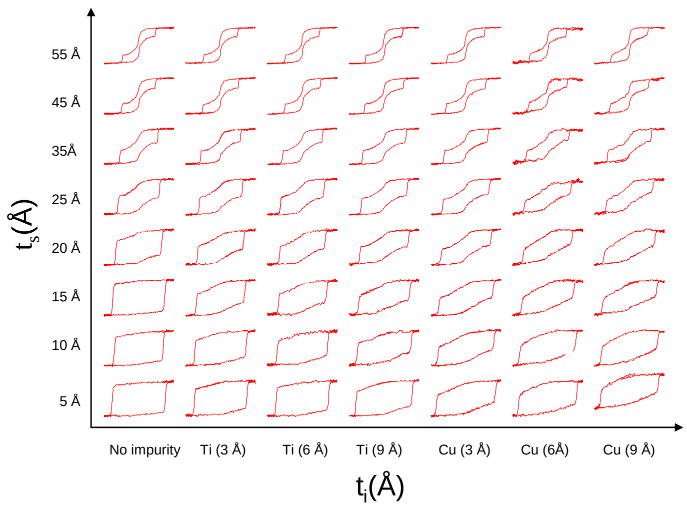
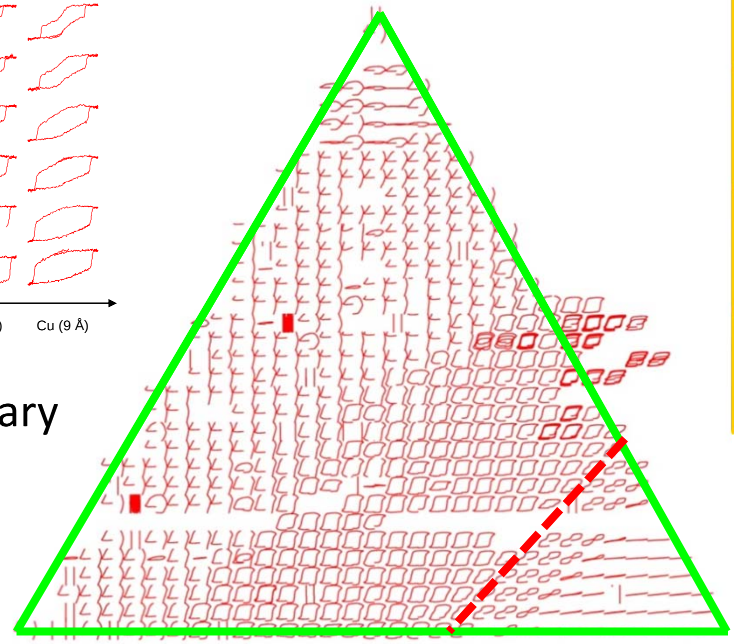
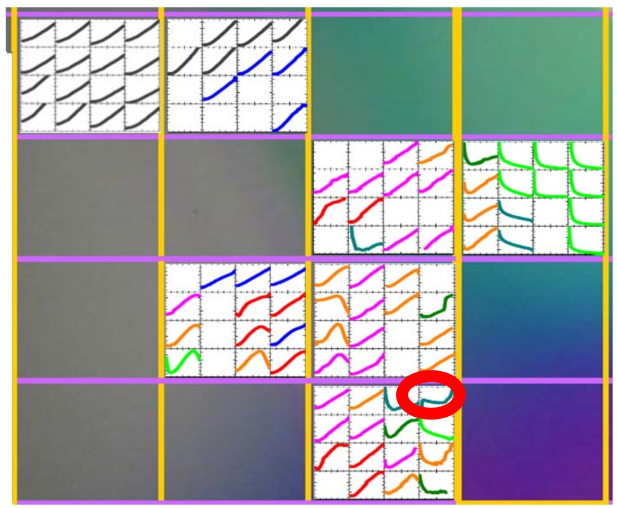
High-throughput Experimentation and Machine Learning for Materials Discovery



Permanent magnet library

Ferroelectric library





Superconductor library

Ichiro Takeuchi
University of Maryland

Outline

- **Combinatorial approach to materials discovery: mapping of complex energy and property landscapes**

- **Search for new superconductors and supervised machine learning of experimental database**

- **Structural phase distribution mapping: unsupervised machine learning of diffraction data**

- **Active learning is the future of high-throughput experimentation**

Acknowledgement

Univ. of Maryland
V. Stanev
T. Gao
Y. Iwasaki
R. Greene

NIST
A. G. Kusne

SLAC
A. Mehta

Duke University
S. Curtarolo

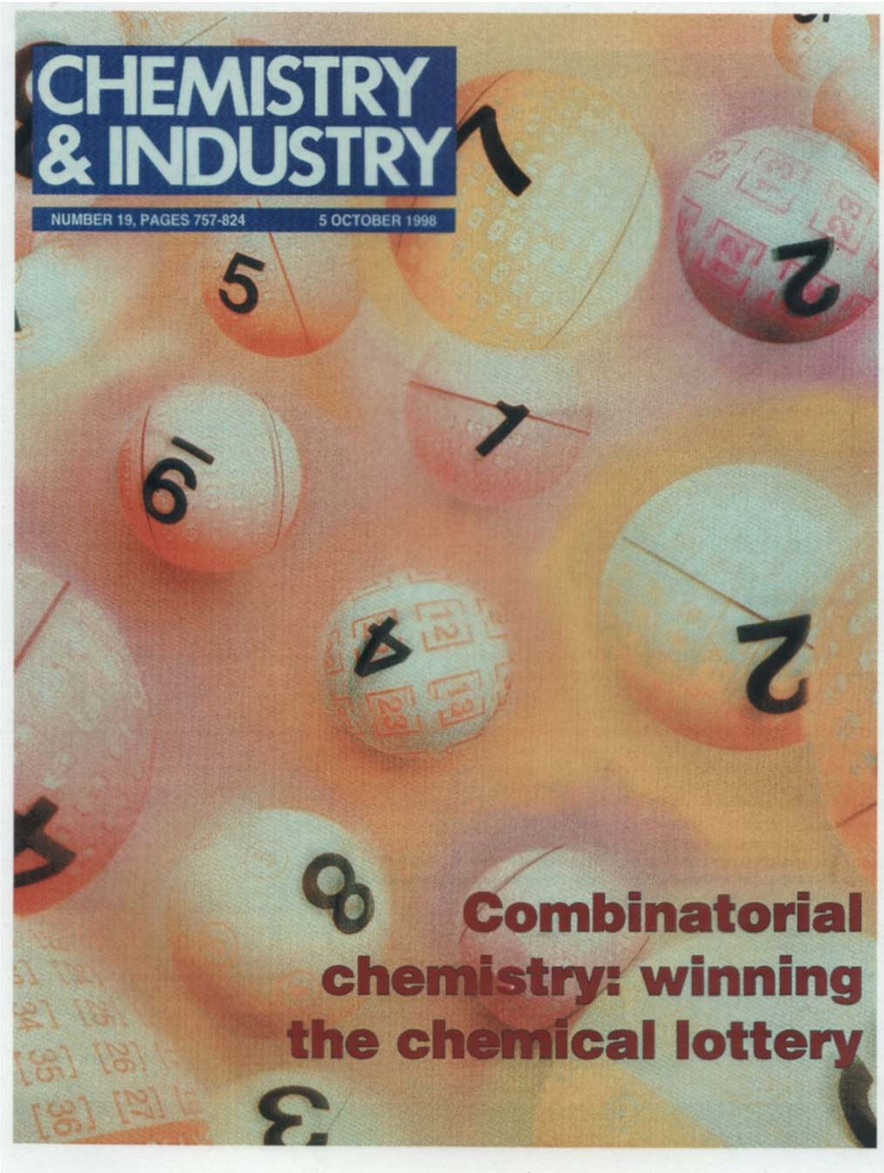
Support

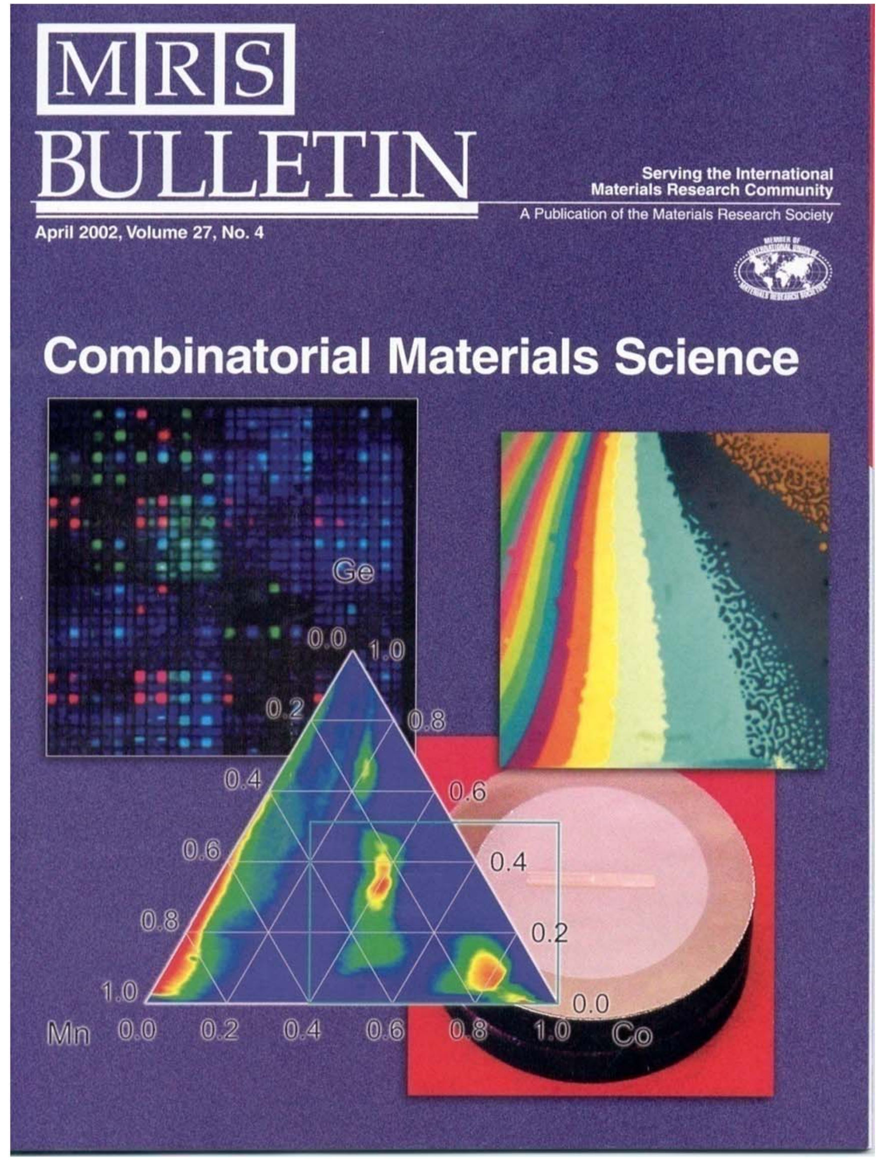
ONR, AFOSR, DOE

MANY AT A TIME Program head Andreas Marzinik (front to back) and lab specialists Raphael Gattlen and Urs Rindisbacher of Novartis Pharma AG, Basel, Switzerland, pipette coupling reagent into 96-well reaction blocks.

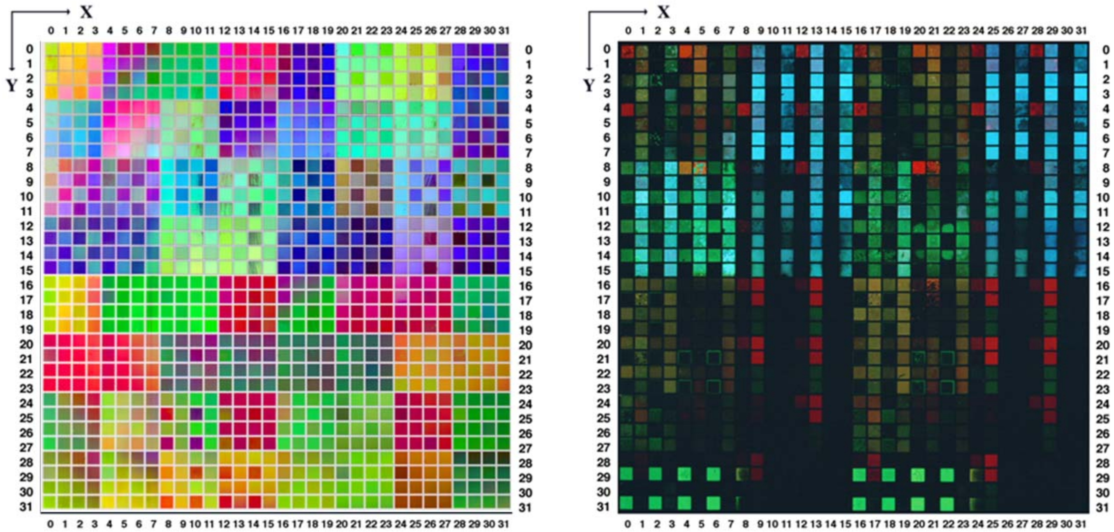
COMBINATORIAL CHEMISTRY

Chemical & Engineering News, August 2001

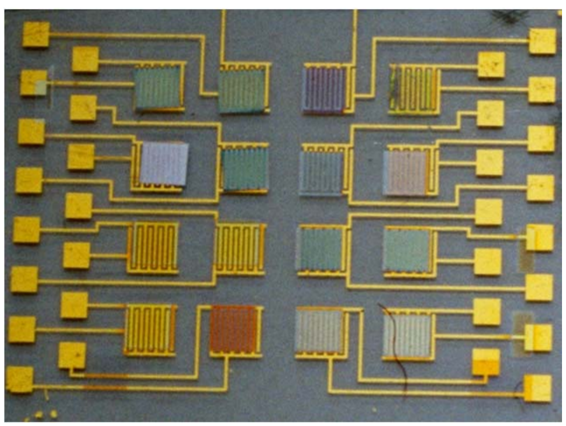

CHEMISTRY
& INDUSTRY
NUMBER 19, PAGES 757-824
5 OCTOBER 1998
Combinatorial chemistry: winning the chemical lottery


MRS BULLETIN
Serving the International Materials Research Community
A Publication of the Materials Research Society
April 2002, Volume 27, No. 4
Combinatorial Materials Science
Ge
Mn
Co
0.0
0.2
0.4
0.6
0.8
1.0

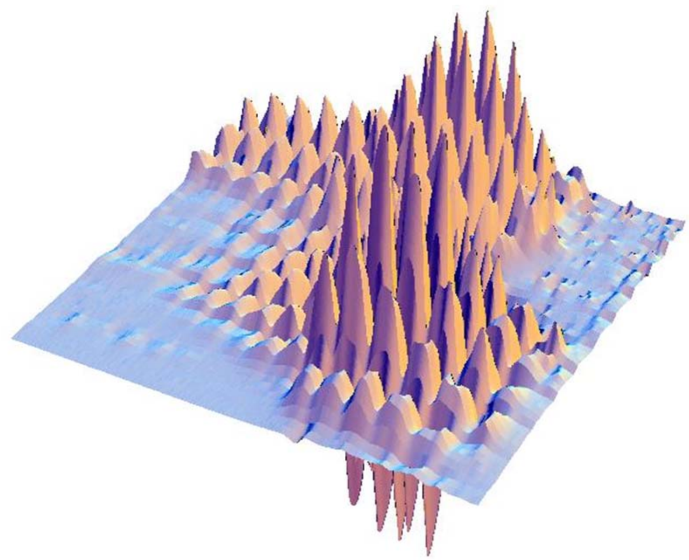
Combinatorial Libraries of Inorganic Materials



Luminescent materials libraries, Science **279**, 1712 (1998)



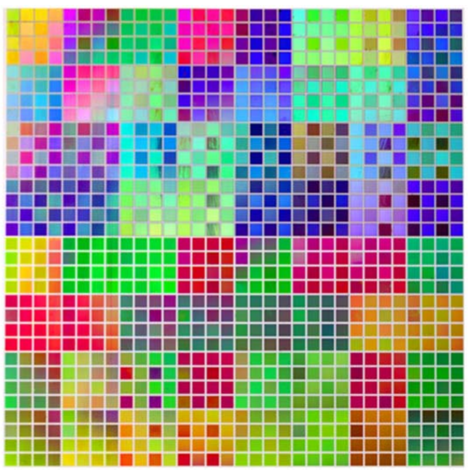
Semiconductor gas sensor library, "electronic nose", Appl. Phys. Lett. **83**, 1255 (2003)

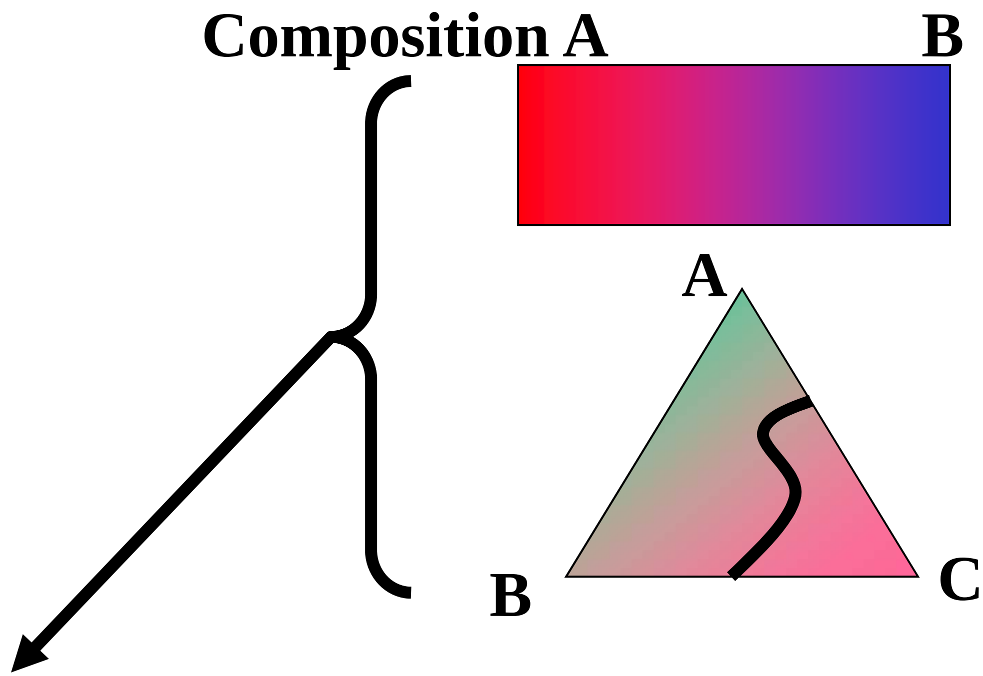


Magnetic shape memory alloy library, Nature Materials **2**, 180 (2003)

Various combinatorial experimental designs:

discrete libraries vs composition spreads

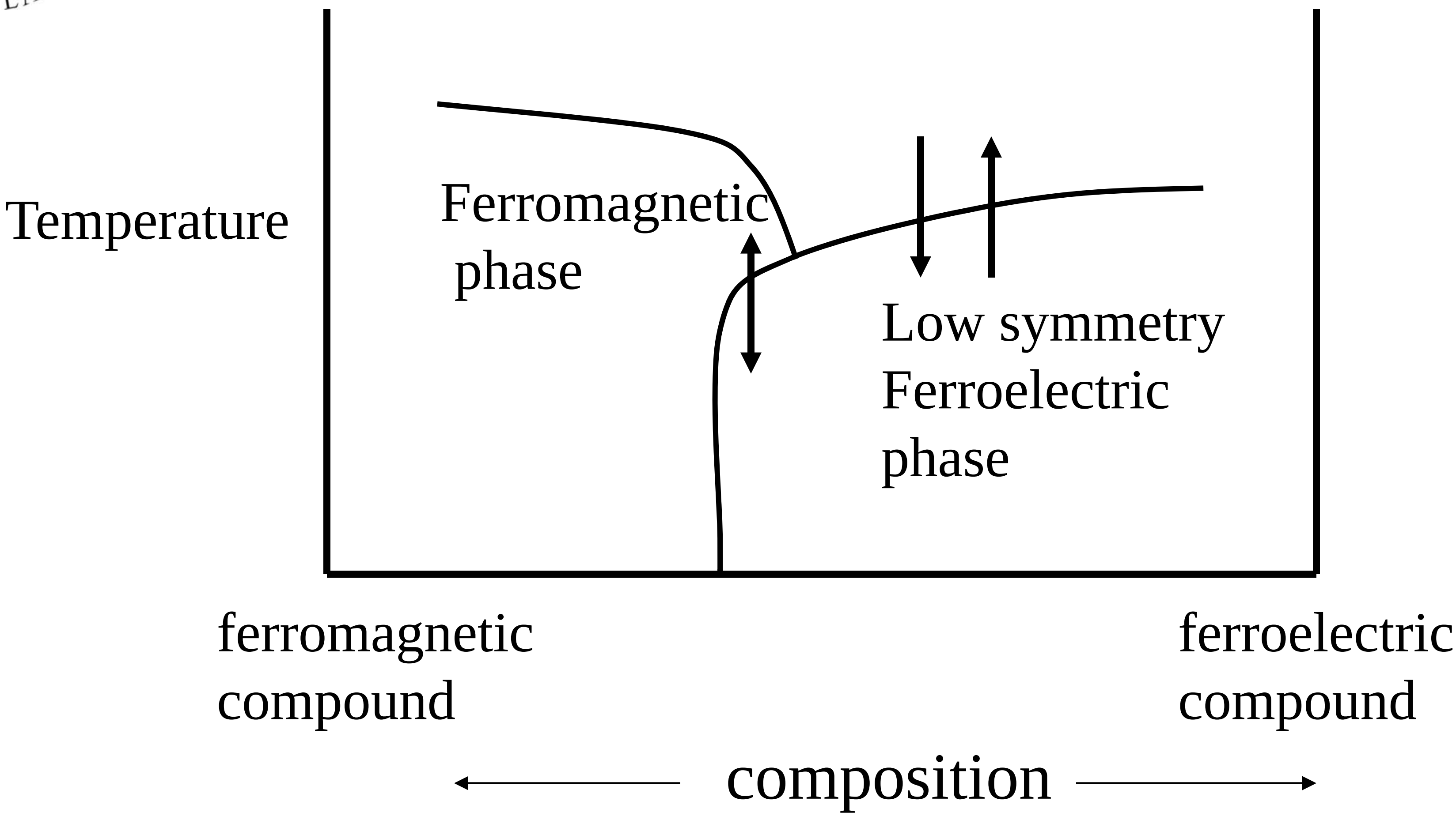




- **Composition spreads allow continuous mapping of physical properties and phase boundaries**

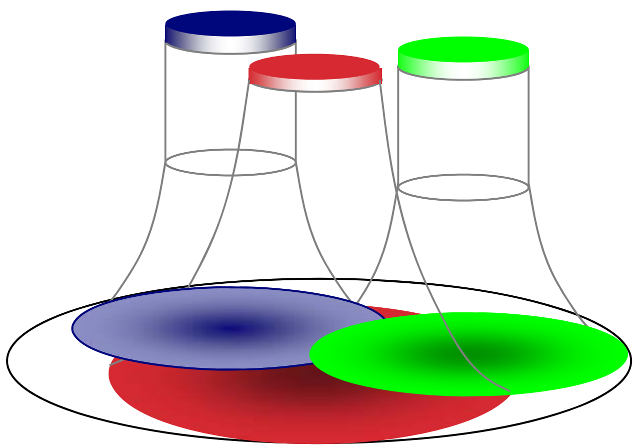
- **Run to run variation in ordinary experiments is removed**

Exploration of Interesting Phase Boundaries

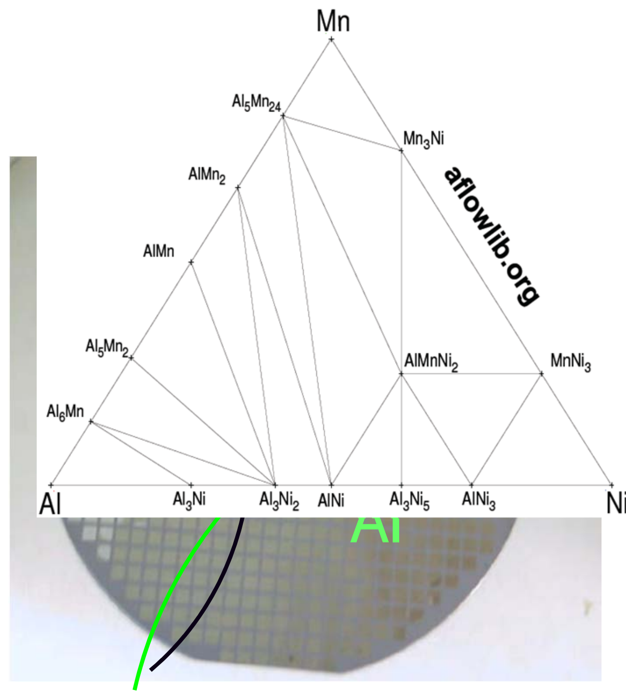


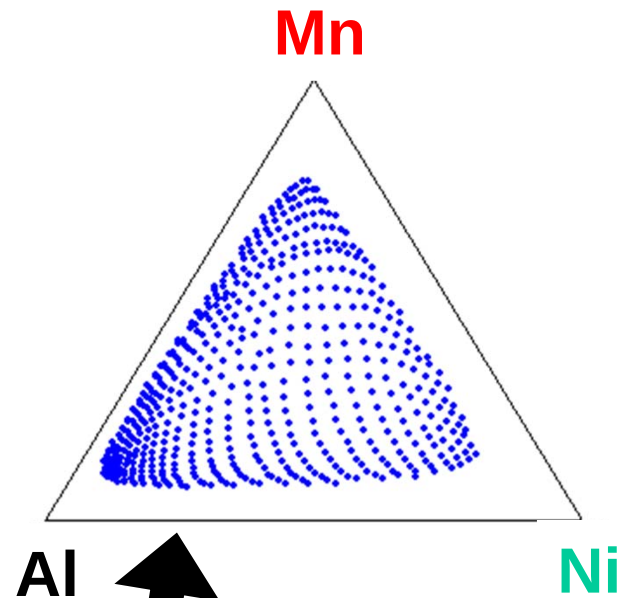
Shape memory alloys, piezoelectric materials, multiferroic materials, magnetostrictive materials, etc.

Composition Spreads of Ternary Metallic Alloy Systems



Co-sputtering scheme





Phase diagram

Composition is mapped using an electron probe (WDS)

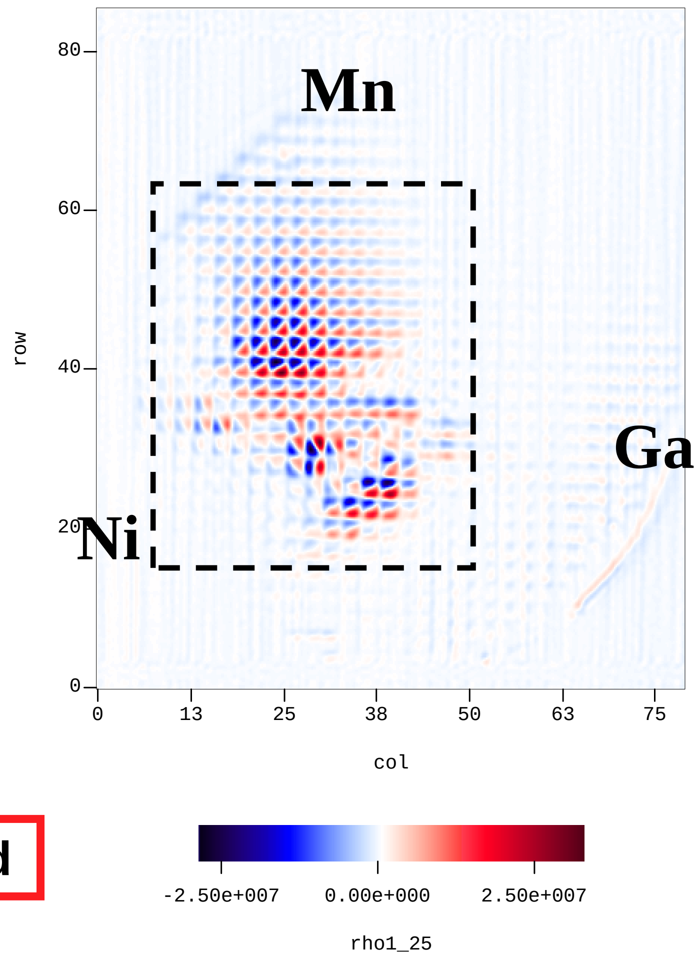
Review article: Green *et al*., JAP **113**, 231101 (2013)

Rapid mapping of magnetic properties: scanning SQUID

SQUID assembly
inside vacuum

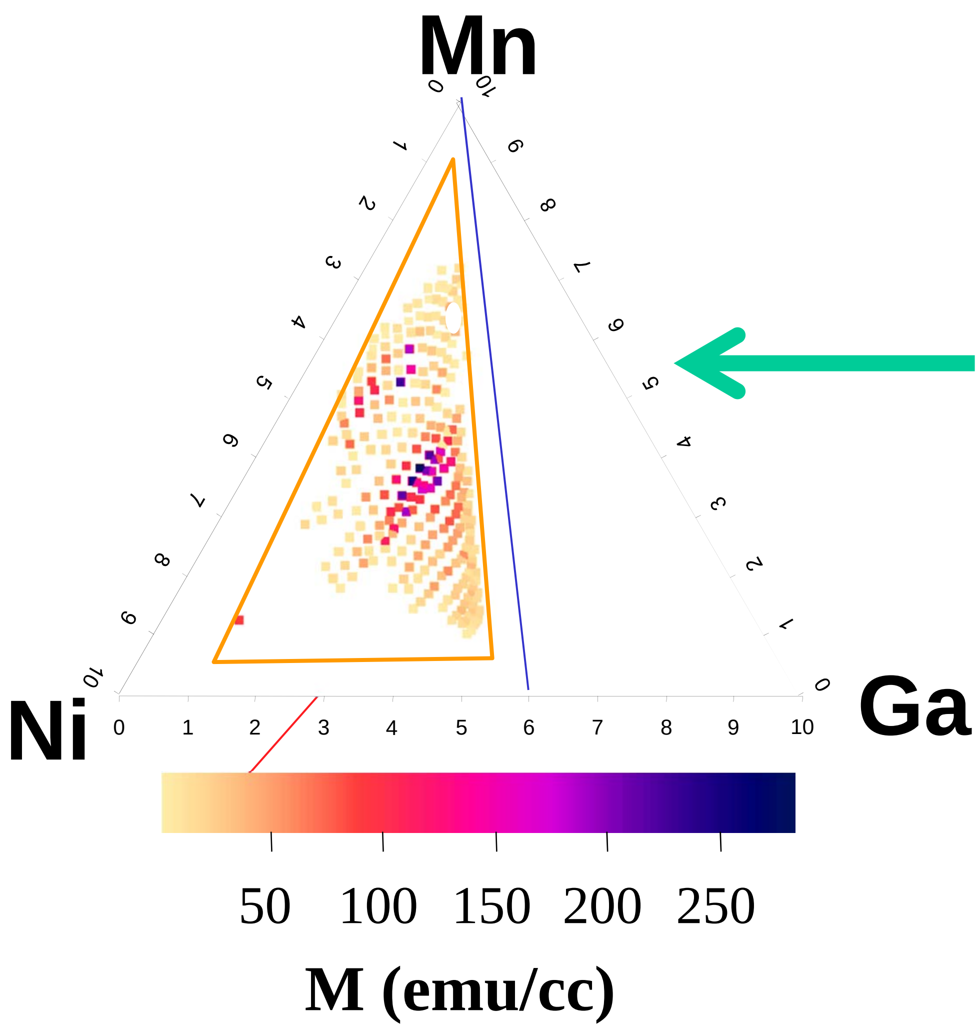
Room temperature samples are measured

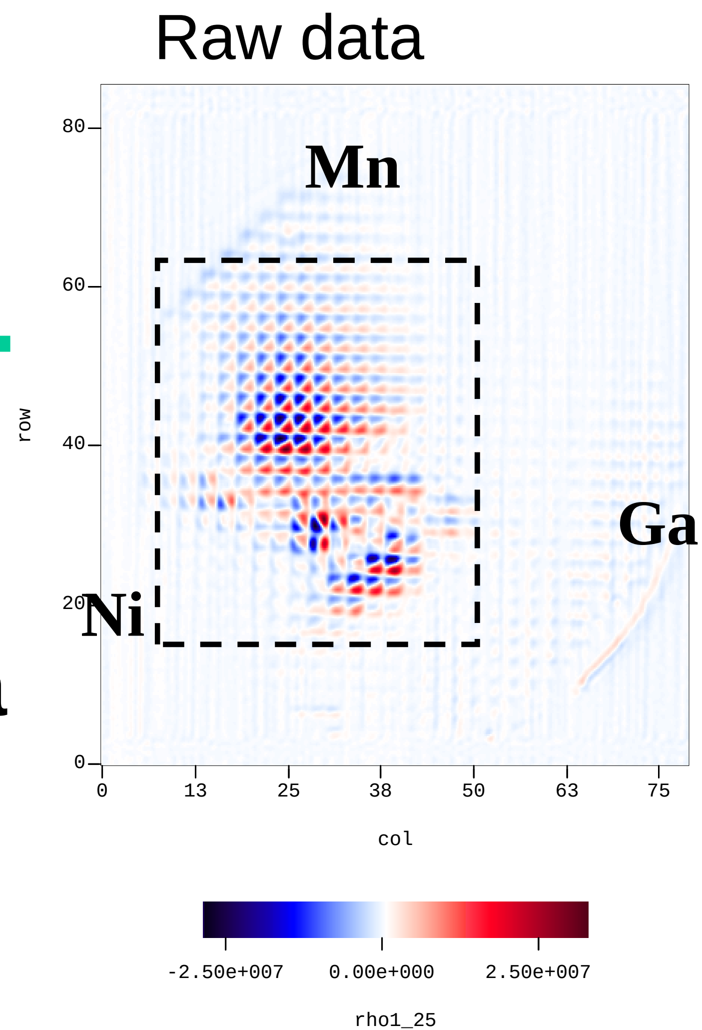
Raw data



Nature Materials **2**, 180 (2003)

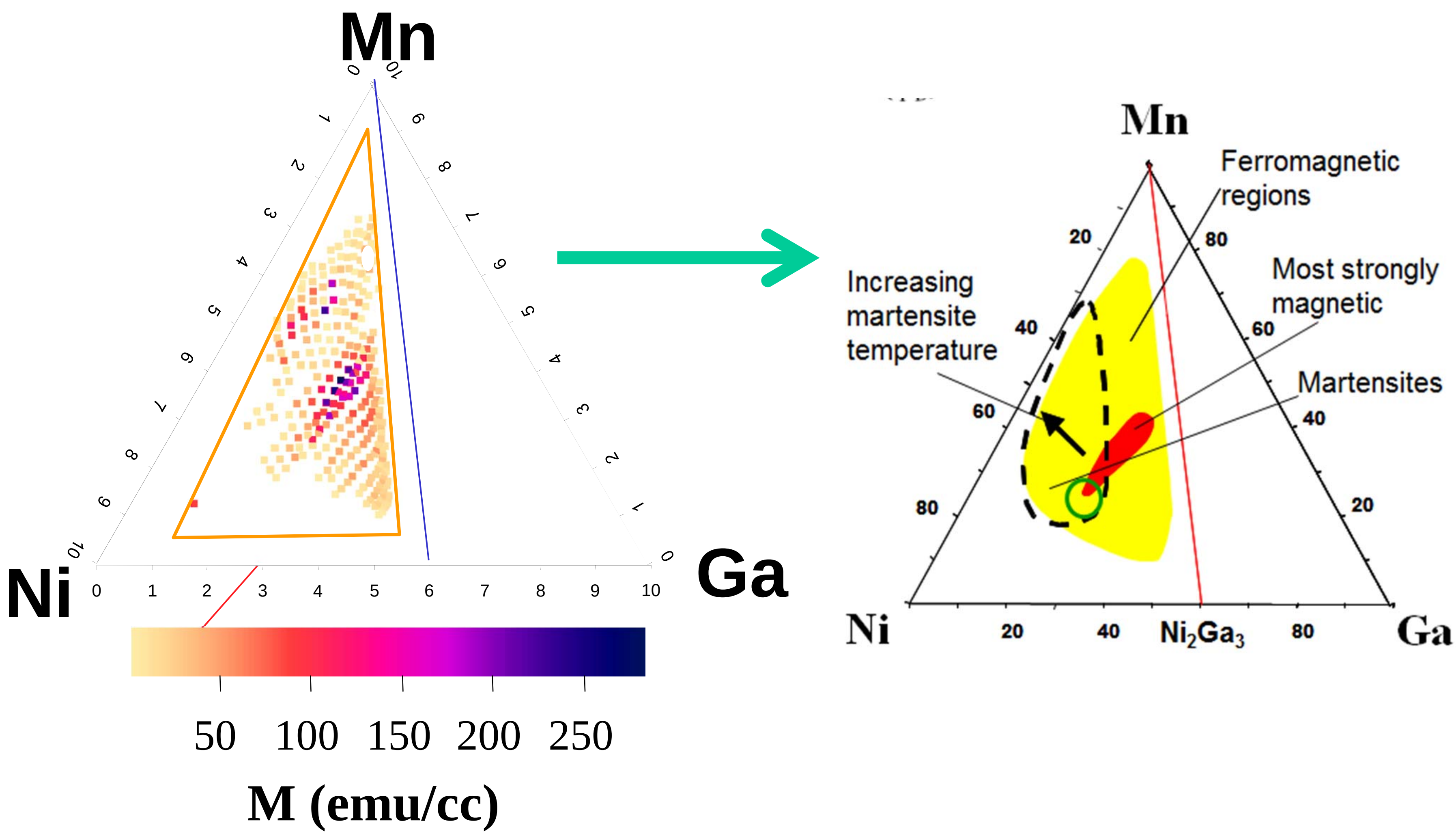
Rapid mapping of magnetic properties: scanning SQUID





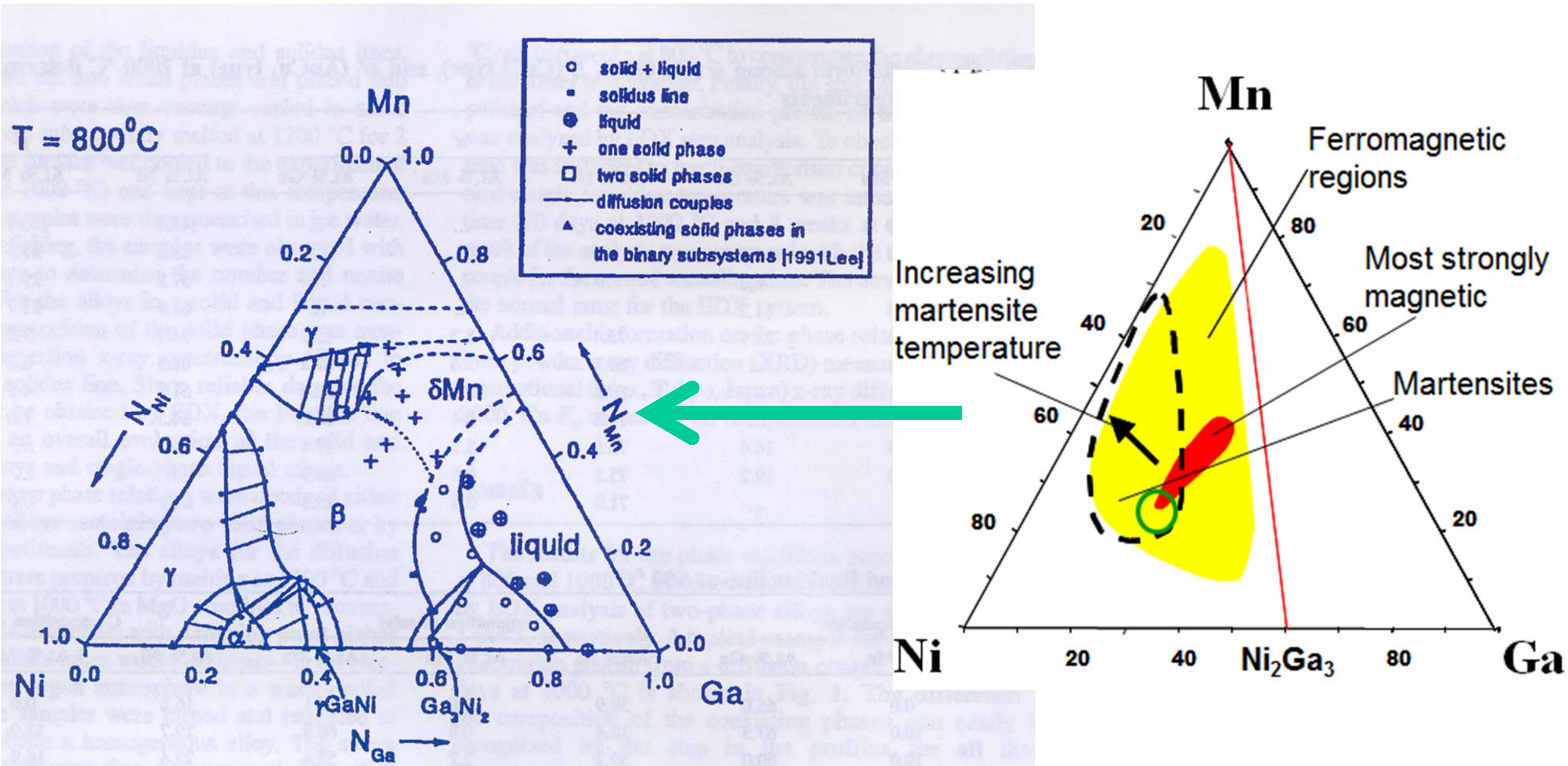
Nature Materials **2**, 180 (2003)

Rapid mapping of magnetic properties: constructing functional phase diagram



Nature Materials **2**, 180 (2003)

Rapid mapping of magnetic properties: comparison with phase diagram



C. Wedel and K. Itagaki,
Journal of Phase Equilibria 22, 324 (2001)

Nature Materials **2**, 180 (2003)

Targeting superconductors predicted by theory

Possible high-temperature superconductors predicted from electronic structure and data-filtering algorithms

M. Klintenberg, O. Eriksson *

Department of Physics and Astronomy, Uppsala University, Box 516, SE-75120 Uppsala, Sweden

Computational Materials Science 67 (2013) 282–286

PRL **105,** 217003 (2010) PHYSICAL REVIEW LETTERS week ending 19 NOVEMBER 2010

Prediction: FeB_4 is a superconductor with T_c ~ 15-20 K

New Superconducting and Semiconducting Fe-B Compounds Predicted with an *Ab Initio* Evolutionary Search

A. N. Kolmogorov,[1] S. Shah,[1] E. R. Margine,[1] A. F. Bialon,[2] T. Hammerschmidt,[2] and R. Drautz[2]

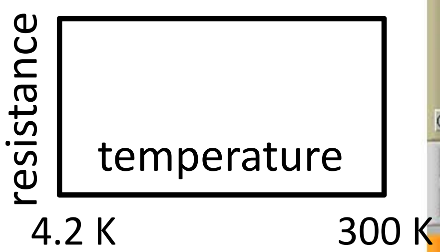
[1]*Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, United Kingdom*

[2]*Atomistic Modelling and Simulation, ICAMS, Ruhr-Universität Bochum, D-44801 Bochum, Germany*

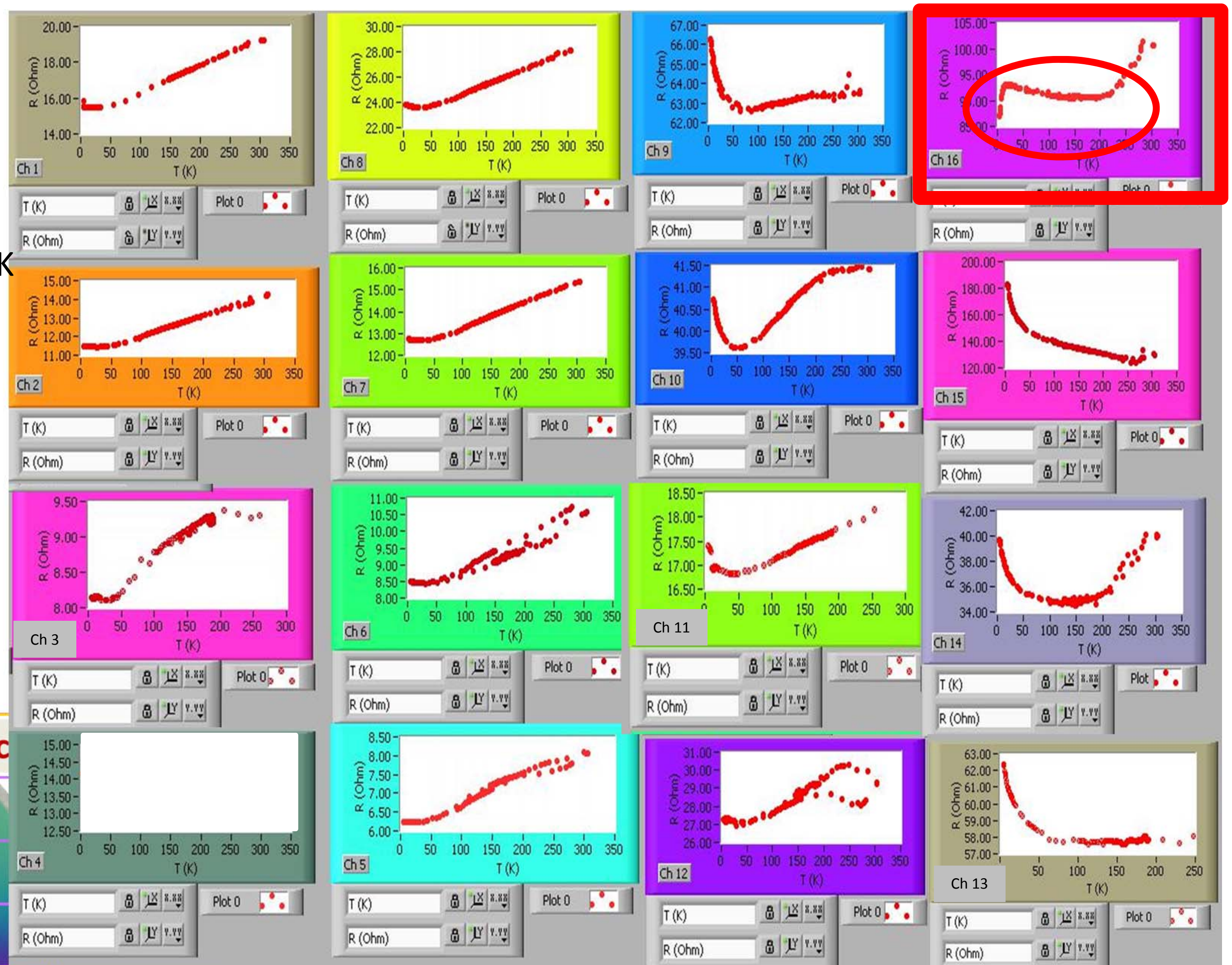
(Received 20 April 2010; published 19 November 2010)

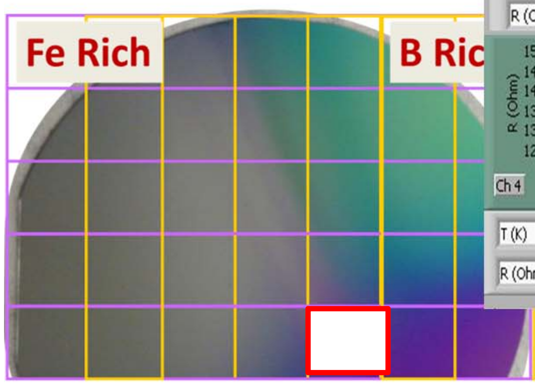
New candidate ground states at 1:4, 1:2, and 1:1 compositions are identified in the well-known Fe-B system via a combination of *ab initio* high-throughput and evolutionary searches. We show that the proposed *oP*12-FeB_2 stabilizes by a break up of 2D boron layers into 1D chains while *oP*10-FeB_4 stabilizes by a distortion of a 3D boron network. The uniqueness of these configurations gives rise to a set of remarkable properties: *oP*12-FeB_2 is expected to be the first semiconducting metal diboride and *oP*10-FeB_4 is shown to have the potential for phonon-mediated superconductivity with a T_c of 15–20 K.

Fe-B composition spread: FeB_x(x =2-4), 16 spots on one 1 cm^2 chip



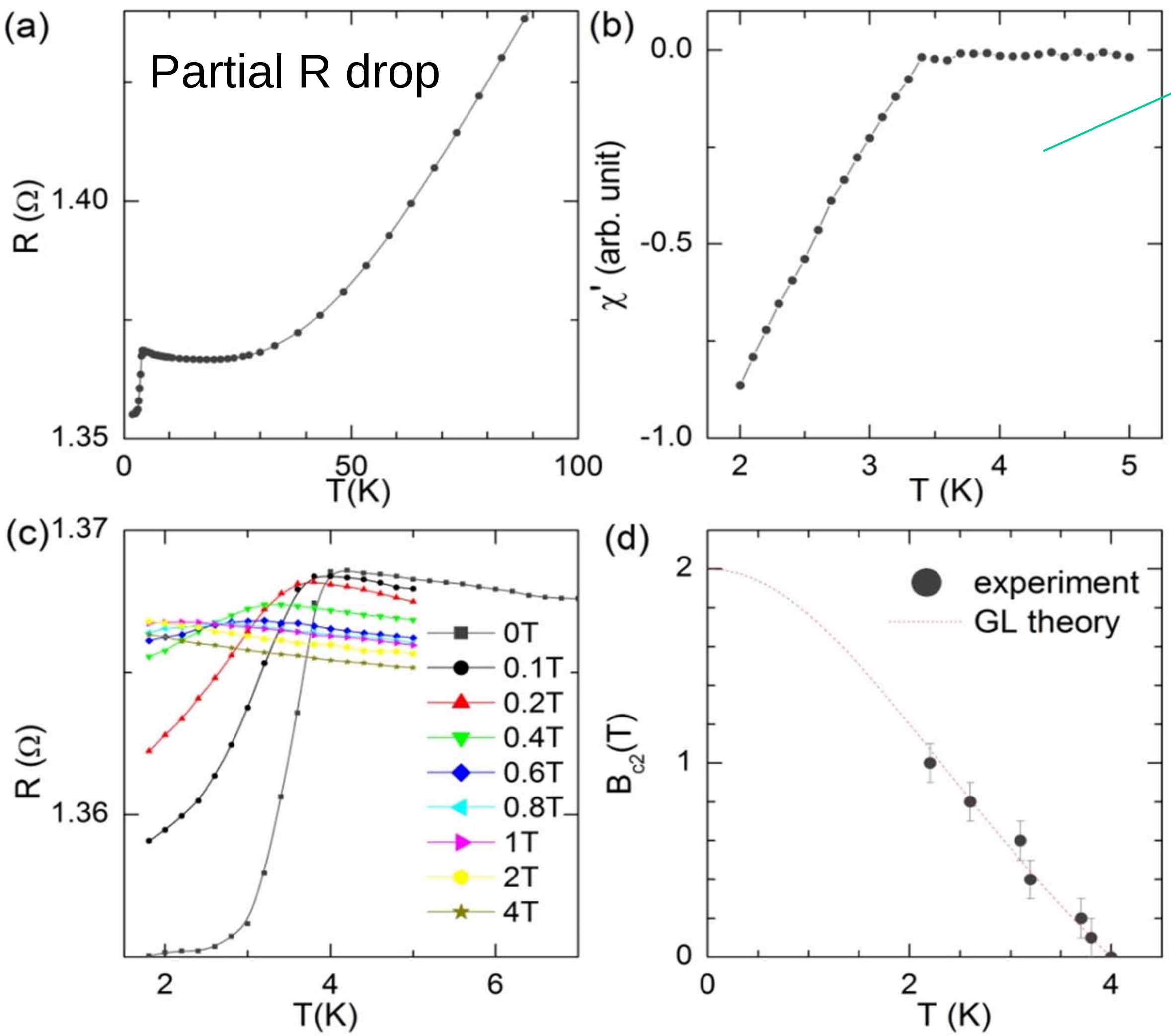
Middle region:
$FeB_2 – FeB_4$



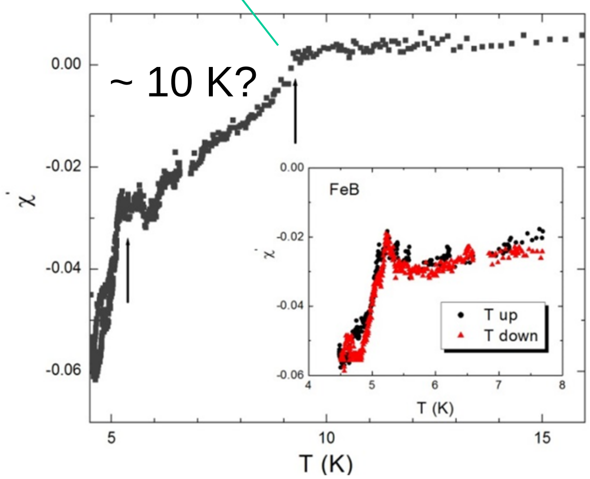


more Fe ⟷ more B

FeBx: it looks like a real superconductor



Susceptibility shows diamagnetism



$B_{c2}(T) = B_{c2}(0)[1-(T/T_c)^2]/ [1+(T/T_c)^2]$

gives $B_{c2}(0)$ = 2 T

-> Type II BCS superconductor

Superconducting phase was detected in 2 spread wafers

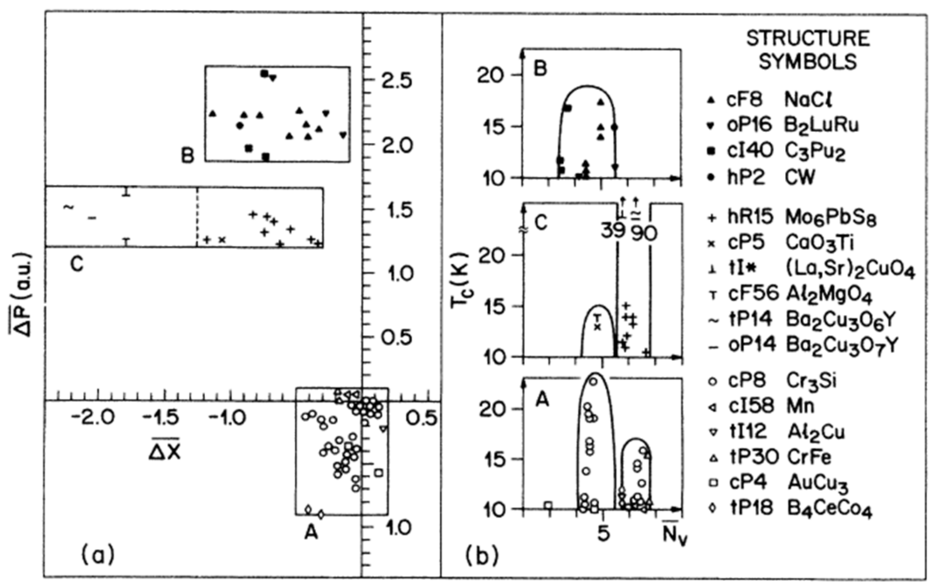
APL Materials 1, 042101 (2013)

Mining superconducting databases

PHYSICAL REVIEW B VOLUME 37, NUMBER 4 1 FEBRUARY 1988

Quantum structural diagrams and high-T_c superconductivity

P. Villars* and J. C. Phillips



	Cr*	Mn*	Rh*	Fe	Co	Ni	Cu	Ag	Au	Zn	Cd	Hg
Li		Cl F	Cl N	F Cl	F Cl	O Cl	O F	Cl	S Br Cl N	N O	N O	
Na		I Cl	I Cl Se	F I	F Se I	O Se I	O Te F Se	Cl Se Te	S Br Cl Se	N Te O Se	N Te O Se	Se
K				F	F	O	O F	Cl	S Br Cl	N O	N	
Rb	I			F	F	O	O F	Cl	S Br Cl	N O	N	
Cs				F I	F	O	O F	Cl	S Br Cl	N O		
Be	F	F	O P	F	O F	O F	N O	N O	N S O	N O	N O	N O
Mg	F Br Cl	Cl Br	Cl Br	F Br Cl	O Cl F Br	O Br Cl	N O	S	S	N O	N O	
Ca	I Br	I Cl	I	F Cl I	O I F	O I	N O			N O	N O	
Sr				F	O F	O	N O		S	N O	N	
Ba				F Br	O F	O	N O		S	N O	N	

Superconductors grouped by 3 golden coordinates

“Predictions of new compounds”

Based on ~ 600 superconductors (1988)

Visualization and mining of NIMS SuperCon database

MatNavi
NIMS Materials Database

	A	B		
28988	153347	Al20.3V2		
28989	153348	Ga0.2V2Al20		
28990	153349	Y1V2Al20		
28991	153350	La1V2Al20	La	1
28992	153351	C14H10Sr1.5	C	14
28993	153352	C14H10Ba1.5	C	14
28994	153353	C22H14K2.9	C	22
28995	153354	C22H14K3.3	C	22
28996	153355	C22H14Rb3.1	C	22
28997	153358	Sr0.8Eu0.2Fe1.78Co0.22As2	Sr	0.8
28998	153359	Sr0.54Eu0.46Fe1.78Co0.22As2	Sr	0.54
28999	153360	Tl0.58Rb0.42Fe1.72Se2	Tl	0.58
29000	153361	Fe1Se0.98	Fe	1
29001	153362	Fe1Se0.94	Fe	1
29002	153363	La1Fe1As1O0.89	La	1
29003	153364	Pr1Fe1As1O0.896	Pr	1
29004	153365	Nd1Fe1As1O0.9	Nd	1
29005	153366	Tb1Fe1As1O0.8	Tb	1
29006	153689	U1Ni2.05Al2.98	U	1
29007	154544	Mo1Sn0.03Sr2Y1Cu2O8-z	Mo	1
29008	159192	Tm0.84Lu0.16Fe3Si5	Tm	0.84
29009	160000	Tl1Ba2Ca3Cu4O11	Tl	1
29010	181284	Nb0.8Pd0.2	Nb	0.8
29011	181287	Nb0.69Pd0.31	Nb	0.69
29012				

29000 entries in MatNavi (2014)

29000 entries

(Cuprates: more than 10000)

↓

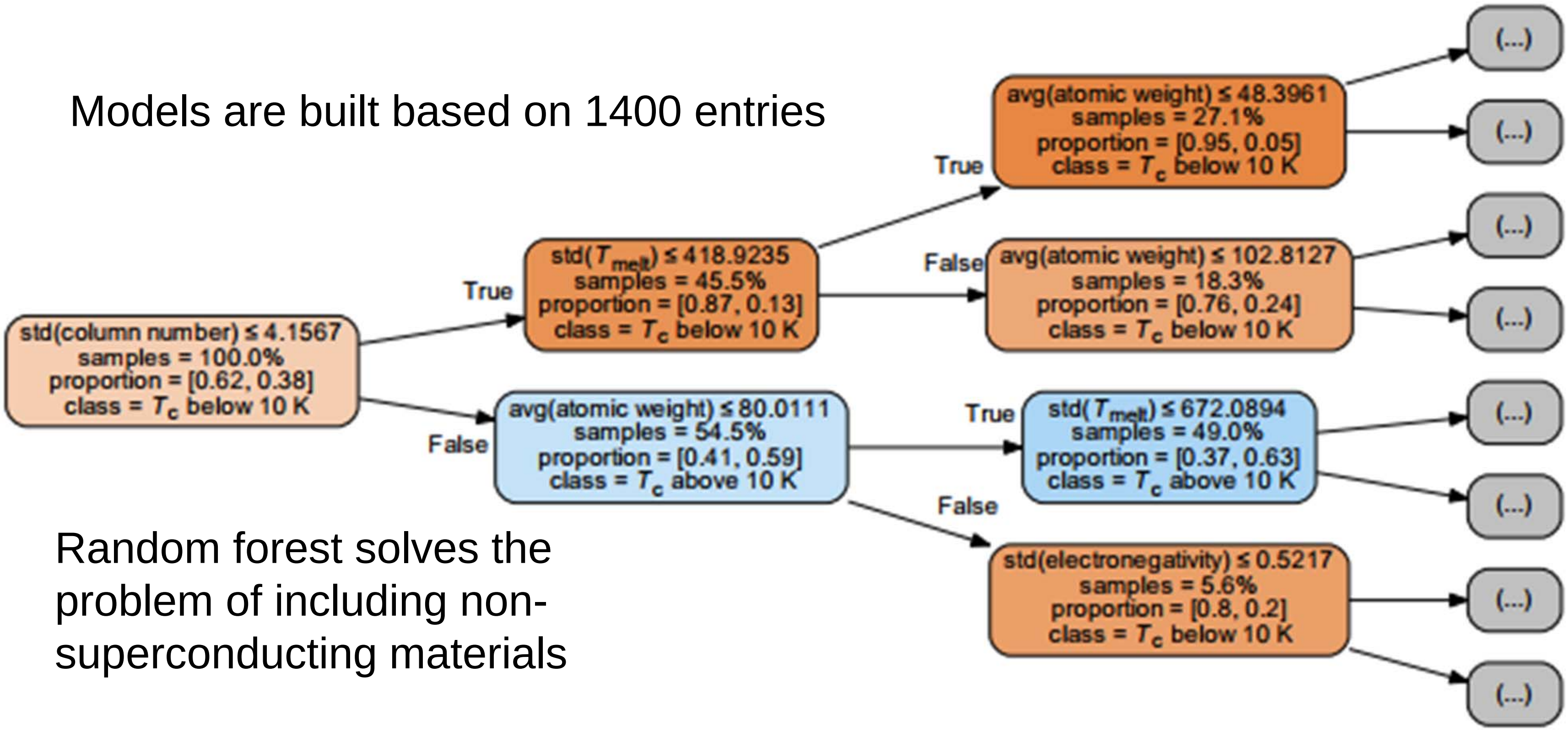
Removing misentries, duplicates, etc. results in 14000 entries

↓

Data mining via supervised machine learning using Magpie descriptors

Machine learning modeling of superconducting critical temperature (random forest)

Models are built based on 1400 entries

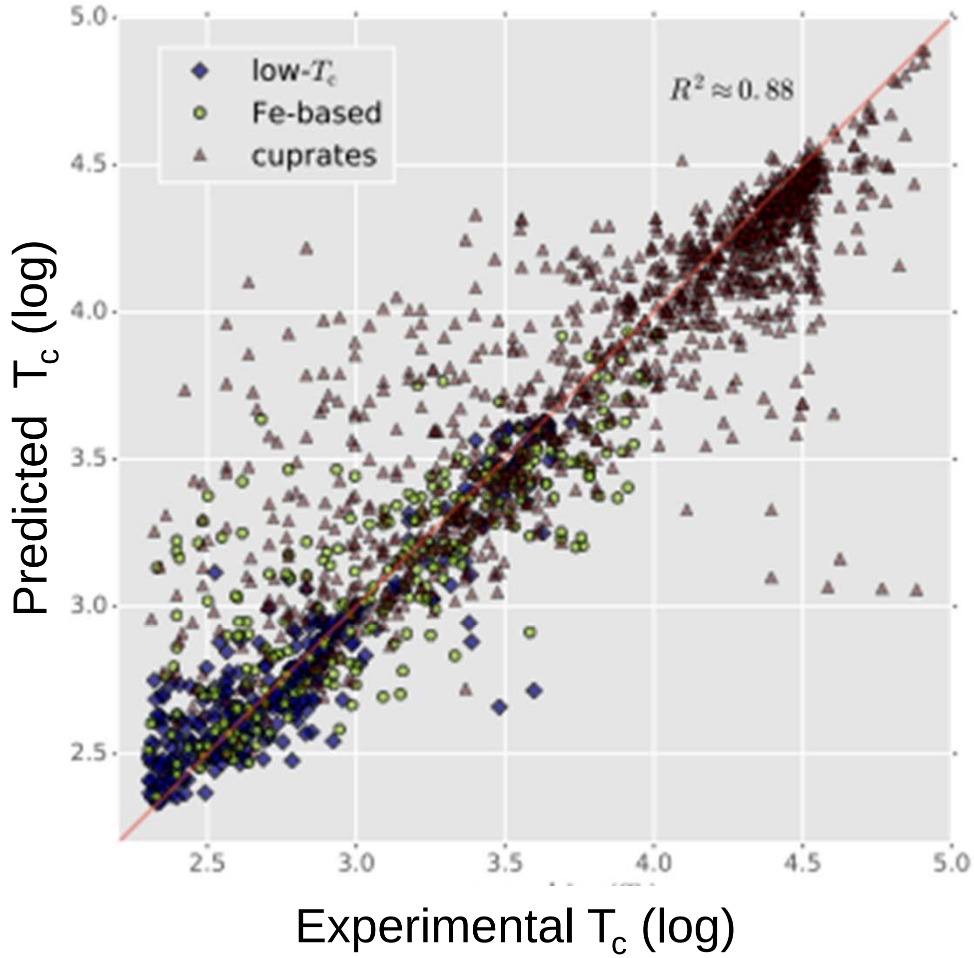


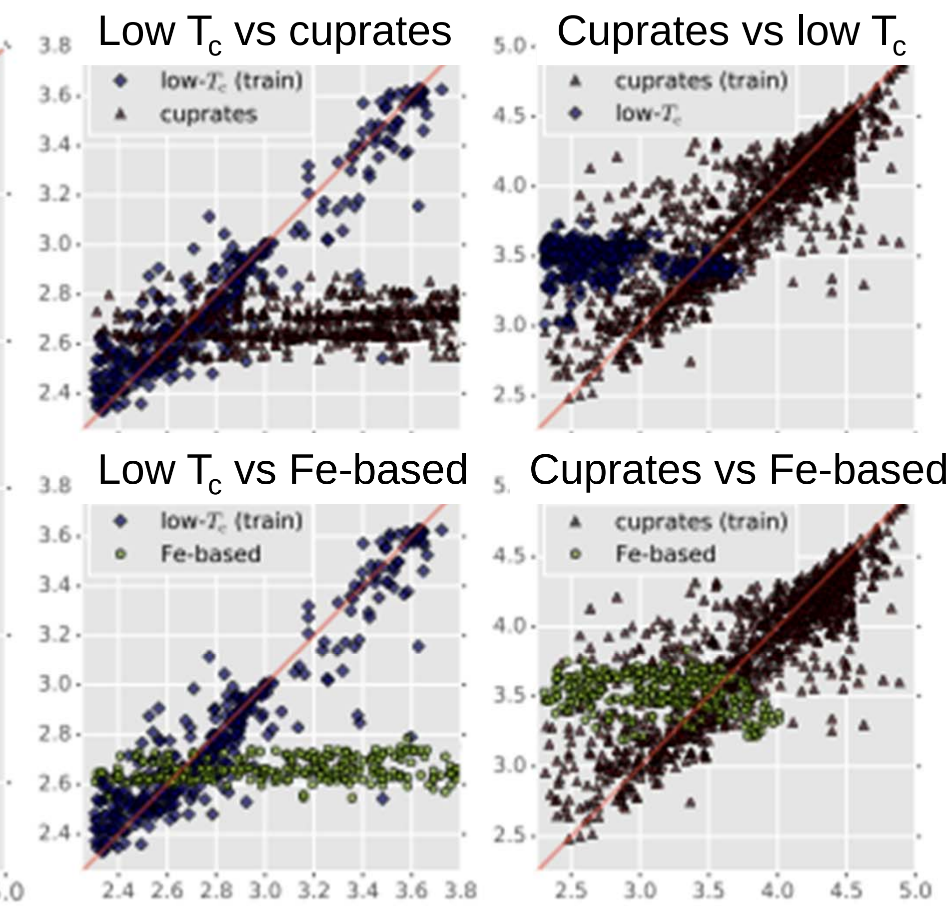
Random forest solves the problem of including non-superconducting materials

Stanev, et al.
arXiv preprint arXiv:1709.02727

Machine learning modeling of superconducting critical temperature

Overall prediction accuracy is pretty good

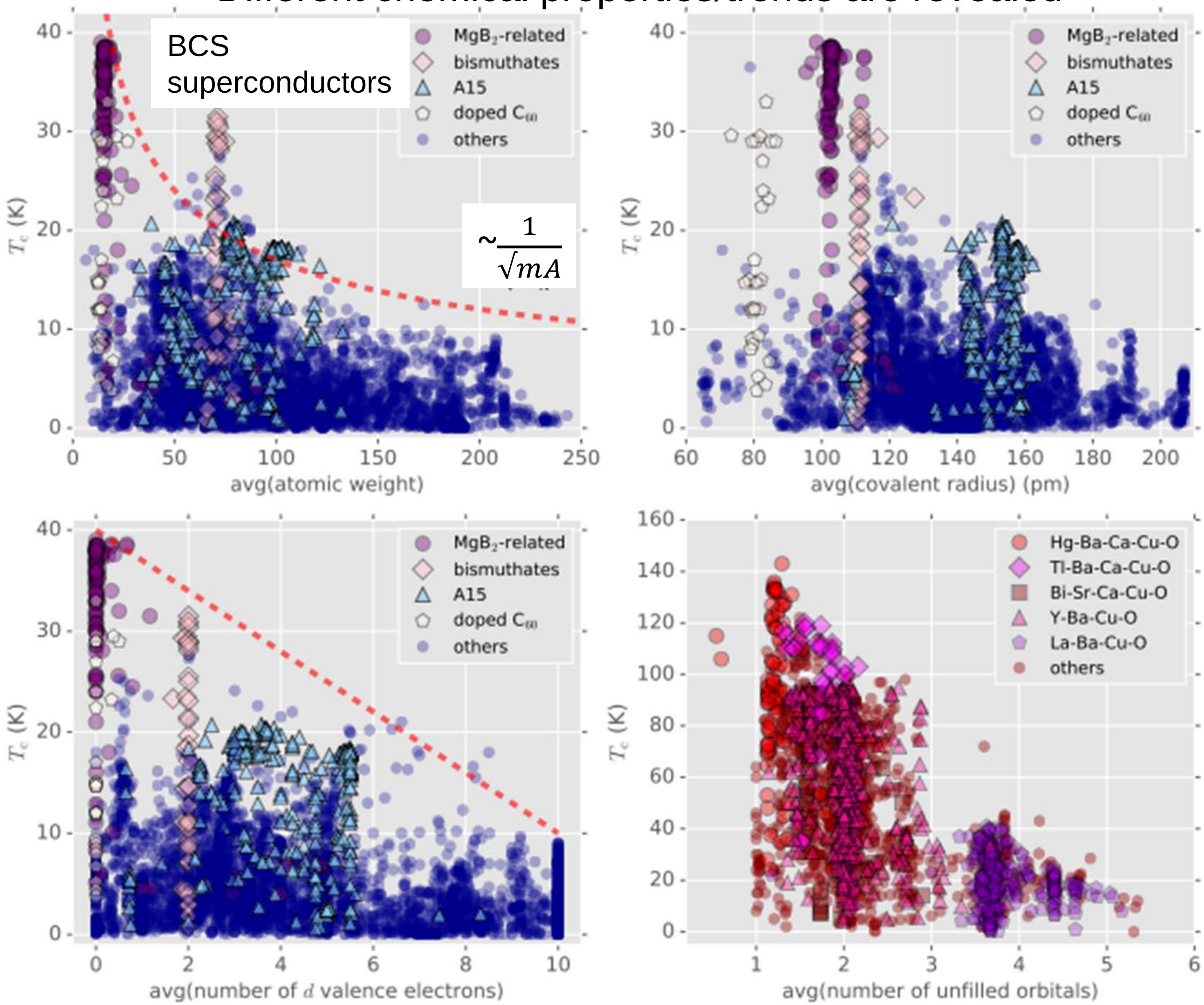




Different classes of superconductors
can be distinguished

Machine learning modeling of superconducting critical temperature

Different chemical properties/trends are revealed

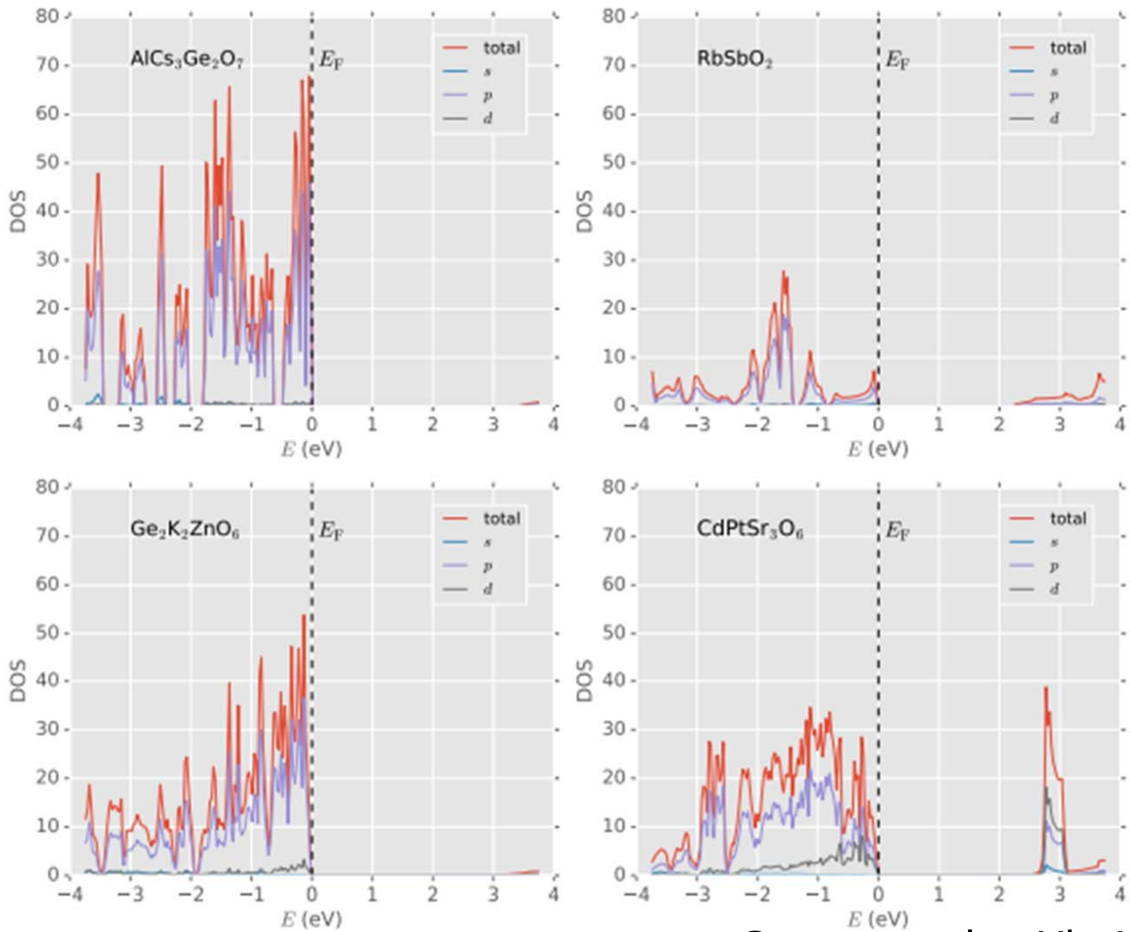


Machine learning modeling of superconducting critical temperature

compound	Predicted T_c range
CdK_2SiO_4	30 K - 40 K
$Cd_2IrNa_3O_6$	20 K - 30 K
$CdPtSr_3O_6$	30 K - 40 K
CdO_4Rb_2Si	30 K - 40 K
$Ge_2K_{16}Sr_4O_{36}$	30 K - 40 K
GeK_2ZnO_4	30 K - 40 K
$Ge_2K_2O_6Zn$	20 K - 30 K
$GeK_{0.6}Na_{0.4}ZnO_4$	20 K - 30 K
$PtSr_3ZnO_6$	30 K - 40 K
$KSbO_2$	30 K - 40 K
$RbSbO_2$	30 K - 40 K
$AlCs_3Ge_2O_7$	30 K - 40 K
$AsRbO_2$	30 K - 40 K
$AgAuBa_4O_6$	30 K - 40 K
$Au_2Sr_5O_8$	20 K - 30 K
$AsBeCsO_4$	20 K - 30 K
K_2SiZnO_4	20 K - 30 K
$FRbSeO_2$	20 K - 30 K
$CsFSeO_2$	20 K - 30 K
$K_2Si_2ZnO_6$	20 K - 30 K
$Ca_3Ge_6Na_6O_8$	20 K - 30 K
$CsSbO_2$	20 K - 30 K
$AgCrO_2$	20 K - 30 K
$K_4Na_2Tl_2O_6$	20 K - 30 K
$BCdCsO_3$	30 K - 40 K
CsF_3MoZnO_3	20 K - 30 K
$FKTeO_2$	20 K - 30 K
$HEu_{0.5}Ge_{1.5}K_{1.5}O_5$	20 K - 30 K
$K_{0.8}Li_{0.2}Sn_{0.76}O_2$	30 K - 40 K
$BiNa_3Ni_2O_6$	20 K - 30 K
$BiCa_2Na_3O_6$	20 K - 30 K
$Ba_5Br_2Ru_2O_9$	30 K - 40 K
$H_2Ge_3K_3TbO_{10}$	20 K - 30 K
$BaGe_3K_4O_9$	30 K - 40 K
$Ba_6Ga_7KZn_4O_{21}$	20 K - 30 K

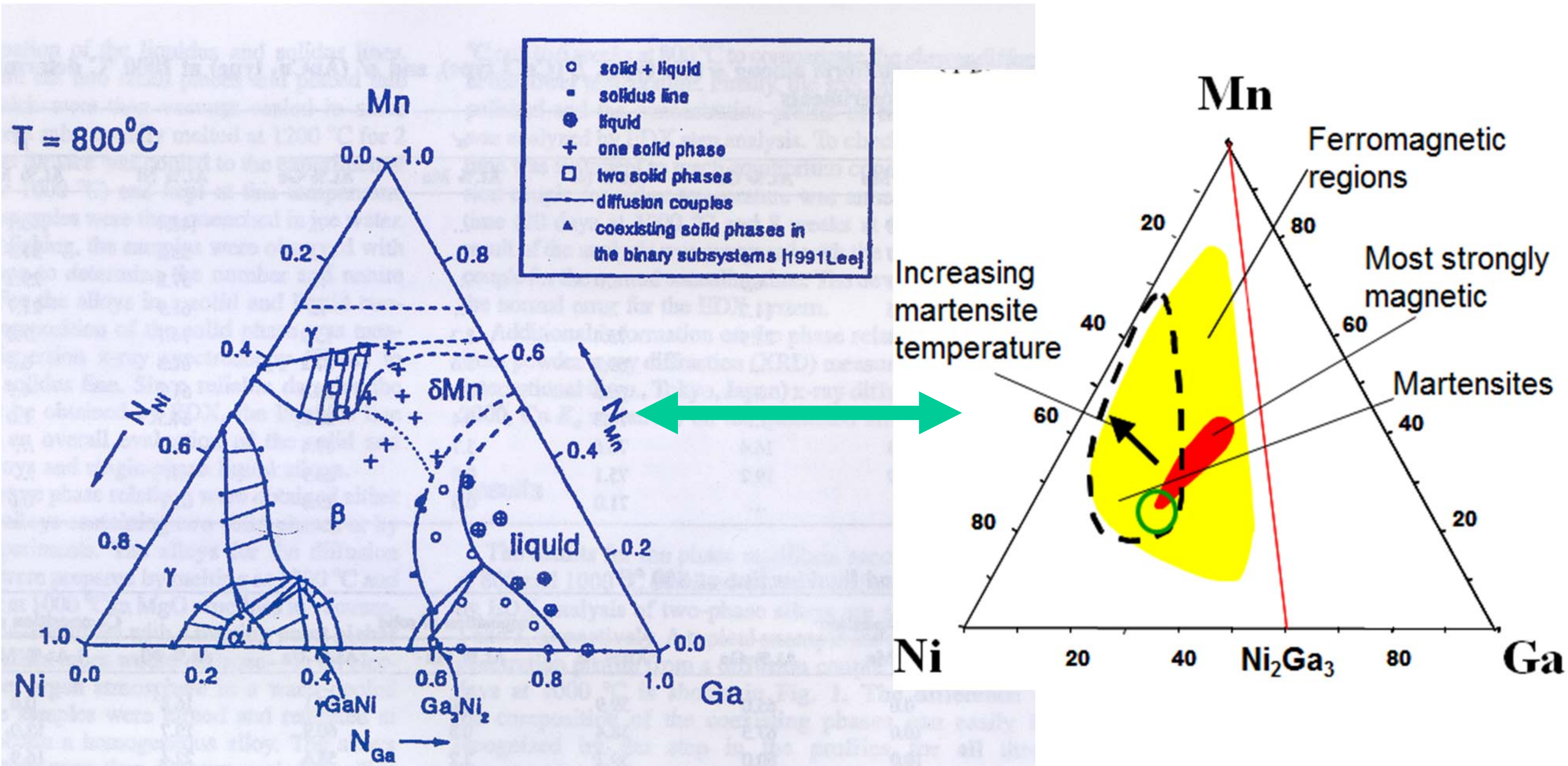
Combining SuperCon (experimental database), machine learning, ICSD, and AFLOW, we predict possible new superconductors

DOS of predicted superconductors show unusual structure



Stanev, et al. arXiv:1709.02727

Rapid mapping of structural phase diagram



C. Wedel and K. Itagaki,
Journal of Phase Equilibria 22, 324 (2001)

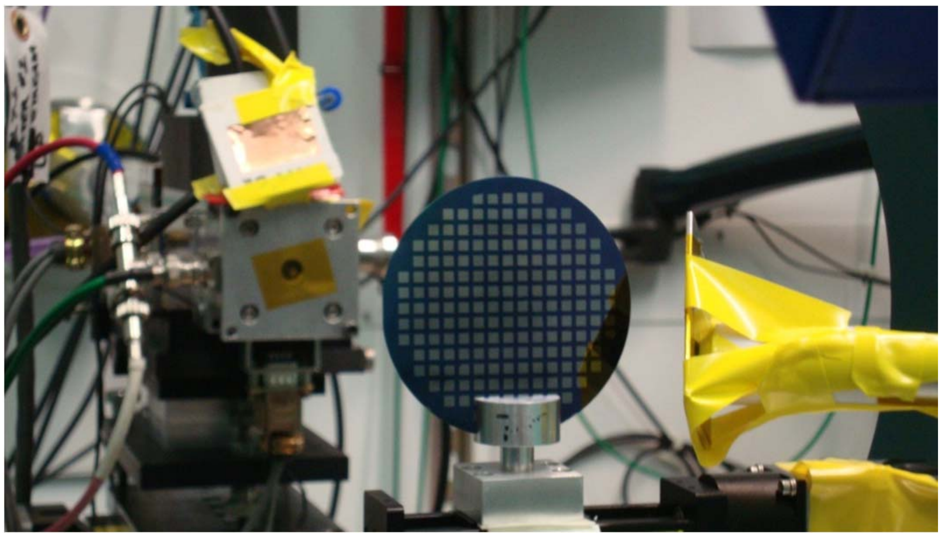
Nature Materials **2**, 180 (2003)

Structural property mapping of combinatorial wafers

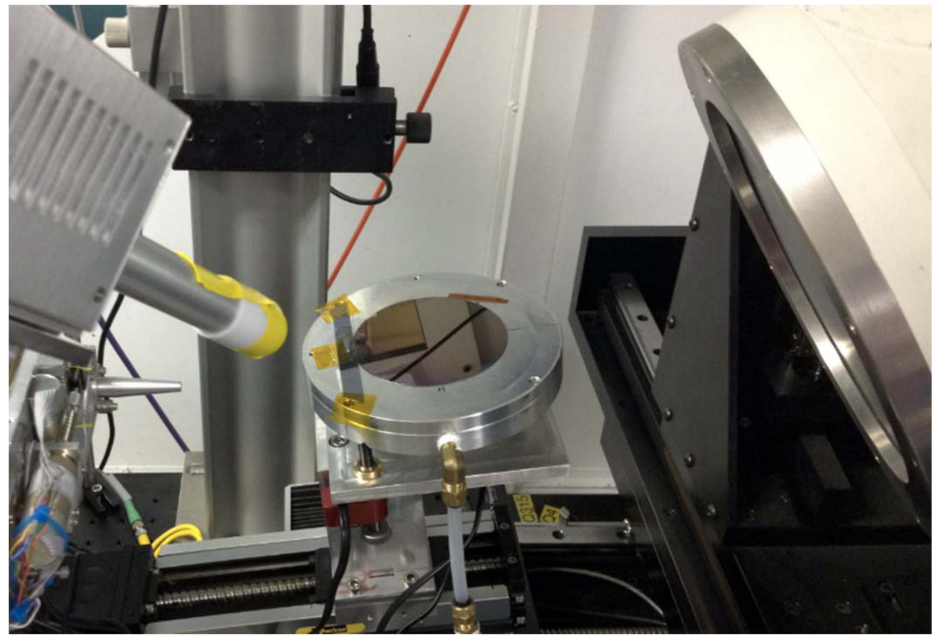
Synchrotron diffraction set up at SSRL

The entire 3” wafer (300 spots) can now be measured in 2 hrs

Transmission set up



Reflection set up

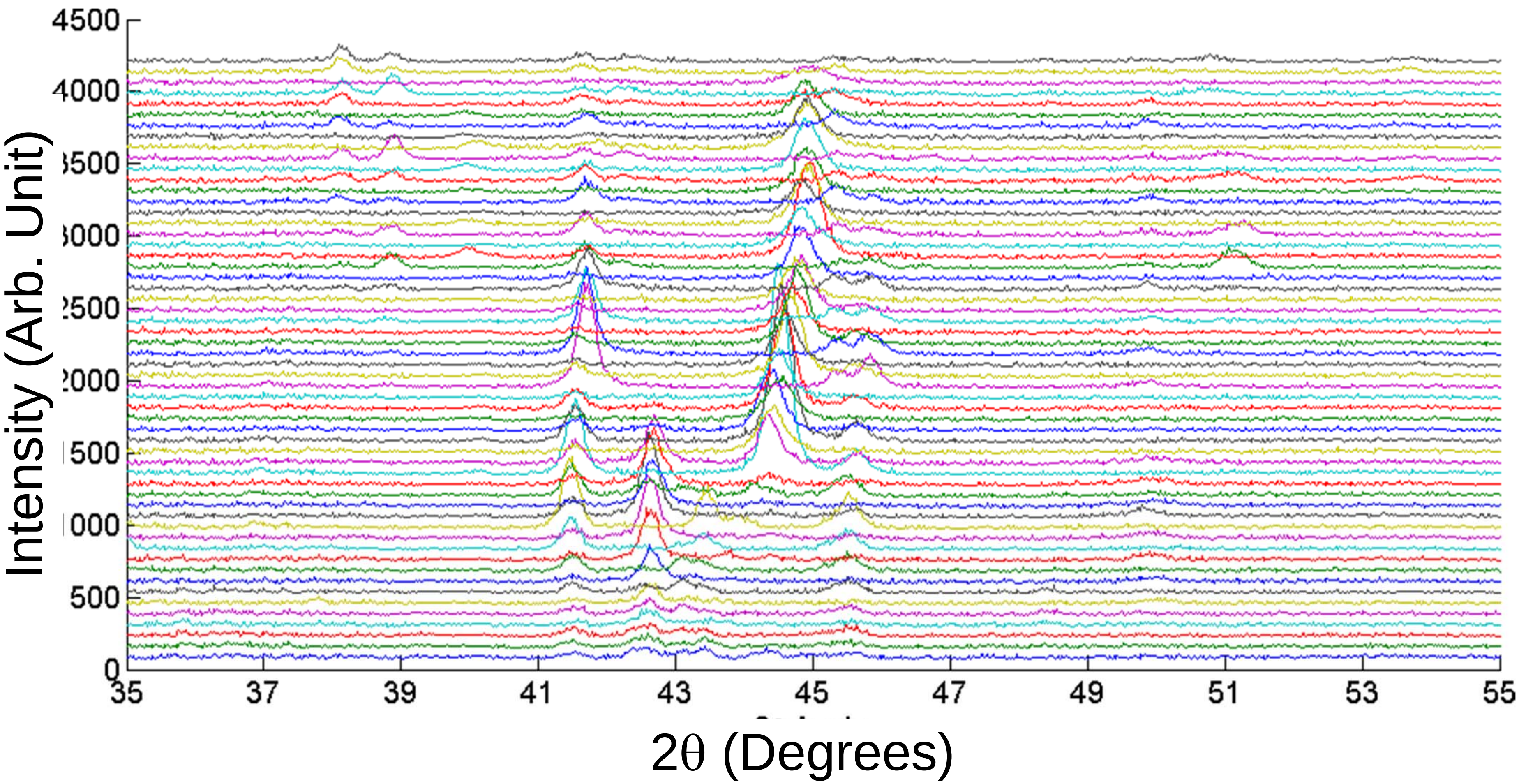


XRF carried out simultaneously

Each wafer produces: 300 MB to 2 GB of image data

w/ A. Mehta

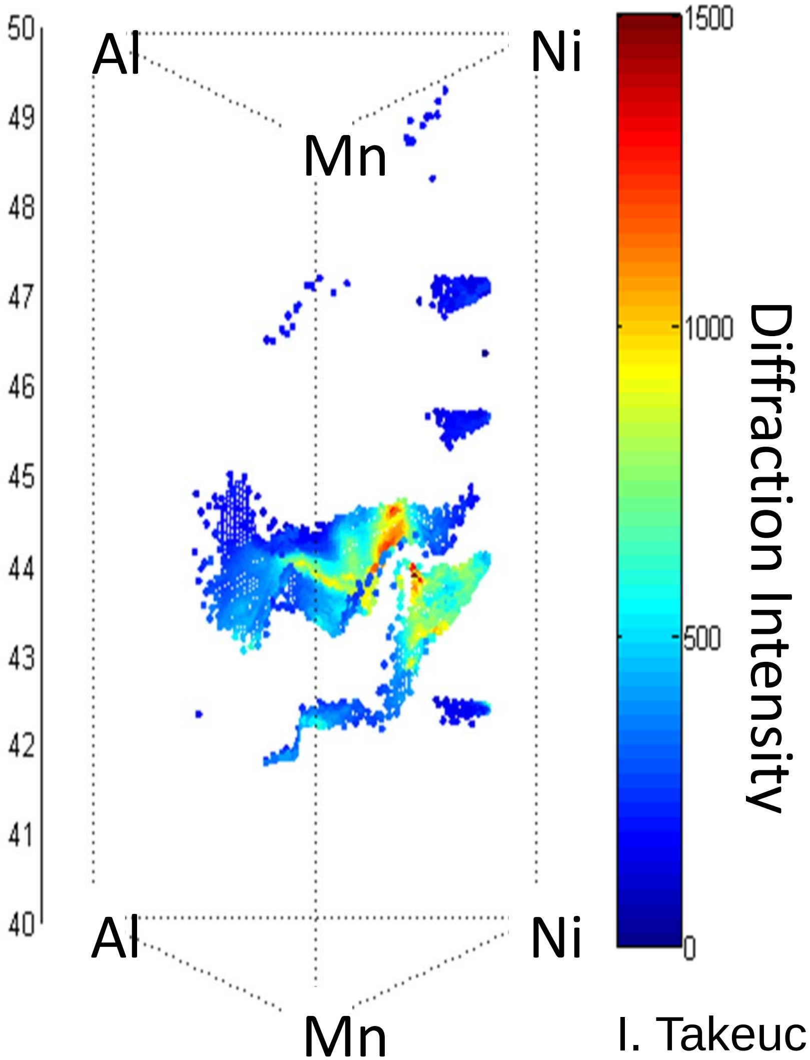
Hundreds of XRD Patterns are difficult to analyze by hand



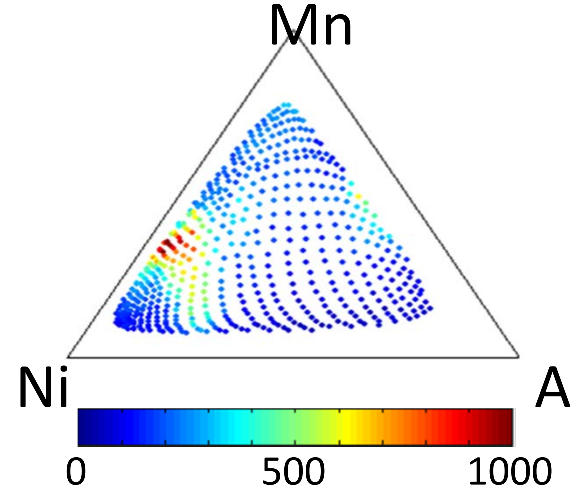
The same is true for any spectral data (Raman, FTIR, etc.)

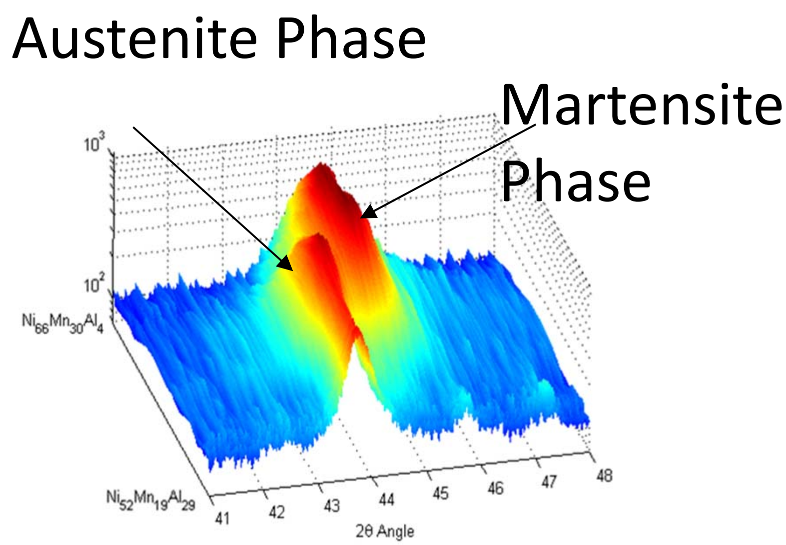
First step is visualization

Peak Angle and Intensity



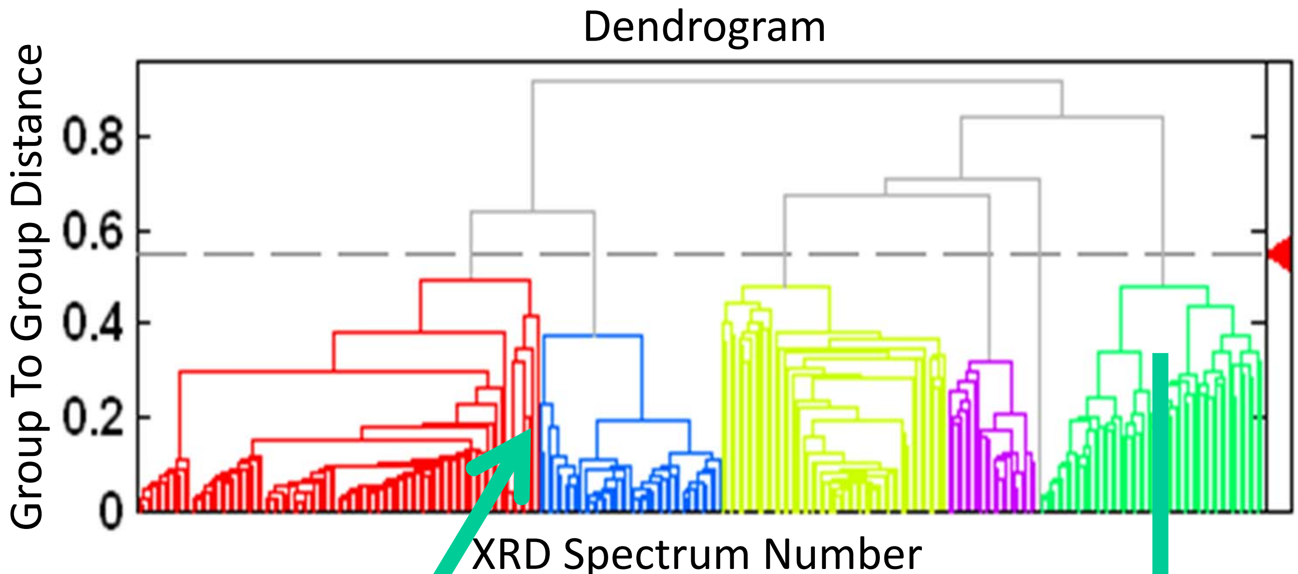
Intensity Cross Section

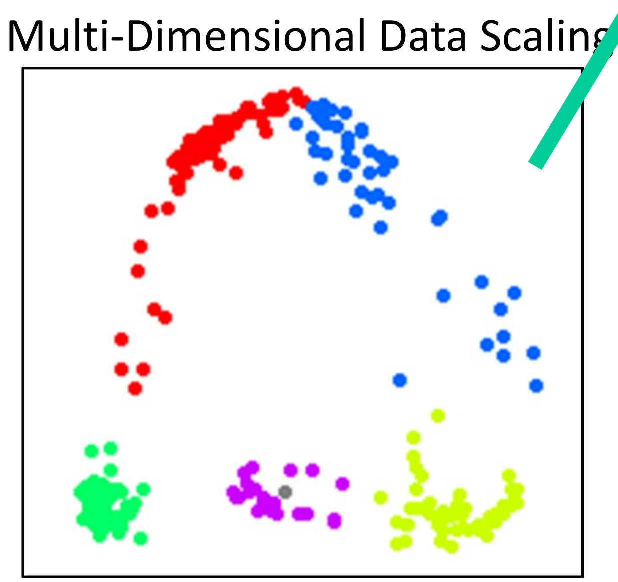


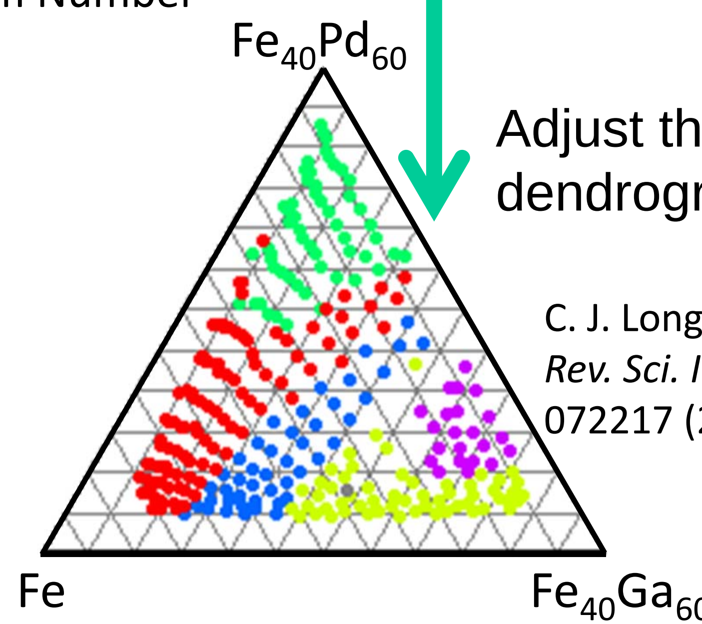


I. Takeuchi *et al*., Rev. Sci. Instrum. 76, 062223 (2005)

Analyze all spectra together using cluster analysis: Look for similarities between spectra





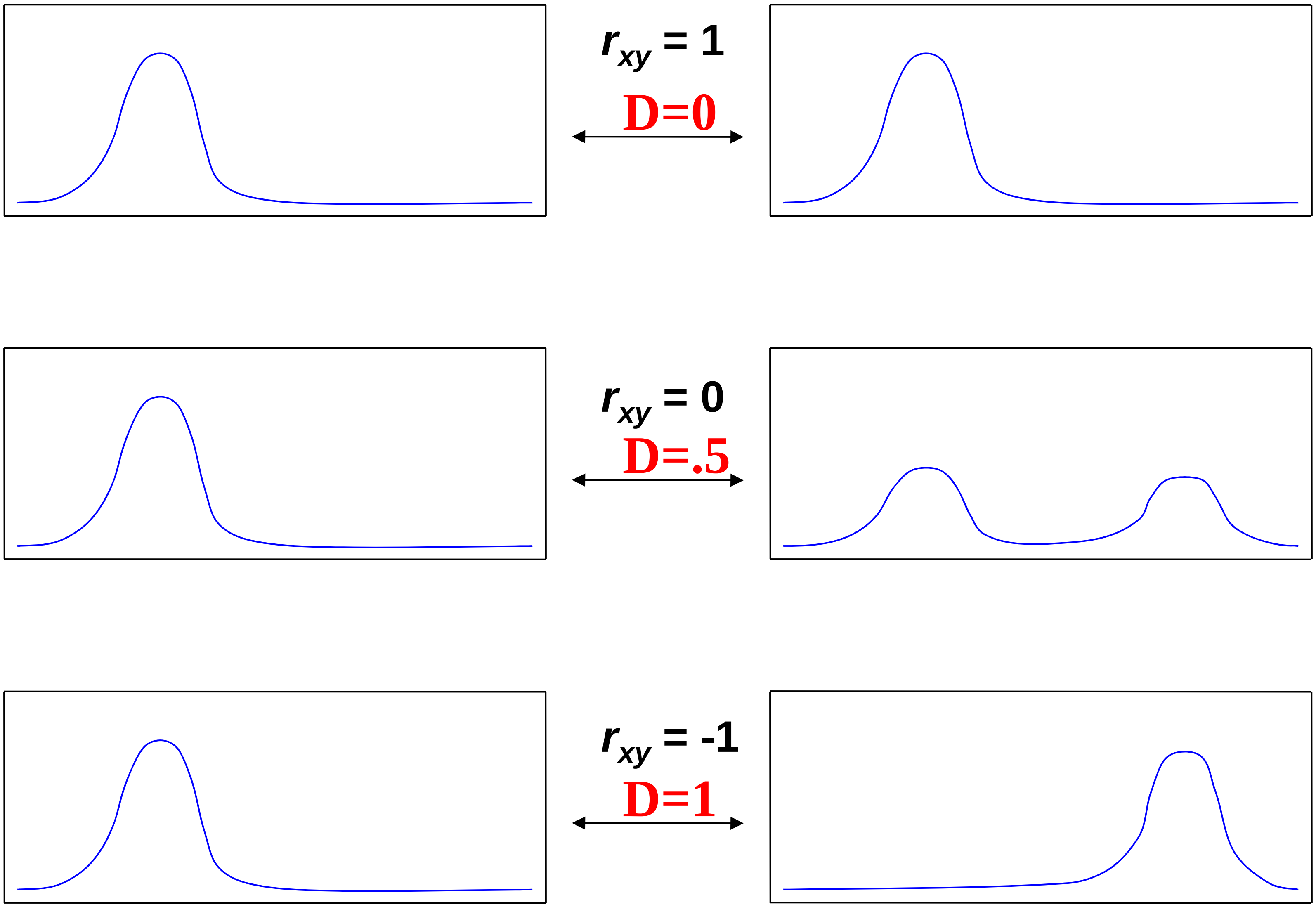


Adjust the dendrogram

C. J. Long, et al. *Rev. Sci. Instrum*. 78, 072217 (2007)

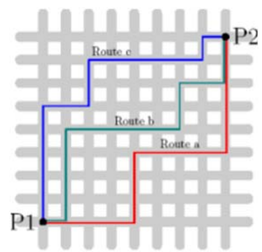
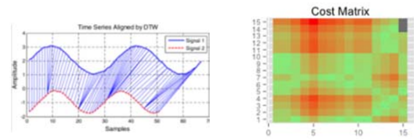
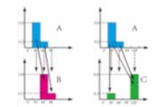
Initial construction of structural phase diagram

Comparing XRD patterns



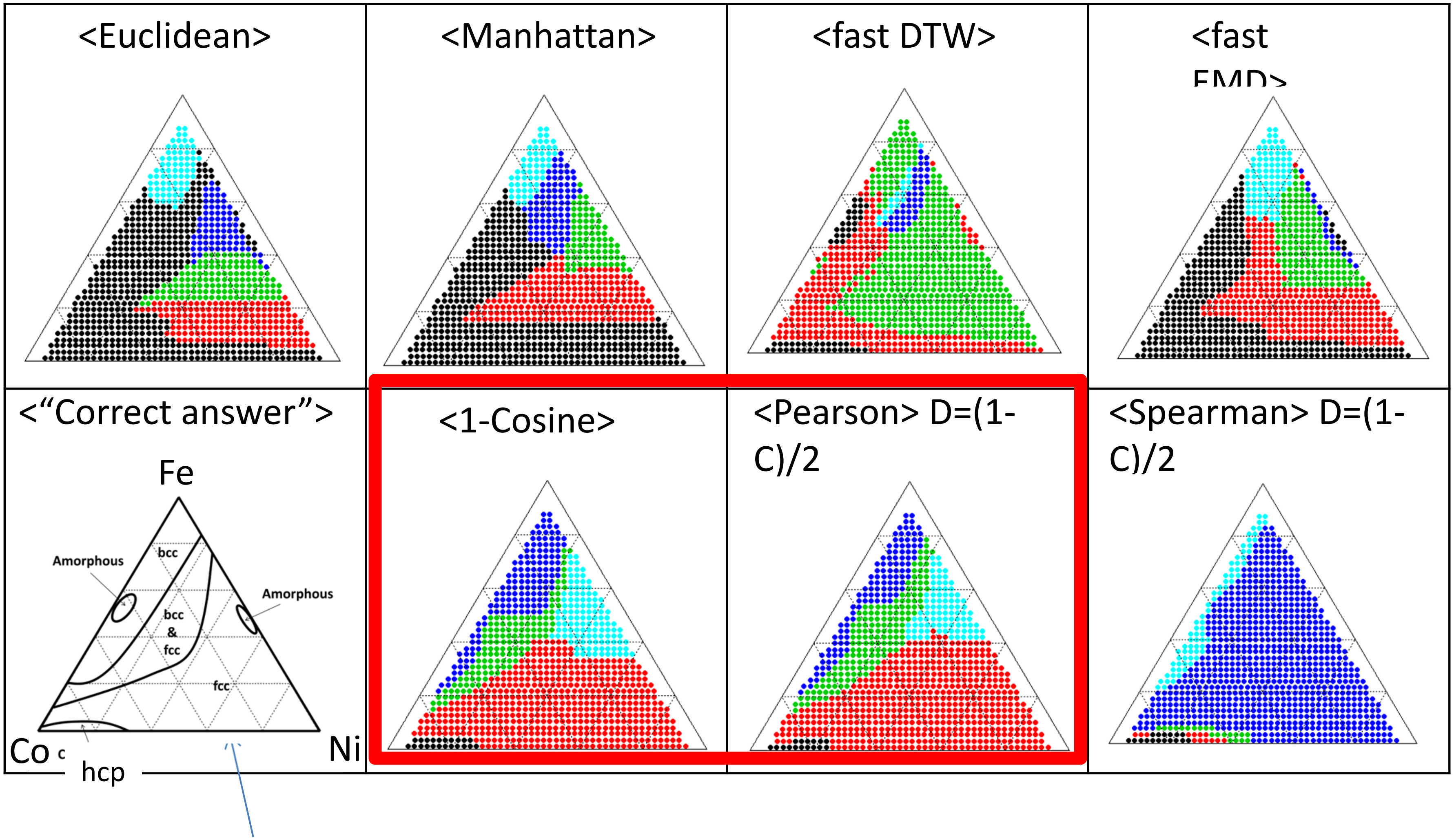
D = “distance”

Different metrics ("distance" between data points (vectors))

<Euclidean>	**<Manhattan>**	**<fast DTW>**	**<fast EMD>**
$d=\sqrt{\sum_k^n (x_k-y_k)^2}$	$d_1=\sum_k^n \lvert x_k-y_k\rvert$  	To find a minimum pass in the cost matrix  This is often used in speech recognition field	This is just transportation problem between two data series $EMD(P,Q)=\frac{\sum_{i=1}^m\sum_{j=1}^n d_{ij} f_{ij}^{\min}}{\sum_{i=1}^m\sum_{j=1}^n f_{ij}^{\min}}$  f is minimum transportation cost This is often used in image recognition field
	< 1 - Cos> D= 1-C $C_{\sin}=\frac{\sum x_i y_i}{\sum x_i^2 \sum y_i^2}$ This is often used in text mining field	**<Pearson> D=(1-C)/2** Known as normal correlation coefficient (parametric) $C_{cor}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$	**<Spearman> D=(1-C)/2** This is non-parametric index. d_i is difference between ranking of data $C_{spearman}=1-\frac{6\sum_i d_i}{N^3-N}$

Iwasaki et al., npj Computational Materials 3 (1), 4

Different metrics calculated and clustered using hierarchical clustering: XRD patterns from Fe-Co-Ni composition spread

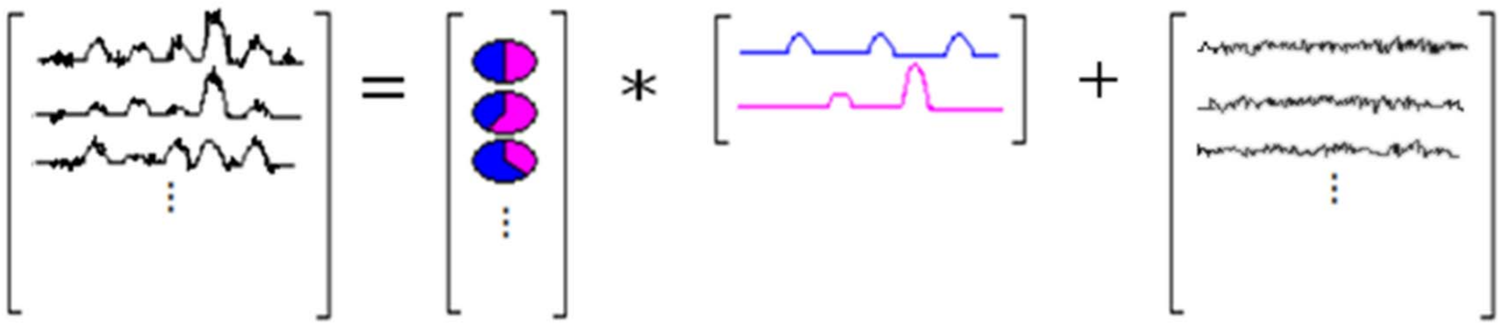


Y.K. Yoo et al. / Intermetallics 14 (2006) 241–247

Iwasaki et al., npj Computational Materials 3 (1), 4

Another analysis method:

Non-Negative Matrix Factorization (NMF):
(The Basic Idea)

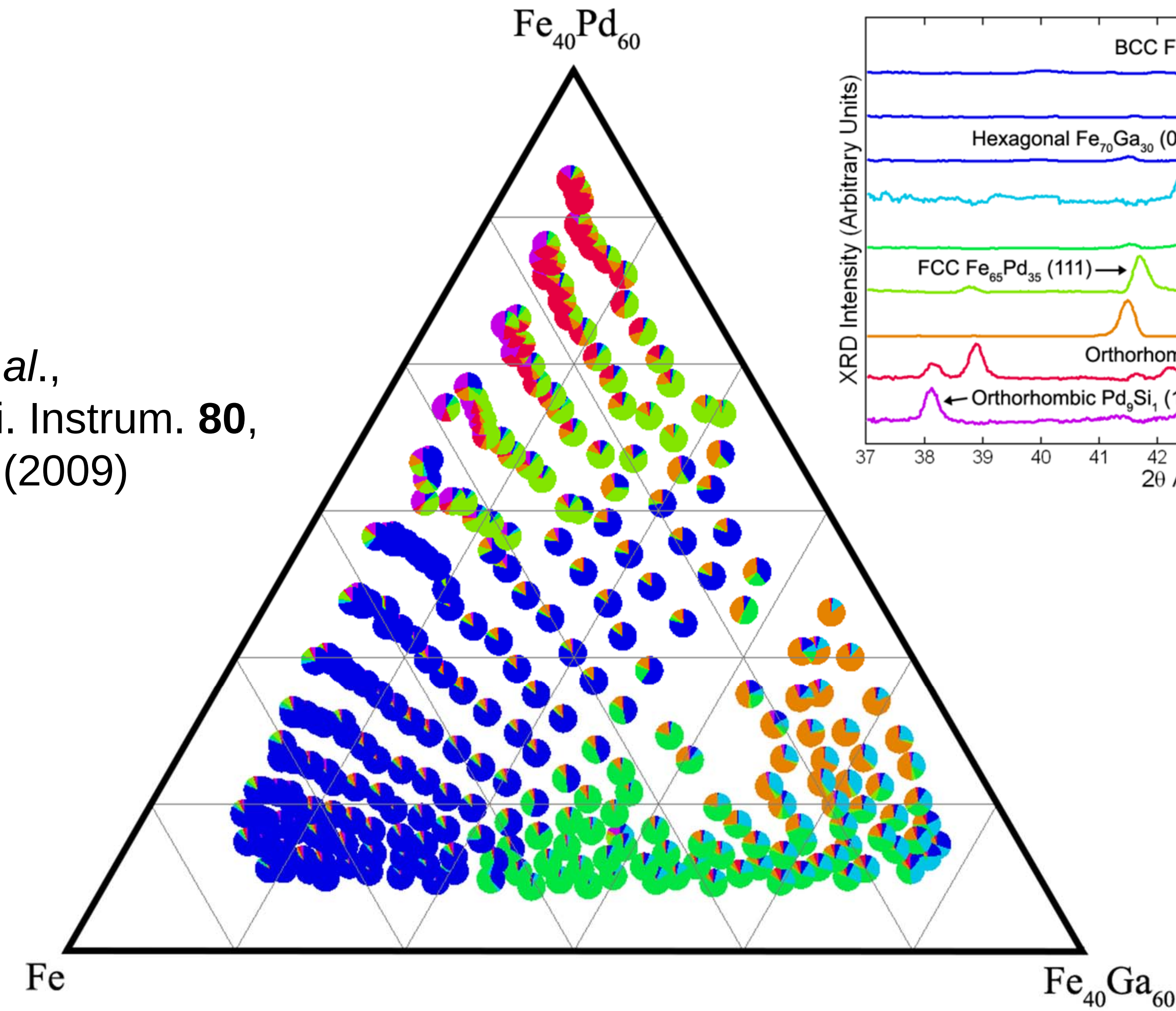


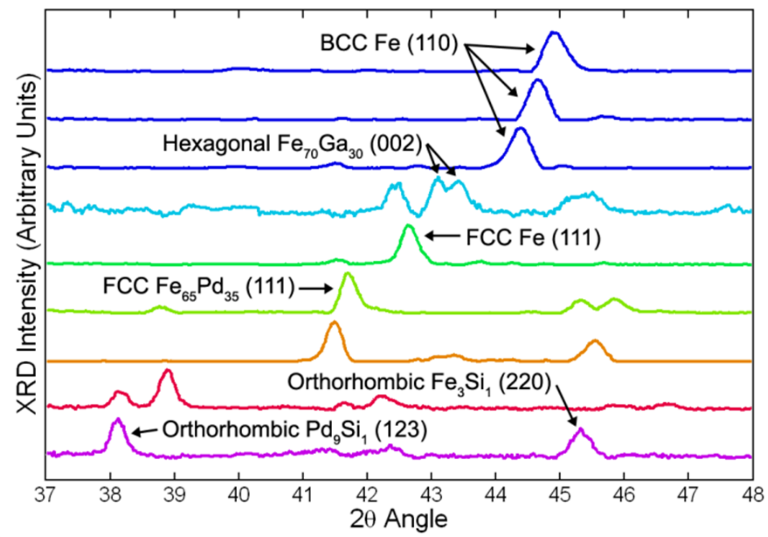
$$S = W * B + E$$

Experimental Spectra — **Weights** — **Basis Spectra** — **Residual Error**

Experimental Spectra are Deconvolved

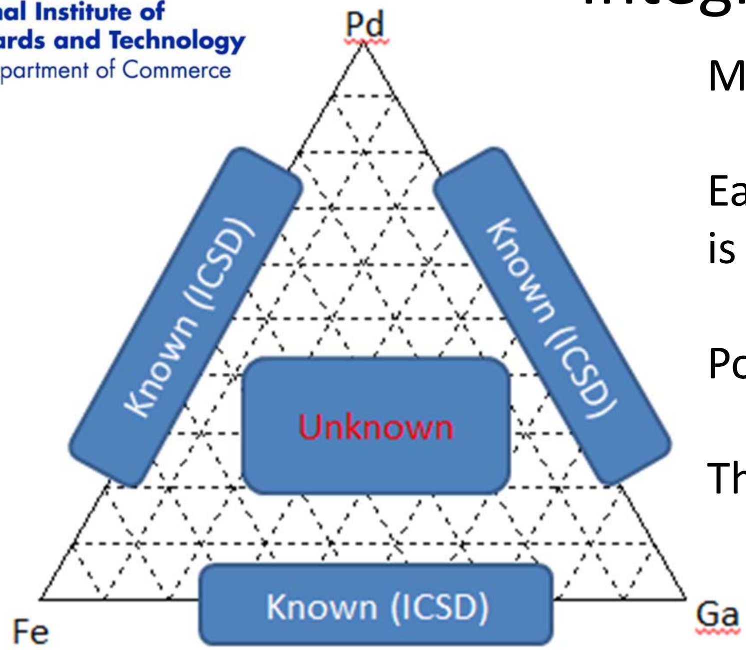
Working toward a structural phase diagram using NMF





Long *et al.*,
Rev. Sci. Instrum. **80**,
103902 (2009)

Integrating ICSD with combi XRD data



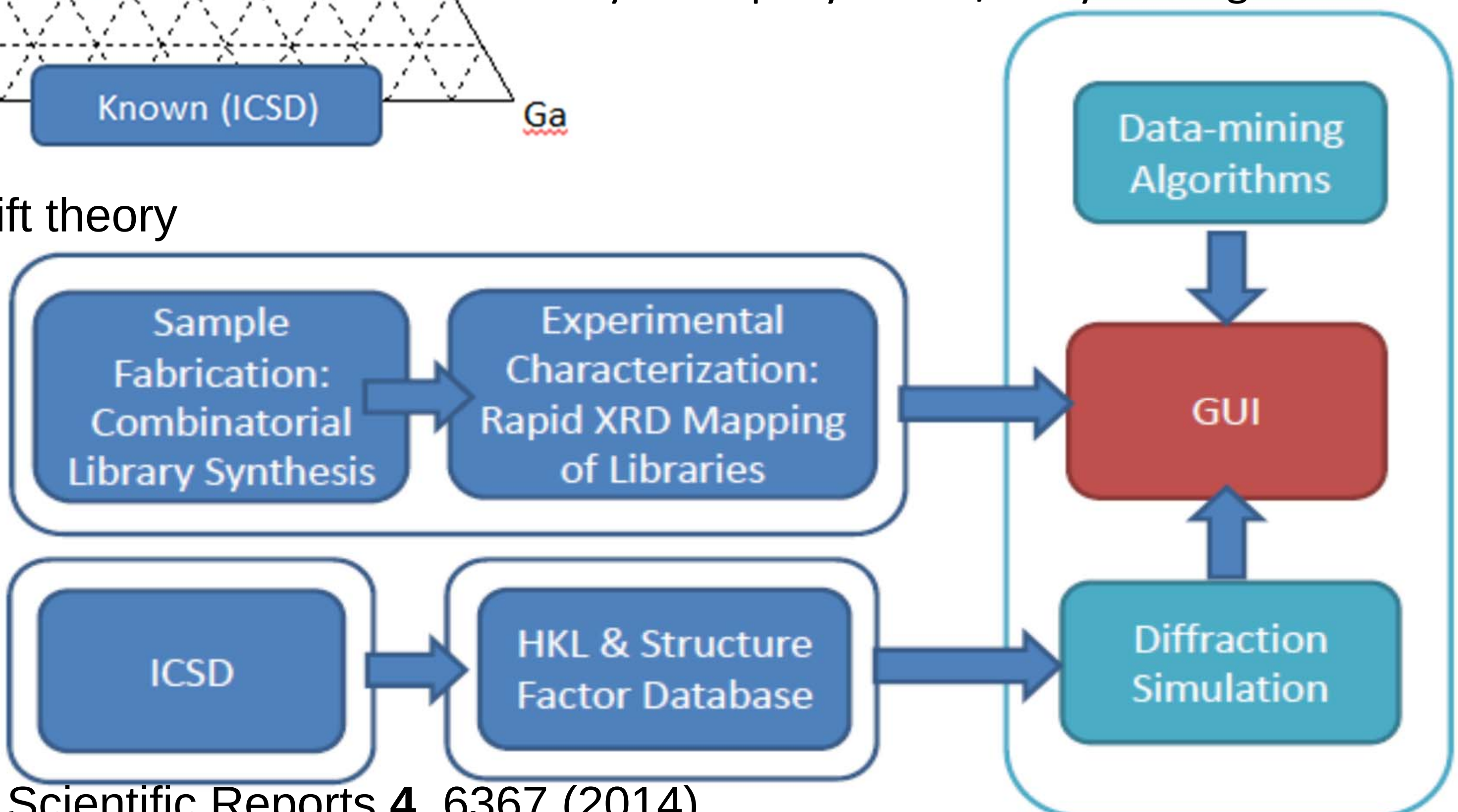
Most ternary phase diagrams are not known

Each point on the ternary phase diagram
is one X-ray spectrum (expt)

Points on binary lines are simulated spectra from ICSD

They are rapidly mined/analyzed together

mean shift theory



Scientific Reports **4**, 6367 (2014)

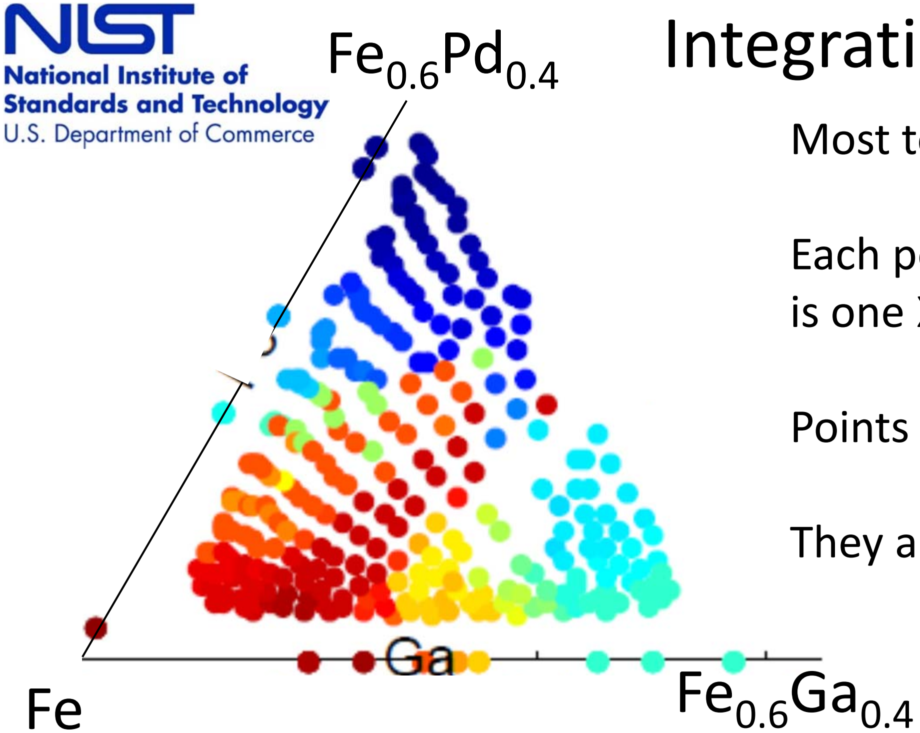
Integrating ICSD with combi XRD data

Most ternary phase diagrams are not known

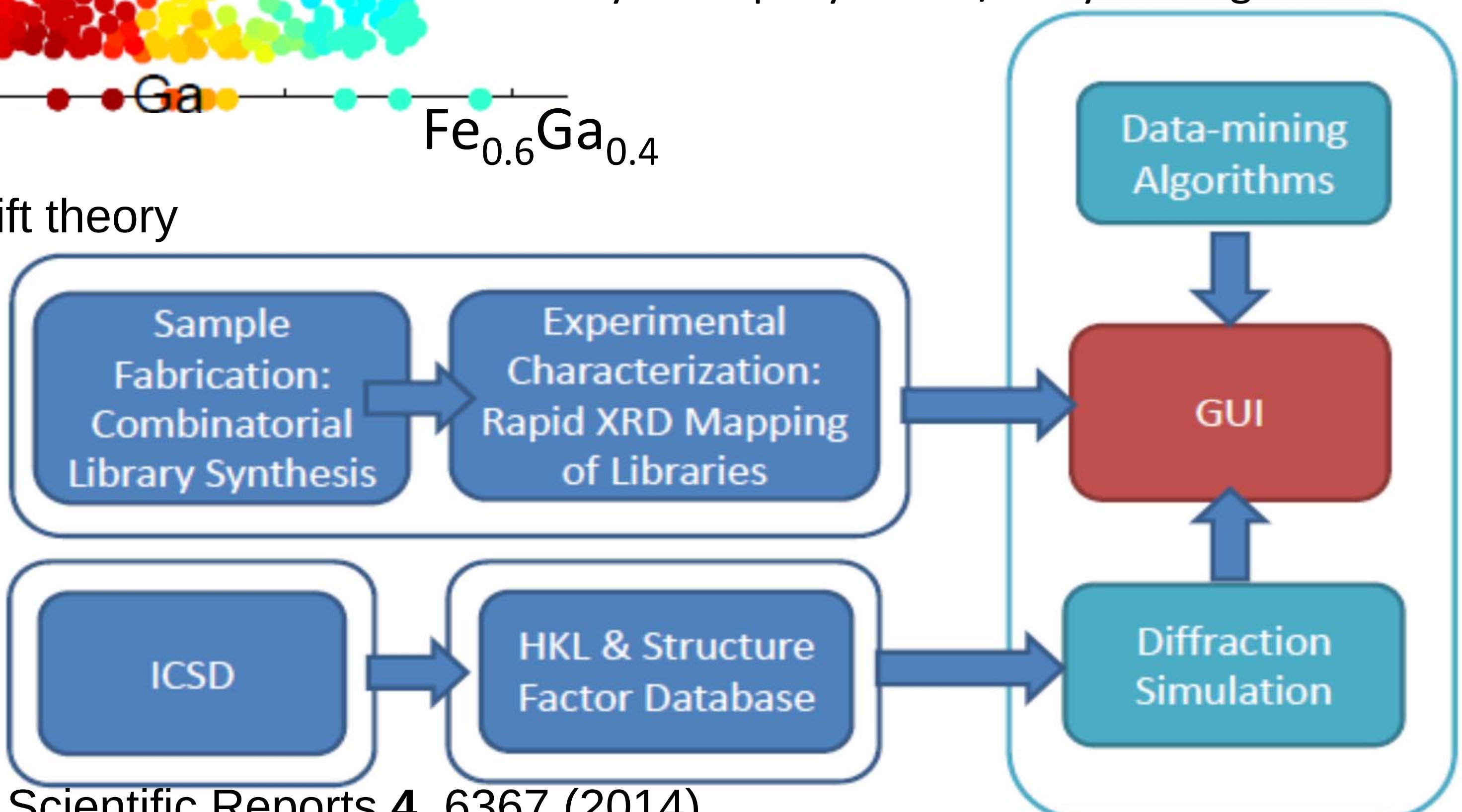
Each point on the ternary phase diagram is one X-ray spectrum (expt)

Points on binary lines are simulated spectra from ICSD

They are rapidly mined/analyzed together



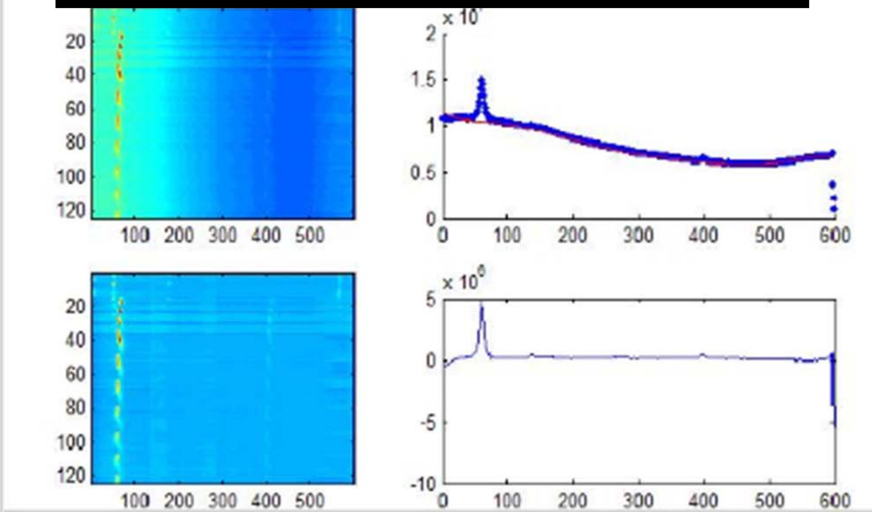
mean shift theory



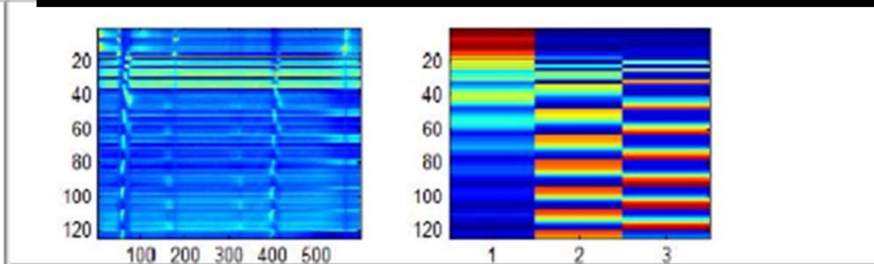
Scientific Reports **4**, 6367 (2014)

Real time analysis of combinatorial library data (synchrotron XRD); Integration with ICSD

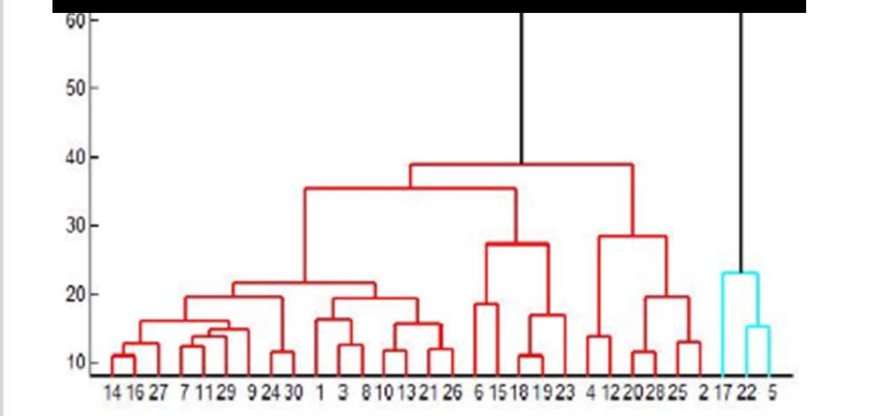
Diffraction data (integrated Pilatus images) plotted for different compositions as they are taken

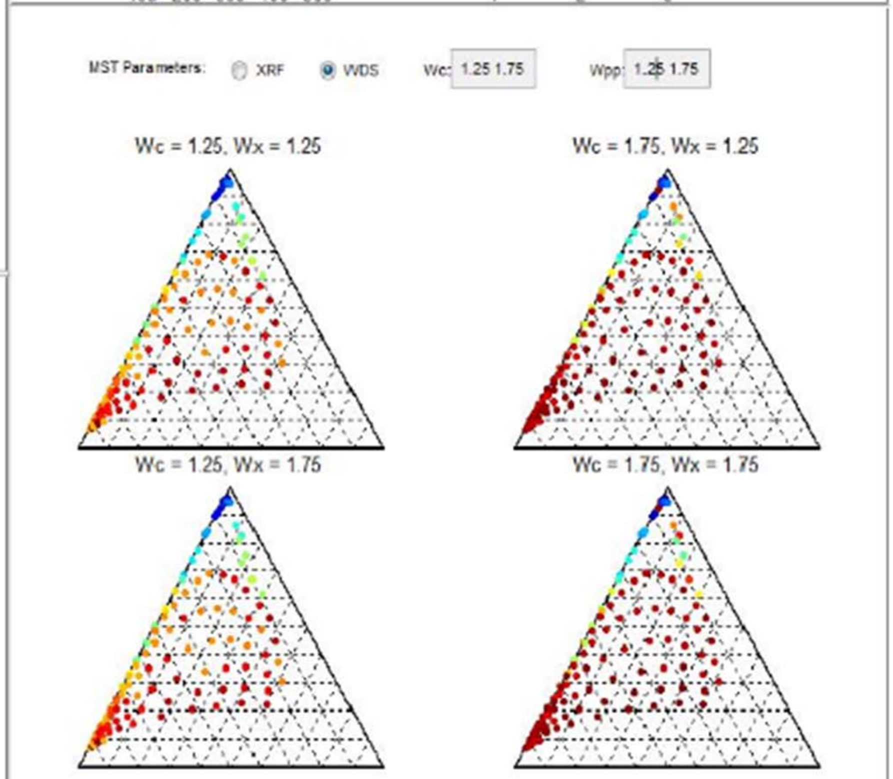


X-ray fluorescence data are also processed real time



Hierarchical clustering is used to group similar spectra together





Spectra are mapped on ternary phase diagram real time

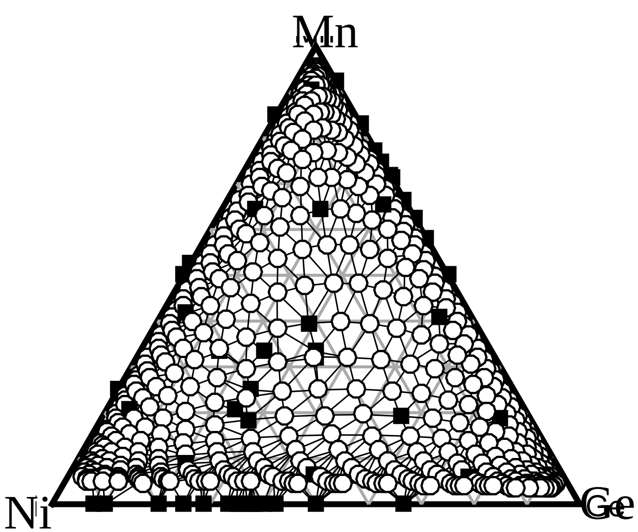
Scientific Reports **4**, 6367 (2014)

Phase mapping based on <u>active learning</u>

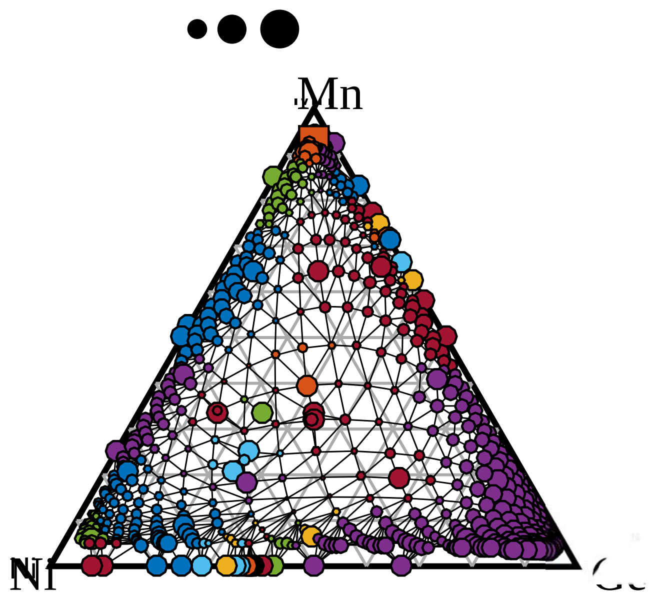
- Semi-Supervised Learning: Propagate knowledge

from measured/known samples to unmeasured samples

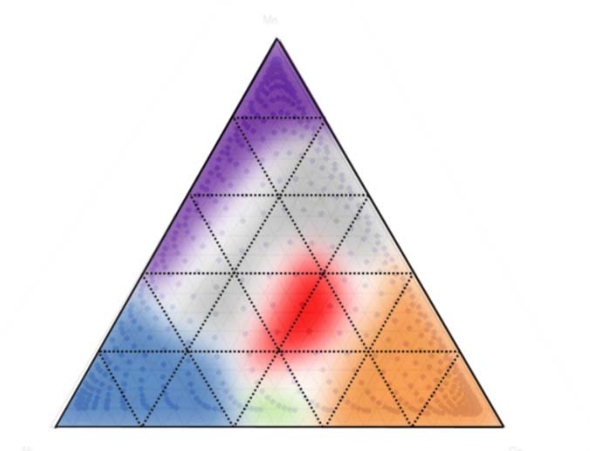
- Quantify Certainty of Prediction
- Graph-based phase separation algorithm used



■ Coarse Measurement
● Measured (Active)
○ Not Measured

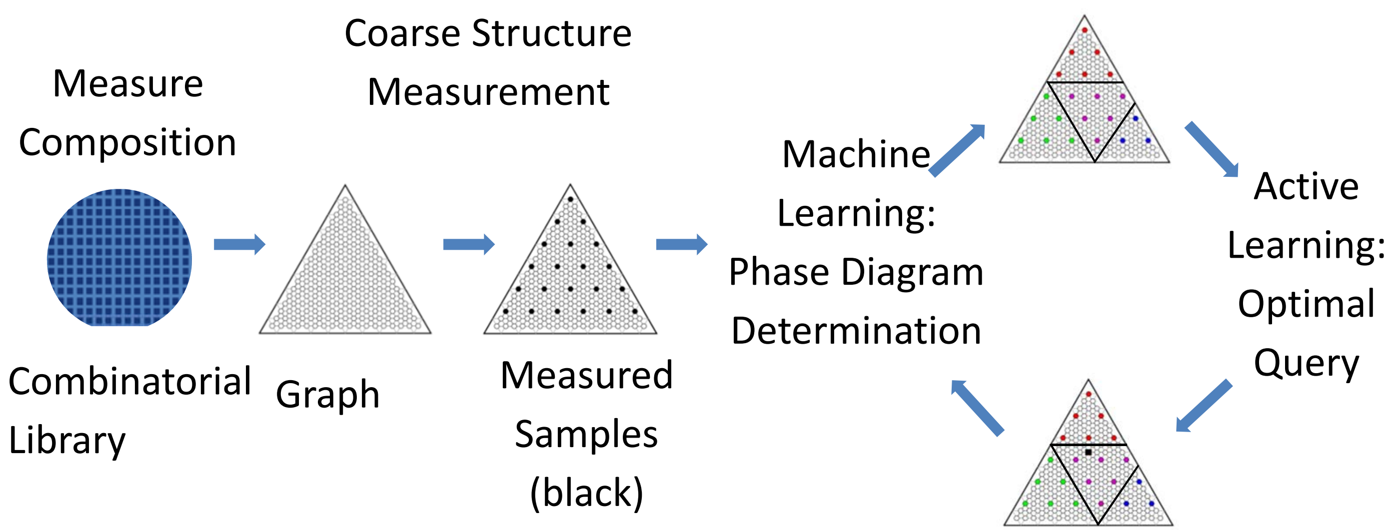


■ Sample to Query
●●● Certainty of Cluster Label (0-1)



Runs Live at SLAC X-Ray Beamline

Active Learning: Flow Chart

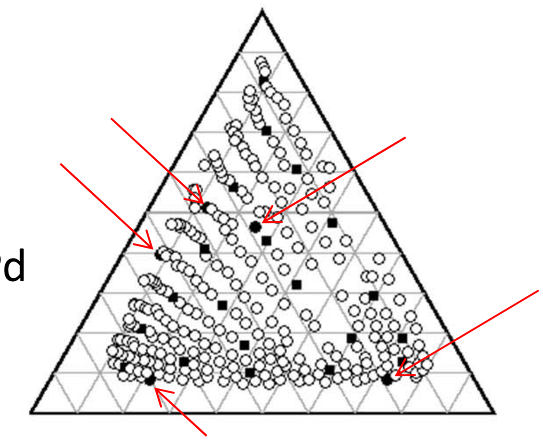


Active Learning Implementation:

- Select query that minimizes ‘risk’
- Risk : estimated expected classification error

Active Learning: Example

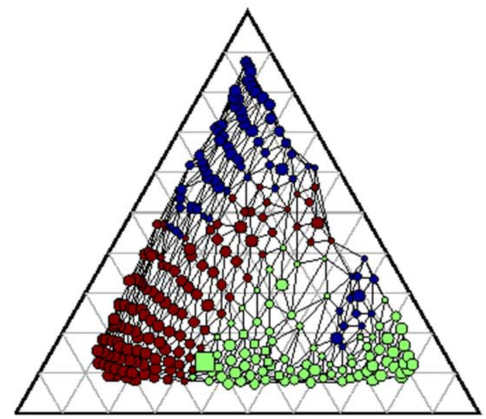
Ternary Spread Demonstration : Fe-Ga-Pd



Coarse Measurement

Measured (Active)

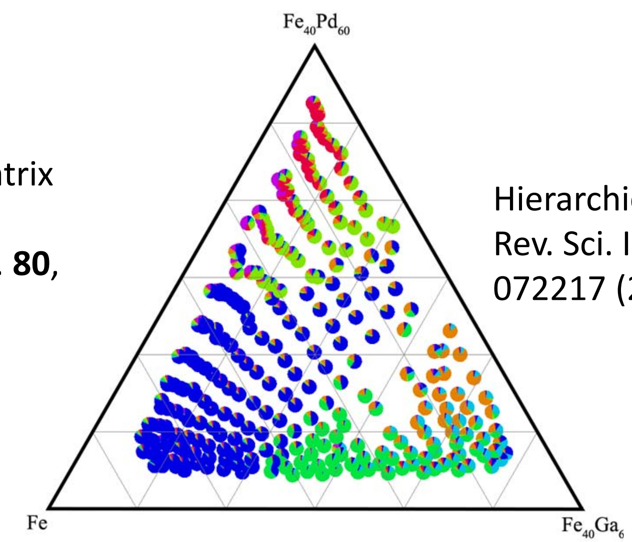
Not Measured



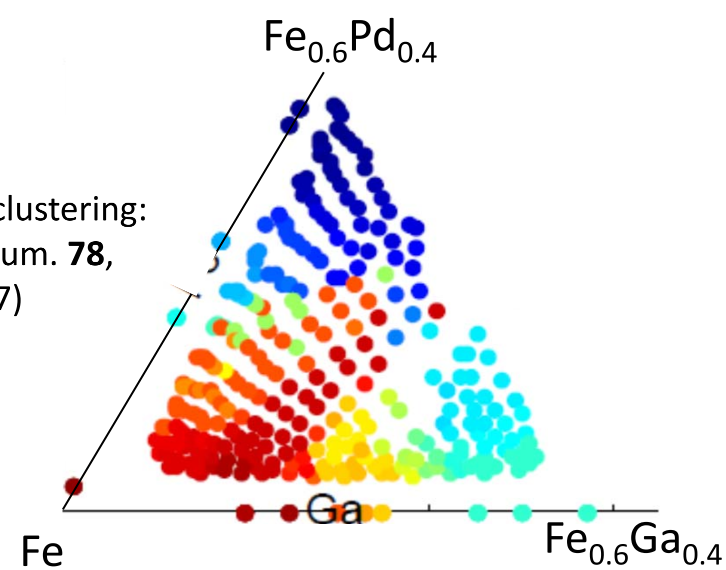
Sample to Query

Certainty of Cluster Label (0-1)

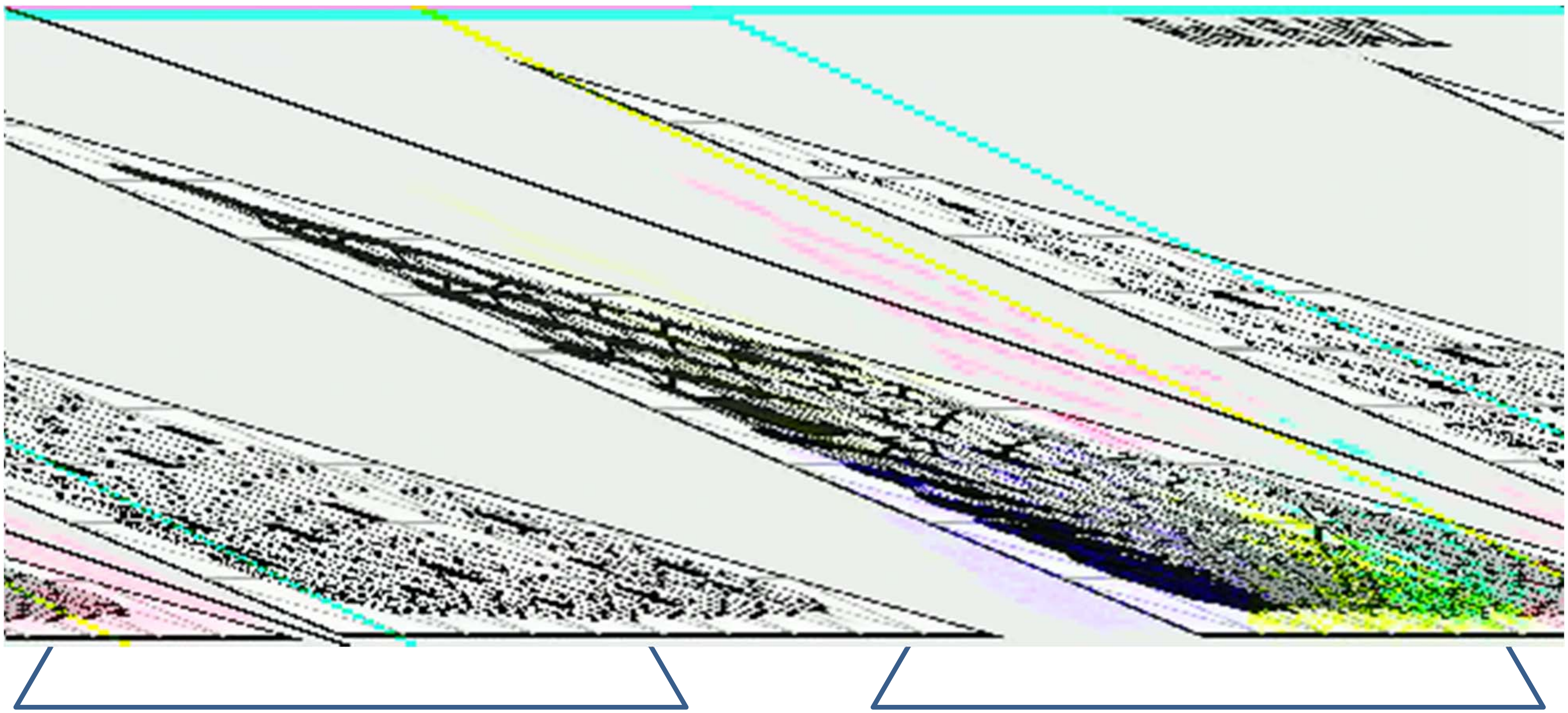
Previous results: non-negative matrix factorization: Rev. Sci. Instrum. **80**, 103902 (2009)



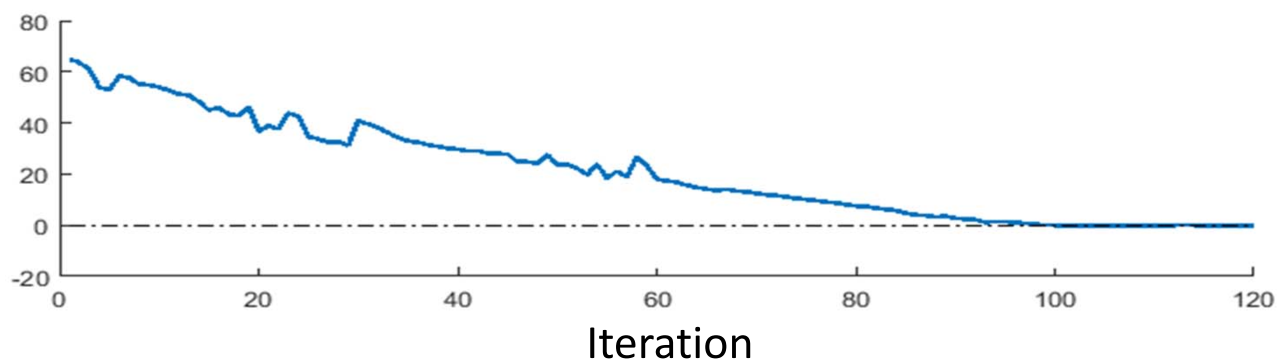
Hierarchical clustering: Rev. Sci. Instrum. **78**, 072217 (2007)



Active learning: algorithm finds the boundaries for you

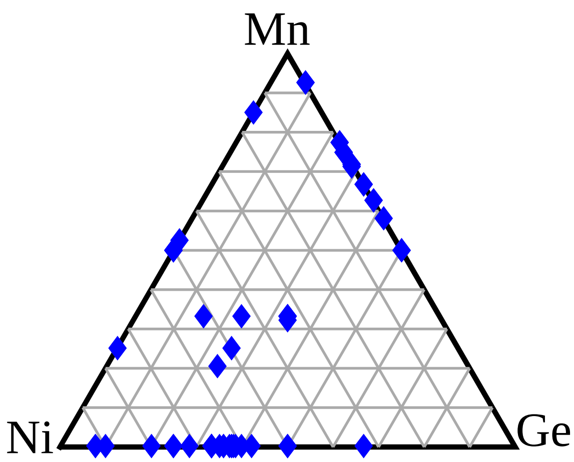


Risk: Estimated expected classification error

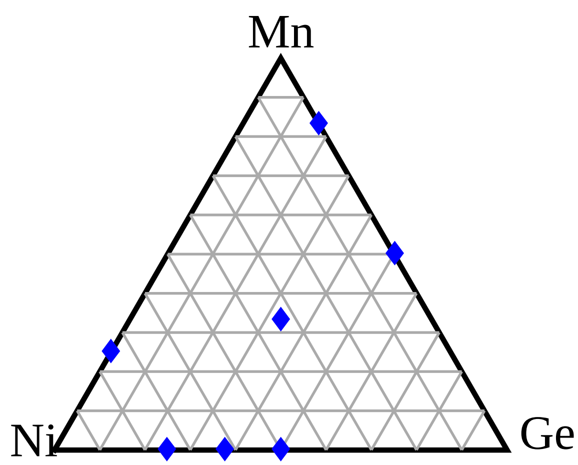


Active Learning Phase Mapping

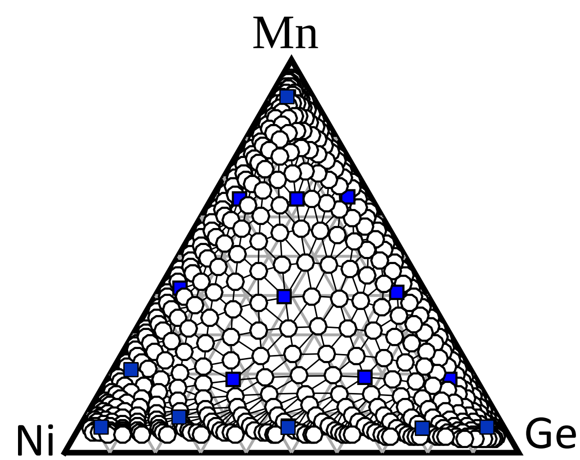
Ni-Mn-Ge



Data from ICSD

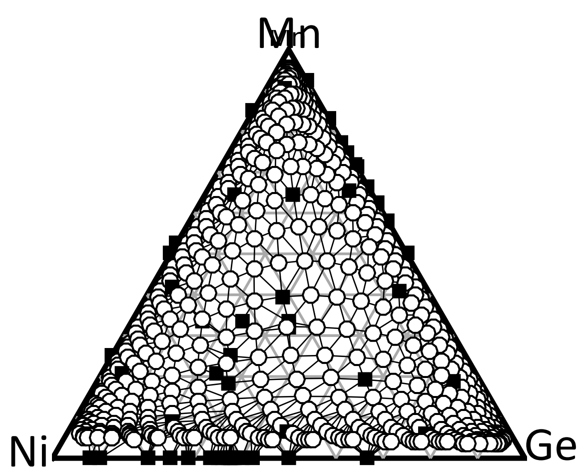


Data from AFLOW



Initial measurements (16)

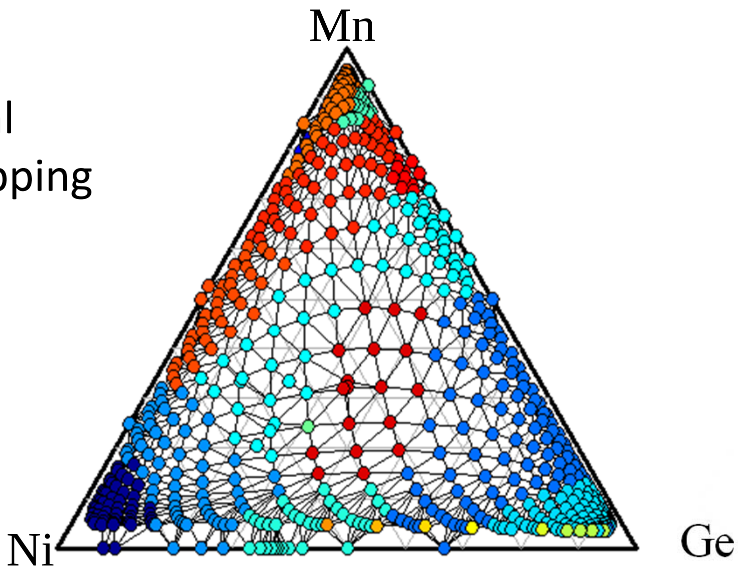
Combined initial information



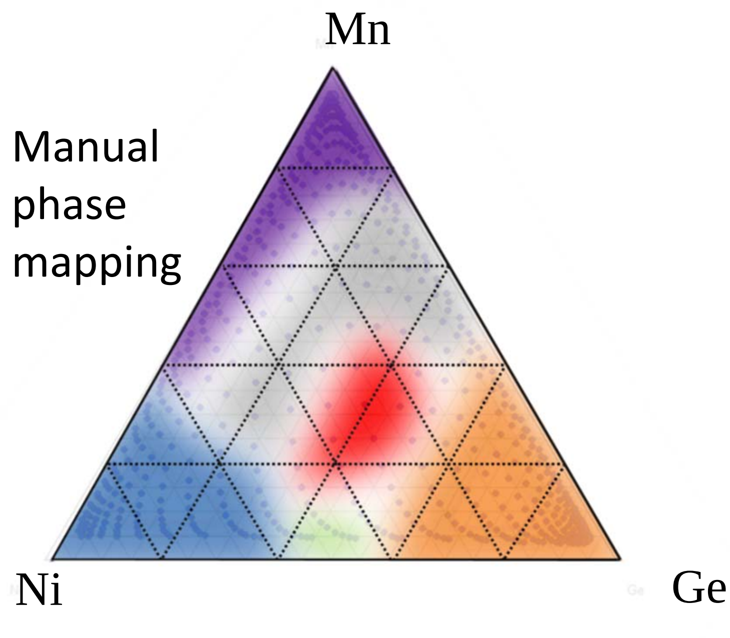
Active Learning Phase Mapping

Ni-Mn-Ge

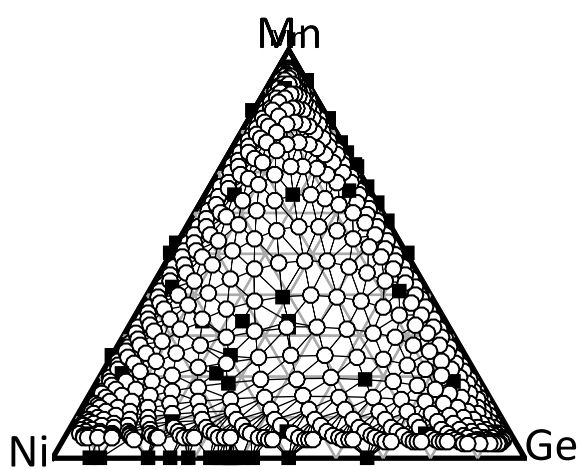
Final mapping



Manual phase mapping



Combined initial information



- F-measure used for accuracy

- 10% of samples required for measurements to get 80% accuracy for the whole

Evolution of the combinatorial strategy (and a future forecast)

Circa 1990 | **2000** | **2010** | **2017 and beyond**

Challenges:

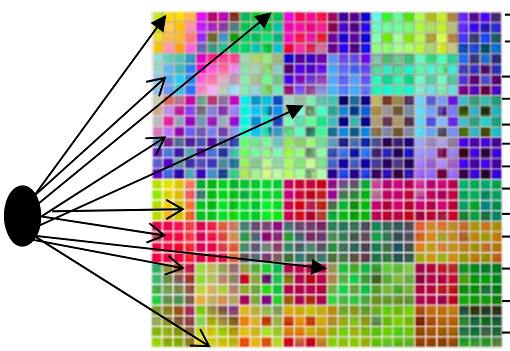
How to make hundreds of samples fast

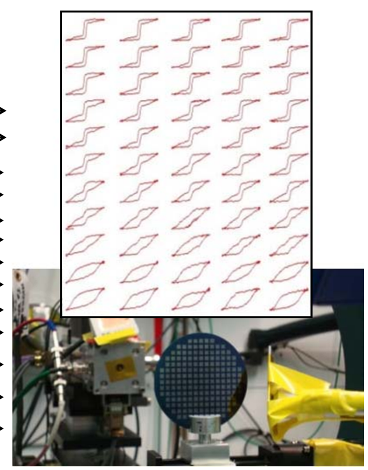
How to measure large number of samples: speed, number and quantitative accuracy

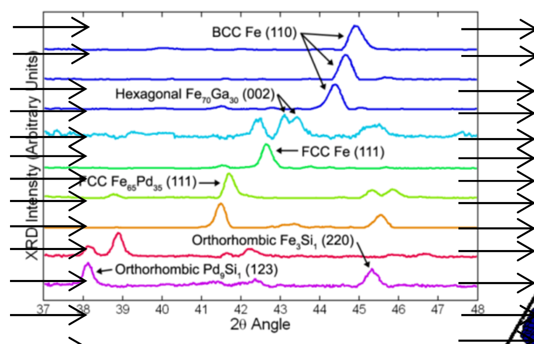
How to quickly analyze large amount of data

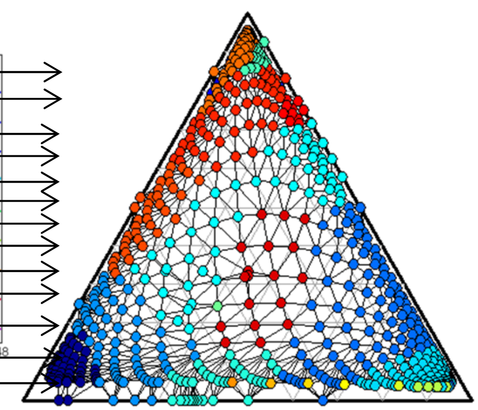
Machine learning to control and reduce number of samples

# of samples: x 100-1000









# of samples: x 0.1 – 0.2

# of samples: x 0.1 – 0.2 Reducing the number of data points

Summary

Combinatorial (high-throughput) experimentation + theory can speed up discovery of materials

Machine learning can greatly enhance combinatorial experimentation

Annual Machine Learning for Materials Research Workshop and Summer Camp, Univ. of Maryland Sponsored by UMD and NIST

https://www.nanocenter.umd.edu/events/mlmr/

June, 2017