

# The ABC of Computational Text Analysis

## #2 TEXT AS DATA

Alex Flückiger

Faculty of Humanities and Social Sciences  
University of Lucerne

March 10, 2023

# Outline

- recap + reading
- methodological foundation 😊
- first computational text analysis

# Recap last Lecture

## computer as ...

- ... an intelligent device
- ... a tool for a *new social science*

## datafication

- abundance of data
- exploit new form of data

# Reading

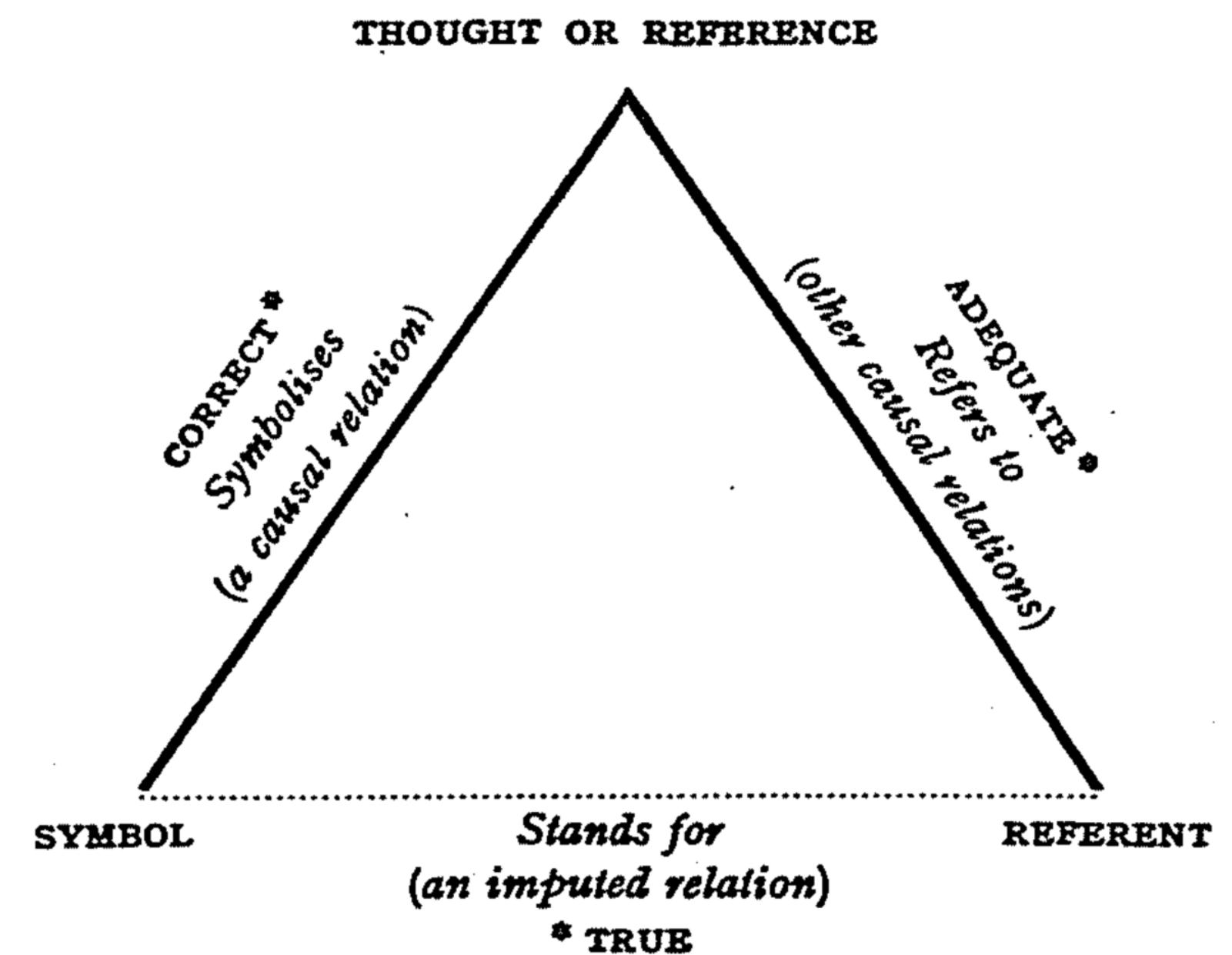
Computational Social Science (Lazer et al. 2009)

- data-driven
- network analysis + text analysis
- historical perspective vs. real-time dynamics
- issues: limited access to data and new methods

# Semiotic Triangle

## Loose coupling between

- World
- Cognition
- Language



*Semiotic Triangle* (Ogden and Richards 1923)

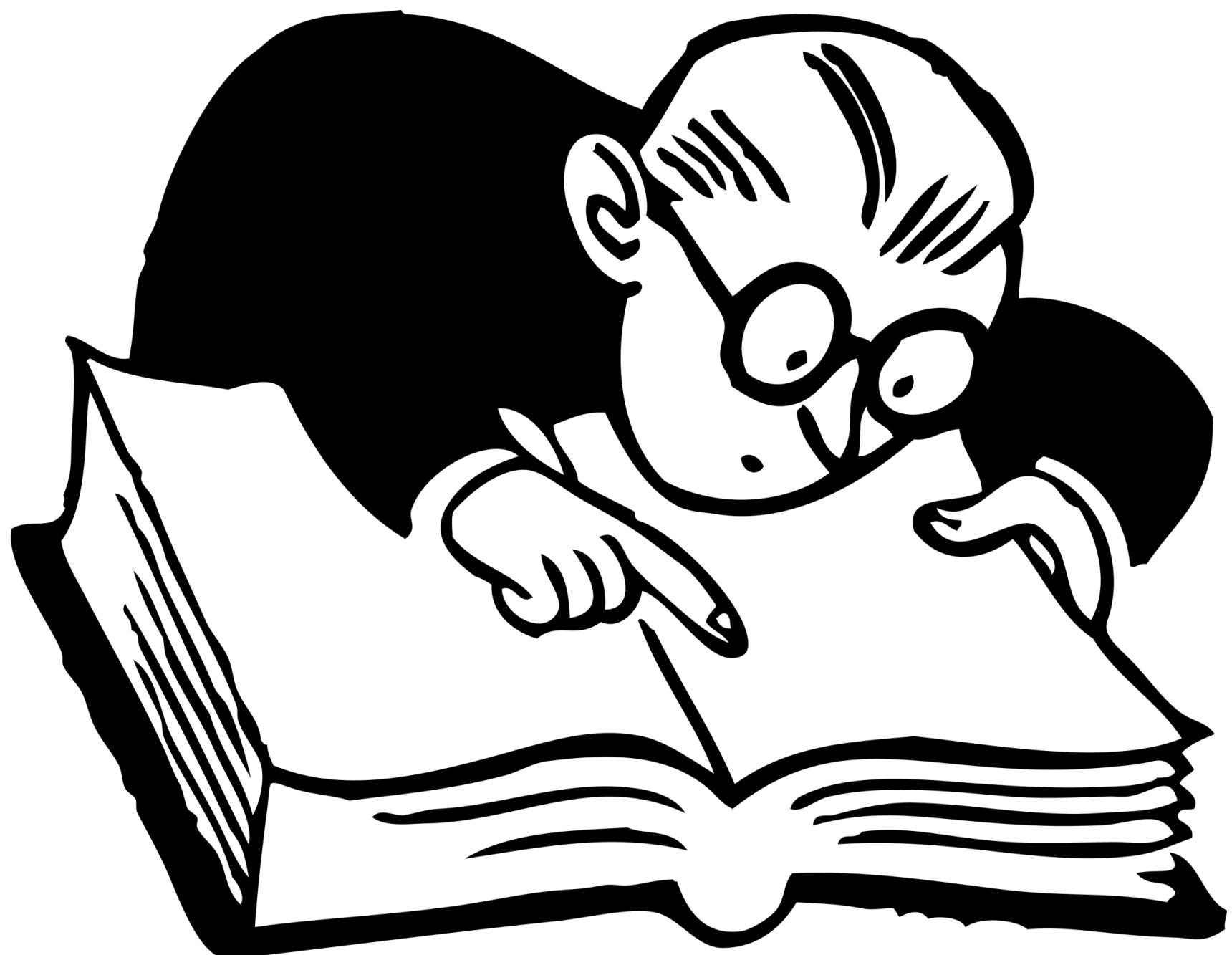
«Language shapes the way we think,  
and determines what we can think about.»

—**Benjamin Lee Whorf**

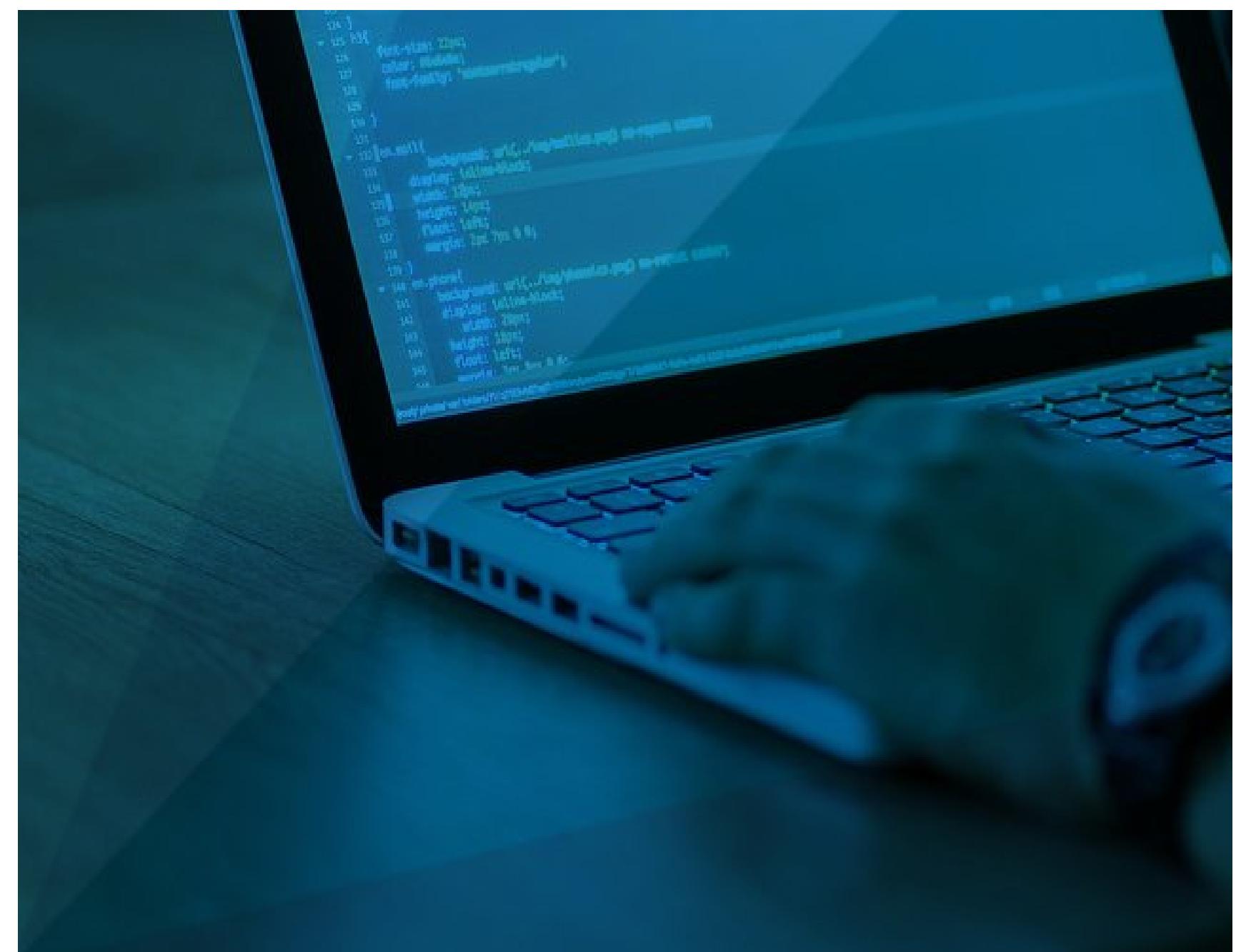
# Working with Texts

# A micro and macro perspective I

individual cases vs. collective trends



*close reading (src)*



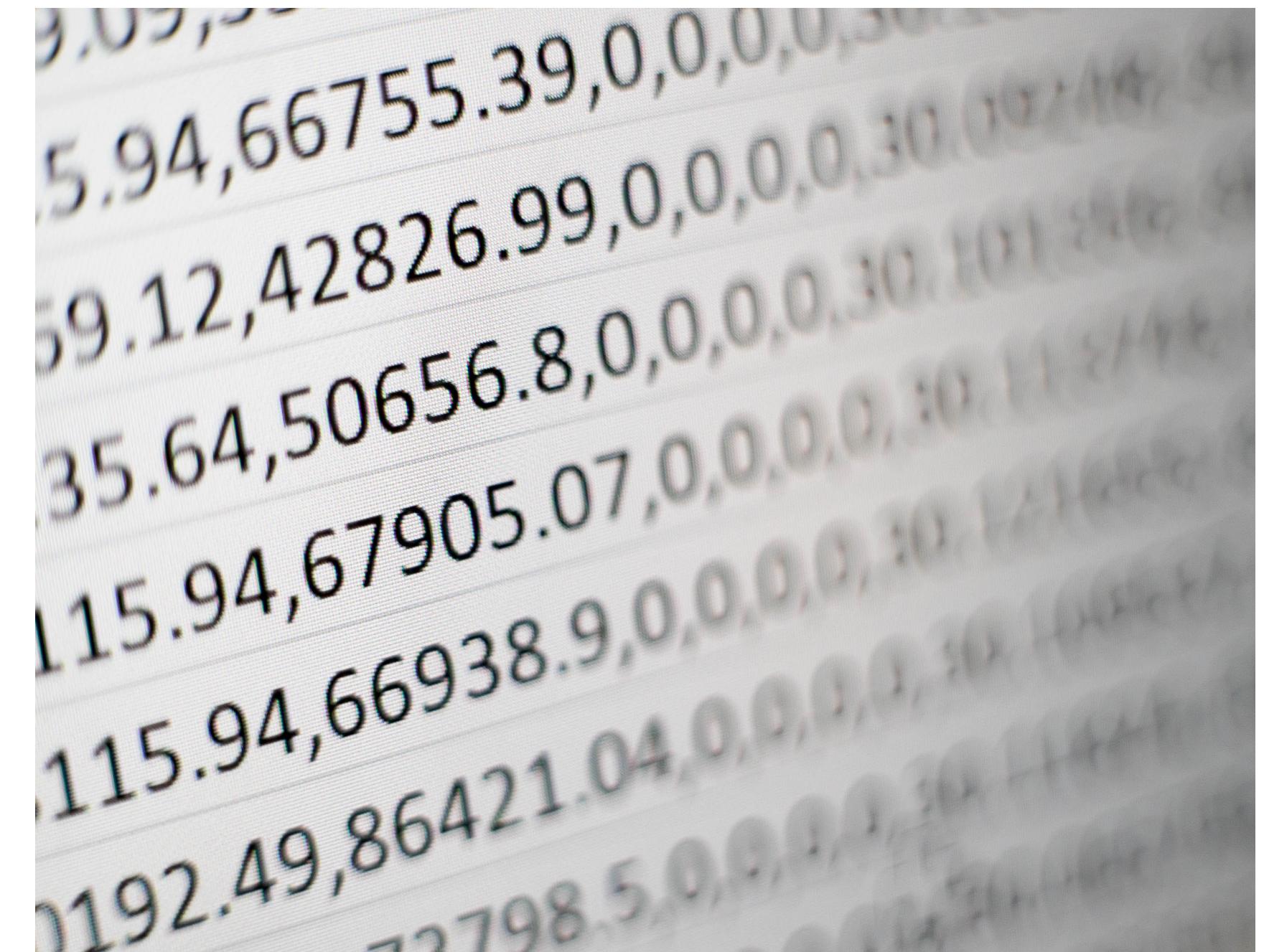
*distant reading (src)*

# A micro and macro perspective II

## scalability vs. abstraction

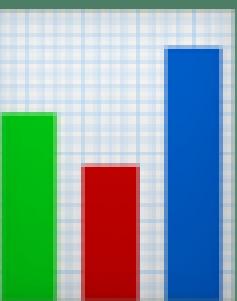


*tons of text (src)*



*meaning of numbers (src)*

From micro to macro  
...and back again



# Two Research Paradigms

data exploration vs. hypothesis testing (Evans and Aceves 2016)

- add nuance
- develop new narratives
- verify hypothesis

Numbers do not talk



Thus, quantification and qualitative analysis go well together.

# Text as Data

- synonymy
- ambiguities
- compositionality of meaning
- discrete symbols
- unstructured, messy data

(see also Grimmer and Stewart 2013)

# Unstructured Text? 🤔

collection > documents > paragraphs > sentences > words



*Texts carry more meaning than a soup (=set) of words*

# Data Formats

# In-class Task: File Types

- What file formats do you know?
- Open files of different types in a text editor.  
Which ones look good?

# File Formats

- machine-readability
  - raw: txt, csv, tsv
  - formatted: docx, pdf, html, xml
- open vs. proprietary
- digital sustainability

Let's Dive into it! 

# Counting ngrams

**Google Ngram Viewer** (Michel et al. 2011)

- historical perspective with ngrams
- >5.2 million books
- rise and fall of cultural ideas and phenomena

# In-Class Task: Investigate the Environmental Discourse

- What other terms have been used to describe nature?  
e.g. environment
- What environmental issues are debated the strongest? When?  
e.g. nuclear power plant
- Are there any differences between languages?  
i.e. similar words with non-equivalent curves over time



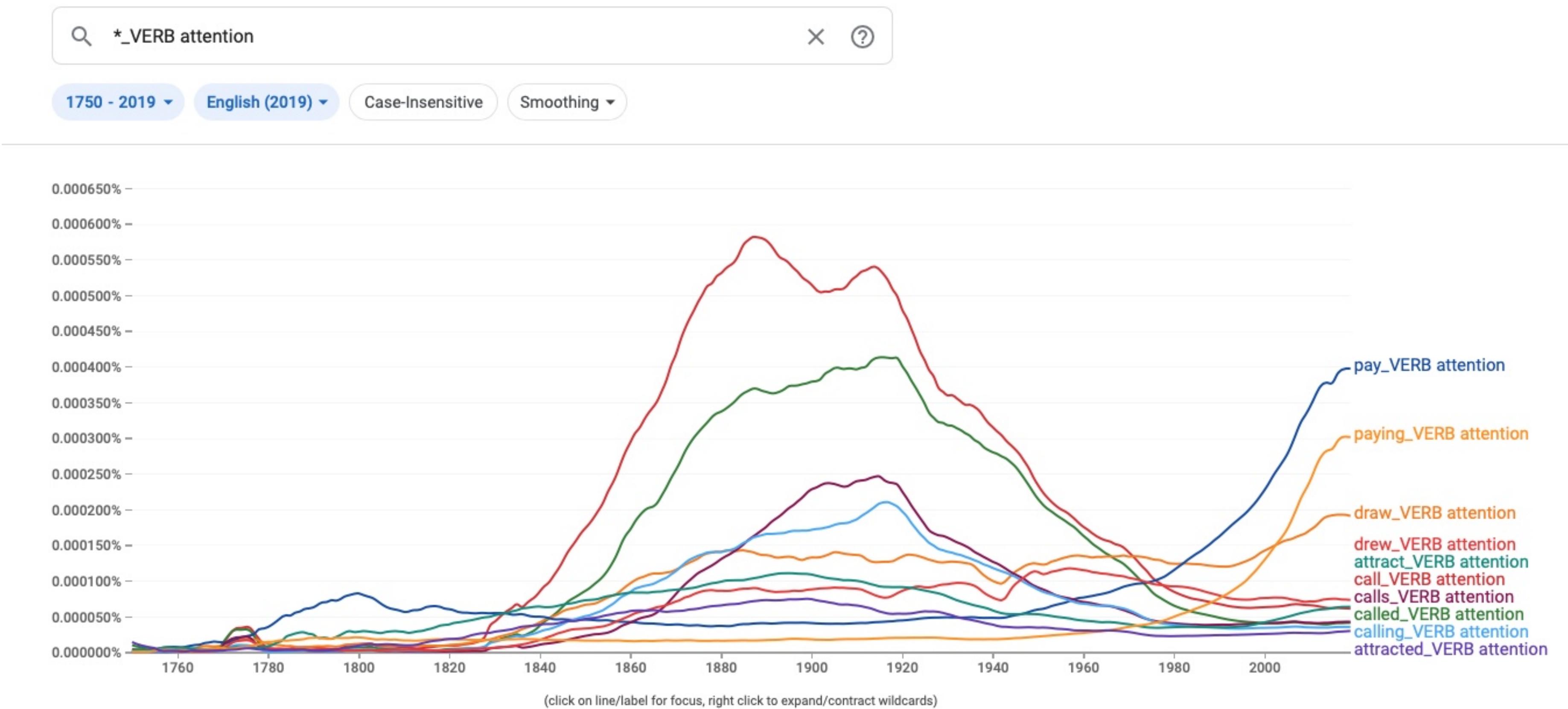
**What do you conclude from your observations?**

# Refine your Queries

check out case-sensitivity, wildcards (\*) and operators 

Operator	Description
+	sums multiple expressions into one to <b>aggregate trends</b> .
-	subtracts the expression on the right from the expression on the left, giving you a way to <b>measure one ngram relative to another</b> .
/	divides the expression on the left by the expression on the right, which is useful for <b>isolating the behavior of an ngram with respect to another</b> .
*	multiplies the expression on the left by the number on the right, making it easier to compare ngrams of very different frequencies. (Be sure to enclose the entire ngram in parentheses so that * isn't interpreted as a wildcard.)

# The raise of the Ngram pay attention



*Google Ngram Viewer: Evolution of the phrase 'attention'*

Remember 

Has the language evolved over time or the social perception?

**Both, most likely.**

Similarly, language may vary across regions and communities.

# No Culturomics but Meaning-Making

**phenomena in collective memory**

- semantic drifts (meaning)
- lexical shifts (frequency)

**Read, read, read** to complement **stats** with context!

# Questions of Interpretation

## possible reasons of decreasing frequency

- loosing interest
- becoming an established fact
- new reference

The Great War → World War I

- selection of data sources

# A Word of Caution

## The unknowns of Google Ngram Viewer

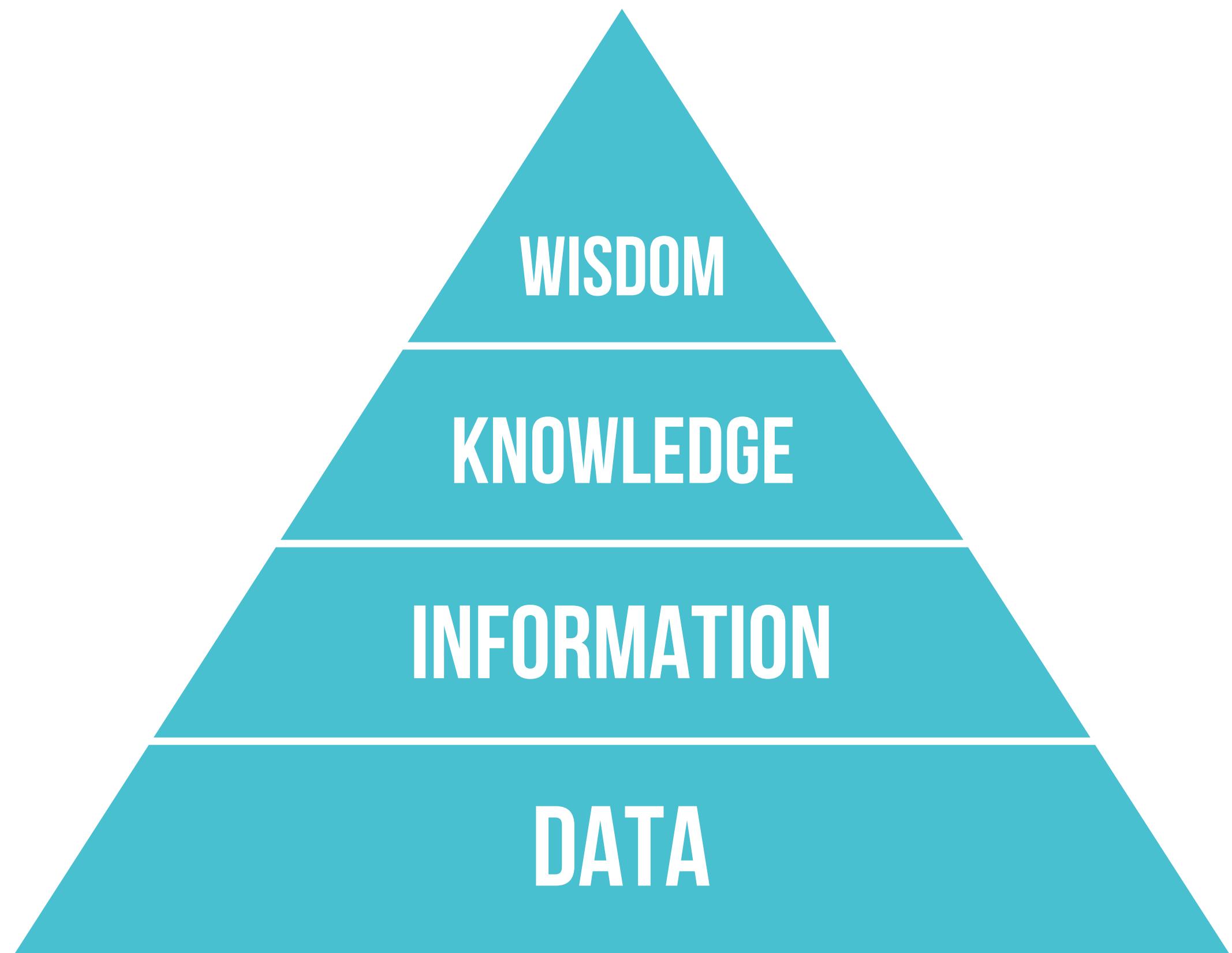
- index of books
  - genre, authors, quantity
- artifacts of digitalization



use better alternative: [bookworm HathiTrust](#)

# Interacting with Data

It is a lense, not a map.



*DIKW pyramid (Wikipedia)*

# Prepare your System

1. backup files + update system 
2. start installation with this [guide](#) 



Questions?

# References

- Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42 (1): 21–50. <https://doi.org/10.1146/annurev-soc-081715-074206>.
- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. <https://doi.org/10.1093/pan/mps028>.
- Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23. <https://doi.org/10.1126/science.1167742>.
- Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. <https://doi.org/10.1126/science.1199644>.
- Ogden, Charles Kay, and Ivor Armstrong Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism. Supplementary Essays by B. Malinowski and F.G. Crookshank*. New York: Harcourt. <https://books.google.com?id=i3MIAQAAIAAJ>.