Wildfire Smoke and Air Quality: How Machine Learning Can Guide Forest Management

"Tackling Climate Change with Machine Learning" Workshop at NeurIPS 2020

Lorenzo Tomaselli, *Ph.D. Student @ Statistics and Data Science – CMU*

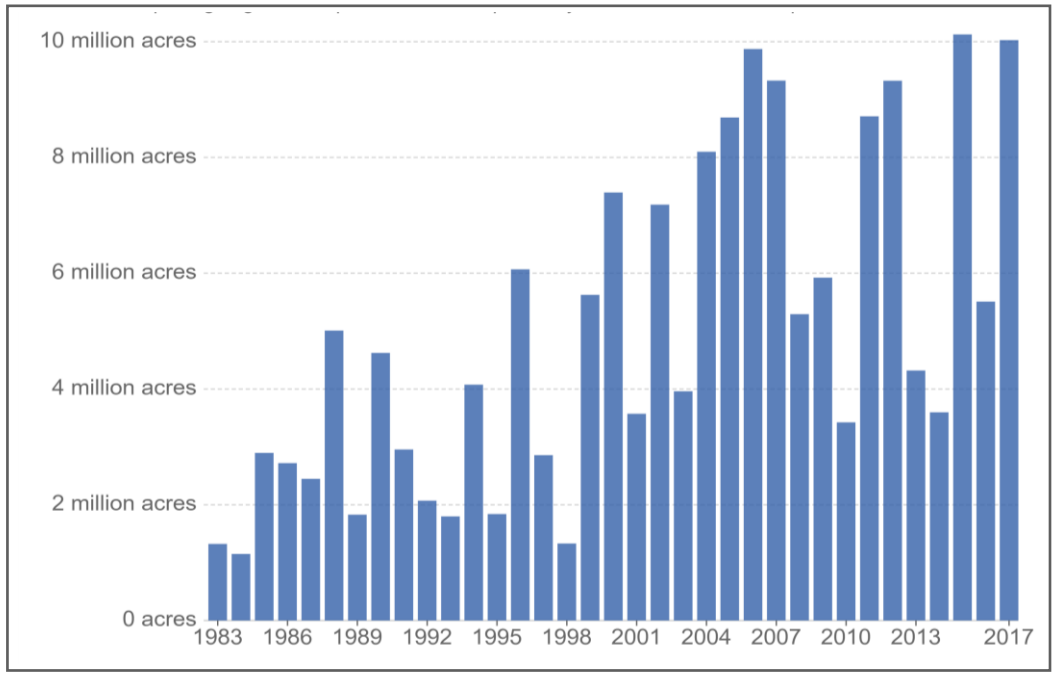
Coty Jen, *Assistant Professor @ Chemical Engineering – CMU*

Ann B. Lee, *Professor @ Statistics and Data Science – CMU*

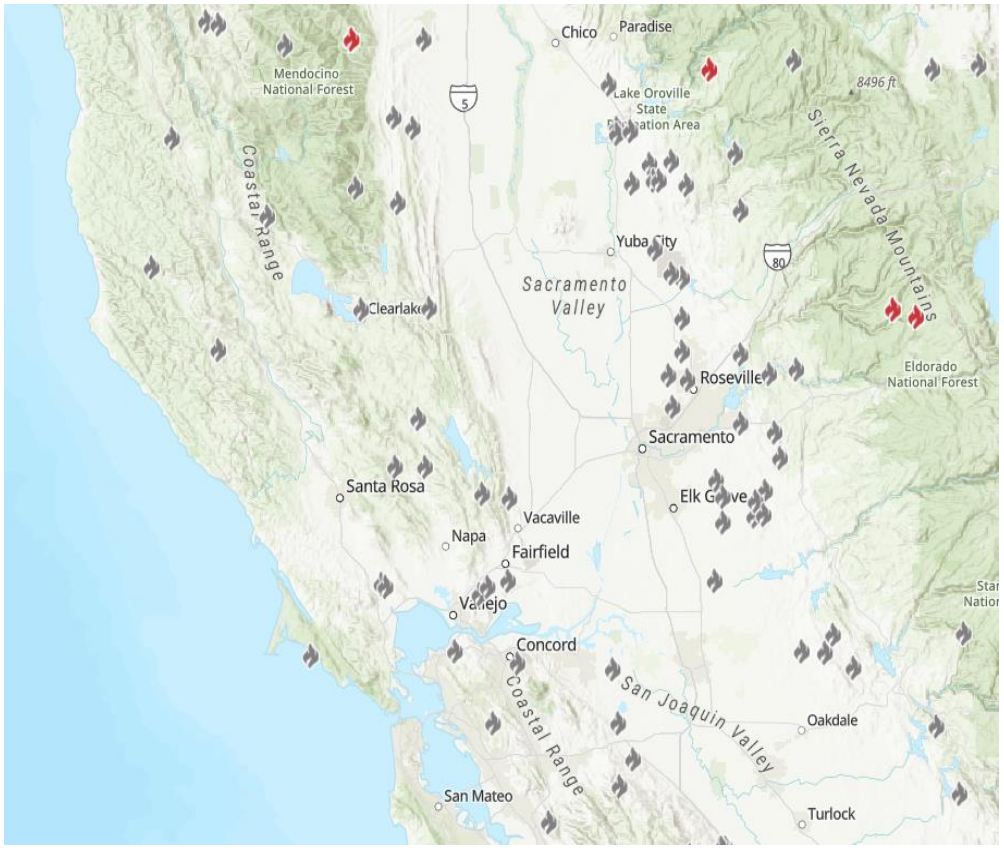
Wildfires are Getting More Frequent Over the Years

Larger and more frequent wildfires due to climate change and other factors (fire suppression, droughts etc.)

According to CAL FIRE, in 2020 there have been 9,177 events with 4,194,148 acres burned. [https://www.fire.ca.gov/incidents/2020/]



https://www.nifc.gov/fireInfo/fireInfo_stats_totalFires.html



https://modis.gsfc.nasa.gov/

Prescribed Burns to Control Wildfires

- Prescribed burns most effective way to manage forests in the US
- However, burning large amounts of built-up fuel may negatively impact air quality
- **All smoke is not the same**. How can we tell the difference between smoke?

https://www.swgafarmcredit.com/prescribed-burning/

Prescribed Burns to Control Wildfires

- Prescribed burns most effective way to manage forests in the US
- However, burning large amounts of built-up fuel may negatively impact air quality
- **All smoke is not the same**. How can we tell the difference between smoke?

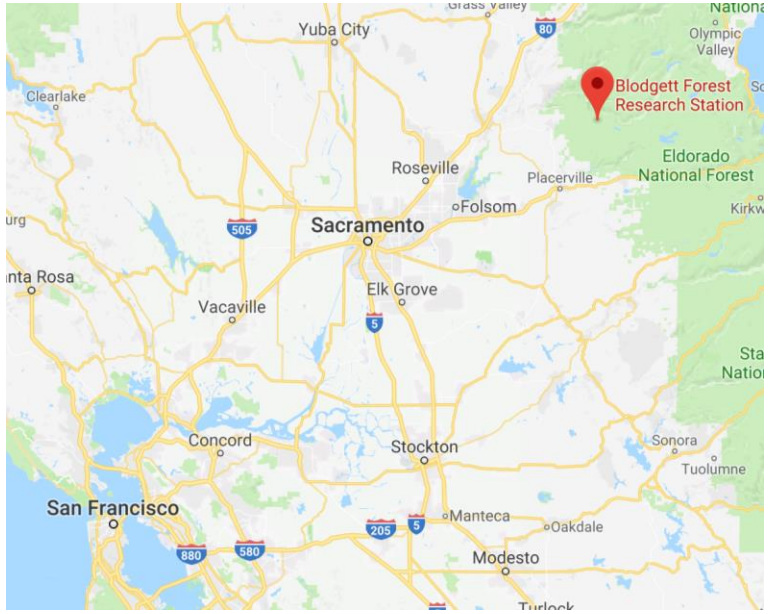
Our goal:
To help forest managers assess impact of prescribed burns on air quality.

https://www.swgafarmcredit.com/prescribed-burning/

Data Collection of Ground and Air Smoke Samples

Our CMU Center for Atmospheric Particles Studies (CAPS) team collected 54 smoke samples at Blodgett Forest Research Station (BFRS):

- Ground: 33 smoke samples
- Air: 21 smoke samples collected at ~100 m altitude

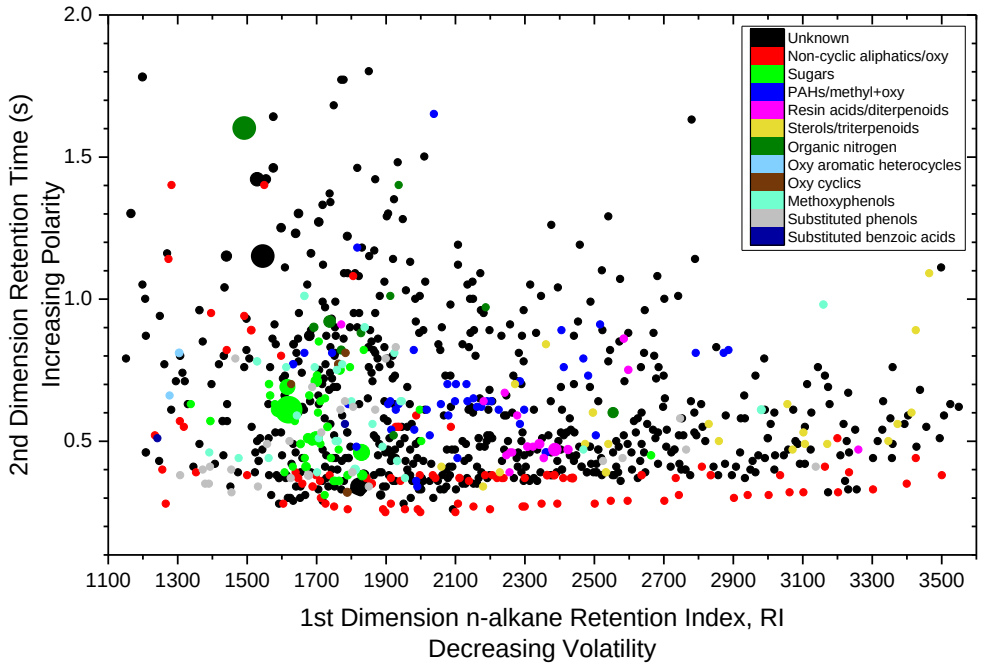


Courtesy of Coty Jen

Chemical Fingerprinting to Identify Chemical Compounds in Smoke Samples

Chemical fingerprint analysis to obtain:

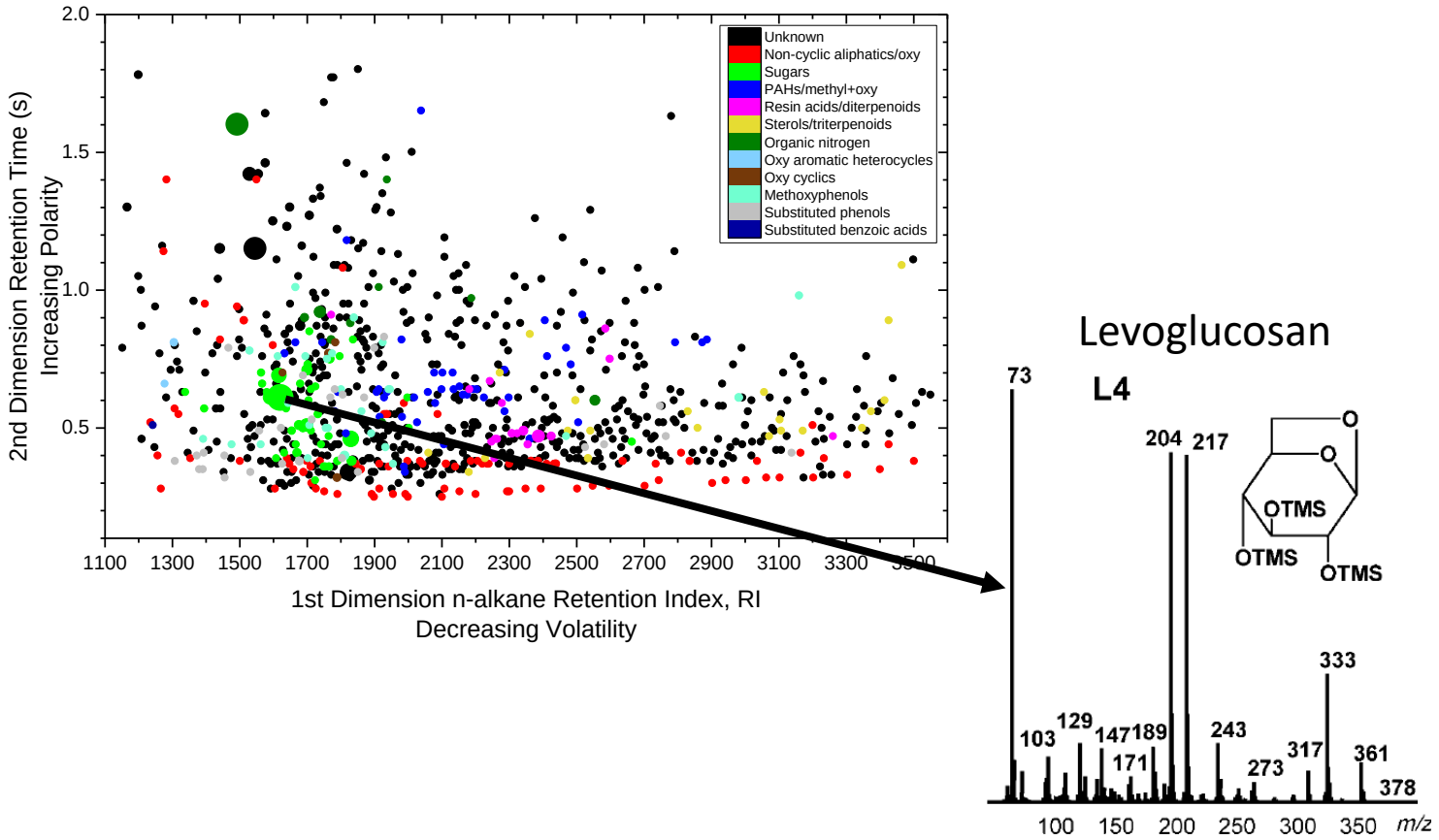
- **gas chromatogram** (GC) of smoke sample
- **mass spectrum** (MS) of each compound.



Chemical Fingerprinting to Identify Chemical Compounds in Smoke Samples

Chemical fingerprint analysis to obtain:

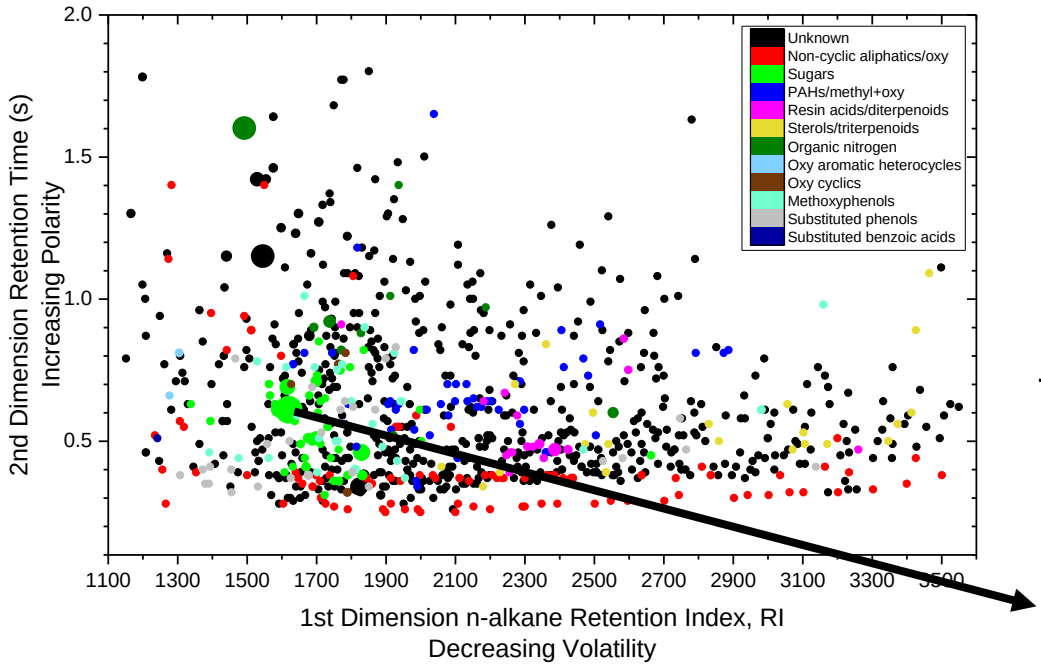
- **gas chromatogram** (GC) of smoke sample
- **mass spectrum** (MS) of each compound.



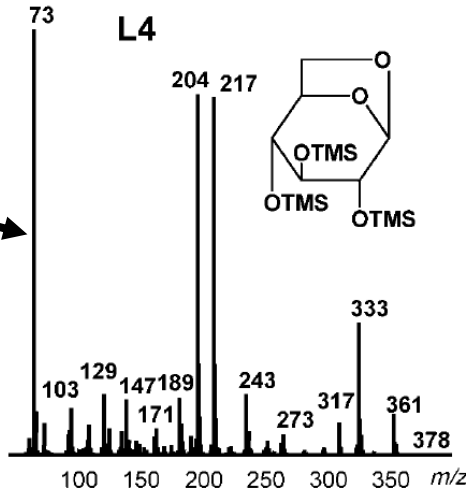
Chemical Fingerprinting to Identify Chemical Compounds in Smoke Samples

Chemical fingerprint analysis to obtain:

- **gas chromatogram** (GC) of smoke sample
- **mass spectrum** (MS) of each compound.



Levoglucosan



Each smoke sample is a **distribution over compounds** (weighted set):

$$S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$$

- $\mathbf{x}_i \in \mathbb{R}^p$ is mass spectrum ($p \sim 500, m \sim 1000$);
- the weights $w_{\mathbf{x}_i}$ represent amount of each compound

Comparing Distributions of High-dimensional Data

Each smoke sample is a distribution over high-dimensional data (MS of compounds), $\mathbf{x}_i \in \mathbb{R}^{500}$:

$$S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$$

Statistical challenge:

- How to compare such distributions
- How to relate differences to chemical compounds

Comparing Distributions of High-dimensional Data

Each smoke sample is a distribution over high-dimensional data (MS of compounds), $\mathbf{x}_i \in \mathbb{R}^{500}$:

$$S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \dots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$$

Statistical challenge:

- How to compare such distributions
- How to relate differences to chemical compounds

In brief:

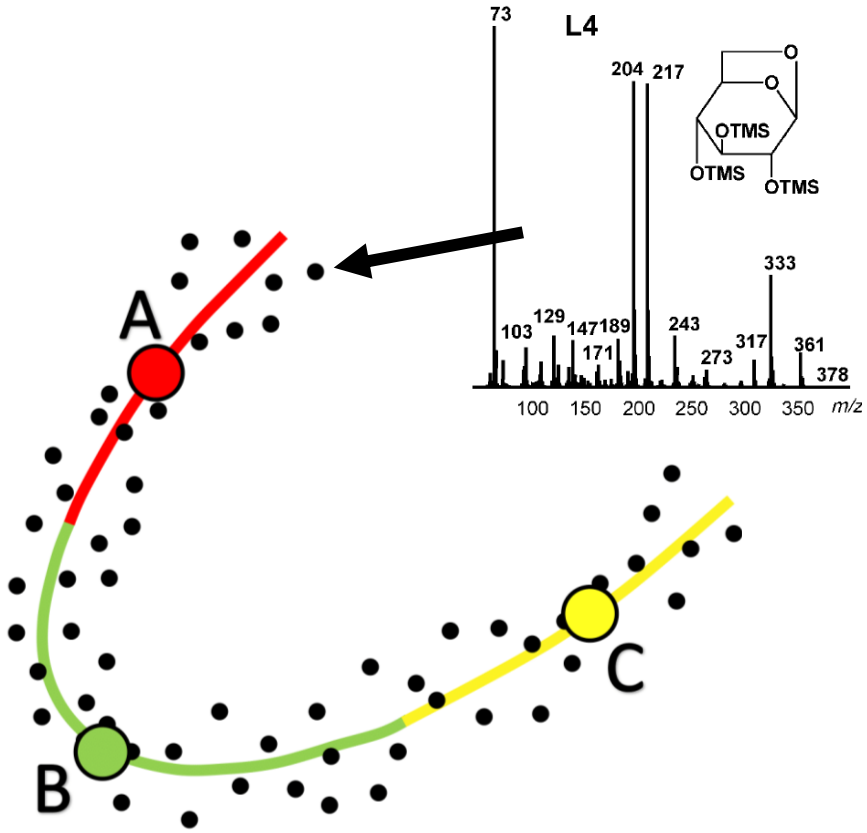
Our goal is to define a scientifically **interpretable metric** between smoke samples

Construct Codebook of Compound "Words" that Reflects Data Structure

Smoke sample as distribution over high-dimensional data:

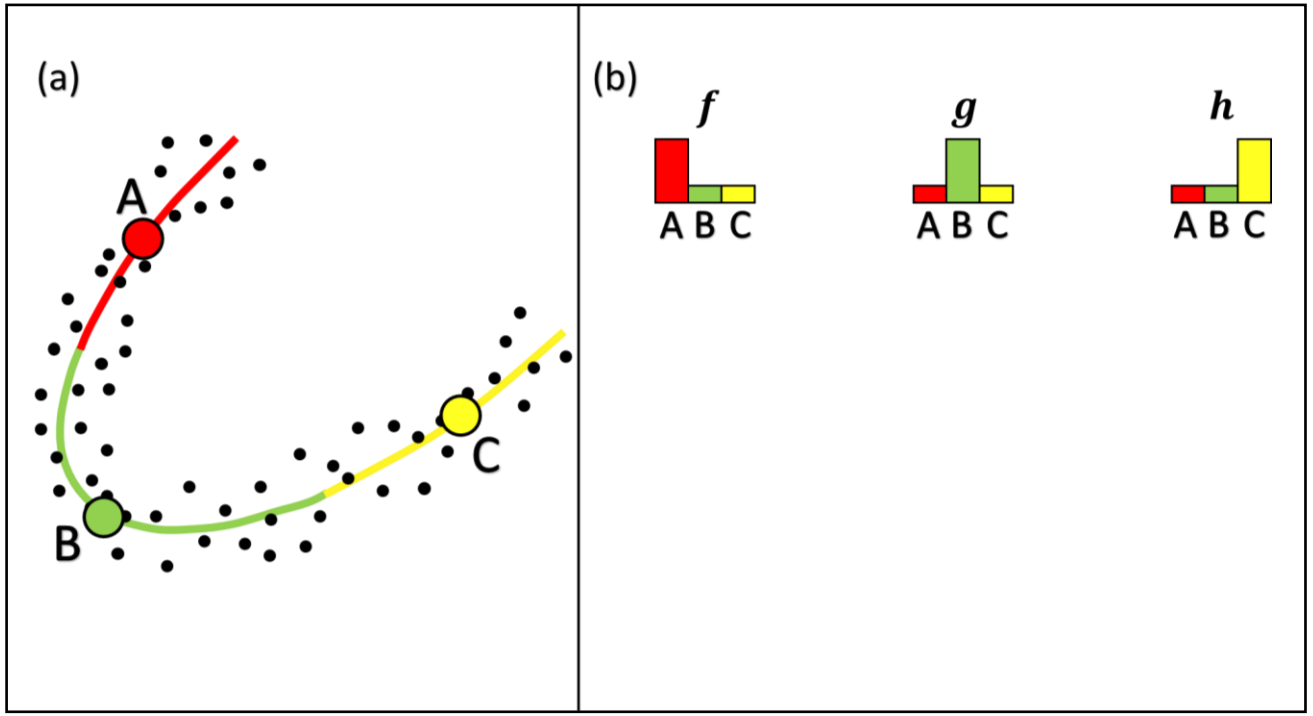
$$S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$$

- Data are high-dimensional but have **sparse structure**
- Construct a **codebook of K compound "words"** by spectral clustering and connectivity analysis



A Metric between Smoke Samples Should Account for Data Geometry

Smoke samples can be represented as **histograms** over the K compound “words”.

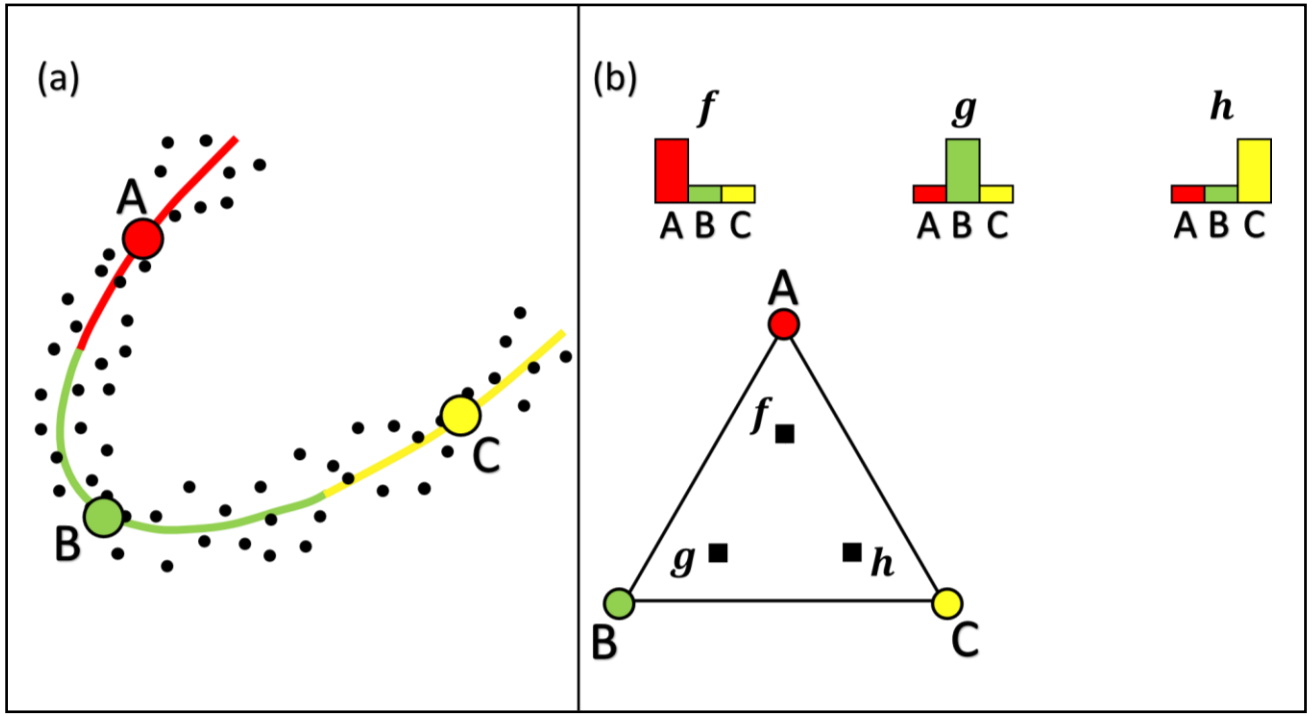


A Metric between Smoke Samples Should Account for Data Geometry

Smoke samples can be represented as **histograms** over the K compound “words”.

Histograms as **points in the simplex of K vertices**

However, this arrangement does not reflect data geometry/interbin relationship

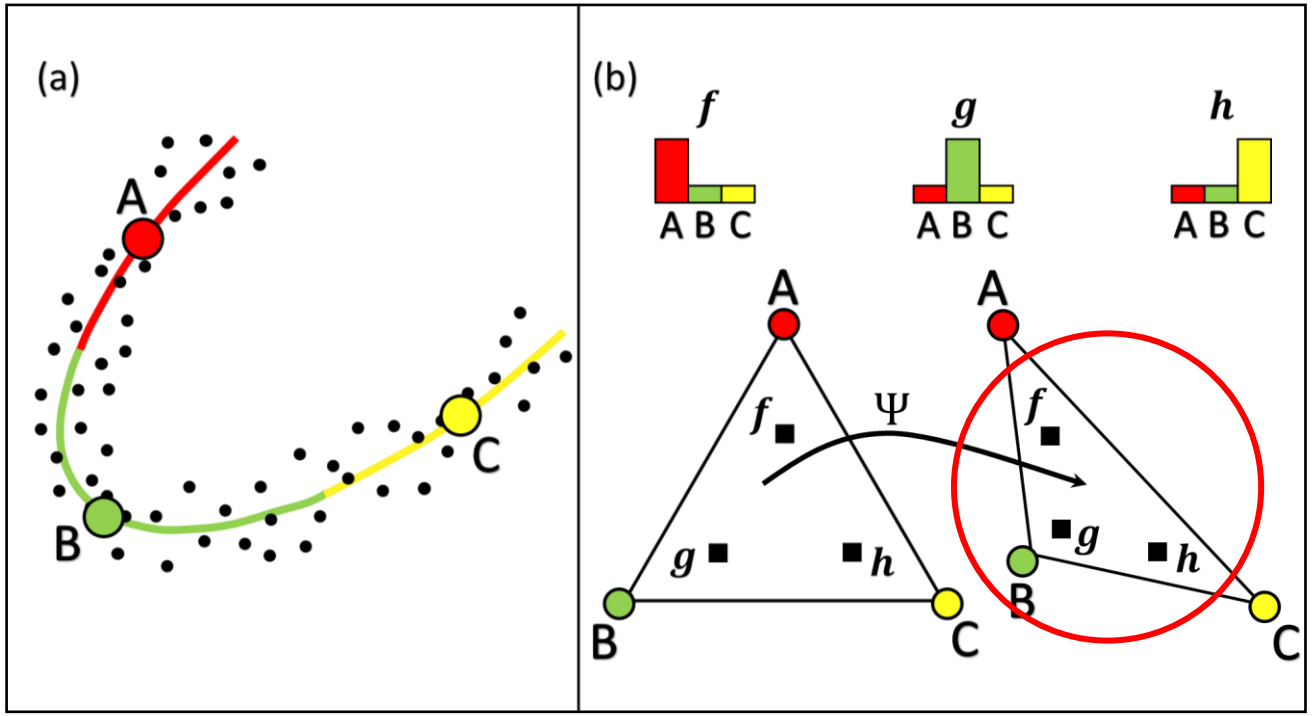


Our Proposed Metric Is Interpretable and Reflects Data Geometry

Using **diffusion map**:

$$\mathbf{x} \mapsto \Psi(\mathbf{x}) = \left(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_D(\mathbf{x})\right) \in \mathbb{R}^D$$

In this embedding (red circle), Euclidean distances reflect connectivity.



Our Proposed Metric Is Interpretable and Reflects Data Geometry

Using **diffusion map**:

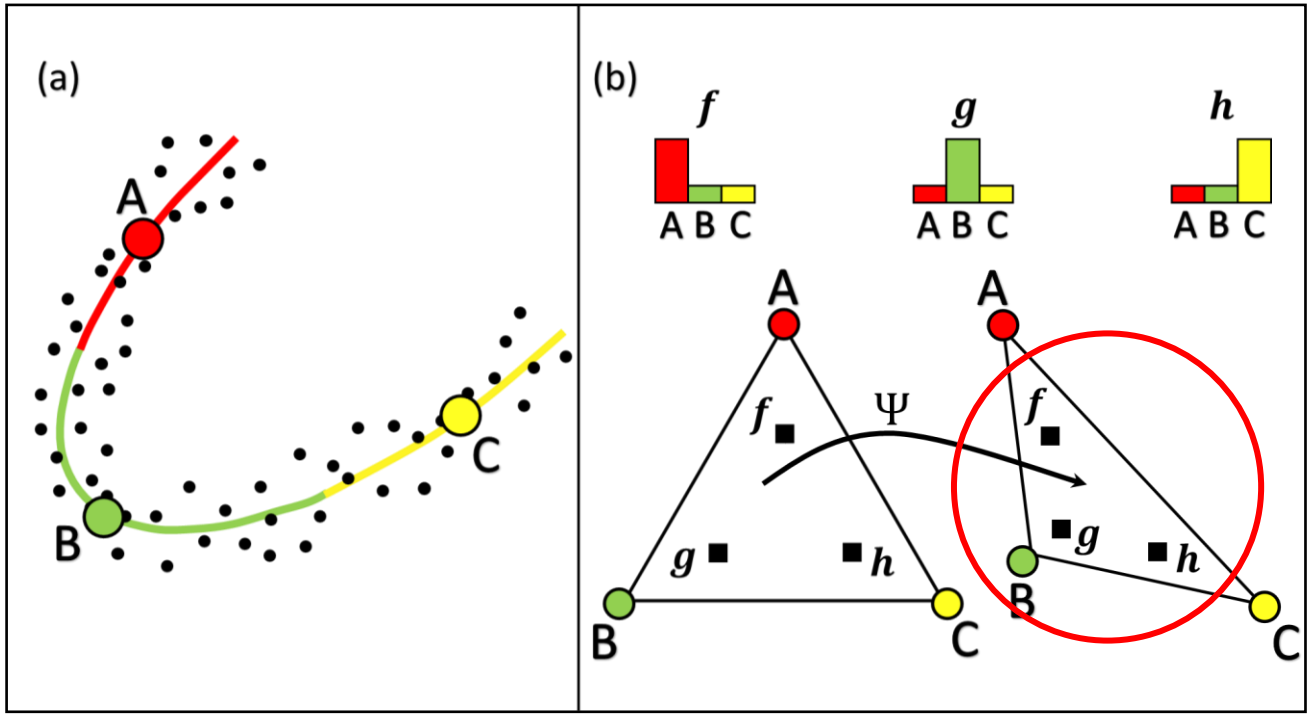
$$\mathbf{x} \mapsto \Psi(\mathbf{x}) = \big(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \dots, \psi_D(\mathbf{x})\big) \in \mathbb{R}^D$$

In this embedding (red circle), Euclidean distances reflect connectivity.

This induces a distance between smoke samples, or their histogram representations:

$$\mathcal{D}(S_i, S_j) = \left\| \sum_{i=1}^{K} (f_i - g_i) \cdot c_i \right\|$$

with $\{c_i\}_{i=1}^{K}$ being the K words (e.g. A, B, C).



Our Proposed Metric Is Interpretable and Reflects Data Geometry

Using **diffusion map**:

$$\mathbf{x} \mapsto \Psi(\mathbf{x}) = \big(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \dots, \psi_D(\mathbf{x})\big) \in \mathbb{R}^D$$

In this embedding (red circle), Euclidean distances reflect connectivity.

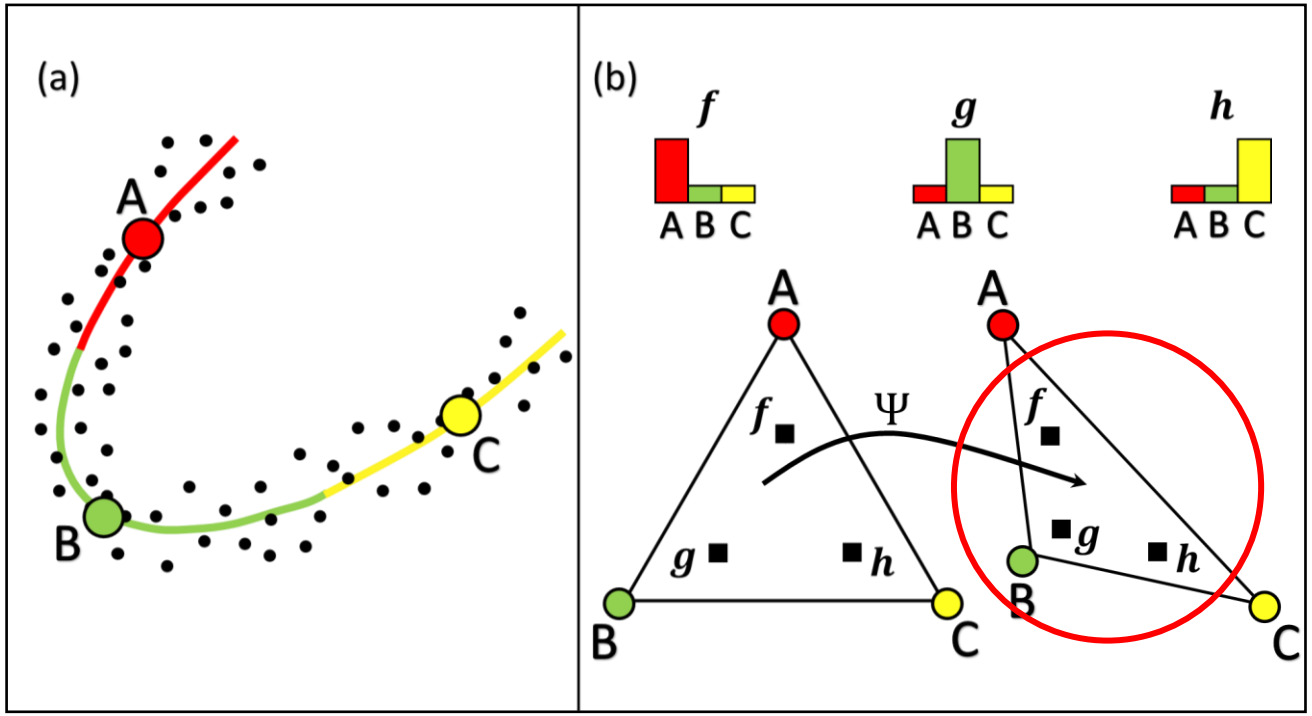
This induces a distance between smoke samples, or their histogram representations:

$$\mathcal{D}(S_i, S_j) = \left\| \sum_{i=1}^{K} (f_i - g_i) \cdot c_i \right\|$$

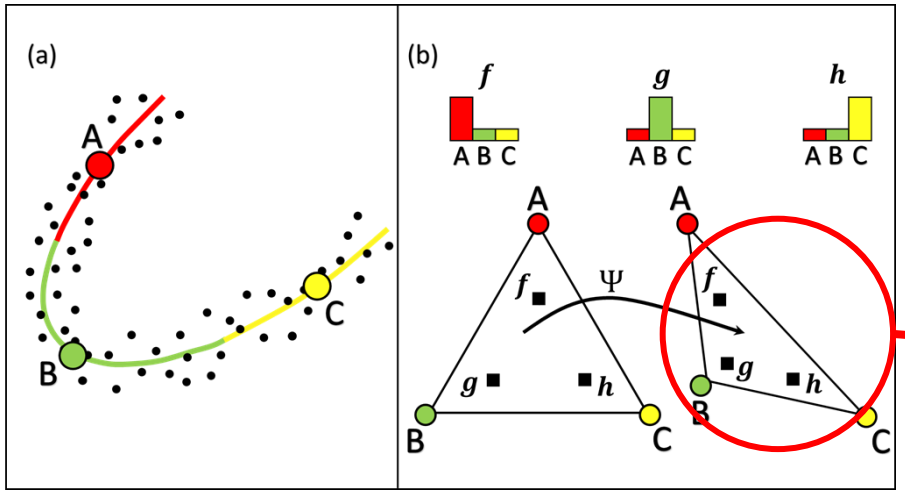
with $\{c_i\}_{i=1}^{K}$ being the K words (e.g. A, B, C).

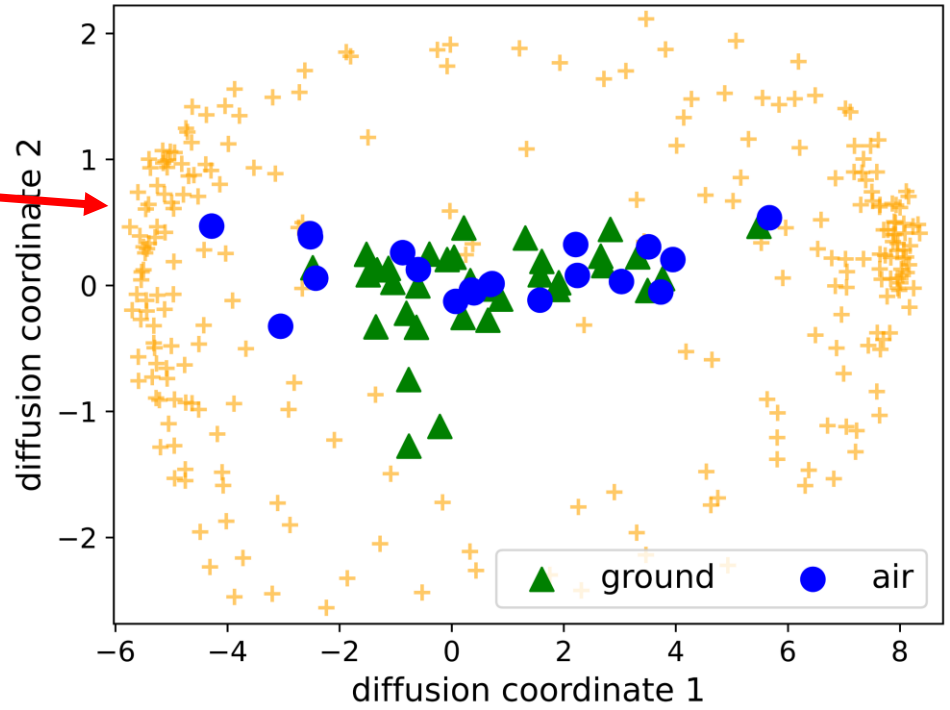
Our proposed metric:

- reflects data geometry
- is interpretable thanks to the codebook
- can serve as input to kernel ML algorithms

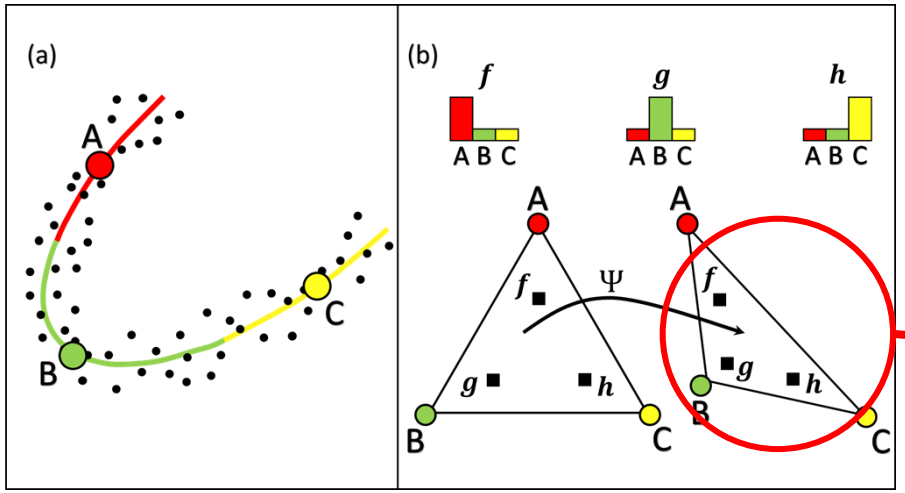


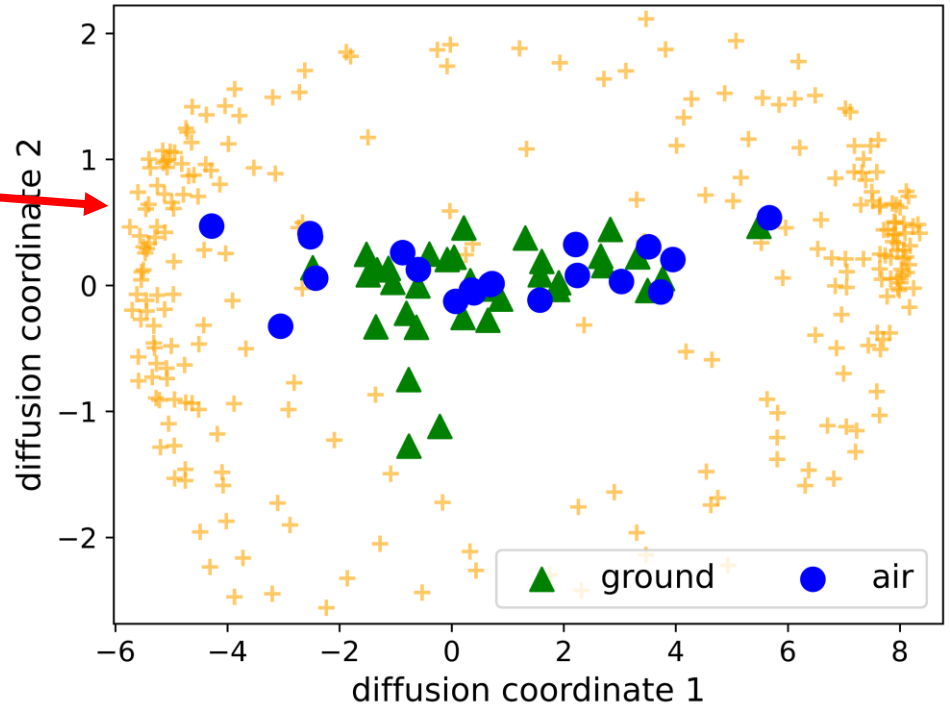
Our Method Differentiates Between Smoke Sample!





Interpretable Results Help Develop Forest Management Plans





- Can relate differences back to compound space for scientific interpretability
- Interpretable results can help forest managers design prescribed burns that minimize negative air quality impact