

Lesson 2

Basic concepts

Andrea Saltelli

Introduction to Statistics
March 30, 2023 UPF-BSM



Basic concepts – source, population, sample. Definition of data types – qualitative vs quantitative, discrete vs continuous, ordinal vs nominal.

What is a mean?

- We consider means as everyday objects
- Yet means were – in a sense – discovered
- There is more to a mean than meets the eye

Mean cost of a flat per square metre

Mean distance between the earth and the sun

Mean OECD–PISA score for a given country

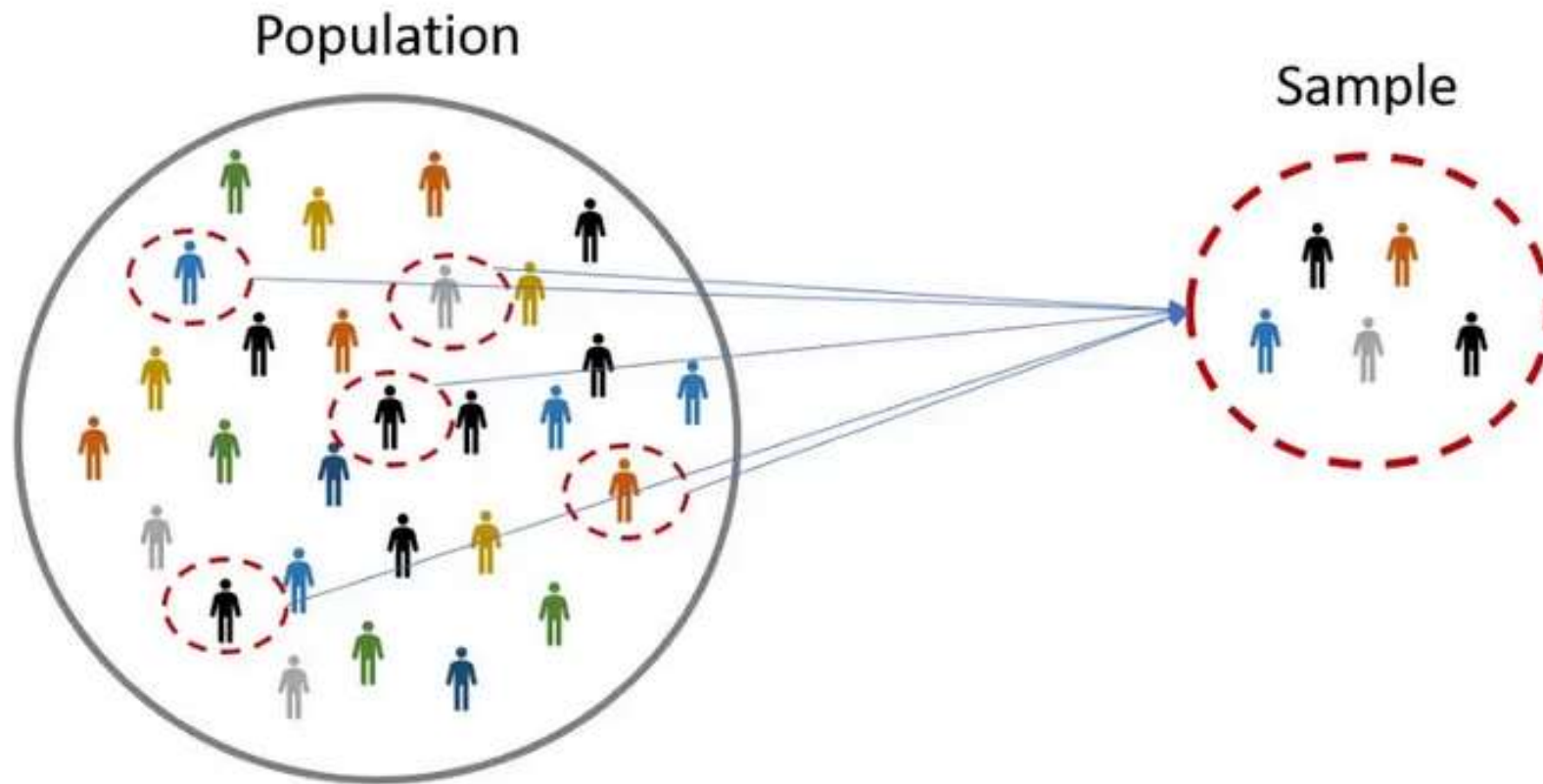
Mean consumption per household of a given
commodity

Gross domestic product, consumer price indices,...

To distinguish these objects we need to know more about where these are coming from. Key questions are

What population is the mean referring to?

Is the mean computed on the population or is it an estimate based on a sample?



Source: <https://medium.com/analytics-vidhya/statistics-population-and-sample-993a488572ac>

Mean for population data: $\mu = \frac{\sum X}{N}$

Mean for sample data: $\bar{X} = \frac{\sum X}{n}$

Mean cost of a flat per square metre

Computed on all flats or on some?

Mean OECD-PISA score for a given country

Computed on all students or on some?

Mean consumption per household of a given
commodity, consumer price indices, inflation rate
...

All these are in general sample based country
averages built by statistical offices

These sample averages are different object than:

Mean distance between the earth and the sun

Mean height of the persons in this room

Sample based estimates of a given socioeconomic variable is often the result of a survey

When the sample extends to the entire population it is called a *census*

The first census; Babylonian
~3600 BC, “taken every six or
seven years and counted the
number of people, livestock,
quantities of butter, honey,
milk, wool and vegetables” [*];
Egyptians and Romans followed

(Much earlier than the Hammurabi code
~1750BC)



<http://persiababylonia.org/archives/>

[*]Australian Statistical Office, <https://www.abs.gov.au/>

In modern Europe, census were first seen in Germany, and become regular with Frederick William I of Prussia, in the XVIII century

The modern use of the term statistics comes from the German Statistik (data about the state), in turn coming from Latin status

status = "a station, position, place; order, arrangement, condition", modern Latin statisticum (collegium) = state affairs, Italian Statista,

'A Brush with Catastrophe': Inside the 2020 Census Meltdown

How the pandemic and Trump took a fun house mirror to America's once-in-a-decade look at itself.

POLITICO

By ZACK STANTON

09/10/2020 06:48 PM EDT



Who is counted or not counted in a census is the locus of political conflicts

➔ For apportioning house seats, funding of social programmes...

If using a sample, is the sample representative?

“For decades, psychologists have done research on college students in “Western, Educated, Industrialized, Rich and Democratic” (WEIRD) societies and assumed that the conclusions hold for all people” (Narayanan, 2022)

Narayanan, Arvind. 2022. “James Baldwin Lecture Series: ‘The Limits Of The Quantitative Approach To Discrimination.’” Department of African American Studies. 2022. <https://aas.princeton.edu/events/2022/james-baldwin-lecture-series-limits-quantitative-approach-discrimination>.

To decide if a sample is representative or not a possible question is:

Is the sample random?

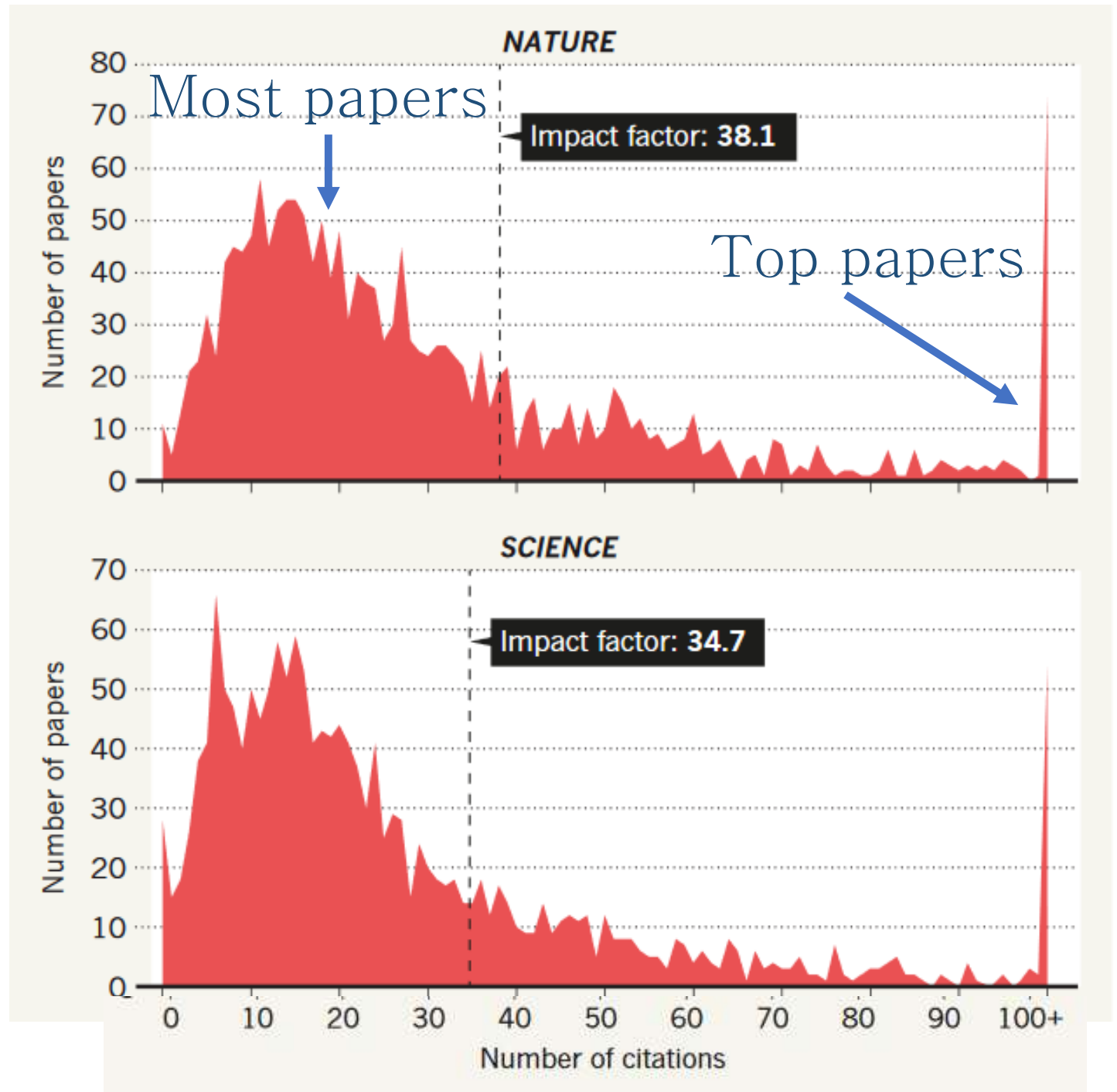
Truly random samples are hard to come by in the social science

Samples may be ‘*biased*’ by data availability, predisposition of the analyst (framing), disciplinary traditions, purpose or goal of the analysis ...

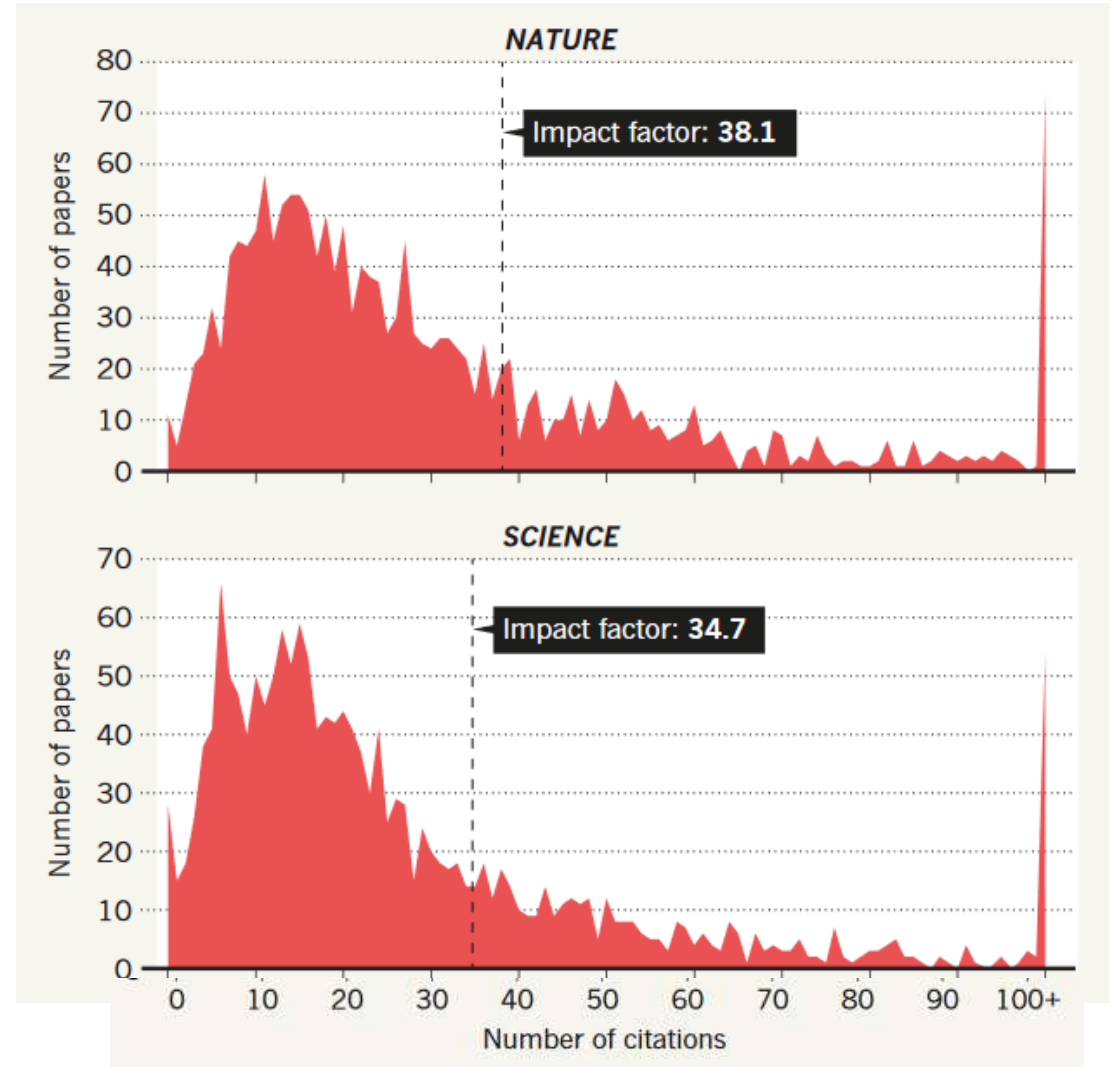
Back to the mean now: the example of the impact factor of academic journals

The average paper is cited much less than the journal's impact factor

Source: E.
Callaway, 2014
Publishing elite turns
against impact factor,
Nature, 535, 210-211.



The mean may not be the most significant statistics to draw from a ‘*distribution*’ with a long ‘*tail*’,



Variance and standard deviation: measures of dispersion

Population mean

Sample mean

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

and

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$


Why $n-1$ and not just n ?

Variance of the population

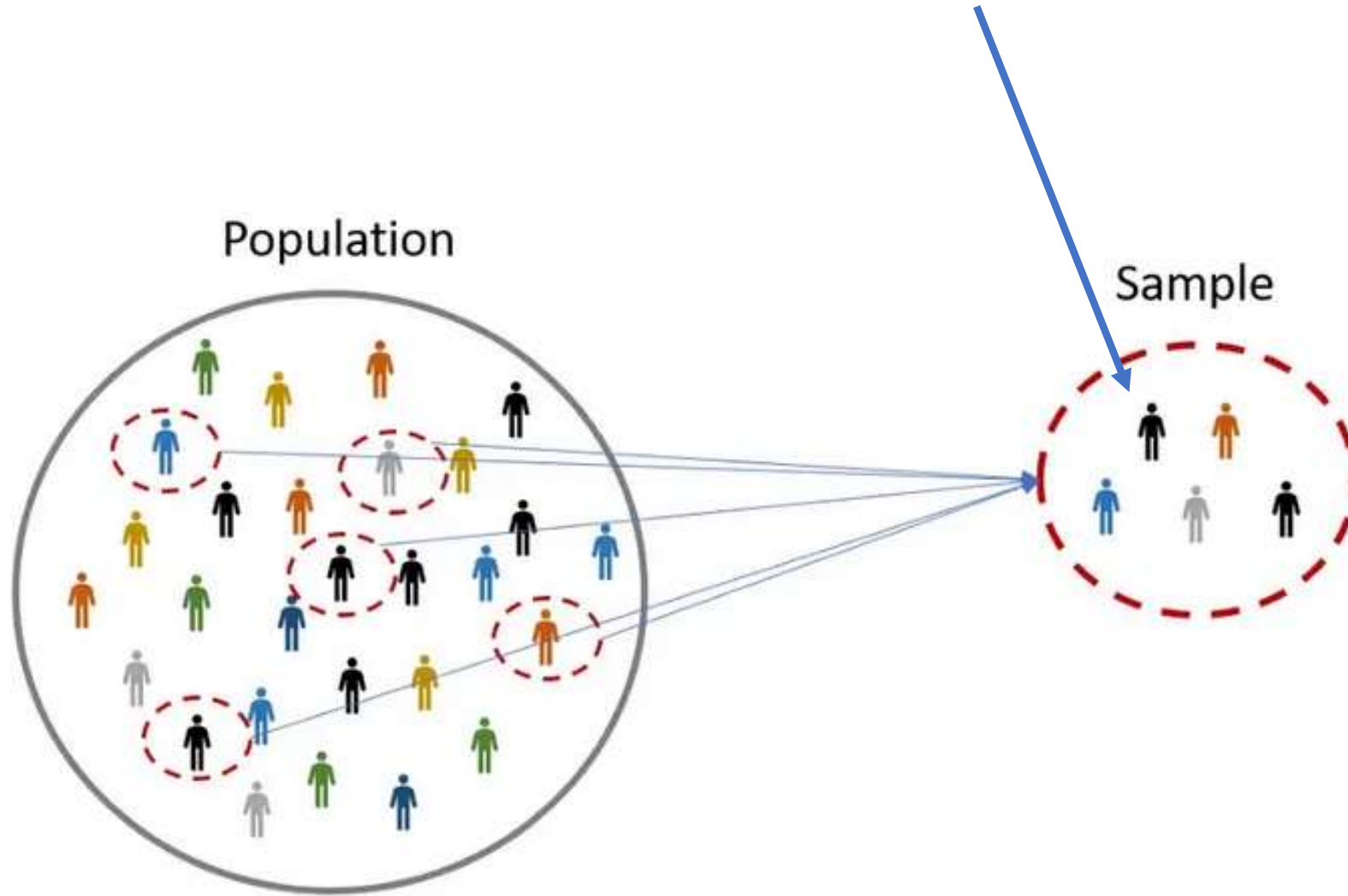
Size population

Variance of the sample

Size sample



More definitions: element or member of a sample



More definitions: a characteristic of the study that takes different values for different element is called a *variable*



element



Variable=hight

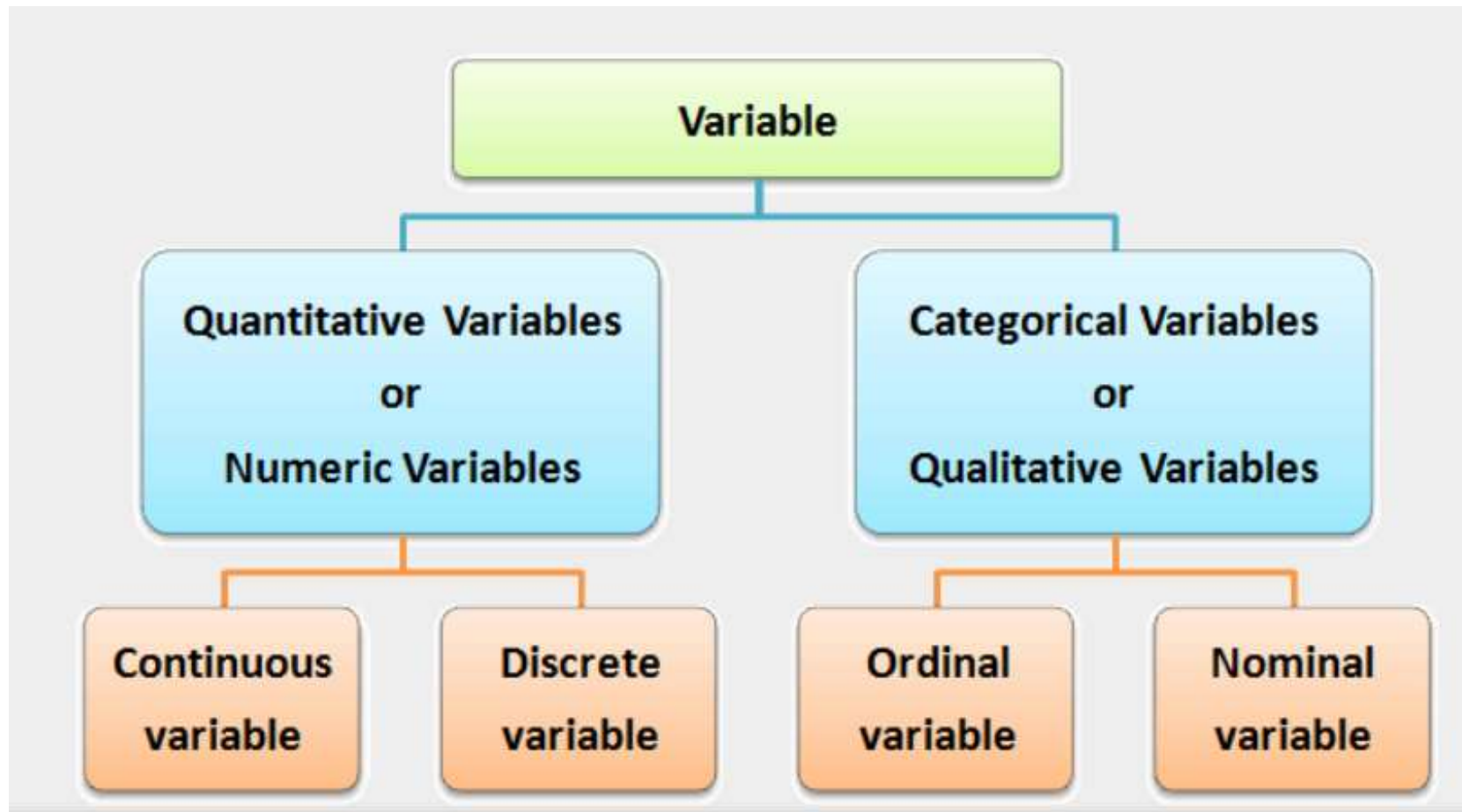
More definitions: the value of a variable for a given element is called *observation*

Element	Variable =weight (Kg)
John	71
Mary	68
...	...

More definitions: a collection of observations on one or more variables is called a *data set*

	Weight (Kg)	Eye colour	...
John	71	brown	...
Mary	68	green	...
...

Quantitative *Qualitative
or categorical*



Credits: <https://prinsli.com/quantitative-variables/>

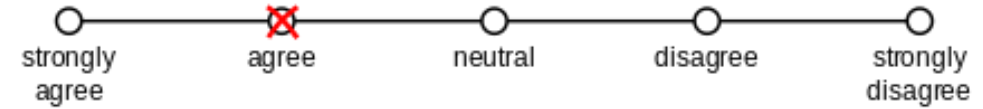
↓	↓	↓	↓
1.67m, 66 Kg	33 visitors on Friday	Low, medium, high	Red, green

**Ordinal
variable**

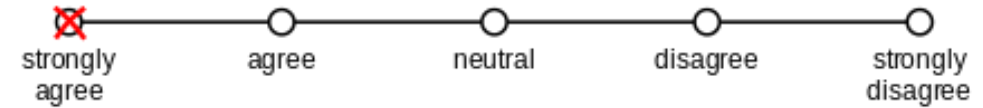
An example: A Likert scale for questionnaires

Tip: use odd number of levels

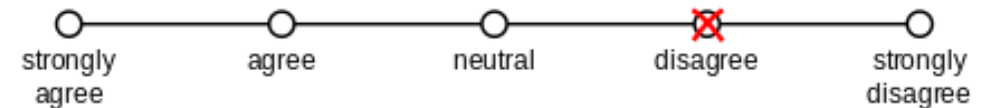
1. The website has a user friendly interface.



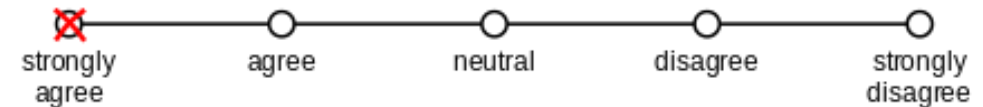
2. The website is easy to navigate.



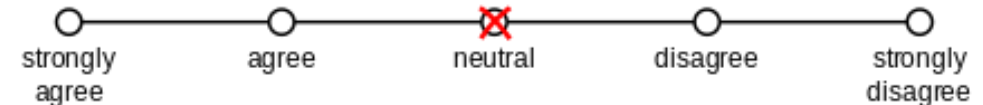
3. The website's pages generally have good images.



4. The website allows users to upload pictures easily.



5. The website has a pleasing color scheme.



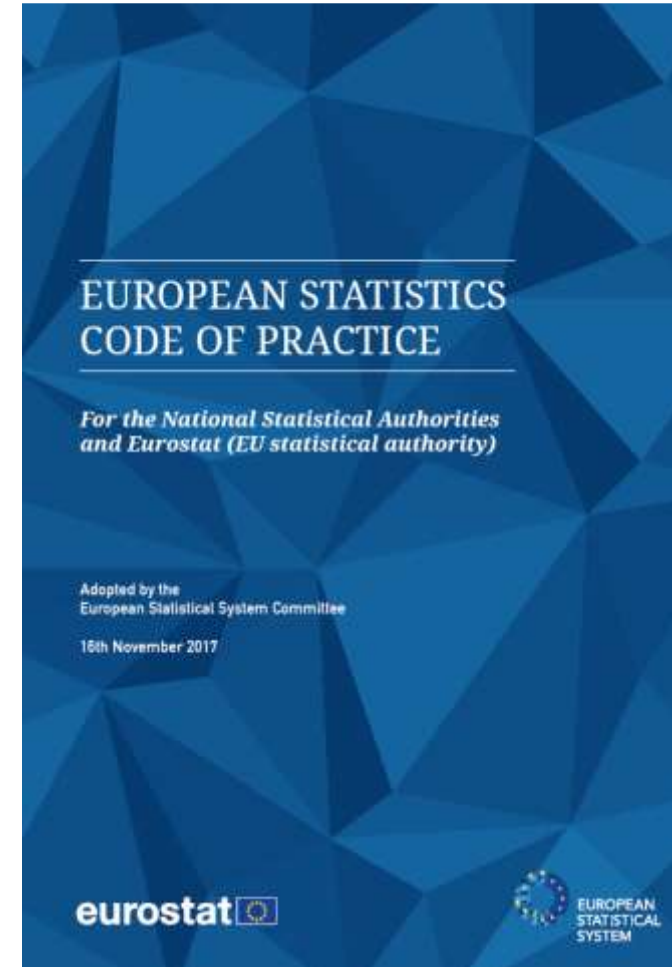
Source: Wikipedia Commons
https://en.wikipedia.org/wiki/Likert_scale#/media/File:Example_Likert_Scale.svg

When there is a *dataset* there ought to be *metadata*

U.S. Census Bureau Statistical Quality Standards



<https://www2.census.gov/about/policies/quality/quality-standards.pdf>

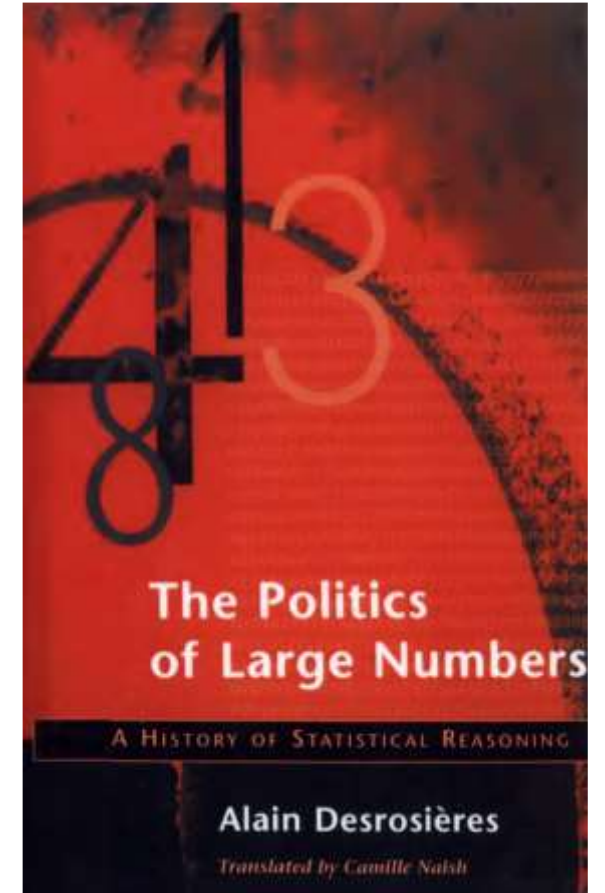


Back again to the mean: is it just a name?

Expressions such as

“one carrying the other”
“the strong carrying the weak”
“year in, year out”
“one inside another”
“grosso modo”
“when all is said and done”
“all things considered”

(XVIII century) → some kind of average behaviour (p. 72)



Is the mean just a name? Or ‘thing’

➔ Quetelet and the discovery of the mean

Mean as a ‘thing’ beside a ‘name’:

The regularity in the value
certain numbers, such as a
country’s births, deaths,
marriages, suicides ...

As well as of their *distribution* ...



Raymond Quetelet
(1796–1874)

Others made the same the discovery:

The regularity in the difference male versus female births as a proof of divine providence (John Arbuthnot)

Johann Peter Süssmilch (1706–1767):
a divine order

(Desrosières p. 74)

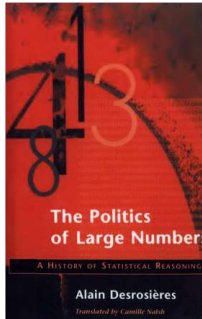


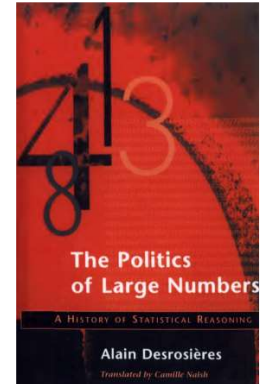
Image credit: Wikipedia Commons



John Arbuthnot
(1667–1735)

Quetelet discovery of the *average men* → social statistics, statistical offices, first congress of statistics (1853)

Society then also can exist as a 'thing'
(Émile Durkheim (1858–1917))



Durkheim ← Jean-Jacques Rousseau's idea that there is a *general will* superior to the *will of everyone*

More definition: cross sectional data, time series, panel data

Cross sectional data: elements of a population or a sample at one point in time

Time series: One elements traced ad different points in time

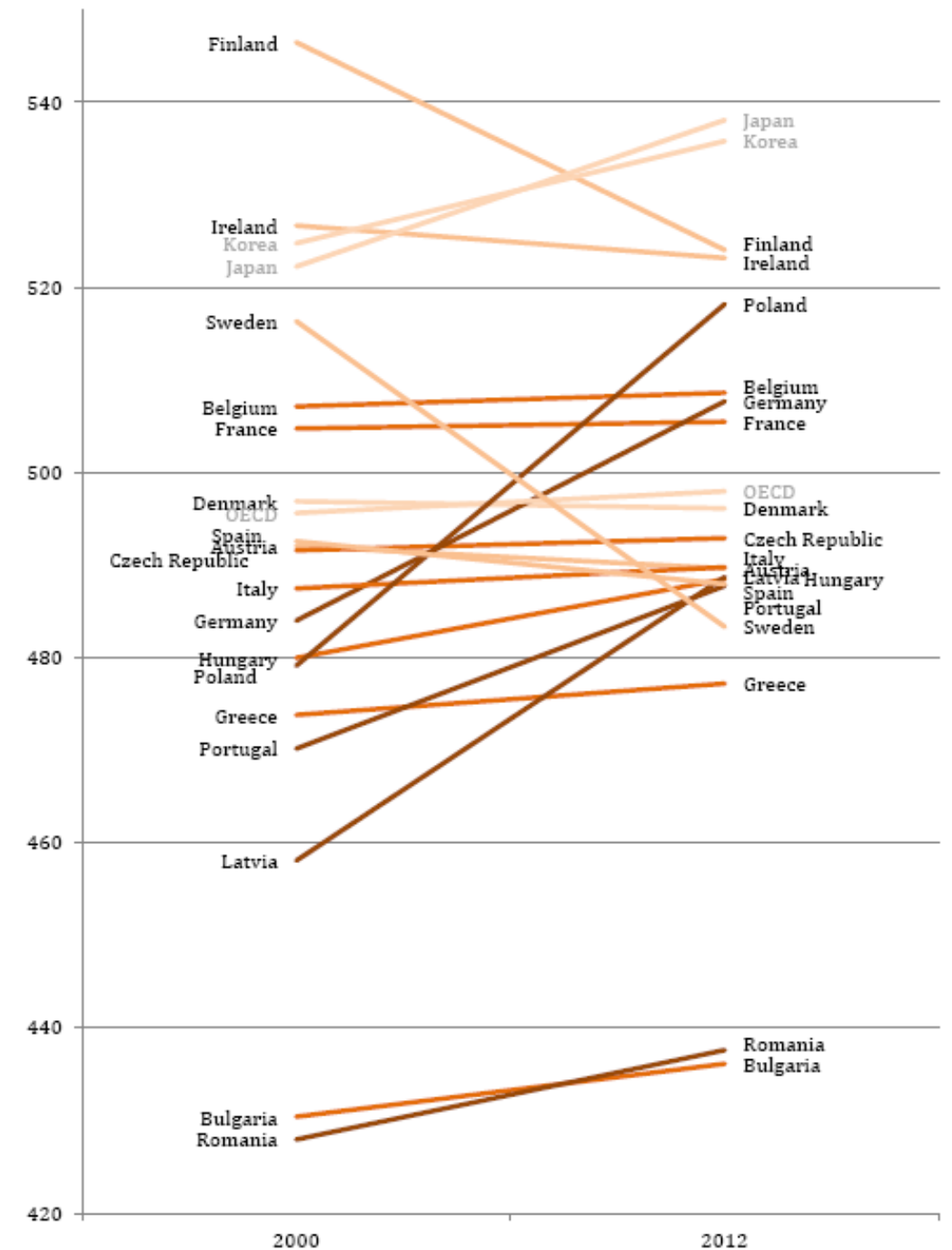
Panel: data from many units, over many points in time



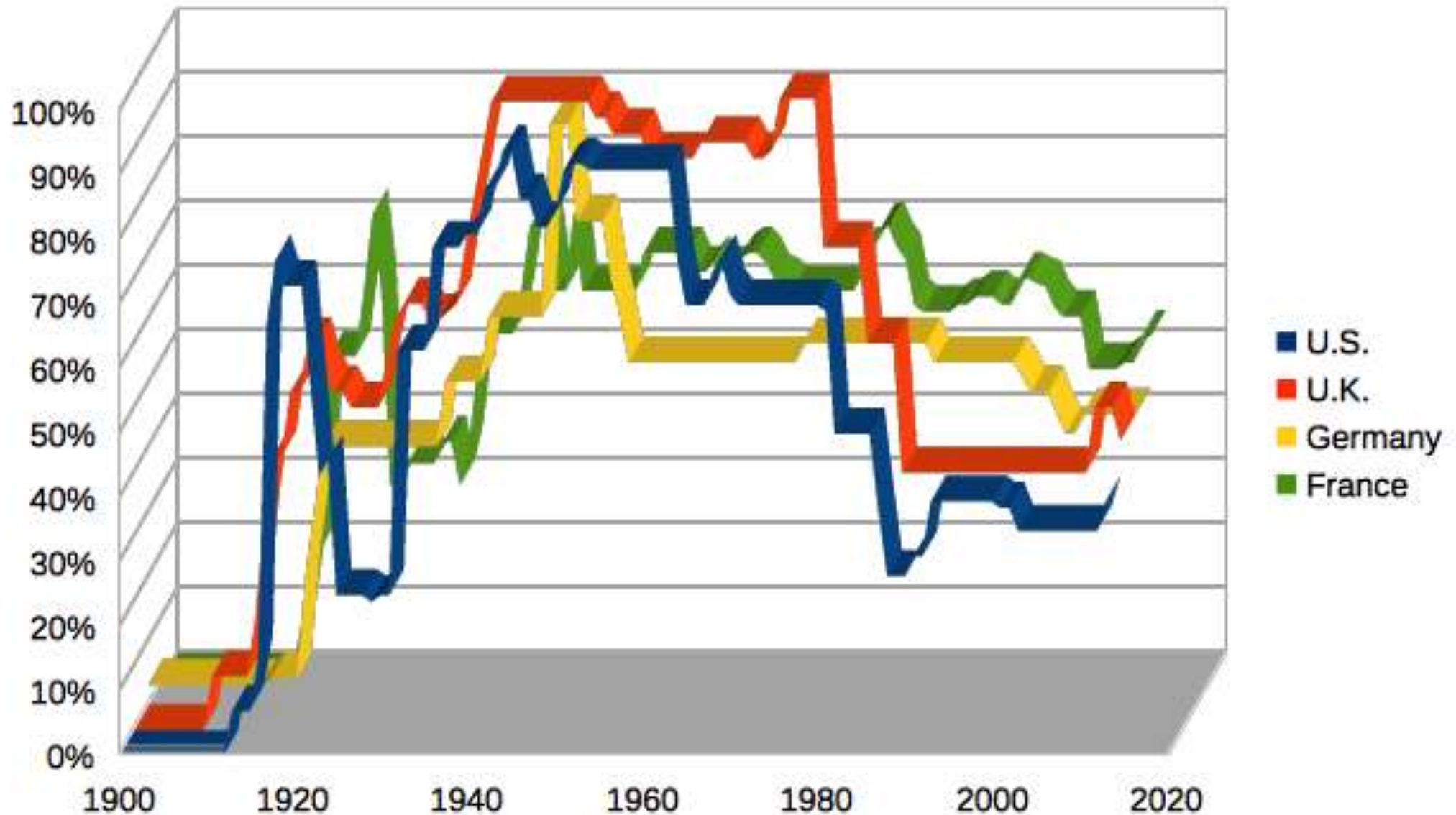
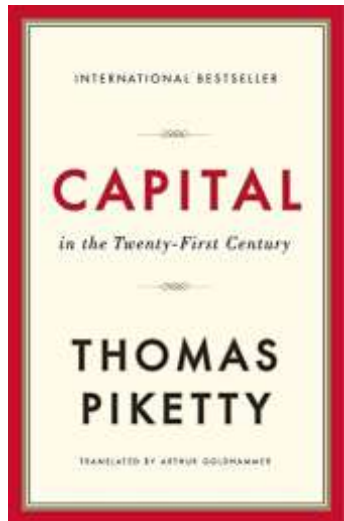
Cross section: snapshot of (students / universities) performance at one point in time

OECD-PISA at two points in time (2000, 2012) to produce a time series

Source: Woessmann, L. (2014), "The economic case for education", EENEE Analytical Report 20, European Expert Network on Economics of Education (EENEE), Institute and University of Munich



Top income
tax rate
1900–2013
(p. 499)



https://es.wikipedia.org/wiki/Archivo:Piketty_14.1_top_marginal_income_tax_rate.png

Organizing data

Row data, or ungrouped data; 30 students
& how they reach the university






walk, bus, bike, walk, bike, bus, walk, car, walk, bike,
bike, bus, walk, walk, walk, car, bus, walk, bus, bus,
walk, car, car, walk, walk, train, bike, bus, walk, walk

walk, bus, bike, walk, bike, bus, walk, car, walk, bike, bike, bus, walk, walk, walk, car, bus, walk, bus, bus, walk, car, car, walk, walk, train, bike, bus, walk, walk








Tally and frequency; sum=30

(source <https://thirdspacelearning.com/gcse-maths/statistics/tally-chart/>)

Transport	Tally	Frequency	Relative frequency
Walk		13	$13/30=0.43$
Bus		7	$7/30=0.23$
Car		4	$4/30=0.13$
Bike		5	$5/30=0.17$
Train		1	$1/30=0.03$

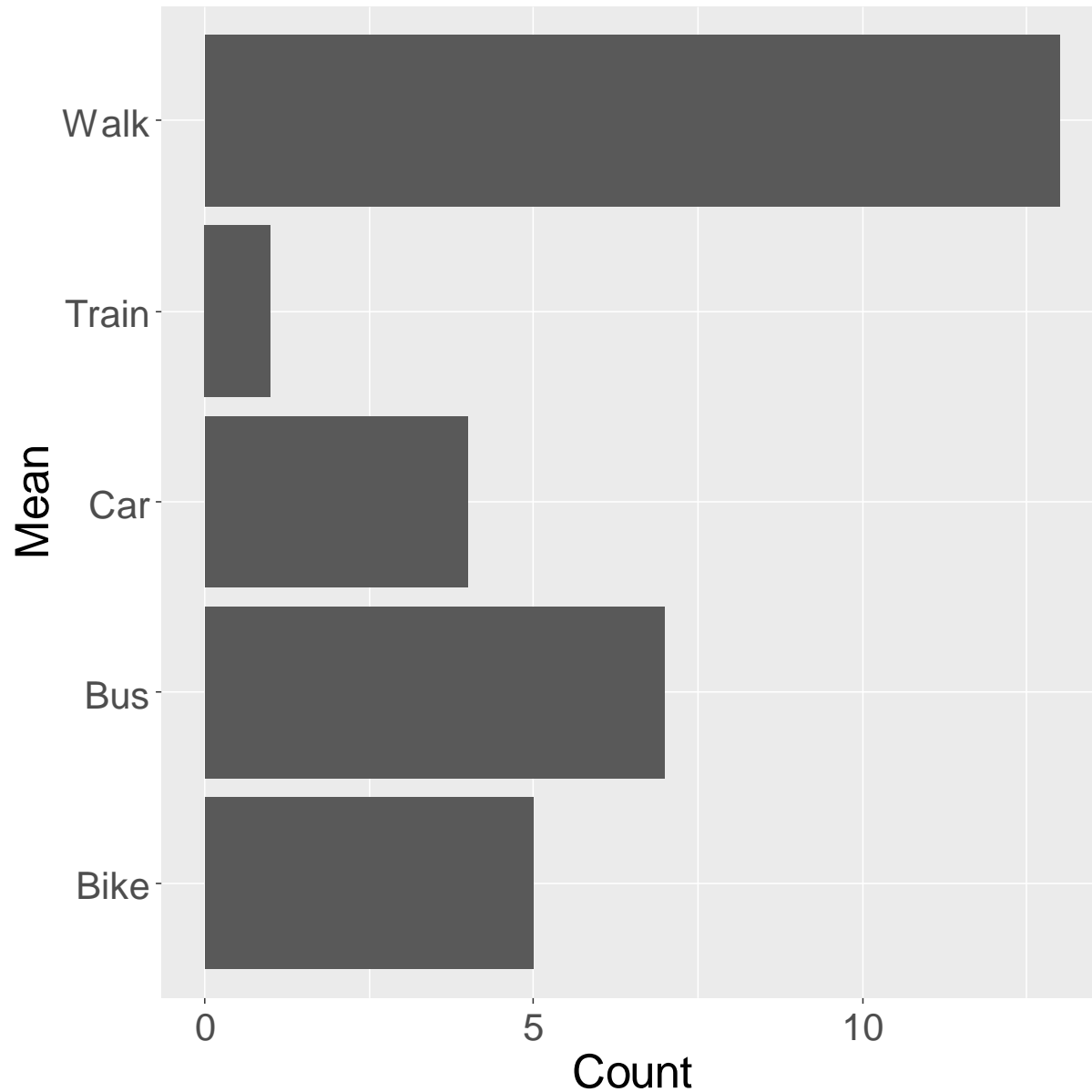
Tally and frequency; sum=30

(source <https://thirdspacelearning.com/gcse-maths/statistics/tally-chart/>)

Transport	Tally	Frequency
Walk		13
Bus		7
Car		4
Bike		5
Train		1

Relative frequency
$13/30=0.43$
$7/30=0.23$
$4/30=0.13$
$5/30=0.17$
$1/30=0.03$

Relative frequency;
sum=0.99

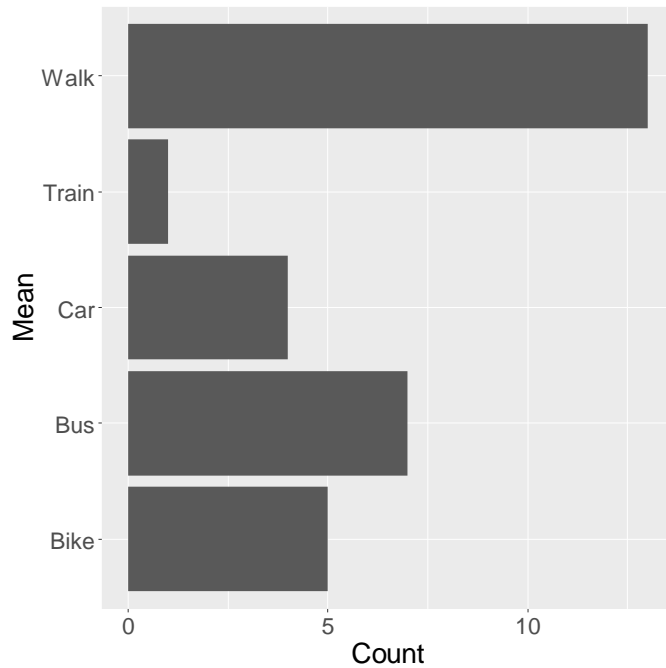


Transport	Tally	Frequency
Walk		13
Bus		7
Car		4
Bike		5
Train		1

Bar chart

```
library(ggplot2)
```

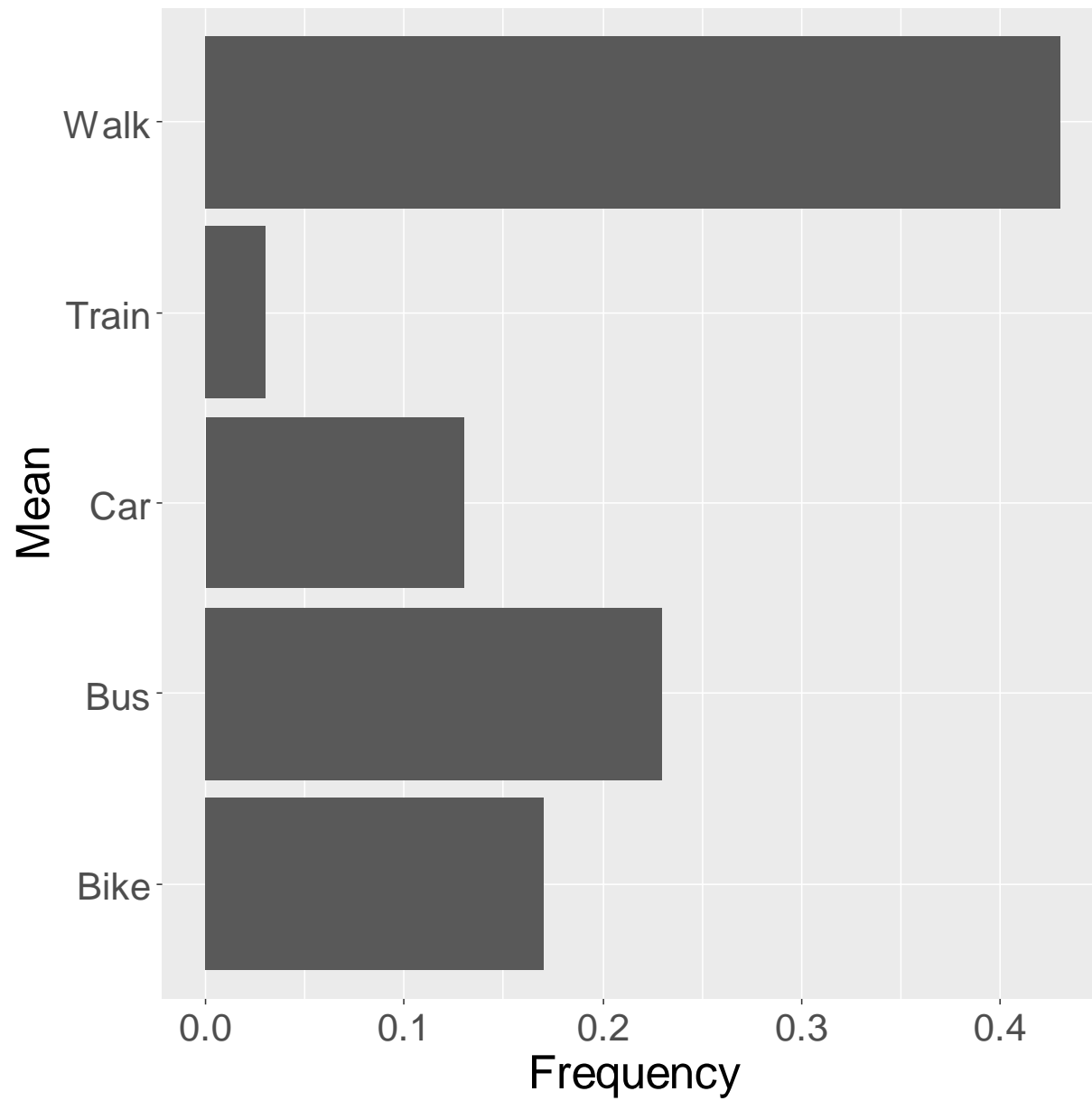
```
Commuting_Data <- read.table("C:/Users/Andrea/R/To_Uni.dat", header=T, sep="\t")
ggplot(Commuting_Data, aes(Mean, Count)) +
  geom_bar(stat = "Identity") +
  coord_flip() + theme(text = element_text(size = 20))
```



To_Uni.dat - Notepad

File	Edit	Format	View	Help
Mean	Count			
walk	13			
Bus	7			
Car	4			
Bike	5			
Train	1			

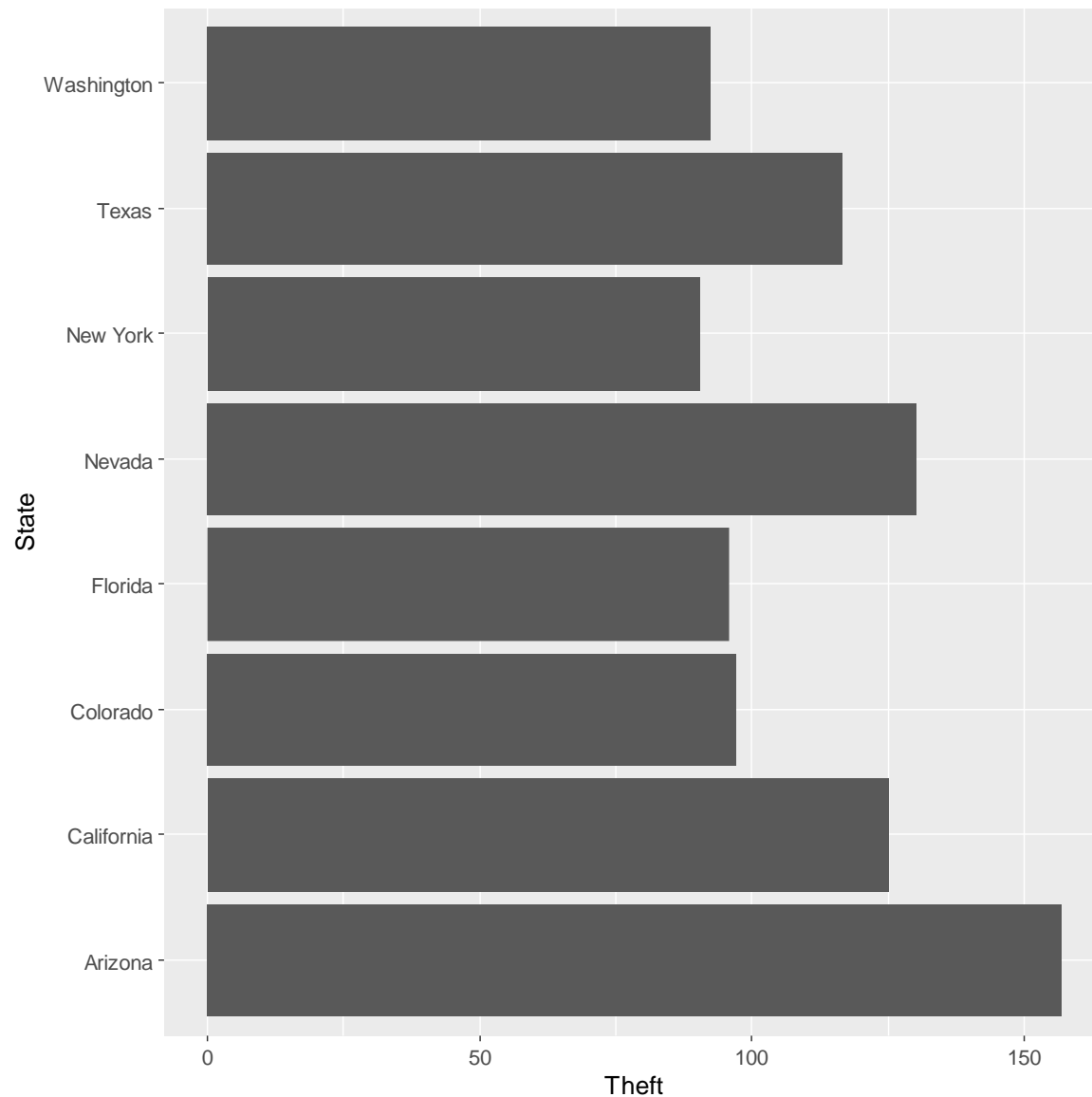
```
Commuting_Data <- read.table("C:/Users/Andrea/R/To_Uni.dat", header=T, sep="\t")
ggplot(Commuting_Data, aes(Mean, Count)) +
  geom_bar(stat = "Identity") +
  coord_flip() + theme(text = element_text(size = 20))
```



Relative frequency
$13/30=0.43$
$7/30=0.23$
$4/30=0.13$
$5/30=0.17$
$1/30=0.03$

Relative
frequency
chart

```
ggplot(Commuting_Data_Freq, aes(Mean, Frequency)) +
  geom_bar(stat = "Identity") +
  coord_flip() + theme(text = element_text(size = 20))
```



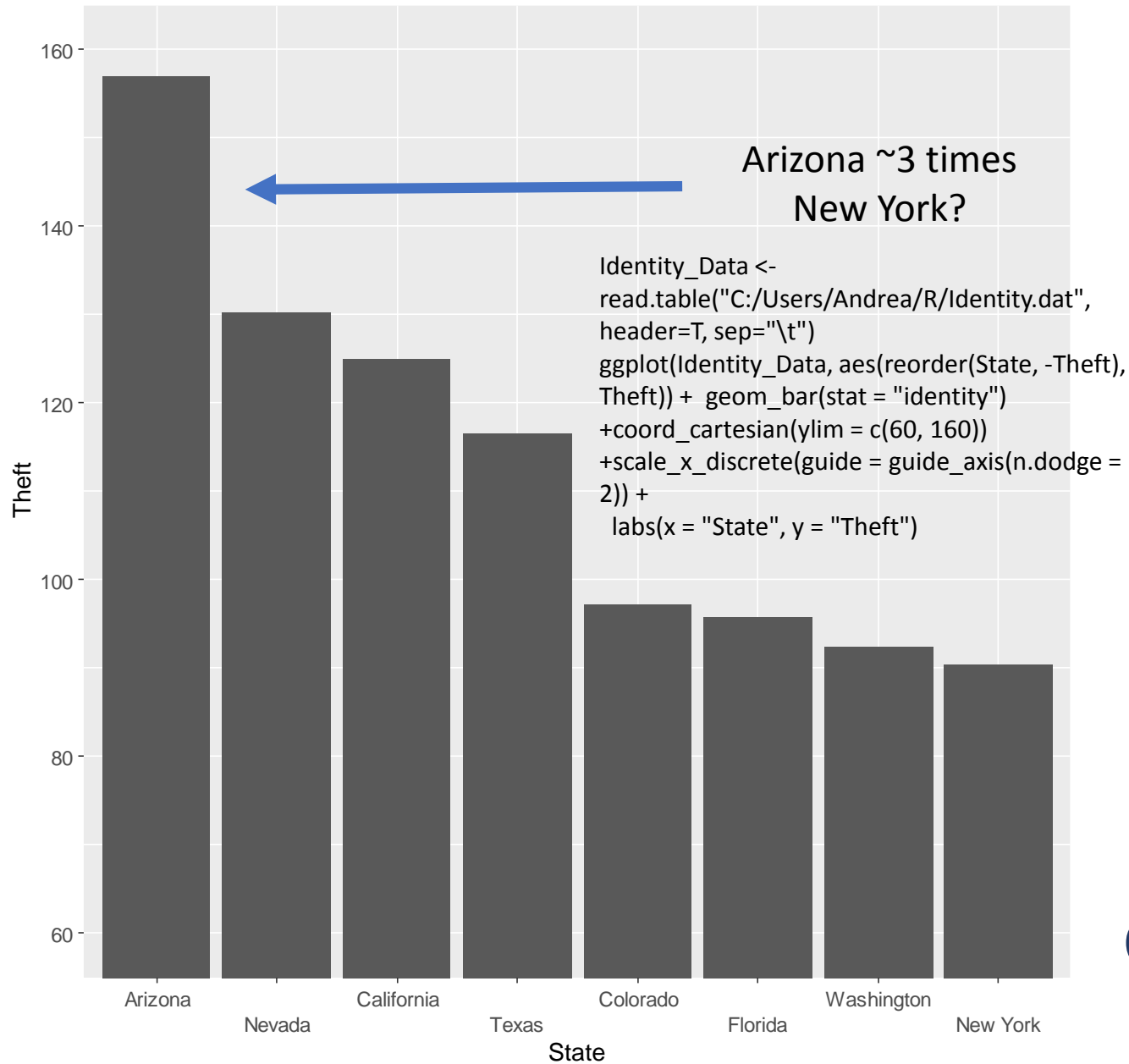
State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

Identity theft in different US states, cases/1000 people

Source: Mann, Prem S. 2010. Introductory
Statistics. Wiley, p.61.

Correct

```
ggplot(Identity_Data, aes(reorder(State, -Theft), Theft)) + geom_bar(stat = "identity")
+coord_cartesian(ylim = c(0, 160)) +scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
labs(x = "State", y = "Theft")
```



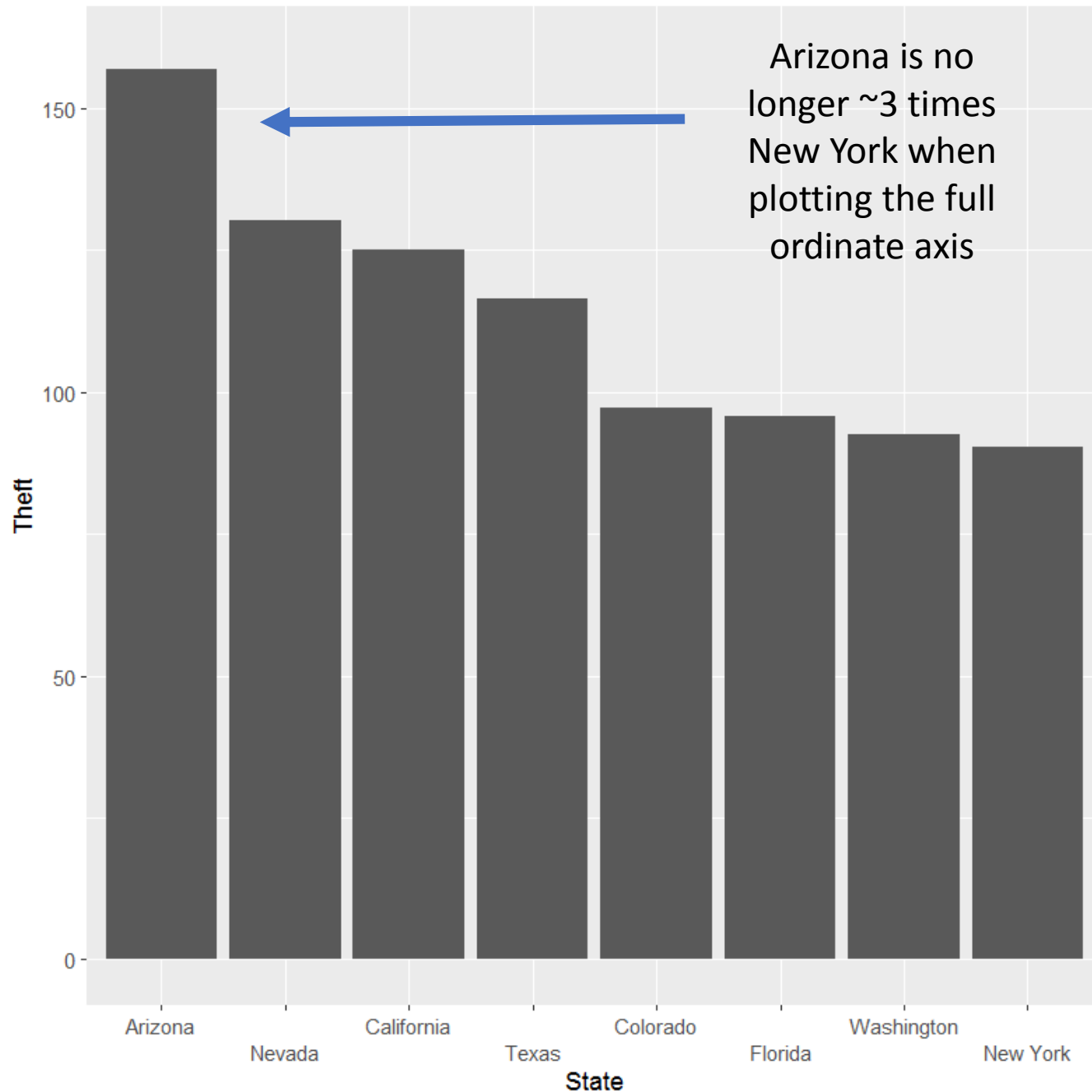
State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

Identity theft in different US states, cases/1000 people

Source: Mann, Prem S. 2010. Introductory
Statistics. Wiley, p.61.

Correct?



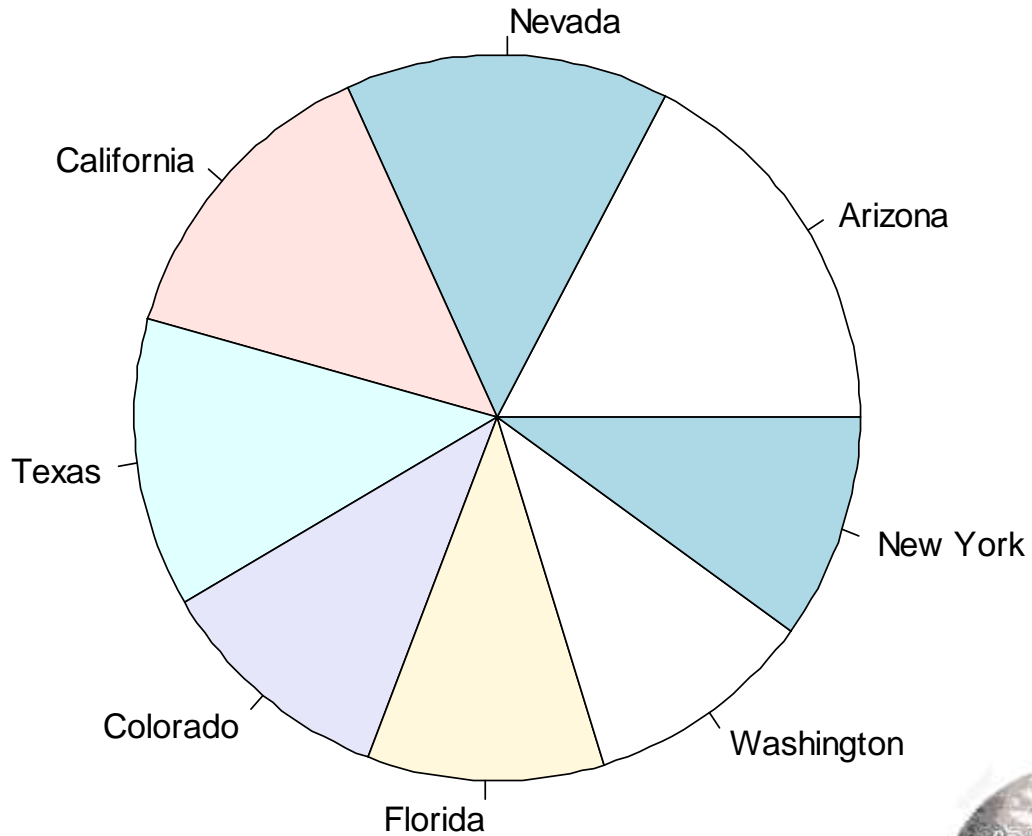


State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

Identity theft in
different US states,
cases/1000 people

Source: Mann, Prem S. 2010. Introductory
Statistics. Wiley, p.61.

Pie Chart of States



State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

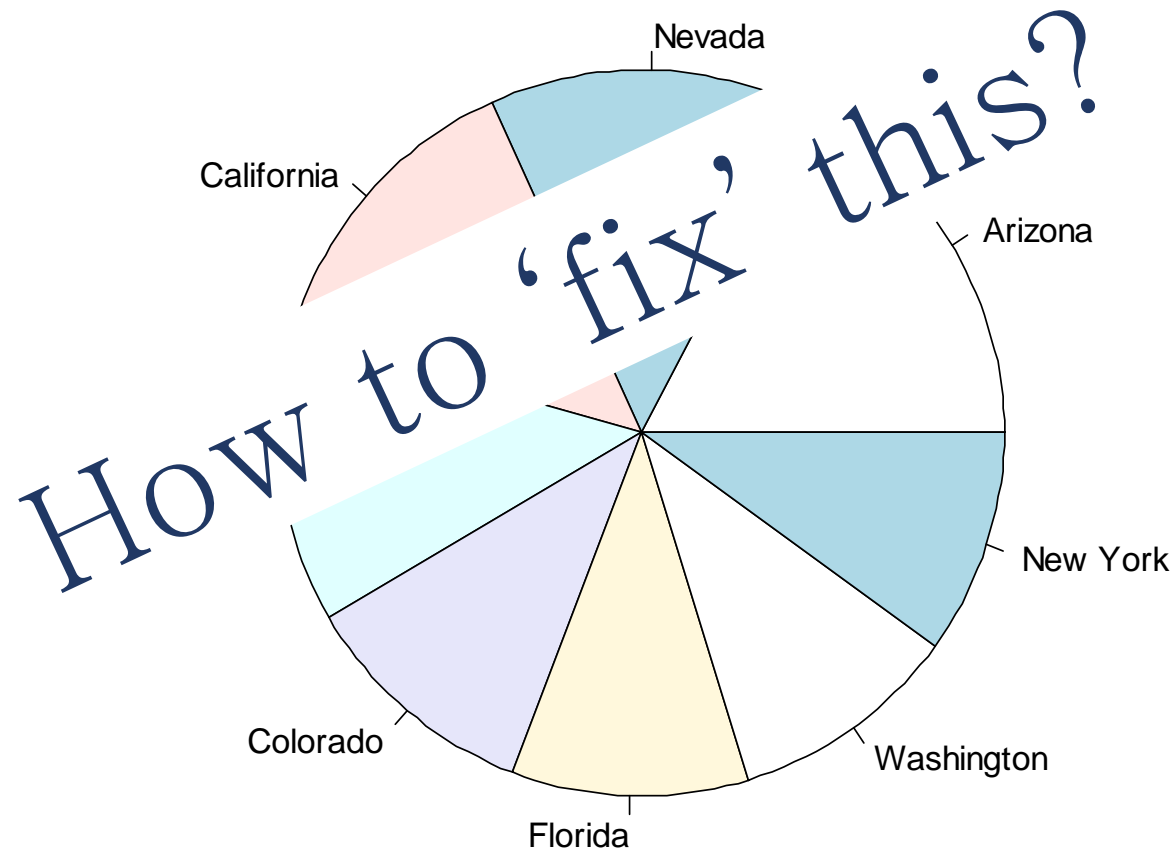
Identity theft in
different US states,
cases/1000 people

Source: Mann, Prem S. 2010. Introductory
Statistics. Wiley, p.61.



Correct?

Pie Chart of States



```
xdata <- c(156.9,130.2,125.0,116.5,97.2,95.8,92.4,90.3)
labs <-
c("Arizona","Nevada","California","Texas","Colorado","Florida","Washington","New_York")
pie(xdata,labs, main="Pie Chart of States")
```

State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

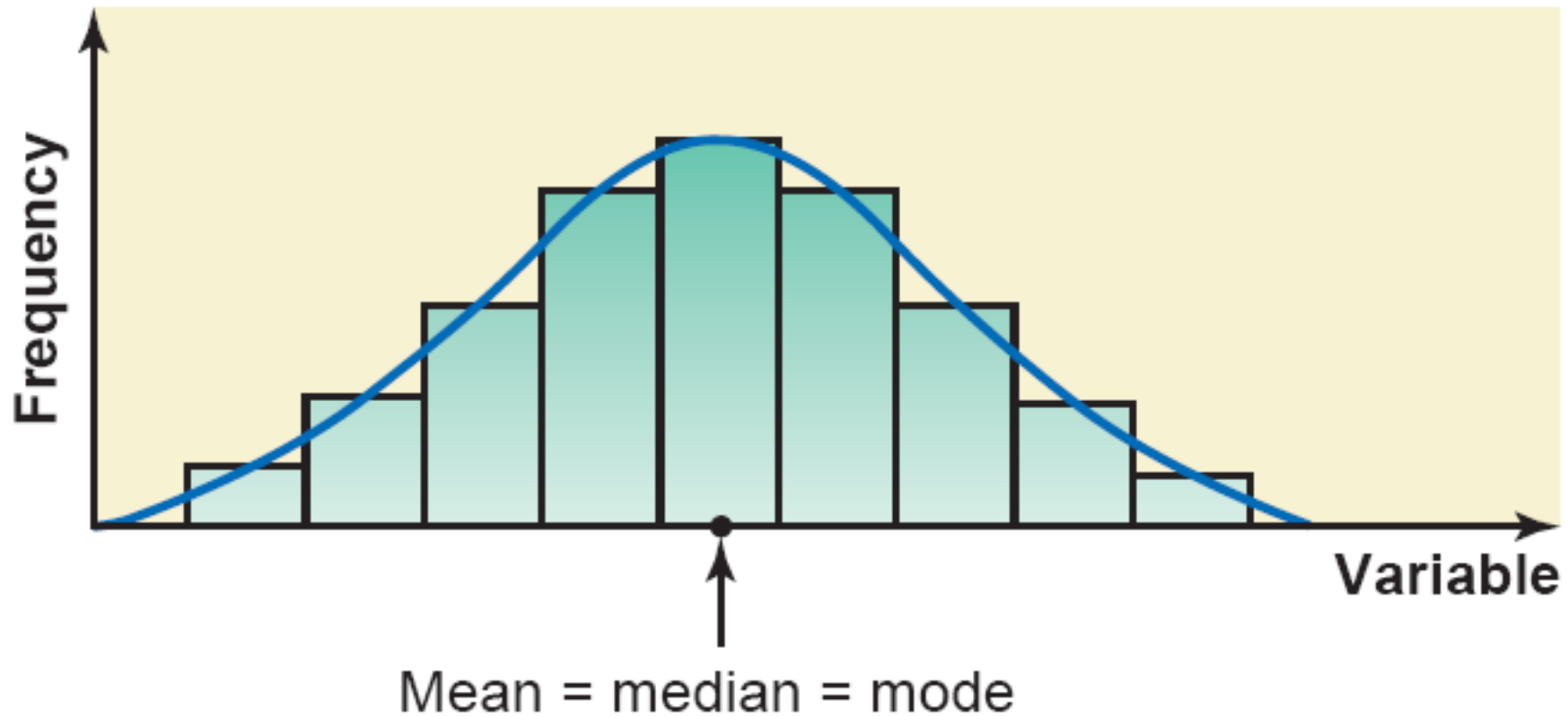
Identity theft in
different US states,
cases/1000 people

Source: Mann, Prem S. 2010. Introductory
Statistics. Wiley, p.61.



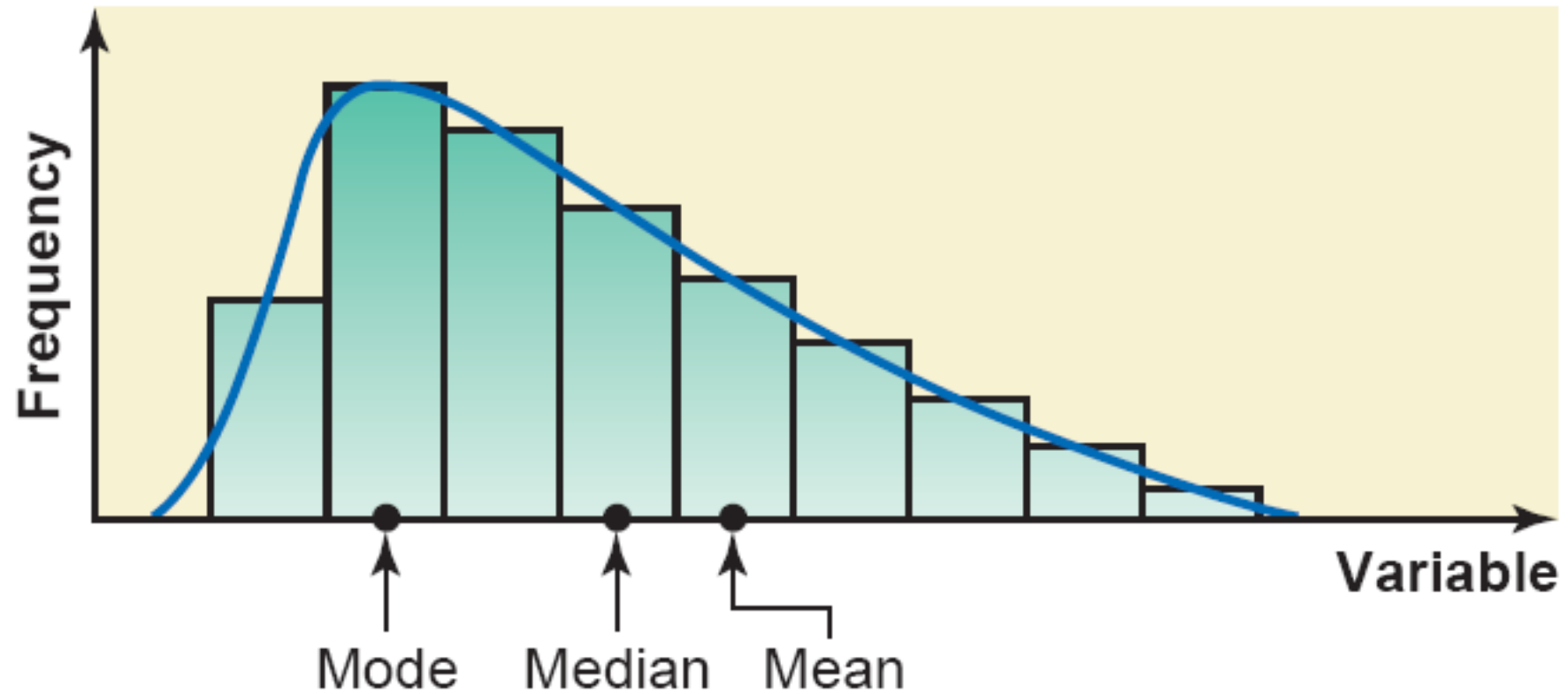
More elementary concepts:
mode, median and percentiles

A symmetric distribution



Source: Mann, Prem S. 2010. Introductory Statistics. Wiley.

A non-symmetric distribution



Source: Mann, Prem S. 2010. Introductory Statistics. Wiley.

Median: order the data and identify the midpoint

For our identity theft data

State	Theft
Arizona	156.9
Nevada	130.2
California	125.0
Texas	116.5
Colorado	97.2
Florida	95.8
Washington	92.4
New York	90.3

4th value 97.2
5th value 116.5

Median = $(97.2 + 116.5) / 2 = 106.85$

From lowest to highest

The median can be a more ‘democratic’ way (than the mean) to describe a population in the presence of skewed distributions: the gross domestic product (GDP) may tell a different story than median household income

In voting theory a good practice is the majority judgment method (Balinski & Laraki), based on the median of the scores given by different voters to the candidates

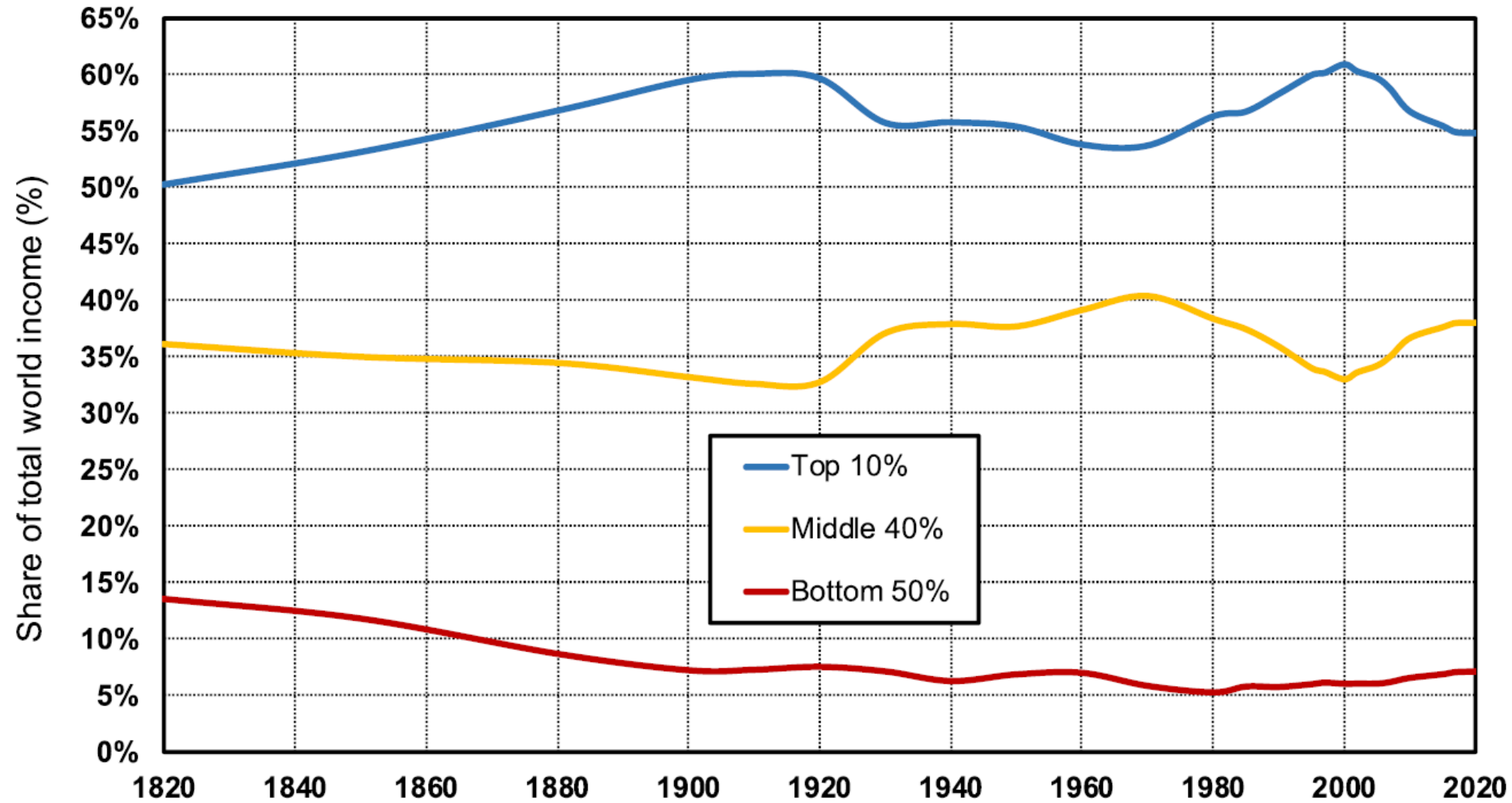
MAJORITY JUDGMENT

Measuring, Ranking, and Electing



MICHEL BALINSKI AND RIDA LARAKI

Percentiles are also important



Source: GLOBAL INCOME INEQUALITY, 1820–2020: THE PERSISTENCE AND MUTATION OF EXTREME INEQUALITY, by Lucas Chancel and Thomas Piketty, *Journal of the European Economic Association* 2021 19(6):3025–3062

Sources

Specific queries, specific sources, example:

- US GDP? US Bureau of economic analysis
- US consumer price index? Bureau of Labor Statistics...

- [Bureau of Economic Analysis](#)
- [Bureau of Justice Statistics](#)
- [Bureau of Labor Statistics](#)
- [Bureau of Transportation Statistics](#)
- [Census Bureau](#)
- [DAP Public Dashboard](#)
- [Data.gov](#)
- [Economic Research Service](#)
- [Energy Information Administration](#)
- [Internal Revenue Service Tax Statistics](#)
- [National Agricultural Statistical Service](#)
- [National Center for Education Statistics](#)
- [National Center for Health Statistics](#)
- [National Center for Science and Engineering Statistics](#)
- [Office of Personnel Management](#)
- Social Security Administration [Office of Research Evaluation and Statistics](#)
- [USAspending.gov](#)

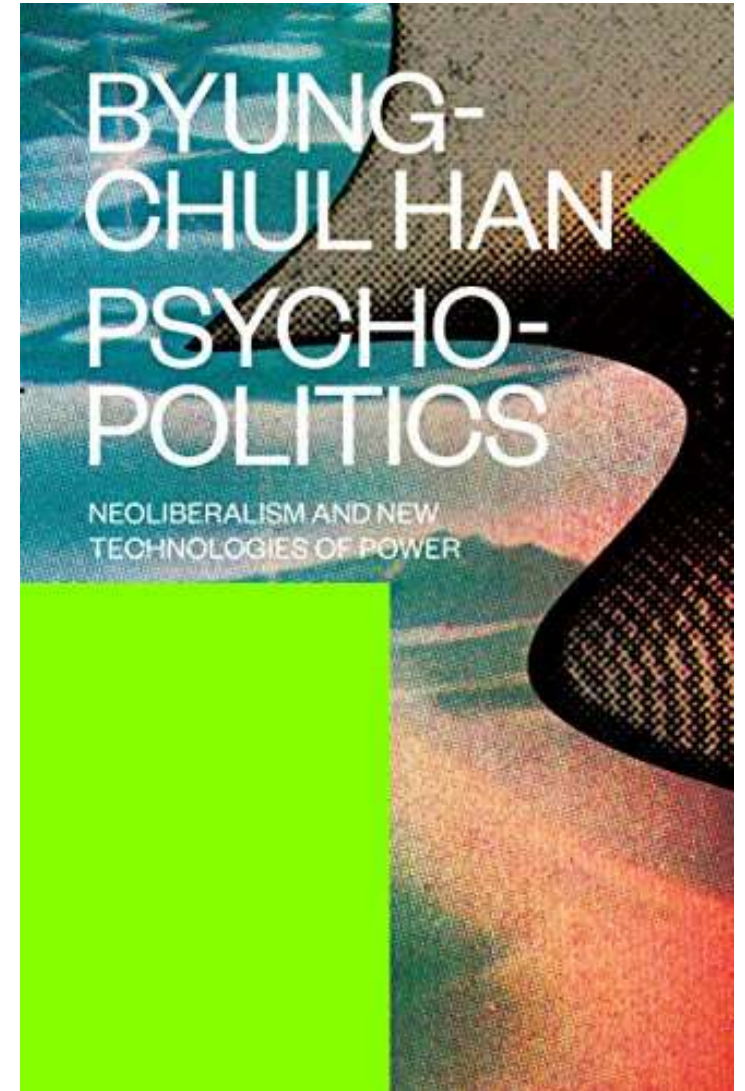


<https://www.usa.gov/statistics>

Something on
sociology of data

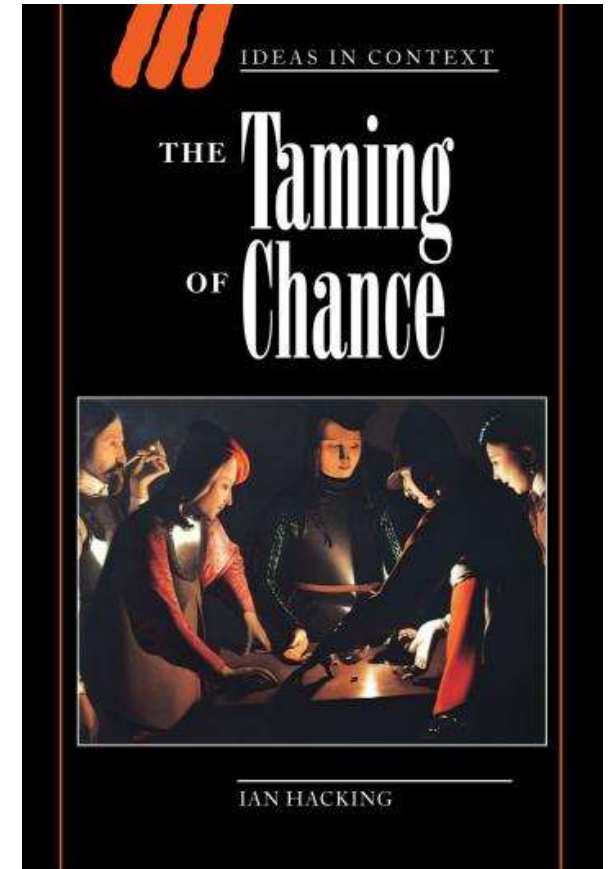
Sources? Data are everywhere!

A second enlightenment or
'dataism'?



So if data exploded during the XIX century as noted by sociologists and historians ...

... They are now exploding again. Why?



“The role [of statistical indicators] has increased significantly over the last two decades.

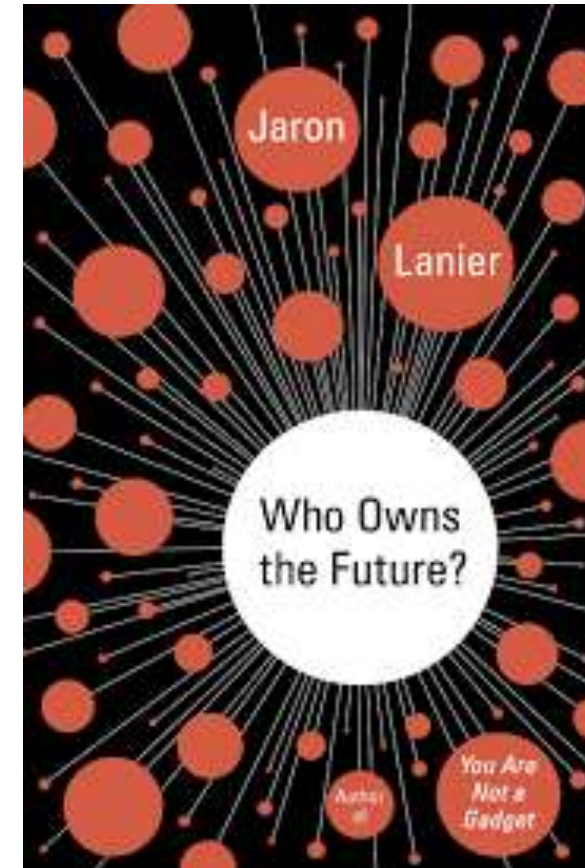
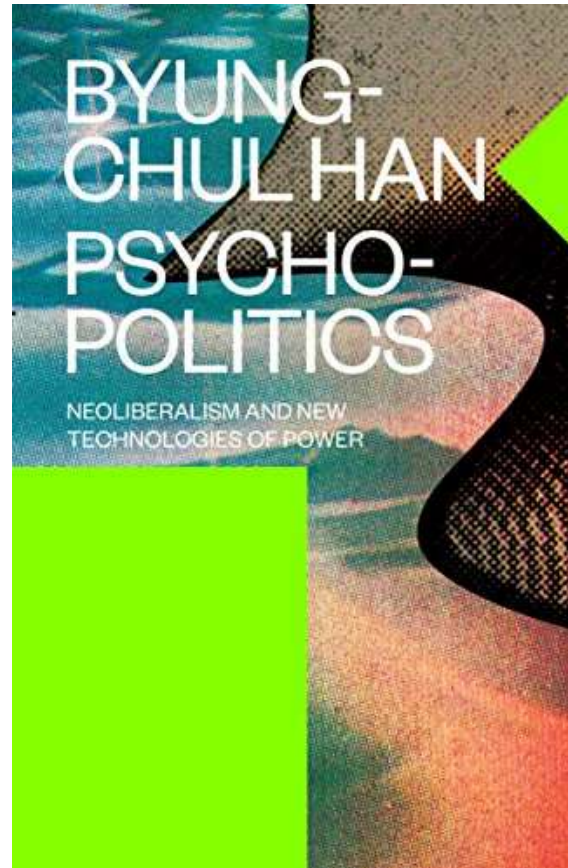
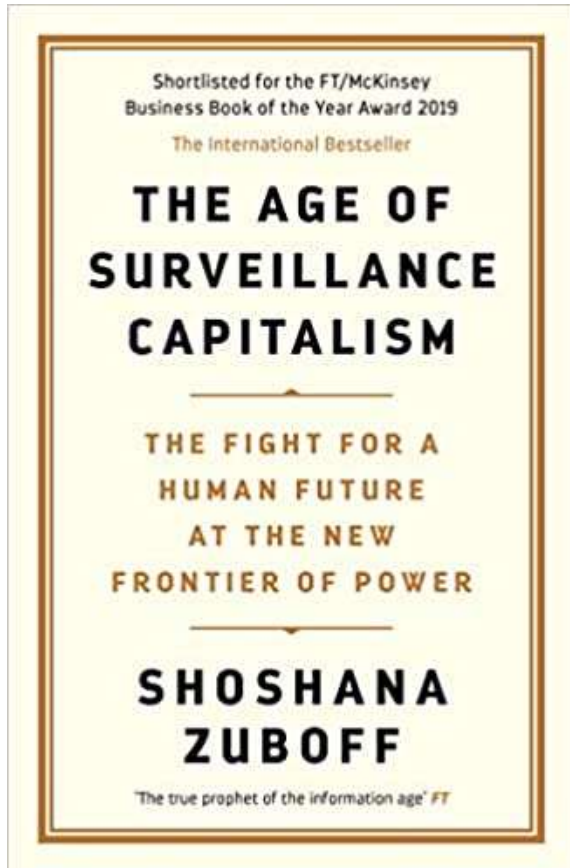


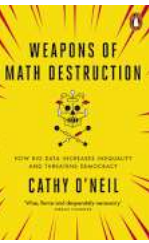
Jean-Paul Fitoussi,
Joseph Stiglitz and
Amartya Sen

This reflects improvements in the level of education in the population, increases in the complexity of modern economies and the widespread use of information technology”

CMEPSP (2009). Commission on the Measurement of Economic Performance and Social Progress, URL: [http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+ Commission+ report](http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report), last accessed June 2017.

For other scholars there are more data because
“data is the new oil” → sociology of quantification





NETFLIX

UNLIMITED TV SHOWS & MOVIES

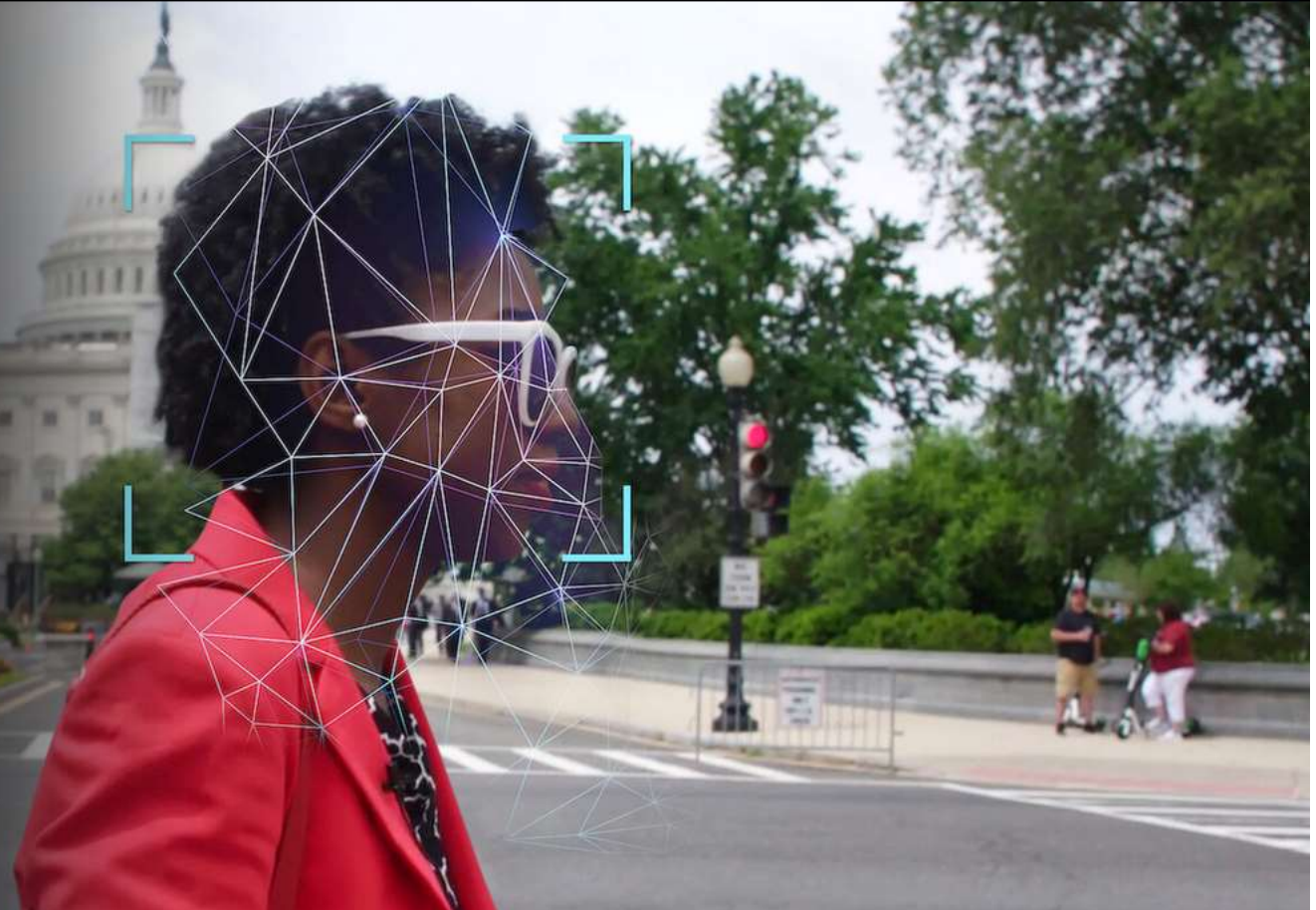
JOIN NOW

C O D E D B I A S

Coded Bias

2020 | 12+ | 1h 25m | Science & Nature Docs

This documentary investigates the bias in algorithms after M.I.T. Media Lab researcher Joy Buolamwini uncovered flaws in facial recognition technology.





Algorithmic Justice
League

<https://www.ajl.org/>

The New York Times

Bloomberg
Business

Forbes

TIME

FORTUNE

TED

WIRED

The Telegraph

A useful illustration of strategies of capture, starring O'Neil, Zuboff, Lanier, and GAFA technologists...



... such as Tristan Harris, former design ethicist at Google, explaining from inside how social media pursue addiction to maximize profit and manipulates people's behaviour



Tip: In your quest for sources start from academia, as someone might have gone through the problem before you

Google Scholar

WIKIPEDIA
The Free Encyclopedia

English

6 585 000+ articles

日本語

1 354 000+ 記事

Русский

1 875 000+ статей

Deutsch

2 751 000+ Artikel

Italiano

1 785 000+ voci

فارسی

941 000+ مقاله

Português

1 096 000+ artigos

Français

2 477 000+ articles

Español

1 823 000+ artículos

中文

1 323 000+ 条目 / 條目



EN ▾



Scopus Preview

Welcome to Scopus Preview

[What is Scopus ↗](#) [Blog ↗](#)

☒ Articles ☐ Case law

The End



<https://mstdn.social/@AndreaSaltelli/>