Data Analysis & Visualisation

Week 2

Raoul Grouls, 26-2-2024

Recap leerdoelen les 1

- de gestalt principes & five guidelines toepassen op visualisaties
- een virtual environment activeren met pdm
- nieuwe features extraheren met behulp van regular expressions
- Een script vanaf de terminal opstarten
- click gebruiken voor command line arguments bij een script
- begrijpt de opzet van een project (src folder, data/raw en data/processed, pyproject.toml, notebooks) en kan dit zelf opzetten
- kan een eigen git-repo maken
- Regular expressions toepassen:
    - start ^
    - end $
    - or: [Bb]
    - ranges [a-zA-Z]
    - any char .
    - zero or more a*
    - one or more a+
    - not in range [^a-z]
    - shortcuts (\w, \s, \d)
    - lookbehind (?<=...)
    - lookahead (?=...)

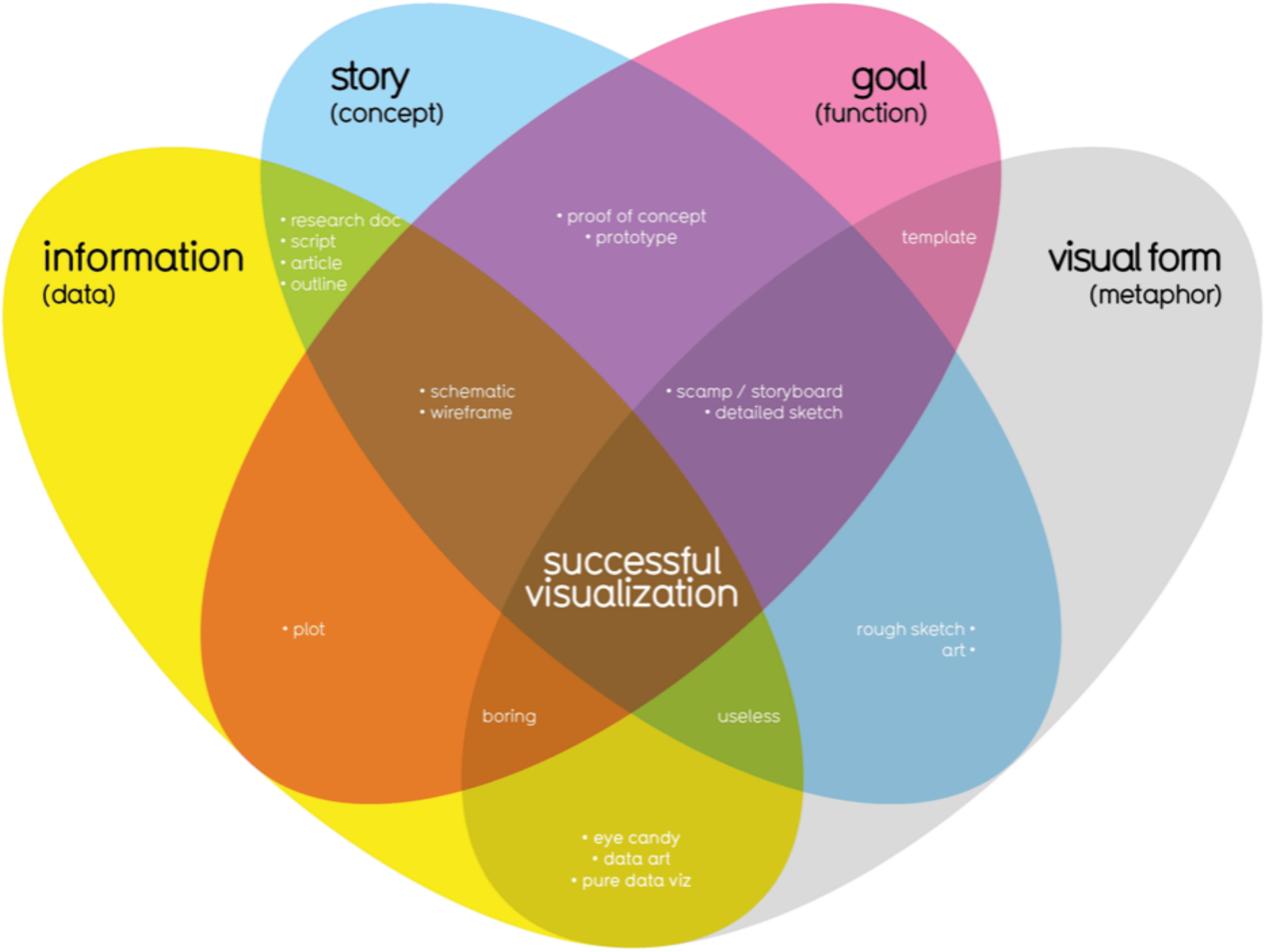
leerdoelen les 2

- Leren toepassen van visualisatie principes
- Omgaan met venv, pdm, path, scripts
- Oefenen met nieuwe features extraheren met behulp van regular expressions
- Vergelijken van categorieen met behulp van data visualisaties:
    - Barplots
    - Barbell plot
    - heatmaps
- Werken met palettes (en list comprehensions)
- Pandas
    - Pandas groupby & aggregate
    - Pandas cut

rollover for more detail

What Makes a Good Visualization?

explicit (implicit)



story
(concept)

goal
(function)

information
(data)

visual form
(metaphor)

- research doc
- script
- article
- outline

- proof of concept
- prototype

template

- schematic
- wireframe

- scamp / storyboard
- detailed sketch

successful visualization

- plot

rough sketch •
art •

boring

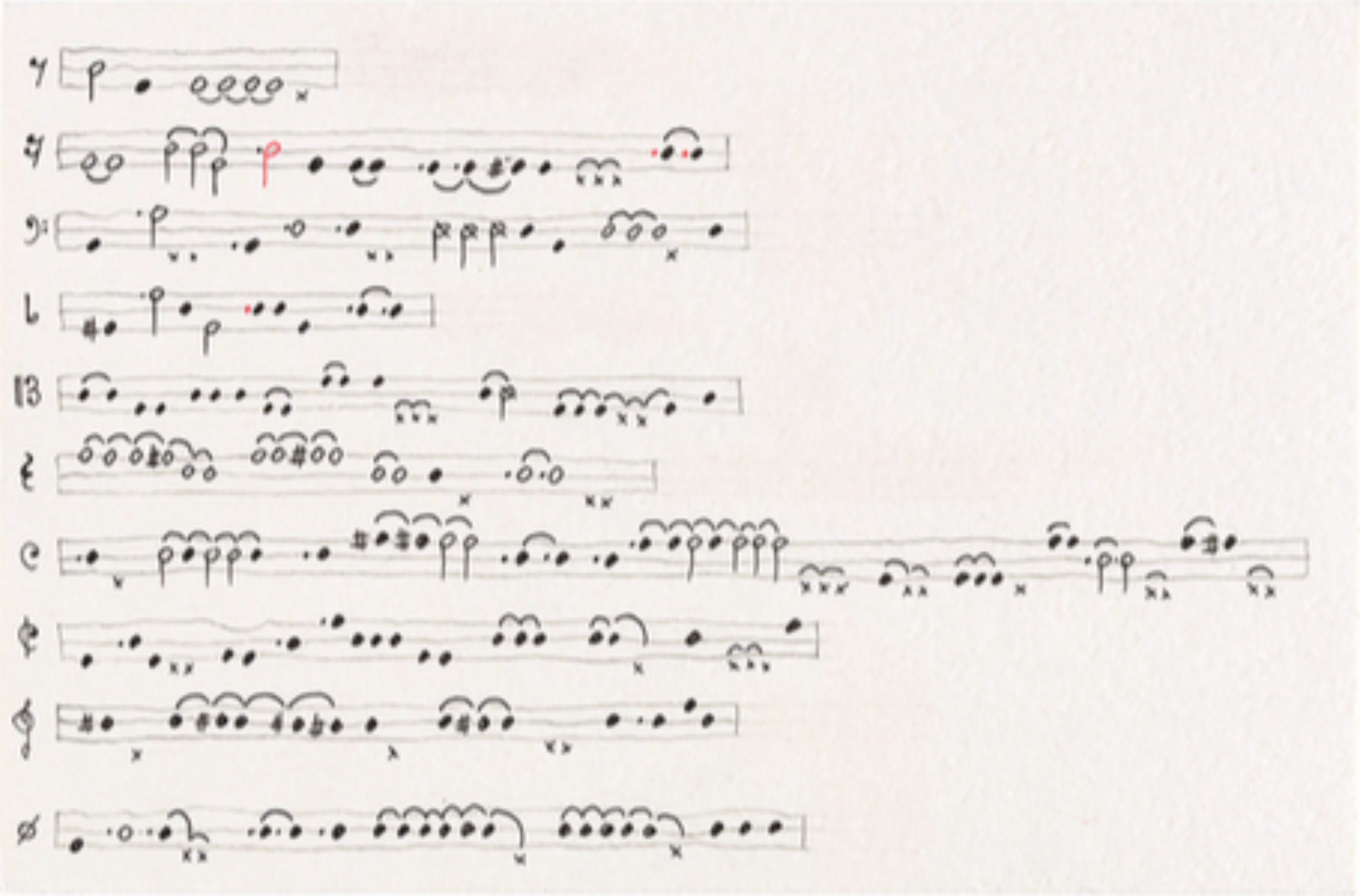
useless

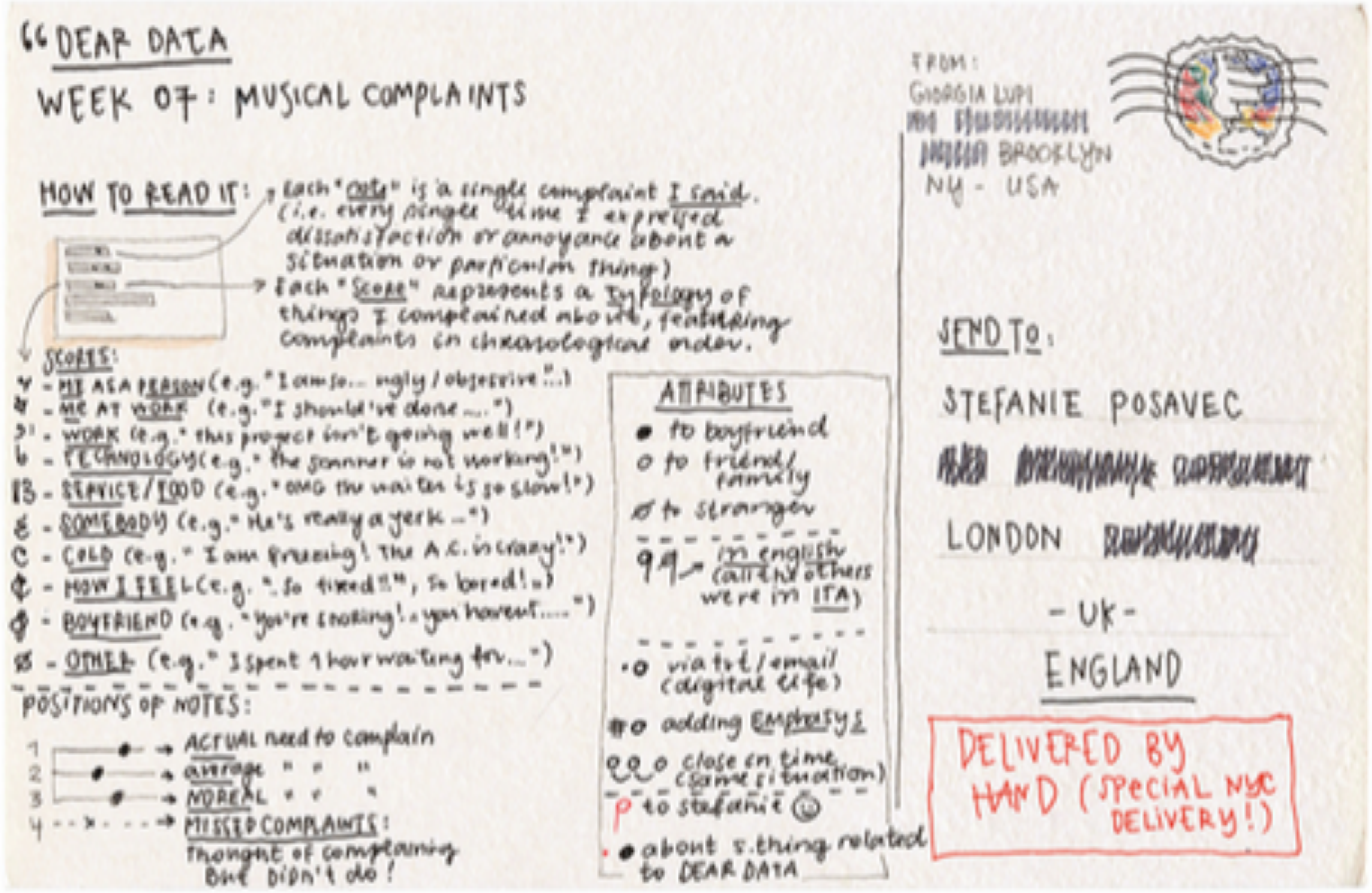
- eye candy
- data art
- pure data viz

David McCandless
InformationisBeautiful.net

taken from new book
Knowledge is Beautiful

find out more
bit.ly/KIB_Books





"DEAR DATA

WEEK 07: MUSICAL COMPLAINTS

HOW TO READ IT: → Each "note" is a single complaint I said. (i.e. every single time I expressed dissatisfaction or annoyance about a situation or particular thing)
→ Each "score" represents a typology of things I complained about, featuring complaints in chronological order.

SCORES:
- Y - ME AS A PERSON (e.g. "I am so... ugly / obsessive"...)
- 4 - ME AT WORK (e.g. "I should've done ....")
- 9: - WORK (e.g. "this project isn't going well!")
- 6 - TECHNOLOGY (e.g. "the scanner is not working!")
- 13 - SERVICE/FOOD (e.g. "OMG the waiter is so slow!")
- ξ - SOMEBODY (e.g. "He's really a jerk ...")
- C - COLD (e.g. "I am freezing! The A.C. is crazy!")
- ¢ - HOW I FEEL (e.g. "...so tired!!", "so bored!")
- 𝄞 - BOYFRIEND (e.g. "You're snoring!" "you haven't....")
- Ø - OTHER (e.g. "I spent 1 hour waiting for...")

POSITIONS OF NOTES:
1 → ACTUAL need to complain
2 → average " "
3 → NORMAL " "
4 → MISSED COMPLAINTS: thought of complaining but didn't do!

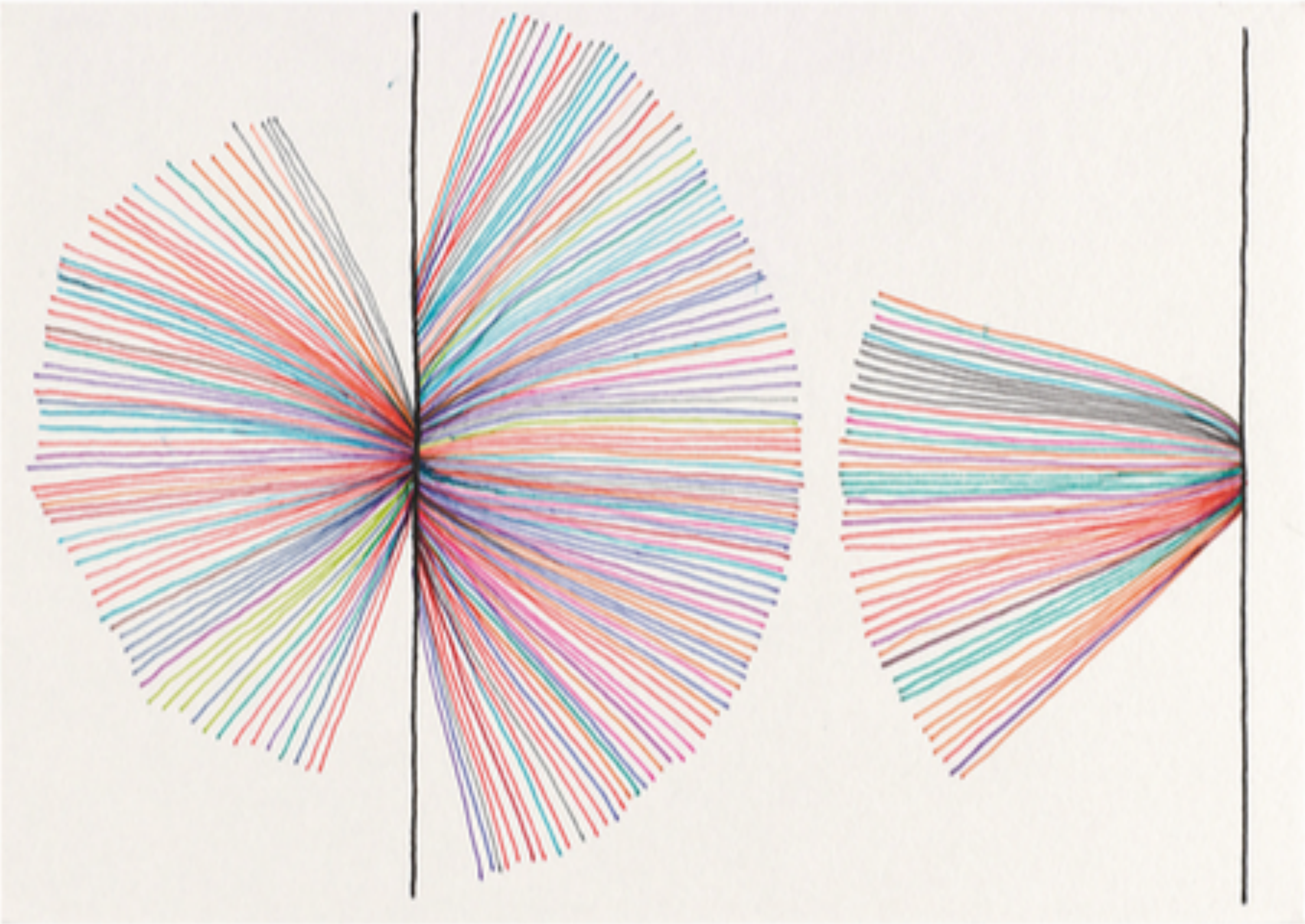
ATTRIBUTES
- ● to boyfriend
- ○ to friend/family
- ø to stranger
- 𝅘𝅥 → in english (all the others were in ITA)
- ·○ via txt/email (digital life)
- #○ adding Emphasys!
- close in time (same situation)
- p to stefanie ☺
- ·● about s.thing related to DEAR DATA

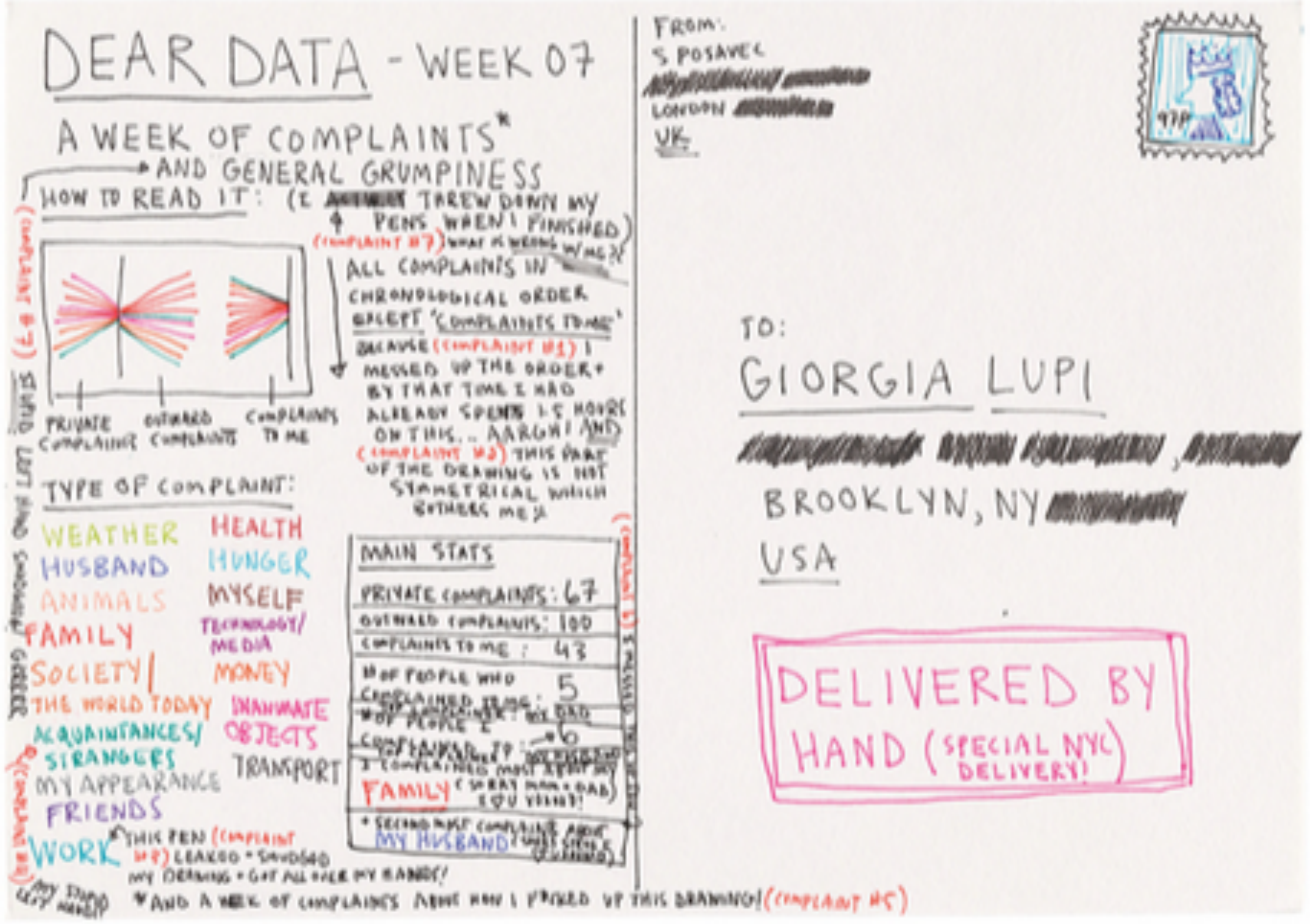
FROM:
GIORGIA LUPI
[redacted]
[redacted] BROOKLYN
NY - USA

SEND TO:
STEFANIE POSAVEC
[redacted]
LONDON [redacted]
- UK -
ENGLAND

DELIVERED BY HAND (SPECIAL NYC DELIVERY!)

What better visual reference than a musical score to show the repetitiveness of Giorgia's protests and the "level" of complaint: whether they are justified or totally out of place.





DEAR DATA - WEEK 07

A WEEK OF COMPLAINTS*
→ AND GENERAL GRUMPINESS

HOW TO READ IT: (I [redacted] THREW DOWN MY PENS WHEN I FINISHED) (COMPLAINT #7) WHAT IS WRONG W/ ME?!

PRIVATE COMPLAINTS
OUTWARD COMPLAINTS
COMPLAINTS TO ME

ALL COMPLAINTS IN CHRONOLOGICAL ORDER EXCEPT 'COMPLAINTS TO ME' BECAUSE (COMPLAINT #1) I MESSED UP THE ORDER + BY THAT TIME I HAD ALREADY SPENT 1.5 HOURS ON THIS... AARGH! AND (COMPLAINT #2) THIS PART OF THE DRAWING IS NOT SYMMETRICAL WHICH BOTHERS ME!!

TYPE OF COMPLAINT:
- WEATHER
- HUSBAND
- ANIMALS
- FAMILY
- SOCIETY
- THE WORLD TODAY
- ACQUAINTANCES/STRANGERS
- MY APPEARANCE
- FRIENDS
- WORK
- HEALTH
- HUNGER
- MYSELF
- TECHNOLOGY/MEDIA
- MONEY
- INANIMATE OBJECTS
- TRANSPORT

MAIN STATS
PRIVATE COMPLAINTS: 67
OUTWARD COMPLAINTS: 100
COMPLAINTS TO ME: 43
OF PEOPLE WHO COMPLAINED TO ME: 5
TOP COMPLAINER: MY DAD
OF PEOPLE I COMPLAINED TO: 6
TOP COMPLAINEE? MY HUSBAND
I COMPLAINED MOST ABOUT MY FAMILY (SORRY MUM + DAD LOVE YOU!!!)
* SECOND MOST COMPLAINED ABOUT: MY HUSBAND (SORRY [illegible])

(COMPLAINT #6) STAINED LEFT HAND SMUDGING [illegible]

* THIS PEN (COMPLAINT #4) LEAKED + SMUDGED MY DRAWING + GOT ALL OVER MY HANDS!

MY STAINED LEFT HAND!!

* AND A WEEK OF COMPLAINTS ABOUT HOW I F'KED UP THIS DRAWING! (COMPLAINT #5)

FROM:
S POSAVEC
[redacted]
LONDON [redacted]
UK

97P

TO:
GIORGIA LUPI
[redacted]
BROOKLYN, NY [redacted]
USA

DELIVERED BY HAND (SPECIAL NYC DELIVERY!)

Note the hand-drawn stamps: these postcards were delivered in person in New York!

What is Consciousness?

Make up your own mind

A field that exists in its own parallel "realm" of existence outside reality so can't be seen.
(Substance Dualism)

A sensation that "grows" inevitably out of complicated brain states.
(Emergent Dualism)

A physical property of all matter, like electromagnetism, just not one the scientists know about.
(Property Dualism)

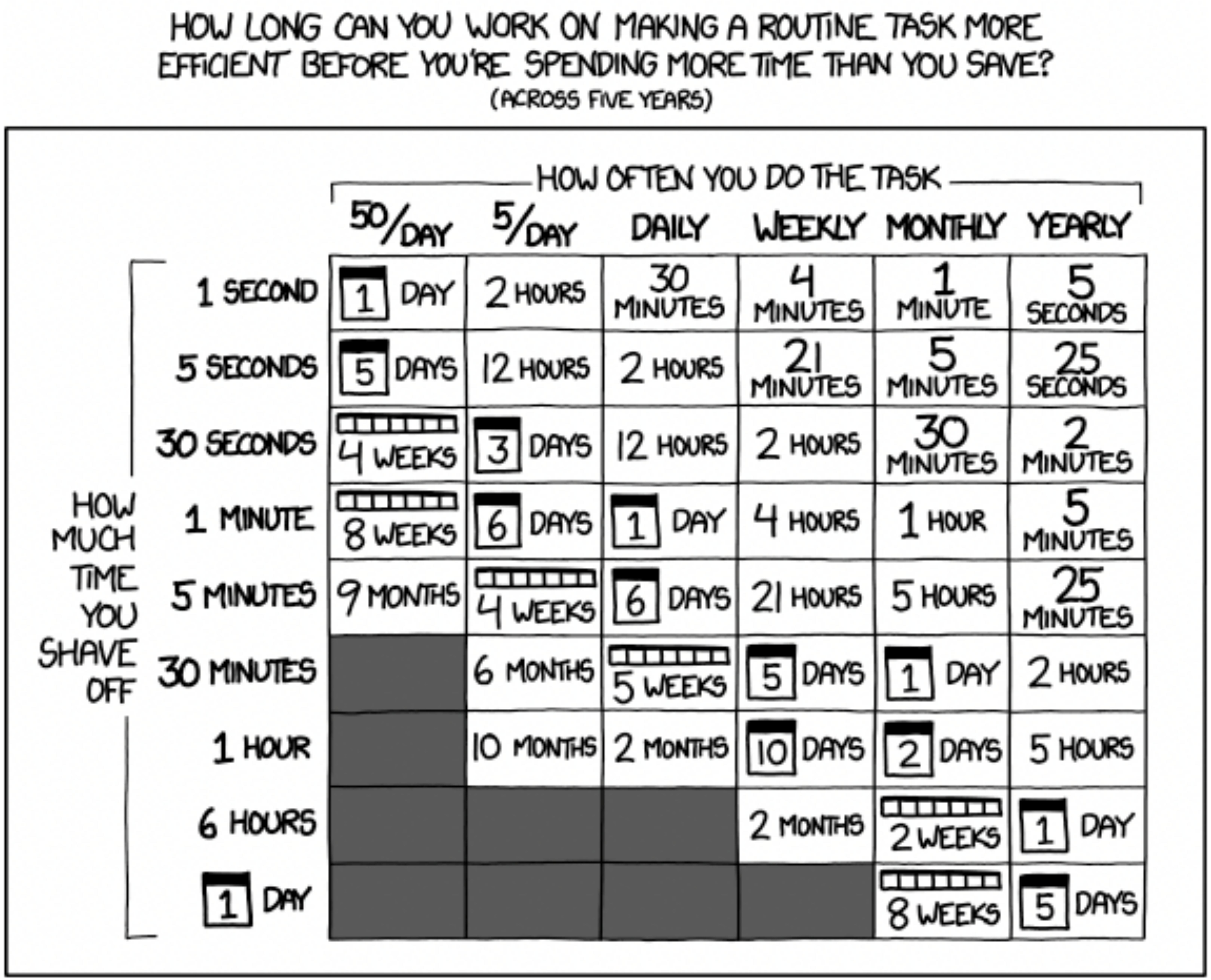
All matter has a psychic part. Consciousness is just the psychic part of our brain.
(Pan Psychism)

Simply, mental states are physical events that we can see in brain scans.
(Identity Theory)

Consciousness and its states (belief, desire, pain) are simply functions the brain performs.
(Functionalism)

Literally just behaviour. When we behave in a certain way, we appear conscious.
(Behaviourism)

An accidental side-effect of complex physical processes in the brain.
(Epiphenomenalism)



HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?

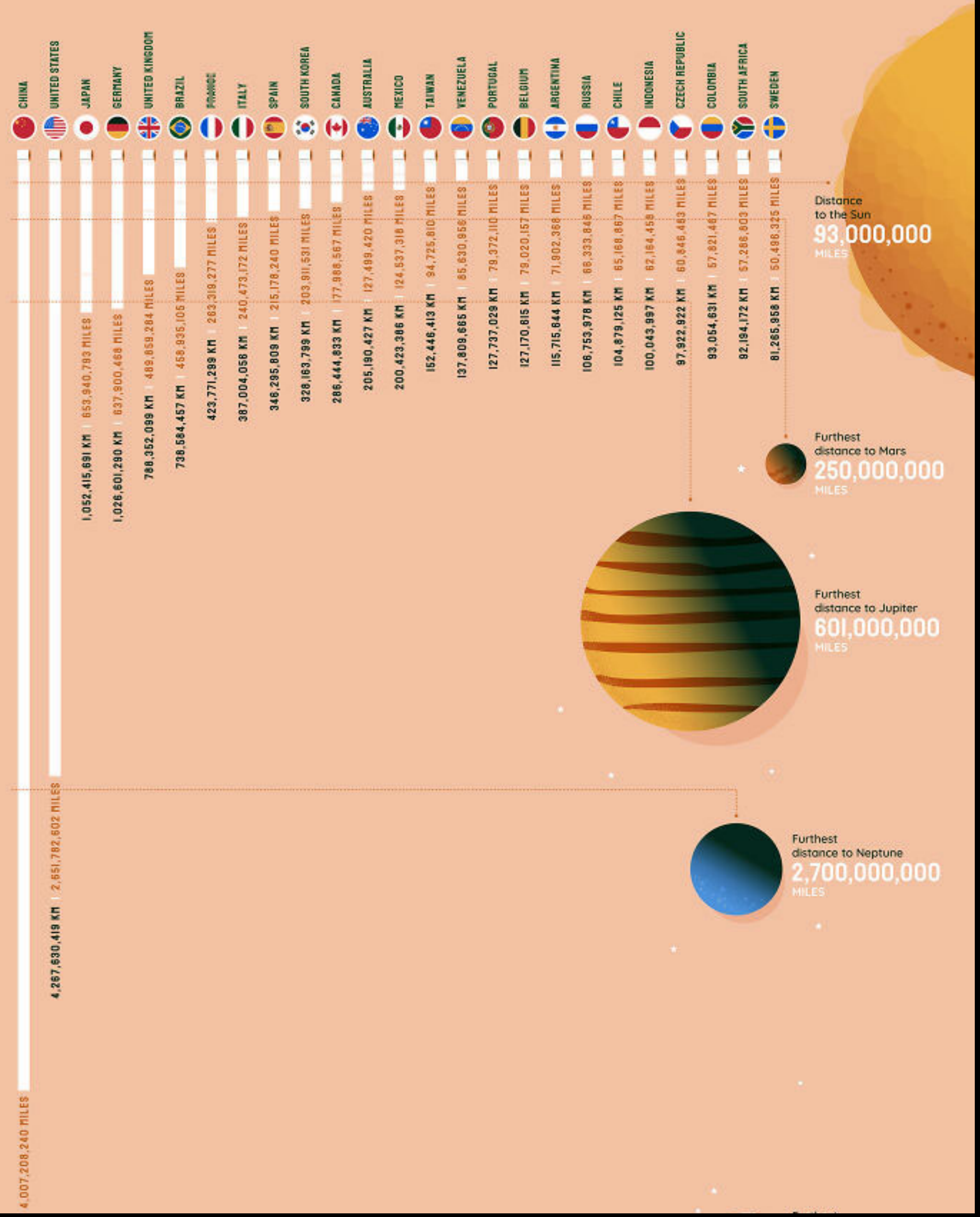
(ACROSS FIVE YEARS)

	HOW OFTEN YOU DO THE TASK					
HOW MUCH TIME YOU SHAVE OFF	50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS

THE COUNTRIES THAT USE THE MOST TOILET PAPER

If you laid out all of the rolls of toilet paper used in each country in one year, what would that look like? For China it would be an incredible 4 billion miles long, which is further than the distance from Earth to Neptune. For the USA it's 2.65 billion miles, while both Japan and Germany use enough toilet paper to stretch out beyond Jupiter.

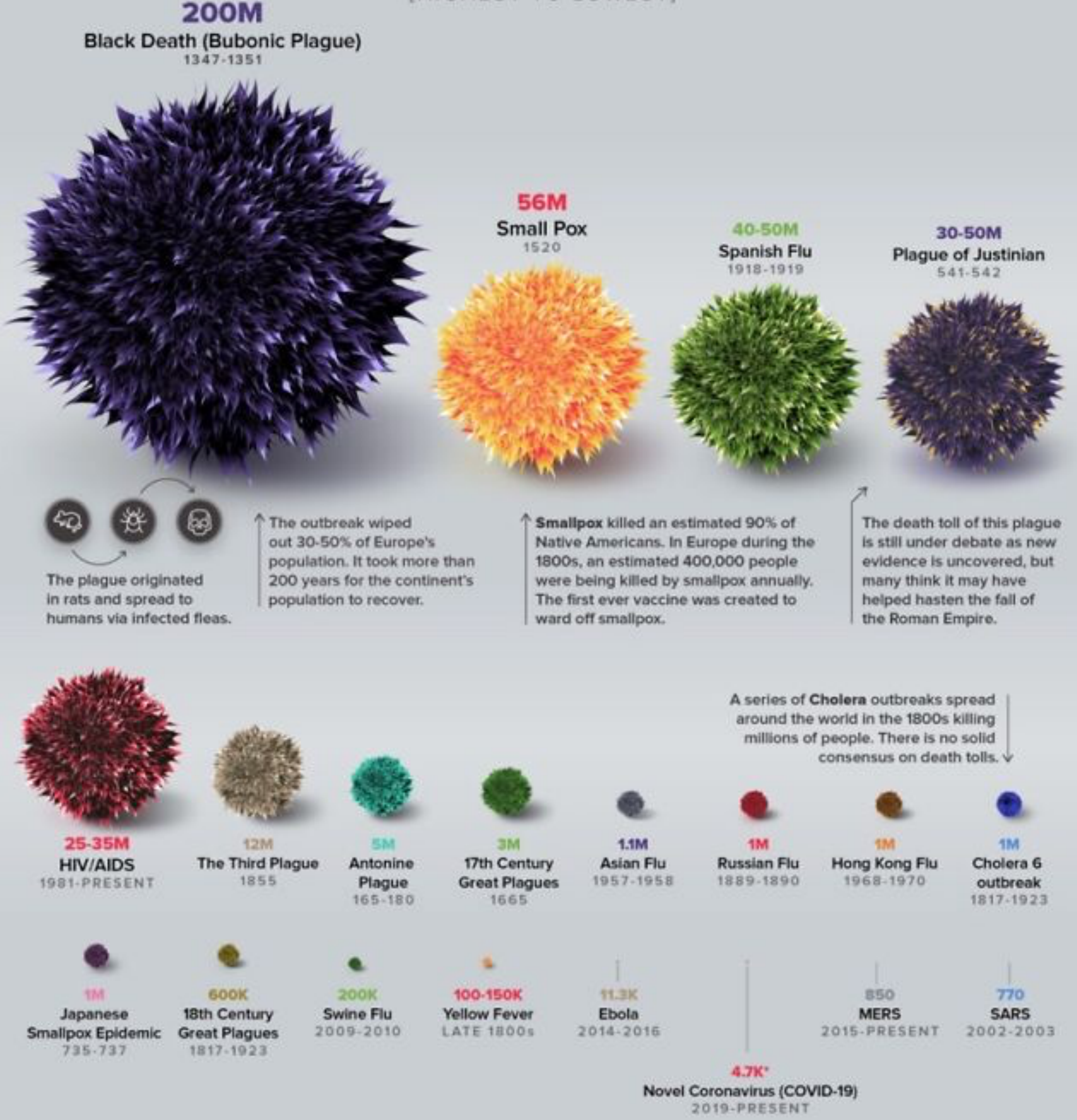
▸ TOILET PAPER USAGE PER YEAR PER COUNTRY



the disease is new to medicine, and data is still coming in.

DEATH TOLL

[HIGHEST TO LOWEST]



200M
Black Death (Bubonic Plague)
1347-1351

The plague originated in rats and spread to humans via infected fleas.

The outbreak wiped out 30-50% of Europe's population. It took more than 200 years for the continent's population to recover.

56M
Small Pox
1520

Smallpox killed an estimated 90% of Native Americans. In Europe during the 1800s, an estimated 400,000 people were being killed by smallpox annually. The first ever vaccine was created to ward off smallpox.

40-50M
Spanish Flu
1918-1919

30-50M
Plague of Justinian
541-542

The death toll of this plague is still under debate as new evidence is uncovered, but many think it may have helped hasten the fall of the Roman Empire.

A series of **Cholera** outbreaks spread around the world in the 1800s killing millions of people. There is no solid consensus on death tolls.

25-35M
HIV/AIDS
1981-PRESENT

12M
The Third Plague
1855

5M
Antonine Plague
165-180

3M
17th Century Great Plagues
1665

1.1M
Asian Flu
1957-1958

1M
Russian Flu
1889-1890

1M
Hong Kong Flu
1968-1970

1M
Cholera 6 outbreak
1817-1923

1M
Japanese Smallpox Epidemic
735-737

600K
18th Century Great Plagues
1817-1923

200K
Swine Flu
2009-2010

100-150K
Yellow Fever
LATE 1800s

11.3K
Ebola
2014-2016

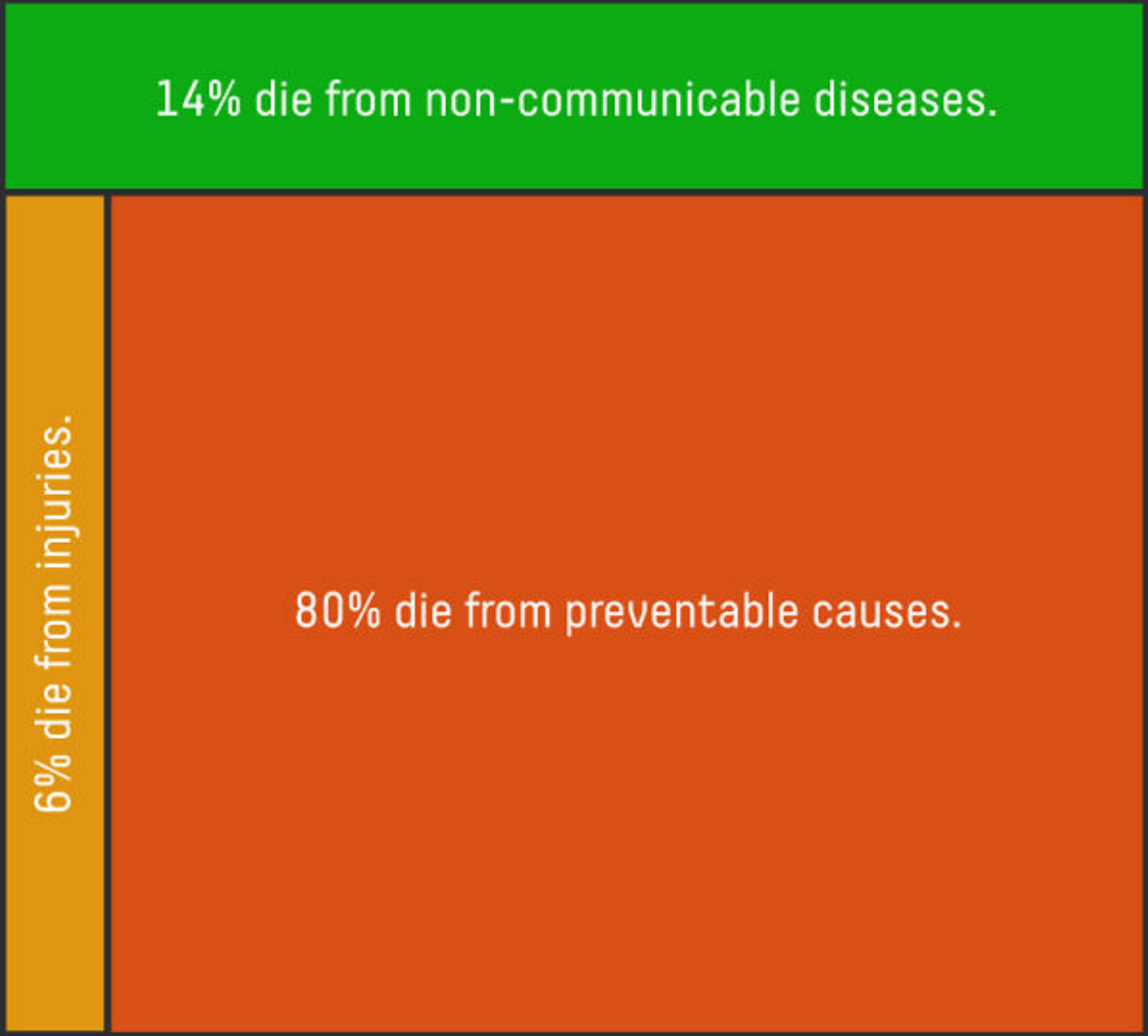
4.7K*
Novel Coronavirus (COVID-19)
2019-PRESENT

850
MERS
2015-PRESENT

770
SARS
2002-2003

Causes of death in children under 5

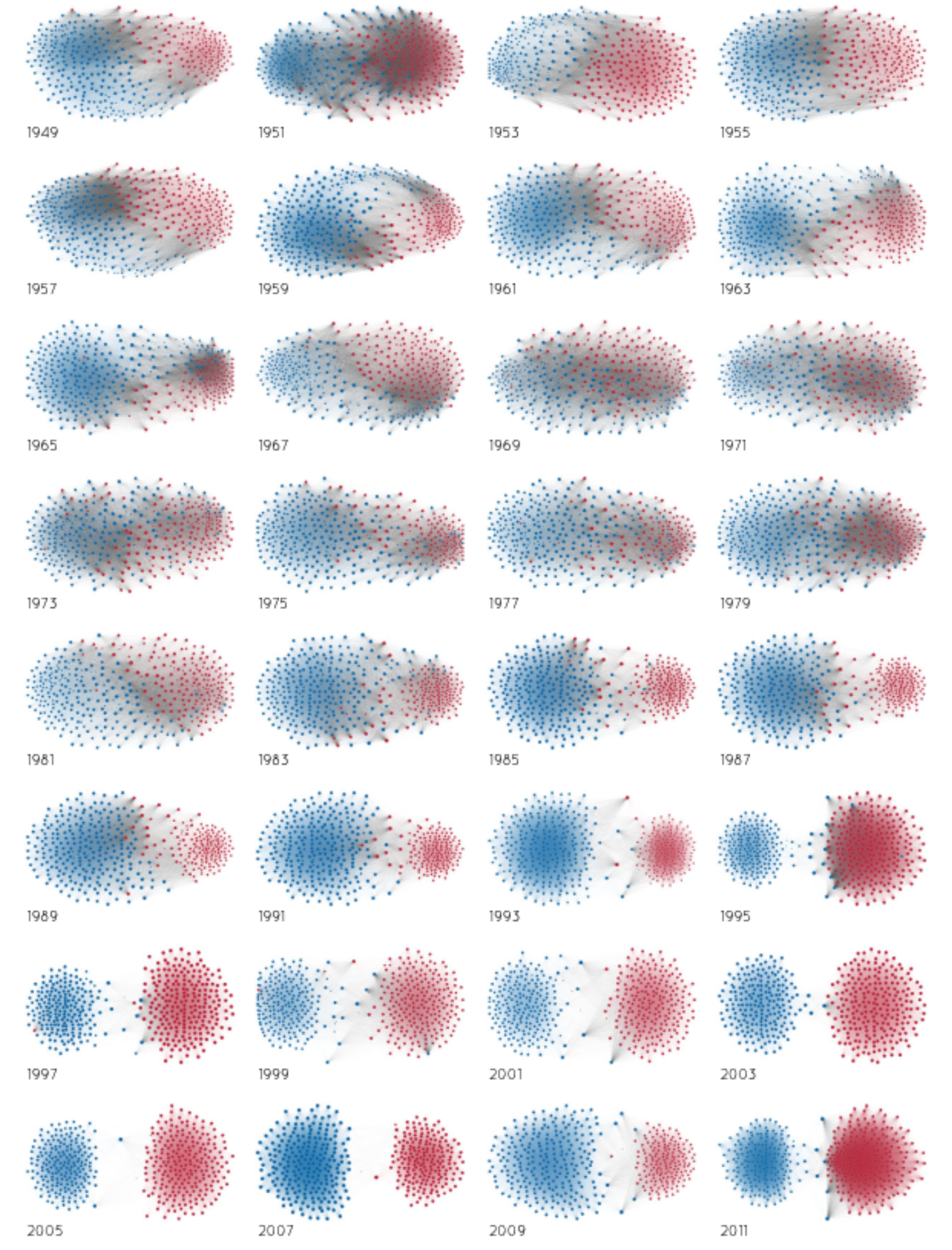
(2013)



Source: IHME

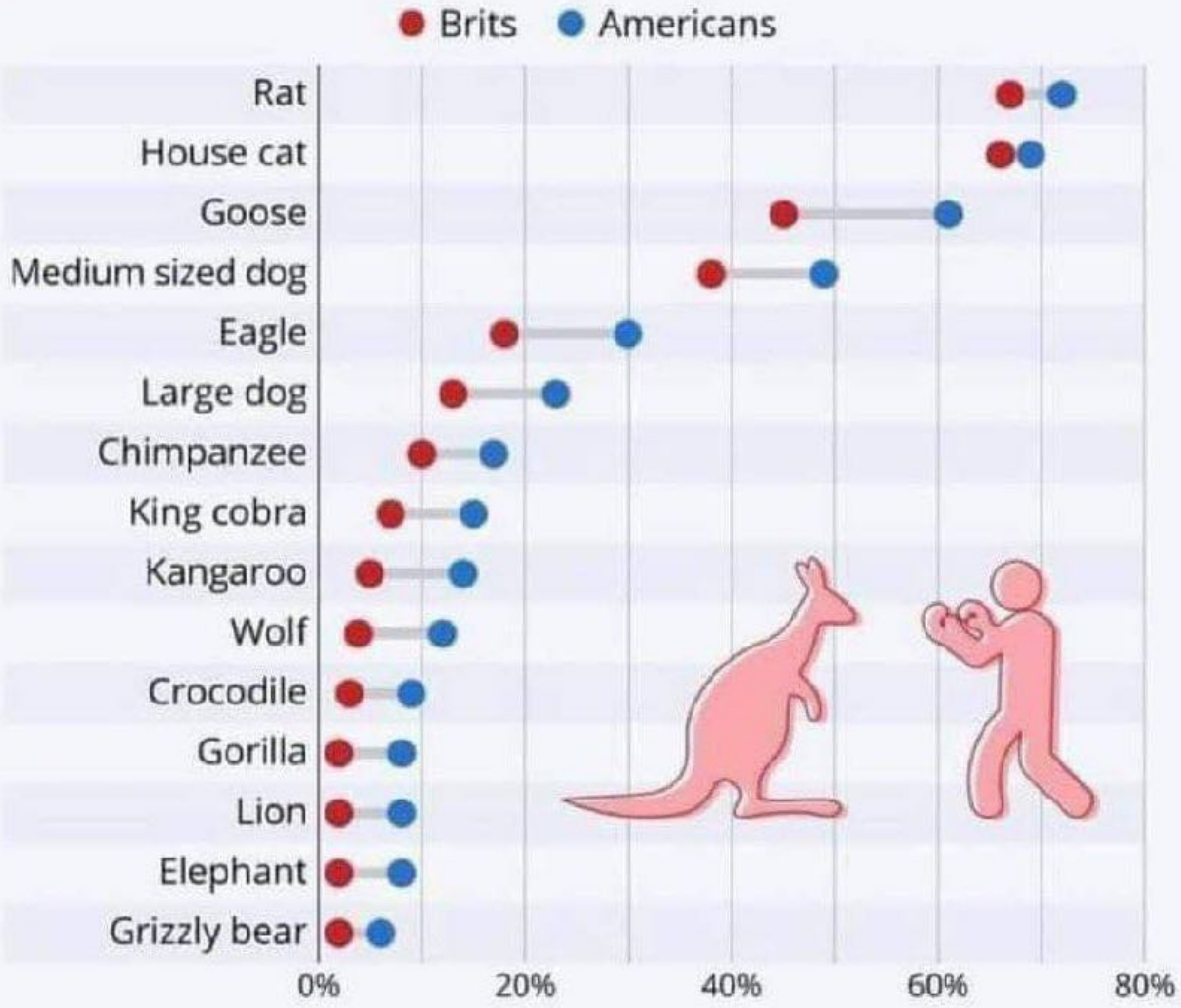
gatesnotes

The Rise of Partisanship and Super-Cooperators in the U.S. House of Representatives



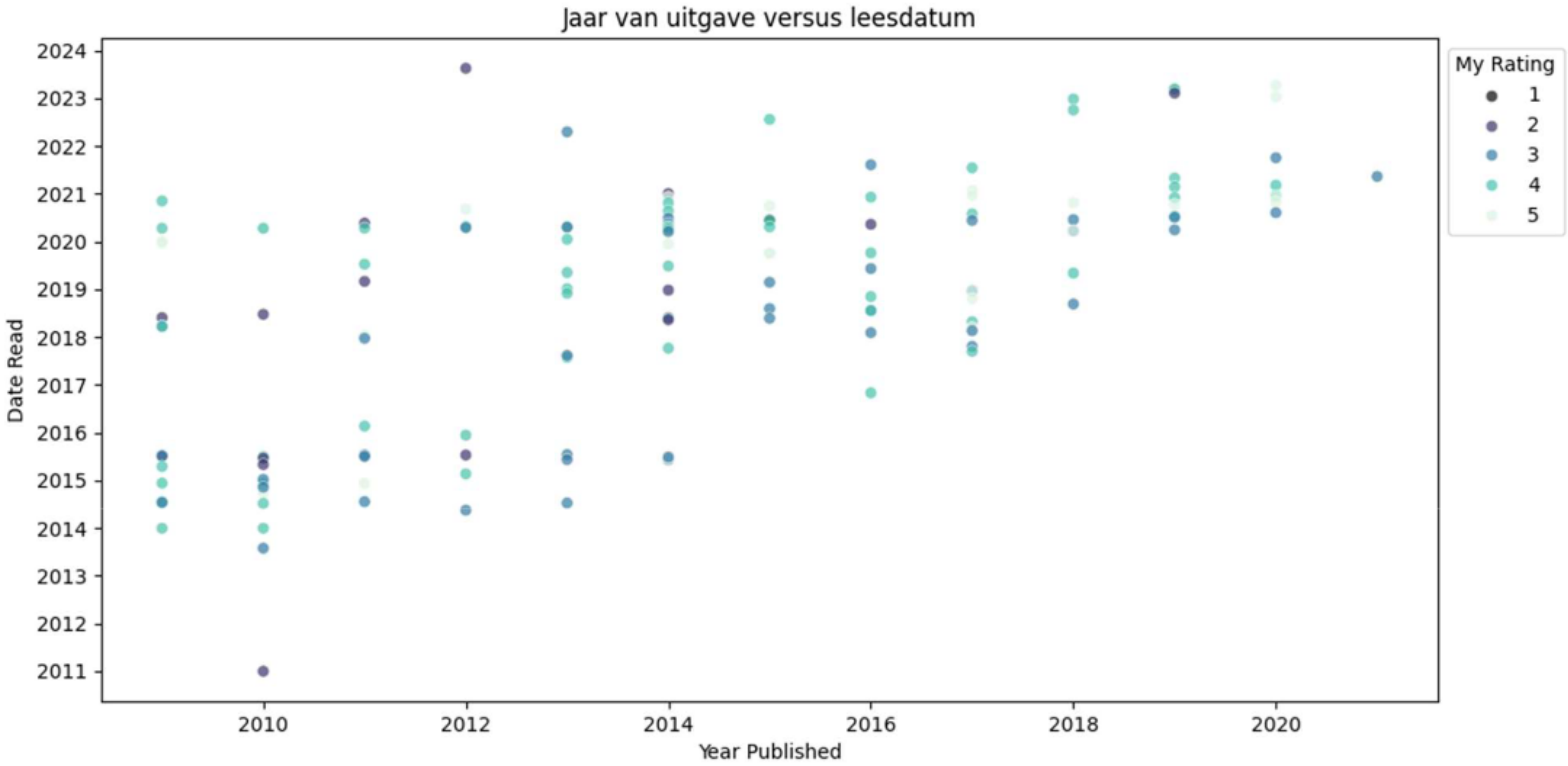
Which Animals Could You Beat in a Fight?

"Which of the following animals, if any, do you think you could beat in a fight if you were unarmed?"



Survey of 2,082 GB adults (conducted 18-19 May 2021) &
1,224 U.S. adults (conducted 12-13 April 2021).
Source: YouGov

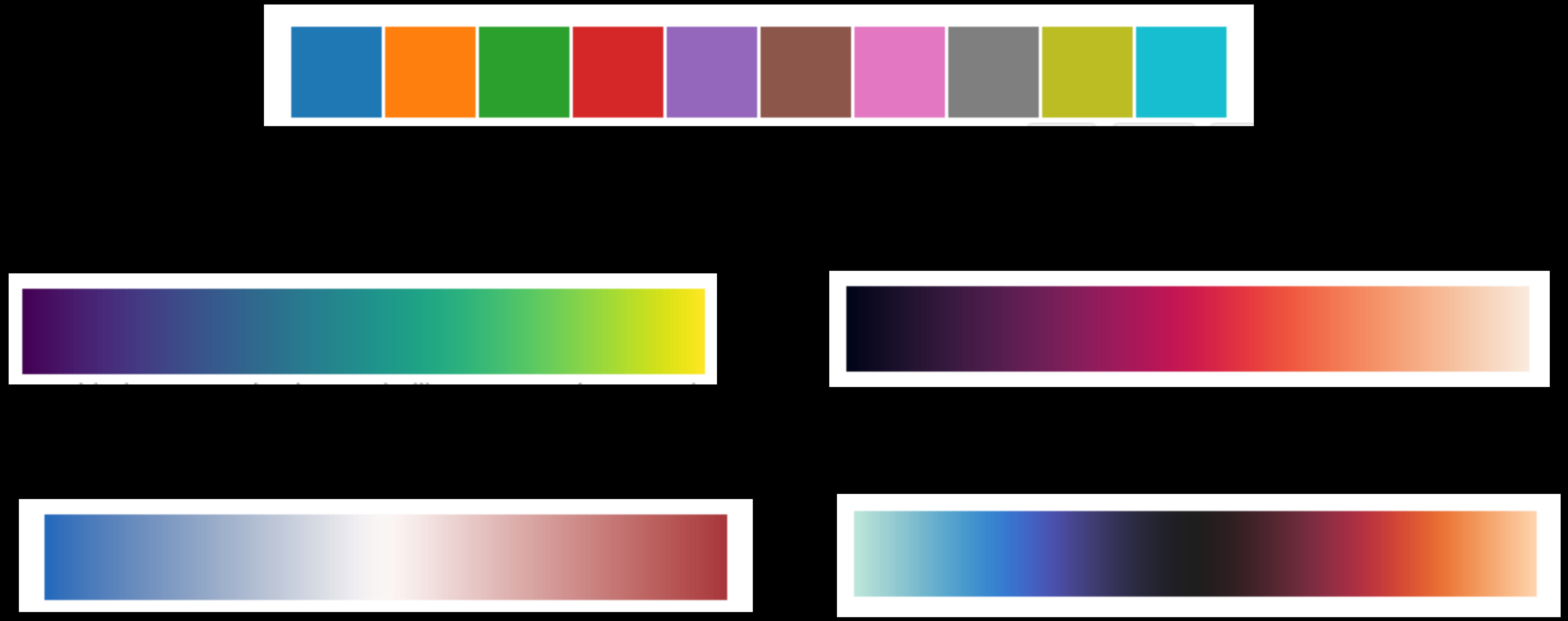
Trend – publicatie jaar vs Date Read



Color palettes

https://seaborn.pydata.org/tutorial/color_palettes.html

- **Qualitative**: representing categorical data
- **Sequential**: perceptually uniform
- **Diverging**: both large low and high values are interesting and span a midpoint value



Introduction regular expressions

- Regular expressions (regex) are a powerful tool for matching patterns in text.
- Used for searching, editing, or manipulating text and data.
- Use https://regex101.com/ to develop and test your regexes
- chatGPT is pretty good at creating and explaining regexes.

Basic symbols

- **^ (Start):** Matches the start of a line.
- **$ (End):** Matches the end of a line.
- **. (Any Char):** Matches any character except a newline.

- Example: To match any line that starts with "A", we use `^A`

The OR operator and Ranges

- **[Bb] (Or):** Matches either "B" or "b".
- **[a-zA-Z] (Ranges):** Matches any letter, regardless of case.
- **[0-9]:** matches any number from 0 to 9

Example:

- To find any line that starts with a lowercase letter or number, use `^[a-z0-9]`

Quantifiers

- **a* (Zero or More):** Matches zero or more occurrences of "a".
- **a+ (One or More):** Matches one or more occurrences of “a".
- **a{3} (Exactly Three):** Matches exactly three occurrences of "a".
- **a{2,5} (Two to Five):** Matches between two and five occurrences of “a".

- **Example**: Combined with ranges: `[a-z]+`
- To match a string that contains four to six a’s in a row, use `a{4,6}`

Negation and Shortcuts

- **[^a-z] (Not in Range):** Matches any character not in the range "a" to "z".
- Shortcuts:
  - **\w (Word Char):** Matches any word character (letter, number, underscore).
  - **\s (Whitespace):** Matches any whitespace character (space, tab, newline).
  - **\d (Digit):** Matches any digit.

- **Question:** How to find lines <u>not</u> starting with any lowercase letter?

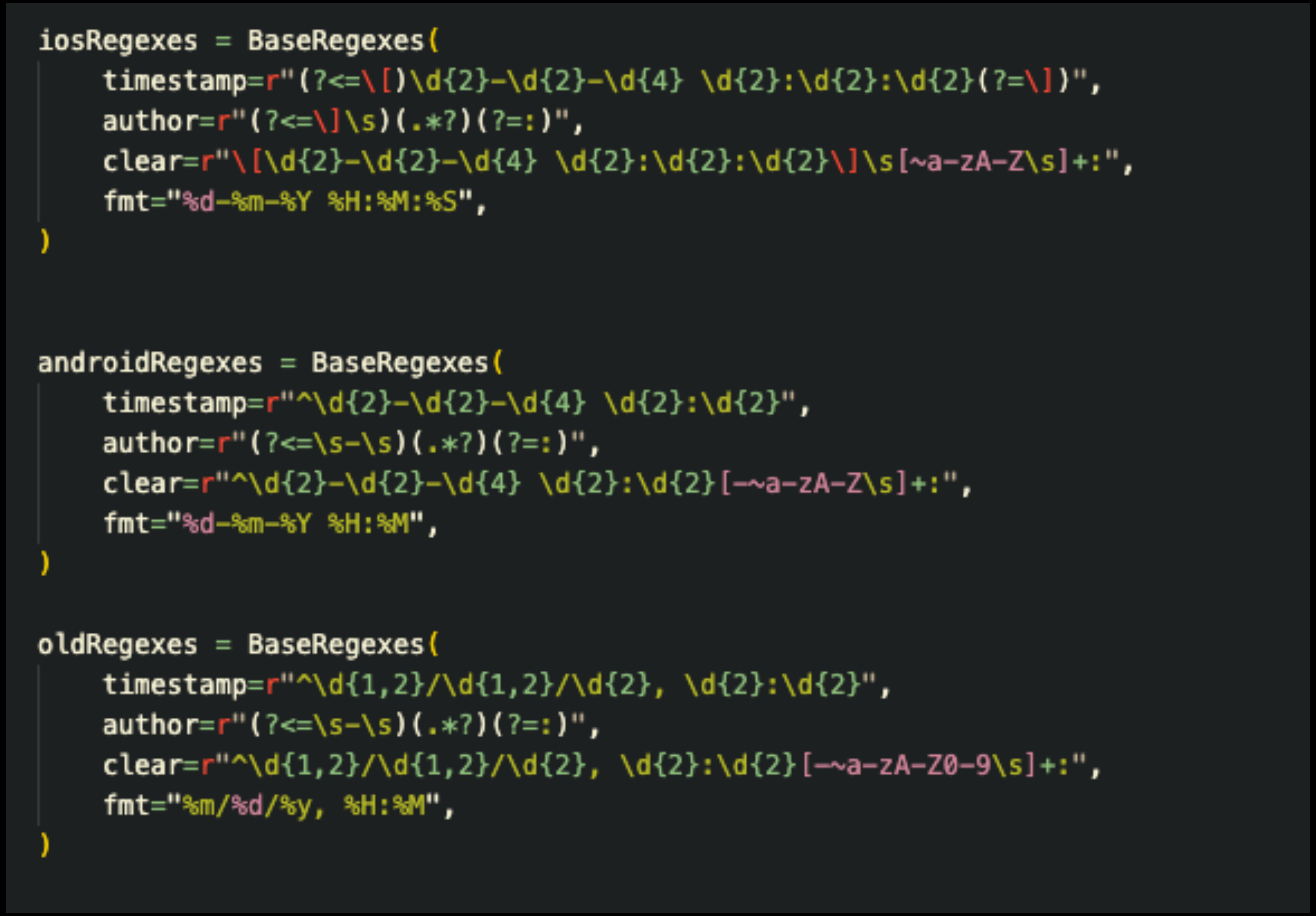
Lookaround

- **Lookahead (?=...):** Matches a group after the main expression without including it in the result.
- **Lookbehind (?<=...):** Matches a group before the main expression without including it in the result.
- **Example:** To find words directly after a colon without including the colon, use `(?<=:)\s*\w+`
- Question: what does the \s* do?

Combining everything

Message format:

- [11-01-2022 08:05:25] ~Raoul: Ah
- Can you explain what the other formats should be, as defined by the regexes?



```python
iosRegexes = BaseRegexes(
    timestamp=r"(?<=\[)\d{2}-\d{2}-\d{4} \d{2}:\d{2}:\d{2}(?=\])",
    author=r"(?<=\]\s)(.*?)(?=:)",
    clear=r"\[\d{2}-\d{2}-\d{4} \d{2}:\d{2}:\d{2}\]\s[~a-zA-Z\s]+:",
    fmt="%d-%m-%Y %H:%M:%S",
)


androidRegexes = BaseRegexes(
    timestamp=r"^\d{2}-\d{2}-\d{4} \d{2}:\d{2}",
    author=r"(?<=\s-\s)(.*?)(?=:)",
    clear=r"^\d{2}-\d{2}-\d{4} \d{2}:\d{2}[-~a-zA-Z\s]+:",
    fmt="%d-%m-%Y %H:%M",
)

oldRegexes = BaseRegexes(
    timestamp=r"^\d{1,2}/\d{1,2}/\d{2}, \d{2}:\d{2}",
    author=r"(?<=\s-\s)(.*?)(?=:)",
    clear=r"^\d{1,2}/\d{1,2}/\d{2}, \d{2}:\d{2}[-~a-zA-Z0-9\s]+:",
    fmt="%m/%d/%y, %H:%M",
)
```

Improving with chatGPT

[…] this is the dataformat, […] this the regex, can you improve it?

\[([0-2]?\d|3[01])-(0?\d|1[012])-(\d{4}) ([01]?\d|2[0-3]):([0-5]?\d):([0-5]?\d)\]

This regex breaks down as follows:

- `([0-2]?\d|3[01])`: Matches days from `1-31`, allowing for a leading zero.
- `(0?\d|1[012])`: Matches months from `1-12`, also allowing for a leading zero.
- `(\d{4})`: Matches any four digits for the year.
- `([01]?\d|2[0-3])`: Matches hours from `0-23`, with optional leading zero.
- `([0-5]?\d)`: Matches minutes and seconds from `0-59`, with optional leading zero.