GEI1002

Computers and the humanities

Introductory Lecture

Concepts and module structure

Data is everywhere

Data is used everywhere. It is important for everyone to be a thoughtful producer and consumer of data.

Data literacy

- Ability to obtain and analyze data (technical skills)
- Asking questions about the ways data was obtained and analyzed (interpretive skills)

This module

- We will focus on the analysis and creation of data visualizations.
- The lectures and tutorials will use examples from the arts and culture (film, literature, etc.) but you can use any topic from the humanities or social sciences for your projects.

Objectives of the module

- To get you thinking about interdisciplinary work
- To provide you with some general skills that are useful for many other fields
- To give you a foundation from which you can learn on your own

General skills you will learn

- How to evaluate datasets and visualizations
- How to create a wide range of visualizations

Two interconnected components

Doing stuff: learning a bit of **survival coding** to understand the computational process of dataviz.

Thinking: stepping back and reflecting on what we have done.

These two steps are not distinct, they feed each other.

This double focus on doing and thinking shapes our assignments and the structure of the module.

Teaching modes

We follow a flipped classroom model, all lectures are delivered through video. In the tutorials we will do two things:

- Discussions in small groups: based on readings
- Coding labs: learn how to create exploratory data visualizations in Python (no coding experience required)

GEI1002

Computers and the humanities

Introductory Lecture

Part II. What is data?

What is data?

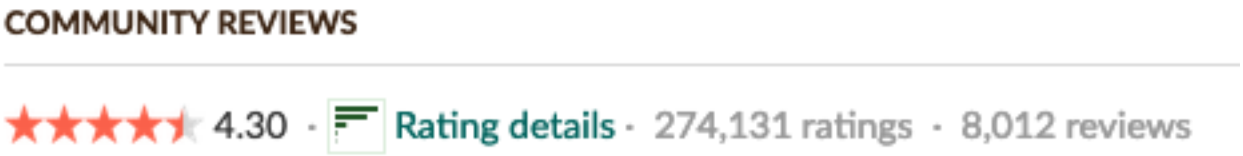
- Systematic observations about a phenomenon
- Variables and values
- Example: a dataset of books
- What are some possible variables?
    - Author
    - Genre
    - Number of pages
    - Average rating

Types of data

- Quantitative and categorical data
- This refers to the possible values a variable can take
- Quantitative data
  - Number of pages: 292, 165
  - Rating: 3.6, 4.1
- Categorical data
  - Genre: Horror, Sci-Fi, Mystery, Romance
  - Author: Arthur Conan Doyle, Agatha Christie

Subjectivity in data

- Subjective vs objective data
- Ratings for *The Adventures of Sherlock Holmes,* from goodreads.com



COMMUNITY REVIEWS

★★★★★ 4.30 · Rating details · 274,131 ratings · 8,012 reviews

- Is this data objective?
    - Remember that quantitative data is not necessarily objective

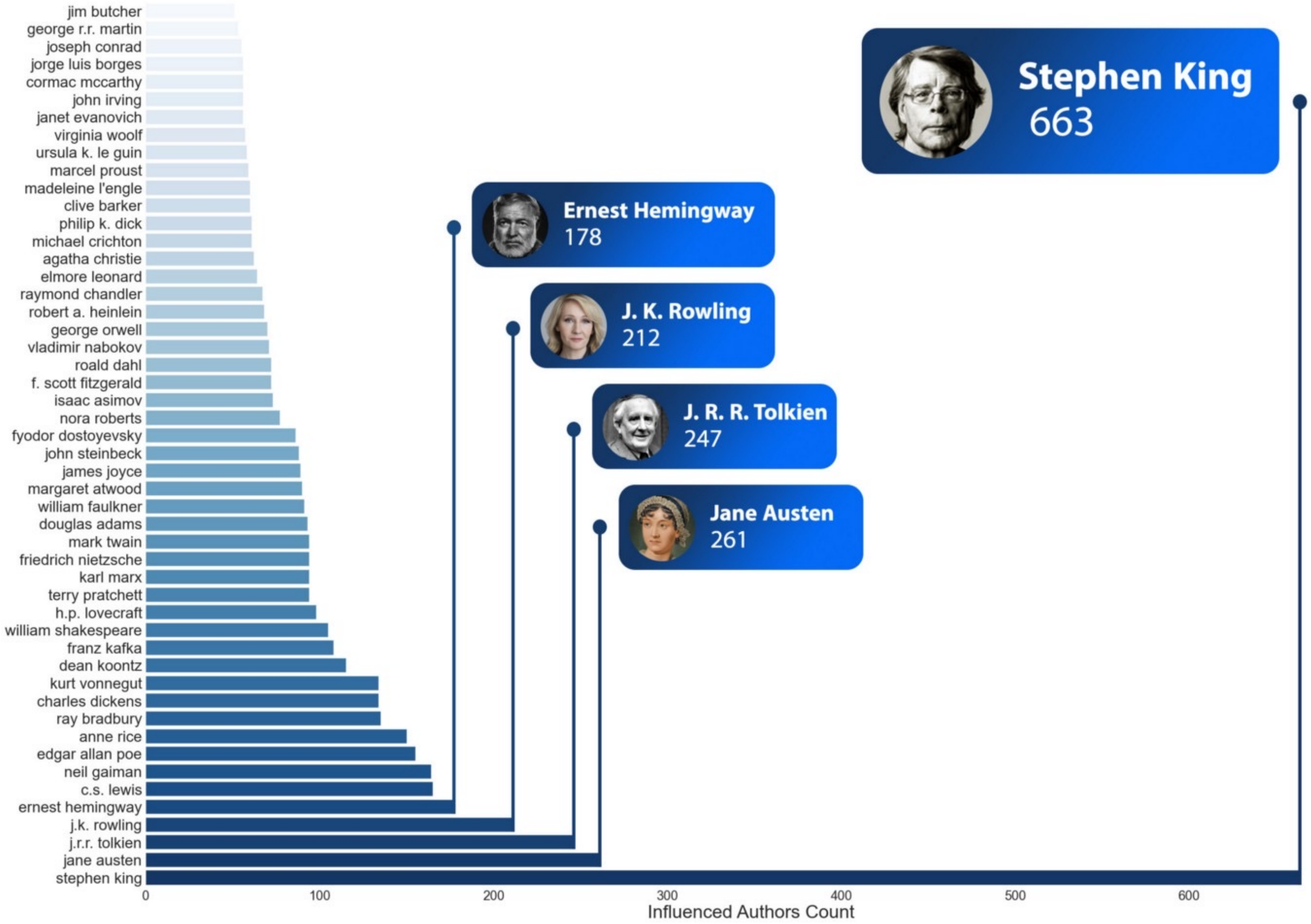
Subjectivity in data

- Subjective data is not useless, but you need to be careful about what claims you can make with it.
- These two statements are not the same:
  - “The Adventures of Sherlock Holmes is a good book”
  - “The Adventures of Sherlock Holmes has a high average rating on goodreads.com”
- Can you use the goodreads.com data to claim that a book is good? It depends on the context.

Subjectivity in data

- And “categorical” data is not necessarily very subjective
- Can you think of an example?

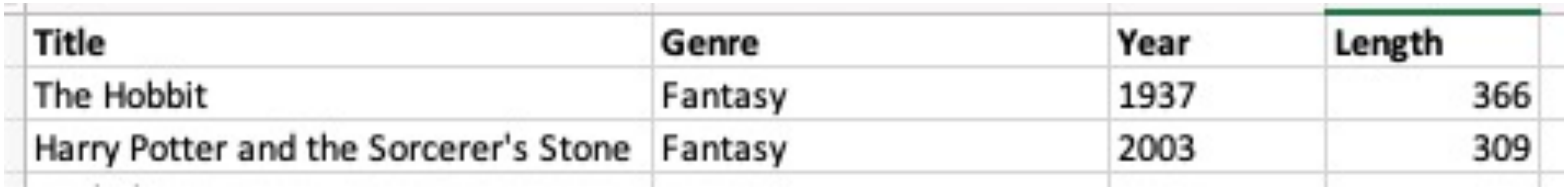
Authors as “categories”



Source: https://towardsdatascience.com/all-the-authors-around-us-an-analytical-look-into-goodreads-authors-dataset-part-one-61697721e58e

Decisions shape the data

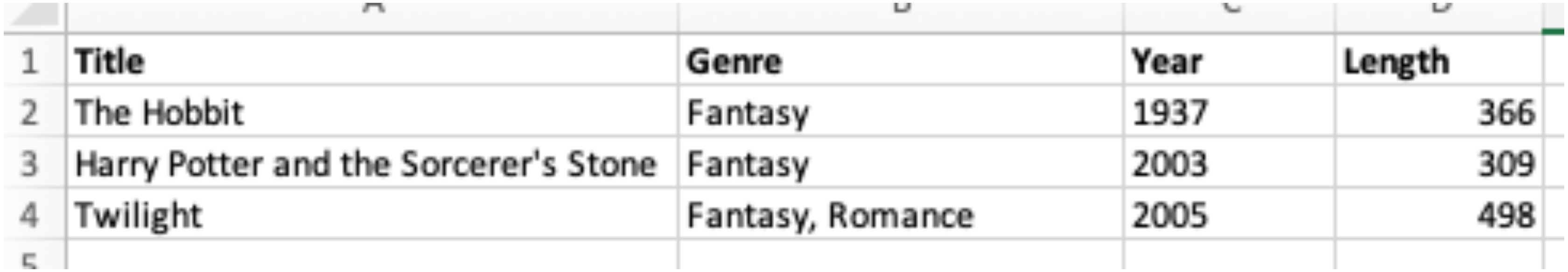
- A dataset is always shaped by the decisions of the people who made it
- Lets imagine a dataset with book genres



Title	Genre	Year	Length
The Hobbit	Fantasy	1937	366
Harry Potter and the Sorcerer's Stone	Fantasy	2003	309

Decisions shape the data

- What about a book that fits into two categories?
- Shall we list both categories?



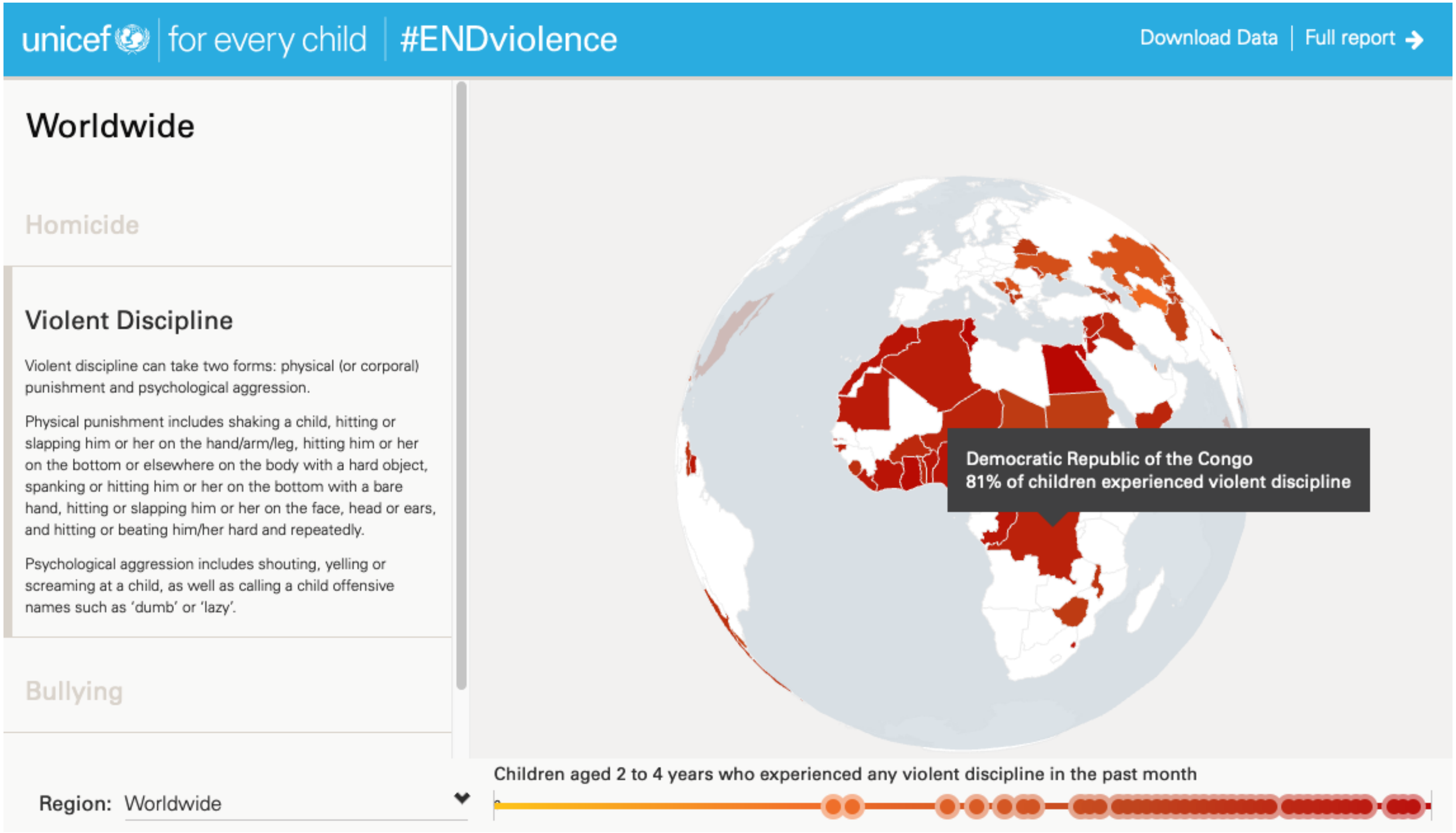
	A	B	C	D
1	Title	Genre	Year	Length
2	The Hobbit	Fantasy	1937	366
3	Harry Potter and the Sorcerer's Stone	Fantasy	2003	309
4	Twilight	Fantasy, Romance	2005	498
5				

- This constraints the types of analysis you can make

Definitions

- Different definitions will lead to different datasets (Thorp, 2021)
- A good data project describes the definitions used, and provides links to the sources

Definitions



unicef | for every child | #ENDviolence

Worldwide

Violent Discipline

Violent discipline can take two forms: physical (or corporal) punishment and psychological aggression.

Physical punishment includes shaking a child, hitting or slapping him or her on the hand/arm/leg, hitting him or her on the bottom or elsewhere on the body with a hard object, spanking or hitting him or her on the bottom with a bare hand, hitting or slapping him or her on the face, head or ears, and hitting or beating him/her hard and repeatedly.

Psychological aggression includes shouting, yelling or screaming at a child, as well as calling a child offensive names such as 'dumb' or 'lazy'.

Bullying
Region: Worldwide
Democratic Republic of the Congo
81% of children experienced violent discipline
Children aged 2 to 4 years who experienced any violent discipline in the past month

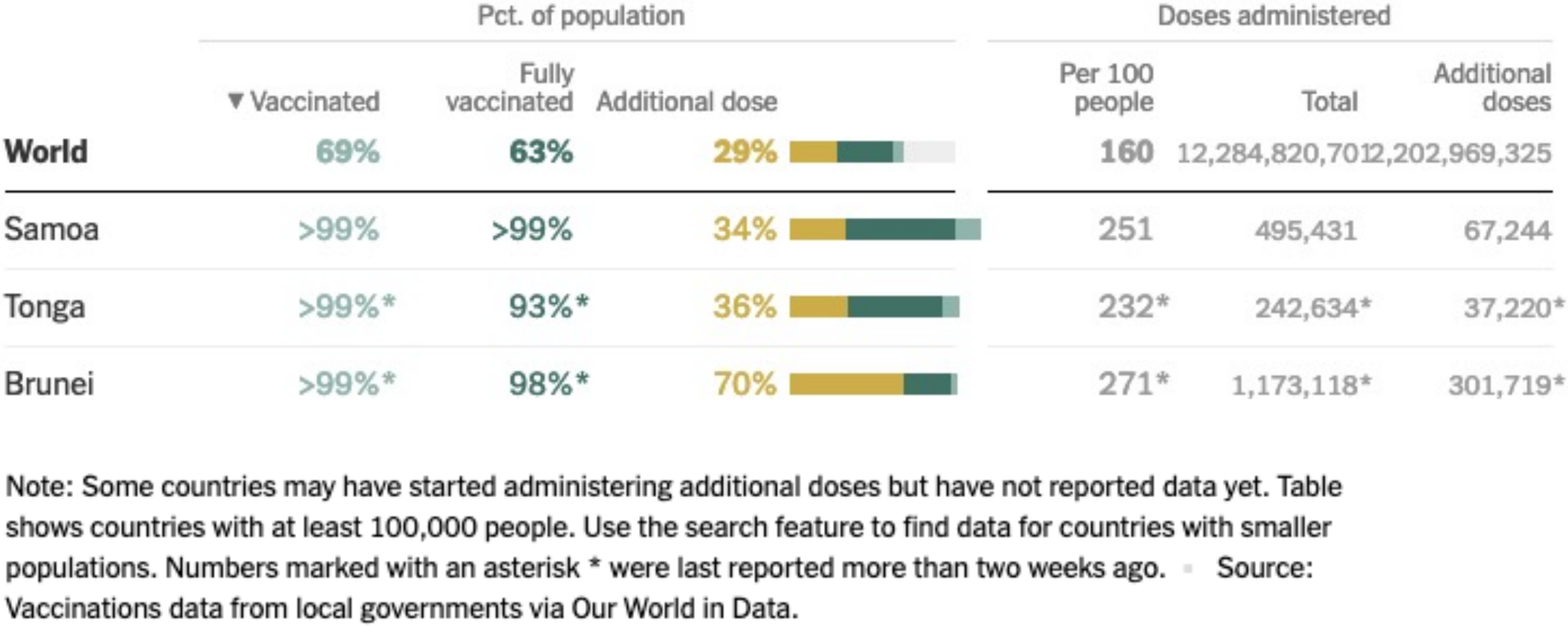
https://works.periscopic.com/unicef-child-violence/#all&criteria=1

Limitations

- There are always limitations in the ways data is collected
- A good data project acknowledges these limitations

Limitations

Covid-19 vaccinations by country



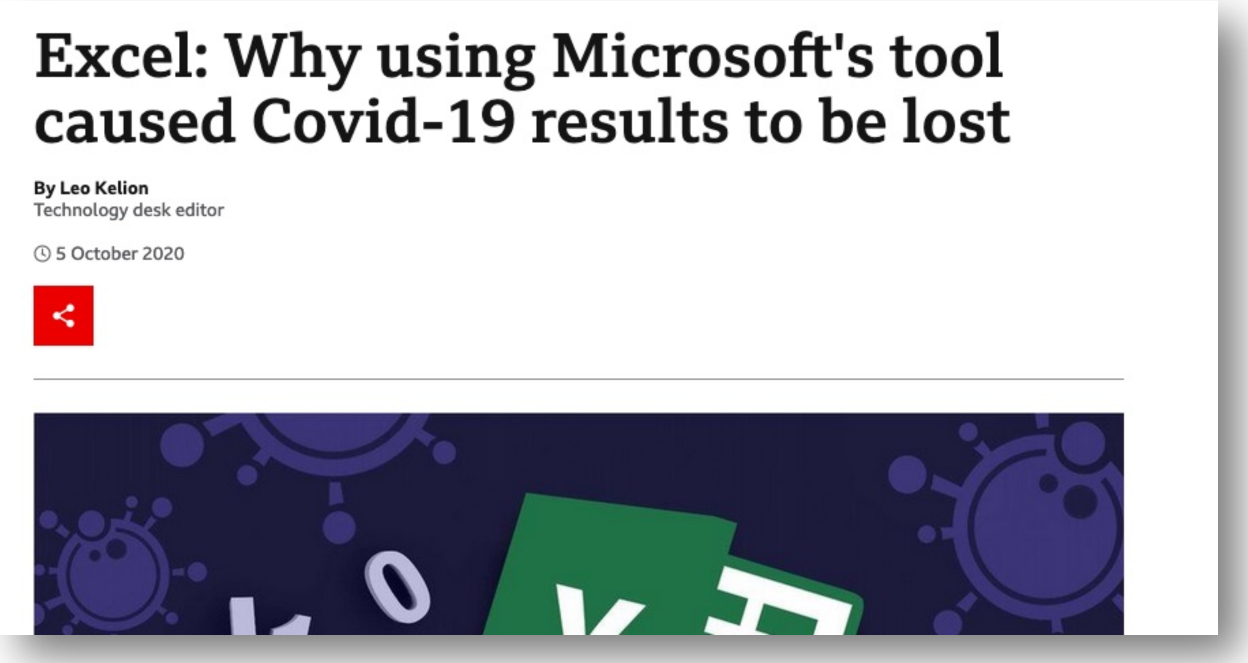
	Pct. of population			Doses administered		
	▼ Vaccinated	Fully vaccinated	Additional dose	Per 100 people	Total	Additional doses
World	69%	63%	29%	160	12,284,820,701	2,202,969,325
Samoa	>99%	>99%	34%	251	495,431	67,244
Tonga	>99%*	93%*	36%	232*	242,634*	37,220*
Brunei	>99%*	98%*	70%	271*	1,173,118*	301,719*

Note: Some countries may have started administering additional doses but have not reported data yet. Table shows countries with at least 100,000 people. Use the search feature to find data for countries with smaller populations. Numbers marked with an asterisk * were last reported more than two weeks ago. Source: Vaccinations data from local governments via Our World in Data.

https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html

Errors in data

- Even the best intended datasets might contain errors



Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October 2020

https://www.bbc.com/news/technology-54423988

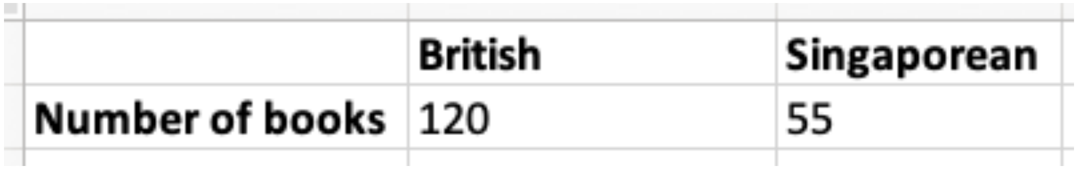
Context

- Pay attention to who made the data and why.
- For example, what is the objective of goodreads.com?
- How is it different from the objective of Unicef’s “#ENDviolence”?

“Tidy” data

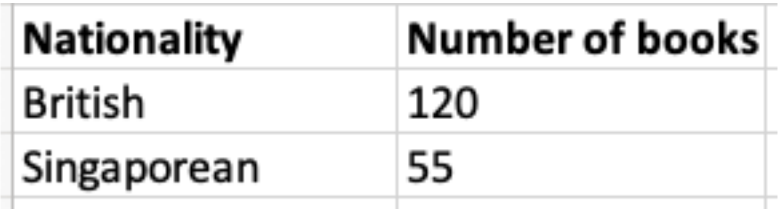
- For this module, we will mostly use spreadsheets with “tidy” data.
- The variables are the columns.
- There is one observation per row.

“Tidy” data



	British	Singaporean
Number of books	120	55

✗



Nationality	Number of books
British	120
Singaporean	55

✓

Summary

- Data refers to systematic observations
- A dataset includes values, variables and observations
- In “tidy” datasets there is one observation per row and the columns represent the variables
- Variables can be quantitative or categorical, and they can be *more* subjective or more objective
- But no dataset is fully objective
- Pay attention to the context of a data project:
    - Objectives
    - Definitions
    - Limitations
    - Assumptions
- A good data project clearly states its limitations and sources.

GEI1002

Computers and the humanities

Introductory Lecture

Part III. Data visualizations

What is data visualization?

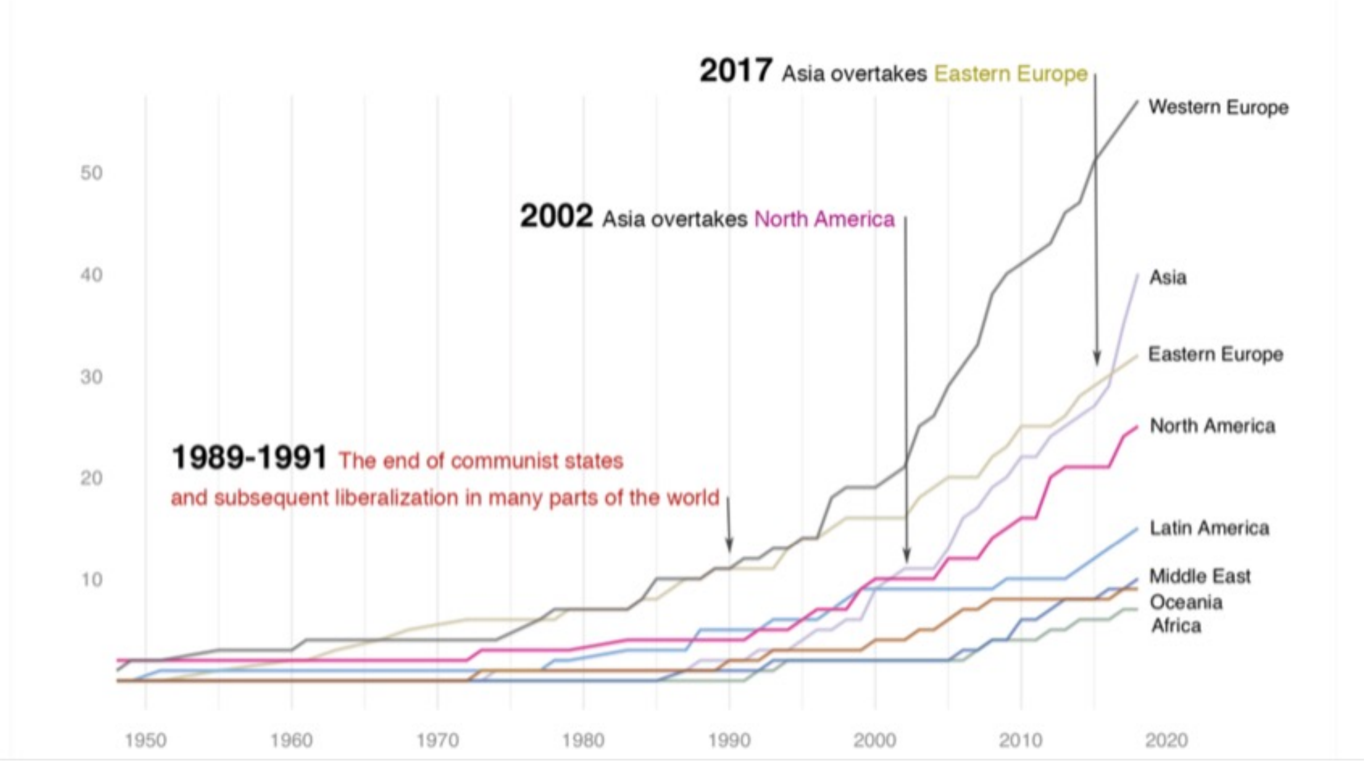
- Representation of data using visual conventions: shapes, colors, distances, symbols.
- Fundamental activity for making sense of data.
- Data visualization ~ infographics, graphs, charts.

Quantitative description

- In this module, we will focus on “quantitative descriptions”, sometimes called “exploratory data analysis”.
- We will not try to prove hypotheses, as this is a more advanced topic.

Quantitative description

Number of art biennales around the world over time



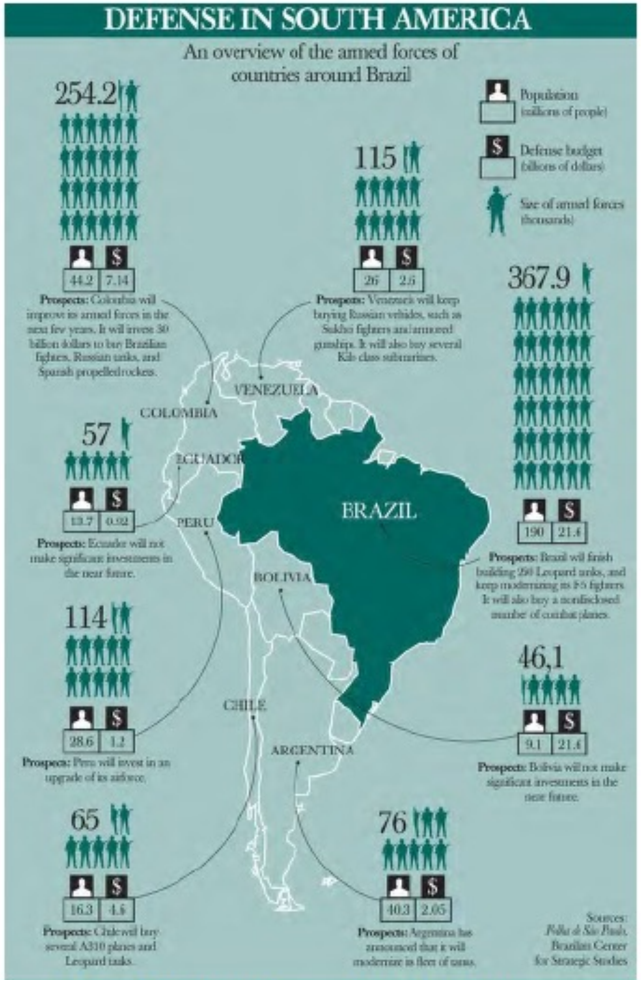
Tifentalle and Manovich (2020)

Choosing a data visualization

- What task is the visualization aimed to help people achieve? (Cairo 2012)

through thinking from the end user for the visualisation

Choosing a data visualization



Prospects: Colombia will improve its armed forces in the next few years. It will invest 30 billion dollars to buy Brazilian fighters, Russian tanks, and Spanish propelled rockets.

Prospects: Venezuela will keep buying Russian vehicles, such as Sukhoi fighters and armored gunships. It will also buy several Kilo class submarines.

Prospects: Ecuador will not make significant investments in the near future.

Prospects: Brazil will finish building 250 Leopard tanks, and keep modernizing its F5 fighters. It will also buy a nondisclosed number of combat planes.

Prospect: Peru will invest in an upgrade of its airforce.

Prospect: Bolivia will not make significant investments in the near future.

Prospects: Chile will buy several A310 planes and Leopard tanks.

Prospects: Argentina has announced that it will modernize its fleet of tanks.

Sources: *Folha de São Paulo*, Brazilian Center for Strategic Studies

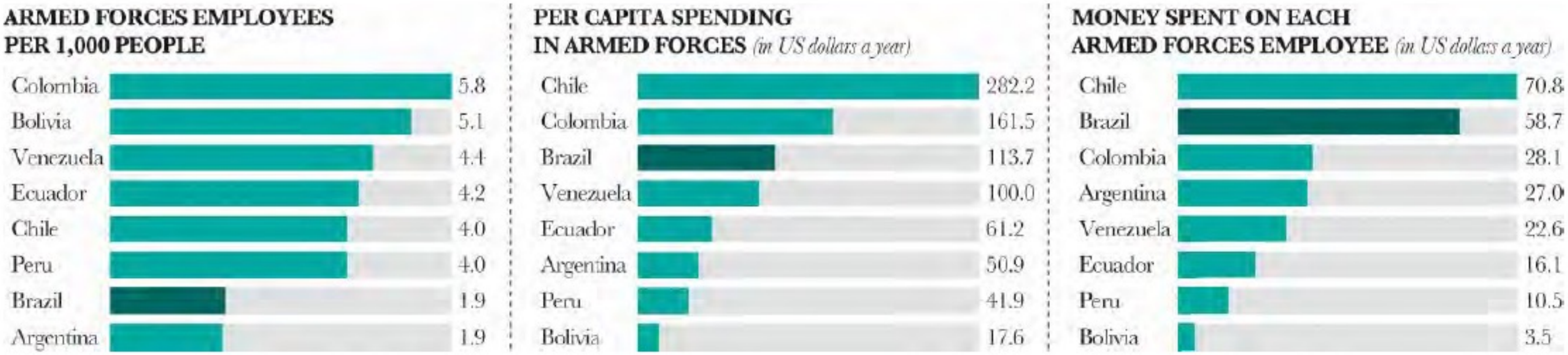
From Cairo (2012)

What tasks does this visualization enable?

Choosing a data visualization

Here is the same data, but visualized as a barchart.

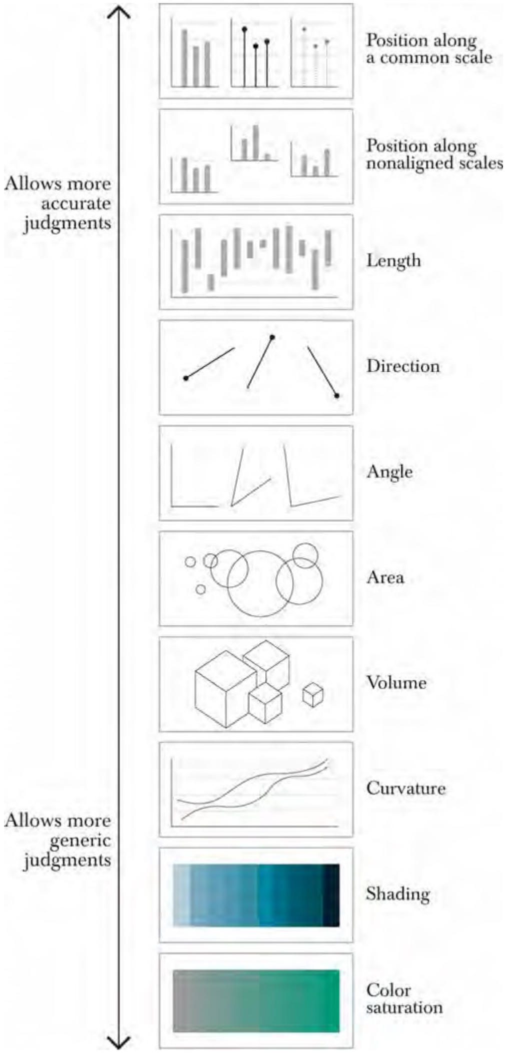
From Cairo (2012)



Choosing a data visualization

- Is the visualization suitable to its context?
- Identifying this is an art and a science.

Choosing a data visualization


Allows more accurate judgments
Allows more generic judgments
Position along a common scale
Position along nonaligned scales
Length
Direction
Angle
Area
Volume
Curvature
Shading
Color saturation

Alberto Cairo (2012)'s guide for choosing a visualization, based on Cleveland and McGill's elementary perceptual tasks (see Files for full PDF).

Good data visualizations?

- There are many “rules” and “best practices”
- Often they are not based on empirical evidence
- It is important to pay attention to context
- What task is the visualization aimed to help people achieve?
- Your argument is more important than getting things “right”.

Summary

- Representation of data using visual conventions: shapes, colors, distances, symbols.
- Key question: what task is the visualization aimed to help people achieve?
- Your argument is more important than getting things “right”.

GEI1002

Computers and the humanities

Introductory Lecture

Part IV. Case studies: using data to study culture

Two types of mediation

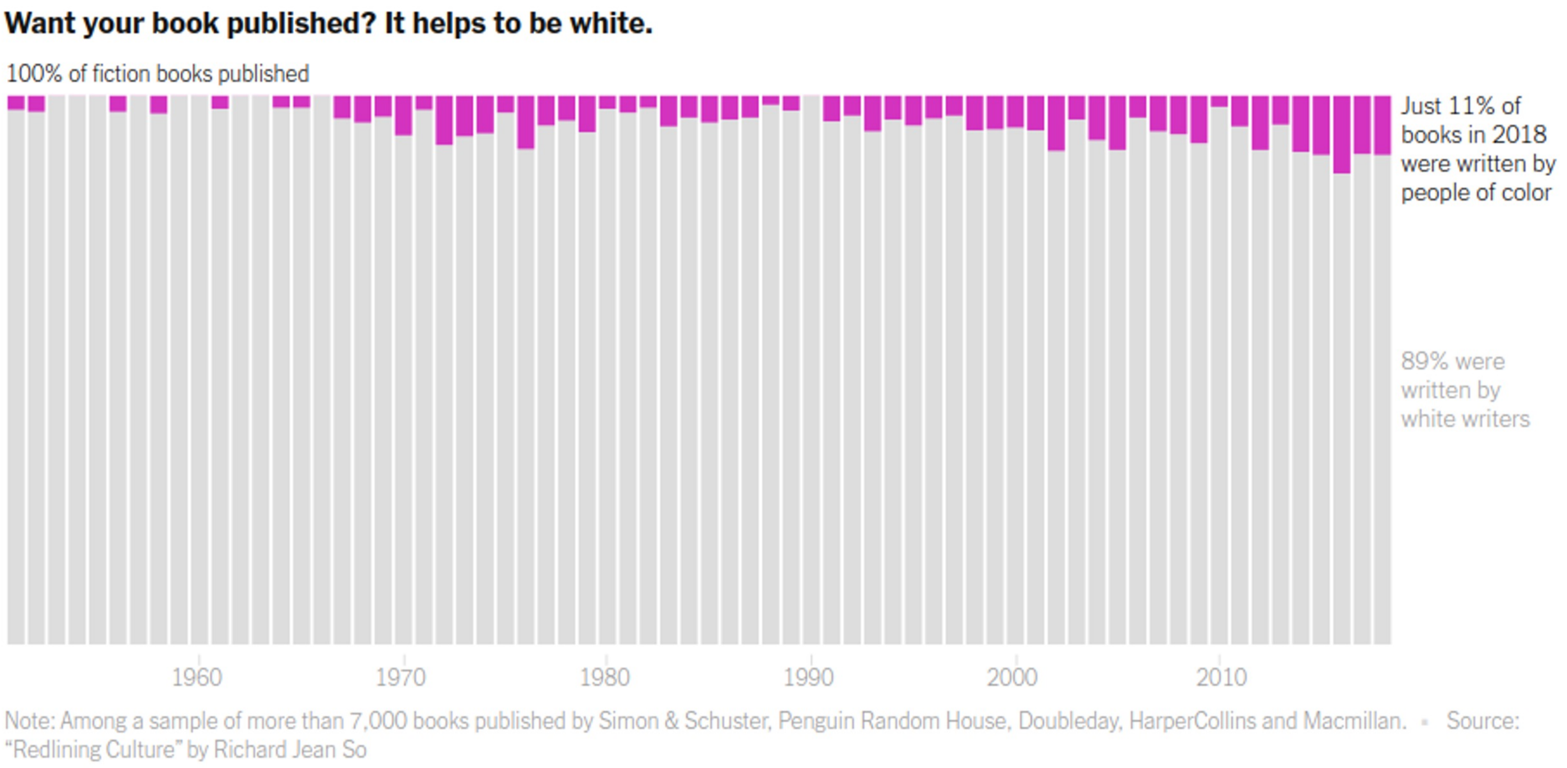
-World to Data
-Data to Image
See Gray et al (2016) for more on this.

*always pay attention to the decisions made by the people who created the data or the visualizations
*if you are the creators, justify your choices!

Two types of mediation

A research project: How many current authors are people of color (PoC)?

The publishing industry



Want your book published? It helps to be white.

100% of fiction books published

Just 11% of books in 2018 were written by people of color

89% were written by white writers

Note: Among a sample of more than 7,000 books published by Simon & Schuster, Penguin Random House, Doubleday, HarperCollins and Macmillan. • Source: "Redlining Culture" by Richard Jean So

Source: "Redlining Culture" by Richard Jean So.

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

Two types of mediation

First, we gathered a list of English-language fiction books published between 1950 and 2018 […]

We also constrained our search to books released by some of the most prolific publishing houses […] After all that we were left with a dataset containing 8,004 books, written by 4,010 authors.

To identify those authors' races and ethnicities, we worked alongside three research assistants, reading through biographies, interviews and social media posts. Each author was reviewed independently by two researchers. If the team couldn't come to an agreement about an author's race, or there simply wasn't enough information to feel confident, we omitted those authors' books from our analysis. By the end, we had identified the race or ethnicity of 3,471 authors.

The publishing industry

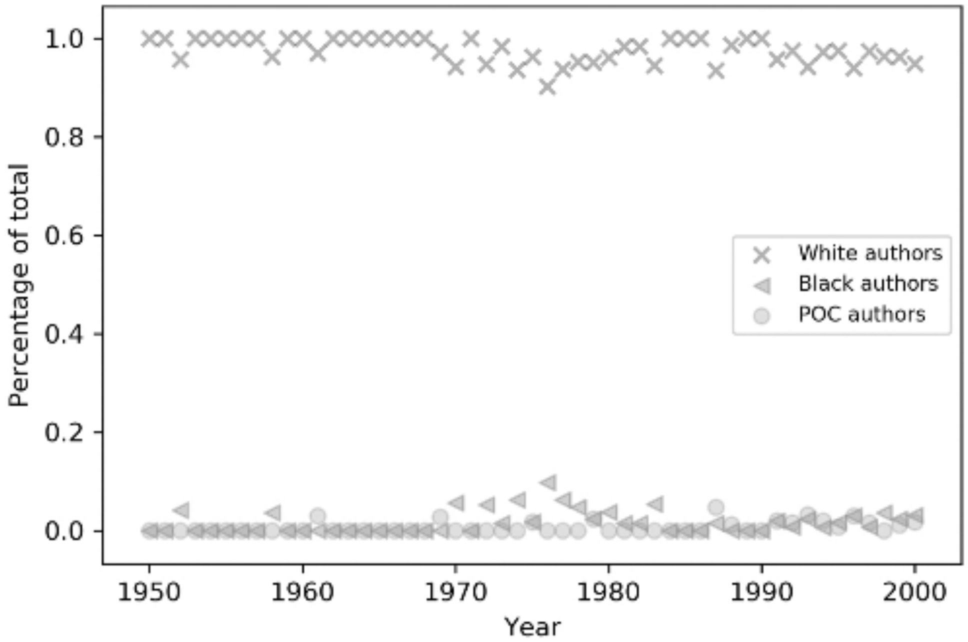


FIGURE 0.1 Percentage of novelists by racial identity (white, black and POC—Asian American, Latinx, and Native American) published at Random House by year between 1950 and 2000.

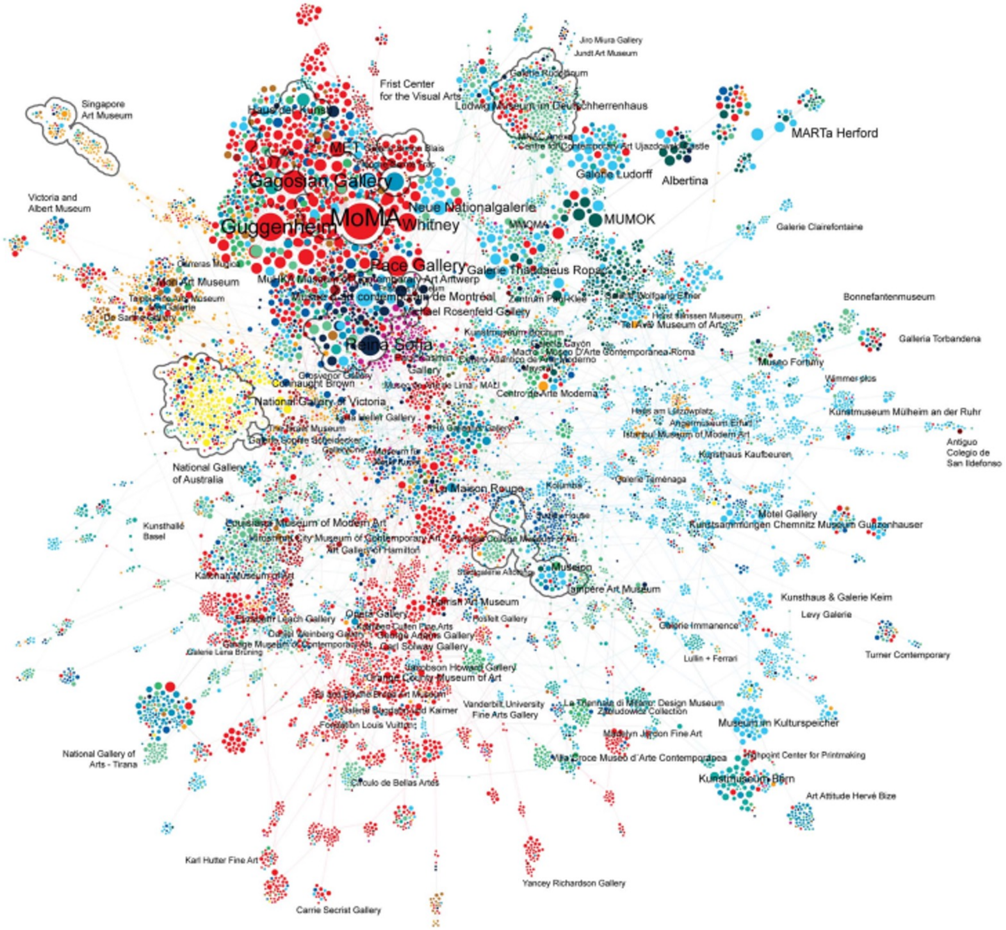
Source: “Redlining Culture” by Richard Jean So (2020).

Data and Culture

We will see many different types of data and visualizations. But we must always ask:

- How was a phenomenon represented as data?
- How was this data visualized?
- What decisions were made?
- Do they make sense within their context?

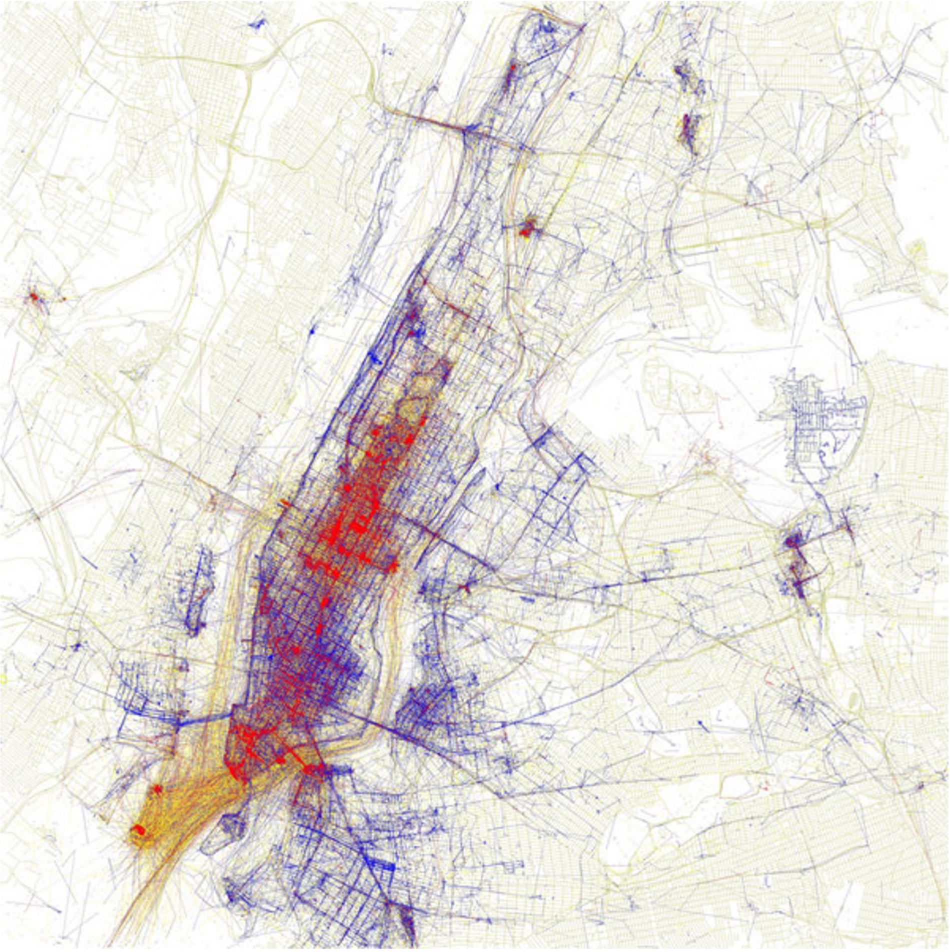
Sucess in art



Network of 12,238 exhibition venues for artists

Fraiberger et al (2018)

Phototrails



Locals and tourists in New York. The visualization compares locations of photos uploaded to Flickr and Picasa. Blue pictures are by locals. Red pictures are by tourists. Yellow pictures might be by either.

Hochman and Manovich (2013)
https://firstmonday.org/ojs/index.php/fm/article/view/4711/3698

GEI1002

Computers and the humanities

Introductory Lecture

Part V. Structure and Assignments

Module Structure

Concepts
1 Data, computation and the humanities
2 What is data?
3 Visualizing data

Tools
4 Gentle introduction to data visualization in Python
5 Working with text I
6 Working with text II
7 Visualizing networks
8 Tools for visualizing networks
9 Visualizing geographical data
10 Tools for visualizing geographical data

Looking beyond
11 Computation and society
12 Group project consultations
13 Group project consultations

All lectures will be video based

Tutorial Sessions

Concepts
1 Data, computation and the humanities
2 What is data?
3 Visualizing data

Tools
4 Data visualization in Python

5 Visualizing text
6 Tools for visualizing text

7 Visualizing networks
8 Tools for visualizing networks

9 Visualizing spatial data
10 Tools for visualizing spatial data

Looking beyond
11 Computation and society
12 Project consultations [no lecture]
13 Project consultations [no lecture]

Tutorial Sessions

#1 Concepts (Week 3)

#2 Python visualizations (Week 5)

#3 Text (Week 7)

#4 Networks (Week 9)

#5 Spatial visualizations (Week 11)

A note on programming

There’s a very gentle introduction to programming in Week 4. This is not a full-fledged programming course, and I hope the simple exercises will get you interested in learning more about programming (if you haven’t already done this).

We will only learn to load Excel files into Python and to visualize them through simple commands (using interactive Jupyter notebooks).

Tools for the module

None of these require programming.

Voyant Tools for textual analysis http://voyant-tools.org/

Google maps for geographical visualizations http://maps.google.com

Gephi for network analysis https://gephi.org/

ASSESSMENT

See the PDF for details on the assessment schedule and description of the assignments.

REFERENCES

Cairo, Alberto. *The Functional Art: An Introduction to Information Graphics and Visualization*. 1st edition. Berkeley, California: New Riders, 2012.

Thorp, Jer. *Living in Data: A Citizen's Guide to a Better Information Future*. New York: MCD, 2021.

Gray, Jonathan et al. 'Ways of Seeing Data: Toward a Critical Literacy for Data Visualizations as Research Objects and Research Devices'. In *Innovative Methods in Media and Communication Research*, edited by Sebastian Kubitschko and Anne Kaun, 227–51. Cham: Springer International Publishing, 2016.

So, Richard Jean. *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction*. New York: Columbia University Press, 2020.

Fraiberger, Samuel P., Roberta Sinatra, Magnus Resch, Christoph Riedl, and Albert-László Barabási. 2018. "Quantifying Reputation and Success in Art." *Science* 362 (6416): 825–29. https://doi.org/10.1126/science.aau7224.

Hochman, Nadav, and Lev Manovich. 2013. "Zooming into an Instagram City: Reading the Local through Social Media." *First Monday*. https://firstmonday.org/ojs/index.php/fm/article/view/4711/3698