Machine learning: An unexplored horizon in Arctic science

Joseph Cook, Vice President of the UK Polar Network (UKPN) Committee explores why machine learning is an unexplored horizon in Arctic science

Machine learning has become a disruptive force in many sectors, underpinning the technology that protects us from scam emails, calculates our credit scores, detects financial fraud and powers self-driving cars. Machine learning enables deep insights to be drawn from datasets that are too large or too complex to be digestible by humans. In this article, I will argue that it also has huge potential in the field of polar science.

Machine learning is an umbrella term for a large family of algorithms that share a common principle – they "learn" how to extract value from a dataset by being shown examples. The examples are known as the "training data". The algorithm explores the training data looking for statistical relationships between variables that it can use to make predictions about other data. A simplistic example is an algorithm designed to identify certain objects in images – apples, for instance.

The training set will contain many images that the researcher labels as "apple" or "not-apple". The algorithm examines each example and decides for itself what are the characteristics of apples it can search for in other images. With every example it sees, the algorithm's criteria are refined. Then, the algorithm can use those same criteria to decide whether apples are present in any image. This is an example of "supervised learning" which means that the training data includes labels as "correct answer" targets for the algorithm. In this example, there are just two labels, but in reality, there may be many.

The alternative is "unsupervised learning" where an algorithm decides for itself how to group the data into distinct classes. The power of these methods comes from the algorithm developing its own criteria for separating data and making predictions. The lack of explicit programming and the ability to iteratively churn through huge datasets makes machine learning a particularly powerful tool for extracting value from "big data".

Polar science is awash with "big data". One particularly important source is optical remote sensing – our "eye in the sky" instruments on drones, planes and satellites that provide measurements of reflected light that we can use to identify features and objects on the Earth's surface. Every pixel in every satellite image contains information about reflected light at several wavelengths (colours), adding up to a huge volume of data – ideal territory for mining information using machine learning.

With access to a reliable and sufficiently large training set, a "supervised classification" algorithm could be trained to scan through satellite imagery across the cryosphere and assess the composition of each individual pixel – assigning labels such as dry snow, wet snow, ice, melting ice, ice with dust, ice with algae, water, crevasse etc – providing the science community with automatically generated, highly detailed maps of the surfaces of glaciers, ice sheets, sea ice and snow fields that can be used to assess glacier change or predict safe routes for ice cap crossings, for example.

One of the strengths of machine learning is that the pipeline from data collection to classification can be automated, meaning the process can easily be repeated at regular intervals, limited only by the frequency that the relevant satellite or aircraft instrument passes overhead.

Between 2016 and 2018, I have been using these techniques to map the surface of the Greenland Ice Sheet from custom-built drones and from space (using data

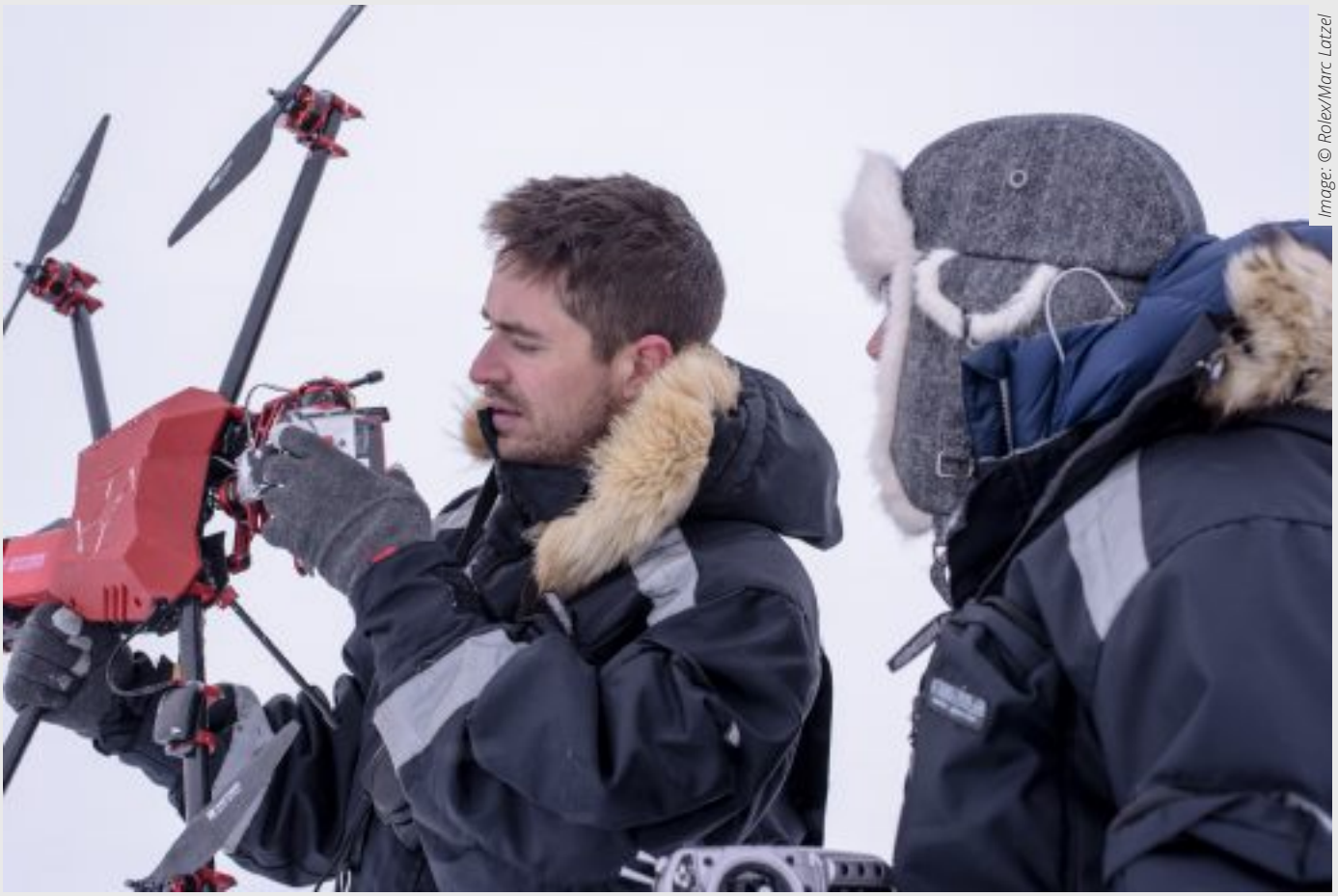

Image: © Rolex/Marc Latzel

from the European Space Agency's Sentinel-2 satellite). My specific aim was to quantify how much of the ice surface was covered by algal blooms and other particulates that have an accelerating effect on ice melt rates and use machine learning algorithms to evaluate changes over space and time. I trained supervised classification algorithms on reflectance measurements made at on-ice camps on the Greenland Ice Sheet (Figure 1), so that the algorithm could "learn" to identify subtly different features of the surface. This is just one application of one type of algorithm – there are countless others making up a huge unexplored opportunity space bridging machine learning with polar science.

While machine learning is a powerful addition to the polar scientist's toolbox, there are also limitations. Firstly, training data is relatively scarce. To generate a suitable training set for classifying snow and ice surfaces based on composition, our eyes in the sky rely upon feet on the ground. Scientists still need to visit field sites and make very detailed measurements of the ice surface composition and optical properties along with reflectance measurements that replicate those made by the relevant satellite instruments. This is the only way to generate the initial labelled training data set required by the algorithms. Field science can be expensive and time-consuming; however, as the technology is adopted more widely and a standard protocol for gathering this data is established, sharing between groups will allow a central repository of training data to grow.

"Machine learning is an umbrella term for a large family of algorithms that share a common principle – they "learn" how to extract value from a dataset by being shown examples. The examples are known as the "training data". The algorithm explores the training data looking for statistical relationships between variables that it can use to make predictions about other data."

The second challenge is that satellite measurements can be obscured by cloud – sometimes limiting the available data to a few images per season. For this reason, some researchers have been making equivalent measurements using planes and drones that fly beneath the cloud base. Another hurdle is computing power. Applying complex algorithms to huge datasets exerts a major computational load and may require

Continued on page 195 ➔

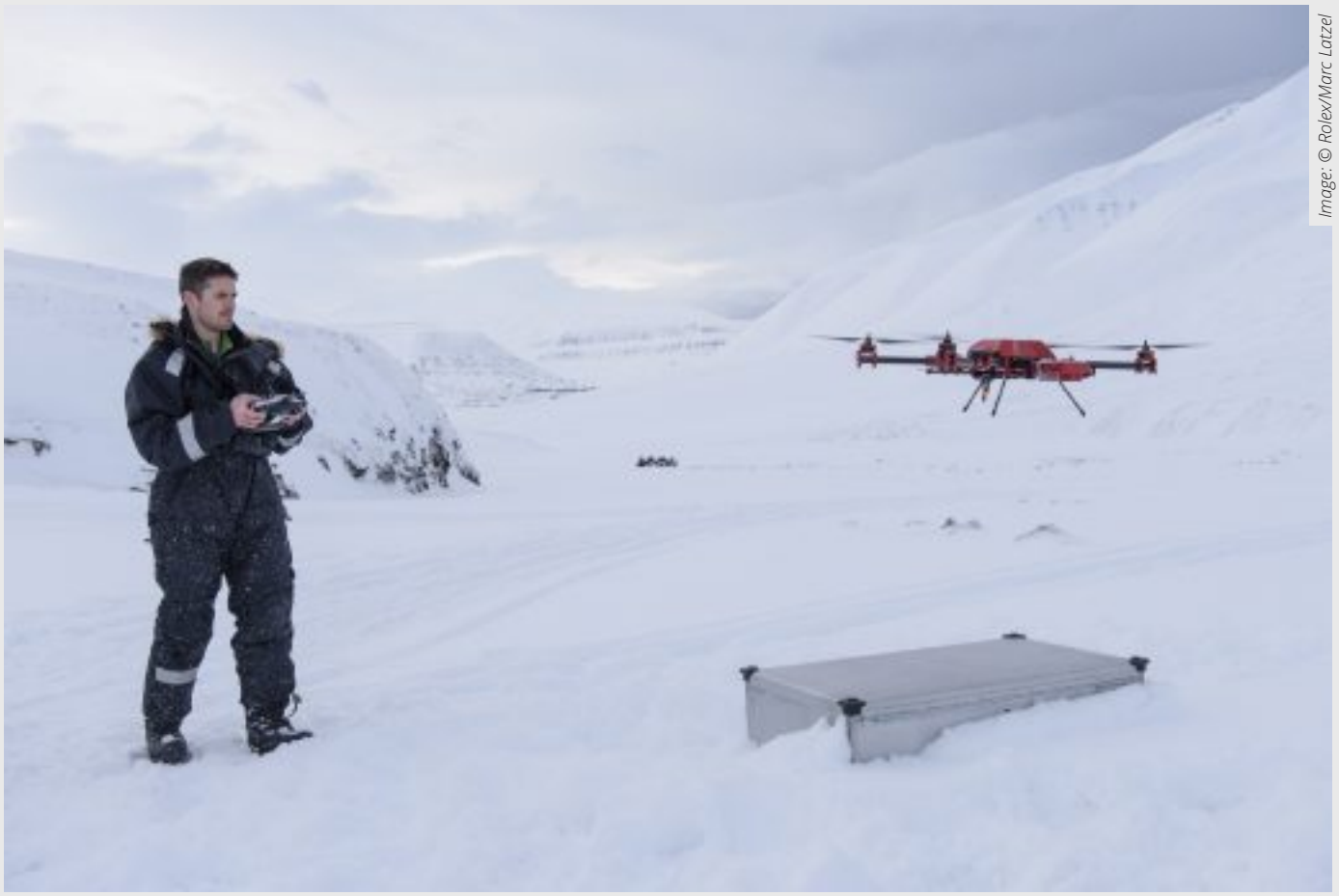

Image: © Rolex/Marc Latzel

← *Continued from page 193*

access to high- performance computing (HPC) clusters, which can often be awkward, expensive and may require specialist support from a research software engineer. However, the current boom in cloud computing services and access to remote servers such as the Amazon Web Services or Microsoft Azure are making computational heavy lifting achievable from a home laptop. A great example is the Google Earth Engine, which is already optimised and organised for geospatial data analysis via an interactive Javascript console running in the web browser or a downloadable Python API (Application programming interface).

As these technologies develop and as the availability of training data increases, machine learning will increasingly influence polar science and be an ever more powerful and accessible strategy for geospatial data analysis across the cryosphere.

Dr Joseph Cook is a polar scientist with interests in applying machine learning to polar science. He is a postdoctoral researcher on NERC's Black and Bloom project, a Rolex Laureate, World Frontiers Forum Pioneer and co-founder and director of the cryosphere focused science communication organisation "Ice Alive". He is also currently Vice President of the UK Polar Network. Please see www.icealive.org or http://tothepoles.wordpress.com for more information or www.github.com/jmcook1186/ for related codes and notebooks. ■

Joseph Cook
Postdoctoral Researcher, University of Sheffield, UK
Vice President of the UK Polar Network (UKPN) Committee
joe.cook@sheffield.ac.uk
tothepoles.wordpress.com
www.twitter.com/tothepoles