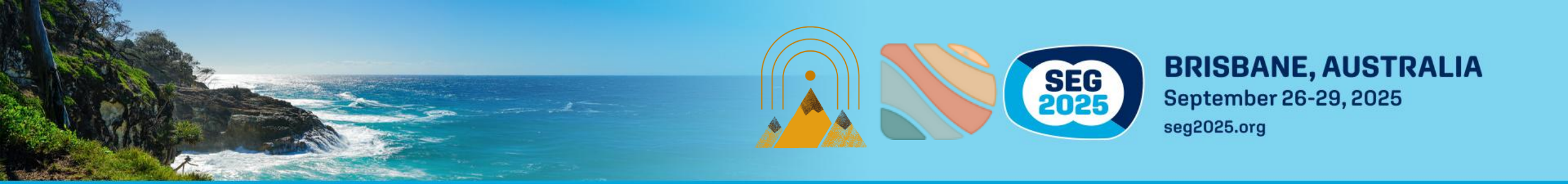


Data Before Algorithms: An Open-Source Approach to Legacy Geochemical Method Categorization

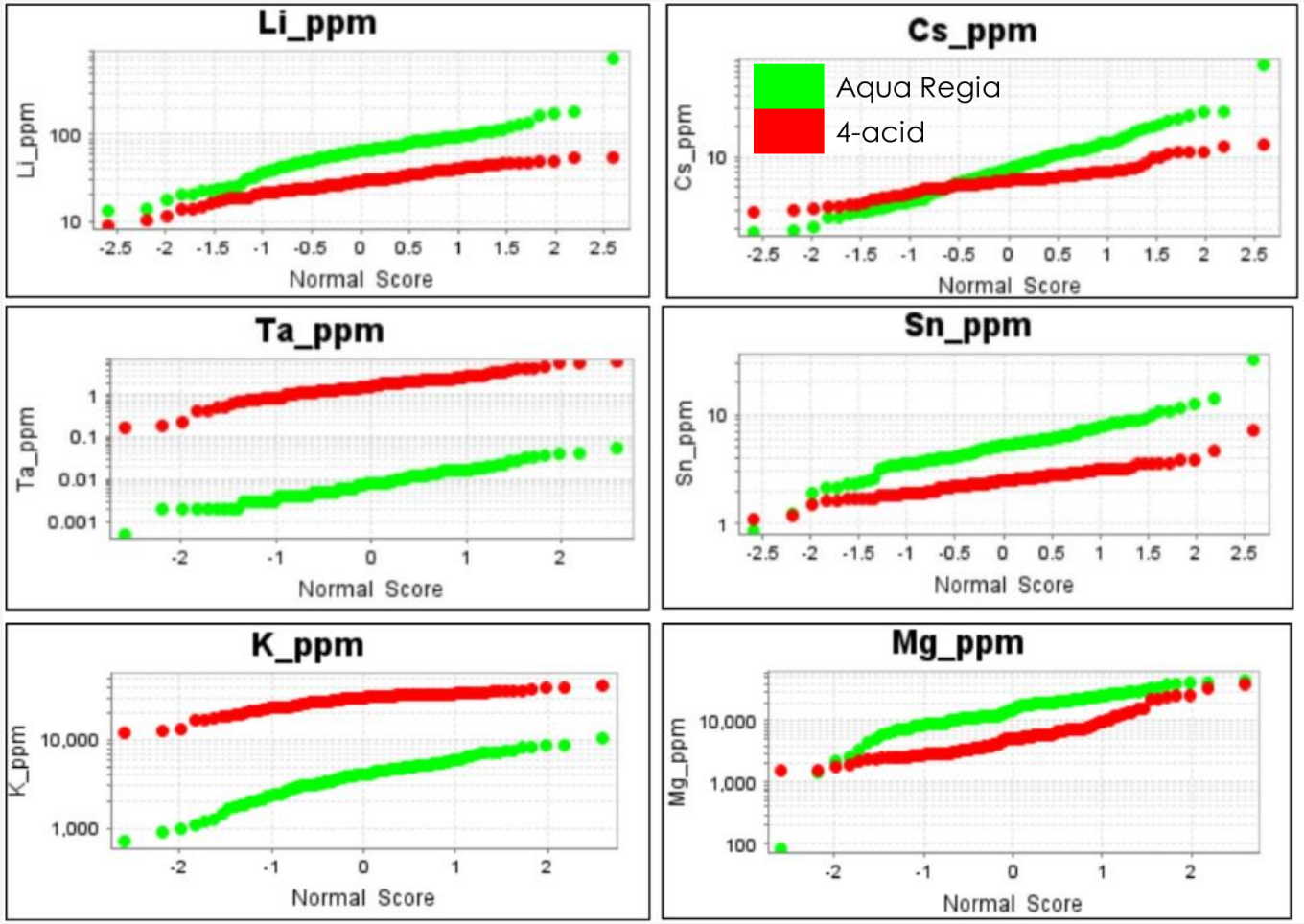
Sam Scher, LKI Consulting Inc. | sscher@lkiconsulting.com
Tom Carmichael, Datarock | tomcarmichael@datarock.com.au

Data acknowledgement

- The authors gratefully acknowledge Fireweed Metals Corp. for making their dataset publicly accessible, which enabled us to demonstrate the application of the code presented herein.
- Fireweed Metals had no involvement in this work, and all interpretations and conclusions are solely those of the authors.

Mixed Methods = Mixed Messages

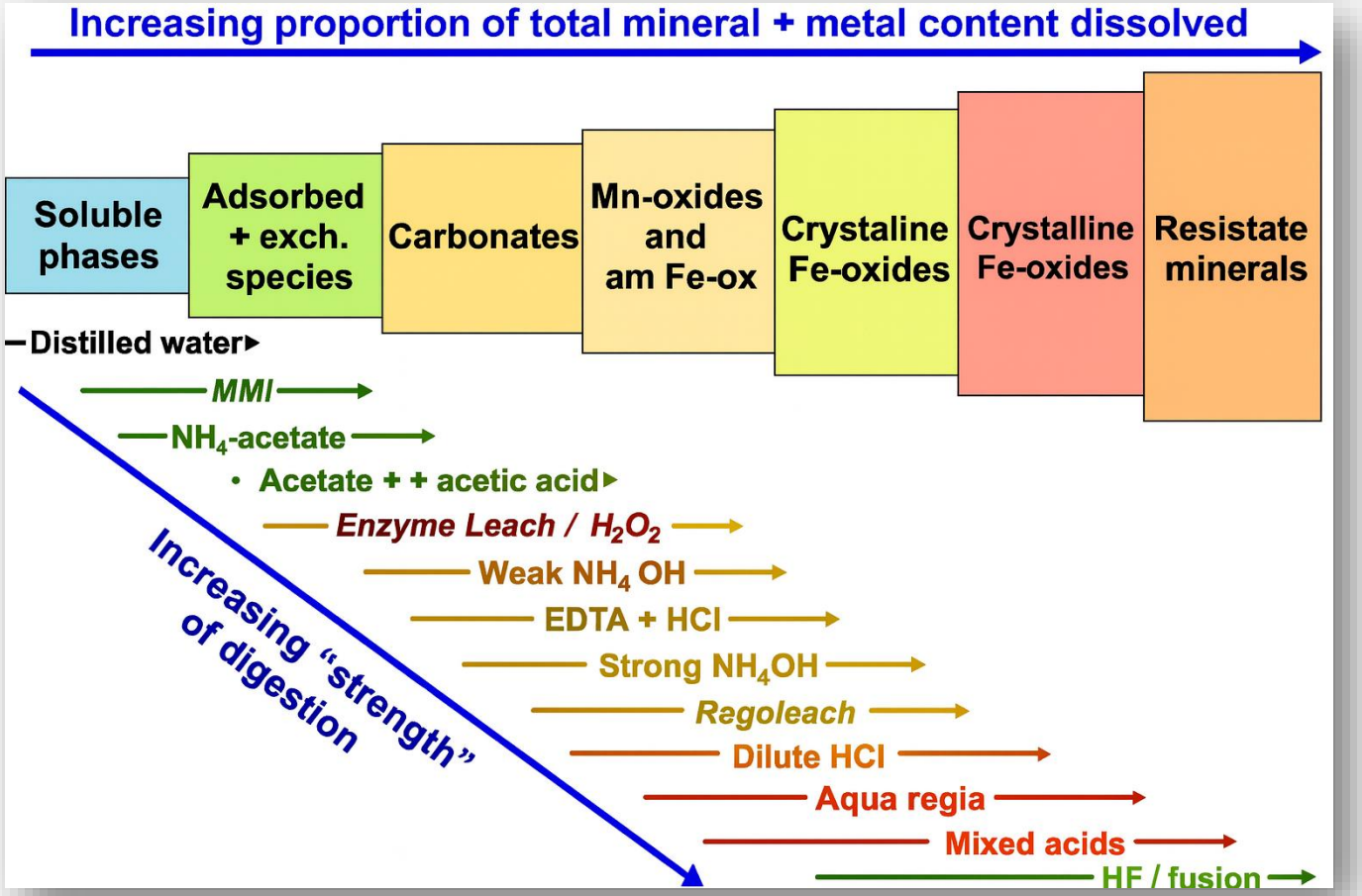
- Legacy geochemical data is riddled with hidden method changes.
- Your maps, models, and interpretations are meaningless if this not addressed.



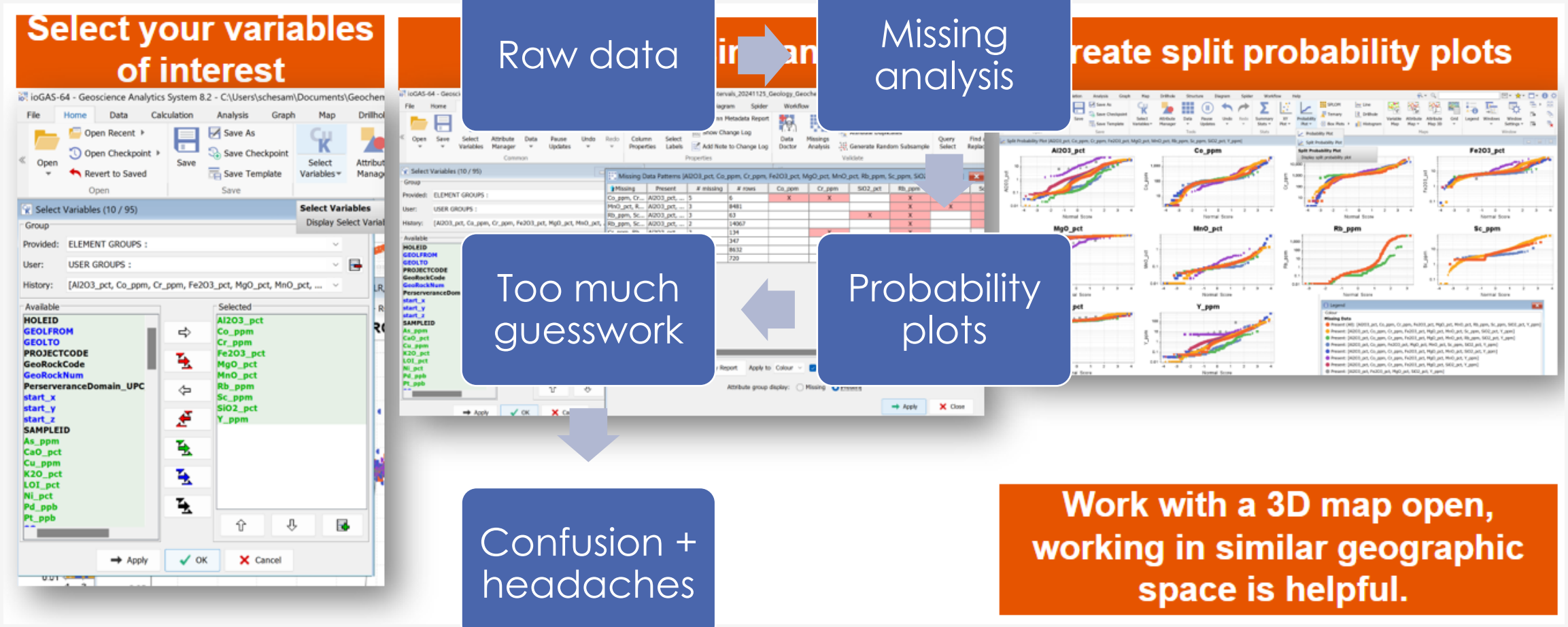
New Age Exploration Ltd., 2024

The whole package matters

- Reported values are not universal truths.
- Different digestions target different chemical features – the same rock can yield different elemental values.
- Detection limits (finish) also imprint distinct data structure signatures.
- Sample prep and varying sample mass affect representativity, shifting element distributions and ratios.

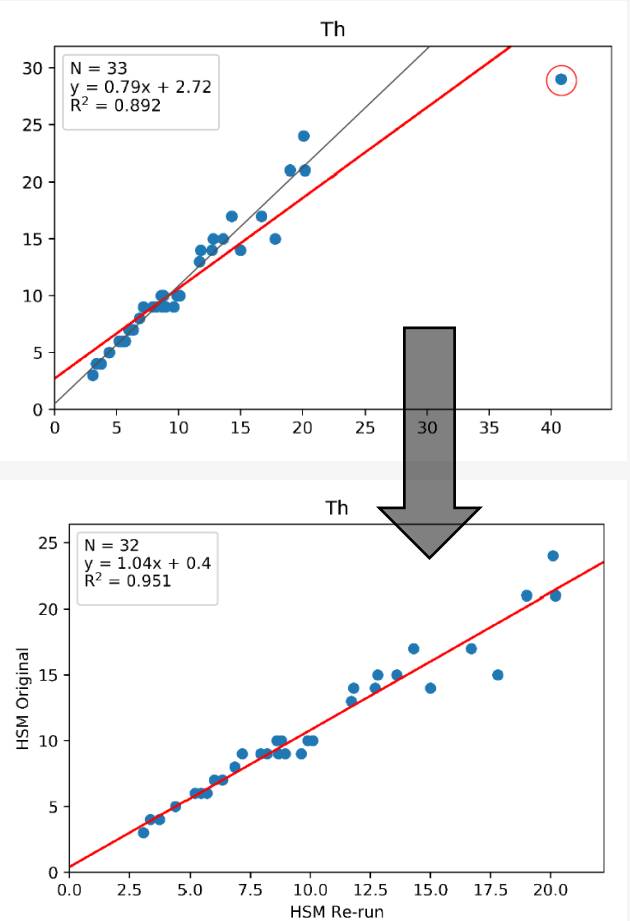


Current modus operandi



Why this matters: data levelling

- Common after-the-fact strategy: force datasets to match by scaling or shifting values.
- **Issues with typical approaches:**
  - Ignore lab drift, batch effects, and detection limits.
  - Treat elements independently → breaks covariance.
  - Mask real geological variability → creates false homogeneity.
- **Statistical conditioning**
  - Recognizes we cannot change sampling or digestion, but we can treat data rigorously.
  - To do this we require knowledge of the metadata of our datasets, even if it is not preserved...



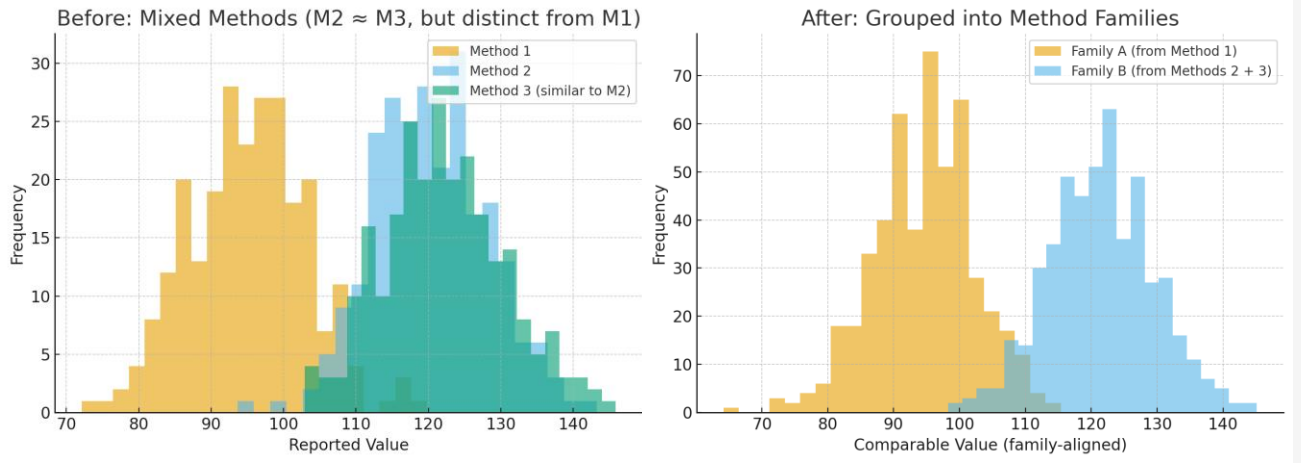
Main & Champion, 2019

"If there is one thing the history of evolution has taught us is that life will not be contained, life breaks free, it expands to new territories and crashes through barriers, painfully, maybe even dangerously, but, uh... well, there it is.“

–Dr. Ian Malcolm

When a geochemist postulates the framework of an idea to mathematical geologist…

- **Geochemist:**
  - Mixing analytical methods confuses geology with process artifacts that leads to false trends, unstable conceptual or statistical models, and unreliable interpretations.
- **Mathematical geologist:**
  - Grouping data into *method families* does not guarantee that everything about the families in the data are the same, but they are sufficiently similar to use together.
  - This restores comparability, stabilizes ratios, strengthens QC, and improves downstream results.



Left (before): Mixed methods create offset distributions and will result in unstable interpretations.
Right (after): Grouped into *method families* that while not identical are close enough to support robust statistical investigation.

Goal: make mixed methods comparable

Inputs

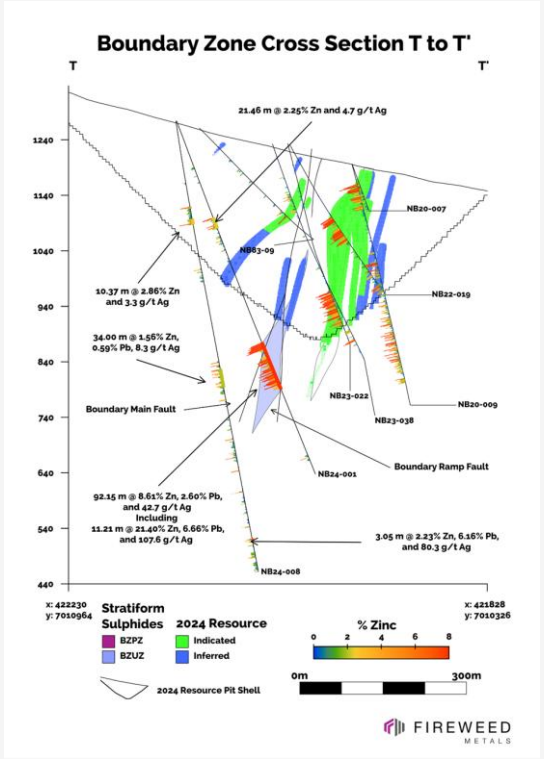
- Legacy assays with mixed/unknown methods.

Outputs

- A method ‘family’, some degree of confidence.

Guardrails of use

- We are not guessing the lab code.
- We will not change or edit the initial dataset to ‘correct’ for anything.
- Note that both of the guardrails are possible if that is of interest.



Ultimately: are these values for the same element from the same database the same? If not, how can I reconstruct the metadata in order to progress my analysis?

The starting point based on data structure

Primary

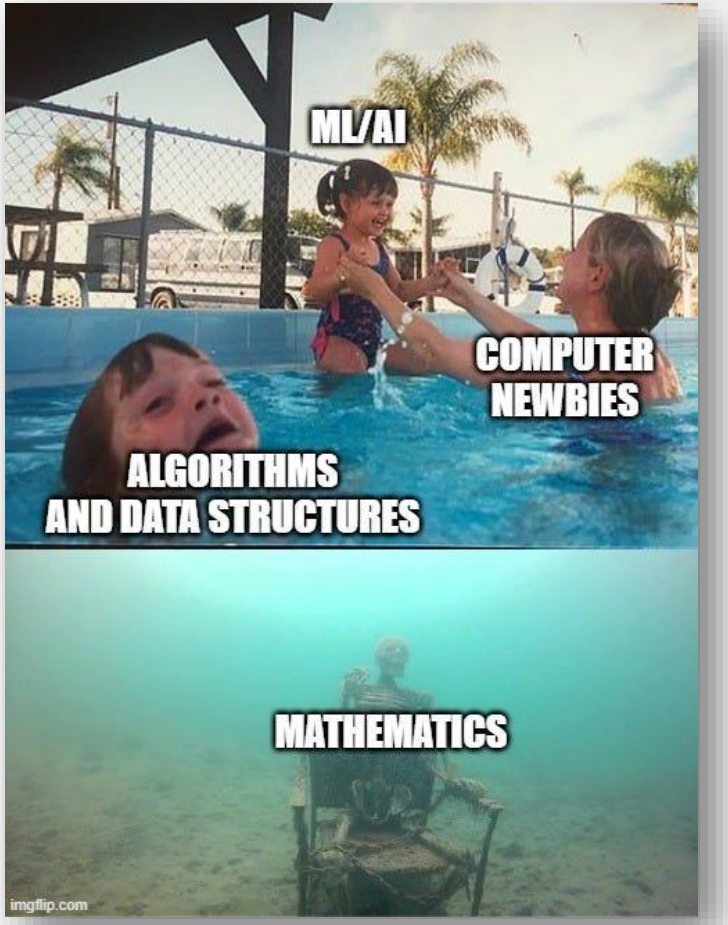
- Suite coverage: which elements are appearing together and which are absent.

Secondary

- Detection limit: detection limit presence/absence?
- Data precision: what is the specificity of the returned data?

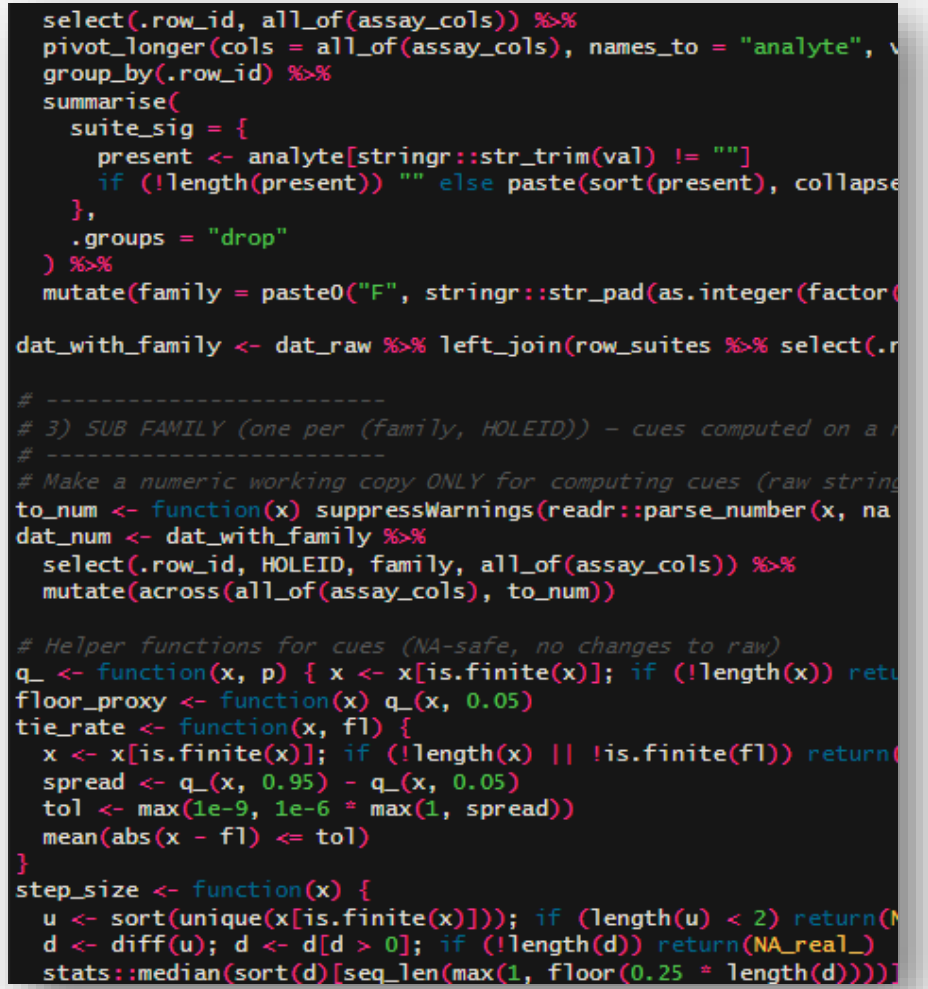
Tertiary

- Recovery fingerprints: robust ratios?
- Correlation shapes: within batch structure?



Workflow overview

- Step 1: Identify obvious splits.
- Step 2: Build feature vectors (e.g., element missingness, detection limits, precision).
- Step 3: Quantify distances and group candidates.
- Step 4: Assign families with confidence.
- Step 5: Validate with QC plots and using legacy datasets with known analytical methods.

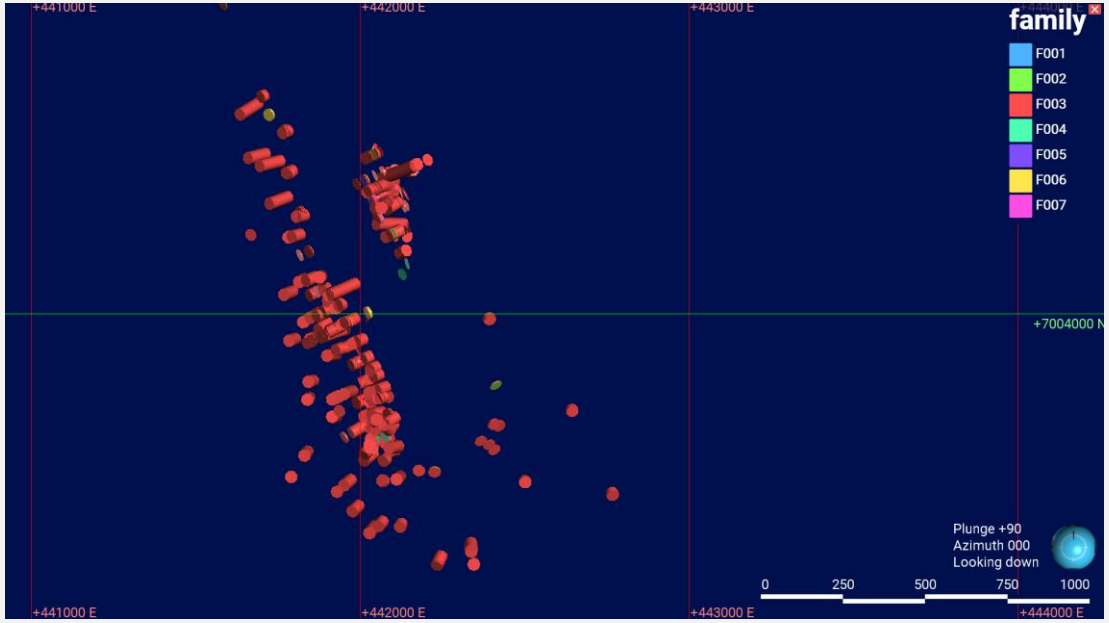


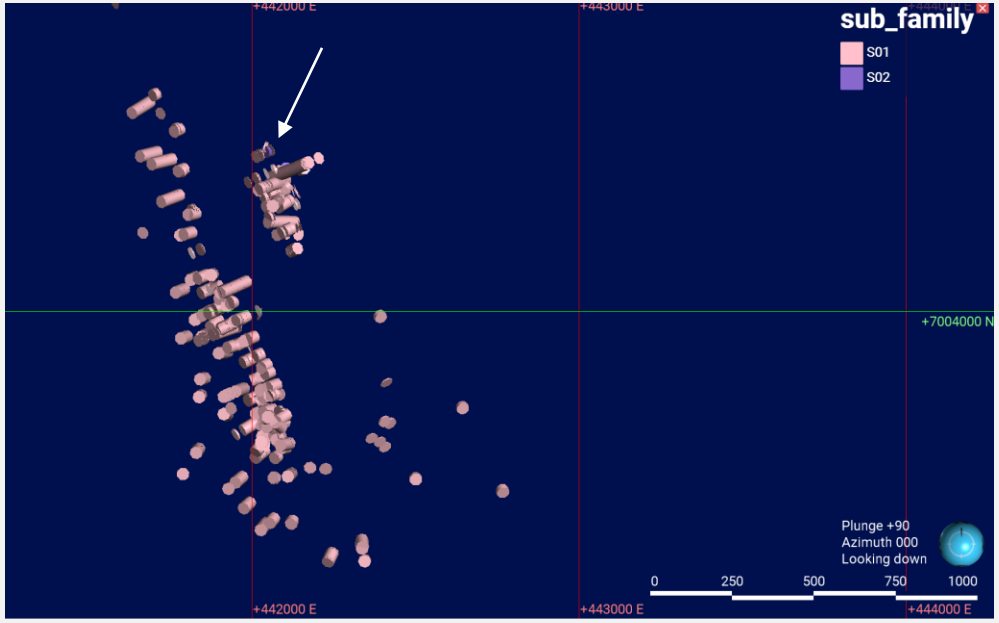
```
  select(.row_id, all_of(assay_cols)) %>%
  pivot_longer(cols = all_of(assay_cols), names_to = "analyte", v
  group_by(.row_id) %>%
  summarise(
    suite_sig = {
      present <- analyte[stringr::str_trim(val) != ""]
      if (!length(present)) "" else paste(sort(present), collapse
    },
    .groups = "drop"
  ) %>%
  mutate(family = paste0("F", stringr::str_pad(as.integer(factor(

dat_with_family <- dat_raw %>% left_join(row_suites %>% select(.r

# ------------------------
# 3) SUB FAMILY (one per (family, HOLEID)) - cues computed on a r
# ------------------------
# Make a numeric working copy ONLY for computing cues (raw string
to_num <- function(x) suppressWarnings(readr::parse_number(x, na
dat_num <- dat_with_family %>%
  select(.row_id, HOLEID, family, all_of(assay_cols)) %>%
  mutate(across(all_of(assay_cols), to_num))

# Helper functions for cues (NA-safe, no changes to raw)
q_ <- function(x, p) { x <- x[is.finite(x)]; if (!length(x)) retu
floor_proxy <- function(x) q_(x, 0.05)
tie_rate <- function(x, fl) {
  x <- x[is.finite(x)]; if (!length(x) || !is.finite(fl)) return(
  spread <- q_(x, 0.95) - q_(x, 0.05)
  tol <- max(1e-9, 1e-6 * max(1, spread))
  mean(abs(x - fl) <= tol)
}
step_size <- function(x) {
  u <- sort(unique(x[is.finite(x)])); if (length(u) < 2) return(N
  d <- diff(u); d <- d[d > 0]; if (!length(d)) return(NA_real_)
  stats::median(sort(d)[seq_len(max(1, floor(0.25 * length(d))))]
```

Families v Sub-Families

Families have the same row-wise exact element suite.

Sub-families are data-driven groups within a family, clustered on detection limits, step size, precision, and variance signatures.

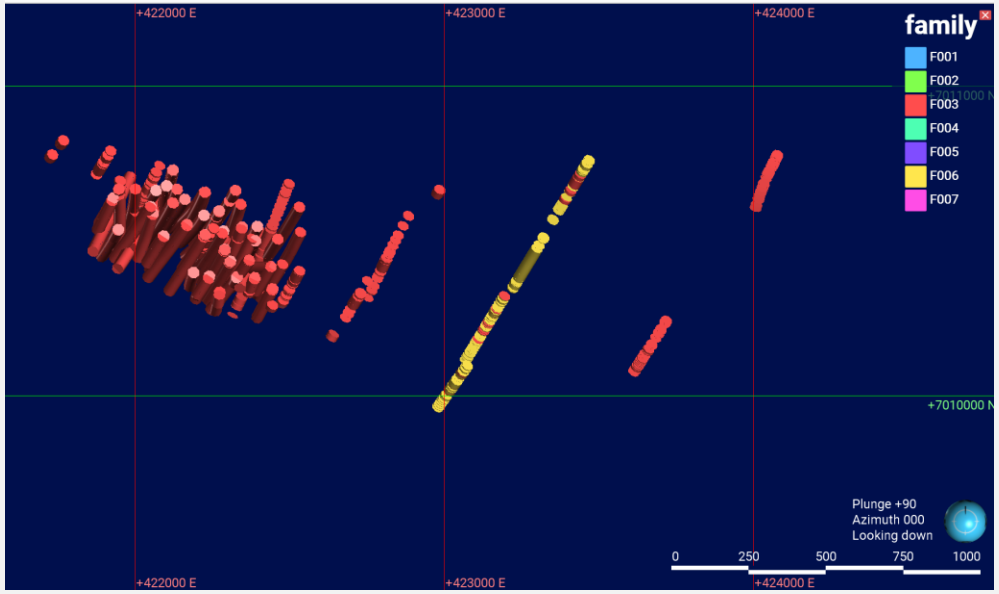


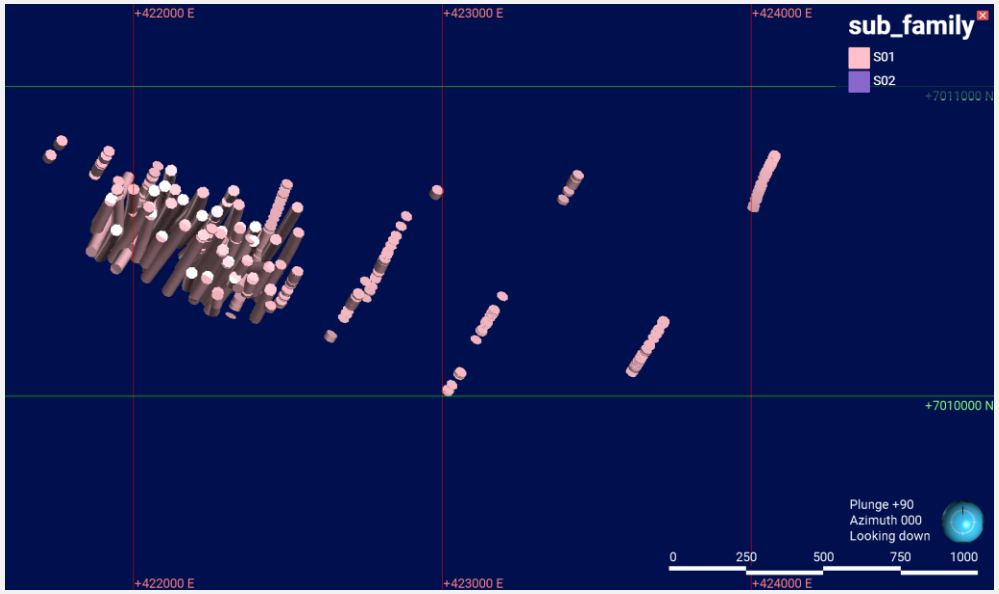


Families v Sub-Families

Families have the same row-wise exact element suite.

Sub-families are data-driven groups within a family, clustered on detection limits, step size, precision, and variance signatures.





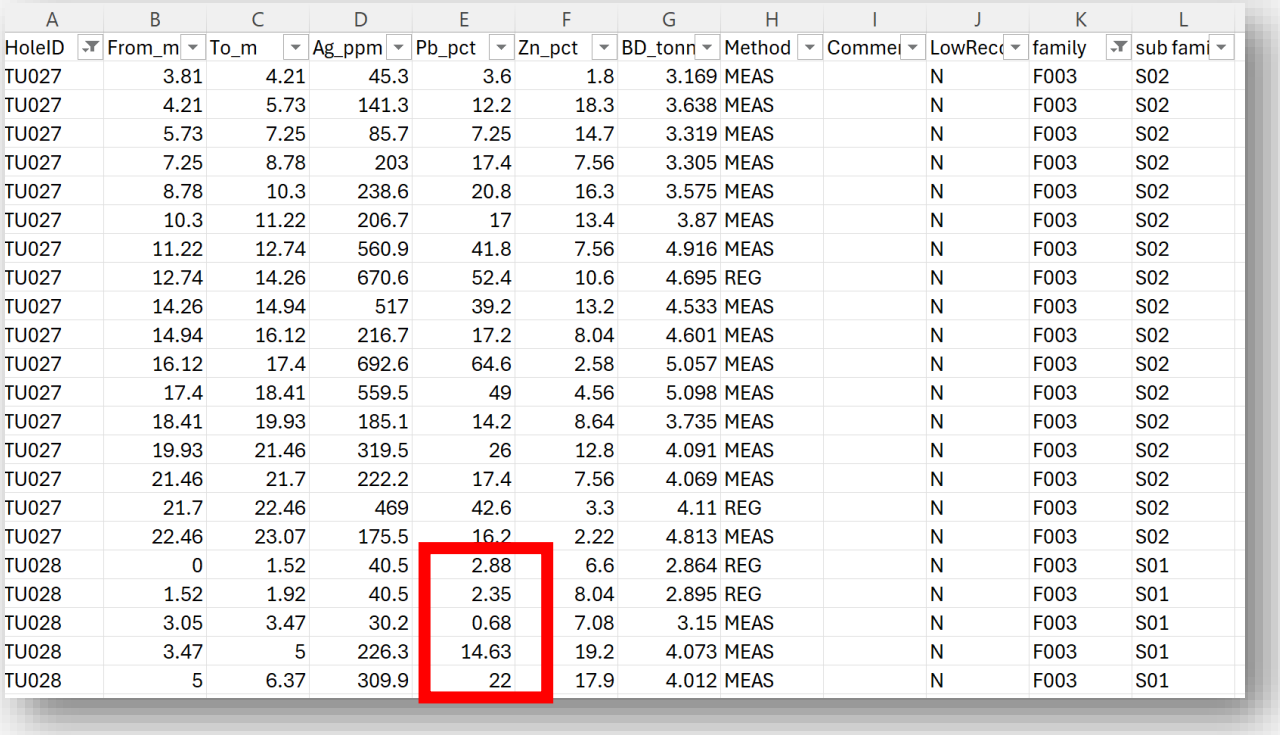
Outputs: families

- Families group rows with the same element suite.
- Breaks datasets into constituent analytical parts.
- Equivalent to a structured missing data analysis.

	A	B	C	D	E	F	G
1	family	rows_in_far	holes_in_fa	suite_sig	Ag_ppm	Pb_pct	Zn_pct
2	F001	12	6	Ag_ppm	1	0	0
3	F002	26	9	Ag_ppm|Pb_pct	1	1	0
4	F003	31379	410	Ag_ppm|Pb_pct|Zn_pct	1	1	1
5	F004	16	11	Ag_ppm|Zn_pct	1	0	1
6	F005	7	5	Pb_pct	0	1	0
7	F006	1047	55	Pb_pct|Zn_pct	0	1	1
8	F007	22	8	Zn_pct	0	0	1

Outputs: sub-families

- Families are a missing analysis grouped by element suite.
- Sub-families dig deeper and highlight variation in precision, detection limits, and step signatures.
- This enables geochemists, mathematical geologists, data scientists, etc. to:
  - Spot batch effects or campaign changes hidden within a family.
  - Separate out “noisy” or lower-quality subsets.
  - Improve confidence in downstream statistics and models.
- Overall, the sub-family export deliverables a more accurate picture of your data improving downstream interpretations.



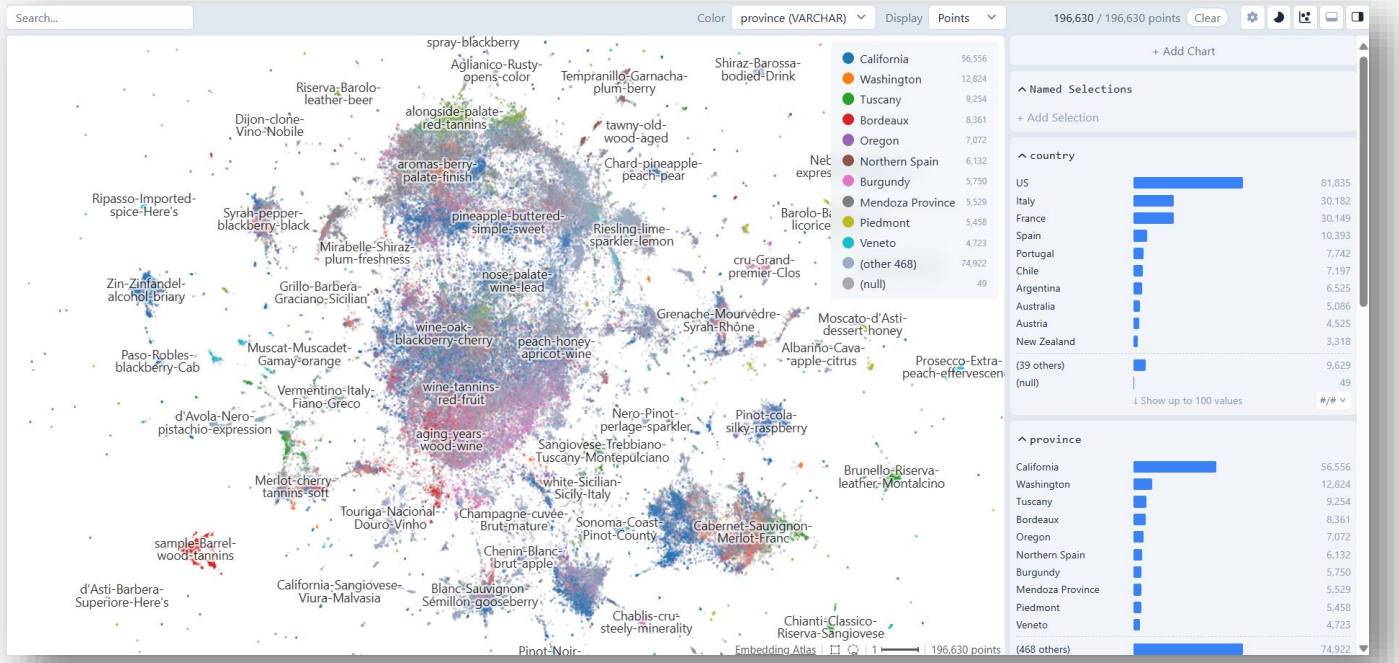
A	B	C	D	E	F	G	H	I	J	K	L
HoleID	From_m	To_m	Ag_ppm	Pb_pct	Zn_pct	BD_tonn	Method	Comme	LowRec	family	sub fami
TU027	3.81	4.21	45.3	3.6	1.8	3.169	MEAS		N	F003	S02
TU027	4.21	5.73	141.3	12.2	18.3	3.638	MEAS		N	F003	S02
TU027	5.73	7.25	85.7	7.25	14.7	3.319	MEAS		N	F003	S02
TU027	7.25	8.78	203	17.4	7.56	3.305	MEAS		N	F003	S02
TU027	8.78	10.3	238.6	20.8	16.3	3.575	MEAS		N	F003	S02
TU027	10.3	11.22	206.7	17	13.4	3.87	MEAS		N	F003	S02
TU027	11.22	12.74	560.9	41.8	7.56	4.916	MEAS		N	F003	S02
TU027	12.74	14.26	670.6	52.4	10.6	4.695	REG		N	F003	S02
TU027	14.26	14.94	517	39.2	13.2	4.533	MEAS		N	F003	S02
TU027	14.94	16.12	216.7	17.2	8.04	4.601	MEAS		N	F003	S02
TU027	16.12	17.4	692.6	64.6	2.58	5.057	MEAS		N	F003	S02
TU027	17.4	18.41	559.5	49	4.56	5.098	MEAS		N	F003	S02
TU027	18.41	19.93	185.1	14.2	8.64	3.735	MEAS		N	F003	S02
TU027	19.93	21.46	319.5	26	12.8	4.091	MEAS		N	F003	S02
TU027	21.46	21.7	222.2	17.4	7.56	4.069	MEAS		N	F003	S02
TU027	21.7	22.46	469	42.6	3.3	4.11	REG		N	F003	S02
TU027	22.46	23.07	175.5	16.2	2.22	4.813	MEAS		N	F003	S02
TU028	0	1.52	40.5	2.88	6.6	2.864	REG		N	F003	S01
TU028	1.52	1.92	40.5	2.35	8.04	2.895	REG		N	F003	S01
TU028	3.05	3.47	30.2	0.68	7.08	3.15	MEAS		N	F003	S01
TU028	3.47	5	226.3	14.63	19.2	4.073	MEAS		N	F003	S01
TU028	5	6.37	309.9	22	17.9	4.012	MEAS		N	F003	S01

Providing a stable foundation for analytics

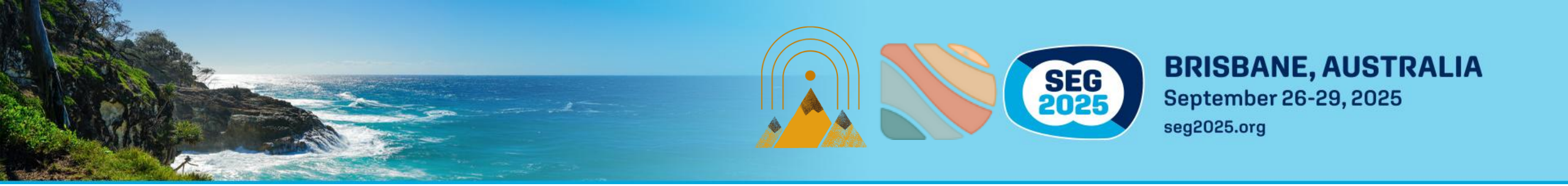
- This code unlocks legacy datasets, expanding usable datasets rather than discarding them.
- Families and sub-families provide clarity so that interpreters can trust what is comparable, ensuring that everything from simple interpretations to advanced analytics have more stable foundations.
- We have made the code is open-source in the spirit of transparent and accessible science.

Says the mathematical geologist to the geochemist…

- I have rewritten this eight times and it is a good **first** step.
- I want to try the Apple Embedding Atlas.
- I want to add a large language model.



Github: Embedding Atlas


SEG
2025
BRISBANE, AUSTRALIA
September 26-29, 2025
seg2025.org

Thank You!