PFAS FORENSIC STATISTICAL METHODS

EXTERNAL ADVISORY GROUP MEETING

Skyler Sorsby [skyler.sorsby@wsp.com]
December 19, 2025

Agenda

- The FALCON method
- Bar charts – and other graphics
- Data quality – and the detection limit
- Principal Components Analysis – a ***canvas*** for the data (**c. 1901**)
- Factor analysis – finding ***exemplar*** signatures / trends (**c. 1976**)
- Clustering – finding ***average*** signatures / groups (**c. 1963**)
- Correlation – ***relating*** signatures to sources or tracers (**c. 1896**)
  - Birds-eye view of machine learning as “advanced mode” correlation (**c. 1896 & 1951**)

What is the FALCON method?

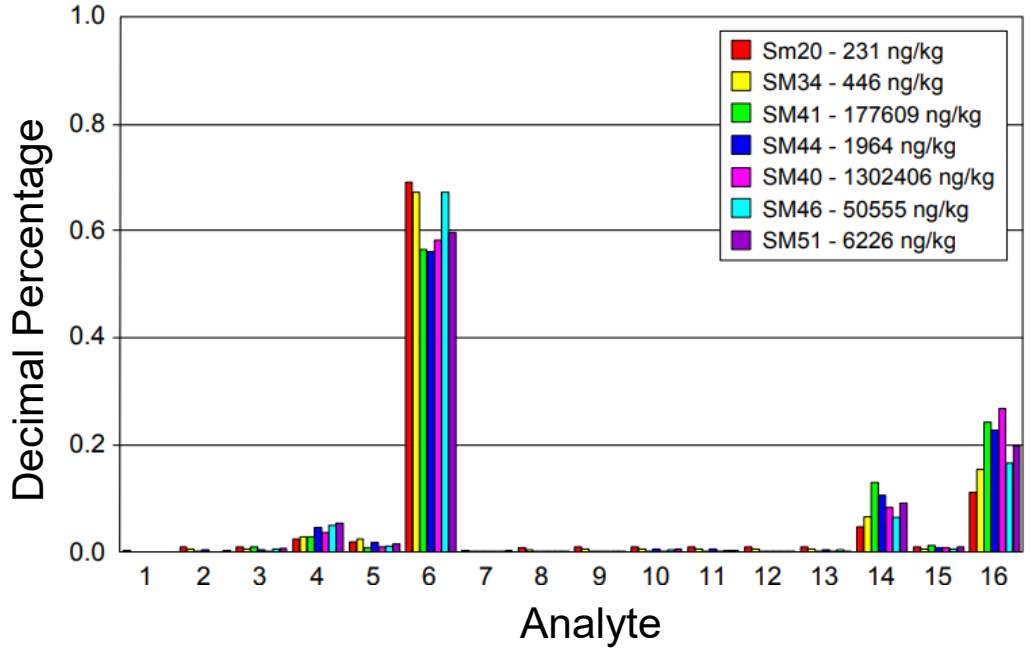
- Systematic comparison of chemical fingerprints to each other, and to sources (Plumb, 2004).
- Steps:
  - **Tabulate** (pick analytes, samples, order).
  - **Normalize** (divide rows by sum to get decimal percentages).
  - **Visualize** (FALCON traditionally uses bar charts).
  - **Statistics** (discussed later).
- Demo with two GeoTracker landfill sites, 11 total data points.

United States Environmental Protection Agency | Office of Research and Development | Office of Solid Waste and Emergency Response | EPA/600/5-04/054 May 2004

EPA **Technical Support Center Issue**

Fingerprint Analysis of Contaminant Data:
A Forensic Tool for Evaluating Environmental Contamination

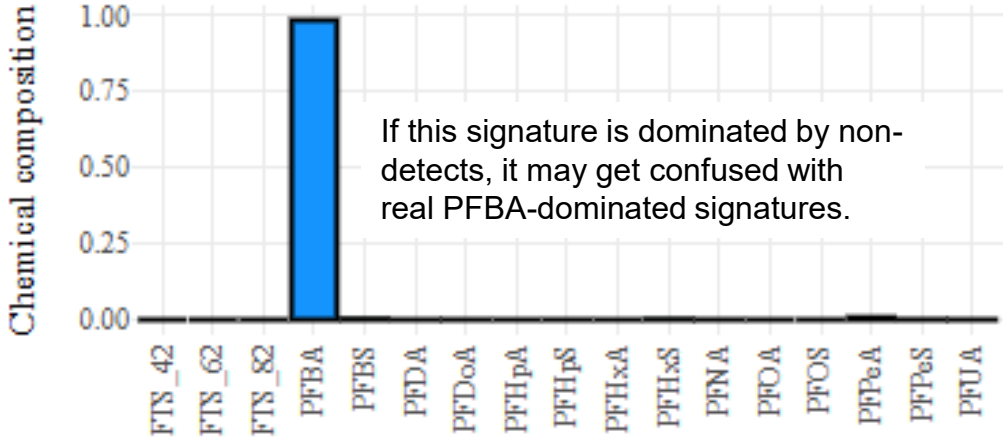
— Russell H. Plumb, Jr.*



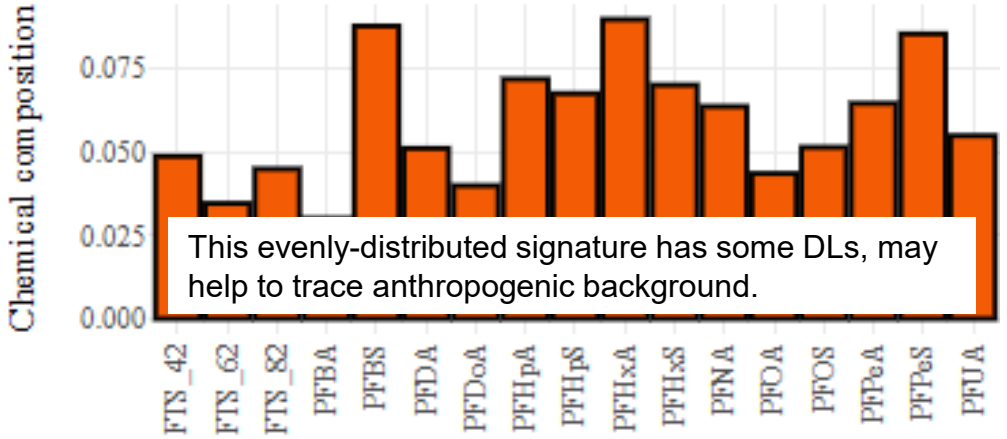
https://www.epa.gov/remedytech/fingerprint-analysis-contaminant-data-forensic-tool-evaluating-environmental

Dealing with detection limits

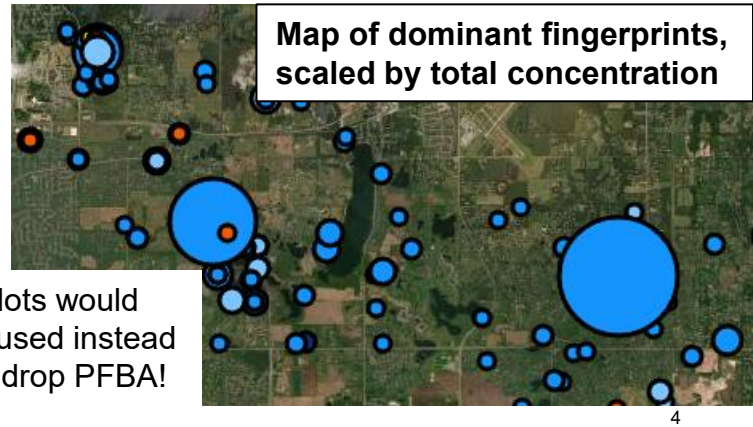
- **Elevated detection limits** may occur if the sample is heavily impacted, or the lab equipment is compromised. Problem: extra noise in the dataset.
- **Fingerprints dominated by non-detects** may be misleading if they resemble distinct patterns
  - Substitution: represent non-detects (NDs) with the method detection limit (MDL), reporting limit (RL), ½ of either, zero, or (advanced) impute.
  - Uniform / low DLs: may simply track as another fingerprint.
  - Analytes with known elevated DLs may be worth excluding – if not a tracer.
- **Regardless,** it is important to vet sensitivity of results (i.e., replace with zero vs. the MDL vs. remove analyte).



If this signature is dominated by non-detects, it may get confused with real PFBA-dominated signatures.



This evenly-distributed signature has some DLs, may help to trace anthropogenic background.

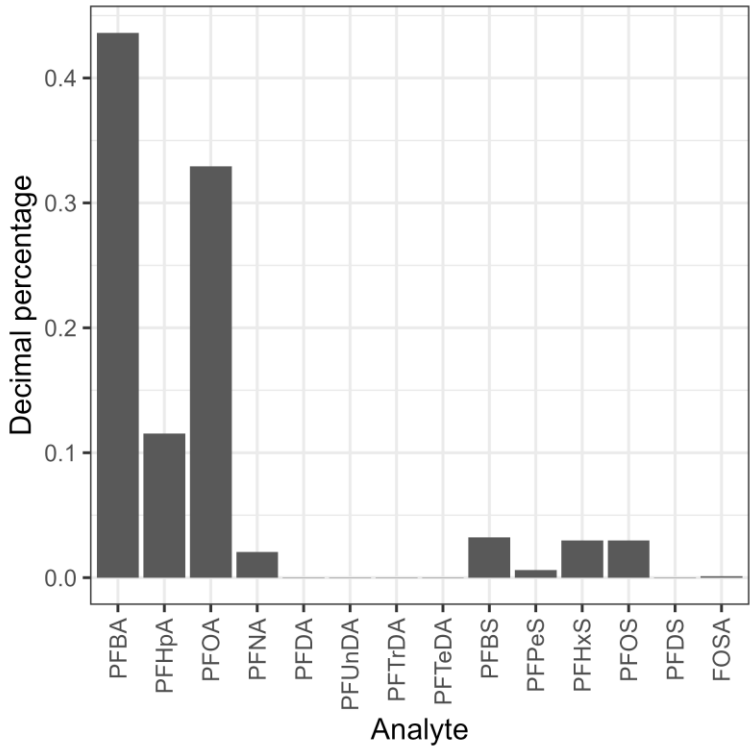


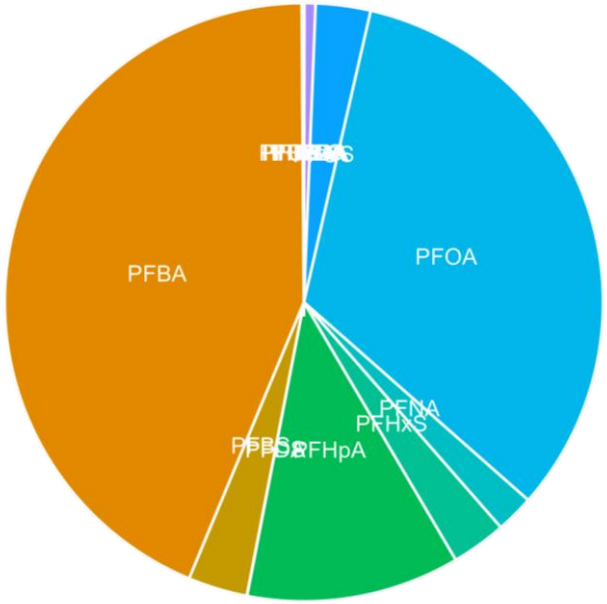
Map of dominant fingerprints, scaled by total concentration

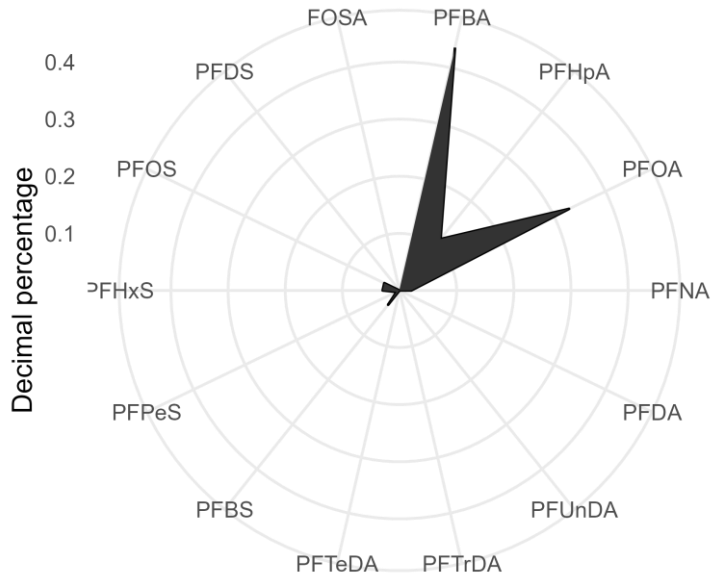
Perhaps some of these dots would turn orange if zero were used instead of the MDL – can't really drop PFBA!

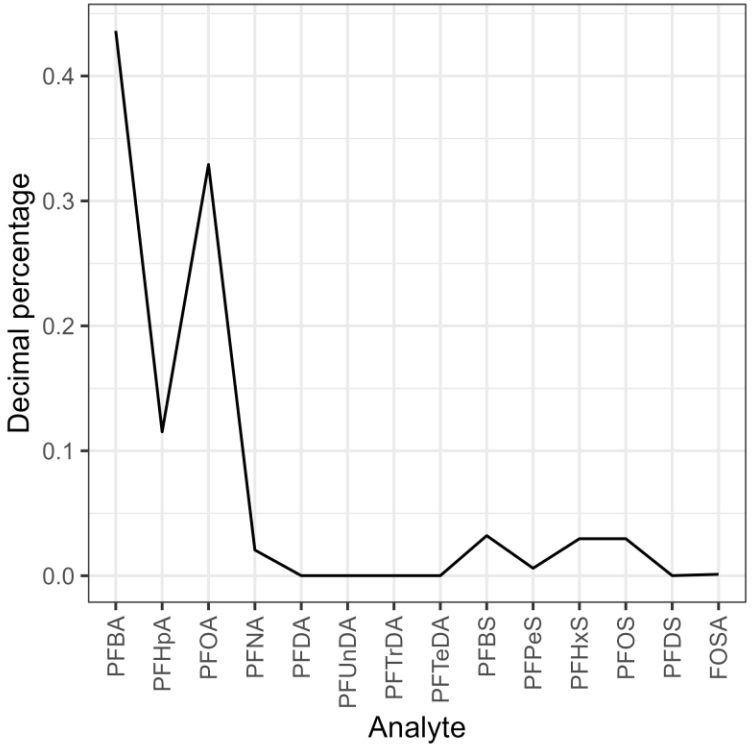
Visualization – sample E-05

- **Bar charts** are traditional for FALCON.
- **Radial diagrams** create potentially distinct shapes
- **Pie charts** are visually appealing – areas may inflate perception of proportions.
- **Parallel Coordinate charts** are essentially unrolled radial diagrams.
- A meaningful order (i.e., by type and chain length) can simplify patterns.



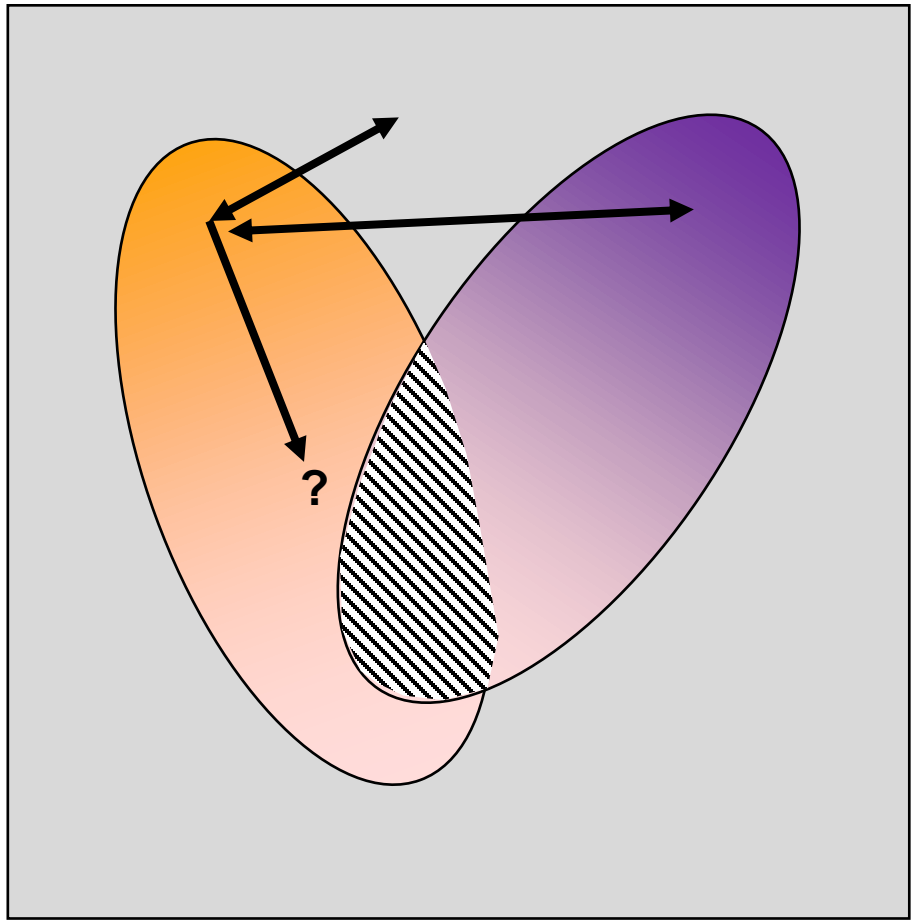






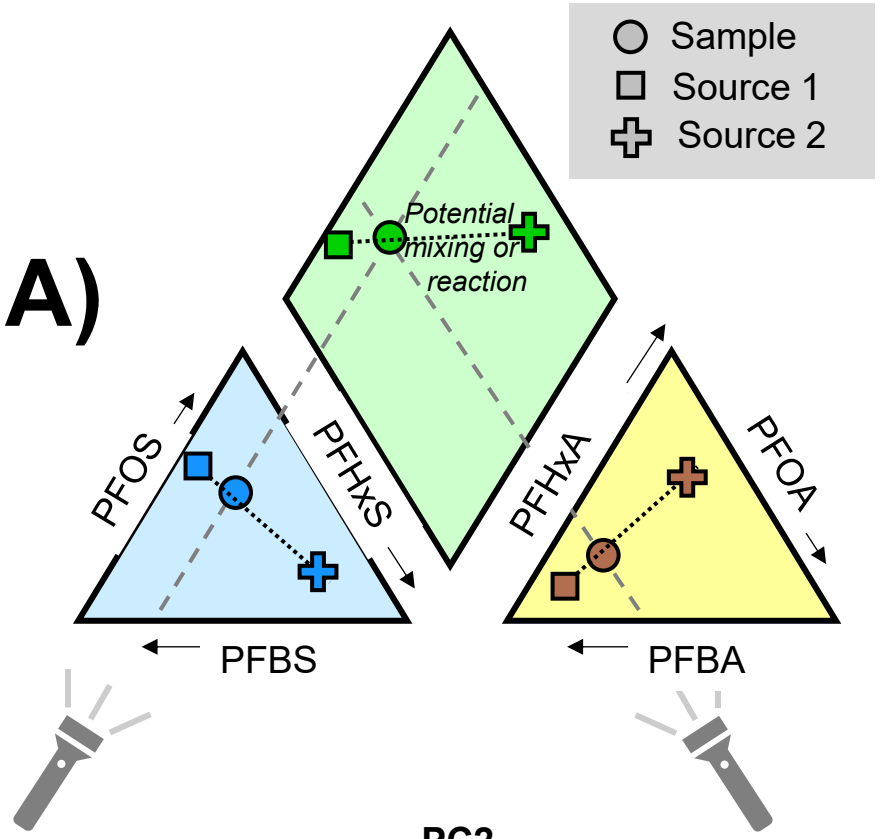
Potential goals of statistical analysis

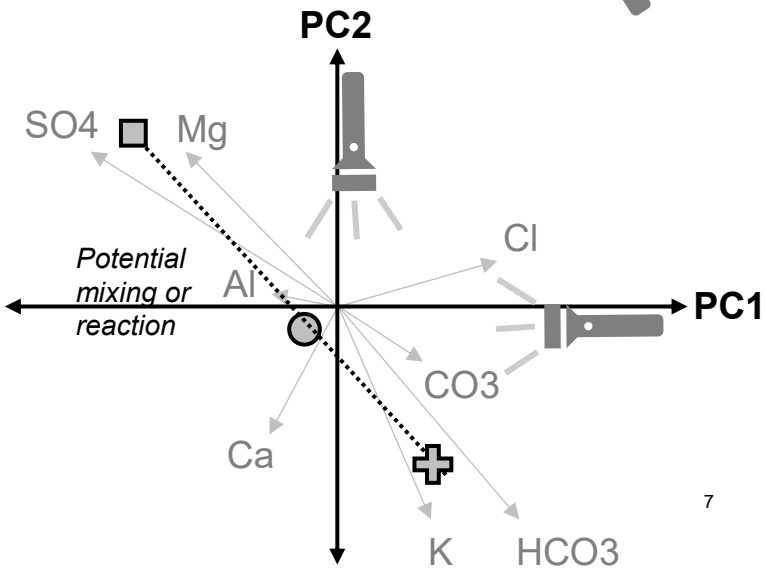
- Plumb (2004) indicates statistical analysis can help to:
  - differentiate the sources from [*anthropogenic*] **background**
  - demonstrate whether contamination detected at some distance from a site is **related to the source**
  - map contaminant **migration** away from a source
  - **differentiate** multiple sources of the same contaminant
  - estimate the **mixing ratio** of two plumes
- Two popular statistical tools are commonly used in environmental forensics to help with these: **matrix factorization** and **cluster analysis**.

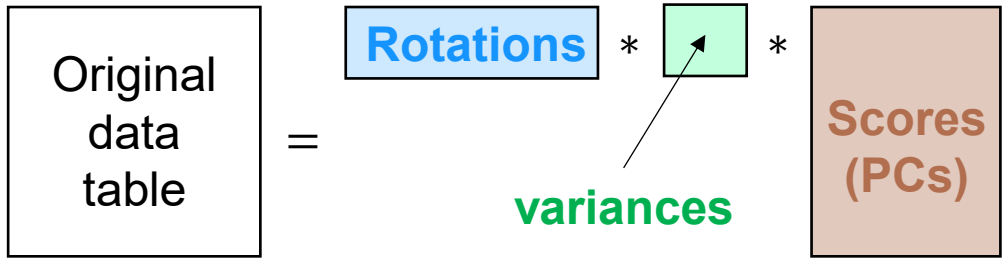


Principal Components Analysis (PCA) as an advanced Piper Diagram

- Geochemists use **Piper Diagrams** to inspect major-ion data & ID mixing, reactions, etc.
- Three parts: two triangles (cations, anions), and a diamond that "projects" through the triangles
  - "**Projection**" = linear combination of eight variables (major ions). Think: multidimensional flashlight.
  - Data are always normalized
- PCA: a *matrix-factorization* technique developed in **1901** that uses **linear algebra identifies the combinations** & collapses to rectangular coordinates (vs. a diamond).

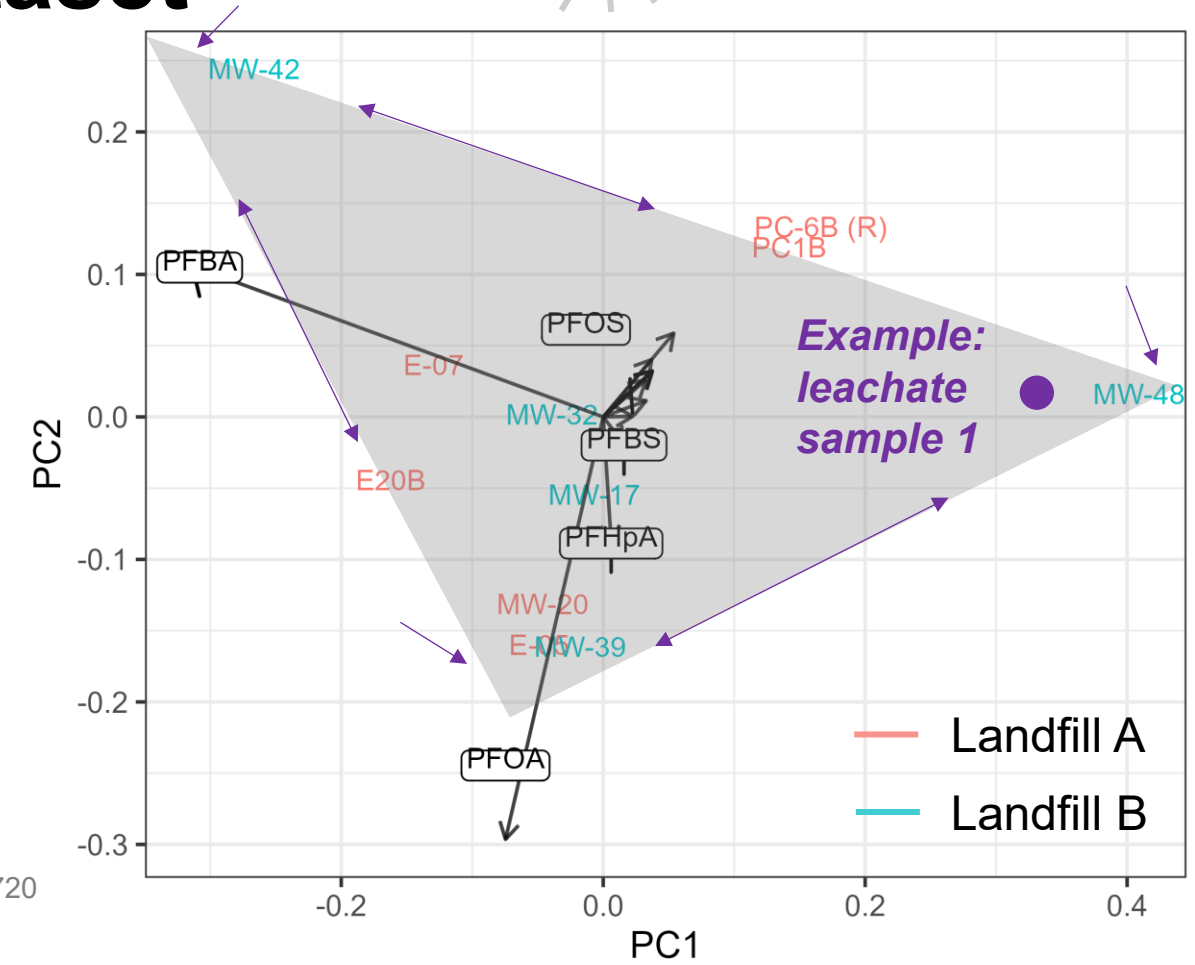






Principal Components Analysis (PCA) as a canvas for the dataset

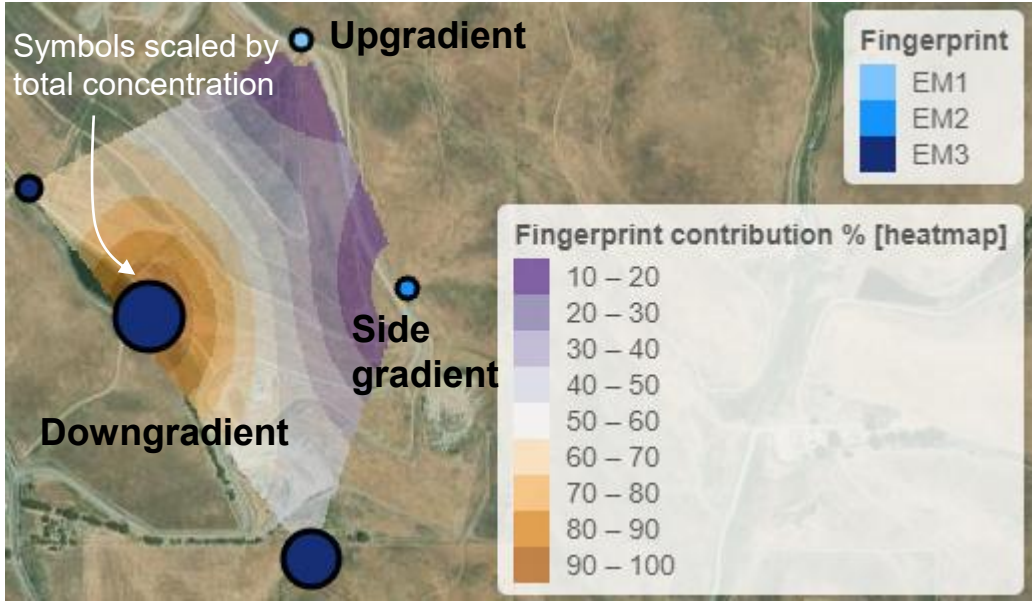
- PCA creates a hybrid line / scatter plot
  - **Sample position** relates to composition (scores/PCs).
  - **Analytes point** in the direction of higher dominance (rotations)
  - **Variances** indicate amount of information captured per PC.
- Percent-normalized data tend to produce:
  - corners (unique signatures)
  - edges (gradation between two signatures).
- Additional data can be “painted” onto the canvas.
- PCA can be over-interpreted.
- Here, potentially three signatures, two from Landfill B might be distinct. General overlap of sample fingerprints.



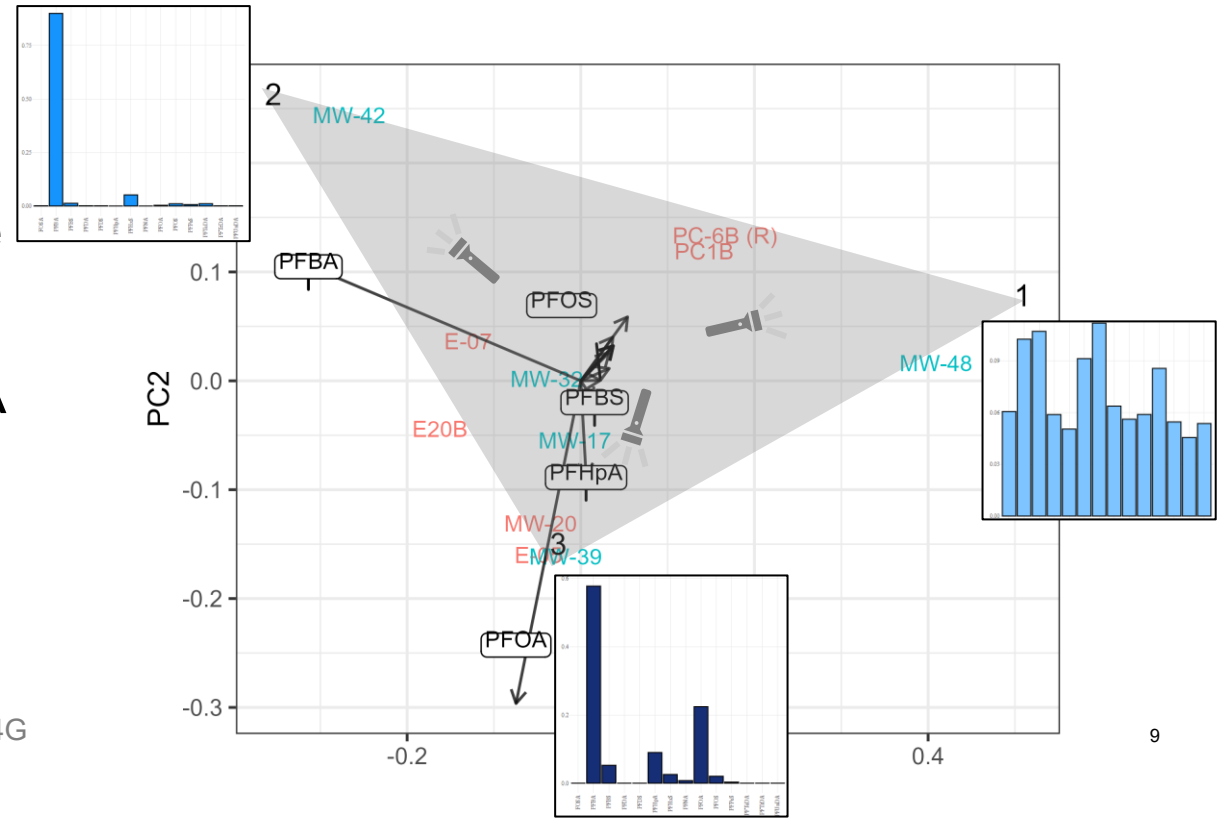
Pearson (1901) https://www.tandfonline.com/doi/abs/10.1080/14786440109462720

Original data table = **Loadings** * **Scores (EM)**

Factor Analysis finds *archetypal* signatures & *gradients*



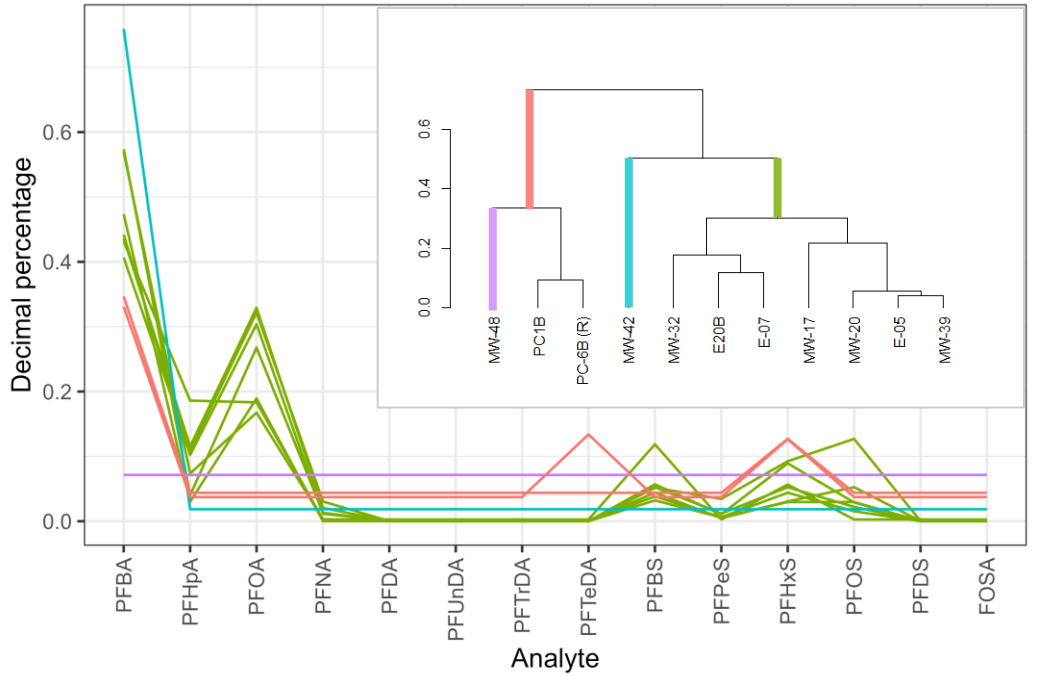
- Factor analysis (**1976**; i.e., NMF, PVA, ALS) is also matrix factorization.
  - finds “pure” **corners** (fingerprints)
  - estimates “**signal strength**” for each sample.
  - Roots in geochemical fingerprinting, longstanding use in forensics. https://pubs.usgs.gov/publication/pp574G
- Can plot as **compositional bar charts** / on a PCA plot, and **heatmaps of signature strength**.
- Aside from sources, signatures may also relate to precursor transformation, F&T, or anthropogenic background.

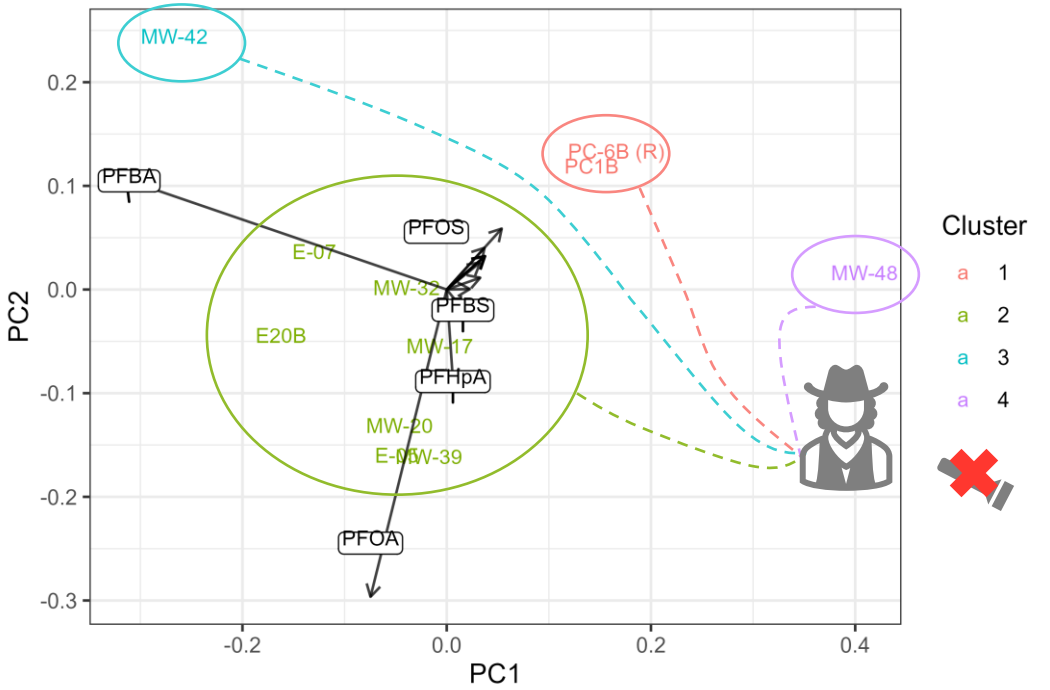


Miesch (1976) https://pubs.usgs.gov/publication/pp574G

Clustering finds *average* signatures and *groups*

- Clustering (K-means, hierarchical clustering [1963], etc.) conceptually **best suited for distinct signatures** with some variability, where the purest signals are at the center of each group.
- Dataset may or may not have clusters. **Dendrogram** tree height is a good way to check.
- Can also **visualize on a PCA plot** (but it is *only* an overlay).
- Not a matrix factorization method. Think: multidimensional cowboy.
- Related: clustering (particularly GMM) has received attention to parse out **anthropogenic background levels**.





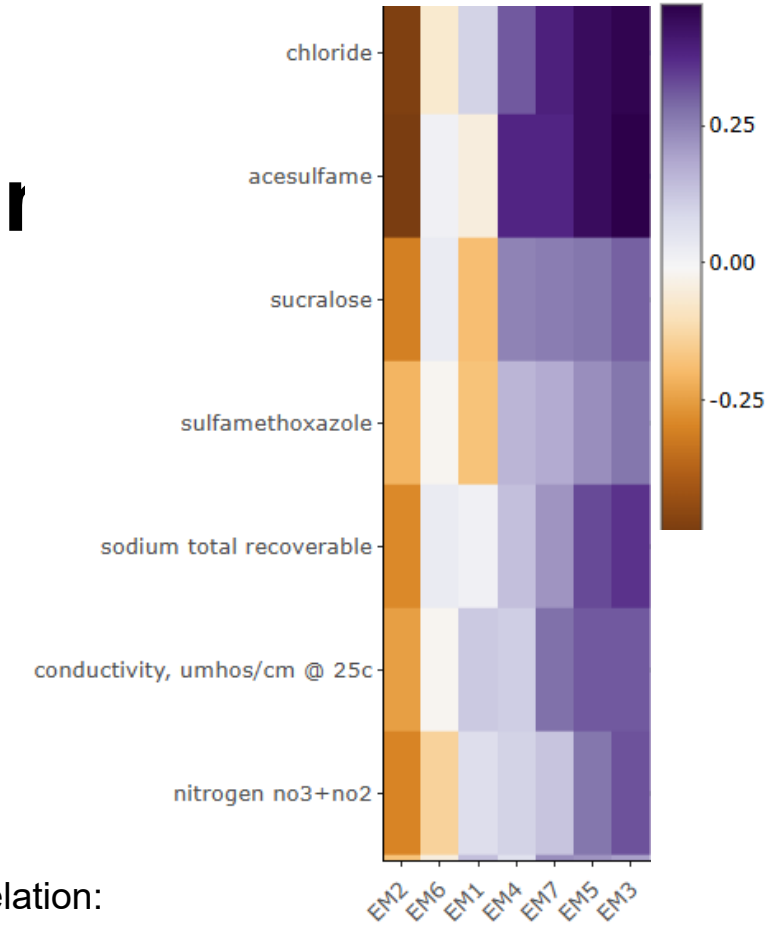
- Anderson and Modiri (2024) https://link.springer.com/article/10.1007/s10661-024-12400-z

Ward (1963) https://www.jstor.org/stable/2282967

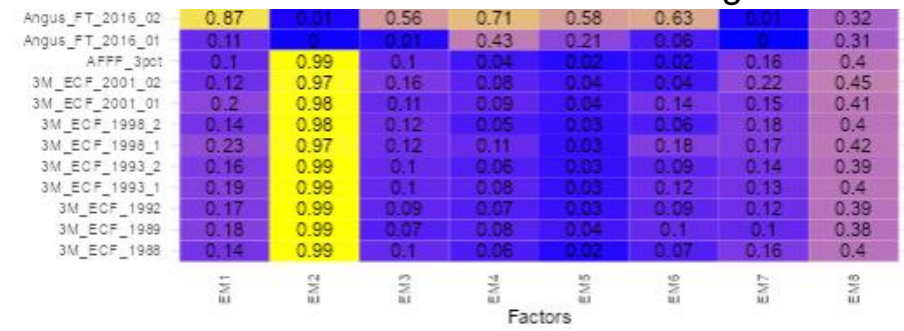
Correlation identifies similarity between samples or data columns

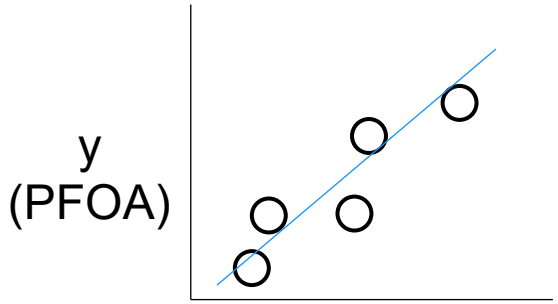
- Two primary correlation metrics:
  - I.e., Pearson (1896; compare columns / analytes). Use to compare signatures vs. tracers, co-contaminants.
  - Cosine-theta (compare rows / samples). Use to compare to reference data (i.e., leachate samples).
- Value ~ 0: little/no correlation
- Value ~ 1: Strong correlation
- Value ~ -1: Inverse correlation

Example: spearman correlation:
Tracers vs. PFAS signatures



Example cosine-theta correlation:
AFFF formulations vs. PFAS signatures

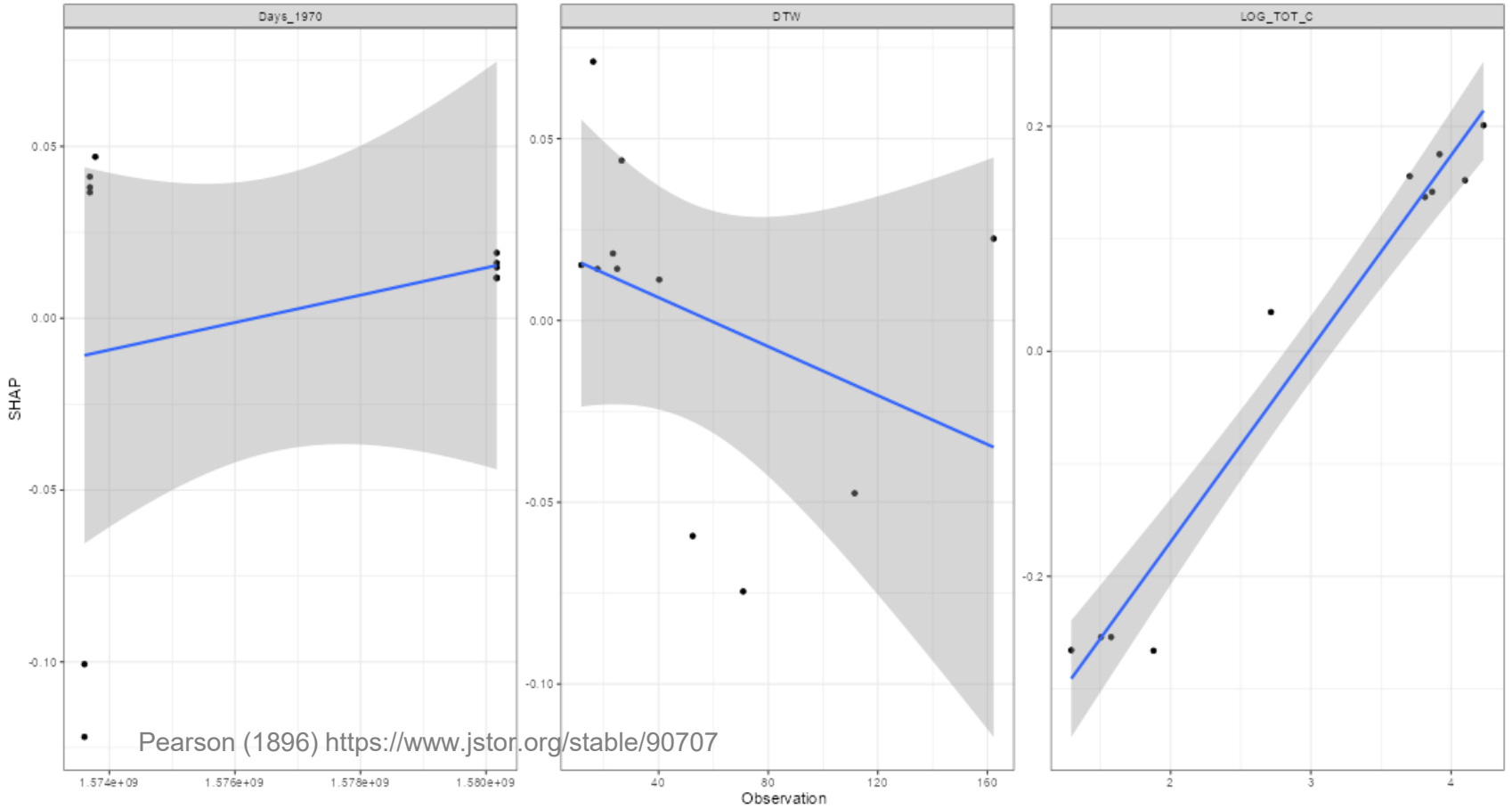




Machine-learning regression is “advanced mode” tracer correlation

- Regression [1896] = advanced correlation.
- Similar to what we learned in basic algebra (y = m*x + b).
- Extended to include lots of inputs! (y = m1*x1 + m2*x2 +…)
- Relates fingerprints to anything that can be measured, i.e.:
  - Soil types
  - Geochemistry
  - Distance from point of known release
  - All considered simultaneously.

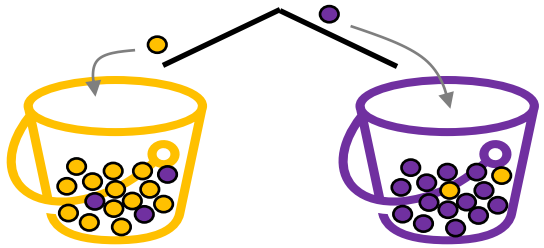
Signature 2 relatively stable over time, decreasing with depth, and strongly linked to elevated PFAS.

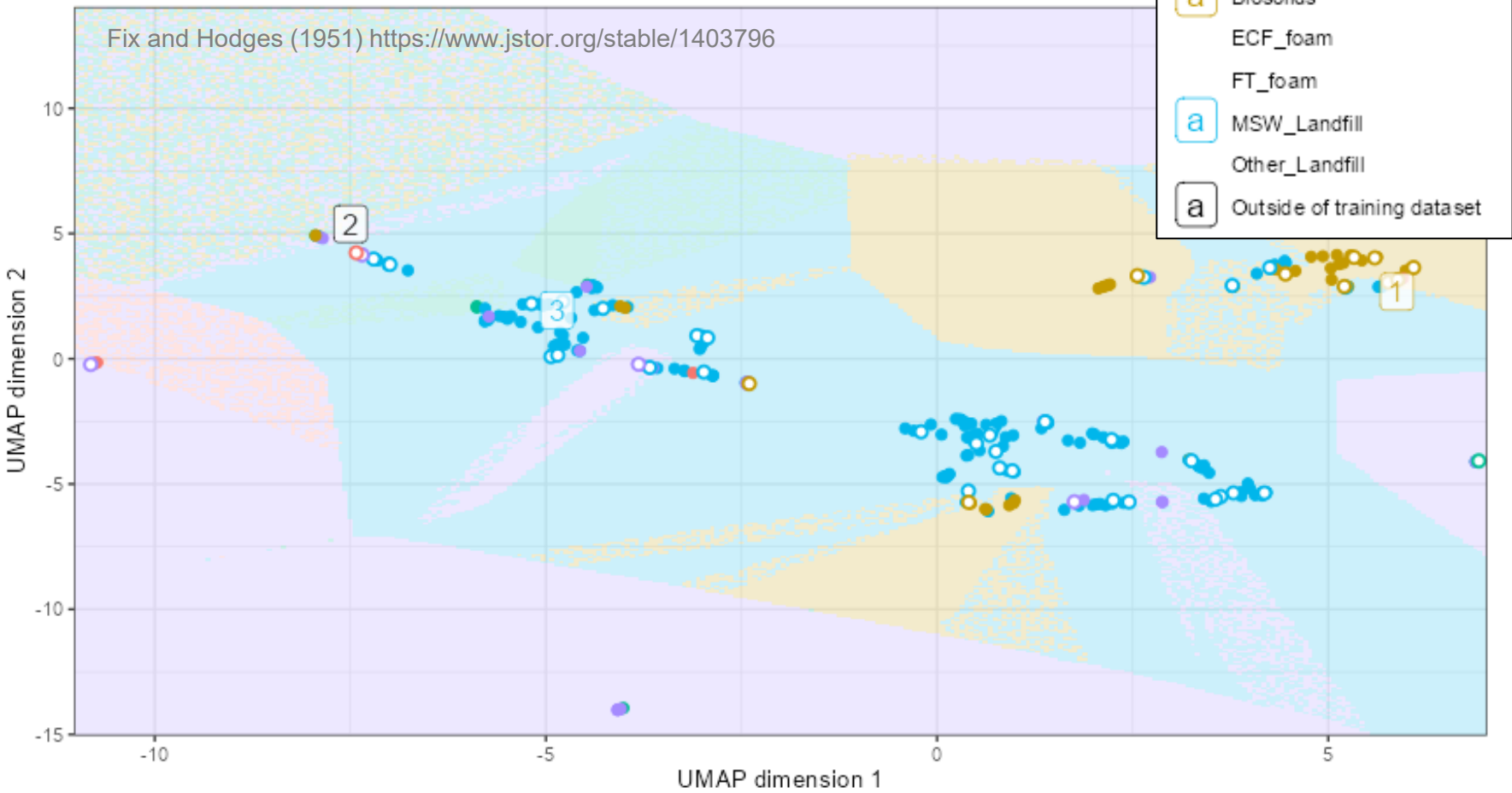


Pearson (1896) https://www.jstor.org/stable/90707

Machine-learning classification is “advanced mode” sample correlation

- Predicting group membership. Think: a sorting machine.
- Most classifiers (like KNN; 1951) assume the data they were trained on are **comprehensive**. Ambient signature mis-ID’d as biosolids? Closest match.
- Signature 2 out-of-sample.
- Signature 3 correctly ID’d as a landfill signature.





Summary

For more info, scan the QR code or email skyler.sorsby@wsp.com

- FALCON method **normalizes bar charts** to total PFAS concentration and focuses on bar chart inspection + additional statistics.
- PCA is essentially a **roadmap of bar-charts**, can compare site vs. reference vs. calculated signatures.
- Factor analysis **derives 'unmixed' bar charts** from site data.
- Clustering identifies **average bar charts and their groups**.
- Correlation is a simple and quick way to **compare site bar charts** vs. tracers or a reference.
- Machine learning methods **may help interpret bar charts but must be handled with care**. Not a silver bullet.