

Innovations Lab: Interdisciplinary Approaches to Biomedical Data Science Challenges

Examples of Mathematical/Statistical Topics Applicable to Precision Medicine

Machine Learning and Pattern Recognition

For large biomedical datasets, which may include clinical images, population informatics, and/or genetic maps, development of machine learning and pattern recognition efforts are of enormous importance. Unsupervised approaches in machine learning are methods being developed to seek out hidden structure or patterns in datasets that have not been previously labeled or annotated. When labels are known (and sometimes they are acquired at great expense), supervised learning methods can be used for training and prediction. In all cases, development of algorithms that jointly optimize statistical accuracy and run-time efficiency are needed.

Network Analysis

Networks arise naturally in social research settings (e.g. contact networks that affect disease transmission) and in genomics research (e.g. interaction relationships between genes, proteins, environment, etc.). Because dependency is typically an inherent feature of network analysis, traditional statistical approaches that assume independence between observations do not hold, and other statistical approaches are needed. In addition, the non-linear structure of networks makes data manipulation and curating, as well as mathematical modeling, challenging.

Causal Analysis

Analysis of causality, or the cause-effect relationship between two events or factors, is important for determining actionable relationships between various factors. These analytic tools can potentially be very useful for complex Big Data problems in precision medicine research.

Missing Data

In complex Big Data collected over large populations, and particularly in the joining of data sets, there will always be scenarios when some data are not collected completely. Developing solutions for precision medicine from such datasets will

require the usage, and perhaps the development or extension, of missing data methods and tools.

Natural Language Processing (NLP)

NLP is the process of converting written or spoken language into useful, digitized data. This is important in research involving electronic health records and/or social media data where ideas need to be captured from language for health research. A key challenge is that the underlying meaning of the language is of more interest than the words themselves. Some of the important tasks in this area are information extraction (ideas, identification of entities, relationships between entities, etc.), finding aspects of linguistic structure (part-of-speech tagging, syntactic structure, etc.), and machine translation.

Visualization: Visual Mining and Progressive data analysis

Visual mining has been shown to be able to augment the thinking of even experienced data analysts to better understand and make judgments about what data to use and which steps to take in complex Big Data settings. For example, uncertainty of data labels occurs frequently in population research setting, especially when merging or integrating disparate data sources. The challenge is how to meaningfully visualize the probabilistic arrangements between the variables and the uncertain labels in a manner which provides useful insight.

One of the problems with complex Big Data is that analyzing the complete dataset can take quite a long time. Progressive data analysis allows the researcher to view intermediate results in real-time, so that problems in analytic approaches can be corrected relatively rapidly. Both intermediate results and measures of confidence can be provided to facilitate real-time decision-making.