

Driving Biomedical Projects

Rather than develop algorithms for causal discovery in a vacuum, we selected 3 very different biomedical problems to use as test beds for our algorithms and to drive the development of new algorithms that meet the needs of biomedical researchers:

- **Cancer driver mutations** – focus on merging multiple omic datasets to discover combinatorial patterns of cancer drivers and pathways
- **Lung disease susceptibility & progression** – focus on analyzing tissue images and omics data to discover pathogenic pathways in chronic obstructive pulmonary disease & idiopathic pulmonary fibrosis
- **Functional connectome of the brain** – focus on using fMRI data to identify characteristic patterns of causal connections among regions of the human brain

Teams of data and biomedical scientists for each project work closely with the core group of algorithm, software, and systems architecture developers to create efficient, user-friendly programs for exploring causality in large complex data.

Training Opportunities

We teach biomedical scientists the concepts and application of causal discovery with graphical models (including what types of problems are appropriate), and we teach data scientists how to incorporate causal discovery methods into computational workflows and how to develop new algorithms, software, and systems for causal discovery. Please visit our Training website for updates on available in-person, online, and downloadable resources (www.ccd.pitt.edu/training/):

- **Summer Short Course**
- **Workshop** – one-day session at various conferences
- **YouTube Channel** – view our Distinguished Lectures in Causal Discovery, short course, & workshops
- **Causal Discovery Library** – download key literature, Technical Reports, & presentation slides
- **Open Learning Initiative** – CMU introductory course on Causal & Statistical Reasoning

We will also create interactive tutorials and a help desk as our software is released.

The CCD is supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. (www.bd2k.nih.gov)

National Institutes of Health

Big Data to Knowledge (BD2K)

Center of Excellence

datascience.nih.gov/bd2k



Center for
Causal
Discovery

University of Pittsburgh

Carnegie Mellon University

Pittsburgh Supercomputing
Center

Yale University

ccd.pitt.edu

bd2kccd@pitt.edu

412.648.9213 (tel)

412.648.9118 (fax)

Algorithm Development Team

CCD investigators are working to optimize and speed up existing causal discovery algorithms capable of modeling hidden variables: constraint-based algorithms, which use tests of conditional independence, and Bayesian algorithms, which allow the specification of structure and parameter prior probabilities. The team also works with the driving biomedical project teams to develop new algorithms designed to address the varied needs of biomedical researchers across disciplines.

Software Development Team

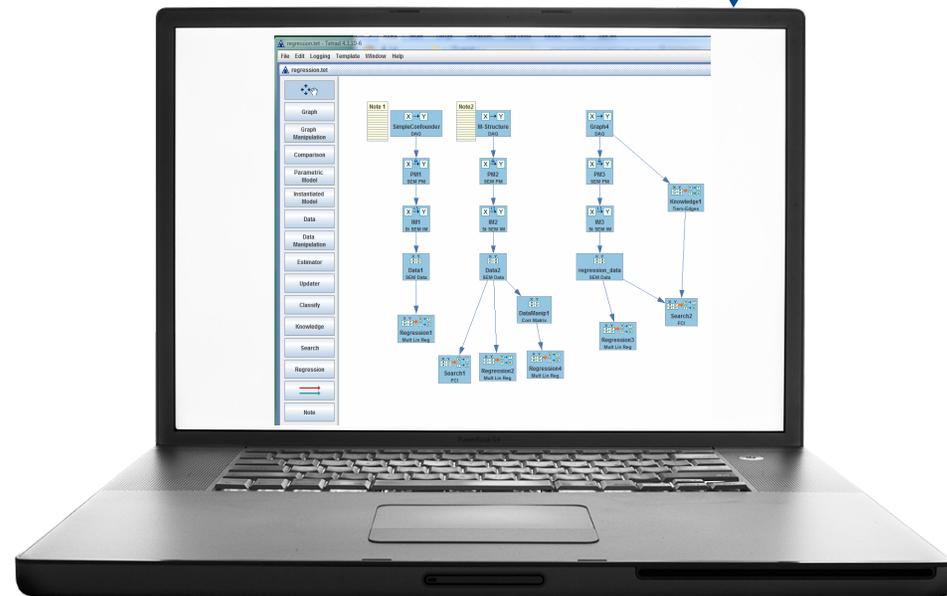
Investigators in the CCD are working to make causal discovery algorithms accessible and useful to a wide variety of biomedical researchers who might not otherwise take advantage of them. These algorithms will be freely available as open-source application programming interfaces (APIs) for “one-stop shopping” for scientists who wish to incorporate causal discovery methods into their research.

Systems Architecture Team

CCD investigators, including those at the Pittsburgh Supercomputing Center, are developing a graphical user interface that can run the APIs on your desktop computer. Backend processing of causal analyses can take place on the desktop machine for tasks that are not too computationally demanding, while more demanding tasks are automatically relayed to and performed on a high-performance computer cluster.

Product Release

By fall 2015, we anticipate releasing our first desktop graphical use interface that will implement at least one CCD algorithm, handle datasets with thousands of discrete and continuous variables, output a causal graph as an XGMML file, and support seamless job submission of analyses to a high-performance computing environment. Releases of software, algorithms, and other CCD products will be announced on our website (www.ccd.pitt.edu/data-science/).



User Team

You as an end user will download free, open-source APIs to your desktop that include wizards for selecting the algorithms best suited to your questions and data and to guide you as you query, view, compare, annotate, store, and share the models generated. Your input on your own causal discovery needs and your feedback on the software and algorithms we create will be invaluable to our work, so please contact Genine Bartolotta (bartgm@pitt.edu) to be added to our ccd-users mailing list. You can also send queries and suggestions directly to Harry Hochheiser, PhD (harryh@pitt.edu). You might also consider becoming a CCD Scientific Catalyst – an advocate for causal discovery in biomedical research (www.ccd.pitt.edu/consortium/scientific-catalysts/).