TMS

PROMOTING THE GLOBAL SCIENCE AND ENGINEERING PROFESSIONS CONCERNED WITH MINERALS, METALS, AND MATERIALS

NSF-IMI COSMIC International Materials Institute

Data Mining and Materials Informatics: a primer

Krishna Rajan

Department of Materials Science and Engineering
NSF Intl. Materials Institute Combinatorial Sciences & Materials Informatics Collaboratory
Iowa State University

What is data?

primary, secondary , derivative

sources of data

What do we mean by mining data?

data correlations – dimensional analysis approach

data correlations – data mining

What can we learn from data mining?

Classification

data mining databases

hierarchy of data

trends in data

Prediction

predicting structure-property relationships

computational informatics vs. computational materials science

data mining for building databases

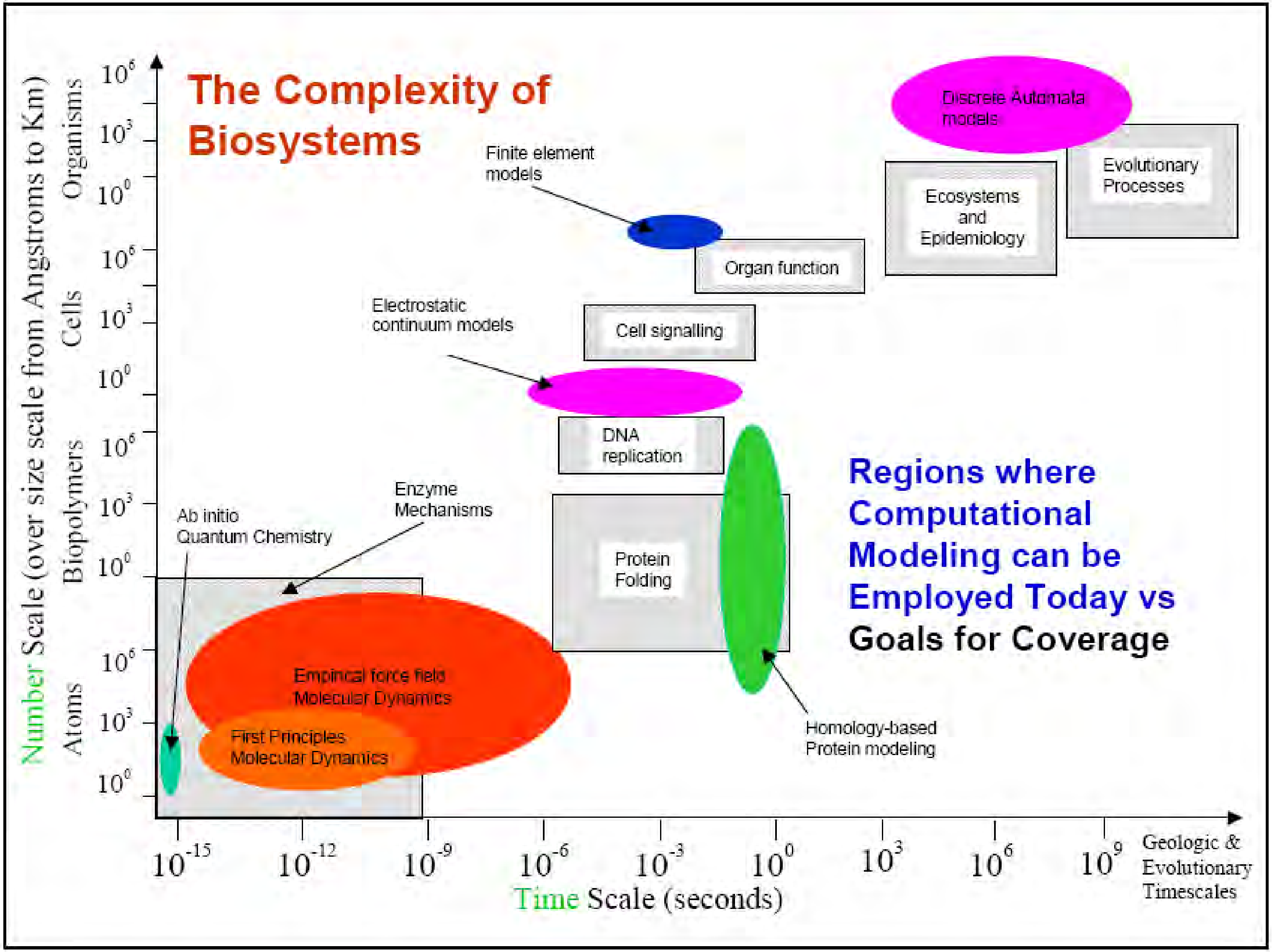
What is it ?

- Searching for patterns of behavior among multivariate data sets
- Can pattern recognition lead to predictions?

Why ?

- Establish new correlations
- Identify outliers
- Enlarge database / virtual libraries
- Evaluate databases
- Establish predictions

COMPLEXITY OF BIOSYSTEMS



2005 and Beyond: A Roadmap for Consolidation and Exponentiation

A Workshop Report July 14-15, 2003

NSF Directorate for Biological Sciences Advisory Committee (BIOAC)

INFORMATICS STRATEGY

Data + Correlations + Theory = Knowledge-base

- Atomistic based calculations
- Continuum based theories

Informatics-Based Design

- Machine learning
- Data compression
- Pattern recognition

Machine learning toolkit:

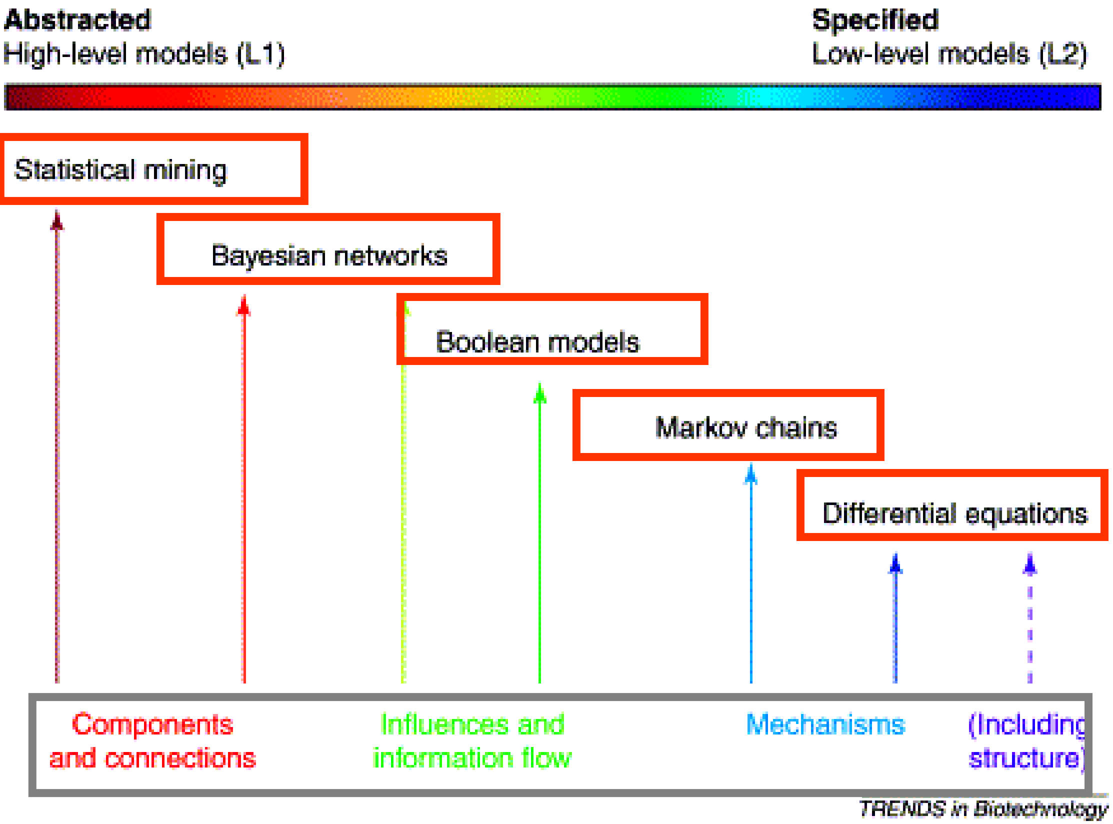
- Classification
- Prediction
- Hybrid tools

- Simulations
- Combinatorially derived datasets
- Digital libraries
- Spectral and imaging libraries

ULTRA LARGE SCALE INFORMATION SPACE

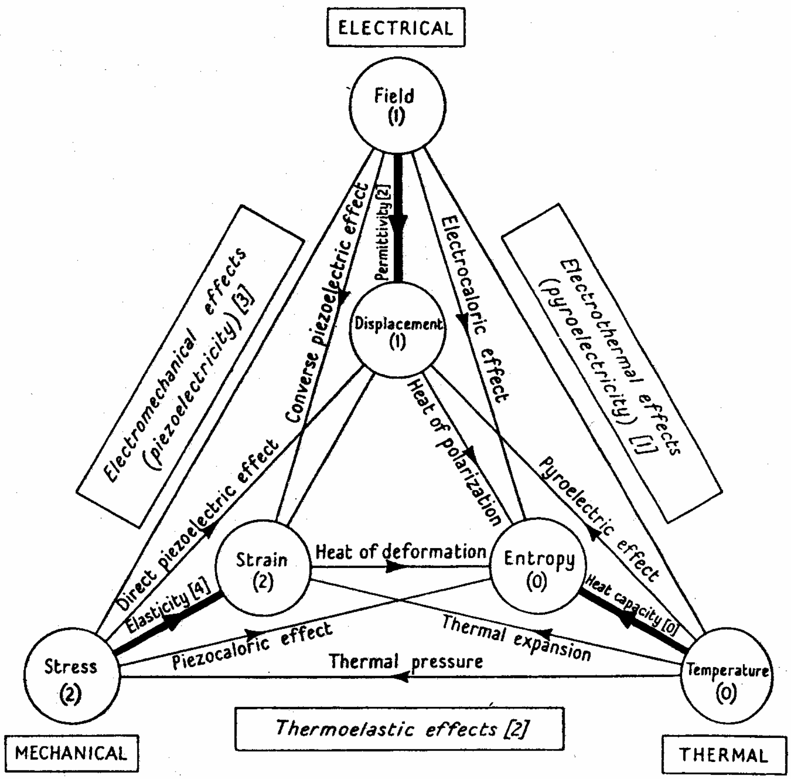
INFORMATICS-BASED DESIGN STRATEGIES

BIOLOGICAL ANALOGUE

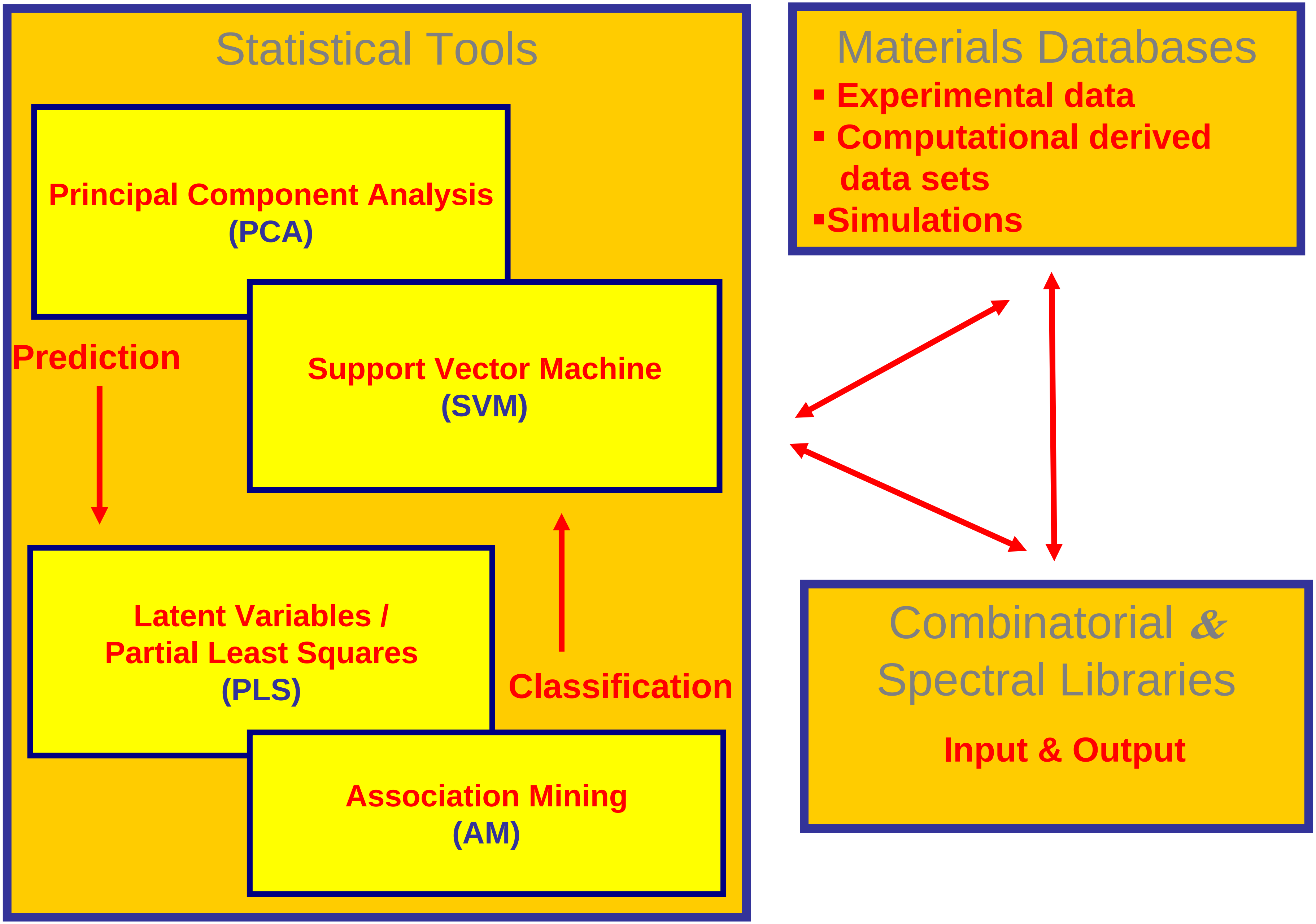


Ideker and Lauffenburger:(2003)

CRYSTAL STRUCTURE ANALOGUE



Nye (1950)


Statistical Tools
Principal Component Analysis
(PCA)
Support Vector Machine
(SVM)
Prediction
Latent Variables /
Partial Least Squares
(PLS)
Classification
Association Mining
(AM)
Materials Databases
▪ Experimental data
▪ Computational derived data sets
▪Simulations
Combinatorial & Spectral Libraries
Input & Output

Requirements

- Global minima
- Categorical data
- Missing / variable data
- Skewed distributions
- Large data sets
- Scalable

Pitfalls

- Local minima
- Categorical data difficult
- Convergence difficult for large data sets
- Few outliers can lead to poor performance

DATA STORAGE

DATA CURATION

DATA REPRESENTATION

KNOWLEDGE DISCOVERY

Database administration & management

- thermodynamic, crystallographic & property data bases
- combinatorial experimental data

Oracle, Unix, Supercomputing, SQL

- taxonomy and ontology of materials science data
- data sharing , networking / cyber infrastructure

JAVA, HTML, Python

- object oriented programming language
- visualization of high dimensional data

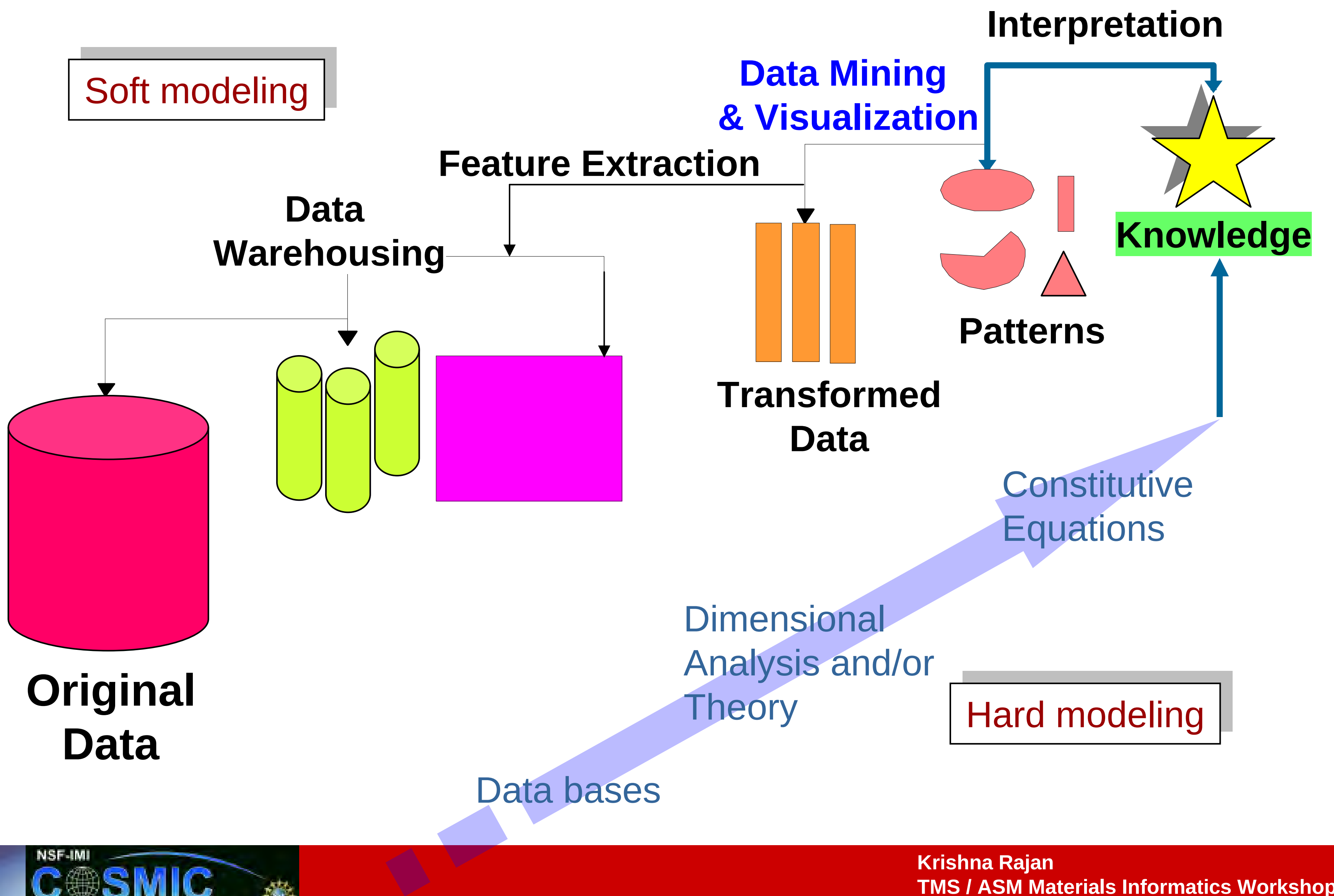
Data mining algorithms

- Clustering analysis
- Quantitative Structure-Activity Relationships (QSAR) for materials design

WHY COUPLE COMPUTATIONAL MATERIALS SCIENCE *AND* INFORMATICS ?

- Accelerated insertion of materials into engineering systems
- Rapid multiscale design and optimization of materials properties
- Establishment of new structure –property correlations among large, heterogeneous and distributed data sets
- Discovery of new chemistries and compounds
- Formulation and / or refinement of new theories for materials behavior
- Rapid identification of critical data and theoretical needs for future problems

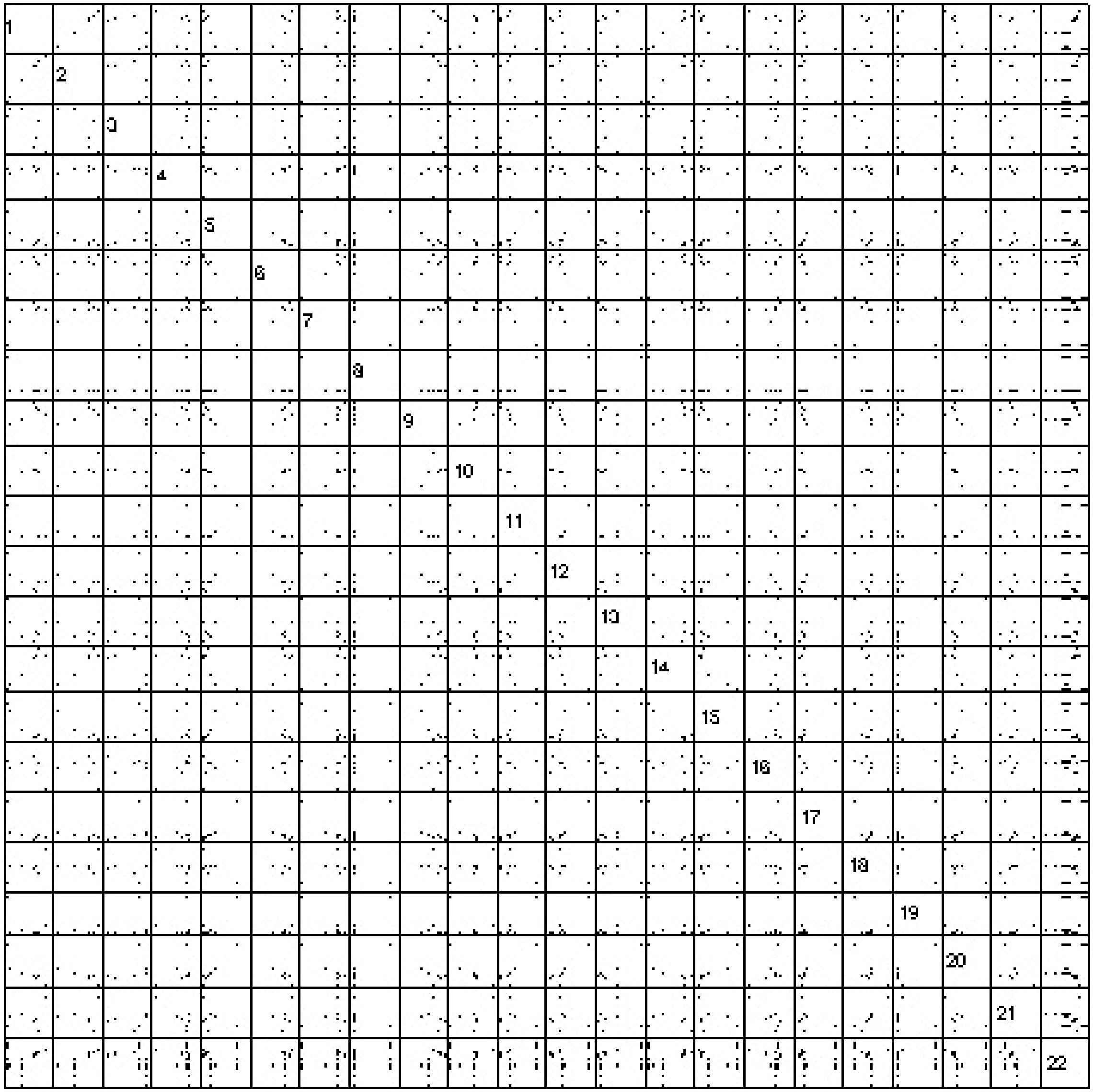
SOFT MODELING vs. HARD MODELING



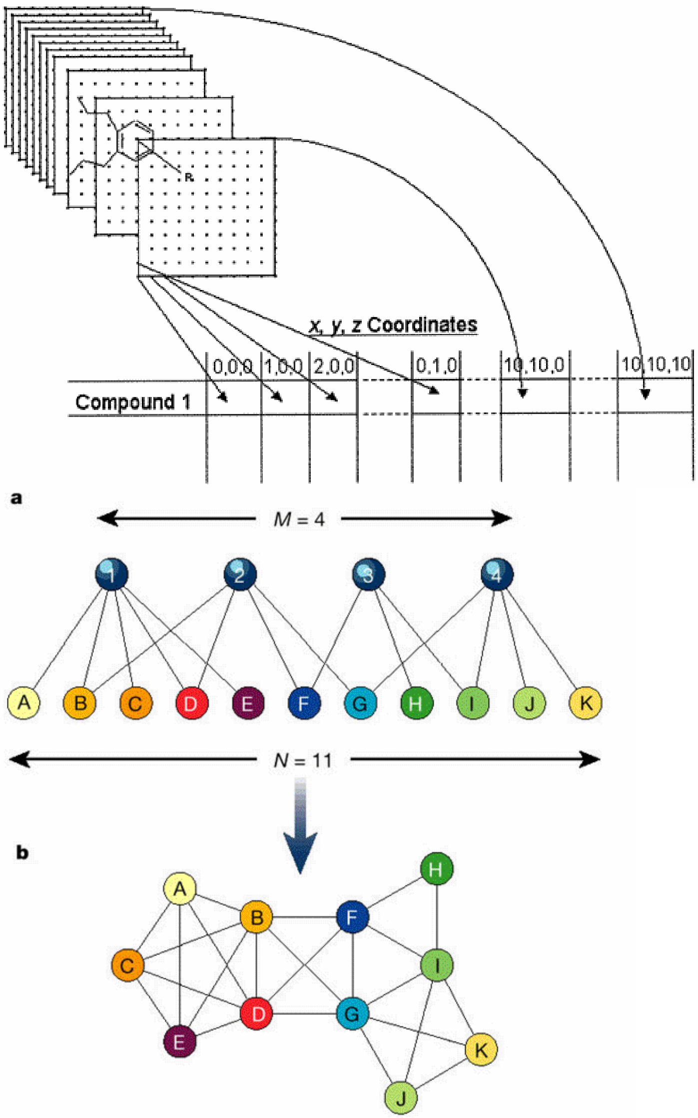
DATA MAP

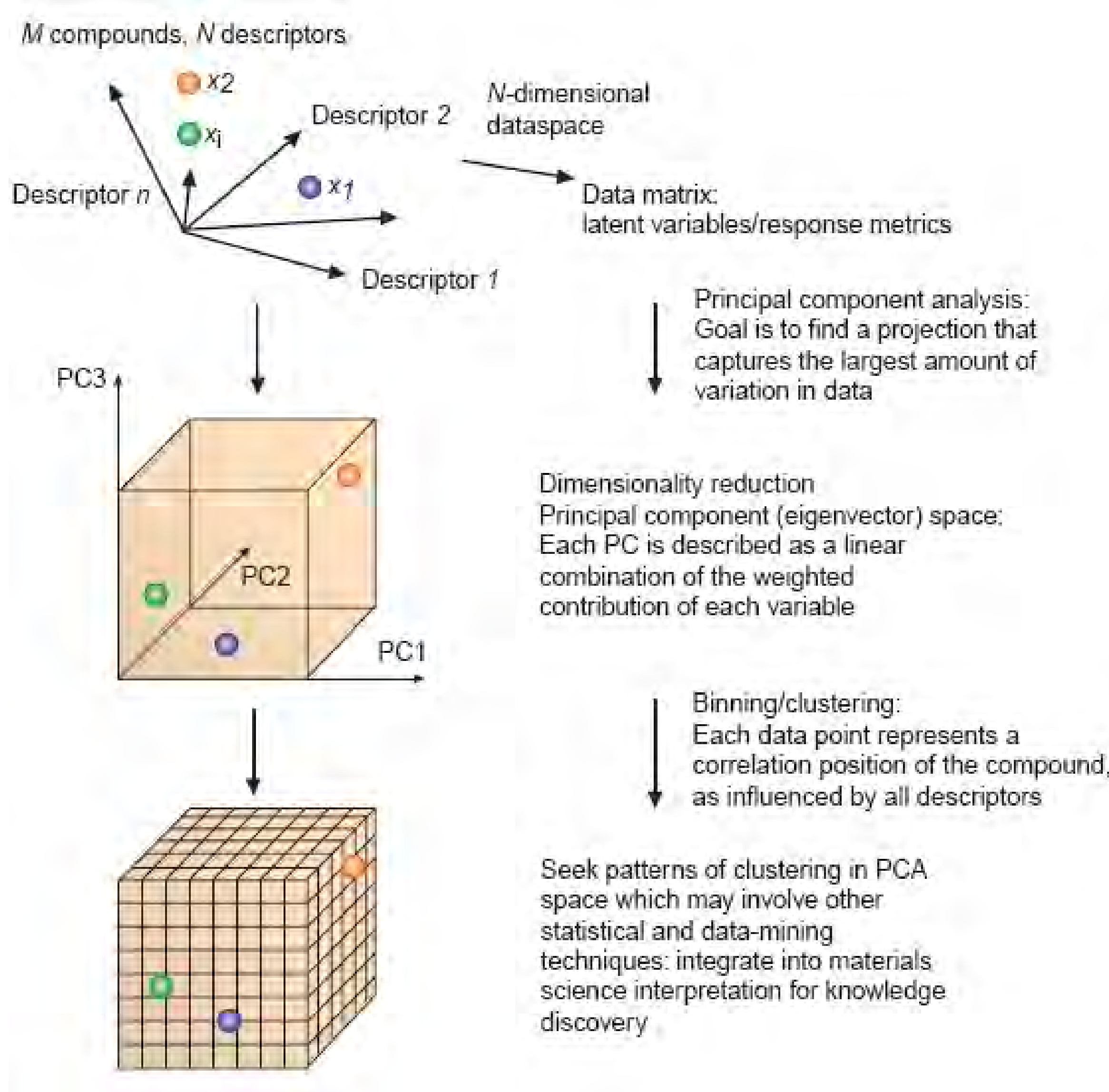
22x22 (484 cells – 2500 data points):

Silicon nitride descriptors



DIMENSIONALITY REDUCTION





$$Sp = \lambda p$$

eigenvector

eigenvalue

$$SP = P\Lambda$$

Eigenvalues on the diagonal of this diagonal matrix

Eigenvectors forming the columns of this matrix

If V is nonsingular, this become the eigenvalue decomposition

Eigenvalue Matrix

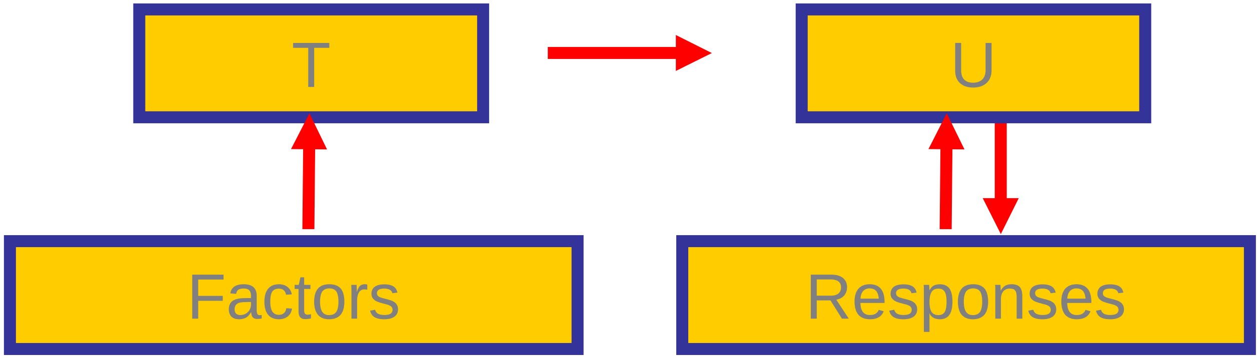
$$S = P\Lambda P^{-1}$$

Loading Matrix=Eigenvector Matrix

PREDICTIONS

PLS

- 'Upgradeable' linear predictive tool
- Quadratic and interaction terms :flexibility



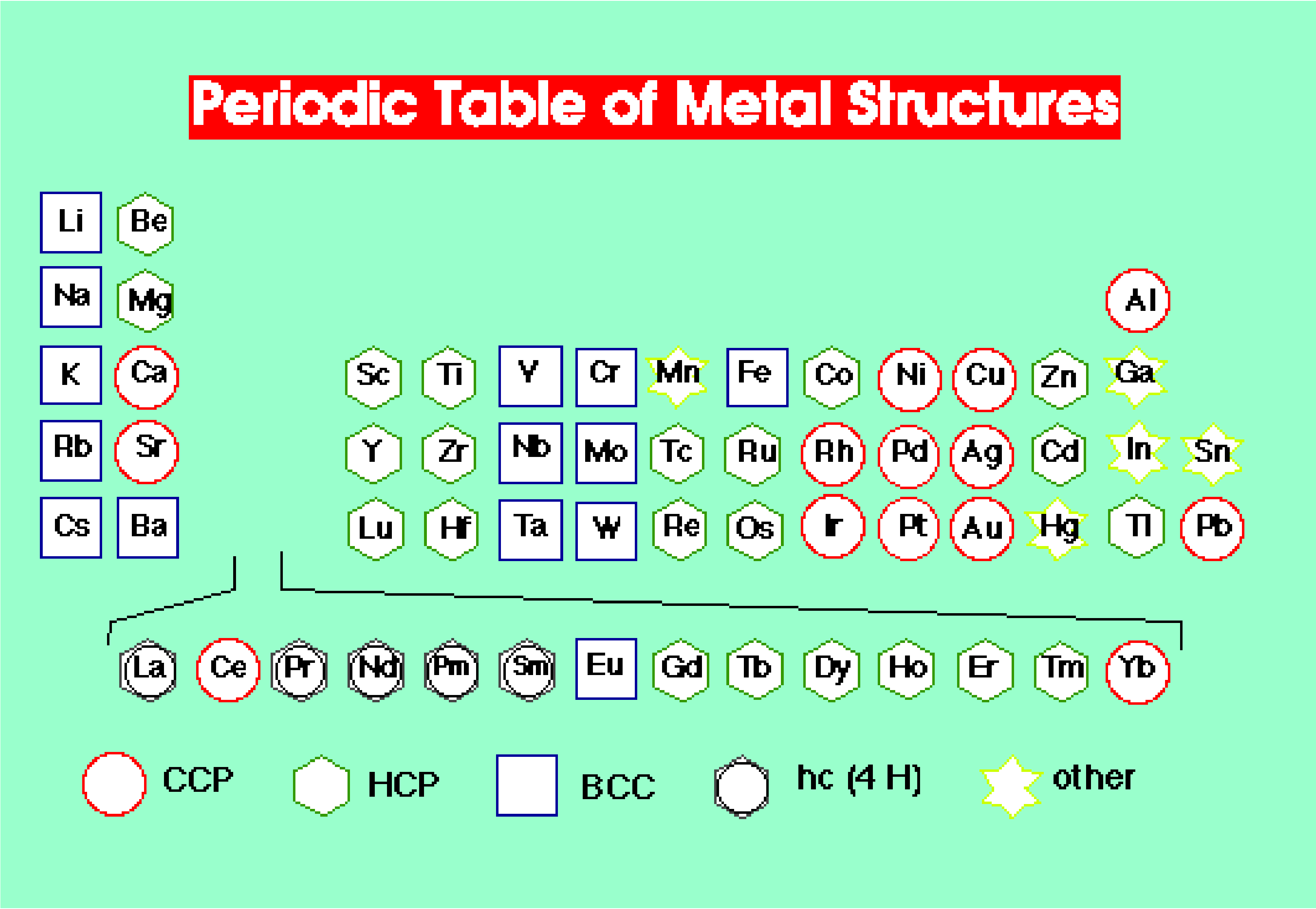
Calculate latent factors from original predictor variables

T=XW where ***T***- latent factors, ***X*** – predictors, ***W*** - weights

Use latent factors to predict response

Y=TQ+E where Y - responses, Q – loadings, E- noise

- ***W*** maximizes covariance between ***Y*** and ***T***


Periodic Table of Metal Structures
Li Be
Na Mg
Al
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn
Cs Ba Lu Hf Ta W Re Os Ir Pt Au Hg Tl Pb
La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb
CCP
HCP
BCC
hc (4 H)
other

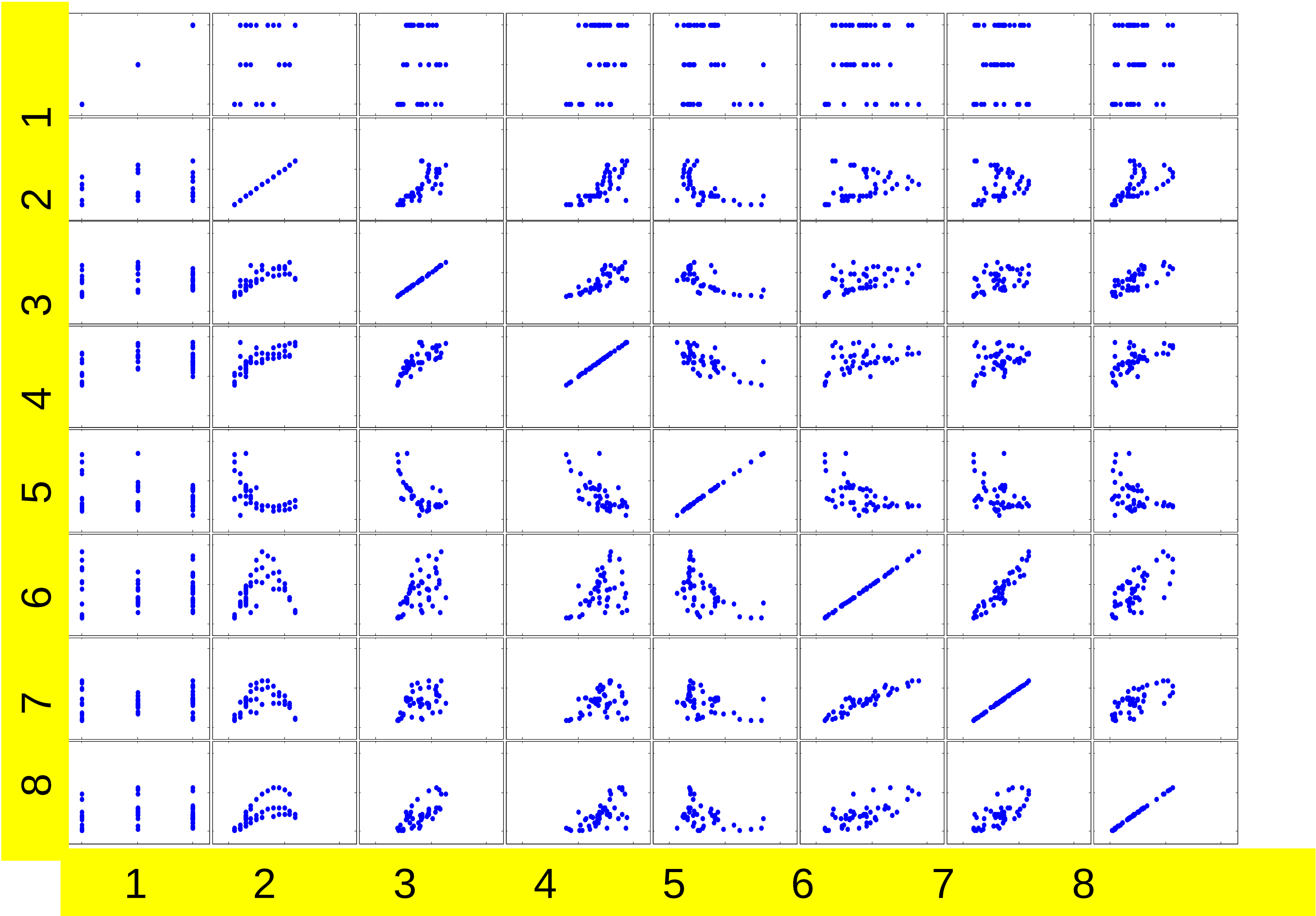
DATA WAREHOUSING

Symbol	Atomic Number	Structure	Number of atoms in a unit cell	Valence electron number, Nv	Pauling electronegativity	First Ionization Potential (eV)	Atomic Radius (A)	Melting Point (K)	Boiling Point (K)	Density @ 293 K (g/cm^3)	Atomic Weight
Li	3	bcc	2	1	0.98	5.39	1.52	453.69	1620	0.534	6.941
β-Be	4	bcc	6	2	1.57	9.32	1.13	1551	3243	1.8477	9.012182
α-Be	4	hcp	6	2	1.57	9.32	1.13	1551	3243	1.8477	9.012182
Na	11	bcc	2	1	0.93	5.14	1.54	370.96	1156.1	0.971	22.989768
Mg	12	hcp	6	2	1.31	7.64	1.6	922	1363	1.738	24.305
Al	13	fcc	4	3	1.61	5.98	1.43	933.52	2740	2.698	26.981539
K	19	bcc	2	1	0.82	4.34	2.27	336.8	1047	0.862	39.0983
α-Ca	20	fcc	4	2	1	6.11	1.97	1112	1757	1.55	40.078
Sc	21	hcp	6	3	1.36	6.56	1.61	1814	3104	2.989	44.95591
Ti	22	hcp	6	4	1.54	6.83	1.45	1933	3560	4.54	47.88
V	23	bcc	2	5	1.63	6.74	1.32	2160	3650	6.11	50.9415
Cr	24	bcc	2	6	1.66	6.76	1.25	2130	2945	7.19	51.9961
Fe	26	bcc	2	8	1.83	7.9	1.24	1808	3023	7.874	55.847
Co	27	hcp	6	9	1.88	7.86	1.25	1768	3143	8.9	58.9332
Ni	28	fcc	4	10	1.91	7.63	1.25	1726	3005	8.902	58.6934
Cu	29	fcc	4	11	1.9	7.72	1.28	1356.6	2840	8.96	63.546
Zn	30	hcp	6	12	1.65	9.39	1.33	692.73	1180	7.133	65.39
Rb	37	bcc	2	1	0.82	4.18	2.475	312.2	961	1.532	85.4678
Y	39	hcp	6	3	1.22	6.5	1.81	1795	3611	4.469	88.90585
Zr	40	hcp	6	4	1.33	6.95	1.6	2125	4650	6.506	91.224
Nb	41	bcc	2	5	1.6	6.77	1.43	2741	5015	8.57	92.90638
Mo	42	bcc	2	6	2.16	7.18	1.36	2890	4885	10.22	95.94
Tc	43	hcp	6	7	1.9	7.28	1.36	2445	5150	11.5	-97.9072
Ru	44	hcp	6	8	2.2	7.36	1.34	2583	4173	12.37	101.07
Rh	45	fcc	4	9	2.28	7.46	1.34	2239	4000	12.41	102.9055
Pd	46	fcc	4	10	2.2	8.33	1.38	1825	3413	12.02	106.42
Ag	47	fcc	4	11	1.93	7.57	1.44	1235.08	2485	10.5	107.8682
Cd	48	hcp	6	12	1.69	8.99	1.49	594.1	1038	8.65	112.411
Sb	51	hcp	6	5	2.05	8.64	1.82	903.89	1908	6.691	121.757
Cs	55	bcc	2	1	0.79	3.89	2.654	301.55	951.6	1.873	132.90543

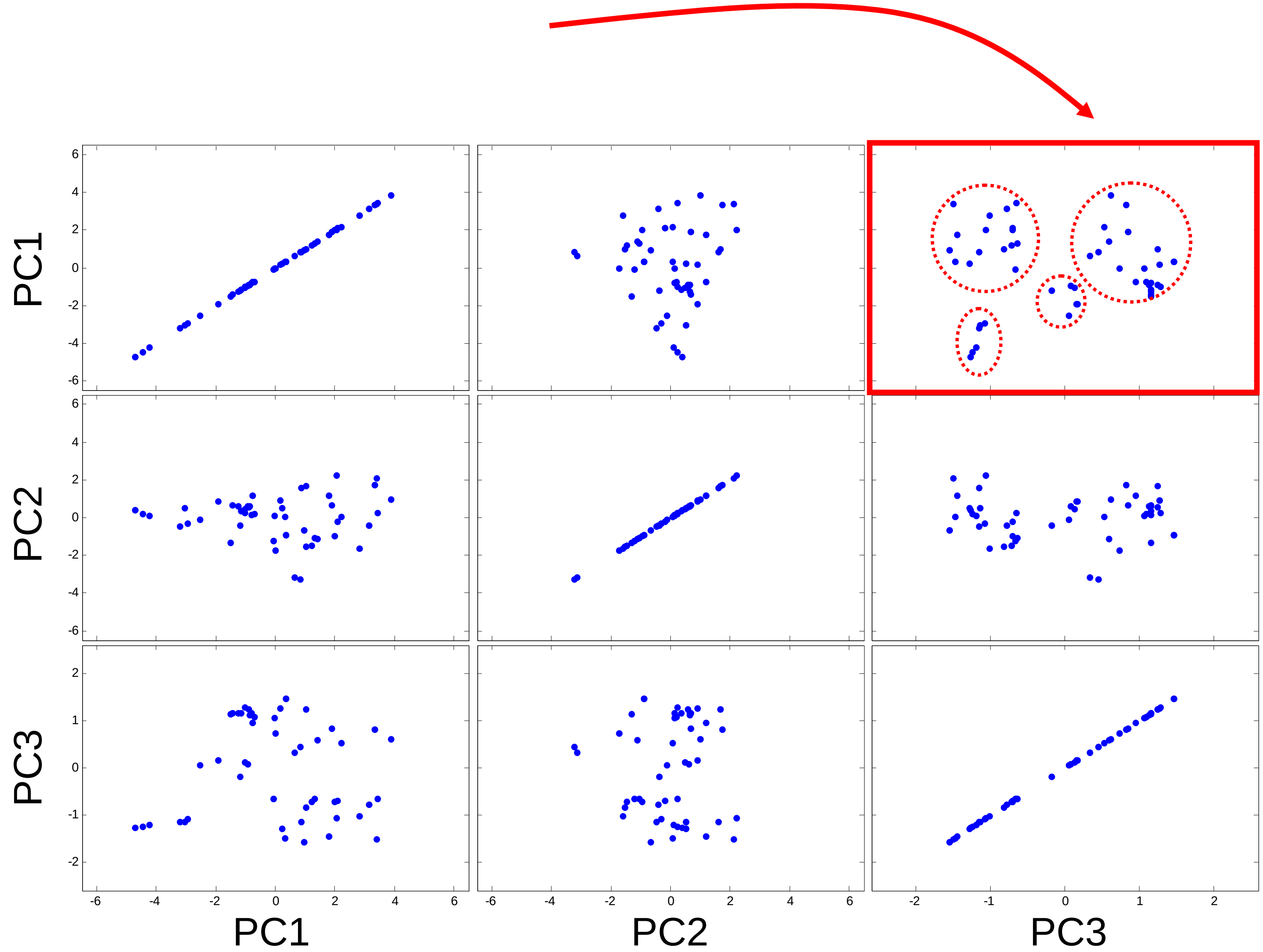
BIVARIATE MAPS

1. # of atoms/unit cell 2. Valence electron # 3. Electronegativity 4. 1st Ionization potential

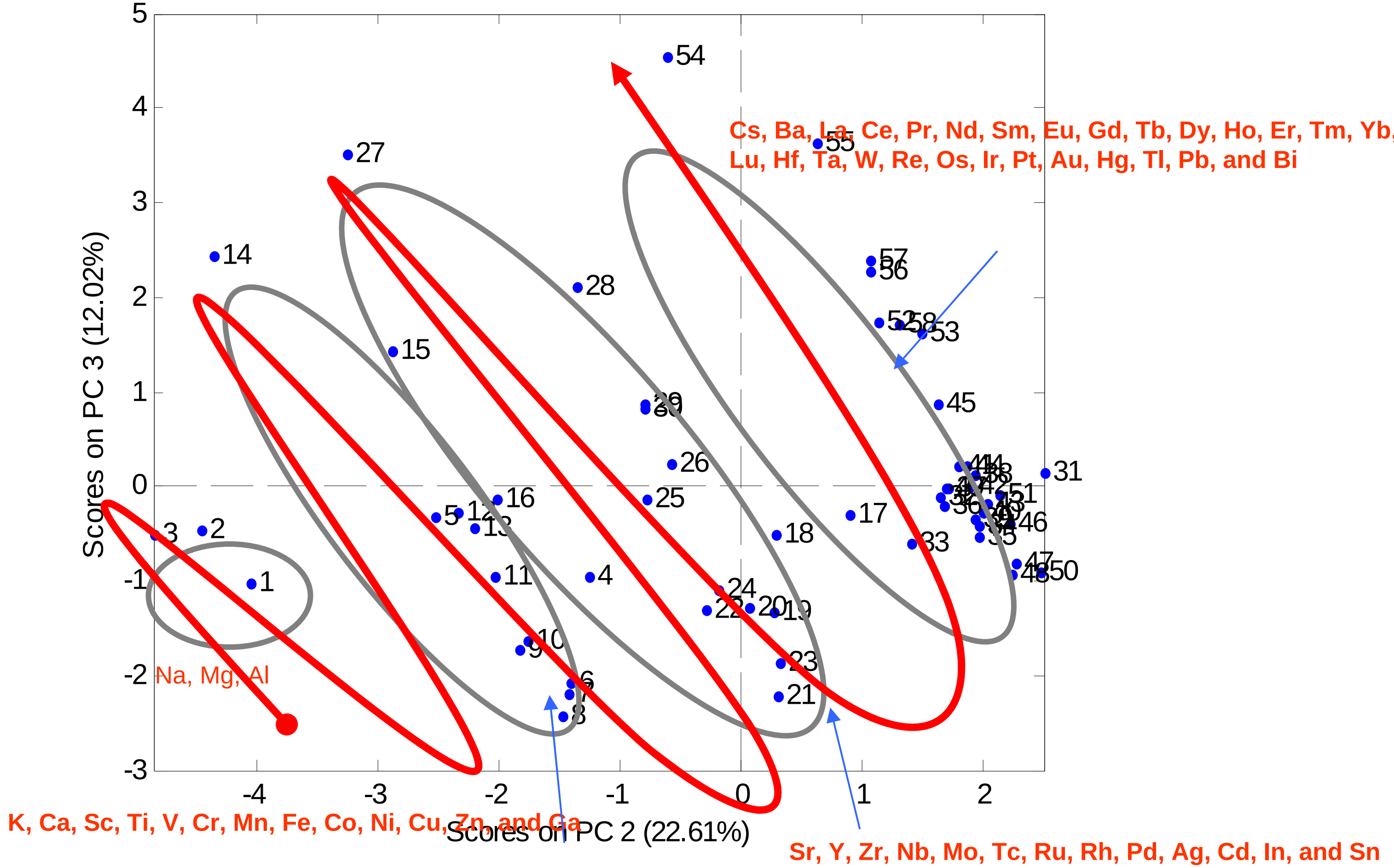
5. Atomic radius 6. Melting point 7. Boiling point 8. Density



SEEKING PATTERNS

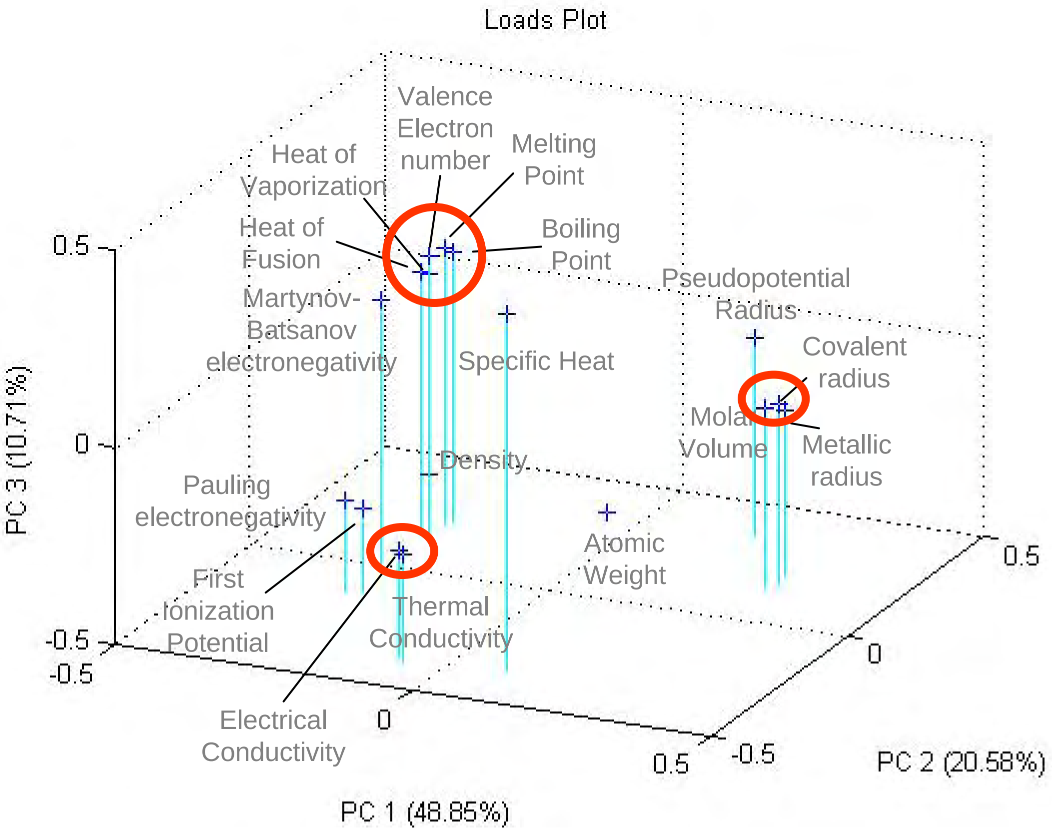


MENDELEEV SEQUENCING



Each cluster represent each row in the periodic table

DEVELOPING PHYSICAL LAWS



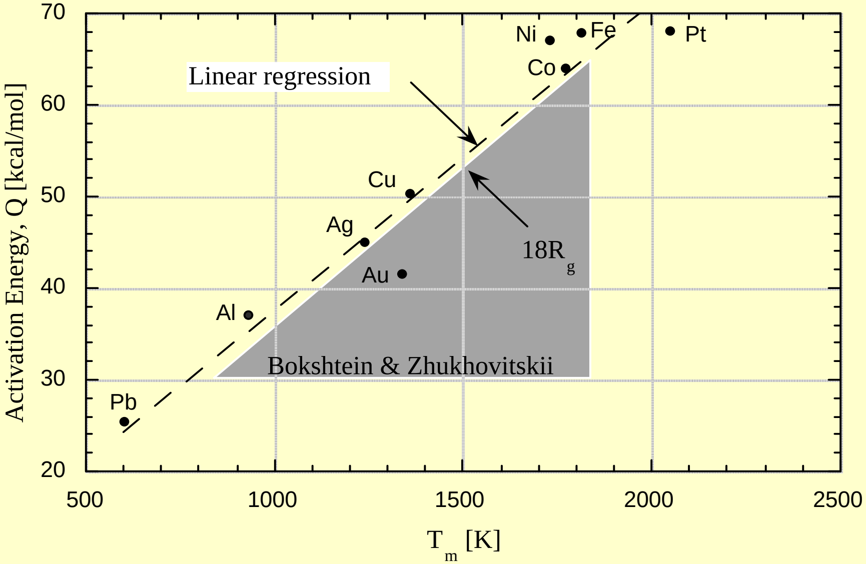
- **Location of property in loading plot indicates influence of property on PC**
  - Atomic weight has no influence on PC1, high influence on PC2 and 3

Clustered properties indicate relationships

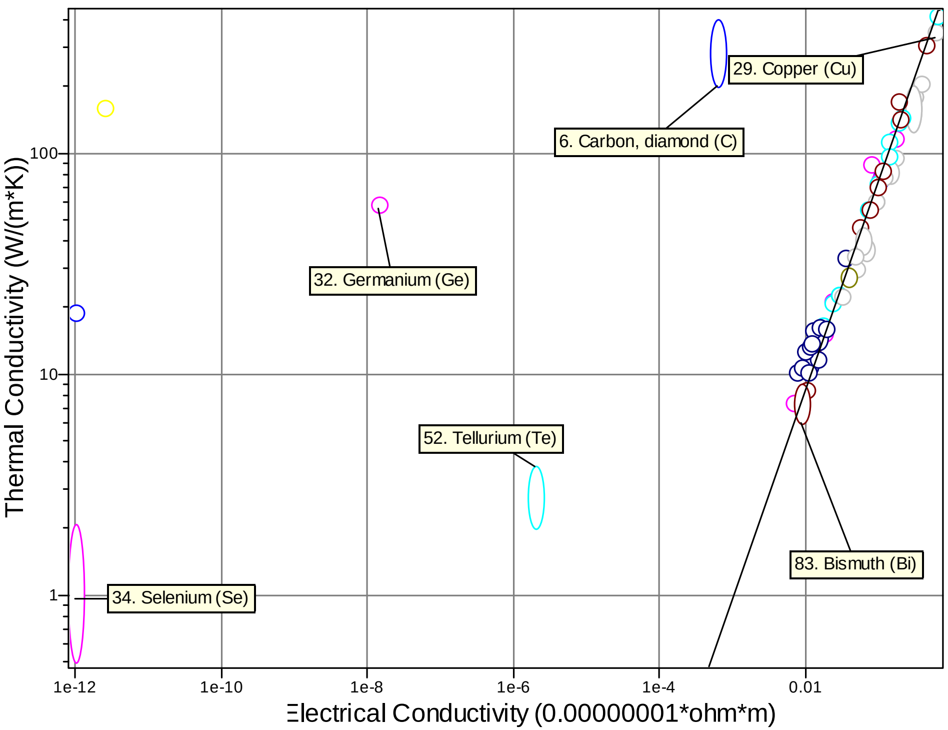
- Electrical and thermal conductivity
- Melting point and density
- Molar volume and atomic radius
- Melting and boiling points, heats of fusion and vaporization, and valence number
- Pauling electronegativity and first ionization potential

DEVELOPING PHYSICAL LAWS

Activation energy for self-diffusion versus melting point

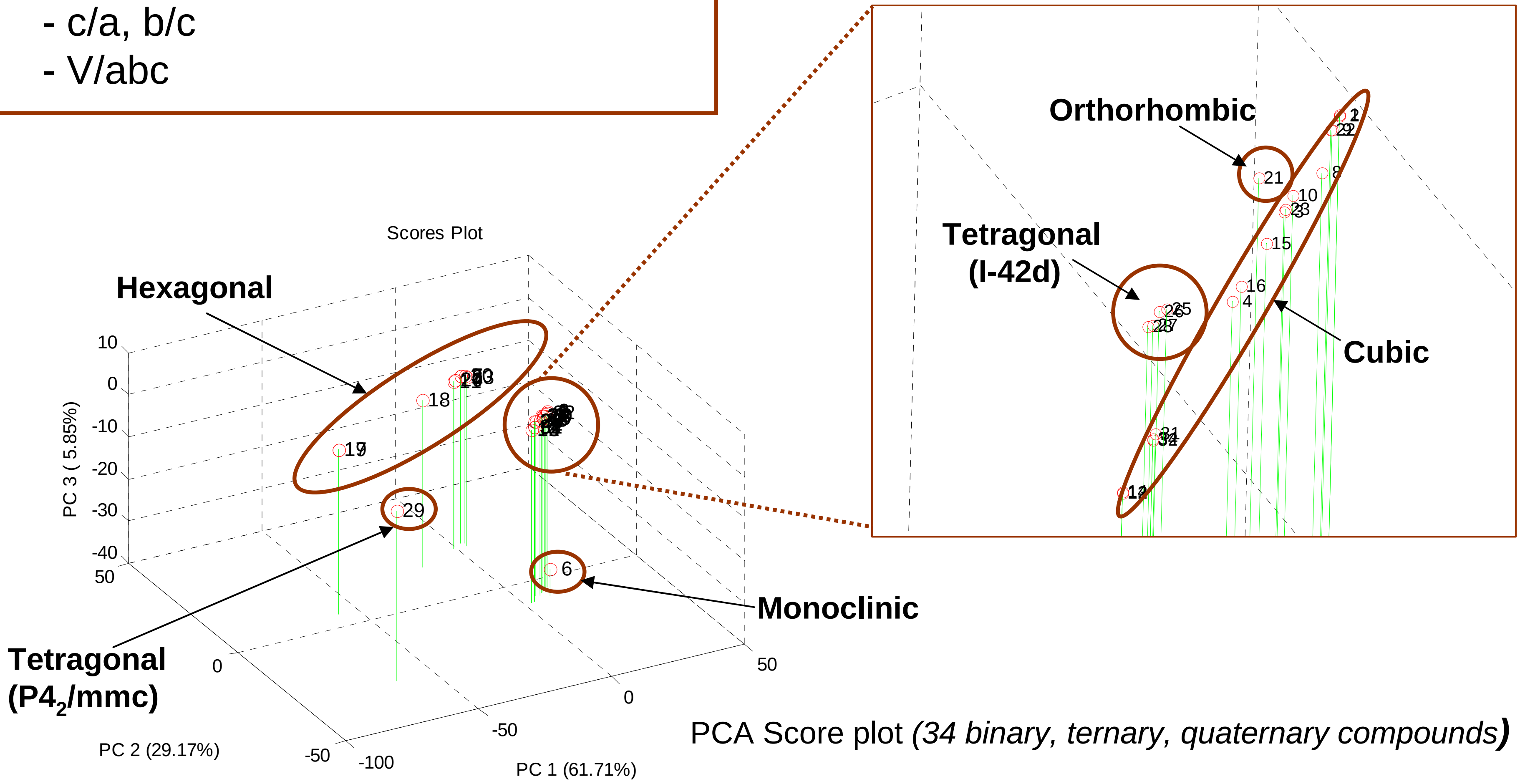


From Glicksman



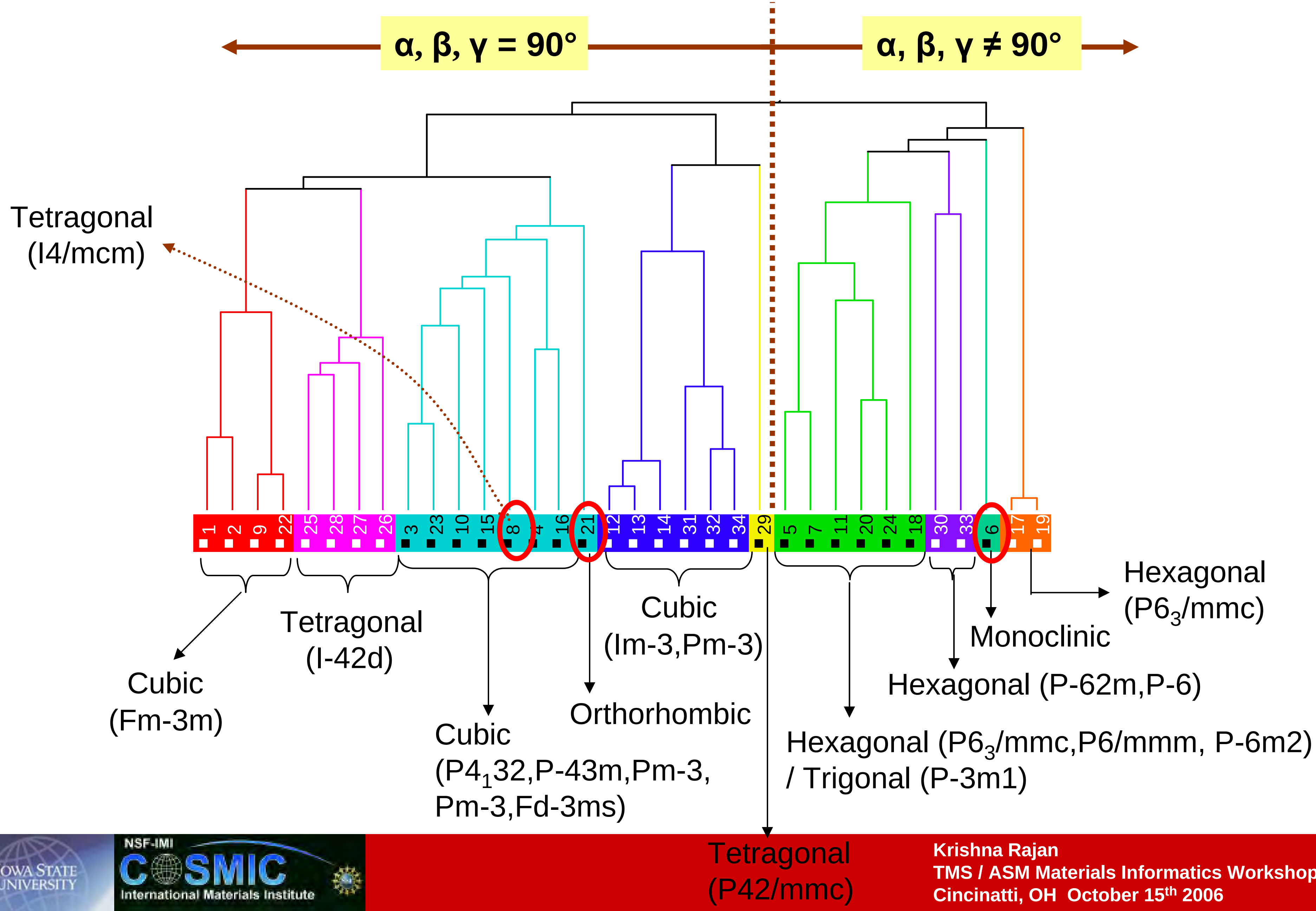
PCA of 34 binary, ternary, quaternary compounds: Score plot

9 descriptors from NIST data
- Lattice parameters (a, b, c, α, β, γ)
- c/a, b/c
- V/abc



PCA Score plot *(34 binary, ternary, quaternary compounds)*

Cluster analysis of 34 binary, ternary, quaternary compounds

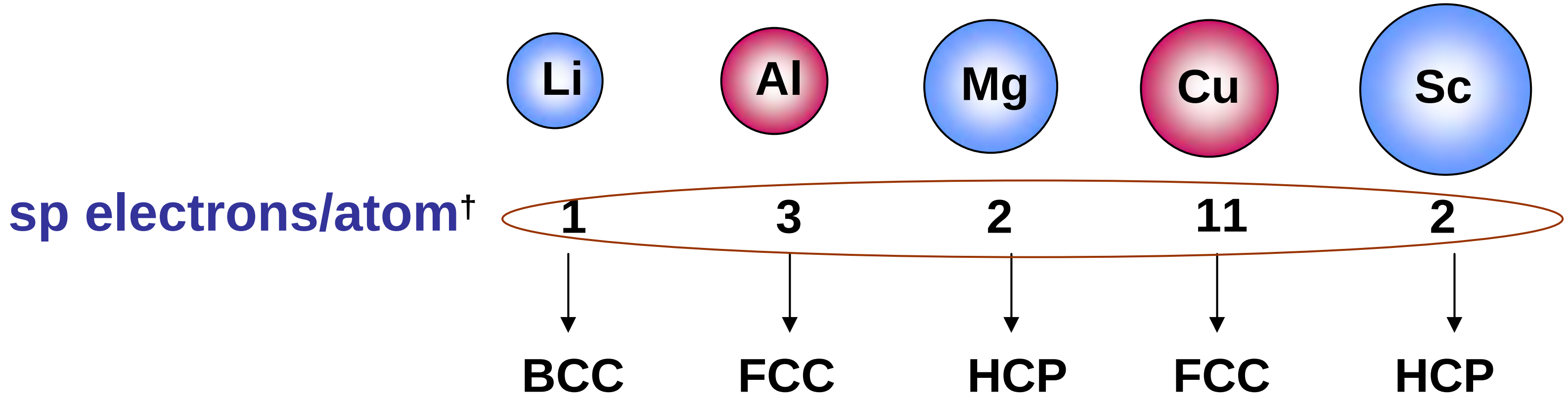


RULE EXTRACTION

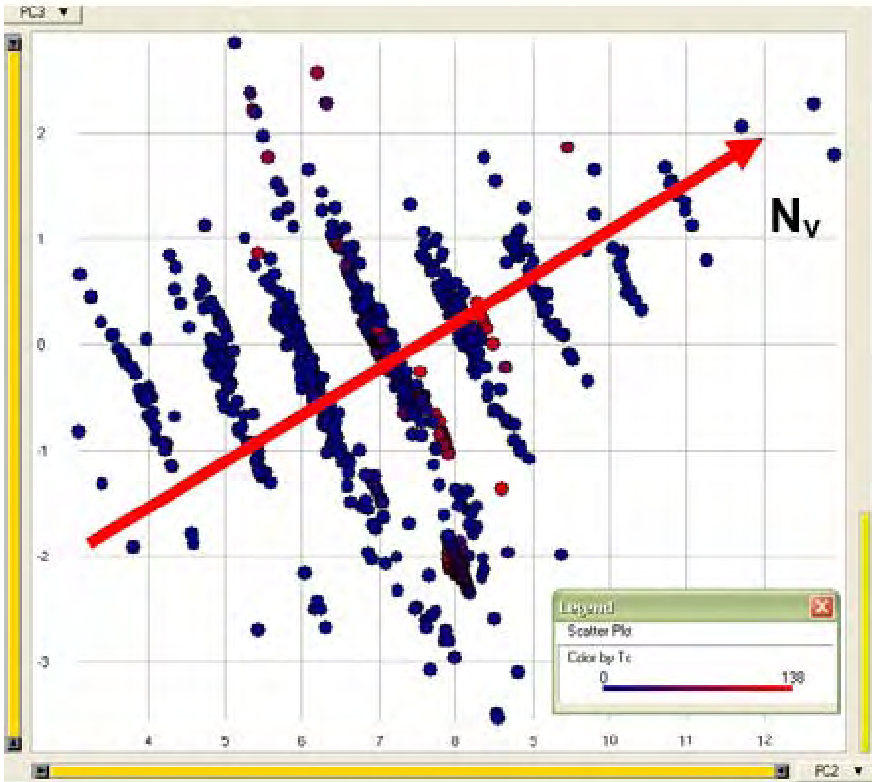
Atomic packing of single element

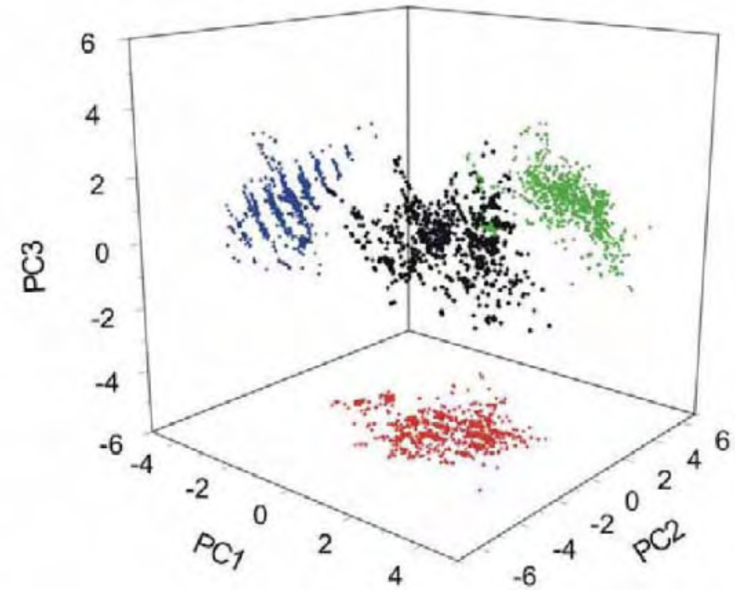
Ex.) Engel's model using "# of sp electrons/atom"†

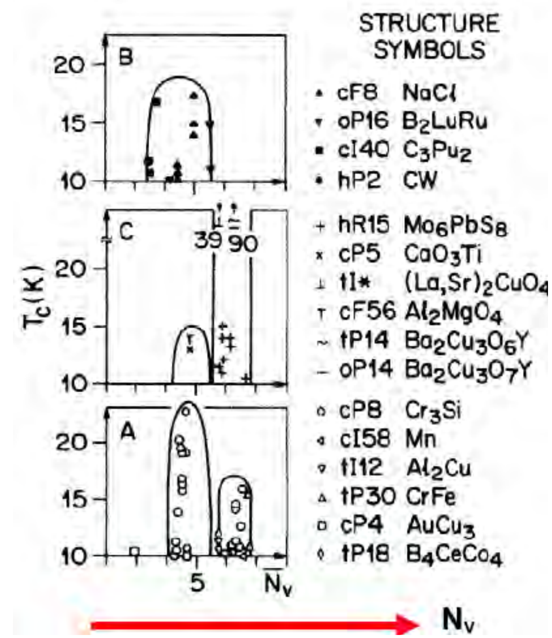
BCC < 1.5
HCP 1.7 – 2.1
FCC 2.5-3



PATTERN DETECTION

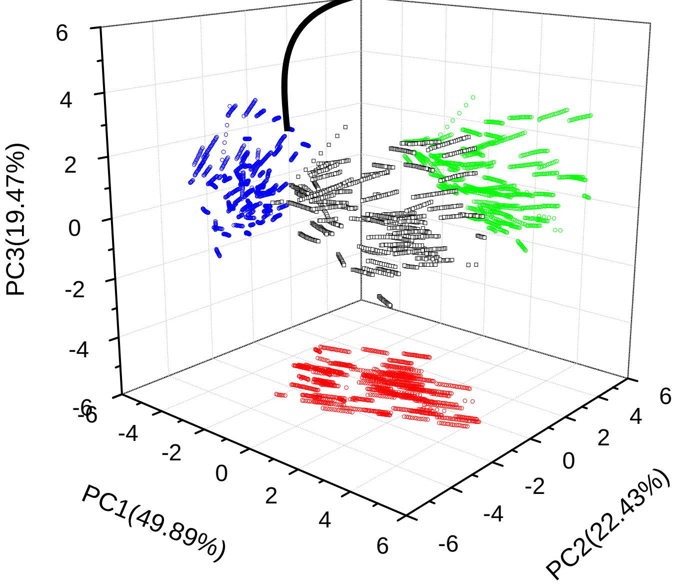


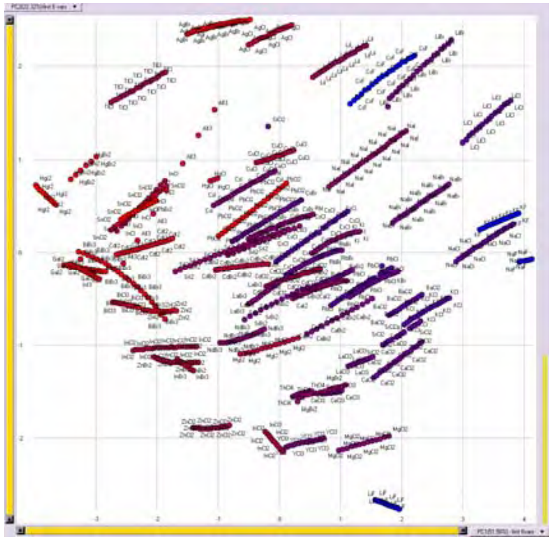




DATA MINING DATABASES

classification

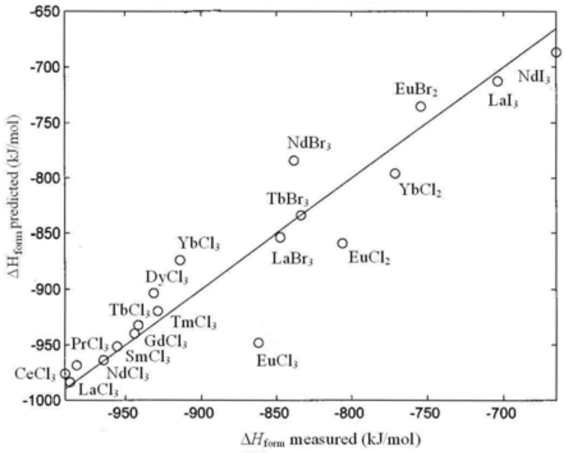




covalent

ionic

prediction



Graduciz, Gaune-Escard
Univ.Novi Sad, Serbia
CNRS, Marseilles

COMPUTATIONAL ISSUES

Focus on properties of signal / macroscopic behavior rather than noise/ error. **Assume complexity !!!**

Establish multivariate database:

Seek **DIVERSITY** in datasets

Data can come across length and time scales

Utilize data dimensionality reduction techniques

Analyze variation and correlation in data

Covariance

Establish correlations across diverse data sets (ie. length & time scales)

Identify outliers: explore cause

Develop predictive models

- Target requirements of missing data
- Quantitatively assess data diversity

Model relationships in data to seek heuristic relationships:

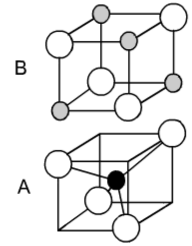
Advanced statistical learning tools can deal with:

- skewed data
- missing data
- differentiate between local and global minima
- ultra large scale datasets
- variable uncertainty

•Singular value decomposition
•Support vector machines
•Association mining
•Fuzzy clustering

CRYSTAL CHEMISTRY DESIGN

Ching et.al J. Amer. Ceram.Soc. ***85*** *75-80 (2002)*



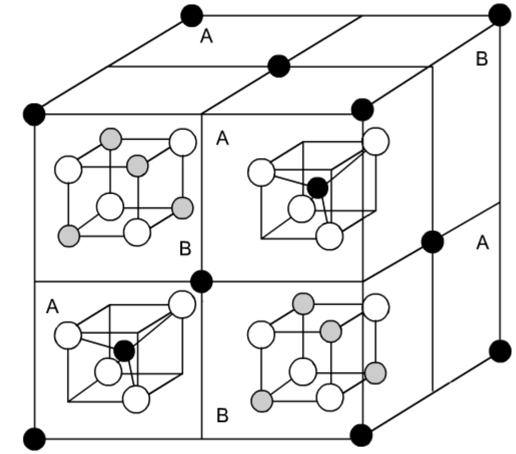


Table I. Summary of Calculated Properties of the Thirty-Nine Single and Double Spinel Nitrides

Crystal	*a* (Å)	*x*	Δ*E* (eV)	E_g (eV)†	Q^*_{tet}	Q^*_{oct}	Q^*_N	Bond order‡ Crystal	A–N	B–N
c-C_3N_4	6.8952	0.3832		1.14	3.70	3.63	5.27	8.647	0.358 (1.584)	0.241 (1.673)
c-Si_3N_4	7.8374	0.3844		3.45	2.65	2.58	6.05	8.670	0.362 (1.831)	0.241 (1.885)
c-Ge_3N_4	8.2112	0.3841		2.22	2.81	2.80	5.90	7.900	0.327 (1.907)	0.220 (1.982)
c-Sn_3N_4	8.9651	0.3845		1.29	2.71	2.70	5.97	6.958	0.284 (2.092)	0.195 (2.165)
c-Ti_3N_4	8.4460	0.3832		0.25 (d), 0.07 (id)	3.09	3.20	5.62	8.474	0.353 (1.949)	0.236 (2.045)
c-Zr_3N_4	9.1215	0.3830		0.40 (d), 0.23 (id)	3.06	3.17	5.65	8.609	0.356 (2.109)	0.240 (2.206)
c-Hf_3N_4	8.7038	0.3815		Metal	3.17	2.97	5.72	4.718	0.220 (1.982)	0.123 (2.121)
c-CSi_2N_4	7.5209	0.3811	−0.65	1.343 (d), 1.259 (id)	4.14	4.44	4.75	11.231	0.299 (1.714)	0.368 (1.832)
c-SiC_2N_4	7.2867	0.3885	3.08	Metal	2.45	3.71	5.53	8.260	0.359 (1.754)	0.225 (1.725)
c-CGe_2N_4	7.7429	0.3700	0.00	1.356	3.67	2.79	5.68	8.284	0.361 (1.616)	0.225 (1.970)
c-GeC_2N_4	7.4289	0.3942	3.84	0.707	2.85	3.67	5.45	7.816	0.317 (1.863)	0.220 (1.723)
c-$SiGe_2N_4$	8.0873	0.3772	−0.26	1.850	3.10	2.91	5.77	9.999	0.564 (1.790)	0.229 (2.000)
c-$GeSi_2N_4$	8.0008	0.3900	0.44	2.635 (d), 2.554 (id)	3.02	2.53	5.98	8.260	0.320 (1.946)	0.238 (1.885)
c-CTi_2N_4	7.8351	0.3637	−1.95	Metal	3.71	3.23	5.46	9.005	0.383 (1.550)	0.248 (2.046)
c-TiC_2N_4	7.5400	0.3937	4.51	0.965 (d), 0.62 (id)	2.97	3.77	5.37	8.030	0.352 (1.883)	0.217 (1.753)
c-$SiTi_2N_4$	8.2168	0.3753	−1.43	Metal	2.51	3.31	5.72	9.075	0.366 (1.785)	0.256 (2.051)
c-$TiSi_2N_4$	8.0470	0.3898	1.08	2.62 (d), 2.51 (id)	2.62	3.65	5.56	10.806	0.303 (1.956)	0.349 (1.896)
c-$TiGe_2N_4$	8.3159	0.3836	0.91	2.269 (d), 1.87 (id)	3.08	2.91	5.77	8.231	0.378 (1.932)	0.217 (2.006)
c-$GeTi_2N_4$	8.4002	0.3829	−0.44	Metal	2.90	3.18	5.68	8.657	0.323 (1.940)	0.253 (2.032)
c-$TiZr_2N_4$	8.9276	0.3800	0.95	0.32 (d), 0.15 (id)	3.14	3.14	5.64	8.482	0.339 (2.017)	0.240 (2.163)
c-$ZrTi_2N_4$	8.6806	0.3868	−0.13	Metal	4.10	1.14	6.40	7.261	0.452 (2.056)	0.152 (2.072)
c-CSn_2N_4	8.3600	0.3636	1.79	Metal	3.65	2.76	5.71	7.234	0.335 (1.650)	0.190 (2.187)
c-SnC_2N_4	7.7625	0.3988	5.65	1.00 (d), 0.99 (id)	2.70	3.71	5.47	7.093	0.273 (2.007)	0.204 (1.772)
c-CZr_2N_4	8.5091	0.3674	1.75	Metal	3.78	3.16	5.48	8.169	0.312 (1.738)	0.236 (2.189)
c-ZrC_2N_4	7.8186	0.3965	6.51	0.39	2.55	4.43	5.15	7.286	0.348 (1.991)	0.188 (1.799)
c-$SiSn_2N_4$	8.6279	0.3715	0.29	1.68	4.62	3.65	5.02	6.740	0.441 (1.816)	0.134 (2.188)
c-$SnSi_2N_4$	8.2479	0.3948	1.15	2.63 (d), 2.58 (id)	2.96	3.33	5.60	10.829	0.312 (2.076)	0.347 (1.909)
c-$SiZr_2N_4$	8.7484	0.3753	0.25	Metal	3.45	3.02	5.63	9.443	0.498 (1.878)	0.227 (2.197)
c-$ZrSi_2N_4$	8.2912	0.3928	1.71	3.315 (d), 2.78 (id)	2.62	3.59	5.55	10.619	0.315 (2.051)	0.338 (1.937)
c-$GeSn_2N_4$	8.7583	0.3795	0.41	Metal	2.88	2.66	5.95	7.184	0.311 (1.965)	0.196 (2.151)
c-$SnGe_2N_4$	8.4615	0.3895	0.06	2.31 (d), 2.28 (id)	2.68	2.83	5.92	7.536	0.291 (2.053)	0.217 (1.997)
c-$GeZr_2N_4$	8.9505	0.3807	0.73	Metal	3.00	3.12	5.69	8.442	0.289 (2.028)	0.255 (2.188)
c-$ZrGe_2N_4$	8.5689	0.3872	0.96	2.64 (d), 2.40 (id)	2.99	2.96	5.77	8.219	0.388 (2.040)	0.213 (2.029)
c-$SnTi_2N_4$	8.6340	0.3888	−0.20	0.0 (semi-metal)	2.75	3.22	5.70	8.430	0.289 (2.087)	0.255 (2.045)
c-$TiSn_2N_4$	8.8175	0.3780	1.06	2.27 (d), 1.88 (id)	3.21	2.76	5.82	7.705	0.384 (1.953)	0.193 (2.176)
c-$SnZr_2N_4$	9.1425	0.3859	0.52	Metal	2.80	3.17	5.72	8.419	0.270 (2.152)	0.261 (2.191)
c-$ZrSn_2N_4$	9.0475	0.3814	0.08	2.63 (d), 2.60 (id)	3.08	2.80	5.83	7.753	0.397 (2.059)	0.191 (2.206)
c-$ZrHf_2N_4$	8.9223	0.3853	−1.523	−0.02 (semi-metal)	2.82	3.09	5.75	4.473	0.163 (2.091)	0.132 (2.143)
c-$HfZr_2N_4$	8.9922	0.3815	0.035	0.24 (d), 0.10 (id)	3.23	2.83	5.78	4.379	0.199 (2.057)	0.116 (2.193)

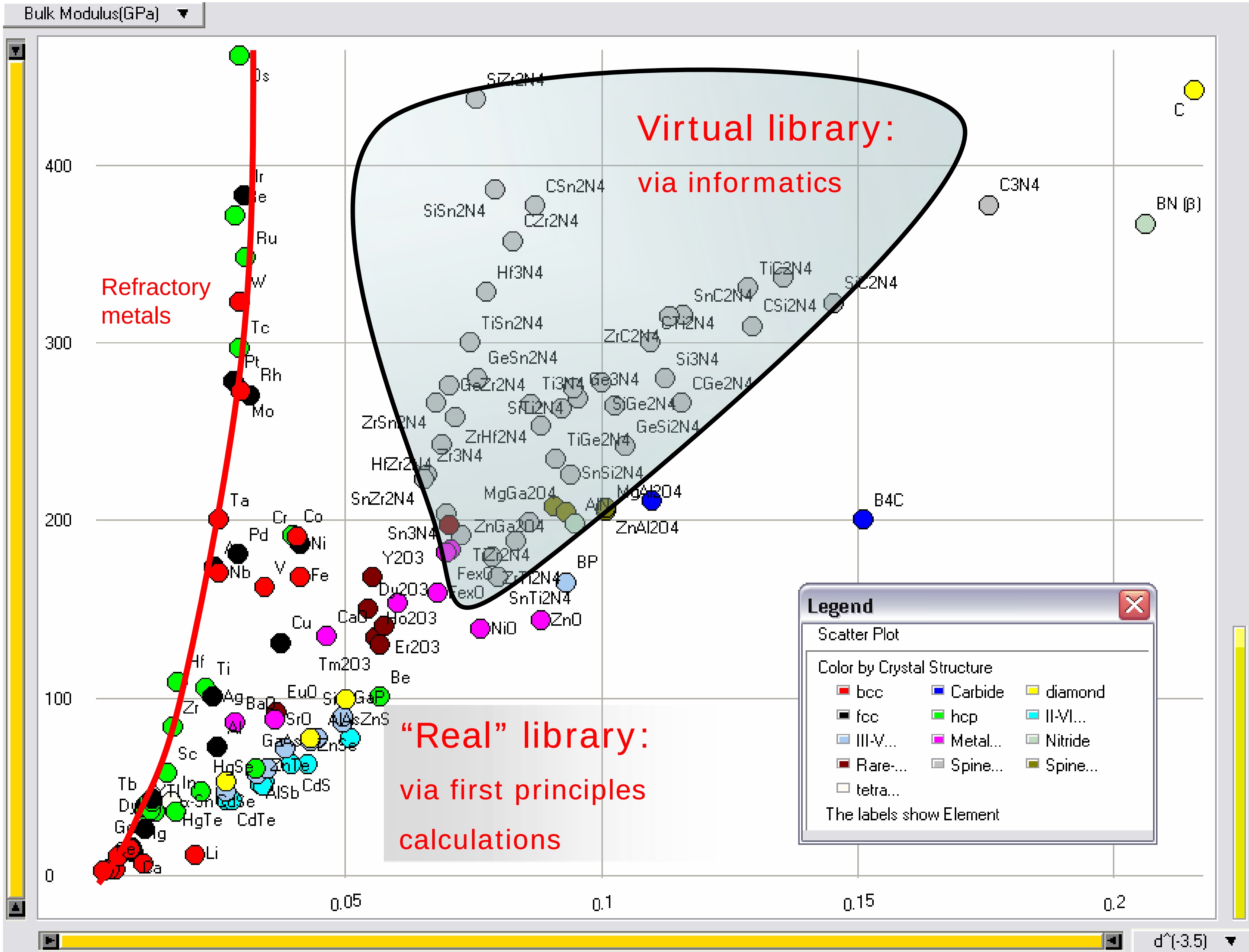
†"d" denotes a direct band gap and "id" denotes an indirect band gap. ‡Bond length given in parentheses, in units of Å.

1. Assess influence of latent variables (i.e. electronic structure parameters) on properties of known data

2. Establish heuristic relationships on database of *all input* variables instead of phenomenological relationships in bivariate manner

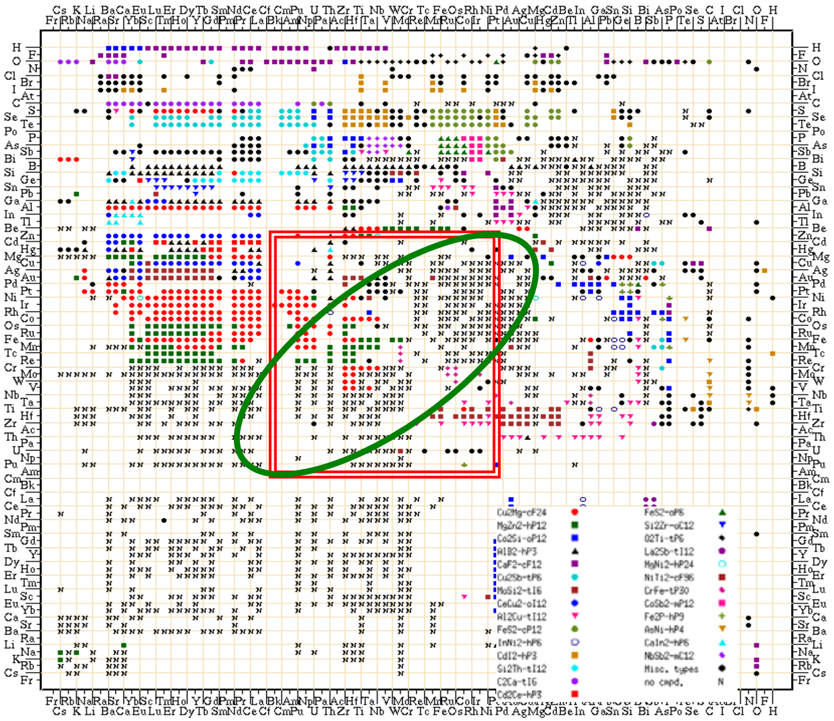
3. Use statistical learning to predict new materials behavior on new multivariate *input* data

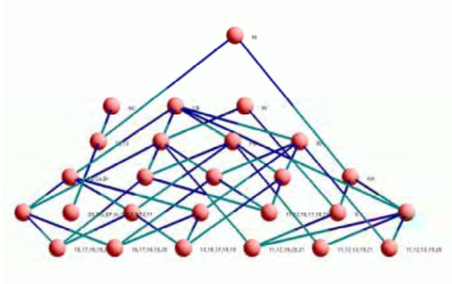
4. Inverse problem approach to formulate quantitative structure-property relationships

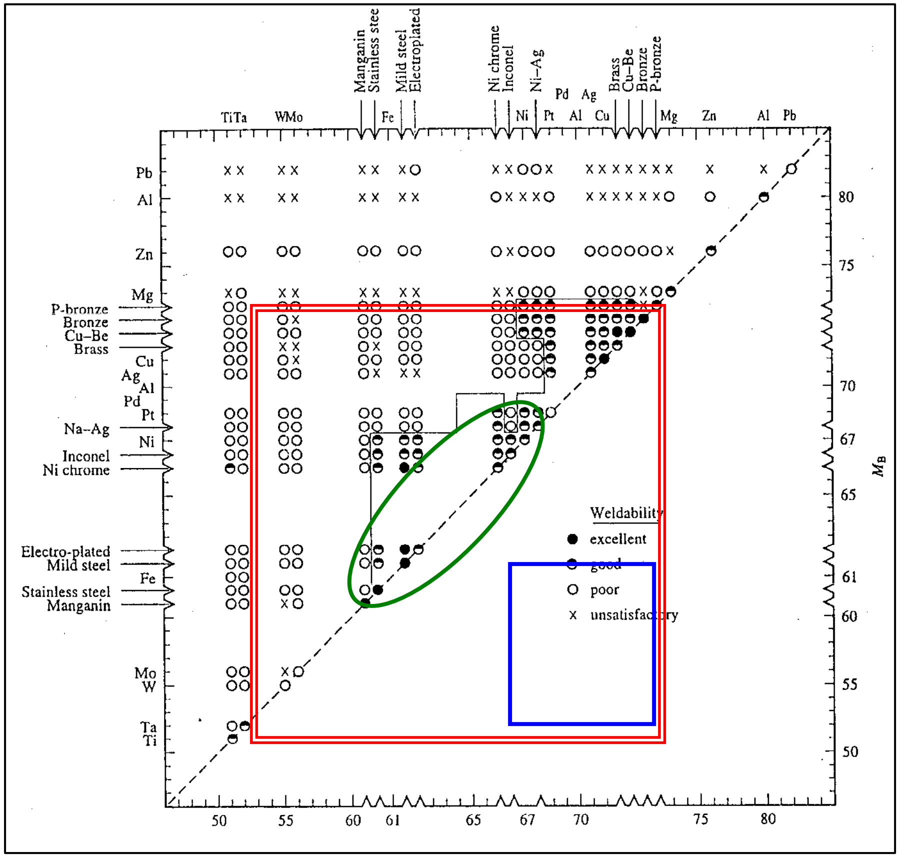

Bulk Modulus(GPa)
Virtual library:
via informatics
Refractory
metals
"Real" library:
via first principles
calculations
400
300
200
100
0
0.05
0.1
0.15
0.2
d^(-3.5)
Legend
Scatter Plot
Color by Crystal Structure
bcc
fcc
III-V...
Rare-...
tetra...
Carbide
hcp
Metal...
Spine...
diamond
II-VI...
Nitride
Spine...
The labels show Element
C
C3N4
BN (β)
B4C
Os
Ir
Ru
W
Tc
Rh
Mo
Ta
Cr
Co
Pd
Ni
Nb
V
Fe
Cu
Hf
Ti
Zr
Sc
Li
SiZr2N4
CSn2N4
SiSn2N4
CZr2N4
Hf3N4
TiC2N4
SnC2N4
SiC2N4
CSi2N4
TiSn2N4
ZrC2N4
Si3N4
GeSn2N4
CGe2N4
ZrSn2N4
ZrHf2N4
GeSi2N4
TiGe2N4
Zr3N4
SnSi2N4
SnZr2N4
MgGa2O4
ZnAl2O4
Sn3N4
ZnGa2O4
Y2O3
BP
SnTi2N4
Ho2O3
Er2O3
NiO
ZnO
Tm2O3
Be
EuO
SrO
CdS
HgTe
CdTe
AlSb

LINKING DISPARATE LENGTH SCALES

Visualizing Associations

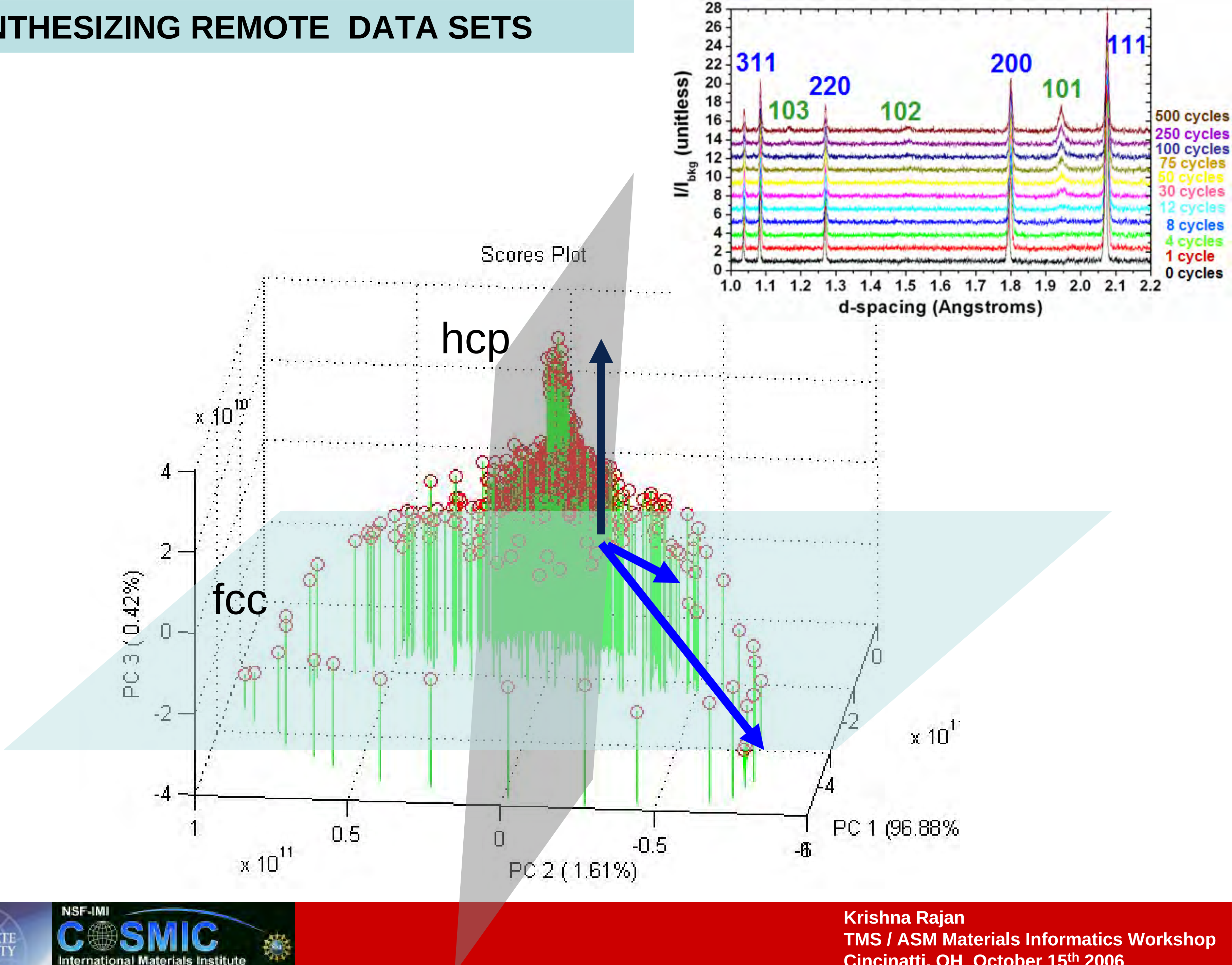






Krishna Rajan
TMS / ASM Materials Informatics Workshop
Cincinatti, OH October 15th 2006

SYNTHESIZING REMOTE DATA SETS



Krishna Rajan

STRATEGY FOR MATERIALS DESIGN

- Search and retrieve
- Refining descriptors
- Developing predictions
- Filling in missing data

- Structure databases
- Diffraction spectra databases
- Hyper spectral imaging databases

Mapping / Visualization of databases:

- Correlations
- Trend analysis

CONCLUSIONS

Science was originally empirical, like Leonardo, making wonderful drawings of nature. Next came the theorists who tried to write down the equations that explained the observed behaviors, like Kepler or Einstein. Then when we got to complex enough systems like the clustering of a million galaxies, there came the computer simulations, the computational branch of science. Now we are getting into the data exploration part of science, which is kind of a little bit of them all"..

Dr. Alex Szalay: Virtual Observatory Project

The New York Times

May 20th 2003

The Facts of the Matter: Finding, understanding, and using information about our physical world (2000)

Information science based design of materials……next stage of the scientific discovery process

National Science Foundation
WHERE DISCOVERIES BEGIN

Division of Materials Research:
Cyber infrastructure and cyber discovery in materials science (2006)