

Edge Data Analysis/Data Management Tools and Data Sets

Last Modified on 10/27/2021 11:14 am EDT

Edge is pleased to provide a comprehensive, categorized digest highlighting a number of these tools, data sets and other resources for the research community, many of which are available through various [EdgeMarket contract vehicles](#).

IEEE DataPort

- IEEE DataPort provides a unified data and collaboration platform which researchers can leverage to efficiently store, share, access, and manage research data, accelerating institutional research efforts. Researchers at subscribing institutions gain access to the more than 2,500 research datasets available on the platform and the ability to collaborate with more than 1.25 million IEEE DataPort users worldwide. The platform also enables institutions to meet funding agency requirements for the use of and sharing of data.
- https://njedge.net/wp-content/uploads/2021/09/Edge_IEEE-DataPort_Overview.pdf

Data Management, Data Sharing and Data Transfer

- The DMPTool is a free, open-source, online application that helps researchers create data management plans. These plans, or DMPs, are now required by many funding agencies as part of the grant proposal submission process. <http://dmptool.org>
- Zotero is a free, easy-to-use tool to help collect, organize, cite, and share research <https://www.zotero.org>
- Globus (<https://www.globus.org>) is an infrastructure for transferring large amounts of research data between one university and other universities participating in the Globus System. Researchers have been using Globus for years, either on departmental servers or on their desktop workstations or laptops. Primary access to the Globus data transfer service is through the web interface at [Globus.org](https://www.globus.org)
- IEEE Dataport
- Dataverse - Harvard-based tool to share, cite, reuse and archive research data. <https://dataverse.org>

Research Collaboration Platforms

- CodeOcean - Cloud-based executable repository and computational reproducibility platform allowing researchers to share, discover and run published code. (<https://codeocean.com>)

Datasets and Data Research Services

- Wharton Research Data Services (WRDS)WRDS provides researchers with one location to access over 350 terabytes of data across multiple disciplines including Accounting, Banking, Economics, ESG, Finance, Healthcare, Insurance, Marketing, and Statistics. (<https://wrds-www.wharton.upenn.edu>)
- <https://ieee-dataport.org>

Basic document sharing and collaboration tools

- Box - A cloud-hosted platform that allows researchers to store and share documents, photos, research materials and other files for collaboration. Box allows users to simultaneously edit Microsoft Office documents.
- Jupyter Notebook is an open-source web application that allows one to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. JupyterHub

brings the power of notebooks to groups of users. (<https://jupyter.org/>)

- Kubernetes is an open-source container-orchestration system for automating application deployment, scaling, and management of containerized applications. <https://kubernetes.io>

Survey tools

- Qualtrics - Qualtrics is a powerful, full-featured web-based platform for creating, sharing and conducting online surveys. <https://www.qualtrics.com>

Visualization tools

- MATLAB - High-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numerical computation. MATLAB also provides 2-D and 3-D plotting and visualization capabilities. Ancillary toolboxes also provide advanced image processing functionality for biomedical and remote sensing applications.
- Mendeley - A platform comprising a social network, reference manager, article visualization tools. https://www.mendeley.com/?interaction_required=true
- Gnuplot - Gnuplot is a popular plotting software. <http://www.gnuplot.info/>
- Tableau - Data Visualization tools <https://www.tableau.com/>
- Looker - Data visualization tools & software for interactive analytics <https://looker.com/product/visualizations>

Analysis Tools:

- R is free to download and use, and all the codes are open. At the same time, users can easily add their own programs (once they are familiar with statistics and programming).
- SAS is an integrated statistical package. It is a powerful statistical-analysis and data-management system for complex data sets. It is especially strong in analysis of variance (ANOVA), the general linear model, and their extensions. <https://www.sas.com>
- SPSS performs statistical analysis on quantitative data. The graphical user interface makes statistics analysis easier, including most complex models. Mendeley - A platform comprising a social network, reference manager, article visualization tools. https://www.mendeley.com/?interaction_required=true
- Stata is a command-based statistical package that offers a lot flexibility for data analysis. The program language keeps a simple structure, so is easy to learn, allowing users to focus on the statistical modelling. <https://www.stata.com>
- Excel is good is for the simplest descriptive statistics, or for more than a very few columns. It is easy to use for basic data analysis, and is much more convenient for data entry and shape manipulating.
- NVivo is a qualitative data analysis package. It helps researchers organize and analyze complex non-numerical or unstructured data, both text and multimedia. The software allows users to classify, sort, and arrange thousands of pieces of information. It also accommodates a wide range of research methods. It supports

documents in many languages. <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>

- Mathematica - Fully integrated technical computing software. Wolfram Mathematica is a modern technical computing system spanning most areas of technical computing — including neural networks, machine learning, image processing, geometry, data science, visualizations, and others. The system is used in many technical, scientific, engineering, mathematical, and computing fields. (<https://www.wolfram.com/mathematica/>).
- Quantum GIS (QGIS) is a free and open-source cross-platform desktop geographic information system application that supports viewing, editing, and analysis of geospatial data. is open source GIS software, available for both Windows and Mac OS. <https://qgis.org/en/site/>

ML, AI, Deep Learning

- Tensorflow - An end-to-end open source machine learning platform <https://www.tensorflow.org>
- Caffe is a deep learning framework. <http://caffe.berkeleyvision.org>
- Amazon SageMaker is a service that enables a developer to build and train machine learning models for predictive or analytical applications in the Amazon Web Services (AWS) public cloud.

Other HPC tools

- SLURM - Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters. <https://slurm.schedmd.com/overview.html>
- Apache Spark is a unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. (<https://spark.apache.org>)
- OpenMP – OpenMP OpenMP is an application programming interface that supports multi-platform shared memory multiprocessing programming in C, C++, and Fortran, on many platforms, instruction set architectures and operating systems (www.openmp.org)

Software Version Control and Project Hosting

- GitHub - Online software project hosting using the Git revision control system. (<https://github.com>)

Programming Tools

- Intel High-performance Fortran/C/C++ compilers and MKL math library. <https://software.intel.com/en-us/parallel-studio-xe>
- Python - is an interpreted, high-level, general-purpose open-source programming language. Python is used in many domains, and widely used in scientific and numeric computing. <https://www.python.org>
- Java

Debugging Tools

- Allinea's Distributed Debugging Tool (DDT). It can be used to debug serial and parallel (OpenMP and/or MPI) codes. <http://www.allinea.com/products/ddt>
- MAP - A performance profiler from Allinea. <http://www.allinea.com/products/map>

- GDB - Open Source standard GNU debugging environment (<http://www.gnu.org/software/gdb/>)

Cloud

- AWS Storage Services
 - **AWS Research Resources:** <https://aws.amazon.com/blogs/publicsector/resources-researchers-institutions-work-remotely/>
 - **AWS Educate Resources:**
- Amazon EC2 P2 Instances - Powerful, Scalable GPU instances for high-performance computing.
<https://aws.amazon.com/ec2/instance-types/p2/>
- Google for Education Resources:
 - https://edu.google.com/programs/?modal_active=none
 - https://edu.google.com/programs/credits/teaching/?modal_active=none
 - https://edu.google.com/programs/credits/research/?modal_active=none
 - <https://cloud.google.com/public-datasets>

Grant Reports

- The Grants Office reports: <https://www.grantsoffice.com>