



CSCAR CONSULTING FOR STATISTICS,
COMPUTING & ANALYTICS RESEARCH
UNIVERSITY OF MICHIGAN

ARC ADVANCED
RESEARCH COMPUTING
UNIVERSITY OF MICHIGAN



Data Science Initiative

CSCAR: Consulting for Statistics, Computing, and Analytics
Research

CSCAR evolved out of the University of Michigan Statistical Laboratory that began in the 1940's.

We assumed roughly our present form in the early 1980's under the name "Center for Statistical Consultation and Research".

Major expansion starting in 2016 with support of the Data Science Initiative.

CSCAR's mission is to support research that uses data and computation, through consulting, training, and provision of analysis services.

Consulting: guidance provided to researchers on issues arising in specific projects with the goal of empowering them to perform their analyses independently.

Training: workshops in methods and tools for data analysis and research computing

Analysis services: CSCAR analysts plan/conduct/implement/report analyses (usually via recharge at an effort percentage)

Open to researchers from all disciplines/all skill levels

Mechanics

- Call CSCAR or use our web-request form to schedule a 1 hour appointment with a consultant
- Remote appointments via Bluejeans, Skype, ...
- Walk-in to CSCAR (Rackham Building) without an appointment (only GSRA consultants available without appointment)
- Send a question via email to ds-consulting@umich.edu
- Sign up for a free workshop; register for a fee-based workshop (details at cscar.research.umich.edu)
- Email cscar@umich.edu to discuss hiring a CSCAR analyst for a project

CSCAR Staff

- ~14 staff consultants (~10 FTE's)
- Most have PhDs (in Statistics, Biostatistics, Math, Computer Science, Psychology, ...)
- 6 GSRA's (Biostatistics, Statistics, ISR)
- Selected/trained on technical skills, communication skills, breadth of research experience, self-management, ...

Core/foundational support

- Formulation of research aims
- Development of plans for data collection and analysis
- Statistical study design including power analysis and sample size assessment
- Data visualization/statistical graphics
- Interpretation and presentation of quantitative findings
- Strategies for using distributed/high performance computing infrastructure
- Identifying/compensating for bias/variation in data
- Profiling/optimizing/verifying code
- Uncertainty assessment
- Predictive methods
- Data modeling
- Implementation and optimization of algorithms
- Methods for high dimensional data
- Causal inference

Domain expertise

- Remote sensing/geospatial
- Distributed data processing
- Image/sound/video/text
- Genetics/genomics
- Machine learning
- Psychometrics
- Survey methods
- Observational studies
- Administrative data

Practical skills

- Software packages (many)
- U-M Infrastructure (Flux, Armis, MiDesktop)
- Methods for reproducible research
- Working with sensitive data
- Analysis plans for funding proposals, responding to reviewers
- Data management

Data Science Skills Series/ARC workshops

- R/Dplyr
- Python/Pandas
- R/Stan
- Python/Numpy
- Go for Data Processing
- Python/ArcGIS
- Python machine learning
- Hadoop/Spark
- Flux/batch computing
- Python toolz/streaming
- Linux command line
- Python databases
- R data exploration
- Python regression analysis
- R ggplot
- Python mixed models
- R data.table and big data sets
- Python survival analysis
- Python missing data and imputation
- Python databases
- Golang leveledb
- Python profiling and optimization
- Python numerics: numexpr, theano, ...
- Python/Matplotlib

Fee-based workshops

- Statistics - a review
- Structural Equation Modeling
- Analysis of sample surveys
- Regression analysis
- Data analysis with R
- Multivariate analysis
- SPSS
- Stata
- SAS

CSCAR consultants can join your team and get hands-on with your project. Some success stories:

- Health services/outcomes research : processed more than 1 billion claims records, developed propensity-matched, time-dependent, survival regression for a non-randomized treatment, projected to US population, addressed dependent censoring (Medical school assistant professor)
- Satellite image processing: massive data restructuring and normalization (LSA graduate student)
- Cluster analysis and mapping of 400 million+ genomic sequences (Medical school assistant professor)
- Developed custom multi-level regression Hamiltonian MC sampler for massively crossed linguistics experiment (LSA postdoc)
- Isolation of individual driving characteristics in >1TB of naturalistic driving behavior (MIDAS Challenge Project)

Funding for dataset acquisition

DADS: Data Acquisition for Data Science

Funding can be used to purchase licensed or commercial data; to fund preparation of existing data; to pay for storage costs

Data should in general be open to all U-M researchers; appropriate controls for sensitive data can be accommodated.

More information at: arc.umich.edu/dads

Trials management

CSCAR maintains a web application for managing trials:

<http://cscar-randomization.appspot.com/dashboard>

- Sequential randomization
- Balance with respect to covariates
- Access based on UM credentials
- Support for logging and data export
- Free to use, open source; source code is available:

<https://github.com/kshedden/randomization>