Why Bad Data Breaks Predictive Models: Microscale Textures, Hyperspectral Unmixing, and Mine Profitability

2nd HSI Symposium
October 8, 2025

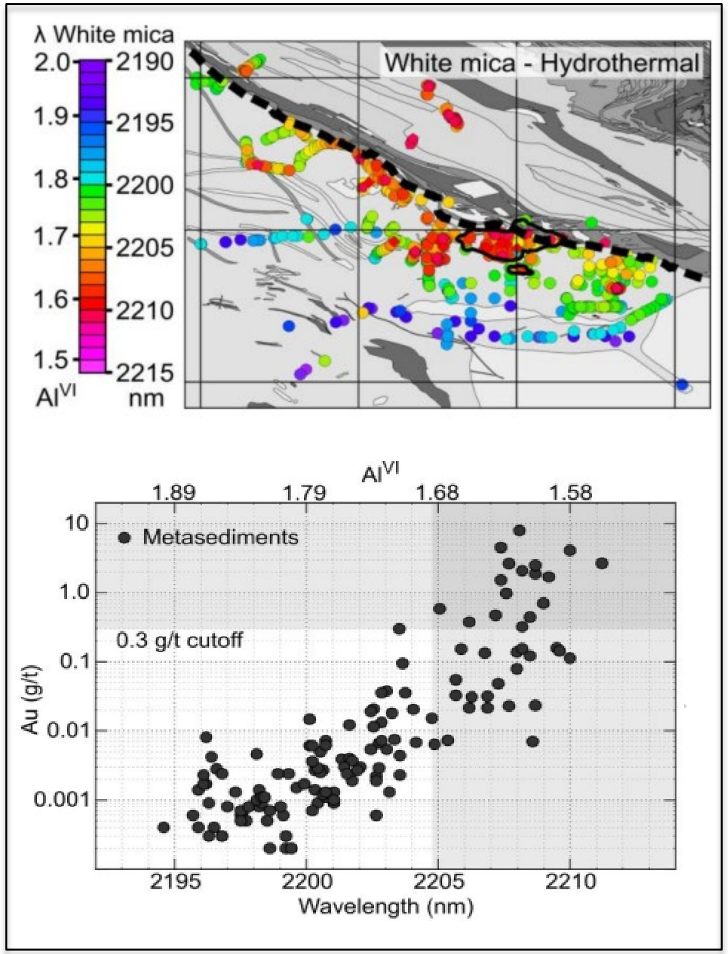
Sam Scher | sscher@lkiconsulting.com
Brigette Martini | brigette.martini@angloamerican.com

Microscale textures control macroscale outcomes

- Microscale mineral textures drive:
  - Geometallurgical domain boundaries
  - Ore grade distribution
  - Comminution/Hardness
  - Metallurgical recovery
  - Leach efficacy

- Recovery forecast (and thus mine profitability $$) depends on the fidelity of the data describing those textures.

- Most 'traditional' rock quality measures are at large scales (e.g. one BWI measure per 500m in a block model); the promise of hyperspectral has always been continuous, consistent, spatially cohesive data at sample intervals orders of magnitude higher than traditional workflows.



Lypaczewski et al., 2019

Where does the data breaks down

- Current process:

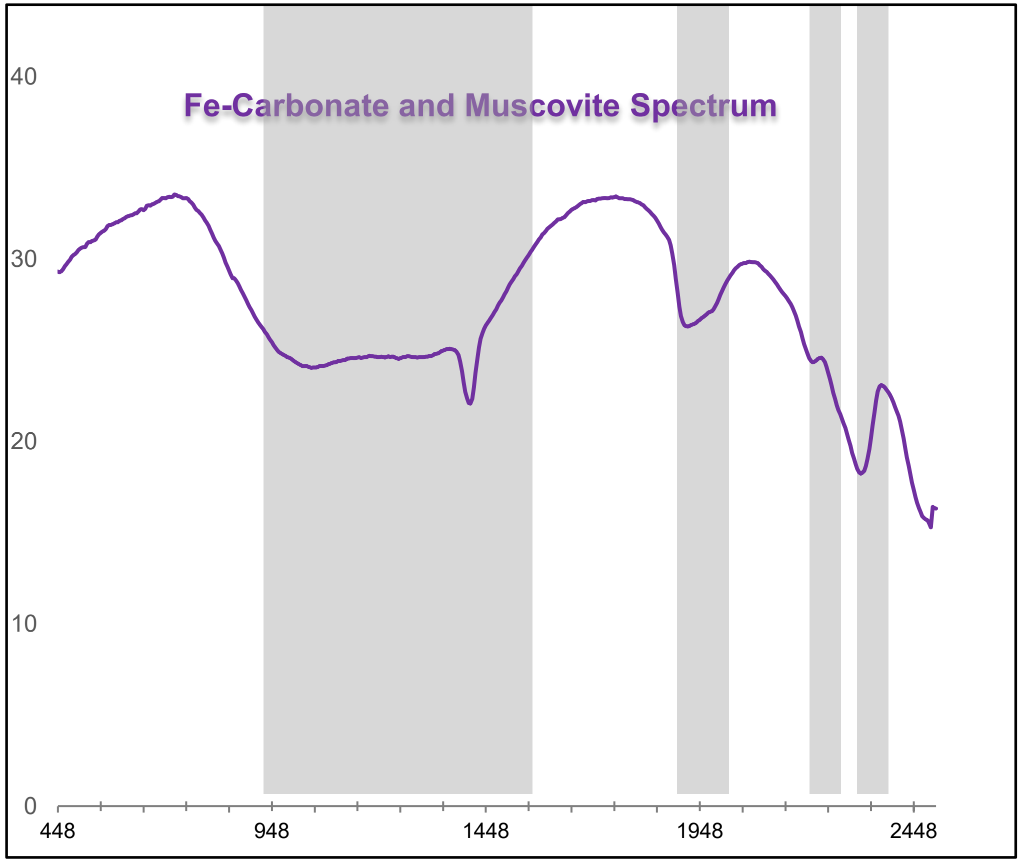
- At each step, assumptions compound with the end users (e.g., geochemists, 3D modelers, data scientists) relying on the hyperspectral core imaging (HSI) vendor to provide data that approximates true rock mineralogy.

What is spectral unmixing

LKI Consulting

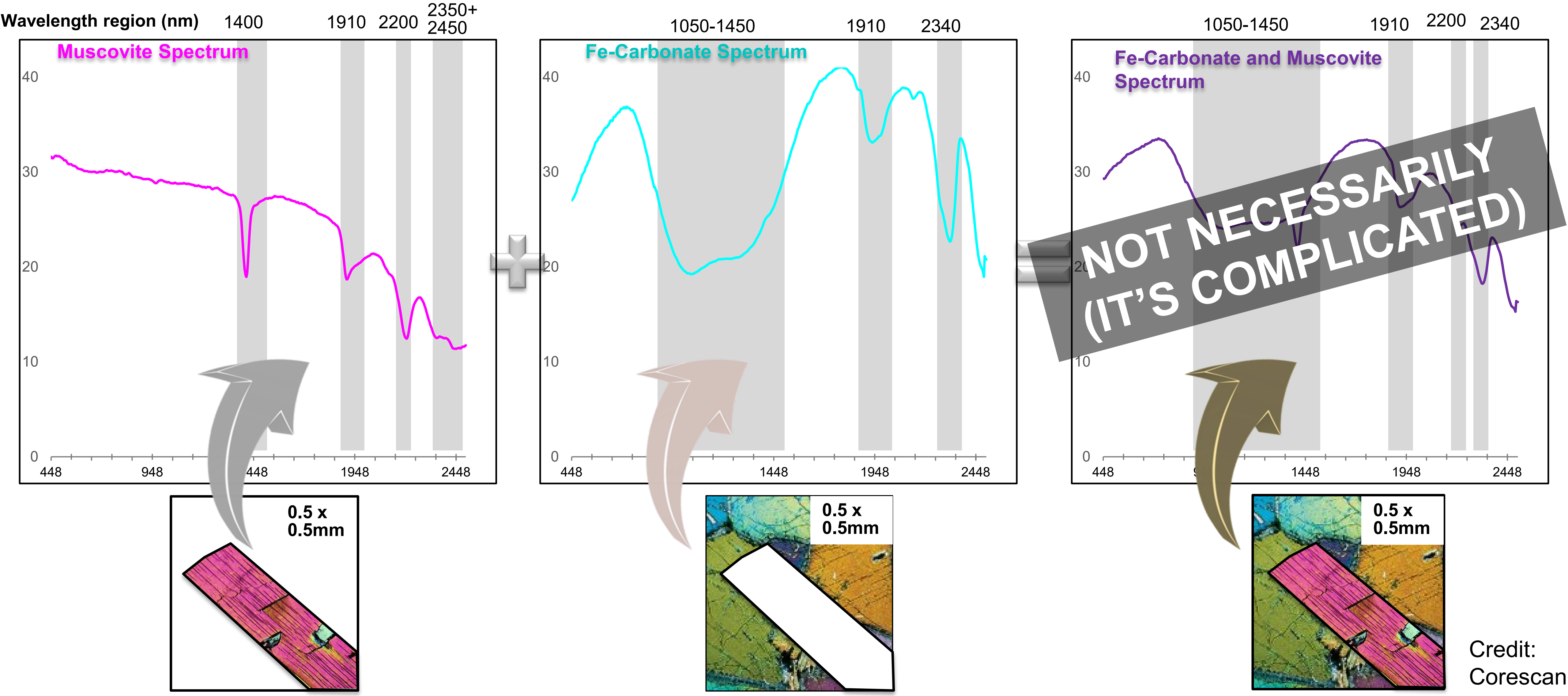
What is spectral unmixing?

- A spectrum represents reflected energy measured across a given wavelength range.
- Unmixing attempts to estimate proportions of multiple minerals within one pixel.
- While conceptually simple, mixtures themselves are not straightforward; not withstanding the underlying statistics of applying unmixing algorithms.
- For practical reasons (and guided by the dominance of proximal point spectroscopy up until only the last ~7-10 years) we mostly treat mixing phenomenology as linear.
- In reality, mineral mixtures are not as linear as we might hope – a fact that is now inordinately experienced with the advent of core imaging.
- Fine-grain intimate mixtures, albedo effects and amorphous content are just some of the sources of non-linear mixing.



Applying the principles of linear unmixing, reveal a relatively simple Fe-carbonate and muscovite spectrum.

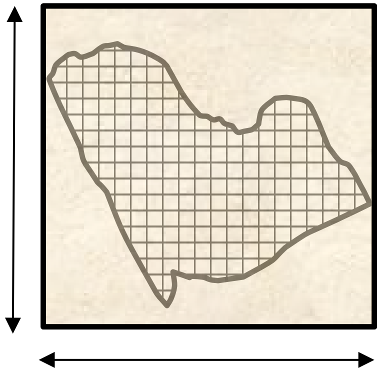
Spectral mixing is not as straightforward as it seems

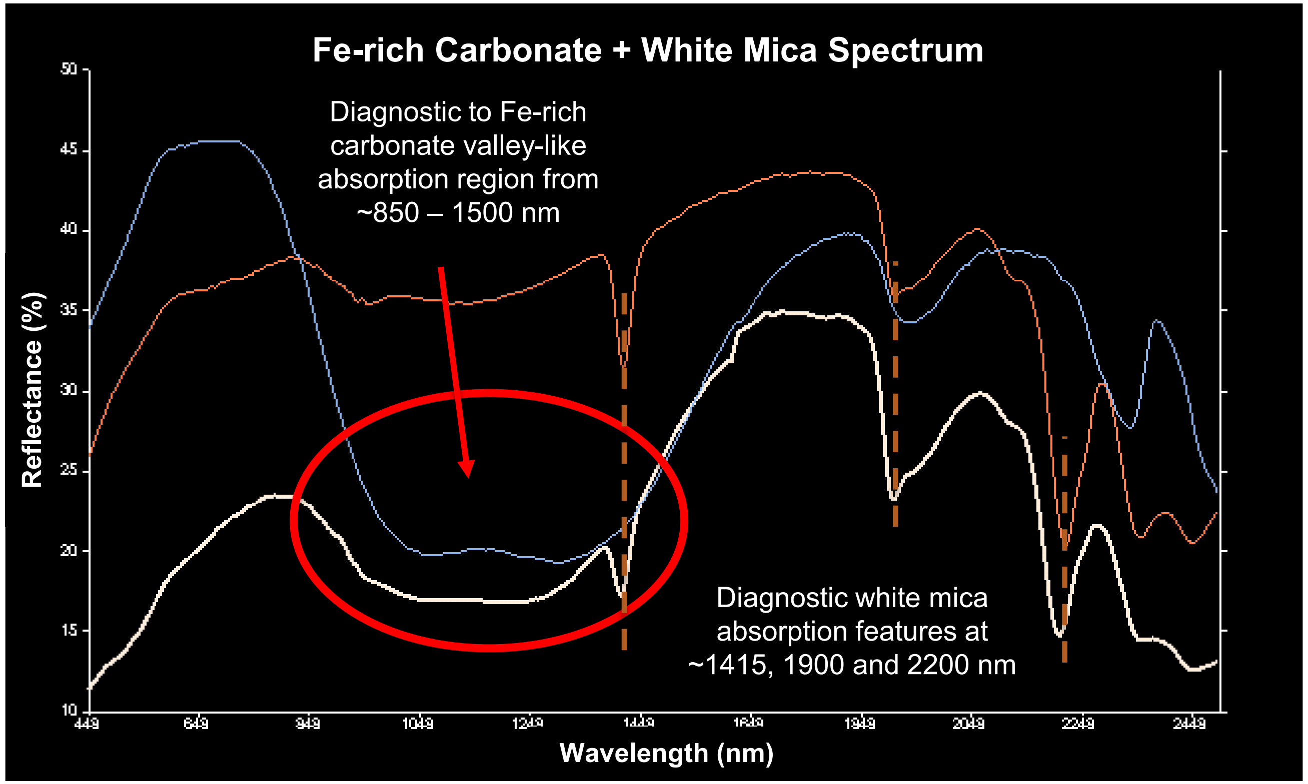


Credit: Corescan

Types of mixture: linear

- **Linear mixture**
  - Materials in field of view are optically separated.
  - No multiple scattering between components.
  - Combined signal is the sum of the fractional area times the spectrum of each component.
  - Also called an ‘areal’ mixture.’
- Intimate mixture
- Coatings
- Molecular mixture



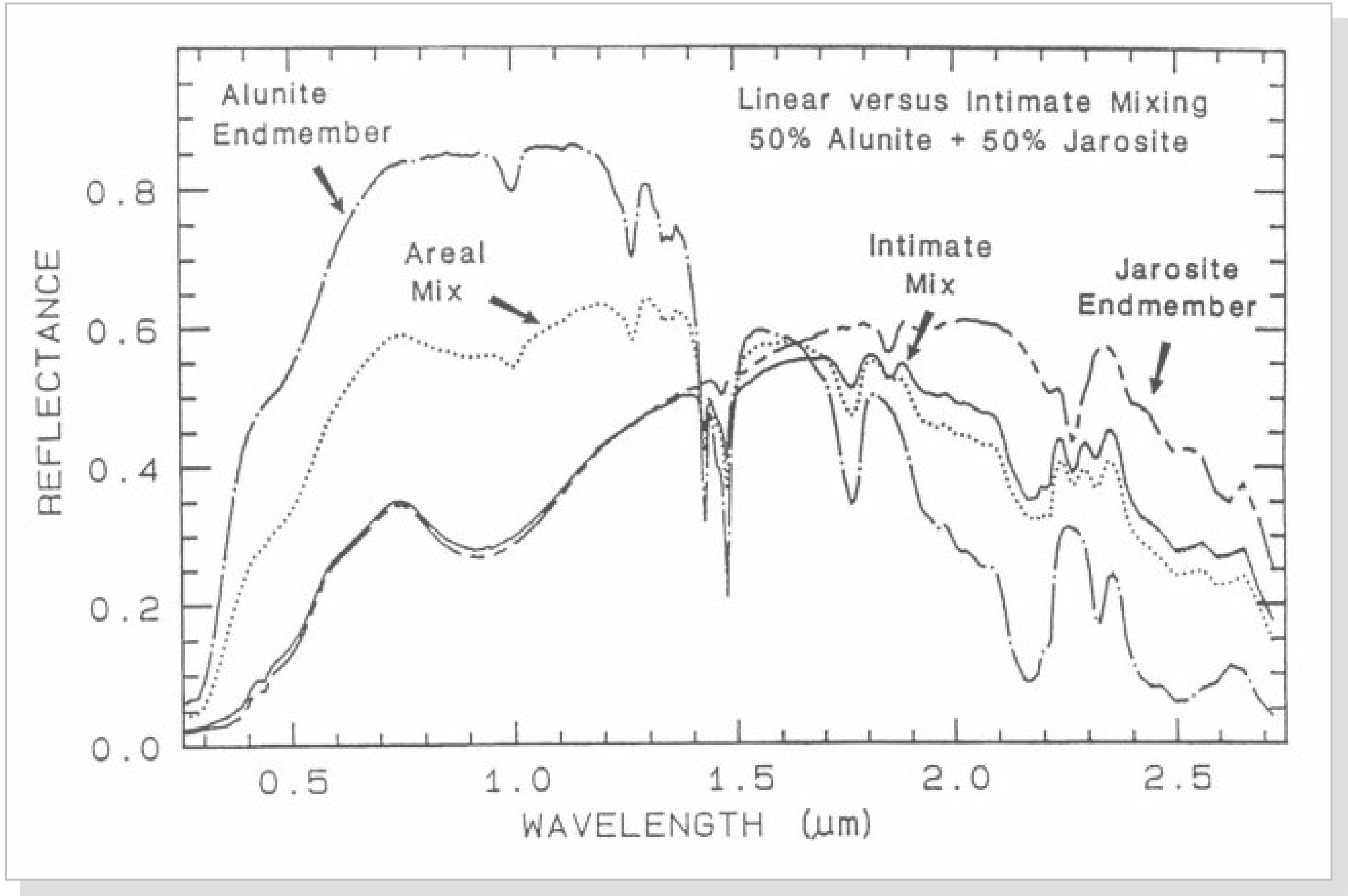


White – Project sample spectrum
Orange – Reference library white mica spectrum
Blue – Reference library siderite spectrum

Credit: Corescan

Types of mixtures: intimate mixtures

- Linear mixture
- **Intimate mixture**
  - When different materials are in intimate contact in scattering surface.
  - Depending on optical properties of each component this can result in a signature that is a highly non-linear combination of each endmember.
- Coatings
- Molecular mixture

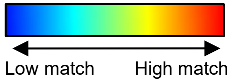


Reflectance spectra from alunite (long dash + dot line) and jarosite (long dash + 2 small dash line). Two different mixture types of these two minerals are also shown: 1) an areal mixture (dotted line), and 2) an intimate mixture (solid line) (Rencz and Robert, 1999).

Credit: Corescan

Types of mixtures: coatings

- Linear mixture
- Intimate mixture
- **Coatings**
  - Occurs when a thin layer of one material covers an in-situ mineral in the rock.
  - Coatings can be naturally occurring, in-situ in the rock (such as hematite dusting) or caused by surface weathering, contamination, or oxidation.
  - Each coating is another layer to transmit the photons through.
  - Many of these mixtures are non-linear.
- Molecular mixture



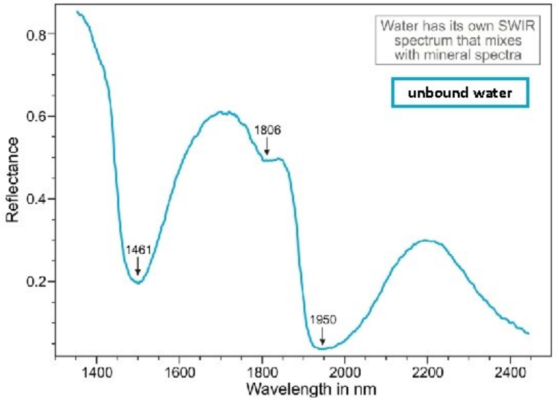
Gypsum coating from weathering

Same sample – cut to expose fresh surface

Credit: Corescan

Types of mixtures: molecular

- Linear mixture
- Intimate mixture
- Coatings
- **Molecular mixture**
  - Occurs on a molecular level (such as a mixture of two liquids or a liquid and solid) where different molecules co-occur but are not chemically bonded.
  - The combined signal that returns to the detector is non-linear.



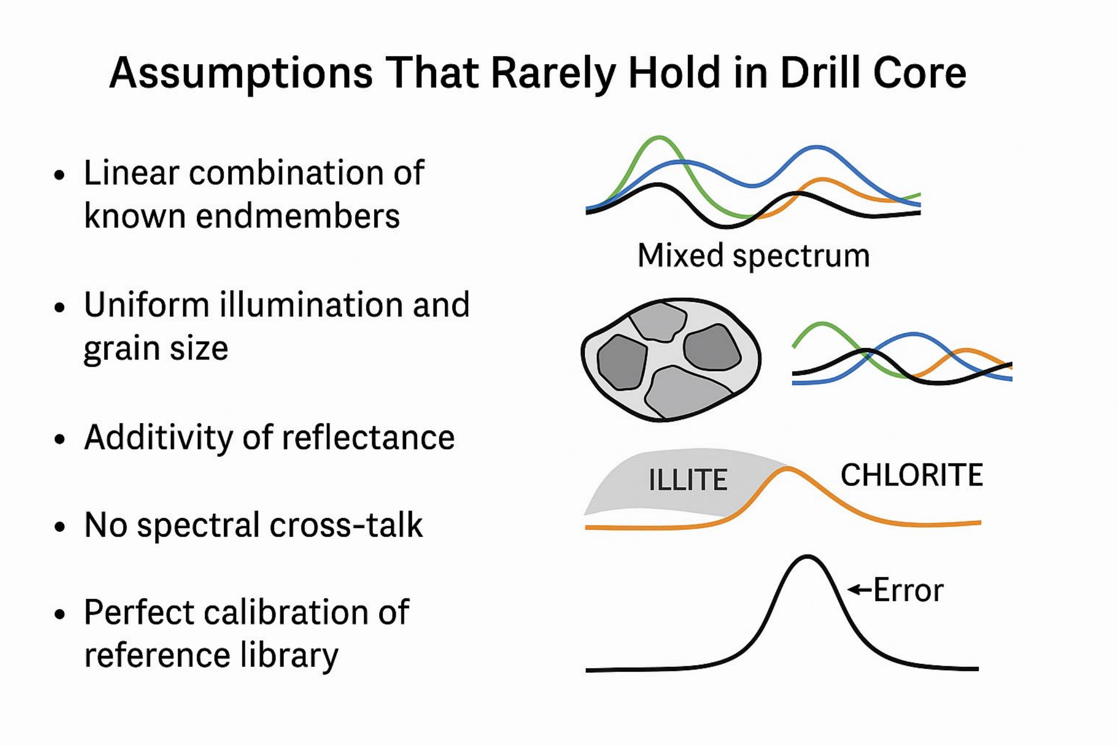
(from Laukamp, 2015)

Quartz with fluid inclusions produces a spectrum which displays the broad, negative slope indicative of quartz with sharp absorption features at ~1450nm, ~1810nm, and ~1950nm, indicating unbound water. The two are not chemically bonded, but the signal represents both materials (although this resulting signature is not a linear mixture of the two materials).

Credit: Corescan

Mathematical Assumptions

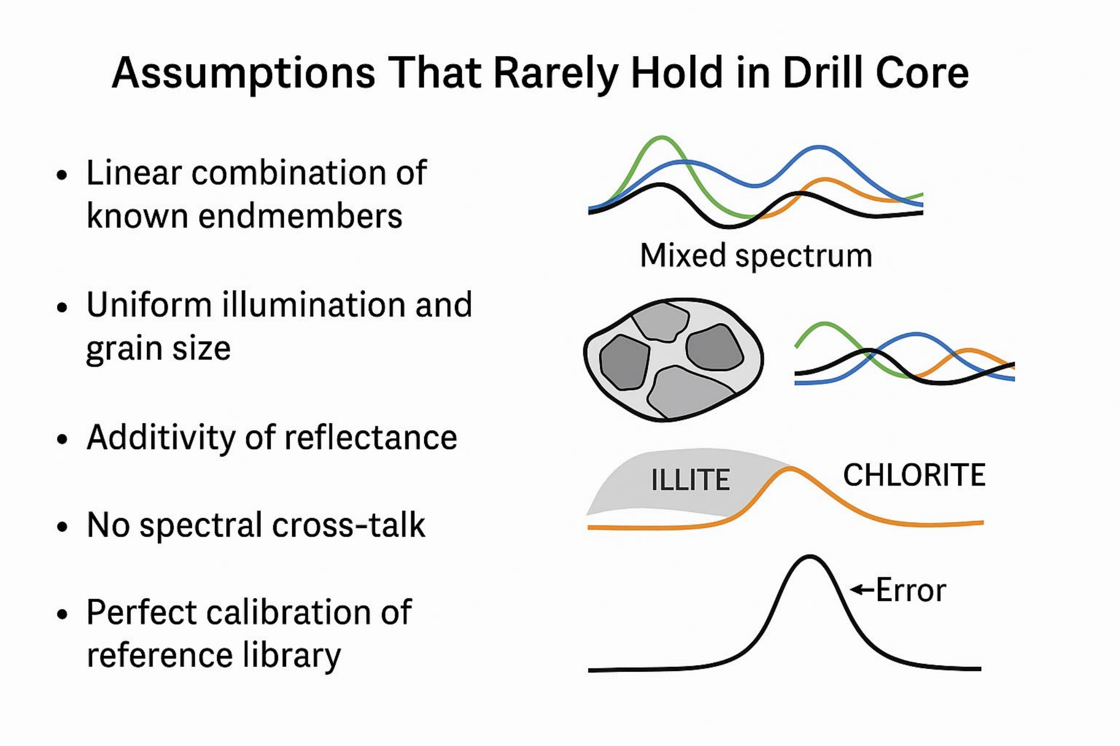
- Linear combination of known endmembers: the mixed spectrum (analyzed by the instrument) is assumed to be a weighted sum of a few pure mineral spectral endmembers.
    - This fails because mineral grains are rarely optically isolated whereby light scatters between them and coatings or inclusions alter reflectance in non-linear ways. Therefore, “linearity” is violated.
- Uniform illumination and grain size: each pixel or scan point is assumed to have even lighting and uniform particle size.
    - This fails because in drillcore, surface roughness, saw marks, and variable grain sizes change brightness and spectral contrast, distorting apparent absorption depths and shifting wavelengths slightly.
- Additivity of reflectance
- No spectral cross-talk
- Perfect calibration of reference library



Mathematical Assumptions

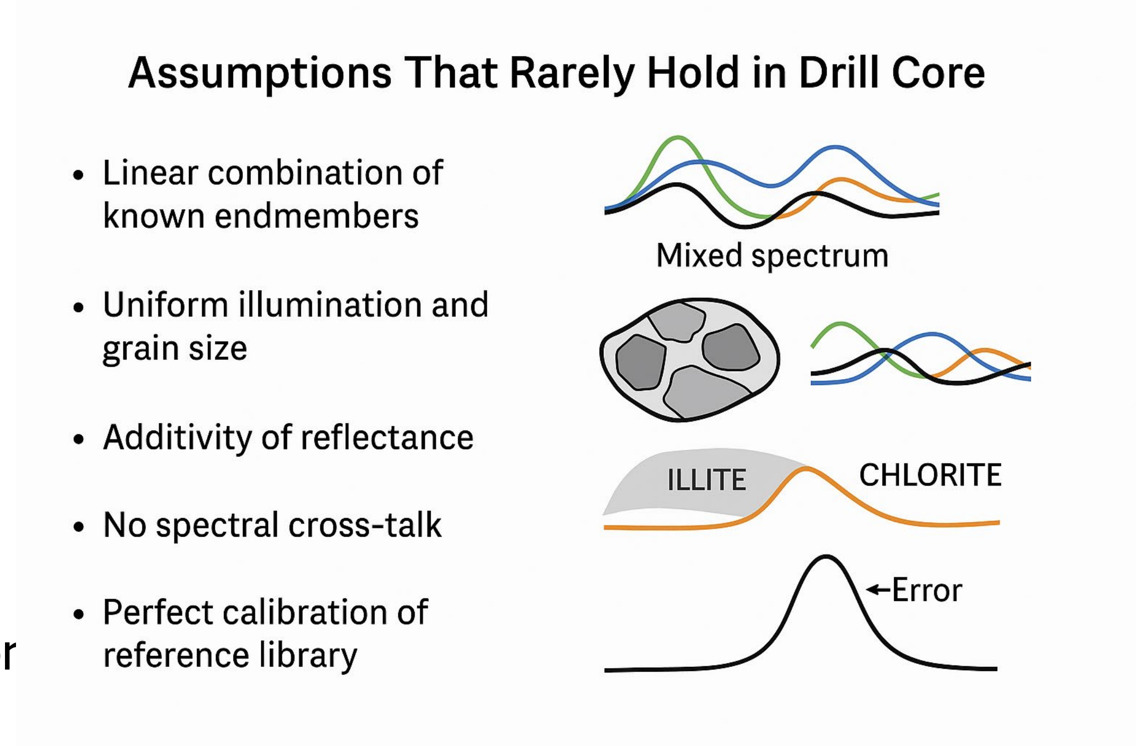
LKI Consulting

- Linear combination of known endmembers
- Uniform illumination and grain size
- Additivity of reflectance: the reflectance from a mixture equals the sum of the reflectance of its parts, weighted by abundance.
    - This fails because, in reality, reflectance is non-linear because photons interact multiple times within grains. Minerals with different albedos (e.g., dark biotite and bright kaolinite) interact in complex ways that cannot be summed linearly.
- No spectral cross-talk: each mineral's diagnostic absorption features are distinct and non-overlapping.
    - This fails because many phyllosilicates (e.g., illite, muscovite, montmorillonite, chlorite) have overlapping OH and Al–OH bands between 2200–2350 nm. When bands overlap, the algorithm can't distinguish who "owns" the absorption — causing misassignments or spurious correlations.
- Perfect calibration of reference library



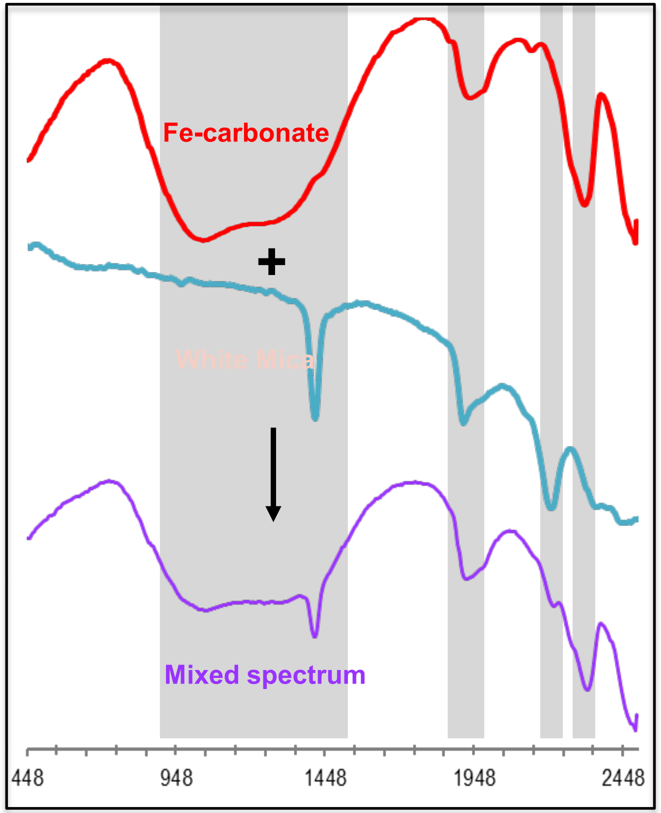
Mathematical Assumptions

- Linear combination of known endmembers
- Uniform illumination and grain size
- Additivity of reflectance
- No spectral cross-talk
- Perfect calibration of reference library: the "pure" spectra used for unmixing are perfectly measured, under the same conditions, and representative of the minerals in your sample.
    - This fails because library spectra are often collected under different lighting, grain size, or instrument settings; even small calibration offsets (e.g., 2–3 nm) can shift absorption minima enough to misclassify minerals.

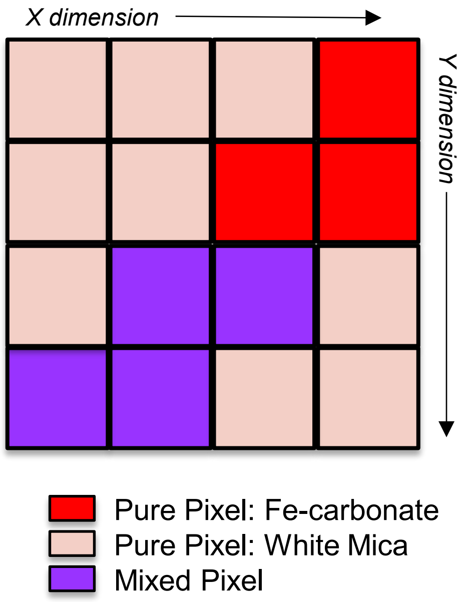


Quantification of hyperspectral data

- Quantification* of mineralogy from hyperspectral data is usually done in one of two ways:
  - Point counting (records presence/absence of each mineral in an analysis spot/spectrum)
  - Spectral unmixing (tries to record relative abundance of each mineral within an analysis spot/individual spectrum)



Imaging spectroscopy



Point counting

Fe-Carbonate = 7/16 = 43.75%

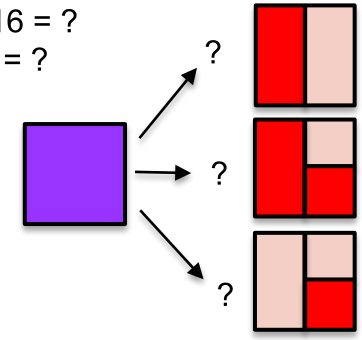
White Mica = 13/16 = 81.25%

Values easily normalized to 100%

Spectral Unmixing

Fe-Carbonate = (3 + ?)/16 = ?

White Mica = (9 + ?) /16 = ?

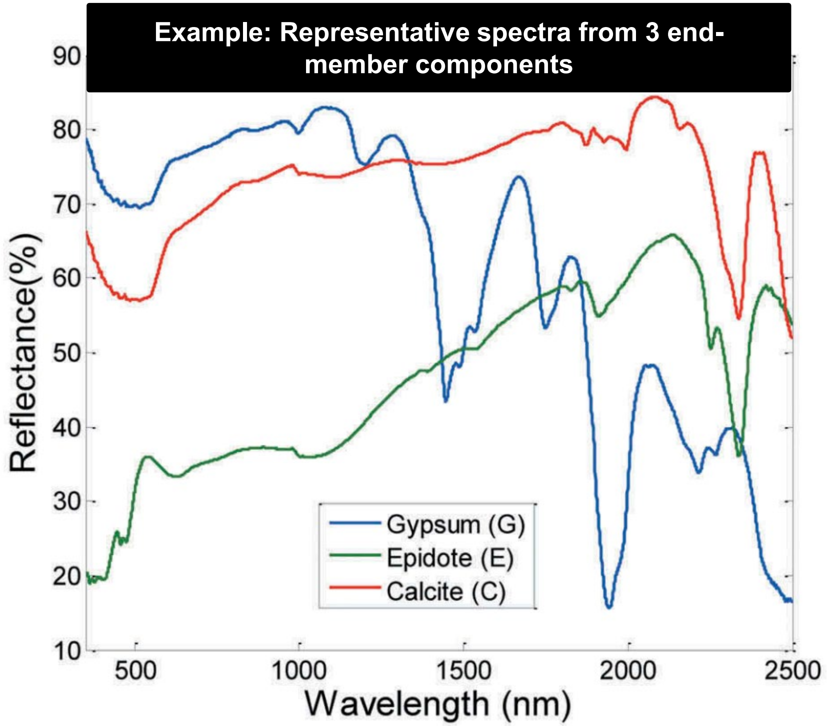


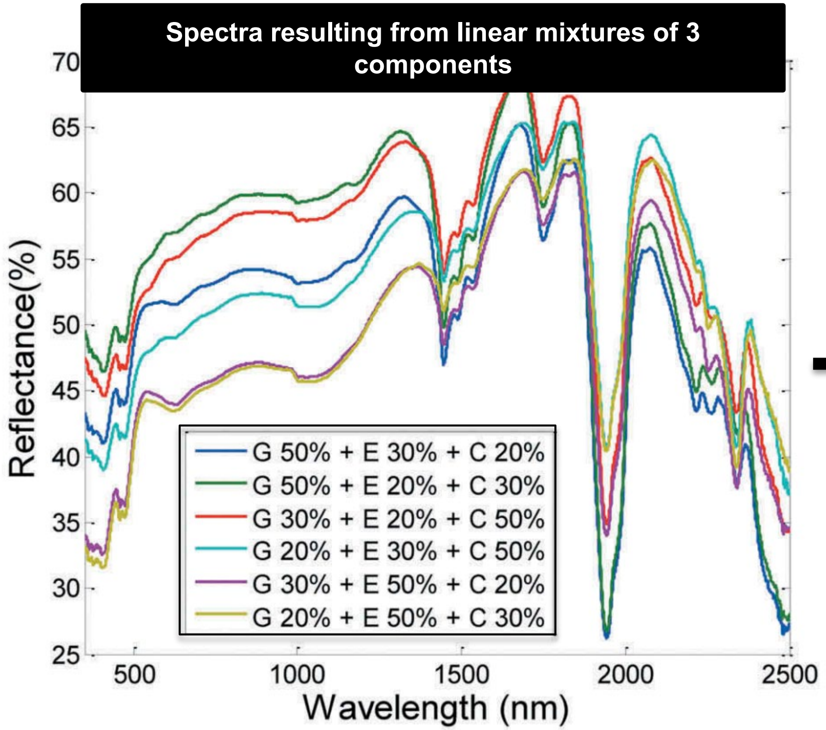
**more accurately described as a semi-quantified method, due to the potential for spectrally inactive mineral components present in the sample.*

Credit: Corescan

Spectral unmixing – estimating abundance

- A spectrum comprised of more than one mineral is very common when analyzing natural geological materials.
- Unmixing algorithms do exist for simple binary (or less commonly ternary) mixtures of two end-members
  - In an ideal case, an unknown spectrum could be decomposed into its constituent endmembers allowing for a quantification of the proportion of each mineral/endmember present in the pixel (e.g., spectral unmixing).





Once the theoretical spectra have been generated by a spectral mixing algorithm, these results can then be compared to unknown spectra to estimate mineral proportions of the same end-member components

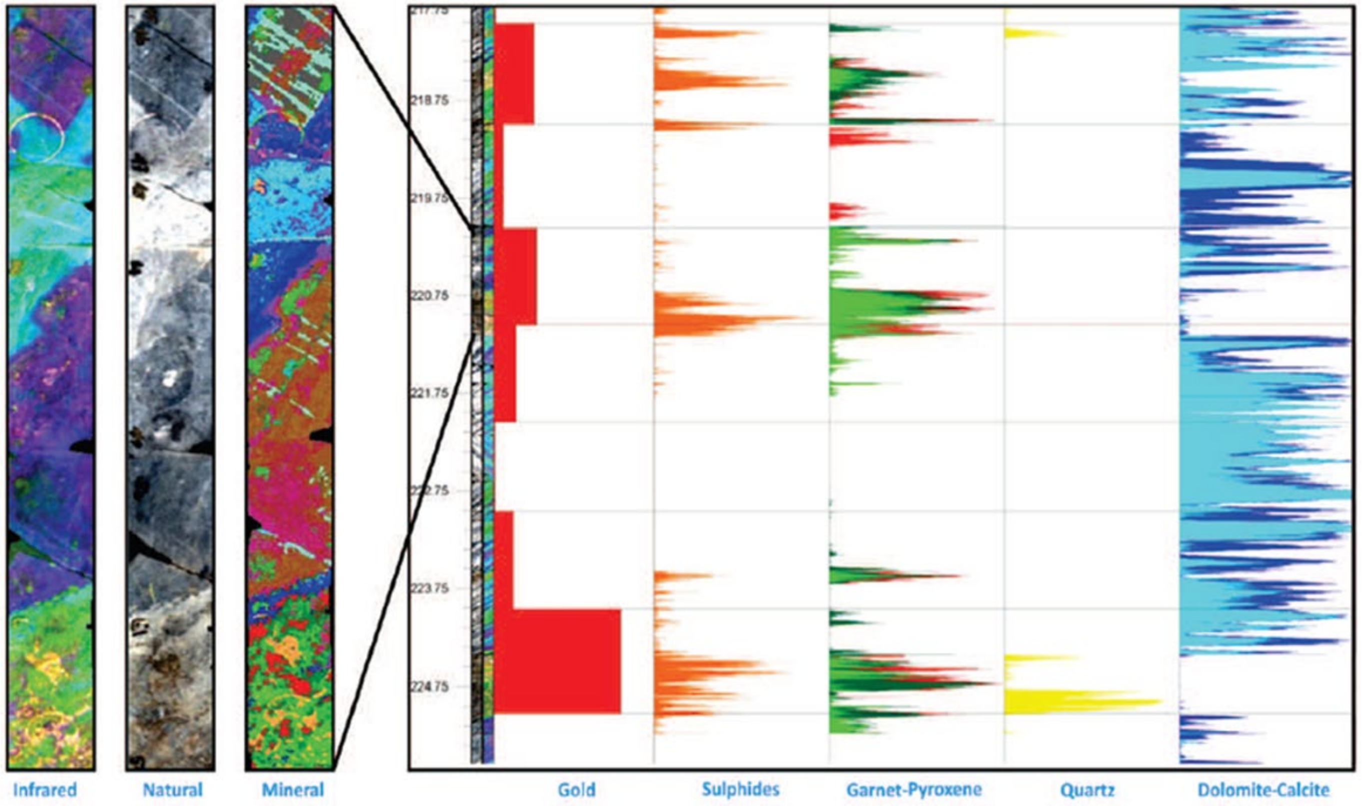
(Zhao and Zhao, 2019)

Note: This methodology relies on many assumptions that may not be applicable in natural geological materials. Be aware!

Credit: Corescan

How It's Being Used (and Misused)

- Vendors present pixel-perfect mineral abundance maps.
- No uncertainty, no residual maps, no diagnostics.
- Users interpret colors as quantitative truth.



Specim

The underlying statistics

…why we care.

LKI Consulting

Why this matters: linking mineralogy, texture and met

LKI Consulting

- Processing energy costs (especially grinding) dominate mine operating expenses; small improvements in hardness and/or recovery prediction have huge financial and environmental implications.
- Texture affects liberation, recovery and mill throughput, yet these are rarely quantified in block models.
- Can quantitative rock textures extracted from hyperspectral imagery improve predictions of key geometallurgical parameters, specifically comminution hardness (SPI®) and copper flotation recovery?

Model descriptions.

Model	Predictive variables datasets	Predicted variable
HG	Logging Geochemistry	SPI®
HGM	Logging Geochemistry Hyperspectral mineralogy	SPI®
HGT	Logging Geochemistry Textural clustering	SPI®
HGMT	Logging Geochemistry Hyperspectral mineralogy Textural clustering	SPI®
RG	Logging Geochemistry	Cu flotation recovery
RGM	Logging Geochemistry Hyperspectral mineralogy	Cu flotation recovery
RGT	Logging Geochemistry Textural clustering	Cu flotation recovery
RGMT	Logging Geochemistry Hyperspectral mineralogy Textural clustering	Cu flotation recovery

Merrill et al., 2023

Why getting the math wrong breaks this

- If spectra are “unmixed” linearly into artificial mineral proportions, closure effects and noise redistribution would destroy the textural signal.
- Closed datasets (where components sum to 1 or 100%) create false negative correlations and distort the statistical space, turning geology into geometry.

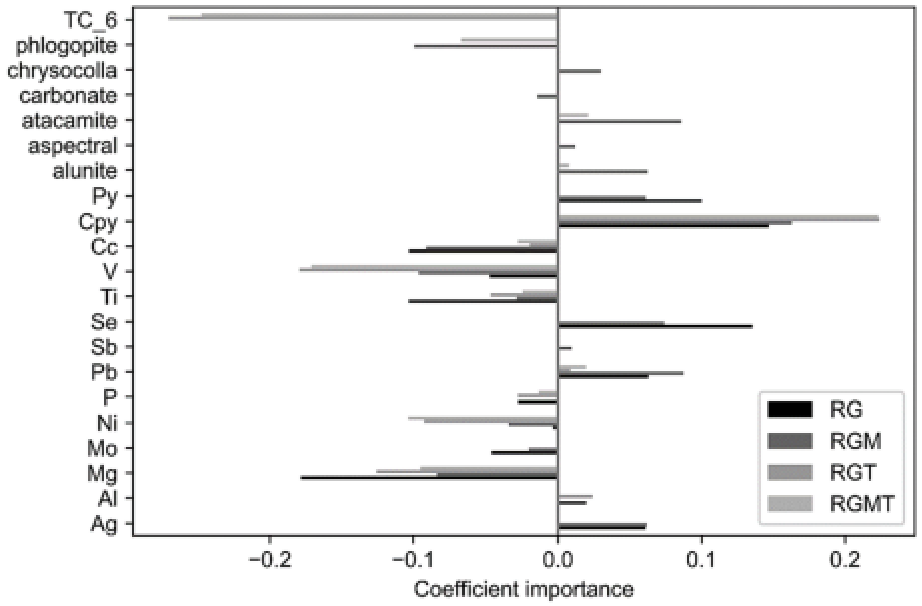
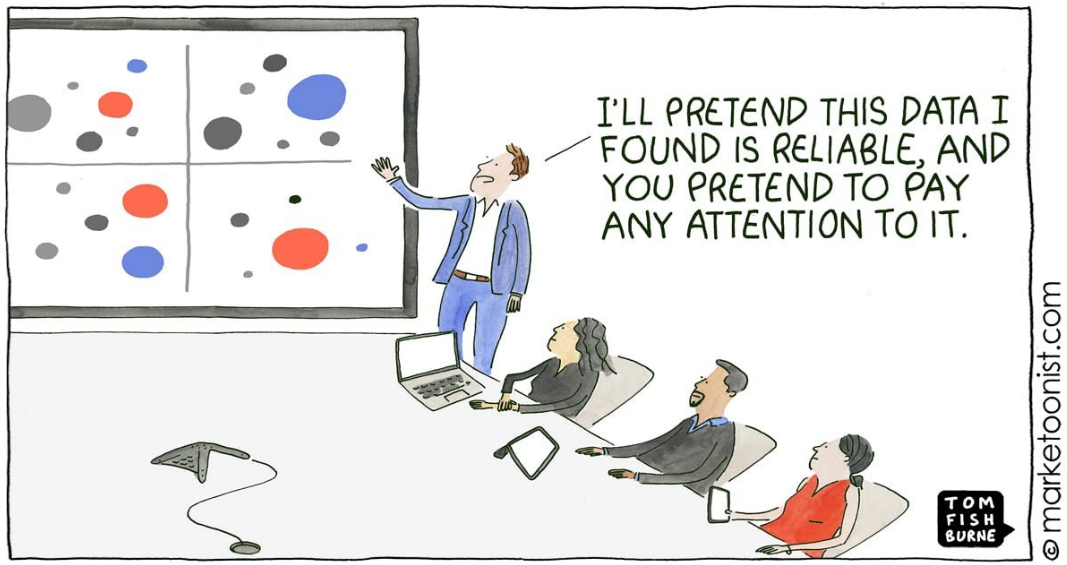


Fig. 13. Recovery prediction model coefficient comparison ($\alpha = 0.075$).

Merrill et al., 2023

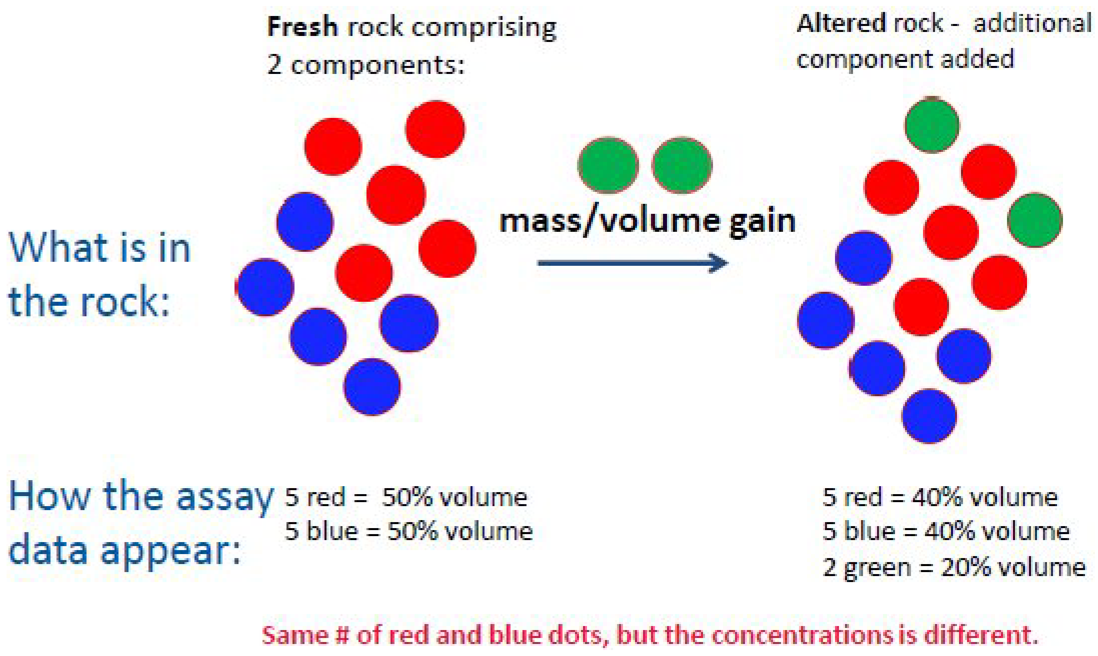
What is compositional data

- Compositional data are **vectors of proportions** that represent parts of a whole, for example, mineral percentages or oxide weight %.
- The defining property is that all components are **relative** and constrained to sum to a constant (often 1 or 100%).
- Because of that constraint, information lies in the **ratios between components**, not their absolute values.



Closure

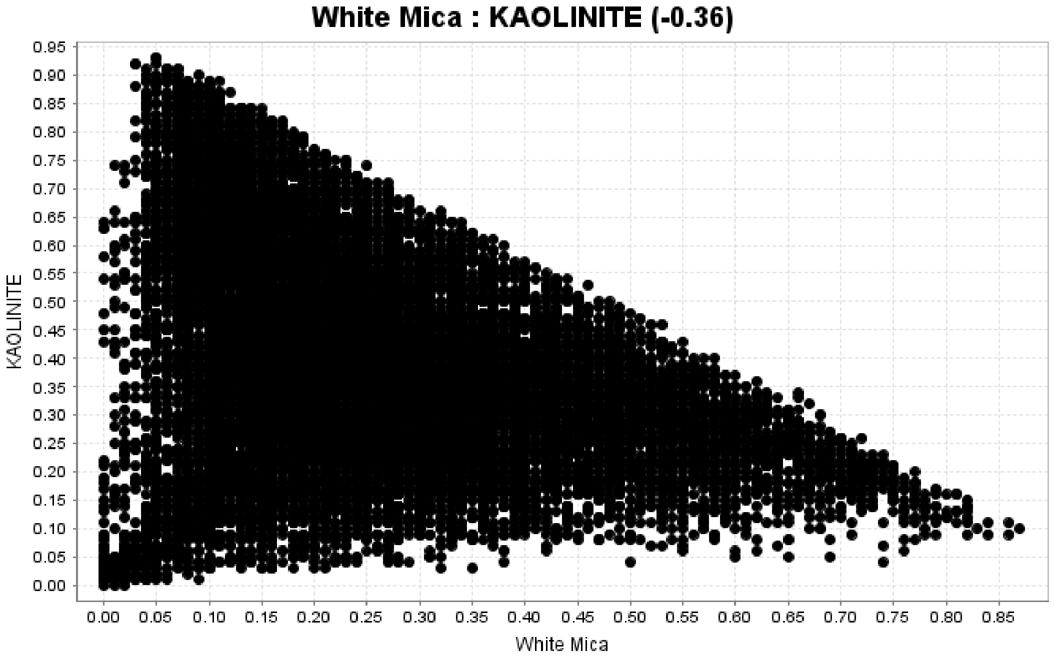
- Closure refers to this **constant-sum constraint** that forces dependence among variables.
- When one component increases, at least one other must decrease, even if they are unrelated in reality.
- This mathematical artifact introduces **spurious correlations** and distorts statistical distances, making standard methods like PCA or regression invalid unless the data are transformed (e.g., using log-ratio transformations such as CLR or ILR).



Closure Effects and Compositional Pitfalls

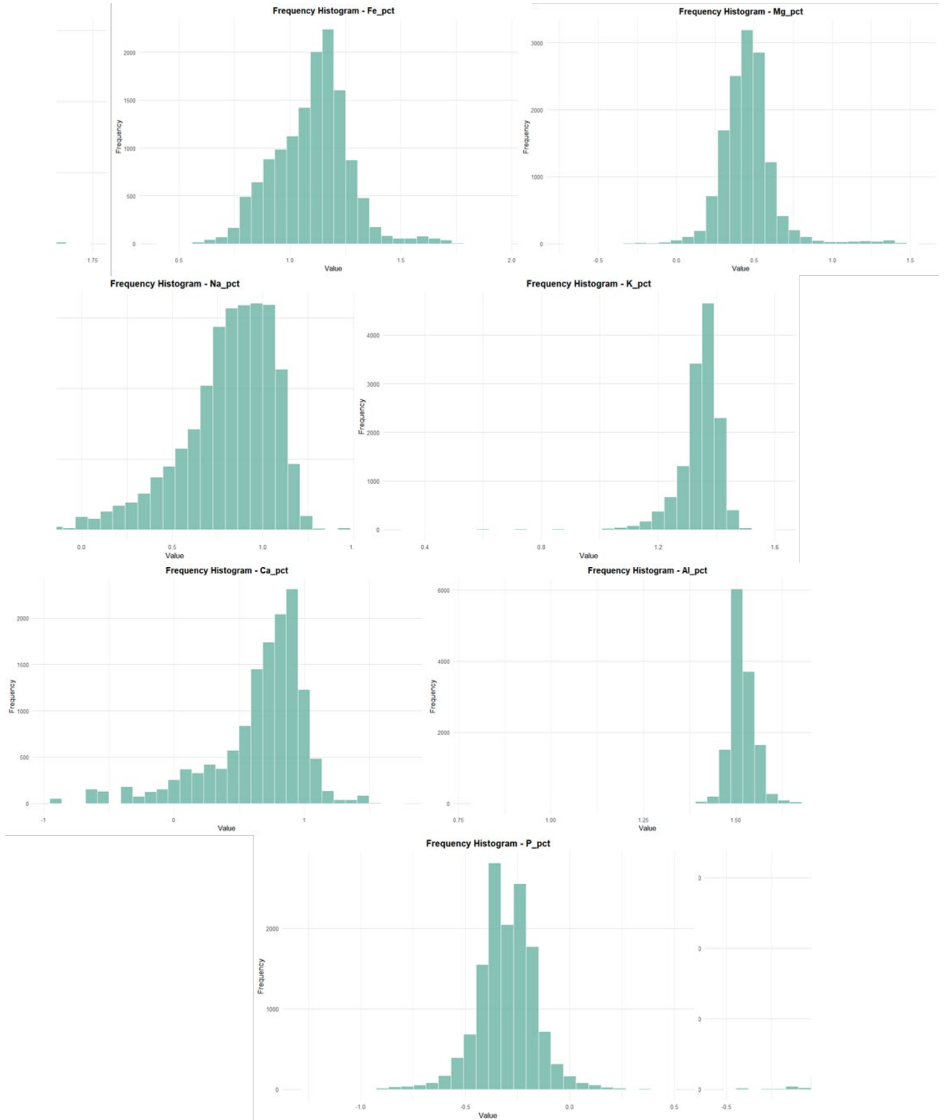
Spectral data are compositional data

- Instead of leaving hyperspectral data open, decisions have been made across the industry to deliver outputs with closed compositions (i.e., sum = 100%).
- This, in turn, forces correlations that may not actually exist.
- Ignoring the principles of compositional data analysis makes any statistical interpretation invalid.



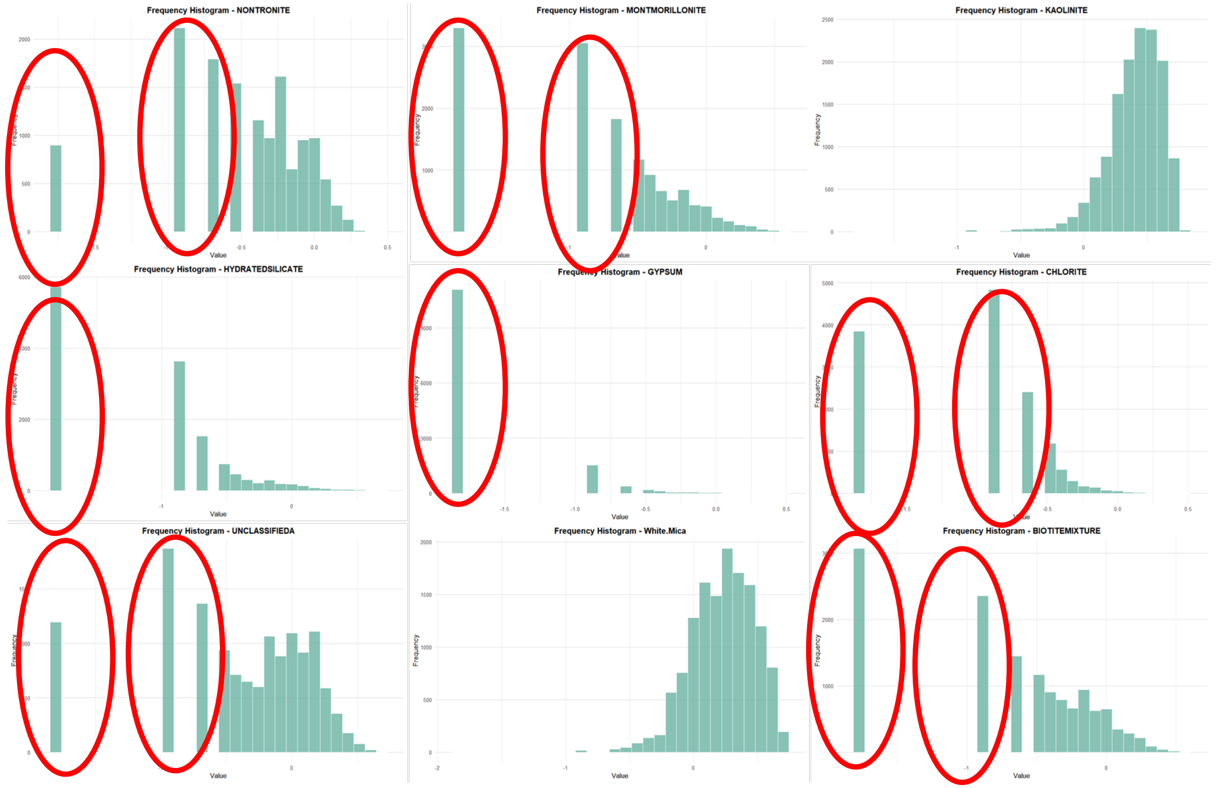
This white mica v kaolinite scatterplot is a perfect example of data that has a large closure issue.

Normalization of compositional data



- Histograms of selected 4-acid ICP-MS elemental data after transformation.
- The distributions are approximately Gaussian, with symmetric shapes and minimal tail effects.
- This behavior demonstrates the ability of this dataset to normalize cleanly and provide statistically robust input for methods such as UMAP, PCA, t-SNE, etc.

Normalization of compositional data



- Histograms of selected hyperspectral mineral abundances after transformation.
- Unlike the data shown in the prior slide, these distributions remain highly irregular, exhibiting multimodality, heavy skew, and artificial spikes at discrete values (e.g., –1, 0).
- These features are symptomatic of forced unmixing algorithms, where closure effects generate **spurious negative correlations** and **noise** in minor components.
- The prevalence of discretized peaks, long tails, and lack of symmetry indicate that the data cannot be normalized effectively and are therefore unfit for quantitative multivariate approaches that require continuous, well-behaved variables.

Why artifacts break statistical models

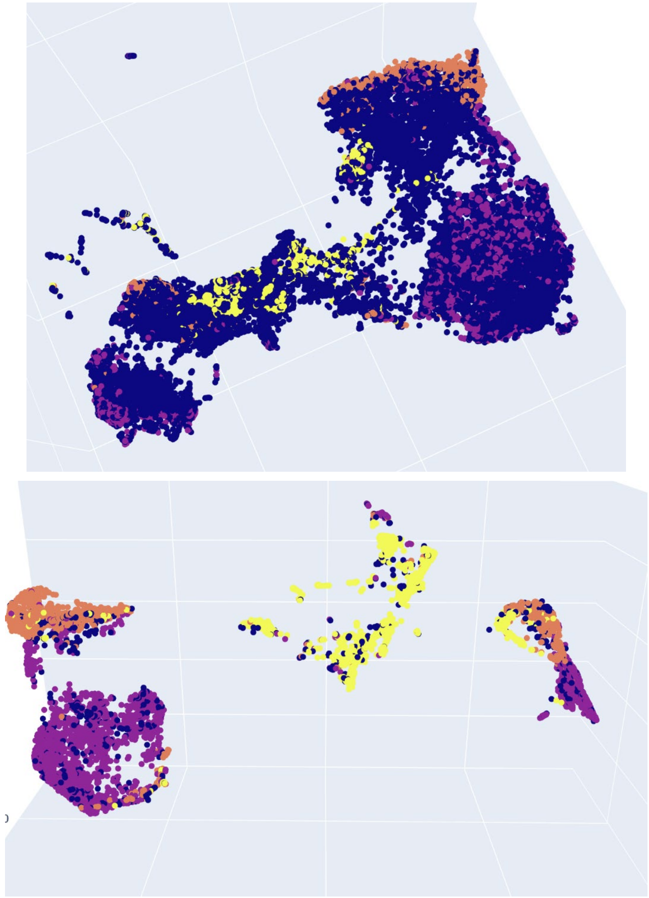
LKI Consulting

Using ML to create geomet or alteration models

- UMAP: local neighborhoods disrupted by noise/closure
- DBSCAN: density clusters reflect algorithmic artifacts, not geology
- PCA: false loadings due to negative correlations
- Result: apparent “zones” or trends that arise from how the data are scaled and constrained, not from real mineral variation.

Using ML to create predictive or proxy models

- Predictive models assume valid, independent inputs.
- Unmixed data violate independence, thereby creating spurious proxies.
- Result: degraded model fidelity, misplaced confidence



Top: example of a dataset with noise/artefacts.
Bottom: the same dataset with the artefacts removed. stackoverflow

Building a geomet domain (not alteration) model

Purpose

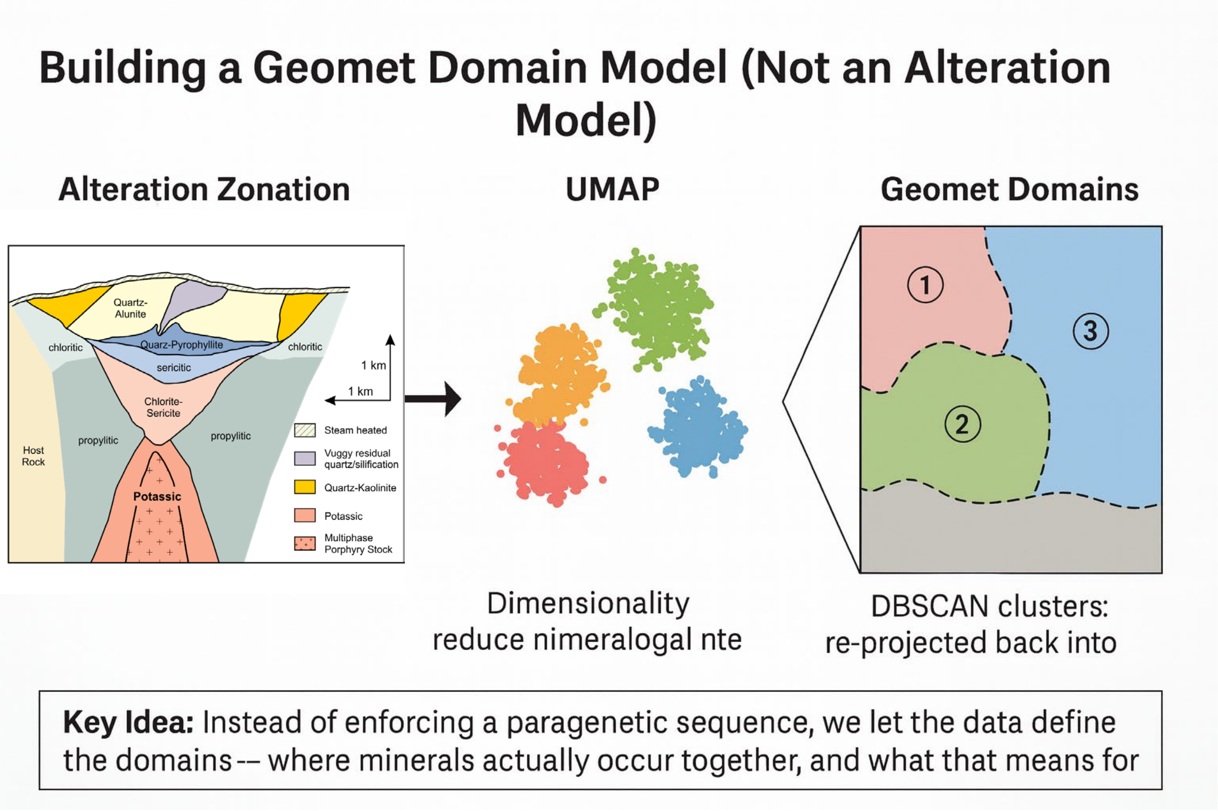
- A geomet domain model defines areas of consistent mineral co-occurrence and shared processing behavior, not classical alteration zonation.
- The goal is to guide representative metallurgical testwork and predict processing response, not to reconstruct fluid evolution.

Conceptual Difference

- **Alteration model:** built from paragenetic relationships and spatial zonation (e.g., K-feldspar/biotite → white mica → chlorite → montmorillonite).
- **Geomet domain model:** built from statistical relationships among minerals that *co-exist,* regardless of their relative age.
- Example: late argillic overprint within a "potassic" domain is not noise, but rather it is metallurgically meaningful because it affects comminution and recovery.

Building a geomet domain model (example workflow)

- Start with pixel- or interval-scale mineral abundances or probabilities (hyperspectral, XRD, assay proxies).
- Reduce dimensionality using UMAP. This preserves local mineralogical relationships without assuming linearity.
- Apply DBSCAN clustering to group samples by co-occurrence patterns, not by depth or mapped zones.
- Each cluster = a geomet domain representing unique mineralogical assemblages and expected metallurgical behavior.
- Validate with metallurgical testwork (e.g., SPI®, recovery) and adjust sampling strategy accordingly.



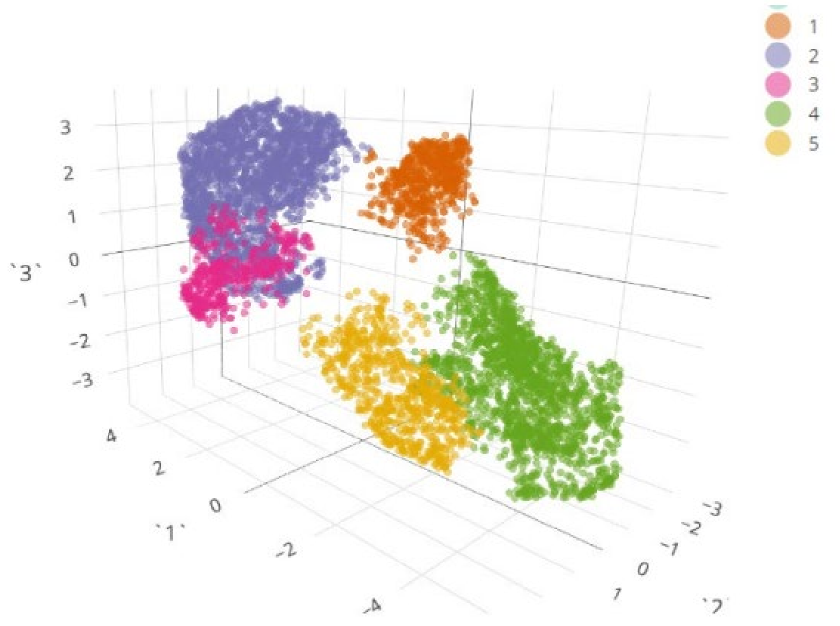
When bad math breaks good data

The Reality Check

- When we tried to integrate the hyperspectral data into the geomet domain model, unmixing artifacts destroyed the UMAP structure.
- Pixels “closed” to 100% mineral proportions introduced false relationships and density distortions…. UMAP clustered noise, not mineralogy.
- DBSCAN then grouped the noise into “domains” that did not exist.

What Actually Worked

- The only stable, geologically meaningful models were those anchored by geochemistry.
- Any hyperspectral variable below ~10% abundance was effectively noise and had to be excluded.
- Geochemistry governed the domains, not the spectral data.



UMAP and DBScan of a dataset using major minerals only and underpinned by geochemistry.

The outsized role of spectrally active minerals in linear unmixing

LKI Consulting

The Spectral Bias

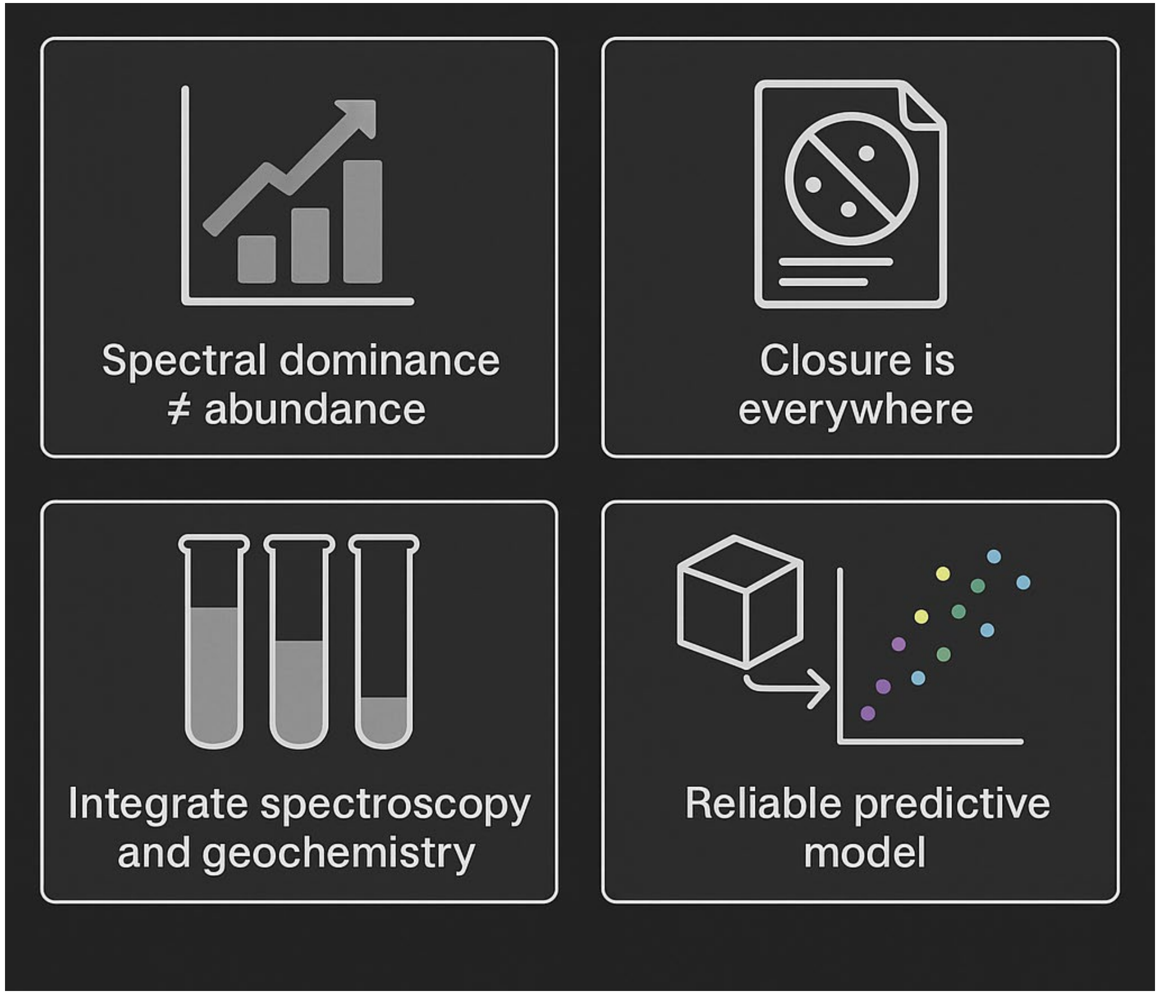
- The “major minerals” that dominated the model were not necessarily the most abundant in the rock.
- Most spectral algorithms weight by reflectance, not mass fraction… the ‘brightest’ or highest albedo minerals win.
- Minerals like kaolinite, white mica, and chlorite are well known to overwhelm the spectral response, drowning out ‘darker’ or optically subtle phases (e.g., biotite, amphibole, sulfides).

The Consequence

- The hyperspectral dataset became optically biased, not mineralogically representative.
- Those high-albedo minerals dictated cluster structure, steering the UMAP and DBSCAN outputs even when they were not necessarily volumetrically significant.
- In effect, the physics of light replaced the chemistry of rocks as the control on the model.

Reclaiming rigor

Path forward



- There is extreme value in hyperspectral data, however, as an industry we need to make it more clear to our clients that it is quantitative and, in some cases, even semi-quantitative data.
- We are still forcing non-linear spectral behavior into linear models. In those same models we are not always correcting for the albedo.
- Most percent mineral datasets are compositional and not statistical. Leaving it in the hands of the end users without educating them will destroy our industry.
- We need to ground hyperspectral data in mineralogy, geochemistry, and math that actually describes *rocks* – not just the minerals that comprise them.
- What worked statistically with point spectral data is not always fit for purpose for core imaging data that is an order of magnitude larger in both volume and relative to absolute spatial coverage.