# Intersection of Dataset Size and Number of Variables In Bayesian Network Structure Learning

By James Mann in the School of Biology

LAIDLAW SCHOLARS

University of St Andrews

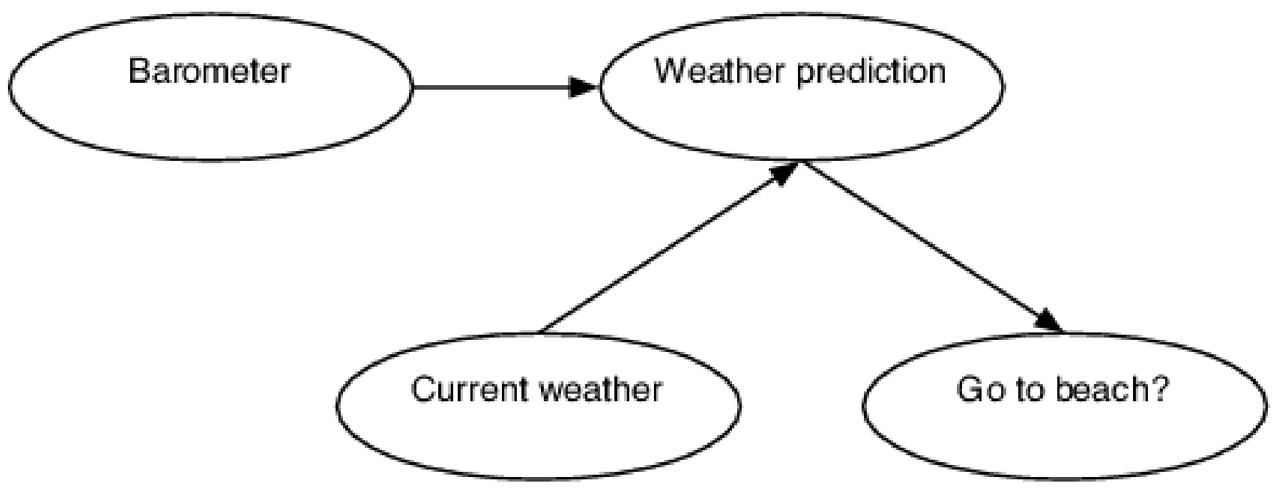
## What is a Bayesian Network ?

Bayesian Networks are a flexible type of mathematical model which has been used to predict Parkinson's disease risk, model ecosystems, and much more. A Bayesian Network is made up of nodes and links. Each node represents a different variable (see ovals in diagram on left) and links indicate which variables influence other variables.

Bayesian Networks can be used for prediction, take the network on the right: with a barometer reading and the current weather this Bayesian network can be used to predict the probabilities of a given weather condition in the future (weather prediction node).

Often a machine learning approach is used to find the structure of an underlying Bayesian Network given a set of data. This means a computer is used to find the network structure that best fits a given dataset. This is known as Bayesian Network learning.

## Example Bayesian Network For Weather

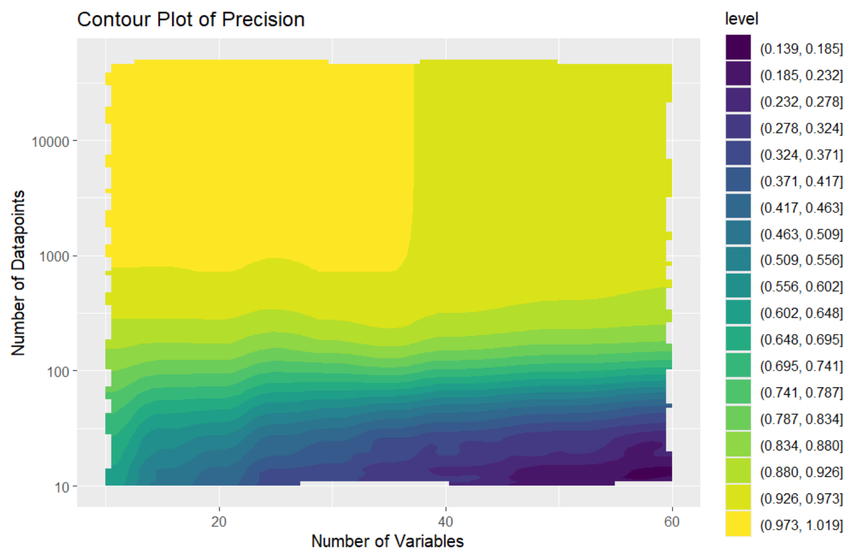


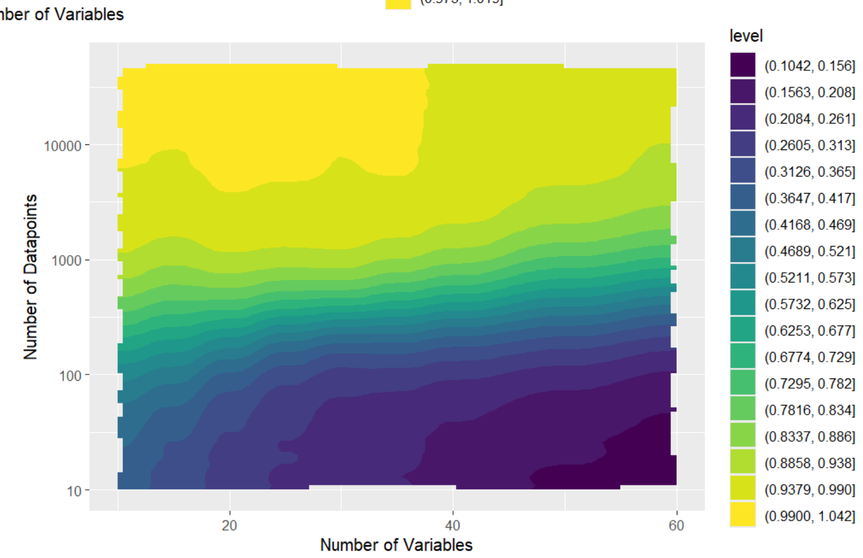

## Research Summary

My research project investigated the quality of Bayesian Network learning given the number of variables in the network and the number of datapoints in the sample the network was learned from. I used a simulation approach where a random Bayesian Network was created, and synthetic data generated from it. Then a Bayesian Network was learned from this data. Finally, the learned network was compared to the original network. The closer to the original network the learned network was the more effective the learning has been.

These simulations returned 2 metrics for a given number of variables and dataset size. Recall and Precision. Higher recall and higher precision is better. The results of the simulation are shown as heatmaps/contour plots on the right. A key takeaway is that larger datasets are better for learning and higher variable networks are harder to learn effectively for.

## Heatmaps Showing Precision and Recall for Various Variable and Datapoint Counts




I would like to thank Lord Laidlaw and the Laidlaw Foundation for the funding and support that made this work possible.

Thank you to my supervisor Dr. V Anne Smith for her insights and encouragement.