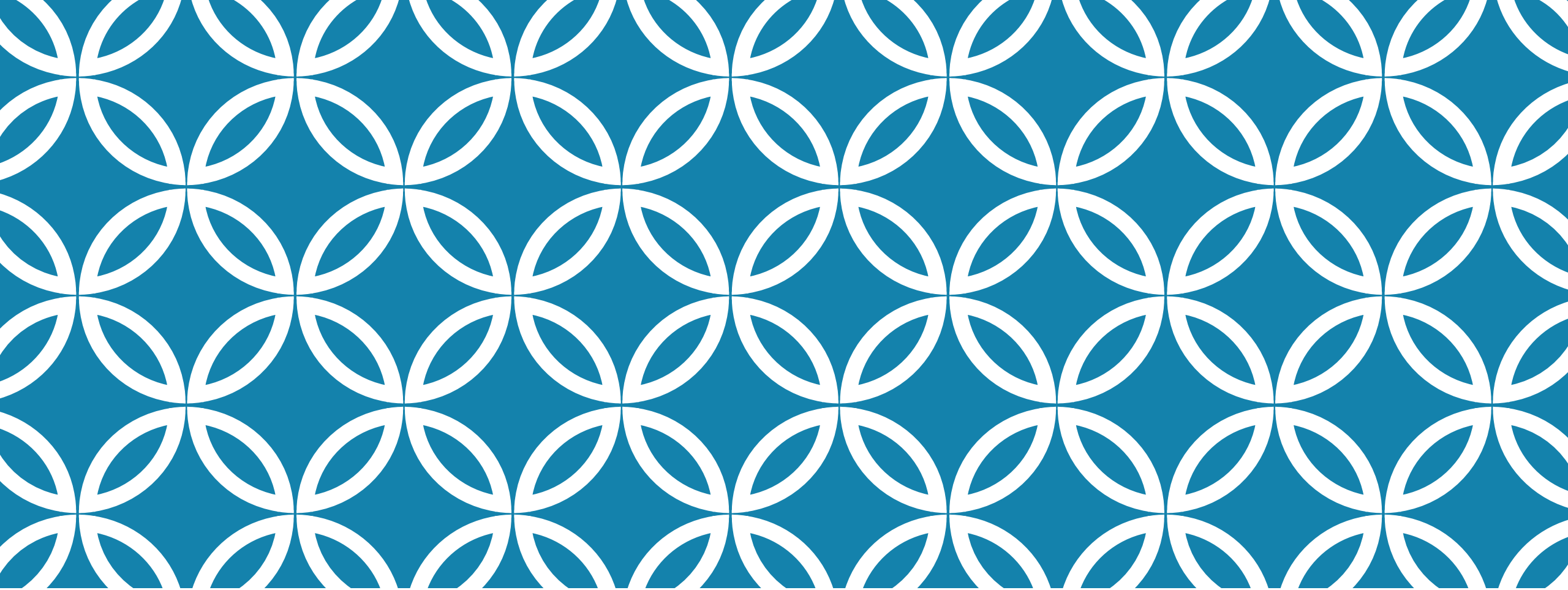

TO READ OR NOT TO READ A JULIA PROJECT

Sean's Group: Frank Mai, Sean Seruya, William Zabet, and Robert Wolfe

OVERVIEW & OBJECTIVE

To Read or Not to Read visualizes a small library of books

Data visualization compares how similar the books in a set are to each other

Comparisons are based on analysis of book text and metadata

Similar books are closer to each other spatially; the center of the user's literary taste is located at the origin (0,0) of a cartesian plane

Unread books can be entered and placed spatially relative to the other books of the library; the closer to the origin, the more likely the user is to like them

TEXT ANALYSIS

Read into strings a set of text files containing the text of books in the library

Measure language properties of each book, including:

Language Complexity (based on mean sentence length, mean word length, and amount of variation in vocabulary)

Language Overlap (based on amount of words in a book which occur in other books in the library vs. only in the book)

METADATA

Use HTTP, Gumbo, and Cascadia packages to pull metadata from Goodreads

Metadata collected includes title, author, and Goodreads rating

Metadata used to display information about book in visualization

DATA STRUCTURES

Words read into a dictionary, with words as keys, and occurrences of words as values

Dictionaries for each book merged into a dictionary containing all words in library

Master dictionary compared to individual dictionaries to determine language overlap

Individual book dictionaries exported to sorted arrays of tuples, used to analyze complexity of language

Database used to hold information about books and library

DATA STORAGE

Text of book is read in, parsing occurs, calculations performed

Metadata collected from Goodreads

Calculated language scores, metadata, and other variables recorded in database (currently a .csv file)

Table exists for each user library (set of ~10 books)

DATA VISUALIZATION

Two variables represented on x and y axes

Defaults to language complexity and language overlap

Center of user's literary profile located at the origin

Ideally, includes a hover-over effect for each book in the visualization which displays metadata about the book

CONSTRAINTS AND LIMITATIONS

Can only work with books with publicly available text (*i*.e., those out of copyright)

Placing books on a two-dimensional plane uses only two variables at a time, requiring use of composite scores for language complexity, overlap, etc.

Desktop application for now; web and mobile interfaces will take too long to implement given the time constraints

STRETCH GOALS

Visualization comparing two (or more) users' libraries to each other (or to a set of known books, like the top ten books on Project Gutenberg)

Include user ratings of books as a point of comparison

Allow user to choose the variables represented on the axes of the visualization

USES OF APPLICATION

Determining whether a user is likely to enjoy a new book

Determining how unusual a book is in comparison to the rest of a writer's oeuvre

Understanding how similar or different two users' literary tastes are (stretch goal)

TECHNOLOGY: WHY JULIA?

Easy to read in entire text files to strings

Easy to parse strings into dictionaries and arrays

Easy to manipulate strings using regex

Access to Python libraries via PyCall

Robust data visualization packages

QUESTIONS?

Or no questions.