Ana Trisovic

anatrisovic.com | linkedin.com/in/ana-trisovic | github.com/atrisovic | twitter.com/atrisovic
anatrisovic@g.harvard.edu | (617)-230-1653

Summary

My primary research interests are computational reproducibility, software sustainability, data provenance, research preservation and reuse. I am enthusiastic about data science, machine learning and open science.

Areas of Expertise & Skills

- **Data Engineering**
  Data Wrangling, Data visualization, High-dimensional datasets, Big Data, Large-scale data analysis, Working with missing data, Workflows & Automation, Data integration, Time Series (longitudinal) data analysis and filtering, Data Management, Metadata, Research dissemination, Computational reproducibility
- **Machine Learning and Artificial Intelligence (ML/AI)**
  Statistical modeling and analysis, Clustering, Principal component analysis (PCA), Classification, Linear regression, Non-linear regression, Logistic regression, Ridge regression, Outlier detection, T-SNE, Support Vector Machine, Feature Extraction, Feature Engineering
- **Computational & Programming**
  Cloud (AWS, OpenStack Cloud), Graph databases (Neo4j/CYPHER), Relational databases (SQL), Python (advanced), Python ecosystem (numpy, pandas, scipy, seaborn), R (basic), C++ (basic), Git/GitHub, DevOps (CD/CI, Docker), Linux scripting (bash), Slurm, HTCondor
- **Web Development & Design**
  HTML/CSS, Javascript, Flask, Adobe Photoshop
- **Other**
  LaTeX, Vim, documentation & guides (readthedocs, Jupyter Book)
- **Management & Organization**
  Conference organization committee, Chairing working groups, Develop large-scale scientific research programs, Secure research grants (NIH, Alfred P. Sloan Foundation, AWS), Presenting research findings, Keynotes, Project management, Supervision

Notable projects

- **Health and climate data integration for biostatistics research**
  Large-scale integration and aggregation using high-performance computing (HPC) of diverse datasets including health data from the Centers for Medicare and Medicaid Services, demographic data from the US Census, climate and weather, and environmental exposure data (air pollution and wildfires).
- **Concurrence of extreme exposure**
  Investigating the trends, frequency, the associations and interrelationships among in concurrence of extreme exposure events such as wildfires, extreme heat days, and air pollution, which are projected to increase in frequency, intensity, and duration due to global warming.
- **Spacial confounding and health benchmarking data**
  Developing a set of semi-synthetic benchmark datasets for causal analysis under multiple climate and environmental scenarios, with a goal to enable evaluating, comparing and verifying the performance of causal inference methods and different learning approaches and techniques.
- **Feature development at Dataverse research repository**
  Developing new features in the Dataverse software to accommodate research software through citation and dependency capture, and geospatial, environmental and climate data (NetCDF and HDF5 formats).

- **Reproducibility of computational research and standardization**
  Conducting large-scale studies on AWS to identify typical challenges of computational research and develop new standards and recommendations for research software, workflow automation and data dissemination (FAIR principles).

Education

University of Cambridge, Newnham College — Cambridge, United Kingdom
PhD in Computer Science — *2014 – 2018*

Union University, School of Computing — Belgrade, Serbia
BSc in Computer Science — *2010 – 2014*

University of Belgrade, Faculty of Mechanical Engineering — Belgrade, Serbia
BSc in Mechanical Engineering — *2010 – 2013*

Experience

Harvard University — Cambridge, USA
Research Associate — *Feb 2022 – Present*

- Undertake research and software development toward advancing the quality, reproducibility, and reuse of statistical analysis on air pollution and health.

Harvard University — Cambridge, USA
Sloan Postdoctoral Fellow — *Sept 2019 – Feb 2022*

- Conduct computational experiments and software development to support better documentation, sustainability, and reuse of research outputs disseminated through data repositories.

The University of Chicago — Chicago, USA
CLIR Postdoctoral Fellow — *Sept 2018 – Sept 2019*

- Collaborate with the researchers at the Energy Policy Institute at the University of Chicago (EPIC) to facilitate data analysis, reproducibility, and openness in energy, environmental, and climate research.

CERN — Meyrin, Switzerland
Project Associate — *Sept 2017 – Sept 2018*

- Conduct software development for the CERN Analysis Preservation and CERN Open Data (opendata.cern.ch) platforms for sharing particle-physics analyses and experimental data.

CERN — Meyrin, Switzerland
Technical Student — *July 2013 – July 2014*

- Develop an LHCb particle collision display, a stand-alone application for education and outreach that was later used by hundreds of high-school students through the event International Masterclass in Physics.

Microsoft Development Center — Belgrade, Serbia
Data Science Associate — *March 2013 – July 2013*

- Analyze Azure cloud data to evaluate the efficiency of the load balancer and propose improvements.

Selected Publications

1. Mauricio Tec, Ana Trisovic, Michelle Audirac, and Francesca Dominici. `SpaCE`: The **Sp**atial **C**onfounding (Benchmarking) **E**nvironment. *Submitted to CLeaR (Causal Learning and Reasoning)*, 2023
2. Ana Trisovic. Cluster Analysis of Open Research Data: A Case for Replication Metadata. *International Journal of Digital Curation*, 2023
3. Daina Bouquin, Oliver Bertuch, Elena Colon-Marrero, and Ana Trisovic. Advancing Software Citation Implementation. *Computing Research Repository (CoRR) arXiv*, 2023
4. Lee Whanhee, Xiao Wu, Seulkee Heo, Joyce Mary Kim, Kelvin C. Fong, Ji-Young Son, Matthew Benjamin Sabath, Ana Trisovic, Danielle Braun, Jae Yoon Park, Yong Chul Kim, Jung Pyo Lee, Joel Schwartz, Ho Kim, Francesca Dominici, Ziyad Al-Aly, and Michelle L. Bell. Air Pollution and Acute Kidney Injury in the US Medicare Population: A Longitudinal Cohort Study. *To appear in Environmental Health Perspectives*, 2023

5. Ana Trisovic, Thomas Pasquier, Matthew K Lau, and Mercè Crosas. A Large-Scale Study on the Quality and Reproducibility of Open Research Outputs in R. *Nature Scientific Data*, 2022
6. Daniel Garijo, Hervé Ménager, Lorraine Hwang, Ana Trisovic, Michael Hucka, Thomas Morrell, Alice Allen, Task Force on Best Practices for Software Registries, and SciCodes Consortium. Nine best practices for research software registries and repositories. *PeerJ Computer Science*, 2022
7. Ana Trisovic, Philip Durbin, Tania Schlatter, Gustavo Durand, Sonia Barbosa, Danny Brooke, and Mercè Crosas. Advancing Computational Reproducibility in the Dataverse Data Repository Platform. *The 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS)*, 2020
8. Xiaoli Chen, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Ana Trisovic, et al. Open Is Not Enough. *Nature Physics*, 2019
9. Anna E. Woodard, Ana Trisovic, Zhuozhao Li, Yadu Babuji, Ryan Chard, Tyler Skluzacek, Ben Blaiszik, Daniel S. Katz, Ian Foster, and Kyle Chard. Real-Time HEP Analysis With FuncX – a High-Performance Platform for Function as a Service. *The 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP)*, 2020