David Nicholson

dnicholson329@gmail.com • 412-607-6313 • 11291 Chatterly Loop Apt 104, Manassas VA 20109

Summary

Data Scientist with 15+ years of programming experience and 6+ years of experience in data analytics and visualization. My data analytics experience consisted of using document embeddings to gain further insight into biomedical research. Recently, my experience has grown to include working with large language models, clustering, and other analytics to help government clients better understand biomedical research.

Skills & Proficiency

Github/Gitlab • Python • R • SQL • Google BigQuery • Data Analysis • Machine Learning • Deep Learning • Natural Language Processing • Transformers • Large Language Models • Text Mining • Topic Modeling • Clustering • Knowledge Graphs • Document Embeddings • Data Visualization • ETL Pipelines • API Framework • Databases • Algorithms • Software Development • Parallel Processing • Google Cloud Platform • Dashboards • Continuous Integration (CI/CD) • Docker

Professional Experience

Data Scientist
Digital Science & Research Solutions, Ltd.
June 2022 - Present

- Constructed a software package to enable fast clustering of document sets, cluster hierarchy generation, and cluster metadata (labels from LLMs, various metrics such as Growth Rate, etc.).
- Maintained an ETL pipeline design to extract and update over 100GB of document data for the National Institute of Health (NIH) clientele.
- Designed and implemented a data analysis pipeline designed to cluster an 818k+ document set centered on women's health for a client affiliated with the National Institute of Health (NIH).
- Constructed an ETL pipeline designed to detect cancer drug treatments using a 64k+ biomedical document set for a pharmaceutical client.
- Back-tested vector databases were used to determine which database is most optimal for handling 10M+ document embedding vectors.
- Constructed a pipeline that used deep learning and dimensionality reduction models to uncover research topics and trends within a 20M+ document set for government clients.

Graduate Researcher Scientist
University of Pennsylvania
August 2016 - June 2022

- Designed and implemented parallel processing pipelines that achieved a 3x speed-up when analyzing terabytes of biomedical text.
- Used weak supervision for a 1.5x speed-up when training deep learning models (recurrent neural networks and transformers) to extract biomedical relationships from biomedical text.
- Applied a k-nearest-neighbor model to provide scientists with a web service that lists journals linguistically similar to a preprint of interest.
- Applied a time series analysis to discover over 20,000 different time points where words have changed their semantic meaning.

Publications

- **Unmasking The Language Of Science Through Textual Analyses On Biomedical Preprints And Published PapersOn Biomedical Preprints And Published Papers**
  Nicholson, D. N. (2022)
- **Changing Word Meanings in Biomedical Literature Reveal Pandemics and New Technologies**
  Nicholson, D. N. Alquaddoomi, F., Rubinetti, V., Greene, C. S. (2023)
- **Characterization of the Genome and Silk-gland Transcriptomes of Darwin's Bark Spider (Caerostris darwini)**
  Babb, P. L., Gregorič, M., Lahens, N. F.,**Nicholson, D. N.**, Hayashi, C. Y., Higgins, L., Kuntner, M., Agnarsson, I., Voight, B. F. (2022)
- **Examining Linguistic Shifts between Preprints and Publications**
  Nicholson, D. N., Rubinetti, V., Hu, D., Thielk, M., Hunter, L. E., Greene, C. S. (2022)
- **Expanding a Database-derived Biomedical Knowledge Graph via Multi- Relation Extraction from Biomedical Abstracts**
  Nicholson, D. N., Himmelstein, D. S., Greene, C. S. (2022)
- **Ten important roles for academic leaders to promote equity, diversity, and inclusion in data science.**
  Moore JH, Truong VQ, Robbins AB, **Nicholson D. N.**, Williams-Devane CL. (2021)
- **Constructing Knowledge Graphs and Their Biomedical Applications**
  Nicholson, D. N., Greene, C. S. (2020)
- **The Nephila Clavipes Genome Highlights the Diversity of Spider Silk Genes and their Complex Expression**
  Babb, P. L., Lahens, N. F., Correa-Garhwal, S. M., **Nicholson, D. N.**, Kim, E.J., Hogenesch, J.B., Kuntner, M., Higgins, L., Hayashi, C. Y., Agnarsson, I., Voight, B.F. (2017)

Education

Doctor of Philosophy (Ph.D.), Genomics and Computational Biology; University of Pennsylvania (Philadelphia, PA)

Postbaccalaureate Program (Penn Prep); University of Pennsylvania (Philadelphia, PA)

Bachelor of Science, Computer Science; University of Maryland Baltimore County (Baltimore, MD)