USING OPTICAL CHARACTER RECOGNITION TO SCRAPE BIOMEDICAL RESEARCH DATA

Ariel Langer, Coding it Forward Data Science Fellow

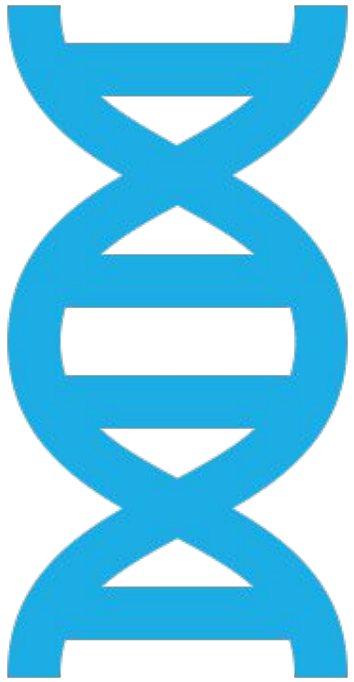
ABOUT ME

- Rising junior at University of California, Berkeley
- Majoring in Data Science with a concentration in Economics
- Passionate about using data science for social good

WHERE DID I WORK THIS SUMMER?

At the National Eye Institute (NEI) on a genomic medicine initiative called eyeGene® under the guidance of Kerry Goetz

WHAT IS EYEGENE®



- a genomic medicine initiative run by NEI
- partnered with clinics to screen individuals with inherited eye diseases for gene variants
- collected patient level genomic data
- facilitates research into genomic causes of rare eye diseases
- created patient registry for researchers to identify patients for future studies and vice versa

THE DATA SET AT A GLANCE

4,107 patients in the dataset

114,833 rows in the dataset

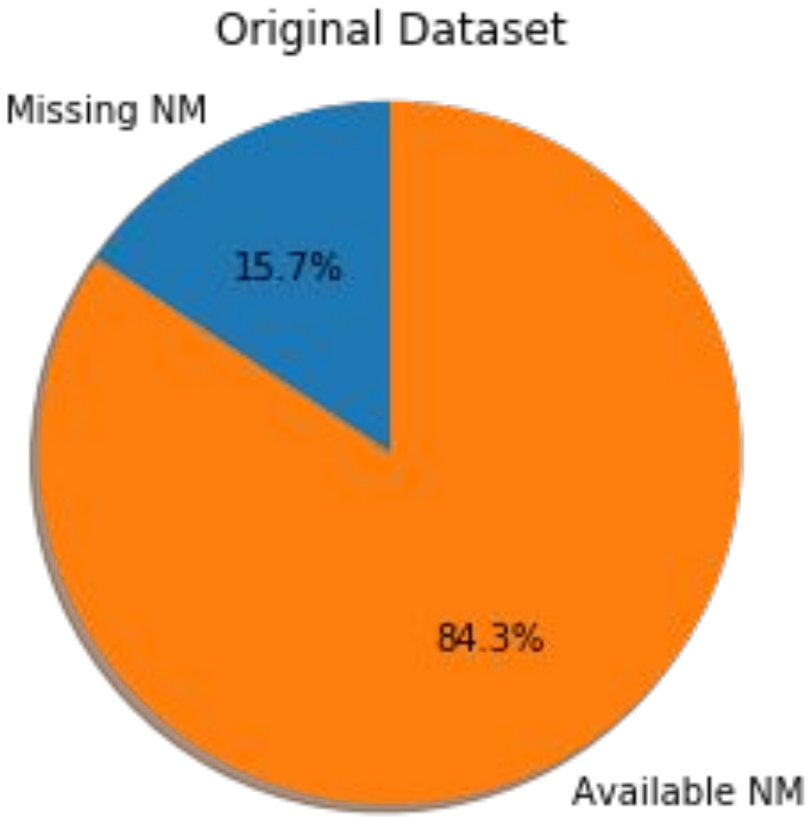
18,034 rows are missing an NM number

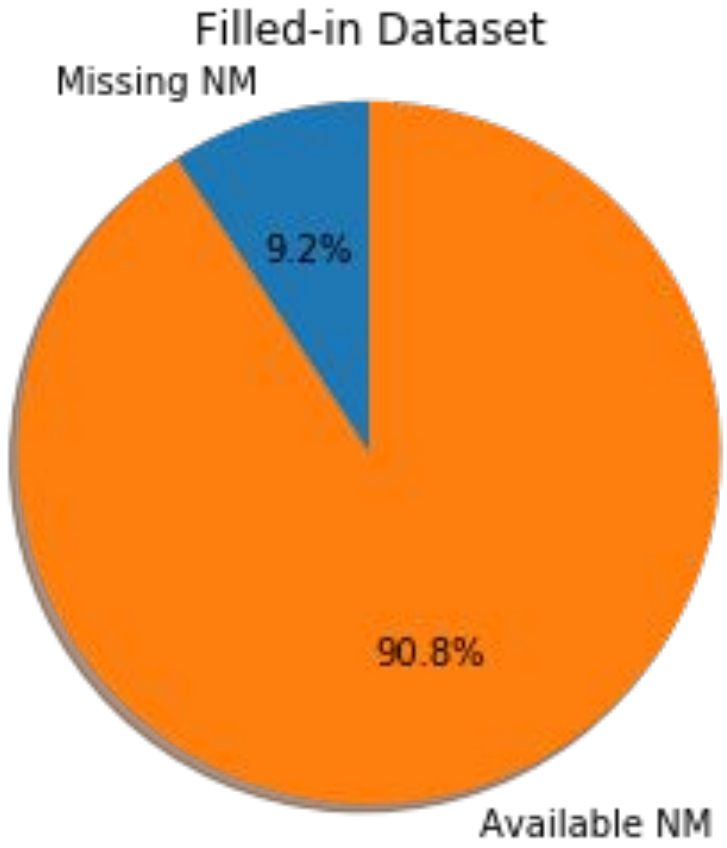
10,283 pdfs of patient results

MISSING DATA SOLUTIONS

- used optical character recognition (OCR) on the 10,283 pdfs
- turned pdfs machine readable text
- pattern matching to look for NM number & patient ID
- filled in missing data points
- checked accuracy of scraped results

RESULTS


Original Dataset
Missing NM
15.7%
84.3%
Available NM


Filled-in Dataset
Missing NM
9.2%
90.8%
Available NM

ENSURING CONTINUED PROJECT VIABILITY

Code lives in Jupyter notebooks

Explanations for each cell block

Thorough walk through of processes completed

Generalized so that code can be applied to future datasets as more genomics data is populated in BRICS

FUTURE STEPS

The dataset will be shared with other public sources of clinical genetic data such as Leiden Open Variation Database (LOVD) and ClinVar to further knowledge of the genetic causes of inherited eye conditions.

SPECIAL THANKS TO KERRY GOETZ, MELISSA REEVES, & THE REST OF NEI