Advanced Geoscience Targeting via Focused Machine Learning
Applied to the QUEST Dataset
Justin Granek and Eldad Haber
Department of Earth and Ocean Science, University of British Columbia

Mineral Prospectivity Mapping
Over the last 20 years mineral exploration has begun adopting advanced data mining techniques to assimilate large data sets and identify prospective targets. The goal is to learn the mapping function which can predict the existence or absence of economic mineralization from a compilation of geoscience datasets.

Mathematically, the problem can be stated as follows: Given geoscience data X, and known mineralization y, find a mapping function f(X) to approximate the relationship between the data and the mineralization occurrences such that it can be used to predict mineral potential on a new dataset X^new.

Support Vector Machines
Since they were first proposed in the 1990s, support vector machines (SVM) have gained considerable popularity for numerous machine learning applications. As a maximum margin algorithm, the basic goal for any SVM is to learn the function which optimally separates the data in feature space. In the simplest algorithm, the basic goal for any SVM is to learn the function which can predict the existence or absence of economic mineralization.

Problem Characteristics
- Unbalanced Training Data: Since mineral occurrences are rare, there are far fewer positive training samples than negative.
- Training Label Uncertainty: No mineralization does not always mean there is none—only that it has not yet been found.
- Training Data Uncertainty: All data have uncertainty, whether from instruments, subjectivity, or variations in data quality.
- Big Data: Large regions and many data = large databases!

Uncertainty Management
Training Label Uncertainty: can be added to the standard SVM formulation via the addition of weights on the penalty term inversely proportional to the training label uncertainty (ν).

Training Data Uncertainty: can be incorporated via the addition of an extra penalty term in the SVM objective function. Under the assumption of gaussian errors and known variances, 2, the additional term aims to minimize the normalized difference between the observed data, X, and true data, X. With the addition of these two terms the optimization problem therefore becomes:

minimize 1/2||w||^2 + (1/2) max{0, 1 - f(X_w + b) + ε} + 1/2 \sum \text{diag}(\Sigma) (X - X^*)^2