

David Shumway

Software Engineer & Data Scientist

Email: davidshumway@gmail.com
 Github: [@davidshumway](https://github.com/davidshumway)
 LinkedIn: [@david-shumway-0b4661159](https://www.linkedin.com/in/david-shumway-0b4661159)
 StackEx: [@2504705/there](https://stackoverflow.com/users/2504705/there)

Phone: (773) 759-5970
 Address: 160 N. Elizabeth St.,
 Unit 1909, Chicago, IL 60607

Employment History

Graduate Research & Teaching Assistant, University of Illinois at Chicago - Chicago, IL 2018-Present

- Innovated with solutions in knowledge graphs, AI, GIS, ETL, data science, visual analytics, machine learning, and domain adaptation toward solving real-world problems.
- Taught and led discussions for 11 semesters, engaging with groups of 5-250 students in courses such as Languages & Automata, Software Development for Mobile, Data & Web Semantics, Big Data Mining, Software Engineering, and Software Design.
- Mentored students in 4 semesters with NSF-REU, fostering research skills and academic growth.

Co-Owner & Lead Developer, Finger Lakes Business Systems - Plattsburgh, NY 2010-2018

- Collaborated with 750+ clients through Amazon MTurk, completing 1000+ unique projects and 3+ million microtasks while maintaining a 99.9% task approval rating, and continued off-site with key clients.
- Engineered in-house web scraping, data management, workflow optimization, and distributed systems solutions, writing 375K+ lines of in-house code in JavaScript, PHP, HTML/CSS, SQL, and Python.
- Extensive use of the Google Chrome API and browser extension programming, including browser tabs, background scripts, event scripts, pop-up scripts, and storage.
- Created a simple open-source browser extension to help streamline the MTurk worker interface, published to the Chrome Store and Firefox Add-Ons, reaching up to 10K weekly users.

Summer Intern, École Polytechnique Fédérale de Lausanne, DSIL - Lausanne, Switzerland 2017

- Enhanced Kamusi.org's Neo4j natural language database by analyzing and integrating new datasets.

Digital Archivist & Technical Specialist, Paul Brunton Philosophic Foundation - Burdett, NY 2010-2013

- Curated 100K+ scanned documents and designed LAMP, JavaScript/HTML, Bash, and VBA solutions.
- Automated scanning process, resulting in 4x efficiency and saving two years of manual work.

Technical Freelancer, Driver, Retail Associate, Librarian 2000-2010

- Completed freelance work (Rent-A-Coder, Freelancer.com, MTurk) while working full-time in other roles.

Education

Ph.D. in Computer Science, University of Illinois at Chicago (expected Fall 2025) 2018-Present

- Advisors: Isabel Cruz (in memoriam), Cornelia Caragea.
- Completed PhD qualifying exam in *Advances in Scientific Workflow Management System Provenance Storage and Querying* (2020). 1.5 oral, 2.0 written (1-5, 1=best).
- Completed PhD coursework requirement (2021).
- GRE (entrance): 159 verbal, 149 quantitative, 4.0 analytical.
- 41 GPA credit hours, 3.51/4.0 GPA. 54 PhD Thesis Research credit hours.

B.S. in Computer Science, State University of New York at Plattsburgh 2015-2017

- Dean's List honors in all five semesters.
- 17th Annual Computer Science Department Academic Excellence Award recipient (2016).
- 97 GPA credit hours, 3.86/4.0 GPA.

No degree, Utah Valley University, Germanna Community College, Northeastern University 1998, 2001-2002

- Architecture, Arts and Sciences, & Engineering Focuses.
- 80 GPA credit hours.

Technical Skills

Languages: Python, JavaScript, PHP, HTML, CSS, Java, R, C, C++, C#, Assembly, VBA, VB.NET

Databases & Knowledge Graphs: MySQL, MariaDB, MySQL Workbench, PostgreSQL, SPARQL, RDF, OWL, Protégé, WebProtégé, PhpMyAdmin, SQLite, Virtuoso, Neo4j, Cypher, pgAdmin, rdflib

Other: Android, Bash/Shell, Linux, Git, Pandas, Numpy, GeoPandas, Jupyter, D3.js, Google Colab, matplotlib, scikit-learn, PyTorch, TensorFlow, Docker, Mass Open Cloud, REST, AJAX, cURL, Apache web server, Nginx, MySQL Workbench, Node.js, browser extensions, Chrome, Django, Flask, Apache Cordova, Scrapy, LaTeX, PySpark, Dask

HW/SW: Built and maintained PCs and laptops including Linux/Windows installation and maintenance.

Research Focus

Construction and use of knowledge graphs and ontologies, exploration of knowledge graph and ontology learning, and knowledge graph embeddings. Applying machine learning techniques in two key directions: 1) detecting bacteria species and small molecules in MALDI-TOF spectra, and 2) predicting concentrations of fecal indicator bacteria in coastal waters with transfer learning and domain adaptation.

Research Experience

Project Title: AI methods for reducing diver exposure to biological hazards

Research Advisors: Dr. Isabel Cruz, Dr. Cornelia Caragea

Duration: August 2021 - Present

Grant: ONR, #00014-21-1-2286

Project Title: A new paradigm for the creation and mining of microbial libraries for drug discovery

Research Advisor: Dr. Isabel Cruz

Duration: August 2018 - August 2021

Grant: NIH, #R01GM125943

Research Interests

Data science, web and data semantics, knowledge graphs, ontologies, embeddings, big data, machine learning, domain adaptation, transfer learning, databases, data mining, information retrieval, web search, information visualization, computational biology, open-source, agile development.

Teaching History

University of Illinois at Chicago - Chicago, IL

2018-Present

Instructor

Data and Web Semantics, CS586 (2021)

Teaching Assistant

Digital Literacy, CS100 (2019)

Languages and Automata, CS301 (2019, 2023, 2024)

Software Design, CS342 (2018)

Software Engineering I, CS440 (2019)

Software Development for Mobile Platforms, CS478 (2020, 2021)

Big Data Mining, CS494 (2020)

Data and Web Semantics, CS586 (2019, 2020, 2021)

Graduate Coursework – UIC (relevant)

Computer Algorithms, CS401

Introduction to Machine Learning, CS412

Visualization and Visual Analytics I, CS424

Systems Performance and Concurrent Computing, CS463

Numerical Analysis, MCS471

Software Development for Mobile Platforms, CS478

Computer Algorithms II, CS501

Data and Web Semantics, CS586

Research Methods, CS590

Algorithms for Big Data, CS594

Visual Data Science, CS594

Undergraduate Coursework – SUNY Plattsburgh (relevant)

Artificial Intelligence, CSC345

Operating Systems, CSC433

Data Mining, CSC442

Software Design Studio, CSC446

Climate Change Science, ENV406

Sustainable Transportation, GEG380

Urban Geography and Planning, GEG307

Linear Algebra I, MAT202

Manuscripts & Publications

† co-author

* primary author

‡ supporting role

Published

† Gautam, N., Shumway, D., Kowalczyk, M., Khanal, S., Caragea, D., Caragea, C., ... & Dorevitch, S. (2023, April). Leveraging Existing Literature on the Web and Deep Neural Models to Build a Knowledge Graph Focused on Water Quality and Health Risks. In *Proceedings of the ACM Web Conference 2023* (pp. 4161-4171).

Submitted / Revised

† Elahi, A., Shumway, D., Gautam, N., Kowalczyk, M., Caragea, D., Caragea, C., & Dorevitch, S. (2024). "Predicting Surface Water Bacteria Levels Using Transfer Learning and Domain Adaptation." Submitted to *IJCAI (2023), DSAA (2023), The Web Conference (2024)*.

* Shumway, D., Elahi, A., Gautam, N., Kowalczyk, M., Caragea, D., Caragea, C., & Dorevitch, S. (2023). "A Domain Adaptation Approach for Predicting Recreational Water Quality at Data-Scarce Coastal Locations." Submitted and revised for *Environmental Science & Technology*.

No Submission

* Shumway, D., Elahi, A., Gautam, N., Kowalczyk, M., Caragea, D., Caragea, C., & Dorevitch, S. (2023). "Developing a recreational and occupational waterborne illness knowledge graph toward biological risk assessment in coastal waters."

* Shumway, D. (2023). "Use of large-scale public datasets toward improving coastal water quality predictions."

* Aggarwal, M., Shumway, D., Cruz, I. (2020). "Bacteria Identification in MALDI-TOF Mass Spectrometry Data Using a Convolutional Neural Network Approach."

* Shumway, D. (2020). "Storage and Querying of Provenance in Scientific Workflow Management Systems." Completed as part of PhD qualifying exam (UIC).

† Zhao, W., Shumway, D. (2020) "A Review of QA4IE." Completed as part of Research Methods graduate course (UIC, CS590).

* Shumway, D. (2020). "A public community-driven database for sharing and execution of matrix-assisted laser desorption ionization time-of-flight mass spectrometry bacterial and protein spectral data and workflows." Completed as part of Research Methods graduate course (UIC, CS590).

* Shumway, D. (2019). "Bacterial identification in a distributed mass spectrometry sensor network using streaming big data." Completed as part of Algorithms for Big Data graduate course (UIC, CS594).

† Zhao, K., Shumway, D., Xingbo, W., Cruz, I. (2019). "Efficient Multi-user Semantic Desktop Framework that Supports Data-aware Storage."

* Shumway, D. (2018). "ADEPT - An ontology for Atomic Distributed Experiments: Modeling Scientific Workflow Processes as Always-Available Cloud Services to aid in Workflow Reproducibility, Composition, Sharing, Resource Conservation, and Testability." Completed as part of Data and Web Semantics graduate course (UIC, CS586).

Grant Writing

‡ Dorevitch, S., Cruz, I., Catlett, C. (2021). "Microbial hazard risk estimation and communication for Navy divers."

‡ Cruz, I., Miranda, F., Derrible, S., Siciliano, M., Cailas, M., Sambanis, A., ... & McHenry, K. (2021). "ICE-PRESUR: Illinois Center of Excellence-Planning a Resilient and Equitable State Using Real-time data."

† Cruz, I., Shumway, D. (2021). "EarthCube Capabilities: Semantic Data Connections to Enable Distributed Science and Capability Building."

† Zhao, K., Shumway, D., Xingbo, W., Cruz, I. (2019). "Efficient Knowledge Graph Processing in Modern Memory and Storage Hierarchy."

* Shumway, D. (2019). "A collaboration and visualization tool to help meet sustainability goals by individuals and groups within local and global contexts." Completed as part of Grant Writing (UIC, UPP493).

Presentations

A public community-driven database for sharing and execution of matrix-assisted laser desorption ionization time-of-flight mass spectrometry protein fingerprints and workflows. Spring 2020. Presentation for PhD Written Critique and Proposal (WCP) PhD Qualifying Exam, UIC. Spring 2020.

A public community-driven database for sharing and execution of matrix-assisted laser desorption ionization time-of-flight mass spectrometry protein fingerprints and workflows. Spring 2020. Final presentation for Research Methods, UIC.

Bacterial identification in a distributed mass spectrometry sensor network using streaming big data. Spring 2019. Final presentation for Algorithms for Big Data, UIC.

ADEPT Ontology: Atomic Scientific Workflow Processes as Always-Available Cloud Services. Fall 2018. Presentation for Data and Web Semantics, UIC.

Natural Drug Discovery Web-Based Data Visualization: Project Proposal. Fall 2018. Presentation for Visual Data Science, UIC.

The Salvation of Paris: Notes on Sustainable Transit in Paris from Taras Grescoe's Straphanger. December 2017. Final presentation for Sustainable Transportation, SUNY Plattsburgh.

Containers, Docker, Linux: Notes from the 2016 LinuxCon Conference. August 2016. Presented to the Software Engineering Club at SUNY Plattsburgh.

Health and Well-Being in the Built Environment: Notes on Christopher Alexander. May 2016. Final presentation for Public Speaking, SUNY Plattsburgh.

Open Source Contributions

- Publicly shared numerous small projects and code snippets on GitHub and GitHub Gists demonstrating a commitment to community and collaboration in software development.
- Contributed to open-source issue tracking through new issue documentation and creation of pull requests.

Conference Attendance

LinuxCon, Toronto. (2016).

UIC Urban Forum. (2019).

EarthCube Annual Meeting, Virtual attendee. (2021, 2022).

International Conference of Biomedical Ontologies, Virtual attendee. (2022).

Notebook Reviewer

2nd and 3rd Annual EarthCube Call for Notebooks. (2021, 2022).

Portfolio & Example Projects

[AI methods for reducing diver exposure to biological hazards](#). Lead Developer, 2021-Present.

Constructed water quality datasets by combining multiple environmental data, developed and tested machine learning, transfer learning, and domain adaptation models, including supervised, semi-supervised, and unsupervised methods, and helped to build a waterborne illness ontology.

Tech: Python, LaTeX, SPARQL, RDF, OWL, Protégé, WebProtégé, Virtuoso, Dask, rdflib, OBO ROBOT, Overleaf

Visualization & Data Analytics: Pandas, matplotlib, GeoPandas, D3.js, Plotly, Google Colab + Google Drive, Google Sheets

Machine Learning & AI: time-series prediction, domain adaptation, transfer learning, supervised & unsupervised learning, PyTorch, TensorFlow, scikit-learn, Linear Regression, Random Forest, Decision Tree, SVM, MLP, CNN, XGBoost, LSTM, ADAPT, EasyAdapt, subspace alignment, CORAL, GRU, GNN, BERT, LLM

Popular APIs & Datasets: UMLS, Infectious Disease Ontology, WaterQualityData.us, EPA BEACON, OpenStreetMap API, NOAA Global Hourly, NOAA Tides & Currents, NOAA NDBC, EPA ECHO, EN4 Met Office Hadley Centre Ocean Salinity

[FDIC Quarterly Banking App](#). Independent project, 2023.

Single-page app using Flask, JavaScript, and DataTables.js to analyze FDIC Banks API reporting data from 5K+ US financial institutions, with mechanisms for automated retrieval, caching, and CSV download of selected views.

Tech: Flask, DataTables.js, FDIC Banks API

[MALDI-DB: A public repository for MALDI-TOF mass spectra](#). Lead Developer, 2018-2021.

Built a community-driven public repository for preprocessing, analysis, search, and hosting of MALDI-TOF mass spectra and their small molecules, utilizing RPLumber, PostgreSQL, Nginx, Django, D3.js, and Docker.

Files: <https://github.com/idbac/malidb>

Development screenshots: <https://github.com/idbac/malidb/tree/master/mdb/media>

[Zoomie: Round-robin breakout scheduling in Zoom](#). Independent project, 2020.

Simple browser extension built for a meditation group which utilizes Zoom's web-chat client, enabling moderators to run round-robin partnered breakout room events (Firefox, Chrome).

[Chicago Public Schools Enrollment App](#). Visual Design Studio (CS526), 2018.

Single-page app built the UIC Department of Minority Affairs toward recruitment of high school students from within the City of Chicago. Developed 3 out of 5 visualizations, the search functionality, and tied the app together into a single page with linked views.

Tech: HTML, CSS, JavaScript, D3.js, Leaflet, Chicago Public Schools dataset

Files: <https://github.com/davidshumway/CPS>

Portfolio & Example Projects (Continued)

Finger Lakes Business Systems. Lead Developer, 2010-2018.

Tech: JavaScript, Linux, MySQL, PHP, HTML, CSS, Python, Google Chrome API, Chrome/Firefox browser extensions, Amazon Mechanical Turk (AMT), Chart.js

Files (available on request):

- AMT marketplace-related browser extensions
- AMT marketplace-related MySQL databases
- AMT marketplace-related PHP scripting
- AMT marketplace-related Apache Cordova mobile app
- AMT task-related browser extensions
- AMT task-related MySQL databases
- AMT task-related PHP scripting
- AMT task-related Python scripting (Scrapy)
- Browser extensions, SQL, and PHP scripting related to clients outside of AMT

Popular APIs: Google Places (nearby search, text search, autocomplete, radar search), Google Maps, Yelp Businesses (v1-v3), Facebook Graph, DBpedia, Foursquare, BBB, Wikipedia, GeoNames, Twitter, Airbnb, Upcdatabase.org, Smoots.com

Tools for Amazon's Mechanical Turk. Independent project (open-source), 2013-2018.

Developed Firefox and Chrome browser extensions to improve worker productivity on the (AMT) website.

Tech: Firefox / Chrome browser extension, HTML, CSS, JavaScript

Trumpocalypse. Team project, Software Design Studio (CSC446), Fall 2017.

Politically-themed satirical text-based game in the style of Oregon Trail, built using the Python Pygame library. Completed roughly 350 out of 500 team commits.

École Polytechnique Fédérale de Lausanne. Summer Intern, Summer 2017.

Tech: Neo4j, Cypher, Ubuntu, node.js, PostgreSQL

Files (available upon request):

- Cypher queries for working with Kamusi's Neo4j graph database
- PostgreSQL queries for analysis of the PanLex language database

Paul Brunton Philosophic Foundation. Technology Specialist & Developer, 2010-2013.

Automated the (EPSON) scanning process, resulting in 4x efficiency and saving two years of manual work.

Created MySQL database full-text search form within Microsoft Word using VB.NET and Visual Studio.

Utilized tesseract open-source OCR library to extract text from typewriter-written documents, saving the extracted text to a MySQL database, and implemented tools for 1) LibreOffice auto-complete using macros, and 2) duplicate hard-copy document checking. Utilized ImageMagick to automatically crop scanned white space from thousands of scanned typewritten documents, and created and used an interactive tool to validate cropped white space using thumbnails (HTML/JavaScript/CSS), providing performance gains of up to 5x over manual image cropping.

Tech: Linux, Apache, MySQL, PHP, HTML, CSS, JavaScript, VBA, VB.NET, Tesseract, Bash, Visual Studio, LibreOffice, Microsoft Word

Files (available on request):

- Epson scanning automation tool: Initially VBA macros, and later combined with VB.NET front-end
- Microsoft Word, MySQL database full-text search form (VB.NET)
- Tesseract OCR via shell scripting, MySQL database
- ImageMagick cropping and crop-review tool (LAMP, HTML, JavaScript, CSS)

Freelance Technology Specialist. Technology Specialist & Developer. 2005-2010.

Provided freelance tech services in person and on freelance sites such as Rent-A-Coder, Freelancer.com, and MTurk.com.

Tech: Linux, Apache, MySQL, PHP, HTML, CSS, JavaScript, Audio transcription

Online: <https://www.freelancer.com/u/telosMed> (sample work history)