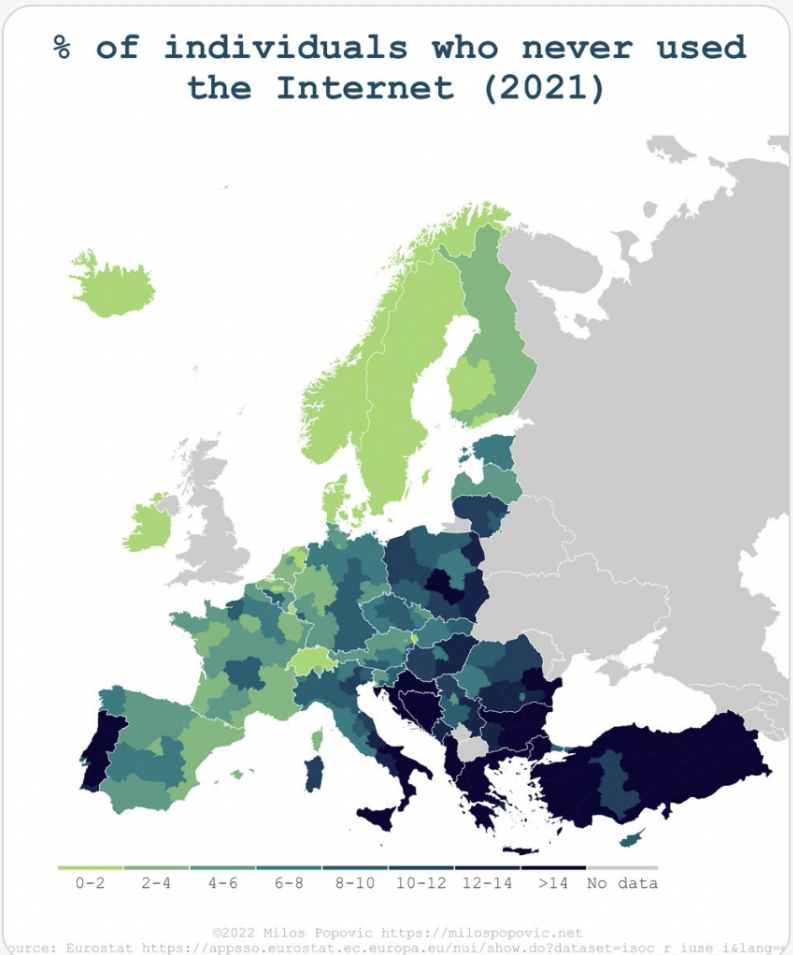
Data communication and visualization

Jae Yeon Kim

Milos Popovic @milos_agathon · 22h

My new **map** shows % of people who never used the **Internet** in **Europe** 😱😱😱

#internet #www #digital #RStats #DataScience #dataviz **#maps**

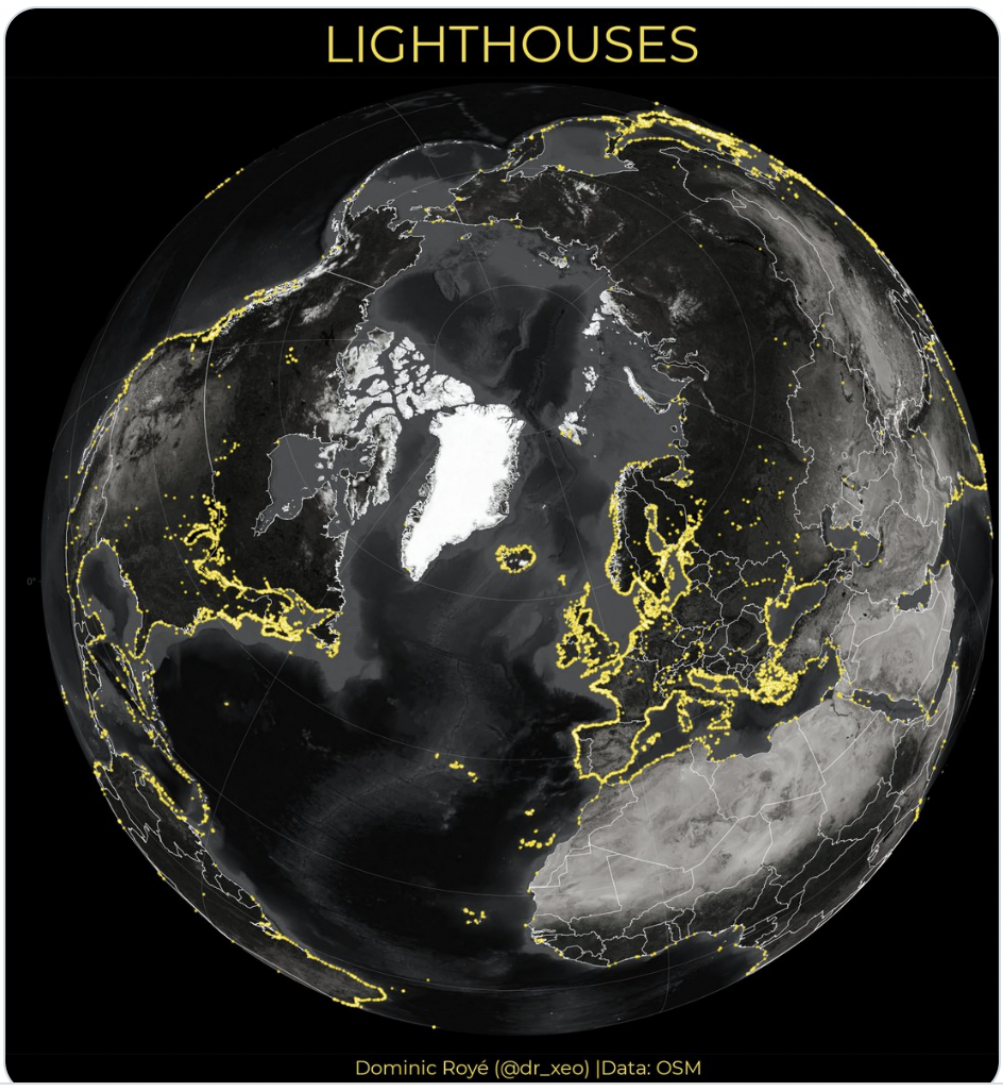


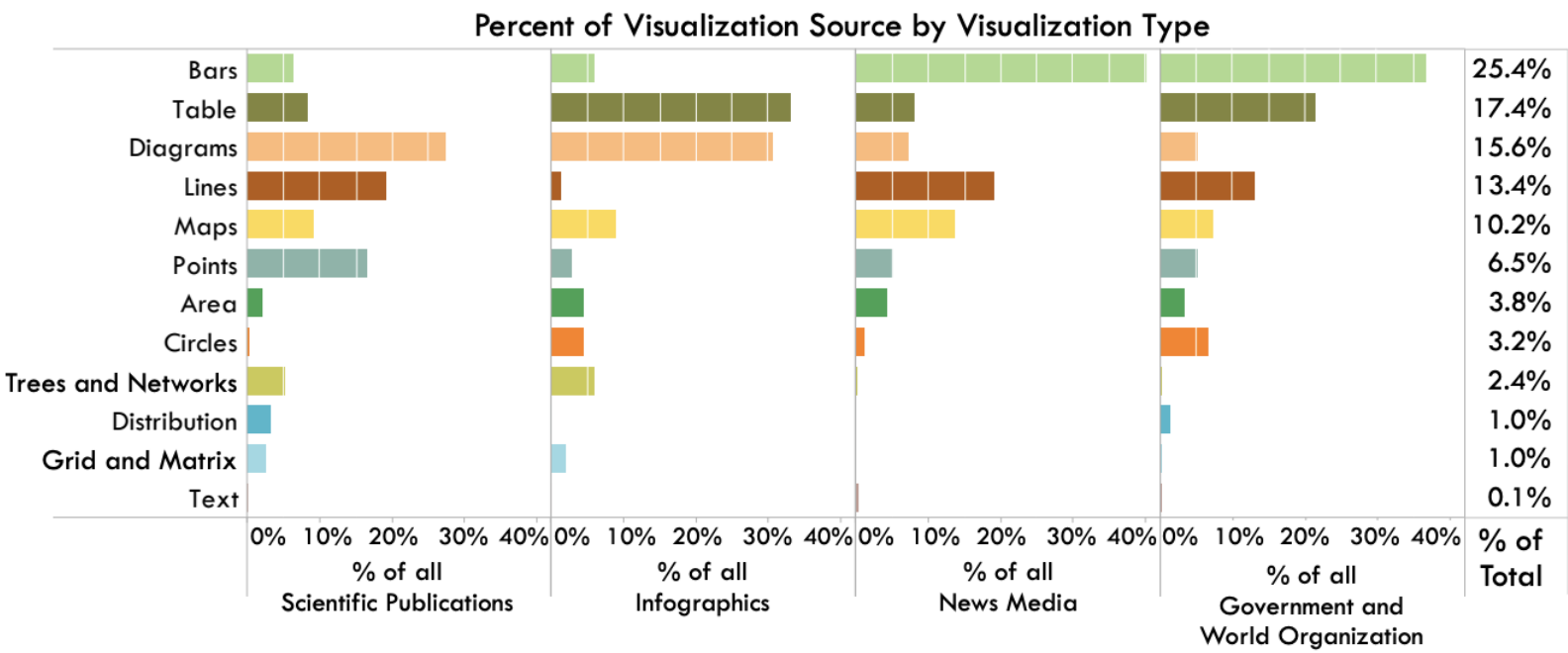
9 48 140

Data science · See more

Dr. Dominic Royé @dr_xeo · 15h

World's Lighthouse. #rstats #dataviz




Percent of Visualization Source by Visualization Type
Bars
Table
Diagrams
Lines
Maps
Points
Area
Circles
Trees and Networks
Distribution
Grid and Matrix
Text
25.4%
17.4%
15.6%
13.4%
10.2%
6.5%
3.8%
3.2%
2.4%
1.0%
1.0%
0.1%
0% 10% 20% 30% 40%
% of all Scientific Publications
% of all Infographics
% of all News Media
% of all Government and World Organization
% of Total

Data visualization = turing (i.e., statistical) information into visual contexts (e.g., plots)

Data visualization = means not ends (= data communication)

Objectives:

(1) Learn (lectures; freameworks)

and

(2) practice (workshops; workflows)

communicating data via writing (#1)

and programming (#2)

Cognitive science +
R programming +
data writing =
this course

Why not tables?

aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Psychological Science in the
Public Interest
2021, Vol. 22(3) 110–161

DOI: 10.1177/15291006211051956
www.psychologicalscience.org/PSPI
SAGE

The Science of Visual Data Communication: What Works

Steven L. Franconeri[1], Lace M. Padilla[2], Priti Shah[3], Jeffrey M. Zacks[4], and Jessica Hullman[5]

[1]Department of Psychology, Northwestern University; [2]Department of Cognitive and Information Sciences, University of California, Merced; [3]Department of Psychology, University of Michigan; [4]Department of Psychological & Brain Sciences, Washington University in St. Louis; and [5]Department of Computer Science, Northwestern University

Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;

(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;

(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

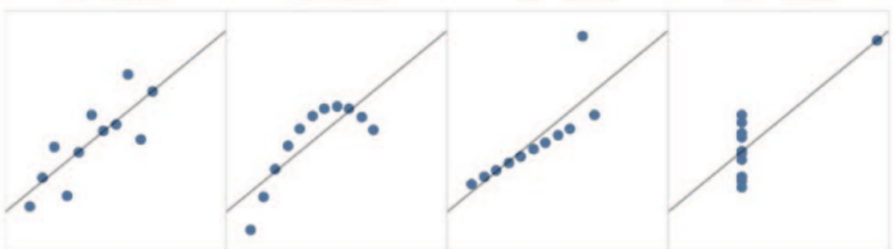
Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

A few simple types of statistical analysis are now considered.

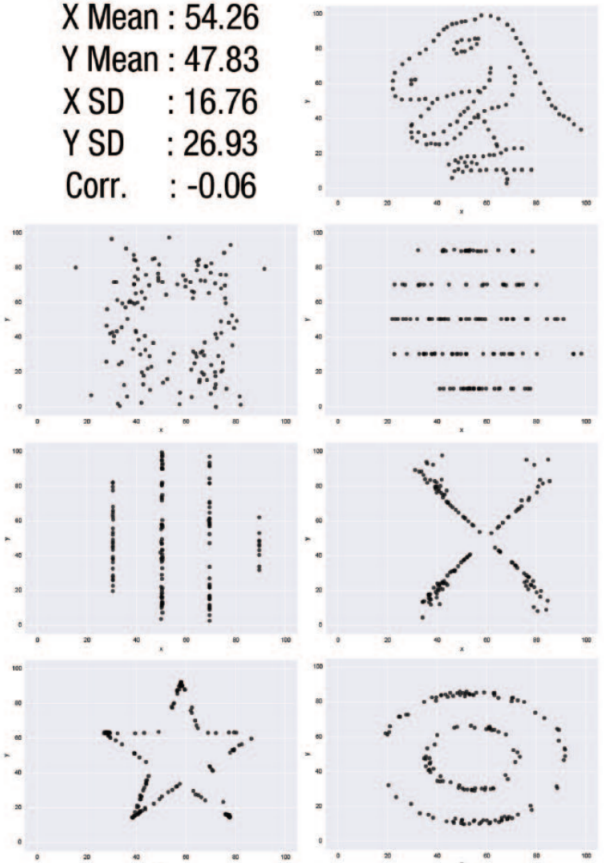
2. *Regression analysis—the simplest case*

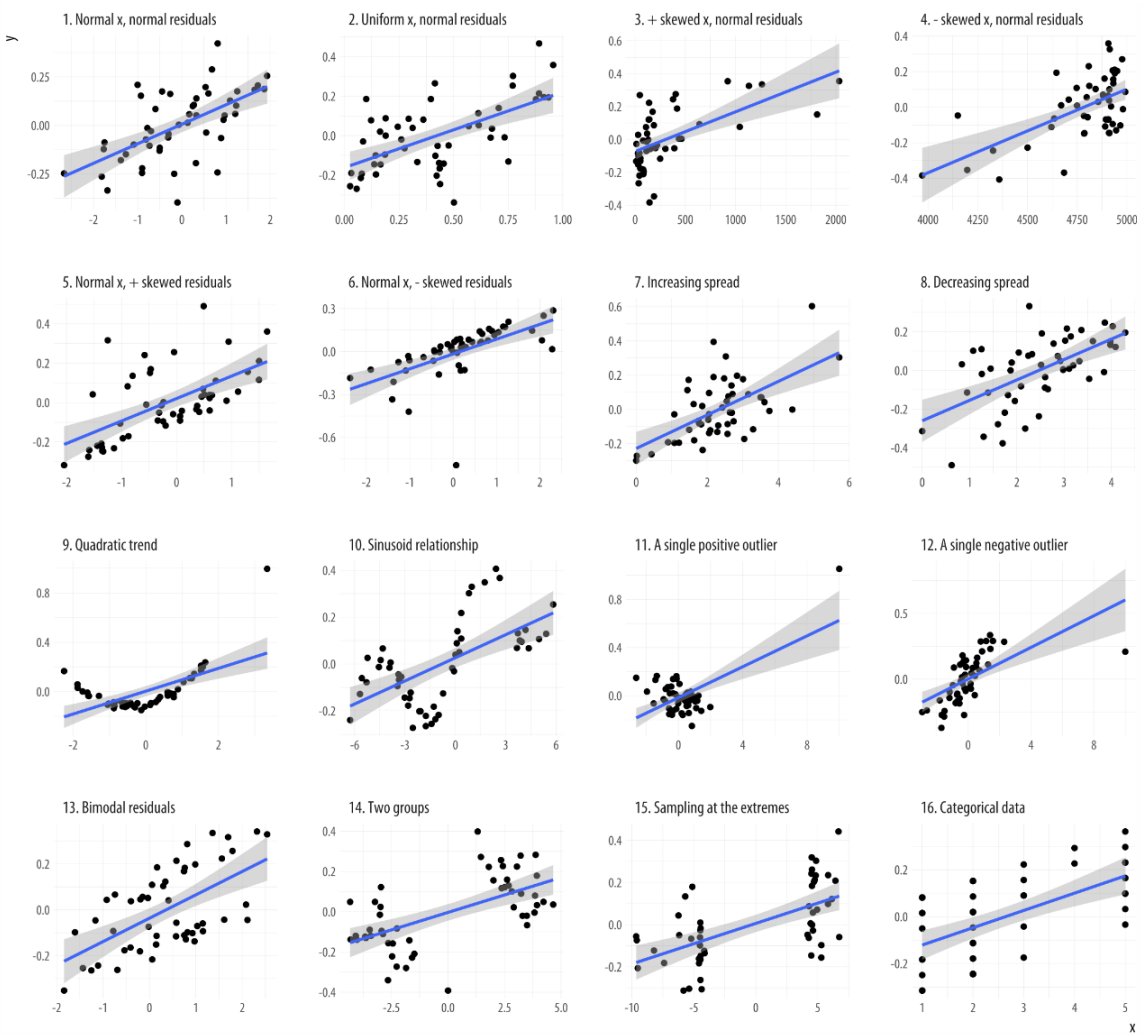
Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor)

x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Stdev	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
	R	0.82	**R**	0.82	**R**	0.82	**R**	0.82



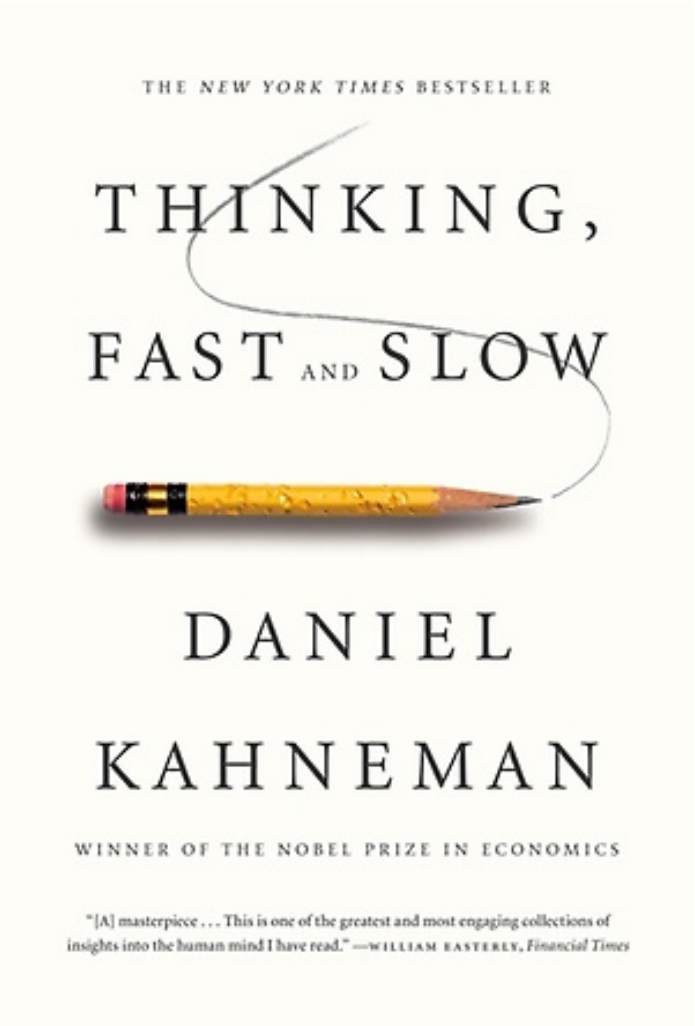
X Mean : 54.26
Y Mean : 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

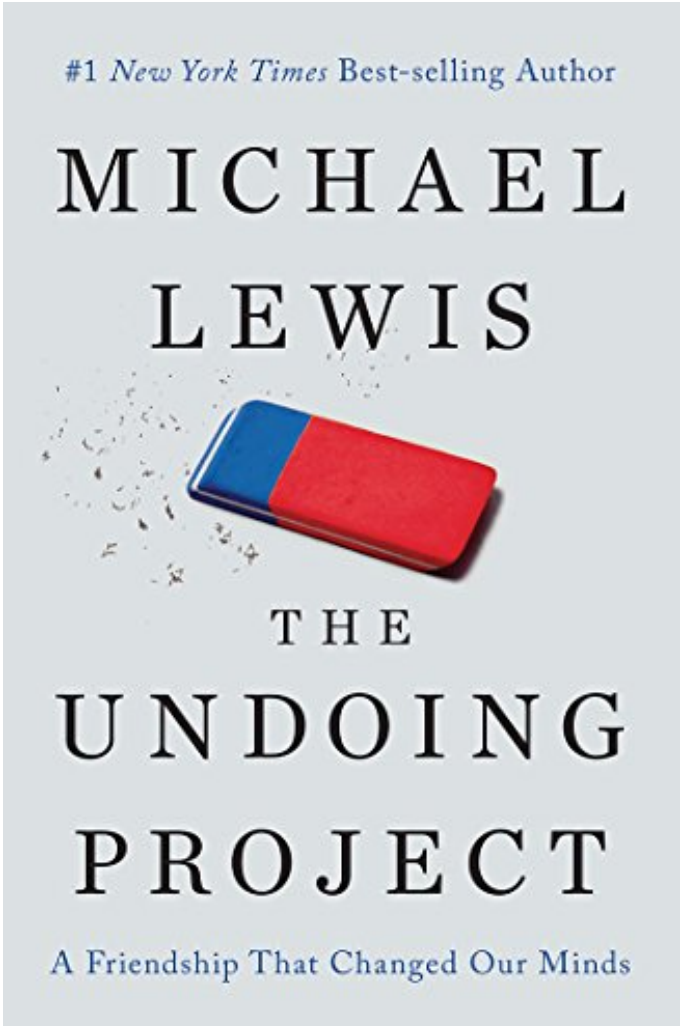



1. Normal x, normal residuals
2. Uniform x, normal residuals
3. + skewed x, normal residuals
4. - skewed x, normal residuals
5. Normal x, + skewed residuals
6. Normal x, - skewed residuals
7. Increasing spread
8. Decreasing spread
9. Quadratic trend
10. Sinusoid relationship
11. A single positive outlier
12. A single negative outlier
13. Bimodal residuals
14. Two groups
15. Sampling at the extremes
16. Categorical data
y
x

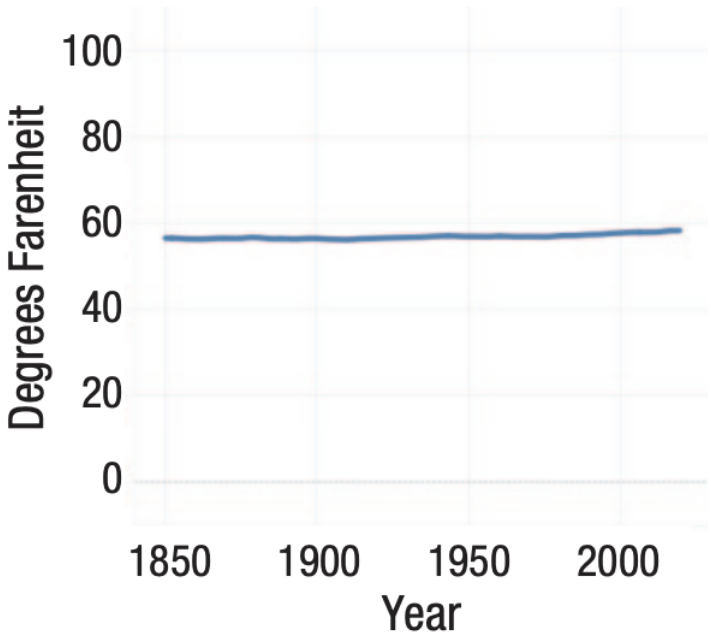
A picture is worth a thousand words???

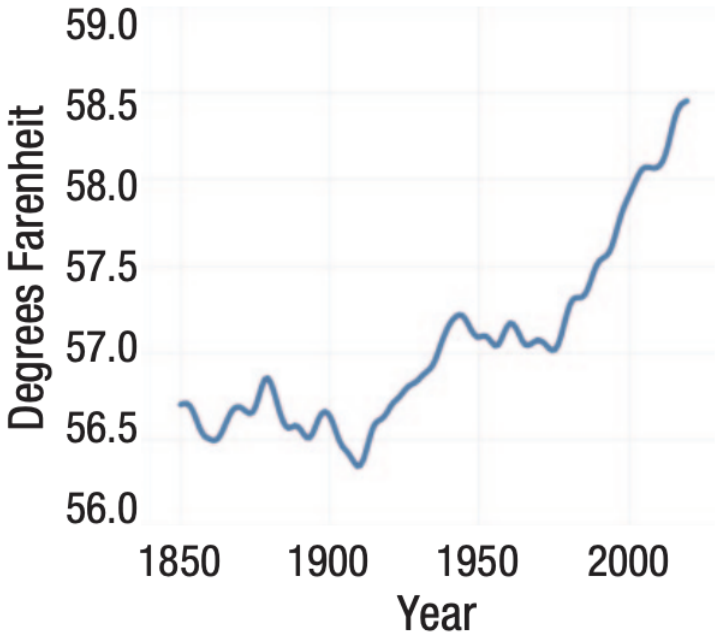
It DEPENDS


THE NEW YORK TIMES BESTSELLER
THINKING, FAST AND SLOW
DANIEL KAHNEMAN
WINNER OF THE NOBEL PRIZE IN ECONOMICS
"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, Financial Times


#1 New York Times Best-selling Author
MICHAEL LEWIS
THE UNDOING PROJECT
A Friendship That Changed Our Minds

Stretching the *y*-axis scale of the left graph drastically increases the slope of the perceived trend at right, which feels dishonest.





Here, taking a small visual increase (left) and stretching it (right) is the "honest" way to show climate-change data.

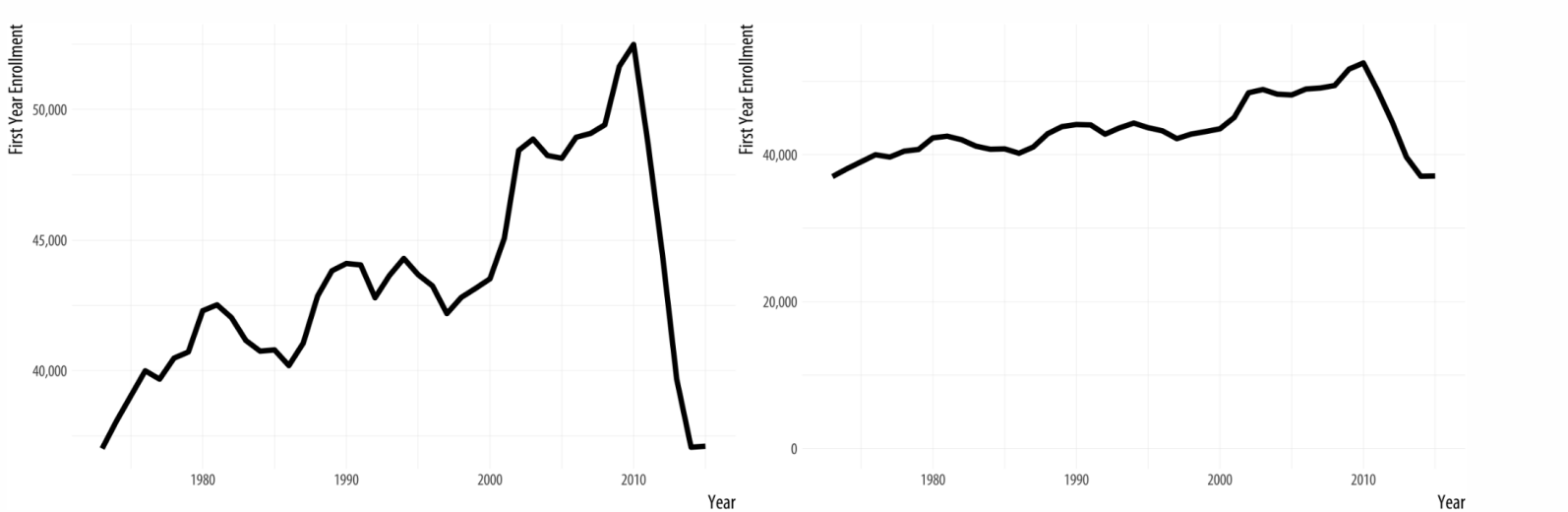
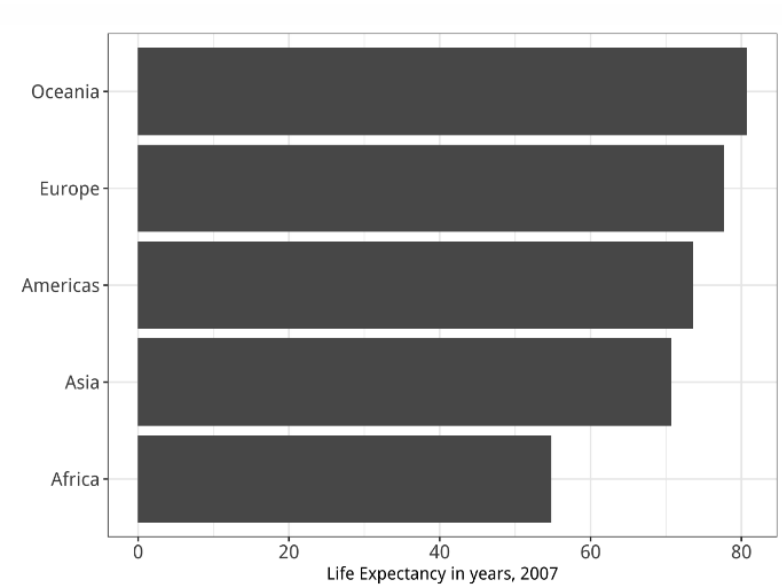
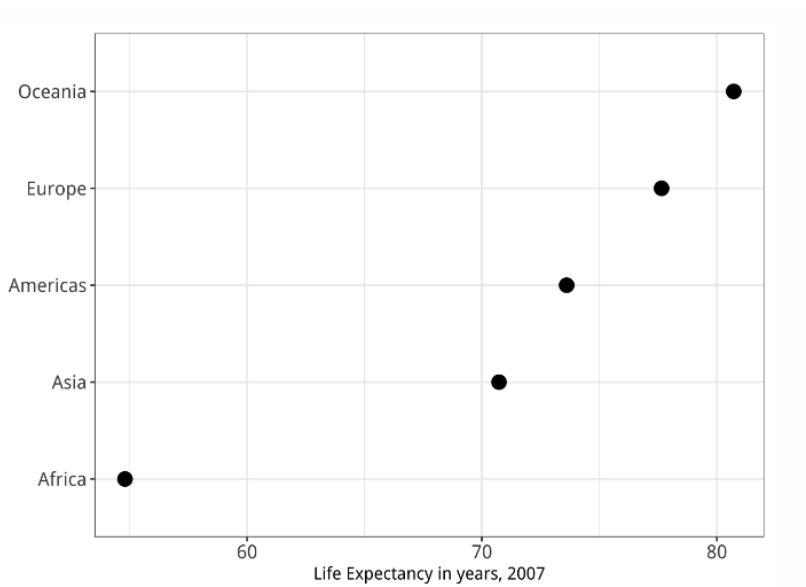


Figure 1.27: Two views of the rapid decline in law school enrollments in the mid-2010s.


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

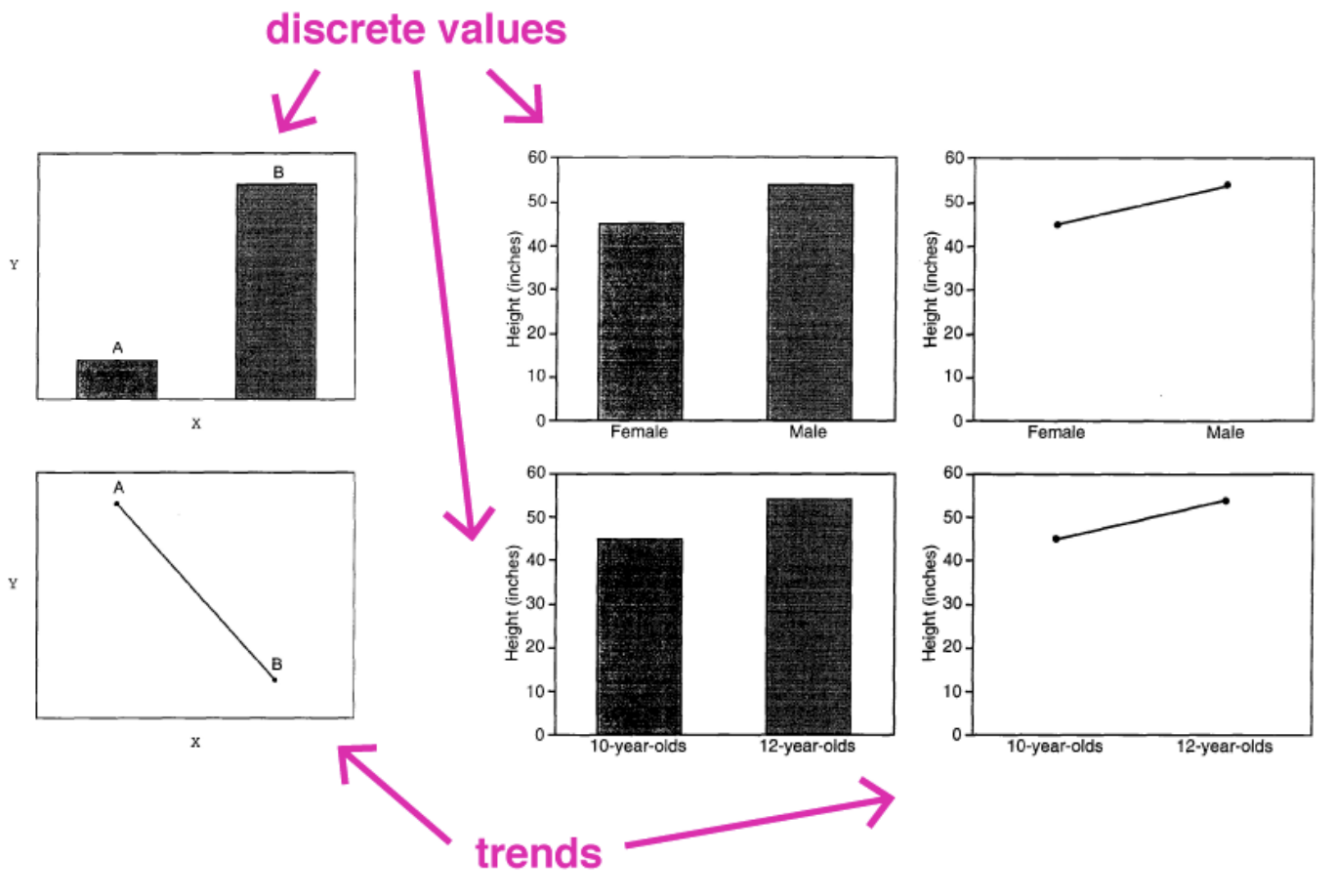

Oceania
Europe
Americas
Asia
Africa
60
70
80
Life Expectancy in years, 2007

Memory & Cognition
1999, 27 (6), 1073-1079

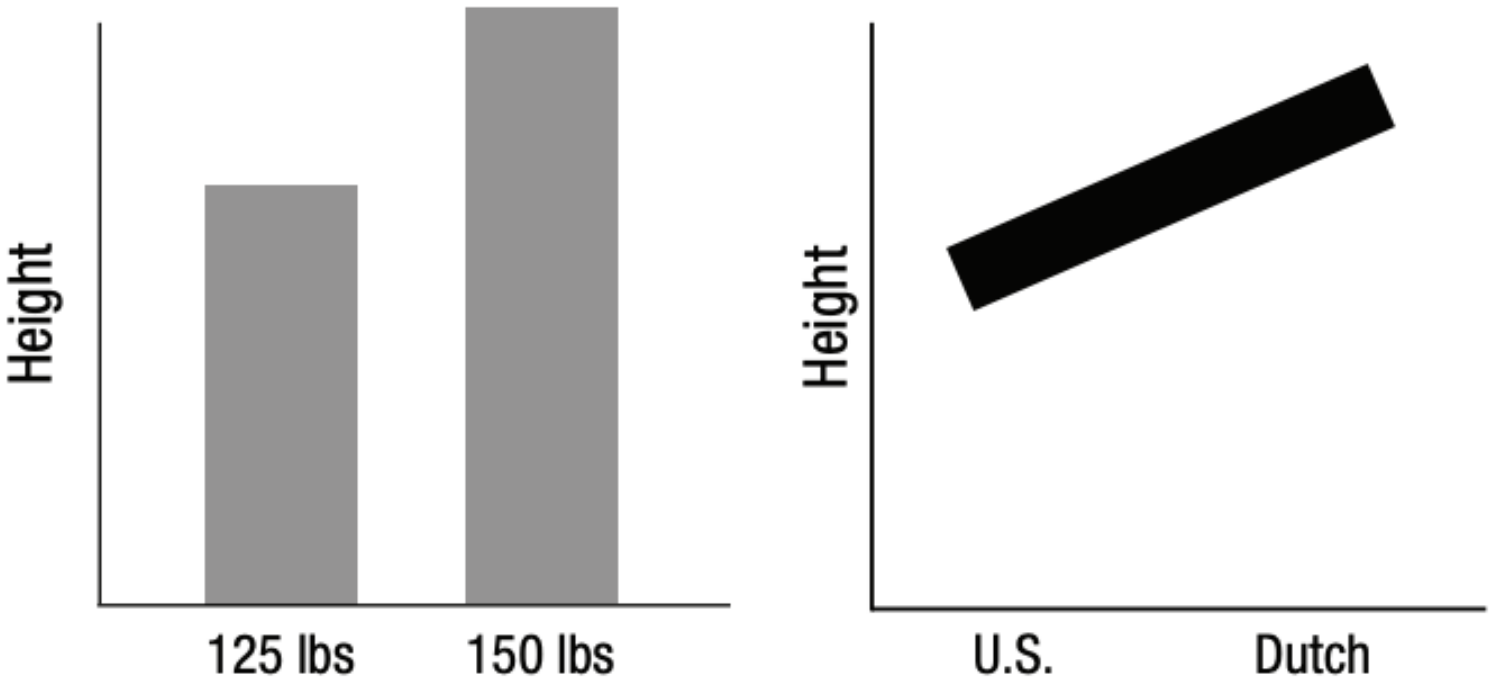
Bars and lines: A study of graphic communication

JEFF ZACKS and BARBARA TVERSKY
Stanford University, Stanford, California

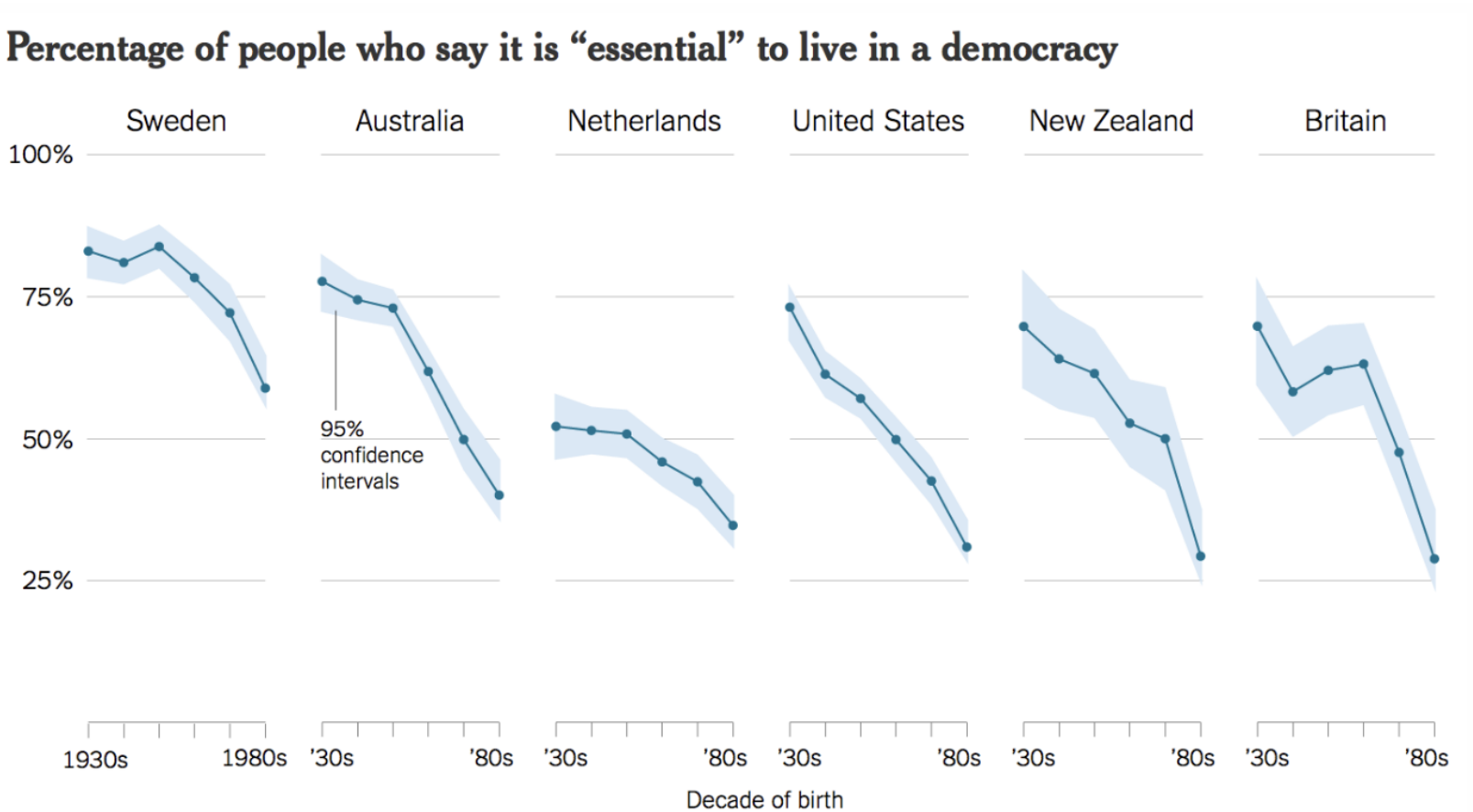
Interpretations of graphs seem to be rooted in principles of cognitive naturalness and information processing rather than arbitrary correspondences. These predict that people should more readily associate bars with discrete comparisons between data points because bars are discrete entities and facilitate point estimates. They should more readily associate lines with trends because lines connect discrete entities and directly represent slope. The predictions were supported in three experiments—two examining comprehension and one production. The correspondence does not seem to depend on explicit knowledge of rules. Instead, it may reflect the influence of the communicative situation as well as the perceptual properties of graphs.

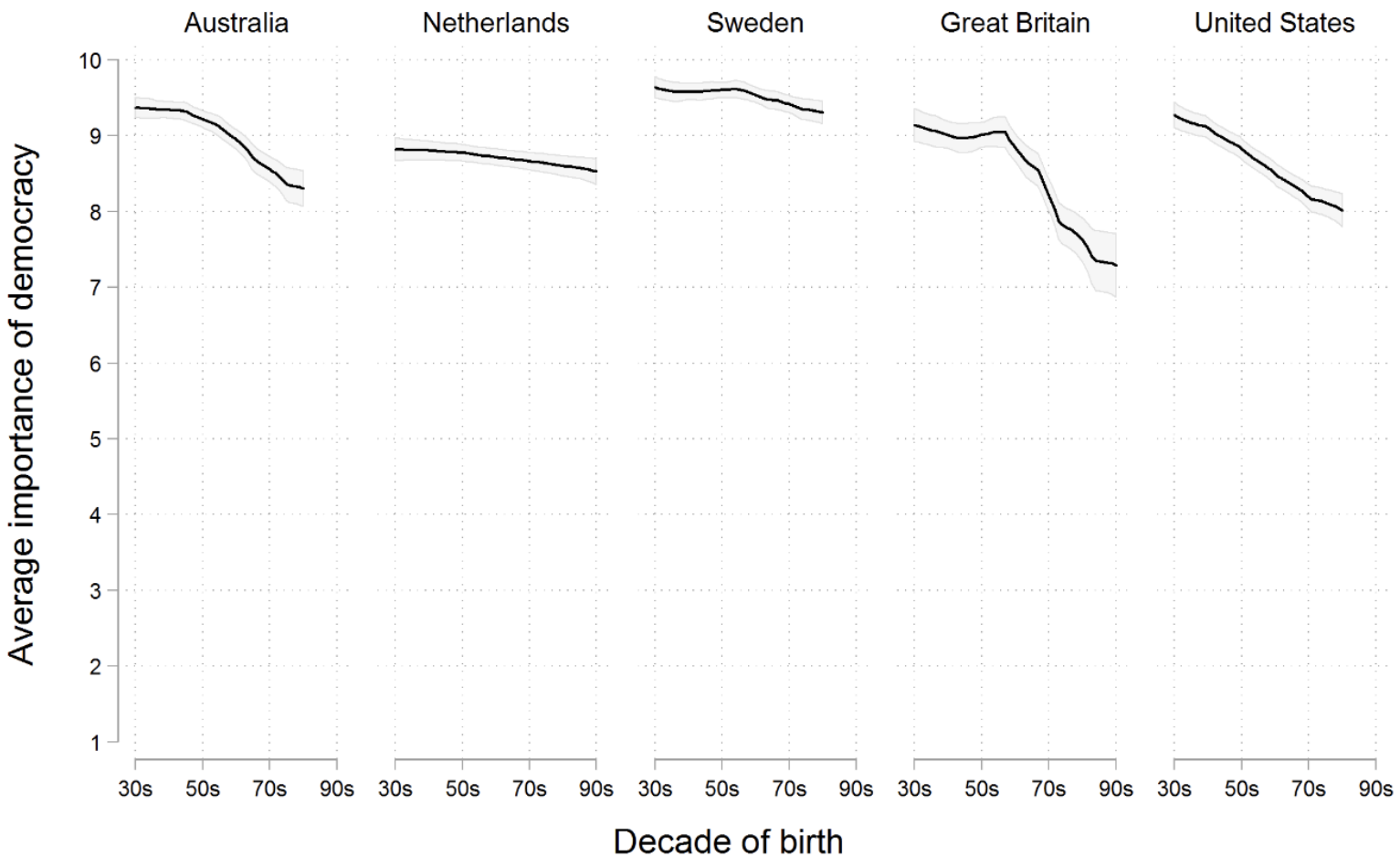


Zacks and Tversky found that participants described contrasts between the x-axis variables more when presented with bar charts, and relationships between the x-axis variables more with line charts.



The choice of graph can substantially influence conclusions made from the same data


Percentage of people who say it is “essential” to live in a democracy
Sweden
Australia
Netherlands
United States
New Zealand
Britain
100%
75%
50%
25%
95% confidence intervals
1930s
1980s
'30s
'80s
Decade of birth



Graph by Erik Voeten, based on WVS 5

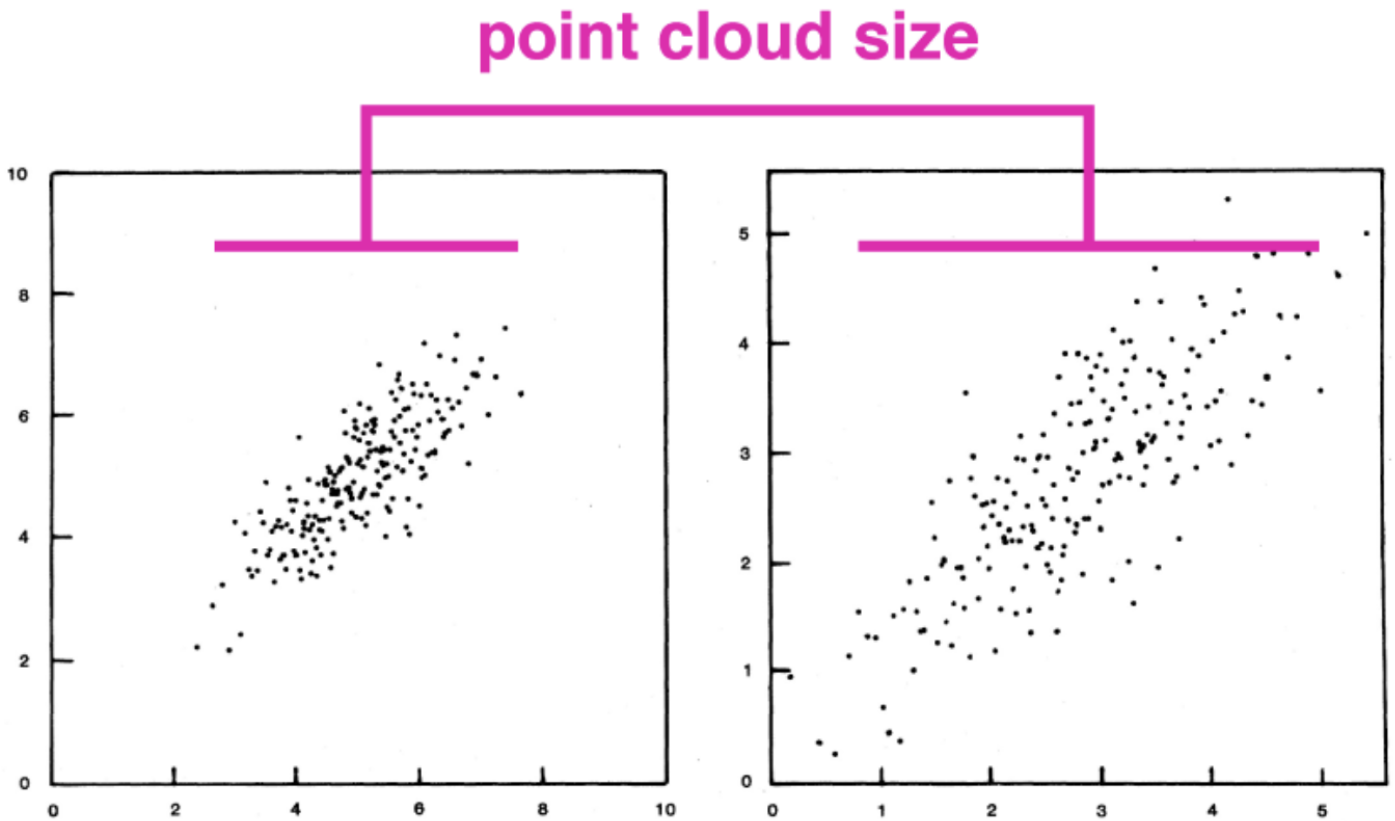
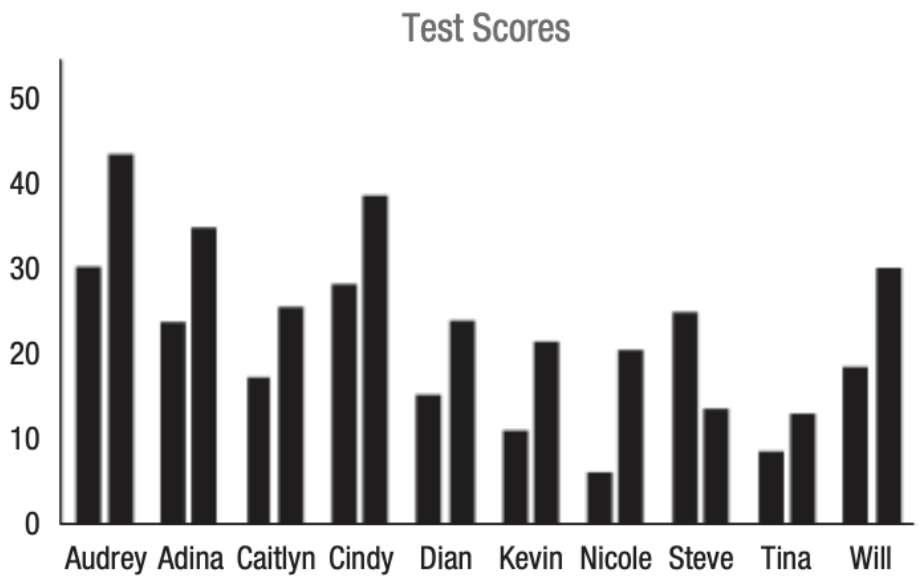


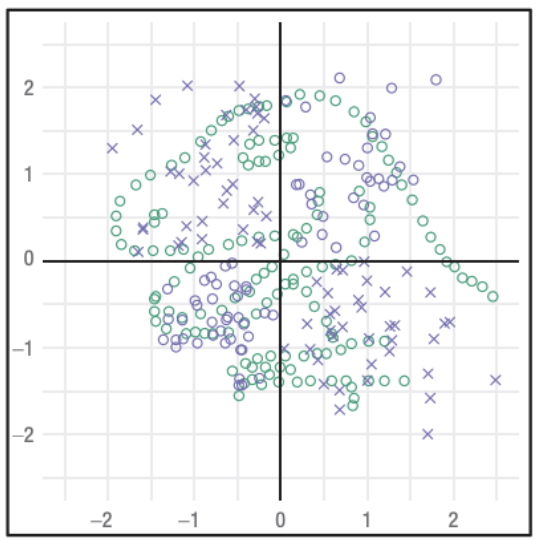
Fig. 1. Reductions of two scatterplots used in the three types of experiments. The left panel is point-cloud size 2 and the right panel is point-cloud size 4. In both panels $w(r) = .4$ and $r = .8$.

Cleveland and coauthors found that altering point cloud size alone could lead to inaccuracies in evaluating correlation in scatterplots.

In-class discussion

1. Form a group of 4-5.
2. Discuss when tables are more effective and when figures are more effective in communicating data.


Test Scores
50
40
30
20
10
0
Audrey
Adina
Caitlyn
Cindy
Dian
Kevin
Nicole
Steve
Tina
Will


2
1
0
−1
−2
−2
−1
0
1
2

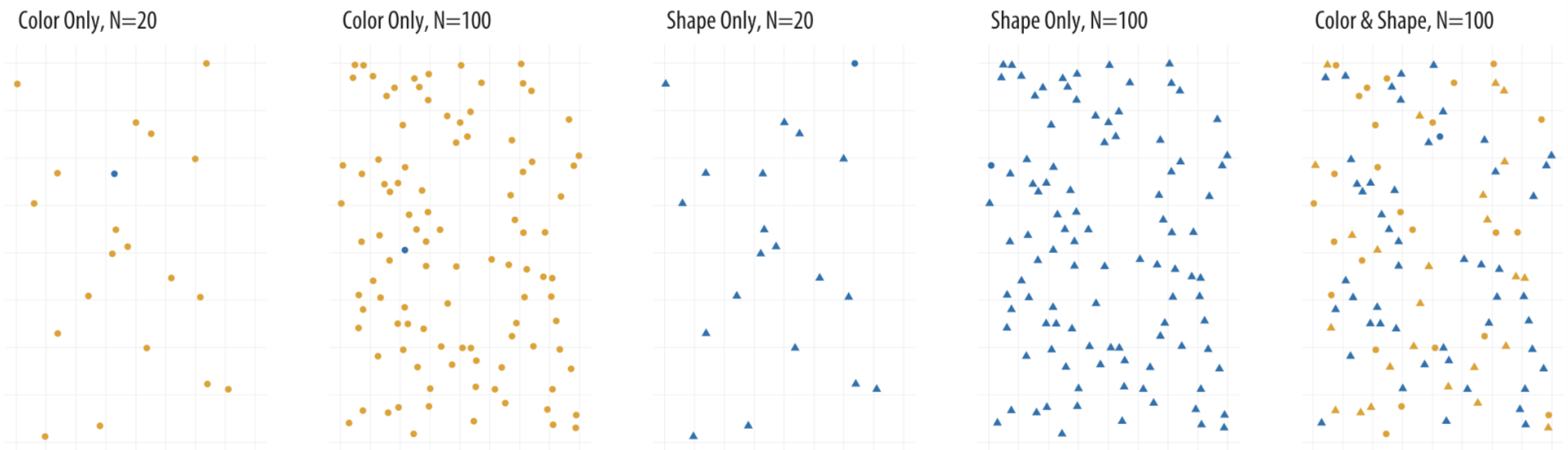
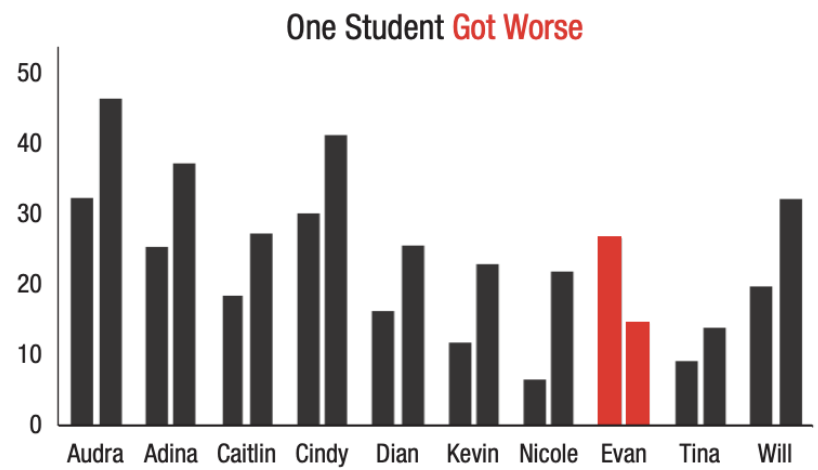
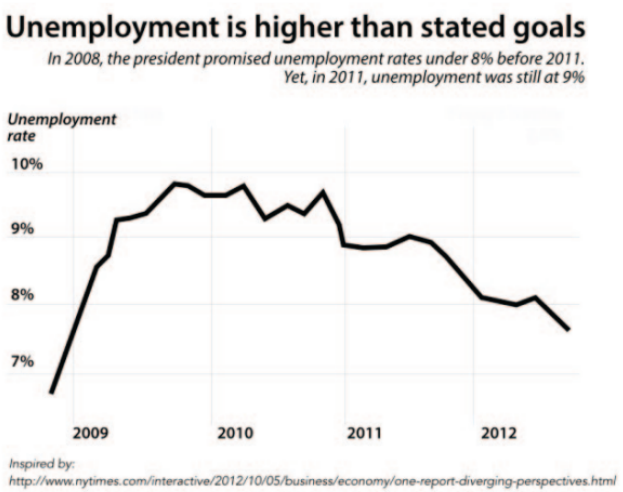
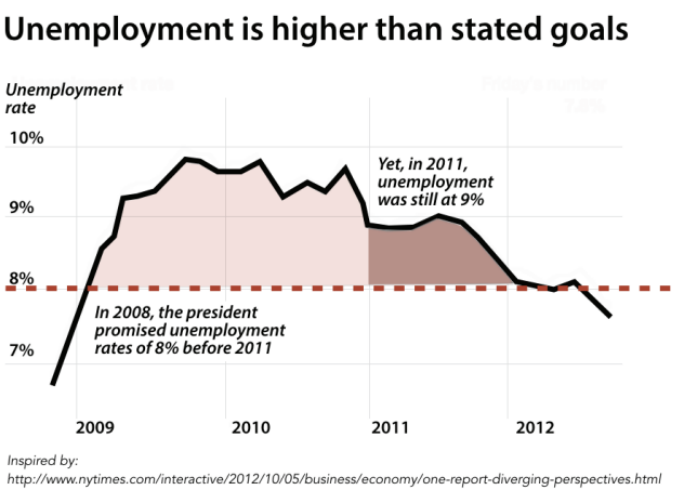


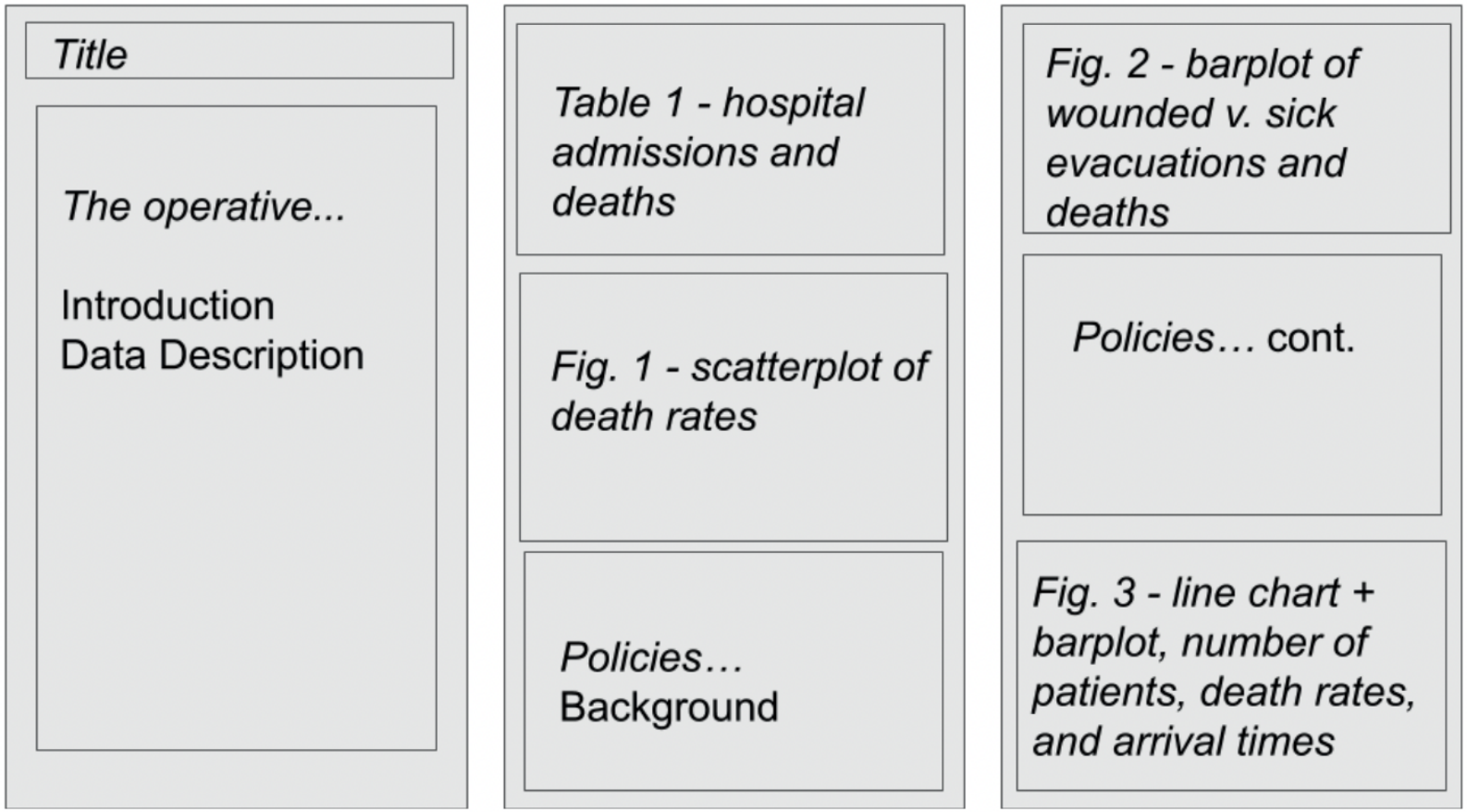
Figure 1.18: Searching for the blue circle becomes progressively harder.


One Student Got Worse
50
40
30
20
10
0
Audra
Adina
Caitlin
Cindy
Dian
Kevin
Nicole
Evan
Tina
Will

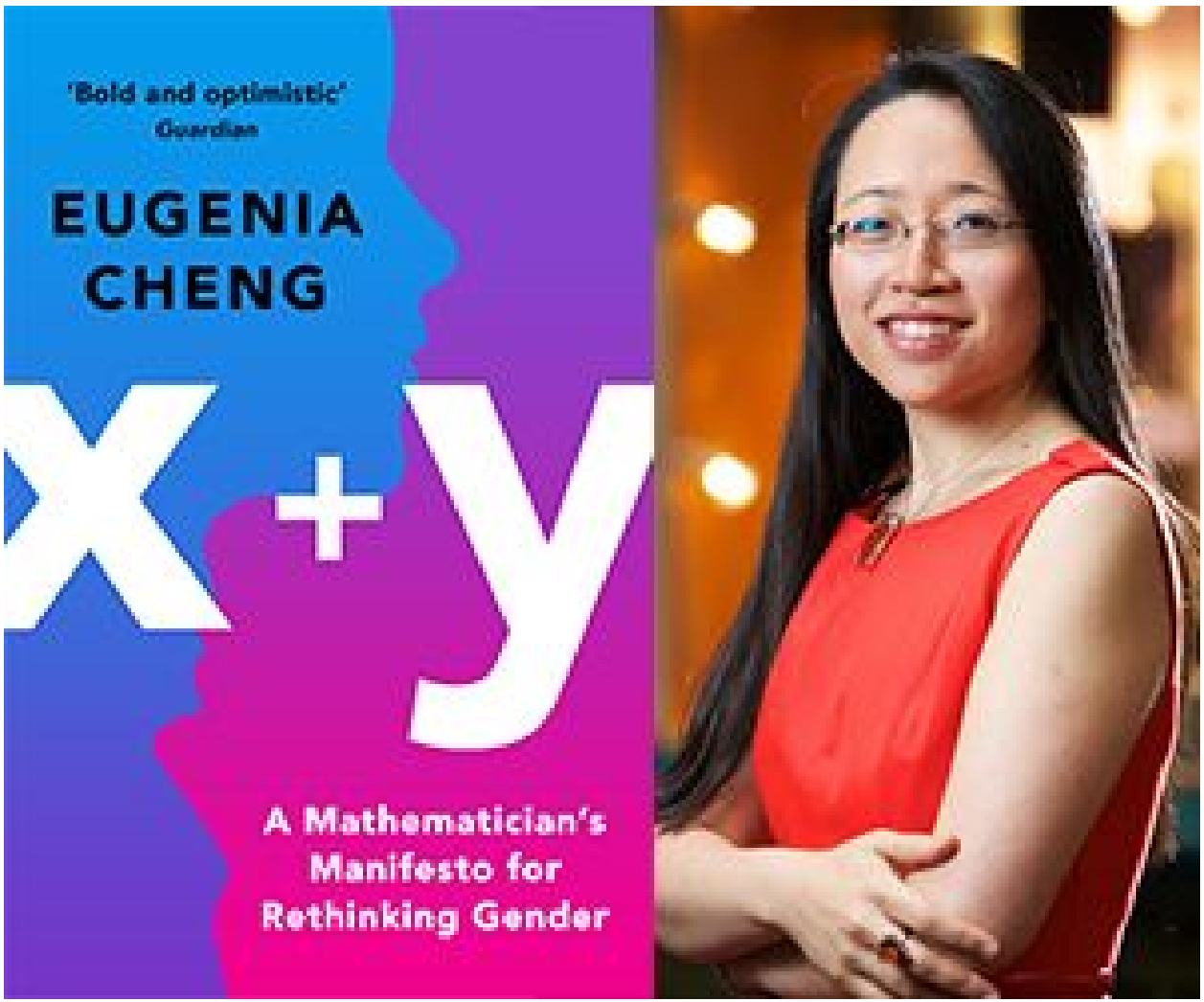

Unemployment is higher than stated goals
In 2008, the president promised unemployment rates under 8% before 2011.
Yet, in 2011, unemployment was still at 9%
Unemployment rate
10%
9%
8%
7%
2009
2010
2011
2012
Inspired by:
http://www.nytimes.com/interactive/2012/10/05/business/economy/one-report-diverging-perspectives.html


Unemployment is higher than stated goals
Unemployment rate
10%
9%
8%
7%
Yet, in 2011, unemployment was still at 9%
In 2008, the president promised unemployment rates of 8% before 2011
2009
2010
2011
2012
Inspired by:
http://www.nytimes.com/interactive/2012/10/05/business/economy/one-report-diverging-perspectives.html

Three-step fraemwork of data communication (Nolan and Stoudt 2021)



1. Map the organization
2. Identify the statistical elements
3. Examine the eargument


'Bold and optimistic'
Guardian
EUGENIA
CHENG
x + y
A Mathematician's
Manifesto for
Rethinking Gender

In-class practice

1. Randomly form a group of 4-5 people

2. In Table 1, M is male, F is female. The cells indicate their incomes.

3. (1) Calculate some statistics, (2) draw a plot based on the statistical information, and (3) explain whether the plot helps make a good decision about whether there's a gender gap in this population.

ID	**M**	**F**
1	5000	100000
2	4000	100
3	6000	300
4	4000	100
5	1140	200

Table 1

Syllabus review

KSS survey (every Weds)

K: Keep

S: Start

S: Stop

Let's have fun!