# A Tool for Mapping Topics to Public GitHub Repositories

Nima Tayefeh, Neh Patel, Kyung Han, Chris Lee

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

# Background and Motivation

- Over 80% of the cloud based version control is owned by GitHub
- GitHub has a search tool that can look for keywords
- Want a way to map all repos/commits/files to these keywords
- Some current tools exist online but these techniques require text encodings and neural networks
- Focused mostly on covid for our own experiments but any keyword can be used
  - ['vaccin', 'covid', 'quarantin', 'corona', 'lockdown', 'social dist' ]

# Dataset

- Utilize the World of Code mappings

| Map | Description | Size |
|---|---|---|
| P2p | Mapping containing project and deforked project names | Finds 209,048,151 projects |
| P2c | Mapping project names to commit hash SHA | Finds 131,175,609 commits |
| P2b | Mapping commits to blob SHA | Finds 6,602,635,571 blobs |
| b2f | Mapping blob SHA to filenames | Not saved into a temporary file due to large file size |

# Experiment Setup

- Numerous different command line options to pick which functionality to run

| Option | Functionality |
| --- | --- |
| -c [outputFile.txt] | - Maps search queries to all commit SHAs that contain any of the queries in their commit messages.<br>- Writes to the optional output file or found_commits.txt |
| -p [outputFile.txt] | - Maps search queries to all project repository names that contain any of the queries in their name<br>- Writes to the optional output file or found_projects.txt |
| -r [outputFile.txt] | - Maps search queries to all blob SHAs contain any of the queries in the content of the readme<br>- Writes to the optional output file or found_readmes.txt |

# Commit Mapping

- Utilizes the Oscar library in Python to use the World of Code API within Python
- Uses the P2c mapping file to take each commit SHA and convert to a Commit object where the message attribute can be checked

*zcat allProjects.txt | grep "query1\|query2 \|query3"*

# Project Mapping

- Find project name substring
- Utilizes the grep tool
- Runs a system command using python

# Readme Mapping

- Uses project blobs
- Run in chunks of 5000
- Only check readme files
- Use showCnt to get the content of the files
    - decode64

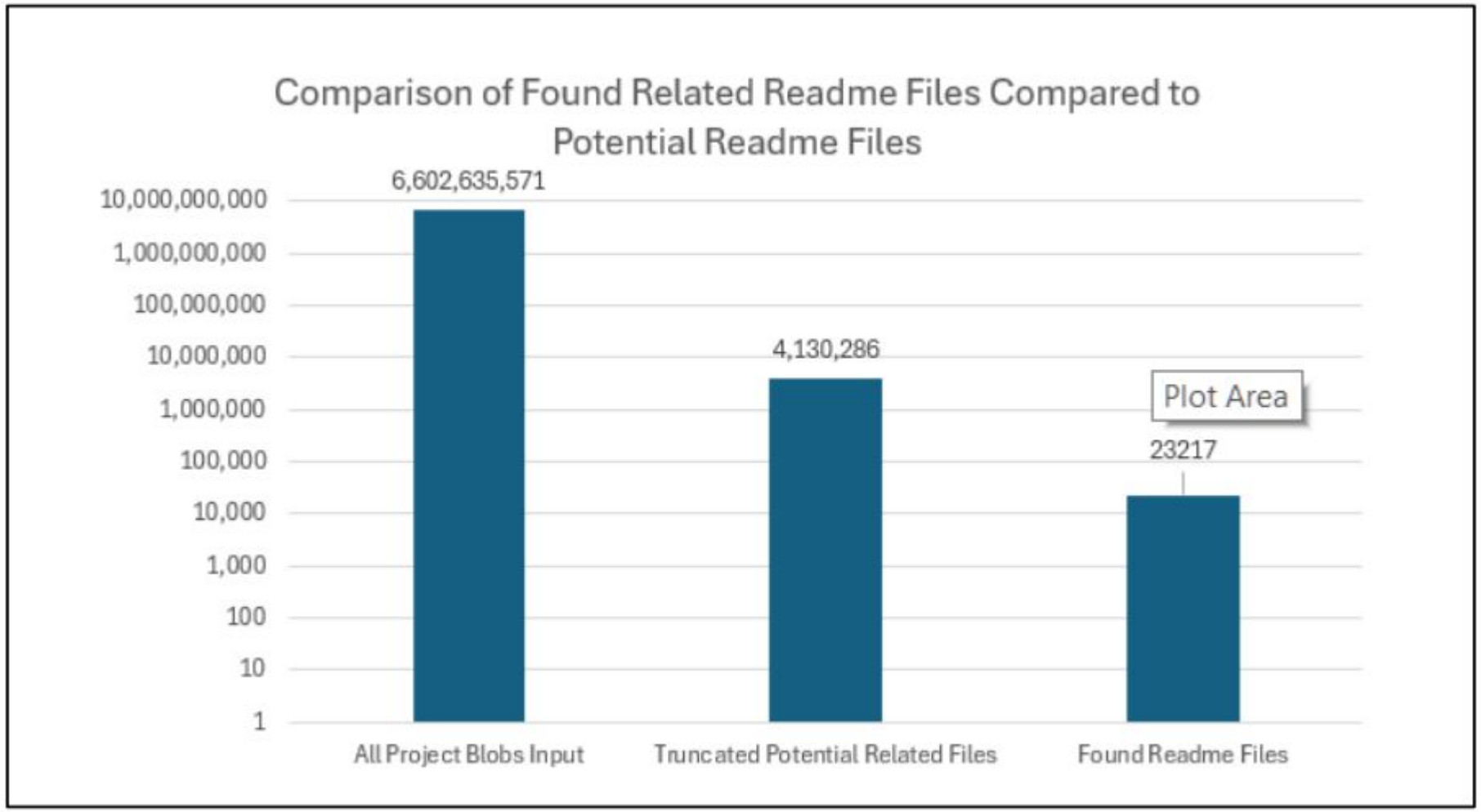
# Results

| All Project Blobs | Potential Related Files | Found Readme Files |
|---|---|---|
| 6,602,635,571 | 4,130,286 | 23217 |

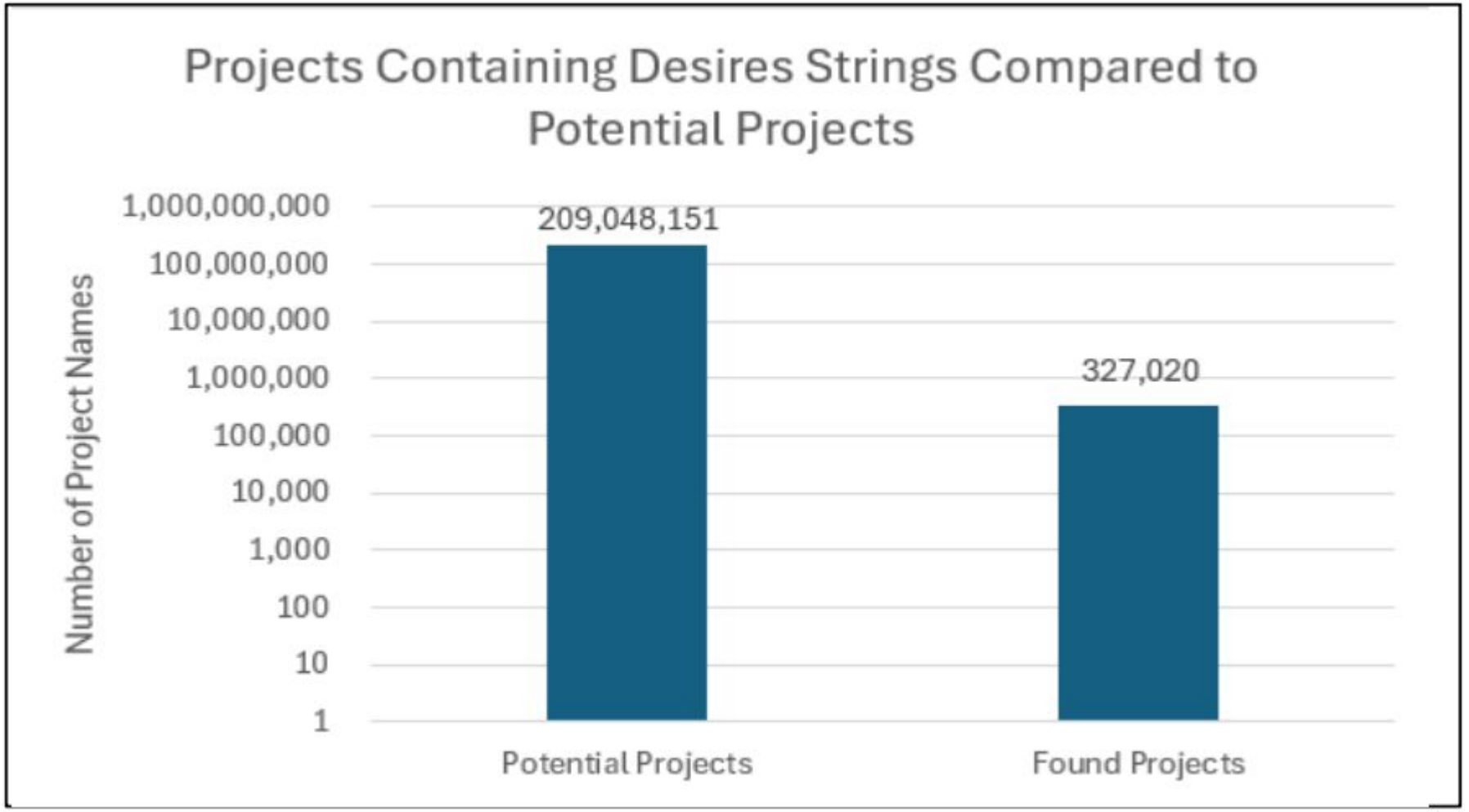
| Potential Projects | Found Projects |
|---|---|
| 209,048,151 | 327,020 |

| All Commits Input | Truncated Commits Input | Found Commits |
|---|---|---|
| 131,175,609 | 39,686,000 | 529,767 |

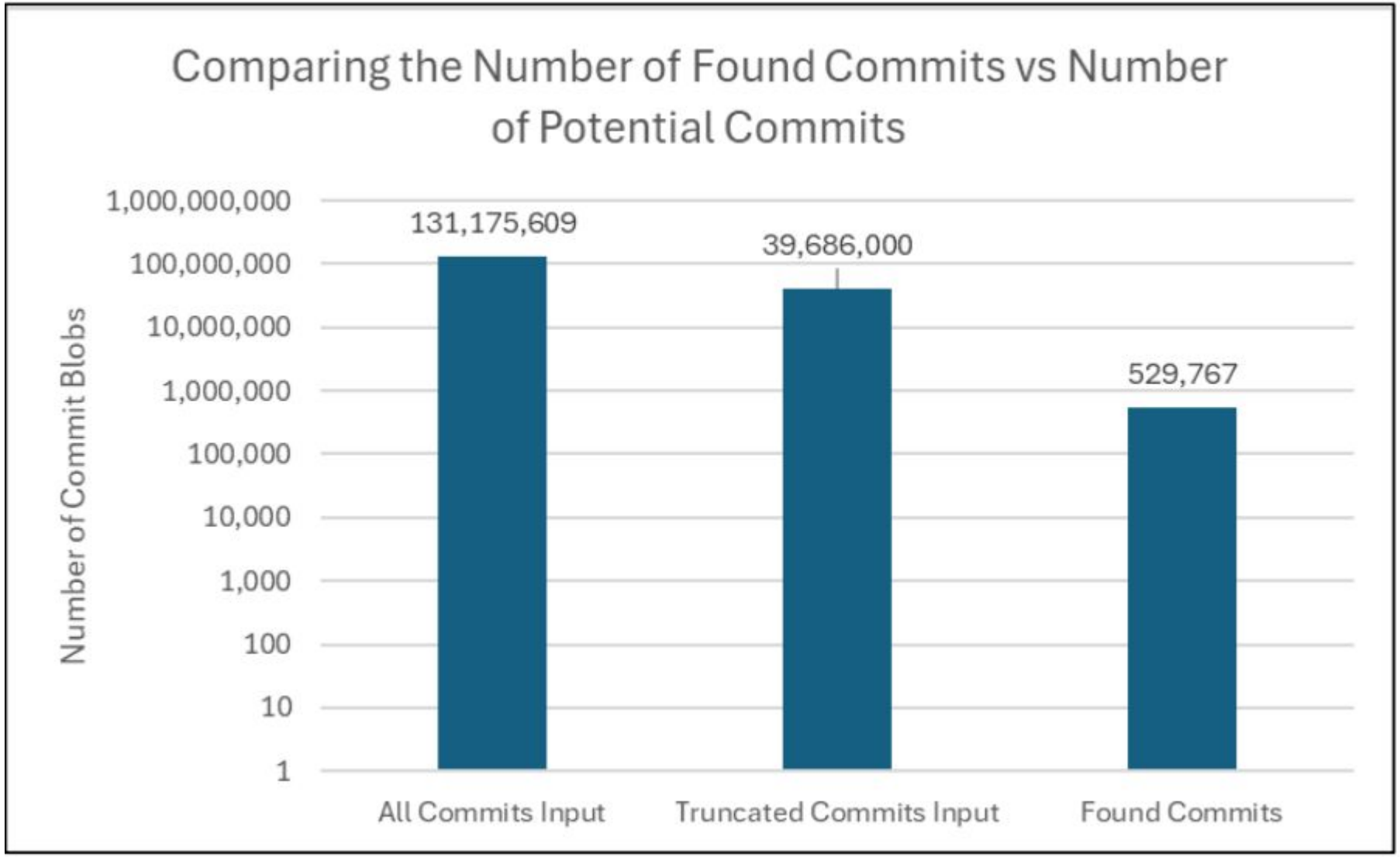
# Results & Figures



# Results & Figures



# Results & Figures



# Conclusions and Future Work

- Very effective
- Largest implication is time
  - large file sizes and linear search
- Multiple SSH
  - Running the getValues and showCnt does another SSH into a different server, which drastically can slow things down
- Required using VSCode's Live Share option for collaboration due to large runtimes and shared shell sessions
- Use the encoding and neural network approach to find similarity scores