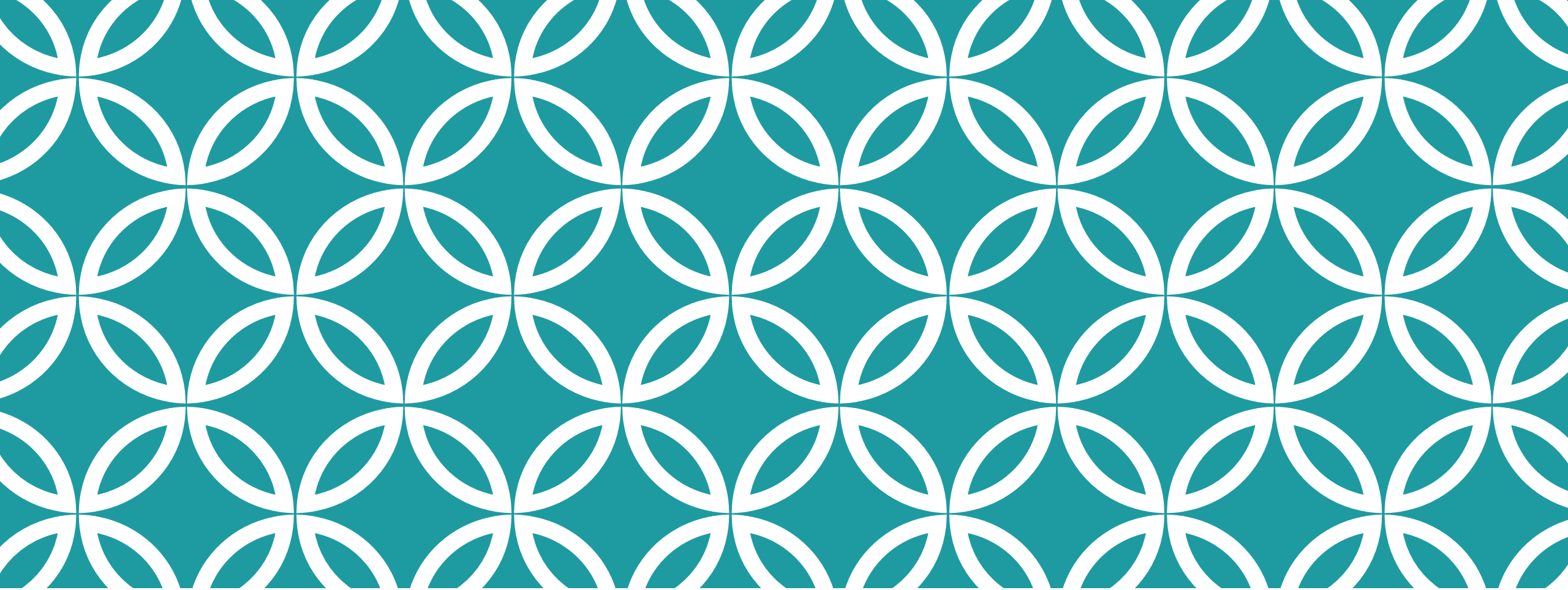

OPENREFINE FOR LIBRARIANS

Dr Maria del Mar Quiroga

VALA Tech Camp

July 2023

MOTIVATION

Cassie Kozyrkov • 2nd
Chief Decision Scientist at Google, Inc.
3mo •

Why #data wrangling is 10% skill, 90% anger management...

#DataScience #statistics #rstats #AI #analytics

Image via Keith McNulty.

PHIADELPHIA	PHILADELPHIA	PHILADLPHIA
PHIALDELPHIA	PHILADELPHIA PA	PHILADPHIA
PHIDELPHIA	PHILADELPHIA,	PHILADRLPHIA
PHIELADELPHIA	PHILADELPHIA, PA	PHILAELPHIA
PHIILADELPHIA	PHILADELPHIA'	PHILDADELPHIA
PHILA	PHILADELPHIAP	PHILDADLPHIA
PHILA.	PHILADELPHIAPHIA	PHILDAELPHIA
PHILAD	PHILADELPHILA	PHILDELPHIA
PHILADALPHIA	PHILADELPHIOA	PHILDEPPHIA
PHILADEDLPHIA	PHILADELPIA	PHILIADELPHIA
PHILADELAPHIA	PHILADELPOHIA	PHILIDELPHIA
PHILADELHIA	PHILADELPPHIA	PHILLA
PHILADELHPIA	PHILADEPHA	PHILLADELPHIA
PHILADELLPHIA	PHILADEPHIA	PHILLY
PHILADELOHIA	PHILADEPHILA	PHILOADELPHIA
PHILADELPH	PHILADEPLHIA	PHLADELPHIA
PHILADELPHA	PHILADERLPHIA	PHOLADELPHIA
PHILADELPHAI	PHILADLELPHIA	PHPILADELPHIA
PHILADELPHI	PHILADLEPHIA	PIHLADELPHIA

Problem	OpenRefine
Data is often very messy	- allows you to identify and amend messy data
It's important to know exactly what you did to your data	- captures all actions applied to your raw data - everything is easily reversed - uses a copy of data, does not modify original
Data cleaning steps often need to be repeated on multiple files	- keeps track of your actions and allows them to be applied to different datasets
Clustering algorithms are complex to implement	- makes it easy to introduce and use

https://openrefine.org/

A free, open source, powerful tool for understanding and cleaning messy data

Works best with tabular data (xlsx, csv, tsv)

'Facets' to get an overview of large datasets

'Clustering' algorithms to fix inconsistencies in data

Helps 'tidy' your data (e.g., split columns)

- Autosaves every 5 mins
- Unlimited undo/redo

Reconciliation services to match data with external databases

- Works with large-ish datasets (100,000 rows)
- Can adjust memory allocation to accommodate larger datasets

Runs a small server on your computer and uses your web browser to interact with it

Your data is kept private on your computer

Mostly point-and-click, but GREL for more sophisticated actions

SESSION OUTLINE

1. Introduction and Overview (10 minutes)
2. Working with OpenRefine
   i. Importing data (10 minutes)
   ii. Layout of OpenRefine (10 minutes)
   iii. Faceting and filtering (20 minutes)
   iv. Clustering (10 minutes)
   Stretch our bodies!
   v. Columns and sorting (10 minutes)
   vi. Introduction to transformations (10 minutes)
3. Connecting OpenRefine with web services (30 minutes)
4. Exporting data (5 minutes)

IMPORTING DATA

OpenRefine does not manipulate your data directly!

- The data you import and all the changes you make are stored in a **project**
- You can stop working on a project and continue later
- You can also export a project and continue working on it on a different computer
- You can upload or import files in a variety of formats including:

  - TSV (tab-separated values
  - CSV (comma-separated values)
  - TXT
  - Excel
  - JSON (javascript object notation)
  - XML (extensible markup language)
  - Google Spreadsheet

LET'S GET OUR HANDS DIRTY!

- Please follow along on your computer
- If at any point anything is unclear, please ask! Chances are there will be at least one other person with the same question
- Don't save your questions for the end, interrupt me when they are relevant
- If you are familiar with any of the content (or if you are a very fast learner) please help your neighbors