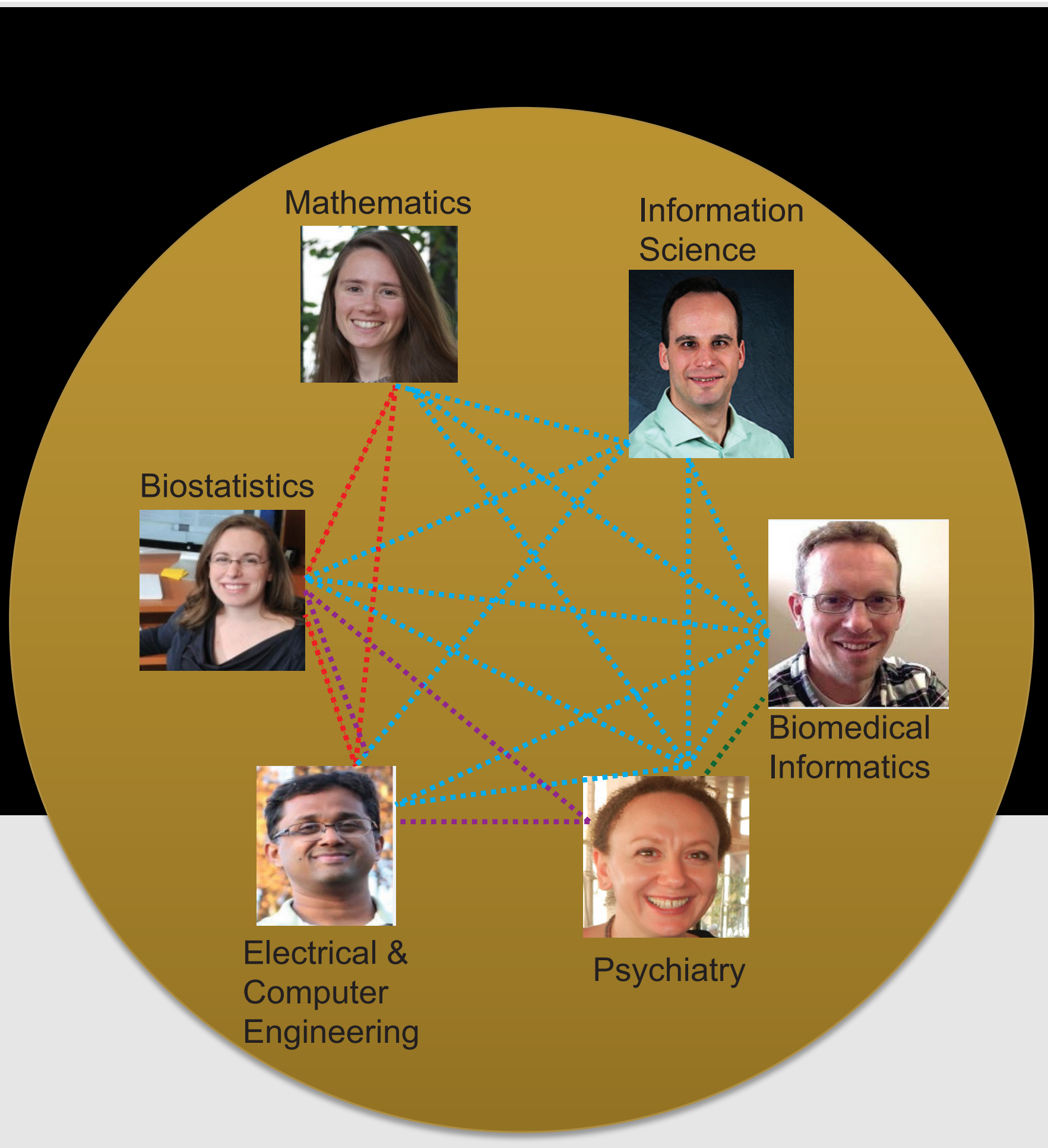
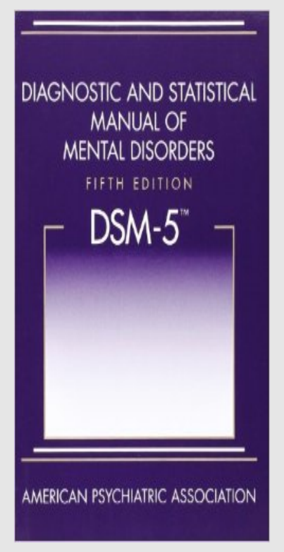
Interactive Ensemble Clustering for Mixed Data with Application to Mood Disorders

Rachael Blair Hageman, PhD[1]; Brian Chapman, PhD[2]; Arianna Di Florio, MD, PhD[3]; Ellen Eischen, PhD[4]; David Gotz, PhD[5]; Mathews Jacob, PhD[6]

[1]Department of Biostatistics, University of Buffalo; [2]Department of Biomedical Informatics Research and Department of Radiology, University of Utah; [3]Department of Psychiatry, University of North Carolina School of Medicine; [4]Department of Mathematics, University of Oregon; [5]School of Information and Library Science, University of North Carolina; [6]Department of Electrical and Computer Engineering, University of Iowa



Introduction


DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS
FIFTH EDITION
DSM-5
AMERICAN PSYCHIATRIC ASSOCIATION

Mental disorders are among the most elusive conditions in medicine and defy simple models, be they biological, psychological, social, or any simplistic admixture. [CMW14] In contrast to current classifications in other areas of medicine, those used in psychiatry, including the Diagnostic and Statistical Manual of Mental Disorders (DSM) [Ass13], rely on clinical manifestations (signs) and subjective reports (symptoms) rather than on the underlying causes and mechanisms.

Three landmark naturalistic studies funded by the National Institute of Mental Health (NIMH) provided some sobering statistics in this respect: psychiatric interventions are **effective in less than 25% of patients** presenting with an acute episode.17, 27 Diagnoses of mental health conditions are currently characterized by the following:

- Based on little objective evidence (almost arbitrary)
- No biological markers
- Co-morbidity (one person having multiple diagnoses)
- Heterogeneity within diagnosis (two patients with the same diagnosis can have two different sets of symptoms, with little or no clinical overlap)

In this project, we aim to develop a novel quantitative, big-data approach to enable precision diagnosis and treatment in this challenging application domain, with the ultimate goal providing data-driven tools that help clinicians significantly focus patient diagnosis and improve mental health outcomes.

Implications for Precision Medicine and Mental Health

“NIMH is committed to new and better treatments, but this will only happen by developing a more precise diagnostic system. Going forward, we will be supporting research projects that look across current categories -- or sub-divide current categories – to begin to develop a better system.” Tom Insel, MD

My Blog
Tom Insel, M.D.
Subscribe to RSS Feed for Blog
Home > About NIMH > Director's Blog > Posts from 2013
Director's Blog: Transforming Diagnosis
Recent Posts
In Vitro Veritas? September 16, 2013
Accessing and Assessing Science September 5, 2013
Antipsychotics: Taking the Long View August 28, 2013
Infantile Amnesia August 19, 2013
By Thomas Insel on April 29, 2013
In a few weeks, the American Psychiatric Association will release its new edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). This volume will tweak several current diagnostic categories, from autism spectrum disorders to mood disorders. While many of these changes have been contentious, the final product involves mostly modest alterations of the previous edition, based on new insights emerging from research since 1990 when DSM-IV was published. Sometimes this research recommended new categories (e.g., mood dysregulation disorder) or that previous categories could be dropped (e.g., Asperger's
Publications by the Director
Selected publications by NIMH Director Thomas Insel
Science News
Jay Giedd on PBS documentary "Brains on Trial"

This project aims to develop an appropriate big-data analysis pipeline for mental health data so as to inform precision diagnosis and treatment in this difficult application domain. Ultimately, the proposed quantitative analysis methods will provide data-driven tools – include interactive visualizations - that will significantly
improve patient diagnosis and mental health outcomes. In particular, the project aims to

1. Enable more precise, personalized **diagnoses**
2. Enable more precise, personalized **treatment decisions**
3. Enable the identification of **new treatment strategies**

Novel and Innovative Methods

The proposed methodology hypothesizes that a more precise and personalized classification of mental health diseases can be obtained through the development of novel integrated clustering methods that identify clinically significant structures within large population datasets. Such methods must:

1. Link treatment response to a very large number of variables, including clinical, biological (e.g., genomics), and psycho-social factors;
2. Overcome several real-world data challenges, including sparsity, ambiguity/uncertainty, missing data, high dimensionality, and heterogeneity;
3. Enable interactivity, allowing the user to inject prior knowledge into the classification process.

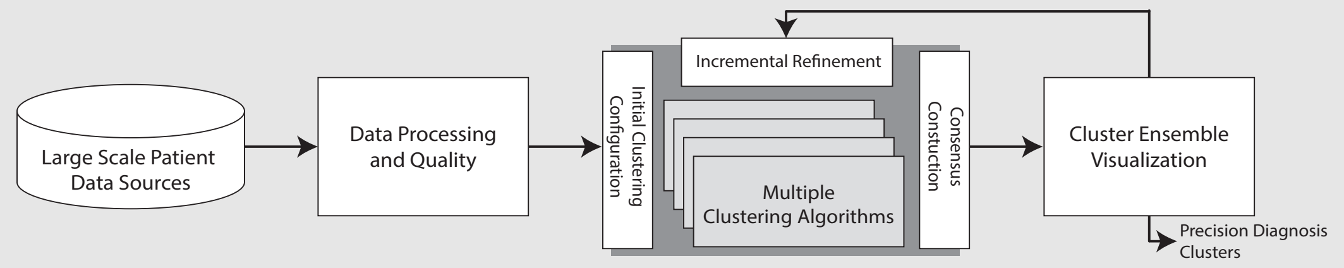

Large Scale Patient Data Sources
Data Processing and Quality
Initial Clustering Configuration
Incremental Refinement
Multiple Clustering Algorithms
Consensus Constuction
Cluster Ensemble Visualization
Precision Diagnosis Clusters

Figure 1: An overview of our proposed framework showing the key components and workflow.

The key components of our proposed methodology include development of:

1. A weighted ensemble approach that combines self-organizing maps and topological data analysis (including persistent homology)
2. A bootstrapping framework to estimate error
3. An efficient interactive clustering algorithm that can handle missing and mixed data-types
4. Interactive visualizations of clusters, which will ultimately aid clinicians

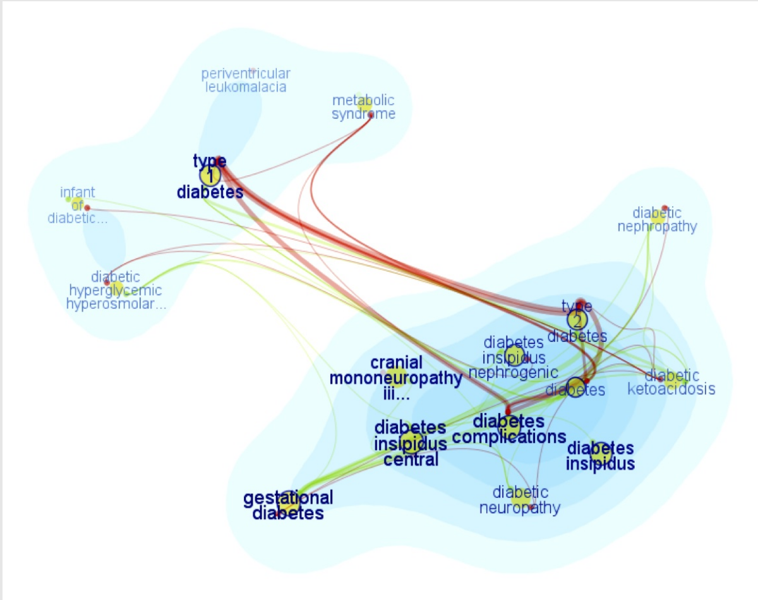


Figure 3: The FacetAtlas visualization [CSL+10], developed by Dr. Gotz and collaborators, uses multi-dimensional scaling (MDS) to position data items within a 2D visualization canvas based on pairwise similarity. Kernel density estimation (KDE) is used to compute the blue contours to summarize the dataset, while prominent nodes are rendered individually as landmarks. This view shows diagnosis data with dominant clusters corresponding to Type 1 and Type 2 diabetes.

One-Year Plan

- In the first year, the primary focus is on data heterogeneity, which arguably represents one of the most fundamental statistical concerns.
- This phase focuses on mood disorders (bipolar disorder and major depression) and their genetic and clinical heterogeneity as a simpler problem to develop and test the proposed methodologies. Reasons for this choice include the fact that mood disorders impact over 20% of U.S. adults at some point in life with an economic impact of $210 billion [GFS+15] and that the distinction between unipolar and bipolar depression remains subjective (thus with substantial potential for exploration of objective boundaries, based on the investigations of large data sets).
- Goals for year one include:
  - Identification of data sources
  - Preliminary investigation of algorithms to handle imperfect data, including missing and noisy data
  - Preliminary development of weighted ensemble clustering algorithms, focused on self-organizing maps and persistent homology
  - Static (not interactive) visualization of the clustering results using dimension reduction

Five-Year Goals

Data Science Methodology

The long-term goal is to produce novel algorithms that can address clustering and imputation under the same umbrella, thus being robust to missing and corrupted data. The algorithms will also be capable of accounting for additional priors, which will enable adaptive clustering.

In addition, the project aims to produce novel bootstrapping and consensus algorithms robust to imperfect data to determine the optimal parameters

Visualization Aim

Novel visualization algorithms will facilitate interactive clustering of data, coupled with adaptive clustering.

Precision Medicine Aim

A key goal is to predict treatment response failure (Validation) and to suggest potential paths for biologic exploration for alternative therapies

Broader Impact

Methodology will be applicable to other (currently) inappropriately clustered diseases

Broader Impacts

Mental Health:

Although the specific focus of the project is mood disorders, the methodologies will be applicable to a broad range of other mental health disorders which suffer from similar challenges of diagnosis and treatment. Therefore, the novel methods outlined in this proposal will lay the foundation for the future development of data-driven, personalized medicine tools for mental health. In particular, the proposed approach promises to enable more accurate and targeted diagnosis, and provide personalized evidence for treatment. Given the widespread impact of mental disorders, these techniques have the potential to significantly improve health outcomes for millions of patients.

Other Areas of Medicine:

The methodology is generalizable to other areas of medicine, beyond mental health, where similar diagnosis and treatment challenges are faced.

References

[Ass13] Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5, 5 edition ed. American Psychiatric Publishing, Washington, D.C, May 2013.

[CMW14] Craddock, N., and Mynors-Wallis, L. Psychiatric diagnosis: Impersonal, imperfect and important. British Journal of Psychiatry 204, 2 (2014), 93-95.

[CSL+10] Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu, FacetAtlas: multifaceted visualization for rich text corpora, IEEE Transactions on Visualization and Computer Graphics 16 (2010), no. 6, 1172{1181.

[GFS+15] Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., and Kessler, R. C. The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010). The Journal of Clinical Psychiatry 2010, February (2015), 155-162.

This project is supported by NSF Grants DMS-1557642, DMS-1557593 ,DMS-1557589, DMS-1557576, and DMS-1557668.