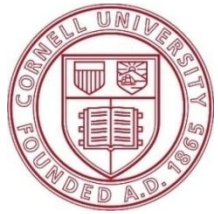


Constraint Reasoning and Kernel Clustering for Pattern Decomposition With Scaling

Ronan LeBras	Computer Science
Theodoros Damoulas	Computer Science
Ashish Sabharwal	Computer Science
Carla P. Gomes	Computer Science
John M. Gregoire	Materials Science / Physics
Bruce van Dover	Materials Science / Physics

Sept 15, 2011 **CP'11**

Motivation

Cornell Fuel Cell Institute

Mission: develop **new materials** for **fuel cells**.

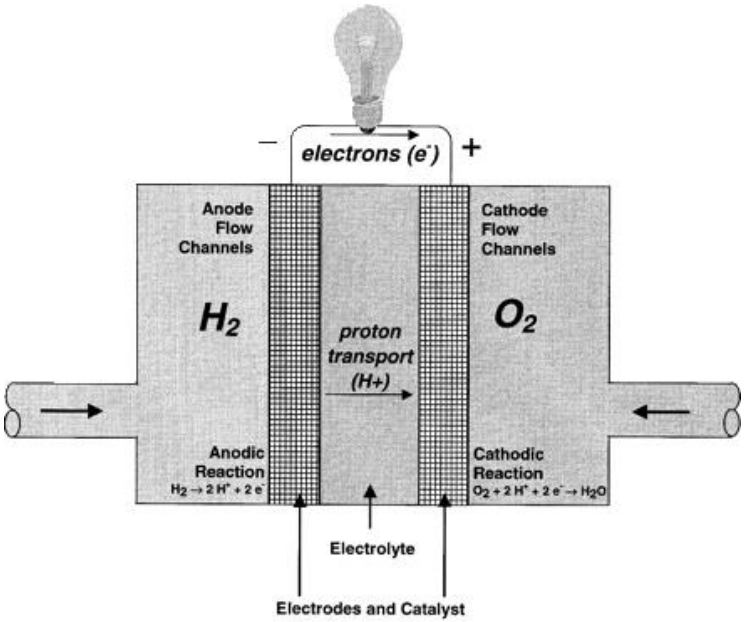


Figure 1. Fuel cell schematic.
Source: Annual Reveiws of Energy and the Environment. http://energy.annualreviews.org/cgi/content/full/24/1/281

*An **Electrocatalyst** must:*

1) *Be electronically **conducting***
2) ***Facilitate** both reactions*

***Platinum** is the best known metal to fulfill that role, but:*

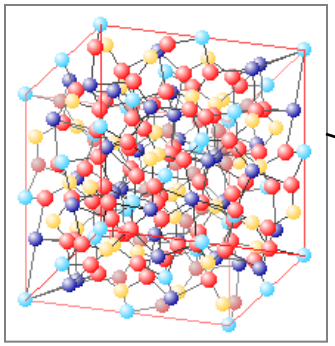
1) *The reaction rate is still considered slow (causing **energy loss**)*
2) *Platinum is fairly **costly**, **intolerant** to fuel **contaminants**, and has a **short lifetime**.*

*Goal: Find an **intermetallic compound** that is a better catalyst than Pt.*

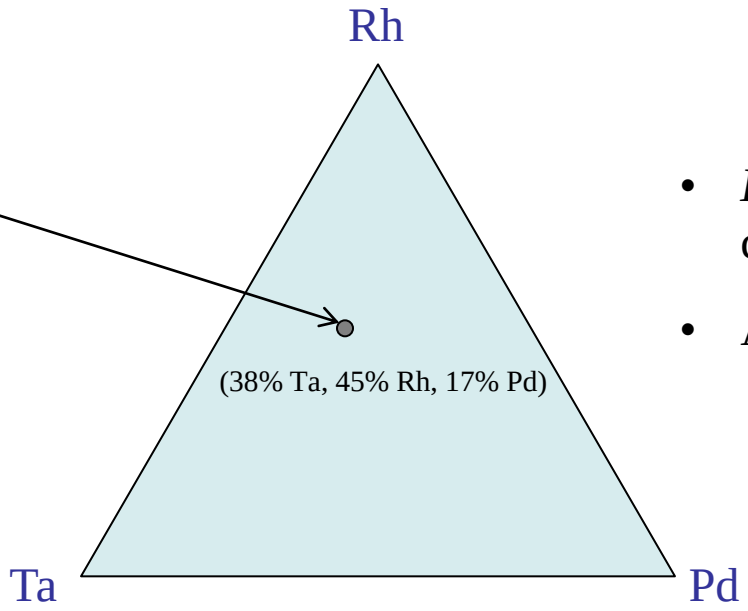
Motivation

Recipe for finding alternatives to Platinum

1) *In a vacuum chamber, place a silicon wafer.*
2) *Add three metals.*
3) *Mix until smooth, using three sputter guns.*
4) *Bake for 2 hours at 650ºC*



[Source: *Pyrotope*, Sebastien Merkel]

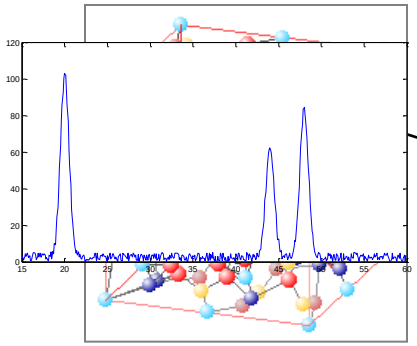


- *Deliberately* inhomogeneous composition on Si wafer
- Atoms are intimately mixed

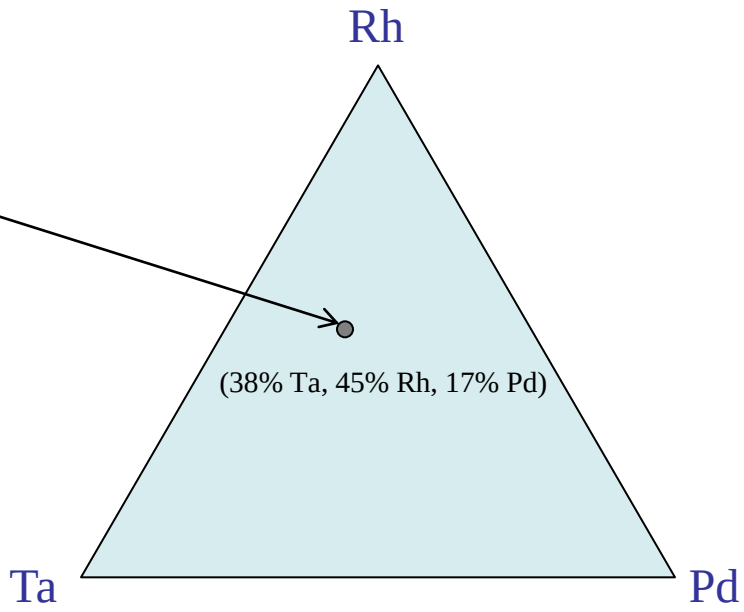
Motivation

Identifying crystal structure using **X-Ray Diffraction** at CHESS

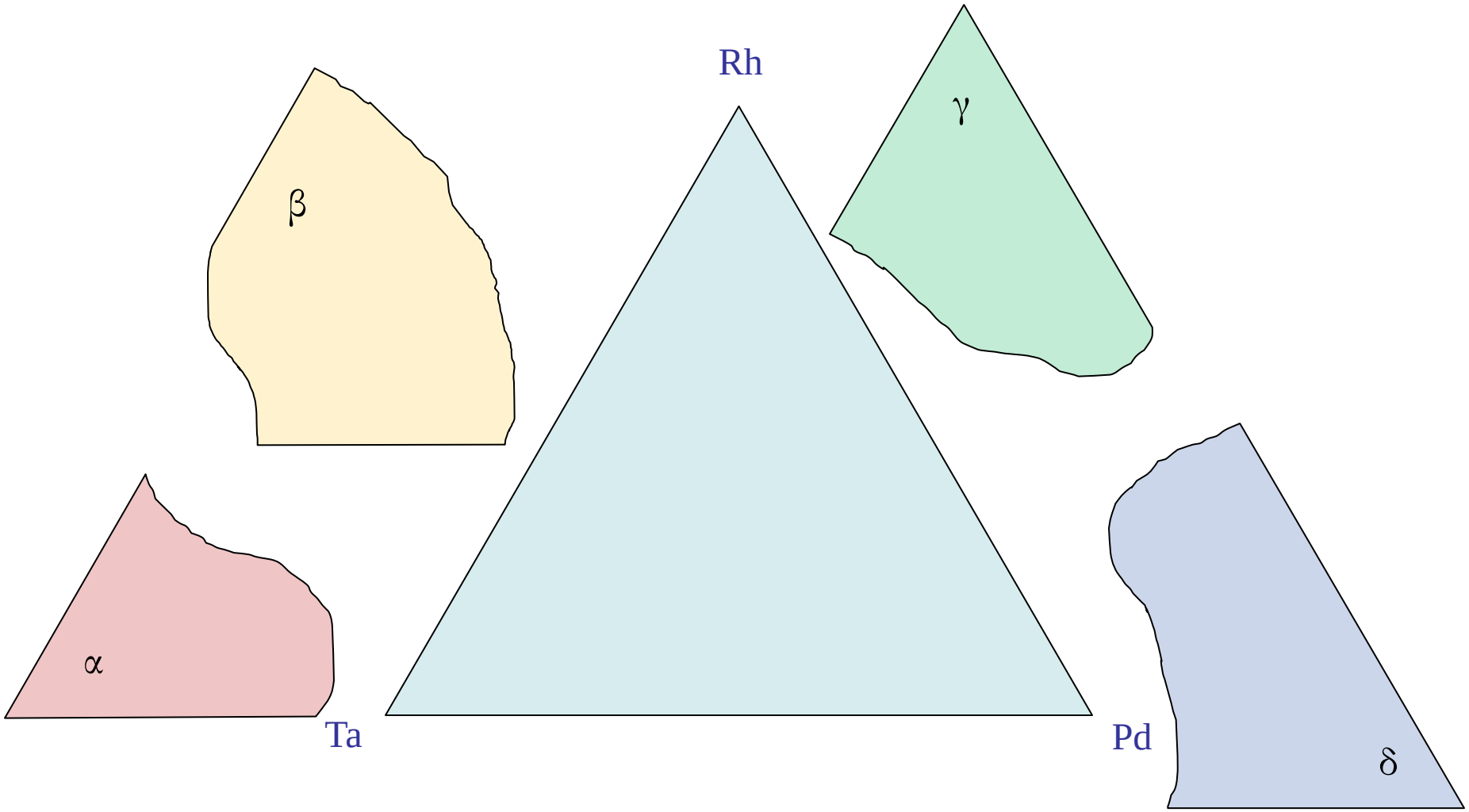
- XRD pattern **characterizes** the underlying **crystal** fairly well
- **Expensive** experimentations: Bruce van Dover's research team has access to the facility **one week every year**.



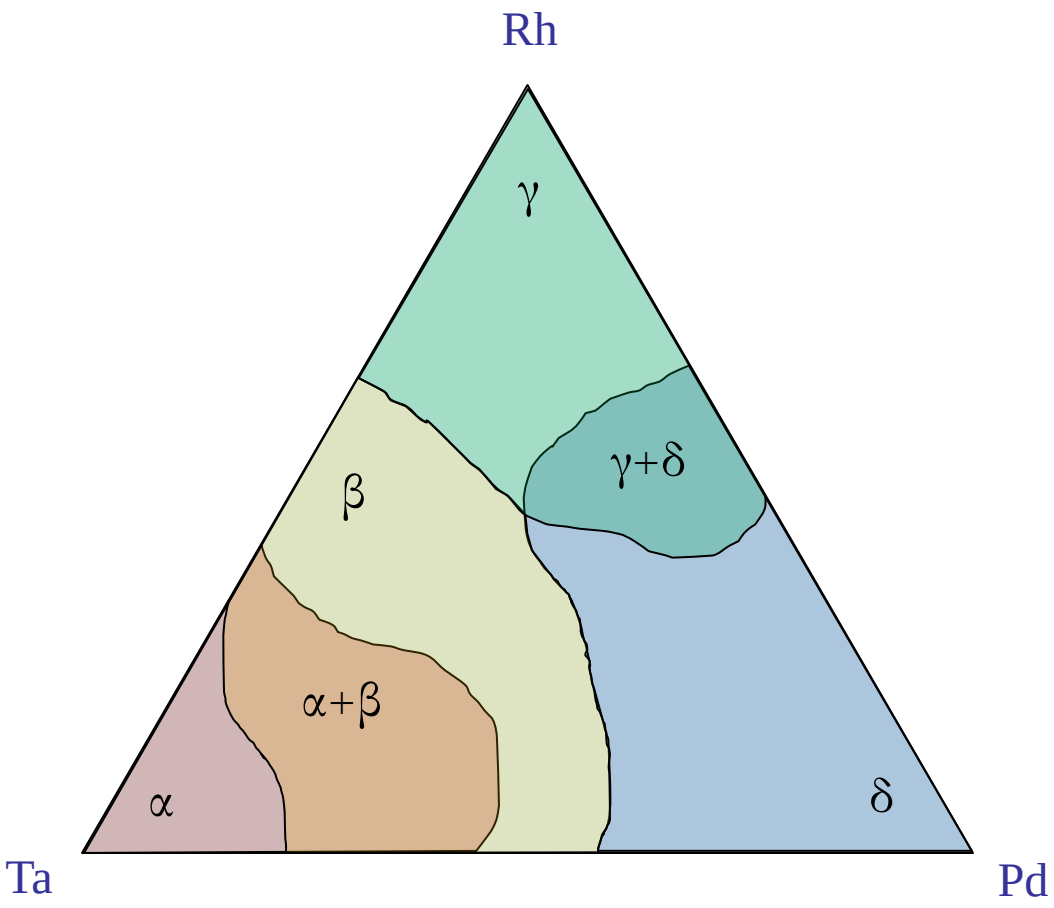
[Source: *Pyrotope*, Sebastien Merkel]



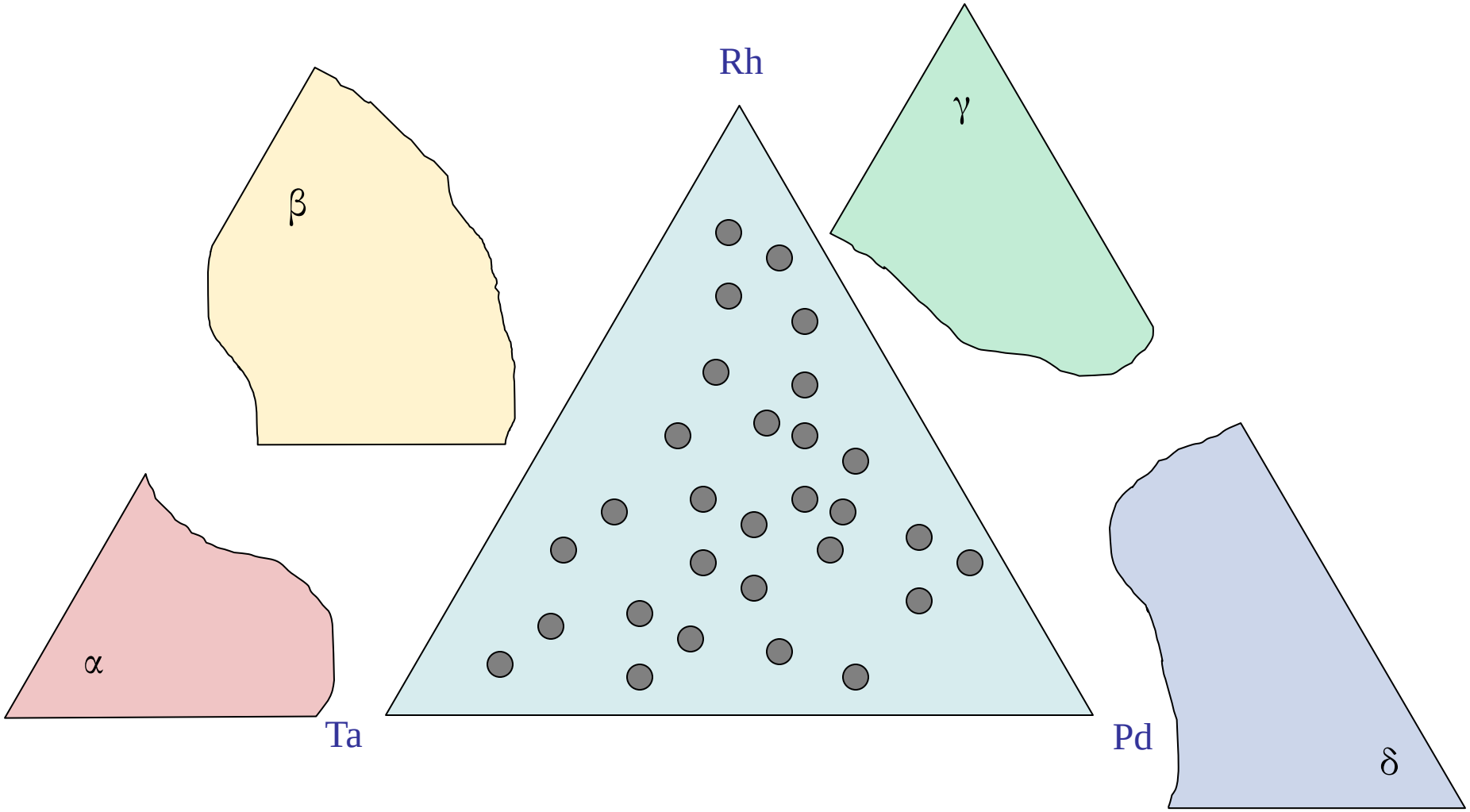
Motivation



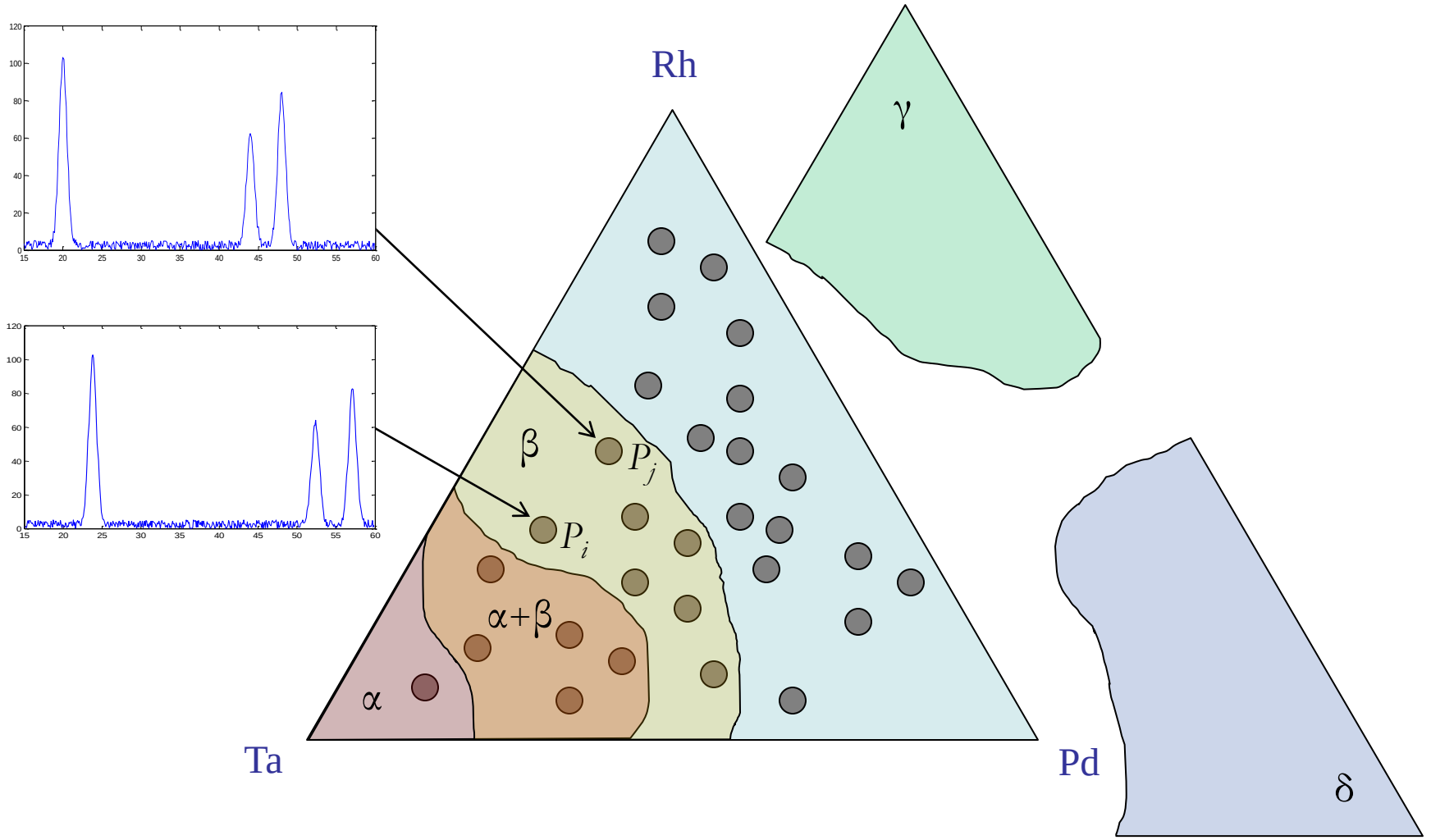
Motivation



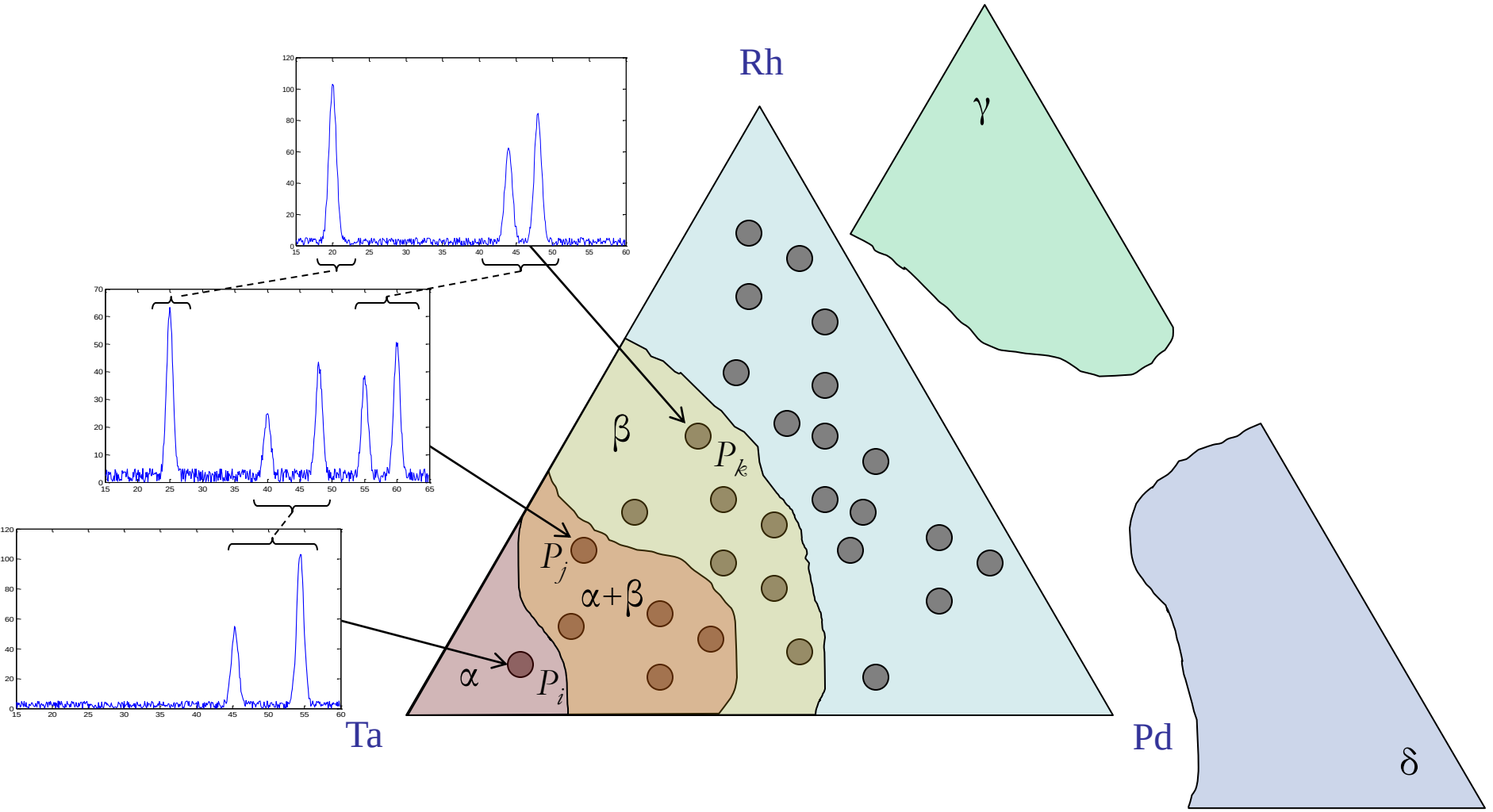
Motivation



Motivation

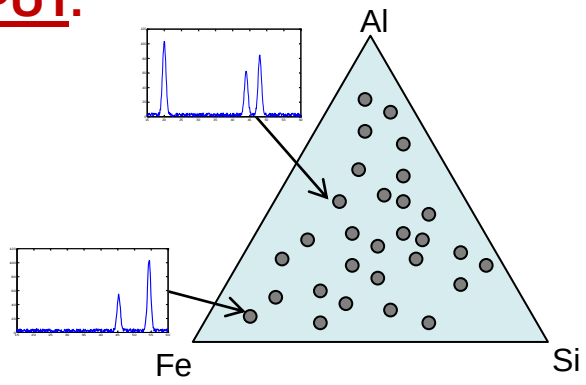


Motivation



Motivation

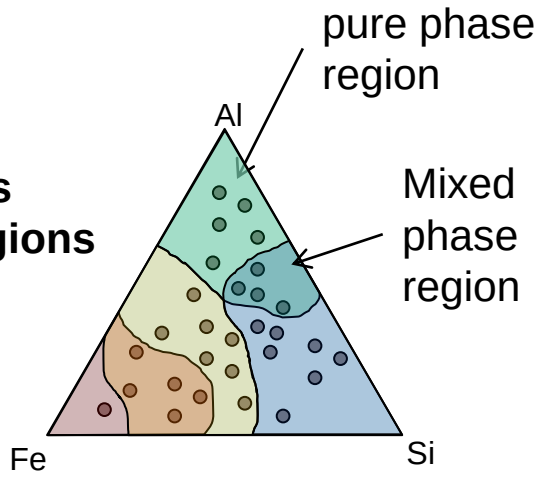
INPUT:



OUTPUT:

***m* phase regions**
- ***k* pure regions**
- ***m-k* mixed regions**

XRD pattern characterizing pure phases



Additional Physical characteristics:

- Peaks shift by ≤ 15% within a region
- Phase Connectivity
- Mixtures of ≤ 3 phases
- Small peaks might be discriminative
- Peak locations matter, more than peak intensities

Motivation

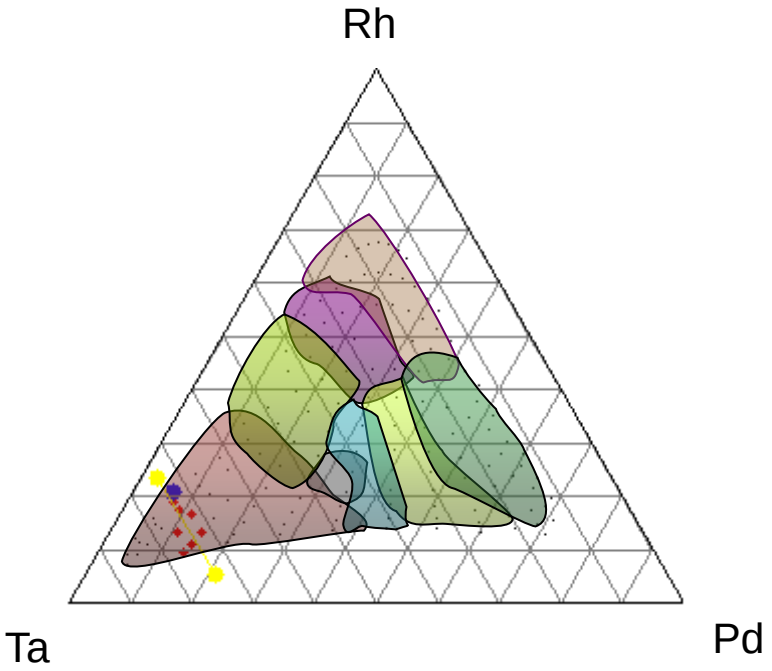


Figure 1: Phase regions of Ta-Rh-Pd

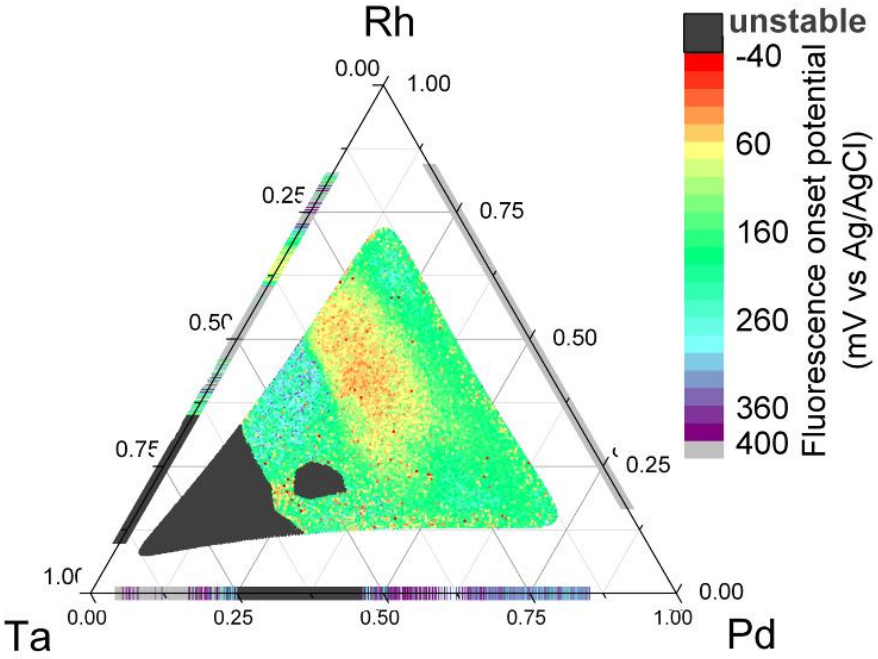


Figure 2: Fluorescence activity of Ta-Rh-Pd

Outline

- Motivation
- Problem Definition
  - ✓ Abstraction
  - ✓ Hardness
- CP Model
- Kernel-based Clustering
- Bridging CP and Machine learning
- Conclusion and Future work

Problem Abstraction: *Pattern Decomposition with Scaling*

Input

- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*

Problem Abstraction: *Pattern Decomposition with Scaling*

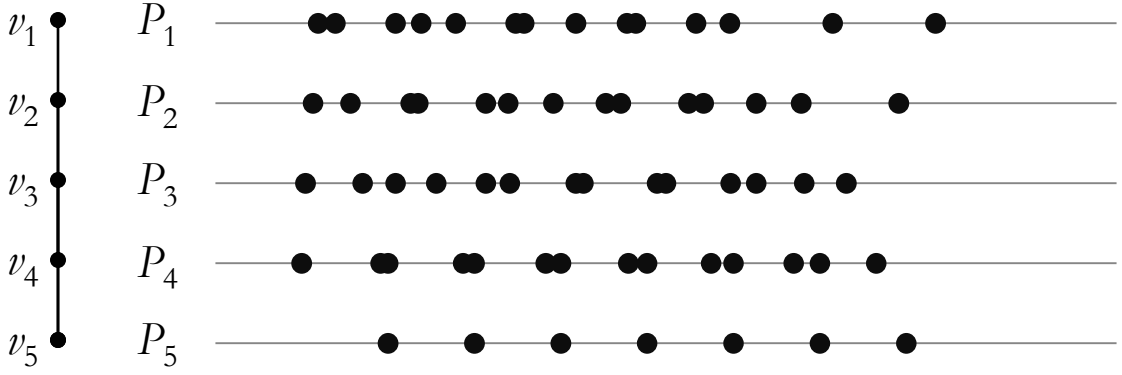
Input

- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*

Problem Abstraction: *Pattern Decomposition with Scaling*

Input

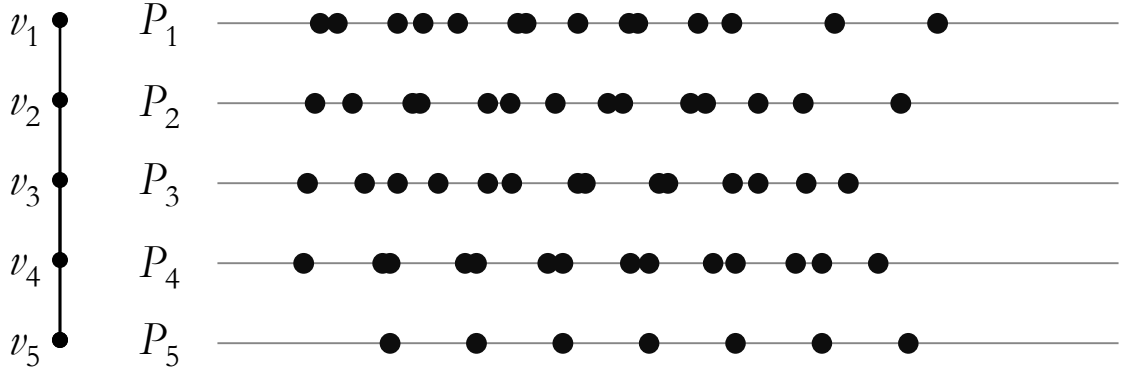
- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*



Problem Abstraction: *Pattern Decomposition with Scaling*

Input

- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*



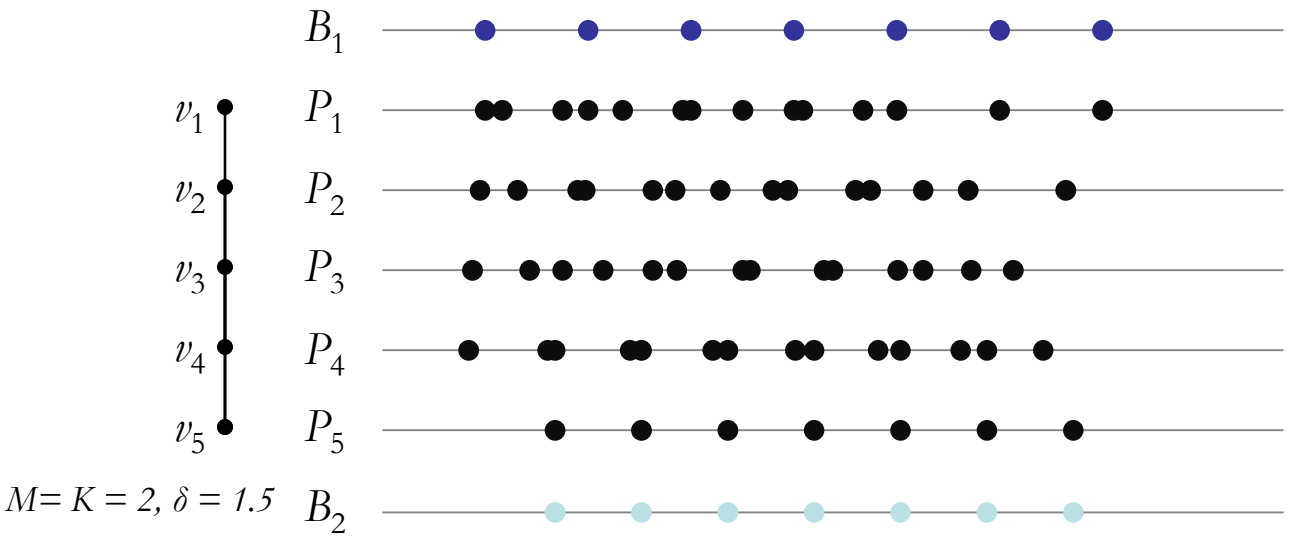
$M = K = 2, \delta = 1.5$

Problem Abstraction: *Pattern Decomposition with Scaling*

Input – $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
– $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
– $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*

Output *Determine whether there exists a collection* $\mathcal{B}$ *of* K basis patterns *over* S *and* scaling factors $s_{ik} \in \{0\} \cup [1/\delta, \delta]$ *such that:*

(a) $\forall i$: $P_i = \bigcup_{k=1}^{K} s_{ik} B_k$
(b) $\forall i$: $|\{k \mid s_{ik} > 0\}| \leq M$
(c) $\forall k$: *the subgraph of* G *induced by* $V_k = \{v_i \in V \mid s_{ik} > 0\}$ *is connected*



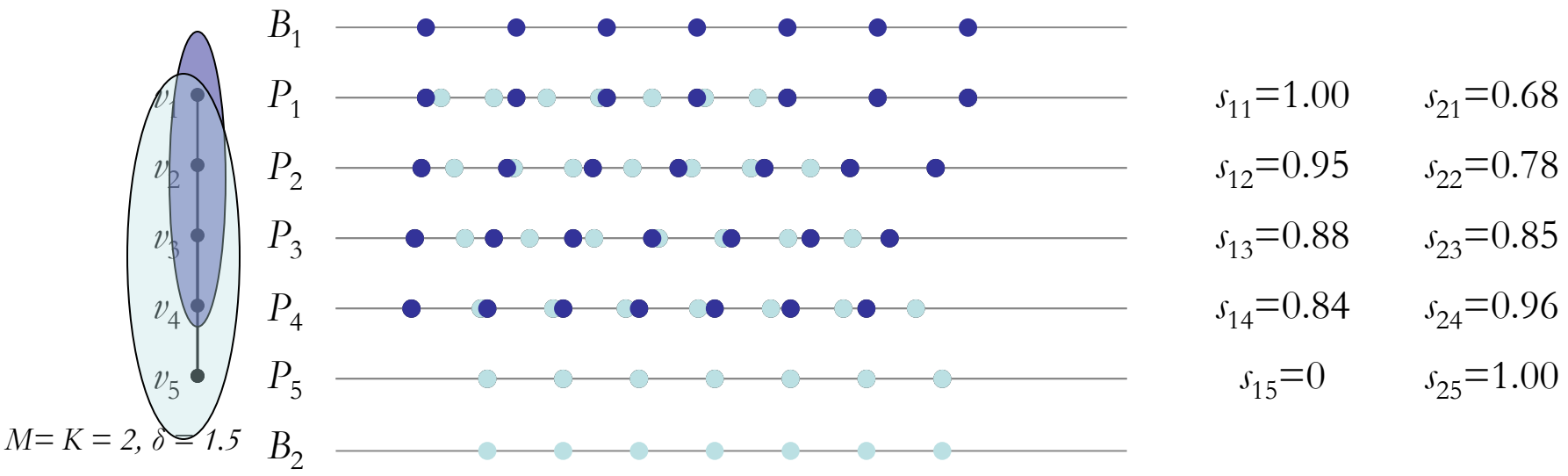
Problem Abstraction: *Pattern Decomposition with Scaling*

Input

- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*

Output *Determine whether there exists a collection* $\mathcal{B}$ *of* K basis patterns *over* S *and* scaling factors $s_{ik} \in \{0\} \cup [1/\delta, \delta]$ *such that:*

(a) $\forall i$: $P_i = \bigcup_{k=1}^{K} s_{ik} B_k$
(b) $\forall i$: $|\{k \mid s_{ik} > 0\}| \leq M$
(c) $\forall k$: *the subgraph of* G *induced by* $V_k = \{v_i \in V \mid s_{ik} > 0\}$ *is connected*



Problem Hardness

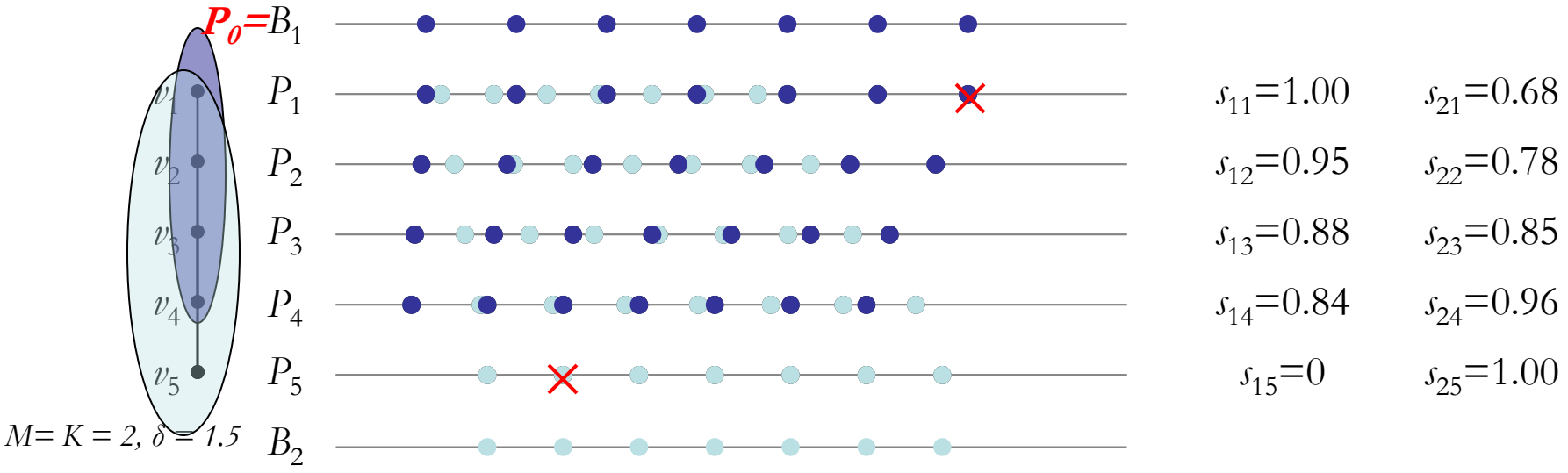
Assumptions: ***Each B_k appears by itself in some v_i / No experimental noise***

The problem can be solved* in **polynomial time**.

Assumption: ***No experimental noise***

The problem becomes ***NP-hard*** (reduction from the *"Set Basis"* problem)

Assumption used in this work: ***Experimental noise in the form of missing elements in P_i***



Outline

- Motivation
- Problem Definition
- CP Model
  - ✓Model
  - ✓Experimental Results
- Kernel-based Clustering
- Bridging CP and Machine learning
- Conclusion and Future work

CP Model

Variables	Description
p_{ki}	Normalizing element for phase k in pattern P_i
a_{ki}	Whether phase k is present in pattern P_i
q_k	Set of normalized peak locations of phase B_k

$$(a_{ki} = 0) \iff (p_{ki} = 0) \qquad \forall\ 1 \le k \le K, 1 \le i \le n \tag{1}$$

$$1 \le \sum_{K} a_{si} \le M \qquad \forall\ 1 \le i \le n \tag{2}$$

$$(p_{ki} = j) \Rightarrow (q_k \subseteq r_{ij}) \qquad \forall\ 1 \le k \le K, 1 \le i \le n, 1 \le j \le |P_i| \tag{3}$$

$$(p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1) \Rightarrow (r_{ij} \subseteq q_k)$$

$$\forall\ 1 \le k \le K, 1 \le i \le n, 1 \le j \le |P_i| \tag{4}$$

CP Model (continued)

$$(p_{ki} = j \wedge p_{k'i} = j' \wedge \sum_{s=1}^{K} a_{si} = 2)$$
$$\Rightarrow (member(r_{ij}[j''], q_k) \vee member(r_{ij'}[j''], q_{k'}))$$
$$\forall\ 1 \leq k, k' \leq K, 1 \leq i \leq n, 1 \leq j, j', j'' \leq |P_i| \quad (5)$$

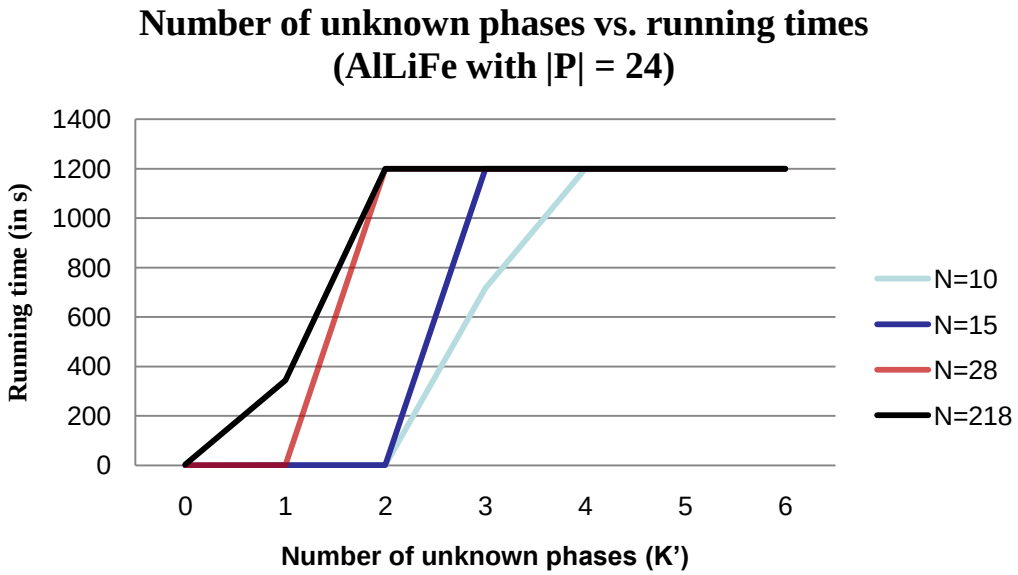
$$(p_{ki} = j) \Rightarrow (p_{ki'} \neq j') \qquad \forall\ 1 \leq k \leq K, (i, j, i', j') \in \Phi \quad (6)$$

$$\Phi = \{(i, j, i', j') \mid \frac{P_i[j]}{P'_i[j']} < 1/\delta \vee \frac{P_i[j]}{P'_i[j']} > \delta, i < i'\}$$

$$basisPatternConnectivity(\{a_{ki} | 1 \leq i \leq n\}) \qquad \forall\ 1 \leq k \leq K \quad (7)$$

Advantage: Captures physical properties and relies on peak location rather than height.
Drawback: Does not scale to realistic instances; poor propagation if experimental noise.

CP Model – Experimental Results



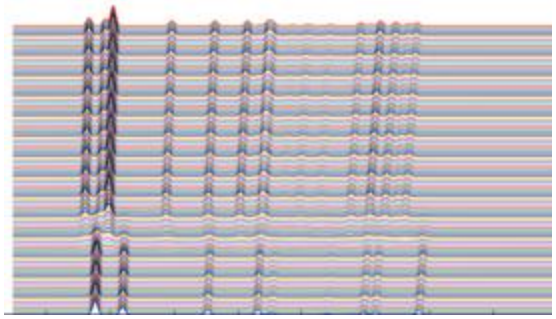
For realistic instances, K' = 6 and N ≈ 218...

Outline

- Motivation
- Problem Definition
- CP Model
- Kernel-based Clustering
- Bridging CP and Machine learning
- Conclusion and Future work

Kernel-based Clustering

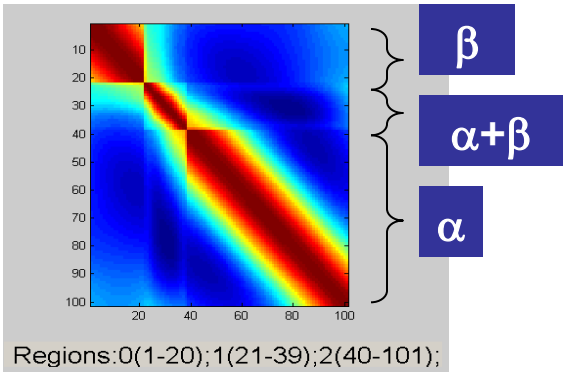
Set of features: X = α, α+β, β



Similarity matrices:

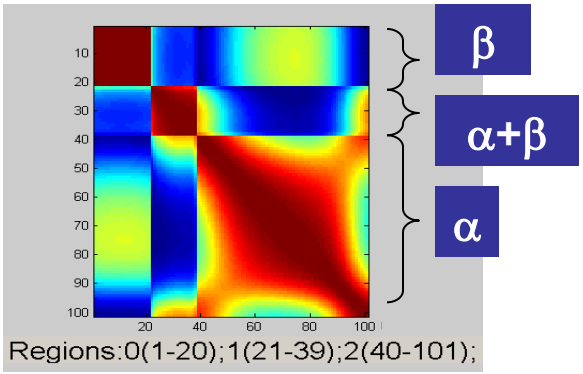
[X.X^T]

$$k_{DTW}(x_i, x_j) = \exp\left(-\frac{||c_i - c_j||^2}{2\sigma^2}\right)$$



Better approach: take shifts into account

Red: similar
Blue: dissimilar

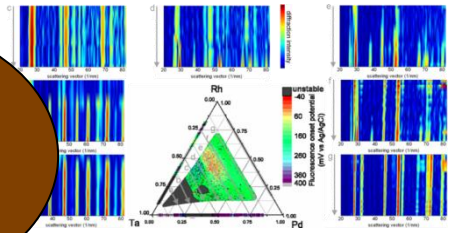


Method: K-means on Dynamic Time-Warping kernel

Goal: Select groups of samples that belong to the same **phase region** to feed the CP model, in order to extract the underlying phases of these sub-problems.

Bridging CP and ML

Goal: a robust, *physically meaningful*, scalable, automated solution method that combines:



Underlying Physics

A language for Constraints enforcing "local details"

Machine Learning for a "global data-driven view"

$$a_{ki} = 0 \iff p_{ki} = 0 \qquad \forall\, 1 \le k \le$$

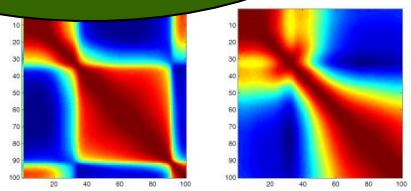
$$1 \le \sum_{s=1}^{K} a_{si} \le 3 \qquad \forall\, 1 \le i \le n$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow q_k \subseteq r_{ij} \qquad \forall\, 1 \le k \le K$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow r_{ij} \subseteq q_k \qquad \forall\, 1 \le k \le K, 1 \le$$

$$P(k,k',i,j,j') \rightarrow \begin{cases} member(r_{ij}[j''], q_k) \\ \vee \\ member(r_{ij'}[j''], q_{k'}) \end{cases} \qquad \forall\, 1 \le k < k$$

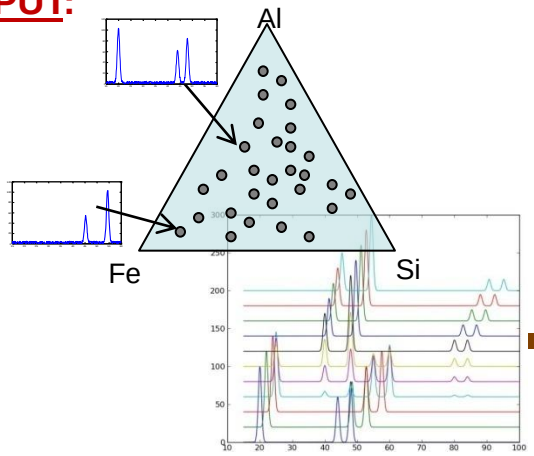
Constraint Programming model



Similarity "Kernels" & Clustering

Bridging Constraint Reasoning and Machine Learning: Overview of the Methodology

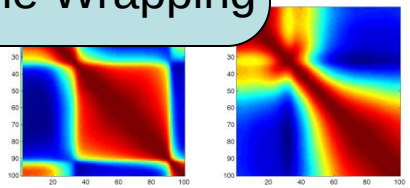
INPUT:



Peak detection

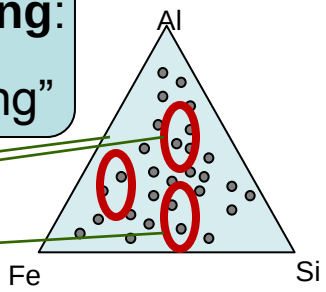
Machine Learning:

Kernel methods, Dynamic Time Wrapping



Machine Learning:

Partial "Clustering"



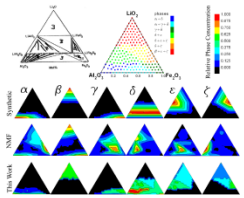
Fix errors in data

CP Model & Solver on sub-problems

Full CP Model guided by partial solutions

- α only
- γ only
- α, α+β, β

OUTPUT



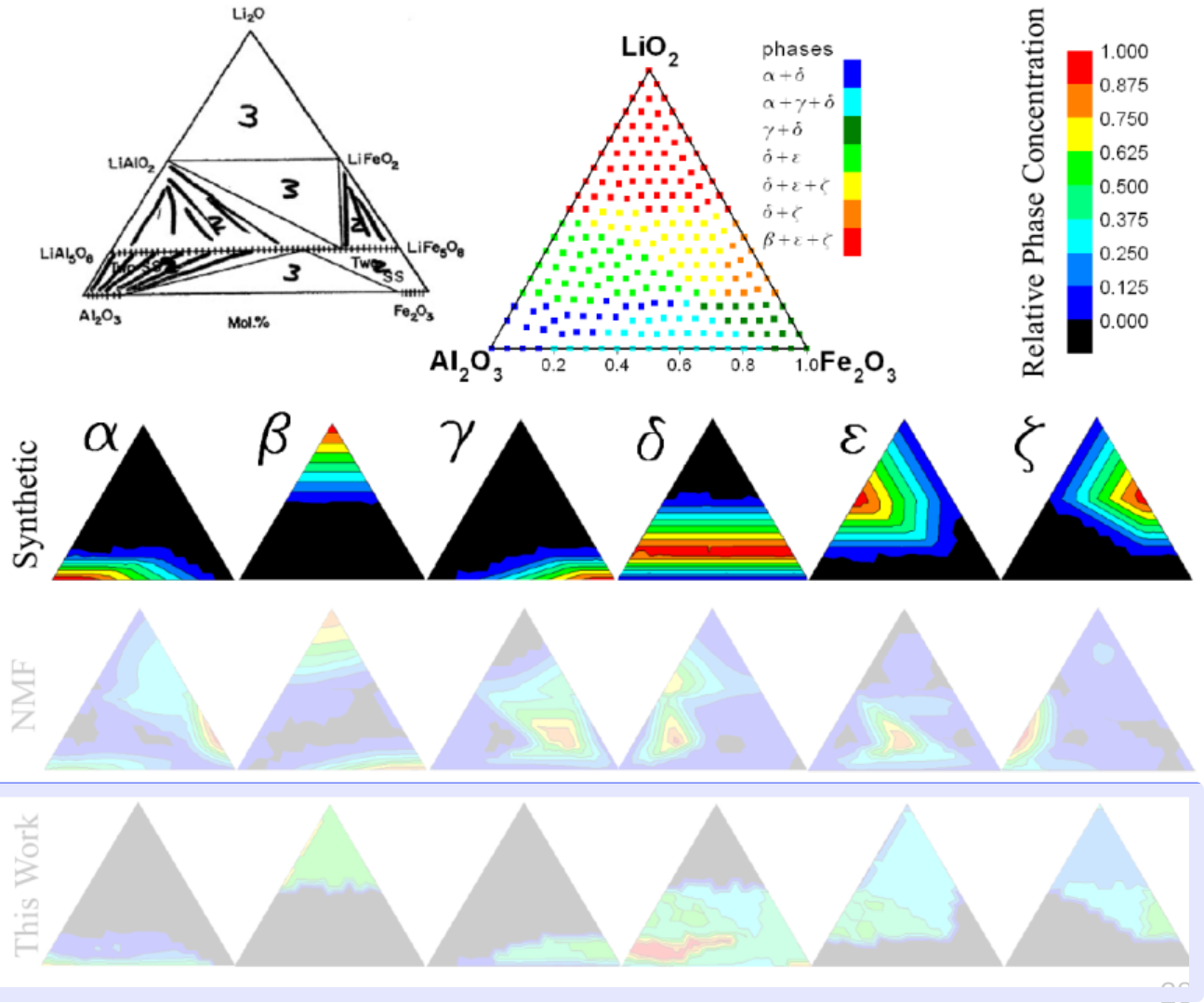
Experimental Validation

Al-Li-Fe instance with 6 phases:

Ground truth (known)

Previous work (NMF) violates many physical requirements

Our CP + ML hybrid approach is much more robust



Conclusion

Hybrid approach to clustering under constraints

- More robust than data-driven “global” ML approaches
- More scalable than a pure CP model “locally” enforcing constraints

An exciting application in close collaboration with physicists

- Best inference out of expensive experiments
- Towards the design of better fuel cell technology

Future work

Ongoing work:

Spatial Clustering, to further enhance cluster quality

Bayesian Approach, to better exploit prior knowledge about local smoothness and available inorganic libraries

Active learning:

Where to sample next, assuming we can interfere with the sampling process?

When to stop sampling if sufficient information has been obtained?

Correlating catalytic properties across many thin-films:

In order to understand the underlying **physical mechanism** of **catalysis** and to find promising intermettalic compounds

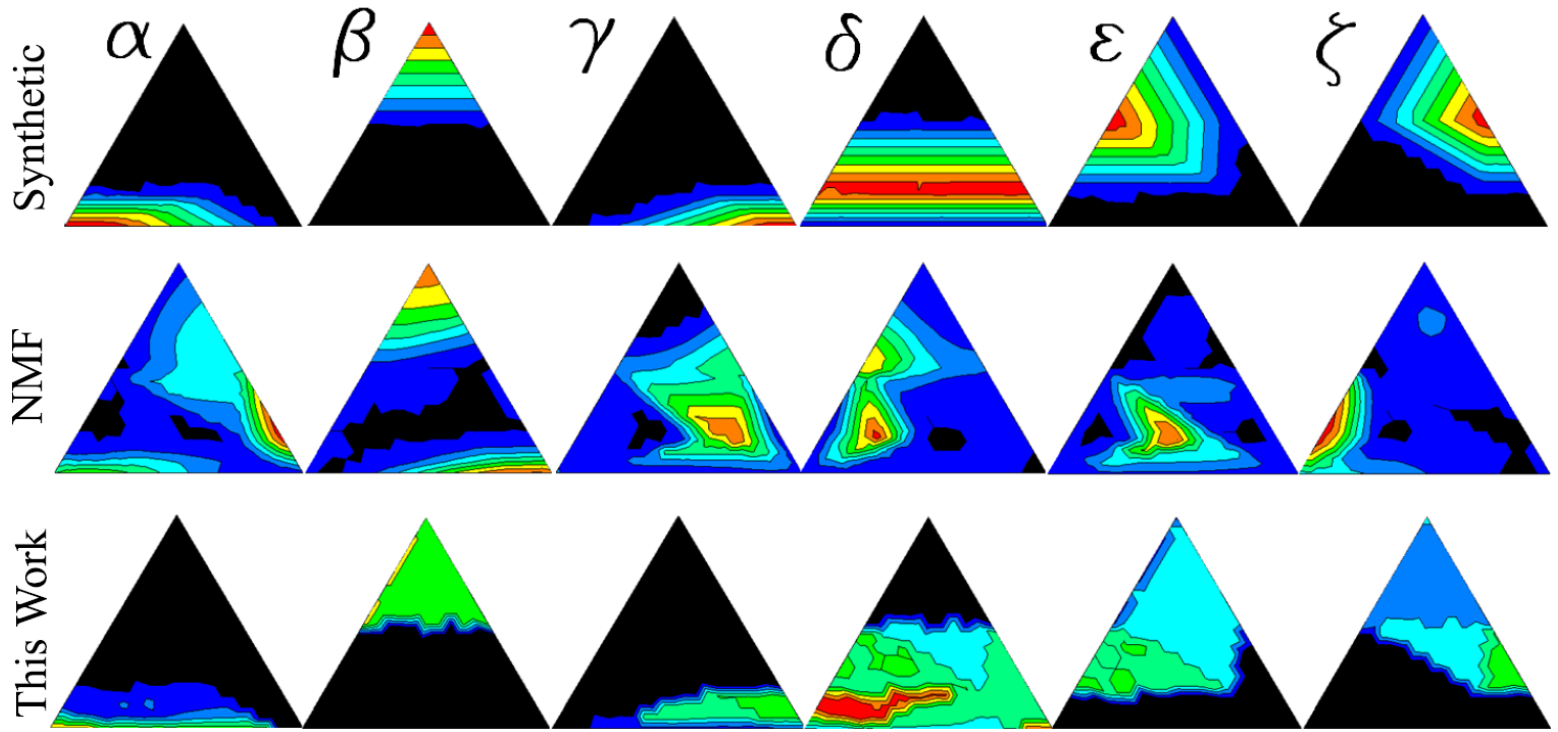
The end

Thank you!

Extra slides

Experimental Sample

Example on Al-Li-Fe diagram:



Applications with similar structure

Flight Calls / Bird conservation

Identifying bird population from sound recordings at night.

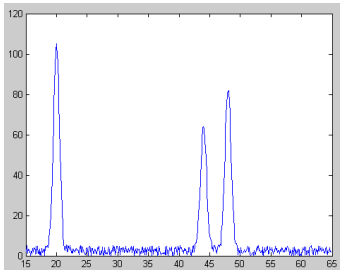
Analogy: basis pattern = species
samples = recordings
physical constraints = spatial constraints, species and season specificities…

Fire Detection

Detecting/Locating fires.

Analogy: basis pattern = warmth sources
samples = temperature recordings
physical constraints = gradient of temperatures, material properties…

Previous Work 1: Cluster Analysis *(Long et al., 2007)*

$x_i =$ 

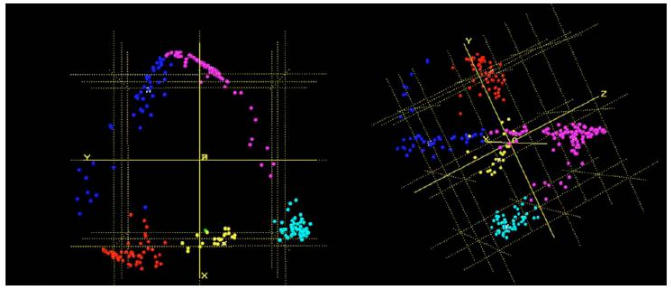
(Feature vector)

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}}$$

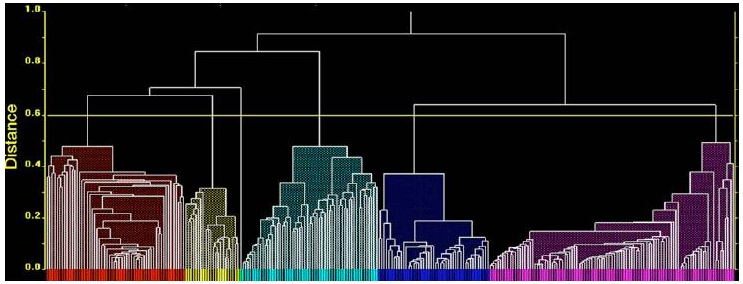
(Pearson correlation coefficients)

$$D=(1-C)/2$$

(Distance matrix)



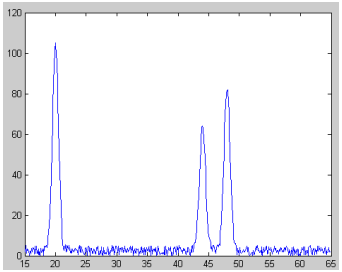
(PCA – 3 dimensional approx)



(Hierarchical Agglomerative Clustering)

Drawback: Requires sampling of pure phases, detects phase regions (not phases), overlooks peak shifts, may violate physical constraints (phase continuity, etc.).

Previous Work 2: NMF *(Long et al., 2009)*

$x_i =$ 

$\longrightarrow$ $X = A.S + E$ $\longrightarrow$ $Min \left\| E \right\|$

(Feature vector)

(Linear positive combination (A) of basis patterns (S))

(Minimizing squared Frobenius norm

Drawback: Overlooks peak shifts (linear combination only), may violate physical constraints (phase continuity, etc.).