Data Mining: Advanced Techniques

Introduction

- 主讲教师: 陈佳伟, sleepyhunt@zju.edu.cn
  - https://jiawei-chen.github.io/
- TA: 陈思睿, chenthree@zju.edu.cn

Grading Policy

- Literature review (60%)
  - Written assignment (>3000 words)
- 2-3 homework (30%)
- 1-2 In-class quiz (10%)

Potential Audiences

- What can we obtain from this course?
  - Basic **knowledge** about the data mining, machine learning, graph neural network, recommender systems, LLMs, etc.
    - Knowledge for research
    - Knowledge to `fight’ the interviewer
  - Something about **how to do research**: basic thoughts and Methodology
    - How to find the idea
    - How to write the paper
    - how to conduct the analyses
  - Some promising **directions** on these topics
  - **Grades!**

Outline (Preview)

- 1-2th lesson: Basic Data mining & Data processing
- 3-4th lesson: Frequent
- 5-6th lesson: Classificat
- 7-8th lesson: Clustering

Outline

- 1-2th lesson: Basic Data mining & Machine Learning
- 3-4th lesson: Graph Mining & Graph neural network
- 5-6th lesson: Recommender Systems
- 7-8th lesson: Large Language Models

Why these topics?

- Topics on KDD 2023
  - Machine Learning (Over 90% papers)
  - Graph Mining (≥90/313)
  - Recommender Systems(≥37/313)

- Graph and Recommendation are most influential topics in data mining.

Why these topics?

- Topics on KDD 2024
  - Machine Learning (Over 90% papers)
  - Graph Mining (≥130/410)
  - Recommender Systems(≥46/410)
  - LLM+X (≥40/410)

Why these topics?

- Best paper on KDD 2024

Best Paper Award – Research

CAT: Interpretable Concept-based Taylor Additive Models

Viet Duong, Qiong Wu, Zhengyi Zhou, Hongjue Zhao, Chenxiang Luo, Eric Zavesky, Huaxiu Yao, Huajie Shao

Best Student Paper Award – Research

Dataset Regeneration for Sequential Recommendation

Mingjia Yin, Hao Wang, Wei Guo, Yong Liu, Suojuan Zhang, Sirui Zhao, Defu Lian, Enhong Chen

Best Paper Award – Applied Data Science

LiGNN: Graph Neural Networks at LinkedIn

Why these topics?

- Top-10 most influential papers in KDD 2022
  - 1. GraphMAE: Self-Supervised Masked Graph Autoencoders
  - 2. Graph Attention Multi-Layer Perceptron
  - 3. Learned Token Pruning for Transformers
  - 4. GraphWorld: Fake Graphs Bring Real Insights for GNNs
  - 5. Contrastive Cross-domain Recommendation in Matching
  - 6. Graph Neural Networks for Multimodal Single-Cell Data Integration
  - 7. Proton: Probing Schema Linking Information from Pre-trained Language Models for Text-to-SQL Parsing

https://www.paperdigest.org/2023/04/most-influential-kdd-papers-2023-04/

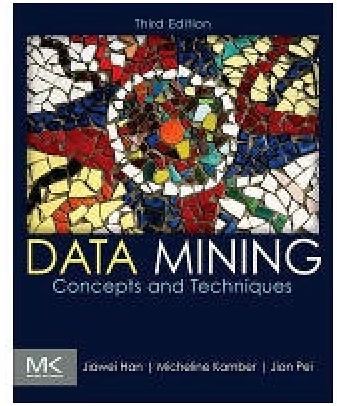
Why these topics?

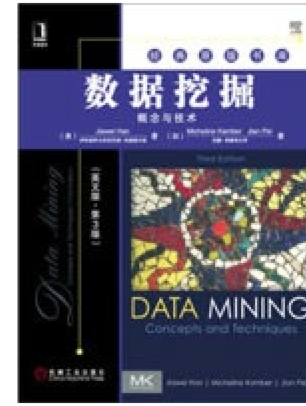
- Top-10 most influential papers in KDD 2022
  - 8. Graph-Flashback Network for Next Location Recommendation
  - 9. TwHIN: Embedding The Twitter Heterogeneous Information Network for Personalized Recommendation
  - 10. FederatedScope-GNN: Towards A Unified, Comprehensive and Efficient Package for Federated Graph Learning

https://www.paperdigest.org/2023/04/most-influential-kdd-papers-2023-04/

Textbooks & Recommended Readings

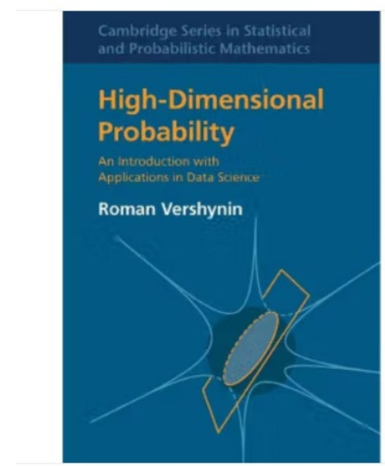
- Textbook:
  - J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Morgan kaufmann, 2011. 机械工业出版社(2012)





Advanced Data Mining Topics?

Machine learning?
Graph Mining?
Recommender systems?
Large language models?



Textbooks & Recommended Readings

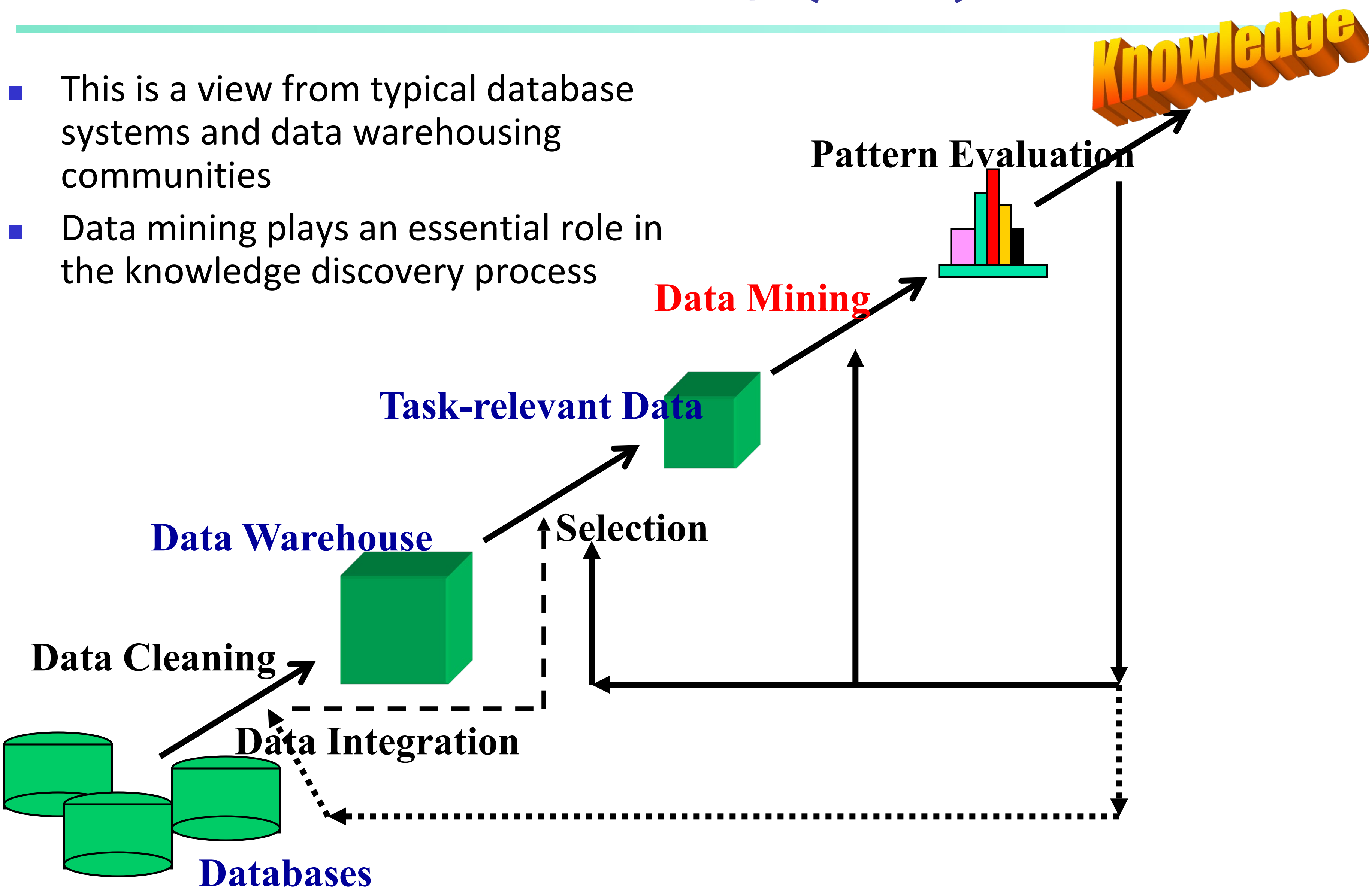
- **机器学习基础：**
  - 《统计学习方法》第二版 李航 部分章节（感知机学习对偶形式、kd树、隐马尔可夫模型、条件随机场、潜在语义分析、潜在狄利克雷分配、概率潜在语义分析不用看）
  - 《Understanding Machine Learning: From Theory to Algorithms》看1-6章
- **数学基础**
  - [《凸优化》](https://book.douban.com/subject/21249088/)部分章节（看第一章到第五章以及9.1、9.2、9.3、9.4、9.5章节；2.6、3.4、3.5、3.6、4.6不用看；其中第五章对偶是重点）
- **矩阵求导相关内容，** 参考[知乎] (https://zhuanlan.zhihu.com/p/24709748)
- **深度学习**
  - [《神经网络与深度学习》](https://book.douban.com/subject/35044046/)
  - 李宏毅的视频
- Pytorch基础
  - Pytorch官网的tutorial

Basic Data Mining

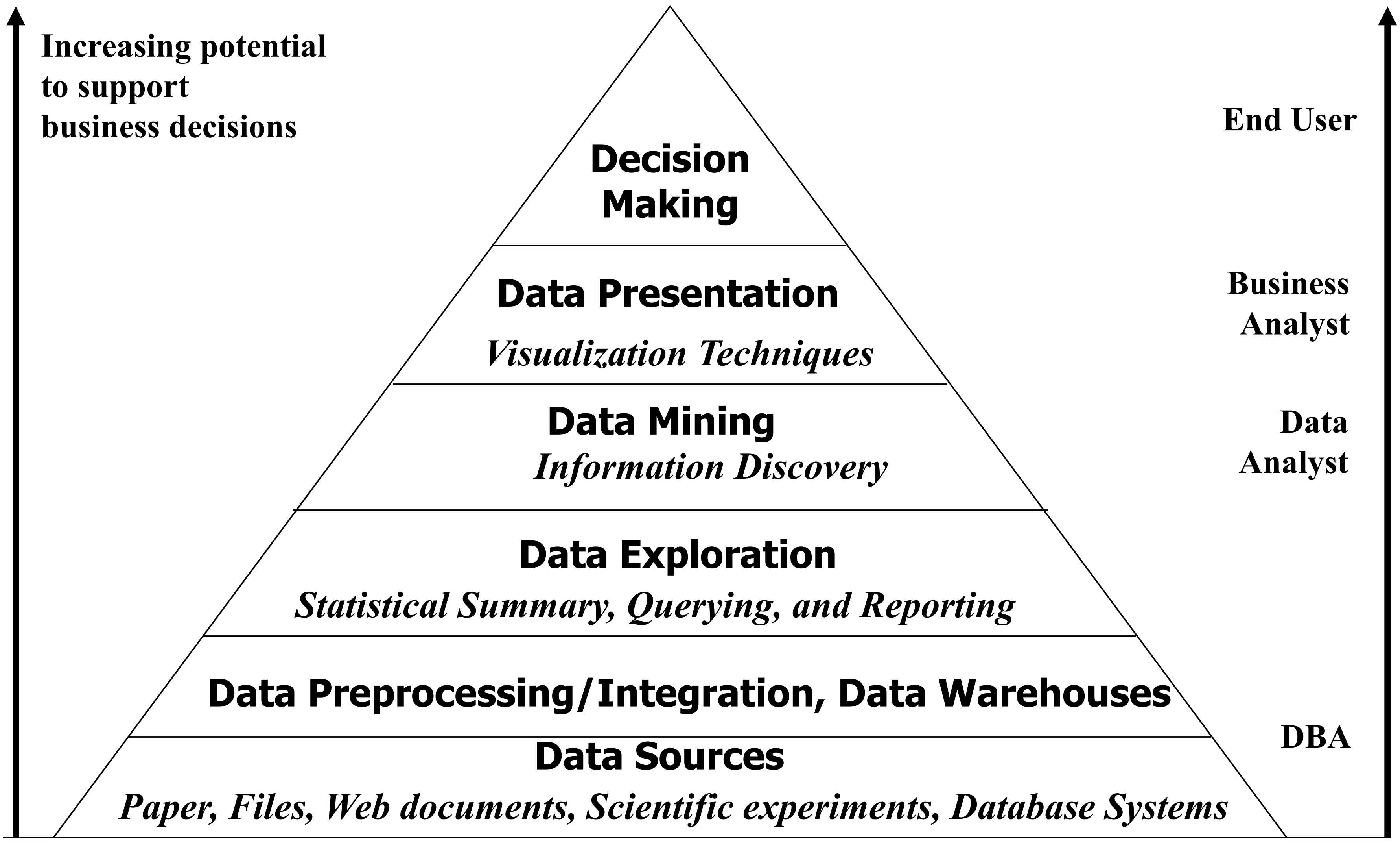
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

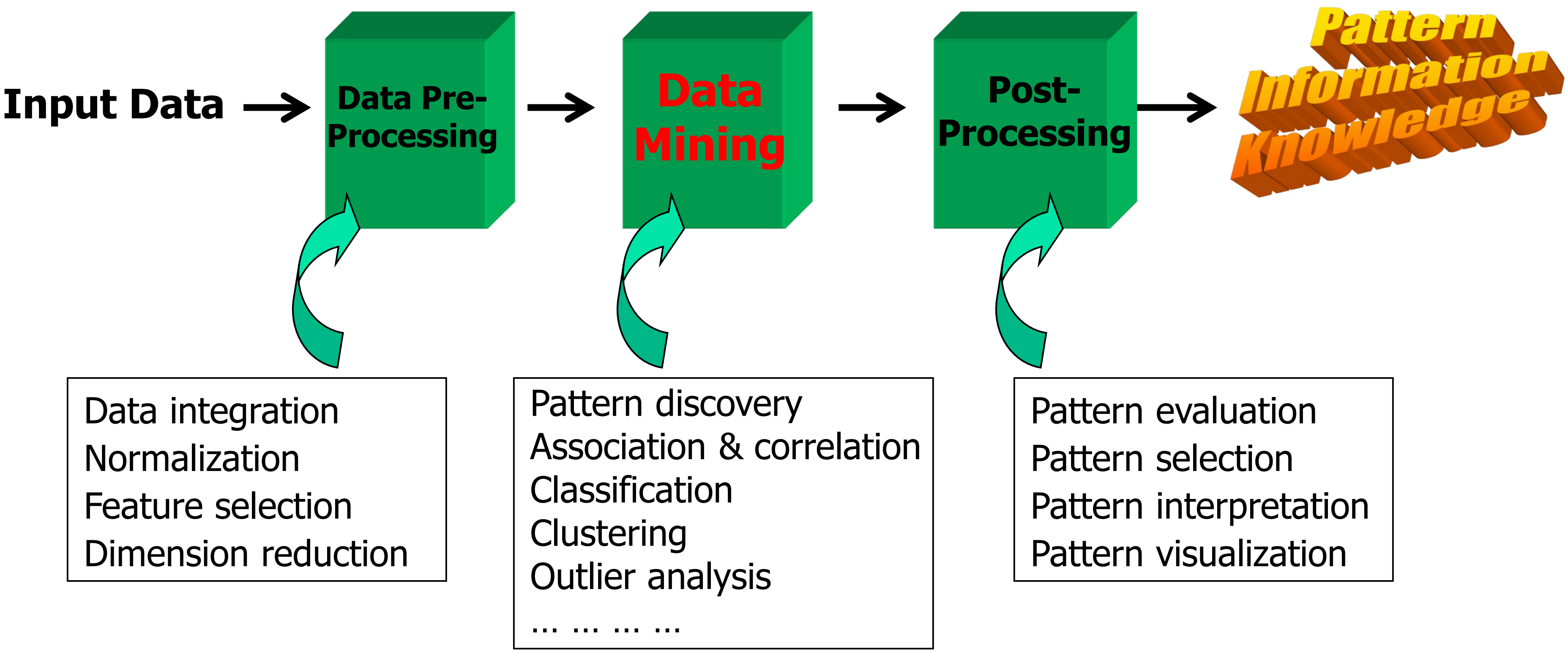
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence

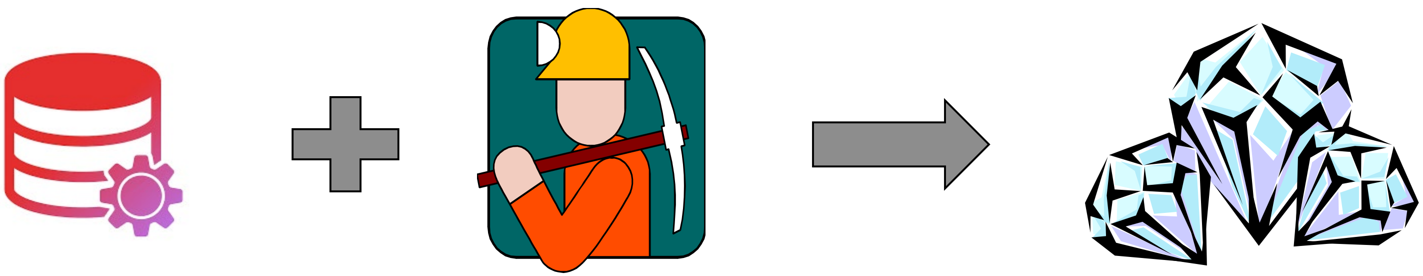


KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

Data Mining: Conclusions



Data **Mining** **Knowledge**

Data:

Data streams and sensor data

Time-series data, temporal data, sequence data (incl. bio-sequences)

Structure data, graphs, social networks and multi-linked data

Object-relational databases

Heterogeneous databases and legacy databases

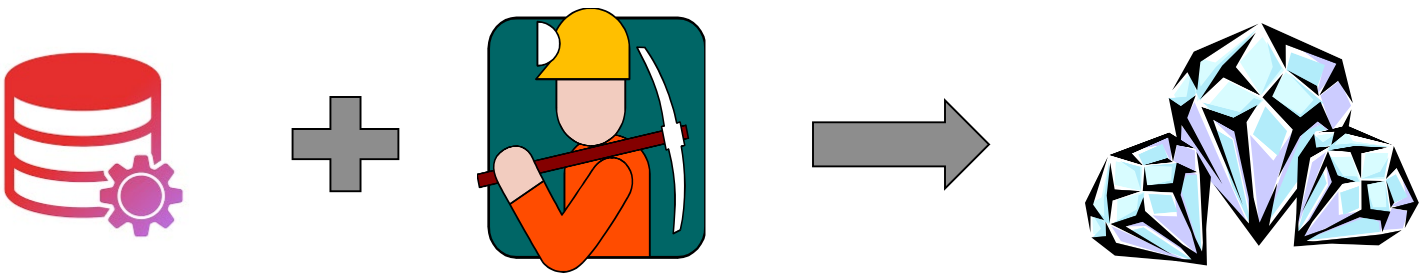
Spatial data and spatiotemporal data

Multimedia database

Text databases

The World-Wide Web

Data Mining: Conclusions



Data **Mining** **Knowledge**

Knowledge:

Models

Patterns

Causal relations

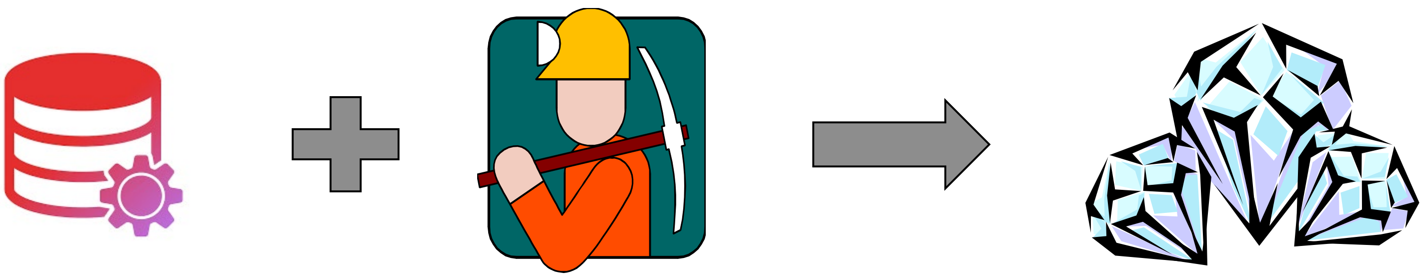
Rules

Decisions

Agents

……

Data Mining: Conclusions



Data **Mining** **Knowledge**

Mining Techniques:

Machine learning!
Statistics
Visualization
Database Technology
Algorithm
…..

Data Mining is indeed a BIG area!
Why KDD is not so hot as Neurips or ICML?

Basic Machine Learning

- Machine learning: basic concepts
- Types of learning
- Foundation of machine learning
- Promising directions

What is machine Learning?

- "Learning is any process by which a system improves performance from experience" - Herbert Simon

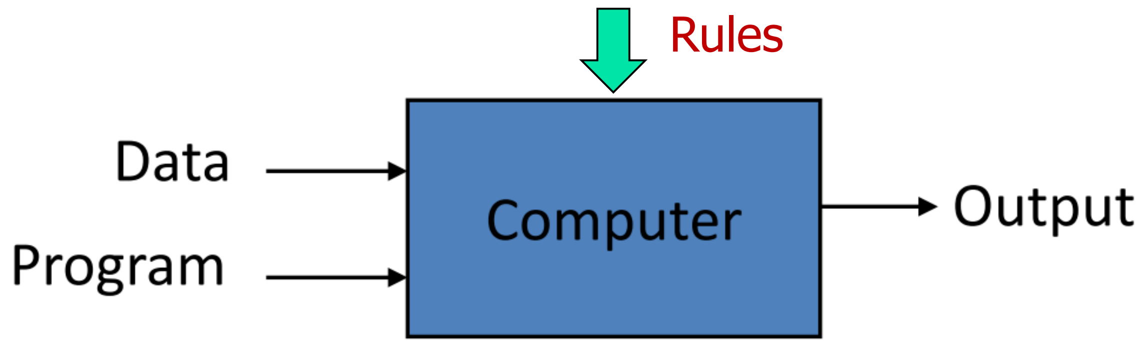
- Definition by Tom Mitchell (1998)

  Machine Learning is the study of algorithms that
  - improve their performance P
  - at some task T
  - with experience E

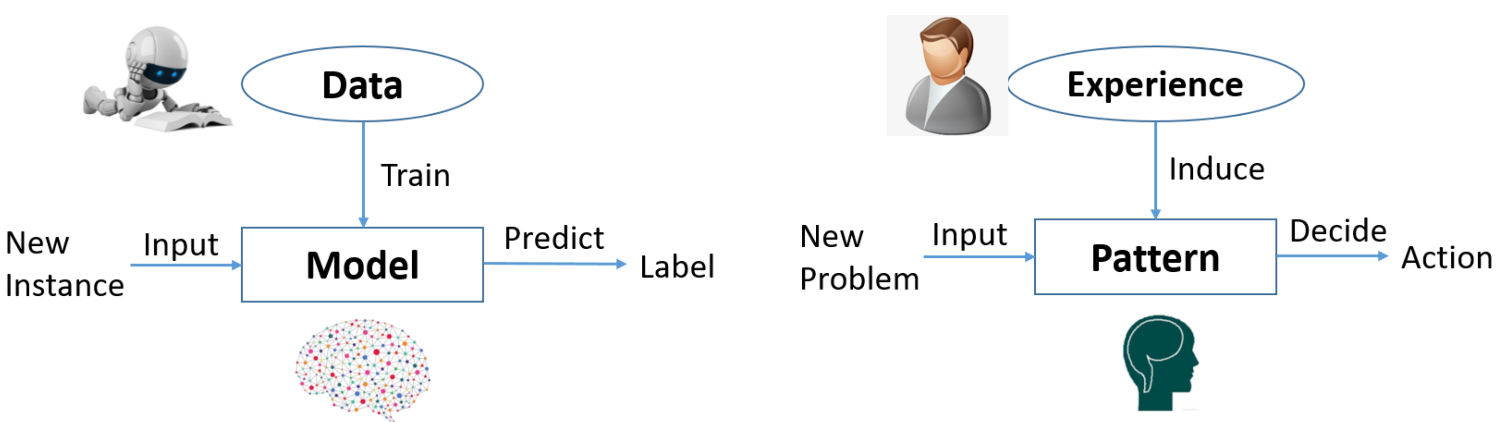
Slide credit: Eric Eaton

Why machine Learning?

- Traditional Programming

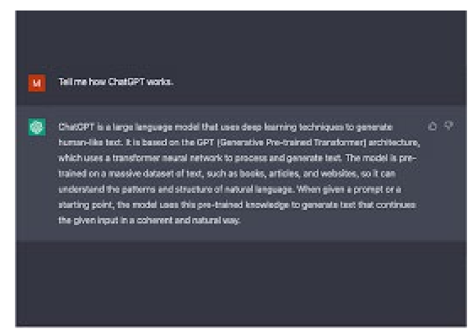


- Machine Learning
  - Intelligence: does not need human explicit programming
  - Universality: can handle complex data and diverse tasks



When Do We Use Machine Learning?

- ML is used when:
  - Human expertise does not exist (navigating on Mars)
  - Humans can’t explain their expertise (speech recognition)
  - Models must be customized (personalized recommendation)
  - Models are based on huge amounts of data (conversation)



Slide credit: Eric Eaton

Basic Machine Learning

- Machine learning: basic concepts
- Types of learning
- Foundation of machine learning
- Promising directions

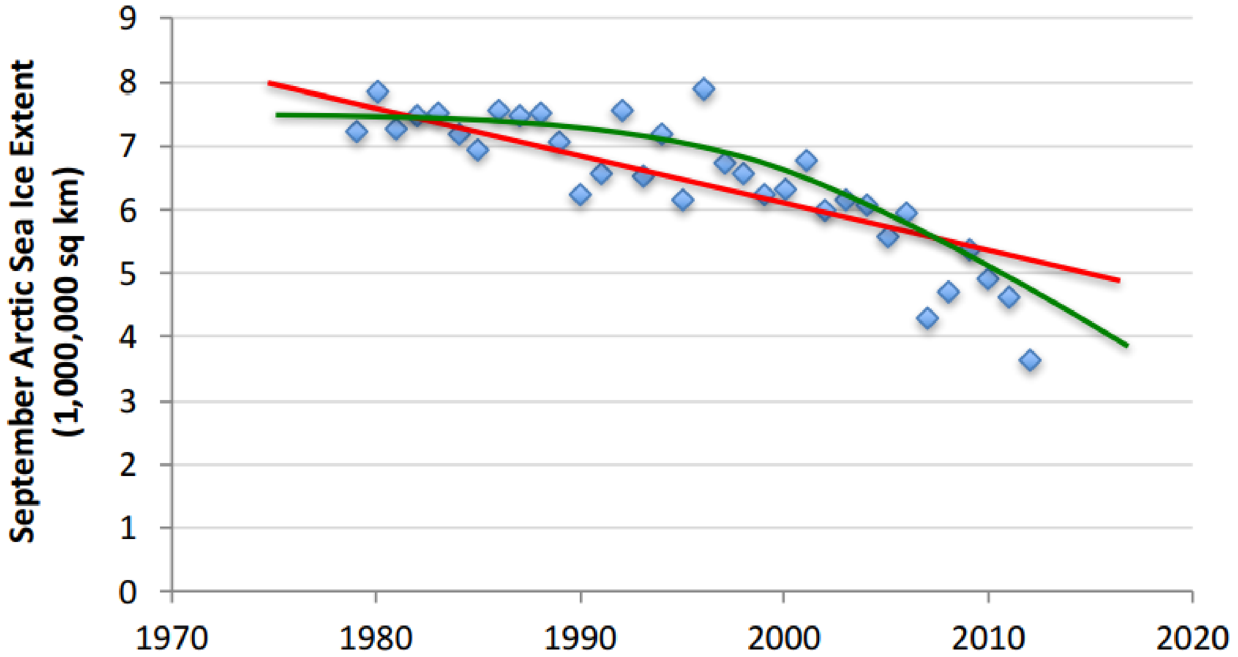
Basic Machine Learning

- Supervised learning
  - – Given: training data + desired outputs (labels)
- Unsupervised learning
  - – Given: training data (without desired outputs)
- Reinforcement learning
  - – Rewards from sequence of actions

Slide credit: Eric Eaton

Supervised learning: Regression

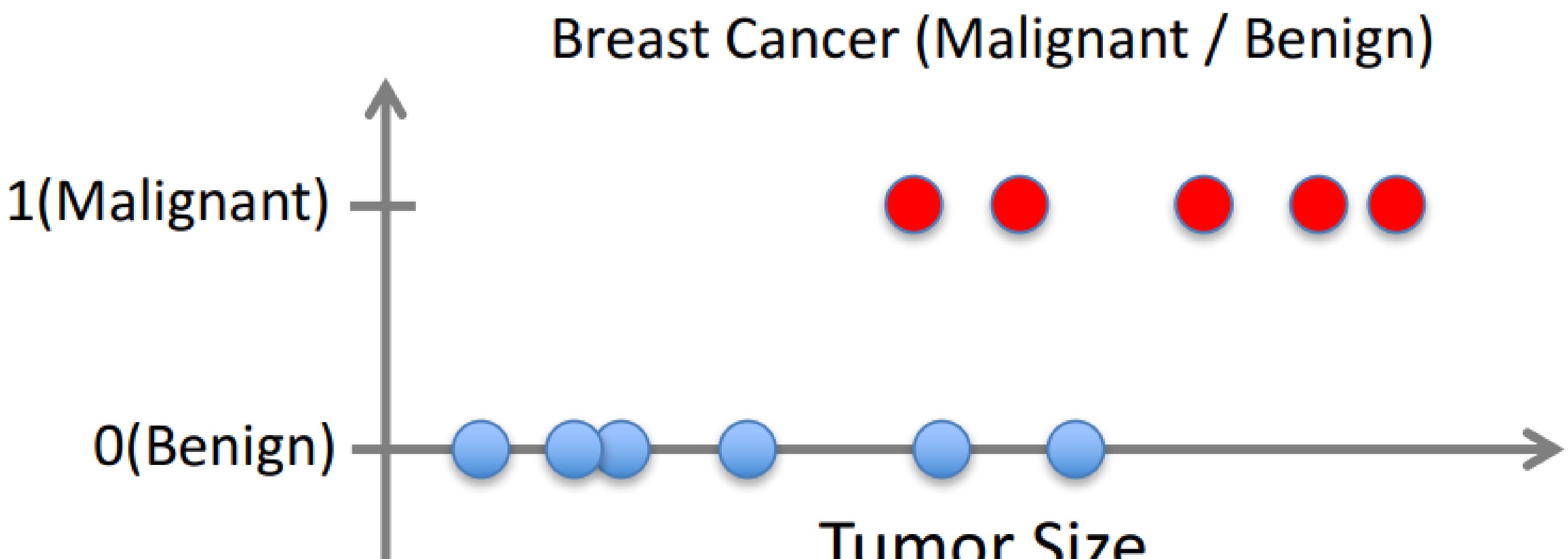
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued



Slide credit: Eric Eaton

Supervised learning: Classification

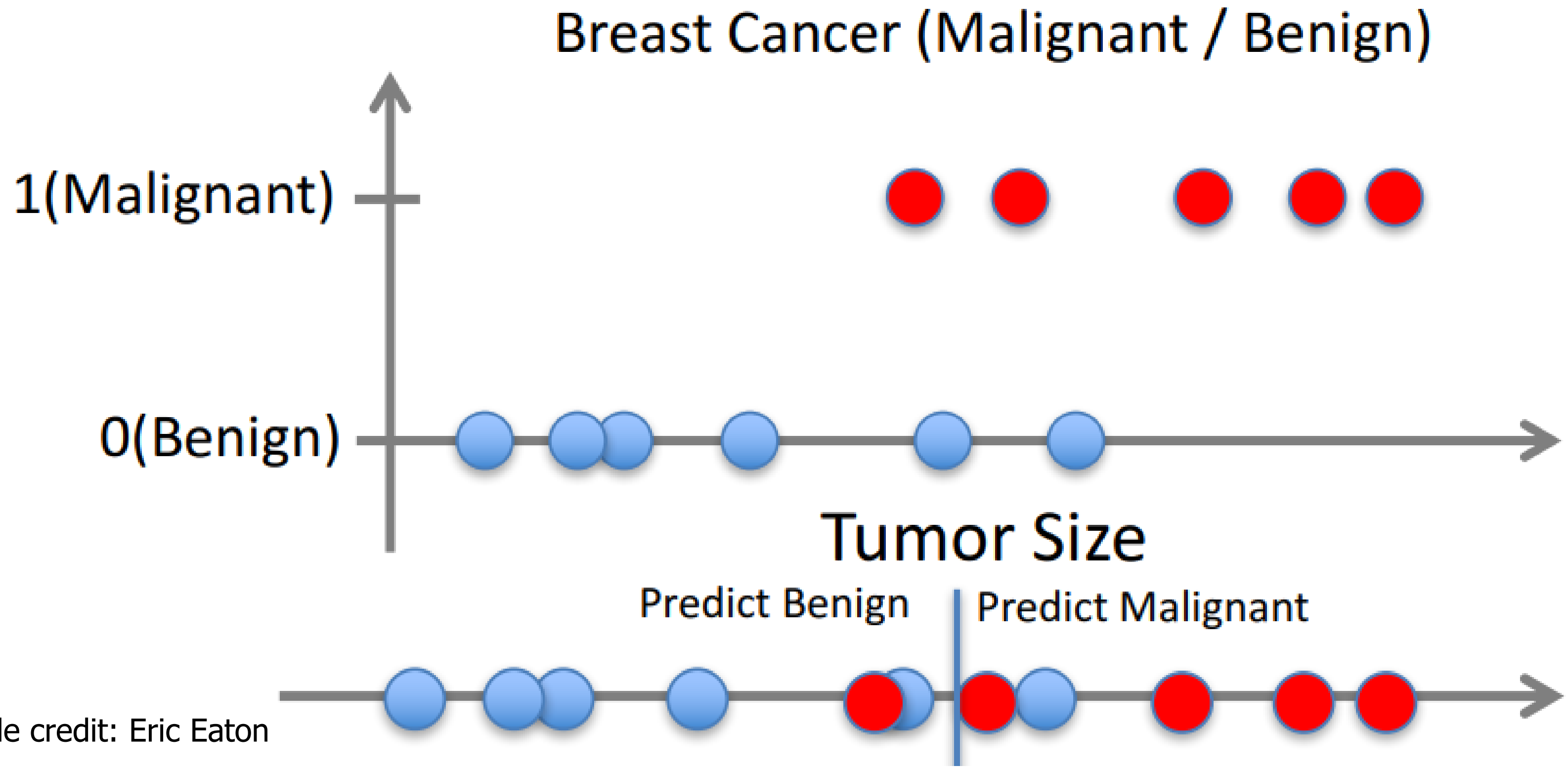
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical



Slide credit: Eric Eaton

Supervised learning: Classification

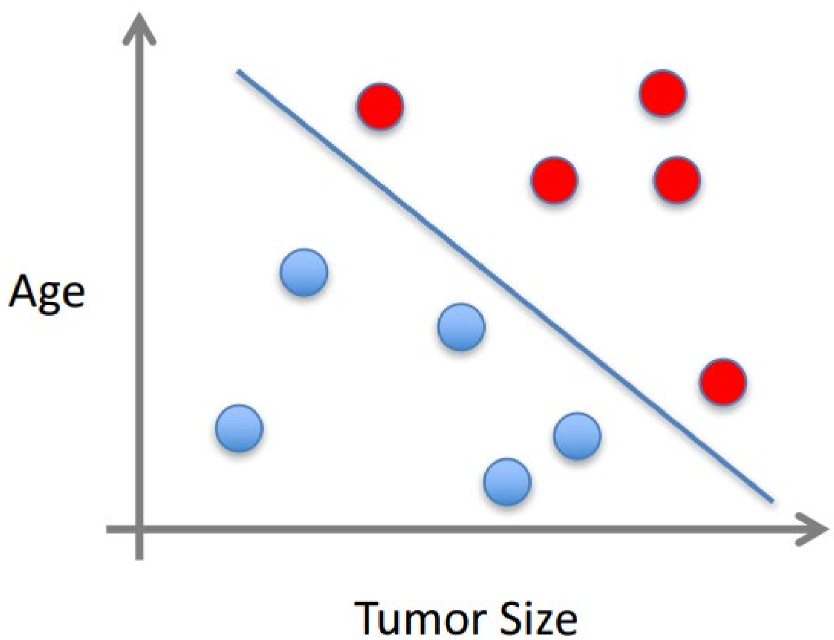
- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical



Slide credit: Eric Eaton

Supervised learning

- x can be multi-dimensional
  - Each dimension corresponds to an attribute

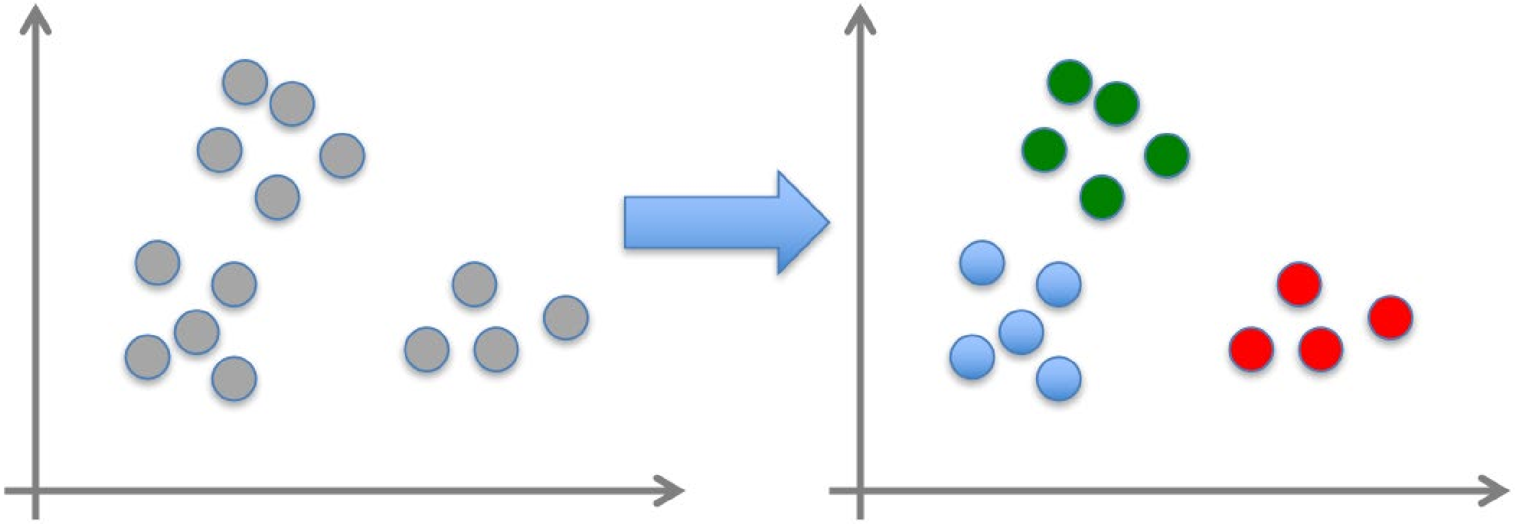


- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

Slide credit: Eric Eaton

Unsupervised learning

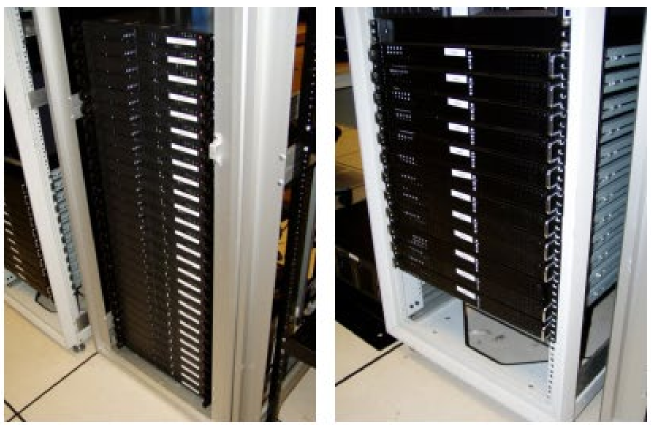
- Given $x_1, x_2, \ldots, x_n$(without labels)
- Output hidden structure behind the x's
  - Clustering
  - Learning Representation



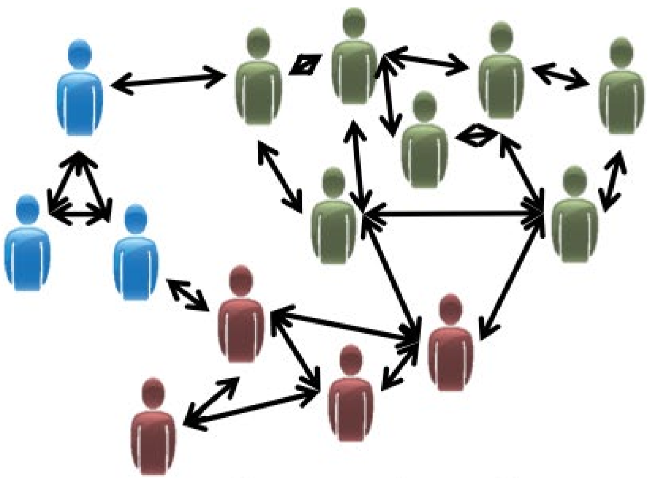
Slide credit: Eric Eaton

Unsupervised learning: Clustering

- Clustering: some examples



Organize computing clusters



Social network analysis

Slide credit: Eric Eaton

Unsupervised learning: Representation

- Learning representation from the data

---Self-supervised learning:

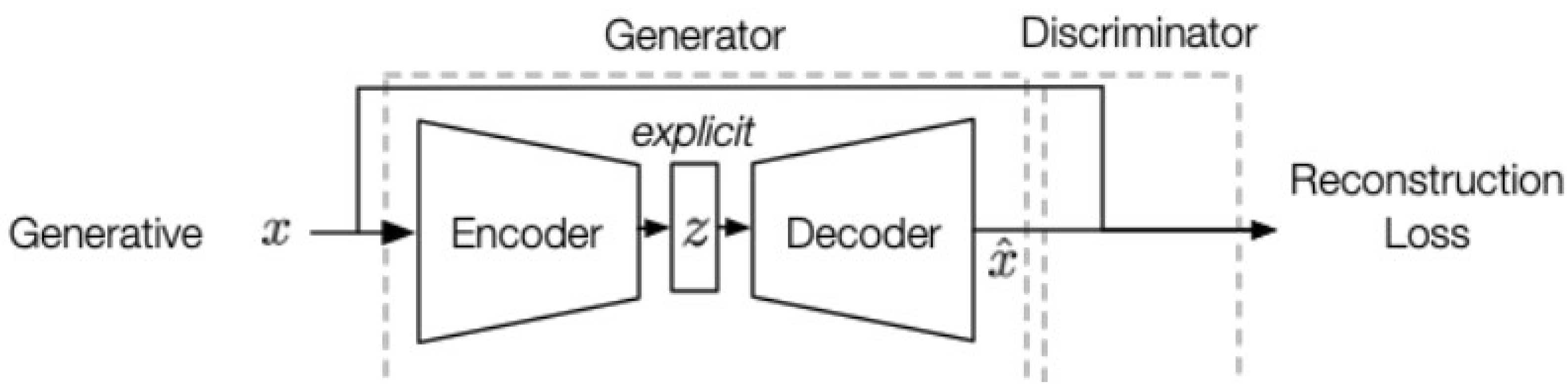
  - Obtain "labels" from the data itself by using a "semiautomatic" process.
  - Predict part of the data from other parts.

- "Construct a supervised task from the data"

TKDE'21: Self-supervised Learning: Generative or Contrastive
NIPS'23: Understanding contrastive learning via distributionally robust optimization

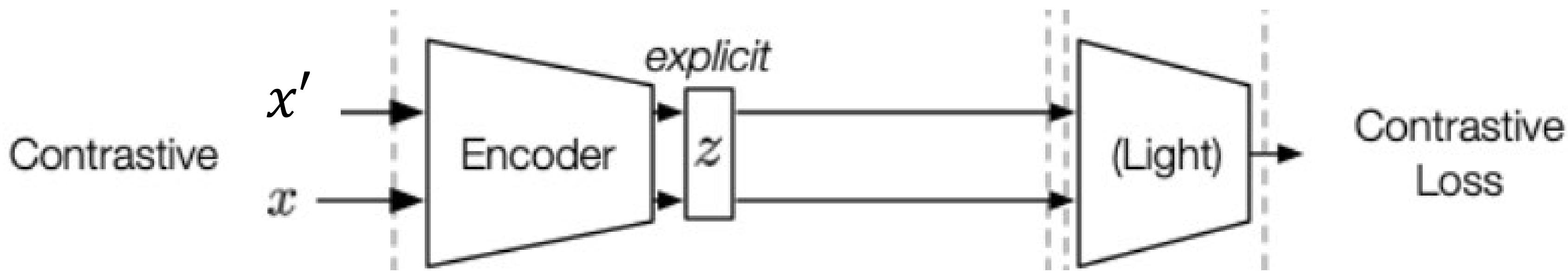
Unsupervised learning: Representation

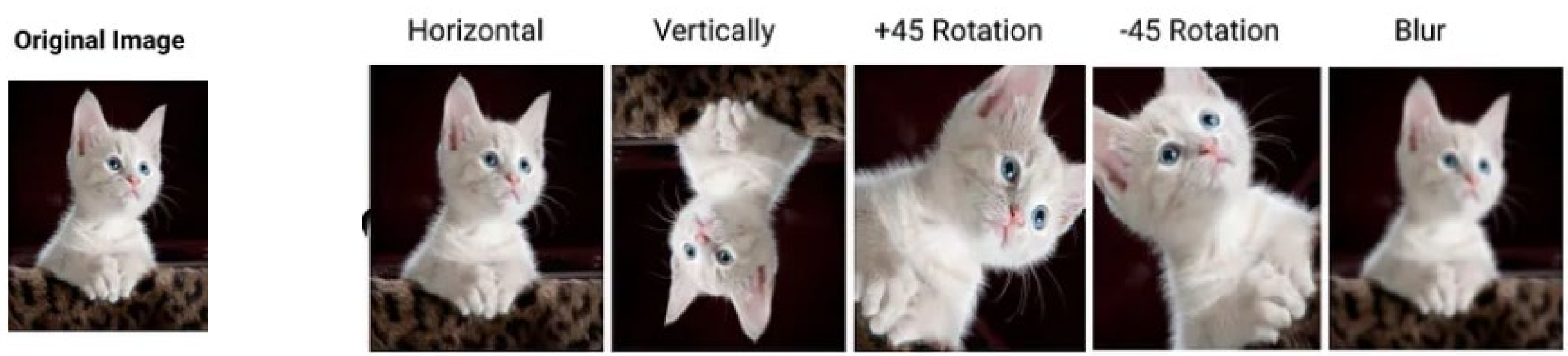
- Generative models:
  - train an encoder to encode input x into an explicit vector z and a decoder to reconstruct x from z
  - such as: Auto-Encoder, VAE, Diffusion Model



Unsupervised learning: Representation

- Constrastive models:
  - train an encoder to encode input x and x' into an explicit vector z to measure similarity
  - x' from data augmentation, e.g., Rotation, Cropping, Noise addition, Scaling

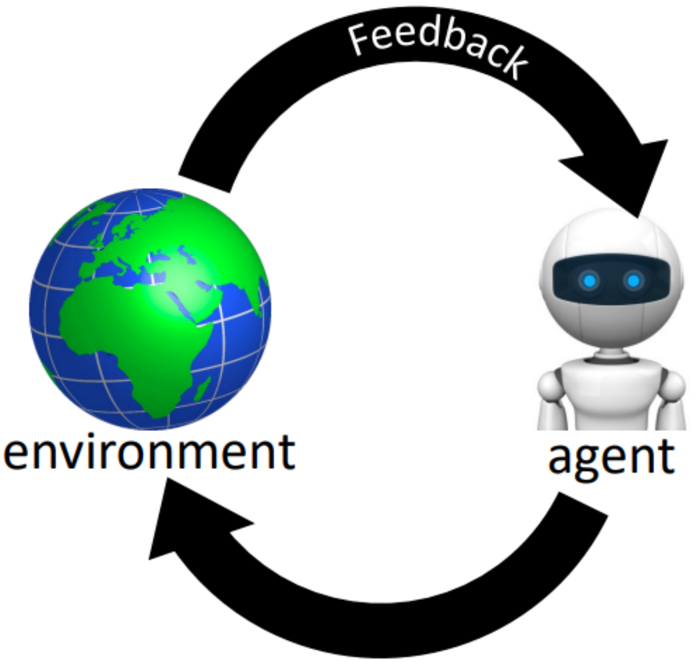




Reinforcement learning

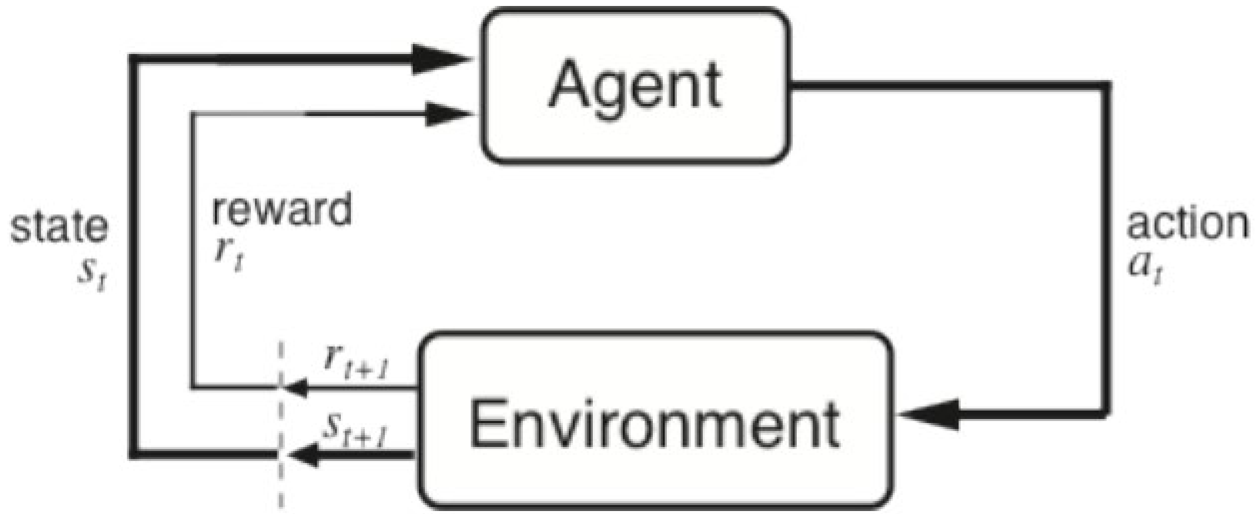
- learn how to perform a task from interactions with the environment

- Examples:
  - Playing chess (interact with the game)
  - Robot grasping an object (interact with the object/real world)
  - Recommender systems (system interacts with users)
  - Reinforcement learning from human feedback



Slide from CIS 4190

Reinforcement learning



- Agent and environment interact at discrete time steps
  - Agent observes state at step t, s_t
  - Produces action, a_t
  - Gets resulting reward, r_t

Basic Machine Learning

- Machine learning: basic concepts
- Types of learning
- Foundation of machine learning
- Promising directions

Foundation of machine learning

- The Statistical Learning Framework
  - Domain set $\mathcal{X}$, the set of objects that we may wish to label, e.g., a set of points represented by a vector of features
  - Label set $\mathcal{Y}$, $\mathcal{Y}$ can be $\{0,1\}$ for two-side classification or $\mathcal{R}$ for regression
  - Training data: S $= \left((x_1, y_1) \ldots (x_m, y_m)\right)$ is a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$ sampled from the training distribution D
  - *Learner:* learning a prediction function $h: \mathcal{X} \to \mathcal{Y}, h \in \mathcal{H}$

Foundation of machine learning

- The Statistical Learning Framework
  - *Loss function: measures the error between the prediction and the label* $l: \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0,1]$
    - $l(h, x, y) = \mathbf{I}[h(x) \neq y]$ for classification
    - $l(h, x, y) = (h(x) - y)^2$ for regression
  - True risk: how likely the learned h to make an error when labeled data are randomly drawn **according** to D
    - $L_D(h) = \underset{(x,y)\sim D}{\mathbf{E}}[l(h, x, y)]$
  - Objective of ML: find a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes the true risk $L_D(h)$

Foundation of machine learning

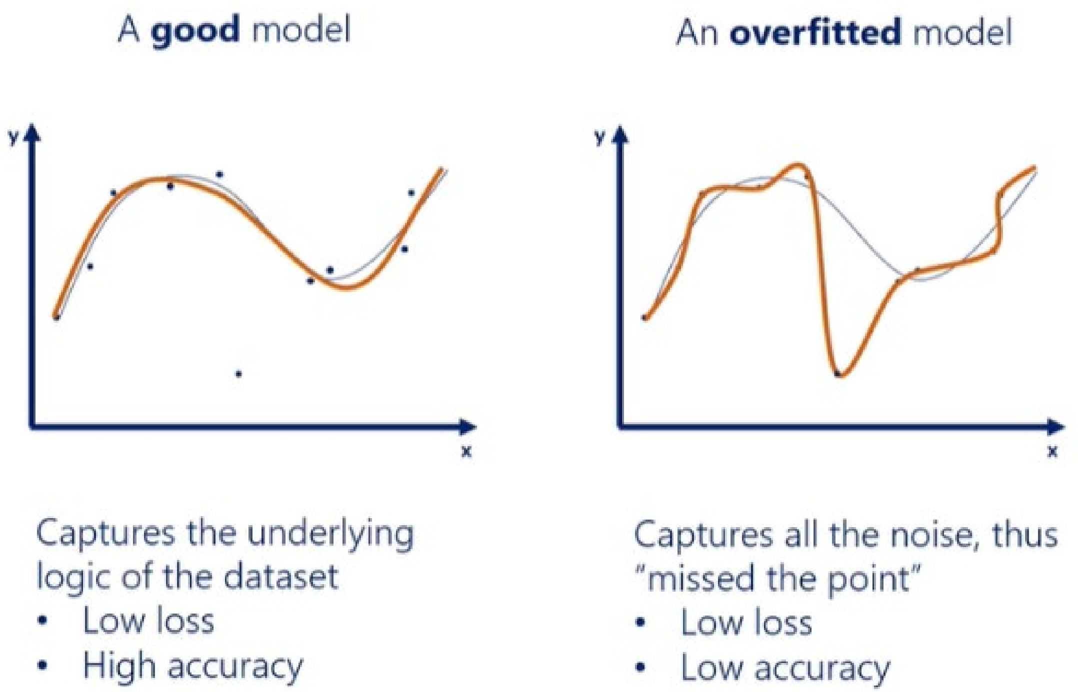
- Empirical risk:
  - Since data distribution D is not available, the model is learned on training data with optimizing:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, x_i, y_i)$$

  - Note that $L_S(h) \neq L_D(h)$
  - $L_S(h) = 0$ does not suggests good performance.

Foundation of machine learning

- Phenomenon of overfitting:



- Controlling the space of $\mathcal{H}$
- A theory to build the relation between $L_S(h)$ and $L_D(h)$

Foundation of machine learning

- PAC learning theory:

Theorem 1. For any finite hypothesis space of $\mathcal{H}$, given a training set S sampled i.i.d from D, then for any learned function $\tilde{h} = \text{argmin}_{h\in\mathcal{H}} L_S(h)$, with probability $1-\delta$, satisfies:

$$L_D(\tilde{h}) \le L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$$

Where h^* denotes the optimal function in $\mathcal{H}$ with $h^* = \text{argmin}_{h\in\mathcal{H}} L_D(h)$

Foundation of machine learning

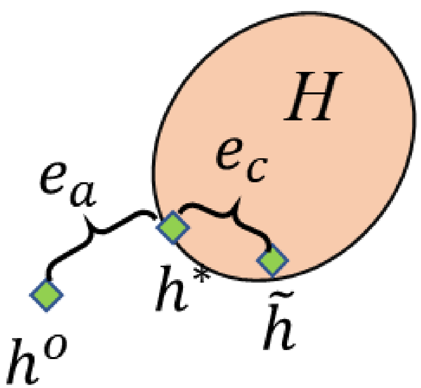
$$L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$$

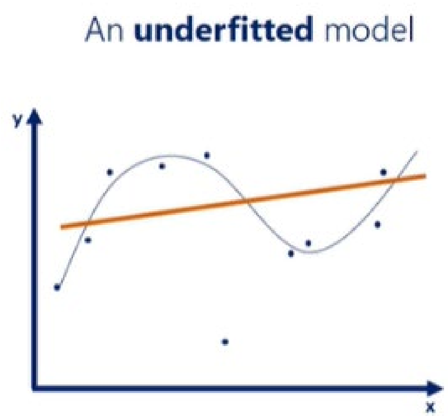
- The relations between $L_D(\tilde{h})$ and $L_D(h^*)$. Optimizing $L_S(h)$ would be a reasonable strategy.
- The difference ϵ is bounded, *w.r.t.* m, $|\mathcal{H}|$.
- $m \uparrow,\ \epsilon \downarrow.\quad m \to \infty, \epsilon \to 0.$
- $|\mathcal{H}| \uparrow,\ \epsilon \uparrow.$
- *Overfitting: large* $|\mathcal{H}|$*, small* m*, large* ϵ*.*

Foundation of machine learning

$$L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$$

- *Bias-complexity decomposition*
  - $e_a = L_D(h^*)$, Approximation Error or inductive bias.
  - $e_c = L_D(\tilde{h}) - L_D(h^*)$, Estimation Error.
- When $|\mathcal{H}|$ is small

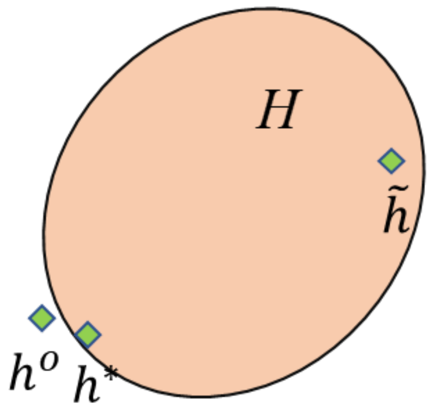


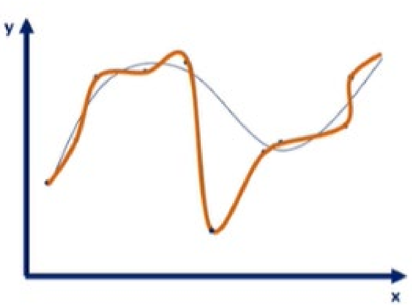


Foundation of machine learning

$$L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$$

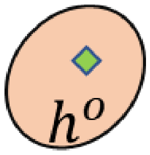
- *Bias-complexity decomposition*
  - $e_a = L_D(h^*)$, Approximation Error or inductive bias.
  - $e_c = L_D(\tilde{h}) - L_D(h^*)$, Estimation Error.
- When $|\mathcal{H}|$ is large





Foundation of machine learning

- *$\mathcal{H}$ is important ---> design better ML model according to the task, e.g., better neural architecture, using valuable knowledge, etc.*



- *Data is important ---> enrich the dataset, e.g., data augmentation.*

- *Optimizer is important ---> design a better and suitable optimizer.*

$$\tilde{h} = \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$$

Foundation of machine learning

- *Proof of Theorem 1.* $L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$

  - Part 1: establish the relations between $L_D(h)$ and $L_S(h)$.

LEMMA 4.5 (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

$$L_D(h) = \mathop{\mathbf{E}}_{(x,y)\sim D}[l(h,x,y)] \qquad L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, x_i, y_i)$$

We have: $P(|L_S(h) - L_D(h)| > \epsilon) \leq 2\exp(-2m\epsilon^2)$

Foundation of machine learning

- *Proof of Theorem 1.* $L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$

  Based on $P(|L_S(h) - L_D(h)| > \epsilon) \leq 2\exp(-2m\epsilon^2)$

  We further have:

  $P(\{\exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}) \leq 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)$

  due to the fact: $P(A_1 \cup A_2 \ldots \cup A_n) \leq \sum_{i=1}^{n} P(A_i)$

  Thus, we have:

  $$P(\{\forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon\}) > 1 - \delta$$

  Where $\delta = 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)$

Foundation of machine learning

- *Proof of Theorem 1.* $L_D(\tilde{h}) \leq L_D(h^*) + \sqrt{\frac{2}{m}\log(\frac{2|\mathcal{H}|}{\delta})}$

  - Part 2: establish the relations between $L_D(\tilde{h})$ and $L_D(h^*)$.

Based on $|L_S(h) - L_D(h)| \leq \epsilon$

We have:

$$L_D\left(\tilde{h}\right) \leq L_S\left(\tilde{h}\right) + \epsilon$$
$$\leq L_S(h^*) + \epsilon$$
$$\leq L_D(h^*) + 2\epsilon$$

Based on $\delta = 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)$*, the theorem got proved.*

Basic Machine Learning

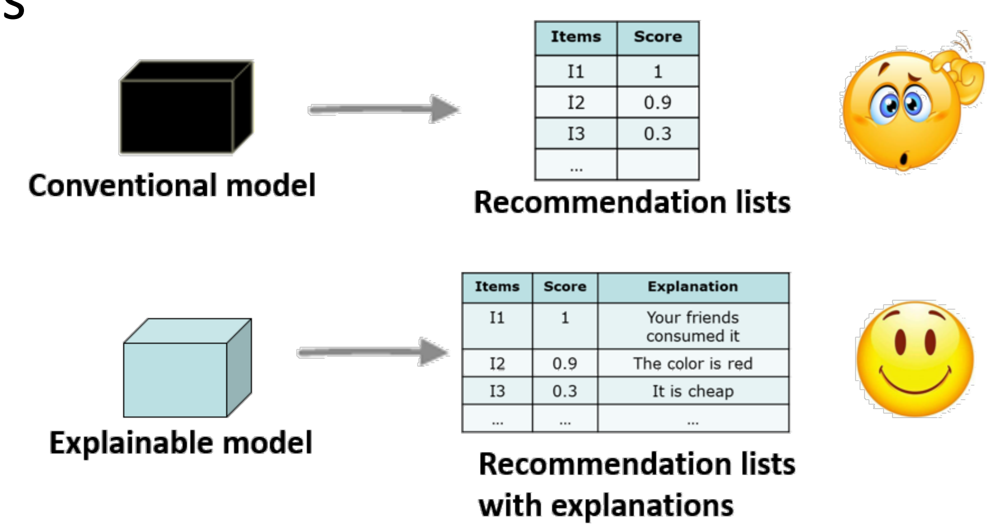
- Machine learning: basic concepts
- Types of learning
- Foundation of machine learning
- Promising directions

Promising directions

- Model robustness
  - Distribution shift --- IID不满足
  - Noisy data --- 样本不太对
  - Attacks --- IID不满足
  - Long-tailed data --- 学习器自身问题，很“懒”
  - Outliers --- 极端值炸裂
- Tailored for specific model and task

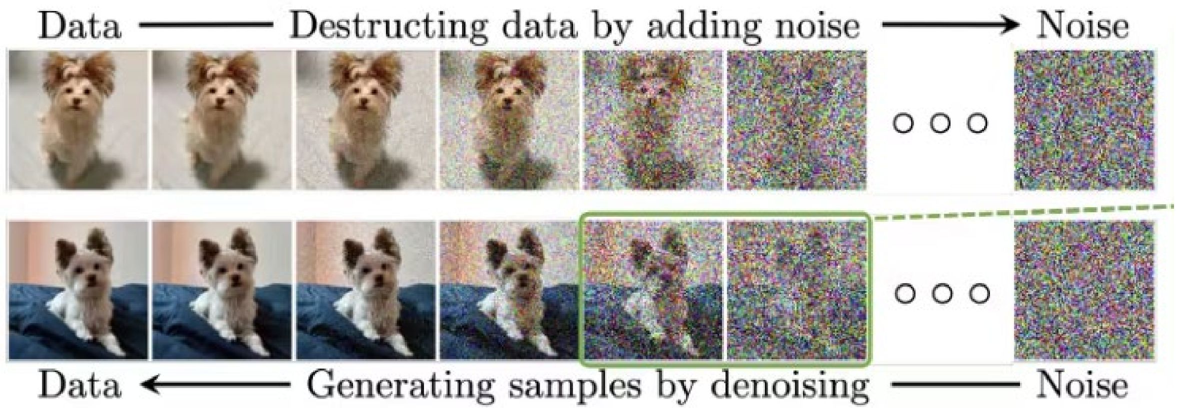
Promising directions

- Model interpretability and Explainability
  - Interpretability is the degree to which a human can understand the cause of a decision
    - Understanding mechanism of deep model or LLM
  - Explainability is how to take ML model explain the behavior in human terms

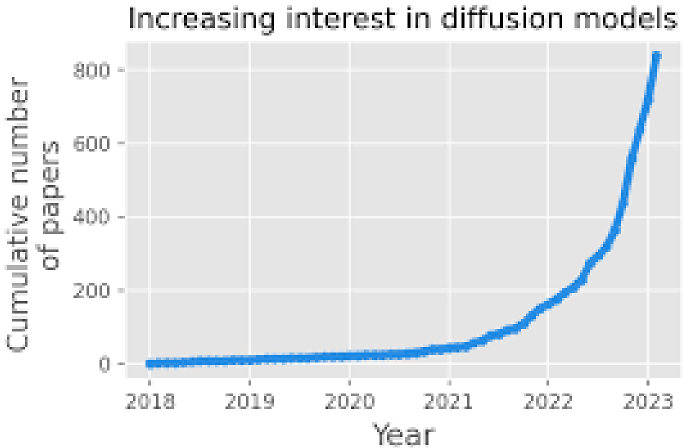


Promising directions

- Diffusion model



  - Diffusion models offer several advantages over traditional generative models
  - Becomes very hot topic
  - Many things can be done



Promising directions

- Large language model
  - Prompt
  - Fine-tuning
  - Empirical & Theoretical analyses
  - Application

list some types of data augmentations of an image

Data augmentation is a technique used to increase the diversity of a dataset by applying various transformations to the original images. This helps improve the performance of machine learning models, particularly in computer vision tasks. Here are some common types of image data augmentations:

1. Rotation: Rotating the image by a certain angle.
2. Flipping: Flipping the image horizontally or vertically.
3. Scaling: Resizing the image by a specific scale factor.
4. Cropping: Removing parts of the image to create a smaller region of interest.
5. Translation: Shifting the image along the x or y axis.

Promising directions

- Large language model
  - Prompt
  - Fine-tuning
  - Empirical & Theoretical analyses
  - Application

list some types of data augmentations of an image

Data augmentation is a technique used to increase the diversity of a dataset by applying various transformations to the original images. This helps improve the performance of machine learning models, particularly in computer vision tasks. Here are some common types of image data augmentations:

1. Rotation: Rotating the image by a certain angle.
2. Flipping: Flipping the image horizontally or vertically.
3. Scaling: Resizing the image by a specific scale factor.
4. Cropping: Removing parts of the image to create a smaller region of interest.
5. Translation: Shifting the image along the x or y axis.

Promising directions

- Large language model
  - Math reasoning
  - LLM+X, recsys, graph, etc.
  - Fine-tuning LLM, RLHF; DPO; LORA; etc.
  - Evaluation & Benchmark
  - Hallucination & Boundary of ability
    - 知道自己知道的； 知道自己不知道的；
    - 不知道自己知道的；不知道自己不知道的。
  - Retrieval-augmented LLMs

THANK YOU!