(Getting Started with)

Text Analysis & Data Viz

Maggie Murphy & Jo Klein

Slides at: go.uncg.edu/r7f48y

1

Text Analysis

Text Analysis. Why?

Humanities scholar have always analyzed texts. This includes:

- collecting/selecting texts in order to explore an hypothesis;
- looking for patterns (of words, ideas, symbols, rhetorical or formal structures, etc) within an individual text and/or within sets of texts;
- discovering relationships (of development, dependence, seriality, association, intention, allusion, intertextuality, etc) between parts of texts, whole texts, or sets of texts;
- interpreting the significance of these patterns, relationships, and texts;
- developing arguments for the larger significance of these interpretations

Adapted from "Text Analysis" by Natalie M. Houston in *Digital Pedagogy in the Humanities*

Text Analysis. Why?

However, certain changes have expanded how we can read and analyze texts:

- large scale digitization changes our access both to specific texts and to new quantities of texts;
- relational databases and full text search expand the kinds of research queries that can be pursued;
- new media forms and new interfaces transform how we understand and perform acts of reading;
- growth in computational power and storage offer new ways of curating, displaying, and using collections of texts for analysis;
- tools for data visualization and multimodal composition offer new ways of exploring texts and building arguments.

Adapted from “Text Analysis” by Natalie M. Houston in *Digital Pedagogy in the Humanities*

Text Analysis. Why?

If we want to look at patterns across large numbers of texts, we are limited by the scope of what we can read and categorize ourselves.

Close Reading

Careful, sustained analysis of individual texts

- Focuses on a (relatively) small amount of text
- Often does the work that a computer cannot
- Possibly more subjective
- Deeper comprehension of language

Distant Reading

Computational analysis of a large corpus of texts

- Focuses on a comparatively (or actually) huge amount of text
- Often does the work that a human cannot
- Possibly more objective
- Semantic comprehension of language

Adapted from "Text Analysis" by Kirsten Bussiere in *Digital Humanities: A Primer*

Text Analysis. What?

“Text analysis” refers to an umbrella of computational techniques that allow researchers to:

- Find or organize works around given parameters
- Visualize features of single texts
- Measure/classify textual features and relationships in a text or corpus
- Identify distinctive vocabulary in a corpus
- Model textual forms and their characteristics
- Model social boundaries and networks
- Learn from unsupervised modeling of unlabeled data

Word count, such as by

- Sentence
- Paragraph
- Document

Term frequency of

- Individual words
- N-grams
- Types of words

Sentiment (affect)

Concordance (context)

Adapted from “Seven ways humanists are using computers to understand text” by Ted Underwood

Text Analysis. What?

“Text analysis” refers to an umbrella of computational techniques that allow researchers to:

- Find or organize works around given parameters
- Visualize features of single texts
- Measure/classify textual features and relationships in a text or corpus
- Identify distinctive vocabulary in a corpus
- Model textual forms and their characteristics
- Model social boundaries and networks
- Learn from unsupervised modeling of unlabeled data

Model to explain
For testing hypotheses about data

Model to predict
For uncovering new patterns that lead to new hypotheses

See Galit Shmueli, “To Explain or to Predict?”, *Statistical Science* 25, no. 3 (August 1, 2010), https://doi.org/10.1214/10-STS330.

Adapted from “Seven ways humanists are using computers to understand text” by Ted Underwood

Text Analysis. How?

Ultimately, text analysis is performed with algorithms developed by computer scientists, statisticians, and linguists.

- Humanists interested in computational analysis of text might learn programming languages such as Python or R, in order to use relevant libraries and packages, like NLTK, Pandas, and NumPy, to clean, transform, analyze, and visualize* text
- Additionally, humanists may work *with* computer scientists and statisticians (usually on grant-funded projects) when more sophisticated knowledge and skills are needed
- There are also a number of "out-of-the-box" tools that can be used for computational analysis of text without programming skills

Some text adapted from "Text Analysis Resources" by Digital Humanities at Berkeley

Text Analysis. How?

There are a number of useful resources for learning programmatic text analysis skills, or for exploring existing tools. Here are some suggestions (including a few that we will explore in greater detail today)

- Tapor3: a directory of tools and resources for text analysis
- Early Print Lab: adapts a number of tools and techniques for exploring a specific corpus
- DataBasic: simple digital tools that introduce concepts for working with data, including text
- Voyant Tools: a dashboard of digital tools for text analysis
- Constellate: a new tool from JSTOR for learning, teaching, and performing text analysis
- Programming Historian: a growing set of lessons/tutorials for text analysis and other digital humanities applications.

Text Analysis. Where?

So as historians, where does the text we might want to systematically analyze with the help of computational algorithms potentially come from?

Different sources of “text”:

- Books
- Magazines
- Newspapers
- Correspondence
- Official records
- Audio transcripts
- Social media posts
- *What else?*

These are very general categories to get us thinking not just about history, from, genre, purpose, audience, and content, but also about the ***structural features*** *of different kinds of texts.*

Text Analysis. Where?

For us to analyze text, it needs to be digital and *machine-readable*. Where can we get that kind of text to work with?

- Digitize print materials and run through OCR
    - This requires a lot of labor and “cleaning” of your resulting data
- Computationally scrape/“mine” text from digital sources
    - This requires additional programming skills and can run into copyright/licensing issues depending on the source
- Use AI-generated transcripts of audio material
    - In addition to transcription mistakes, there are specific cleaning issues, like stripping timestamps
- Use existing digitized corpora
    - https://libguides.library.arizona.edu/dighumantools/text

Text Analysis. When?

We've finally made it to when! There are lots of ways to approach this question, so here are my thoughts about text analysis for emerging historians!

- You have a research question that you think you can explore by systematically examining structure, features, or content of text (whether relating to a genre, a time period, a creator, etc.)
- You don't have a research question, but you want to develop programming skills for exploring text data
- You don't have a research question, but you want to investigate an existing digital corpus that aligns to your research interests

There is nothing wrong with exploring, but just keep in mind that not all research questions require, or are suited for, text analysis.

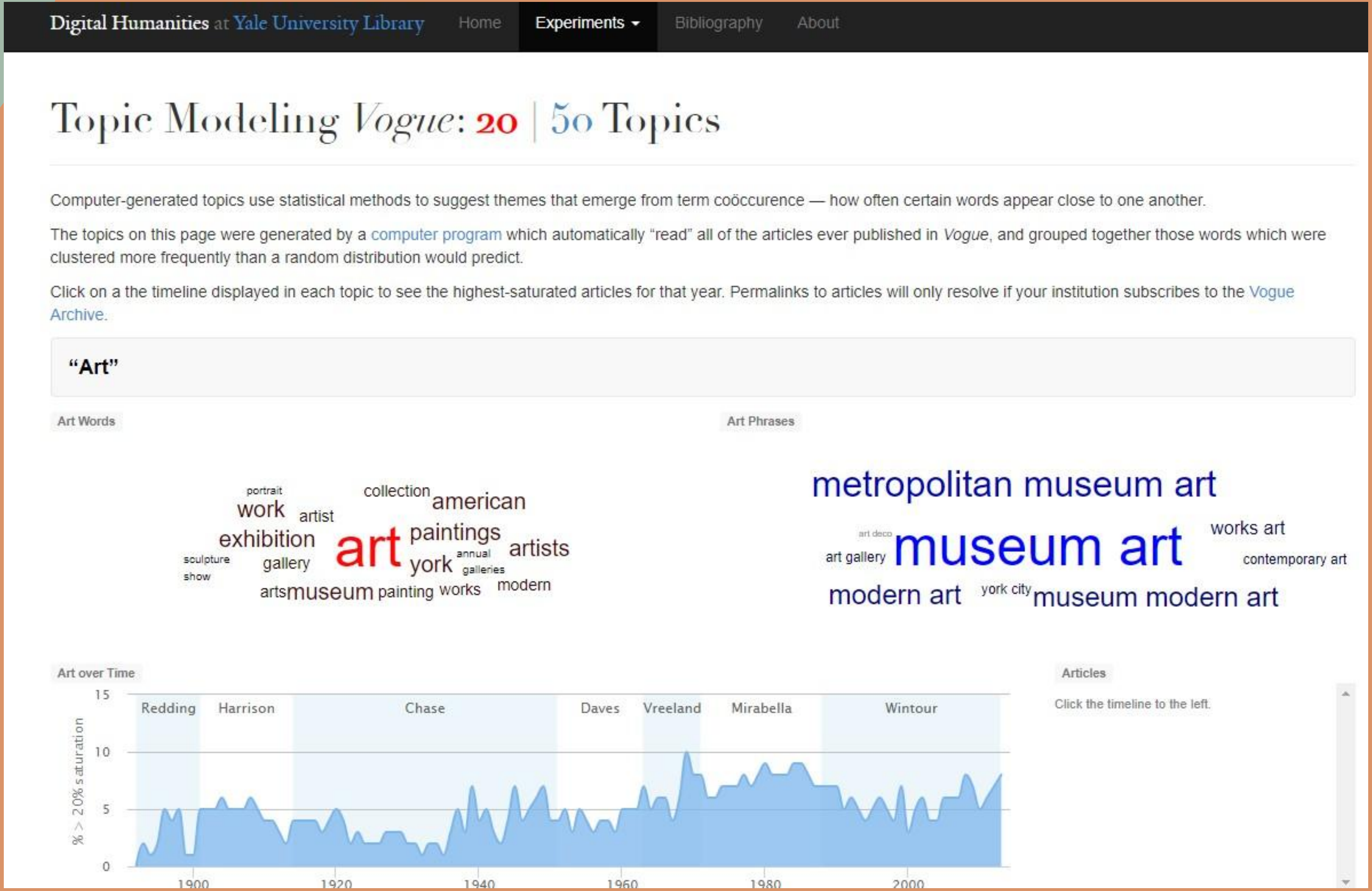
Text Analysis. When?

More stuff to keep in mind around timing/project planning:

- Text analysis *always* requires significant cleaning of data, which takes time and can require many iterations before you can get meaningful results for interpretation
- We often think of cool interactive visualizations as the product of text analysis, but not all programmatic output is ready for presentation
  - Your interpretation and argumentation for significance might come in the form of “regular” written scholarship
  - Many cool interactive visualizations of analyzed text require additional skills, tools, and resources (psst Jo!)

Examples

http://networks.viraltexts.org/1836to1860-simple/



Topic Modeling *Vogue*: 20 | 50 Topics

Computer-generated topics use statistical methods to suggest themes that emerge from term coöccurence — how often certain words appear close to one another.

The topics on this page were generated by a computer program which automatically "read" all of the articles ever published in *Vogue*, and grouped together those words which were clustered more frequently than a random distribution would predict.

Click on a the timeline displayed in each topic to see the highest-saturated articles for that year. Permalinks to articles will only resolve if your institution subscribes to the Vogue Archive.

"Art"

Click the timeline to the left.

http://dh.library.yale.edu/projects/vogue/

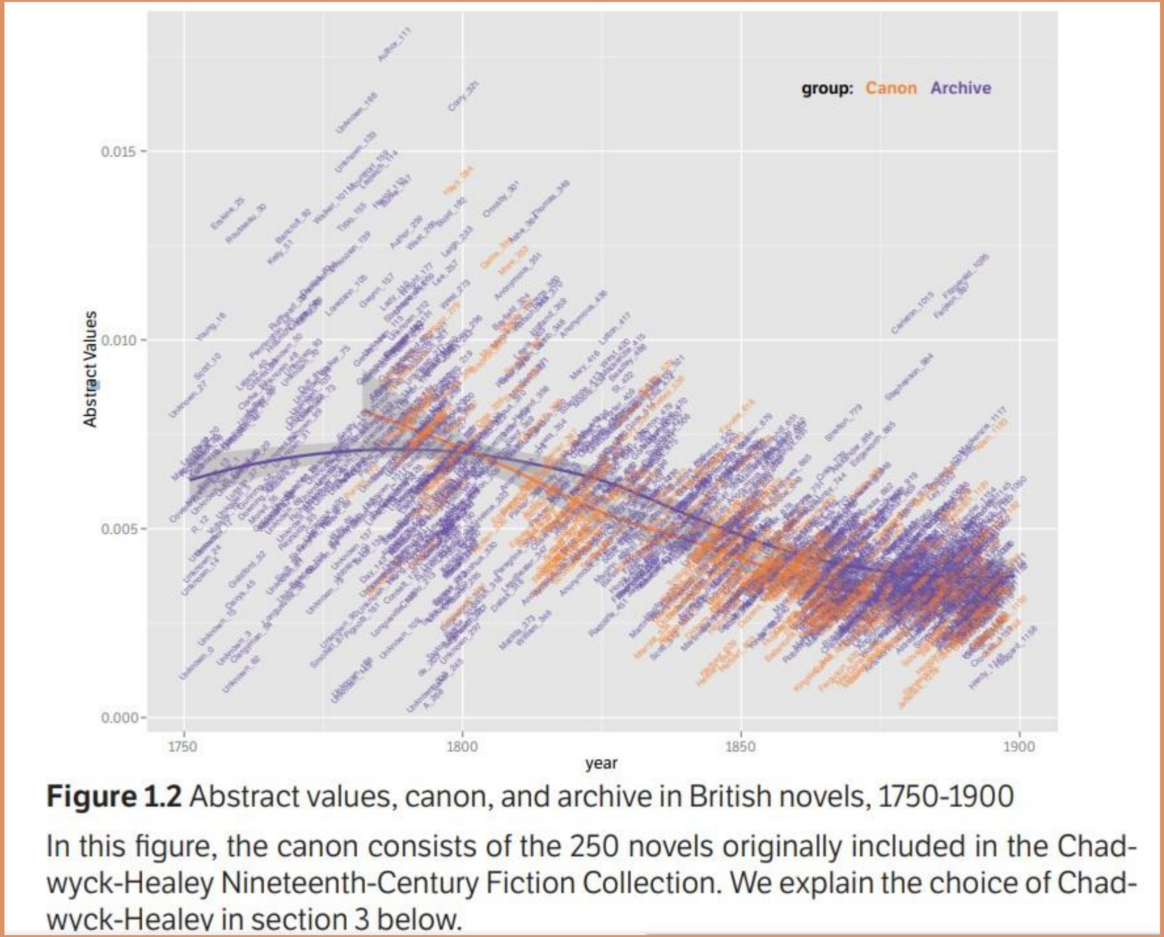
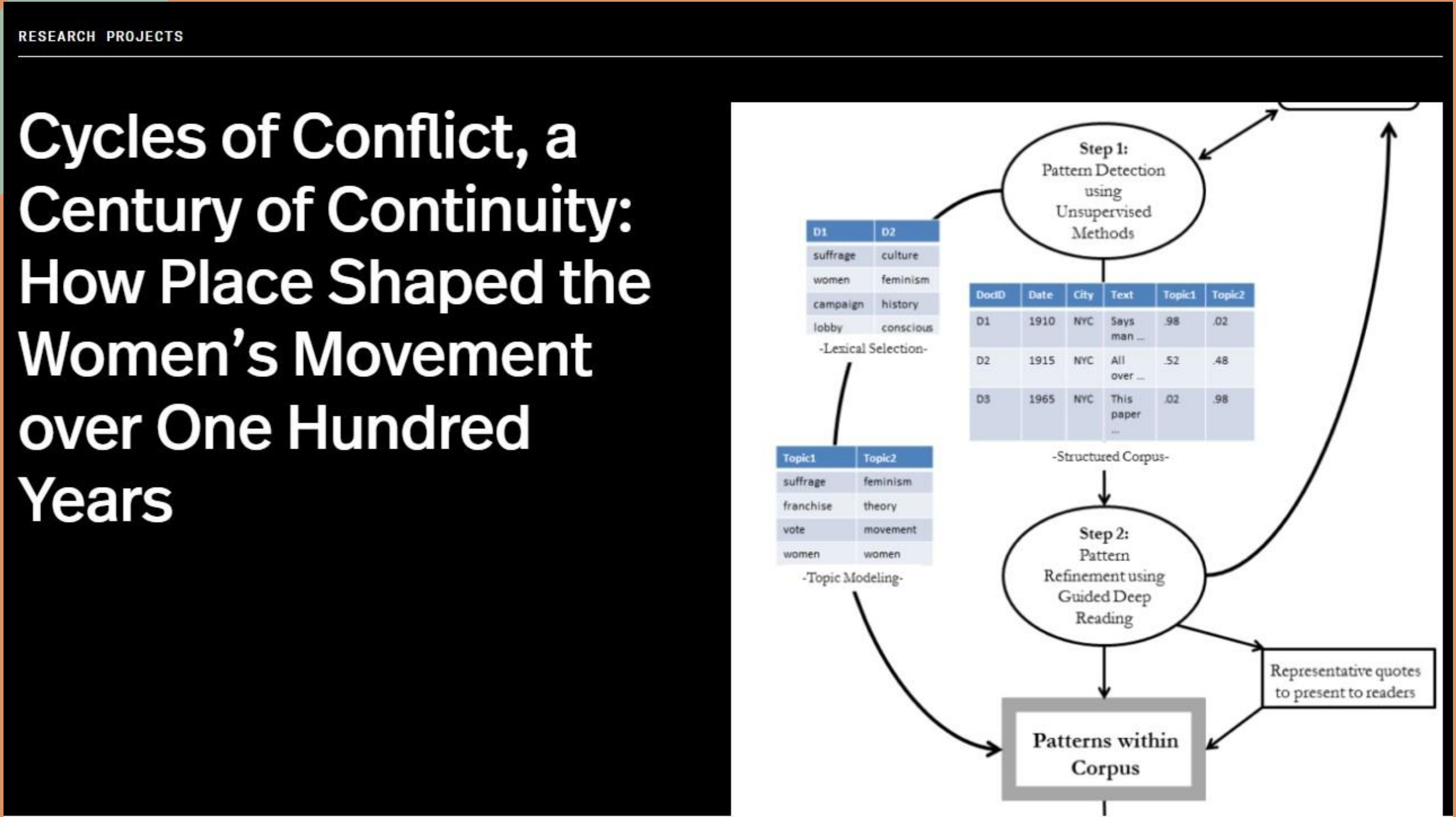


Figure 1.2 Abstract values, canon, and archive in British novels, 1750-1900

In this figure, the canon consists of the 250 novels originally included in the Chadwyck-Healey Nineteenth-Century Fiction Collection. We explain the choice of Chadwyck-Healey in section 3 below.

https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf



RESEARCH PROJECTS

Cycles of Conflict, a Century of Continuity: How Place Shaped the Women's Movement over One Hundred Years

D1	D2
suffrage	culture
women	feminism
campaign	history
lobby	conscious

-Lexical Selection-

DocID	Date	City	Text	Topic1	Topic2
D1	1910	NYC	Says man ...	.98	.02
D2	1915	NYC	All over ...	.52	.48
D3	1965	NYC	This paper ...	.02	.98

-Structured Corpus-

Topic1	Topic2
suffrage	feminism
franchise	theory
vote	movement
women	women

-Topic Modeling-

Step 1:
Pattern Detection using Unsupervised Methods
Step 2:
Pattern Refinement using Guided Deep Reading
Representative quotes to present to readers
Patterns within Corpus

https://cssh.northeastern.edu/nulab/cycles-of-conflict/

Constellate

We are evaluating Constellate this semester; only its free features will be available to use after Spring 2022

Voyant Tools

Let’s grab some text: https://libguides.library.arizona.edu/dighumantools/text

2

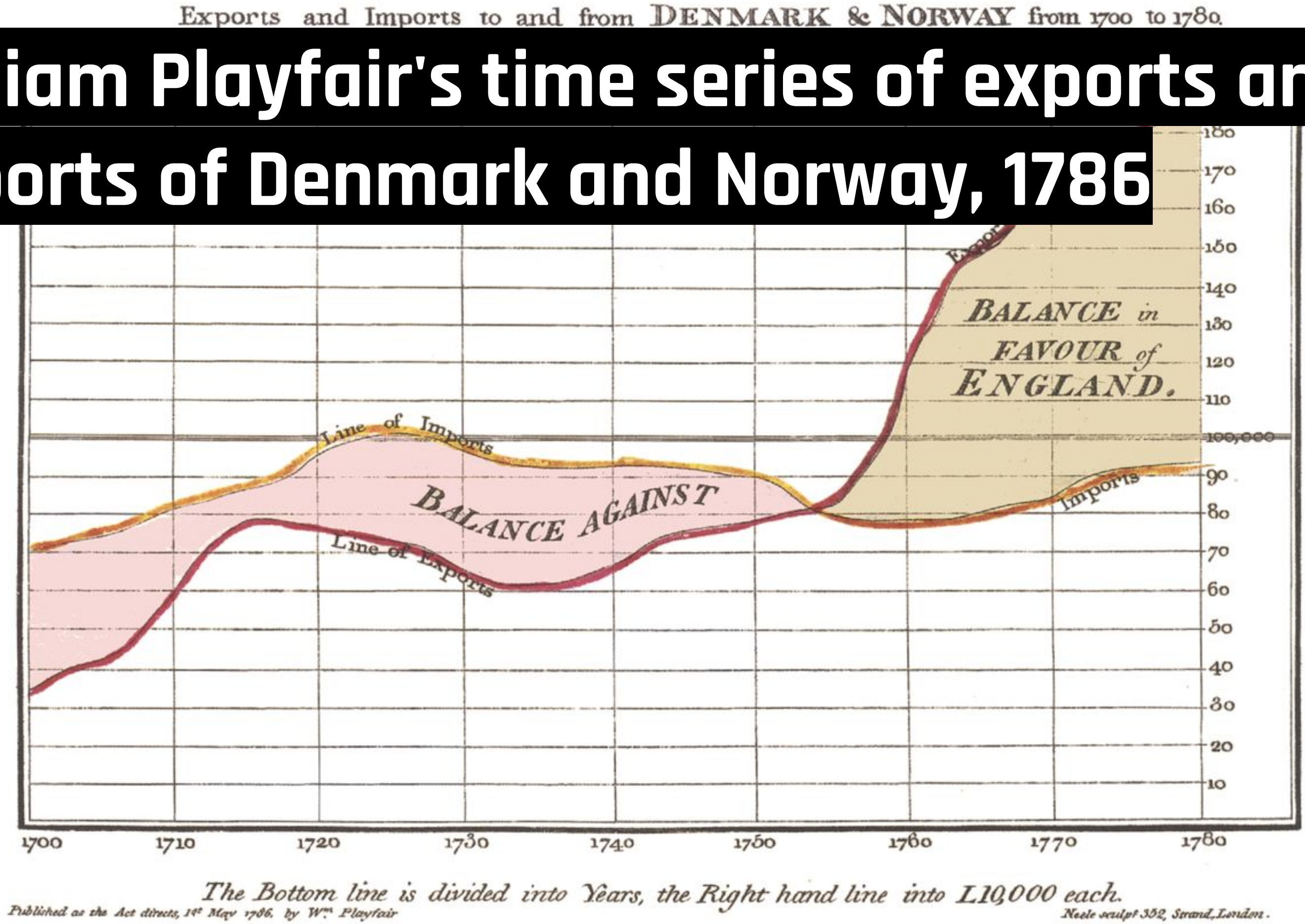
Data Visualization

How we represent data, visually

Data visualizations help us...

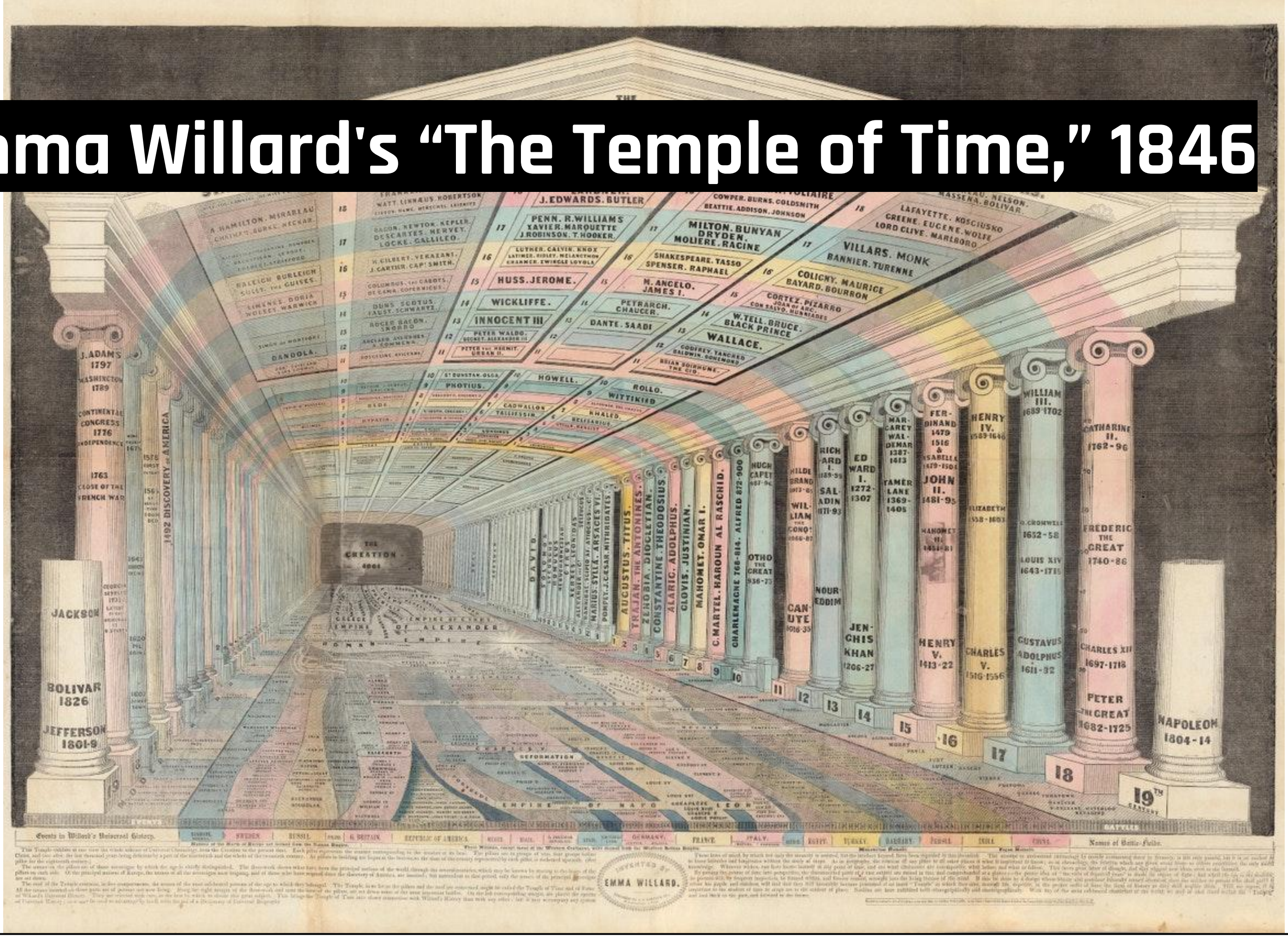
1. See patterns
2. Make comparisons
3. Explore relationships
4. Summarize information

William Playfair's time series of exports and imports of Denmark and Norway, 1786



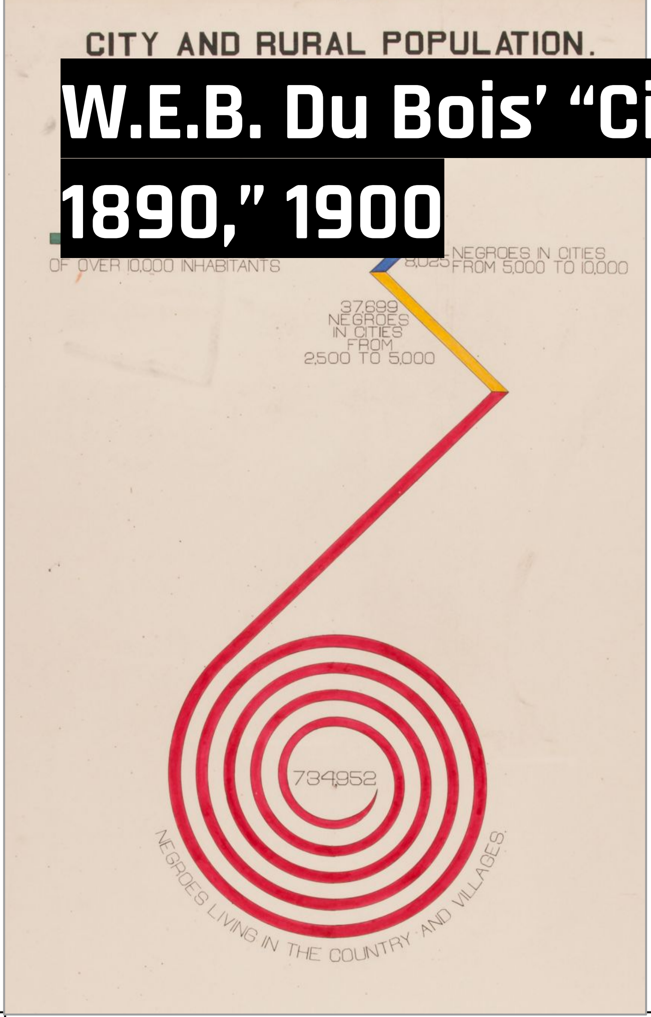
Source: commons.wikimedia.org (retrieved 15 Feb, 2022)

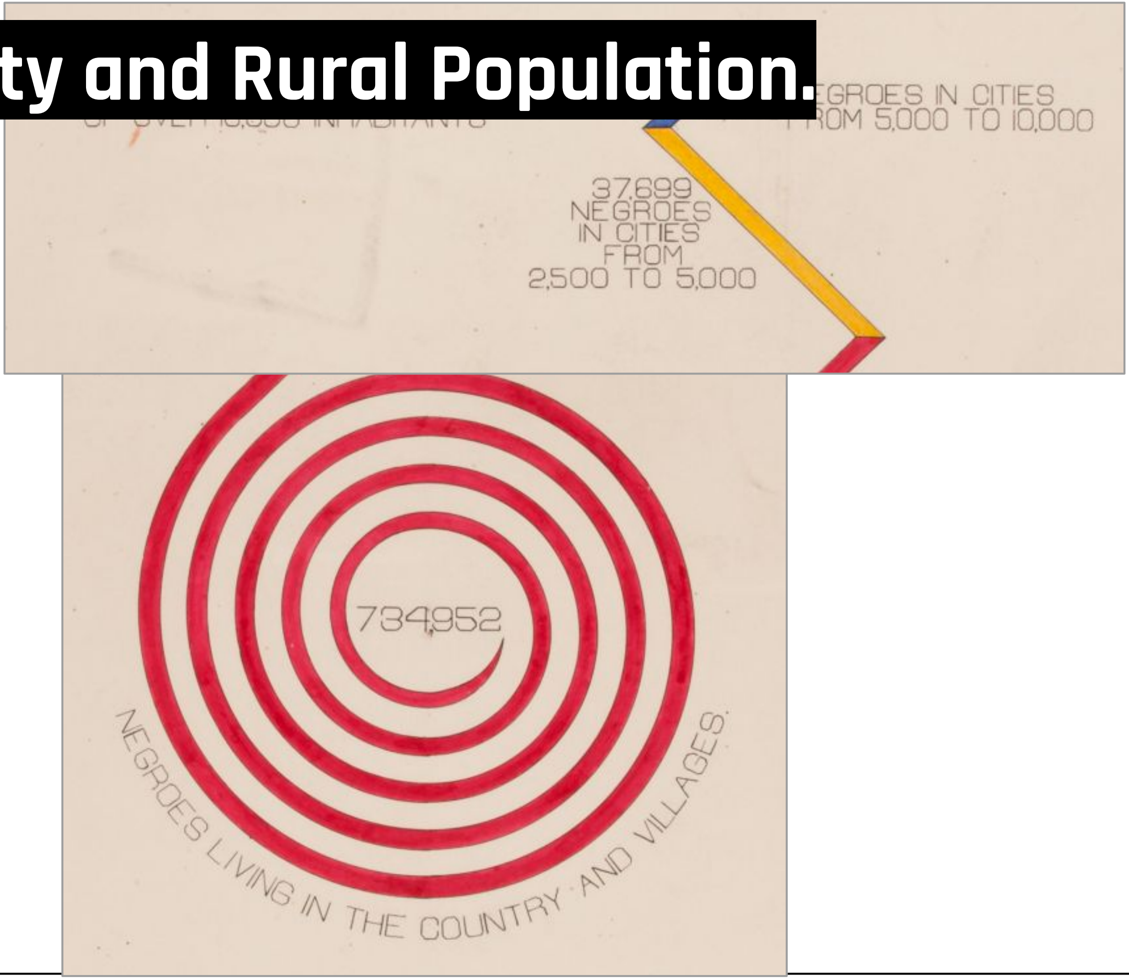
Emma Willard's “The Temple of Time,” 1846



Source: David Rumsey Map Collection (retrieved 15 Feb, 2022)

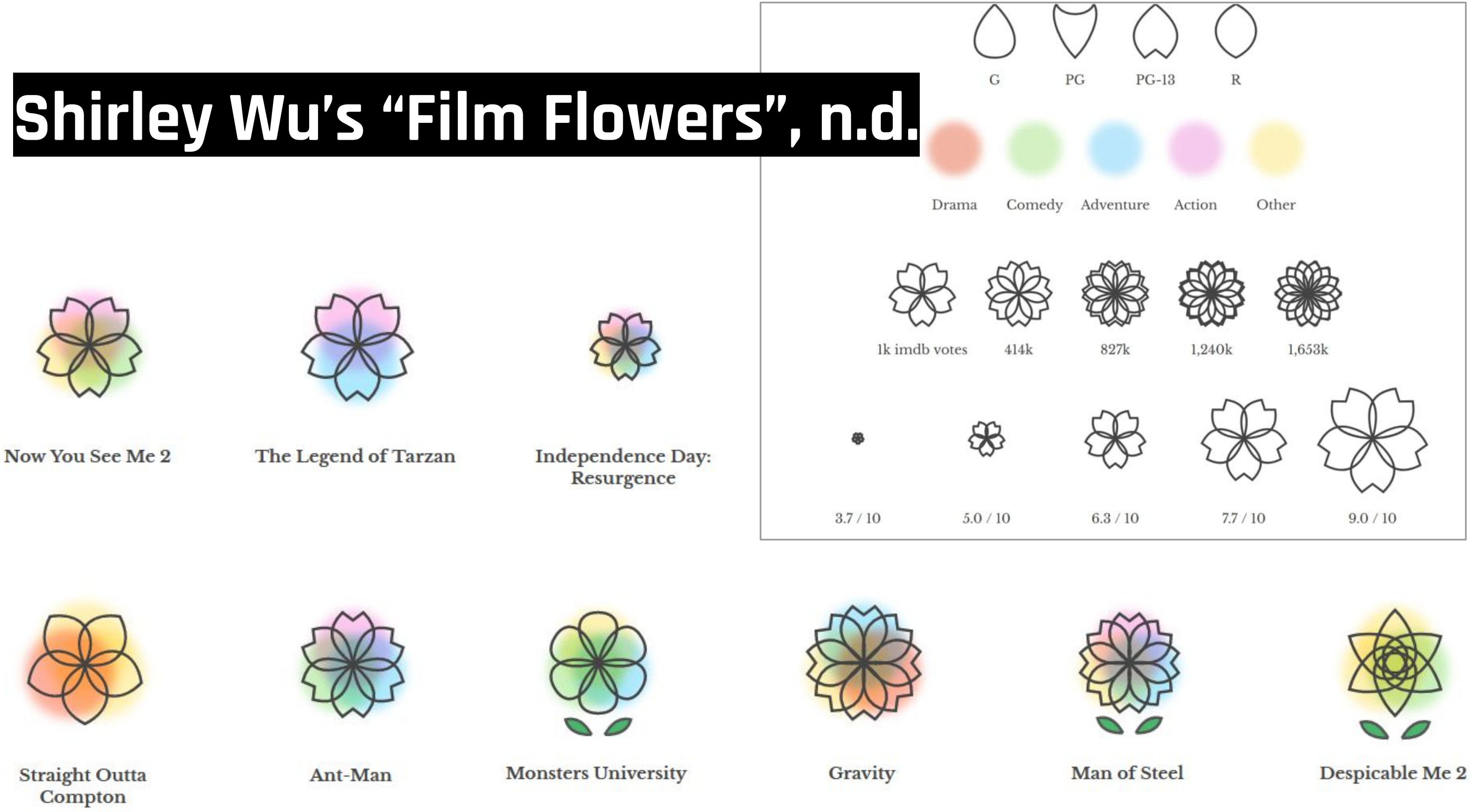
W.E.B. Du Bois' "City and Rural Population. 1890," 1900





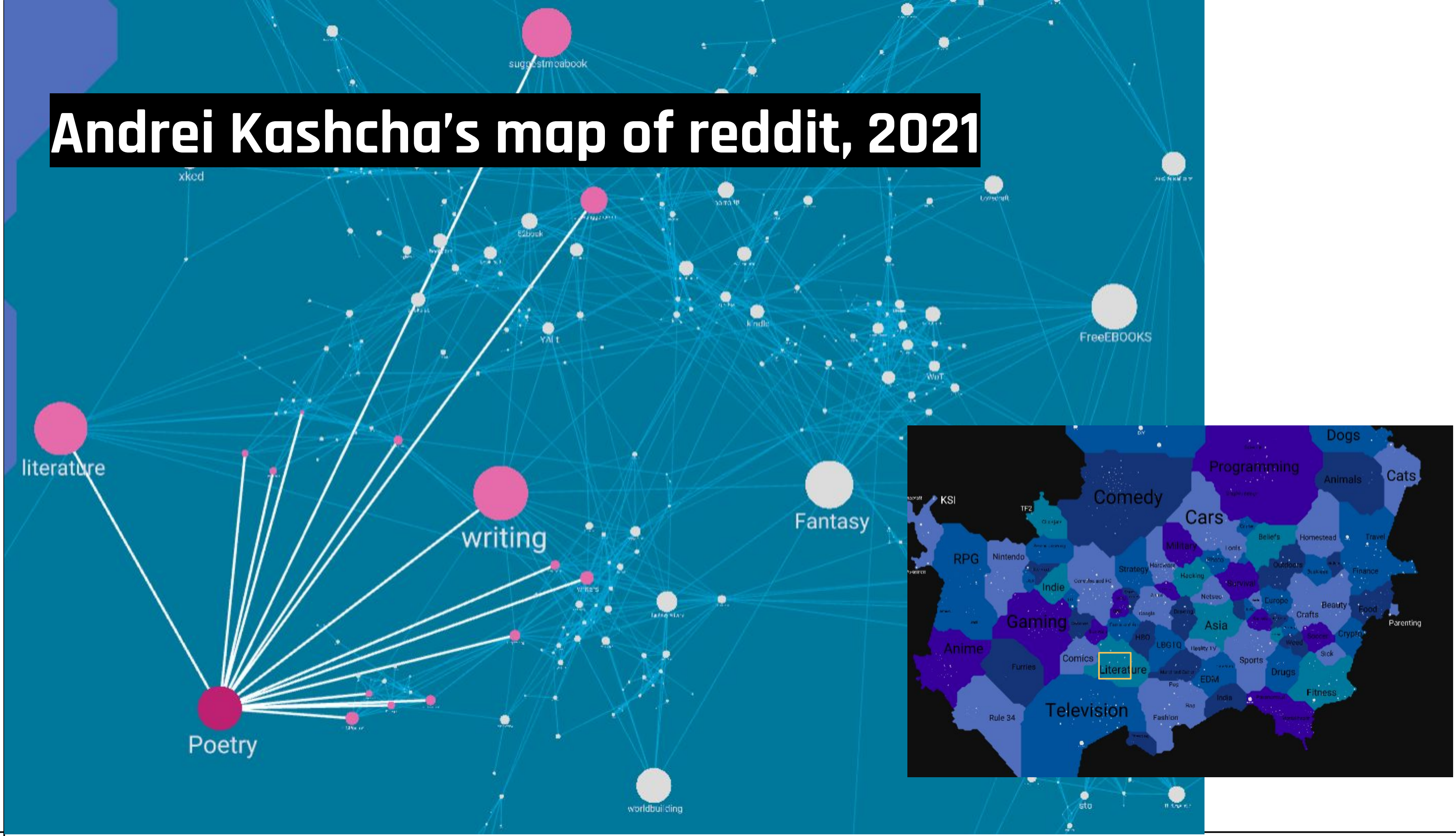
Source: Nightingale (retrieved 16 Feb, 2022)

Shirley Wu's "Film Flowers", n.d.



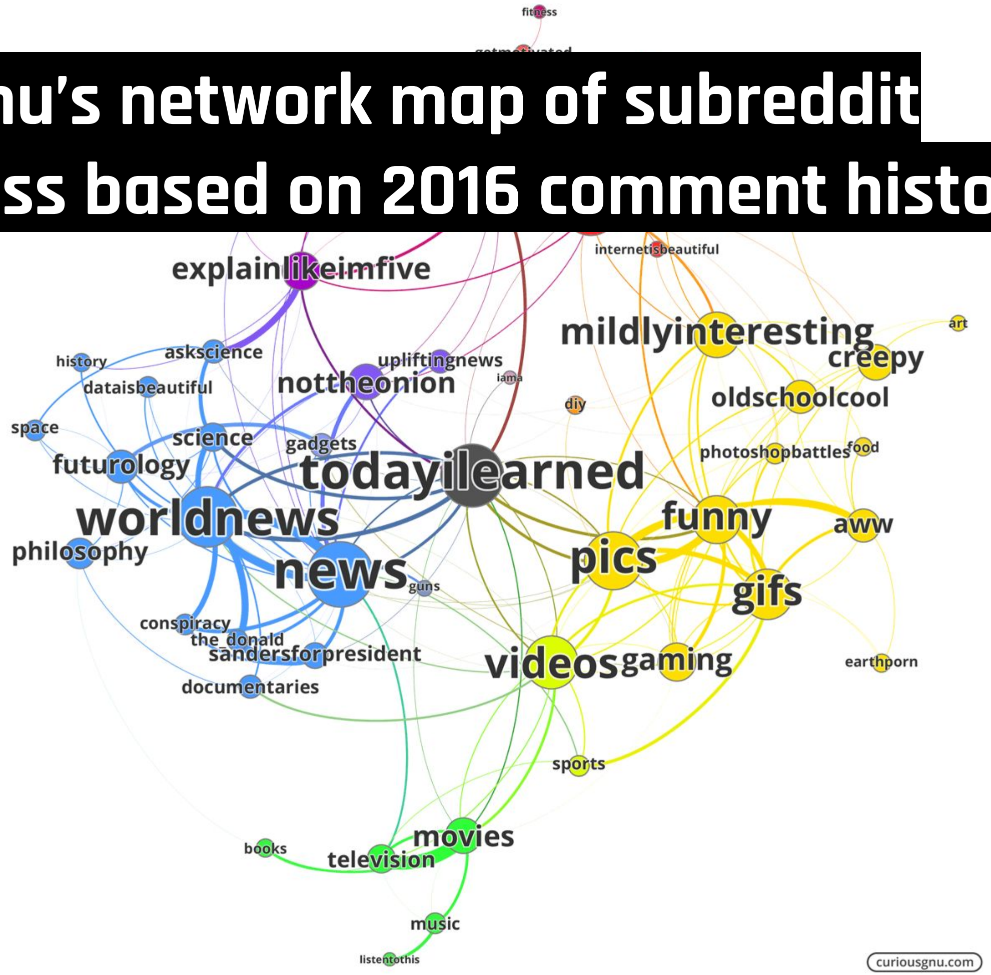
Source: shirleywu.studio (retrieved 16 Feb, 2022)

Andrei Kashcha's map of reddit, 2021



Source: anvaka.github.io (retrieved 16 Feb, 2022)

CuriousGnu's network map of subreddit relatedness based on 2016 comment history, 2016



Source: curiousgnu.com (retrieved 16 Feb, 2022)

Elements

Scale

Angle

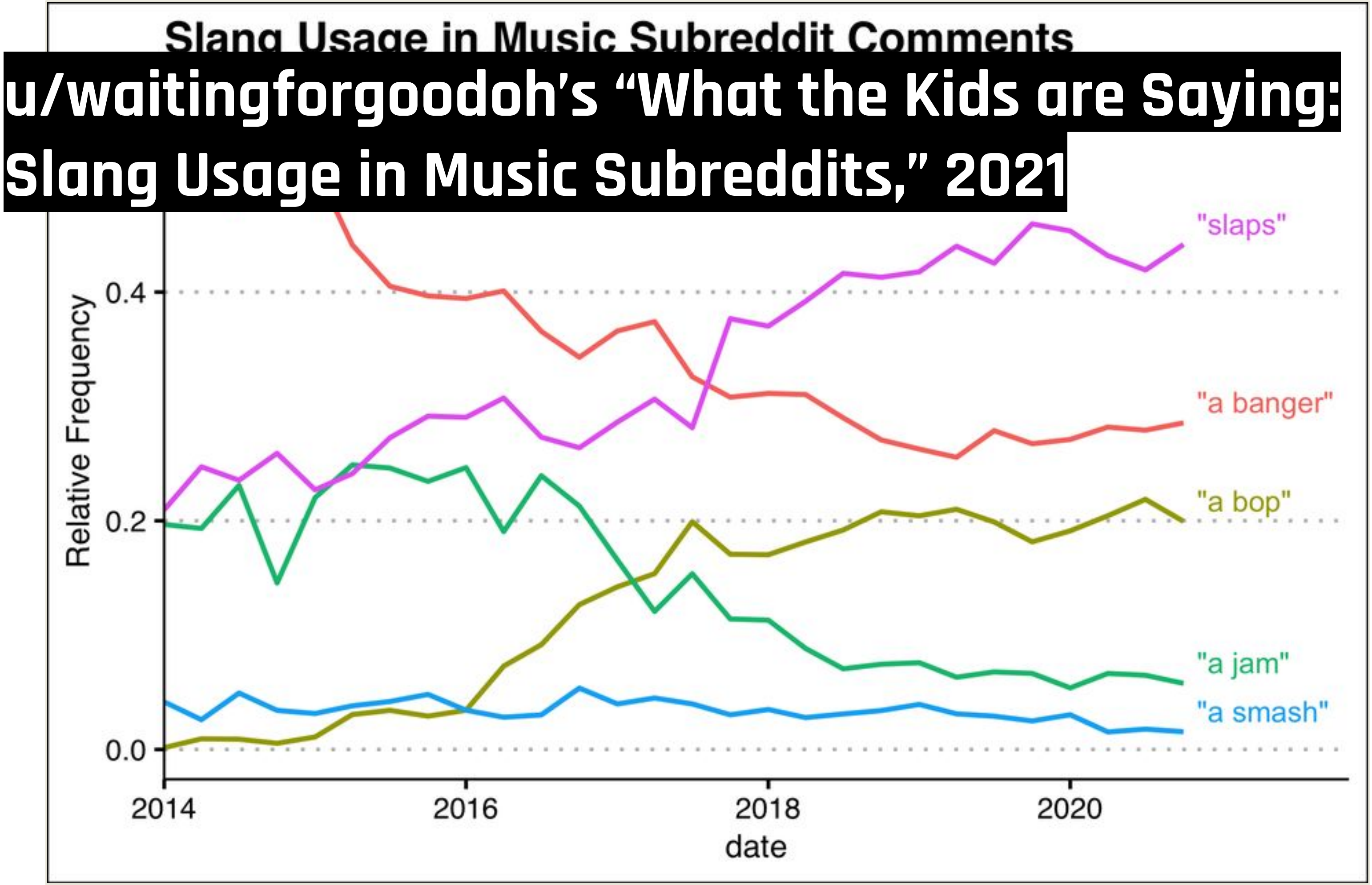
Color

Size

Shape

Text

u/waitingforgoodoh's "What the Kids are Saying: Slang Usage in Music Subreddits," 2021



Source: reddit.com (retrieved 16 Feb, 2022)

Charts, maps, and graphs, oh my!

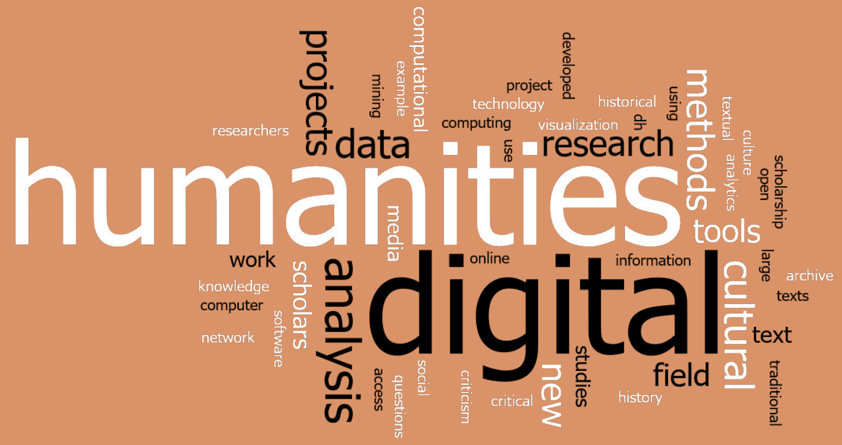
Arc Diagram
Area Graph
Bar Chart
Box & Whisker Plot
Brainstorm
Bubble Chart
Bubble Map
Bullet Graph
Calendar
Candlestick Chart
Chord Diagram
Choropleth Map
Circle Packing
Connection Map
Density Plot
Donut Chart
Dot Map
Dot Matrix Chart
imgflip.com

Two tools to help you pick:

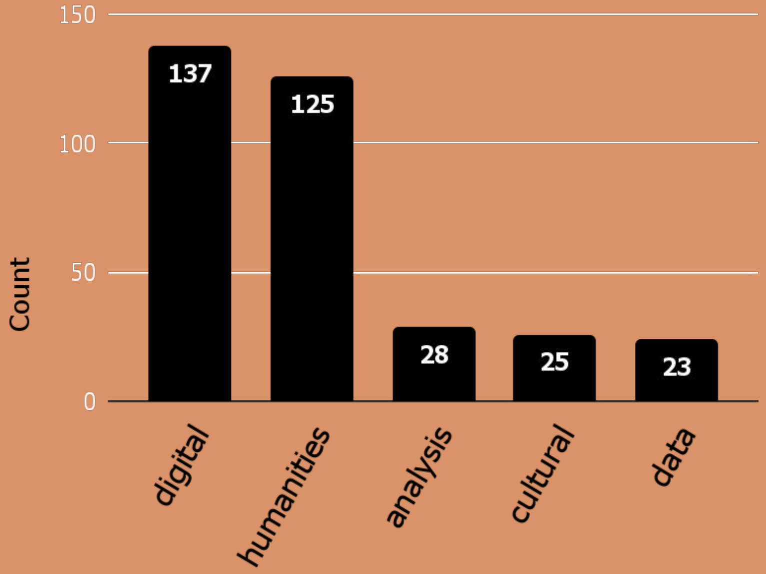
- From Data to Viz decision tree
- The Data Visualisation Catalogue

Source: datavizcatalogue.com (retrieved 16 Feb, 2022)

Form ~~vs.~~ & function

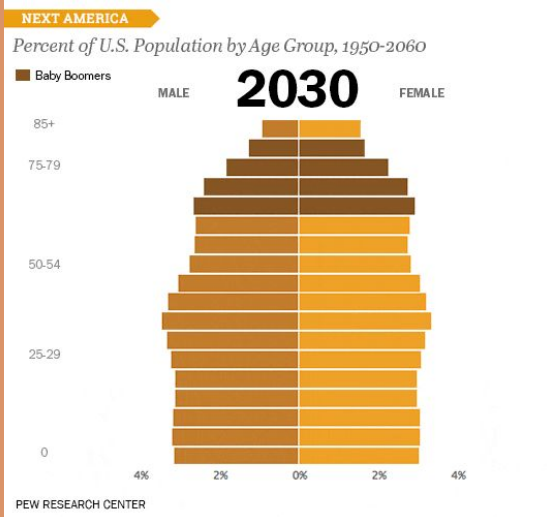


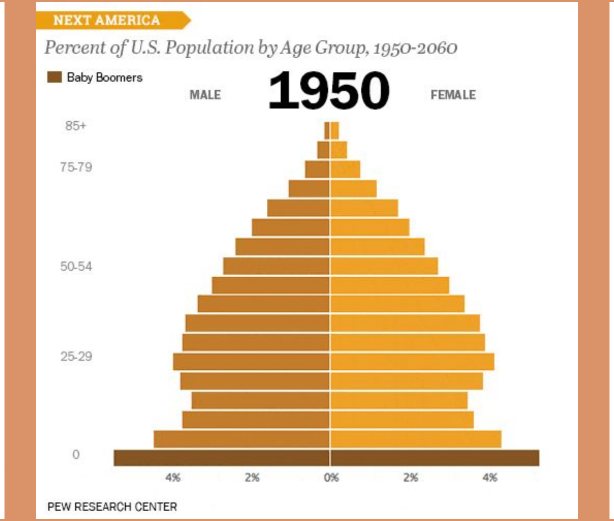
Top five words in the Digital Humanities Wikipedia article, by frequency

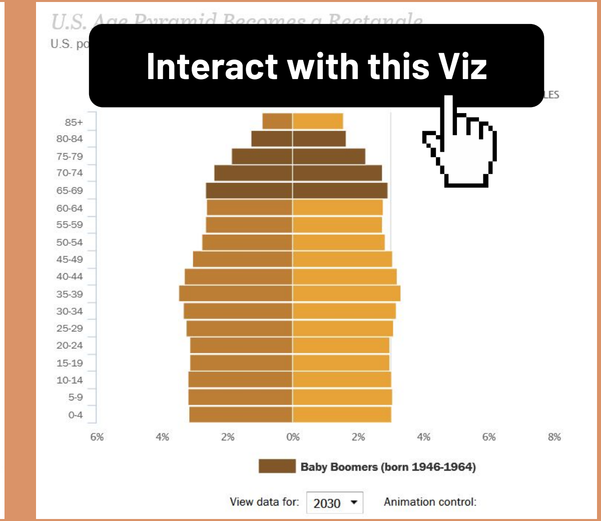


Viz made using Voyant Tools and Google Sheets, with data from Wikipedia retrieved 15 Feb, 2022

Static vs. dynamic, interactive, or live viz







Interact with this Viz

Source: left and center: tableau.com; right: pewresearch.org (retrieved 16 Feb, 2022)

Tools

- “Data” tools: R, **Python**, **Voyant Tools,** Tableau
- Spreadsheet tools: MS Excel, **Google Sheets**, Mac Numbers, Libre Calc
- Web-/Browser-based, often “freemium” tools: **Datawrapper**, RAWgraph
- Visualization-specific tools: Gephi, TimelineJS, GIS tools, ImagePlot
- More at
  - Free Data Visualization Tools for Your Research Toolbelt [slides];
  - “Find Digital Tools” on the Sam Houston State University Digital History Research Guide

Other ways to represent data

“Concretization”

Haptic data: data sculptures, textiles, 3D prints

Sonification

Audible data: sound and music

Source: (upper left) Nightingale; (upper right) Smithsonian Magazine; (lower left) Journal of High Technology Law

Other resources/further reading

Data viz tips & best practices

- Core Principles of Data Visualization Cheatsheet - PolicyViz
- Play Your Charts Right: Data Visualization Tips - Geckoboard
- Data Vis as Guide Dog - Erica Gunn via Nightingale

Examples, tutorials, learning resources

- How to Create a Network Graph Visualization of Reddit Subreddits - Max Woolf's Blog (**BigQuery**, **R/ggplot2 & igraph**)
    - Also available as a Jupyter notebook
- Long, J.D., Teetor, P. (2019). ***R*** *cookbook* (2nd Ed.). https://rc2e.com/
- Chang, W. (2018). ***R*** *graphics cookbook: Practical recipes for visualizing data* (2nd Ed.). https://uncg.on.worldcat.org/oclc/1103606028
- Plotly Python Open Source Graphing Library (**Python/Plotly**)

THANKS!

Maggie Murphy (she/her)
mmurphy@uncg.edu
Schedule an appt: bit.ly/maggiecal

Jo Klein (they/he)
ejklein@uncg.edu
Schedule an appt: go.uncg.edu/joklein-appt

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**