Jason Portenoy

jason.portenoy@gmail.com | jasport.org

Education

University of Washington, Information School — *Seattle, WA*

PhD, Information Science — 2021

- Dissertation: "Harnessing Scholarly Literature as Data to Curate, Explore, and Evaluate Scientific Research"
- Committee: Jevin D. West (Chair), Emma Spiro, Bill Howe, Benjamin Mako Hill

MS, Information Science — 2017

Brown University — *Providence, RI*

BS, Neuroscience — 2008

Work Experience

OurResearch — *Fully Remote*

Senior Data Engineer — OpenAlex — Feb 2023 -- Sept 2024

- OpenAlex is the world's largest free and open Scientific Knowledge Graph (SKG), created by the nonprofit OurResearch. Meant to represent the entire global research ecosystem, it comprises more than 250 million scholarly publications, billions of citation links between them, and terabytes of linked metadata. My primary role was to understand the strengths and limitations of the data, and to help our users do the same.
- As a key team member of this six-person startup, I had an active role in most operations: maintenance and development of a variety of databases (Postgres, Elasticsearch, Redshift, Databricks/Spark); backend development and data pipeline/quality improvement (Python, SQLAlchemy, Flask); writing and maintaining technical and general documentation; working with the press to provide and analyze data; engaging with the community through webinars, virtual and in-person presentations, social media, online forums, customer support tickets, and open houses; monitoring infrastructure and database operations; and tracking KPIs.

University of Washington — *Seattle, WA*

Data Engineer — Center for an Informed Public (CIP) — March 2022 -- Sept 2022

- Managed and maintained the data infrastructure supporting the research and operations of the CIP, an interdisciplinary center for research and action around misinformation and disinformation.
- Responsibilities included oversight and management of a collection of physical and virtual servers, and pipelines and tools for data collection, storage, and analysis, management of a Kubernetes cluster to run the Center's computing operations, and maintenance of multiple databases (Postgres, Airflow) with billions of rows and dozens of terabytes of data.

Research Analyst / Full Stack Developer — Sept 2021 -- March 2022

- With the Gordon and Betty Moore Foundation, developed grantexplorer.org — a tool to explore federal research grants over time — using React, D3, and FastAPI for interactive visualizations, and Elasticsearch and Gensim language models to intelligently assist with keyword queries.

Research Assistant / Teaching Assistant Sept 2013 -- June 2021

- Research projects with Center for an Informed Public, Military Suicide Research Consortium, National Academy of Sciences, JSTOR, Science History Institute, Pew Charitable Trusts
- TA experience: Advanced Data Science Methods, Calling BS in the Age of Big Data, Design Thinking, Information Systems Analysis and Management, Client Side Web Development, Capstone

Allen Institute for AI (AI2) *Seattle, WA*

Visiting Scientist / Full Stack Developer — Semantic Scholar Team March 2020 -- July 2022

- Built AI-enabled scholarly recommendation systems and wrote research publications.
- Using Pytorch and additional Python libraries, I fine-tuned a pretrained BERT language model to encode document similarity, which I used to create embeddings for 3.5 million documents.
- Built a custom React+D3 web app to provide recommendations and collect data for user studies.

Chan Zuckerberg Initiative *Redwood City, CA*

Visiting Researcher May 2019 -- Sept 2019

- Analyzed networks of biomedical researchers to improve suggestions of new papers in research feeds.

Microsoft Research and AI *Redmond, WA*

Summer Research Intern June 2017 -- Sept 2017

- Developed methods to identify when people make commitments in their written communications, and understand what kind of commitments they make.
- Used email data from Microsoft's internal terabyte-scale Cosmos data store to build machine learning classifiers.
- Mentors: Paul Bennett, Ryen White, Eric Horvitz

University of Washington eScience Institute *Seattle, WA*

Data Science for Social Good Summer Fellow June 2015 -- Aug 2015

- Partnership with the Bill and Melinda Gates Foundation and other organizations to help understand and address the problem of family homelessness in western Washington.

Skills

Python, Pandas, JavaScript, React, Typescript, R, SQL, D3, HTML, CSS, Linux, Git

Machine Learning and AI, Databases, APIs, Network Analysis, Visualization, High Performance Computing

Publications

Alperin, J.P., **Portenoy, J.,** Demes, K., Larivière, V., & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses. *arXiv preprint*. https://doi.org/10.48550/arXiv.2404.17663

Portenoy, J., Radensky, M., West, J., Horvitz, E., Weld, D., & Hope, T. (2022). Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery. *CHI 2022*. https://doi.org/10.1145/3491102.3501905

Portenoy, J., & West, J. D. (2020). Constructing and evaluating automated literature review systems. *Scientometrics*. https://doi.org/10.1007/s11192-020-03490-w

Portenoy, J., Hullman, J., & West, J. D. (2017). Leveraging Citation Networks to Visualize Scholarly Influence Over Time. *Frontiers in Research Metrics and Analytics*, 2, 8. https://doi.org/10.3389/frma.2017.00008.

Other publications while at UW Information School

Kim, L., **Portenoy, J. H.**, West, J. D., & Stovel, K. W. (2020). Scientific journals still matter in the era of academic search engines and preprint archives. *Journal of the Association for Information Science and Technology*, 71(10), 1218–1226. https://doi.org/10.1002/asi.24326

Hope, T., **Portenoy, J.**, Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., Hearst, M. A., & West, J. (2020). SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *EMNLP 2020*. http://arxiv.org/abs/2005.12668

Portenoy, J., & West, J. D. (2019). Supervised Learning for Automated Literature Review. *BIRNDL 2019 Workshop at SIGIR 2019*.

Portenoy, J., Kim, L., West, J., & Stovel, K. (2019, September). Do Journals Still Matter in an Era of Online Academic Search? *Metascience 2019*. https://doi.org/10.17605/OSF.IO/5PE73

Portenoy, J., & West, J. D. (2017). Visualizing Scholarly Publications and Citations to Enhance Author Profiles. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 1279–1282. https://doi.org/10.1145/3041021.3053058

Portenoy, J., & West, J. D. (2016). Dynamic Visualization of Citation Networks Showing the Influence of Scholarly Fields over Time. *Semantics, Analytics, Visualization. Enhancing Scholarly Data at WWW '16*, 147–151. https://doi.org/10.1007/978-3-319-53637-8_14

West, J. D., & **Portenoy, J.** (2016a). 10 The Data Gold Rush in Higher Education. In Big Data Is Not a Monolith (p. 129). MIT Press.

West, J. D., & **Portenoy, J.** (2016b). Delineating Fields Using Mathematical Jargon. *BIRNDL 2016 Joint Workshop on Bibliometric-Enhanced Information Retrieval and NLP for Digital Libraries*, 13, 14.

Kinsley, R. P., & **Portenoy, J.** (2015, January). Perspectives of Emerging Museum Professionals on the Role of Big Data in Museums. *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS)*. Hawai'i International Conference on System Sciences. https://doi.org/10.1109/HICSS.2015.249

Publications from a past life as a biomedical researcher

Chadha, M., **Portenoy, J.**, Boolbol, S. K., Gillego, A., & Harrison, L. B. (2012). Is There a Role for Postmastectomy Radiation Therapy in Ductal Carcinoma In Situ? *International Journal of Surgical Oncology*, 2012, e423520. https://doi.org/10.1155/2012/423520

Sheu, R., Lussier, D., Rosenblum, A., Fong, C., **Portenoy, J.**, Joseph, H., & Portenoy, R. K. (2008). Prevalence and Characteristics of Chronic Pain in Patients Admitted to an Outpatient Drug and Alcohol Treatment Program. *Pain Medicine*, 9(7), 911–917. https://doi.org/10.1111/j.1526-4637.2008.00420.x

Portenoy, J., & Teno, J. M. (2007). Hispanic Language Version of the Family Evaluation of Hospice Care. *Journal of Pain and Symptom Management*, 34(5), 459–461. https://doi.org/10.1016/j.jpainsymman.2007.08.003

Code Repositories

- infomap_large_network. Run Infomap community detection on very large networks. https://github.com/h1-the-swan/infomap_large_network
- Autoreview. https://github.com/h1-the-swan/autoreview
- Coauthorship force-directed visualization. https://github.com/h1-the-swan/nodelink_vis_coauthorship
- Nautilus visualization. https://github.com/h1-the-swan/nautilus-vis
- Cluster comparison visualization. https://github.com/h1-the-swan/nodelink_vis_cluster_compare
- Article timeline visualization. https://github.com/h1-the-swan/d3-article-timeline
- Article citations visualization. https://github.com/h1-the-swan/d3-article-citations
- Jargon distance. Calculate the jargon distance measure between documents and visualize the results. https://github.com/h1-the-swan/jargon_distance
- Pajek Tools. Convert network data to Pajek format. https://github.com/h1-the-swan/pajek_tools

Other materials

- https://www.grantexplorer.org/ -- GrantExplorer is a tool to explore federal research grants over time using interactive visualizations and NLP to intelligently assist with keyword queries.
- http://scholar.eigenfactor.org/ -- The home for the nautilus visualization showing scholar influence over time.
- http://scholar.eigenfactor.org/hicss -- Website showing work with the Hawaii International Conference on System Sciences to show their influence over time. Linked to on the official HICSS website (https://hicss.hawaii.edu/).
- http://www.misinformationresearch.org/ -- Website I created for the National Academy of Sciences to show applications of my research for the field of Misinformation, including an automated literature review and visualizations.
- https://scisight.apps.allenai.org/clusters -- SciSight visualization for groups of researchers working on COVID-19 research.
- White, R. W., Bennett, P. N., Horvitz, E. J., Ghotbi, N., **Portenoy, J. H.**, Hasegawa, M. M., Jha, A., & Modak, C. Y. (2019). Automated extraction and application of conditional tasks (United States Patent No. US20190129749A1). https://patents.google.com/patent/US20190129749A1/en

Mentoring

- Teaching data science skills to beginners: Community Data Science Workshop
- Teaching computing skills to researchers: Software Carpentry
- Teaching high-performance computing to undergrads: UW Research Computing Club
- Mentoring HCI projects to high-schoolers: Paul Allen Computing Challenge