KYOTO UNIVERSITY

Machine Learning

- Basic Ideas and Recent Advances -

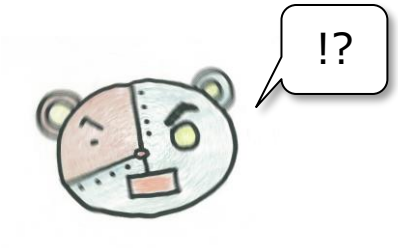
Hisashi Kashima
kashima@i.kyoto-u.ac.jp

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

Introduction:
Basic ideas of machine learning and recent advances

1. What is machine learning?
2. Machine learning applications
3. Recent advances:
   - Deep learning
   - Graph deep learning
   - Treatment effect prediction

What is machine learning?

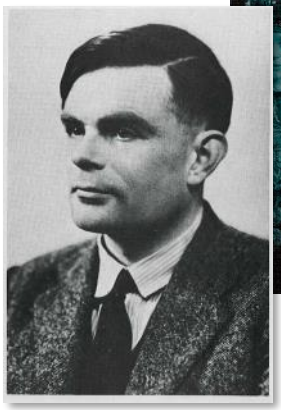


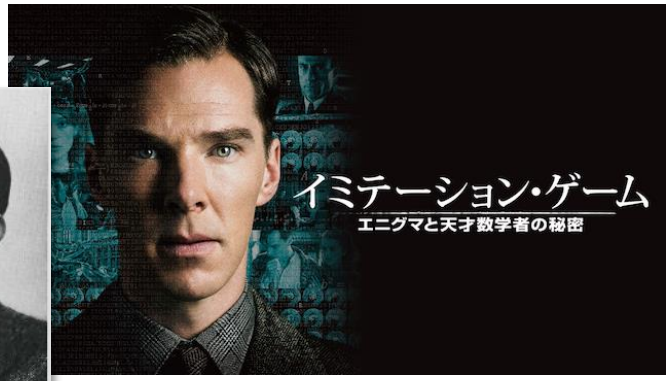
“The third A.I. boom”: Machine learning is a core technology

- You can see many successes of “Artificial Intelligence”:
  - Q.A. machine beating quiz champions & Go program surpassing top players
  - Machine vision better at recognizing objects than humans
  - Image and illustration generation indistinguishable from reality or professionals
  - Text generation and reasoning skills beyond the average human
- Current A.I. boom owes machine learning
  - Especially, deep learning

What is machine learning? :
A branch of artificial intelligence

- Originally started as a branch of artificial intelligence
  - has its more-than-50-years history
    - … almost as old as the history of computers.
  - Computer programs that “learns” from experience
    - Based on logical inference in its early stage





What is machine learning? : A data analytics technology

- Rise of “statistical” machine learning
  - Successes in bioinformatics, natural language processing, and other business areas
  - Victory of IBM’s Watson QA system, Google’s Alpha Go
- Recently rather considered as a data analysis technology
  - “Big data” and “Data scientist”
    - Data scientist is “the sexiest job in the 21st century”
- Success of deep learning
  - The 3rd AI boom

What can machine learning do?: Prediction and discovery

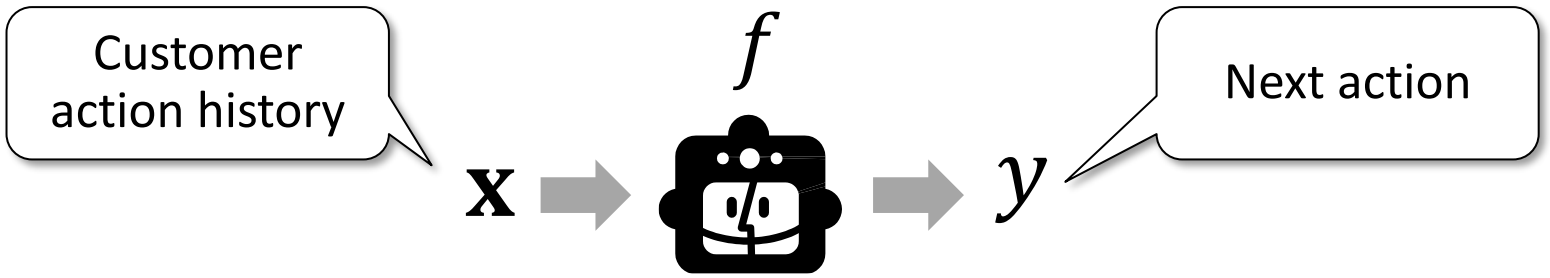
- Two categories of the use of machine learning:

1. Prediction (supervised learning)
   - "What will happen in future data?"
   - Given past data, predict about future data

2. Discovery (unsupervised learning)
   - "What is happening in data in hand?"
   - Given past data, find insights in them

Prediction machine:
A function from a vector to a scalar

- We model the intelligent machine as a mathematical function
- Relationship of input and output $f: \mathbf{x} \rightarrow y$
  - Input $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top \in \mathbb{R}^D$ is a D-dimensional vector
  - Output y is one dimensional
    - Regression: real-valued output $y \in \mathbb{R}$
    - Classification: discrete output $y \in \{C_1, C_2, \ldots, C_M\}$

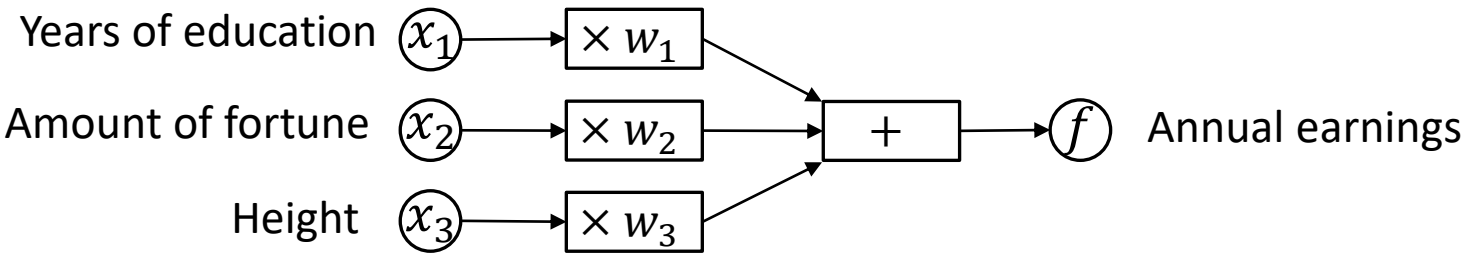


A model for regression:
Linear regression model

- Model f takes an input $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top$ and outputs a real value

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_D x_D$$

  - Model parameter $\mathbf{w} = (w_1, w_2, \ldots, w_D)^\top \in \mathbb{R}^D$

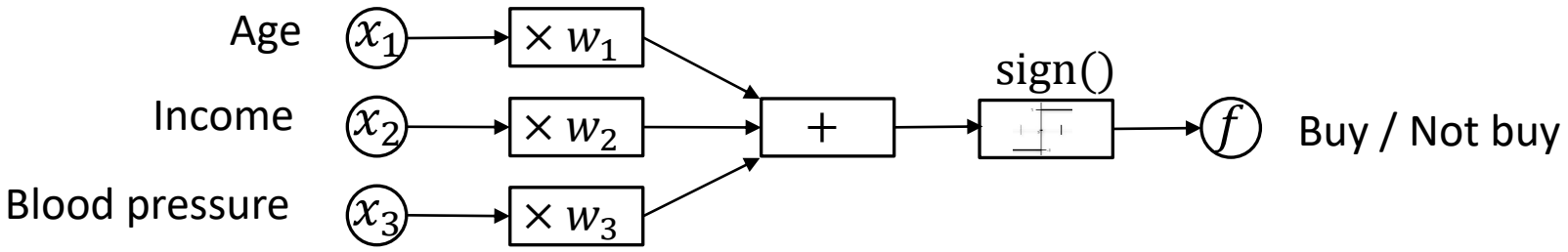


A model for classification: Linear classification model

- Model f takes an input $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top$ and outputs a value from $\{+1, -1\}$

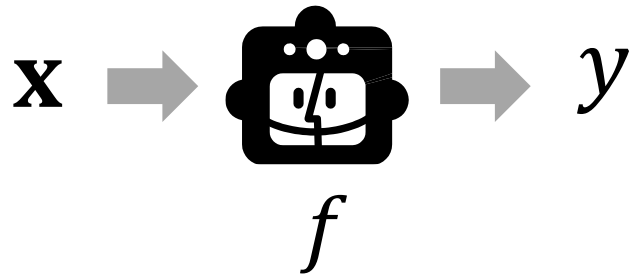
$$f(\mathbf{x}) = \mathrm{sign}(w_1 x_1 + w_2 x_2 + \cdots + w_D x_D)$$

  - Model parameter $\mathbf{w} = (w_1, w_2, \ldots, w_D)^\top \in \mathbb{R}^D$:
    - w_d : contribution of x_d to the output (if $w_d > 0$, $x_d > 0$ contributes to $+1$, $x_d < 0$ contributes to -1)



Formulations of machine learning problems: Supervised learning and unsupervised learning

- What we want is the function f
  - We estimate f from data
- Two learning problem settings: supervised and unsupervised
  - Supervised learning: input-output pairs are given
    - $\left\{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right)\right\}$: N pairs
  - Unsupervised learning: only inputs are given
    - $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\right\}$: N inputs



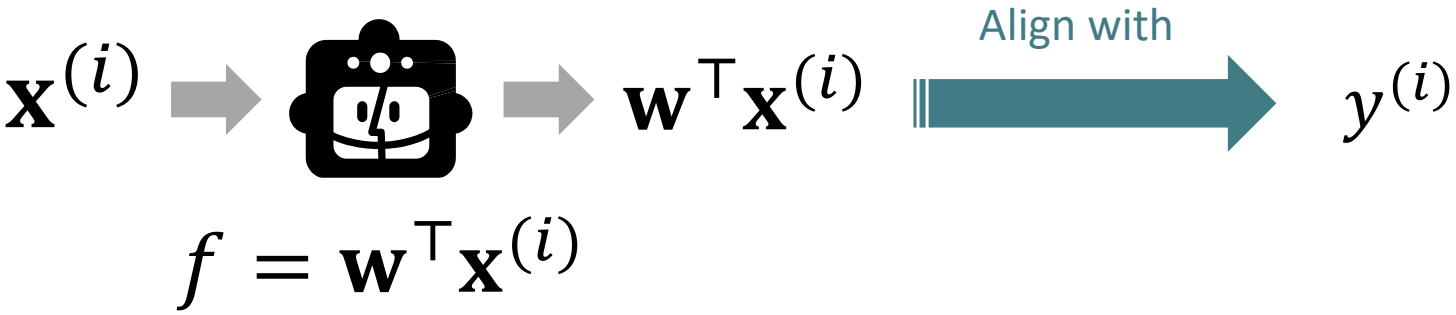
How do machines "learn" ?
Obtaining parameters that reproduce data

- In supervised learning, we want to obtain a function:

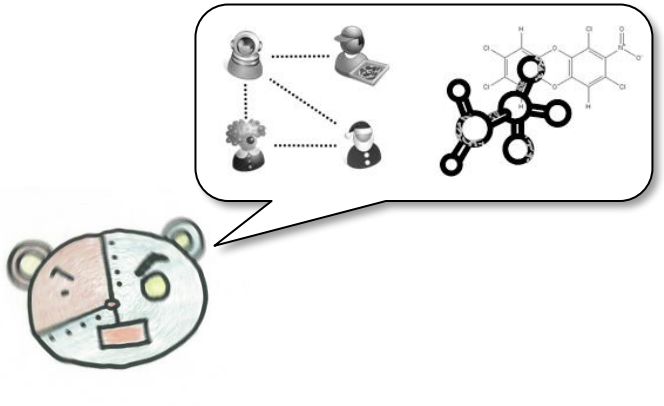
$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_D x_D = \mathbf{w}^\top \mathbf{x}$$

  - Fixing parameters $\mathbf{w}$ determines the model

- "Learning" = finding parameters that best reproduce data
  - Training data: $\left\{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right)\right\}$



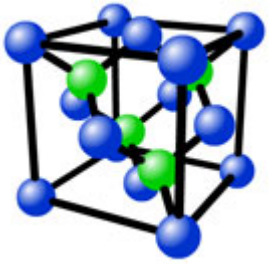
Machine learning applications



Growing ML applications:
Emerging applications from IT areas to non-IT areas

- Recent advances in ML offer:
  - Methodologies to handle uncertain and enormous data
  - Black-box tools
- Not limited to IT areas, ML is wide-spreading over non-IT areas
  - Healthcare, airline, automobile, material science, education, …




Various applications of machine learning: From on-line shopping to system monitoring

- Marketing
  - Recommendation
  - Sentiment analysis
  - Web ads optimization

- Finance
  - Credit risk estimation
  - Fraud detection

- Science
  - Biology
  - Material science

- Web
  - Search
  - Spam filtering
  - Social media

- Healthcare
  - Medical diagnosis

- Multimedia
  - Image/voice understanding
- System monitoring
  - Fault detection

Recommender systems: Personalized information filter

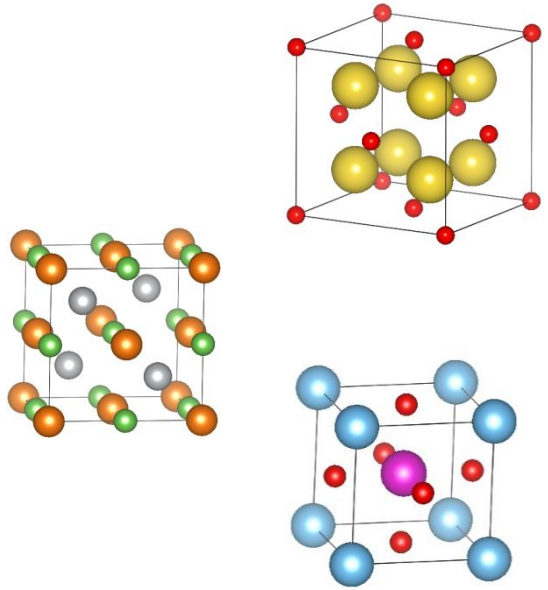
- Amazon offers a list of products I am likely to buy (based on my purchase history)

Ubiquitous recommender systems: Recommender systems are present everywhere

- A major battlefield of machine learning algorithms
  - Netflix challenge (with $100 million prize)
- Recommender systems are present everywhere:
  - Product recommendation in online shopping stores
  - Friend recommendation on SNSs
  - Information recommendation (news, music, …)
  - …

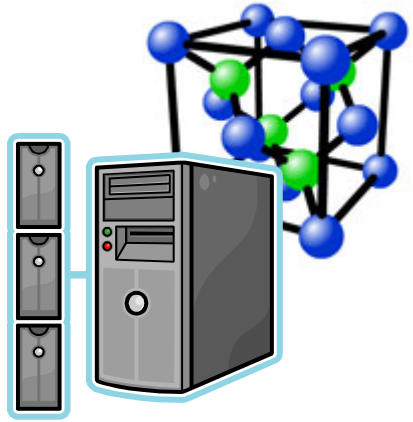
An application of supervised regression learning: Discovering new materials

- Material science aims at discovering and designing new materials with desired properties
  - Volume, density, elastic coefficient, thermal conductivity, …
- Traditional approach:
  1. Determine chemical structure
  2. Synthesize the chemical compounds
  3. Measure their physical properties



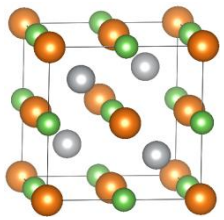
Computational approach to material discovery:
Still needs high computational costs

- Computational approach: First-order principle calculations based on quantum physics to run simulation to estimate physical properties
- First-order calculation still requires high computational costs
  - Proportional to the cubic number of atoms
  - Sometimes more than a month…



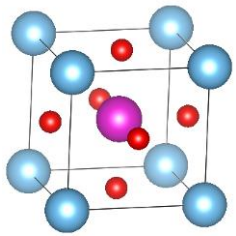
Data driven approach to material discovery: Regression to predict physical properties

- Predict the result of first-order principle calculation from data



Compound A

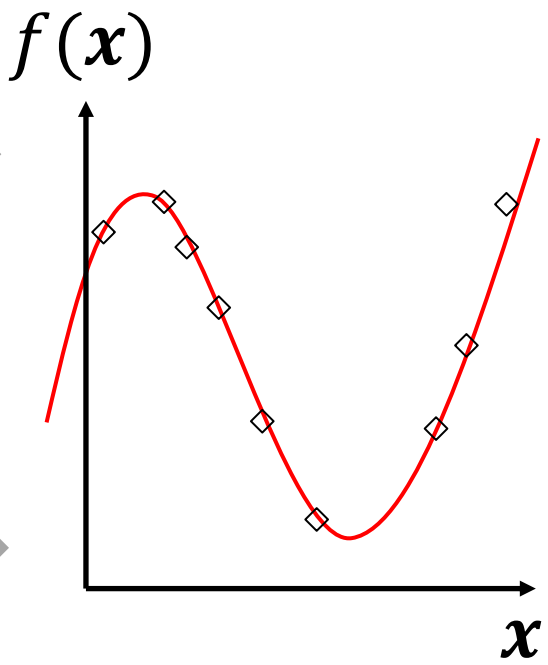
$$\boldsymbol{x}_{\mathrm{A}} = \begin{pmatrix} 1.5 \\ 140 \\ 3 \\ 72 \\ \vdots \\ 0.32 \end{pmatrix}$$



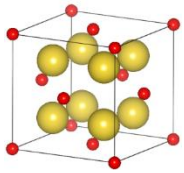
Compound B

$$\boldsymbol{x}_{\mathrm{B}} = \begin{pmatrix} 1.8 \\ 210 \\ 5 \\ 64 \\ \vdots \\ 0.15 \end{pmatrix}$$

Feature vector representation of chemical compounds



Estimate regression models of physical properties from data



New compound

$f(\boldsymbol{x})$

Physical properties

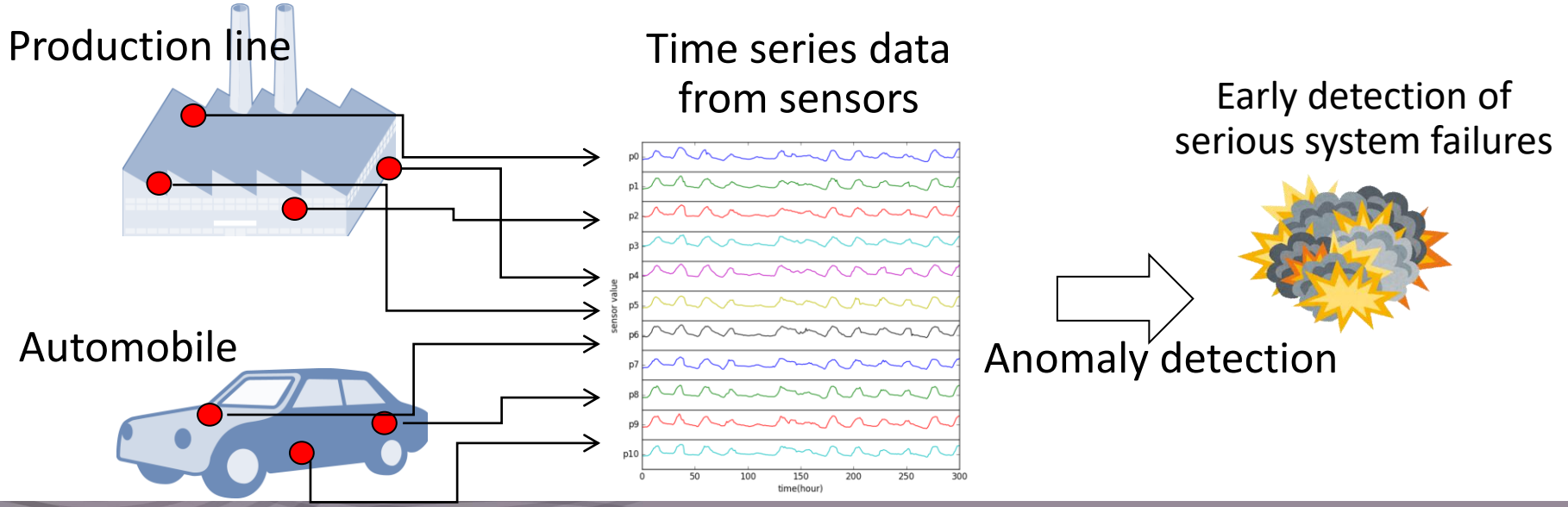
1.39
128
0.62

Predict physical properties of new compounds

Anomaly detection

Anomaly detection:
Early warning for system failures reduces costs

- A failure of a large system can cause a huge loss
  - Breakdown of production lines in a factory, infection of computer virus/intrusion to computer systems, credit card fraud, terrorism, …
- Modern systems have many sensors to collect data
- Early detection of failures from data collected from sensors



Anomaly detection techniques: Find “abnormal” behaviors in data

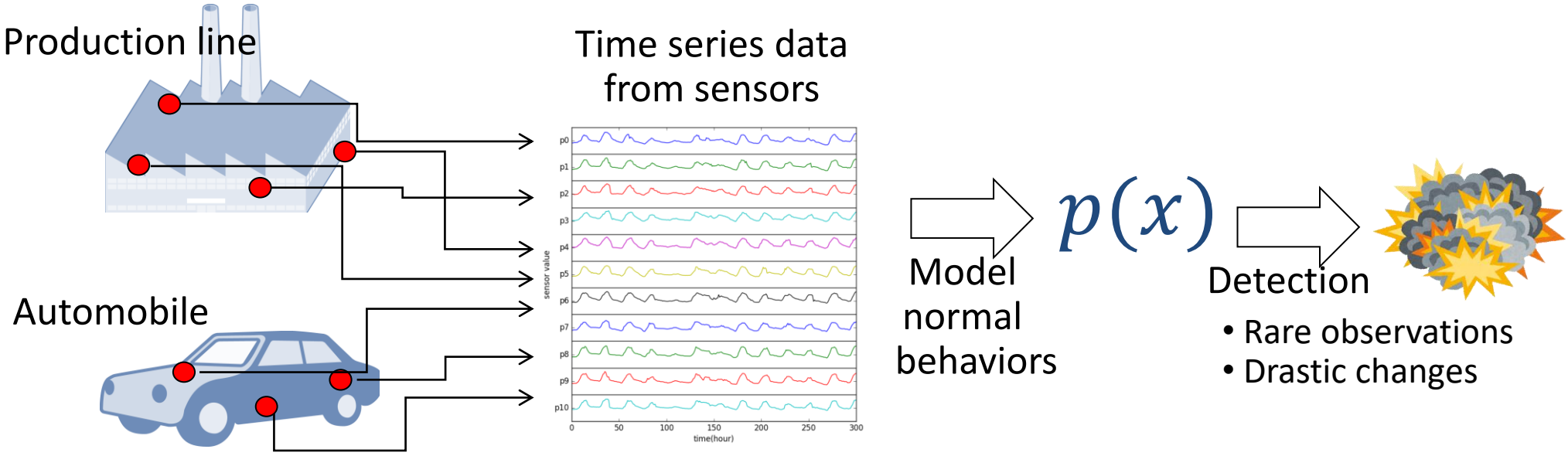
- We want to find precursors of failures in data
  - Assumption: Precursors of failures are hiding in data
- Anomaly: An “abnormal” patterns appearing in data
  - In a broad sense, state changes are also included: appearance of news topics, configuration changes, …
- Anomaly detection techniques find such patterns from data and report them to system administrators

Difficulty in anomaly detection: Failures are rare events

- If target failures are known ones, they are detected by using supervised learning:
  1. Construct a predictive model from past failure data
  2. Apply the model to system monitoring
- However, serious failures are usually rare, and often new ones
  → (Almost) no past data are available
- Supervised learning is not applicable

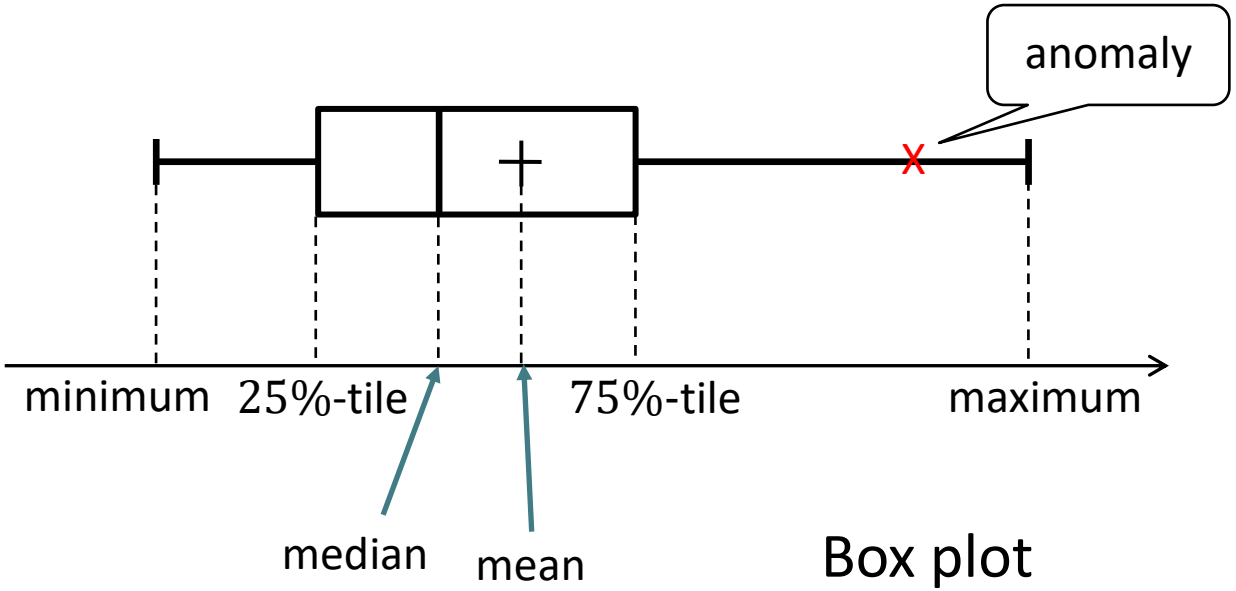
An alternative idea:
Model the normal times, detect deviations from them

- Difficult to model anomalies → Model normal times
  - Data at normal times are abundant
- Report "strange" data according to the normal time model
  - Observation of rare data is a precursor of failures



A simple unsupervised approach: Anomaly detection using thresholds

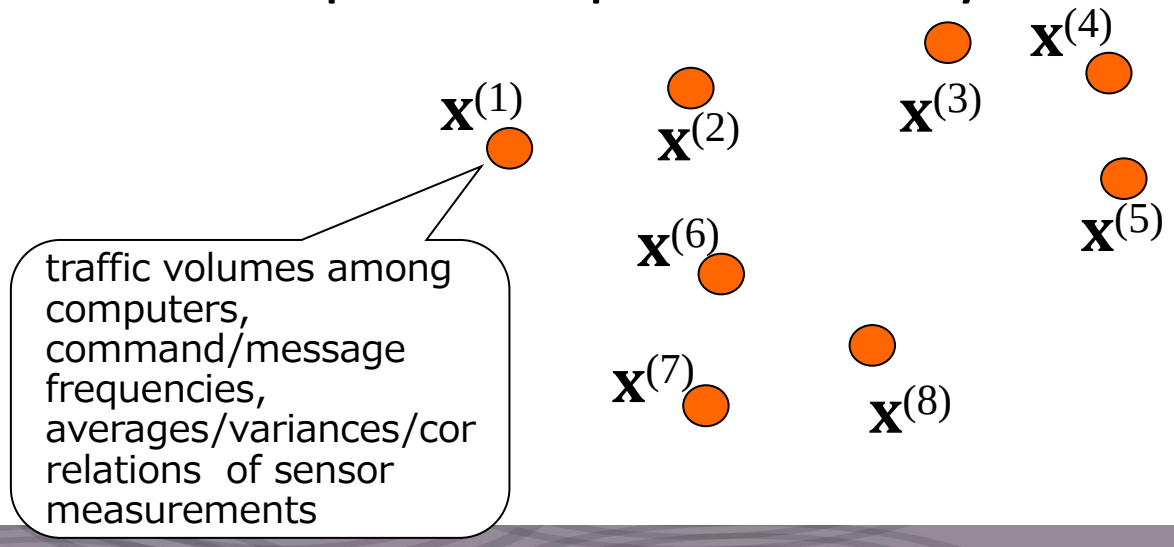
- Suppose a 1-dimensional case (e.g. temperature)
- Find the value range of the normal data (e.g. 20-50 ℃)
- Detect values deviates from the range, and report them as anomalies （e.g. 80℃ is not in the normal range)



Box plot

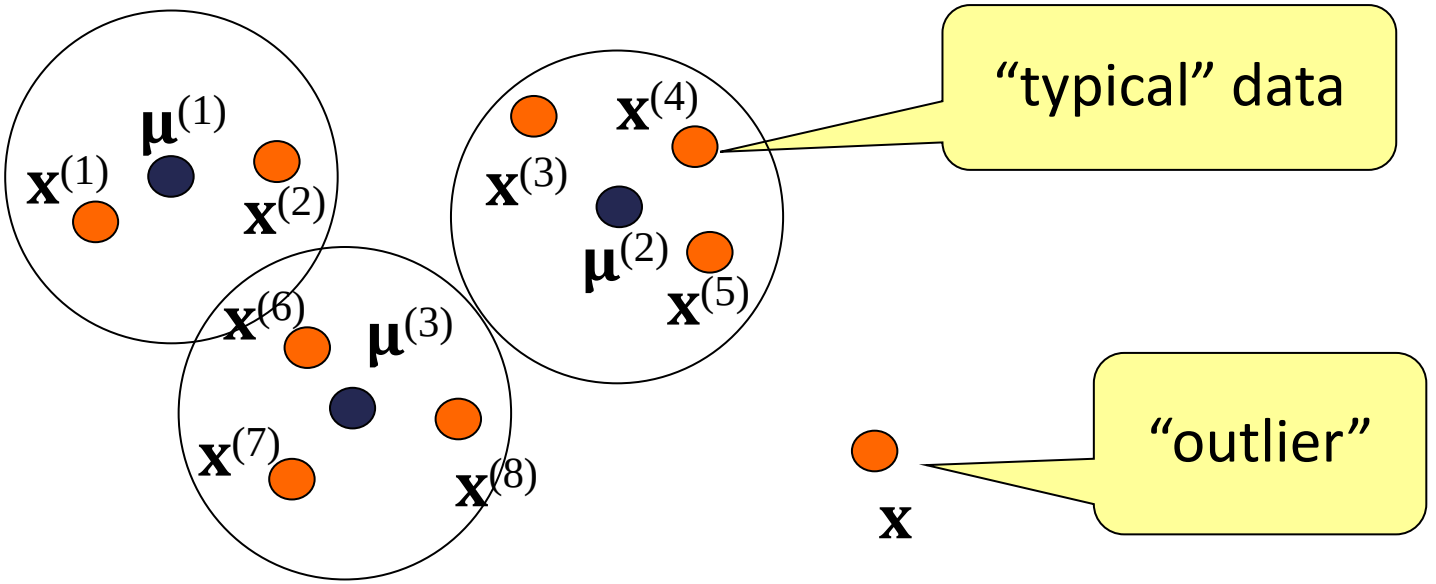
Clustering for high-dimensional anomaly detection: Model the normal times by grouping the data

- More complex cases:
  - Multi-dimensional data
  - Several operation modes in the systems
- Divide normal time data $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ into K groups
  - Groups are represented by centers $\{\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \ldots, \boldsymbol{\mu}^{(N)}\}$



Clustering for high-dimensional anomaly detection: Find anomalies not belonging to the groups

- Divide normal time data $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ into K groups
  - Groups are represented by centers $\{\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \ldots, \boldsymbol{\mu}^{(K)}\}$
- Data $\mathbf{x}$ is an "outlier" if it lies far from all of the centers
  = system failures, illegal operations, instrument faults



Anomaly detection in time series: On-line anomaly detection

- Most anomaly detection applications require real-time system monitoring
- Data instances arrive in a streaming manner:
  - $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(t)}, \ldots$: at each time t, new data $\mathbf{x}^{(t)}$ arrives
- Each time a new data arrives, evaluate its anomaly
- Also, models are updated in on-line manners:
  - In the one dimensional case, the threshold is sequentially updated
  - In clustering, groups (clusters) are sequentially updated

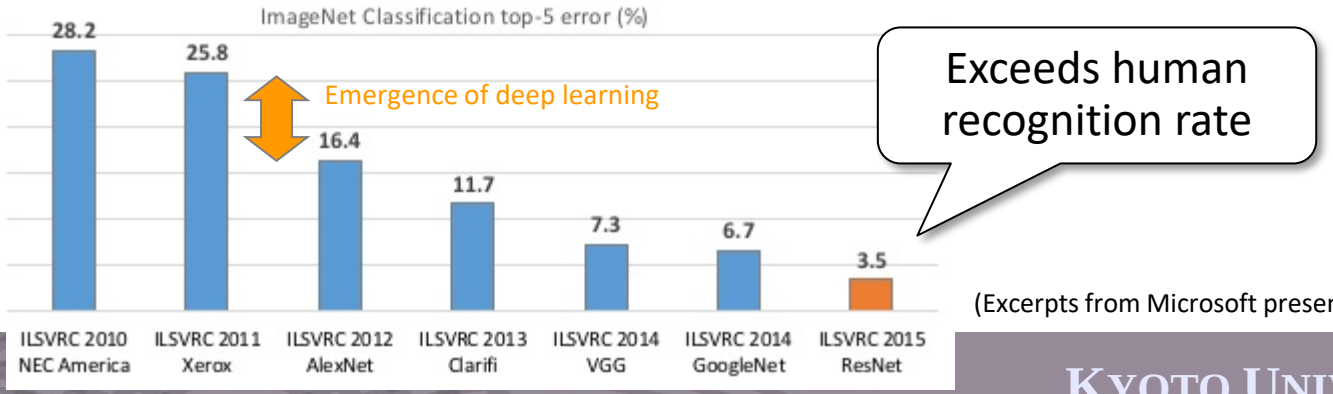
Limitation of unsupervised anomaly detection: Details of failures are unknown

- In supervised anomaly detection, we know what the failures are
- In unsupervised anomaly detection,
  we can know something is happening in the data,
  but cannot know what it is
  - Failures are not defined in advance
- Based on the reports to system administrators,
  they have to investigate what is happening, what are the reasons, and what they should do

Recent Advances in Machine Learning

Emergence of deep learning:
Significant improvement of prediction accuracy

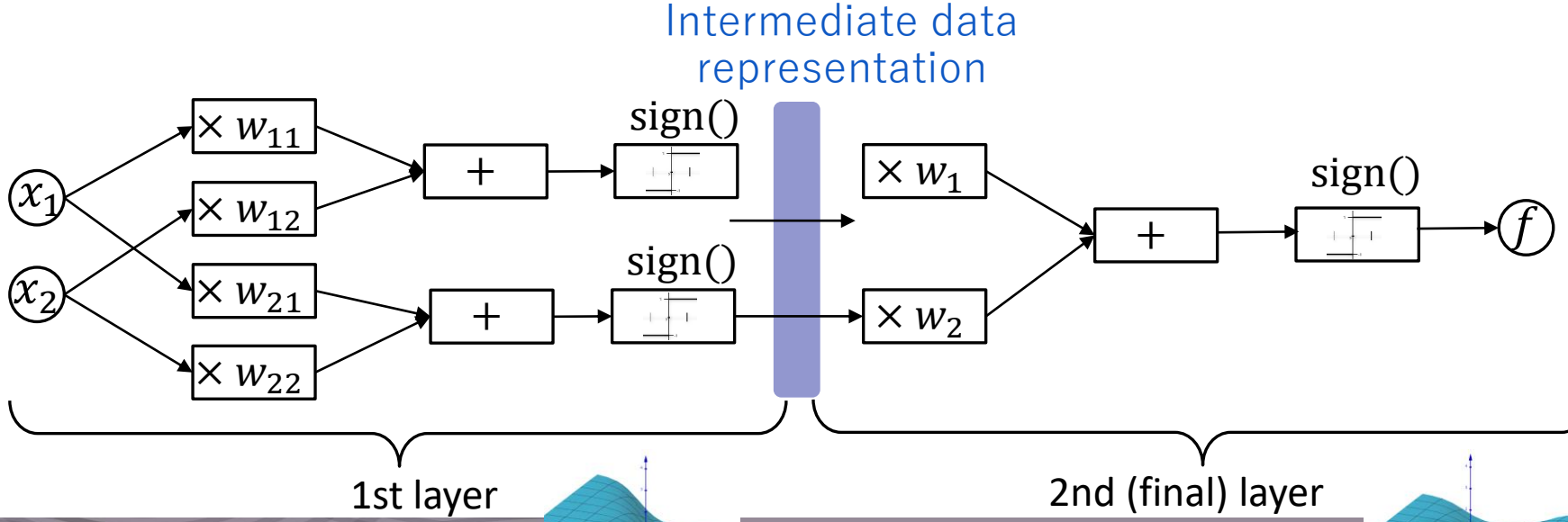
- Artificial neural networks were hot in 1980s, but burnt low after that…
- In 2012, a deep NN system won in the ILSVRC image recognition competition with 10% improvement
- Major IT companies (“Big Tech”) invest much in deep learning technologies
- Big trend in machine learning research



(Excerpts from Microsoft presentation materials)

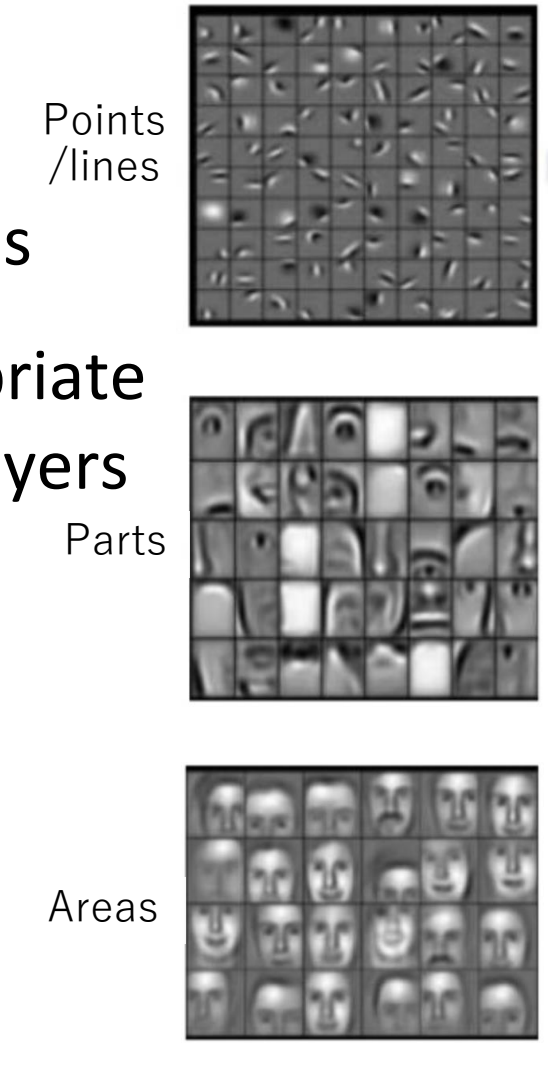
Deep neural network:
Stacked linear classifiers are more powerful

- Essentially, multi-layer neural networks are regarded as stacked linear classification models
  - First to semi-final layers bear feature extraction, and final layer makes predictions
- Deep stacking adds significant non-linearity to the model, ensuring enhanced representational power



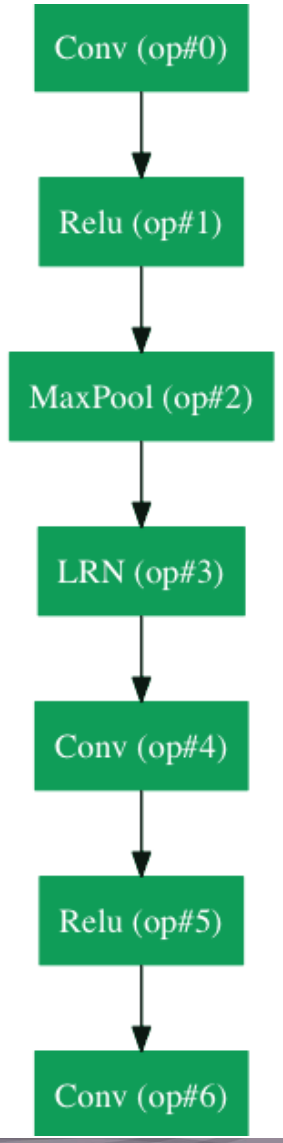
Representation learning in deep neural network: Obtaining task-useful abstractions in the middle layers

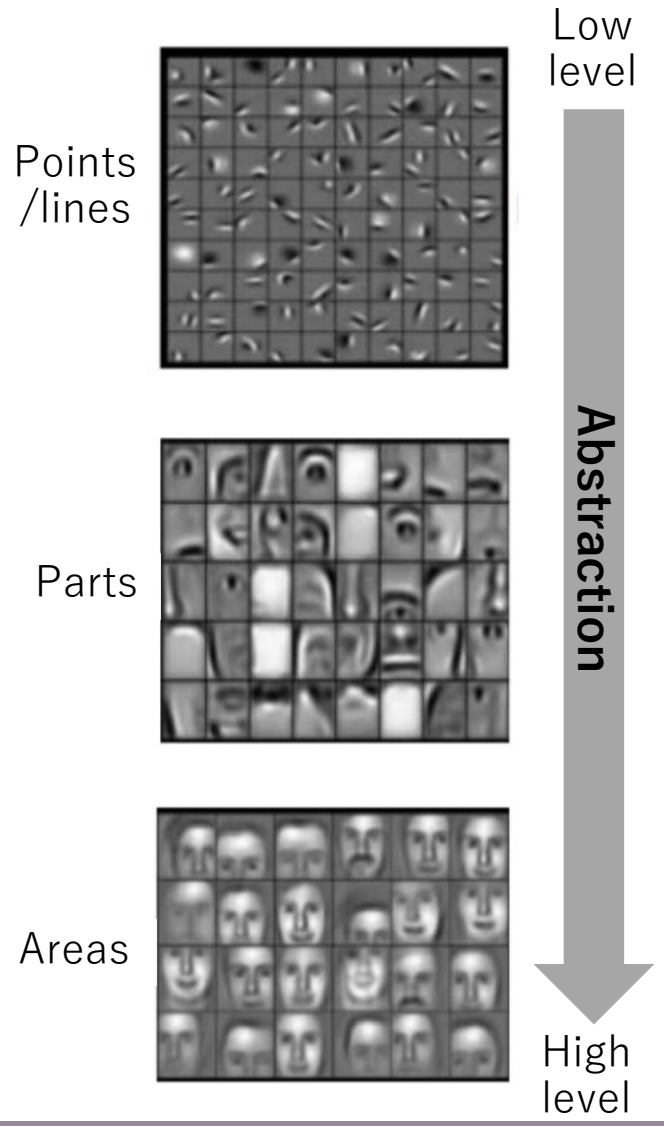
- Deep learning acquires abstract data representations that suit its purpose
  - via dozens to thousands transformations
- Representation learning: acquires appropriate intermediate expressions in the middle layers
- Recycling middle layers: pre-training, transfer learning, meta-learning, ...



https://adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html

Representation learning in deep neural network:
Obtaining task-useful abstractions in the middle layers



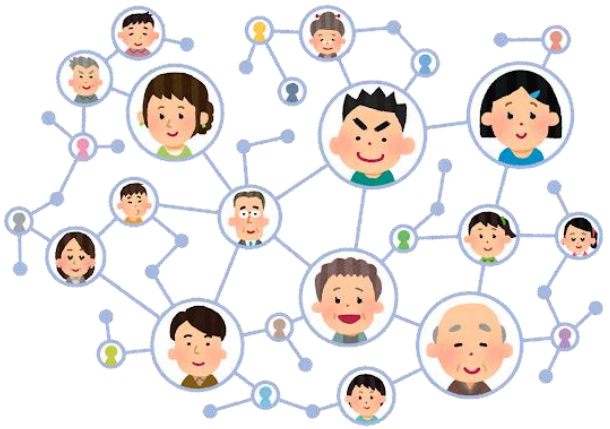


https://adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html

What is the difference of DL from the past NNs?: Deep structures, new techniques, & accessible frameworks

- Differences from the ancient NNs:
  - Far more computational resources are available now
  - New structures:
    - from wide-and-shallow to narrow-and-deep structures
    - CNN, LSTM, attention (Transformers), diffusion models,…
  - New techniques: Dropout, ReLU, batch normalization, adversarial learning, …
  - Easy-to-use frameworks: Pytorch, TensorFlow, …
- They serve as “platforms” for development of modern data-driven systems

Machine Learning with Graphs



Expansion of machine learning: Analysis of a wide variety of data formats

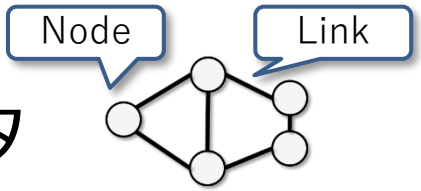
- Machine learning has adapted to new types of data that arise over time
  - with inductive bias based on data domains/types

i.e., prior knowledge

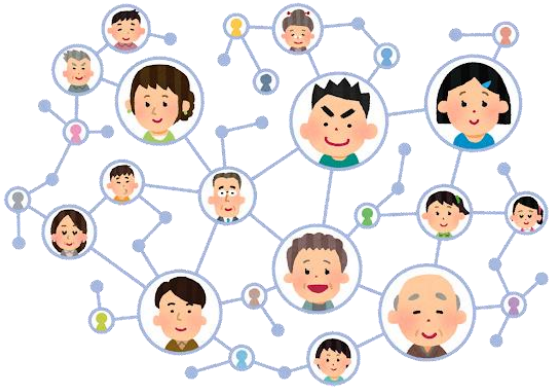
- Three "V"s in big data:
  1. Volume (莫大なデータ)
  2. Velocity (高速度でのデータ入出力)
  3. Variety (多種多様なデータ源と種類)
     - Heterogeneous data integration and feature engineering

Graphs:
Can express various "relationships" in the world

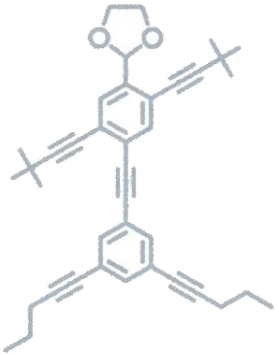
- グラフ構造データ：構成要素（頂点）と
  要素間の「関係」（リンク）によって表現されたデータ



- 世の中にあるグラフ構造データ：文書、構文木、Web、
  XML/HTML、化合物、ソーシャルネットワーク、DNA／タンパク質
  配列、RNA生体ネットワーク、引用関係、企業間取引、…



Social networks

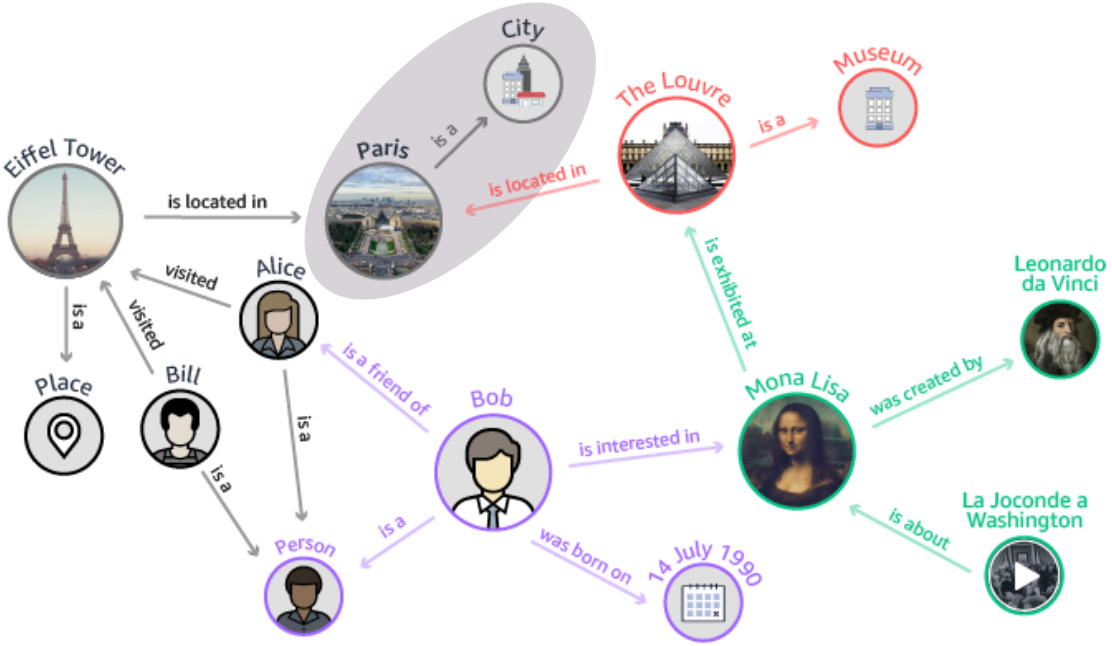


Chemical compounds

DNA sequences

Knowledge graphs:
Graph representation of world knowledge

- 自然言語処理の問題では様々な知識は暗黙的に用いられる
  - 猫が哺乳類だったり液体であることなどは断りなく用いられる
- 世界の知識を「オブジェクト」間の「関係」の集合として捉えたもの：
  - Is_a(Paris, City)：「パリ」と「町」は「である」の関係にある

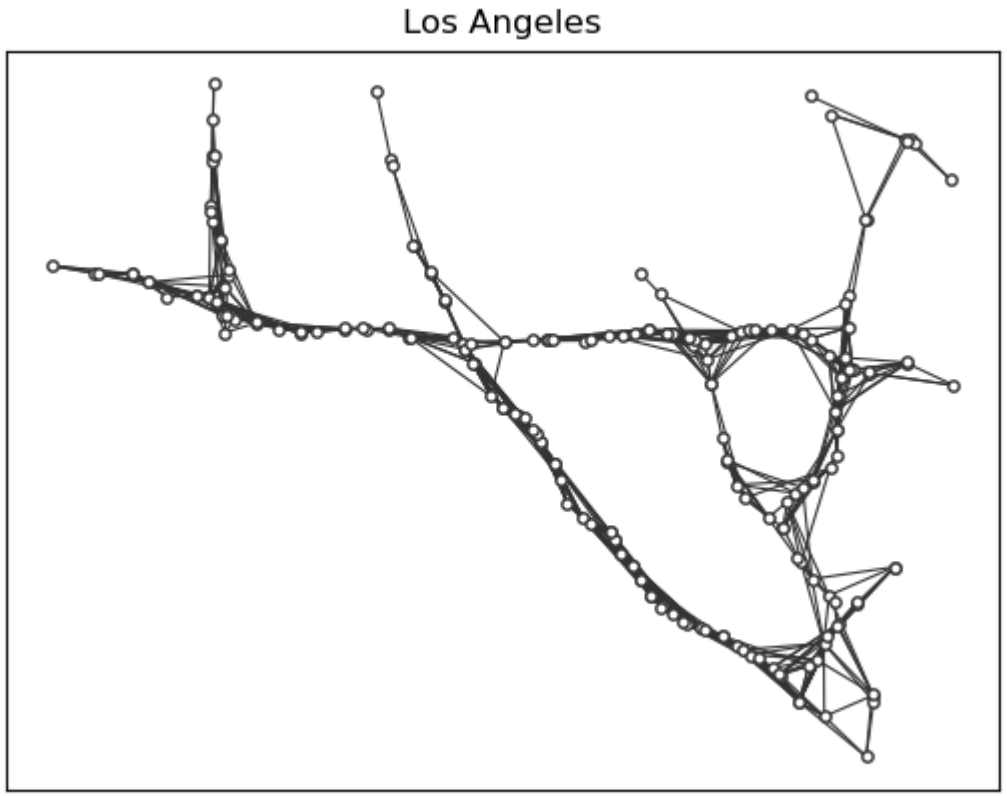


https://www.dlf.pt/ddetail/hmiJmoT_knowledge-graphs-use-case-diagram-knowledge-graph-example/

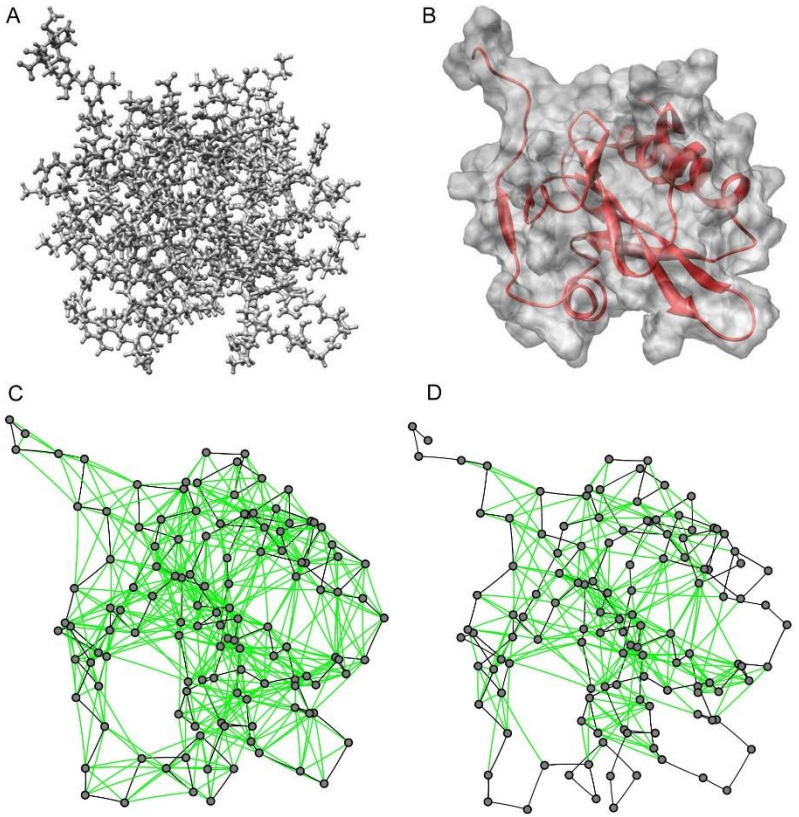
Spatial graph:
Graphs representing spatial adjacency structure

- 平面・空間中のオブジェクトの隣接関係を表現したグラフ

Traffic network



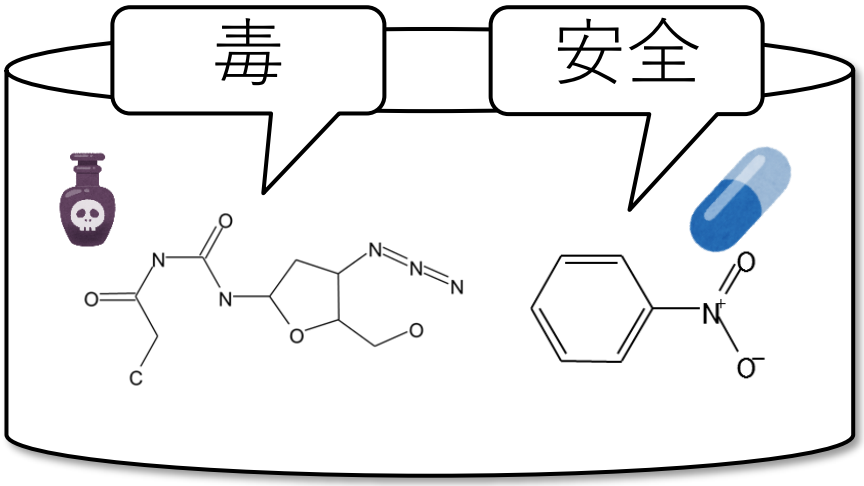
Protein structure



https://sunflower.kuicr.kyoto-u.ac.jp/~sjn/folding/methodology.html

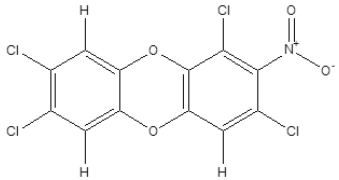
Graph learning:
Prediction with graph-structured data

毒性予測
（グラフ分類問題）

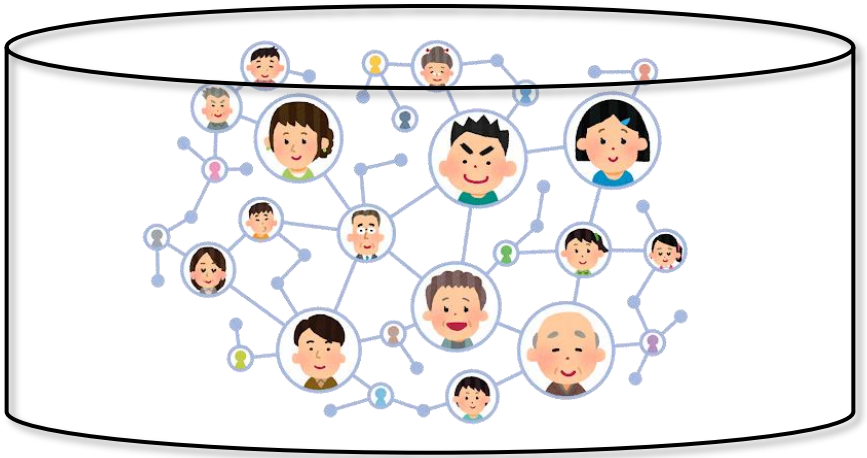


予測

これは毒？安全？



SNSの友人推薦
（リンク予測問題）



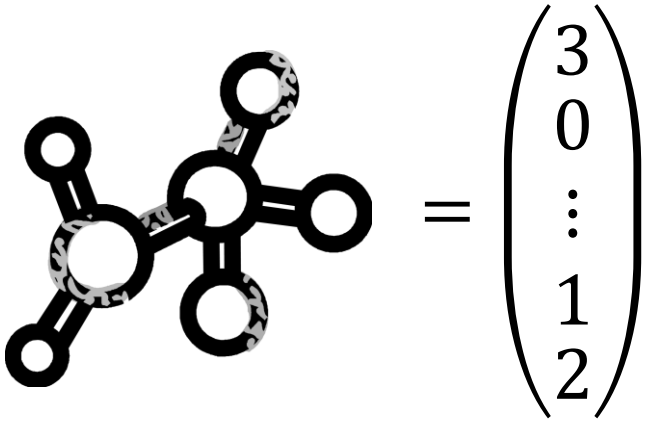
予測

二人は友達？

Challenges in graph learning:
Obtaining vector representations of graphs

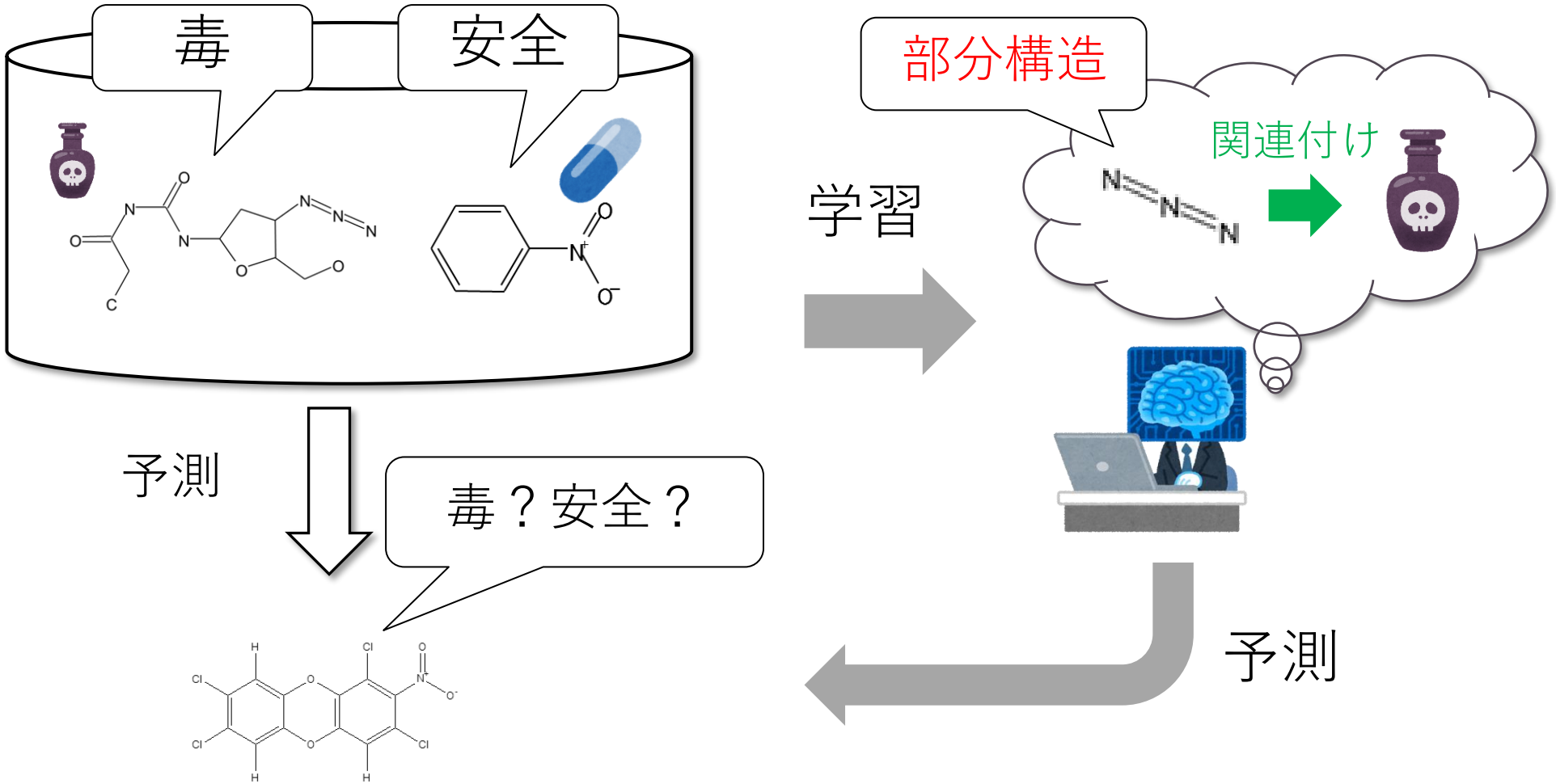
- どんなデータもベクトルとして表現できればあとは同じはず
- 機械学習（を使う各ドメイン）ではそれぞれのデータ形式に合わせて、よい表現（ベクトル）の設計を試行錯誤してきた
  - 画像のSIFT、音声のMFCC、化合物の分子フィンガープリント、…
- グラフの表現学習問題：グラフはそのままではベクトルではない
  - 同じグラフは同じ表現に
  - 異なるグラフは異なる表現に
  - 似たグラフは似た表現に

  なってほしい…



Basic strategy for graph learning:
Capturing *substructures* of graphs

- 入力グラフの「部分構造」と出力の関係を学習



Graph deep learning:
Neural network that can handle substructures of graphs

- グラフの部分構造特徴抽出にニューラルネットワークを利用
- グラフ畳み込みニューラルネットワーク：
  - 画像畳み込みニューラルネットワーク（CNN）：
    各ピクセルがその近傍ピクセルの情報を取り込む
  - グラフ畳み込み：各頂点が周辺頂点の情報を取り込む

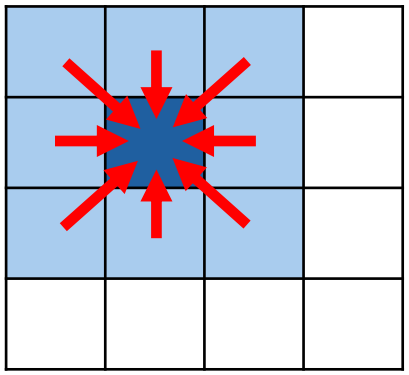
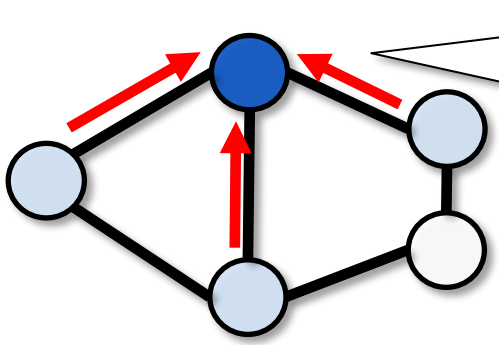


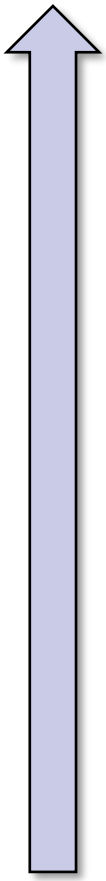
Image convolution

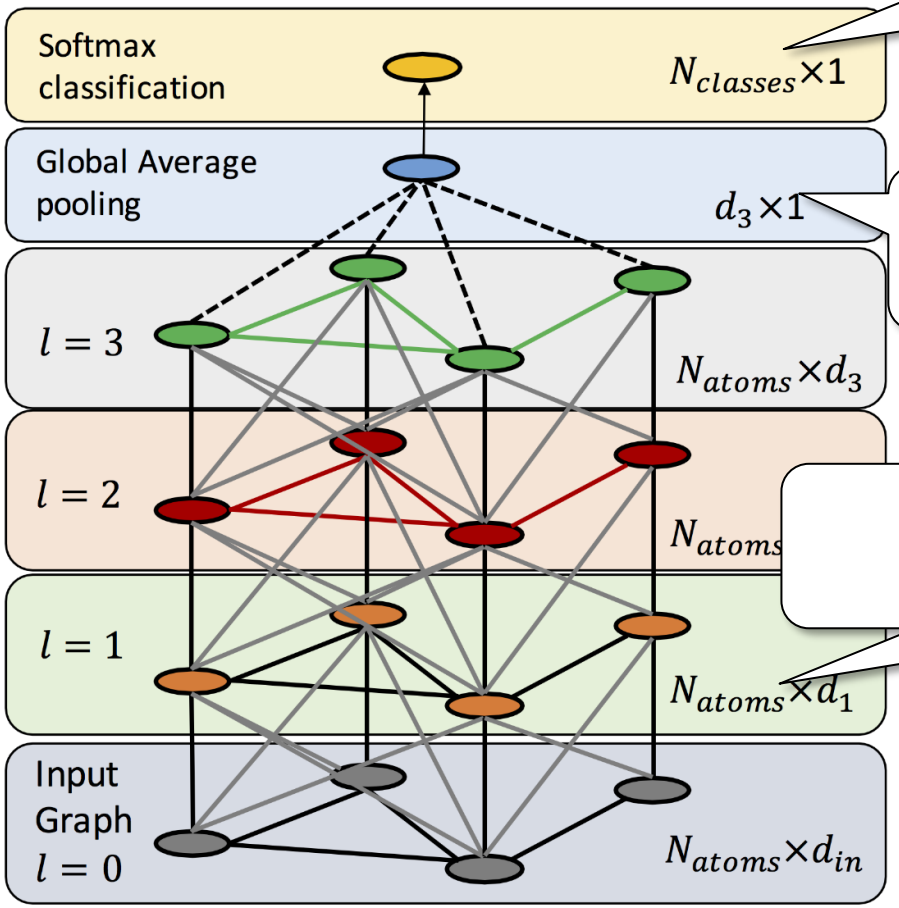


Graph convolution

Structure of a Graph Neural Network (GNN):
One layer corresponds to aggregation of neighbor nodes

入力から
出力を予測

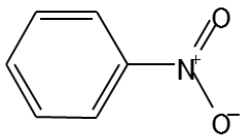




グラフ全体の表現を
もとに予測

全頂点の表現を集計して
グラフ全体の表現をつくる

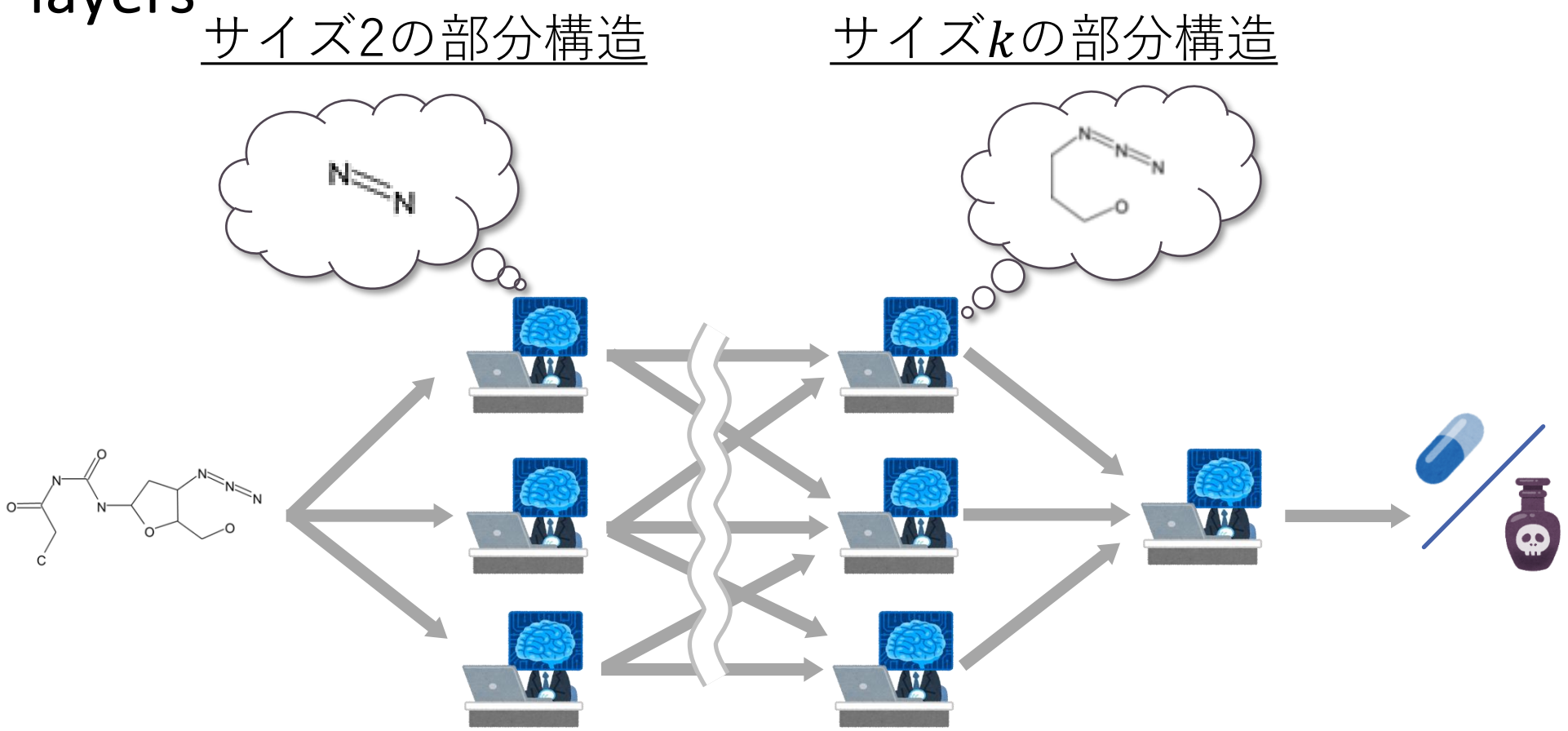
頂点表現の1回の更新が
GNNの1層に対応



Pope *et al*. *CVPR* (2019) の図を改変

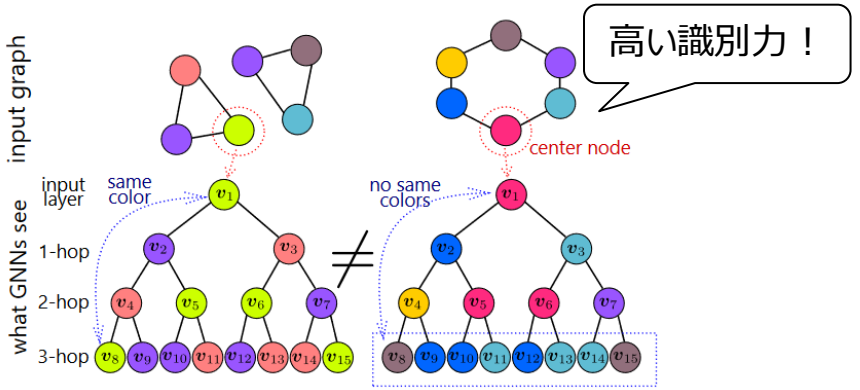
Recognition of graph structure by GNN:
Recognition of large substructures by multiple layers

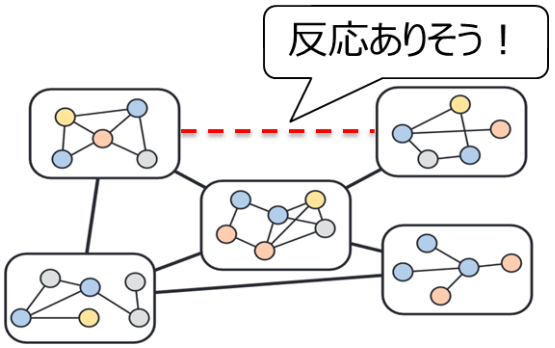
- GNNs acquire larger structural features with multiple layers

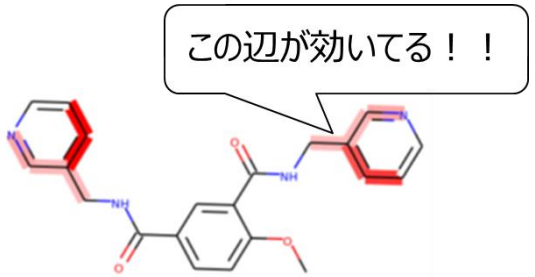


Advances in graph deep learning: Performance improvements, expanded applicability, XAI

- グラフ深層学習の表現力強化
  - ランダム属性追加
- グラフのグラフ上でのリンク予測
  - 「グラフのグラフ」ニューラルネット
  - 化合物（グラフ）同士の反応予測
- グラフの重要箇所説明
  - グラフXAI
  - 薬剤の活性箇所を特定







Sato et al. (2021), Harada et al. (2020), Akita et al. (2018)

Advanced applications of GNNs
Supporting research & development

- Science of Science：膨大な学術知識を定量分析し、エビデンスに基づく科学技術イノベーション政策や研究開発戦略を実施
- 大規模な学術知識グラフデータセットを公開する流れ：
  - Semantic Scholar (米国・Allen AI Institute)
  - Aminer (中国・清華大学)
  - Academic Knowledge Graph (米国・Microsoft Research)
- 大規模学術論文データを解析して得られた知識・知見を技術ロードマッピングや技術予測に活用する

Advanced applications of GNNs
AI for prediction & discovery of new research trends

- 大規模学術論文データを解析して得られた知識・知見を技術ロードマッピングや技術予測に活用する
- 既存研究の将来予測：学術知識グラフの成長モデルに基づく研究分野・トピック・論文・研究者の将来予測
- これから出現する研究の発見：コンセプト間グラフの予測に基づく新規トピック出現予測

※ 学術知識グラフ：
大規模学術論文データから構築した
引用や共著関係等知識のグラフによる表現

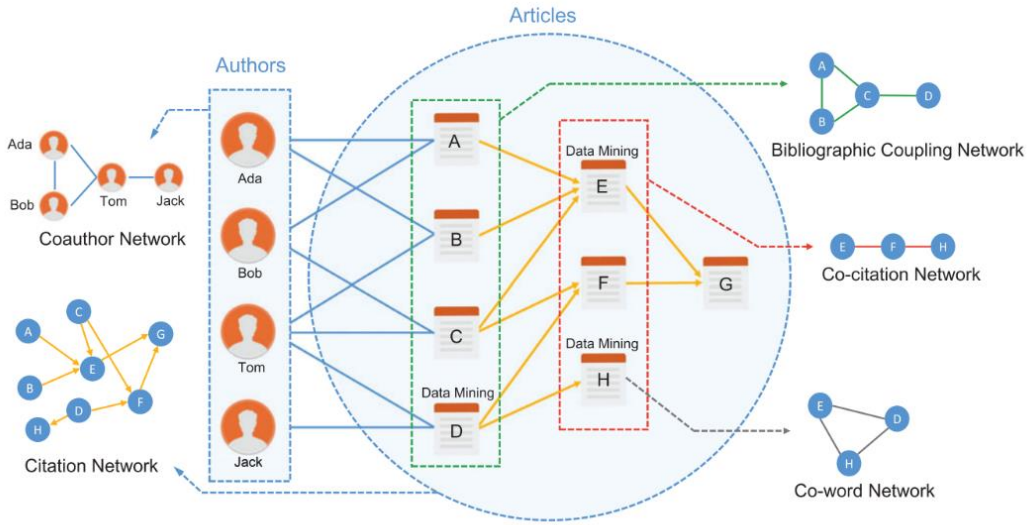


Fig. 5. Four most popular types of scholarly networks.

Xia, Feng, et al. "Big scholarly data: A survey." *IEEE Transactions on Big Data* 3.1 (2017): 18-35.

Treatment Effect Prediction

“Deeper” predictive modeling

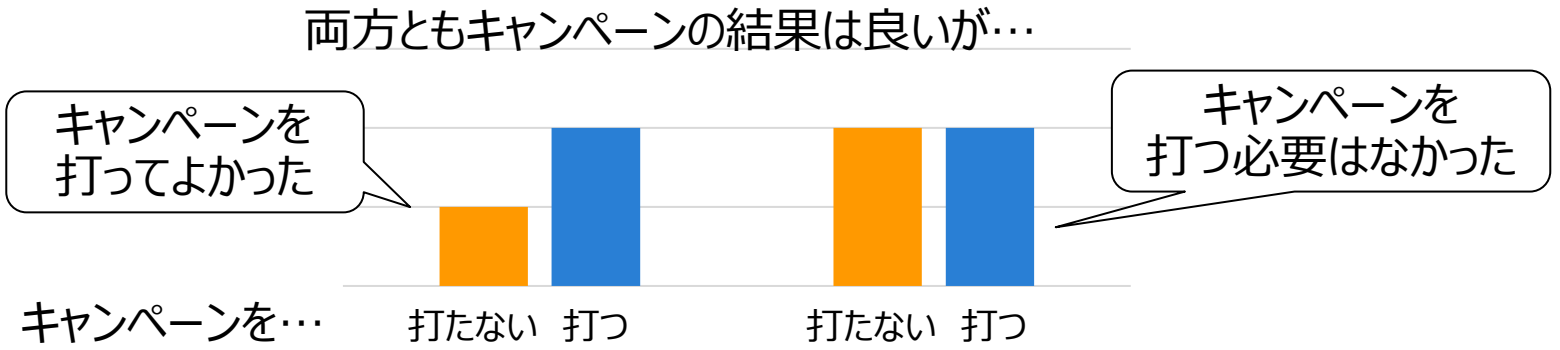
Predictions taking into account the effects of treatments

- In many applications, predictions aim at making decisions
- We want to encourage a specific behavior (outcome) by applying some action (treatment to a person (subject)
- Examples:
  - Issue a coupon to a user, and expect a purchase behavior

  - Provide a certain treatment to a patient, expecting the patient to be cured

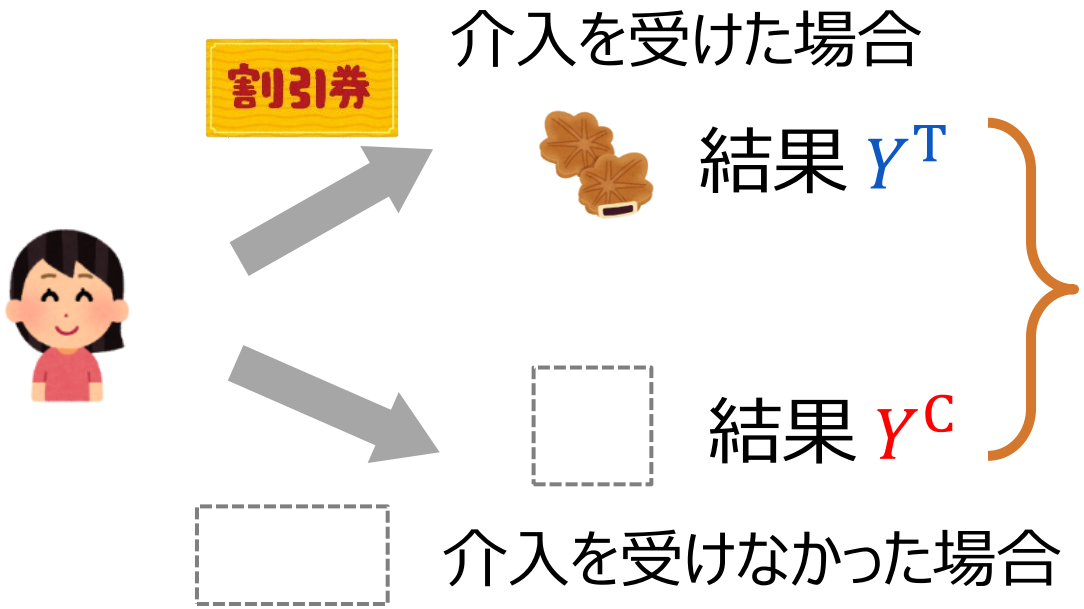
An issue in usual predictive modeling:
Not accounting for the effects of decision-making

- 従来の予測モデリングにもとづく意思決定：
  - 過去のキャンペーン対象ユーザが購入したかどうかのデータから各ユーザが購入するかどうかを予測するモデルを学習
  - このモデルが予測する購入確率が高そうな人にキャンペーンをうつ
- 疑問：キャンペーンを打たなくても買う人はいるのでは？
  - キャンペーンによる効果（＝購買可能性の増分）をみていない



Treatment Effect:
Quantification of strength of causal relationships

- 因果効果＝介入を受けた場合と受けなかった場合の結果の違い
- 介入効果：介入を受けた場合の結果Y^{T}と、受けなかった場合の結果変数Y^{C}の差 $Y^{\mathrm{T}} - Y^{\mathrm{C}}$　T：Treatment　C：Control
  - 観測されない「反事実」との比較（あちら側の世界の出来事  ）



Difference from ordinary prediction problems: Predicting treatment effects from past treatment results

- Ordinary prediction:
  - $\mathbf{x}$：Features of subject (性別、年齢など)
  - y：outcome

  are given as training data

- Treatment effect prediction:
  - $\mathbf{x}$： Features of subject (性別、年齢など)
  - z：Treatment
  - y： outcome

    購入の有無 or 購入金額

  are given as training data

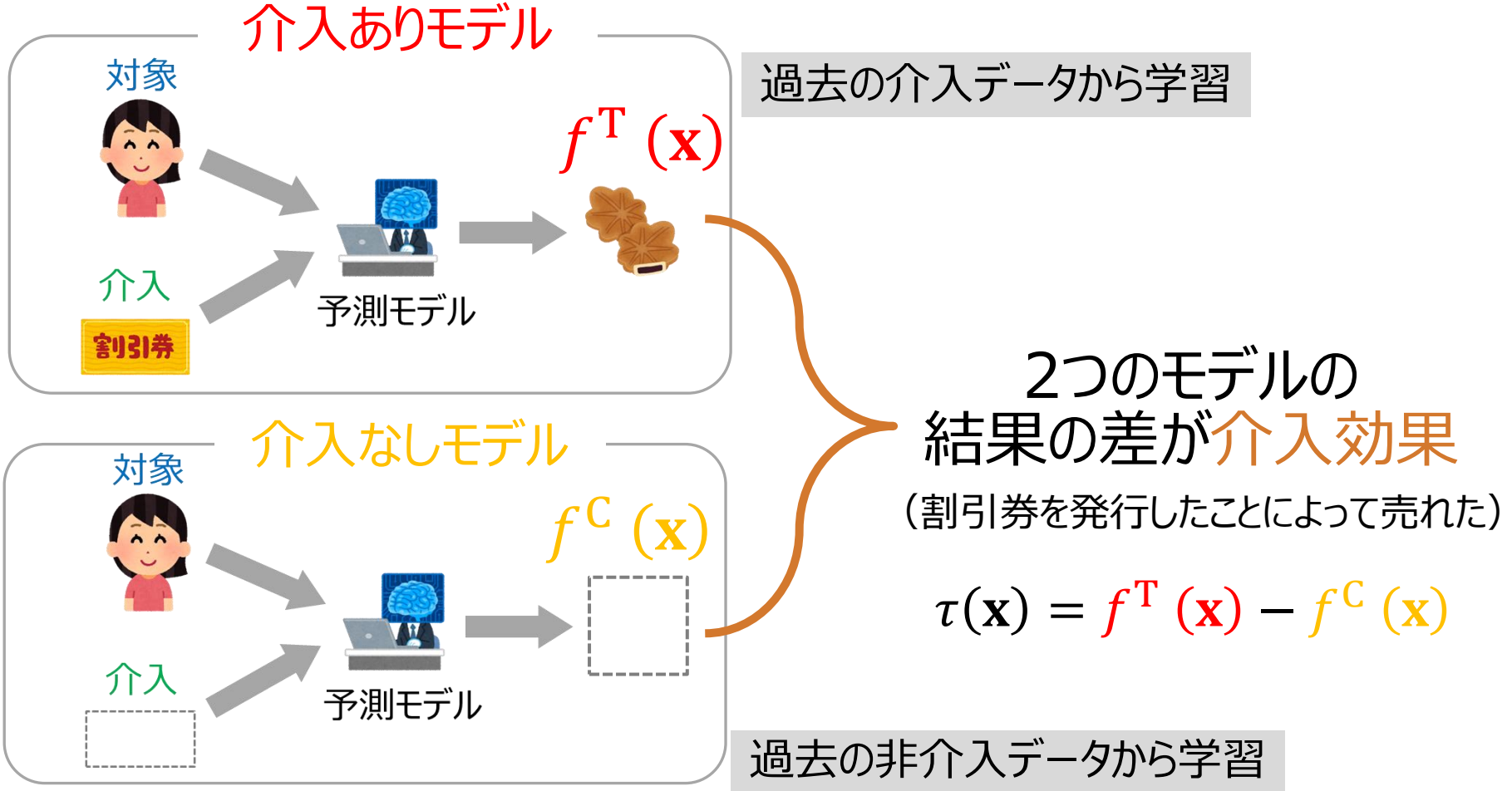
$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

Training data

$$\{(\mathbf{x}_i, z_i, y_i)\}_{i=1}^{N}$$

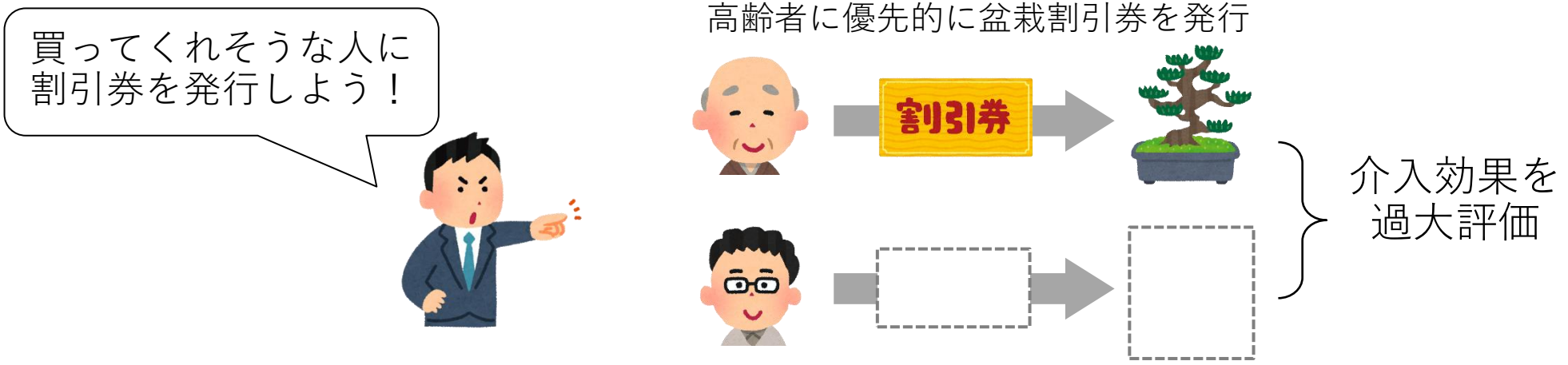
Prediction of treatment effects:
Predict difference in outcomes with & without treatment

- 介入がある場合とない場合の結果の予測モデルをつくる



Difficulty in treatment effect prediction:
Bias in past decisions in data collection biases estimation

- 過去のデータには、その時の介入判断に偏りがある
  - いかにも買いそうな人に割引券を発行した（or その逆）
- 偏ったデータからは偏った介入効果が導かれうる
  - 介入効果を過大評価・過小評価する恐れ
  - 理想的にはランダムに介入された（RCT）データが必要



Difficulty in treatment effect prediction:
Bias in past decisions in data collection biases estimation

- 偏ったデータからは介入効果を過大評価・過小評価するリスクがある
- 理想的にはランダムに介入された（RCT）データが必要
  - もしくはA/Bテストとも呼ばれる
  - ただし、いつでも実行可能ではない

- 偏ったデータから、偏りのない因果効果を推定する方法
  - 統計学や計量経済学で発展

Approaches to treatment effect prediction:
Estimating unbiased treatment effects from biased data

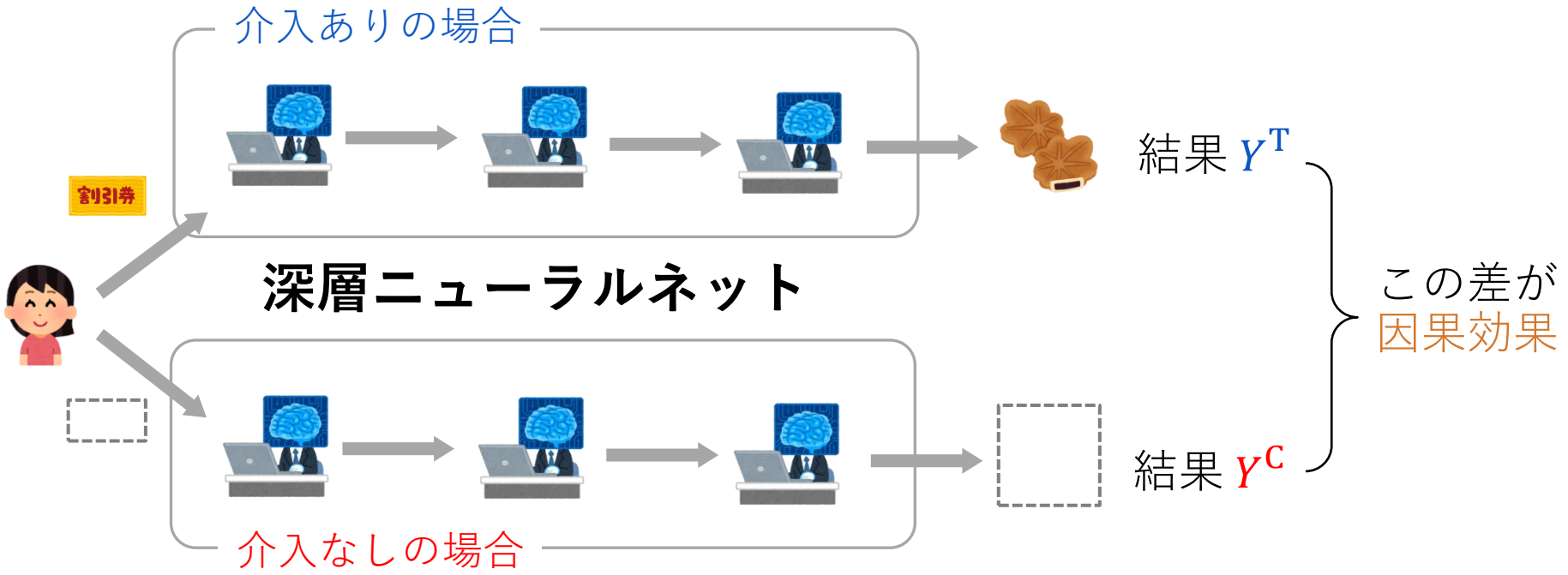
- 偏ったデータから、偏りのない介入効果を推定する方法
- マッチング：ドッペルゲンガーを探せ作戦
  - よく似たデータをペアにして比較

- 傾向スコア重みづけ：あたかもRCT作戦
  - 傾向スコア：介入の確率
    - RCTの倍の確率で介入されたデータは1/2の重み
    - 半分の確率で介入されたデータは2倍の重み

Treatment effect prediction using deep learning: DNN helps improving prediction performance

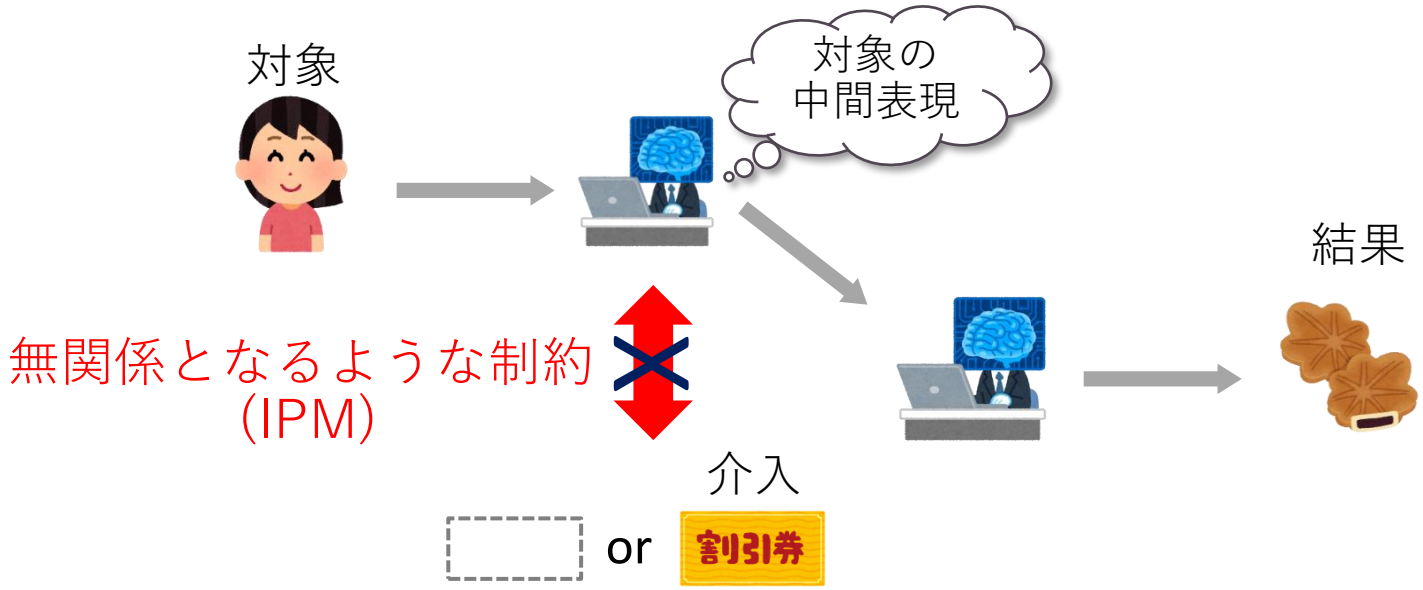
- 介入効果予測でも深層学習を使いたい！
  - モデルを複雑にして予測精度を上げる
  - 偏ったデータからは偏った予測結果が得られるので補正したい

偏ったデータをもとに
RCTデータで予測を当てる勝負



Ideas of “deep” treatment effect prediction:
Representation where subject & treatment are independent

- データのバイアスの問題：データが偏ると結果も偏る
- 傾向スコア重みづけは「あたかもランダム化試験」
  - 入力から介入の有無が予測できないようにする
- 深層学習では、中間表現において対象と介入の関係を断ち切る



Shalit et al. (2021),

Conclusion

Summary:
Basic ideas of machine learning and recent advances

1. What is machine learning?
2. Machine learning applications
3. Recent advances:
   - Deep learning
   - Graph deep learning
   - Treatment effect prediction