Big Data Knowledge Discovery Project: Executive Summary

The Big Data Knowledge Discovery (BDKD) project has achieved its ambitious goal of bringing the power of probabilistic machine learning and large-scale cluster computation to researchers in the natural sciences, and in doing so, has unlocked new opportunities for knowledge discovery.

The project has been a collaboration between the machine learning expertise of Data61 (formerly NICTA), the compute infrastructure capabilities of SIRCA, and three leading university research groups; The Comparative Ecology Lab at Macquarie University, The School of Geosciences at Sydney University and The Photonics Dynamical Systems Group at Macquarie University. These researchers have helped Data61 and SIRCA design and build new data-driven knowledge discovery tools which, in turn, have been used to generate novel scientific insights across a broad range of disciplines.

The promise of big data and machine learning is the ability to automatically extract knowledge and insight in ways that would be impossible for a human. However there are substantial engineering, algorithmic and mathematical challenges to achieving this promise, especially for a group of scientists without a specialised support team.

The tools built during the BDKD project are designed to address these challenges. They are a set of machine learning, data management and computational-cluster deployment systems that can be used independently in an existing scientific workflow, or put together to form an end-to-end pipeline. This includes

- Machine learning tools that can identify patterns and relationships in large volumes data, and can predict the unknown variables of scientific models from (potentially large amounts) of real-world data. These tools are unique in that they compute the confidence in their predictions while still being able to process large amounts of data. This understanding of uncertainty is critical for making decisions based on the scientific outcomes.
- An 'active sampling' tool that uses machine learning to infer the benefit of taking a measurement in an experiment. It can then direct the experimenter to redesign their tests to ensure measurements are recorded only in areas where there is low confidence in a prediction. It can even be hooked up directly to the experimental apparatus and automate the data collection entirely. The benefit for the scientist is a greatly improved prediction accuracy for the same number or fewer measurements, saving both time and money.
- An easy-to-use cluster compute platform that allows scientific users to deploy and run their computation/analytic software at scale on the cloud, as well as create, update, manage, retrieve and share large datasets in the cloud. This software has already demonstrated how it could revolutionise the productivity of some research by reducing the time needed to perform routine data calculations by an order of

magnitude.

The development of these and other tools, as well as the partnership with Data61 and SIRCA, has helped the university collaborators produce new data-driven scientific results in the fields of ecology, photonics and geoscience. This collaboration has also led to novel machine learning applications and research. A few highlights of these results are

- A state-of-the-art machine learning technique to efficiently simulate the evolutionary cycle of a forest. It enables the researchers to investigate how varying environmental conditions such as the frequency of bushfires can lead to vastly different ecosystems.
- A reconstruction of the motion of the Earth's tectonic plates, created by fusing information from a wide range of data sources such as magnetic signatures baked into ancient rock samples. The reconstruction offers a unique glimpse of the planet's surface billions of years ago.
- An entirely new geological map of the seafloor from training a machine learning algorithm on 15,000 samples of deep marine sediments. The new map completely changes our understanding of what's on the ocean floor, revealing a complex patchwork where previously large continuous belts of seafloor sediments were mapped.
- New software to analyse huge amounts of data from a complex laser system, looking for chaotic behaviour that could be used for new types of cryptographically secure communication.
- A set of novel algorithms for inferring the parameters of experimental laser systems from experimental data, as well as an active sampling control system that can automatically control an experimental laser in order to build a model of its parameter space.
- New machine learning methods that can incorporate existing scientific models into data-driven predictions. These algorithms can estimate unknown variables in models of natural processes, and know how confident they are about their answers.

The software tools created in this project have all been publically released with liberal licenses and ongoing support that will ensure they continue to be of use to scientists in all disciplines and industry. These tools have been publicised by the project, and are starting to see take-up in other areas of research. They also continue to be used by the various teams for many different tasks beyond the scope of the project.

The BDKD project has demonstrated the impact of the cross-pollination of ideas and techniques resulting from the interaction between the machine learning researchers and natural scientists. The scientists on the project now have the tools and understanding to make use of data-driven techniques within their scientific workflows. Similarly, the machine learning researchers have developed tools that incorporate, or work alongside, classical scientific models as an additional source of information to exploit for making predictions. The bi-directional flow of information between the groups stands as an example of how data-driven knowledge discovery will shape developed in this project proliferate into other fields.

Sample Images

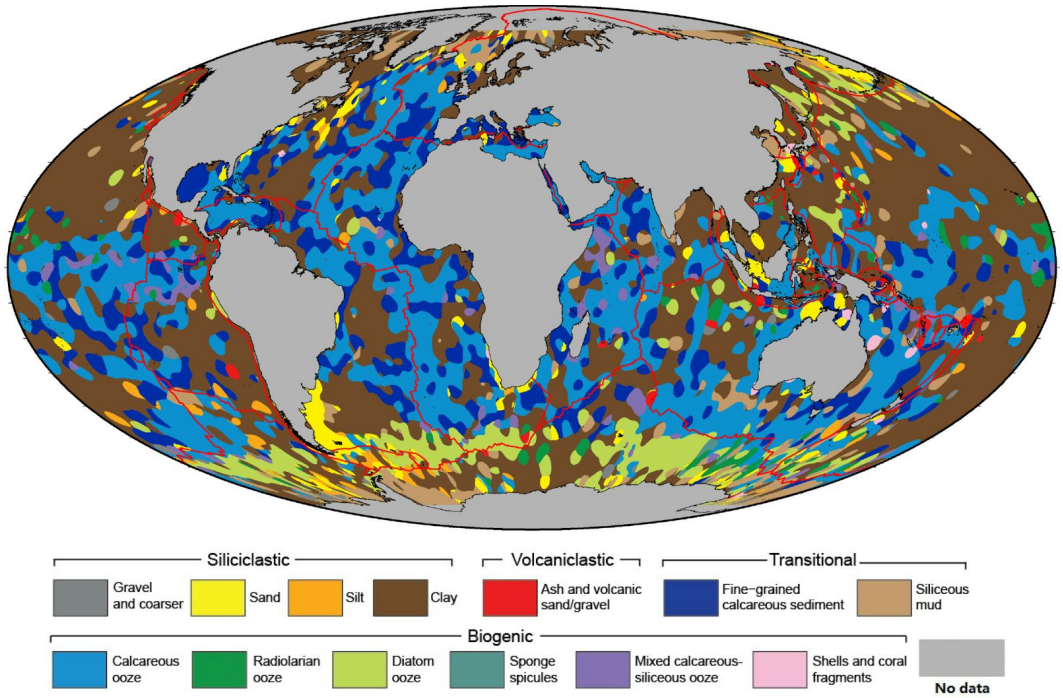

Figure: *Digital map of major lithologies of seafloor sediments in the world's oceans based on descriptions of 14,500 seafloor samples interpolated using a support vector machine learning algorithm.*

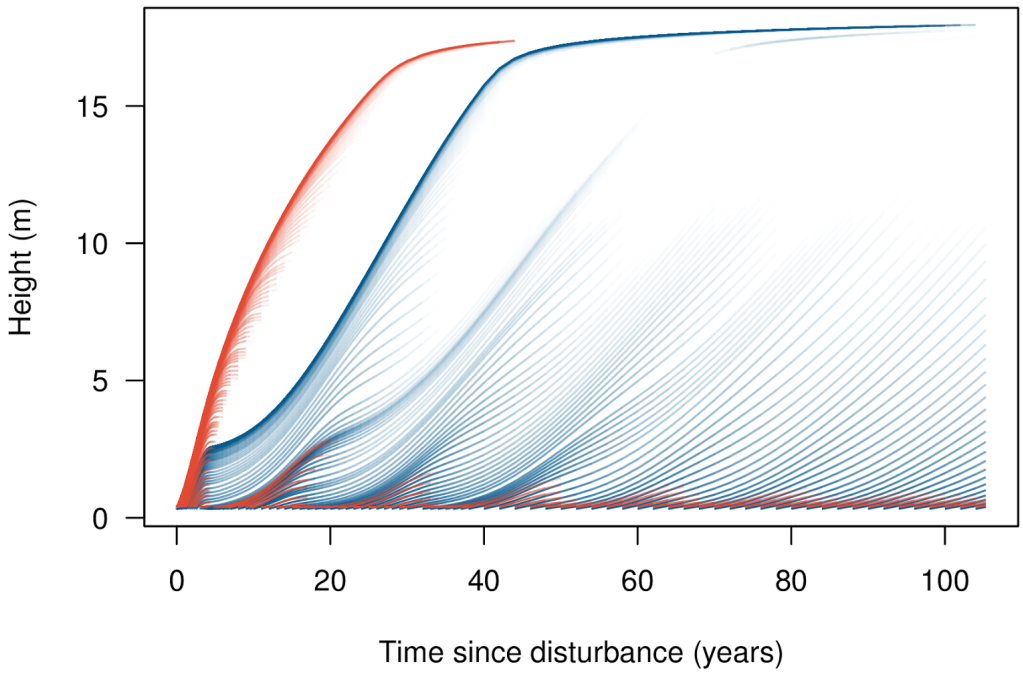

Figure: *Height trajectories for a collection of trees growing within a patch for two different species having different traits (indicated by colours) plotted against time since the most recent bushfire.*

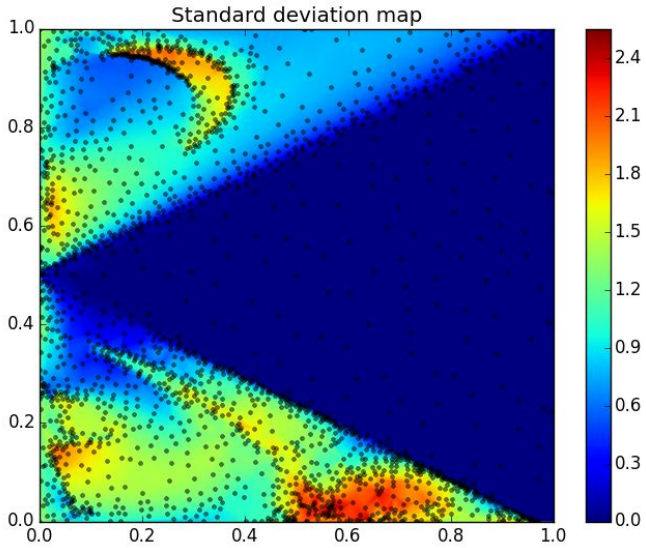

Figure: *Software algorithms that learn the behaviour of a laser system as new observations are acquired substantially reduce the number of observations required to understand its performance characteristics.*

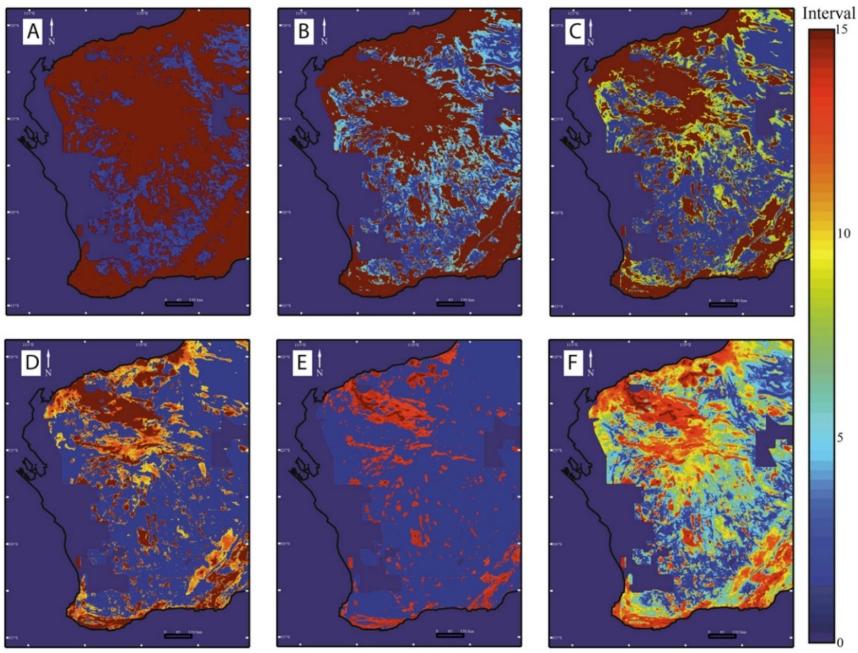

Figure: *Predictive confidence map for Western Australia derived from geophysical data using bracketed intervals based on the probability of a pixel to contain iron ore; (A) intervals 1–3 (lowest probability, 0–21.1%); (B) intervals 4–6 (low probability, 21.1–42.2%); (C) intervals 7–9 (medium probability, 42.2–63.3%); (D) intervals 10–12 (high probability, 63.3–84.4%); (E)*

intervals 13–15 (highest probability, 84.4–100%), and; (F) all intervals. The dull blue colour (interval 0) represents areas where there is no geophysical data.