Big Data Applications: Opportunities and Challenges

Yuzuru Tanaka[1) 2)]
Research Supervisor, JST CREST Program
on “Big Data Applications”
Professor Emeritus, Hokkaido University
[1)] MaDIS, NIMS (National Institute for Materials Science)
[2)] Faculty of Engineering, Hokkaigakuen University

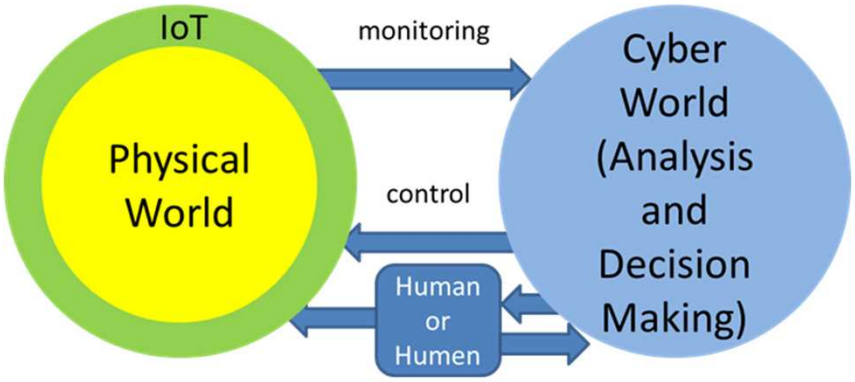
Opportunities

Two Potential Categories of Applications

- Urban-scale Social Cyber-Physical Systems for Secure, Sustainable, and Better Social Life
  - Optimizing social services such as
    - transportation / water supply and sewerage system / energy supply and consumption / traffic accident prevention / snow removal / …
  - Disaster management (preparedness, mitigation, response, and recovery)
  - Terror prevention
- Data-driven Sciences: Paradigm Shift from "X" Science to "X" Informatics for varieties of "X"
  - X: bio / biomedical / chemical / geo / brain / cosmological / meteorological / pharmaceutical / epidemiological / materials / …
- Cf. NSF's focused 2 areas for big data applications
  - Smart and connected communities
  - Harnessing data for 21st Century Science and Engineering

Urban-scale Social Cyber-Physical Systems with Humans in the Loop

- Real-time monitoring and control of the situation through IoT
  - weather, traffic & mobility, road condition, people's behaviors, energy consumption, CO_2, precipitation, earthquake, tsunami, epidemic, …
- Real-time assessment of the situation
  - Quantitative assessment
  - Geo-Visualization of states, events, and flows
  - Identification of their anomalies
- Prediction of the future situation
  - Data assimilation of simulation and observation
  - Machine learning
- Decision Making and Action to the Physical World
  - Based on real-rime assessment and/or prediction

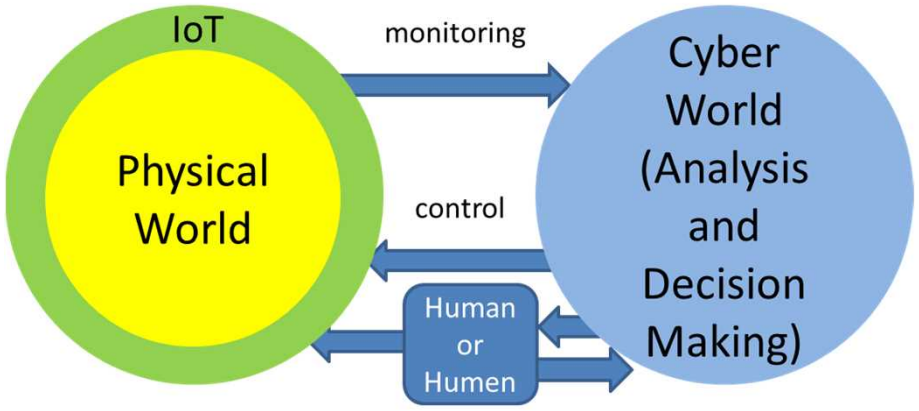


Data-driven Sciences

- Forerunners
  - Bio Informatics
  - Biomedical Informatics
- Followers
  - Materials Informatics

- (End-to-End) In-Silico-Science :
  - No physical transfer of mass → Open Science → Citizen Science

Bio-Bank,
High Throughput Experiment,
3D print



End-to-End In-Silico Science

JST CREST programs on Big Data

- 2013-2020
- Each winning project: 5.5 years
- CREST Program on Big Data Applications
  - PO: Yuzuru Tanaka (Hokkaido Univ.)
  - Collaboration between CS and/or Math researchers and domain science researchers is mandatory.
  - For either creating a new societal and/or economic value or discovering new scientific knowledge
- CREST Program on Big Data Core-Technologies
  - PO: Prof. Masaru Kitsuregawa (NII)

Design Policy of CREST Program on "Big Data Applications"

- Designing a good portfolio to cover challenging big data applications.
- Choosing a flagship project from each area.
- Promoting cross-disciplinary synergy, especially among young researchers.
- Clarifying the fundamental common denominator technologies, and integrate them into an open science platform.

Portfolio of Domain Sciences and Flagship Projects

2 Projects awarded in 2013

Pharmacy: Drug Discovery

- Development of a knowledge-generating platform driven by big data in drug discovery through production processes.
  - PI: Kimito Funatsu(Professor, The University of Tokyo)

- Innovating "Big Data Assimilation" technology for revolutionizing very-short-range severe weather prediction
  - PI: Takemasa Miyoshi(Team Leader, RIKEN)

Meteorology: 30 min ahead Forecasting of Localized Severe Rain

4 Projects awarded in 2014 (2)

Epidemiology: pandemic forecast ng

- Detecting premonitory signs and real-time forecasting of pandemic using big biological data
  - PI: Hiroshi Nishiura(Professor, Graduate School of Medicine, Hokkaido University)

- Statistical Computational Cosmology with Big Astronomical Imaging Data
  - PI: Naoki Yoshida(Professor, Department of Physics / Kavli IPMU, The University of Tokyo)

Cosmology: Discovery of new Super Novae and 3D Mapmaking of the Dark Matter Distribution

4 Projects awarded in 2014 (1)

- Establishing the most advanced disaster reduction management system by fusion of real-time disaster simulation and big data assimilation
  - PI: Shunichi Koshimura(Professor, International Research Institute of Disaster Science, Tohoku University)

Tsunami Disaster Prevention and Mitigation

- Exploring etiologies, sub-classification, and risk prediction of diseases based on big-data analysis of clinical and whole omics data in medicine
  - PI: Tatsuhiko Tsunoda(Professor, Medical Research Institute, Tokyo Medical and Dental University)

Personalized/ Precision Medicine

3 Projects awarded in 2015

- Data-driven analysis of the mechanism of animal development
  - PI: Shuichi Onami(Team Leader, Quantitative Biology Center, RIKEN)

- Knowledge Discovery by Constructing AgriBigData
  - PI: Masayuki Hirafuji(Project Professor, Graduate School of Agricultural and Life Sciences, The University of Tokyo)

- Knowledge Discovery through Structural Document Understanding
  - PI: Yuji Matsumoto(Professor, Graduate School of Information Science, Nara Institute of Science and Technology)

Developmental Biology: Automatic Digitization of Development Processes

e-Agriculture: Phenotyping

Literature-based Knowledge Discovery

Fundamental Common Denominator Technologies

- Varieties of Data Science Algorithms: applicability and restrictions
- Literature-based knowledge discovery: from big data to big mechanism
- Data assimilation of real-time observation and physical-model based ensemble simulation for the high-precision real-time prediction of the near future
  - Continuous system modeling (well studied) / discrete system modeling (not well studied yet)
- Exploratory visual analytics to cope with the heterogenous nature of available training data sets.
  - Interactive segmentation of heterogenous data to sets of homogeneous data, and analysis of each of them
  - Definition, management, and execution of such analysis process scenarios.
- Integration Platform: Cyber Research Infrastructure
  - Hands-on portals for 9 projects
- Ontology-based management of resources, analysis scenarios, users, and projects.

Advisory Board

International Advisors

• Costantino Thanos DB, Cyber Research Infrastructure	Research Director, Institute of Information Science and Technologies
• Norbert Graf Personalized Medicine	Professor, Director, Saarland University Hospital
• Nicolas Spyratos DB, Big Data Analytics	Professor Emeritus, University of Paris Sud 11
• Nigel Waters GIS	Professor Emeritus, University of Calgary
• Randolph Goebel ML, Visual Analytics	Professor, University of Alberta

Local Advisors

• Hajime Amano	President, ITS Japan
• Ryosuke Shibasaki	Professor, Center for Spatial Information Science & Institute of Industrial Science, The University of Tokyo
• Masafumi Shimoda	Business Strategy Advisor, DNA Chip Research Inc.
• Ryosuke Suzuki	Consultant, Nomura Research Institute, Ltd.
• Koichi Takeda	Professor, Graduate School of Informatics / Director, Future Value Creation Research Center Nagoya University
• Yasumasa Nishiura	Professor, WPI Advanced Institute for Materials Research, Tohoku University
• Tomoko Matsui	Professor, The Institute of Statistical Mathematics
• Satoru Miyano	Professor, Human Genome Center, Institute of Medical Science, The University of Tokyo

Symposiums on Big Data Applications

- September Symposium (1day)
  - 2 keynote speakers
    - Michele Sebag in 2017
    - Christos H. Papadimitriou in 2018
  - progress report by each of 9 PIs
- January Symposium (2 days)
  - 1 keynote speaker
    - Dennis Tsichritzis in 2015
    - Christos Faloutsos in 2018
    - Stuart Kaffman in 2019?
  - Each project session
    - 1 invited speaker
    - Progress report by PI and members
- + Joint Symposium with CREST Program on Big Data Core Technologies (NSF-JST, DATAIA-JST)

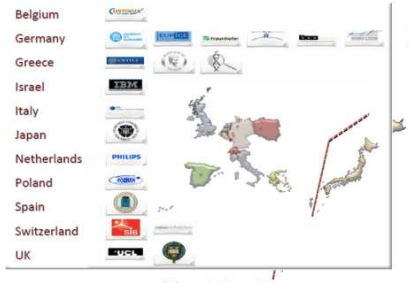
My Involvement in Big Data projects (1)

- Biomedical Science: Personalized Medicine for cancer
  - EU FP projects for integrated IT support of clinical trials on cancer
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010): 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015): 29 teams

The ACGT Consortium

CBI学会2015年大会 10月29日

The p-medicine Consortium

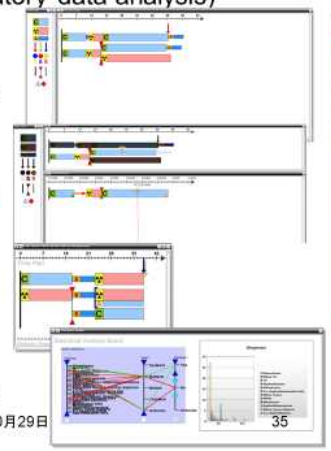


CBI学会2015年大会 10月29日

Trial Outline Builder (TOB) (2010)

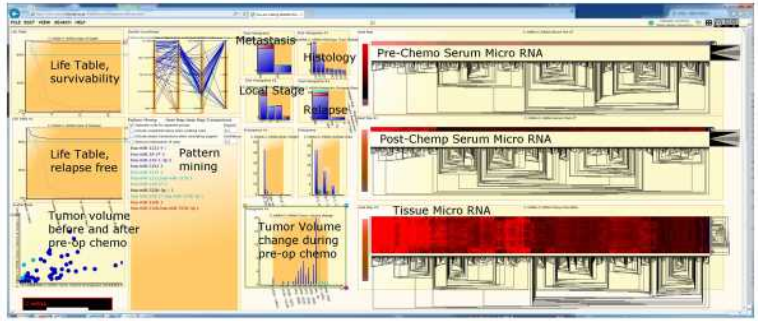
(Web-top integrated environment for planning trials, patient data acquisition, and exploratory data analysis)

- Trial Plan Editing
  - Copy-and-paste of trial event types to design both a trial flow graph and a set of some additional events outside the flow.
  - A click of each event opens its CRF editor
- Patient Treatment View : CRF input for each patient through the TOB
  - Possibly with the specification of some additional outside-of-flow events
- Query & Analysis View: Querying the DB
  - for specific cases for their statistical analysis or the visualization of correlations among specified items



CBI学会2015年大会 10月29日 35

TOB for analyzing the Effect of Pre-op Chemotherapy



Each found pattern may work as a new biomarker to identify those patients who are helped or not helped by the preop chemotherapy

CBI学会2015年大会 10月29日 63

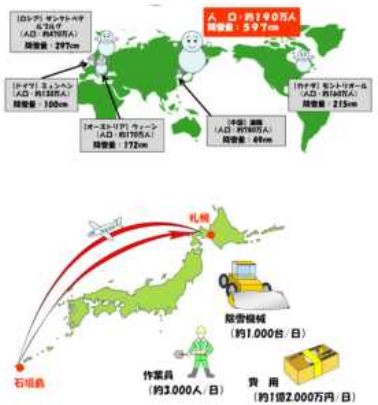
Exploratory Visual Analytics

My Involvement in Big Data projects (2)

- Social Science: Urban-scale Monitoring and Service Optimization
  - MEXT initiative project on Social CPS (Cyber-Physical System) for Efficient Social Services (09/2012-03/2017)
    - Project Consortium (NII (National Institute of Informatics), Hokkaido Univ., Osaka Univ., Kyushu Univ.)
    - Hokkaido team focuses on smart snow removal.

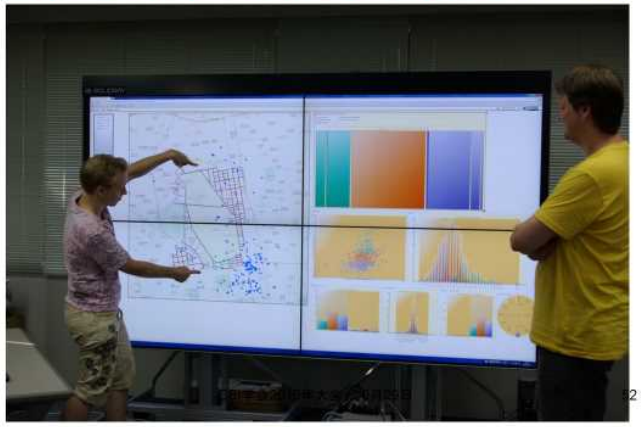
Snow Removal in Sapporo as a Large-scale Complex Social Service

- Population: 1,920,739
- Annual snowfall: 597cm
  - The largest annual snowfall among the cities with more than 1M people in the world
- Annual budget for snow plowing and removing (2010):
  14,729,000,000 yen
  (147,000,000 $)
  2nd last season:
  22,000,000,000 yen
  (220,000,000 $)
- Total distance of snow plowing and removing during a single night: 5,328km



CBI学会2015年大会 10月29日 38

Geospatial Digital Dashboard for Exploratory Visual Analytics (2013)



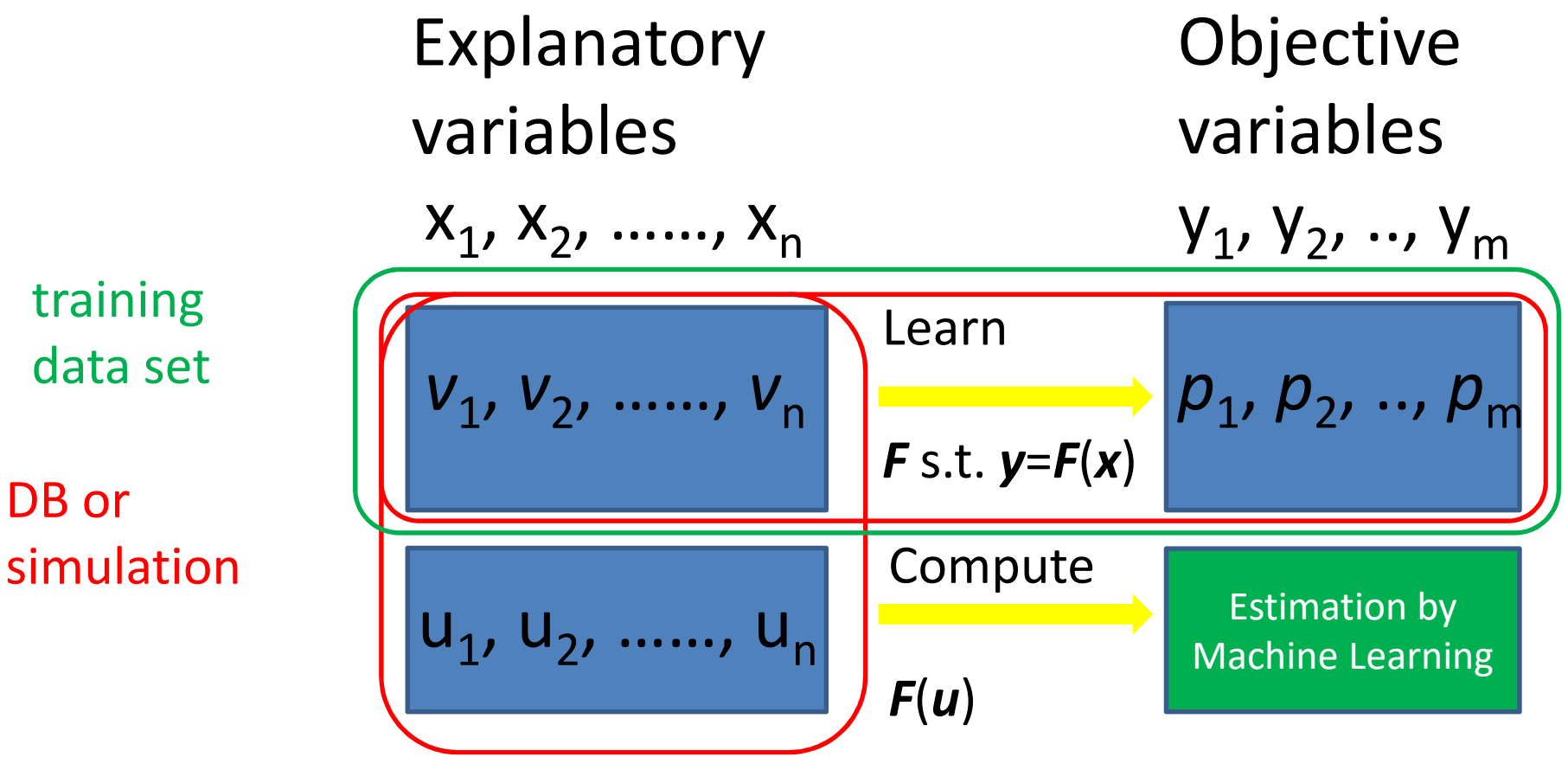
Exploratory Visual Analytics

- Material Science: Collaboration with Dr. Keisuke Takahashi at NIMS (National Institute for Materials Science) (2014-)

Materials Informatics

- Current status: emerging period
  - Computational (and/or experimental) materials science with the help of ML-based data analysis
- 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
- Main targets: natural materials with modifications

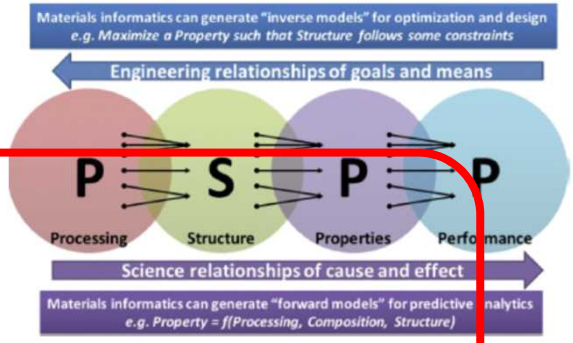
ML for Speed-Up



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

What ML to learn?

3 Major Goals



- Materials Discovery:

  Find the material with maximum performance
  - DFT to compute F : Structure → Performance
  - ML (regression) to learn F as an explicit function
  - Inverse Problem: arg max $F(\boldsymbol{x})$

experiments: years
simulations: hours, days
machine learning: seconds
(for candidates discovery)

- Measurement Analysis:

  Identify the material structure from its measurement result
  - (Measurement Data) + Simulation Data: F^*: Structure → Property
  - ML (Deep Learning) to learn 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

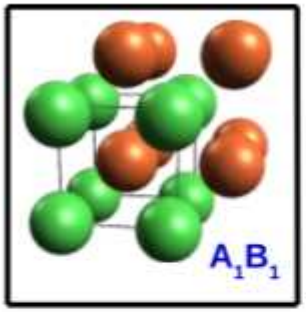
High Speed Estimation of Lattice Constants

THE JOURNAL OF CHEMICAL PHYSICS 146, 204104 (2017)

Descriptors for predicting the lattice constant of body centered cubic crystal

Keisuke Takahashi,[1,2,a] Lauren Takahashi,[3] Jakub D. Baran,[4] and Yuzuru Tanaka[1]
[1]*Center for Materials research by Information Integration (CMI[2]), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
[2]*Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan*
[3]*Freelance Researcher, Central Ward, Sapporo 064, Japan*
[4]*Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom*

(Received 24 February 2017; accepted 11 May 2017; published online 24 May 2017)



A: Ag, Al, As, Au, Co, Cr, Cu, Fe, Ga, Li, Mg, Na, Ni, Os, Pd, Pt, Rh, Ru, Si, Ti, V, W, Zn

B: Atomic numbers 1-42, 44-57, 72-83.

Material	Predicted	Calculated	Error	Experiment	Error
FeAl	2.92	2.90	0.7%	2.91 [15]	0.3%
FeTi	3.08	2.96	3.9%	2.98 [16]	3.3%
CoTi	3.07	2.99	2.6%	3.00 [17]	2.3%
MgAg	3.41	3.36	1.5%	3.31 [18]	2.9%
ScAl	3.47	3.37	2.9%	3.39 [19]	2.3%

Exp: months or years
Comp.: hours
ML: seconds

To find good descriptors

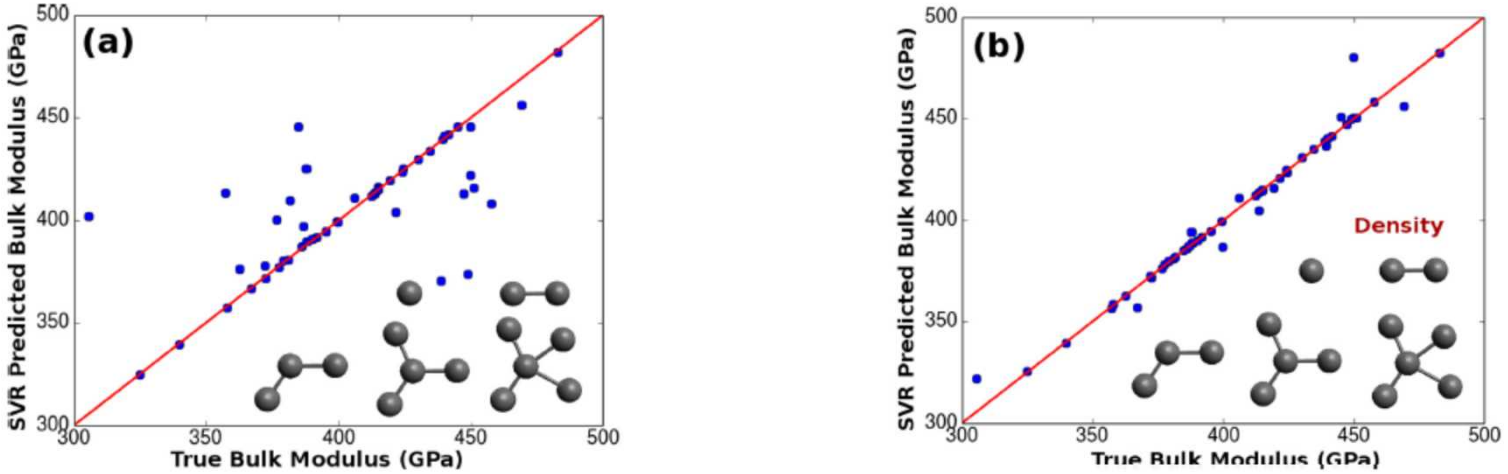
PHYSICAL REVIEW B 95, 054110 (2017)

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.

2D Magnets
(Journal of Physics: Condensed Matter)


Miyasato

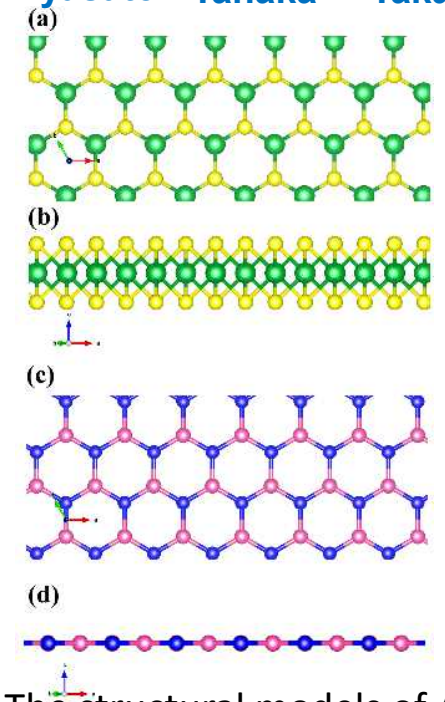

Tanaka


Takahashi

216 2D Materials Data + ML (4 discriptors)

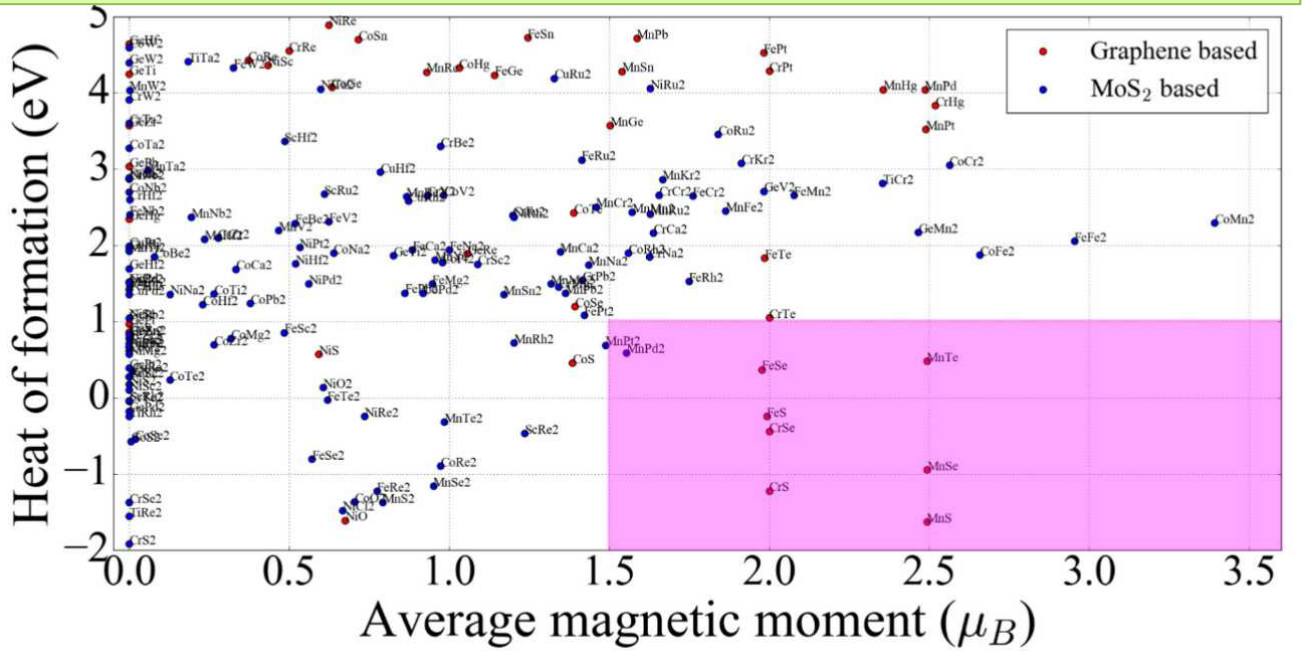
Prediction 254 2D Materials with High Magnetic Moment

Checking by DFT



The structural models of AB2 in top (a) and side (b) view and graphene based AB in top (c) and side (d) view

8 undiscovered stable 2D materials with high magnetic moments

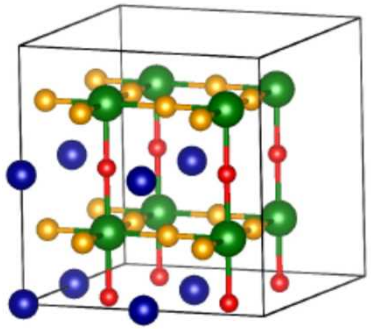


Searching for hidden perovskite materials

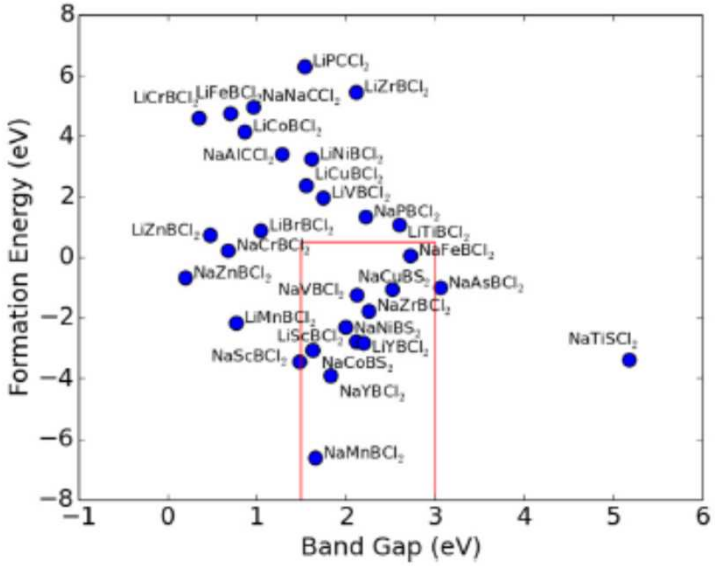
(ACS Photonics: Keisuke Takahashi, Lauren Takahashi, Itsuki Miyazato, and Yuzuru Tanaka)

To find perovskite materials within the ideal band gap and formation energy ranges for solar cell applications

- 15,000 perovskite materials data for ML (random forest) to predict the band gap
- 18 physical descriptors are revealed to determine the band gap.
- 9,328 perovskite materials with potential for applications in solar cell materials are predicted.
- The selected Li and Na based perovskite materials within predicted 9,328 are evaluated with DFT.
- 11 undiscovered Li(Na) based perovskite materials are found.



Atomic model of perovskite materials, ABC2(C1,C2)D. Atomic color code; Blue:A, Green:B, Yellow: C, Red:D.

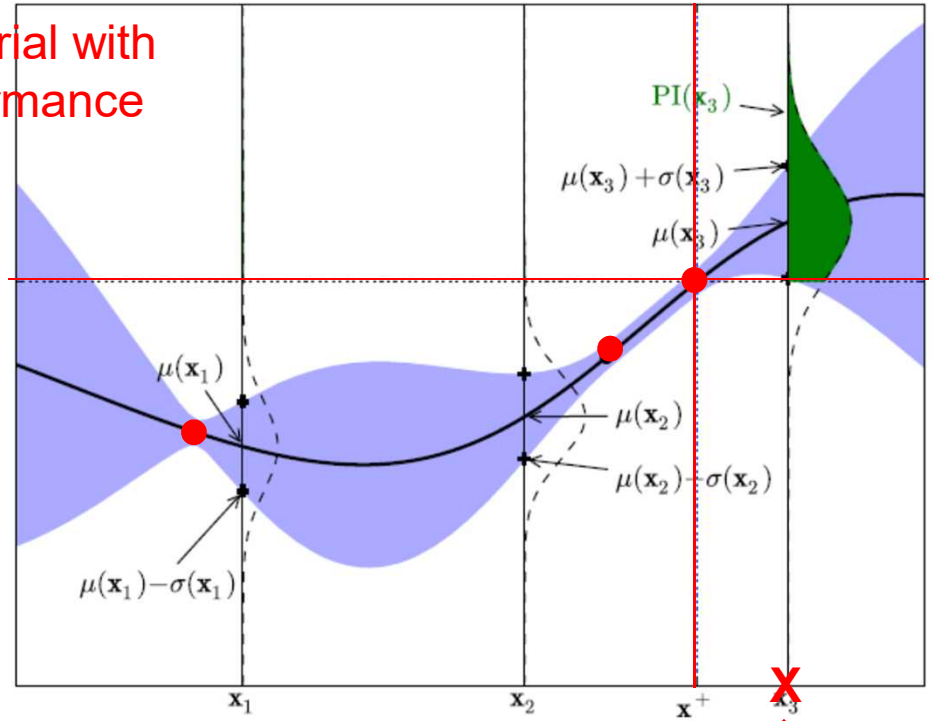


Emerging MI: Progressive Exploratory Search: Bayesian Optimization

To find the material with maximum performance

Gaussian Process Regression with 3 sample points

3 sample points



current maximum

X: search space

$$\begin{aligned} \mathrm{PI}(\mathrm{x}) &= P(f(\mathrm{x}) \geq f(\mathrm{x}^+)) \\ &= \Phi\left(\frac{\mu(\mathrm{x}) - f(\mathrm{x}^+)}{\sigma(\mathrm{x})}\right), \end{aligned}$$

Next point to sample
for maximizing the expected improvement PI(x), i.e., size of the green area
→ guiding experiment or DFT computation

Bayesian Optimization

- No guarantee that the physics follows the Gaussian distribution assumption.
- Exploration may somehow finally reaches to a local maximal, not a global maximum.
- Question:
  - Can this method find $Nd_2Fe_{14}B$, starting from $SmCo_5$?
  - Probably not, since they follow different physics.
  - How about $Sm_2(Co, FE, Cu, Zr)_{17}$, starting from $SmCo_5$?
  - ?

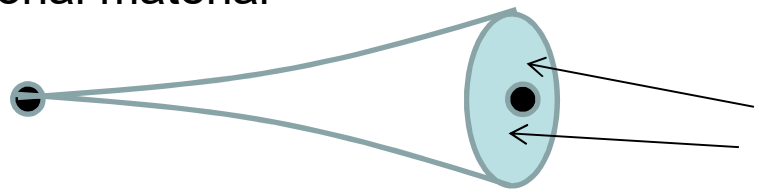
How to find a needle in a haystack?

In the neighbor of a found one

An experimentally discovered champion functional material

its neighborhood as an exploration space

Current Focus of Computational and/or Experimental materials Scientists

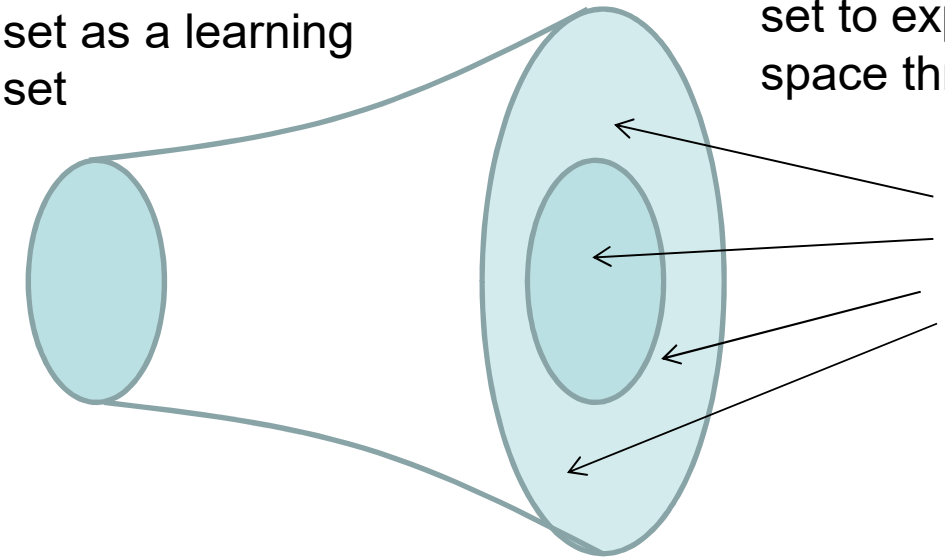


discovery of desired materials

Brute force exhaustive filtering

Simulation / experiment data set as a learning data set

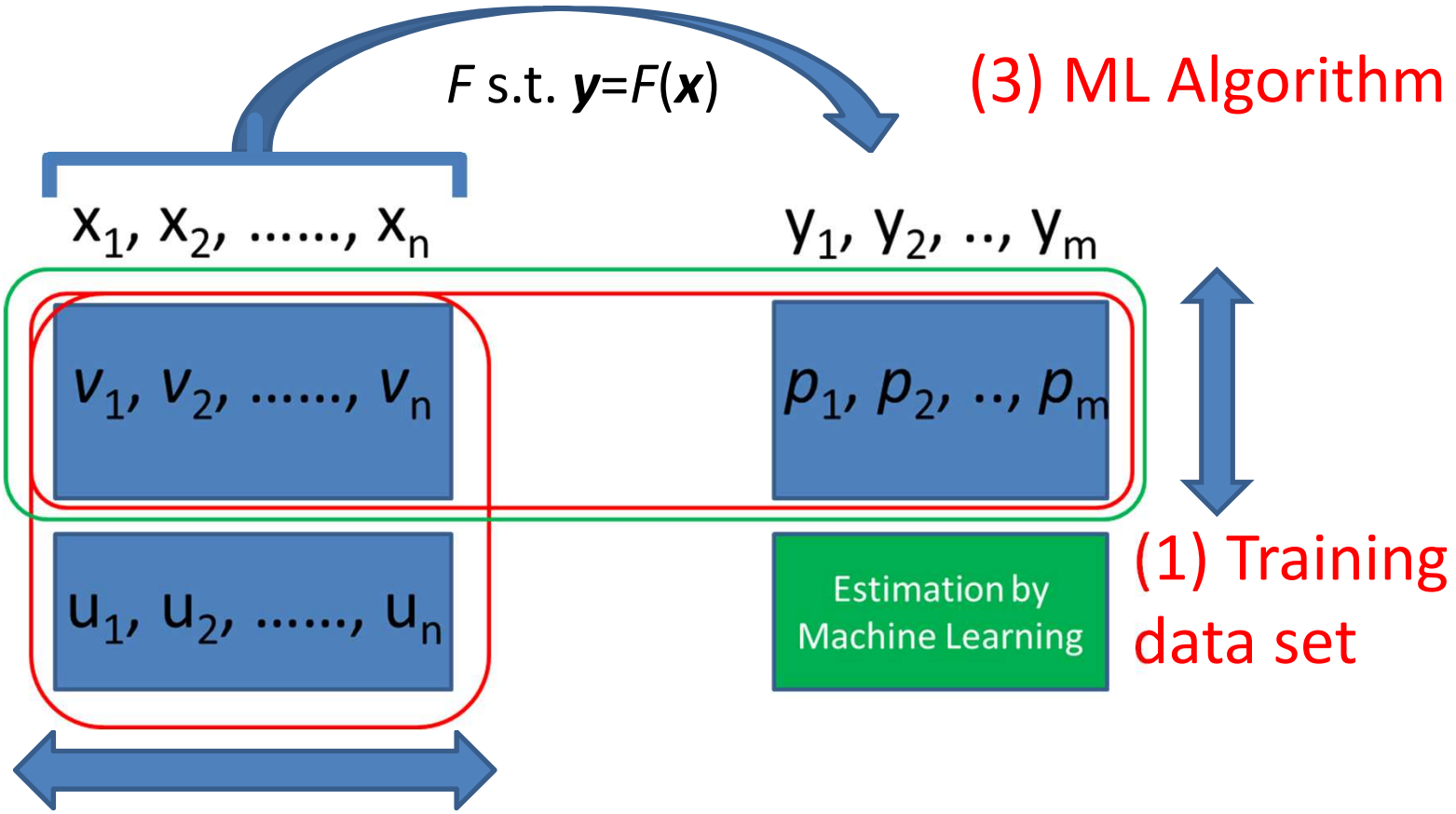
Simulation / experiment data set to expand an exploration space through ML



discovery of desired materials

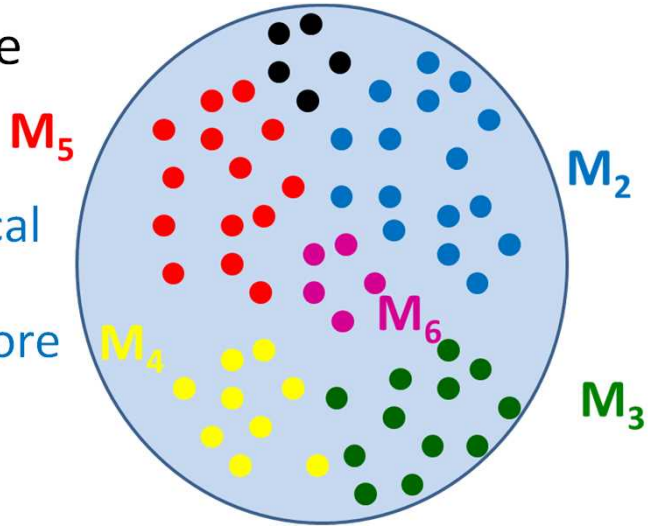
Challenges

3 Things to Consider



(1) Training Data Set

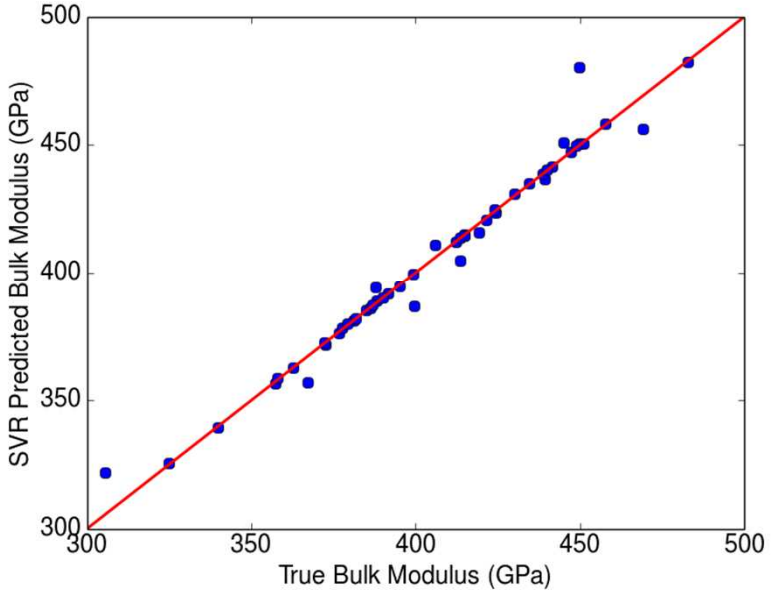
- Heterogeneous (also observed in urban-scale traffic in different road links, and in chemo-response of patients and tumors)
  - Different groups follow different mathematical models.
  - Appropriate segmentations are required before analysis!
- Size of each homogeneous data set
  - Inorganic materials: $10^3 \sim 10^4$
  - Difficult to provide more than 10^5 data
    - No more variations of structures and components
    - Both DFT computations and experiments are time-consuming

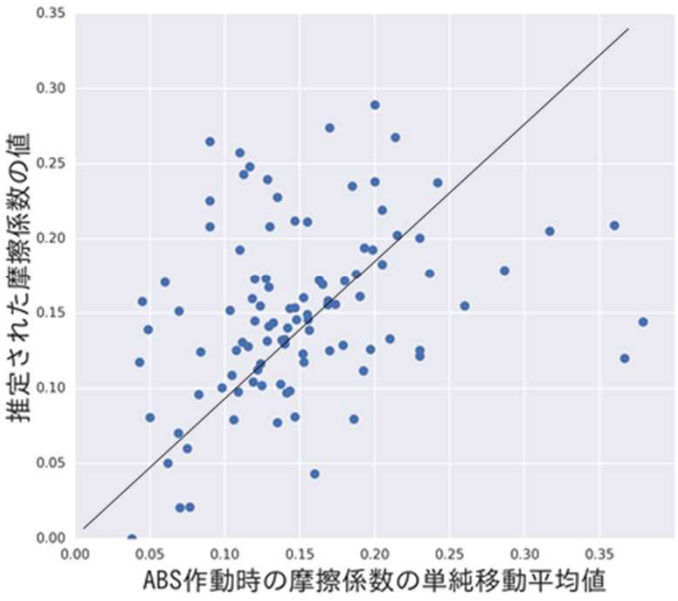


How to increase the size of the homogeneous training data set? Is it really necessary?

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

- SVR works well to find a hidden physical order which follows a math model.
  - Different from data sets in other research areas





ML is to find out a hidden physical order whose mathematical model is not known yet, and to give its approximate function.

(2) Set of Descriptors

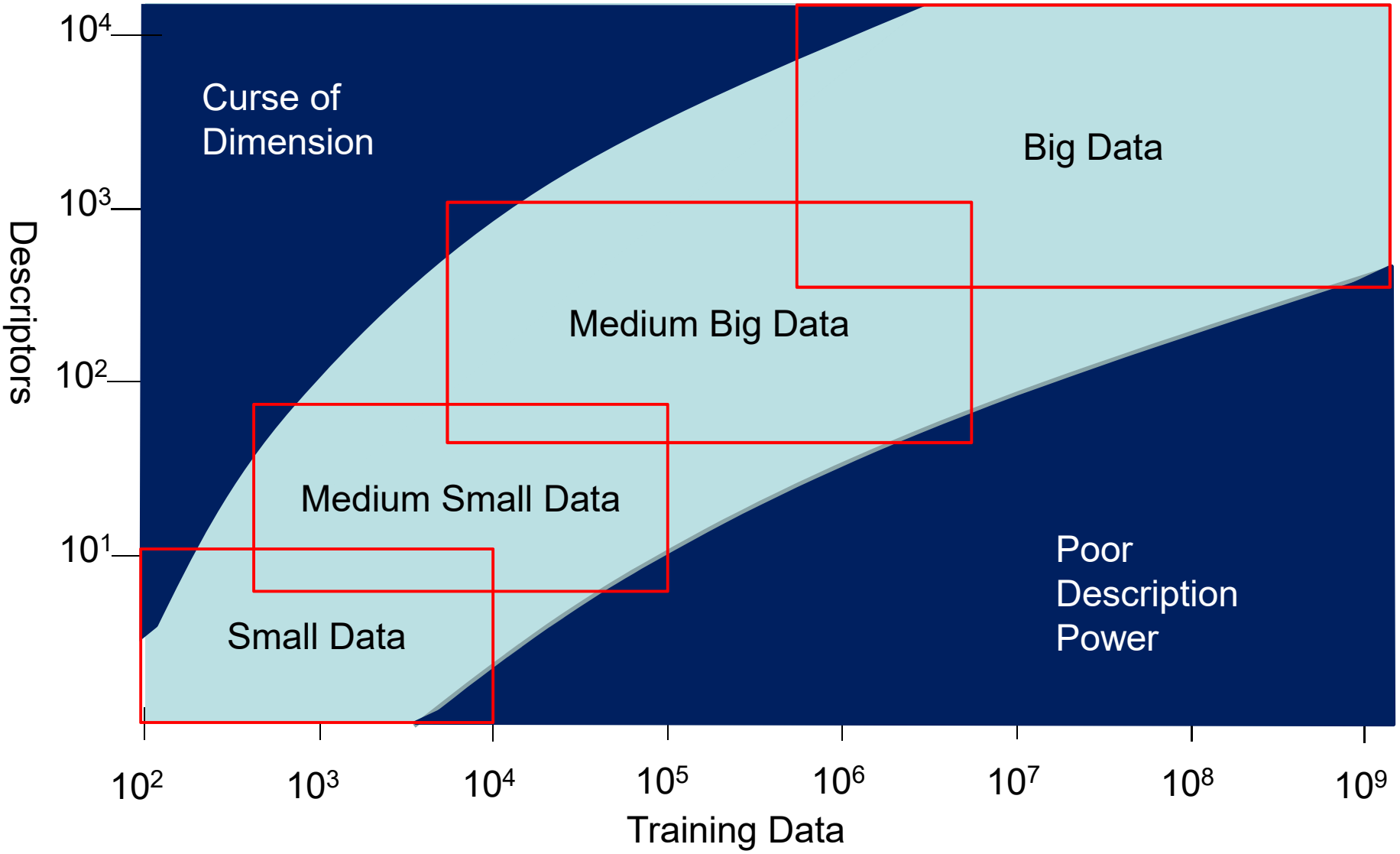
- No systematic way to define an arbitrarily large set of descriptors
  - cf. Genome systematically provides genes, their expressions, and gene alterations as descriptors.
- A large training data set requires a large set of descriptors,
- while a small training data set needs to use only a small set of good descriptors to avoid the curse of dimensions.

How to **systematically define** a large set of descriptors?

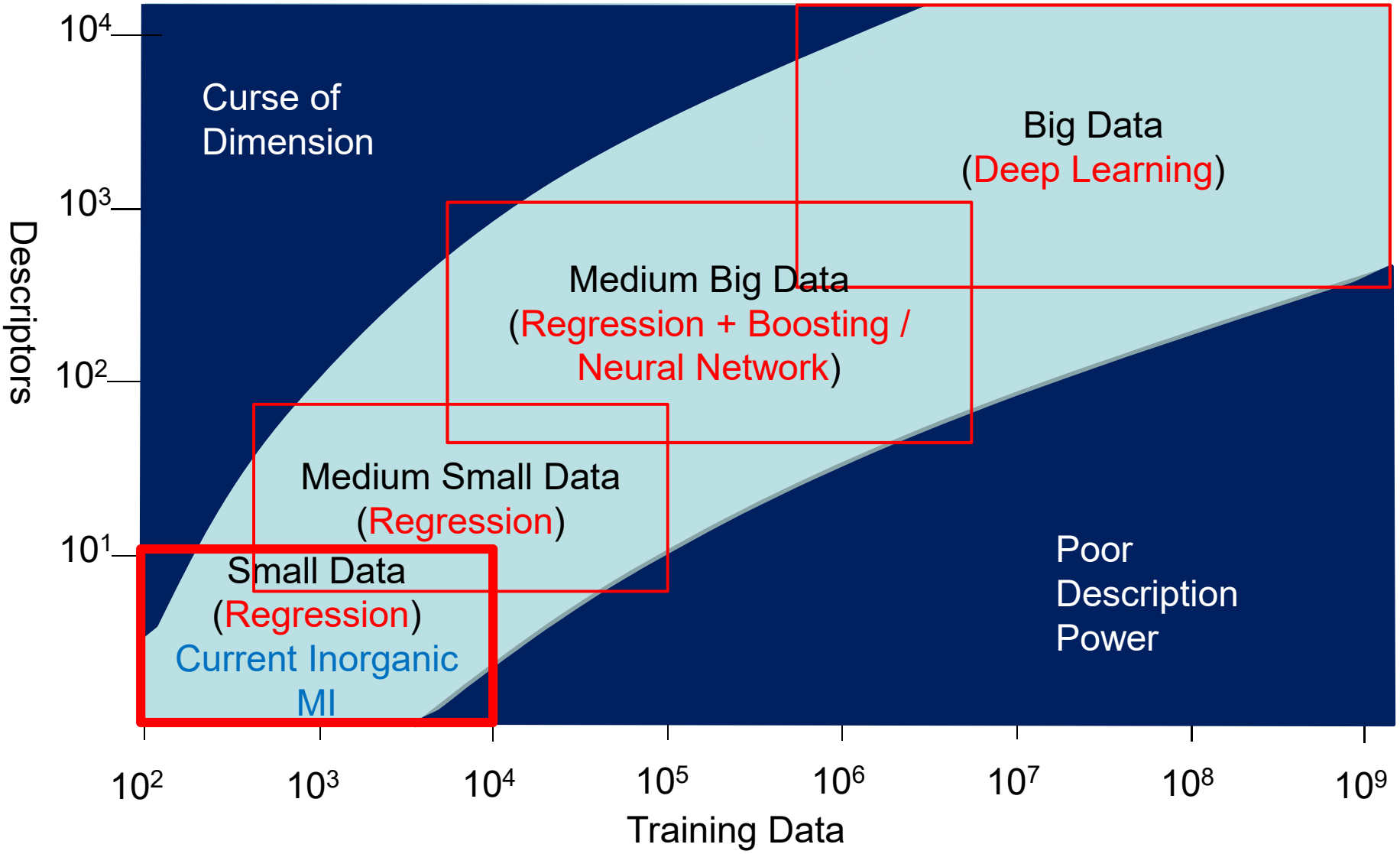
How to define a **small set of good descriptors** for a small set of training data to avoid the curse of dimensions?

Big Data vs. Small Data

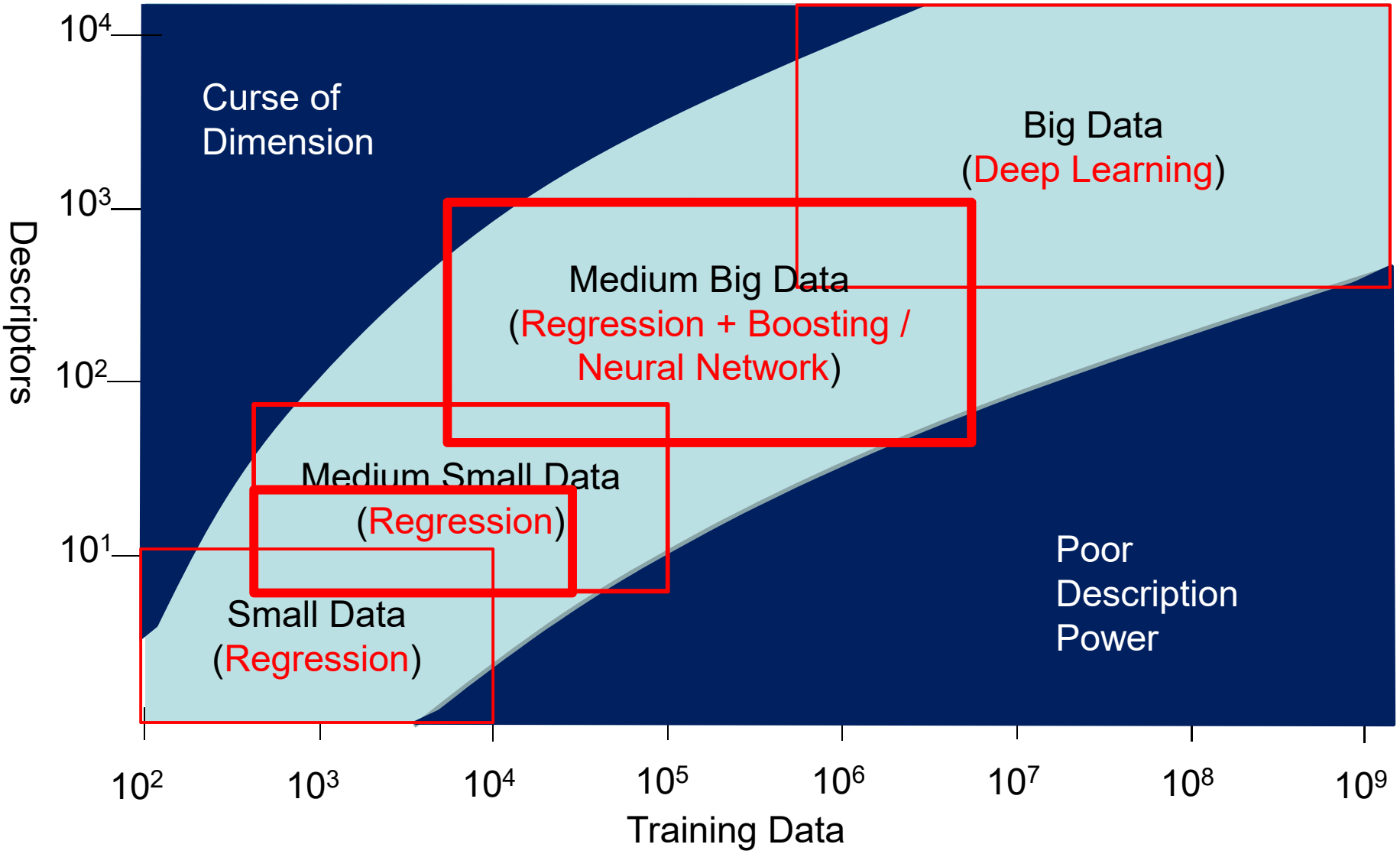
after segmented into homogeneous data sets



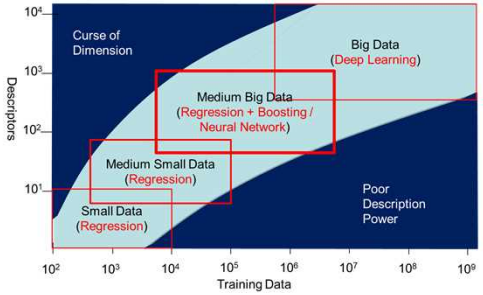
(3) ML Algorithm



Potential MI Scenarios to Come

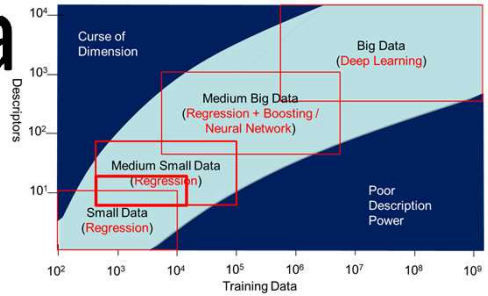


Medium Big Data



- Big Data, or Small Data?
  - Organic materials may result in big data ($\geq 10^6$)
  - Inorganic materials result in small data ($\leq 10^4 \sim 10^5$)
- Some people try to increase the data size.
  - combinatorial design
  - organometallic materials whose skeletal polymers increase the variety.
- Systematic Definition of Descriptors
  - some researchers focus on organometallic materials.
    - The SMILE representations of their skeletal polymers enable them to systematically define descriptors.

Medium Small Data



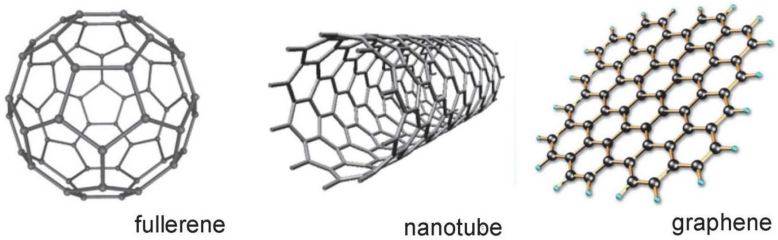
- Target:
  - 10^3-10^4 homogeneous simulation data and/or HTE data
  - Less than 10^1 governing well-designed descriptors
  - Heterogeneous data consisting of those homogeneous ones.
- Method:
  - First, segmentation
    - What kind? → (new segmentation algorithm based on item-set mining)
  - Then, regression
    - SVR-based machine learning to reveal hidden orders as math functions
- Numerical solution to inverse problems

Medium Small Data: Design Parameters as Descriptors

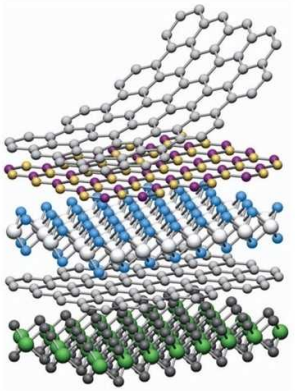
- Designers class of materials
  - Artificially designed materials
  - No more than 10 design parameters
  - Data can be acquired through HTE or HTC
  - # of materials in the class $>10^3$
  - Simultaneous materials discovery for varieties of functions
- Design framework: combinatorial design
  - Multilayered 2D materials
  - scaffolding + modifiers
    - Scaffolding: functional / nonfunctional
    - modifiers to give functions
    - Different scaffoldings define different classes.

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

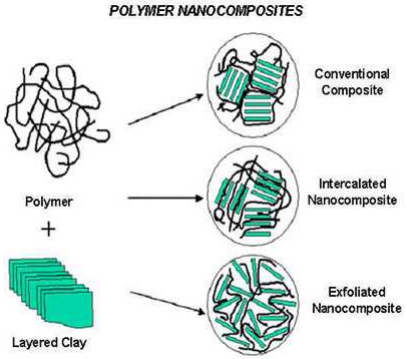
- Carbon-based ones:

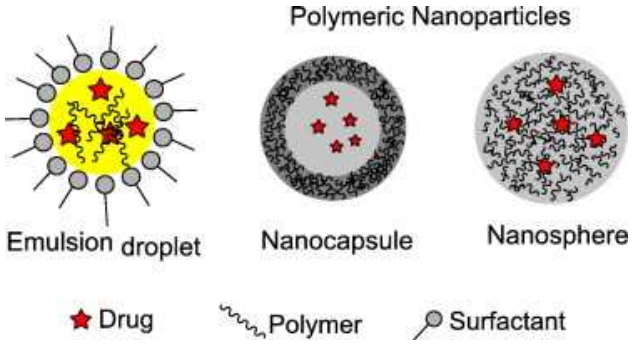


- 2D materials + layered structures
  - Intralayer modifier
  - Interlayer modifier



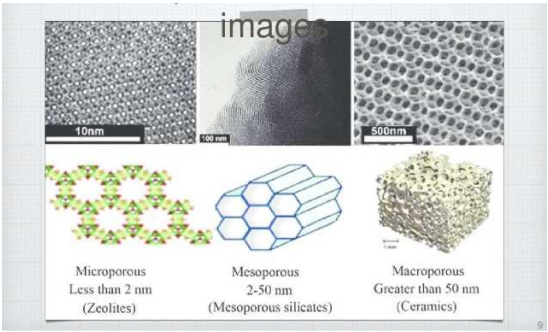
- Polymer nanocomposites/nanoparticle



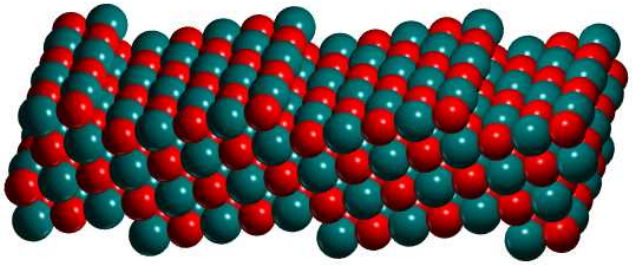


(Scaffolding + Modifiers) Framework: Candidates of Scaffolding (2)

- Nano pores

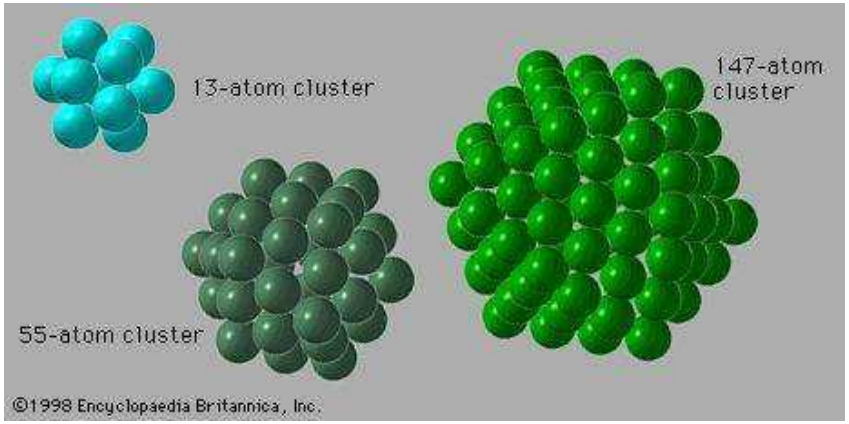


- Crystal surface

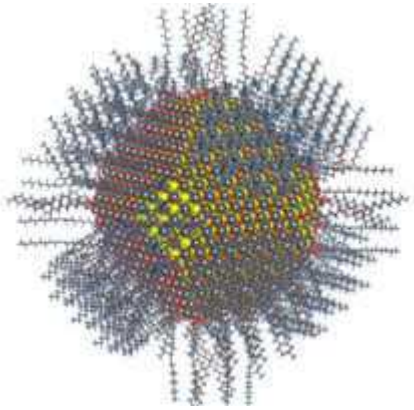


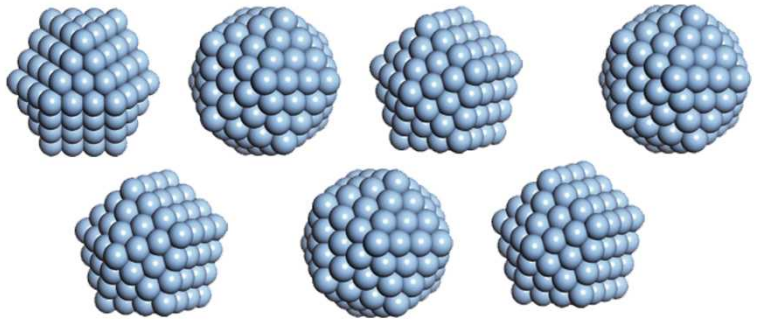
(Scaffolding + Modifiers) Framework
(1) Modifiers

- Single atom
- Atom cluster



- Nano particle

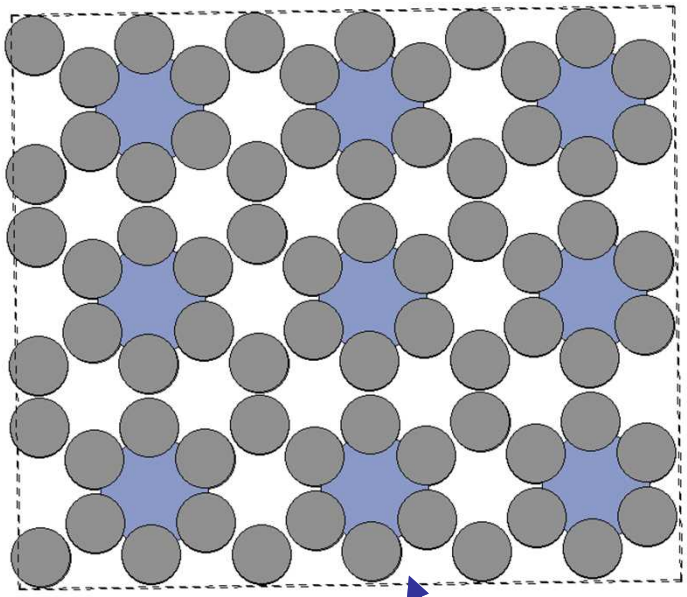


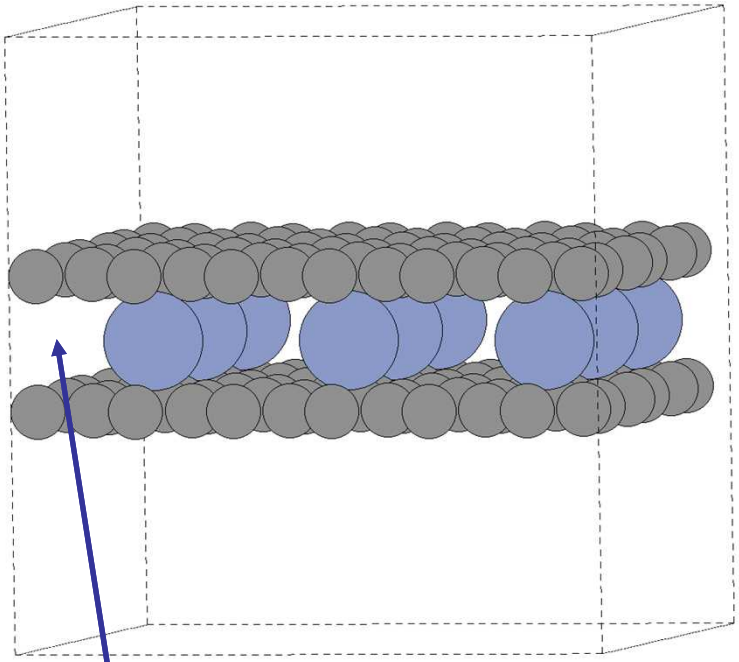


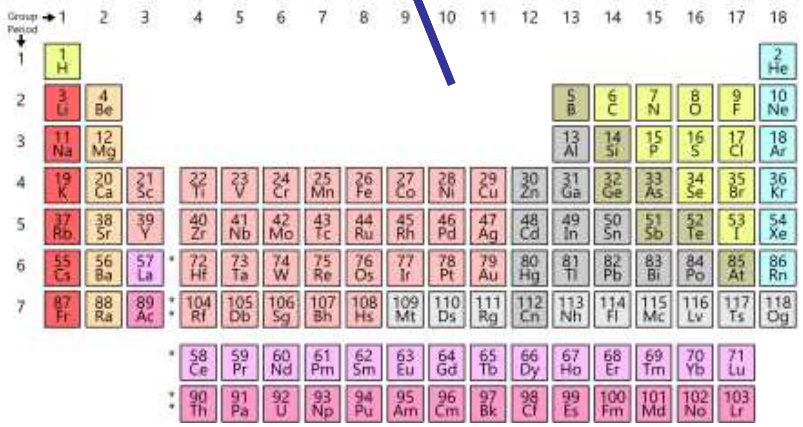
Special focuses on red ones

- Others:
  - DFT computation becomes difficult.
  - No translational symmetry
- Biggest interest on
  - Double layered 2D Materials with metallic atoms or clusters as interlayer modifiers
    - For ML-based analysis, the scaffolding 2D material is fixed.
  - Nano particles
    - Design parameters can be well defined

Single atom between graphene layers



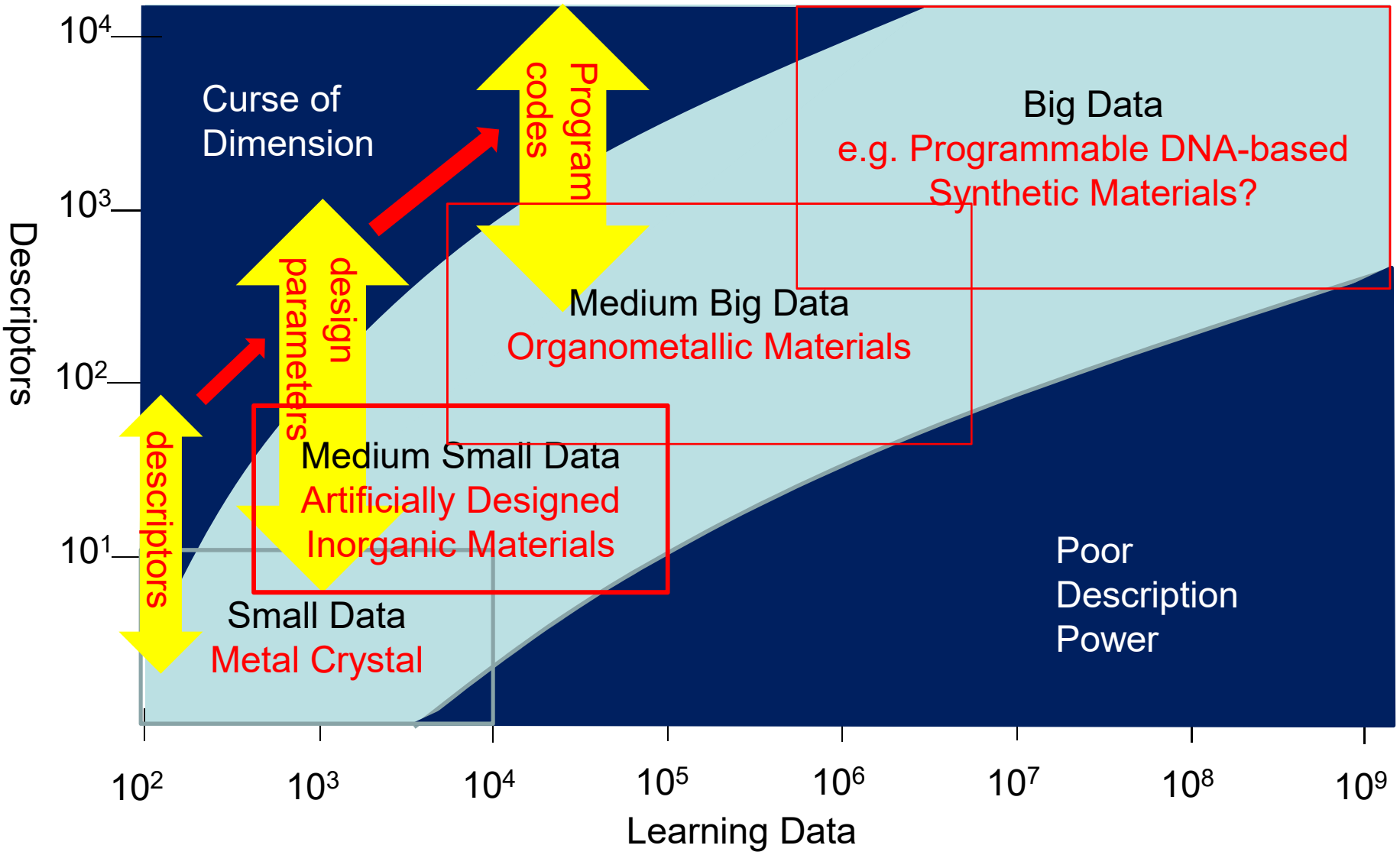




Van der Waals force

Properties of single atoms are well preserved

Landscape of Materials Informatics



Take Home Messages

- How to deal with the heterogeneity of data in practice?
  - Exploratory visual analytics
  - From description to design
- Implications both from the nature of inorganic materials and from ML
- Target: Designers classes of materials
  - 10^3-10^4 DFT data and/or HTE data
  - ≤ 10 governing well-designed descriptors
- Method: Segmentation → Regression
- Open Question
  - What kinds of designers classes of materials can effectively exploit both DFT and ML for the exhaustive filtering of its whole search space?