DETECTING COMMON DATA ELEMENTS IN CENTRAL REPOSITORY DATASETS

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, lead ICO); National Eye Institute (NEI); National Institute of Minority Health and Health Disparities (NIMHD); National Institute of Allergy and Infectious Diseases (NIAID); Center for Scientific Review (CSR); National Library of Medicine (NLM)

Co-mentors: Ken Wilkins, Jenna Norton, Rebecca Rodriguez (DK); Kerry Goetz (EI); Luca Calzoni, Deborah Duran (HD); Melanie Laffin (AI); Adrian Vancea (SR); Robin Taylor (LM)

CIVIC DIGITAL FELLOWSHIP

NIH

SARAH MACHARG

Stanford University

PROJECT LANDSCAPE - CDES

- National Library of Medicine (NLM) Common Data Element (CDE) Repository
  - CDEs can be exported as JSON files ⇒ easily machine readable
- Common Data Elements: Standardized measurements to facilitate reuse and interoperability of research
- Ongoing efforts to standardize and endorse CDEs, and to develop CDEs for important emerging fields like social determinants of health

https://cde.nlm.nih.gov/

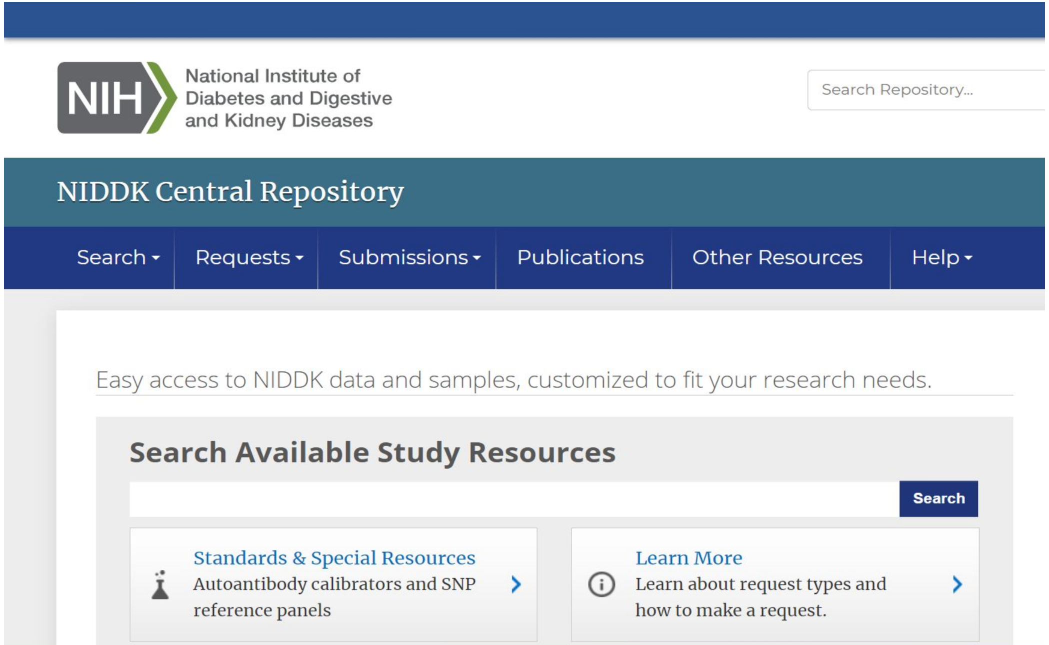
PROJECT MOTIVATION

- Long-term motivation: Standardize and promote use of CDEs, make NIH-funded data more FAIR:
    - **F**indable
    - **A**ccessible
    - **I**nteroperable
    - **R**eusable
- Immediate motivation: NIDDK Central Repository hosts data related to key COVID risk factors - obesity, inflammation, social determinants of health
    - Need to integrate research on these fields ⇒ Need to make existing data on these factors interoperable

CIVIC DIGITAL FELLOWSHIP

PROJECT FOCUS – SCOPE TO NIDDK CR

- NIDDK Central Repository, as one example of an NIH data repository
  - Data and biospecimen resource, controlled access
  - Ongoing and completed multi-center and large single-center studies
  - Data requests require IRB approval



https://repository.niddk.nih.gov/home/

CIVIC DIGITAL FELLOWSHIP

PROJECT LANDSCAPE – NIDDK CR

Study Name

Number of Subjects in Study Archive:
Study Design:
Conditions:
Division:
Duration:
Recruitment Centers:
Treatment:
Available Genotype Data:
Image Summary:
Transplant Type:
Does it have dialysis patients:

General Description

Resources Available

Study Datasets Only

Request

Publications (105)

Resources Available

Study Datasets Only

Request

Publications (105)

Study Documents

DSIC
Forms
Manual of Operations
Data Dictionary (PDF) – 402.6 KB
Roadmap (PDF) – 216.4 KB

CIVIC DIGITAL FELLOWSHIP

PROJECT OBJECTIVES

- **Primary Objective:** Develop Natural Language Processing (NLP) pipeline to extract Common Data Elements (CDEs) from the NIDDK Central Repository
- **Secondary Objectives:**
  - Explore current state of Common Data Element repository and NIDDK Central Repository
  - Draft recommendations for defining and storing Common Data Elements
  - Draft recommendations for NIDDK Repository data collection, storage, and accessibility
    - What should we require of funded researchers when they submit their data to the repository?
    - How can we improve public-facing and controlled-access information on data and metadata to facilitate future research?

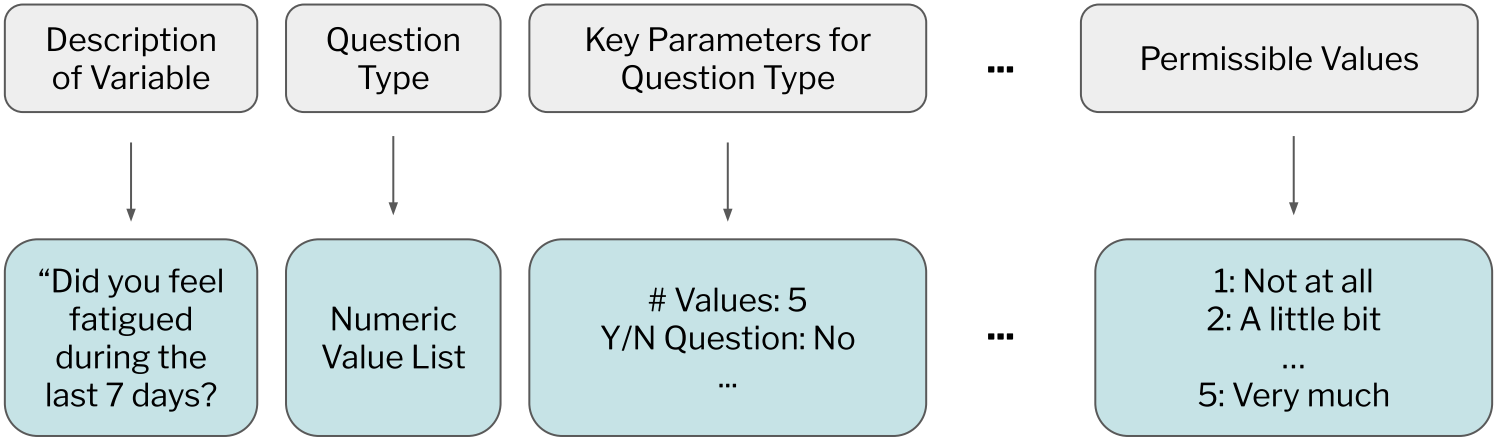
HOW DO WE REPRESENT A CDE?

Stakeholder research identified key CDE features:

- Unique identifiers for data element and its permissible values
- Clear context for what an answer means
- Clear units of measurement
- Title and short definition
- Where has the CDE been used?
- Datatype, parameters
    - Is it a multiple choice question?
    - What is the range of allowable values?
- Measurement protocol, preferred text for questionnaires

REPRESENTING A DATA ELEMENT

- Stakeholder needs + CDE repository structure → a vector of features
- Extract features from CDE repository and NIDDK CR datasets (SAS metadata)



CIVIC DIGITAL FELLOWSHIP

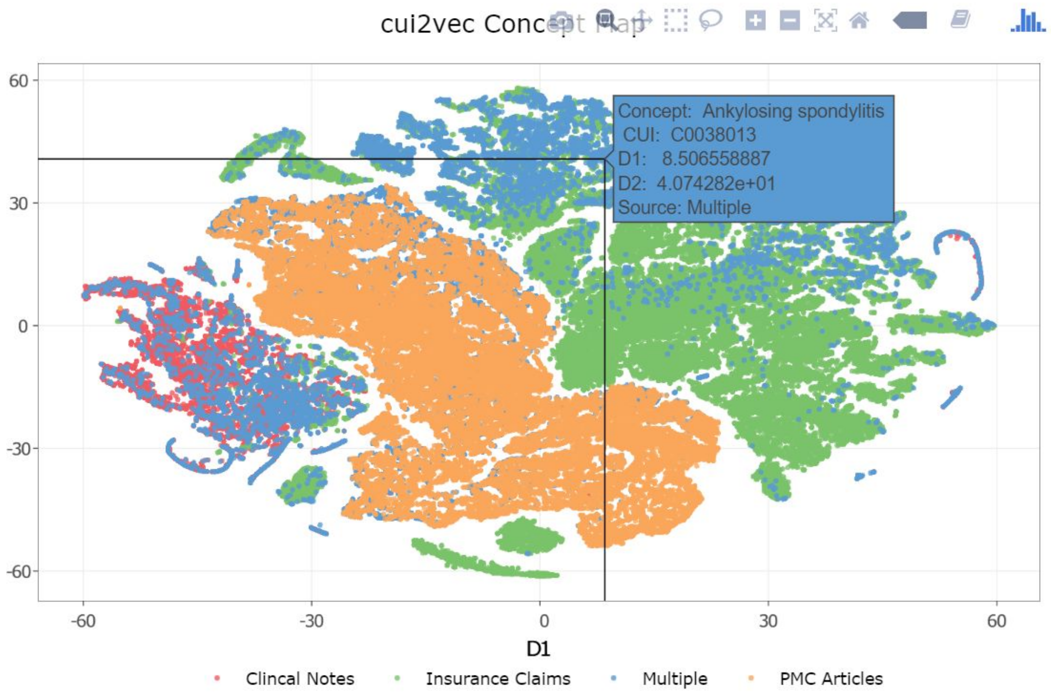
PROCESSING MEDICAL TEXT - SCISPACY

- Comparison requires numbers, but variable and value descriptions are free text...
- CUI - Concept Unique Identifiers
  - A common dictionary for medical terminology
  - Matches synonyms, alternative spellings to one “concept”
- scispaCy: Python package to automatically detect CUIs in free text

The scale of the [amount ENTITY] of [headache ENTITY] the [participant/subject ENTITY] is experiencing. /// [Headache severity scale ENTITY] /// Pain Severity (complete one of the following [scales ENTITY]) /// [Severity ENTITY] ///

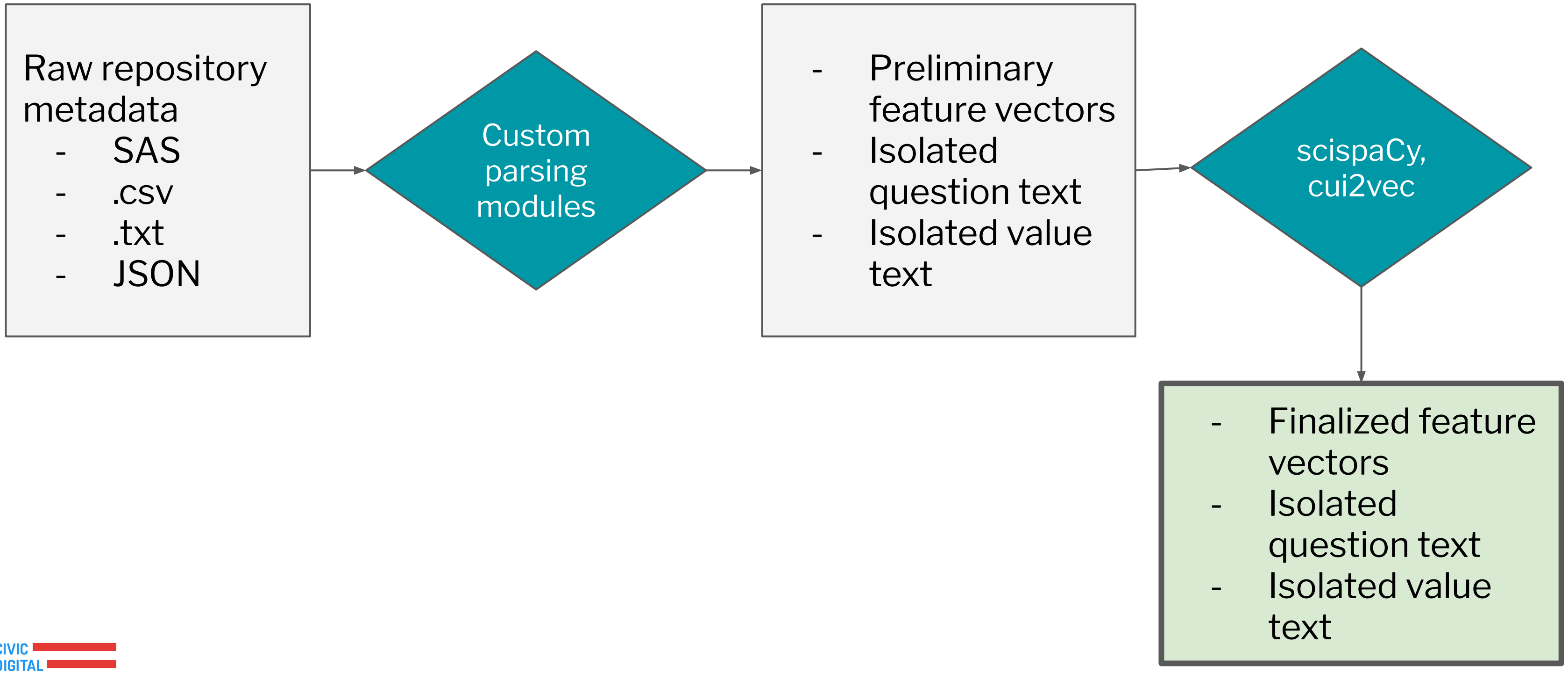
PROCESSING MEDICAL TEXT – CUI2VEC

- cui2vec: Vectors representing 100,000+ CUIs
  - Numerical representation of concepts that encodes meaning
  - Similar concepts have spatially similar vectors
  - Allows for mathematical comparison with cosine similarity
- Represent variable text as a matrix of CUI vectors



CIVIC DIGITAL FELLOWSHIP

DATA PREPROCESSING SUMMARY



COMPARING CDES AND DATASET VARIABLES

- Key insight (maybe): Comparing feature vectors is a **Record Linkage** problem
- Lack of labeled data → unsupervised record linkage task
- Fellegi-Sunter Model:
    - Expectation-Maximization (EM) Algorithm (applied to FS by Winkler)
    - Models probability of whether pair (a, b) is a match based on the comparison between their feature vectors
- Custom comparison for CUI matrices - use top-scoring CUI matches between a CDE and a dataset variable to give a score between 0-1

CIVIC DIGITAL FELLOWSHIP

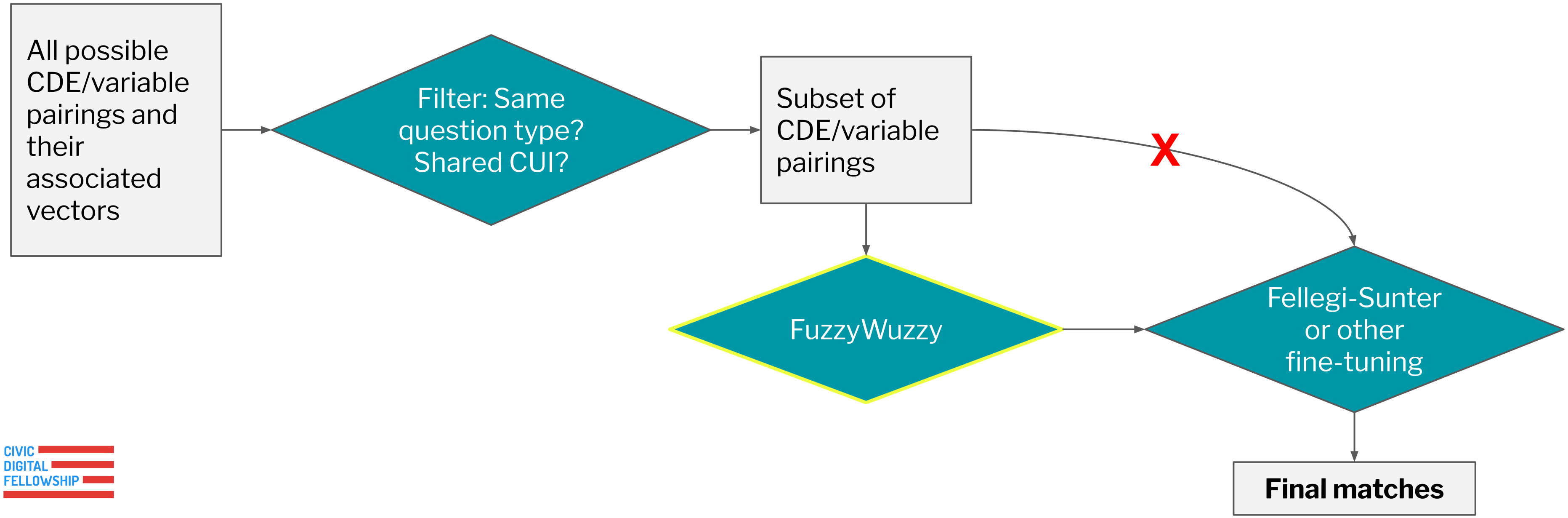
ALTERNATIVE COMPARISON METHODS

- Record Linkage with Fellegi-Sunter → Terrible results!
    - Needed to narrow down the number of potential matches
- FuzzyWuzzy: “fuzzy” string comparison - flexible comparison of text that doesn’t exactly match

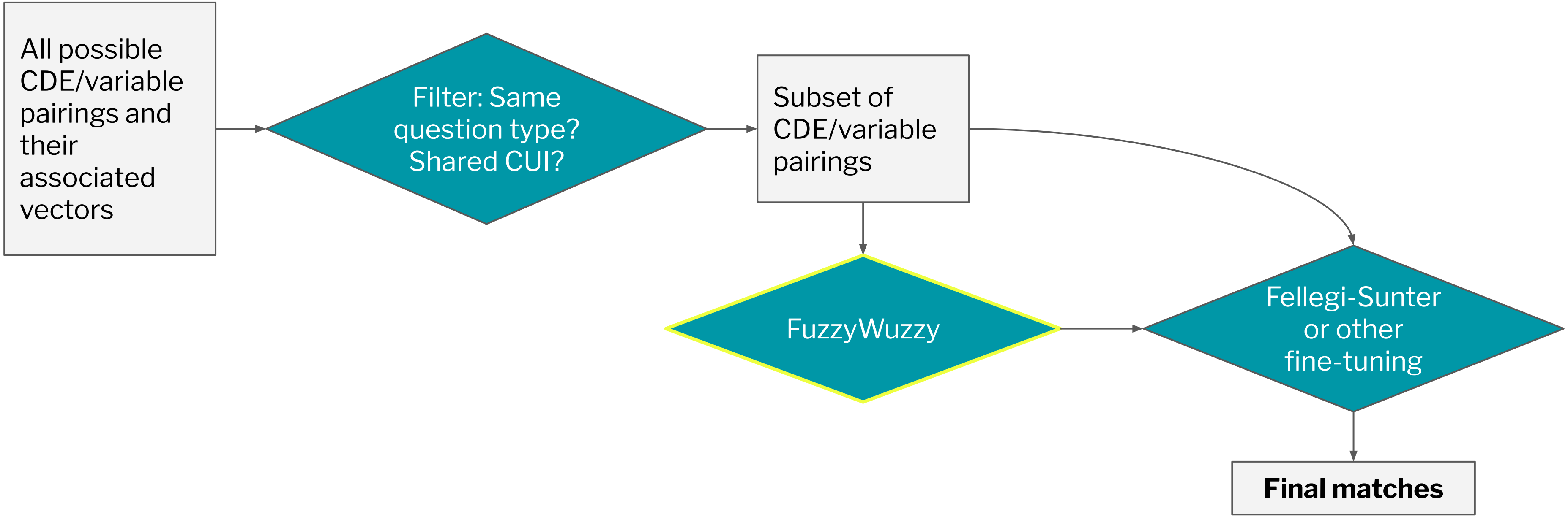
```
>>> choices = ["Atlanta Falcons", "New York Jets", "New York Giants", "Dallas Cowboys"]
>>> process.extract("new york jets", choices, limit=2)
    [('New York Jets', 100), ('New York Giants', 78)]
>>> process.extractOne("cowboys", choices)
    ("Dallas Cowboys", 90)
```

ALTERNATIVE COMPARISON METHODS

- Presort by question type and CUI keywords (as parsed with my custom package and scispaCy), then use FuzzyWuzzy



CIVIC DIGITAL FELLOWSHIP


All possible CDE/variable pairings and their associated vectors
Filter: Same question type? Shared CUI?
Subset of CDE/variable pairings
FuzzyWuzzy
Fellegi-Sunter or other fine-tuning
Final matches

ALTERNATIVE COMPARISON METHODS

- Mixed results, but far superior to Fellegi-Sunter method on the same presorted data
- Post-FuzzyWuzzy fine-tuning needs more investigation

```
FOIF: J11. Now, how many drinks on average... those occasions?
['During the past 30 days, on the days when you drank, about how many drinks did you drink on the average? NOTE: One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. A 40 ounce beer would count as 3 drinks, or a cocktail drink with 2 shots would count as 2 drinks.', 'Alcohol use last month drinking day average drinks numbe
r', 'During the past 30 days, on the days when you drank, about how many drinks did you drink on the average?']
```

```
FOIF: J11. Now, how many drinks on average... those occasions?
['On those day that you engage in moderate to strenuous exercise, how many minutes, on average, do you exercise [SAMHSA]', 'O
n those days that you engage in moderate to strenuous exercise, how many minutes, on average, do you exercise?']
```

A good match and a comically bad one

CIVIC DIGITAL FELLOWSHIP

MOVING FORWARD

- **In Progress:** Using Fellegi-Sunter record linkage technique to narrow down list of candidate pairs produced by FuzzyWuzzy (results currently mixed)
- **To Do:**
    - Explore other methods to refine FuzzyWuzzy candidate pairs
    - Explore new features to use in Fellegi-Sunter record linkage, or new record linkage approaches
    - Label subset of dataset variables with corresponding CDE match(es)
        - SME-labelled data allows for quantitative performance analysis
    - Expand parsing capabilities to other resources outside of the CDE and NIDDK Central Repositories

RECOMMENDATIONS - CDES

- Eliminate/combine duplicate or near-duplicate elements
- Standardize use of predefined fields, expand explanations of field use on repo website
    - Ex: units of measurement should go in the units field, not the definition
- Specify parameters
    - Ex: Minimum and maximum values for numeric measurements often left blank
- Standardize value encoding and listing of permissible values
    - Ex: combine all the different ways to encode “Yes” and “No”, use the same definition/labels for Yes/No across CDEs

RECOMMENDATIONS – NIDDK CR

- Require grantees to submit data in a standardized format and level of detail
    - Draft specific requirements or create template
- Uniform folder structure for controlled access data
- Consistent provision and storage of formats/metadata
- Offer public-facing data in machine-readable formats, aggregate metadata rather than spreading across documents
- Tag CDEs (building off of this project) for increased findability of studies

ACKNOWLEDGEMENTS

A special thank you to:

- Rachel Dodell, Chris Kuang and Ariana Soto @ Coding it Forward
- Jessica Mazerik and the NIH team
- All my mentors:
    - Ken Wilkins
    - Rebecca Rodriguez
    - Kerry Goetz
    - Robin Taylor
    - Jenna Norton
    - Melanie Laffin
    - Adrian Vancea
    - Luca Calzoni
    - Debbie Duran

CIVIC DIGITAL FELLOWSHIP