

Data Analysis & Visualisation

Raoul Grouls, 6 September 2024

Hello world

- Raoul Grouls, data alchemist, <https://the-pttrn.nl/>
- Research @ HAN, ‘internal representations of the world in AI models’
- Teaching @ HAN, HU

raoul.grouls@han.nl

<https://github.com/raoulg>

Leerdoelen les 1

De student begrijpt:

- Gestalt principles: proximity, similarity, enclosure, closure, continuity, connection
- preattentive processing (first 200ms)
- five guidelines: show the data / reduce clutter / integrate text / avoid spaghetti / start with grey
- Wat de motivatie is achter een virtual environment

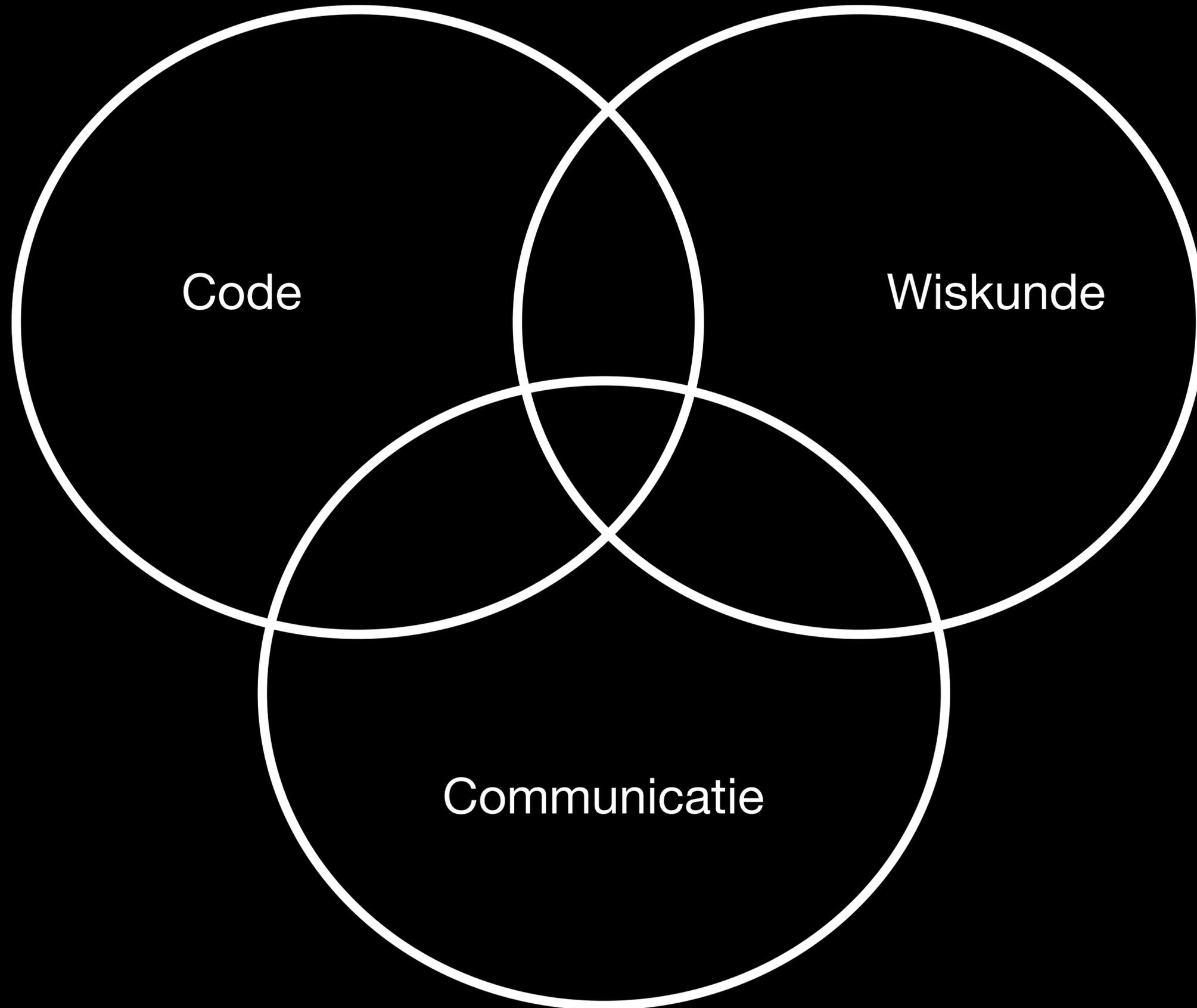
De student kan:

- de gestalt principes & five guidelines toepassen op visualisaties

- een virtual environment activeren met rye
- nieuwe features extraheren met behulp van regular expressions
- Een script op de VM 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:

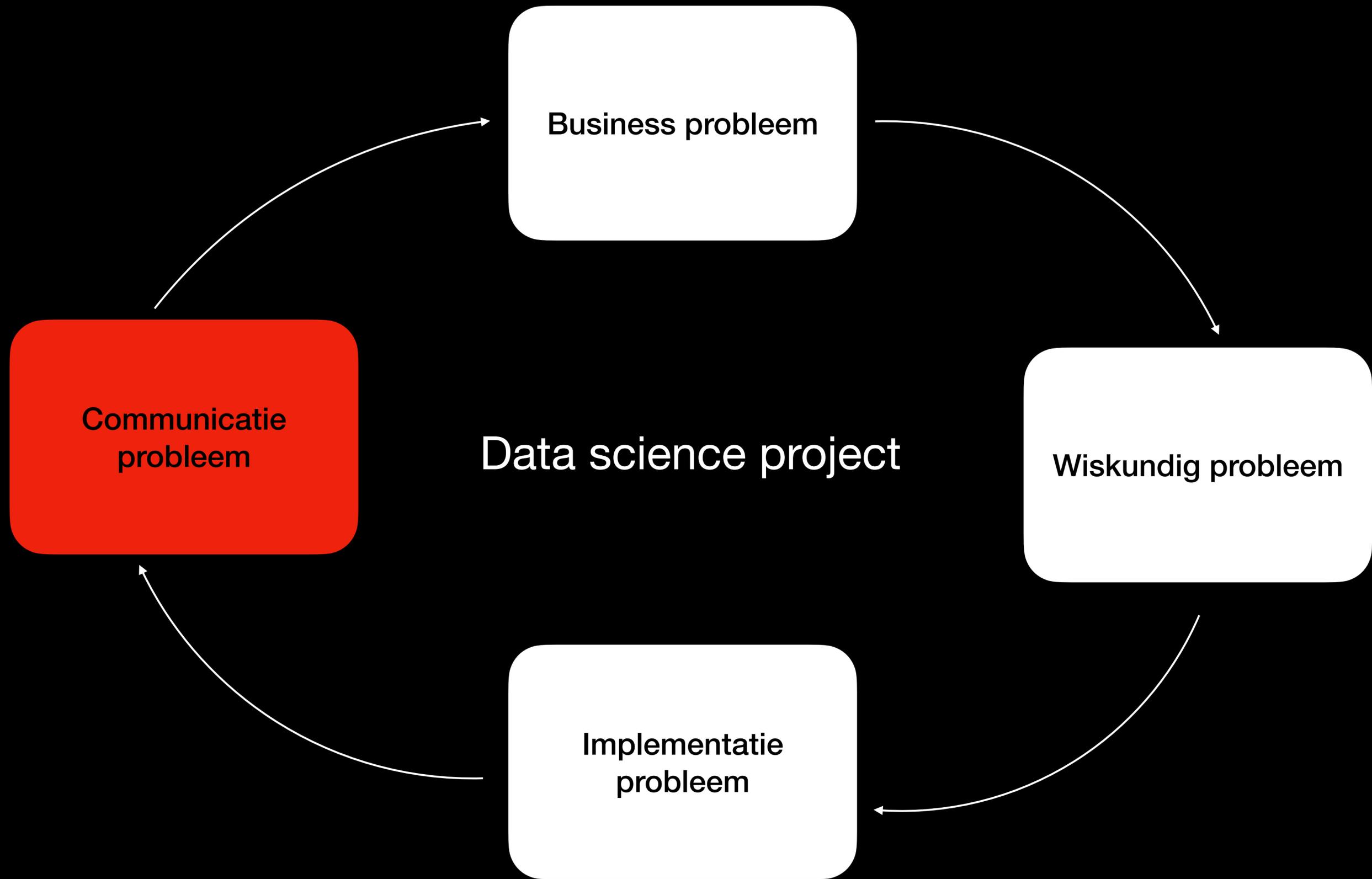
- start `^`
- end `$`
- or: `[Bb]`
- ranges `[a-zA-Z]`
- any char `.`
- zero or more `a*`
- one or more `a+`
- not in range `[^a-z]`
- shortcuts `(\w, \s, \d)`
- lookbehind `(?<=...)`
- lookahead `(?=...)`



Code

Wiskunde

Communicatie



The first technology



Stateful communication

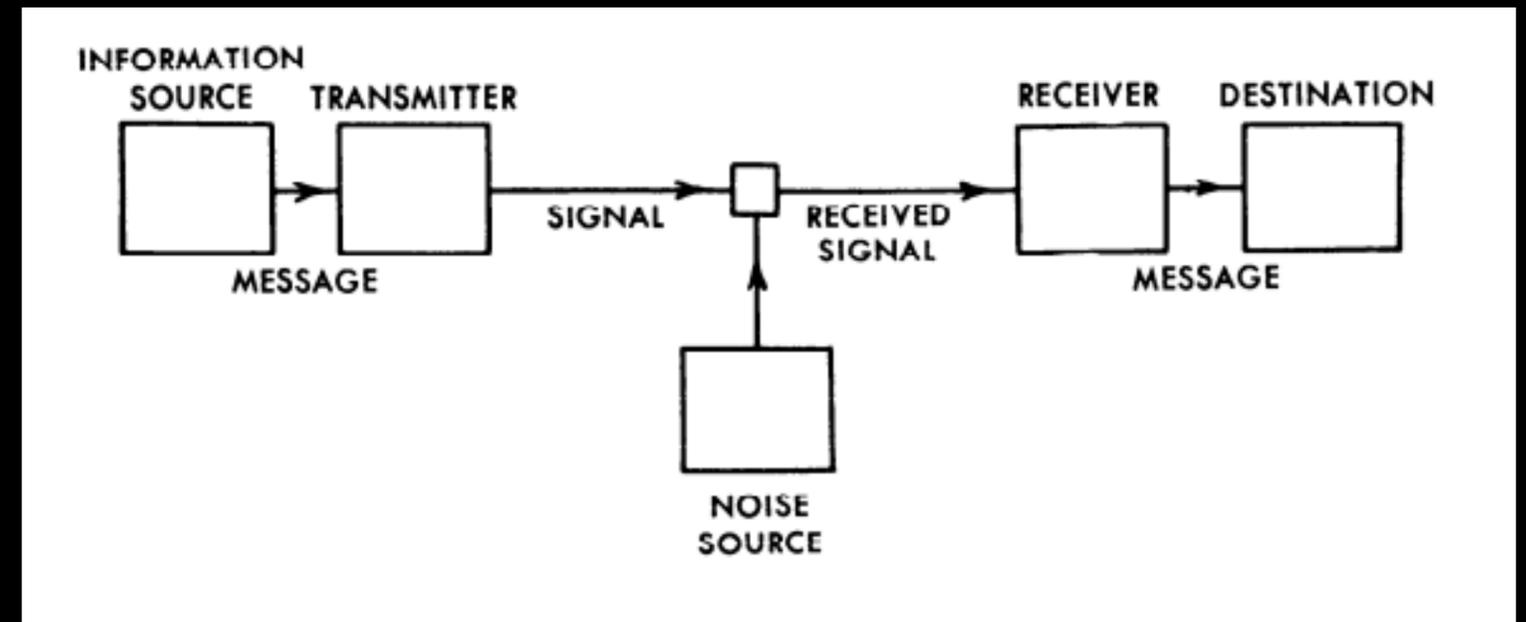
The mathematical theory of communication, Shannon [3]

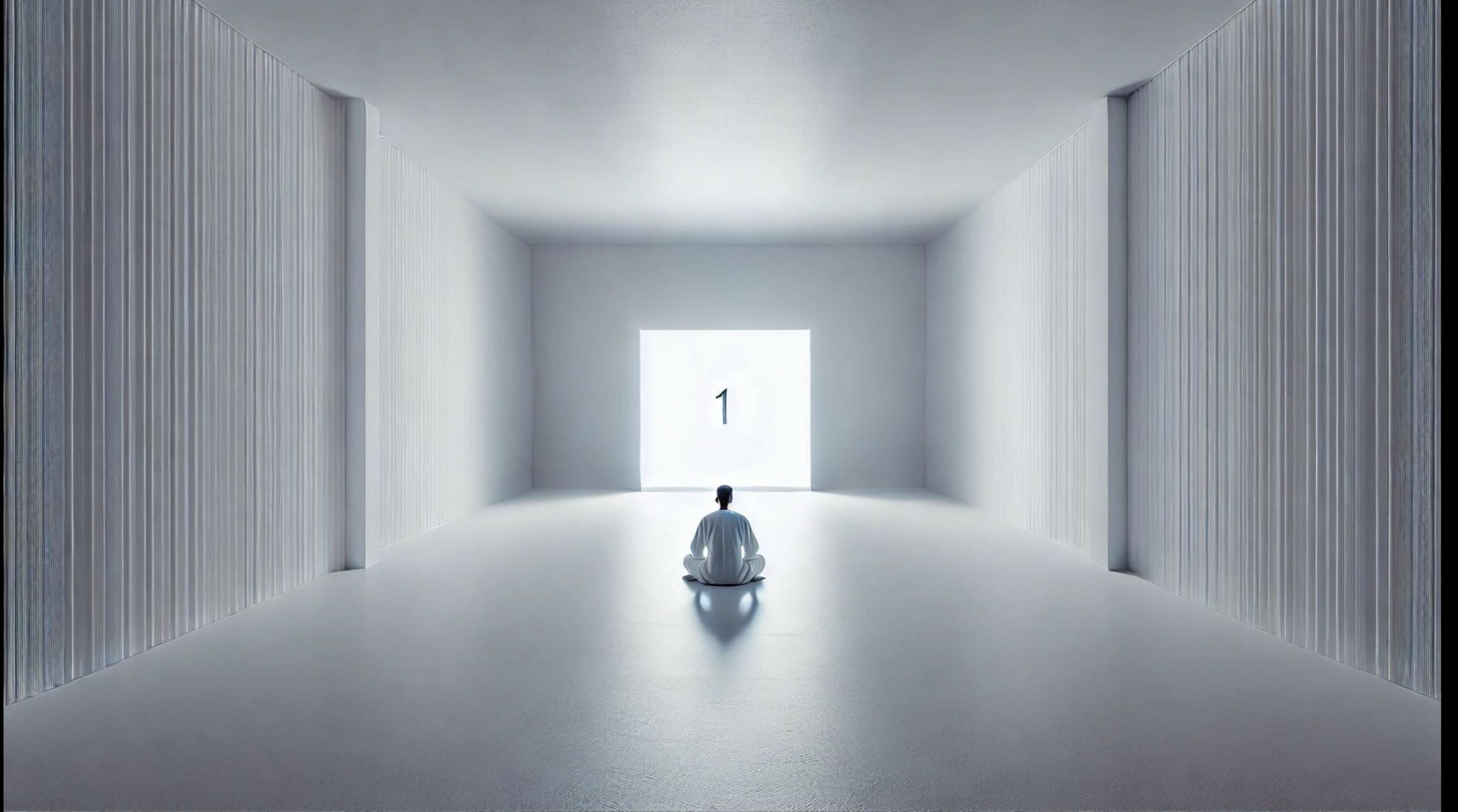
The transducer (encoder/decoder) can be described by two functions:

$$y_n = f(x_n, \alpha_n)$$

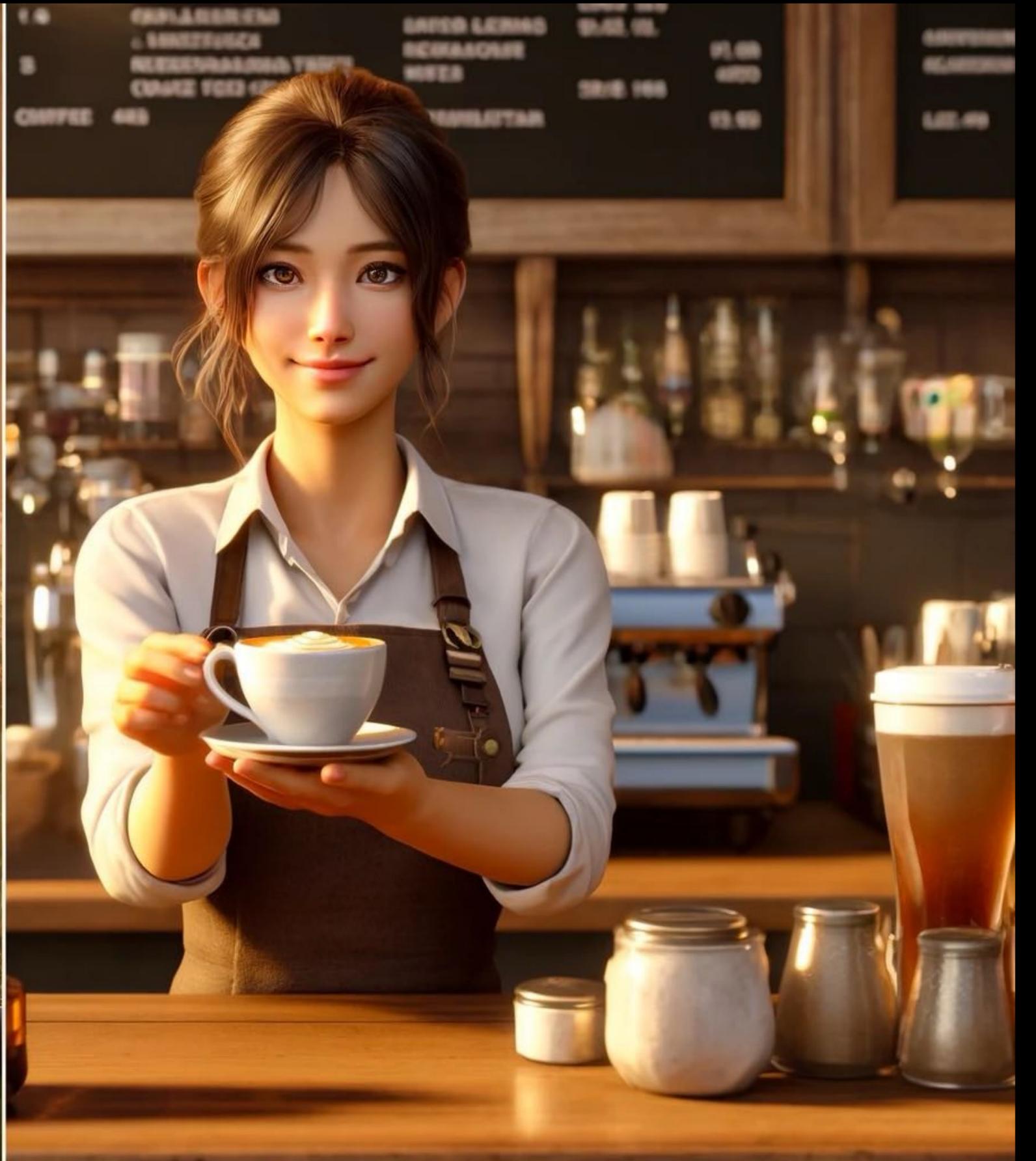
$$\alpha_{n+1} = g(x_n, \alpha_n)$$

Where x_n is the n th input, α_n is the state of the transducer, and y_m is the output signal.





A single bit of information lands differently, depending on state α



Homomorphism

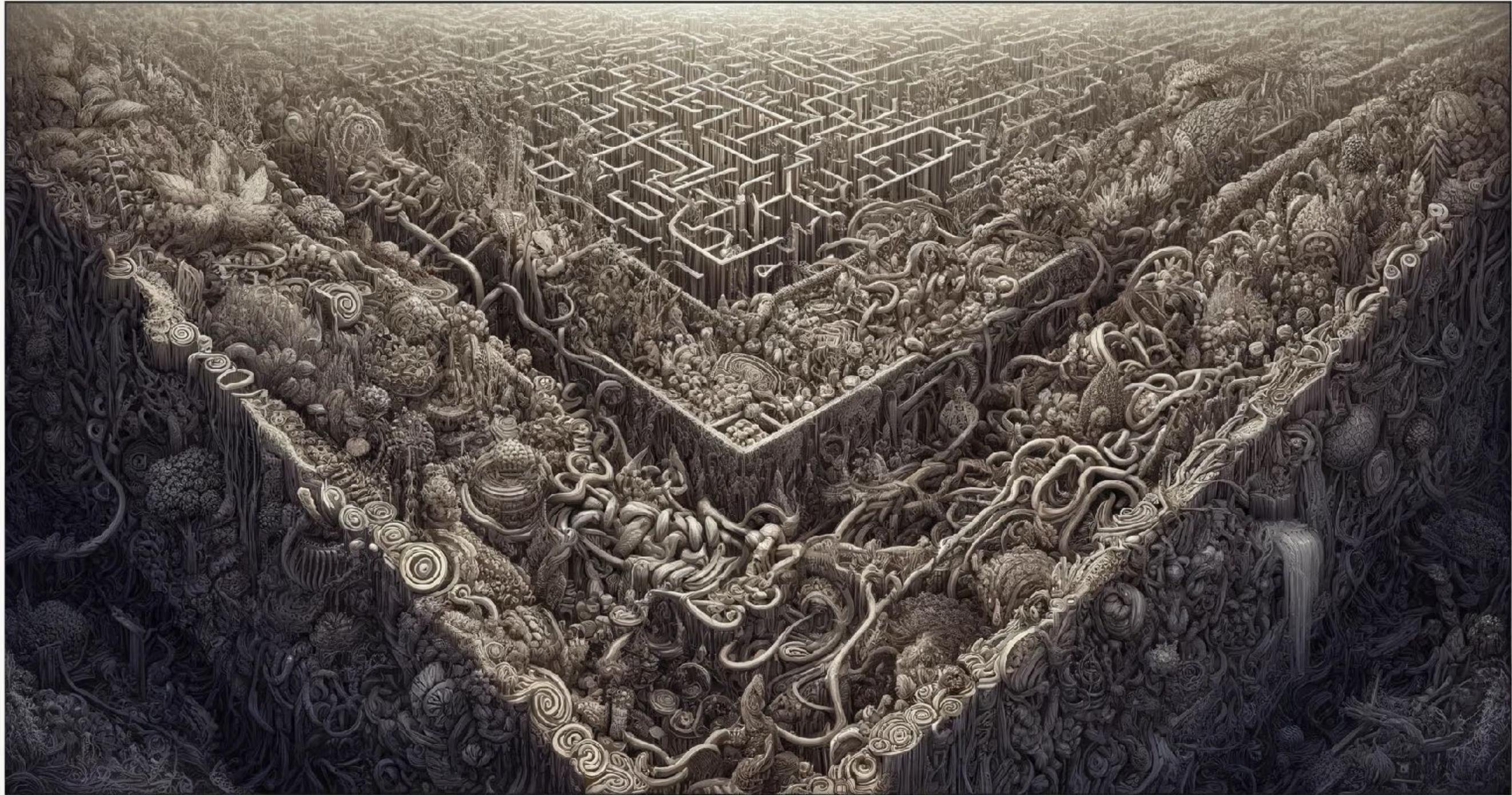
- A homomorphism is a mapping that preserves algebraic structure between domain and range
- $\phi(x \star y) = \phi(x) \star \phi(y), \forall x, y$

Homomorphism

- For example, lets take
 - $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
 - $\mathbb{Z}_3 = \{0, 1, 2\}$, the set of integers modulo 3
 - $\phi: \mathbb{Z} \rightarrow \mathbb{Z}_3, \phi(x) = x \bmod 3$
 - $\phi(5) + \phi(7) = \phi(5 + 7) = 0$
- ϕ is a homomorphism between \mathbb{Z} and \mathbb{Z}_3 for addition because the structure that matters is preserved.



Reading is a homomorphism between ideas



Lounen mns/hd' - 011ce | BtSurel

Geineahioeemethod |



To/tonle level

Some maps are less useful...



...than others

Model of the Earth, Leonardo Da Vinci

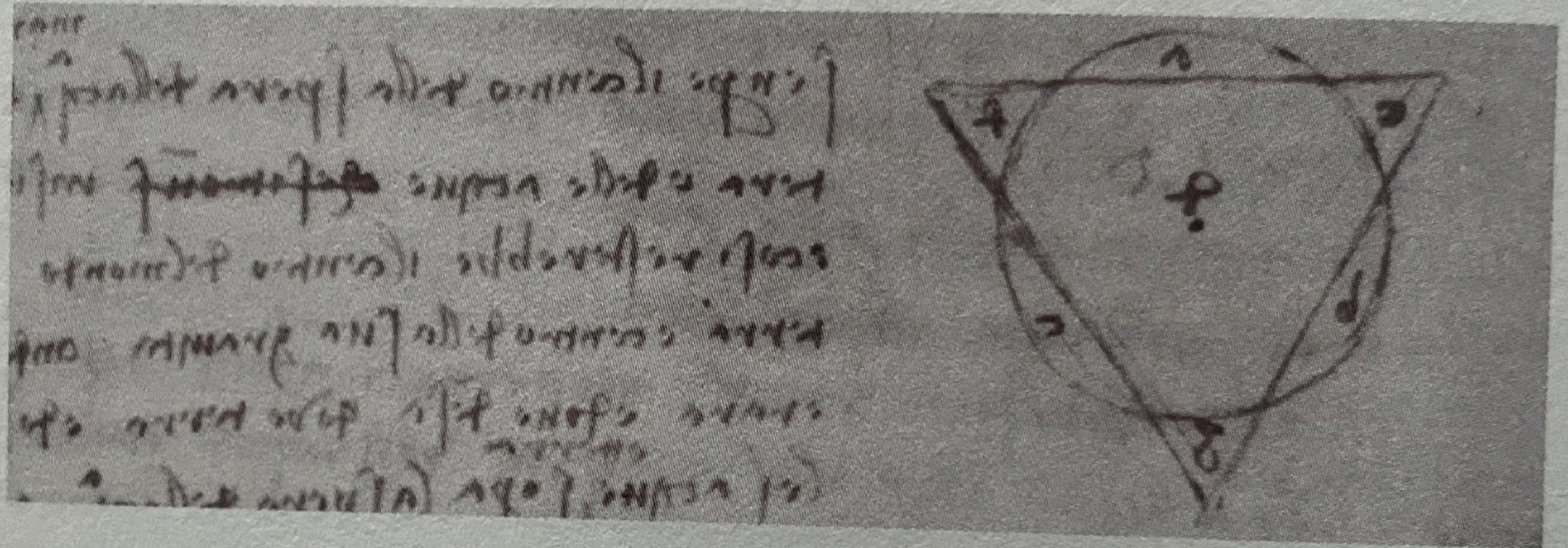


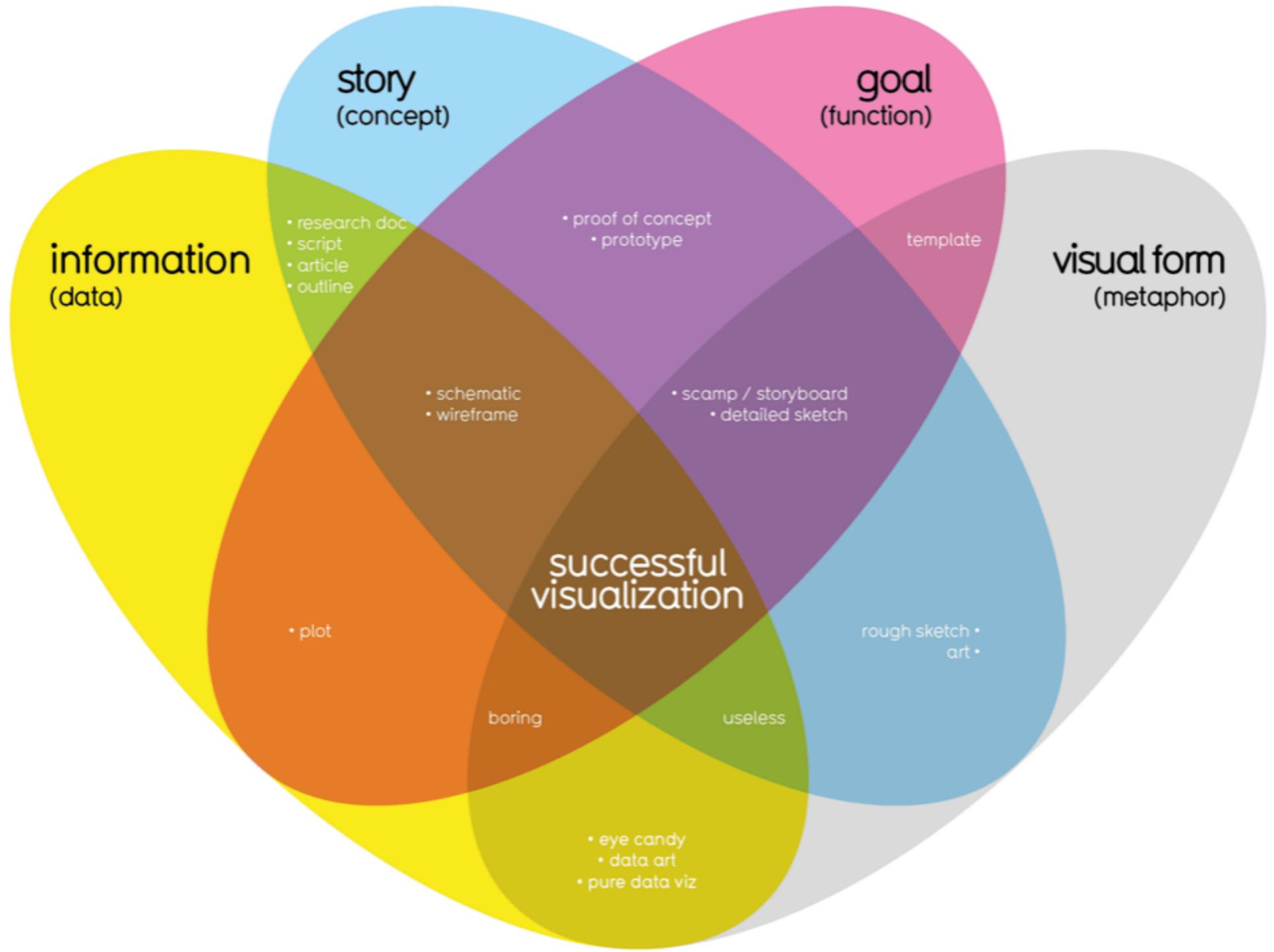
FIG. 2-7. Geometric model of the Earth.
Codex Leicester, folio 35v (detail).

Models don't need to be "correct" to be useful or insightful.
Just homomorphic enough

rollover for more detail

What Makes a Good Visualization?

explicit (implicit)



Gestalt principles

GESTALT PRINCIPLES

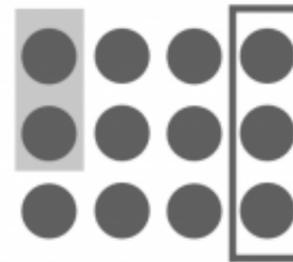
Proximity



Similarity



Enclosure



Connection



Continuity



Closure



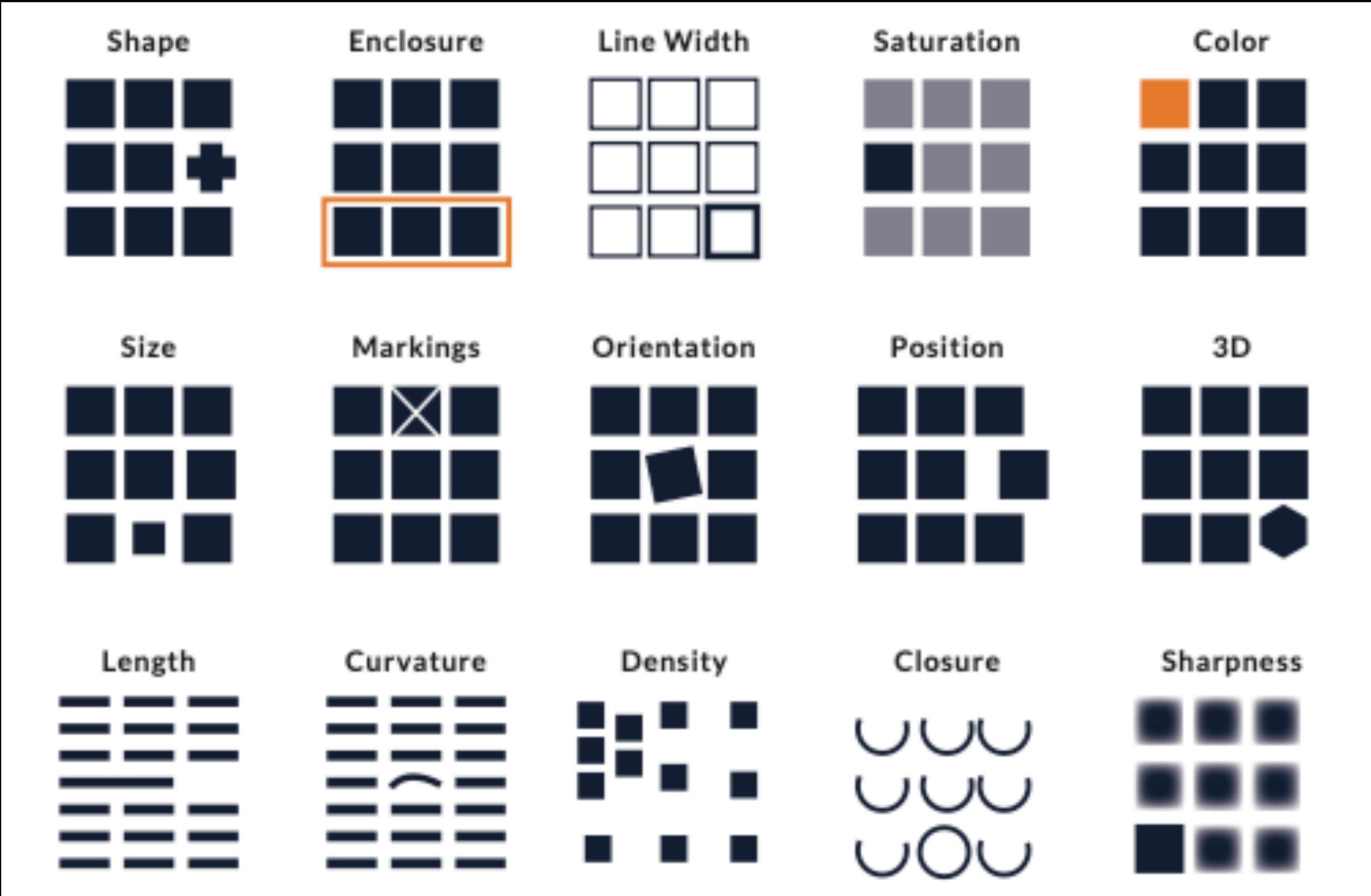
Figure-Ground



Common Fate



Preattentive processing



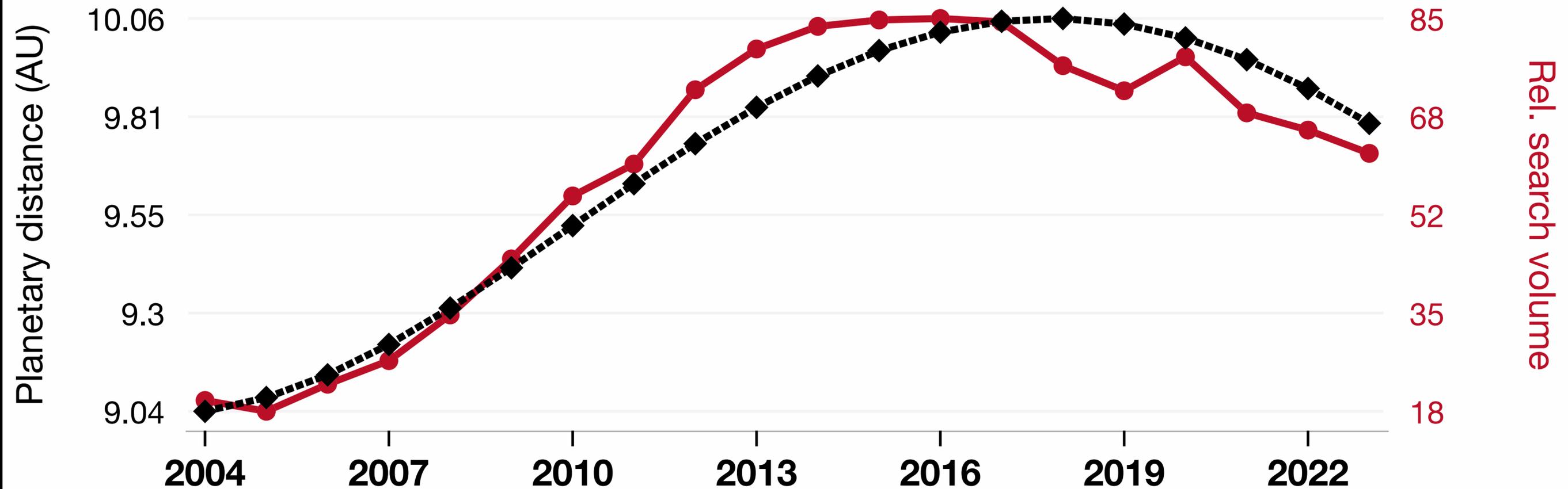
Five guidelines

0. Avoid spurious correlations
1. Show the data
2. Reduce the clutter
3. Integrate graphics and text
4. Avoid the spaghetti chart
5. Start with grey

The distance between Saturn and the Sun

correlates with

Google searches for 'how to make baby'



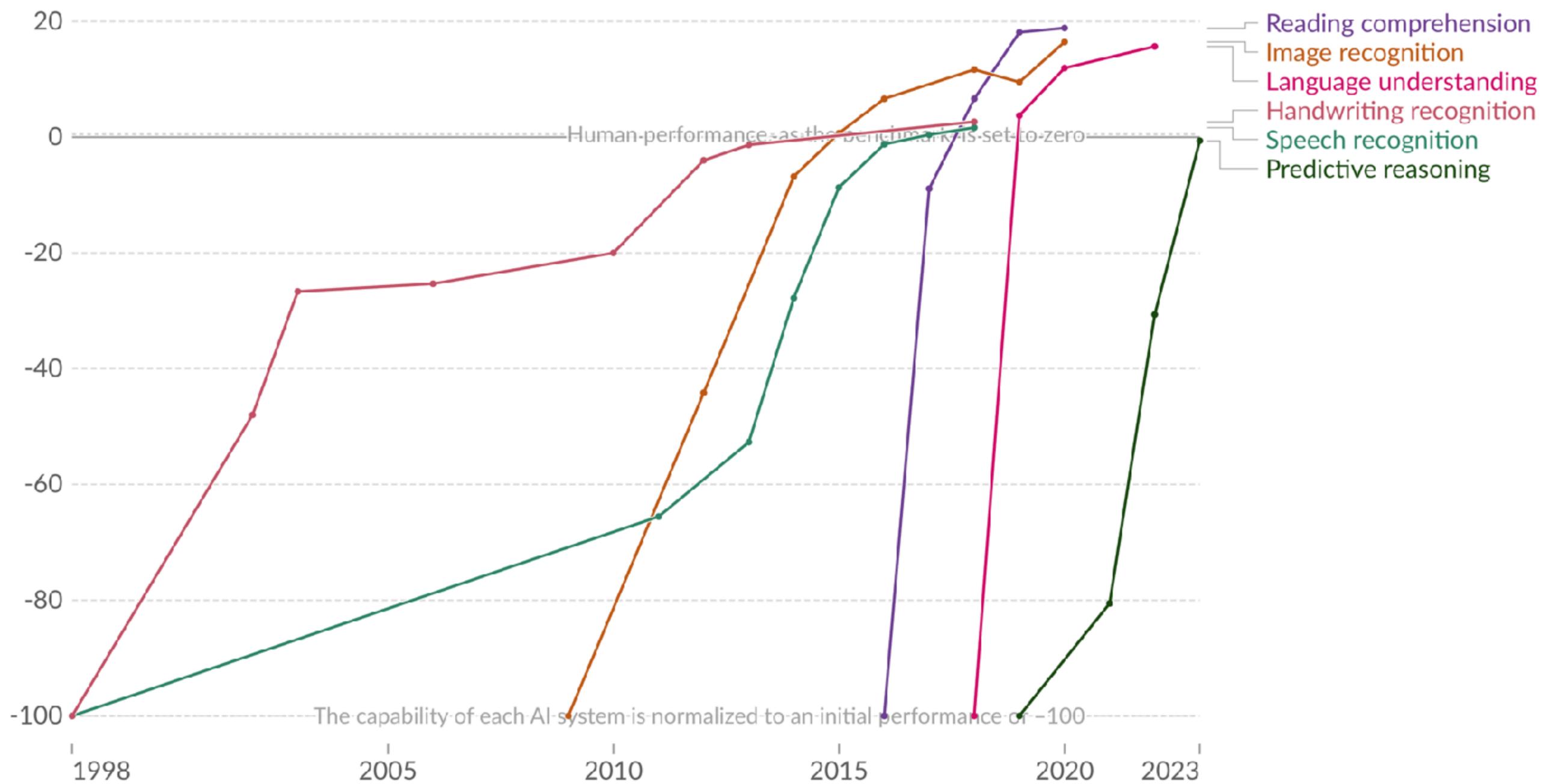
◆ The average distance between Saturn and the Sun as measured on the first day of each month · Source: Calculated using Astropy

● Relative volume of Google searches for 'how to make baby' (Worldwide, with quotes) · Source: Google Trends

2004-2023, $r=0.967$, $r^2=0.935$, $p<0.01$ · tylervigen.com/spurious/correlation/1522

Test scores of AI systems on various capabilities relative to human performance

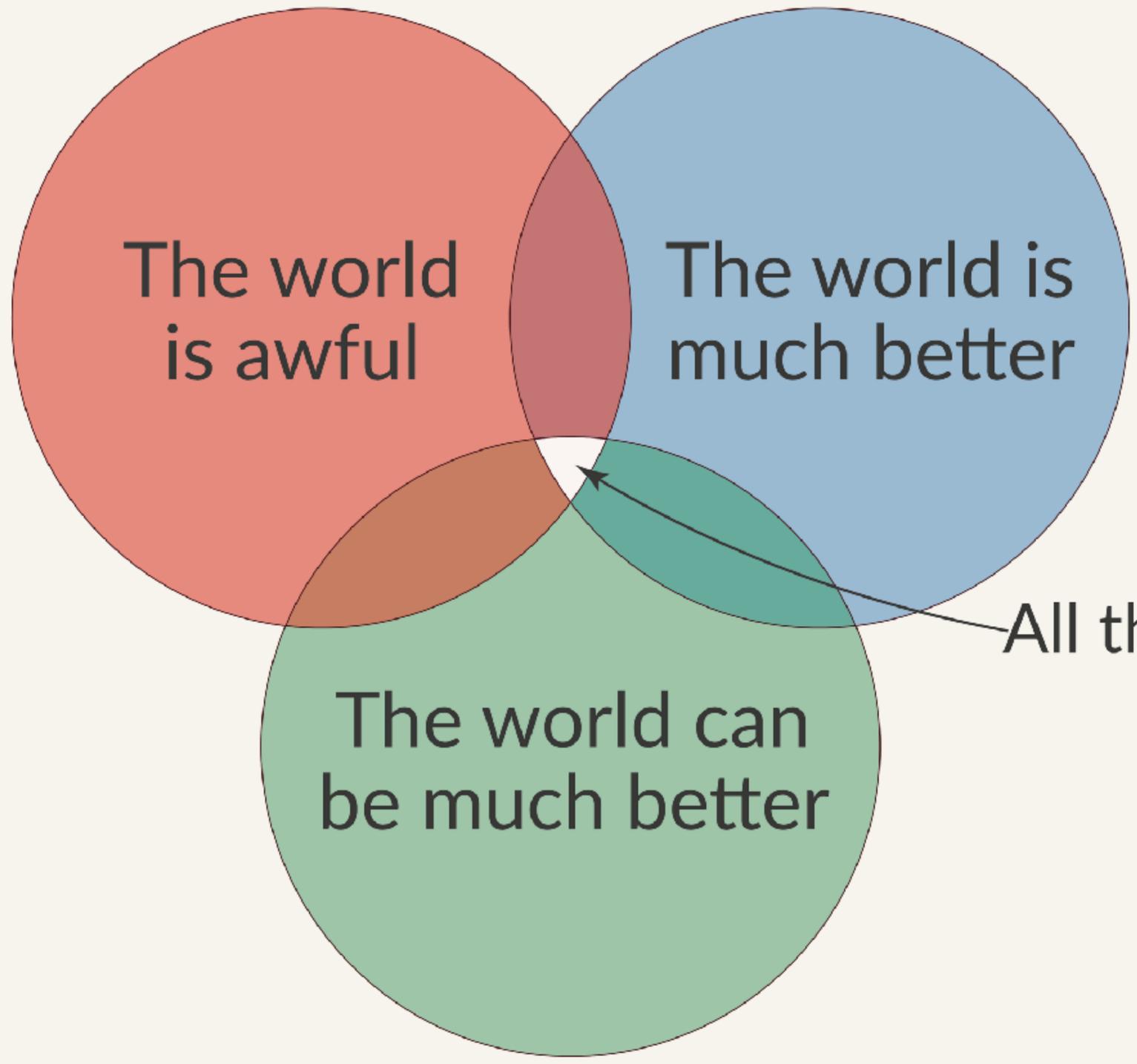
Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

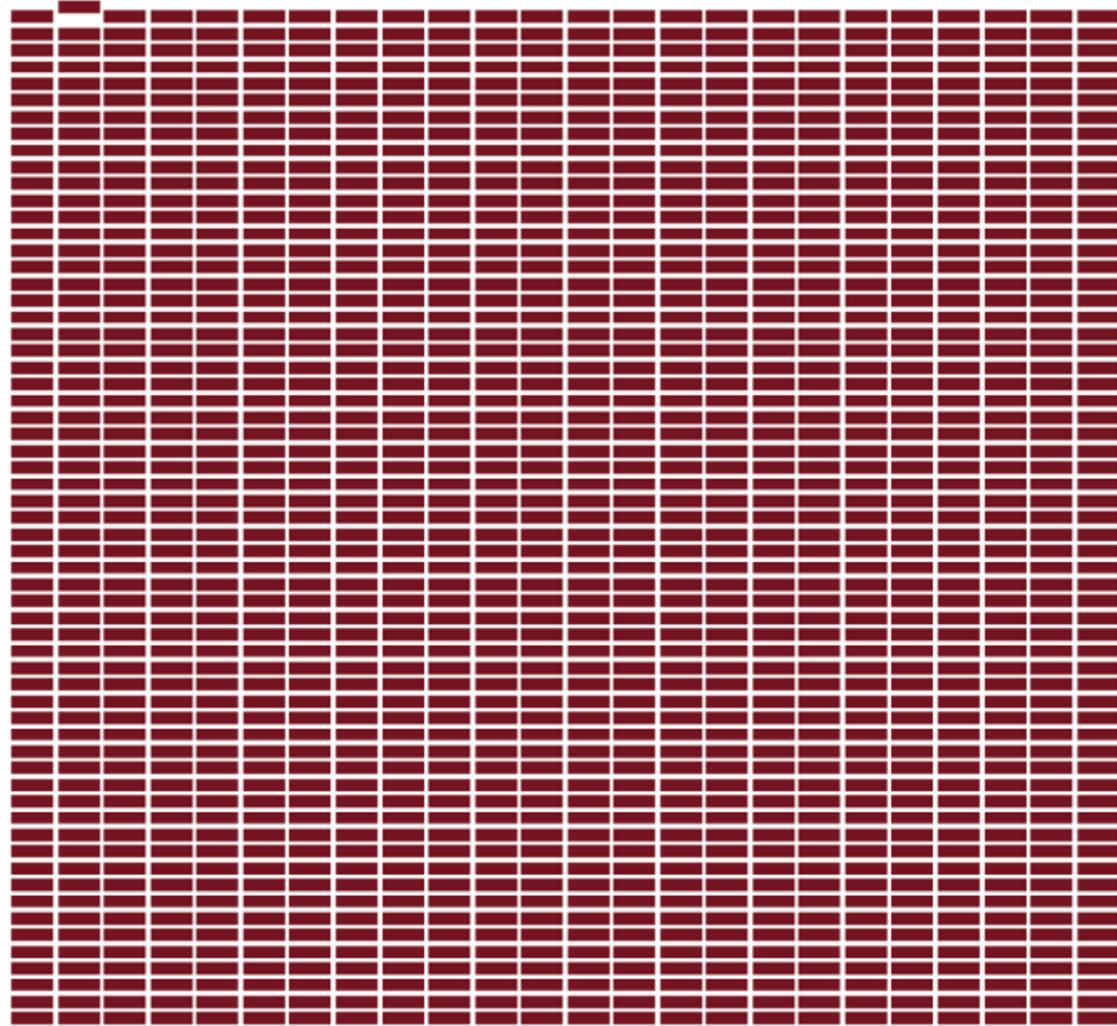


All three statements are true at the same time

The world is awful

Globally 4.3% of all children die before they are 15 years old.
This means every day 16,000 children die.

Each rectangle represents the death of 11 children. That's the number of children who die every minute.



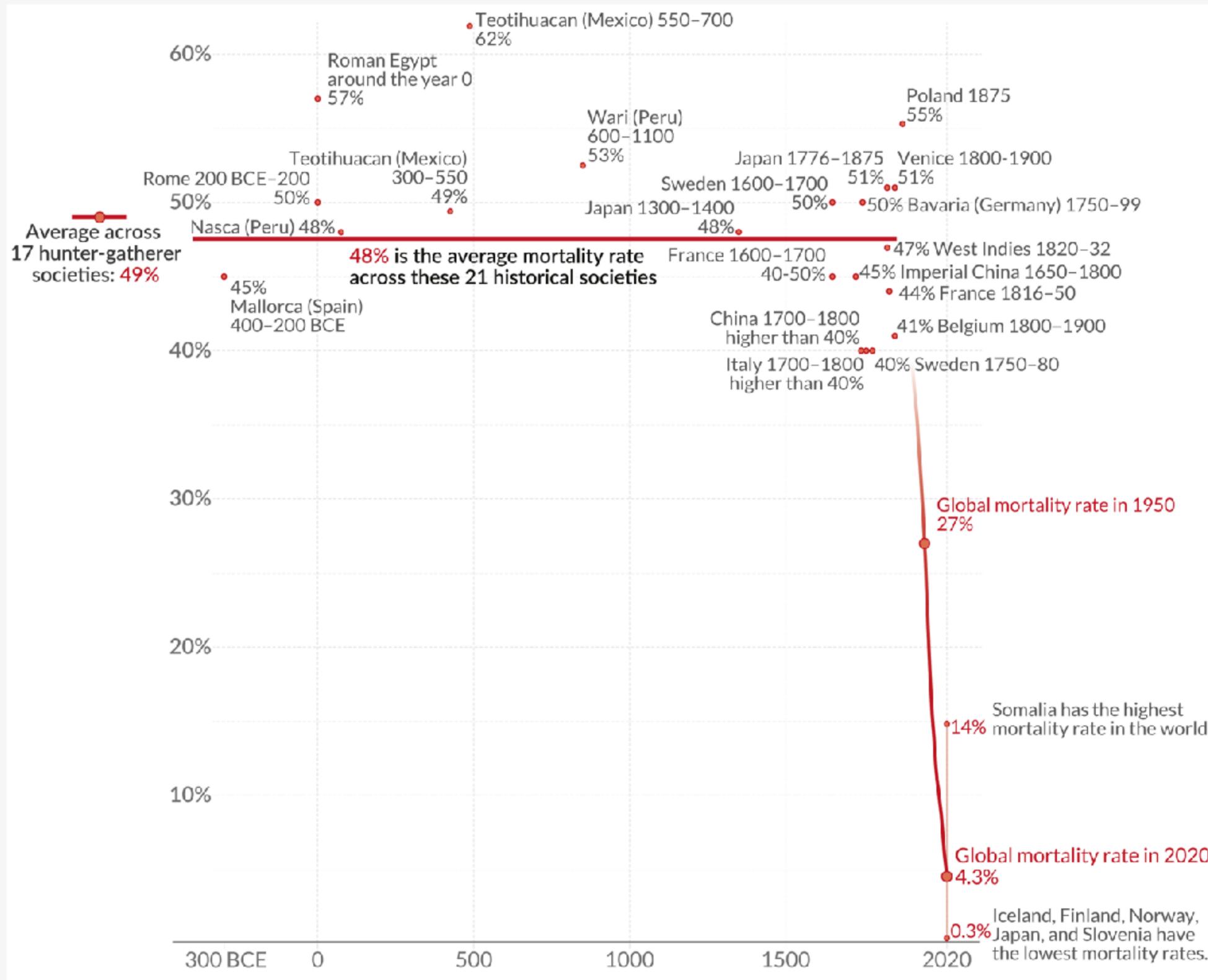
Data: 2020 data from UN IGME.

[OurWorldinData.org](https://ourworldindata.org) - Research and data to make progress against the world's largest problems.

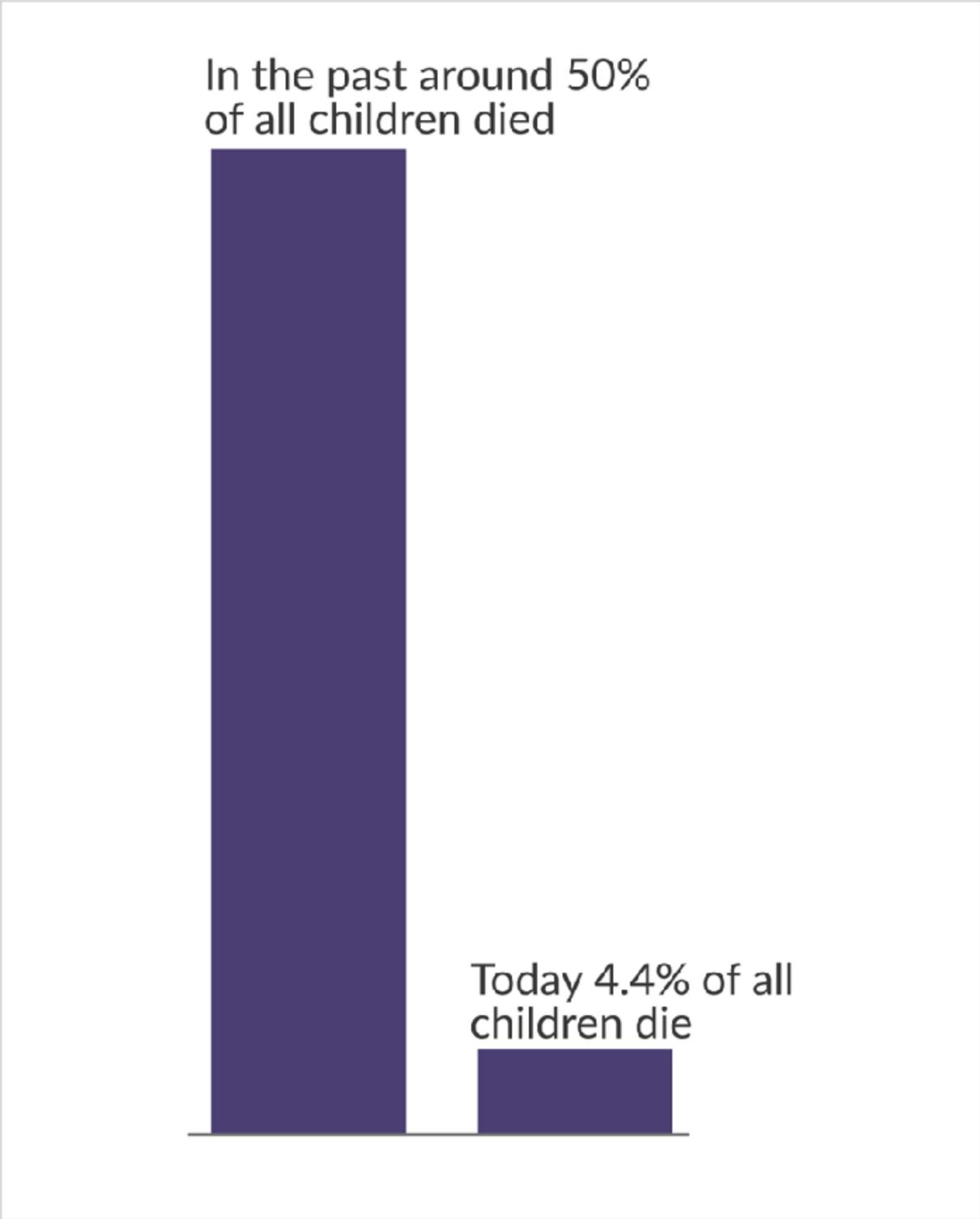
Licensed under CC-BY by the author Max Roser

The long-run history of child mortality

Shown is the share of children who died before reaching the end of puberty.
The exact age cut-off differs slightly between studies, but is around the age of 15.



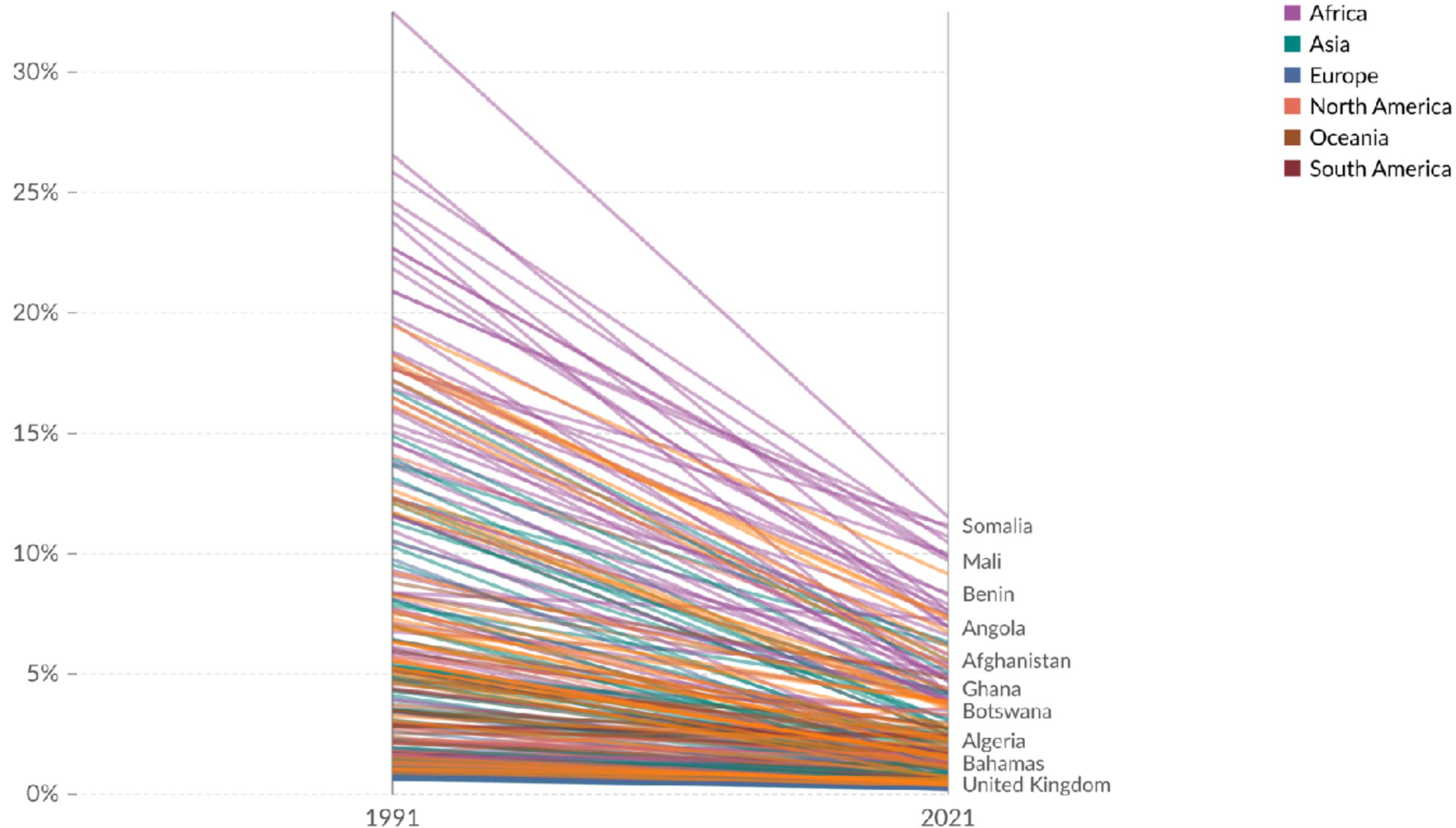
The world is much better



Data: 2020 data from UN IGME and historical data from Volk and Atkinson (2013).
OurWorldinData.org - Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the author Max Roser

Child mortality rate, 1991 to 2021

The change in the estimated share of newborns¹ who die before the age of five.

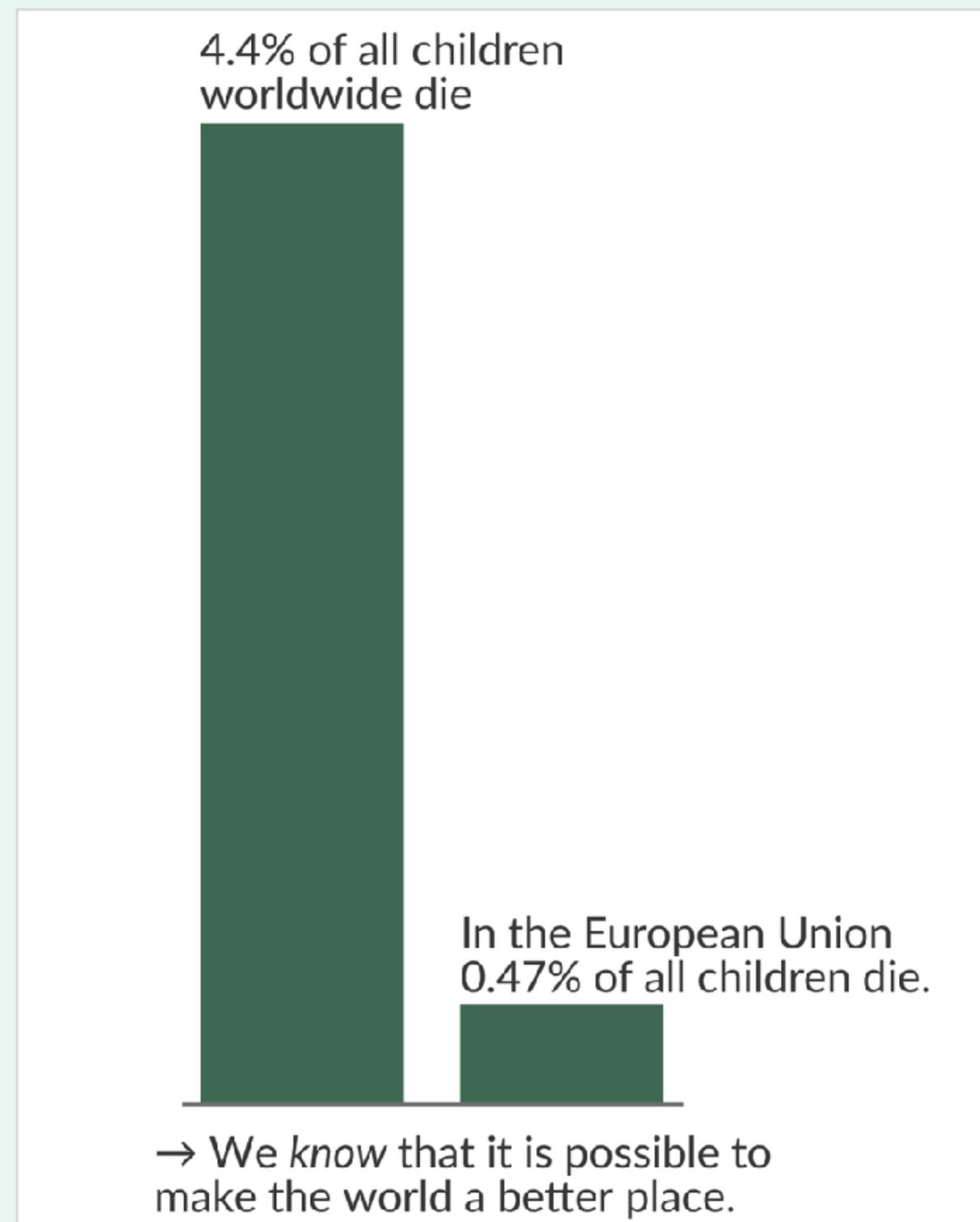


Data source: United Nations Inter-agency Group for Child Mortality Estimation (2023)

OurWorldInData.org/child-mortality | CC BY

1. Newborn: A newborn is defined as a baby born alive, and usually refers to neonates – under 28 days old. Read more in our article: [How do statistical organizations define age periods for children?](#)

The world can be much better



Regular Expressions

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.

OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!



BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!

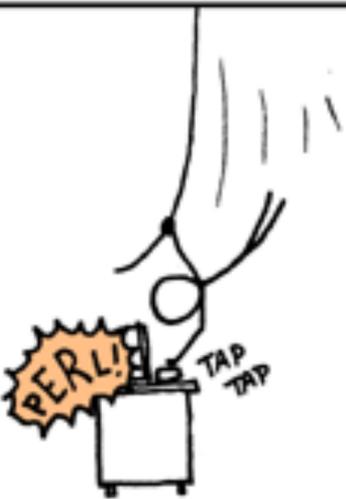


IT'S HOPELESS!

EVERYBODY STAND BACK.



I KNOW REGULAR EXPRESSIONS.



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
- LLMs are pretty good at creating and explaining regexes.
- <https://www.alf.nu/RegexGolf> if you like some puzzles

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 not 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?

```
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 [0 1])`: Matches days from 1–31, allowing for a leading zero.
- `(0 ? \d | 1 [0 1 2])`: Matches months from 1–12, also allowing for a leading zero.
- `(\d { 4 })`: Matches any four digits for the year.
- `([0 1] ? \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.