Visualization of Understudied Genes

Amihan Camson, Dr. Alexander J. Stokes
Department of Molecular Biosciences and Bioengineering
University of Hawaii at Manoa
Hawaii Data Science Institute

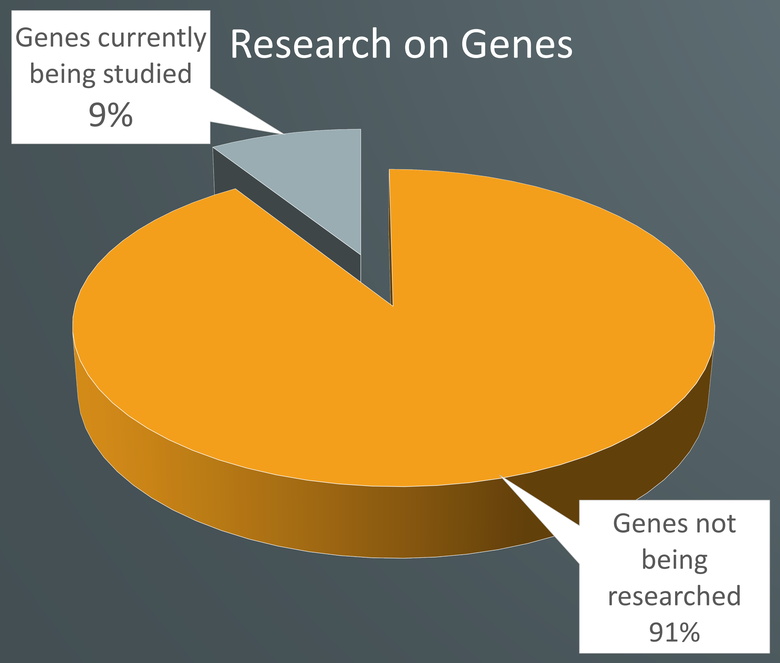
Introduction

U.S. National Library of Medicine Strategic Plan

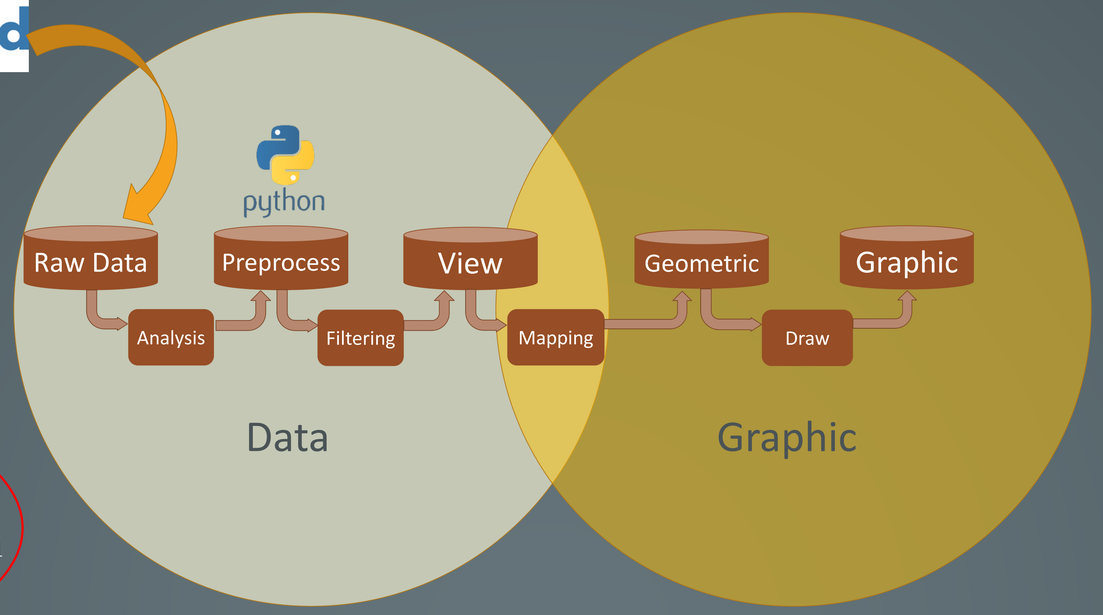
GOAL 1: Accelerate discovery and advance health by providing the tools for data-driven research

GOAL 2: Reach more people in more ways through enhanced dissemination and engagement pathways

GOAL 3: Build a workforce for data-driven research and health



Methods



Adapted from Lewis Chou

Query NCBI to extract list of gene IDs which are orthologs to the human gene

Use the list of IDs to retrieve synonyms differing across orthologs

Use synonyms to query PubMed and return PubMed IDs of articles about the genes

Use PubMed IDs of articles to return the date of publication

Graph the count of publications denoted by the ortholog along a timescale

Conclusion

This visualization is a catalogue of the publications of genes over time. This can serve to draw attention to gaps in knowledge and influence where further investigation is needed.

Further Research

- Automated data extraction from NCBI and cloud storage of data
- Text mining to identify sentences from PubMed abstracts
- Use natural language processing to transform unstructured text in documents into normalized, structured data suitable for analysis or to drive machine learning algorithms

Acknowledgements

I would like to thank my PI, Dr. Stokes and Sean Cleveland for guidance and support.

References

Stoeger, T., Gerlach, M., Morimoto, R. I., & Nunes Amaral, L. A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology*, *16*(9), e2006643. https://doi.org/10.1371/journal.pbio.2006643

Rodriguez-Esteban, R., & Jiang, X. (2017). Differential gene expression in disease: A comparison between high-throughput studies and the literature. *BMC Medical Genomics*, *10*(1), 1–10. https://doi.org/10.1186/s12920-017-0293-y