マテリアルズ・インフォマティクス：機能無機材料探索における機会と課題

Materials Informatics: Opportunities and Challenges for Inorganic Materials

田中　譲

物質・材料研究機構MaDIS・MIリサーチアドバイザー

TANAKA.Yuzuru@nims.go.jp

北海道大学・工学研究院・名誉教授　/　触媒科学研究所・客員教授

tanaka.yzr@ist.hokudai.ac.jp

アルバータ大学コンピューティング・サイエンス学科・非常勤教授

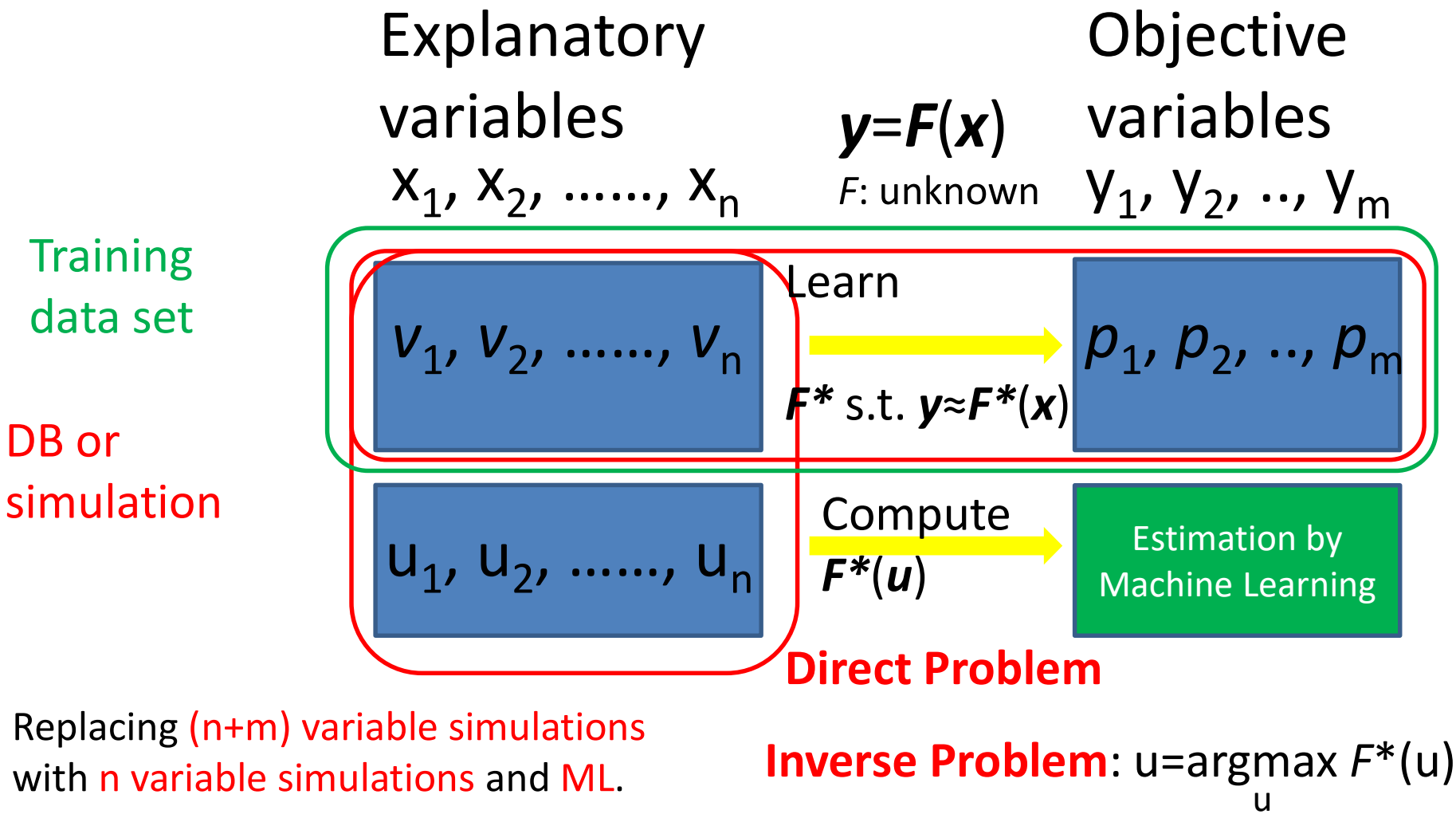
yuzuru@ualberta.ca

Materials Informatics: Opportunities

Paradigm Shift
From X Science to X Informatics

- X?
  - bio / biomedical / **chemical** / geo / brain / cosmological / meteorological / pharmaceutical / **materials** / …
  - Bioinformatics has been a first big success.
  - Precision medicine and materials informatics are followers.
- chemistry / pharmacy / materials science
  - Discovery of new functional materials / drugs
    - empirically based: years
    - simulation based (ab initio calculation): hours, days
    - machine learning: seconds ← just for candidates discovery

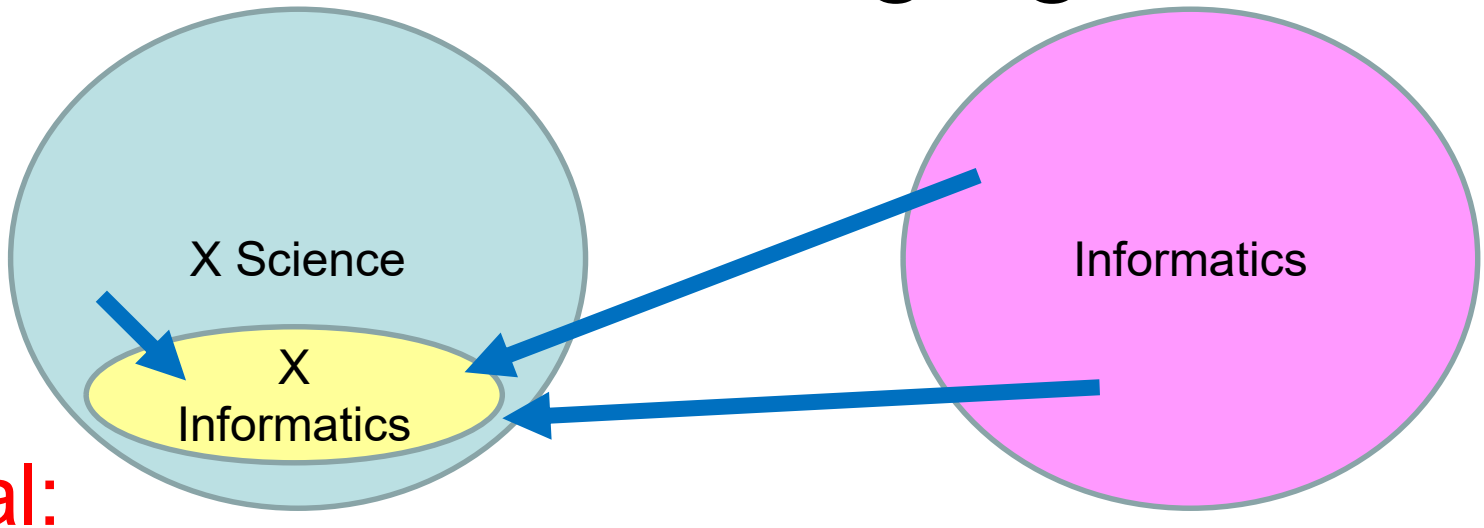
ML (Machine Learning) for Speed-Up



Replacing (n+m) variable simulations with n variable simulations and ML.

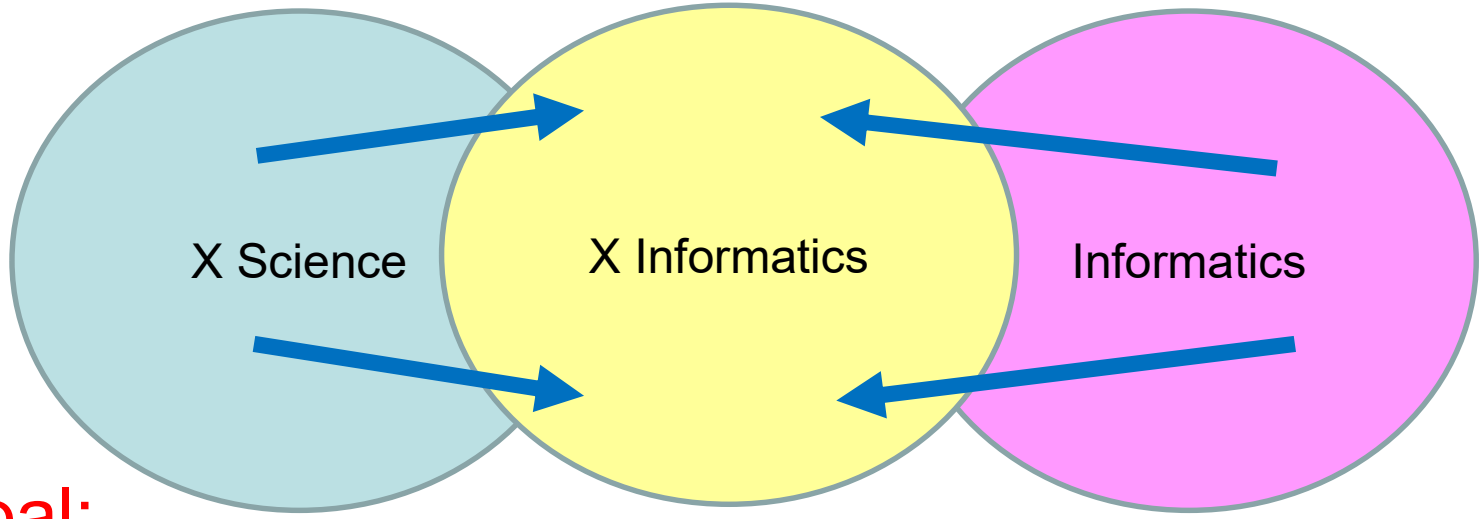
Inverse Problem: $u=\underset{u}{\operatorname{argmax}} F^*(u)$

X Informatics: Emerging Period



- Goal:
  - same as before
- Method:
  - domain methods + ML
- Information scientists help X scientists.

X Informatics: Mature Period



- Goal:
  - same as before
  - Unexplored artificially designed domain of materials
- Method:
  - domain methods + (math. modeling / design) + ML
- X informaticians

Materials Informatics

- Current status: **end** of the emerging period
  - 2 major objectives:
    (1) To replace DFT computation with ML for **speed-up**
    (2) To optimally **guide the exploration** of the target space to decide which material to choose next for DFT computation or experiment ← **Bayesian Optimization**

- **Now, facing an intrinsic big barrier**, especially in inorganic materials informatics.
  - Almost the end of initial successes
  - Need for "Transcending Incrementalism"

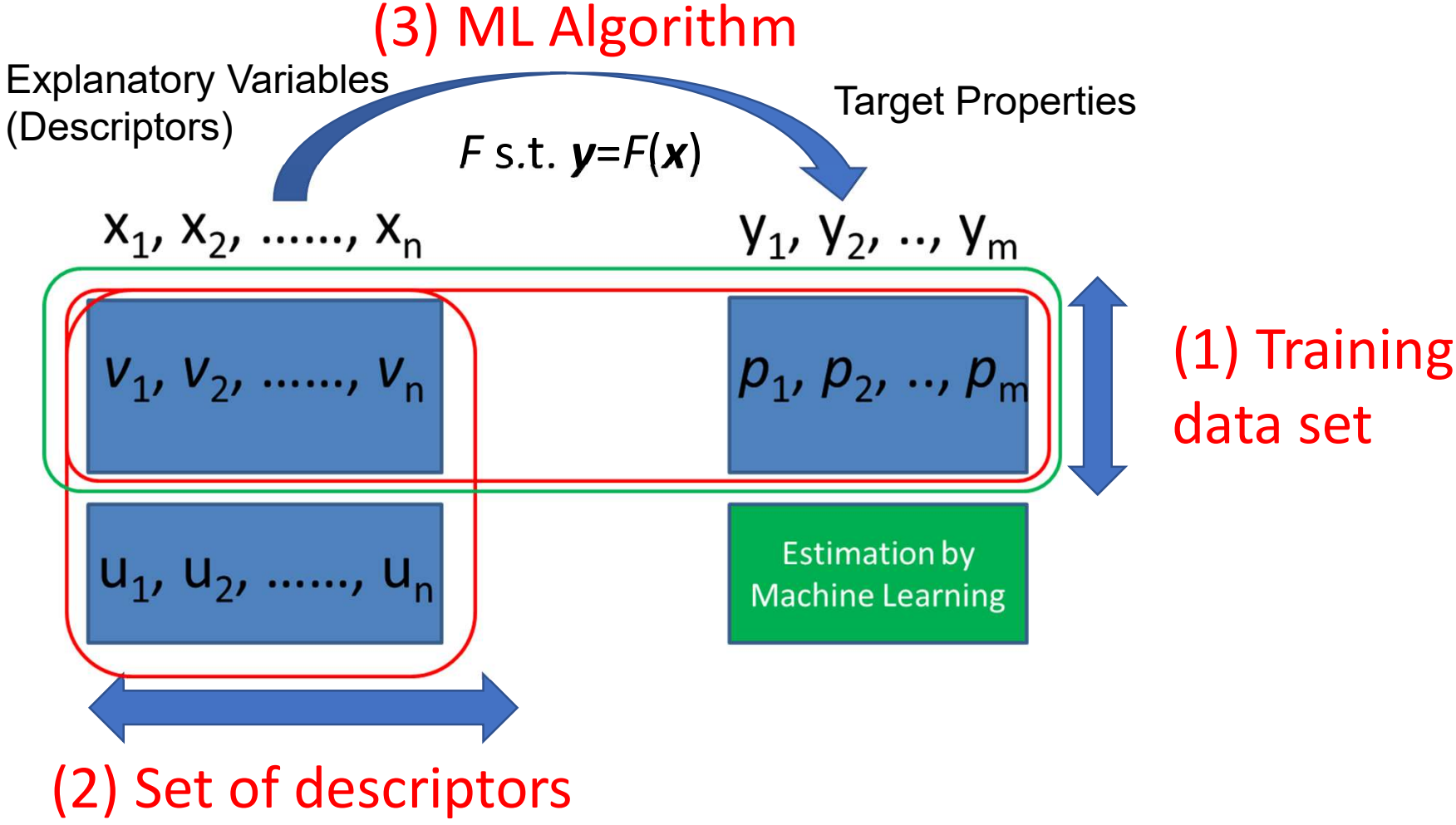
What ML to learn?

3 Major Goals

This Talk

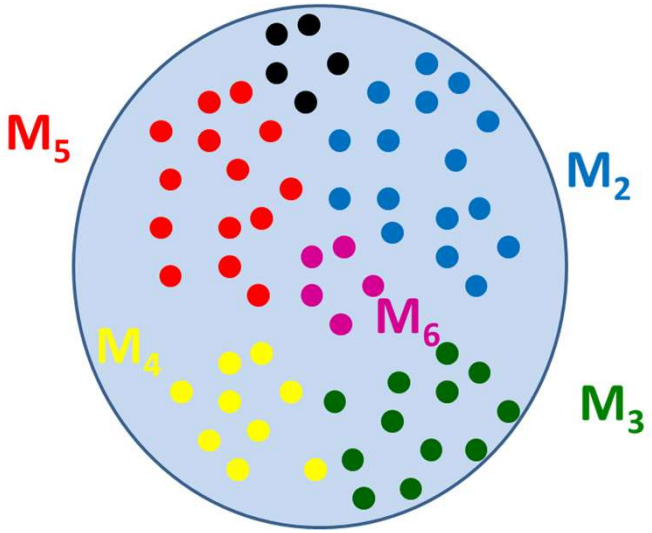
- **Materials Discovery**: Find the material with maximum performance
  - DFT to compute F : Structure → Performance
  - ML (regression) to learn a good approximation F^* of F as an explicit or computable function
  - Inverse Problem: $\mathrm{argmax}_{\boldsymbol{x}} F^*(\boldsymbol{x})$
- **Measurement Analysis**: Identify the material structure from its measurement result
  - (Measurement Data) + Simulation Data: F: Structure → Property
  - ML (Deep Learning) to learn a good approximation $(F^{-1})^*$ of F^{-1} as a computation mechanism (F should be bijective, otherwise Deep Learning does not converge.)
  - Evaluate $(F^{-1})^*$ for a given measurement chart or image to identify its structure.
- **Literature-based Knowledge Discovery**
  - Network of conditional or unconditional causality relations as a directed graph or a catalytic reaction network → discovery of new knowledge through inference

3 Things to Consider



(1) Training Data Set

- Heterogeneous!
  - Different groups follow different math models.
- Small homogeneous data set
  - Inorganic materials: $10^3 \sim 10^4$



How to segment the whole data set into a set of smaller homogeneous data sets?

Neither simple classification nor clustering algorithm does not give such segmentation!

(2) Set of Descriptors (cont.)

(1) How to **systematically define** a large set of descriptors?

(2) The target materials may have different number and different order of components, and hence **different number and order of component properties**! How to define descriptors of each compound independent from the order of its component description?

(3) How to deal with **non-stoichiometric compounds**? e.g. $Nd_2(Fe_xCo_{(1-x)})_{14}B$

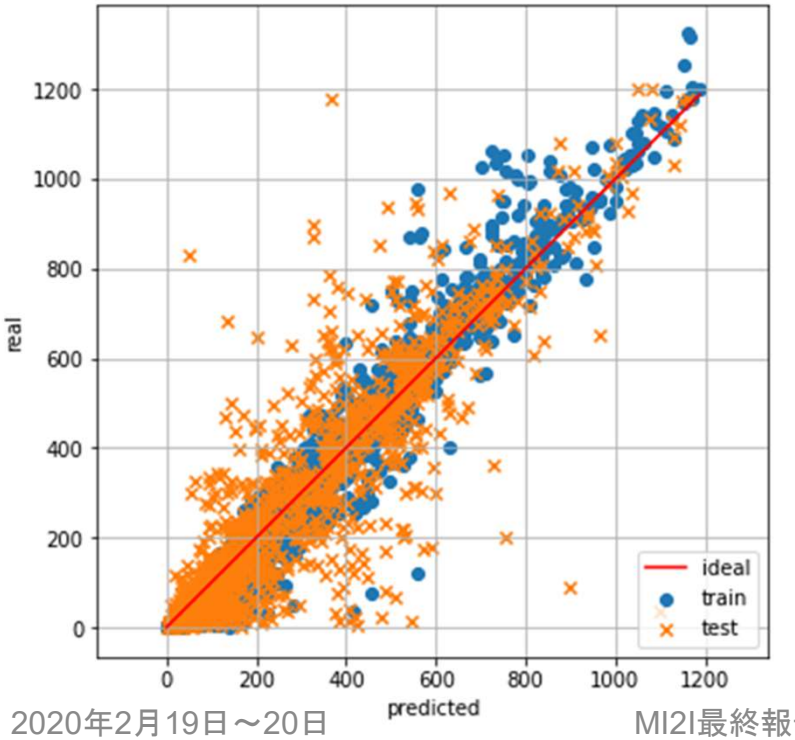
Descriptors of a compound independent from the order of its component description

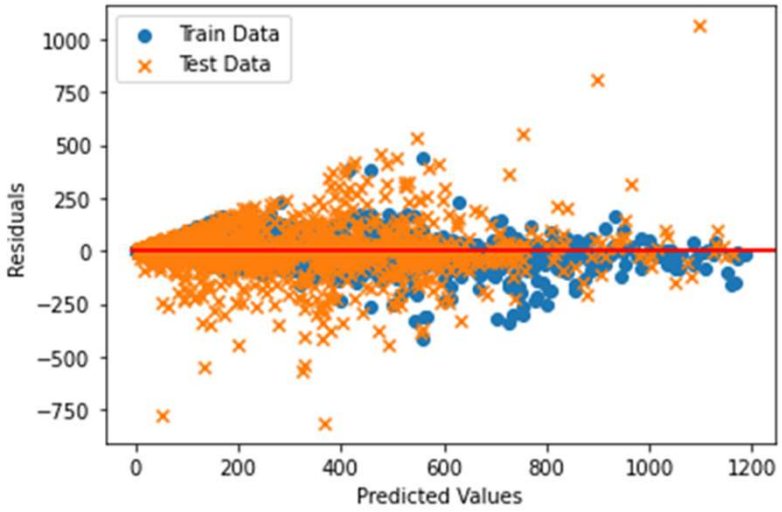
142 descriptors defined below:

- Property mol% descriptors (13)
  - Total mol% of the constituent elements that satisfy each atomic condition C defined in Pymatgen
- Group mol% descriptors (19)
  - Tatal mol% of the constituent elements in each group
- Atom mol% descriptors (103)
  - The mol% of each atom in the compound
- Block mol% descriptors (7)
  - Total mol% of constituent elements in each of s-block, p_L-block, p_H-block, d_L-block, d_H-block, f_L-block, f_H-block, where x_L-block and x_H-block respectively denote the less-than-half-filled and more-than-half-filled x-blocks.

Regression does not work well for the whole heterogeneous data set.

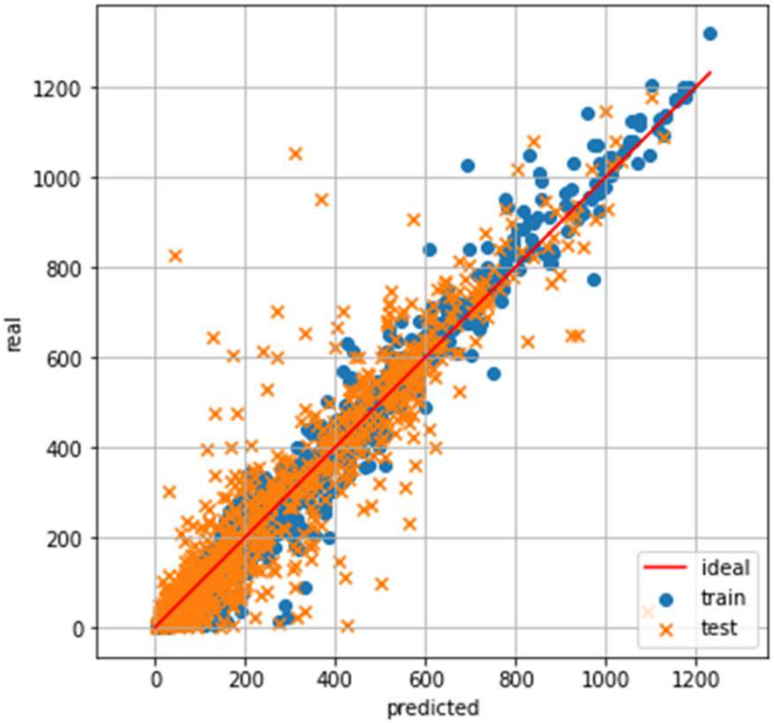
- Example: 4294 materials in AtomWork Adv. with Curie temperature Tc
- RF-regression of Tc (R2: 0.79434 in 3-fold CV)

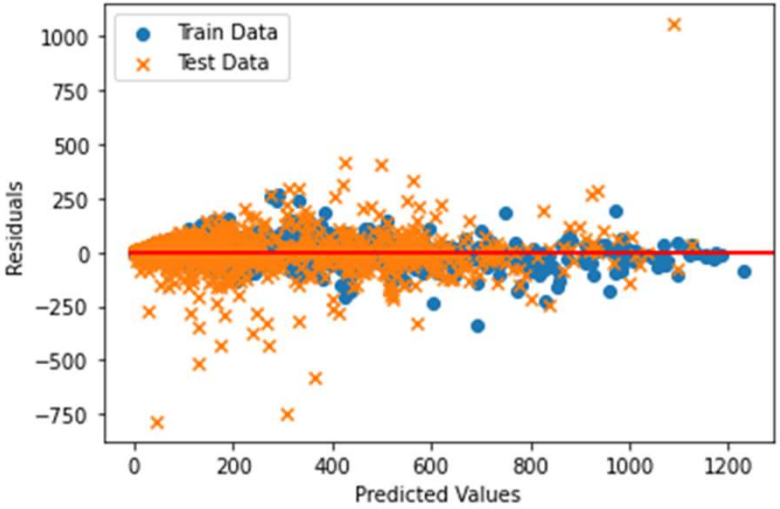




Regression of Tc for those with an f-block element

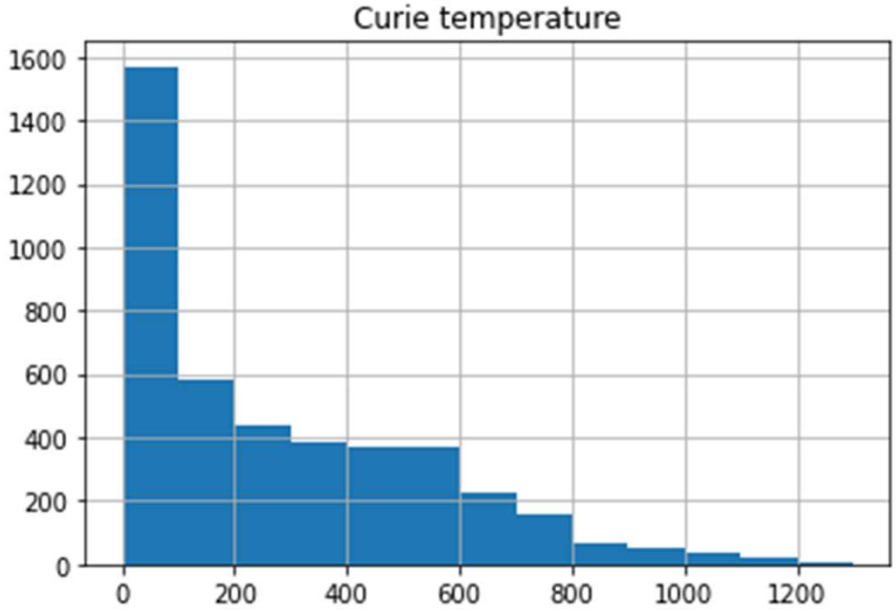
- 3135 materials in AtomWork Adv.
- RF-regression of Tc (R2: 0.86225 in 3-fold CV)





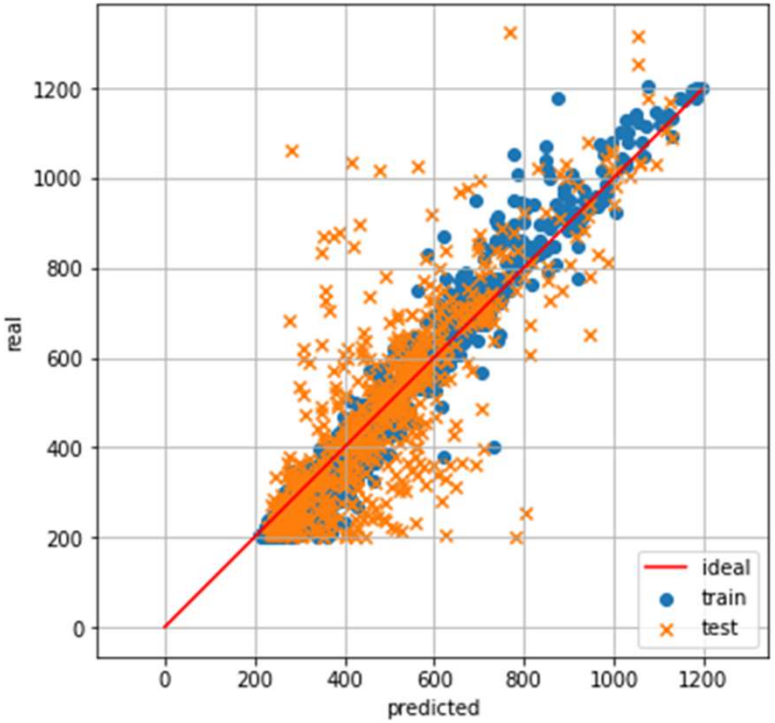
We focus on the high Tc region (Tc ≥ 200)

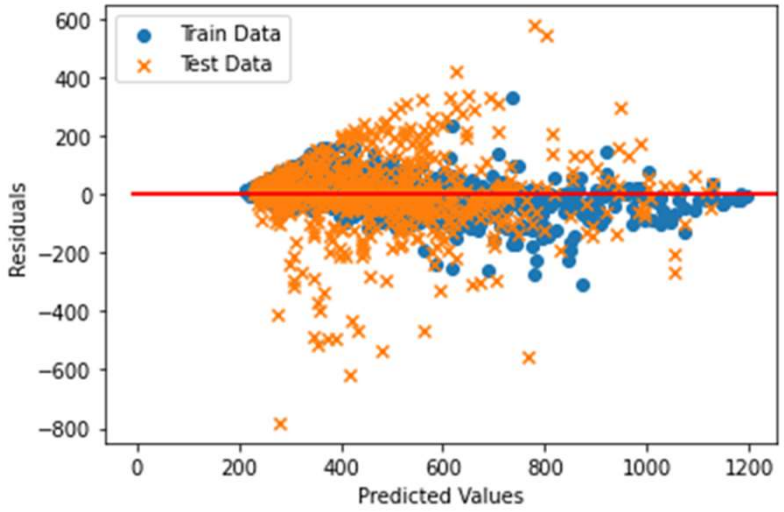
- Too much concentration of data within the low Tc region.
- We are interested in predicting Tc in the high Tc region (Tc ≥ 200) . Training data should be selected from this region.



Regression of Tc (≥ 200) does not work well for the whole heterogeneous data set.

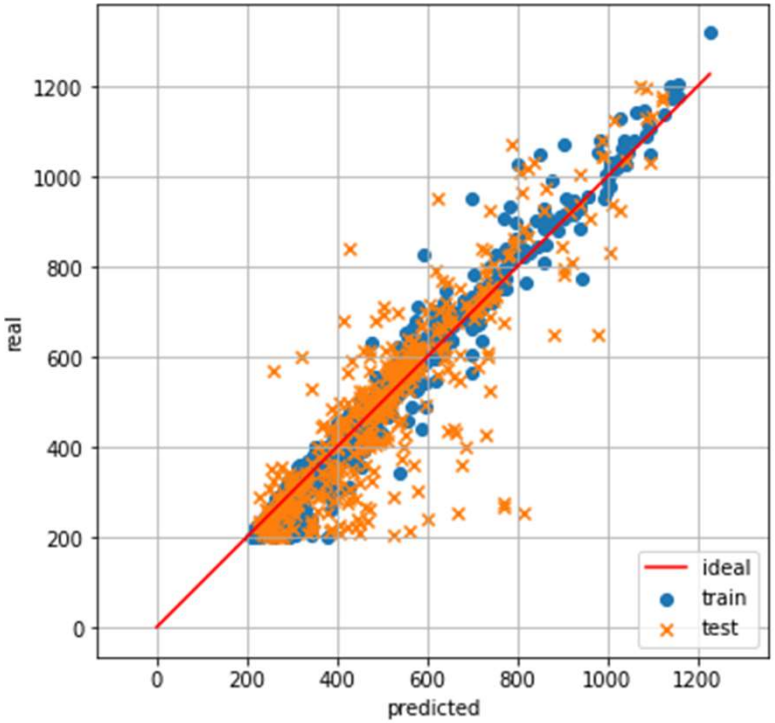
- Example: 2138 materials in AtomWork Adv. with Curie temperature Tc ≥ 200
- RF-regression of Tc (R2: 0.64103 in 3-fold CV)

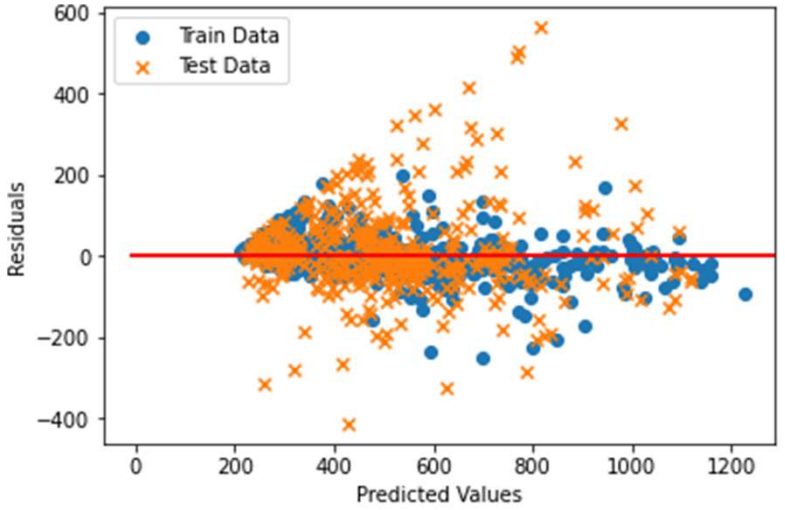




Regression of Tc (≥ 200) for those with an f-block element

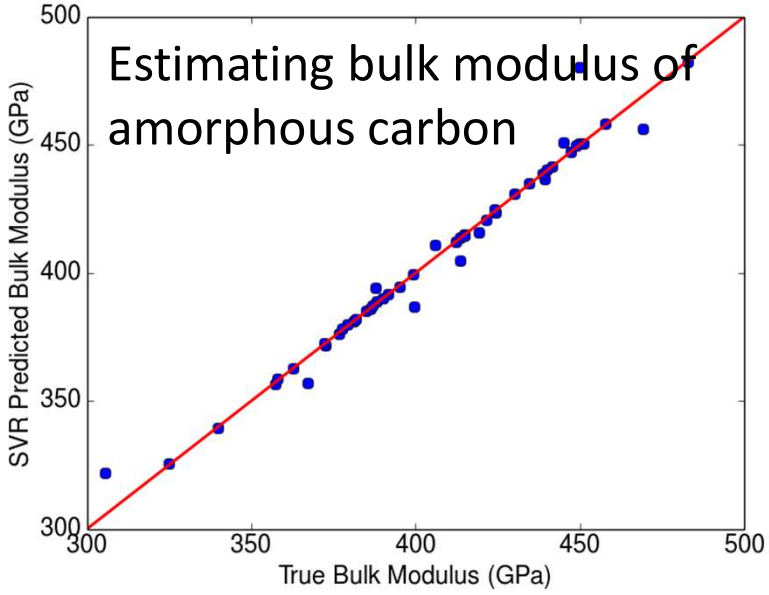
- 1385 materials in AtomWork Adv. with Tc ≥ 200
- RF-regression of Tc (R2: 0.79003 in 3-fold CV)

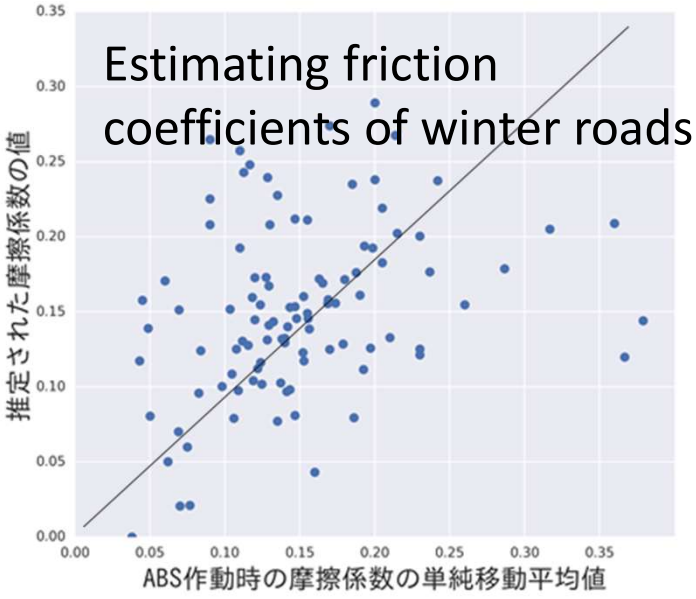




Once segmented to homogeneous systems, each follows a math model.

- *SVR works well to find a hidden physical order.*
  - Different from data sets in other research areas





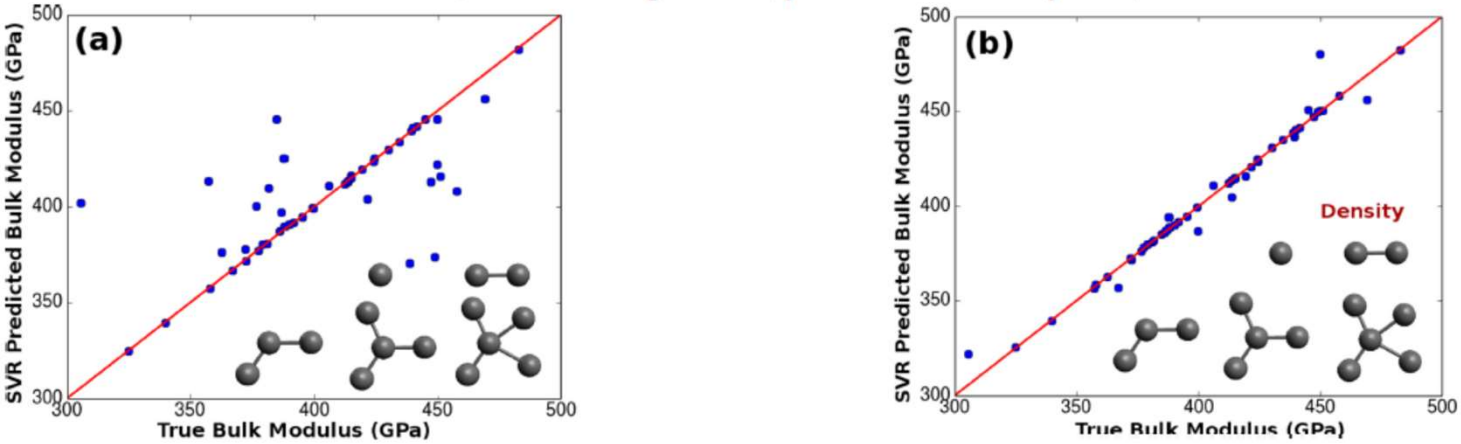
To find good descriptors

PHYSICAL REVIEW B 95, 054110 (2017)

Unveiling descriptors for predicting the bulk modulus of amorphous carbon

Keisuke Takahashi*
Center for Materials Research by Information Integration (CMI²), National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan and Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan

Yuzuru Tanaka
Meme Media Laboratory, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan
(Received 5 August 2016; published 14 February 2017)



Predicted bulk modulus against true bulk modulus with descriptors:
(a) the number of bonds in each C atom and (b) the number of bonds in each C atom with **density**. Structure models of bond type in amorphous carbon are also shown.

(3) ML Algorithm

- Deep Learning
  - Applicable to a large learning data set, not to any data set smaller than 10^4.
  - Not transparent or debuggable
  - *F*: computable, not explicitly obtained
- Regression
  - Linear Regression, SVR
    - *F*: explicitly obtained as a function
      - differentiable, integrable
      - Numerical solution to inverse problems
  - Random Forest Regression
    - *F*: implicitly obtained as a computation

Materials Informatics:
Challenges for Inorganic Materials

My Mission at MI^{2}I since April 2017

(1) Integrated Exploratory Visual Analytics Platform

- Joint work with Dr. Jun Fujima
- MADS (Materials Acquisition by Data Science)
  - poster DSG06

(2) New MI Methods for Inorganic Crystal Materials

- Joint work with Dr. Keisuke Takahashi, Dr. Jun Fujima, and Dr. Hiori Kino
- Small and heterogeneous data sets
- No systematic way to define appropriate descriptors
  - poster DSG10

(3) MI research advisor

Joint Research with

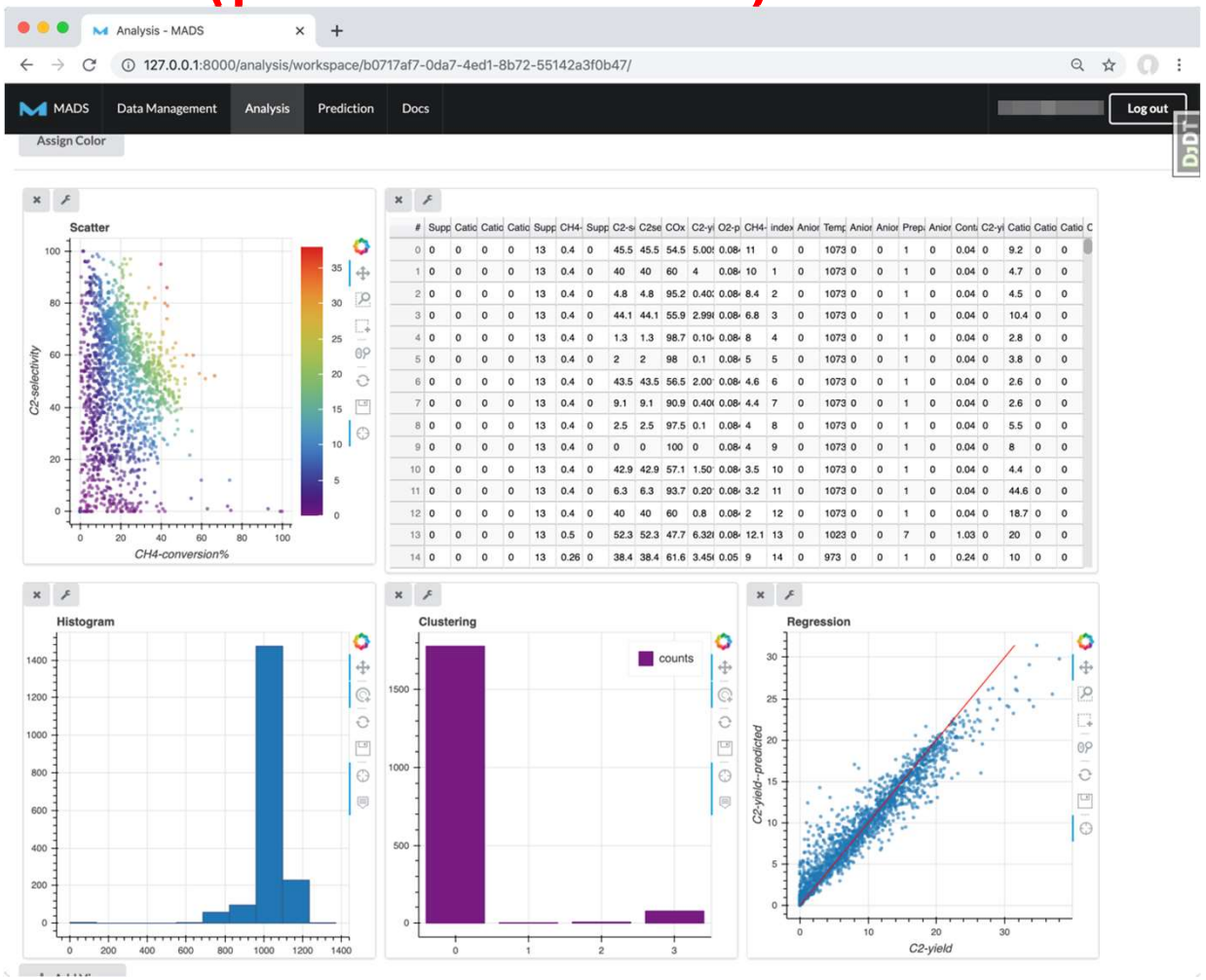
- 藤間淳
  - Fraunhofer研究所より招聘
  - 物質・材料研究機構MaDIS　MI^{2}I特別研究員
- 高橋啓介
  - 他専攻博士課程学生時代に共同研究開始
  - 研究代表者としてCREST獲得
  - 北海道大学理学研究院　准教授
- 木野日織
  - 磁性材料の専門知識で助言

How to apply ML?

- Not a straightforward way!
- Trial and error, back and forth process
  - Data curation
  - Data segmentation
  - Choice of an appropriate ML algorithm
  - Tuning of meta parameter values
  - Designing an appropriate analysis scenario

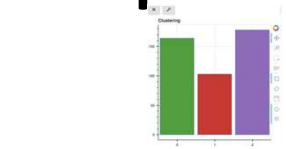
- Need for an Exploratory Visual Analytics Platform

Integrated Exploratory Visual Analytics Platform (poster DSG06)

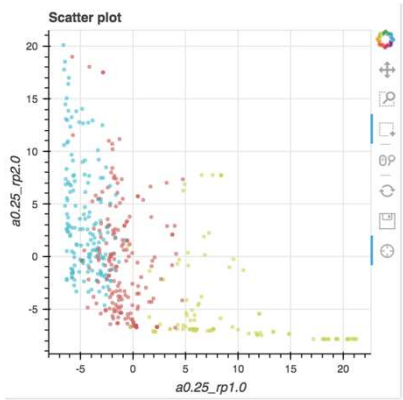


Brushing and linking in Coordinated Multiple Views

structure id	total energy	descriptors

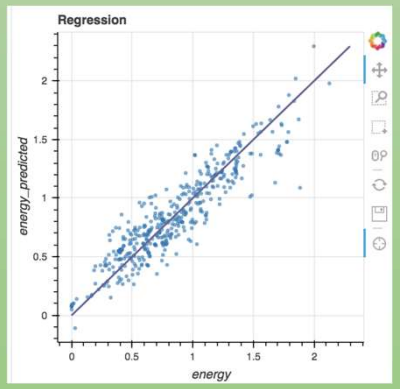


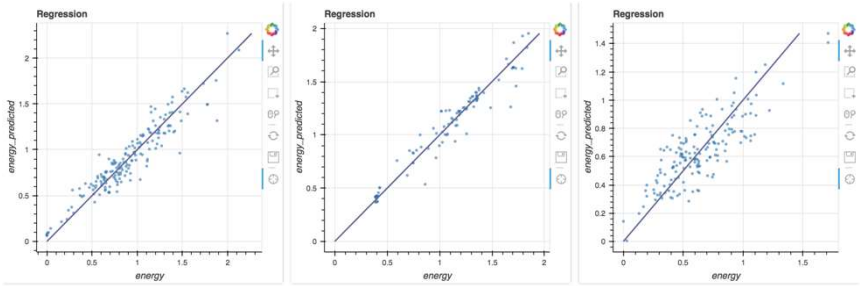
Segmentation
by clustering



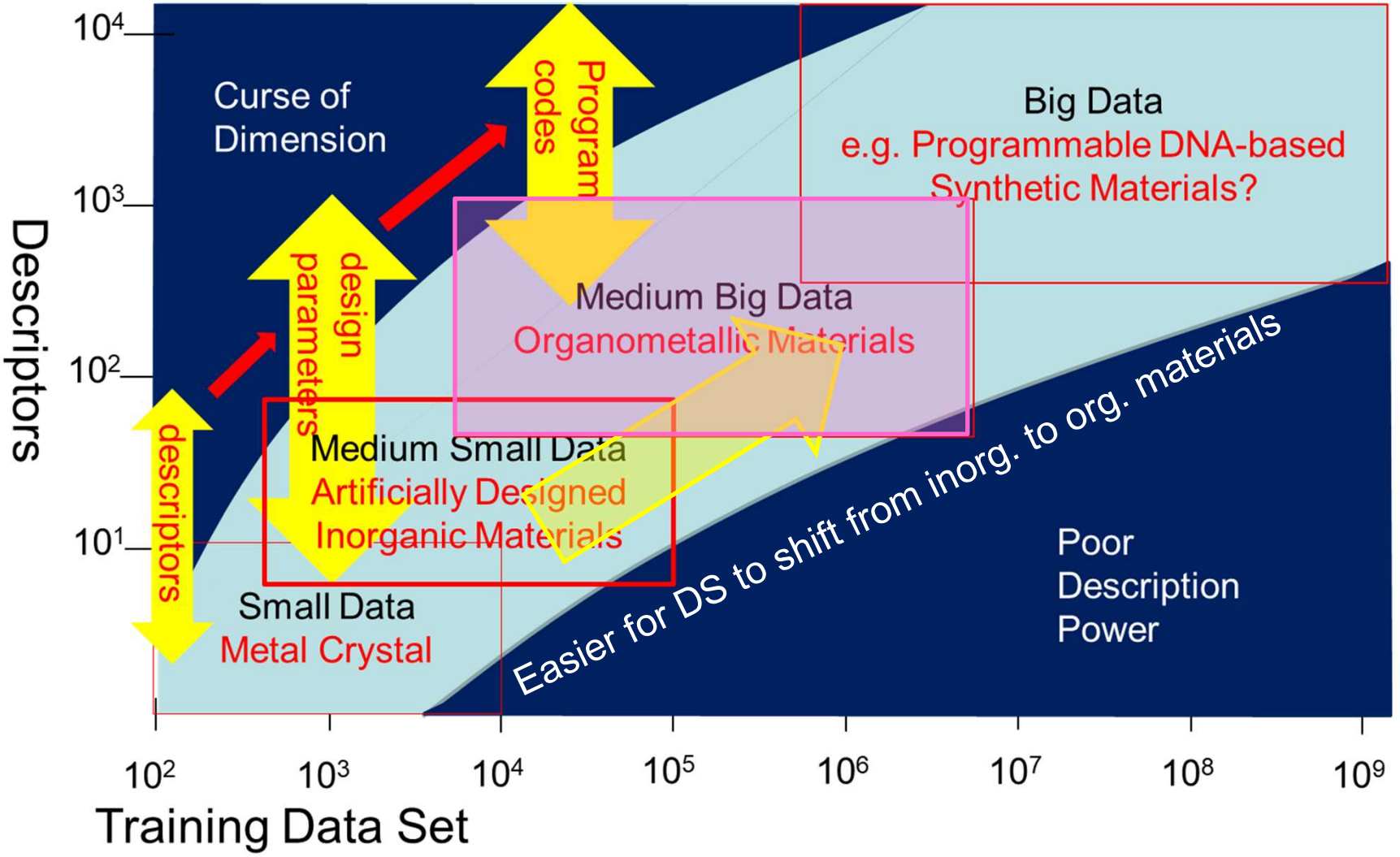
Regression

descriptors → total energy





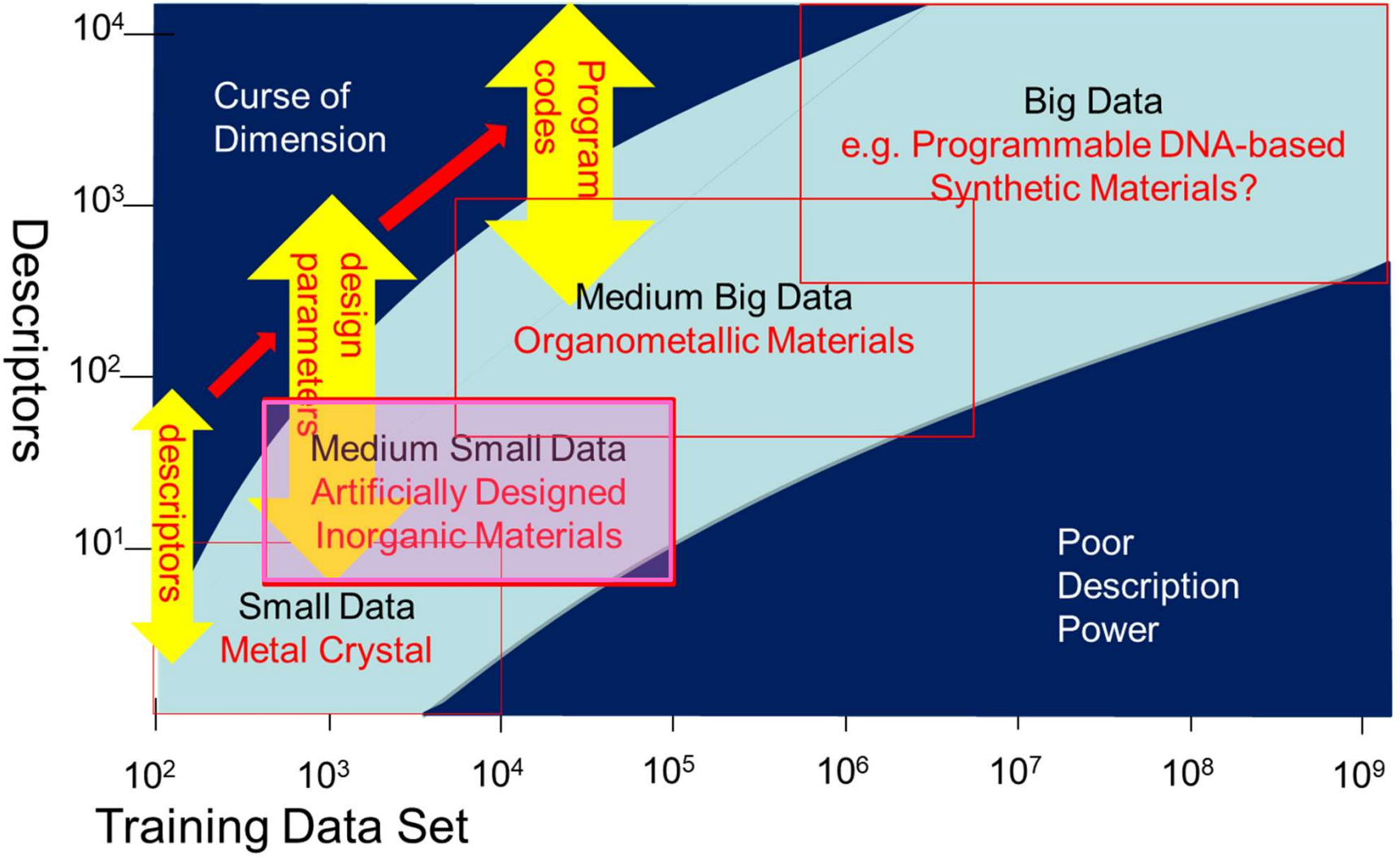
Landscape of Materials Informatics



Organometallic Materials

- Polymer genome as systematically defined descriptors
  ← SMILE representation
- Graph convolution NN
  ← Chemical bond graph

Landscape of Materials Informatics

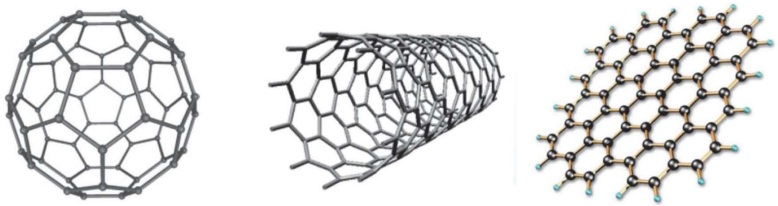


Medium Small Data

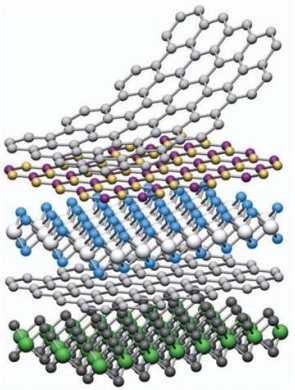
- **Designers' class of materials**
  - Artificially designed materials
- **Design framework**: combinatorial design
  - Multilayered 2D materials
  - **scaffold + modifiers**
    - Scaffold: functional / nonfunctional
    - modifiers to give functions
    - Different scaffoldings define different classes.
- ***Conflict**: Materials scientists acquire only those data appropriate for their target materials with a specific function.*

(Scaffold + Modifiers) Framework: Candidates of Scaffold (1)

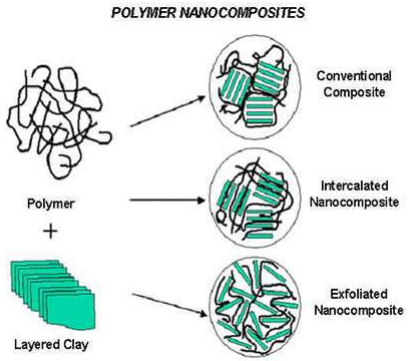
- Carbon-based ones:

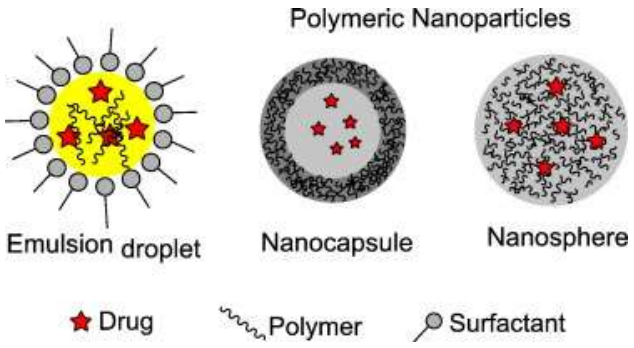


- 2D materials + layered structures
  - Intralayer modifier
  - Interlayer modifier



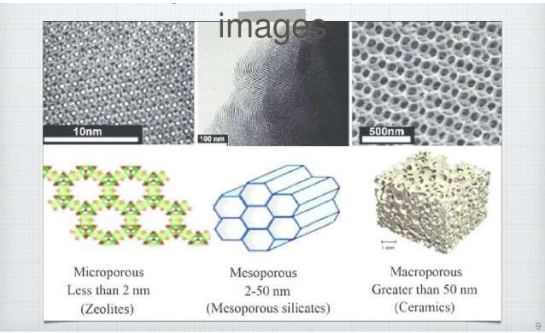
- Polymer nanocomposites/nanoparticle



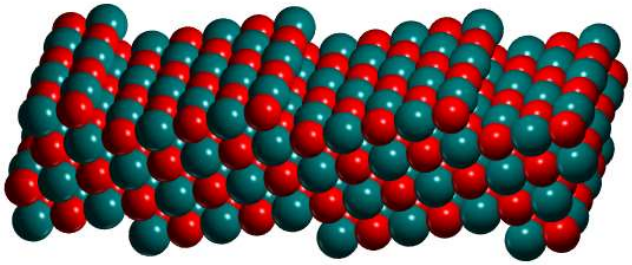


(Scaffold + Modifiers) Framework: Candidates of Scaffold (2)

- Nanopore

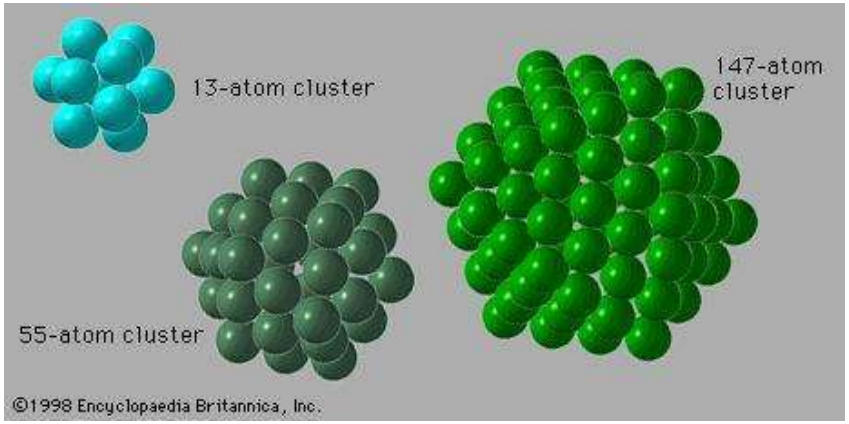


- Crystal surface

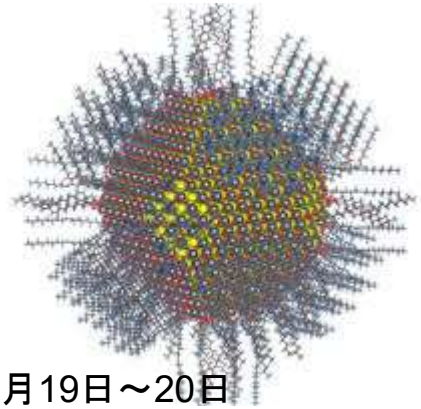


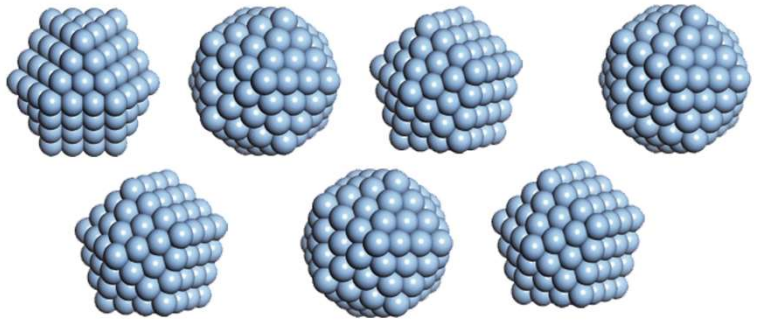
(Scaffold + Modifiers) Framework Modifiers

- Single atom
- Atom cluster

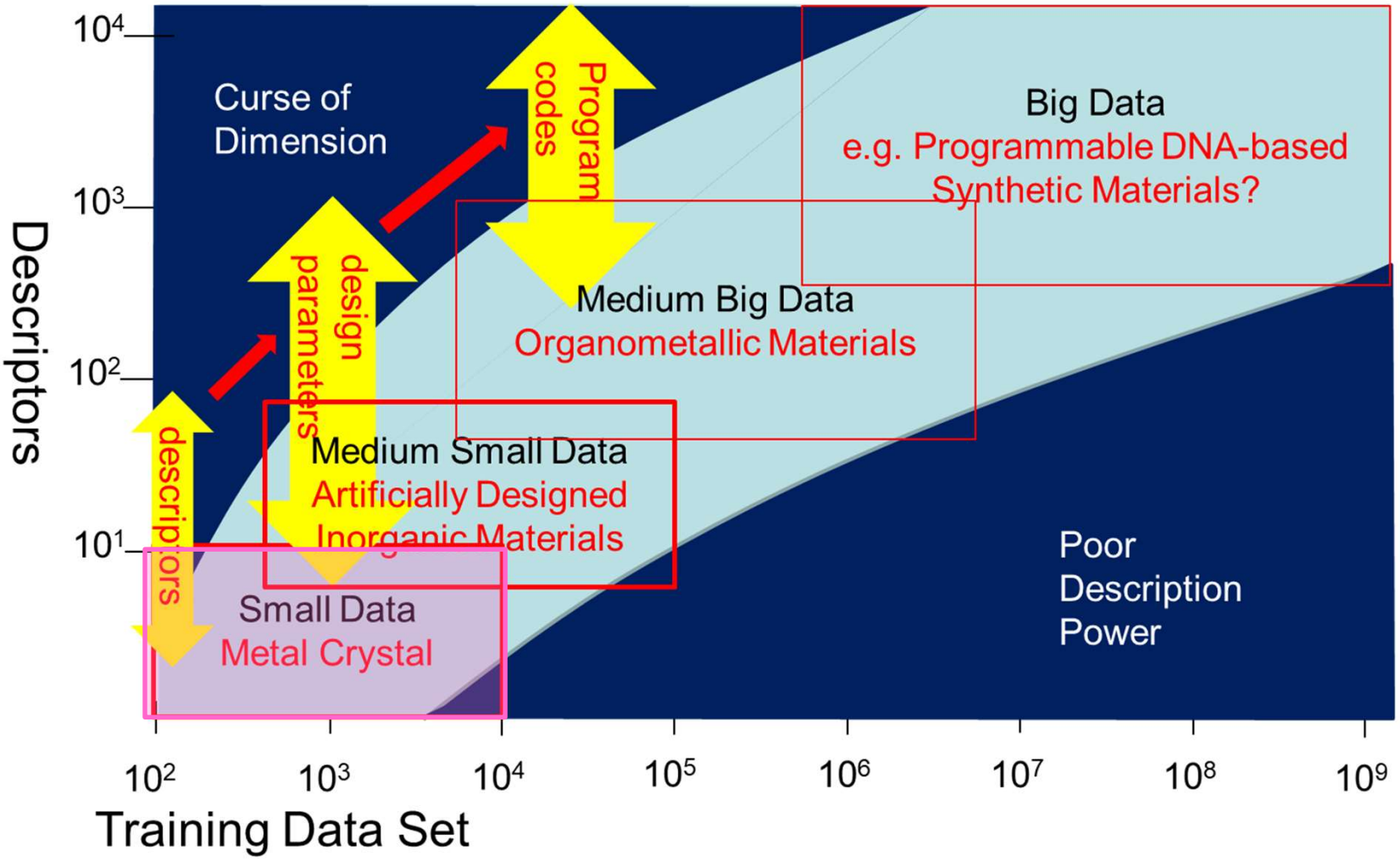


- Nano particle



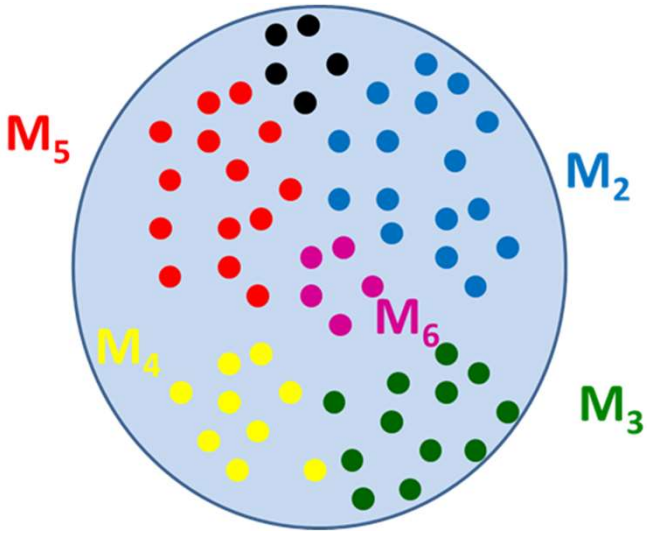


Landscape of Materials Informatics



Small Data: Metal Crystals

- Whole DB:
  - large & heterogeneous
- Homogeneous Subset
  - Small



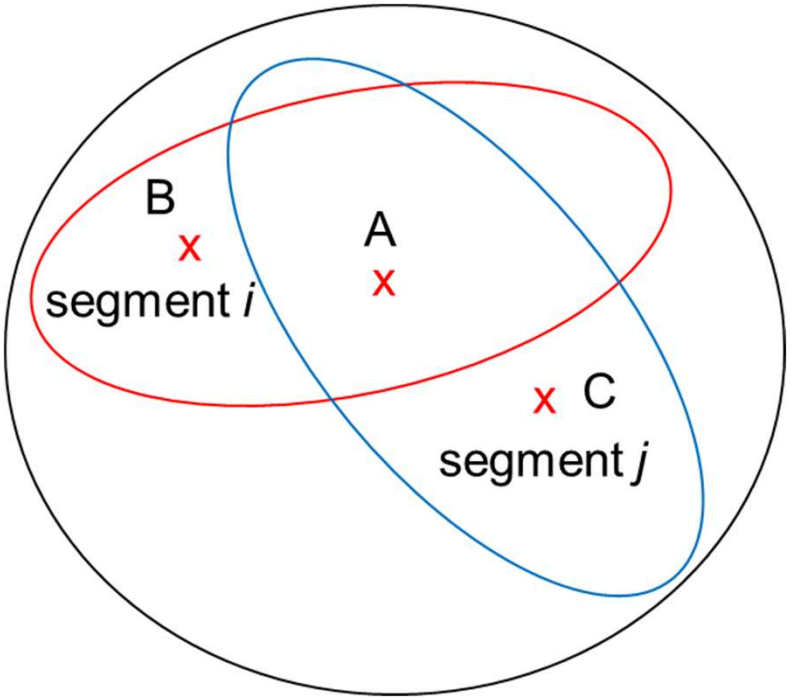
- *How to segment the whole heterogeneous data set into homogeneous sets?*
- *How to define appropriate descriptors?*

Desired Segmentation

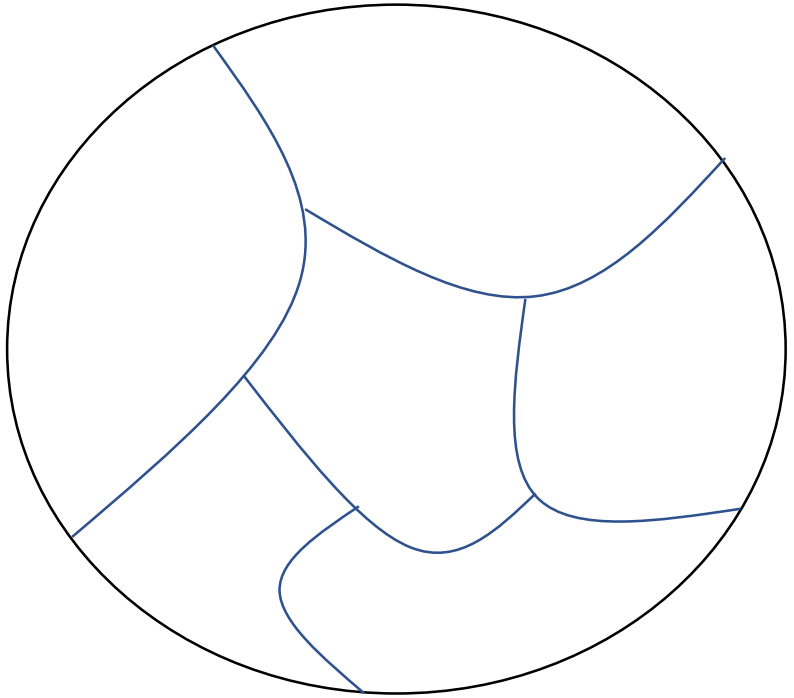
- Materials in each segment should share the same math model, i.e., they share **qualitatively the same mechanism**.
- Our hypothesis:
  - **Quantitative difference** among data in the same segment may be **well-approximated by the regression estimation**.

Segmentation may differ from partition.

Segmentation



Partition



The data point A follows both the model *i* and model *j*, while the data points B and C respectively follow the model *i* and model *j,* but not both.

What kind of mechanisms to focus on for inorganic crystal materials?

- (Strongly) Correlated Electron Systems
  - responsible for a large number of physical phenomena, including ferromagnetism, antiferromagnetism and superconductivity
  - Exchange interaction
    - Direct exchange interaction / Superexchange interaction / RKKY interaction
  - Spin-orbit interaction
- Crystal Field & Orbitals Splitting in Energy

focusing on qualitative properties, but not on quantitative properties

- Space Group / (Oriented) Site Symmetry / Electron Configuration (blocks / subblocks)

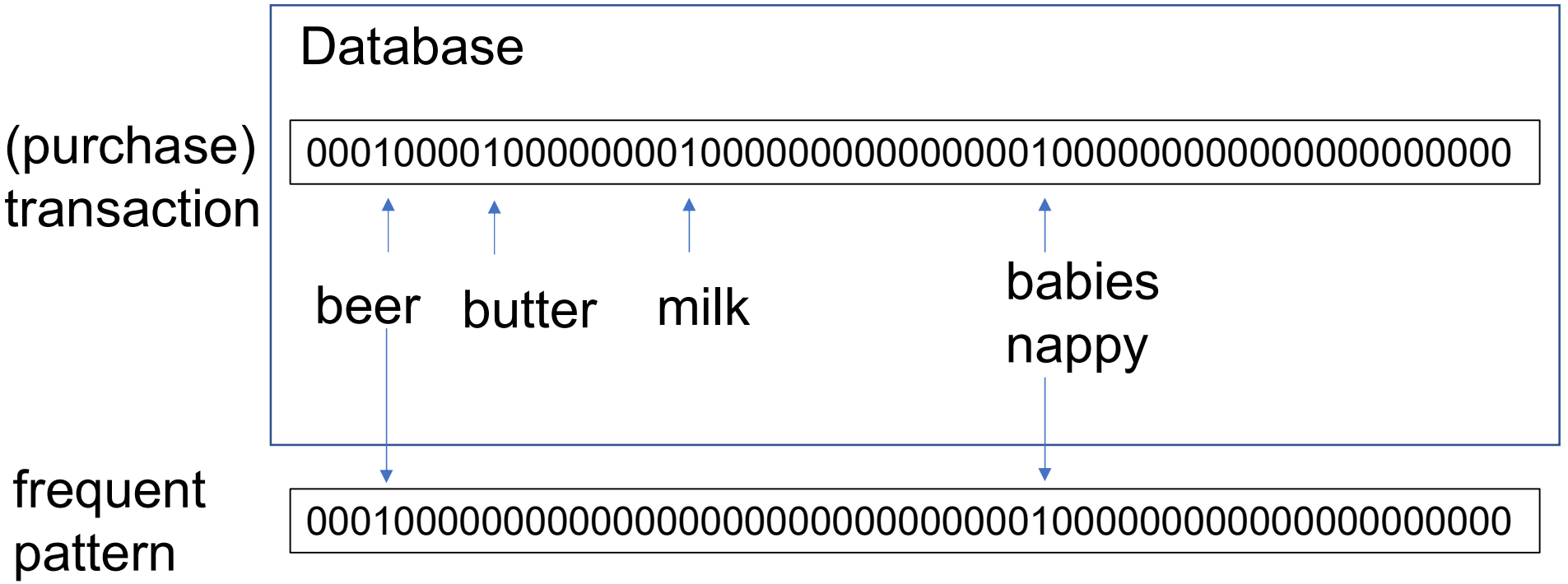
Data Segmentation and Descriptors

- Data segmentation algorithm
  - Which ML algorithm can segment the heterogeneous data set into smaller homogeneous ones?

  - **Item set mining**
    - Each **discovered frequent pattern** identifies the set of all the data with this pattern → **data segmentation**
    - **Focusing on cooccurrence of qualitative properties**
- Descriptors
  - What kinds of descriptors can be used for such a segmentation algorithm?

  - **Crystallographic parameters** (space group, site symmetry, characters, …)
  - **Component elements**: block, sub-block, number of outer shell electrons, …

Item Set Mining (to mine frequent cooccurrence relations)



Support: n　This pattern is included in n transactions at least.

The biggest issue:
It may often result in a huge number of mined patterns?

Data Base

- Materials in AtomWork Adv. with
  - Curie temperature Tc≥200 and
  - site symmetries of elements
- 2138 records
- Data Curation
  - Focusing only on observed Tc: Discard those data with “calculated” in their remark field.
  - Use of modified mean of Tc after grouping those data by substance id

What determines the Properties of Crystalline Inorganic Materials?

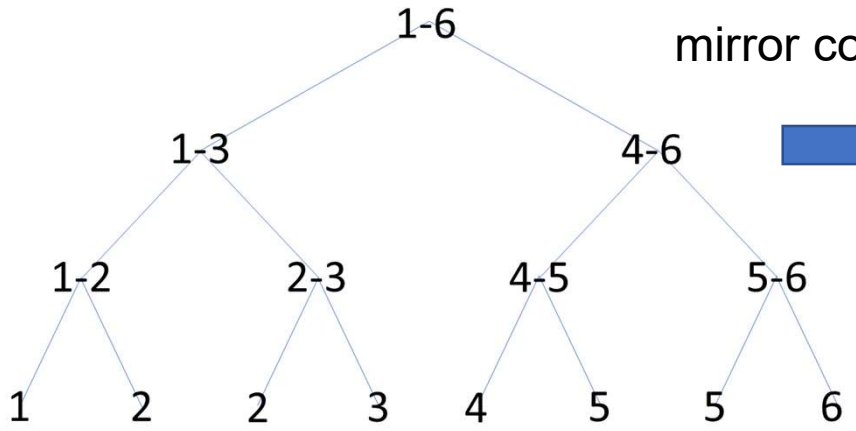
- Structure
  - Crystal Structure: Space Group, Point Group
- Component
  - Element: outermost electron shell, valence, block, sub-block, …
  - Composition ratio, multiplicity, ..
  - Position of each element: (oriented) site symmetry, coordination number, …
  - …
- Dimension (size, scale)

- Our initial focus on red ones

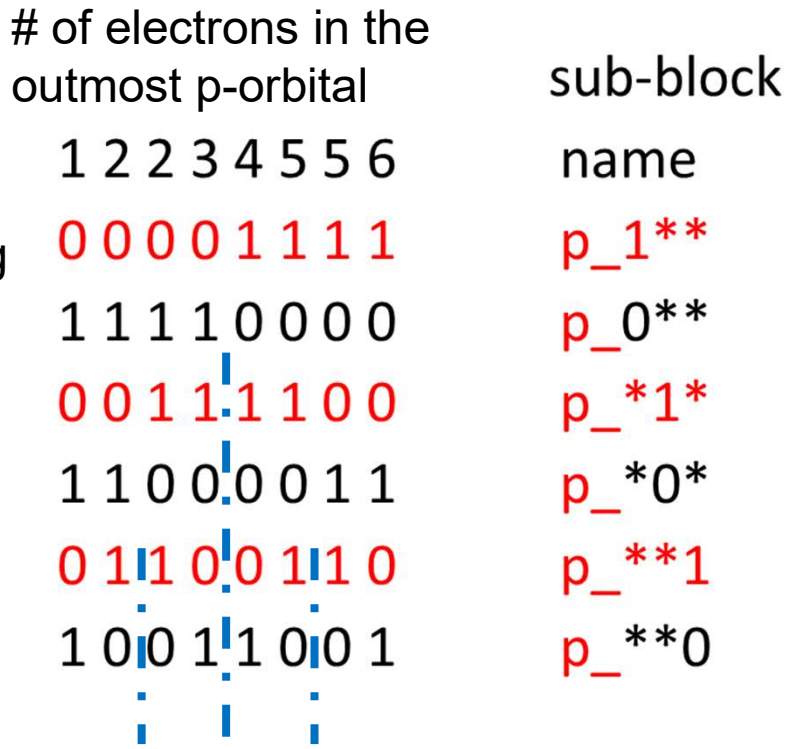
Bitmap Coding of Site Symmetry: subgroup coding of site symmetry

- For each of s-, p-, d-, and f-blocks
  - Hierarchical binary division of each block with respect to the number of electrons in the corresponding orbital.
    - The 1st level division into L and H subblocks:
      - L (Lower) subblock implies that the number of electrons in the corresponding orbital is less than or equal to 1/2 of its maximum electron capacity.
      - H (Higher) subblock implies that the number of electrons in the corresponding orbital is more than 1/2 of its maximum electron capacity.
- Each site symmetry → point group (by neglecting dots)
  - e.g. 4m.m → 4mm
  - Bitmap coding of the equivalent point group as well as all of its sub groups (to consider the crystal distortion effects)

Subblock Coding for p-block

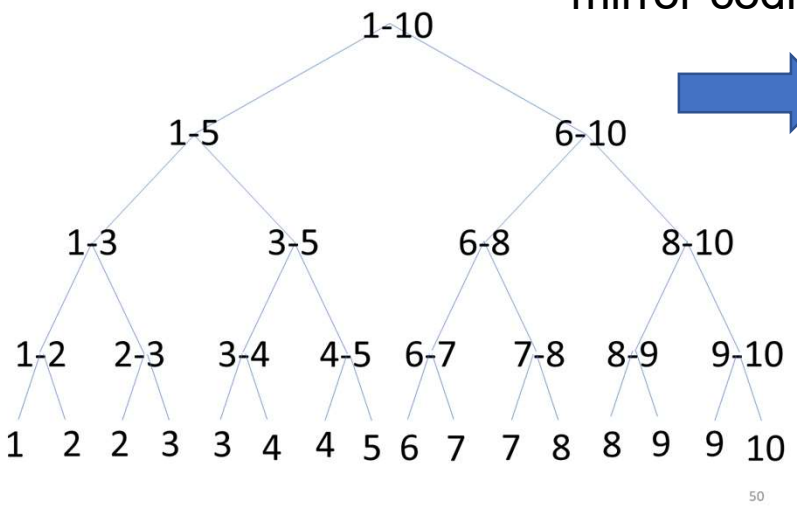


mirror coding

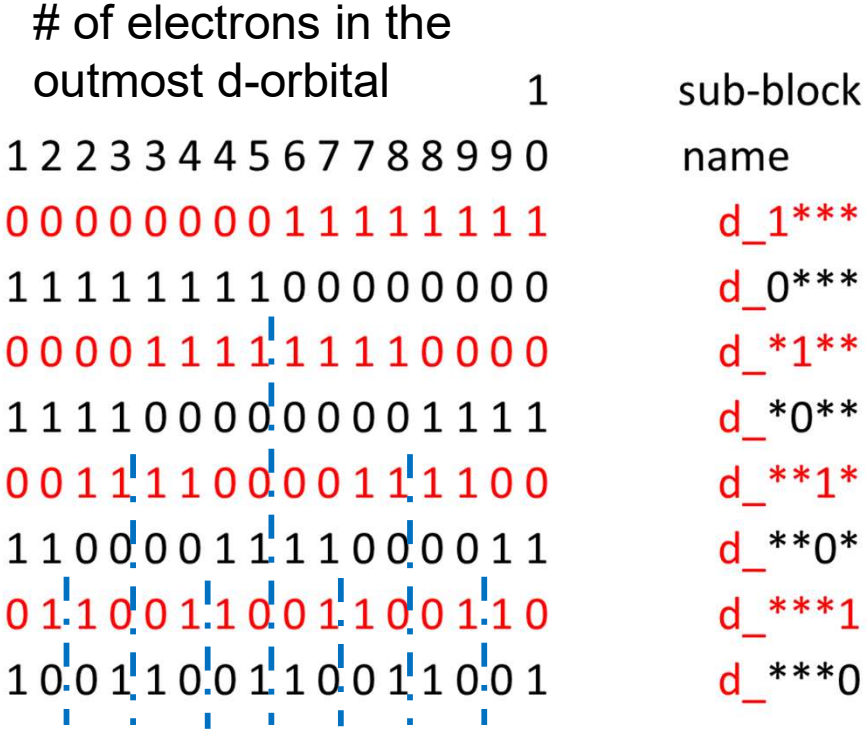


example: p^3 : p_0**, p_*1*, p_**0

Subblock Coding for d-block

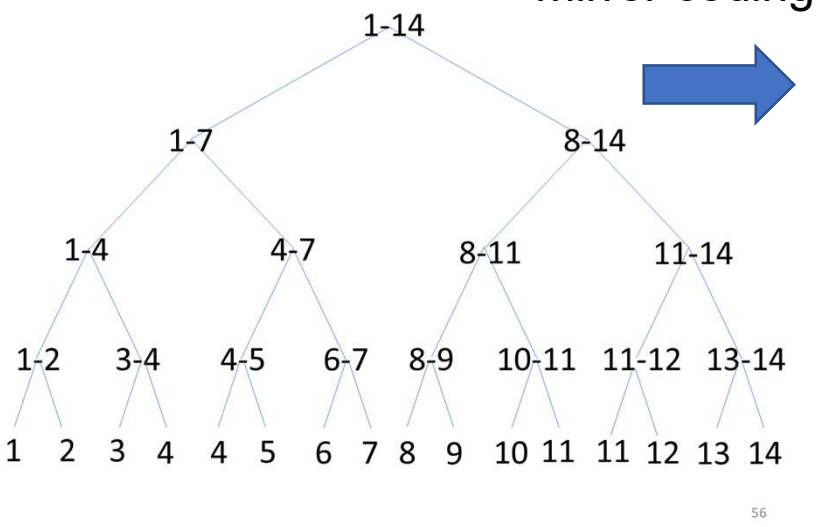


mirror coding



1-5: d_0***, 6-10: d_1***, 3-8: d_*1**, 1-3/8-10: d_*1**, 2-4/7-9: d_**1*

Subblock Coding for f-block



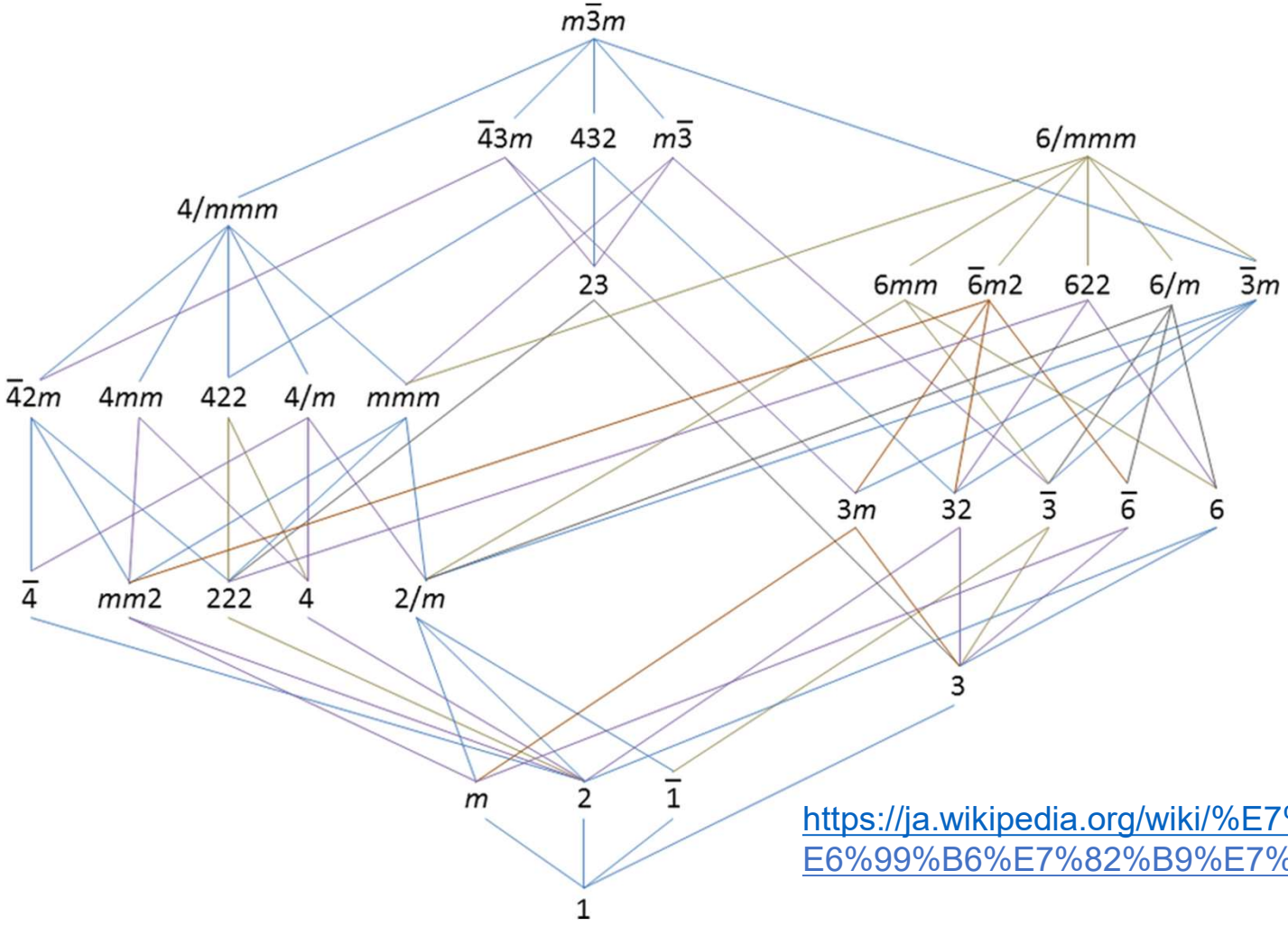
mirror coding

# of electrons in the outmost f-orbital

1 2 3 4 4 5 6 7 8 9 10 11 12 13 14	sub-block name
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1	f_1***
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0	f_0***
0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0	f_*1**
1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1	f_*0**
0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0	f_**1*
1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1	f_**0*
0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0	f_***1
1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1	f_***0

1-7: f_1***, 8-14: f_0***, 4-11: f_*1**, 1-4/11-14: f_*0**, 3-5/10-12: f_**1*

Hierarchy of Point Groups



https://ja.wikipedia.org/wiki/%E7%B5%90%E6%99%B6%E7%82%B9%E7%BE%A4

Example Bitmap Coding of an Inorganic Crystal Material (1)

FePd3

space group: 221
Pm-3m

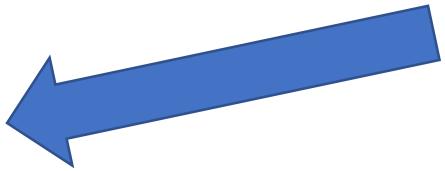
element	multiplicity	site symmetry
Pd	3	4/mm.m
Fe	1	m-3m

Fe: [Ar] $3d^6 4s^2$

Site symmetry: m-3m, 4/mmm, -43m, 432, m-3, -3m, 23, -42m, 4mm, 422, 4/m, mmm, -4, mm2, 222, 4, 2/m, 3m, 32, -3, m, 2, -1, 3, 1

X

Subblock for d^6
d_1***, d_*1**, d_**0*, d_***0, d_****



bitmap code for Fe:
sym(m-3m, d_1***), sym(m-3m, d_*1**), …, sym(m-3m, d_****),
sym(4/mmm, d_1***), sym(4/mmm, d_*1**), …, sym(4/mmm, d_****),
sym(-43m, d_1***), sym(-43m, d_*1**), …
…
sym(1, d_1***), sym(1, d_*1**), …, sym(1, d_***0), sym(1, d_****)

Example Bitmap Coding of an Inorganic Crystal Material (2)

FePd3

Bitmap Code

Codes for more than one element in the same block are superimposed.

Point group & its subgroup: m-3m, 4/mmm, -43m, 432, m-3, -3m, 23, -42m, 4mm, 422, 4/m, mmm, -4, mm2, 222, 4, 2/m, 3m, 32, -3, m, 2, -1, 3, 1

Bitmap code for Fe:

sym(m-3m, d_1***), sym(m-3m, d_*1**), …, sym(m-3m, d_****),
sym(4/mmm, d_1***), sym(4/mmm, d_*1**), …, sym(4/mmm, d_****),
sym(-43m, d_1***), sym(-43m, d_*1**), …
…
sym(1, d_1***), sym(1, d_*1**), …, sym(1, d_***0), sym(1, d_****)

Prioritized fundamental mined patterns for materials with and without f-block elements

- Minimum support: 200
  - To obtain patterns appearing in no less than 200 materials.
- **385,933 patterns** for materials with f-block elements
- **1,237,261 patterns** for materials without f-block elements
  - **Too many** patterns! The **well-known drawback of the item-set mining**.
- **Some patterns may reflect the materials structures**, or equivalently hidden mechanisms**, to determine Tc**, while **others reflect some structures that have nothing to do with ferromagnetism**.
- Mined patterns can be **prioritized** in the descending order of **their regression R2 scores**,
  - where, for each mined pattern and its associated data segment, i.e., the set of all the data satisfying this pattern, we cross validate the RF-regression of Tc to obtain the regression score.

Prioritized mined patterns for **1385** materials with Tc (≥ 200) and f-block elements

	pattern	support	r2	correlatior	C(k)	C_p(k)
0	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,d_***0)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(2/m)']	227	0.890894	0.945658	227	0.163899
1	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'pg(-1)']	222	0.884817	0.94272	319	0.230325
2	['sym(m,p_**0)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(mmm)']	216	0.881959	0.939885	410	0.296029
3	['sym(m,p_**0)', 'sym(-1,d_*1**)', 'sym(m,f_0***)', 'sym(m,f_***0)', 'pg(mmm)']	227	0.881062	0.939288	444	0.320578
4	['sym(-1,d_0***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(2,f_***0)', 'pg(-1)']	205	0.880092	0.941081	485	0.350181
5	['sym(1,d_0***)', 'sym(-1,d_*1**)', 'sym(-1,d_**0*)', 'sym(2,f_0***)', 'sym(2,f_*0**)', 'sym(2,f_***0)', 'pg(2/m)']	201	0.879723	0.941847	486	0.350903
6	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(2,f_0***)', 'sym(2,f_*0**)', 'sym(2,f_***0)', 'pg(-1)']	206	0.87951	0.938911	486	0.350903
7	['sym(-1,d_0***)', 'sym(-1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_***0)', 'pg(2/m)']	267	0.879421	0.939599	496	0.358123
8	['sym(1,d_0***)', 'sym(-1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(2/m)']	214	0.879216	0.942319	496	0.358123
9	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'sym(1,f_***0)', 'pg(2/m)']	203	0.879069	0.939018	496	0.358123
10	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_***0)', 'pg(mm2)']	298	0.878996	0.937761	541	0.390614
11	['sym(m,p_***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(mmm)']	214	0.877957	0.940487	551	0.397834
12	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'pg(1)']	243	0.877807	0.93836	566	0.408664
13	['sym(-1,d_0***)', 'sym(-1,d_**0*)', 'sym(1,f_**0*)', 'sym(1,f_***0)', 'pg(-1)']	207	0.877555	0.940219	566	0.408664
14	['sym(1,d_0***)', 'sym(-1,d_**0*)', 'sym(m,f_***0)', 'pg(2/m)']	308	0.877252	0.937585	575	0.415162
15	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'pg(2/m)']	221	0.875985	0.937279	575	0.415162
6623	['sym(mm2,d_0***)', 'sym(m,d_*1**)', 'sym(2,d_*1**)', 'sym(mm2,d_**0*)', 'sym(m,f_****)', 'pg(mm2)']	405	0.807531	0.898895	1380	0.99639
6624	['sym(1,d_****)', 'sym(m,f_****)', 'pg(222)', 'pg(mm2)']	966	0.807529	0.899073	1381	0.997112
6625	['sym(1,d_1***)', 'sym(3,d_**0*)', 'sym(3,f_*1**)', 'sym(3,f_**0*)', 'pg(-3)', 'pg(2/m)']	216	0.807527	0.900008	1381	0.997112
7696	['sym(m,d_*1**)', 'sym(2,d_*1**)', 'sym(mm2,f_0***)', 'sym(mm2,f_***0)', 'pg(4/mmm)']	269	0.804522	0.899983	1381	0.997112
7697	['sym(1,f_*1**)', 'pg(2/m)']	862	0.804521	0.897935	1385	1

Each highlighted cell indicates that this pattern covers at least one new material. We call such a pattern a **fundamental pattern**.

Prioritized fundamental mined patterns for **1385** materials with Tc (≥ 200) and f-block elements

	pattern	support	r2	correlation	C(k)	C_p(k)	
0	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,d_***0)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(2/m)']	227	0.890894	0.945658	227	0.163899	
1	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'pg(-1)']	222	0.884817	0.94272	319	0.230325	
2	['sym(m,p_**0)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(mmm)']	216	0.881959	0.939885	410	0.296029	
3	['sym(m,p_**0)', 'sym(-1,d_*1**)', 'sym(m,f_0***)', 'sym(m,f_***0)', 'pg(mmm)']	227	0.881062	0.939288	414	0.320578	
4	['sym(-1,d_0***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(2,f_***0)', 'pg(-1)']	305	0.880092	0.941081	485	0.350181	
5	['sym(1,d_0***)', 'sym(-1,d_*1**)', 'sym(-1,d_**0*)', 'sym(2,f_0***)', 'sym(2,f_*0**)', 'sym(2,f_**0*)', 'pg(2/m)']	201	0.879723	0.941847	486	0.350903	
7	['sym(-1,d_0***)', 'sym(-1,d_**0*)', 'sym(1,f_***0)', 'pg(2/m)']	267	0.879421	0.939599	496	0.358123	
10	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_***0)', 'pg(mm2)']	298	0.878996	0.937761	541	0.390614	
11	['sym(m,p_***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(mmm)']	214	0.877957	0.940487	551	0.397834	
12	['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,f_*0**)', 'sym(1,f_**0*)', 'pg(1)']	243	0.877807	0.938836	566	0.408664	
14	['sym(1,d_0***)', 'sym(-1,d_**0*)', 'sym(m,f_***0)', 'pg(2/m)']	308	0.877252	0.937585	575	0.415162	
23	['sym(mm2,p_***)', 'sym(m,d_****)', 'sym(2,f_***0)', 'sym(m,f_**0)', 'pg(4/mmm)']	213	0.873752	0.939633	587	0.423827	
24	['sym(-1,d_0***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'pg(mmm)']	238	0.873621	0.938011	642	0.463538	
26	['sym(1,p_***)', 'sym(-1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(mmm)']	211	0.872795	0.937256	643	0.46426	
30	['sym(2,p_0**)', 'sym(mm2,d_**0*)', 'sym(m,f_*1**)', 'sym(2,f_*1**)', 'pg(4/mmm)']	213	0.871613	0.93582	707	0.510469	
31	['sym(1,d_0***)', 'sym(1,d_**0*)', 'sym(1,f_0***)', 'sym(1,f_***0)', 'pg(-1)']	398	0.871424	0.935083	714	0.515523	
33	['sym(-1,d_0***)', 'sym(-1,d_**0*)', 'sym(1,f_*1**)', 'pg(2/m)']	221	0.871389	0.934697	756	0.545848	
35	['sym(mm2,p_**0)', 'sym(2/m,d_*1**)', 'sym(2/m,d_**0*)', 'sym(mm2,f_****)', 'pg(4/mmm)']	213	0.871174	0.934408	789	0.569675	
36	['sym(-1,d_0***)', 'sym(1,d_*1**)', 'sym(-1,d_**0*)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'pg(2/m)']	211	0.871144	0.934449	791	0.571119	
4937	['sym(-1,d_*1**)', 'sym(3,d_*1**)', 'sym(-1,d_**0*)', 'sym(3,d_**0*)', 'sym(32,f_****)', 'pg(-3m)']	279	0.813334	0.903438	1379	0.995668	
6216	['sym(1,p_**0)', 'sym(1,f_0***)', 'pg(222)', 'pg(mm2)']	365	0.808871	0.900013	1380	0.99639	
6624	['sym(1,d_****)', 'sym(m,f_****)', 'pg(222)', 'pg(mm2)']	966	0.807529	0.899073	1381	0.997112	
7697	['sym(1,f_*1**)', 'pg(2/m)']	862	0.804521	0.897935	1385	1	

No need to consider more than 7,698 patterns out of 385,933 mined patterns.

Actually, we need to consider **only 113 fundamental patterns** out of 7,698 patterns.

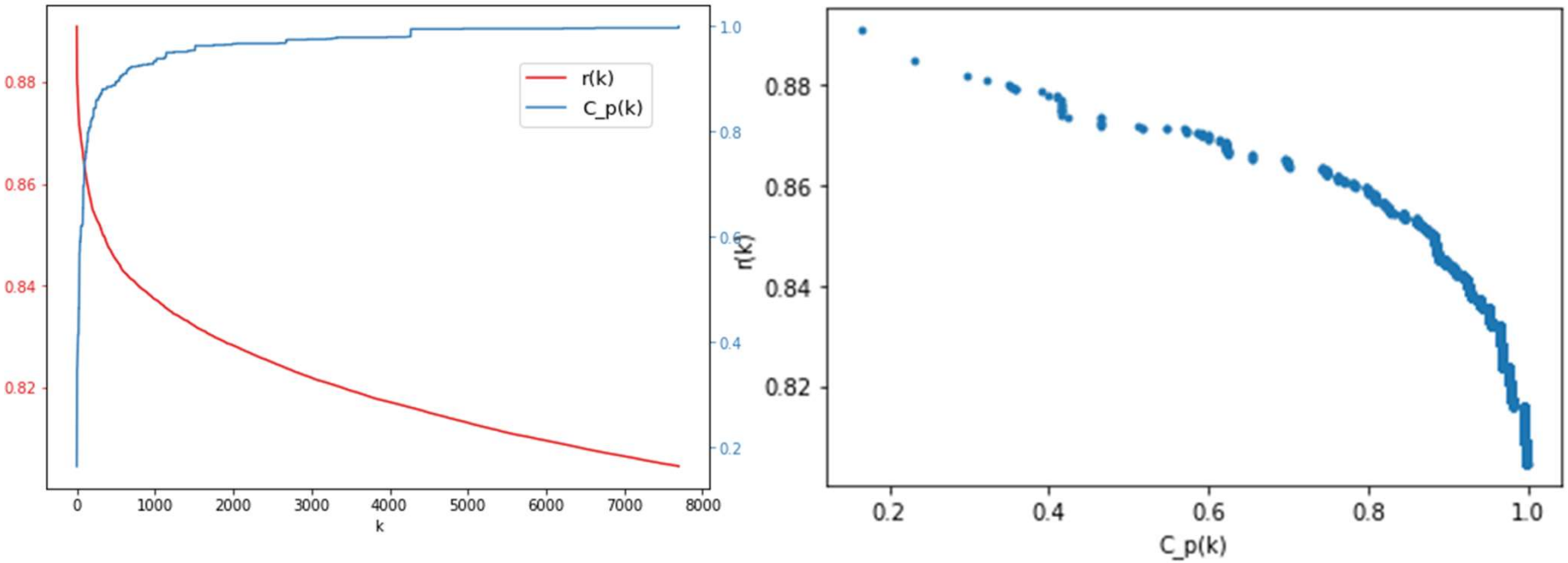
Even the last fundamental pattern gives better R2 score than the score of the regression for the whole data set of materials with an f-block element

Coverage v.s. the minimum R2 score of the regression

- C(k): the k-th coverage, i.e., the number of those materials satisfying one of the patterns listed in the prioritized mined patterns no later than the k-th one.
- C_p(k): the k-th coverage rate, i.e., the rate of C(k) to the total number of the target materials.
- r(k): the k-th regression score, i.e., the R2 score of the regression for the data segment associated with the k-th pattern in the list of prioritized mined patterns.

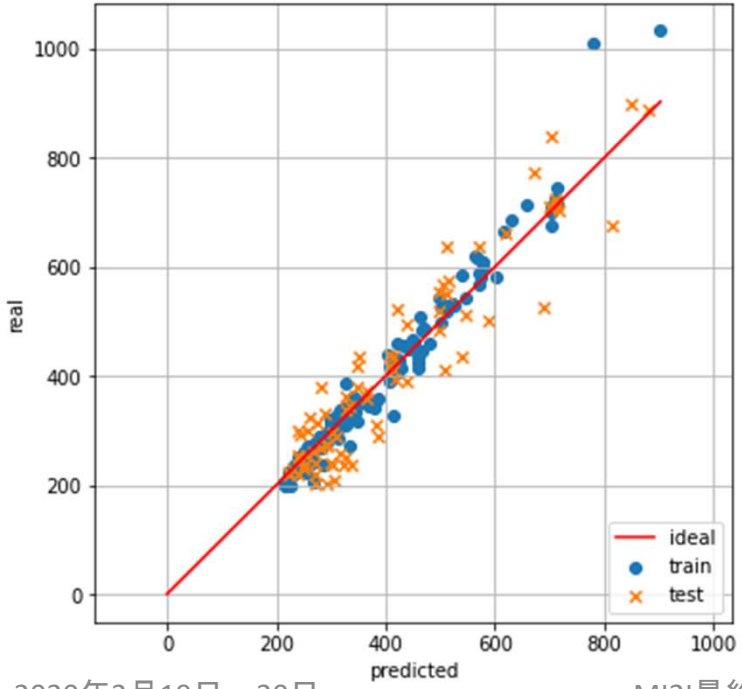
The k-th coverage v.s. the k-th regression score

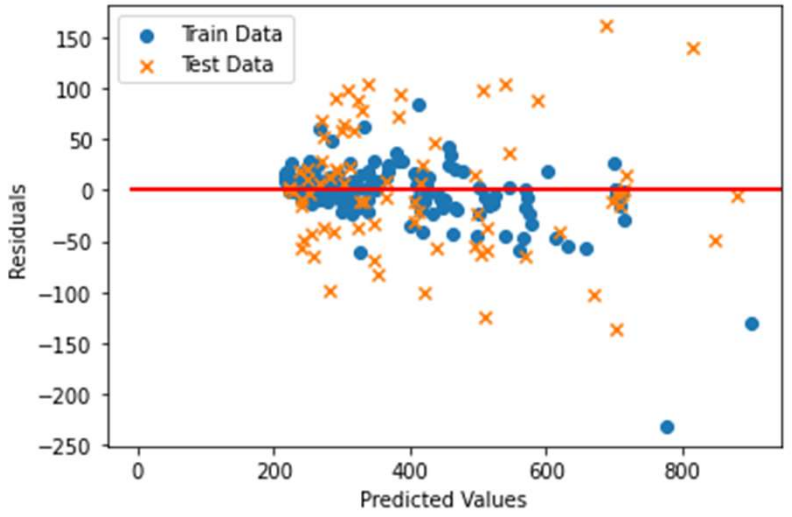
- Ferromagnetic materials with an f-block element



Regression for the data segment of the first fundamental pattern
(ferromagnetic materials with an f-block element)

pattern 0 (the first fundamental pattern)
['sym(1,d_0***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,d_***0)', 'sym(m,f_0***)',
'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(2/m)'],
support: 227,
R2: 0.89090 (c.f. R2: 0.79003 for the whole)

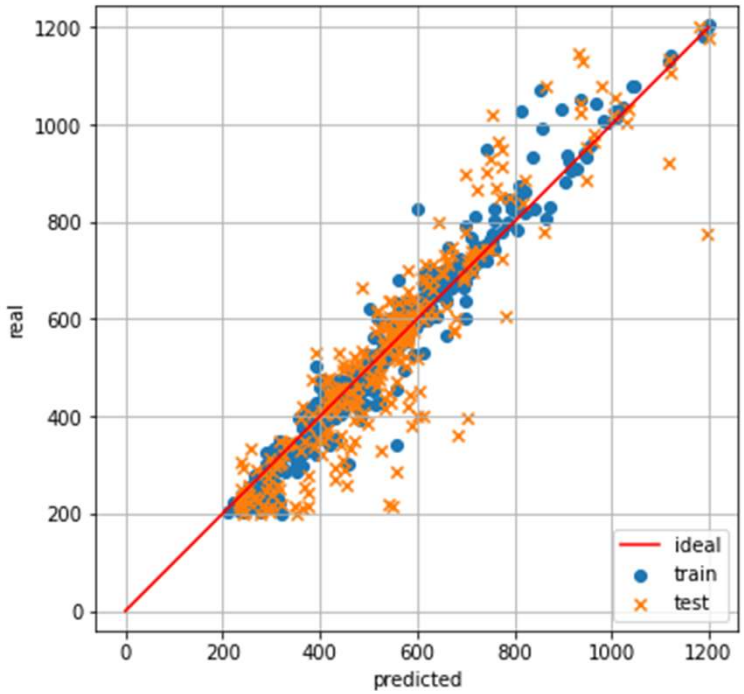


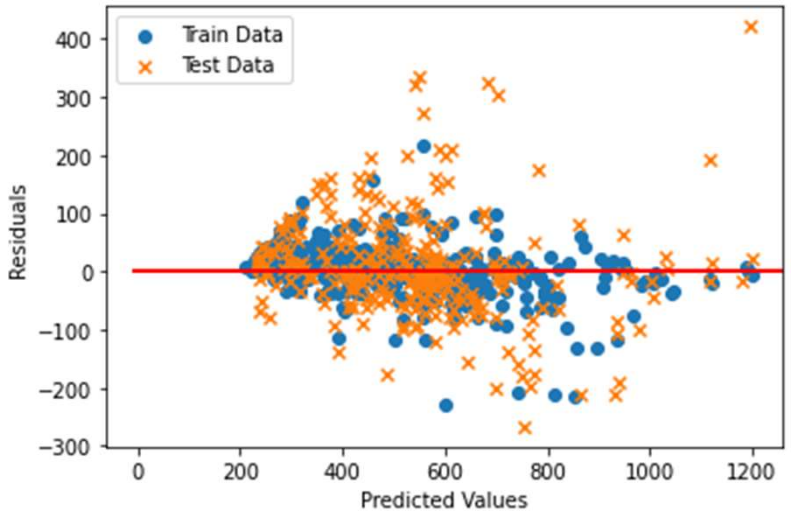


Regression for the data segment of the last fundamental pattern
(ferromagnetic materials with an f-block element)

pattern 7697 (the last fundamental pattern)
['sym(1,f_*1**)', 'pg(2/m)'],
support： 862,
R2: 0.80452　　(c.f. R2: 0.79003 for the whole)

This pattern newly covers only 4 materials, which means that only 4/1385 of materials with f-block elements may have their Tc estimated with this lowest R2.





Estimating Tc of a new material with f-block elements

(when structural information is known.)

- Obtain the point group of its crystal and the site symmetry of each constituent element.
- Construct its bitmap code representation.
- In the ordered list of prioritized 113 fundamental patterns, **find the first pattern** which this code representation satisfies.
- **Apply the RF regression function** obtained for this fundamental pattern to estimate Tc of this material.

Estimating Tc of a new material with f-block elements

(when structural information is not known.)

- (Method 1) When the **structure type of the goal material is known**.
  - Obtain the mined pattern with highest R2 score that is satisfied by all the target materials of this type registered in AtomWork Adv.
    - If such a pattern does not exist, this method is not applicable.
  - Use this pattern to estimate the value.
- (Method 2) **Otherwise**
  - Obtain all the target materials from AtoWork Adv. that include all the constituent elements of the goal material.
  - Obtain the mined pattern with highest R2 score that is satisfied by all the above-mentioned target materials.
    - If such a pattern does not exist, this method is not applicable.
  - Use this pattern to estimate the value.

Estimating Tc of unknown materials of structure type ThM12 (Method 1)

- **Comparison with the calculated Tc of the top 10 materials** with highest magnetization whose magnetization, Tc, and formation energies are obtained through the first-principles calculation by T. Fukazawa, Y. Harashima, Z. Hou, and T. Miyake (PHYSICAL REVIEW MATERIALS **3**, 053807 (2019))
- **We do not know their structural information!**
- **Their structure types are all ThM12.**
- The mined pattern that is satisfied by all the ferromagnetic materials of structure type ThM12 that are registered in AtomWork Adv. is the **pattern 5263**, which is not a fundamental pattern, but has higher priority than the last fundamental pattern.
- We use this pattern for data segmentation.

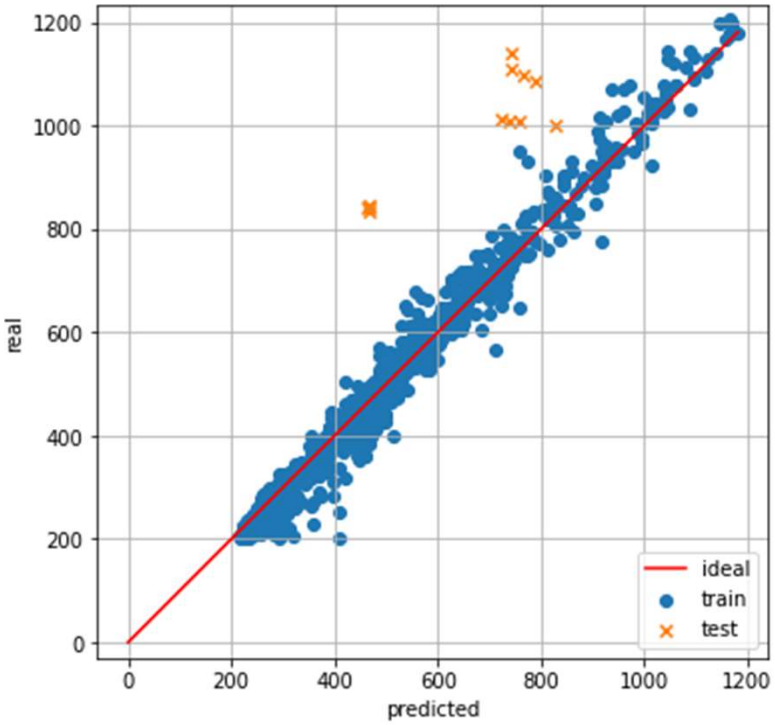
Estimating Tc of each unknown material of structure type ThM12 using the data set of the pattern 5263 except itself

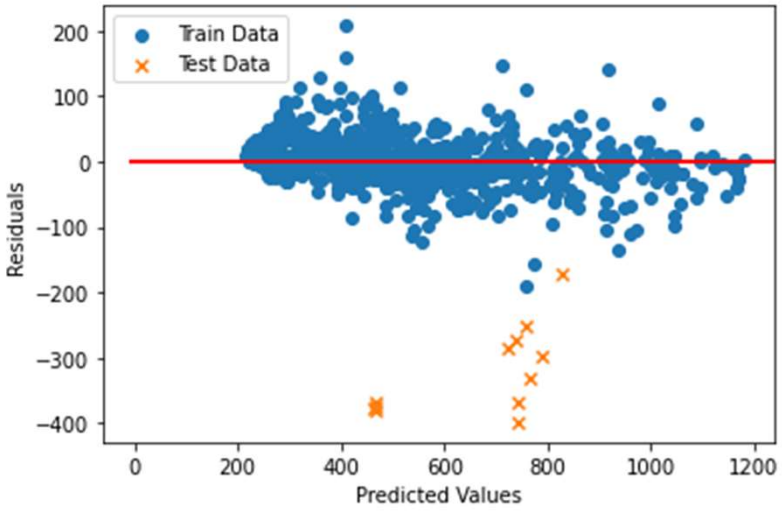
- Comparison with the calculated Tc of the top 10 materials with highest magnetization whose magnetization, Tc, and formation energies are obtained through the first-principles calculation by T. Fukazawa, Y. Harashima, Z. Hou, and T. Miyake (PHYSICAL REVIEW MATERIALS **3**, 053807 (2019))

id	Formula	Obs	Pred	ratio
0	NdFe12	844	465.216587	0.448796
1	Nd(Fe0.9Co0.1)12	1012	725.258733	0.283341
2	Nd(Fe0.8Co0.2)12	1111	743.903133	0.33042
3	(Nd0.9Zr0.1)Fe12	835	468.003301	0.439517
4	(Nd0.9Zr0.1)(Fe0.9Co0.1)12	1011	737.881733	0.270147
5	(Nd0.9Zr0.1)(Fe0.8Co0.2)12	1098	765.810333	0.302541
6	(Nd0.8Zr0.2)Fe12	841	468.128301	0.443367
7	(Nd0.8Zr0.2)(Fe0.9Co0.1)12	1009	756.807533	0.249943
8	(Nd0.7Zr0.3)Fe12	845	465.430445	0.449195
9	(Nd0.8Zr0.2)(Fe0.8Co0.2)12	1086	789.410933	0.273102
10	Nd(Fe0.7Co0.3)12	1143	744.357333	0.348769
11	(Sm0.7Zr0.3)(Fe0.9Co0.1)12	1002	829.0014	0.172653

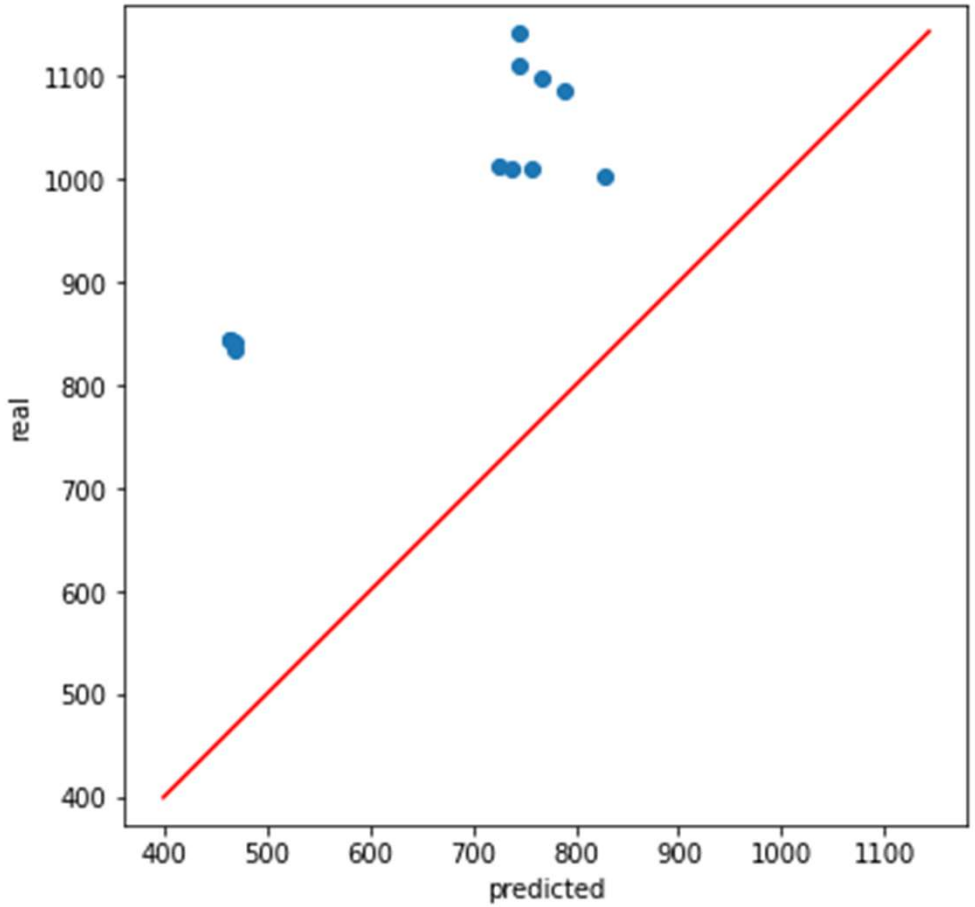
Estimation by regression using pattern 5263

- These materials have the structure type of ThM12.
- It is well known that **the calculated Tc's are higher than measured ones**, and that they are **correlated well**.





Estimating Tc of unknown materials of structure type ThM12 (3)



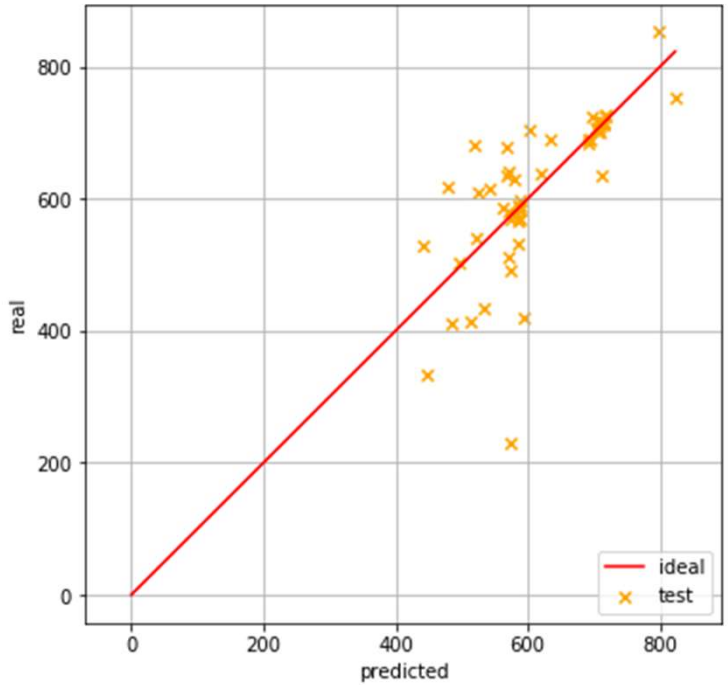
Pearson correlation coefficient between calculated Tc's and estimated ones (points in red): 0.89486

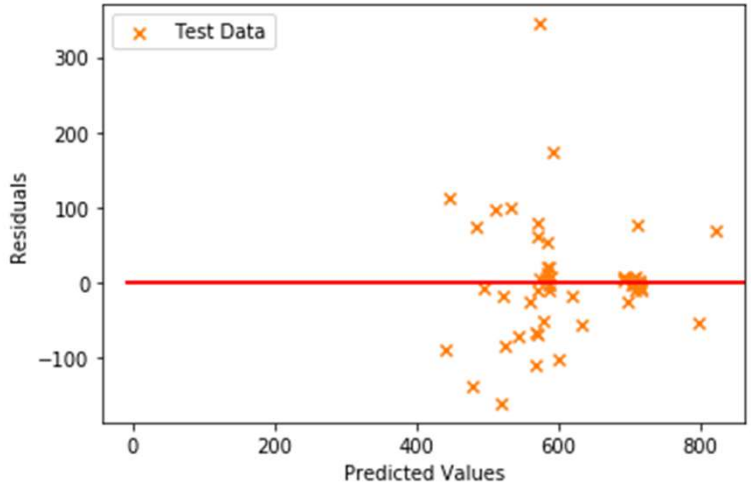
Good correlation!

Estimating Tc of all the **57 known materials** with **Nd, Fe, and B** (Method 2)

- Materials with Nd, Fe, and B with Tc >200: 57
- **Patterns covering all of them** → 4 patterns
  - **189**,"['sym(1,p_***)', 'sym(1,d_*1**)', 'sym(1,d_**0*)', 'sym(1,d_***0)', 'sym(m,f_0***)', 'sym(m,f_*0**)', 'sym(m,f_***0)', 'pg(2/m)']", 247
  - 2582,"['sym(m,d_****)', 'sym(m,f_*1**)', 'sym(2,f_*1**)', 'pg(m)', 'pg(2)']", 644,
  - 3966,"['sym(2,f_*1**)', 'pg(m)', 'pg(2)']", 685
  - 7697,"['sym(1,f_*1**)', 'pg(2/m)']", 862

Estimating Tc of each of **57 known materials** by regression using all the data of the **pattern 189** except itself. (Method 2)





Pearson correlation coefficient: 0.75062
Root Mean Squared Error: 74.867
Mean Absolute Error: 44.880
Average Error Rate: 0.097227

Finding the material of type $\mathbf{A_xNd_{(2-x)}Fe_{14}B}$ with the **highest Tc** (1)

- Using the **regression model** obtained for **pattern 189** with **247 materials** for training.
- For each

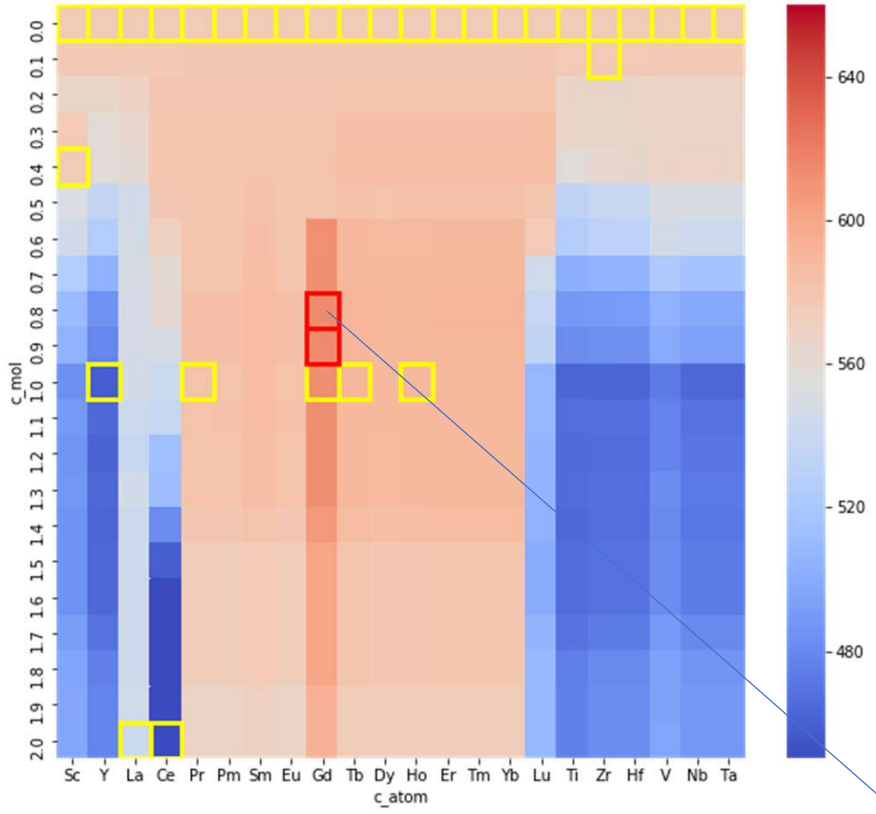
  $x = 0, 0.1, 0.2, \ldots, 1.0$,

  and for each

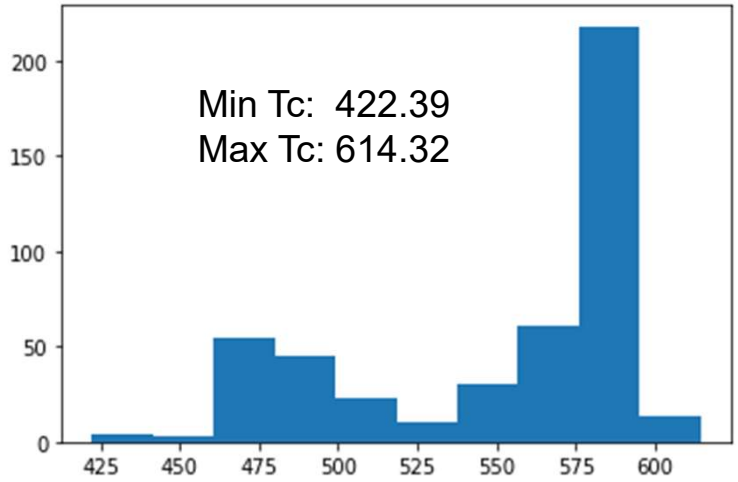
  $A \in \{$ Sc, Y, La, Ce, Pr, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Ti, Zr, Hf, V, Nb, Ta $\}$,

  estimate Tc.

Finding the material of type $\mathbf{A_xNd_{(2-x)}Fe_{14}B}$ with the **highest Tc** (2)



min: 450, max: 660



max: T_C: 614.32

$Gd_{0.8}Nd_{1.2}Fe_{14}B$, $Gd_{0.9}Nd_{1.1}Fe_{14}B$

Problems with the data set

- Example:
  - Ni3Al
  - Space Group: Pm-3m,
  - Curie temperature (average)=50.73
- Data dispersion among different research papers
  - Phys. Status Solidi A,1992,133,,231-235: Tc=43
  - J. Phys. Soc. Jpn.,1992,61,,3313-3321: Tc=87.5
  - Phys. Rev. B: Condens. Matter,1989,40,,11488-11492: Tc=68
  - and 25 more.
- Different Space-Group estimation in some research papers
  - space group=Pmmn in
  - J. Less-Common Met.,1991,170,,171-184
  - Phys. Status Solidi A,1992,133,,231-235
  - Phys. Rev. B: Condens. Matter,1996,53,,1129-1137

The Tc value 610 of $Sc_{0.4}Nd_{1.6}Fe_{14}B$ (a wrong value!)

- This was confused with the Tc of $Nd_2(Fe_{0.9}, Sc_{0.1})_{14}B$ reported in Table 1 of the following paper.
- H. C. Ku and L. S. Yen: Magnetic properties of the new permanent magnet compounds $Nd_2(Fe_{0.9}, M_{0.1})_{14}B$ (M = SC, Ti, V, Cr, Mn, Co, Ni), *Journal of the Less-Common Metals, 127* (19874) 3 -48.

TABLE 1
Crystallographic and magnetic properties of $Nd_2(Fe_{0.9}M_{0.1})_{14}B$ compounds

Compound	*Lattice parameters*			*Density*	*Curie temperature*	*Saturation moment*	*Anisotrop field*
	a(Å)	*c*(Å)	*V*(Å^3)	ρ(g cm^{-3})	T_C(K)	$\mu_s(\mu_B$ f.u.$^{-1})^a$	H_a(T)^a
$Nd_2(Fe_{0.9}Sc_{0.1})_{14}B$	8.814	12.19	947	7.43	610	23.6	≈3.6
$Nd_2(Fe_{0.9}Ti_{0.1})_{14}B$	8.798	12.19	944	7.50	410	18.5	≈2.6
$Nd_2(Fe_{0.9}V_{0.1})_{14}B$	8.800	12.20	945	7.55	415	18.1	≈1.7
$Nd_2(Fe_{0.9}Cr_{0.1})_{14}B$	8.796	12.20	944	7.57	550	23.4	≈2.3
$Nd_2(Fe_{0.9}Mn_{0.1})_{14}B$	8.827	12.25	955	7.51	350	15.5	≈2.7
$Nd_2Fe_{14}B^b$	8.792	12.19	942	7.62	585[b]	32.1[b]	≈6.3[b]
$Nd_2(Fe_{0.9}Co_{0.1})_{14}B$	8.796	12.17	942	7.66	668[c]	31.2	≈4.1
$Nd_2(Fe_{0.9}Ni_{0.1})_{14}B$	8.799	12.20	945	7.63	670	26.7	≈4.1

[a] At room temperature and maximum field 1.2 T.
[b] From refs. 3, 5 and 10.
[c] From ref. 12.

Concluding Remarks (1)

- How to apply ML to the design and discovery of new functional inorganic materials?
  - **Designers' class of new materials**
  - **Crystal inorganic materials**
    - Heterogeneous small data set
    - No systematic way to define descriptors
- New preliminary trial
  - **Item set mining for the data segmentation** in MI
  - Use of **site symmetries, electron configuration**, and their **co-occurrence relations**.
  - Once segmented to homogeneous data sets, RF regression can well predict the target property value.

Concluding Remarks (2)

- In MI for crystalline inorganic materials, **site symmetry and subblock co-ocurrence information might be used to find out homogeneous data subsets** from a heterogeneous data set.
- Mined patterns require further analyses for their physical interpretations.
- Looking for **appropriate applications**.

- Requiring more intensive collaboration with materials scientists.
- Please send me a mail to tanaka.yzr@ist.hokudai.ac.jp, if you are interested in collaboration.