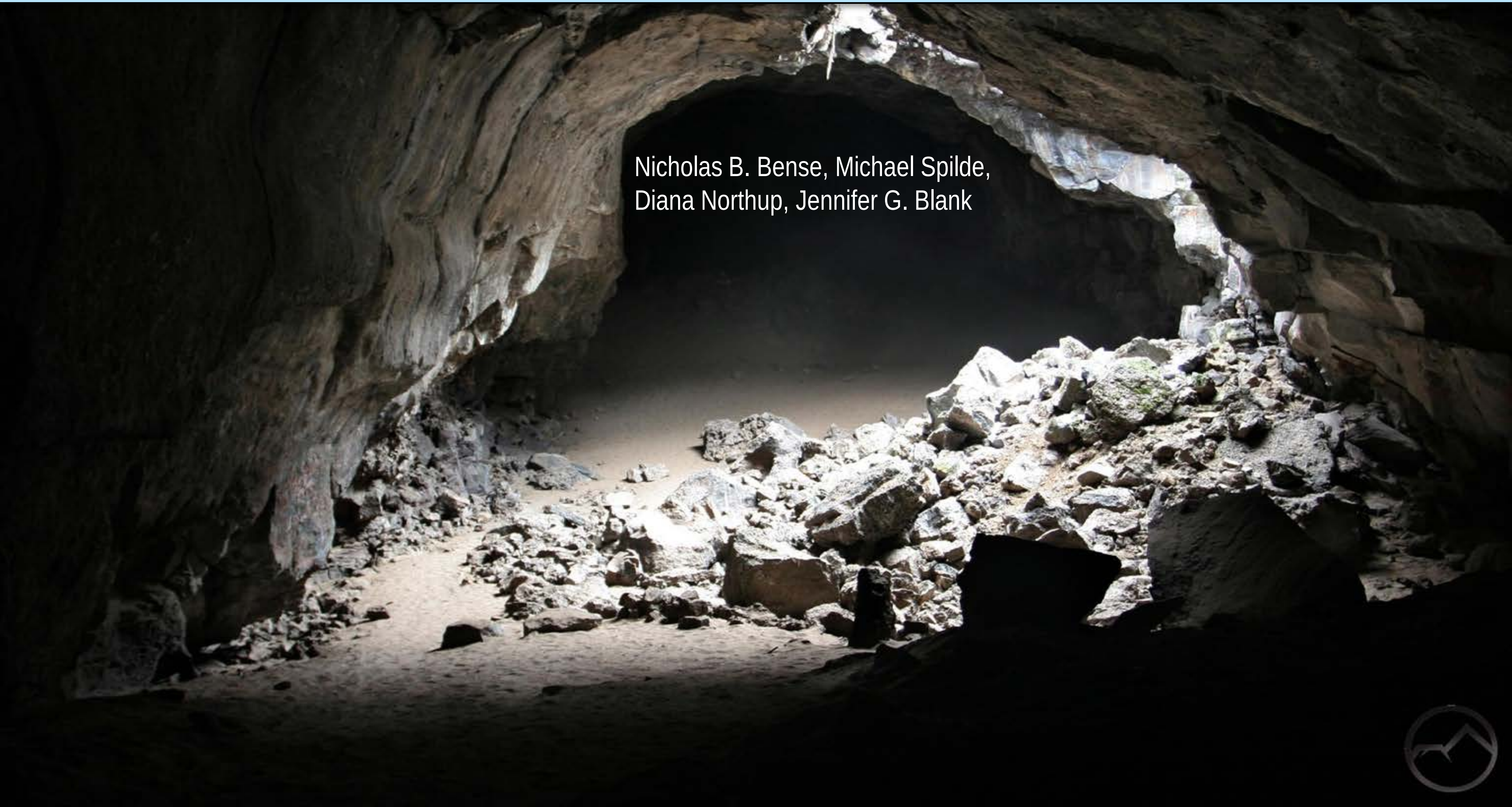
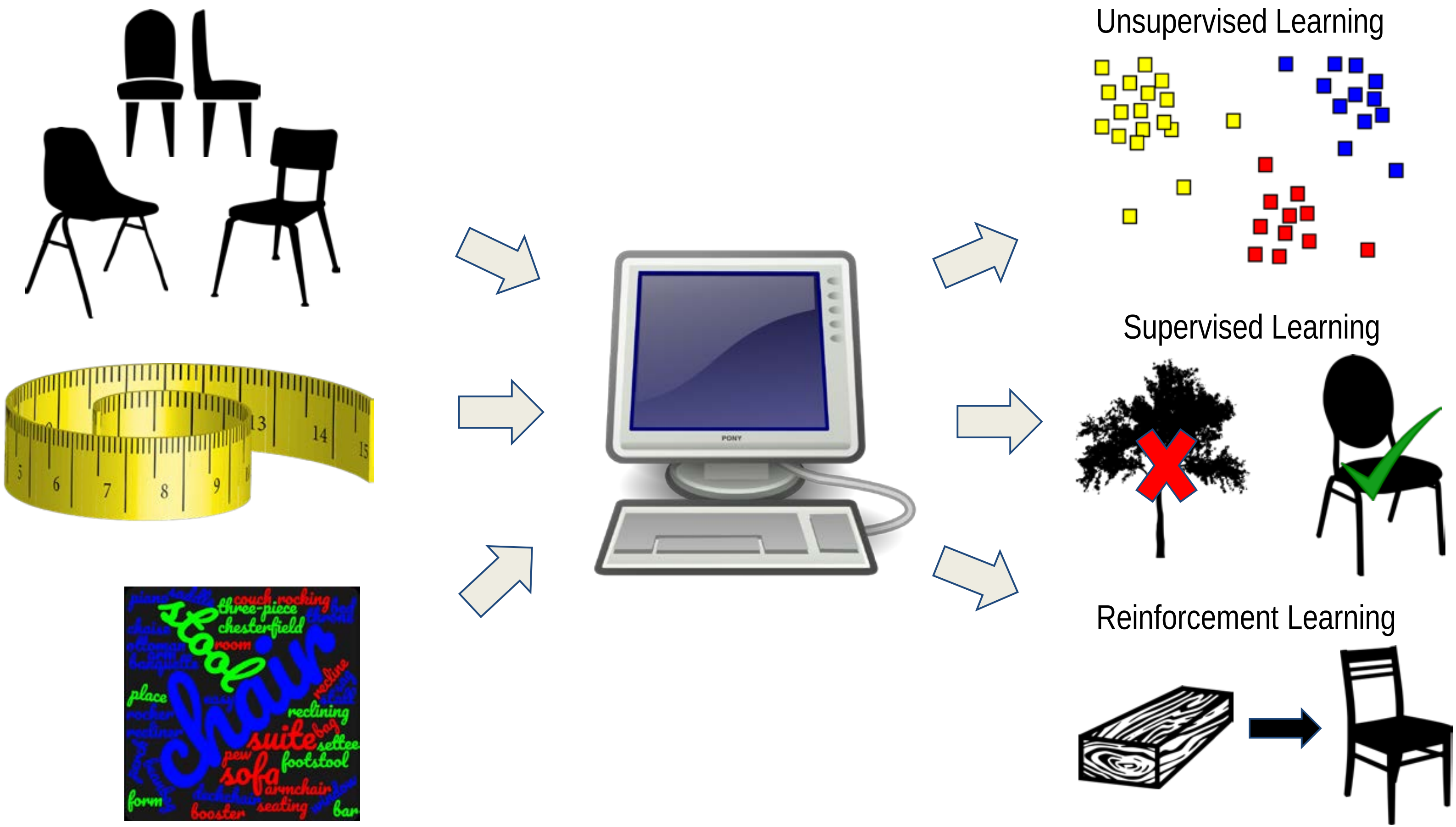
Probing Putative Biomarker Assignments in SEM Imagery from BRAILLE Investigations at Lava Beds National Monument (N. CA, USA) Using Machine Learning Tools

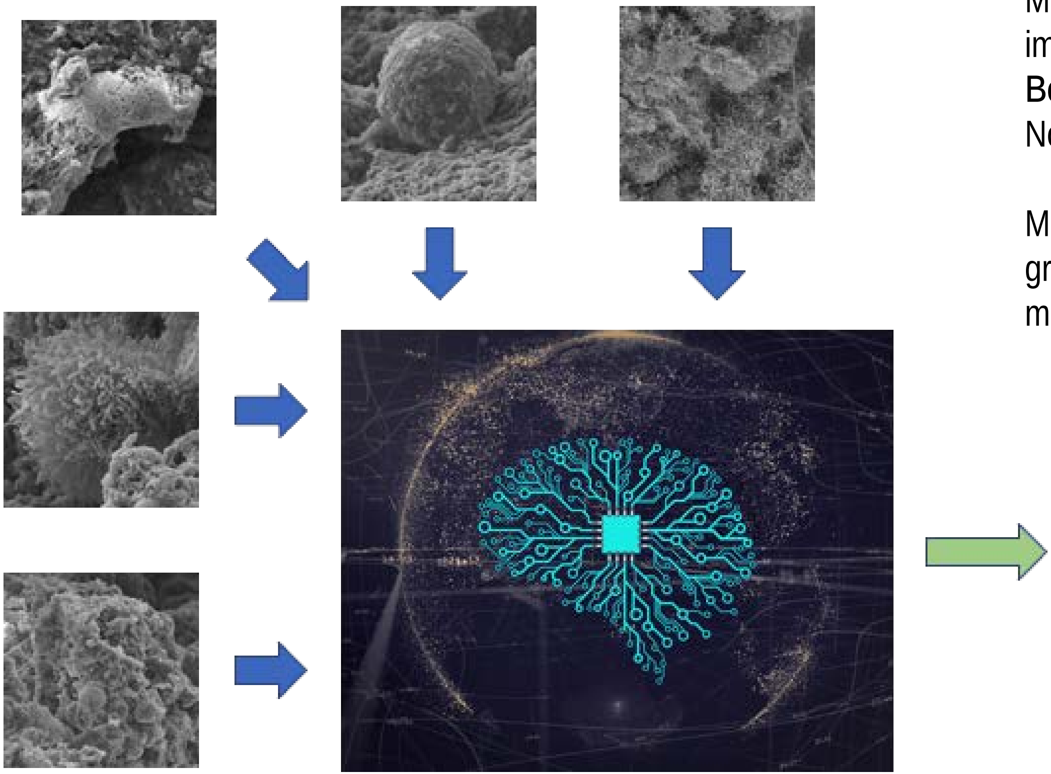
Nicholas B. Bense, Michael Spilde,
Diana Northup, Jennifer G. Blank



Machine Learning 101



Mineral to Microbe Continuum (MMC) Correlations

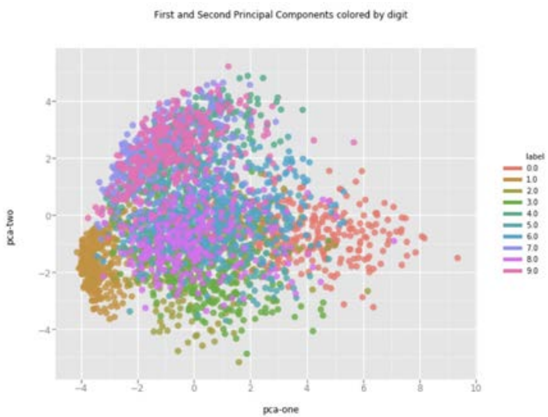


Mineral-Microbe morphologies in SEM images of lava tubes, sampled from **Lava Beds National Monument (LABE)**, Northern CA, USA

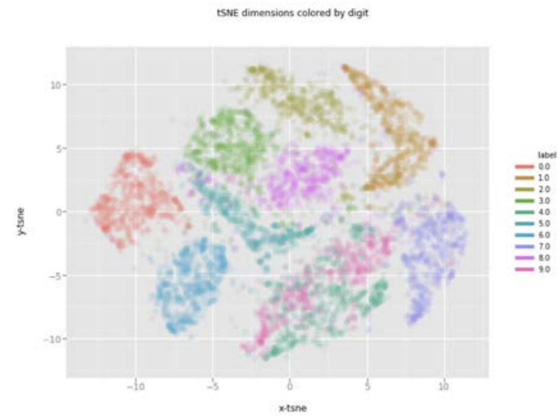
Mineral to Microbe Continuum (MMC): gradient of life activity across mineral/bio-mineral features

Unsupervised Clustering Techniques

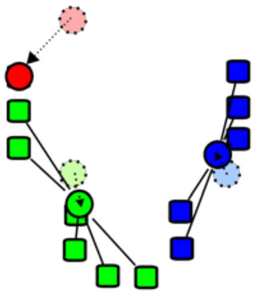
- Principal Component Analysis (PCA)
  - Used to visualize low dimensionality representation of data variance
  - Limited to linear projections
- t-Distributed Stochastic Neighbor Embedding t-SNE)
  - Probabilistic characterization of local structure
  - Prevents low dimensional “crowding problem”
  - Well suited for visualization of high-dimensionality
- K-Means Clustering
  - Optimizes grouping of data into user specified number of groups (“K”)
  - Reassigns data points to new group means with every round until no changes are made



PCA

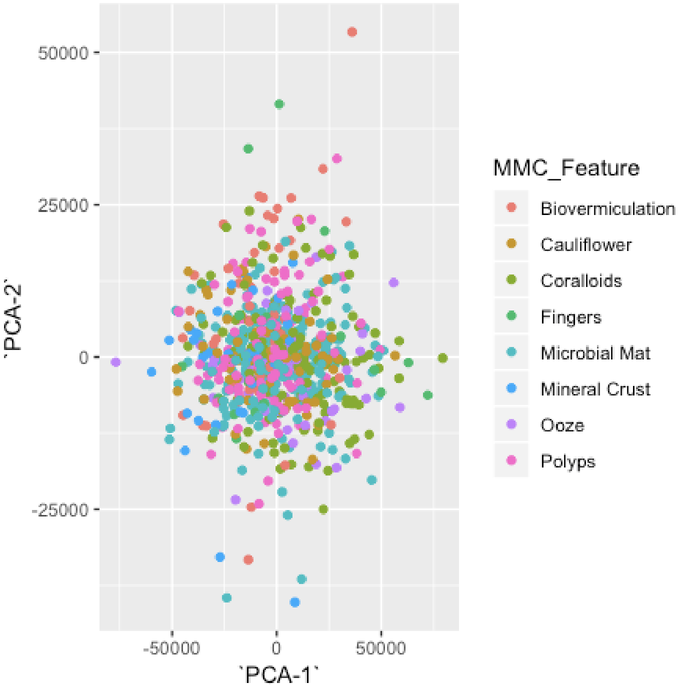


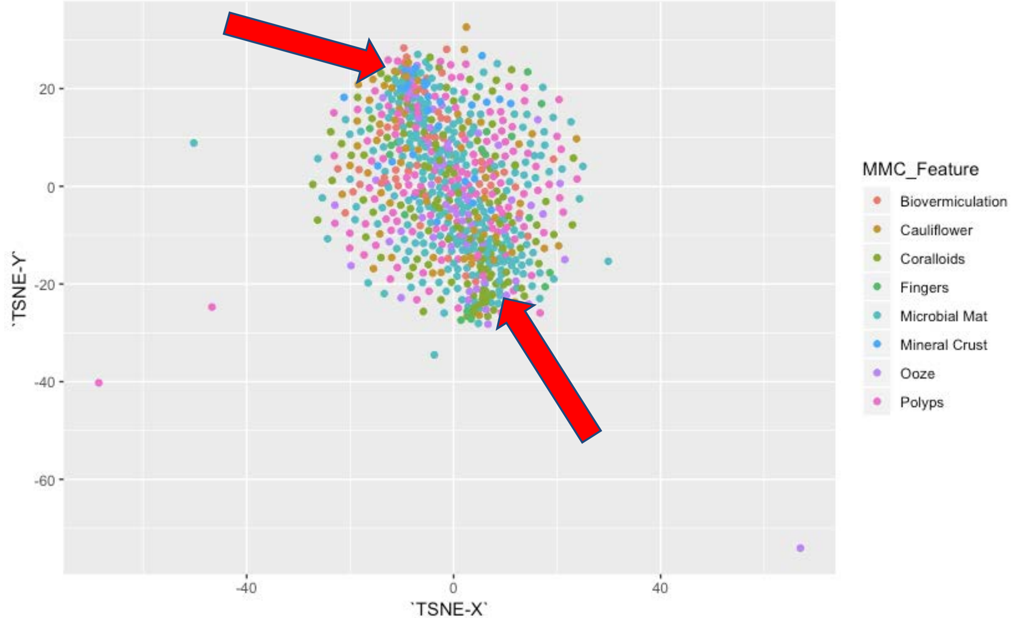
tSNE



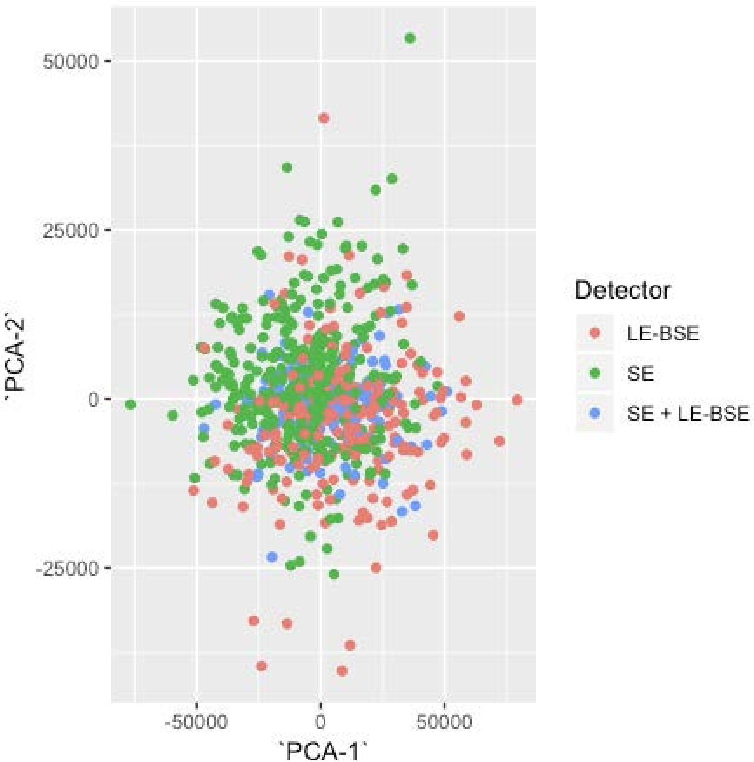
K-Means

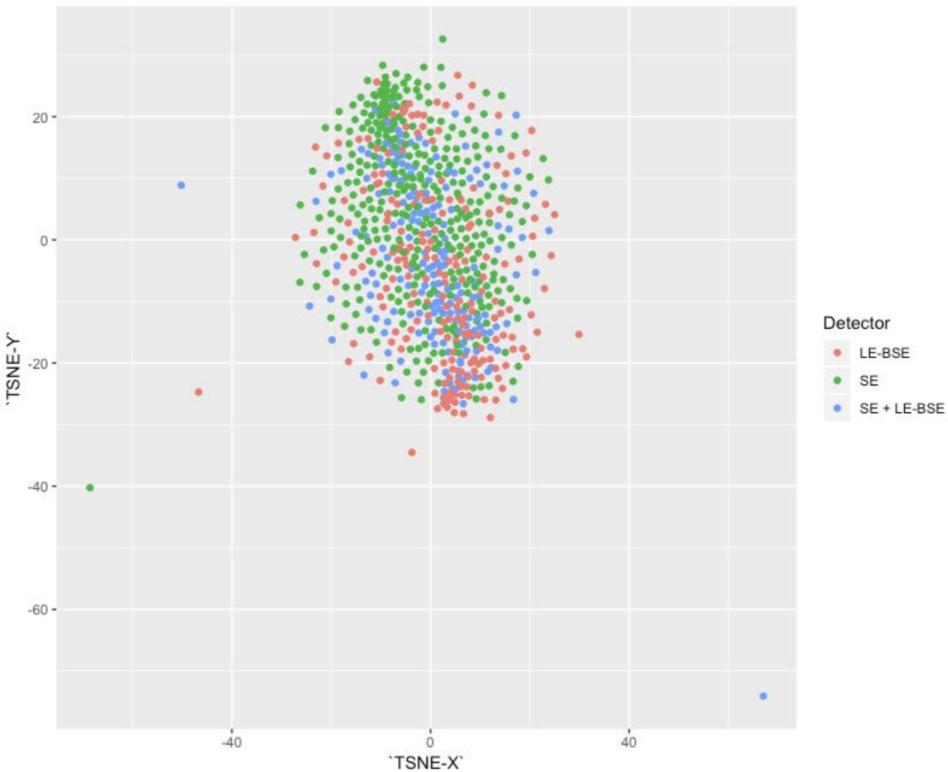
PCA & tSNE – MMC Features (n = 770)



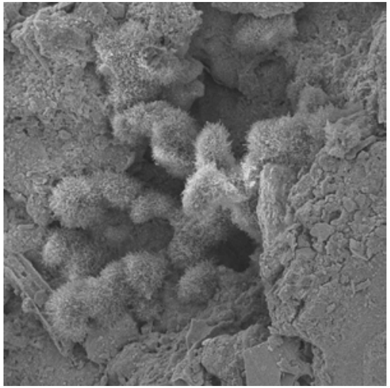


PCA & tSNE – Detector Bias

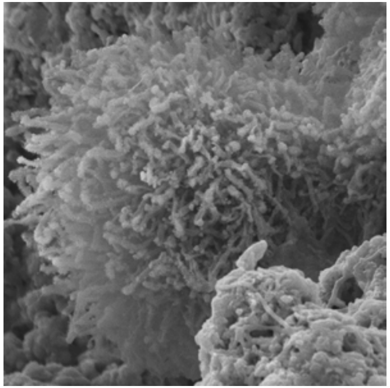




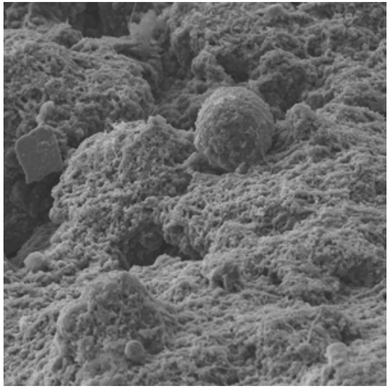
Magnification vs. Morphology



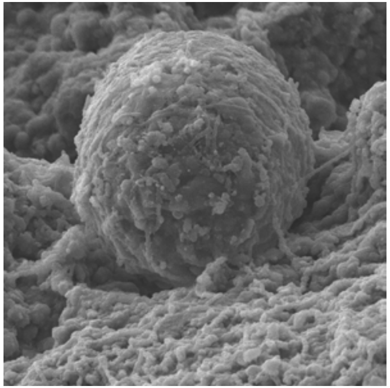
500x



4000x



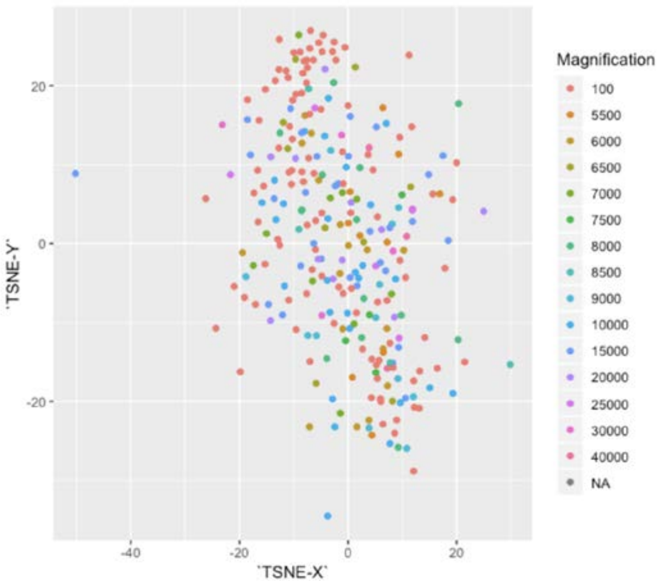
1000x



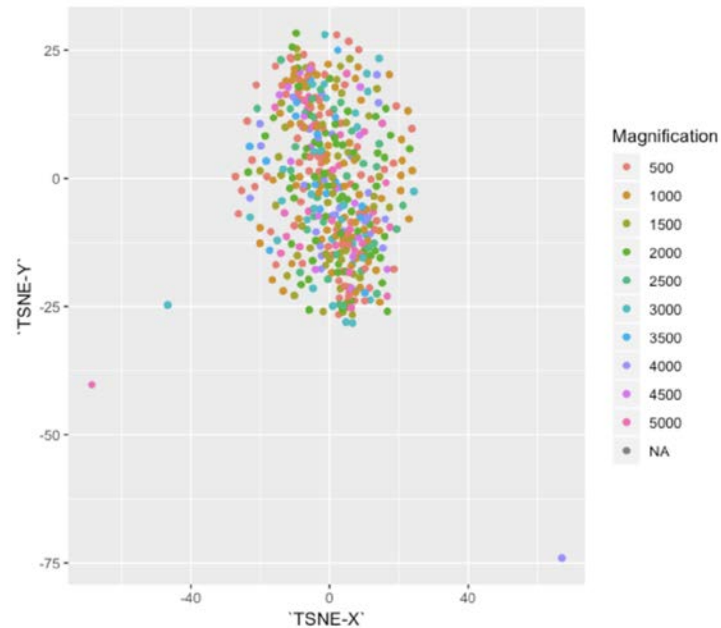
3000x

tSNE – Magnification Plots

Boundary Magnification Ranges (40% of Images)



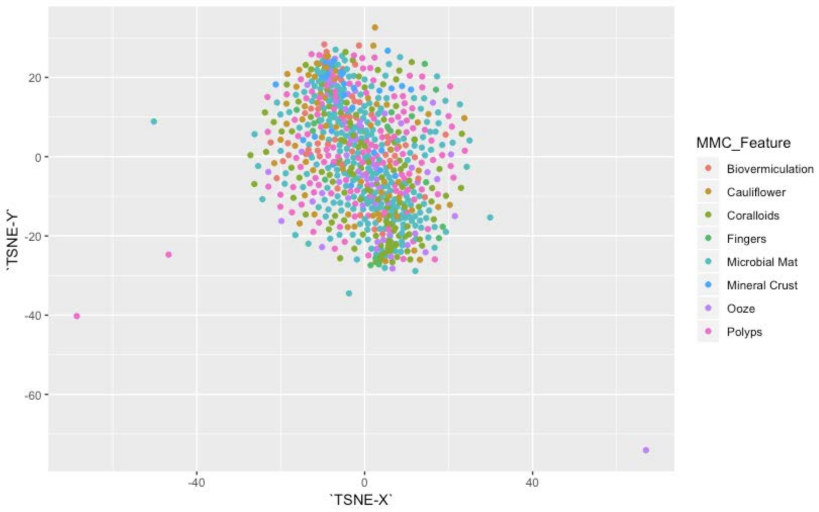
Median Magnification Ranges (60% of Images)



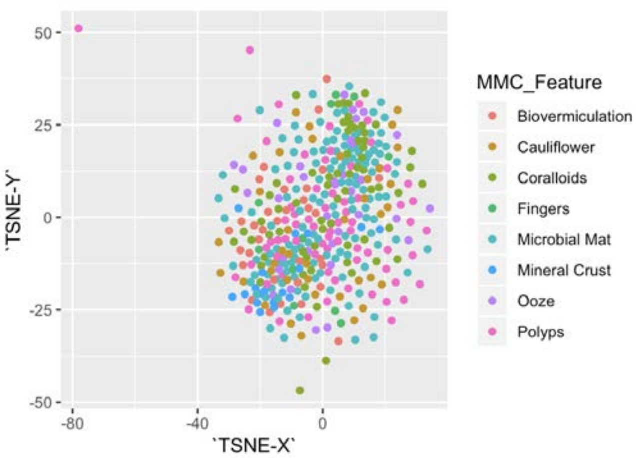
Min. 1st Qu. Median Mean
3rd Qu. Max.
100 500 2000

Magnification tSNE Comparison

All Images (n = 770)

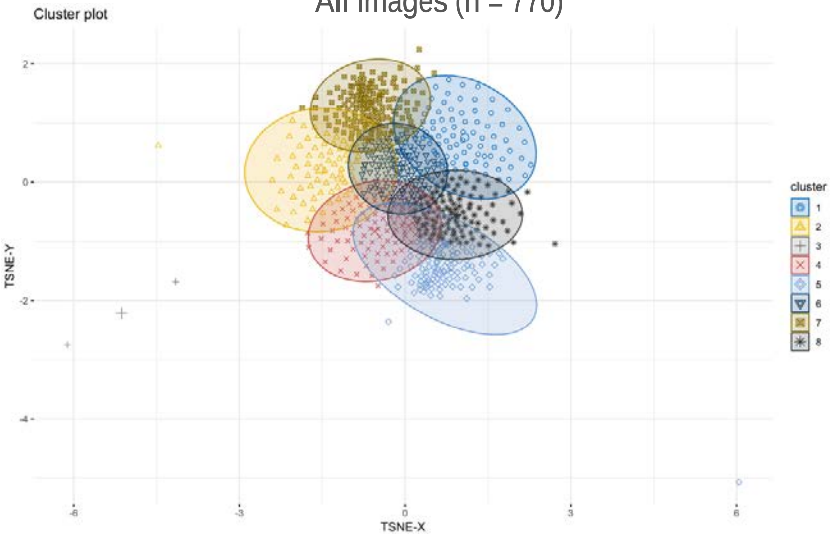


Magnification Subset (n = 401)

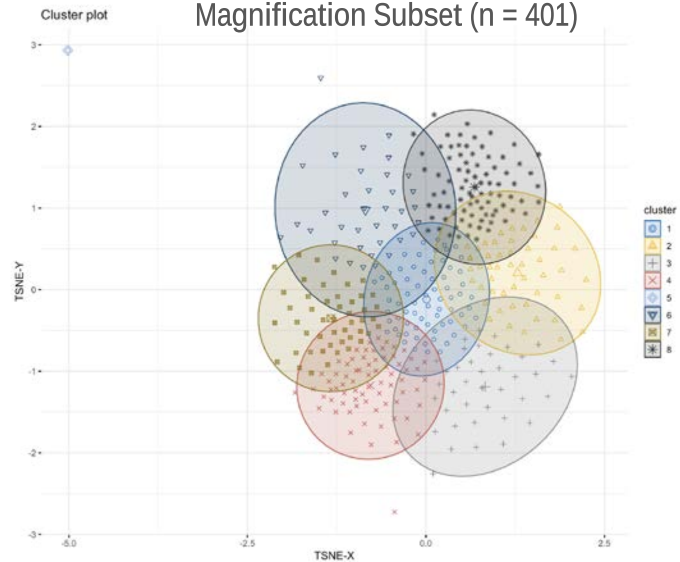


K-Means Clustering on TSNE

All Images (n = 770)



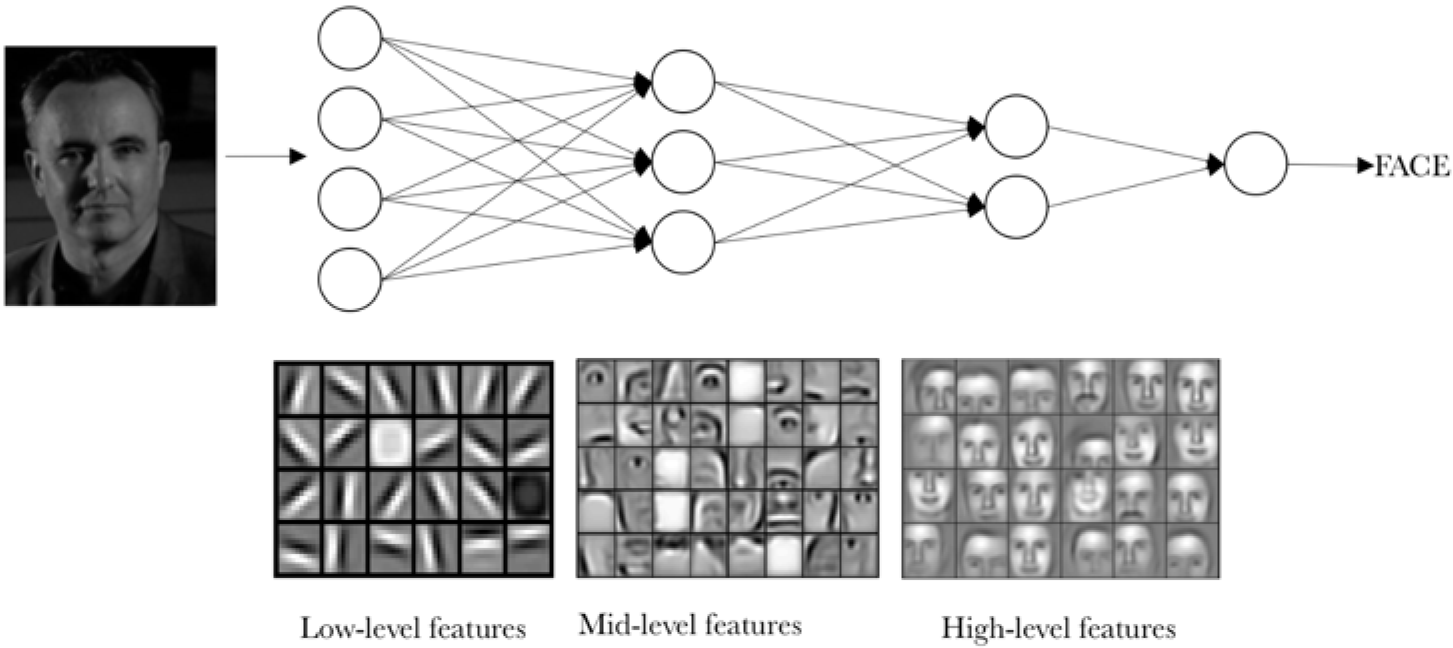
Magnification Subset (n = 401)



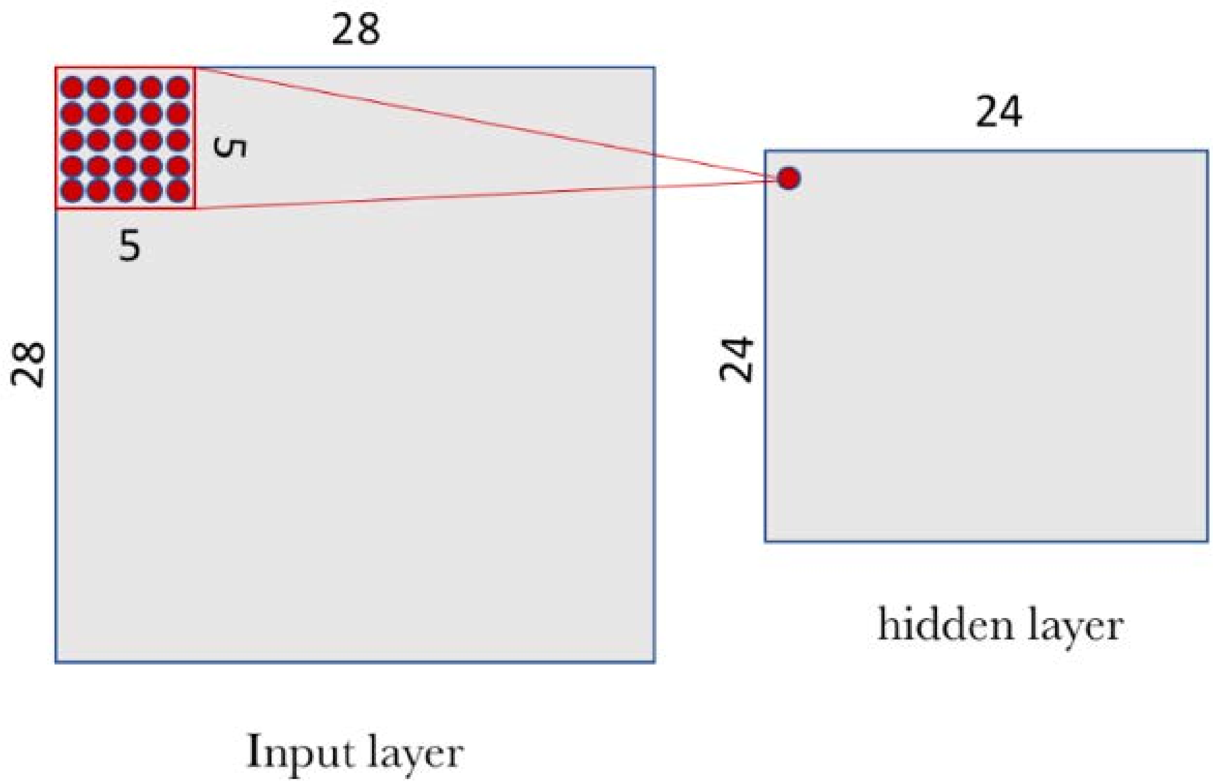
K-Means Cluster vs. MMC Category Corrected Rand Index: **0.006367637**
K-Means Cluster vs. MMC Category Chi-Squared P-Value: **2.721e-07**
MMC vs. TSNE-X + TSNE-Y
MANOVA P-Value: **8.545e-08**
Response TSNE-X P-Value: **0.03237**
Response TSNE-Y P-Value: **1.421e-**

K-Means Cluster vs. MMC Category Corrected Rand Index: **0.02526194**
K-Means Cluster vs. MMC Category Chi-Squared P-Value: **7.529e-08**
MMC vs. TSNE-X + TSNE-Y
MANOVA P-Value: **8.177e-07**
Response TSNE-X P-Value: **6.703e-06**

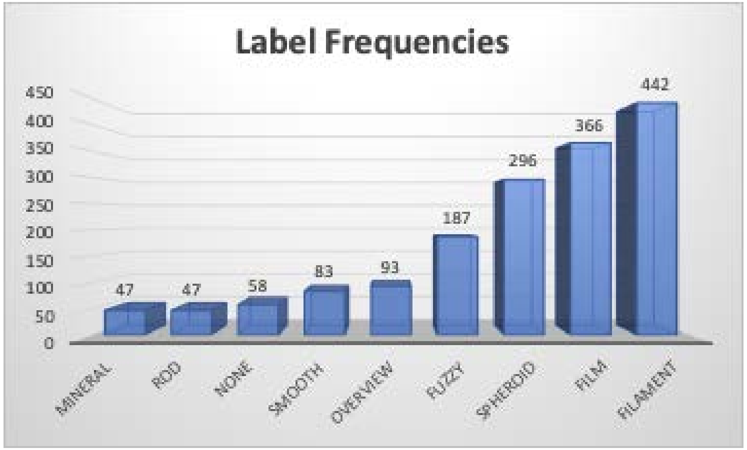
Convolutional Neural Networks (CNNs) for Image Classification



Convolution Layers



CNN Results

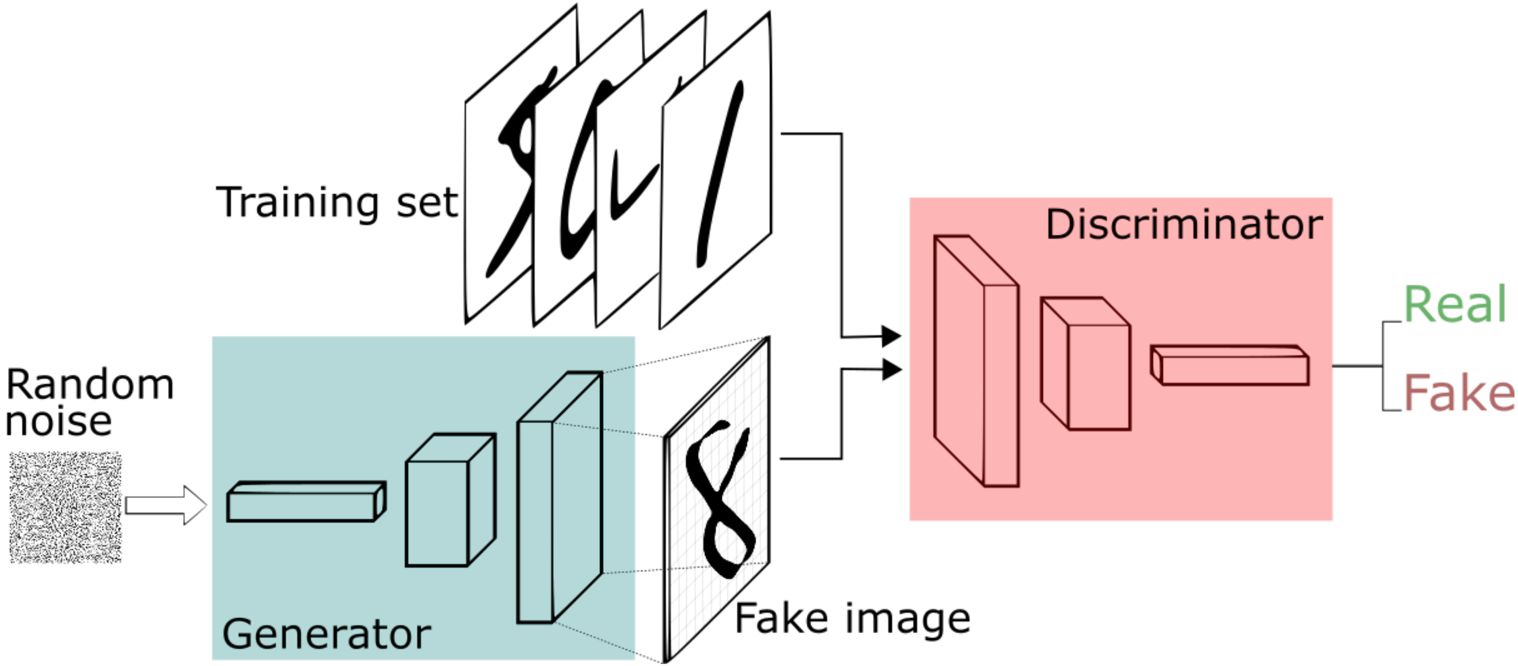


- Binary vs. multiclass vs. multilabel
- Overfitting/Biased Classes
- Possibly insufficient training data

Accuracy

	Binary	Multiclass
Cross-Validation	75%	80%
Hold-Out Testing	55%	50%

Generative Adversarial Networks



Next Steps

- Strengthen supervised classification
  - Generative Adversarial Networks
  - Leverage pre-built models

- Develop methodology for generating unbiased datasets
  - Raster scans
  - Random fields

- Applications
  - Automatically detect microscopic life
  - Correlate morphology with distinct taxonomies
  - Automatically detect biomineralization features/extinct life

Acknowledgements

The Institute of Meteoritics

NASA GeneLab

OPEN SCIENCE FOR LIFE IN SPACE

sky mind