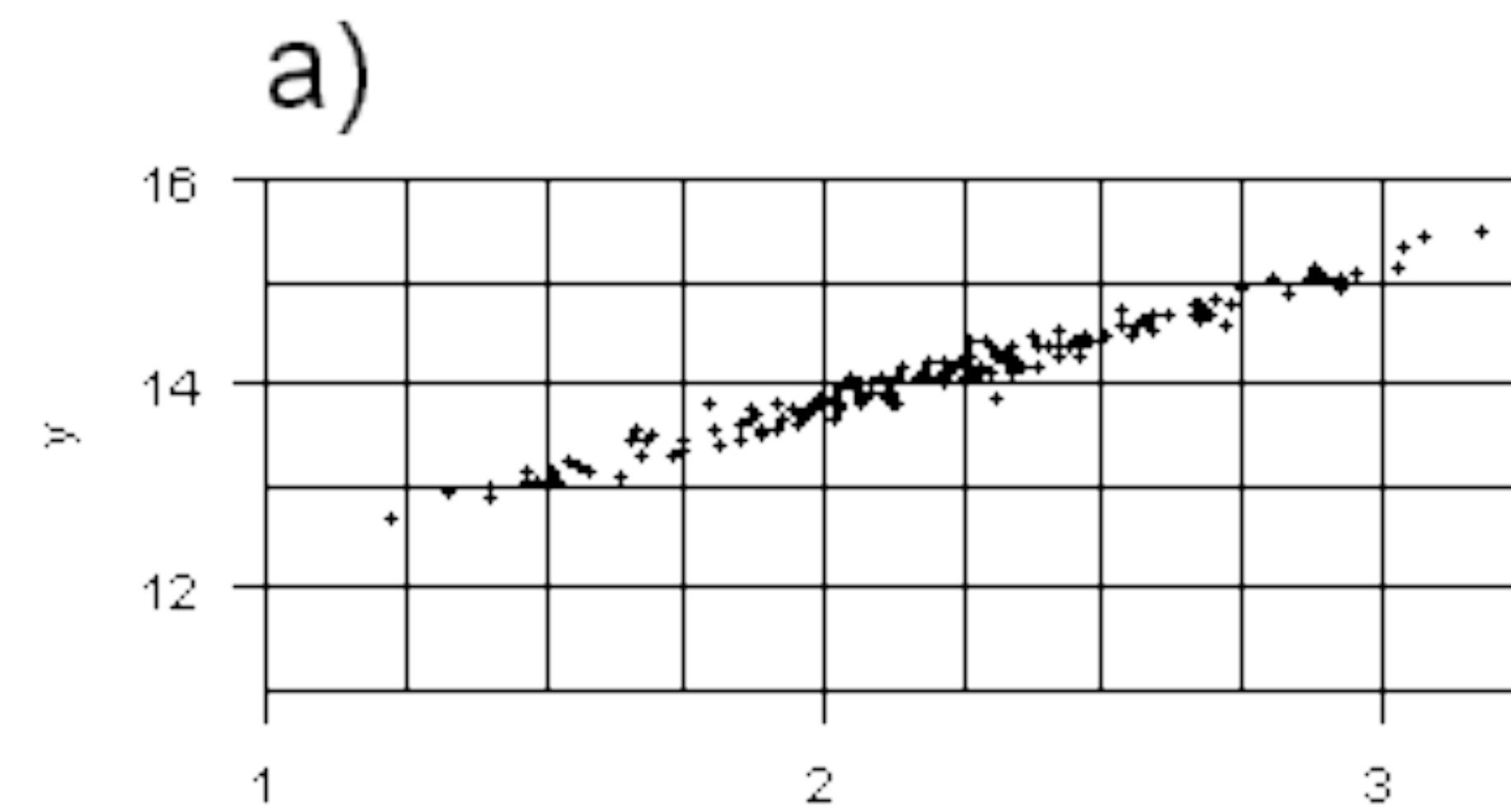




Calories per 100g

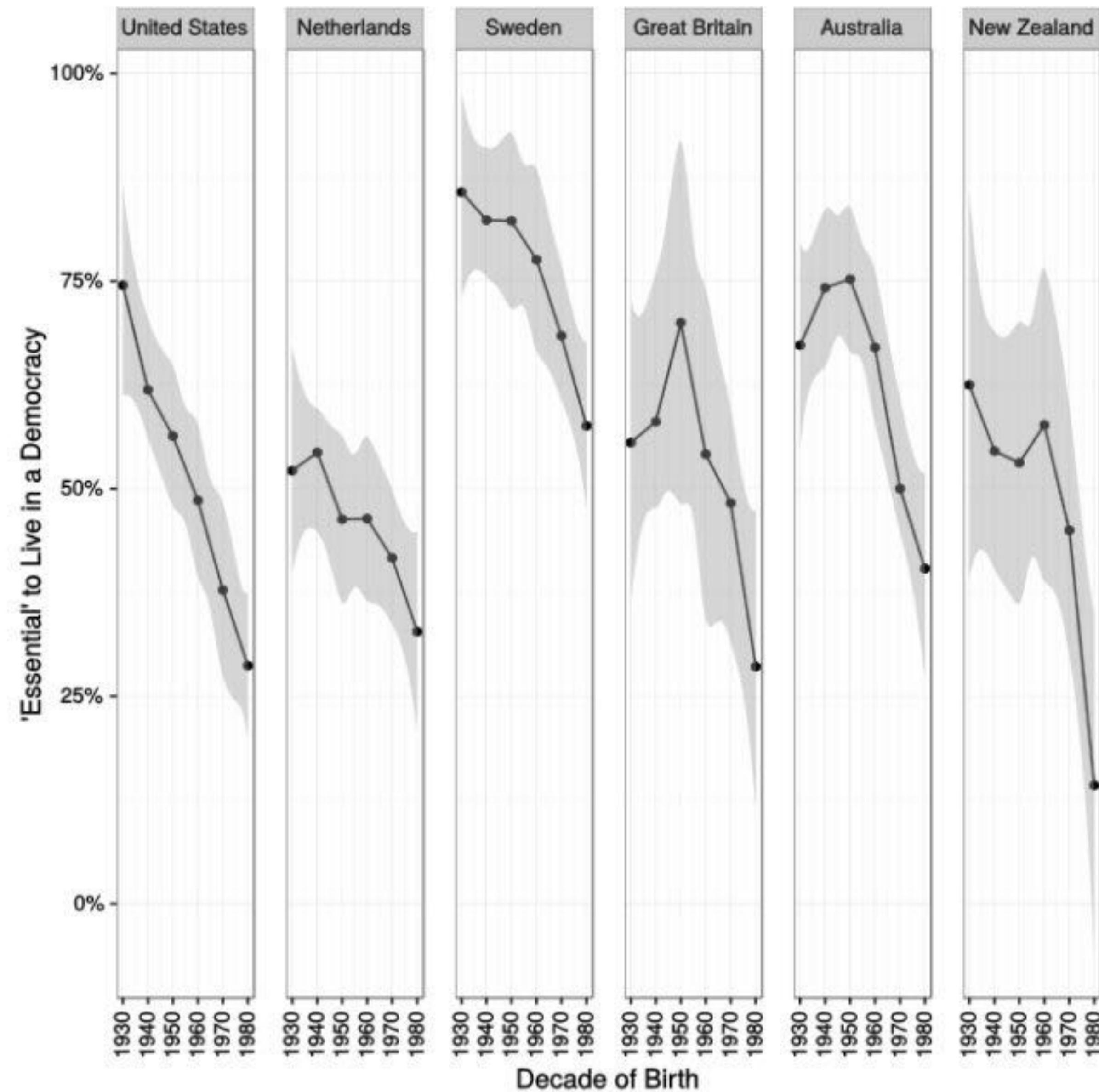




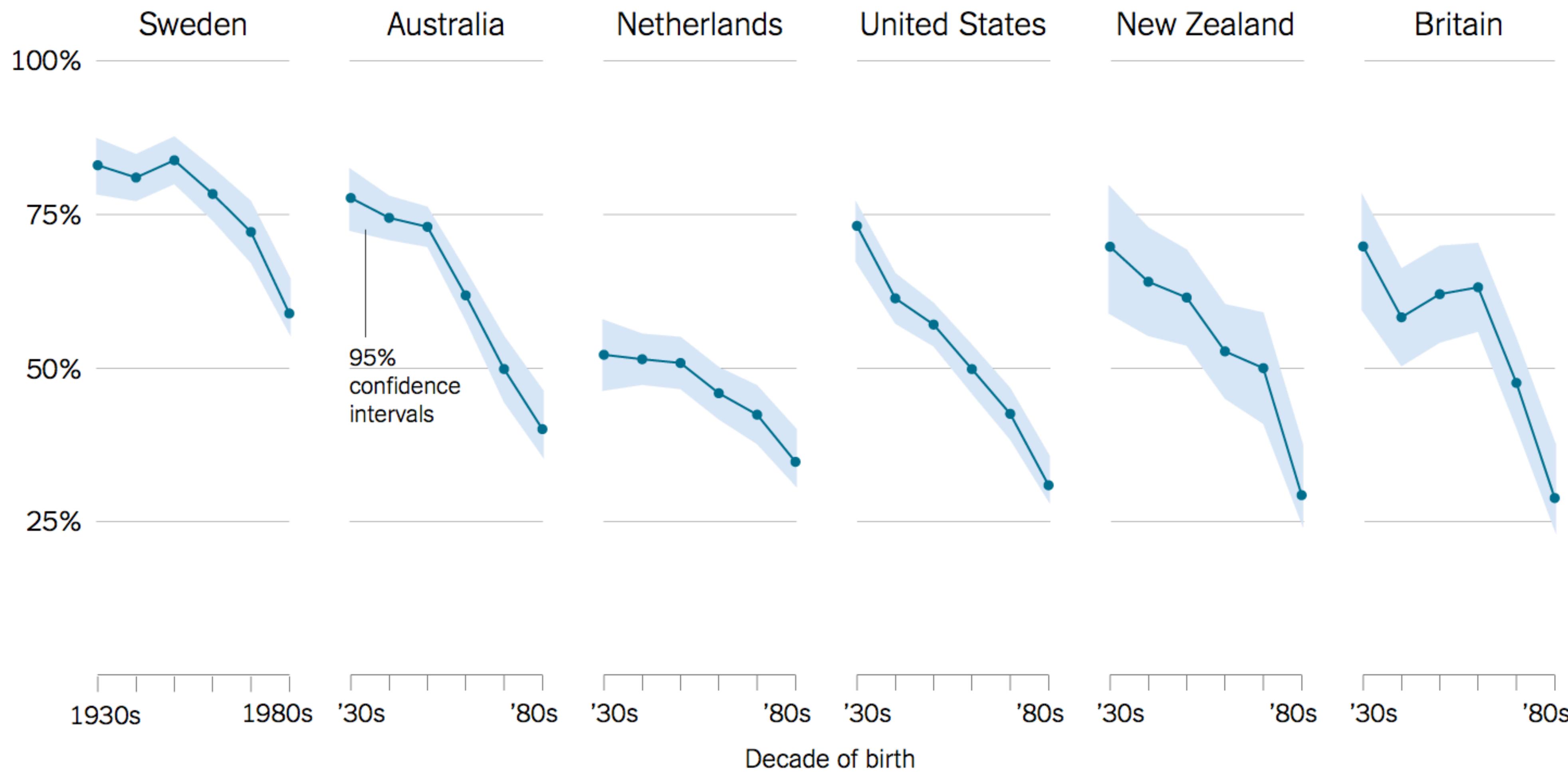


DATA

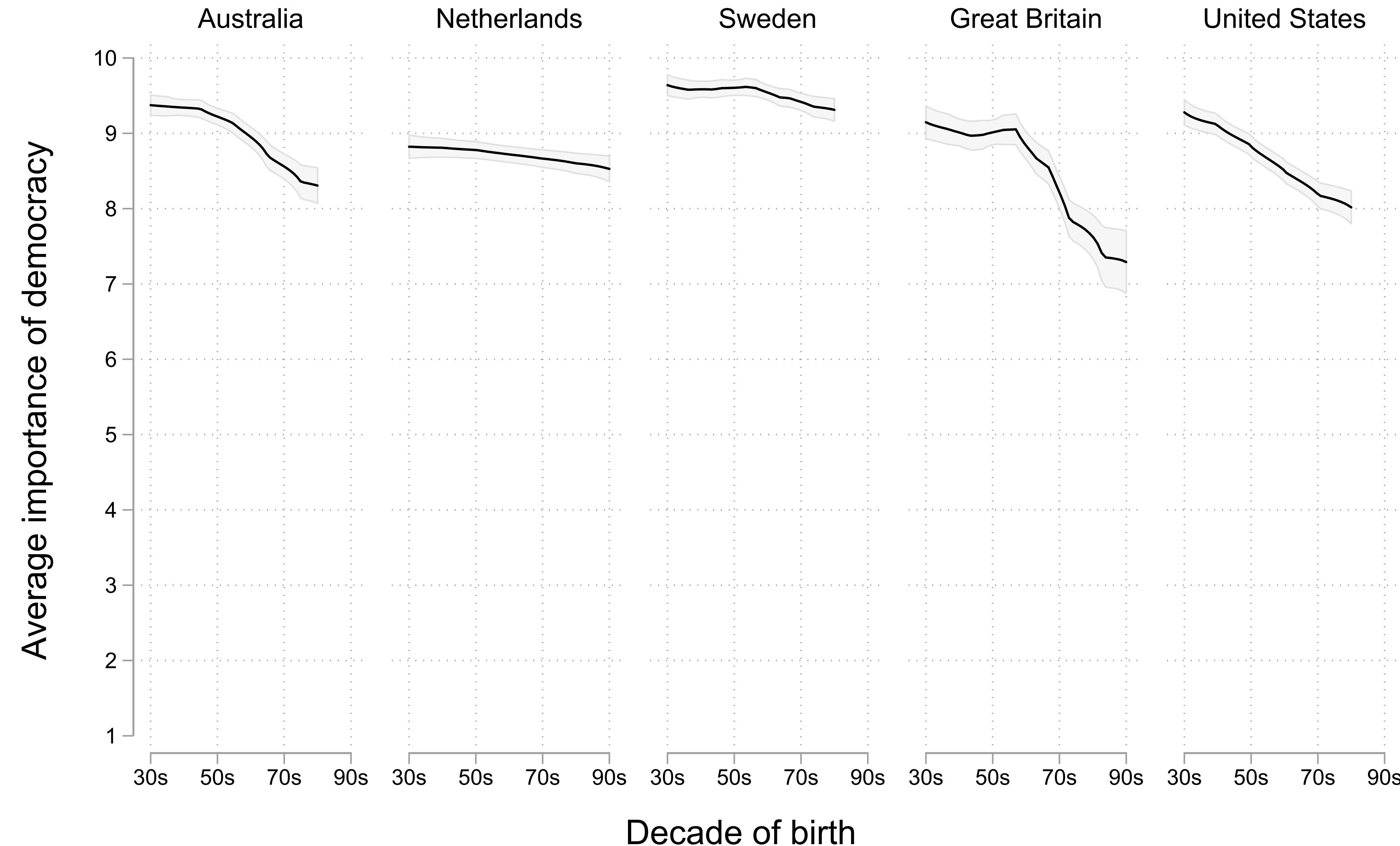
**JUNK-FREE
JUNK CHARTS**



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

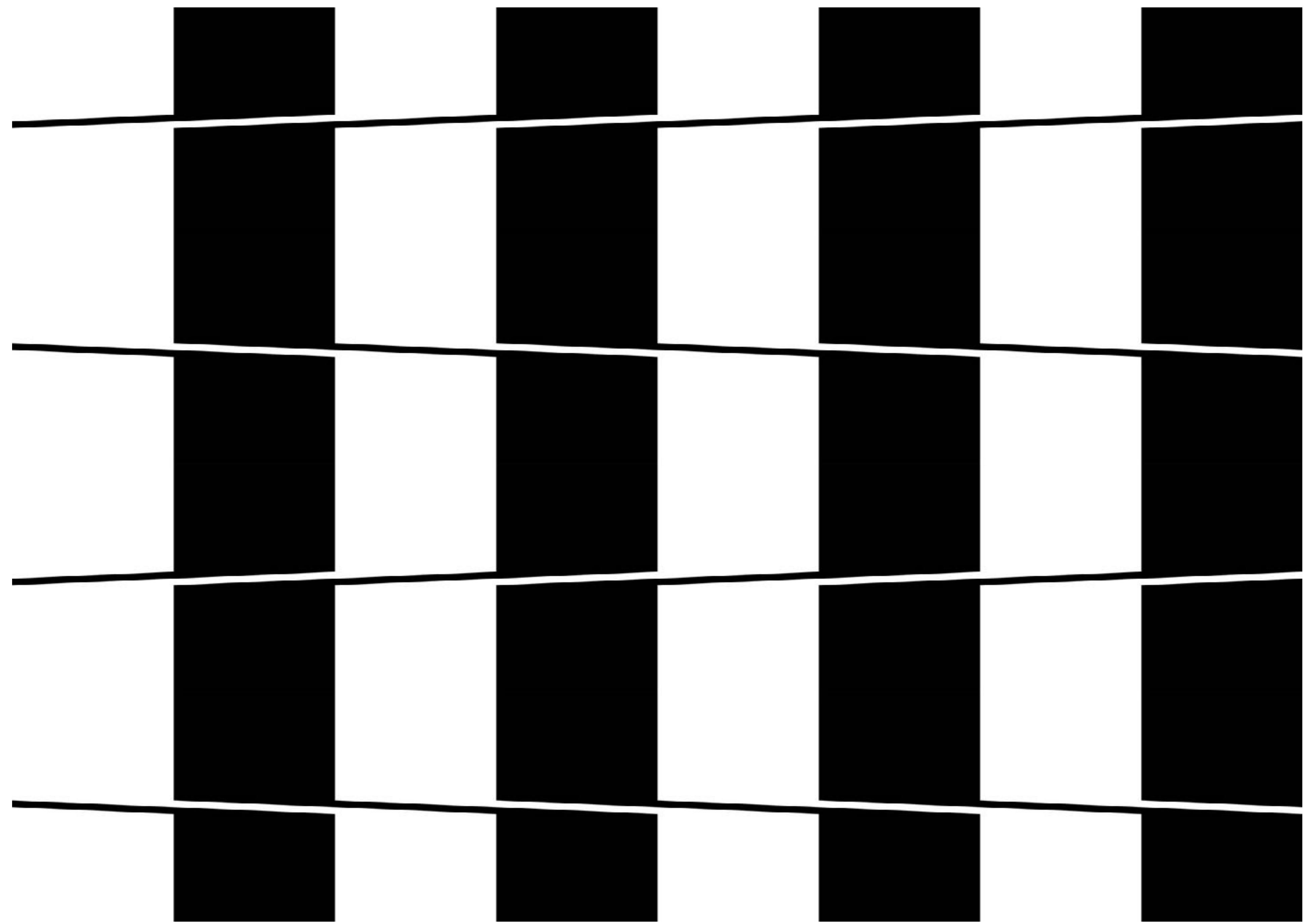
PERCEPTION

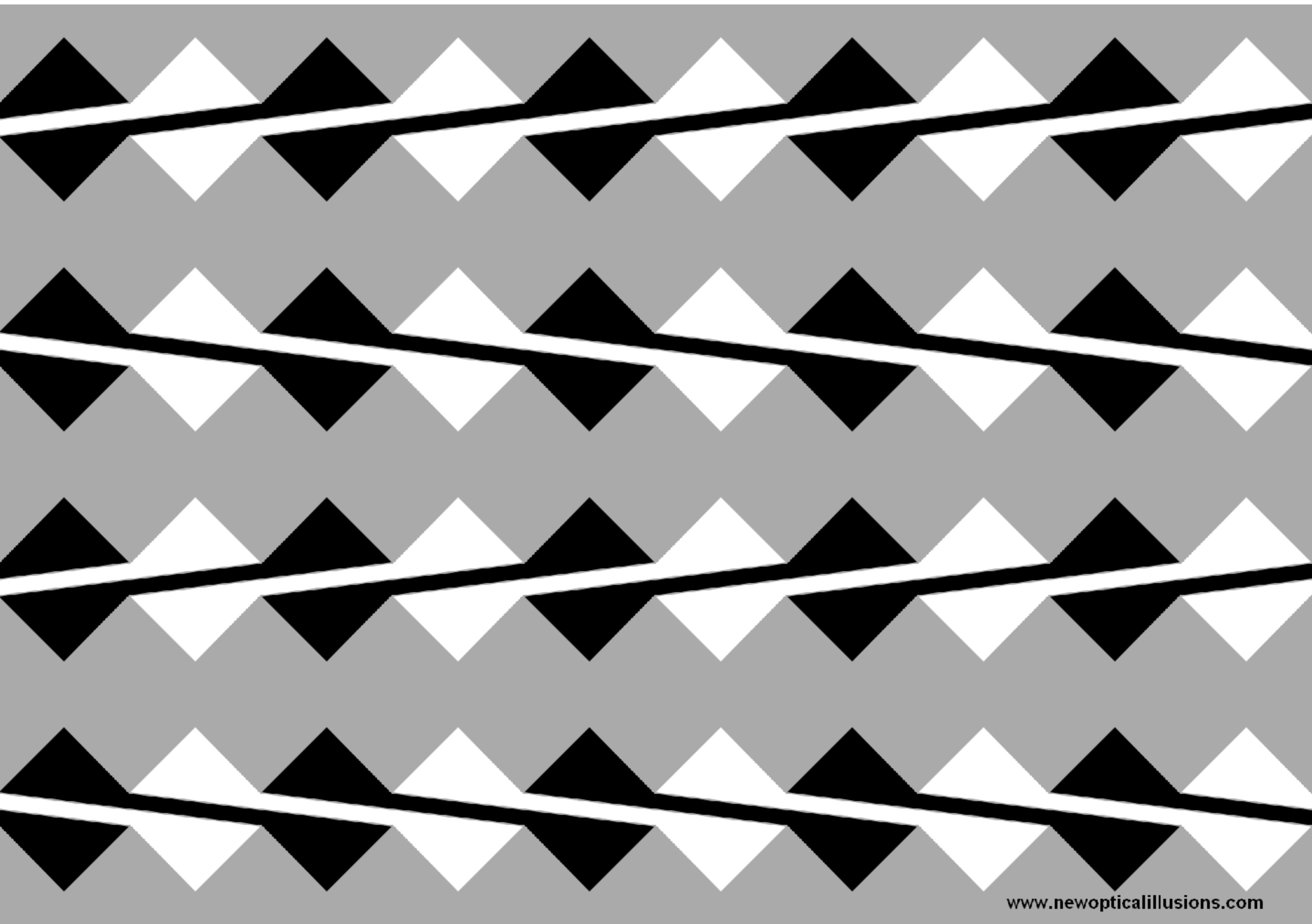
SIGHT AND INSIGHT

**Edges &
Contrasts**

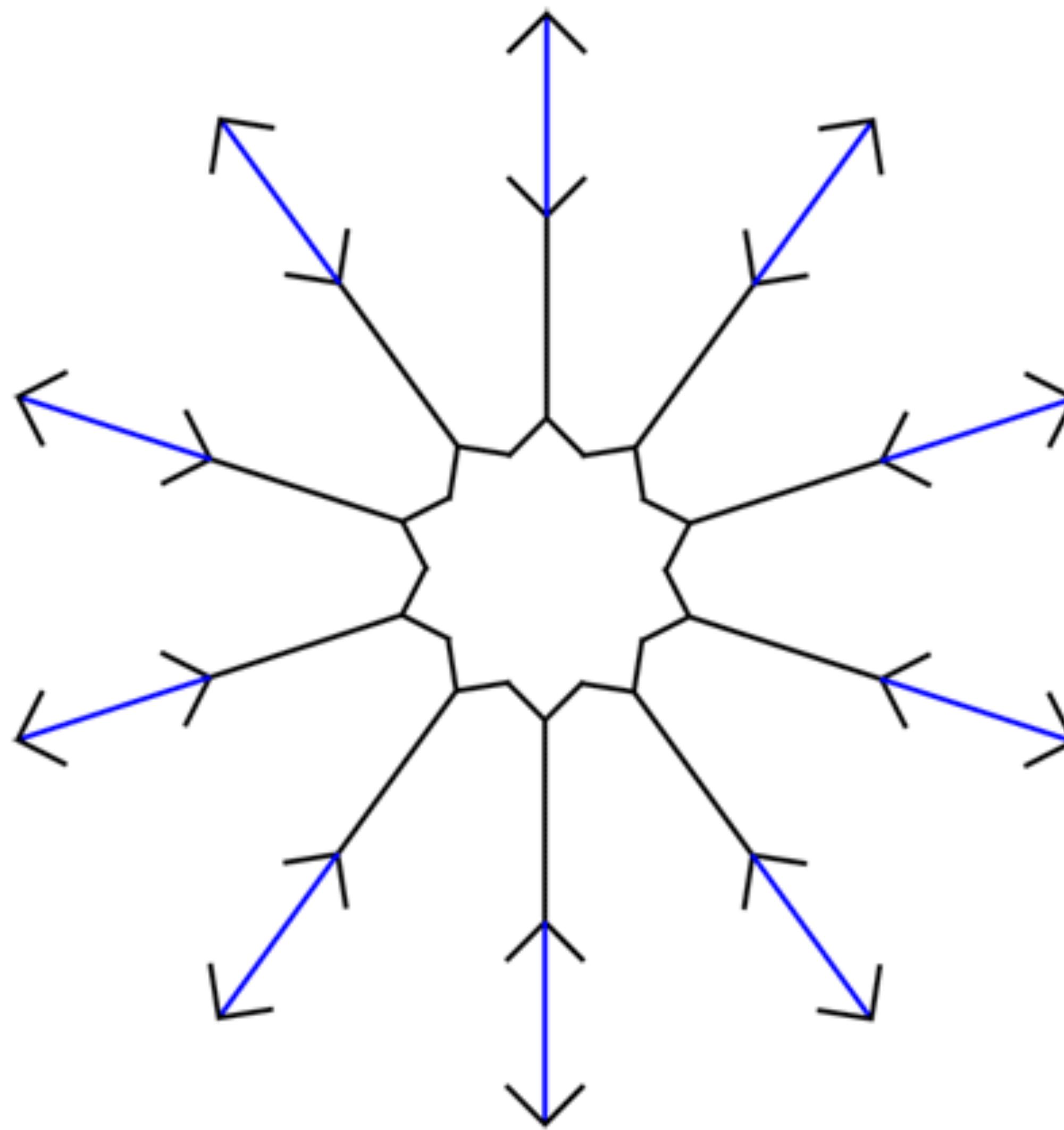


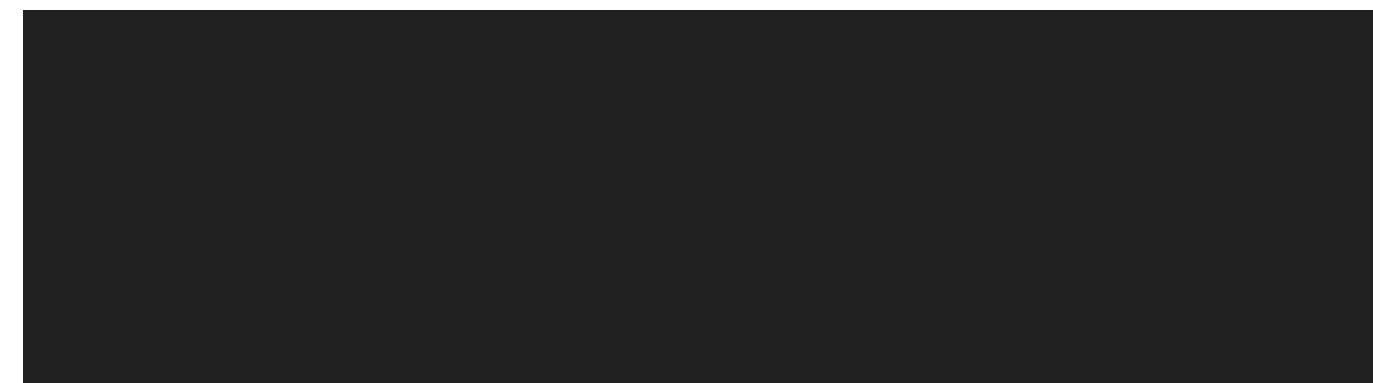
fin





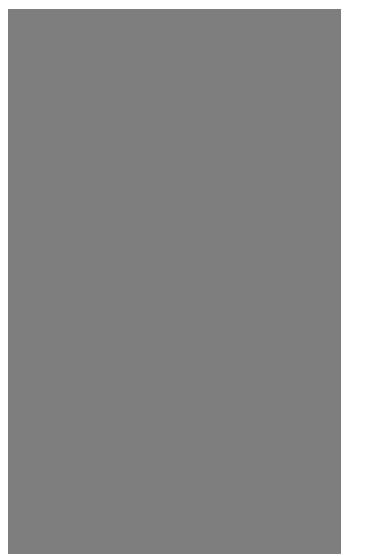
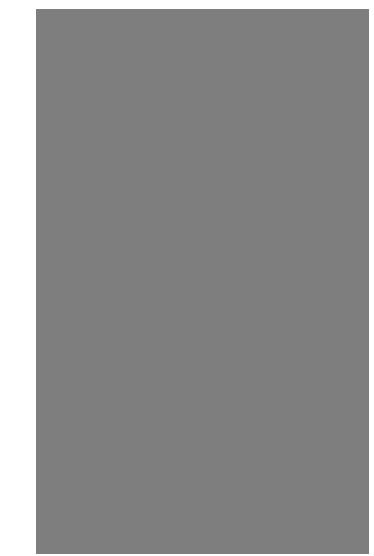
Sarcone's Dynamic Müller-Lyer Illusion

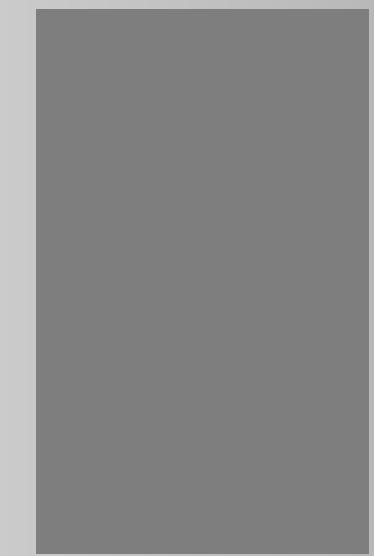
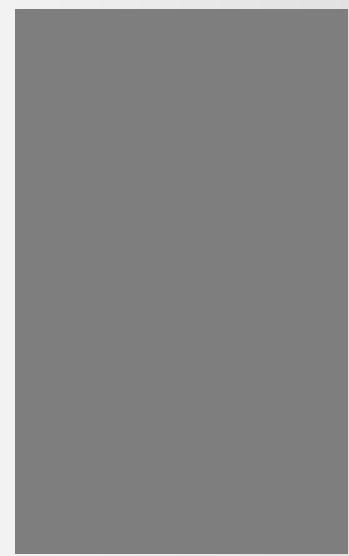


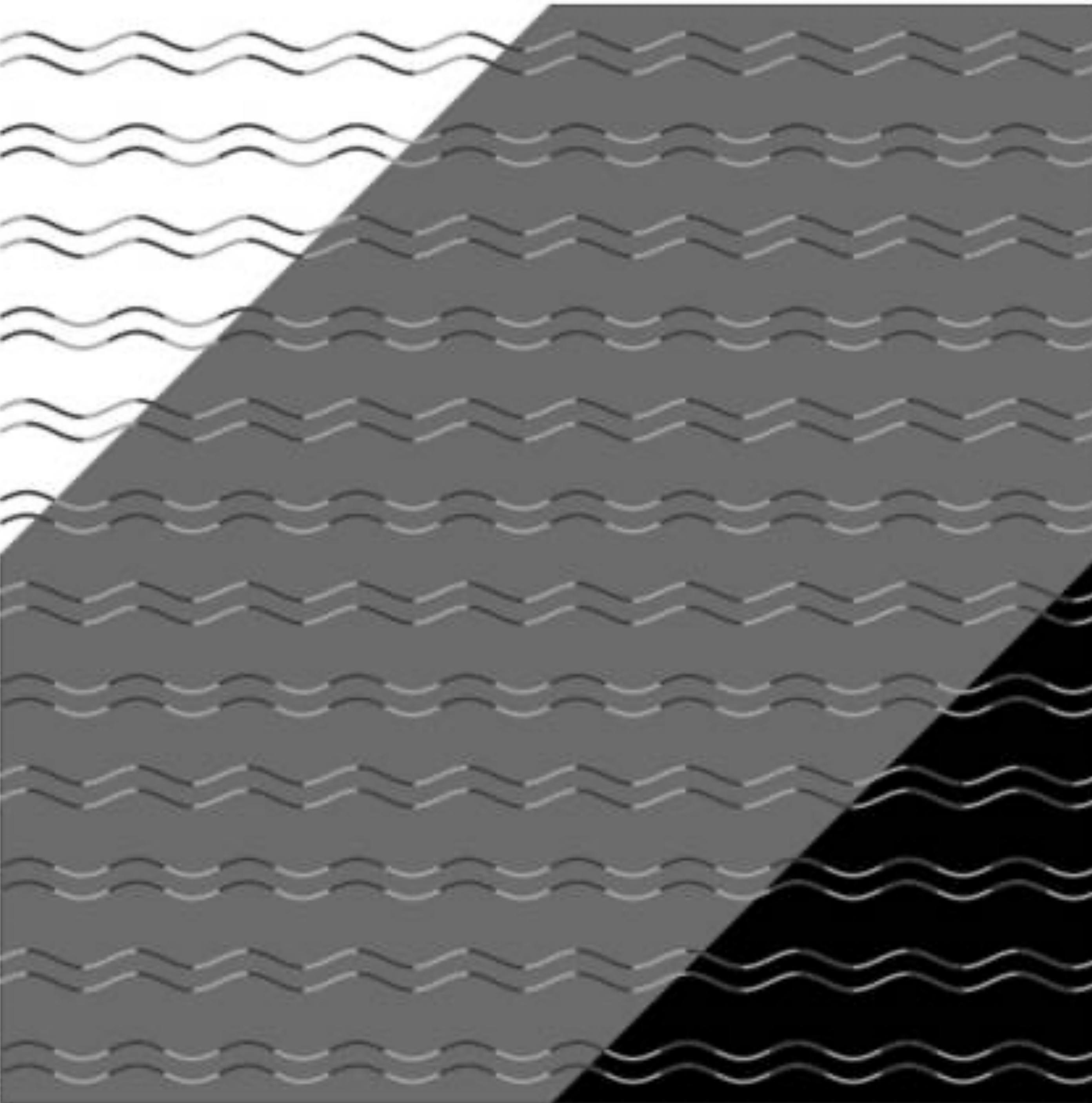


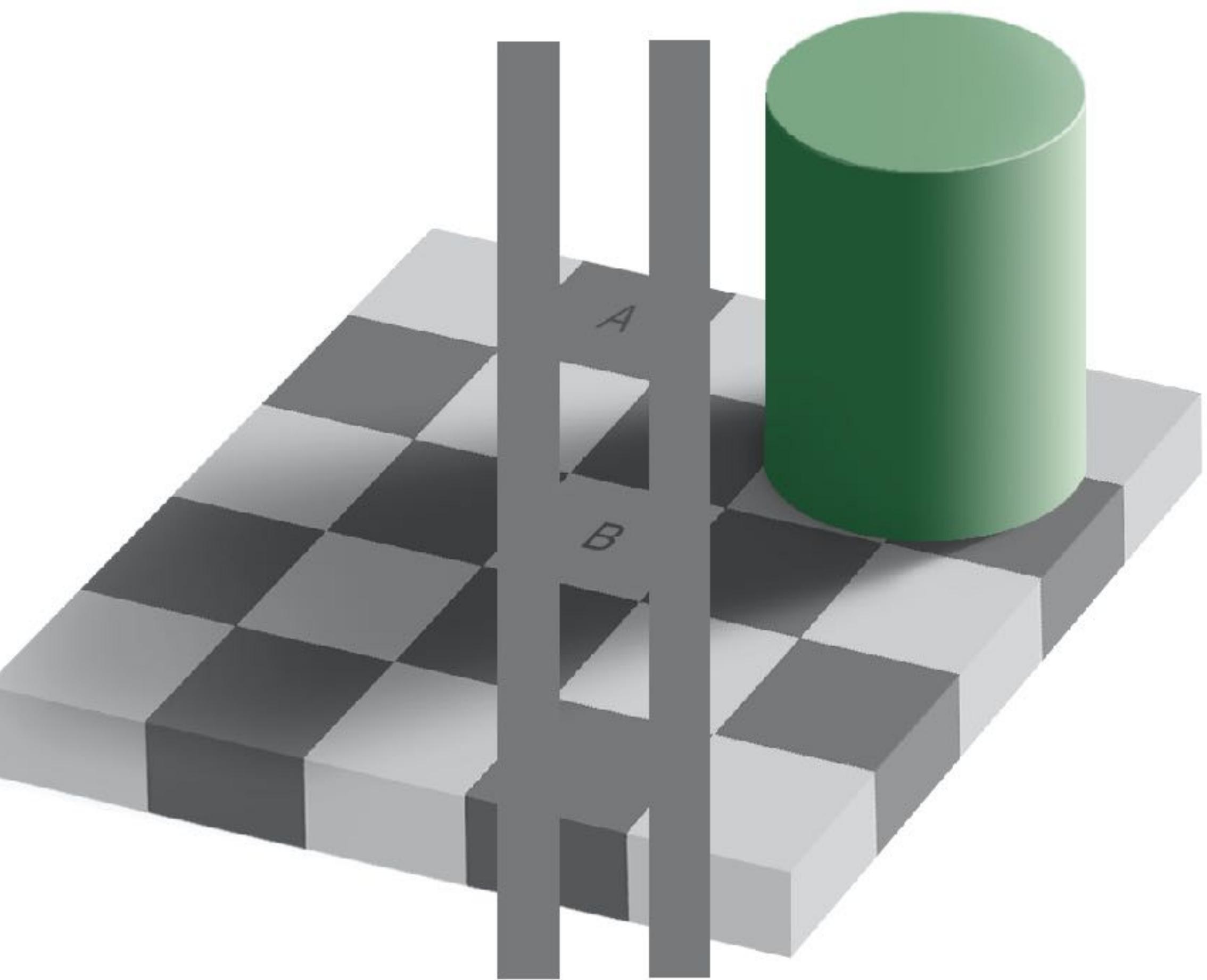
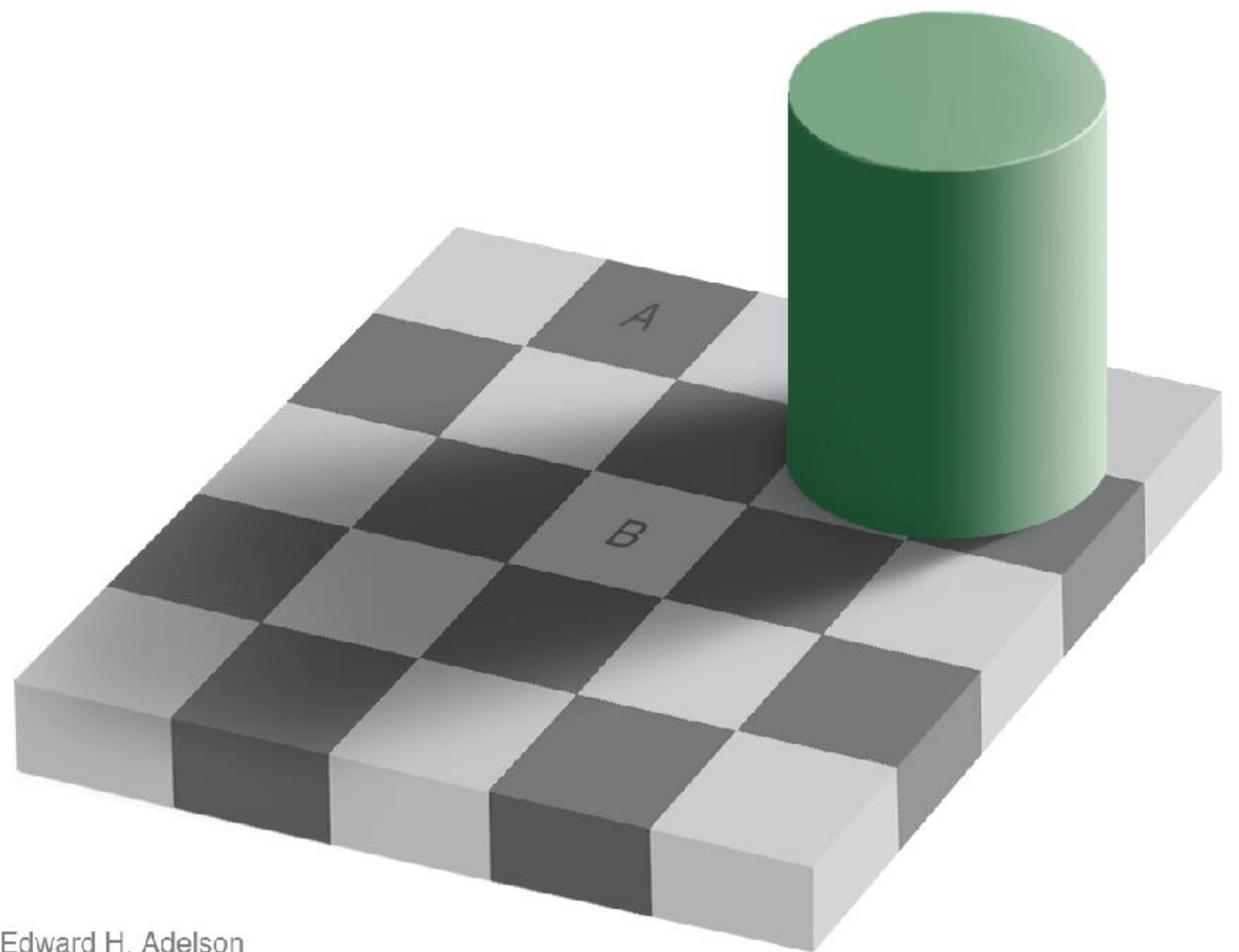




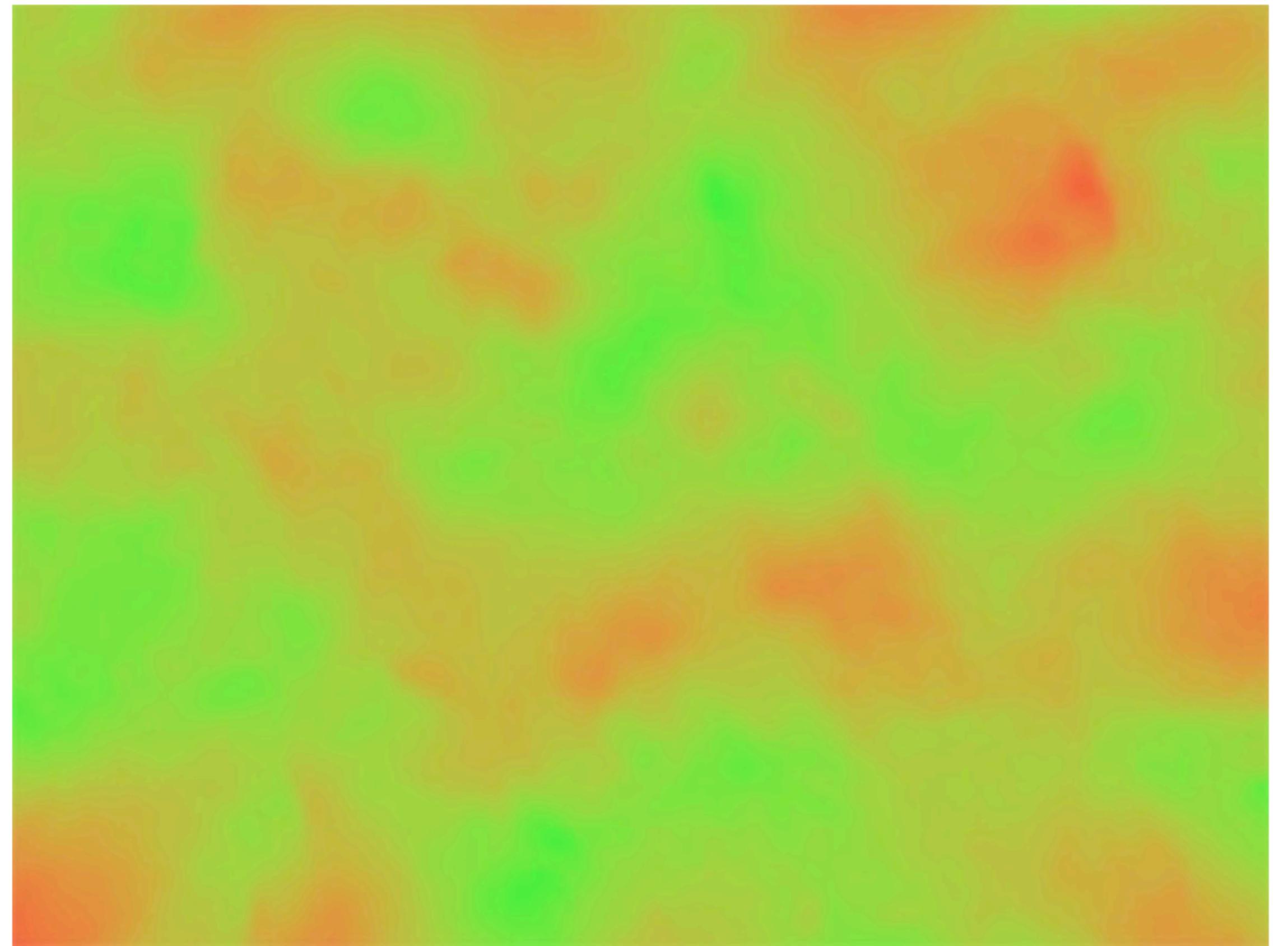


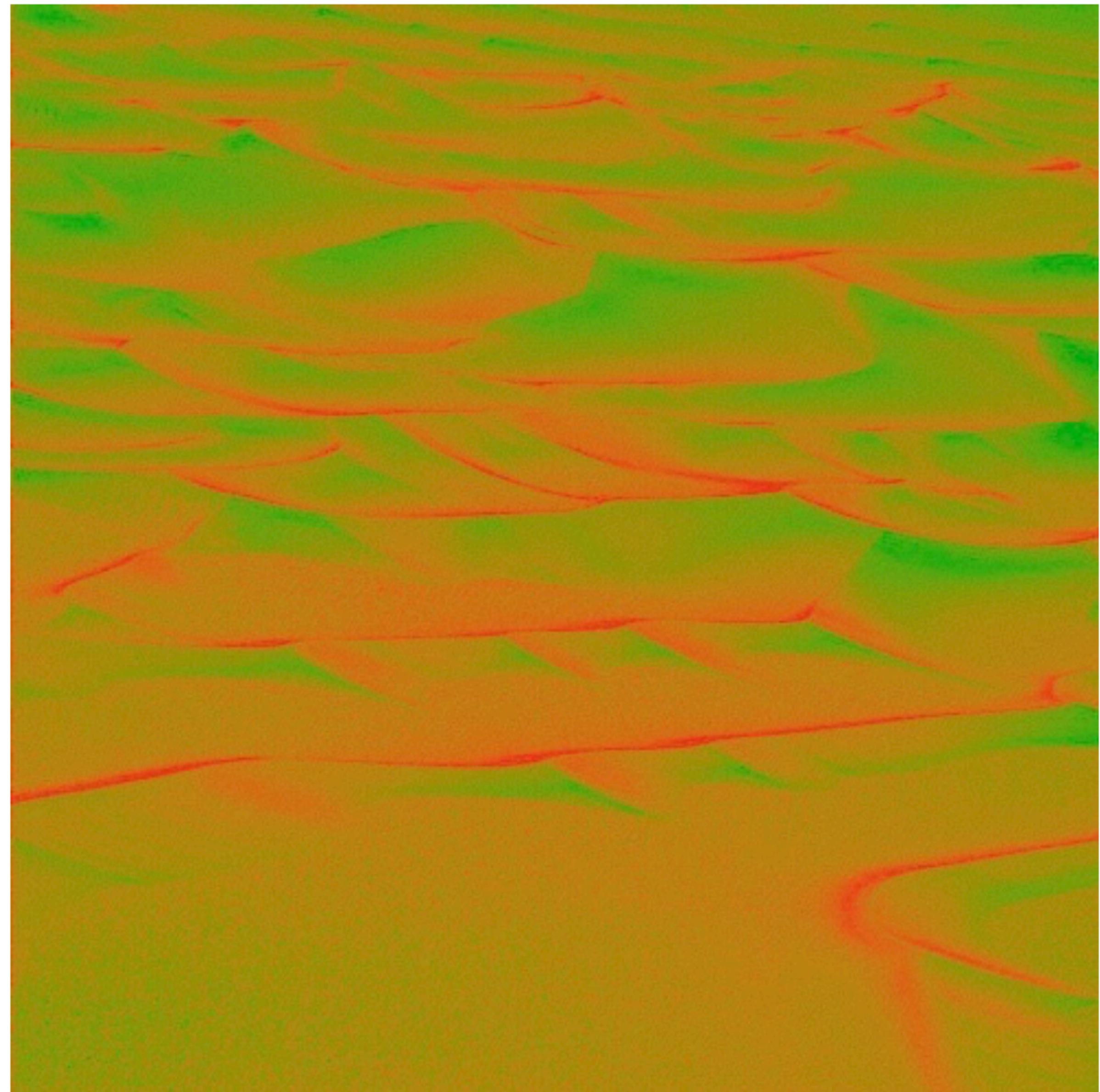


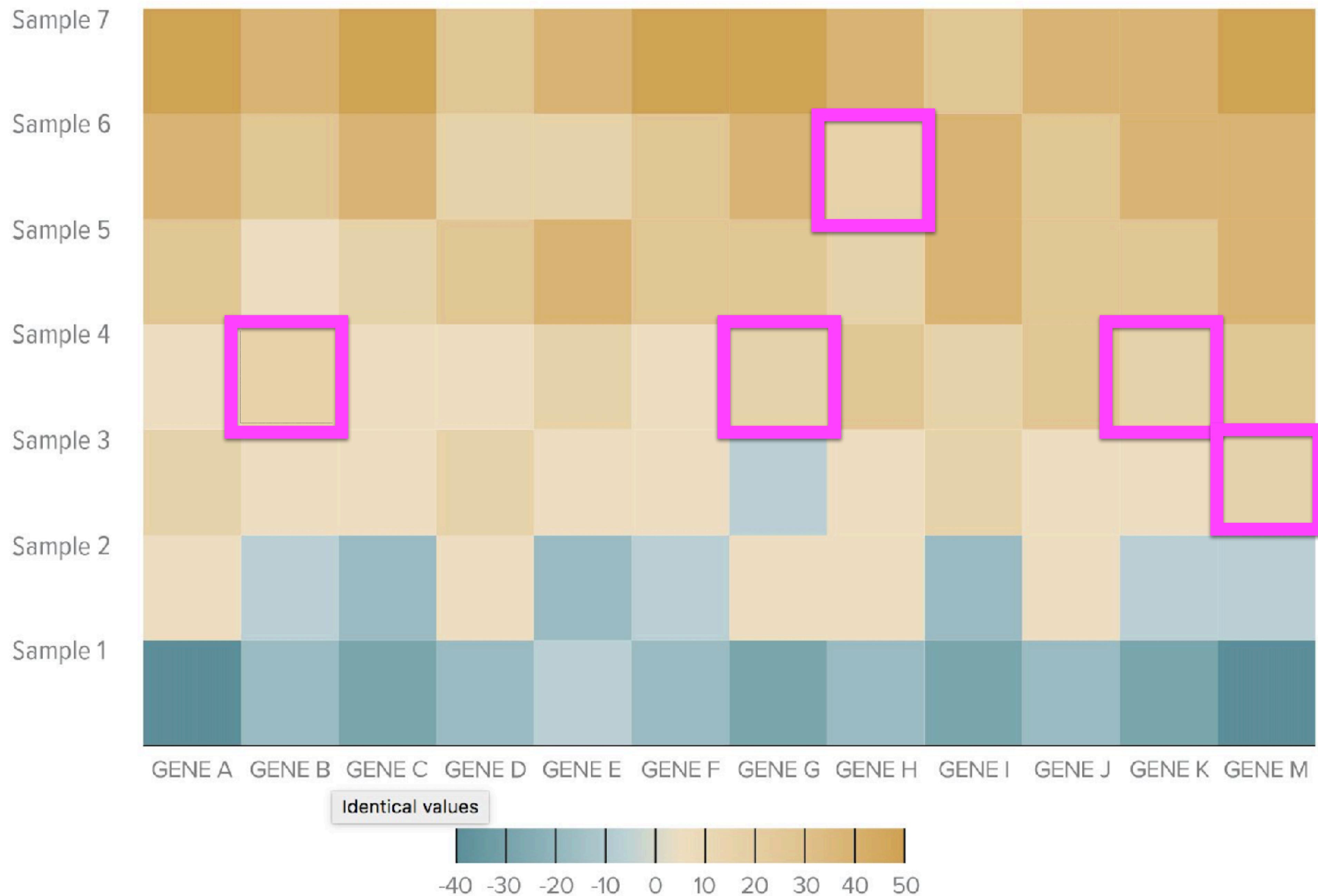


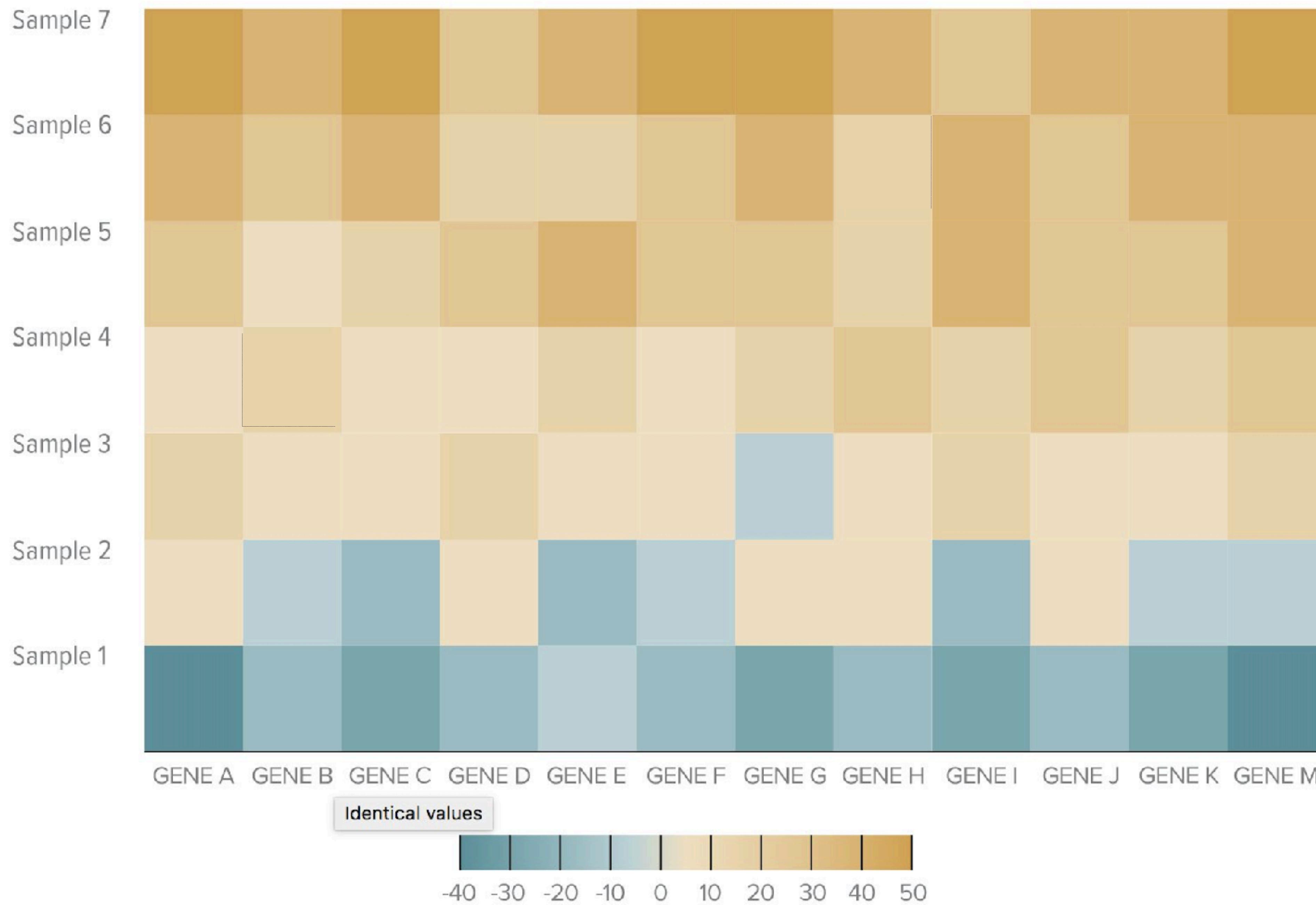


Edward H. Adelson





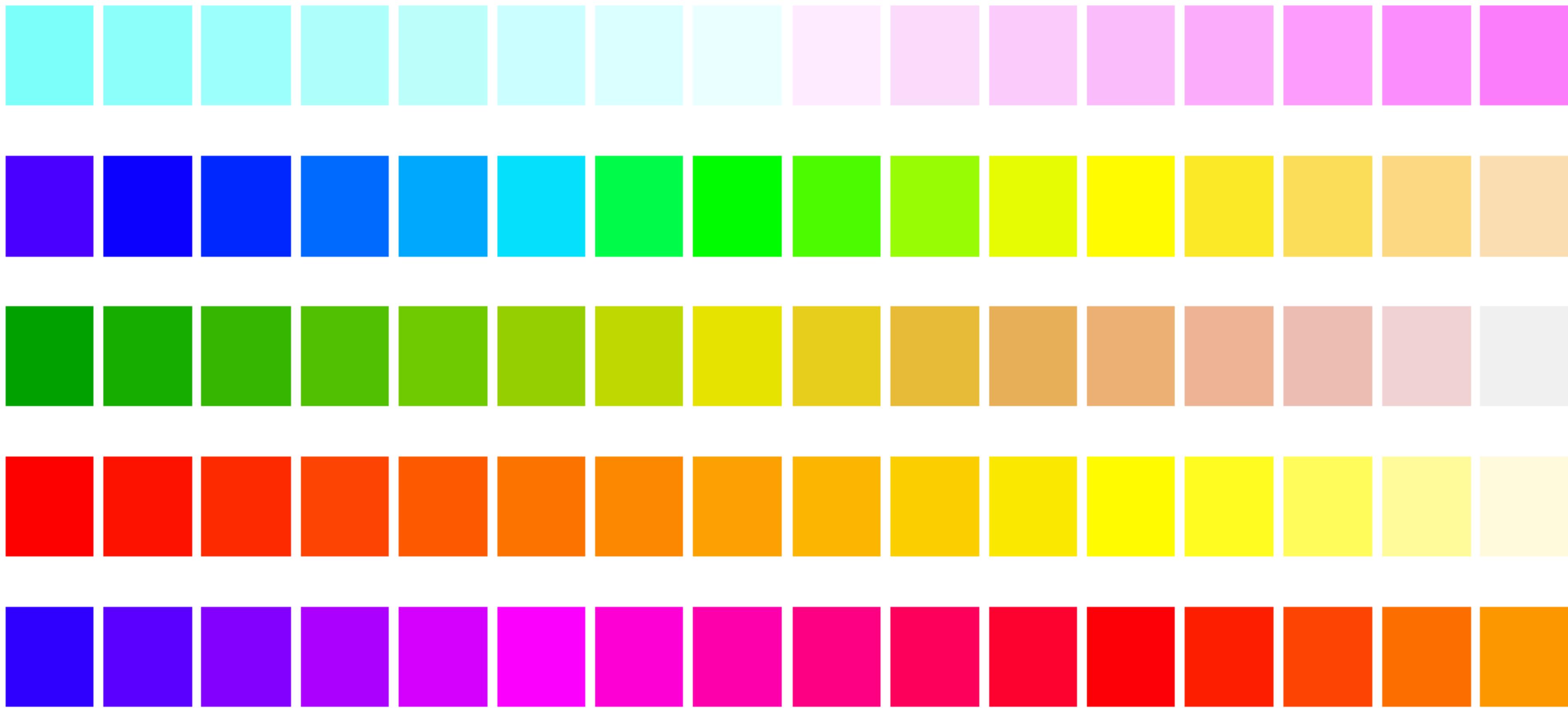




Color

+





viridis



magma



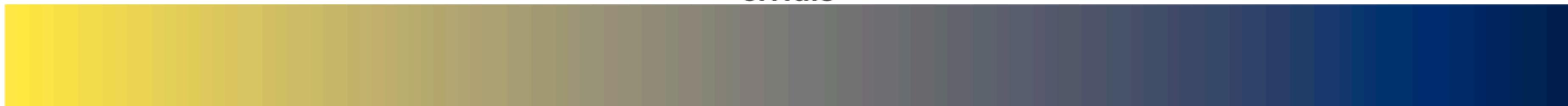
plasma



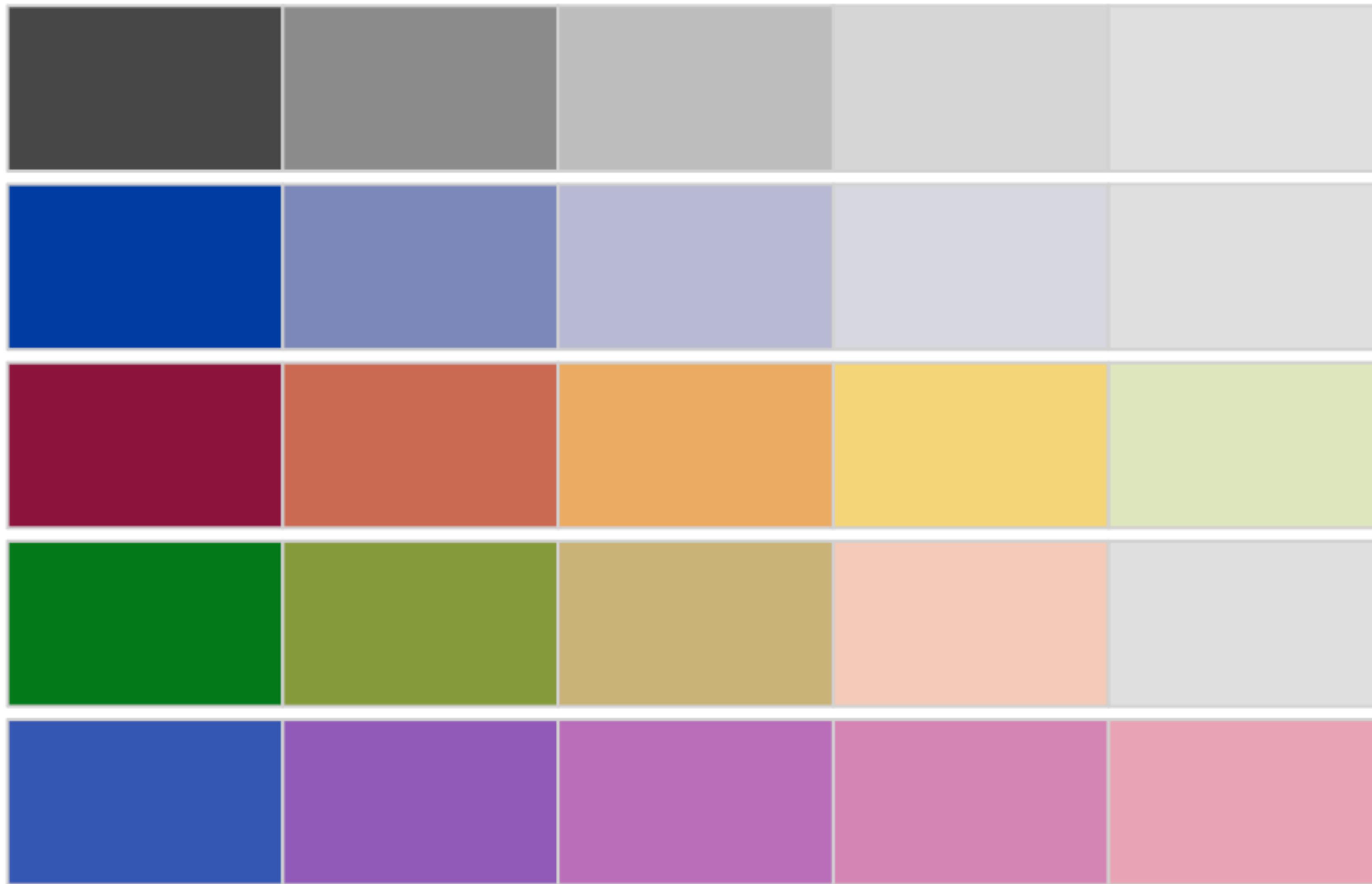
inferno

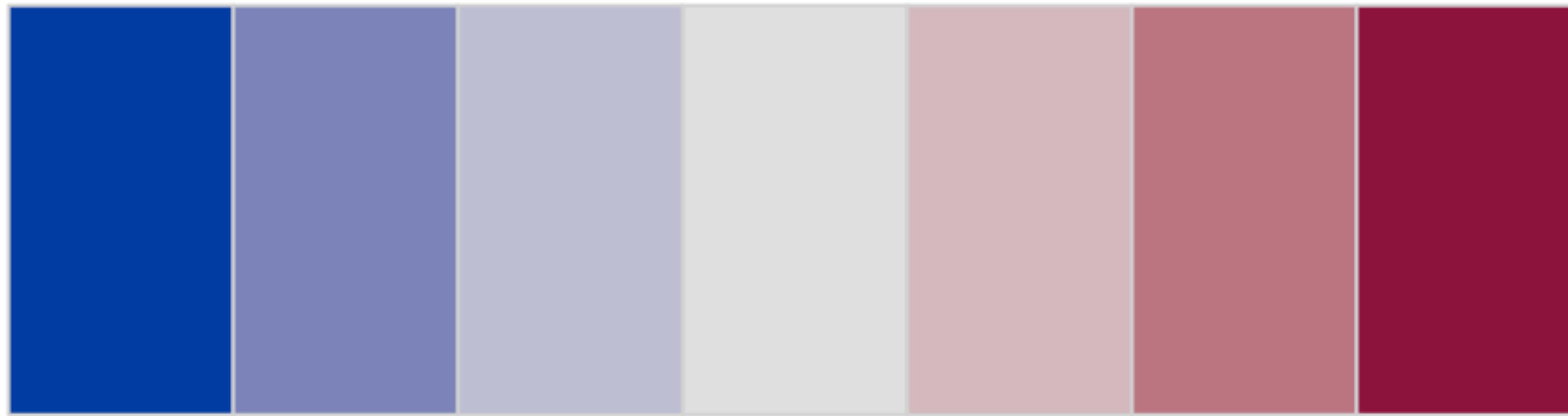


cividis





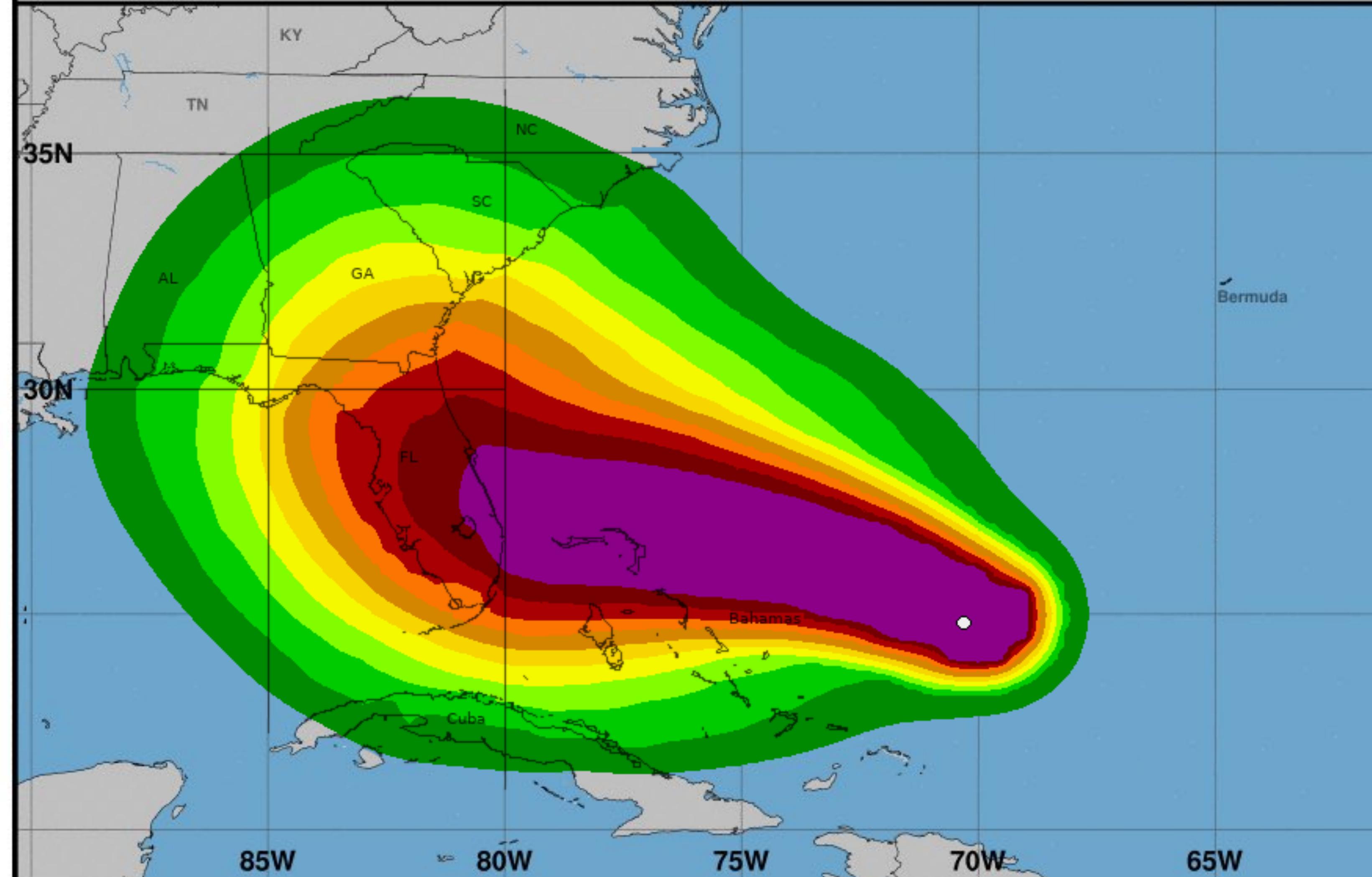






Tropical-Storm-Force Wind Speed Probabilities

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
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)

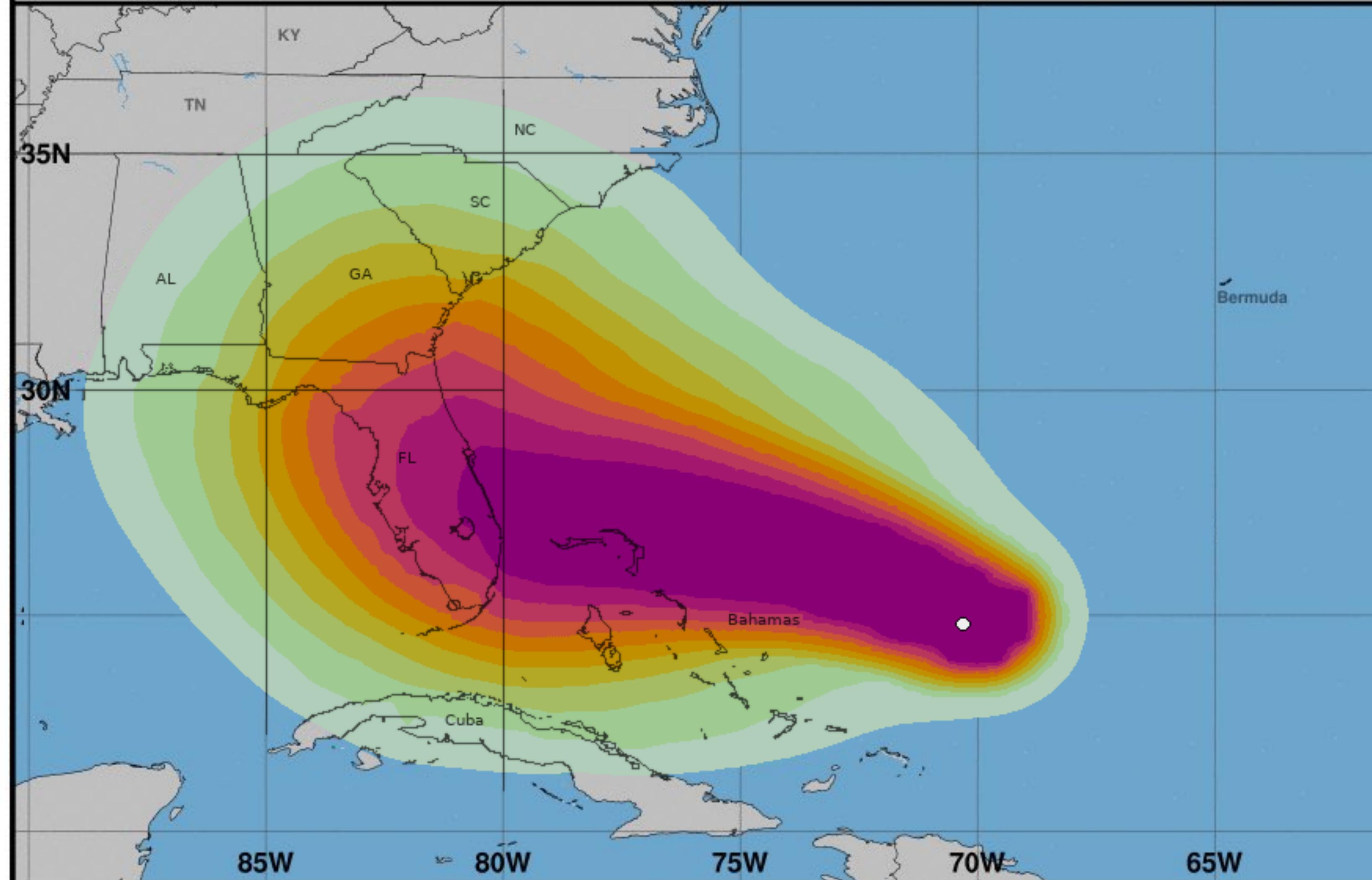


See: Achim Zeileis
<https://bit.ly/35zBIYL>



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



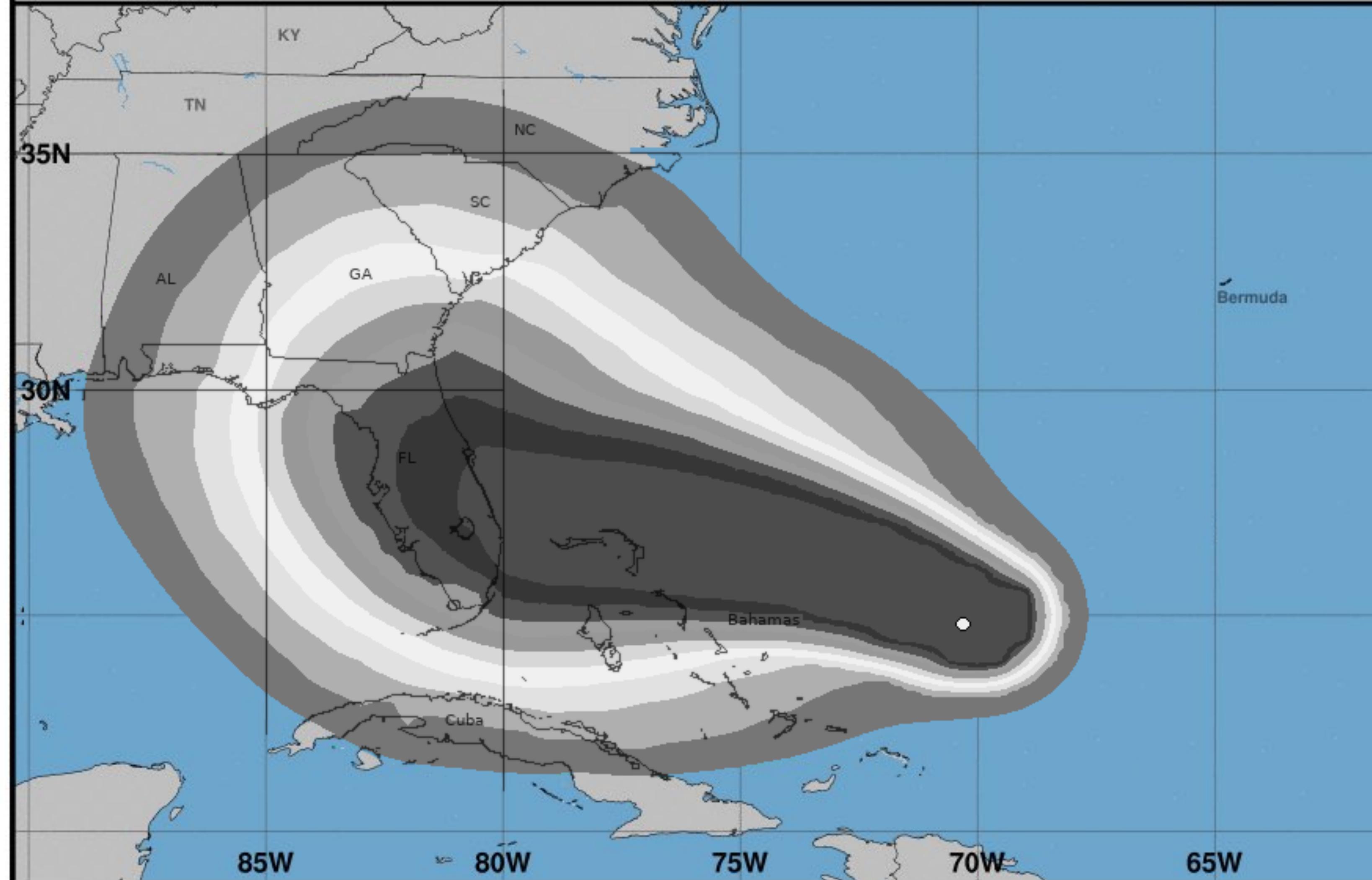
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)



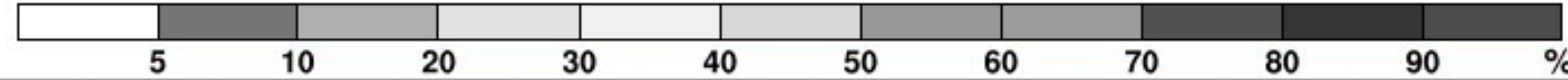


Tropical-Storm-Force Wind Speed Probabilities

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



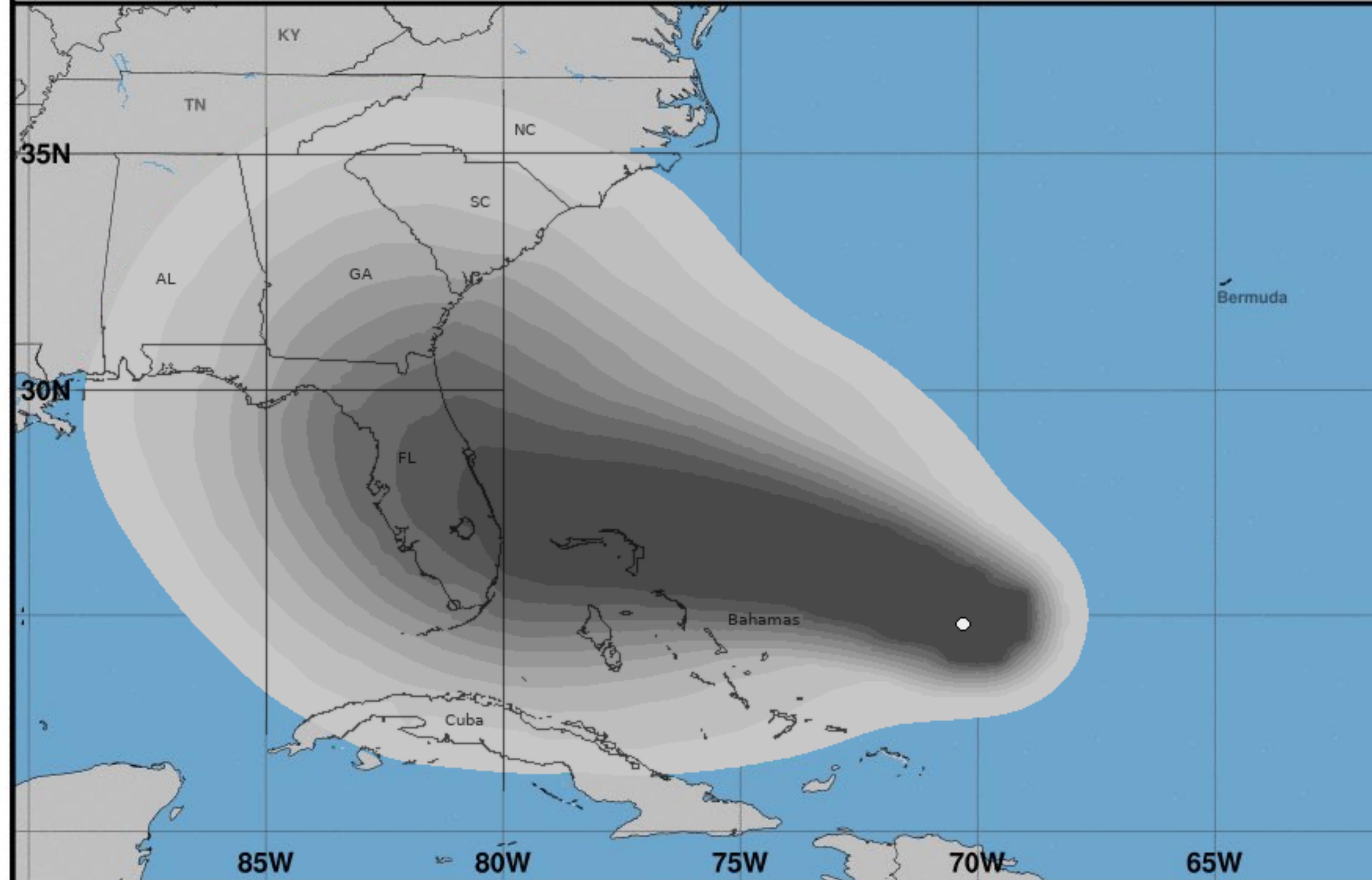
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





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



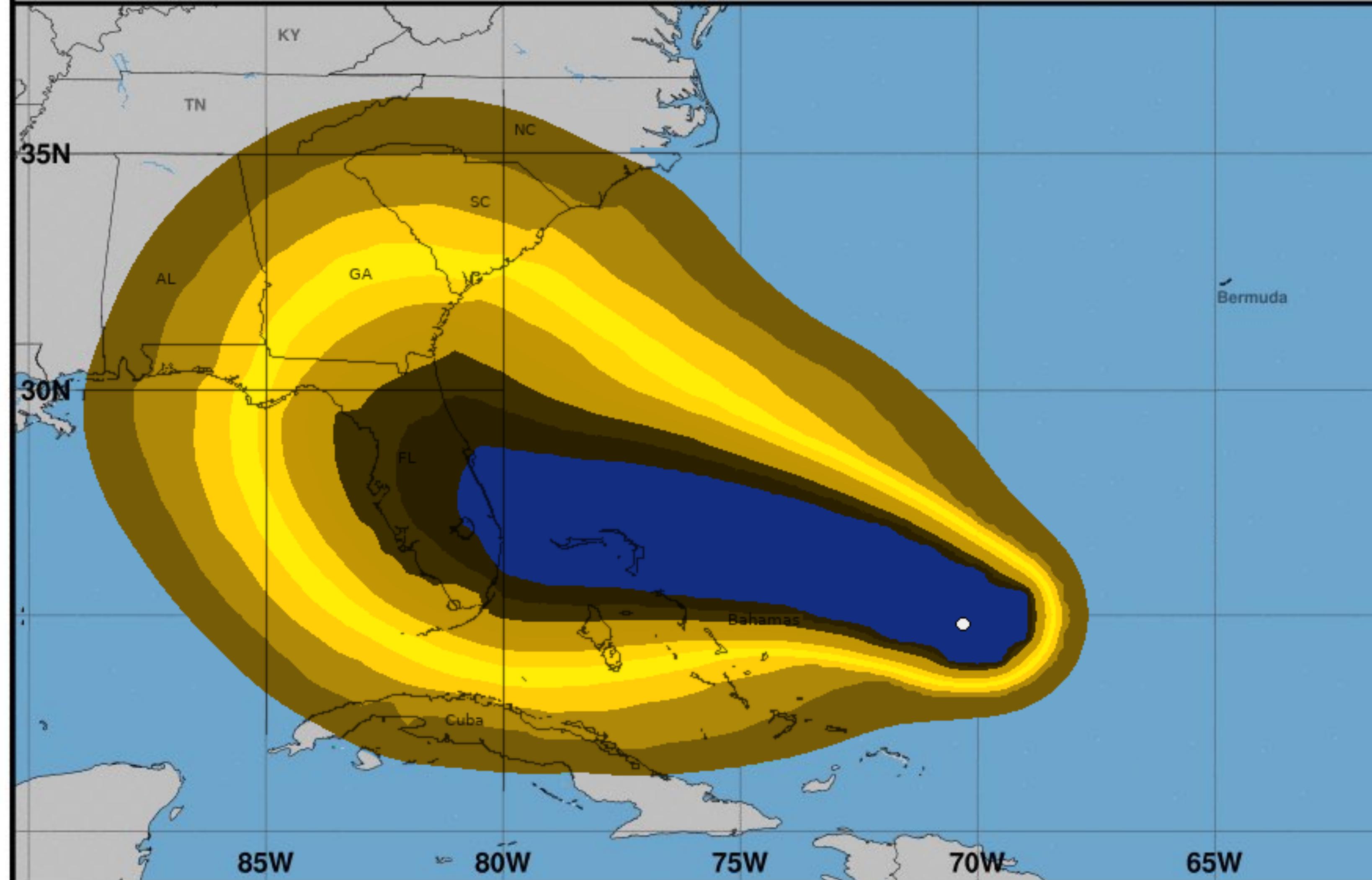
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





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



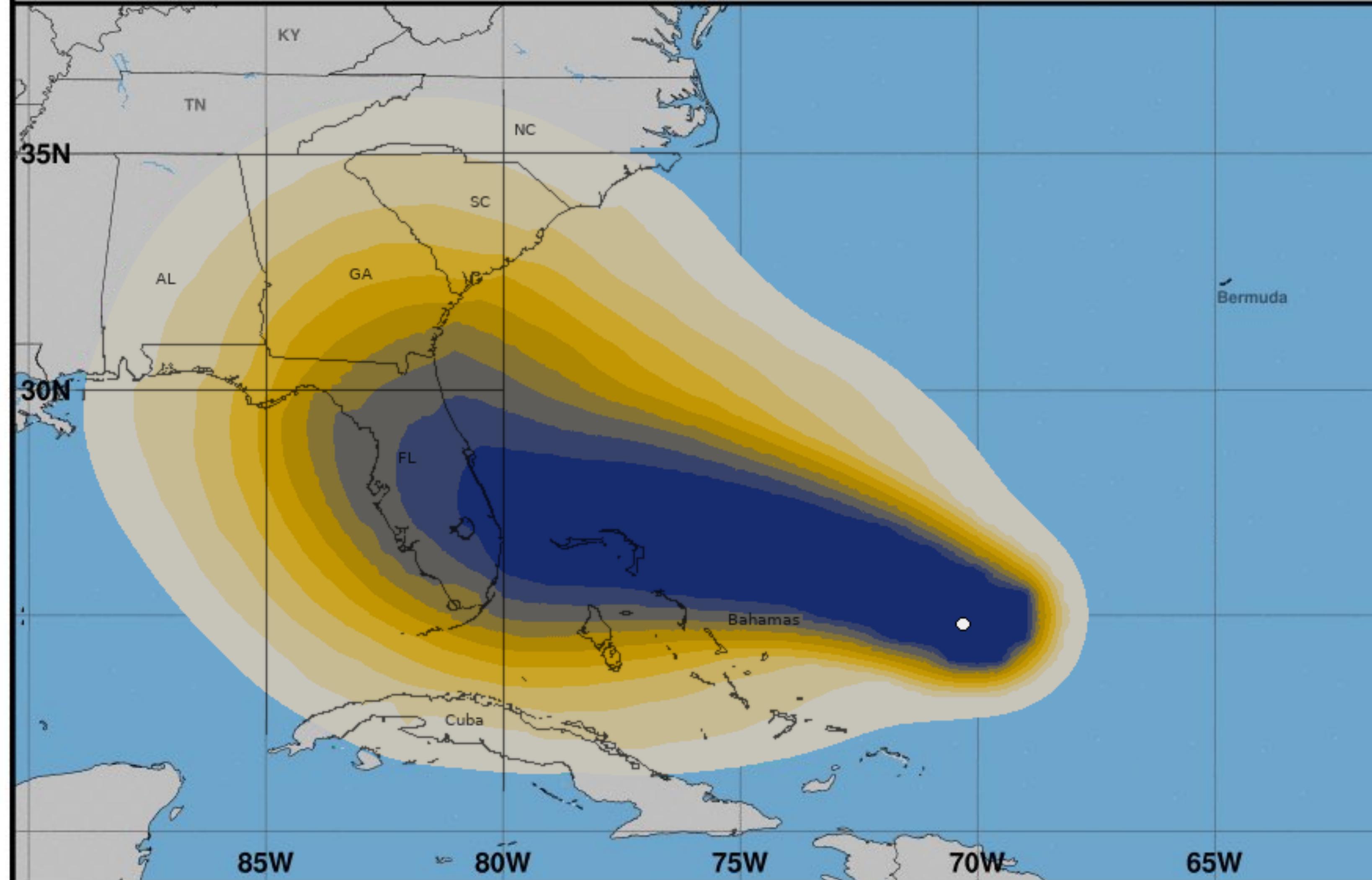
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





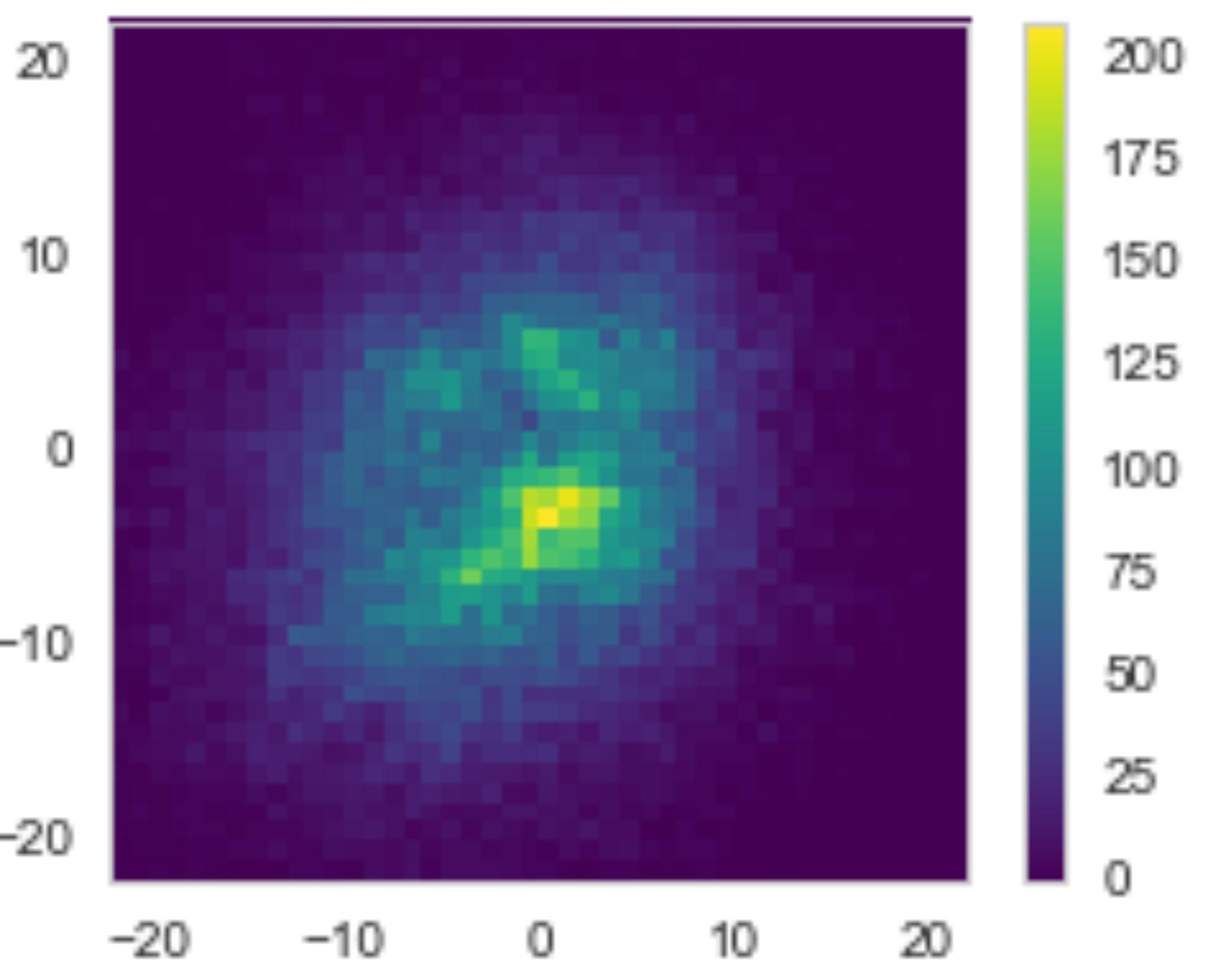
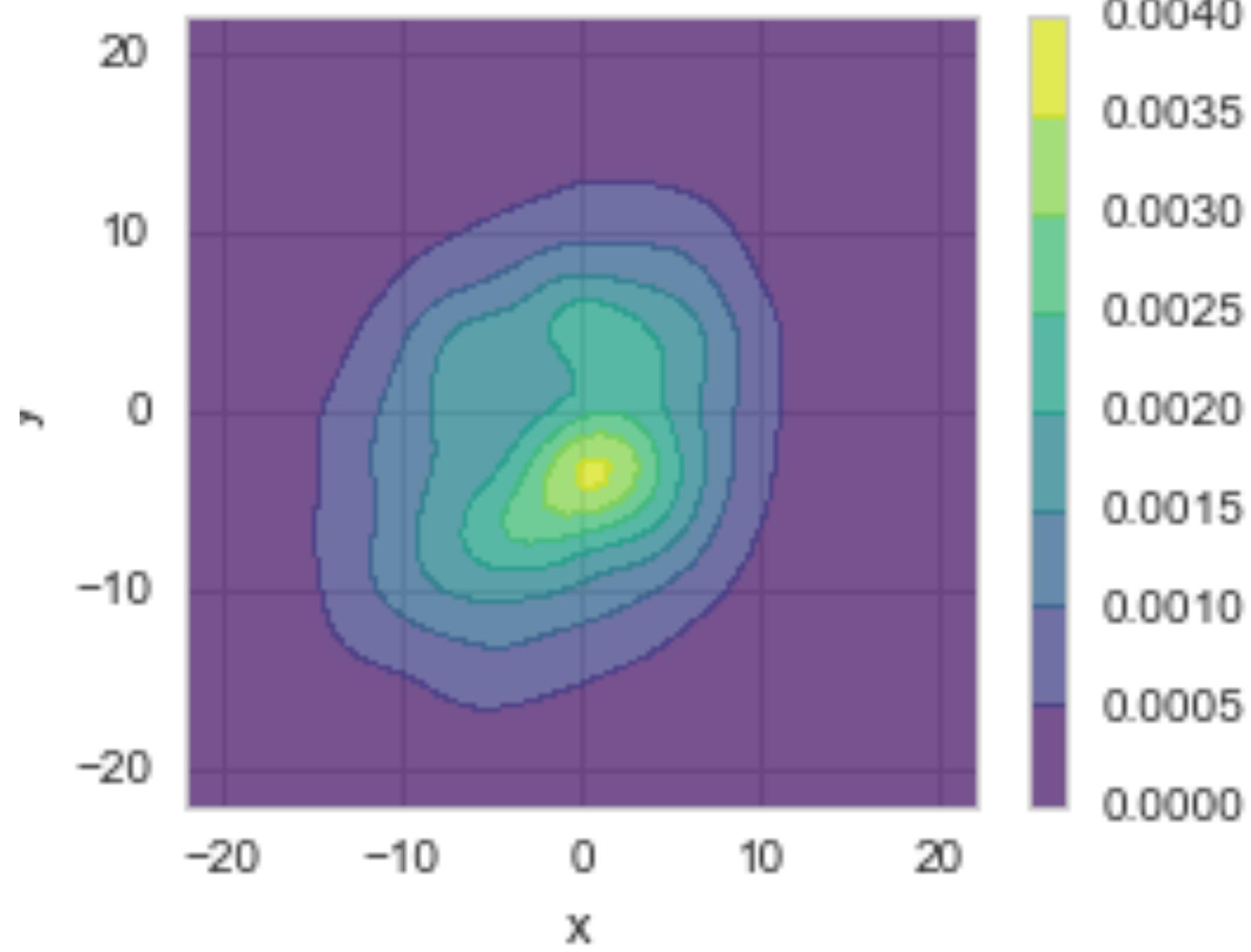
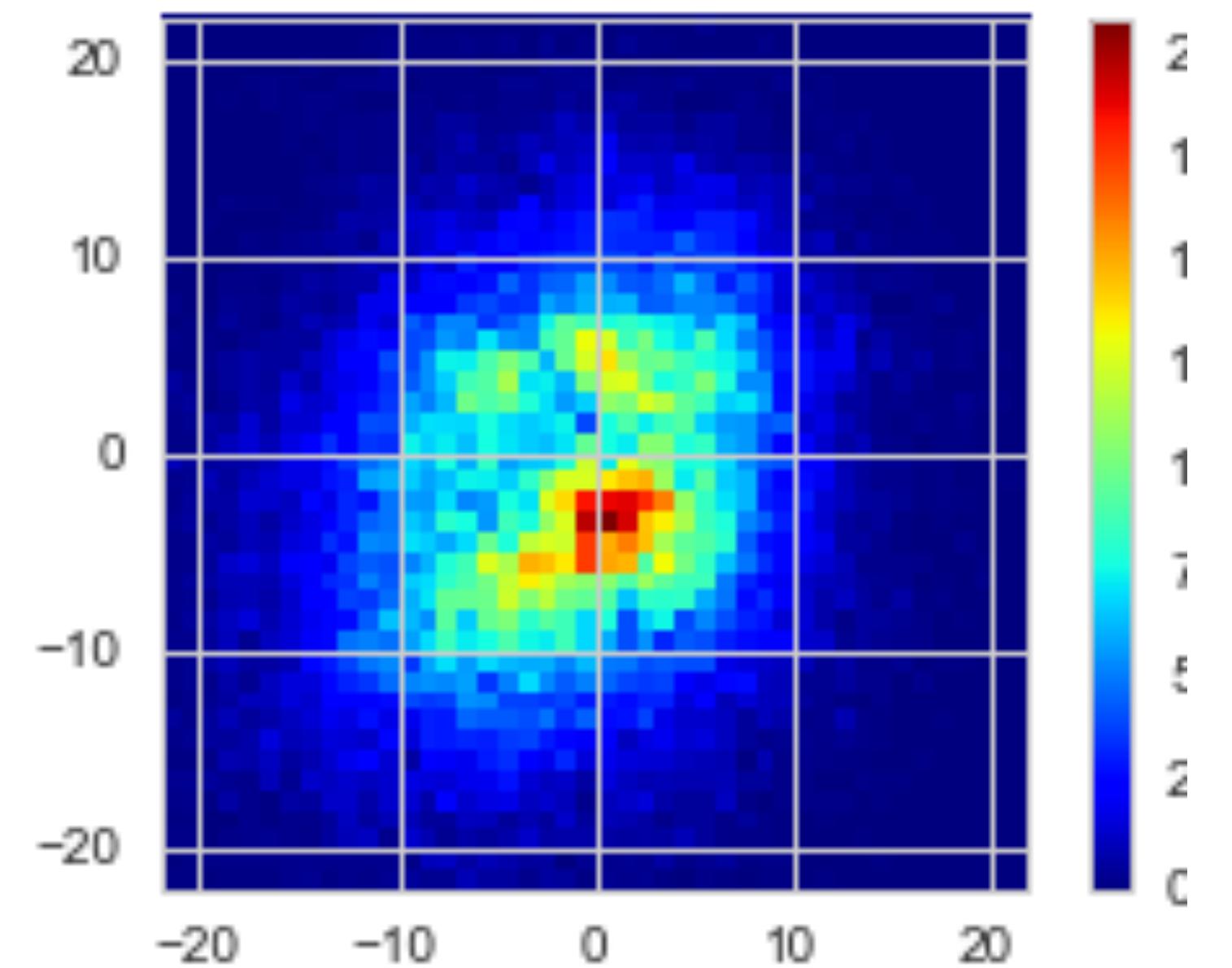
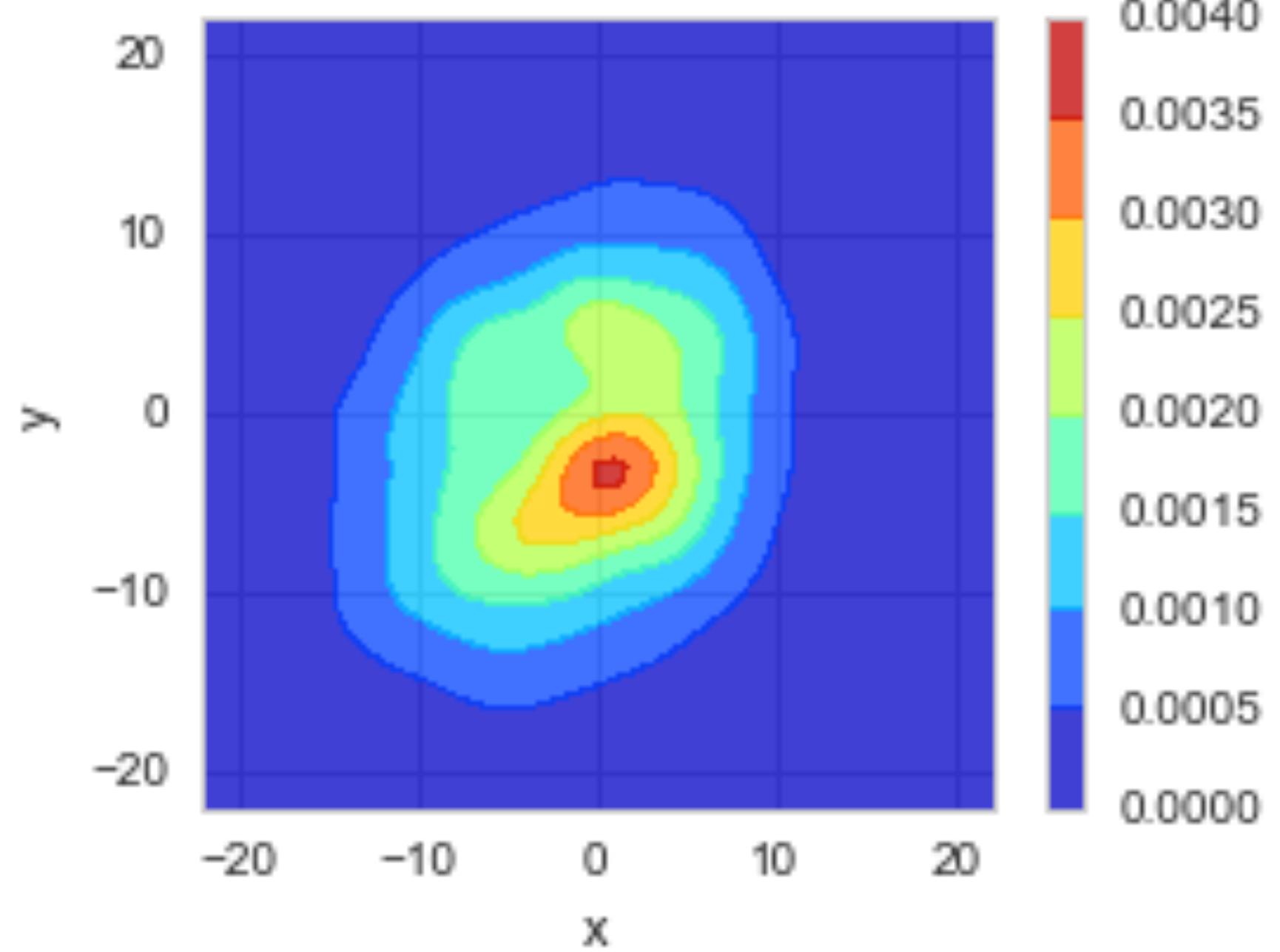
Tropical-Storm-Force Wind Speed Probabilities

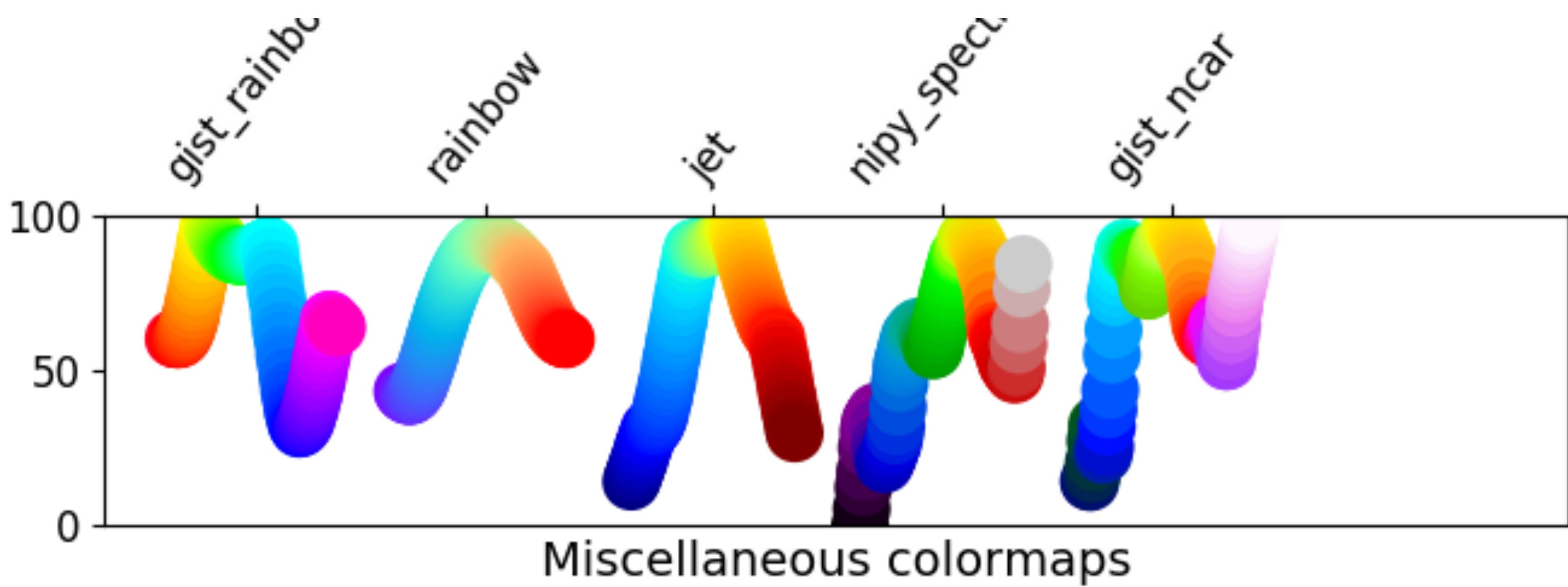
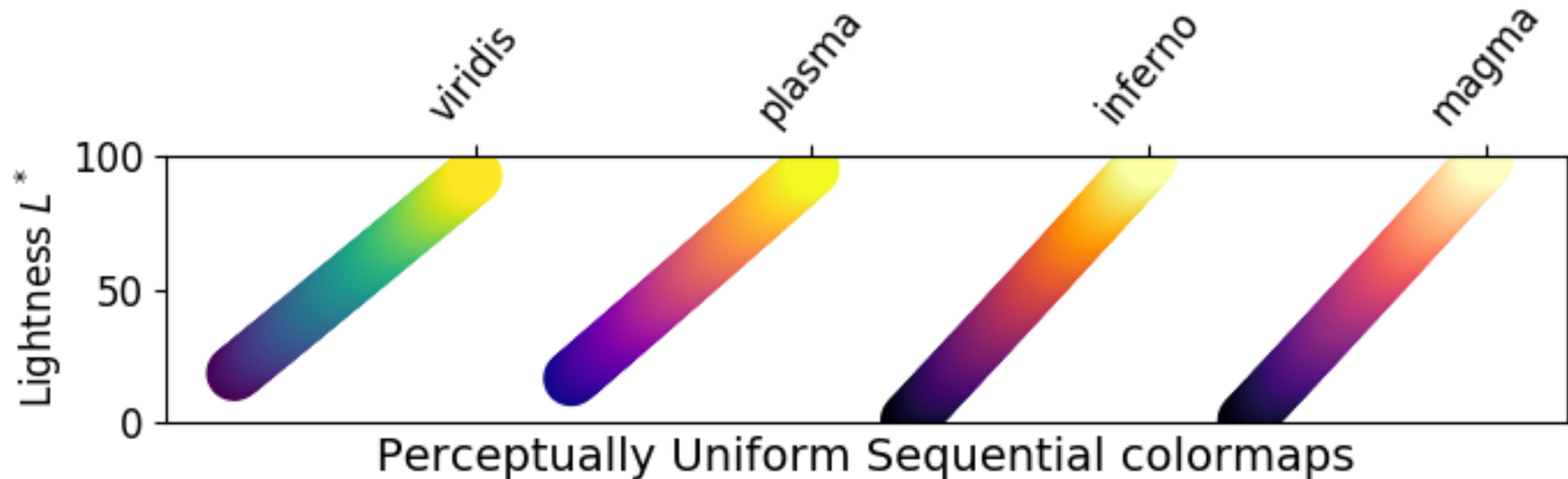
For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04

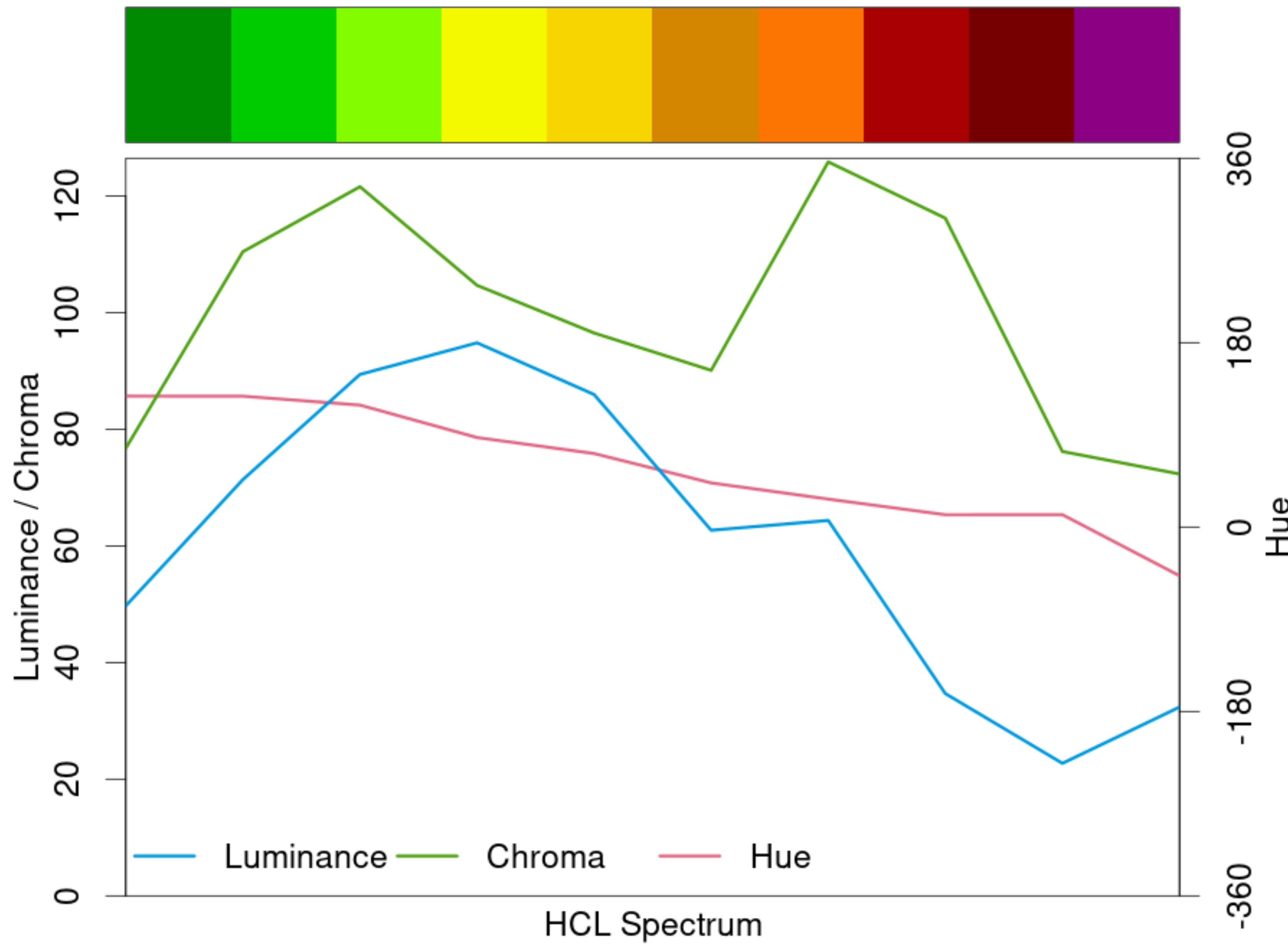


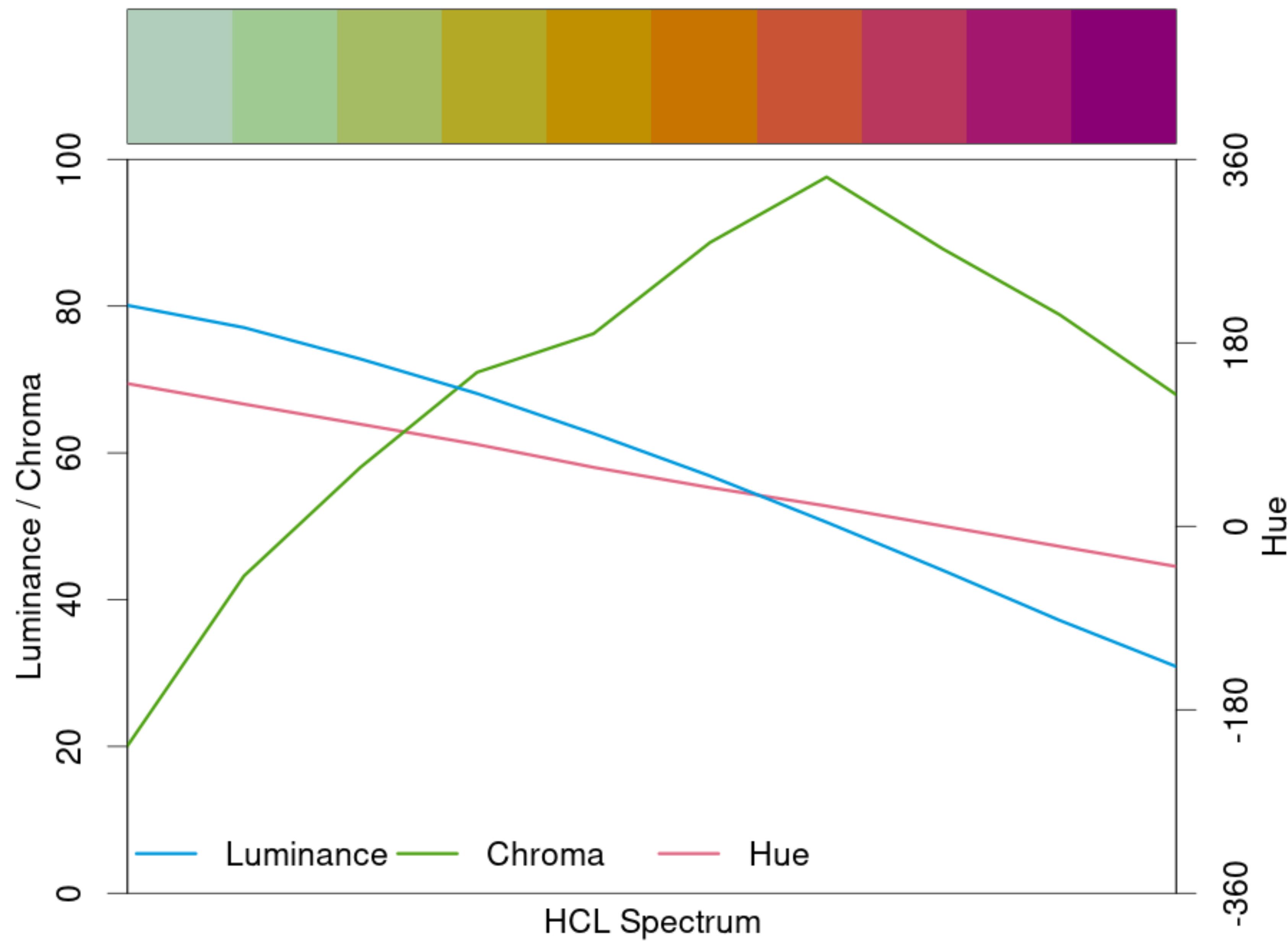
Probability of tropical-storm-force winds (1-minute average ≥ 39 mph) from all tropical cyclones
○ indicates Hurricane Dorian center location at 2 PM EDT FRI AUG 30, 2019 (Forecast/Advisory #26)





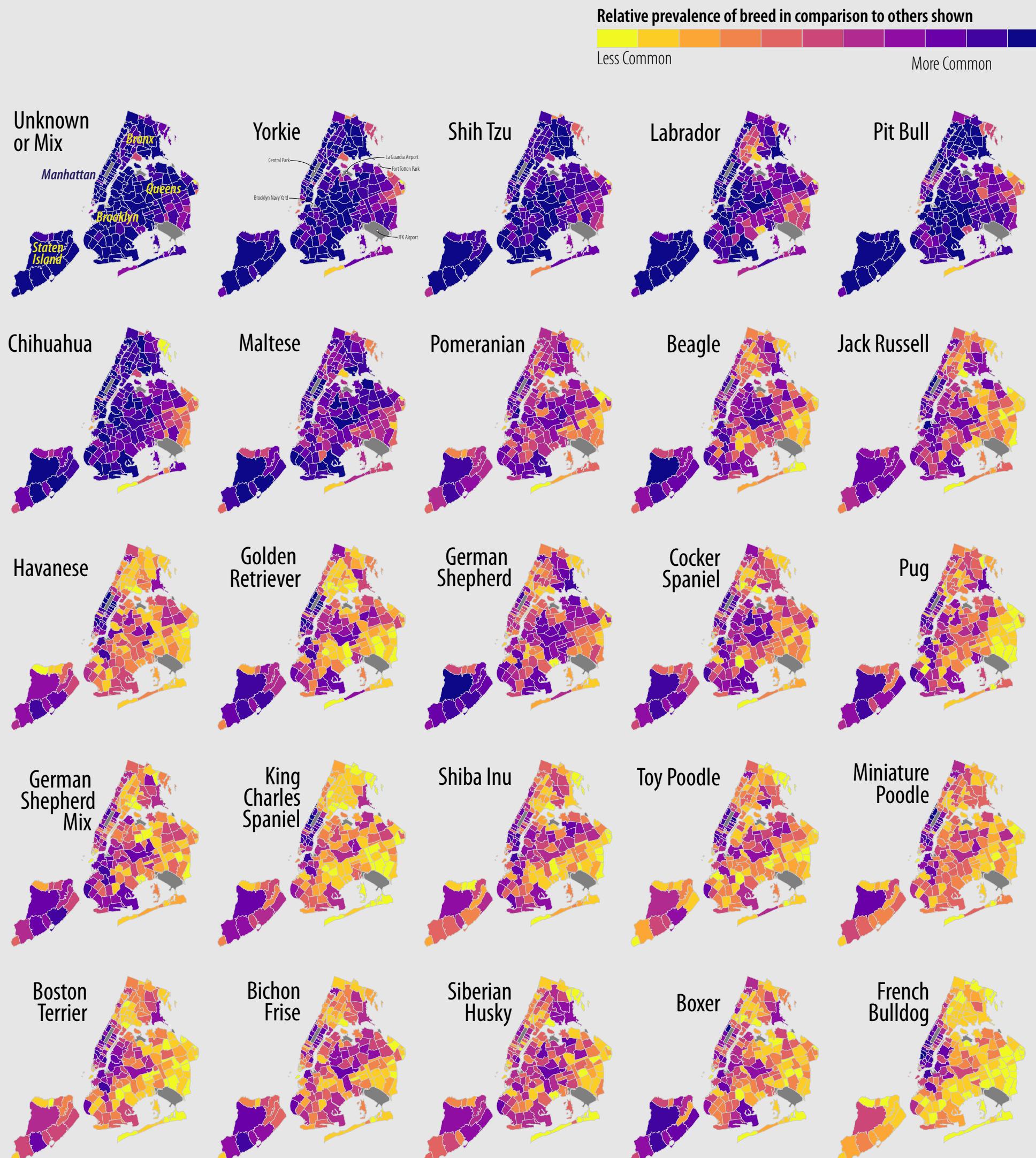






Dogs of New York

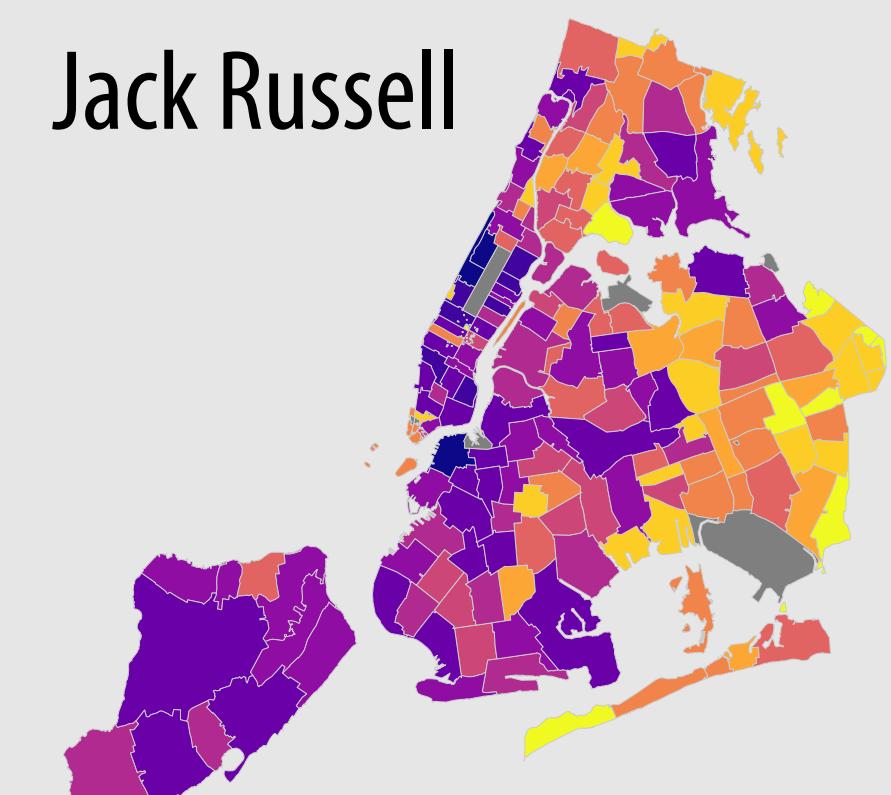
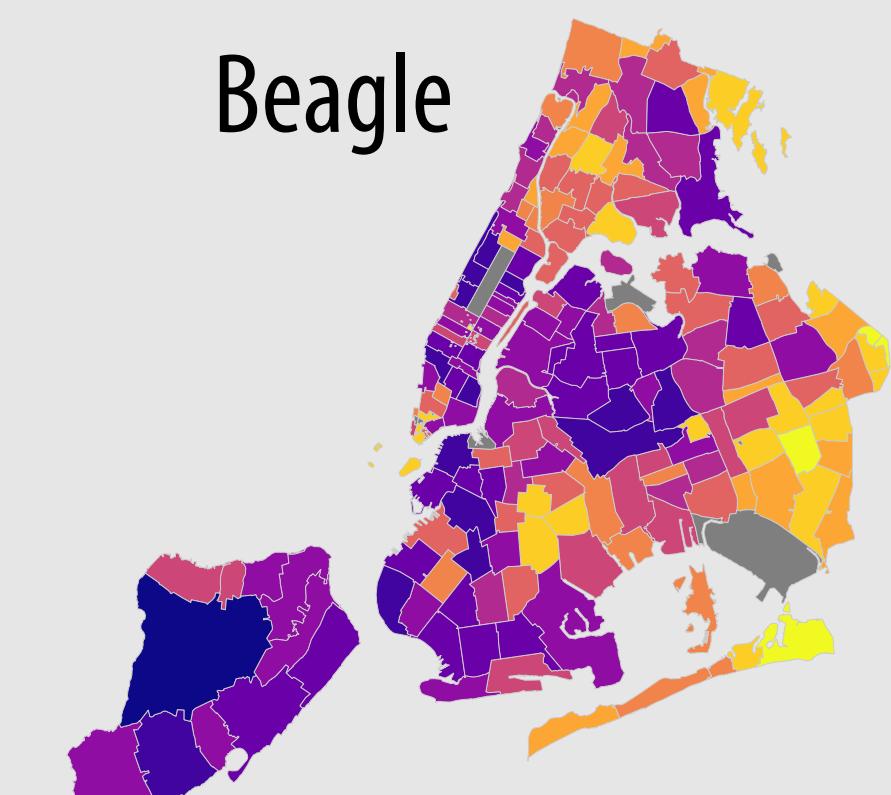
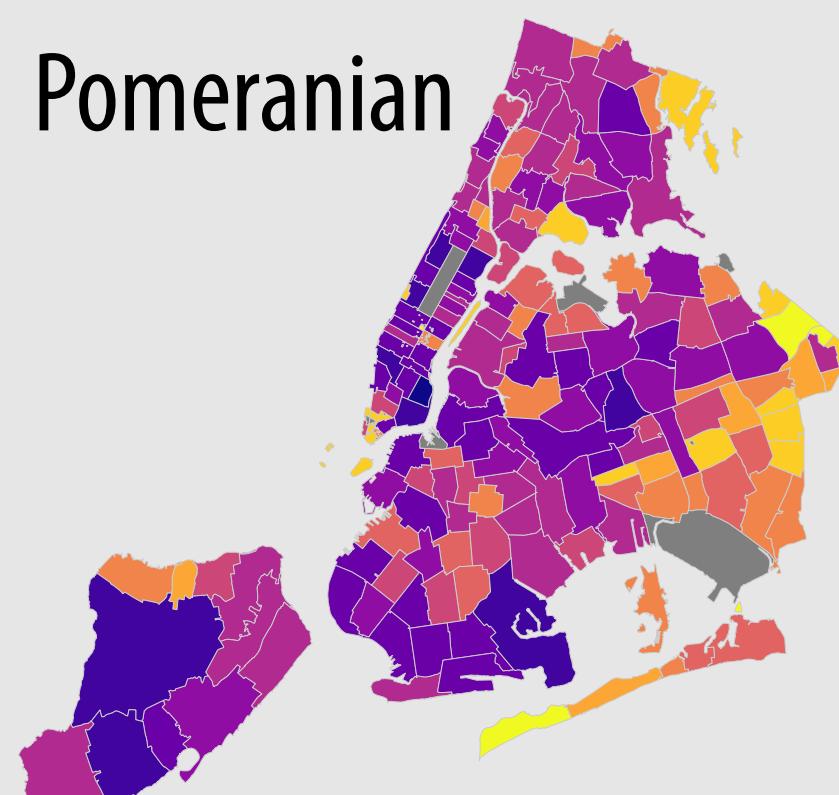
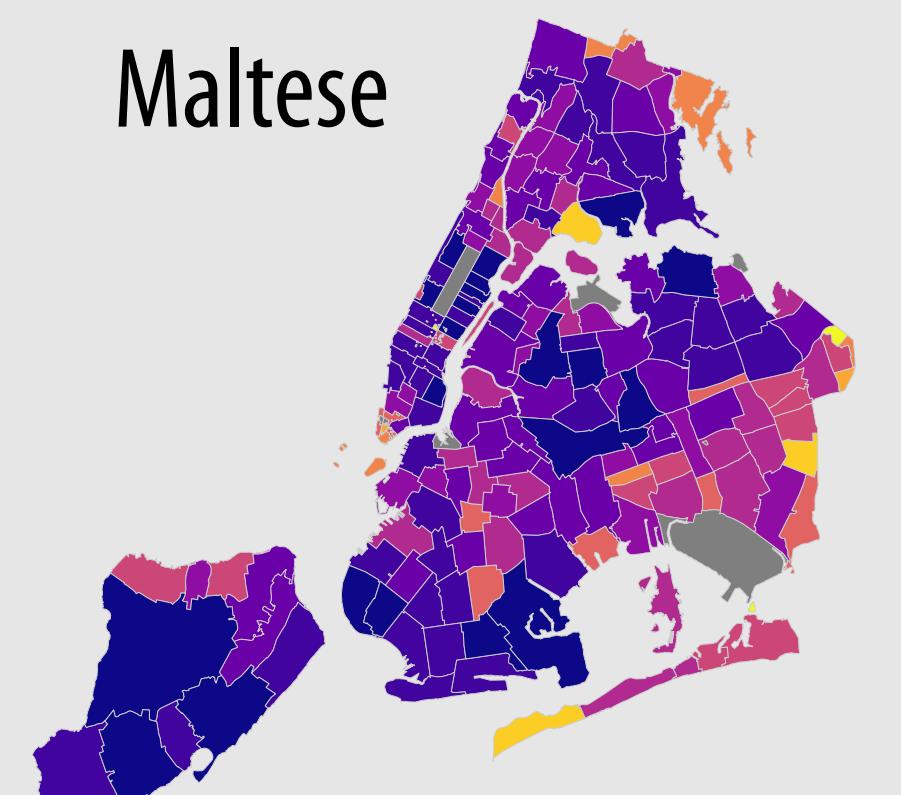
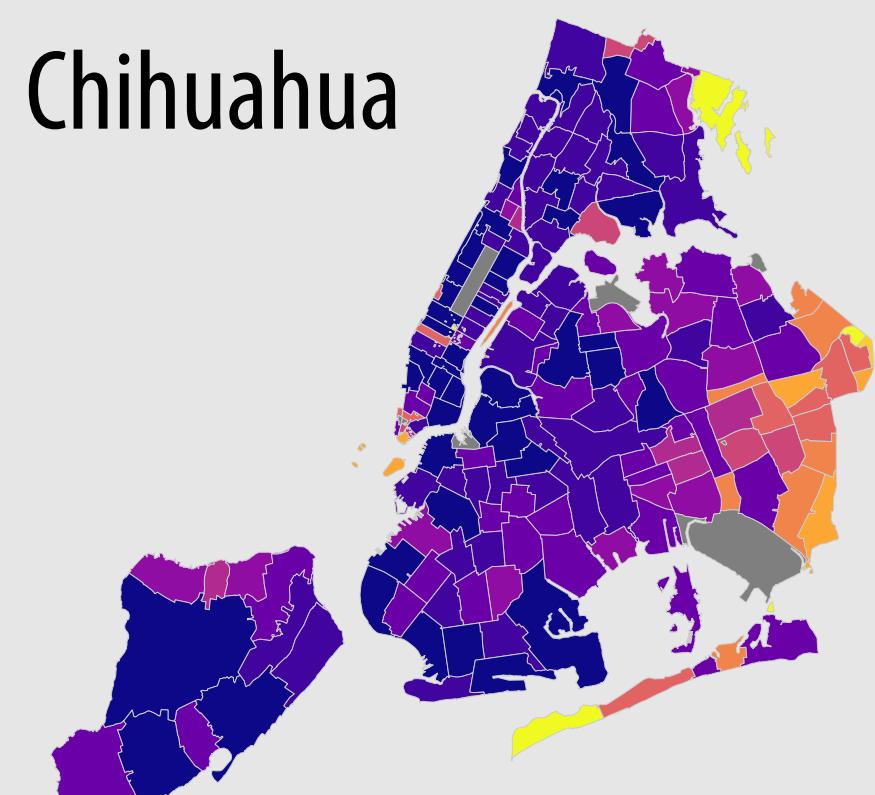
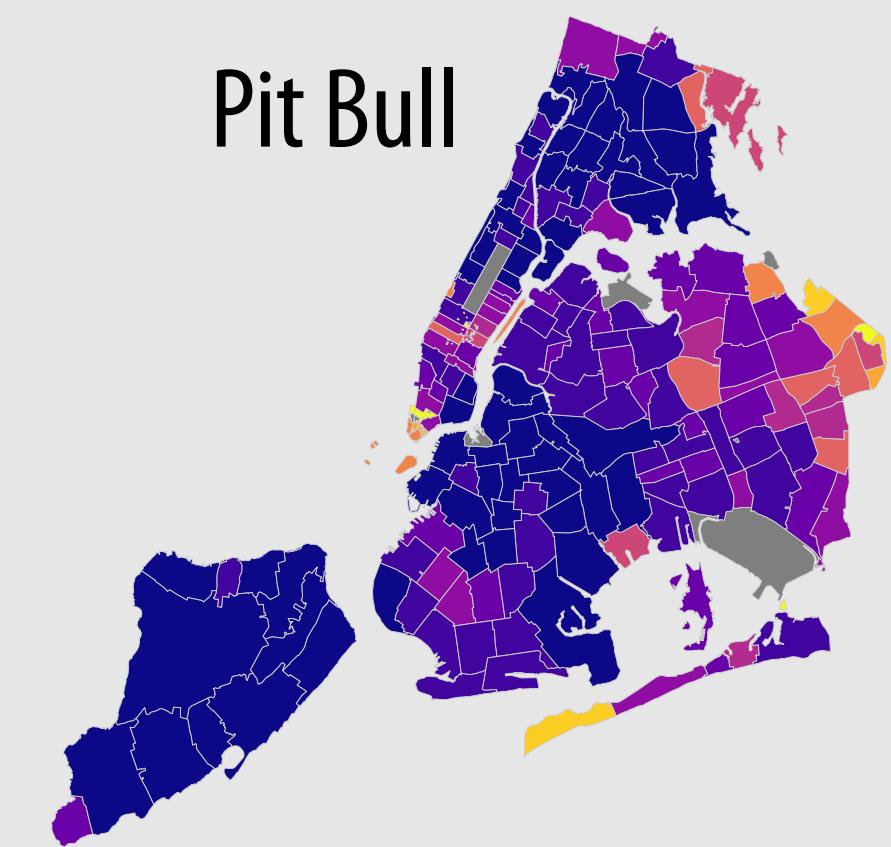
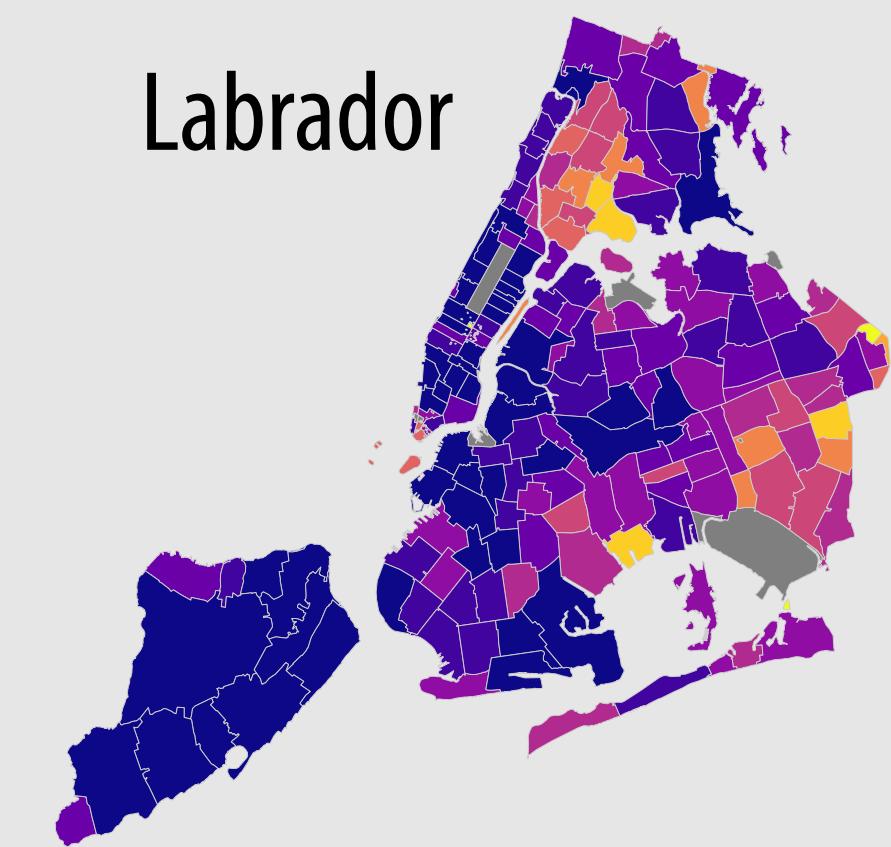
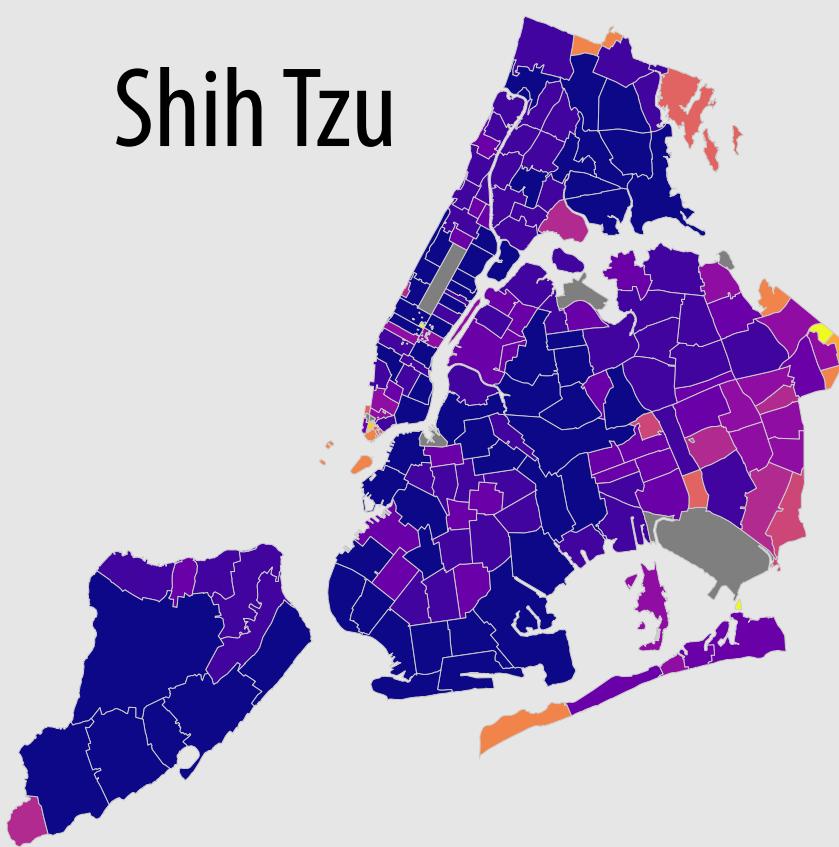
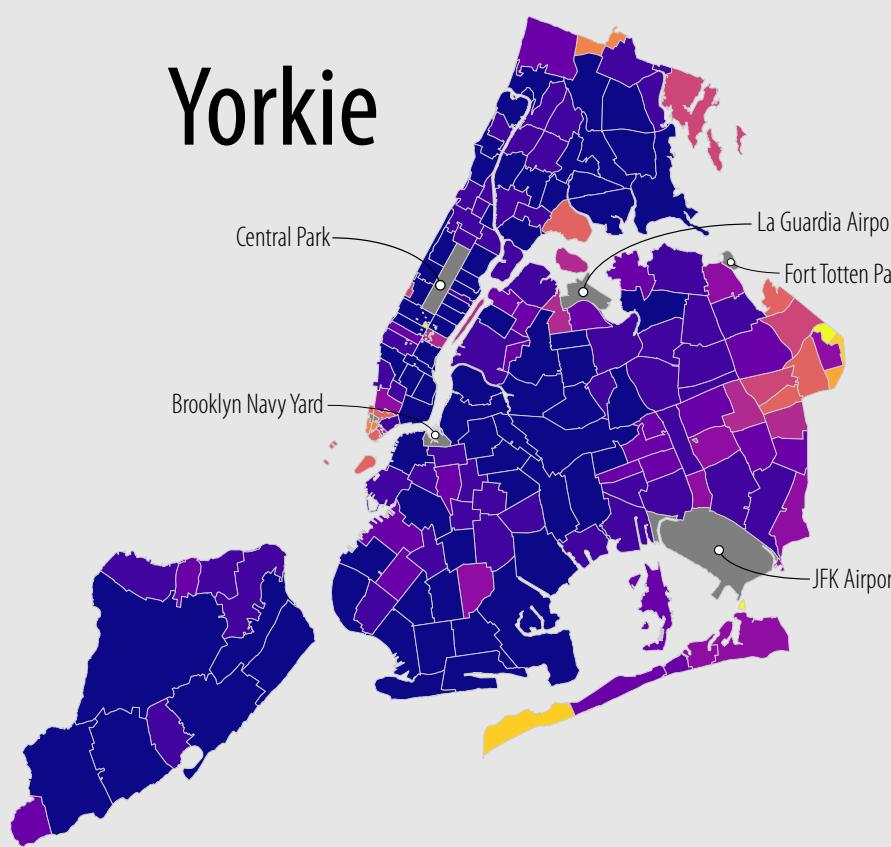
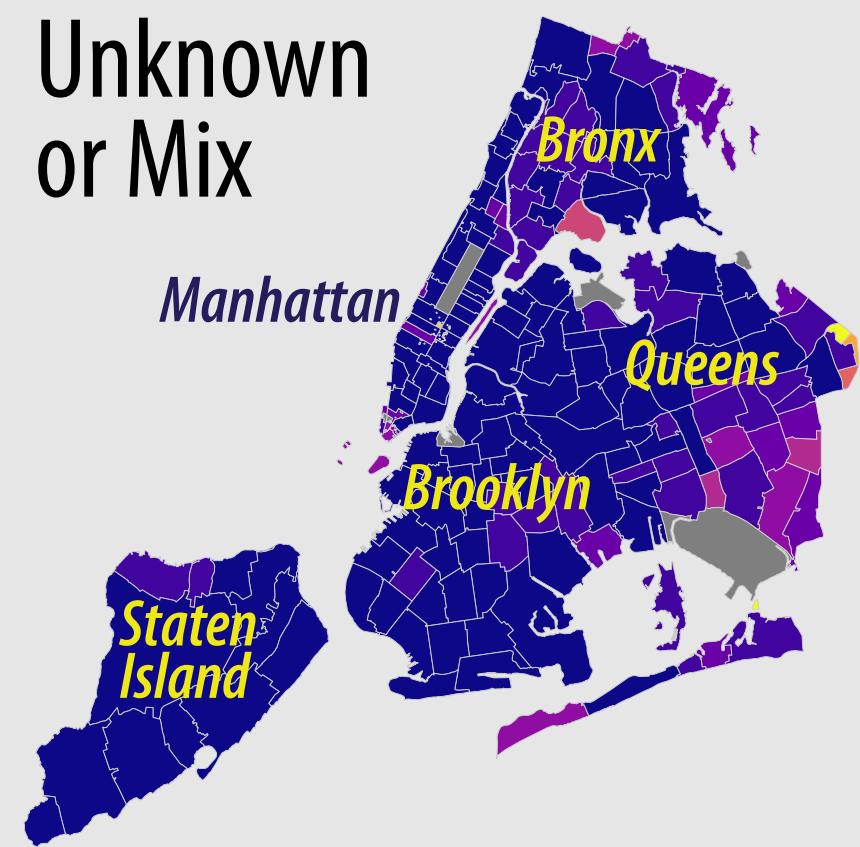
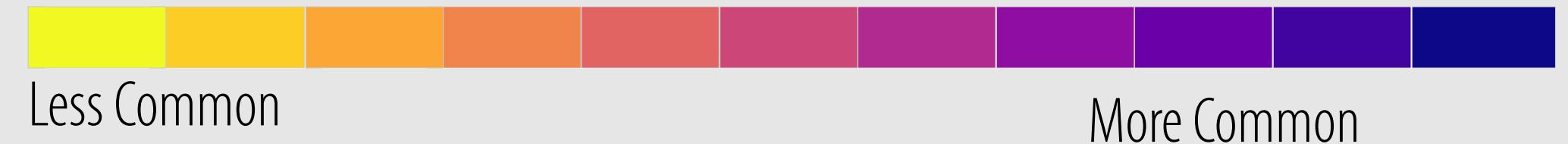
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.



Dog Breeds in New York

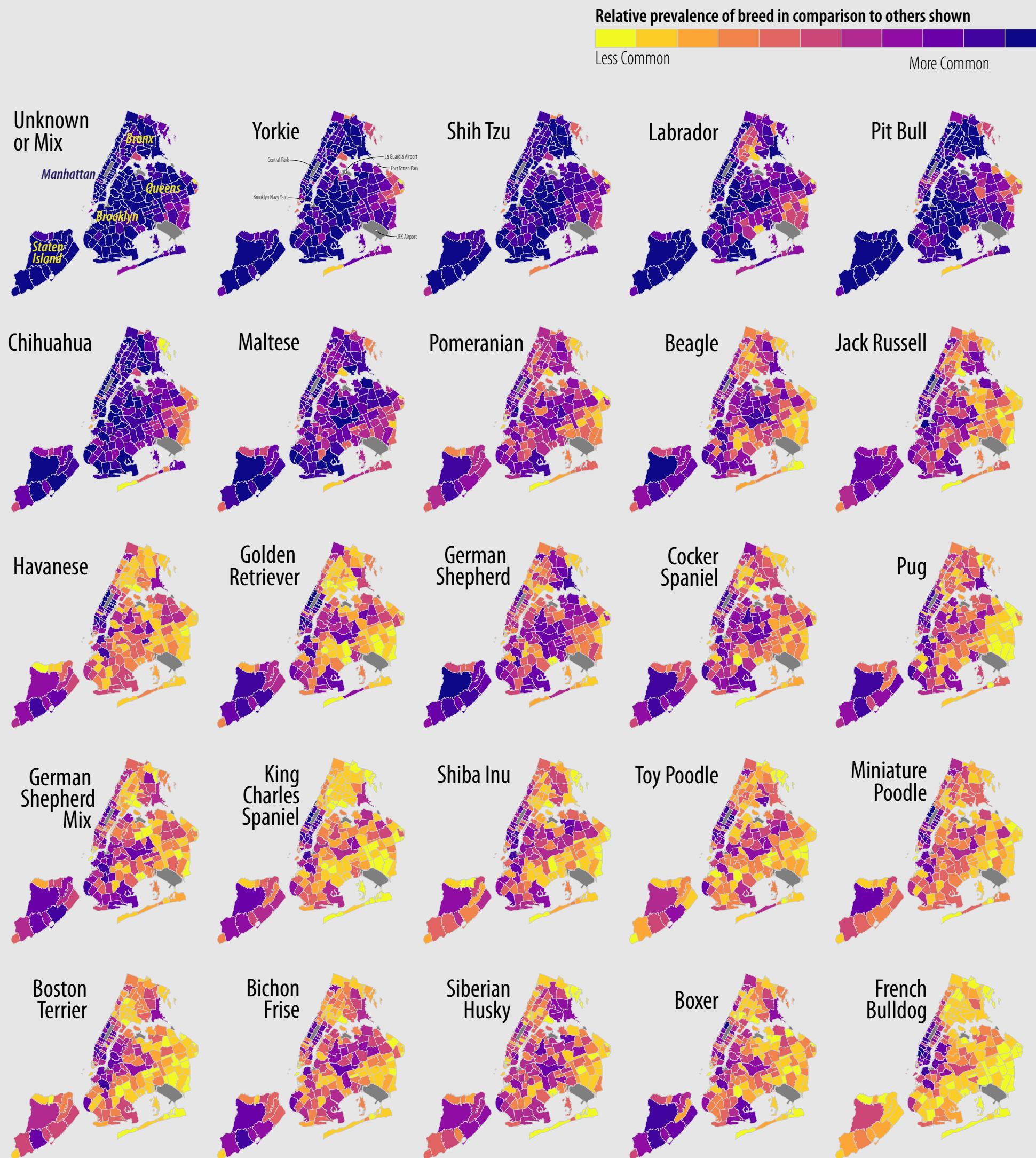
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.

Relative prevalence of breed in comparison to others shown



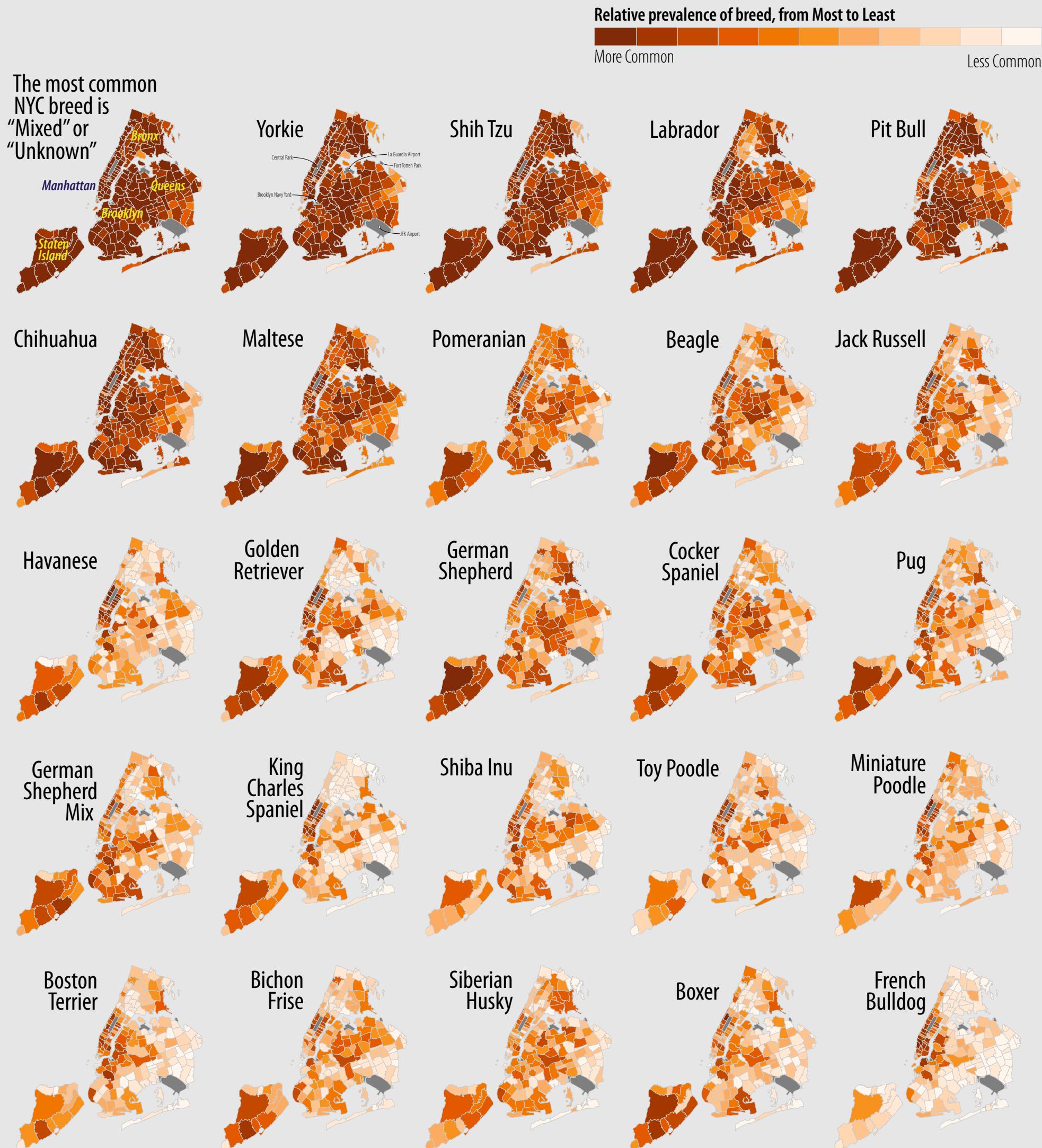
Dogs of New York

Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.

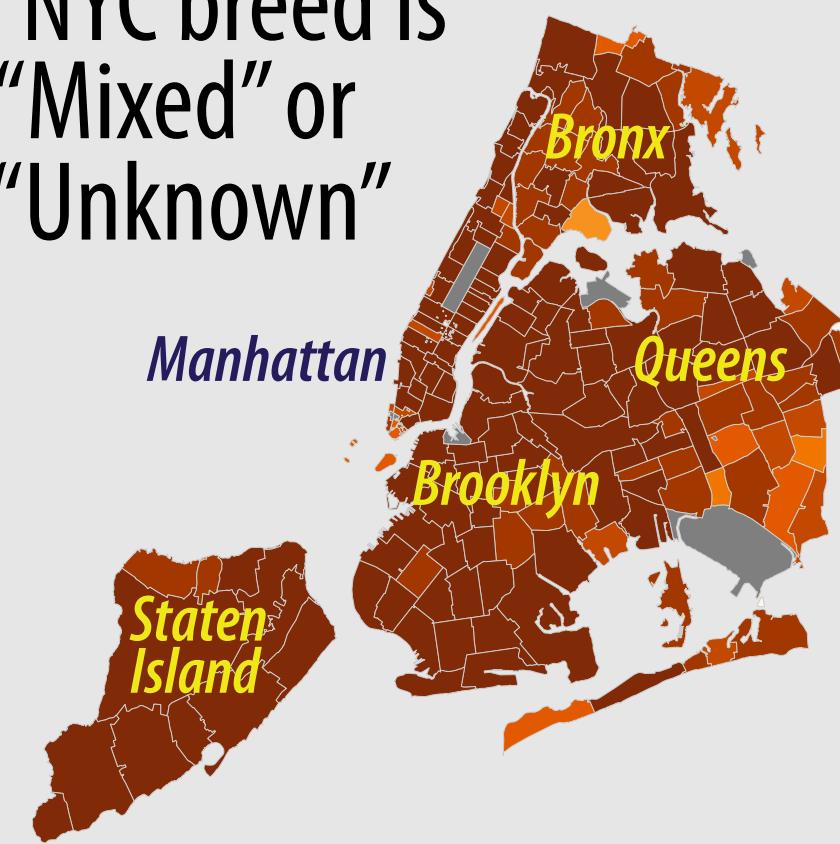


Dogs of New York

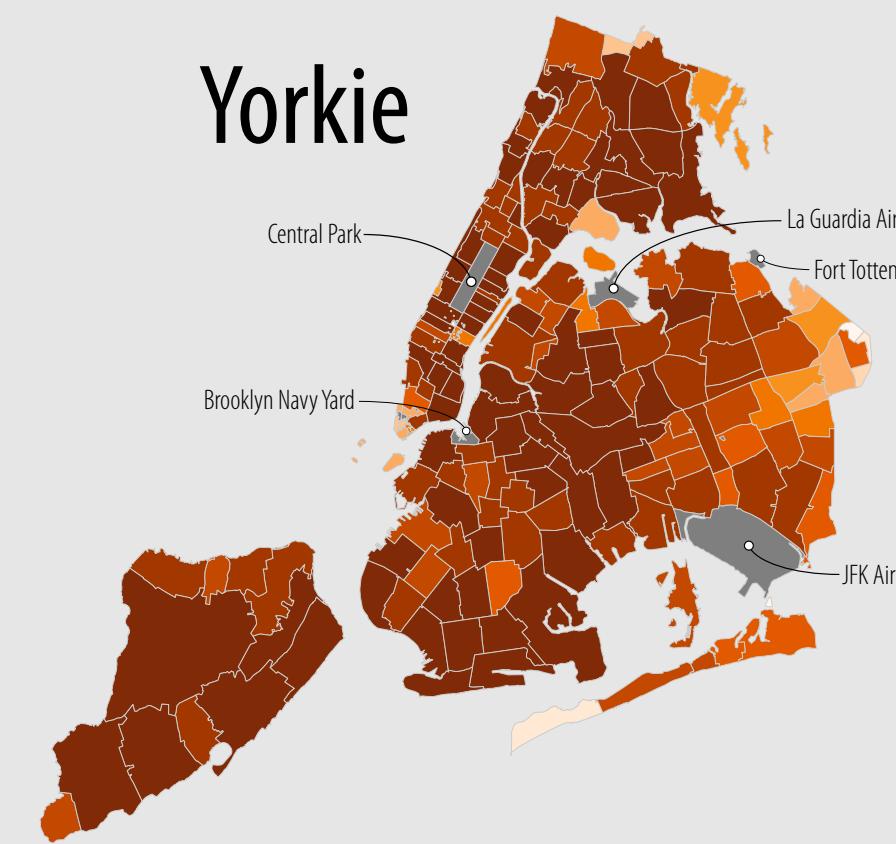
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.



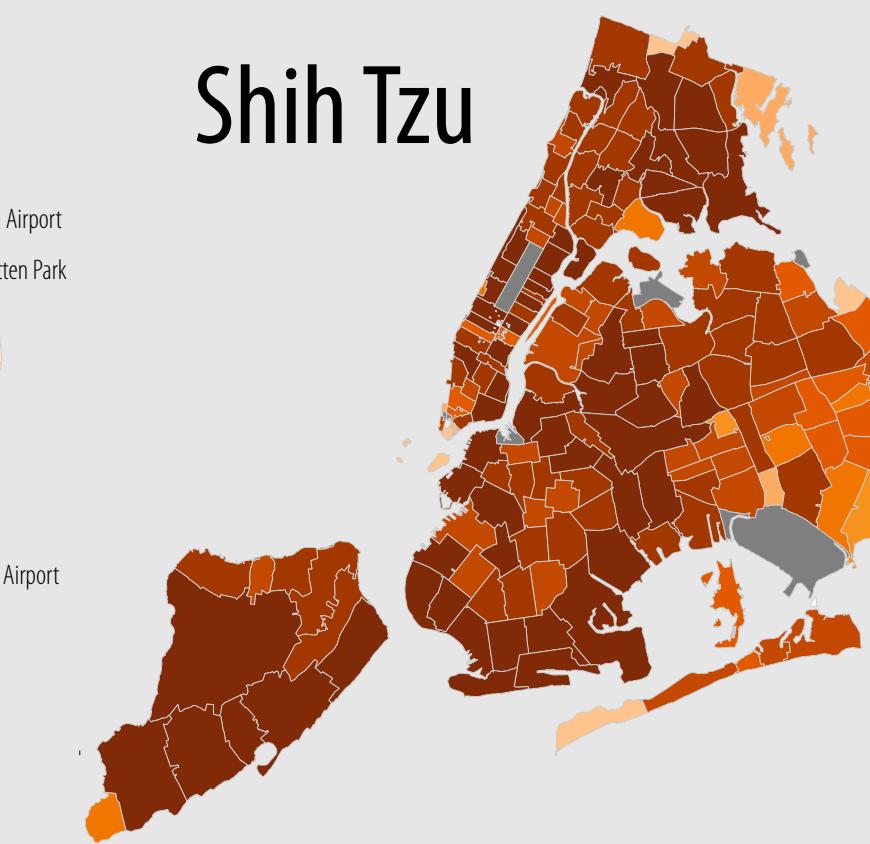
The most common
NYC breed is
“Mixed” or
“Unknown”



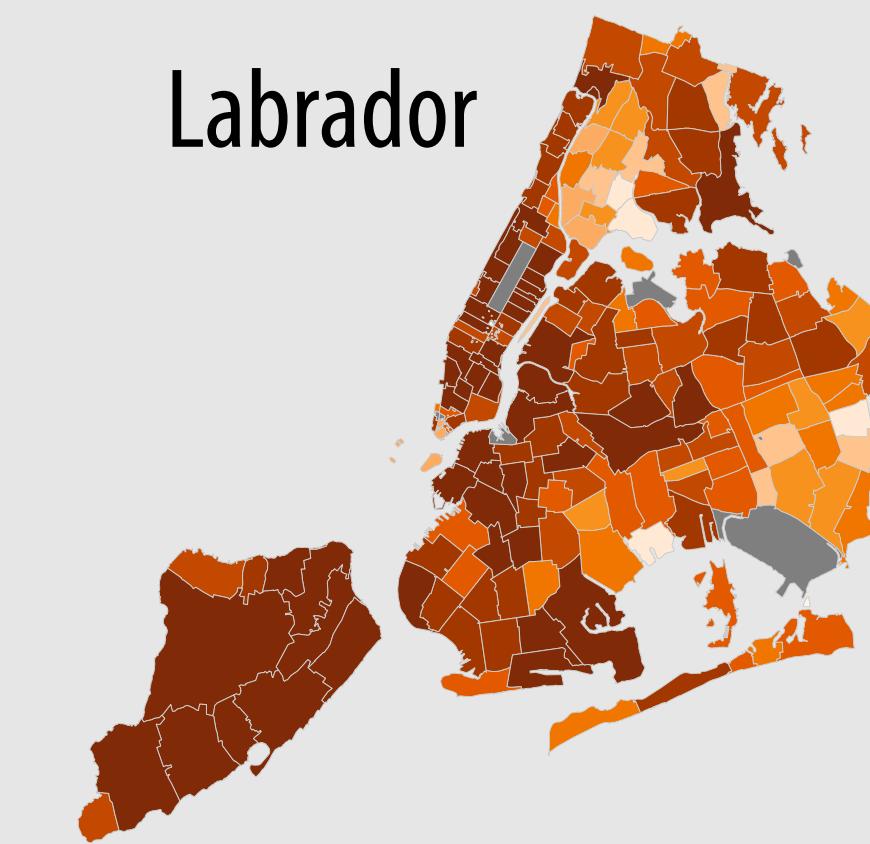
Yorkie



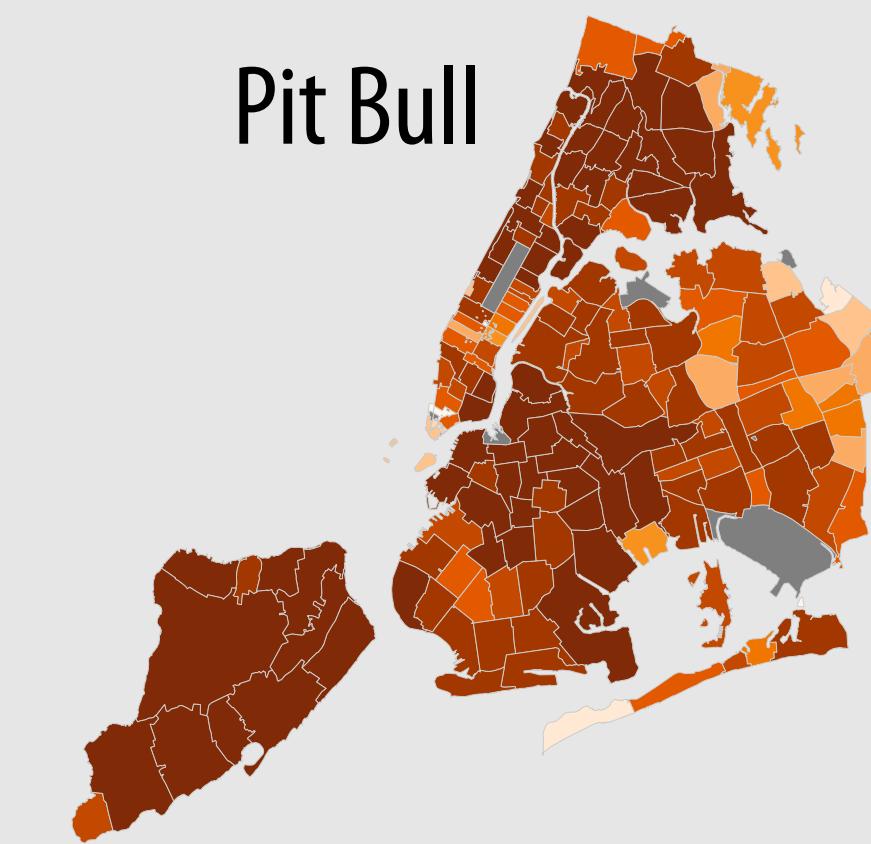
Shih Tzu



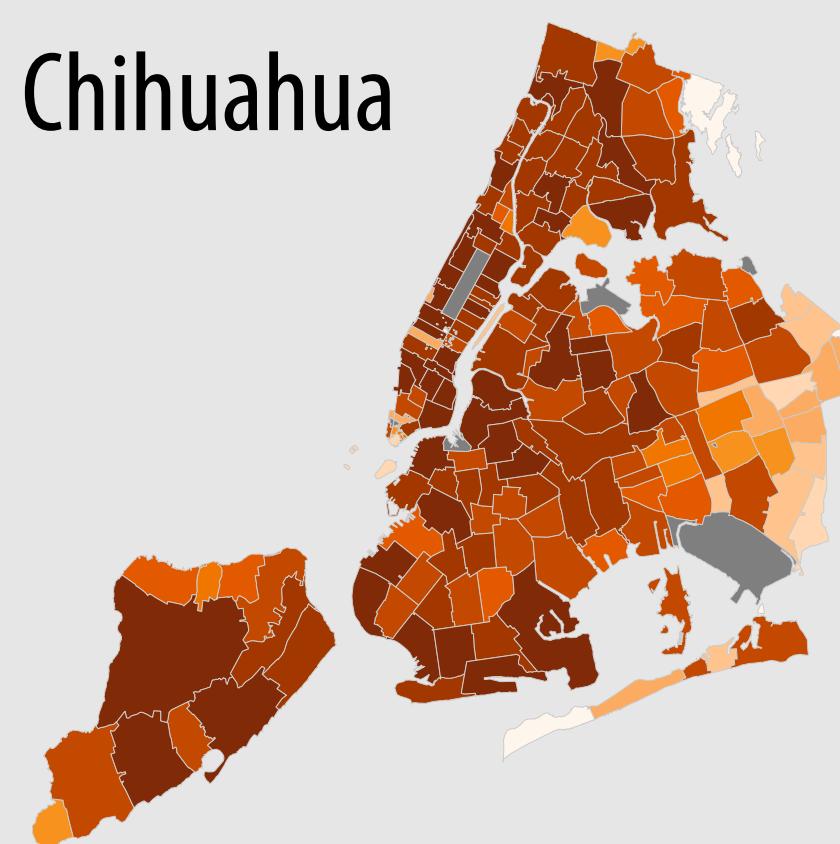
Labrador



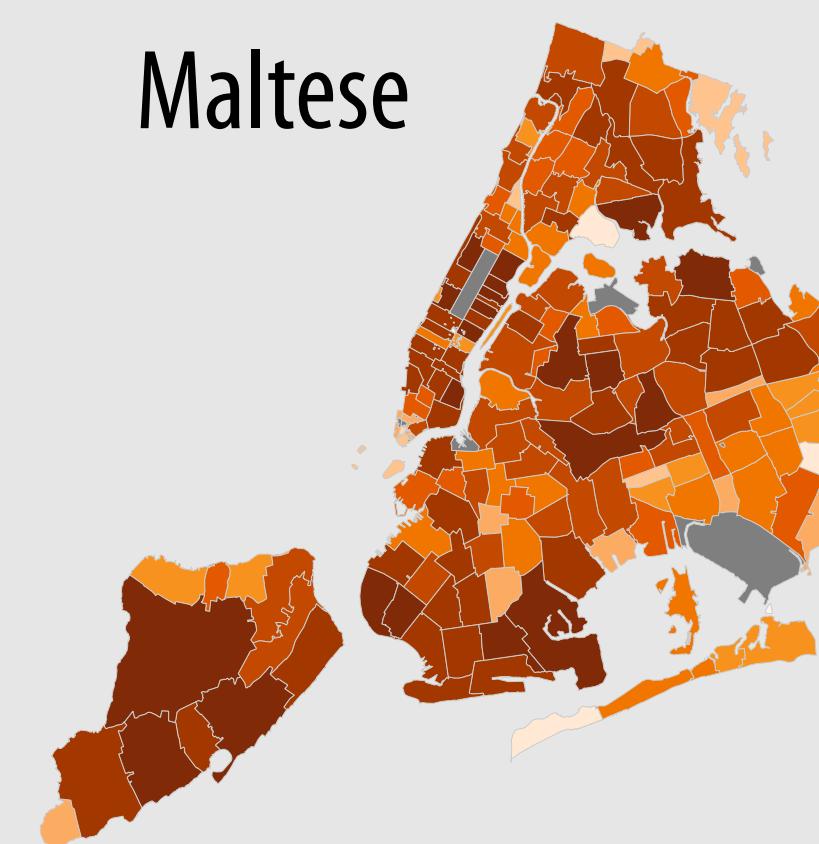
Pit Bull



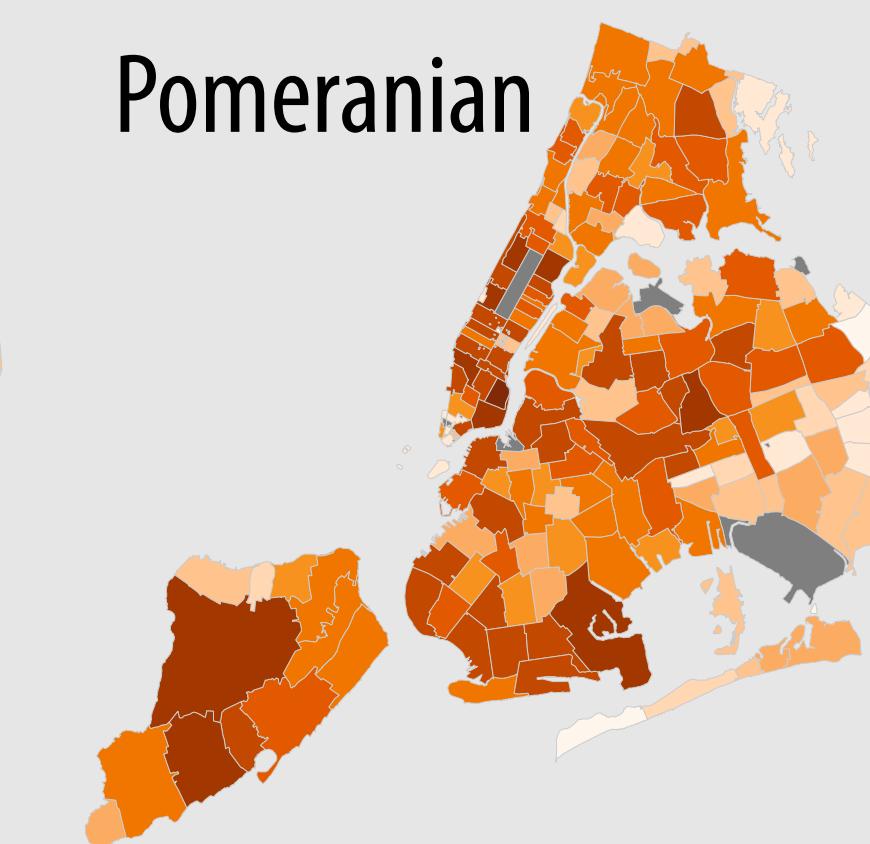
Chihuahua



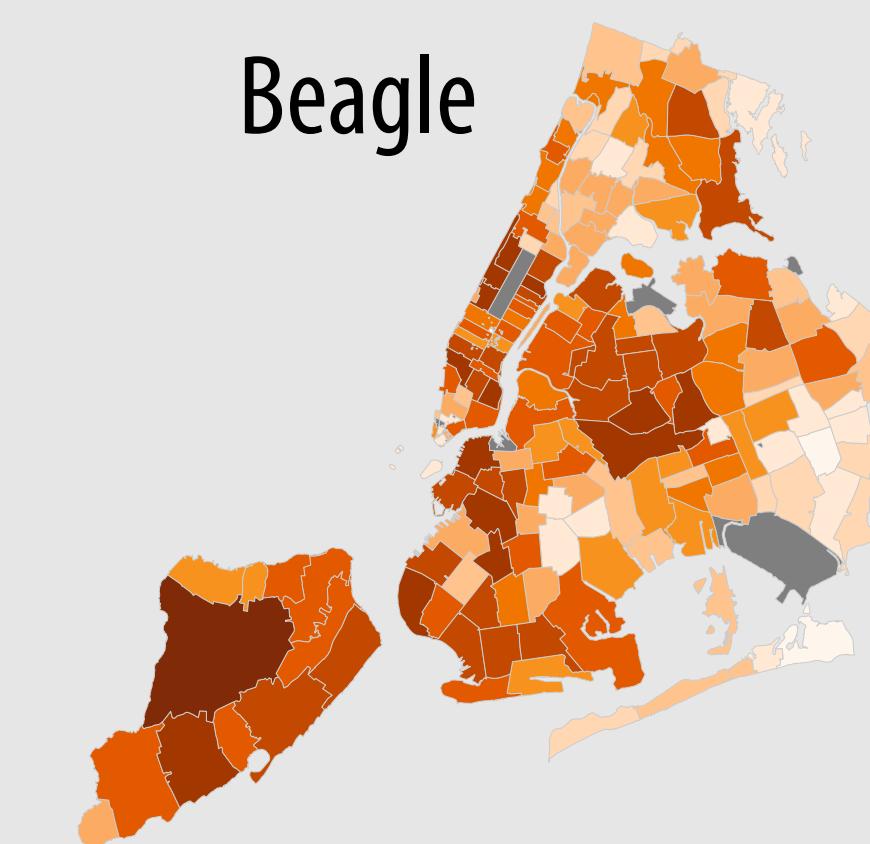
Maltese



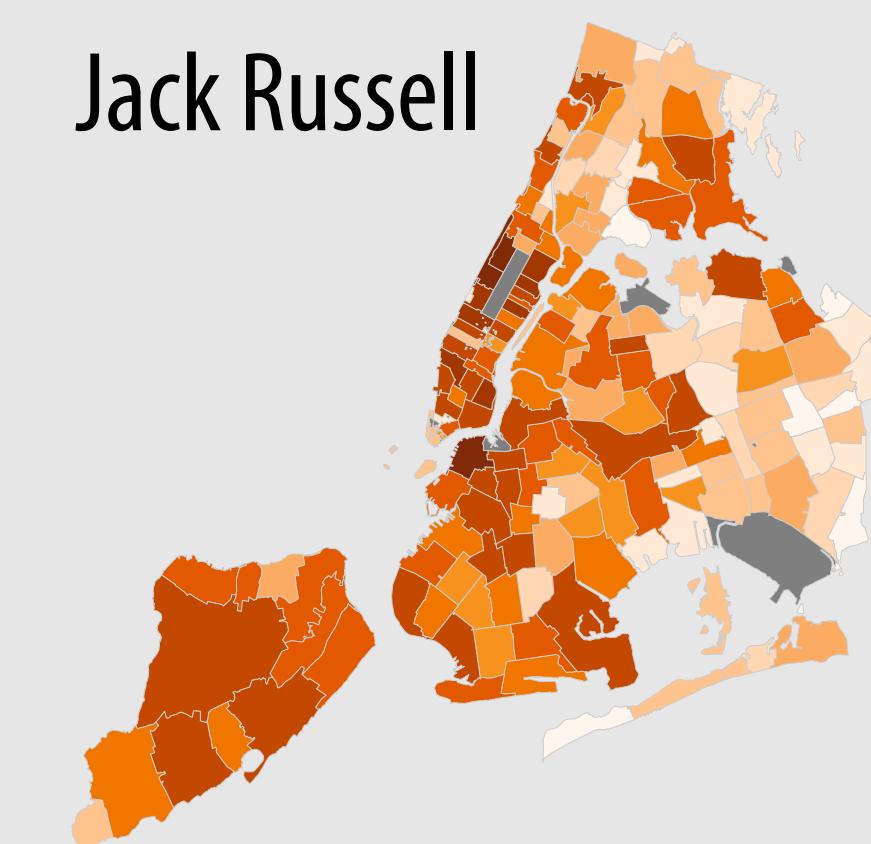
Pomeranian



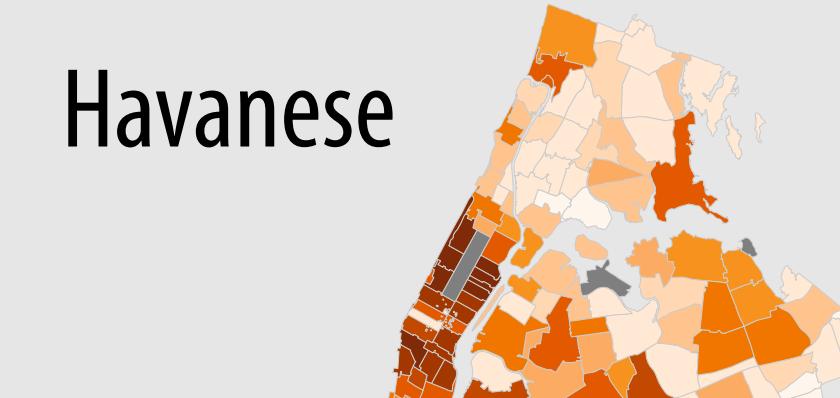
Beagle



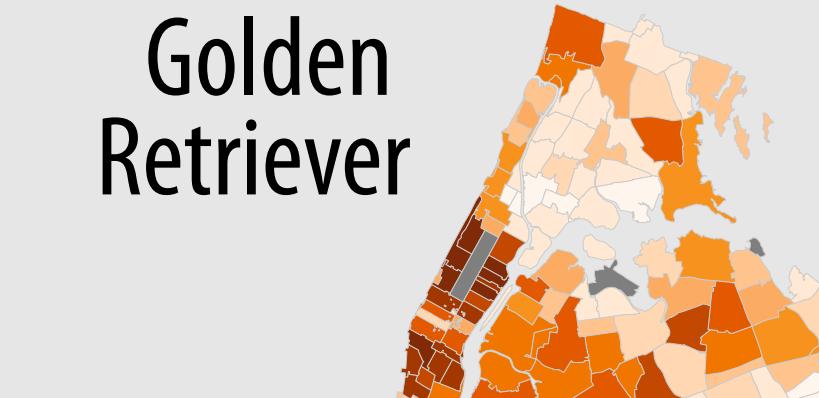
Jack Russell



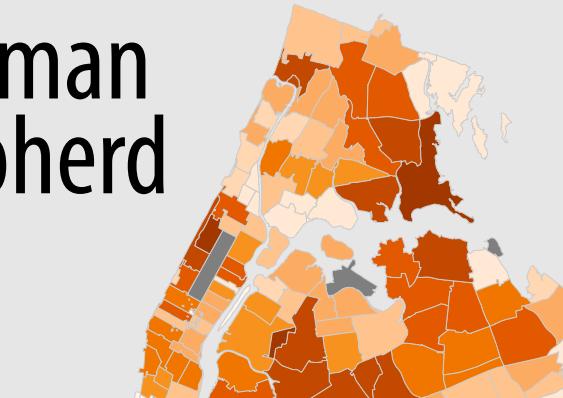
Havanese



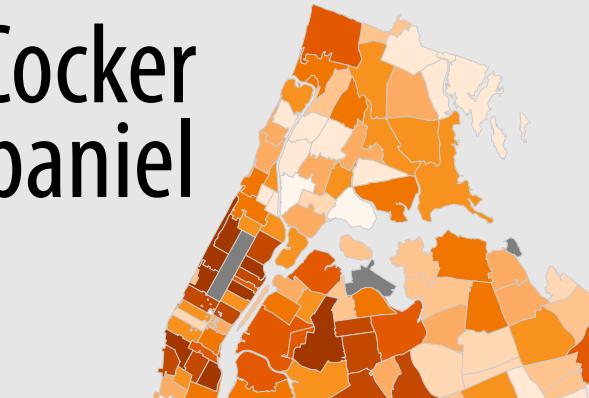
Golden Retriever



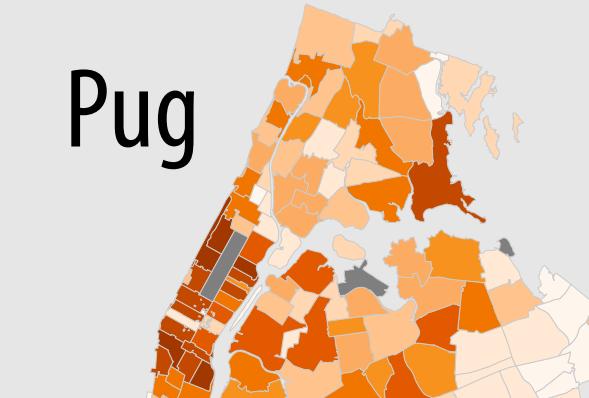
German Shepherd



Cocker Spaniel



Pug



Relative prevalence of breed, from Most to Least

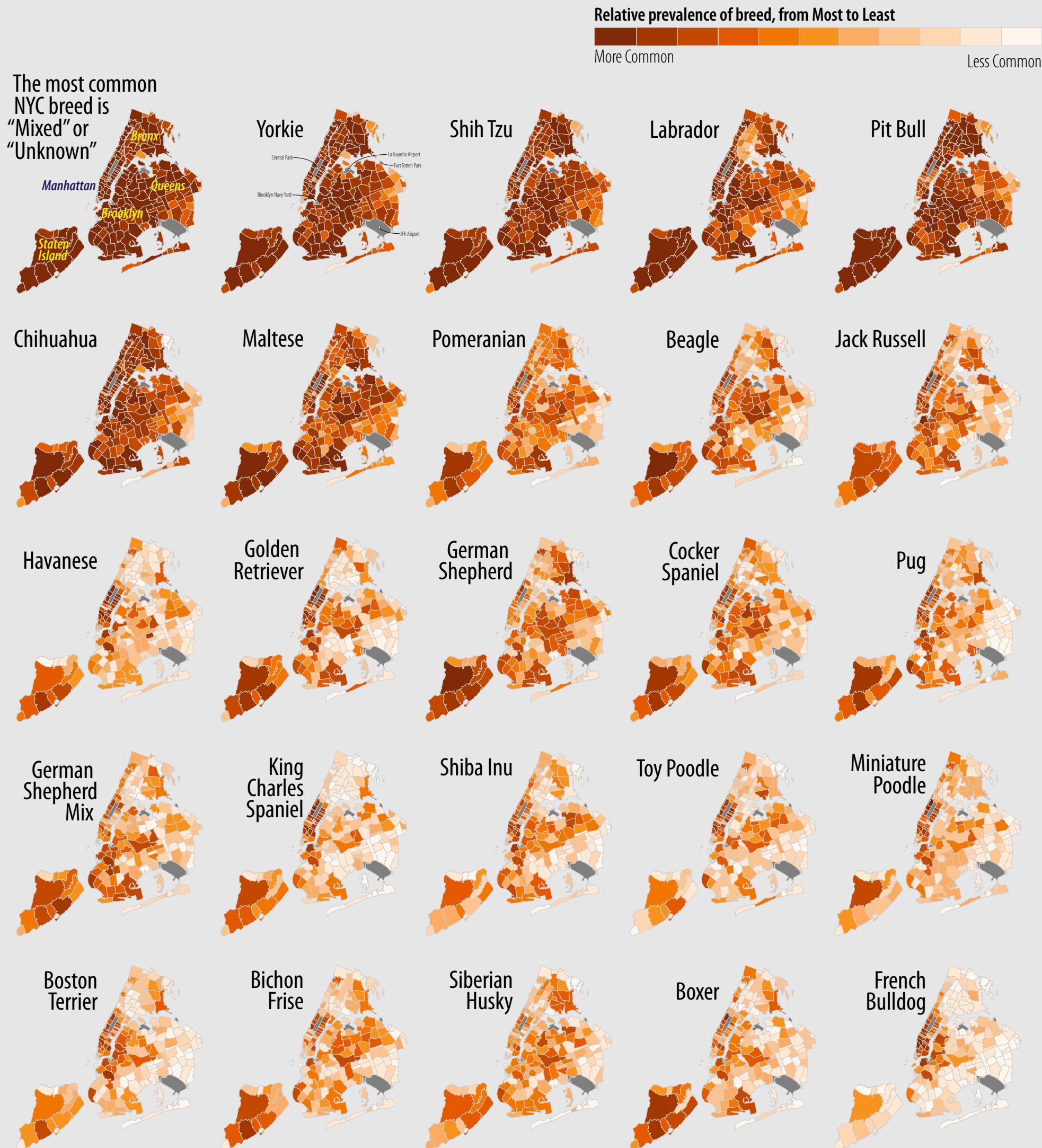


More Common

Less Common

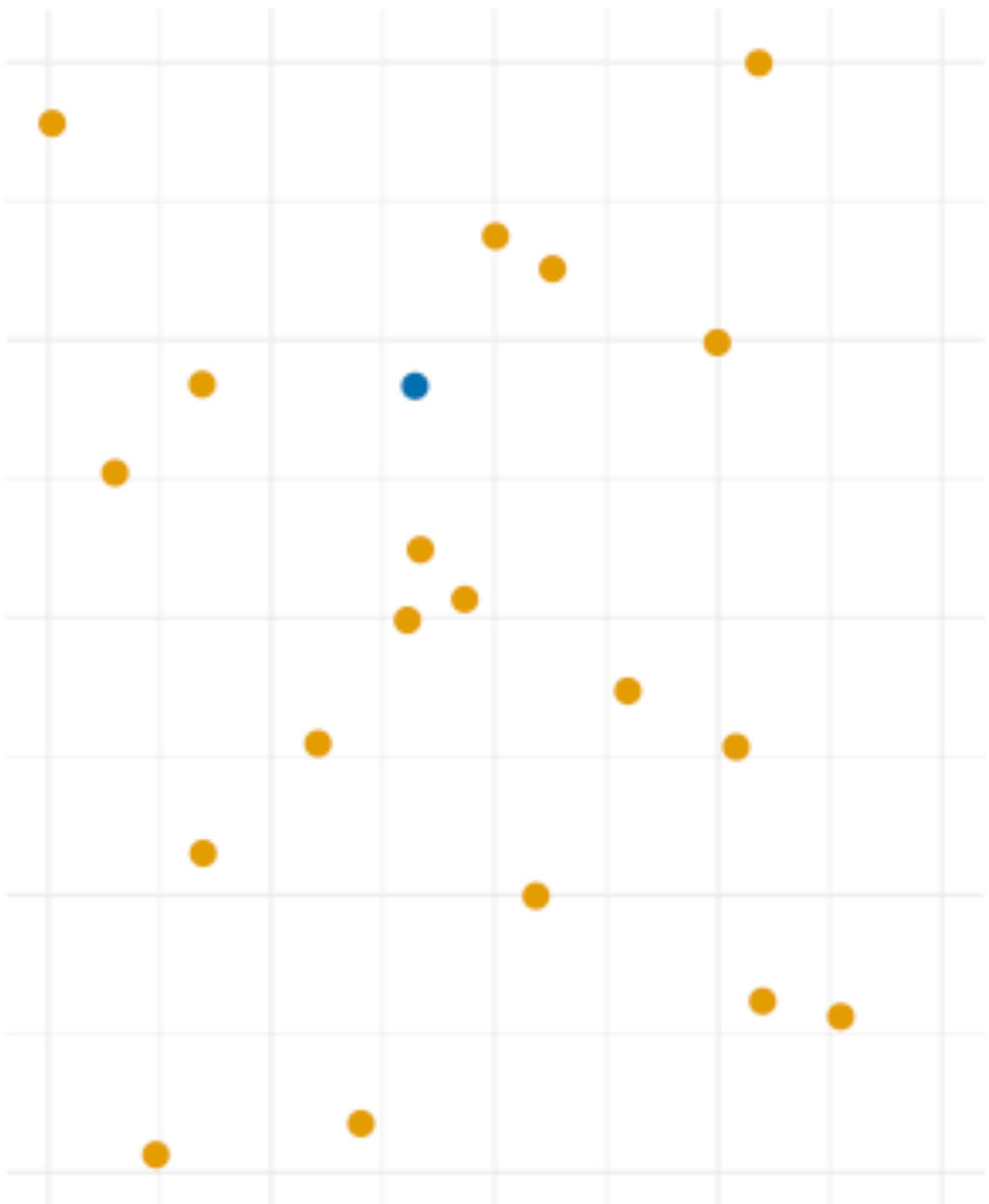
Dogs of New York

Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.

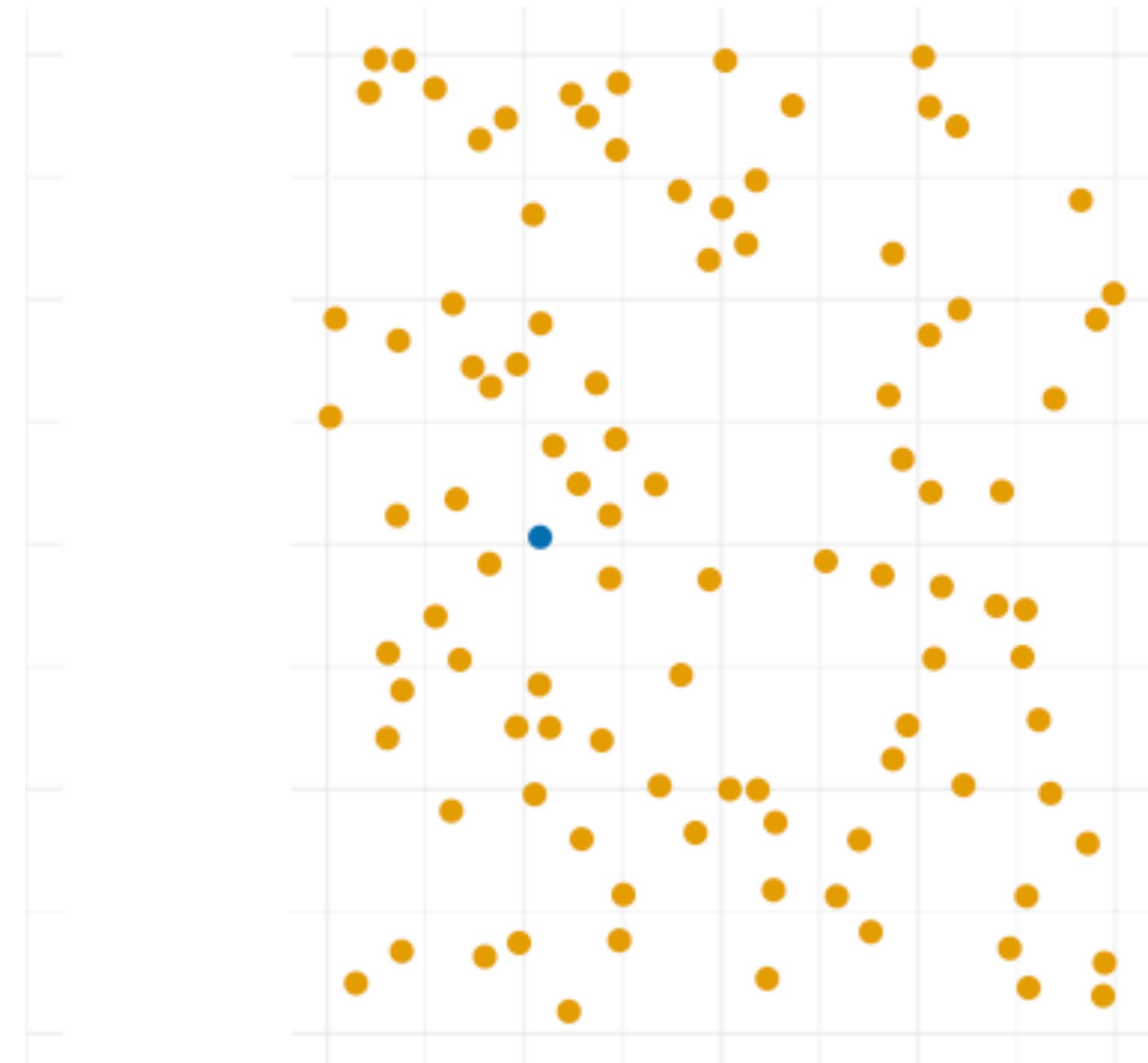


PRE-ATTENTIVE PROCESSING

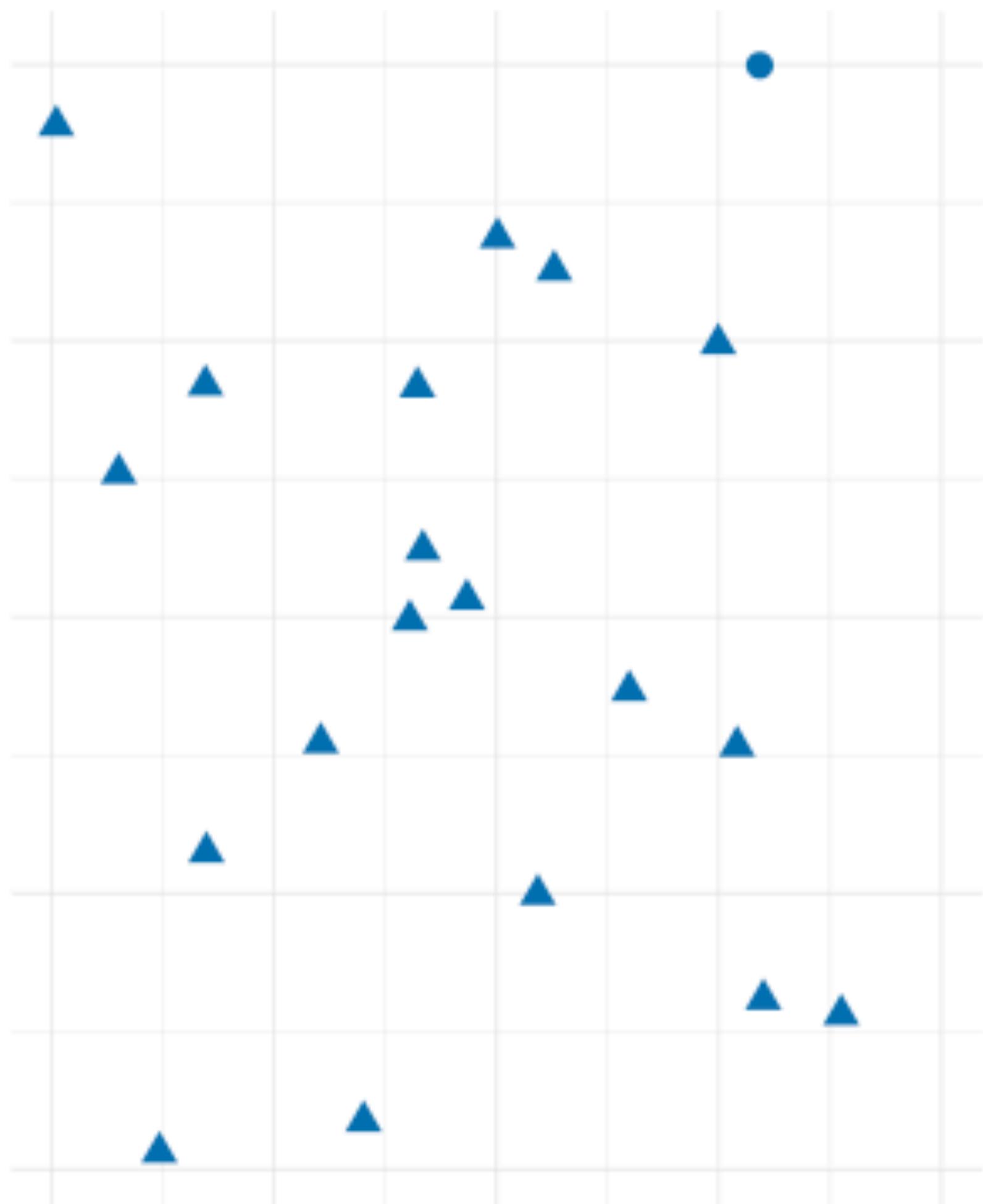
Color Only, N=20



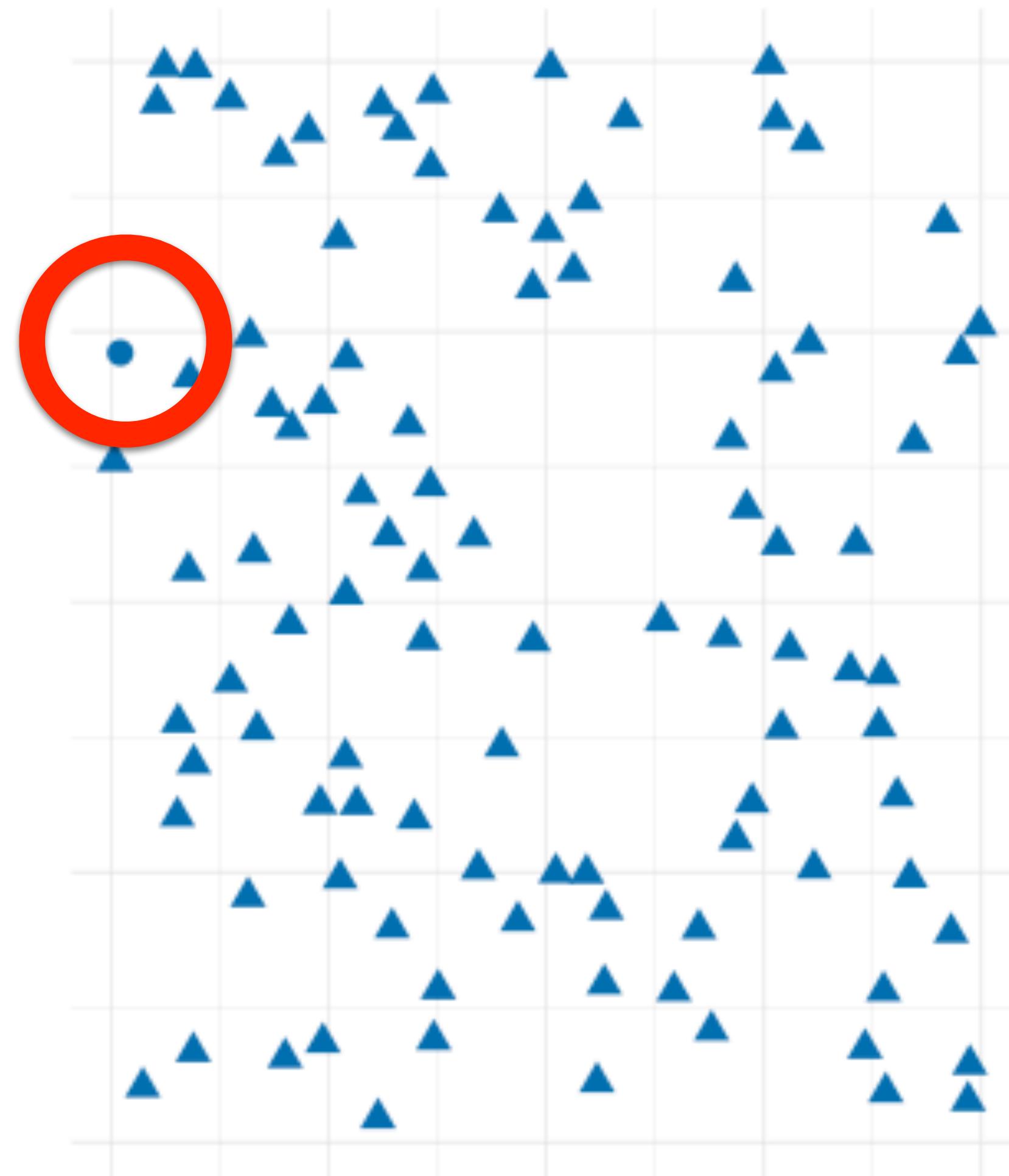
Color Only, N=100



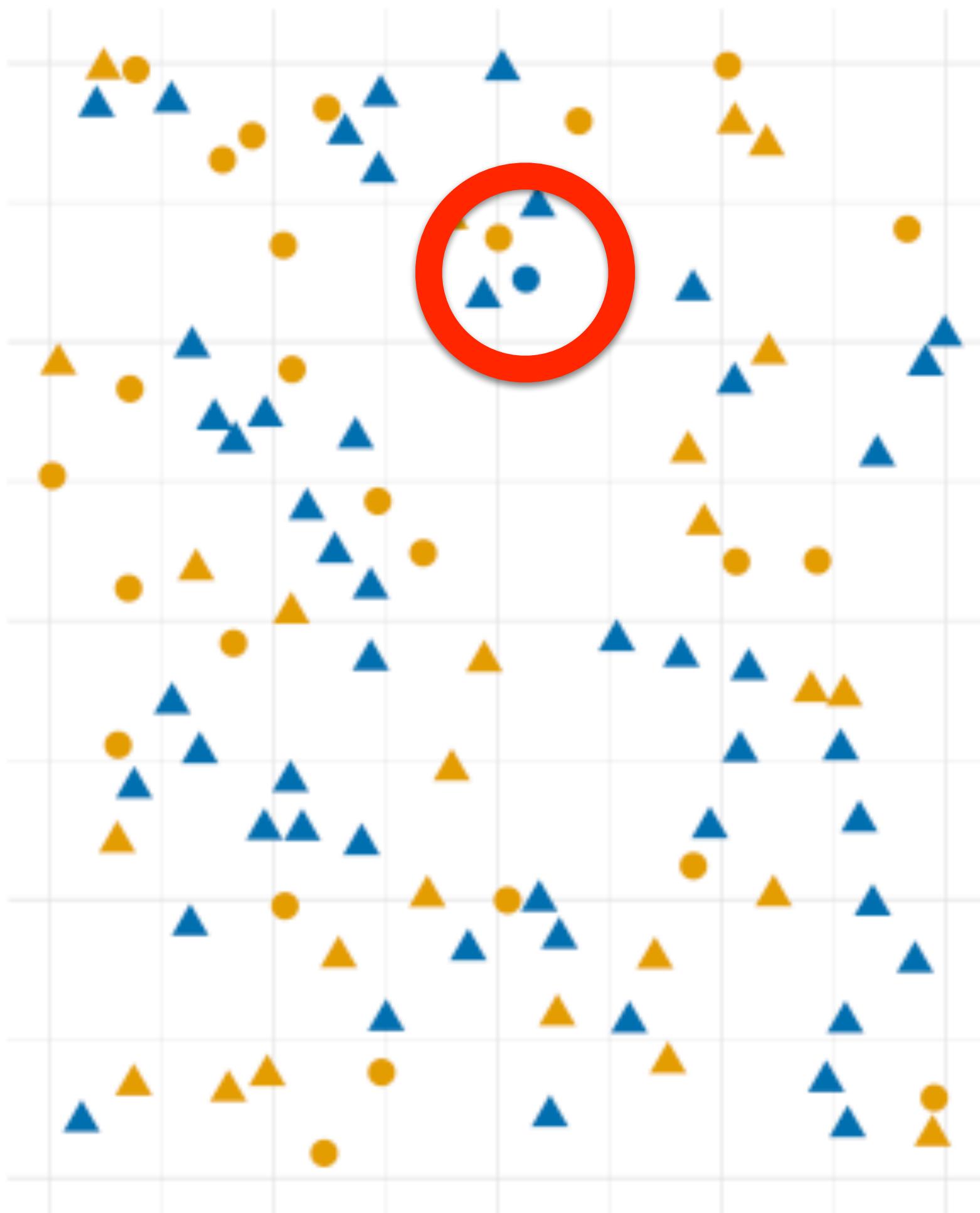
Shape Only, N=20

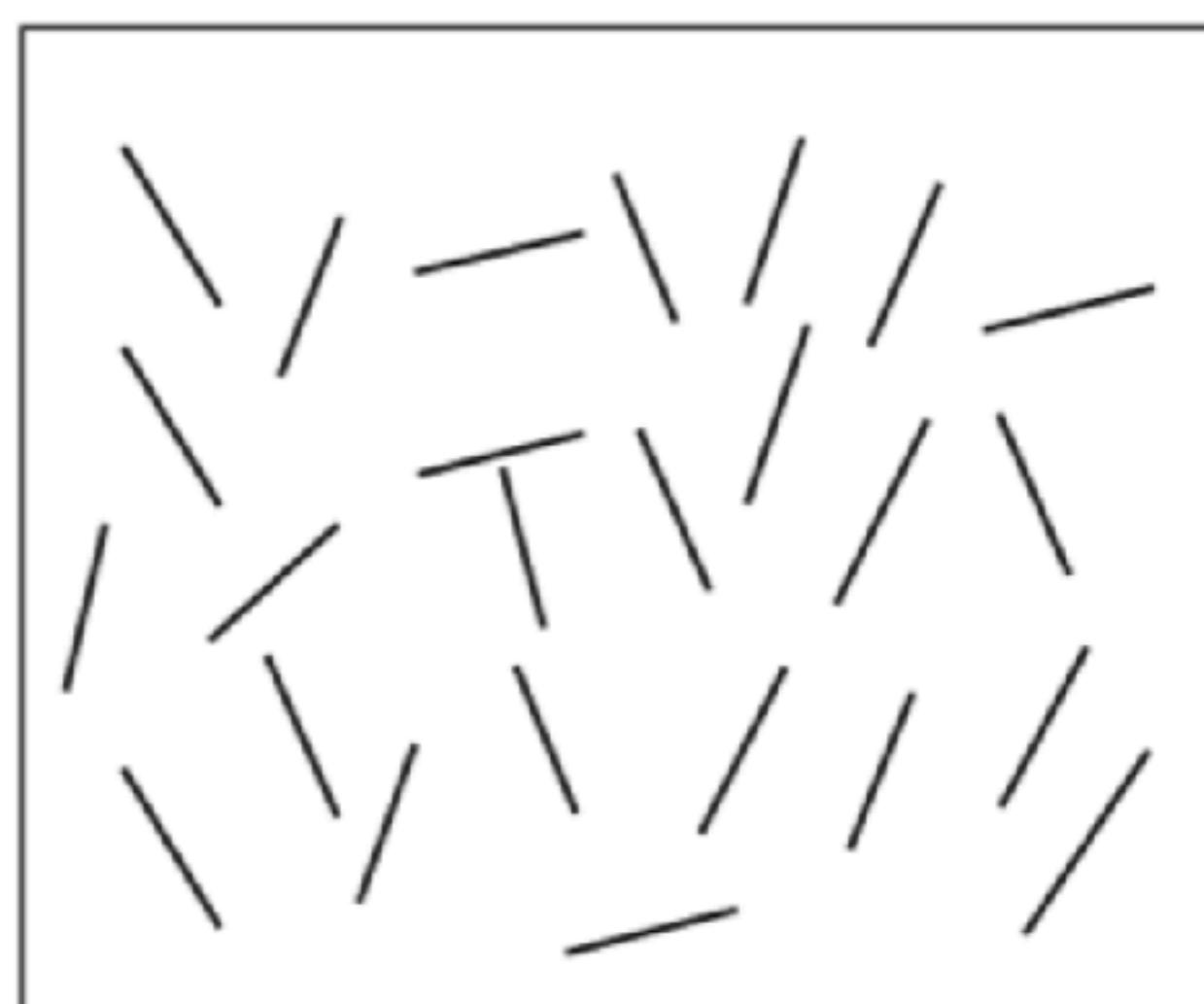
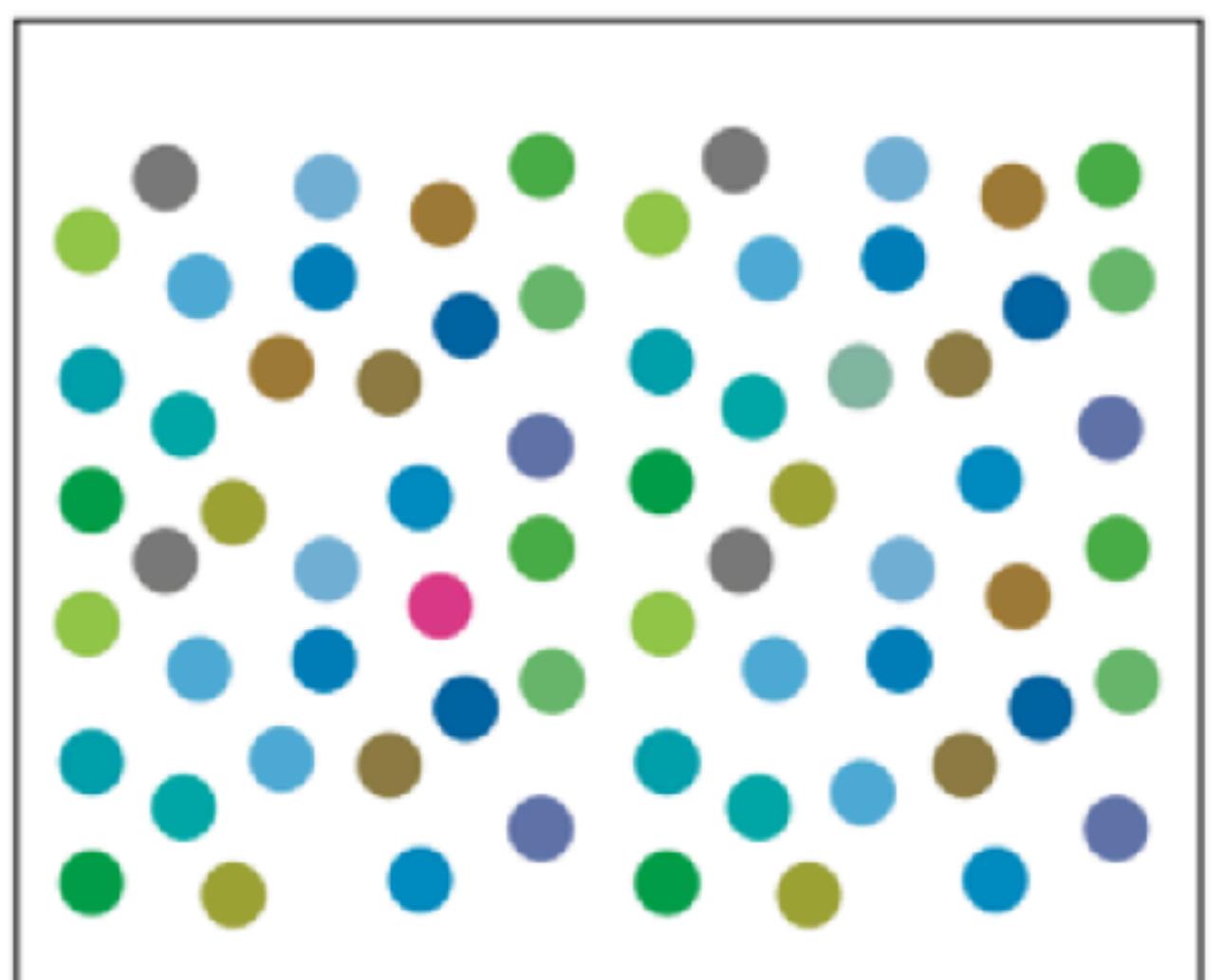
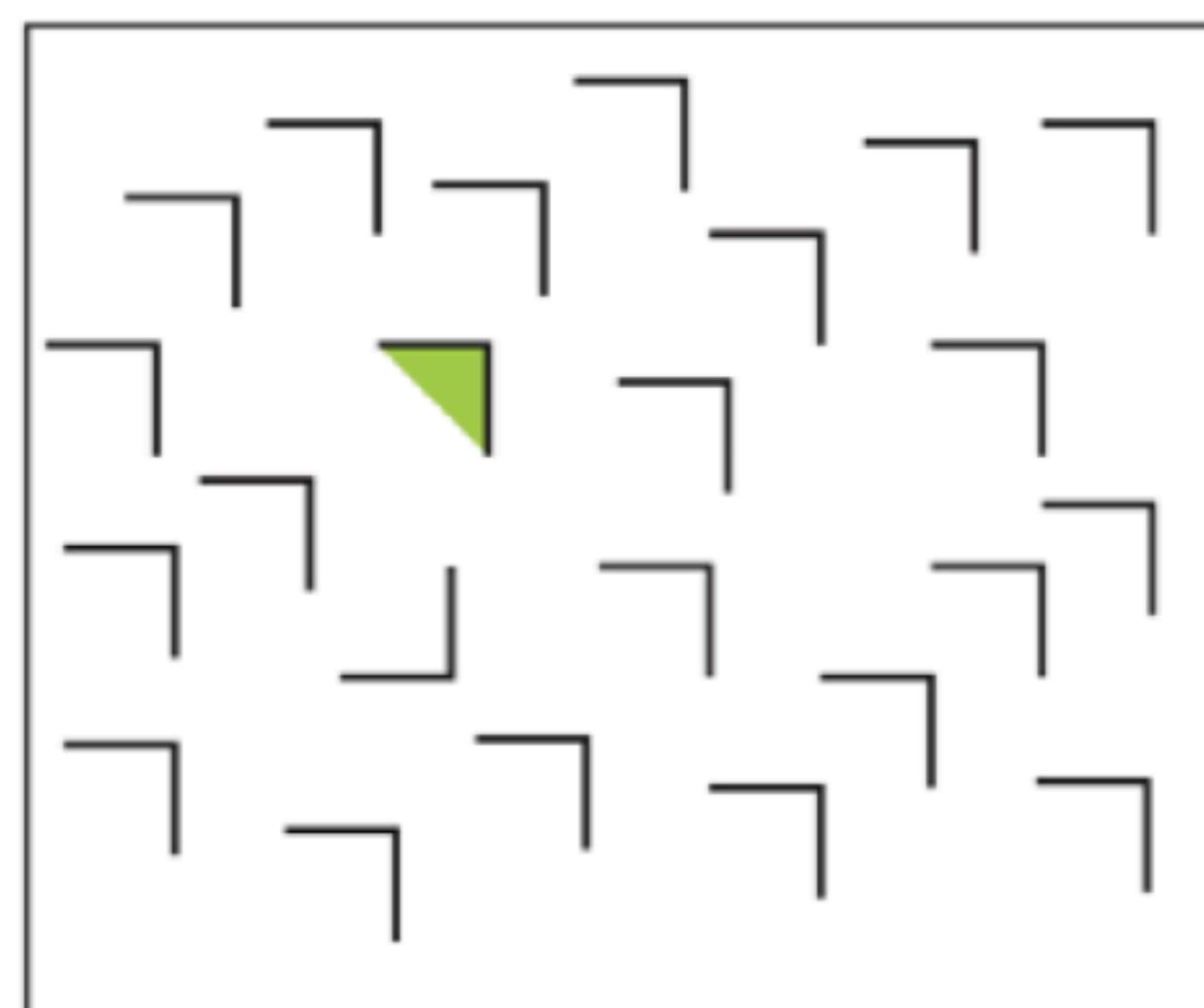
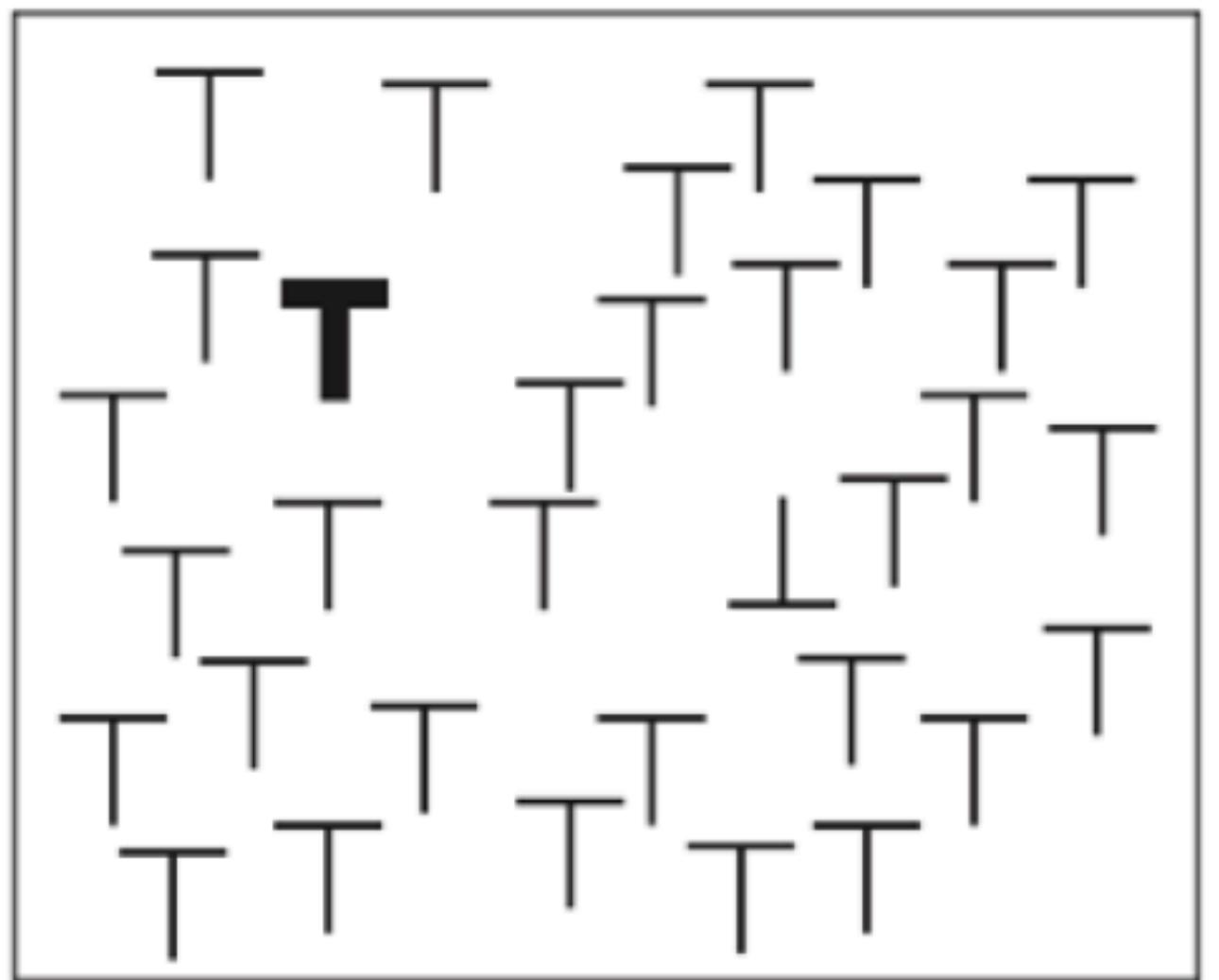


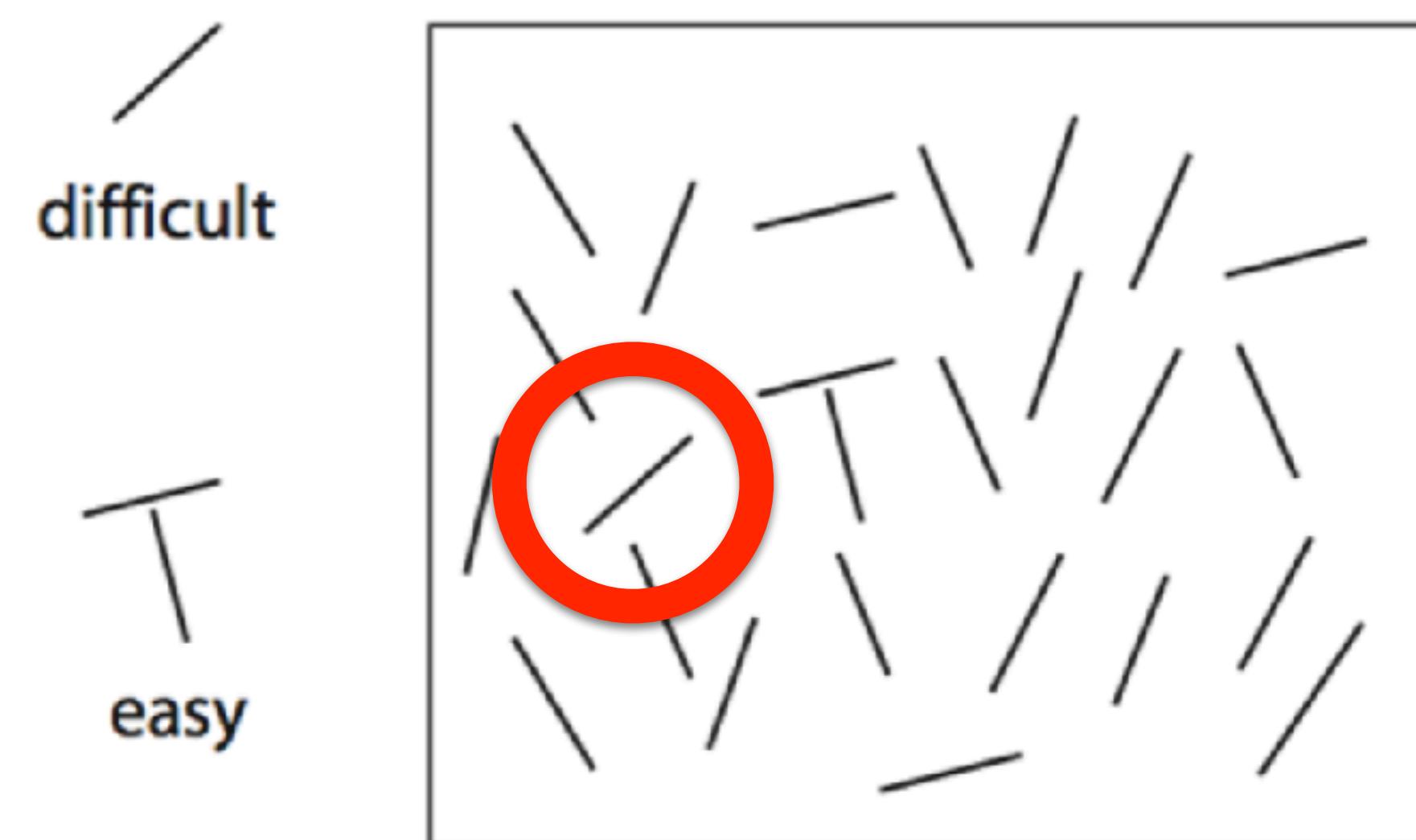
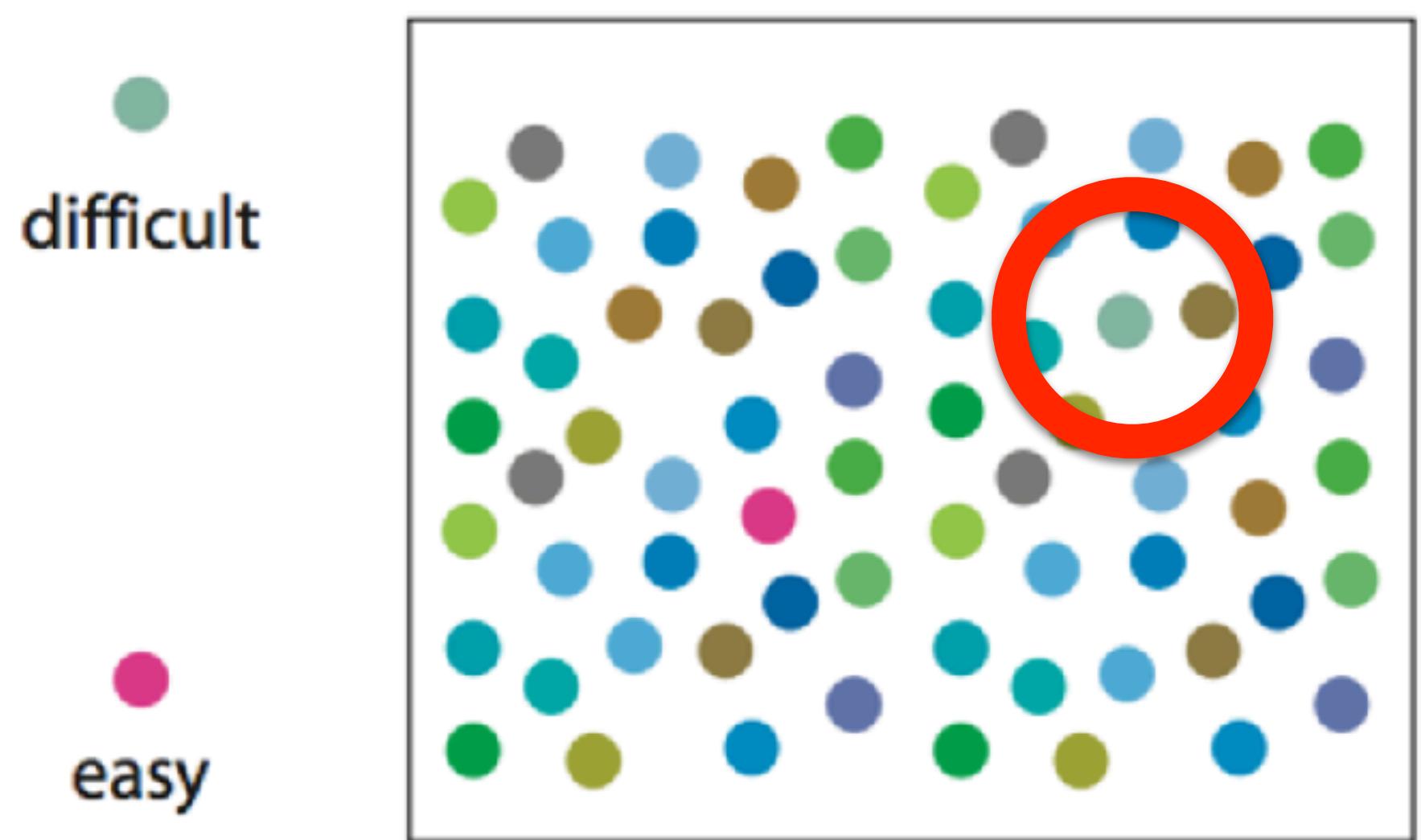
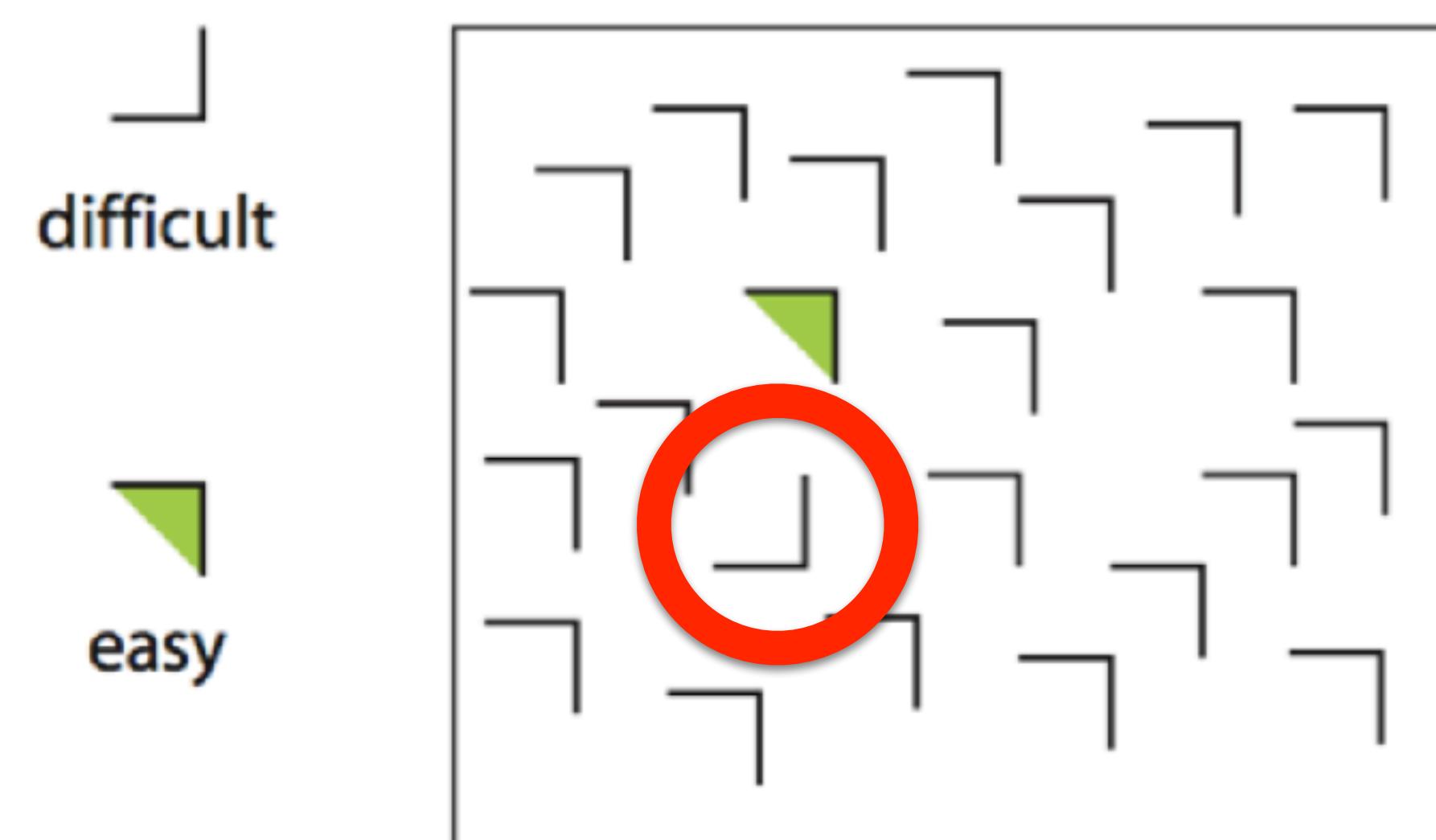
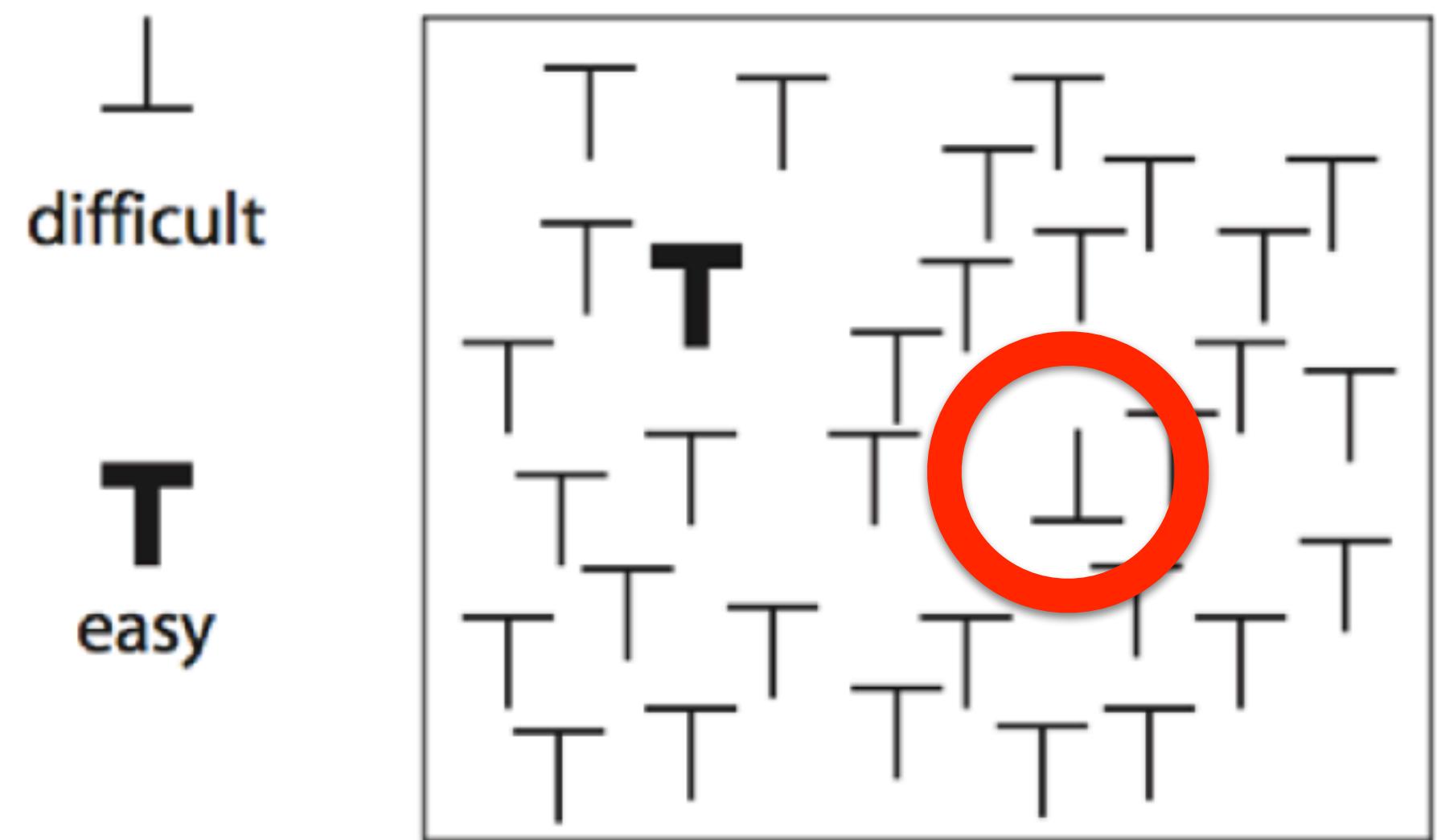
Shape Only, N=100

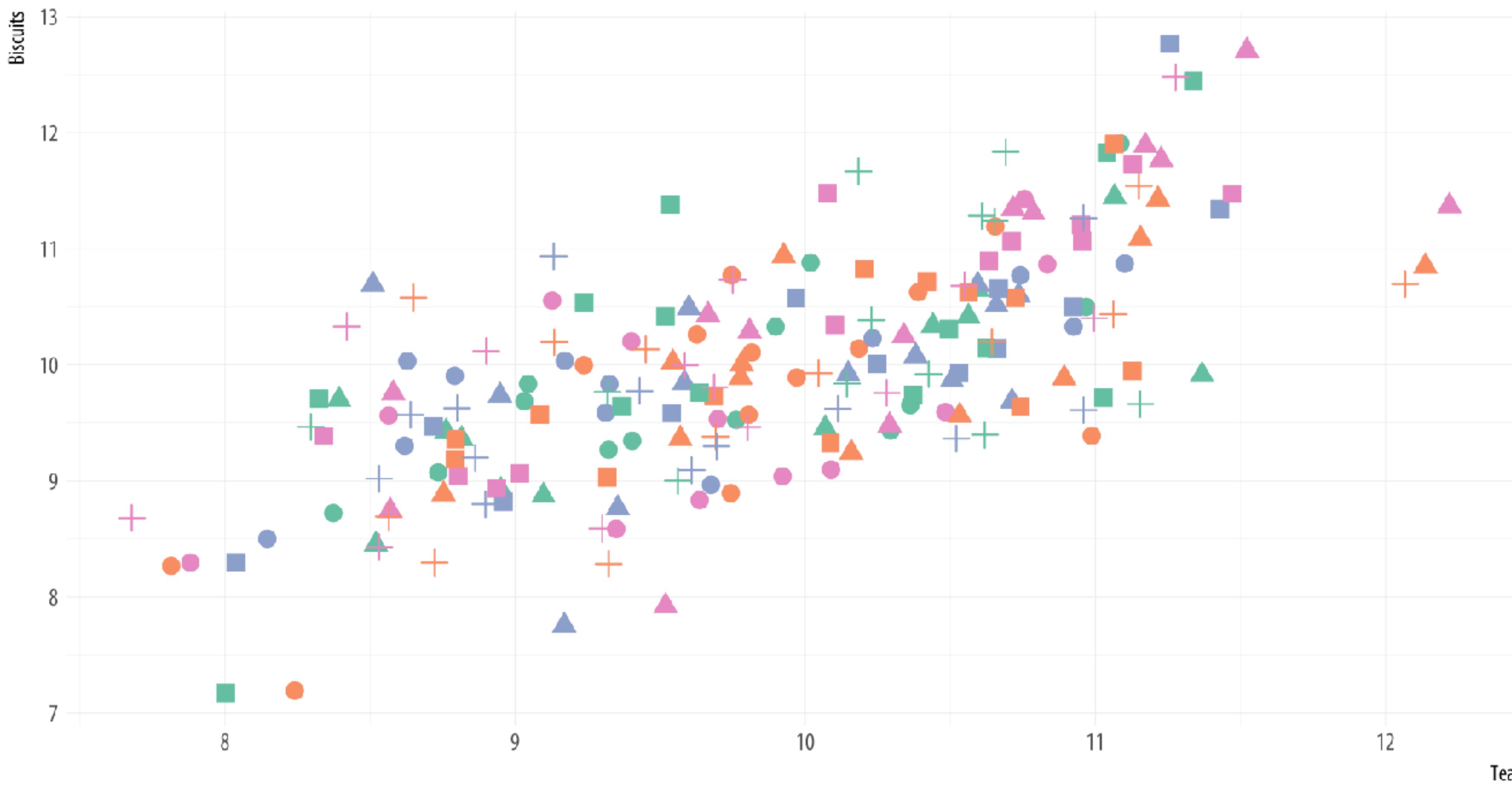


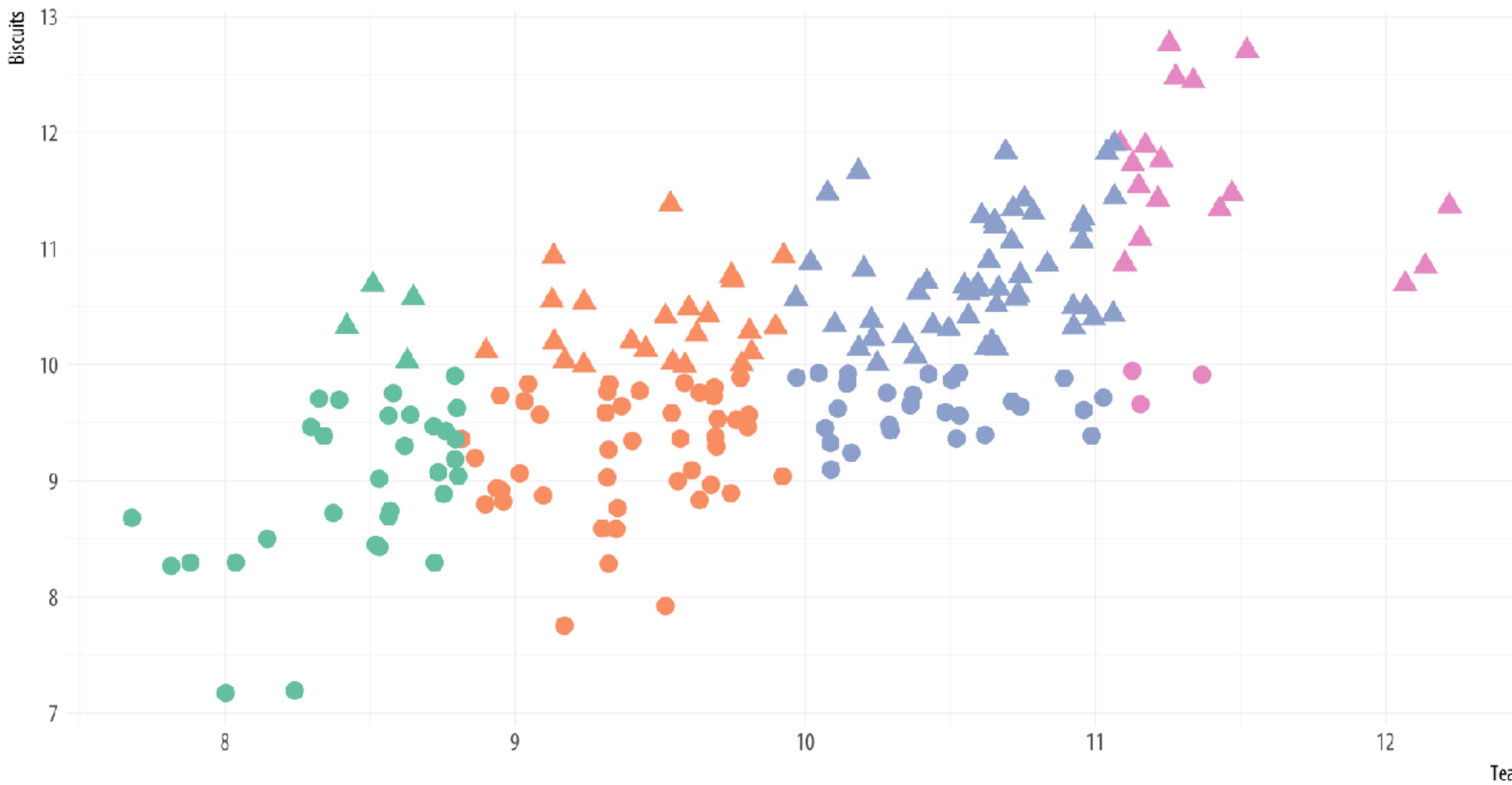
Color & Shape, N=100



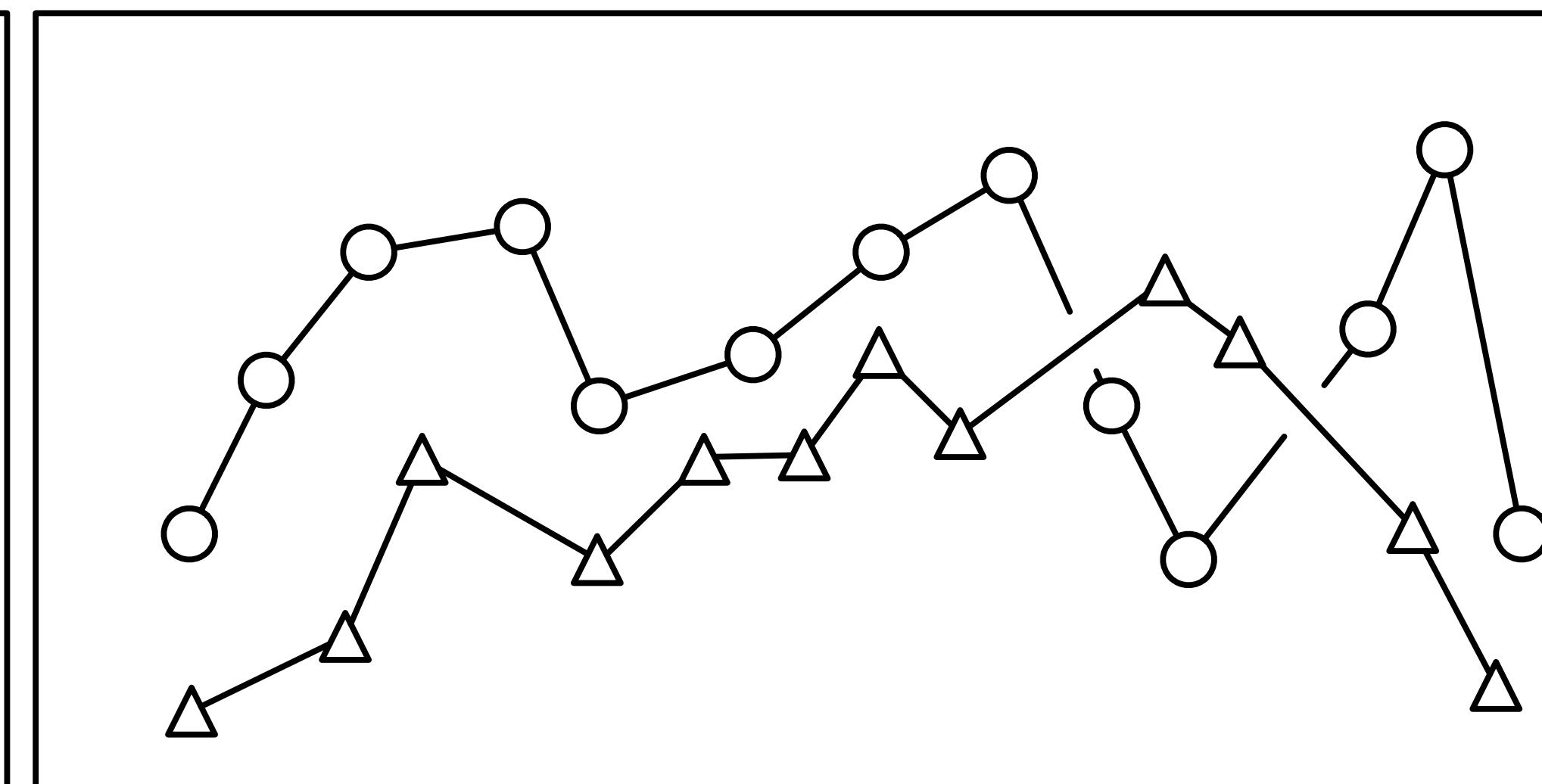
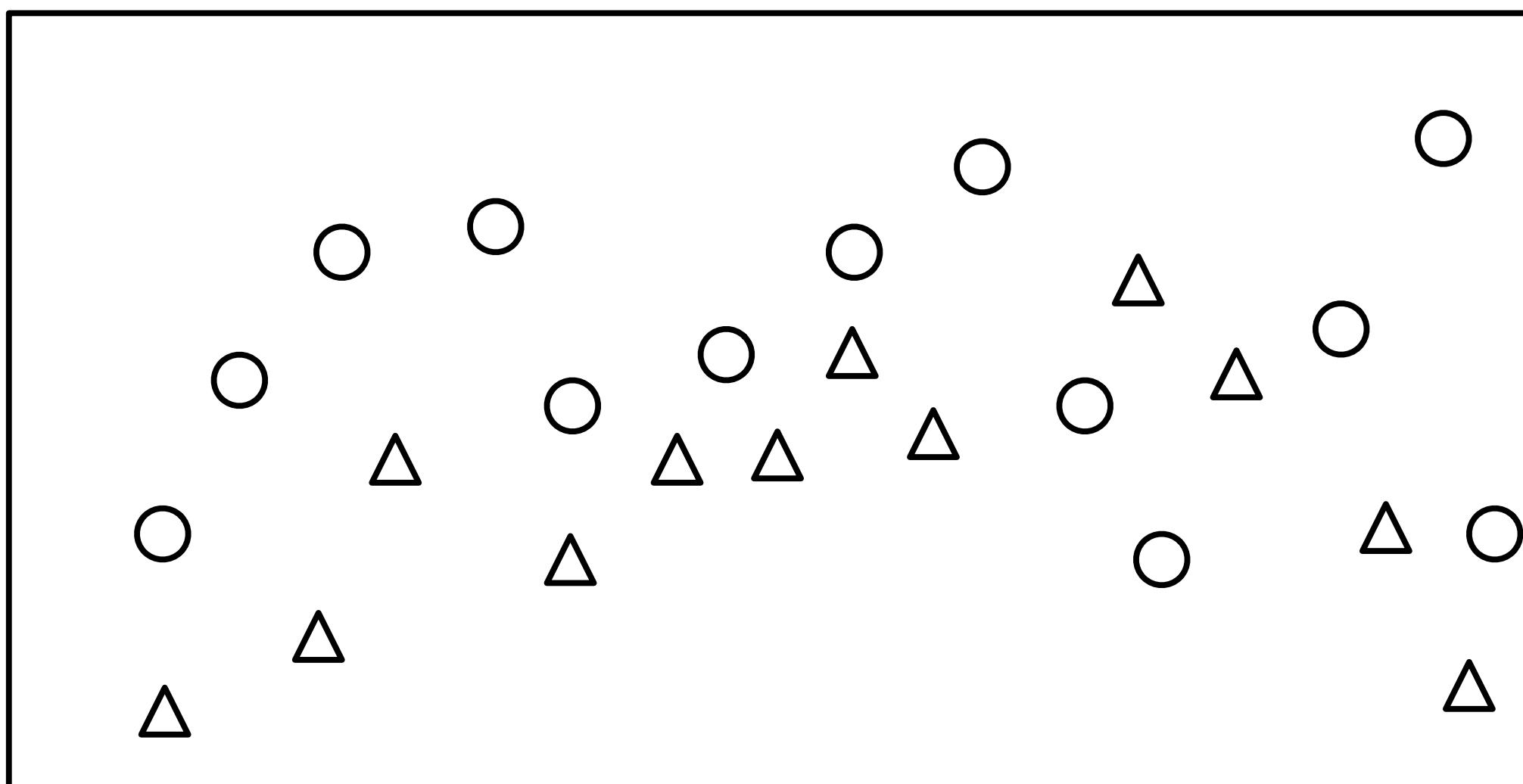
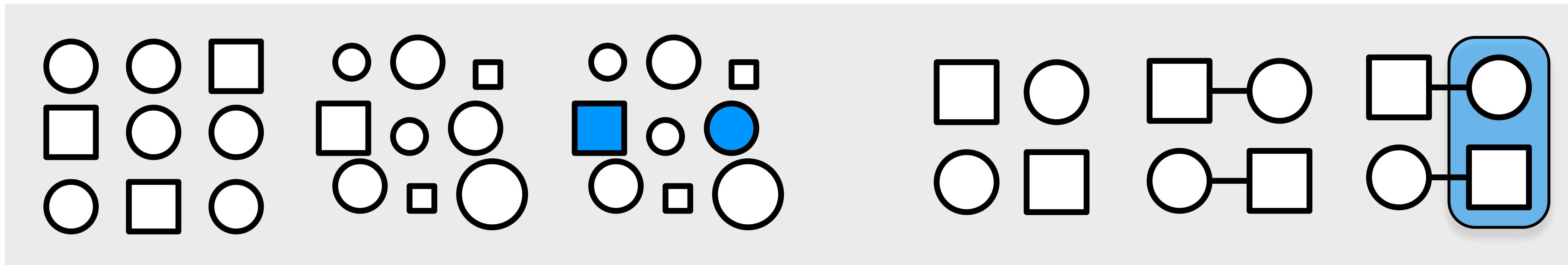
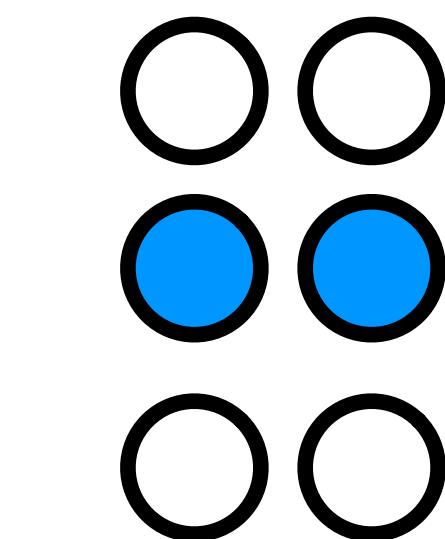
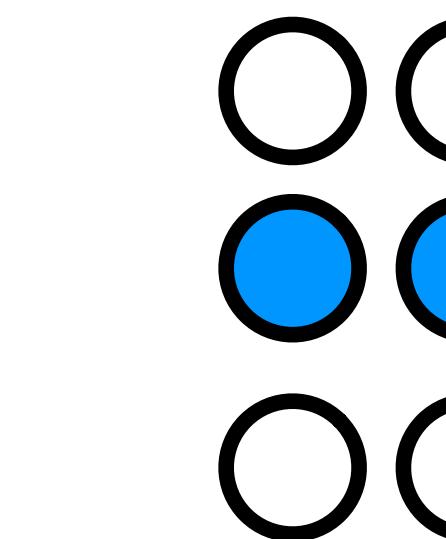
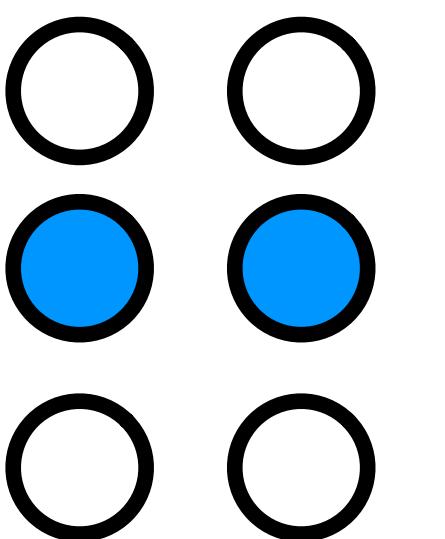
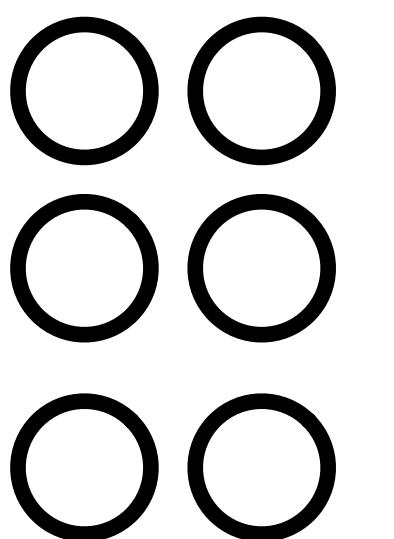
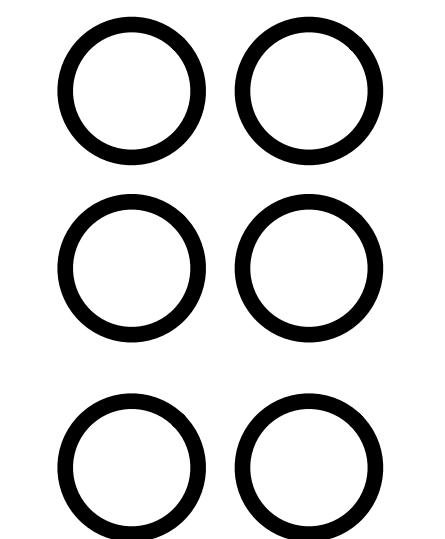
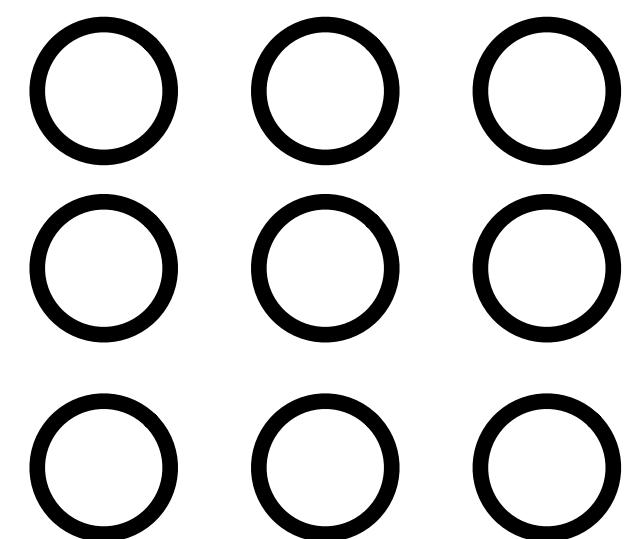




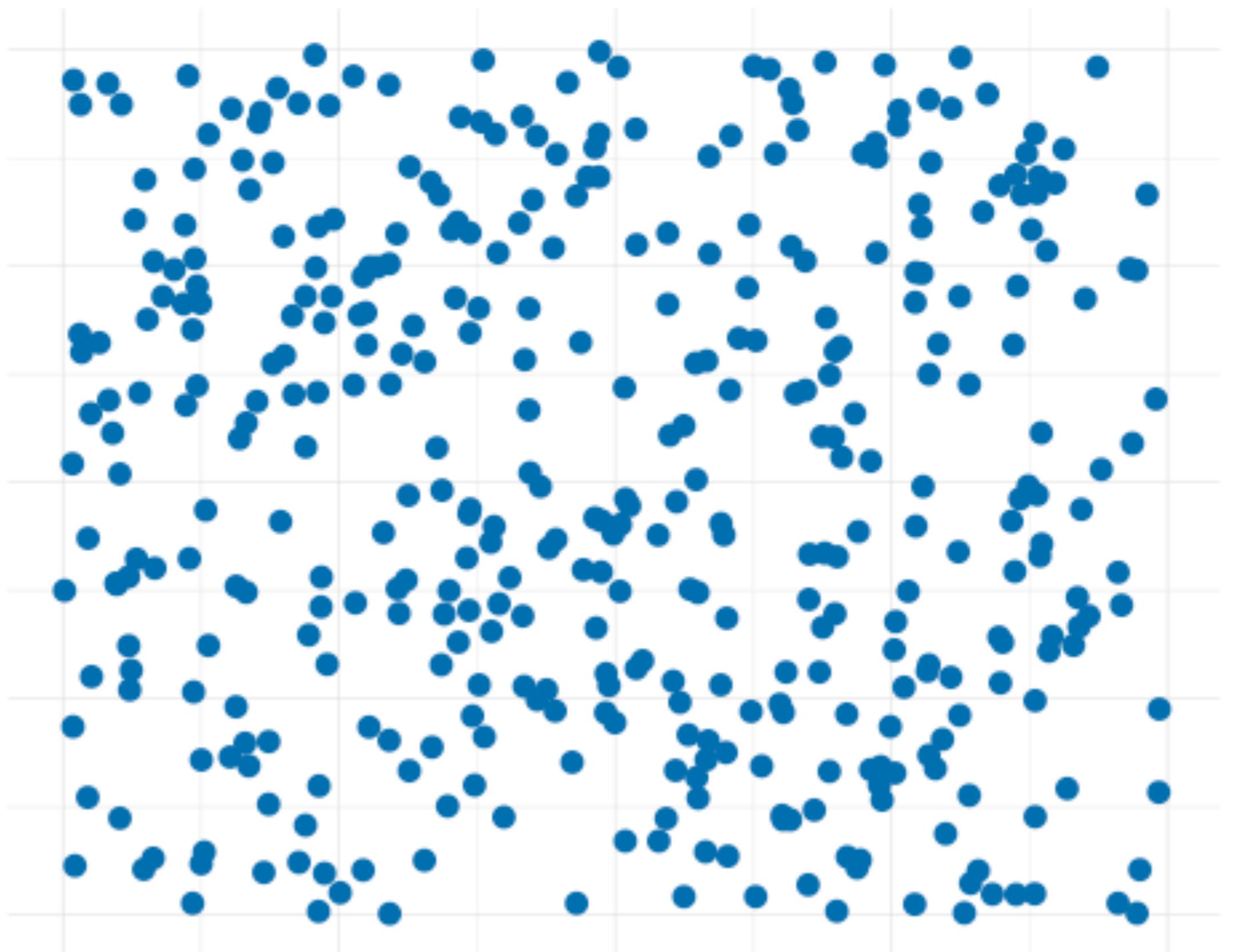




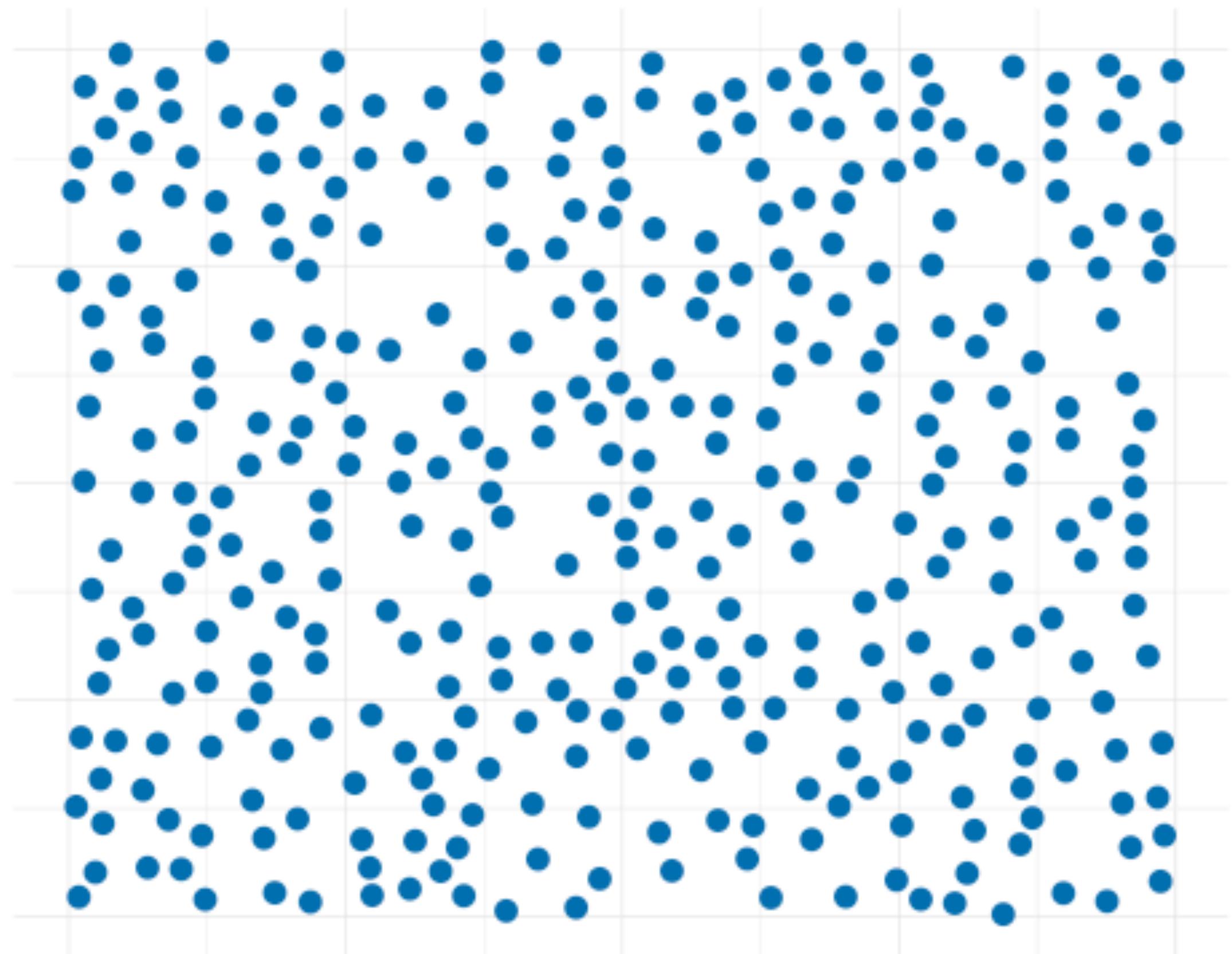
GESTALT INFERENCES



Poisson



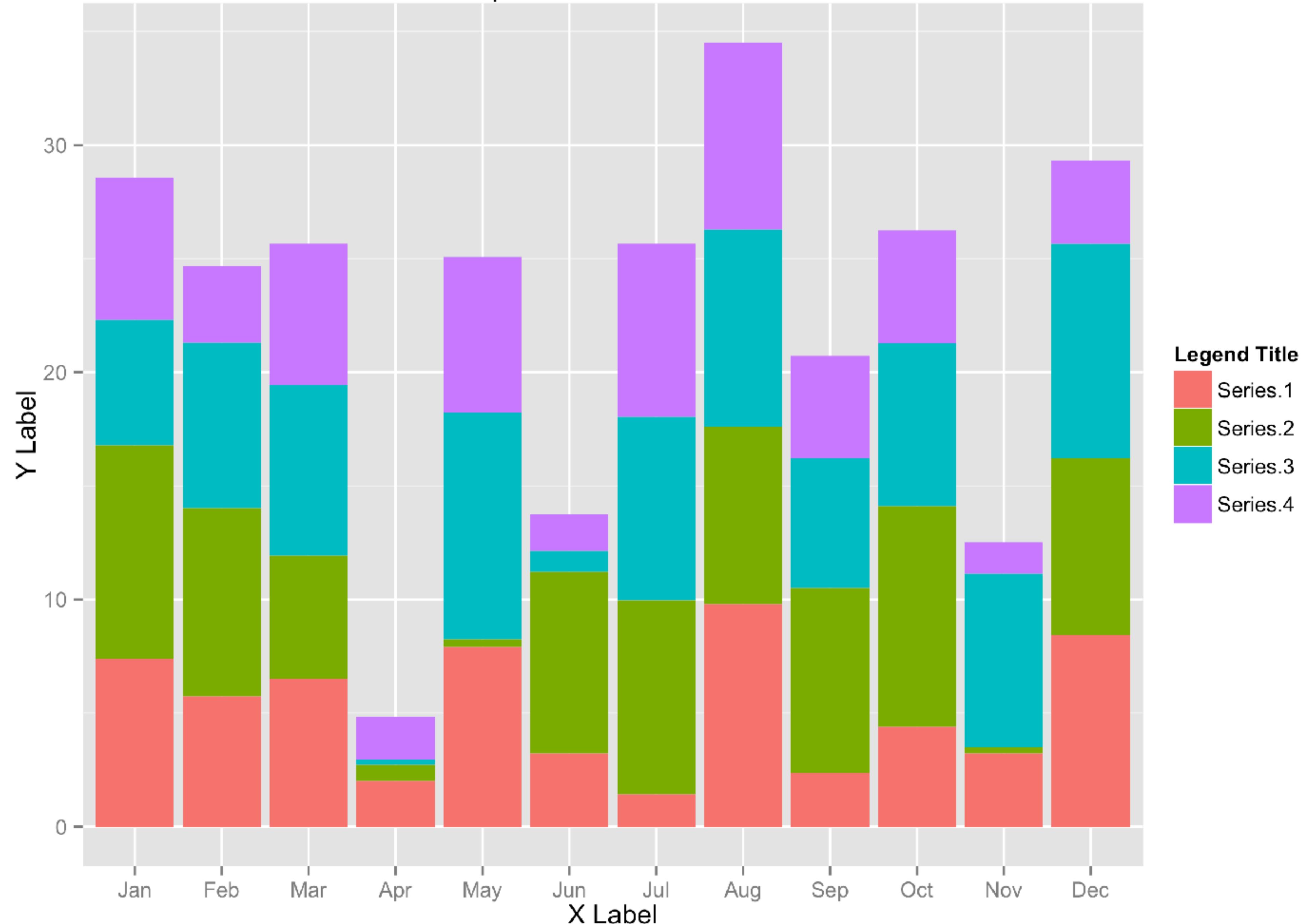
Matérn

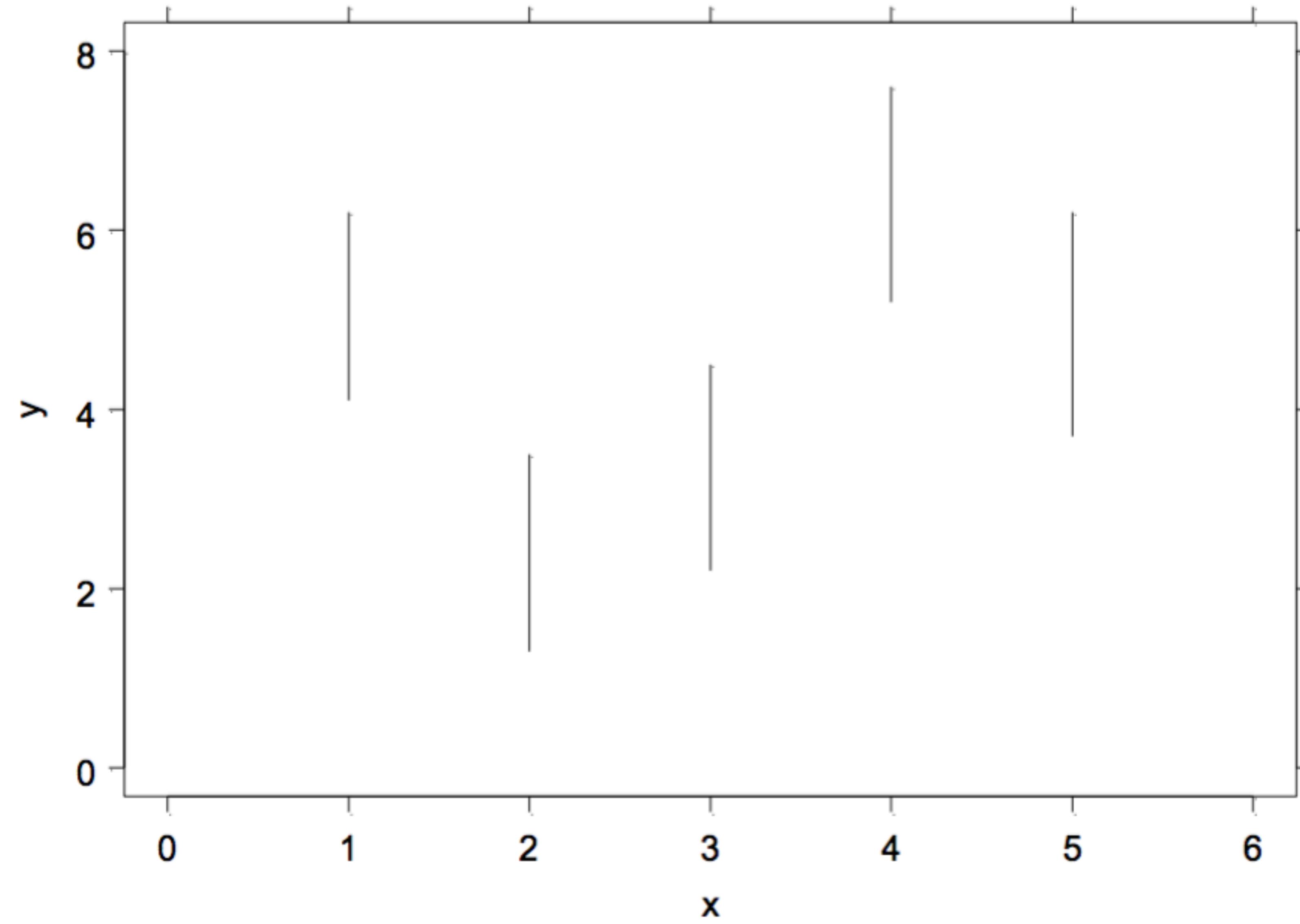


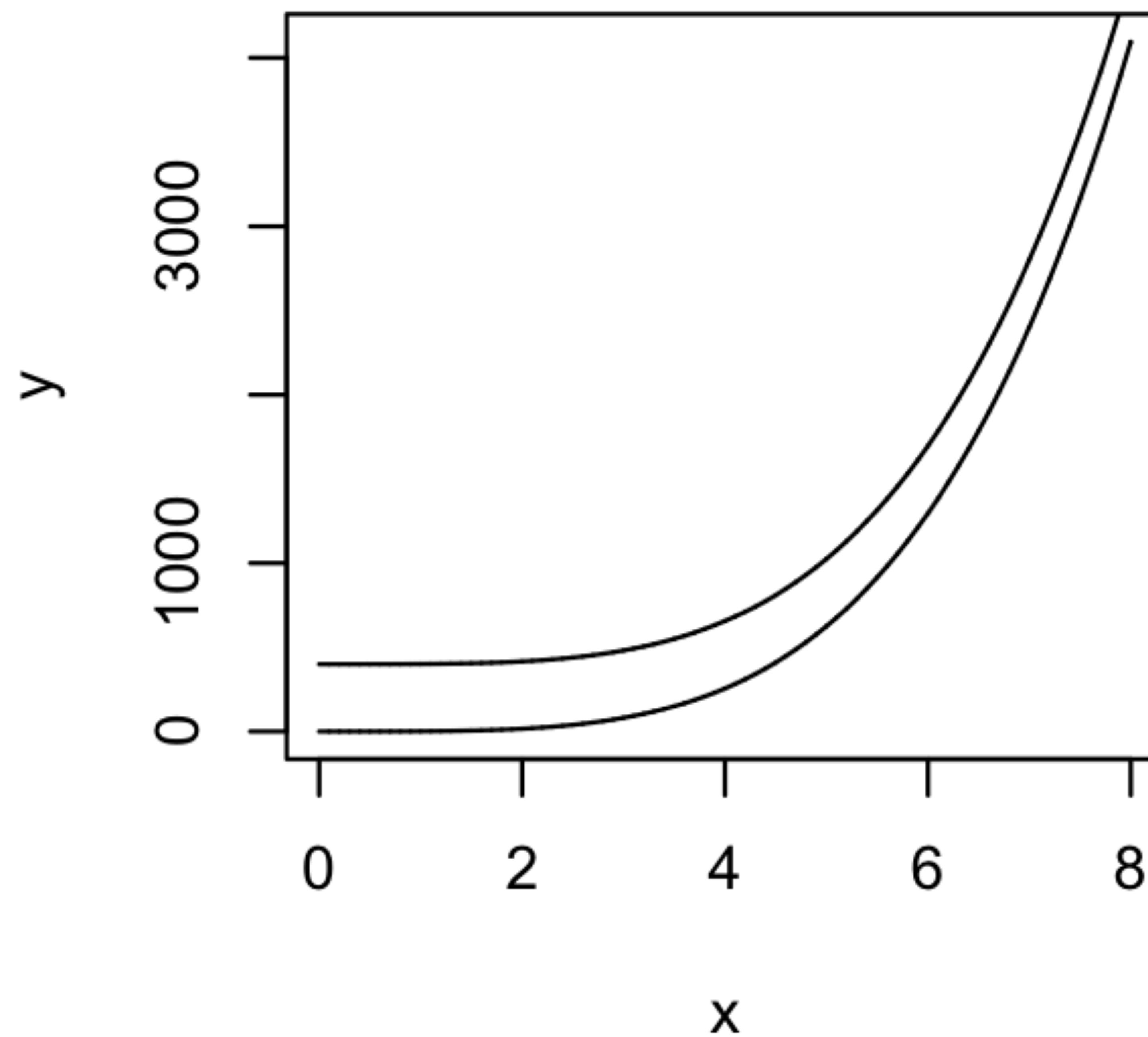
MAPPINGS FOR DATA

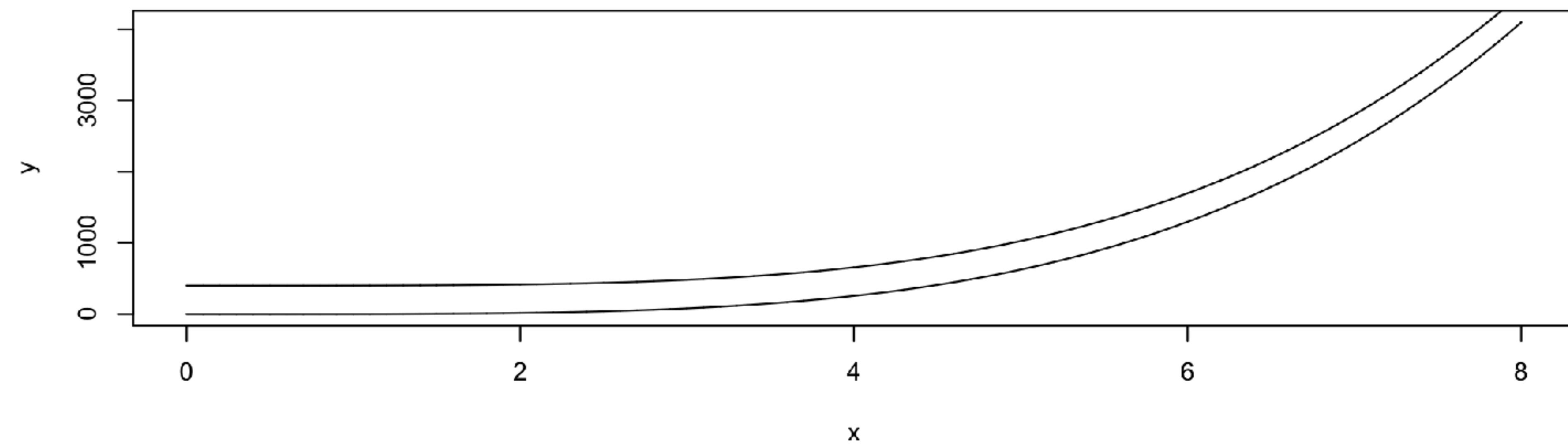
Visual Tasks in Decoding Graphs

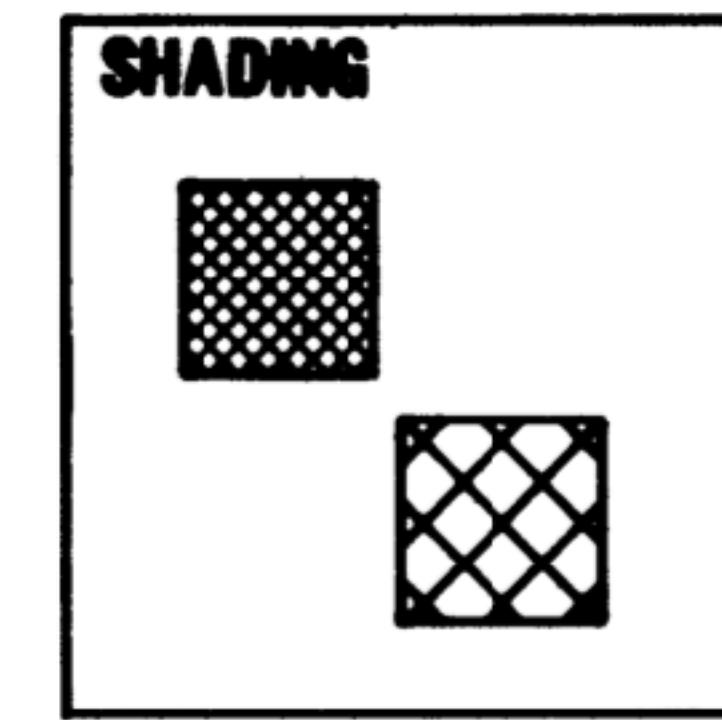
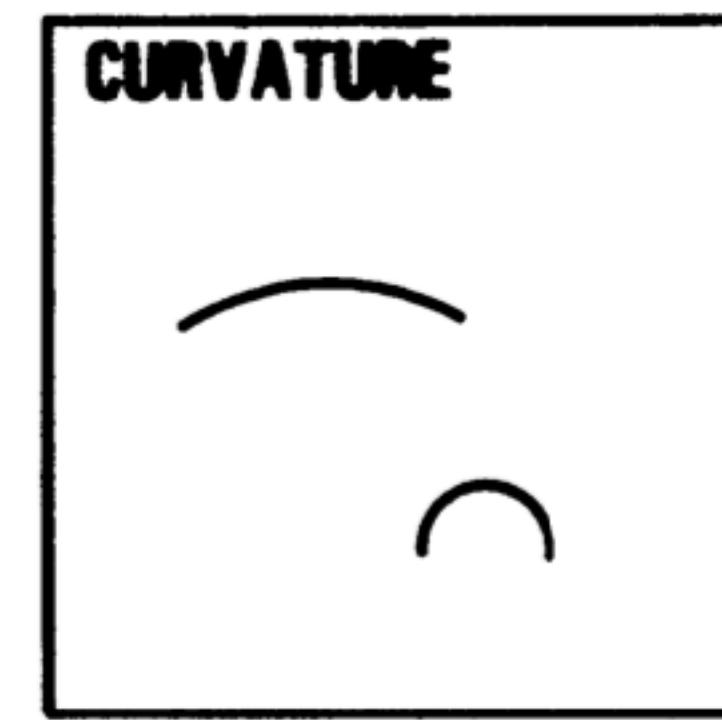
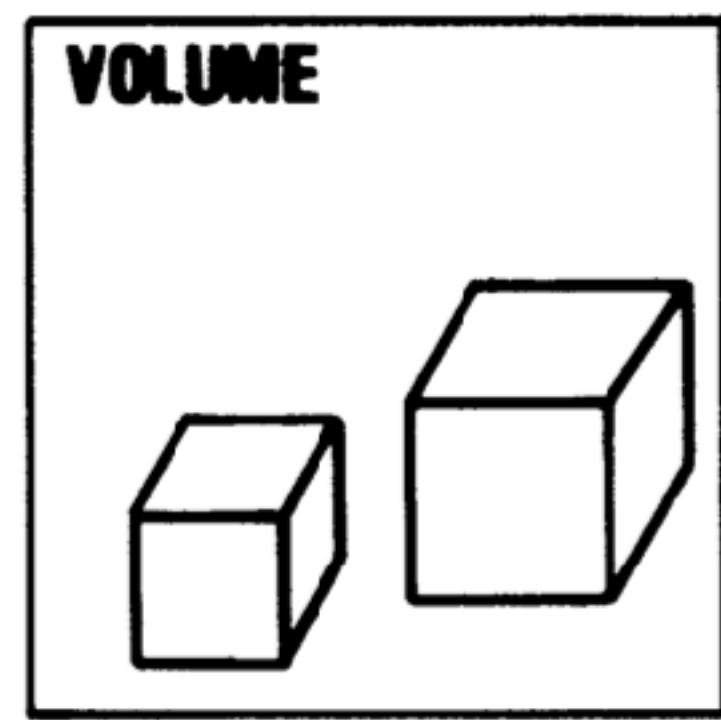
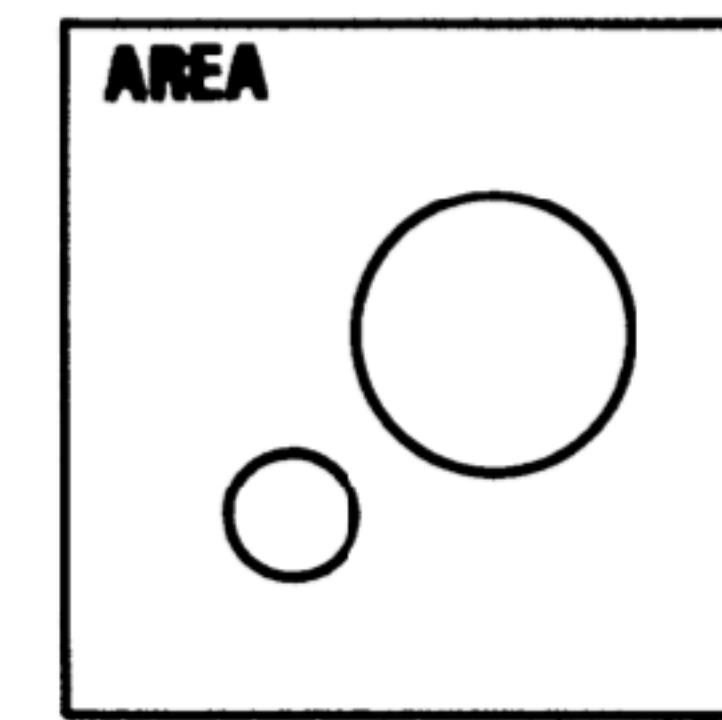
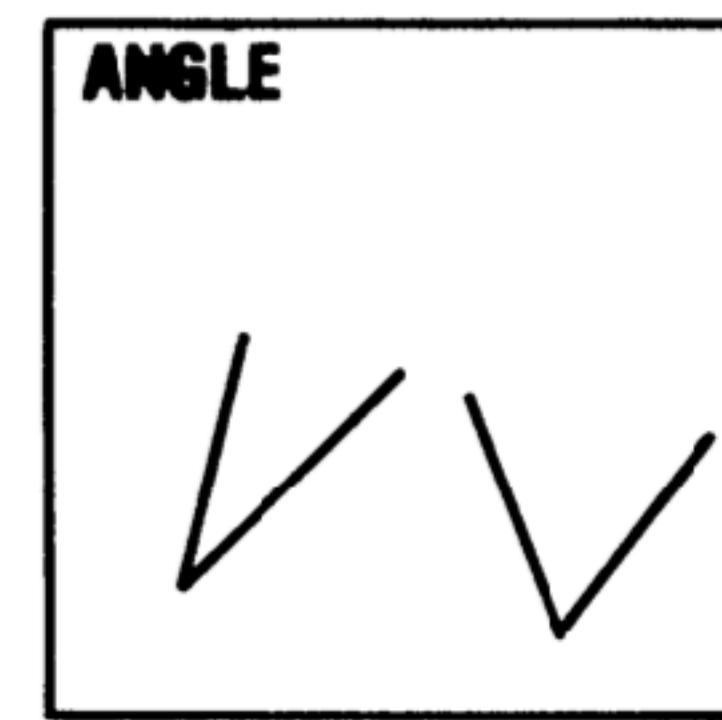
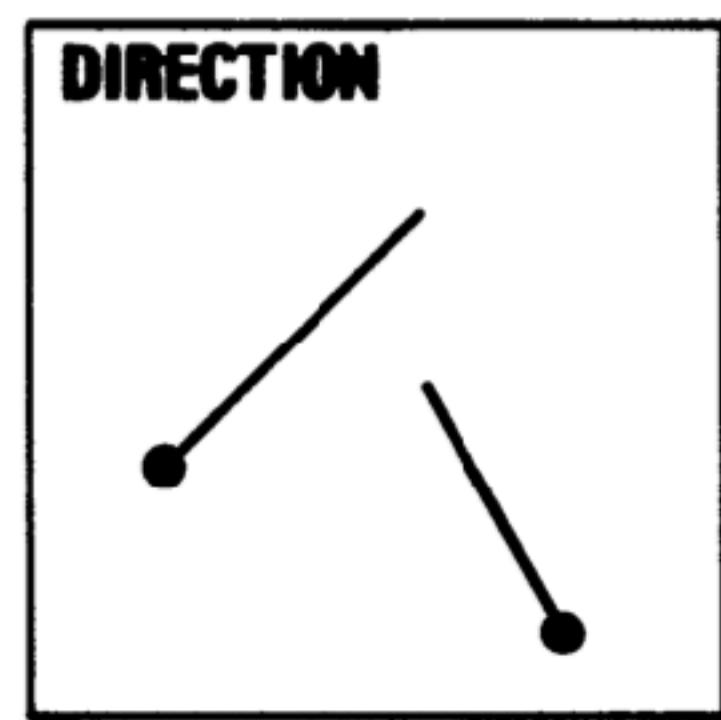
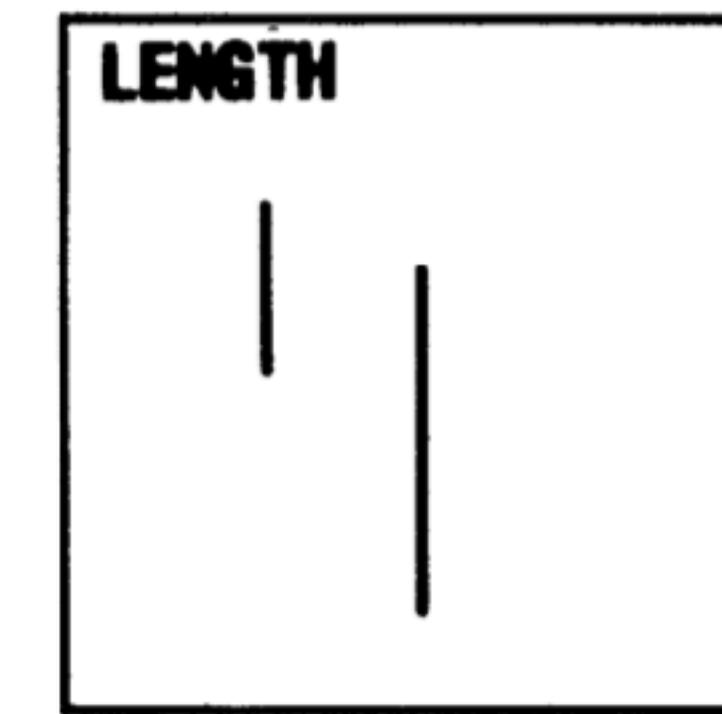
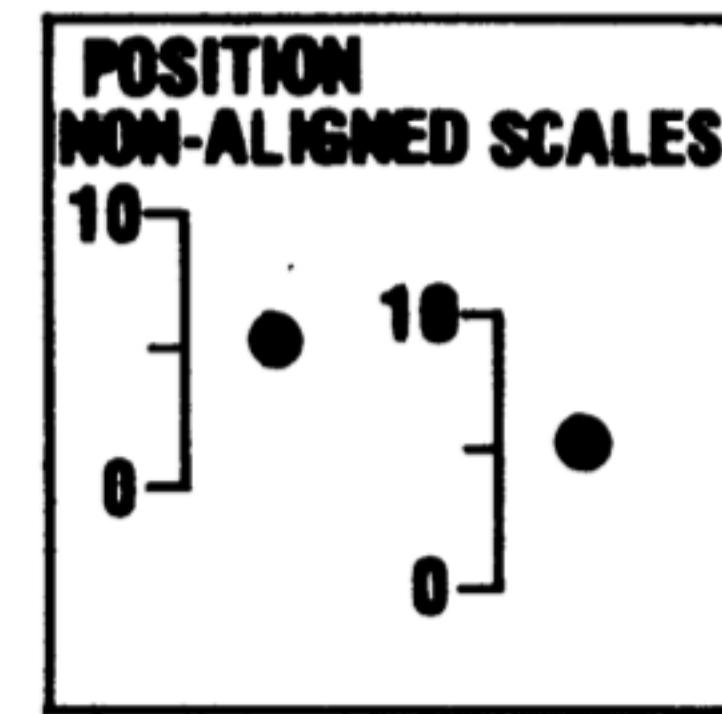
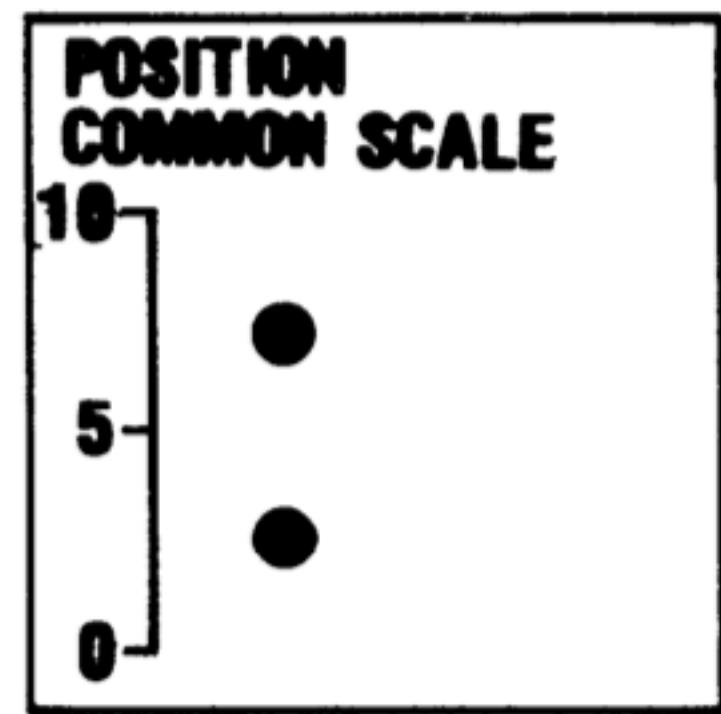
An Example Stacked Column Chart











COLOR SATURATION

Figure 1. Elementary perceptual tasks.

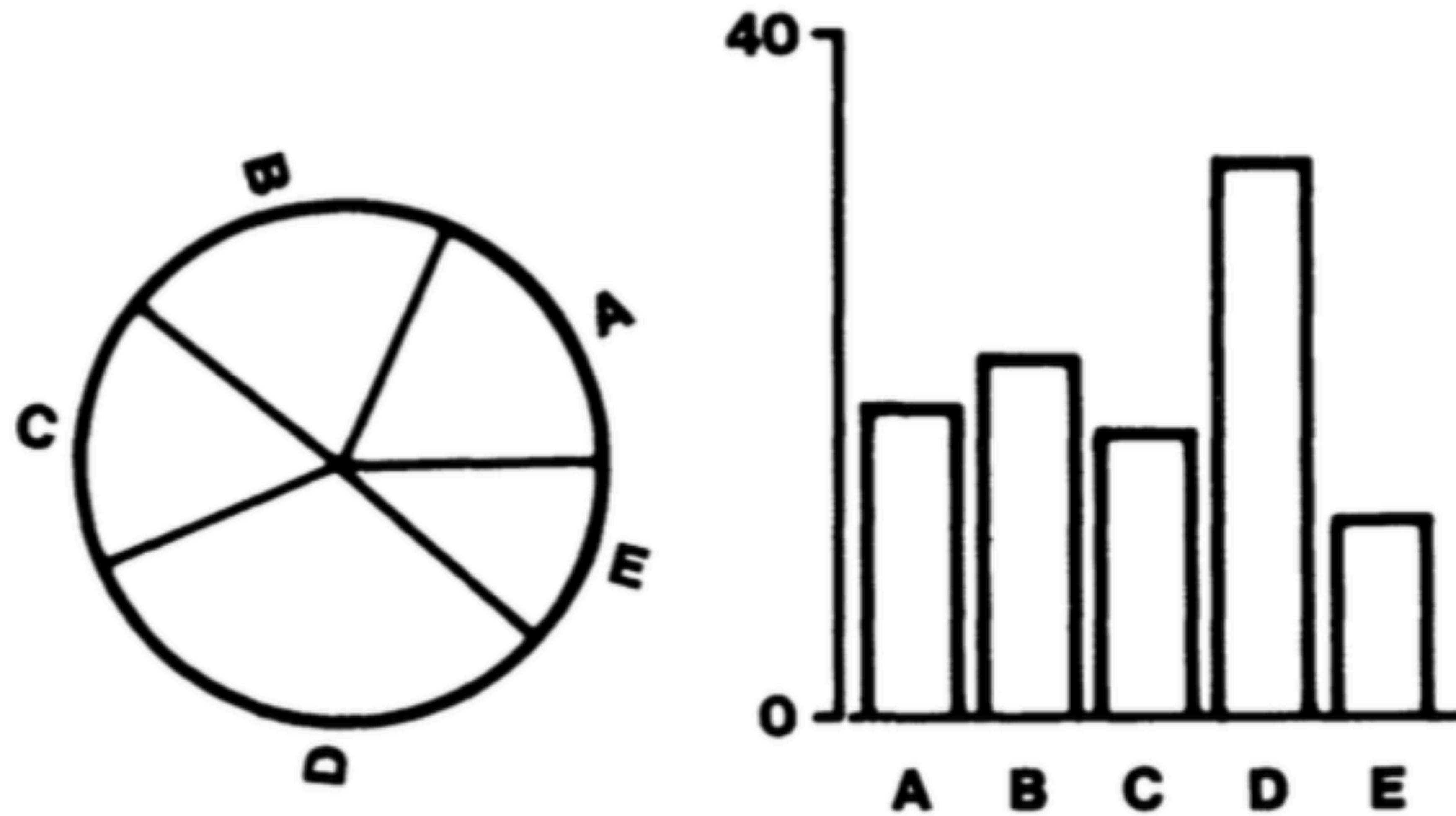


Figure 3. Graphs from position-angle experiment.

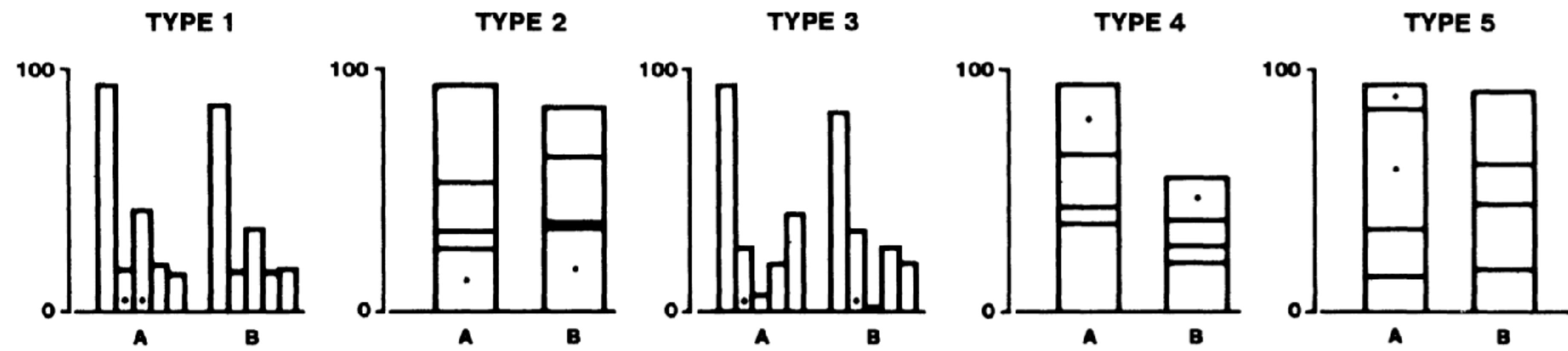
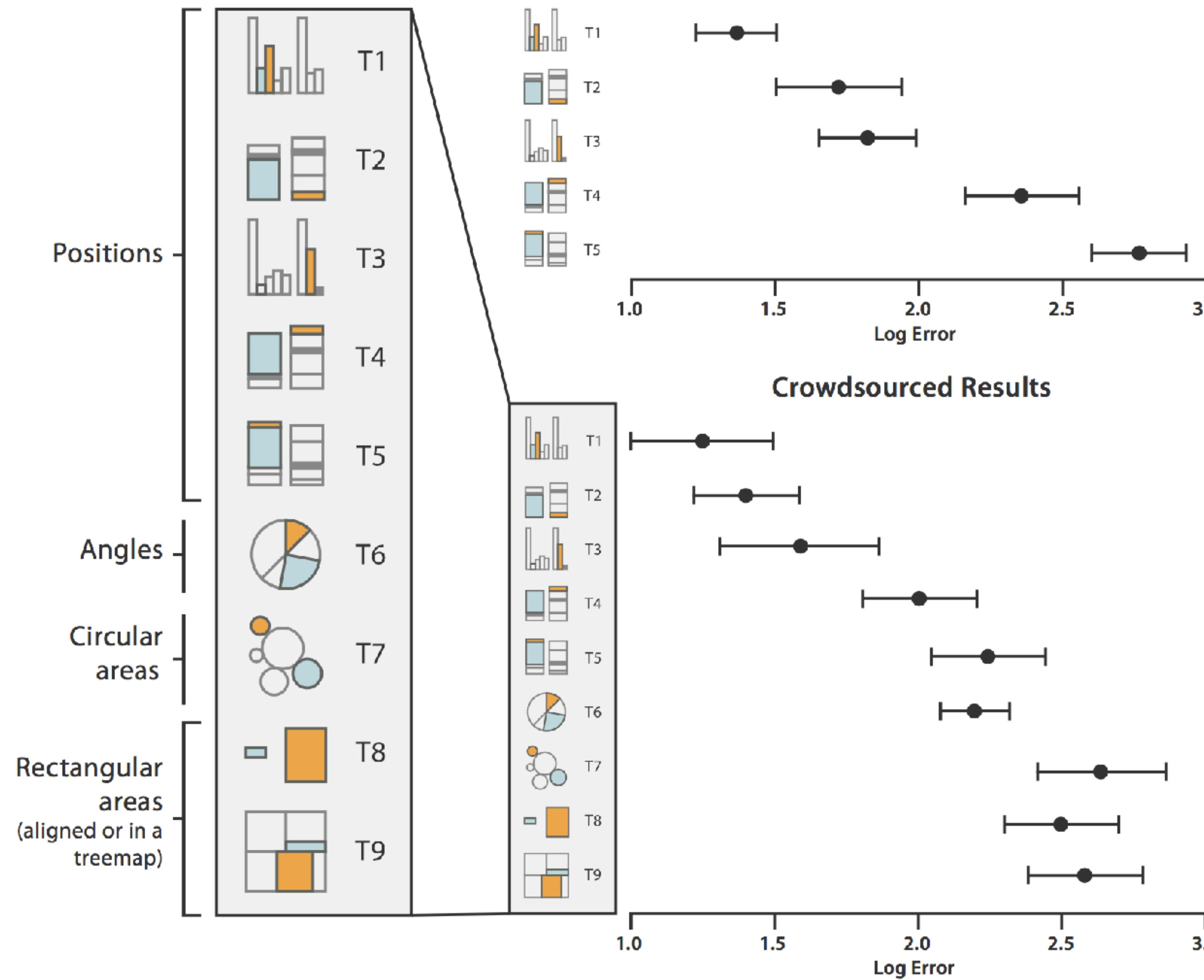


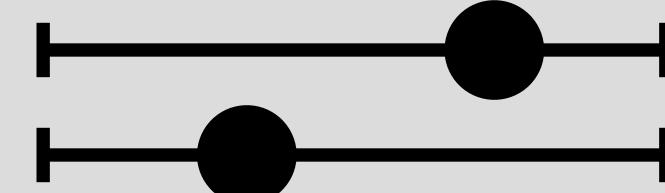
Figure 4. Graphs from position-length experiment.

Cleveland & McGill's Results



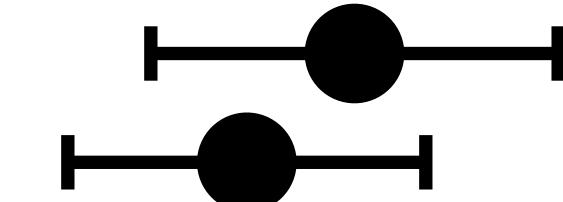
A Rough Hierarchy of Mappings

Position on a common scale



Better

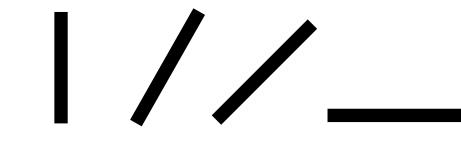
Position on unaligned scale



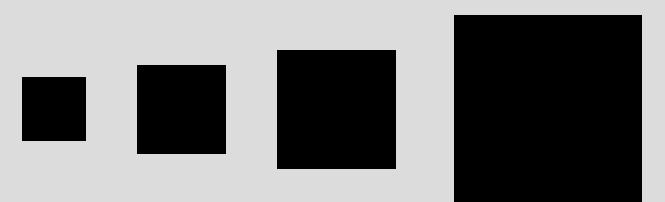
Length (1D as size)



Tilt or Angle



Area (2D as size)



Effectiveness

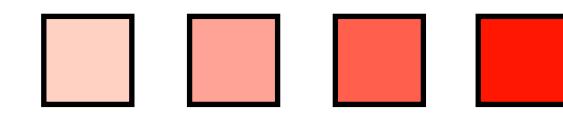
Depth (3D as Position)



Color luminance [brightness]



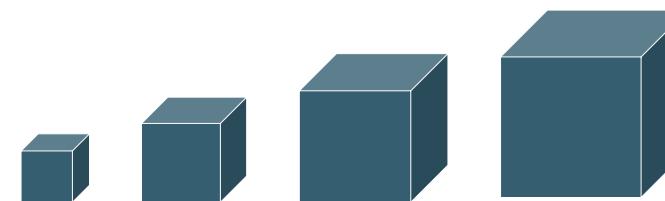
Color saturation [intensity]



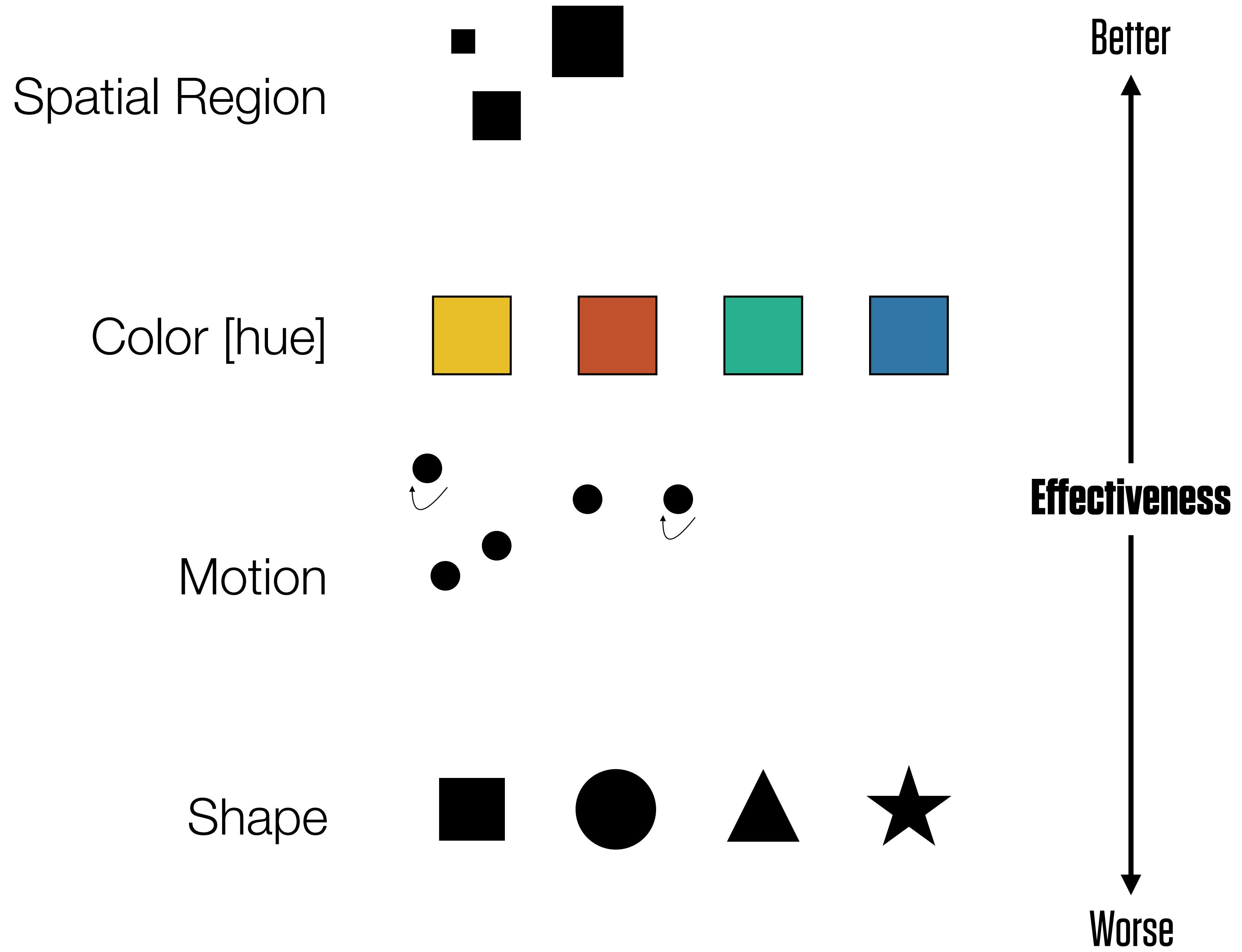
Curvature



Volume (3D as size)



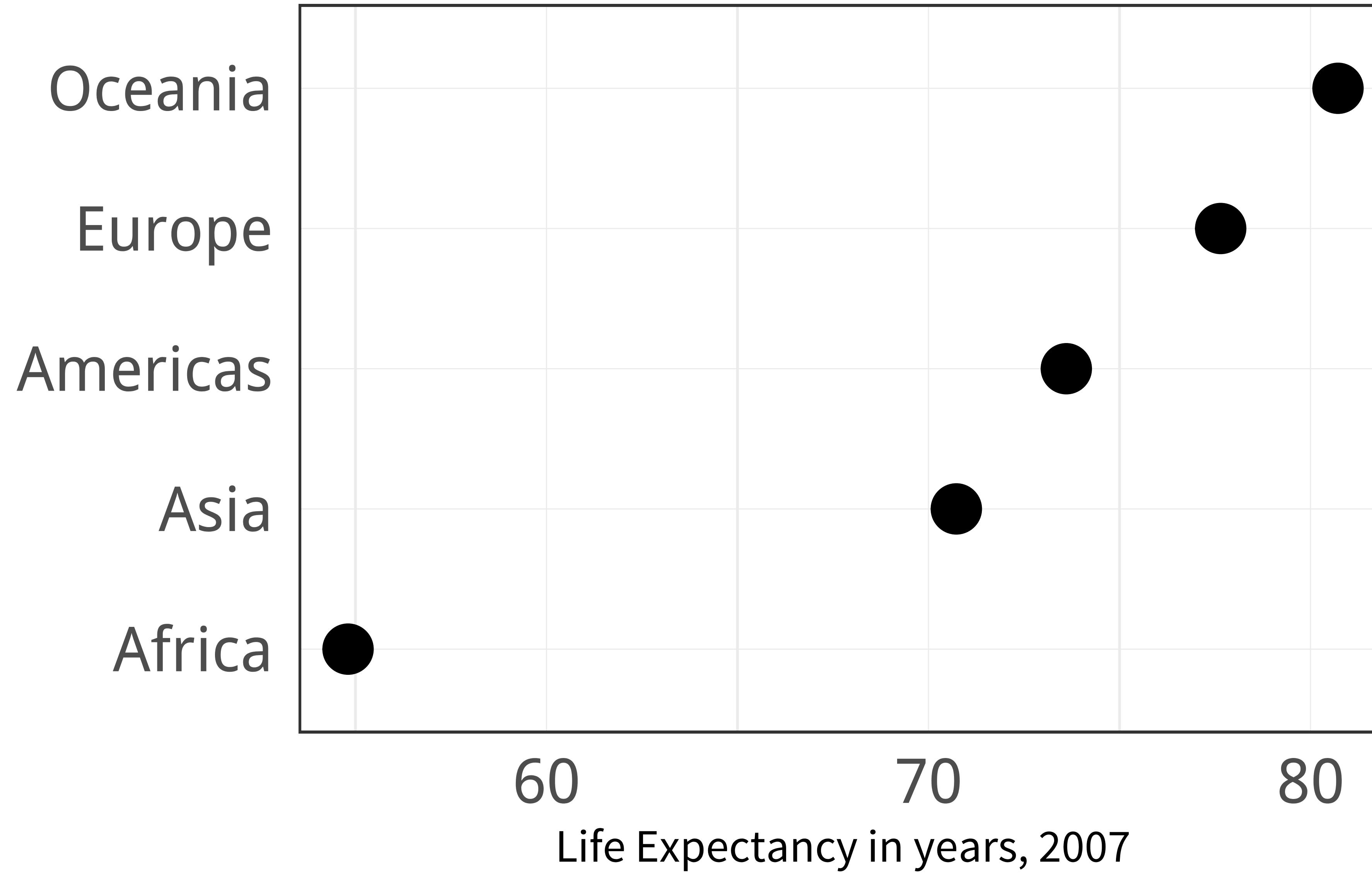
Worse

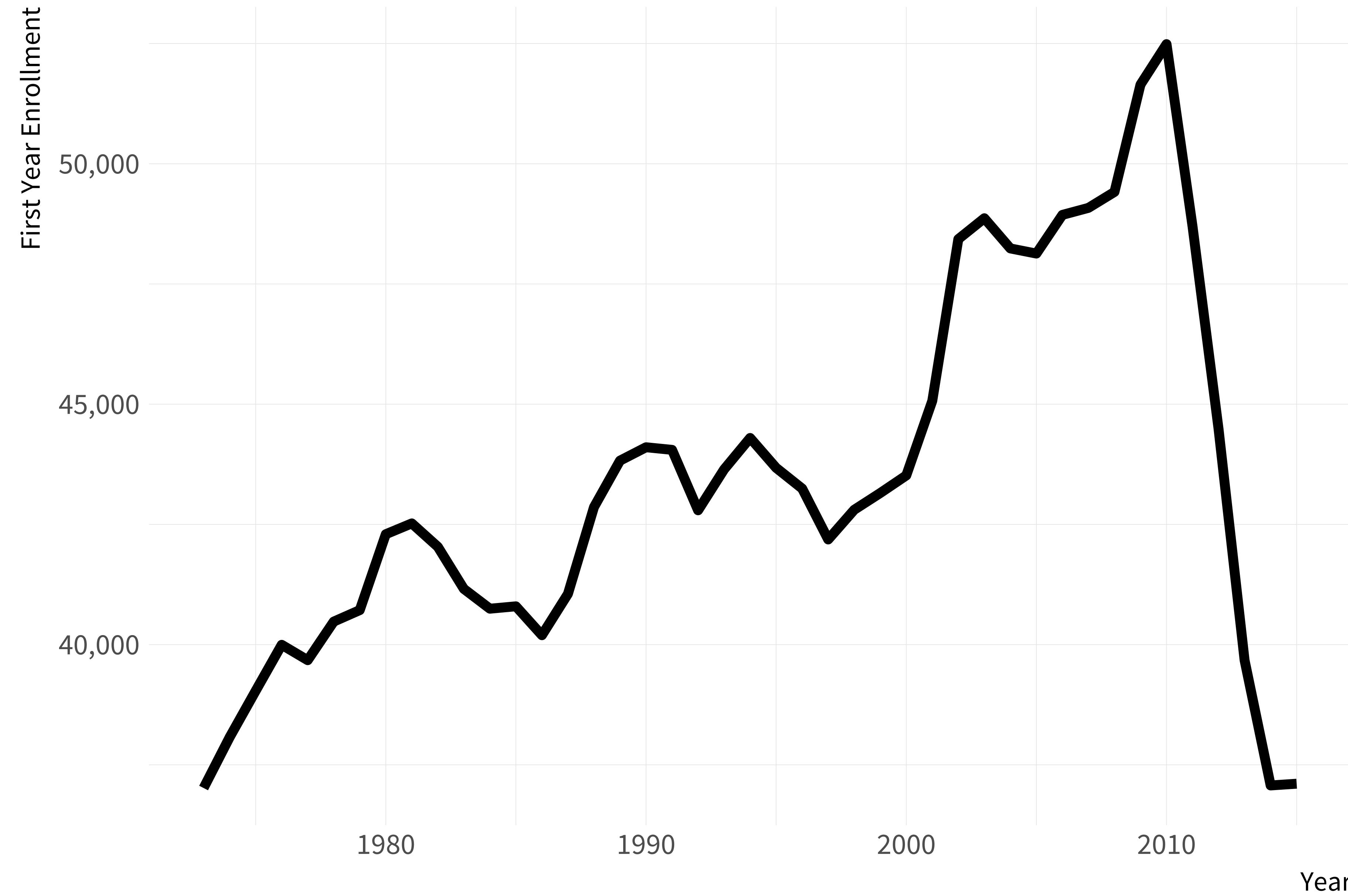


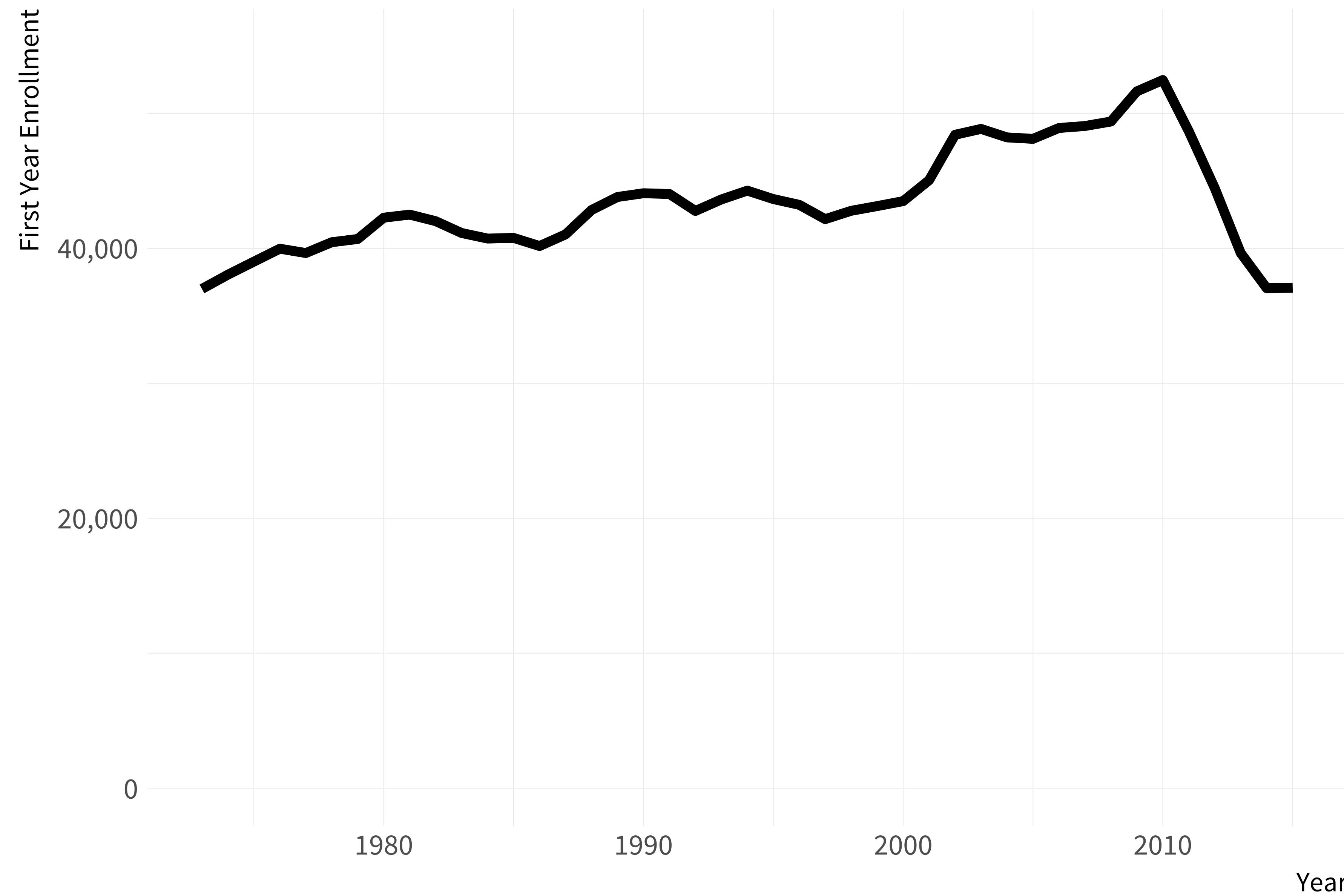
**HONESTY AND
GOOD JUDGMENT**

**Example:
Baselines**



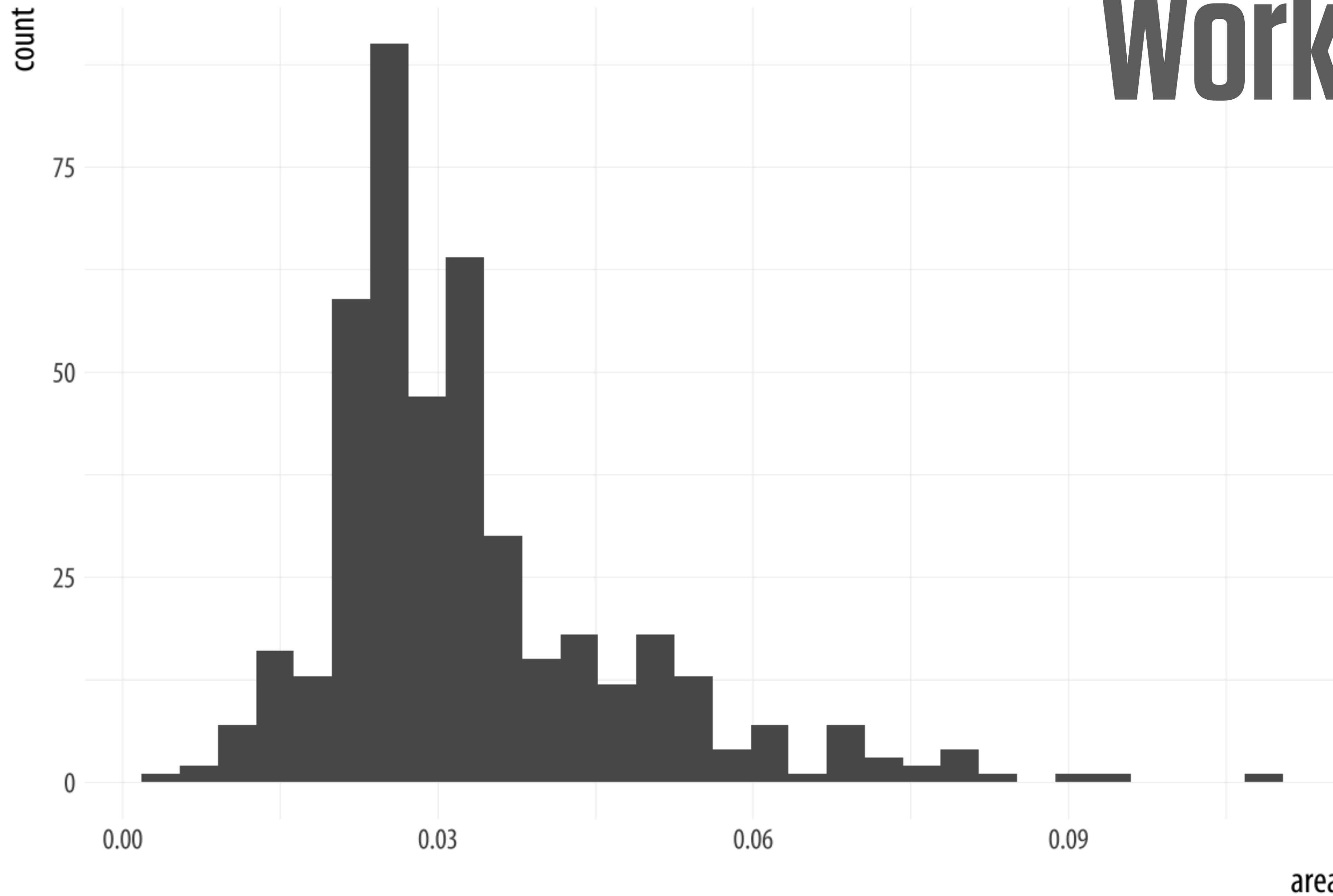






In Practice

Workhorses



lifeExp

80

60

40

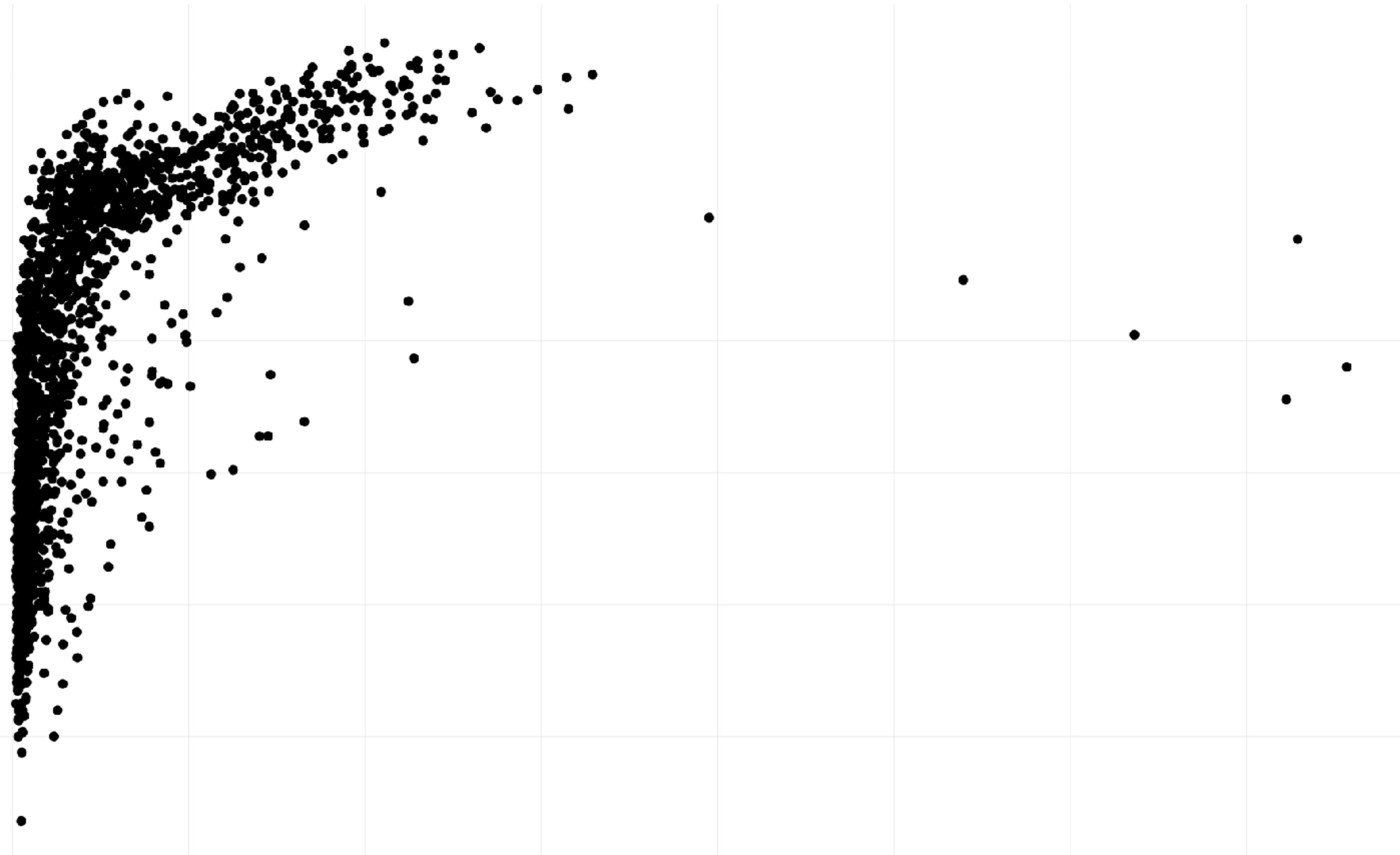
0

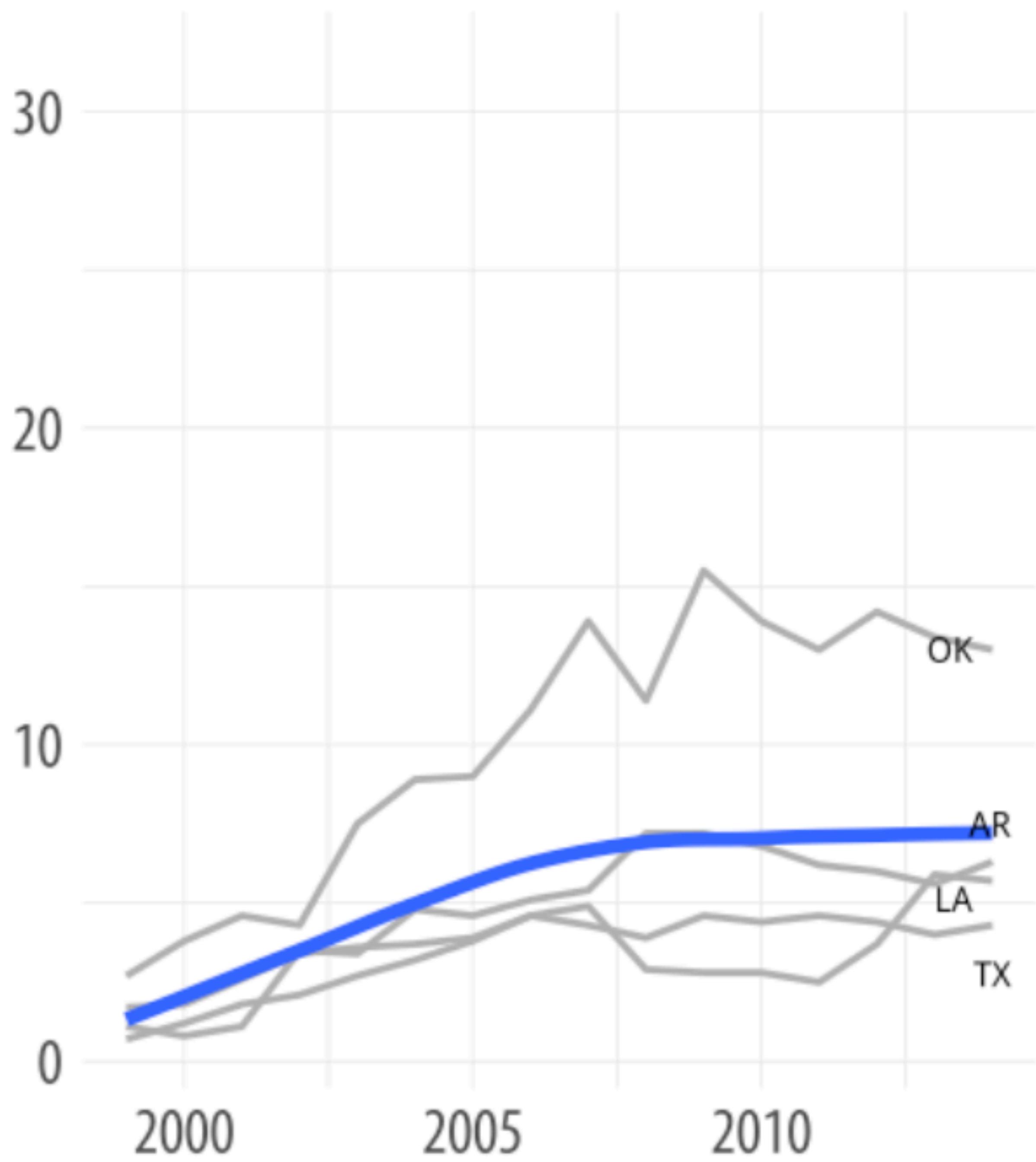
30000

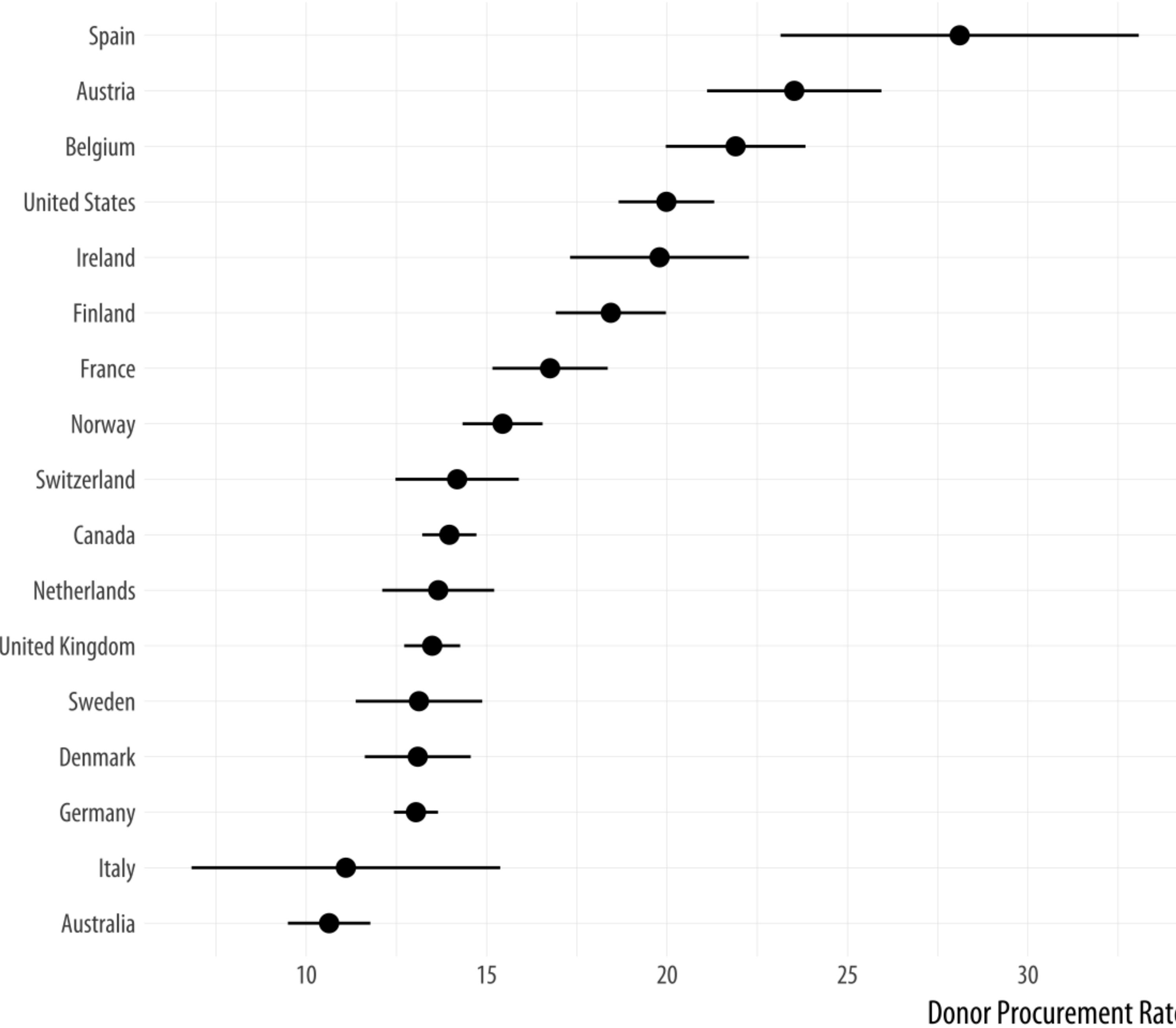
60000

90000

gdpPerCap

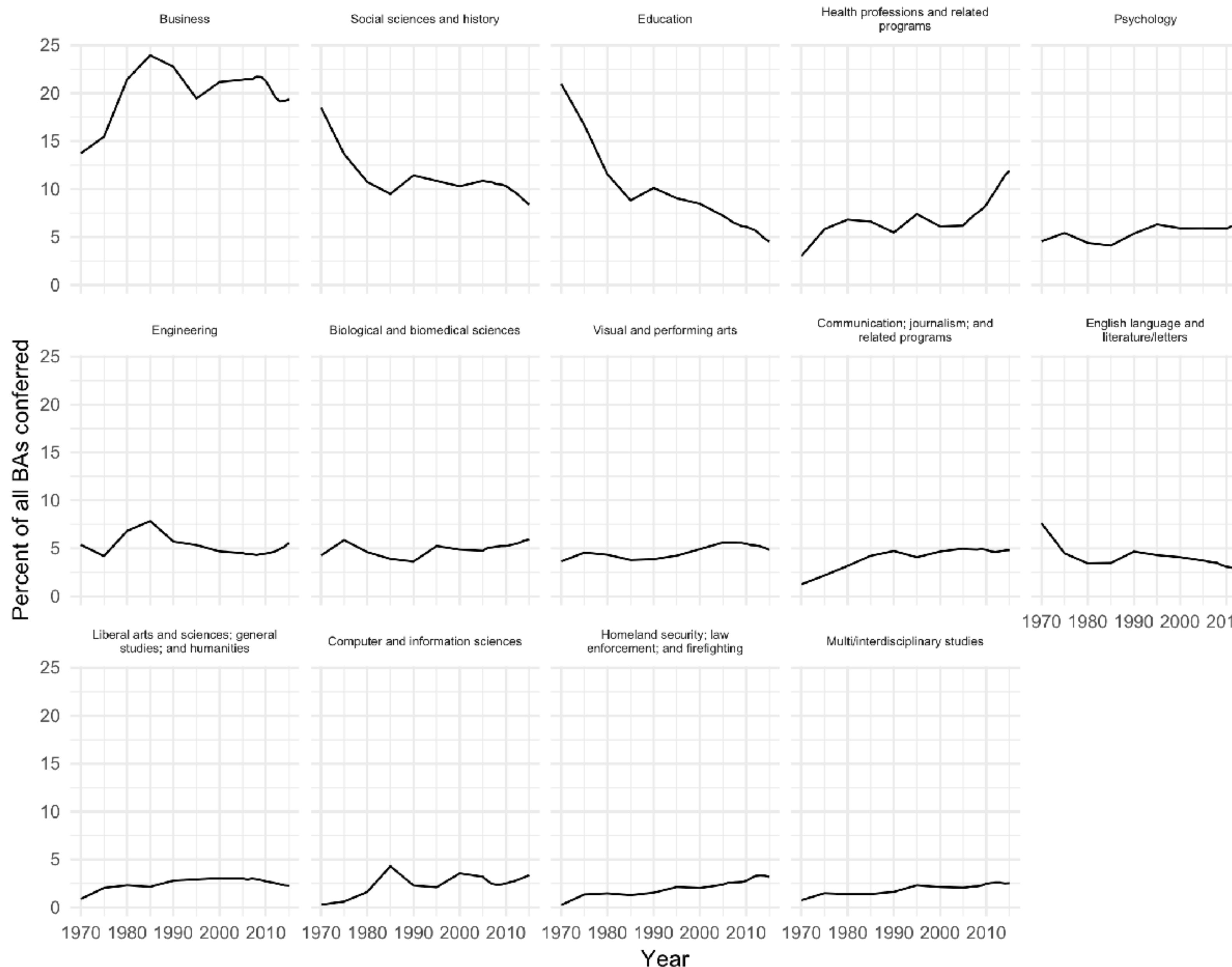






US Trends in Bachelor's Degrees Conferred, 1970-2015, for Areas averaging more than 2% of all degrees

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



Data from NCES Digest 2017, Table 322.10.

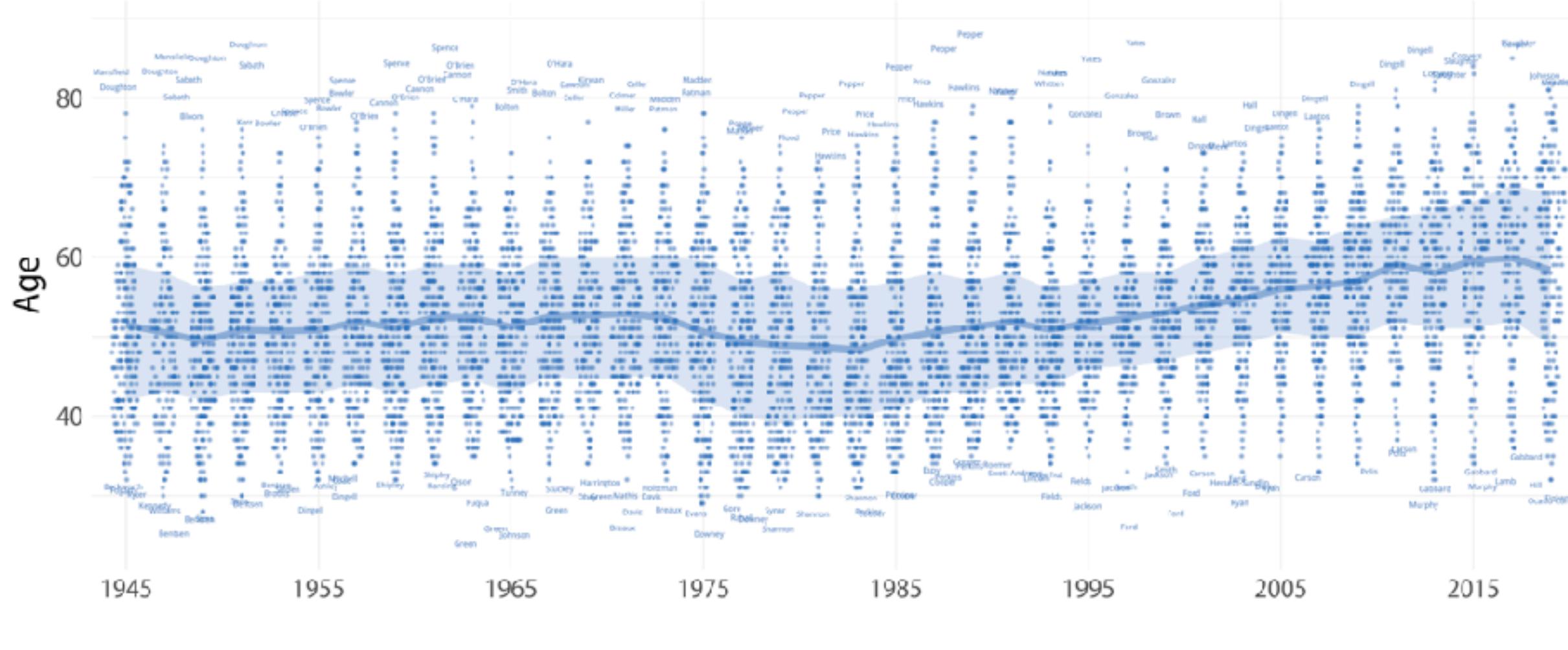
Show Ponies

Age Distribution of Congressional Representatives, 1945-2019

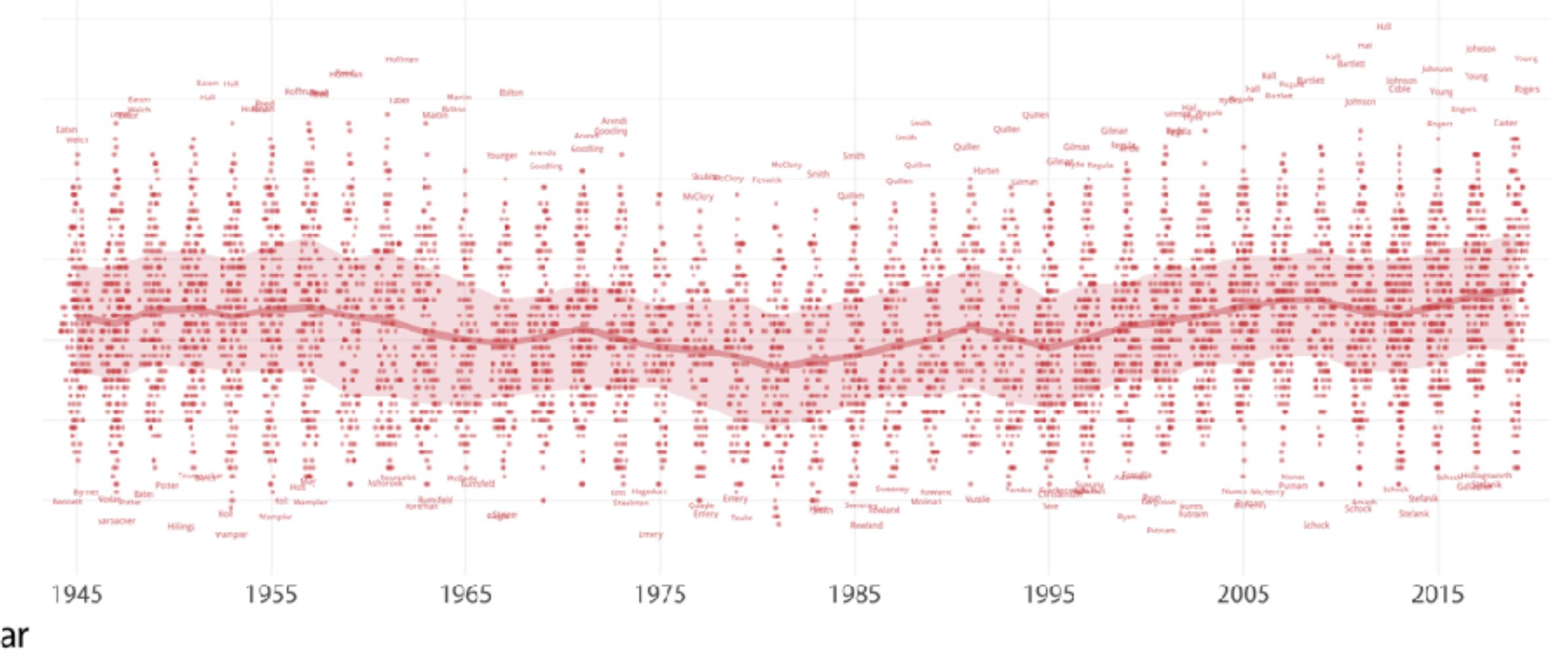
Trend line is mean age; bands are 25th and 75th percentiles of the range.

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

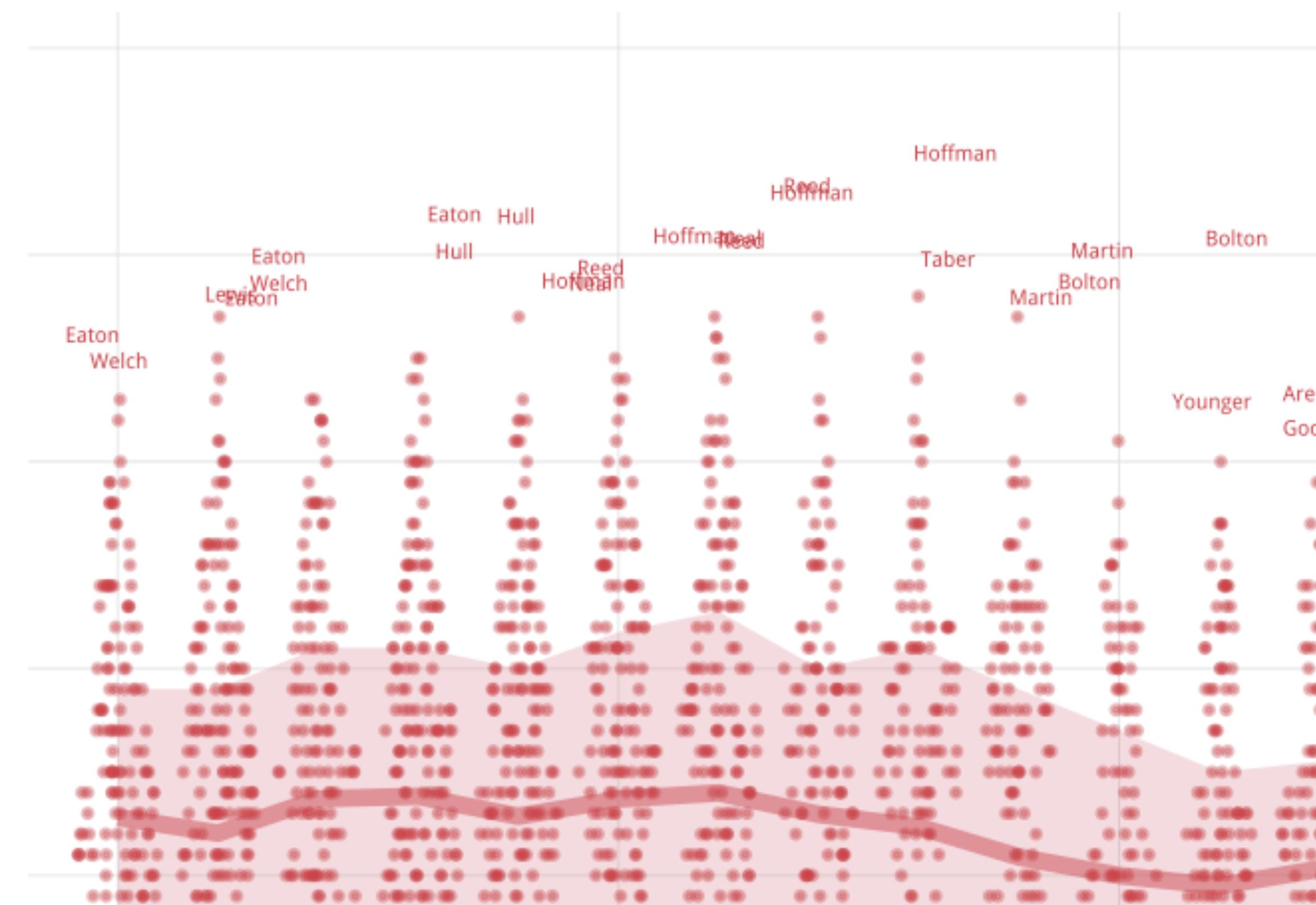
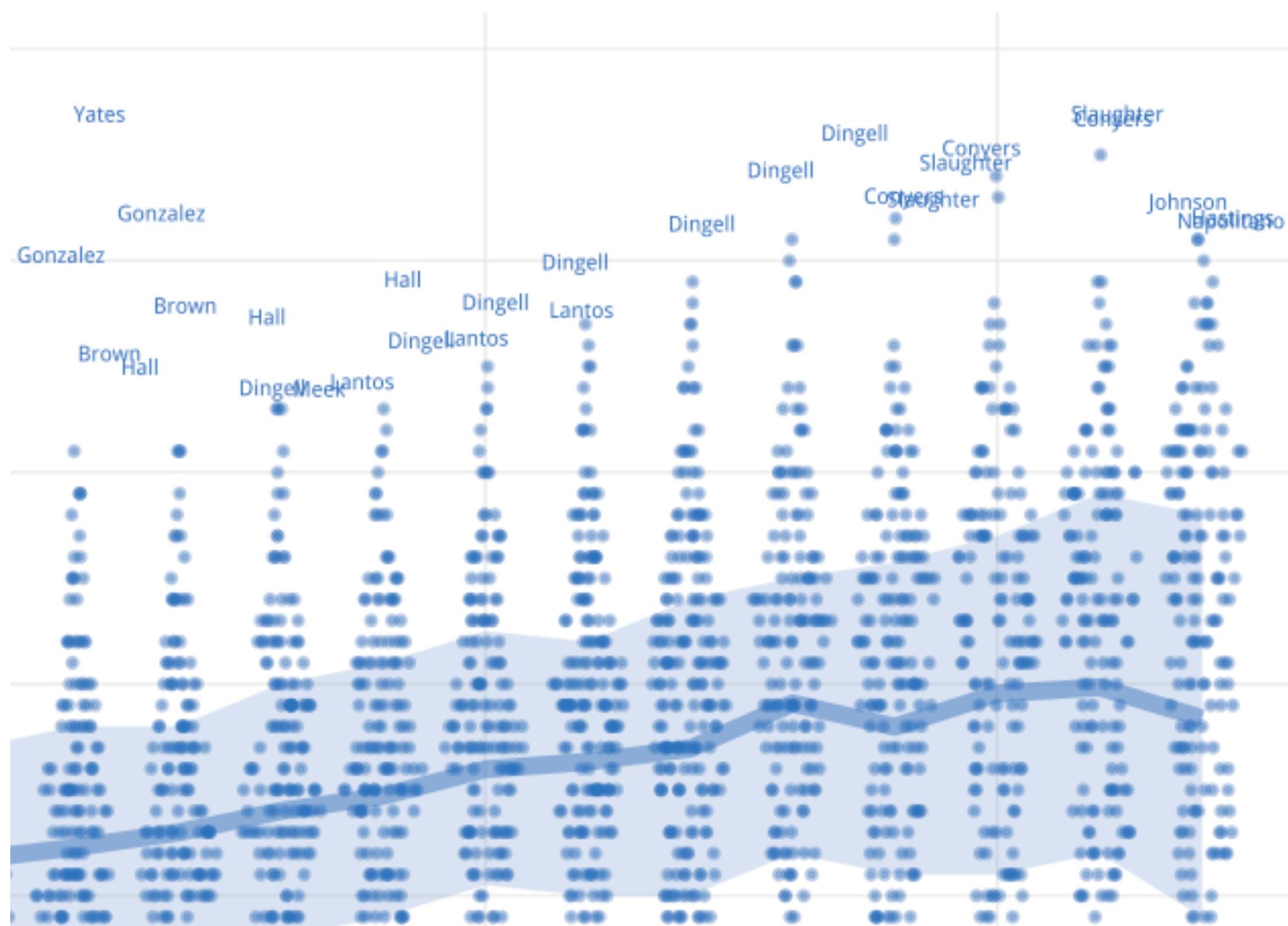
Democrats



Republicans



Show Ponies



Relative prevalence of breed, from Most to Least



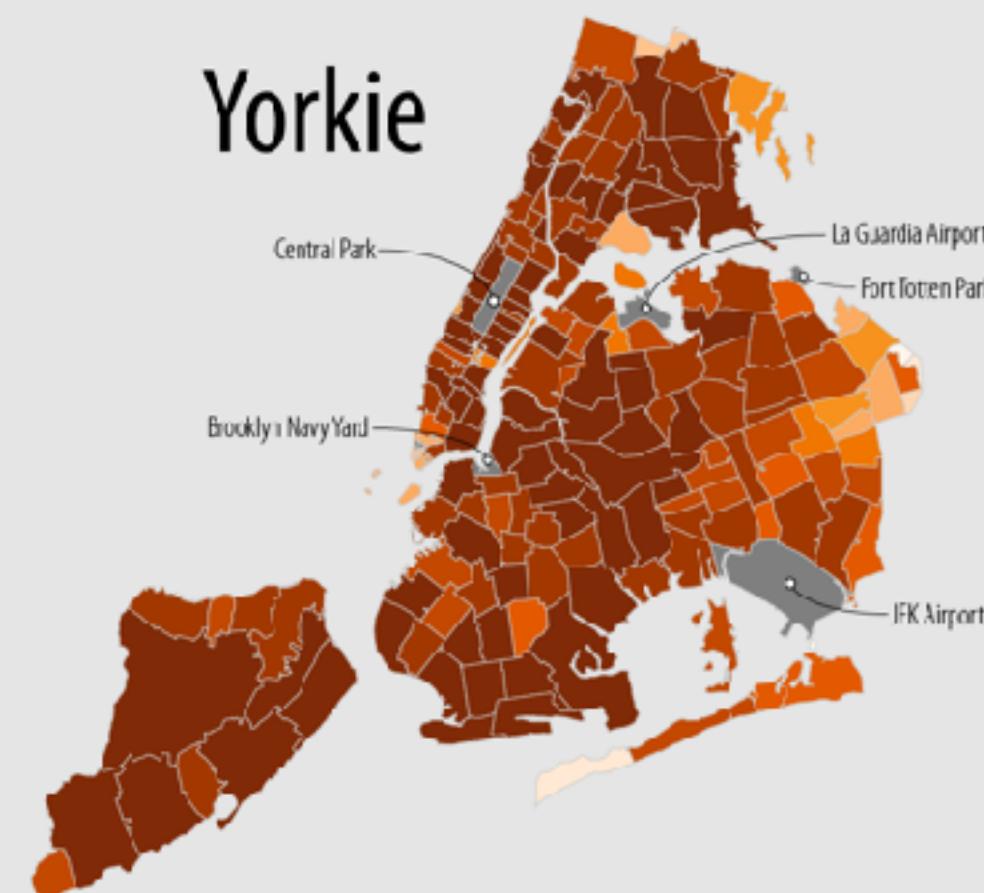
More Common

Less Common

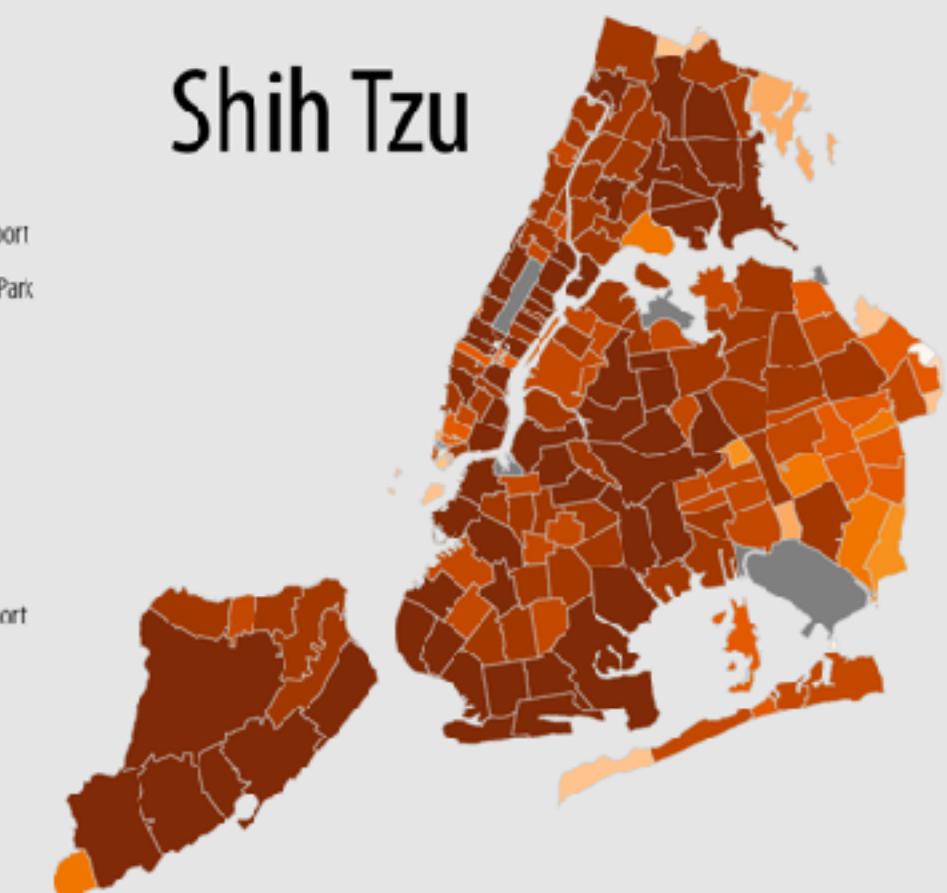
The most common
NYC breed is
“Mixed” or
“Unknown”



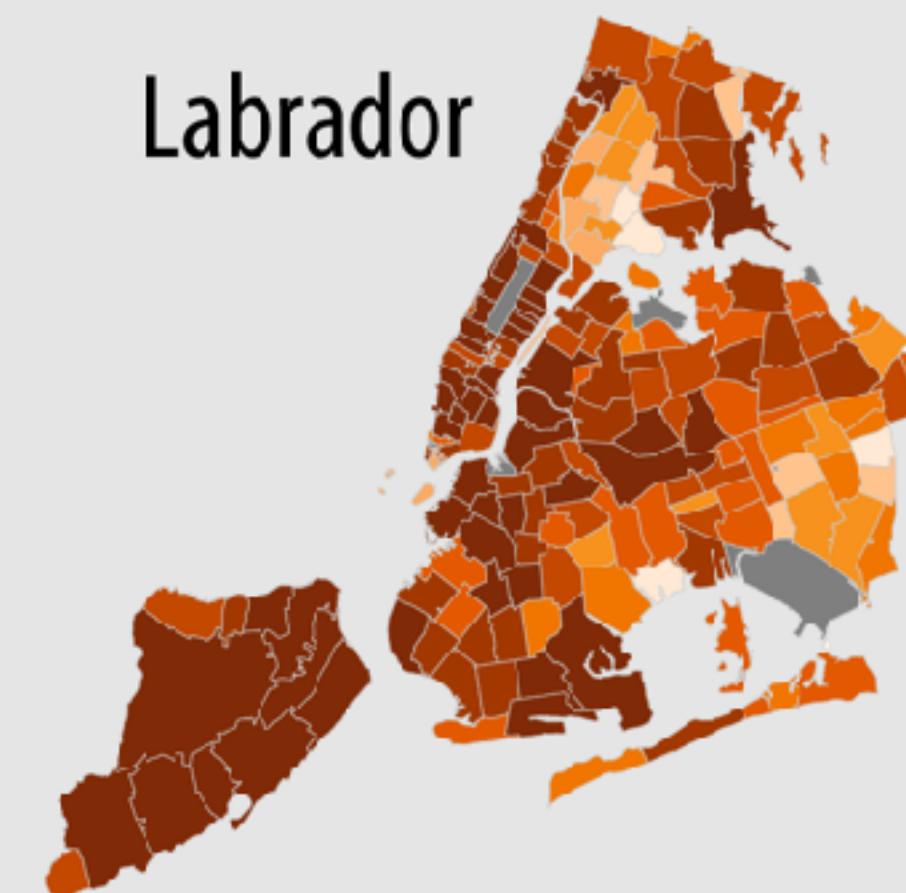
Yorkie



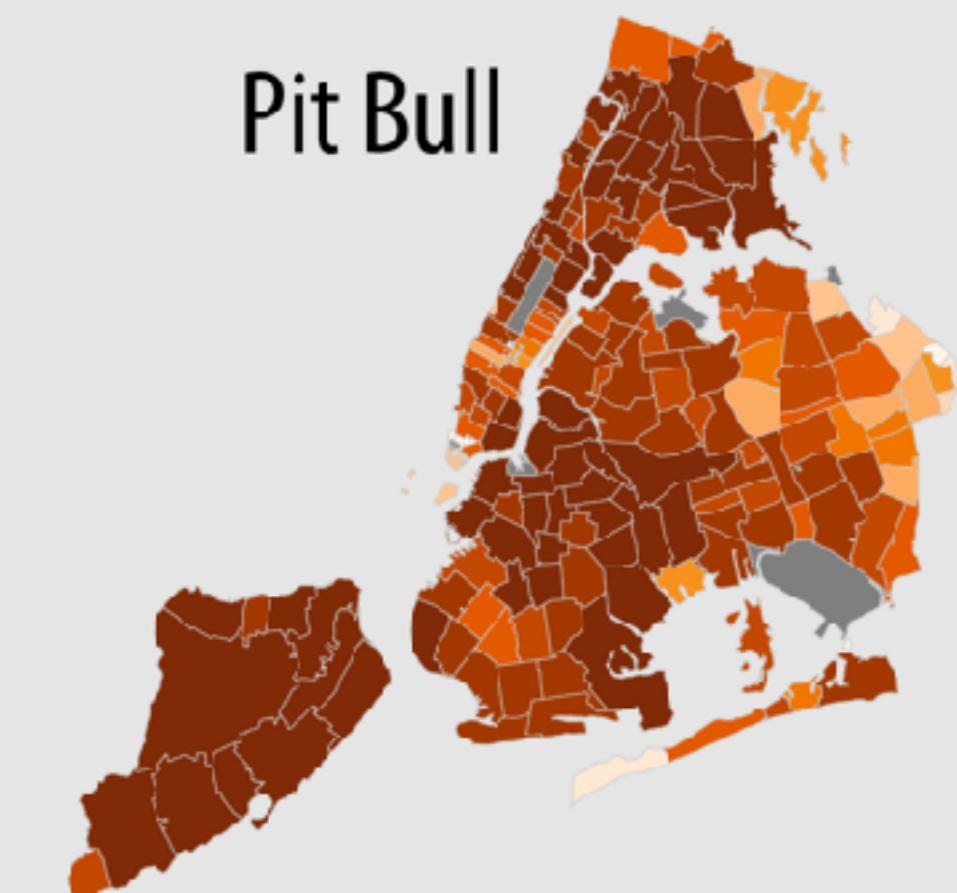
Shih Tzu



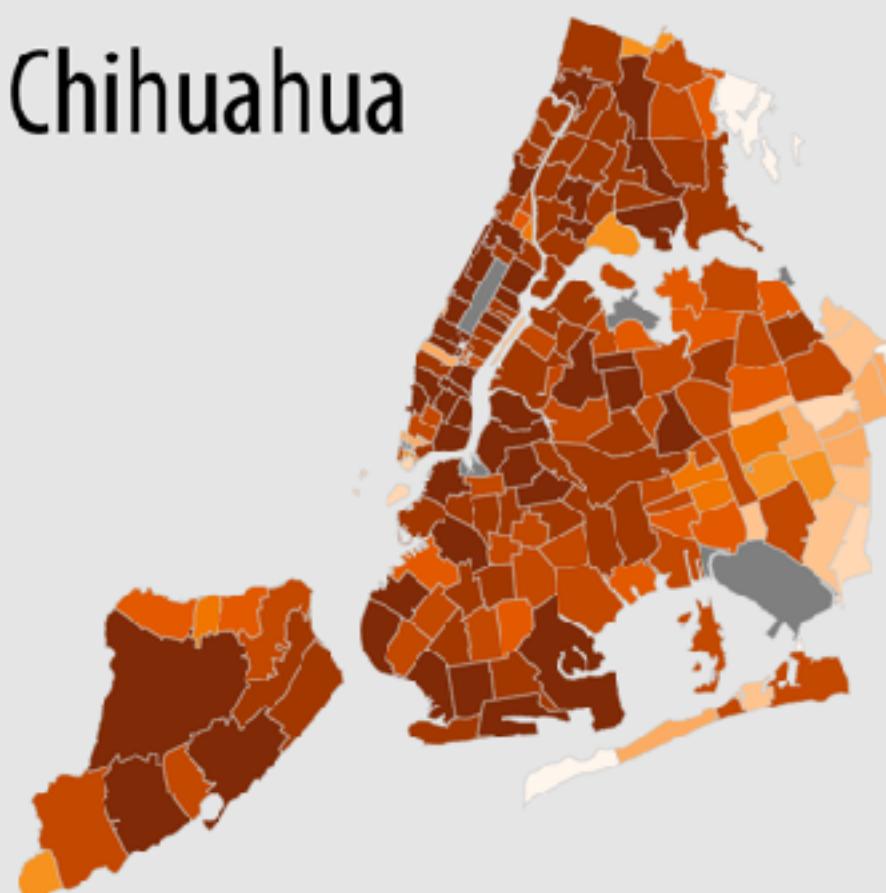
Labrador



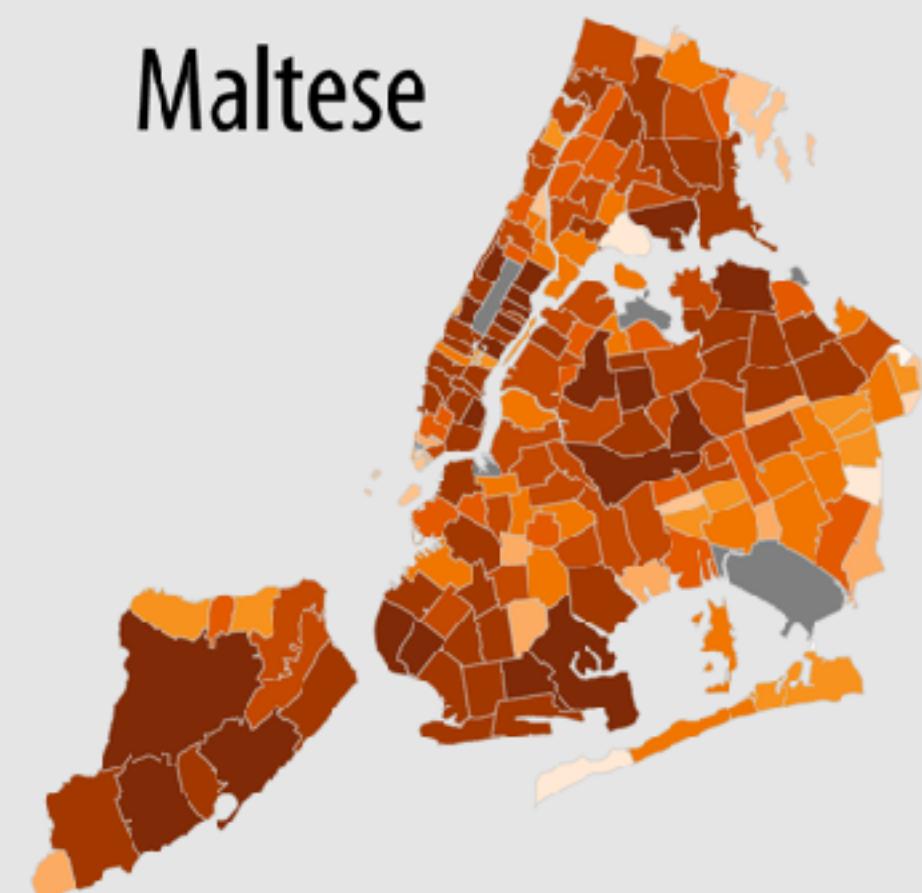
Pit Bull



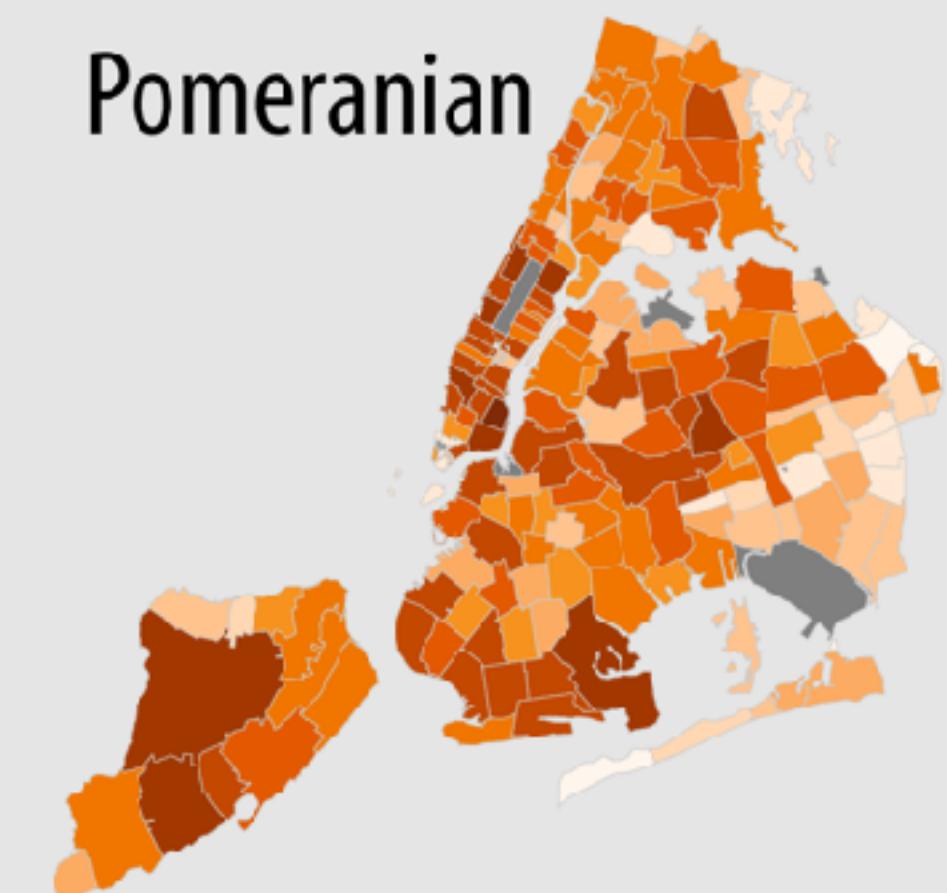
Chihuahua



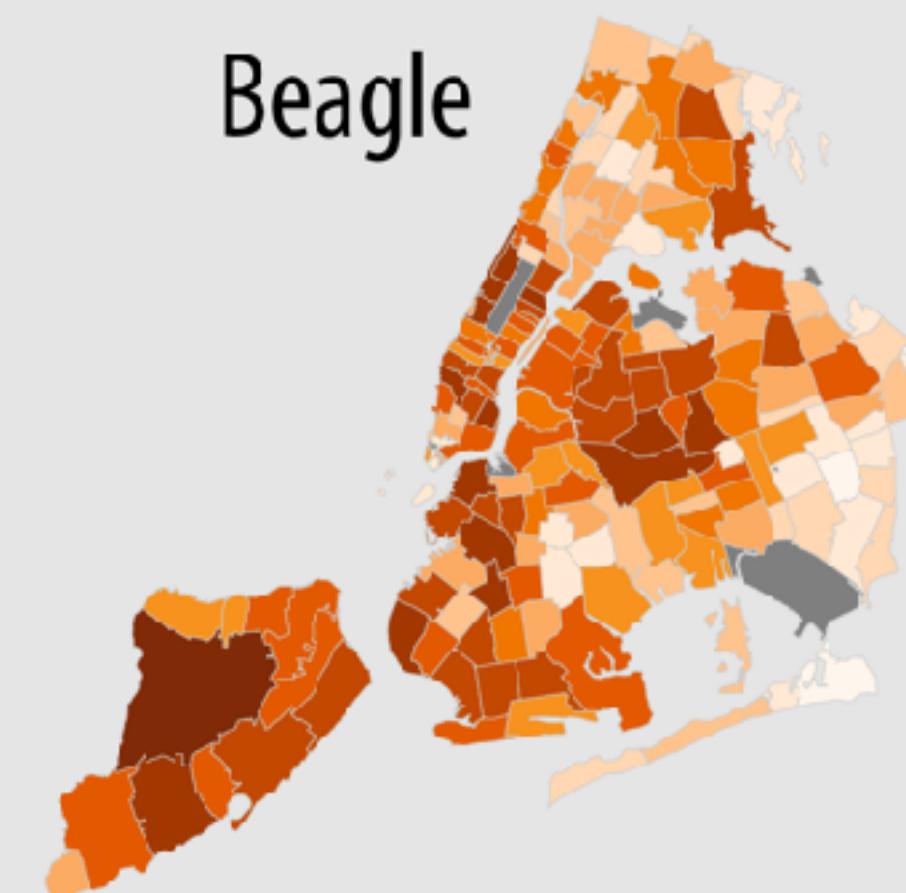
Maltese



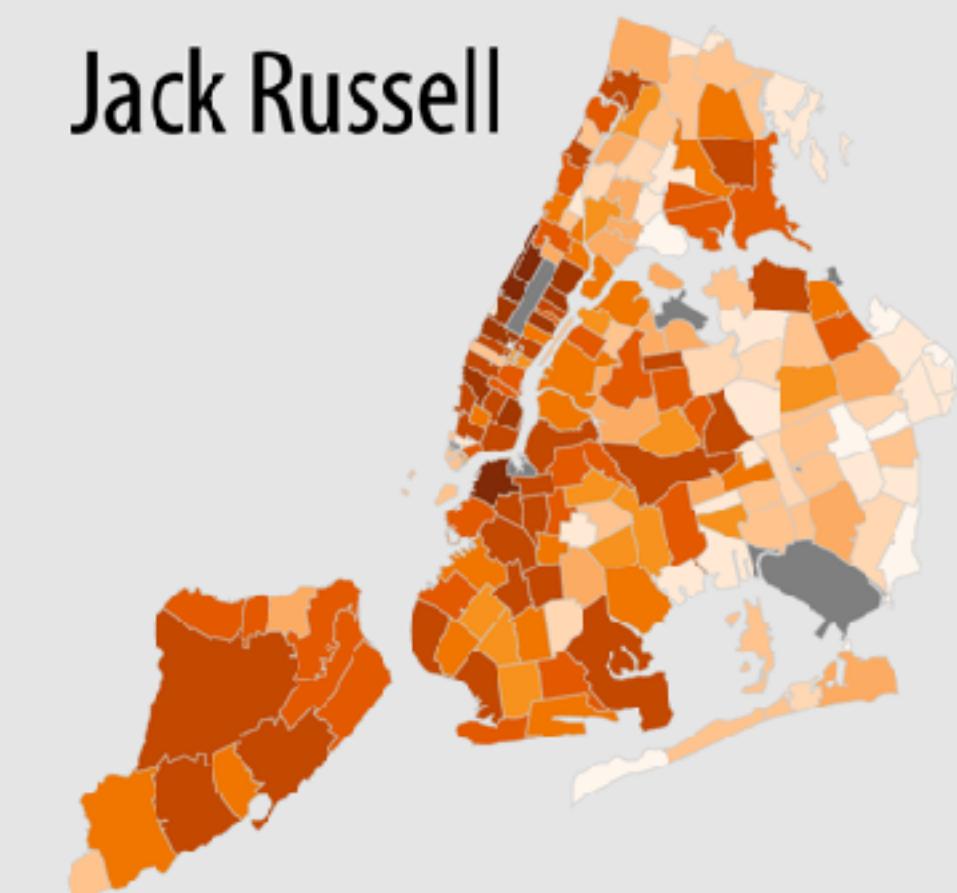
Pomeranian



Beagle

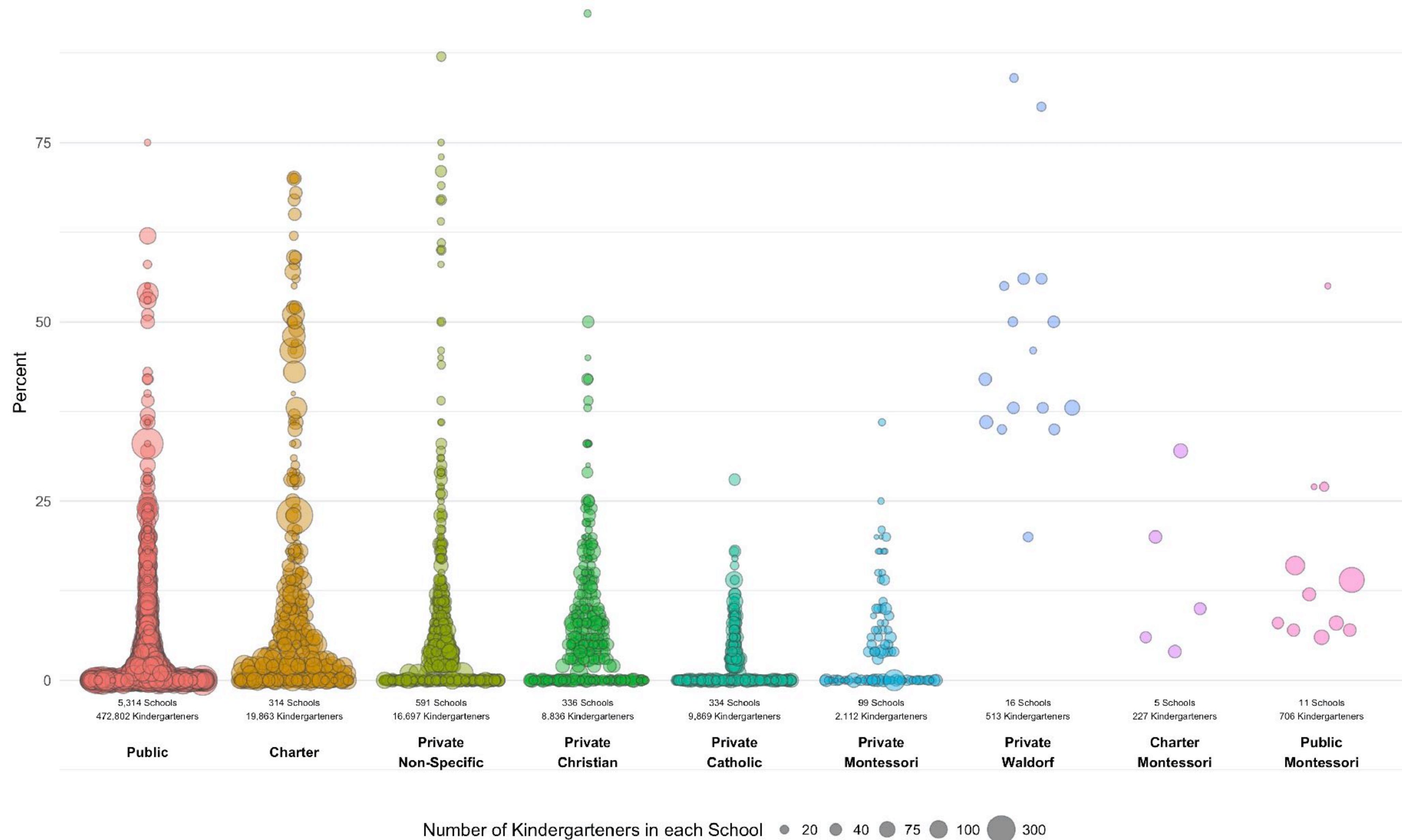


Jack Russell

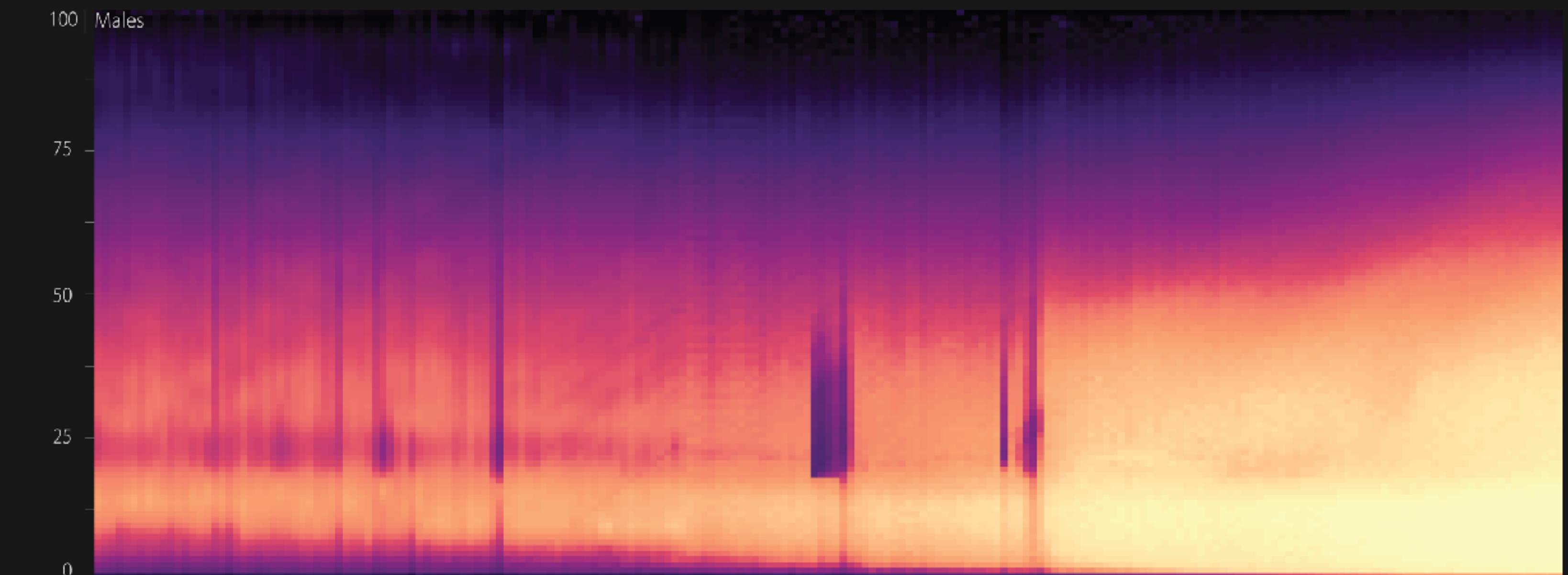


Vaccination Exemption Rates in California Kindergartens

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



MORTALITY IN FRANCE 1816 – 2016



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

"Unicorns"

