COS 126/CDH

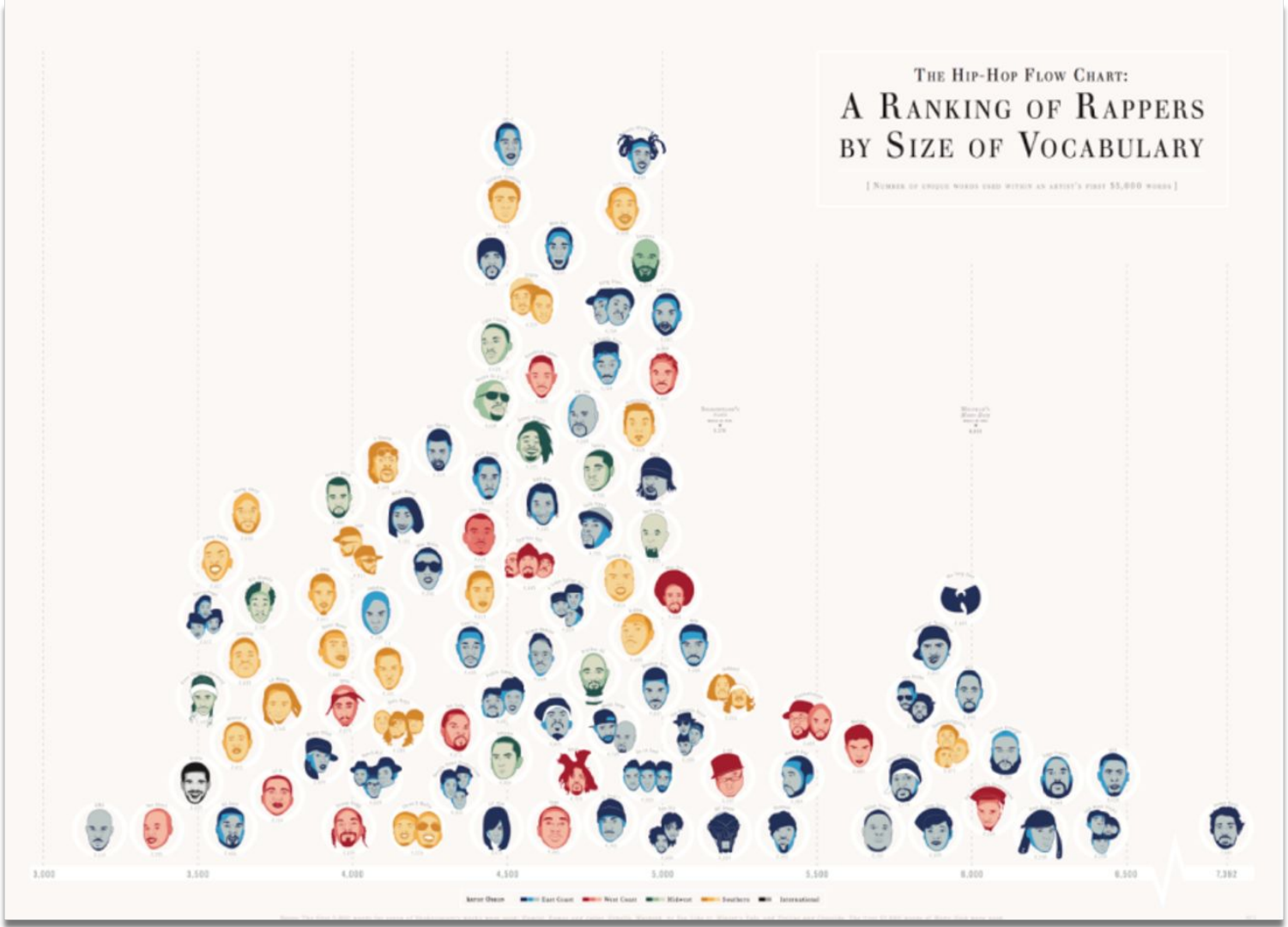
Using Humanities Data

THE CENTER
FOR DIGITAL
HUMANITIES
@PRINCETON

THE CENTER
FOR DIGITAL
HUMANITIES
@PRINCETON

Workshop Agenda

1. Asking questions of data
2. Overview of three humanities datasets
3. Breakout rooms
    a. Choose one of the three datasets you're most interested in and work in a group to define what and how you can ask questions of the data computationally
4. Additional Resources

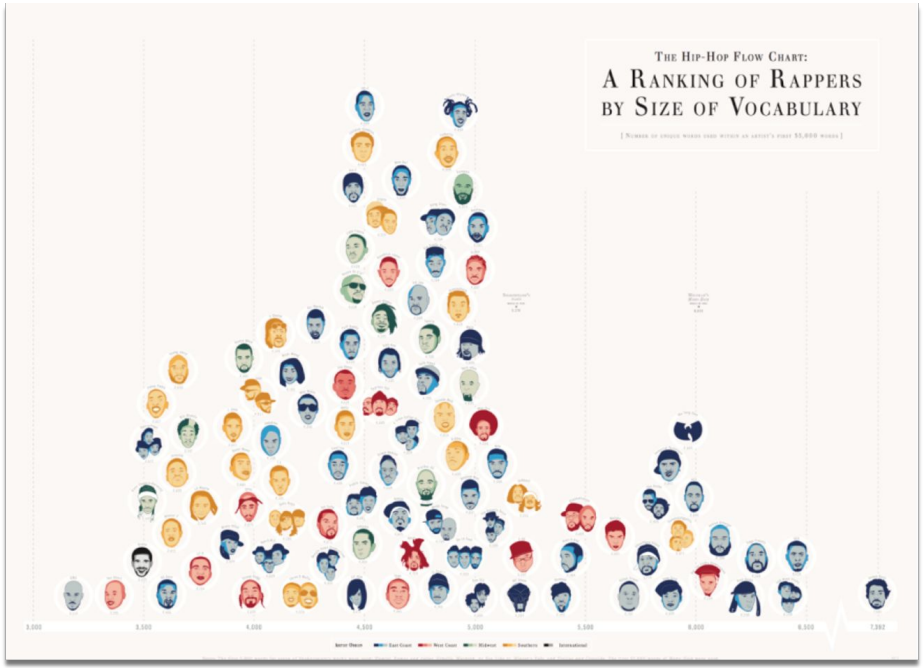


Asking Questions of Data

- What's in your data?
  - What do you have?
  - What's missing?

- What do you want to know?
  - What questions do you have?
  - How would you test these questions computationally with the data you have?

Scoping a question



Rappers ranked by size of vocabulary

Not about:

- Style/Genre
- Popularity
- Album Sales
- Flow

Data

Translating questions to computation

How do you measure “vocabulary” computationally?

- How do rappers rank by the “size of vocabulary?”
  - Here size = number of unique words (strings that only appear once in an individual rapper’s lyrics) by %
- Who uses the most complex words?
  - Complex = longest
  - Find length of strings and rapper who uses the longest strings by %
- Who uses the most neologisms (new words)?
  - Compare corpus to dictionary corpus
  - Find what words in your corpus are NOT in the dictionary
  - Rank rappers by % of neologisms

Featured Humanities Datasets

1. The Endangered Languages Project Data

2. Library Lending Data: Shakespeare and Company

3. The Bechdel Test for Movies

1. The Endangered Languages Project

This project explores the languages spoken worldwide that are in jeopardy of disappearing. These languages are evaluated based on their vitality (the level of risk the language faces).

Data

1. The Endangered Languages Project

The dataset is what was used to make the map and info pop-ups. Columns I (More on vitality) and J (Language context) are qualitative data and the data is gathered from multiple sources.

Code Authority	Language code	Language	AKA	Vitality	Number of speakers	Classification	Variants and dialects	More on vitality	Language context	Location	Continent	Coordinates

Standard

- Geographic plot of where languages are most endangered
  - Customize the map by visualizing the number of speakers (color, scale or change dot size, etc.)

Sprinkle

- Support searching through the dataset:
  - E.g. given a language status (Endangered, Threatened), list all languages under that category

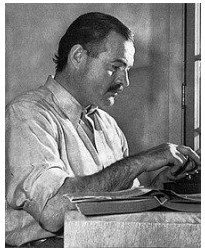
Sparkle

- Create a more powerful search
  - E.g. create a search with multiple criteria

2. Library Lending Data: Shakespeare & Company Project

gertrude stein

james joyce



ernest hemingway

aimé césaire

simone de beauvoir

SHAKESPEARE AND COMPANY — PROJECT —

In **1919**, an American named Sylvia Beach opened Shakespeare and Company, an English-language bookshop and lending library in **Paris**. It became the home away from home for a community of expatriate writers and artists now known as the *Lost Generation*. In 1922, she published James Joyce's *Ulysses* under the Shakespeare and Company imprint, a feat that made her—and her bookshop and lending library—famous around the world. In 1941, she closed Shakespeare and Company after refusing to sell her last copy of Joyce's *Finnegans Wake* to a Nazi officer.

2. Library Lending Data: Shakespeare & Company Project

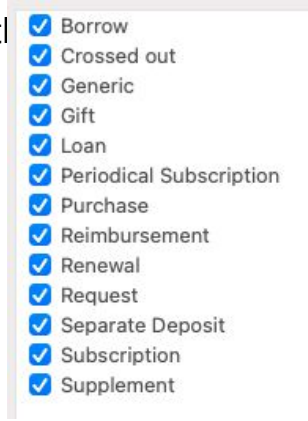
3 separate datasets that include library members, books, and interactions between them
Demographics and addresses for some members; bibliographic data; membership information (subscriptions, renewals, etc) with financials; book borrowing activity.

members dataset, information on ~5,600 lending library members in the following fields:
member name, title, gender, individual or organization, *has lending library card*, birth year, deatl year, *membership years*, VIAF URL, Wikipedia URL, nationalities, addresses, postal codes, arrondissements, longitude and latitude coordinates

books dataset, information about ~6,000 books
title, author, editor, contributor, translator, illustrator, introduction, preface, photographer, *year of publication*, format, uncertain, eBook URL, volume/issue, event count, *borrow count, purchase count, circulation years*

events dataset, information about ~35,000 lending library events in the following fields:
event type*, start date, end date, member name, *subscription price*, deposit amount, duration, *duration in days,* volume limit, category, purchase date, reimbursement amount, *book borrow status*, borrow duration in days, book purchase price, currency, item URL, *title*, volume, author, year of publication, notes, source type, source citation, source manifest, source image.

*event types



2. Library Lending Data: Shakespeare & Company Project

Standard

- Plot home addresses of members (can do more complex things like coloring based on gender and changing size of dot based on number of events etc.)

Sprinkle

- analyze how different aspects of a member's profile (e.g., age, nationality, postcode) correlate with their borrowing habits (e.g. use multiple datasets)

Sparkle

- Given a member's events, make recommendations of other books they might like
- Use the addresses to generate walking tours of Paris
  - Ex.: Given location in Paris, produce shortest tour (TSP!) of 5 nearest addresses in member dataset

3. Bechdel Test in Movies

Data is from a 2014 article "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" by FiveThirtyEight.

Alison Bechdel originally created the test in a 1985 comic. It's criteria are simple:

- The movie has to have at least two women in it
- who talk to each other
- about something other than a man

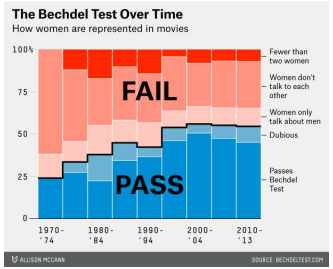
Each movie was evaluated as to whether it passed or failed all three tests.

https://en.wikipedia.org/wiki/Bechdel_test

https://lectures.princeton.edu/lectures/2018/alison-bechdel

3. Bechdel Test in Movies

This dataset was created by merging the data from the Bechdel Test Movie List and general information about the movie industry from the Numbers using movies released between 1990 and 2013.*



You have a lot of variables to explore with this dataset:

year	imdb	title	test	clean_test	binary	budget	domgross	intgross	code	budget_2013$	domgross_2013$	intgross_2013$	period code	decade code	director	director_gender	genre	rating	country	language

Standard

- Plot / correlate different columns (e.g., different color point for PASS/FAIL or the more specific reasons for fairly, for different budgets / domestic gross, rating)
- Analyze the effect that the gender of the director or the gender had on whether or not a movie passed/failed the Bechdel test

Sprinkle

- Given a start and end year, create a timeline plot of the % of movies that pass the Bechdel test
- Of the all different pieces of information available, which is most correlated with failing/passing the test?

Sparkle

- Train a classifier using the different pieces of information available as input features and try to predict whether a movie would pass or fail

* Current movie data, Current Bechdel test info using the API

Featured Humanities Datasets — Breakout

- Breakout Room 1:
  - The Endangered Languages Project Data
- Breakout Room 2:
  - Library Lending Data: Shakespeare and Company
- Breakout Room 3:
  - The Bechdel Test for Movies

Going forward

- Know your data
- Ask questions your data can actually answer
- Break down your question into functions you can construct with the skills you have

Center for Digital Humanities, Emily McGinn
https://cdh.princeton.edu

THE CENTER FOR DIGITAL HUMANITIES @PRINCETON