Webinar Recording Consent

This webinar **will be recorded** and made available at a future date.

Your continued participation indicates your consent to be recorded.

Processing Data in OpenRefine

Lubov McKone, Data Management Specialist

November 9, 2023

JOHNS HOPKINS
LIBRARIES

Data Services

JHU DATA SERVICES

Before we start, a bit about ZOOM

- Mute audio and video

- Ask questions!
  - Use the public/private chat

- This webinar will be recorded
- You will receive today’s materials by email and on GitHub

Today's software

OpenRefine

Download here: https://openrefine.org/download.html

Recommended web browsers: Chrome, Edge, Safari

Today's materials

Materials available on GitHub:

https://github.com/jhu-data-services/data-cleaning-openrefine

Repository contains:

- These slides
- Workshop data
- Step-by-step workshop guide
- Resources

JHU DATA SERVICES

HELPING YOU NAVIGATE DATA

WE HELP FACULTY, RESEARCHERS AND STUDENTS

FIND

USE

MANAGE

VISUALIZE

SHARE

FIND OUT MORE

GO TO dataservices.library.jhu.edu

EMAIL dataservices@jhu.edu

SHARE AT archive.data.jhu.edu

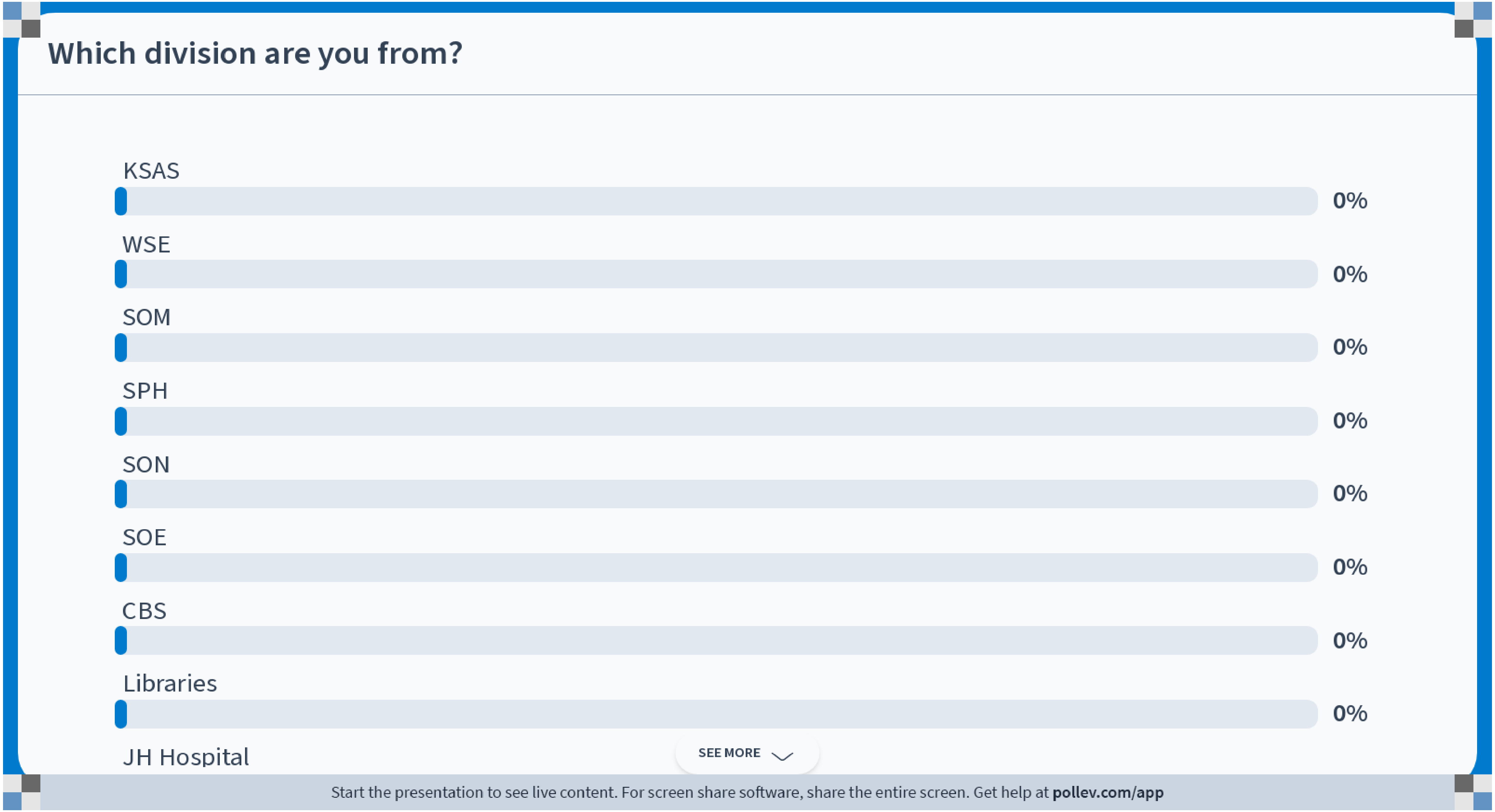
JOHNS HOPKINS LIBRARIES | **Data Services**

Agenda

- Data processing: what and why?
- Introduction to OpenRefine
- Data processing: NUFORC dataset
- Resources

Learning Objectives

- Understand the importance of processing and standardizing data
- Carry out at least three transformations to standardize a dataset
- Become familiar with the reproducible aspects of OpenRefine and how to apply transformations to a new project



Which division are you from?

KSAS 0%

WSE 0%

SOM 0%

SPH 0%

SON 0%

SOE 0%

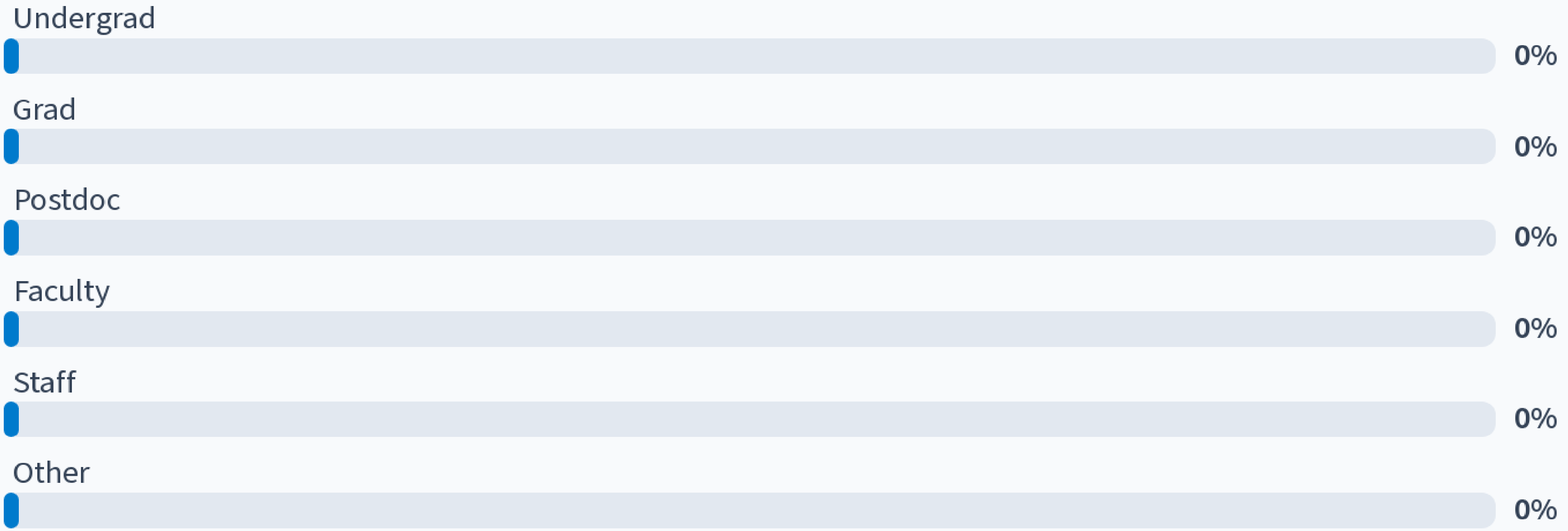
CBS 0%

Libraries 0%

JH Hospital

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

What is your position?



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

What data cleaning or data analysis software do you currently use?

Nobody has responded yet.

Hang tight! Responses are coming in.

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Data Processing: What & Why

80% of data analysis is spent on the process of cleaning and preparing the data (Wickham, 2014)

"Data cleaning" is a consequential step in the research process that we often make opaque by the way we talk about it. (Rawson & Munoz, 2016)

~~Data cleaning~~ → Data processing

"Data cleaning" implies that there is some kind of pure or clean data buried in a thin layer of non-clean data, and that one need only hose the dataset off to reveal the hard porcelain underneath the muck. In reality, the process is more like deciding how to cut into a piece of material, or how much to plane down a surface. It's not that there's any real distinction between good and bad... **Judgement is critical.**" (David Mimo, 2014)

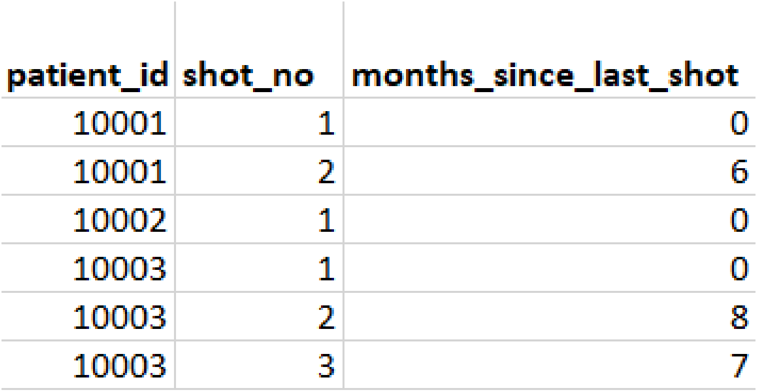
What does data processing involve?

Standardizing & **preparing your data for analysis**

- Coding categorical variables, dates, and missing values consistently & appropriately for your type of analysis
- Addressing misspellings & inaccurate information
- Restructuring your data to suit your type of analysis

Example: Standardizing missing values

You received the data below that was processed by another researcher. The goal of your project is to find the average number of months between shots. What problems might arise from how the data has been processed?



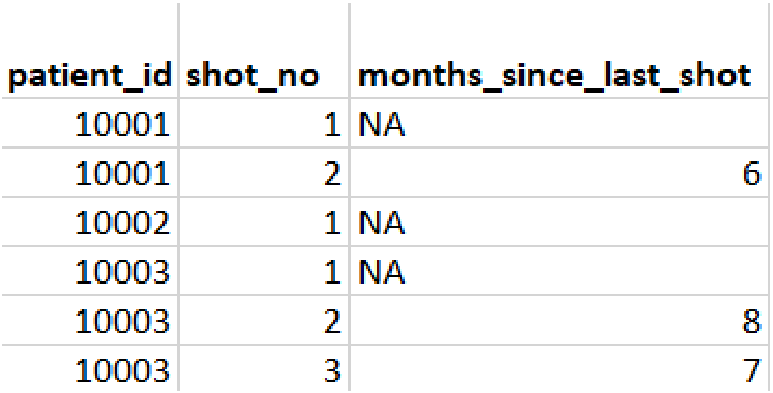
patient_id	shot_no	months_since_last_shot
10001	1	0
10001	2	6
10002	1	0
10003	1	0
10003	2	8
10003	3	7

Average months between shots = 3.5

✗

Example: Standardizing missing values

How would you process the data so that it can accurately answer your research question?



patient_id	shot_no	months_since_last_shot
10001	1	NA
10001	2	6
10002	1	NA
10003	1	NA
10003	2	8
10003	3	7

Average months between shots = 7

✓

Things to look for

- Special characters (e.g. commas in numeric values)
- Numeric values stored as text/character data types
- Duplicate rows
- Misspellings
- Extreme values
- Leading or trailing white space
- Missing data
- Zeros instead of null values

Why process your data?

Makes data reliable and reusable

Saves time!

Facilitates further analysis or visualization, especially in specialty software

Your analysis is only as good as your data

OpenRefine

What is OpenRefine?

https://openrefine.org/

- Free, open source tool for processing data (previously developed by Google, as GoogleRefine)
- Application opens in a web browser, runs on a local server
- Perform actions using graphical user interface (GUI) or writing expressions in General Refine Expression Language (GREL)
- Stores all actions and transformations for a project, can be replicated in new projects

What can OpenRefine do?

- Prepare data for further analysis or visualization
- Good with text data
- Add data from external sources – reconciliation and APIs
- Save your transformations to apply to new dataset – reproducibility!
- Works for medium to large datasets – 100,000s of rows

OpenRefine and Data Security

- OpenRefine is installed locally and stores data locally on your computer
- Does not send data outside of local environment (exception: Reconciliation)
- **Reminder:** it is the researcher’s responsibility to keep data secure

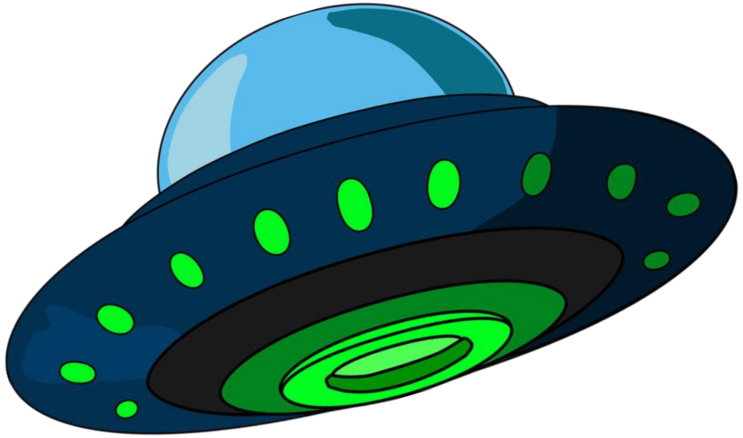
OpenRefine and Accessibility

- Main interface components should be compatible with text-to-speech software
- Many actions (clustering, reconciliation, etc.) may not be compatible with assistive technology
- Developer note in OpenRefine FAQ (scroll down to Accessibility heading)

Dataset: National UFO Reporting Center (NUFORC)

How we will process today's data:

- Standardize values through editing, clustering, and writing GREL expressions
- Remove duplicate rows
- Split one column into two columns
- Reconcile data against external sources
- Add data from an external source



See the **Workshop Guide** for step-by-step instructions

Resources

Data Processing

- Against Cleaning, Katie Rawson & Trevor Munoz
- Towards Data Science: The Ultimate Guide to Data Cleaning
- Tidy Data by Hadley Wickham, Journal of Statistical Software

  DOI: 10.18637/jss.v059.i10
- The Programming Historian: Understanding Regular Expressions

  DOI: 10.46430/phen0033
- Regular Expression Cheat Sheet: https://regexcheatsheet.com/

Using OpenRefine

- OpenRefine Official Documentation
- The Programming Historian: Cleaning Data with OpenRefine

  DOI: 10.46430/phen0023
- Tutorial: Cleaning Data with OpenRefine
- University of Illinois Libguide: OpenRefine

Upcoming Data Services Workshops

- November 14-15th: Interactive Data Visualization in R with Shiny
- November 15th: Joining Data with ArcGIS Online
- November 28th: Best Practices for Data Management & Sharing
- December 5th: All About Sharing Data on the Johns Hopkins Research Data Repository

Register here: https://dataservices.library.jhu.edu/training-workshops/calendar/

JOHNS HOPKINS
LIBRARIES

Data Services

Contact JHU Data Services

GO TO

dataservices.library.jhu.edu

EMAIL

dataservices@jhu.edu

SHARE DATA AT

archive.data.jhu.edu

Helping you

FIND **USE**

MANAGE **VISUALIZE** **SHARE**

DATA

TAKE OUR
SURVEY

https://www.surveymonkey.com/r/openrefine

JOHNS HOPKINS
LIBRARIES | **Data Services**