Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Data & Inference

WEEK 3A

Course outline

Week	Description
1	Measurement, data types, data collection, data cleaning
2	Data manipulation, data exploration, visualization techniques
3	Probability, statistical distributions, descriptive statistics
4	Statistical hypothesis testing, quantifying confidence
5	Time series analysis, autoregression, moving averages
6	Linear regression, parameter estimation, model selection, evaluation

Today's Lecture

No.	Activity	Description
1	Challenge	Using statistics for a fact based world view
2	Discussion	Truth and statistics
3	Case study	GapMinder
4	Analysis	Using descriptive statistics
5	Demo	Calculating descriptive statistics
6	Q&A	Matlab questions and feedback

Using statistics to deliver facts

- Statistics are commonly quoted for measuring quantities
- Business leaders and politicians use statistics now more than ever:
- Population, Wealth, Health, Immigration, ...
- Averages: current snapshot
- Growth rates: trends
- Proportions and probability
- Projections: targets and goals

Lightning strikes (Poll 1)

- Is the probability of being struck by lightning greater for women or men?

Slido.com
#78793

Second chance (Poll 2)

- Men are x times more likely to be struck by lightning than women [US data 2006-2016]
- X = 2,3,4,5

Slido.com
#78793

https://www.accuweather.com/en/weather-news/men-vs-women-which-gender-is-more-likely-to-be-fatally-struck-by-lightning/351789

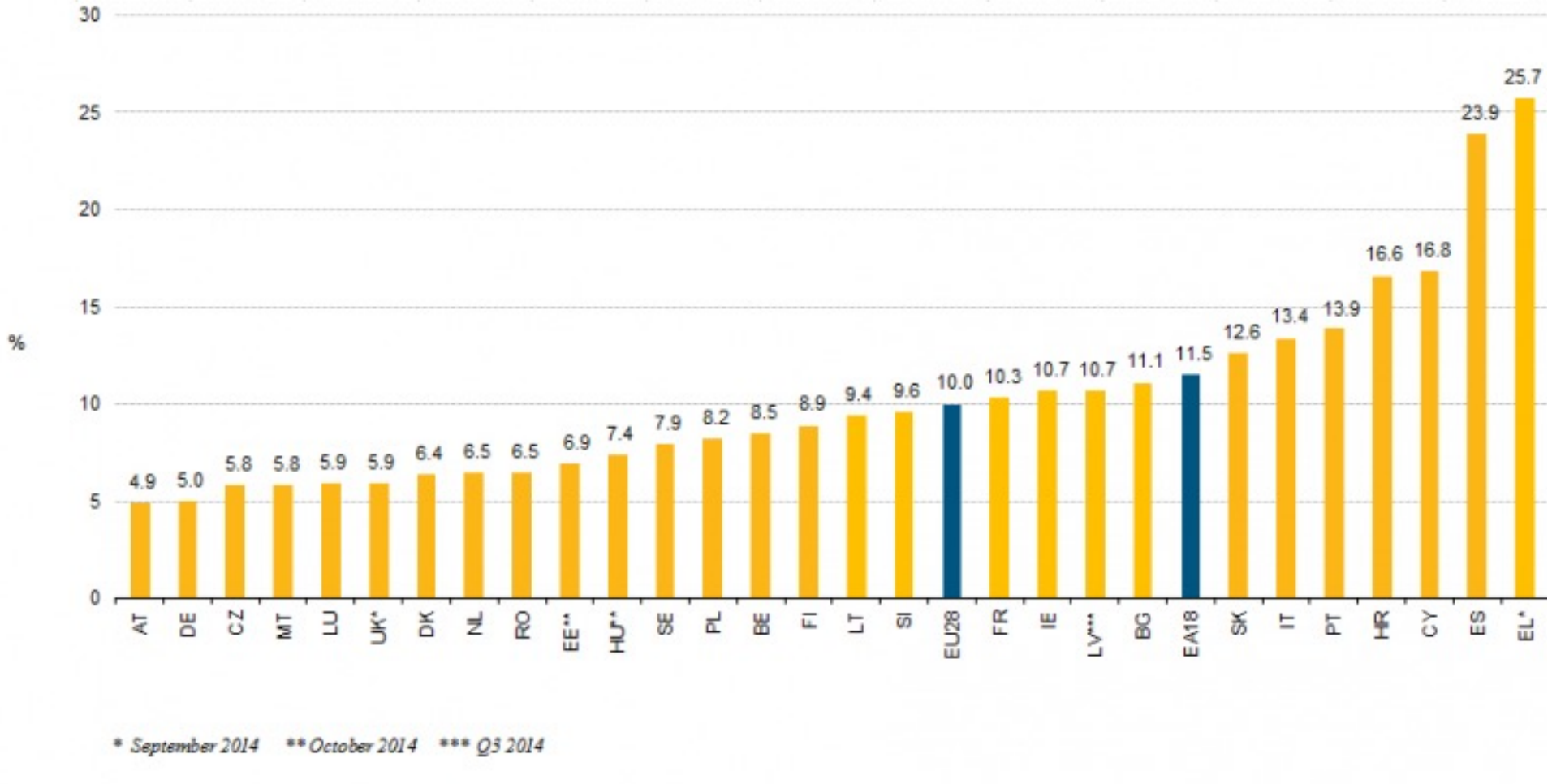
Ed Miliband: “only crisis-hit Spain has higher numbers of young unemployed than the UK”

- The Labour leader told The Sun that “only crisis-hit Spain has higher numbers of young unemployed than the UK”.
- While technically true, this is mainly because the UK has a larger population than most European countries, so is always likely to feature near the top of unemployment (or indeed employment) tables if we look only at the raw numbers.
- UK and Spain have 65m and 46m respectively.

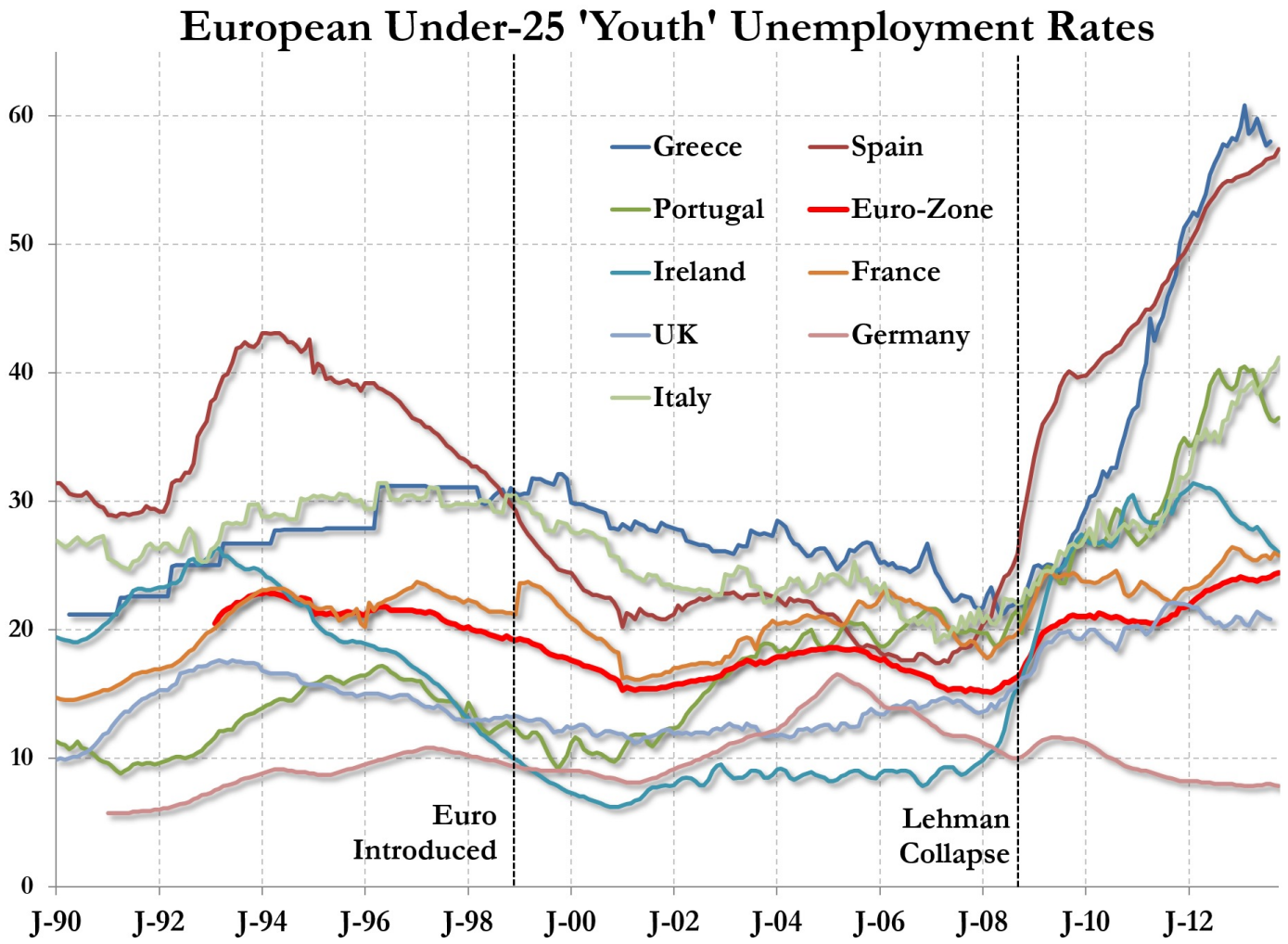
How can we decide for ourselves?

- If we want to know how the UK performs in relation to other countries, we need to look at the youth unemployment rate (the proportion of the economically active that are out of work).
- If we do that, the UK is actually one of the better performing EU nations and far away from Spain.
- www.fullfact.org check these facts and warn when claims could misinform people.

EuroStat: unemployment rates



Youth Unemployment rates



Source: zerohedge

Totals versus Rates

Youth unemployment totals (thousands) November 2012 (under 25's) (seasonally adjusted)	
Spain	959
United Kingdom	951
France	791
Italy *	610
Poland	433
Germany	364
Romania *	194
Greece	186
Portugal	175
Sweden	158
Netherlands	138
Hungary	87
Belgium	84
Slovakia	84
Czech Republic	79
Bulgaria	70
Ireland	67
Denmark	63
Finland	63
Austria	51
Latvia *	32
Lithuania	31
Slovenia	22
Estonia	13
Cyprus	12
Malta	5
Luxembourg	3

Youth unemployment rates (%) November 2012 (under 25's) (seasonally adjusted)	
Greece	59.4
Spain	55.4
Portugal	38.7
Italy	36.8
Slovakia	35.7
Latvia *	31.7
Ireland	29.9
Cyprus	28.5
Poland	28.2
Hungary	27.9
Bulgaria	27.5
Slovenia	26.9
France	26.8
Sweden	24.7
Lithuania	23.8
Romania *	23.0
Czech Republic	21.2
United Kingdom	20.7
Belgium	19.7
Estonia	19.5
Finland	18.9
Luxembourg	18.6
Malta	16.3
Denmark	14.2
Netherlands	9.7
Austria	8.6
Germany	8.0

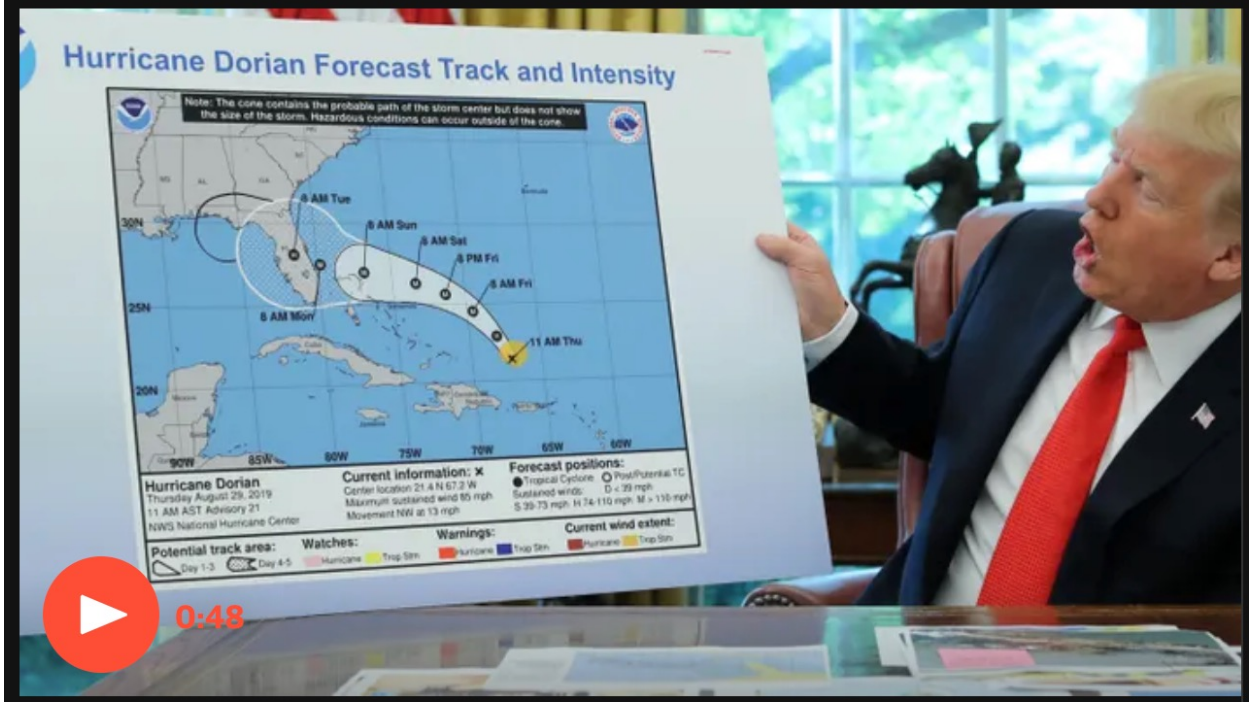
Source: www.fullfact.org

Nick Clegg: 3 million British jobs depend on the EU

- When discussion ‘Brexit’, Nick Clegg claimed that 3 million British jobs “rely directly” on EU membership.
- The 3 million figure actually refers to the number of jobs linked directly and indirectly to firms that export to the EU (including, for example, those providing logistics, legal or financial services to exporters).
- It doesn’t tell us about how many jobs – if any – would be lost if Britain exited the EU.

Sharpie-Gate

Donald Trump displays Hurricane Dorian map apparently doctored with marker pen – video



Hurricane Dorian headed to Alabama (according to Trump). This fact was denied by US National Weather Service.

https://www.theguardian.com/us-news/video/2019/sep/05/donald-trump-displays-hurricane-dorian-map-doctored-with-marker-pen-video

Lies, damned lies and statistics

- Mark Twain attributed the phrase to the 19th-century British Prime Minister Benjamin Disraeli (1804–1881):

- "There are three kinds of lies: lies, damned lies, and statistics.”

Truth and statistics

- Mistakes can easily be made during data analysis
- Numbers that suit an individual's purpose can be selected
- Data should support decisions
- Decisions often come first and data that is supportive found afterwards
- Data can also be manipulated

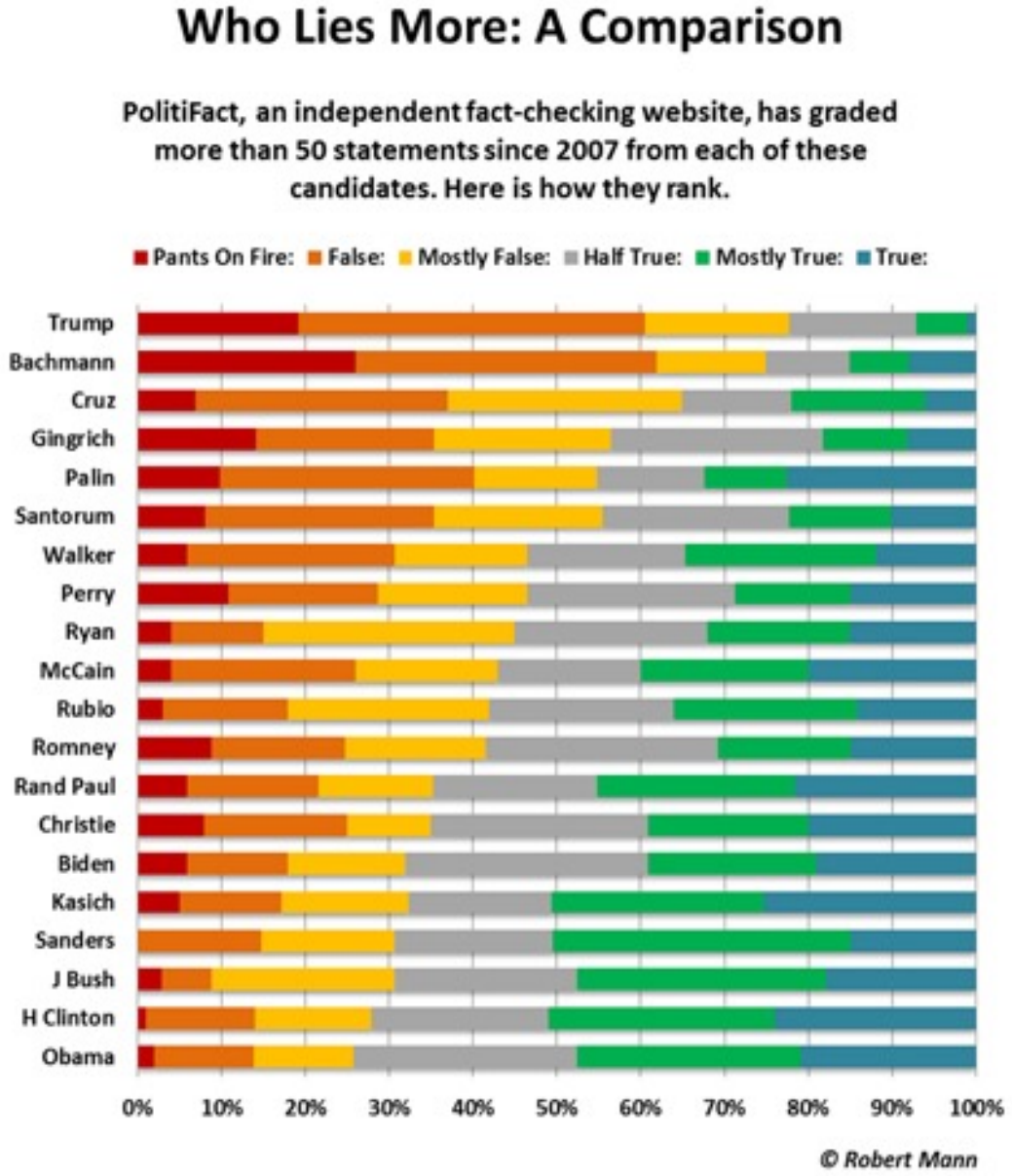
Caveman effect

- Much of our understanding of prehistoric people comes from caves.
- For example, we are familiar with cave paintings made nearly 40,000 years ago.
- Contemporary paintings on trees, animal skins or hillsides would likely have been washed away long ago.
- Similarly, evidence of fire pits and burial sites are most likely to remain intact to the modern era in caves.
- Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives.

President Roosevelt Election

- A sampling frame error occurs when the wrong sub-population is used to select a sample.
- A classic frame error occurred in the 1936 presidential election between Roosevelt and Landon.
- The sample frame was from car registrations and telephone directories.
- In 1936, many Americans did not own cars or telephones and those who did were largely Republicans.
- The results wrongly predicted a Republican victory.

US Politicians: “alternative facts”



GapMinder

- https://www.gapminder.org/videos/200-years-that-changed-the-world/

- https://www.youtube.com/watch?v=jbkSRLYSojo

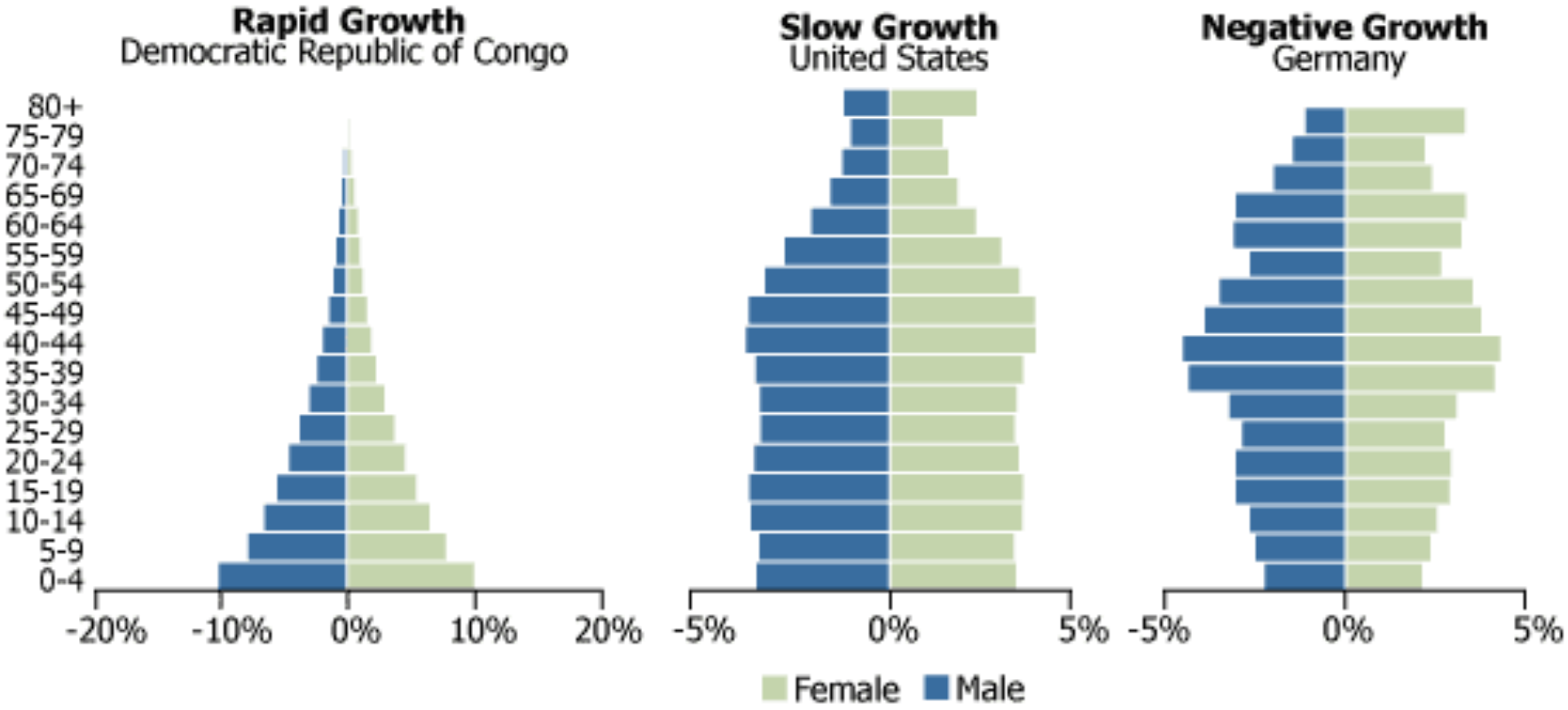
Wealth and Health

- “Your health is your wealth”
- “The greatest wealth is health” – Virgil
- www.gapminder.org/world

Summarizing Data

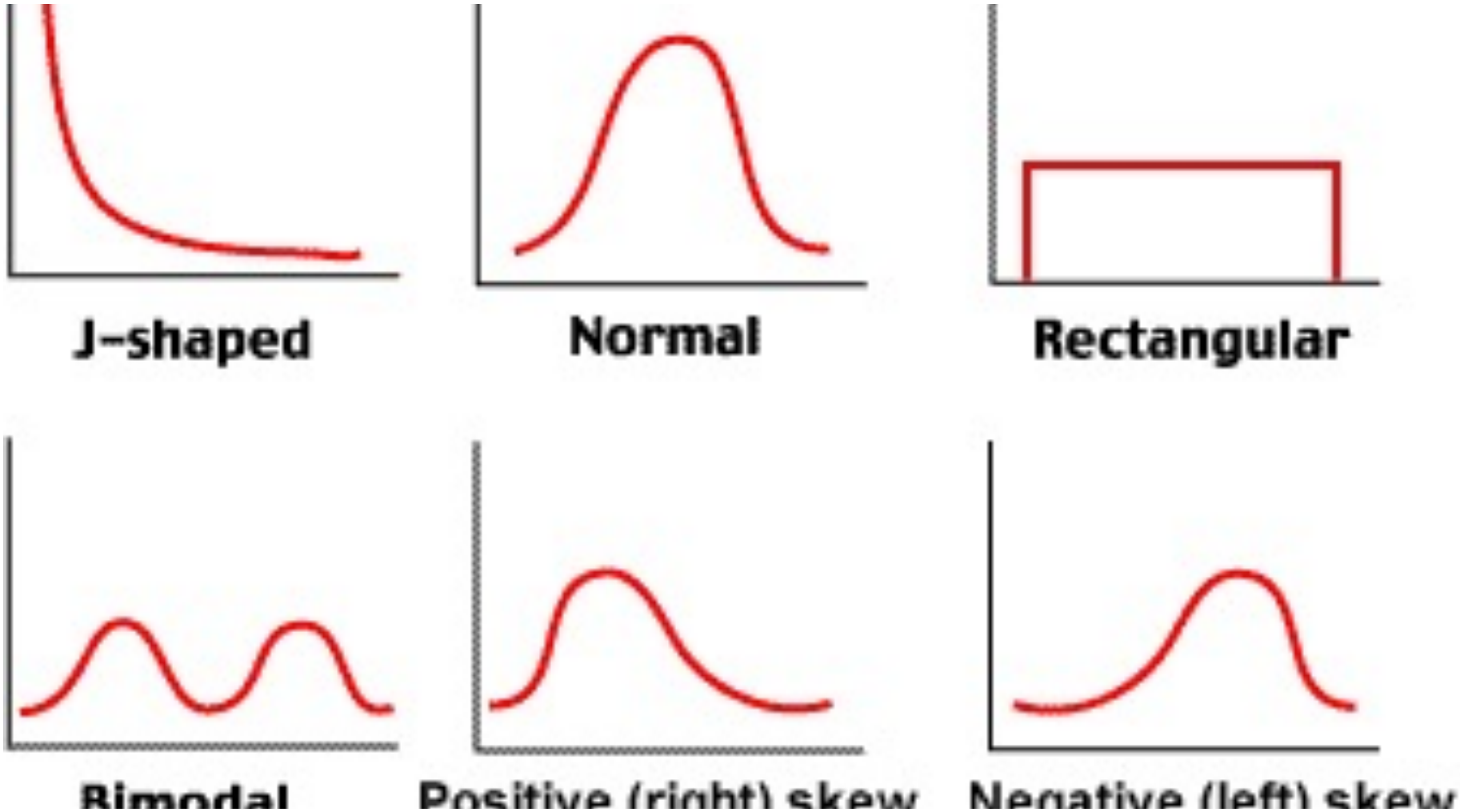
- Statistics relates to collecting reliable numerical data and then analyzing and interpreting them
- For discrete data, the first step is to measure the frequency with which each value occurs and construct a histogram
- For continuous data, bins are formed using ranges and this allows us to measure the frequency of values falling in each bin

Age Histograms



Source: UN World Population Prospects, The 2006 Revision.

Types of distributions



Theoretical Distributions

- Normal distribution - heights
- Log-normal distribution – financial returns
- Poisson distribution – waiting times
- Student's t-distribution – hypothesis testing

Normal distributions

- Normally distributed random variable with mean μ and variance σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- All higher order moments are given in terms of μ and variance σ^2:
- Easily manipulated:
  - if $x \sim N(0, \sigma_x^2)$ and $y \sim N(0, \sigma_y^2)$, then $x + y \sim N(0, \sigma_x^2 + \sigma_y^2)$
- Central limit theorem: sum of a large number of IID random variables (with finite mean and variance) is normally distributed
- For linear models:
  - Normal distributions are preserved by principle of superposition
  - Normally distributed forecast errors: Maximum likelihood gives least squares
  - Useful for calculating prediction intervals
- Problems for nonlinear systems:
  - Use of normal distributions neglects possibility of asymmetric distributions
  - Fat tailed distributions imply larger probability of worse case scenarios (risk management)

Normal distributions II

- Distributions that are normal or Gaussian have the following characteristics:
  - Approximately 68% of the values fall between the mean and one standard deviation (in either direction)
  - Approximately 95% of the values fall between the mean and two standard deviations (in either direction)
  - Approximately 99.9% of the values fall between the mean and three standard deviations (in either direction)

Human Body Temperature

- Having an idea of what is “normal” is useful for identifying outliers and issuing alerts

Description	Temperature Range
Hypothermia	< 35.0
Normal	36.5 – 37.5
Fever	37.5 – 38.3
Hyperthermia	37.5 – 38.3
Hyperpyrexia	> 40.0 – 41.5

- A temperature of over 37.5C (99.5F) is classified as a fever [Source: NHS, UK]
- Ebola screening: 37.5 ^{o}C

History of normal temperature

- ‘Normal' benchmark for body temperature was established by a 19th century German physician called Dr Carl Wunderlich
- The NHS says that a normal temperature is around 37C (98.6F), although it depends on:
  - The person
  - Their age
  - What they've been doing
  - The time of day
  - Which part of the body you take the temperature

Spanish influenza 1918 (Poll 3)

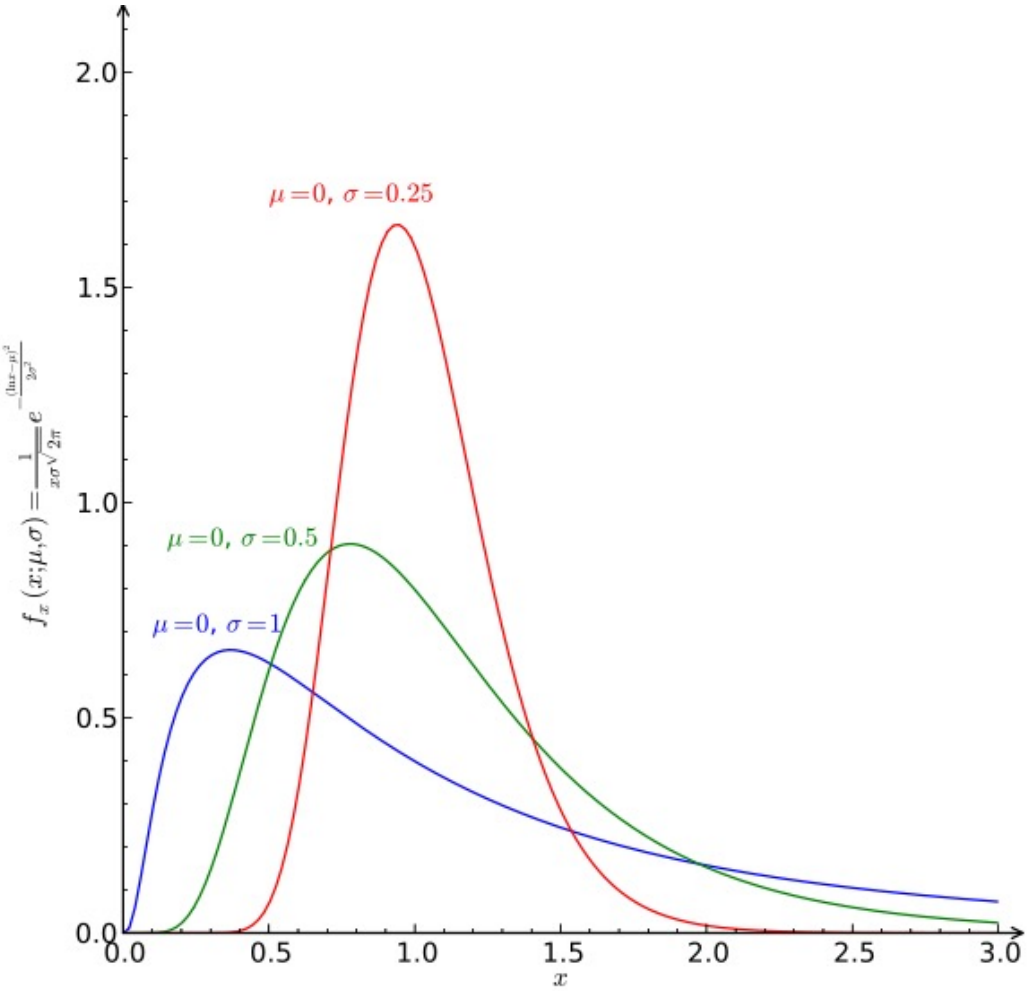
- It is estimated that about 500 million people or one-third of the world's population became infected.
- What was the death rate?

Slido.com
#78793

Distribution Constraints

- Many phenomena (such as wind speed, rainfall, price, …) cannot be negative and therefore an appropriate distribution must be chosen
- Both lower and upper bounds may exist
- Models that take account of these constraints tend to perform better

Log normal distribution



Measures of location

- Mean: simple average of measurements

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Median: middle value obtained by first placing the measurements in ascending order of magnitude
- Mode: value which occurs with the greatest frequency

Measures of central tendency

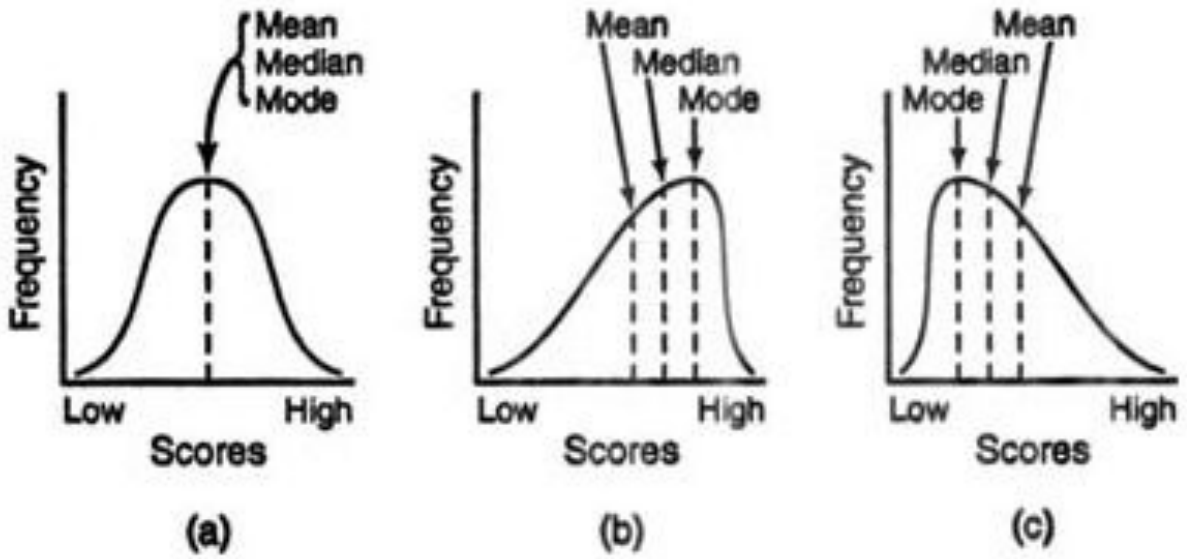
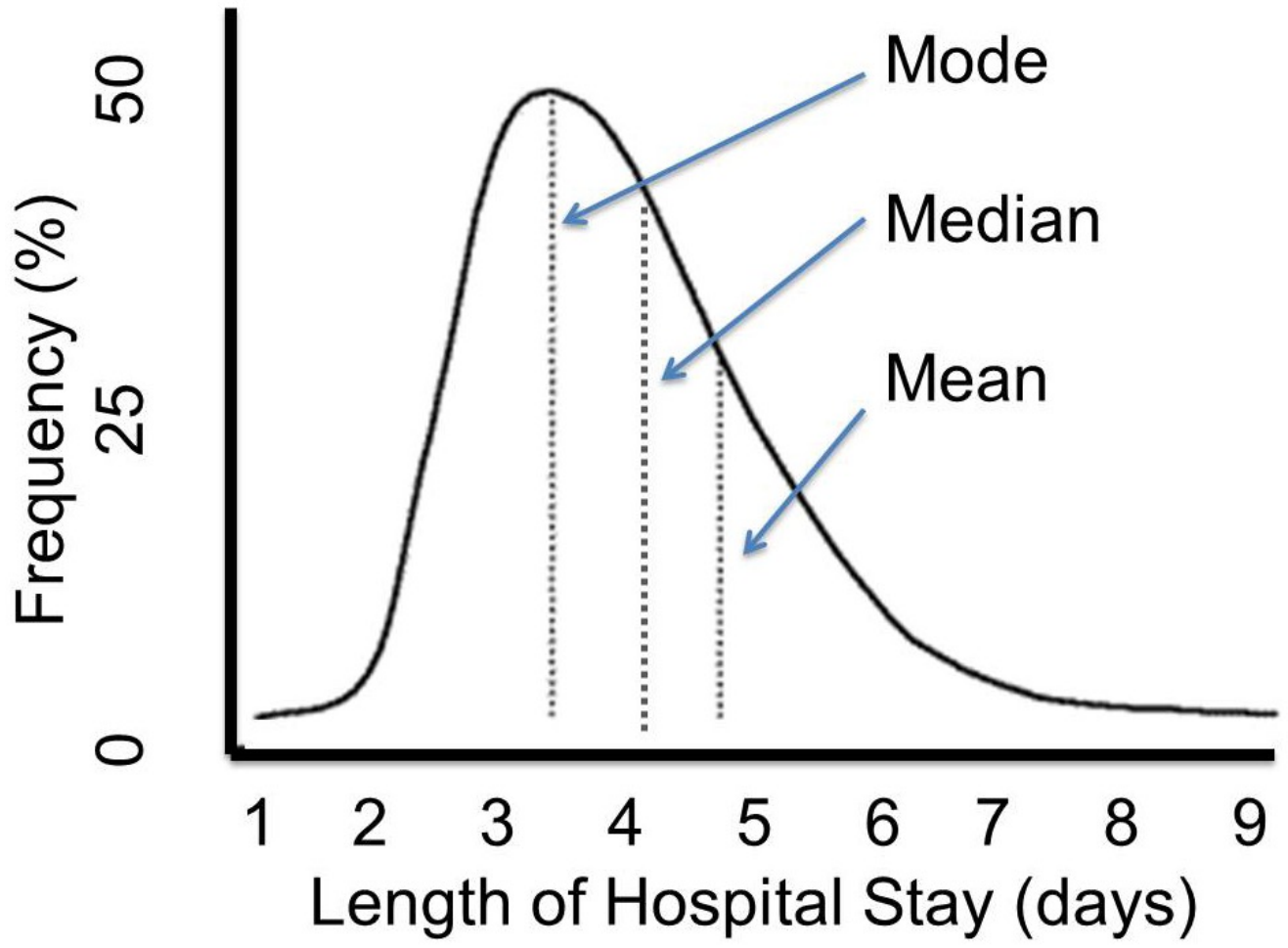


Figure 3
Measures of Central Tendency

Skewed distribution



Measures of variability

- Standard deviation
- Variance
- Inter-quartile range (middle 50%)
- 5% and 95%
- 2.5% and 97.5%

Descriptive Statistics

- Centrality: Mean, median, mode;
- Variability, standard deviation and variance;
- Ranking and ordering data
- Statistics from imperfect data
- Outputting statistical summary tables

Matlab functions

- mean, median, mode
- std, var
- sort, prctile
- nansum, nanmean, nanstd
- disp, fprintf
- sprint, blotter

Data, Inference & Applied Machine Learning

WEEK 3B

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Making data visually accessible	10
2	Discussion	How to visualize data	10
3	Case study	Dashboards	10
4	Analysis	Visualization techniques	20
5	Matlab Demo	Statistical distributions	20
6	Q&A	Matlab questions and feedback	10

Making data visually accessible

- Select sensible units so that numbers are relatively small (in units of ones, tens or hundreds).
- Small numbers are easy to visualize.
- Use ratios if normalization is appropriate.
- Use rates if there is a need to communicate the speed of change.

Why make data visual?

- To communicate information clearly and efficiently to users via the information graphics selected.
- To stimulate viewer engagement and attention.
- To help users in analyzing and reasoning about data and evidence.
- To make complex data more accessible, understandable and usable.

Class Poll 1

- Where would you like to work?
- 1. Government
- 2. Non-governmental Organization (NGO)
- 3. Foundation
- 4. Company
- 5. Start-up

Slido.com
#96496

Data visualization

- Data visualization is both an art and a science
- Tables allow users to look-up a specific measure of a variable.
- Charts: show patterns or relationships in the data for one or more variables.
- Spatiotemporal data is the most challenging

Dashboard

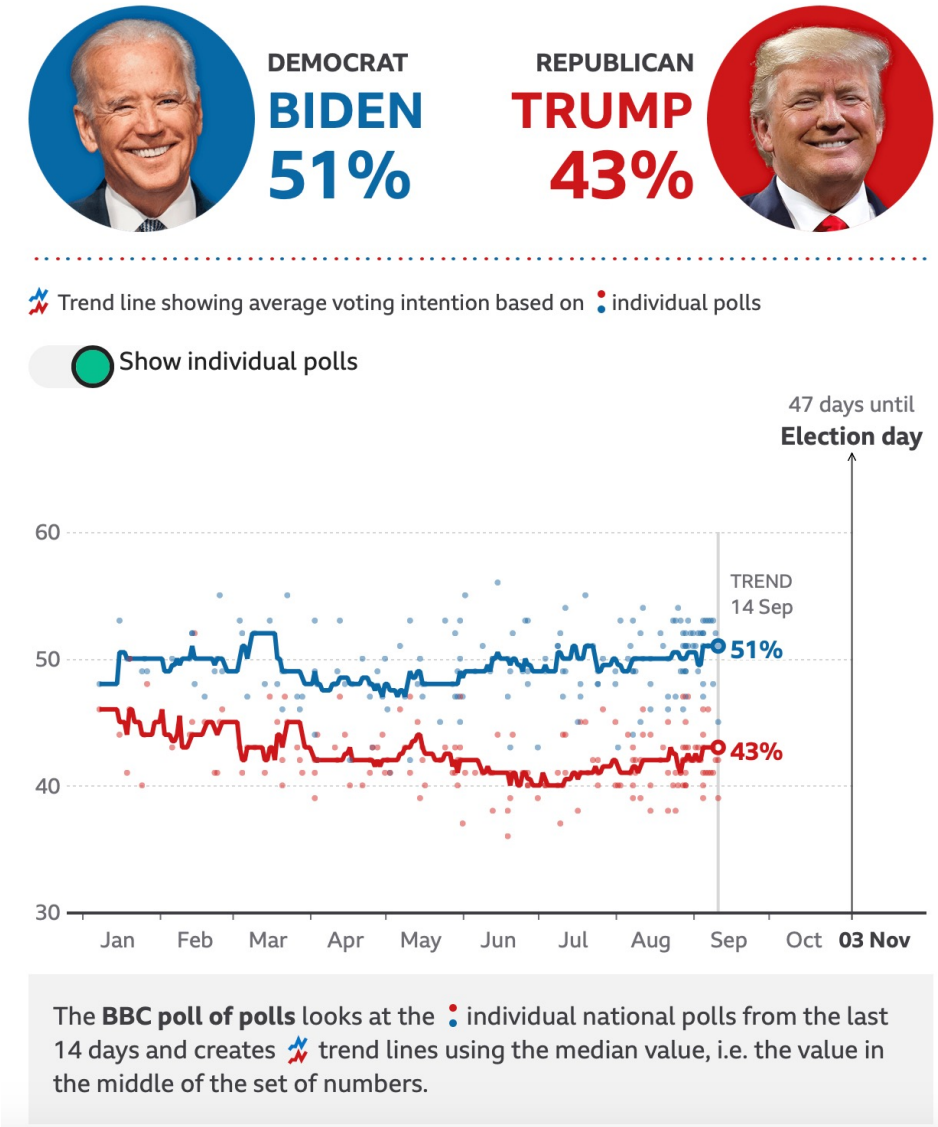
- A dashboard is a screen that displays information.
- A well designed dashboard provides the user with an overview, offering access to the most important data, insights, functions and controls.
- It can help the user to be more efficient and facilitate decision-making.

Class Poll 2

- Who do you think will win the US 2024 Presidential Election?
- Biden or Trump

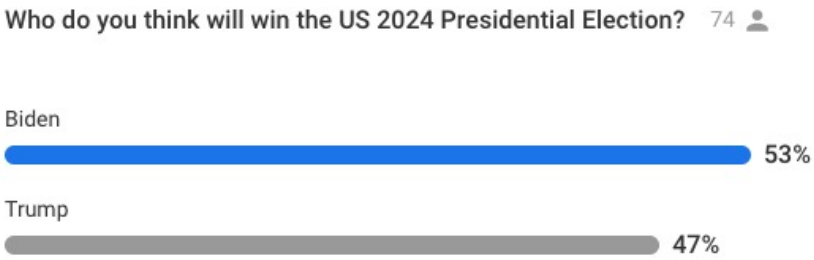
Slido.com
#96496

Election Polls



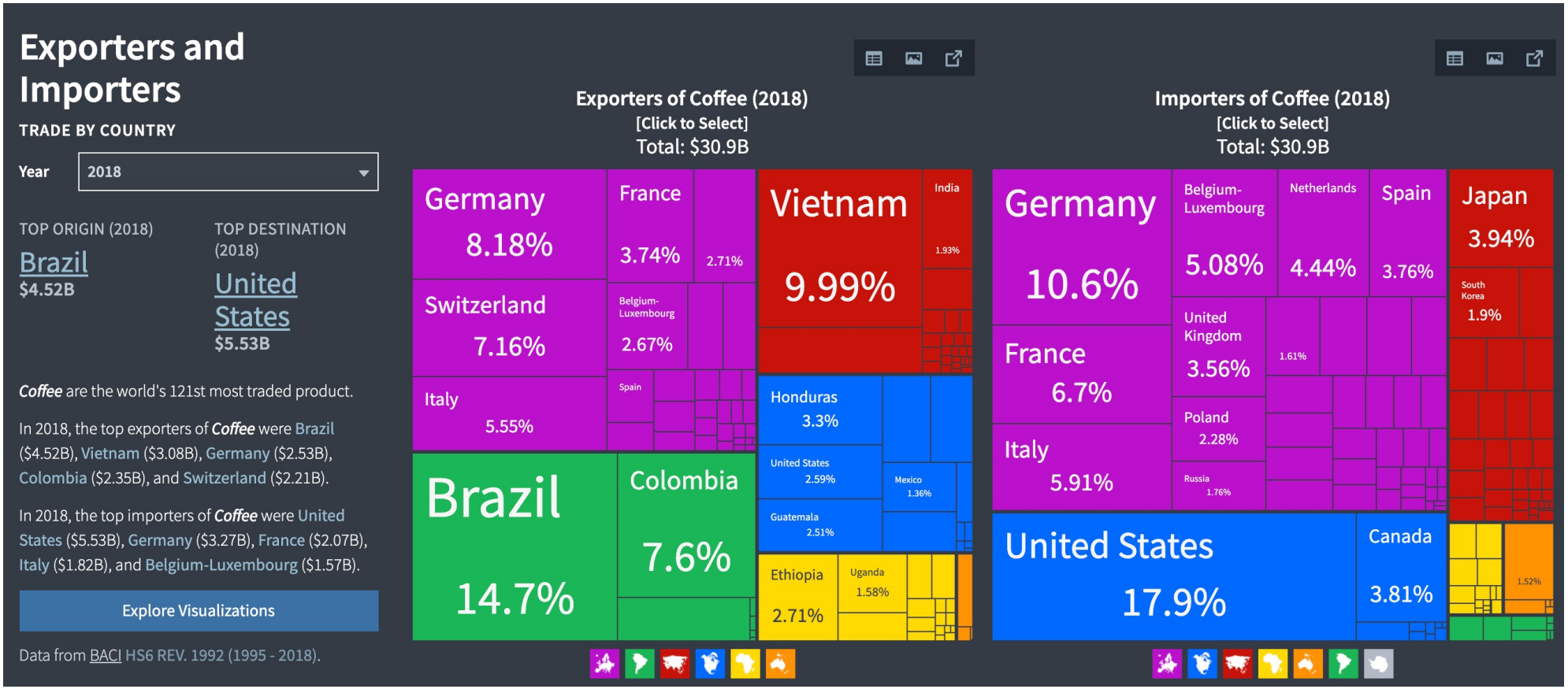
National polls are a good guide as to how popular a candidate is across the country as a whole, but they're not necessarily a good way to predict the result of the election.

In 2016, for example, Hillary Clinton led in the polls and won nearly three million more votes than Donald Trump, but she still lost - that's because the US uses an electoral college system, so winning the most votes doesn't always win you the election.



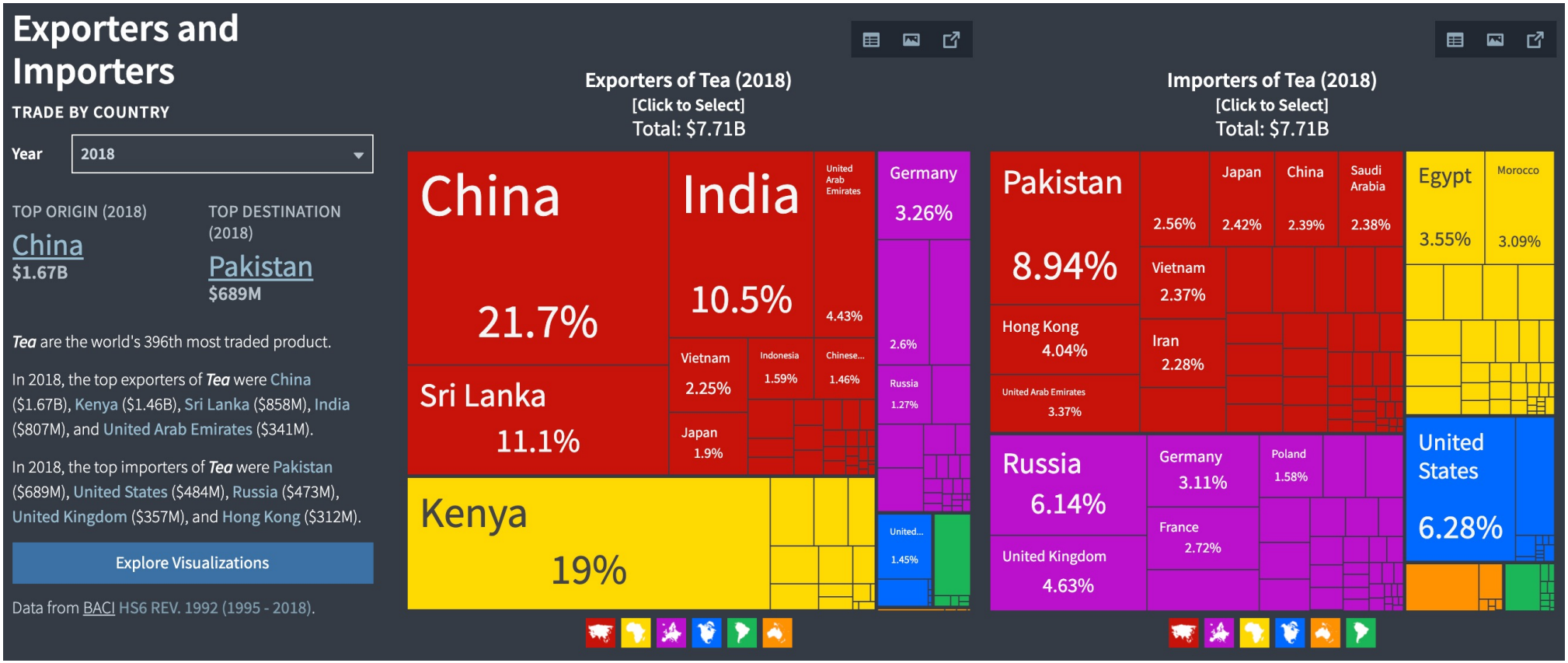
https://www.bbc.co.uk

Coffee Trade



https://oec.world/en/profile/hs92/coffee

Tea Trade



https://oec.world/en/profile/hs92/tea

COVID-19 Dashboard

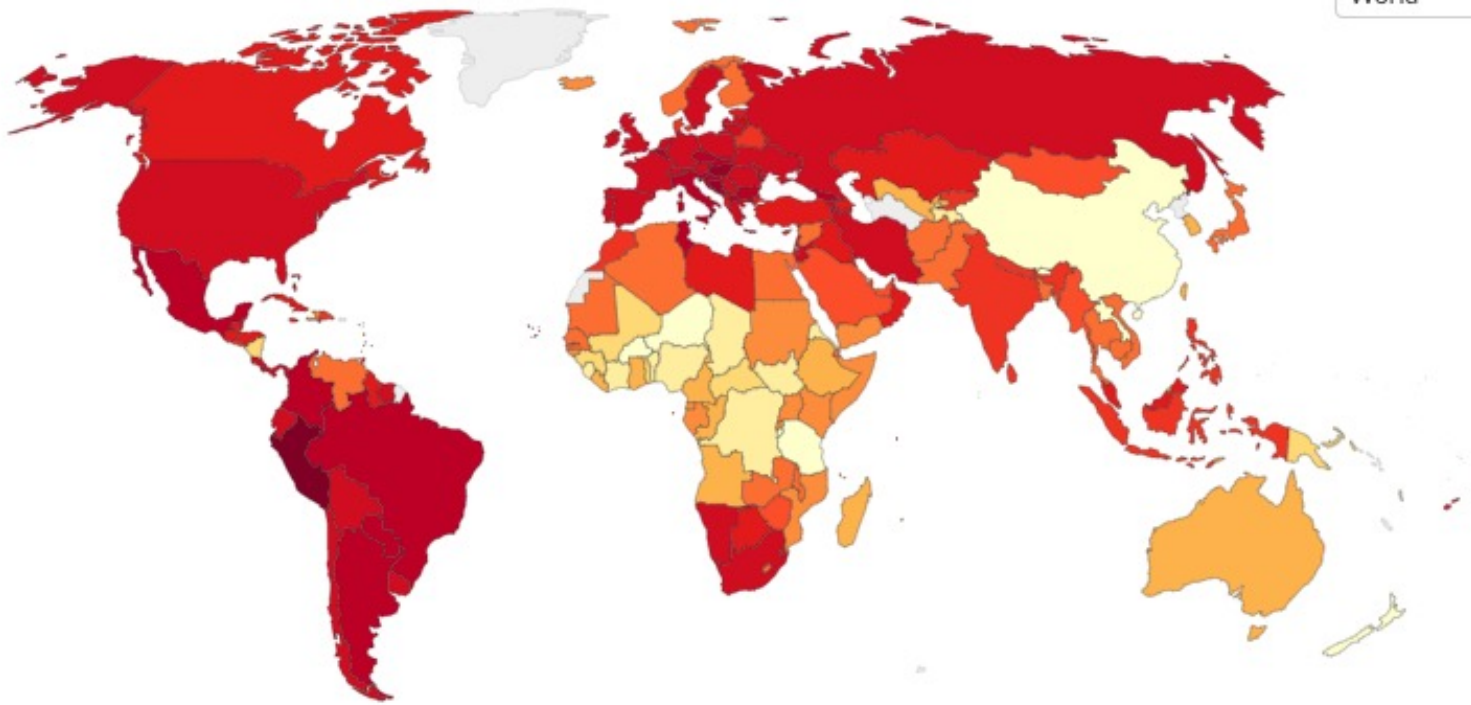
https://coronavirus.jhu.edu/map.html

Cumulative confirmed COVID-19 deaths per million people, Sep 8, 2021

Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19.

Our World in Data

World



No data 0 10 20 30 50 100 200 300 500 1,000 2,000 3,000 5,000 10,000

Source: Johns Hopkins University CSSE COVID-19 Data

Jan 22, 2020 — Sep 8, 2021

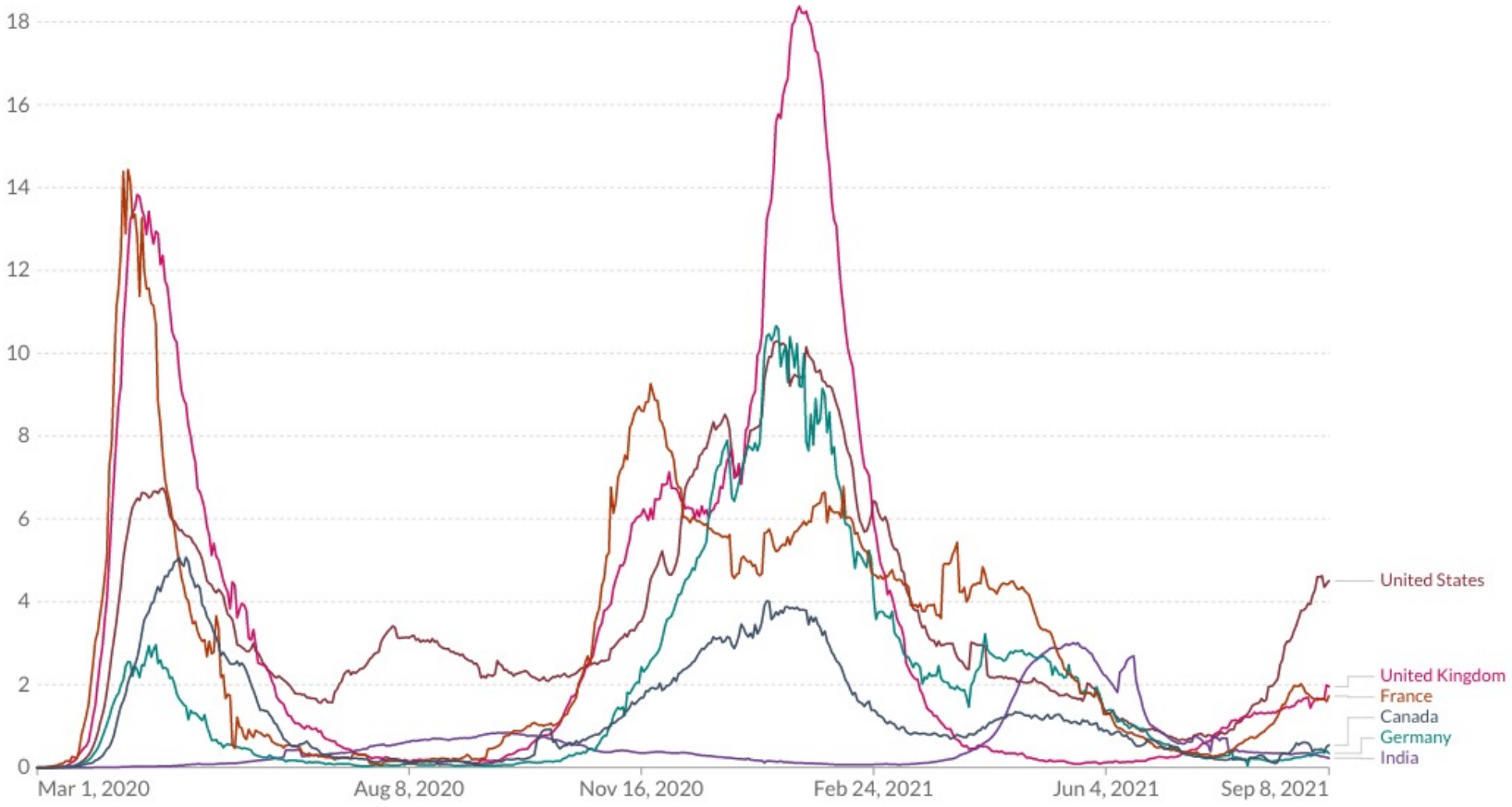
CHART | MAP | TABLE | SOURCES | DOWNLOAD

Daily new confirmed COVID-19 deaths per million people

Shown is the rolling 7-day average. Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19.

Our World in Data

LINEAR LOG



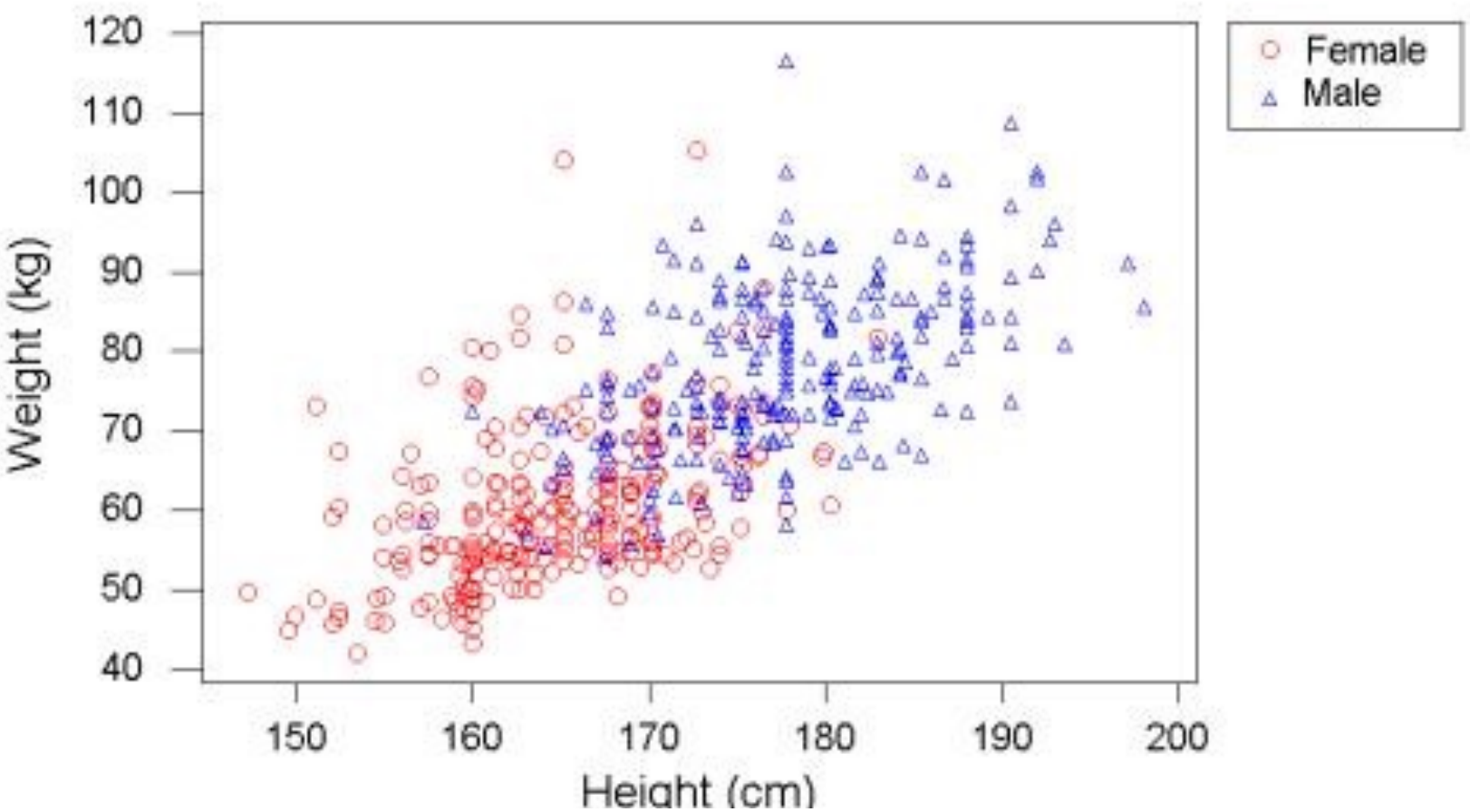
Source: Johns Hopkins University CSSE COVID-19 Data

Jan 28, 2020 — Sep 8, 2021

Dashboard Assignment

- Observatory of Economic Complexity (OEC): Use GeoMaps to find the largest exporters and importers of cars, computers, etc.
- OurWorldinData: Explore COVID for your country
- Pantheon is a project that uses biographical data to expose patterns of human collective memory.
- Assignment:
  - Study differences between TreeMaps, Matrices, Scatterplots, Charts and Maps.
  - Explore functionality: time slides, country selection, linear/log, smoothing with moving averages, filters

Weight versus height

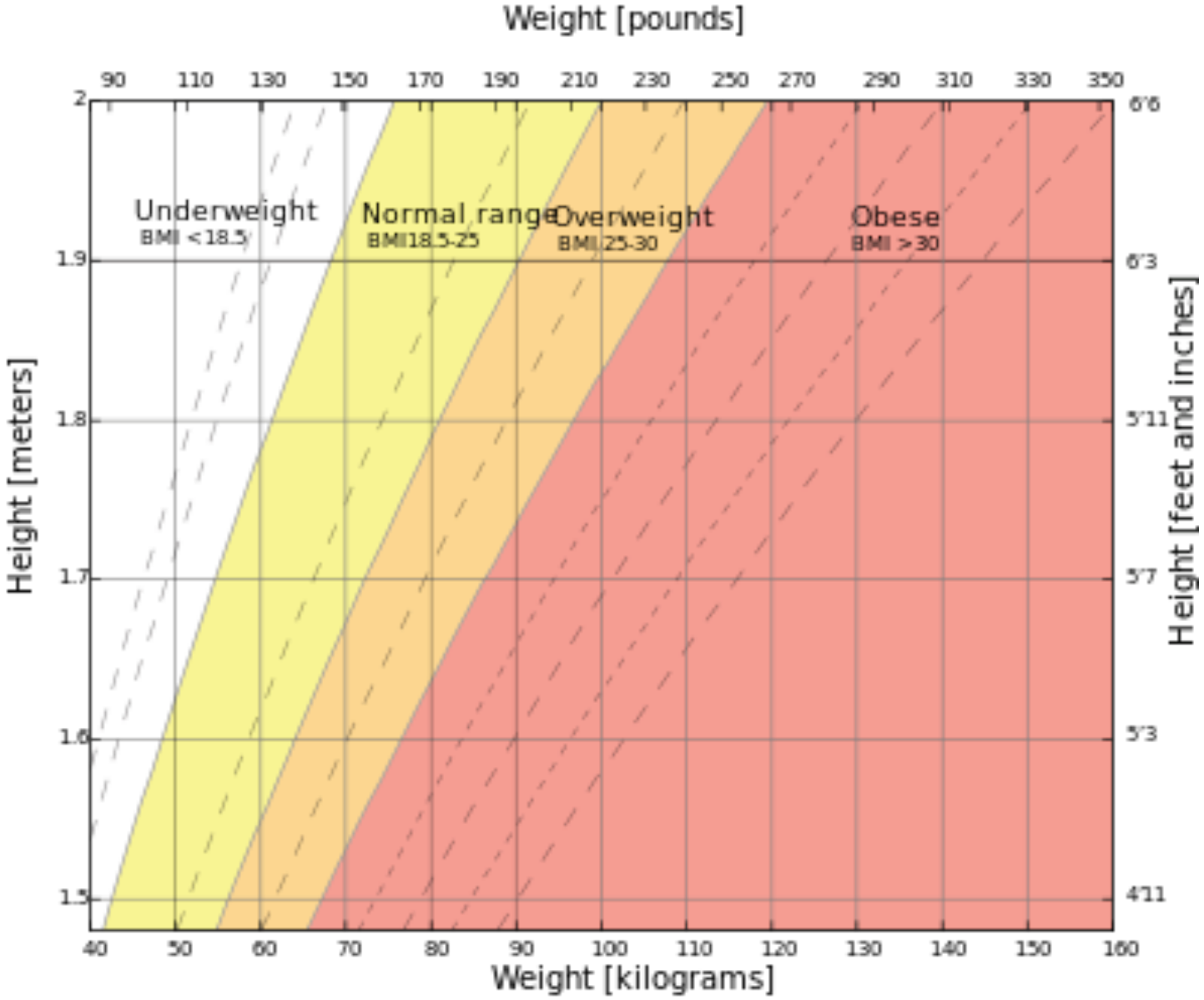


Heinz et al., 2003. Exploring relationships in body dimensions.

Body Mass Index

- Body Mass Index (BMI) is an individual's body mass divided by the square of their height [units of kg/m^2]:
- BMI = Mass / $Height^2$
- BMI is a reliable indicator of body fatness for most people and is used to screen for weight categories that may lead to health problems
- Devised in early 1800s by the Belgian polymath Adolphe Quetelet during the course of developing "social physics"

BMI graph


Weight [pounds]
90 110 130 150 170 190 210 230 250 270 290 310 330 350
Underweight
BMI <18.5
Normal range
BMI 18.5-25
Overweight
BMI 25-30
Obese
BMI >30
Height [meters]
2 1.9 1.8 1.7 1.6 1.5
Height [feet and inches]
6'6 6'3 5'11 5'7 5'3 4'11
40 50 60 70 80 90 100 110 120 130 140 150 160
Weight [kilograms]

A Taxonomy

Chart

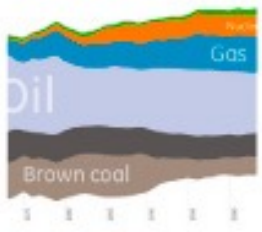
Quantities,
Distributions,
Correlations

Map

Geography

Network

Interconnections

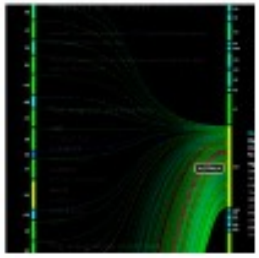


Time Series

Temporality

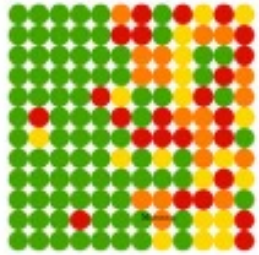
Hierarchy

Tree Structure



Flow

Movement and
Exchange



Matrix

Composition

Infographic

Explanation and
Communication

visualizing.org

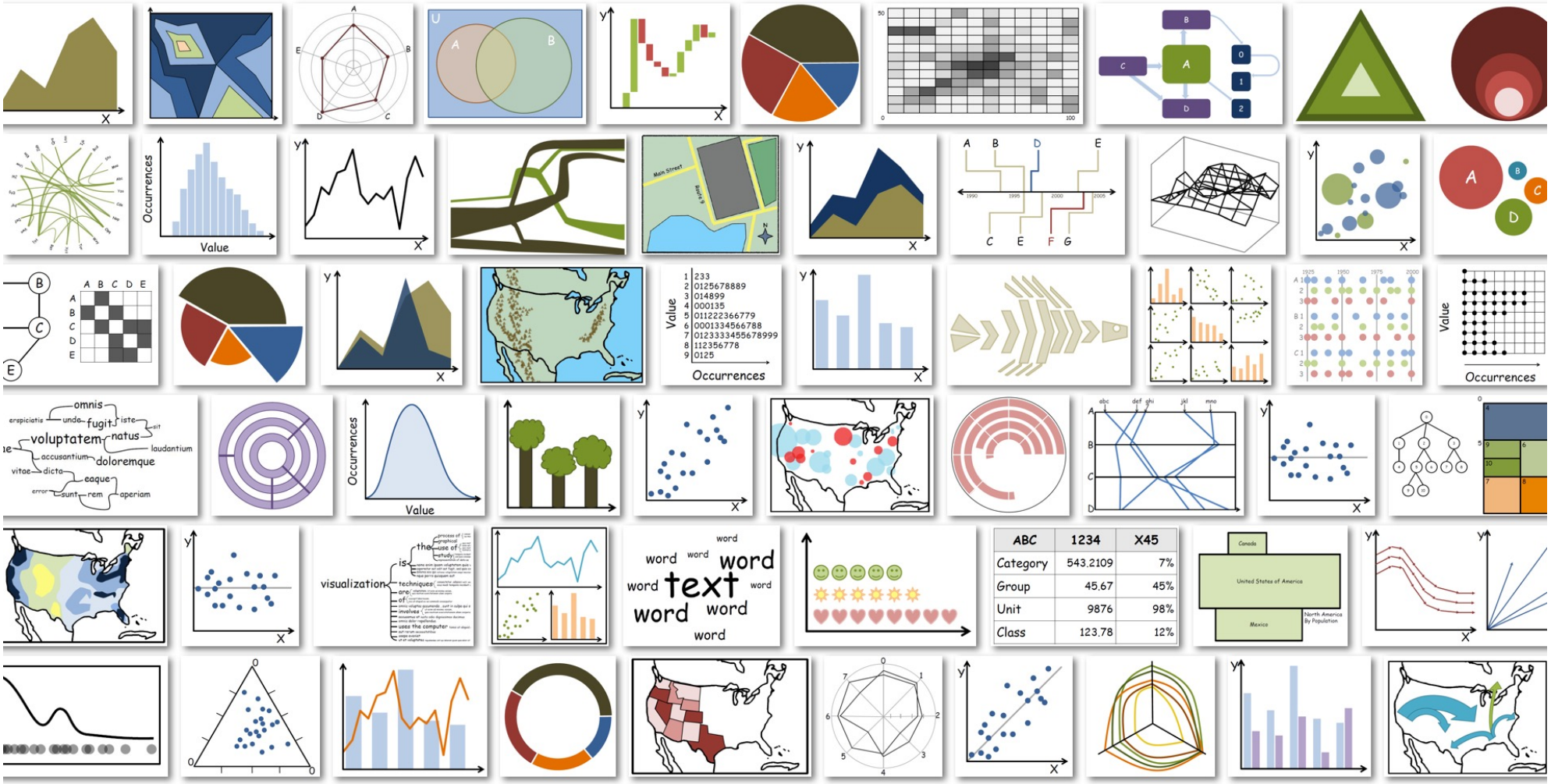
The Visual Display of Quantitative Information

- In his 1983 book The Visual Display of Quantitative Information, Edward Tufte defines 'graphical displays' and principles for effective graphical display in the following passage:
- "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency”

Graphical displays should:

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Types of visualizations



Source: Michelle Borkin, Harvard SEAS.

Discussion on Visualization

- Storytelling
- Writing reports
- Where do graphics fit in?
- "A picture is worth a thousand words"
- Where data meets design and creativity
- What are the key components of good visualization?

Class Poll 3

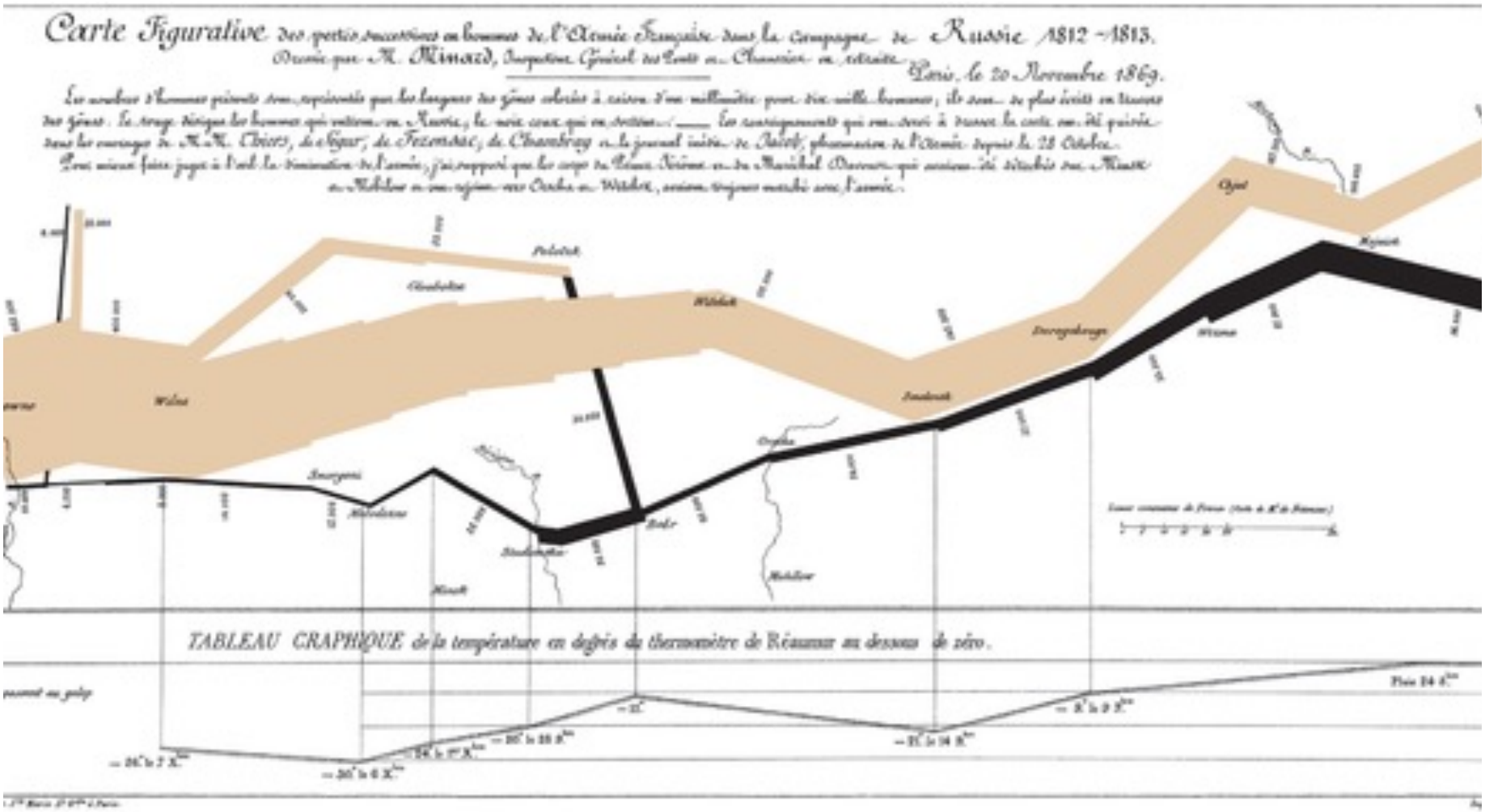
- What does “infographic” make you think of?
- Please type one word.

Slido.com
#96496

Infographic

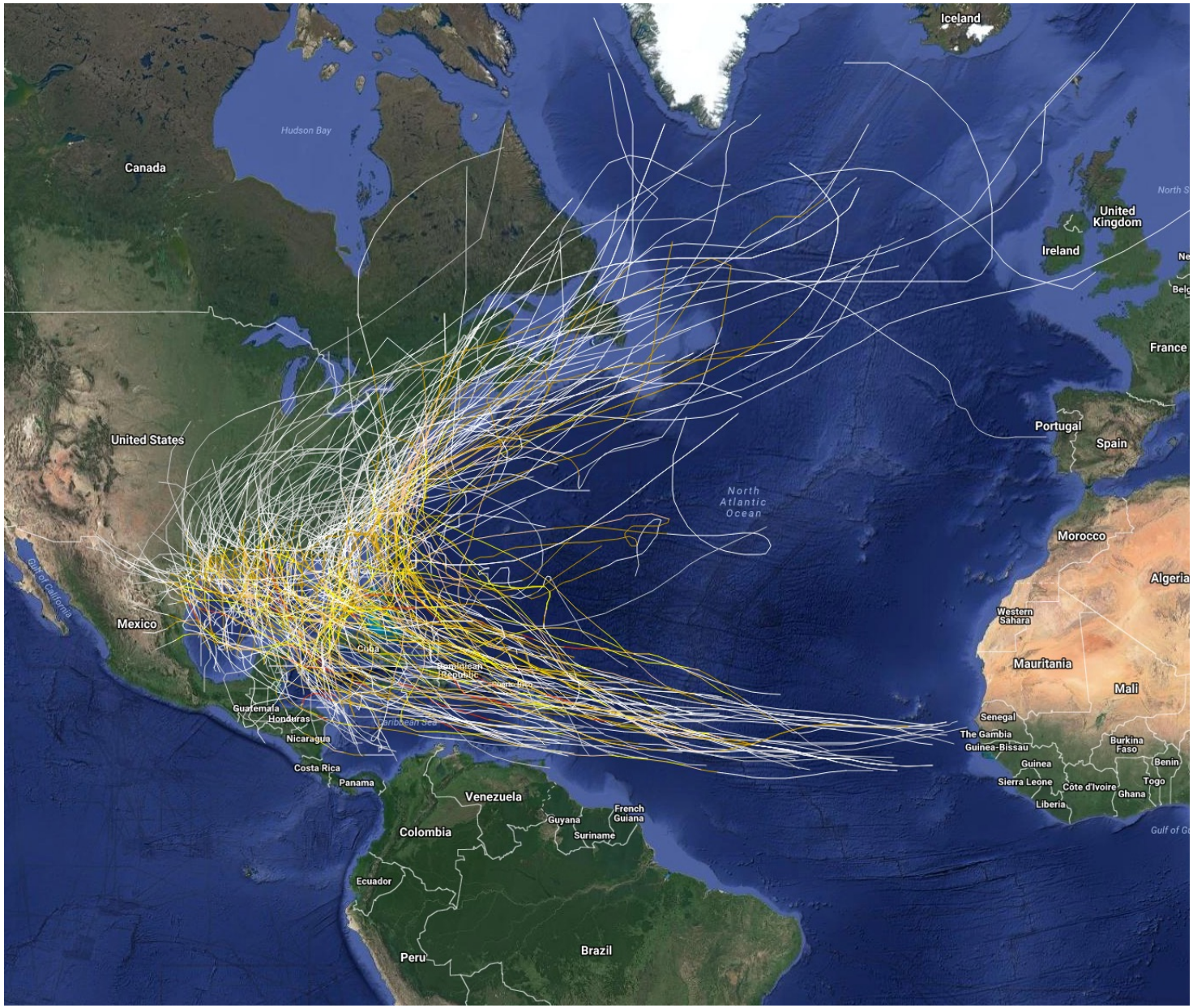
- A visual representation of information or data, e.g. as a chart or diagram:
- “a good infographic is worth a thousand words”
- Origin: 1960s (as adjective): blend of information and graphic.

Oldest Infographic

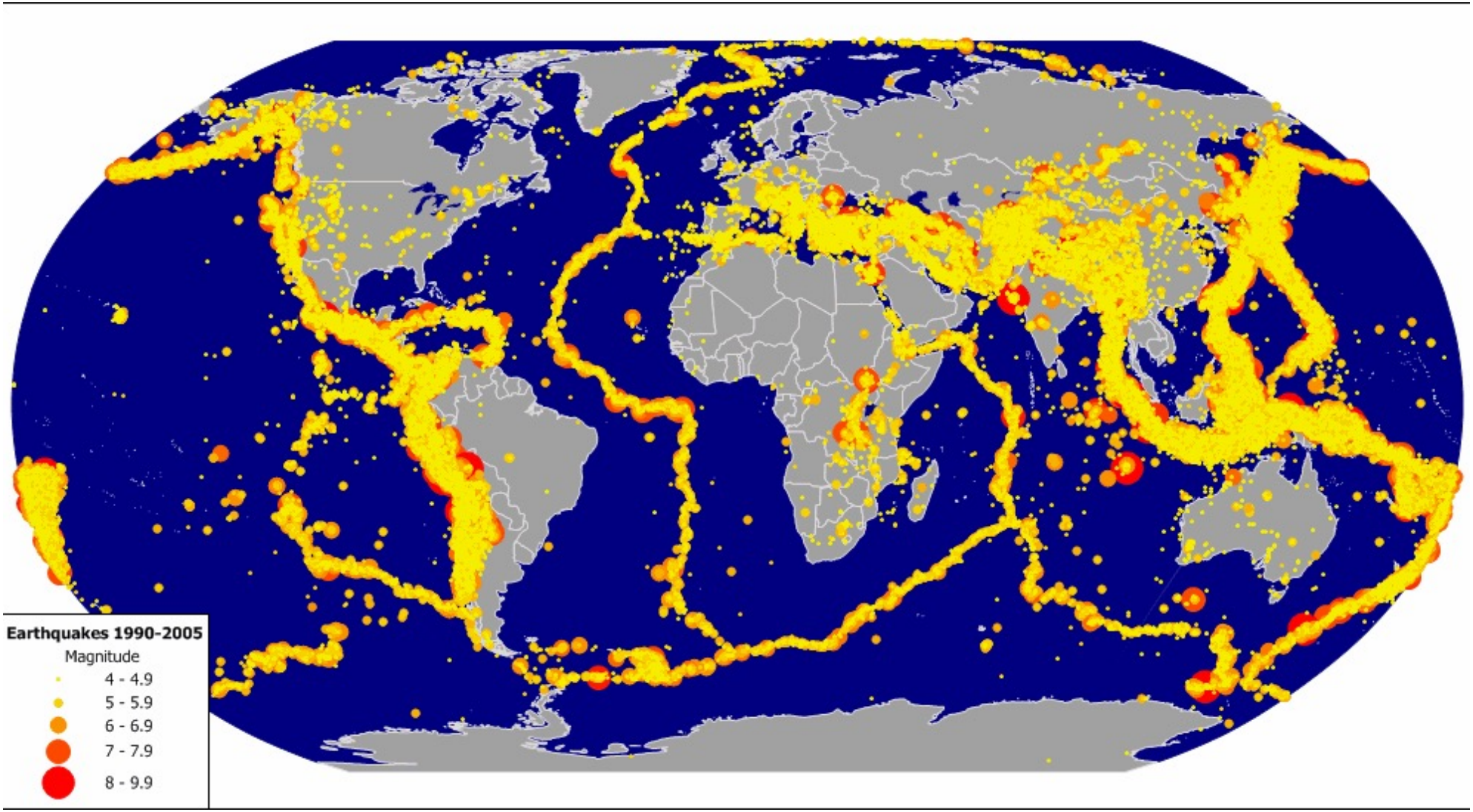


Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates.

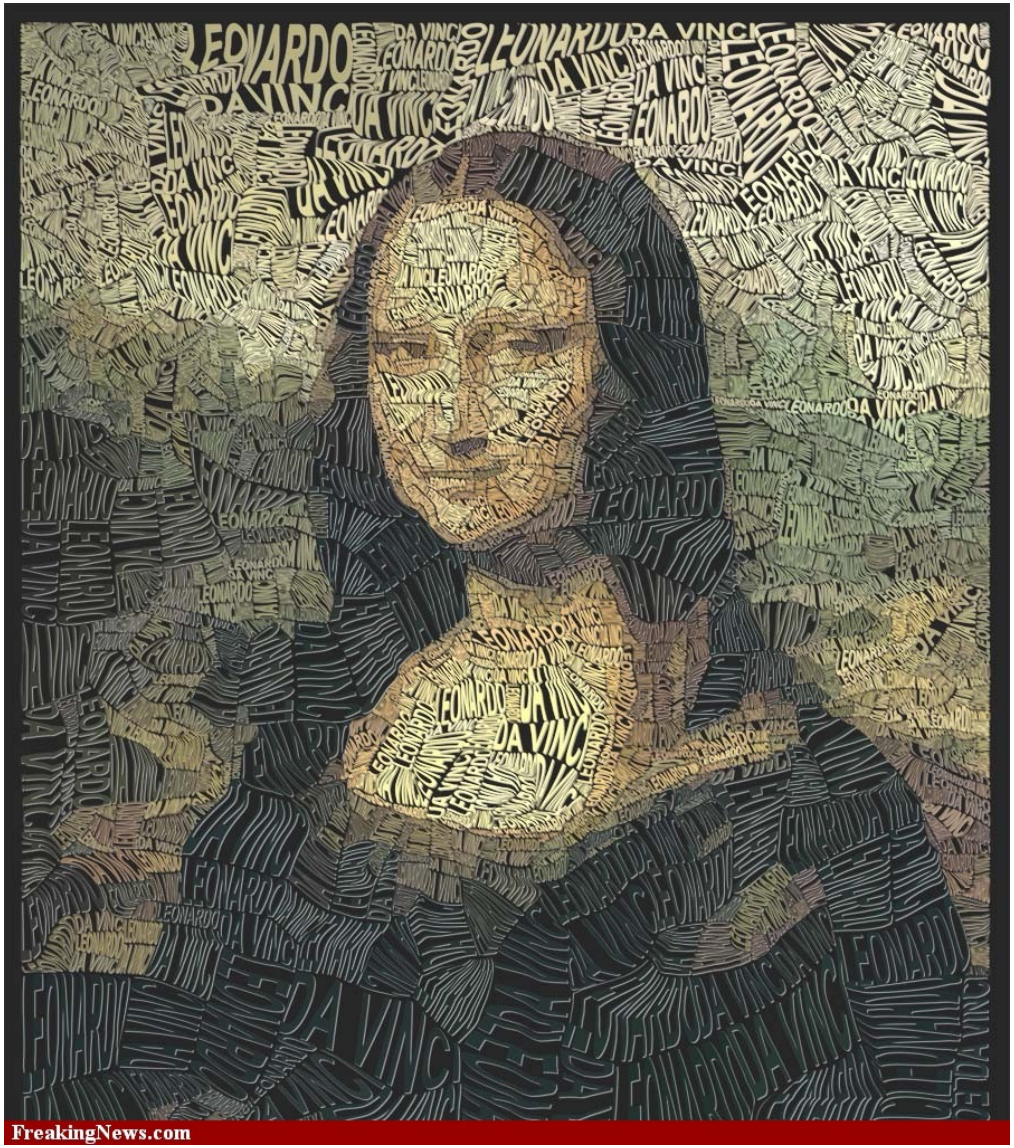
Hurricane Tracks



Earthquake Visualization



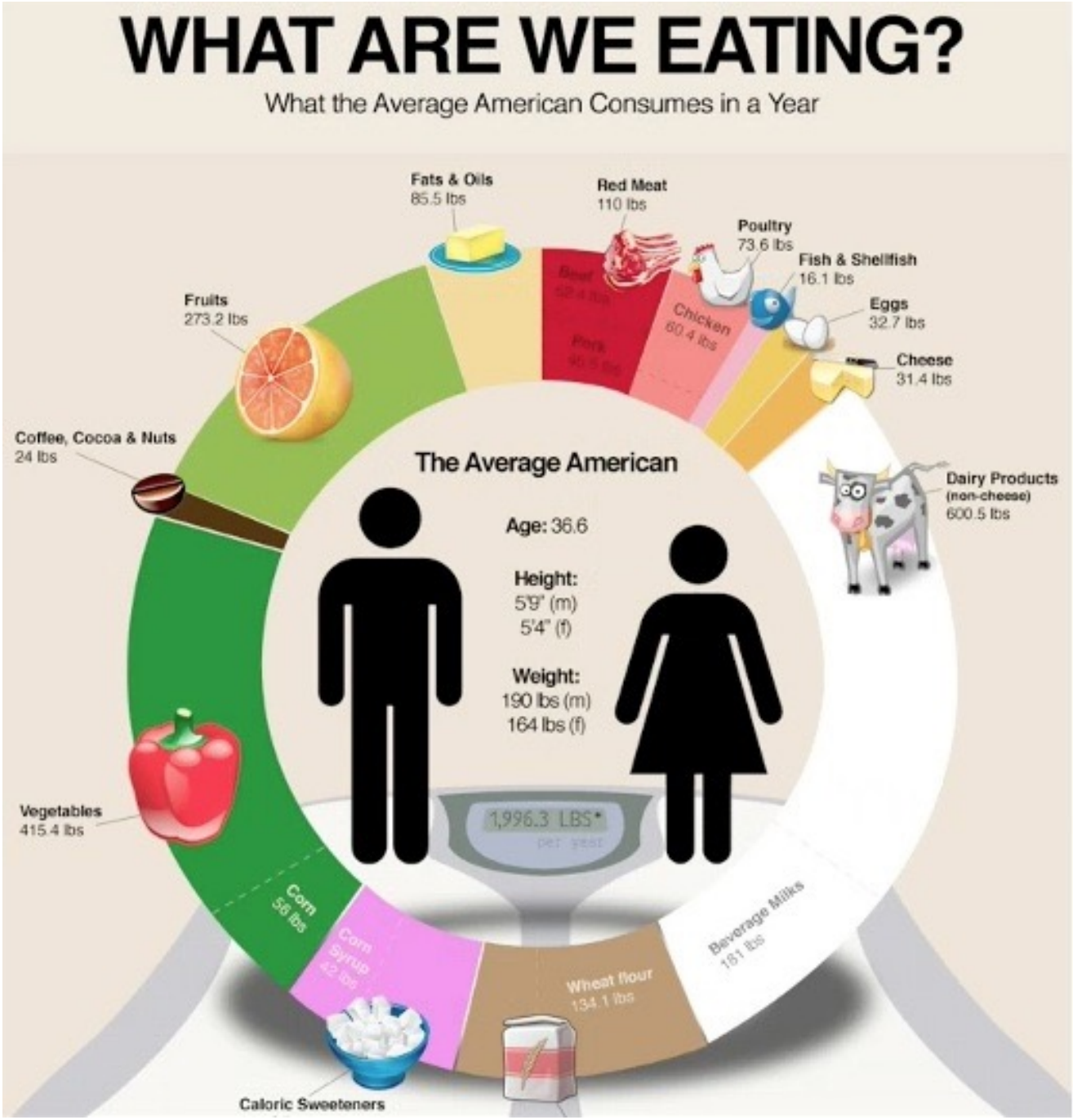
A picture from words



Worst Dinner Guest

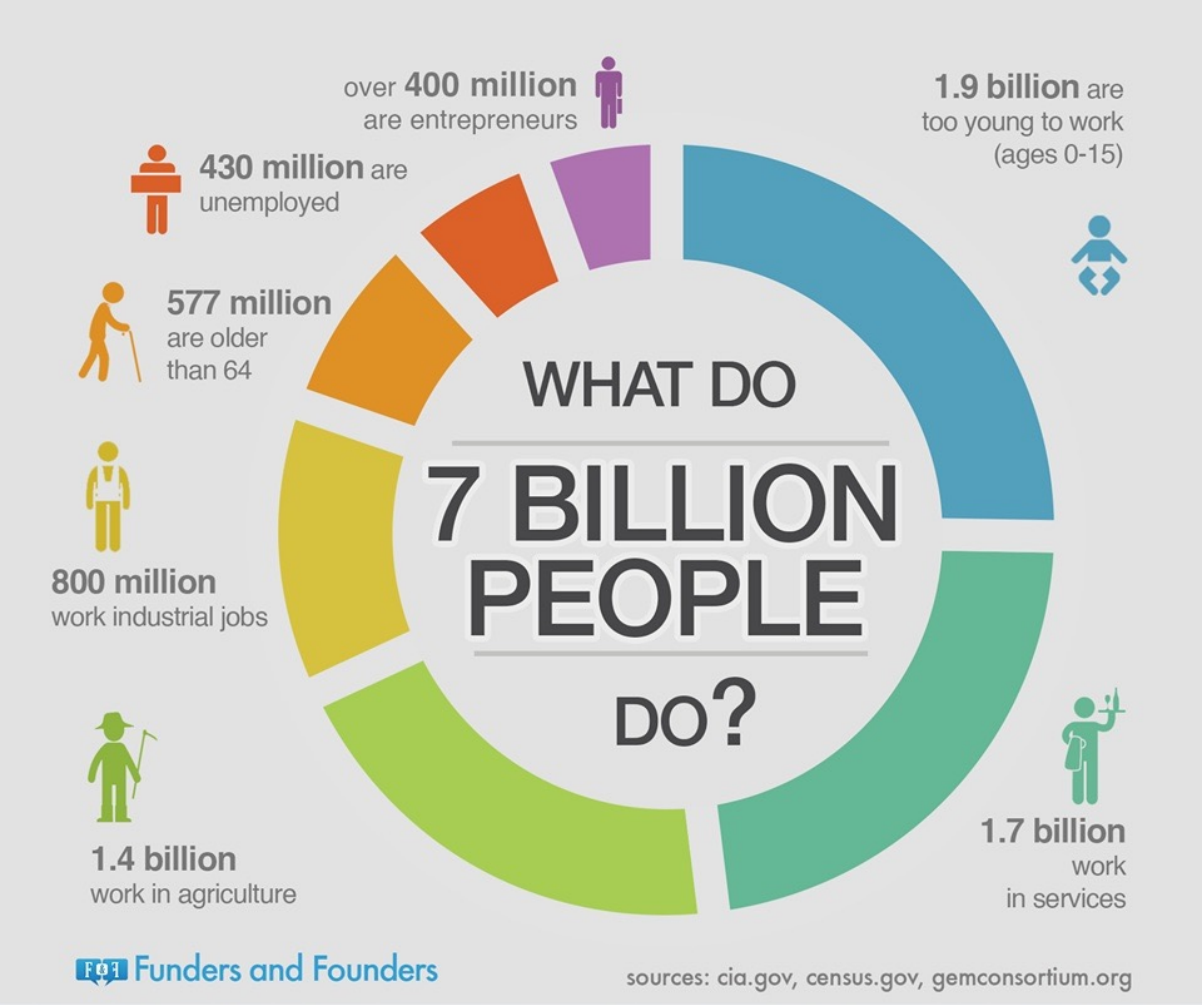
Wordle


Data
Inference
analysis
techniques
model
course
statistical
research
project
models
design
assumptions
using
real-world
quantifying
information
practical
empirical
utilize
business
students
testing
undertaking
visualization
cleaning
skills
confidence
question
Statistical
form
collection
descriptive
learn
textarea
hypothesis
Participants
errors
datasets
Matlab
assigned
involve
input
details
discussing
distributed
Select
averages
relationships
Computer
likely
Linear
Produce
investigating
Mathematics
Learning
support
identify
hands-on
discovering
practice
strong
discipline
high-level
display:none
available
statistically significant
avoid
submit
better
Course
interpreting
manipulation
Outcomes
decision
plan
understanding
Programming
measurement
summary
applications
expertise
useful
statistics
implement
completing
publicly
given
primarily
visual
objective
quality
give
series
Estimate
Visualize
overview
type
seek
text
construct
test
computational
assignments
action
description
Statistics
parameters
evaluation
autoregression
optimal
Identify
types
key
answer
highlighting
aspects
name
mistakes
Probability
throughout
exploratory
outcomes
structured
Load
POST
enable
demonstrations
advantages
common
selection
decision-making
missing
supporting
characteristics
diagnostic
Measurement
goal
Physics
Time
Design
modeling
Decide
regression
work
moving
collecting
method
example

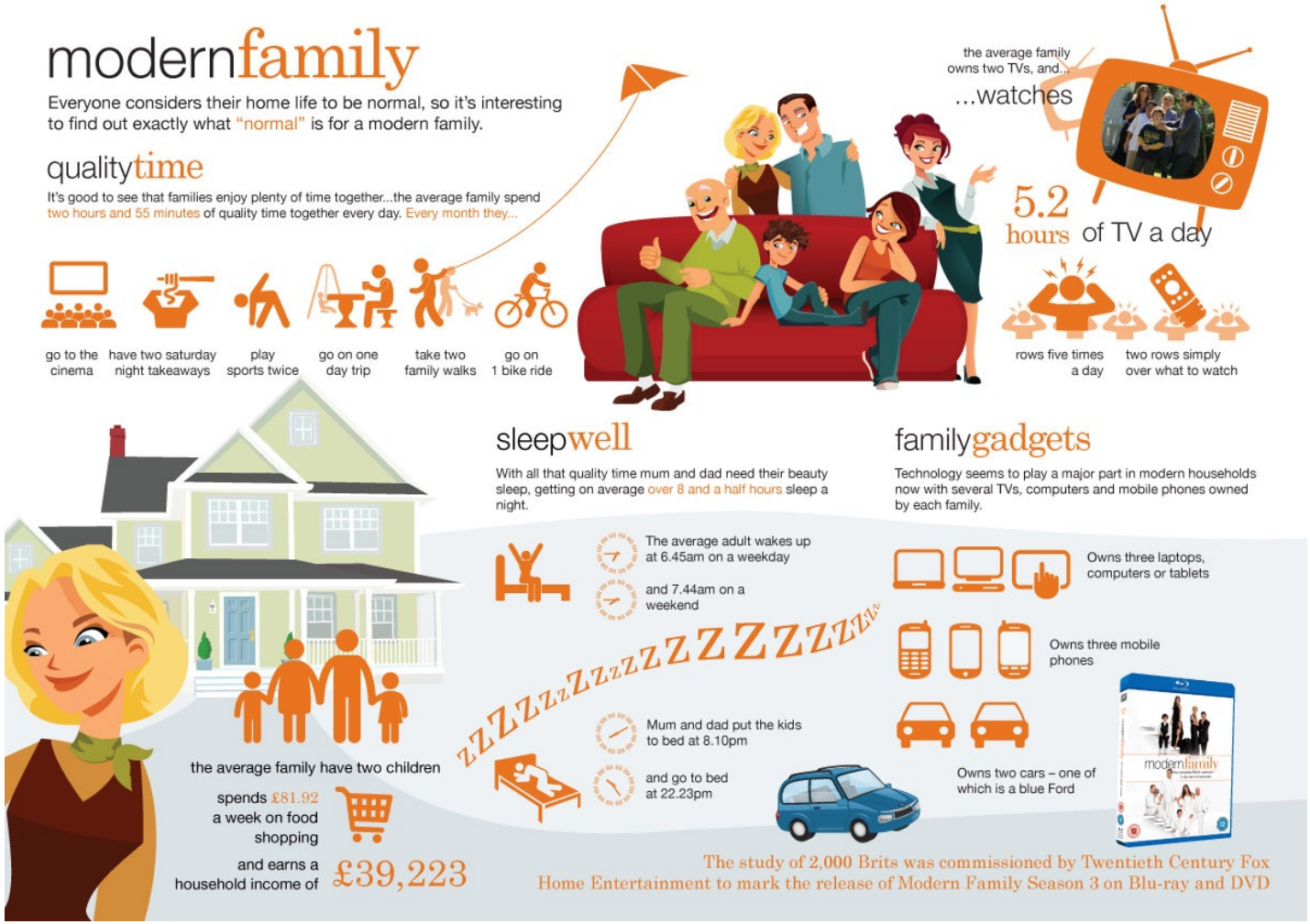


Source. www.lynda.com

What do people do?



Modern UK family



Source: http://news-by-design.com

Matlab - Statistical distributions

- Random number generators;
- Histograms, Pie-charts
- Boxplots, qq plots
- Exploring data in 2D
- Exploring data in 3D

Matlab functions

- randn, rand
- hist, bar, barh
- qqplot, boxplot
- imagesc
- mesh, surf,
- bar3, hist3

Q&A