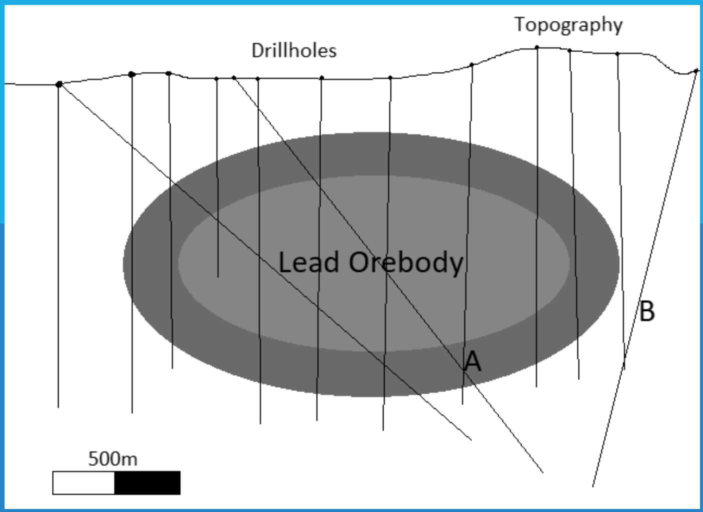
Geochem Dataset : QAQC for Machine Learning

NIXIS CARRERO| 30TH SEPT 2025

Table of Content

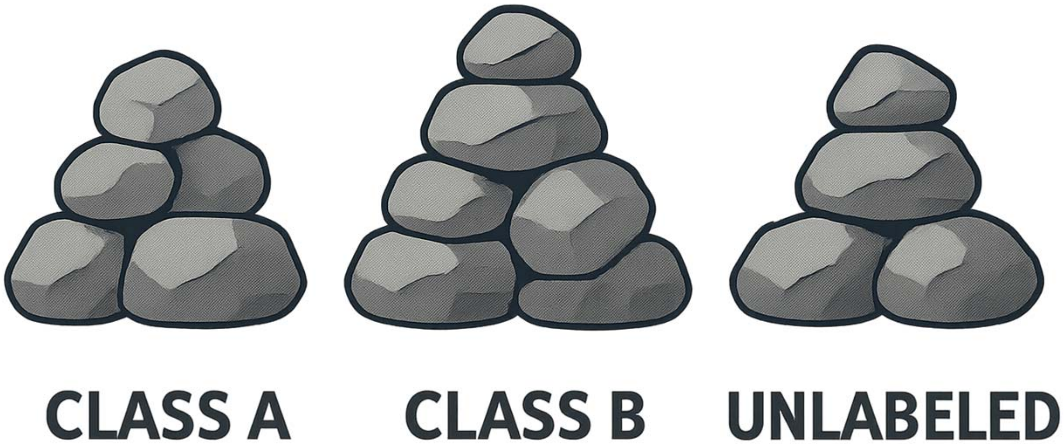
What are we talking about?



Objective

– **Can we use** the same geochemical data and **labels** to generate a predictive model for future drill holes which can label samples on whether they are in **class A** or **class B?**

– More data has been acquired since the geochemist completed her work - can we **predict labels** onto these data points (**labelled "?"**).



Data | Methodology

The Data

- **Data Summary:**
  - **Samples**: 4,771
  - **Assays (8)**: As, Au, Pb, Fe, Mo, Cu, S, Zn
  - **Labels**: A, B, ?
  - **Metadata**: Unique_ID, holeid, from, to
- **Issues Detected**
  - Wrong datatype
  - Missing values (notably As ~31%)
  - Truncated values at detection limits (e.g., “<0.005”)
  - Invalid placeholders (e.g., -999)

Geochem assays

(4771, 13)

	Unique_ID	holeid	from	to	As	Au	Pb	Fe	Mo	Cu	S	Zn	Class
0	A04812	SOLVE003	561	571.0	NaN	0.066	1031.00	61380.0	138.2000	3.600	3586.0000	43.6000	A
1	A03356	SOLVE003	571	581.0	NaN	0.152	1982.00	50860.0	75.4000	4.800	1822.0000	36.4000	A
2	A04764	SOLVE003	581	591.0	NaN	0.068	1064.80	57940.0	29.2000	3.000	740.4000	36.6000	A
3	A04626	SOLVE003	591	601.0	NaN	0.074	891.60	48620.0	63.0000	4.200	820.8000	39.6000	A
4	A05579	SOLVE003	601	611.0	NaN	0.043125	801.25	51025.0	56.0625	4.875	745.6875	32.3125	A

```
--- Info Summary ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4771 entries, 0 to 4770
Data columns (total 13 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Unique_ID  4771 non-null   object
 1   holeid     4771 non-null   object
 2   from       4771 non-null   int64
 3   to         4771 non-null   float64
 4   As         3268 non-null   float64
 5   Au         4765 non-null   object
 6   Pb         4756 non-null   float64
 7   Fe         4709 non-null   float64
 8   Mo         4741 non-null   float64
 9   Cu         4746 non-null   float64
 10  S          4761 non-null   float64
 11  Zn         4762 non-null   float64
 12  Class      4771 non-null   object
dtypes: float64(8), int64(1), object(4)
```

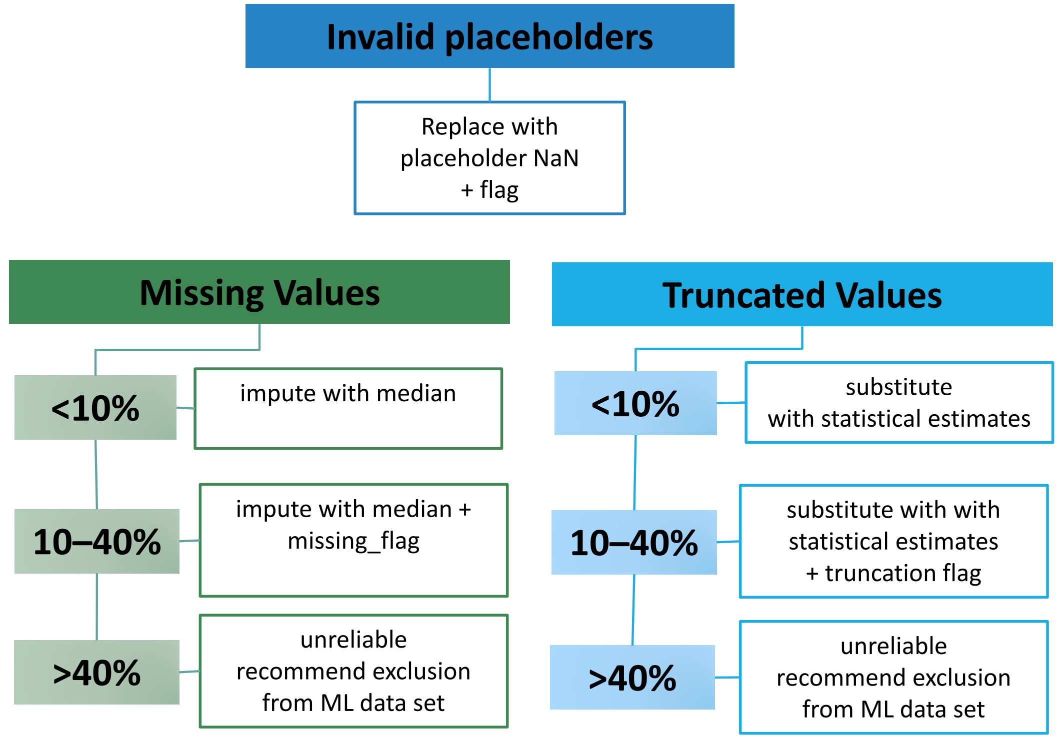
```
Missing values per assay:
Unique_ID      0
holeid         0
from           0
to             0
As          1503
Au             6
Pb            15
Fe            62
Mo            30
Cu            25
S             10
Zn             9
Class          0
dtype: int64
```

```
Missing percentage per assay:
Unique_ID     0.00
holeid        0.00
from          0.00
to            0.00
As           31.50
Au            0.13
Pb            0.31
Fe            1.30
Mo            0.63
Cu            0.52
S             0.21
Zn            0.19
Class         0.00
dtype: float64
```

		3.6	Low ppm values
		3586	Often thousands ppm; sometimes reported in %
Zn	Zinc assay (ppm)	43.6	ppm level, consistent ranges
Class	Rock classification label (target variable)	A	Imbalanced distribution (60% A, 24% B, 15% unknown)

Data Cleaning

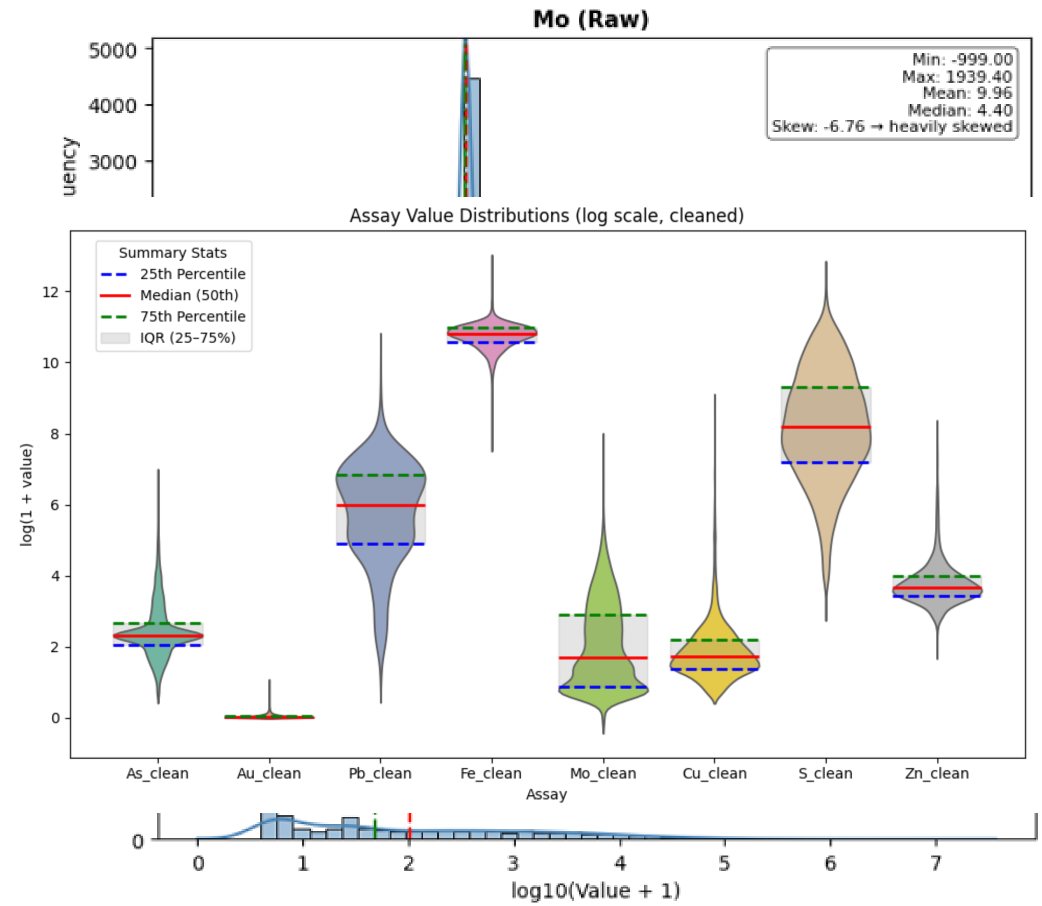
- True **missing values** may arise from **unsaved intervals or lab reporting gaps** (e.g., <-999).
- Some assay results are reported as **truncated values** (e.g., <0.005), meaning the true concentration is below the **detection limit** (DL).
- **Use imputation methods** — fill values in a statistically consistent way.
- **Use flags** — **keep as much information as possible** without losing valuable samples




```
QAQC Summary:
As: Truncated 0.0%, Missing 31.5%, Negatives 0 → Retained
Au: Truncated 9.7%, Missing 0.1%, Negatives 0 → Retained
Pb: Truncated 0.0%, Missing 0.3%, Negatives 0 → Retained
Fe: Truncated 0.0%, Missing 1.3%, Negatives 0 → Retained
Mo: Truncated 0.0%, Missing 0.6%, Negatives 28 → Retained
Cu: Truncated 0.0%, Missing 0.5%, Negatives 0 → Retained
S: Truncated 0.0%, Missing 0.2%, Negatives 0 → Retained
Zn: Truncated 0.0%, Missing 0.2%, Negatives 0 → Retained
```

cle
cleaned_data_fo _flag)

EDA (Exploratory Data Analysis)



–**Histograms** reveal the shape of the data, highlight skewness and outliers.

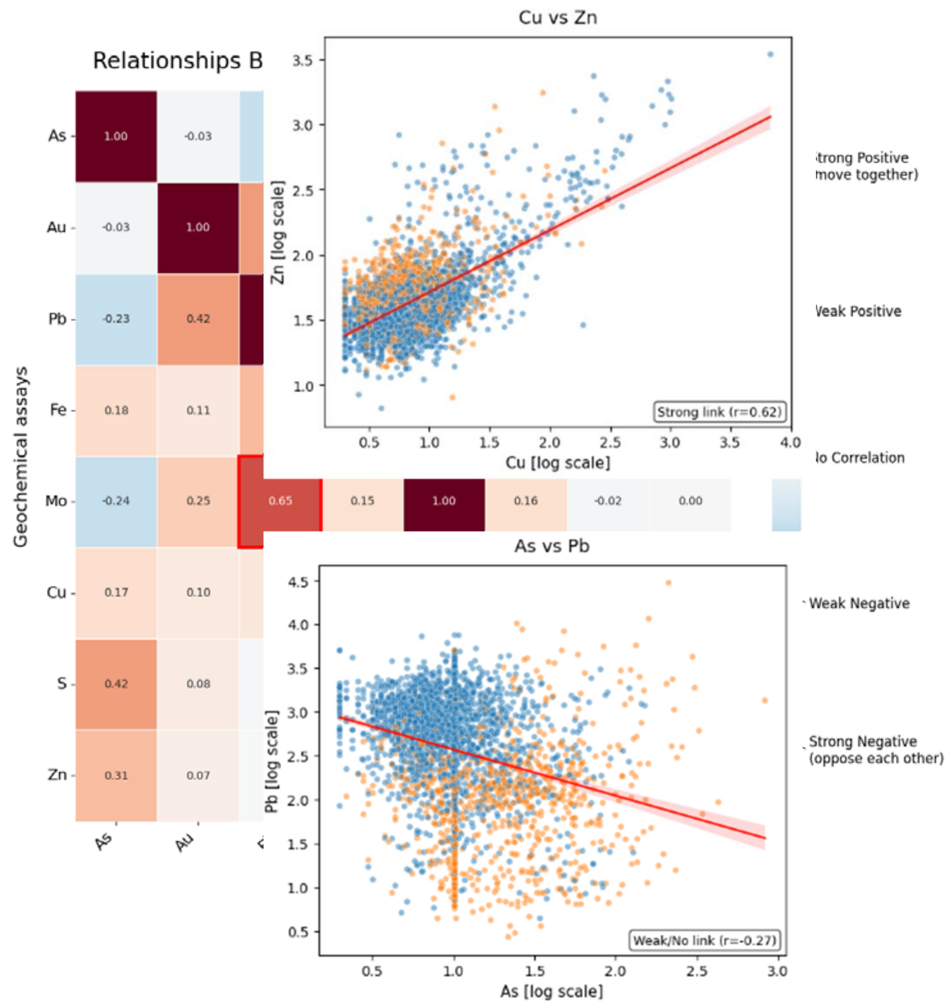
–**Transforming to log** space reduced skew and clarified patterns in the geochemical assays, but not all elements behave the same — some are stable, while others need extra care.

–**Violin plots** reveal the spread and distribution of each element, making patterns and outliers easy to compare across **multiple elements in a single view**.

EDA (Exploratory Data Analysis)

–**Heatmap:** shows correlations between elements (positive and negative), highlighting pairs like Pb–Mo and Cu–Zn that move together.

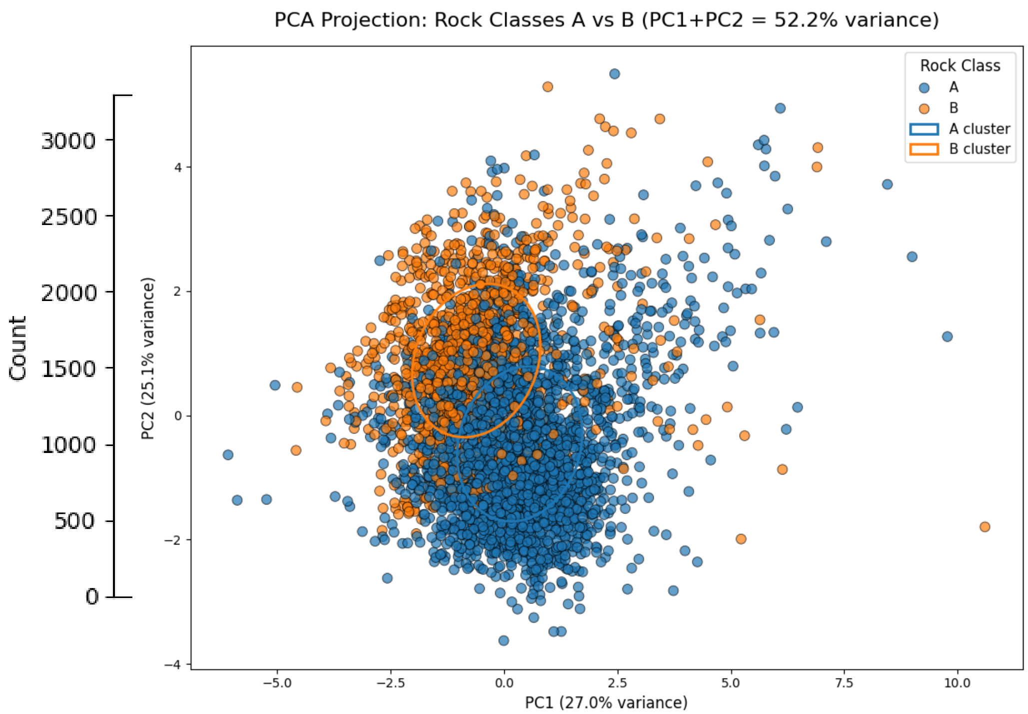
–**Scatter plots:** show relationships by sample and class, useful for confirming patterns and detecting artifacts (e.g., As below detection limit).



EDA (Exploratory Data Analysis)

- **Class distribution:** Dataset is imbalanced (~60% Class A), which could bias models → balancing strategies needed.

- **PCA analysis:** Partial A vs B separation (PC1+PC2 ≈ 52% variance); overlap suggests enrichment with geological/spatial features for stronger predictive models.



Conclusions

Conclusions

1

QAQC is not optional. Clean data is the foundation of reliable machine learning.

2

Smart preprocessing, **like log transforms and structured imputation rules**, makes complex g data interpretable and usable without losing traceability.

3

Predictive labeling is promising, but the dataset by itself is not enough.

To scale, **we need richer data and more context.**

"Clean data enables insight — but robust predictive models require richer datasets"

Thank you

Appendix

https://github.com/Nixis/geochem-assay-qaqc-ml