GET1030

Computers and the humanities

Lecture 2

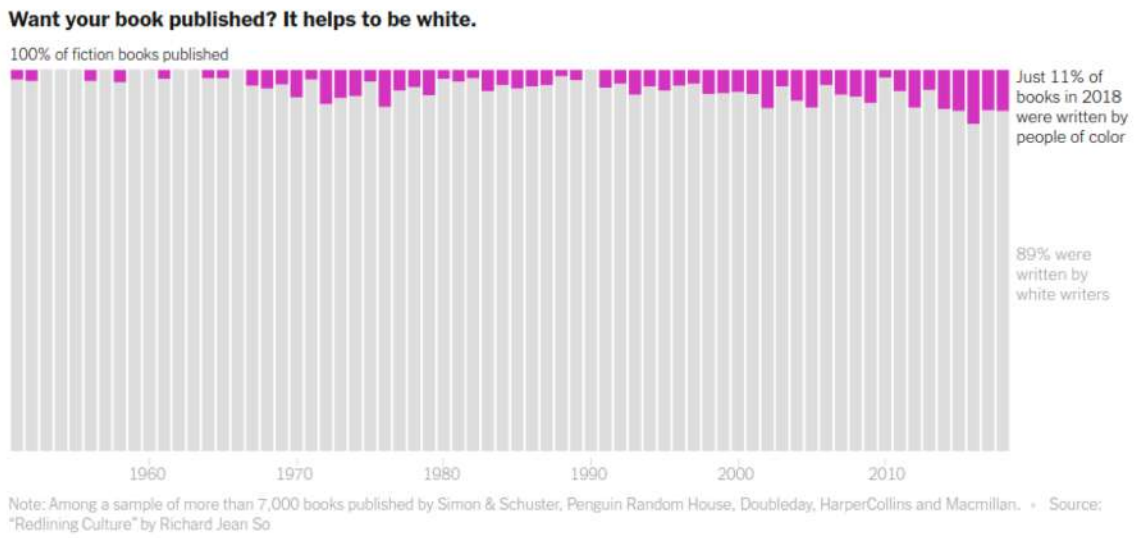
What is data?

Learning Objectives

- To define data
- To consider the relationship between data and interpretation

The publishing industry

https://www.nytimes.com/interactive/2020/12/11/opinion/culture/diversity-publishing-industry.html

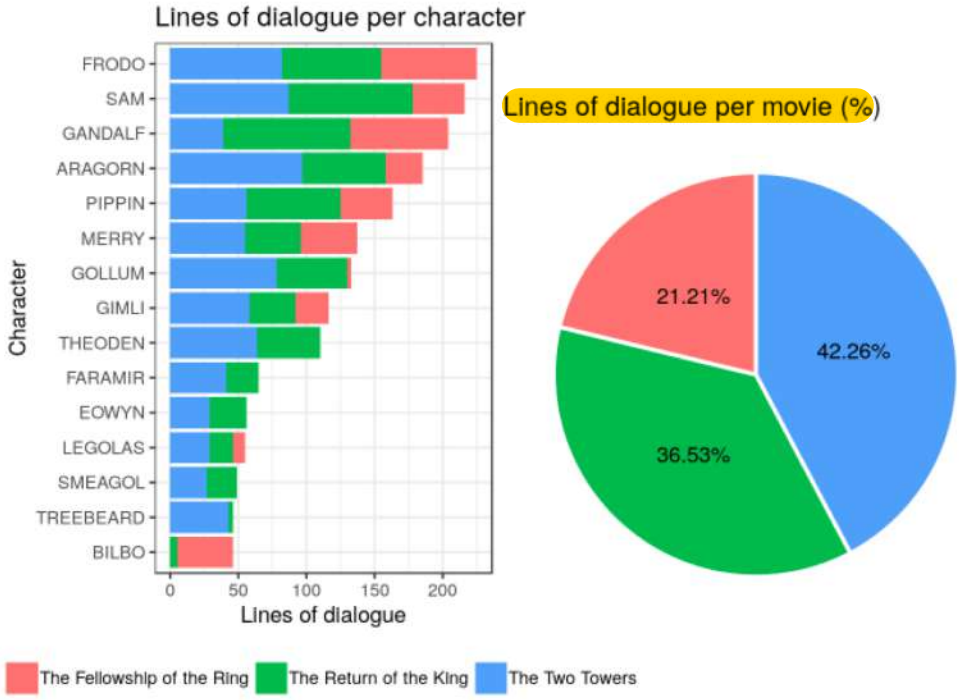


Source: "Redlining Culture" by Richard Jean So.

What data would you need to produce this graph?

Treating LOTR as data

- Do a visualization to show the lines of dialogue per character.



https://www.kaggle.com/xvivancos/analyzing-the-lord-of-the-rings-data

Using data

In this module we will only use spreadsheets to organize our data.

Each row has an object.

Each column has data for one feature of this object.

Decisions

In both examples, we the designers of the visualizations made some decisions.

These decisions are interpretations, they demonstrate what they think is important.

But alternative decisions would have lead to very different visualizations.

GET1030

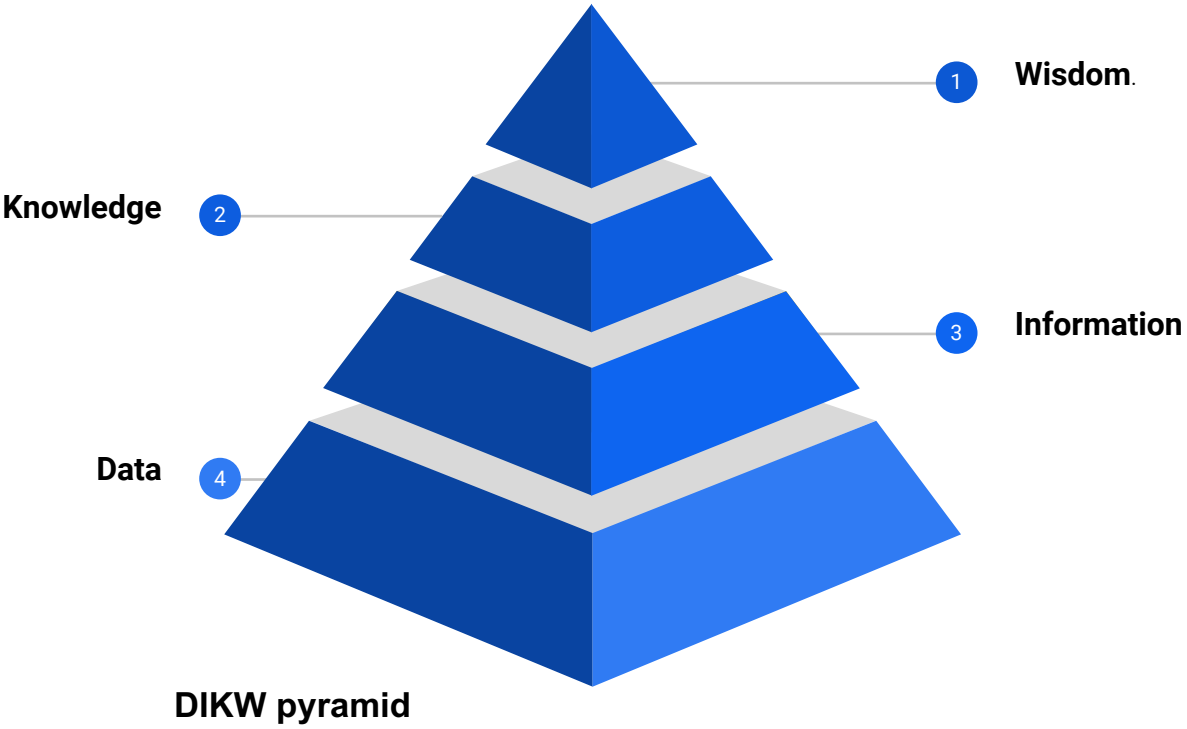
Computers and the humanities

Lecture 2.1

Defining data

What is data?

- Data: potential information



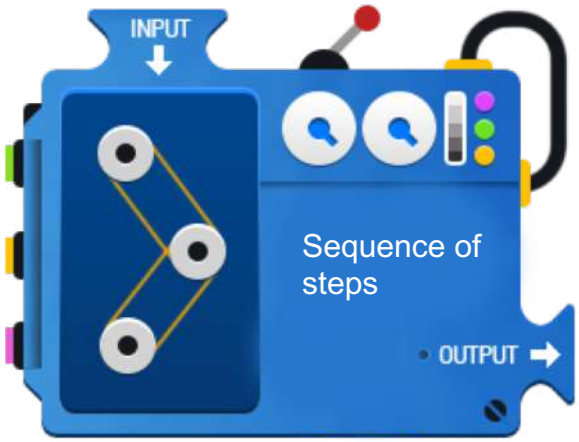
DIKW pyramid

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

T.S. Eliot, *The Rock* (1934)

Using data

Data



Answer

Is this data?

Think of a literary text (say Tolkien's *The Lord of the Rings* series).

Let's imagine we have a digital file containing all these novels.

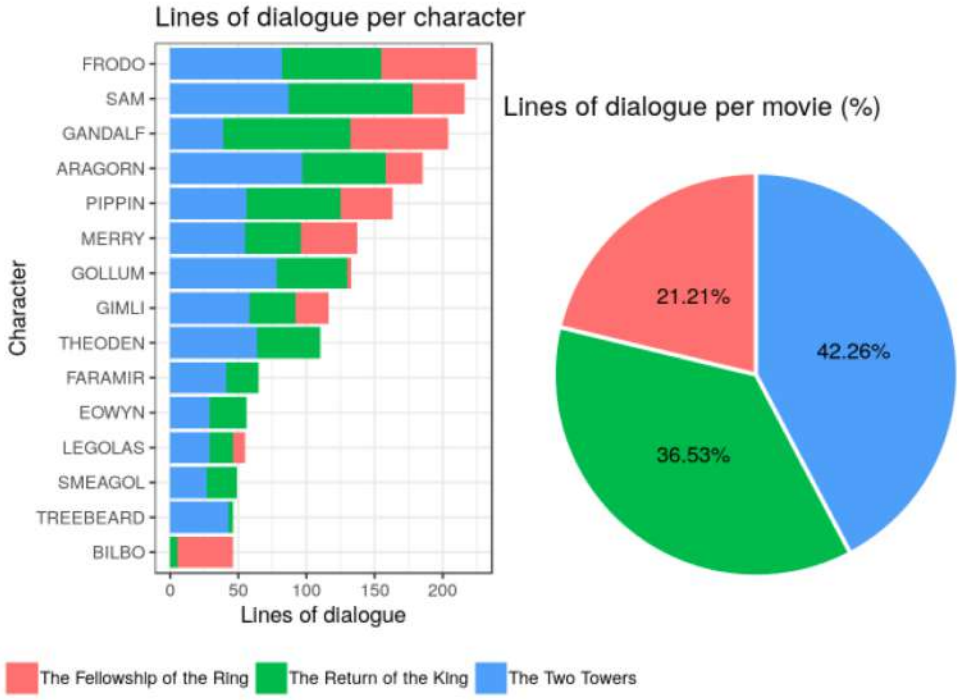
If you read them on your computer, are you treating the novels as data?

Is this data?

- No
- If you read the novels in a digital format, you are treating it as information (according to our definition so far).
- What would you need to do to treat it as data?

Treating LOTR as data

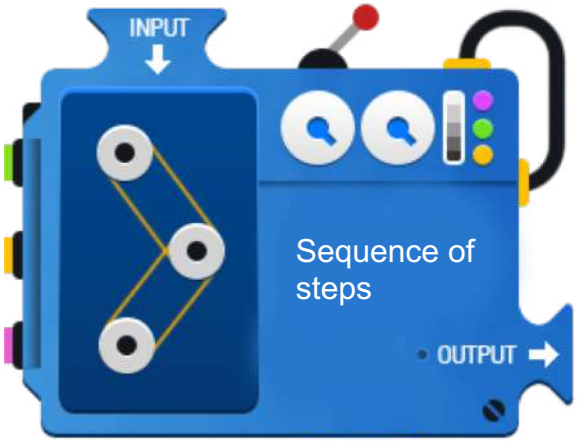
- Do a visualization to show the lines of dialogue per character.

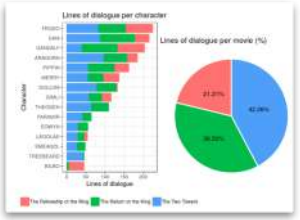


https://www.kaggle.com/xvivancos/analyzing-the-lord-of-the-rings-data

Using data

Data





Answer

Context

- Data is context-dependant
- The question should not be what are data but "when are data" (Borgman, 2015)

That which is data in one context might be information or knowledge in another.

Types of data

- Free text
- Categories
- Coordinates
- Floating point number
- Etc.

They can be manually assigned, or computationally derived.

Types of data

Every category is a preconception - but it is very hard not to impose categories.

Characteristics of a dataset

1. A phenomenon is represented as a set of objects (also called data points, measurements, samples, records) and their features (also called attributes, characteristics, metadata, variables). The features may include already available metadata, as well as measurements of objects' characteristics we generate using algorithms. (The latter process is called feature extraction.)

2. Together, the objects and their features form a dataset.

3. The number of objects in a data set has to be finite.

4. Features are encoded using data types: whole and fractional numbers, categories, spatial coordinates, spatial shapes and trajectories, dates, times, text tags, or free text.

5. Each feature can use only one data type.

6. The number of features in the dataset has to be finite.

Manovich (2020)

Different Objectives

Different objectives across the disciplines

- Prediction (Science)
- Control (Engineering)
- Explanation (Social science)
- Interpretation (Humanities)
- Automation (Machine learning)

Data elsewhere

Data has changed our understanding of nature.

Now, we have access to many different sources of data in digital format, and this makes analysis easier.

But does this mean that every aspect of our lives should be decided by data?

GET1030

Computers and the humanities

Lecture 2.3

Problems of using data

The proxy problem

Sometimes we don't have data for the exact thing that we want to measure and we choose to measure something else.

The proxy problem

Let's imagine I want to use data to grade you on this course and I have access to data on courses you have taken earlier.

Let's say I assume that your previous student performance will be a good indicator of how well you will do in this module, and I assign your grades based on your previous scores for participation, exams and assignments.

What do you think about this?

Potential issues

It locks you in the past (what if you really tried to do better this semester?)

The system would be biased towards people that had a great start (for whatever reason)

So let's imagine that instead of using your data I use data on students similar to you to assign your scores (based on your zip code, your gender, your chosen major).

A cultural example

Films have become more innovative over time.

How do you decide if a film is more innovative?

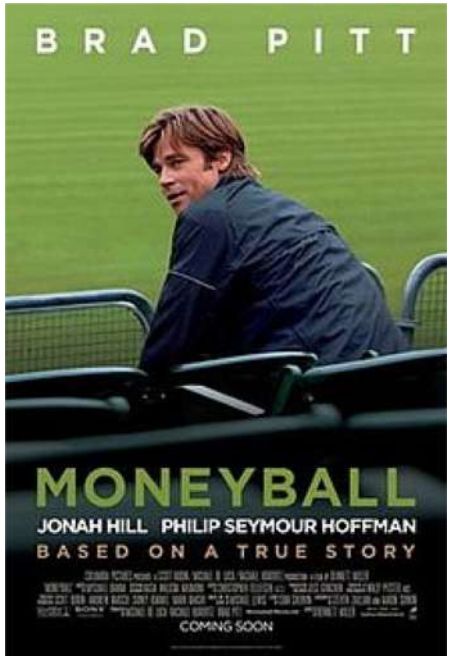
Potential answer: number of keywords.

Potential issues

It is very tempting for businesses and institutions to use data in this way (loans, job applications, health insurance, prison sentencing).

The proxy problem

Areas where data make sense are like data-driven baseball.



The proxy problem

Baseball […] has statistical rigor. Its gurus have an immense data set at hand, almost all of it directly related to the performance of players in the game. Moreover, their data is highly relevant to the outcomes they are trying to predict. This may sound obvious, but as we'll see [...], the folks building Weapons of Math Destruction routinely lack data for the behaviors they're most interested in. So they substitute stand-in data, or proxies. They draw statistical correlations between a person's zip code or language patterns and her potential to pay back a loan or handle a job. These correlations are discriminatory, and some of them are illegal. Baseball models, for the most part, don't use proxies because they use pertinent inputs like balls, strikes, and hits (Cathy O'Neil, *Weapons of Math Destruction*, 17–18).

GET1030

Computers and the humanities

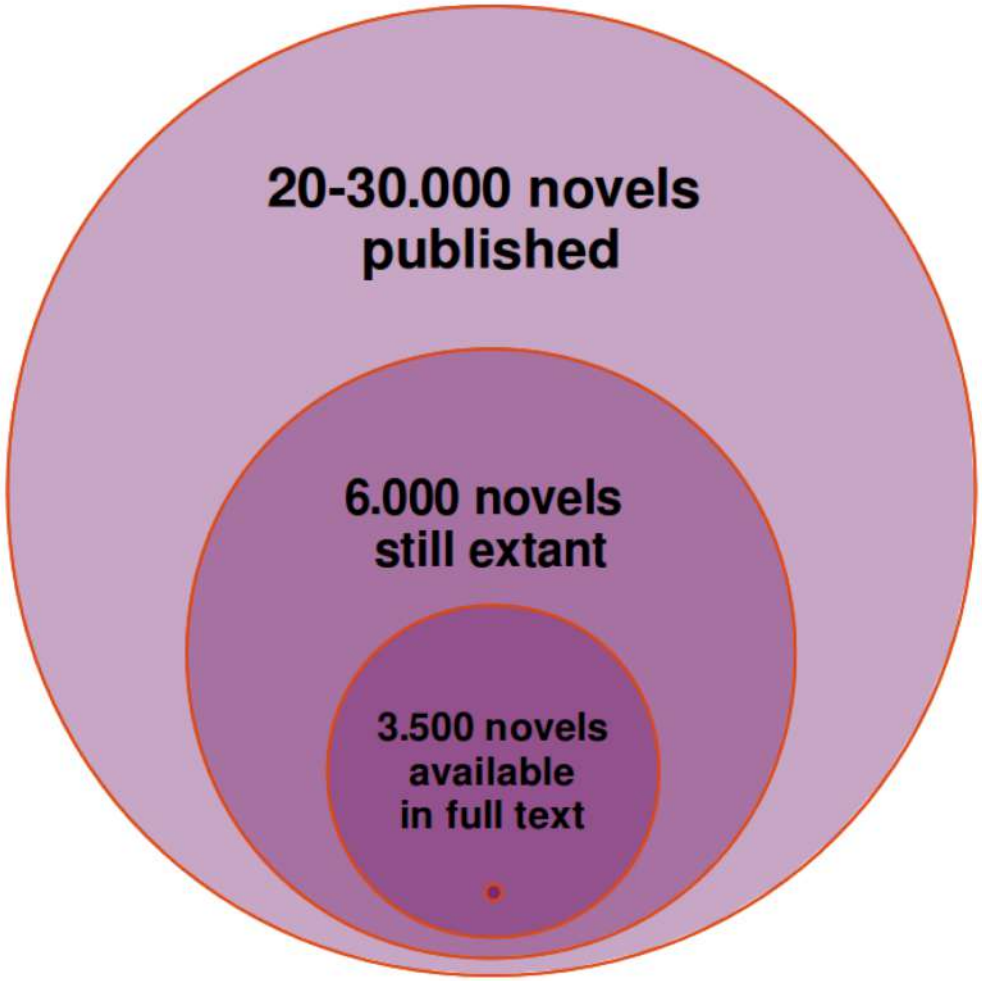
Lecture 2.4

Data in the humanities

Data about culture

- Digitized books, films, artworks, video recordings
- Motion capture
- 3d models
- Surveys
- Production details (film, theatre,books)
- Historical records (posters, sales data)
- Social media (objects shared but also interactions)

Limited data in the humanities



Digitization and availability of British Nineteenth-Century Novels. Image published under a CC-BY license (Schöch).

Types of data

- Big data
- Smart data

Smart data

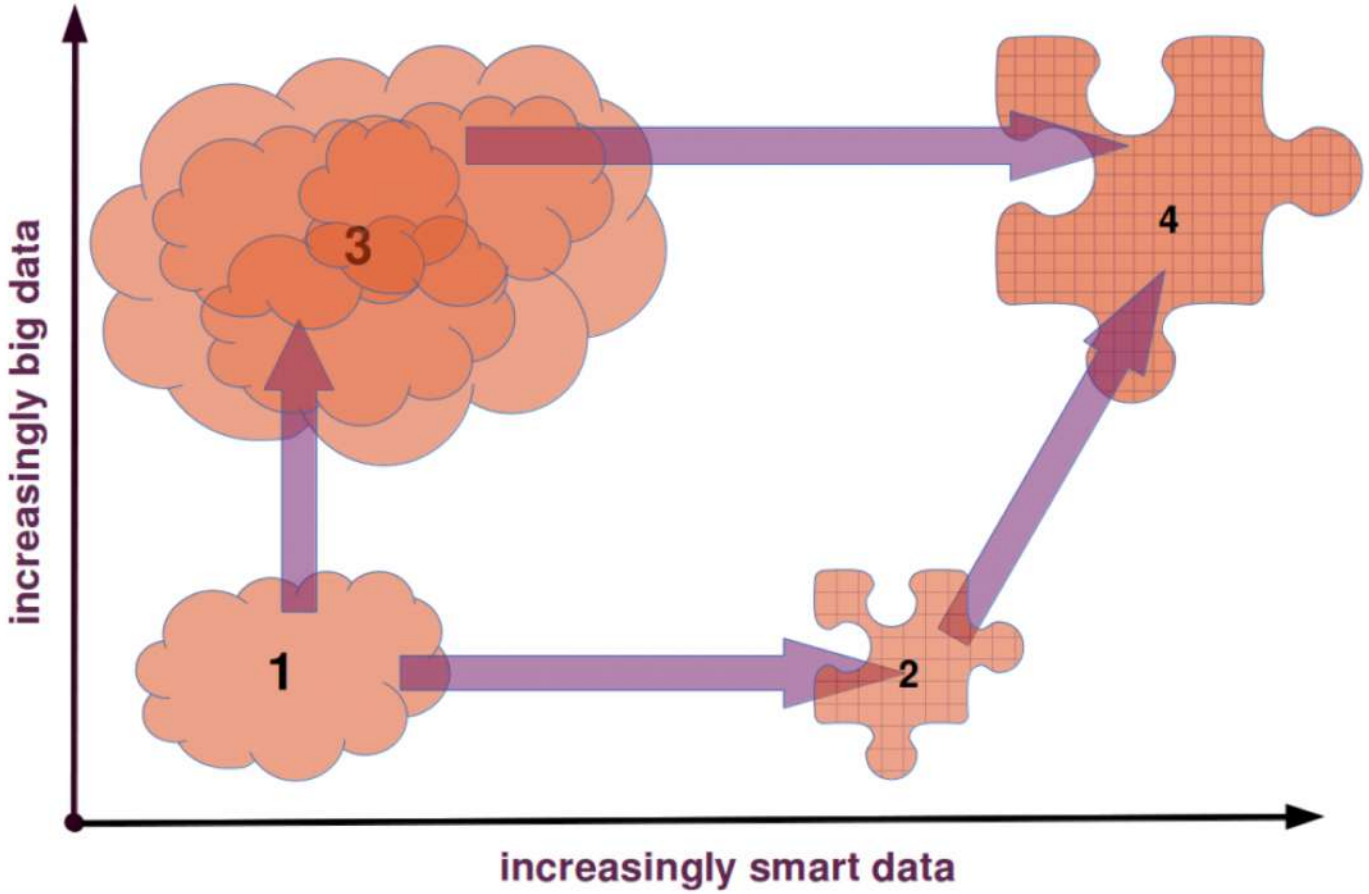
- Markup, annotations and metadata
- Clear data models

Big data

Volume, variety, velocity (veracity and value)

Outliers and errors less important

The future of data for the humanities



From Schöch (2013)

Objectivity and interpretation

- No data in the humanities is fully objective
- It requires active efforts to collect or produce
- Depends on
  - What technology enables
  - Opinions and perspectives of people creating a system

Problems of data

- Standardization
- Incompleteness
- Inaccuracy

What are potential problems in your group projects?

References

Feynman, Richard. 2017. *The Character of Physical Law*. Edited by Alan Sleath. Cambridge: The MIT Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. Crown Publishers.

Schöch, Christof. 2013. "Big? smart? clean? messy? Data in the humanities." *Journal for Digital Humanities 2(3).*