CORNELL UNIVERSITY

Computational Challenges in Material Discovery: Bridging Constraint Reasoning and Machine Learning

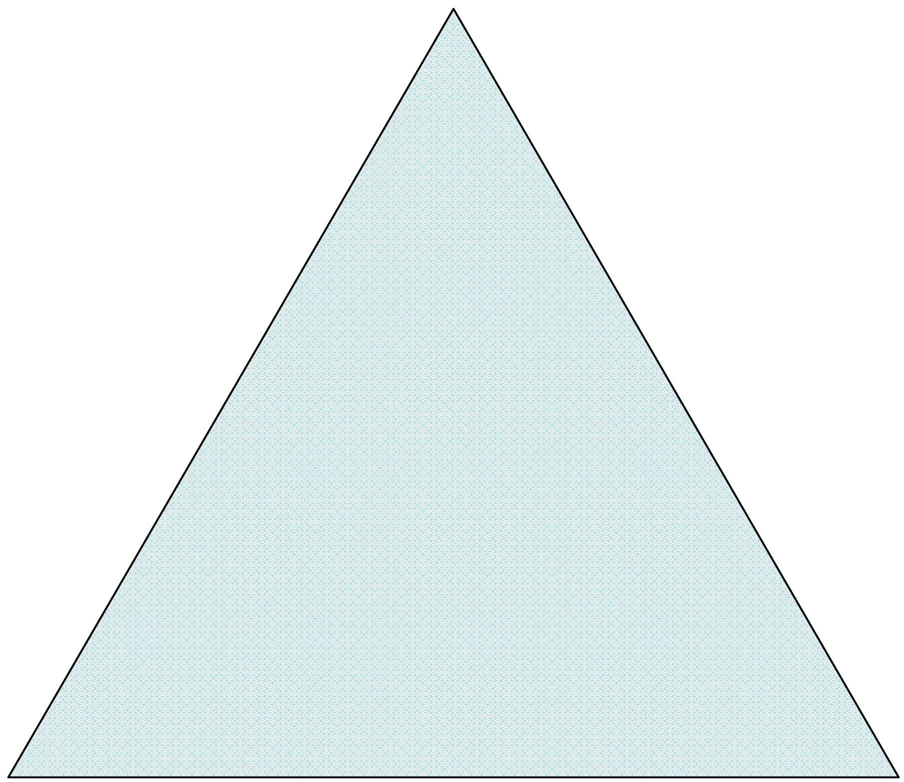
Institute for Computational Sustainability ICS

IISI Intelligent Information Systems Institute

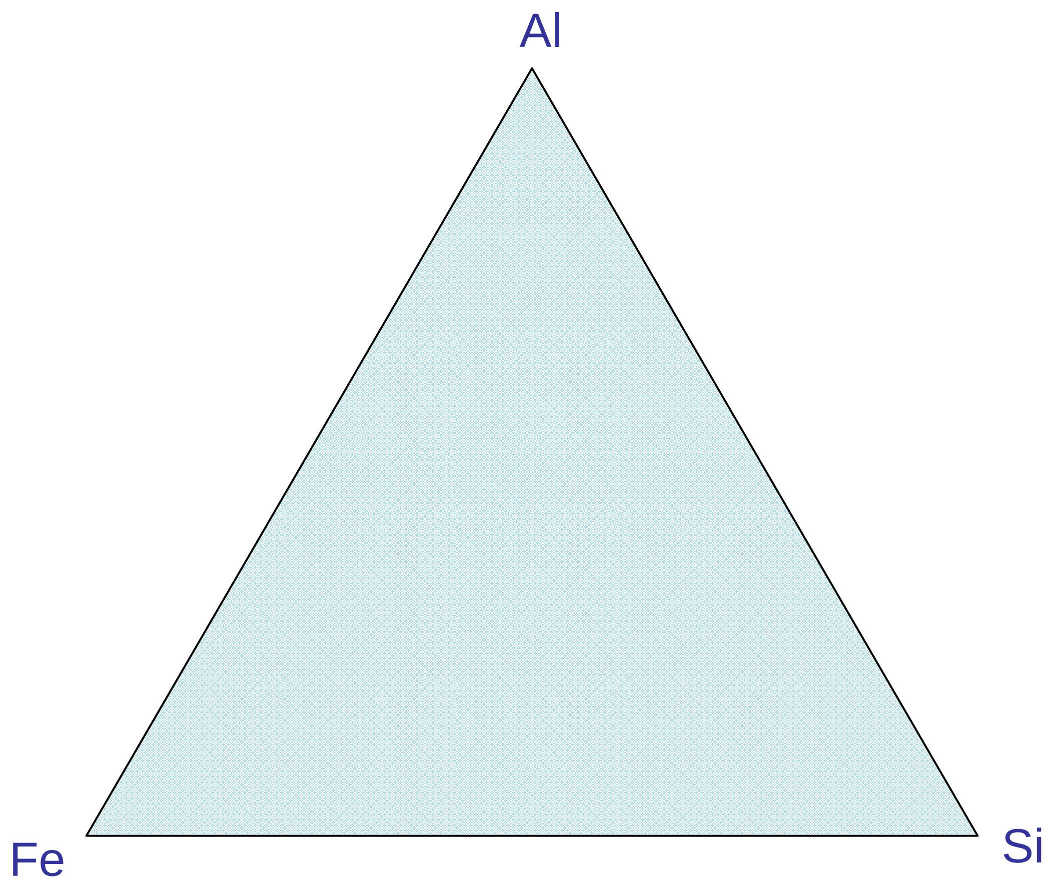
Ronan LeBras	Computer Science
Theodoros Damoulas	Computer Science
John M. Gregoire	Materials Science / Physics
Ashish Sabharwal	Computer Science
Carla P. Gomes	Computer Science
Bruce van Dover	Materials Science / Physics

June 15, 2010 **CROCS'10**

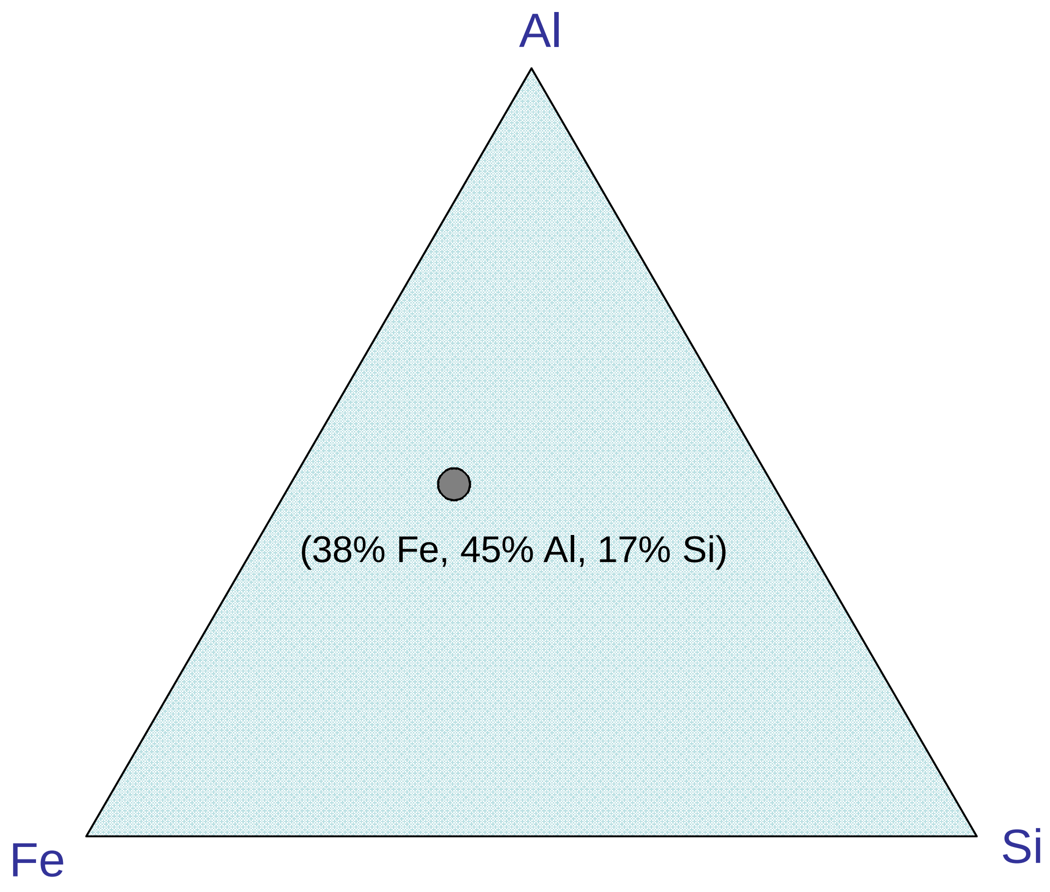
Problem Definition



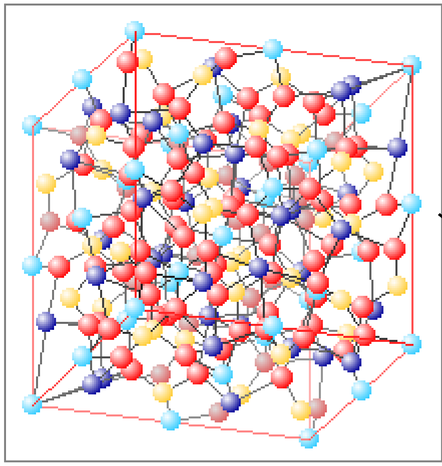
Problem Definition



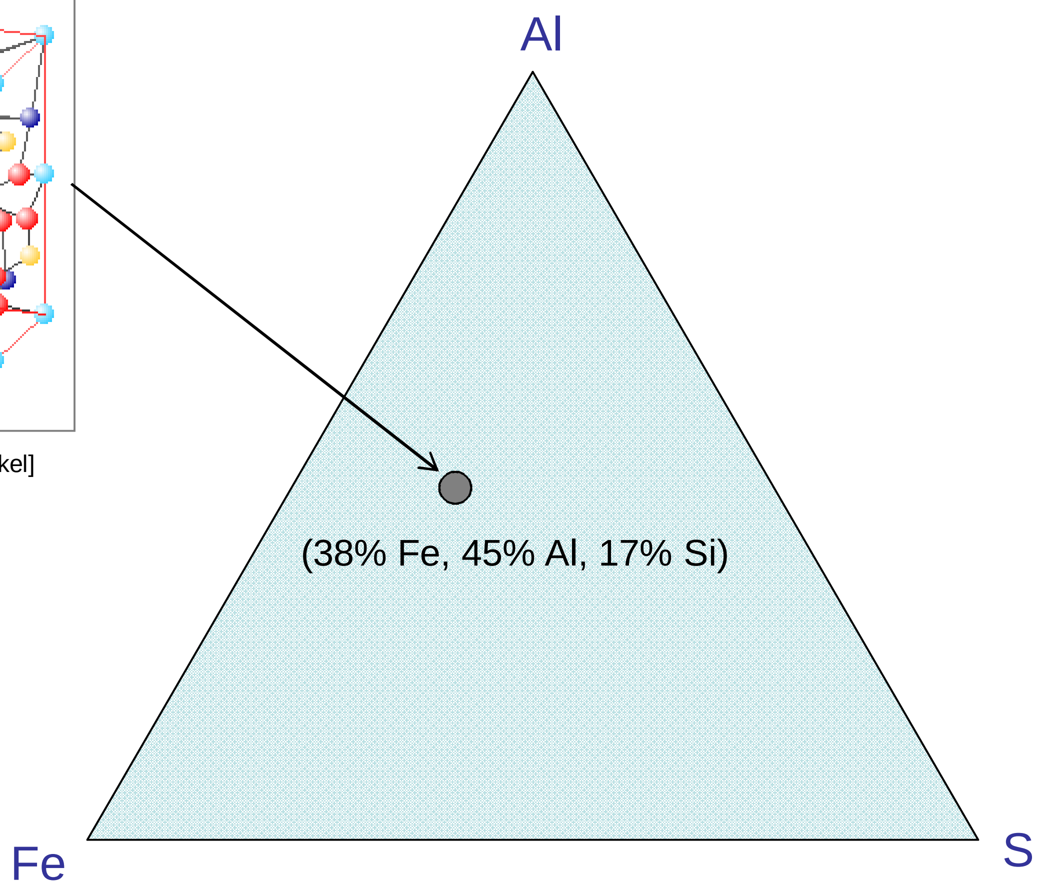
Problem Definition



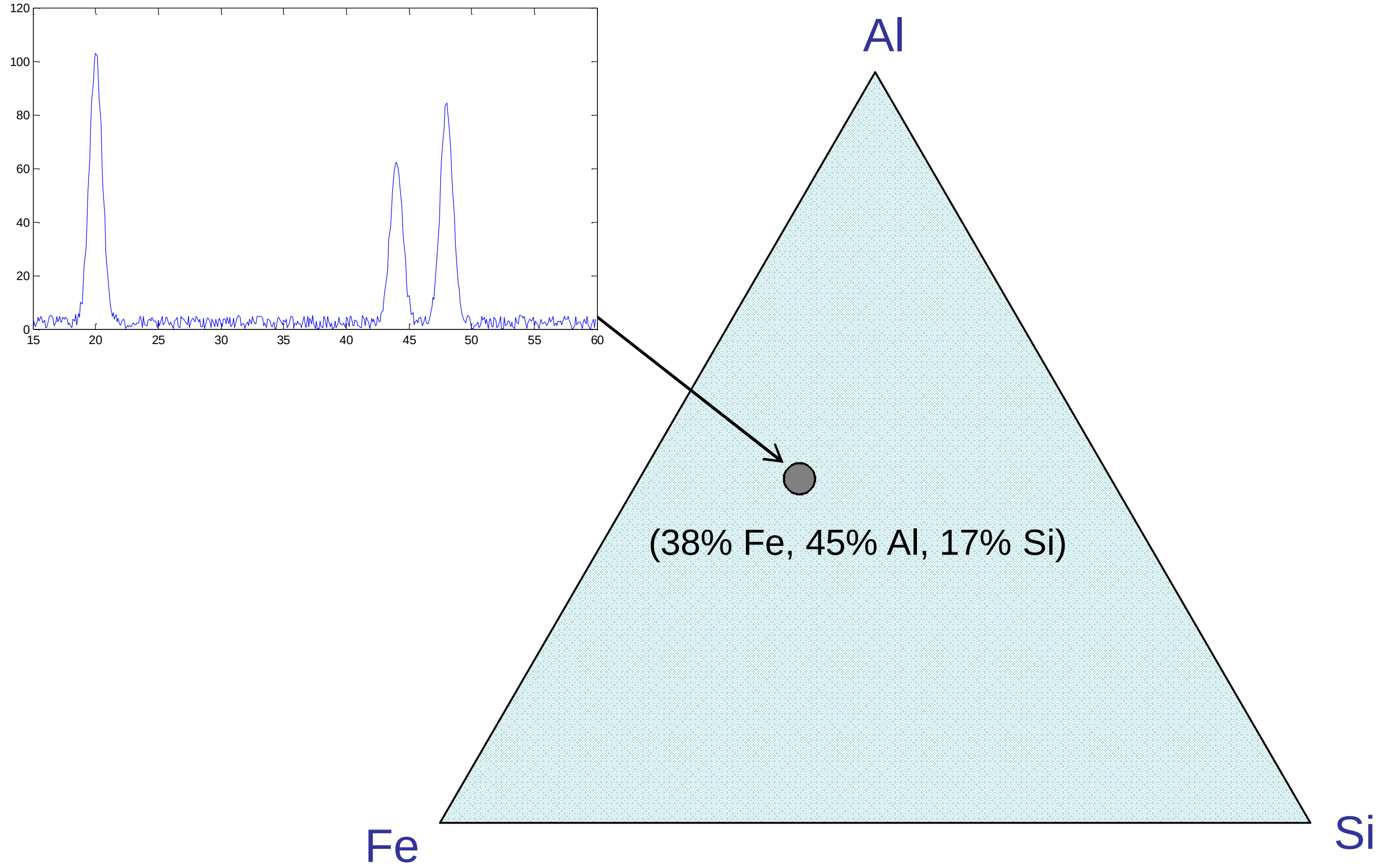
Problem Definition



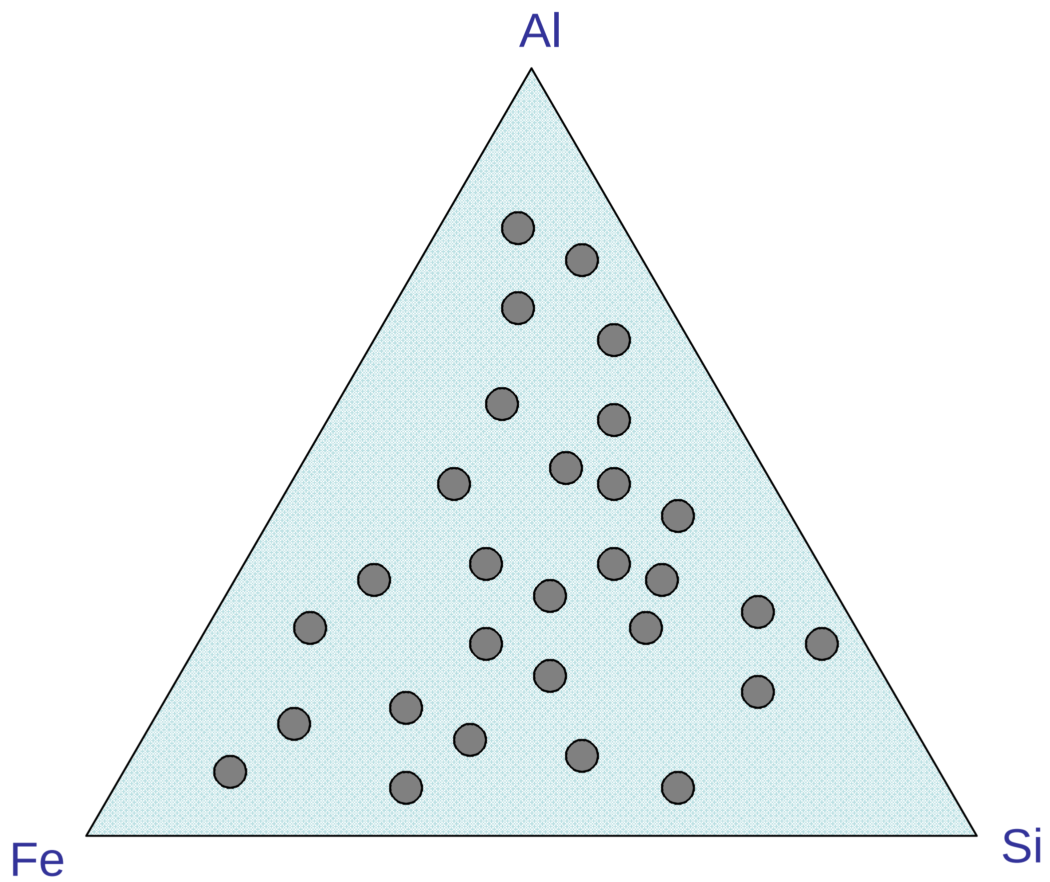
[Source: *Pyrotope*, Sebastien Merkel]



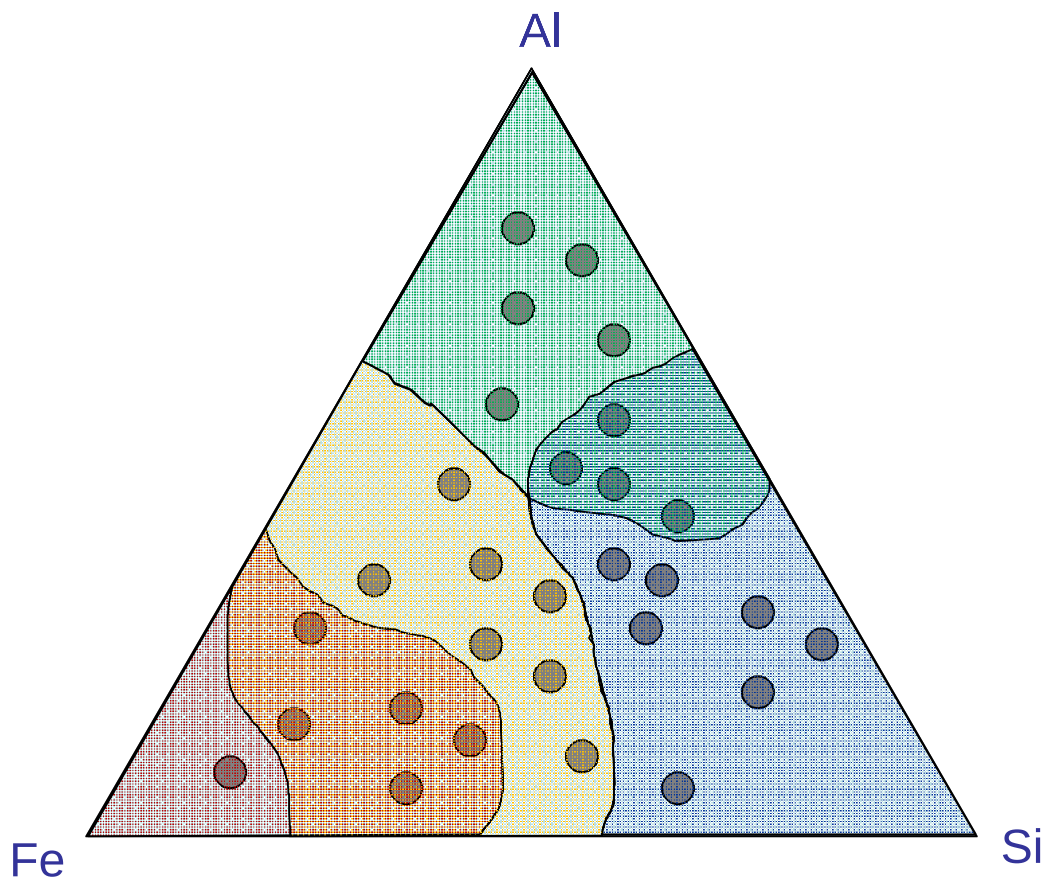
Problem Definition



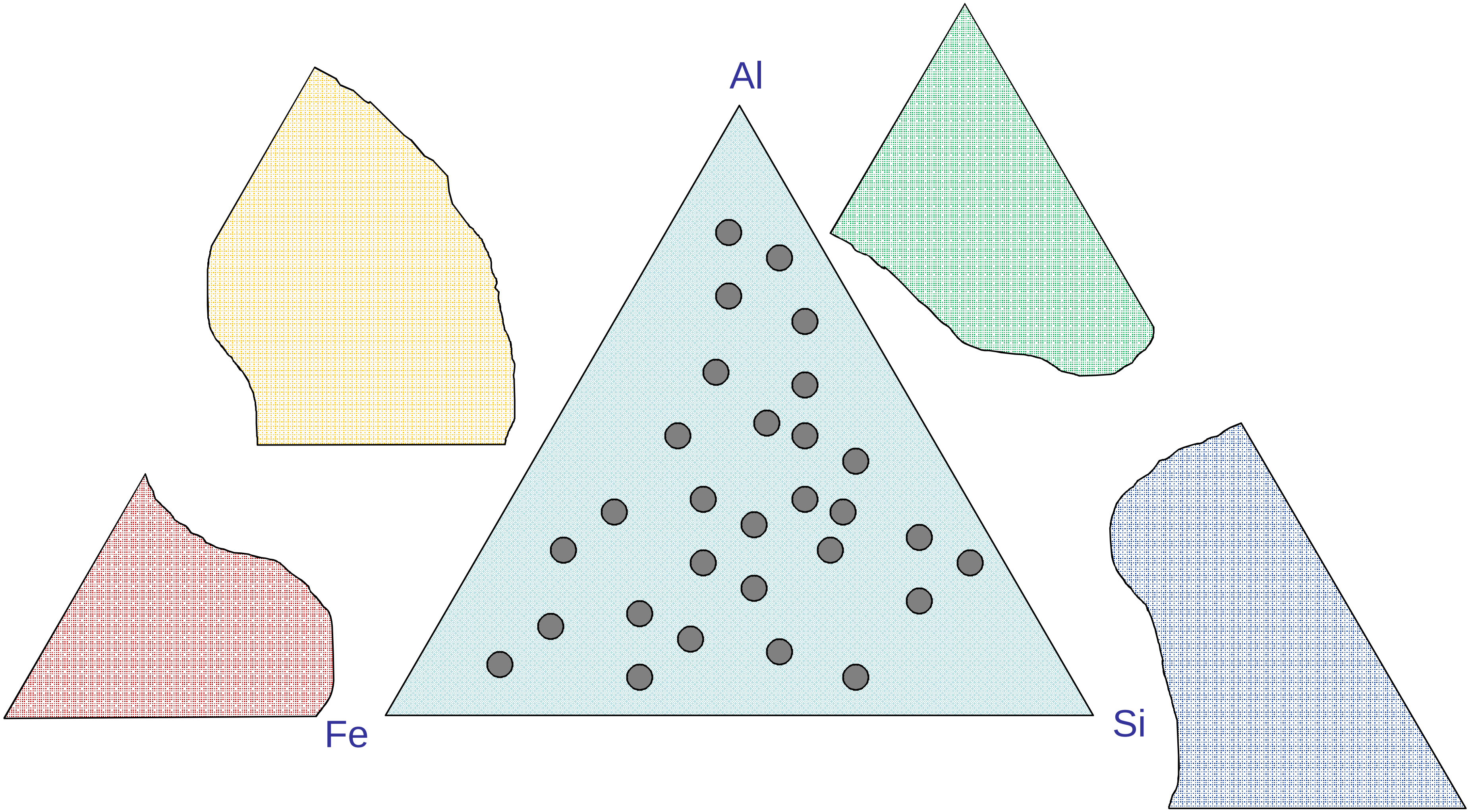
Problem Definition



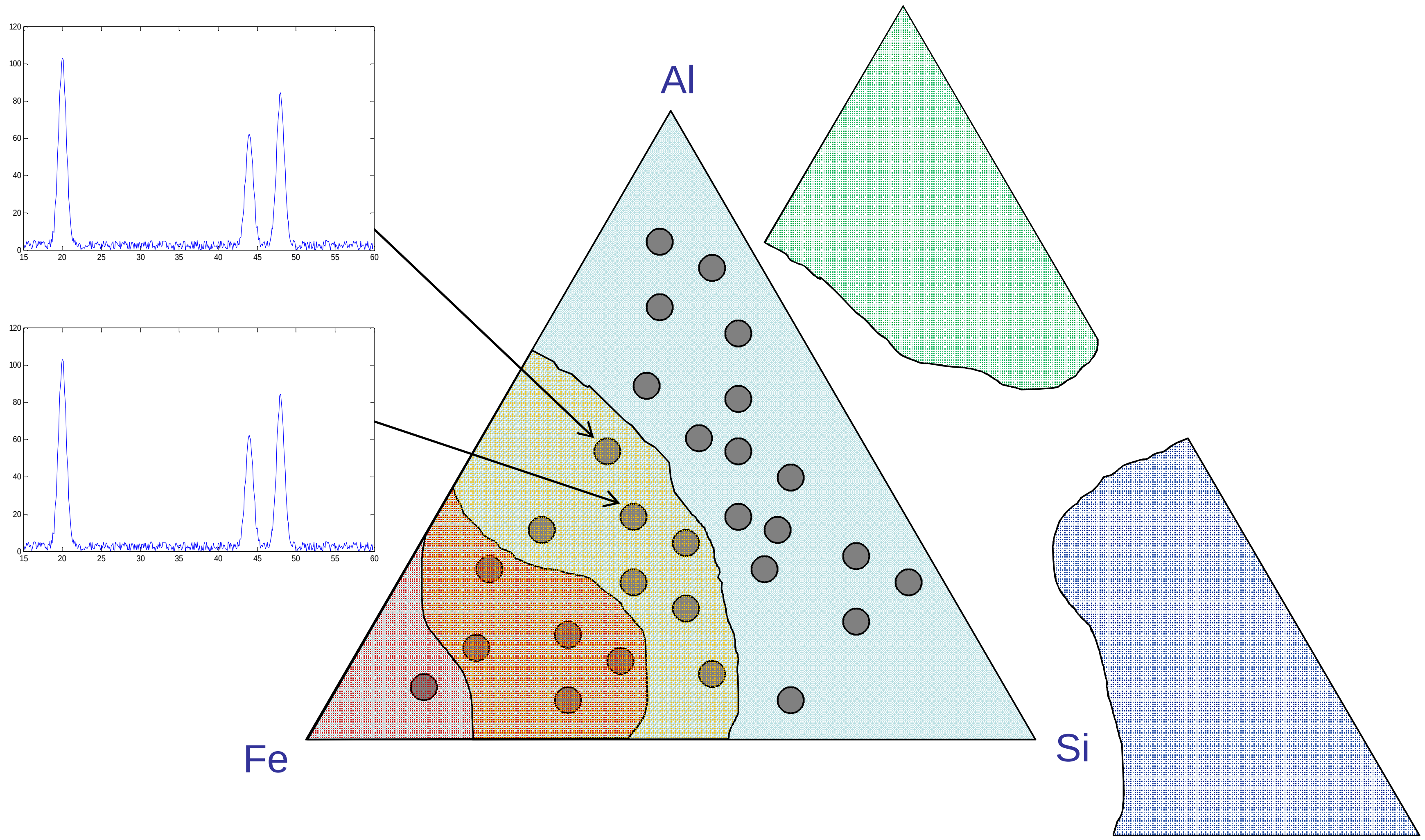
Problem Definition



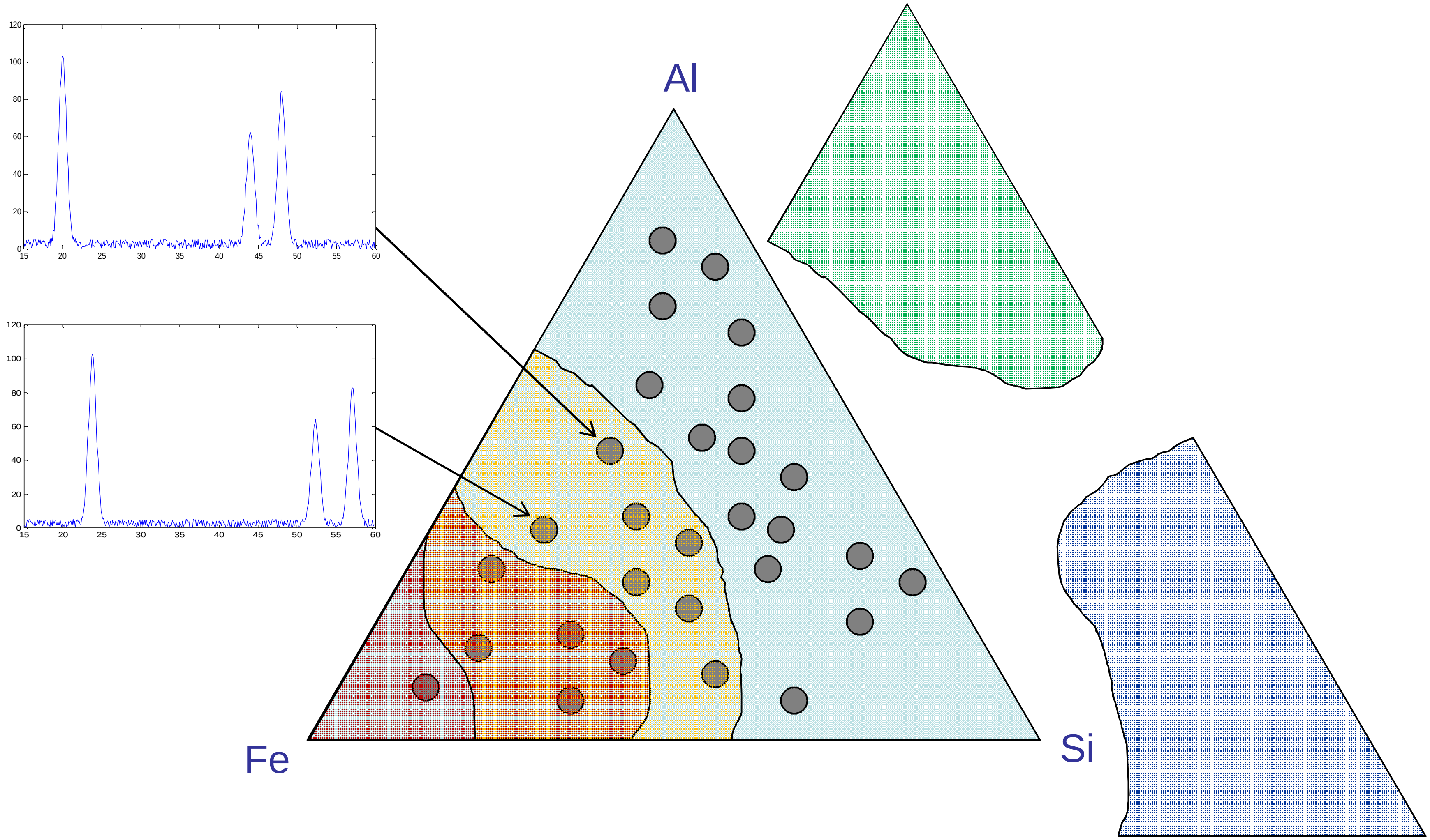
Problem Definition



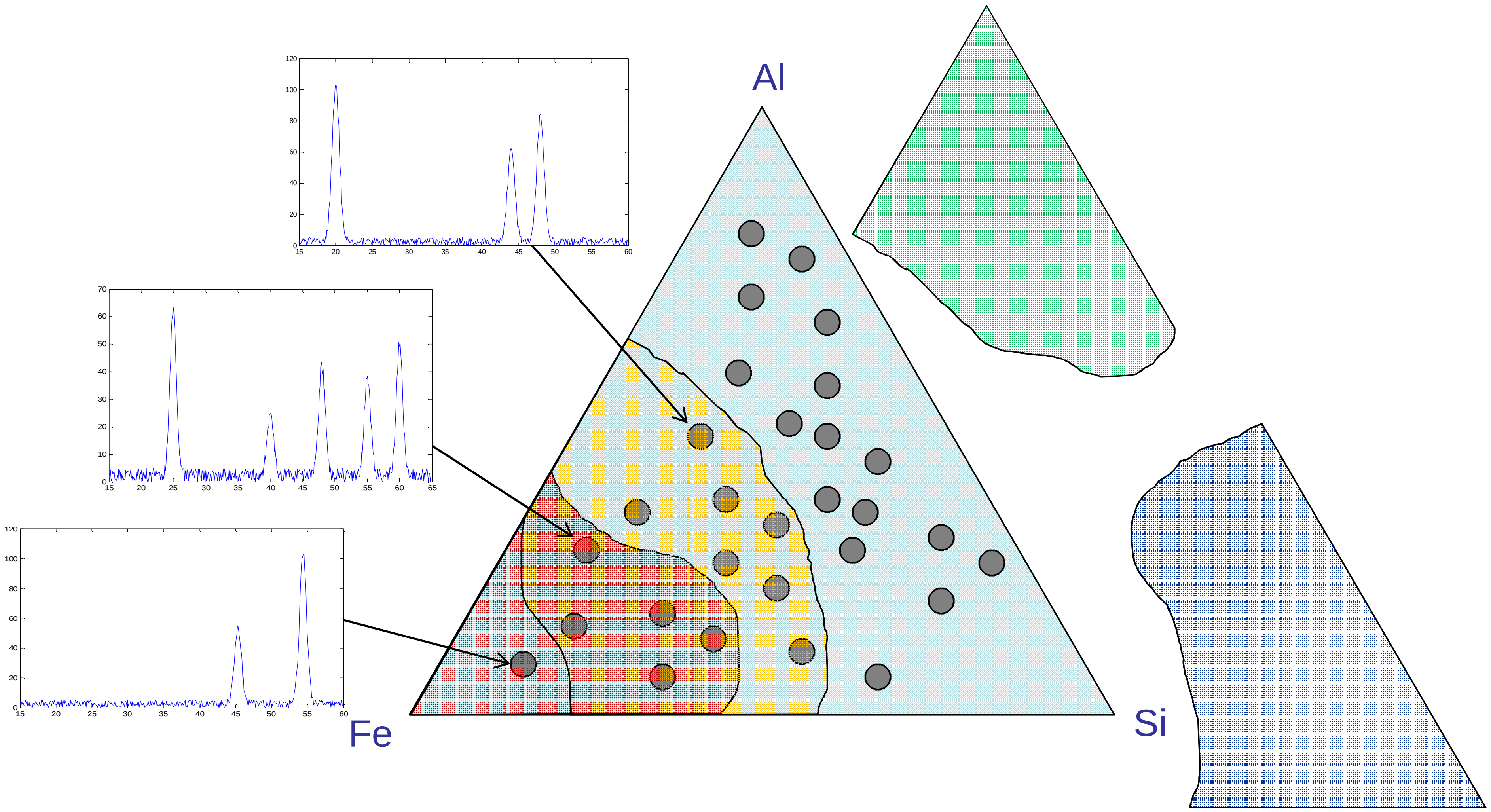
Problem Definition



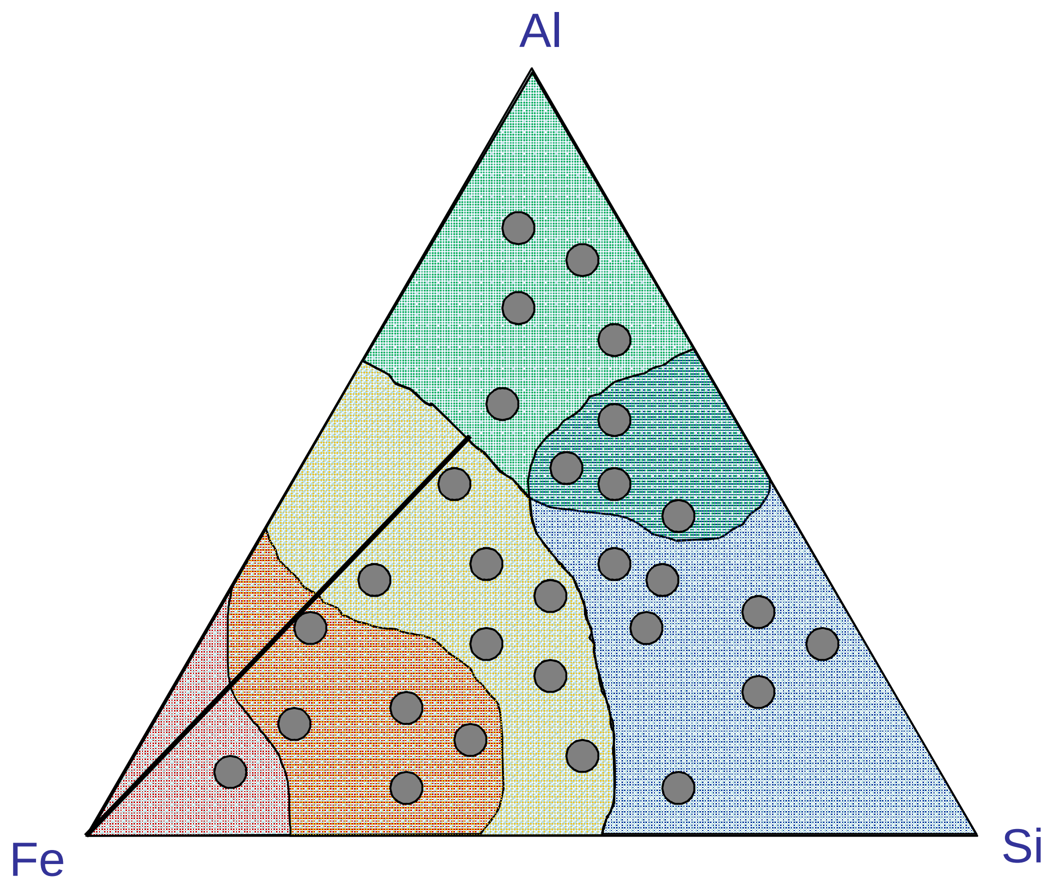
Problem Definition



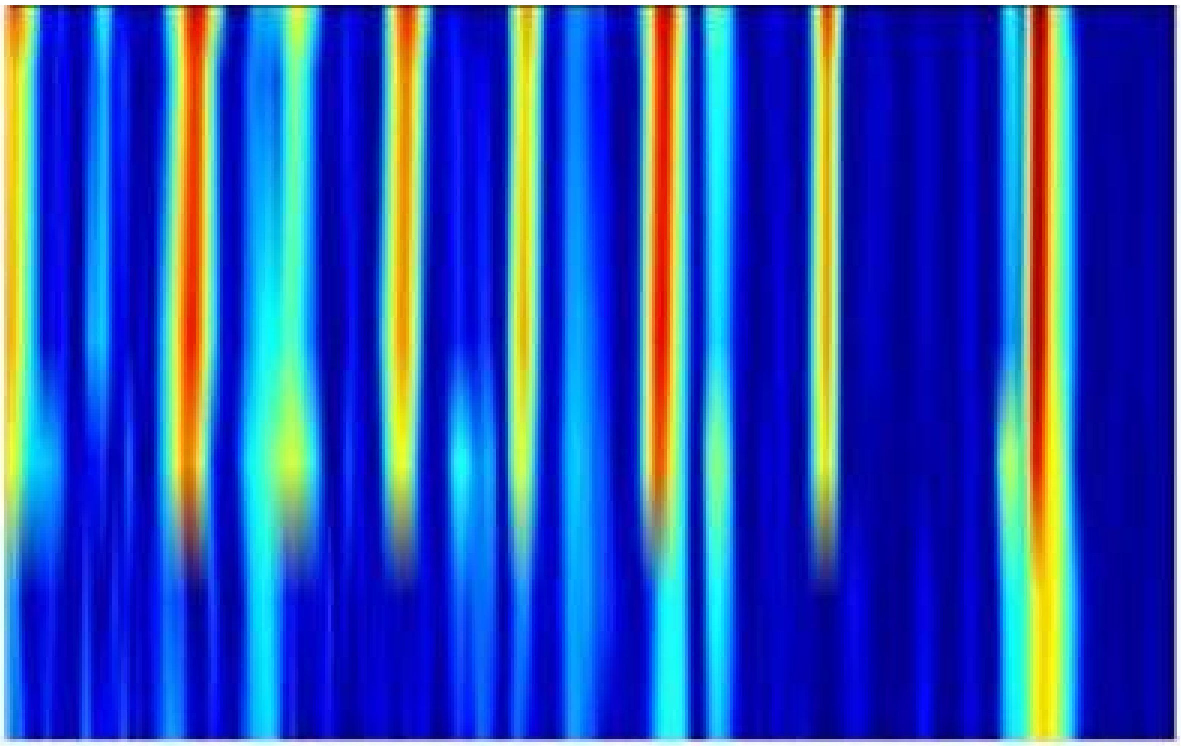
Problem Definition



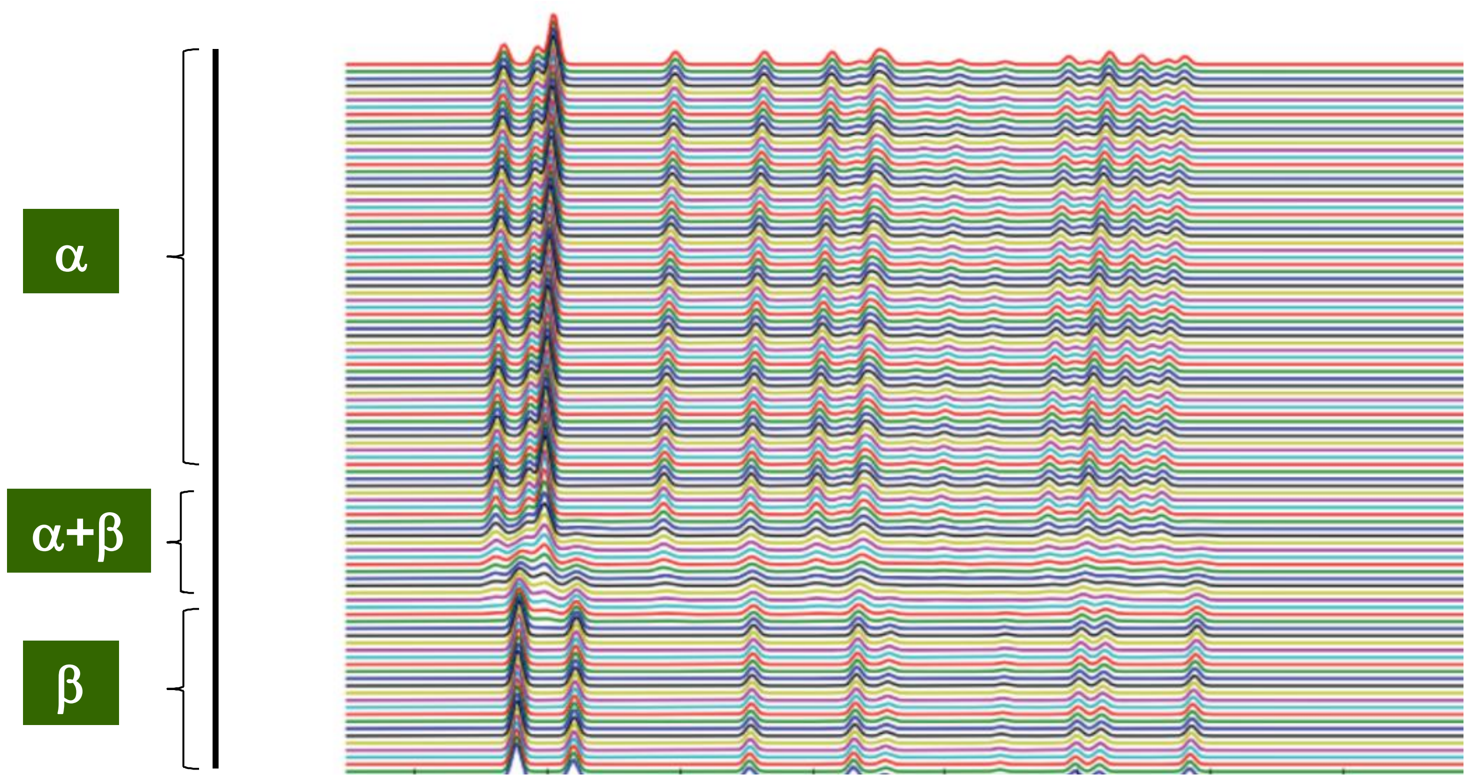
Problem Definition



Problem Definition

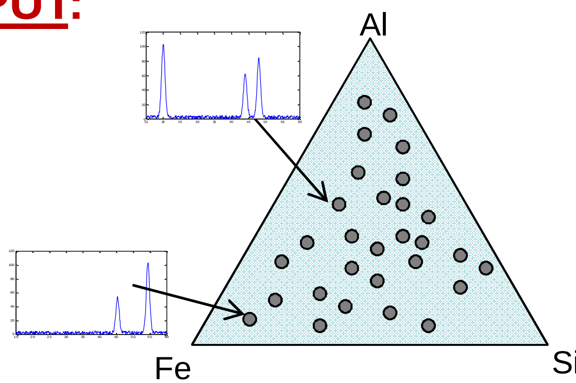


Problem Definition

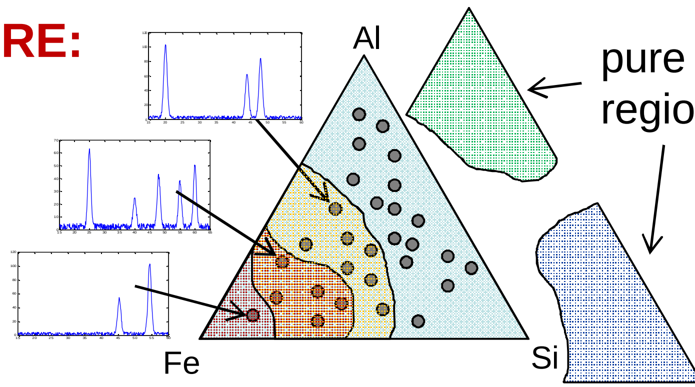


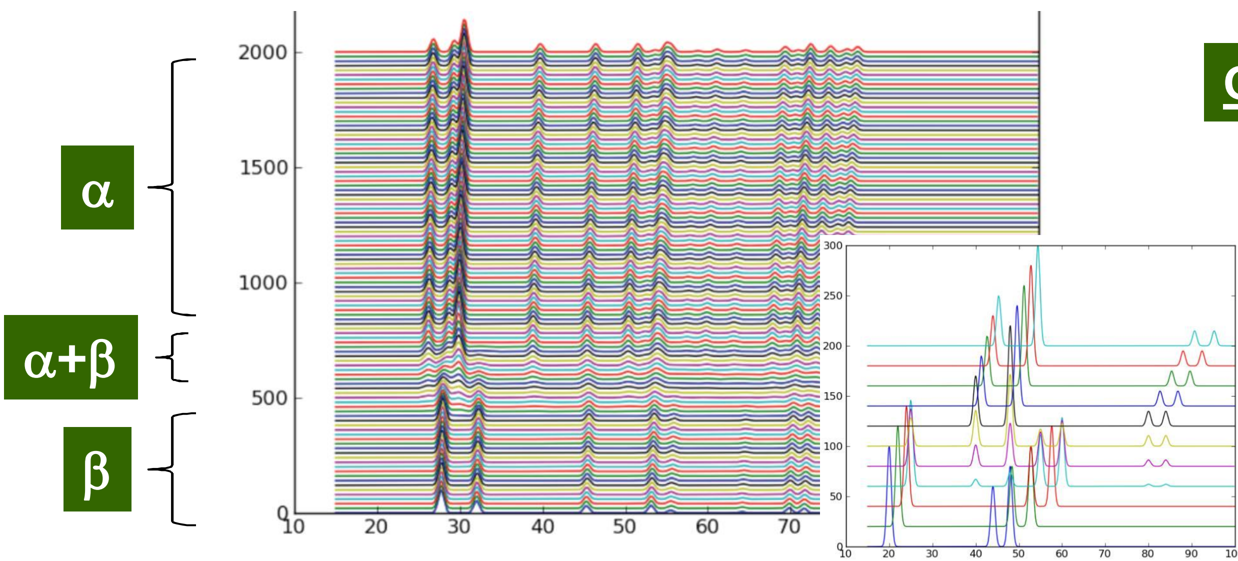
The Problem: Labeling Points with “Phase(s)”

INPUT:



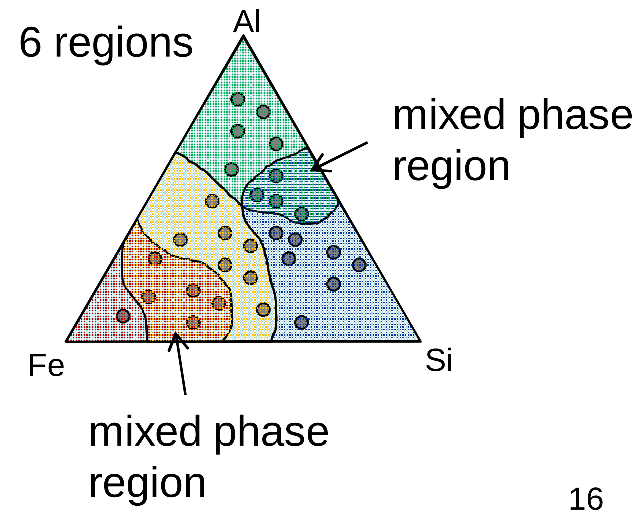
UNDERLYING STRUCTURE:



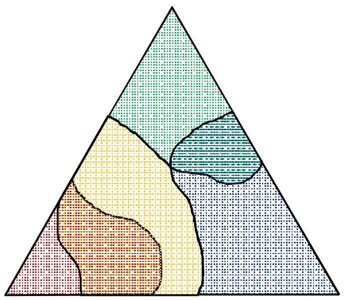


OUTPUT:

NP-hard

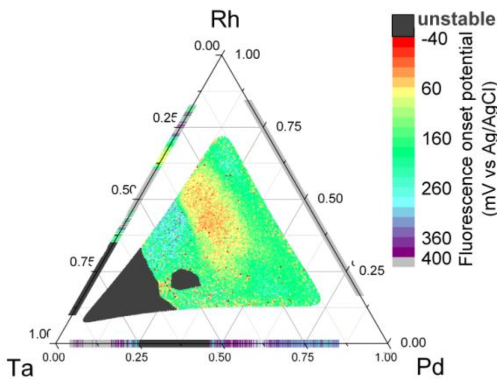


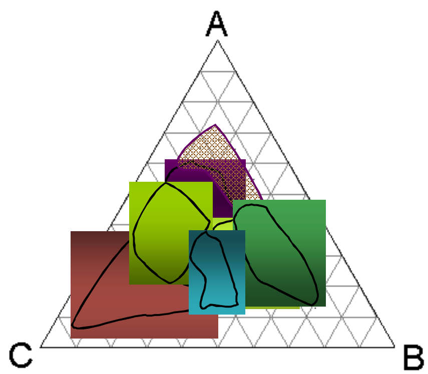
Motivation



Identifying boundaries

Product Substitute, Resource Management…





Identifying new phase regions

Material Property Understanding, Product Substitute…

Ex: Catalysts for fuel cell technology

Automating a laborious manual task

Best data out of expensive experiments…

Outline

- Introduction
  - Problem Definition
  - Motivation
- Key Characteristics, Challenges & Previous work
- Formal Definition & Problem Complexity
- Constraint Programming Model
- Unsupervised Learning
- Integrating both approaches: a new methodology
- Experimental Sample
- Applications with similar structure
- Conclusion

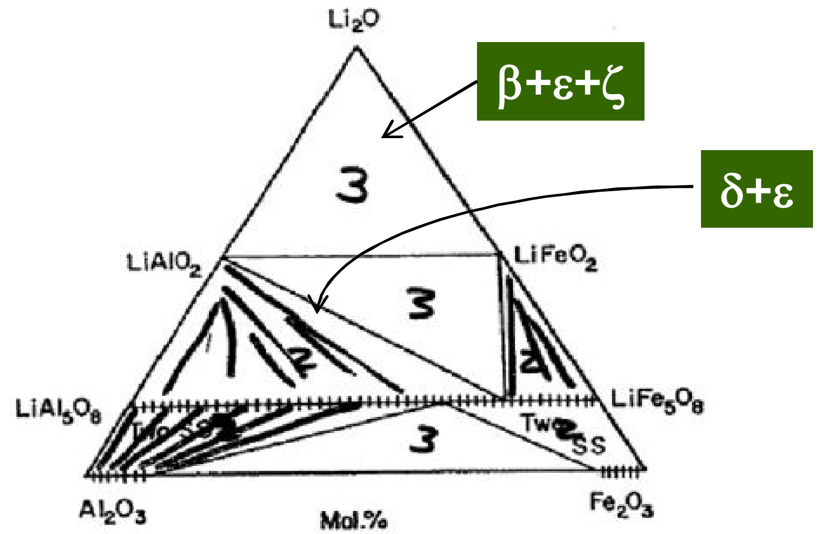
Key Characteristics, Challenges & Previous work

Strong underlying “physics” requirements!

- **Peaks shift within a phase**
- **Intensities fade away**
- **Connectivity**
- **Mixtures of ≤ 3 phases**
- **Small peaks might be discriminative**

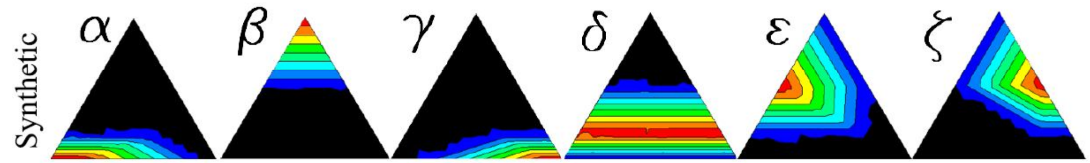
- **Experimentation errors**
- **Large scale**

Previous approaches *unable to model* or enforce these key characteristics!



6 phases

α
β
γ
δ
ε
ζ



Formal Definition & Complexity

The peak location matters ⇨ We discretize the patterns into lists of peaks.

[Formal Definition]

Input: Diffraction patterns $Y_1, \ldots, Y_n$ of n points on the thin-film.

Output: Set of k basis patterns (or *phases*) $X_1, \ldots, X_k$.
Weights $A_1, \ldots, A_n$ and shifts $B_1, \ldots, B_n$ of these basis patterns in the n points.

Theorem: This problem is NP-complete.

Proof: Reduction from the *Normal Set Basis Problem* (which is itself reduced from the *Vertex Cover Problem*).

Outline

- Introduction
- Key Characteristics, Challenges & Previous work
- Formal Definition & Problem Complexity
- Constraint Programming Model
- Unsupervised Learning
    - Global Alignment Kernel
    - K-means clustering
- Integrating both approaches: a new methodology
- Applications with similar structure
- Experimental Results
- Conclusion

Constraint Programming Model

Variables	Description	Type
p_{ki}	Normalizing peak for phase k in pattern c_i	Decision
a_{ki}	Whether phase k is present in pattern c_i	Auxiliary
q_k	Set of normalized peak locations of phase k	Auxiliary

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

$$1 \le \sum_{s=1}^{K} a_{si} \le 3 \qquad \forall\, 1 \le i \le n \tag{2}$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow q_k \subseteq r_{ij} \qquad \forall\, 1 \le k \le K, 1 \le i \le n, 1 \le j \le |c_i| \tag{3}$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow r_{ij} \subseteq q_k \qquad \forall\, 1 \le k \le K, 1 \le i \le n, 1 \le j \le |c_i| \tag{4}$$

$$P(k,k',i,j,j') \rightarrow \begin{cases} member(r_{ij}[j''], q_k) \\ \quad \vee \\ member(r_{ij'}[j''], q_{k'}) \end{cases} \qquad \forall\, 1 \le k < k' \le K, 1 \le i \le n, 1 \le j, j', j'' \le |c_i| \tag{5}$$

where $P(k,k',i,j,j')$ is the proposition: $p_{ki} = j \wedge p_{k'i} = j' \wedge \sum_{s=1}^{K} a_{si} = 2$.

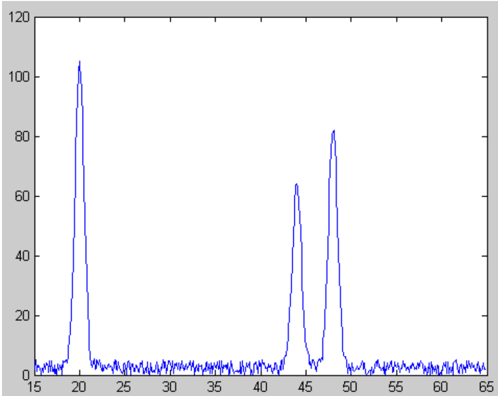
$$p_{ki} = j \rightarrow p_{ki'} \neq j' \qquad \forall\, 1 \le k \le K, (i,j,i',j') \in \Phi \tag{6}$$

$$phaseConnectivity(\{a_{ki} | 1 \le i \le n\}) \qquad \forall\, 1 \le k \le K \tag{7}$$

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

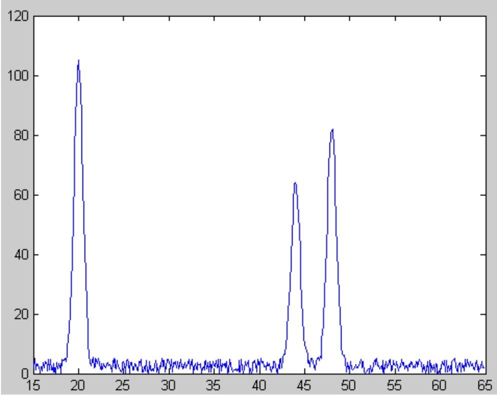
Unsupervised learning: Kernel

Set of features: D =

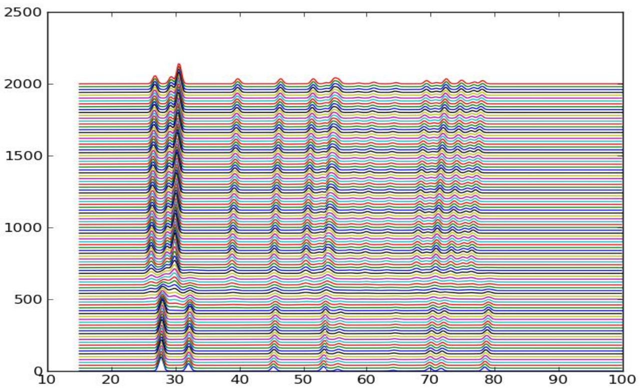


Unsupervised learning: Kernel

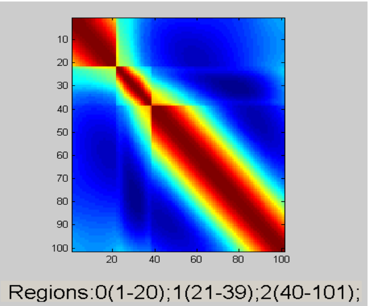
Set of features: D =



Similarity matrix:

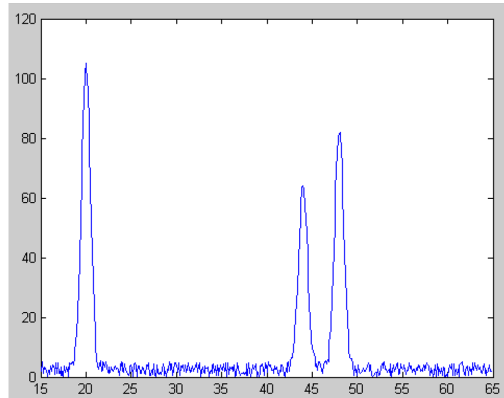


$[D.D^T]$

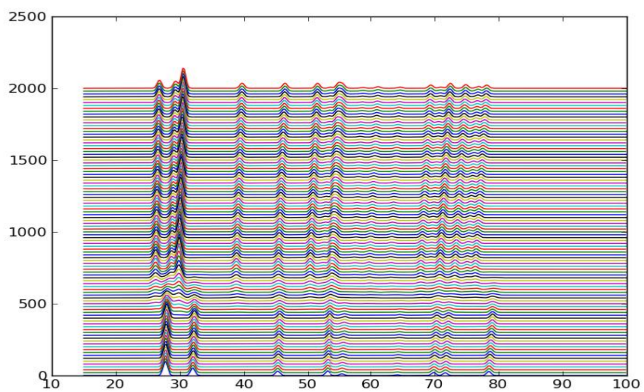


Unsupervised learning: Kernel

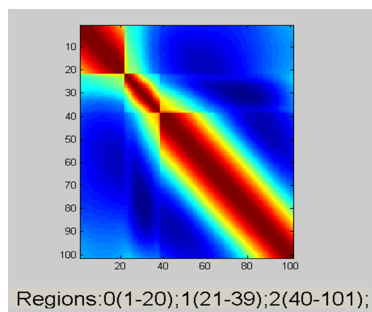
Set of features: D =



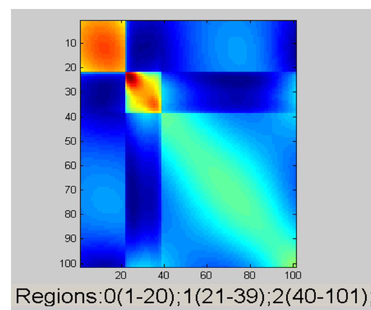
Similarity matrix:



[D.D^T]

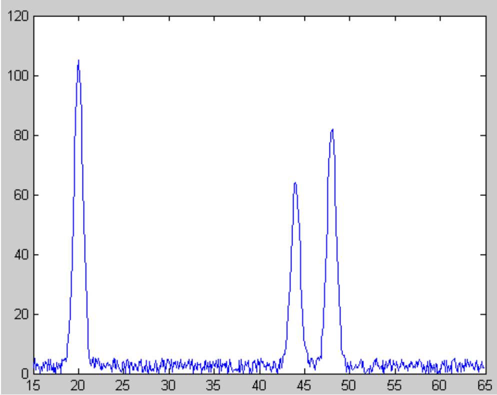


[(D D+s$_1$ D+s$_2$).(D D D)^T]

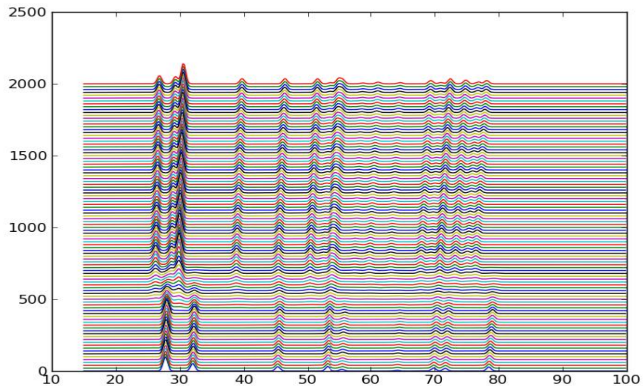


Unsupervised learning: Kernel

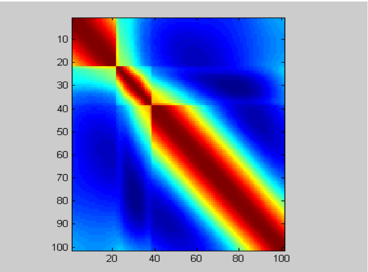
Set of features: D =



Similarity matrix:

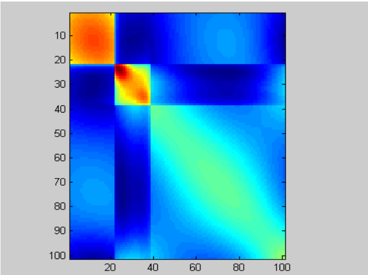


$[D.D^T]$



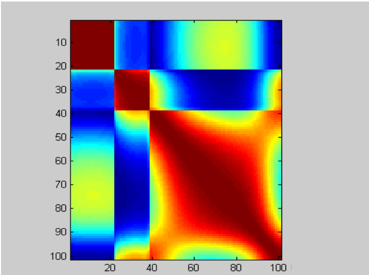
Regions:0(1-20);1(21-39);2(40-101);

$[(D\ D{+}s_1\ D{+}s_2).(D\ D\ D)^T]{=}M$



Regions:0(1-20);1(21-39);2(40-101);

$[M.M^T]$



Regions:0(1-20);1(21-39);2(40-101);

Unsupervised learning: K-means

Purpose

The goal is to select groups of samples that belong to the same phase region and then run the CP approach on this subset, in order to extract the underlying phases of this sub-problem.

Parameter setting

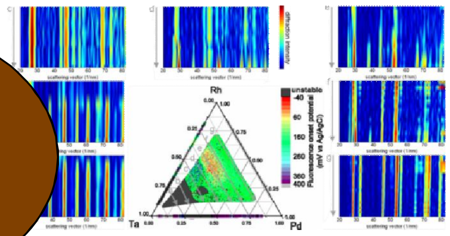
As the number of phase regions is a hidden parameter, we over-segment the kernel by choosing a large number of clusters.

Outline

- Introduction
- Key Characteristics, Challenges & Previous work
- Formal Definition & Problem Complexity
- Constraint Programming Model
- Unsupervised Learning
- Integrating both approaches: a new methodology
- Experimental Sample
- Applications with similar structure
- Conclusion

What's New: Solving it "Properly" Requires…

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



Underlying <u>Physics</u>

A language for <u>Constraints</u> enforcing "local details"

<u>Machine Learning</u> for a "global data-driven view"

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

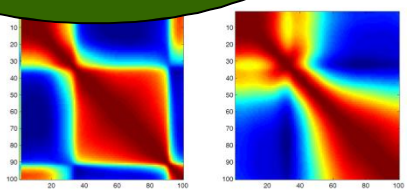
$$1 \le \sum_{s=1}^{K} a_{si} \le 3 \qquad \forall 1 \le i \le n$$

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

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

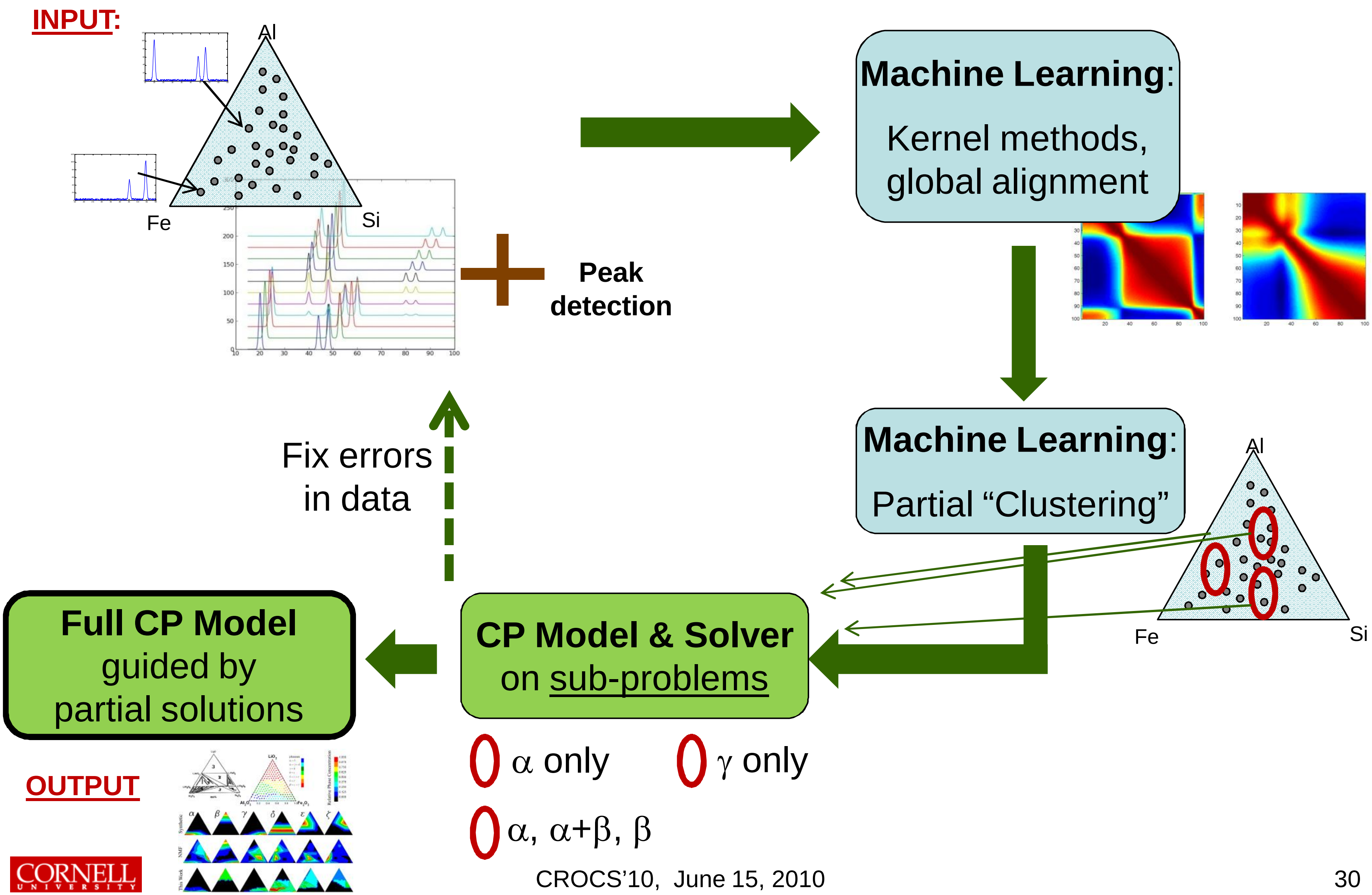
$$P(k,k',i,j,j') \rightarrow \begin{cases} member(r_{ij}[j''], q_k) \\ \vee \\ member(r_{ij'}[j''], q_{k'}) \end{cases} \qquad \forall 1 \le k < k$$

Constraint Programming model



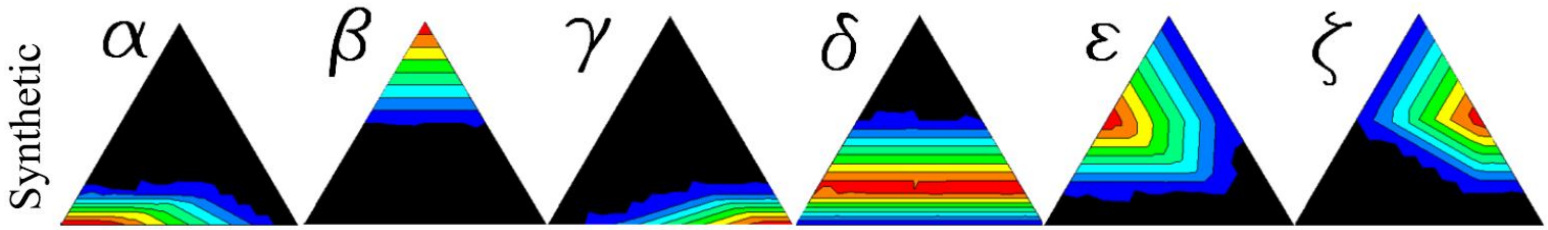
Similarity "Kernels" & Clustering

Bridging Constraint Reasoning and Machine Learning: Overview of the Methodology



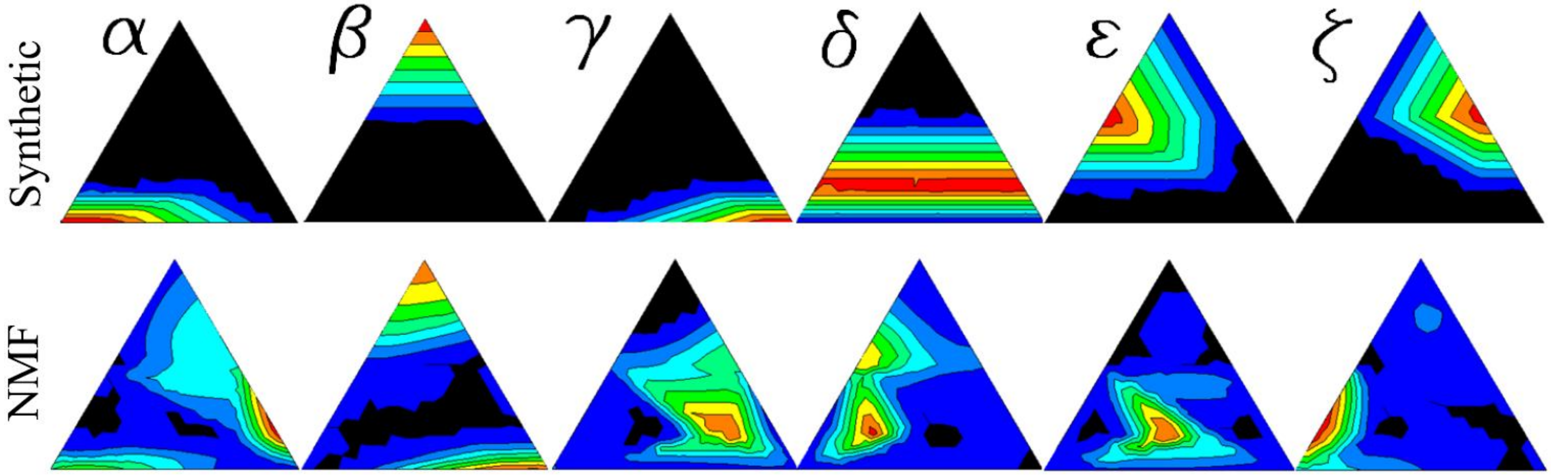
Experimental Sample

Example on Al-Li-Fe diagram:



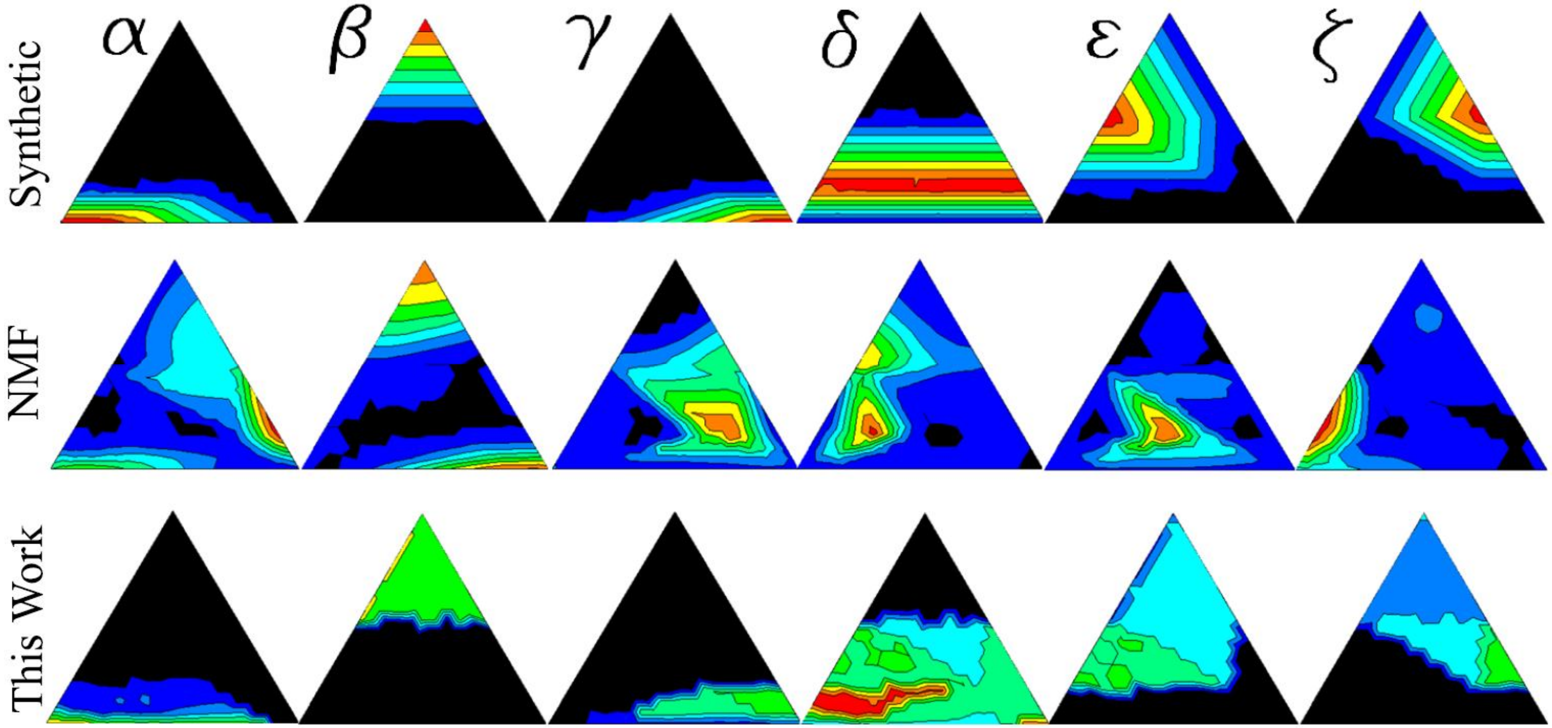
Experimental Sample

Example on Al-Li-Fe diagram:



Experimental Sample

Example on Al-Li-Fe diagram:



Outline

- Introduction
- Key Characteristics, Challenges & Previous work
- Formal Definition & Problem Complexity
- Constraint Programming Model
- Unsupervised Learning
- Integrating both approaches: a new methodology
- Experimental Sample
- Applications with similar structure
- Conclusion

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…

Conclusion

An exciting new problem!

- **Close collaboration with Physicists**

Sustainability impact:

- **Technologies for fuel cell design**
- **Best data out of "expensive" experiment!**

Computer science impact:

New problems at the intersection of constraint reasoning & machine learning

→ clustering under hard & soft constraints (imposed by underlying physics)

Ongoing project