L1: Introduction

Shan Wang

Lingnan College, Sun Yat-sen University

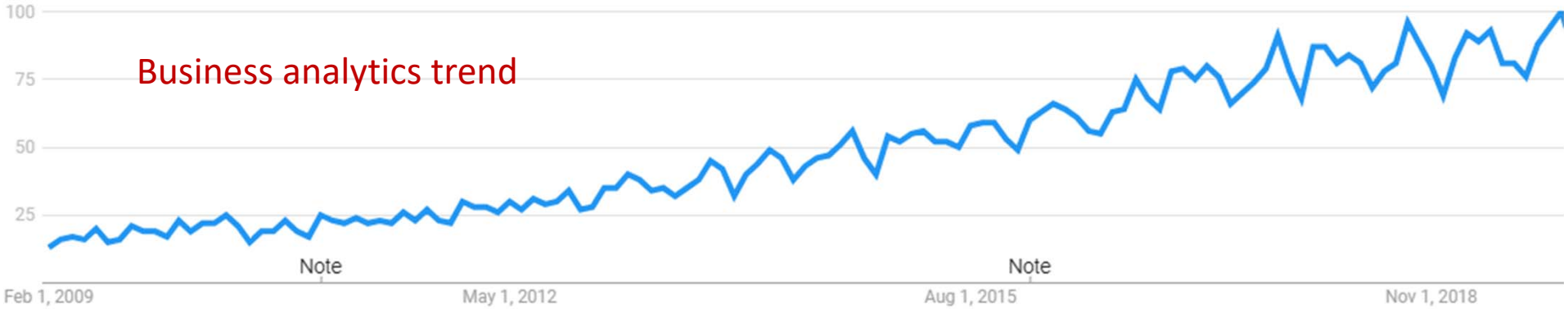
2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

Trends

Machine learning trend

Business analytics trend



https://trends.google.com

What is Business Analytics

- Turn data into information/value.
  - Business managers need to make decisions.
  - They need to make the most informed decisions that they can, and generate value.

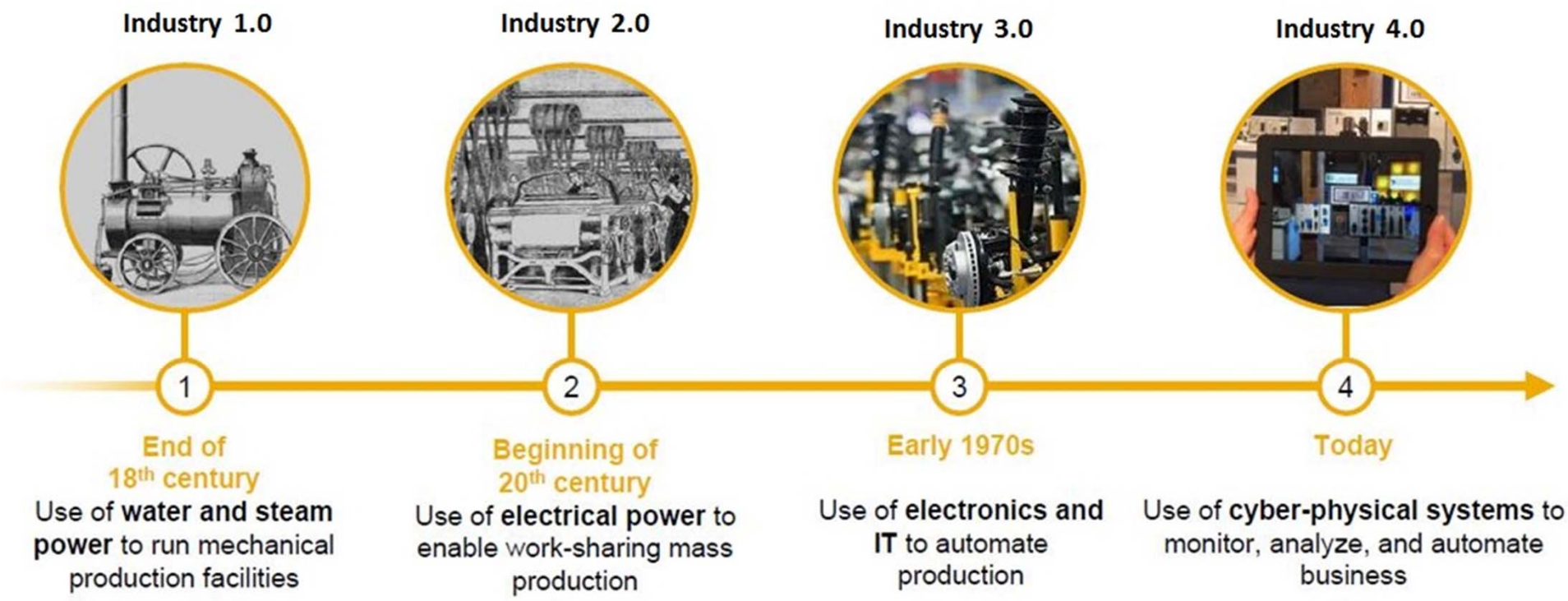
- Decision making under uncertainty.
  - Most of the decisions are based on guesses, rather than "facts."
  - How to make the "best" guess possible as well as how to measure the accuracy of their guesses.

We need machine learning!

Why machine learning?

Industry Revolution 4.0

Four Phases of Industrialization



Big Data

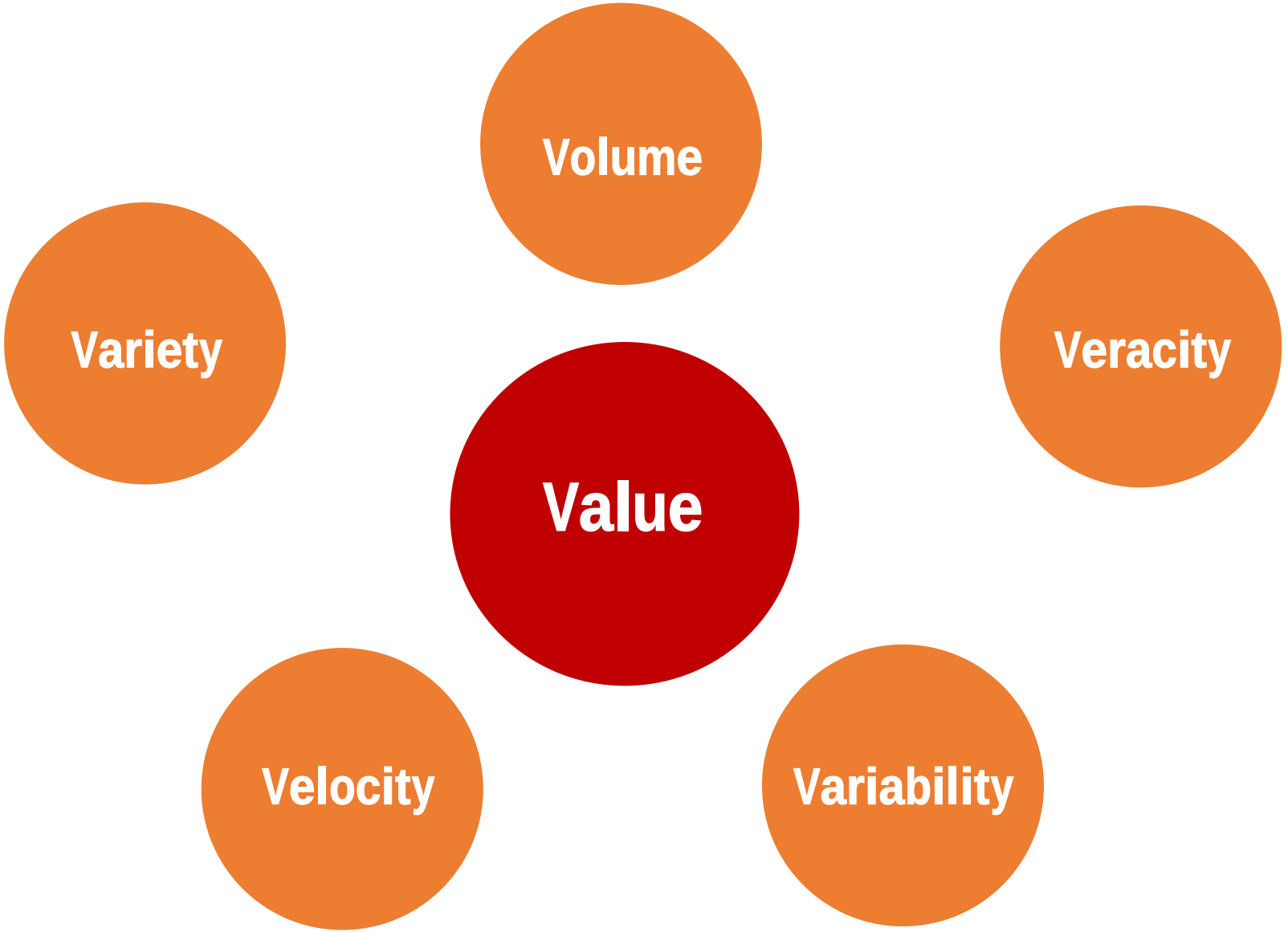

支付宝
youku
You Tube
fr


淘
阿里巴巴
Alibaba.com
Suning.com
苏宁易购
amazon
JD.COM 京东
GROUPON
去哪儿?
Qunar.Com
Expedia
HSBC


10:10 AM
10:37 PM
SAMSUNG
NOKIA
BlackBerry
ThinkPad


Apps
• Straps
• Shirts
• Wrist-worn
• Clips
• Shoe-worn / Foot pods

6 V's of Big Data

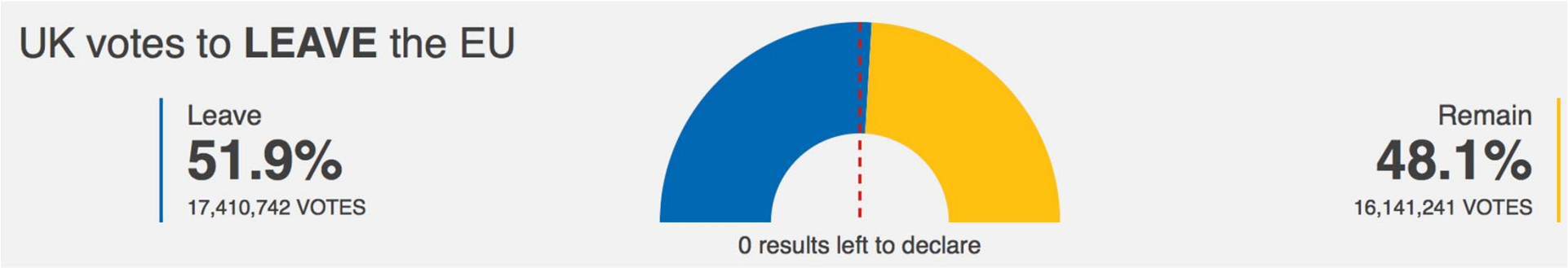


Business value of data

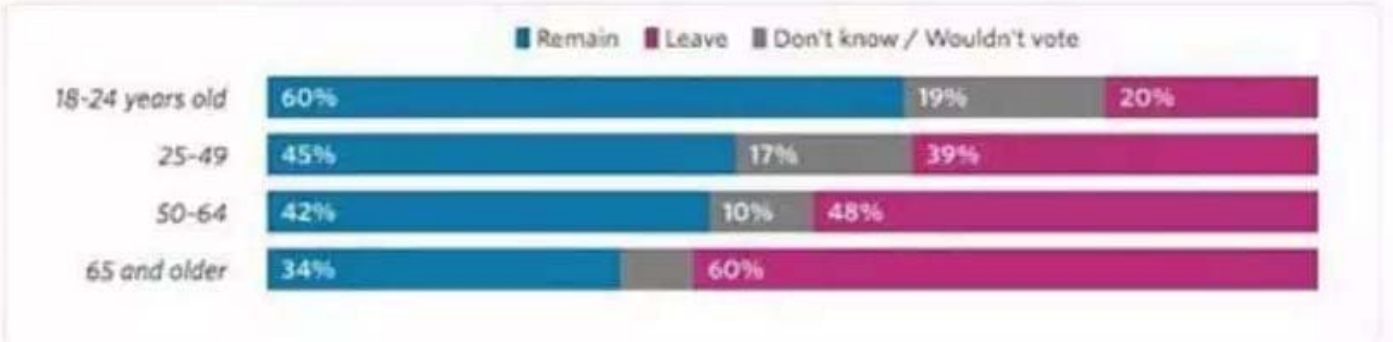
- Estimates suggest that by better integrating big data, **healthcare** could save as much as **$300 billion a year** — that's equal to reducing costs by $1000 a year for every man, woman, and child.
- For a typical Fortune 1000 company, just a 10% increase in data accessibility will result in more than **$65 million** additional net income.
- Retailers who leverage the full power of big data could increase their operating margins by as much as **60%**.
- **73%** of organizations have already invested or plan to invest in big data by 2016.
- **Less than 0.5% of all data is ever analysed and used!**

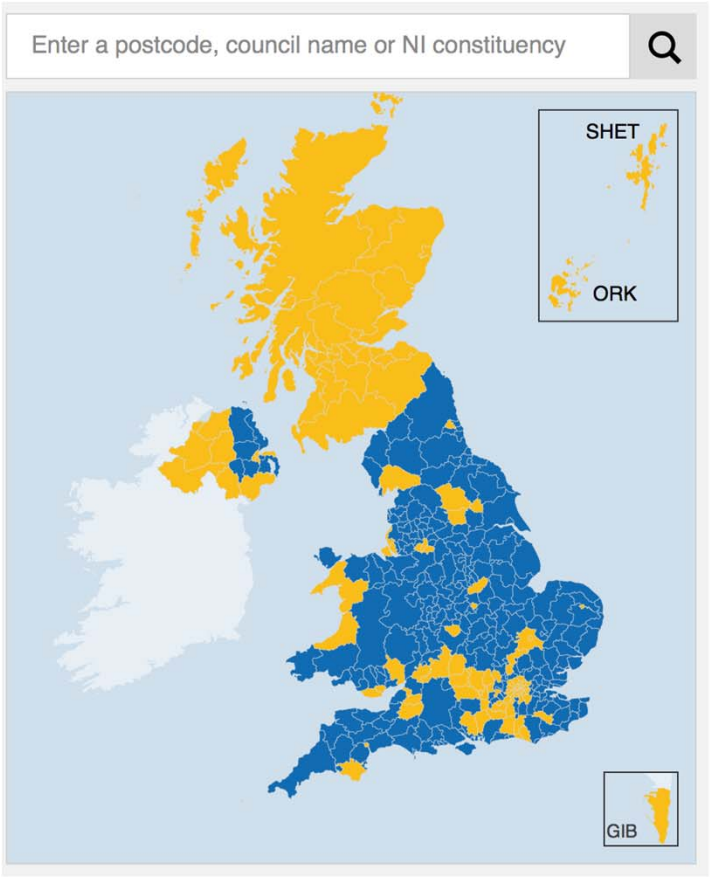
[Forbes on Sep 30, 2015]

Just Being Big May Not Get You There …



Why machine learning?





England

Leave **53.4%**
15,188,406 VOTES

Remain **46.6%**
13,266,996 VOTES

Counting complete

Turnout: 73.0%

Northern Ireland

Leave **44.2%**
349,442 VOTES

Remain **55.8%**
440,437 VOTES

Counting complete

Turnout: 62.9%

Scotland

Leave **38.0%**
1,018,322 VOTES

Remain **62.0%**
1,661,191 VOTES

Counting complete

Turnout: 67.2%

Wales

Leave **52.5%**
854,572 VOTES

Remain **47.5%**
772,347 VOTES

Why machine learning?

- Big Data by itself, regardless of the size, type, or speed, is worthless.
- Big Data + "big" analytics = value
- Big Data brought big challenges
  - Effectively and efficiently capturing, storing, and analyzing Big Data
  - New breed of technologies needed
  - Hadoop, MapReduce, Spark
  - Unstructured data: text mining, social media analysis
- We need machine learning!

What is machine learning?

A subset of artificial intelligence involved with the creation of algorithms which can modify itself without human intervention to produce desired output - by feeding itself through structured data

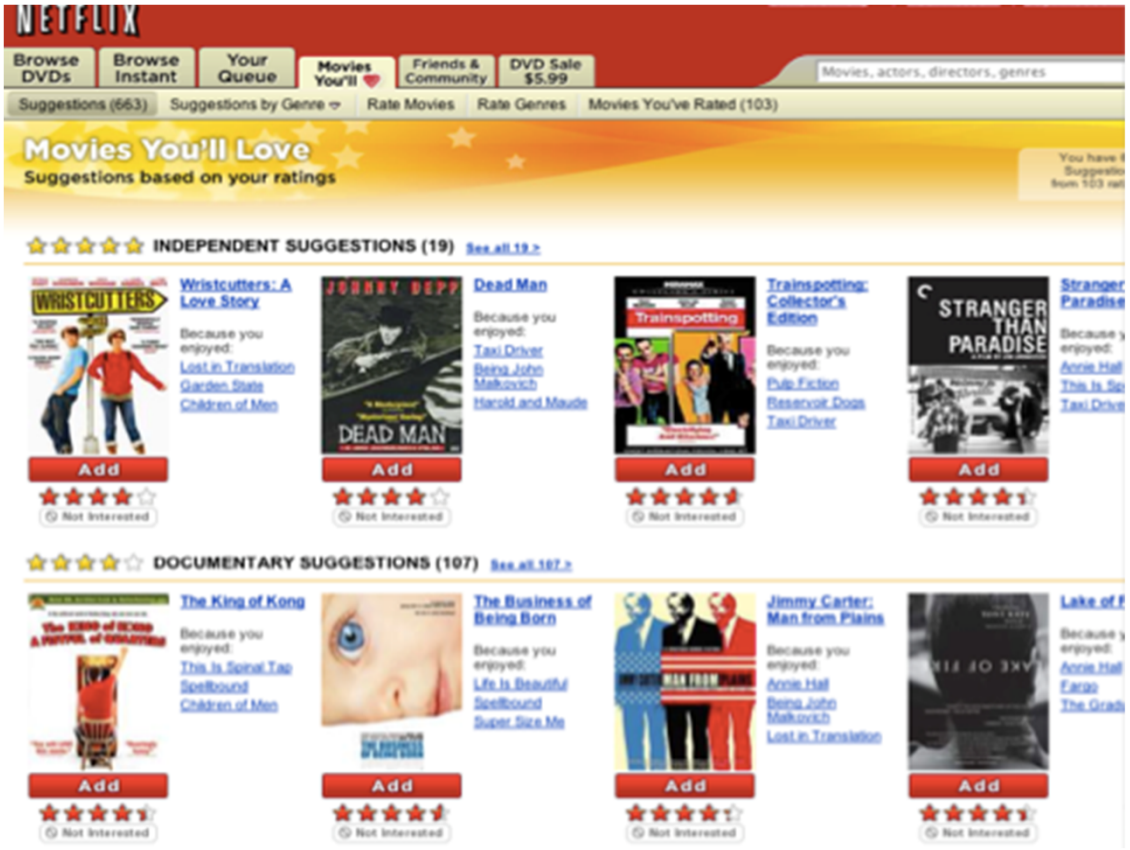
Applications of ML

Target Customer Predictive Analytics

http://www.youtube.com/watch?v=RC5HNTj3Dag&feature=related

Netflix Movie Recommendation

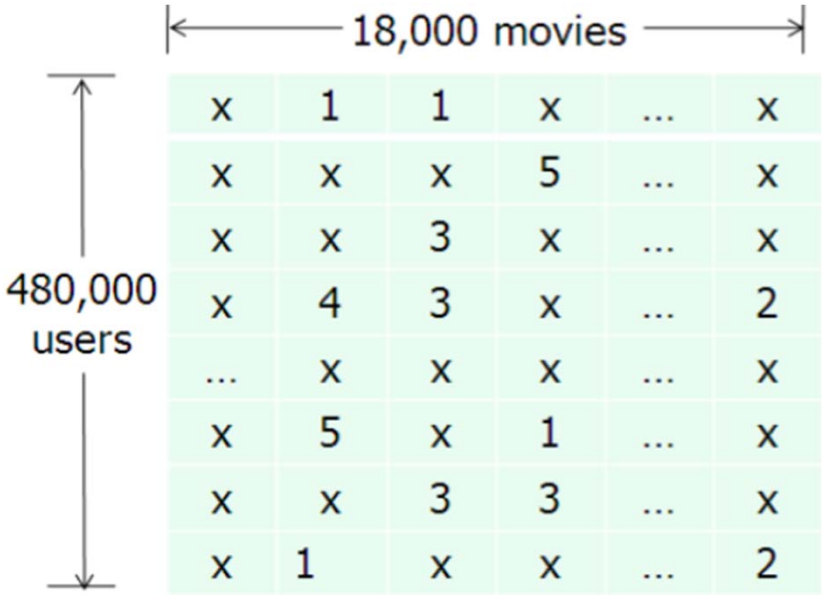
- A US-based DVD retail company (1997 -)
- As of April 2019, ~ 100 million subscribers



- Good recommendation = happy customer = business value (new product: House of Cards)

The Netflix Competition (2006-2009)

- Netflix offers $1M for an improved recommender algorithm
- Training data: 100 million ratings



- Test data
  - Last few ratings of each user (2.8 million)
- Winner BellKor's Pragmatic Theory, using a combination of > 800 models https://www.youtube.com/watch?v=ImpV70uLxyw
  - Two main classes: nearest neighbors and principal component analysis

Rue La La Flash Sales

Event

The 100: Men's Sweaters That Demand a Hot Toddy ▸

CLOSING IN 2 DAYS, 20:32:43

Lazy Sunday Uniform: Leggings, Sweaters, & More ▸

CLOSING IN 1 DAY, 20:32:43

Belle by Sigerson Morrison ▸

CLOSING IN 1 DAY, 20:32:43

Style

sofiacashmere Blue Merino Wool Crew Sweater

~~$150.00~~ **$54.90**

sofiacashmere Heather Grey Cashmere Polo Sweater

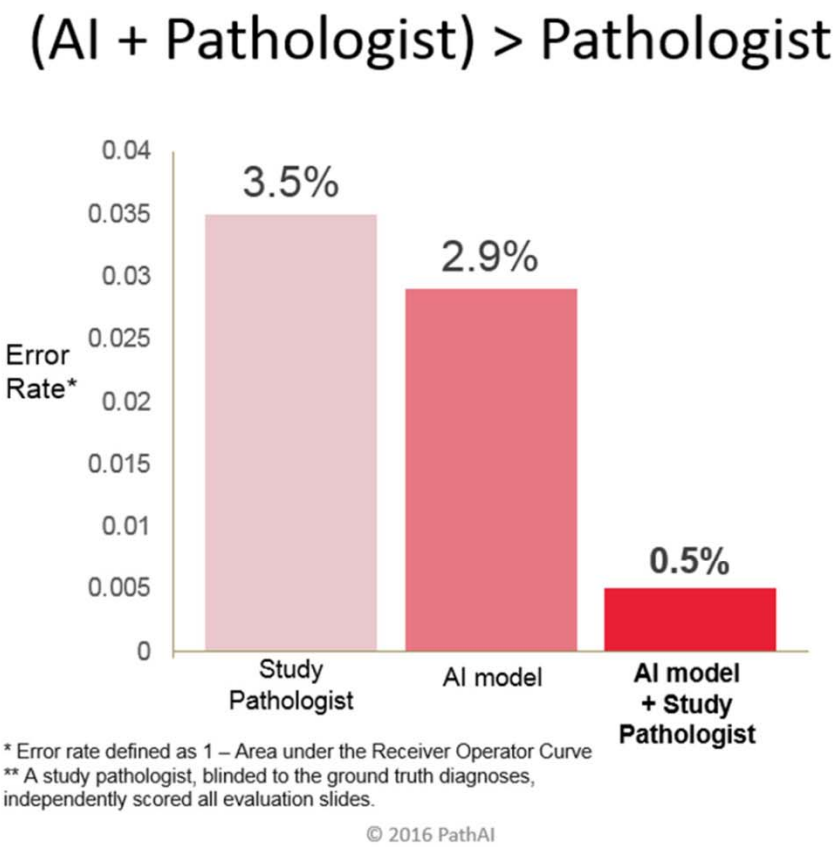
~~$295.00~~ **$84.90**

Cullen Orange Merino Wool V-Neck Sweater

~~$130.00~~ **$49.90**

Supply Chain Analytics in Practice at Rue La La

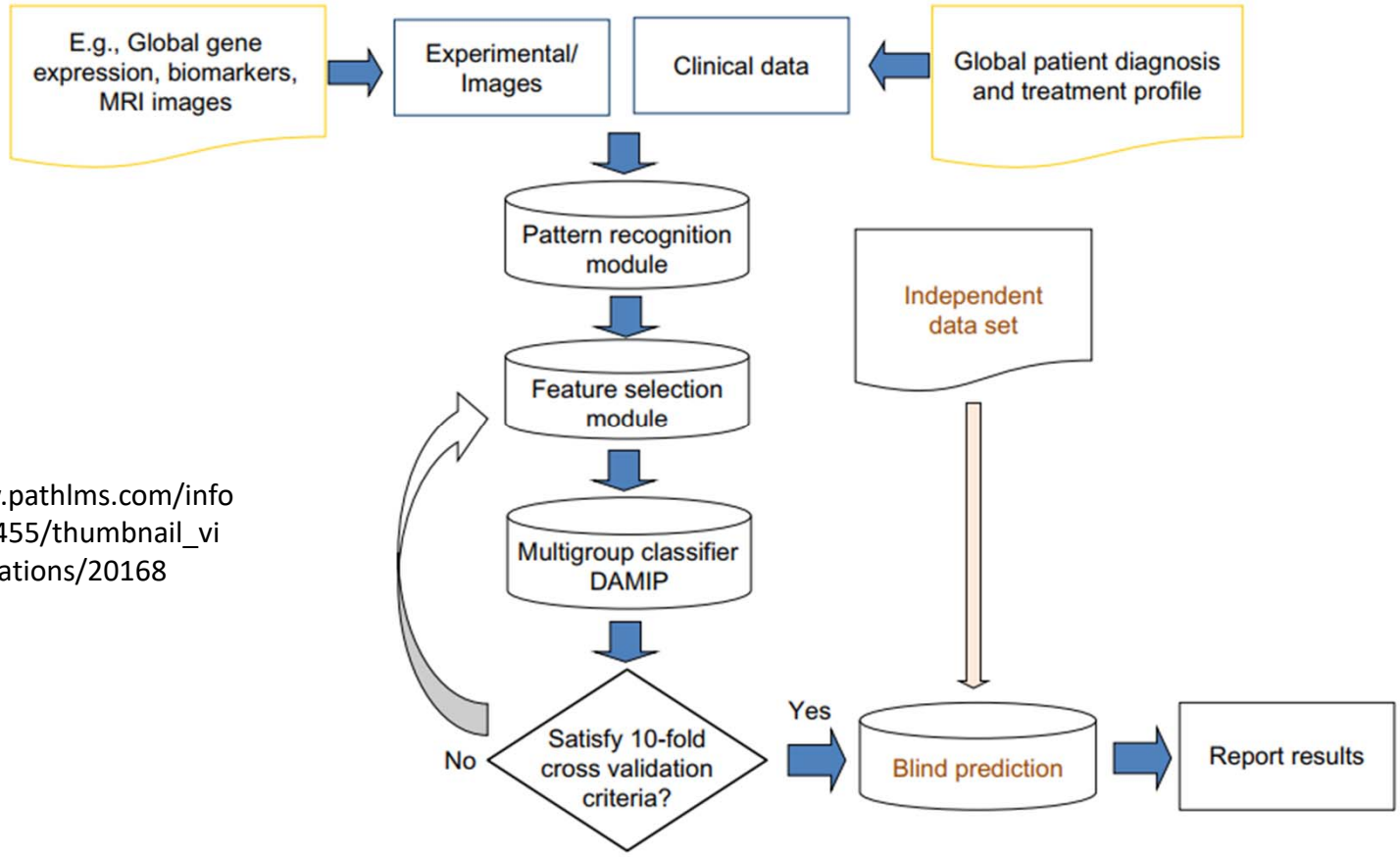
Breast Cancer Diagnoses



D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer,” arXiv preprint arXiv:1606.05718, 2016.

Predicting Vaccine Immunogenicity

- How different individuals respond to vaccination?



https://www.pathlms.com/informs/events/455/thumbnail_video_presentations/20168

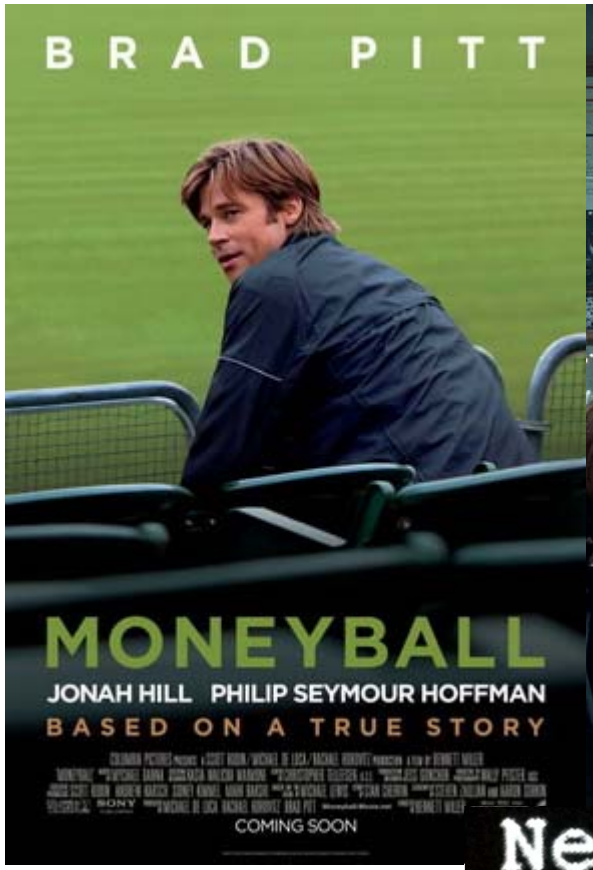
Lee, Eva K., et al. "Machine learning for predicting vaccine immunogenicity." *Interfaces* 46.5 (2016): 368-390.

Game playing

- IBM Deep Blue (1996)
  - 4-2 Garry Kasparov on Chess
  - A large number of crafted rules
  - Huge space search
- Google AlphaGo (2016)
  - 4-1 Lee Sedol on Go
  - Deep machine learning on big data
- AlphaGo Zero (2017)
  - Deep reinforcement learning

Silver, D., et al. Mastering the game of Go with deep neural networks and tree search. Nature 529.7587, 484-489 (2016)
Silver, D., et al. Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017)

Baseball: Moneyball



New York Yankees
$114,457,768
vs
$39,722,689
Oakland Athletics

https://www.bilibili.com/video/av58360594

Text Generation

- Making decision of selecting the next word
- Chinese poem example. Can you distinguish?

Human	Machine
南陌春风早，东邻去日斜。	山夜有雪寒，桂里逢客时。
紫陌追随日，青门相见时。	此时人且饮，酒愁一节梦。
胡风不开花，四气多作雪。	四面客归路，桂花开青竹。

Lantao Yu, Weinan Zhang, et al. Seqgan: sequence generative adversarial nets with policy gradient. AAAI 2017.

Jiaxian Guo, Sidi Lu, Weinan Zhang et al. Long Text Generation via Adversarial Training with Leaked Information. AAAI 2018.

Quick wrap-up

- Business analytics = data → value + decision
- How to turn (big) data into information/value?
- How to make right decisions?
- Machine learning can help us!

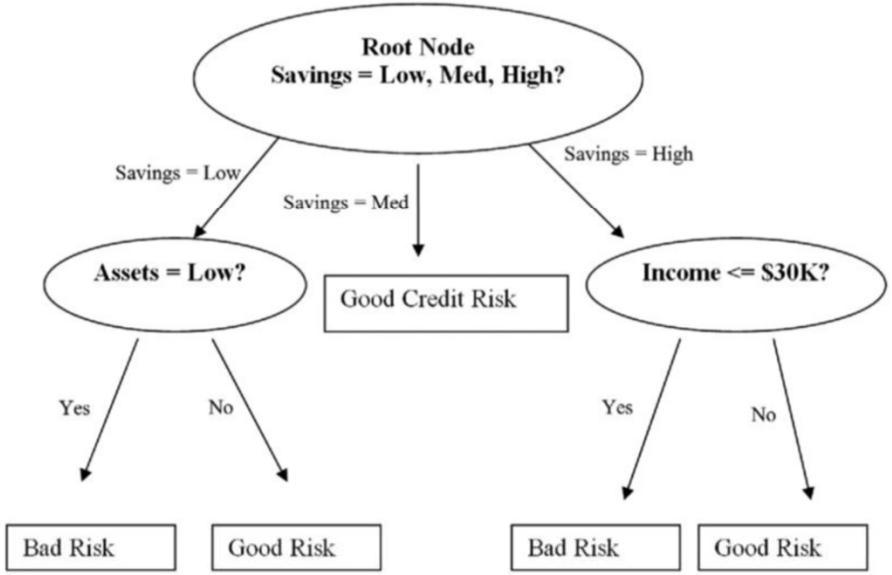
History of ML

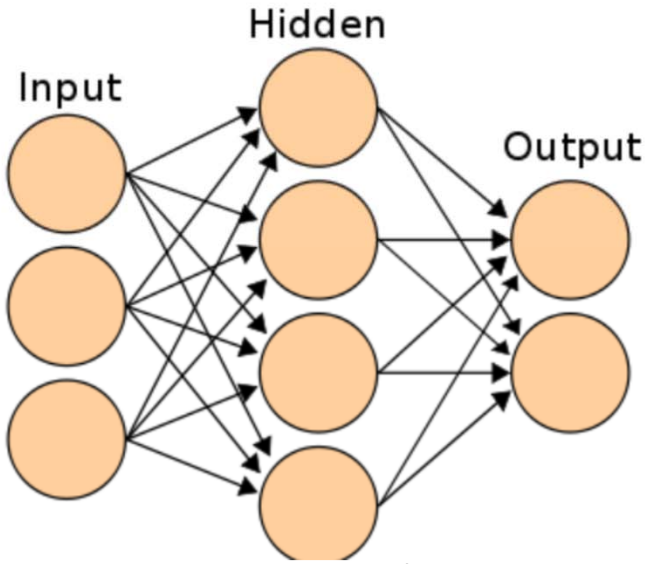
History of machine learning

- 1950s
  - Samuel's checker player
  - Machine learning term created
- 1960s
  - Neural networks: Perceptron
  - Pattern recognition
- 1970s
  - Symbolic concept induction
    - "Logic theorist": We can give machine intelligence if we give them logic
  - Expert systems and the knowledge acquisition bottleneck
    - Only with logic is far from intelligence
    - Machines need knowledge
    - Then find it is hard to teach knowledge summarized by humans to machines
    - It would be better if machines can learn knowledge by themselves!
  - Quinlan's ID3

History of machine learning

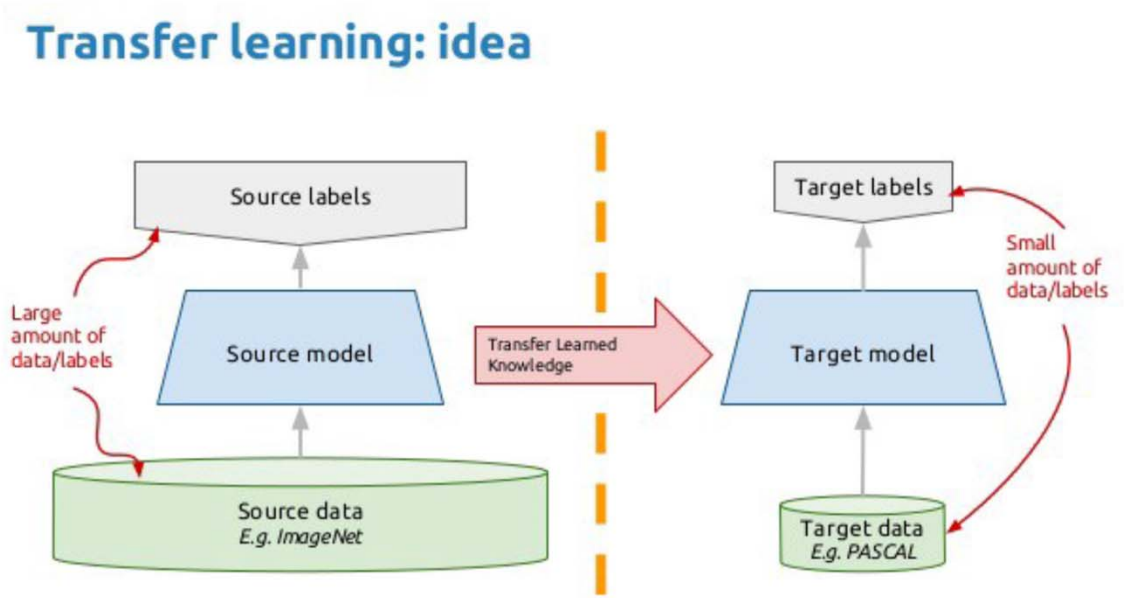
- 1980s
  - Advanced decision tree
    - Learning from samples
    - Simple and efficient
    - Ability to represent knowledge
    - Easy to demonstrate
    - But is hard to learn for large dataset
  - Explanation-based Learning
  - Learning and planning
  - Analogy (类比)
  - Cognitive architectures
  - Resurgence of neural networks
    - Learning from samples
    - Limits: rely heavily on parameters
  - Valiant's PAC Learning Theory (probably approximately correct)





- 1990s
  - Data mining
  - Adaptive software agents and web applications
  - Text learning
  - Reinforcement learning (RL)
  - Ensembles: Bagging, Boosting, and Stacking
  - Bayes Net learning
  - Support vector machines
  - Kernel methods

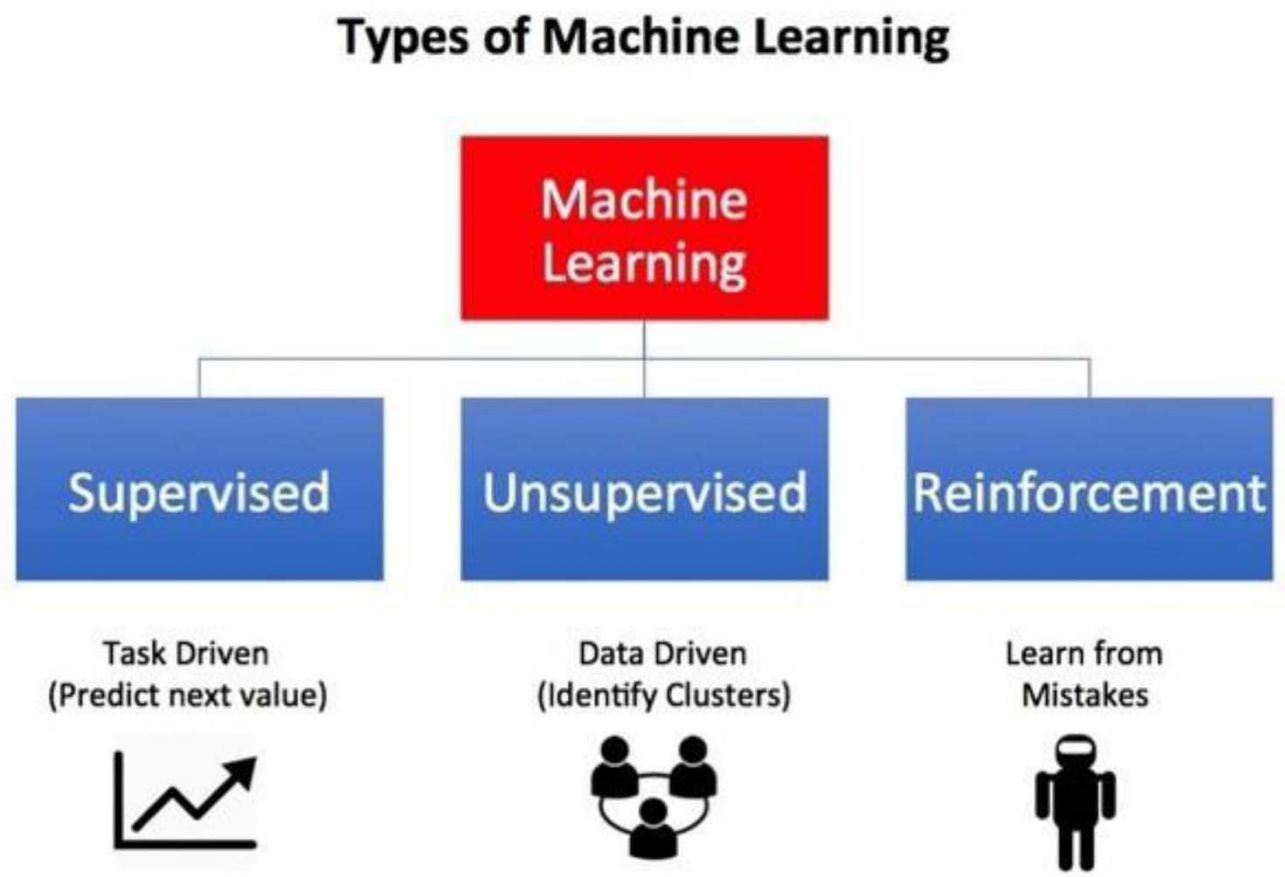
- 2000s
  - Transfer Learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer systems applications
  - Email management
  - Personalized assistants



- 2010s
  - Deep learning
    - Good performances in images/speeches
    - Rely on good parameters (compared to good user previously)
    - Lack of theoretical guarantees but lower threshold to users
  - Learning from big data
  - Learning with GPUs or HPC
  - Multi-task & lifelong learning
  - Deep reinforcement learning
  - Massive applications to vision, speech, text, networks, behavior etc.
  - Meta-learning and AutoML
  - …

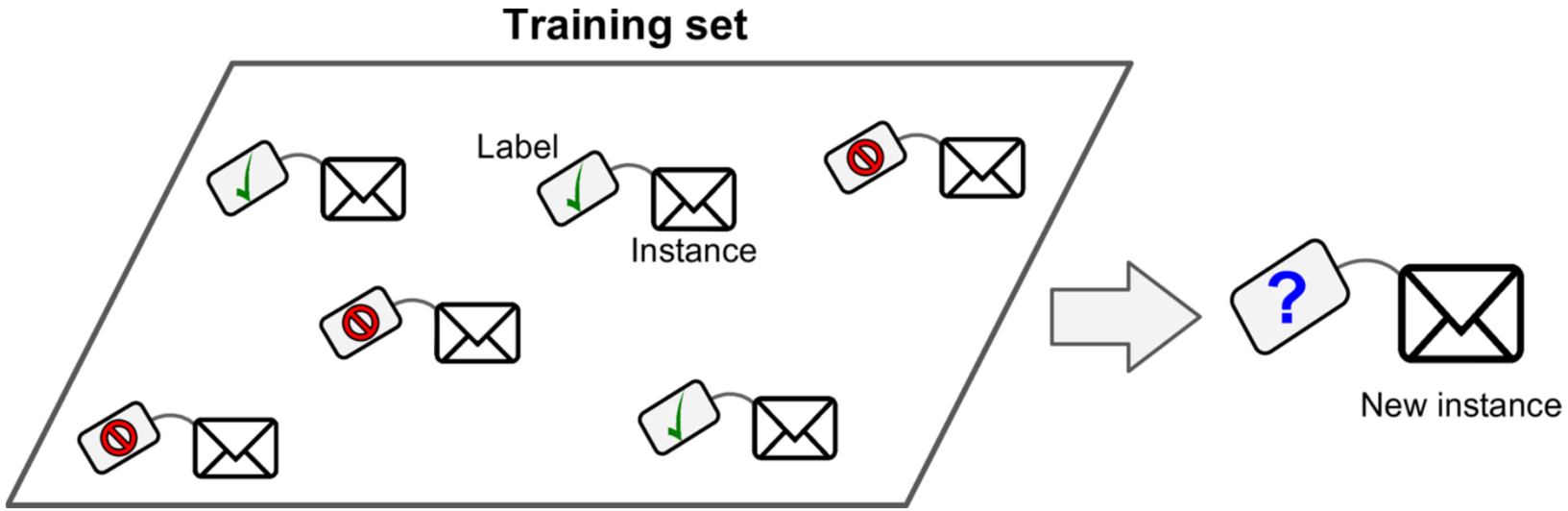
Classifications of ML

Classifications of machine learning



Supervised learning

- Learning a function that maps an input to an output based on example input-output pairs

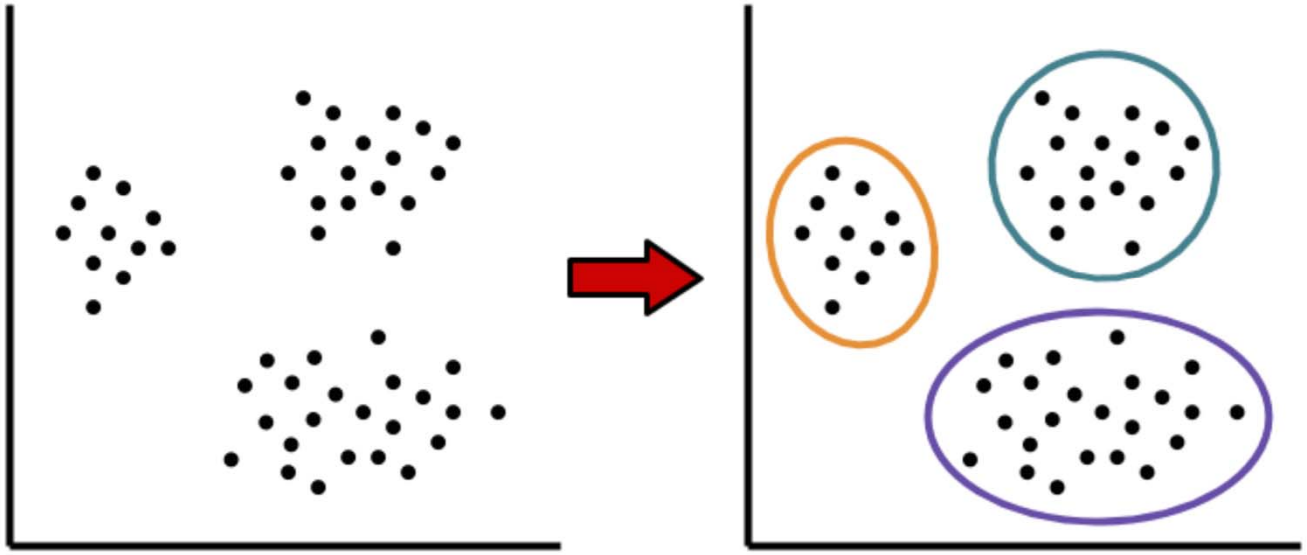


Supervised learning (cont.)

- Linear regression
- Logistic regression
- Decision trees
- Support Vector Machines
- Multi-layer perceptron (one kind of neural networks)
- etc.

Unsupervised learning

- Finding previously unknown patterns in data set without pre-existing labels

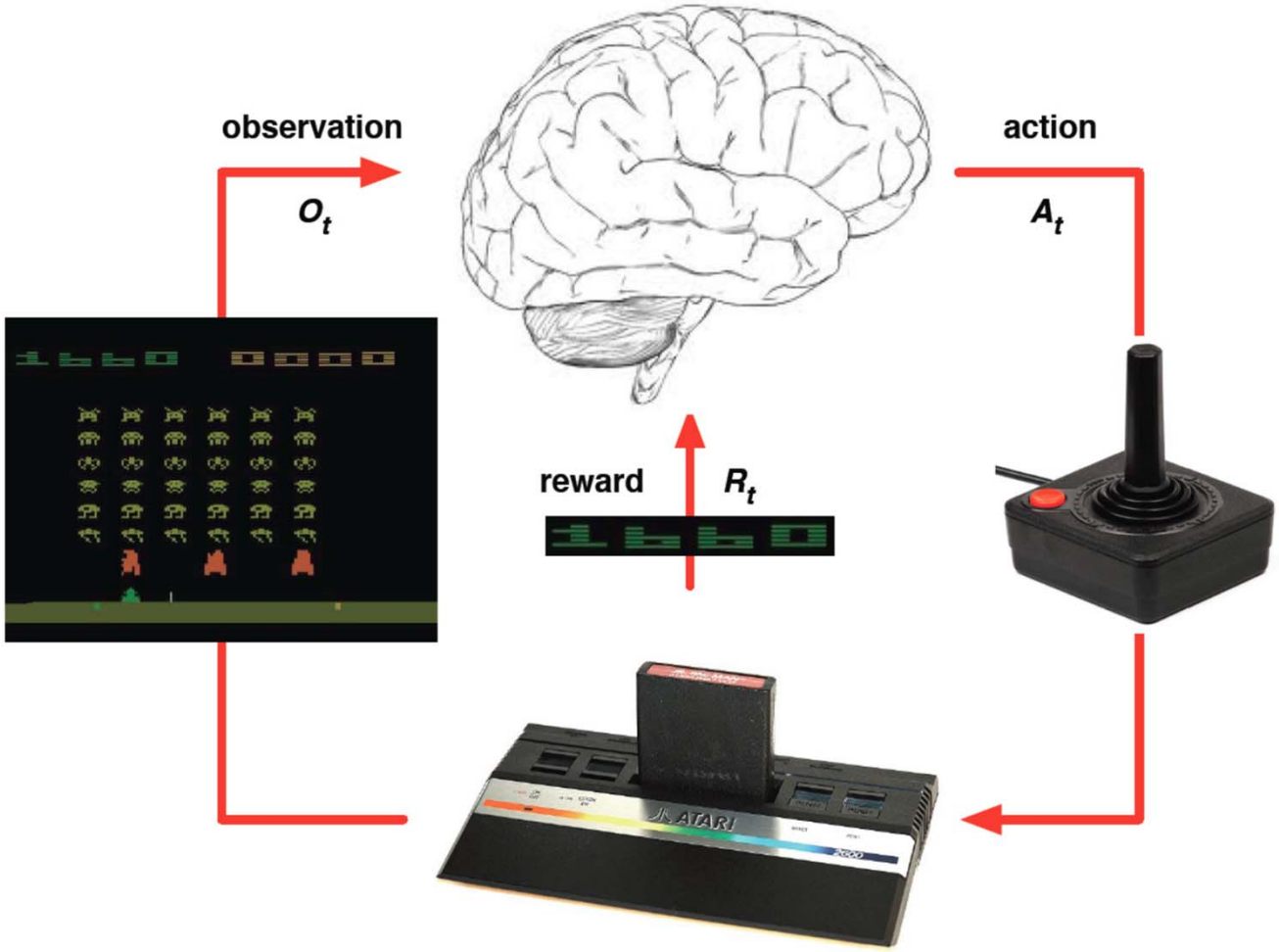


Unsupervised learning (cont.)

- Clustering
- Principal component analysis
- Autoencoders
- Expectation–maximization method (EM)
- etc.

Reinforcement learning

- Taking actions in an environment in order to maximize some notion of cumulative reward



Reinforcement learning (cont.)

- Markov decision process (MDP)
- Approximate dynamic programming (ADP)
- Q-learning
- Deep Q Network (still a kind of neural networks)
- etc.

Deep learning?

- A neural network which is very deep

- When we need?
  - Handle huge amount of data
  - Learn high-level features from data
  - Solve complex problem

- What it needs?
  - High-end machines (GPU etc.)
  - More computational time

Course outline

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

Lecture 1 wrap-up

- ✓What is machine learning
- ✓Machine learning applications
- ✓History of machine learning
- ✓Classifications of machine learning

Reference: VE 445, Shuai LI (SJTU); MSBA 7002, Haipeng SHEN (HKU)

Assignment 1

- A. **Read Home Reading 1A**: Mathematics for Machine Learning
  - Send the page numbers to TA wherever you can't understand
  - TA will collect the information
- B. **Send a life photo** to TA with your name
- Due: TBD
- TA: xiongym3@mail2.sysu.edu.cn

2020 Data Mining and Machine Learning LN3119
https://wangshan731.github.io/DM-ML/

Questions?

Shan Wang (王杉)
https://wangshan731.github.io/