SMT-AIDED COMBINATORIAL MATERIALS DISCOVERY

Stefano Ermon	Computer Science
Ronan Le Bras	Computer Science
Carla P. Gomes	Computer Science
Bart Selman	Computer Science
Bruce van Dover	Materials Science/Physics

June 18, 2012 **SAT 2012**

Motivation

Cornell Fuel Cell Institute

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

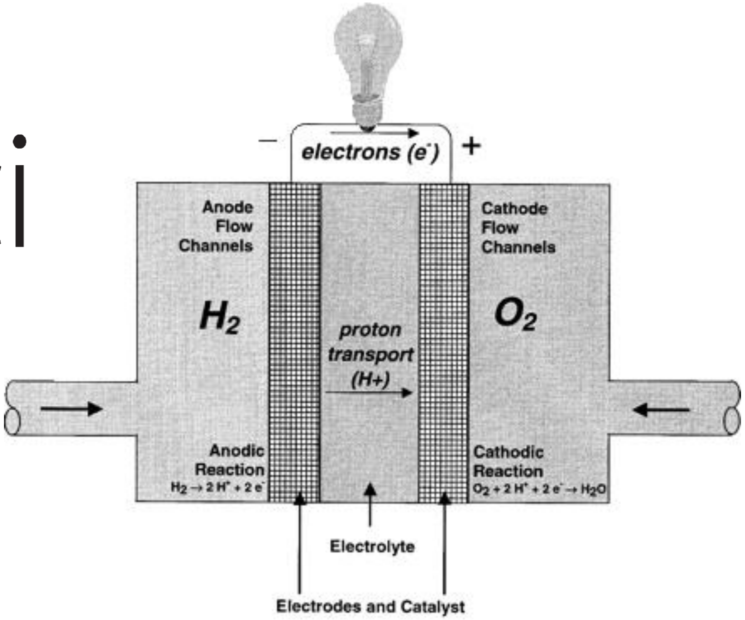


Figure 1. Fuel cell schematic.
Source: Annual Reveiws of Energy and the Environment. http://energy.annualreviews.org/cgi/content/full/24/1/281

An **Electrocatalyst** must:

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

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

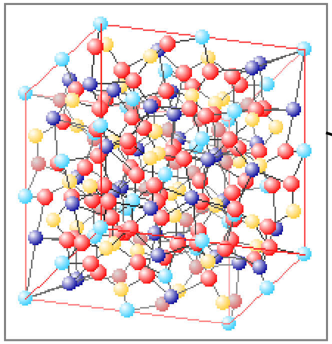
1) The reaction rate is still considered slow (causing **energy loss**)
2) Platinum is fairly **costly**, **intolerant** to fuel **contaminants**, and has a **short lifetime**.

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

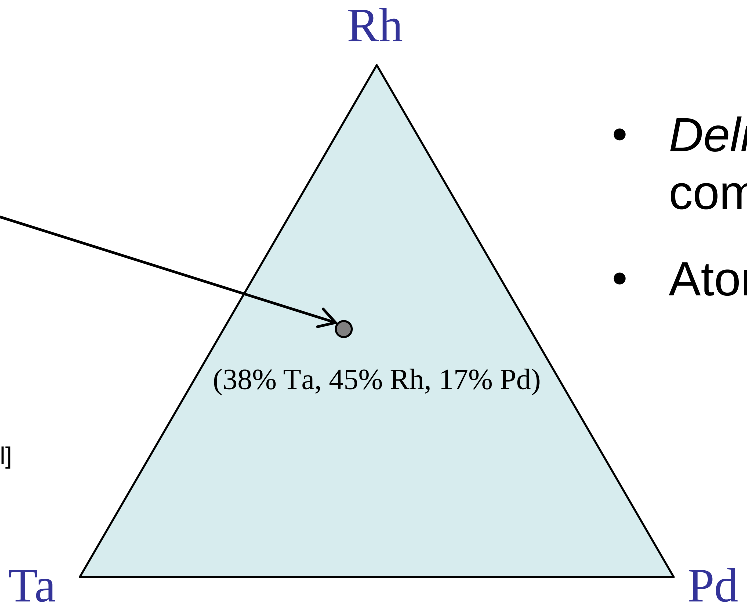
Motivation

<u>*Recipe for finding alternatives to Platinum*</u>

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]

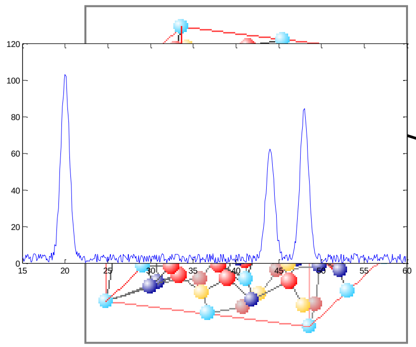


- *Deliberately* inhomogeneous composition on Si wafer
- Atoms are intimately mixed

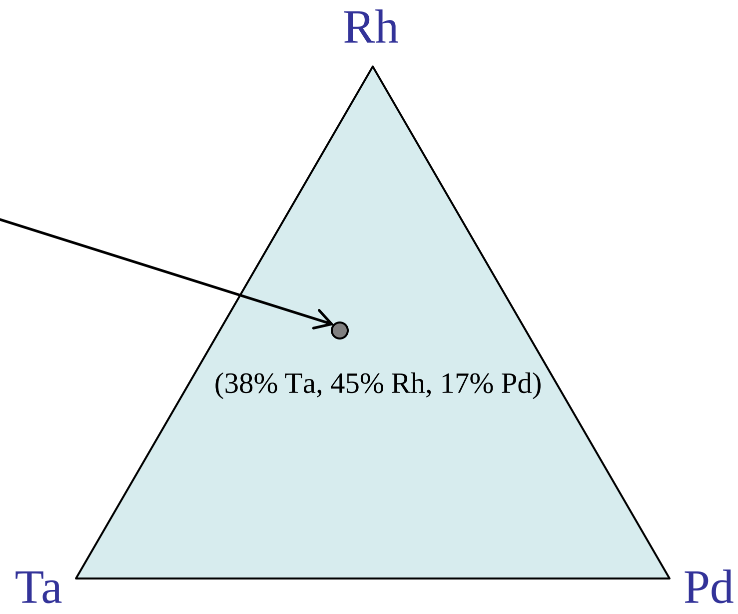
Motivation

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

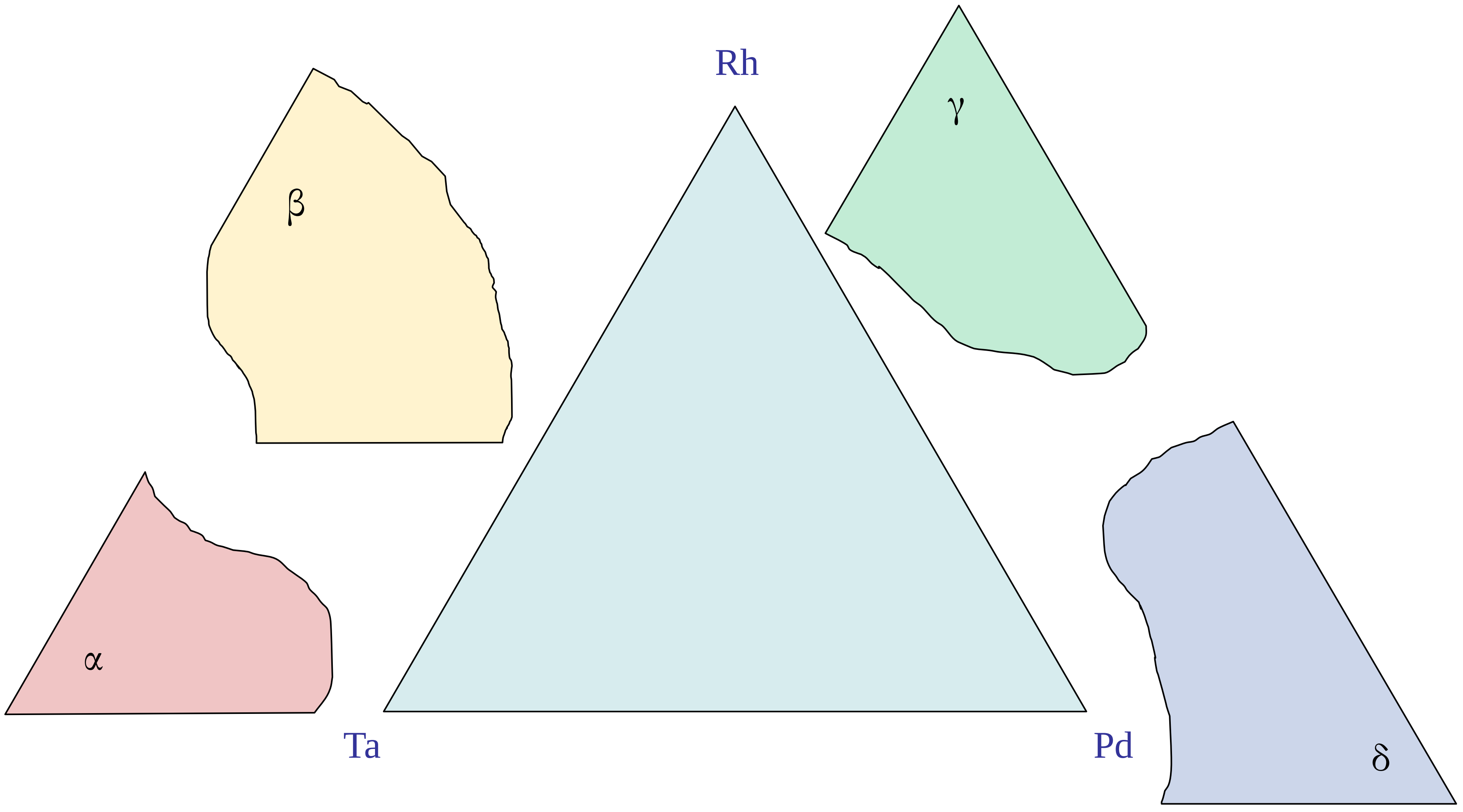
- XRD pattern **characterizes** the underlying **crystal** fairly well
- **Expensive** experimentations: Bruce van Dover's research team has access to the facility **one week every year**.



[Source: *Pyrotope*, Sebastien Merkel]

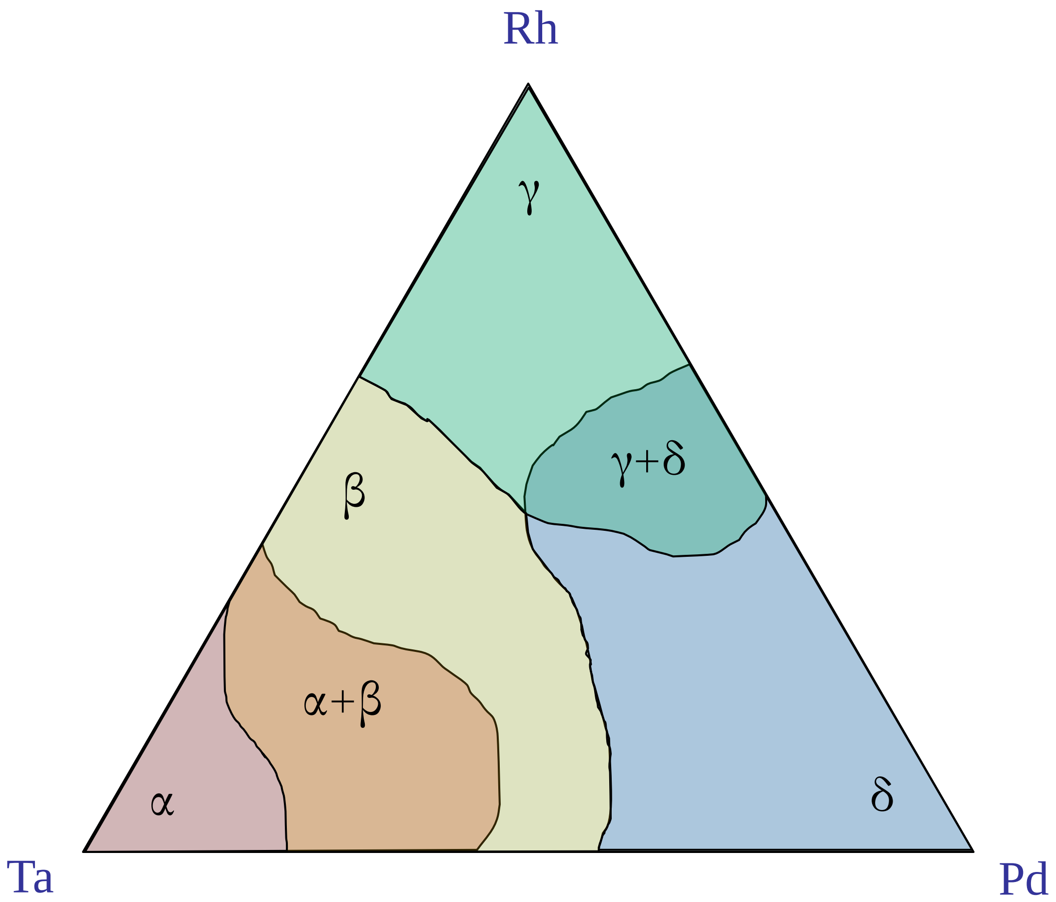


Motivation

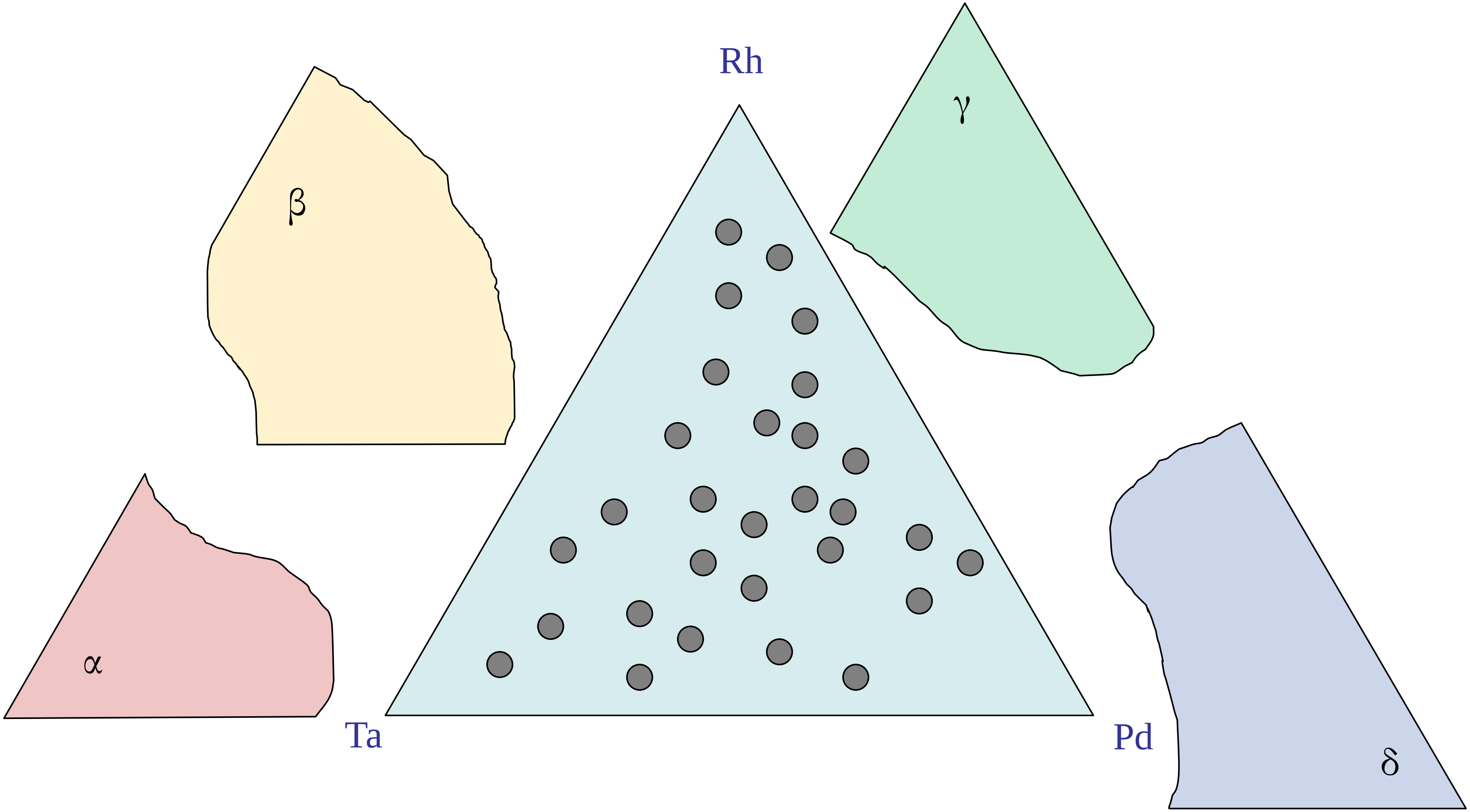


CORNELL UNIVERSITY

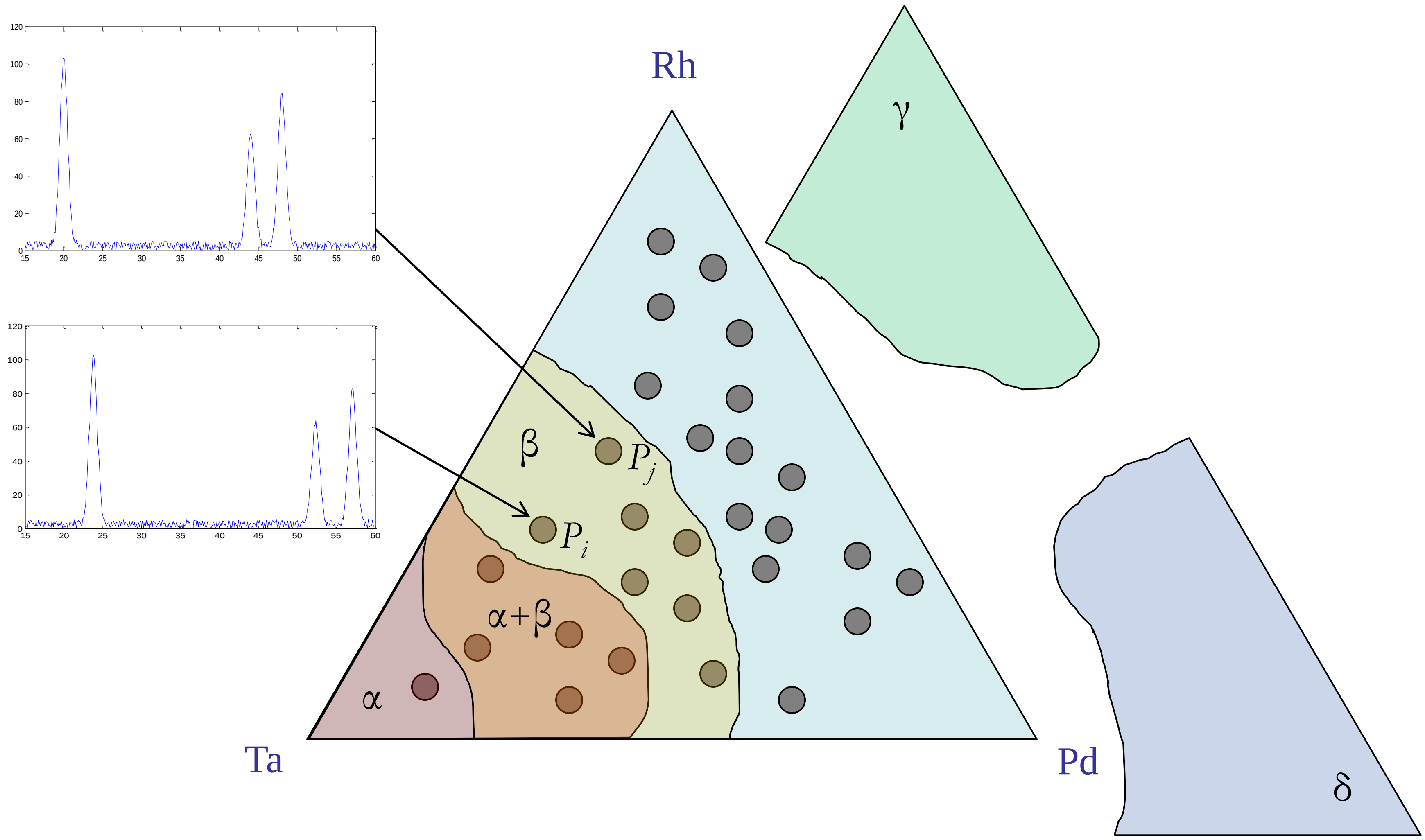
Motivation



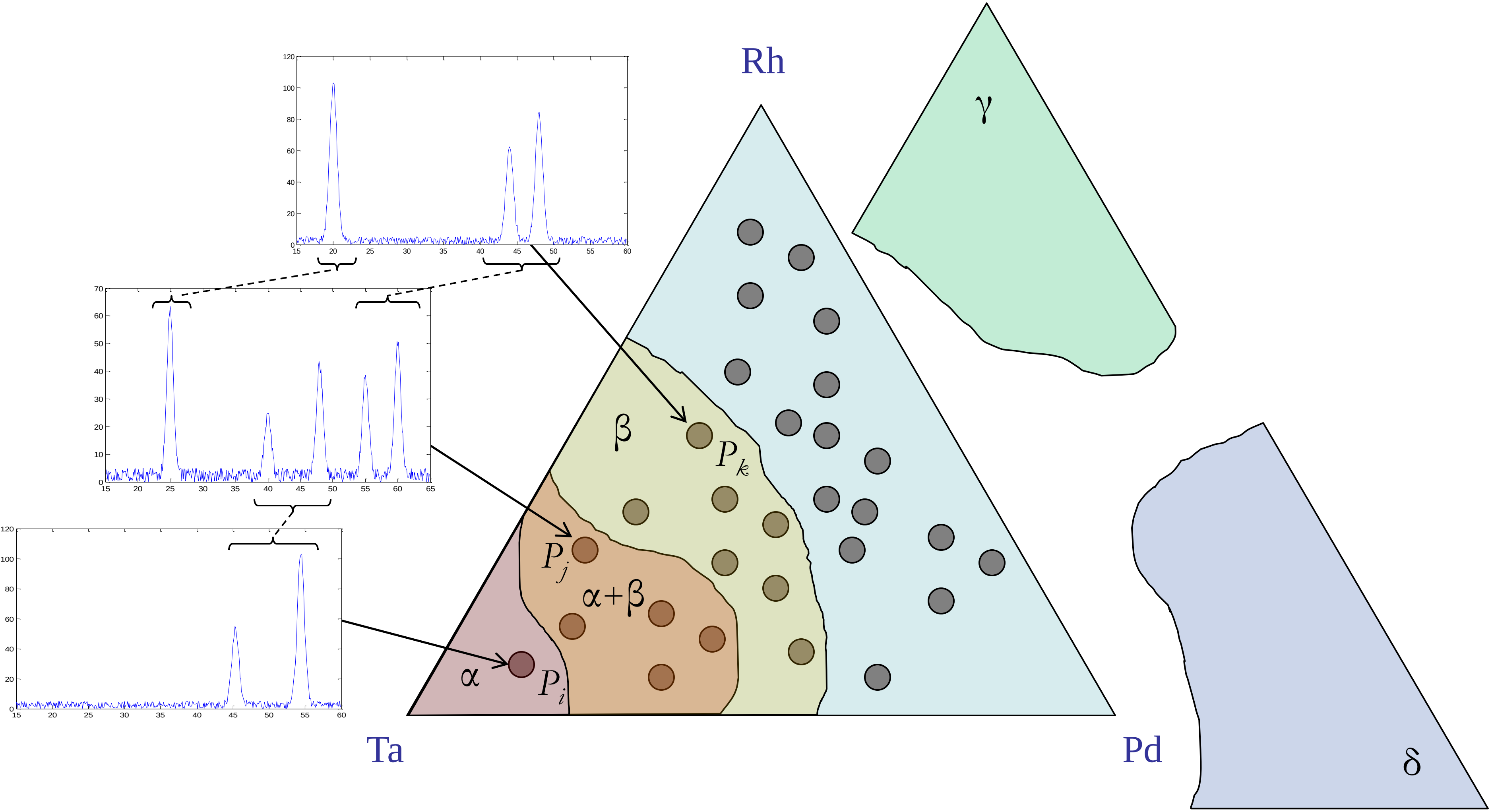
Motivation



Motivation

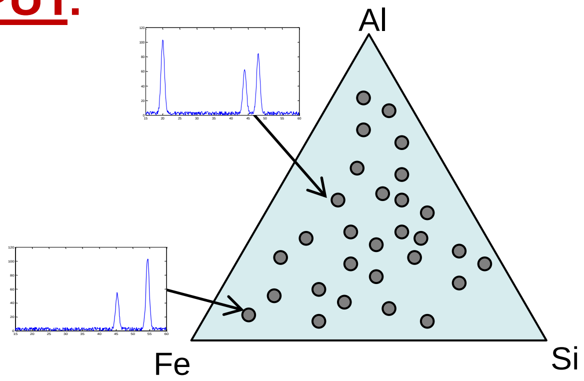


Motivation



Motivation

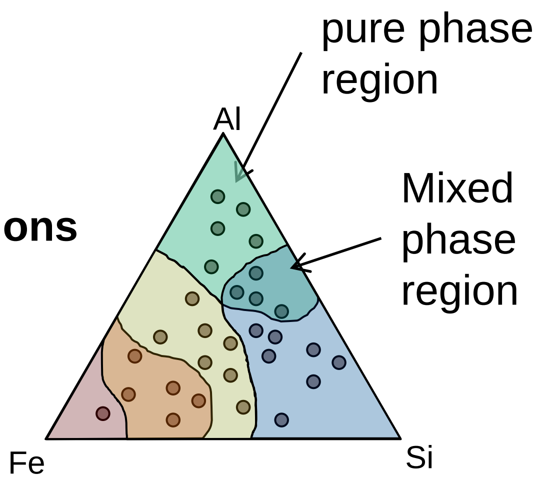
INPUT:



OUTPUT:

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

XRD pattern characterizing pure phases



Additional Physical characteristics:

- Phase Connectivity
- Mixtures of ≤ 3 pure phases
- Peaks shift by ≤ 15% within a region
  - Continuous and Monotonic
- Noisy detection

Motivation

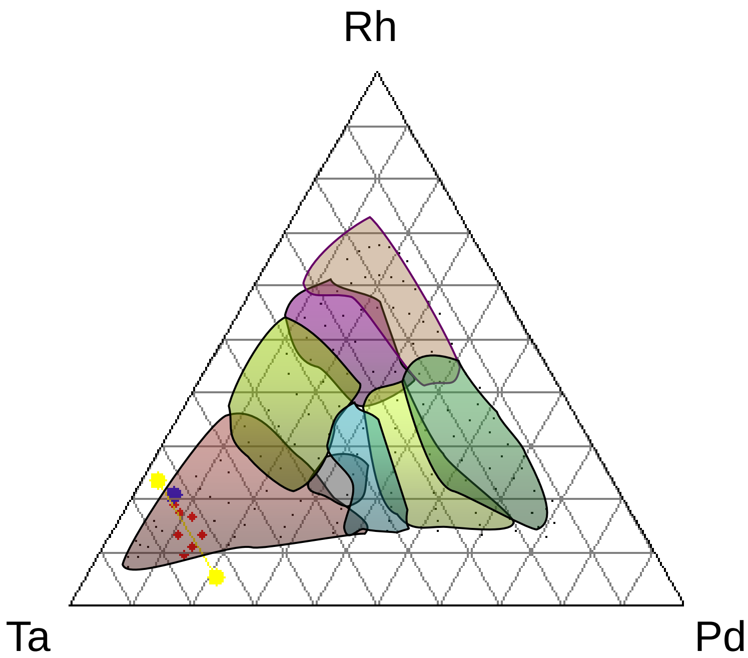


Figure 1: Phase regions of Ta-Rh-Pd

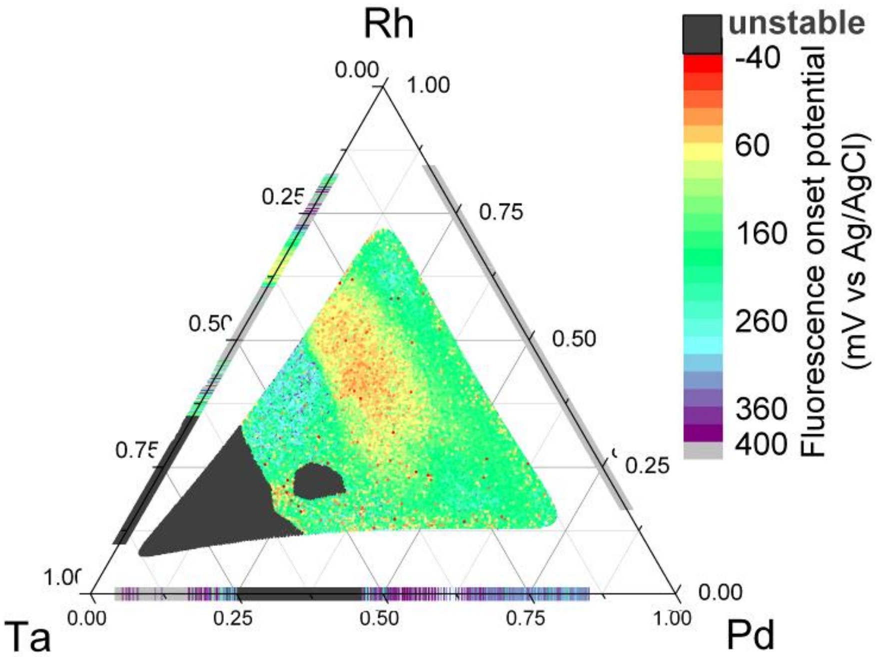
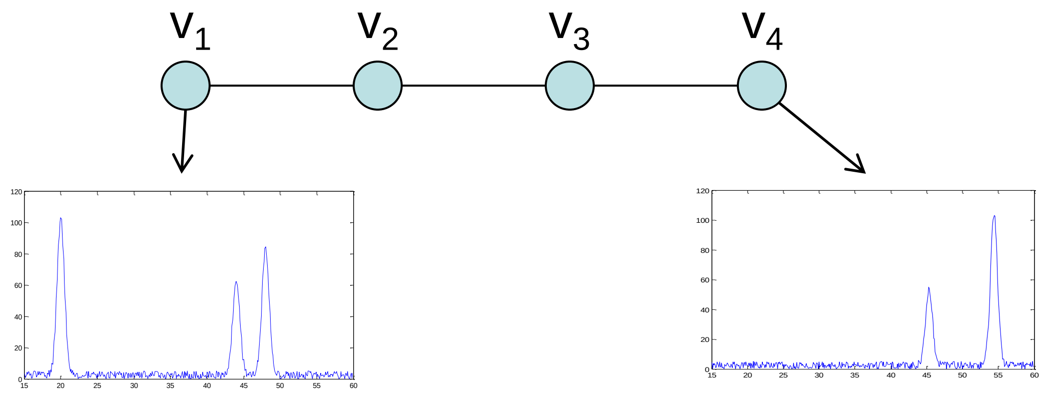


Figure 2: Fluorescence activity of Ta-Rh-Pd

Problem Definition

- **Input:**
  - A graph *G* representing the points on the silicon wafer



  - A real vector $\boldsymbol{D_i}$ per vertex v_i (diffraction patterns)
  - K = user specified number of pure phases
- **Goal**: a basis of K vectors for

$\boldsymbol{D_1}$	$\boldsymbol{D_2}$	…	$\boldsymbol{D_N}$

$$\boldsymbol{D_i} = a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K}$$

Problem Definition

- There is **experimental noise**

$$\boldsymbol{D_i} = a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K} \quad \Rightarrow \quad \boldsymbol{min}\ \|\boldsymbol{D_i} - a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K}\|$$

Minimize norm instead

- **Non-negative** basis vectors and coefficients

$$\boldsymbol{B_i} \geq \boldsymbol{0},\ a_{ij} \geq 0$$

- **At most M (=3) non-zero coefficients** per point

$$|\{j \mid a_{ij} > 0\}| \leq M$$

- Basis patterns appear in **contiguous** locations on silicon wafer

The subgraph induced by $|\{i \mid a_{ij} > 0\}|$ is connected

Problem Definition

- Basis vector can be **shifted**

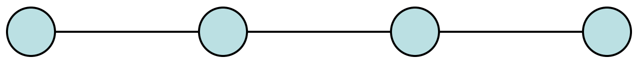
Shift operator

Shift coefficients

$$||\boldsymbol{D_i} - a_{i1}S(\boldsymbol{B_1}, s_{i1}) + \ldots + a_{iK}S(\boldsymbol{B_K}, s_{iK})||$$

- Shifts coefficients are **bounded**, **continuous** and **monotonic**

$$S_{11} \leq S_{12} \leq S_{13} \leq S_{14}$$

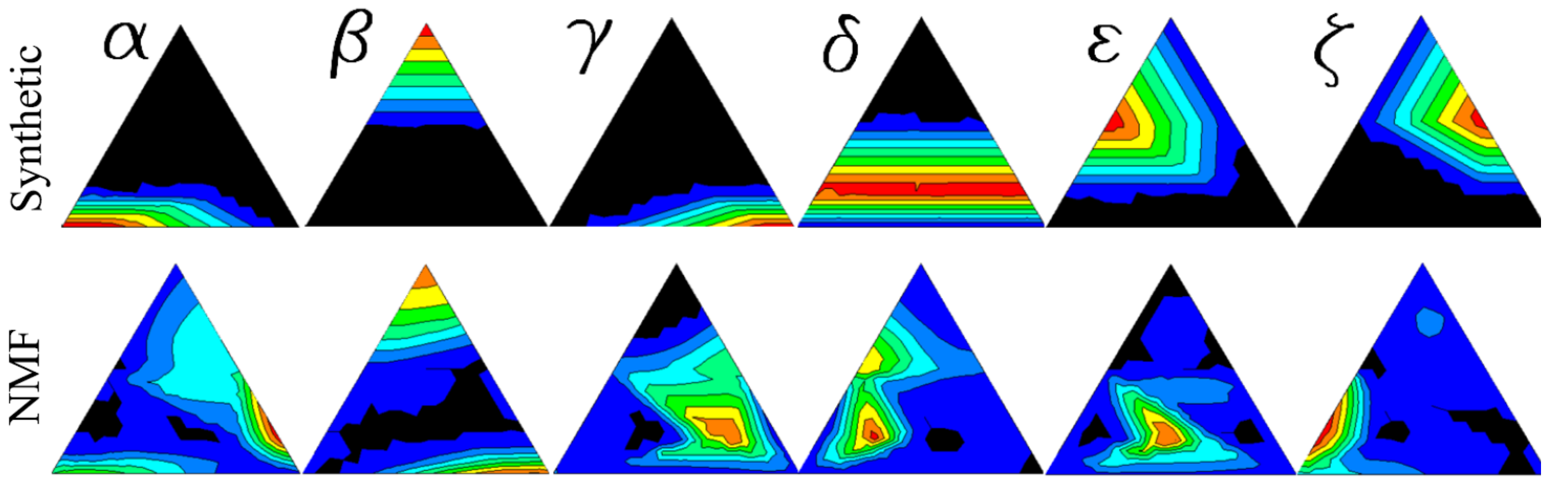


$$|S_{12} - S_{11}| \leq c$$

It is a form of **constrained** Principal Component Analysis (Singular Value Decomposition)

Prior Work/Machine Learning

- Ignore most of the constraints, and shifting
  - **Non-negative matrix factorization**
  - Good scaling
  - Cannot enforce the combinatorial constraints (e.g., connectivity)



[Source:Le Bras et al., 2011]

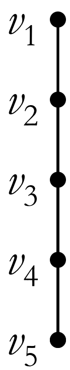
Prior Work / CP

- **Constraint Programming** formulation [Le Bras et al., CP 2011]

  Pattern Decomposition with Scaling:
  - Imitate what humans do. Instead of considering full spectra, **focus on the peaks**.
  - Encoding based on set-variables
  - Does not scale to realistic sized problems
  - Useful in combination with clustering-based heuristic

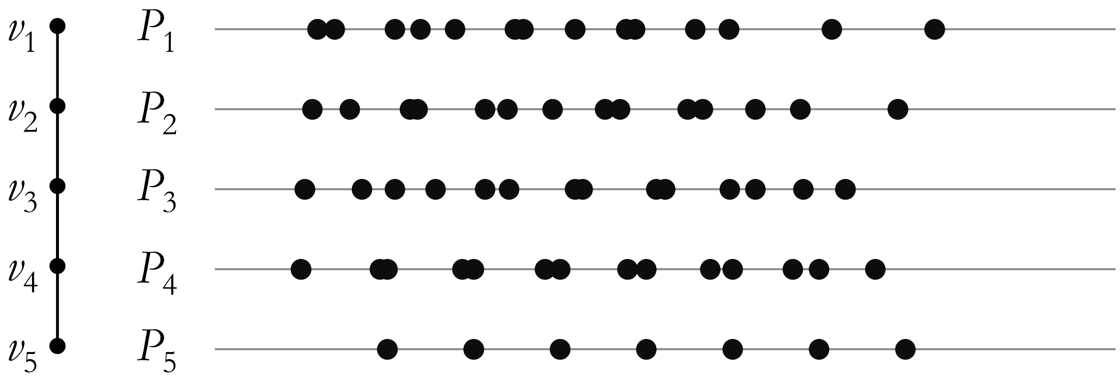
Our Approach

- **Arithmetic based approach** (SMT):
  - Initial graph *G* representing points on the wafer



Our Approach

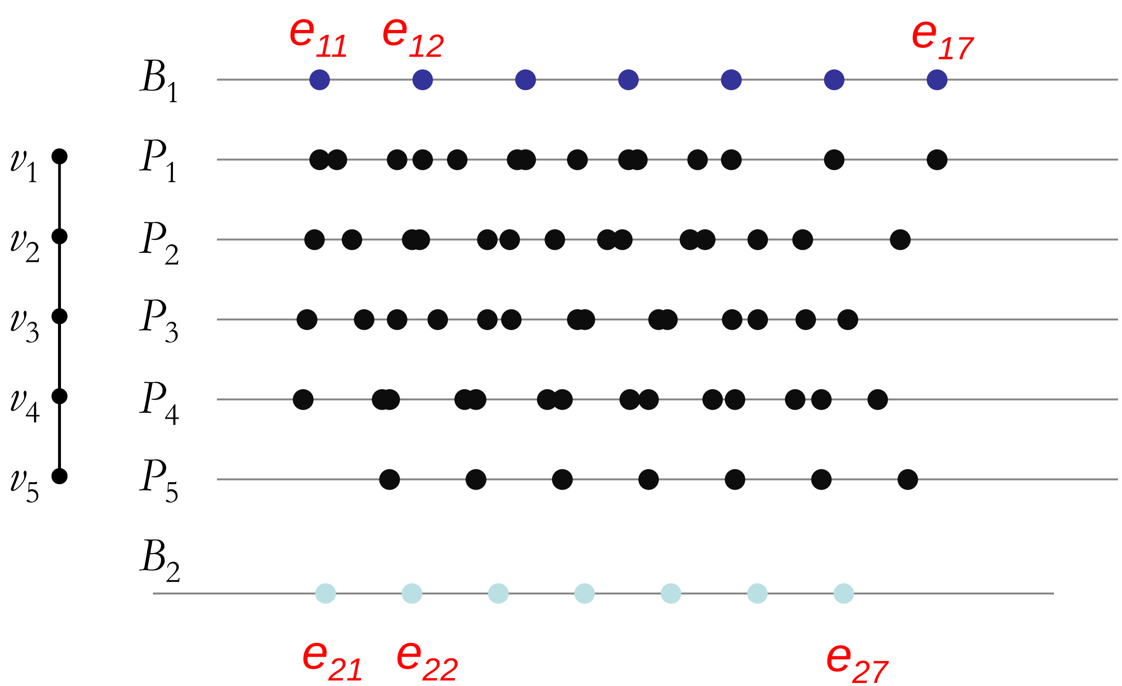
- **Arithmetic based approach** (SMT):
  - Initial graph *G* representing points on the wafer
  - *Peak detection to extract a set of peaks P_i for each diffraction pattern* $\boldsymbol{D_i}$



Our Approach

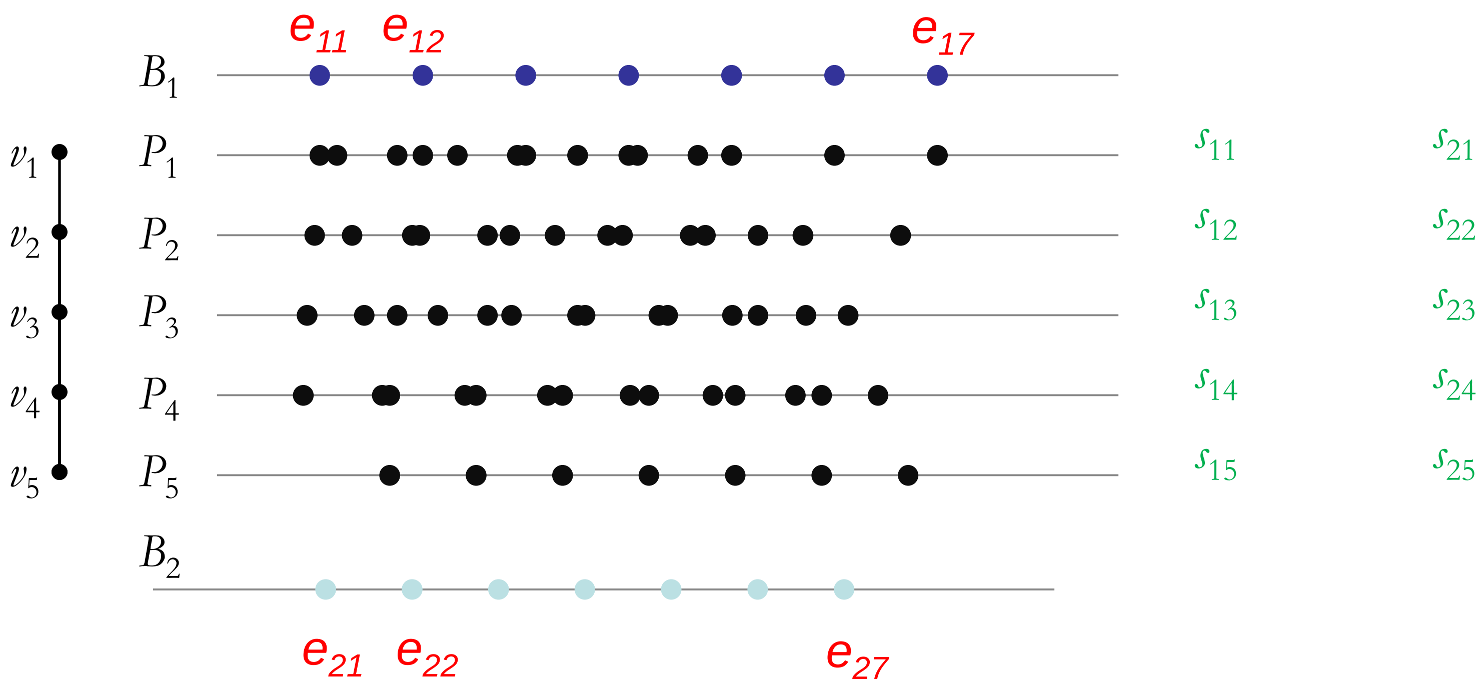
- **Arithmetic based approach** (SMT):
  - Initial graph G
  - *Peak detection to extract a set of peaks P_i for each diffraction pattern* $\mathbf{D_i}$
  - Real variables e_{ij} for the **peak locations** in each B_i

K=2
basis
patterns



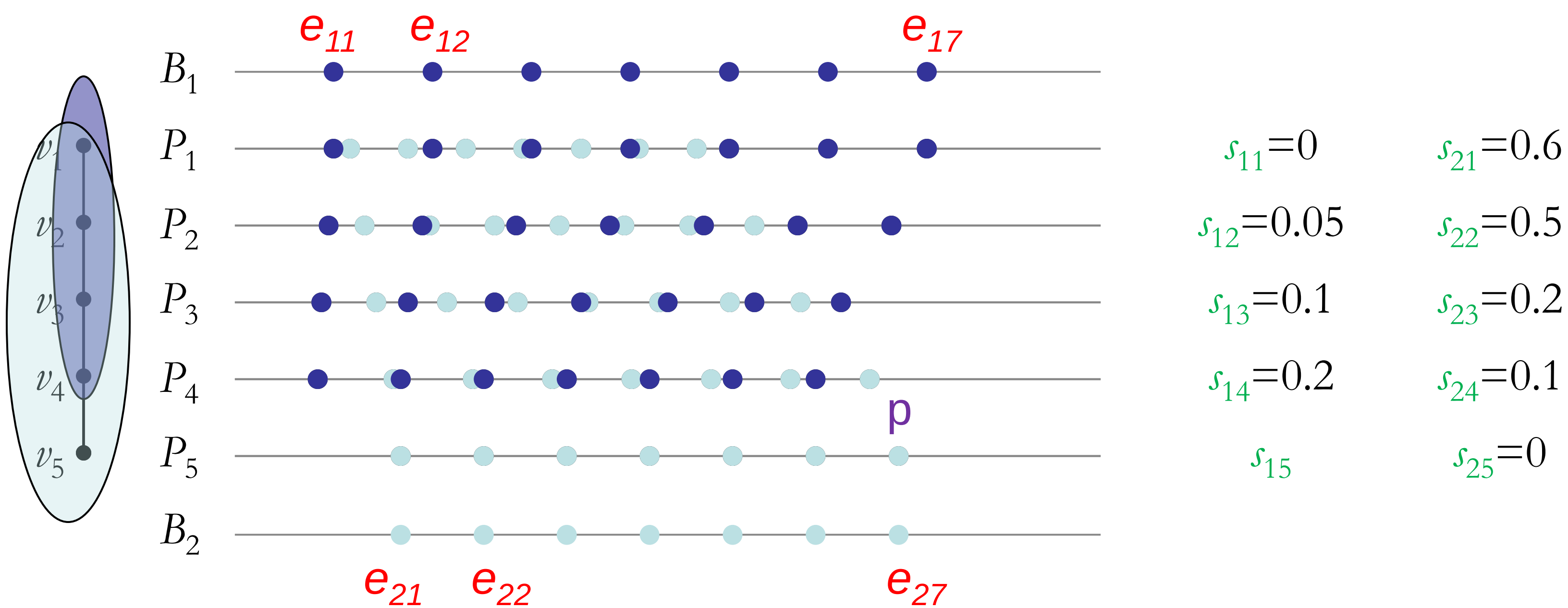
Our Approach

- **Arithmetic based approach** (SMT):
  - Real variables e_{ij} for the **peak locations** in each B_i
  - Real variables for the shift coefficients s_{ij}



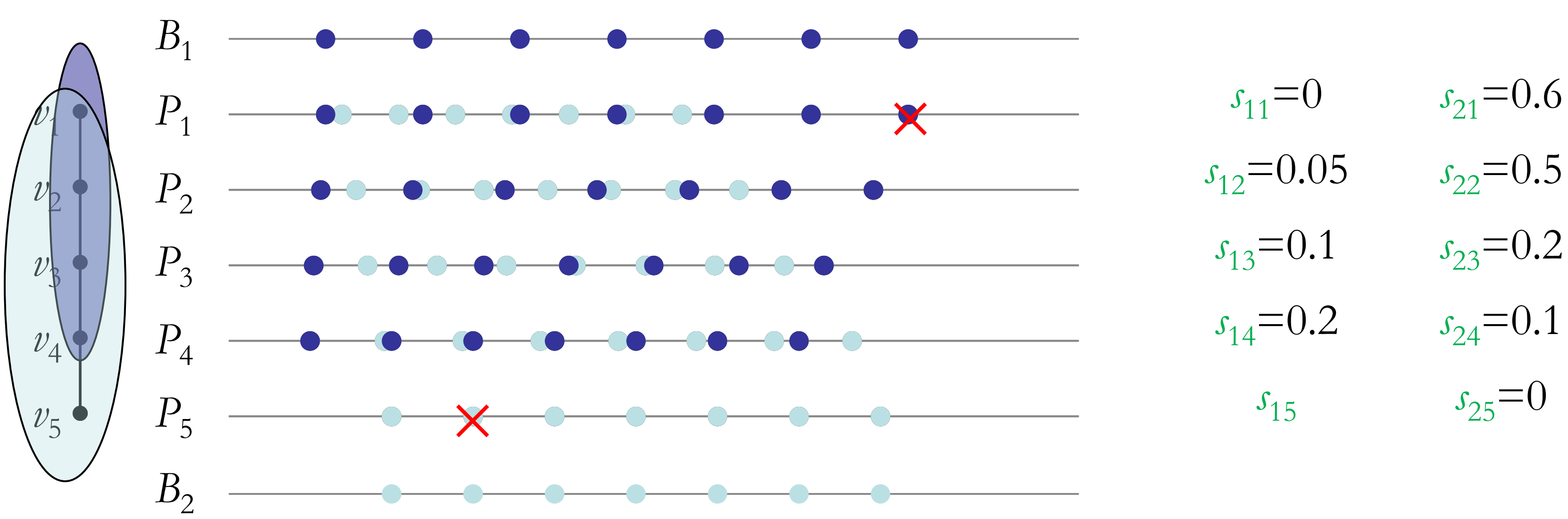
Our Approach

- **Arithmetic based approach** (SMT):
  - Real variables e_{ij} for the **peak locations** in each B_i
  - Real variables for the shift coefficients s_{ij}
  - An observed peak p is "*explained*" if there exists s_{ij}, e_{il} s.t. $|p-(s_{ij} + e_{il})| \leq \varepsilon$



Our approach

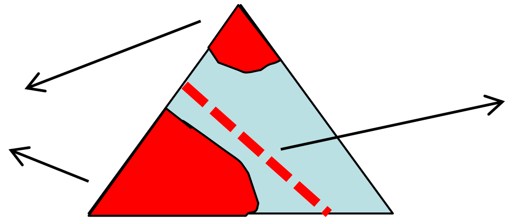
- **Arithmetic based approach** (SMT):
  - Every observed peak must be "*explained*"
  - Bound the number of missing peaks ≤ *T*
  - Minimization by (binary) search on *T*



SMT formulation (continued)

- Arithmetic-based SMT encoding:
  - Linear phase usage constraint (up to M basis patterns per point)
  - Linear constraint for shift monotonicity and continuity ($s_{ij} \leq s_{lm}$)
  - **Lazy connectivity**: add a cut if current solution is not connected

If disconnected regions explained with phase 1



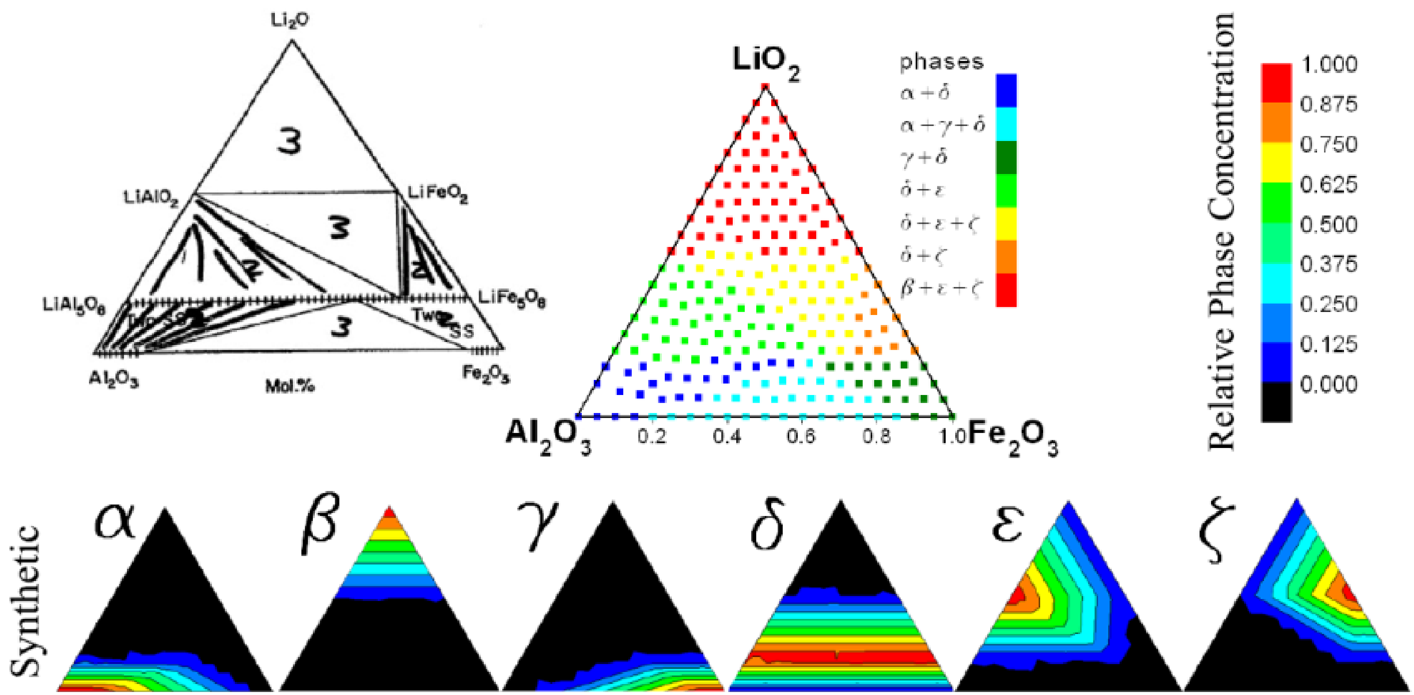
Then Phase 1 must appear in at least one of these points

  - **Symmetry breaking**:
    - Renaming of pure phases
    - Order of the peaks location e_{ij} (per basis pattern)

Quantifier-free linear arithmetic

Experimental Results

- Use synthetic instances from the Al-Li-Fe ternary system
  - Known ground truth
  - Fairly complex system



Runtime

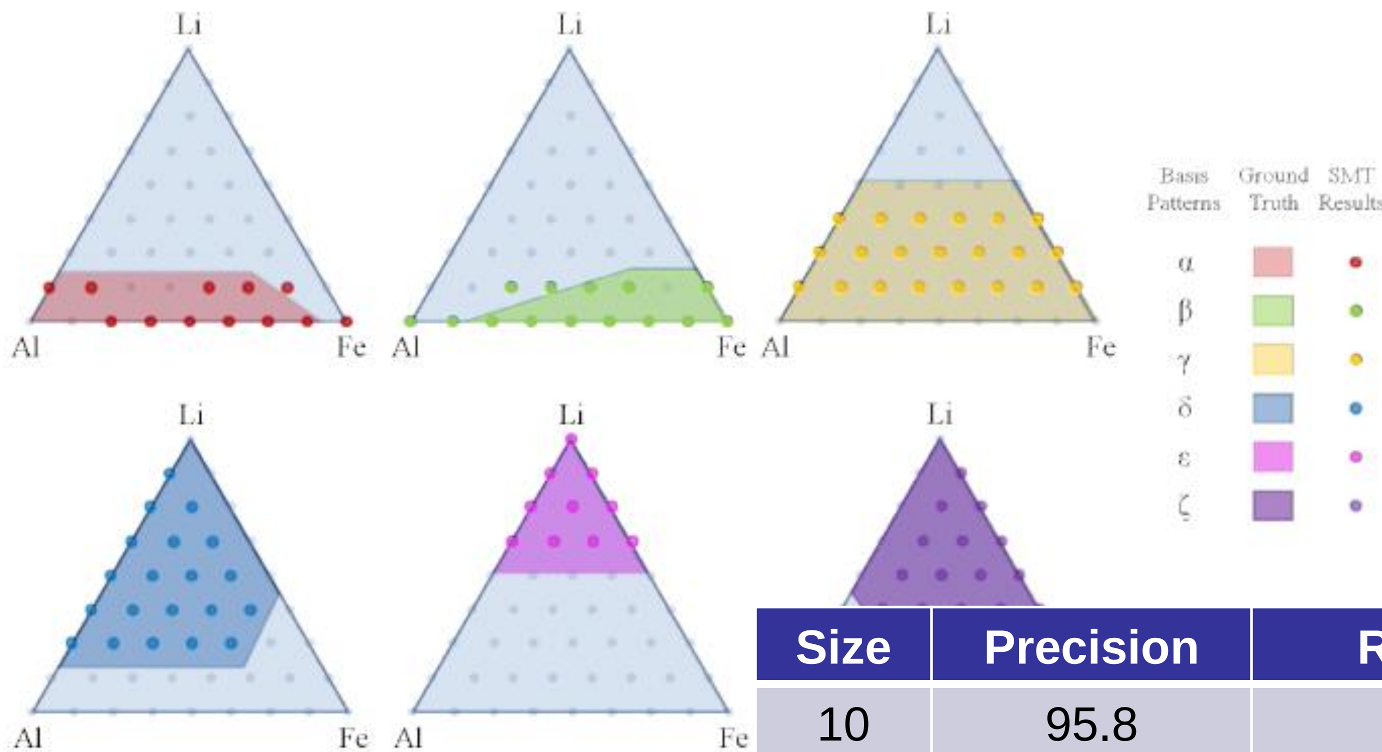
# Points	Unknown Phases	Arithmetic + Z3 (s)	Set-based + CPLEX (s)
10	3	8	**0.5**
	6	**12**	Timeout
15	3	13	**0.5**
	6	**20**	Timeout
18	3	**29**	384.8
	6	**125**	Timeout
29	3	**78**	276
	6	**186**	Timeout
45	6	**518**	Timeout

Z3 scales to realistic sized problems!

Arithmetic encoding translated to CP and MIP:

- MIP is appealing because it can optimize the objective
- They don't scale → SMT solving strategy

Precision/Recall



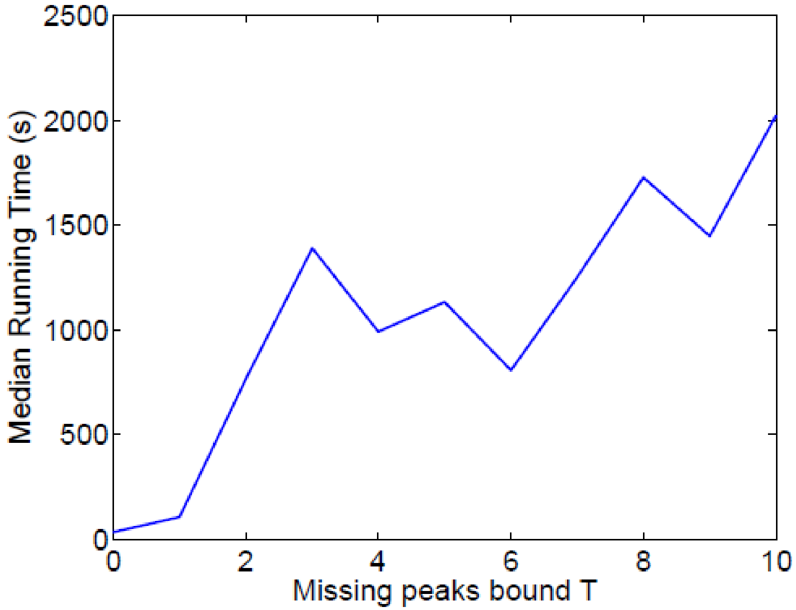
Recovers ground truth

Size	Precision	Recall
10	95.8	100
15	96.6	100
18	97.2	96.6
29	96.1	92.8
45	95.8	91.6

Robustness

- Remove some peaks to simulate experimental noise
- Size = 15 points

Missing Peaks	Precision	Recall
1	96.1	99.6
2	96.3	99.3
3	96.7	99.5
4	95.3	98.9
5	94.8	99.7



Solutions are still **accurate**. Runtime increases approx **linearly**.

Conclusion

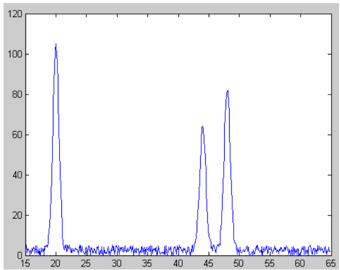
- New **arithmetic-based** encoding for Materials Discovery
- Good performance on synthetic data:
  - Scales to **realistic sized problems** (~50 points)
  - SMT **outperforms previous** one based on set-variables
  - Good accuracy (>90% precision and recall)
  - (likely) due to SMT solving procedure
- Exciting results analyzing and explaining **real-world data**

THANK YOU!

Extra slides

Previous Work 1: Cluster Analysis [Long et al., 2007]

$x_i =$



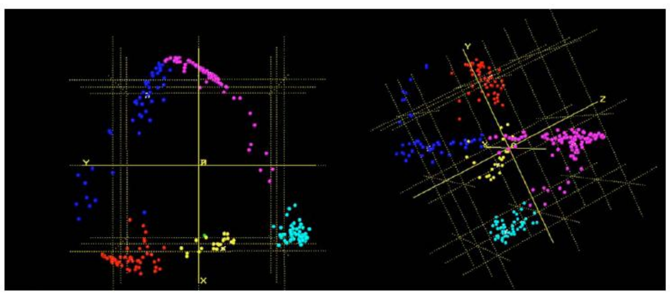
Feature vector

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

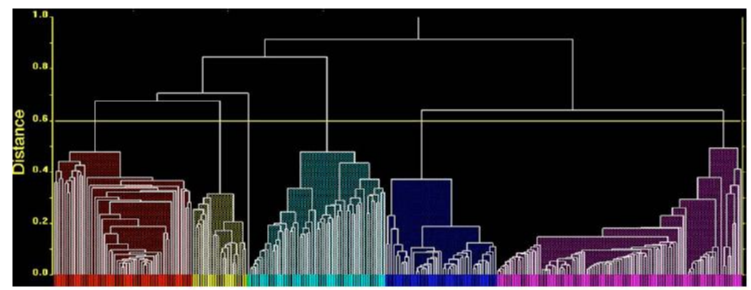
Pearson correlation coefficients

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

Distance matrix



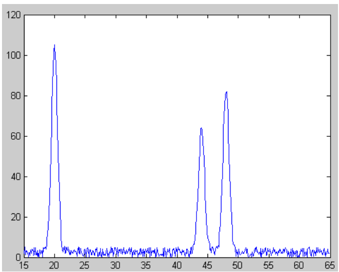
PCA – 3 dimensional approx



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]

$x_i =$ 

Feature vector

$\longrightarrow$ $$X = A.S + E$$

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

$\longrightarrow$ $$Min \ \|E\|$$

Minimizing squared Frobenius norm

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

SMT formulation

- **Parameters**
  - Number of pure phases *K*, tolerance *ε*
- Key components
  - Variables peak positions per base
  - Shifts per point
  - Point *p* is explained by base *k*

SMT formulation

- New arithmetic-based encoding:
  - Real variables e_{ij} for the peak locations in each B_i
  - Real variables for the shift coefficients s_{ij} (per base, per point)
  - An observed peak p is explained if $|p-s_{ij} - e_{ij}| \leq \varepsilon$ (Match the height of the peaks)
  - Bound the number of missing peaks $\leq T$