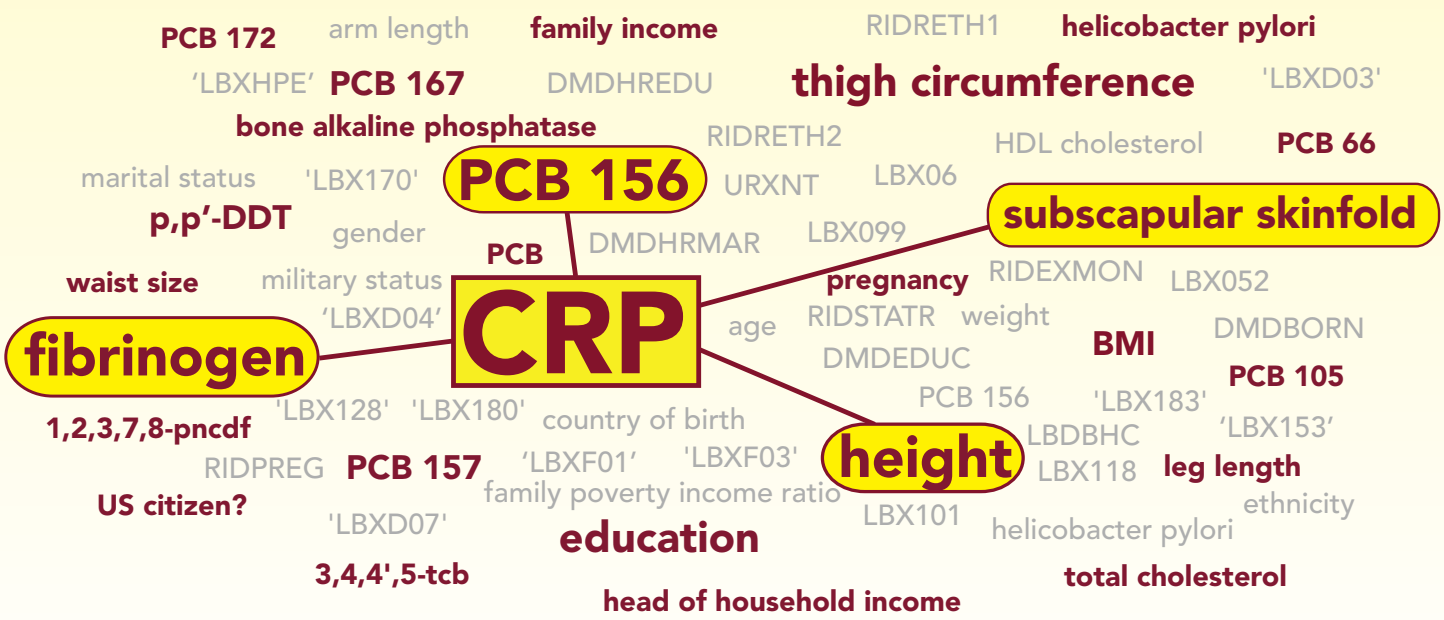
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Discovering Signatures When the Underlying Phenomenon is Poorly Understood

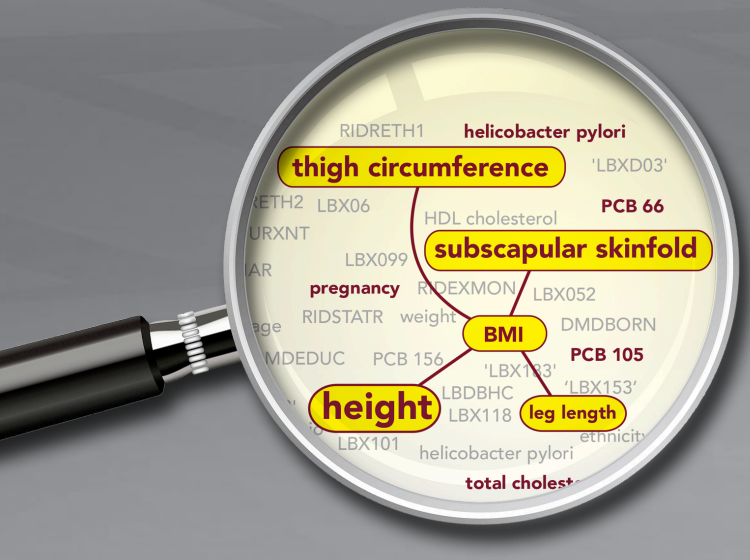
Feature extraction methods suffer from the weakness of assuming that all relevant information has been captured in the data as analyzed. We seek to develop signatures in situations where relevant data may be missing or where the underlying physical or conceptual model is weak or is not completely understood.

CHALLENGE

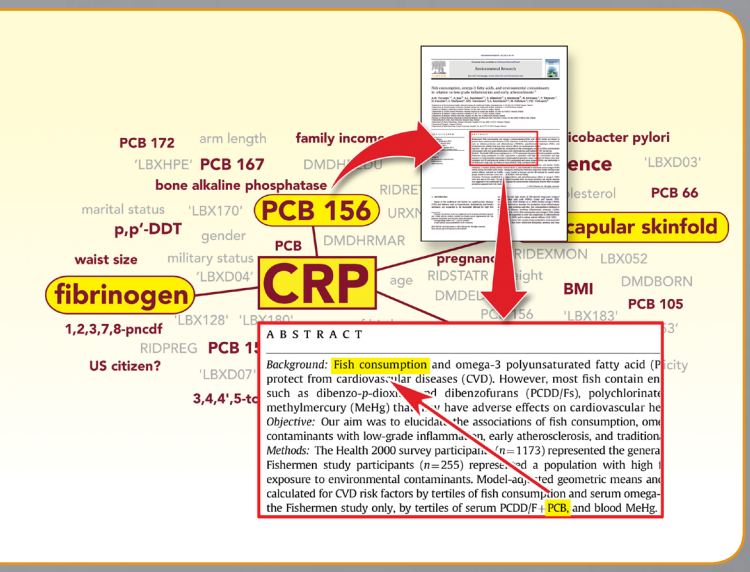
We are developing a framework for discovering signature features of complex or poorly characterized phenomena. Discovering the relationships that exist in a diverse set of features can be challenging, especially if the amount of data available is large, because problem spaces cannot be exhaustively explored. Our methodology allows exploration of large and varied feature spaces, resulting in a small set of features that can help make sense of the information available. The methodology we are developing has applications in environmental, economic, biological, medical, and social media fields.

APPROACH

Employing genetic algorithms and classification trees, our methodology is able to globally search a vast problem space to find relevant features and their relationship to a phenomenon of interest, bypassing some of the common problems associated with simpler decision trees or other clustering algorithms. Our approach separates important features from unimportant ones, narrowing the search and allowing more complete exploration of a smaller problem space. The features found will serve as a departing point for the search and analysis of text data and other related information. The methodology can be applied to searches of online databases containing scientific information, business performance data, and economic indicators among other areas.



Our Methodology Explores the Entire Feature Space, Weighing the Importance of Individual Variables, and Taking into Account Correlations Between Different Variables



A Greedy Search Algorithm Uses Previously Discovered Features to Search for New Ones

We have implemented a greedy search algorithm to discover features that may be potentially relevant to the underlying phenomenon. Because features selected are tested multiple times with different combinations of features and with different testing sets, the effect of noise, spurious correlation, and the inclusion of false features in a solution are minimized. The objective function we have developed combines conditional probability predictions derived from a naïve Bayes' classifier and results from binary classification trees to minimize a loss measure. Our implementation also includes a ranking of relative feature importance to help make sense of the problem space and define sub-regions of interest that are amenable to more thorough exploration. In our approach, simultaneous changes to multiple features appearing in a solution are possible, allowing for implicit estimation of the effect of features in the presence and absence of others. As our next task, we will use initial results to guide a search using text data from technical journals and the web to find and validate relationships between dataset features and the phenomena we are interested in.

IMPACT

Our approach has the ability to globally explore a large problem space that cannot be exhaustively searched and to look for relationships among features that may not be readily apparent to direct search methods. Our methodology has the ability to evaluate many different combinations of features, and explores the effect of simultaneous changes to multiple features on the underlying phenomenon of interest. The algorithm is not easily deterred by noise, and it offers great flexibility because the objective function can be modified to accommodate changes in user priorities.

Kristin H. Jarman
Pacific Northwest National Laboratory
kristin.jarman@pnnl.gov | (509) 375-4539

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*
U.S. DEPARTMENT OF ENERGY