Materials Discovery for Fuel Cells

—————

New Opportunities at the Intersection of Constraint Reasoning and Learning

Ronan Le Bras	Computer Science
Stefano Ermon	Computer Science
Theodoros Damoulas	Computer Science
Rich Bernstein	Computer Science
Carla P. Gomes	Computer Science
Bart Selman	Computer Science
R. Bruce van Dover	Materials Science/Physics

July 5, 2012 **CompSust'12**

Motivation

Cornell Fuel Cell Institute

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

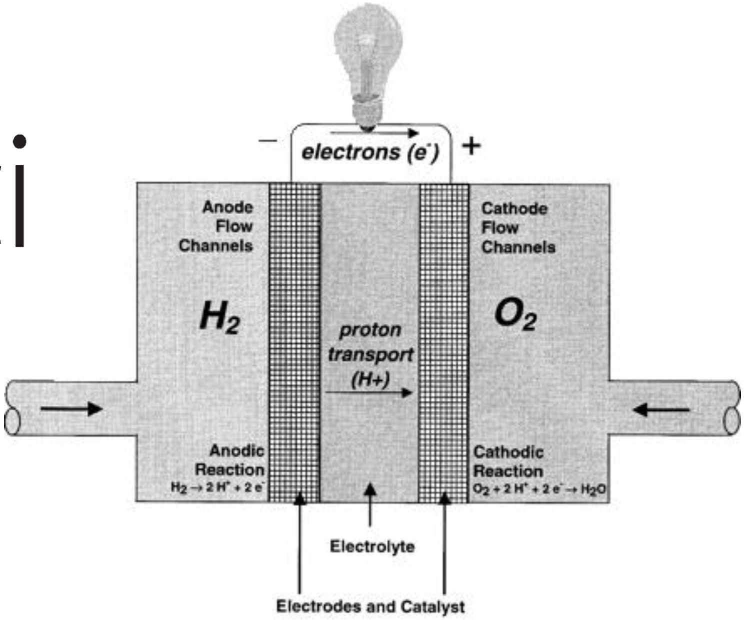


Figure 1. Fuel cell schematic.
Source: Annual Reveiws of Energy and the Environment. http://energy.annualreviews.org/cgi/content/full/24/1/281

An **Electrocatalyst** must:

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

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

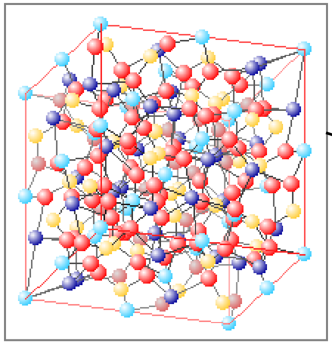
1) The reaction rate is still considered slow (causing **energy loss**)
2) Platinum is fairly **costly**, **intolerant** to fuel **contaminants**, and has a **short lifetime**.

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

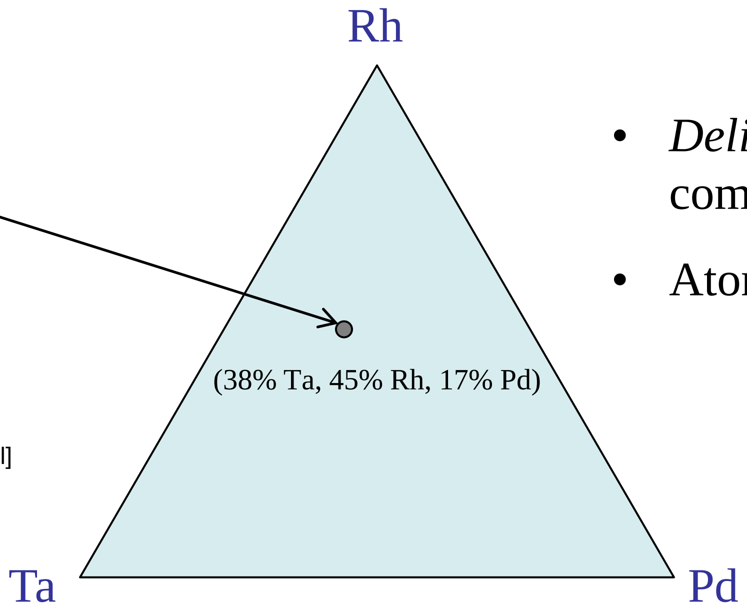
Motivation

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]

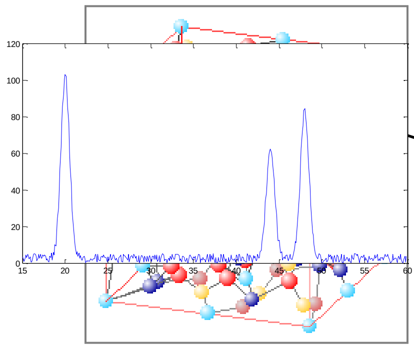


- *Deliberately* inhomogeneous composition on Si wafer
- Atoms are intimately mixed

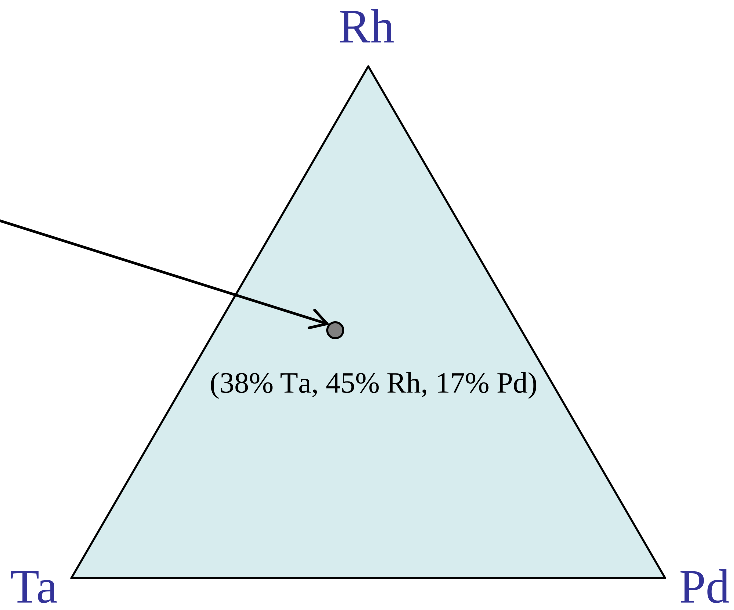
Motivation

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

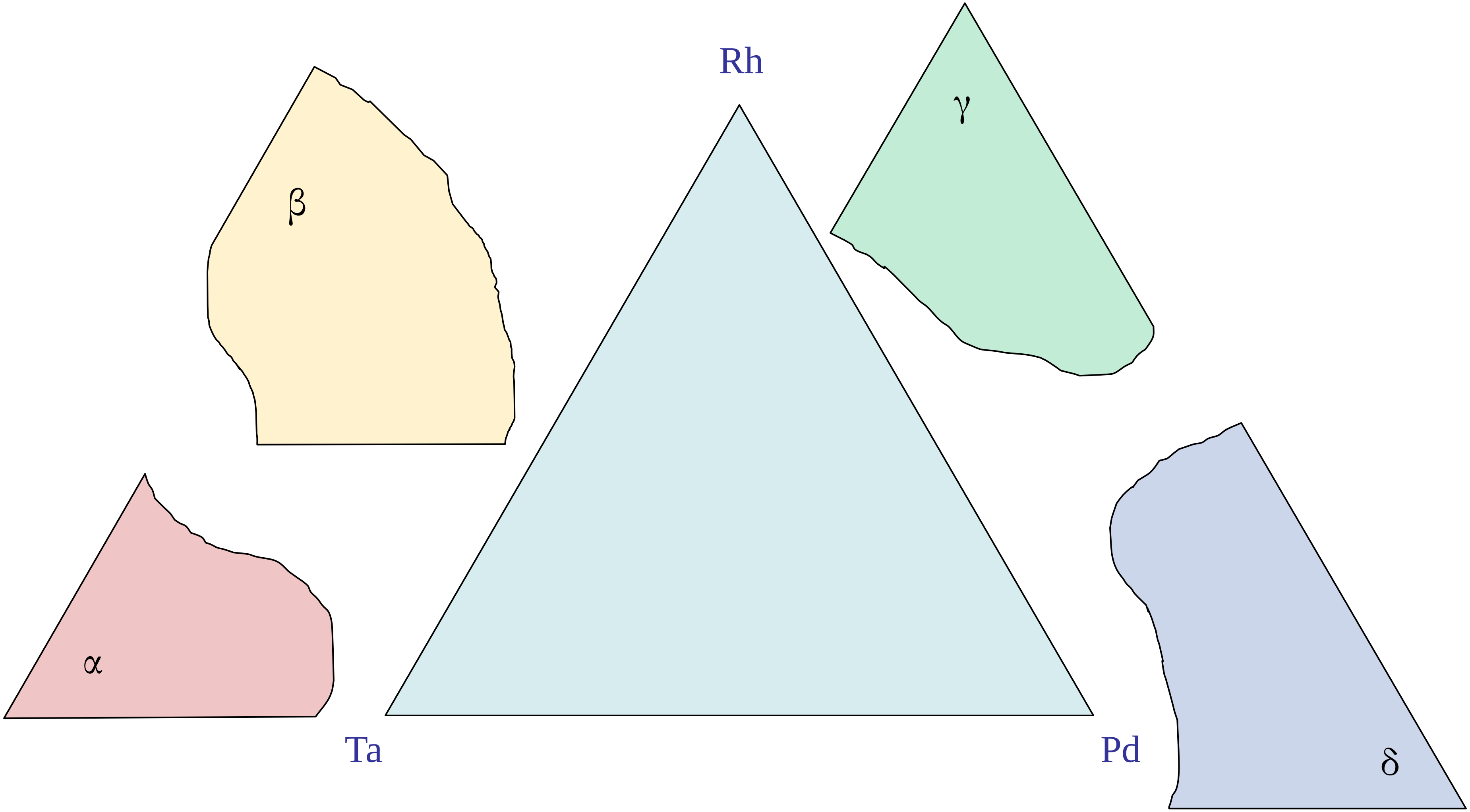
- XRD pattern **characterizes** the underlying **crystal** fairly well
- **Expensive** experimentations: Bruce van Dover’s research team has access to the facility **one week every year**.



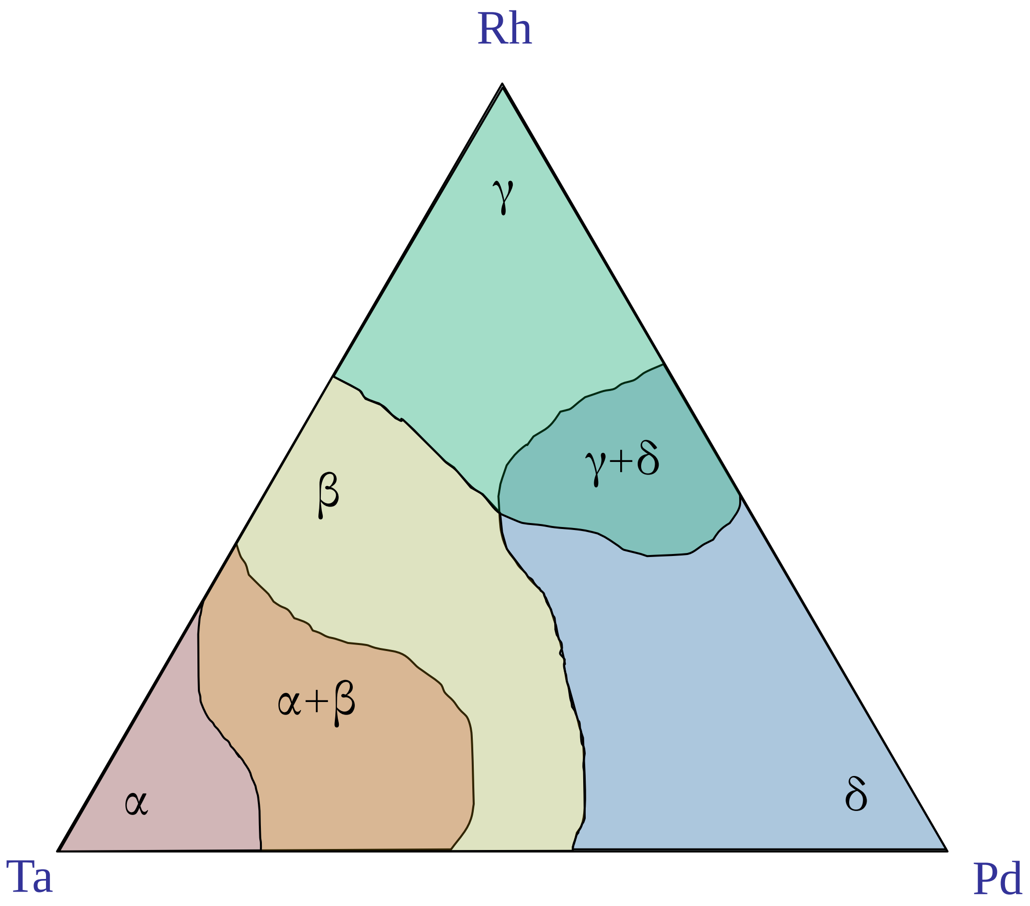
[Source: *Pyrotope*, Sebastien Merkel]



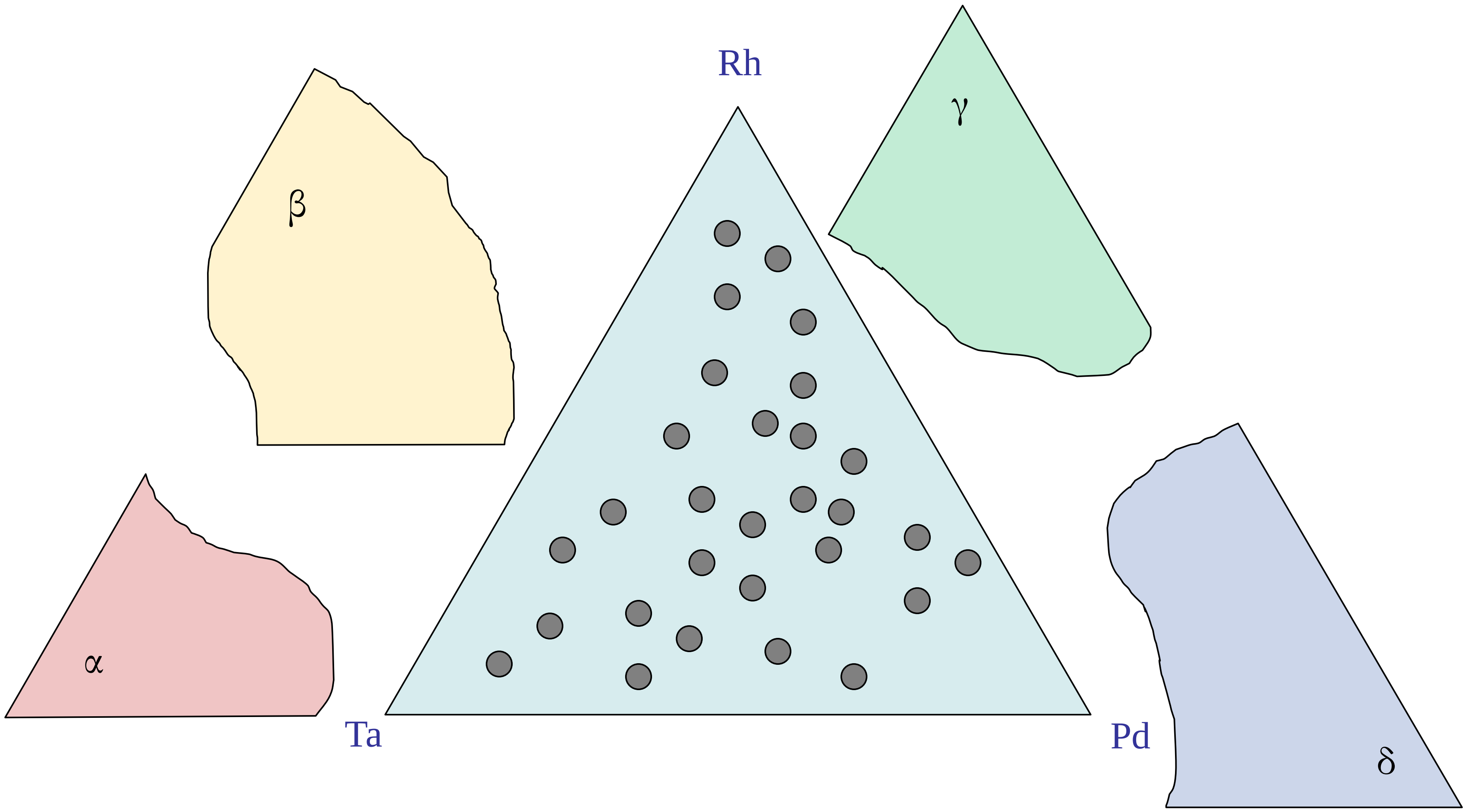
Motivation



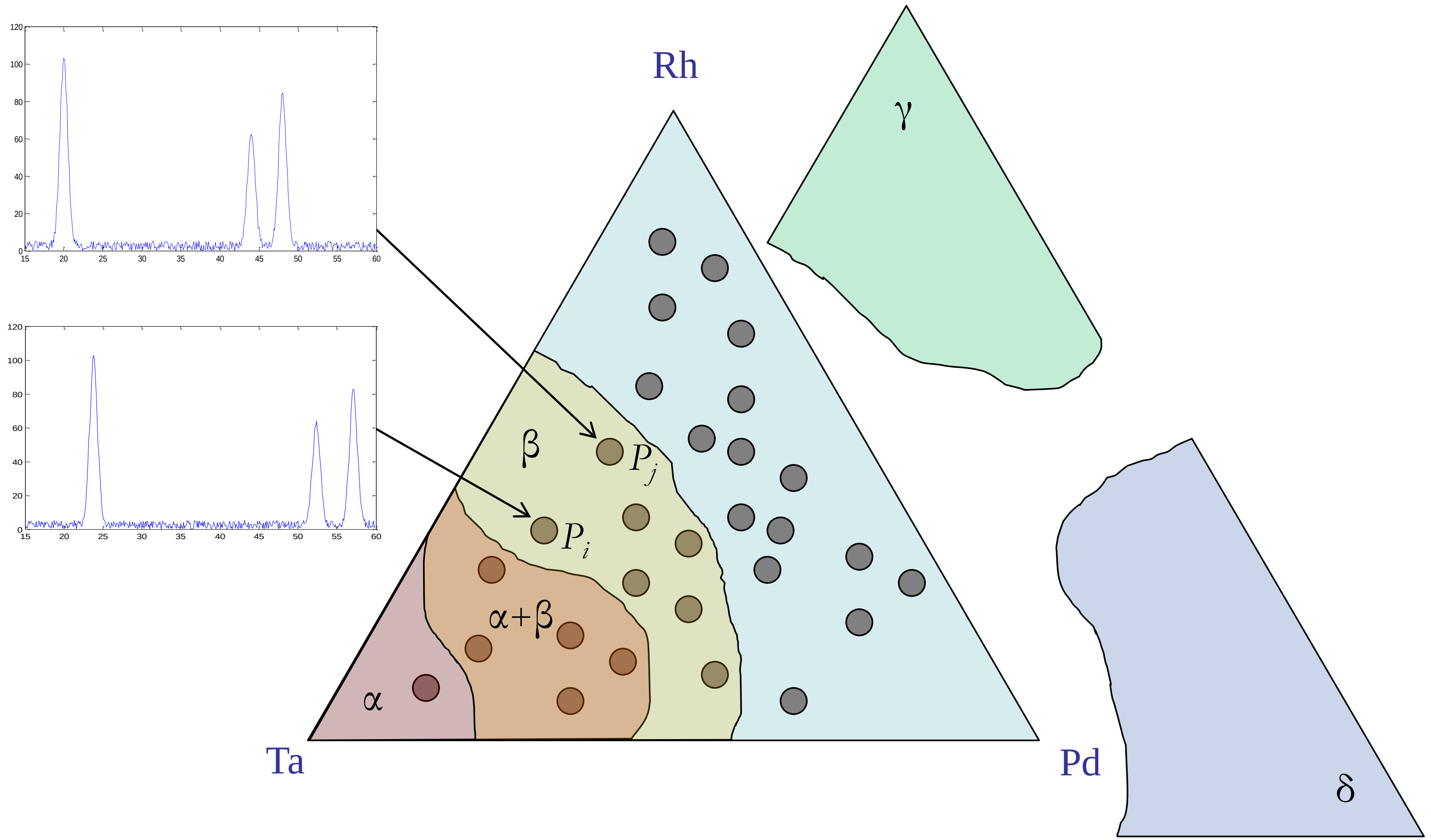
Motivation



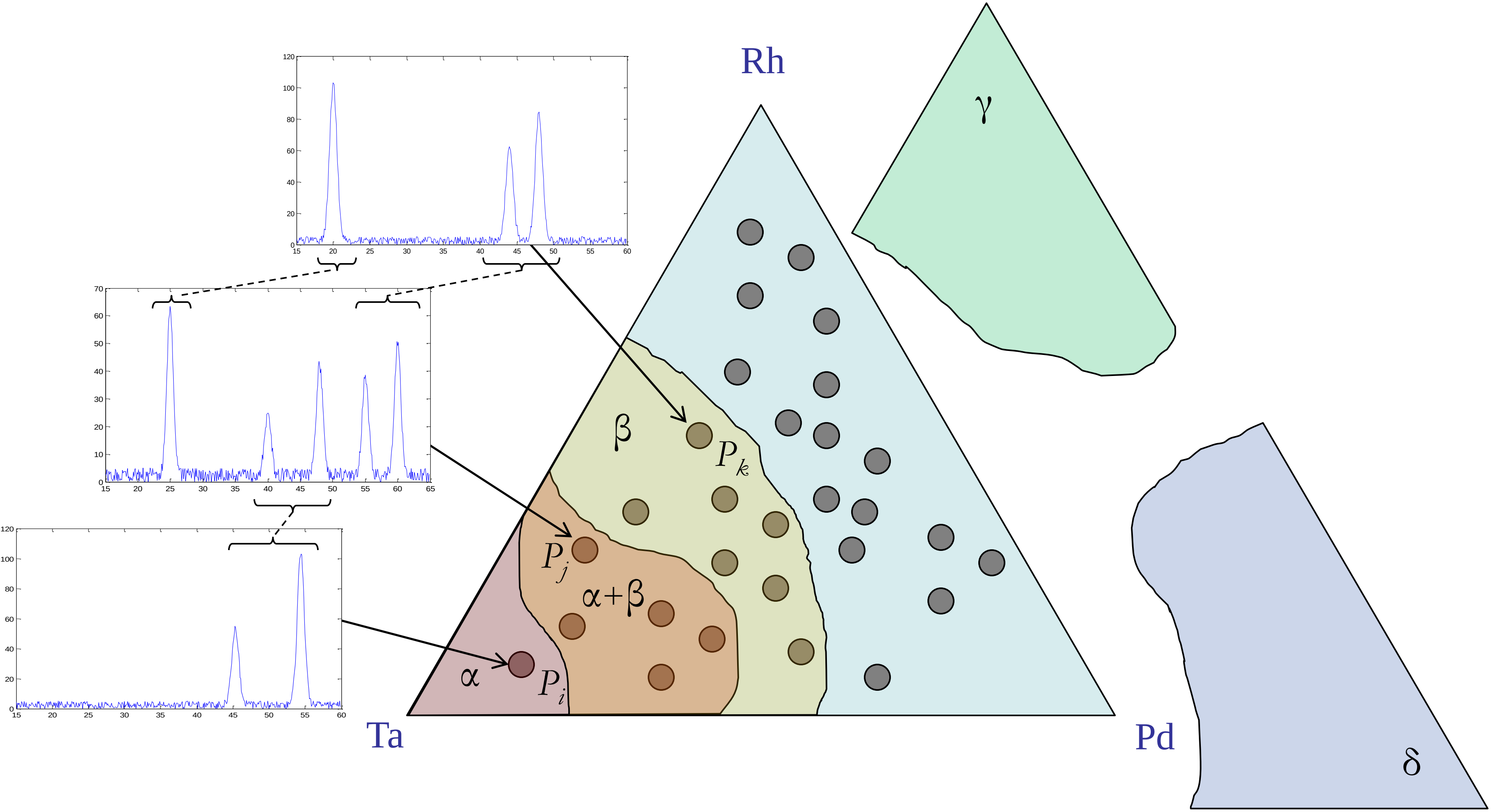
Motivation



Motivation

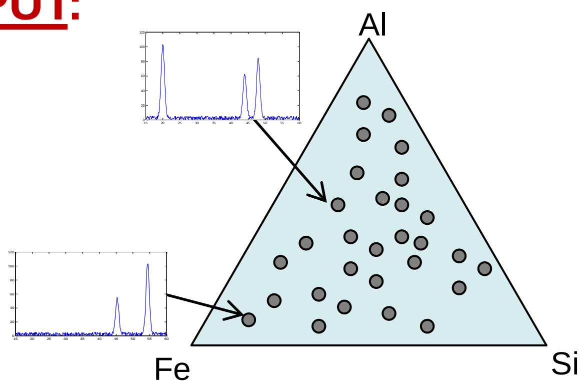


Motivation



Motivation

INPUT:

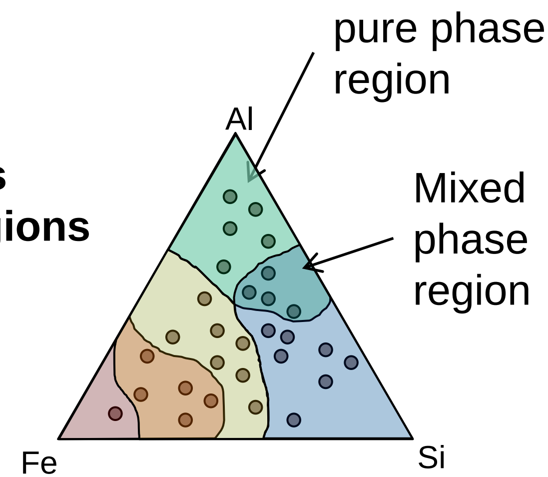


OUTPUT:

***m* phase regions**

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

XRD pattern characterizing pure phases



Additional Physical characteristics:

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

Motivation

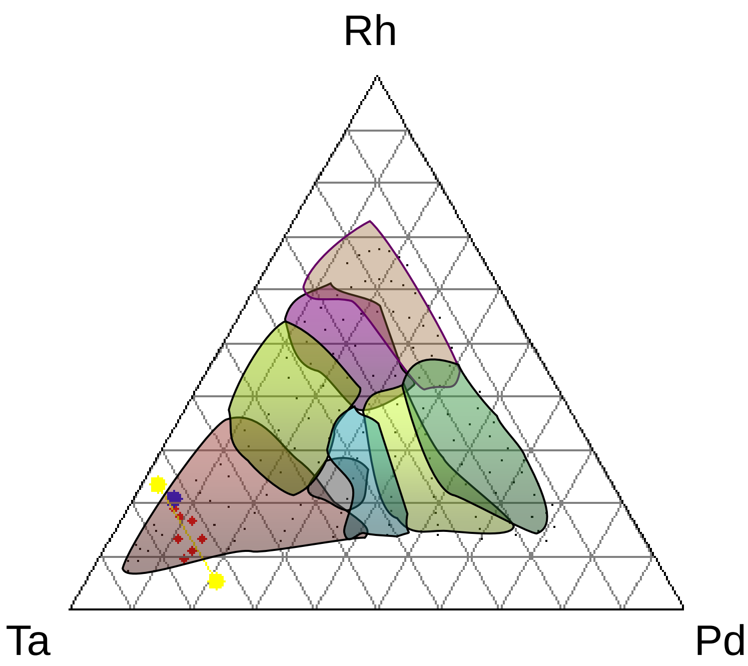


Figure 1: Phase regions of Ta-Rh-Pd

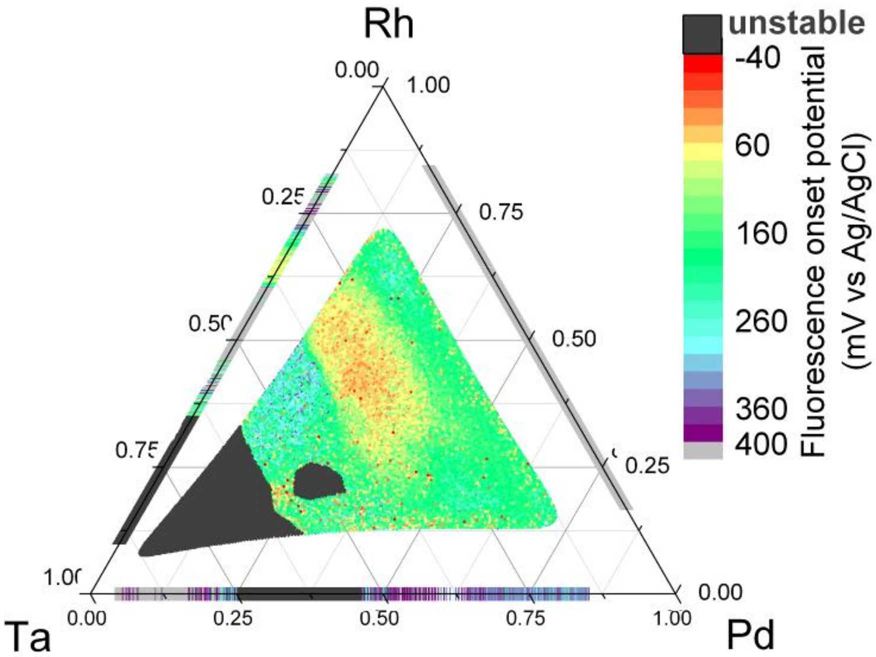


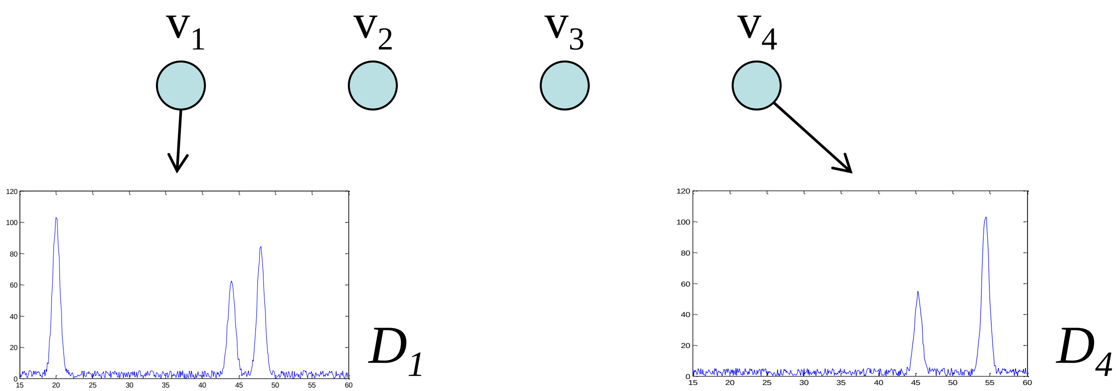
Figure 2: Fluorescence activity of Ta-Rh-Pd

Outline

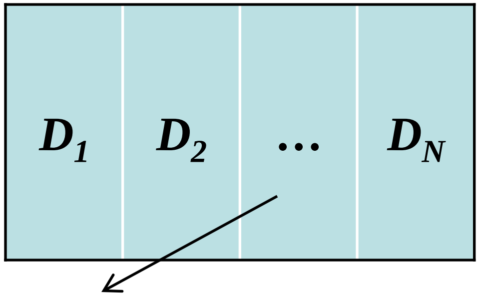
- Motivation
- Problem Definition (Part I)
- Previous Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

Problem Definition (Part I)

- **Input:**
  - A list of 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



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

Problem Definition (Part I)

- There is **experimental noise**

Minimize norm instead

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

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

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

Outline

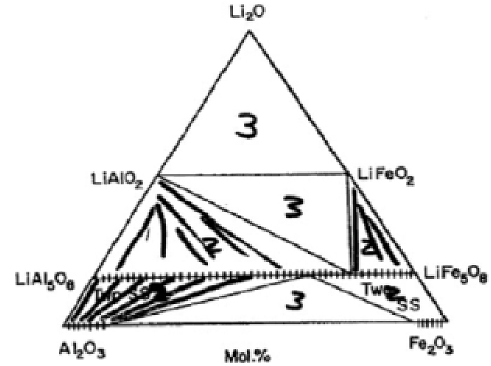
- Motivation
- Problem Definition (Part I)
- Prior Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

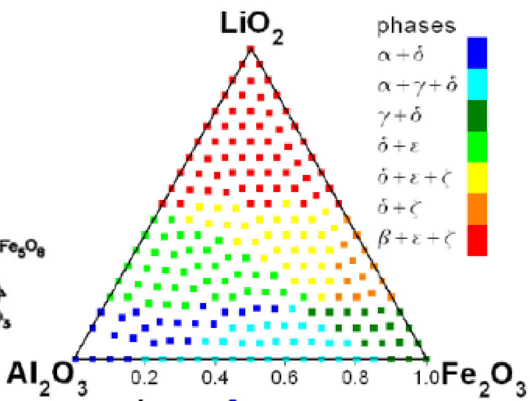
Non-negative Matrix Factorization [Long et al., 2009]

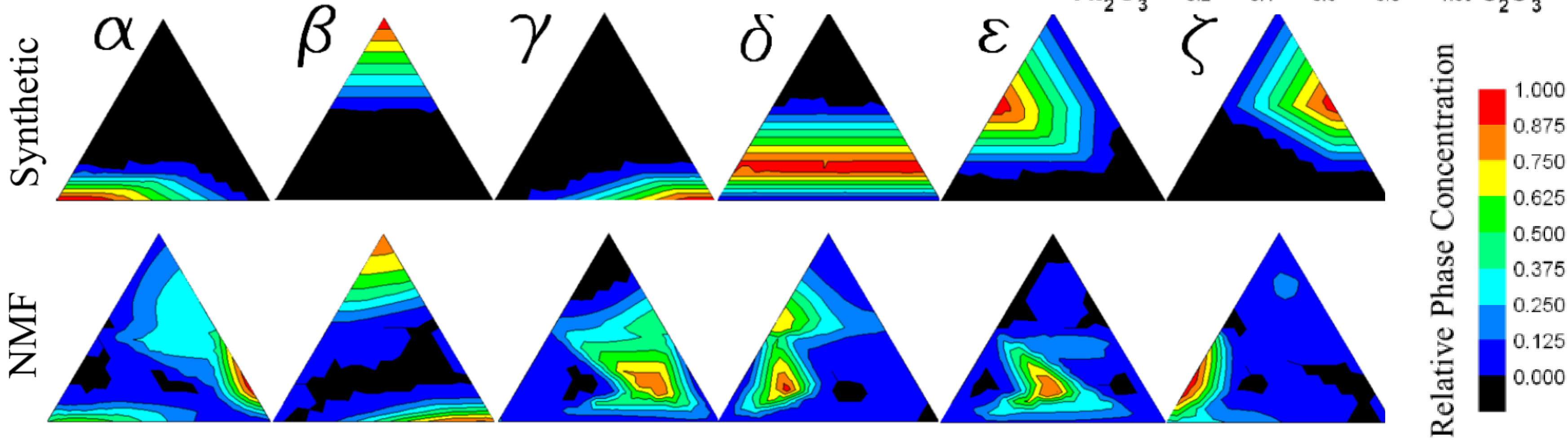
Advantages: scales up very well, accurately solves simple systems

Drawbacks: overlooks critical physical behavior, making the results physically meaningless for more complex systems.

Illustration on synthetic instances from the *Al-Li-Fe* ternary system







Outline

- Motivation
- Problem Definition (Part I)
- Prior Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

Problem Definition (Part II)

- 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

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

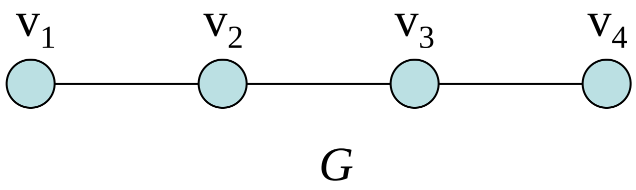
- **At most M (=3) non-zero coefficients** per point

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

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

Build a graph G of the points on the silicon wafer

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



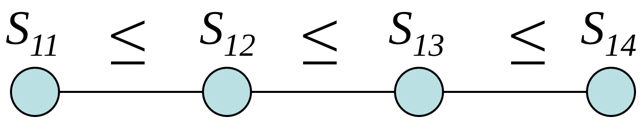
Problem Definition (Part II)

- Basis vector can be **shifted**

Shift operator → S; Shift coefficients → s_{iK}

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

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



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

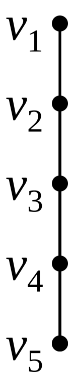
Outline

- Motivation
- Problem Definition (Part I)
- Prior Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

Satisfiability Modulo Theories Approach

- Initial graph G and number K of basis patterns

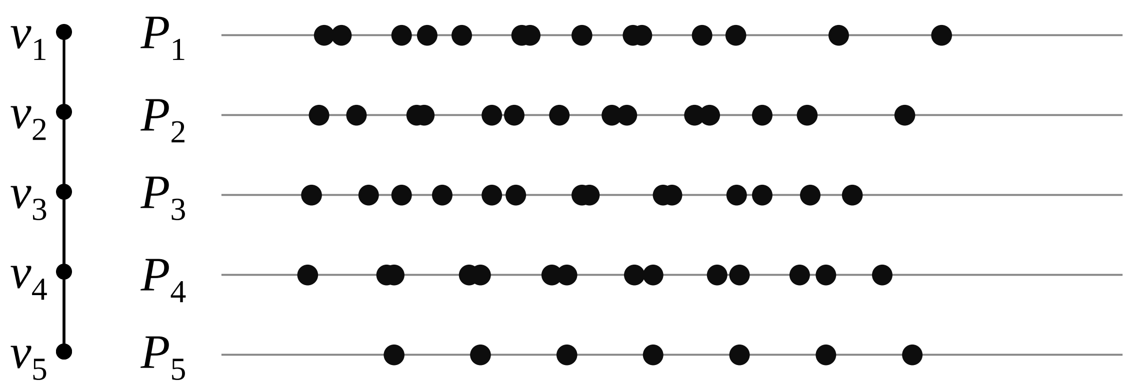
K=2
basis
patterns



Satisfiability Modulo Theories Approach

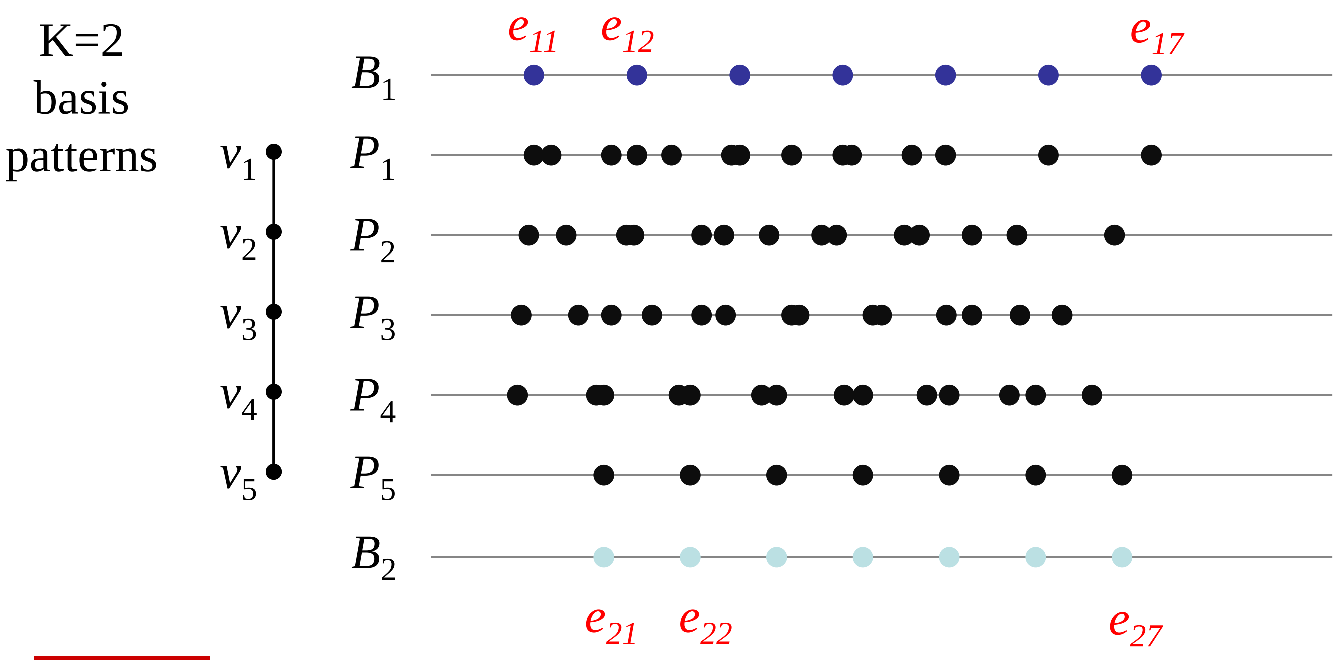
- Initial graph *G* and number *K* of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$

K=2 basis patterns



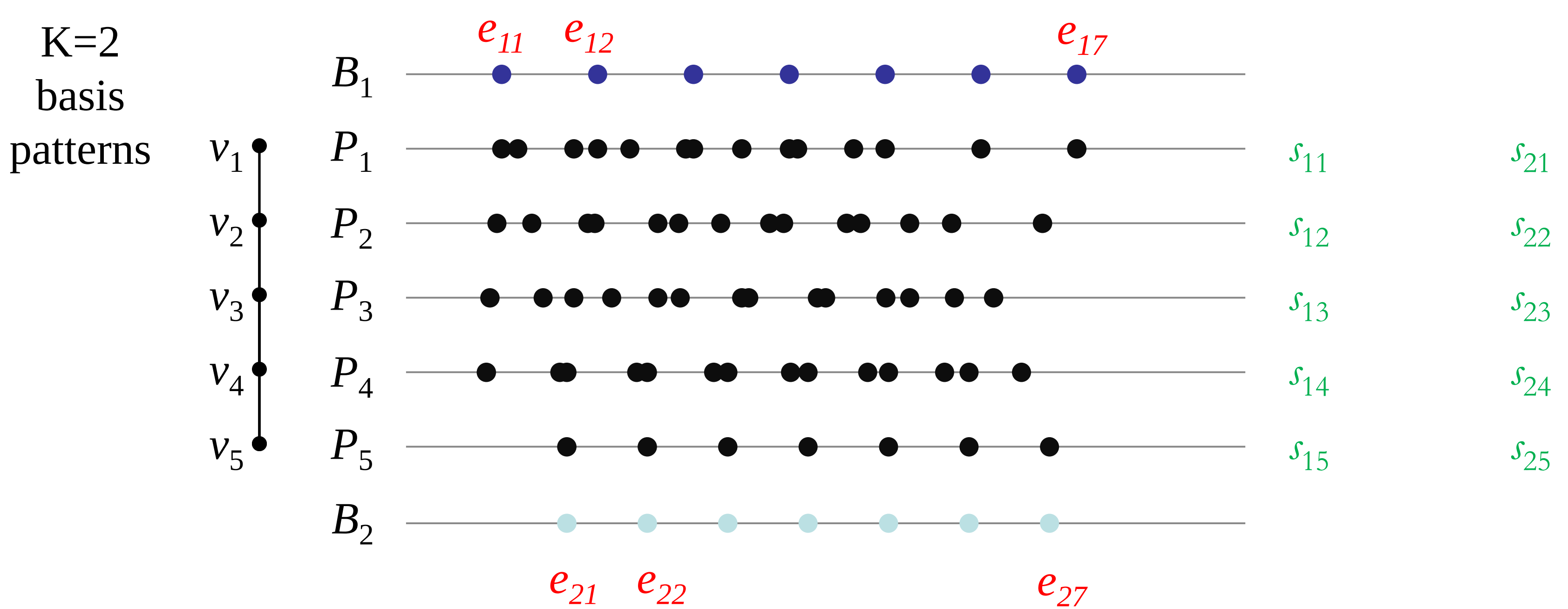
Satisfiability Modulo Theories Approach

- Initial graph *G* and number *K* of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$
- Real variable e_{jk} for the **location of peak** *k* in basis B_j



Satisfiability Modulo Theories Approach

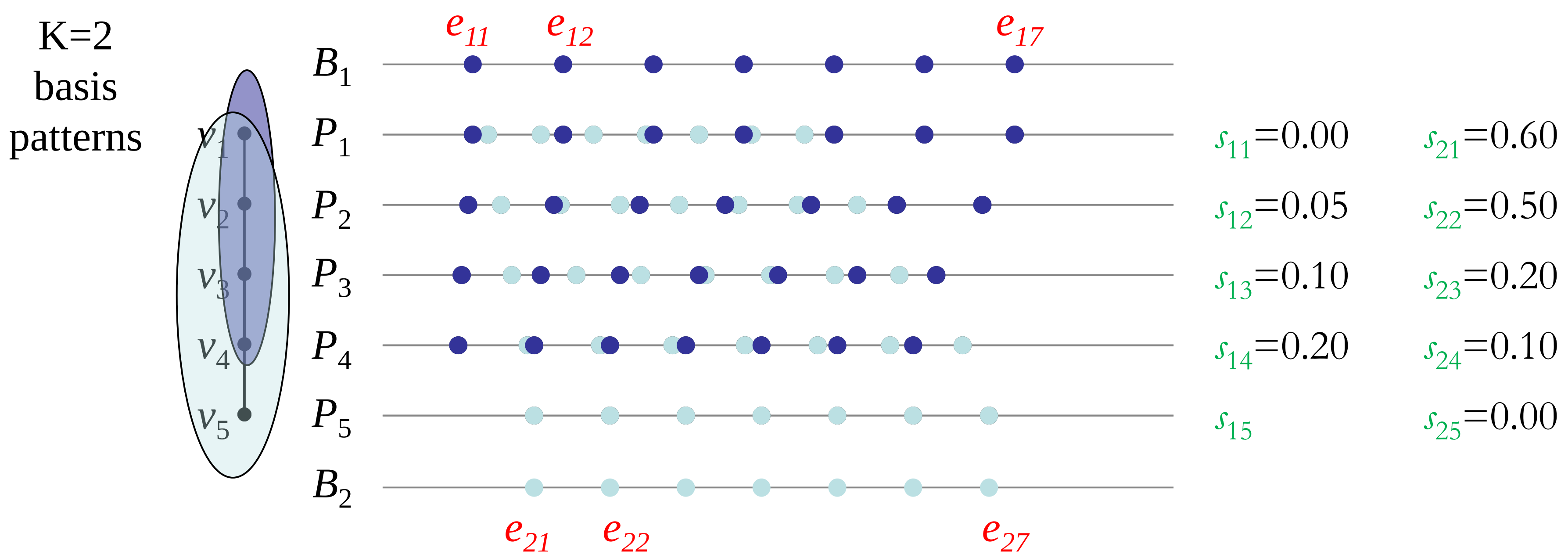
- Initial graph G and number K of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$
- Real variable e_{jk} for the **location of peak** k in basis B_j
- Real variable s_{ij} for the **shift** coefficient of basis B_j in point P_i



CORNELL UNIVERSITY

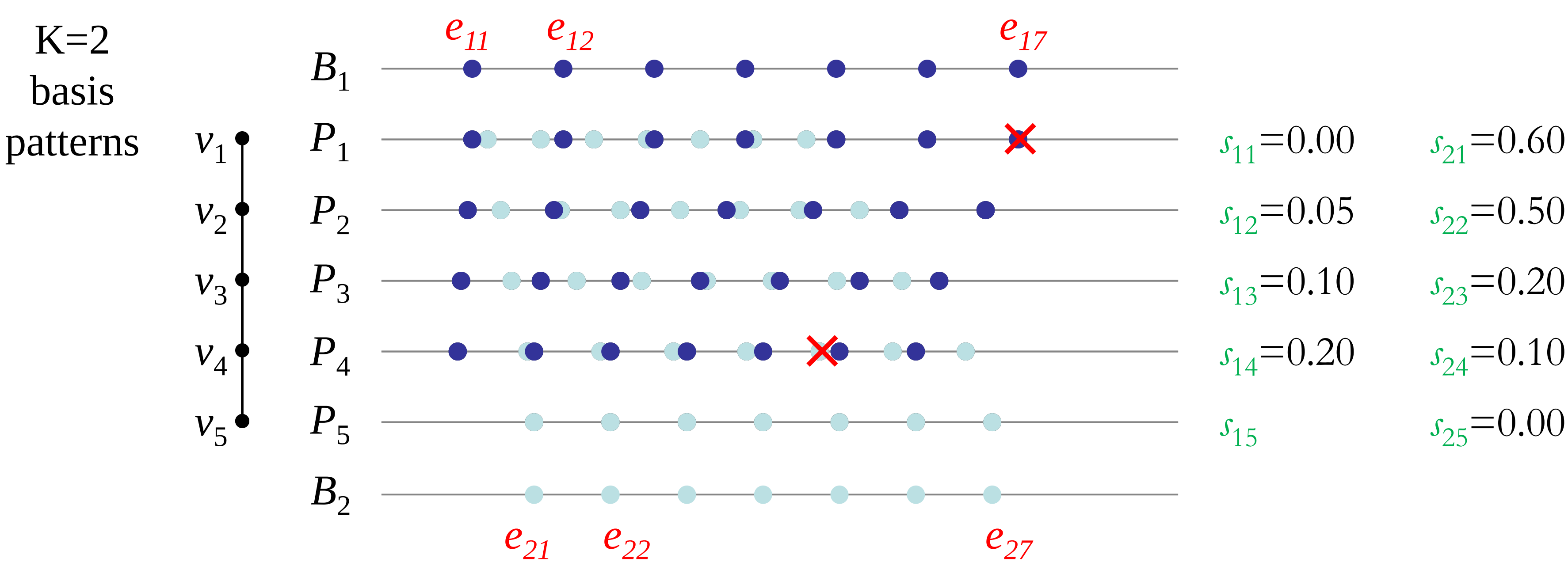
Satisfiability Modulo Theories Approach

- An observed peak p is "*explained*" if there exists s_{ij}, e_{jk} s.t. $|p-(s_{ij} + e_{jk})| \leq \varepsilon$
- Every observed peak must be *explained*



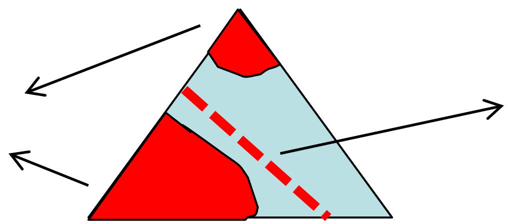
Satisfiability Modulo Theories Approach

- An observed peak p is "*explained*" if there exists s_{ij} ,e_{jk} s.t. $|p-(s_{ij} + e_{jk})| \leq \varepsilon$
- Every observed peak must be *explained*
- Some peaks might be missing (unobserved)
- Bound the number of missing peaks $\leq T$
- Minimization by (binary) search on T



Satisfiability Modulo Theories Approach

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



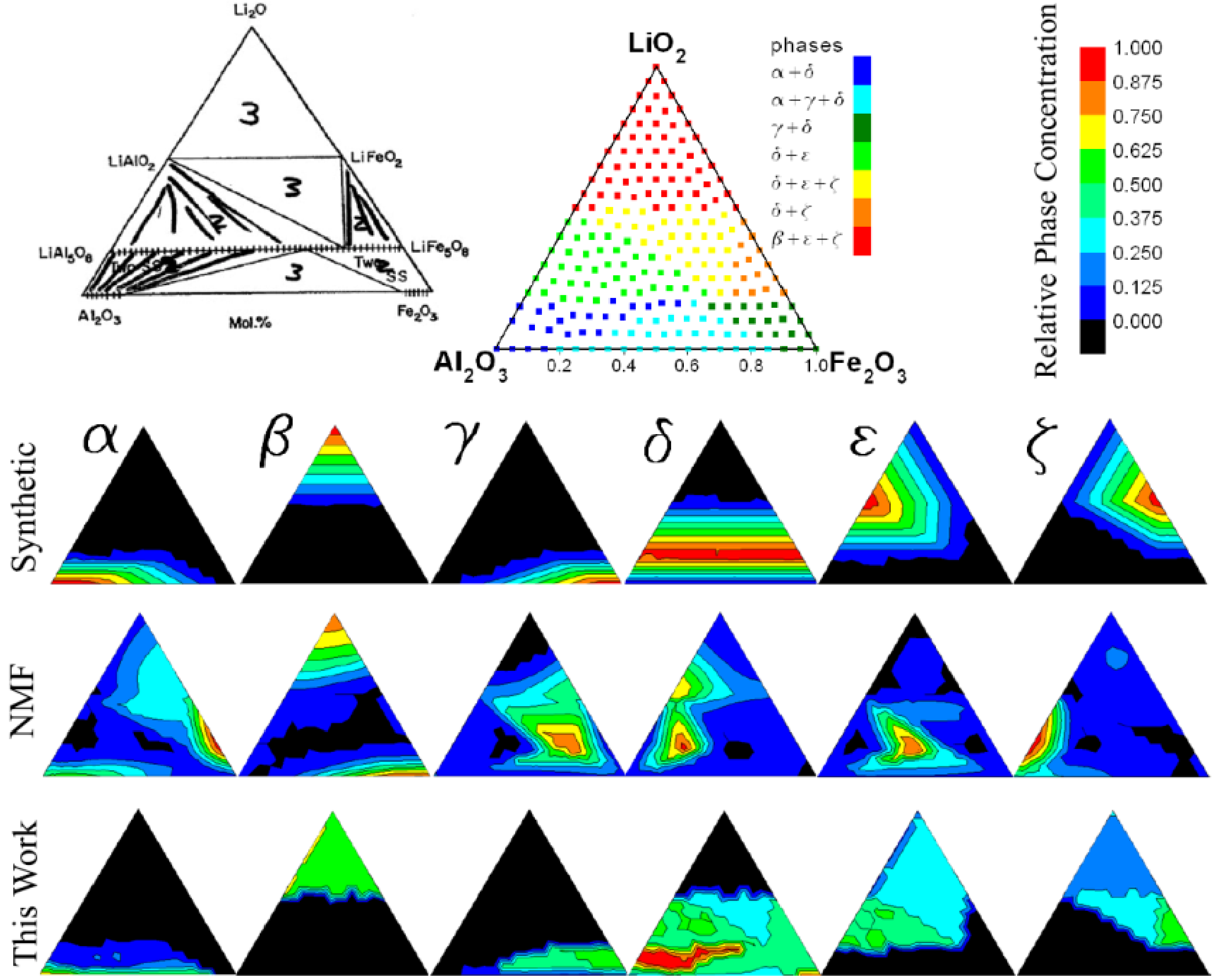
- **Symmetry breaking**:
  - Renaming of pure phases
  - Ordering of the peak locations e_{jk} (per basis pattern)

Quantifier-free linear arithmetic

CORNELL UNIVERSITY

Experimental Results

•Illustration on Al-Li-Fe system



Conclusion and Future work

- Novel **SMT** encoding for Materials Discovery
- Good performance on synthetic data:
  - Scales to **realistic sized problems** (~50 points)
  - Provides **physically-meaningful** solutions
  - **Good accuracy** (>90% precision and recall)
  - Outperforms both Constraint Programming and Mixed Integer Programming direct translations of the SMT model
- **Future work**: online adaptive sampling during data collection
- Exciting results analyzing and explaining **real-world data**

THANK YOU!

Extra slides

Runtime

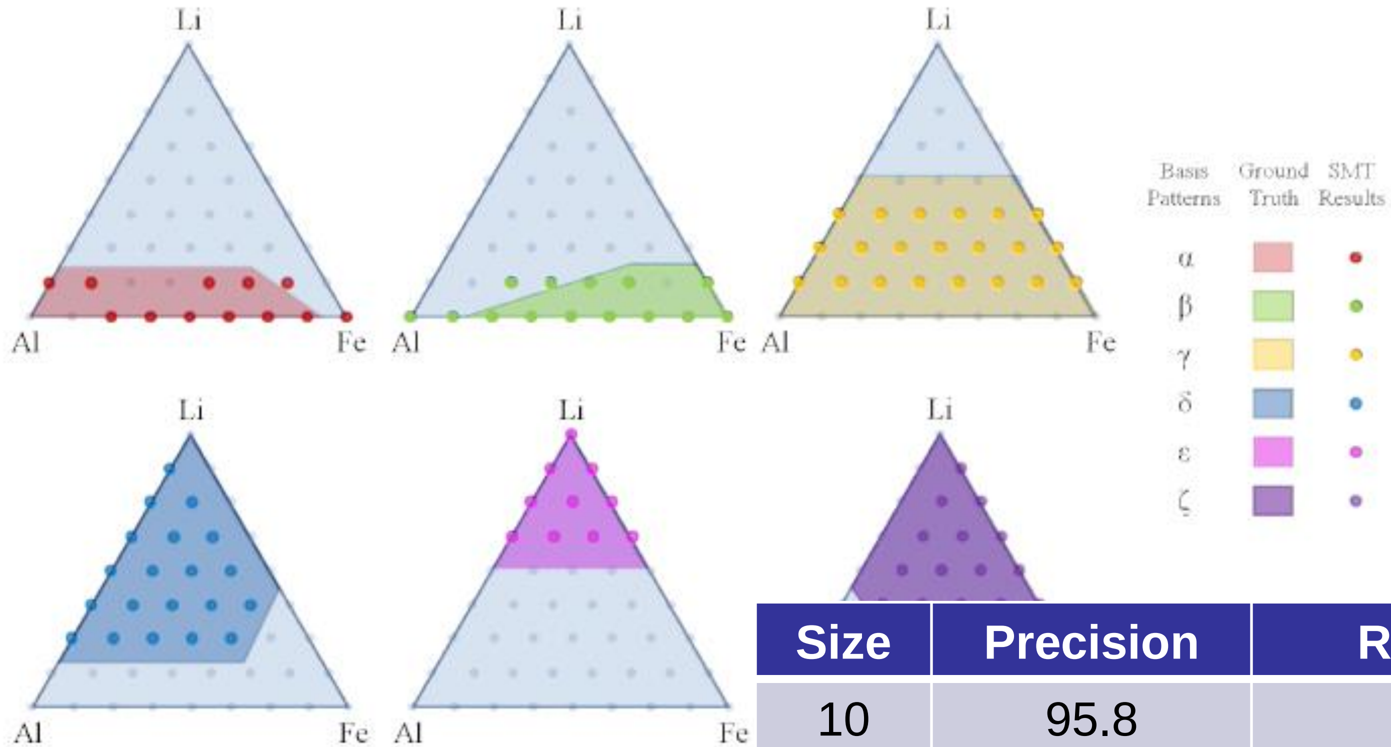
# Points	Unknown Phases	Arithmetic + Z3 (s)	Set-based + CPLEX (s)
10	3	8	**0.5**
	6	**12**	Timeout
15	3	13	**0.5**
	6	**20**	Timeout
18	3	**29**	384.8
	6	**125**	Timeout
29	3	**78**	276
	6	**186**	Timeout
45	6	**518**	Timeout

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



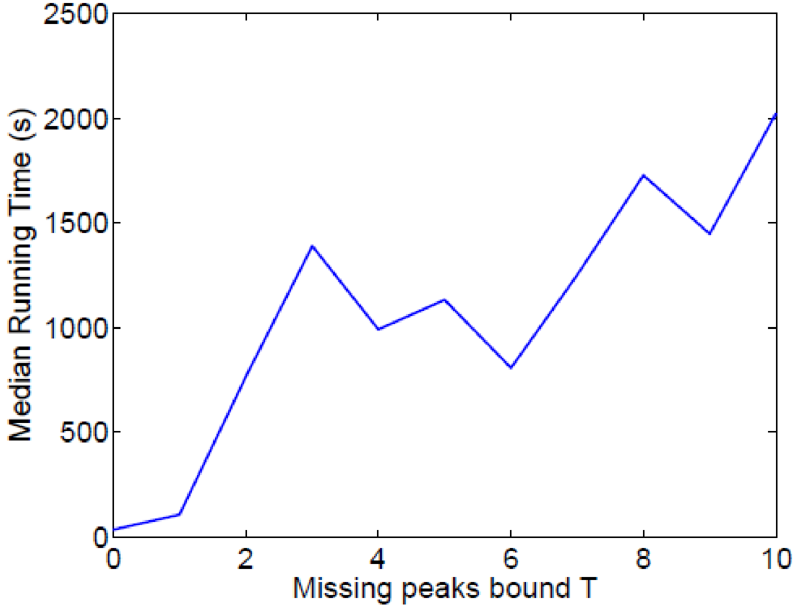
Recovers ground truth

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

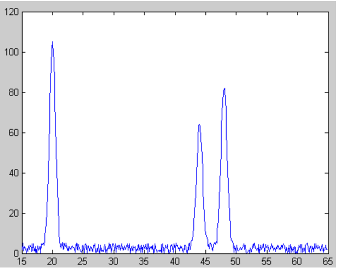
Missing Peaks	Precision	Recall
1	96.1	99.6
2	96.3	99.3
3	96.7	99.5
4	95.3	98.9
5	94.8	99.7



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

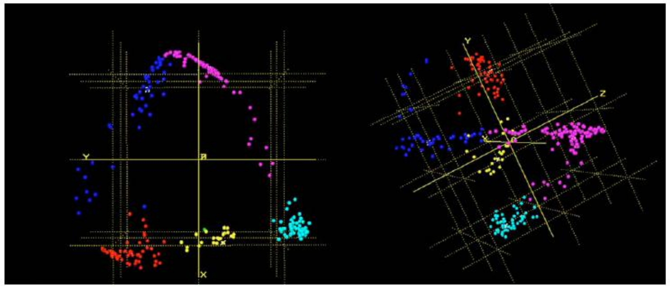
Previous Work 1: Cluster Analysis [Long et al., 2007]

$x_i =$

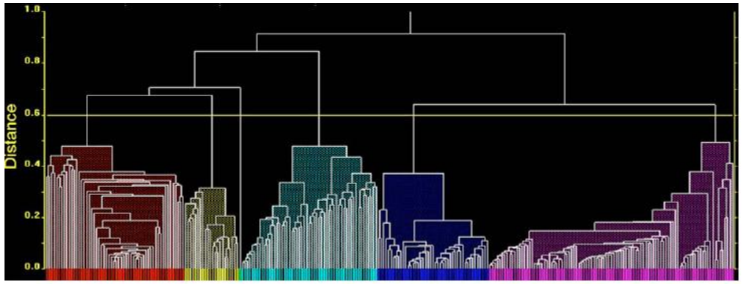


$$C_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}} \longrightarrow D = (1 - C)/2.$$

Feature vector *Pearson correlation coefficients* *Distance matrix*



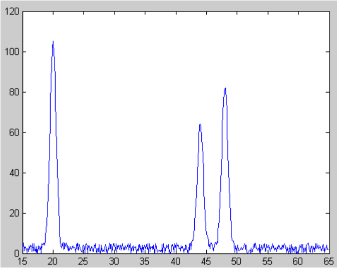
PCA – 3 dimensional approx



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]

$x_i =$ 

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

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.).

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 $\leq T$