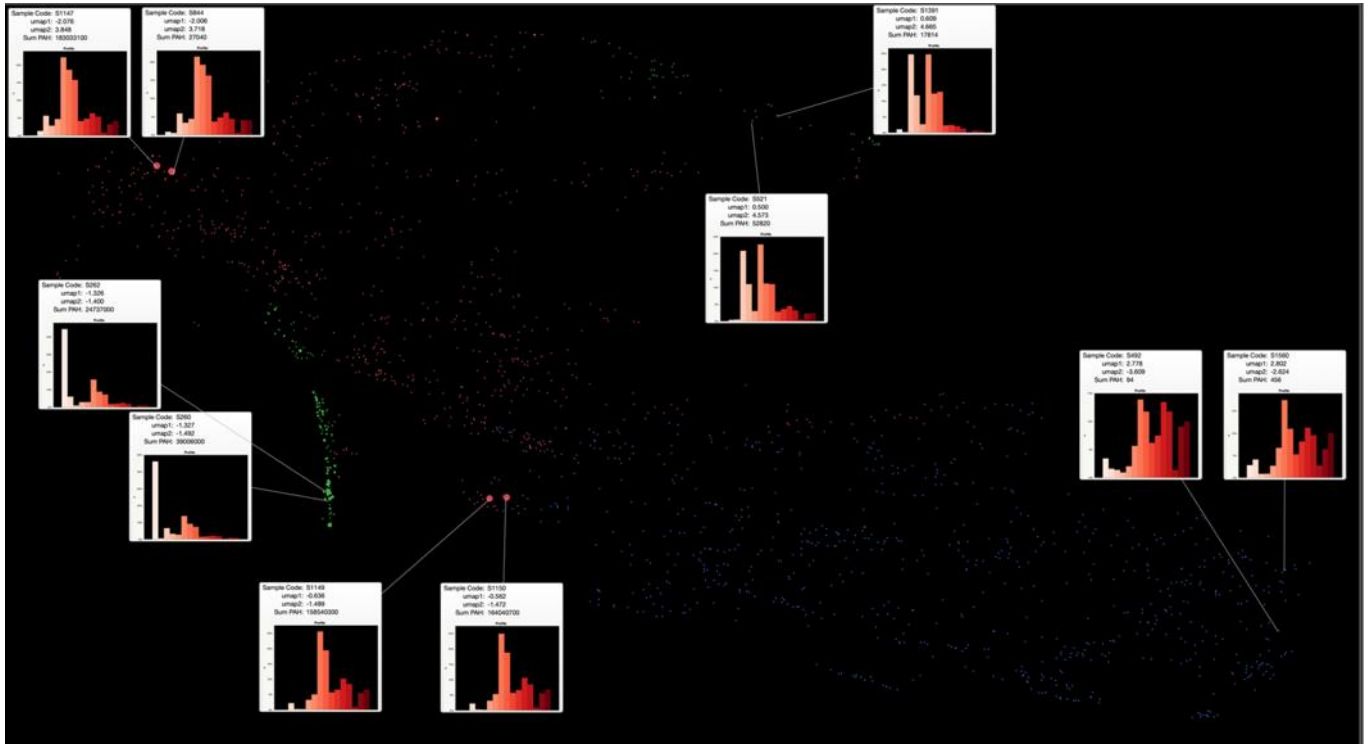
chemistry matters®

Making chemistry data meaningful



Managing data and data discovery for large scale Superfund sites and releases

Court Sandau, PhD, PChem, FRSC and Paul Fuellbrandt, BSc, PAg, PMP

October 12, 2023

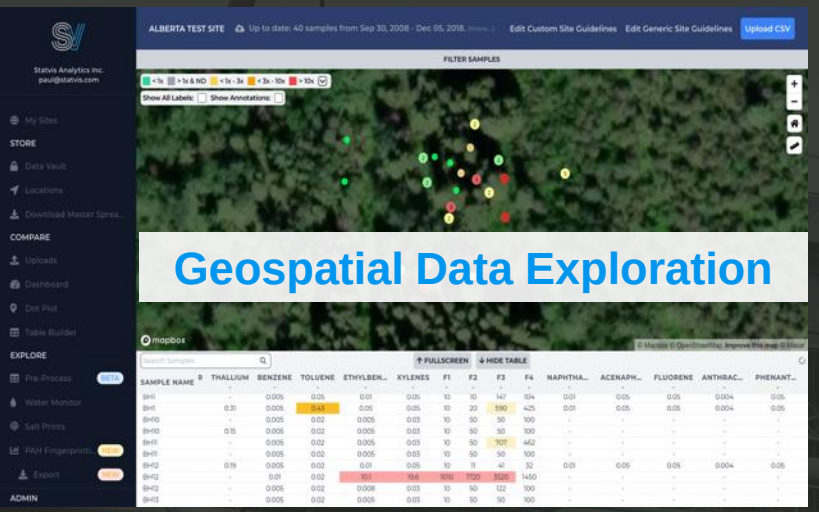
Companies

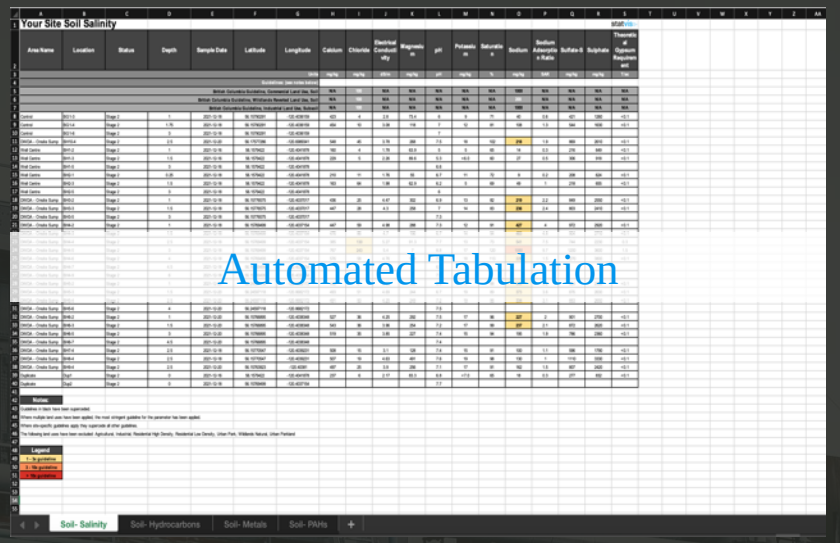
Chemistry Matters Inc.

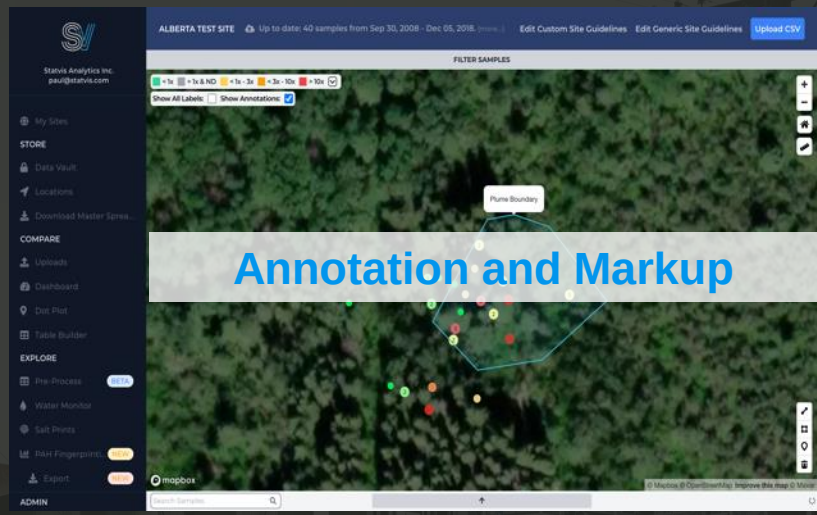
- **Niche chemistry consulting company specializing in environmental forensics, geoforensics, biomonitoring and arson investigations**
- Big environmental datasets have become our specialty to merge chemistry understanding with statistical interpretation and science communication
- Virtual office-based business

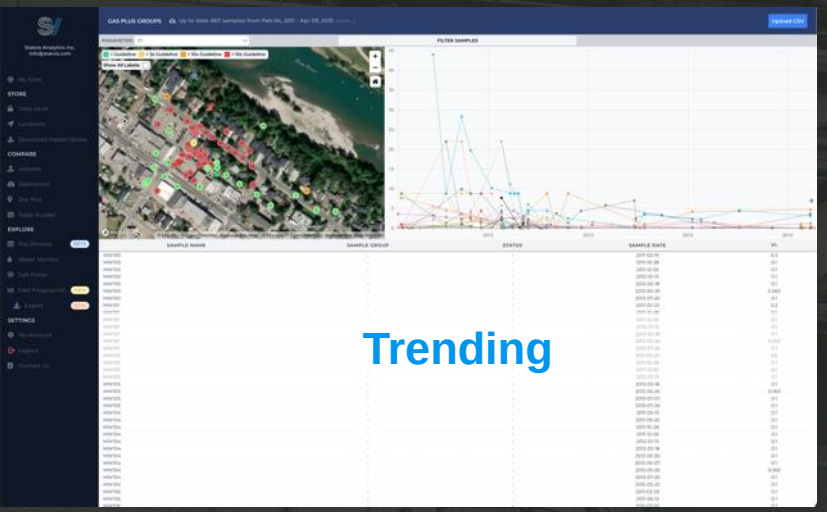
Statvis Analytics Inc.

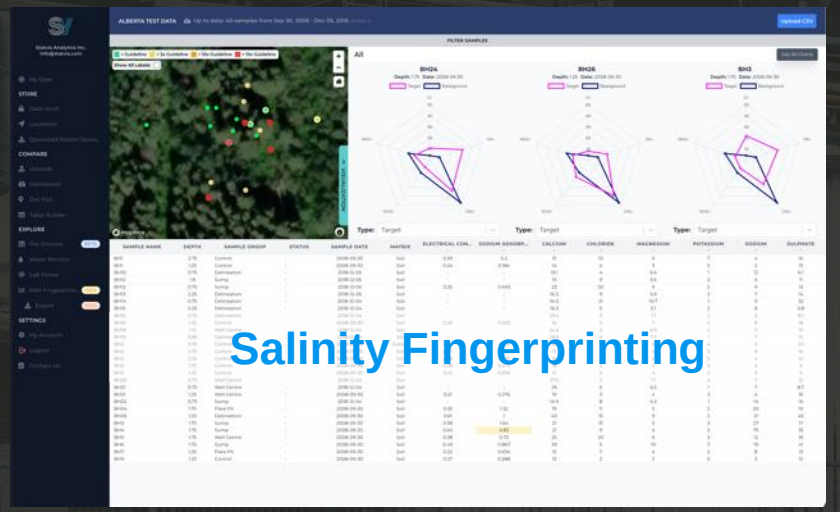
- Software development company specializing in data and statistical visualization software customized for environmental industry
  - **Environmental Data Intelligence System (EDIS)™**
- Turning complex data and statistics into easy-to-understand graphics and pictures
- Helping people communicate science









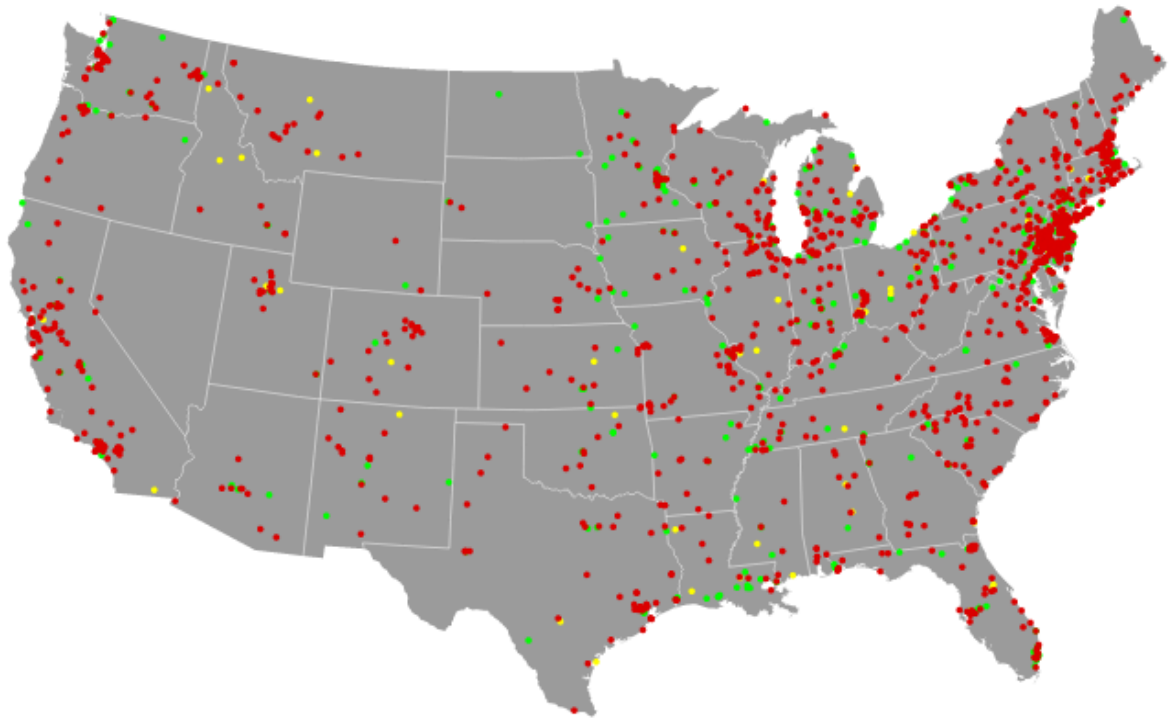


Get in touch at info@Statvis.com to learn more about the only

Environmental Data Intelligence System™

statvis Trust Through Data

United States Superfund Sites

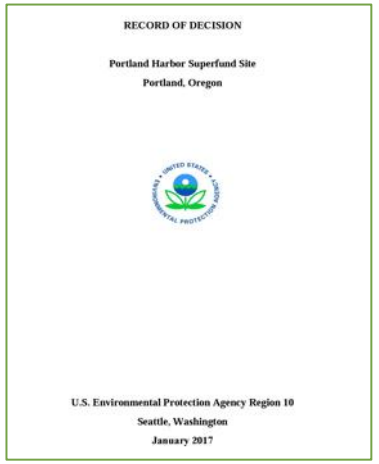


- Sites under EPA control due to hazardous materials that require remediation
- ~1400 sites across US
- All different types of contaminants
  - > 600 chemicals
  - Pd (43%), trichloroethylene (42%), Cr (35%), benzene (34%), PERC (28%), As (28%) and toluene (27%)
  - PCBs, PCDD/Fs, PAHs, PFAS/PFOA
- Some sites have 100+ years of industrial operations

Superfund Sites – Remediation Estimates

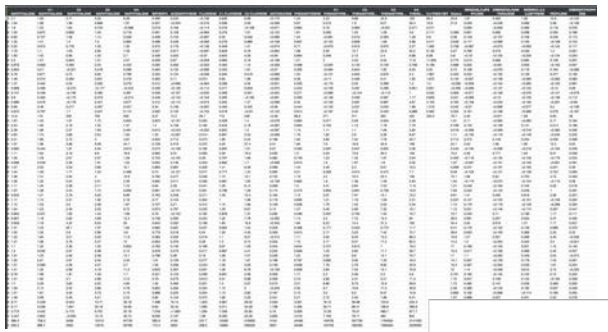
Typical Superfund Site Project Workflow

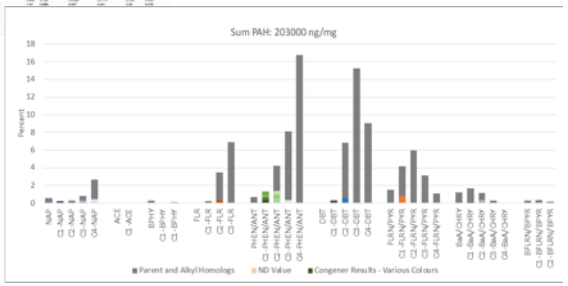
- Review historical reports and data
- Compile and wrangle historical data
- Process data through interactive data workflow
  - Receptor modelling/source apportionment
- Report and present findings
- Depositions, rebuttal reports and testifying

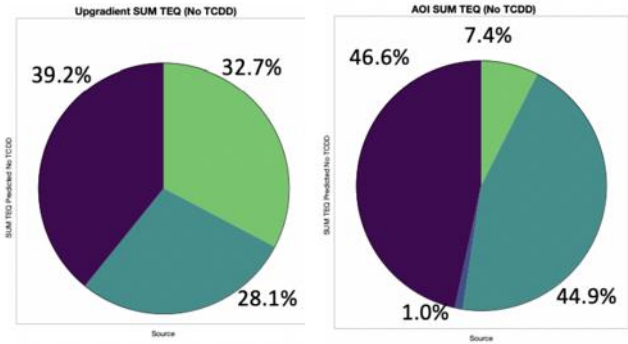

RECORD OF DECISION

Portland Harbor Superfund Site
Portland, Oregon

U.S. Environmental Protection Agency Region 10
Seattle, Washington
January 2017



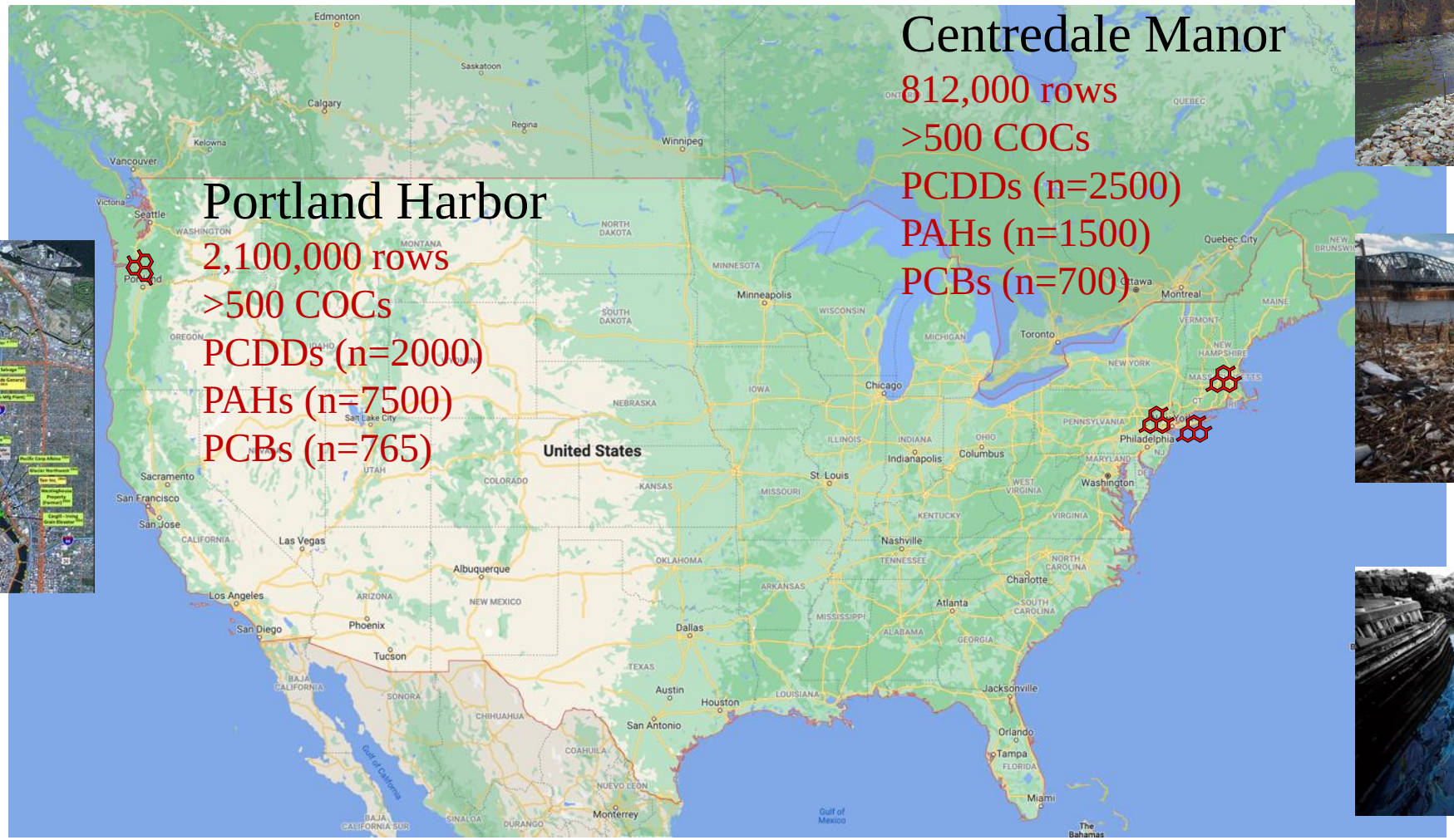




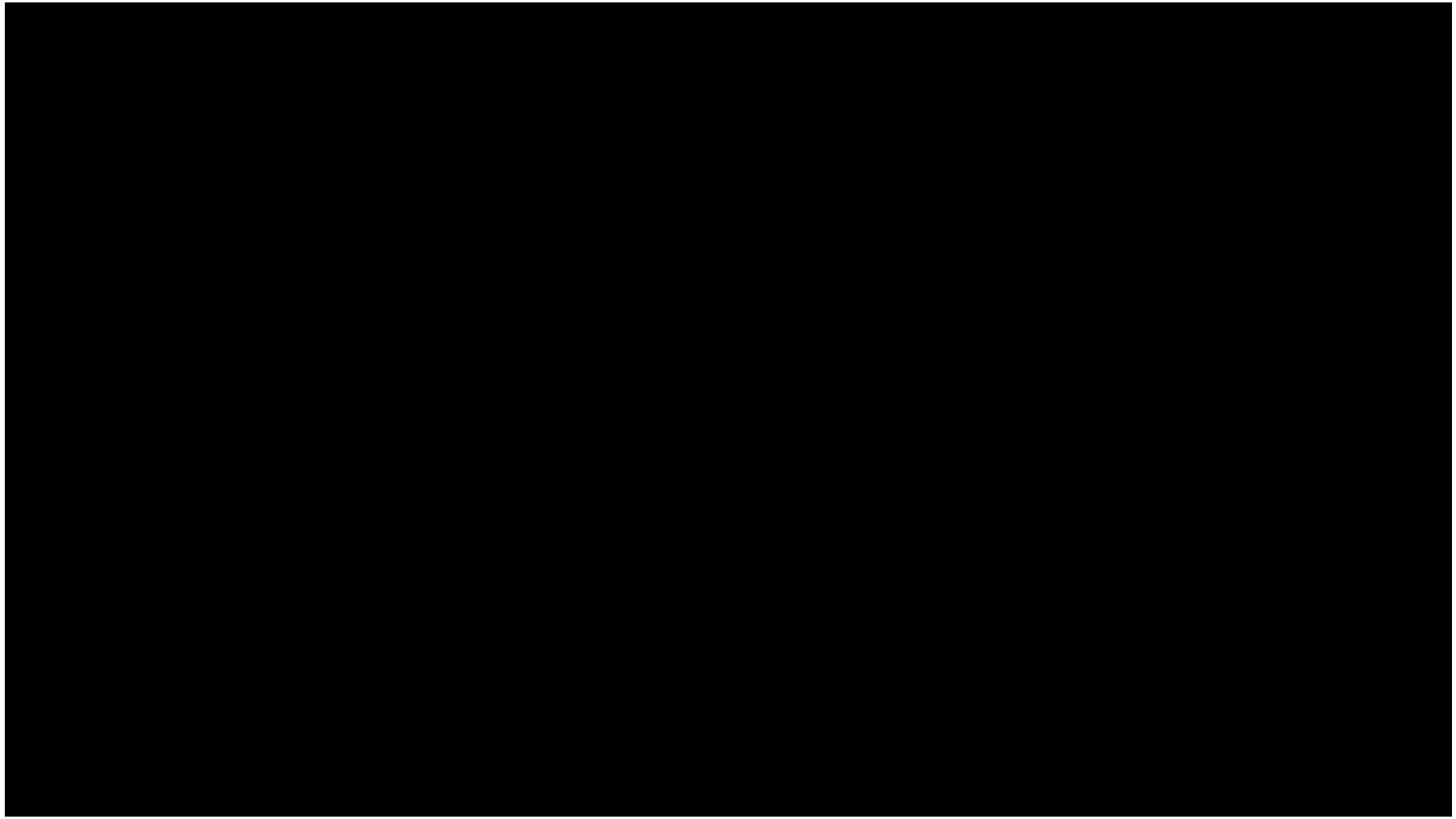
Data Wrangling

- Data import and merge
- Dataset familiarization
- MAR – Missing at Random values
- Duplicates, replicates, repeat analytical results
- Data tidying (unit consolidation, co-elutions)
- Detection Limits – ½ DL, KM, Multiplicative Simple Replacement, Multiple Lognormal Replacement (Maximum Likelihood or Robust Estimates)
- Data screening
  - minimum detects per analyte
  - minimum detects per sample
- Fundamental data coding – up/downstream, labs, investigations, locations
- Initial data review

Superfund Sites – Lots of Data to Work With



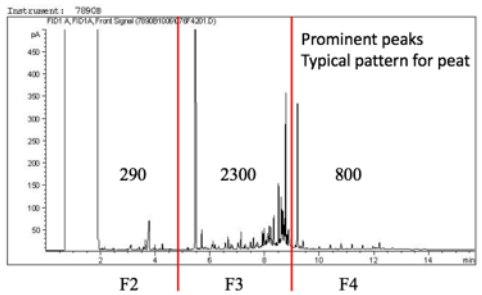
Portland Harbour Superfund Site - PAH Data



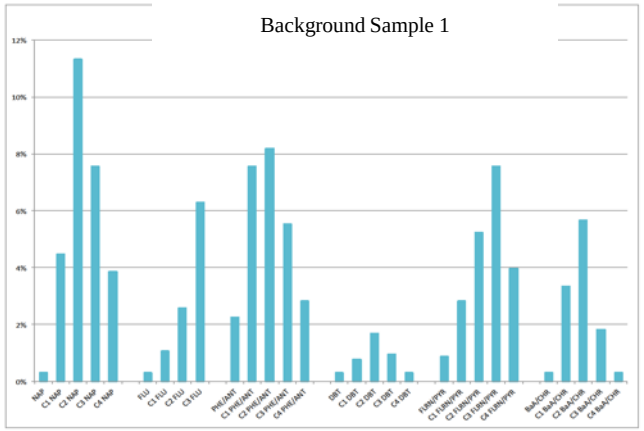
Requirements for Data Discovery

- **Imperative to have data interactivity to discover what the data is telling you and to develop knowledge**
  - ***The goal is 'data mastery'***
- This can only be accomplished with commercial statistical software, like SAS JMP
  - Can be expensive
  - Licensed per user
  - Hard to share with clients or pass along data knowledge
- We are currently developing these data discovery tools for everyday sites in Statvis

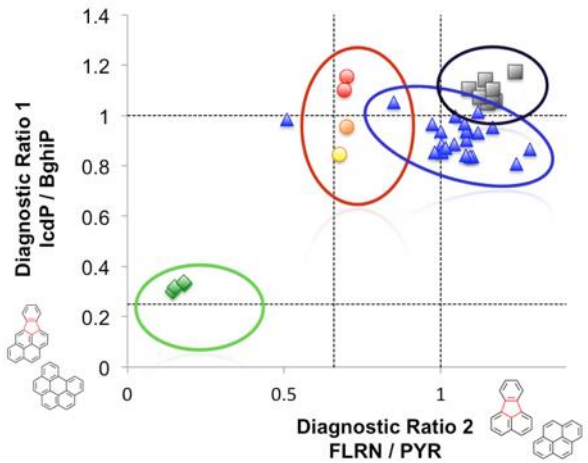
Progression of Chemical Fingerprinting in Environmental Forensics



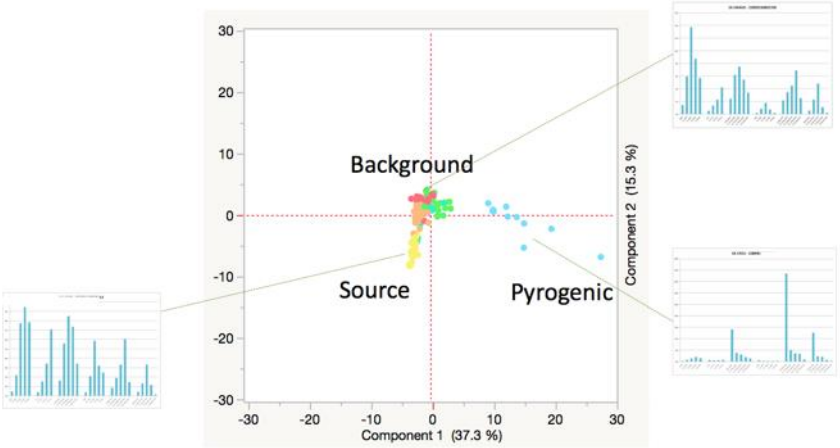
Chromatogram Analysis



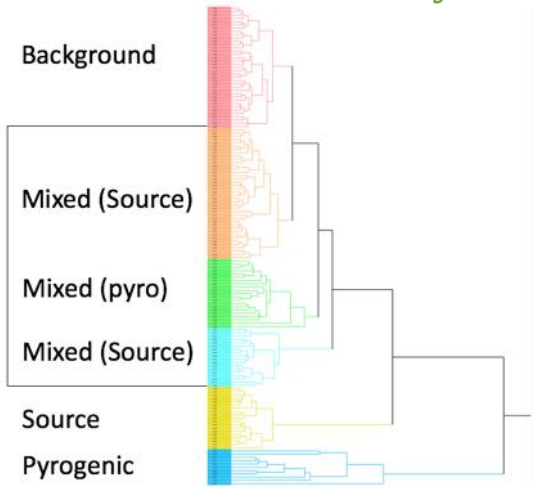
Individual Sample Pattern Analysis



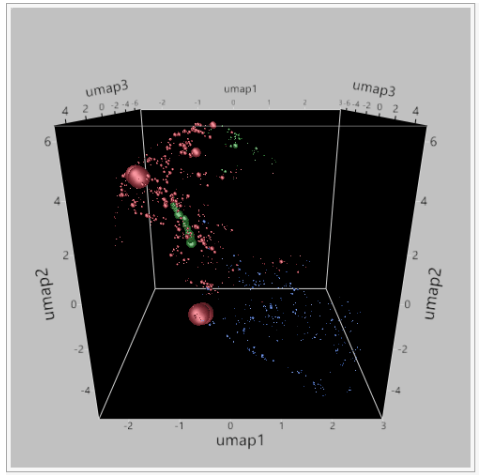
Double Ratio Plots / Diagnostic Ratios



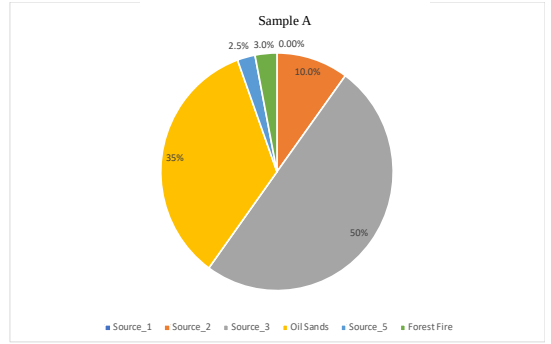
PCA



HCA



UMAP

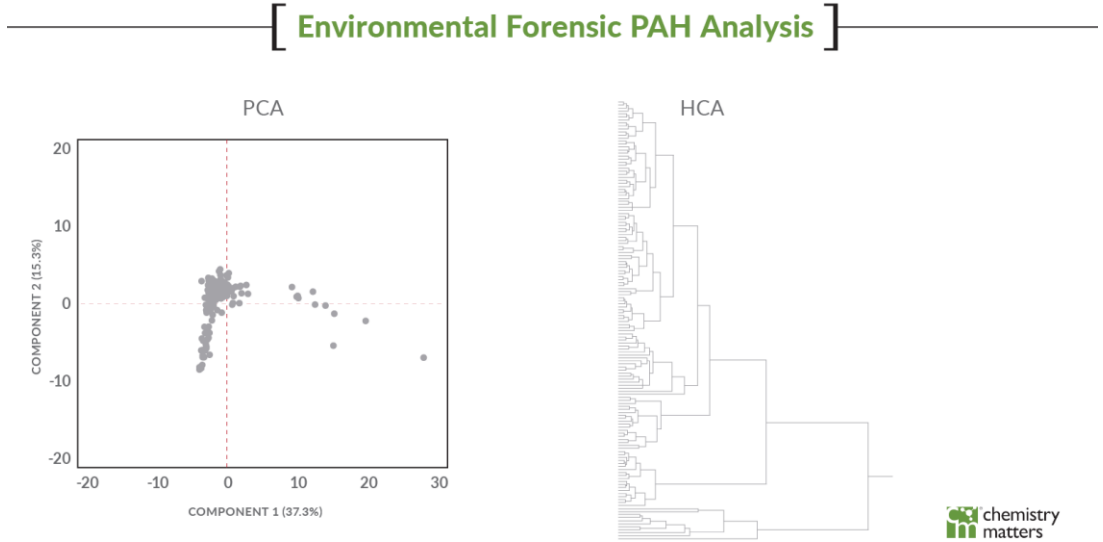


Source Apportionment

Portland Harbour PAH Fingerprinting

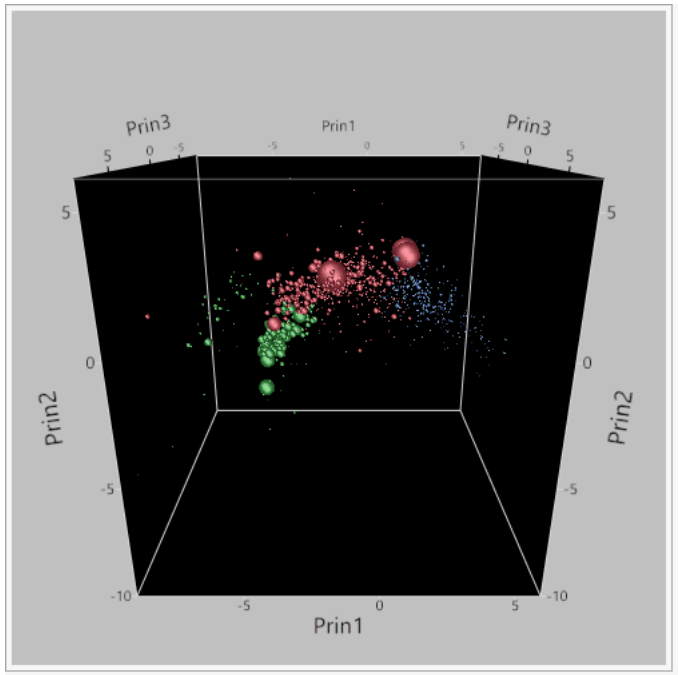
Integration and interactivity is the key to scientific communication

- To build trust and buy-in, data mastery is essential
  - Data at your fingertips
  - Able to ‘show’ the answer
    - Seeing is believing, even with complex data
- Integrated statistical analysis coupled with geospatial analysis with visualization allows learning while you are analyzing
- An interactive data portal that allows the user to interact and demonstrate results all while communicating and explaining its meaning

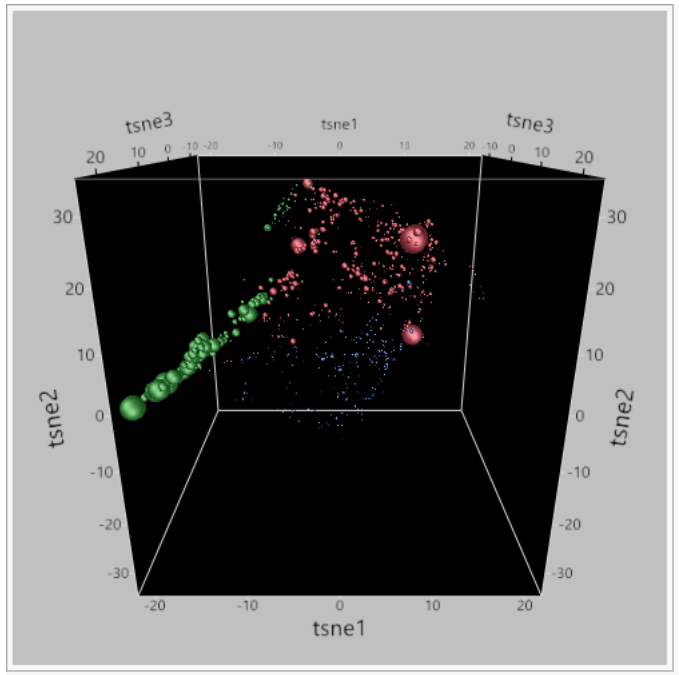


Different Techniques for ~~Dimensionality Reduction~~ Data Visualizations

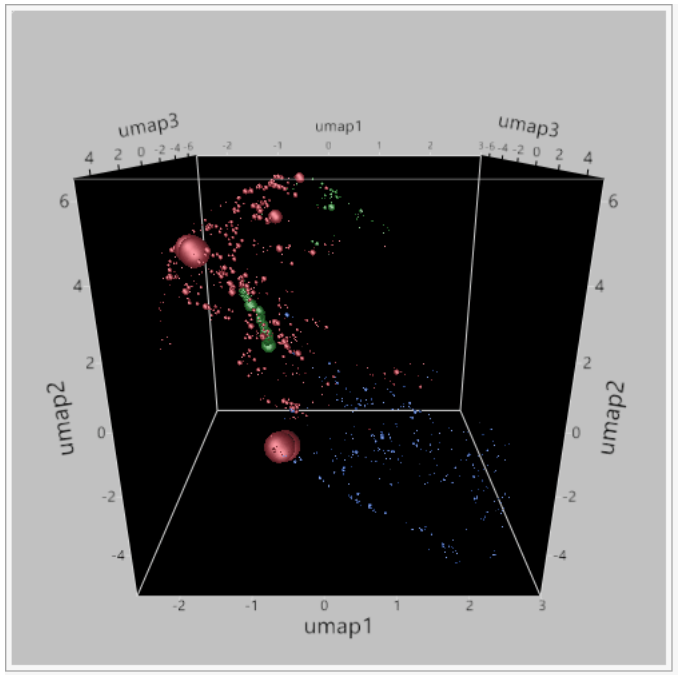
Principal Component Analysis



t-Distributed Stochastic Neig



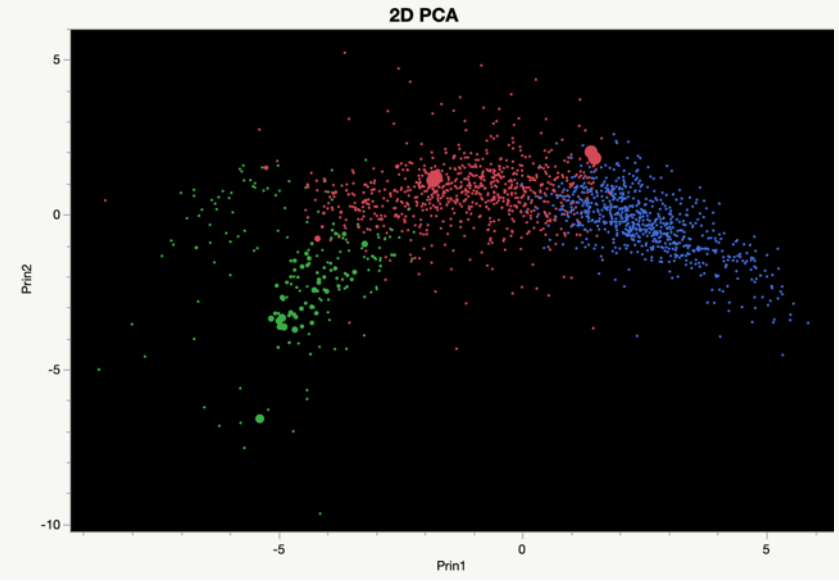
Projection



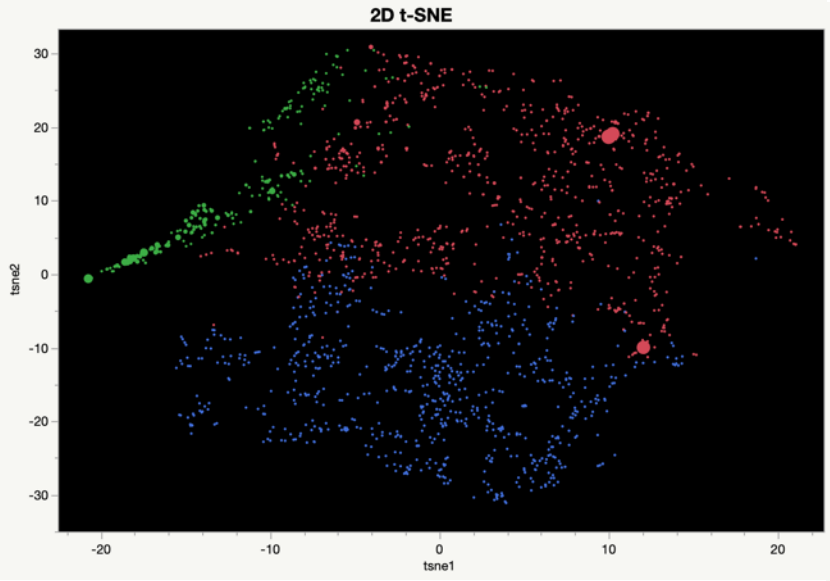
- Each is a different technique used to reduce number of dimensions but keep data variance and/or data structure
- Helps visualize data in 2 or 3D space

Different Techniques for Data Visualizations

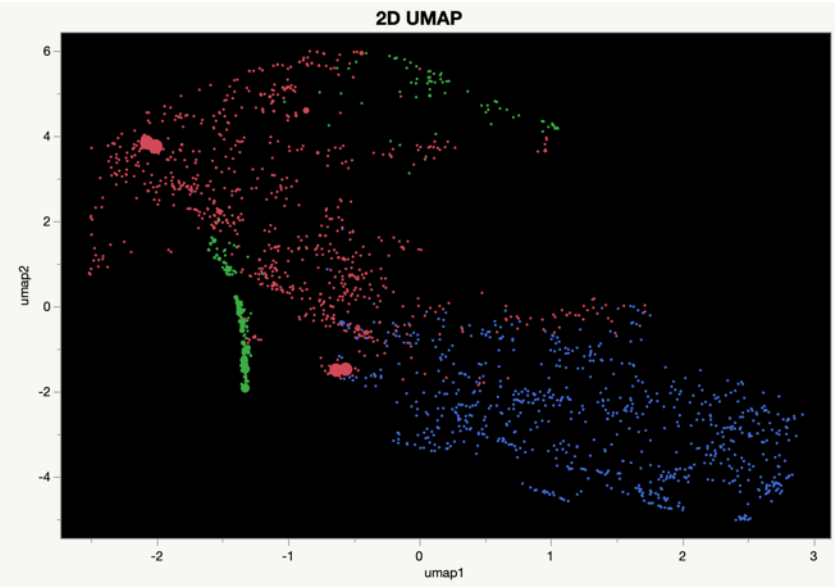
PCA



t-SNE

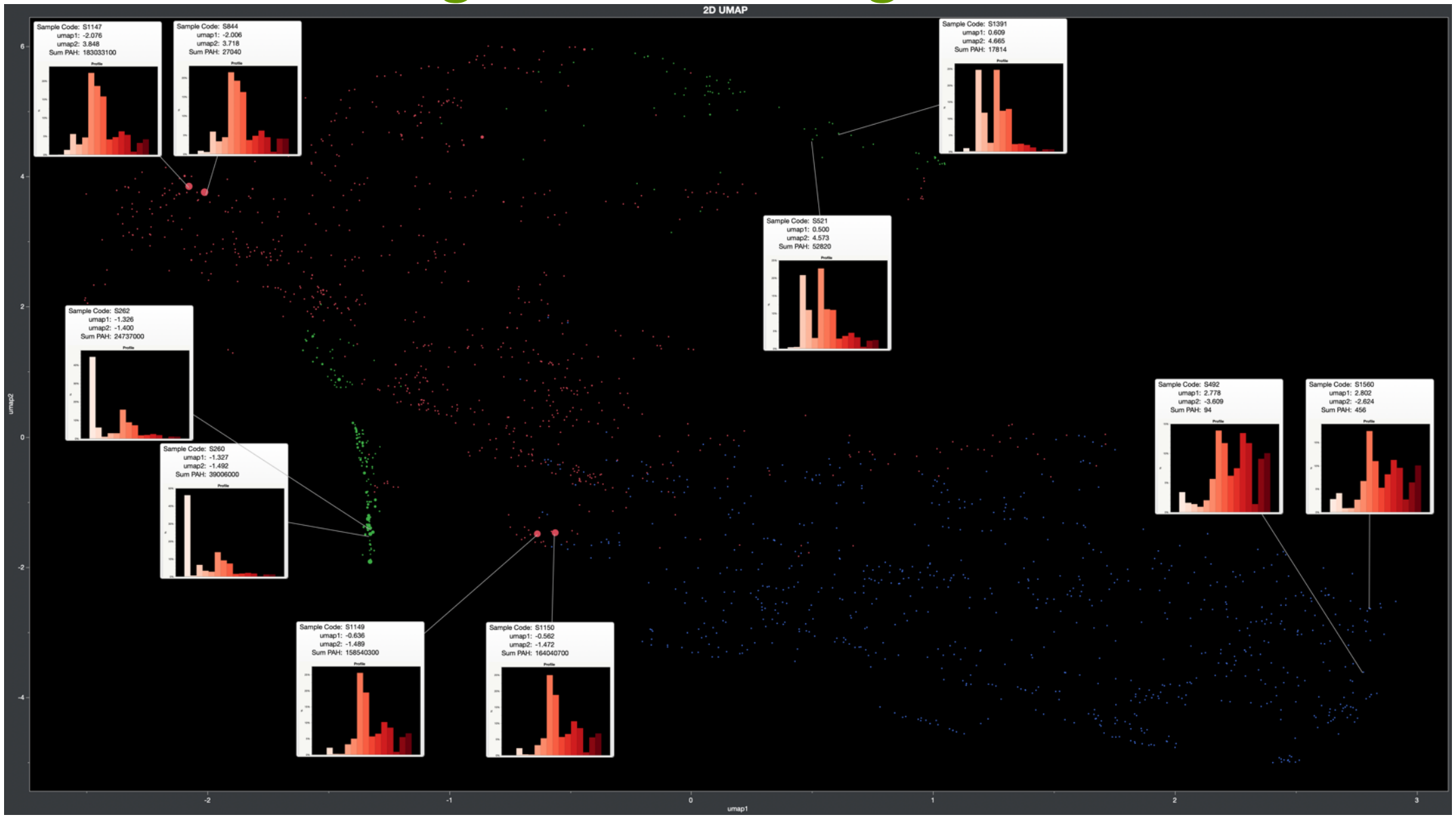


UMAP

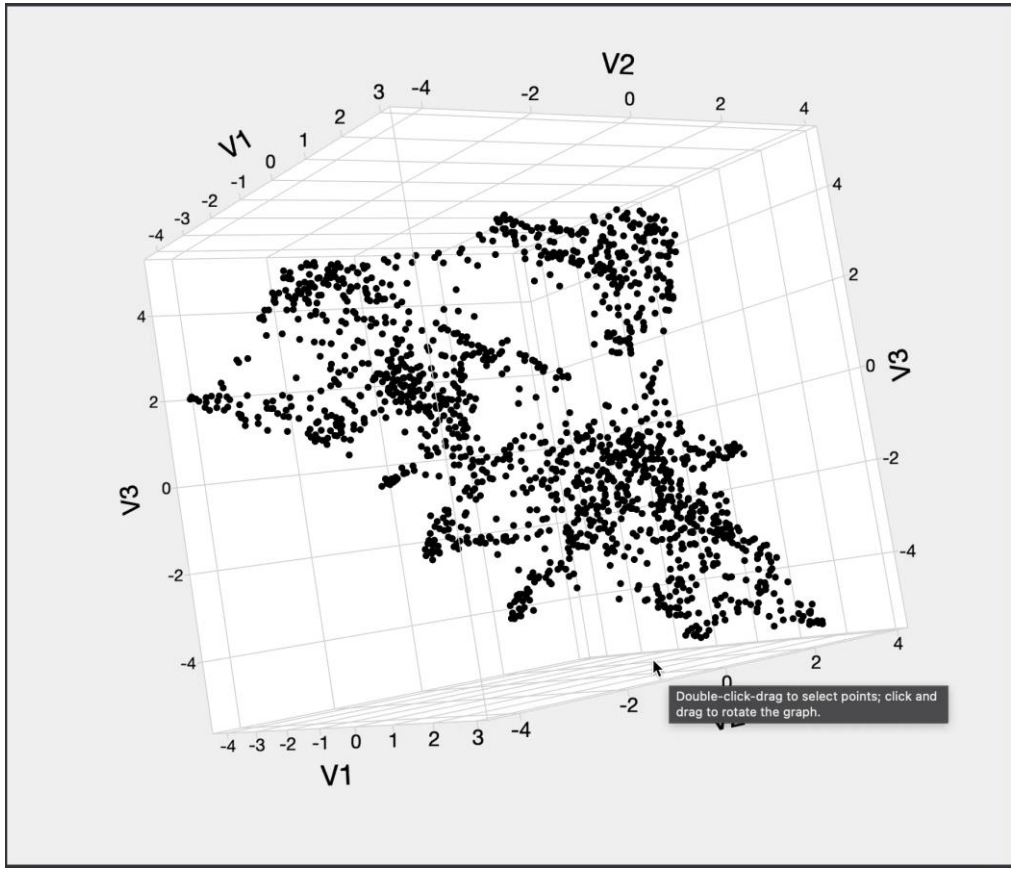


- Provides visualization of all compounds – patterns and concentrations
- Each has different math to derive visualization
- UMAP is newest kid on the block and my personal favourite

Visualizing and Interacting with the Data

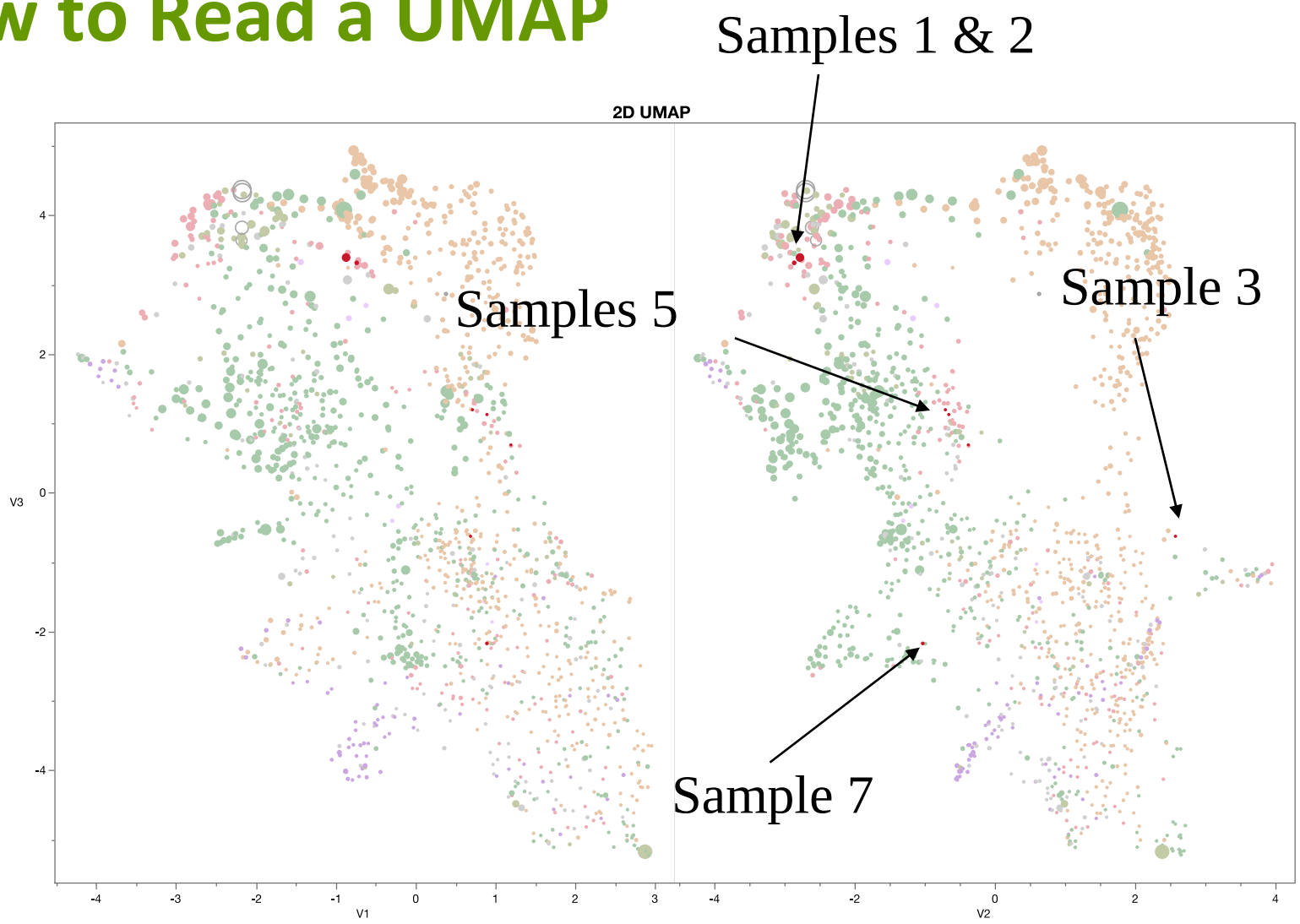


How to Read a UMAP



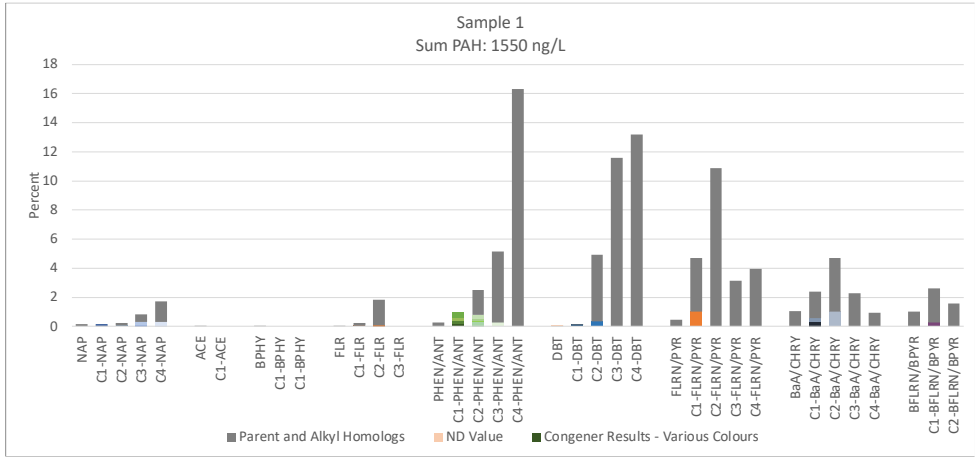
How to Read a UMAP

- Samples close together are more similar
- Samples that are far away from each other are different
- Comparing all samples to Sample 1

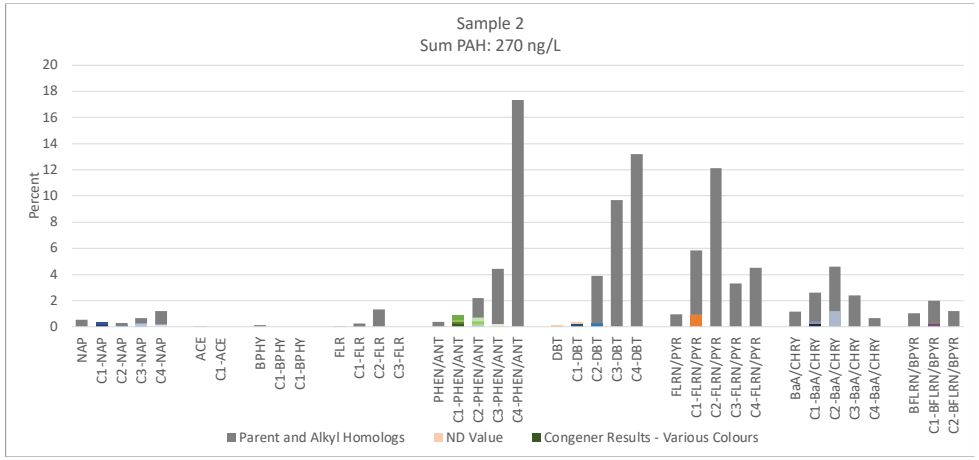


How to Read a UMAP

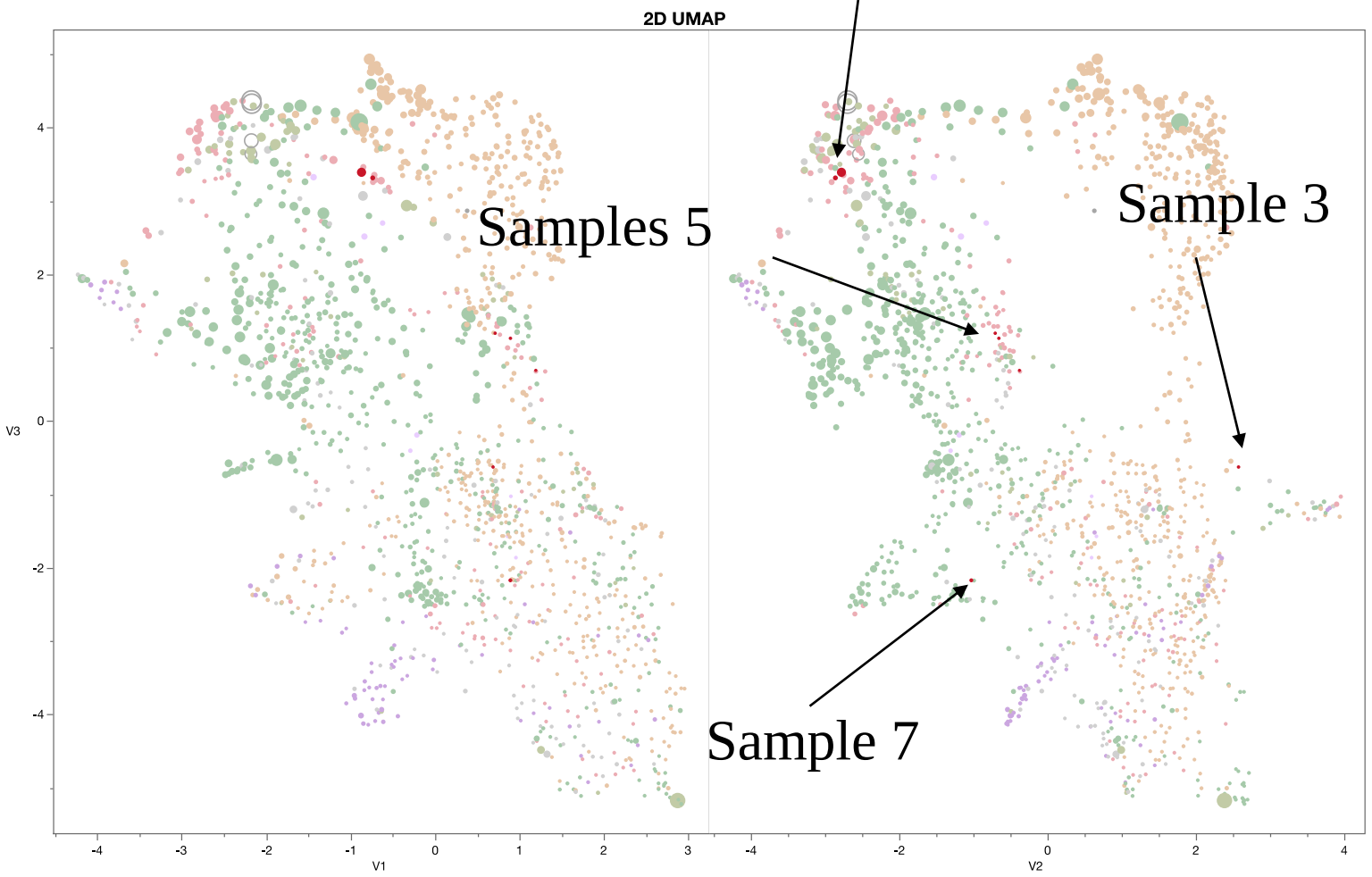
Sample 1



Sample 2

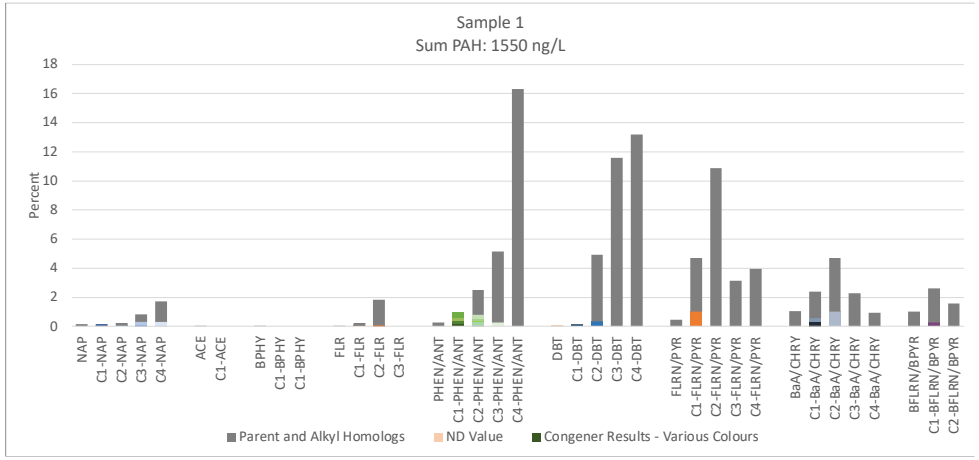


$\cos(\theta) : 0.995$

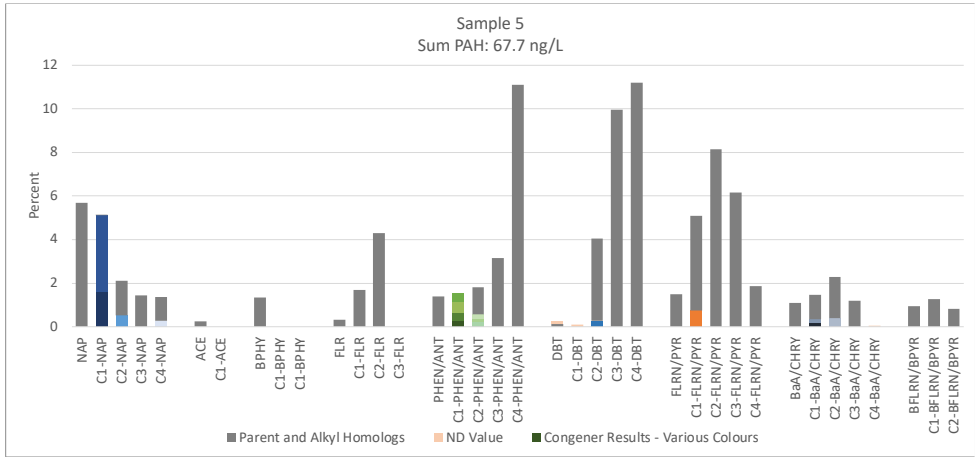


How to Read a UMAP

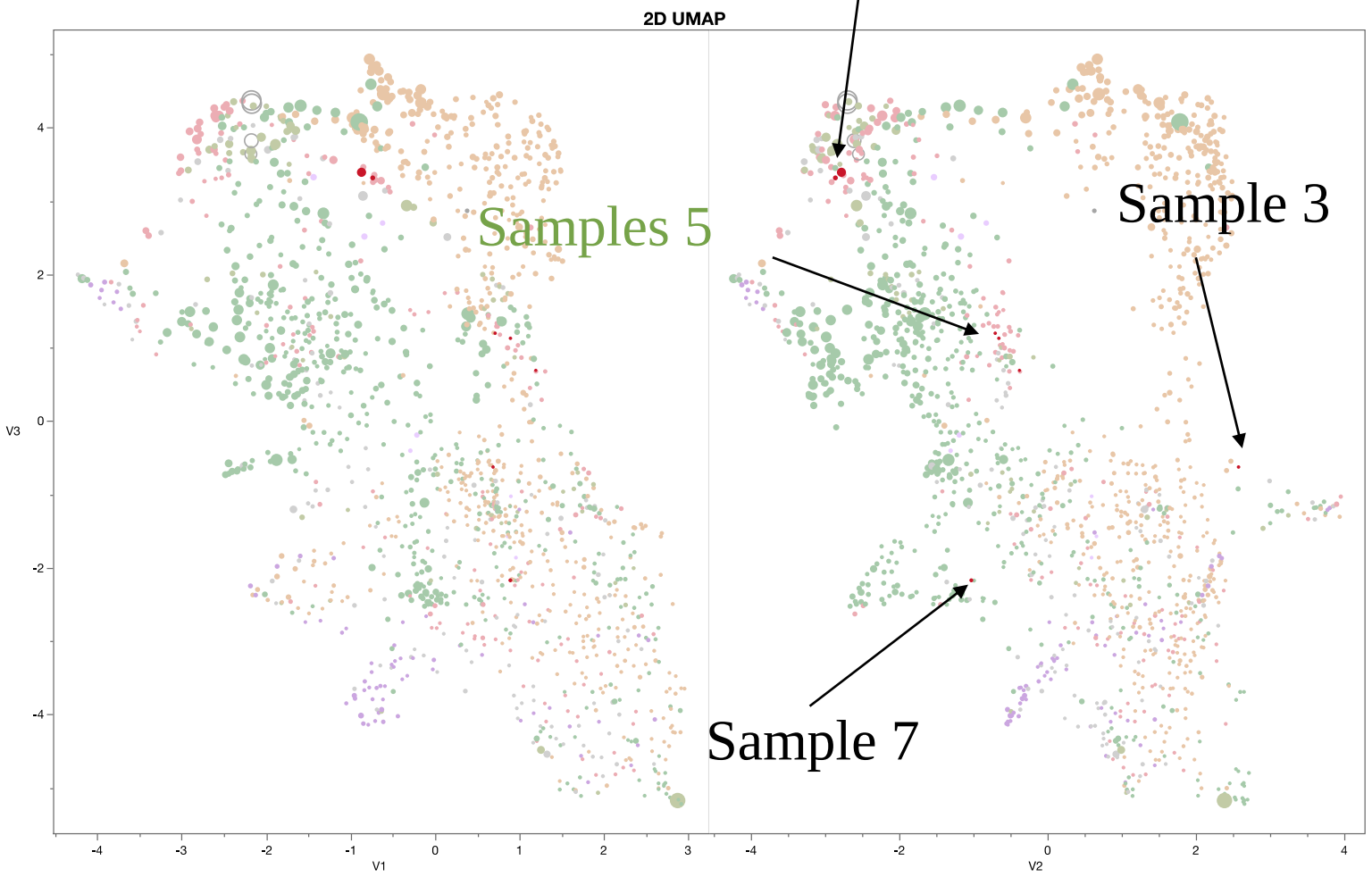
Sample 1



Sample 5

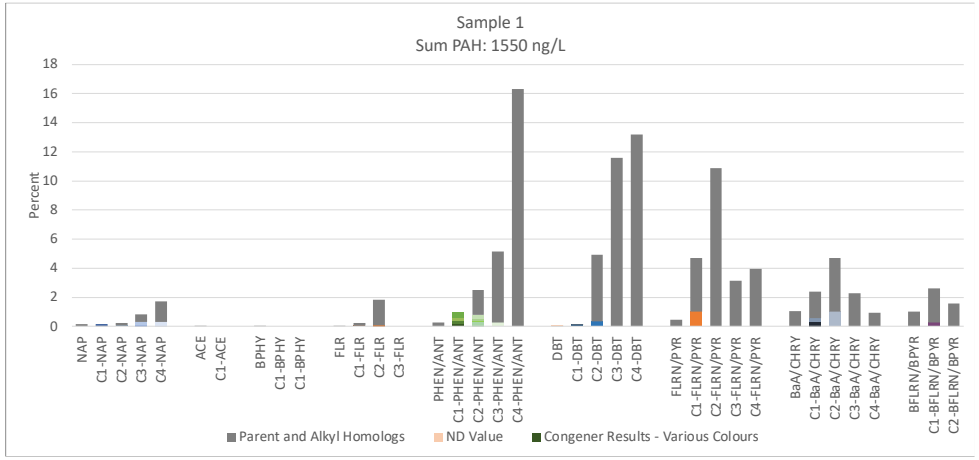


$\cos(\theta) : 0.904$

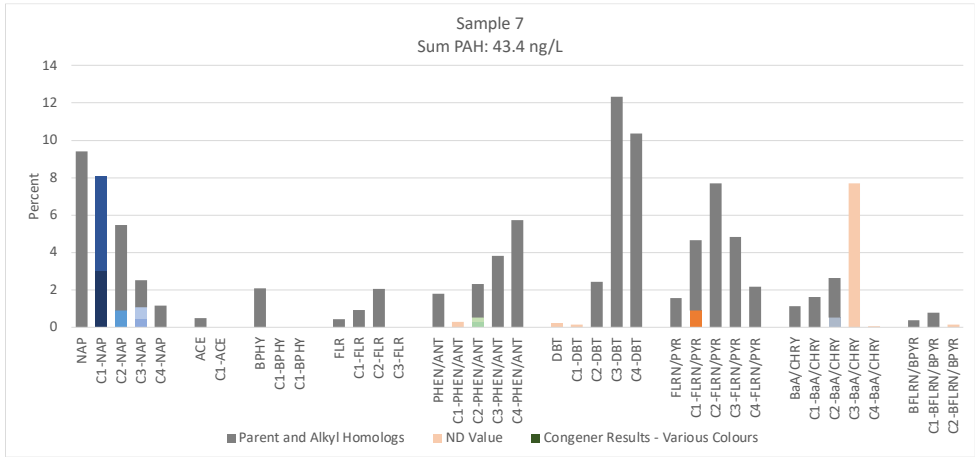


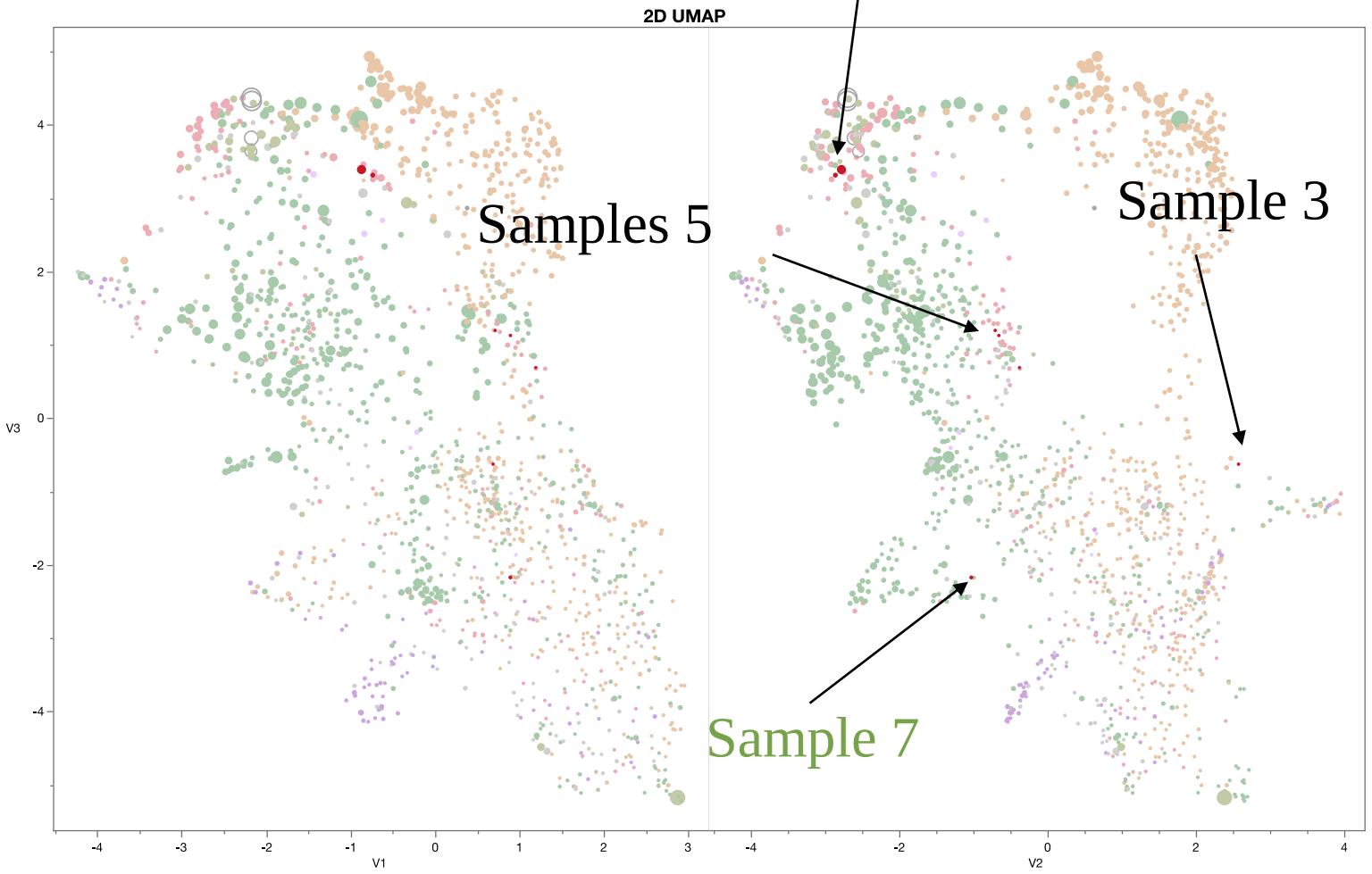
How to Read a UMAP

Sample 1



Sample 7

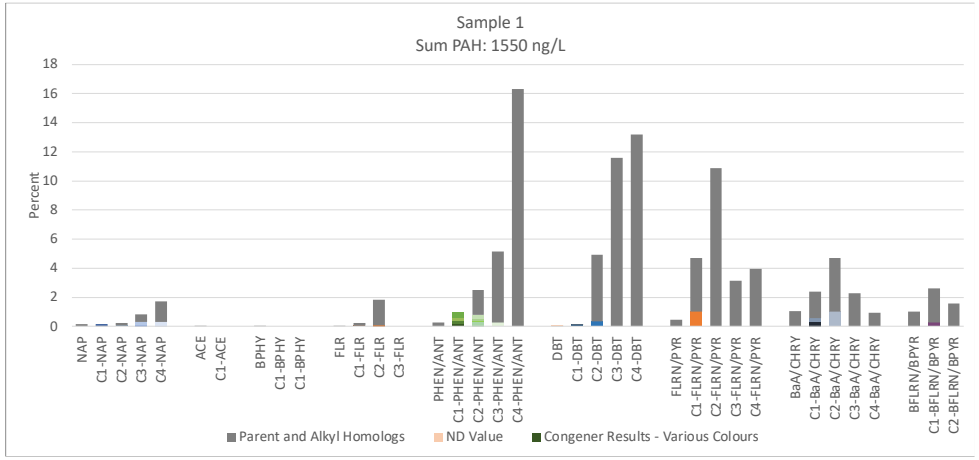




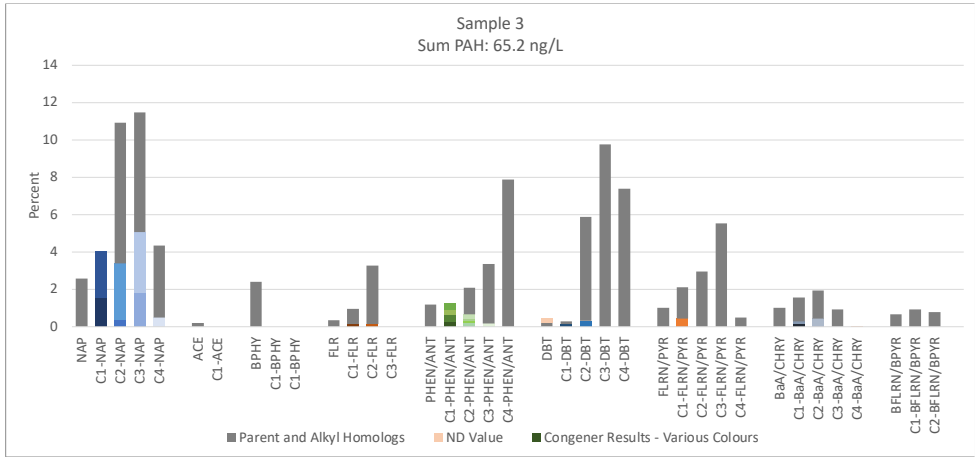
$\cos(\theta) : 0.774$

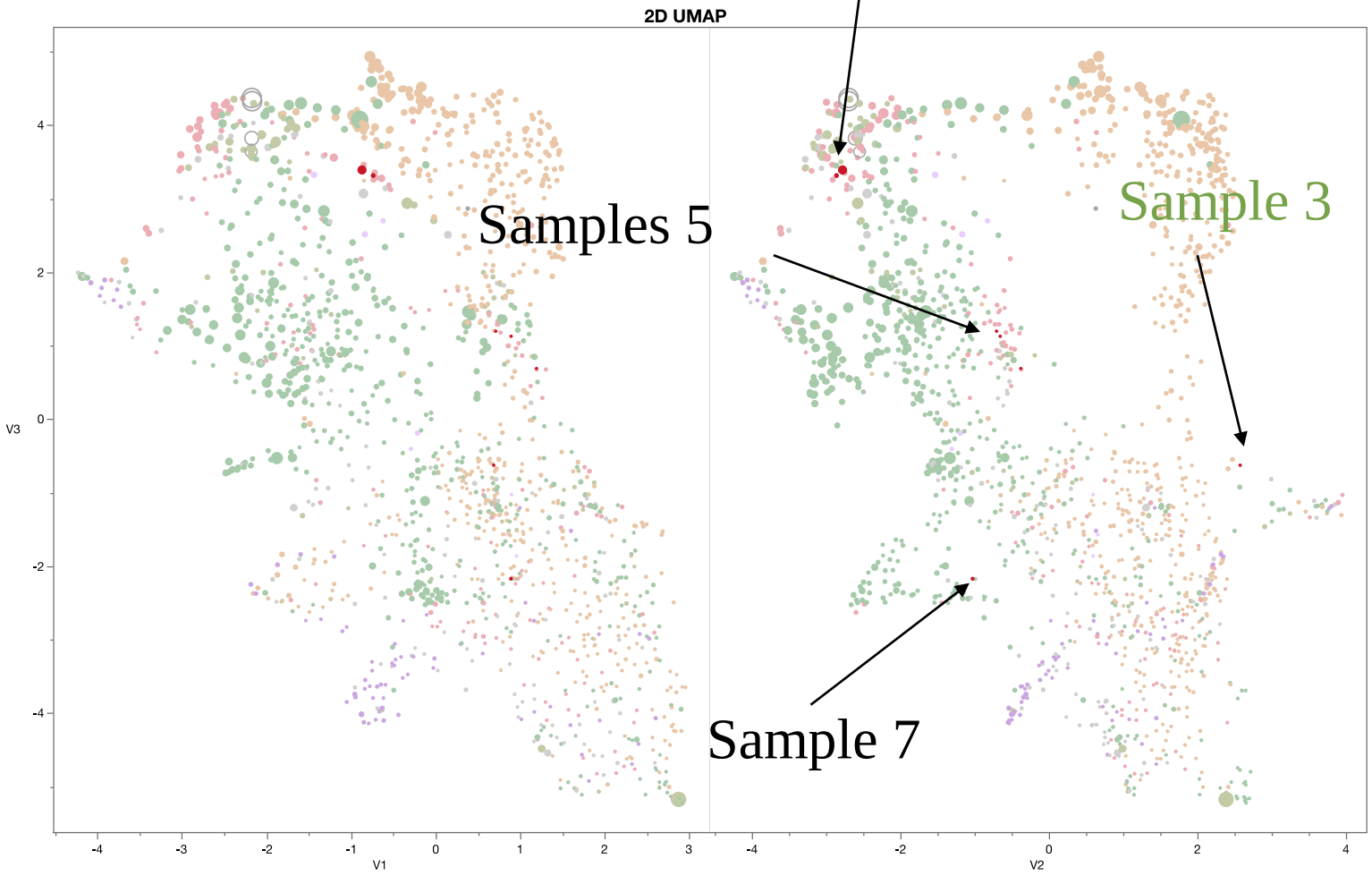
How to Read a UMAP

Sample 1



Sample 3

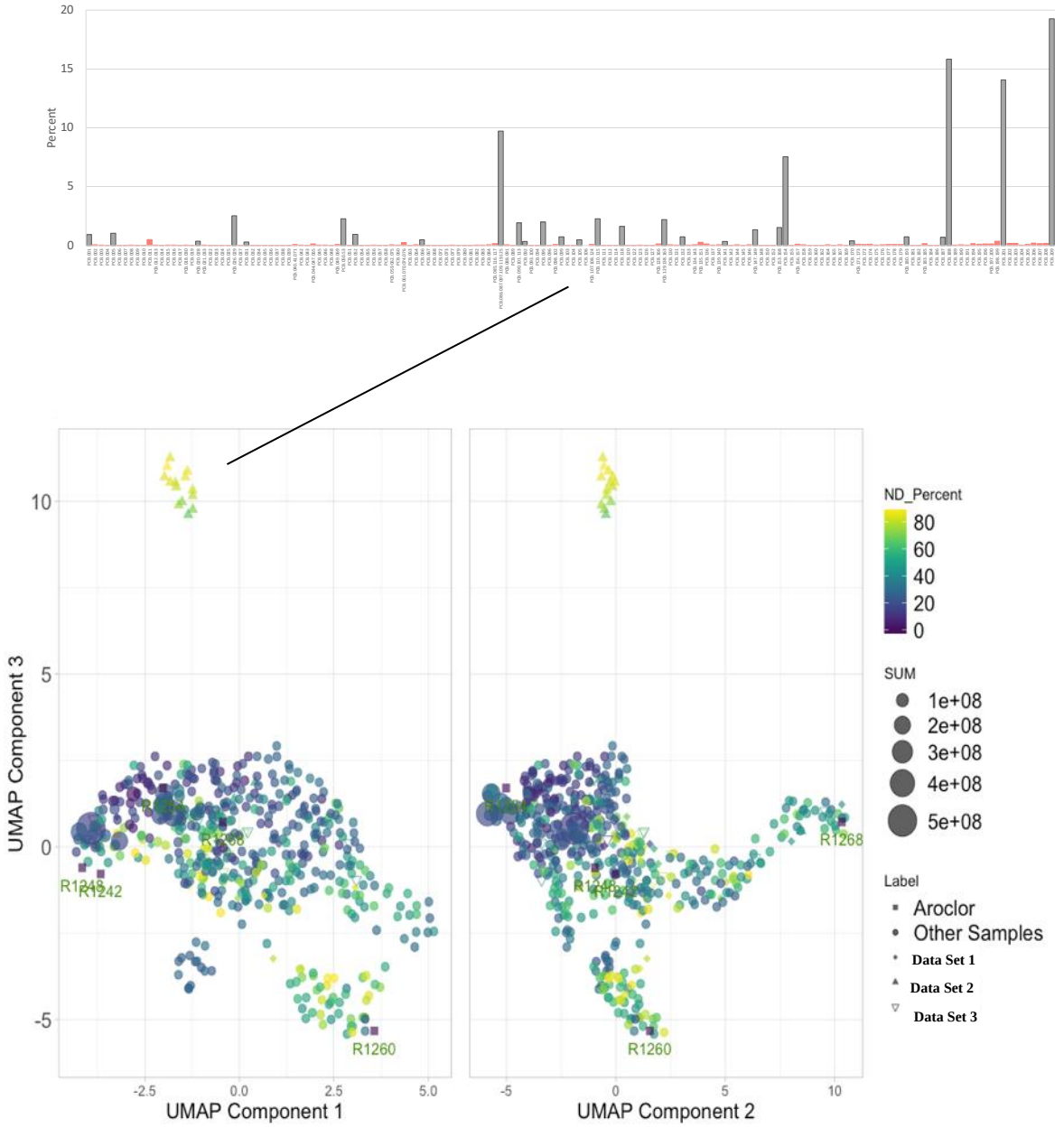




$\cos(\theta) : 0.675$

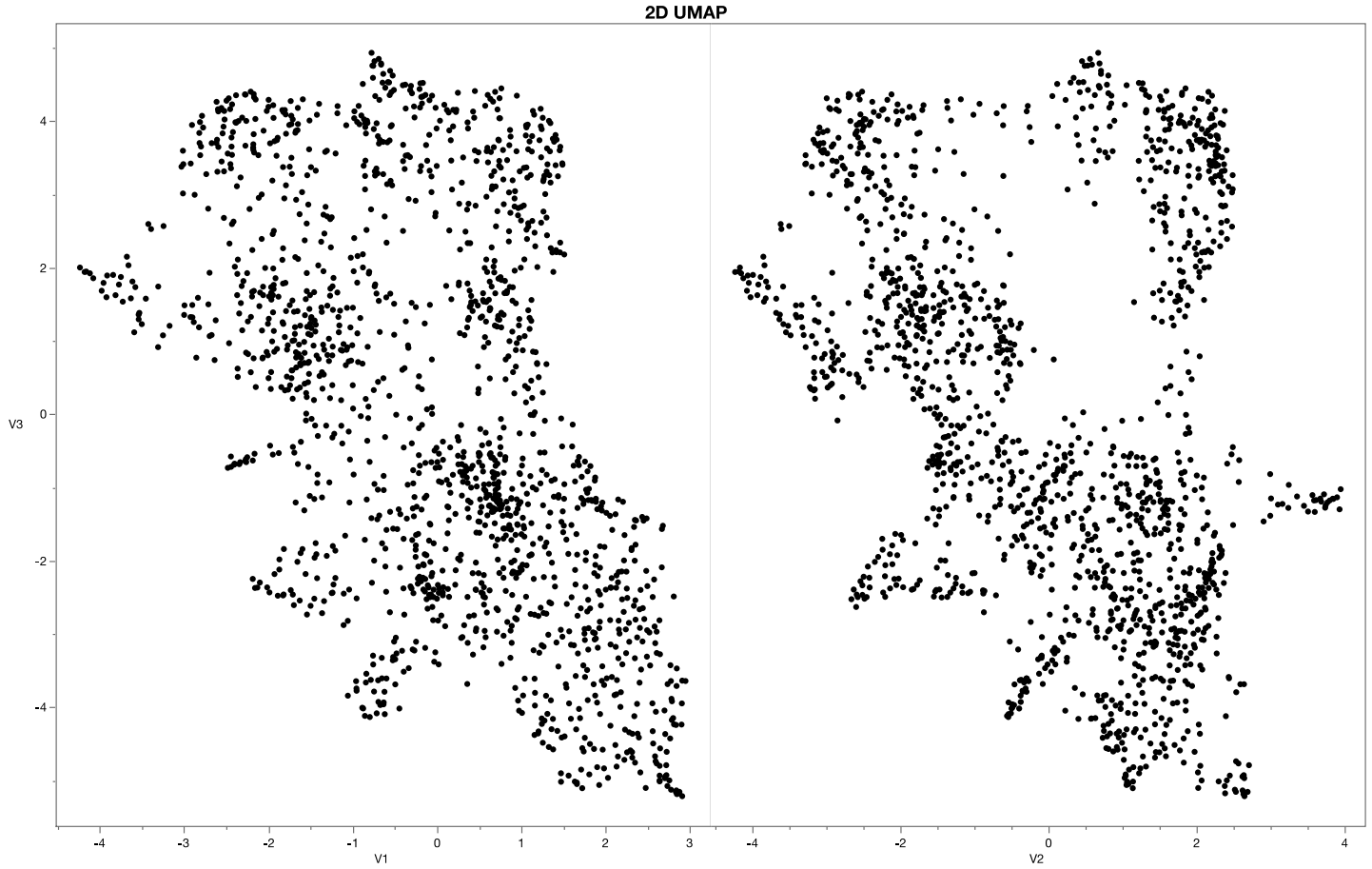
UMAP – Finds Errors

- Preliminary look at data can help
- UMAP can find outliers and bad data
- We found a bad set of results for one site
- Odd fingerprint
- Data before and after this batch of samples looked normal (all same laboratory)
- Samples all collected from similar locations
- This dataset was erroneously reported by laboratory
- It was removed from further analysis



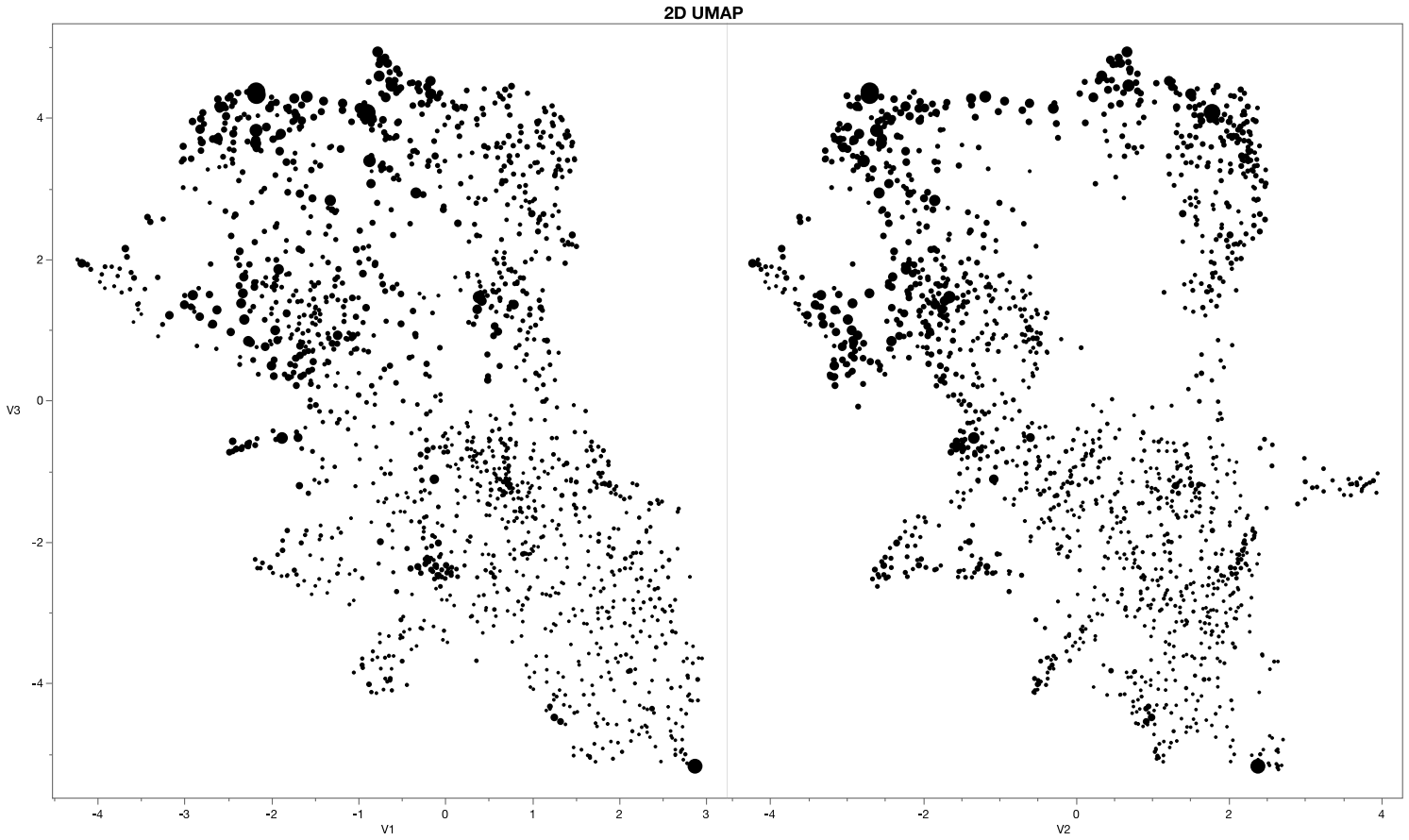
Base UMAP

- Base UMAP shows unique data structure for all compounds (fingerprint) included in the database
- Data nodes (edges, extremes) provide clues as to the potential source-like samples in the dataset
- This dataset has multiple nodes



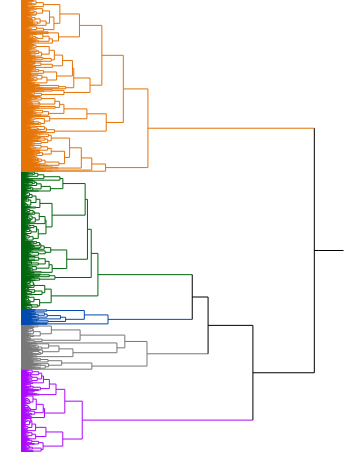
Base UMAP – Size = Concentration

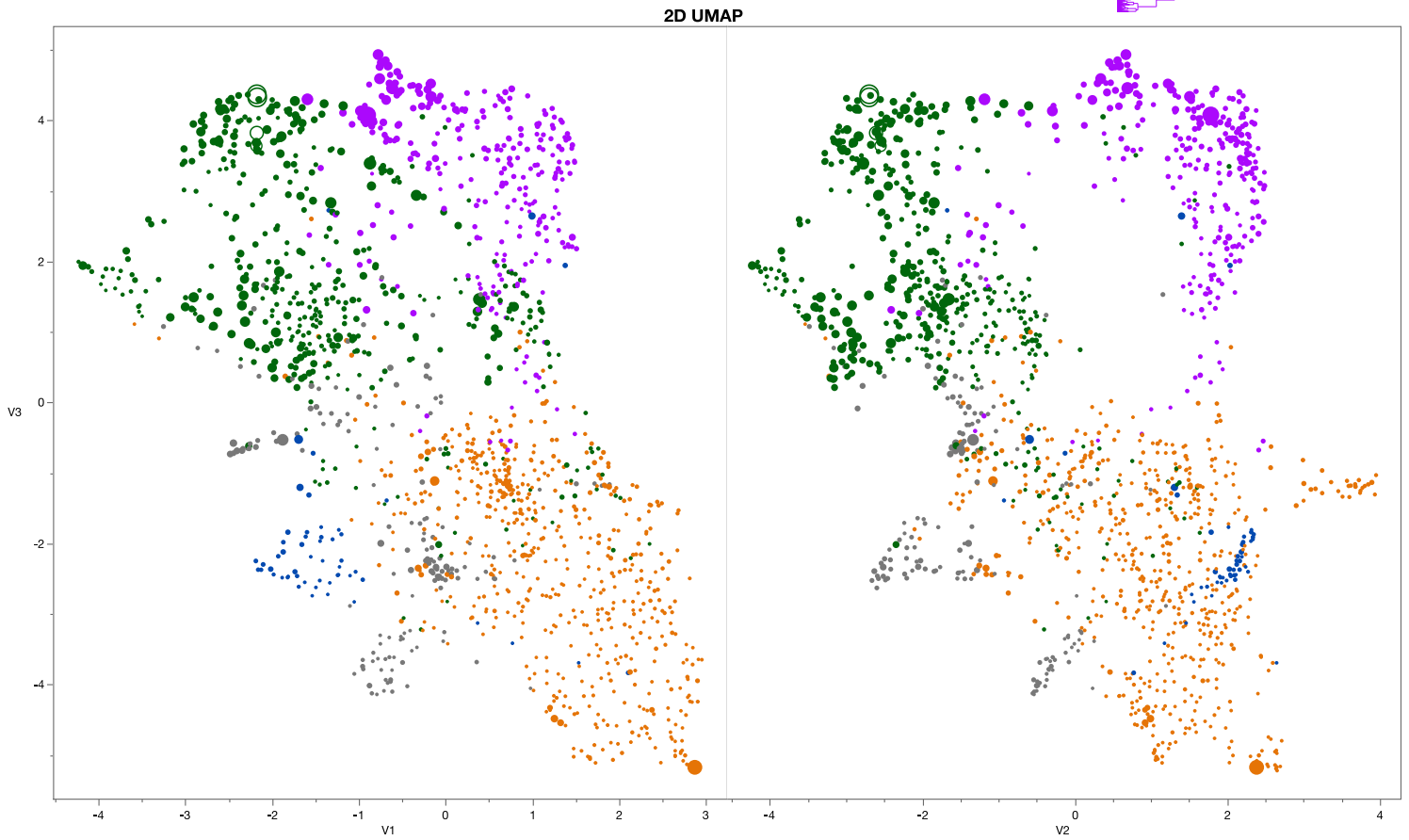
- Adding additional dimension to the dataset shows high concentration samples
- High concentrations generally associated with most source-like samples
  - Shows sources that are most important in this dataset
- Coupled with map, can show where highest concentration samples are located geospatially



UMAP – Add Sample Coding

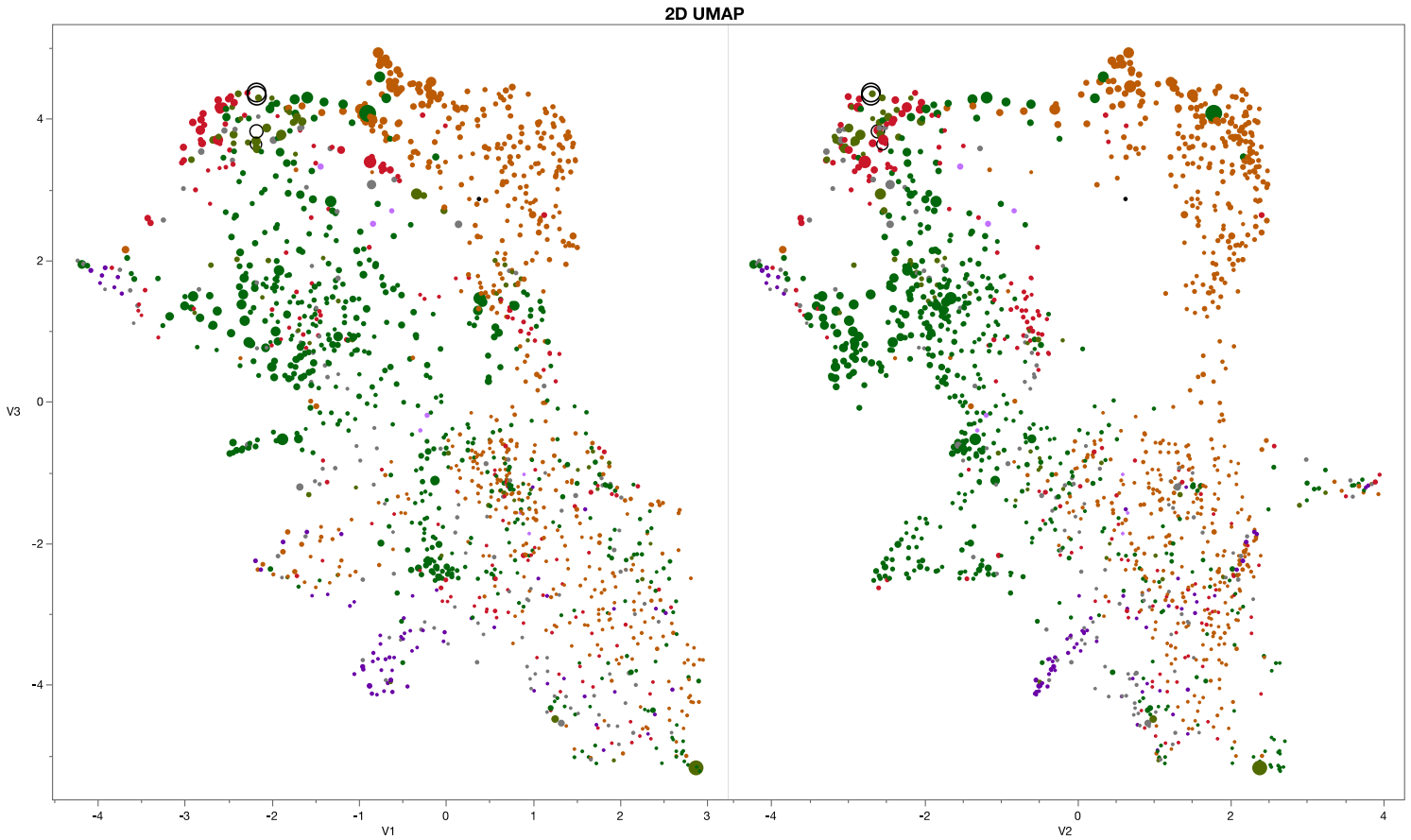
- Coding data to match your dataset helps interpret what UMAP is telling you
  - preliminary hierarchical cluster analysis (HCA)
  - samples coded based on river system
- Other codings
  - sampling events
  - laboratories (to assess variability)
  - receptor modelling results
  - month/year of sampling
- Looking for groups, clusters, reasons for why the data is spread out the way it is





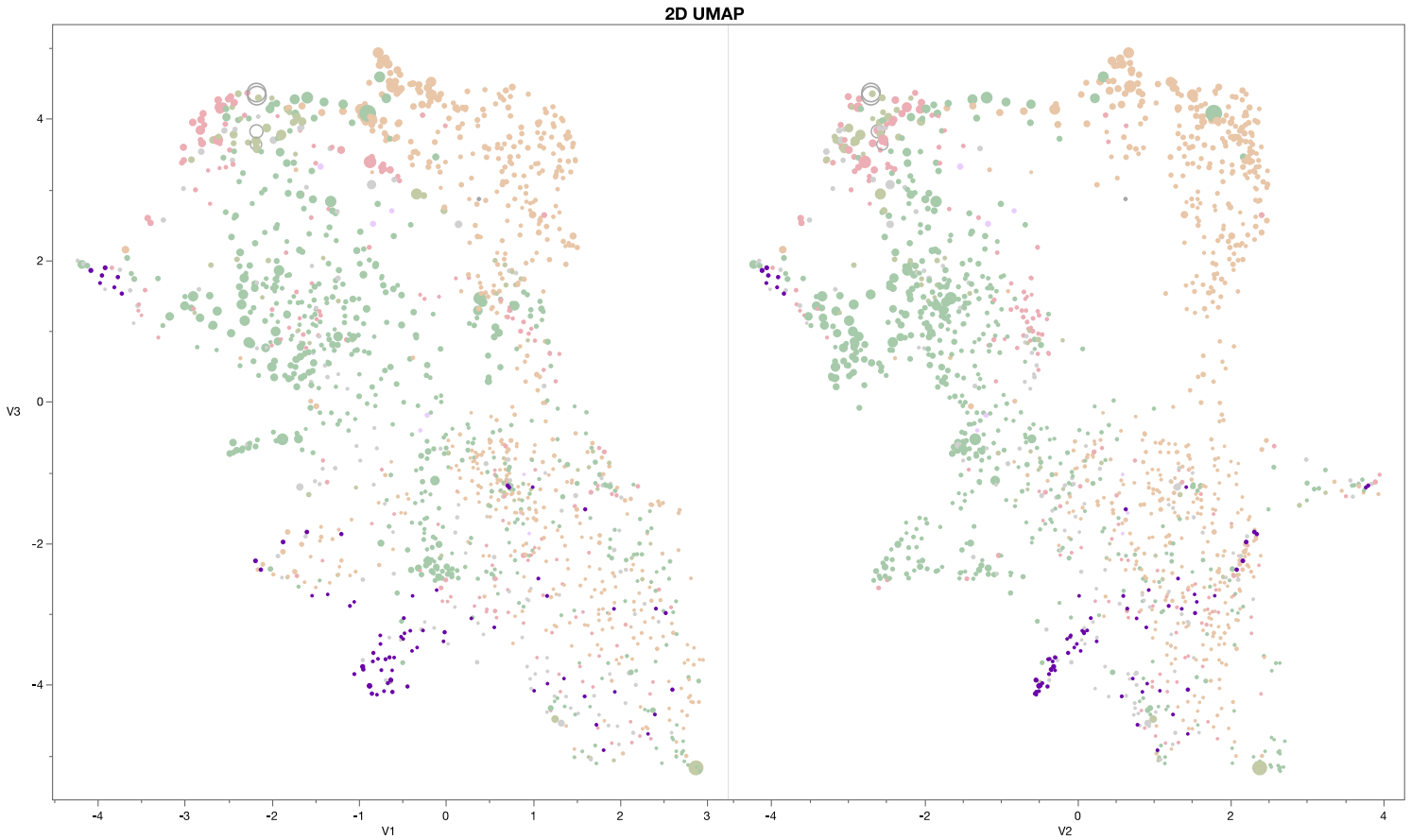
UMAP – Add Sample Coding

- Coding data to match your dataset helps interpret what UMAP is telling you
  - preliminary hierarchical cluster analysis (HCA)
  - samples coded based on river system
- Other codings
  - sampling events
  - laboratories (to assess variability)
  - receptor modelling results
  - month/year of sampling
- Looking for groups, clusters, reasons for why the data is spread out the way it is



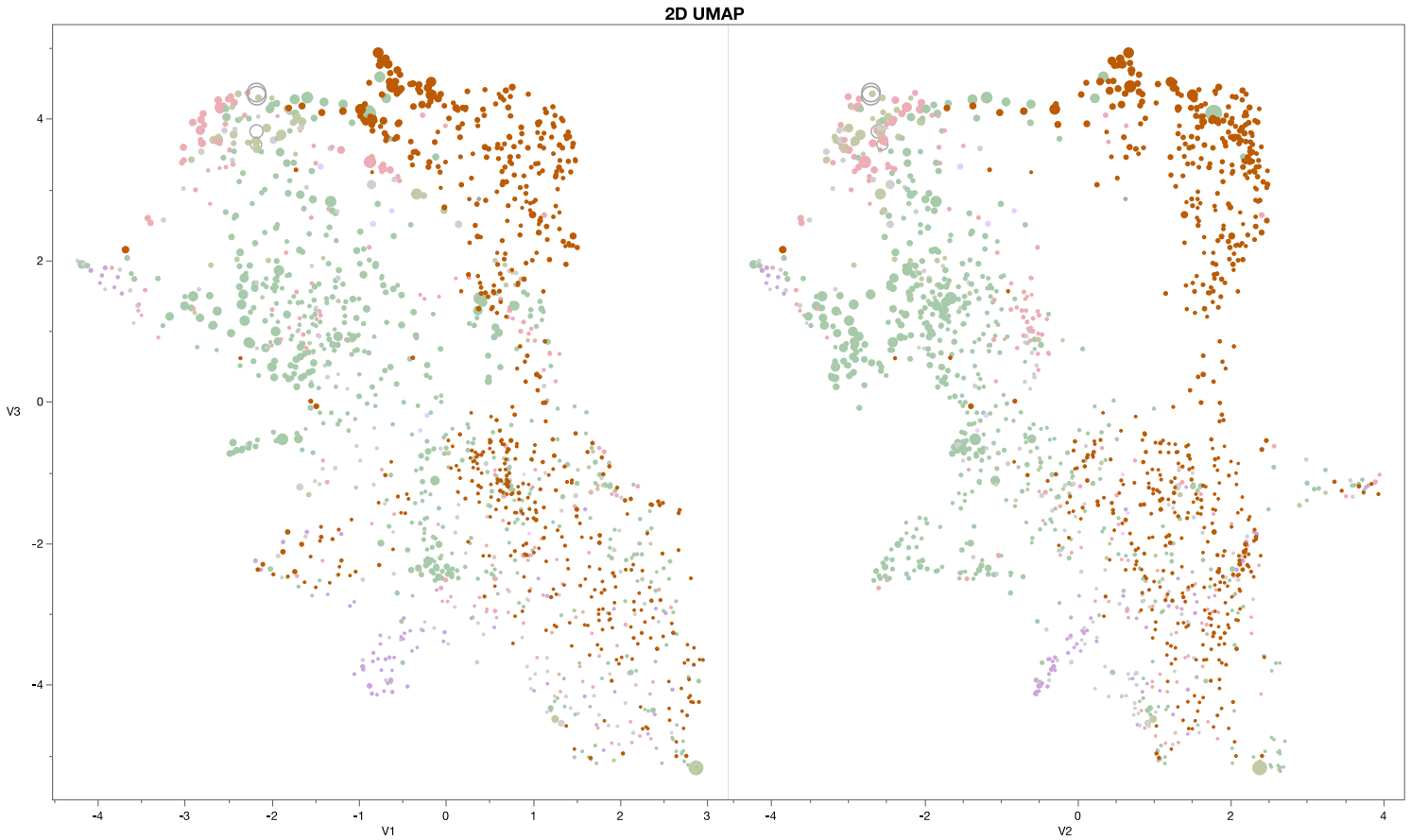
UMAP – Scrolling through River Systems

- Ability to scroll through different river systems helps to identify what the chemical patten of data might be telling you



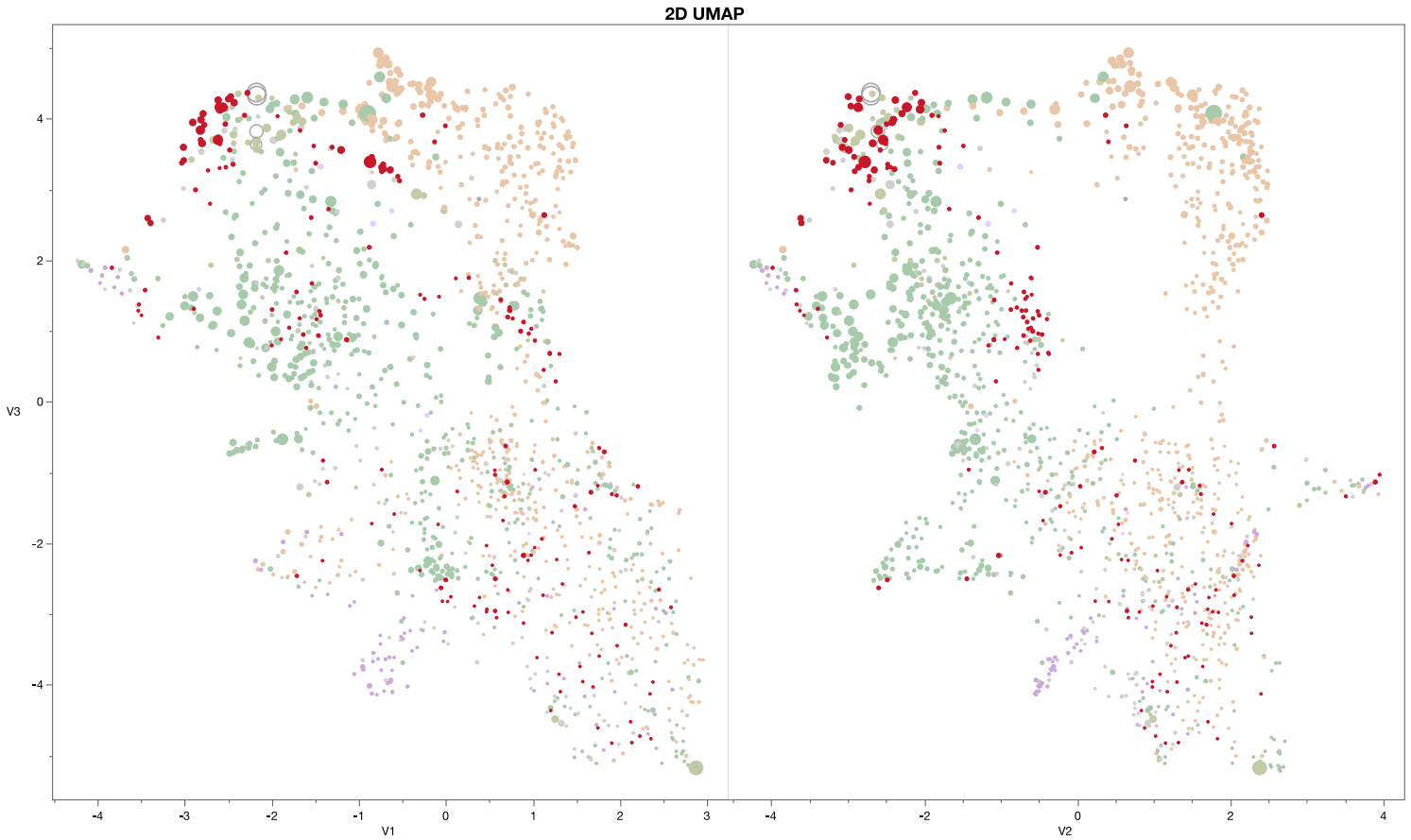
UMAP – Scrolling through River Systems

- Ability to scroll through different river systems helps to identify what the chemical patten of data might be telling you



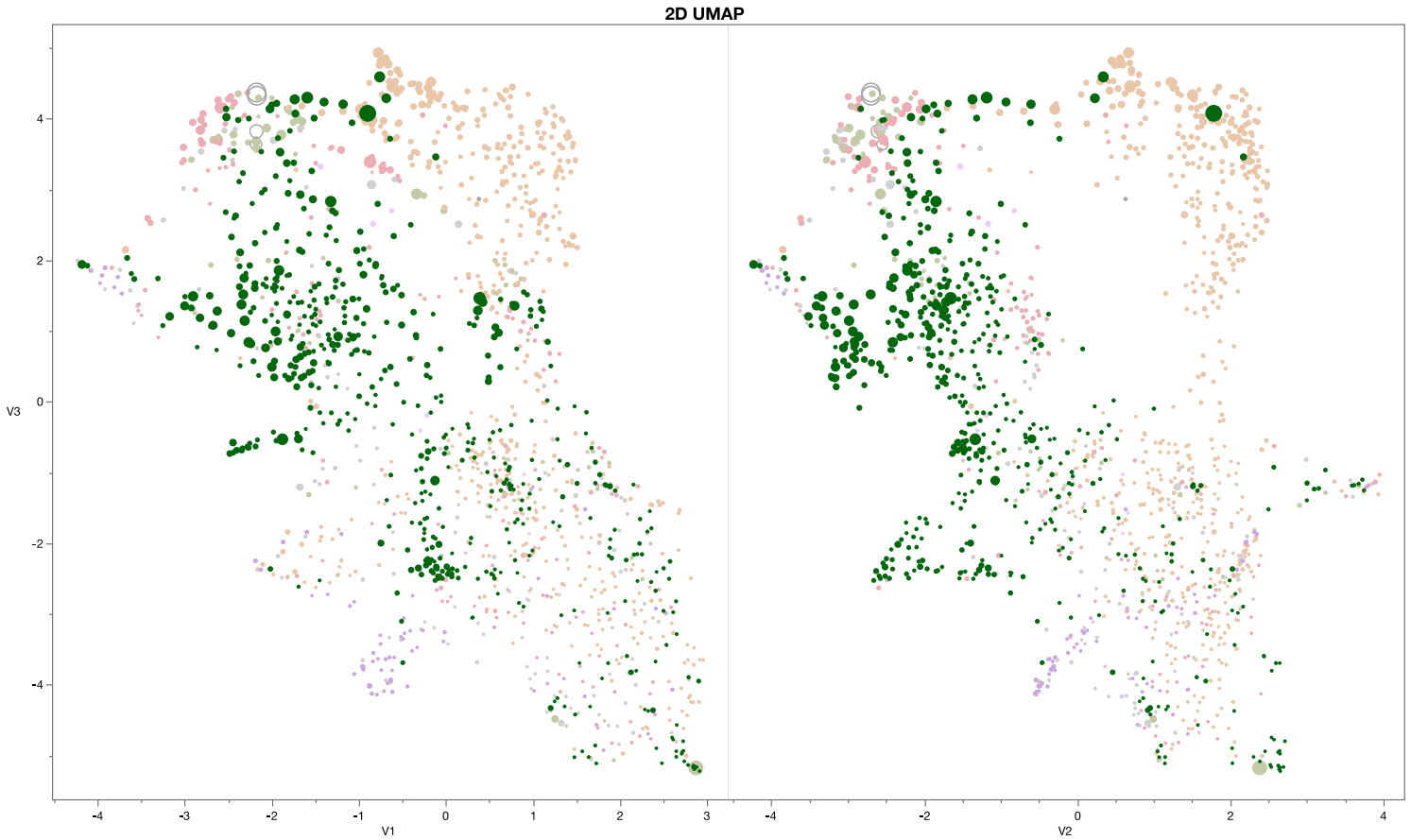
UMAP – Scrolling through River Systems

- Ability to scroll through different river systems helps to identify what the chemical patten of data might be telling you

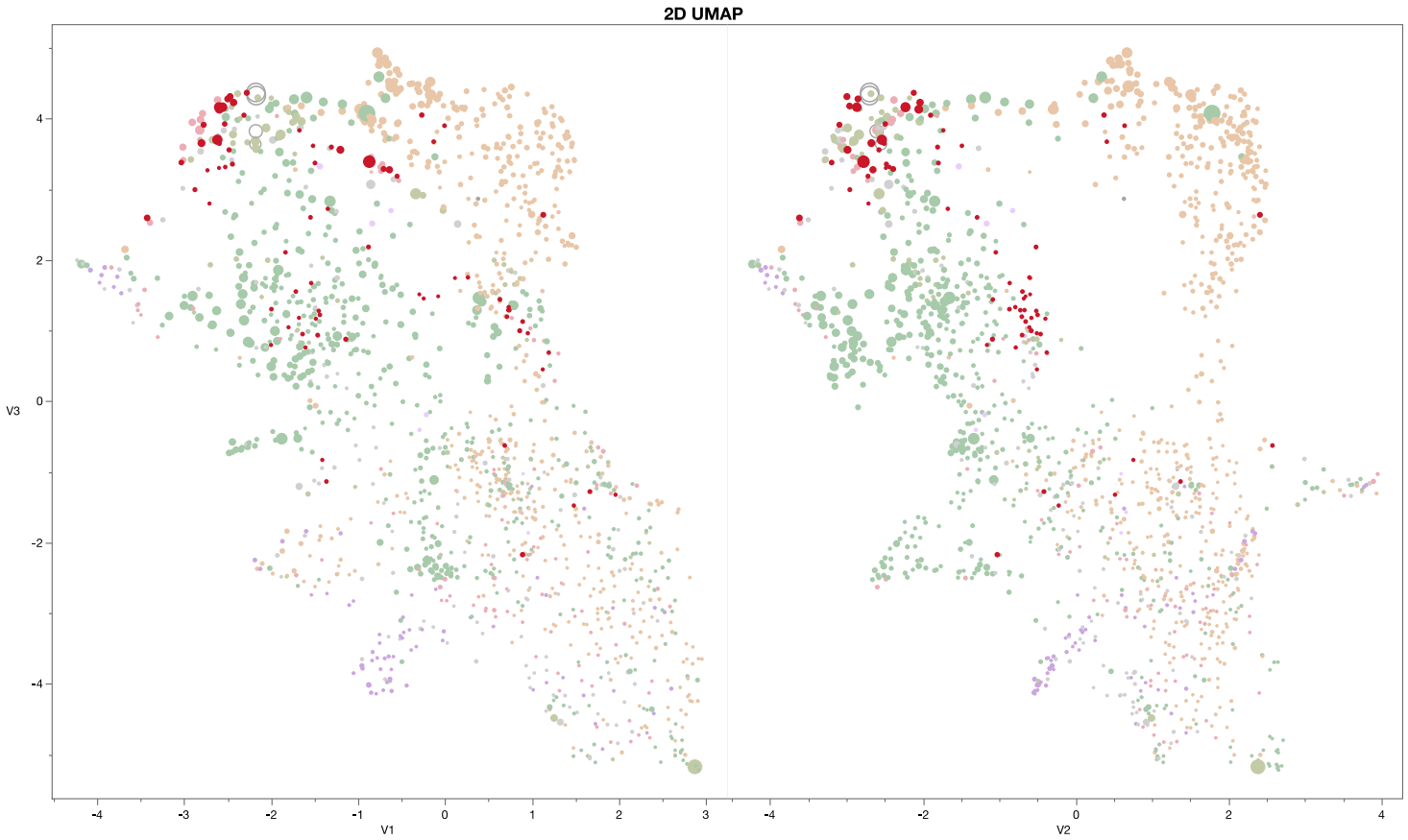


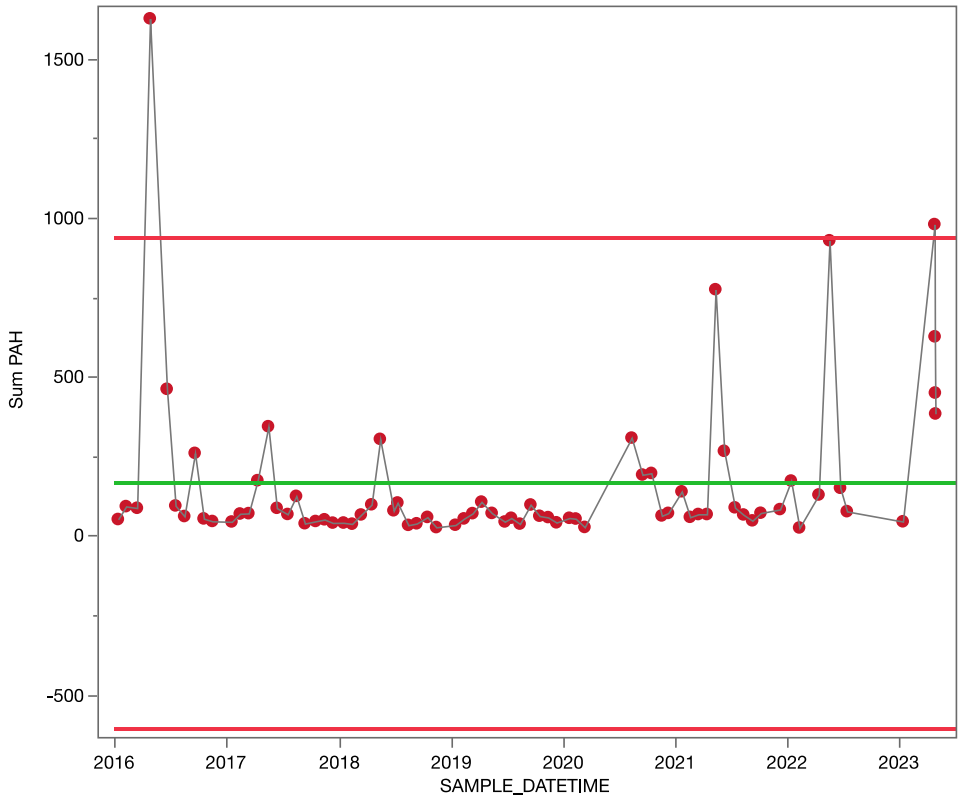
UMAP – Scrolling through River Systems

- Ability to scroll through different river systems helps to identify what the chemical patten of data might be telling you



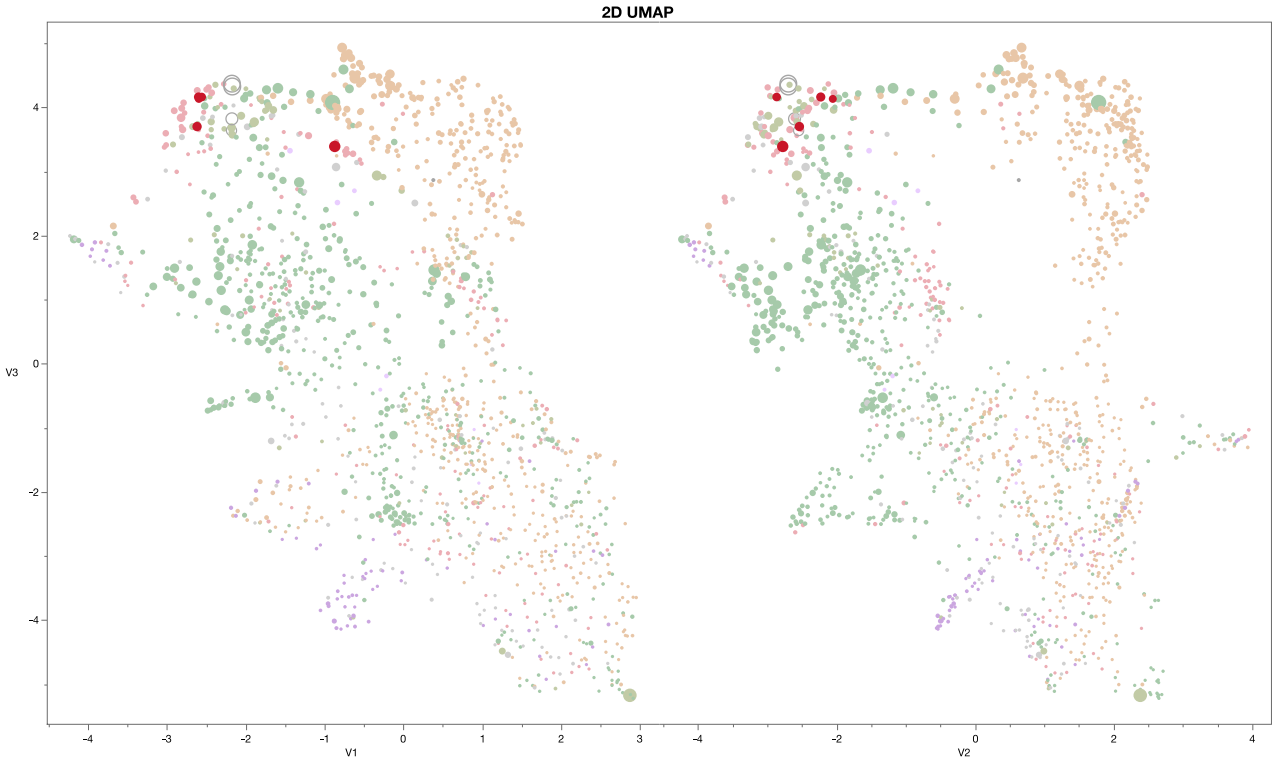
Combining Monitoring with UMAP

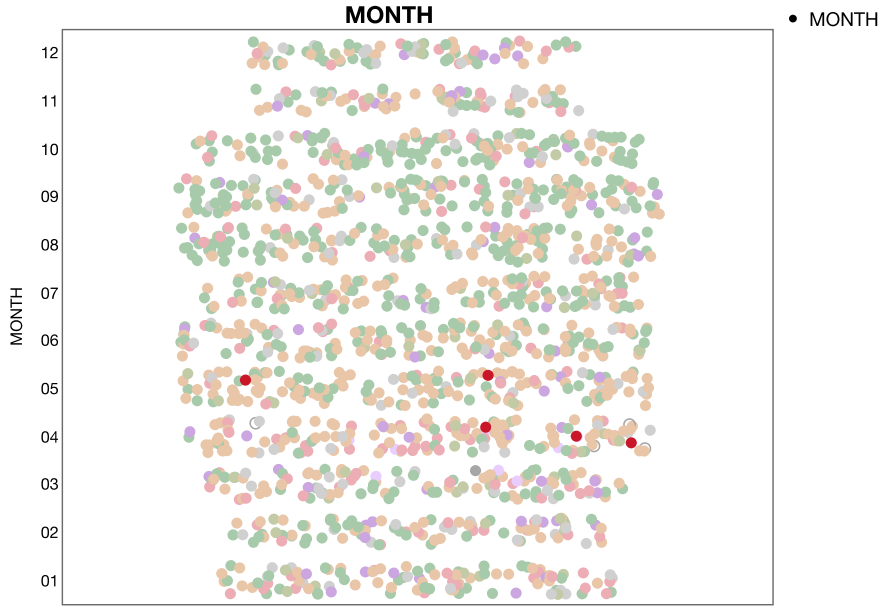


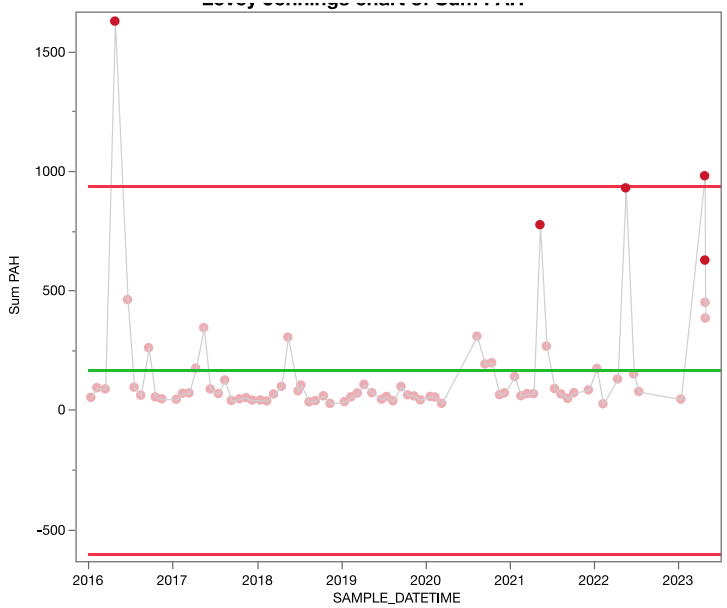


- A sampling location provides monitoring data over time. Examine trends with concentrations and patterns.
  - High concentrations seem to be annual events

Combining Monitoring with UMAP

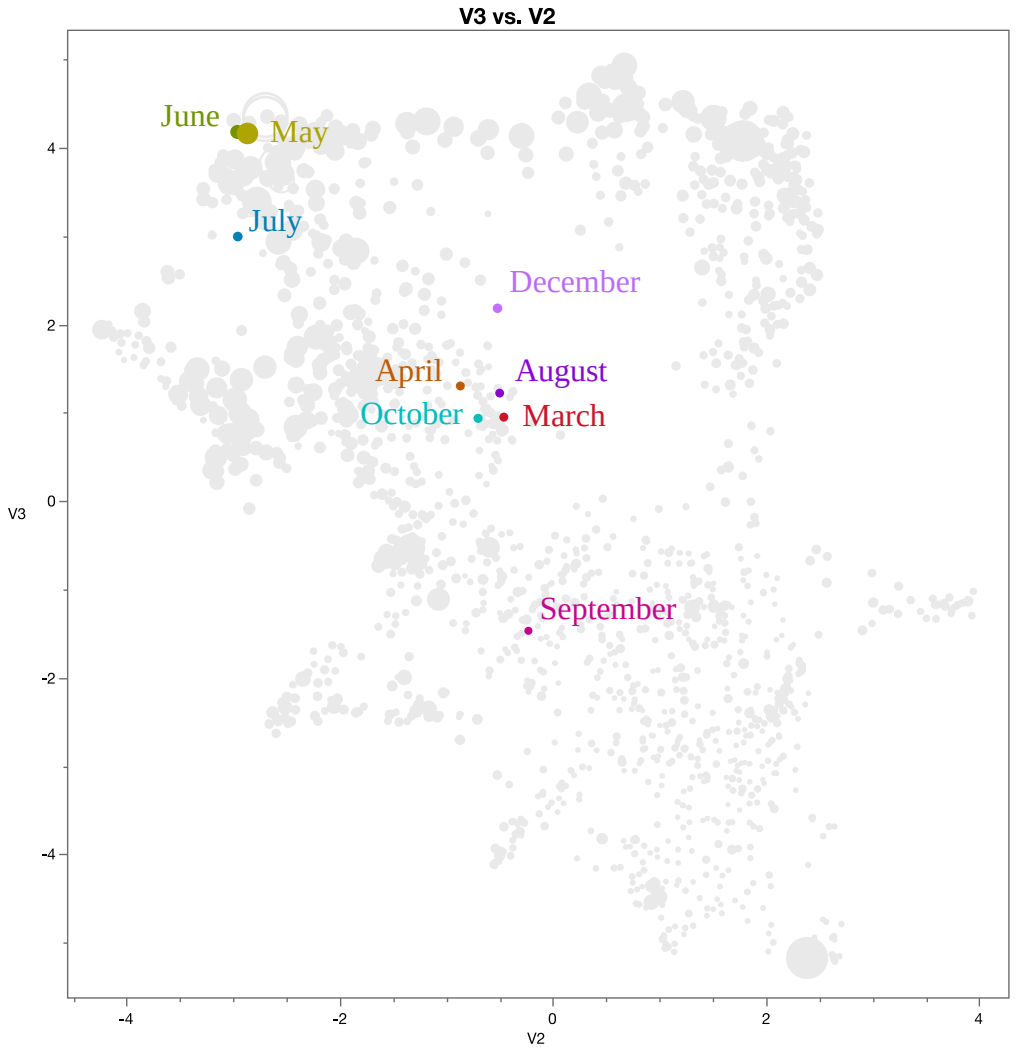


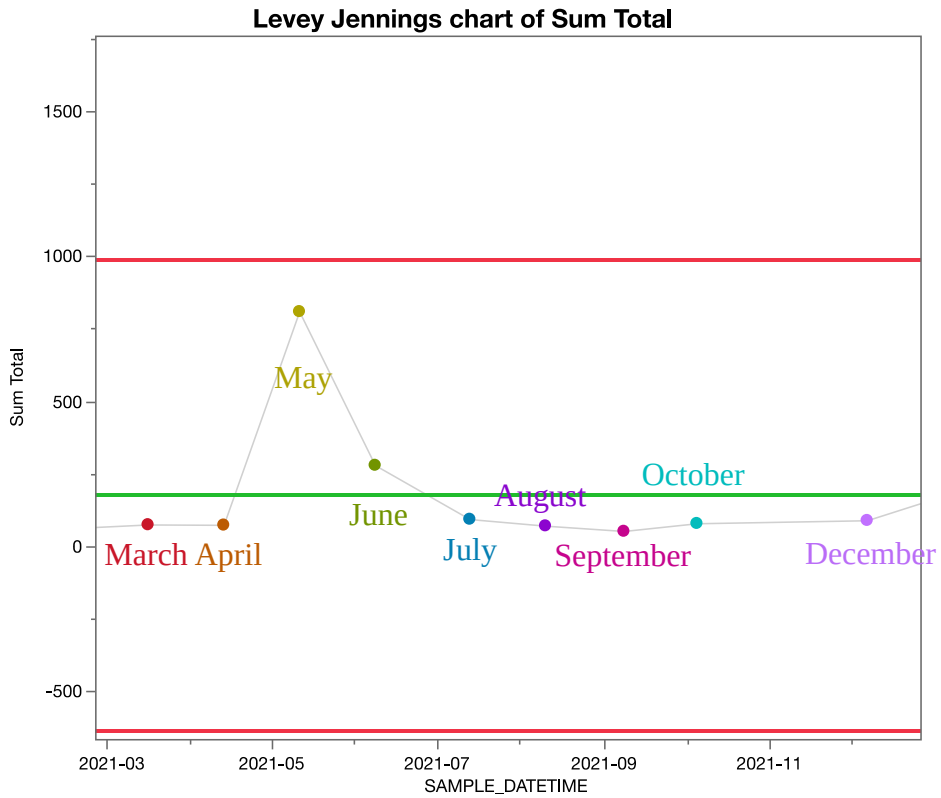




- Spring melt (freshet) causes pulse of PAHs
  - Natural erosion?
  - Washing of source into river with melt?

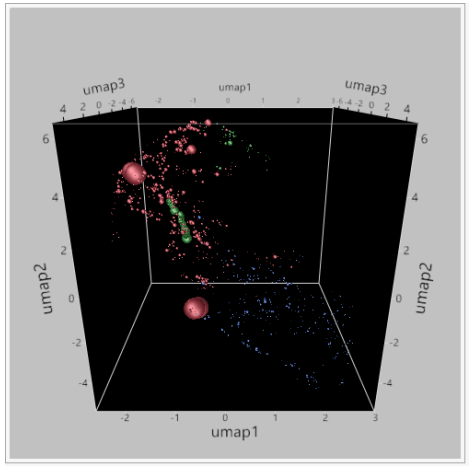
Monthly Patterns Emerge





Concluding Remarks

- Its okay to play with your data, just need to have a playroom that allows it
- Unlock learning about your data with interactive data exploration
- Seeing is believing
  - For you as you are exploring what your data is telling you
  - For client/public communication



Questions?

Contact Info:

Court Sandau

Chemistry Matters Inc.

Emails: csandau@chemistry-matters.com

court@statvis.com

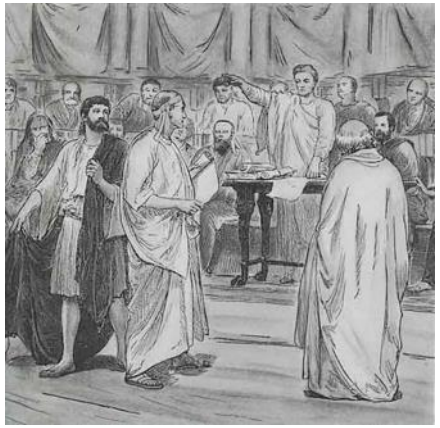
URLs: chemistry-matters.com

statvis.com

Follow us on: LinkedIn

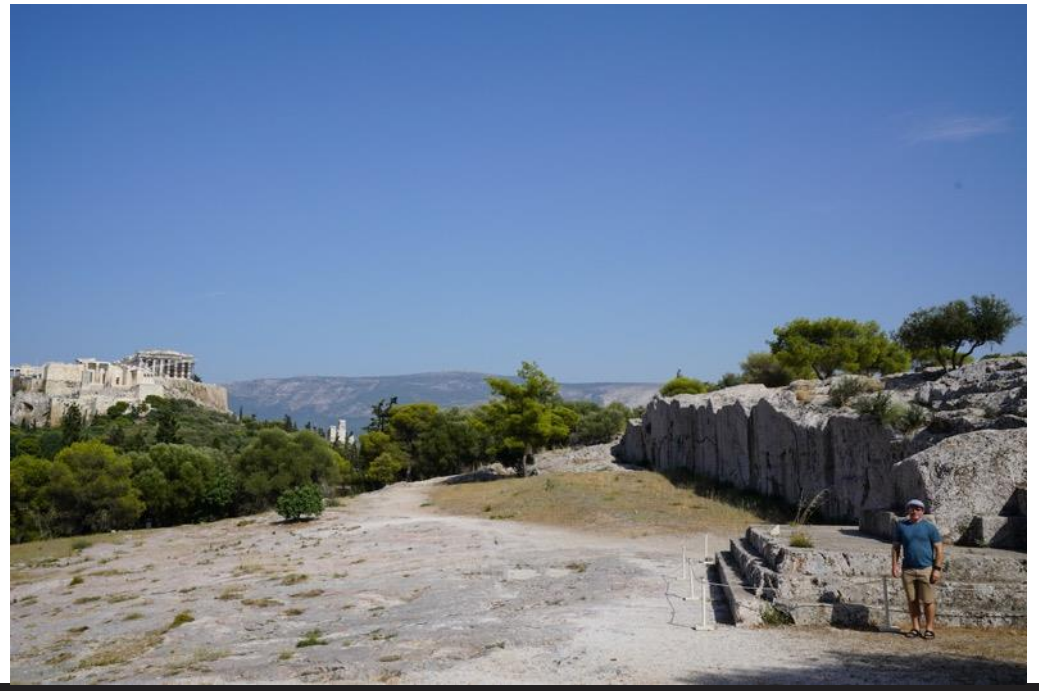
Forensics

- Ancient roman times
- Forensics derived from latin "*forens*":
  - belonging to the forum, public



http://www.crystalinks.com/romelaw.html

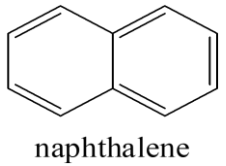
- **Environmental forensics** – *determining identity, source, or timing of contaminant release for the purpose of public communication* and/*or litigation.*

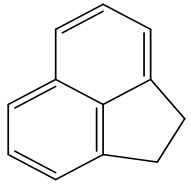


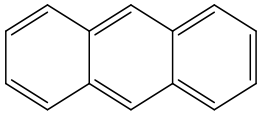
US EPA Priority Pollutants (PAH)

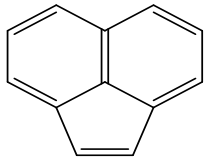
- In 1976, the US Environmental Protection Agency (USEPA) selected 16 PAHs as priority organic pollutants

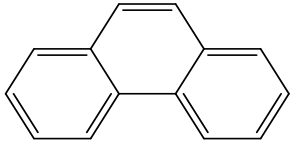
2-ring

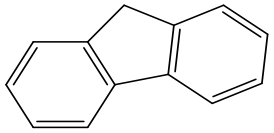

naphthalene


acenaphthene


anthracene


acenaphthylene

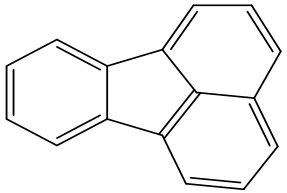

phenanthrene

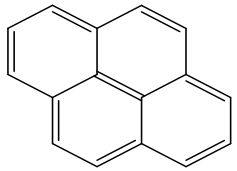

fluorene

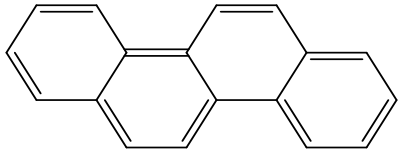
LMW

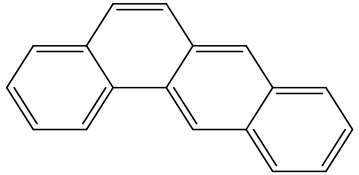
3-ring

HMW

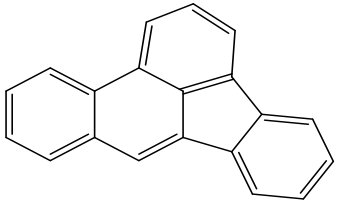

fluoranthene

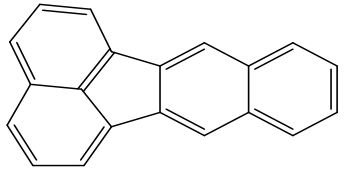

pyrene

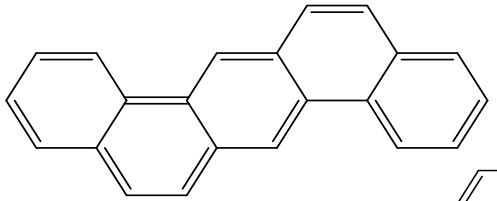

chrysene

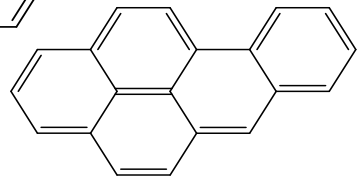

benzo[a]anthracene

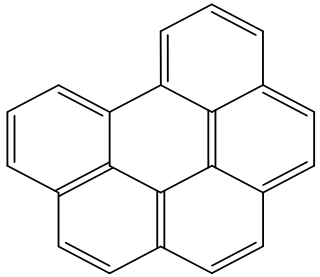
4-ring

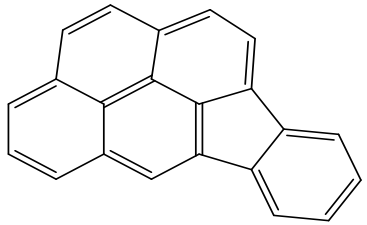

benzo[b]fluoranthene


benzo[k]fluoranthene


dibenz[a,h]anthracene


benzo[a]pyrene


benzo[ghi]perylene

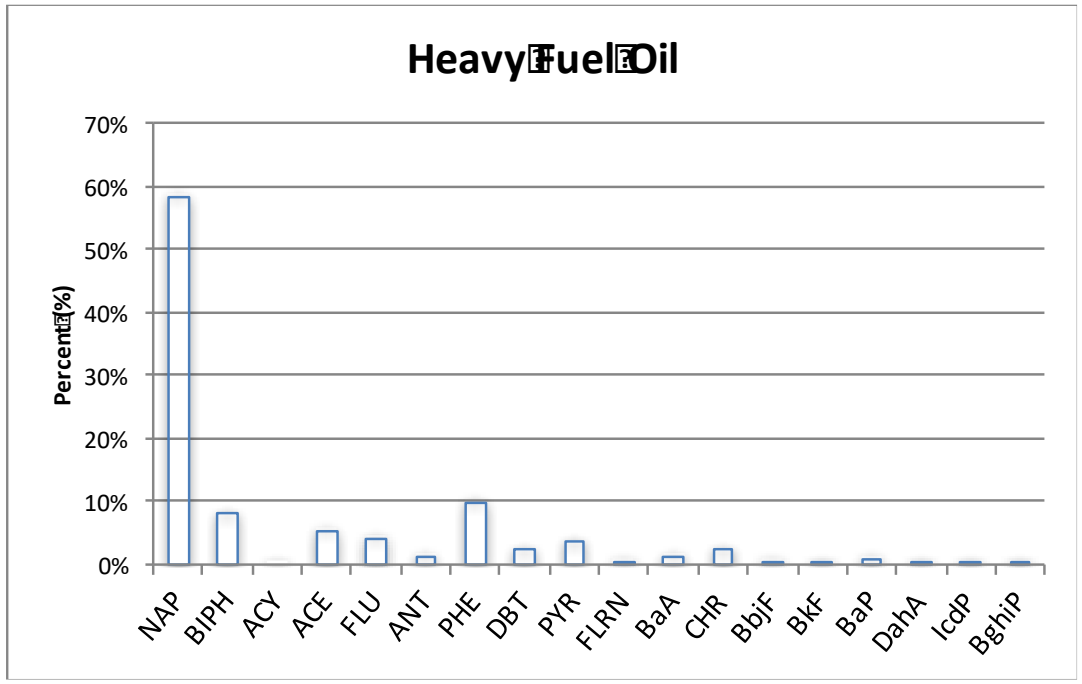

indeno[1,2,3-cd]pyrene

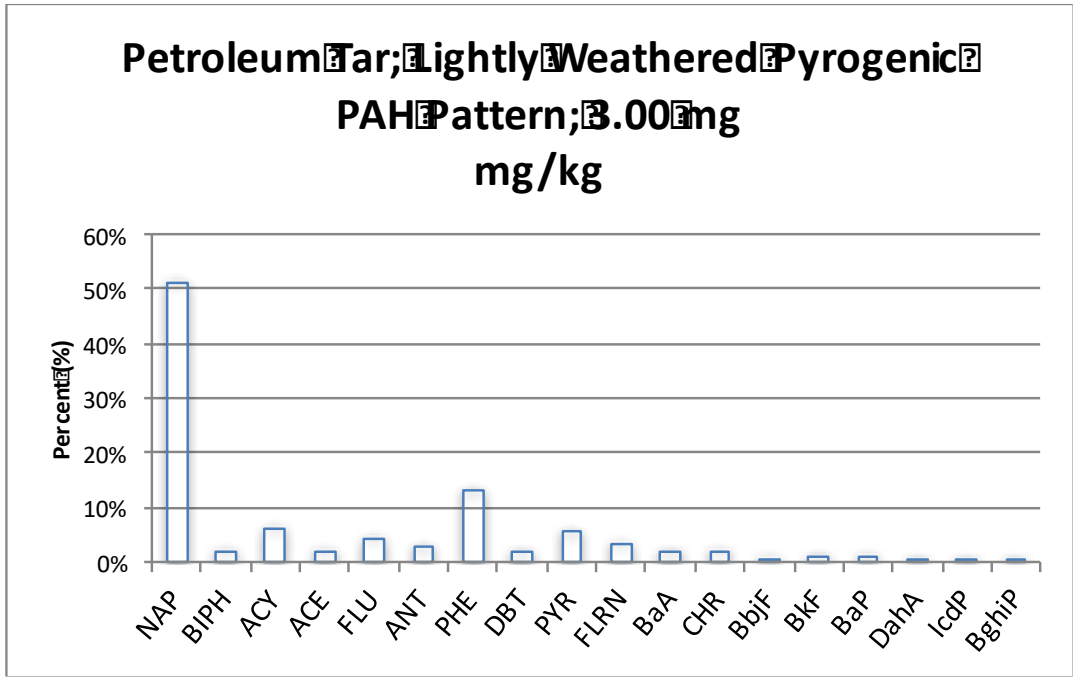
5-ring

6-ring

Keith and Telliard 1979

Forensic Fingerprints



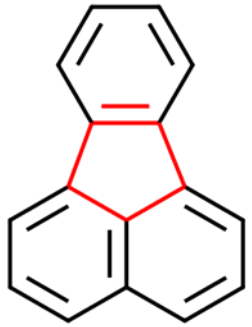


Samples from the same source will have same patterns.

With only 16 compounds, cannot distinguish fuel oil from gas plant tar. Not very forensic! But, need to use the data we have.

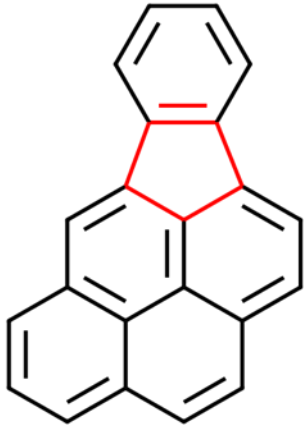
Reliable Diagnostic Ratios

High Temperature Formation



Fluoranthene

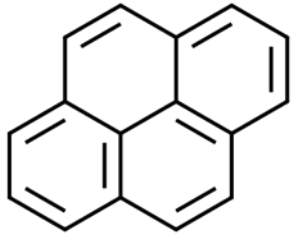
291 kJ/mol



Indeno[123,cd]pyrene

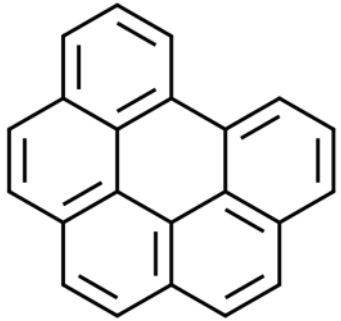
430 kJ/mol

Low Temperature Formation



Pyrene

225 kJ/mol



Benzo[ghi]perylene

300 kJ/mol

- Uses fundamental chemistry properties (heats of formation) which drives their formation ($\Delta_f H^\circ_m$)
- Uses 4 of the 16 priority PAHs
- Uses similar size structures for ratio (behave the same in environment)

Double Ratio Plots to Distinguish Source