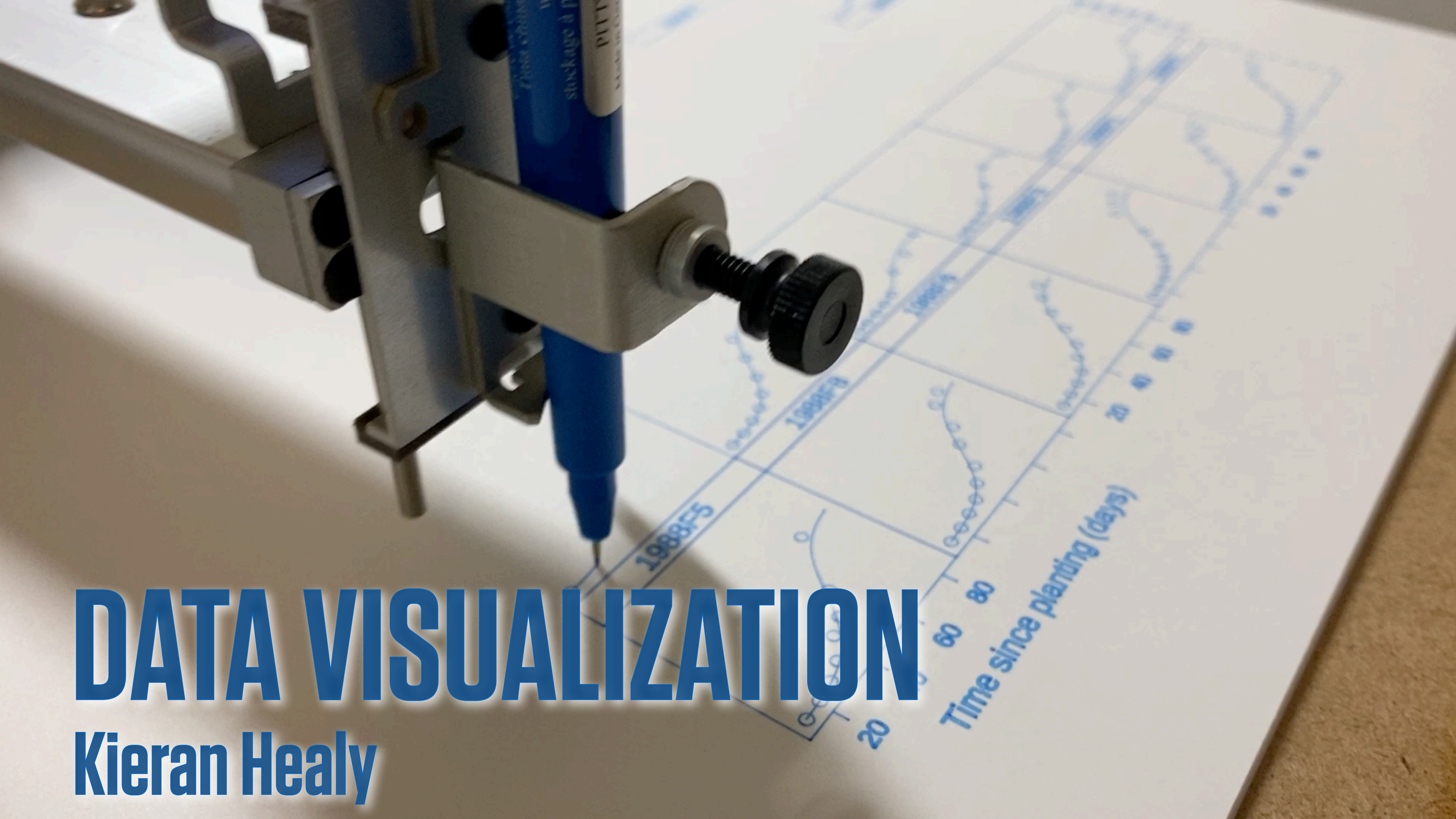
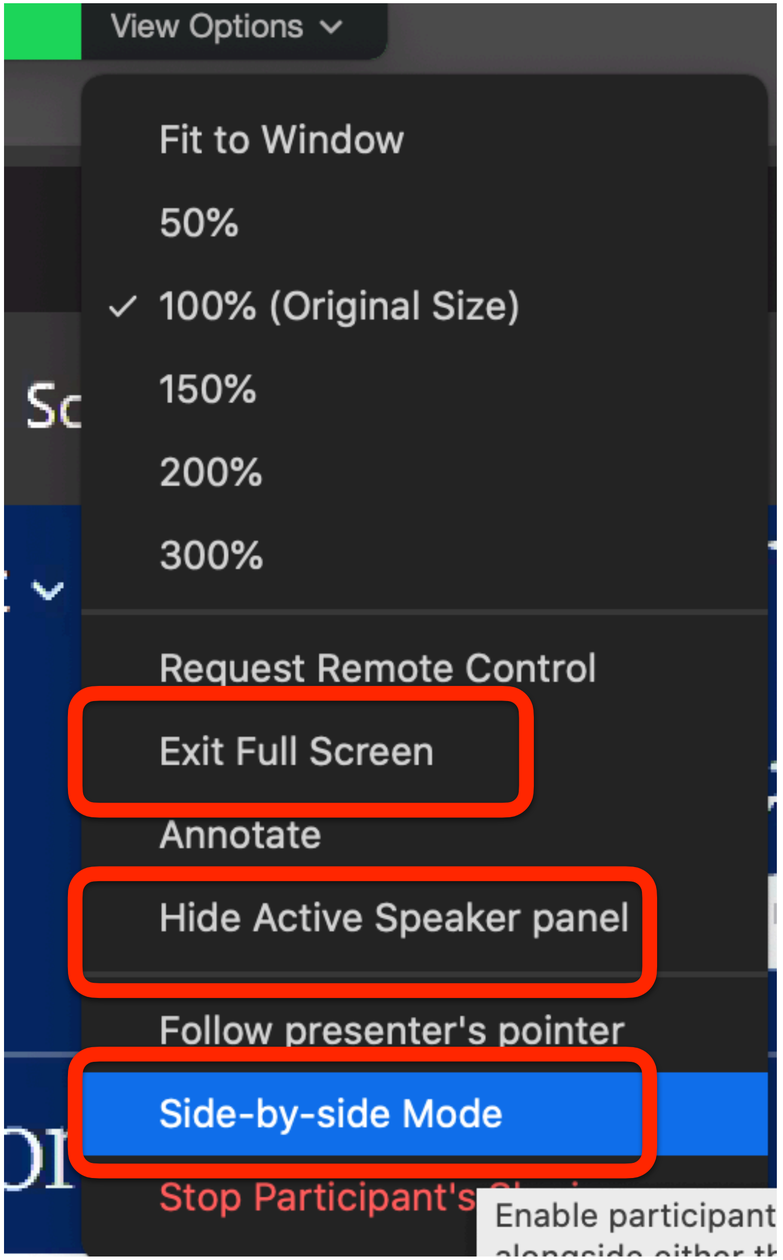


DATA VISUALIZATION

Kieran Healy



For the best Zoom experience, do the following in the "View Options" menu:

- Choose "Full Screen"
- Hide Active Speaker Panel
- Uncheck "Side-by-side Mode"

Housekeeping

Schedule: 11am–2pm

Three 5-minute breaks

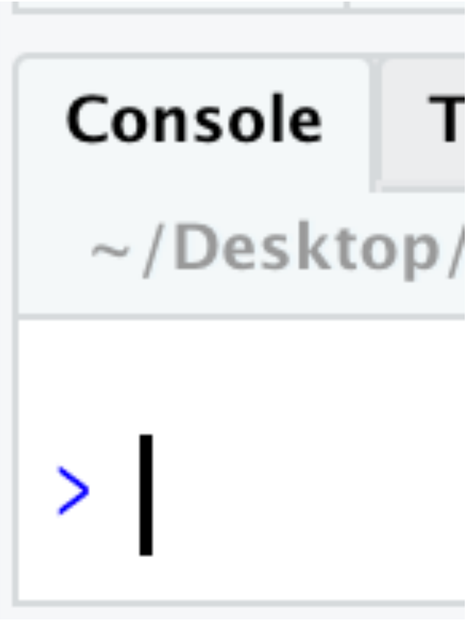
Use Zoom's Chat to ask questions

Use the Slack, too

Housekeeping

RStudio

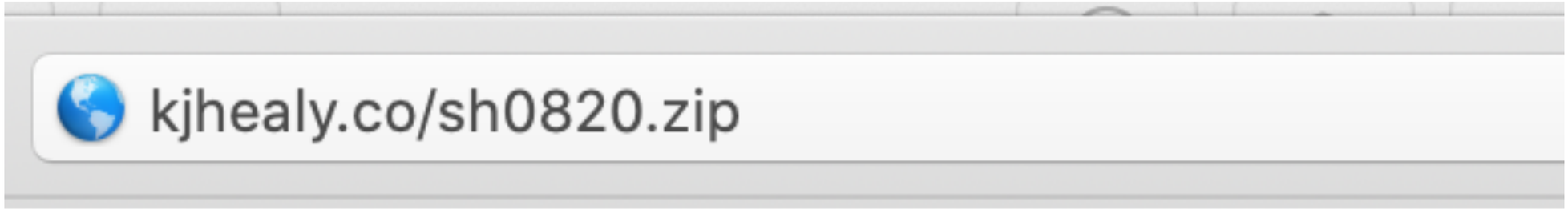
Get Up & Running



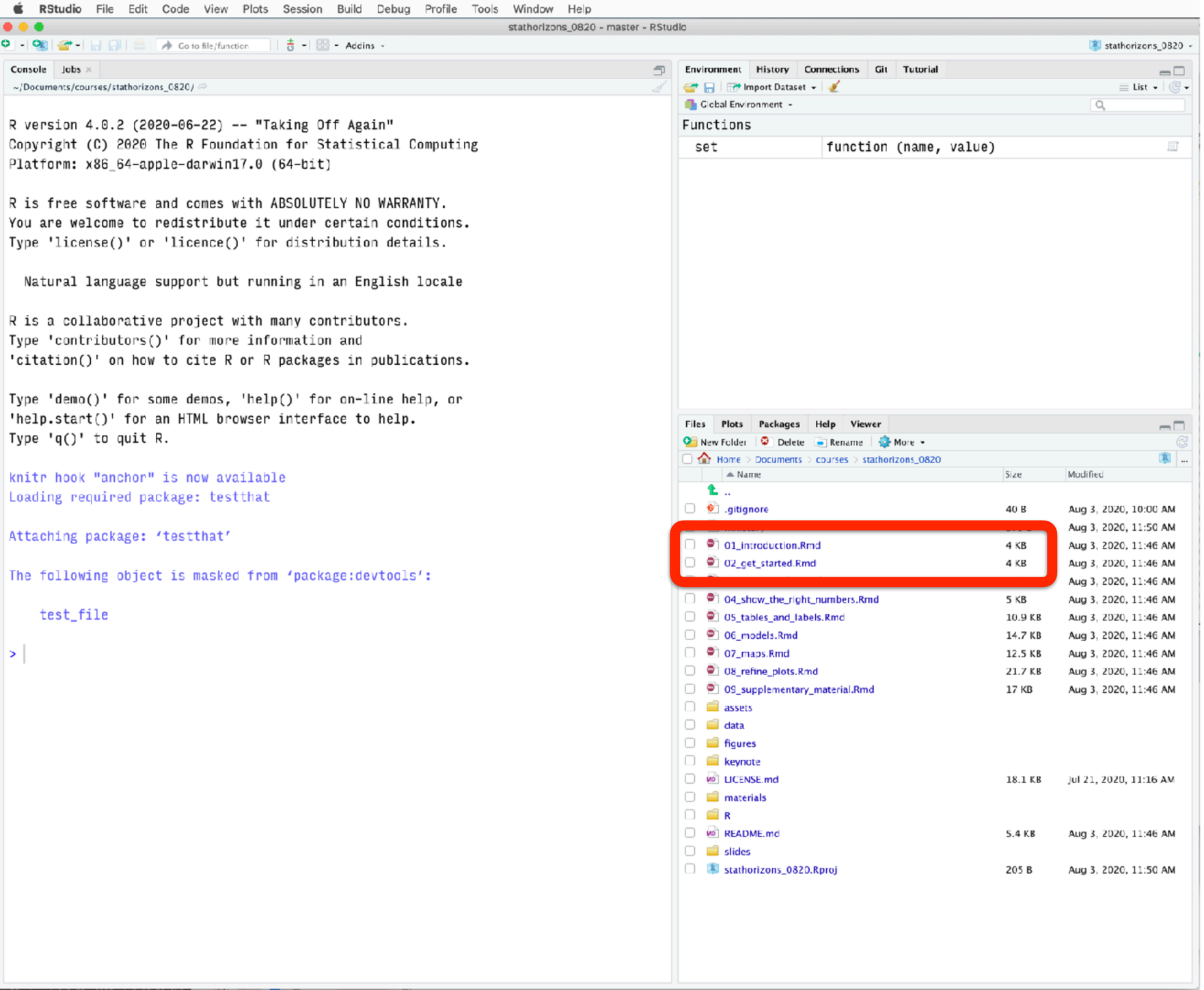
Get the Notes from inside RStudio

```
# Read the options for where to put it
usethis::use_course("http://kjhealy.co/sh0820.zip")
```

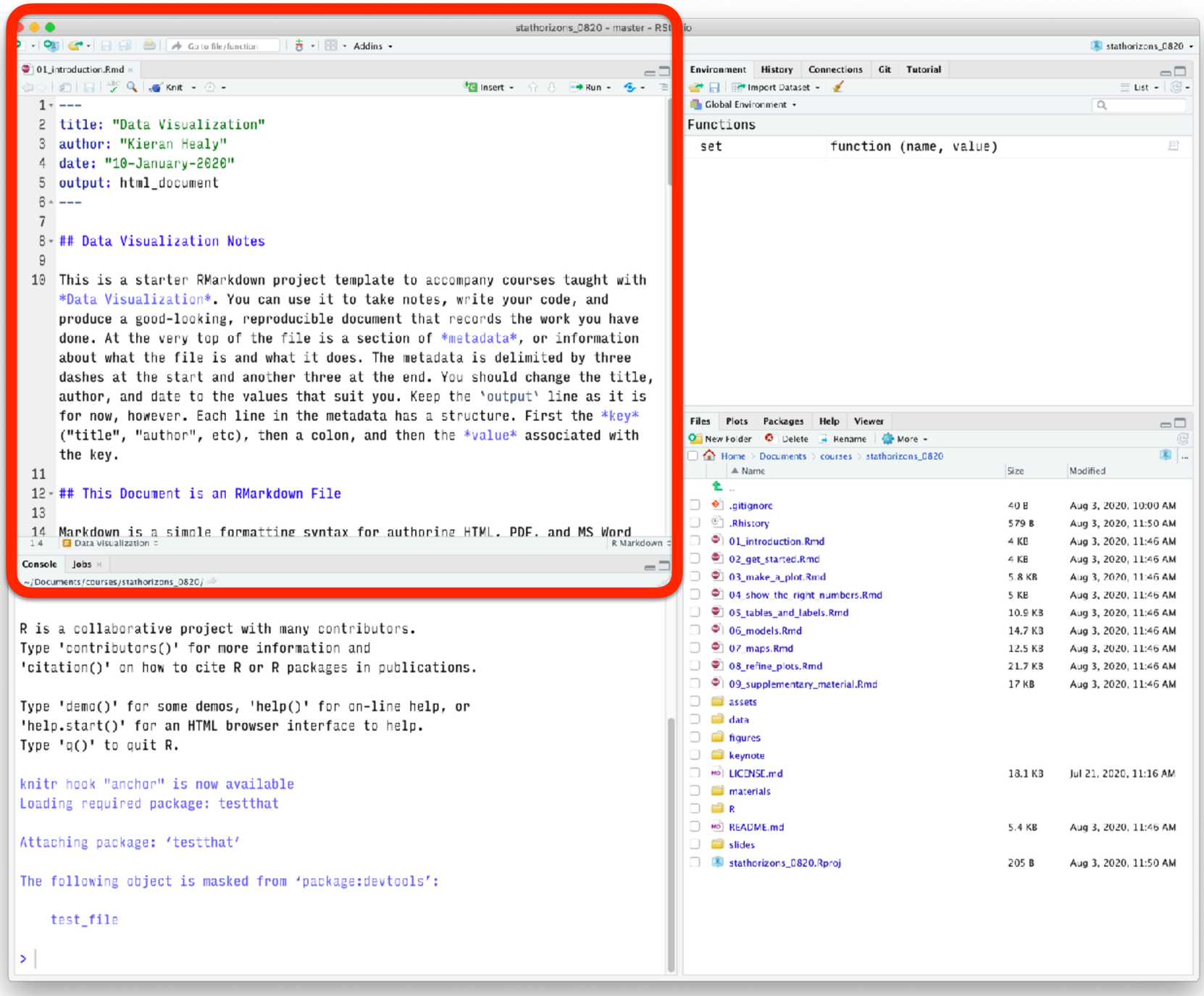
Or just download from your Browser



RStudio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
stathorizons_0820 - master - RStudio
Go to file/function
Addins
stathorizons_0820
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
>
Environment History Connections Git Tutorial
Import Dataset
List
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
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
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


stathorizons_0820 - master - RStudio
01_introduction.Rmd
---
title: "Data Visualization"
author: "Kieran Healy"
date: "10-January-2020"
output: html_document

Data Visualization Notes
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.
This Document is an RMarkdown File
Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word
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
Functions
set
function (name, value)
Files
Plots
Packages
Help
Viewer
Home
Documents
courses
stathorizons_0820
.gitignore
.Rhistory
01_introduction.Rmd
02_get_started.Rmd
03_make_a_plot.Rmd
04_show_the_right_numbers.Rmd
05_tables_and_labels.Rmd
06_models.Rmd
07_maps.Rmd
08_refine_plots.Rmd
09_supplementary_material.Rmd
assets
data
figures
keynote
LICENSE.md
materials
R
README.md
slides
stathorizons_0820.Rproj

stathorizons_0820 - master - RStudio

Go to file/function | Addins

01_introduction.Rmd

Knit | Insert | Run

```
---
title: "Data Visualization"
author: "Kieran Healy"
date: "10-January-2020"
output: html_document
---

## Data Visualization Notes

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

Import Datase

Global Environment

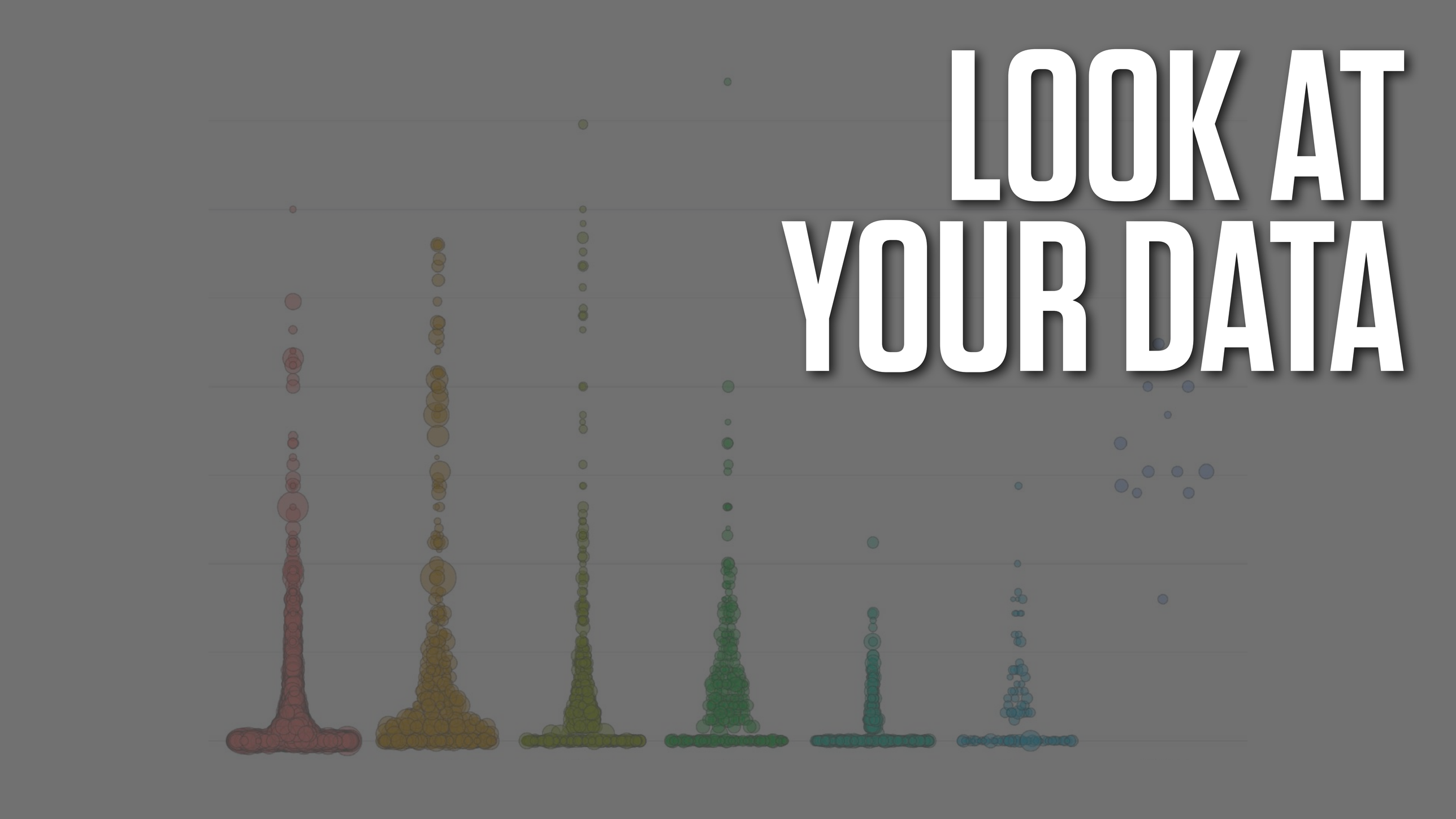
Functions

set

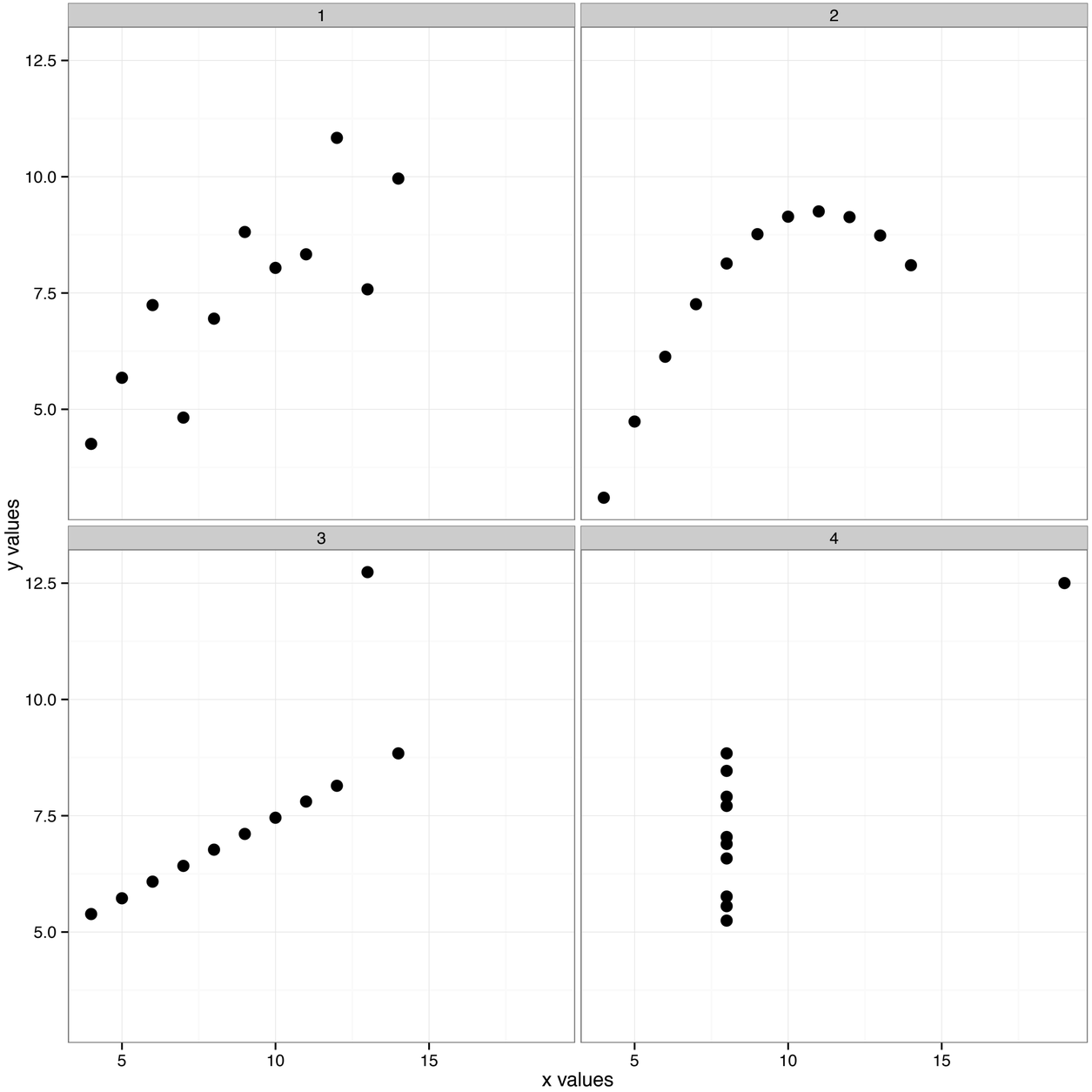
Files | Plots | Packages

New Folder | Delete

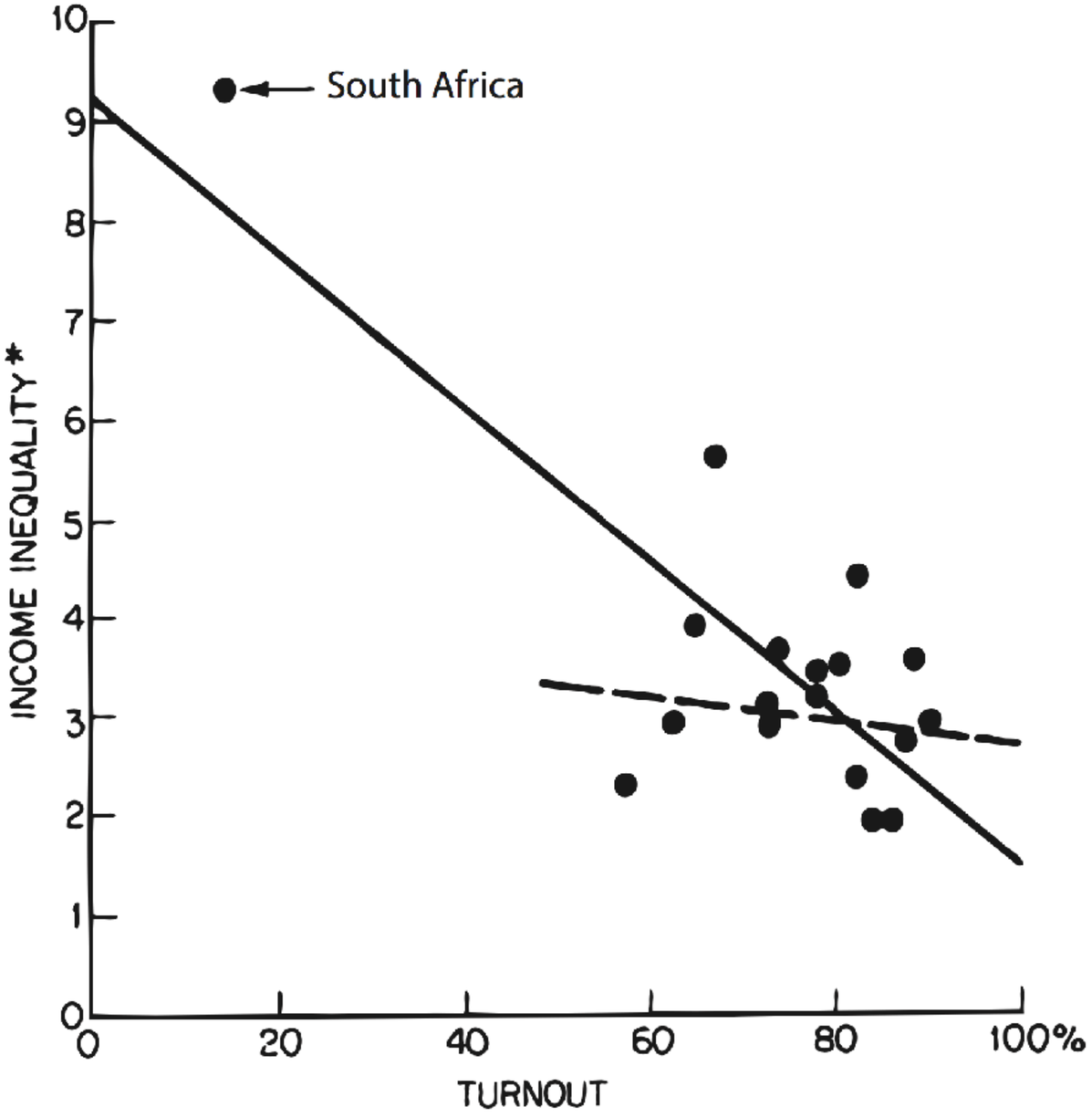
Home > Documents


LOOK AT
YOUR DATA

SEEING THINGS

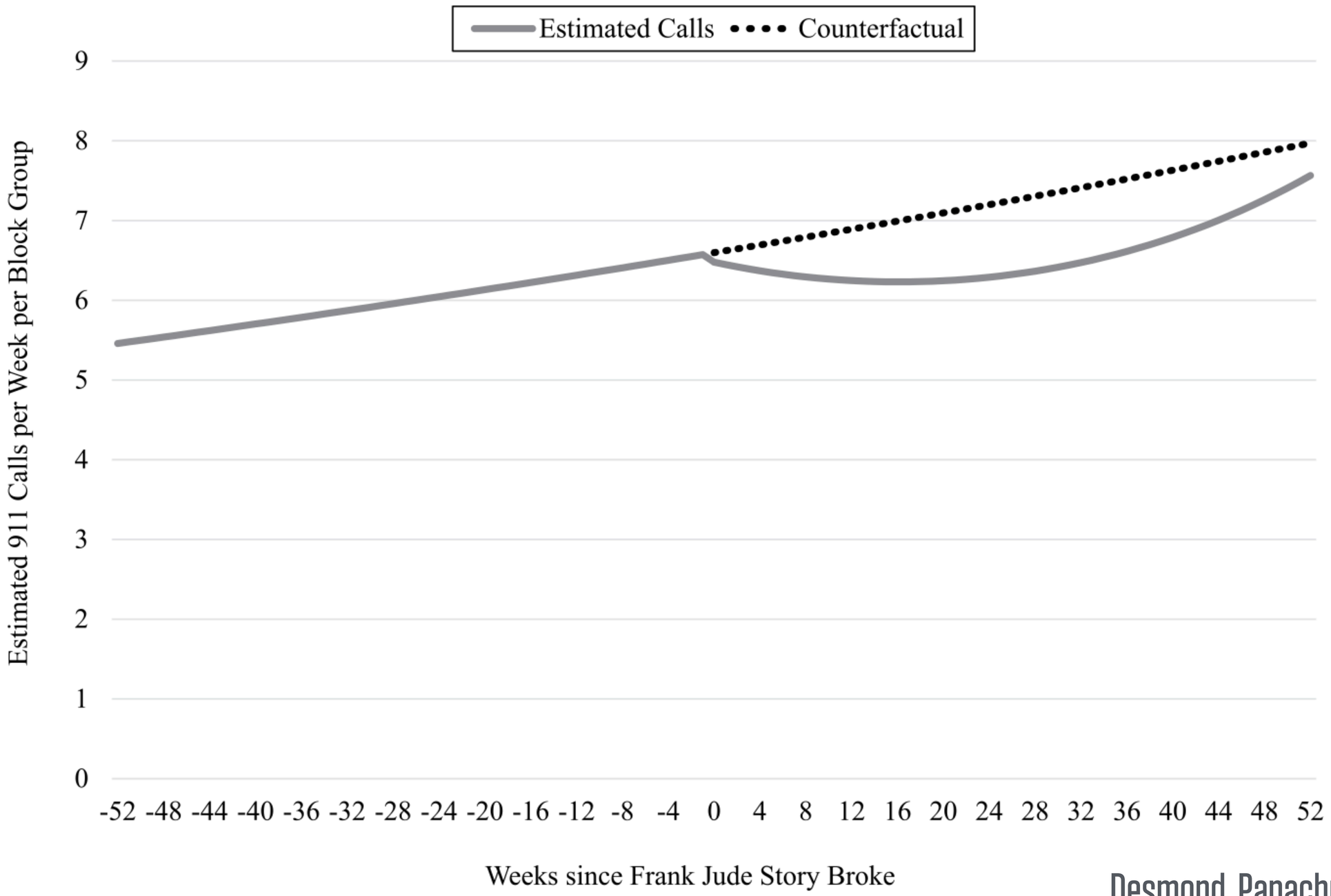


For all panels, N=11; Mean=7.5; Regression: Y=3 + 0.5(X); SE of slope estimate: 0.118, t=4.24; Sum of Squares (X−X): 100; r=0.82.

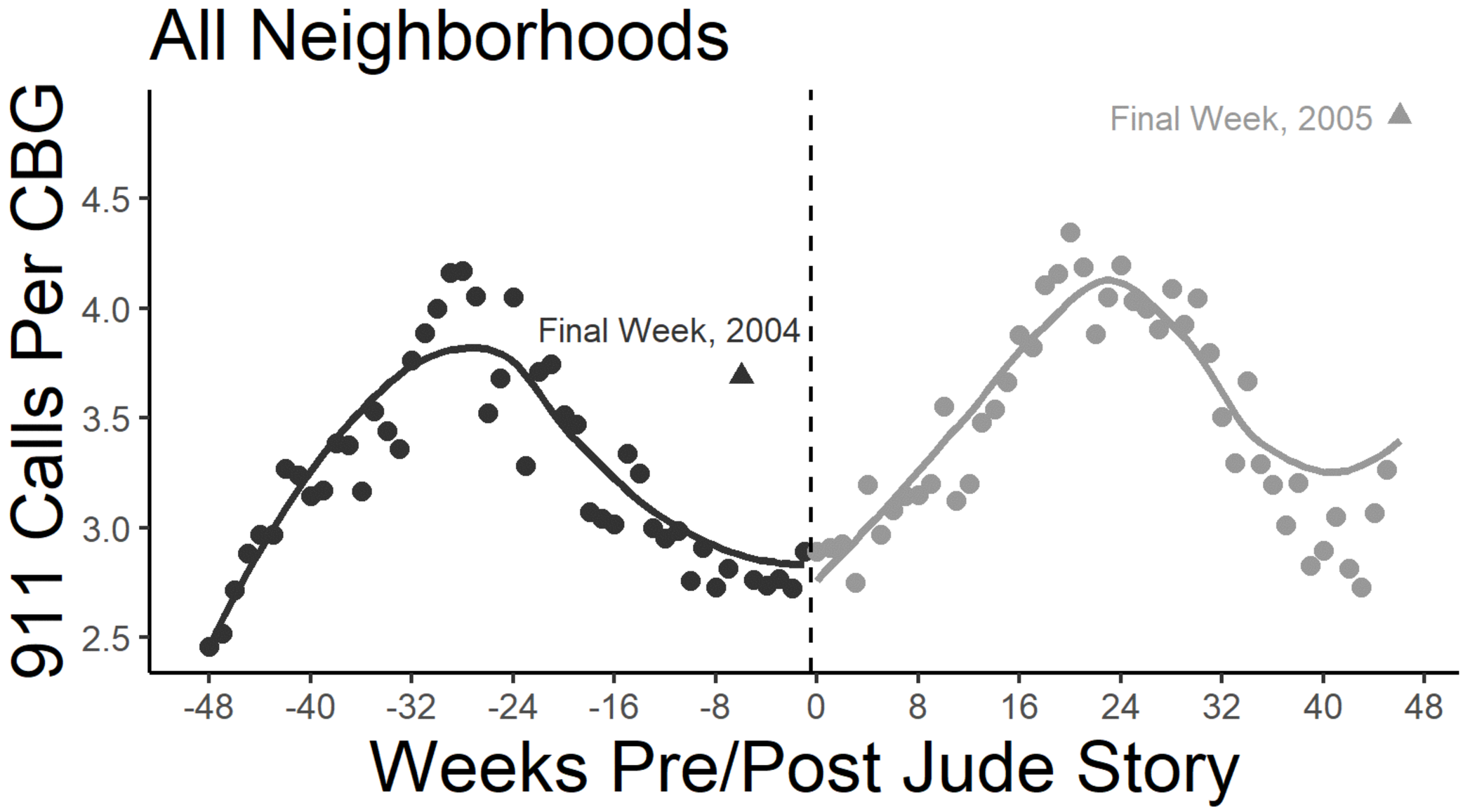


Key. Bivariate slope including South Africa ($N = 18$)

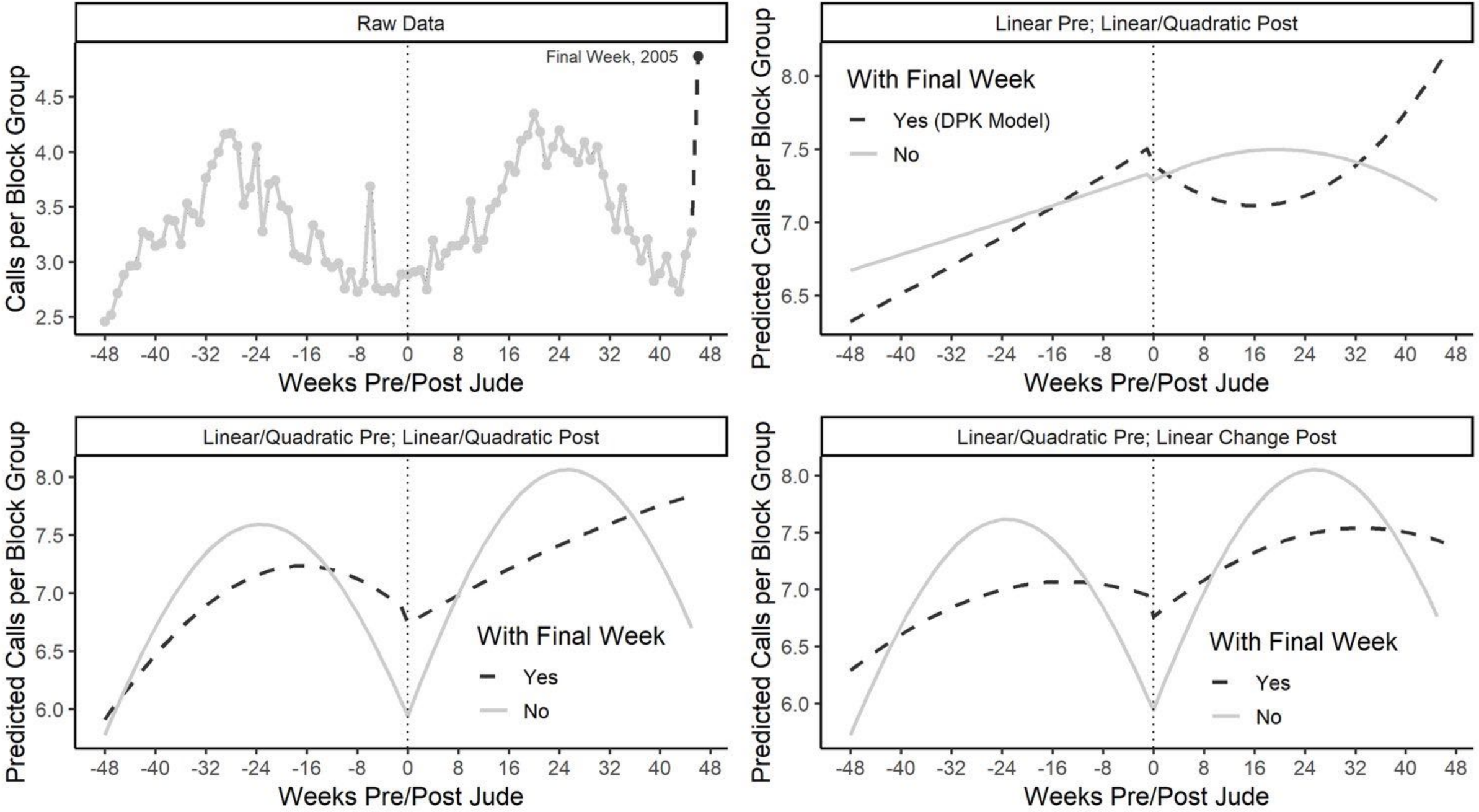
Bivariate slope excluding South Africa ($N = 17$)



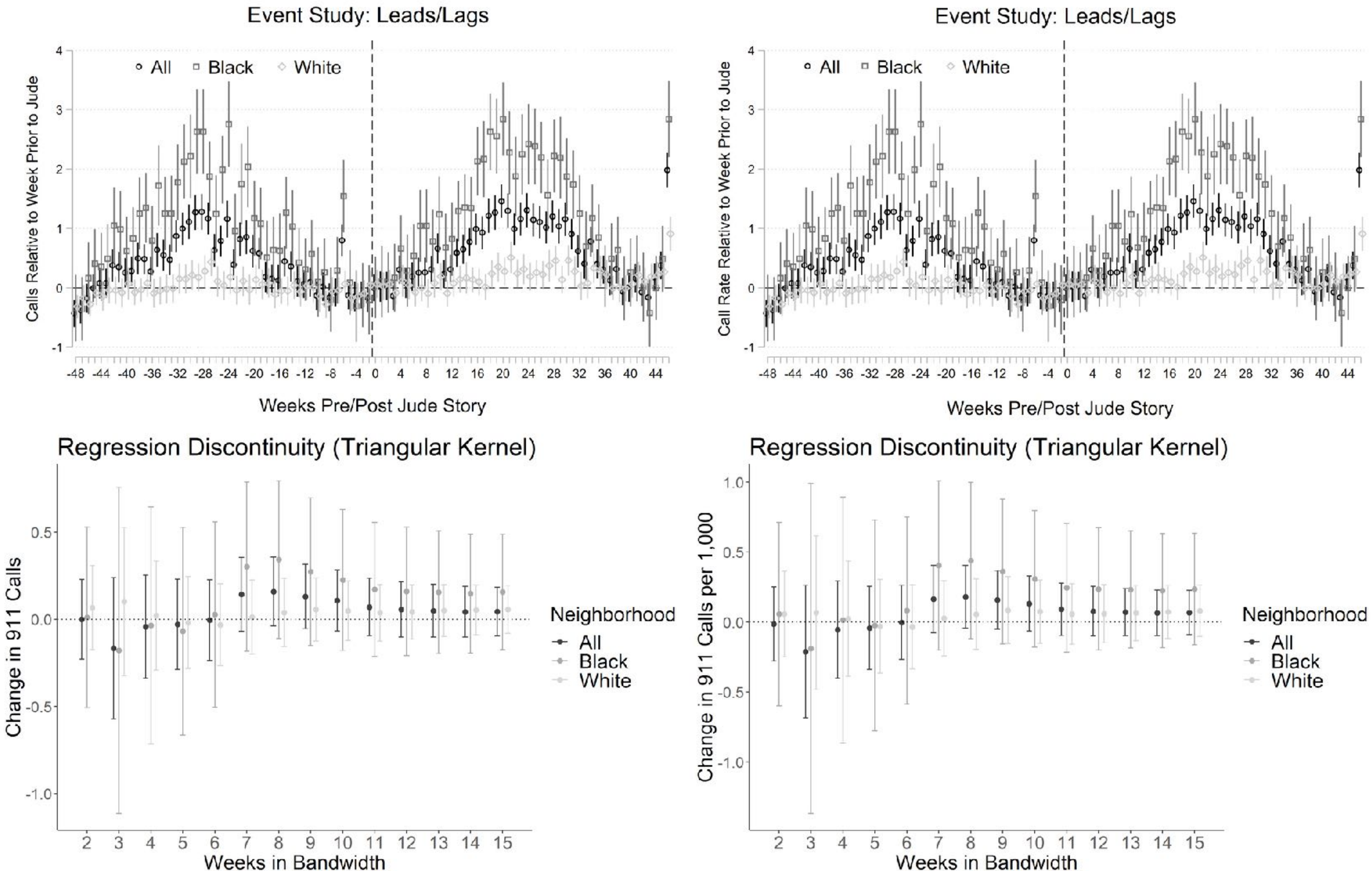
Desmond, Papachristos & Kirk (2016)


All Neighborhoods
911 Calls Per CBG
4.5
4.0
3.5
3.0
2.5
Final Week, 2005
Final Week, 2004
-48
-40
-32
-24
-16
-8
0
8
16
24
32
40
48
Weeks Pre/Post Jude Story

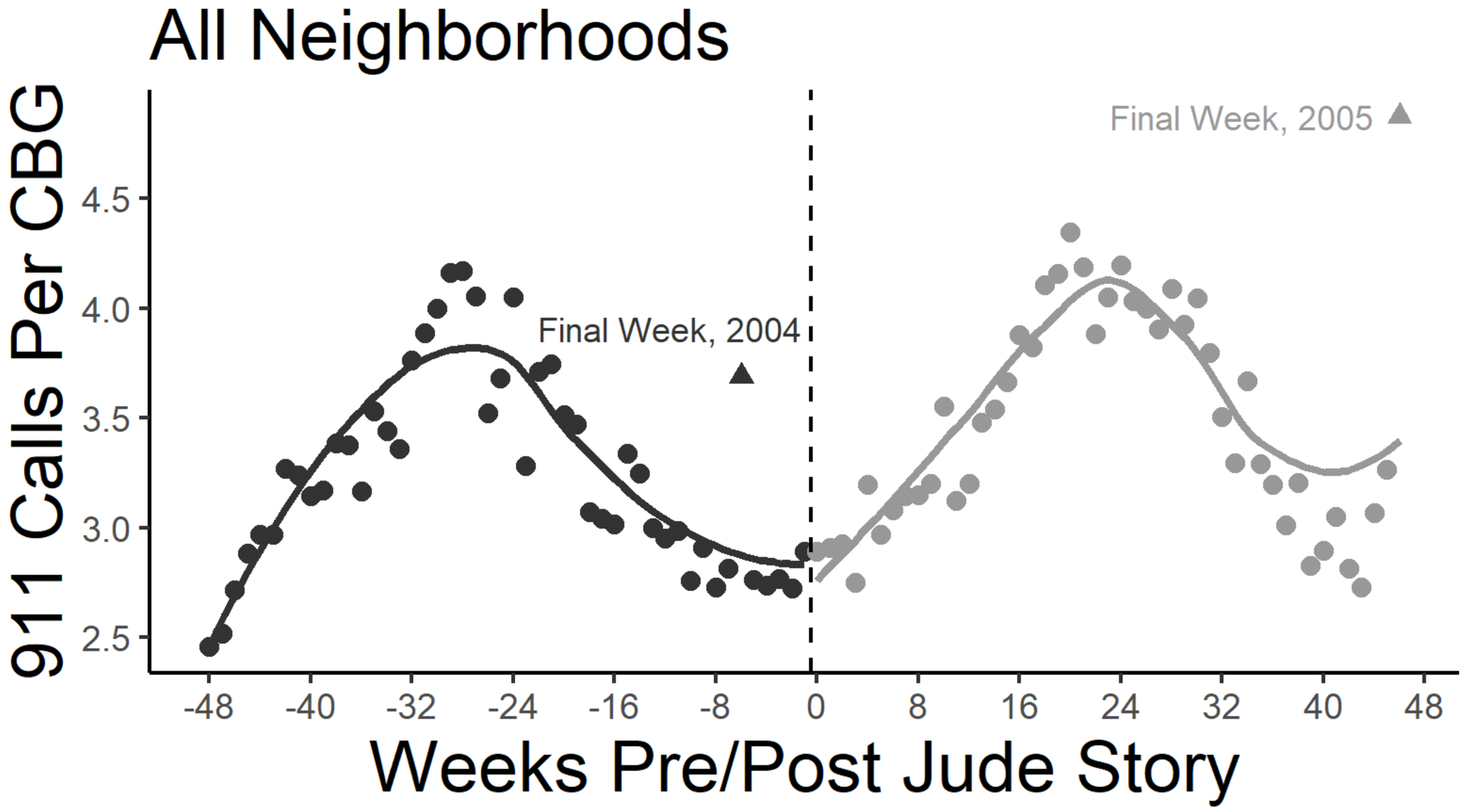
Zoorob (2020)


Raw Data
Final Week, 2005
Calls per Block Group
Weeks Pre/Post Jude
Linear Pre; Linear/Quadratic Post
With Final Week
Yes (DPK Model)
No
Predicted Calls per Block Group
Weeks Pre/Post Jude
Linear/Quadratic Pre; Linear/Quadratic Post
With Final Week
Yes
No
Predicted Calls per Block Group
Weeks Pre/Post Jude
Linear/Quadratic Pre; Linear Change Post
With Final Week
Yes
No
Predicted Calls per Block Group
Weeks Pre/Post Jude

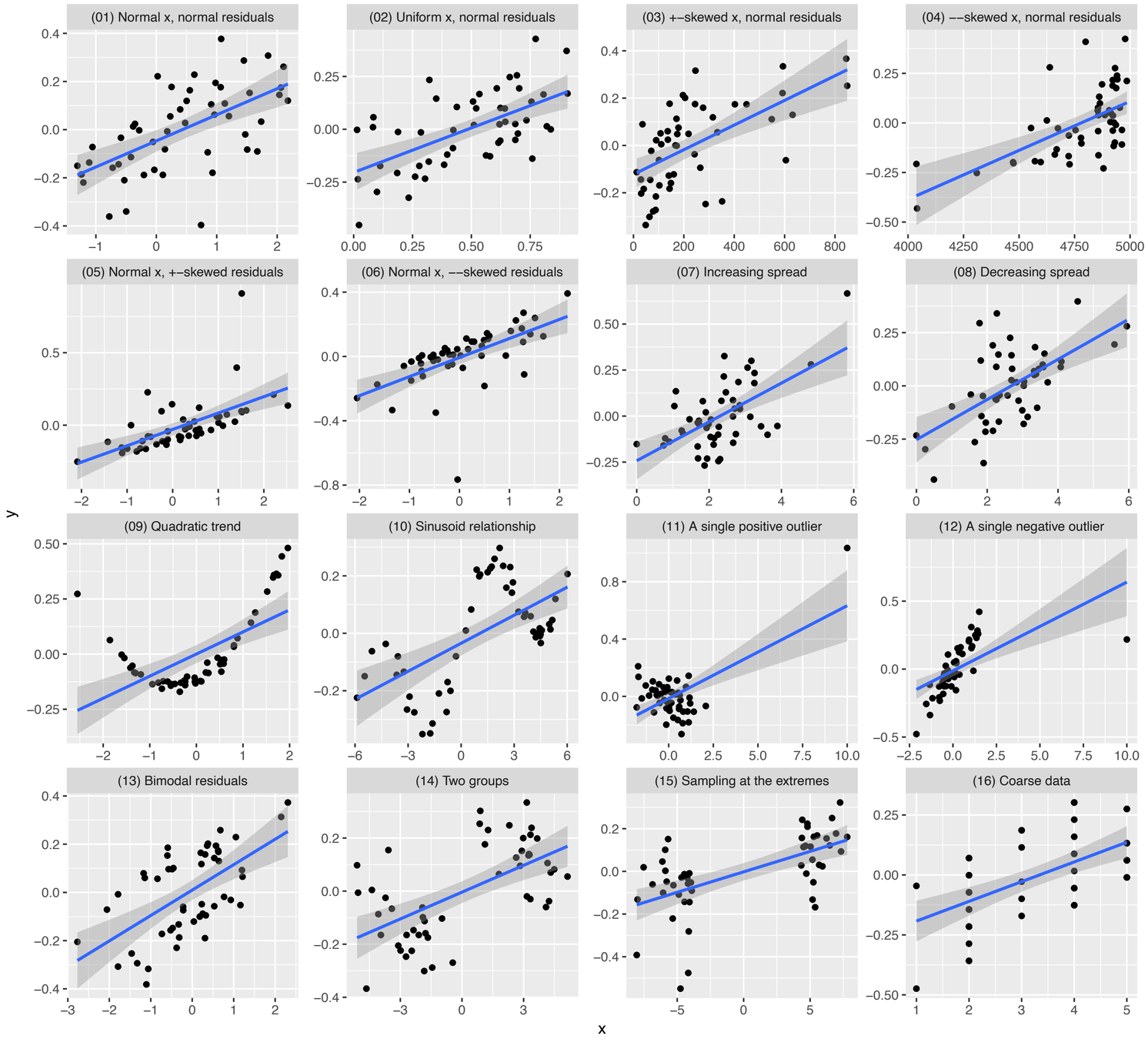
Zoorob (2020)

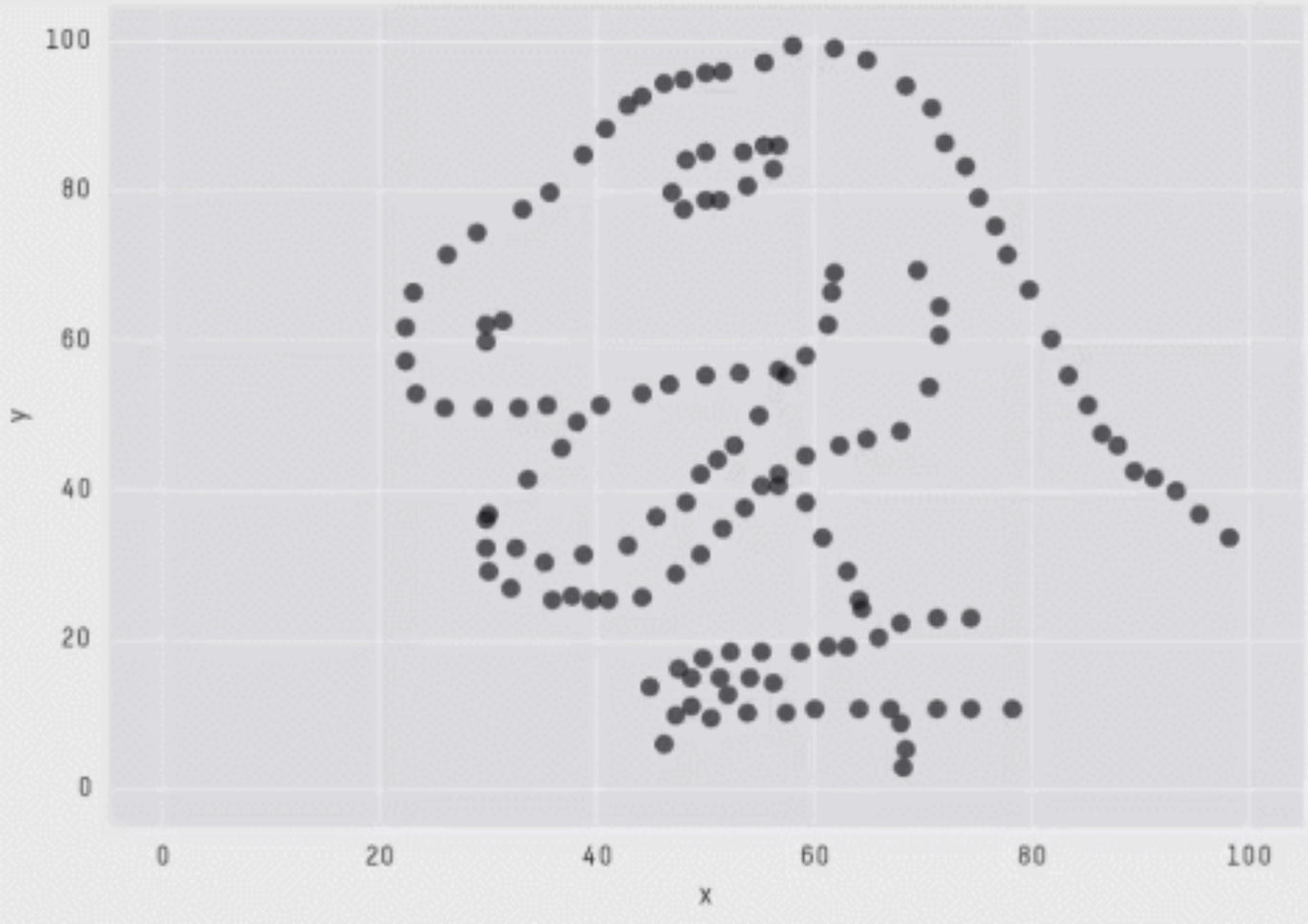

Event Study: Leads/Lags
All
Black
White
Calls Relative to Week Prior to Jude
Weeks Pre/Post Jude Story
Event Study: Leads/Lags
Call Rate Relative to Week Prior to Jude
Weeks Pre/Post Jude Story
Regression Discontinuity (Triangular Kernel)
Change in 911 Calls
Weeks in Bandwidth
Neighborhood
All
Black
White
Regression Discontinuity (Triangular Kernel)
Change in 911 Calls per 1,000
Weeks in Bandwidth
Neighborhood
All
Black
White

Zoorob (2020)


All Neighborhoods
911 Calls Per CBG
4.5
4.0
3.5
3.0
2.5
Final Week, 2005
Final Week, 2004
-48
-40
-32
-24
-16
-8
0
8
16
24
32
40
48
Weeks Pre/Post Jude Story

Zoorob (2020)

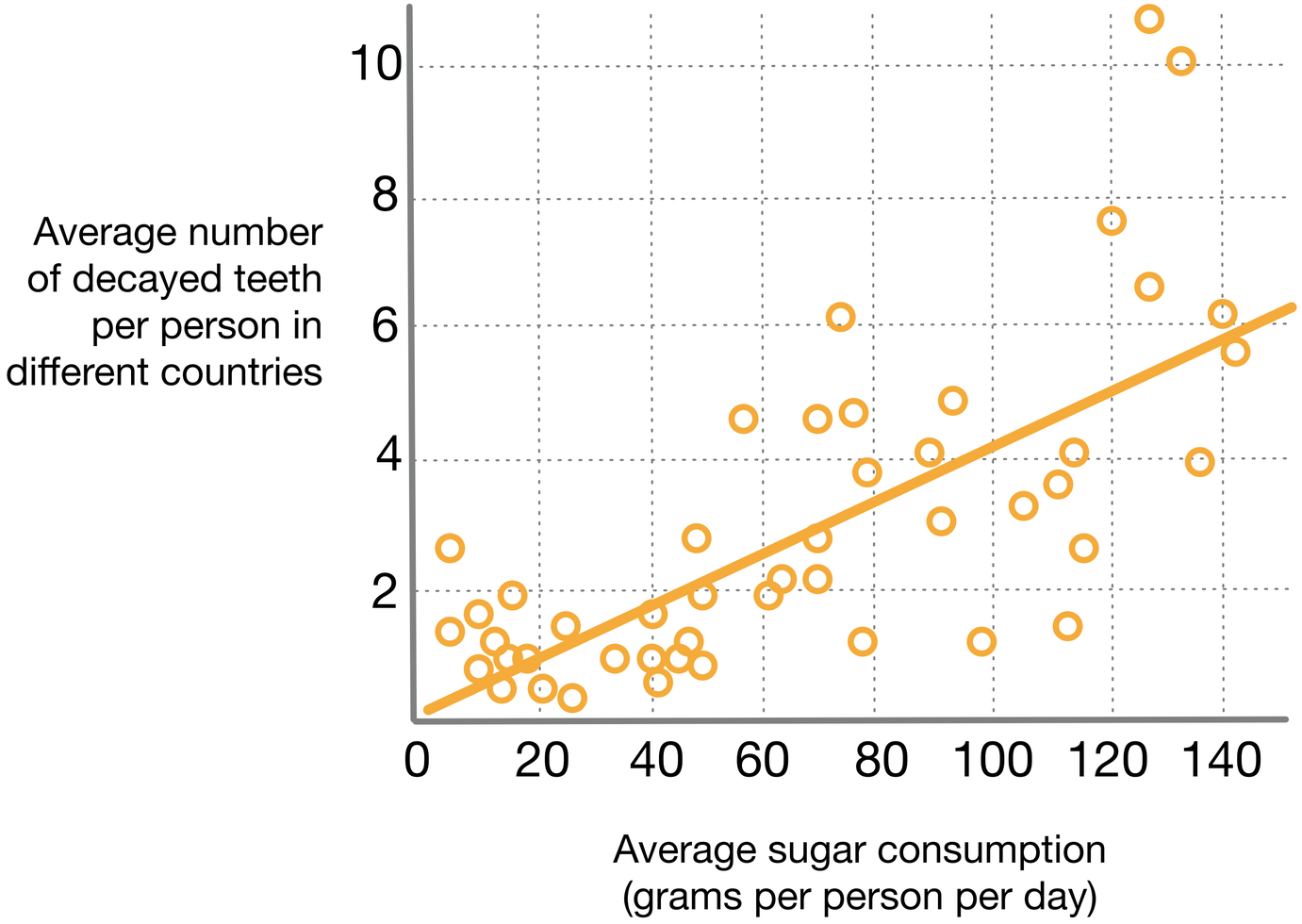

(01) Normal x, normal residuals
(02) Uniform x, normal residuals
(03) +−skewed x, normal residuals
(04) −−skewed x, normal residuals
(05) Normal x, +−skewed residuals
(06) Normal x, −−skewed residuals
(07) Increasing spread
(08) Decreasing spread
(09) Quadratic trend
(10) Sinusoid relationship
(11) A single positive outlier
(12) A single negative outlier
(13) Bimodal residuals
(14) Two groups
(15) Sampling at the extremes
(16) Coarse data
y
x



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

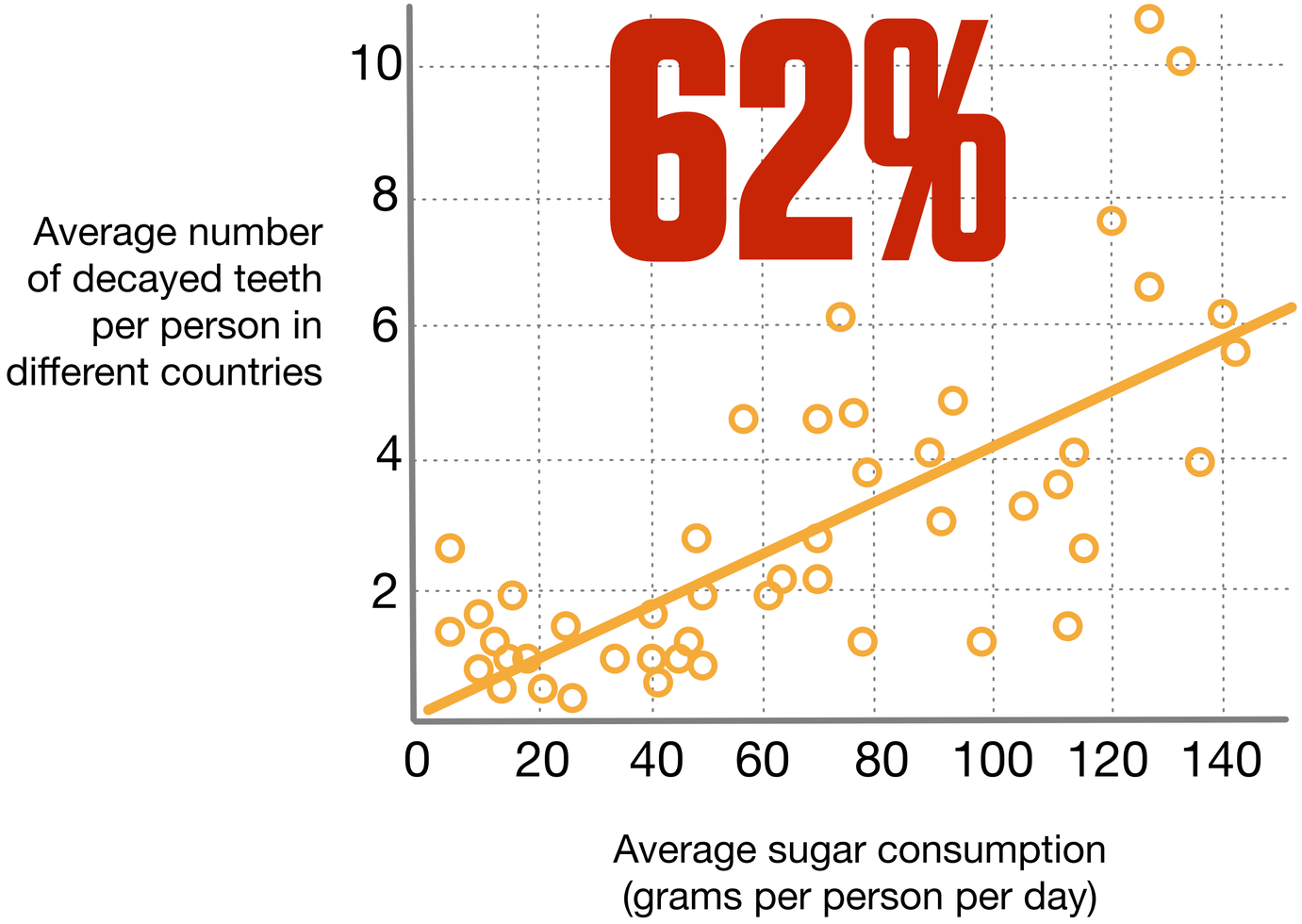
Cairo; Matejka & Fitzmaurice

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



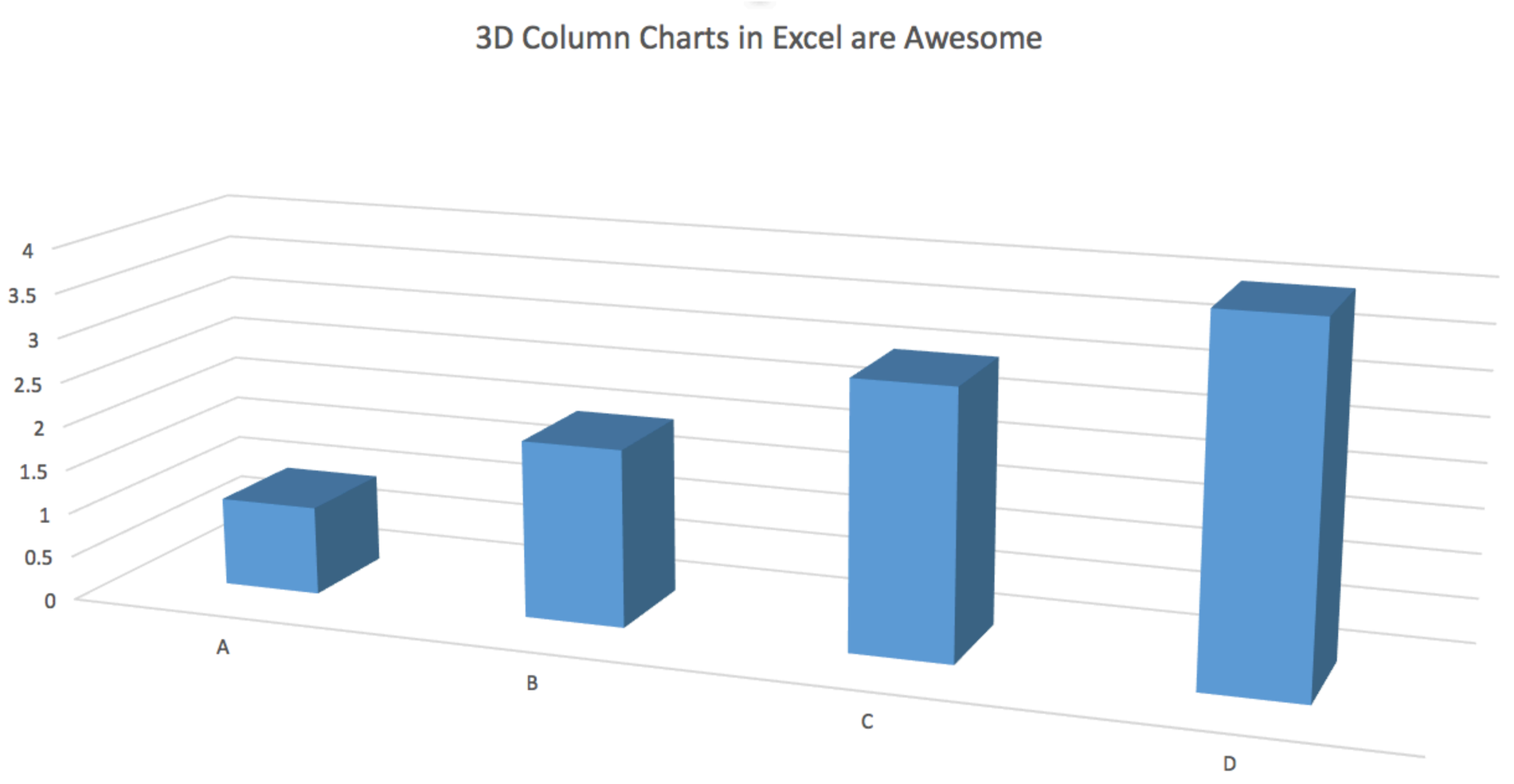
A. In recent years, the rate of cavities has increased in many countries
B. In some countries, people brush their teeth more frequently than in other countries
C. The more sugar people eat, the more likely they are to get cavities
D. In recent years, the consumption of sugar has increased in many countries

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

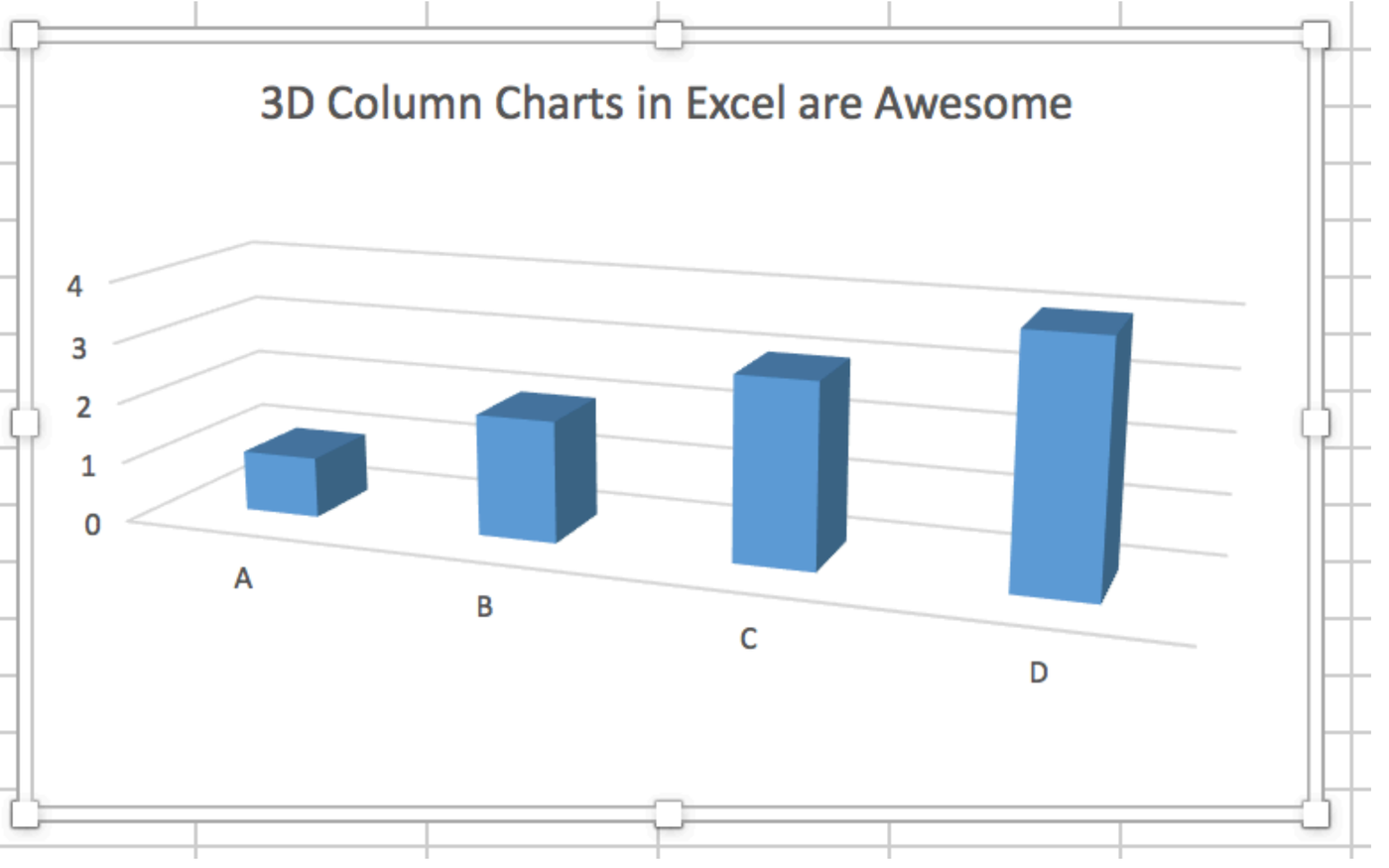


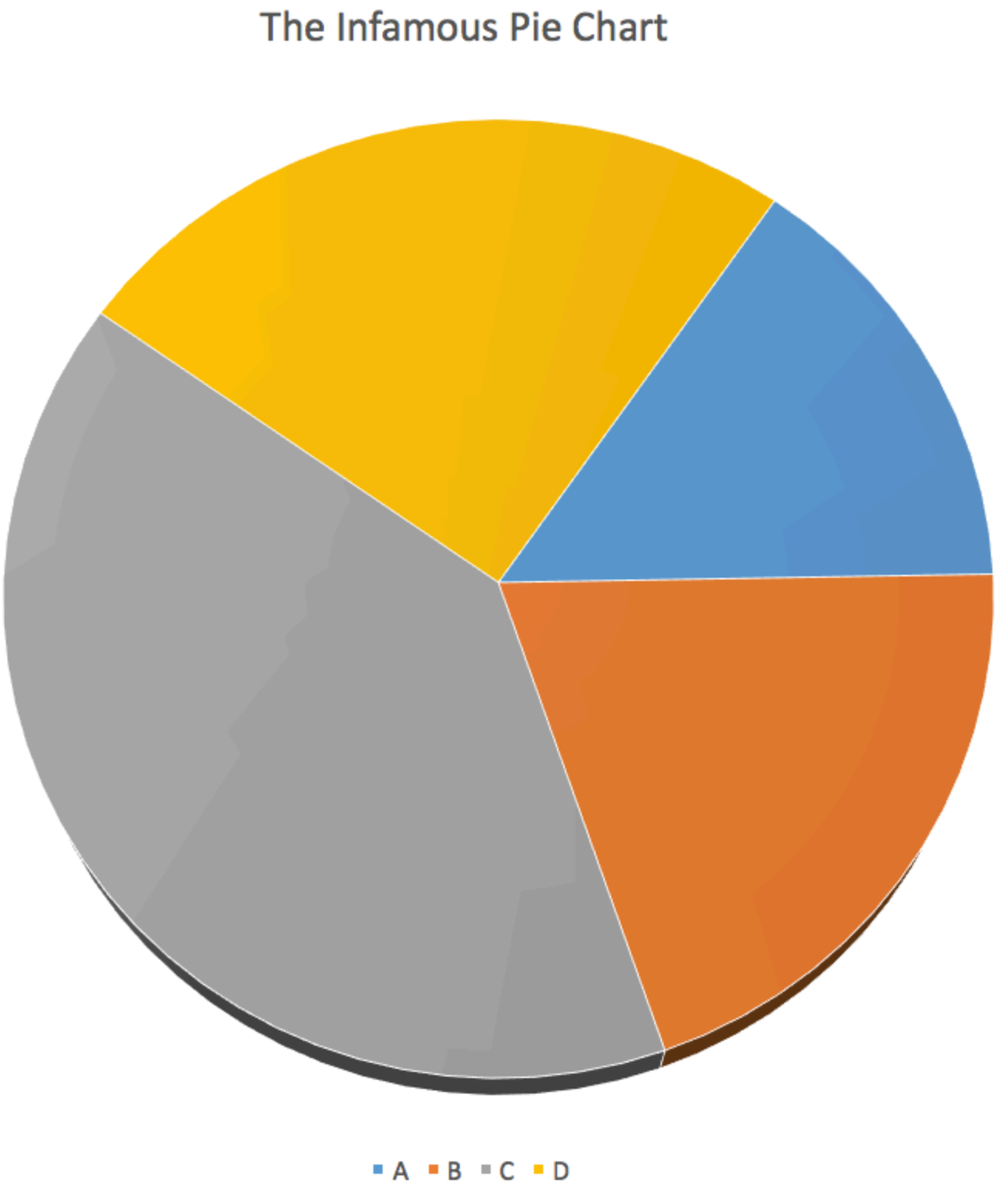
A. In recent years, the rate of cavities has increased in many countries
B. In some countries, people brush their teeth more frequently than in other countries
C. The more sugar people eat, the more likely they are to get cavities
D. In recent years, the consumption of sugar has increased in many countries

NOT SEEING THINGS

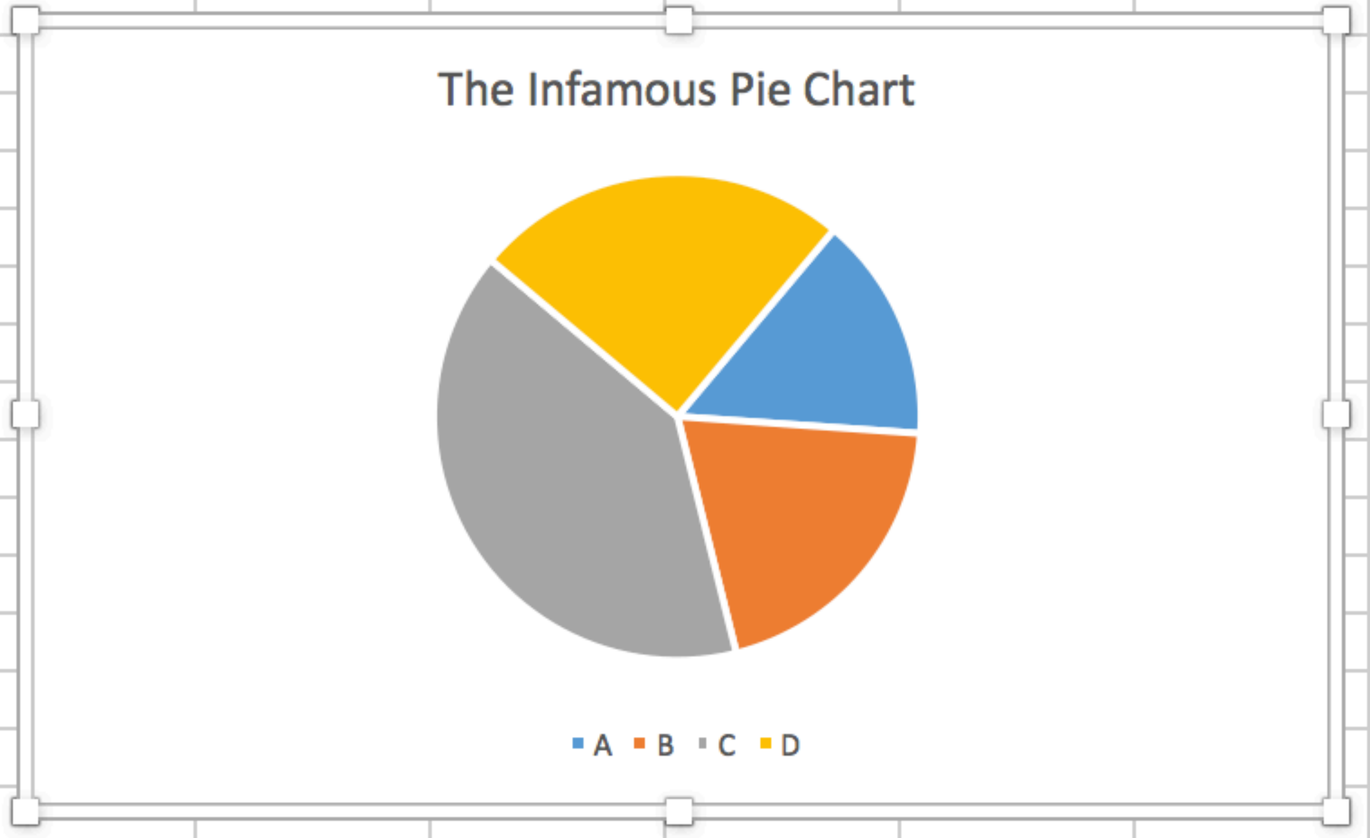

3D Column Charts in Excel are Awesome
4
3.5
3
2.5
2
1.5
1
0.5
0
A
B
C
D

A	1
B	2
C	3
D	4




The Infamous Pie Chart
A
B
C
D

A	15
B	20
C	40
D	25
	100



BAD TASTE

BAD DATA

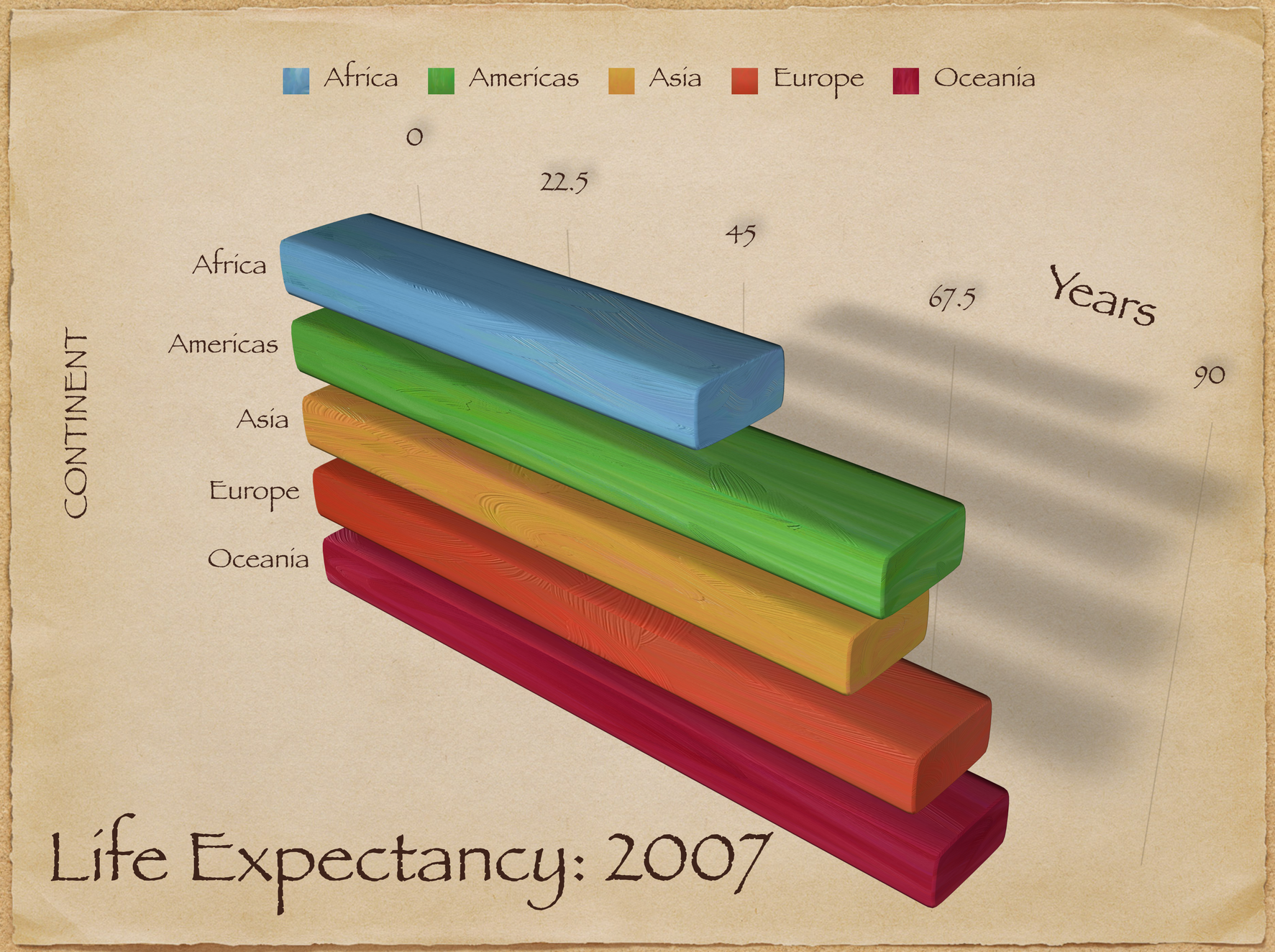
BAD PERCEPTION

TASTE

SIMPLIFY, SIMPLIFY?

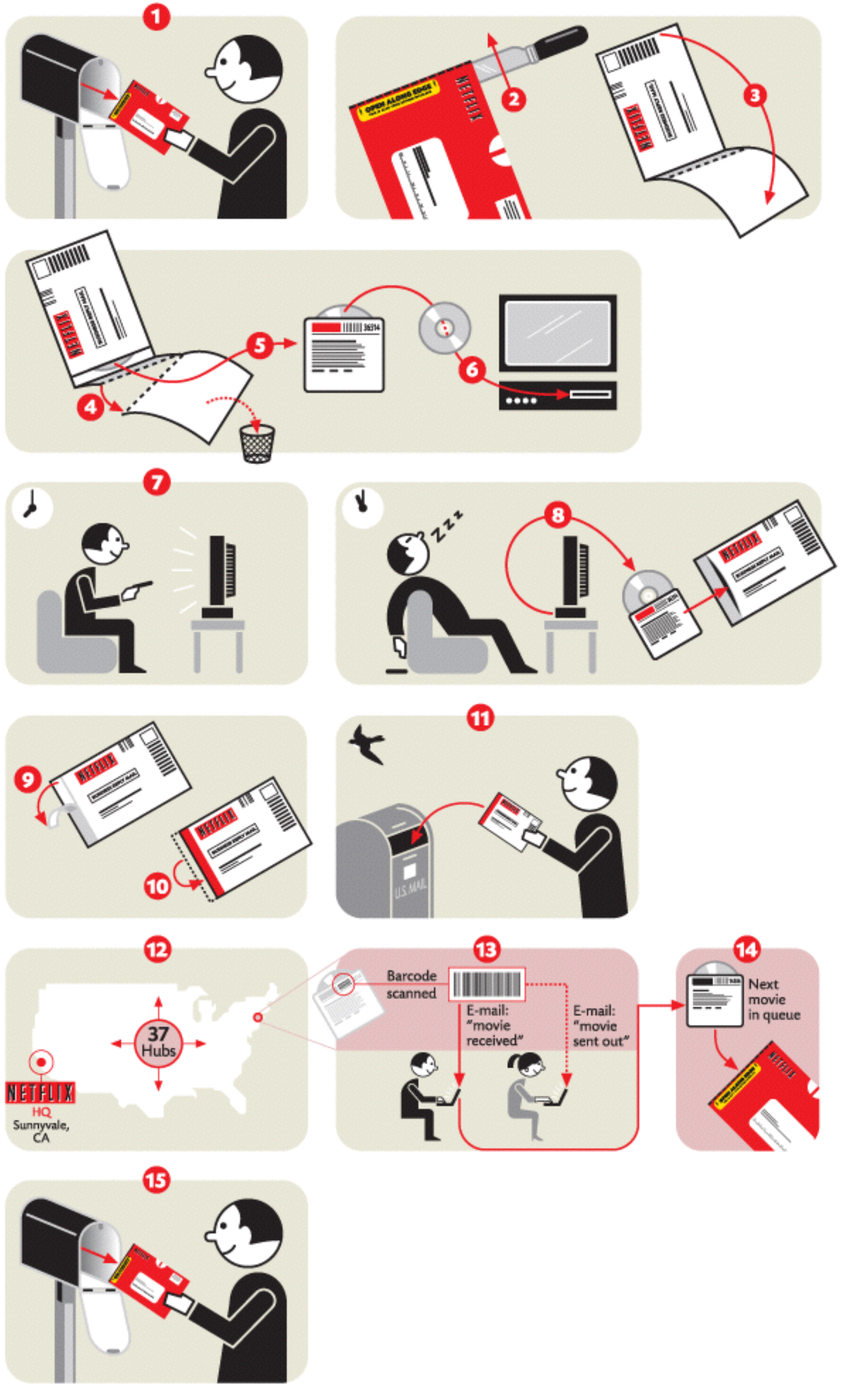
The Visual Display of Quantitative Information

EDWARD R. TUFTE


Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Africa
Americas
Asia
Europe
Oceania
Life Expectancy: 2007

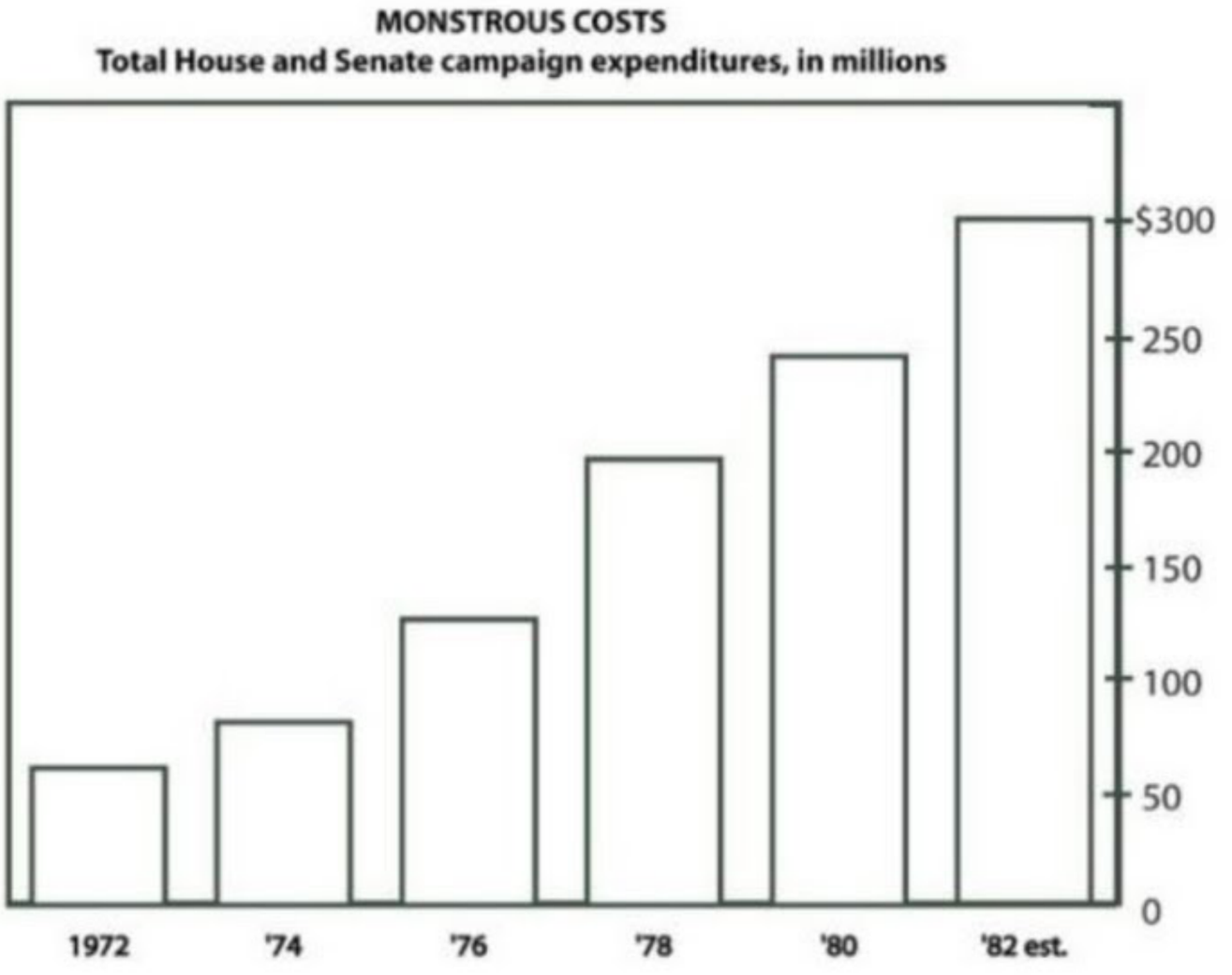
MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

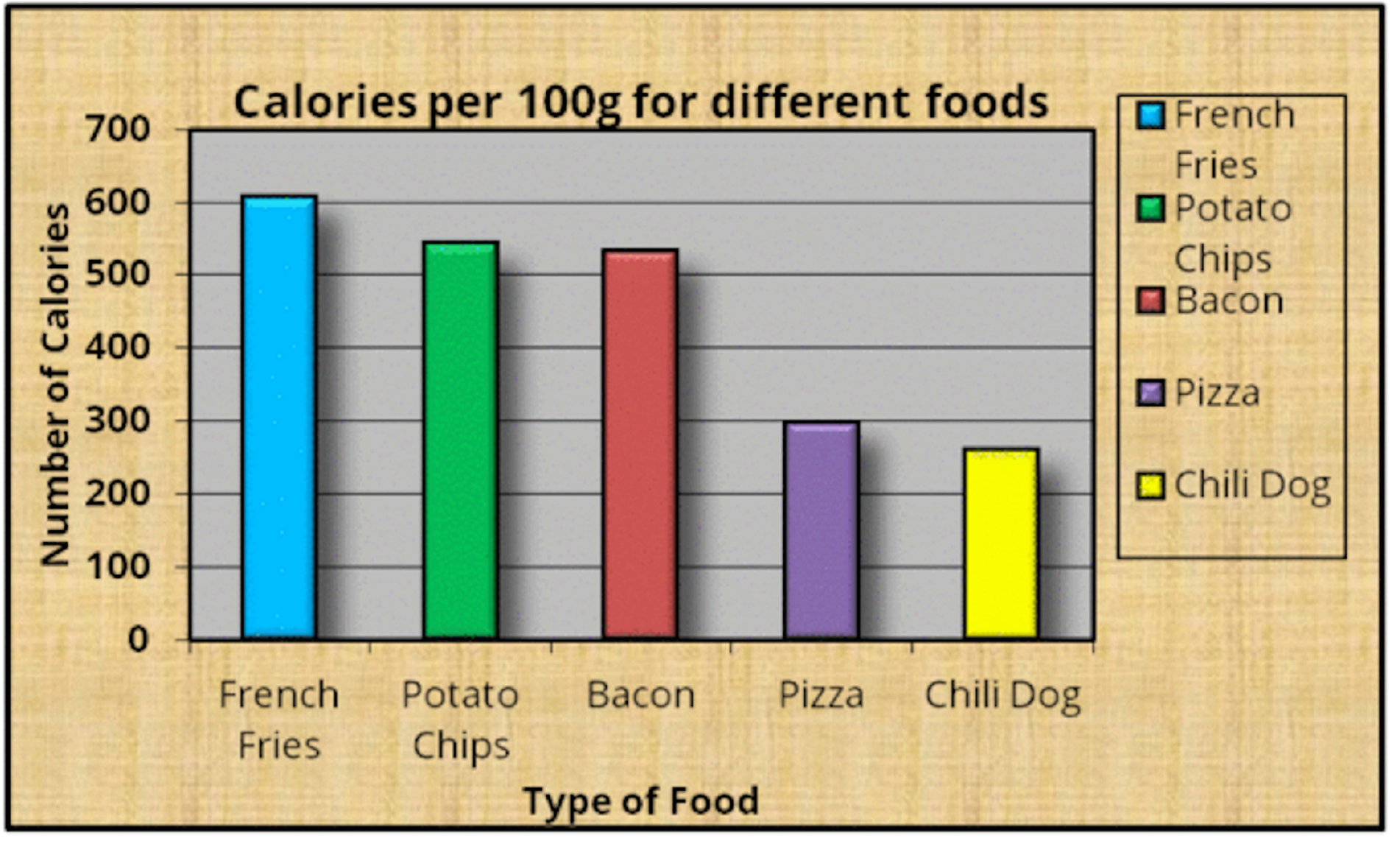
HOW NETFLIX WORKS



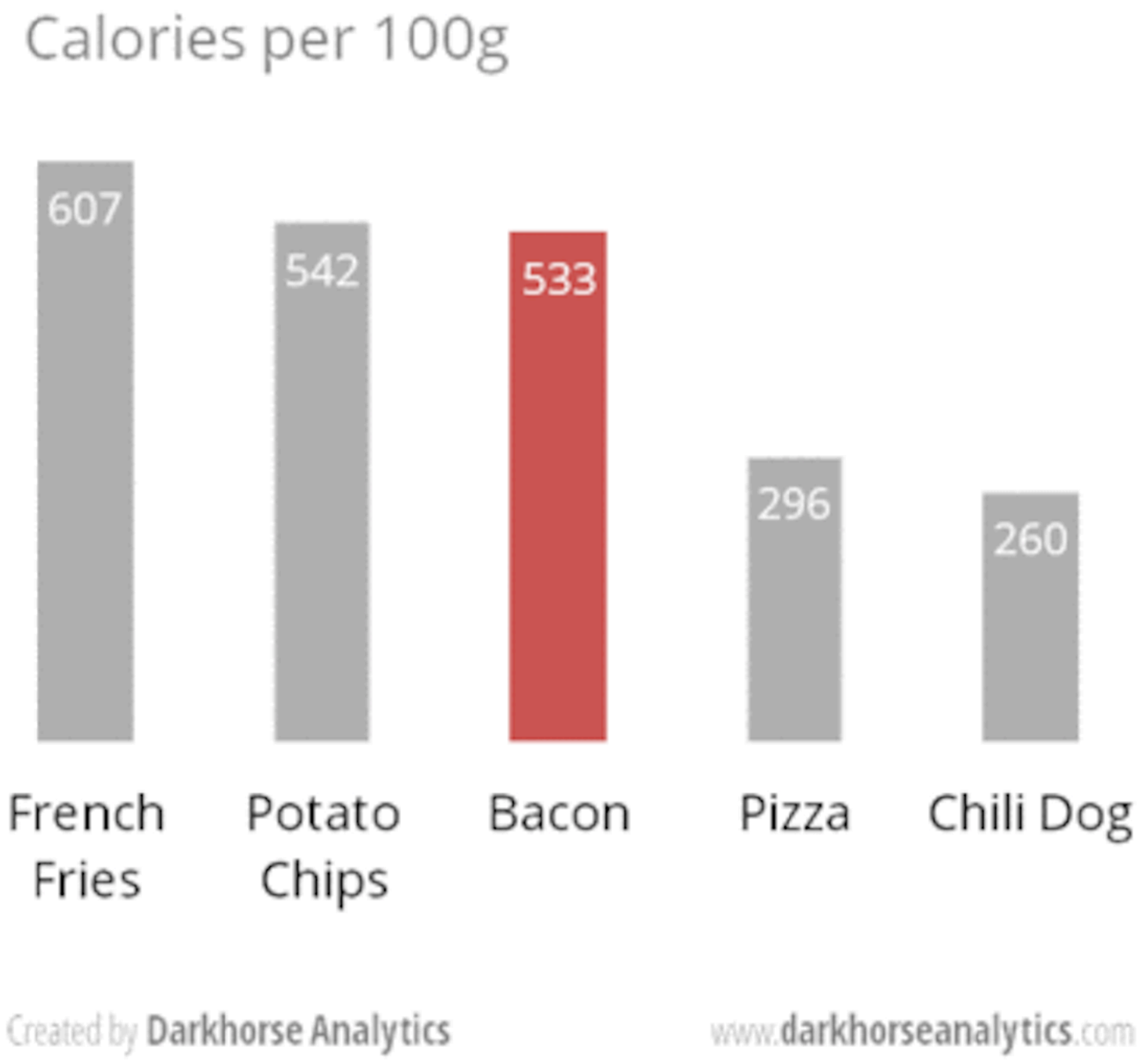
4
5
6
7

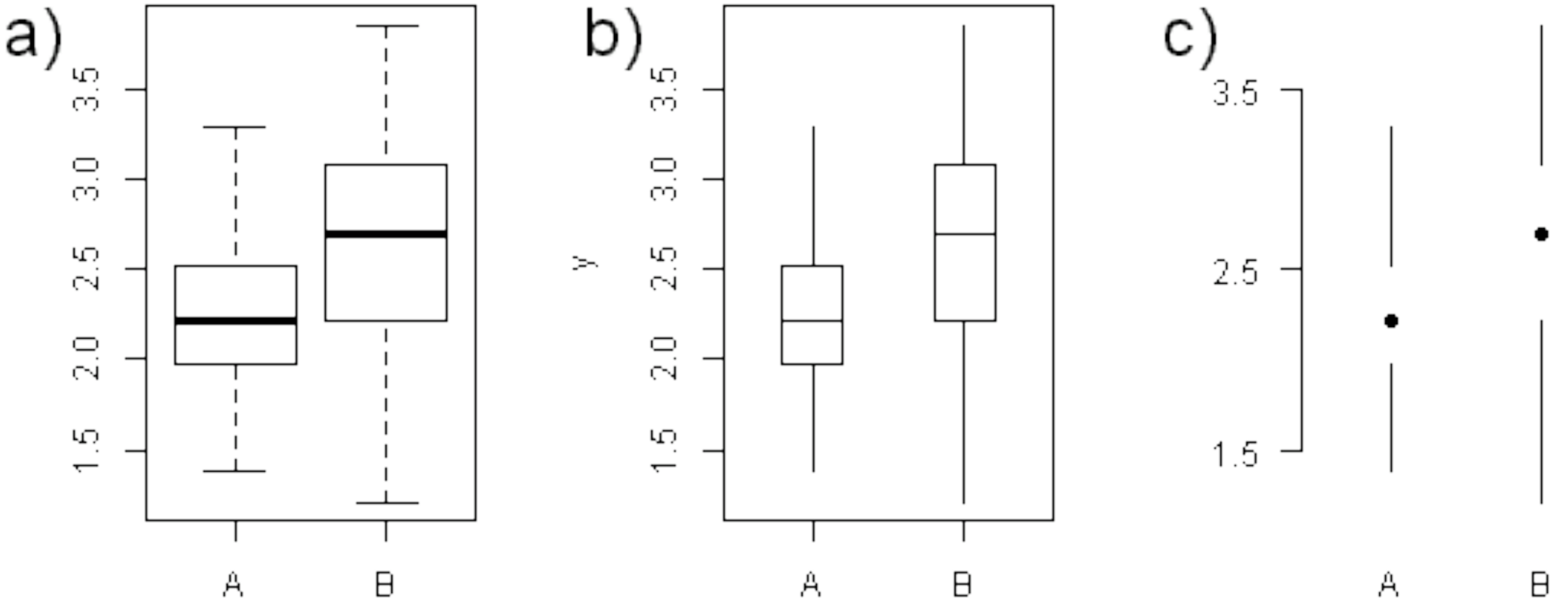
MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

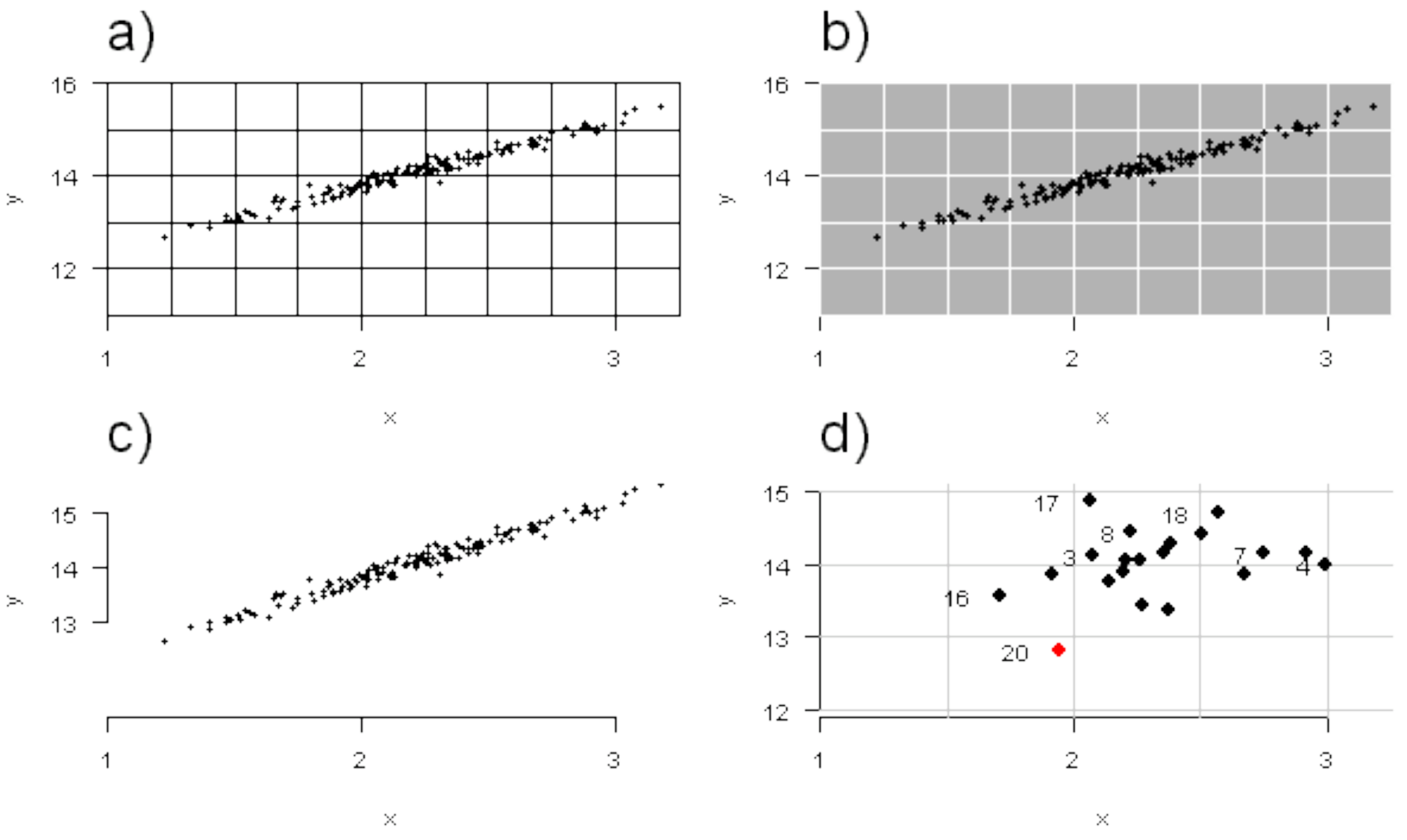

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
0
1972
'74
'76
'78
'80
'82 est.



Created by Darkhorse Analytics www.darkhorseanalytics.com

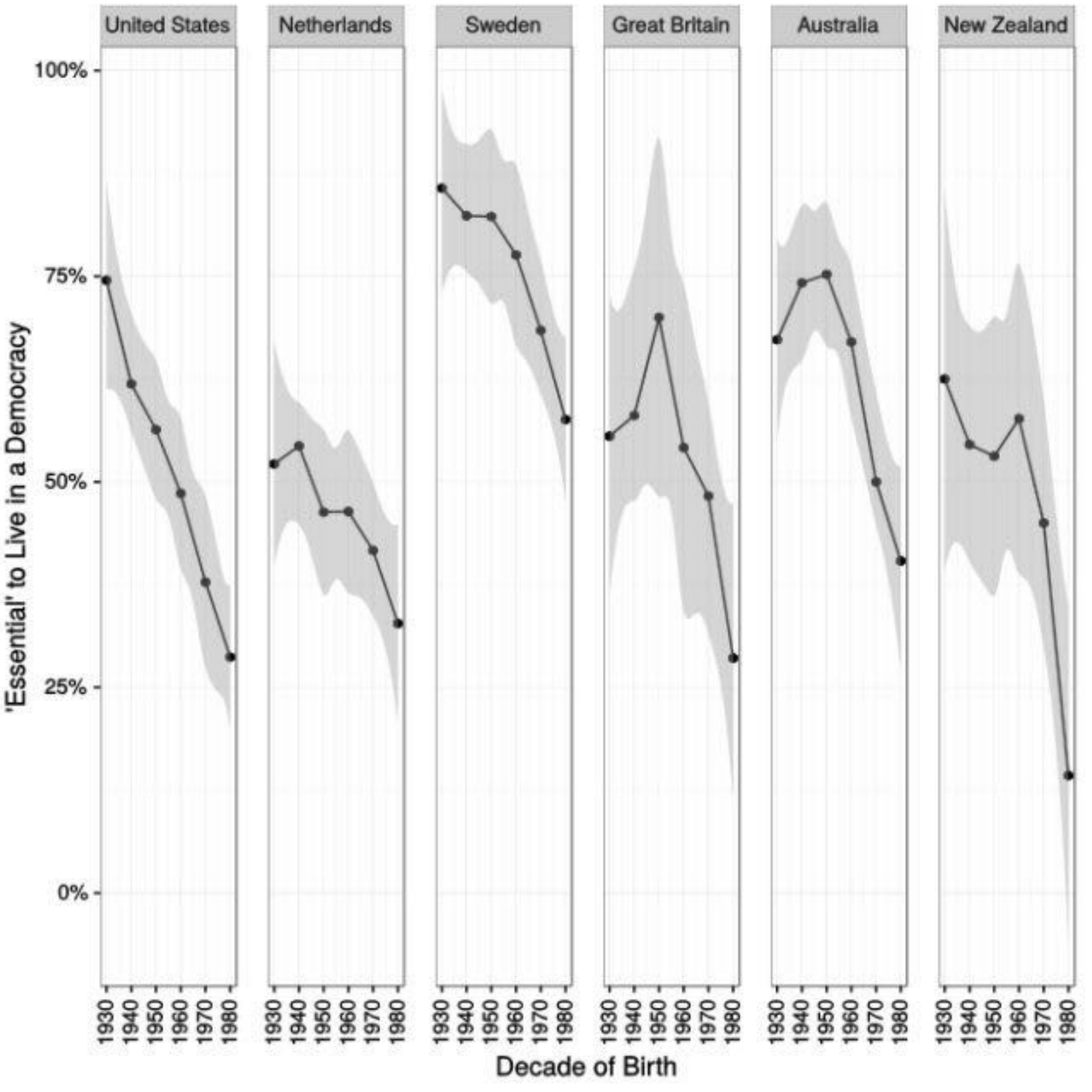

Calories per 100g
607
542
533
296
260
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Created by Darkhorse Analytics
www.darkhorseanalytics.com


a)
b)
c)
y
1.5
2.0
2.5
3.0
3.5
A
B


a)
b)
c)
d)
16
14
12
15
13
1
2
3
x
y
17
18
8
3
7
4
16
20

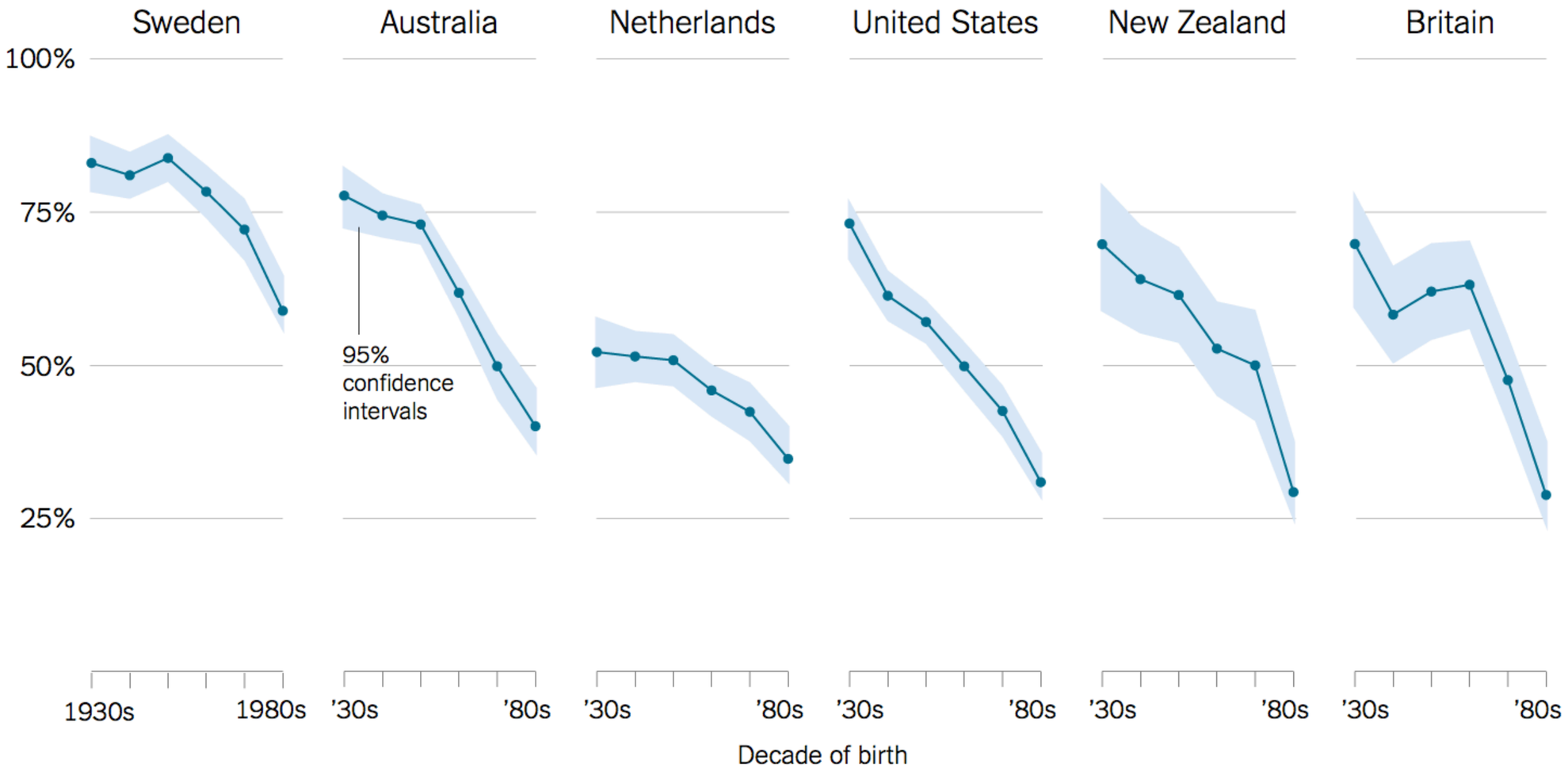
DATA

JUNK-FREE JUNK CHARTS

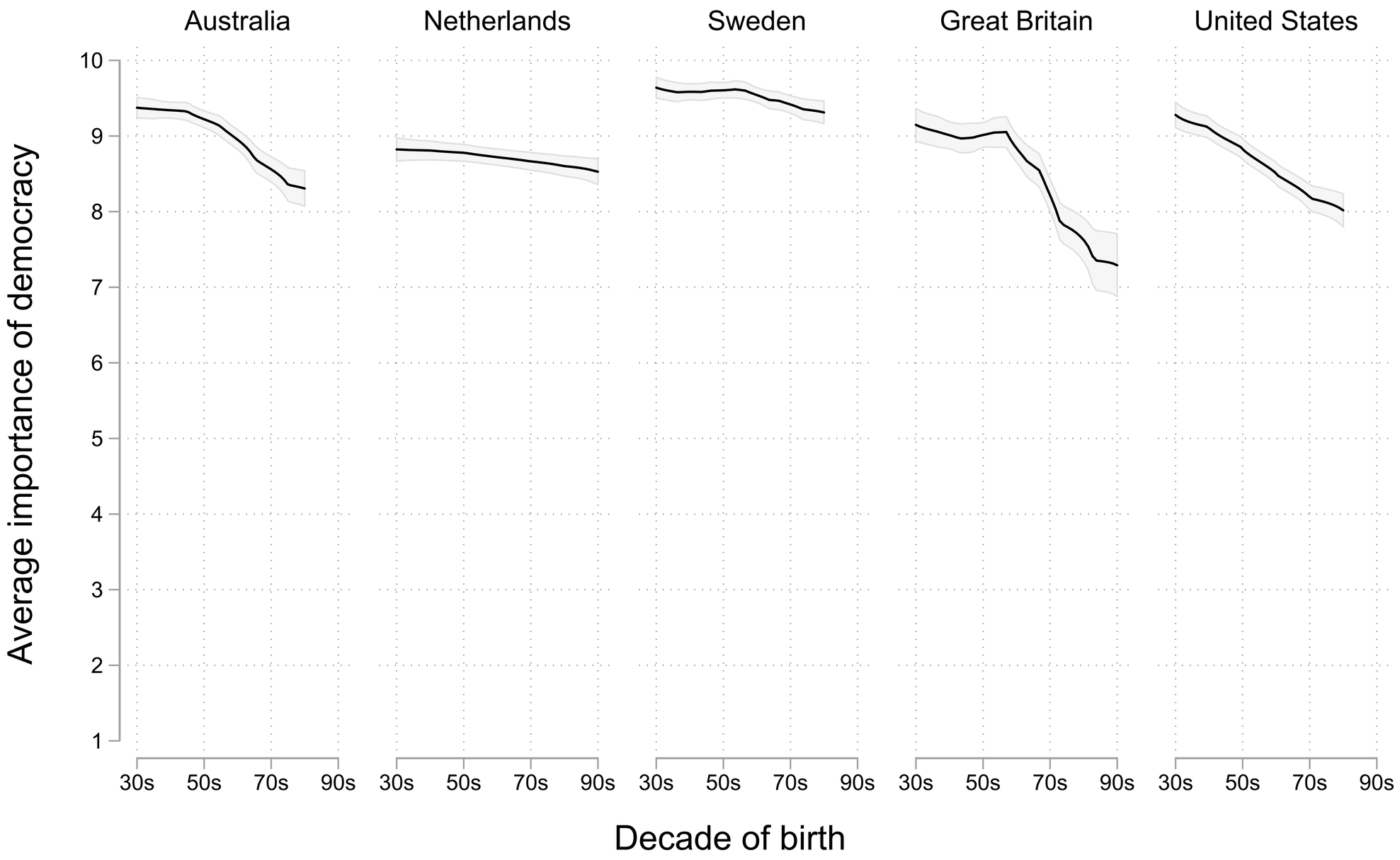


Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy

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

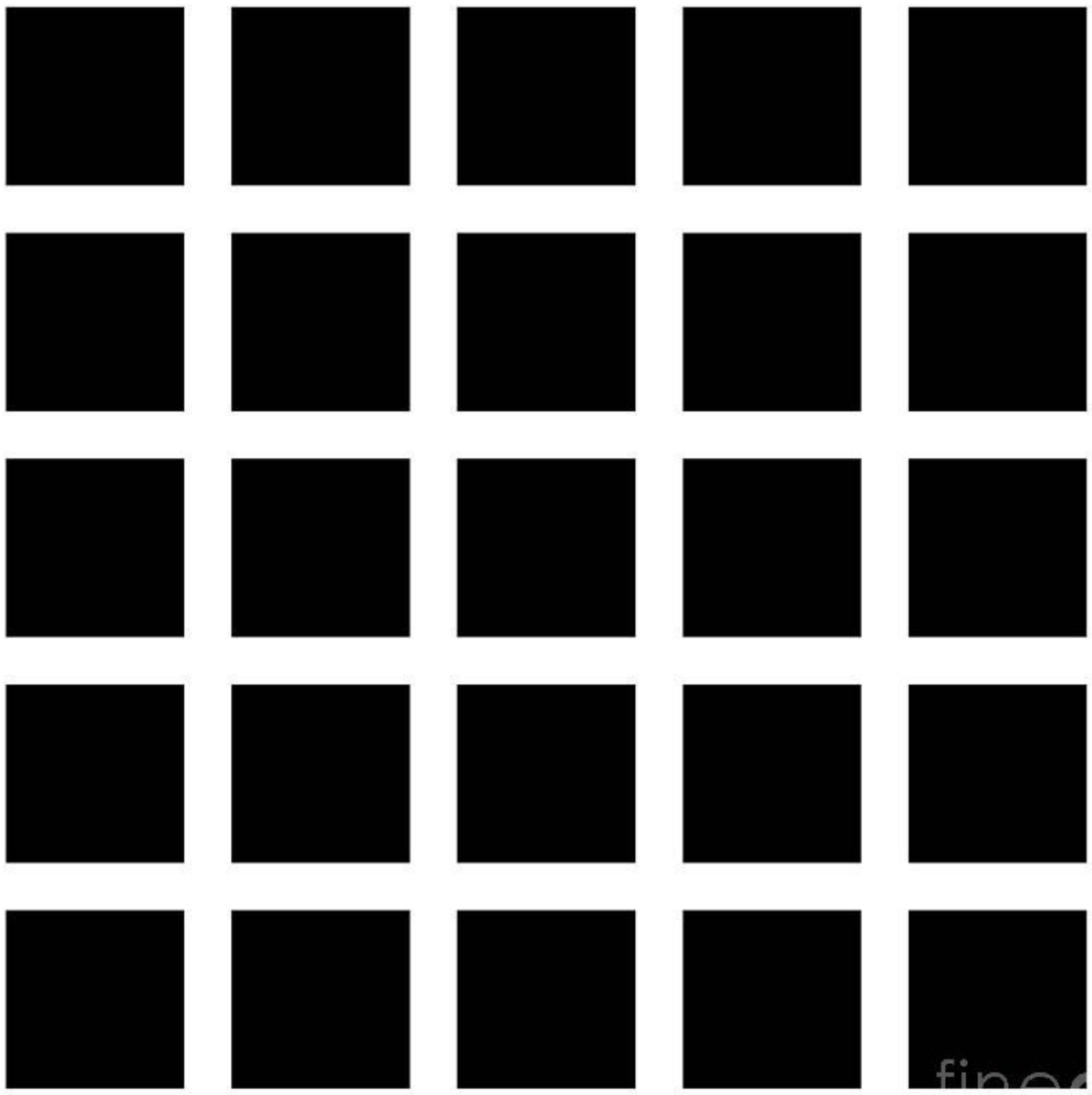


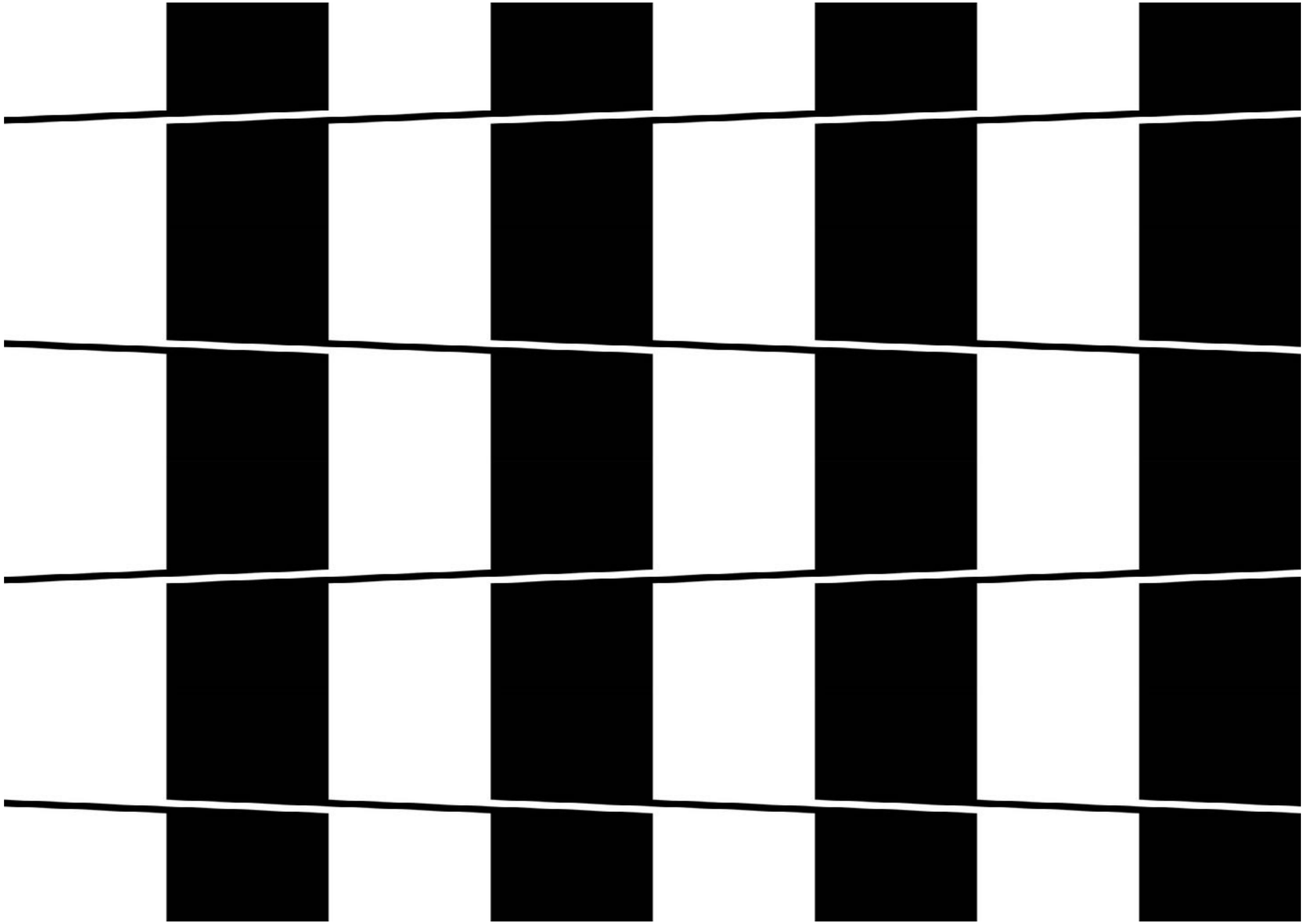
Graph by Erik Voeten, based on WVS 5

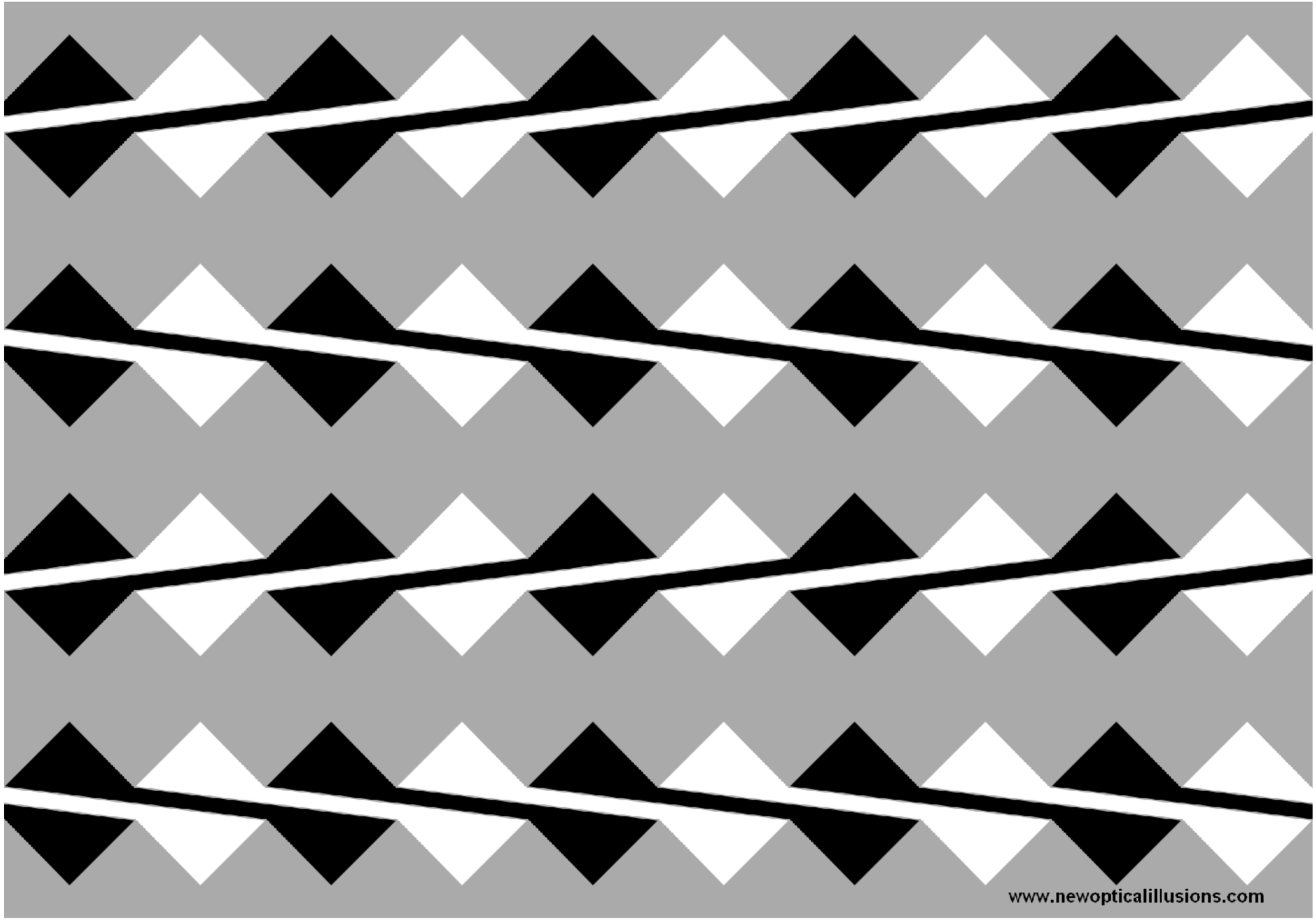
PERCEPTION

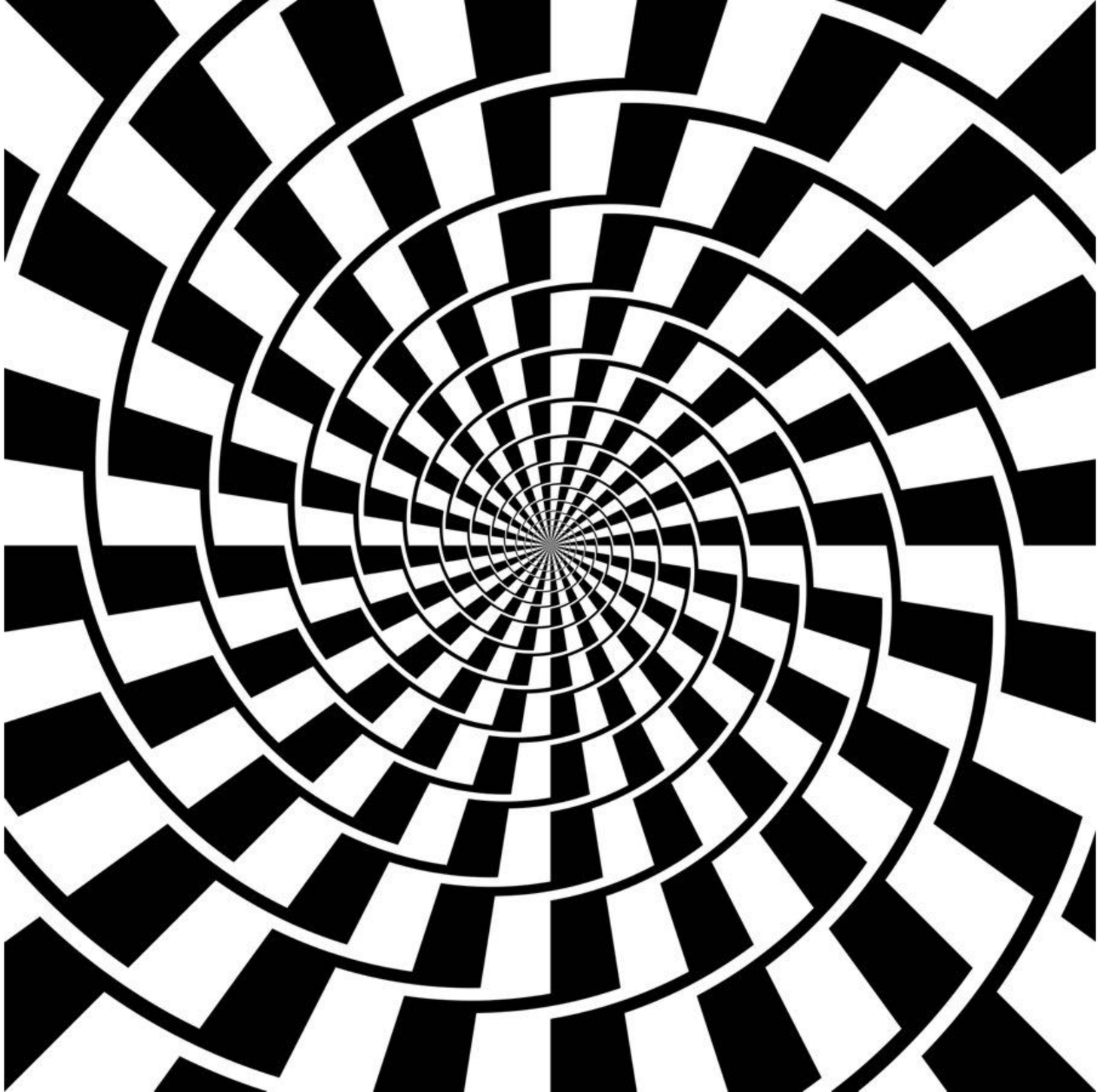
SIGHT AND INSIGHT

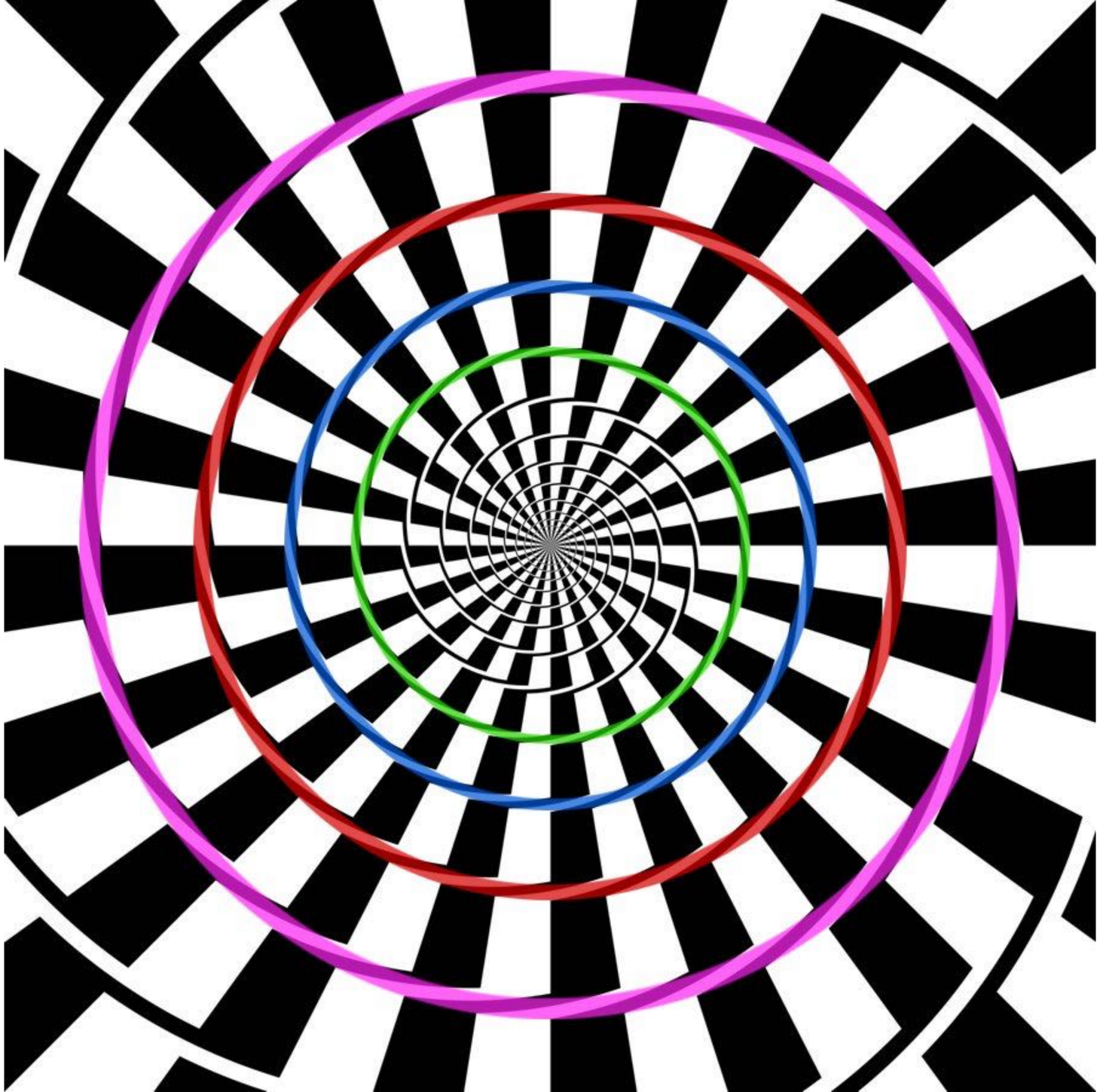
Edges & Contrasts



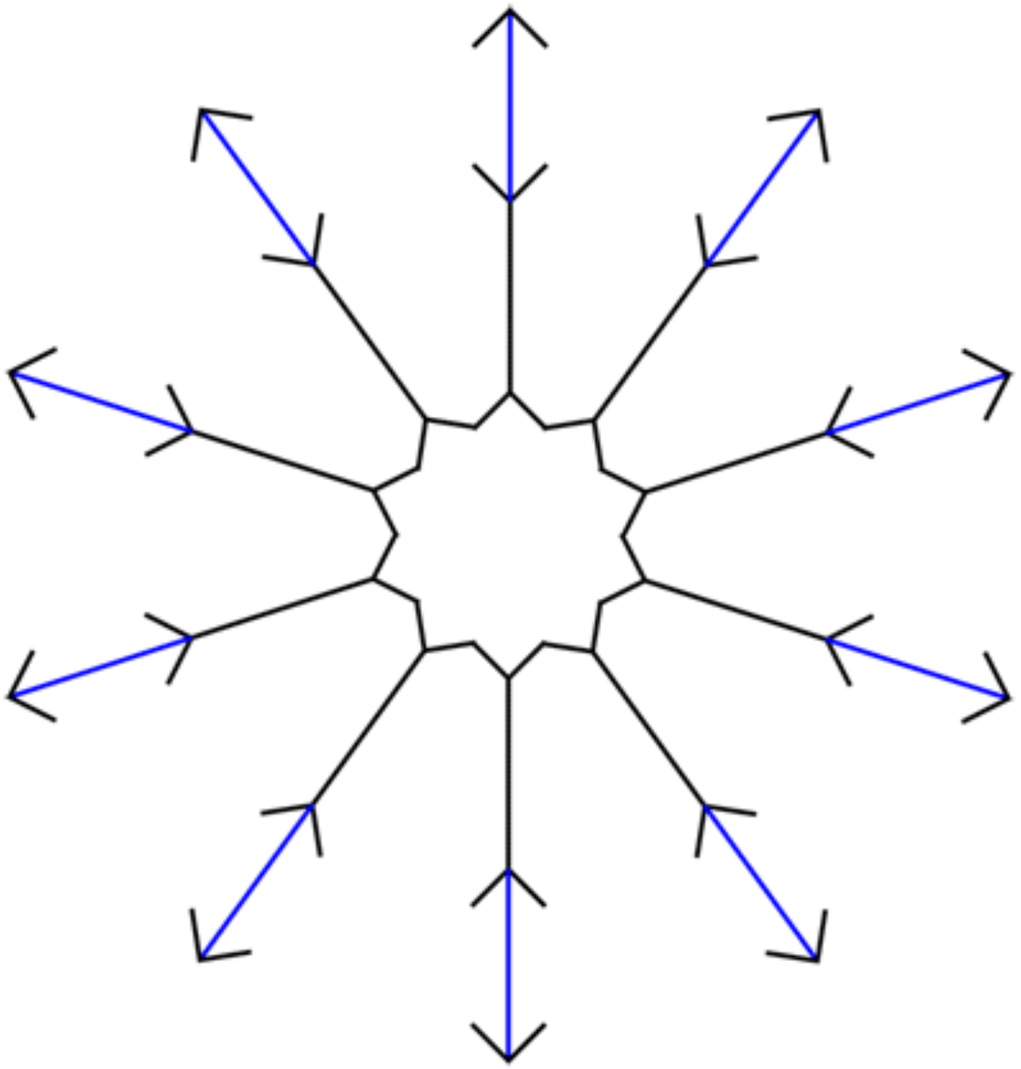



www.newopticalillusions.com

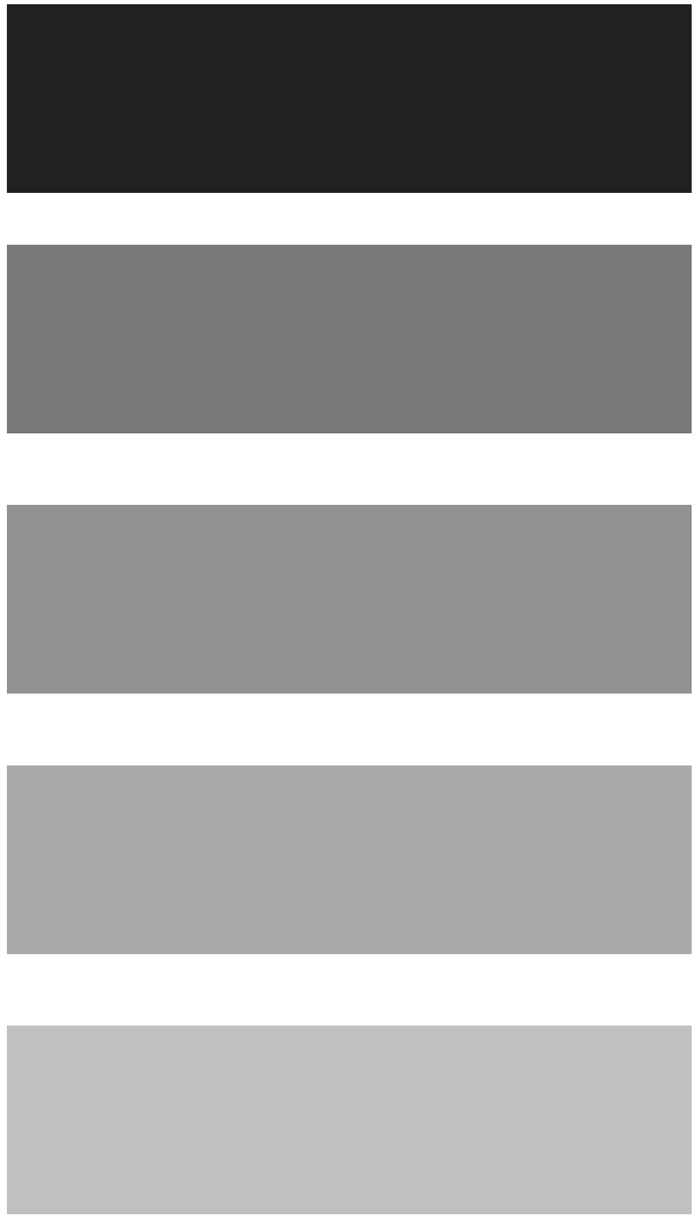


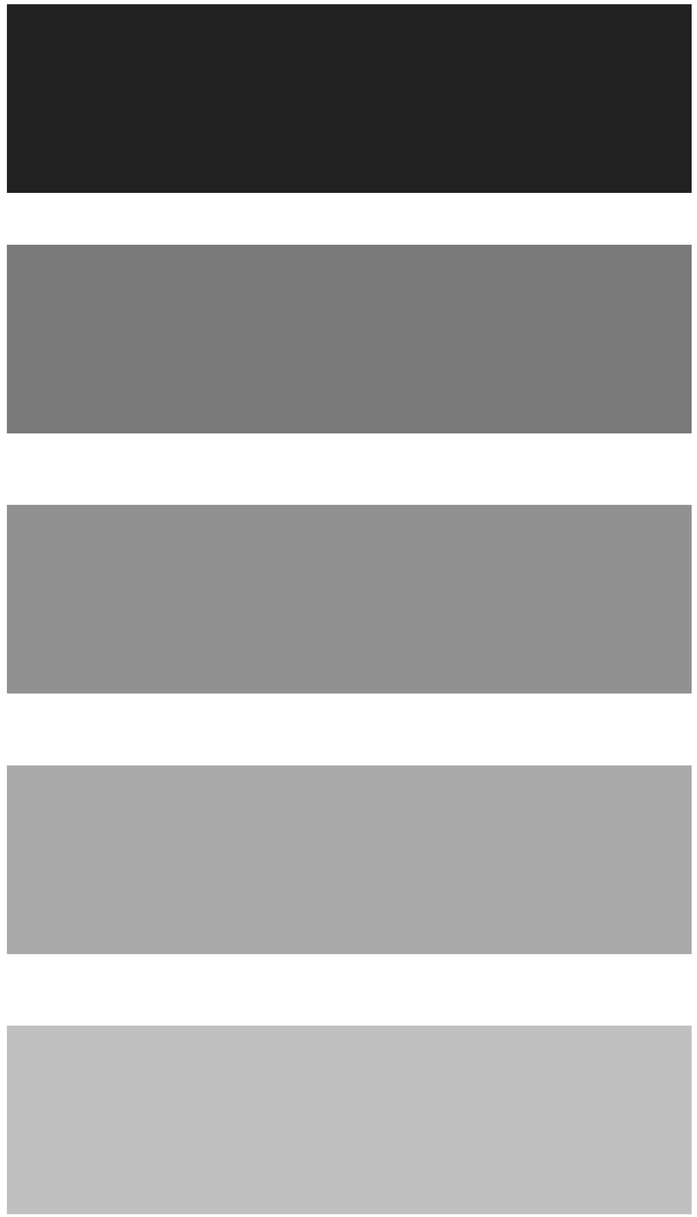


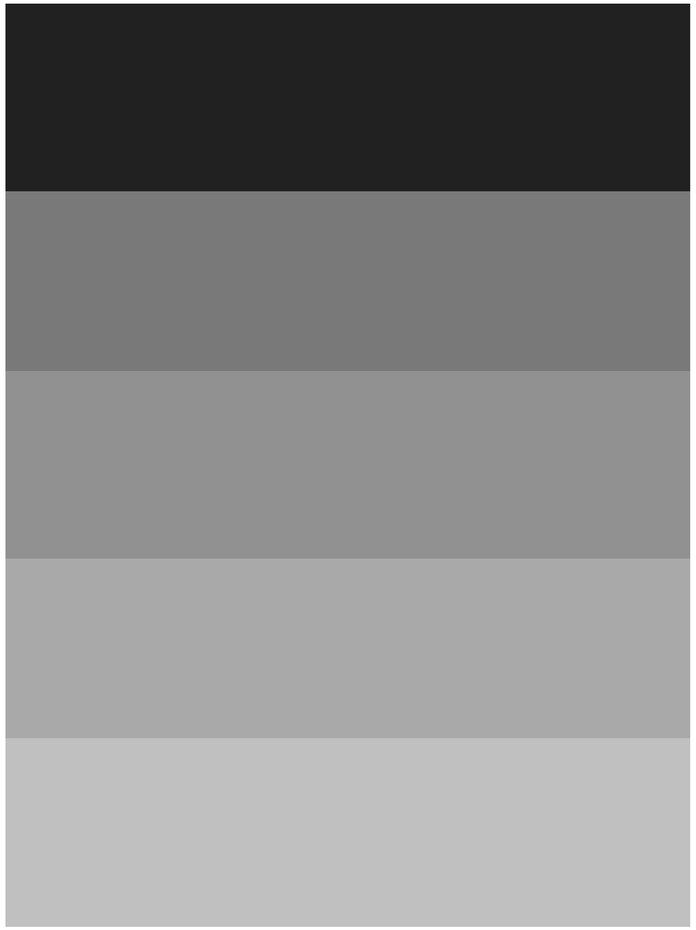
Sarcone's Dynamic Müller-Lyer Illusion

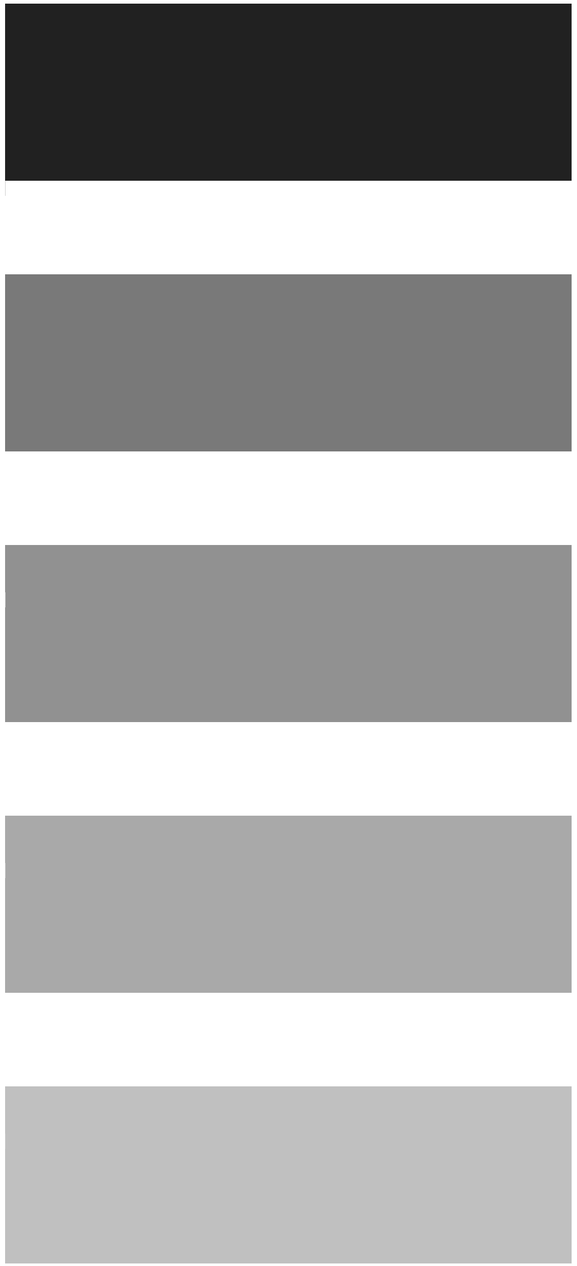


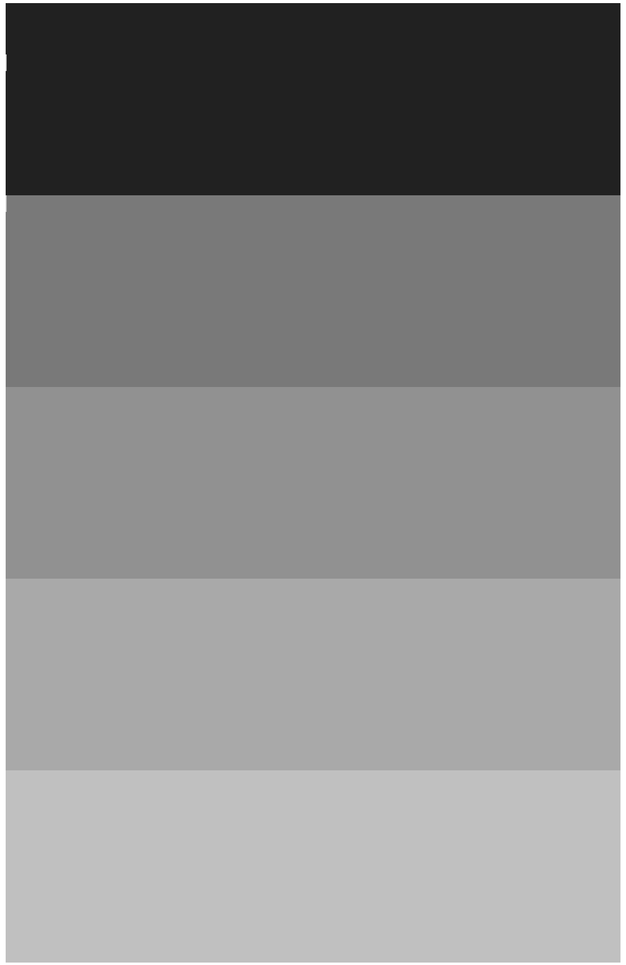
These are two perfectly
geometrical circles

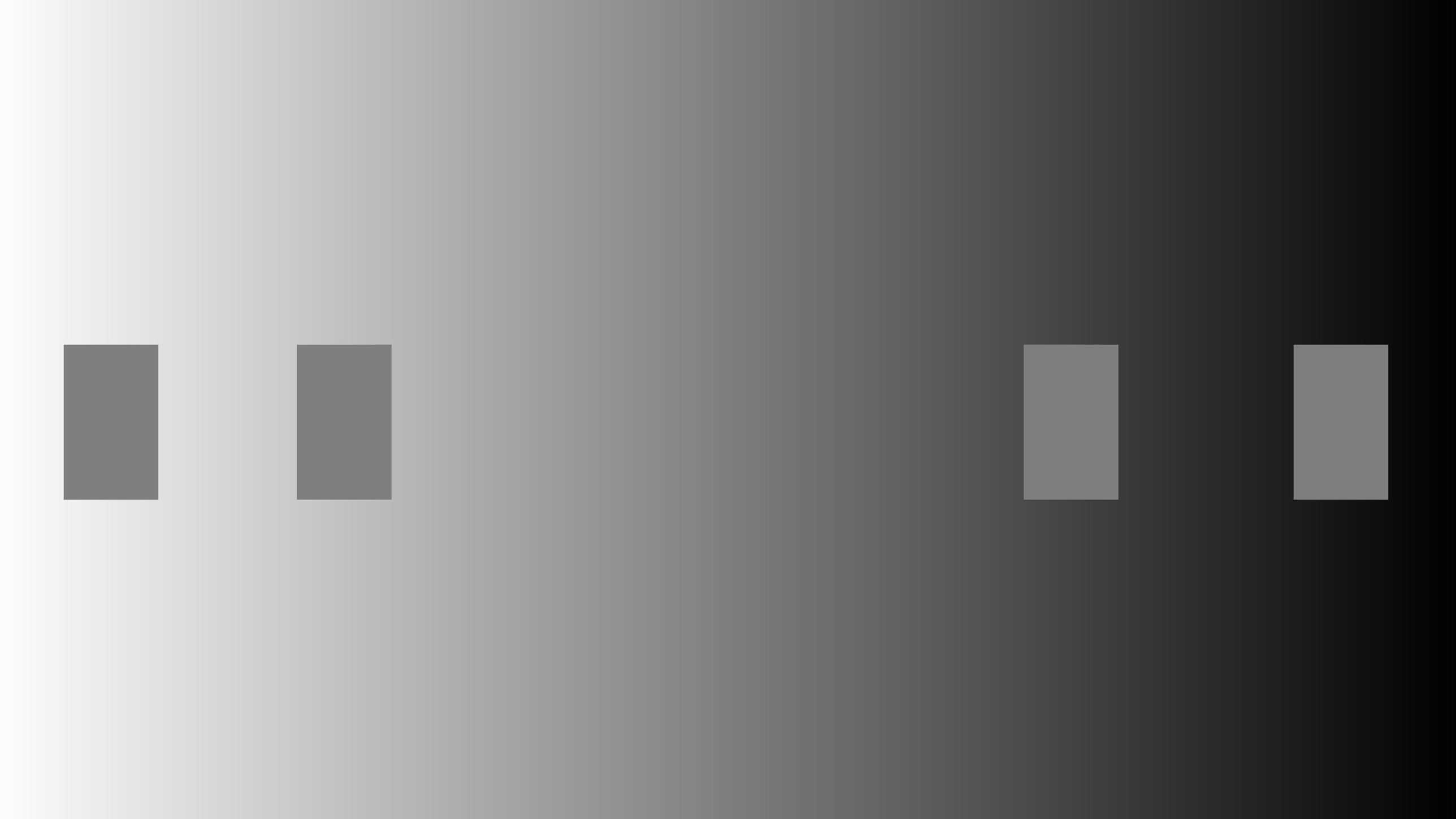


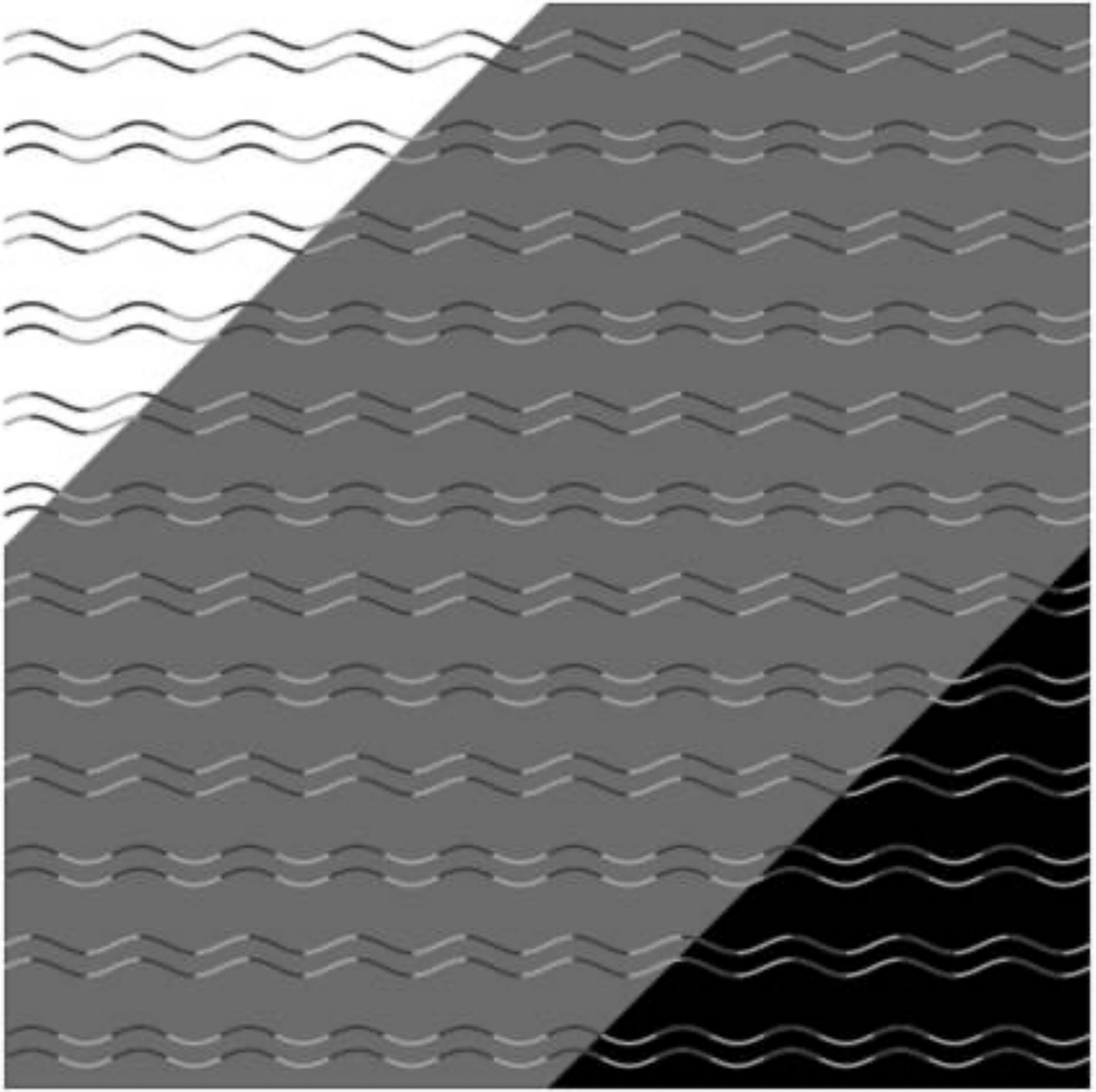


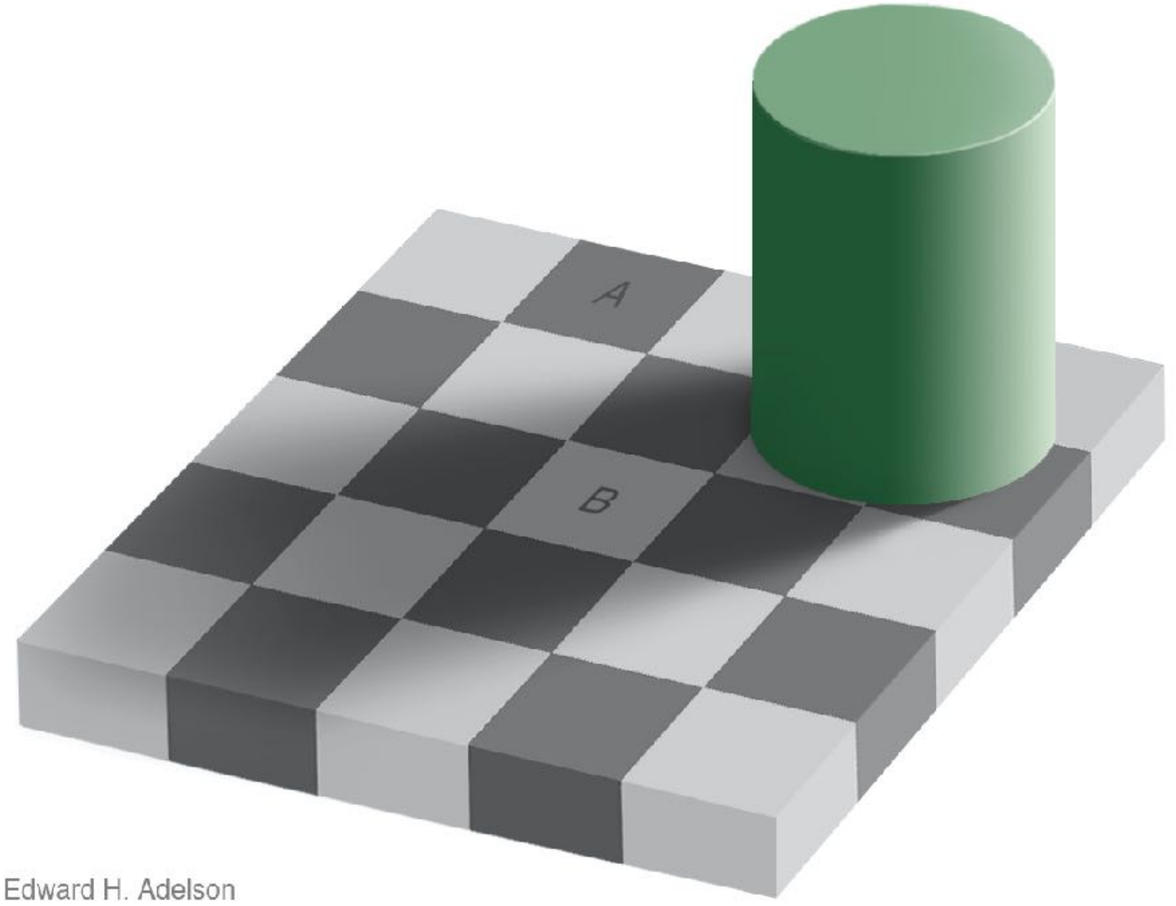




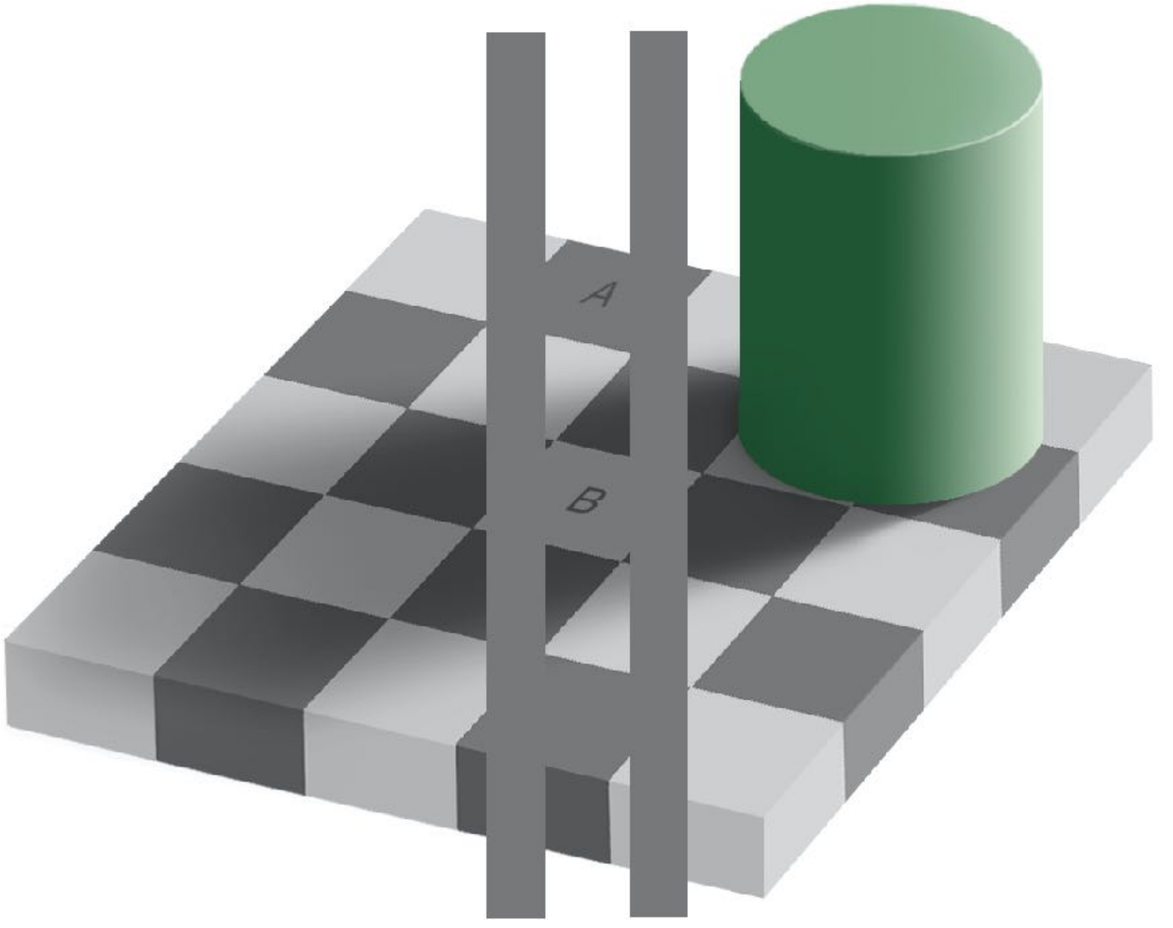


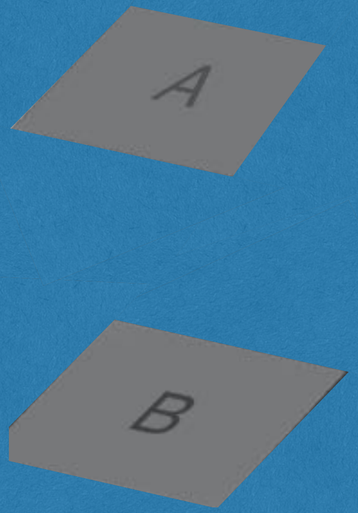


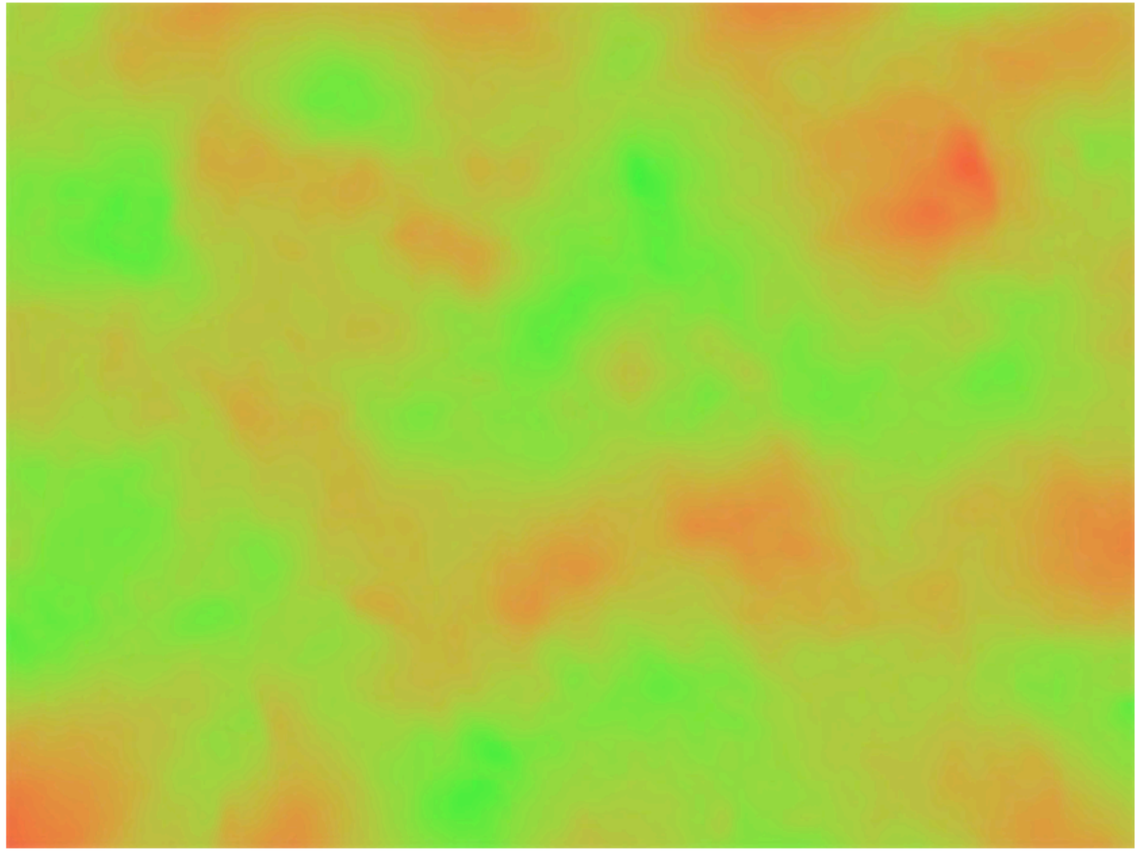


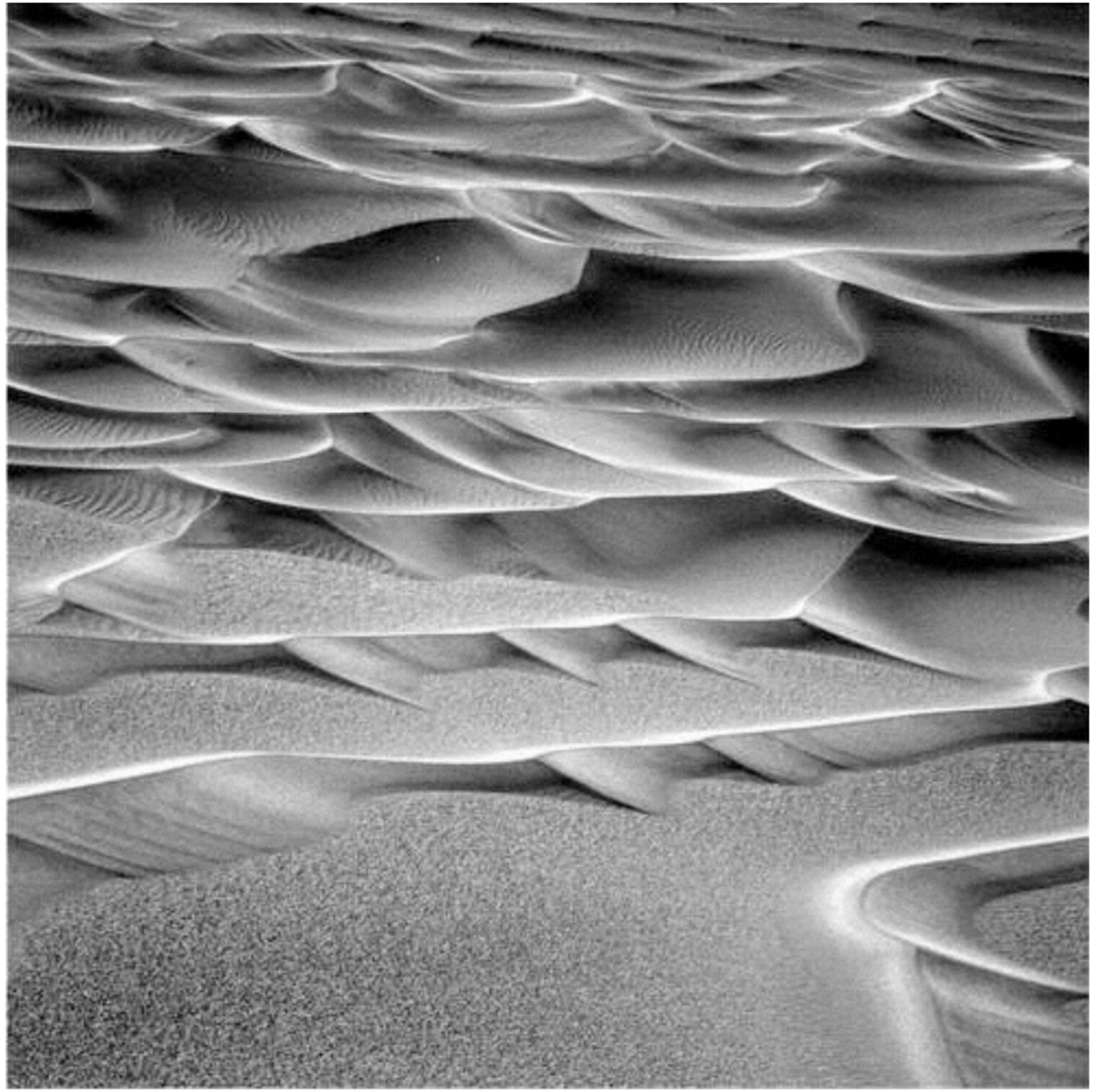


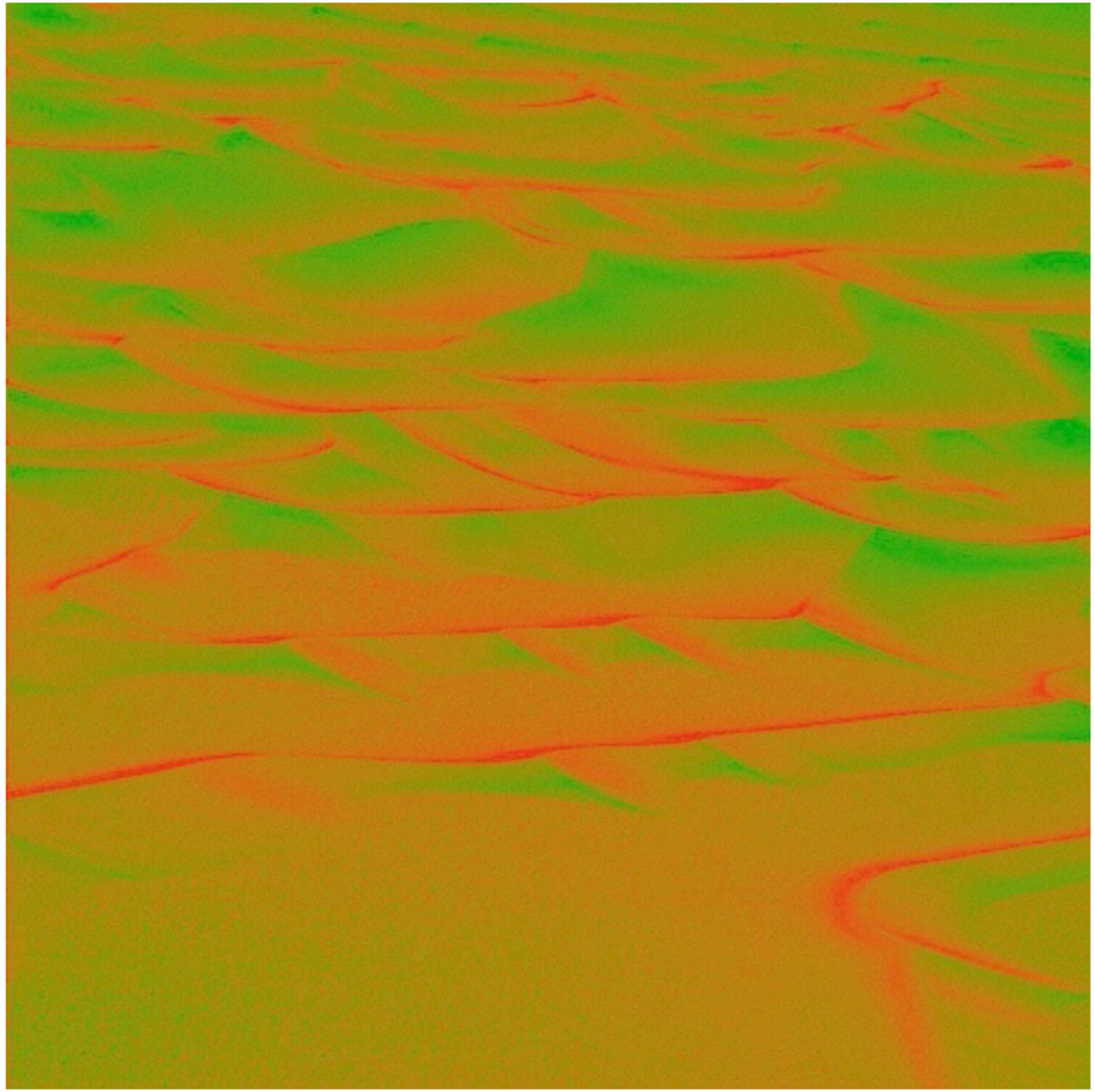
Edward H. Adelson

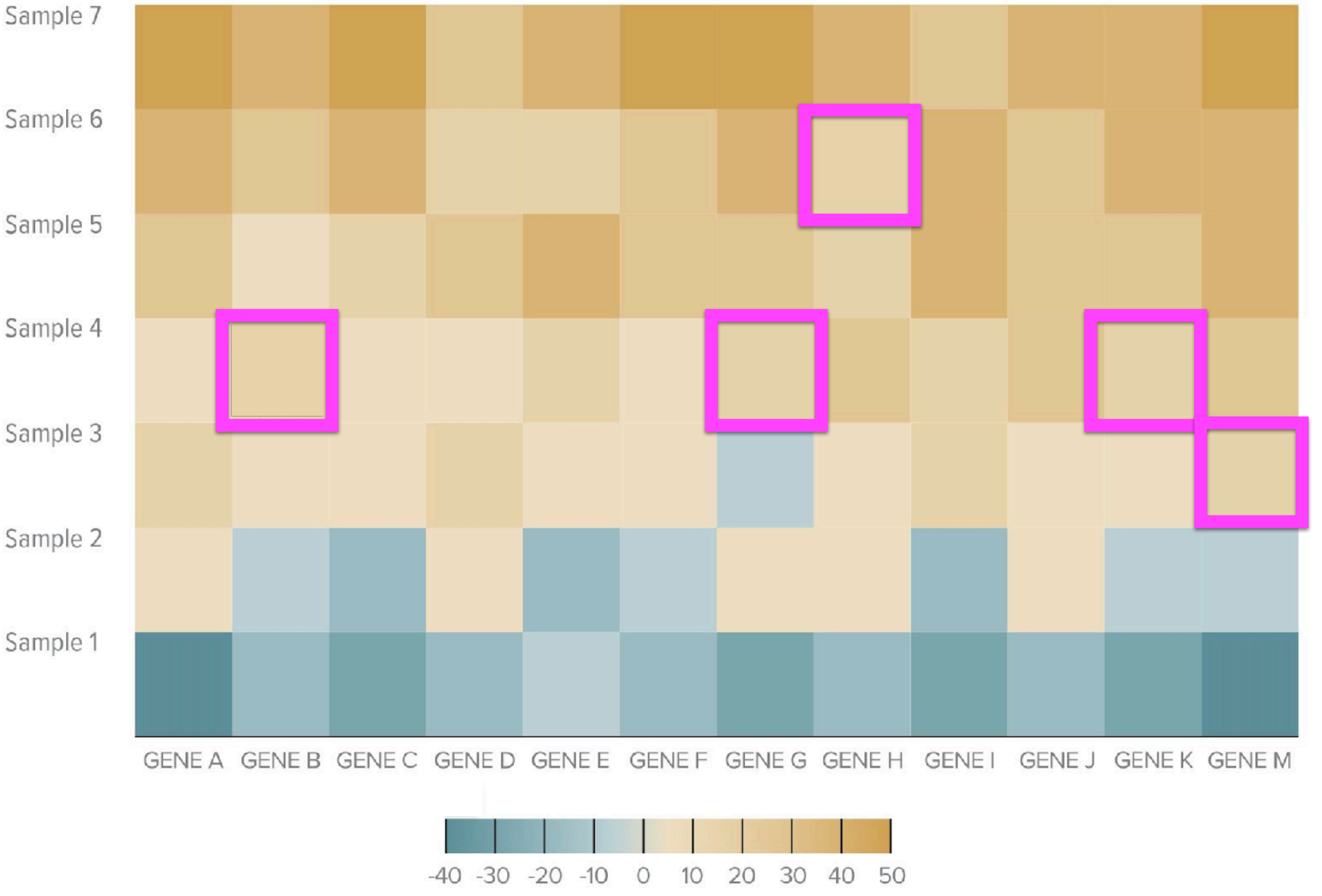



A
B



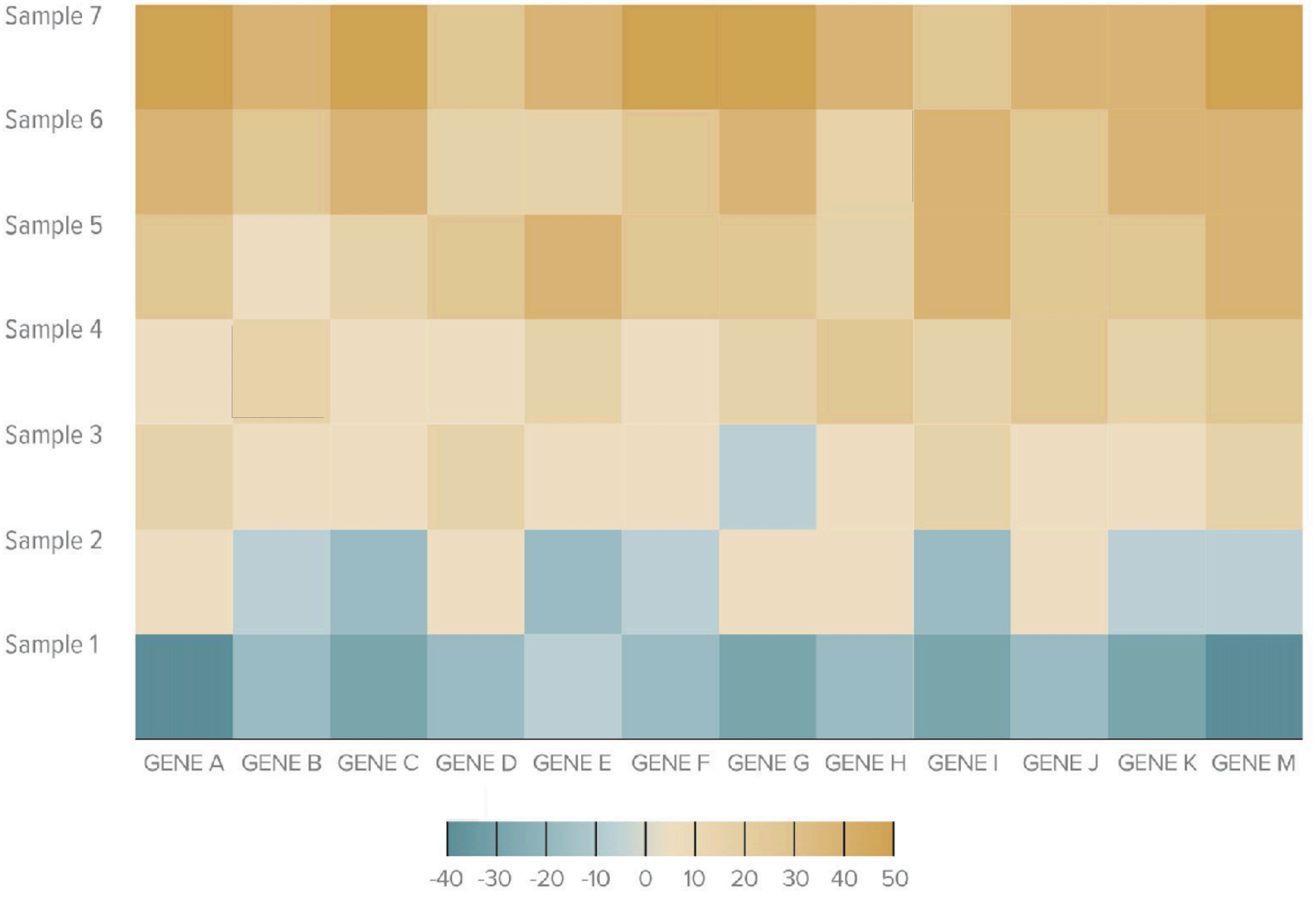






SOURCE: H.E. GRECCO *ET AL* / *NATURE METHODS* 2010

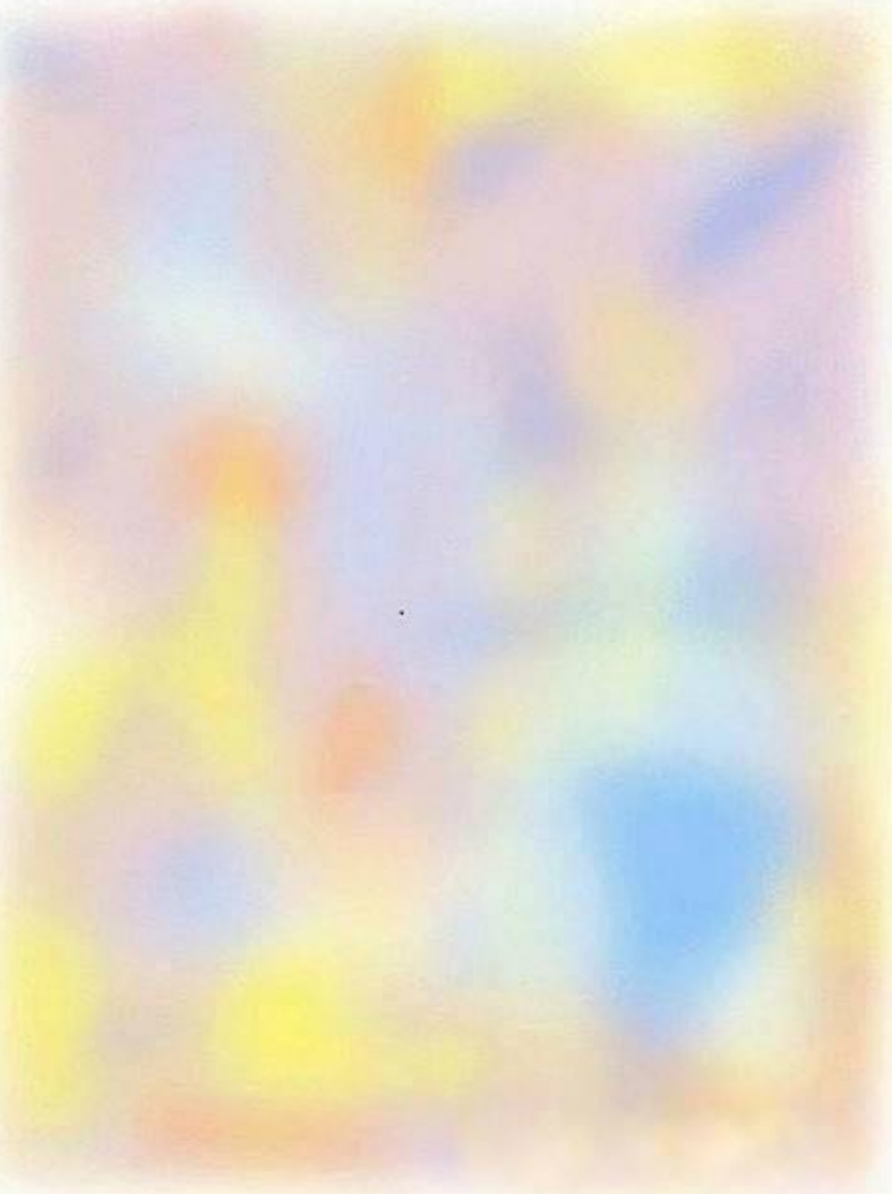
5W INFOGRAPHIC / KNOWABLE

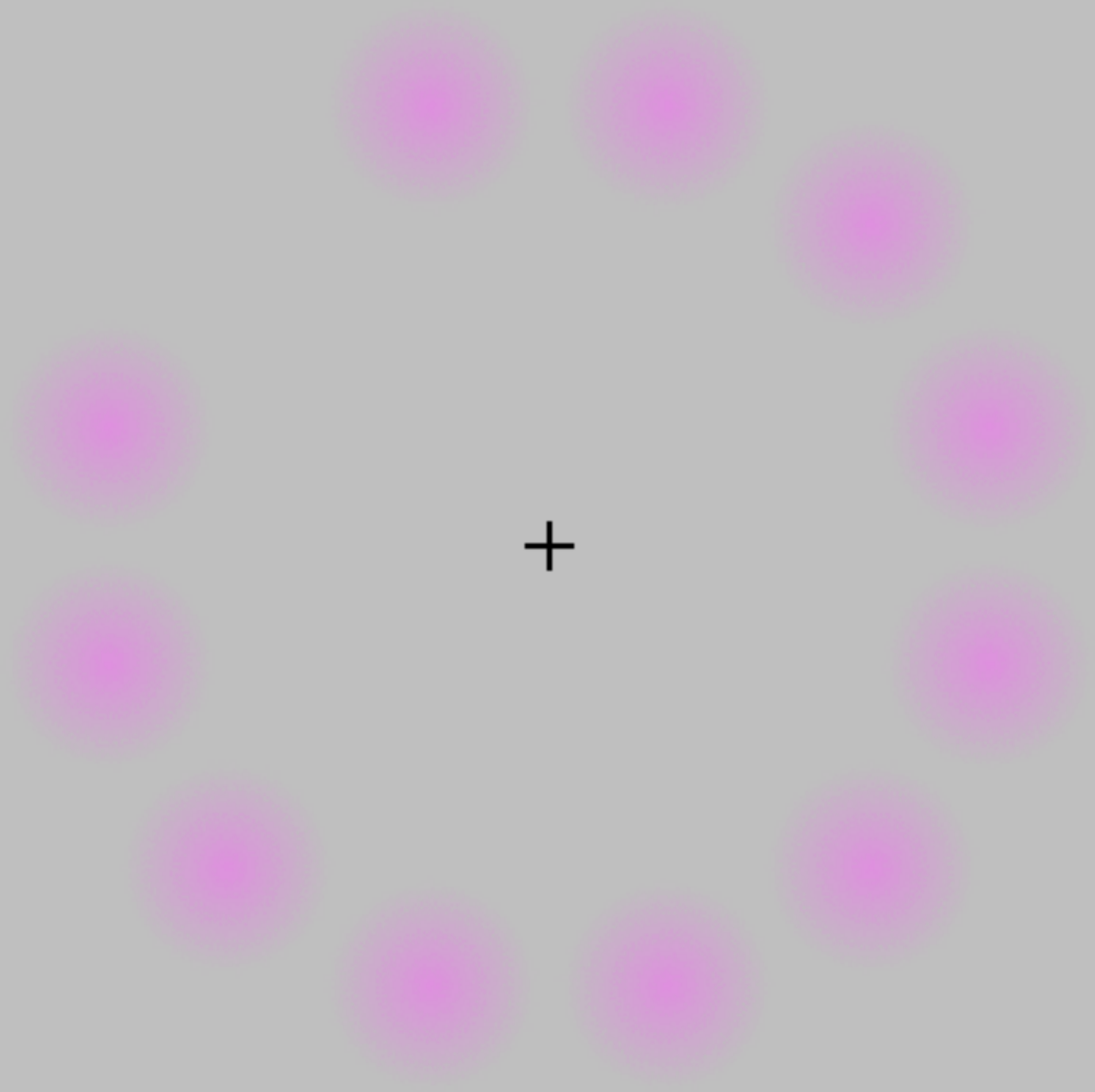


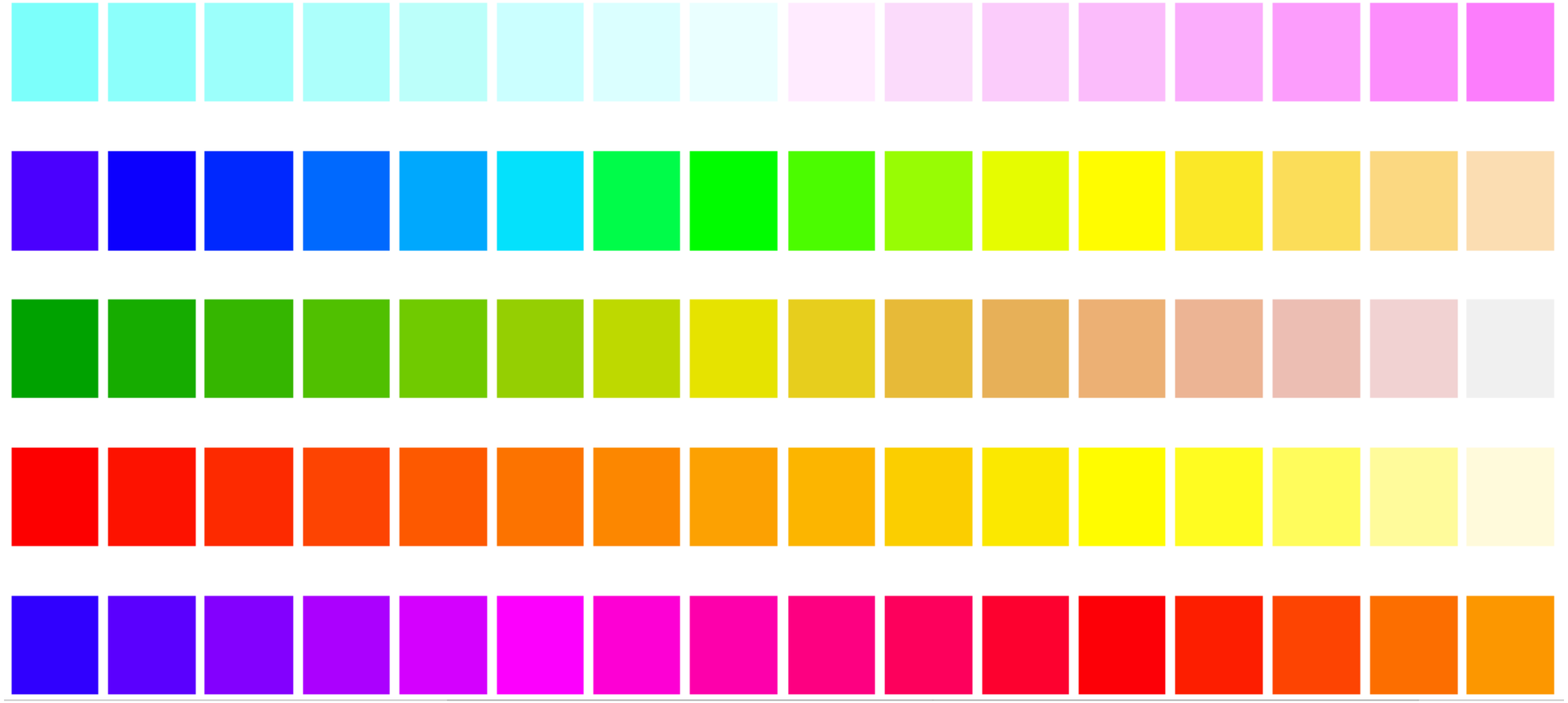
SOURCE: H.E. GRECCO *ET AL* / *NATURE METHODS* 2010

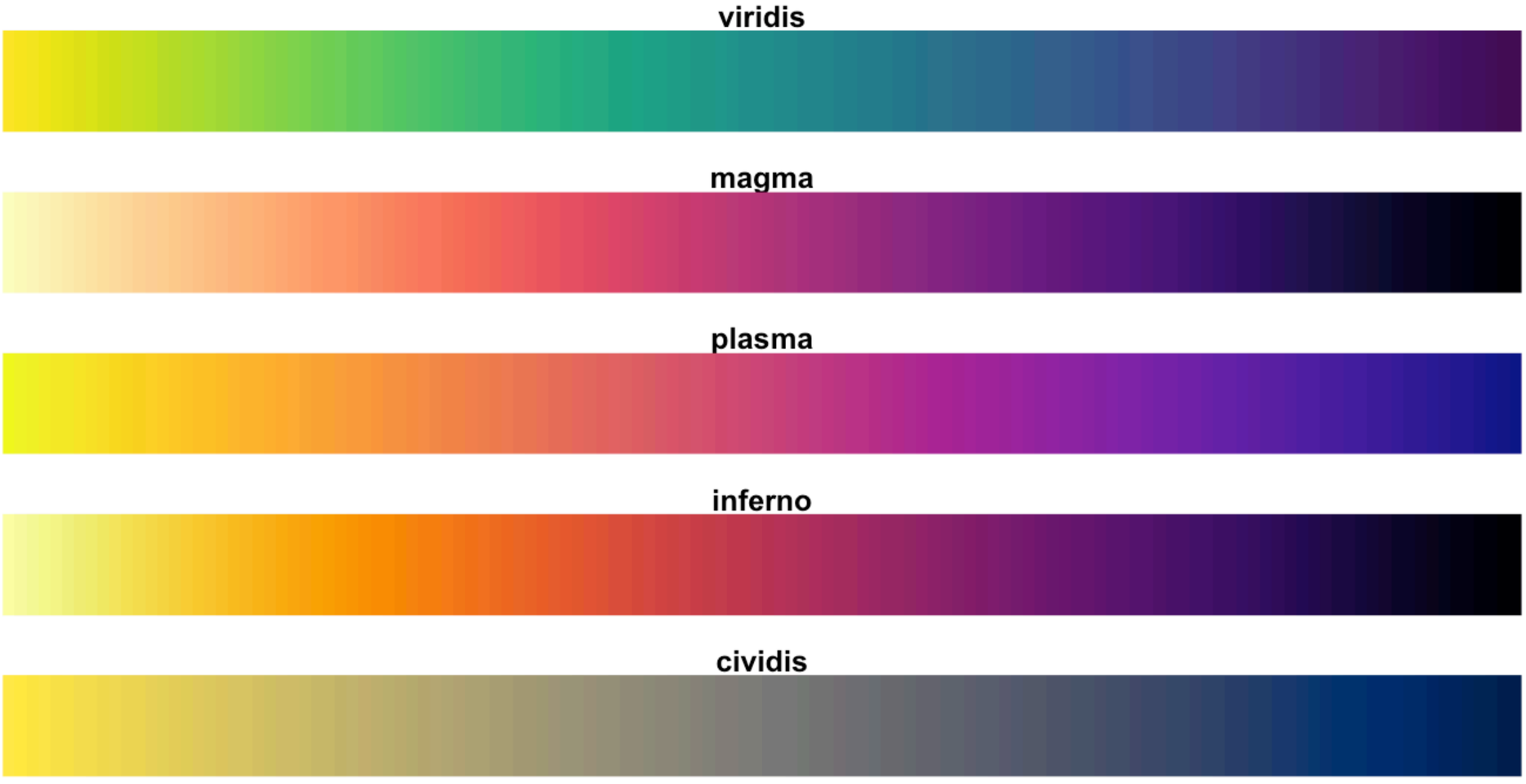
5W INFOGRAPHIC / KNOWABLE

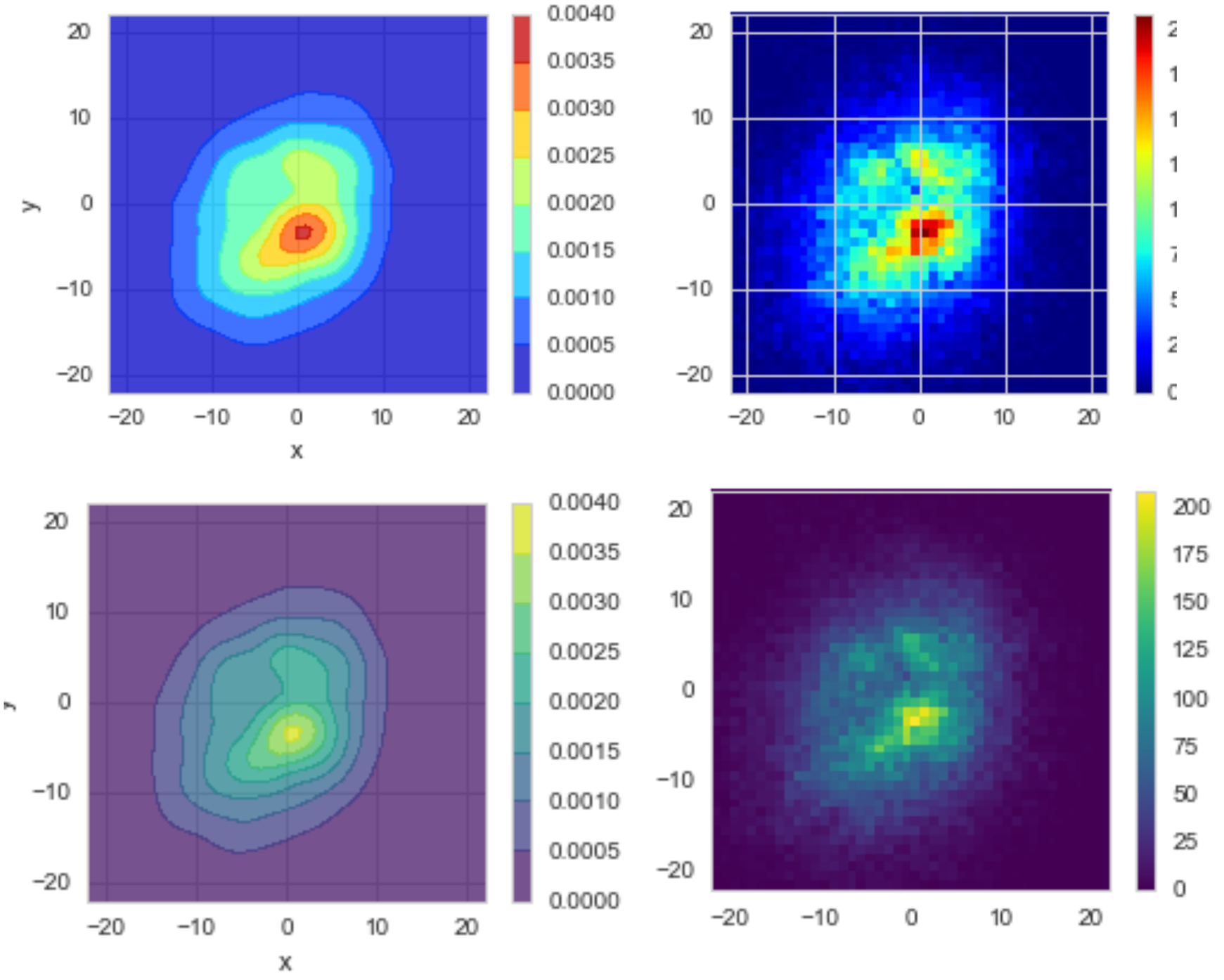
Color

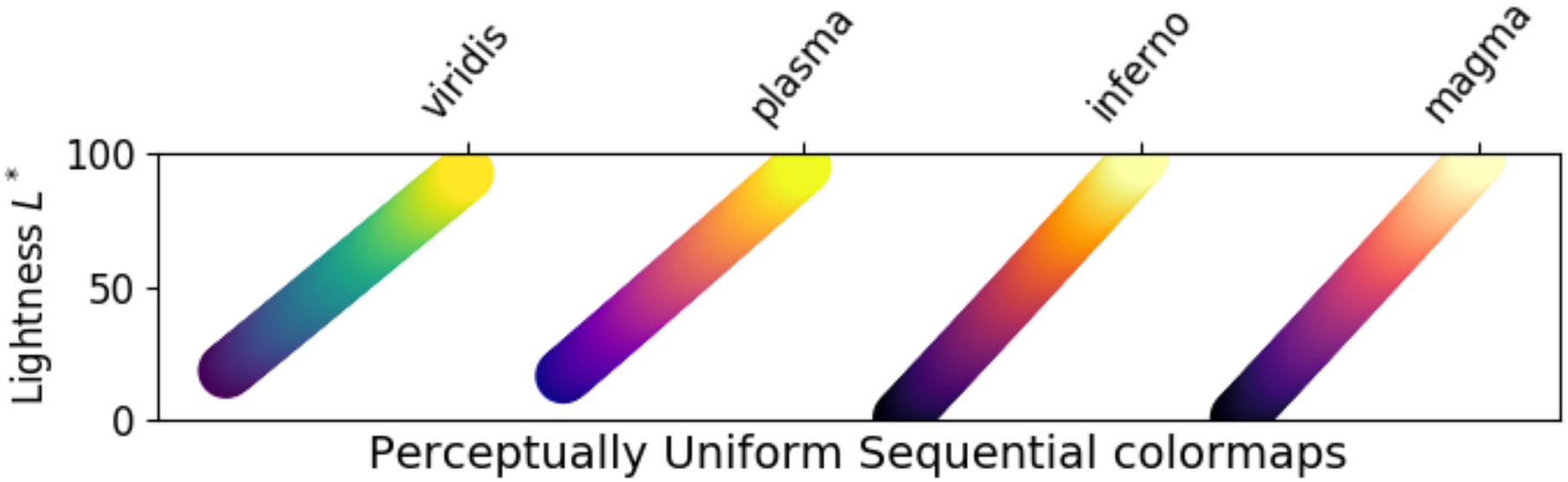


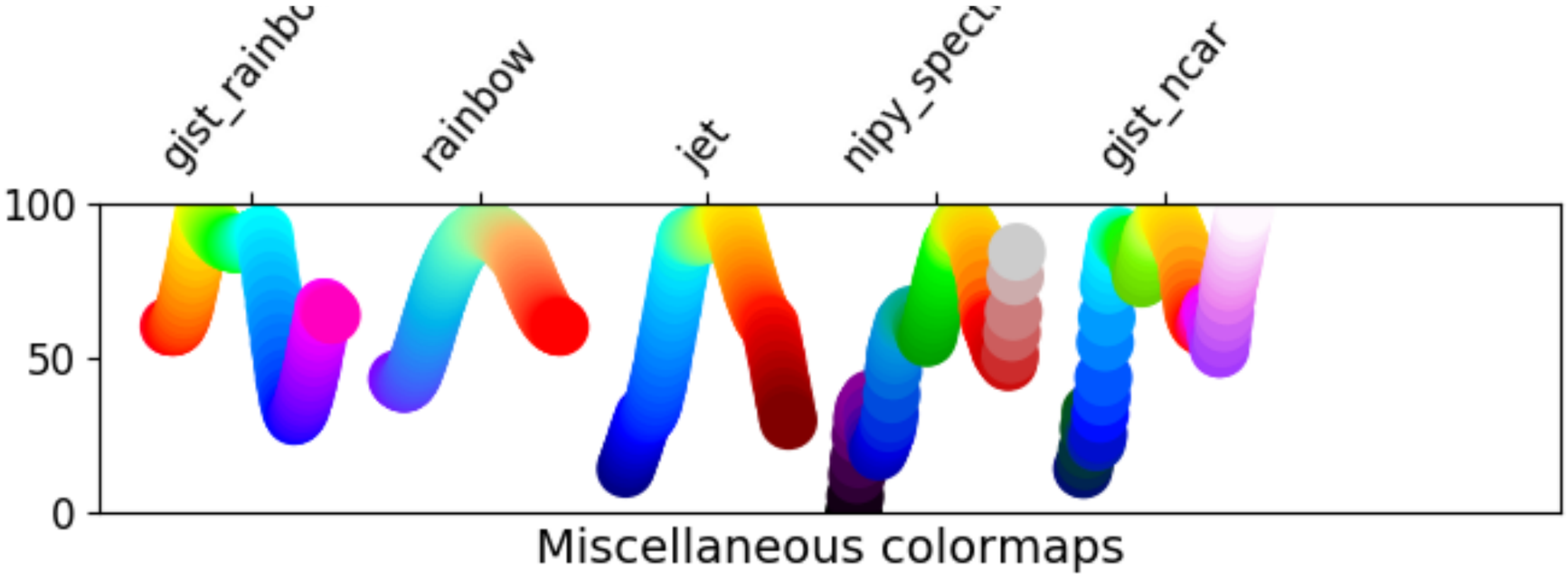


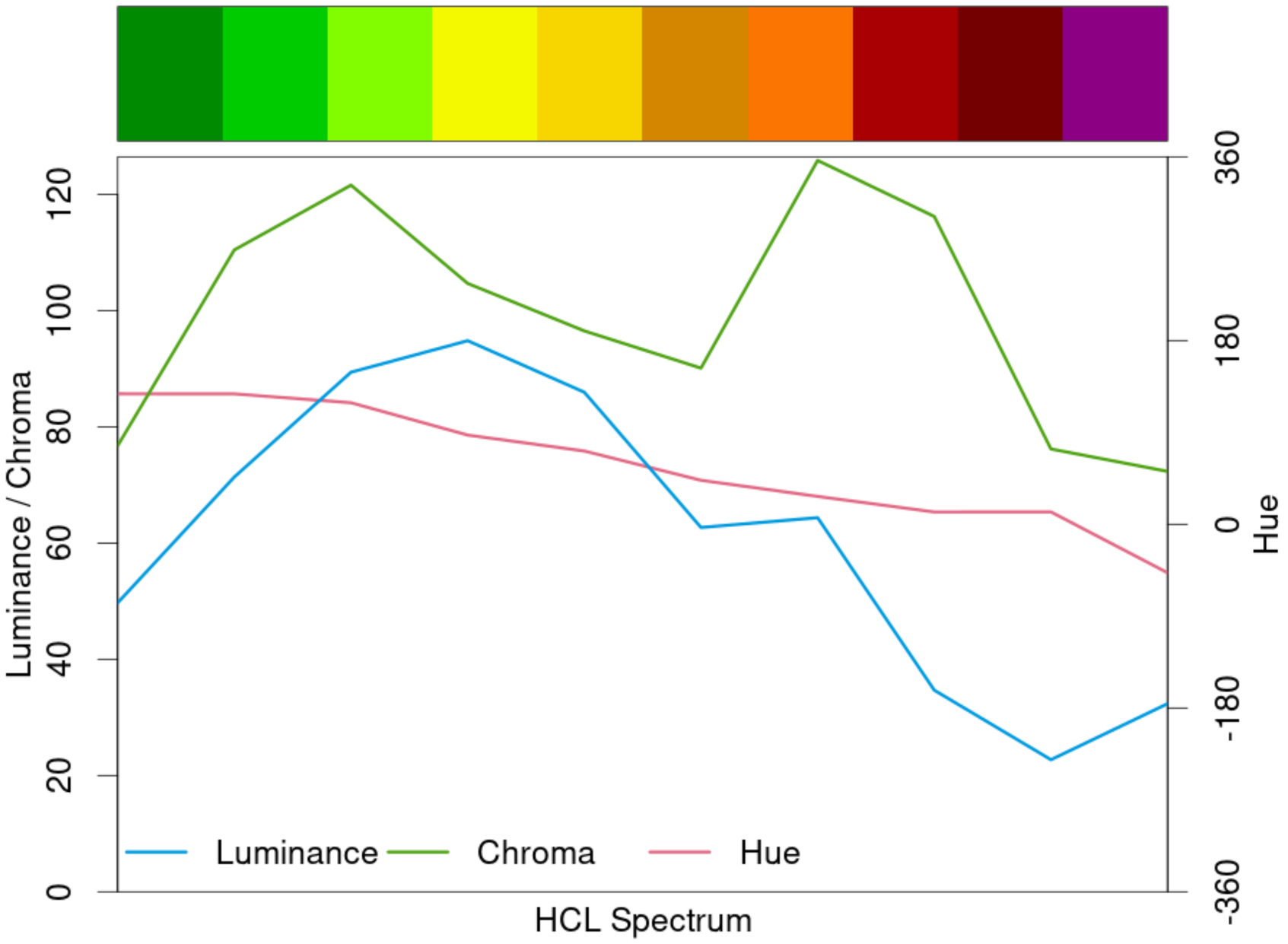


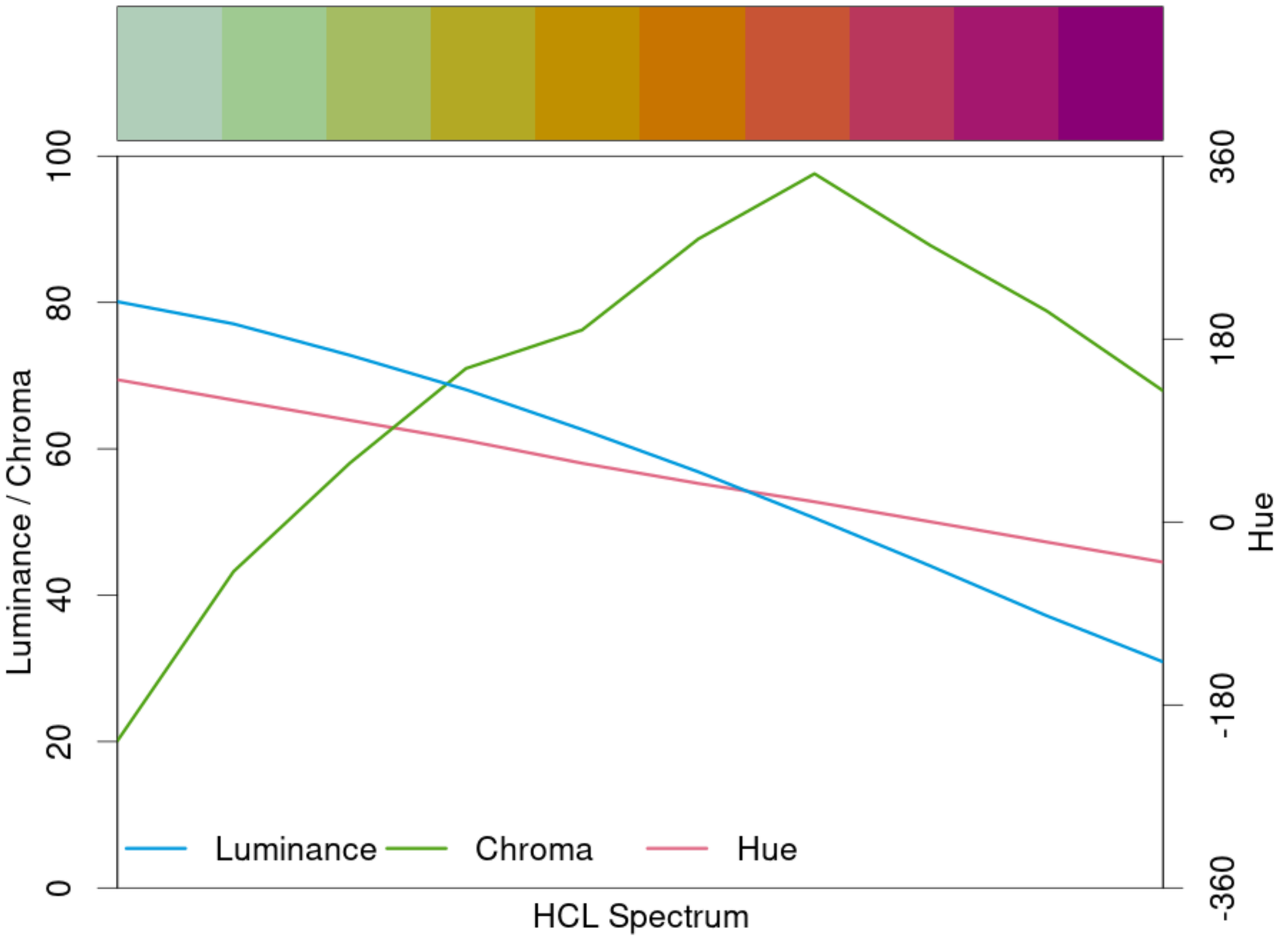

viridis
magma
plasma
inferno
cividis

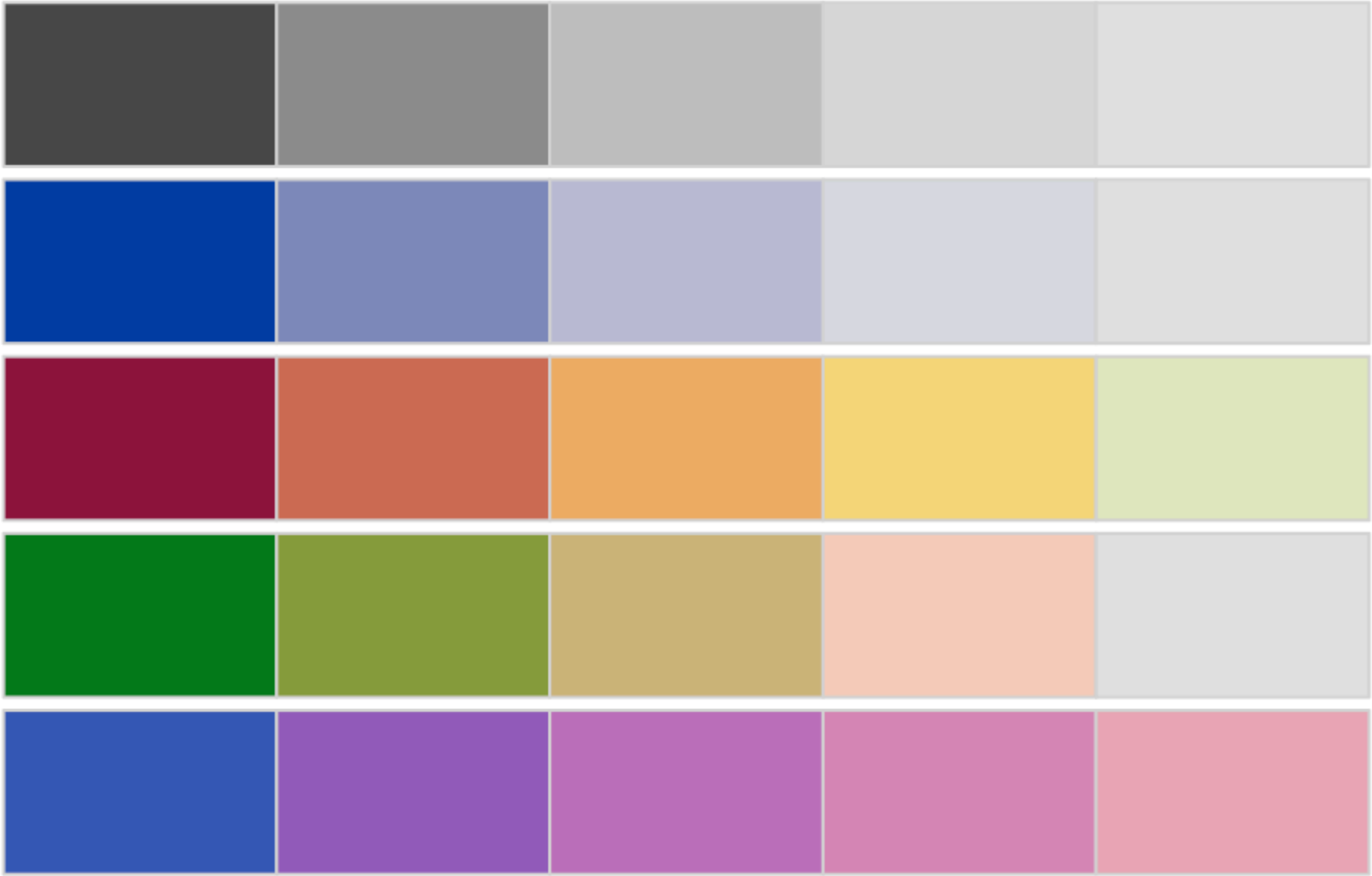

y
x
20
10
0
−10
−20
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
200
175
150
125
100
75
50
25

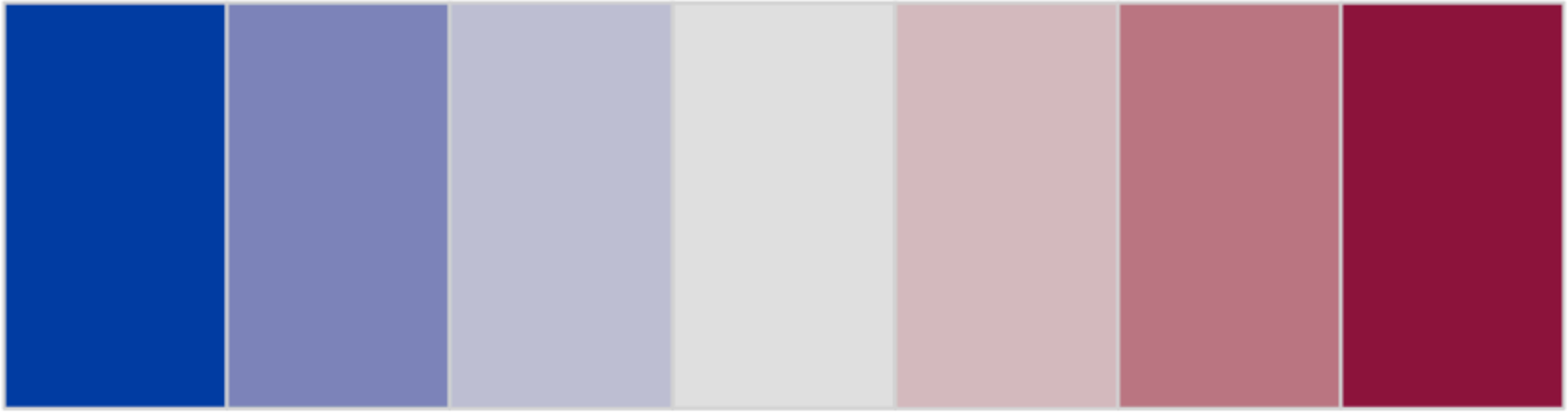

viridis
plasma
inferno
magma
100
50
0
Lightness L^*
Perceptually Uniform Sequential colormaps


gist_rainbo
rainbow
jet
nipy_spect
gist_ncar
100
50
0
Miscellaneous colormaps


Luminance / Chroma
0
20
40
60
80
100
120
Hue
-360
-180
0
180
360
HCL Spectrum
Luminance
Chroma
Hue

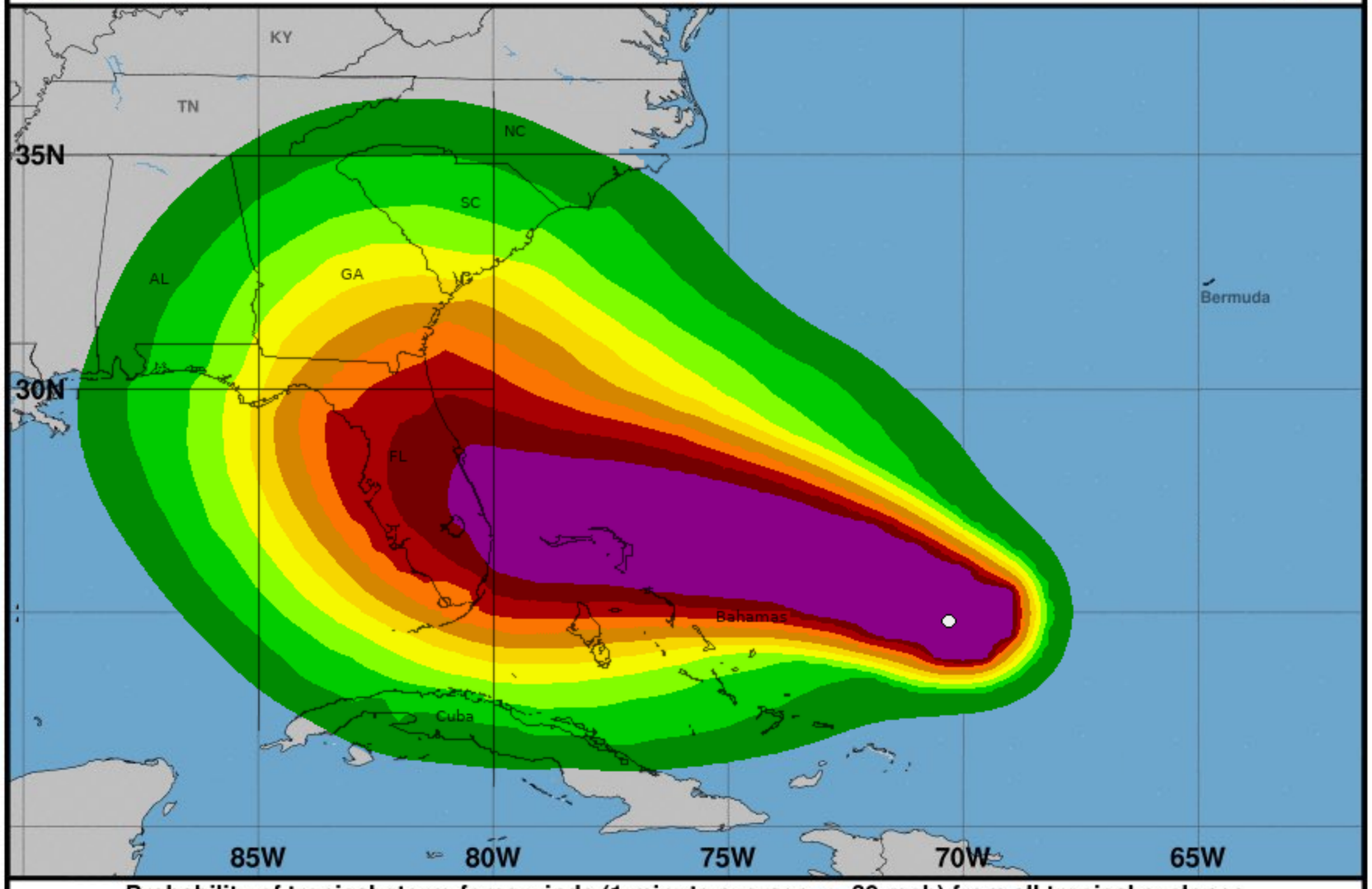

Luminance / Chroma
0
20
40
60
80
100
Hue
-360
-180
0
180
360
Luminance
Chroma
Hue
HCL Spectrum





Tropical-Storm-Force Wind Speed Probabilities

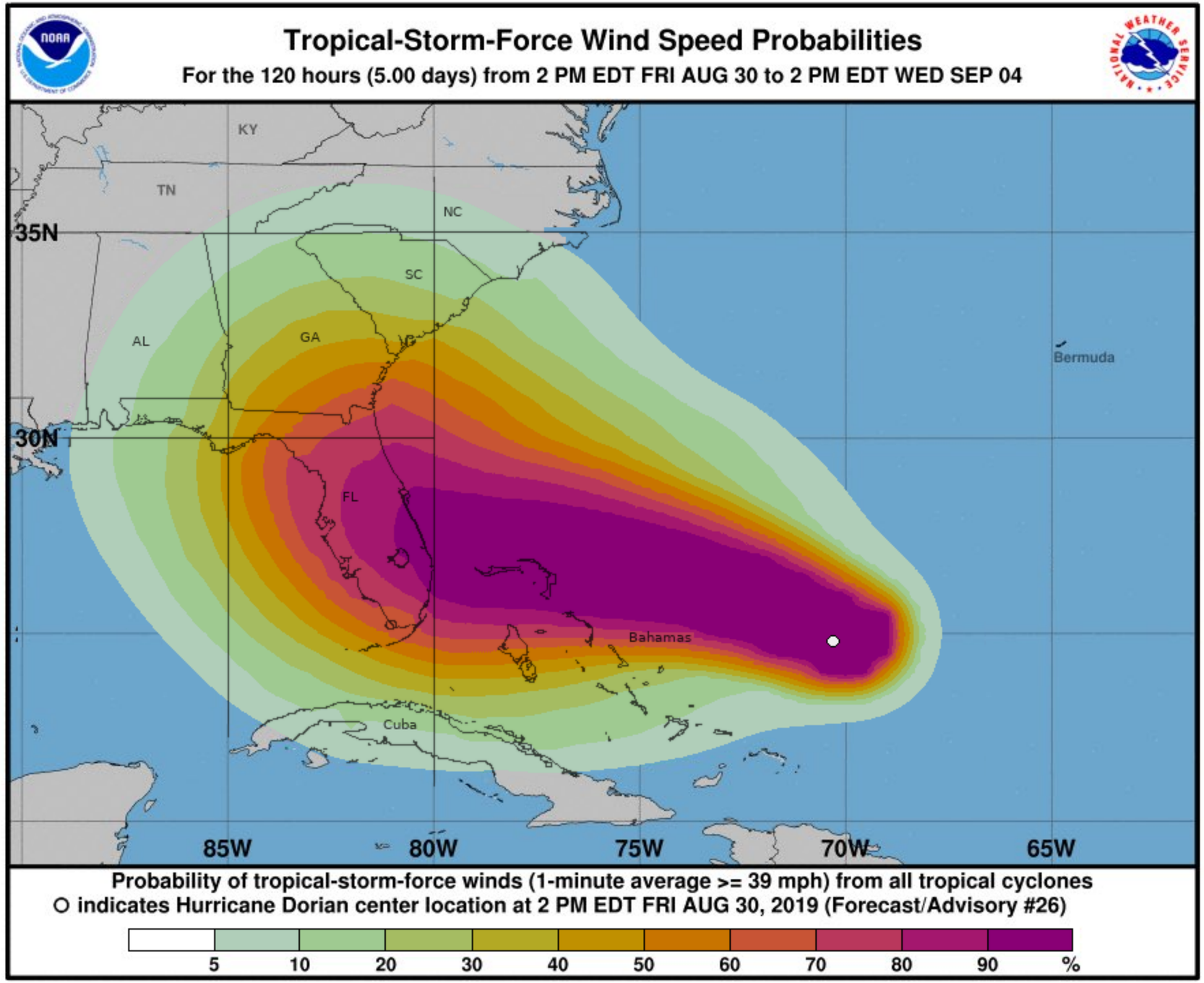
For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04



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)

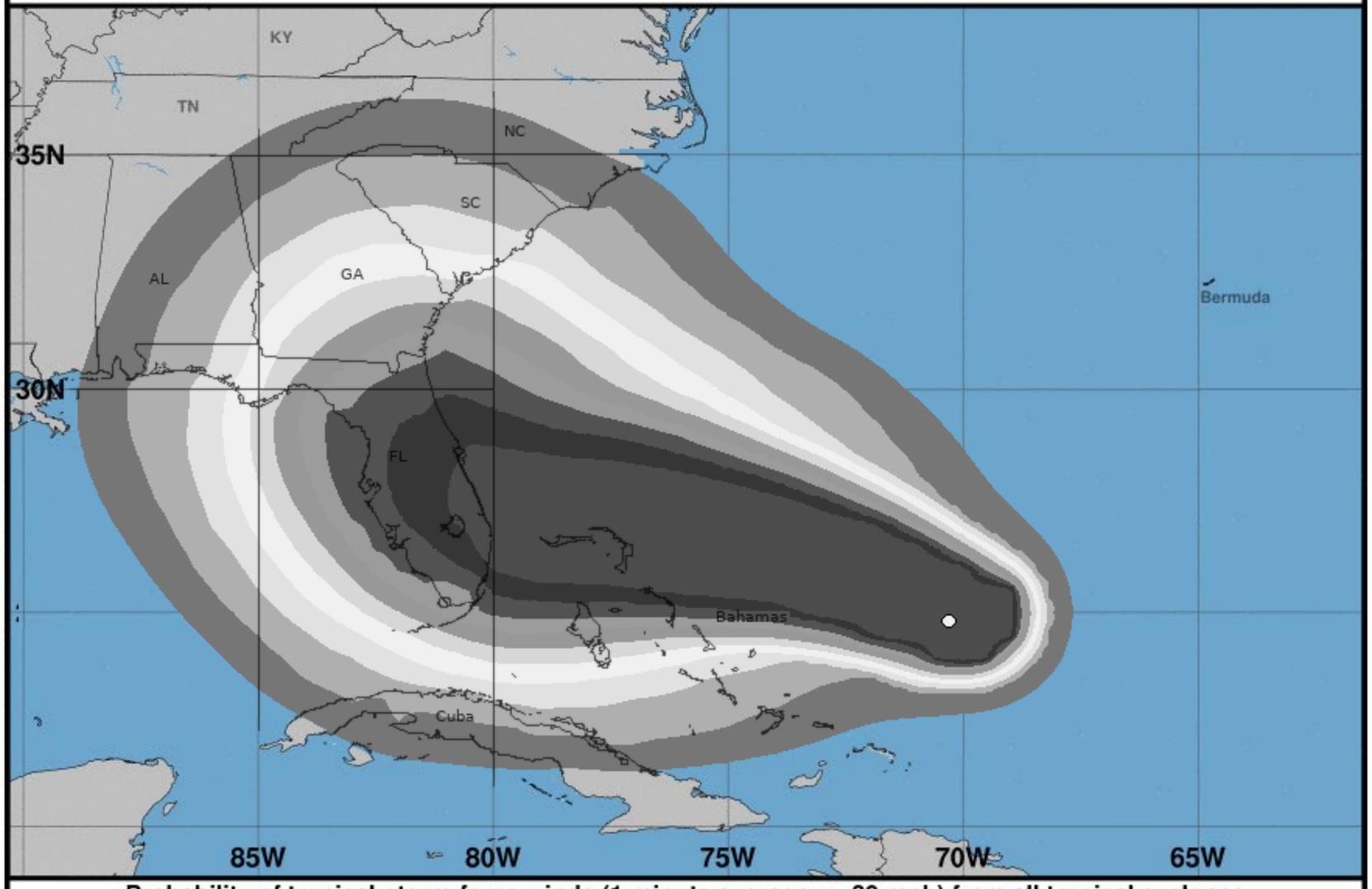
5 10 20 30 40 50 60 70 80 90 %



Achim Zeileis
https://bit.ly/35zBlYL

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



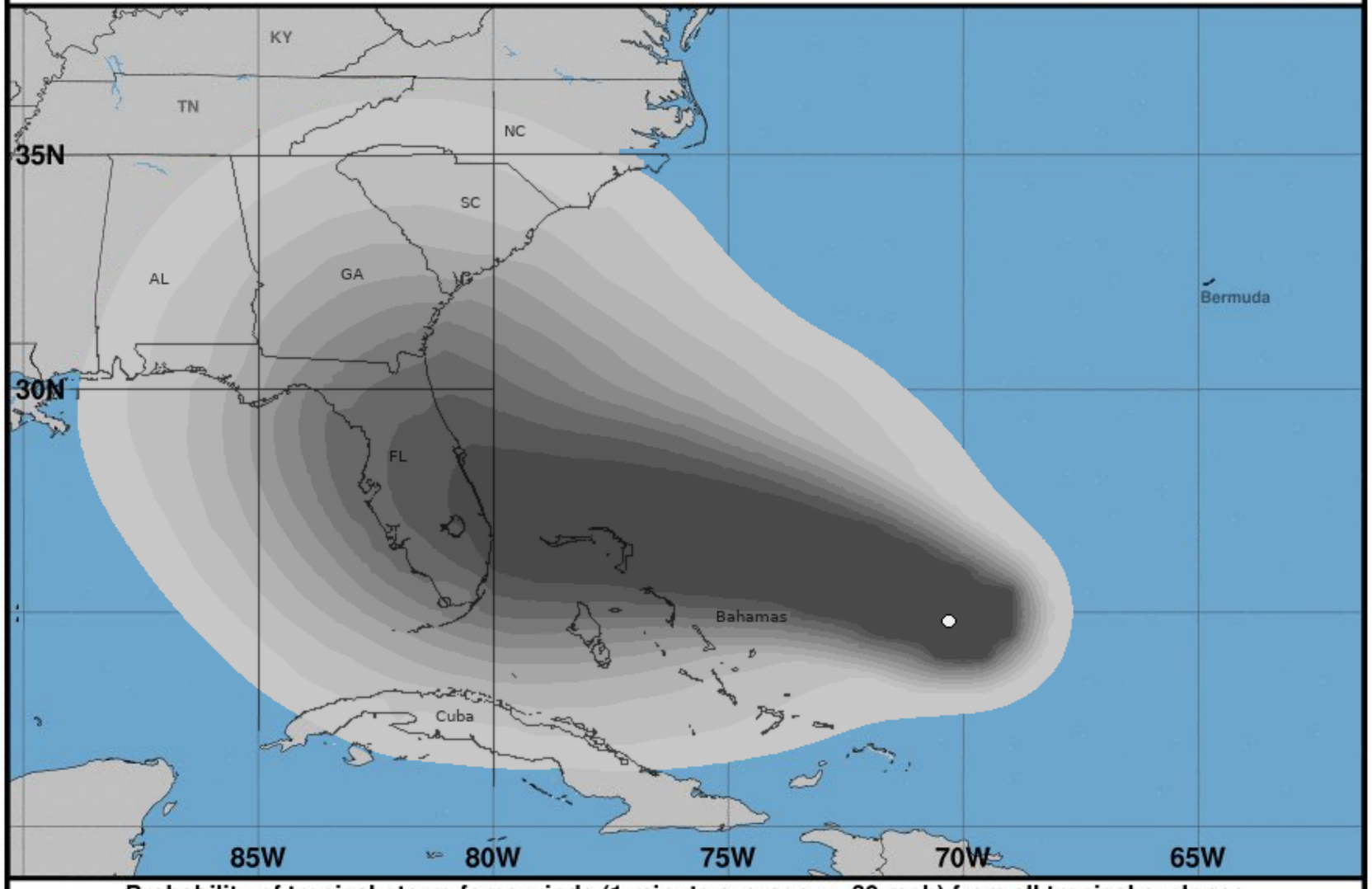
Probability of tropical-storm-force winds (1-minute average >= 39 mph) from all tropical cyclones

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

5 10 20 30 40 50 60 70 80 90 %

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



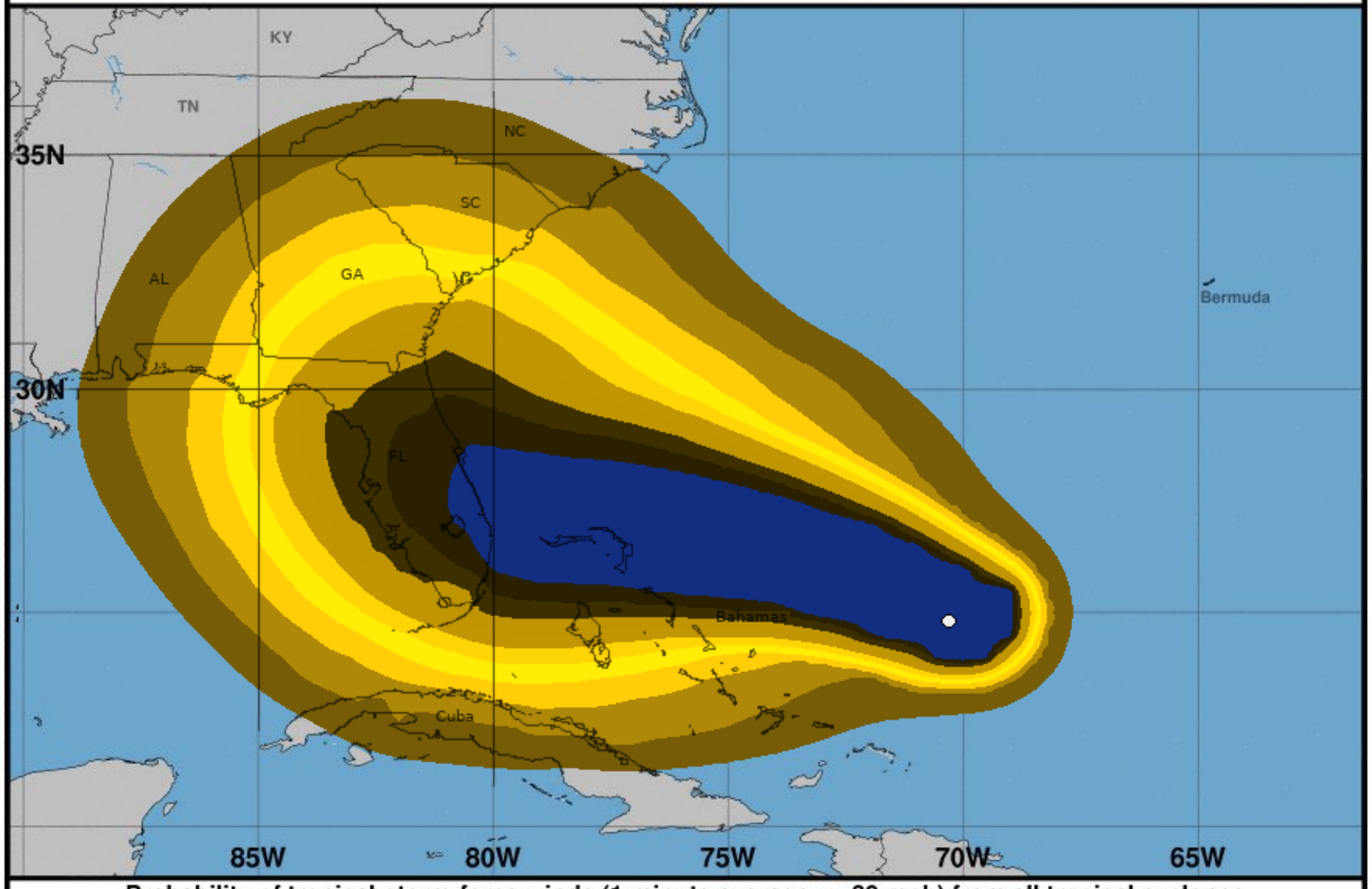
Probability of tropical-storm-force winds (1-minute average >= 39 mph) from all tropical cyclones

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

5 10 20 30 40 50 60 70 80 90 %

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



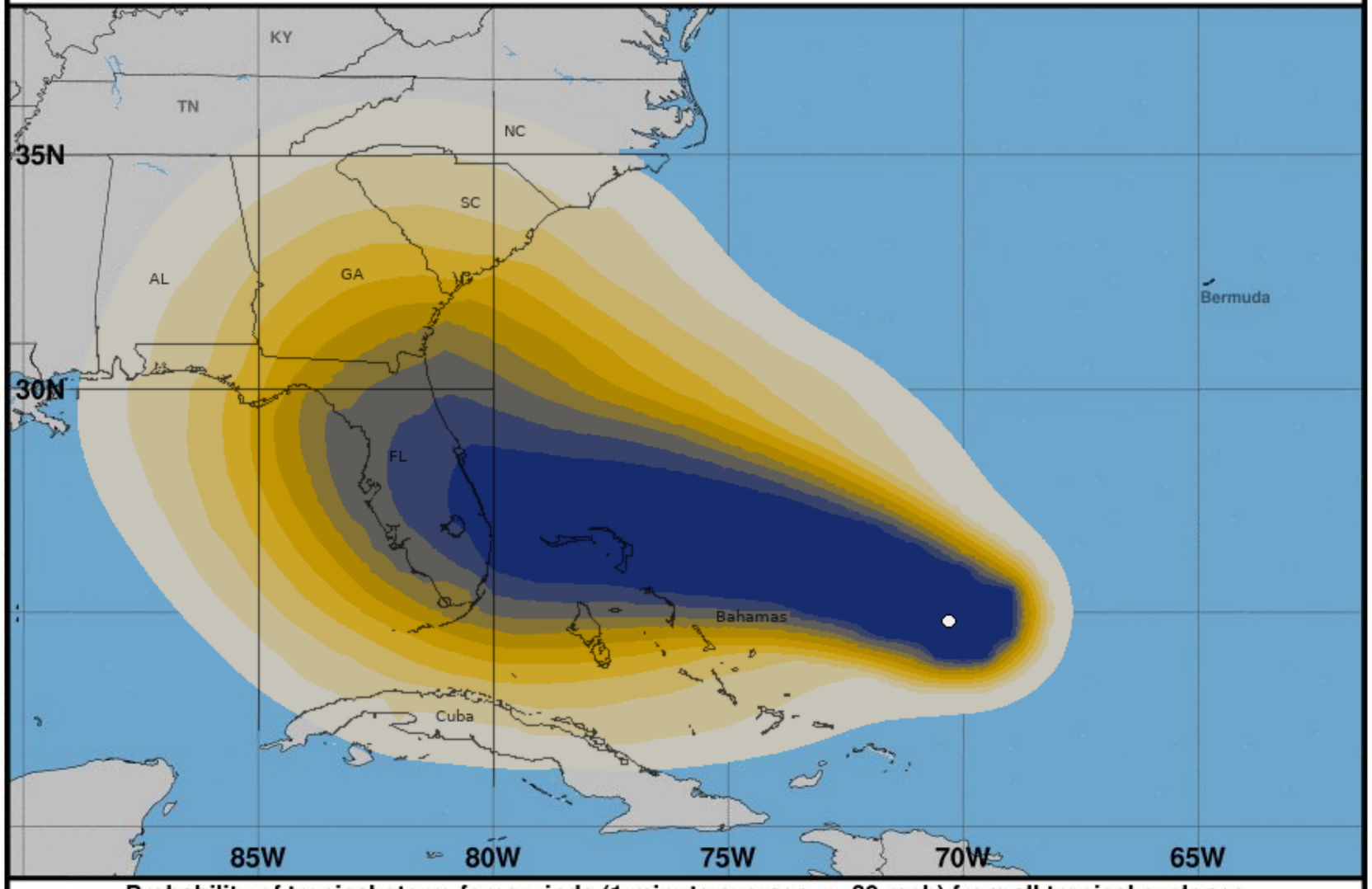
Probability of tropical-storm-force winds (1-minute average >= 39 mph) from all tropical cyclones

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

5 10 20 30 40 50 60 70 80 90 %

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



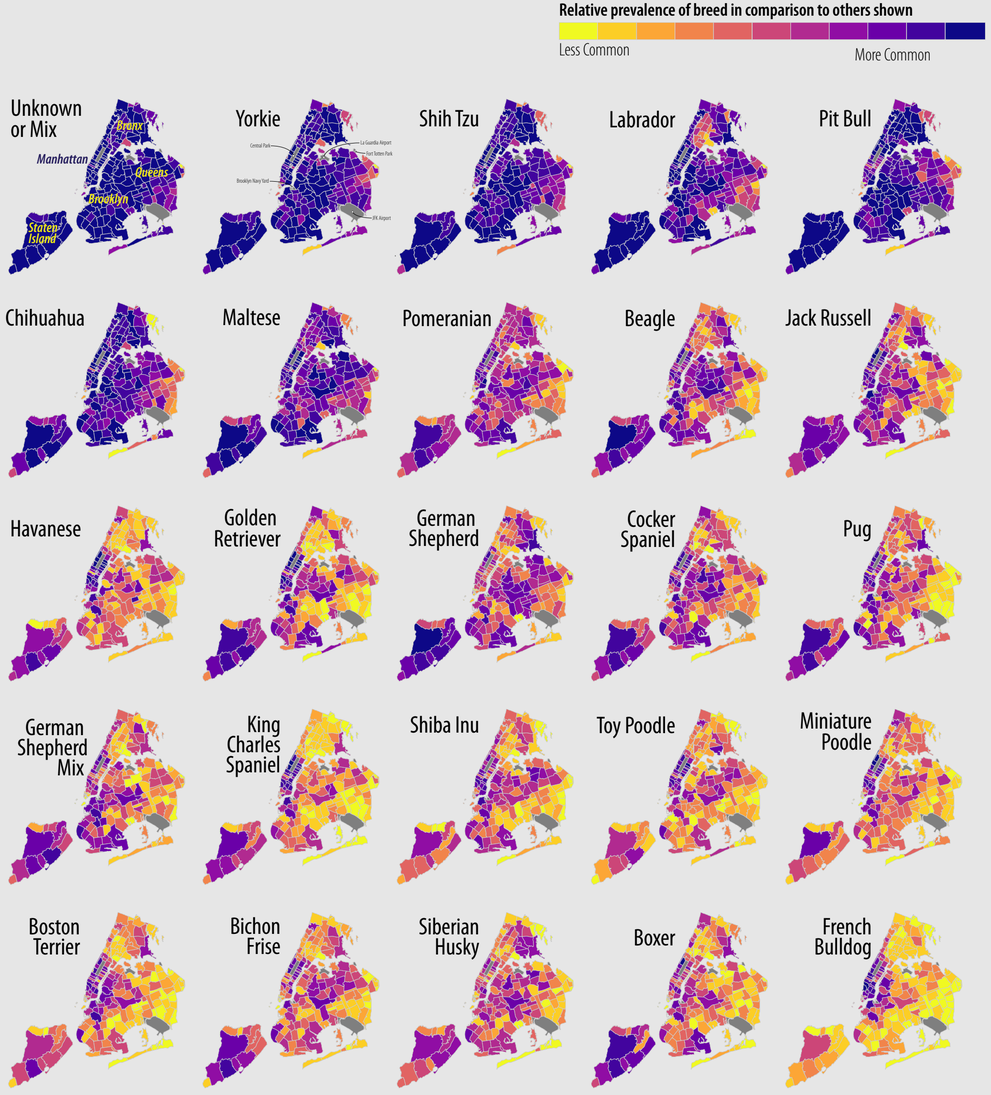
Probability of tropical-storm-force winds (1-minute average >= 39 mph) from all tropical cyclones

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

5 10 20 30 40 50 60 70 80 90 %

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.



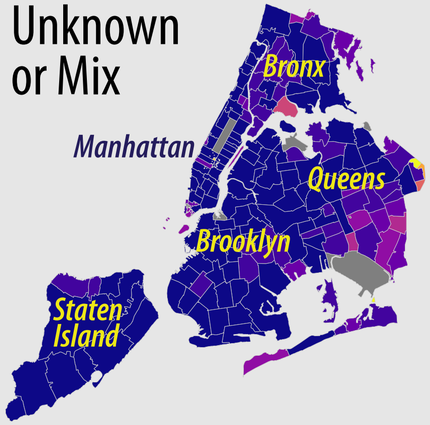
Kieran Healy / socviz.co / *Data Visualization: A Practical Introduction* is published by Princeton University Press

Dogs of New York

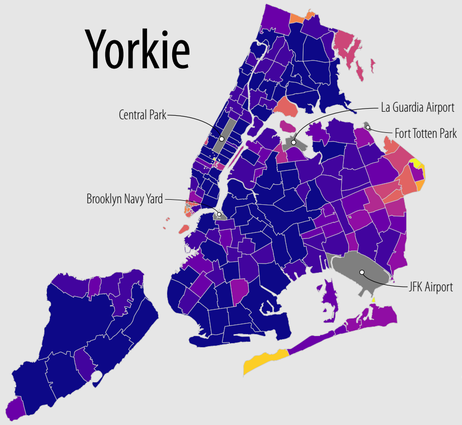
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.

Relative prevalence of breed in comparison to others shown

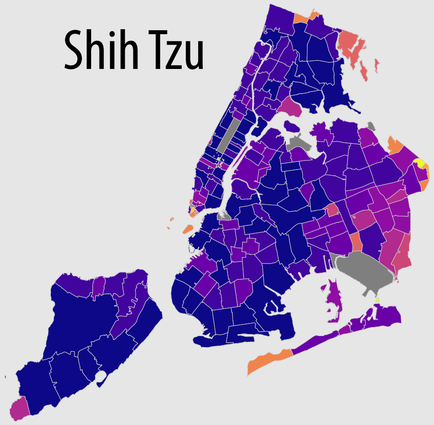
Unknown or Mix



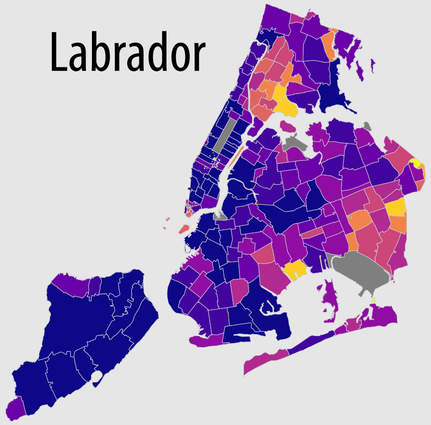
Yorkie



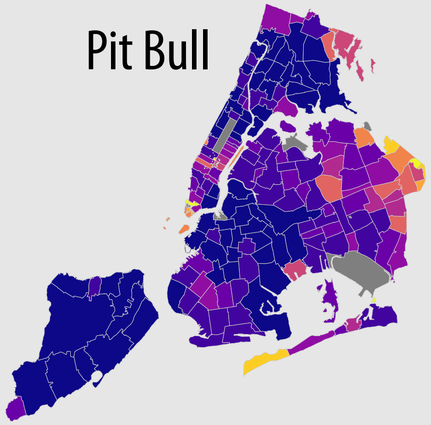
Shih Tzu



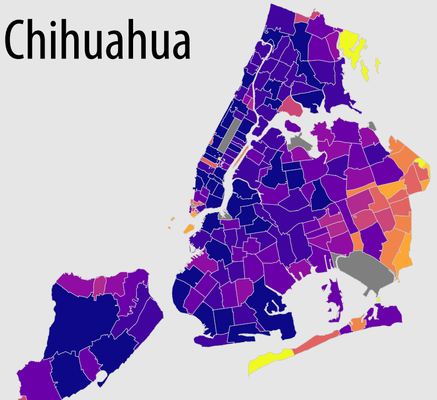
Labrador



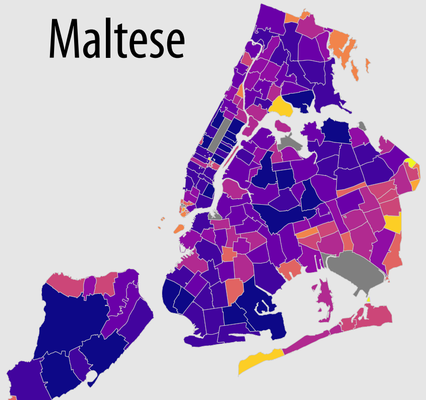
Pit Bull



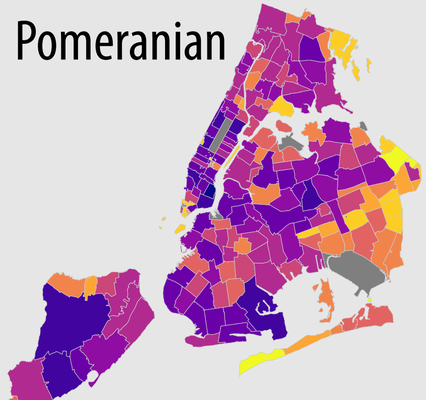
Chihuahua



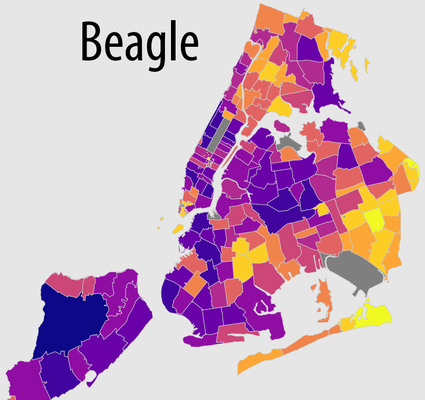
Maltese



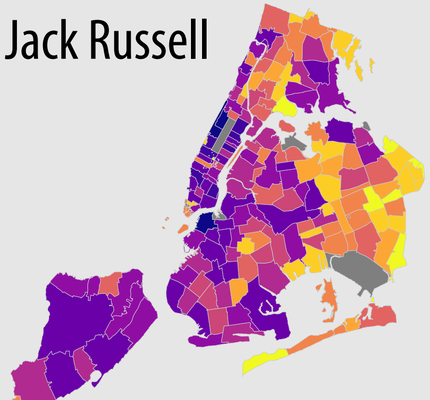
Pomeranian



Beagle

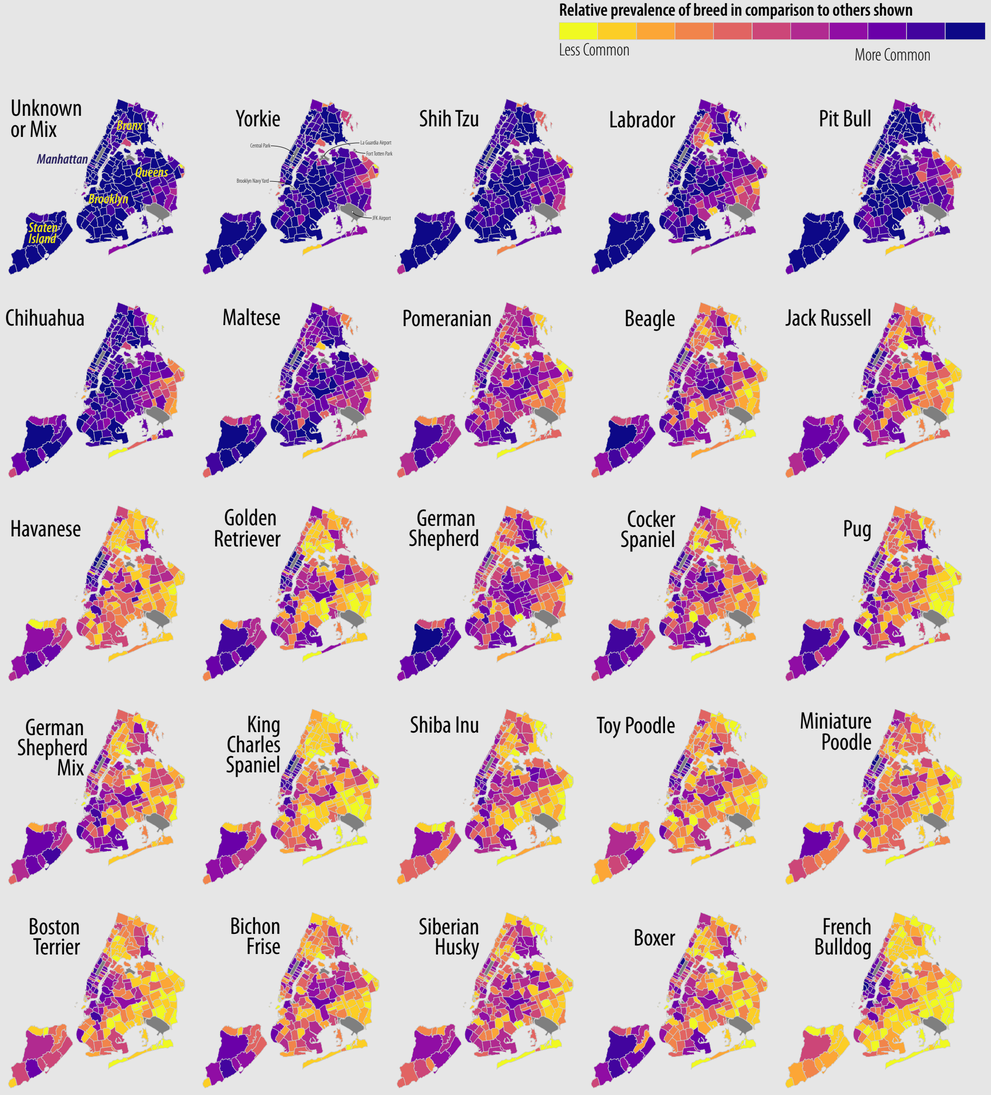


Jack Russell



Dogs of New York

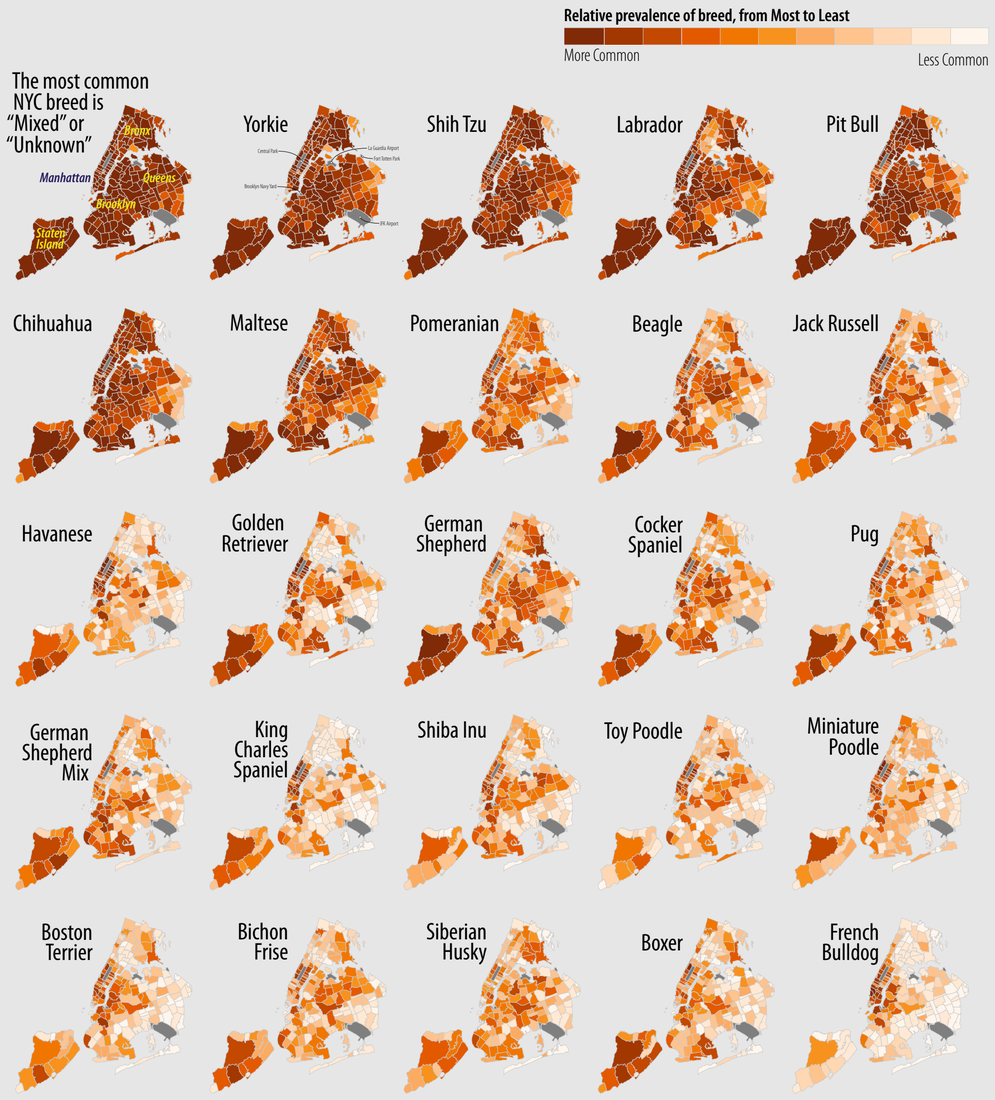
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.



Kieran Healy / socviz.co / *Data Visualization: A Practical Introduction* is published by Princeton University Press

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.

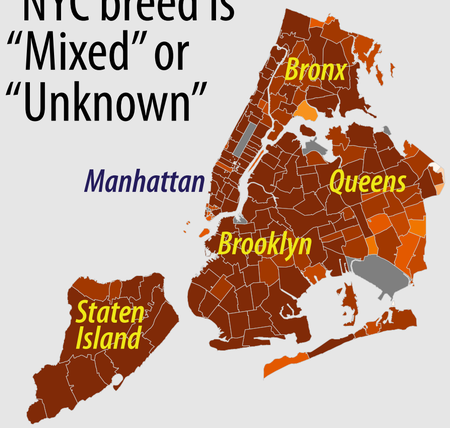


Kieran Healy / socviz.co / *Data Visualization: A Practical Introduction* is published by Princeton University Press

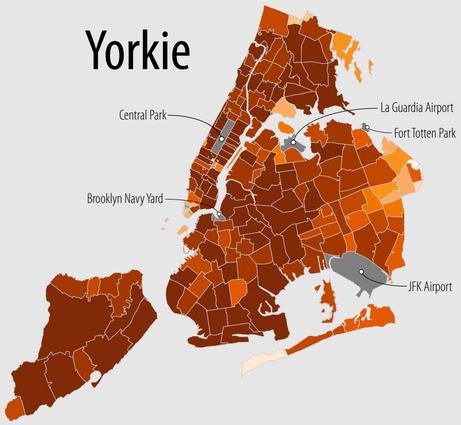
Relative prevalence of breed, from Most to Least

More Common — Less Common

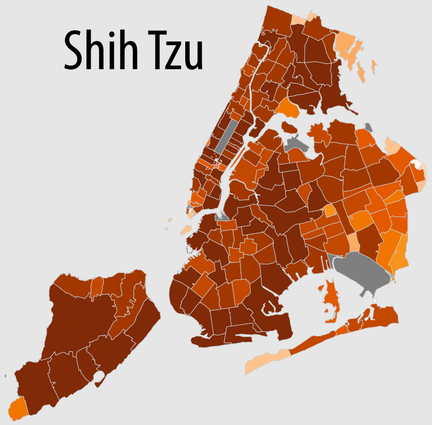
The most common NYC breed is "Mixed" or "Unknown"



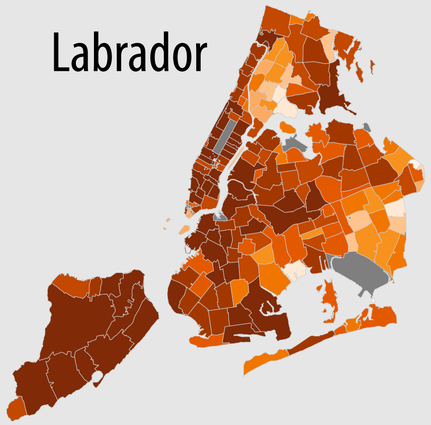
Yorkie



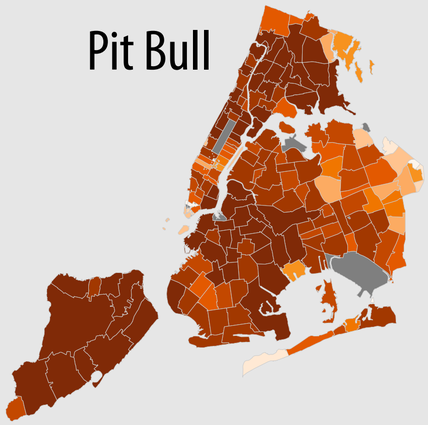
Shih Tzu



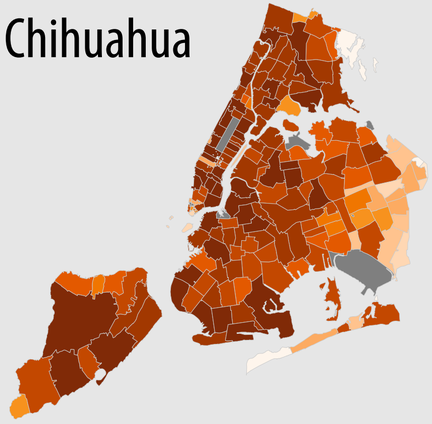
Labrador



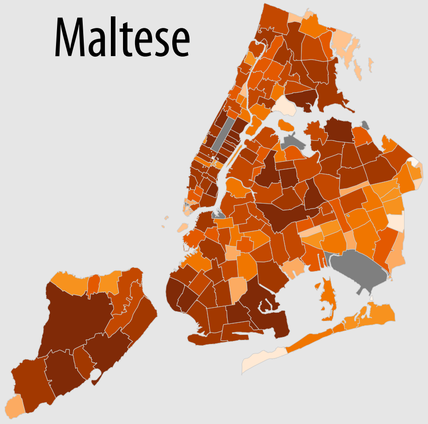
Pit Bull



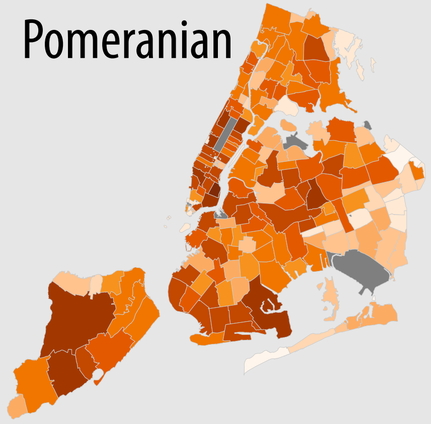
Chihuahua



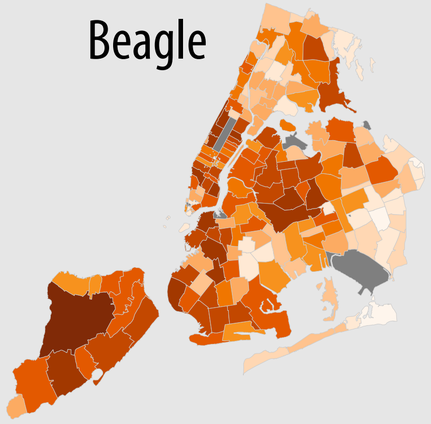
Maltese



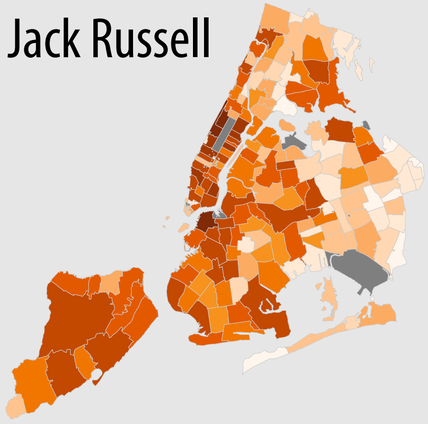
Pomeranian



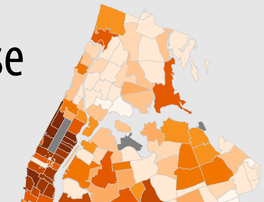
Beagle



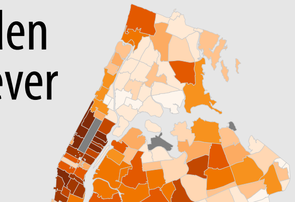
Jack Russell



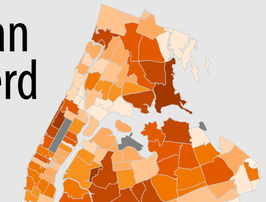
Havanese



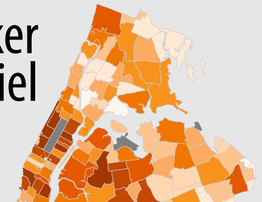
Golden Retriever



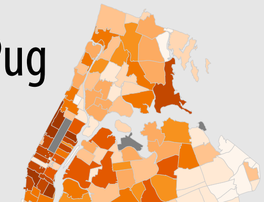
German Shepherd



Cocker Spaniel

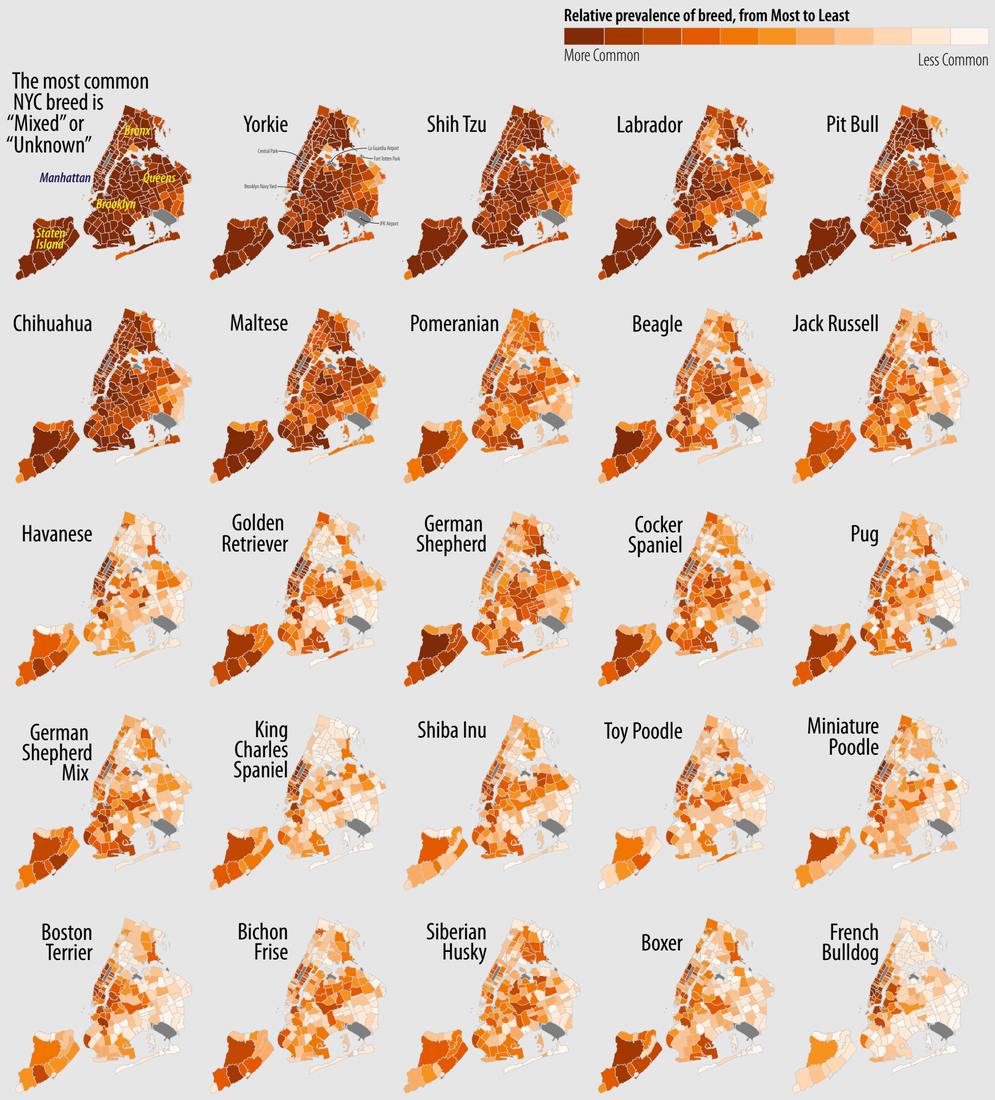


Pug



Dogs of New York

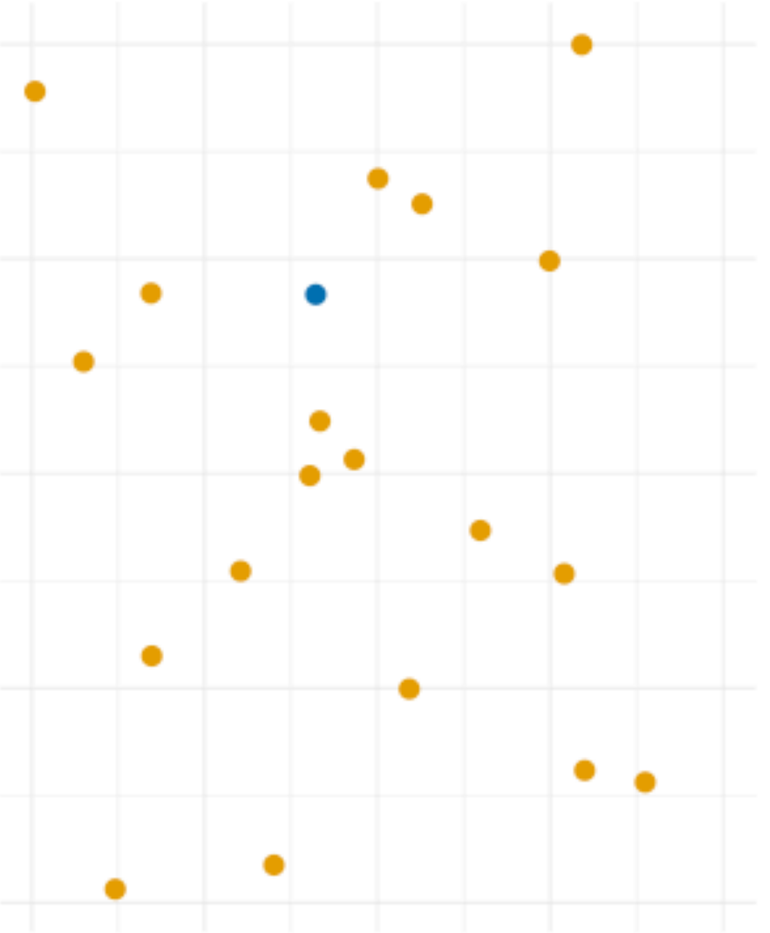
Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.



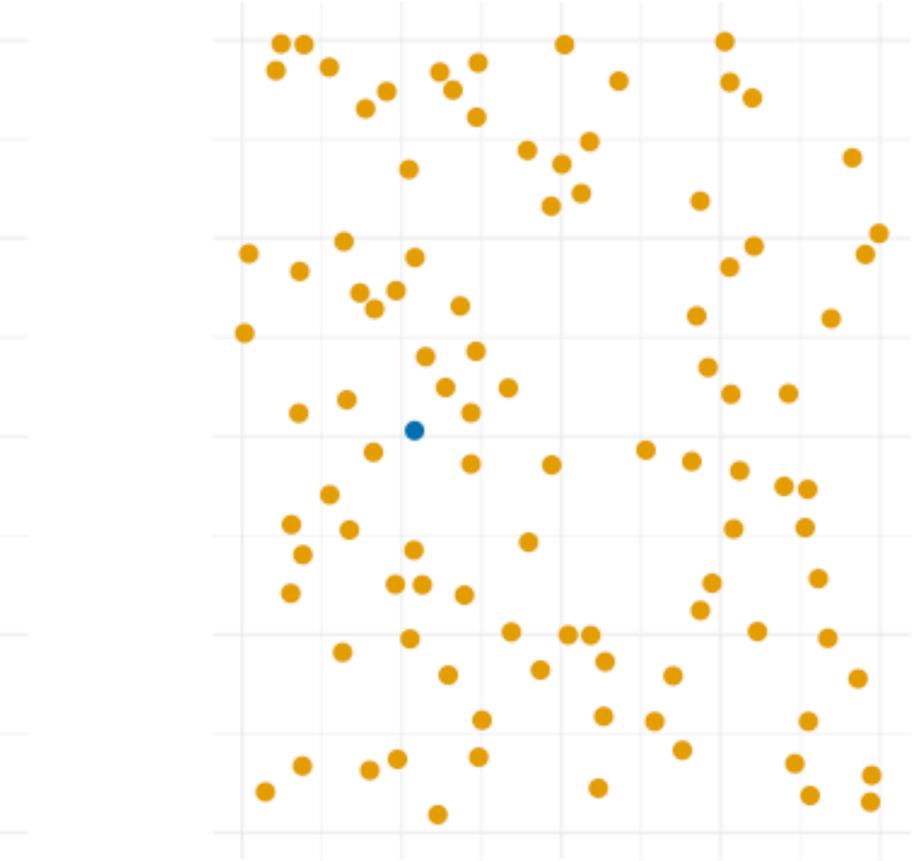
Kieran Healy / socviz.co / *Data Visualization: A Practical Introduction* is published by Princeton University Press

PRE-ATTENTIVE PROCESSING

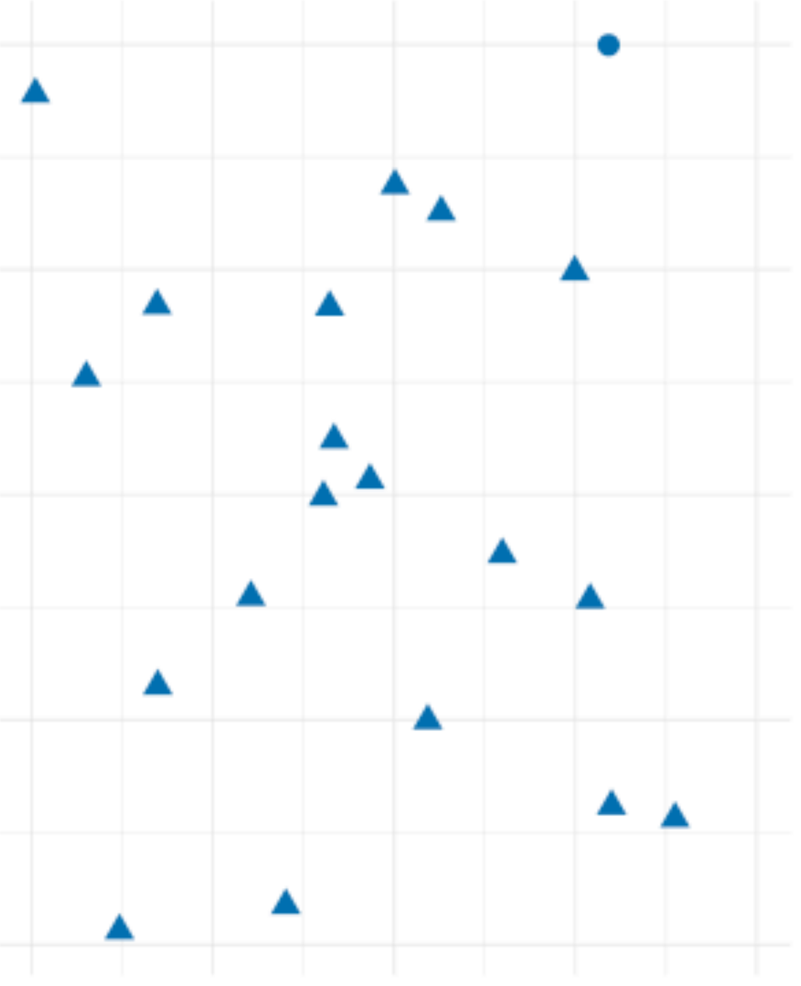
Color Only, N=20



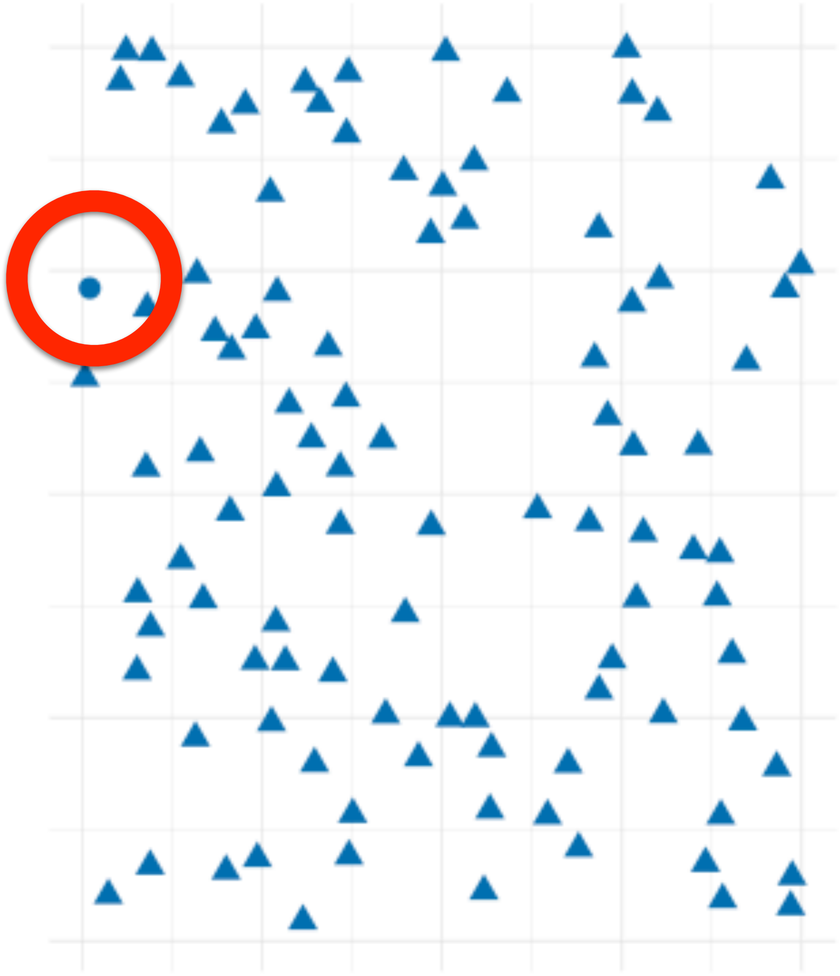
Color Only, N=100



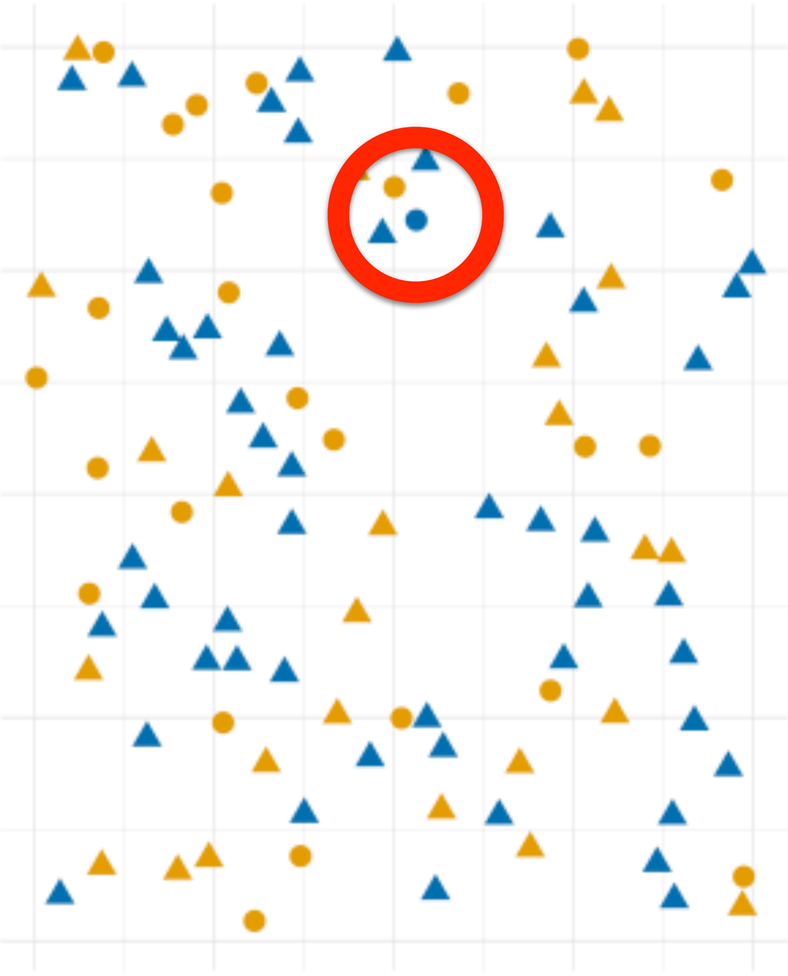
Shape Only, N=20

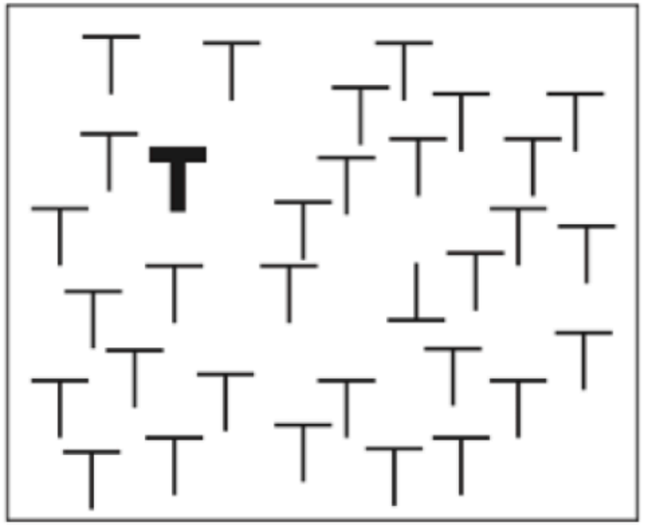


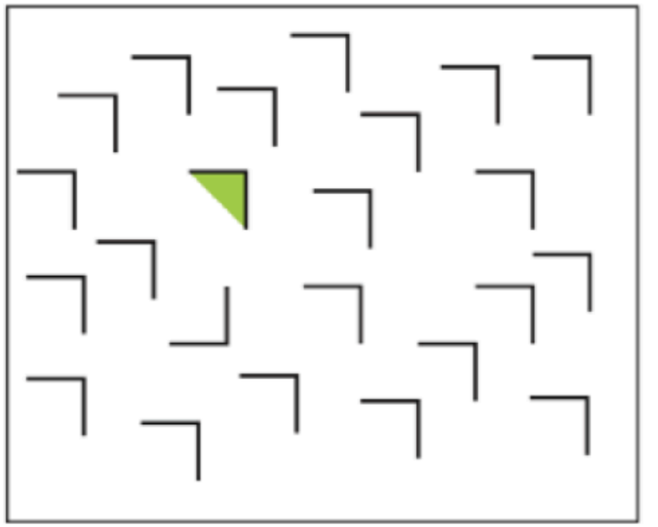
Shape Only, N=100

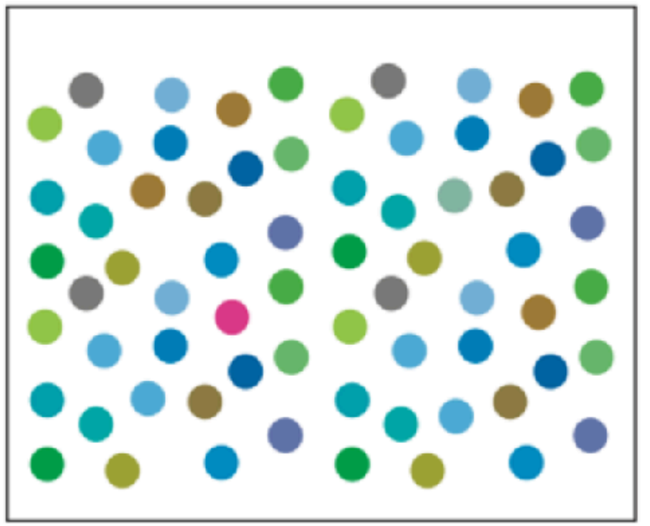


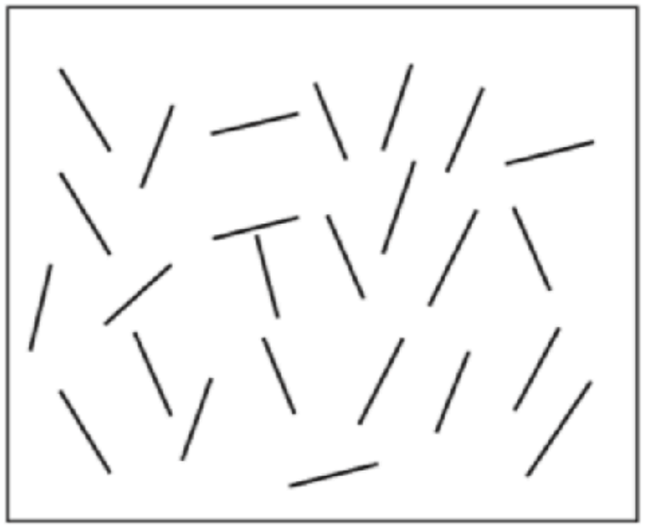
Color & Shape, N=100

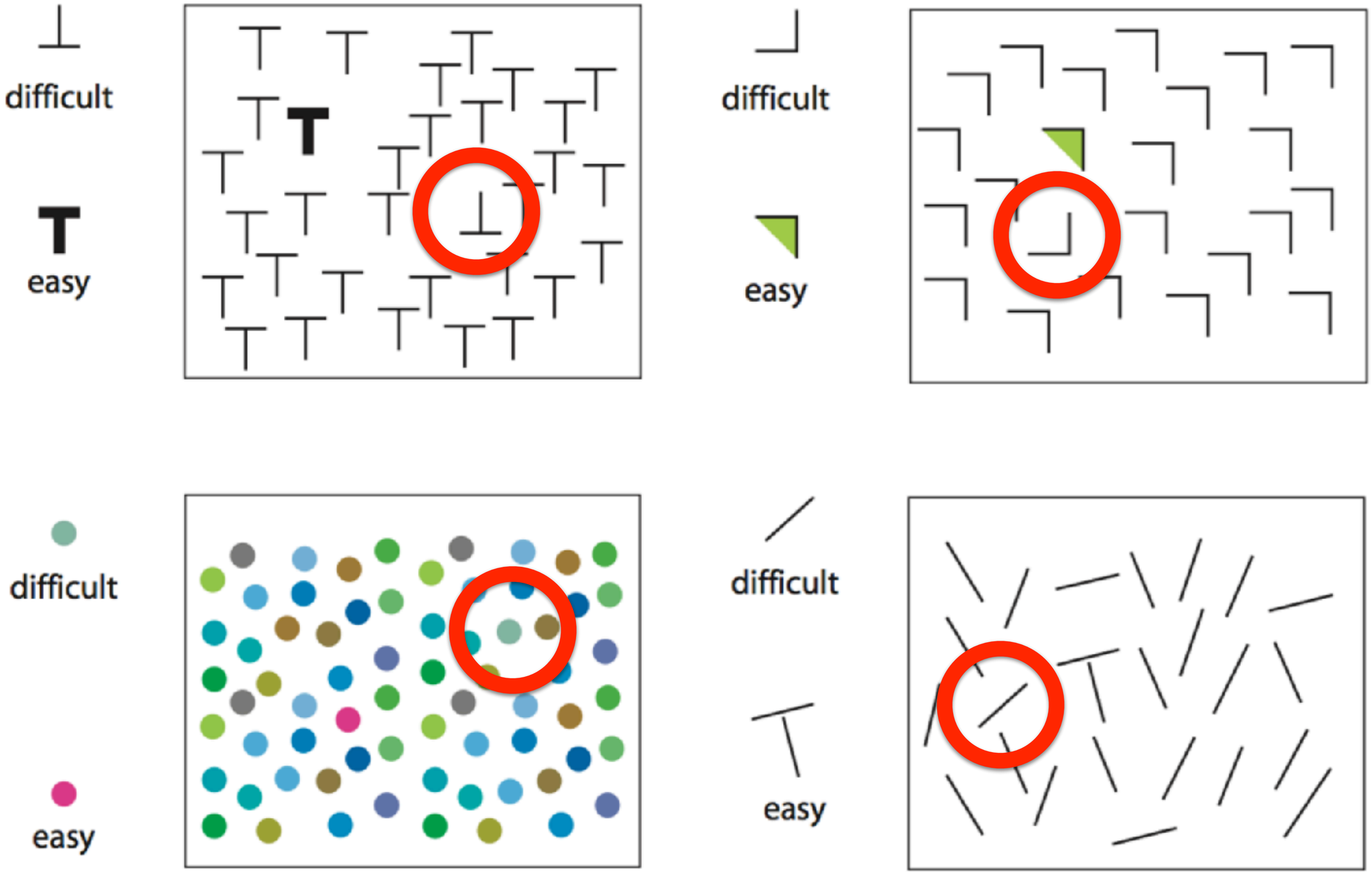


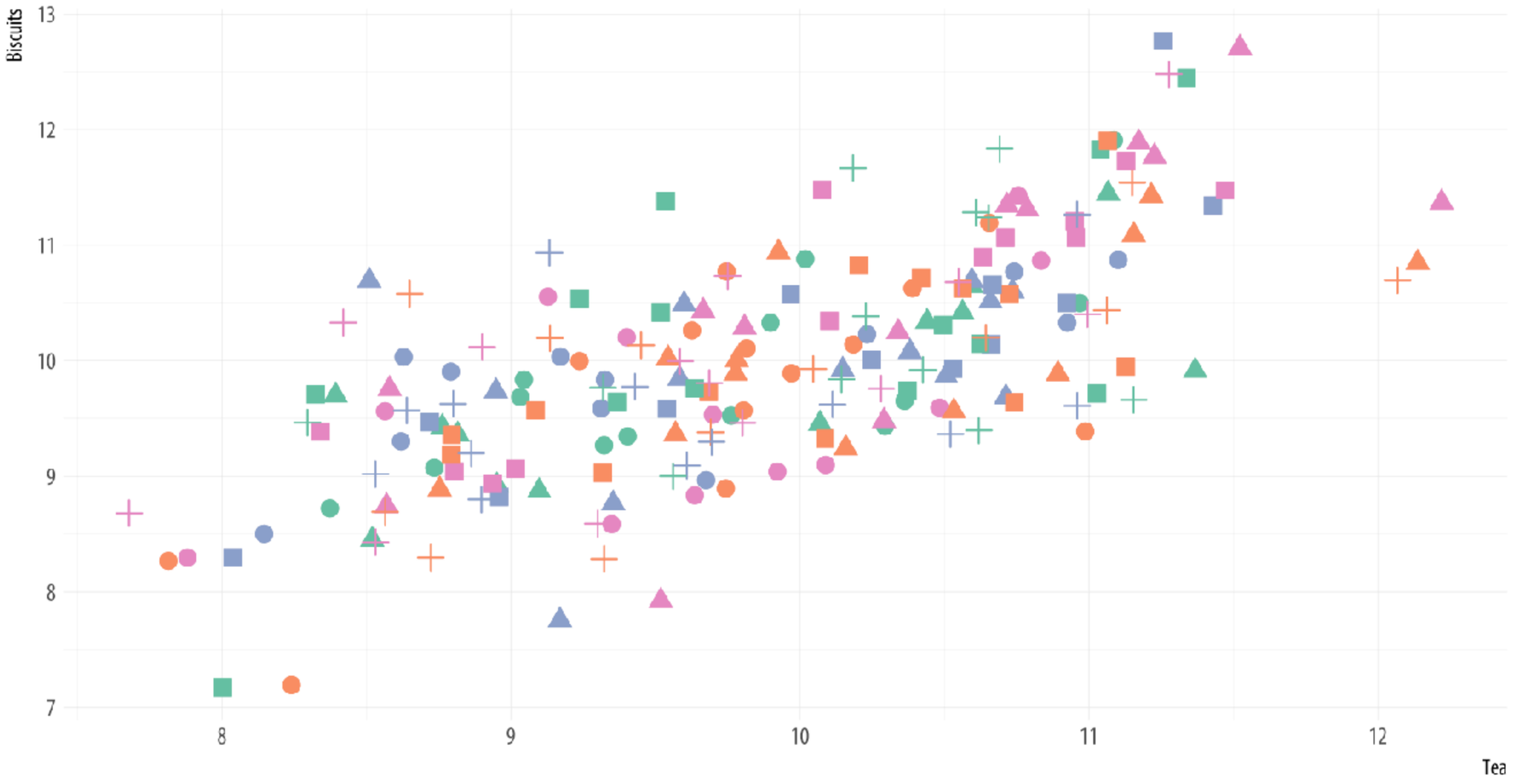


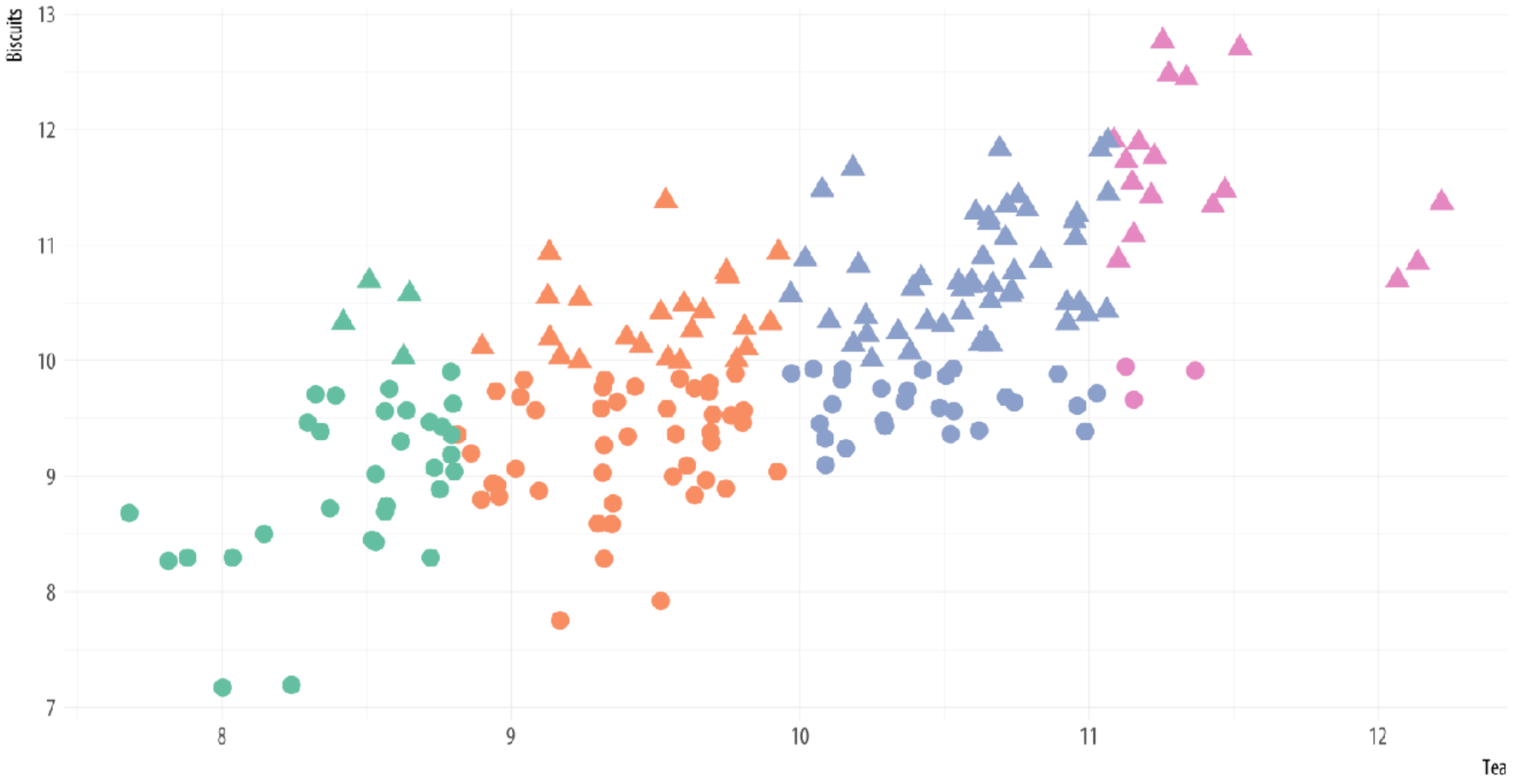






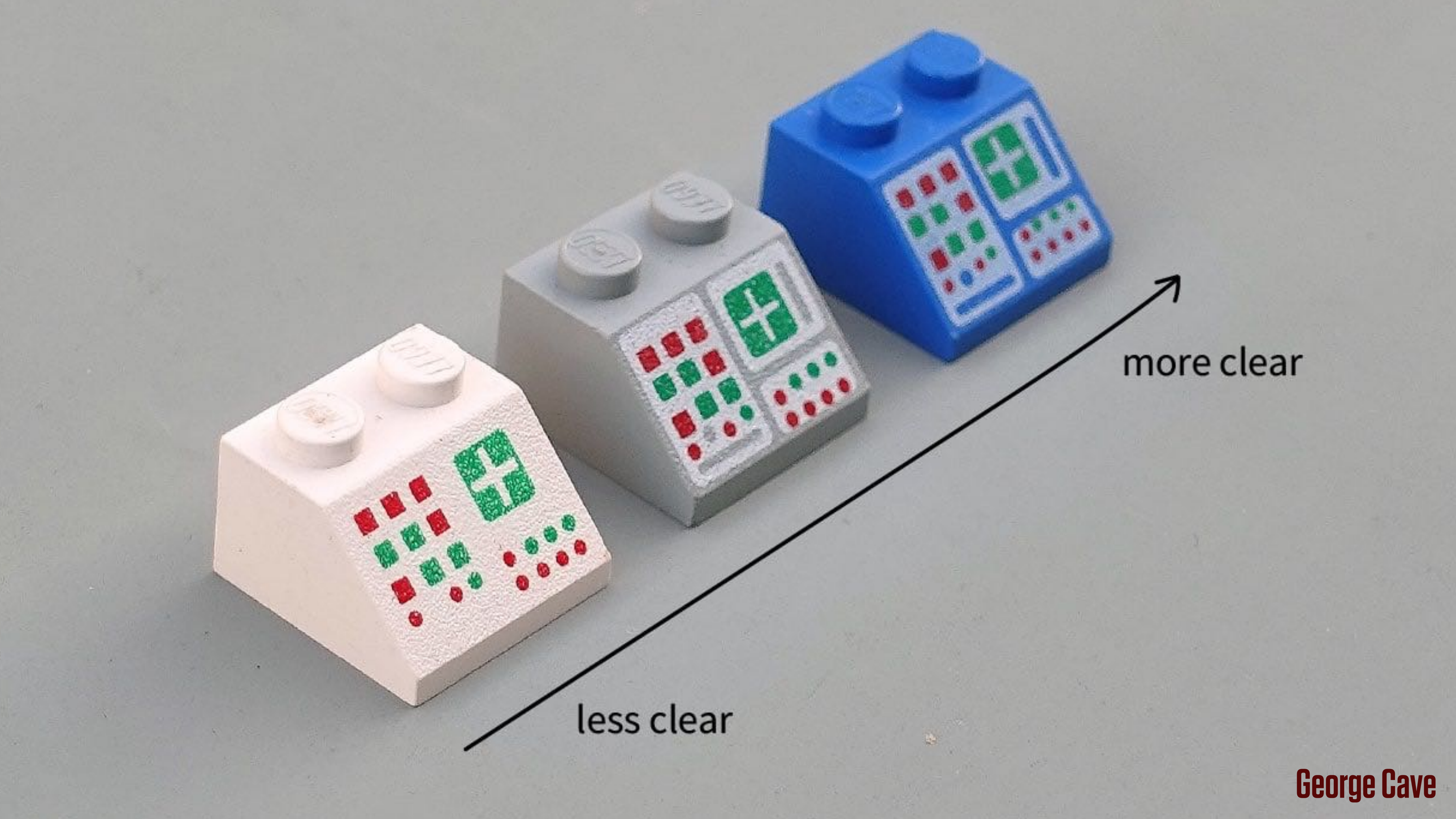

difficult
easy
difficult
easy
difficult
easy
difficult
easy


Biscuits
13
12
11
10
9
8
7
8
9
10
11
12
Tea


Biscuits
13
12
11
10
9
8
7
8
9
10
11
12
Tea

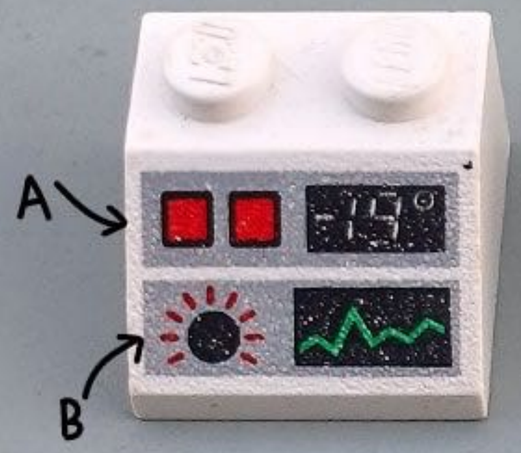
GESTALT INFERENCES

George Cave

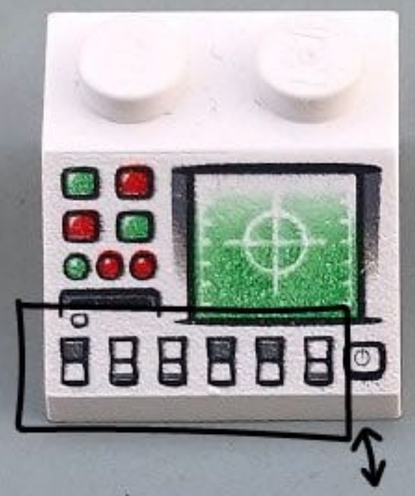


George Cave

by feature

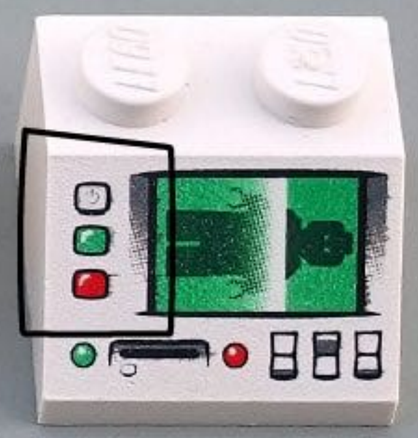


by operation

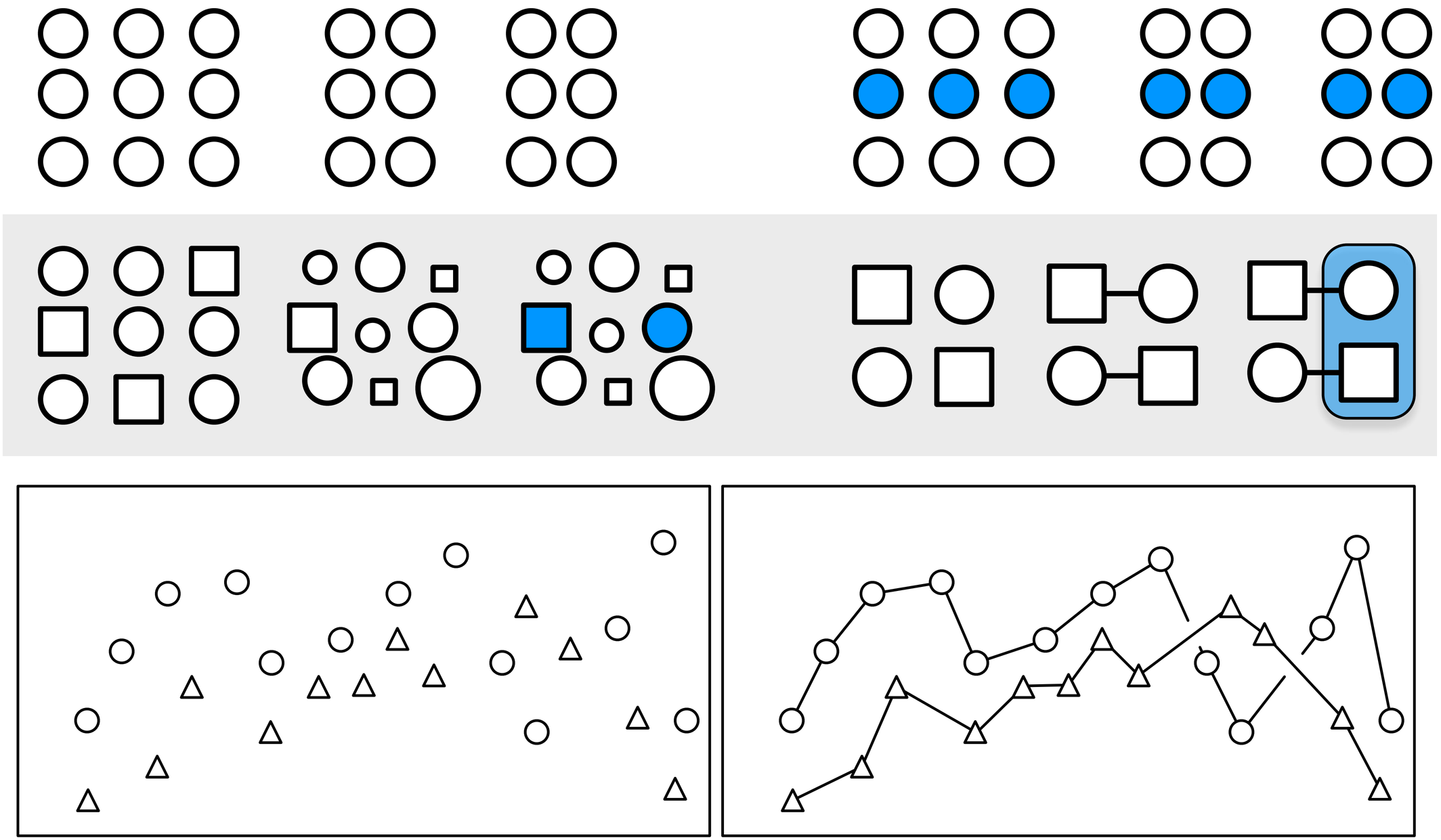


by technology

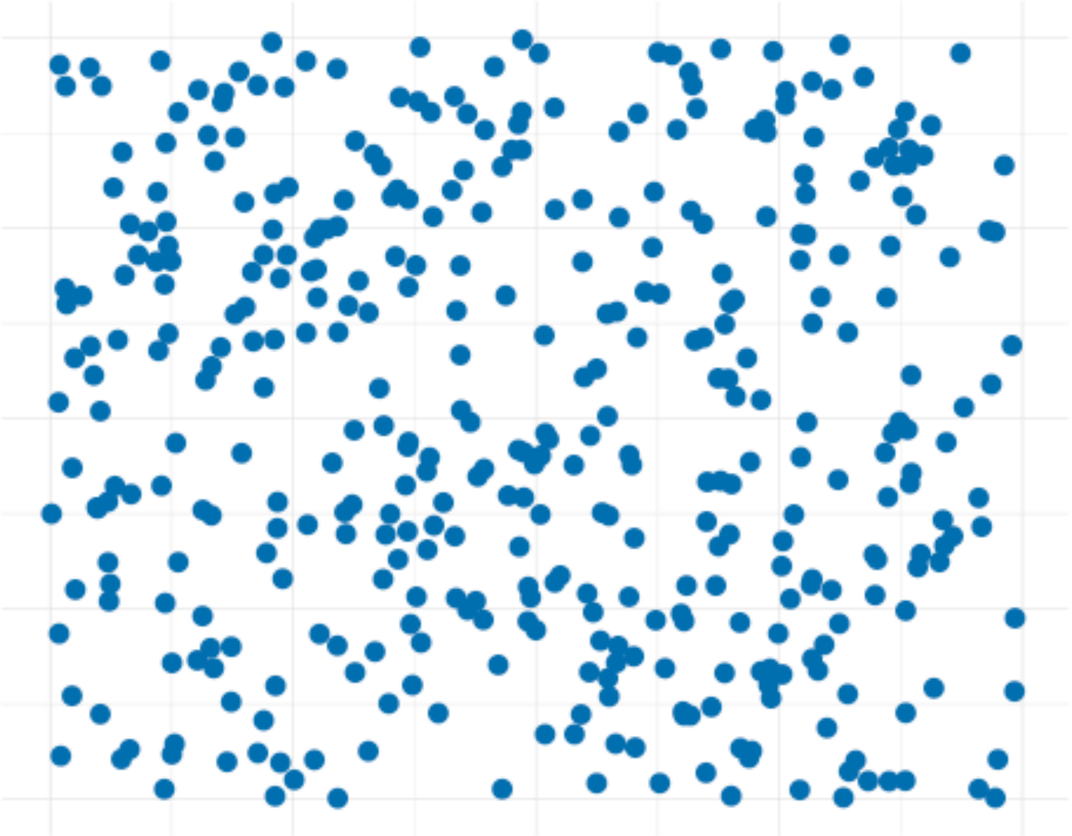
by use-case



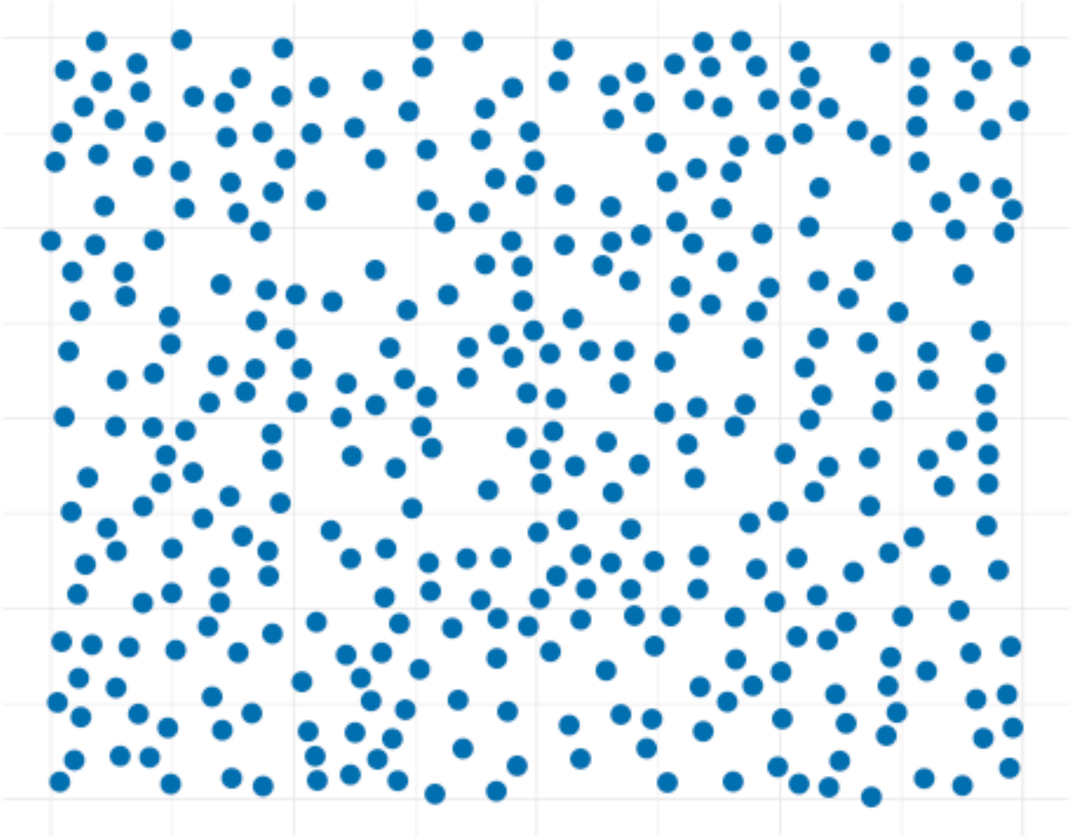
George Cave



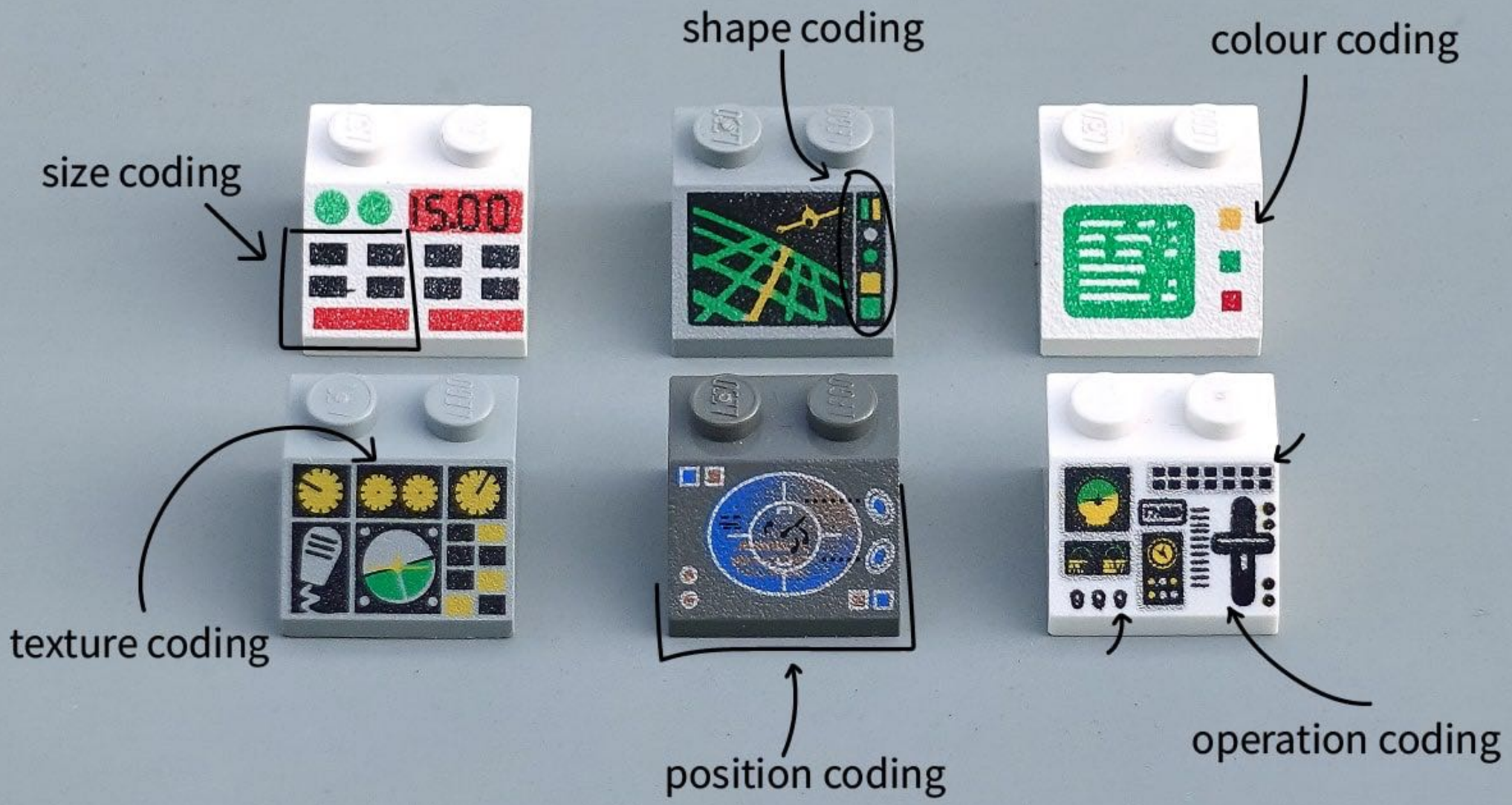
Poisson



Matérn

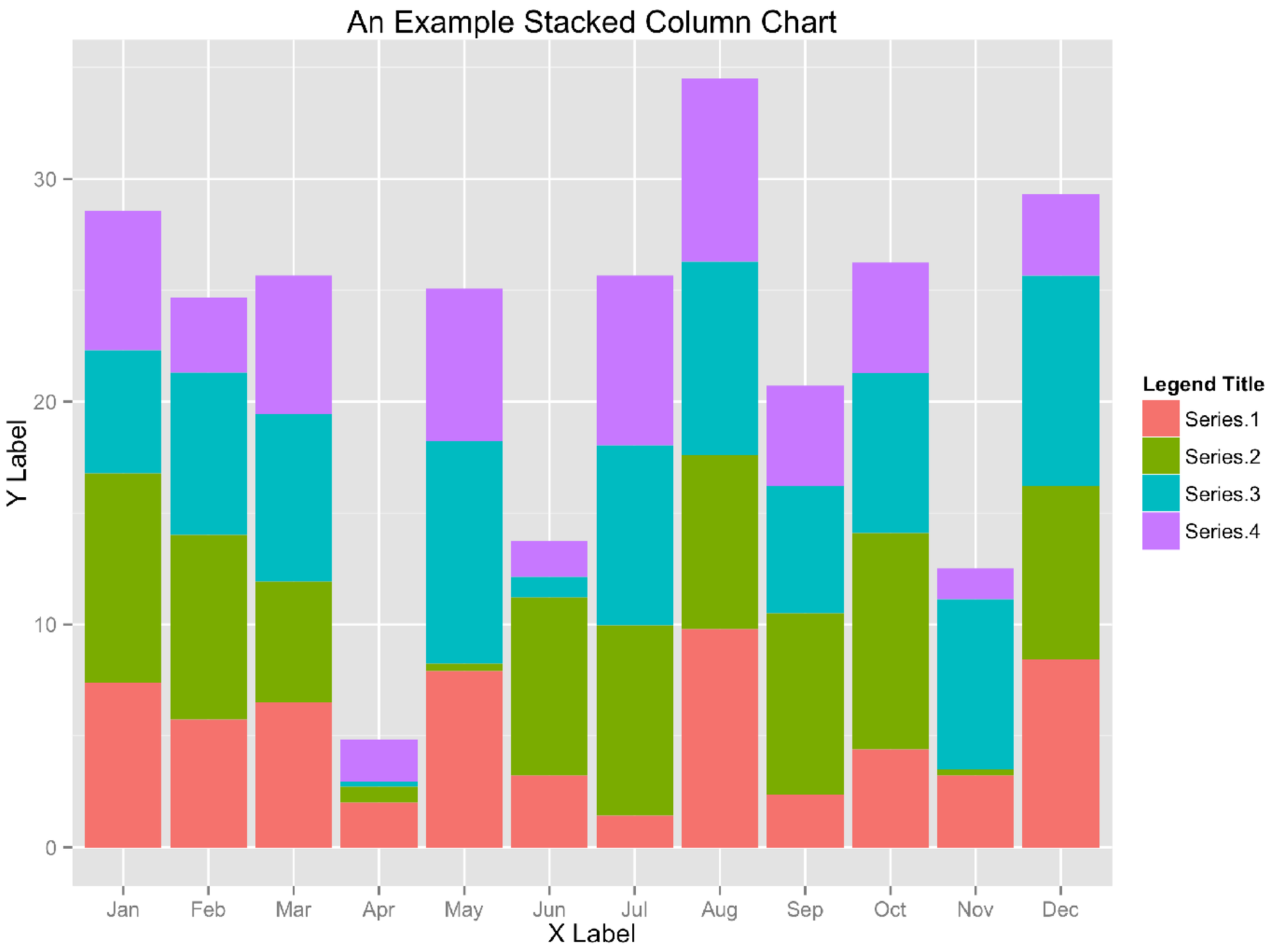


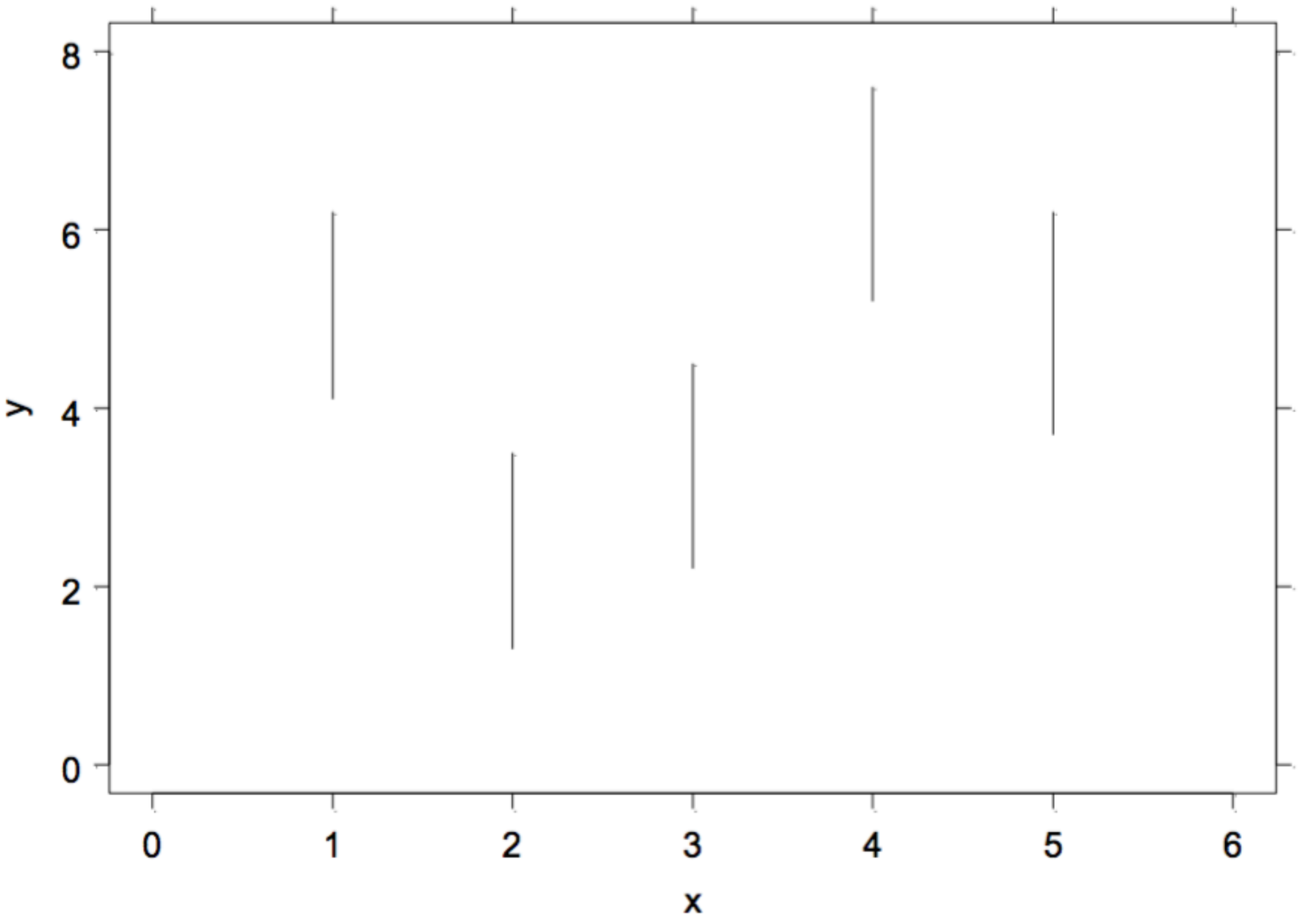
MAPPINGS FOR DATA

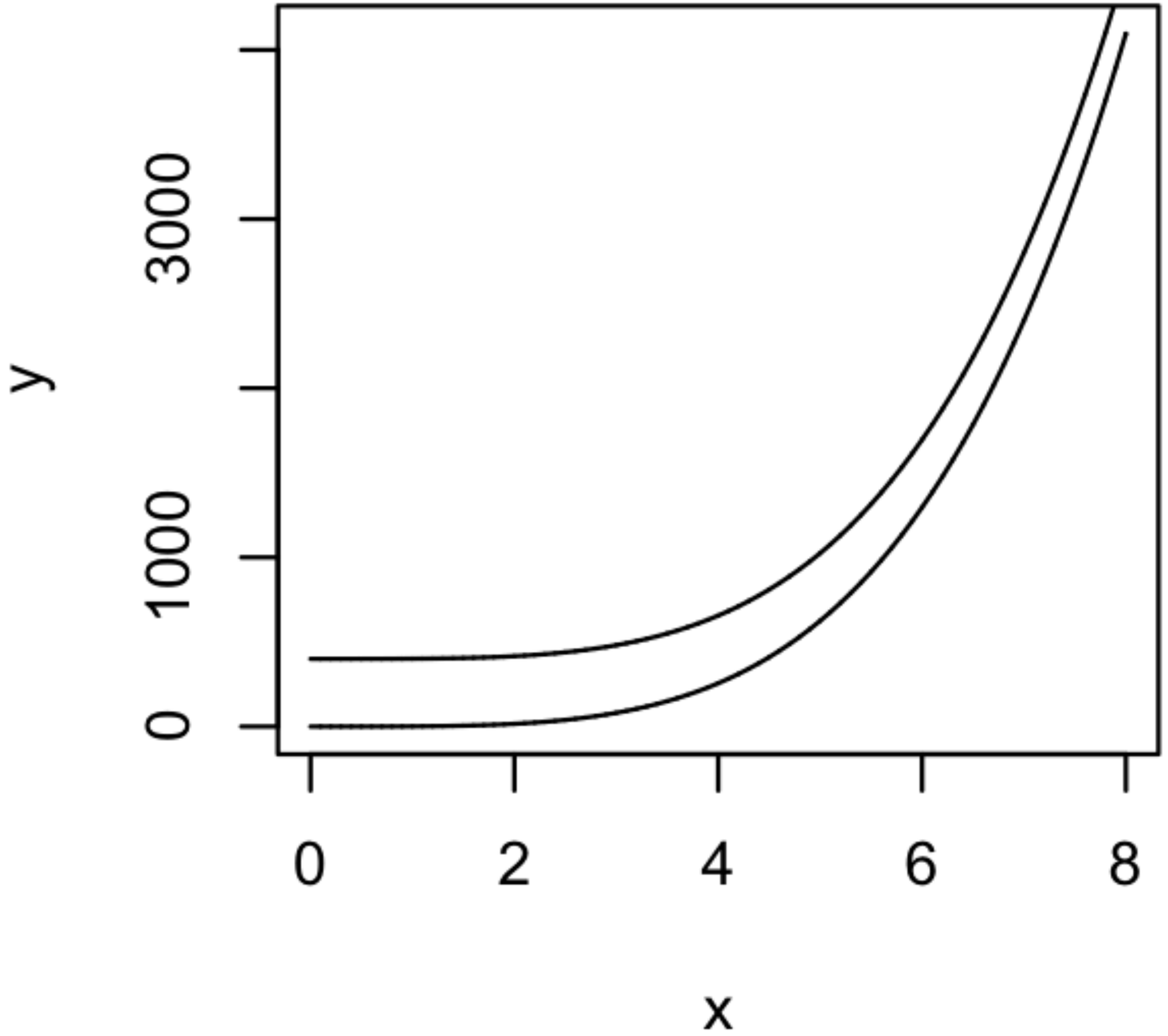

shape coding
colour coding
size coding
15.00
texture coding
position coding
operation coding

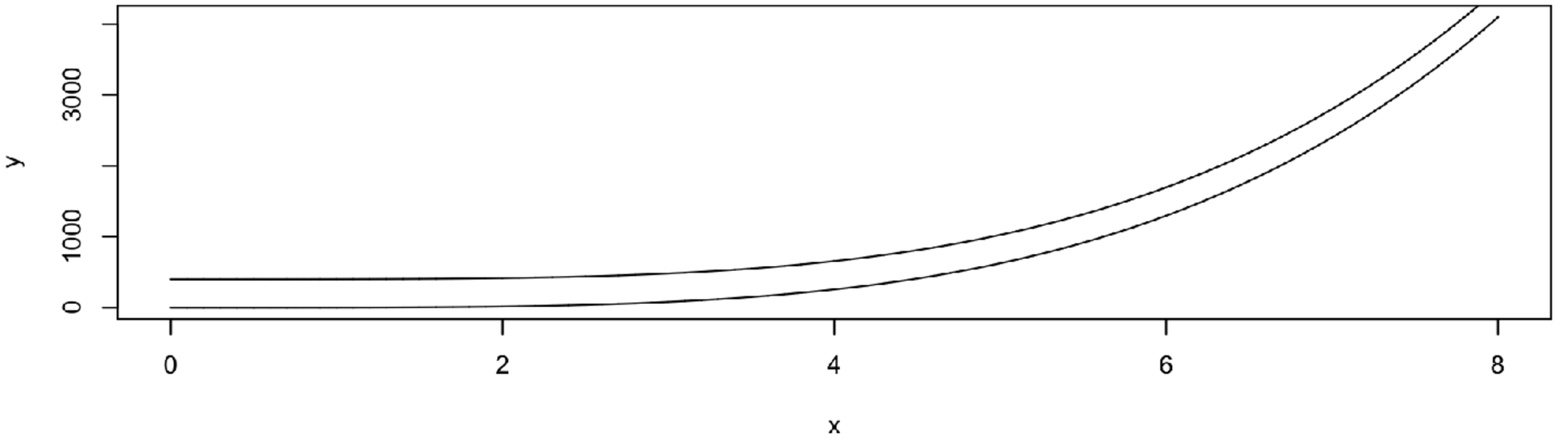
George Cave

Visual Tasks in Decoding Graphs


An Example Stacked Column Chart
Y Label
30
20
10
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
X Label
Legend Title
Series.1
Series.2
Series.3
Series.4


8
6
4
2
0
y
0
1
2
3
4
5
6
x


y
0
1000
3000
0
2
4
6
8
x


y
0
1000
3000
0
2
4
6
8
x

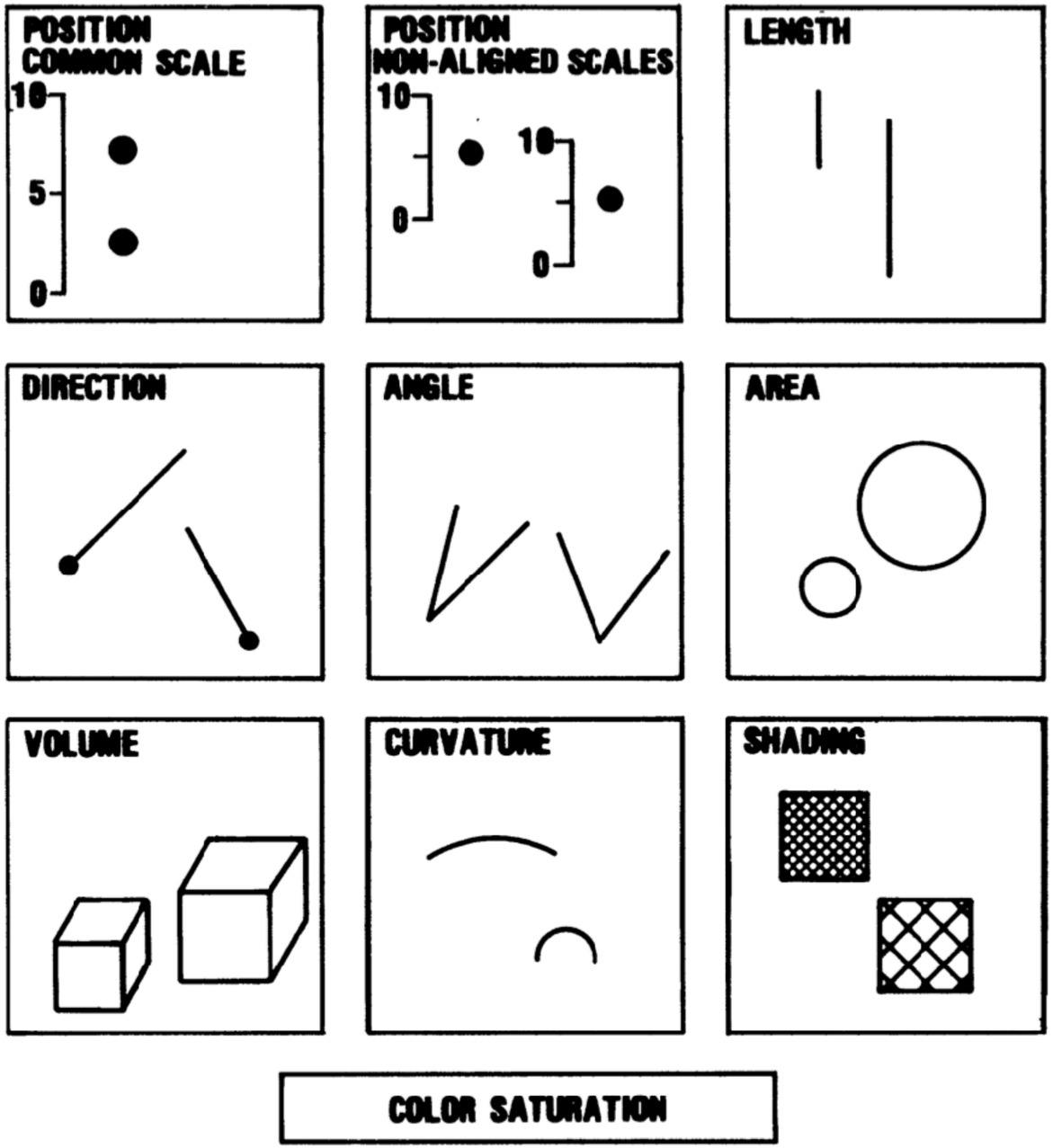


Figure 1. Elementary perceptual tasks.

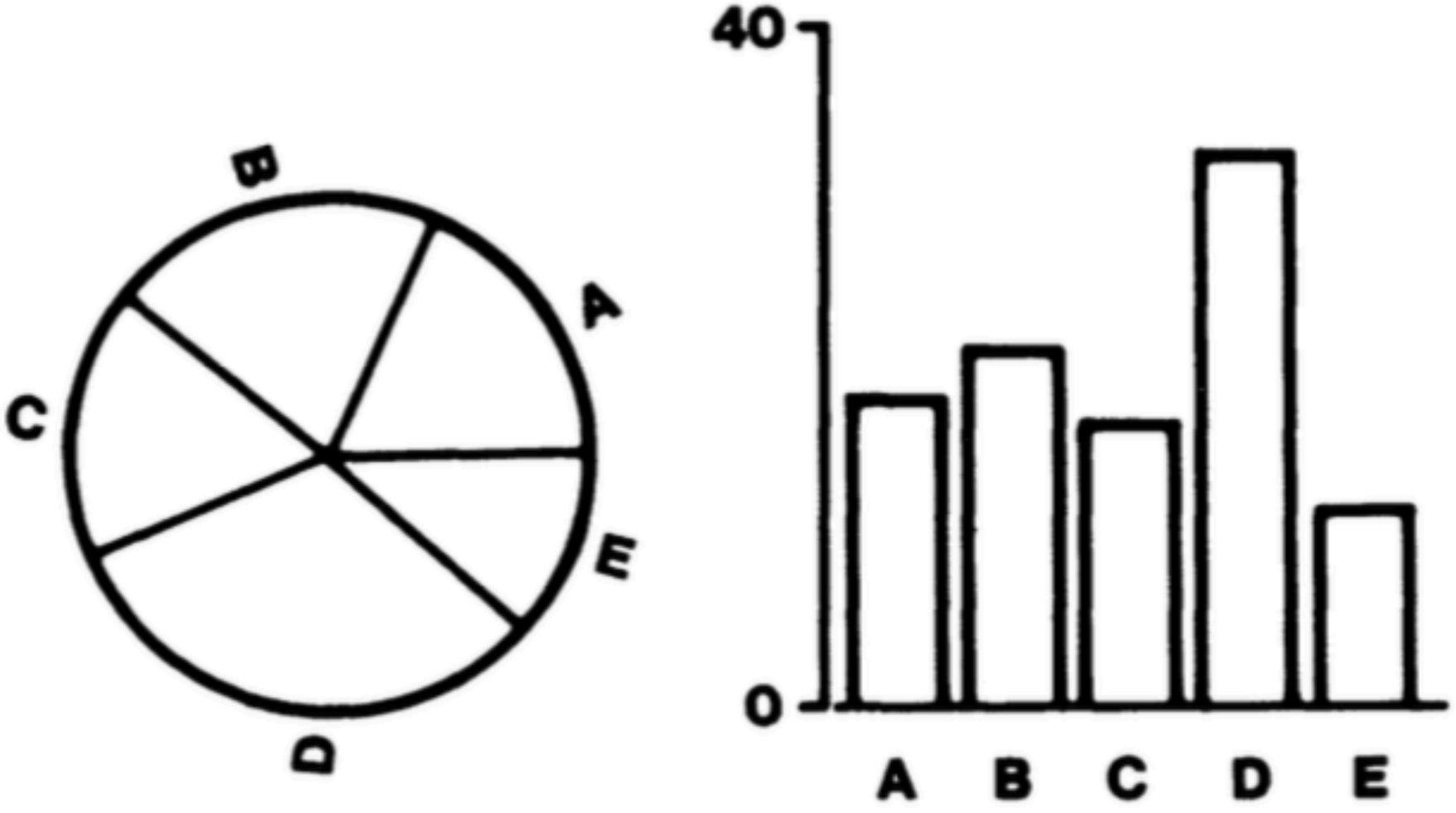


Figure 3. Graphs from position–angle experiment.

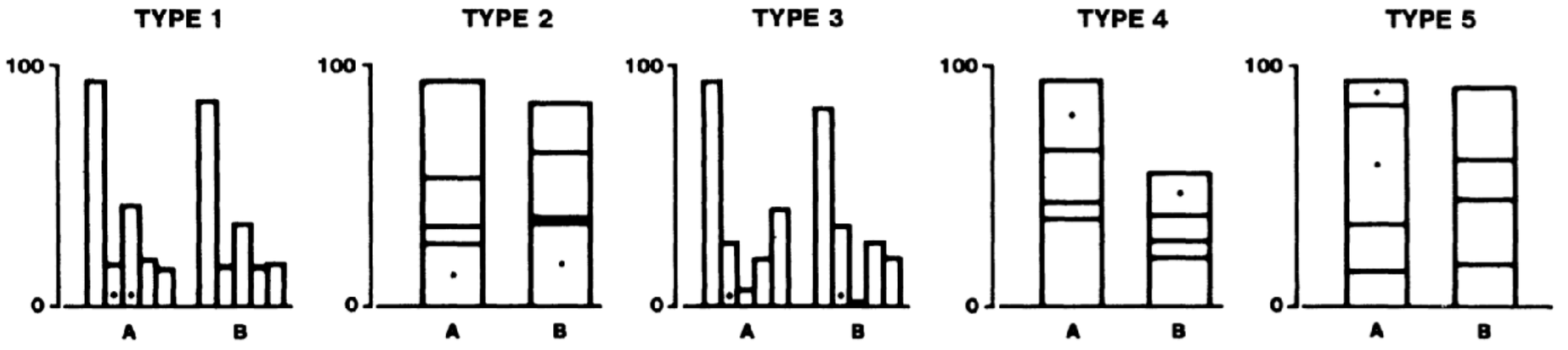
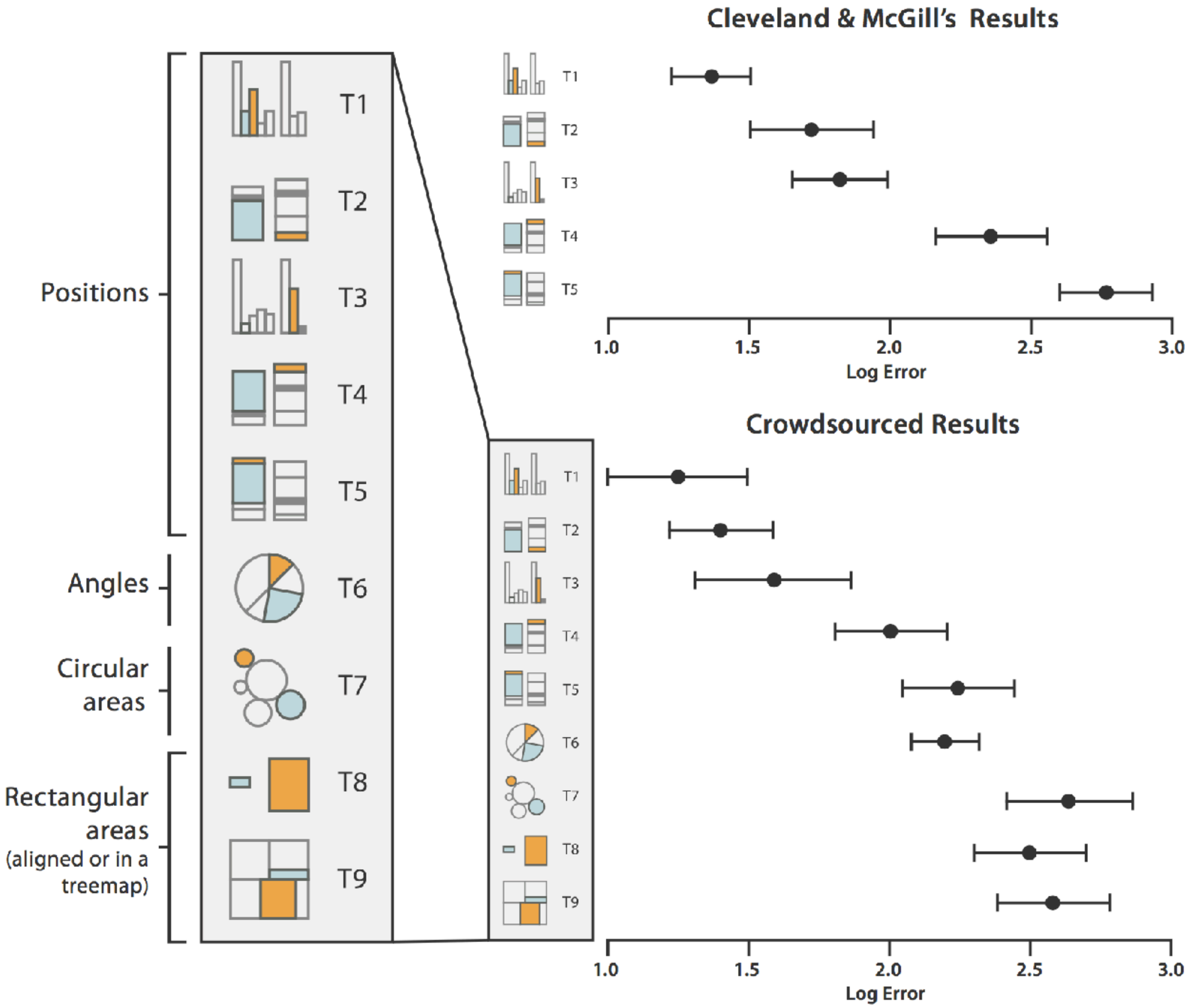
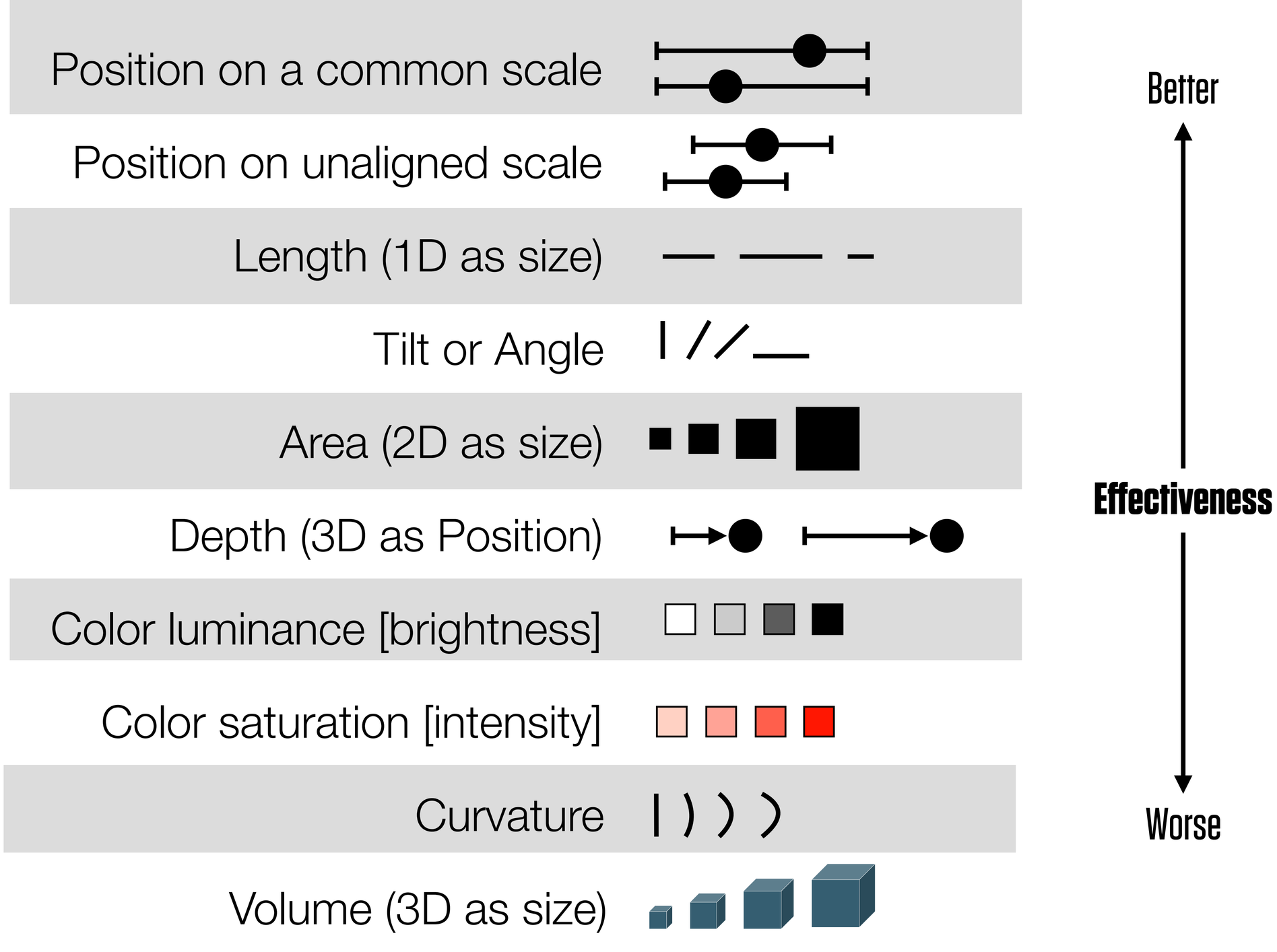


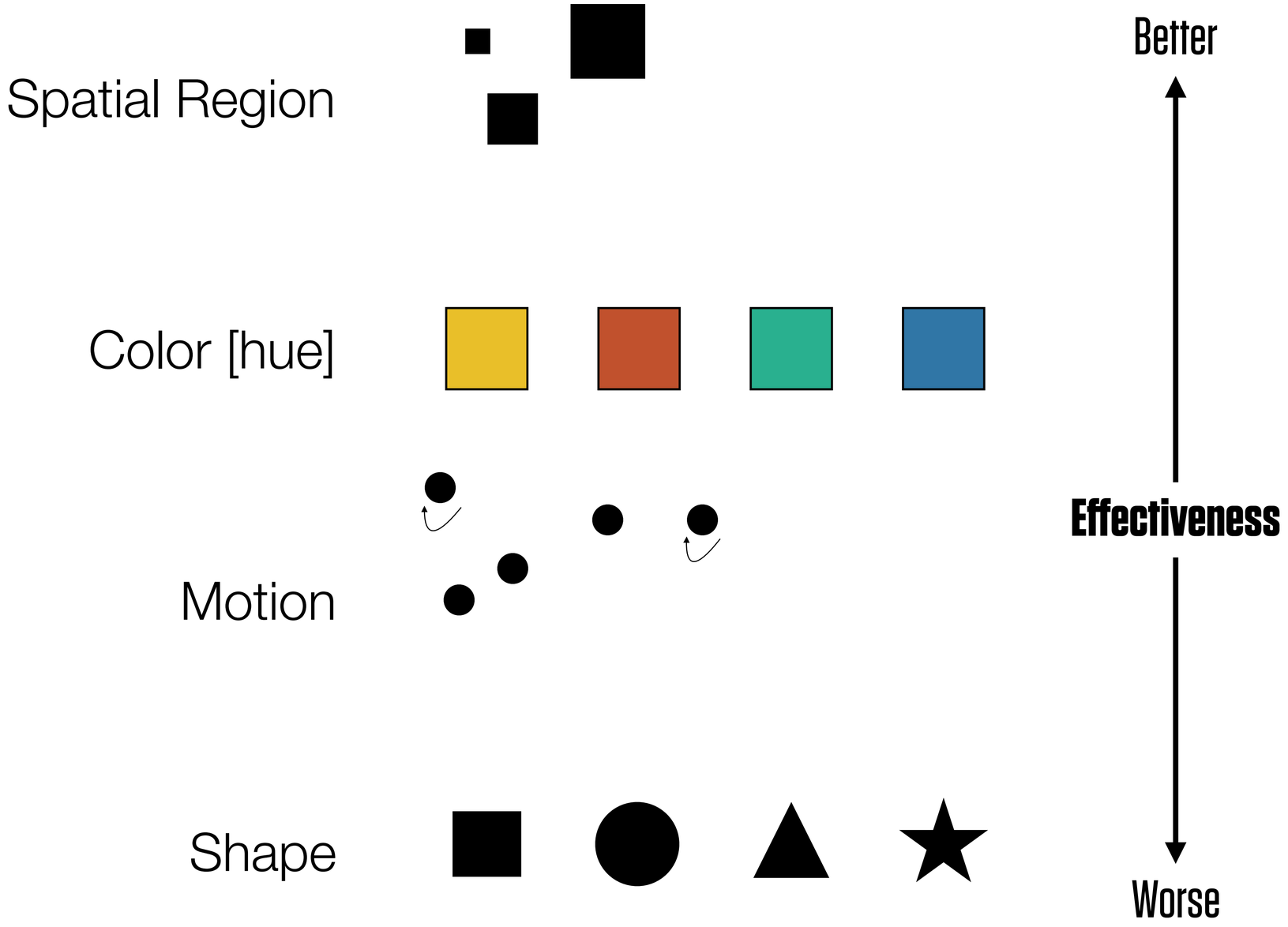
Figure 4. Graphs from position–length experiment.


Cleveland & McGill's Results
Crowdsourced Results
Positions
Angles
Circular areas
Rectangular areas (aligned or in a treemap)
T1
T2
T3
T4
T5
T6
T7
T8
T9
1.0
1.5
2.0
2.5
3.0
Log Error

A Rough Hierarchy of Mappings



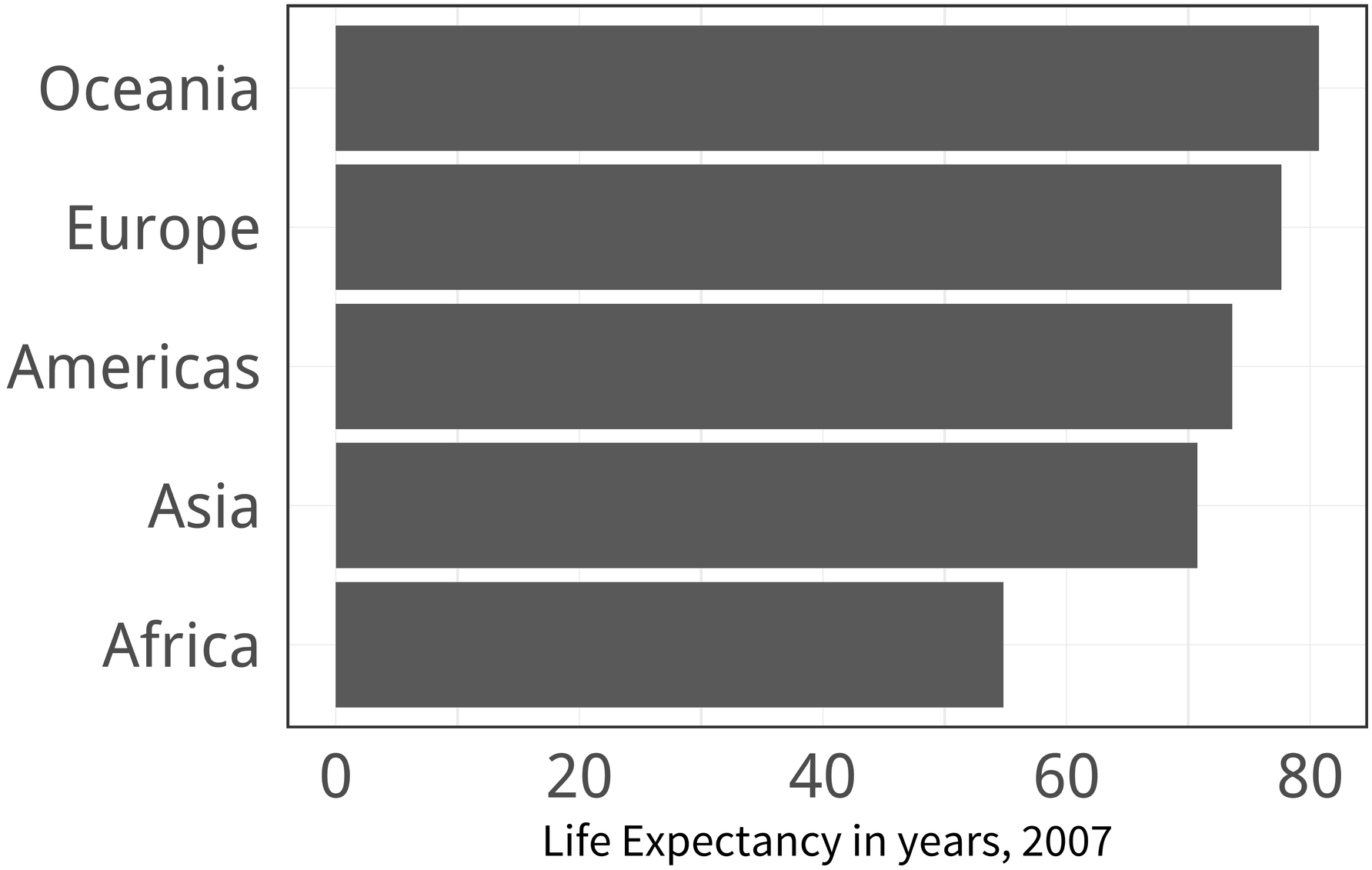
See e.g. Munzer (2014)



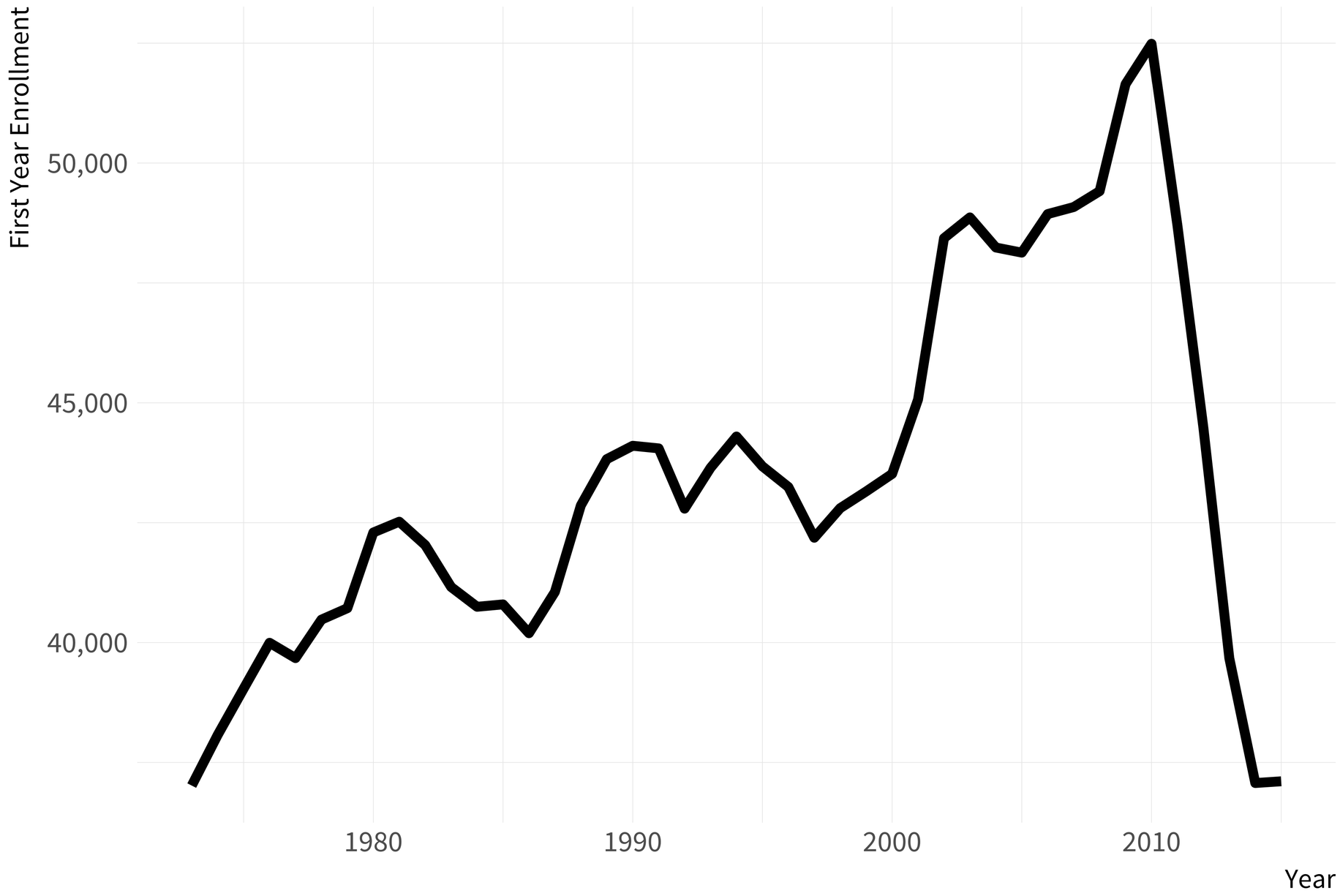
See e.g. Munzer (2014)

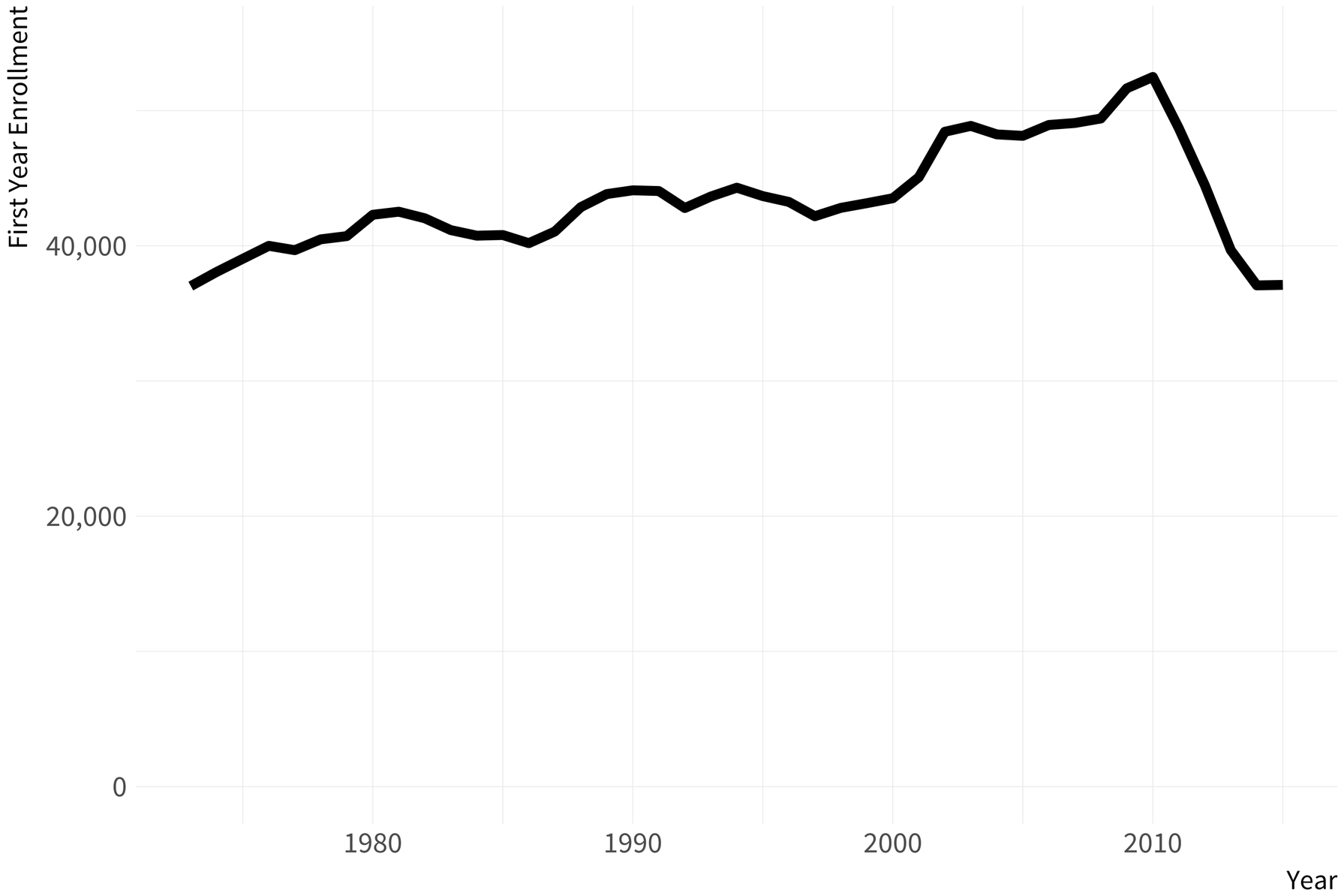
HONESTY AND GOOD JUDGMENT

Example: Baselines


Oceania
Europe
Americas
Asia
Africa
0
20
40
60
80
Life Expectancy in years, 2007

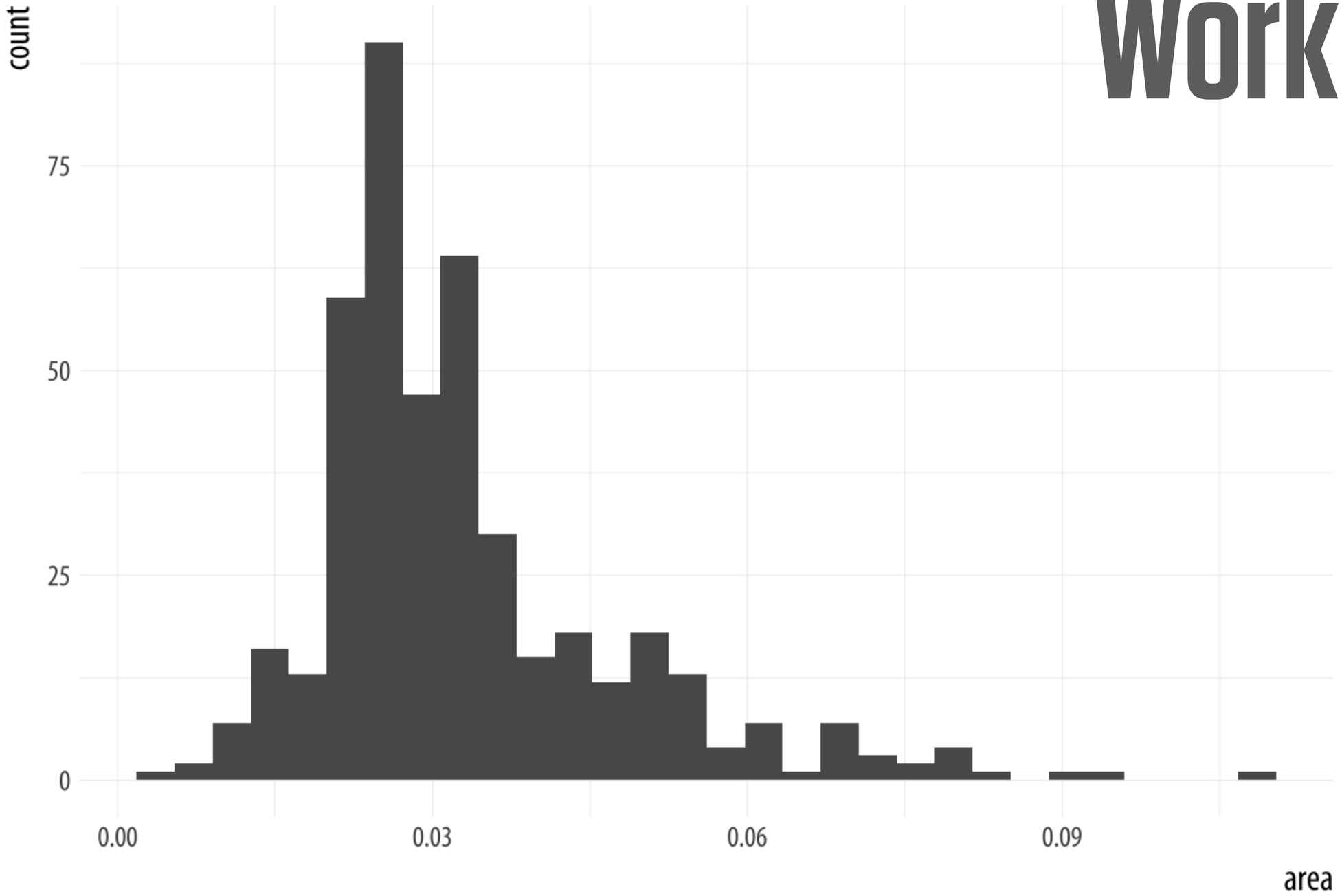
Oceania
Europe
Americas
Asia
Africa
60
70
80
Life Expectancy in years, 2007

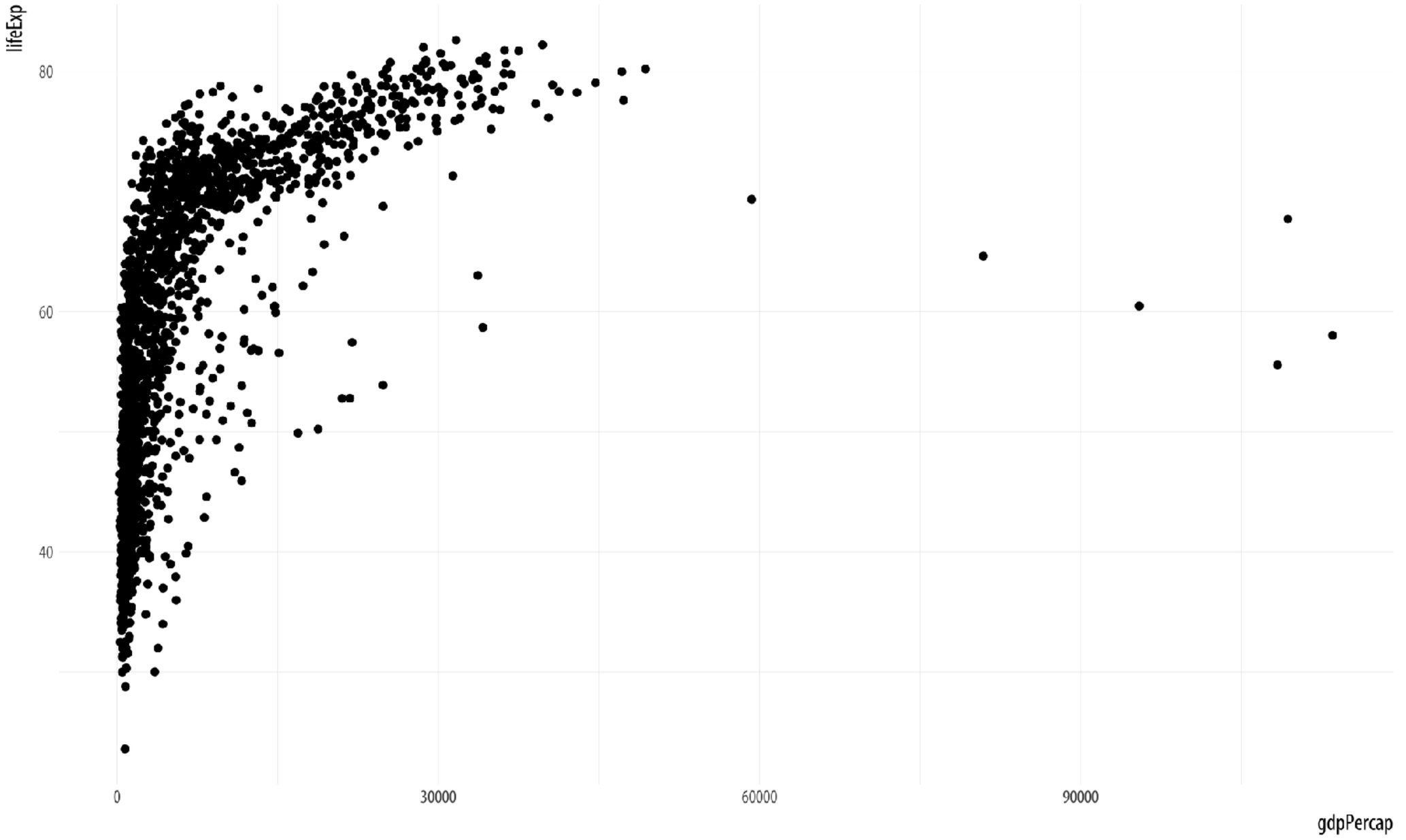

First Year Enrollment
50,000
45,000
40,000
1980
1990
2000
2010
Year

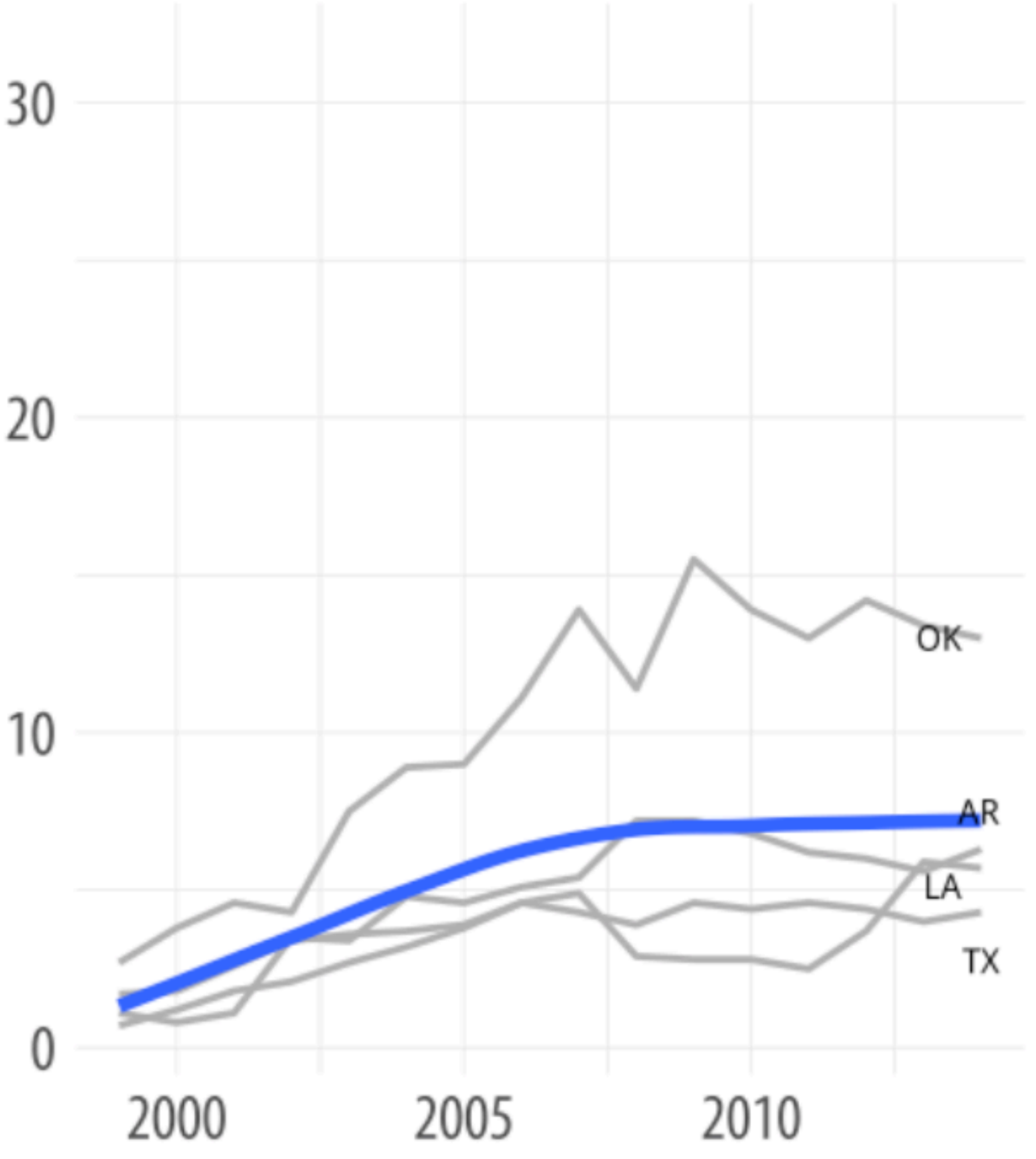

First Year Enrollment
40,000
20,000
0
1980
1990
2000
2010
Year

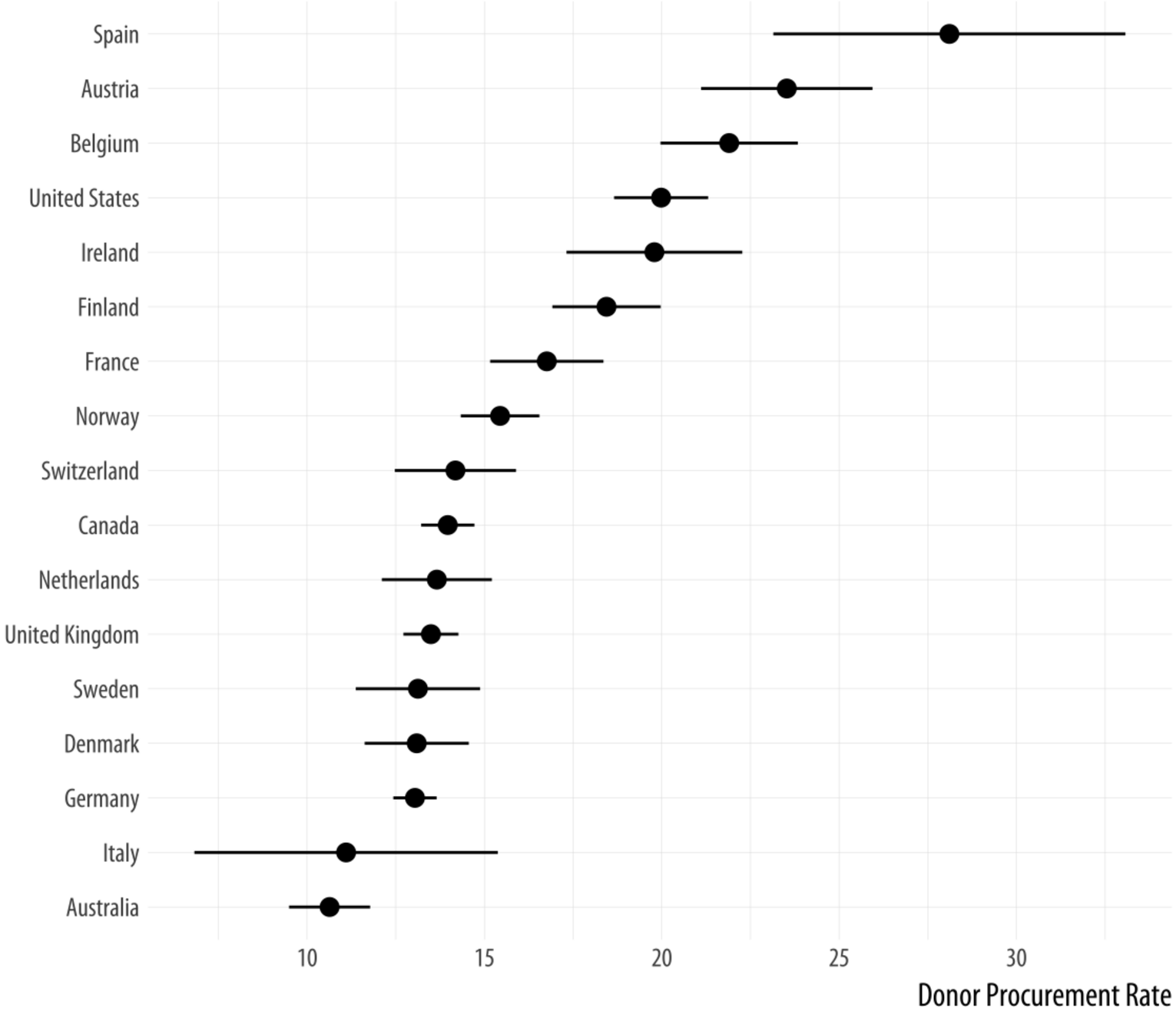
In Practice

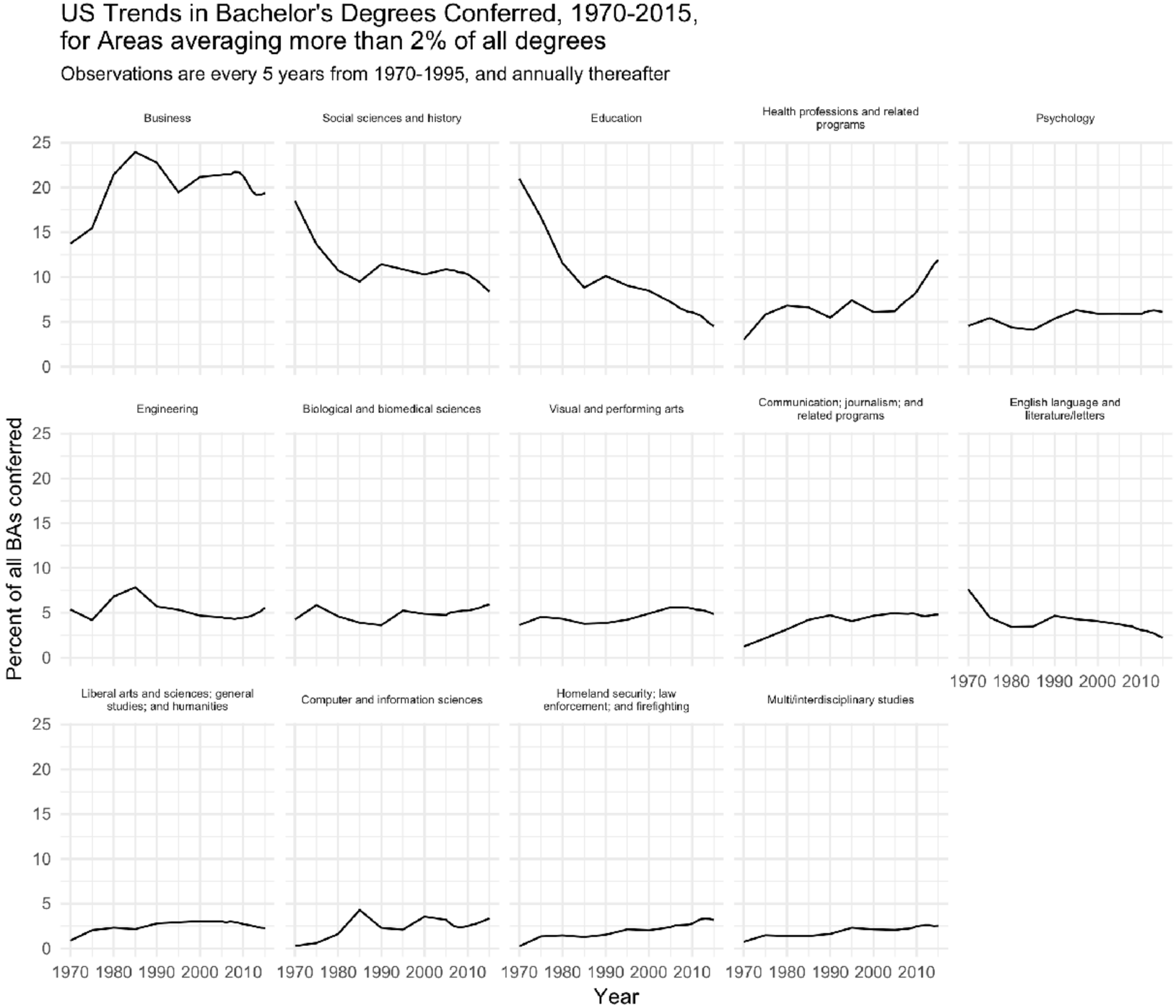
Workhorses




lifeExp
80
60
40
0
30000
60000
90000
gdpPercap

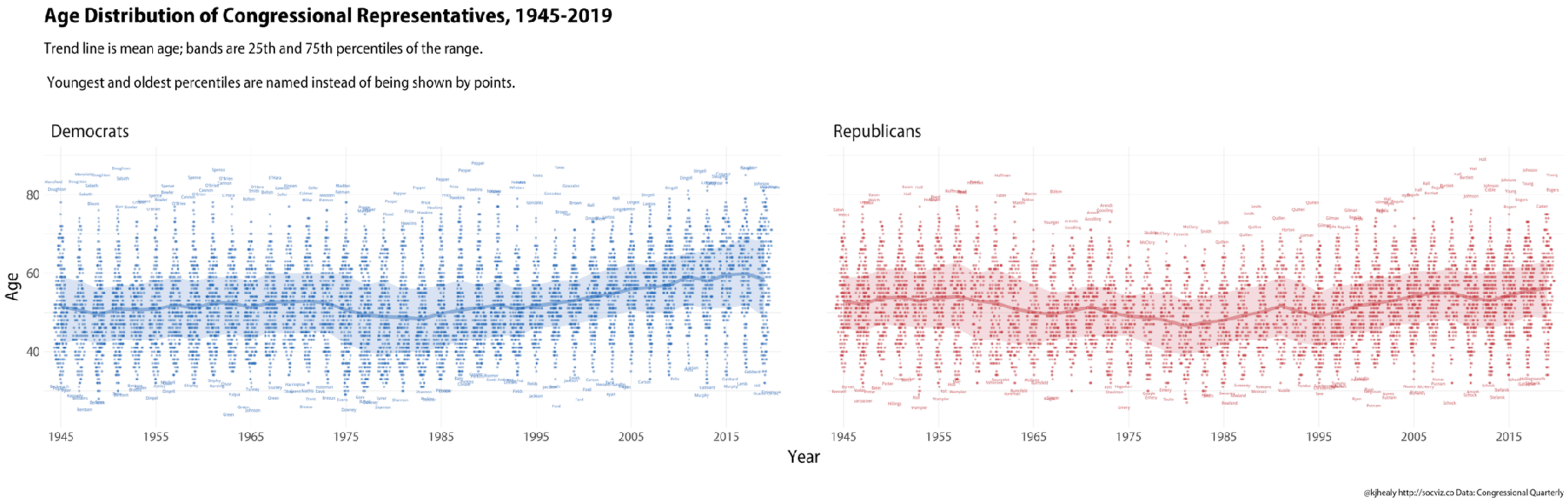

30
20
10
0
2000
2005
2010
OK
AR
LA
TX


Spain
Austria
Belgium
United States
Ireland
Finland
France
Norway
Switzerland
Canada
Netherlands
United Kingdom
Sweden
Denmark
Germany
Italy
Australia
10
15
20
25
30
Donor Procurement Rate

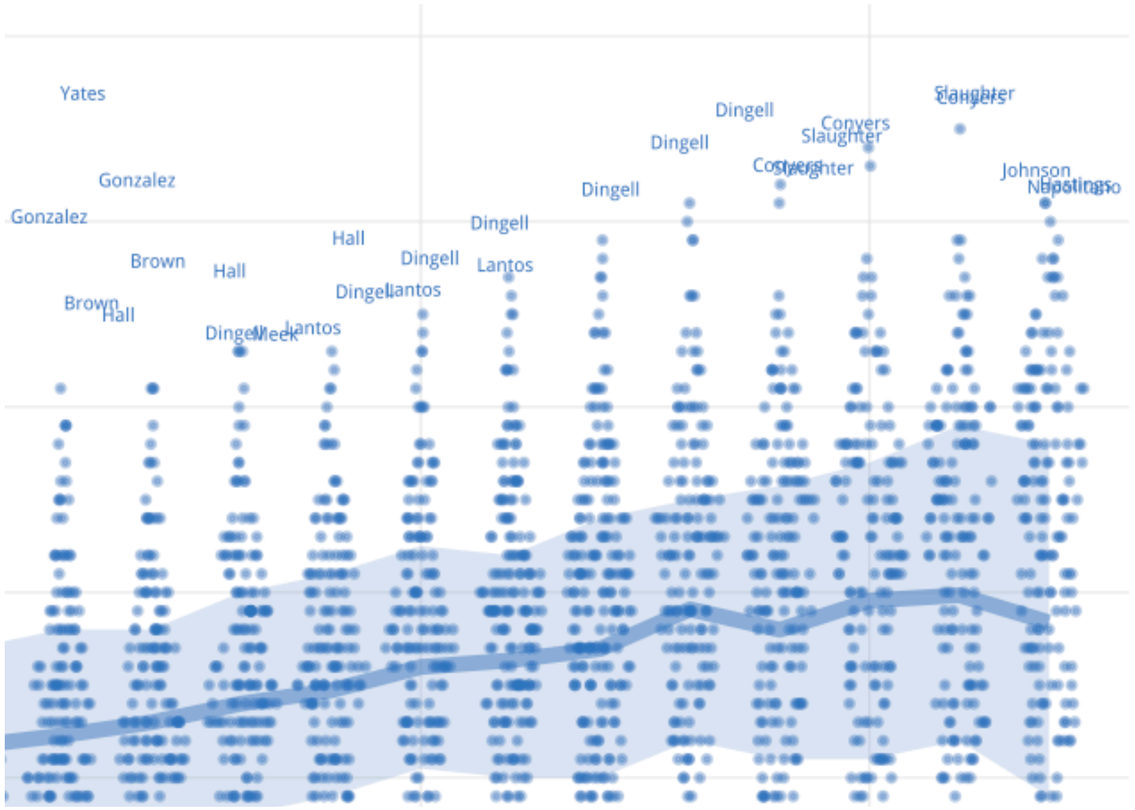


Data from NCES Digest 2017, Table 322.10.

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
Age
80
60
40
1945
1955
1965
1975
1985
1995
2005
2015
Year
@kjhealy http://socviz.co Data: Congressional Quarterly

Show Ponies

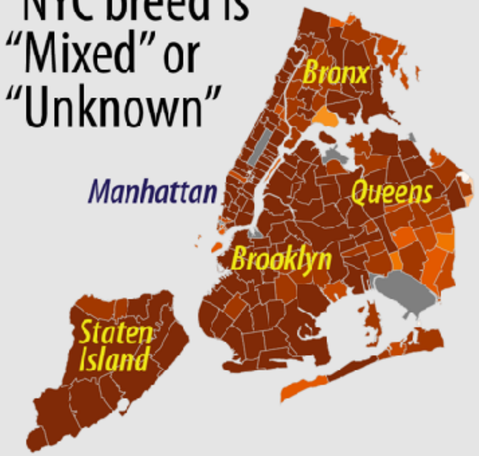

Yates
Gonzalez
Gonzalez
Brown
Brown
Hall
Hall
Hall
Dingell
Meek
Lantos
Dingell
Lantos
Dingell
Lantos
Dingell
Lantos
Dingell
Dingell
Dingell
Convers
Slaughter
Convers
Slaughter
Slaughter
Convers
Johnson

Eaton
Welch
Eaton
Welch
Eaton
Hull
Hull
Reed
Hoffman
Hoffman
Reed
Hoffman
Taber
Martin
Martin
Bolton
Bolton
Younger

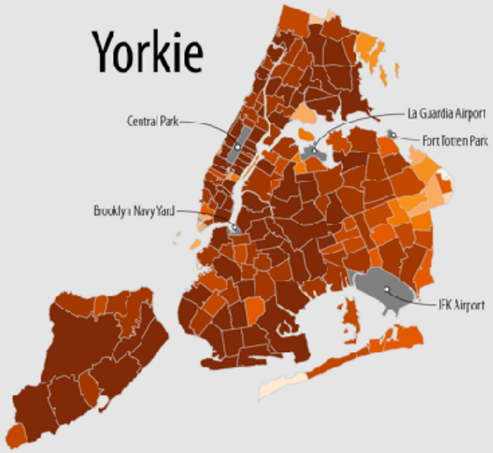
Relative prevalence of breed, from Most to Least

More Common — Less Common

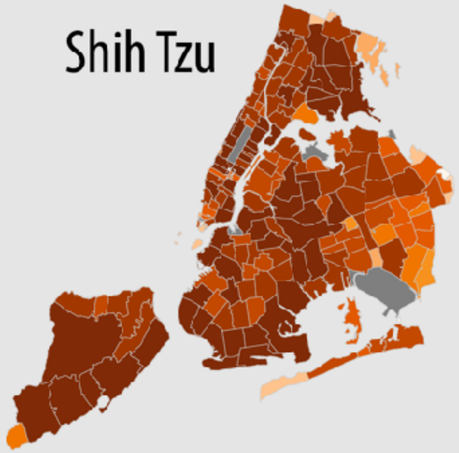
The most common NYC breed is "Mixed" or "Unknown"



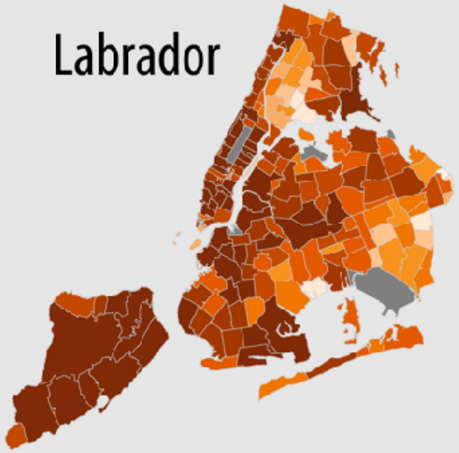
Yorkie



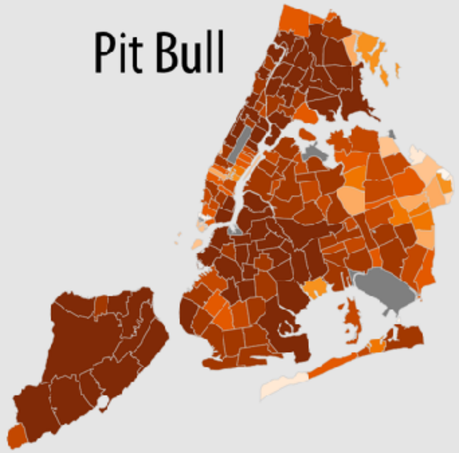
Shih Tzu



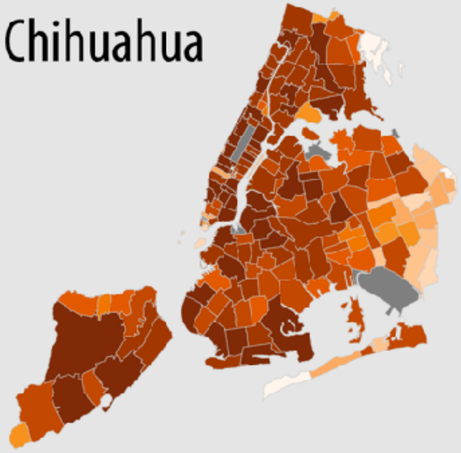
Labrador



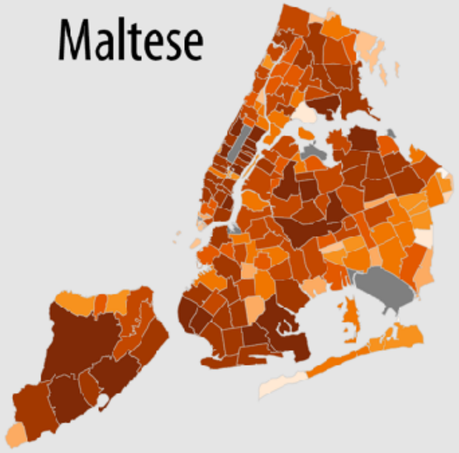
Pit Bull



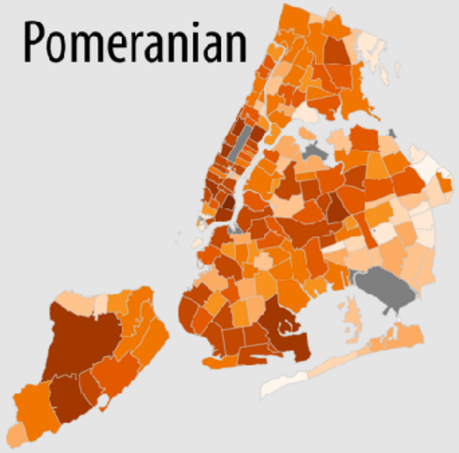
Chihuahua



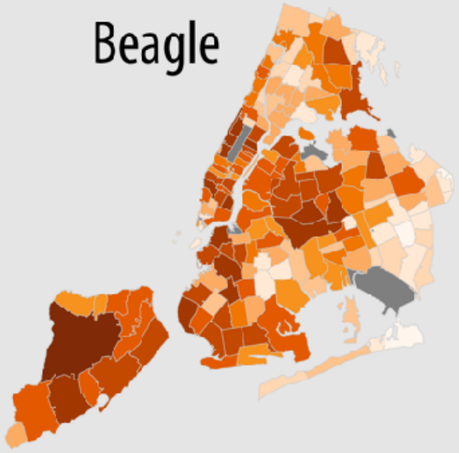
Maltese



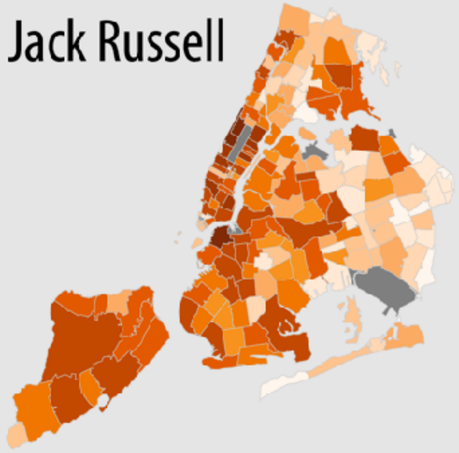
Pomeranian



Beagle

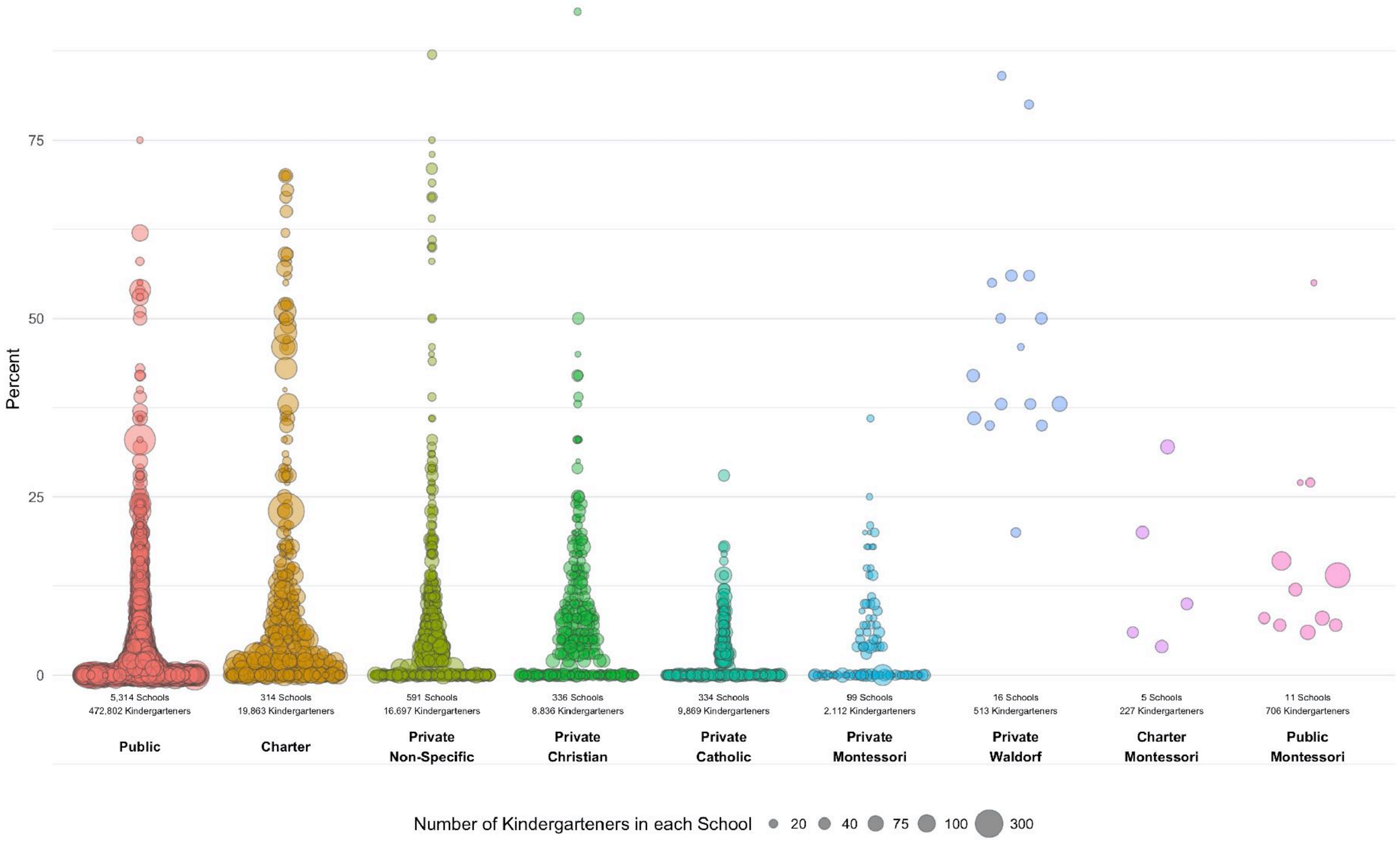


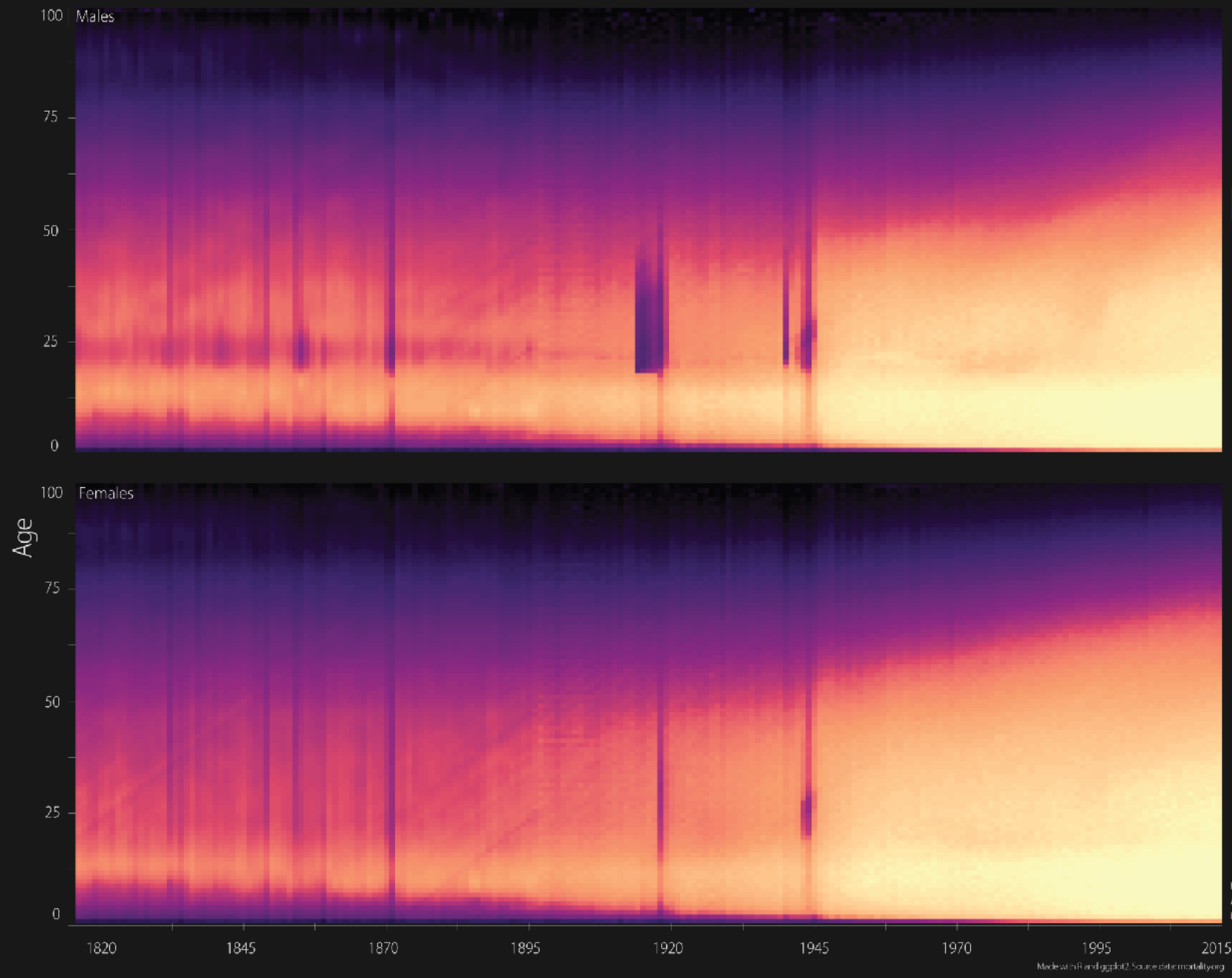
Jack Russell



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

25 50 75 100

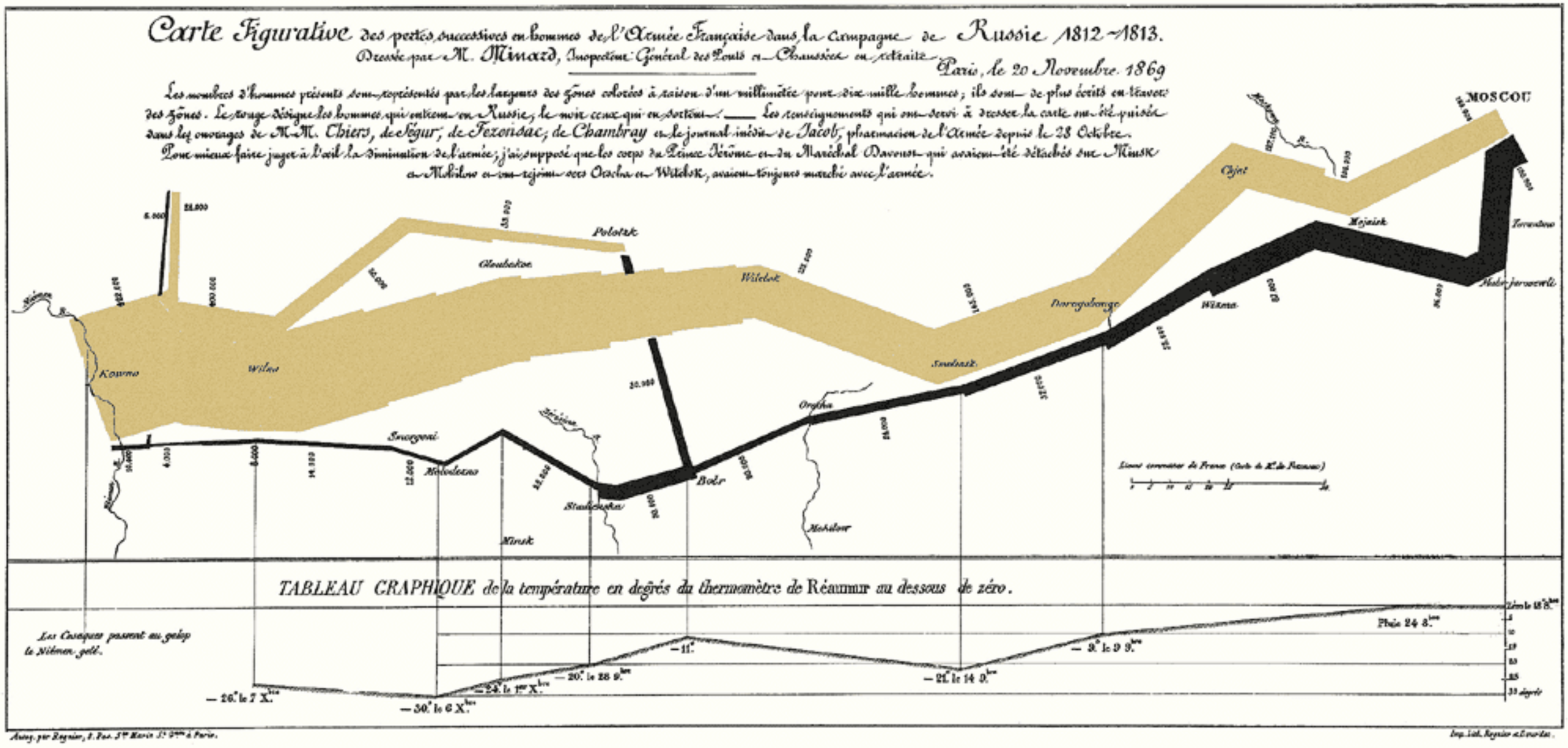
Mortality rates are calculated for each age in each year and binned by percentile. The darker the color at any particular point, the more people of that age die in that year. The lighter the color, the more people of that age survive in that year.

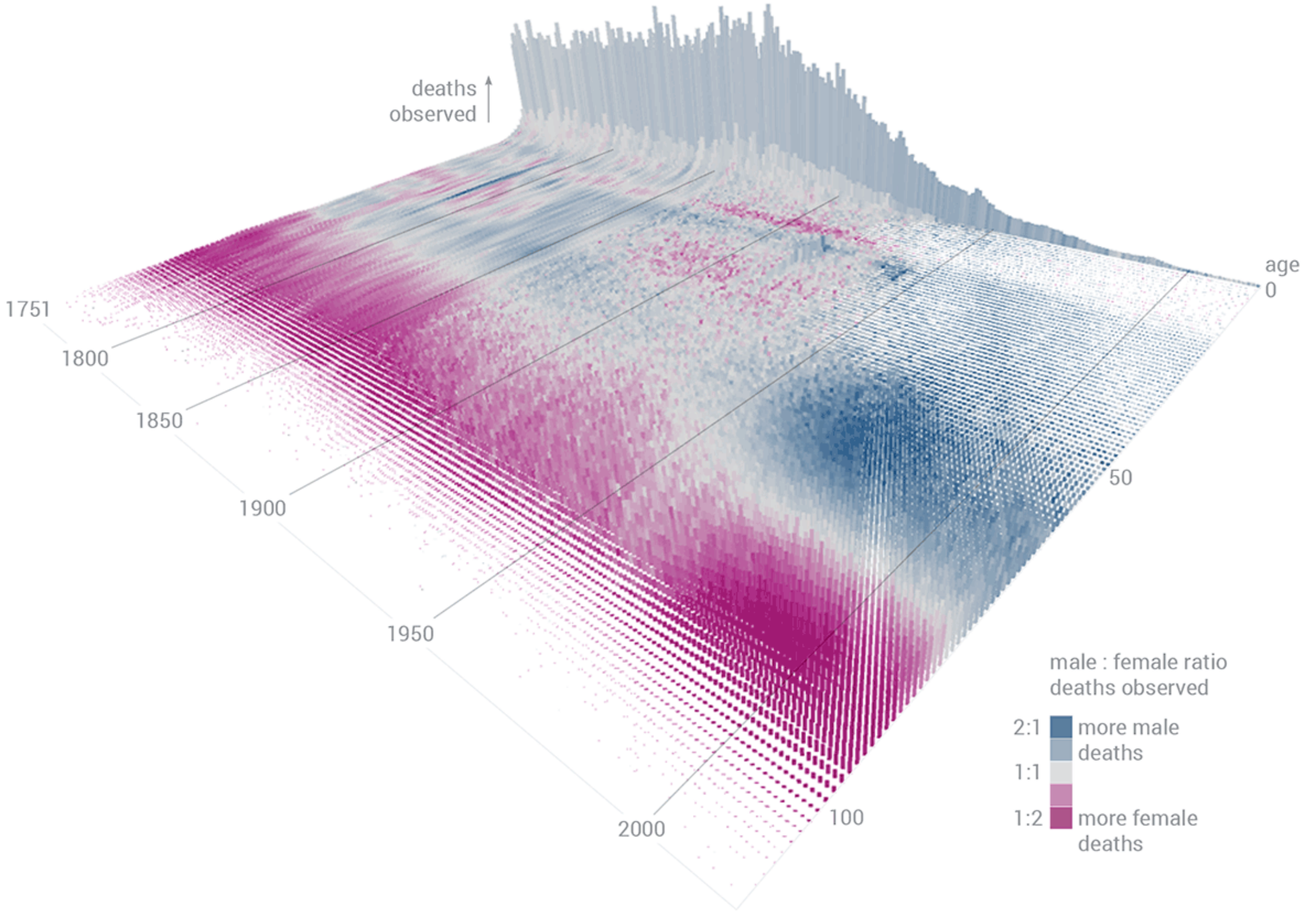
Historical trends are visible, such as the rapid decrease in infant mortality rates after World War II, as well as increased life expectancy overall. Specific events show up as vertical streaks in the graph. The death toll due to wars is evident for Males. Pandemics are also visible, most notably the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

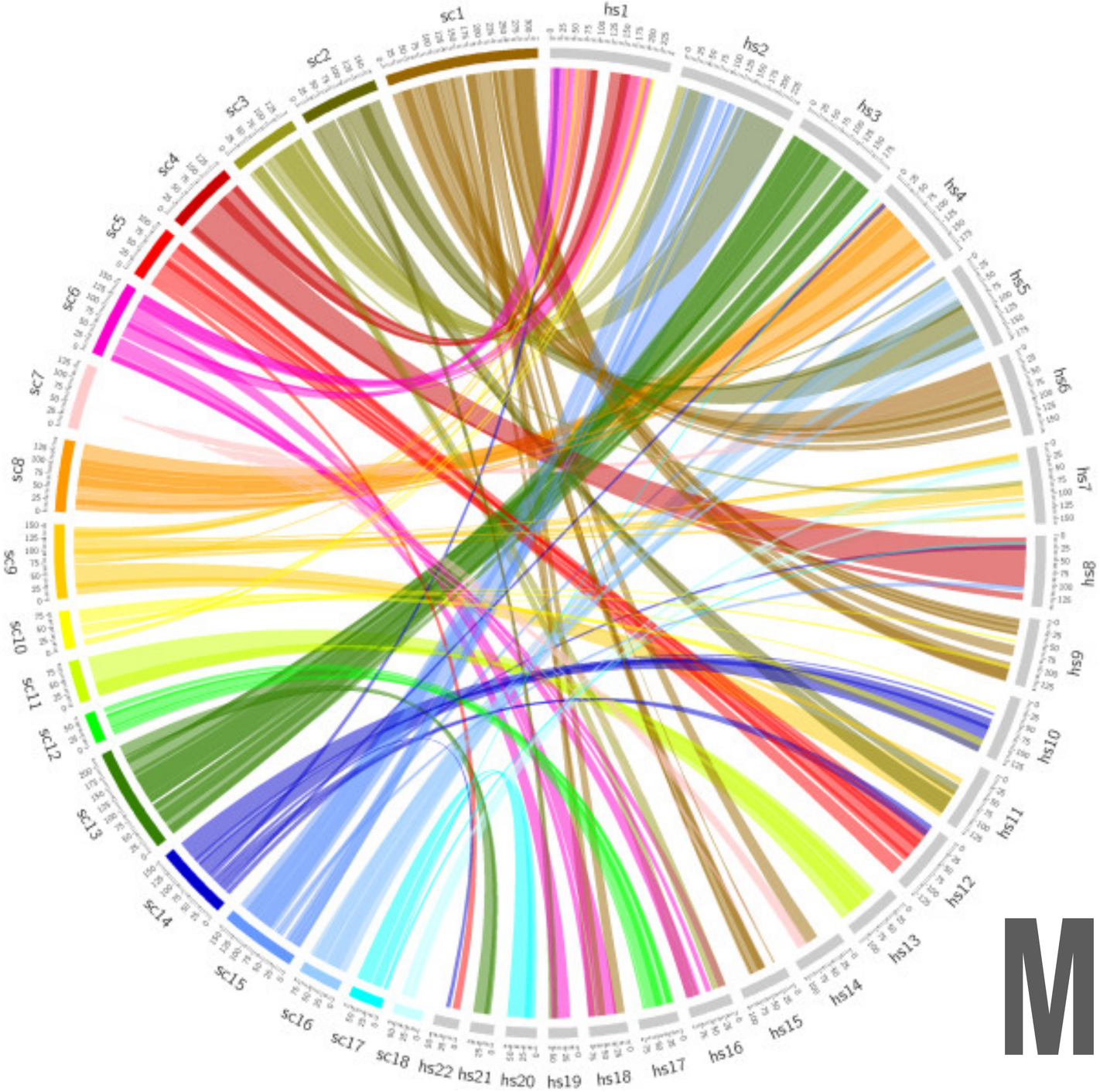
Diagonal streaks in the data are visible in some parts of the graph. These are artifacts due to the estimation of the mortality rate in some years. Single-year-of-age figures prior to 1900 are calculated from five-year age groups, as no single-year data is available in the original mortality tabulations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

Kieran Healy / socviz.co
Data Visualization: A Practical Introduction
is published by Princeton University Press

Unicorns ...


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869
Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orcha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Polotsk
Gloubokoe
Wilebsk
Dorogobouge
Smolensk
Wilna
Mojaisk
Tarantino
Malo-jaroscowetz
Kowno
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bobr
Orscha
Mohilow
Chjat
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26° le 7 X.bre
— 30° le 6 X.bre
— 24° le 1er X.bre
— 20° le 28 9.bre
— 11°
— 21° le 14 9.bre
— 9° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
33 degrés
Imp. Lith. Regnier et Dourdet.


deaths observed
1751
1800
1850
1900
1950
2000
age
0
50
100
male : female ratio deaths observed
2:1
more male deaths
1:1
1:2
more female deaths


sc1
sc2
sc3
sc4
sc5
sc6
sc7
sc8
sc9
sc10
sc11
sc12
sc13
sc14
sc15
sc16
sc17
sc18
hs22
hs21
hs20
hs19
hs18
hs17
hs16
hs15
hs14
hs13
hs12
hs11
hs10
hs9
hs8
hs7
hs6
hs5
hs4
hs3
hs2
hs1

... or Monsters