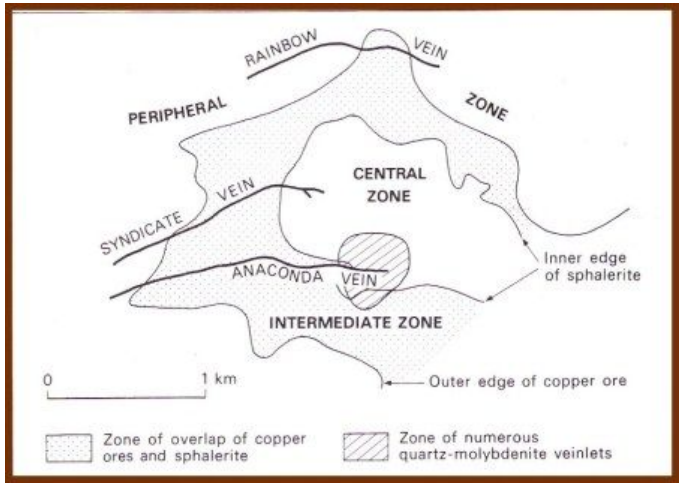
Mining Rock Chemistry Data

Because if anyone can mine data, a geologist can.

Overview

The chemistry of magma and the fluid it gives off as it cools are the most important source of metals in ore deposits.

Ore deposits have changed drastically over Earth's history. Certain deposit types are limited to the Archean or Proterozoic. Can we get age labels (eons) from bulk rock chemistry data for prospecting purposes?



Phanerozoic (0.55 Ga to present) mineralized halos around igneous intrusion (Butte, MT)

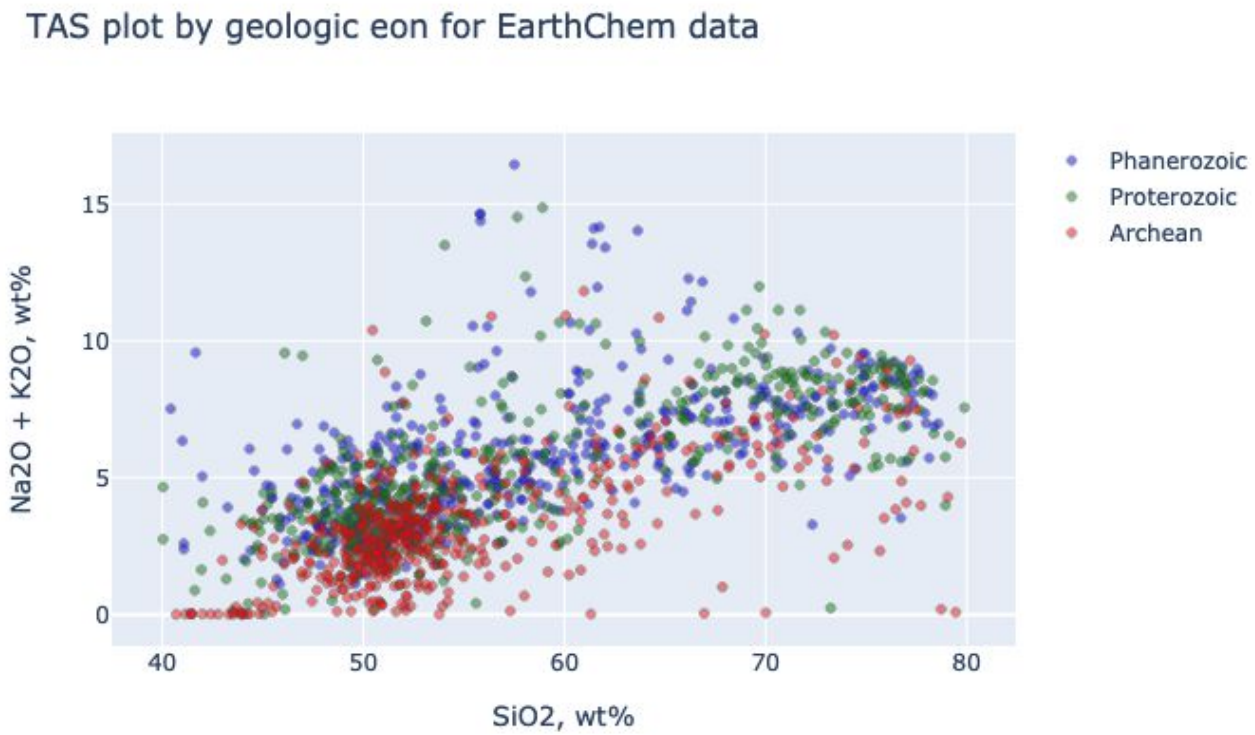
Map credit: Evans, A. M. An Introduction to Ore Geology, Elsevier, NY (1980) p. 148-150.

Archean (4-2.5 Ga) gold/uranium conglomerate (Witwatersrand, South Africa)

Photo credit: INTERNATIONAL GEOLOGICAL CONGRESS

The Data:

- EarthChem database
- ~160,000 usable rock compositions (major elements, lat & long, age)
- Feature engineering: manual.
  - It’s all about the ratios.
  - Except when it’s not.



TAS diagram for rock classification and preliminary data visualization

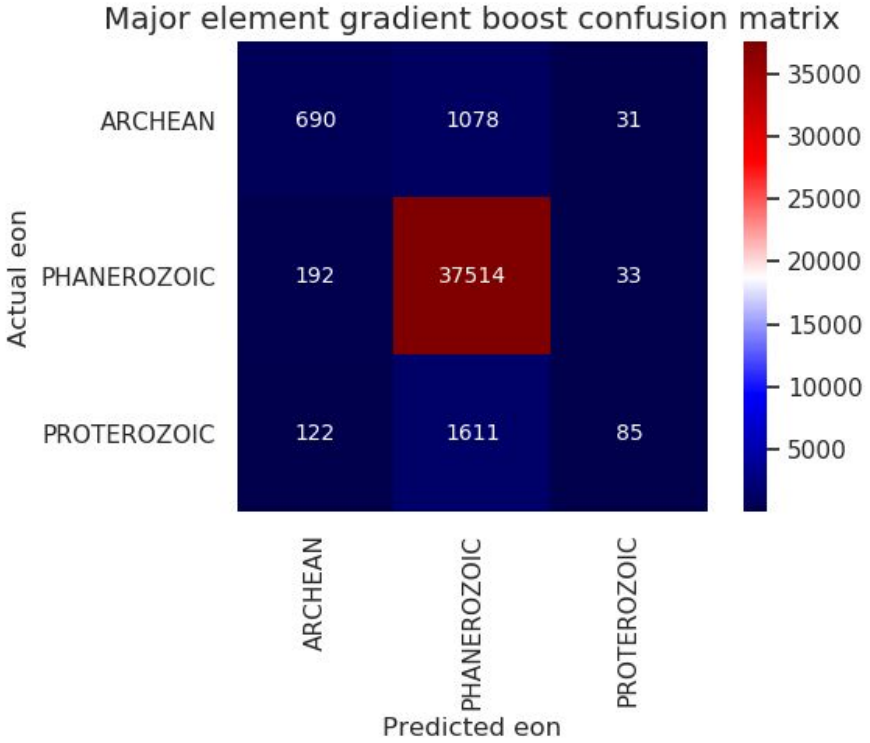
The Analysis 01

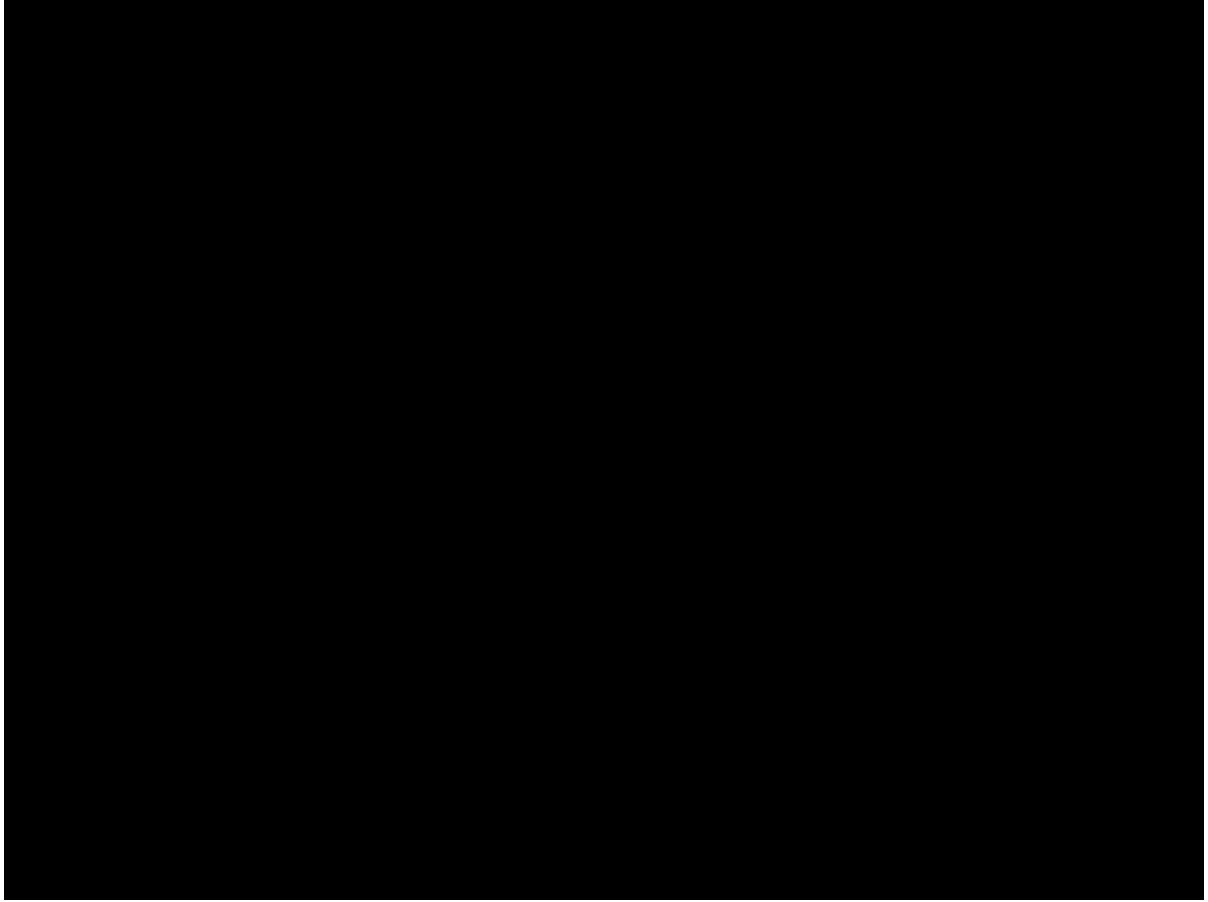
Tools used:

- Logistic regression + regularization
- Linear support vector classifiers
- Decision trees
- Random forest classifiers
- Gradient boosting forests

Techniques & metrics:

- Confusion matrices
- ROC curves (have to convert to binary problem)
- Accuracy, precision, and recall all have their place





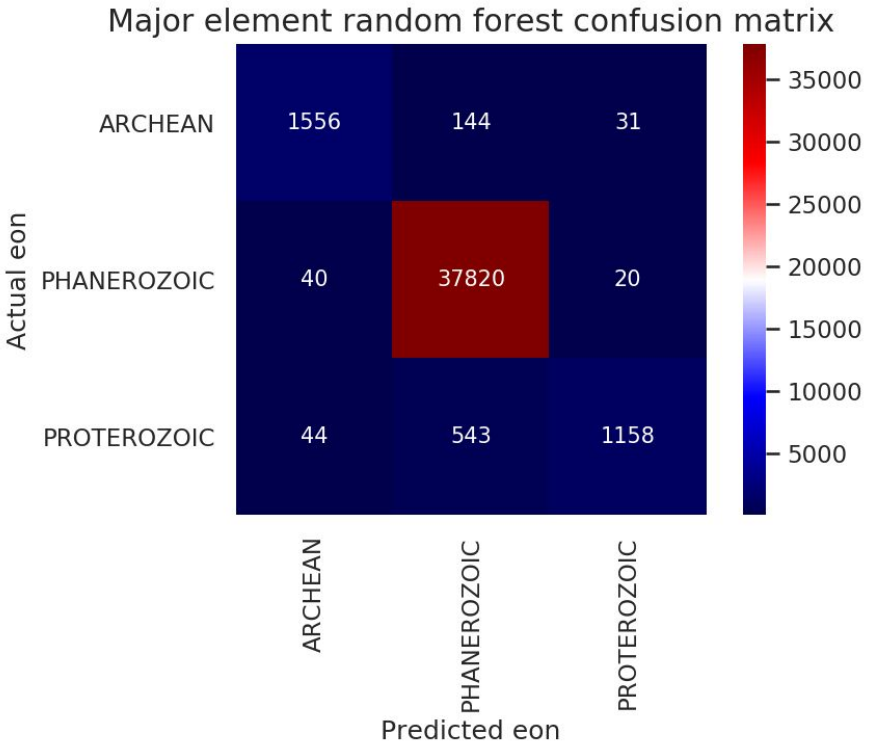
Three major element features most influential in gradient-boosting forest classifier model

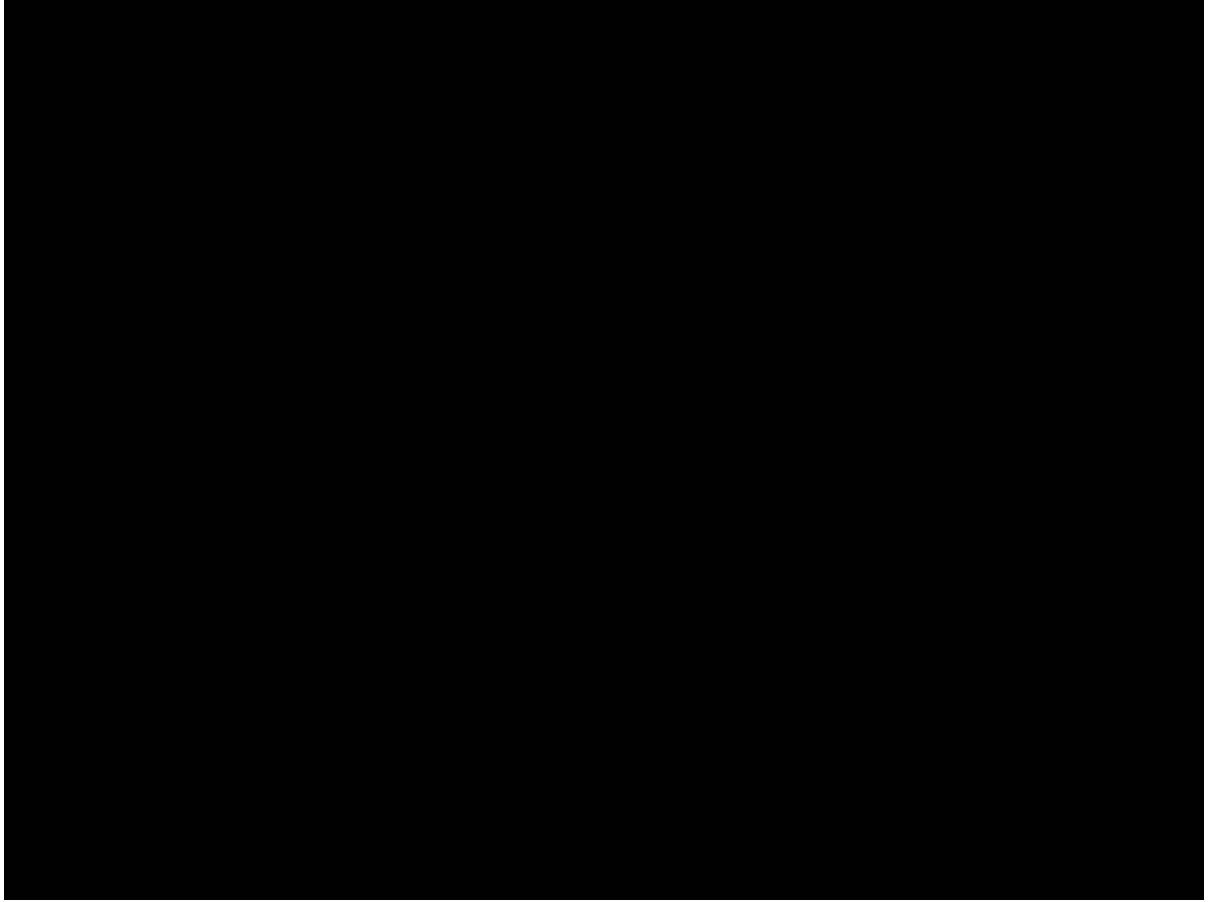
The Analysis 02

You can break this problem wide open*
...if you make latitude and longitude features in your model.

Is that cheating?

* - and by break it wide open, I mean get the recalls all above 0.5



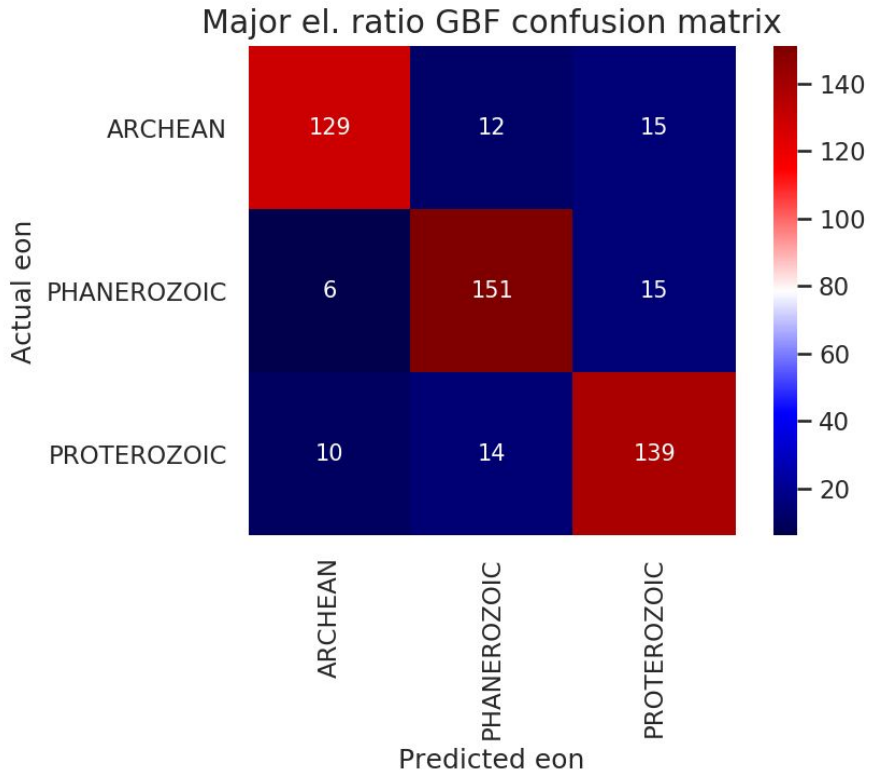


Region in focus: Southern Africa

The Analysis 03

The Southern Africa data sure do clean up nice, don't they?

- Balanced dataset (1,000 analyses per eon)
- Even recall > 0.92…
- Model shown: just five ratio features
- World domination may be too much to ask, but regional hegemony is within our grasp.

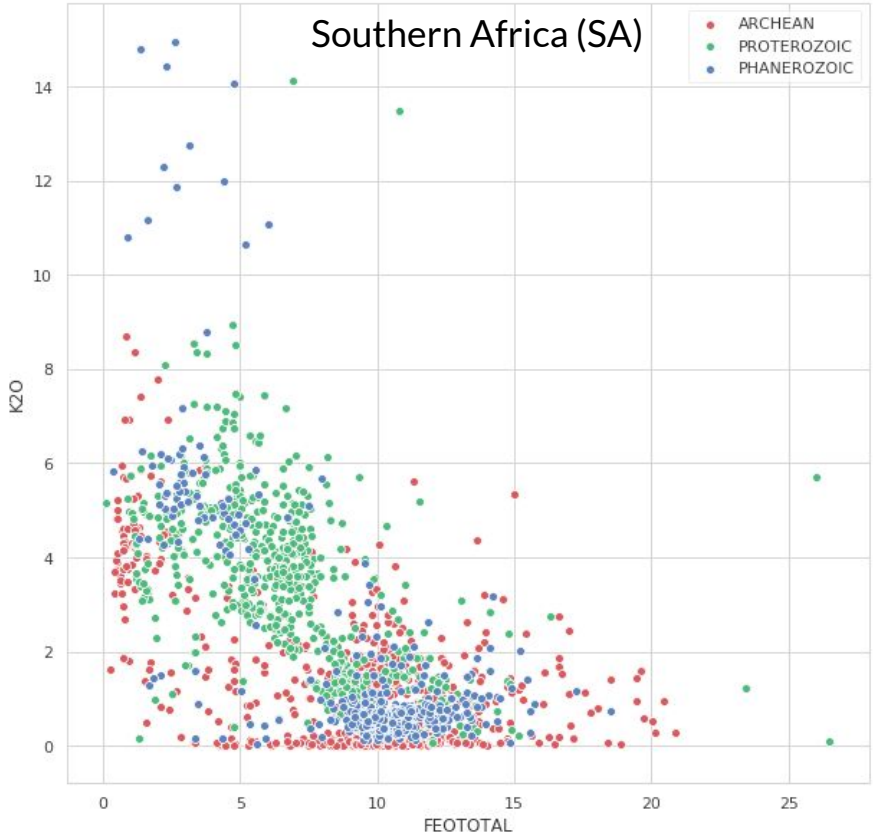


Prospecting - Use Cases & Model Optimization

- Tune models on a regional basis.
- Regions that are still only partly explored (SE Asia, Africa, etc.)
- High recall - make sure that no good sites escape notice.
- High precision - make sure we're only investigating promising sites. Exploration is really expensive.

If Only There Were Time

- Refinement of probability thresholds to adjust precision & especially recall on models for specific use cases.
- Undersampling Phanerozoic compositions when constructing a new generation of models.
- More data visualization for 1 on 1 relationships between variables.
- Clustering analysis to take a new look at the VERY old, VERY divisive problem of rock classification.
- Use this data to construct a tutorial series in data science for geology students.

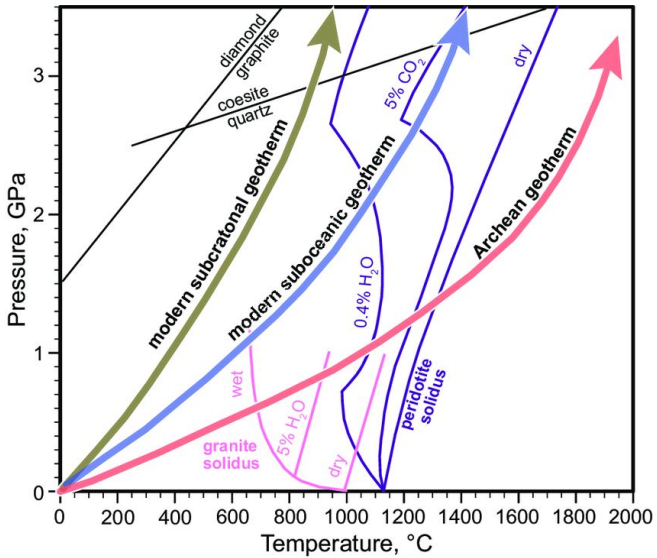


Thank you.

How has Earth changed over time?

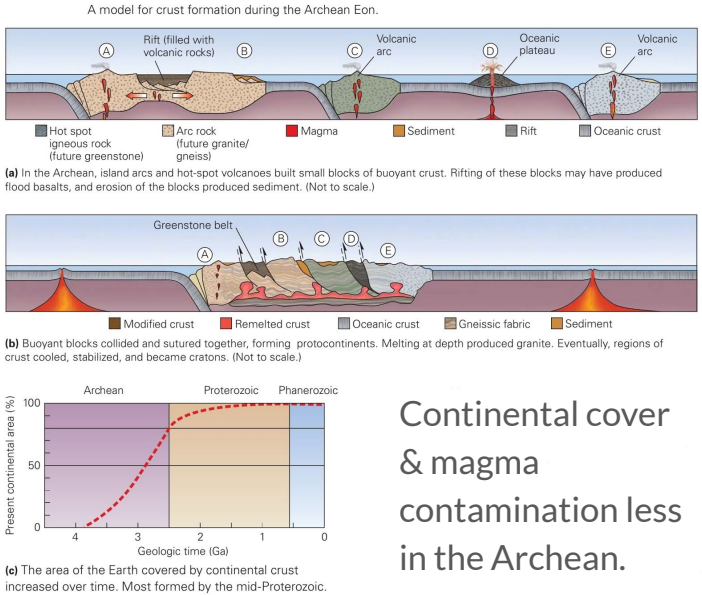
1

The Earth was hotter in the past and melted more easily. Perhaps (perhaps) melt fractions were larger, with a smaller proportion of “flux” elements (Na, K, P).



https://www.researchgate.net/profile/Tatsuki_Tsujimori

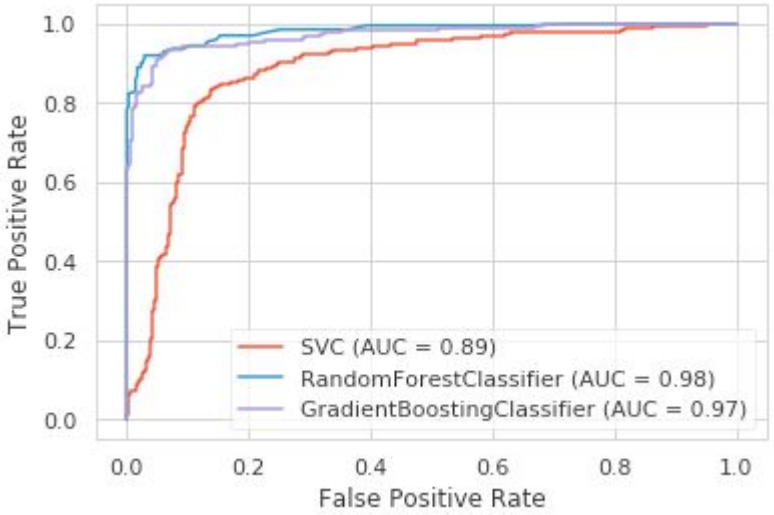
2

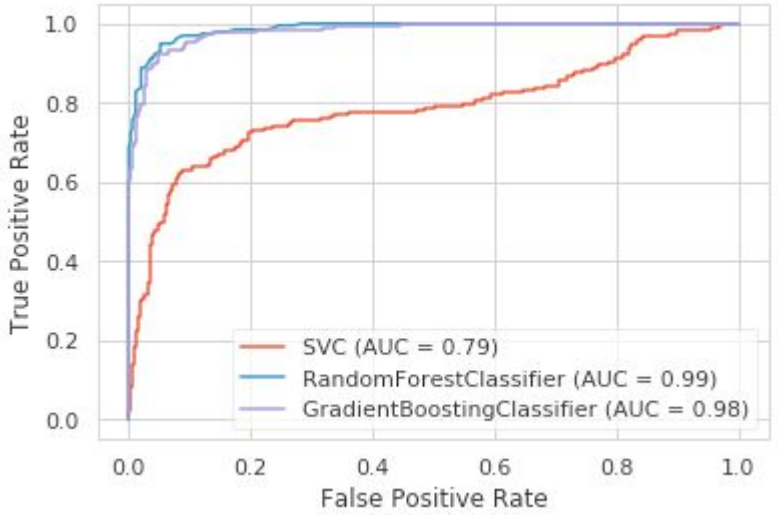


Continental cover & magma contamination less in the Archean.

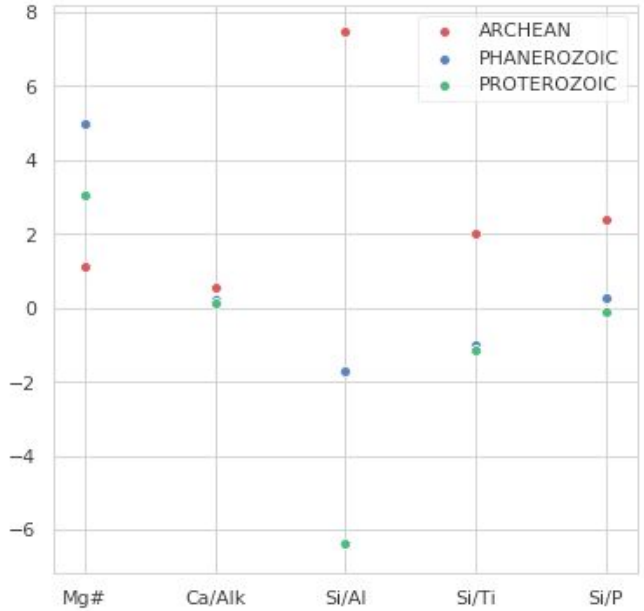
http://geologylearn.blogspot.com

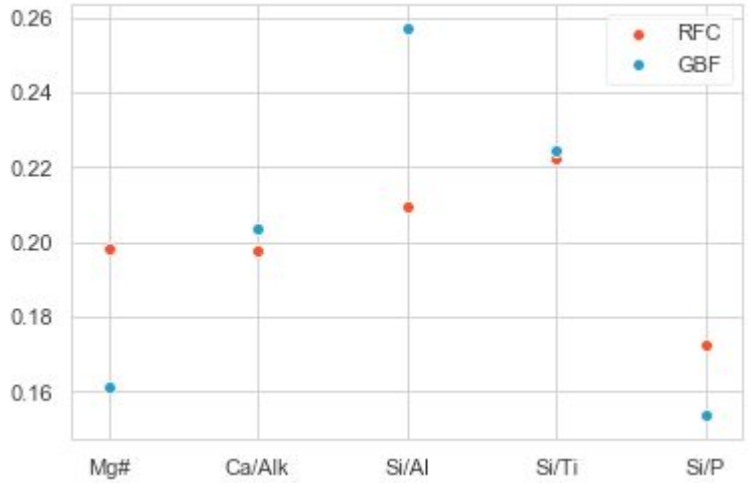
ROC Curves (SA)-Phan / PreC, Arch / ColdEarth





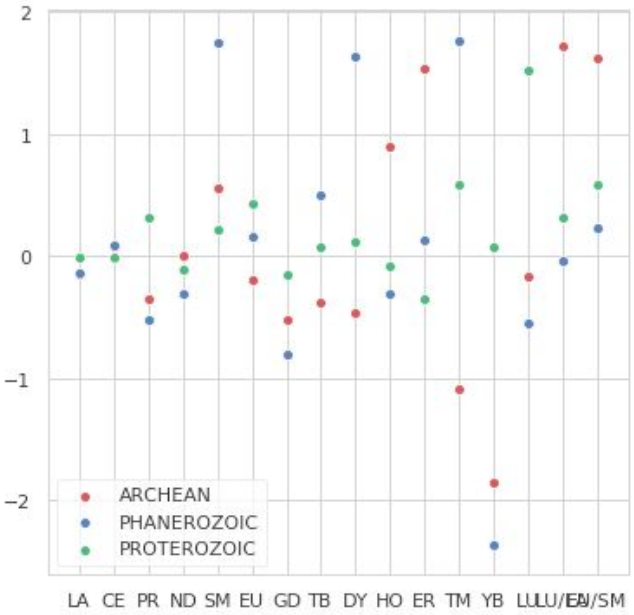
Ratios (SA): SVC coefficients; RFC, GBF importances

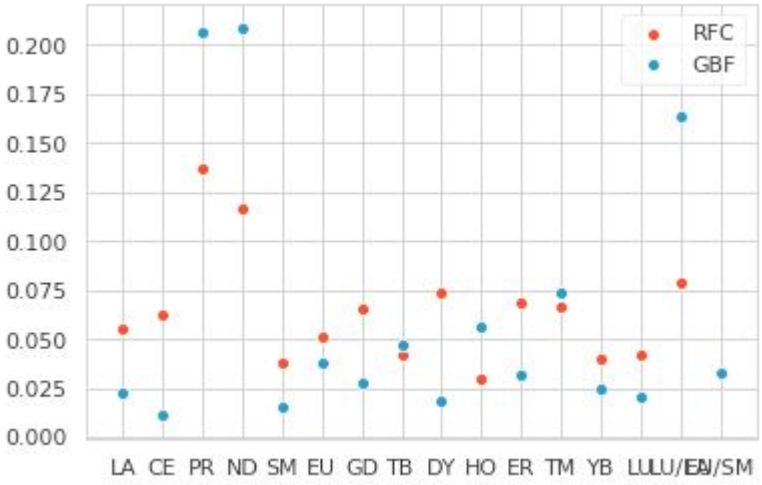




Gradient boosting forests have an OPINION.

Rare Earths (SA): SVC coefficients vs. RFC, GBF importances





Horticulture (SA)