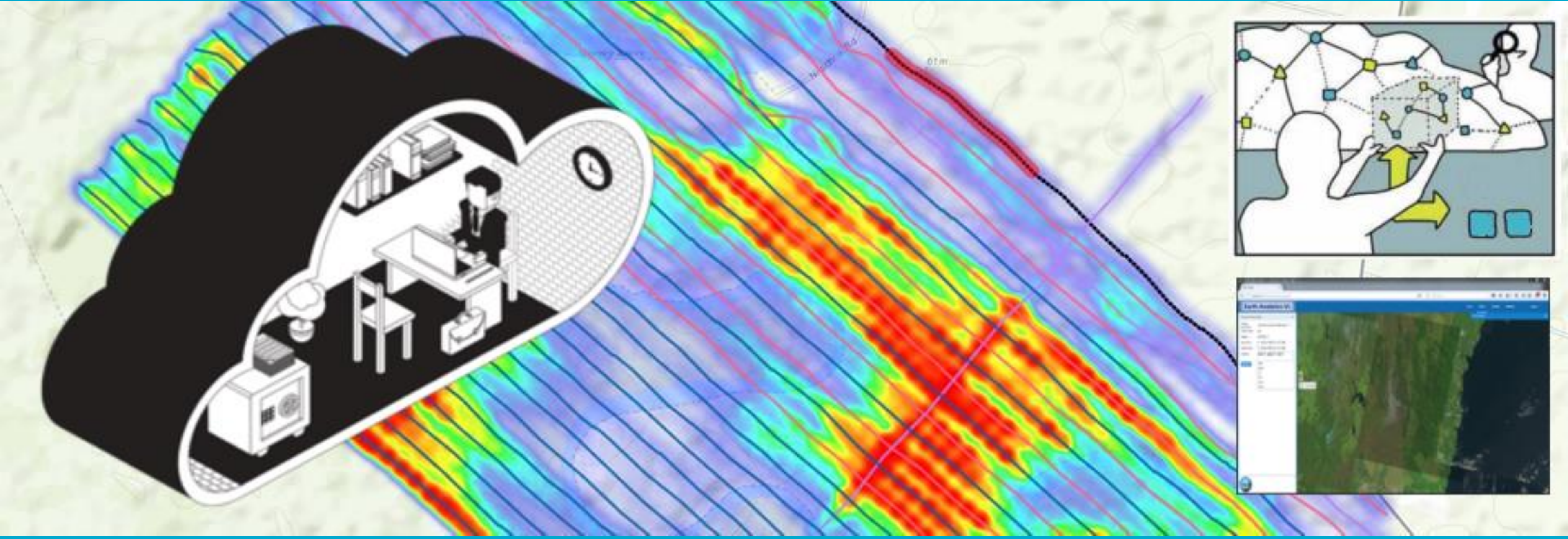


Can we Teach Machines Geology?

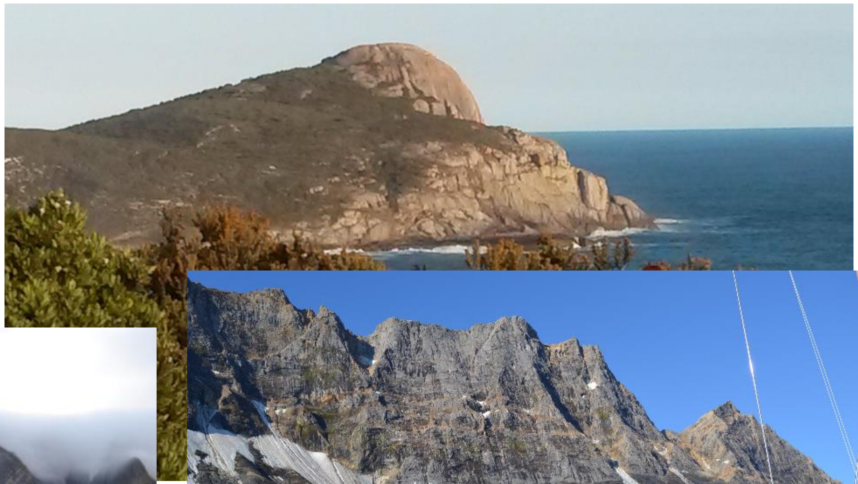
The Role of Data Science in the Geosciences

Jens Klump | Team Leader Geoscience Analytics

28 December 2019

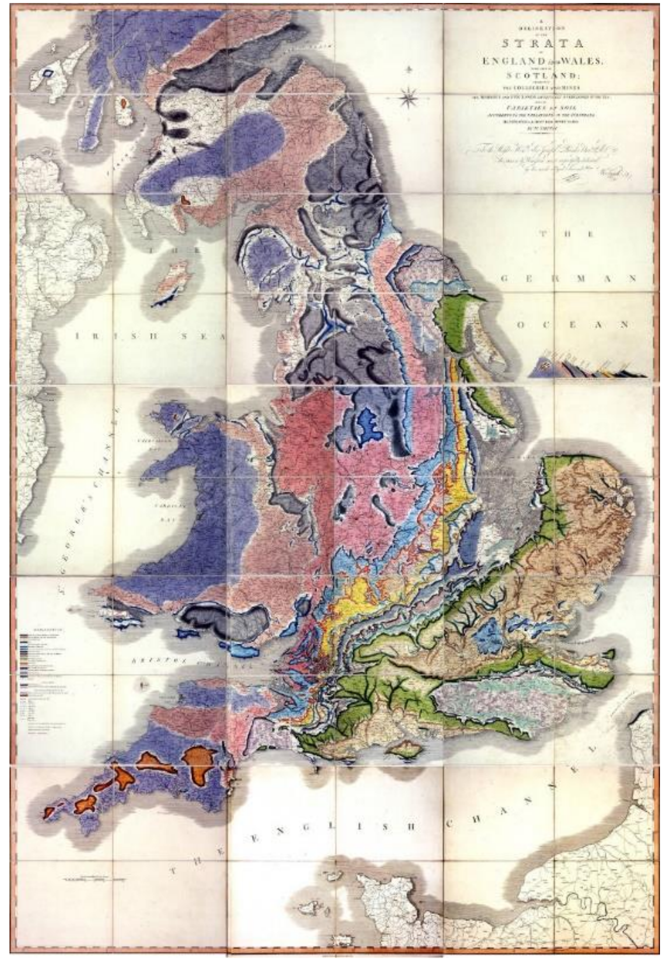
MINERAL RESOURCES
www.csiro.au

How geology presents itself



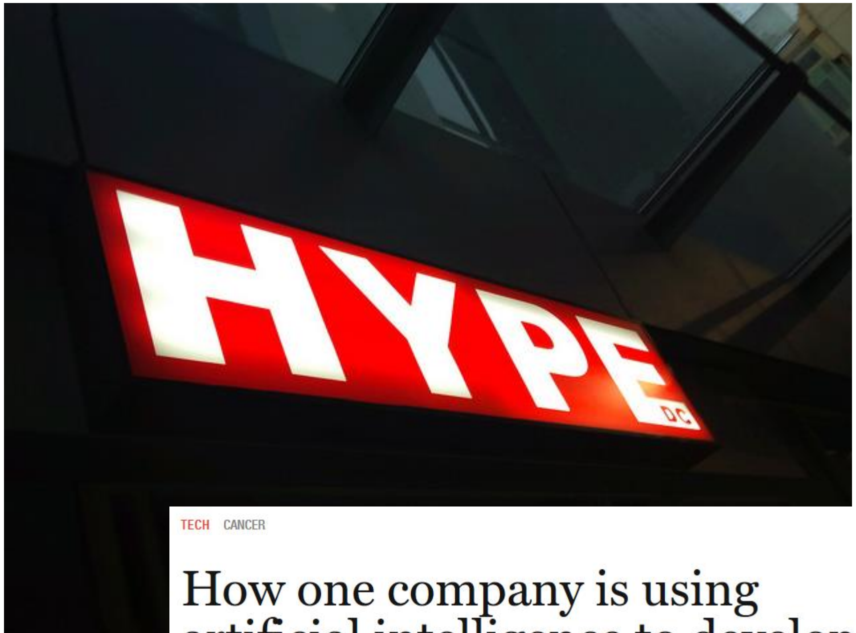
The Big Picture

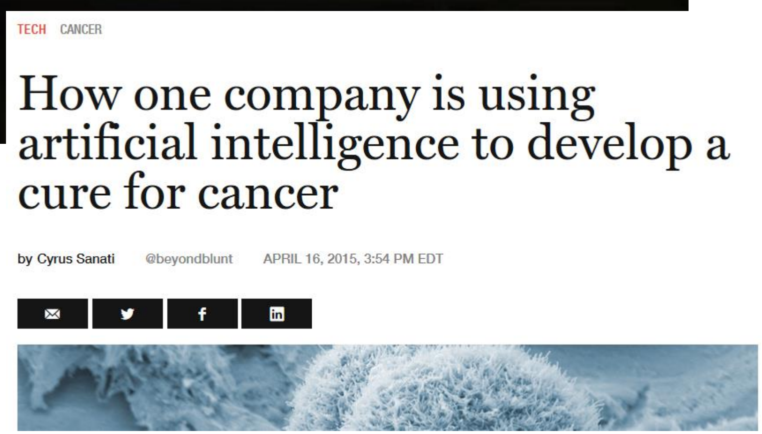
- William Smith, a self-taught civil engineer, connected the dots to create the first modern geological map.
- Visualising his conclusions as a map, from the data he had collected, was a major breakthrough.



Smith, W. (1815). Delineation of the strata of England and Wales. London, UK

The Promise of Machine Learning





TECH CANCER

How one company is using artificial intelligence to develop a cure for cancer

by Cyrus Sanati @beyondblunt APRIL 16, 2015, 3:54 PM EDT

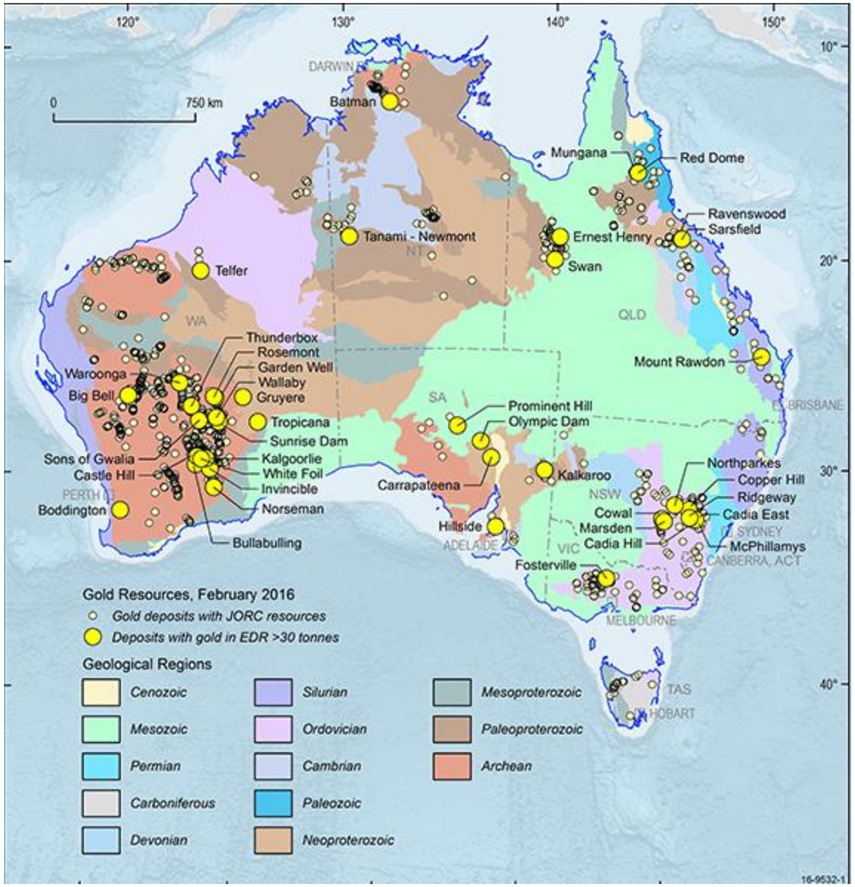
- Machine learning and data science are all the rage today.
- Machine learning is a great tool for the right problem.
- But which problems can be solved by machine learning?
- More specifically, where is the place for machine learning in geology and minerals exploration?

Sparse Data

Sparse Data

How much data do we need?

- Thousands of examples.
- No fewer than hundreds.
- Ideally, tens or hundreds of thousands for “average” modelling problems.
- Millions or tens-of-millions for “hard” problems like those tackled by deep learning.

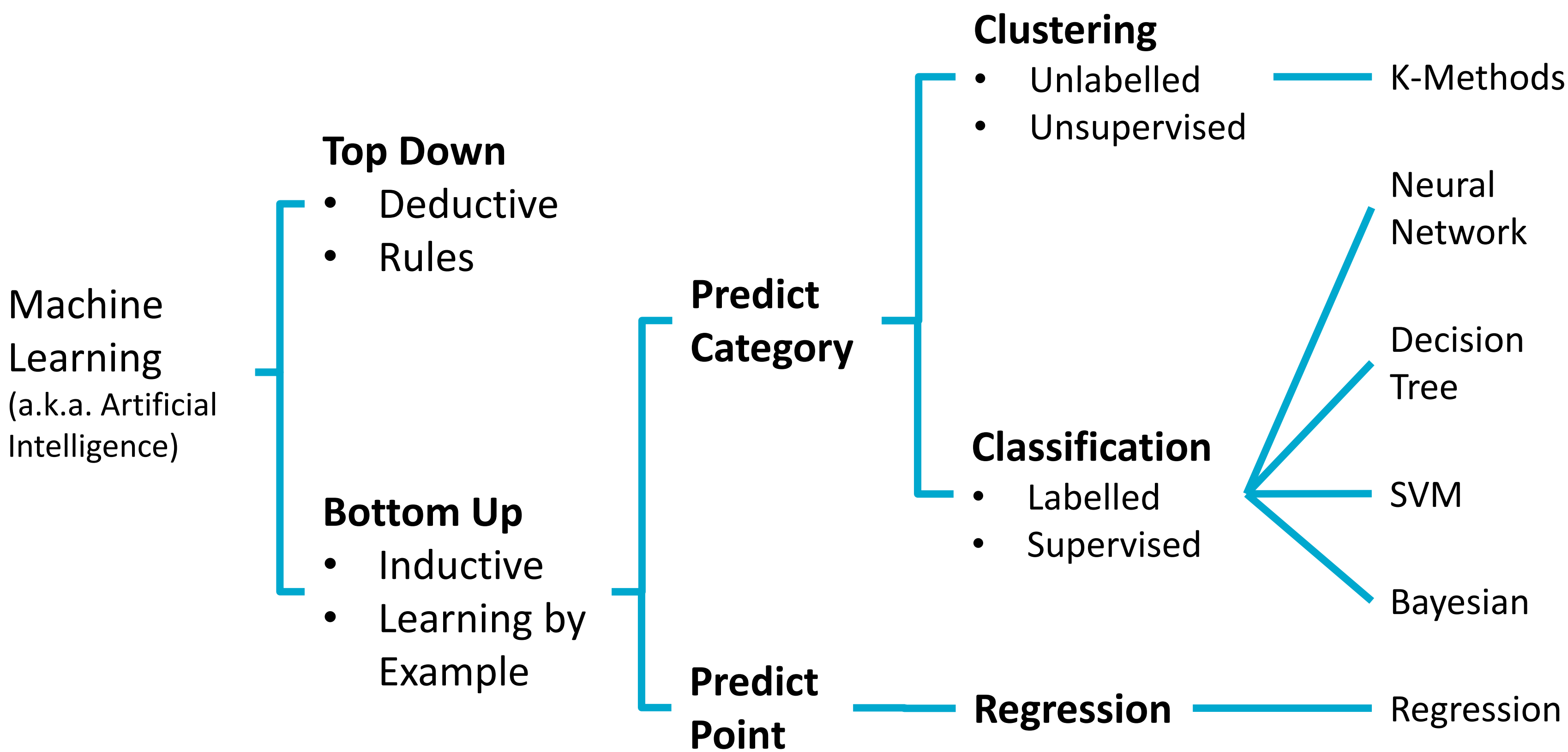


What exactly is machine learning?

What is Machine Learning?

- Machine learning is a field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. (Wikipedia)
  - Data mining
  - Exploratory data analysis
  - Predictive analysis
  - Pattern/anomaly detection

Approaches in Machine Learning



Garbage in, garbage out

- Is the data fit for purpose?
- Is there enough statistical structure to make learning useful?
- Is there scientific reasons to believe the predictions we come up with, or is the computer just hallucinating correlations?

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

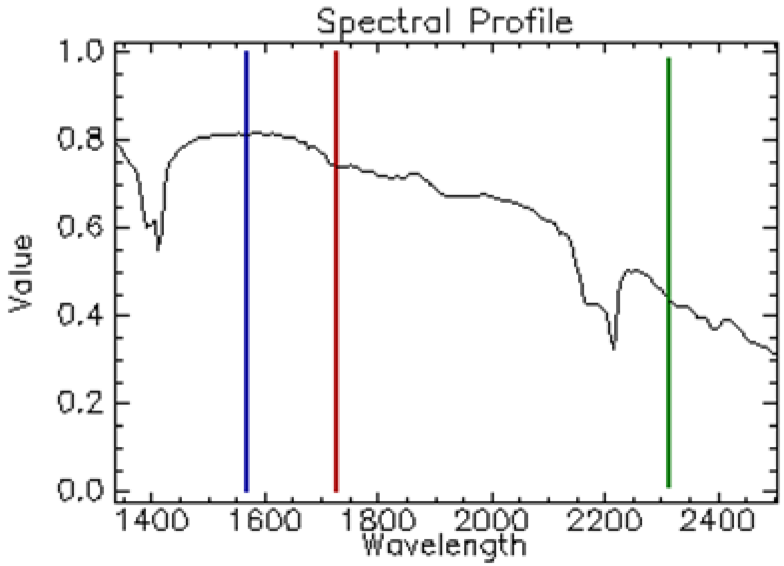
From: Wu & Zhang (2016), arXiv 1611.04135

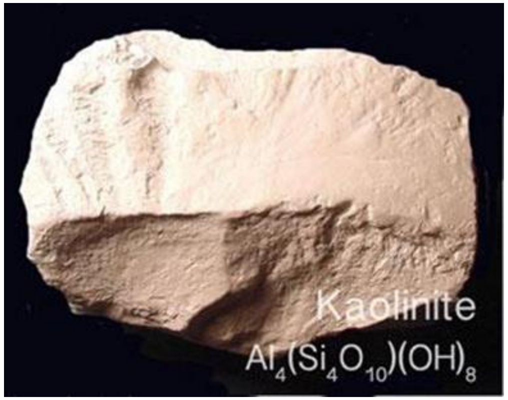
Case Study 1:
Machine Learning

Classifying rock types based on hyperspectral scans of drill cores and geochemical assay data

Case Study 1: Classifying rock based on hyperspectral scans of drill core

- Data: hyperspectral scans of drill core, geochemical assay data.
- Aim: classification of rock types
- Dense dataset (large number of scan lines and point assays)
- Simple model (classification)

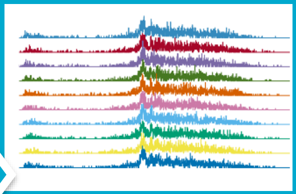


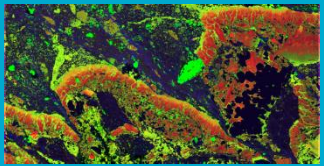


A Rosetta Stone for geochem estimates

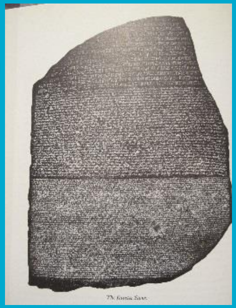
Representative samples

Characterisation





Library of canonical samples + data*



Learning

Estimate

*this can be different data types (e.g. chem vs mineralogy), or same data type at different scales (e.g. XFM vs assay)

New samples/data

Classify
Probabilistically match new sample against canonical samples

Estimate
Generate unknown value + uncertainty

Are we happy with this uncertainty?

NO

Add to representative samples

YES

Job done!

Robertson, Cole, et al.

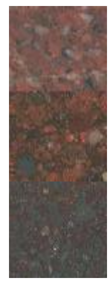
Identifying alteration "geochemically" & "mineralogically"

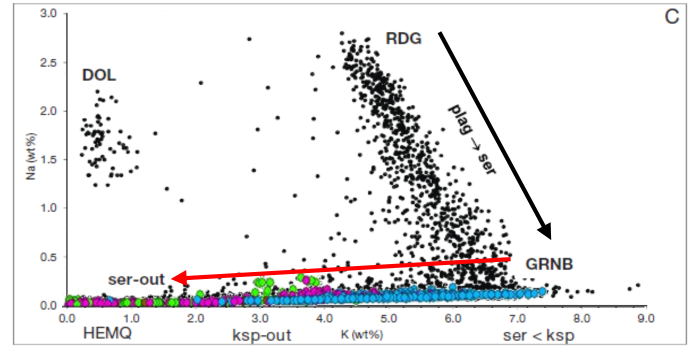
1.1.1 Lithology

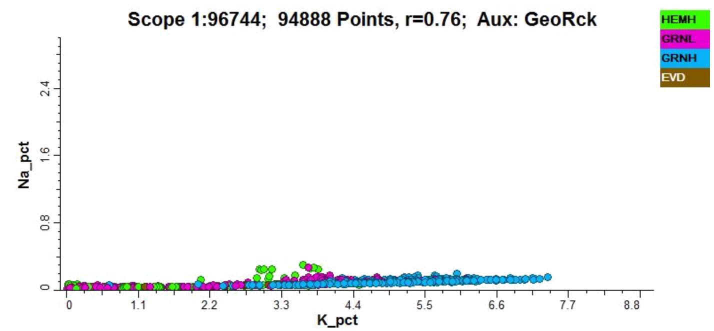
The new legend is based on granite-rich and hematite-rich end members with a continuum between the two based upon the percentage of iron or hematite. This has reduced the number of lithological units from 33 to seven main units and a further four exotic units (Figure 1). The lithological log describes the clast type, abundance, size, shape and alteration. The matrix type, abundance and distribution are also recorded.

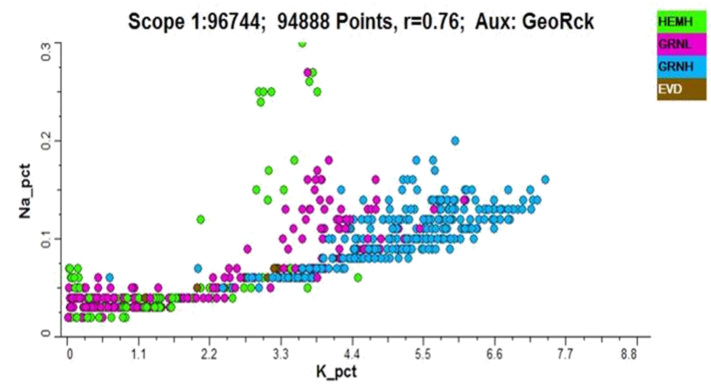
Alias	Criteria	Description
GRN	>90% Granite; <Trace Hem	Undilated, undiluted granite. Crackle Bx or veined.
GRNB	>90% Granite; <10% Hem	Brecciated & Unaltered granite as well as fragmental rocks consisting of unaltered granite-derived components.
GRNH	90-70% Granite; 10-30% Hem	Most commonly breccias but can be generated by replacement.
GRNL	70-40% Granite; 30-60% Hem	Most commonly breccias but can be generated by replacement.
HEMH	40-10% Granite; 60-90% Hem	Most commonly breccias but can be generated by replacement.
HEM	<10% Granite; >90% Hem	Textured or massive; breccias, precipitates, metasomatites.
HEMQ	No Granite; >90% Hem	"Classic" hematite-quartz bx. Must be barren, locally vuggy, porous or silicified. Usually associated with barite. No sulphide, sericite or fluorite.
GRNV	Granite > Volc components	Bx containing granite + um-m volc clasts
HEMV	Hem > Volcanic components	Bx containing Hem + um-m volc clasts.
KASH	Mixed ash/epiclastics	Mixed interbedded epiclastics rocks & volcanic ash
EVD	>90% Volcanic components	Generic dyke; volcanic/sub-volcanic textures. Often chlorite or hematite altered.

Figure 1: ODO Lithological Legend









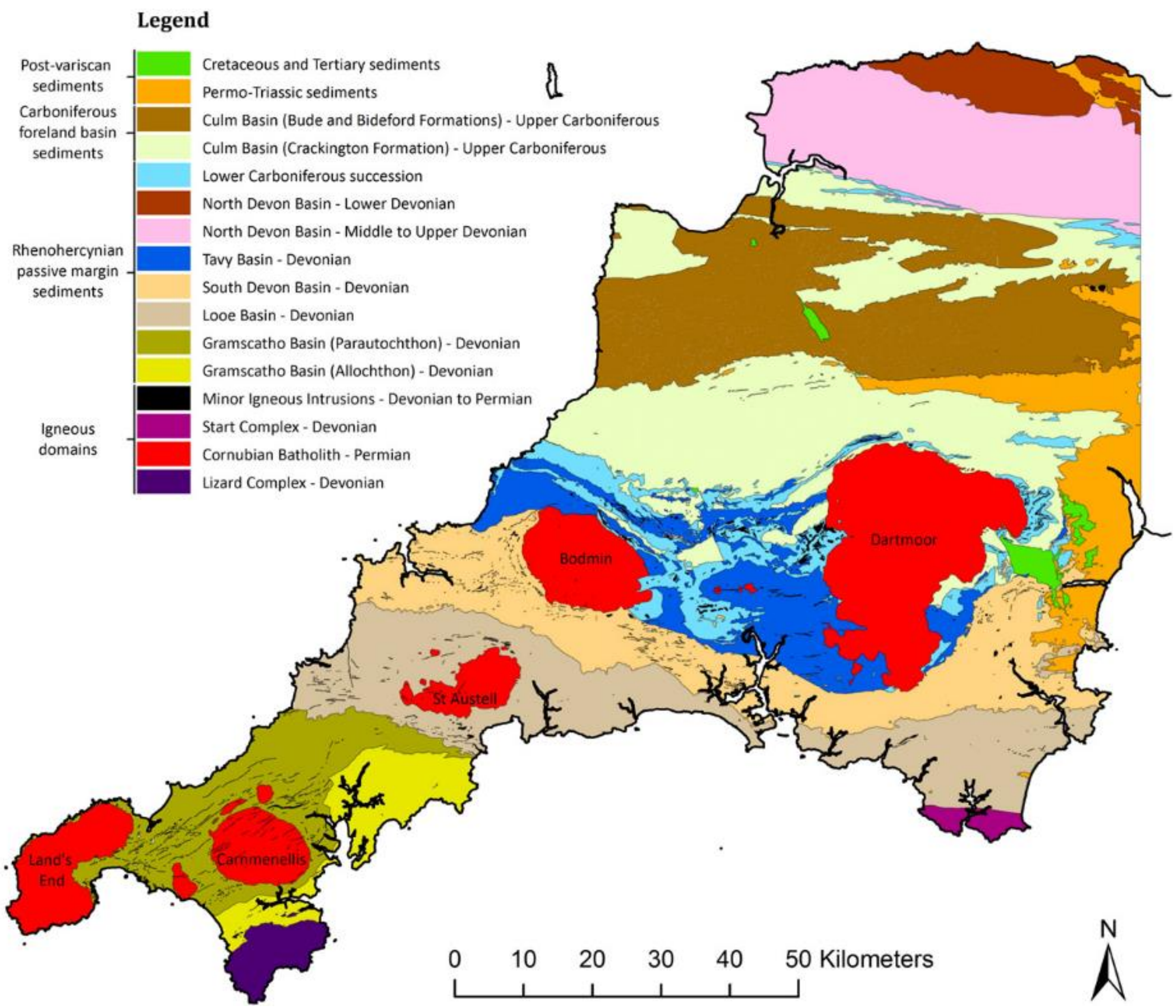
Case Study 2:
Geostats vs. Machine Learning

Estimating soil geochemistry by Kriging and by Random Forest

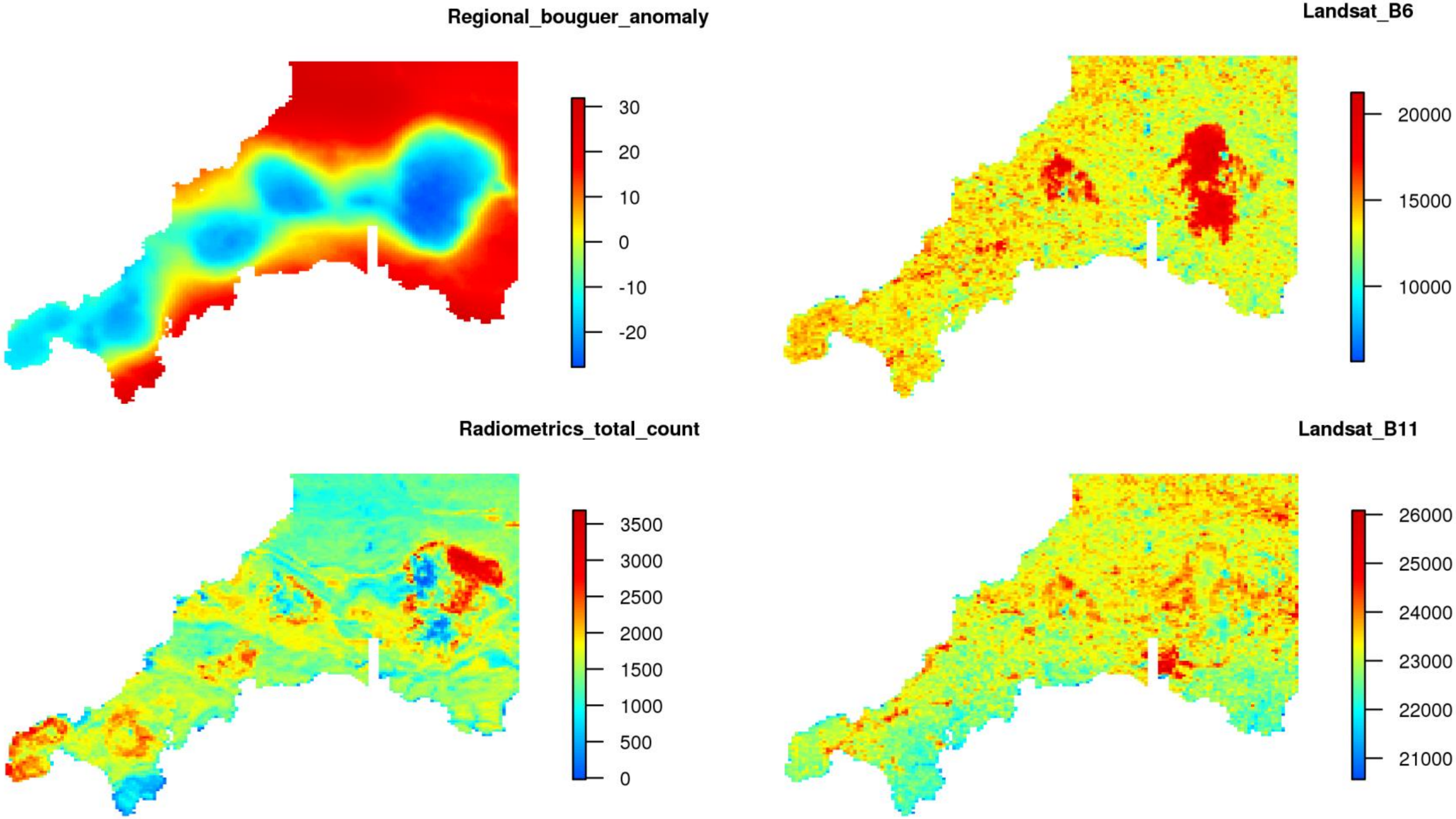
Case Study 2: Soil Geochemistry

- Data: soil geochemistry of southwest England (source: C. Kirkwood, BGS G-BASE)
- Elements used in this study: Al, Ba, Br, Ca, Ce, Co, Cr, Cs, Fe, Ga, Ge, Hf, K, La, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Rb, Sc, Se, Si, Sm, Sr, Ta, Th, Ti, U, V, Y, Zr
- Other elements were excluded due to their hydrothermal mobility or concentrations below detection limits.
- Auxiliary data: Gravity, geomorphology, radiometrics, IR
- Complex model (assumptions about mobility of elements, spatial autocorrelation)
- Sparse data
- Aim: geochemical exploration (outliers)

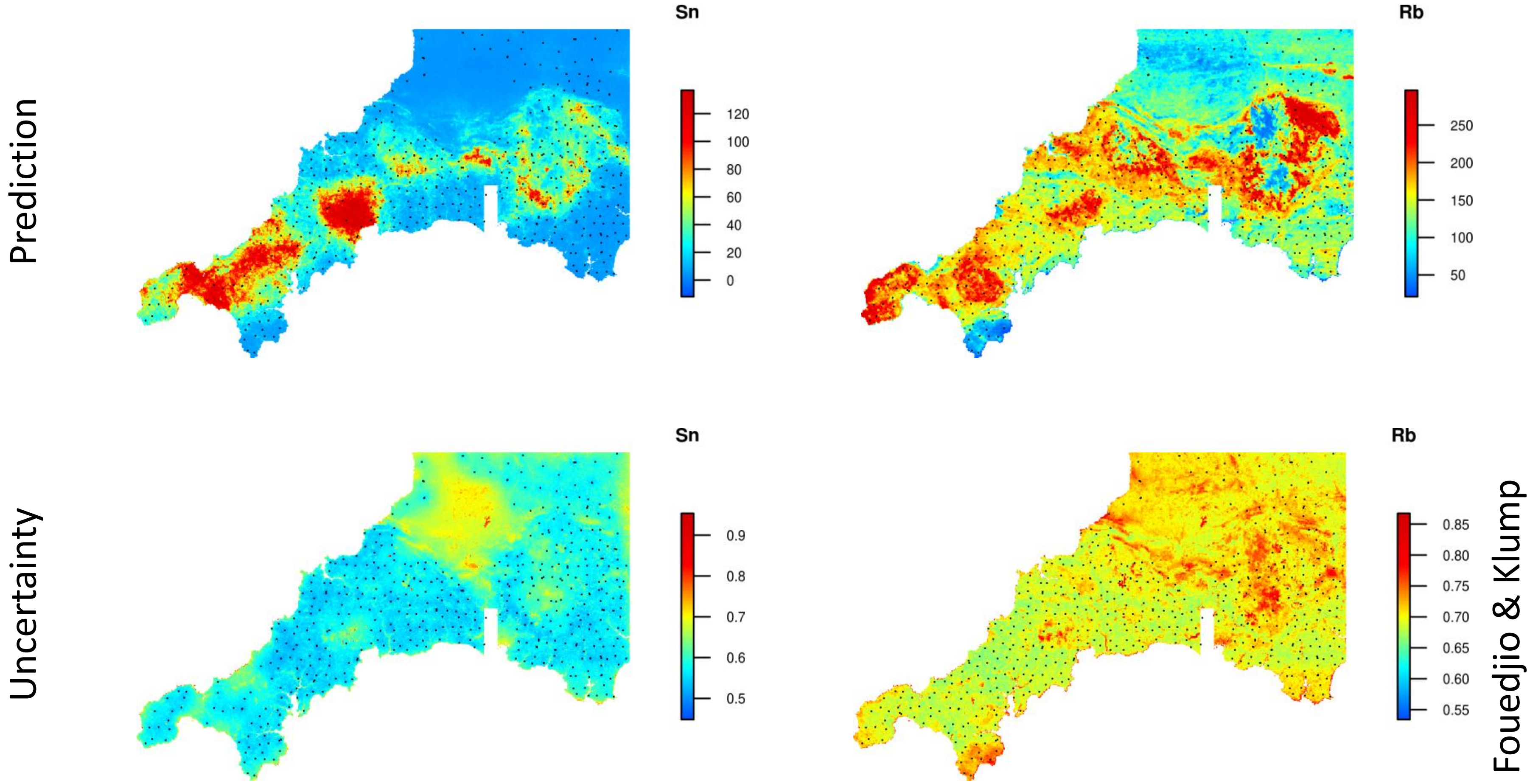
Study Area


Legend
Post-variscan sediments
Carboniferous foreland basin sediments
Rhenohercynian passive margin sediments
Igneous domains
Cretaceous and Tertiary sediments
Permo-Triassic sediments
Culm Basin (Bude and Bideford Formations) - Upper Carboniferous
Culm Basin (Crackington Formation) - Upper Carboniferous
Lower Carboniferous succession
North Devon Basin - Lower Devonian
North Devon Basin - Middle to Upper Devonian
Tavy Basin - Devonian
South Devon Basin - Devonian
Looe Basin - Devonian
Gramscatho Basin (Parautochthon) - Devonian
Gramscatho Basin (Allochthon) - Devonian
Minor Igneous Intrusions - Devonian to Permian
Start Complex - Devonian
Cornubian Batholith - Permian
Lizard Complex - Devonian
Dartmoor
Bodmin
St Austell
Carnmenellis
Land's End
0 10 20 30 40 50 Kilometers
N

Auxiliary Variables

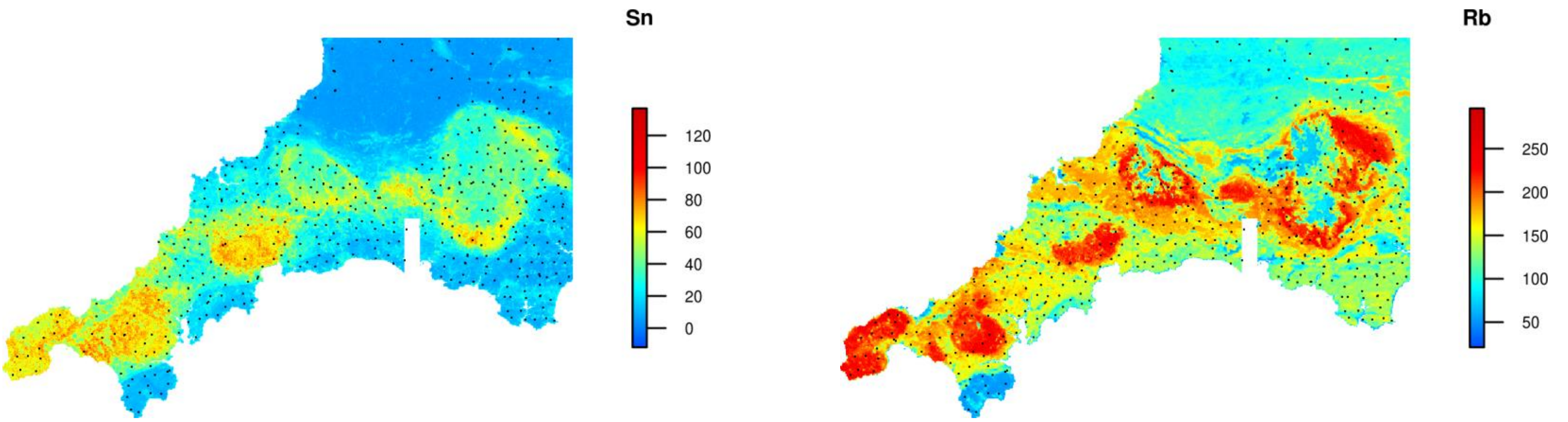


Soil Geochemistry - Kriging

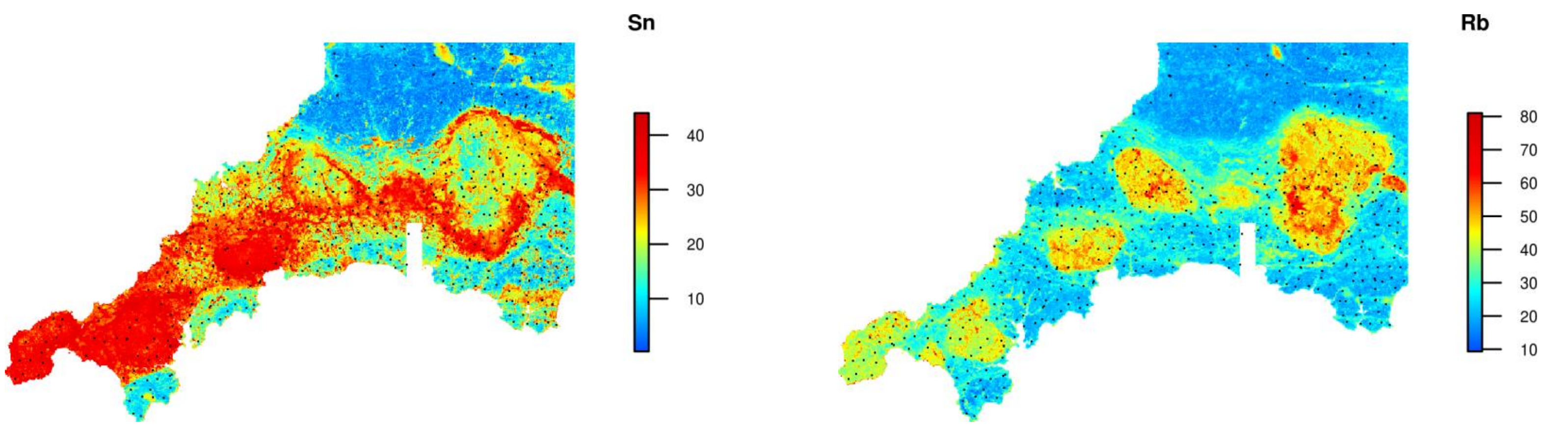


Soil Geochemistry – Random Forest

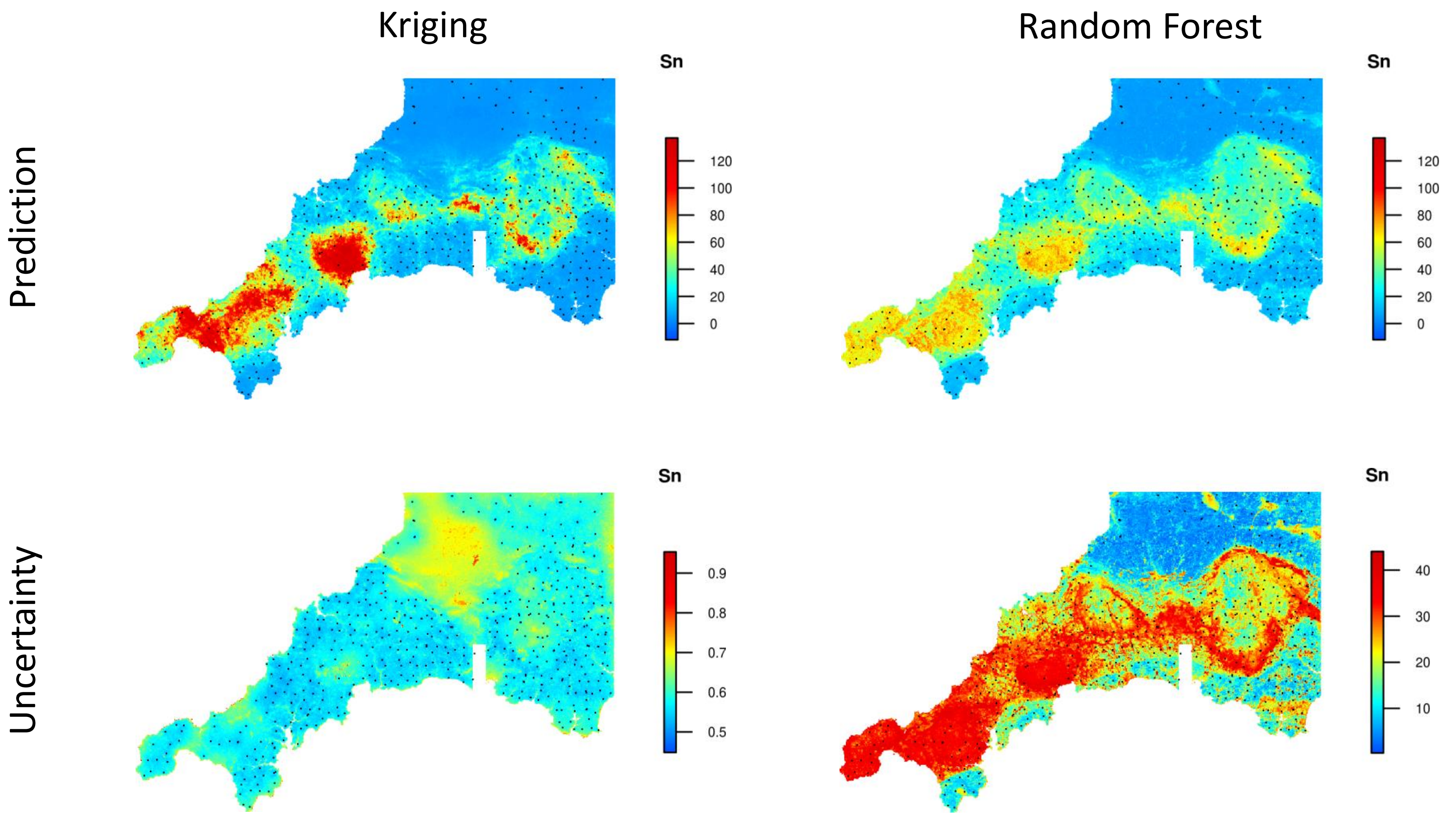
Prediction



Uncertainty



Prediction – Kriging vs. Random Forest



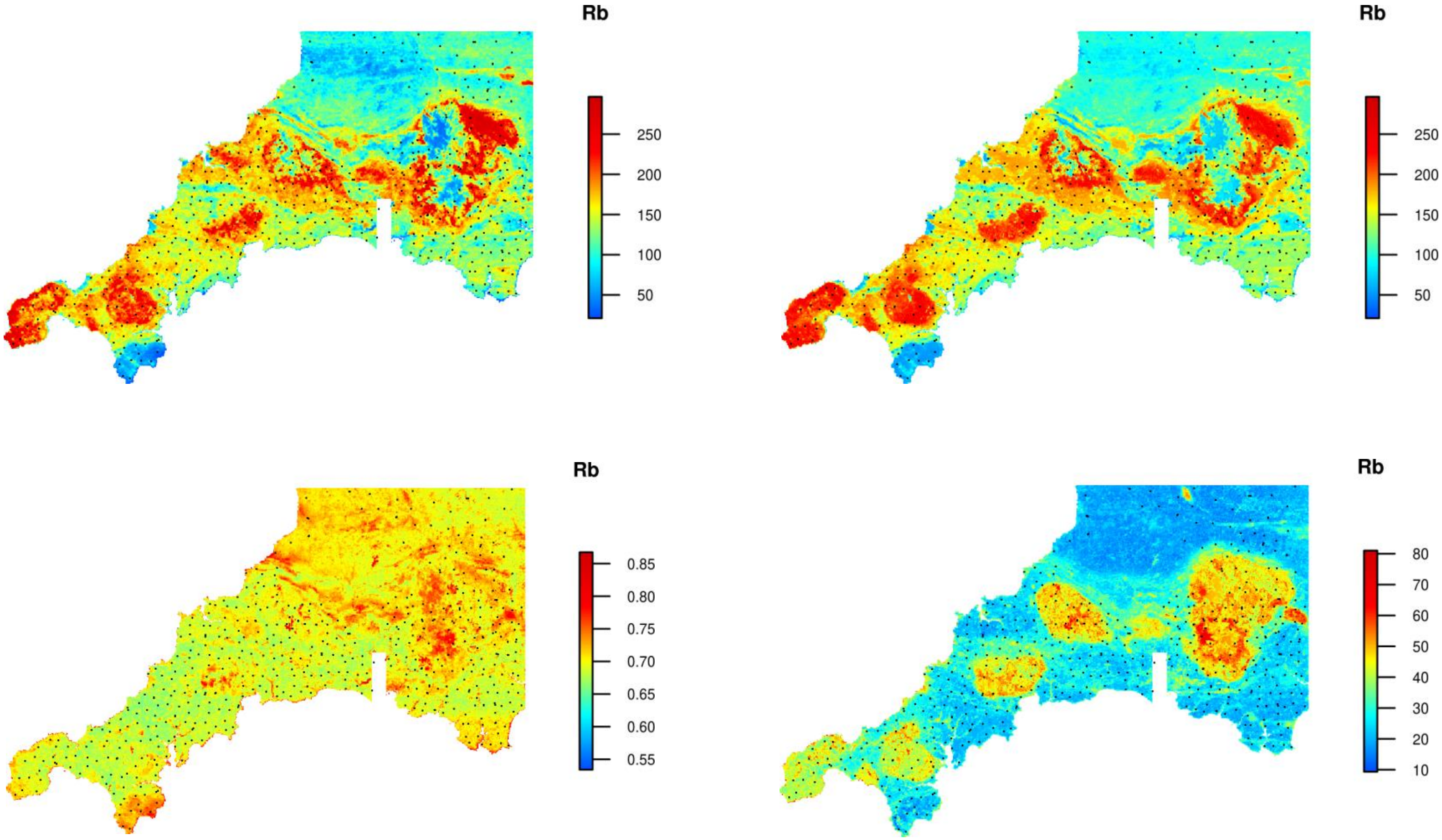
Prediction – Kriging vs. Random Forest

Kriging

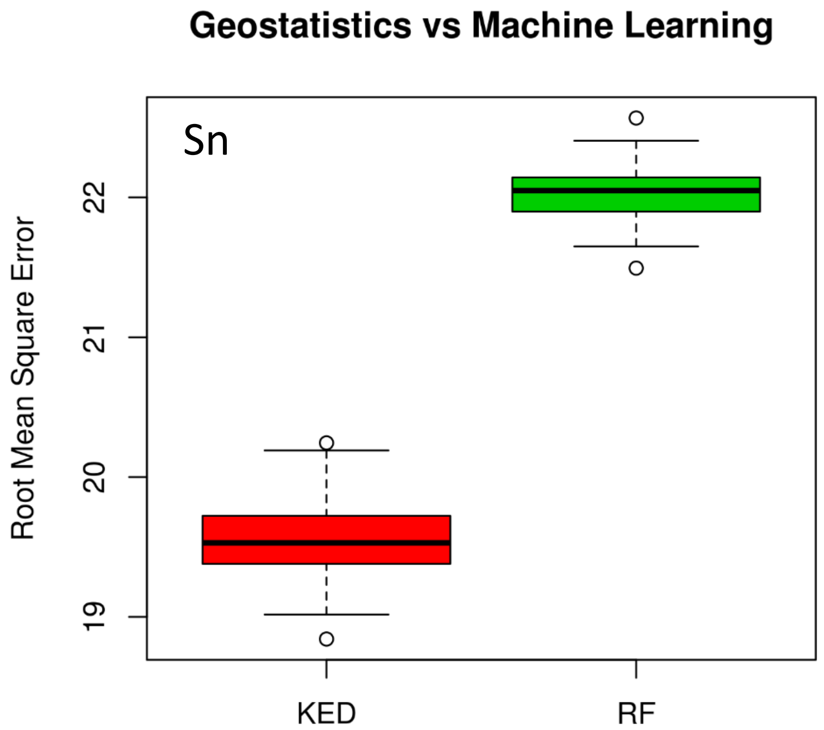
Random Forest

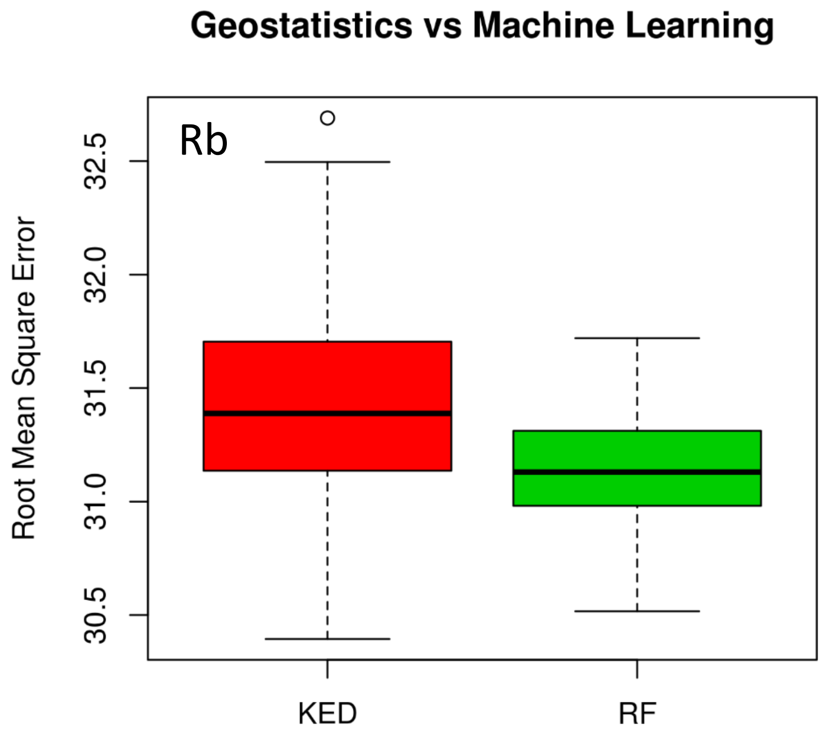
Prediction

Uncertainty



Error estimates

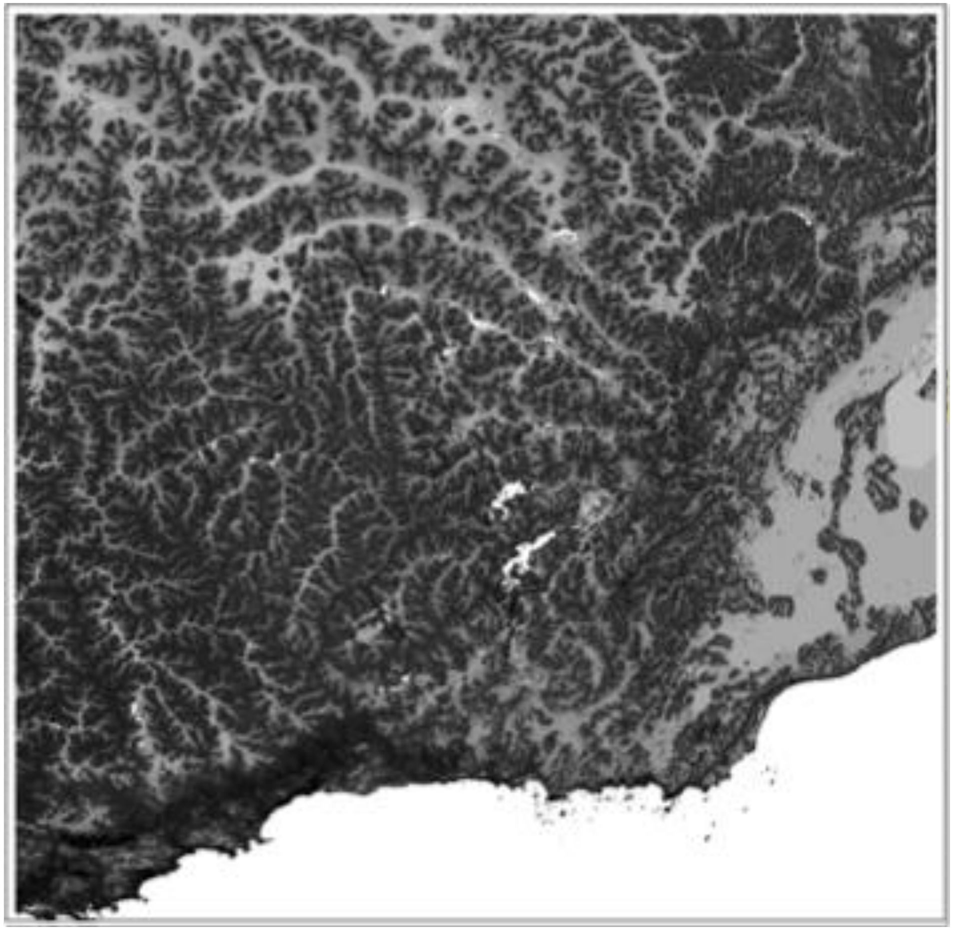




Case Study 3:
Landscape Classification

Classifying Landscape Types from Digital Elevation Data

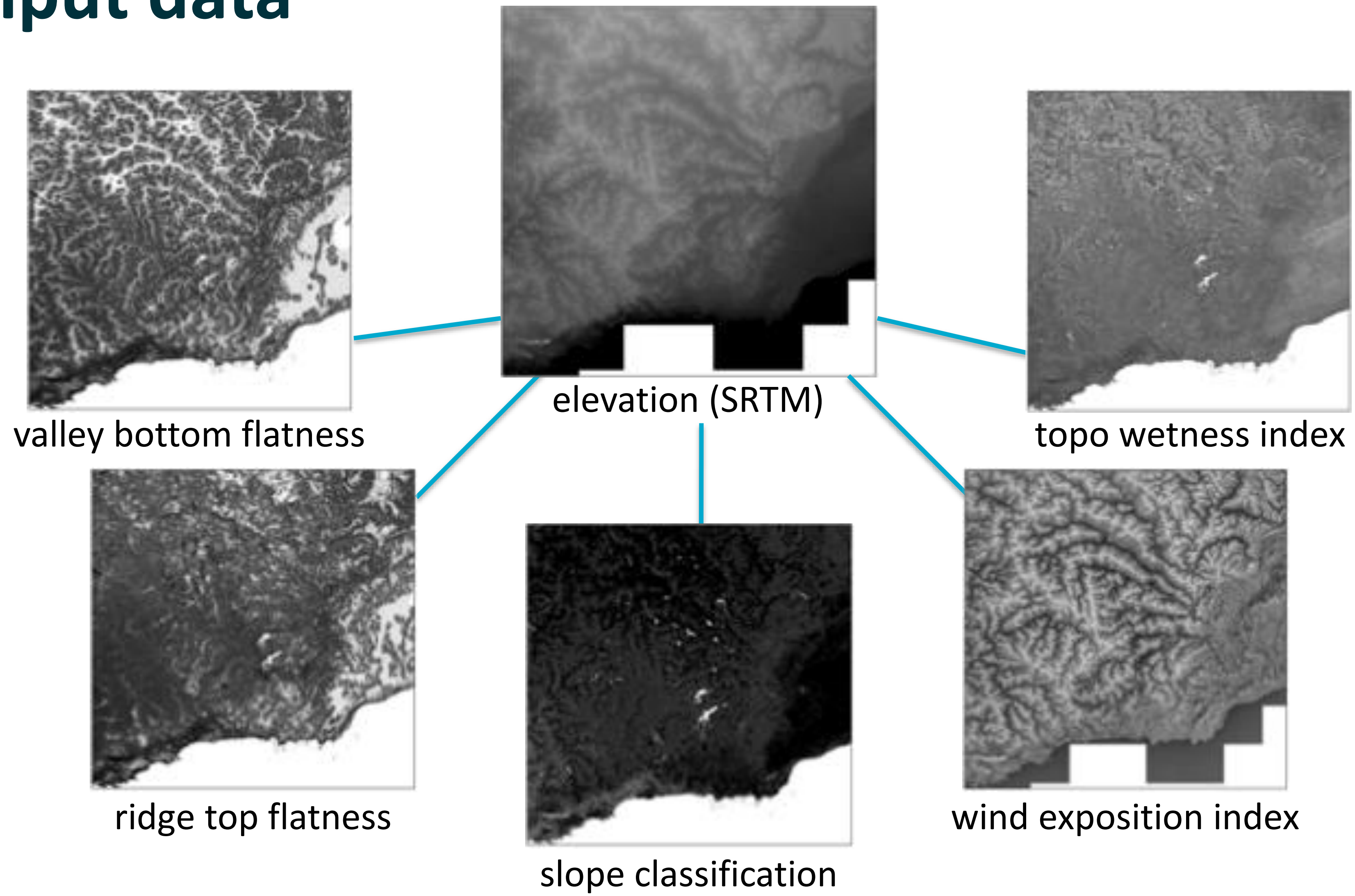
Landscapes as labeled by geologist



flatness map

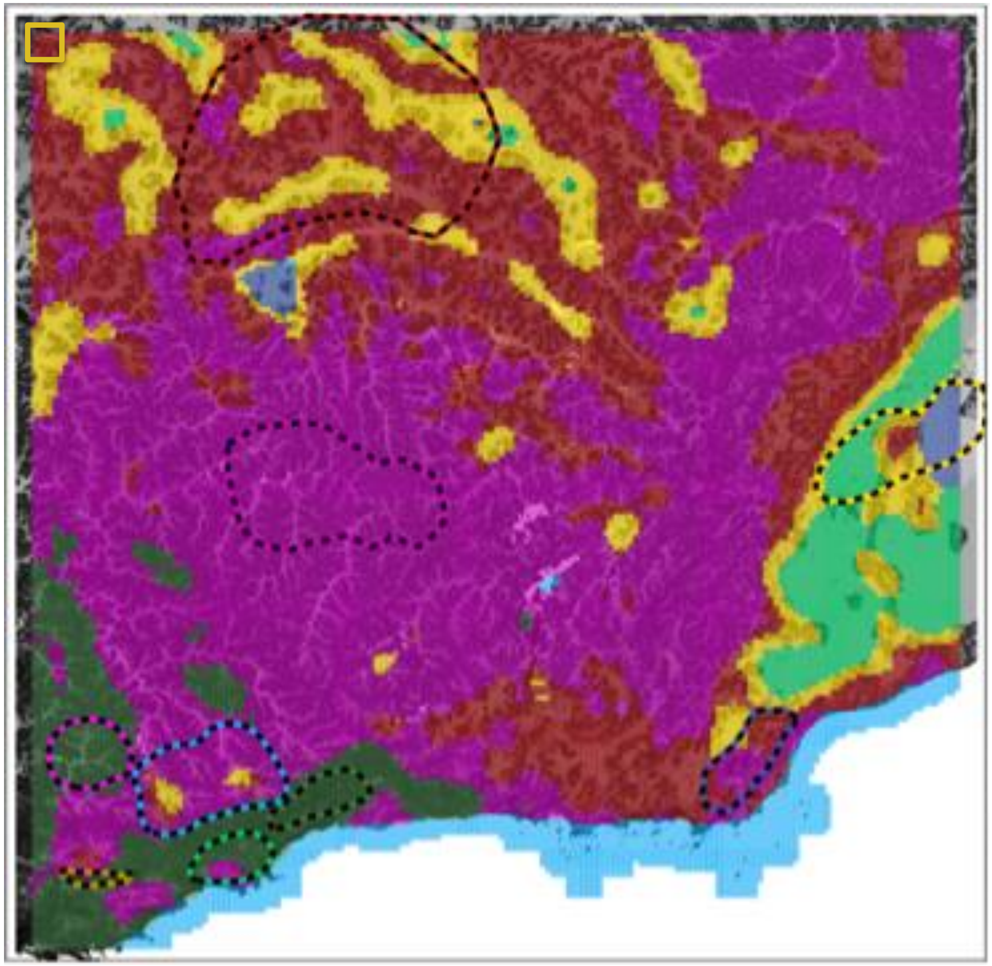
- SRTM covers most of the globe, 1” (30 m) data freely available, many users

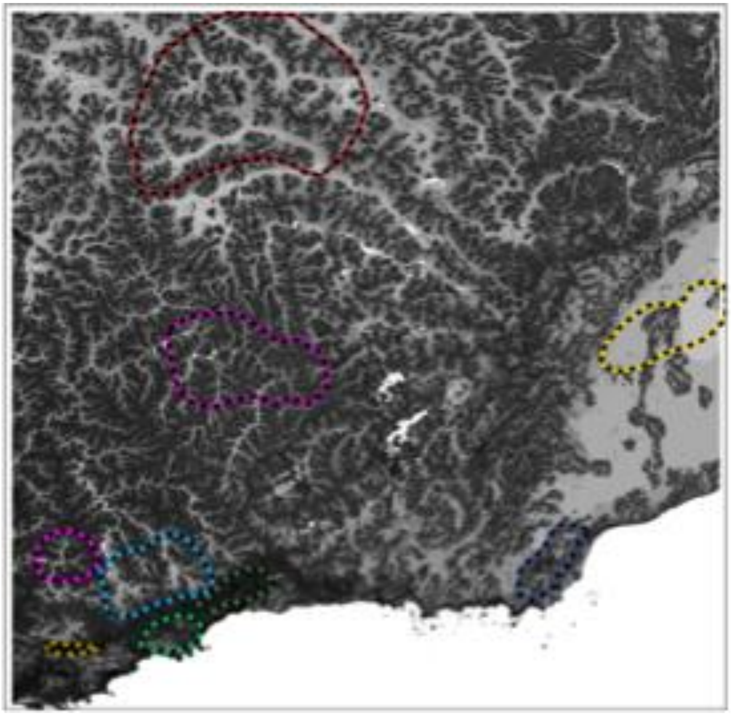
Input data



elevation (SRTM)
valley bottom flatness
topo wetness index
ridge top flatness
slope classification
wind exposition index

unsupervised (k-means 7 classes), flatness

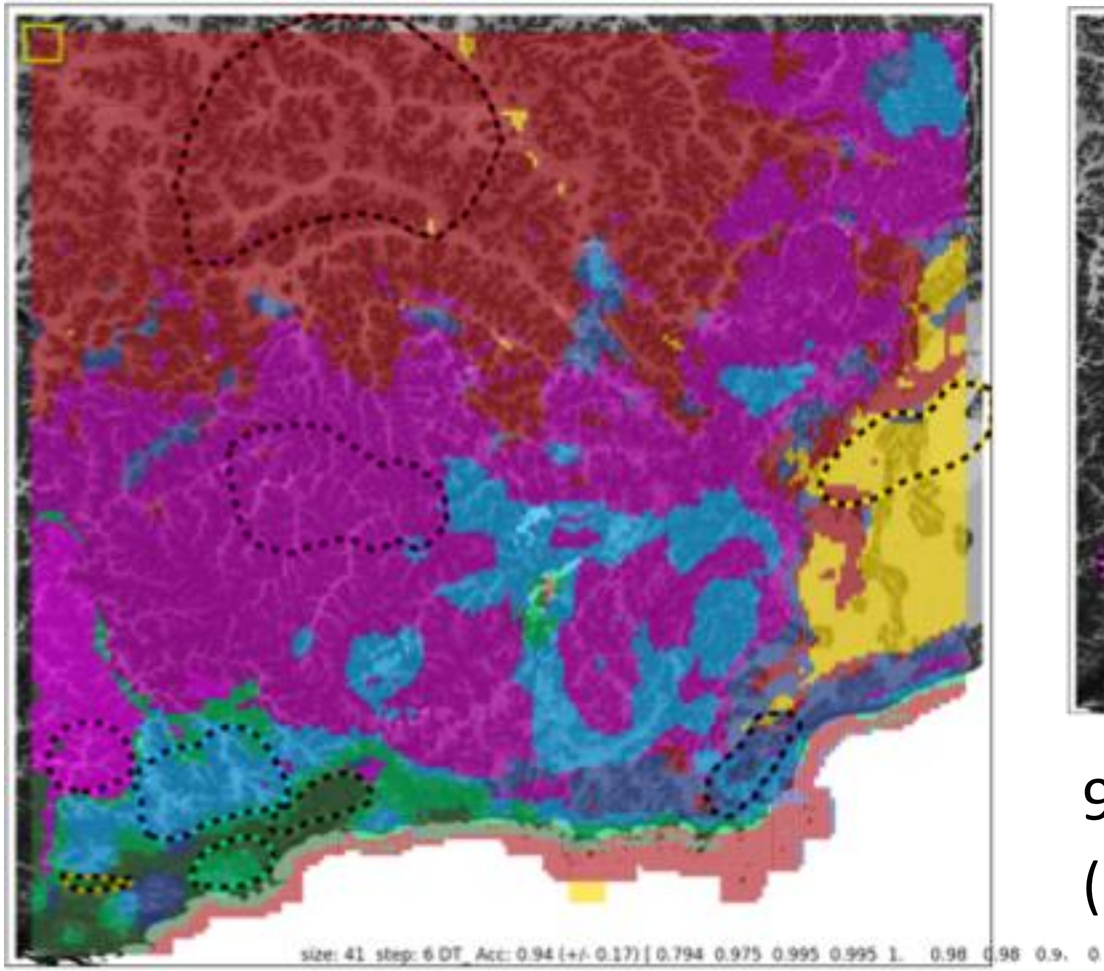


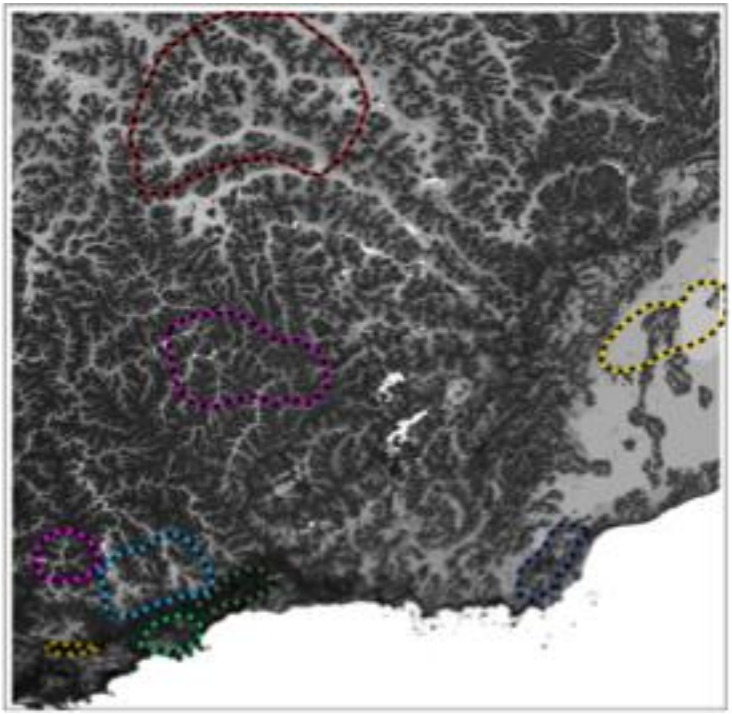


84% accuracy
(10-fold cross-validation)

Albrecht, Gonzalez-Alvarez, Klump & Smith

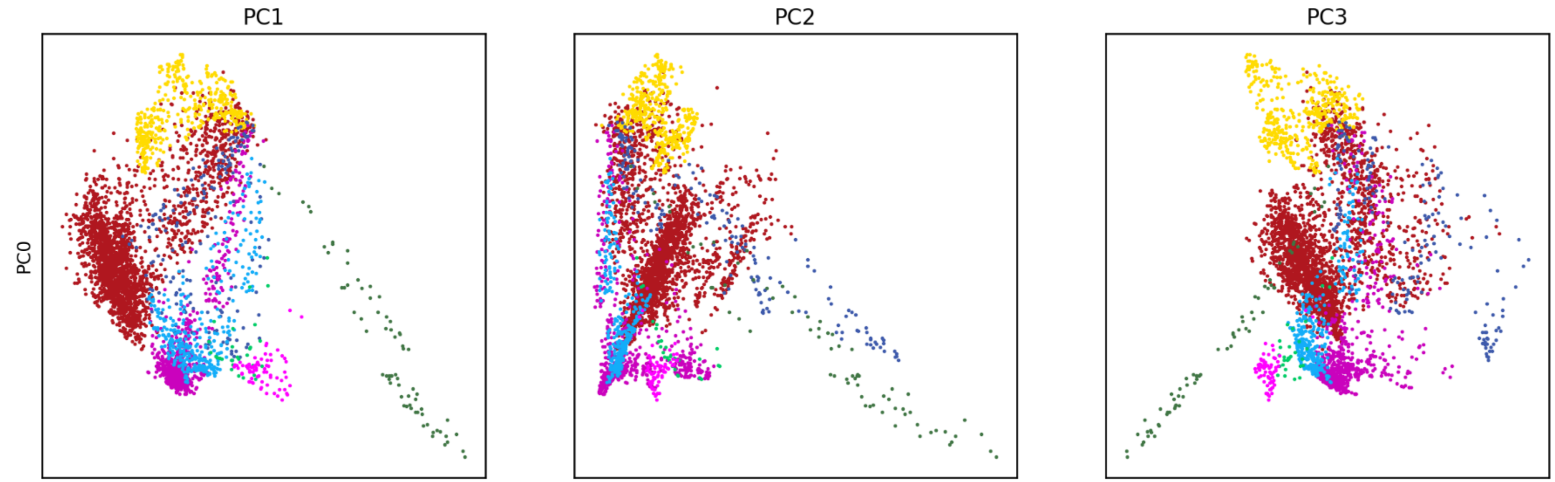
supervised (decision tree), all features

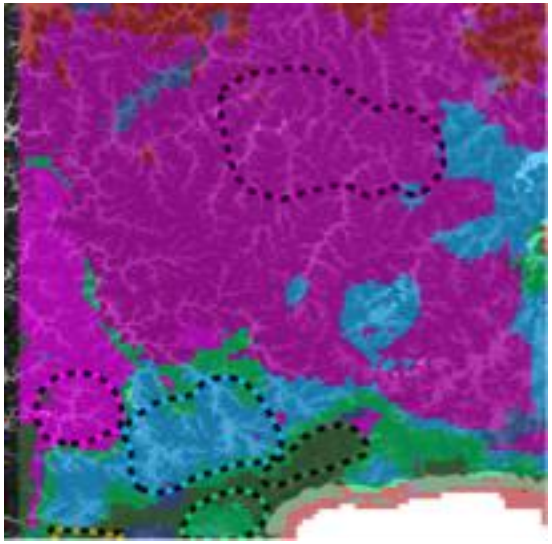


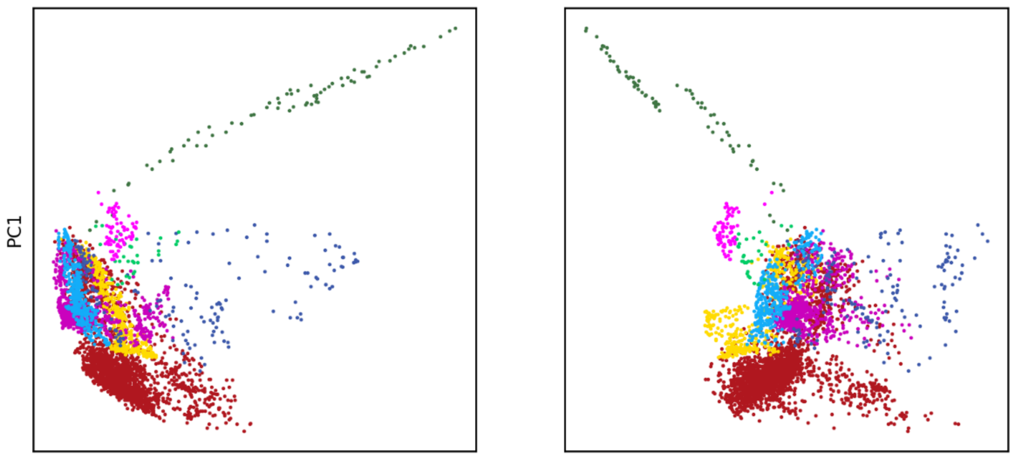


94% accuracy
(10-fold cross-validation)

Principal component analysis







Conclusions:
Lessons Learned

Can we teach geology to machines?

What is the right tool?

High density data		Sparse data
High dimensional data		Low dimensional data
	Question	
Patterns		Outliers
Simple causality No assumptions		Complex causality Model driven
Engineering solution		Analytical solution
	Method	
Machine Learning		**Geostats**

Can we teach machines geology?

- Machine learning is a powerful and versatile tool, but not every problem is a nail.
- Geology often is data sparse, limiting the application of machine learning methods.
- Data rich applications in geology are commonly hyperspectral and potential field methods.
- More research is needed to understand machine learning in a spatial context.
- At present, we do not understand the meaning of uncertainty reported my machine learning in a spatial context. This is problematic in decision making.
- Watch this space!

Thank you

Mineral Resources
Jens Klump
Team Leader Geoscience Analytics

t +61 8 6436 8828
e jens.klump@csiro.au
w people.csiro.au/Jens-Klump