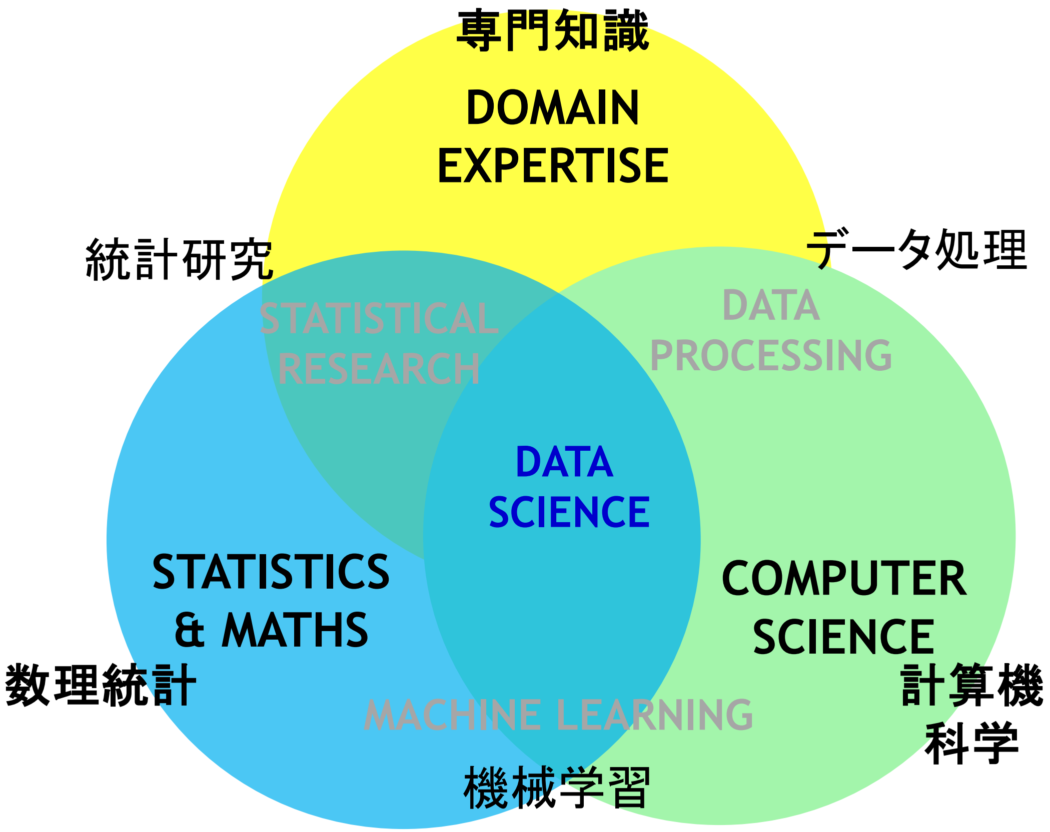
データサイエンスの基礎

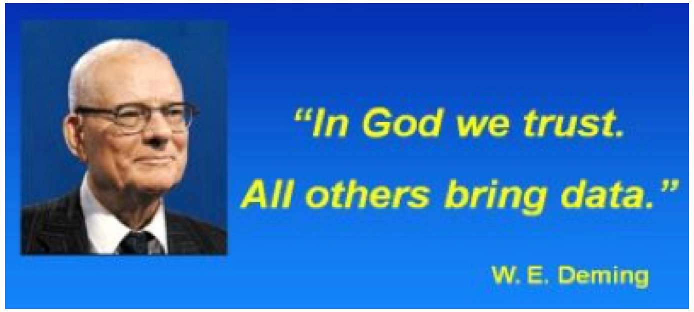
北陸先端科学技術大学院大学
Hieu-Chi Dam

2024年7月8日（第1部 18:00 ~ 19:20）

目次

データサイエンスの定義





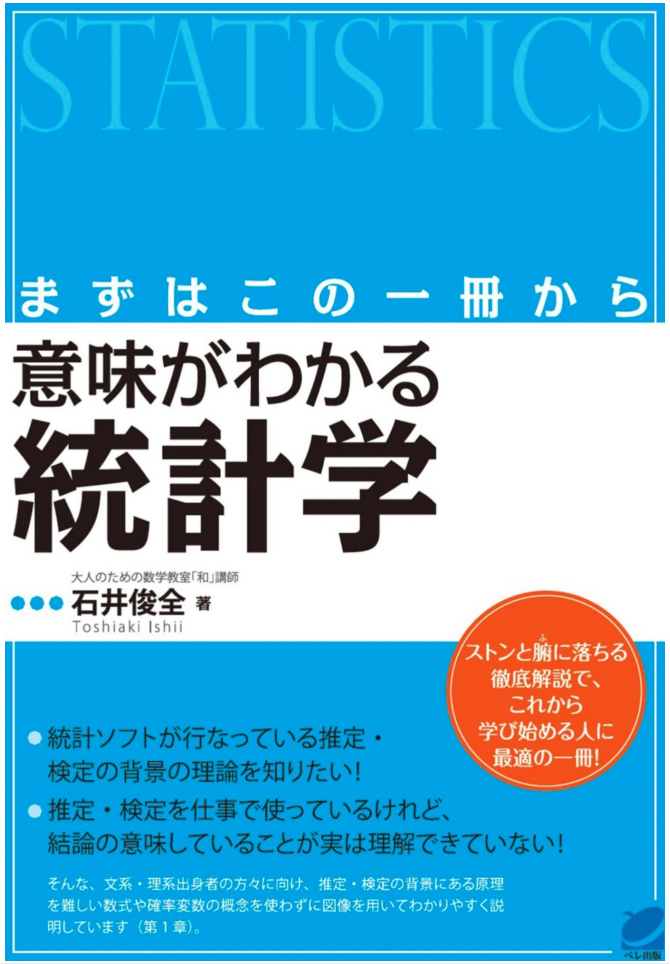
Data Scientist: The Sexiest Job of the 21st Century

(Harvard Business Review, October 2012)

データの理解と活用

1. どのように人々がデータを集め，整理するのかを理解する
2. データの特徴（データ種類，データの特性）を理解する
3. 分析する時のタスクを知る
4. 与えられたタスクに適切なデータ分析手法について知っている
5. 分析ソフトウェアについて知っている
6. 分析から得られた結果について説明できる

参考書の紹介


STATISTICS
まずはこの一冊から
意味がわかる
統計学
大人のための数学教室「和」講師
石井俊全 著
Toshiaki Ishii
ストンと腑に落ちる
徹底解説で、
これから
学び始める人に
最適の一冊!
統計ソフトが行なっている推定・
検定の背景の理論を知りたい!
推定・検定を仕事で使っているけれど、
結論の意味していることが実は理解できていない!
そんな、文系・理系出身者の方々に向け、推定・検定の背景にある原理
を難しい数式や確率変数の概念を使わずに図像を用いてわかりやすく説
明しています（第1章）。


Mathematical
model
データ分析のための
数理モデル入門
本質をとらえた分析のために
本当にそのモデルで
いいんですか？
微分方程式や確率過程、
エージェントベースモデルってなに？
統計検定と数理モデルの関係は？
強化学習じゃなくて制御でもいい？
ディープラーニングよりも遥かに簡単な
モデルで精度が出る？
数理モデル
全体を俯瞰する
全く新しい
教科書!
東京大学先端科学技術
研究センター
江崎貴裕
データサイエンティストの
新たな教養「数理モデル論」
ソシム


Data
hermeneutics
分析者のための
データ解釈学入門
データの本質をとらえる技術
モデリングの前後が疎か
になってませんか？
分析結果は出せるが自信が持てない
統計解析がなにをやっているのかわからない
分析における数理モデルの役割とは何か？
予測しやすい問題と予測しにくい問題とは？
データ
分析者必携の
全く新しい
教科書!
東京大学先端科学技術
研究センター
江崎貴裕
データの取得・分析・
解釈・活用の各段階で
知っておくべき技術を網羅的に解説！
ソシム

イントロ
その２：推論方式の紹介

演繹法，帰納法，およびアブダクション

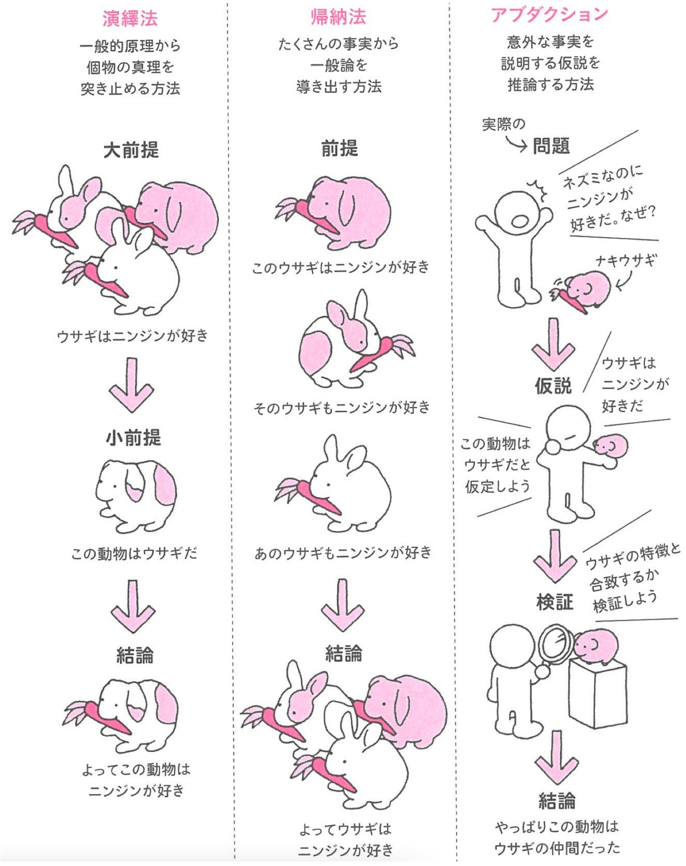
科学のパラダイム:

実験科学　–　第１パラダイム
- 実験，観測などから結論を導く

理論科学　–　第２パラダイム
- ニュートン方程式など

計算科学　–　第３パラダイム
- 複雑現象の計算機シミュレーション，計算機実験など

データ科学　–　第４パラダイム
- 理論，実験，計算機実験，データマイニング



哲学用語図鑑・田中正人（著）・プレシデント社

演繹法，帰納法，およびアブダクション

科学のパラダイム:

実験科学　–　第1パラダイム

- 実験，観測などから結論を導く

理論科学　–　第2パラダイム

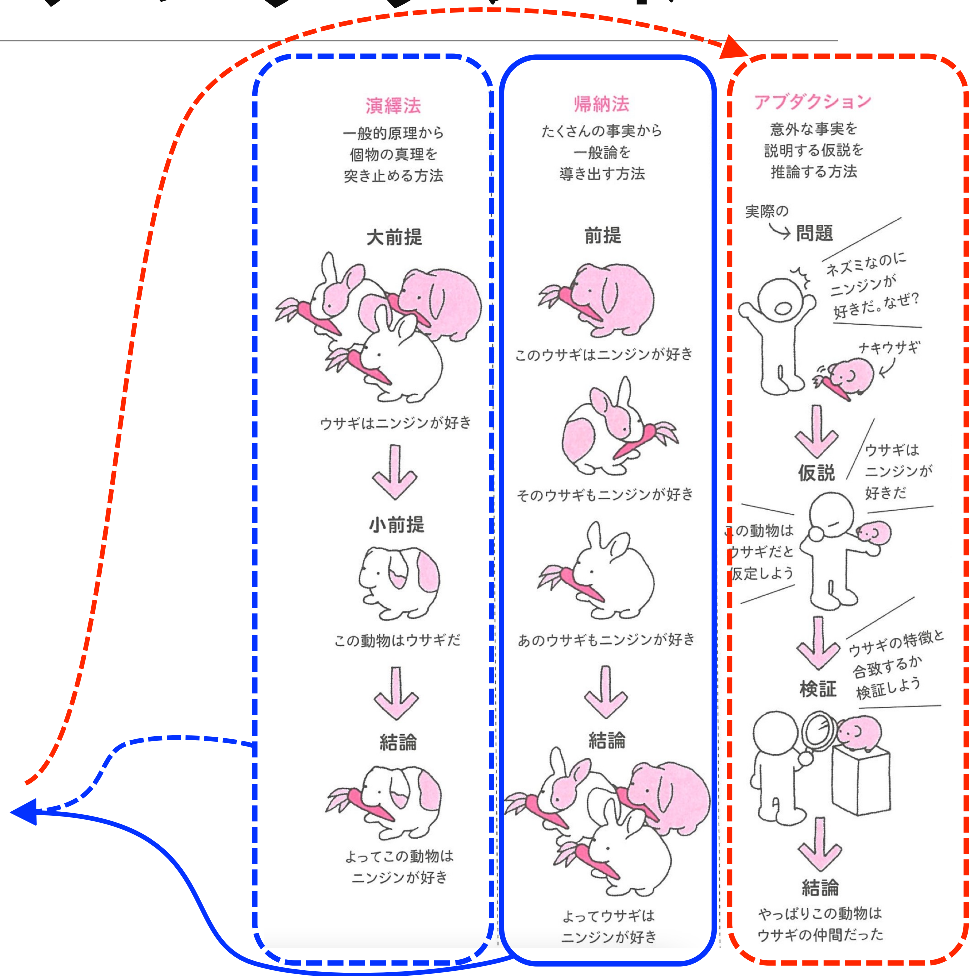
- ニュートン方程式など

計算科学　–　第3パラダイム

- 複雑現象の計算機シミュレーション，計算機実験など

データ科学　–　第4パラダイム

- 理論，実験，計算機実験，データマイニング



哲学用語図鑑・田中正人（著）・プレジデント社

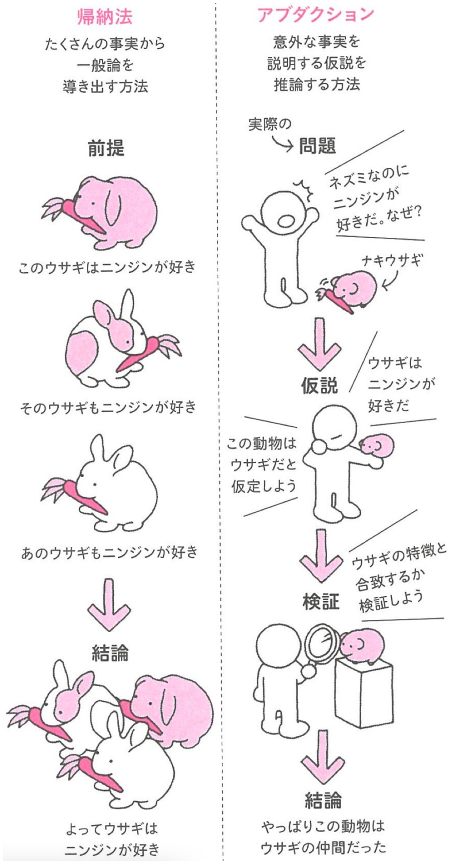
演繹法，帰納法，およびアブダクション

不確かさの定
量評価が必要

妥当性の検証
が必要

大前提が正しければ結
論が正しい（保証される）

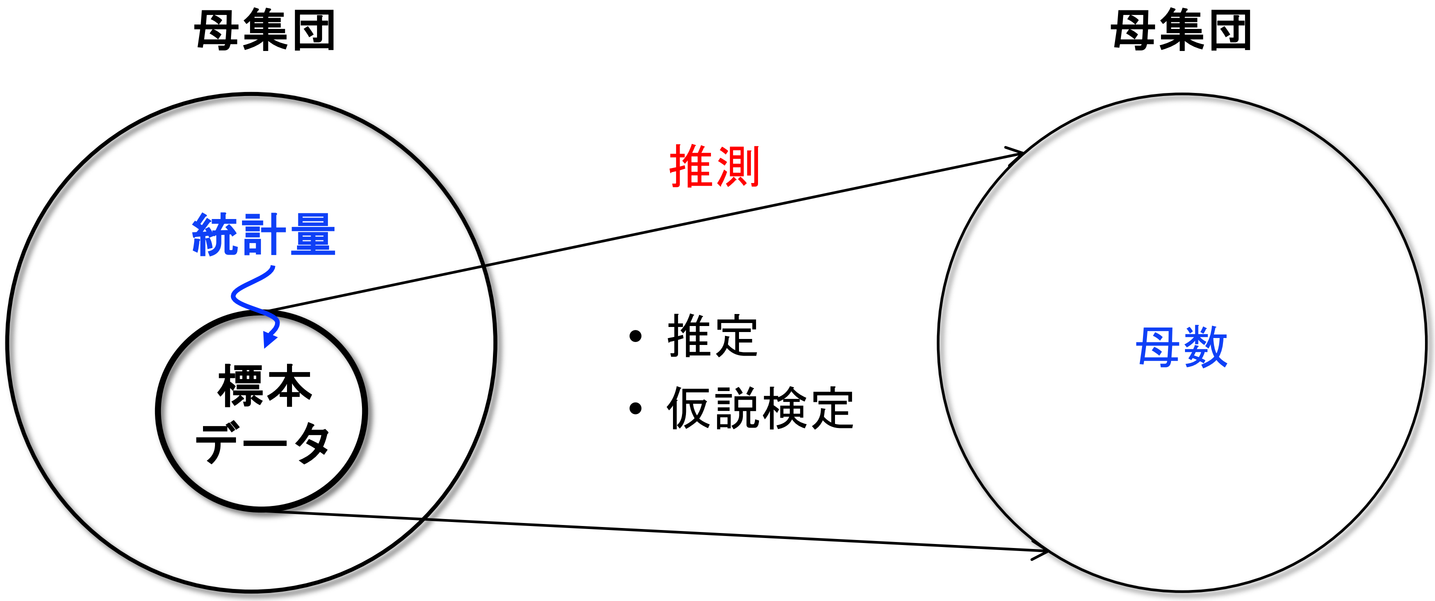
諸前提が正しいであっても
必ずしも結論が正しいと限
らない（保証されない）



哲学用語図鑑・田中正人（著）・プレシデント社

イントロ
その３：数理統計と機械学習の紹介

統計学の基礎:　統計的推測



- 推定
- 仮説検定

統計学的推測とは、標本データを解析して母数に関する結論を導くこと。

統計学の基礎：　統計的推測

- ✓ 母数：母集団全体の特性を表す値（例：母平均、母分散など）。母集団全体のデータに基づいていて、理想的には真の値を示す。

- ✓ 統計量：母集団の一部である標本のデータに基づいて計算される値（例：標本平均、標本分散）。統計量は、**母数を推定**するために使用される。

  - ➢ 推定: すべてのデータをできないため、標本を使用して母数を推定する。

  - ➢ 不確実性: 標本に基づく統計量は、標本の選び方や偶然の影響により、異なる値を有り得るため、統計量には不確実性が伴う。母数は母集団全体に基づいき、（理想的な）確定した真の値。

統計学の基礎：　数学的推論

生データ

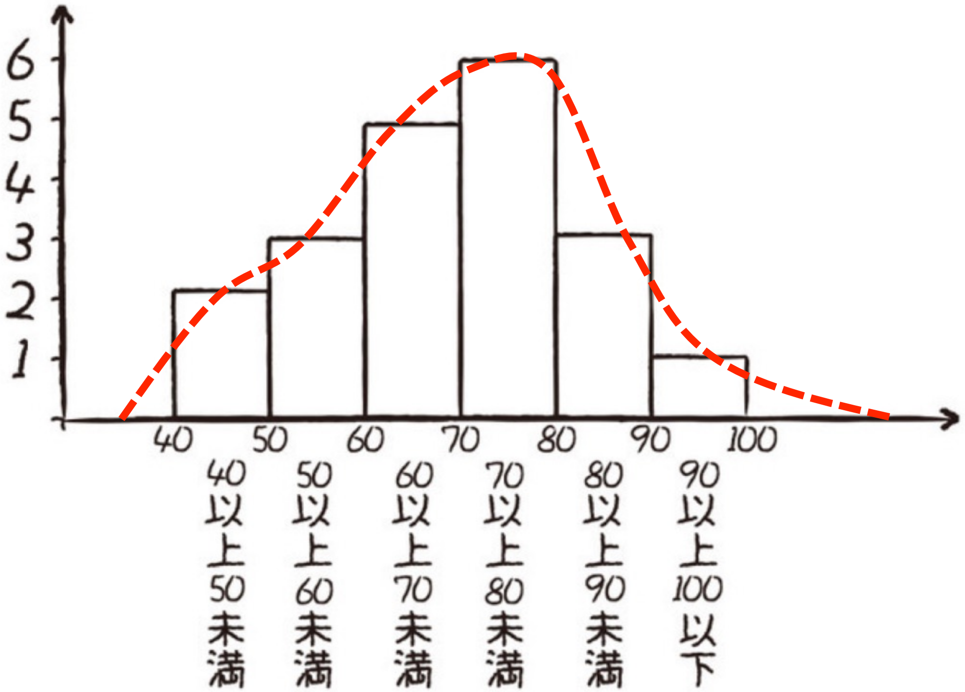
2人 43、47、 3人 52、52、54、 5人 61、67、67、68、69、

6人 70、71、71、73、76、78、 3人 82、84、84、 1人 91

整理された
データ

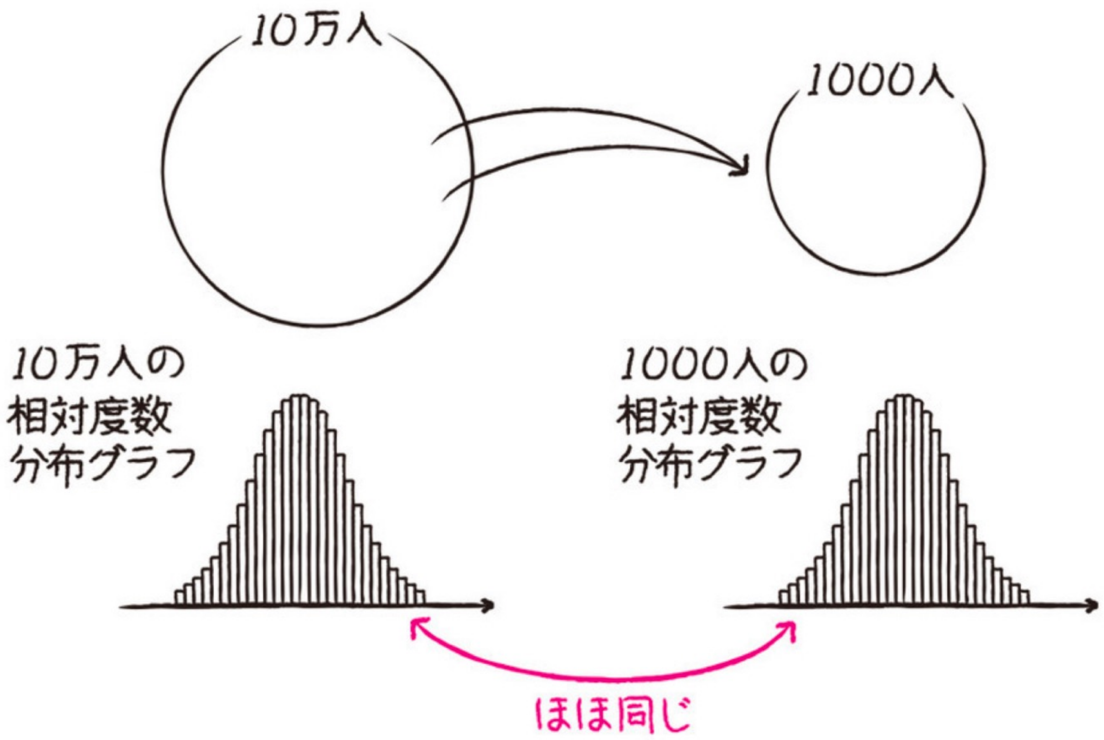
階級	階級値	度数
40 以上　50 未満	45	2
50 以上　60 未満	55	3
60 以上　70 未満	65	5
70 以上　80 未満	75	6
80 以上　90 未満	85	3
90 以上 100 以下	95	1

可視化されたデータ

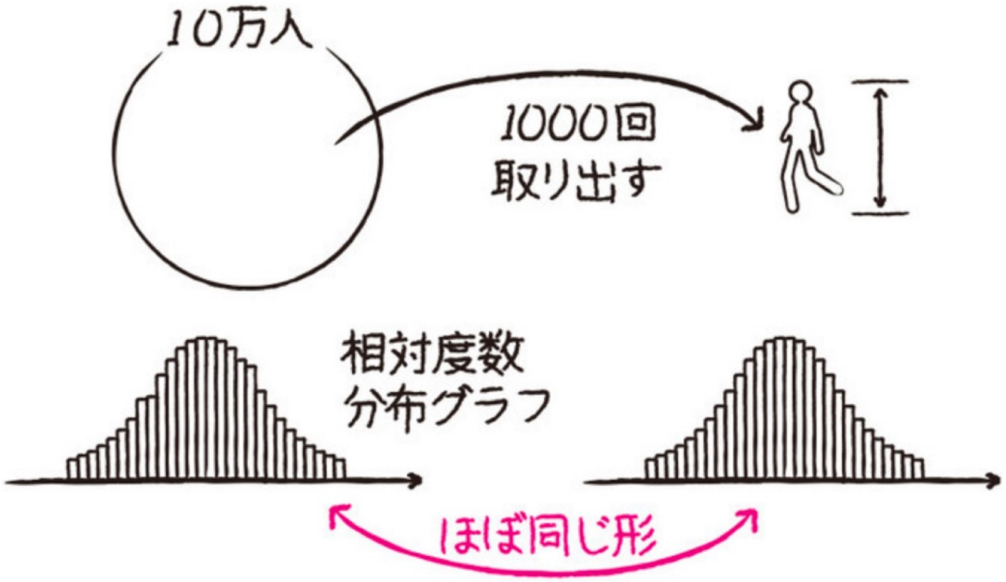


統計学の基礎：　データサンプリング

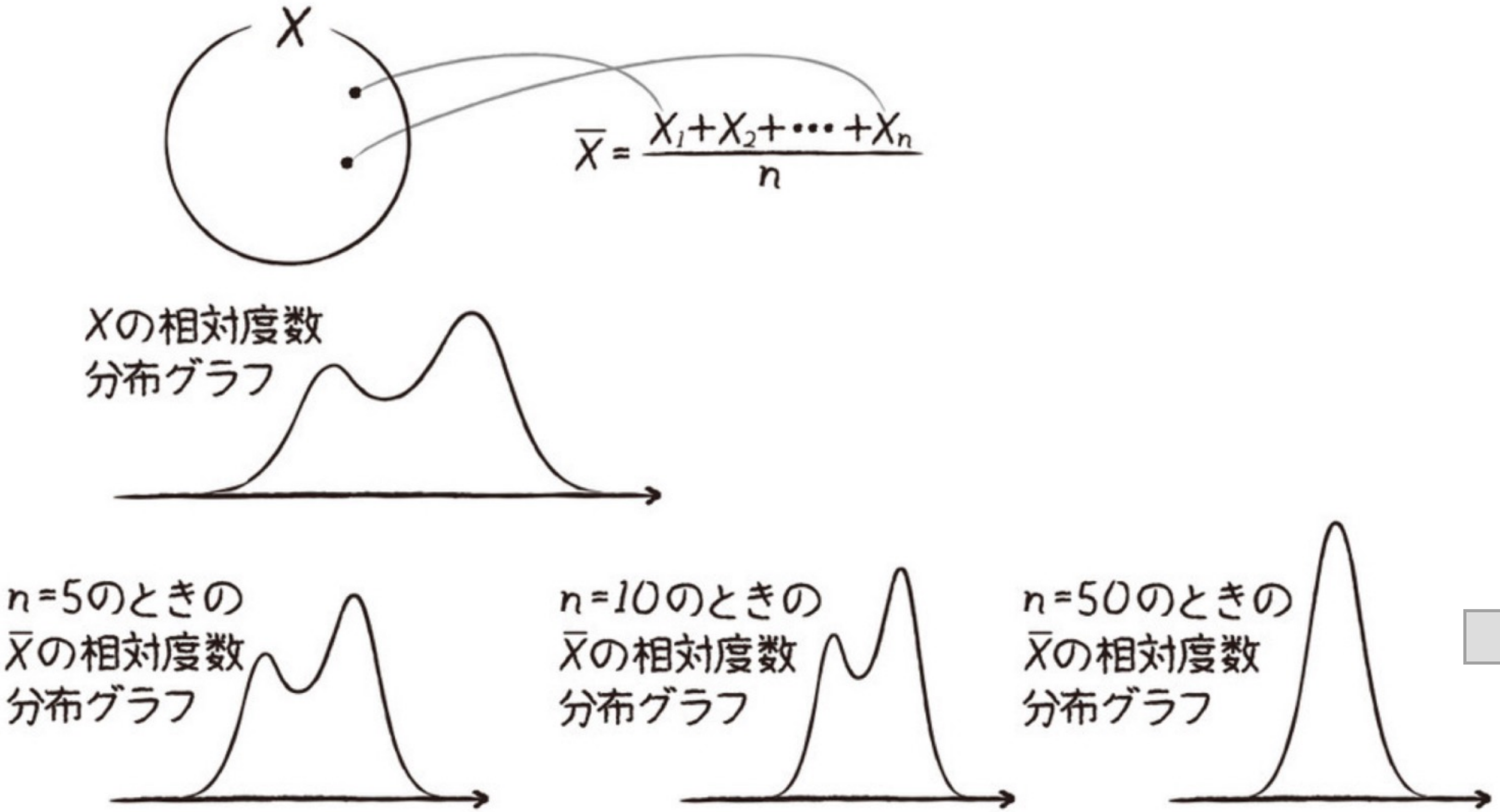
非復元抽出



復元抽出



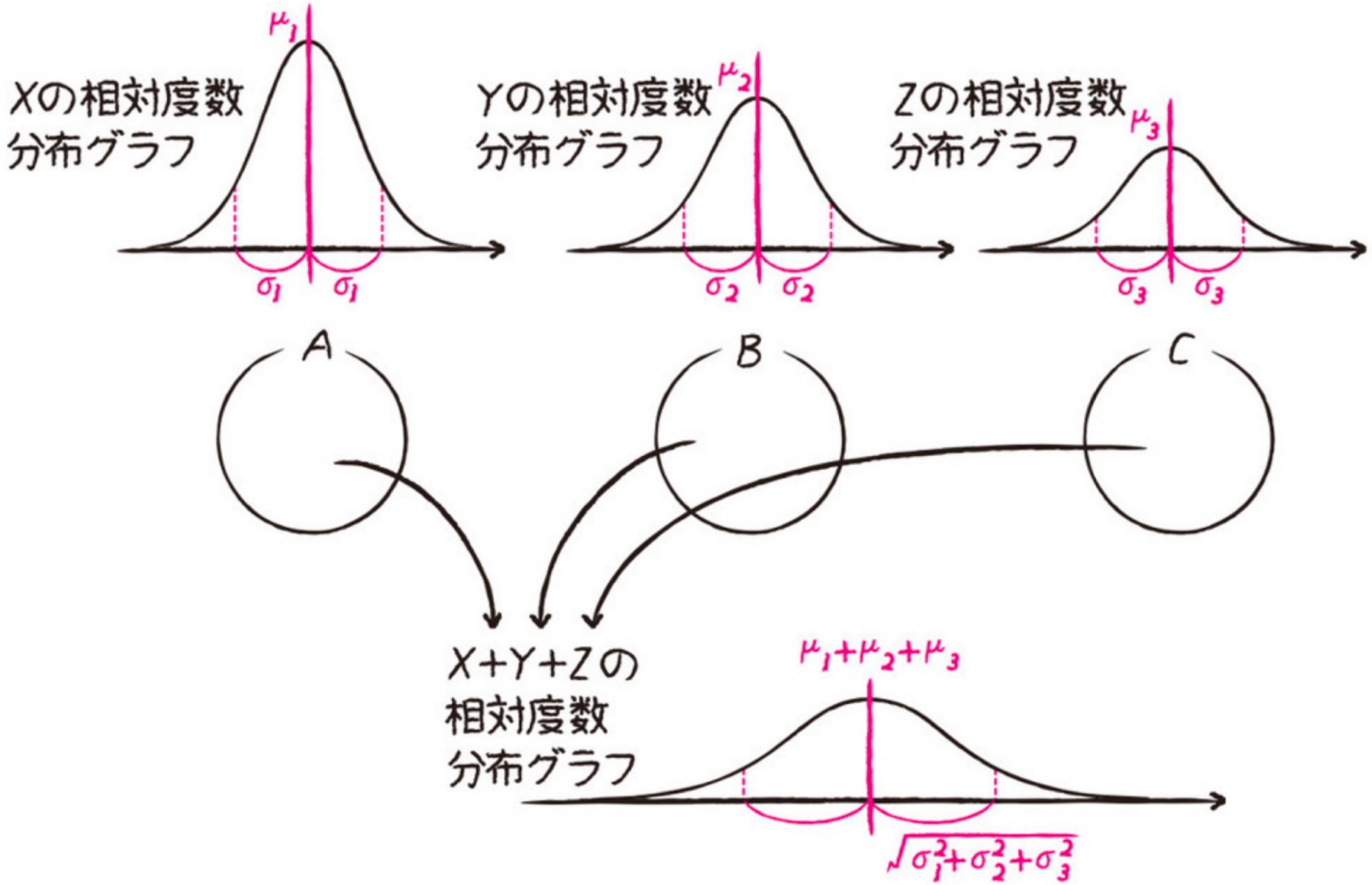
統計学の基礎：　中心極限定理



$n \to \infty$
のとき

正規分布

統計学の基礎：　正規分布



Xの相対度数
分布グラフ
μ_1
σ_1 σ_1
Yの相対度数
分布グラフ
μ_2
σ_2 σ_2
Zの相対度数
分布グラフ
μ_3
σ_3 σ_3
A
B
C
X+Y+Zの
相対度数
分布グラフ
$\mu_1+\mu_2+\mu_3$
$\sqrt{\sigma_1^2+\sigma_2^2+\sigma_3^2}$

統計学の基礎：　推定

不偏推定量による推定

最尤法による推定

統計学の基礎：　推定

与えられたデータは現実に出現している

↓

考えられる数理モデルの集合を検討する

↓

与えられたデータを再現する数理モデルは
比較的高い確率が割り振られている

最尤法による推定

統計学の基礎：コイン投げ

なぜそのように計算できるか?

- 本当なのか? どれぐらい信頼できるか?
- この計算はデータ解析なのか?
- 一種のデータ解析だとしたら、
  このデータ解析が何を基にしているか?
- 背景にはどのような仮説があるか？

100 回: 70 表 & 30 裏

$$p(\text{表}) = \frac{70}{100} = 0.7$$

統計学の基礎:　コイン投げ

100 回: 70 表 & 30 裏

仮定

$$p(\text{表}) = \theta$$

尤度

$$\mathcal{L}(\theta) = {}_{100}C_{70}\theta^{70}(1-\theta)^{30}$$

考え方：

与えられたデータを再現する数理モデルは
比較的高い確率が割り振られている

統計学の基礎：　コイン投げ

100 回: 70 表 & 30 裏

仮定

$$p(\text{表}) = \theta$$

尤度

$$\mathcal{L}(\theta) = {}_{100}C_{70}\theta^{70}(1-\theta)^{30}$$

$$\begin{aligned}\theta &\triangleq \underset{\theta\in[0,1]}{\arg\max} \log p(\mathcal{D}|\theta)\\ &\triangleq \underset{\theta\in[0,1]}{\arg\max} \log \mathcal{L}(\theta)\\ &\triangleq \underset{\theta\in[0,1]}{\arg\max} \log {}_{100}C_{70}\theta^{70}(1-\theta)^{30}\end{aligned}$$

統計学の基礎:　コイン投げ

100 回: 70 表 & 30 裏

仮定

$$p(\text{表}) = \theta$$

尤度

$$\mathcal{L}(\theta) = {}_{100}C_{70}\theta^{70}(1-\theta)^{30}$$

$$\frac{d}{d\theta}\log\mathcal{L}(\theta) = \frac{70}{\theta} - \frac{30}{1-\theta} = 0$$

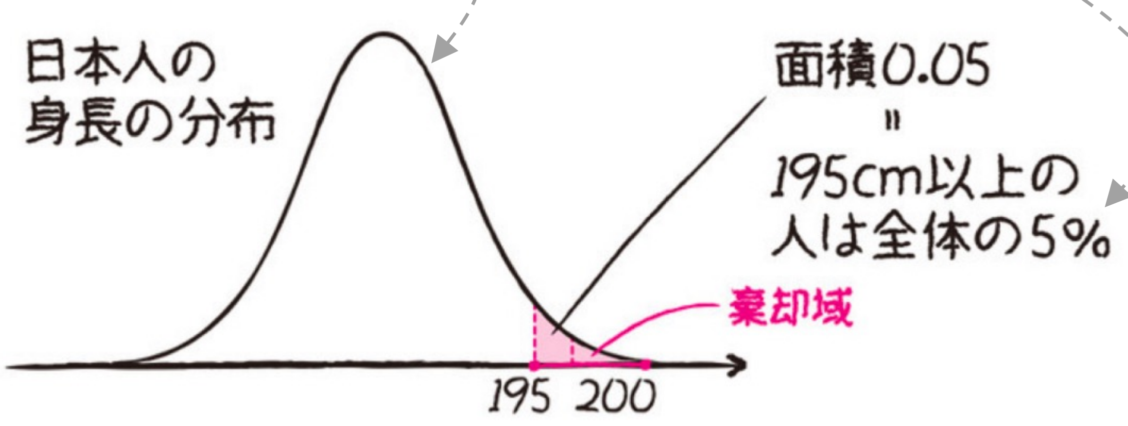
$$\theta = \frac{70}{100} = 0.7$$

統計学の基礎：　仮説検定

状況：街（日本）で向こうを歩いている身長2mの人を見えました.

疑問：歩いてきた人は外国人ではないか？

理由：日本人だとすれば，大きすぎる. 常識から考えて，あり得ないと思う.



統計学の基礎： 仮説検定

「むこうを歩いている人が ← H_0を仮定 帰無仮説

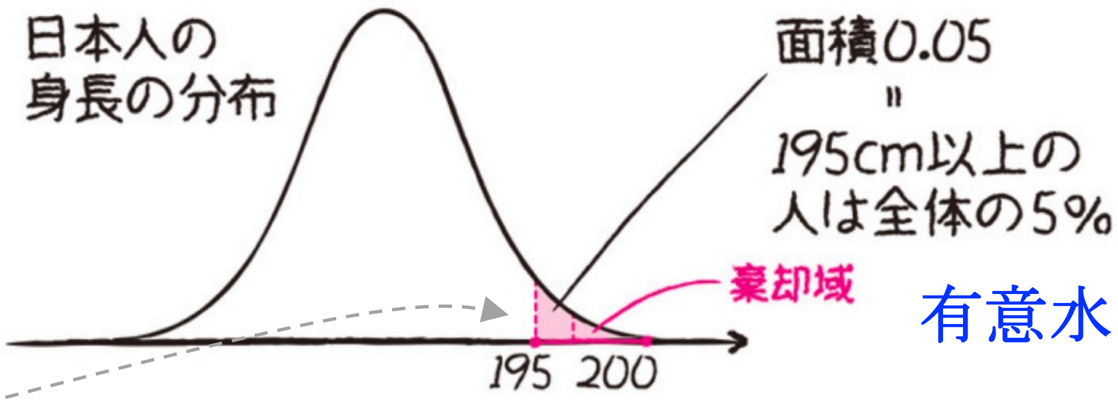
日本人であるとすれば、

大きすぎるなあ。 ← 起こる確率が5％以下の

ありえないよ。 できごとが起きた

日本人じゃないな。 ← H_0棄却

外国人じゃないの」 ← H_1採択 対立仮説



有意水準

事実 ＼ 判断結果	帰無仮説 H_0 を採択	帰無仮説 H_0 を棄却
帰無仮説 H_0 が本当	正しい判断	第 1 種の誤り
帰無仮説 H_0 が嘘	第 2 種の誤り	正しい判断

統計学の基礎:　仮説検定と背理法

状況:街(日本)で向こうを歩いている身長2mの人を見えました.

疑問:歩いてきた人は外国人ではないか？

理由:日本人だとすれば, 大きすぎる. 常識から考えて, あり得ないと思う.

背理法:
① 日本人だと仮定する.
② 日本人なら身長が“それほど”背が高くない.
③ 見えた人の身長が2m
④ ③が(2)と矛盾しているので①が正しくない
(矛盾しない場合は①が正しくないことはまだ言えない)

対立仮説 H_1 ← 証明したいこと

帰無仮説 H_0 を仮定する ← 反対のことを仮定する

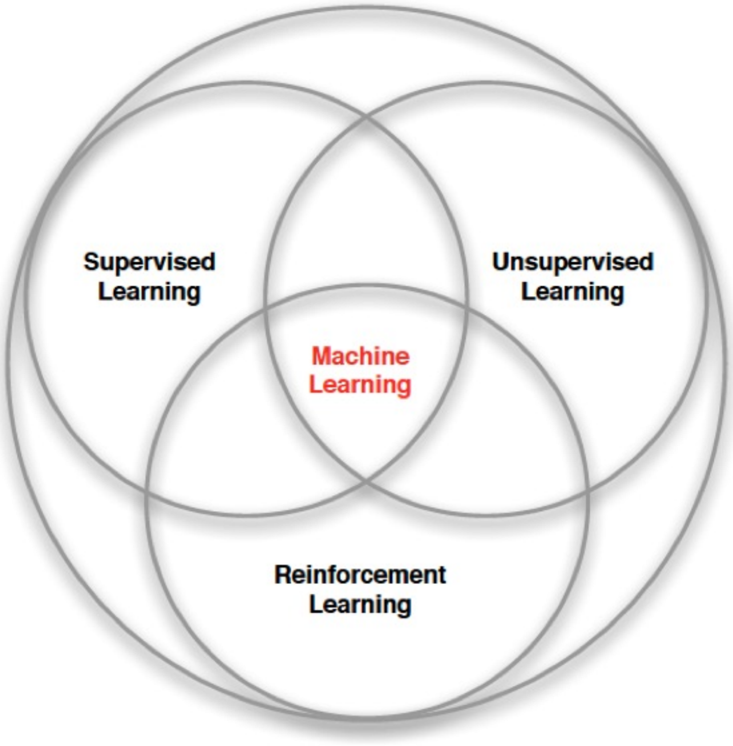
このもとで確率を計算

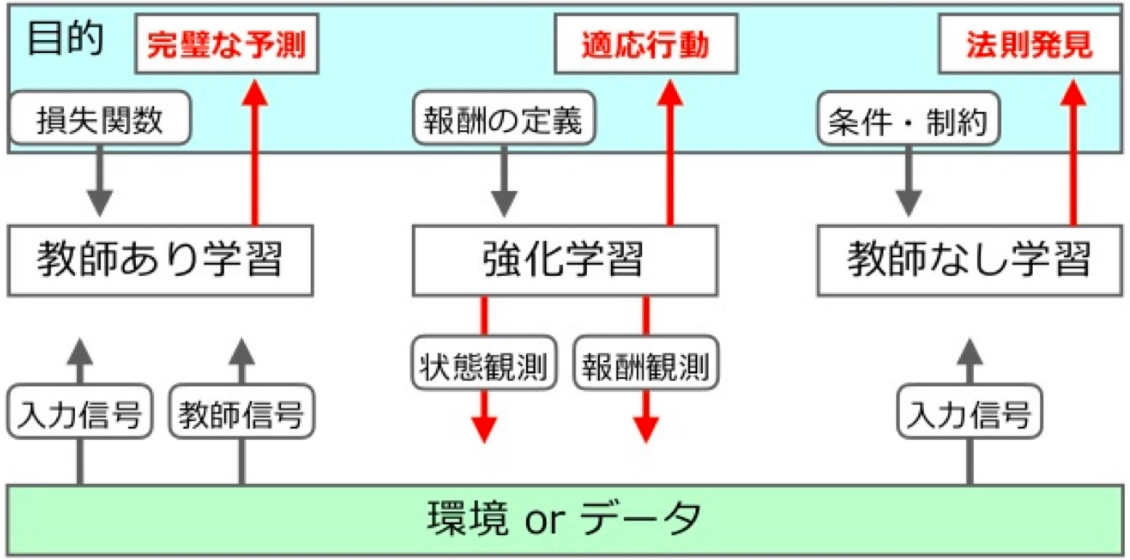
ありえそうもないことが起きている

矛盾を導く

対立仮説 H_1 を採択 ← 証明完了

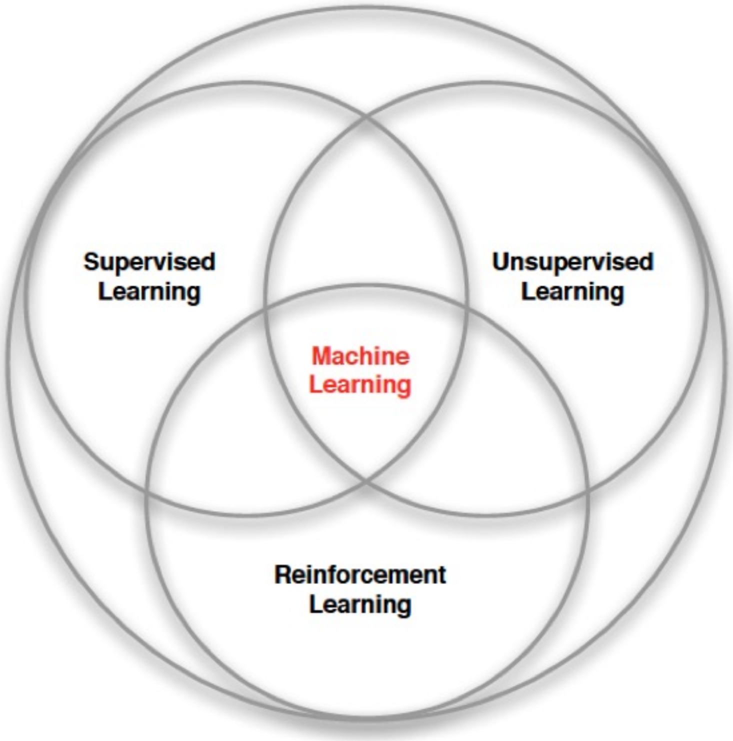
機械学習

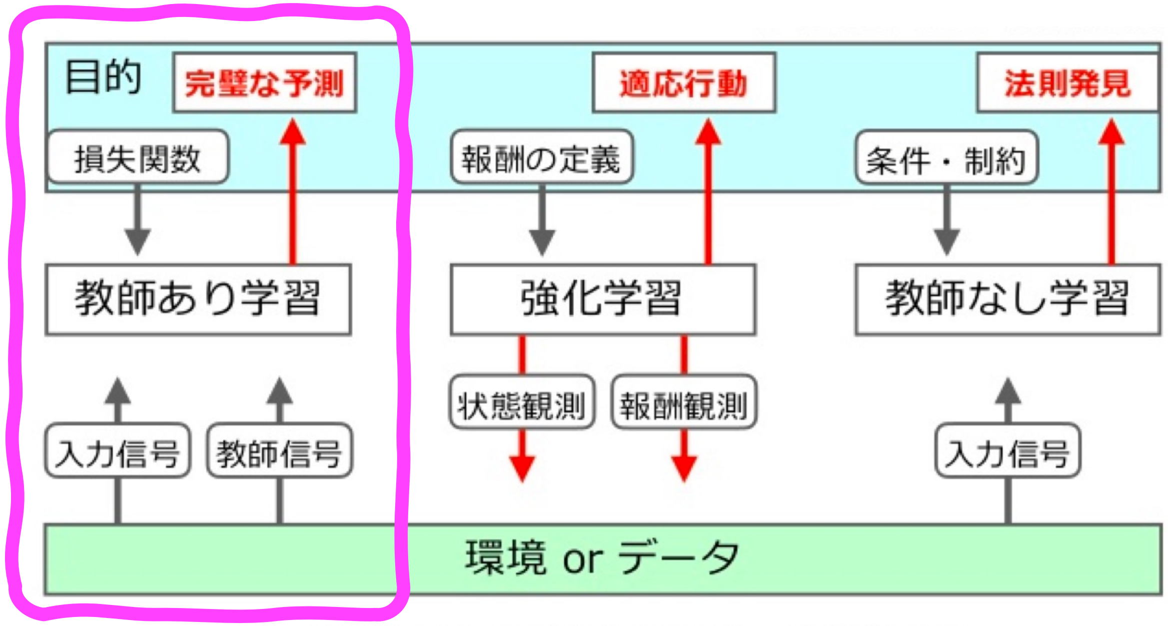




データ所与である前提と異なり，強化学習は
環境を探索して主体的にデータを獲得しつつ行動方策を最適化

機械学習

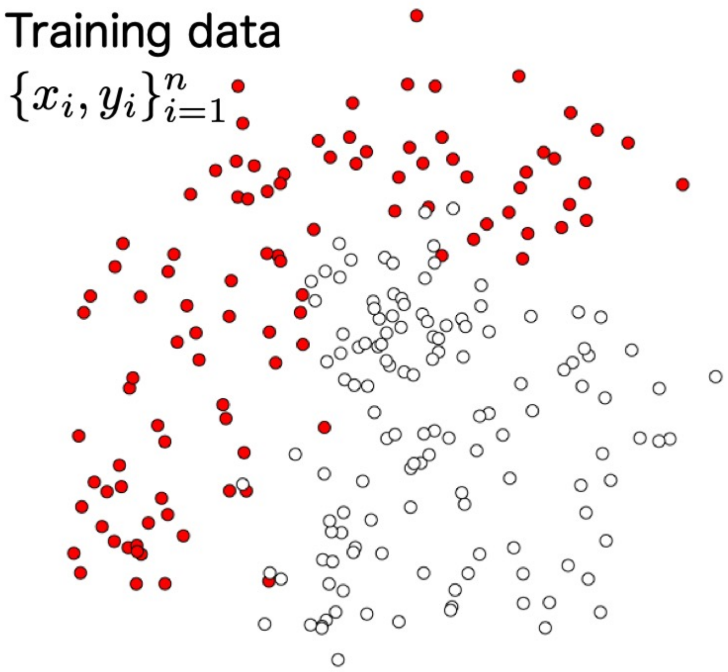

Supervised Learning
Unsupervised Learning
Machine Learning
Reinforcement Learning

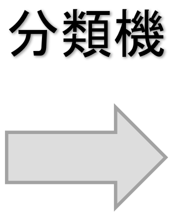

目的
完璧な予測
適応行動
法則発見
損失関数
報酬の定義
条件・制約
教師あり学習
強化学習
教師なし学習
状態観測
報酬観測
入力信号
教師信号
入力信号
環境 or データ
データ所与である前提と異なり，強化学習は
環境を探索して主体的にデータを獲得しつつ行動方策を最適化

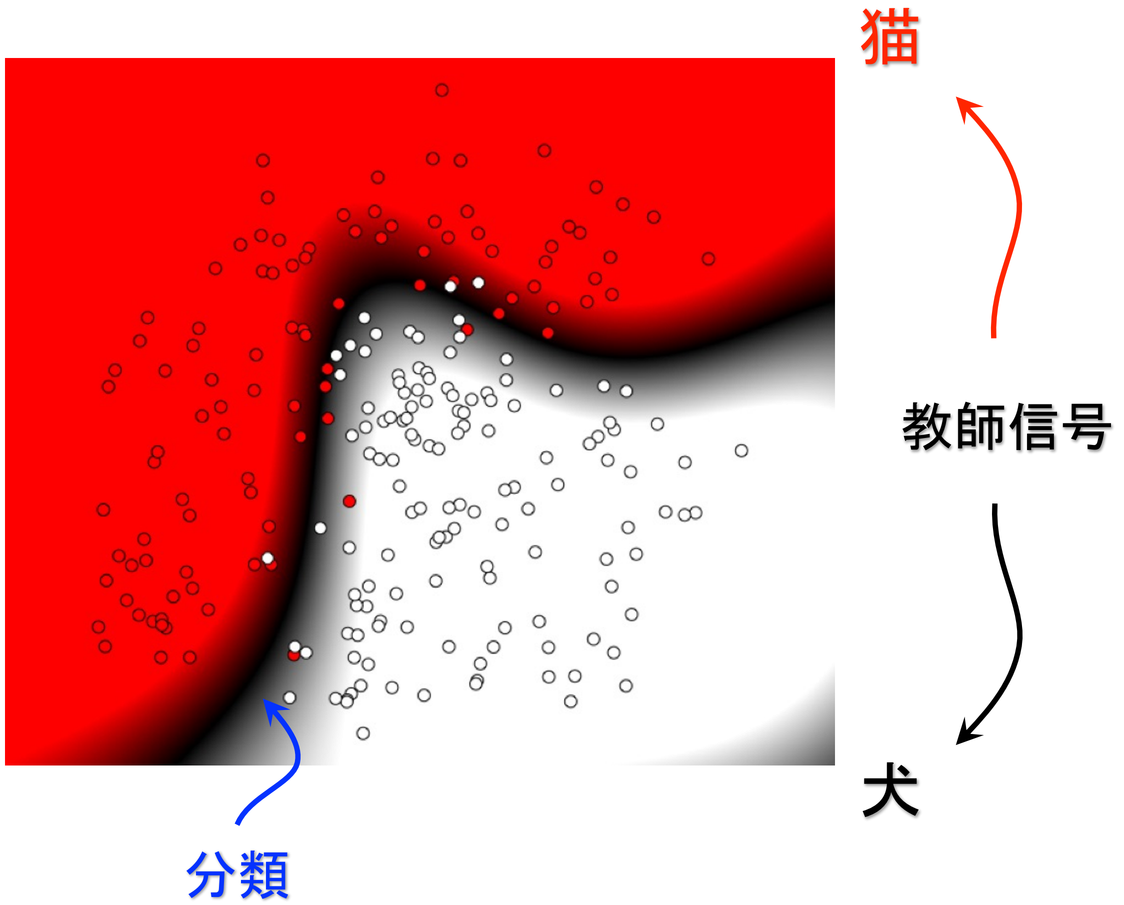
教師あり学習

Training data
$\{x_i, y_i\}_{i=1}^{n}$
データ変換
(属性，特徴量など)
Data source: Kaggle
猫
教師信号
犬

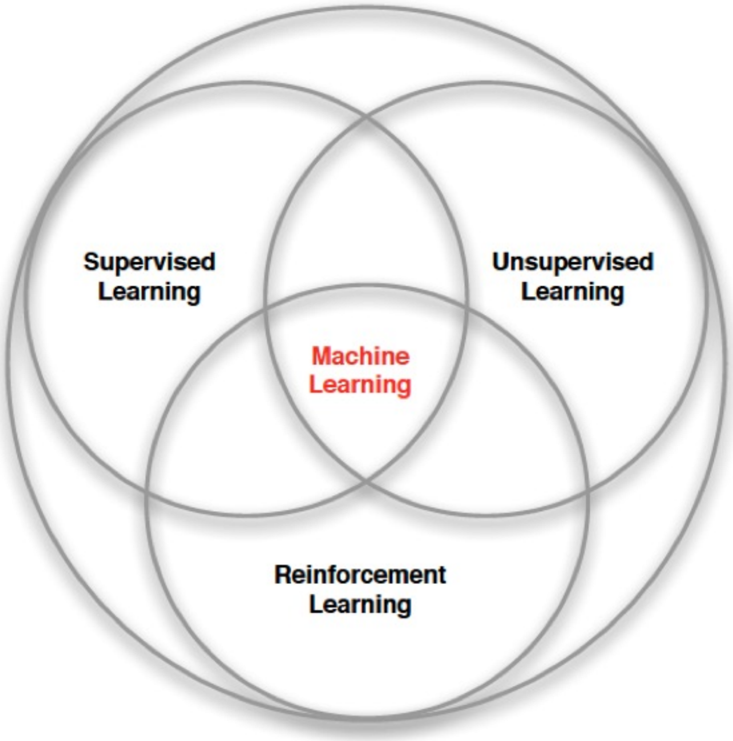
教師あり学習



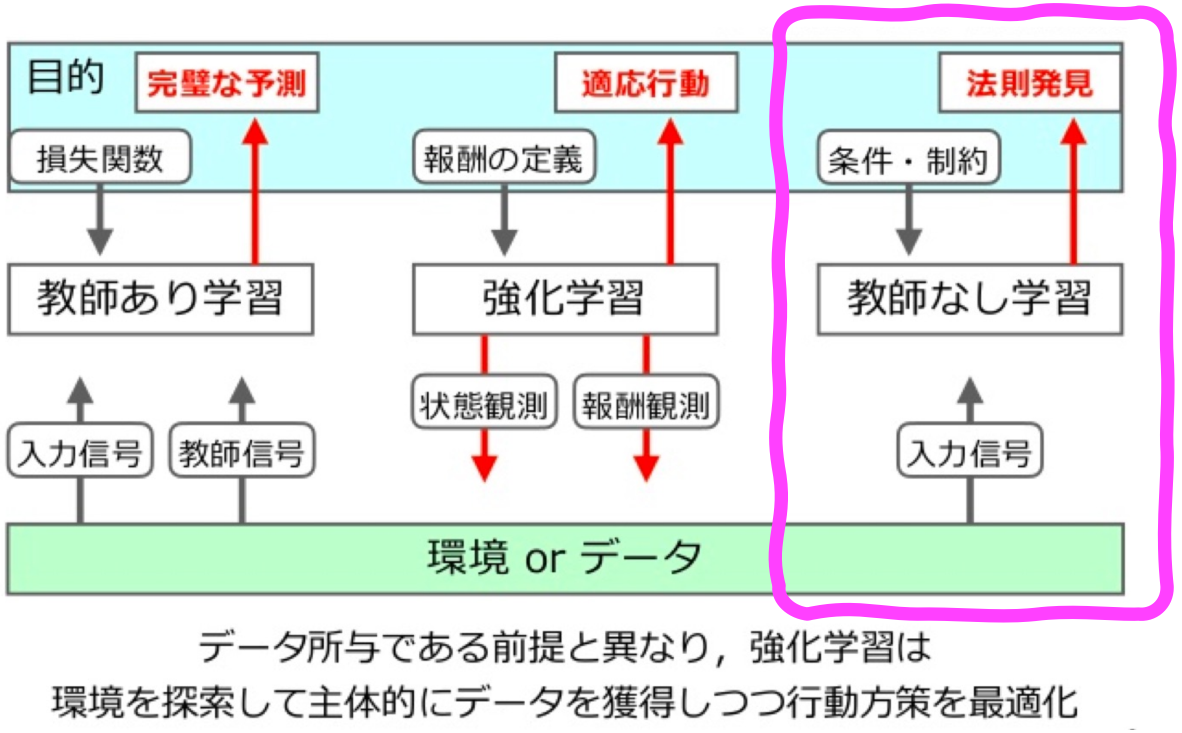




機械学習

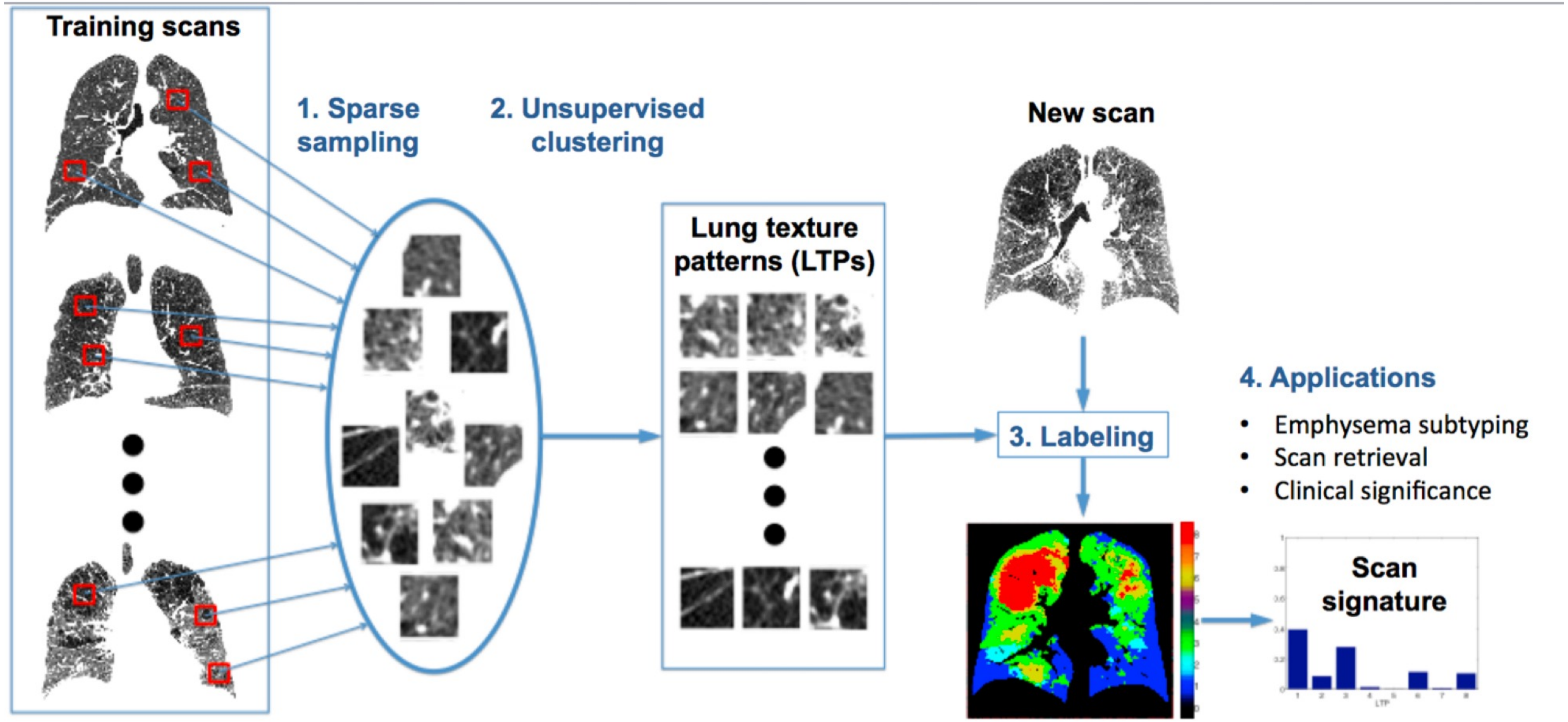


Supervised Learning
Unsupervised Learning
Machine Learning
Reinforcement Learning

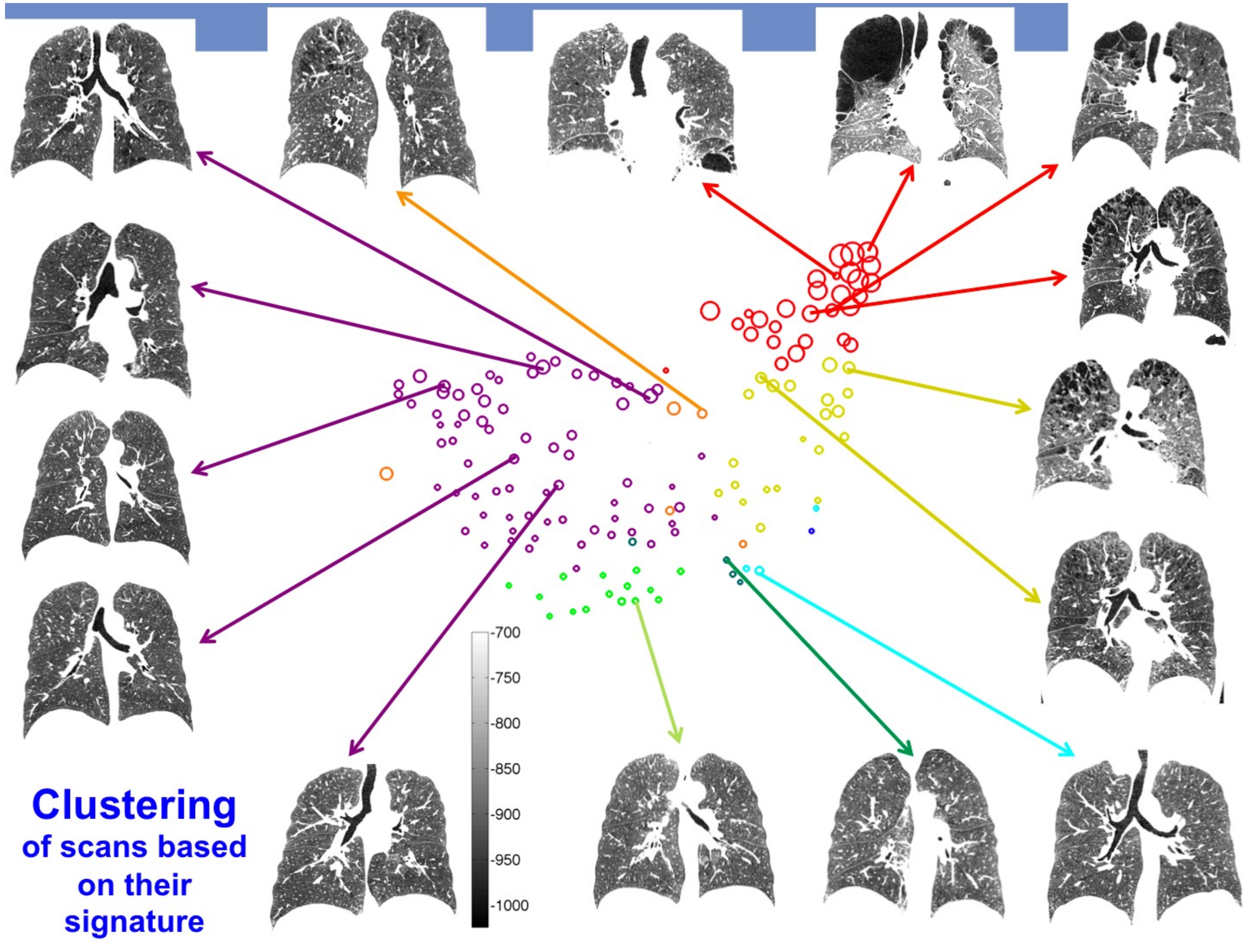


目的
完璧な予測
適応行動
法則発見
損失関数
報酬の定義
条件・制約
教師あり学習
強化学習
教師なし学習
状態観測
報酬観測
入力信号
教師信号
入力信号
環境 or データ
データ所与である前提と異なり，強化学習は
環境を探索して主体的にデータを獲得しつつ行動方策を最適化

教師なし学習

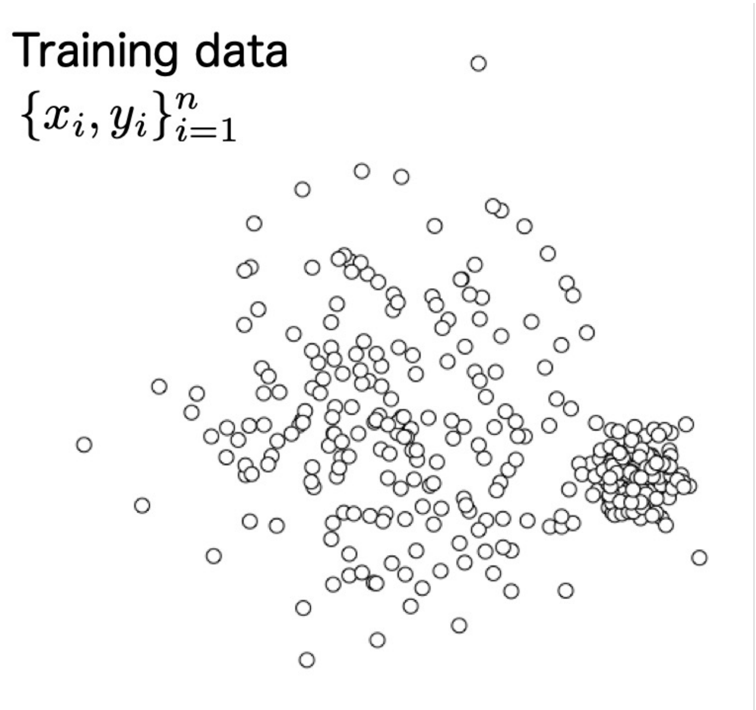

Training scans
1. Sparse sampling
2. Unsupervised clustering
Lung texture patterns (LTPs)
New scan
3. Labeling
4. Applications
• Emphysema subtyping
• Scan retrieval
• Clinical significance
Scan signature

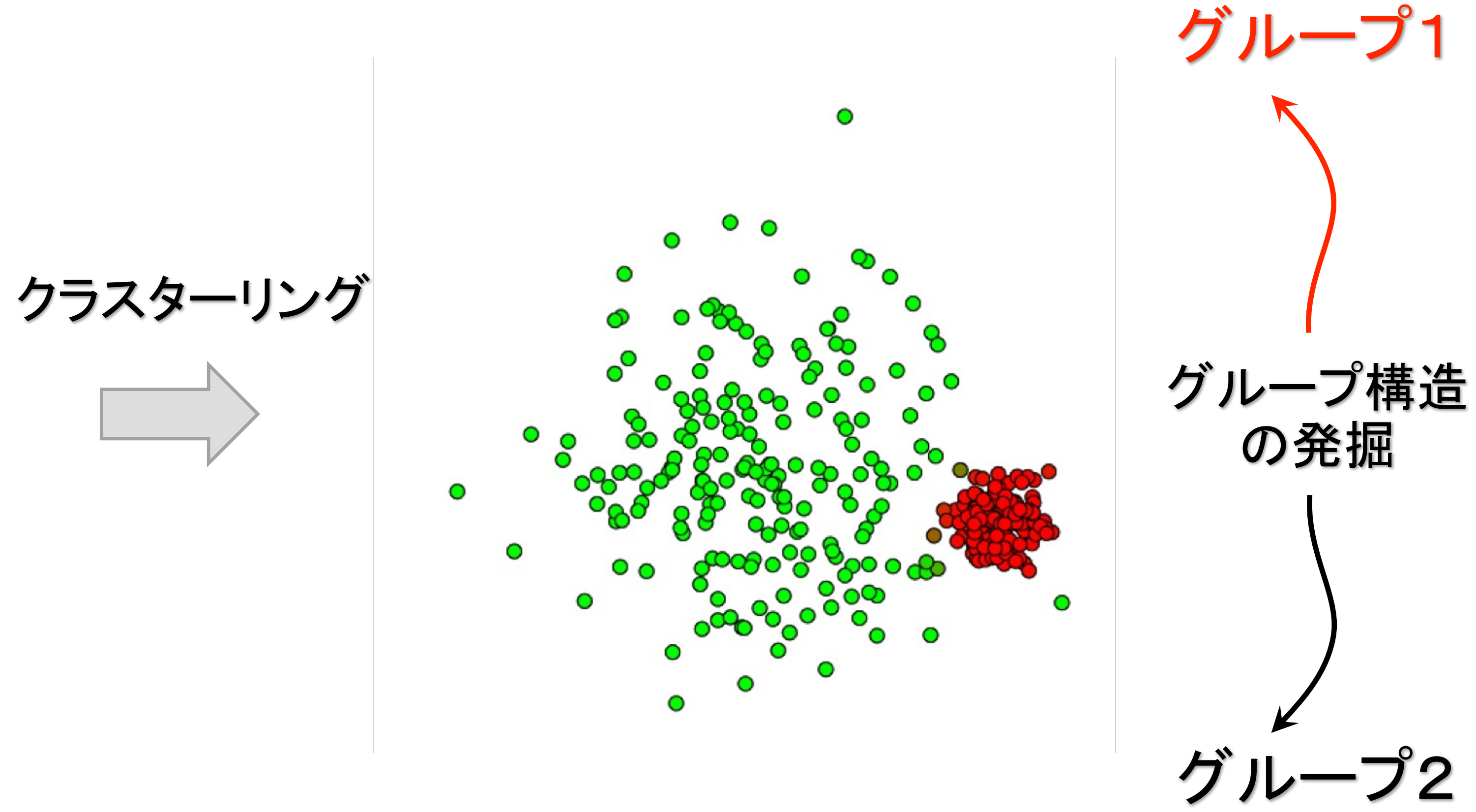
教師なし学習



-700
-750
-800
-850
-900
-950
-1000
Clustering
of scans based
on their
signature

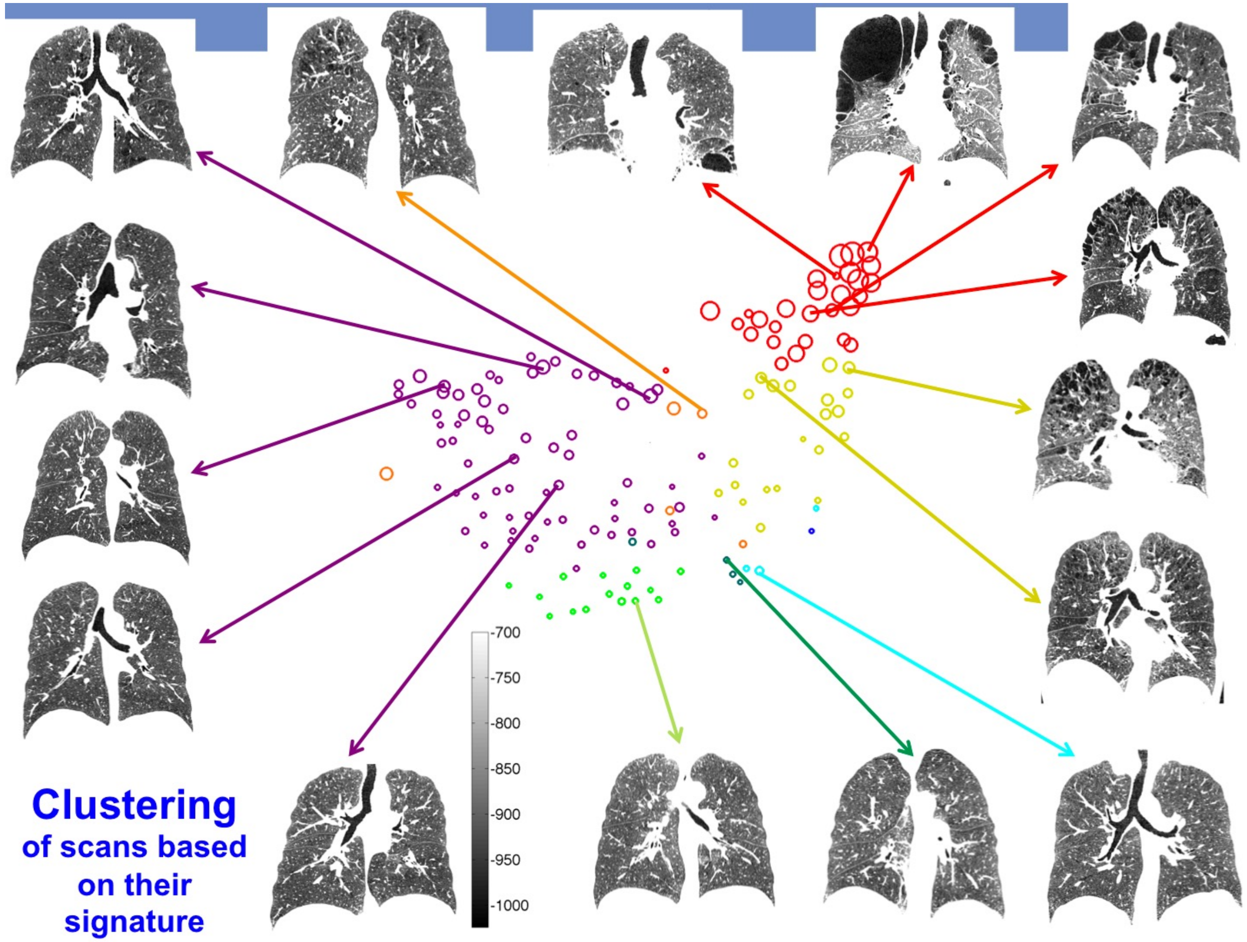
教師なし学習





入力信号だけ、教師信号なし

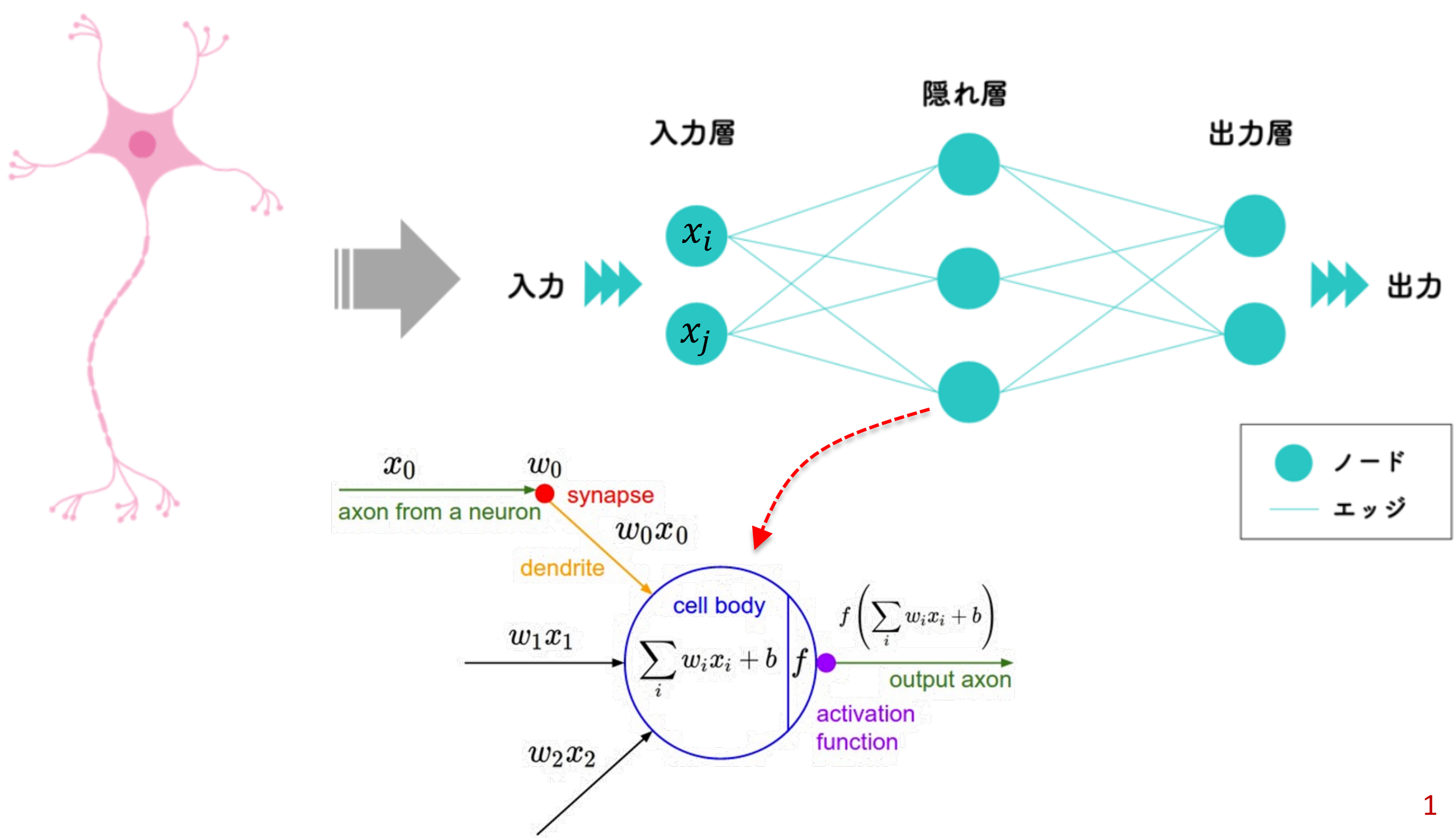
教師なし学習



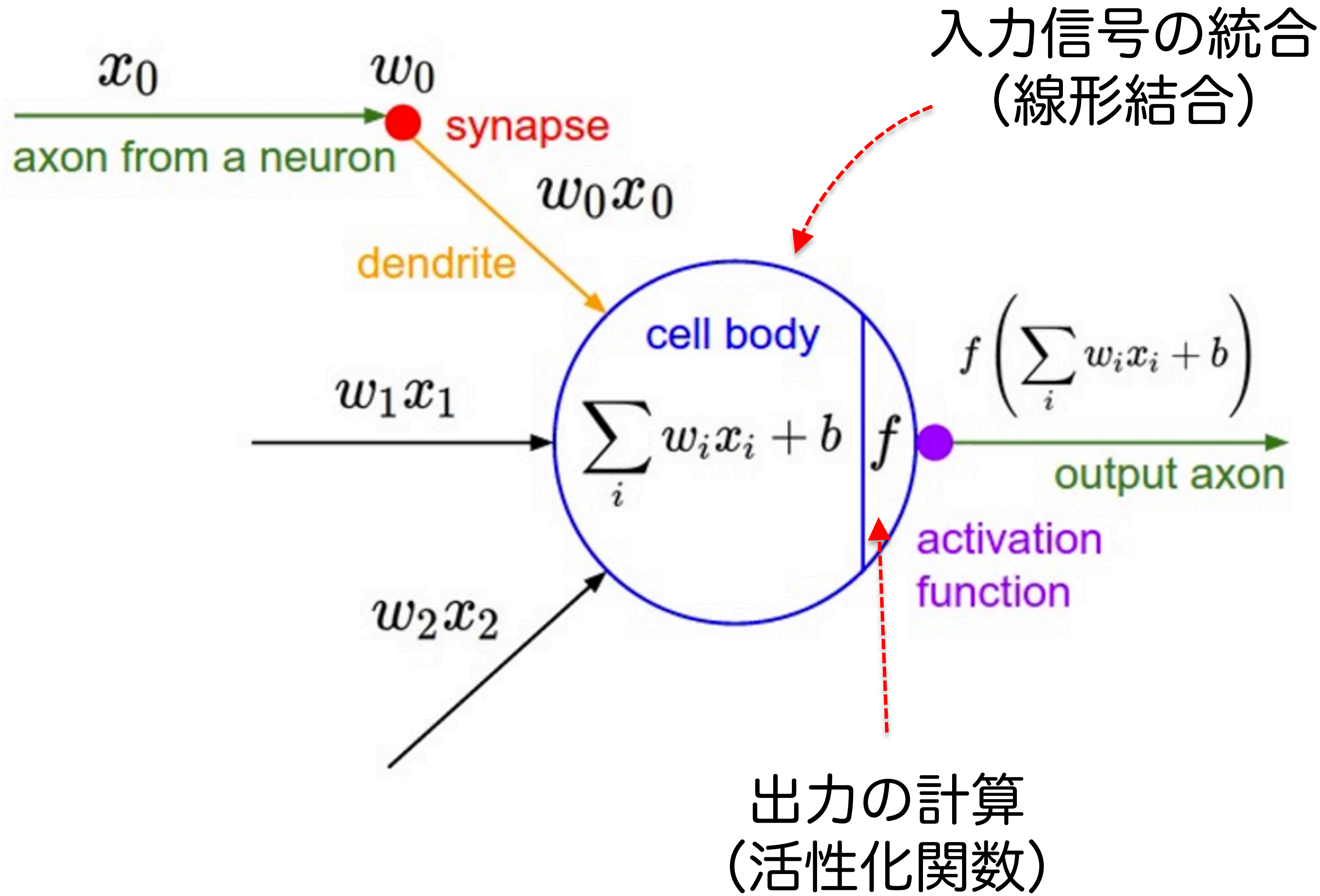
-700
-750
-800
-850
-900
-950
-1000
Clustering
of scans based
on their
signature

ニューラルネットワーク構造

神経細胞(ニューロン)　　　ニューラルネットワーク

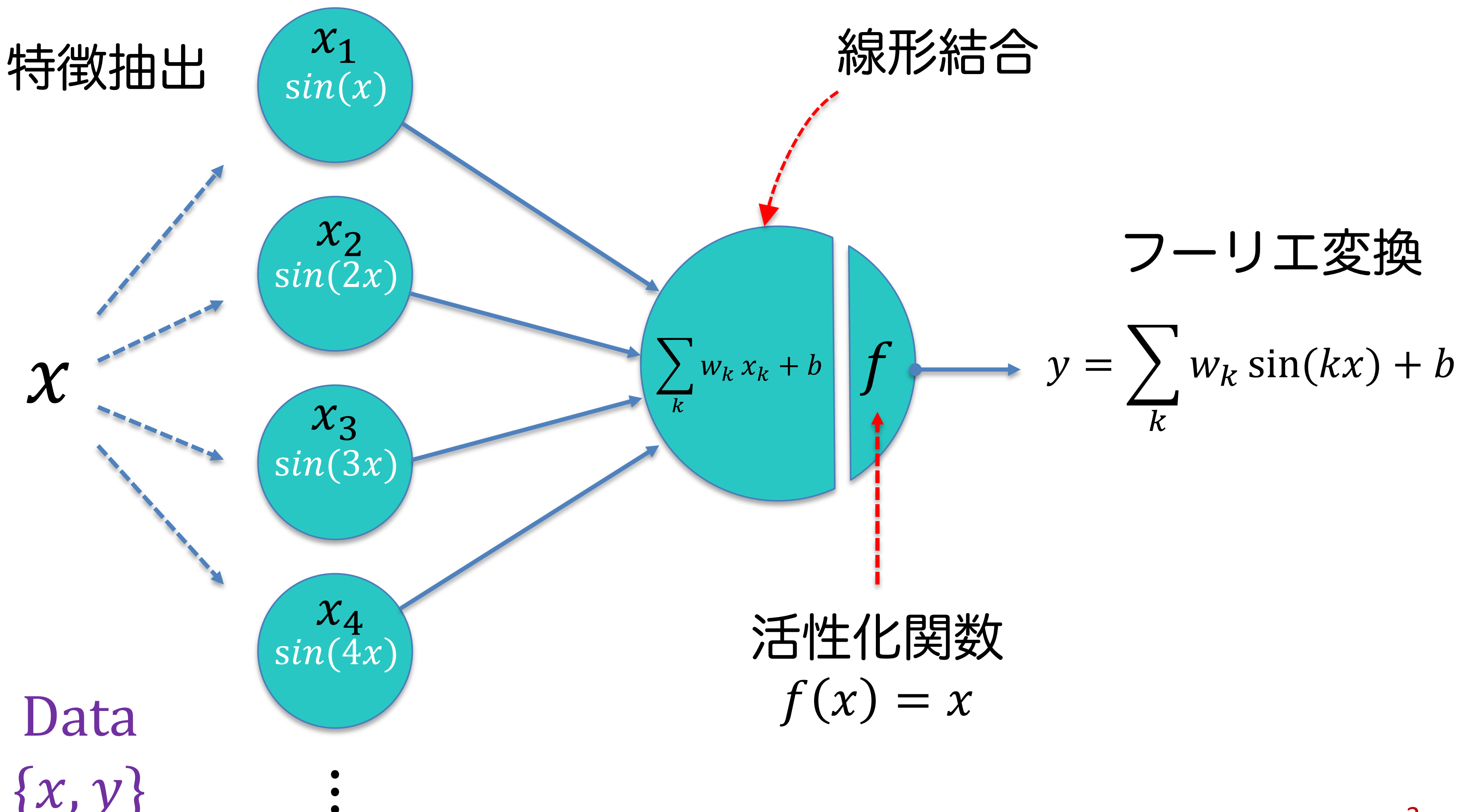


ニューラルノード

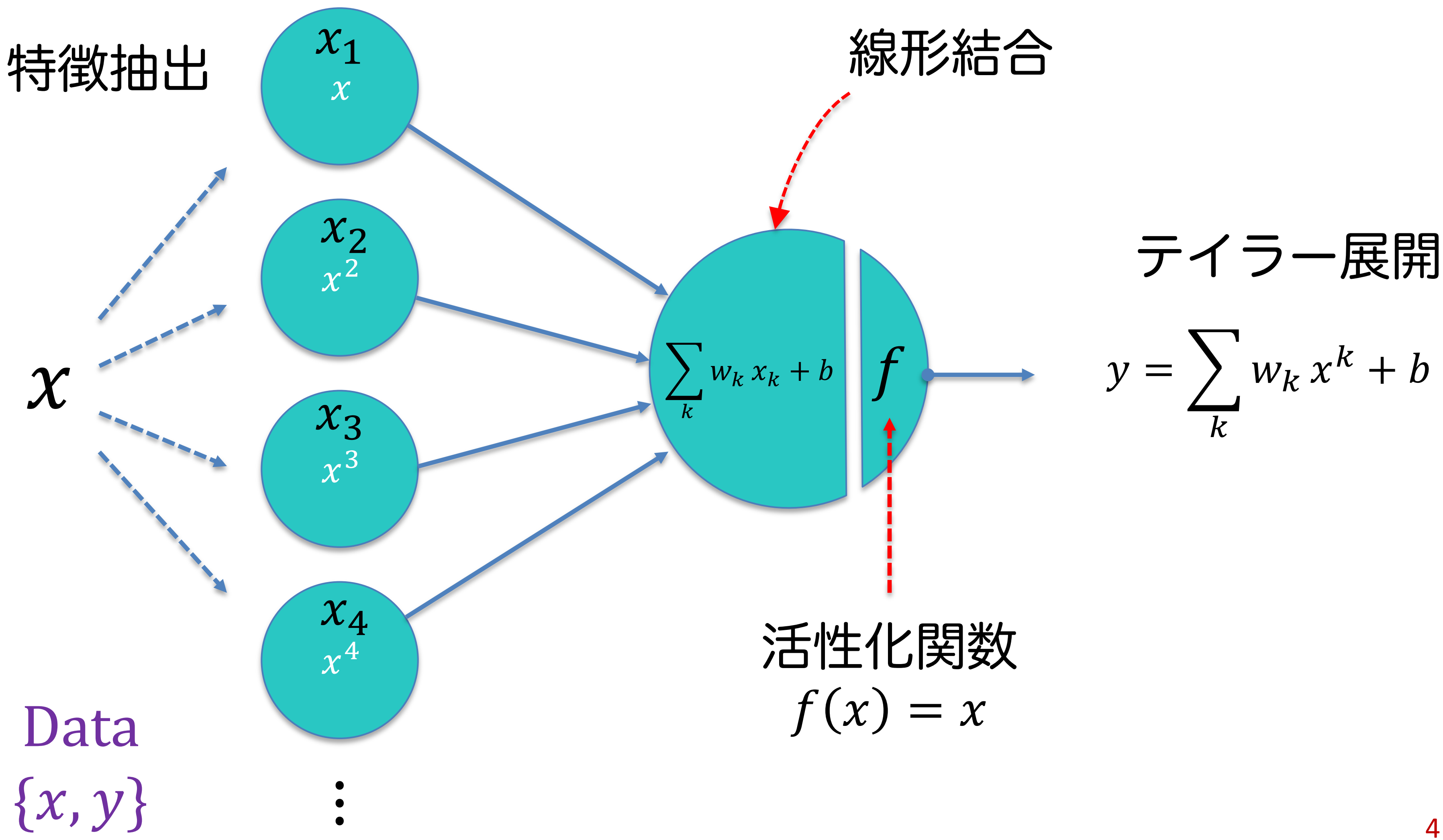


x_0
w_0
synapse
axon from a neuron
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
f
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation function
$w_2 x_2$
入力信号の統合
（線形結合）
出力の計算
（活性化関数）

ニューラルネットの事例 ①

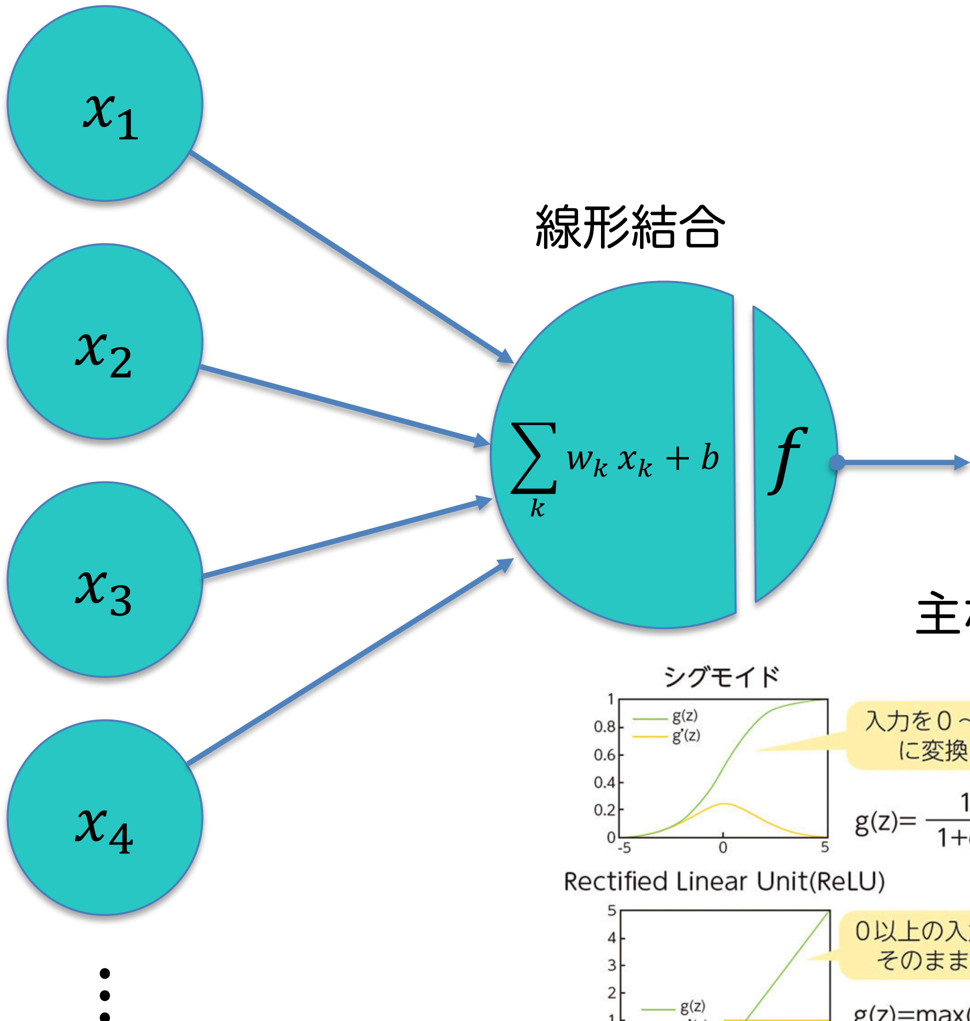


ニューラルネットの事例 ②



活性化関数

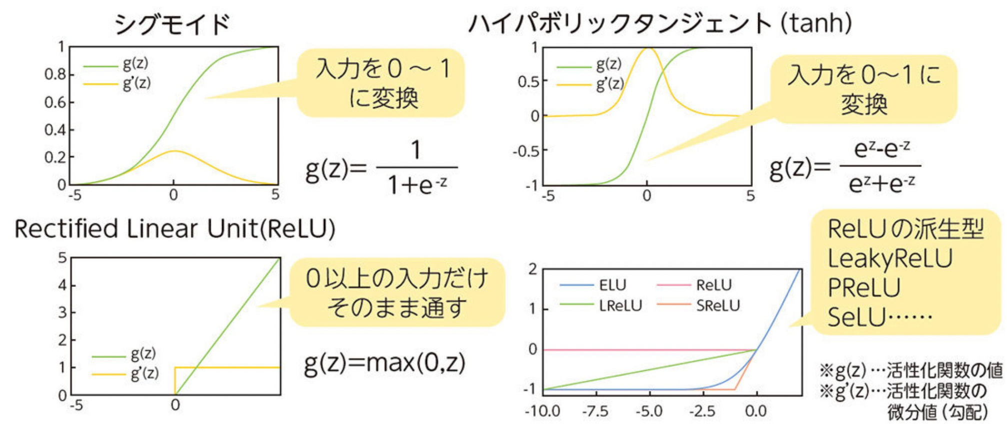
入力



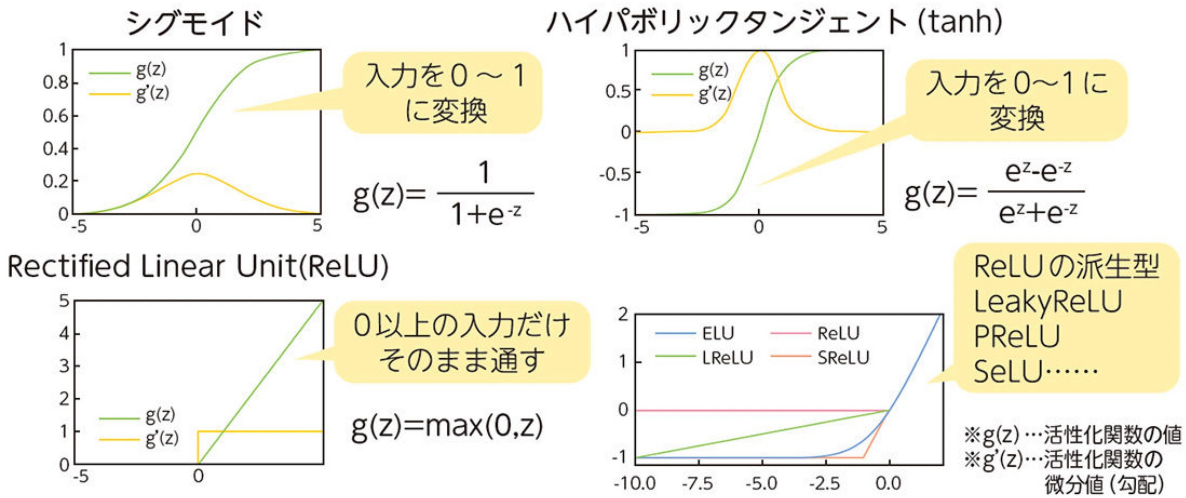
出力

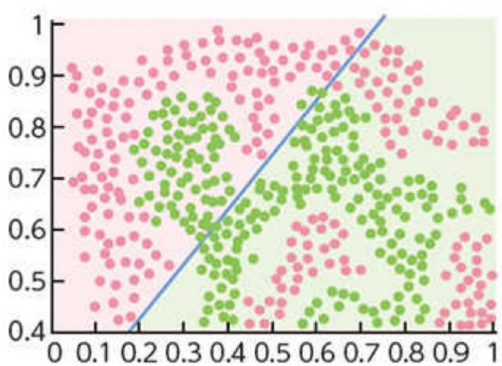
$$y = f\left(\sum_k w_k x^k + b\right)$$

主な活性化関数

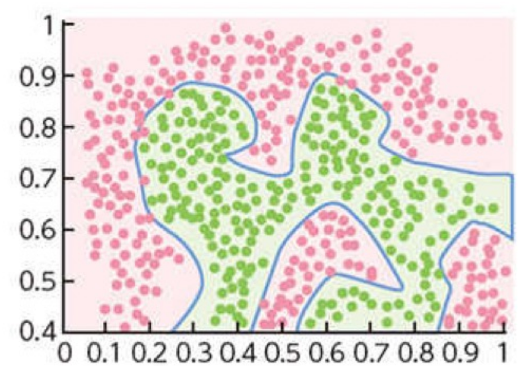


活性化関数の効果





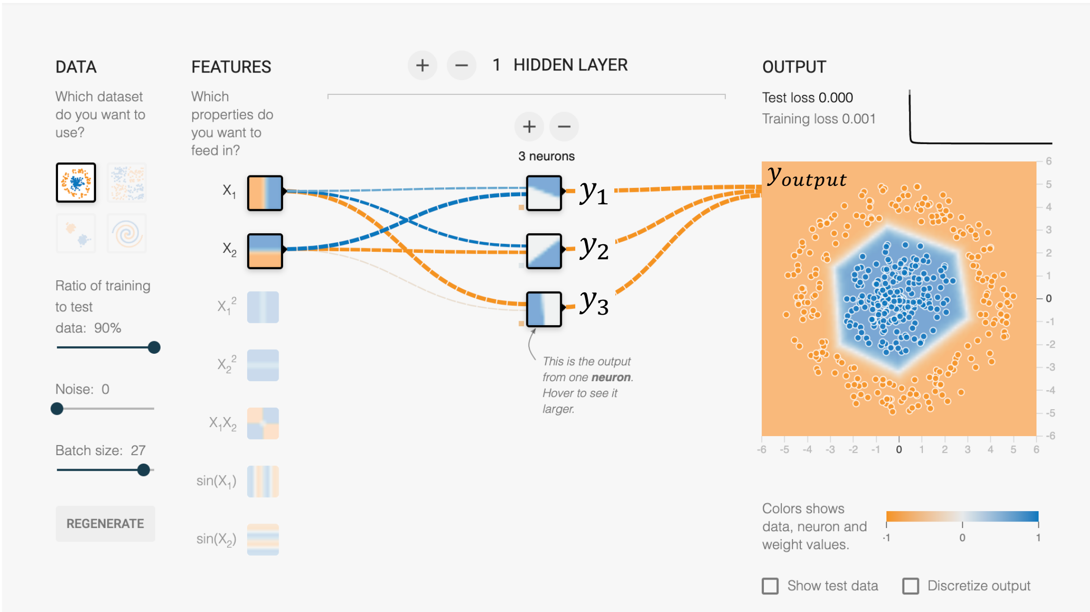
線形活性化関数では
層をディープに
しても線形

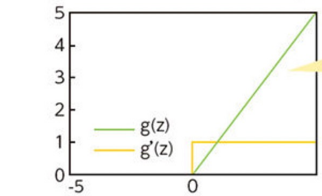


非線形活性化関数では
層をディープに
すると非常に
複雑になる

出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

ニューラルネットを実感しましょう！





$$y_1 = \max(0, w_1^1 x_1 + w_2^1 x_2 + b^1)$$

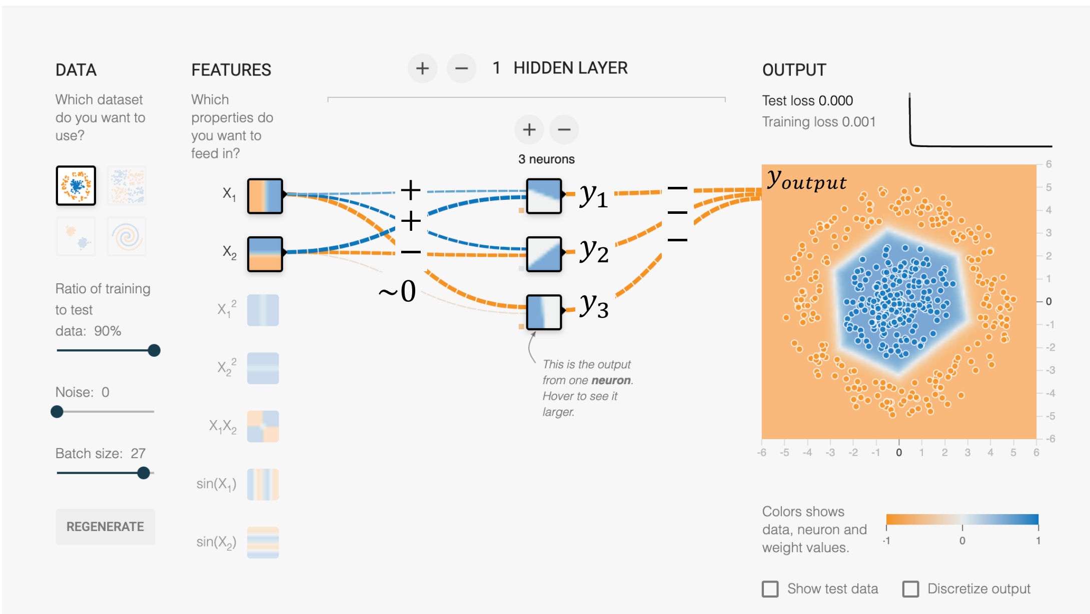
$$y_2 = \max(0, w_1^2 x_1 + w_2^2 x_2 + b^2)$$

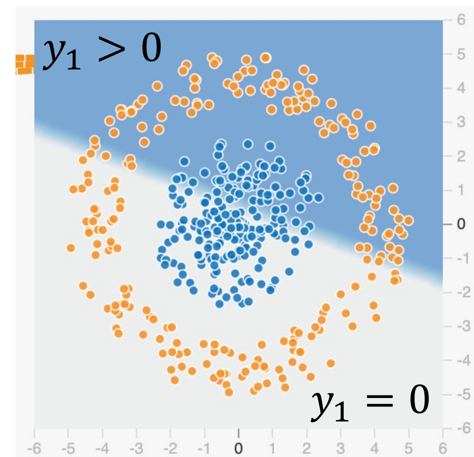
$$y_3 = \max(0, w_1^3 x_1 + w_2^3 x_2 + b^3)$$

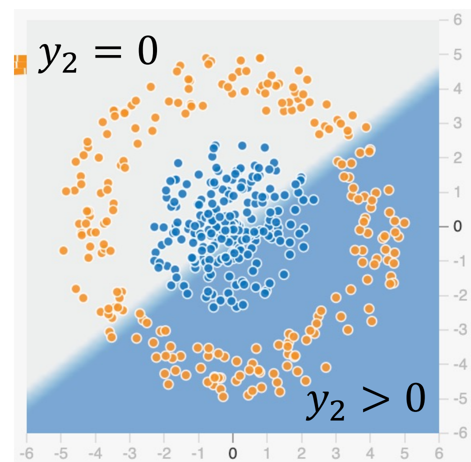
$$y_{output} = \max(0, w_1^4 y_1 + w_2^4 y_2 + w_3^4 y_3 + b^4)$$

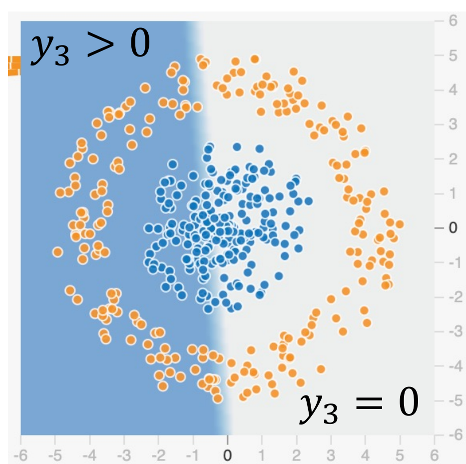
https://playground.tensorflow.org/

ニューラルネットを実感しましょう！



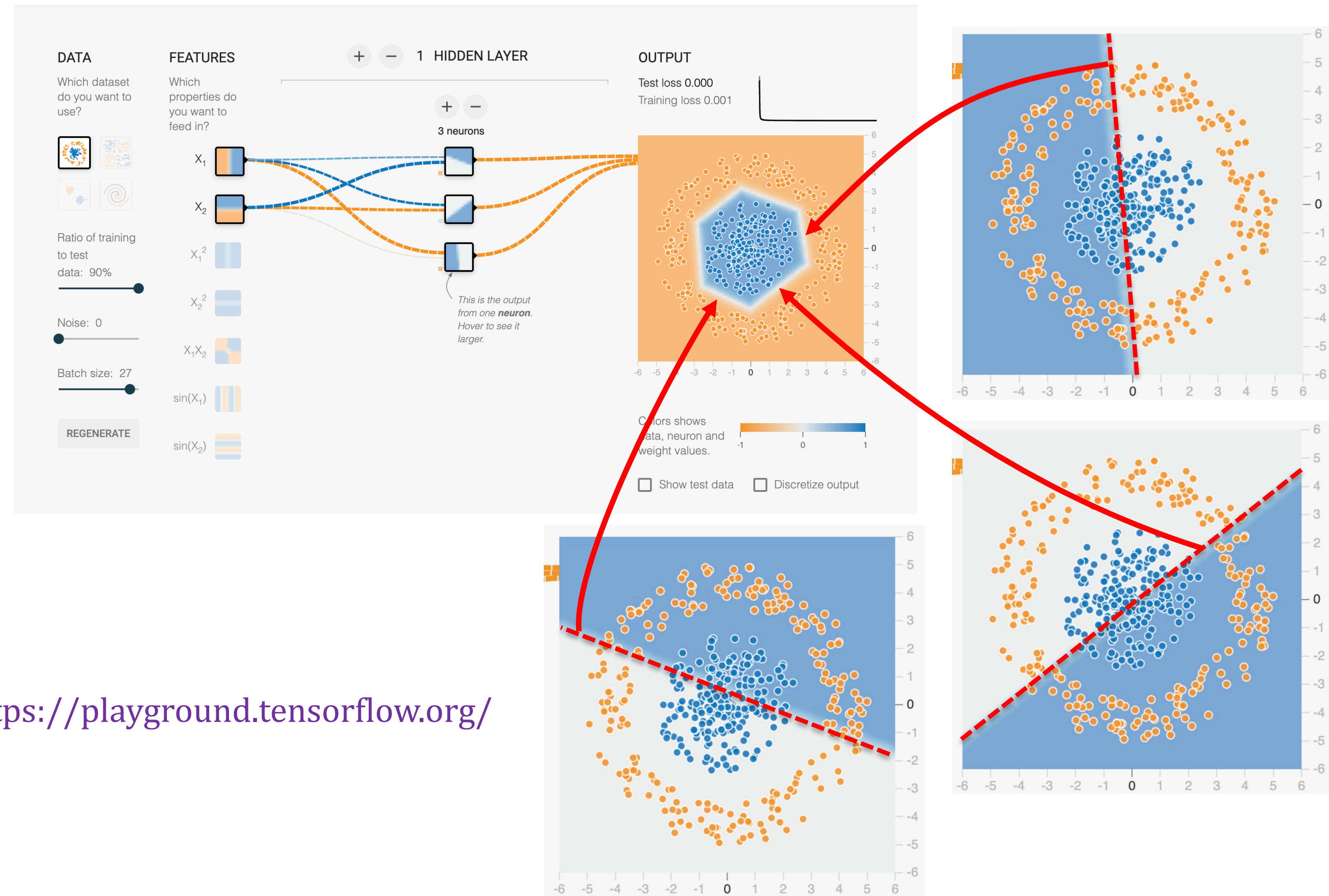






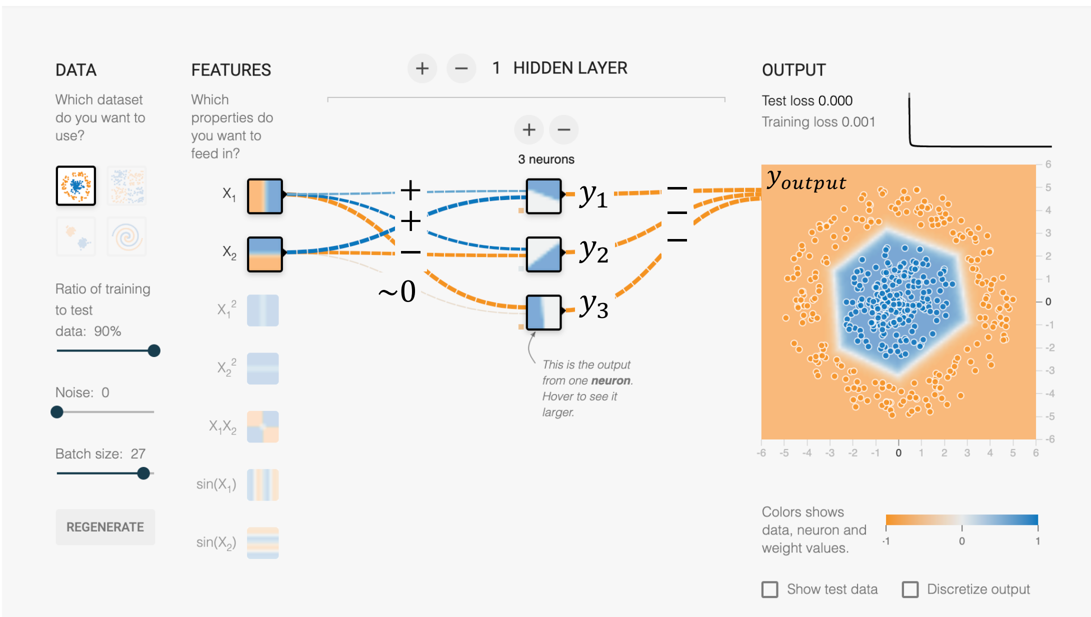
https://playground.tensorflow.org/

ニューラルネットを実感しましょう！



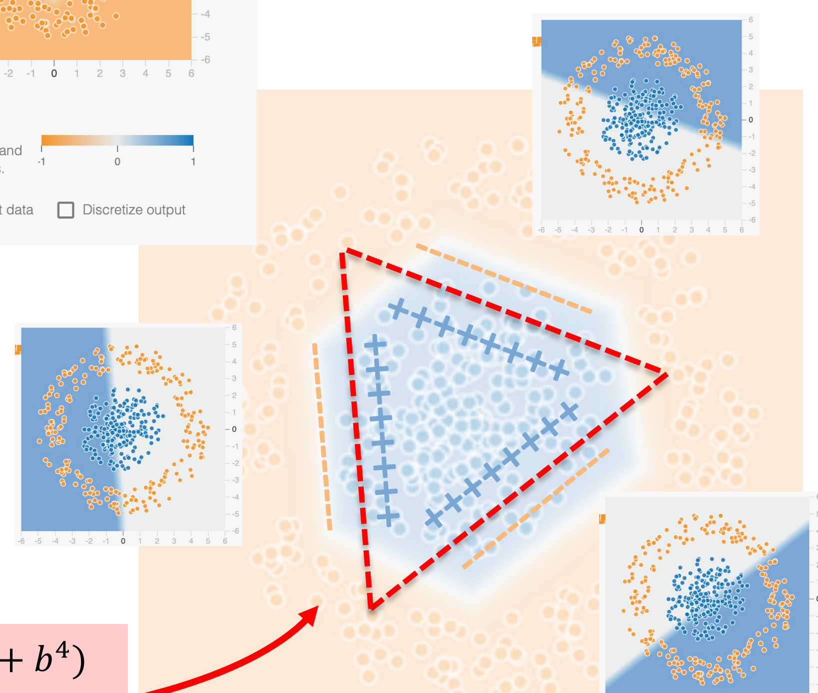
https://playground.tensorflow.org/

ニューラルネットを実感しましょう！



$$y_1 = \max(0, w_1^1 x_1 + w_2^1 x_2 + b^1)$$

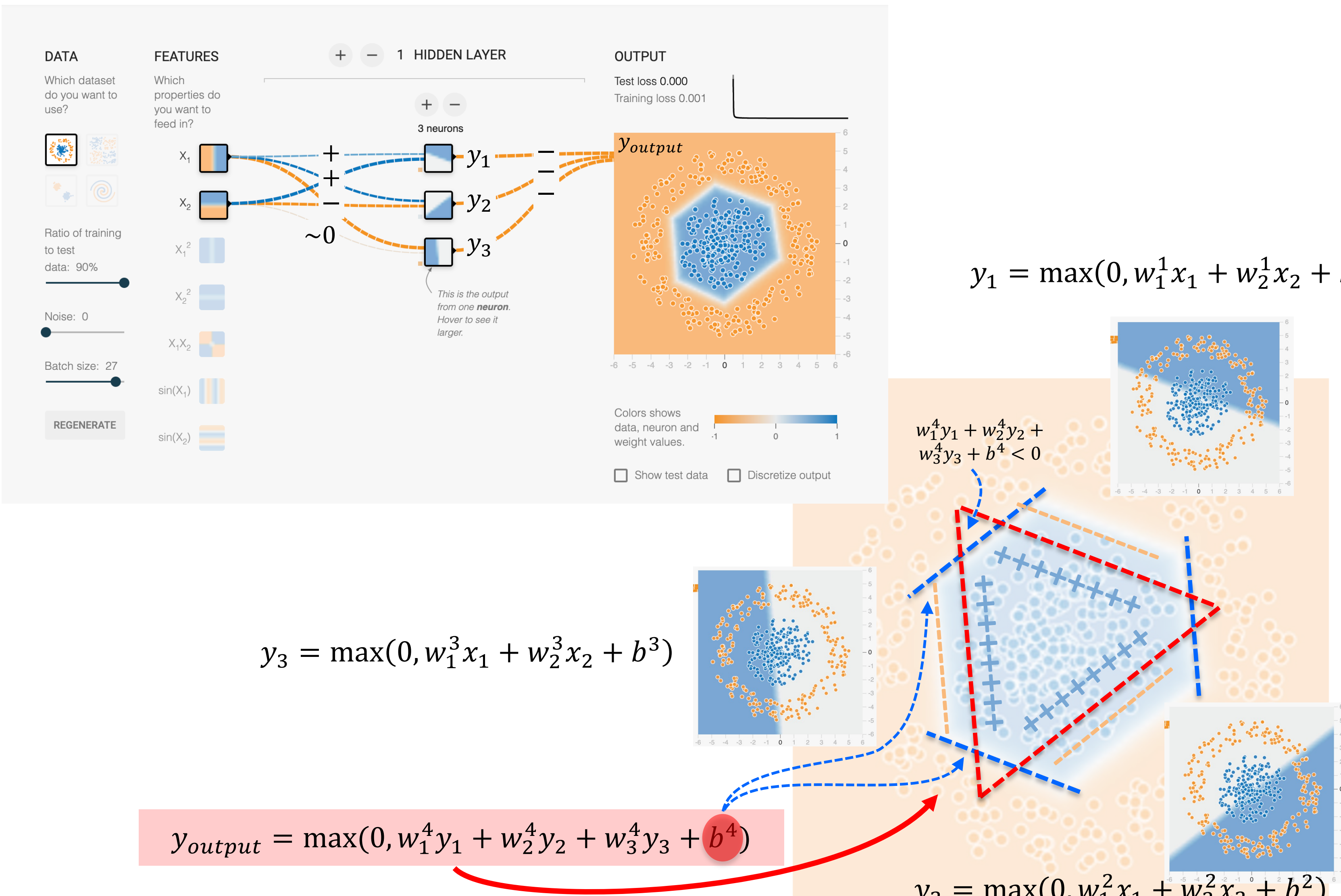
$$y_3 = \max(0, w_1^3 x_1 + w_2^3 x_2 + b^3)$$



$$y_{output} = \max(0, w_1^4 y_1 + w_2^4 y_2 + w_3^4 y_3 + b^4)$$

$$y_2 = \max(0, w_1^2 x_1 + w_2^2 x_2 + b^2)$$

ニューラルネットを実感しましょう！



$$y_1 = \max(0, w_1^1 x_1 + w_2^1 x_2 + b^1)$$

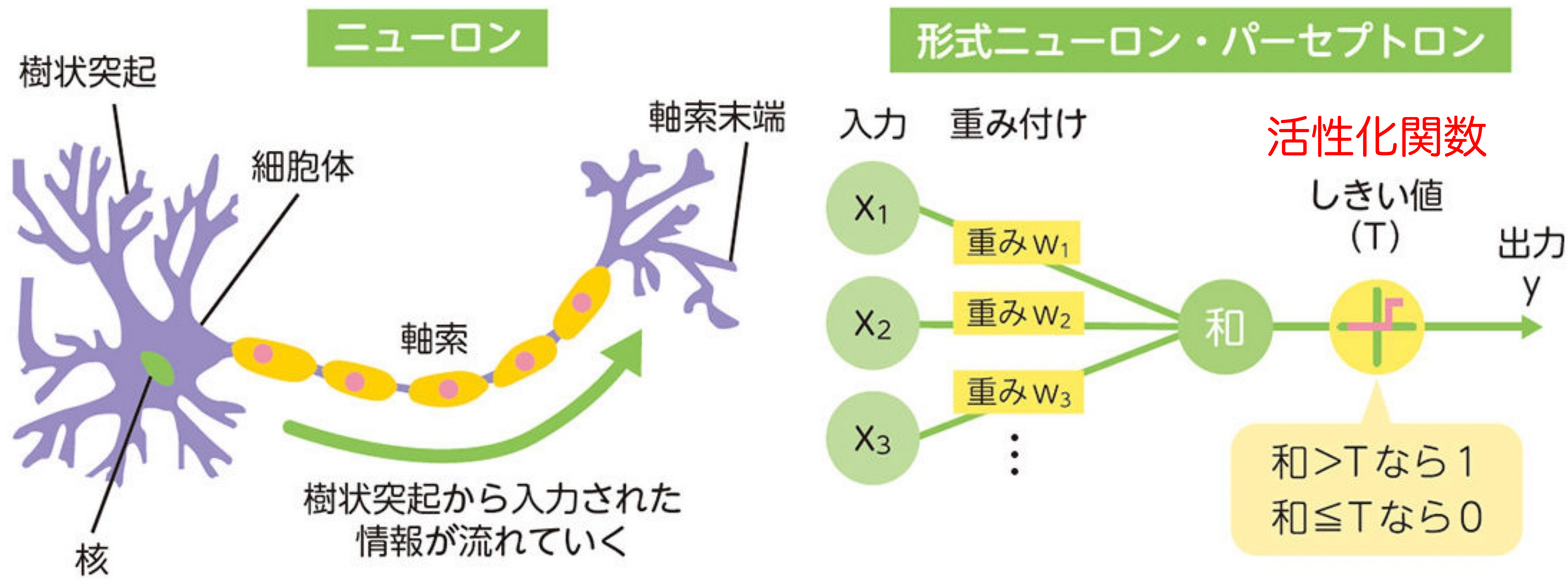
$$y_3 = \max(0, w_1^3 x_1 + w_2^3 x_2 + b^3)$$

$$y_{output} = \max(0, w_1^4 y_1 + w_2^4 y_2 + w_3^4 y_3 + b^4)$$

$$y_2 = \max(0, w_1^2 x_1 + w_2^2 x_2 + b^2)$$

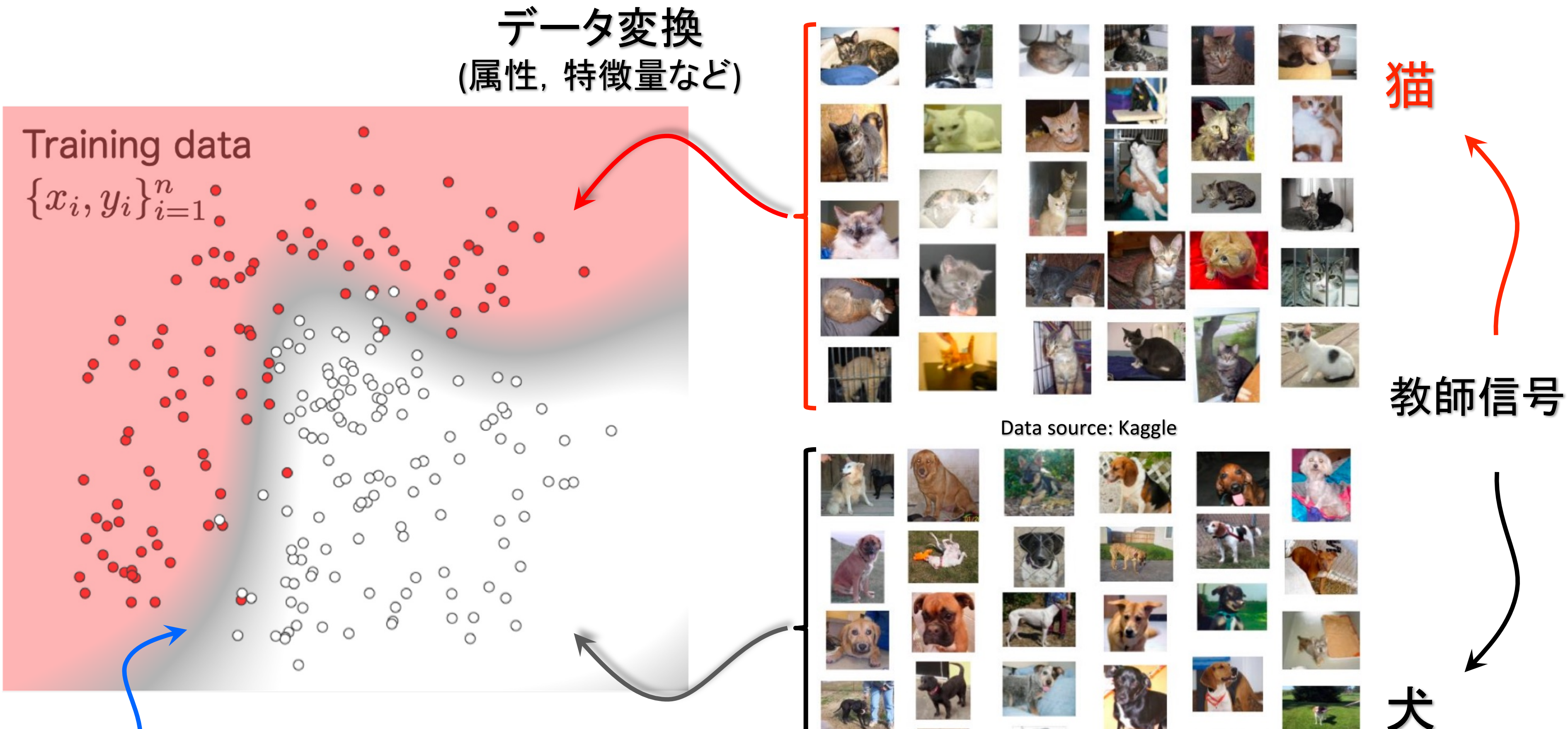
ニューロン・パーセプトロン

■ 1960年代までの研究



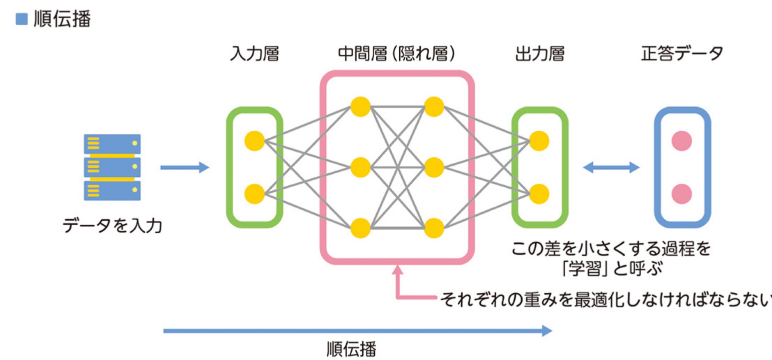
出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

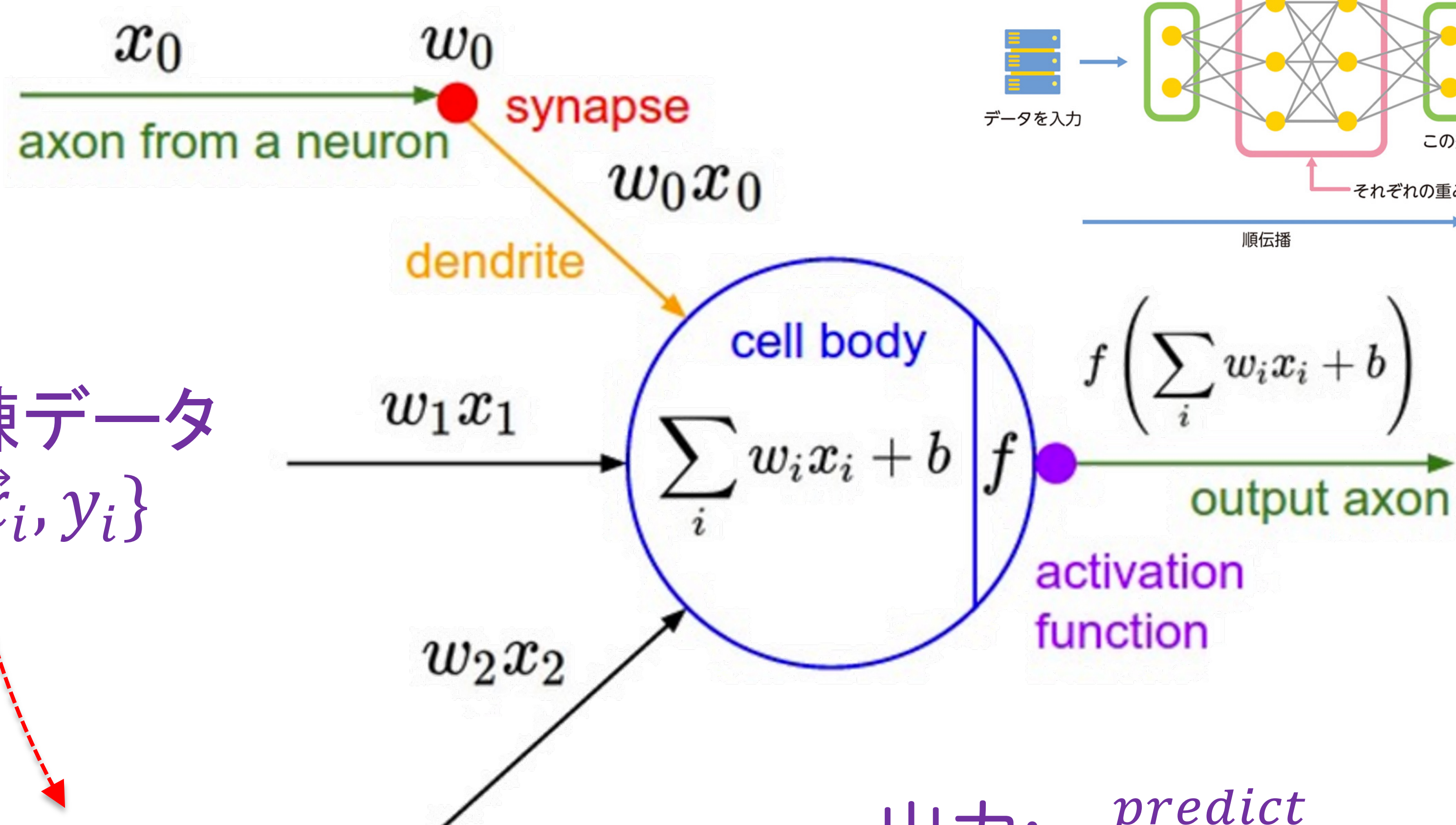
教師あり学習（復習）



$$y = f\left(\sum_k w_k x^k + b\right)$$

ニューラルネットの学習（1）





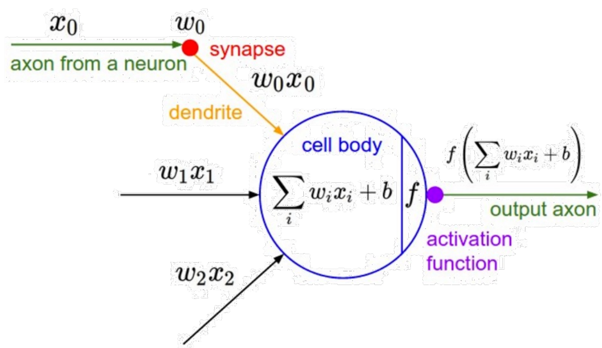
訓練データ
$\{\vec{x}_i, y_i\}$

誤差の和 ＝ **損失関数**
（小さくなるように w_i と b を学習する）

出力: $y_i^{predict} = f(\sum w_i x_i + b)$

誤差: $\left| y_i - y_i^{predict} \right|$

ニューラルネットの学習（2）



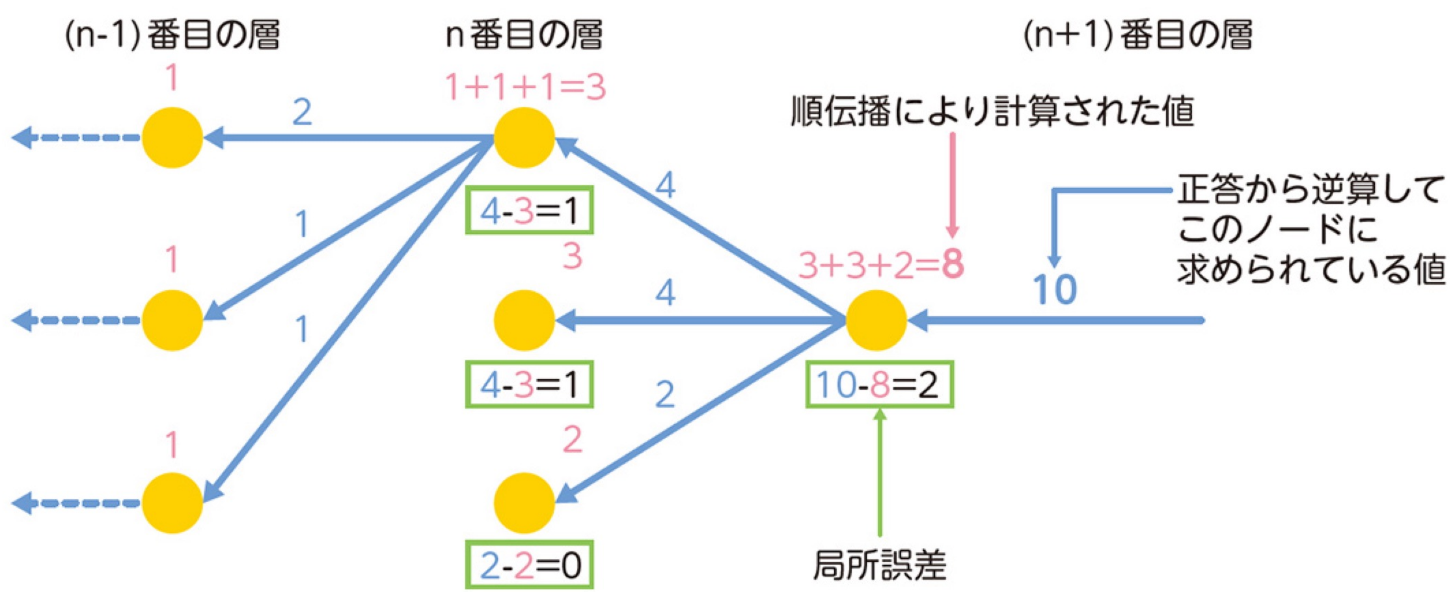
例：$f(x) = x$

$$y_i^{predict} = f\left(\sum w_j x_i^j + b\right) = \sum w_i x_i^j + b$$

$$\Delta y_i = \sum \Delta w_j \, x_i^j + \Delta b$$

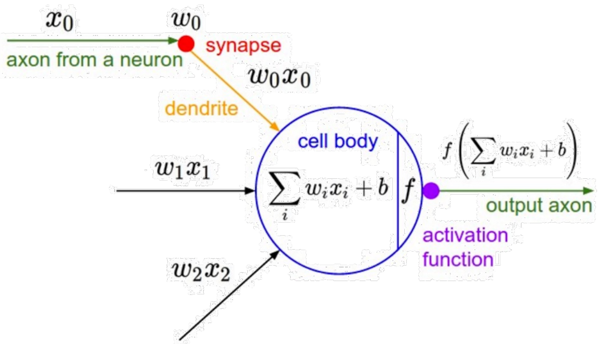
損失が小さくなるように w_j と b を調整する

■ 誤差逆伝播法



出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

ニューラルネットの学習（3）



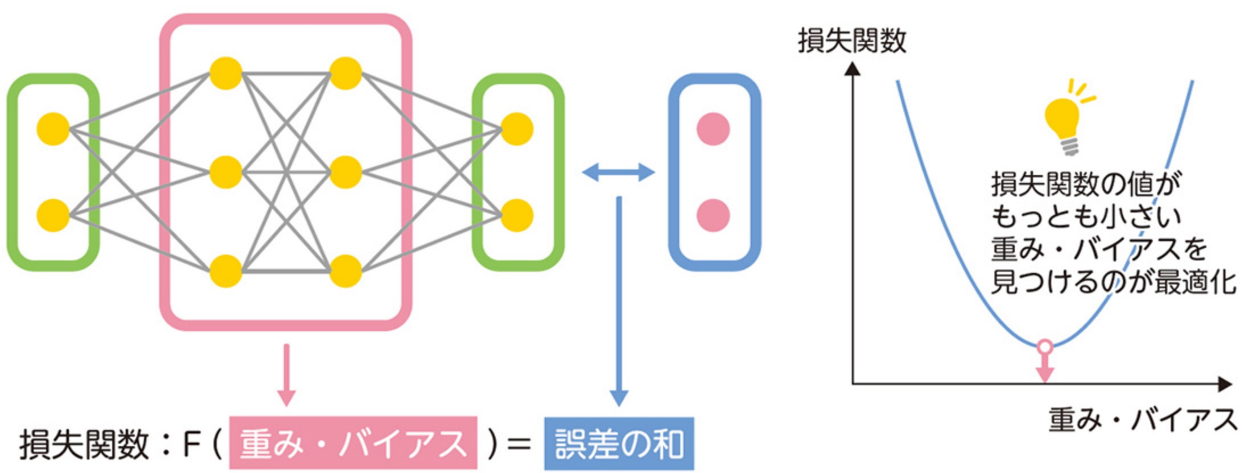
例： $f(x) = x$

$$y_i^{predict} = f\left(\sum w_j x_i^j + b\right) = \sum w_i x_i^j + b$$

$$\Delta y_i = \sum \Delta w_j \, x_i^j + \Delta b$$

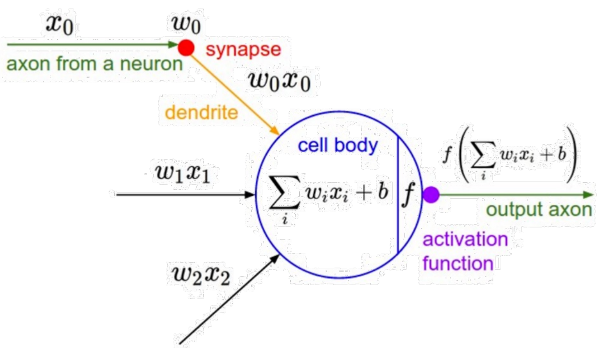
損失が小さくなるように $\underbrace{w_j \text{ と } b}_{\text{重み・バイアス}}$ を調整する

■ 損失関数を最小化する



出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

ニューラルネットの学習（4）



$$y_i^{predict} = f\left(\sum w_j x_i^j + b\right)$$

$$\Delta y_i = \left|y_i - y_i^{predict}\right| = \left|y_i - f\left(\sum w_j x_i^j + b\right)\right|$$

訓練データ
$\{\vec{x}_i, y_i\}$

$$F_i(\vec{w}, b) = \Delta y_i = \left|y_i - \overbrace{f\left(\sum w_j x_i^j + b\right)}^{\text{活性化関数 } f(z)}\right|$$

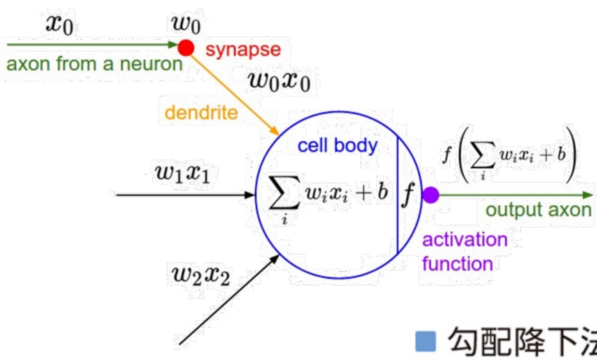
$$\frac{\delta F_i}{\delta w_j} = \frac{\delta F_i}{\delta f}\frac{\delta f}{\delta z}\frac{\delta z}{\delta w_j} = \pm\frac{\delta f}{\delta z}x_i^j \rightarrow \delta F_i = \pm\frac{\delta f}{\delta z}x_i^j\delta w_j$$

$$\frac{\delta F_i}{\delta b} = \frac{\delta F_i}{\delta f}\frac{\delta f}{\delta z}\frac{\delta z}{\delta w_j} = \pm\frac{\delta f}{\delta z} \rightarrow \delta F_i = \pm\frac{\delta f}{\delta z}\delta b$$

効率的に重みの調整（小 δw_j ）で
損失減らしたい（大 δF_i ）

$\overbrace{w_i \text{ を} \delta w_j \text{ や } b \text{ を} \delta b}^{\text{重み・バイアス}}$ 調整する際に
損失が変化する量

ニューラルネットの学習（5）



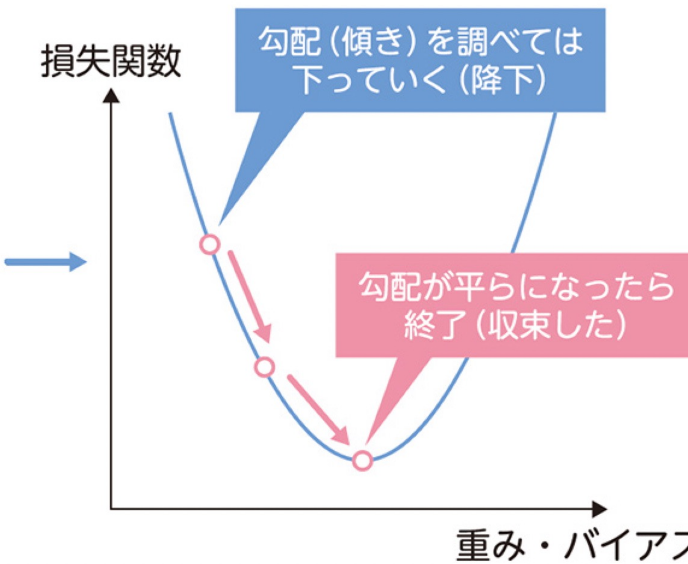
$$y_i^{predict} = f\left(\sum w_j x_i^j + b\right)$$

$$\Delta y_i = \left|y_i - y_i^{predict}\right| = \left|y_i - f\left(\sum w_j x_i^j + b\right)\right|$$

■ 勾配降下法は「山下り」

訓練データ

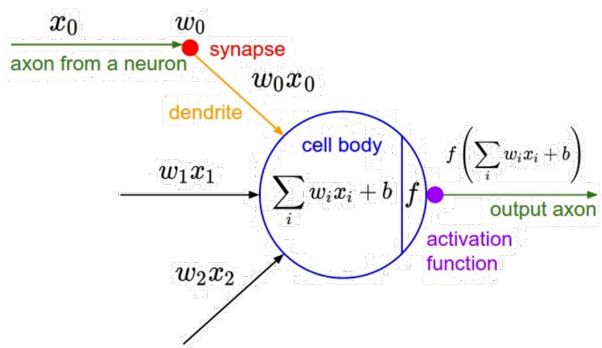
$\{\vec{x}_i, y_i\}$



効率的に重みの調整（小 δw_j ）で
損失減らしたい（大 δF_i ）

w_i をδw_j や b をδb 調整する際に
損失が変化する量

ニューラルネットの学習（6）



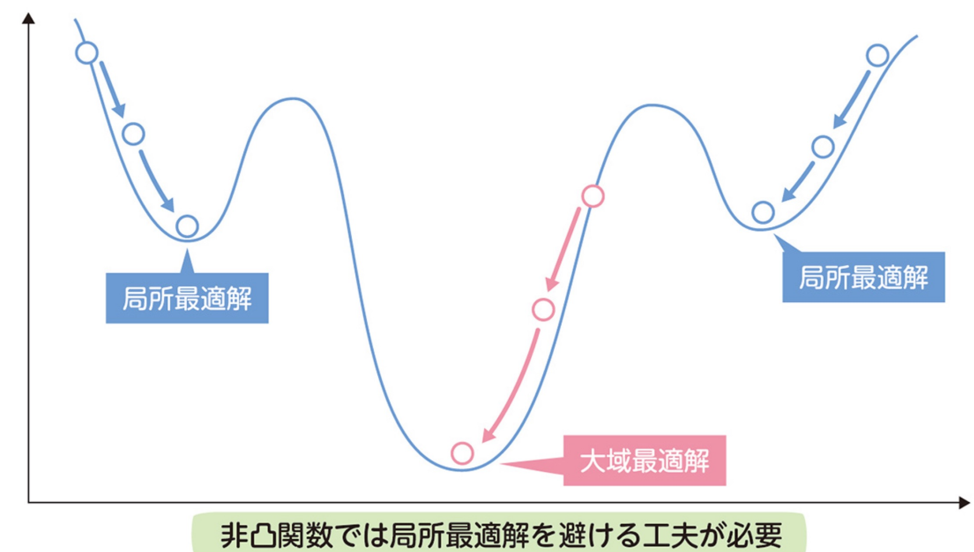
例：$f(x) = x$

$$y_i^{predict} = f(\sum w_i x_i + b) = \sum w_i x_i + b$$

$$\Delta y = y_i - y_i^{predict} = \sum \Delta w_i \, x_i + \Delta b$$

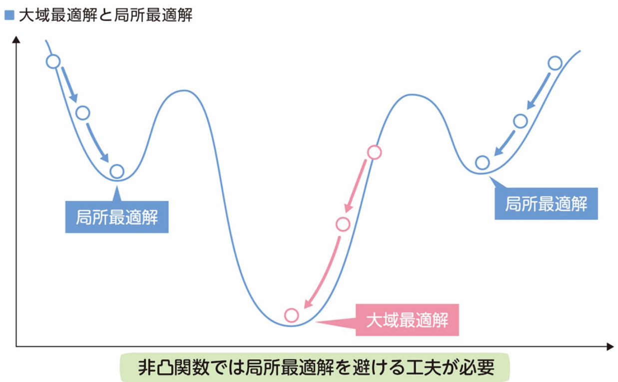
損失が小さくなるように $\underbrace{w_i \text{ と } b}_{\text{重み・バイアス}}$ を調整する

■大域最適解と局所最適解

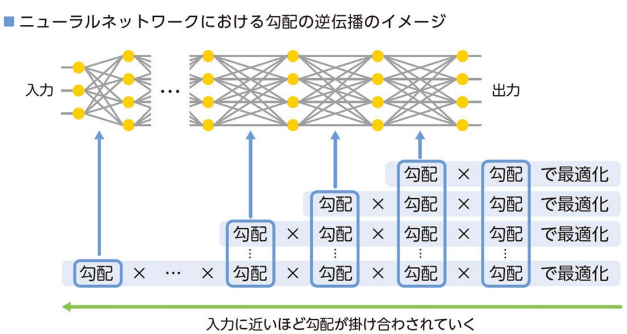


出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

ニューラルネットの学習（7）



局所最適解へのトラップ

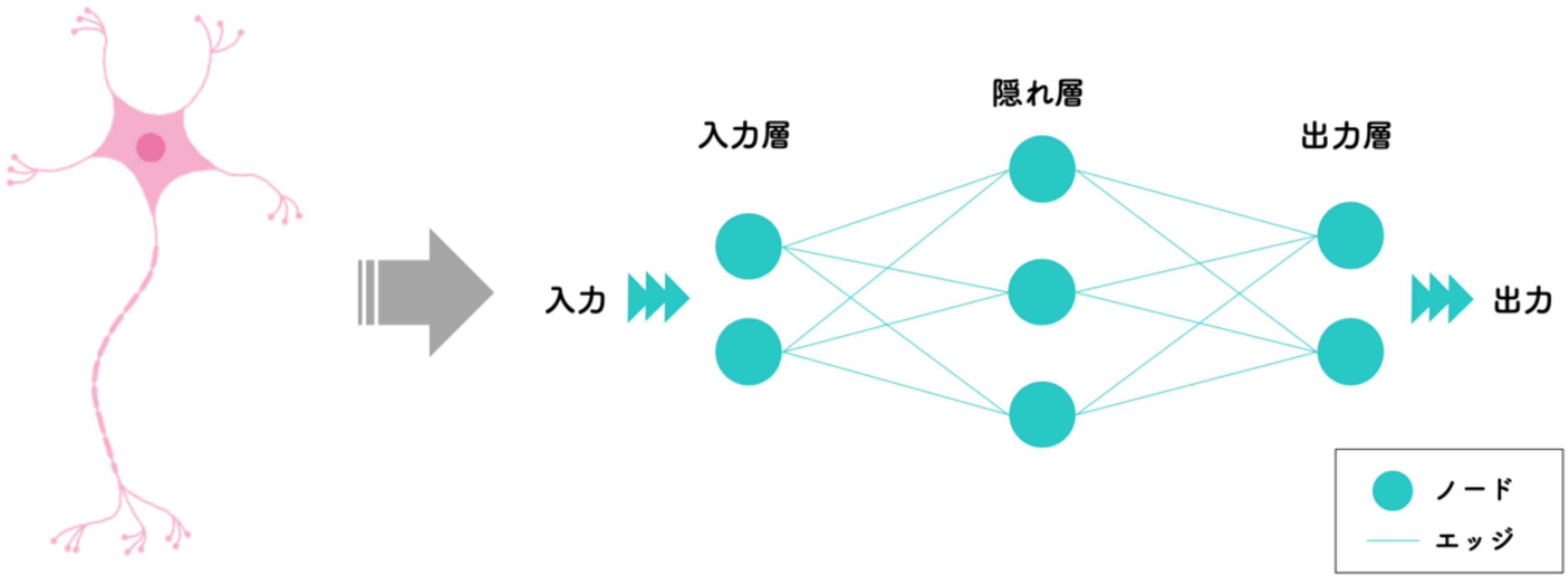


勾配消失問題

- 最適化アルゴリズム
  - SGD（確率勾配降下法）
    一つ一つの学習データの順番を入れ替えながら勾配降下法を適用する
  - Momentum SGD
    勾配を降りている方向に「運動量」をもたせる
  - Adagrad
    パラメターごとに適切に学習率を変化させる
  - RMSprop
    勾配の合計を指数移動平均を利用してAdagradを改良する
  - Adam
    RMSpropとMomentum SGDを組み合わせたアルゴリズム

ニューラルネットのまとめ

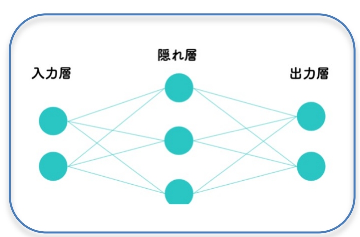
神経細胞(ニューロン)　　　　ニューラルネットワーク



基本機能

$$\vec{y} = f(\vec{x})$$

訓練データ $\left\{\begin{matrix} x_i^1 & y_i^1 \\ \vdots & \vdots \\ x_i^n & y_i^m \end{matrix}\right\}$ → 学習 →



→ 活用 → $\vec{y}_{new} = f(\vec{x}_{new})$

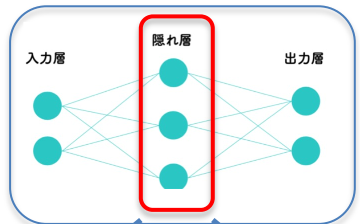
深層学習への展開

ニューラルネット

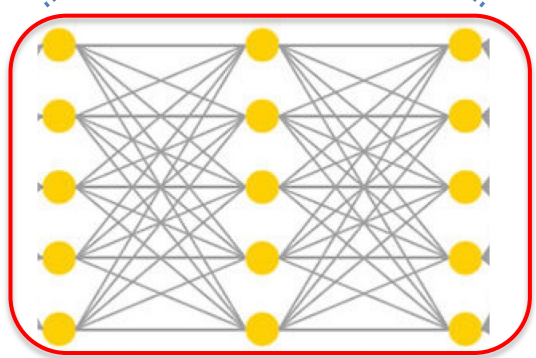
基本機能

$\vec{y} = f(\vec{x})$

訓練データ $\left\{ \begin{matrix} x_i^1 & y_i^1 \\ \vdots & , \quad \vdots \\ x_i^n & y_i^m \end{matrix} \right\}$ 学習



活用 $\vec{y}_{new} = f(\vec{x}_{new})$



②
基本機能の繋がりの設計により柔軟に問題を提起できる

③
$\vec{y}$の定義仕方で適切に問題を提起できる

①
隠れ層の数を増やし、多種の活性化関数の使用によって表現できる関数を複雑化できる

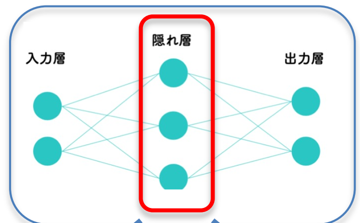
深層学習：複雑関数の表現

ニューラルネット

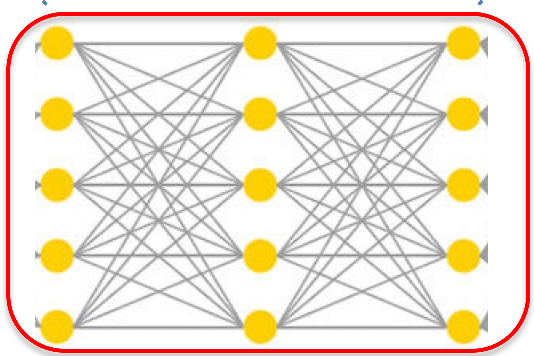
基本機能

$$\vec{y} = f(\vec{x})$$

訓練データ $\begin{Bmatrix} x_i^1 & y_i^1 \\ \vdots & \vdots \\ x_i^n & y_i^m \end{Bmatrix}$ → 学習 →



→ 活用 → $\vec{y}_{new} = f(\vec{x}_{new})$

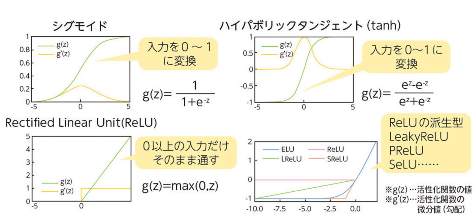


$$y = f^n\left(b_n + \sum w_j^{n-1} f_j^{n-1}\left(b_{n-1} + \sum w_j^{n-2} f_j^{n-2}\left(b_{n-2} + \sum w_j^{n-3} f_j^{n-3}(\cdots\right.\right.\right.$$

k層目の出力　　$k-1$層目の出力

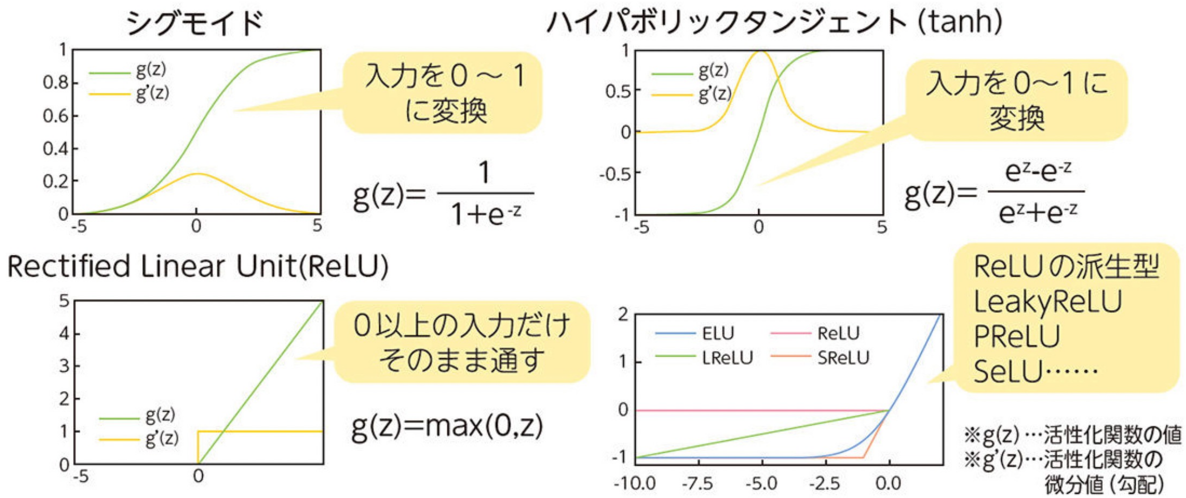
$$x_j^k = f_j^k\left(\sum w_j^k x_j^{k-1} + b_k\right)$$

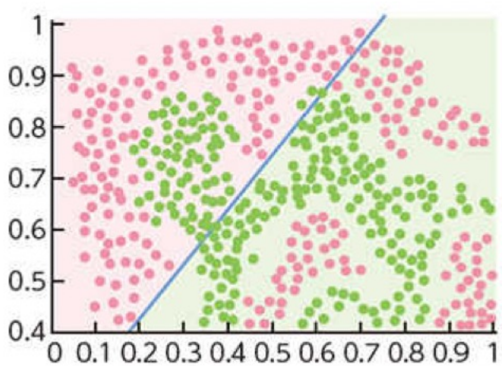
活性化関数　　k層目の重み　　k層目のバイアス



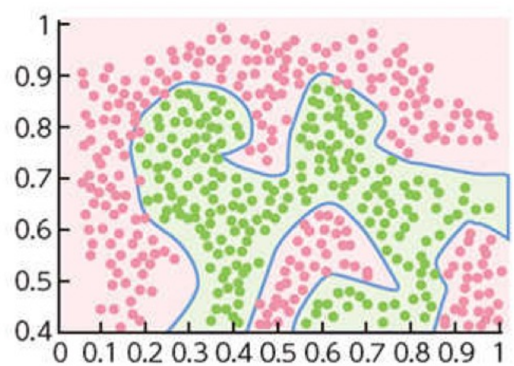
①
隠れ層の数を増やし、多種の活性化関数の使用によって表現できる関数を複雑化できる

活性化関数の効果





線形活性化関数では
層をディープに
しても線形

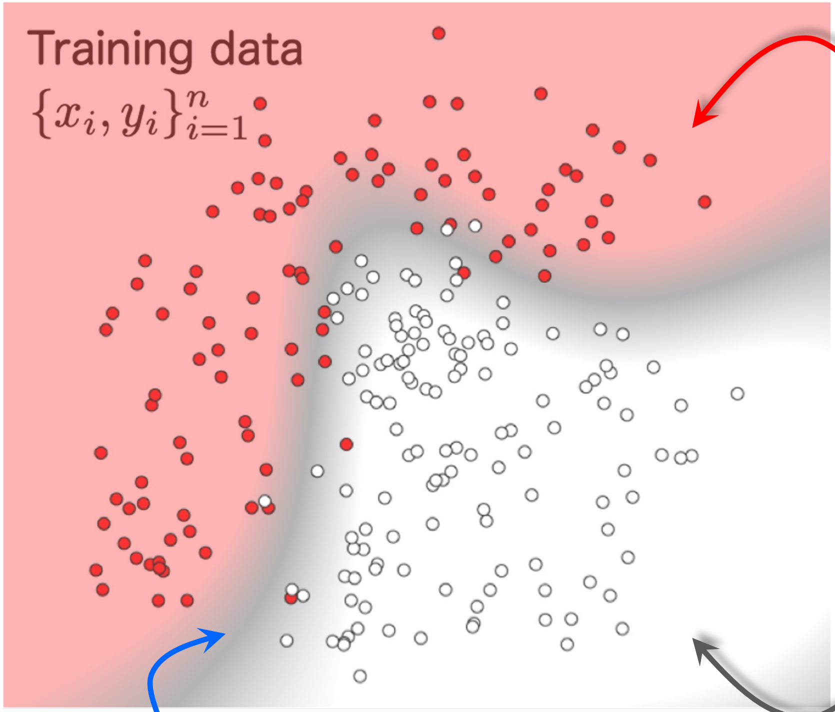


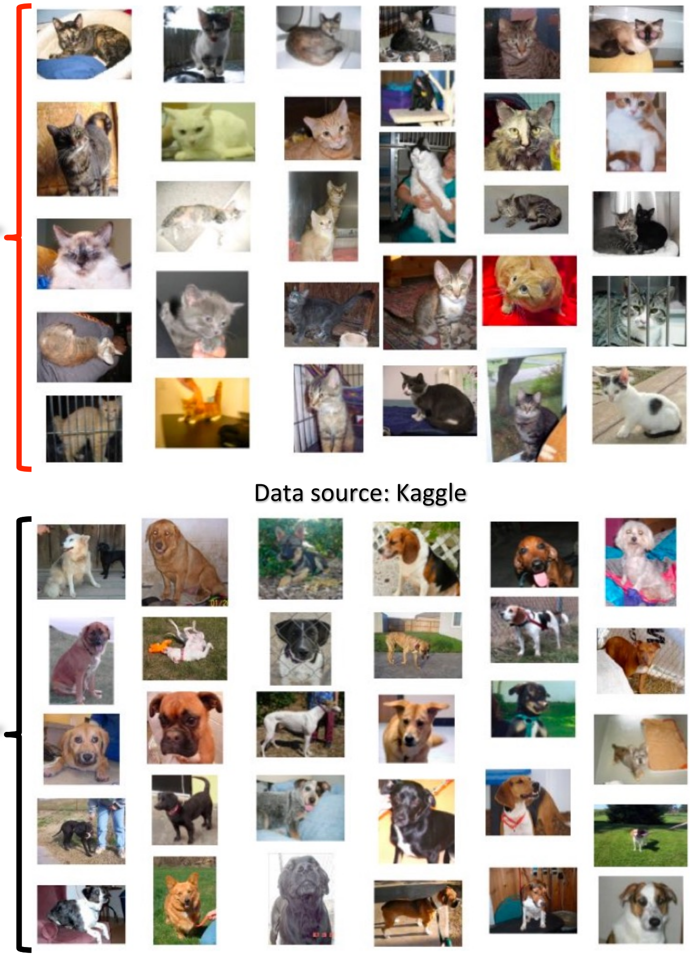
非線形活性化関数では
層をディープに
すると非常に
複雑になる

出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

教師あり学習（復習）

データ変換
(属性，特徴量など)





猫

教師信号

犬

$$y = f\left(\sum_k w_k x^k + b\right)$$

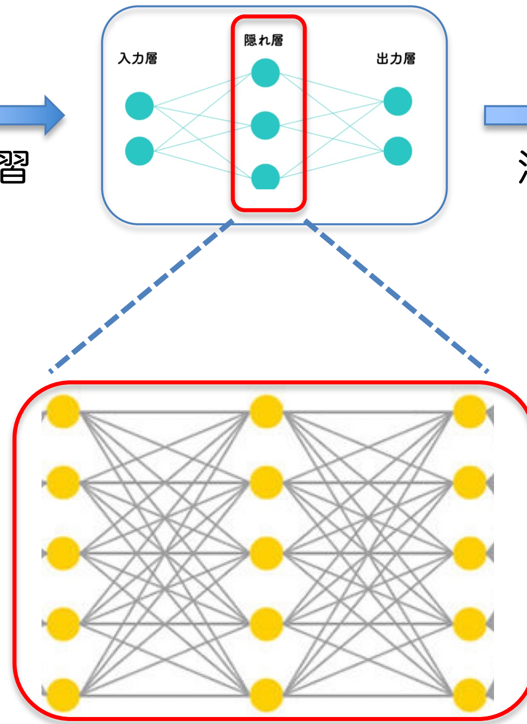
深層学習への展開

ニューラルネット

基本機能

$$\vec{y} = f(\vec{x})$$

訓練データ $\left\{ \begin{matrix} x_i^1 & y_i^1 \\ \vdots\ , & \vdots \\ x_i^n & y_i^m \end{matrix} \right\}$ 学習



活用 $\vec{y}_{new} = f(\vec{x}_{new})$

②
基本機能の繋がりの設計により柔軟に問題を提起できる

③
$\vec{y}$の定義仕方で適切に問題を提起できる

①
隠れ層の数を増やし、多種の活性化関数の使用によって表現できる関数を複雑化できる

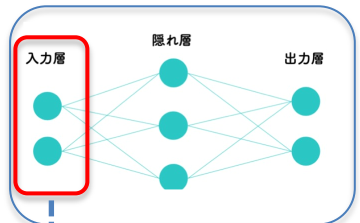
深層学習の画像処理への展開（１）

ニューラルネット

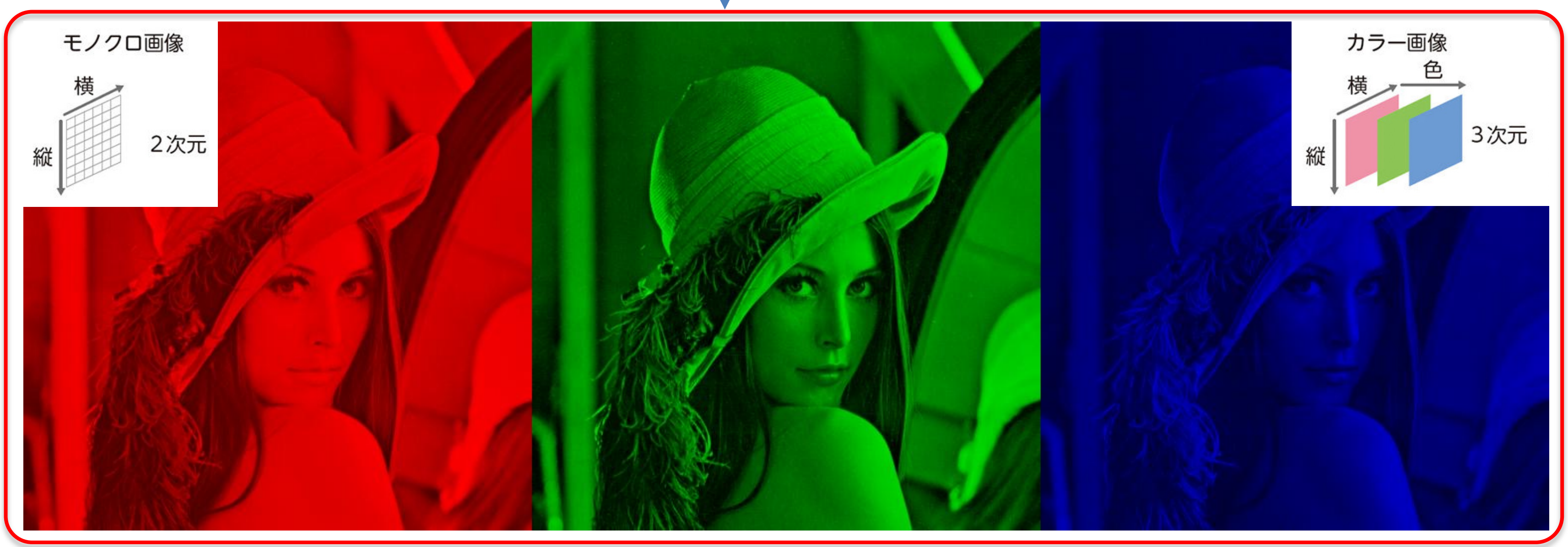
基本機能

$$\vec{y} = f(\vec{x})$$

訓練データ $\left\{ \begin{matrix} x_i^1 & y_i^1 \\ \vdots & , \vdots \\ x_i^n & y_i^m \end{matrix} \right\}$ $\xrightarrow{\text{学習}}$



$\xrightarrow{\text{活用}}$ $\vec{y}_{new} = f(\vec{x}_{new})$



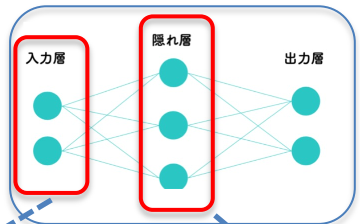
深層学習の画像処理への展開（2）

ニューラルネット

基本機能

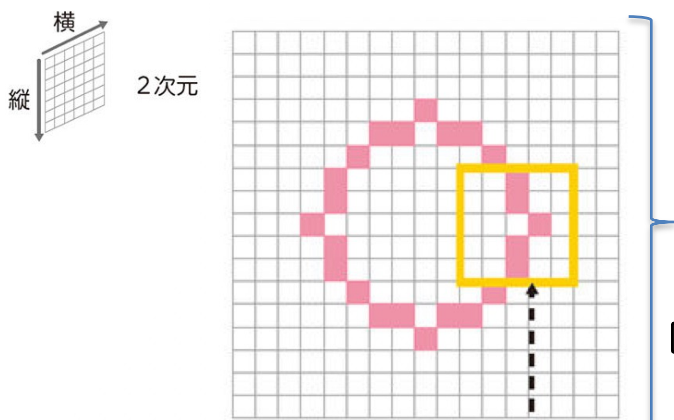
$$\vec{y} = f(\vec{x})$$

訓練データ $\left\{\begin{matrix} x_i^1 & y_i^1 \\ \vdots & \vdots \\ x_i^n & y_i^m \end{matrix}\right\}$ $\xrightarrow[\text{学習}]{}$



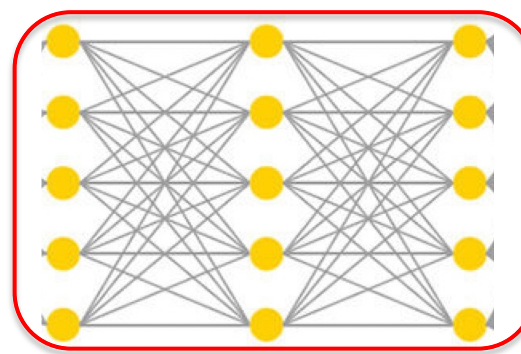
$\xrightarrow[\text{活用}]{}$ $\vec{y}_{new} = f(\vec{x}_{new})$

モノクロ画像



固定グリッド数に変換

特化したネットワークの設計

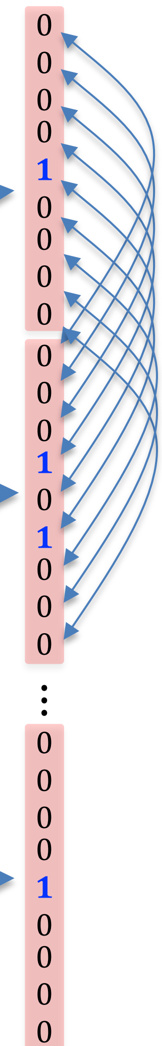


ピクセル同士の「位置関係」が重要

画像処理用ネットワークの設計（1）

白黒画像の行列表現

0	0	0	0	1	0	0	0	0
0	0	0	1	0	1	0	0	0
0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	1	0
0	0	1	0	0	0	1	0	0
0	0	0	1	0	1	0	0	0
0	0	0	0	1	0	0	0	0

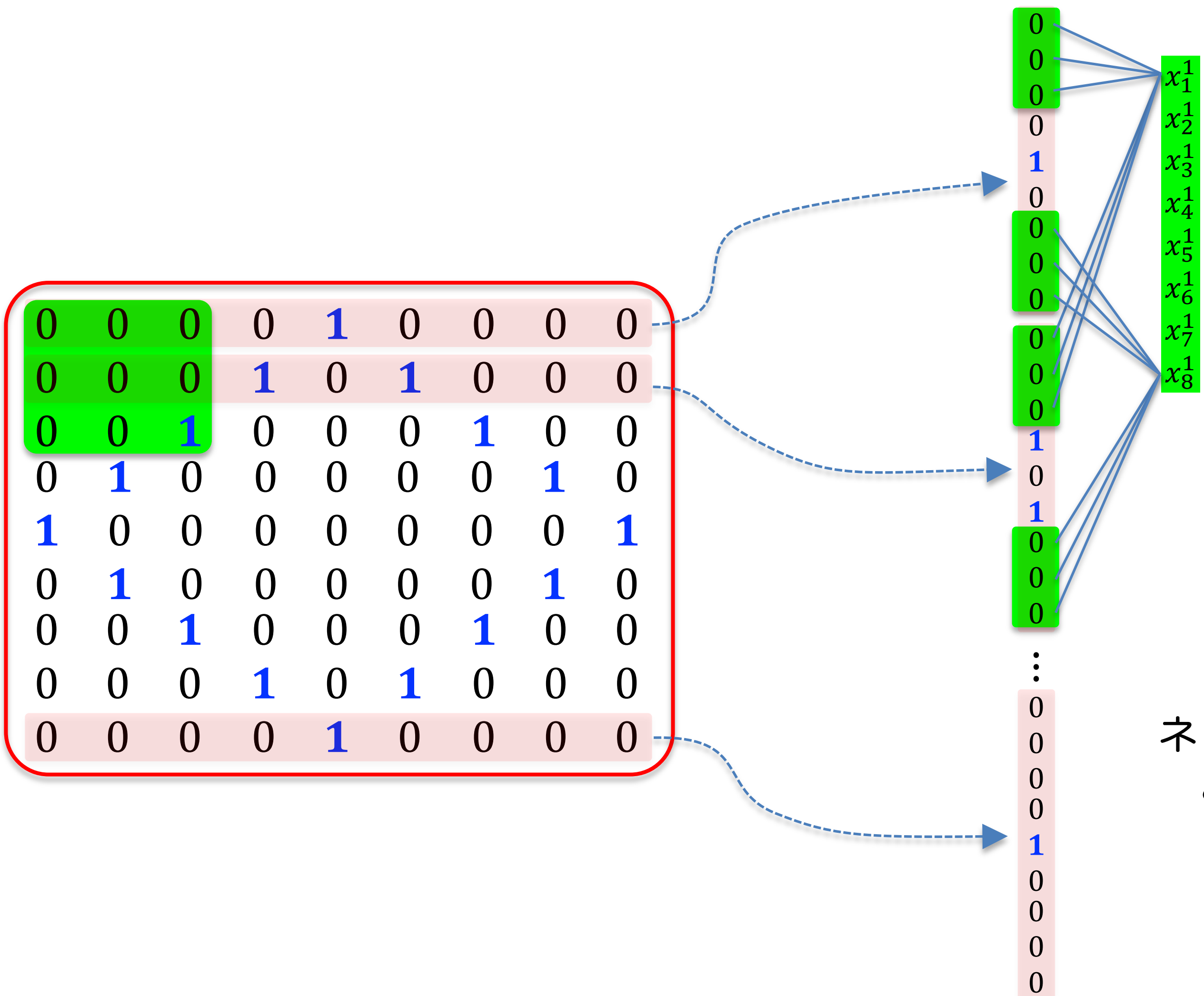


ピクセル同士の
上下近傍関係が存在

ネットワークの繋がり設計に
より画像処理応用に特化

行列表現からベクトル表現へ

画像処理用ネットワークの設計（2）

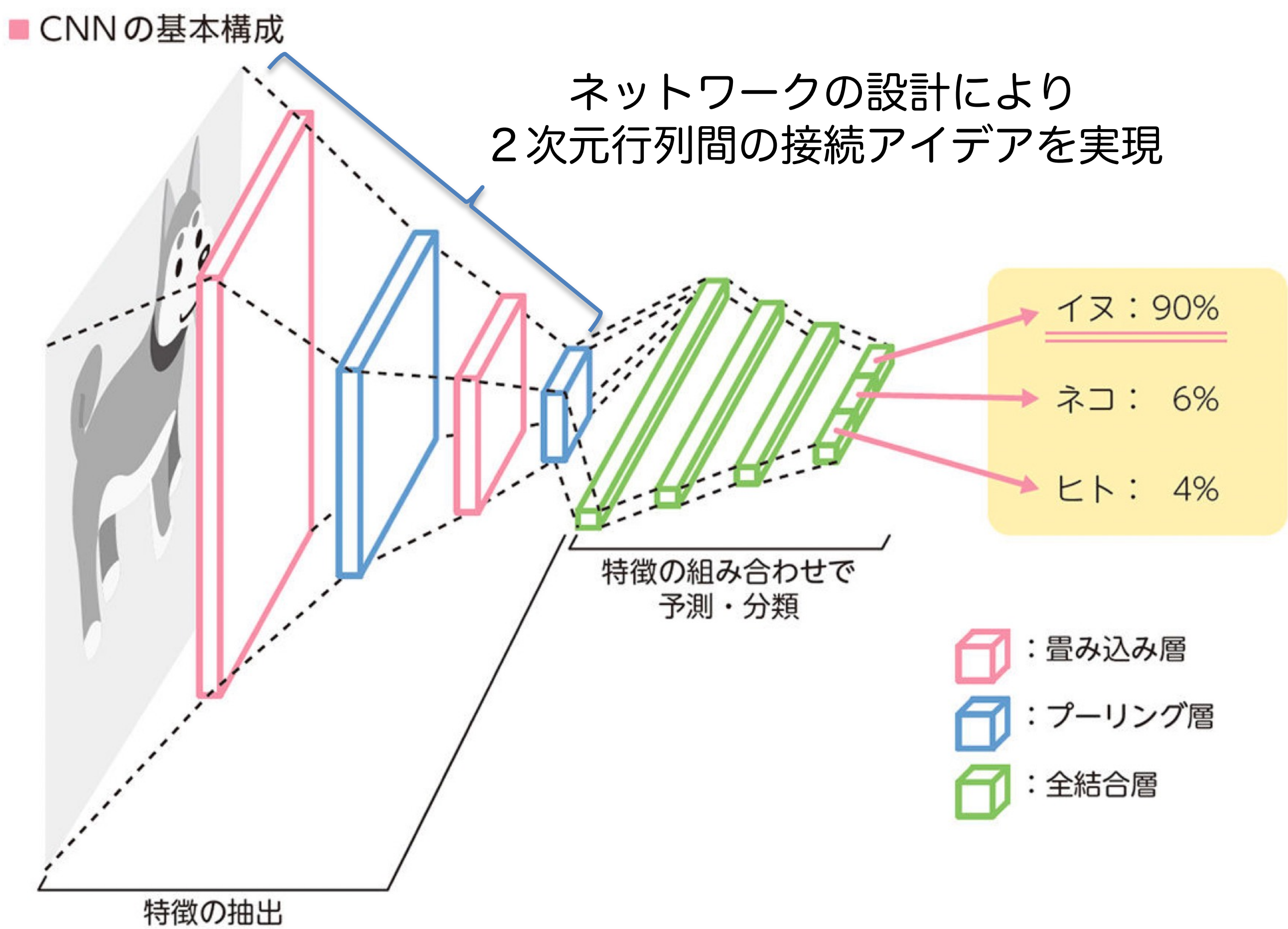


ピクセル同士の
上下近傍関係の考慮

ネットワークの繋がり設計に
より画像処理応用に特化

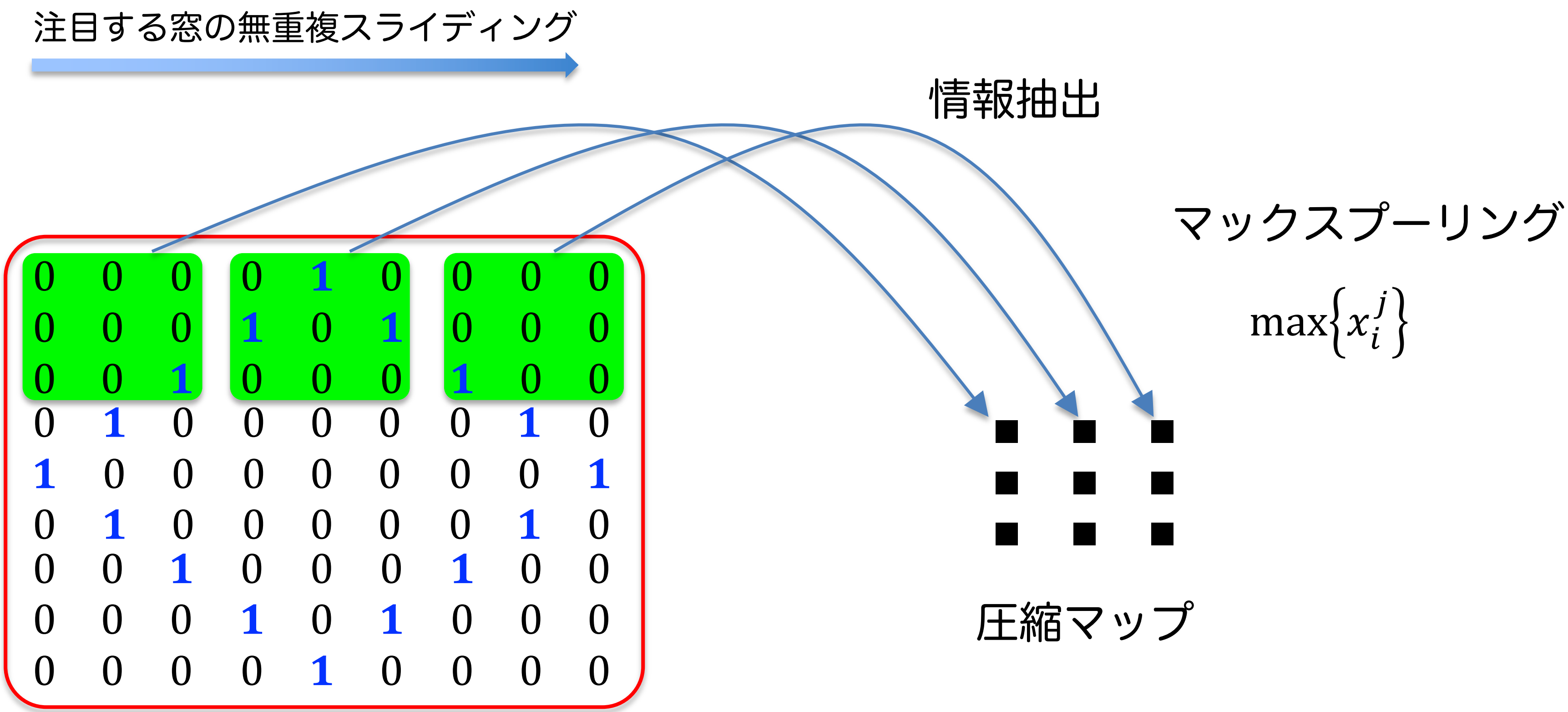
行列表現からベクトル表現へ

畳み込みニューラルネットワーク（１）



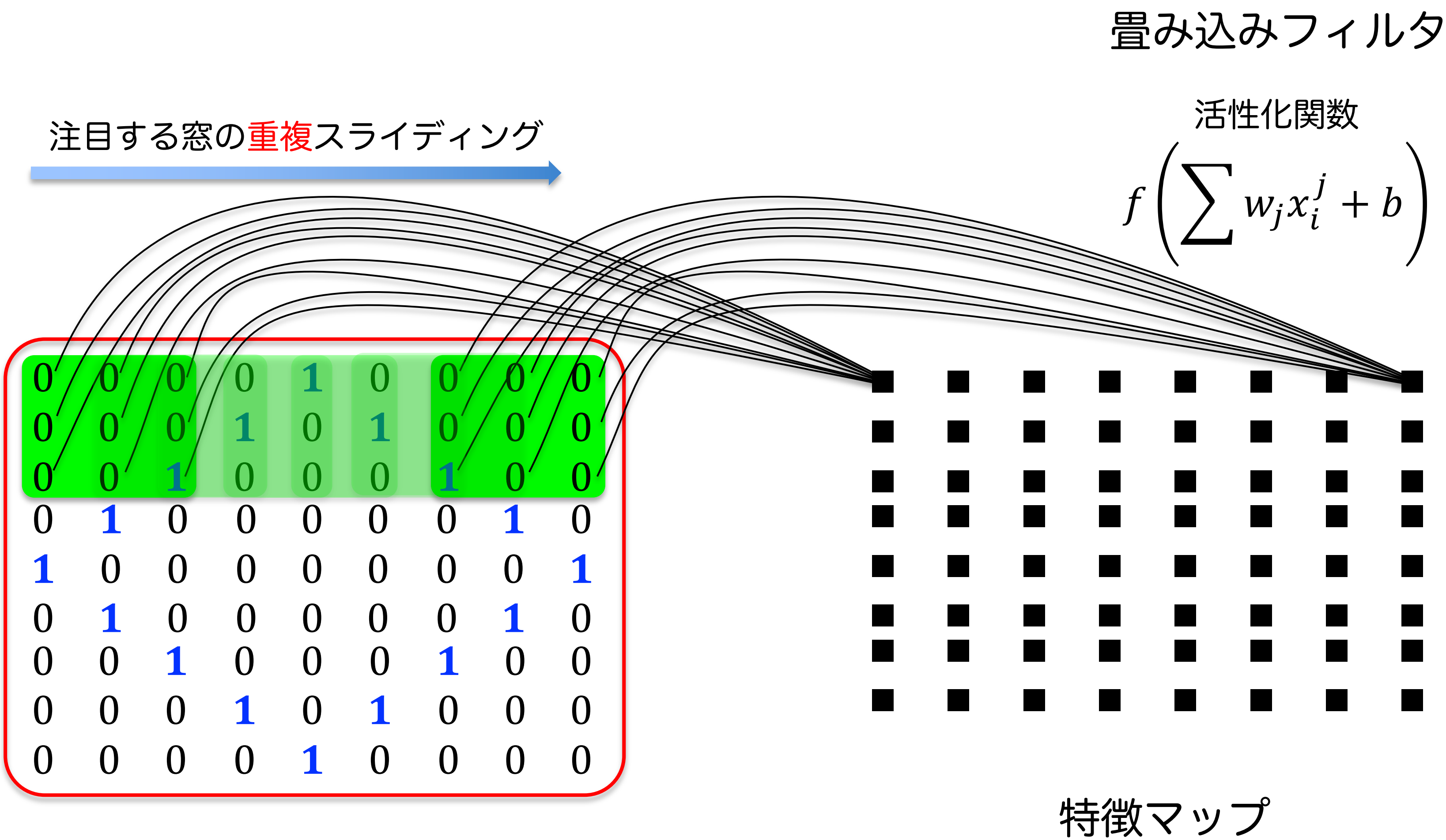
出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

プーリング層



注目する窓の無重複スライディング
情報抽出
マックスプーリング
$\max\left\{x_i^j\right\}$
0 0 0 0 1 0 0 0 0
0 0 0 1 0 1 0 0 0
0 0 1 0 0 0 1 0 0
0 1 0 0 0 0 0 1 0
1 0 0 0 0 0 0 0 1
0 1 0 0 0 0 0 1 0
0 0 1 0 0 0 1 0 0
0 0 0 1 0 1 0 0 0
0 0 0 0 1 0 0 0 0
圧縮マップ

畳み込み層


畳み込みフィルタ
活性化関数
$f\left(\sum w_j x_i^j + b\right)$
注目する窓の重複スライディング
0 0 0 0 1 0 0 0 0
0 0 0 1 0 1 0 0 0
0 0 1 0 0 0 1 0 0
0 1 0 0 0 0 0 1 0
1 0 0 0 0 0 0 0 1
0 1 0 0 0 0 0 1 0
0 0 1 0 0 0 1 0 0
0 0 0 1 0 1 0 0 0
0 0 0 0 1 0 0 0 0
特徴マップ

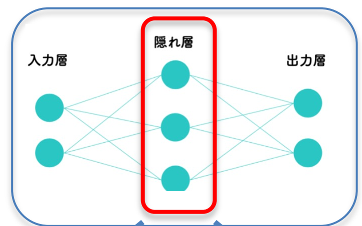
深層学習への展開

ニューラルネット

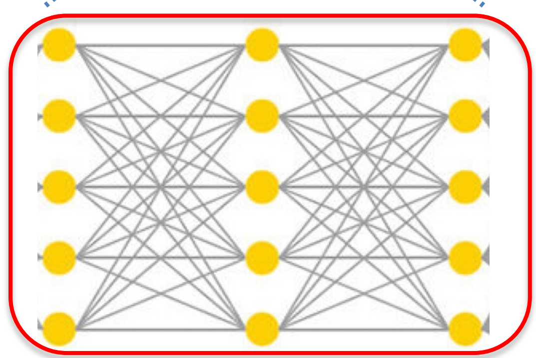
基本機能

$$\vec{y} = f(\vec{x})$$

訓練データ $\begin{Bmatrix} x_i^1 & y_i^1 \\ \vdots\, , & \vdots \\ x_i^n & y_i^m \end{Bmatrix}$ → 学習 →



→ 活用 → $\vec{y}_{new} = f(\vec{x}_{new})$



②
基本機能の繋がりの設計により柔軟に問題を提起できる

③
$\vec{y}$の定義仕方で適切に問題を提起できる

①
隠れ層の数を増やし、多種の活性化関数の使用によって表現できる関数を複雑化できる

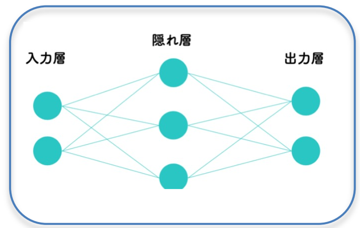
オートエンコーダ（1）

ニューラルネット

基本機能

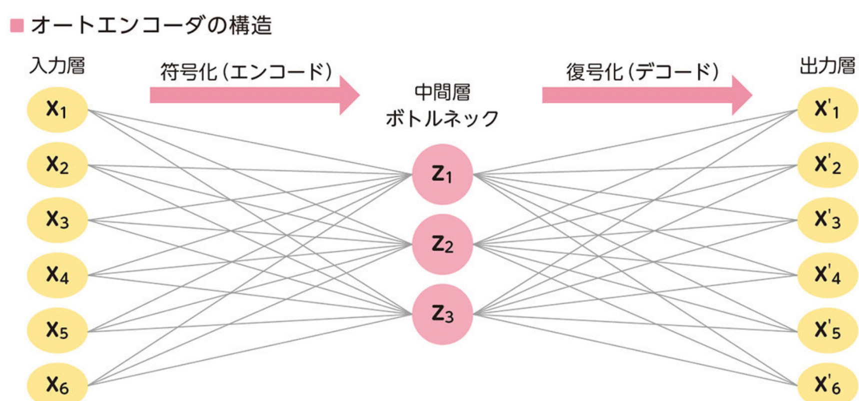
$$\vec{y} = f(\vec{x})$$

訓練データ $\left\{\begin{matrix} x_i^1 & y_i^1 \\ \vdots \;, & \vdots \\ x_i^n & y_i^m \end{matrix}\right\}$ 学習



活用 $\vec{y}_{new} = f(\vec{x}_{new})$

$\vec{y}$ の柔軟な定義による
教師なし学習



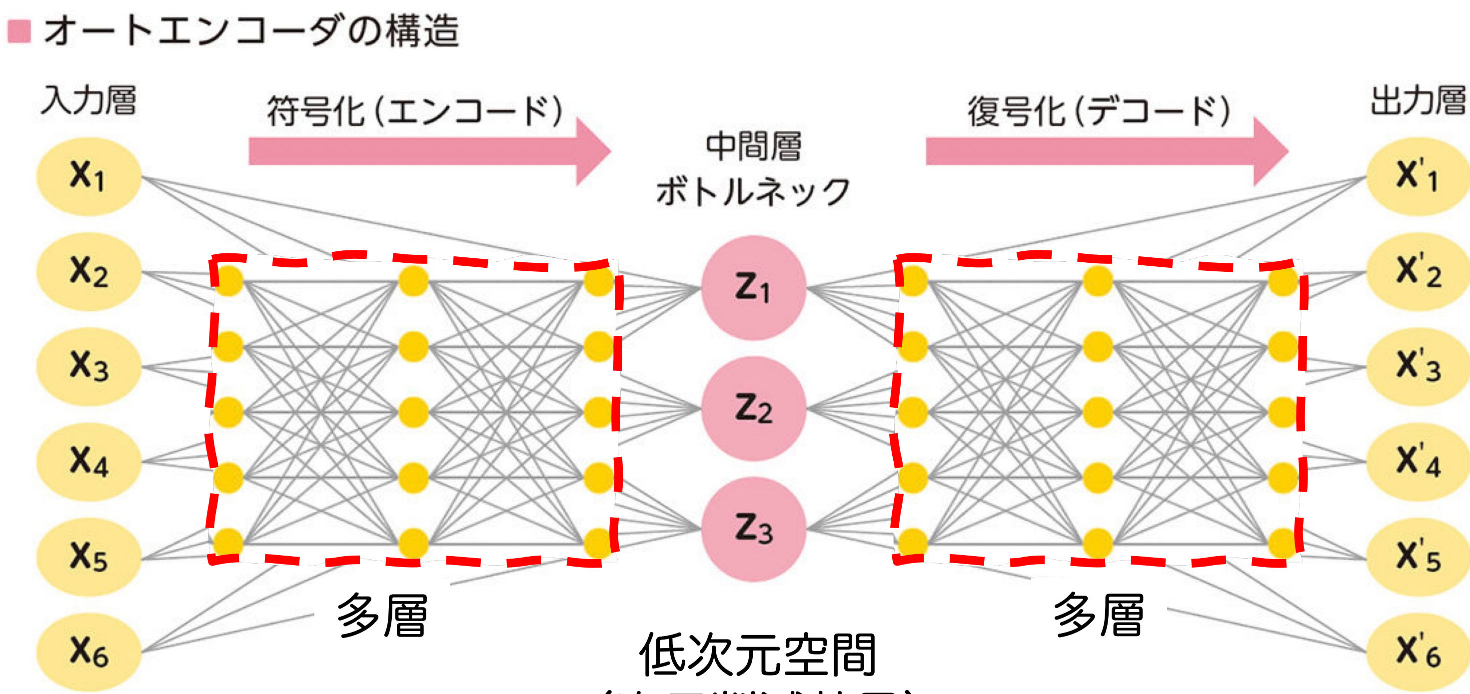
$\vec{y} = \vec{x}$

入力したデータと同じデータが出力されるニューラルネット

③
$\vec{y}$の定義仕方で適切に問
題を提起できる

出典：図解即戦力　機械学習&ディープラーニングのしくみと技術がこれ1冊で
しっかりわかる教科書　株式会社アイデミー (著), 山口 達輝 (著), 松田 洋之 (著)

オートエンコーダ（2）



低次元空間
(次元削減効果)

多次元空間
(重複情報あり)

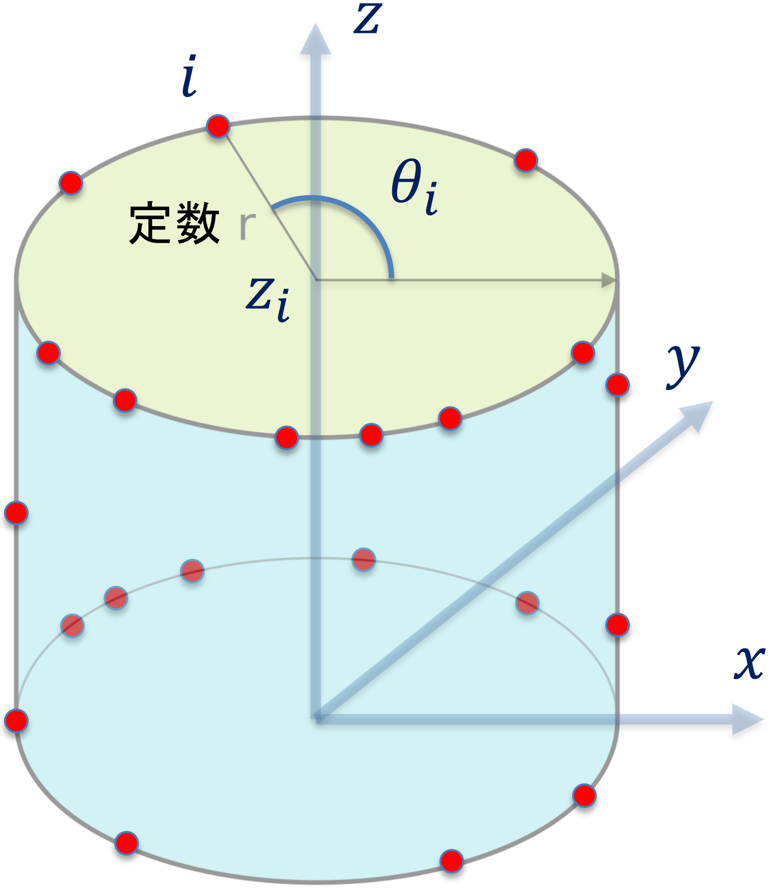
重複情報なし＝潜在変数

多次元空間
(重複情報あり)

$\vec{y} = \vec{x}$

情報が正しく復元できるように圧縮する

潜在変数の例（1）



半径が定数の円柱の表面

データ

直交座標系 $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}$ 3次元

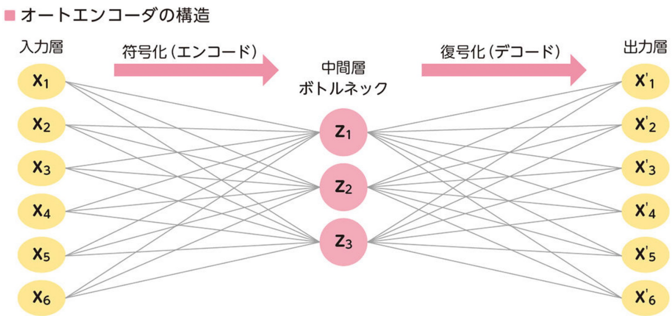
半径が定数の円柱表面にある束縛条件が利用されていない

円柱座標系 $\begin{Bmatrix} \theta_i \\ z_i \end{Bmatrix}$ 2次元

半径が定数の円柱表面にある束縛条件が利用する潜在変数

座標変換
$$\begin{cases} x_i = r\cos\theta_i \\ y_i = r\sin\theta_i \\ z_i = z_i \end{cases}$$

潜在変数の例（２）



深層学習によって自動的にできると期待する
(それほど簡単にうまくいきませんが！！)

データ

直交座標系 $\begin{Bmatrix} x_i \\ y_i \\ z_i \end{Bmatrix}$

3次元
半径が定数の円柱表面にある束縛条件が利用されていない

符号化
（エンコード）

円柱座標系 $\begin{Bmatrix} \theta_i \\ z_i \end{Bmatrix}$

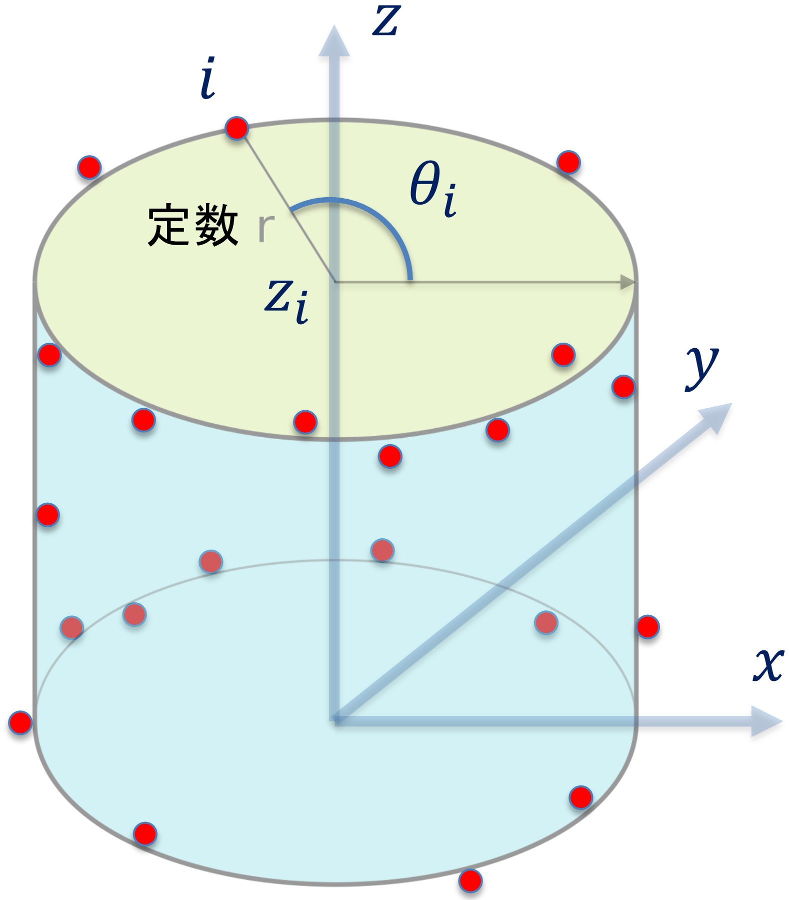
2次元
半径が定数の円柱表面にある束縛条件が利用する潜在変数

復号化
（デコード）

座標変換

$$\begin{Bmatrix} x_i = r\cos\theta_i \\ y_i = r\sin\theta_i \\ z_i = z_i \end{Bmatrix}$$

潜在変数の例（３）



半径が定数の円柱の表面

データ

直交座標系 $\begin{Bmatrix} x_i \\ y_i \\ z_i \end{Bmatrix}$ 3次元

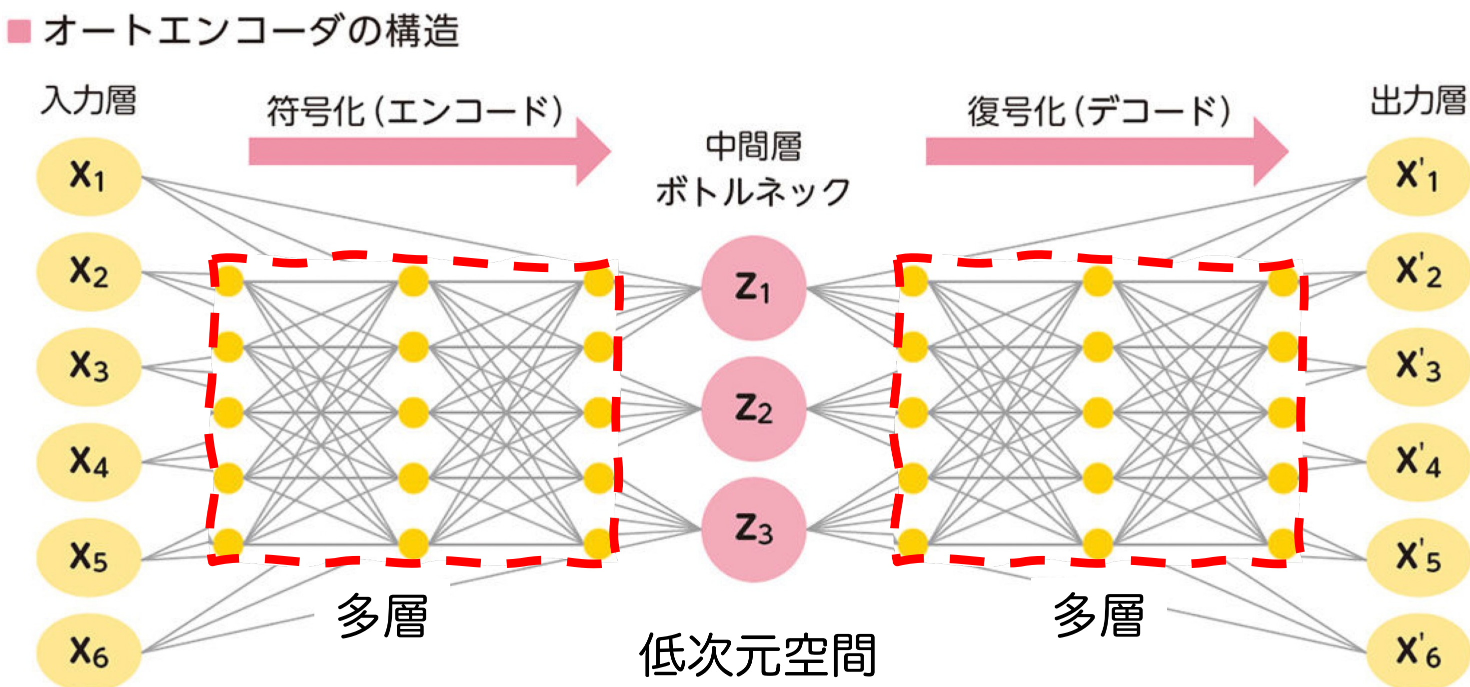
半径が定数の円柱表面にある束縛条件が利用されていない

円柱座標系 $\begin{Bmatrix} \theta_i \\ z_i \end{Bmatrix}$ 2次元

半径が定数の円柱表面にある束縛条件が利用する潜在変数

座標変換
$$\begin{Bmatrix} x_i = r\cos\theta_i + \epsilon_x \\ y_i = r\sin\theta_i + \epsilon_y \\ z_i = z_i \end{Bmatrix}$$

オートエンコーダ（3）



低次元空間
（次元削減効果）

多次元空間
(重複情報あり)

重複情報なし＝潜在変数

多次元空間
（重複情報あり）

$$\vec{y} = \vec{x} + \vec{\epsilon}$$

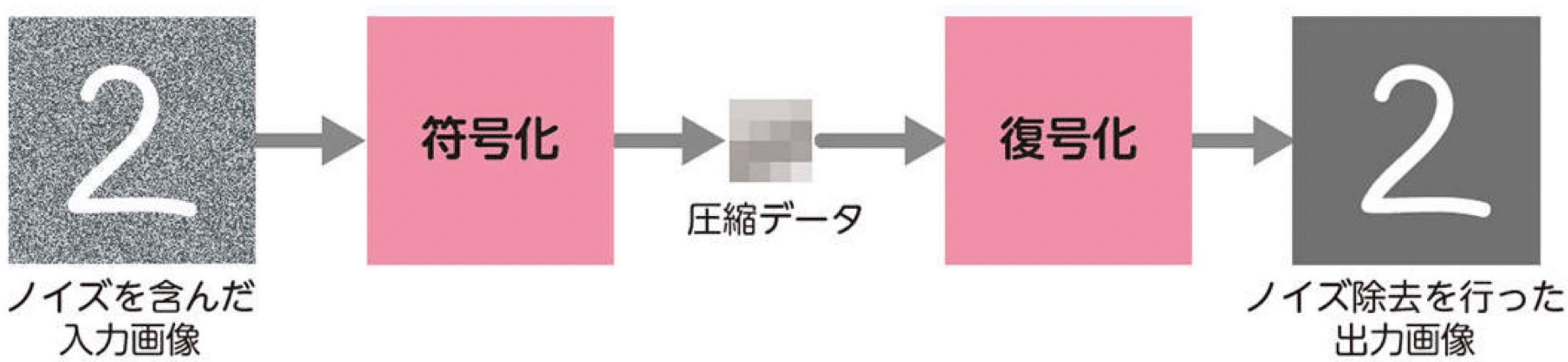
情報が正しく復元できるように圧縮したいが、原理的にできない状況

オートエンコーダの例

Deep Clustering with Convolutional Autoencoders

Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin

Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science, vol 10635. Springer.



多次元空間
(重複情報あり)

低次元空間
(次元削減効果)

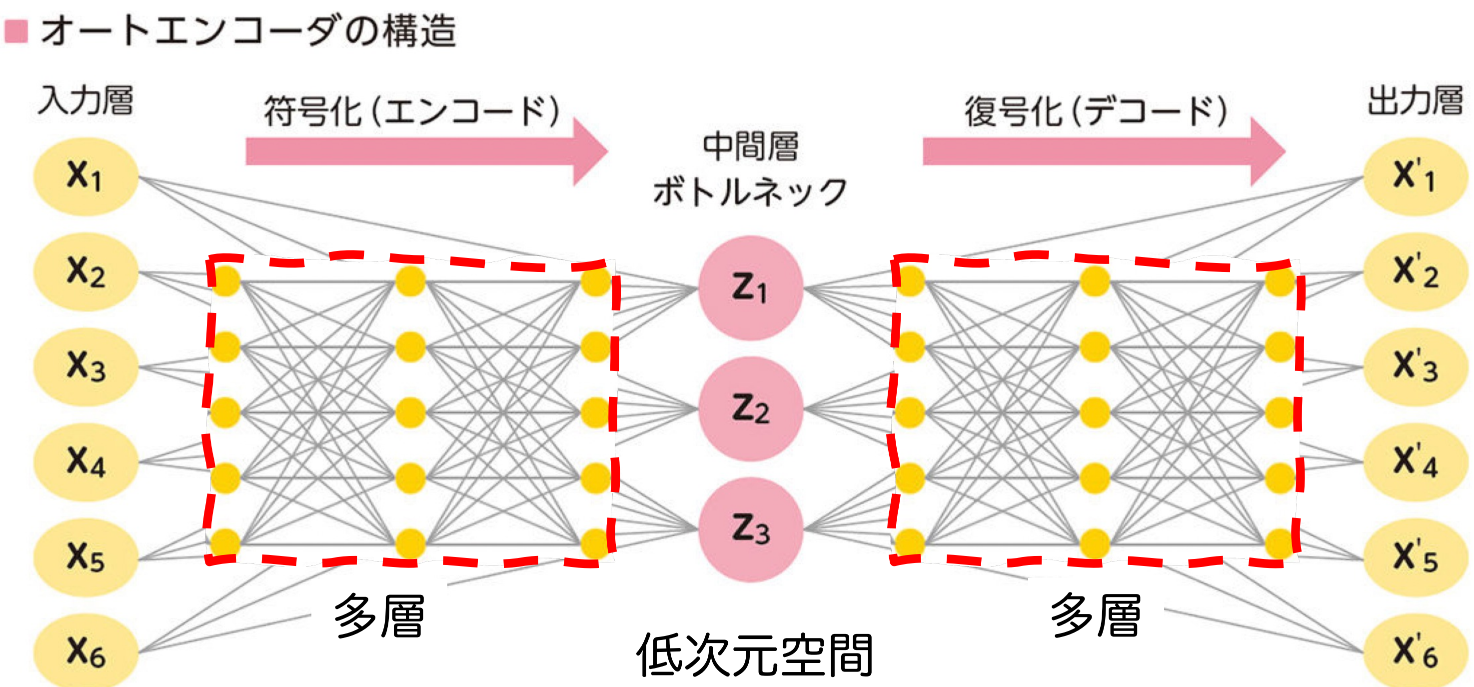
重複情報なし＝潜在変数

多次元空間
(重複情報あり)

$$\vec{y} = \vec{x} + \vec{\epsilon}$$

情報が正しく復元できるように圧縮したいが、原理的にできない状況

オートエンコーダ（2）



低次元空間
（次元削減効果）

多次元空間
(重複情報あり)

重複情報なし＝潜在変数

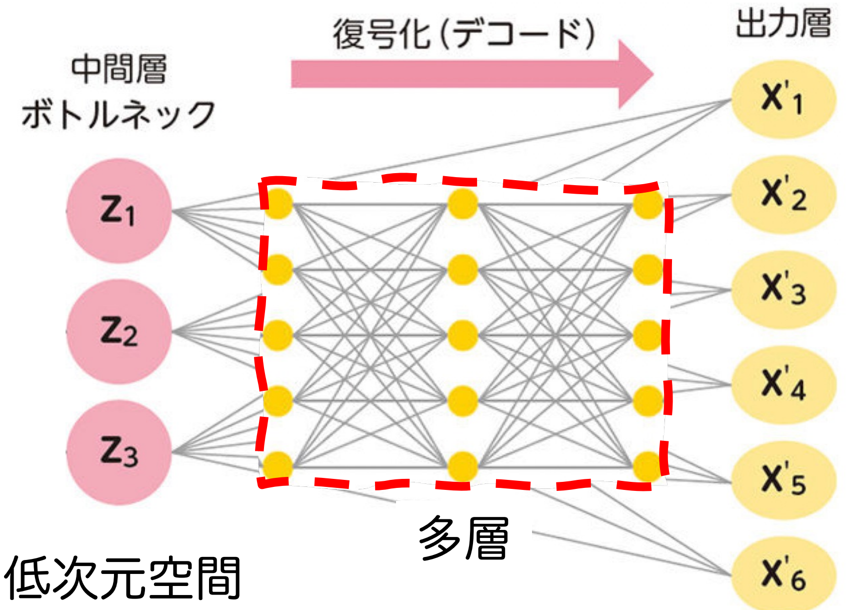
多次元空間
（重複情報あり）

$\vec{y} = \vec{x}$

情報が正しく復元できるように圧縮する

オートエンコーダによるデータ生成

符号化情報を入力すれば入力と
同じ性質を持つ出力を生成できる
データ生成モデル
(期待するが、そうとは限らない！)



低次元空間
(次元削減効果)

多次元空間
(重複情報あり)

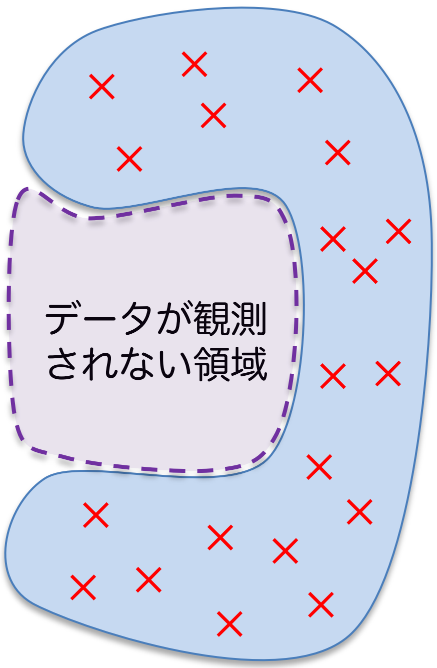
重複情報なし＝潜在変数

出典：
東京大学
松尾研究ホームページ

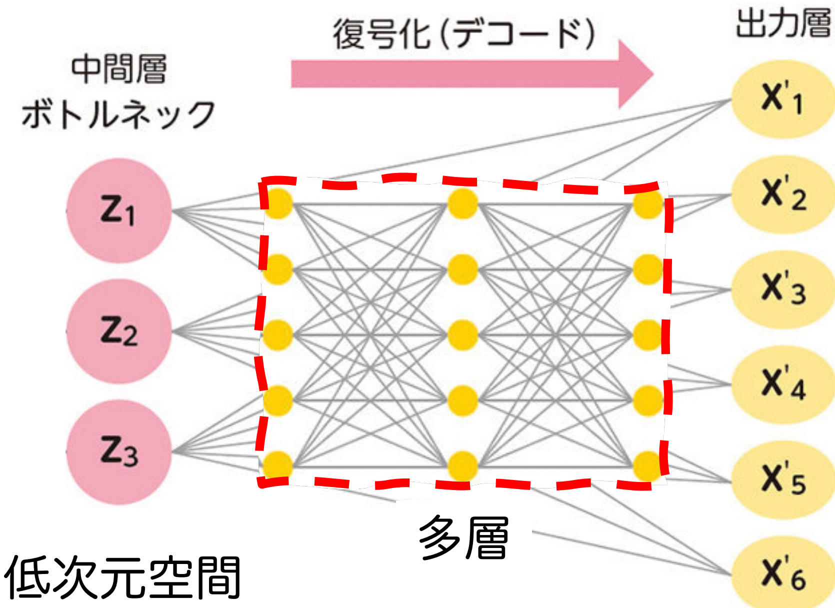
フェイクデータなどを生成できる

オートエンコーダによるデータ生成

潜在変数 $\vec{z}$ 空間



× 訓練データ点



低次元空間
（次元削減効果）

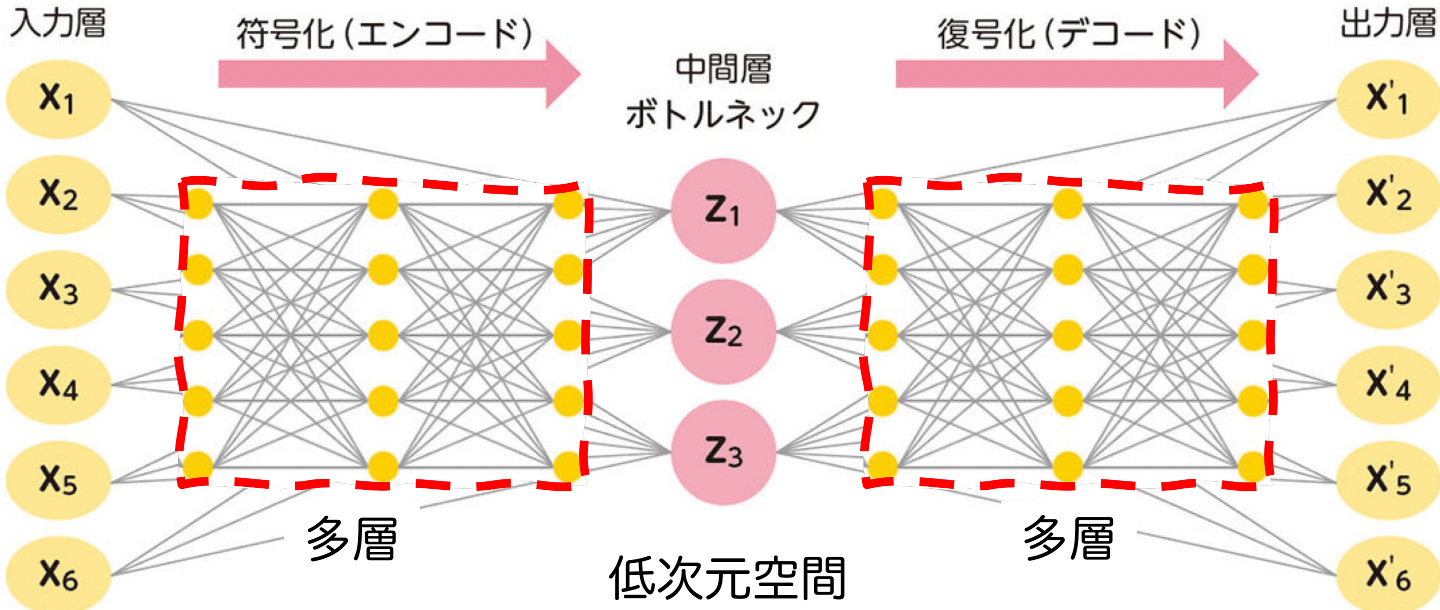
多次元空間
（重複情報あり）

重複情報なし＝潜在変数

外挿問題：ありえないデータの生成？

変分オートエンコーダ (VAE)

$\vec{y} = \vec{x}$



多層

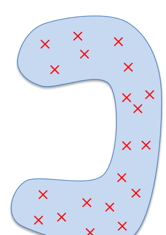
多層

低次元空間
(潜在変数)

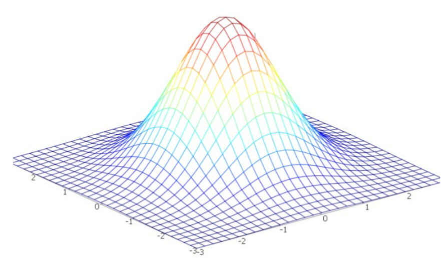
多次元空間

多次元空間

潜在変数空間
分布の形を強制
して学習する



強制



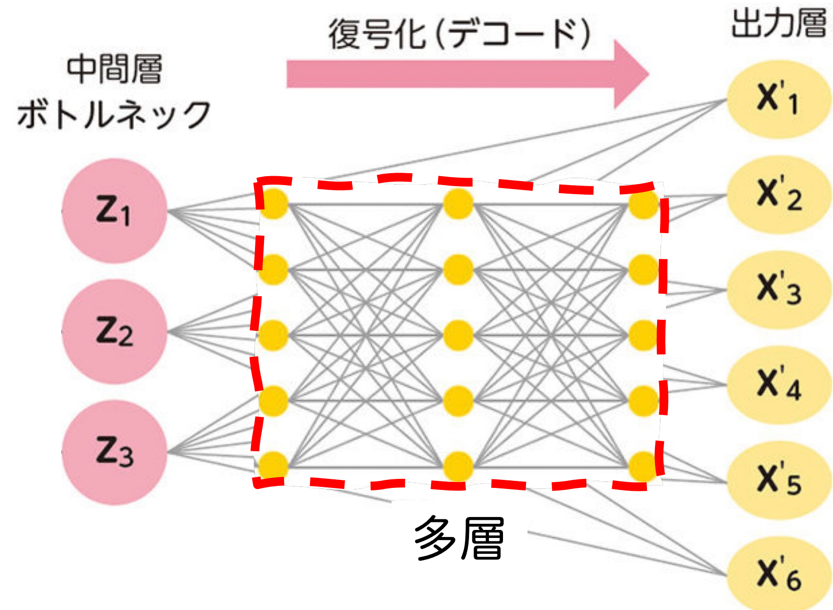
外挿問題：
ありえないデータの
生成をさけたい

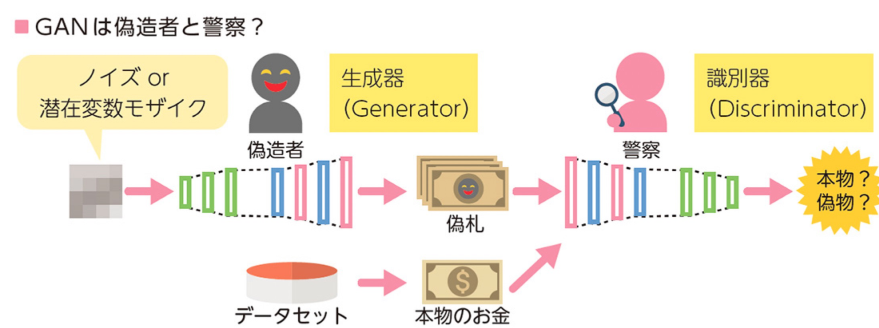
敵対的生成ネットワークGAN

符号化情報を入力すれば入力と
同じ性質を持つ出力を生成できる
データ生成モデル

敵対的

実データとフェイクデータを
識別するモデルを同時に訓練する





データサイエンスの基礎

北陸先端科学技術大学院大学
Hieu-Chi Dam

2024年7月8日（第2部 19:30 ~ 20:45）

目次

1. イントロ
   i. データサイエンスの紹介
   ii. 推論方式の紹介
   iii. 数理統計と機械学習の紹介

2. データサイエンスの材料科学への応用事例紹介

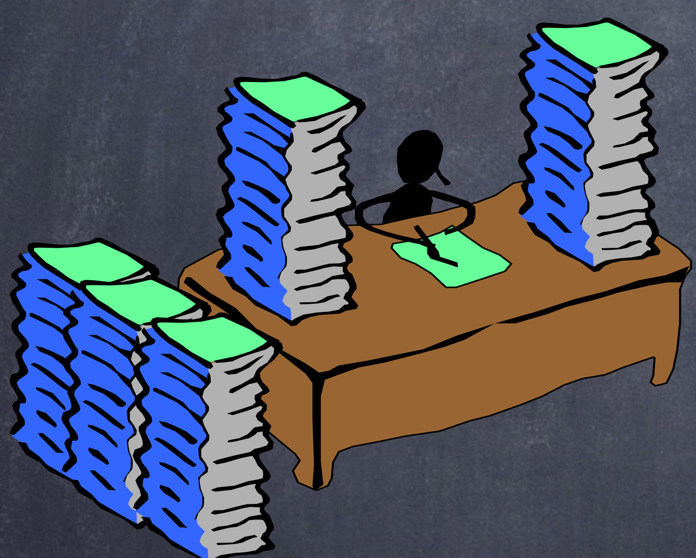
🌿 Data analytics process

「5 iterative steps」

Knowledge discovery and Data mining

可視化

The automatic extraction of non-obvious, hidden knowledge from large volumes/complicated data



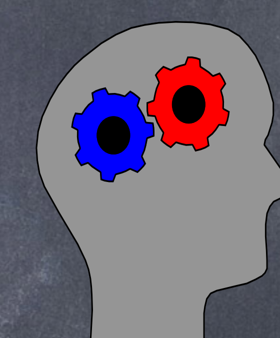
① Data

collection and
representation

② Learning/mining
from data

machine learning and
data mining algorithms

可視化



③ Knowledge

representation
and evaluation

潜む知見

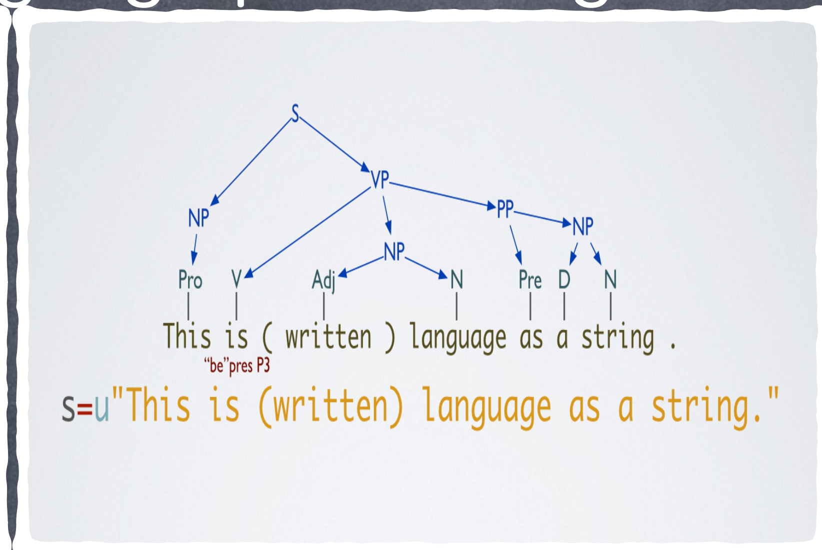
人間に理解可能

① Data

collection and representation

Problem setting
Data instance

Text and Natural language processing



Raw data : characters

character?

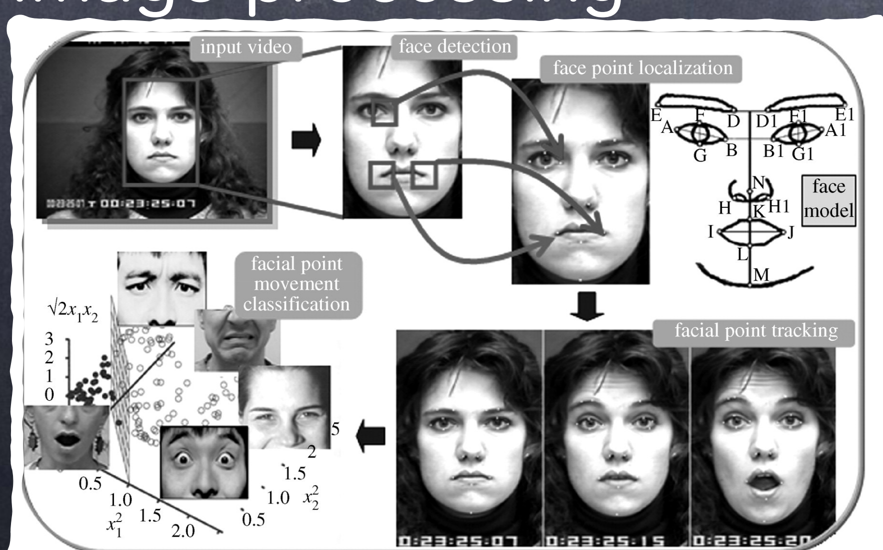
word?

sentence?

paragraph?

document?

Image processing



Raw data : pixels

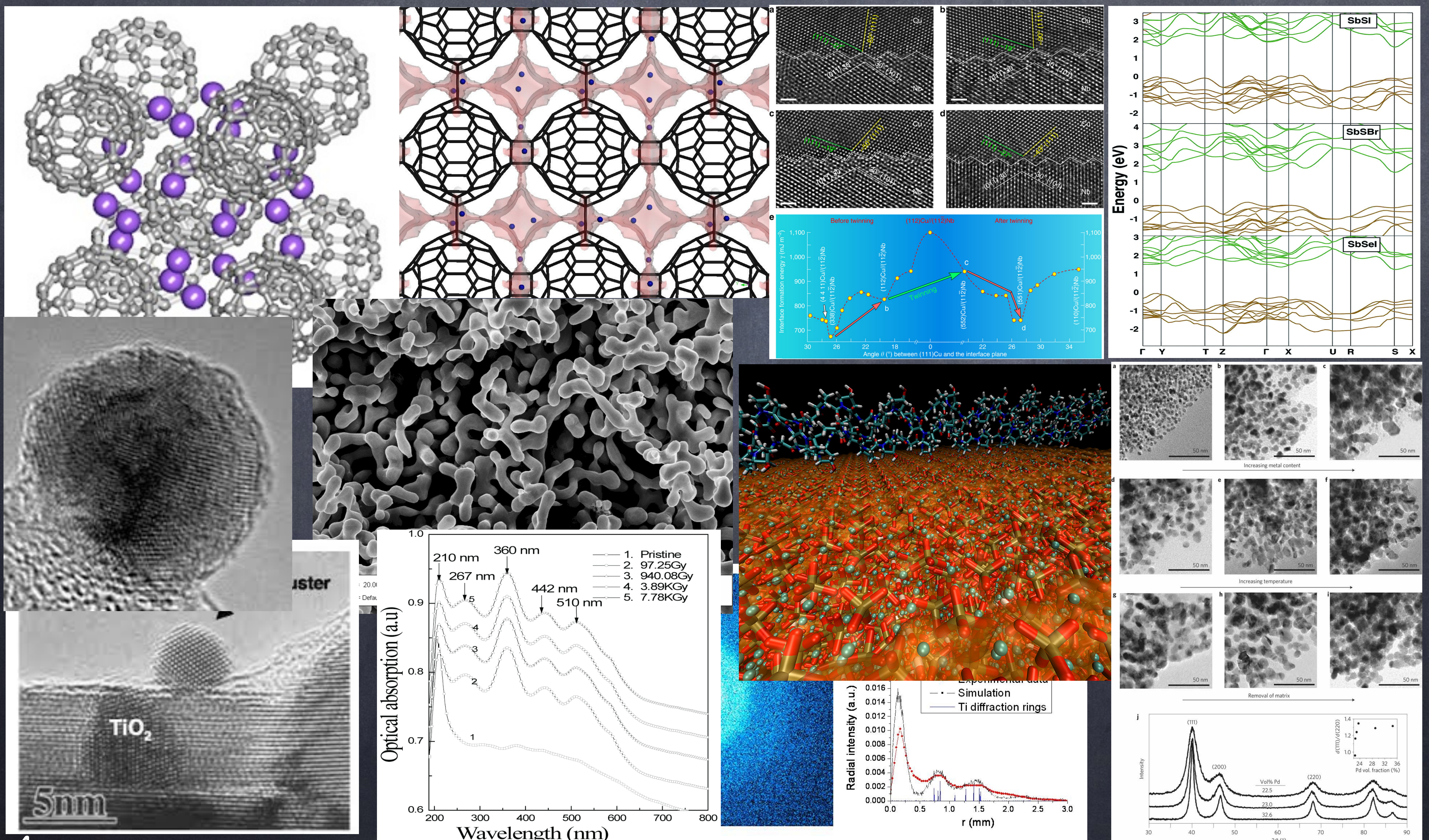
pixel?

pixel window?

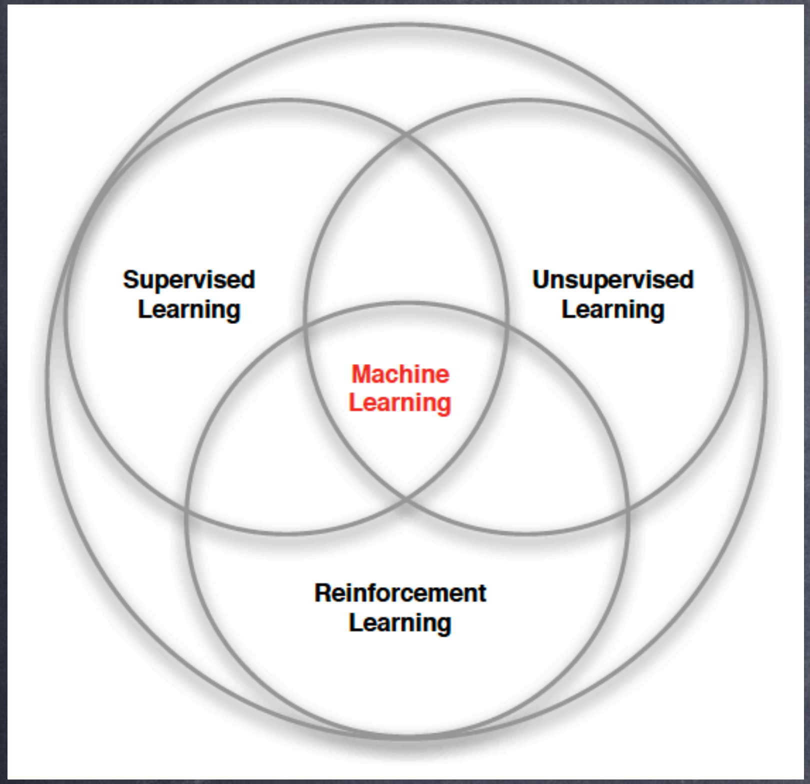
gradient?

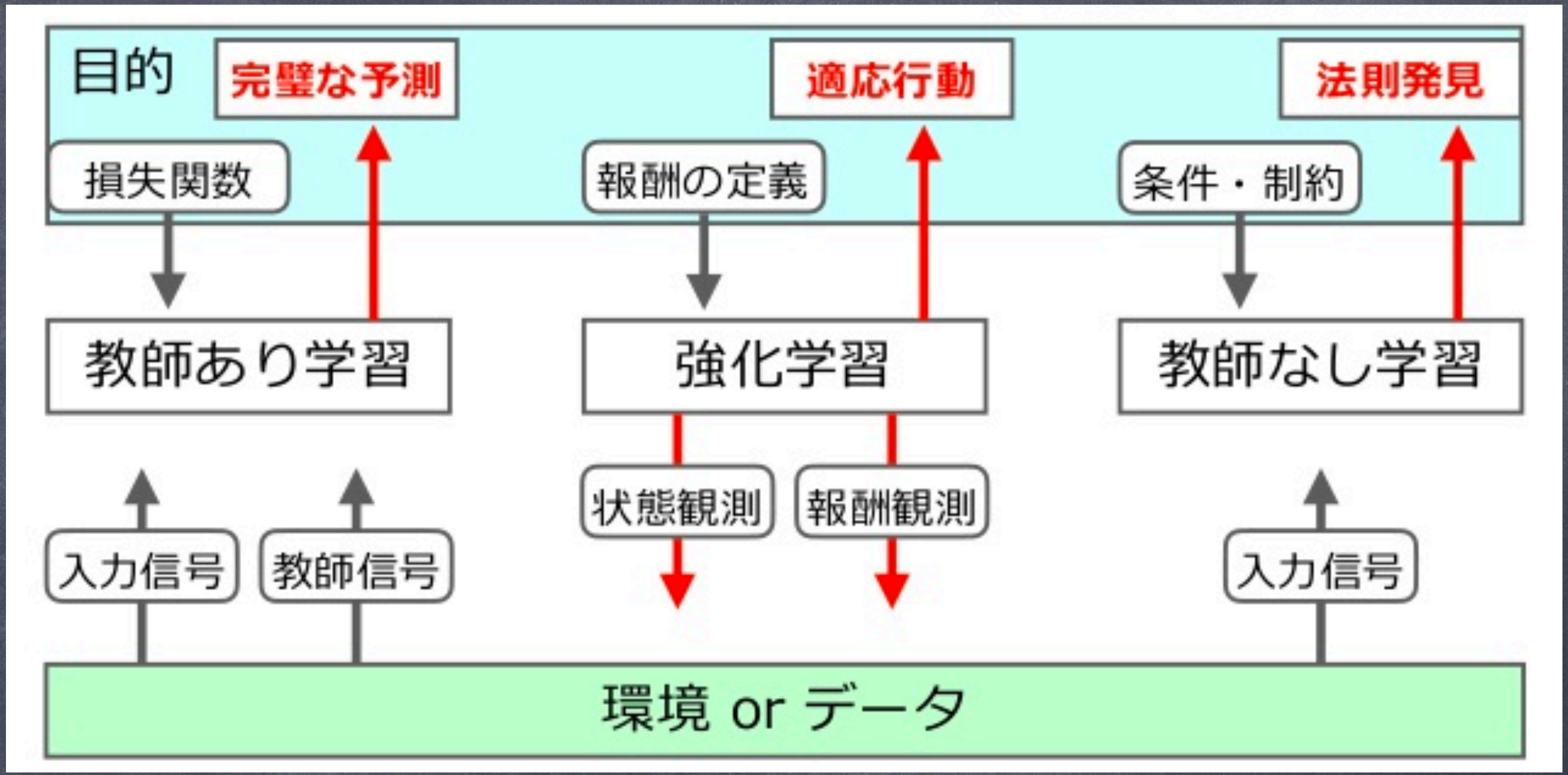
image?

Data in Materials science



Branches of
Machine learning





How to "translate" a materials science problem into machine learning/data mining language?

Representation, similarity measure, and learning

www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf
DeNA Co. Ltd. 甲野 佑 SlideShare

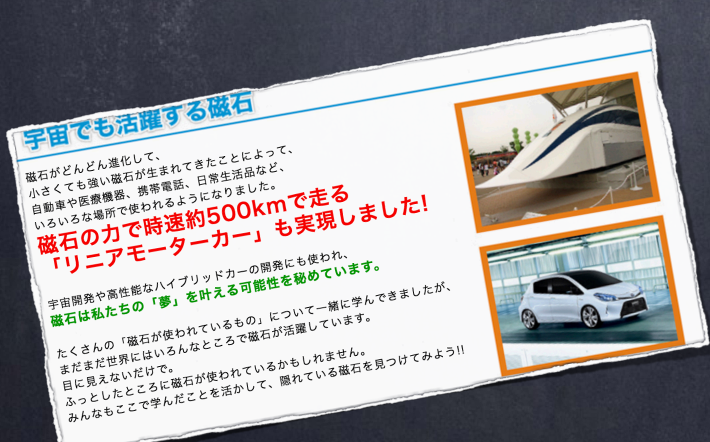
🌿 磁石材料のキュリ温度データの事例

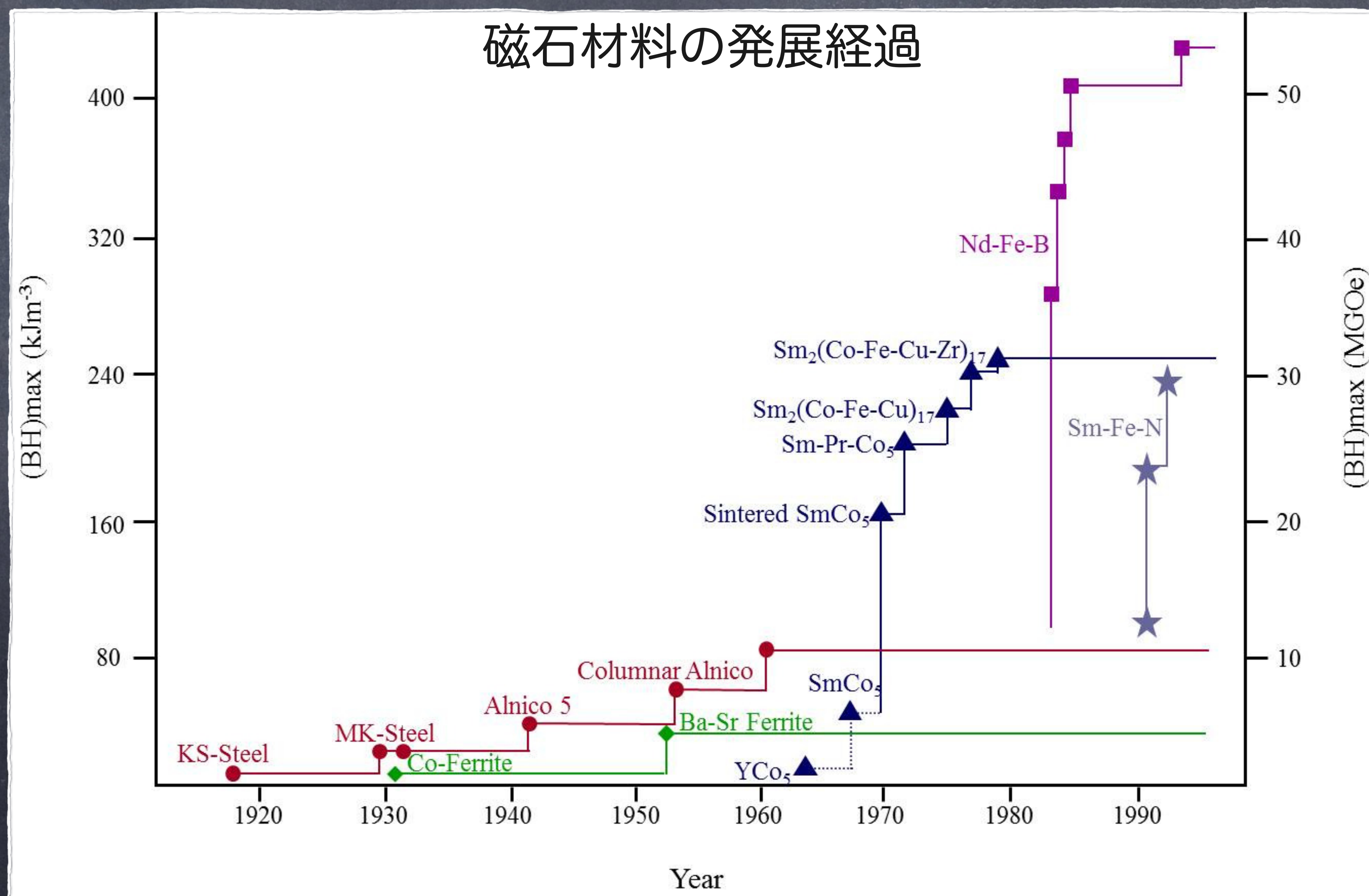
磁力の強い磁石ランキング

1位　ネオジム
2位　サマリウムコバルト
3位　アルニコ
4位　フェライト

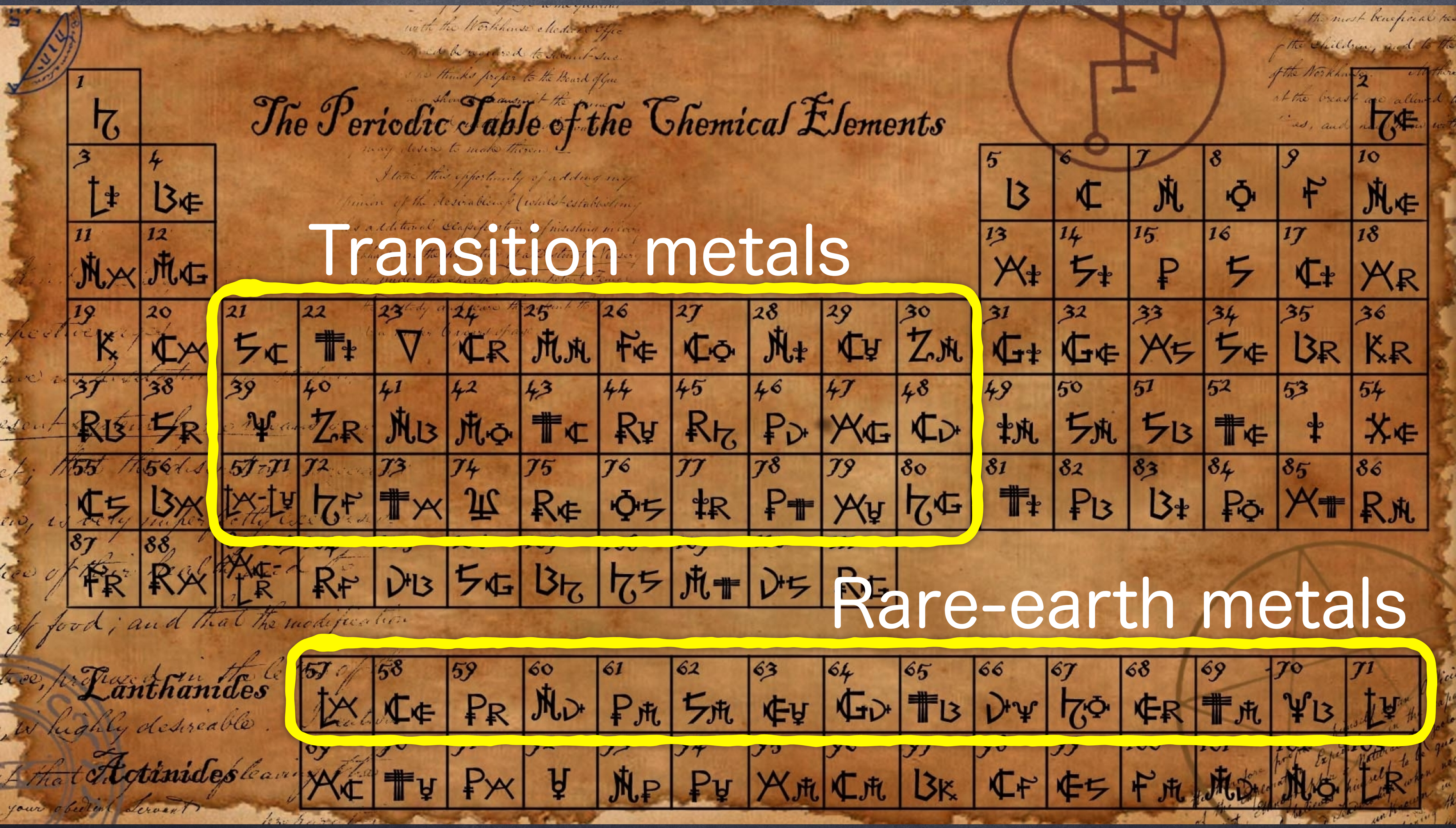
熱に強い磁石ランキング

1位　アルニコ（500℃）
2位　サマリウム・コバルト（300℃）
3位　フェライト（200℃）
4位　ネオジム（80℃）





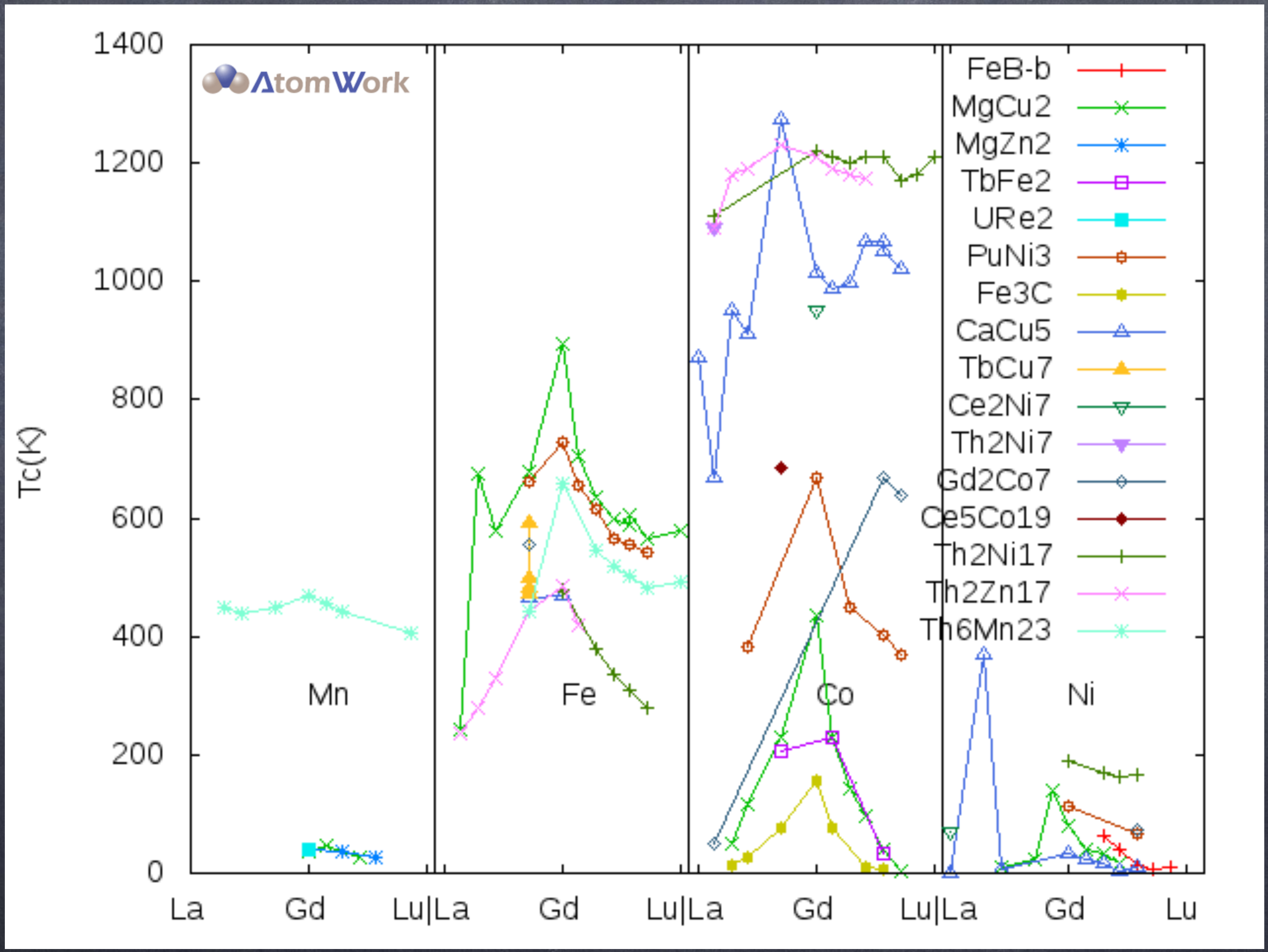
https://www.birmingham.ac.uk/Documents/college-eps/metallurgy/research/Magnetic-Materials-Background/
https://www.magfine.org/magnetkids/secret/utility.html



The Periodic Table of the Chemical Elements
Transition metals
Rare-earth metals
Lanthanides
Actinides

Curie temperature of 3d-4f binary alloys

d-f bimetal alloys from literatures



生データを集めました！

MIの反復５ステップ

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

MIの反復５ステップ

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

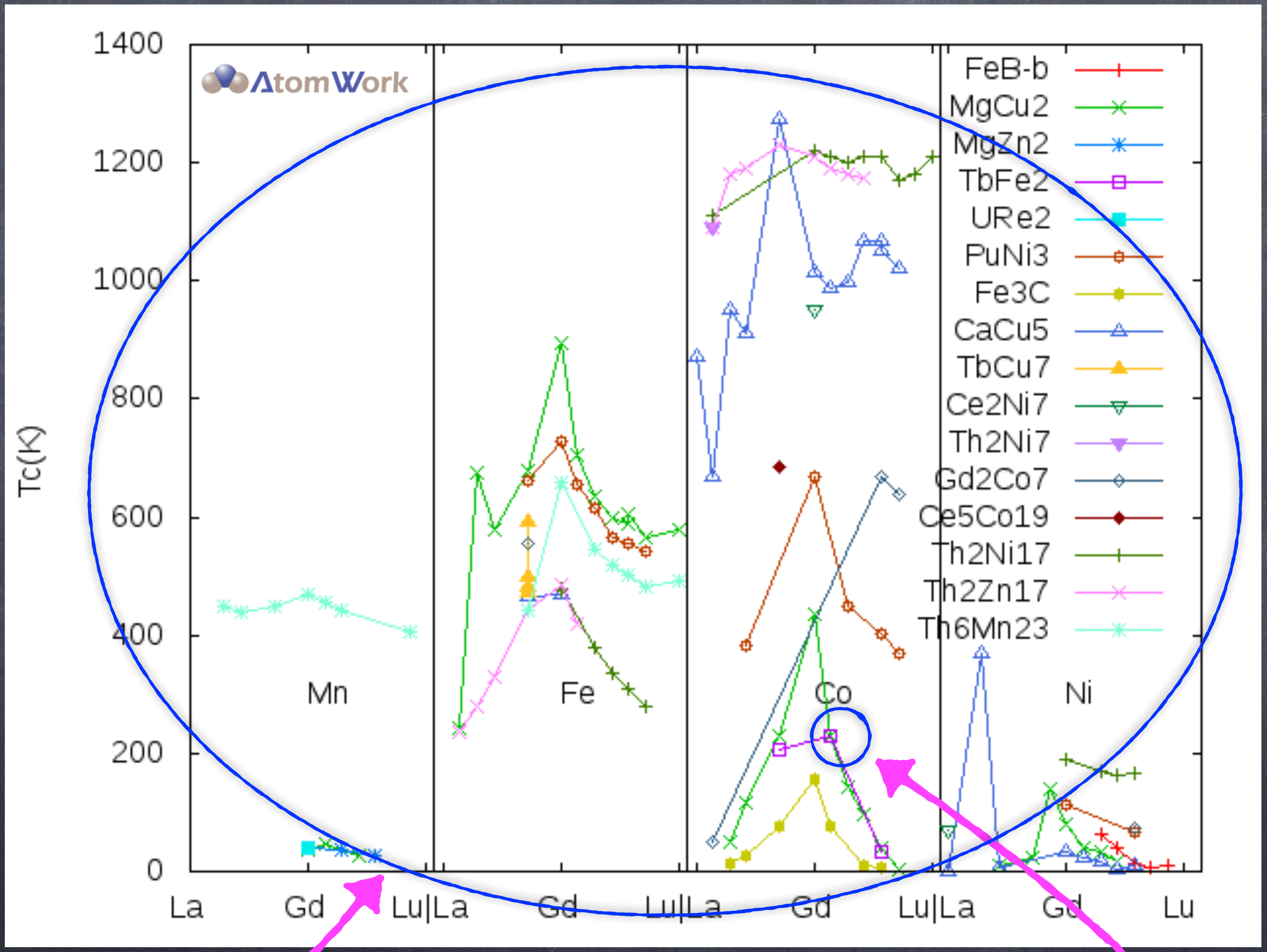
② 記述子の選定・適用

① 問題提起・問題の焼直し

まず
データインスタンスの定義
からスタート！

Curie temperature of 3d-4f binary alloys

d-f bimetal alloys from literatures



データ集合

１データインスタンス

問題提起：キュリ温度を予測したい

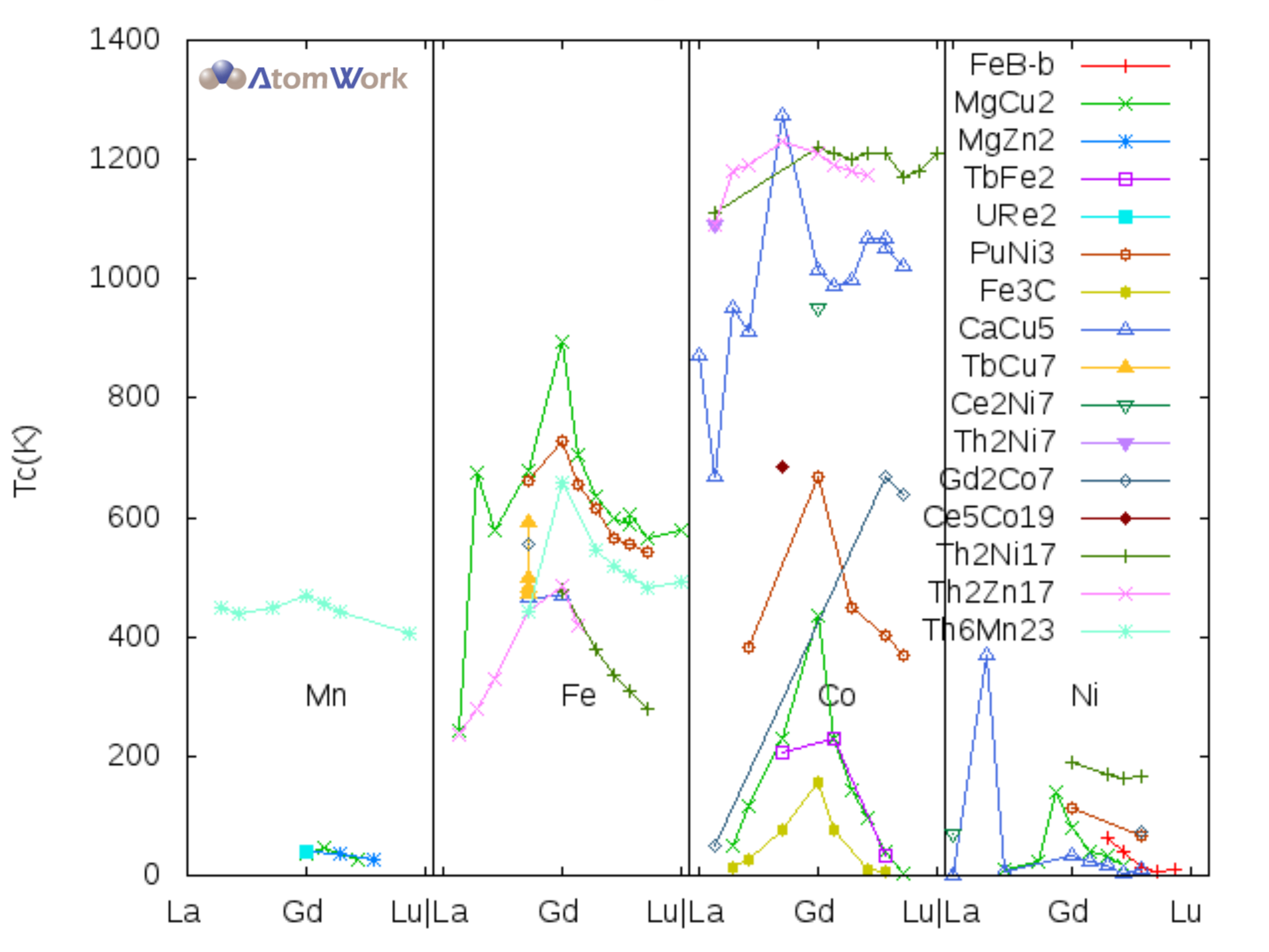
「どうなっている？」

「希土類の種類が
変わるとTcが変わる」

「どうなっている？」

「Tcが希土類金属の属性
の関数だ」

d-f bimetal alloys from literatures



$$T_C^{predict} = f(\Box, \Diamond, \triangle, \cdots)$$

MIの反復５ステップ

予測問題

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

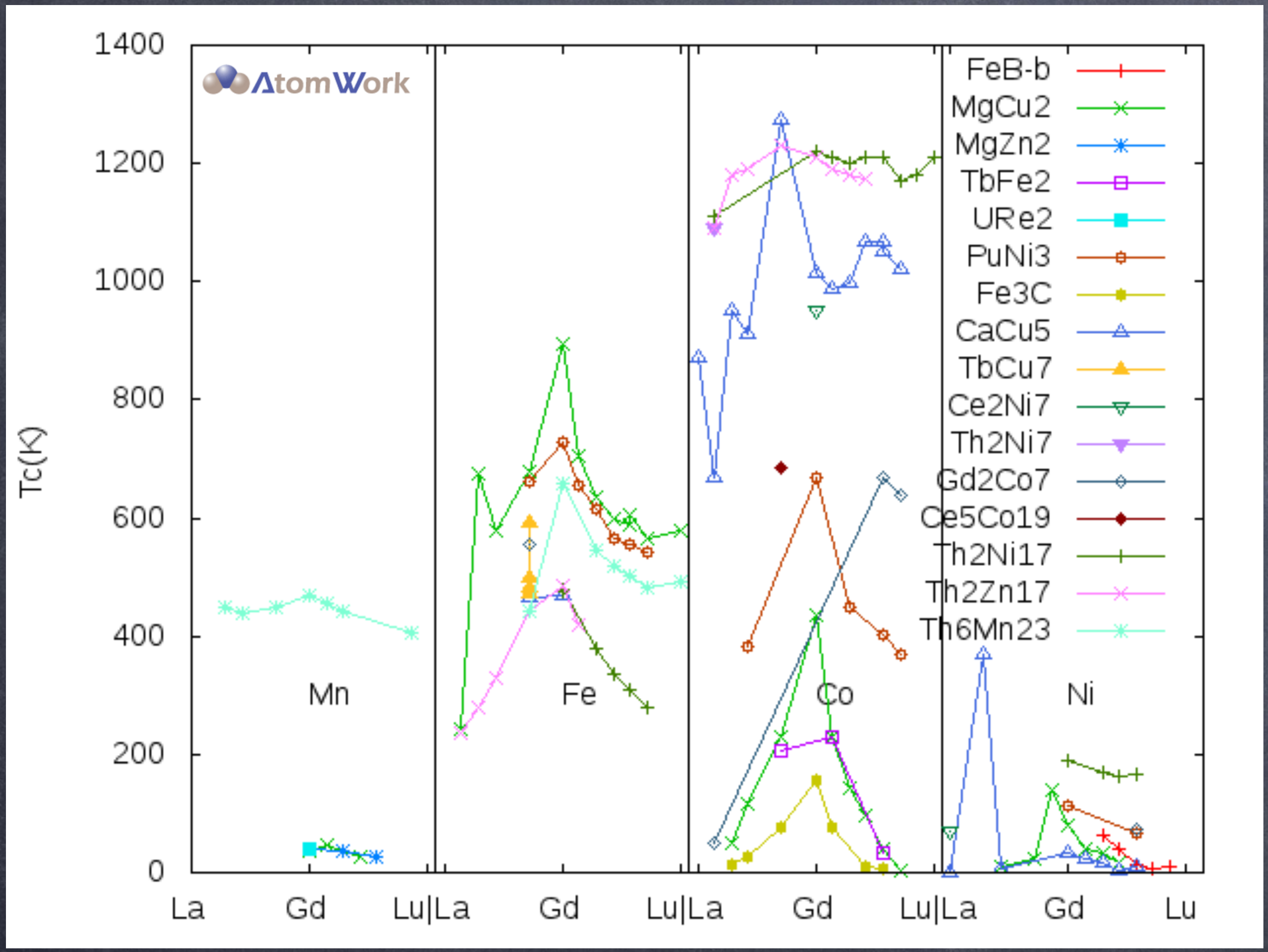
② 記述子の選定・適用

① 問題提起・問題の焼直し

解析用データの準備

２元合金情報の記述

d-f bimetal alloys from literatures



Material structure

$$\left\{\begin{array}{c} a,\, b,\, c,\, \alpha,\, \beta,\, \gamma \\ atom_1,\, x_1,\, y_1,\, z_1 \\ atom_2,\, x_2,\, y_2,\, z_2 \\ \vdots \\ atom_i,\, x_i,\, y_i,\, z_i \\ \vdots \\ atom_n,\, x_n,\, y_n,\, z_n \end{array}\right\}$$

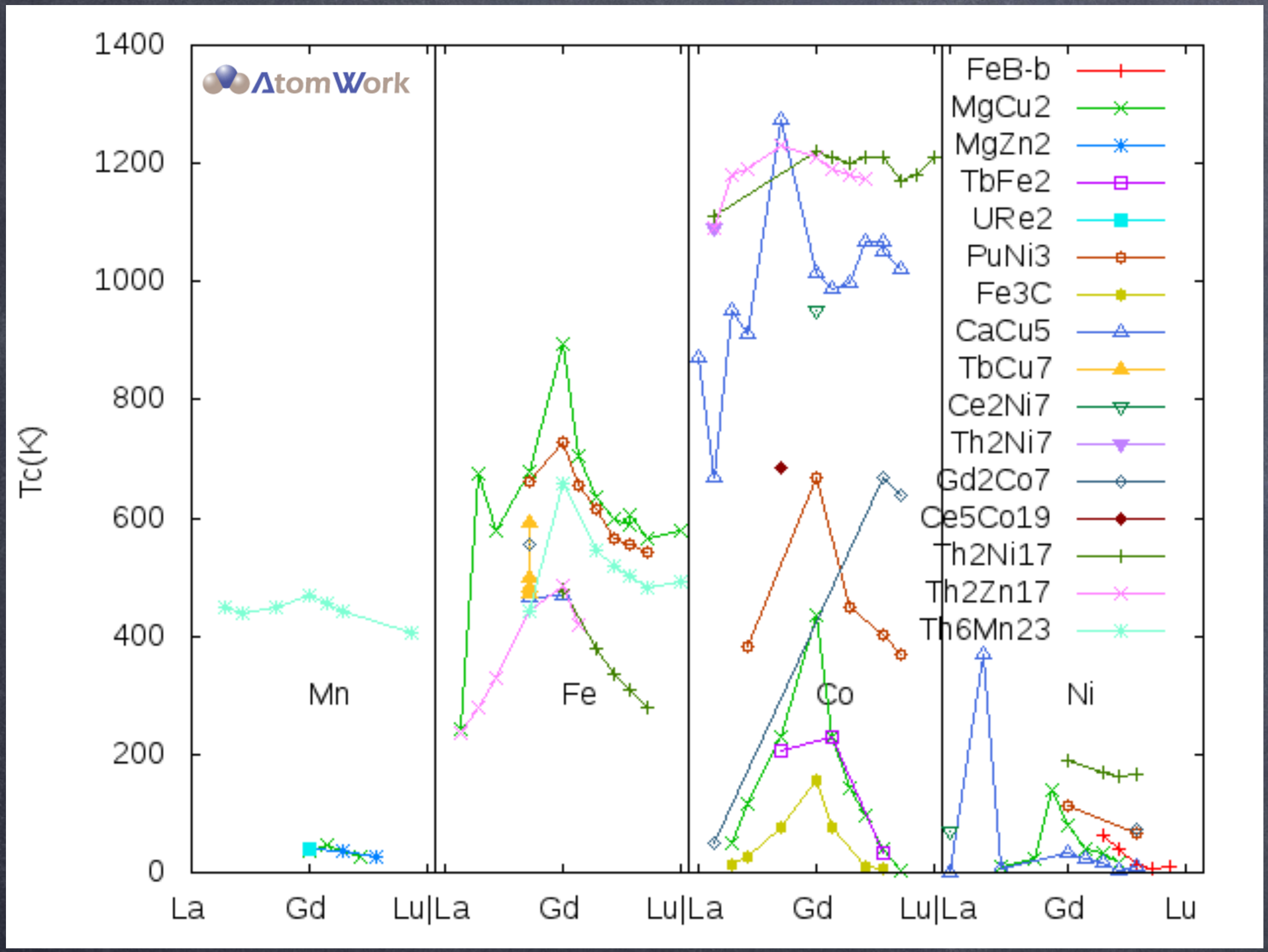
記述子

何が特徴なのか？

$$T_C^{predict} = f(\overbrace{\Box, \Diamond, \triangle, \cdots}^{})$$

２元合金データの記述子の選定

d-f bimetal alloys from literatures



まず，予測できるかをテストする

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

想定する記述子

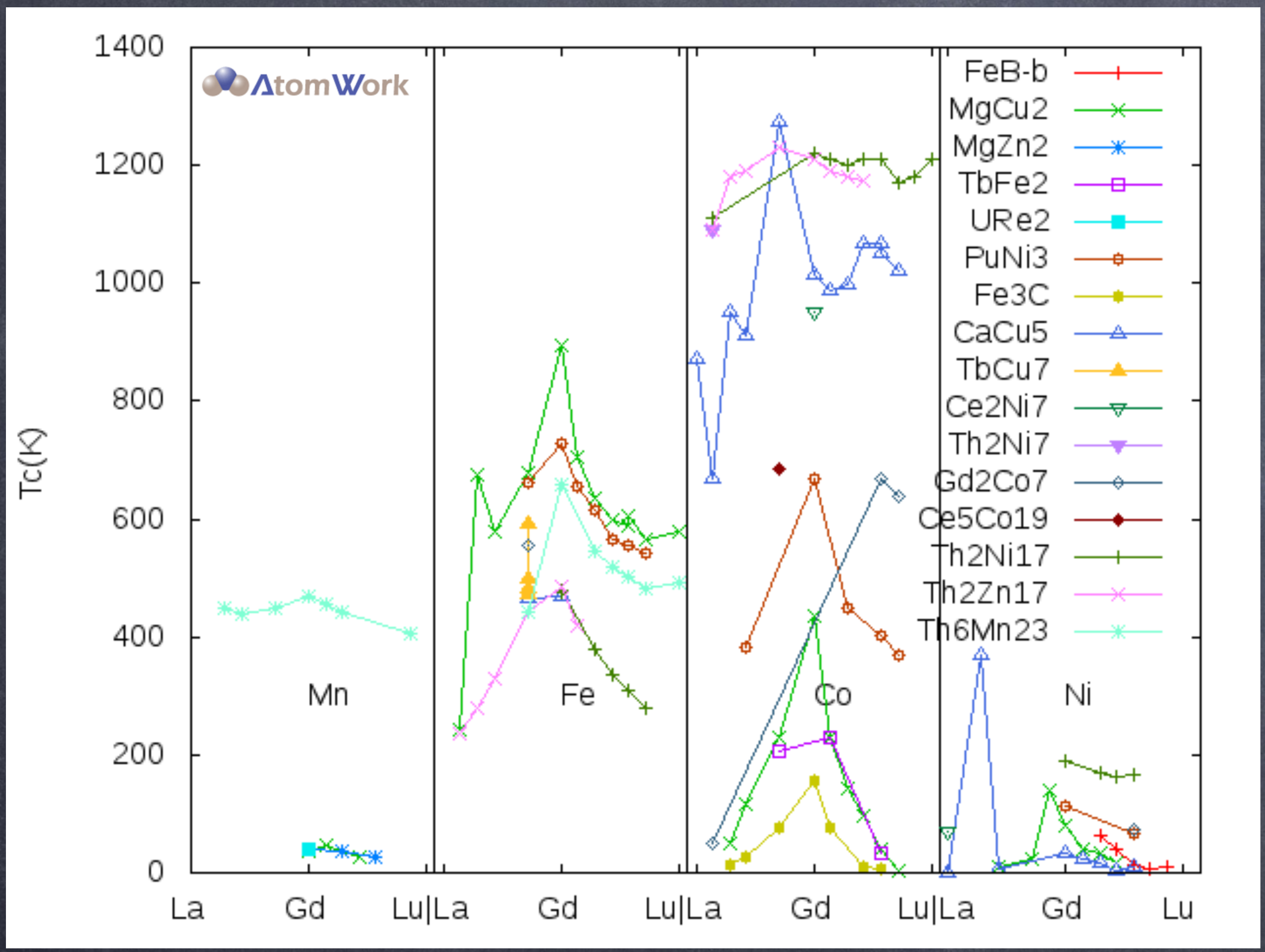
Structural information:
- R-R Nearest neighbor distance
- R-T Nearest neighbor distance
- T-T Nearest neighbor distance
- Number of R surrounding R
- Number of T surrounding T
- Number of R surrounding T
- Number of T surrounding T
- Concentration of R
- Concentration of T

Atomic information:
- Atomic Number of Z_R, Z_T
- Covalent Radius of r_R, r_T
- Electron Negativity of χ_R, χ_T
- Ionization Potential IP_R, IP_T
- Orbital angular momentum L_R, L_T
- Spin angular momentum S_R, S_T
- Total angular momentum J_R, J_T
- Landé g-factor g_{JR}, $J_R g_{JR}$, $J_R(1-g_{JR})$

Empirical descriptors

２元合金データの記述子の適用

d-f bimetal alloys from literatures



解析用データ

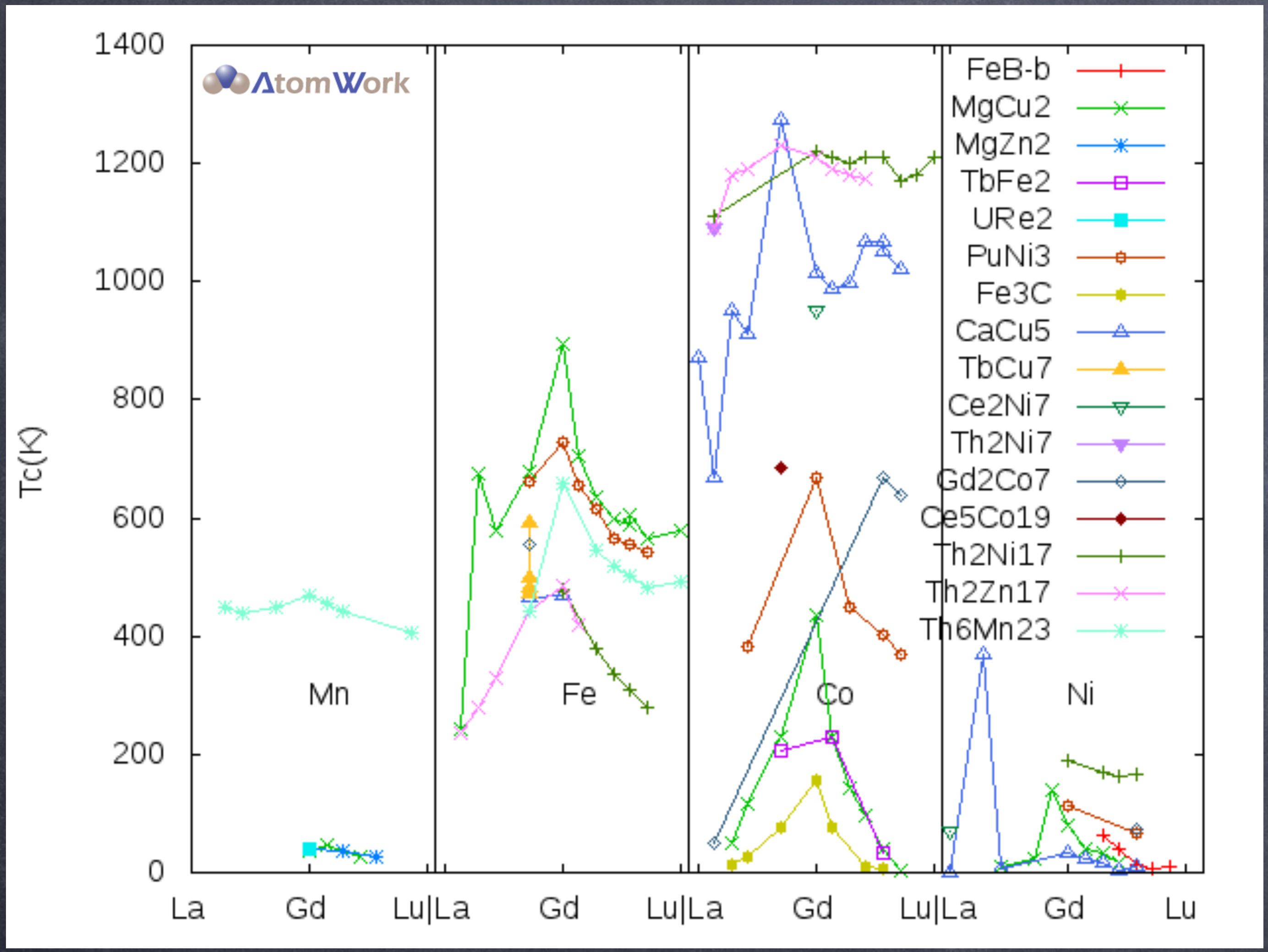
ID	d_1	d_2	…	d_n	T_C
合金 1	d_1^1	d_2^1	…	d_n^1	T_C^1
合金 2	d_1^2	d_2^2	…	d_n^2	T_C^2
…	…	…	…	…	…
合金 n	d_1^n	d_2^n	…	d_n^n	T_C^n

まず，予測できるかをテストする

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

２元合金データの記述子の適用

d-f bimetal alloys from literatures



まず，予測できるかをテストする

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

解析用データ

ID	d_1	d_2	…	d_n	T_C
合金 1	d_1^1	d_2^1	…	d_n^1	T_C^1
合金 2	d_1^2	d_2^2	…	d_n^2	T_C^2
…	…	…	…	…	…
合金 *n*	d_1^n	d_2^n	…	d_n^n	T_C^n

（背後にある隠れた前提）

データ対象間の類似度は

ベクトル演算によって定量評価する

MIの反復５ステップ

予測問題 → 重回帰解析

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

MIの反復５ステップ

母集団 ＆ 標本

重相関関係 ＆ 推定問題

予測問題 → 重回帰解析

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

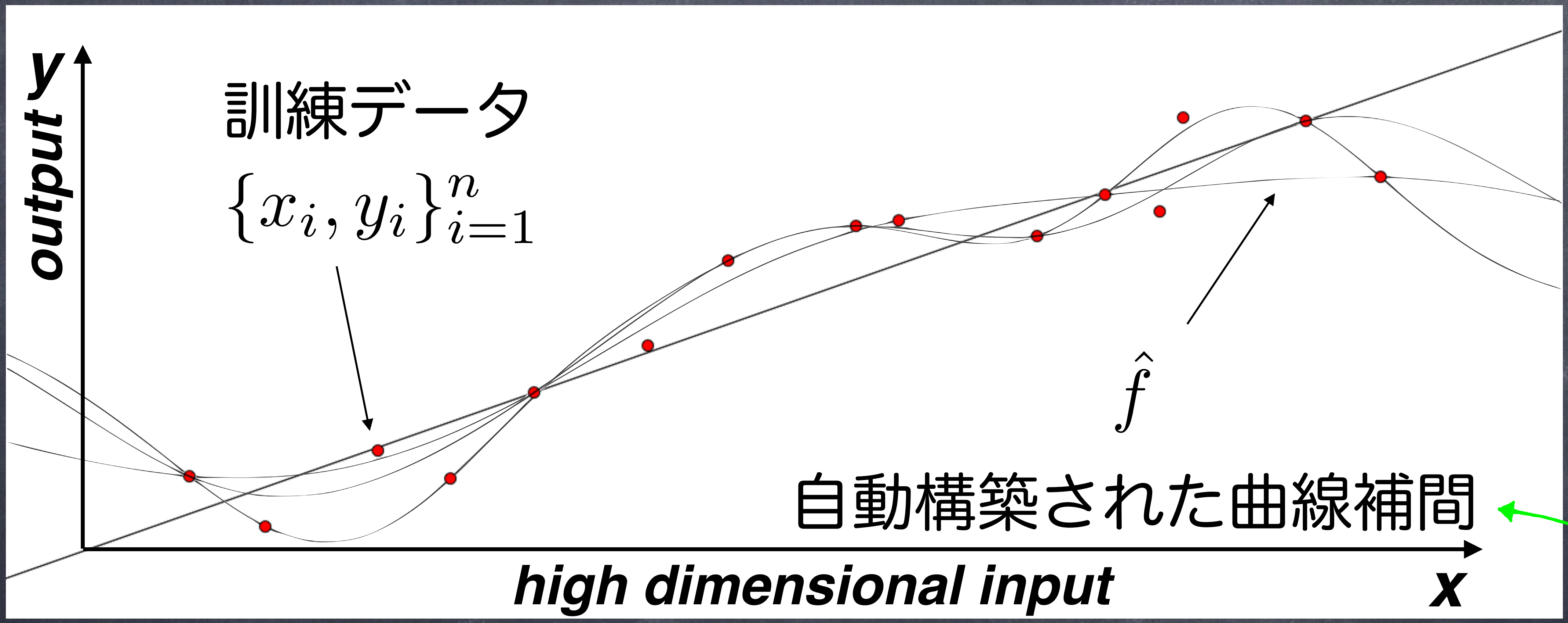
③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

- 回帰解析
  - 線形回帰
    - リッジ回帰
    - LASSO回帰
  - 非線形回帰
    - カーネル回帰
    - ガウス過程回帰
    - ニューラルネット

Supervised learning:
High dimensional curve fitting

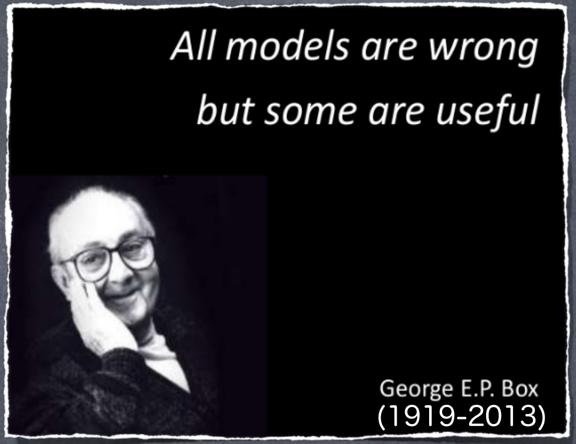


内挿的な予測技術

Tools libraries:

scikit-learn, Weka, DeepLearning (Tensorflow, pyTorch, Chainer, …)
Spark ML, Apple CoreML, Google Cloud AI/Cloud ML, Amazon ML, …

Regression and Data fitting



Data $\{\mathbf{X_i}, \mathbf{Y_i}\}$

Modeling the"relation" $\mathbf{X} \leftrightarrow \mathbf{Y}$ Correlation relation Hypothesis

$\mathbf{Y} \sim \mathbf{f_1}(\mathbf{X})$ $\mathbf{Y} \sim \mathbf{f_2}(\mathbf{X})$ $\mathbf{Y} \sim \mathbf{f_k}(\mathbf{X})$

Data fitting & parameter tuning

Model selection

$\mathbf{Y} = \mathbf{f_1^{\alpha_1,\beta_1,\gamma_1}}(\mathbf{X})$ $\mathbf{Y} = \mathbf{f_2^{\alpha_2,\beta_2,\gamma_2}}(\mathbf{X})$ $\boxed{\mathbf{Y} = \mathbf{f_k^{\alpha_k,\beta_k,\gamma_k}}(\mathbf{X})}$

MIの反復５ステップ

母集団 ＆ 標本

重相関関係 ＆ 推定問題

予測問題 → 重回帰解析

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

回帰解析

線形回帰

リッジ回帰

LASSO回帰

非線形回帰

カーネル回帰

ガウス過程回帰

ニューラルネット

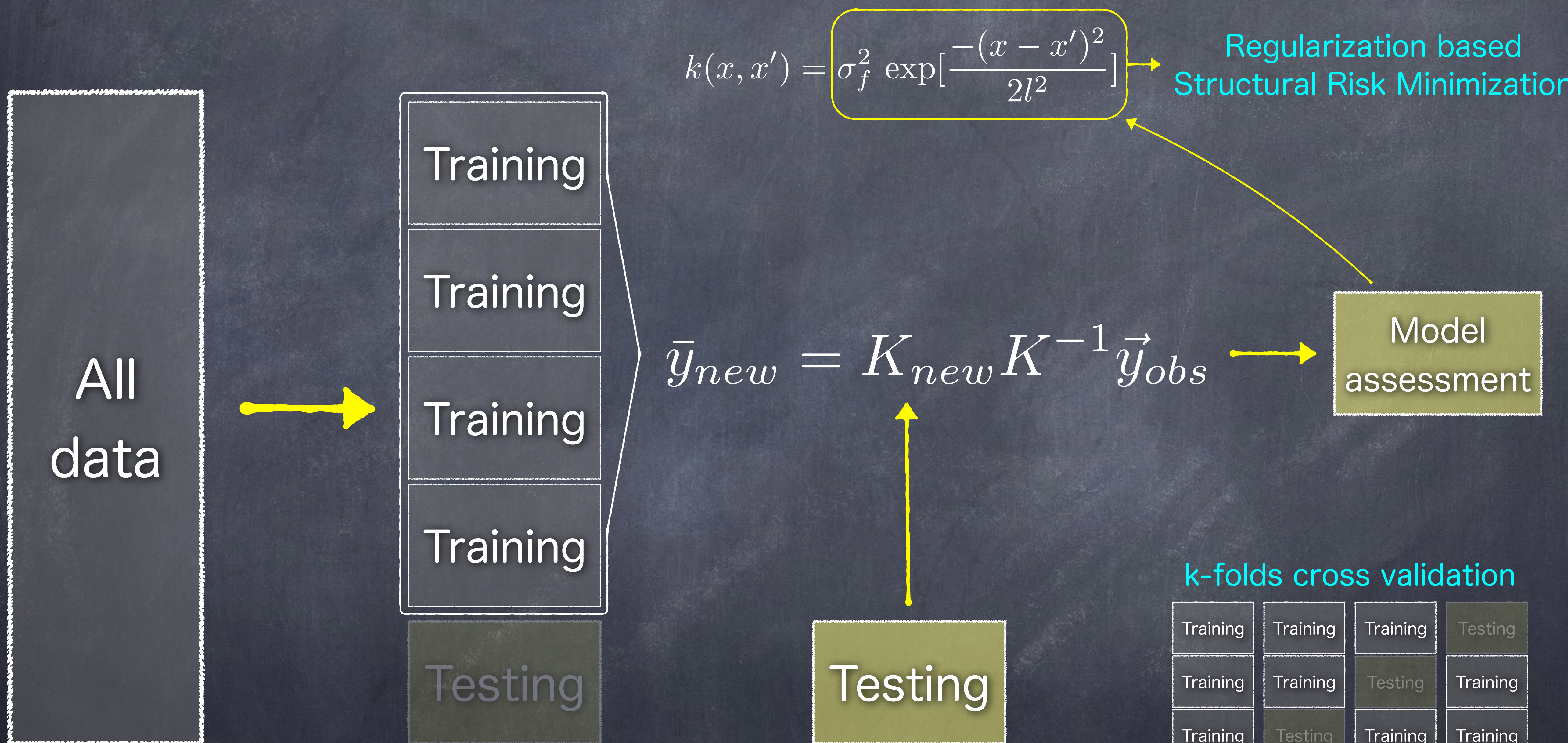
Model assessment

Statistical estimation
交差検定法

1. Predictive accuracy - 予測精度
2. Speed - 予測速度
3. Robustness - 頑健性
4. Scalability - 拡張性
5. Interpretability - 解釈容易性

"Use and Abuse of Regression", George Box,
活用 乱用 Technometrics, Vol. 8, No. 4, (1966)

Model assessment by cross validation



$$k(x,x') = \boxed{\sigma_f^2 \exp[\frac{-(x-x')^2}{2l^2}]}$$

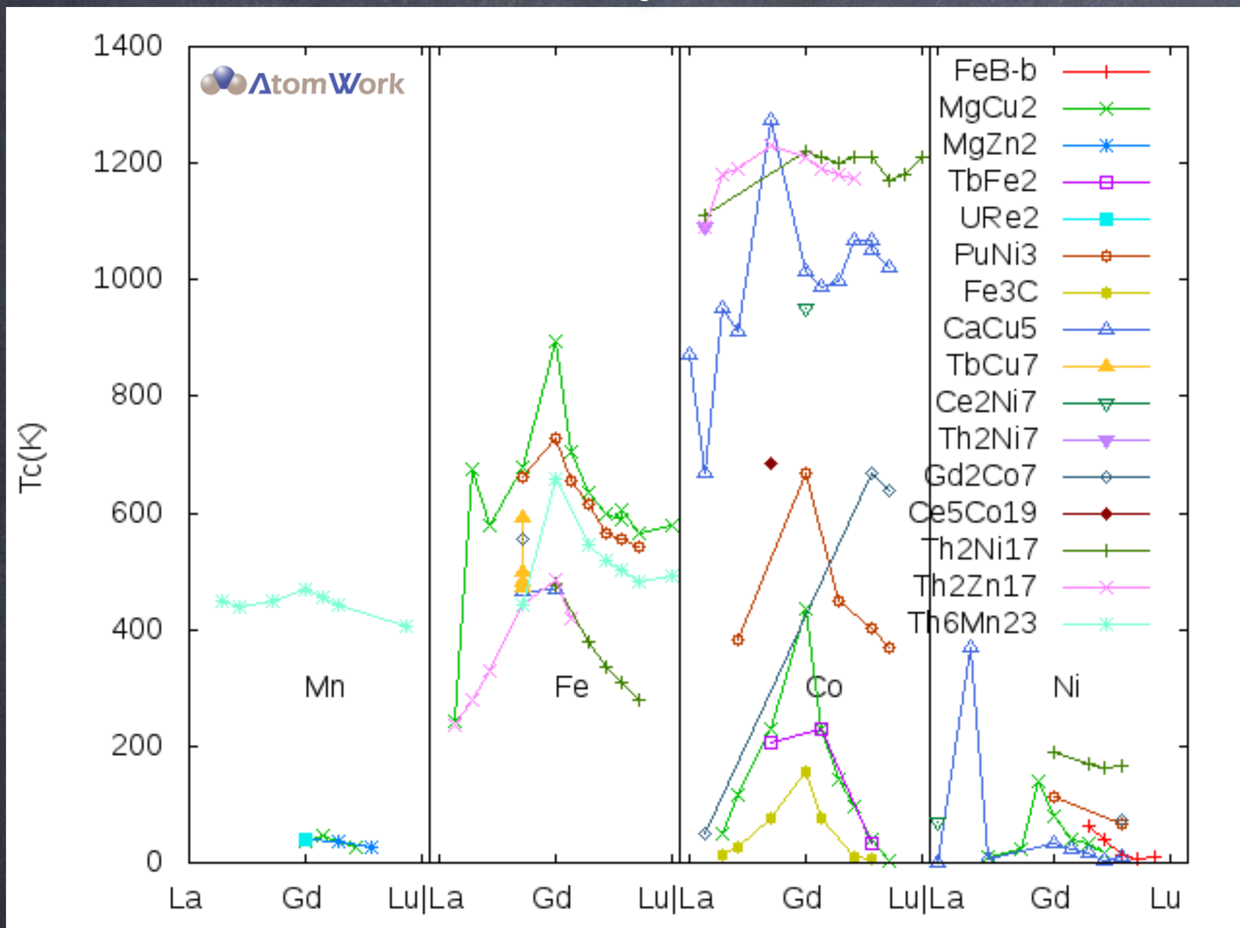
$$\bar{y}_{new} = K_{new} K^{-1} \vec{y}_{obs}$$

k-folds cross validation

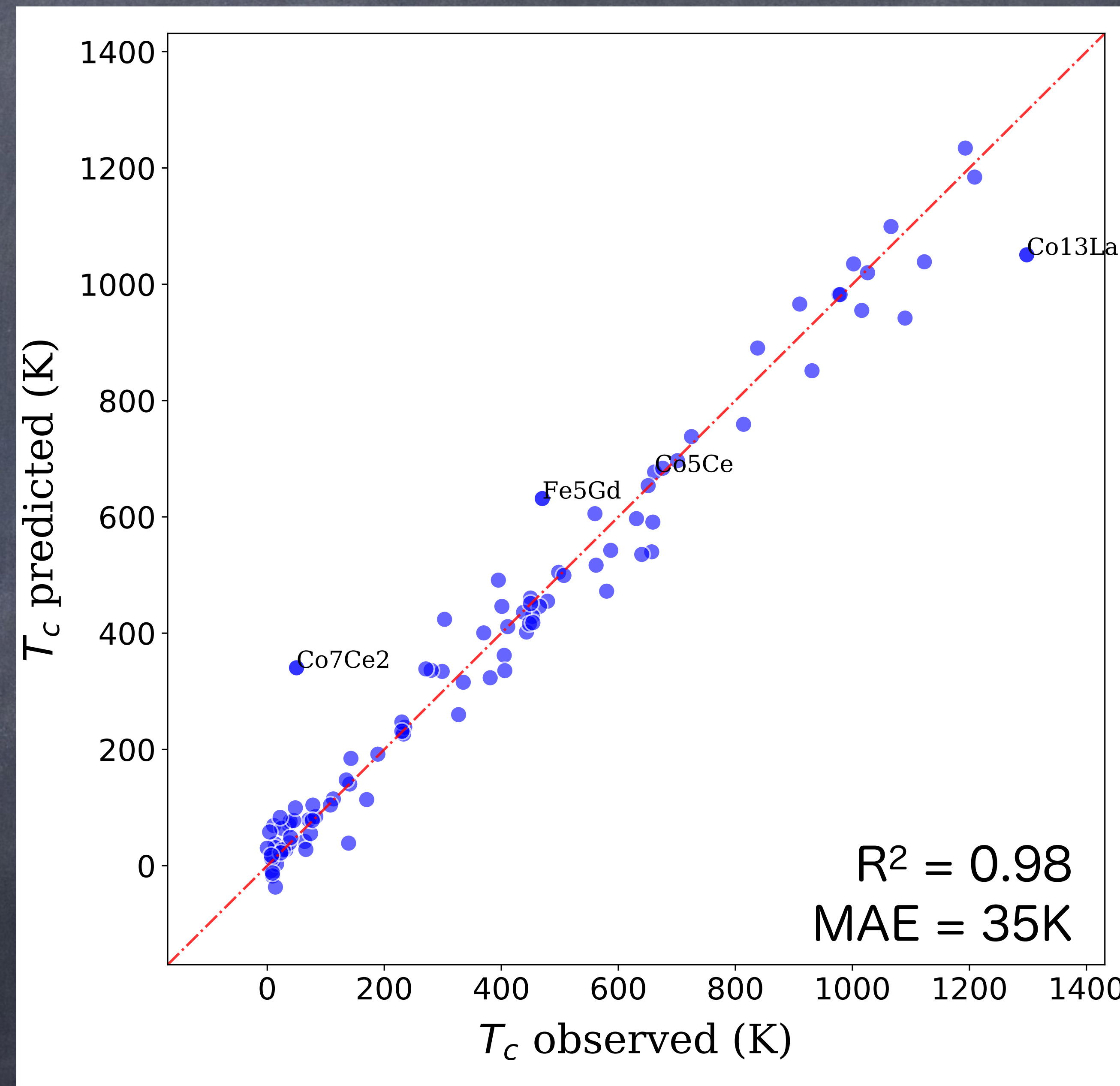
Training	Training	Training	Testing
Training	Training	Testing	Training
Training	Testing	Training	Training
Testing	Training	Training	Training
Training	Training	Training	Training

Prediction of Curie temperature

108 d-f bimetal alloys from literatures



交差検定によるTc予測モデルの評価



交差検定　　帰納法

予測による反証テスト → 相関関係を認める

組成が違う新規合金のTcを予測できるか？

Dam et al., arXiv:1705.00978

H. C. Dam and H. Kino, et al., J. Phys. Soc. Jpn., 87, 113801 (2018)

MIの反復５ステップ

予測問題 → 回帰解析

$$T_C^{predict} = f(d_1, d_2, \cdots, d_n)$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

データには記述子とT_Cとの
相関関係を認める（否定できない）

組成が違う新規合金のT_Cを予測できるか？

外挿問題にぶつかる！

MIの反復５ステップ

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

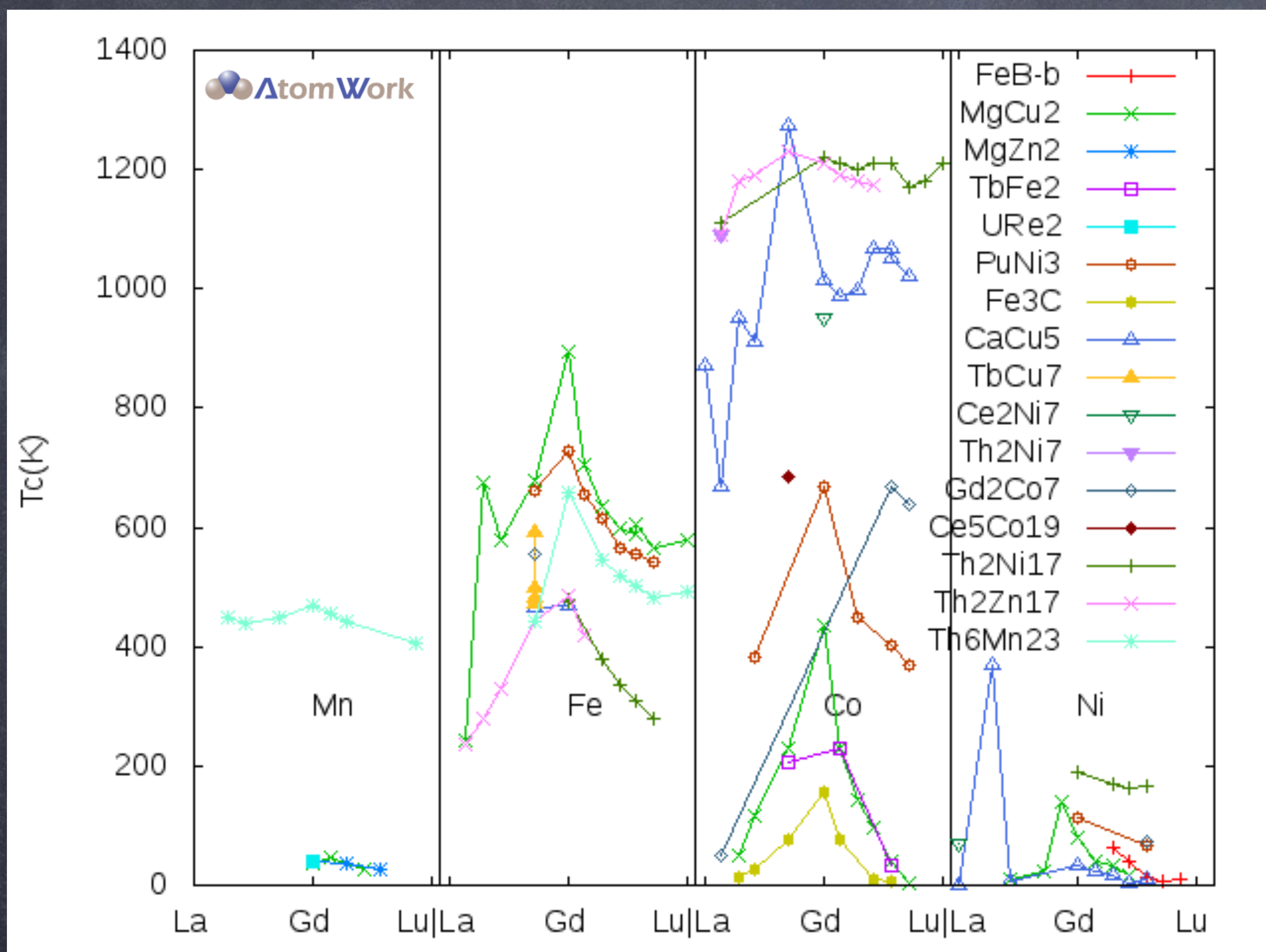
② 記述子の選定・適用

① 問題提起・問題の焼直し

問題の再提起

問題の再提起：キュリ温度の決める機構

d-f bimetal alloys from literatures



Material structure

$$\left\{\begin{array}{c} a,\, b,\, c,\, \alpha,\, \beta,\, \gamma \\ atom_1,\, x_1,\, y_1,\, z_1 \\ atom_2,\, x_2,\, y_2,\, z_2 \\ \vdots \\ atom_i,\, x_i,\, y_i,\, z_i \\ \vdots \\ atom_n,\, x_n,\, y_n,\, z_n \end{array}\right\}$$

記述子

「どうなっている？」と繰り返して問う

記述子とTcとの相関関係が観測された

「なぜ？」と問う

次には

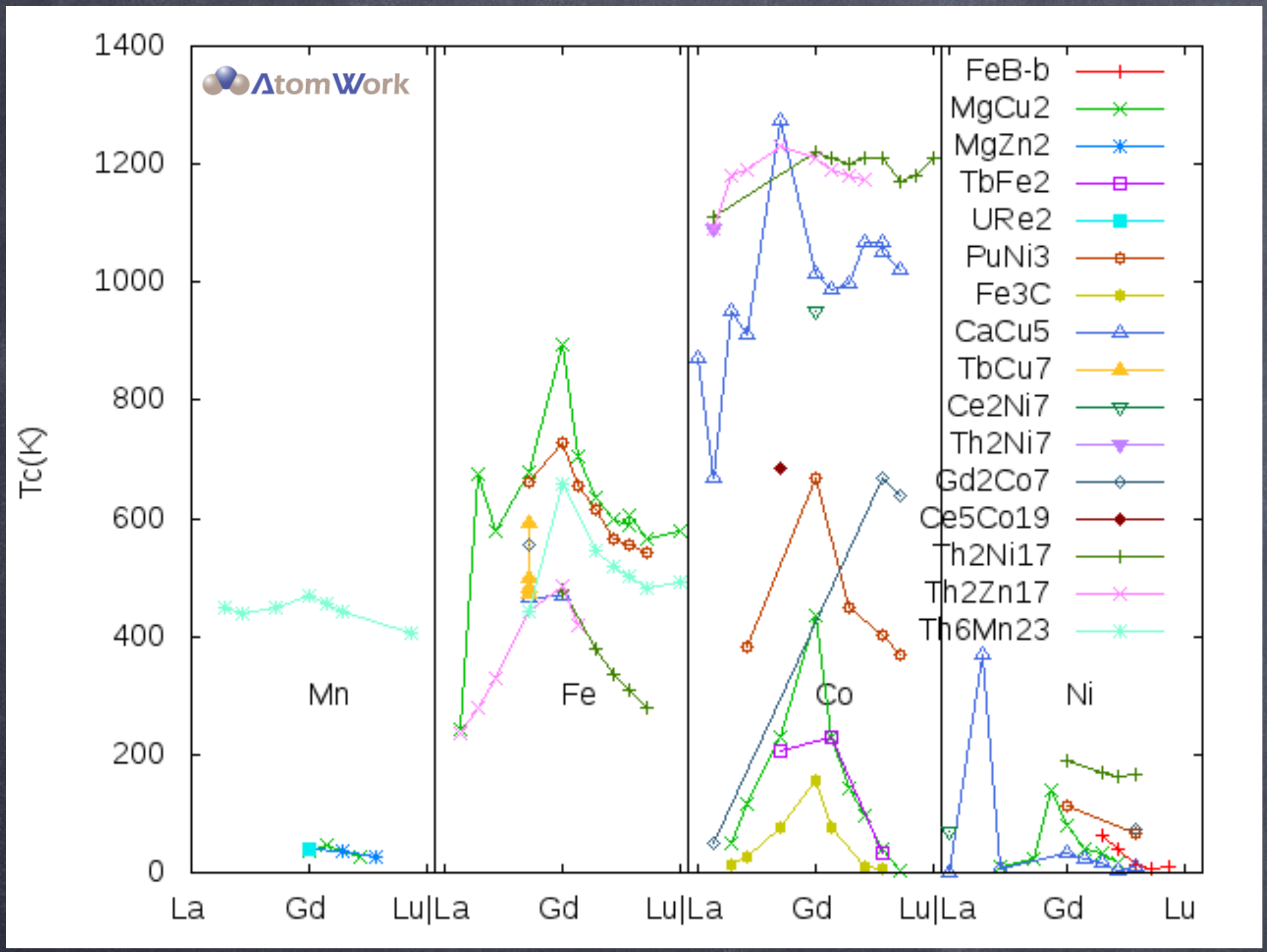
Tcの決める機構を理解するためのヒントが欲しい！

如何にヒントを抽出するか？

$$T_C^{predict} = f(\overbrace{\Box, \Diamond, \triangle, \cdots}^{})$$

Curie temperature of 3d-4f binary alloys

d-f bimetal alloys from literatures



「どうなっている？」と繰り返して問う

記述子とTcとの相関関係が観測された

「なぜ？」と問う

次には

Tcの決める機構を理解するためのヒントが欲しい！

理解可能な構造物性の潜む関係構造を抽出したい！

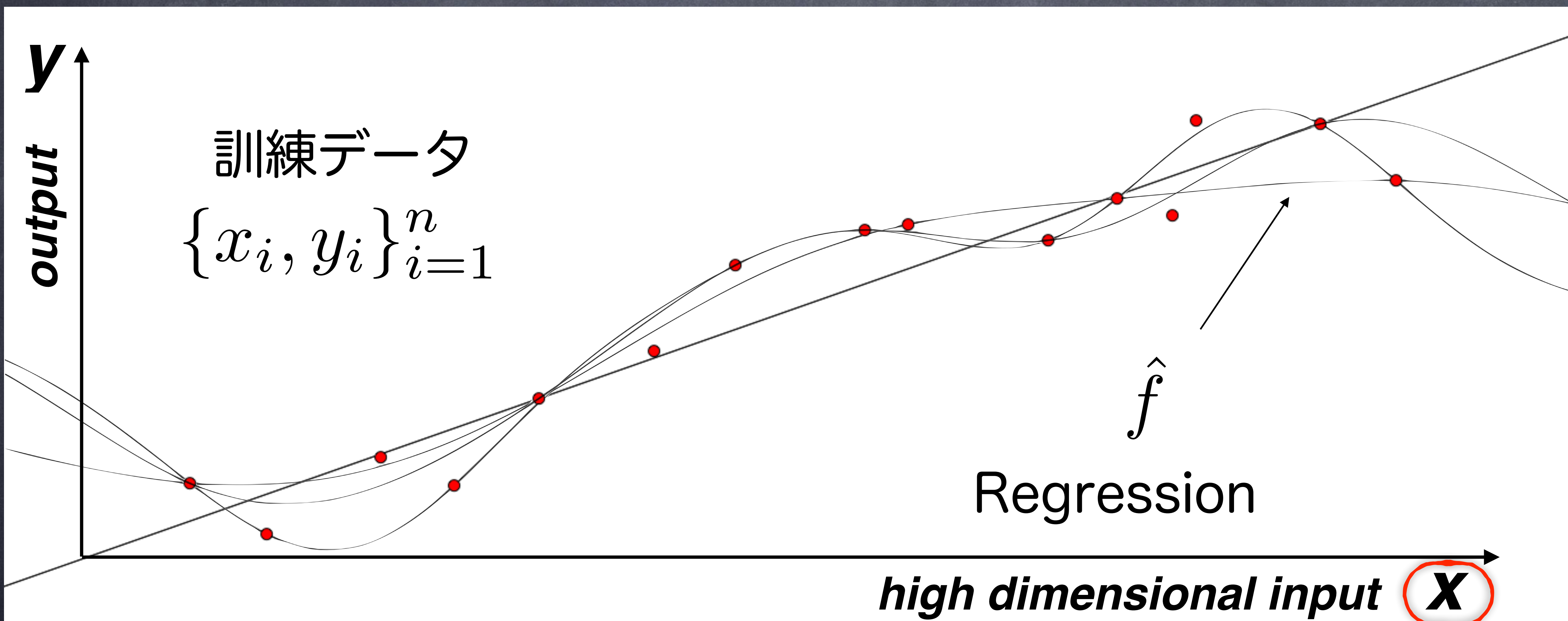
Material structure

$$\left\{ \begin{array}{c} a,\, b,\, c,\, \alpha,\, \beta,\, \gamma \\ atom_1,\, x_1,\, y_1,\, z_1 \\ atom_2,\, x_2,\, y_2,\, z_2 \\ \vdots \\ atom_i,\, x_i,\, y_i,\, z_i \\ \vdots \\ atom_n,\, x_n,\, y_n,\, z_n \end{array} \right\}$$

記述子

$$T_C^{predict} = f\overbrace{(\Box, \Diamond, \triangle, \cdots)}$$

Supervised learning: High dimensional curve fitting



予測精度に効く記述子はどれか？

MIの反復５ステップ

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

もう一度
データインスタンスの定義
からリスタート！

Regression based feature selection

想定する記述子

Structural information:
- R-R Nearest neighbor distance
- R-T Nearest neighbor distance
- T-T Nearest neighbor distance
- Number of R surrounding R
- Number of T surrounding T
- Number of R surrounding T
- Number of T surrounding T
- Concentration of R
- Concentration of T

Atomic information:
- Atomic Number of Z_R, Z_T
- Covalent Radius of r_R, r_T
- Electron Negativity of χ_R, χ_T
- Ionization Potential IP_R, IP_T
- Orbital angular momentum L_R, L_T
- Spin angular momentum S_R, S_T
- Total angular momentum J_R, J_T
- Landé g-factor g_{JR}, $J_R g_{JR}$, $J_R(1-g_{JR})$

ガウス過程回帰

2^{27}

All combinations

$$T_c^{predict} = f_{gp}\boxed{(x^1, \cdots, x^m)}$$

Model data

$$\{\{x^1, \cdots, x^m\}, R^2(T_c^{obs}, T_c^{predict})\}$$

Top N-best prediction accuracy

$$\left\{ \begin{array}{c} R^2_{N-best}(T_c^{obs}, T_c^{predict}), \\ \{x^1, \cdots, x^k\} \end{array} \right\}$$

予測精度に効く記述子はどれか？

Regression based feature selection

データインスタンス：
記述子の組み合わせ

ガウス過程回帰

2^{27}

All combinations

$$T_c^{predict} = f_{gp}\boxed{(x^1, \cdots, x^m)}$$

Model data

$$\{\{x^1, \cdots, x^m\}, R^2(T_c^{obs}, T_c^{predict})\}$$

データ集合：
2^{27} 記述子の全組み合わせ

Top N-best
prediction accuracy

$$\left\{ \begin{array}{c} R^2_{N-best}(T_c^{obs}, T_c^{predict}), \\ \{x^1, \cdots, x^k\} \end{array} \right\}$$

予測精度に効く記述子はどれか？

MIの反復５ステップ

⑤ 最終成果の通信

属性の関係性解析

$$\{\{x^1, \cdots, x^m\}, R^2(T_c^{obs}, T_c^{predict})\}$$

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

② 記述子の選定・適用

① 問題提起・問題の焼直し

解析用データの準備

Regression based feature selection

解析用データ

ID	Descriptors	R^2
Model 1	$\{d_i, d_j, \ldots, d_k\}$	R^2_1
Model 2	$\{d_l, d_m, \ldots, d_n\}$	R^2_2
…	…	…
Model n	$\{d_o, d_p, \ldots, d_q\}$	R^2_n

ガウス過程回帰

2^{27}

All combinations

$$T_c^{predict} = f_{gp}\boxed{(x^1, \cdots, x^m)}$$

Model data

$$\{\{x^1, \cdots, x^m\}, R^2(T_c^{obs}, T_c^{predict})\}$$

Top N-best
prediction accuracy

$$\left\{ \begin{array}{c} R^2_{N-best}(T_c^{obs}, T_c^{predict}), \\ \{x^1, \cdots, x^k\} \end{array} \right\}$$

予測精度に効く記述子はどれか？

MIの反復５ステップ

属性の関係性解析

$$\{\{x^1, \cdots, x^m\}, R^2(T_c^{obs}, T_c^{predict})\}$$

⑤ 最終成果の展開

④ 解析結果の評価・解釈

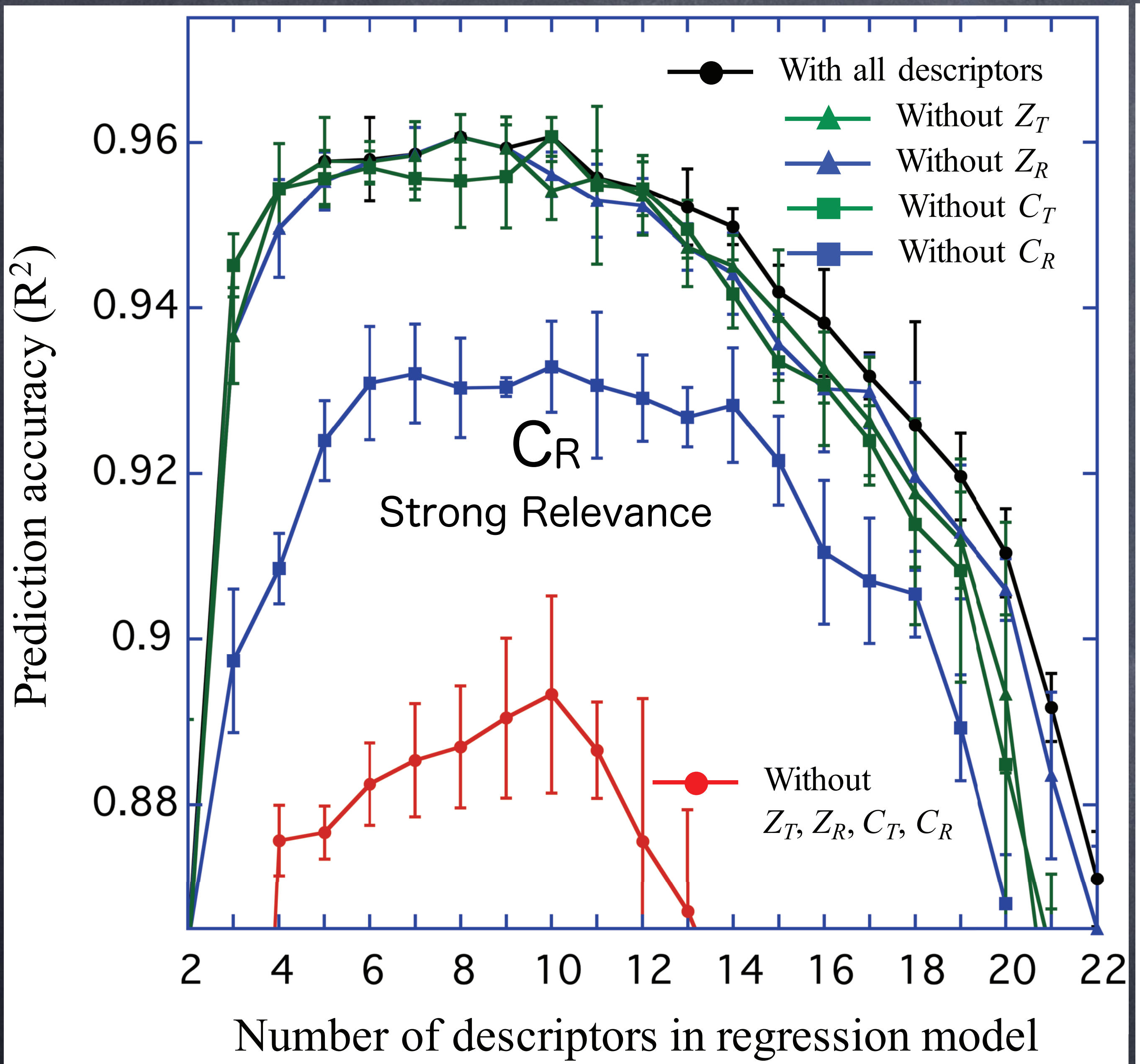
③ 解析手法の選定・適用

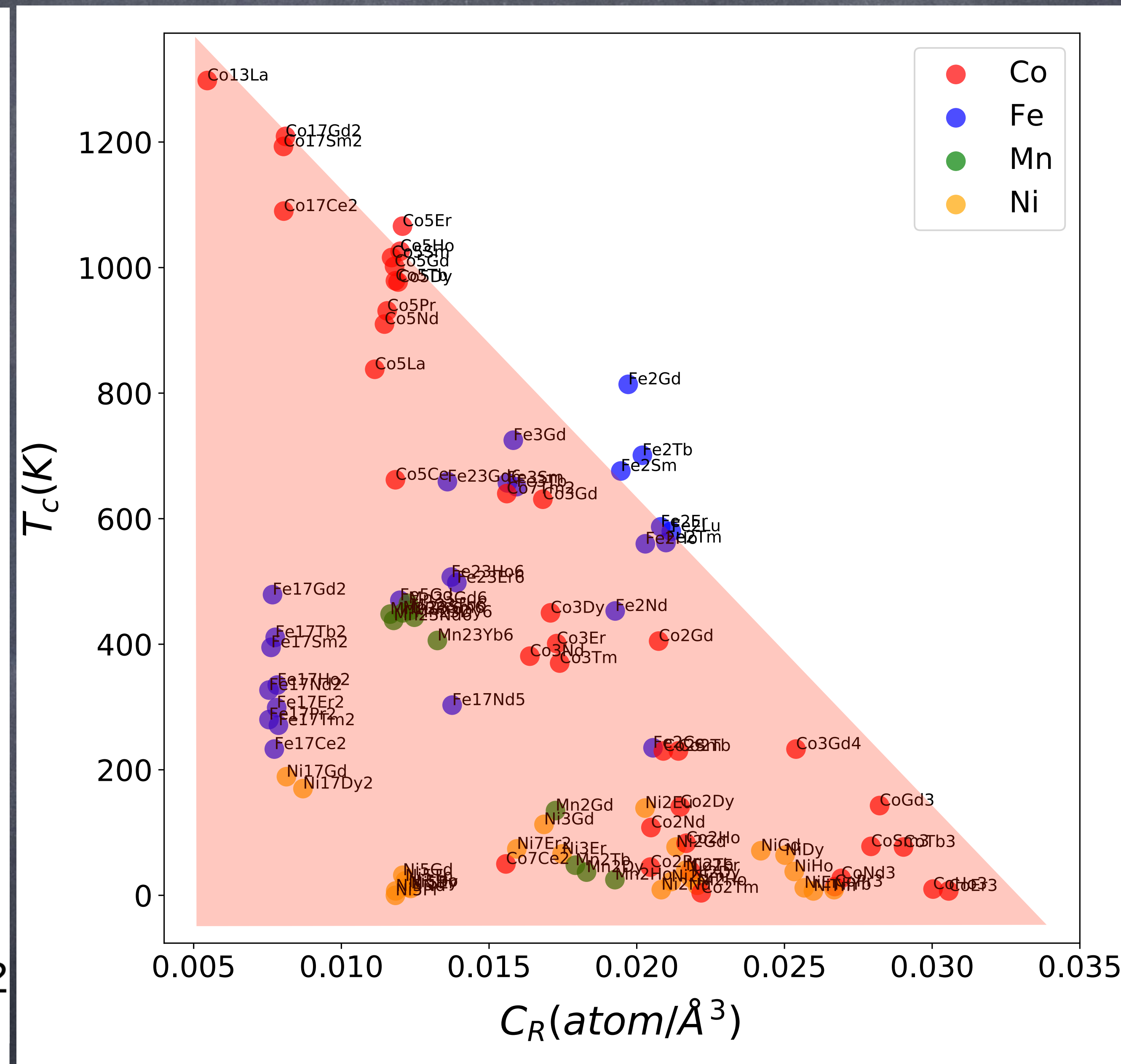
② 記述子の選定・適用

① 問題提起・問題の焼直し

アソシエション解析

Regression based descriptor evaluation

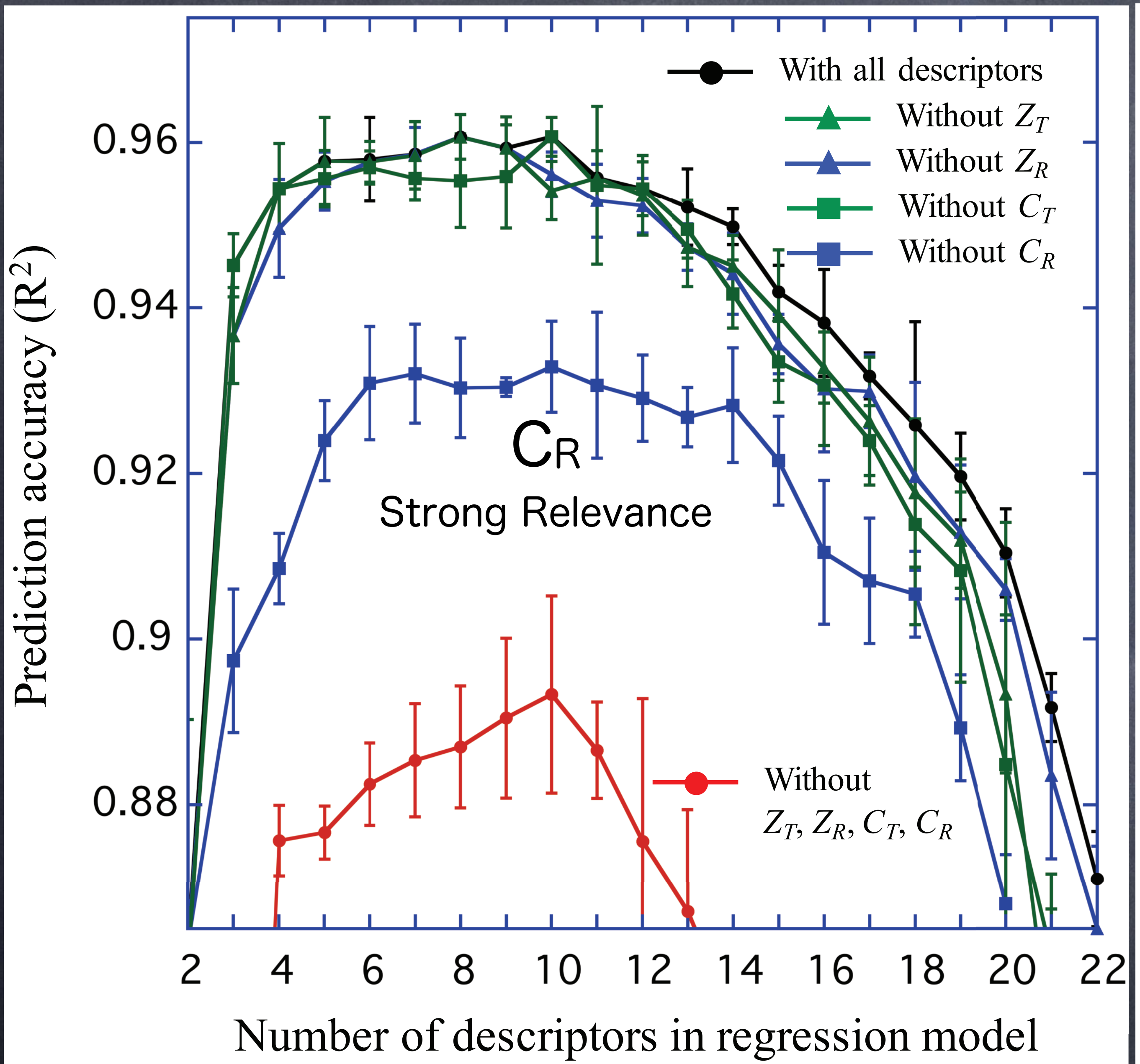


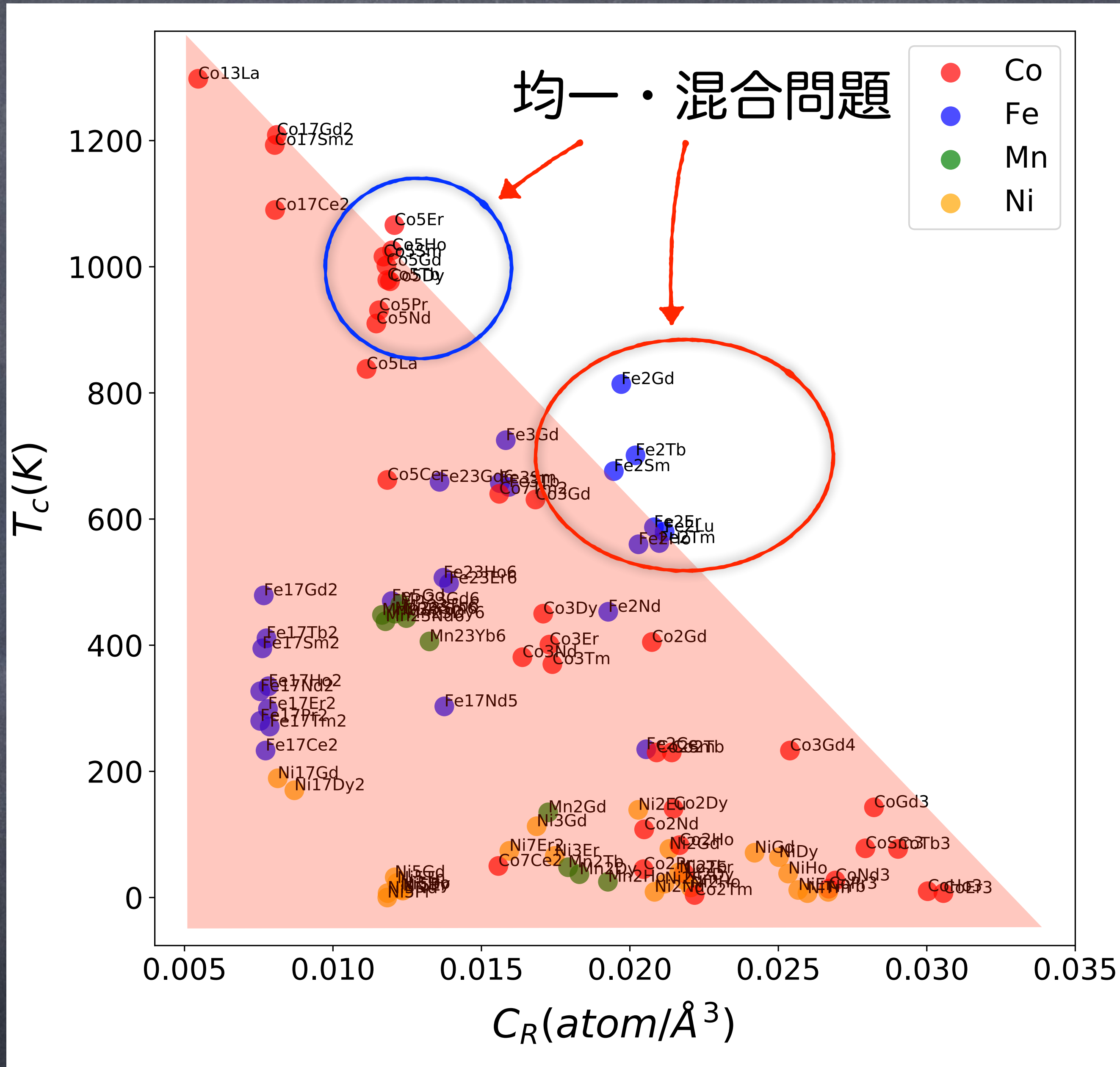


Tcの予測精度と活用する記述子との関係性　　希土類の濃度がTcを決定する機構に重要だ！

J. Phys. Soc. Jpn. 87, 113801 (2018); IUCrJ 5, 6, 830-840 (2018);
J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019).

Regression based descriptor evaluation





Tcの予測精度と活用する記述子との関係性　　希土類の濃度がTcを決定する機構に重要だ！

J. Phys. Soc. Jpn. 87, 113801 (2018); IUCrJ 5, 6, 830-840 (2018);
J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019).

MIの反復５ステップ

⑤ 最終成果の展開

④ 解析結果の評価・解釈

③ 解析手法の選定・適用

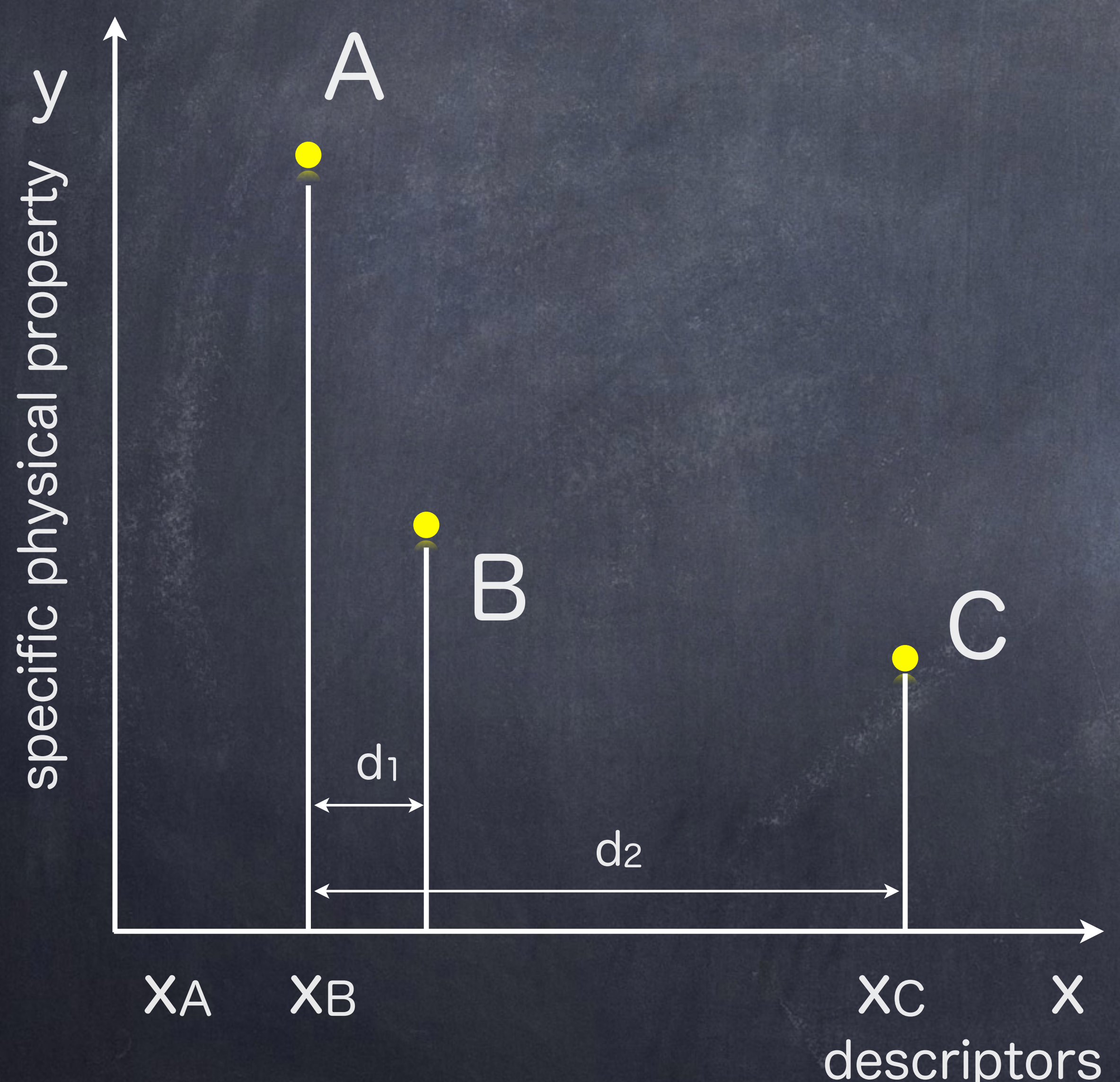
② 記述子の選定・適用

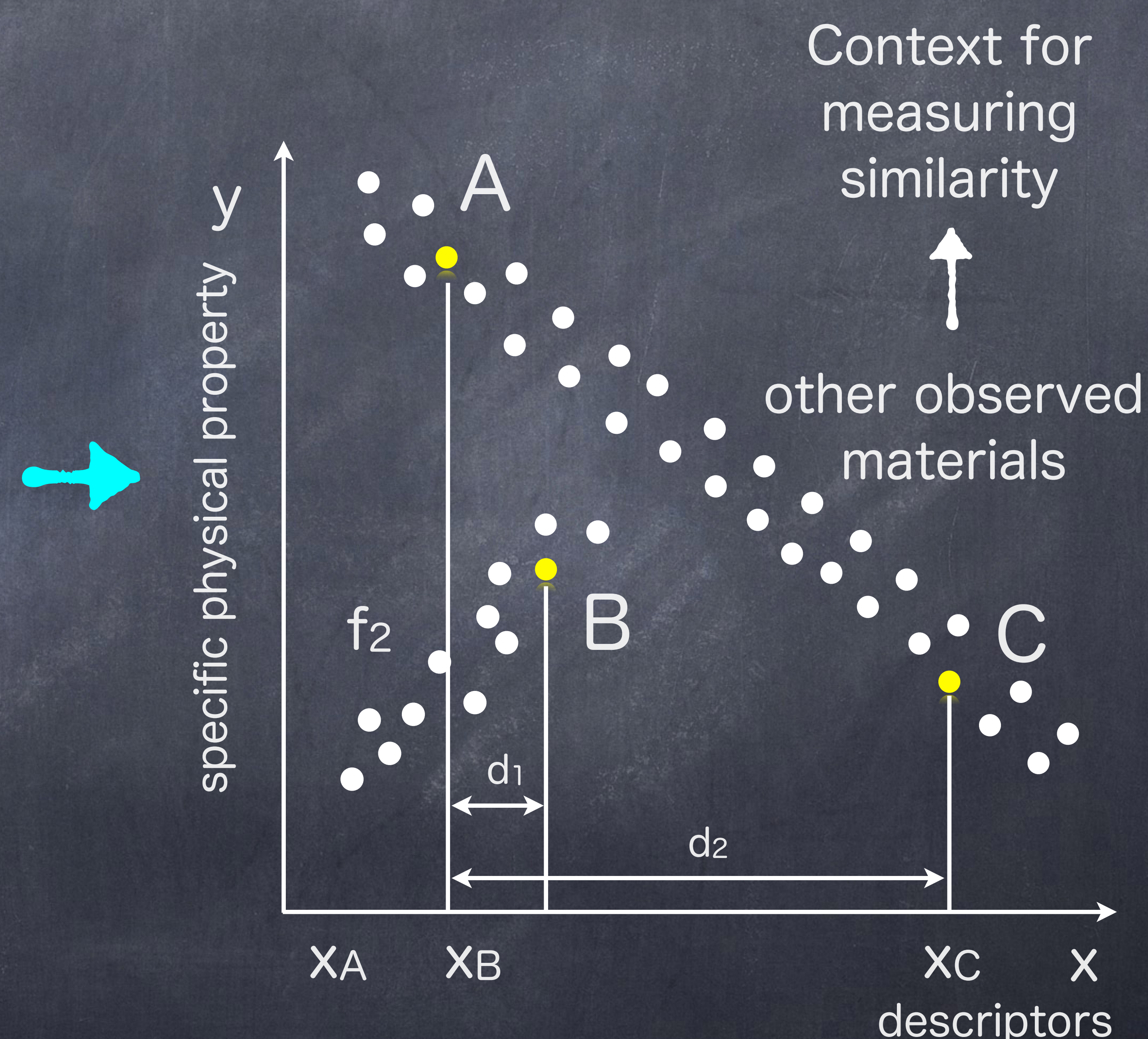
① 問題提起・問題の焼直し

問題の再提起

物質間の類似性をどう評価するか？

Development of context-based method for measuring similarity between materials





Example のまとめ

⑤ 最終成果の展開

交差検定
専門知識による解釈性

④ 解析結果の評価・解釈

1巡目：非線形回帰解析
2巡目：アソシエション解析
3巡目：アンサンブル学習

③ 解析手法の選定・適用

② 記述子の選定・適用

1巡目：専門家が想定した記述子
2巡目：記述子の組み合わせ
3巡目：２巡目で重要だと判定された記述子

① 問題提起・問題の焼直し

1巡目：希土類・遷移金属２元合金のTcの予測問題
2巡目：専門家が想定した記述子の重要性の評価
3巡目：Tcを考慮する際の物質間の類似性

UNCERTAINTY AND DECISION-MAKING FOR MATERIALS DISCOVERY

Hieu-Chi Dam

Japan Advanced Institute of Science and Technology

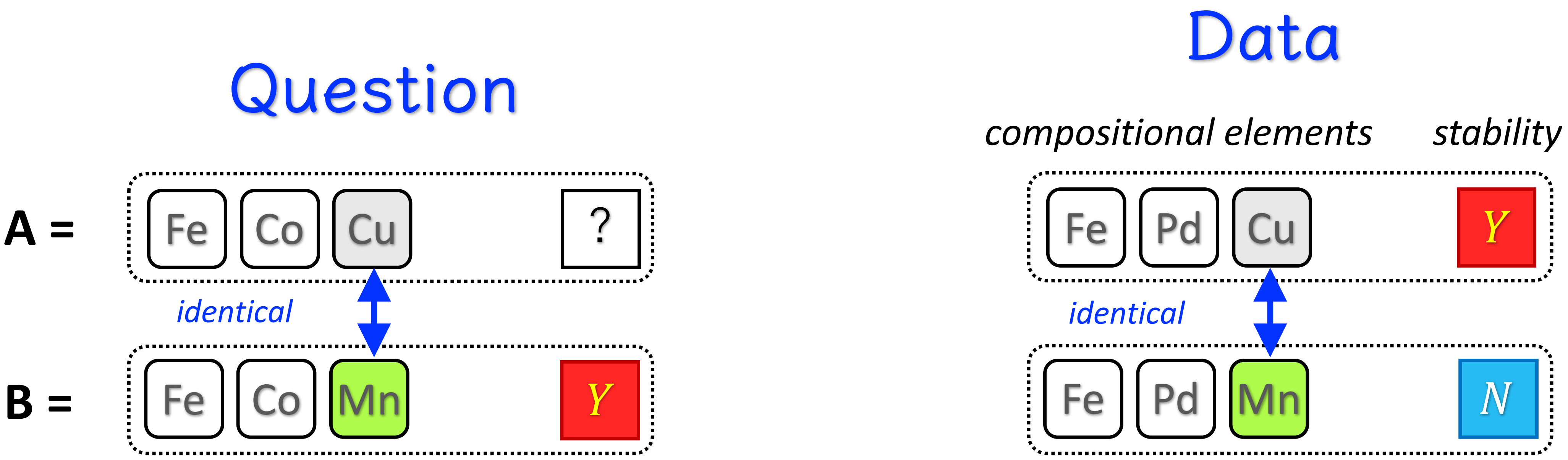
June 29, 2023

Material Dataset: An Example with Alloys

Note: descriptions differ from descriptors

Material	Material descriptions	Stability	Magnetization (T)
Material 1	{Cr,Cu,Mo,Tc, structure}	***N***	0
Material 2	{Co,Cu,Fe,Mn, structure}	***Y***	0.38
…	…	…	…
Material n	{Ag,Fe,Mn,Zn, structure}	***N***	0

Similarity: Error of substitution



N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

The Key Idea:

Transforming Data into Evidence of Similarity

Similarity judgment

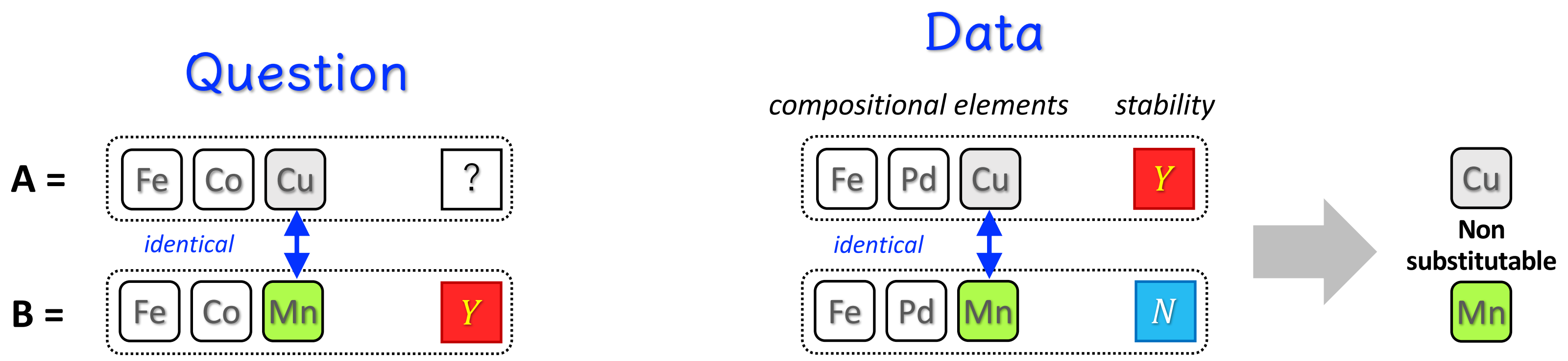
Similarity or dissimilarity data appear in different forms:

- ✓ Ratings of pairs
- ✓ Sorting of objects
- ✓ Communality between associations
- ✓ Errors of substitution
- ✓ Correlations between occurrences

Amos Nathan Tversky
(1937 – 1996)

Amos Teversky, “Features of Similarity", Psychological Review, 84, 4 (1977).

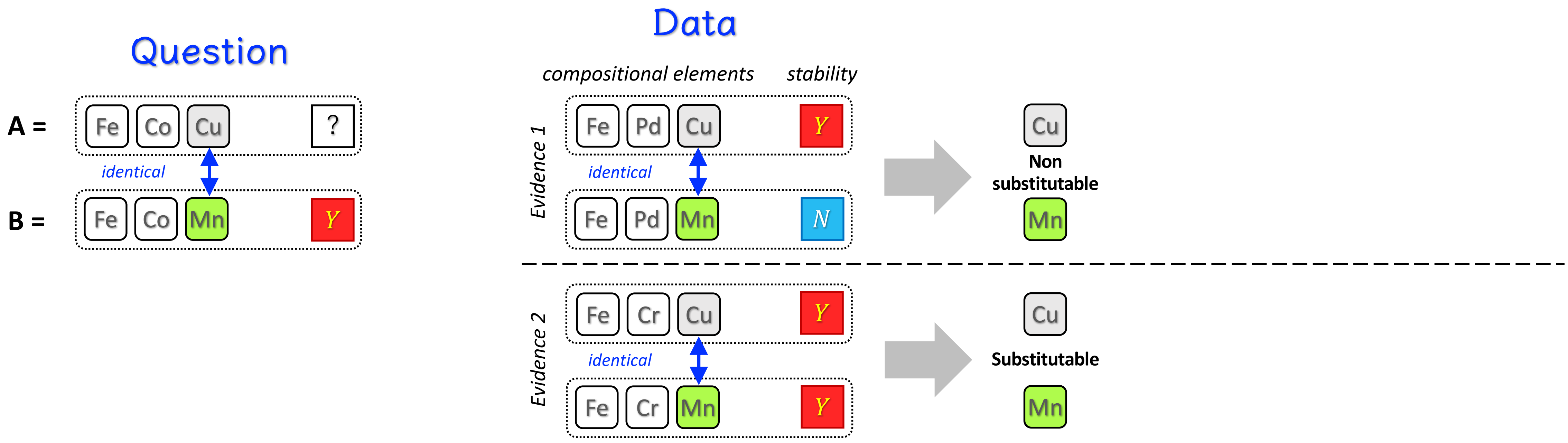
Similarity: Error of substitution



Intuitive Evidence

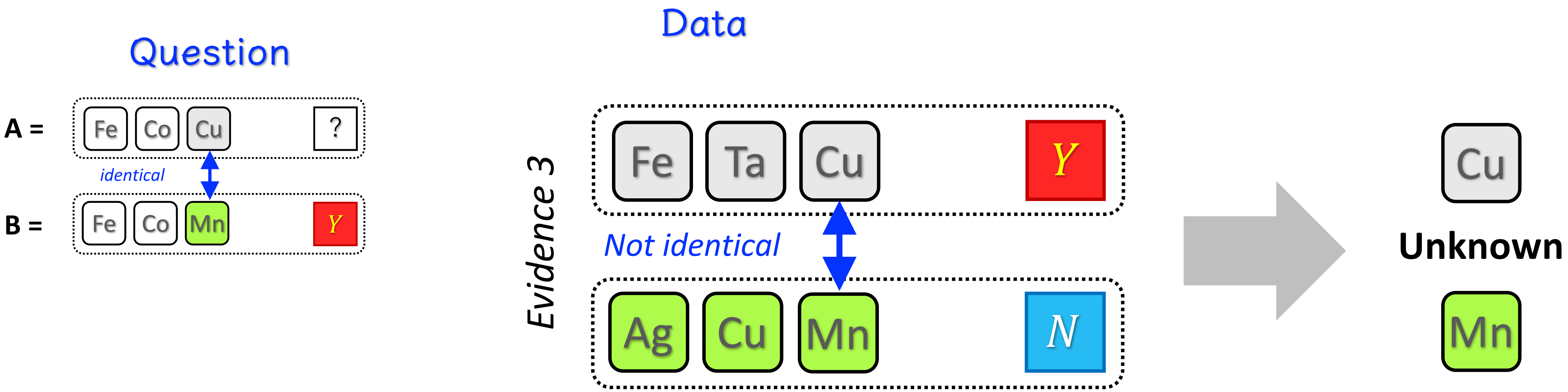
N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Similarity: Error of substitution



N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

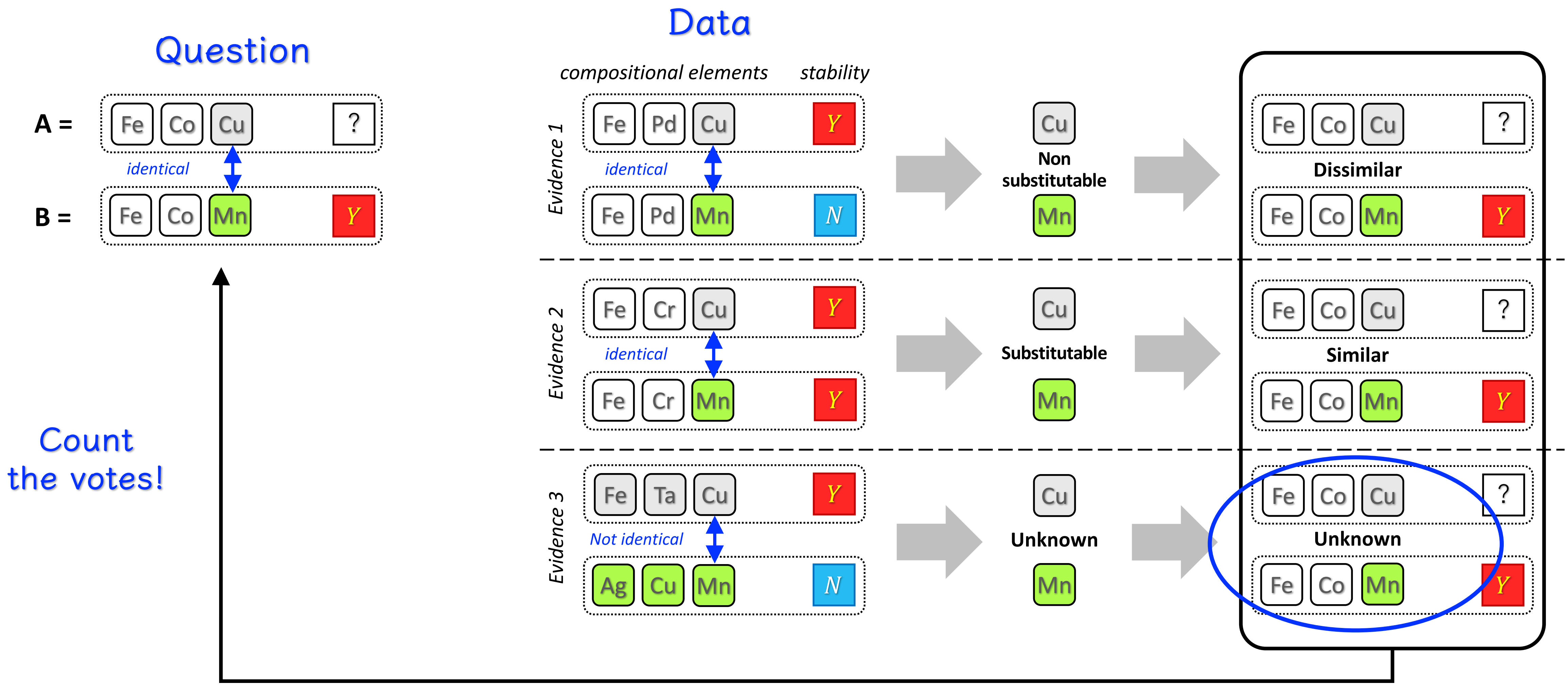
Similarity: Error of substitution



No Evidence is provided!

N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Similarity: Error of substitution



Nguyen et al., IUCrJ, 5, 6 (2018); J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019).

Similarity: Error of substitution

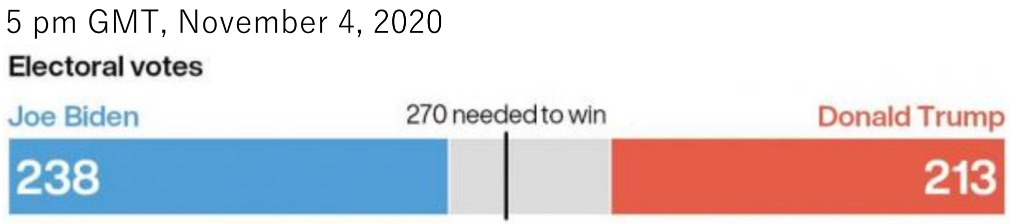
Question

A = Fe Co Cu ?

identical

B = Fe Co Mn *Y*

Uncertainty



"Unknown" ***is not*** 50/50

Evidence 3

Fe Ta Cu *Y*

Not identical

Ag Cu Mn *N*

Cu

Unknown

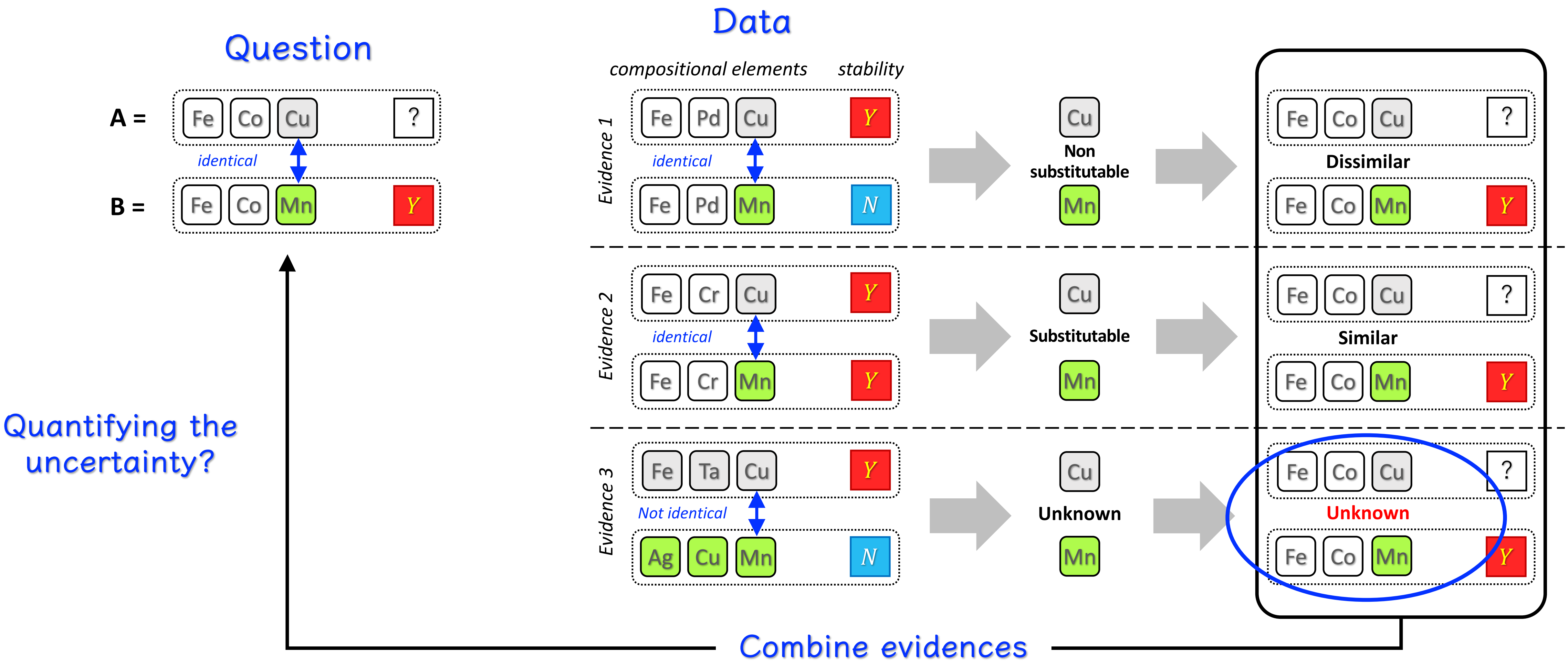
Mn

Fe Co Cu ?

Unknown

Fe Co Mn *Y*

Similarity: Error of substitution



N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Dempster–Shafer theory (belief functions)

Frame of discernments

$$\Omega_{sim} = \{similar, dissimilar\}$$

Pending judgment of
Similarity or Dissimilarity

$$2^{\Omega_{sim}} = \{\emptyset, \{similar\}, \{dissimilar\}, \{similar, dissimilar\}\}$$

Mass function

$$m_{evidence}: 2^{\Omega_{sim}} \to \mathbb{R}$$

$m_{evidence}(\{similar\})$: degree of belief that A and B are similar

$m_{evidence}(\{dissimilar\})$: degree of belief that A and B are dissimilar

$m_{evidence}(\{similar, dissimilar\})$: degree of belief that the situation is unknown

Prof. Arthur Pentland Dempster

A. P. Dempster, **A generalization of Bayesian inference**, Journal of the Royal Statistical Society, Series B, 30 (1968)

Challenges in Material Datasets: Size, Bias, and Conflicts

Summary of the 8 datasets of High Entropy Alloys

Datasets	No. alloys	Observation rate	No. HEAs	HEAs rate	No. candidates
$\mathcal{D}_{ASMI16}$	45 binary alloys	13%	45	(100%)	351 binary alloys
$\mathcal{D}_{CALPHAD}$	243 ternary alloys	9%	243	(100%)	2925 ternary alloys
$\mathcal{D}_{AFLOW}$	117 binary alloys 441 ternary alloys	33% 15%	60 234	(51%) (53%)	351 binary alloys 2925 ternary alloys
$\mathcal{D}_{LTVC}$	117 binary alloys 441 ternary alloys	33% 15%	58 148	(49%) (33%)	351 binary alloys 2925 ternary alloys
$\mathcal{D}_{AFLOW}^{quaternary}$	1,110 quaternary alloys	6%	754	(68%)	17,550 quaternary alloys
$\mathcal{D}_{AFLOW}^{quaternary}$	1,110 quaternary alloys	6%	480	(43%)	17,550 quaternary alloys
$\mathcal{D}_{AFLOW}^{quinary}$	130 quinary alloys	0.16%	129	(99%)	80,730 quinary alloys
$\mathcal{D}_{LTVC}^{quinary}$	130 quinary alloys	0.16%	91	(70%)	80,730 quinary alloys

"HEAs rate": the ratio of No. HEA to No. alloys; "Observation rate": the ratio of No. alloys to No. candidates.

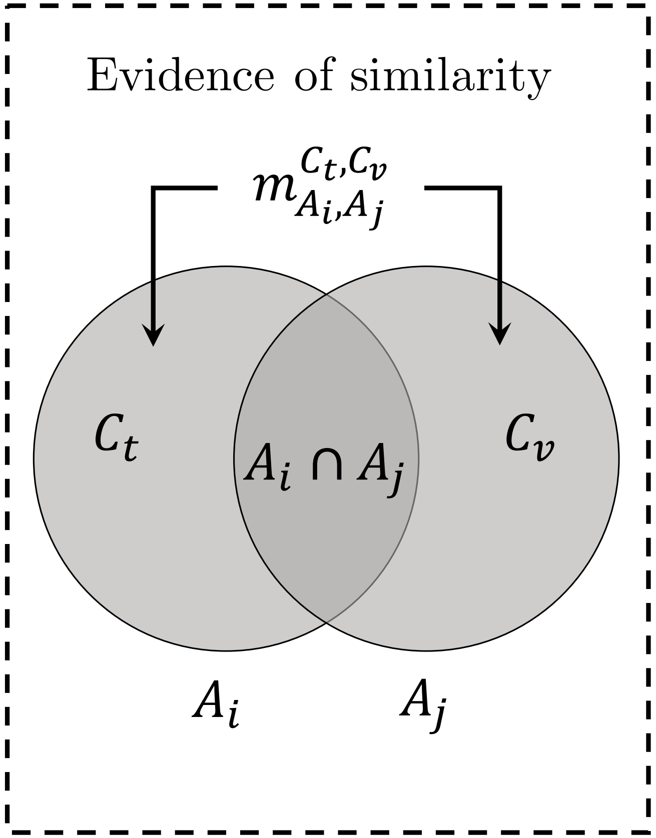
The number of confirmed materials is relatively smaller than the number of candidates

Heavily bias to positive label (HEA phase)

Alman, D., *Entropy* **15** (2013); Zhang, F. et al., *CALPHAD* **45** (2014); Lederer, Y. et al., Acta Mater. 159 (2018)

Modeling Similarity Between Compositions

Quantifying uncertainty
(Due to Limited or Conflicting Data)



Collecting evidence from data

y_{A_i}: property of alloy A_i

y_{A_j}: property of alloy A_j

$$m_{A_i,A_j}^{C_t,C_v}(\{similar\}) = \begin{cases} \alpha & \text{if } y_{A_i} = y_{A_j} \\ 0 & \text{otherwise} \end{cases},$$

confidence of similarity (substitutable)

$$m_{A_i,A_j}^{C_t,C_v}(\{dissimilar\}) = \begin{cases} \alpha & \text{if } y_{A_i} \neq y_{A_j} \\ 0 & \text{otherwise} \end{cases},$$

confidence of dissimilarity (not substitutable)

$$m_{A_i,A_j}^{C_t,C_v}(\{similar, dissimilar\}) = 1 - \alpha$$

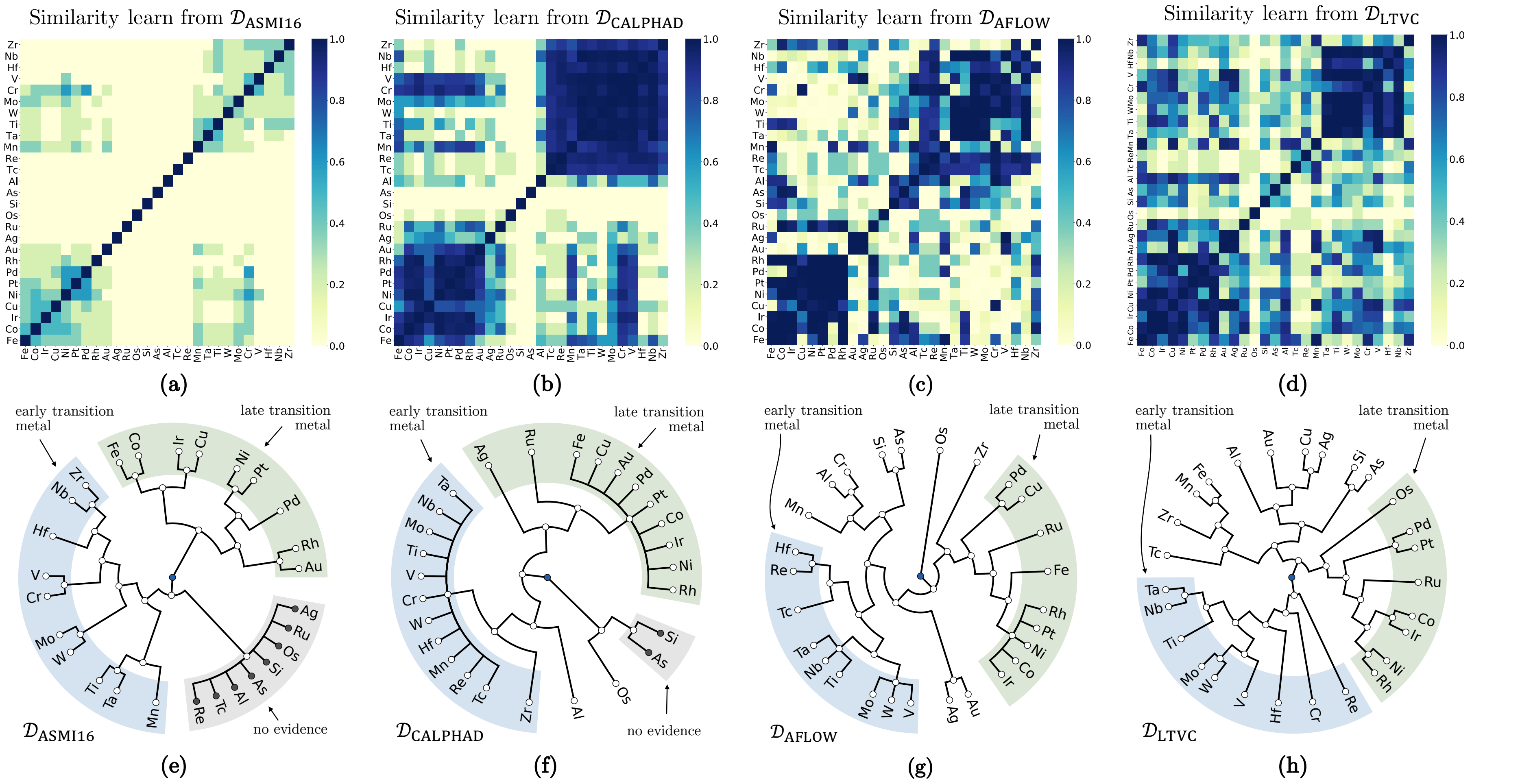
degree of "unknown" (not sure if substitutable)

N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Application:

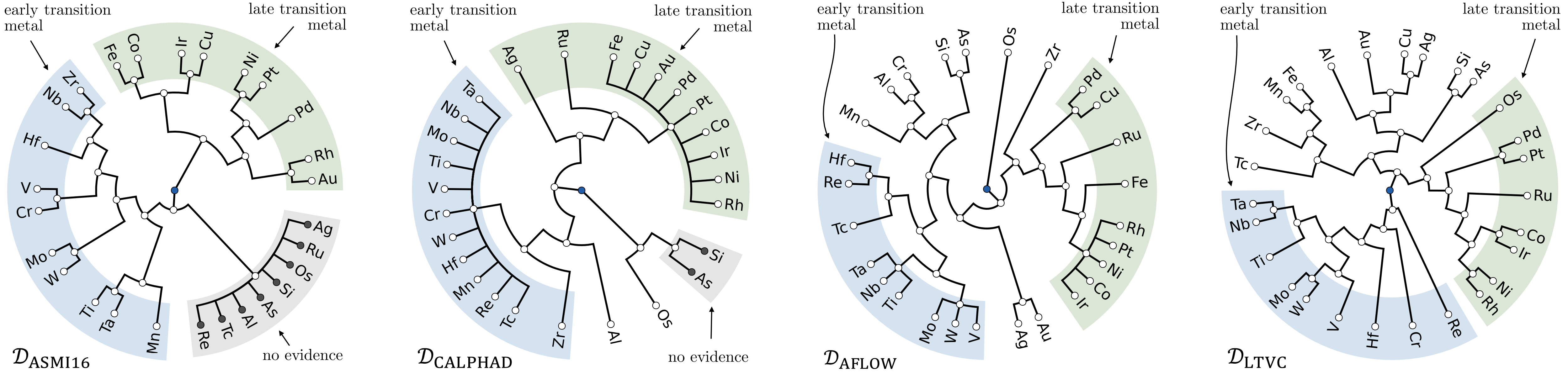
Extracting Insights from Material Similarity to Support Decision Making

Data-Driven Similarities in Forming HEAs



M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Data-Driven Similarities in Forming HEAs



M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Recommending New Element Combinations

To be substituted by other components to create new candidates

confidence of forming HEA

$$m^{A_{new}}_{A_k,C_t\leftarrow C_v}(\{HEA\}) = \begin{cases} s(C_t, C_v) & \text{if } y_{A_k} = HEA \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

confidence of not forming HEA

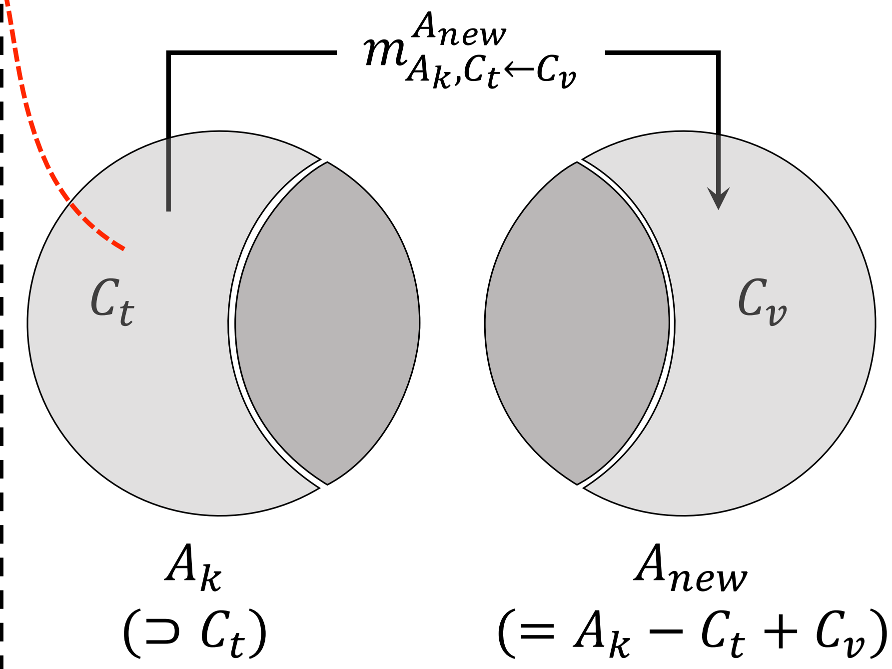
$$m^{A_{new}}_{A_k,C_t\leftarrow C_v}(\{\neg HEA\}) = \begin{cases} s(C_t, C_v) & \text{if } y_{A_k} = \neg HEA \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

degree of "unknown"

$$m^{A_{new}}_{A_k,C_t\leftarrow C_v}(\{HEA, \neg HEA\}) = 1 - s(C_t, C_v), \quad (8)$$

C_t and C_v can differ in size!

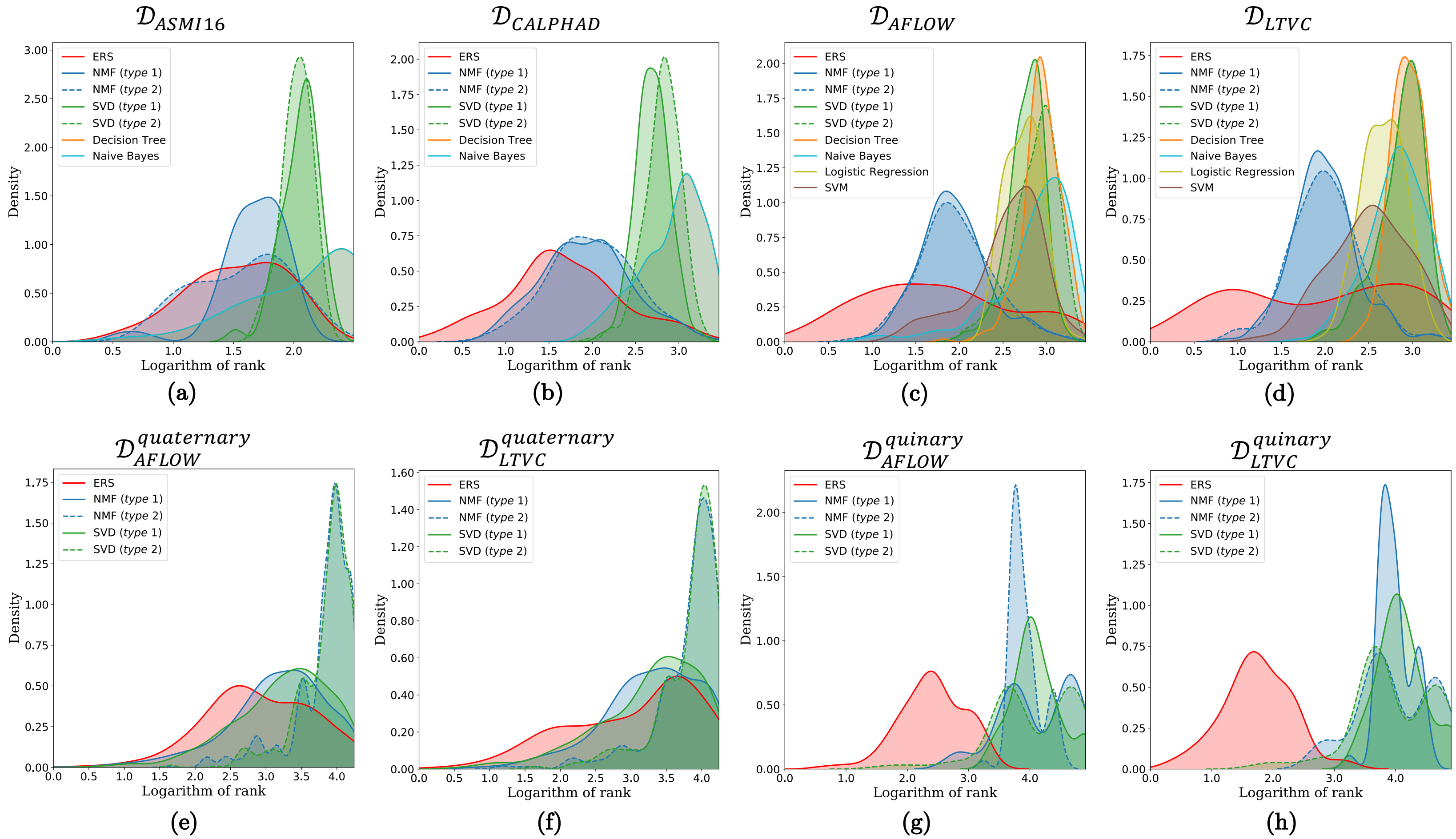
Evidence of element subsitution



Recommending potential HEAs

N. N. T. Ton and M. Q. Ha et al., 2D Mater. 8 015019 (2021); M. Q. Ha et al., Nat. Comp. Sci., 1 (2021).

Recommendation capability



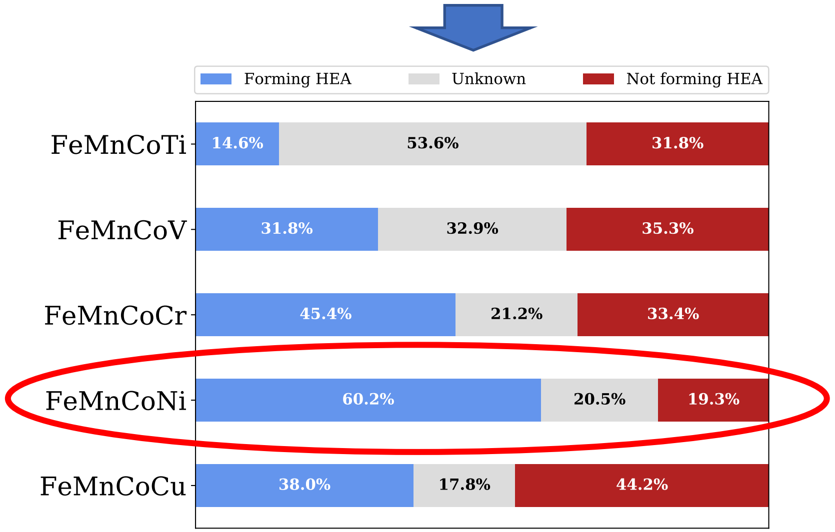
C_t and C_v can differ in size!

Recommending quaternary and quinary HEAs by using binary, ternary HEAs data

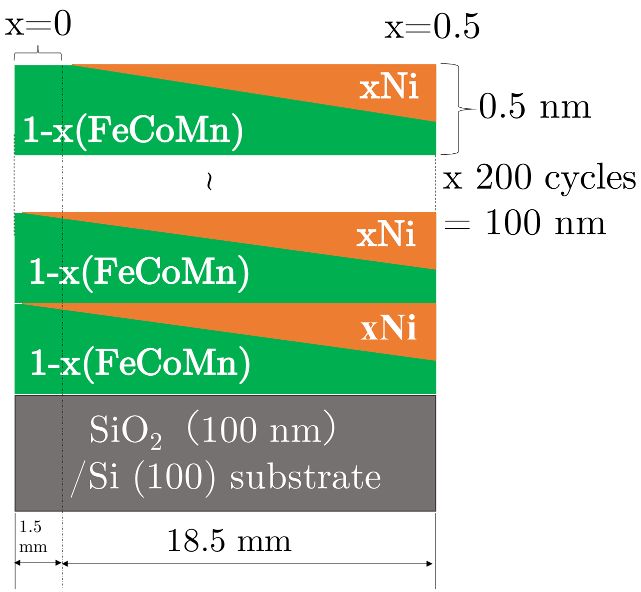
～ 500 times higher recall rate

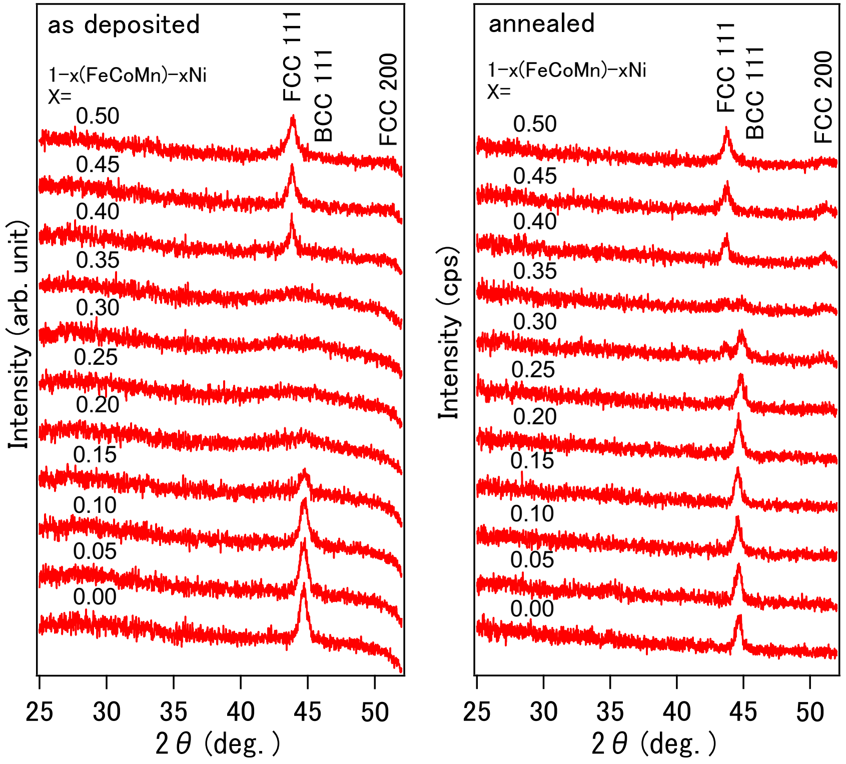
Experimental validation of FeMnCoNi

'Unknown' indicates insufficient information from observed data

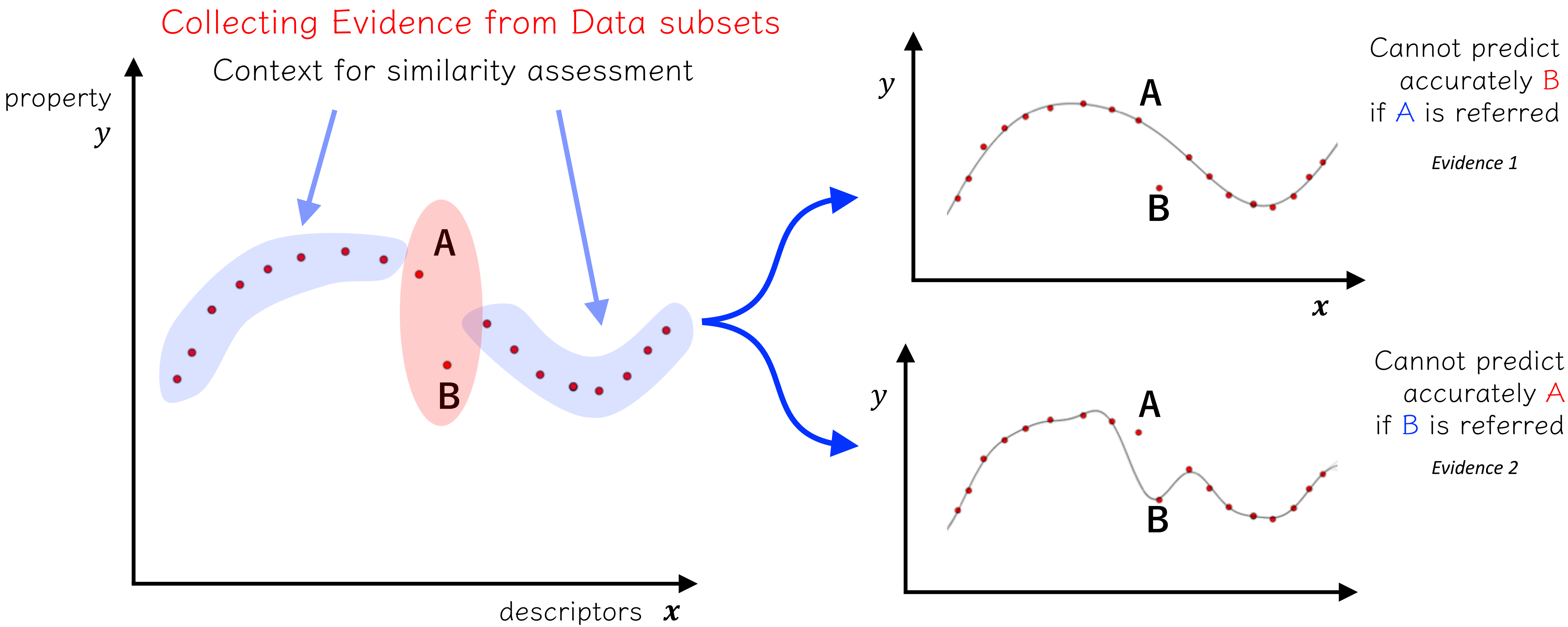


'Unknown' should not be interpreted as a 50/50 probability



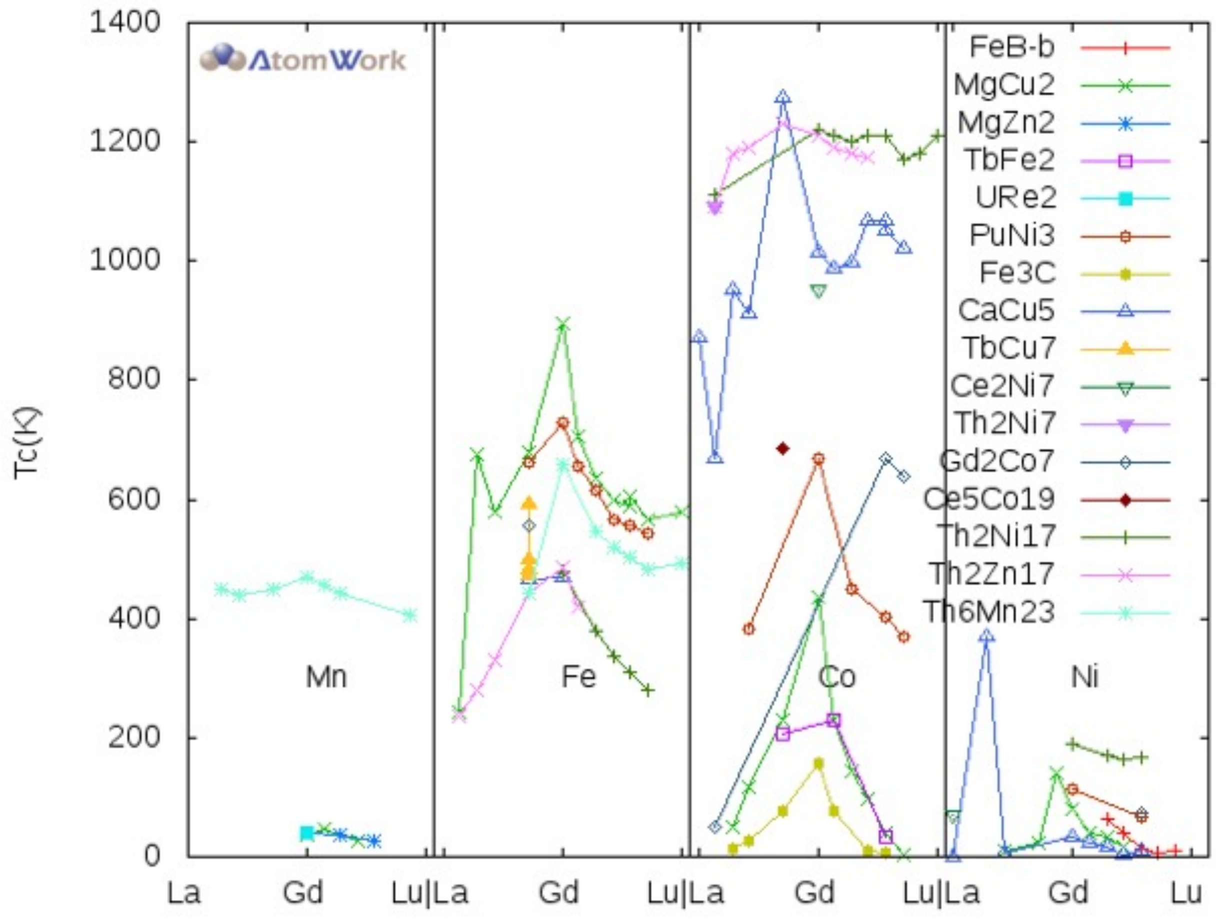


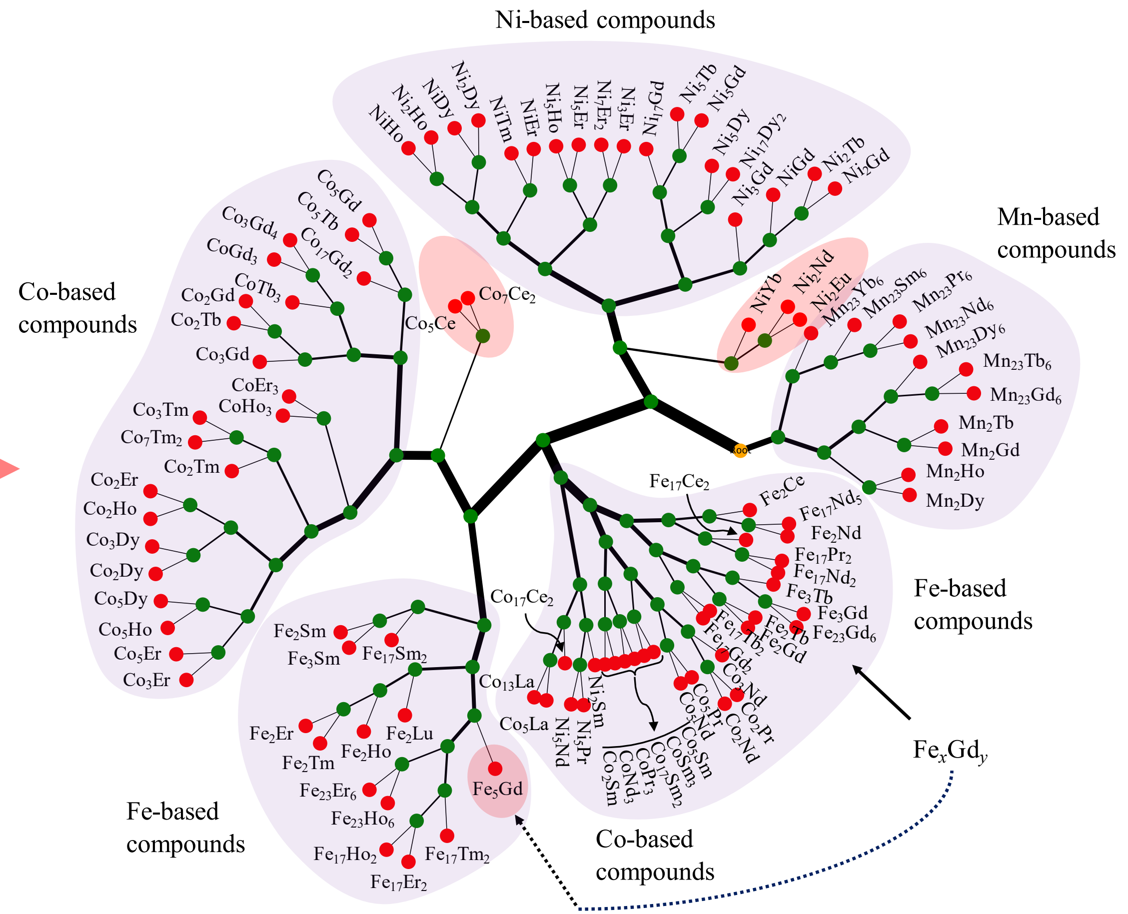
Similarity: Correlation between occurence



Nguyen *et al.*, J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019); M.-Q. Ha *et al.*, J. App. Phys. 133, 5 (2023)

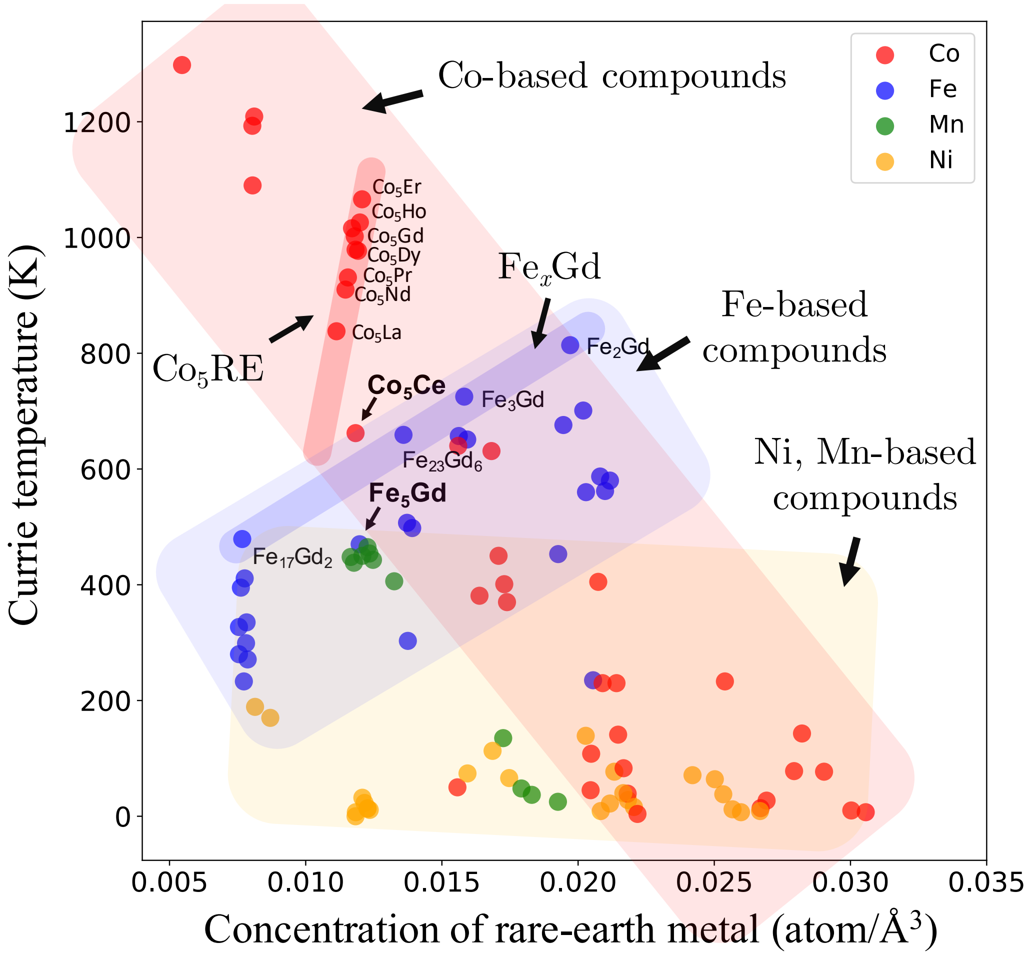
Similarity between
Rare earth – Transition metal alloys

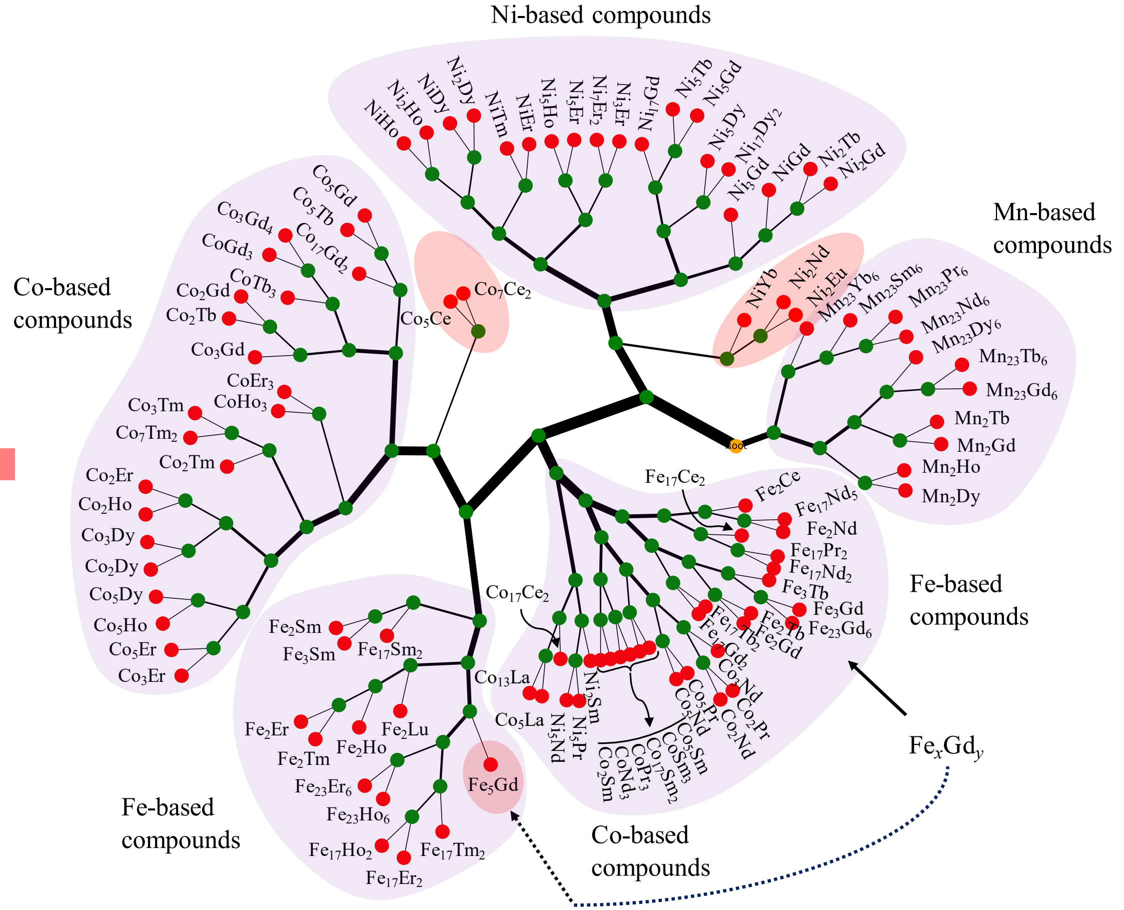




Nguyen et al., J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019).

Gleaning Insights from Material Similarity





Nguyen et al., J. Phys. CS. (CCP-2018) (2019); J. Phys. Mat. 2, 3 (2019).