北京工业大学
BEIJINGUNIVERSITYOFTECHNOLOGY

统计机器学习

Statistical Machine Learning

教师：徐硕
单位：北京工业大学经济与管理学院
Email: xushuo@bjut.edu.cn

课程网址：

http://54xushuo.net/wiki/doku.php?id=zh:courses:ml2025:index

课程主要内容

- 绪论（Introduction）
- 监督学习法（Supervised Learning）
- 支持向量机（Support Vector Machine）
- 序列标注方法（Sequence Labeling）
- 无监督学习法（Unsupervised Learning）
- 概率主题模型（Probabilistic Topic Modeling）

考核方式

- 成绩由平时成绩和英文学术论文报告成绩两部分构成
  - 平时成绩占20%
  - 课程论文报告成绩占80%
    - 老师打分：40%
    - 自己打分：15%
    - 其他同学打分：25%

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

大数据是未来的新油田

Gartner

什么是大数据?

- **Wiki: Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- 大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。
- 维克托《大数据时代》：大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。（理想很丰满，但现实很骨感）

故事：大小之争

- 1936年，Alfred Landon（共和党，兰登） vs. Franklin Roosevelt（民主党，罗斯福）竞选下届总统
- 《The Literary Digest》（文学文摘）承担选情预测（1920、1924、1928、1932成功预测）
- 《The Literary Digest》寄出**1000万份**调查问卷，覆盖当时1/4的选民，回收**240万份**，预测Landon将会以**55:41**的优势获胜
- 实际结果是： Roosevelt以**61:37**的压倒性优势获胜
- 新民意调查的开创者George Gallup（盖洛普）仅仅通过**3000人**的问卷调查，得到准确得多的预测结果

大数据的特征

1. Volume

数据量大

全球在2010年正式进入ZB时代，IDC预计到2029年，全球**数据量将达到500ZB**

2. Variety

数据类型繁多

数据类型不是单一的文本形式，而是结构化数据、半结构化数据和非结构化等多种形式的综合

3. Velocity

实时变化迅速

大数据区分于传统数据最显著的特征。涉及到感知、传输、决策、控制开放式循环的大数据，数据具有高度时效性(动态科技信息、用户动态需求等)

4. Value

沙里淘金，价值密度低

如何通过强大的机器算法更迅速的完成数据的价值“提纯”是目前大数据汹涌背景下亟待解决的难题

故事：园中有金不在金

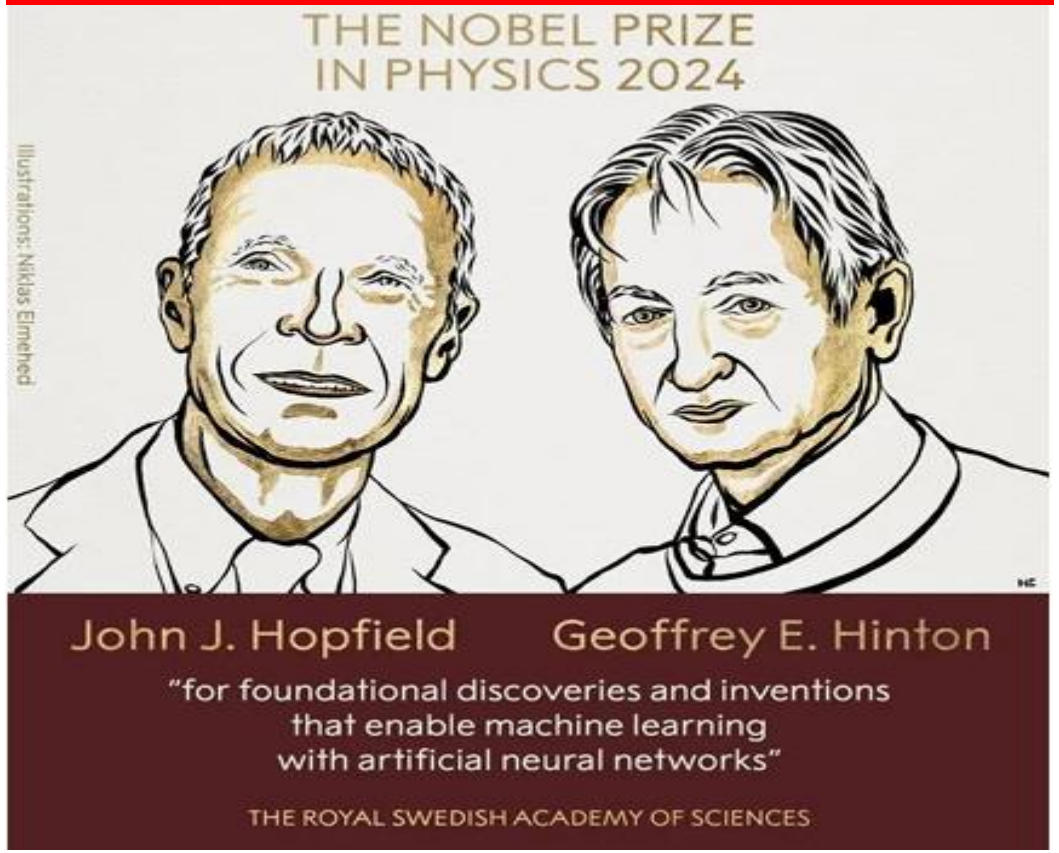
- 寓言故事《园中有金》：
  - 有父子二人，居山村，营果园。父病后，子不勤耕作，园渐荒芜。一日，父病危，谓子曰：园中有金。子翻地寻金，无所得，甚怅然。是年秋，园中葡萄、苹果之属皆大丰收。子始悟父言之理。
- 大数据的价值主要体现在驱动效应上，大数据对经济的贡献，并不完全反映在公司的直接收入上，应考虑对其他行业效率和质量提高的贡献。（**蜜蜂模型**）

科学研究方法论的演变（1/2）

第一范式	第二范式	第三范式	第四范式	第五范式
经验范式	理论范式	模拟和计算范式	数据驱动范式	AI4Science
1600s	1950s	2000s	2020s	当今
科学家基于经验的观察，对万物万象的总结。比如，天文学家开普勒通过观察总结出天体运行的规律。	科学家对经验进行数学抽象和推演。比如，用于描述经典力学的牛顿运动方程，用于描述电场磁场关系的麦克斯韦方程等。	以电子计算机的诞生为契机，科学家开始有能力探究复杂的物理问题。比如，天气预报，大型粒子对撞机，传染病的传播等。	计算能力和传感器的无处不在，机器学习（ML）扮演着非常重要的角色，人们开始用ML方法分析数据，寻找规律，并进行预测。	前四种范式的有机结合，发挥了经验和理论各自的特长，又把AI和计算科学融合在一起，是对科学发现更全面的认知。

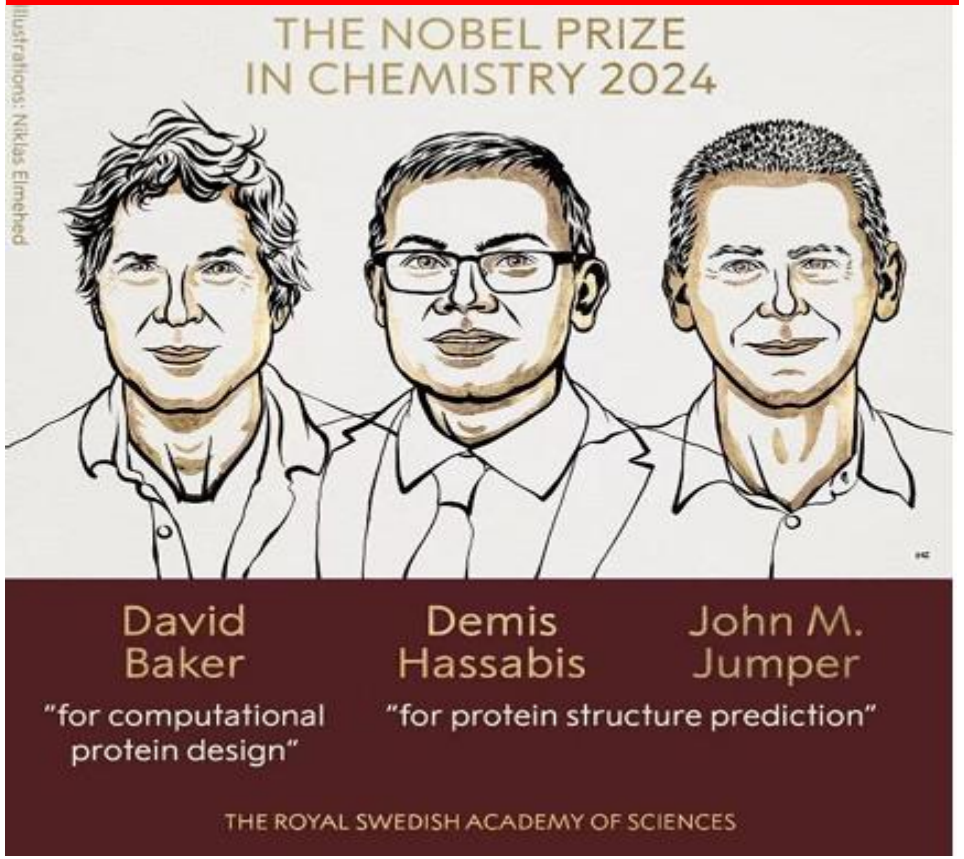
科学研究方法论的演变（2/2）

Science for AI



瑞典皇家科学院将2024年诺贝尔物理学奖授予John J. Hopfield和Geoffrey E. Hinton，以表彰他们“为利用人工神经网络进行机器学习做出的基础性发现和发明”。

AI for Science



瑞典皇家科学院将2024年诺贝尔化学奖授予David Baker，以表彰其在计算蛋白质设计方面的贡献，另一半则共同授予Demis Hassabis和John M. Jumper，以表彰其在蛋白质结构预测方面的贡献。

段子：恐怖的大数据

“杨达才事件”

SODA创新应用大赛（1/3）

SODA创新应用大赛（2/3）

关于SODA　大赛日程　比赛数据　SODA奖项　竞赛规则　大赛组织

2015年上海开放数据创新应用大赛比赛用数据表

城市道路交通指数　地铁运行数据　一卡通乘客刷卡数据　浦东公交车实时数据　强生出租车行车数据　空气质量状况

气象数据　道路事故数据　高架匝道关闭数据　新浪微博交通数据

提供单位
上海市城乡建设和交通发展研究院

具体数据项
状态、 区域、 当前指数、 参考指数、 指数差值

数据格式
CSV

时间范围
20140701-201504030

SODA创新应用大赛（3/3）

上海政府数据服务网已开放交通类数据表

1. 机动车维修企业

提供单位：市交通委

具体数据项

企业名称、经营地址、经营范围、联系电话

2. 车辆统计信息

提供单位：市交通委

具体数据项

车辆类别、统计数据

3. 公交线路首末班车时间

提供单位：市交通委

具体数据项

线路名称、上行首末班车时间、下行首末班车时间

4. 停车场静态信息

提供单位：市交通委

具体数据项

核定收费标准、营业时间、联系电话、车位数、停车场性质、入口地址、出口地址

5. 停车场（库、位）

提供单位：市交通委

具体数据项

备案证号、地址、核定收费标准、时间、电话

6. 全市道路货运搬场企业名录

提供单位：市交通委

具体数据项

许可证号、企业名称、企业地址

7. 全市道路货运出租企业名录

提供单位：市交通委

具体数据项

许可证号、企业名称、企业地址、联系电话

8. 港口经营企业名录

提供单位：市交通委

具体数据项

许可证号、企业名称、企业地址

9. 国内航运企业名录

提供单位：市交通委

具体数据项

许可证号、企业名称、企业地址

10. 全市营运交通线路统计

提供单位：市交通委

具体数据项

营运交通线路类别、统计数

11. 全市营运交通工具统计

提供单位：市交通委

具体数据项

营运交通工具类别、统计数

12. 全市公共停车场库经营统计

提供单位：市交通委

具体数据项

公共停车场类别、统计数

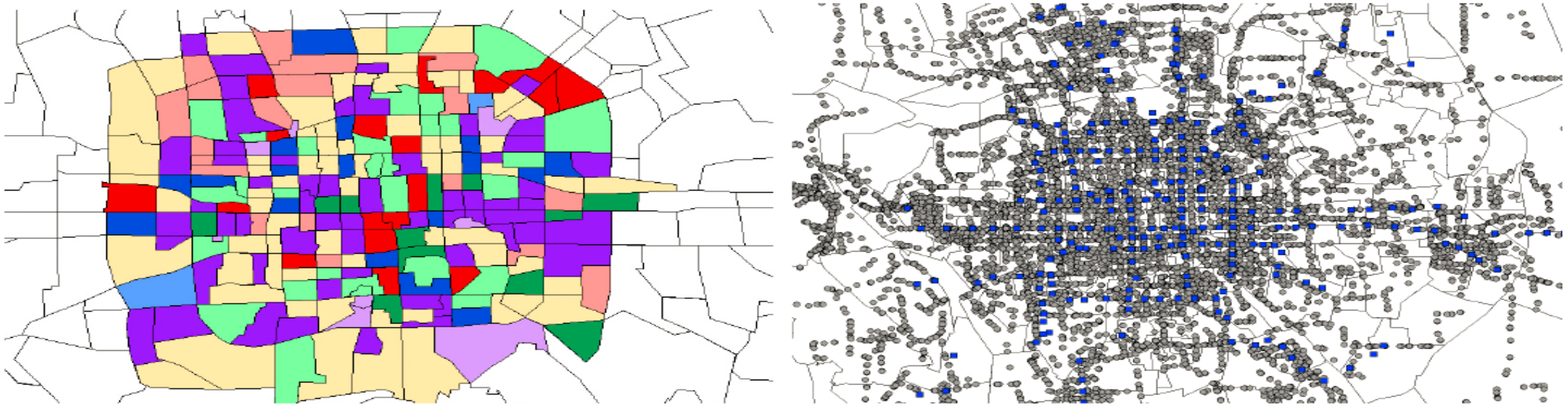
公交一卡通：抓小偷（1/7）

#抓小偷为百姓服务# 7月10日早7时40分，便衣反扒小伟队长带大魁组在349路程庄路口东车站抓获一名男小偷，现案为乘客挽回公交IC卡一张，真应了那句话叫：贼不走空！嫌疑人吴某，27岁，来自河南桐柏县，健壮有力、聪明伶俐，选择不劳而获确实是不应该，目前被公交警方刑事拘留。PS：想起《小花》了



今天 11:00 来自 微博 weibo.com

公交一卡通：抓小偷（2/7）



(a) Functional Regions　　(b) Stations

北京市2014年4-6月三个月间600万乘客的约16亿公交一卡通数据记录，将北京划分为多个小的局部的功能区块，分析了896条公交线经过的44524个公交车站和18条地铁线经过的320个地铁站的数据。

◆Bowen Du, Chuanren Liu, Wenjun Zhou, Zhenshan Hou, and Hui Xiong, 2016. Catch Me If You Can: Detecting Pickpocket Suspects from Large-Scale Transit Records. *KDD 2016*.

公交一卡通：抓小偷（3/7）

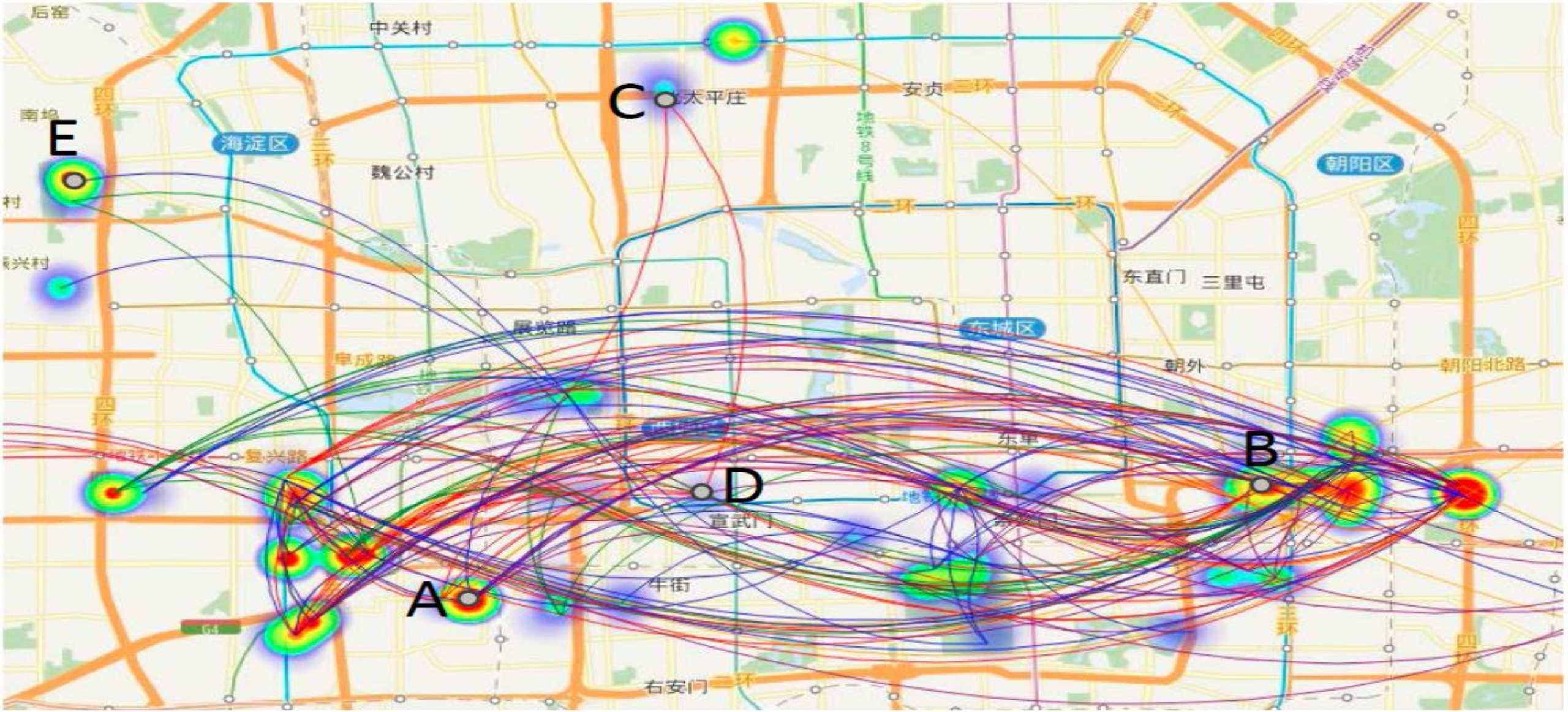
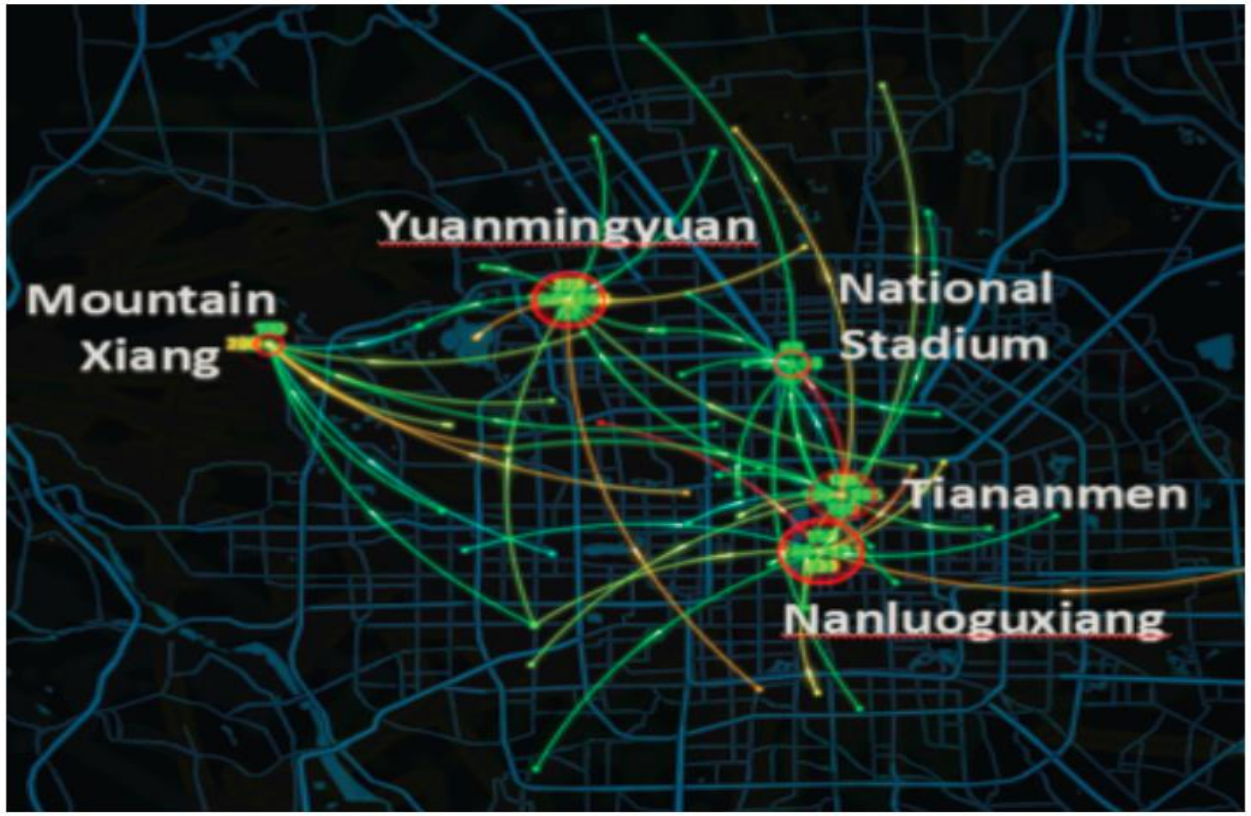


Figure 1: Trajectories of passengers.

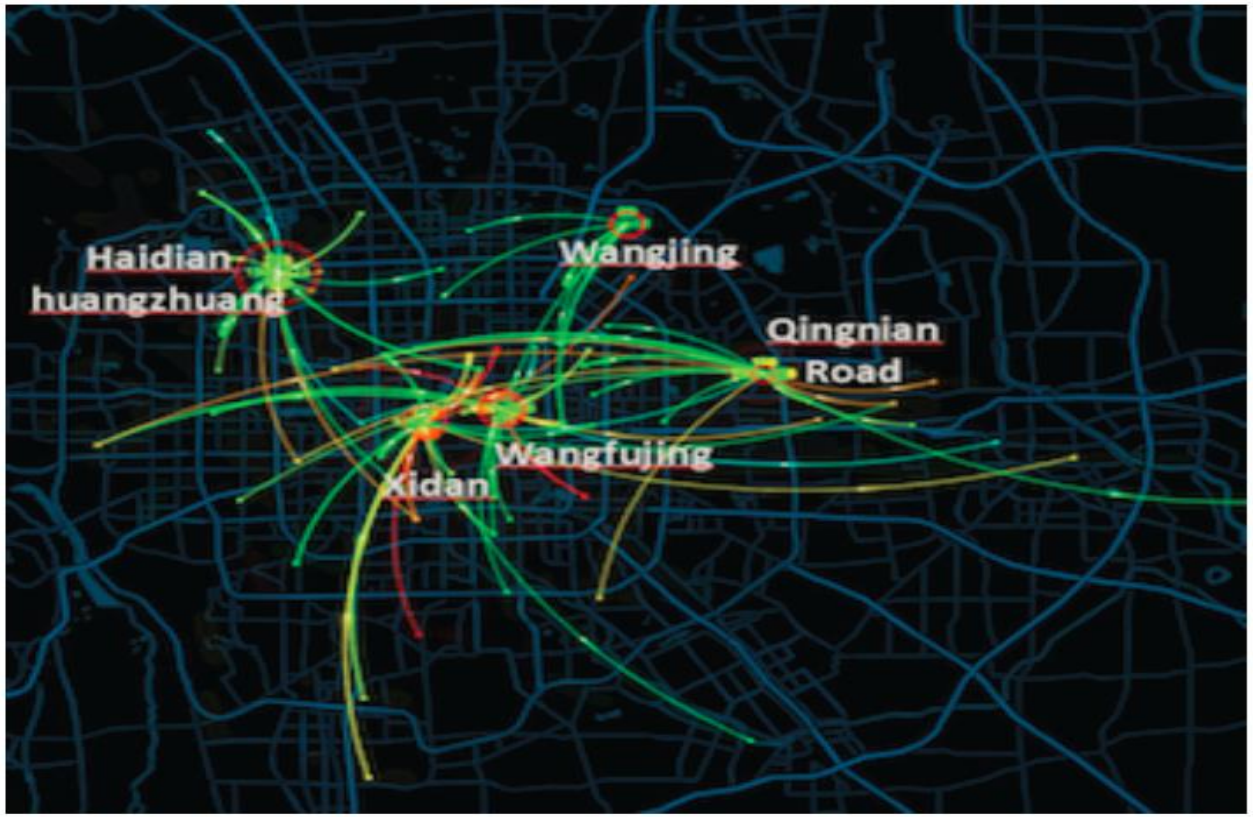
公交一卡通：抓小偷（4/7）

(a) all passengers

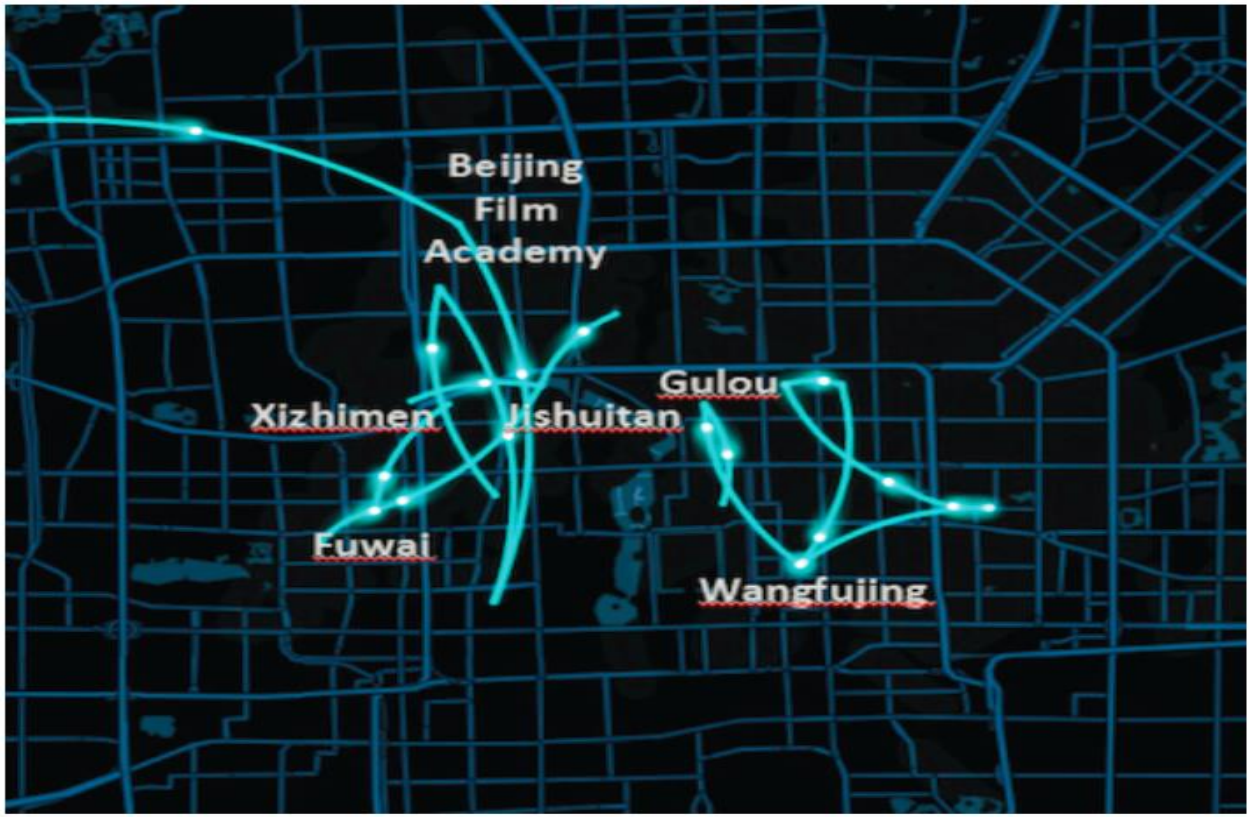


(b) visitors

公交一卡通：抓小偷（5/7）



(c) shoppers



(d) thieves

公交一卡通：抓小偷（6/7）

Passenger 5335568 | Commuter 2310458 | Outlier 13205 | Suspects 315

Passenger Flow

Active Region

Suspects List

ID	Prob.	Freq.	Trip	ShortRides	TravTime	Historical Profile
53174410	0.90	29	5	20	25131	TripFreq 10.06 TripTime 11837.4
52537759	0.88	14	4	8	16047	TripFreq 6.875 TripTime 6897.1
44492021	0.79	10	7	7	27146	TripFreq 5.1 TripTime 13070.4

Selected Suspect

公交一卡通：抓小偷（7/7）

Table 3: A Performance Comparison.

Algorithm	Precision	Recall	F-score	Run Time(s)
CM Methods				
DT	0.002	0.451	0.004	44.81
LR	0.003	0.476	0.006	36.72
SVM	0.005	0.512	0.009	21.31
AD Methods				
LOF	0.004	0.560	0.009	300+
OCSVM	0.015	0.583	0.029	39.67
TS Methods				
LOF+DT	0.011	0.780	0.022	301.18+
LOF+LR	0.016	0.829	0.031	301.16+
OCSVM+DT	0.053	0.878	0.099	41.19
TS-SVM	**0.071**	**0.927**	**0.133**	**41.05**

第一章：绪论

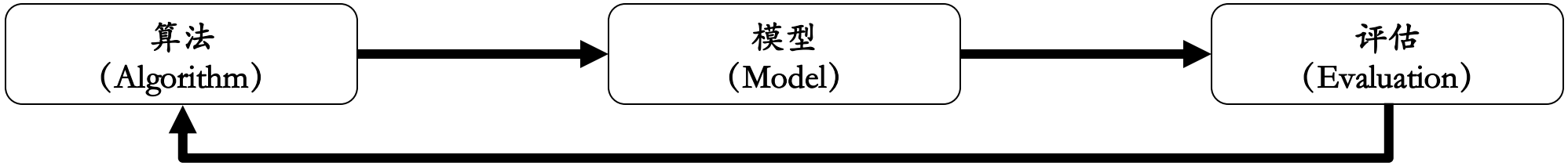
- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

什么是机器学习（1/3）

- 在维基百科上，对机器学习提出以下几种定义：
  - 机器学习是一门研究人工智能的学科，该领域的主要研究对象是人工智能，特别是如何在**经验**学习中改善具体**算法**的**性能**；
  - 机器学习是对能通过**经验**自动改进的计算机**算法**的研究；
  - 机器学习是用数据或以往的**经验**，以此优化计算机程序的**性能**标准。

什么是机器学习（2/3）

- 三个关键词：**算法**、**经验**、**性能**



- 机器学习是数据通过算法构建出模型并对模型进行评估，评估的性能如果达到要求就利用这个模型来测试其他的数据，如果达不到要求就调整算法来重新构建模型，再次进行评估，如此循环往复，最终获得满意的经验来处理其他的数据。

什么是机器学习（3/3）

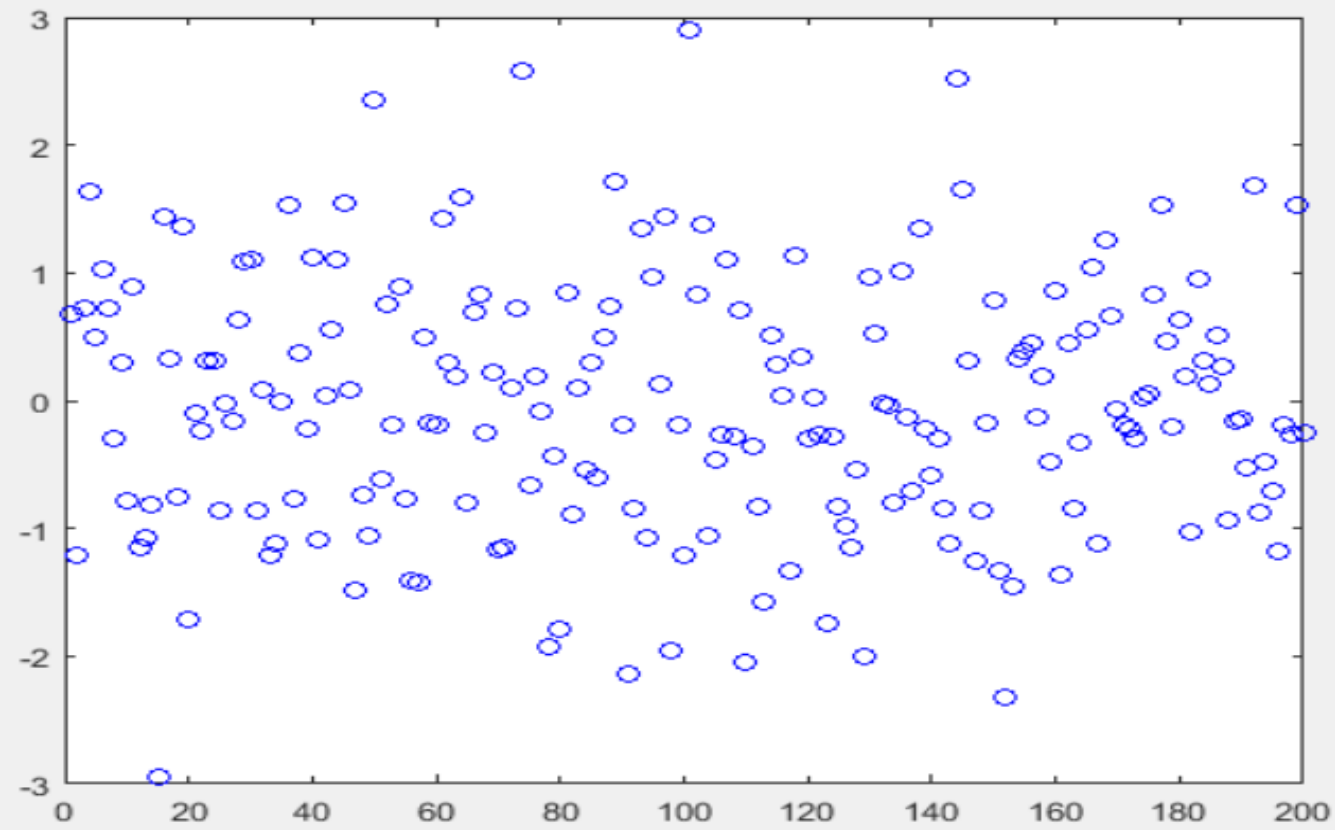
- 统计机器学习的过程符合人类认识事物的规律，总是先接触到个别的事物，而后推及一般，又从一般推及个别，如此循环往复，使认识不断深化
- 这体现了归纳（模型构建）与演绎（预测分析）的辩证统一
- 思想在我国先秦著作《论语》中就曾记载过，比如“举一隅不以三隅返，则不复也”

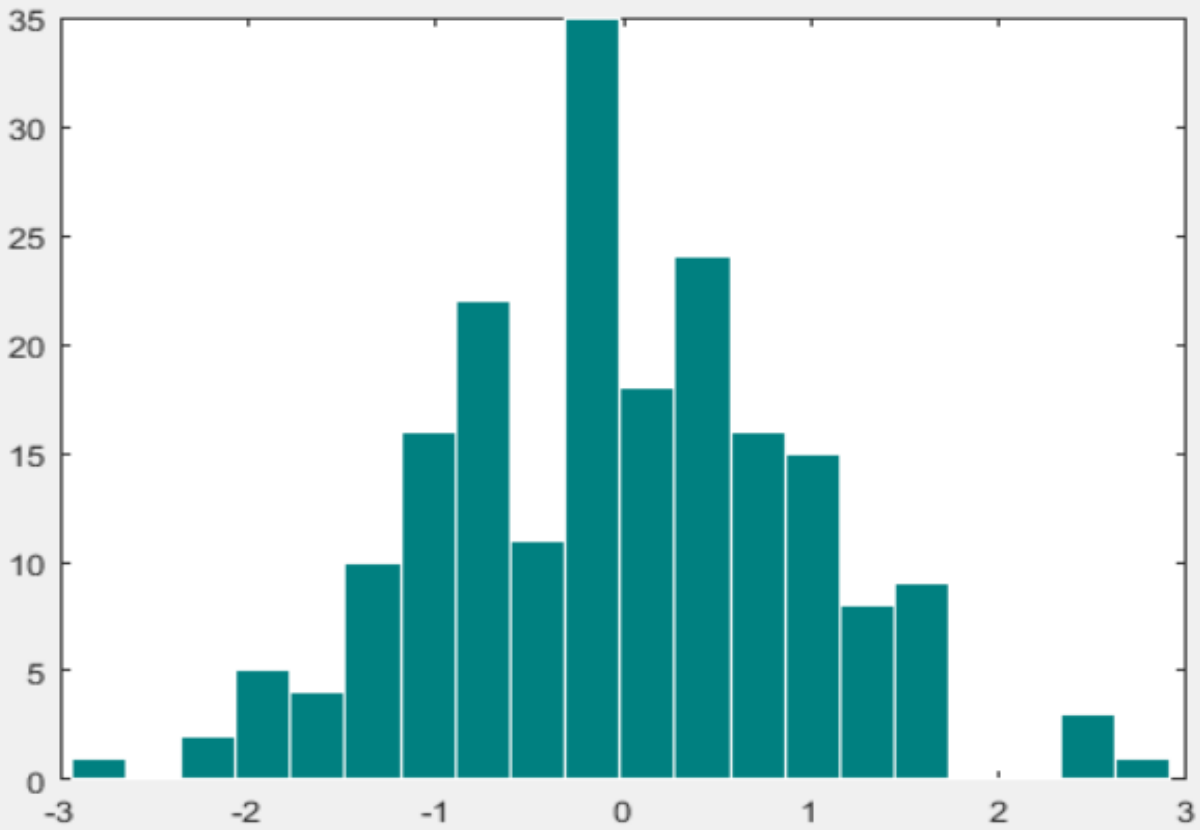
如何构建模型

- 方法一：对数据进行简洁的近似汇总描述

- 方法二：从数据中抽取出最突出的特征，代替数据，并忽略剩余内容

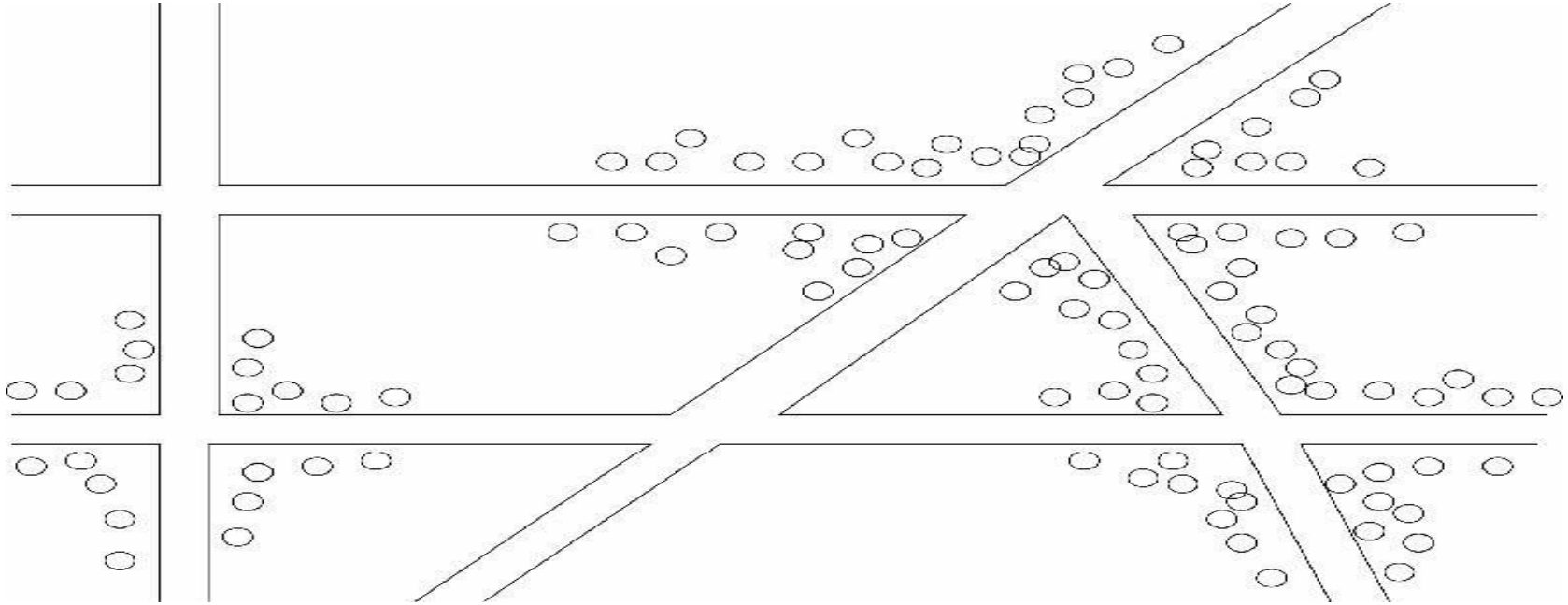
建模方法：数据汇总


3
2
1
0
-1
-2
-3
0
20
40
60
80
100
120
140
160
180
200

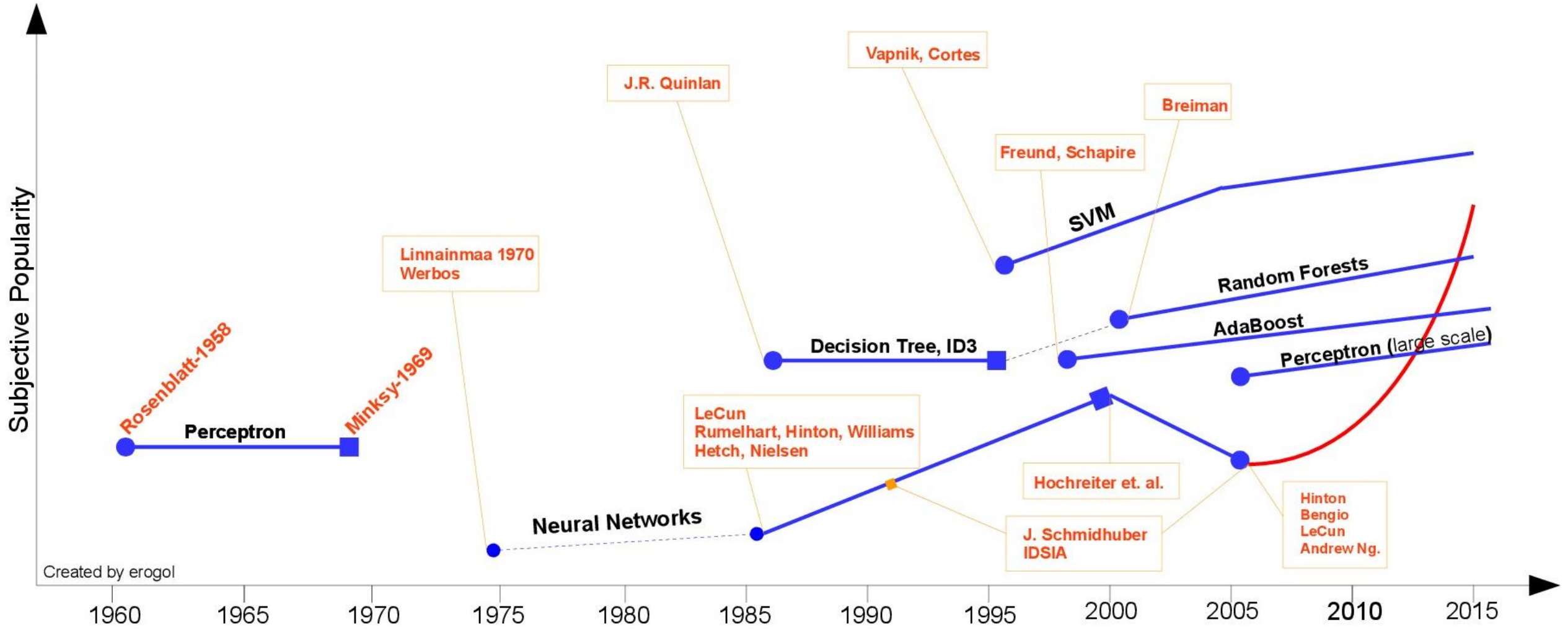

35
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3

伦敦地图标出的霍乱传播情况

- 英国麻醉师John Snow在1854年处理霍乱爆发时，不相信“瘴气”说，绘制了一张疫情地图，标注病例的分布以及水井的位置，直观地揭示了污染的井水与霍乱疫情的关系
- 所采用的流行病学调查方法对当前新冠病毒的溯源工作仍然具有一定的参考价值。



机器学习的发展历史



Source: https://chatbotnewsdaily.com/since-the-initial-standpoint-of-science-technology-and-ai-scientists-following-blaise-pascal-and-804ac13d8151

机器学习的发展历史

- 1949年，Hebb（加拿大）提出赫布学习理论（Hebbian learning theory），用来解释学习过程中大脑神经元所发生的变化，标志着机器学习领域迈出的第一步。
- 赫布学习理论研究的是循环神经网络（RNN）中各节点之间的关联性，这儿的RNN具有把相似神经网络连接在一起的特征，并起到类似于记忆的作用。

- Donal Oolding Hebb, 1949. The Organization of Behavior. New York: Wiley & Sons.

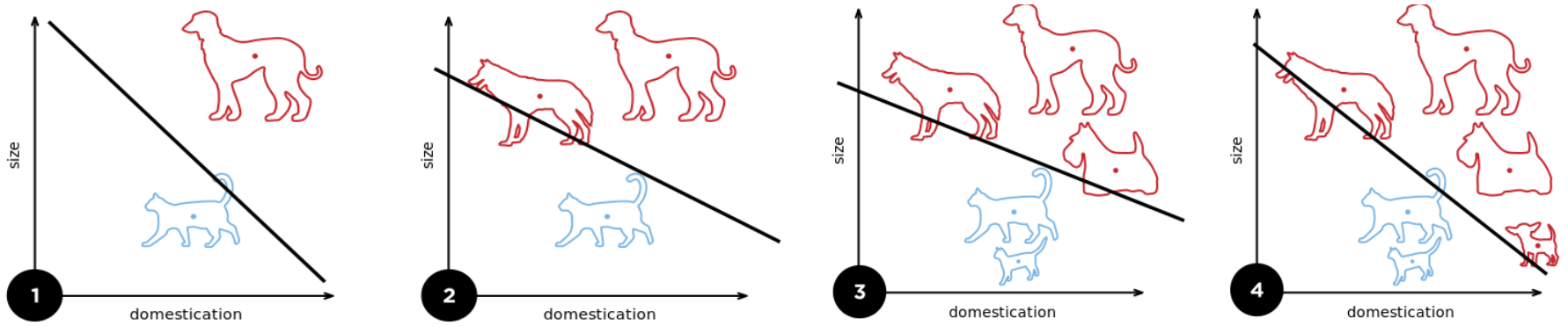
机器学习的发展历史

- 1952年，IBM公司的Samuel（美国，被誉为“机器学习之父”）设计了一款可以学习的西洋跳棋程序。它能通过观察棋子的走位来构建新的模型，并用其提高自己的下棋技巧。
- Samuel用这个程序推翻了以往“机器无法超越人类，不能像人类一样写代码和学习”的传统认识。
- 将“机器学习”定义为：不需要显式编程就可以赋予机器某项技能的研究领域（A field of study that gives computer the ability without being explicitly programmed.）。

◆ Arthur L. Samuel, 1959. Some Studies in Machine Learning using the Game of Checkers. *IBM Journal of Research and Development*, 44: 206-226.

机器学习的发展历史：感知器（1/2）

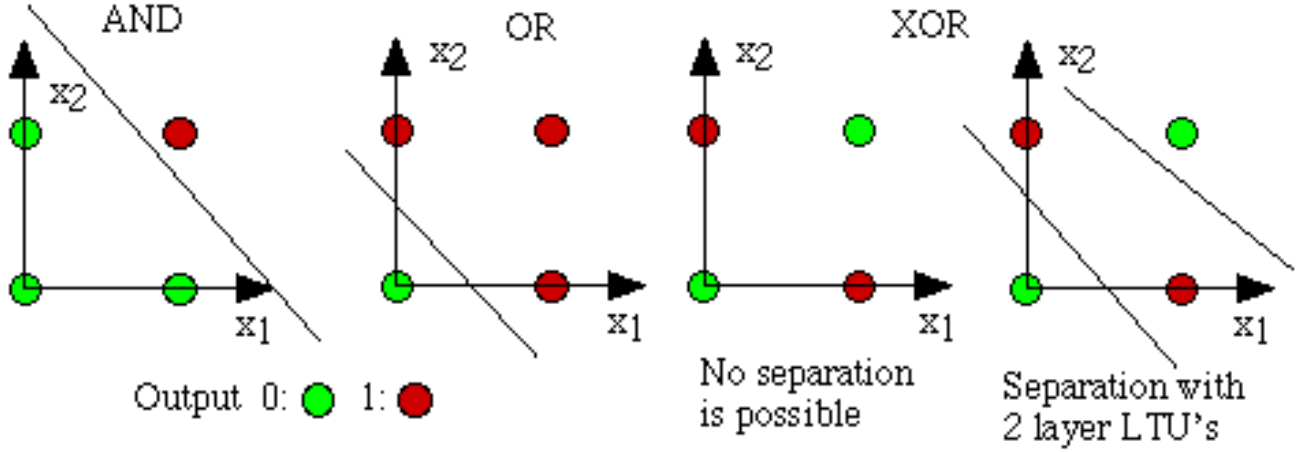
- 1957年，具备神经科学背景的Rosenblatt（美国）提出了感知器（Perceptron）模型，它更接近于如今的机器学习模型。
- 感知器可以在较简单的结构中表现出智能系统的基本属性，也就是说研究人员不需要再拘泥于具体生物神经网络特殊及未知的复杂结构。
- 三年后，Widrow和Hoff提出了差量学习规则（Delta learning rule），被应用于感知器模型中，可以创建出更精准的线性分类器。



- Frank Rosenblatt, 1957. The Perceptorn: A Perceiving and Recognizing Automation. *Report 85-460-4*. Cornell Aeronautical Laboratory.
- B. Widron & M. Hoff, 1960. Adaptive Switching Circuits, pp. 96-104.

机器学习的发展历史：感知器（2/2）

- 1969年，Minsky和Papert提出的异或（XOR）问题，暴露了感知器模型的本质缺陷：无法处理线性不可分问题，此后神经网络研究陷入了长达十多年的停滞。
- 尽管1970年，Linnainmaa首次完整地叙述了反向模式自动微积分算法（反向传播算法BP的雏形），但在当时并没有引起重视。

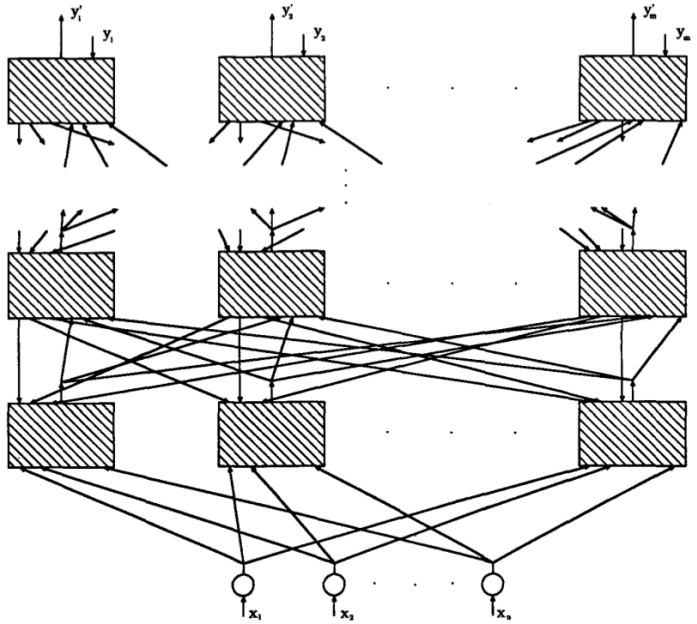


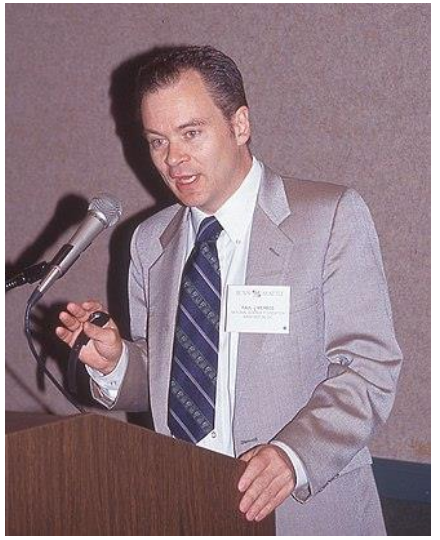
◆ Marvin Minsky & Seymour Papert, 1969. Perceptrons: An Introduction to Computational Geometry.

◆ Seppo Linnainmaa, 1970. The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors (in Finnish). Master's Thesis, University of Helsinki, Finland.

机器学习的发展历史：BP神经网络

- 直到Werbos（美国）于1981年提出将BP算法应用于神经网络以建立多层感知器后，神经网络的发展才得以提速。
- 接着在1985-1986两年间，多位神经网络学者也相继提出了使用BP算法来训练多层感知器的相关想法

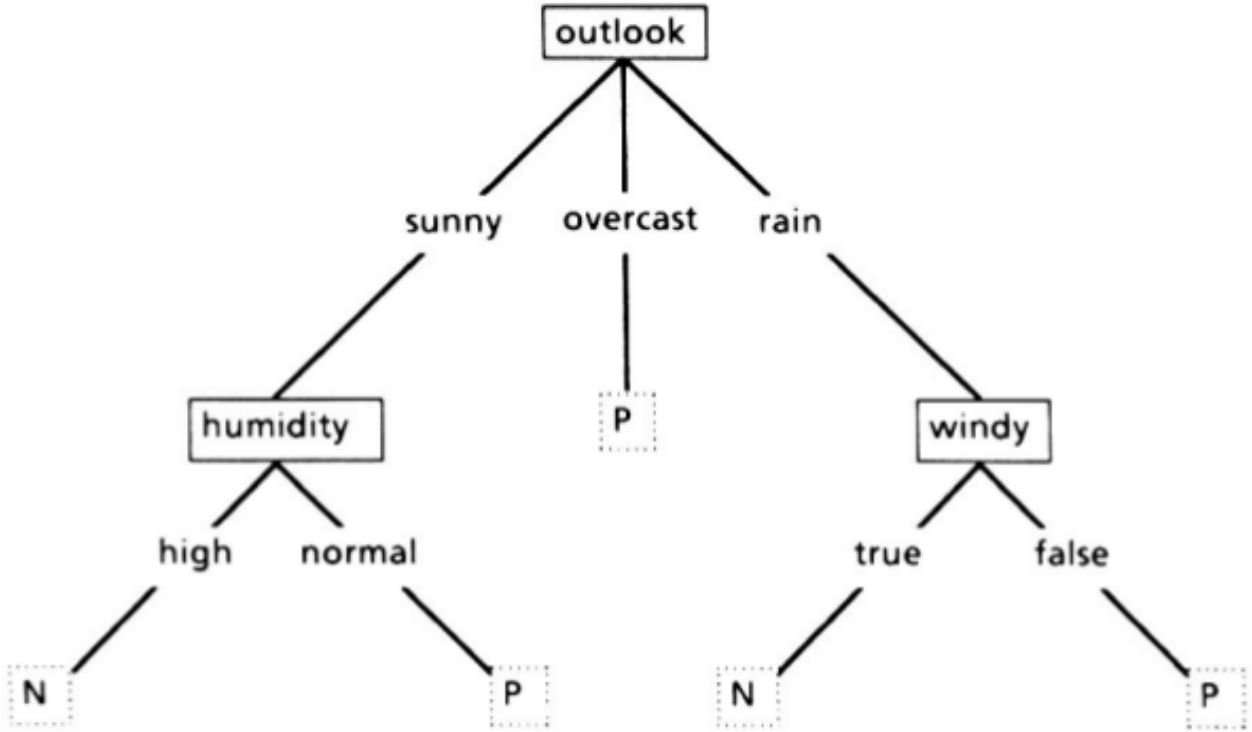




◆ Paul J. Werbos, 1981. Applications of Advances in Nonlinear Sensitivity Analysis. Proceedings of the 10th IFIP Conference, pp. 762-770.

机器学习的发展历史：决策树

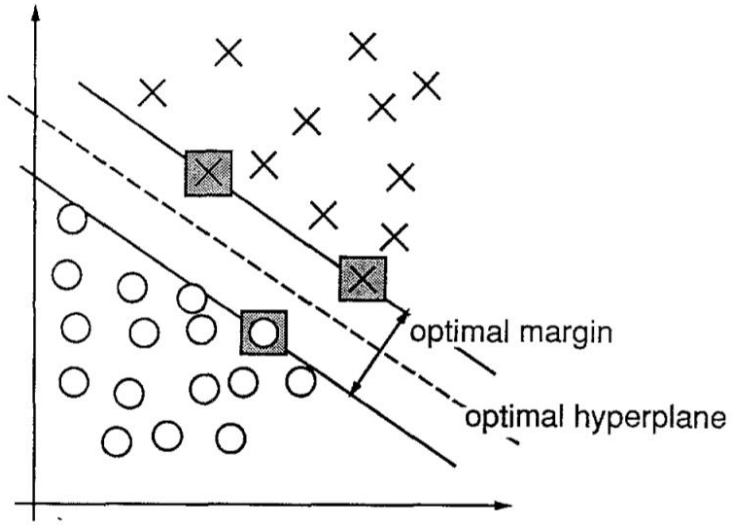
- 1986年，Quinlan（澳大利亚）提出了著名的决策树算法（ID3）
- 与“黑箱式”的神经网络模型不同，决策权采用了简单的规则
- 以后有许多优化改进算法，比如ID4、回归树、C4.5、C5.0等



◆ John Ross Quinlan, 1986. Induction of Decision Trees. *Machine Learning*, Vol. 1, pp. 81-106.

机器学习的发展历史：支持向量机

- 1995年，Vapnik和Cortes提出了支持向量机（SVM）方法,不仅有坚实的理论基础，而且实验结果表现出色。
- 从此以后，机器学习分为了两大流派：神经网络和支持向量机
- 2000年，核版的SVM被提出后，神经网络逐渐处于下风，SVM在此前神经网络垄断的领域中都取得了亮眼的成绩



◆ Corinna Cortes & Vladimir Vapnik, 1995. Support-Vector Networks. *Machine Learning*, Vol. 20, pp. 273-297.

机器学习的发展历史：BP神经网络

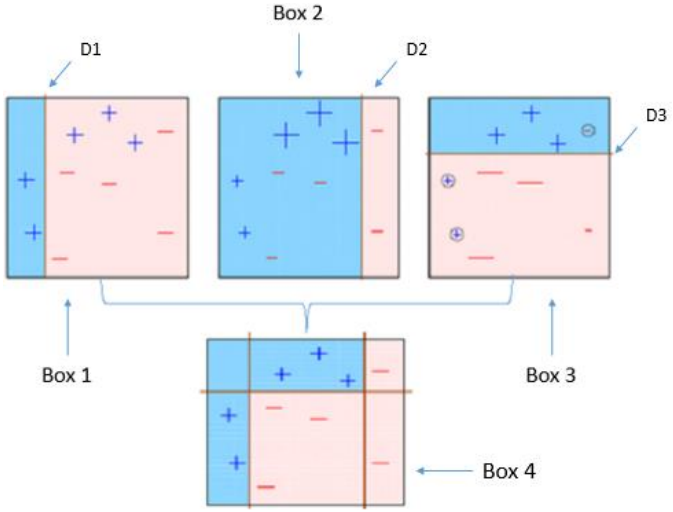
- 在SVM持续发展之际，神经网络再次受到了重创。Hochreiter（德国）及其合作者发现，使用BP算法时，当神经网络的单元饱和之后，系统会发生梯度损失（又称梯度扩散）。
- 简单来说，就是训练神经网络模型的时候，超过一定的迭代次数后，模型将会产生过拟合（over-fitting）。

◆ Sepp Hochreiter, 1991. Untersuchungen zu Dynamischen Neuronalen Netzen (in German). Master's Thesis, Technische Universität München.

◆ Sepp Hochreiter, A. Steven Younger, & Peter R. Conwell, 2001. Learning to Learn using Gradient Descent. *Proceedings of the International Conference on Artificial Neural Networks*, pp. 87-94.

机器学习的发展历史：集成学习（1/2）

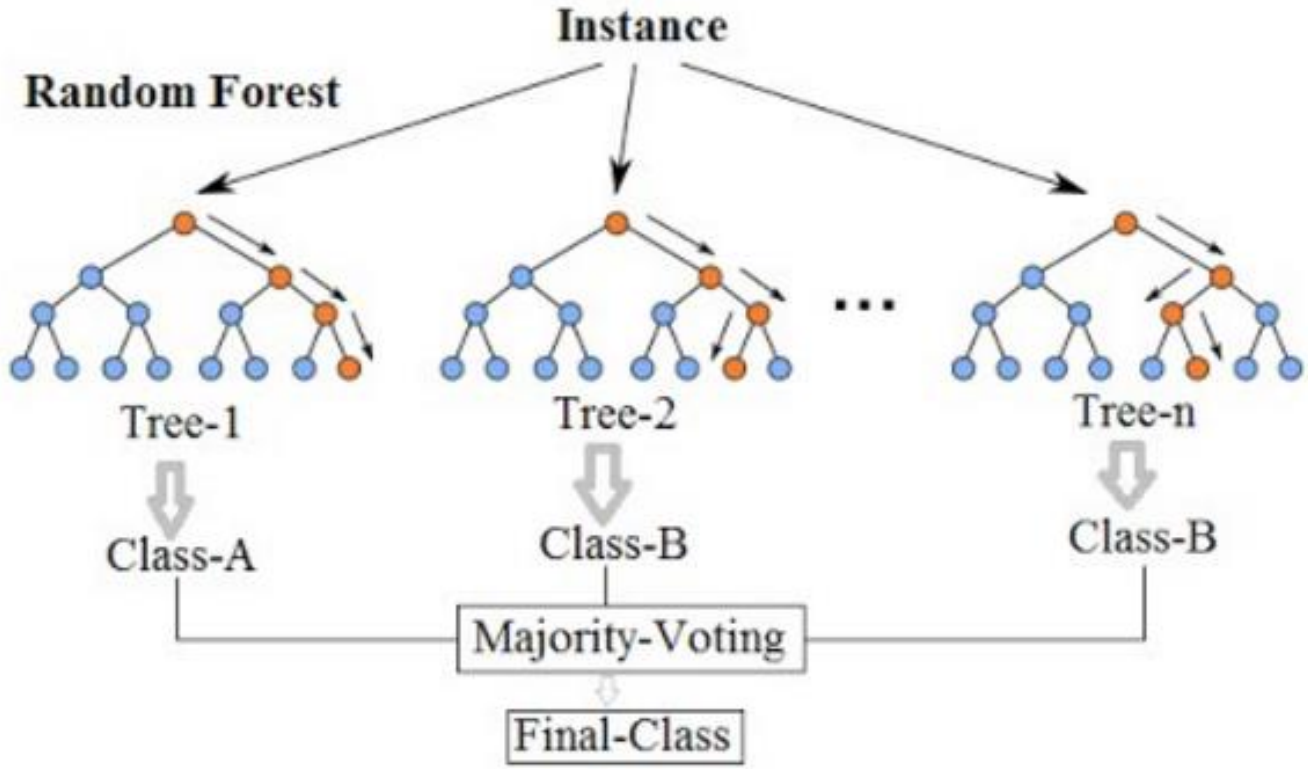
- 在此之前，1997年Freund和Schapire（美国）提出另外一个有效的机器学习模型：Adaboost，该模型为两位学者赢得了哥德尔（Gödel）奖。
- 核心思想：针对同一个训练集训练不同的弱分类器，然后将它们集合起来，构建出更强的最终分类器（强分类器）。
- 实际上，Adaboost可以看作是把多个不同的决策树采用一种非随机方式组合在一起的提升树。



◆ Yoav Freund & Robert E. Schapire, 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119-139.

机器学习的发展历史：集成学习（2/2）

- 2001年，Breiman（美国）提出了另一种多决策树组合模型：随机森林
- 随机森林模型中单个决策树则一个随机子集训练而得，而决策树中每一节点则各选自这一随机子集



◆ Leo Breiman, 2001. Random Forests. *Machine Learning*, Vol. 45, No. 1, pp. 5-32-297.

机器学习的发展历史：深度学习（1/3）

The Milestone Of LSTM 1997

Sepp Hochreiter and Jürgen Schmidhuber publishes a milestone paper on "Long Short-Term Memory" (LSTM). It is a type of recurrent neural network architecture which will go on to revolutionize deep learning in decades to come.

Deep Belief Network 2006


ast learning algorithm for deep belief nets

and Simon Osindero — Yee-Whye T

Science University of Toronto — Department of Compu

College Road — National University of

anada M5S 3G4 — 3 Science Drive 3, Singa

ero}@cs.toronto.edu — tehyw@comp.nus.

act

plementary priors" to
way effects that make
sely-connected belief
n layers. Using com-
e a fast, greedy algo-
rected belief networks
ded the top two lay-
ociative memory. The
d to initialize a slower
-tunes the weights us-
f the wake-sleep algo-
a network with three

remaining hidden layers form
converts the representations in
observable variables such as th
brid model has some attractive

1. There is a fast, greedy le
a fairly good set of par
networks with millions o
layers.
2. The learning algorithm
plied to labeled data by l
both the label and the dat
3. There is a fine-tuning al
lent generative model w

Geoffrey Hinton, Ruslan Salakhutdinov, Osindero and Teh publishes the paper "*A fast learning algorithm for deep belief nets*" in which they stacked multiple RBMs together in layers and called them Deep Belief Networks. The training process is much more efficient for large amount of data.

GPU Revolution Begins 2008

Andrew NG's group in Stanford starts advocating for the use of GPUs for training Deep Neural Networks to speed up the training time by many folds. This could bring practicality in the field of Deep Learning for training on huge volume of data efficiently.

机器学习的发展历史：深度学习（2/3）

ImageNet Is Launched

2009

Finding enough labeled data has always been a challenge for Deep Learning community. In 2009 Fei-Fei Li, a professor at Stanford, launches ImageNet which is a database of 14 million labeled images. It would serve as a benchmark for the deep learning researchers who would participate in ImageNet competitions (ILSVRC) every year.

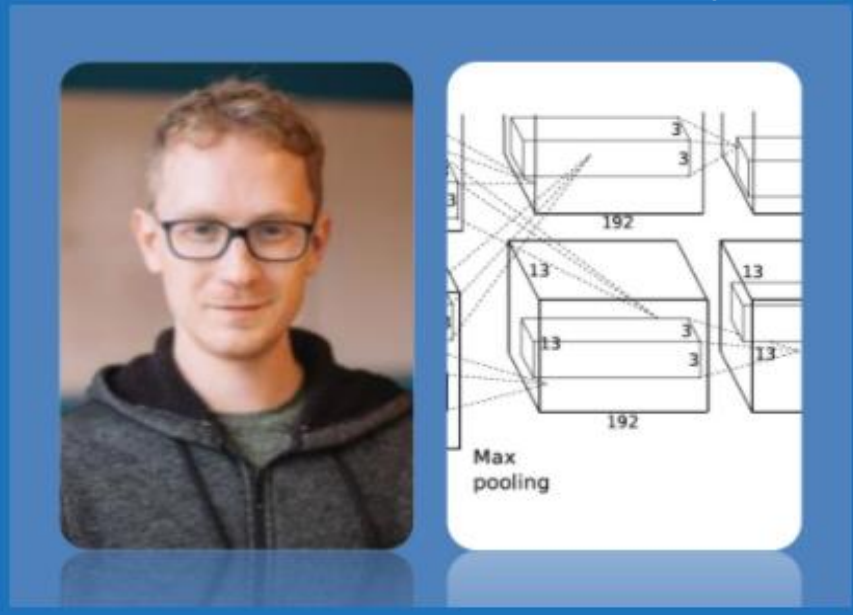
Combat For Vanishing Gradient

2011

Yoshua Bengio, Antoine Bordes, Xavier Glorot in their paper "Deep Sparse Rectifier Neural Networks" shows that ReLU activation function can avoid vanishing gradient problem. This means that now, apart from GPU, deep learning community has another tool to avoid issues of longer and impractical training times of deep neural network.

AlexNet Starts Deep Learning Boom

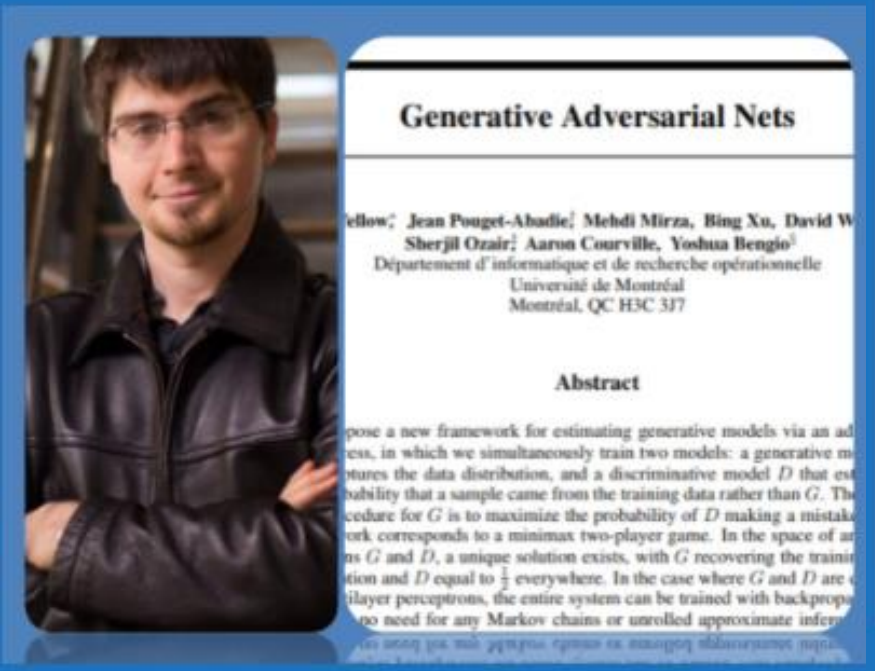
2012



AlexNet, a GPU implemented CNN model designed by Alex Krizhevsky, wins Imagenet's image classification contest with accuracy of 84%. It is a huge jump over 75% accuracy that earlier models had achieved. This win triggers a new deep learning boom globally.

机器学习的发展历史：深度学习（3/3）

The Birth Of GANs 1997



Generative Adversarial Nets

Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David W
Sherjil Ozair, Aaron Courville, Yoshua Bengio
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

pose a new framework for estimating generative models via an ad
ess, in which we simultaneously train two models: a generative m
tures the data distribution, and a discriminative model D that es
bability that a sample came from the training data rather than G. Th
cedure for G is to maximize the probability of D making a mistak
ork corresponds to a minimax two-player game. In the space of ar
ns G and D, a unique solution exists, with G recovering the traini
tion and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are
ilayer perceptrons, the entire system can be trained with backpropa
no need for any Markov chains or unrolled approximate infer

Generative Adversarial Neural Network also known as GAN is created by Ian Goodfellow. GANs open a whole new doors of application of deep learning in fashion, art, science due it's ability to synthesize real like data.

AlphaGo Beats Human 2006

Deepmind's deep reinforcement learning model beats human champion in the complex game of Go. The game is much more complex than chess, so this feat captures the imagination of everyone and takes the promise of deep learning to whole new level.

Trio Win Turing Award 2018

Yoshua Bengio, Geoffrey Hinton, and Yann LeCun wins Turing Award 2018 for their immense contribution in advancements in area of deep learning and artificial intelligence. This is a defining moment for those who had worked relentlessly on neural networks when entire machine learning community had moved away from it in 1970s.

相关会议和期刊

会议

- ICML
- NIPS (NeurIPS)
- CVPR
- KDD
- SDM
- ICDM
- PKDD
- PAKDD
- CIKM
- UAI
- SIGIR
- ICDE

期刊

- Machine Learning
- Journal of Machine Learning Research
- Neurocomputing
- Expert Systems with Applications
- Applied Soft Computing
- Data Mining and Knowledge Discovery (DMKD)
- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- Pattern Recognition
- ACM Trans. on KDD

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

监督学习

- 监督学习是从给定的训练数据集中学习一个函数（模型），当新的数据到来时，可根据这个函数（模型）预测结果；
- 在监督式学习下，输入数据被称为“训练数据”，每条训练数据有一个明确的标识或结果；
- 在建立模型时，监督式学习建立一个学习过程，将预测结果与“测试数据”的实际结果进行比较，不断调整预测模型，直到模型的预测结果达到一个预期的准确率。

分类的定义：以文本分类为例

- 定义：给定分类体系，将数据（文本）分到某个或几个类别中。
- 分类体系：一般人工构造
  - 有层级结构：MeSH主题分类表、中国分类法（CLC）、 国际专利分类法（IPC）
  - 无层级结构：{政治、体育、军事}、{动词、名词、形容词、…}
- 分类模式：
  - 两类问题（binary）：一篇文本属于或不属于某个特定类别；
  - 多类问题（multi-class）：一篇文本属于多个类别中的某一个；
  - 多标识问题（multi-label）：一篇文本同时属于多个类别；

文本分类：应用

- 垃圾邮件的判定（spam or not spam）
  - 类别 {spam, not-spam}
- 新闻出版按照栏目分类
  - 类别 {政治, 体育, 军事, …}
- 词性标注
  - 类别 {名词, 动词, 形容词, …}
- 词义排歧
  - 类别 {词义1, 词义2, …}
- 计算机论文的领域
  - 类别 ACM system
    - H: information systems
    - H.3: information retrieval and storage
- …

分类方法

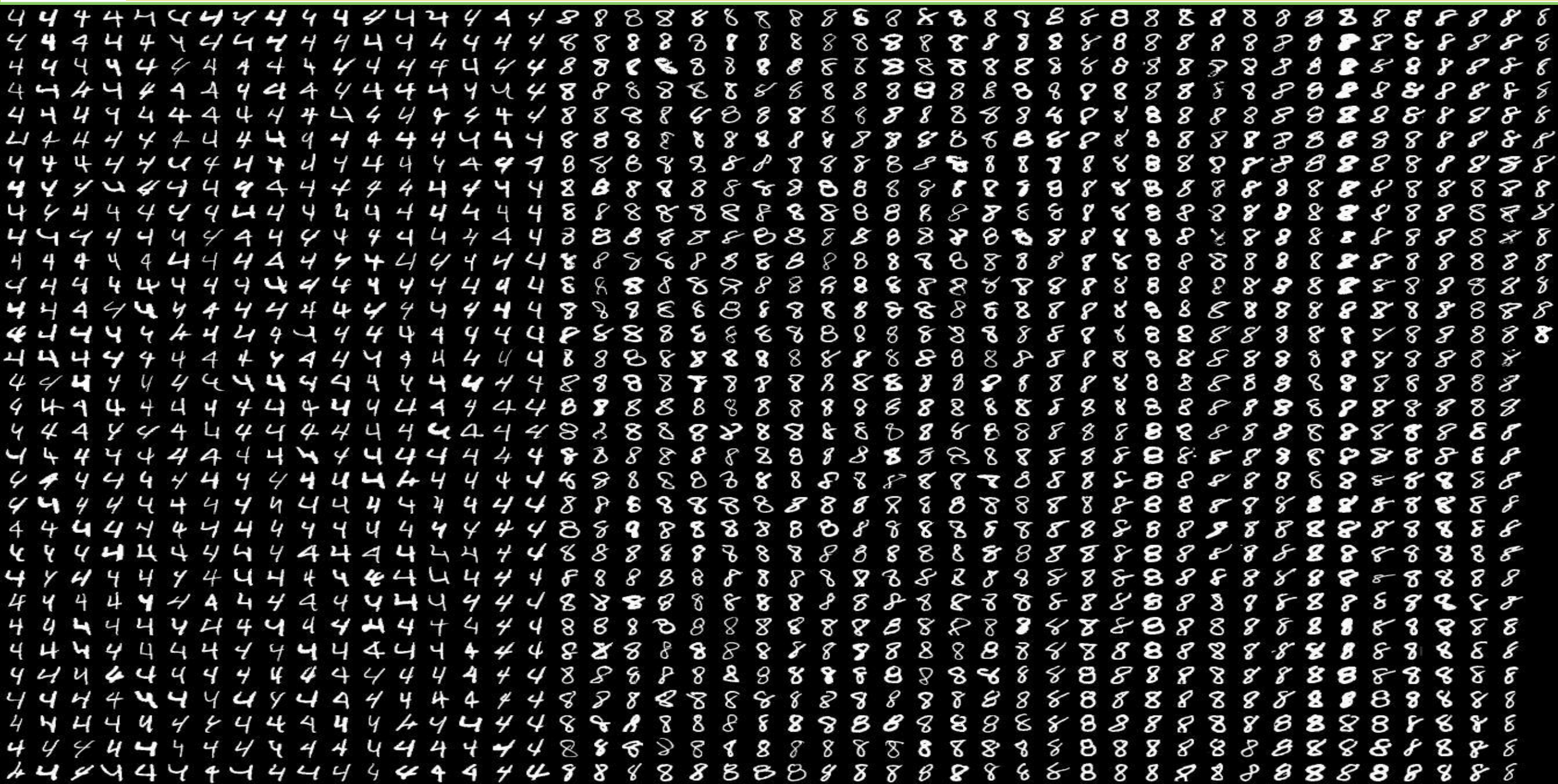
- 人工方法
  - 结果容易理解
    - 足球 and 联赛→体育类
  - 费时费力
  - 难以保证一致性和准确性(40%左右的准确率)
  - 专家有时候凭空想象
  - 知识工程的方法建立专家系统(80年代末期)
- 自动的方法(机器学习)
  - 结果可能不易理解
  - 快速
  - 准确率相对高(准确率可达60%或者更高)
  - 来源于真实文本，可信度高

案例：中文姓名性别自动分类

- 男性：2241
- 女性：1473

男性		女性	
杨学文	蒋卡海	徐贤梅	黎映月
韦柱钦	曹德军	黄霞	陈思思
杨育俭	李溪华	谢秋田	李小琳
李锋	徐东海	甘伟清	林淑萍
陆锡坤	黄国伟	周霞	梁萍
…	…	…	…

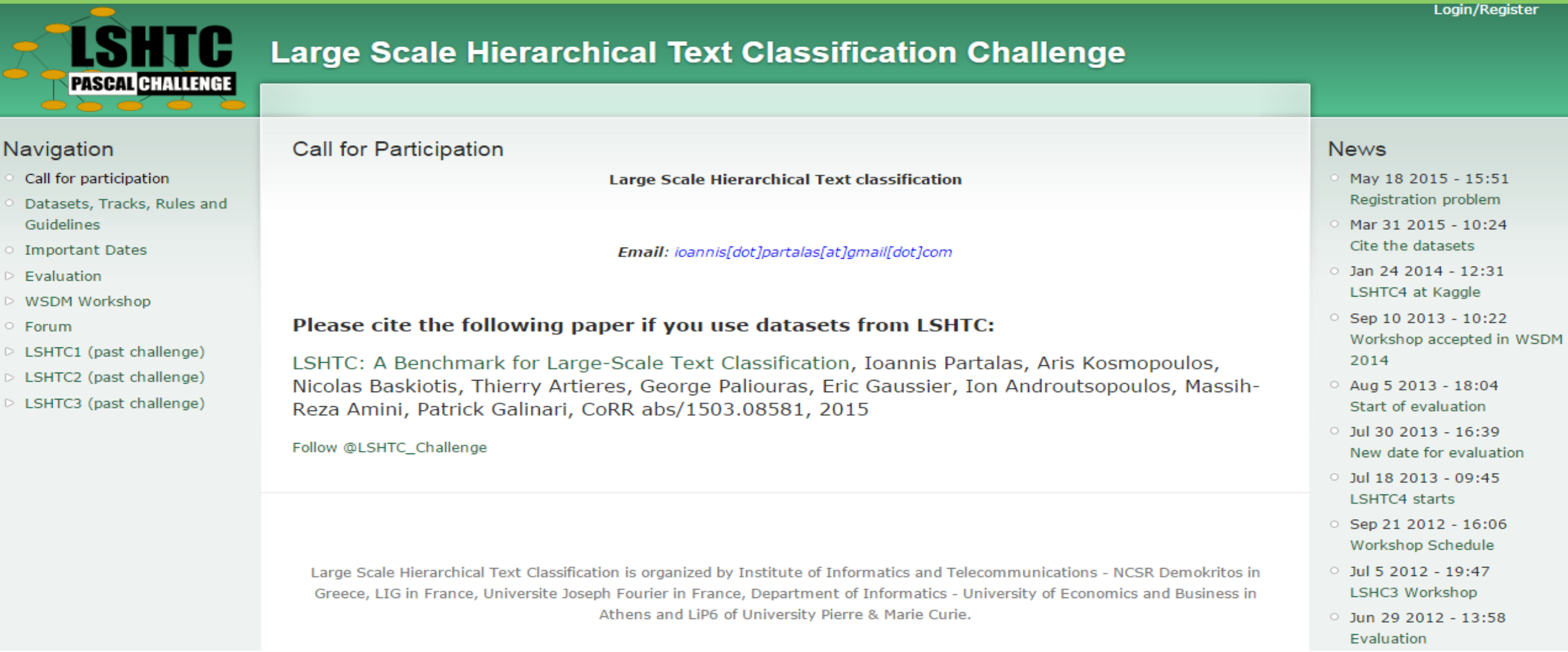
案例：手写数字识别



多标识分类

- 单标识分类问题
  - 类别是互斥的，一个样本不能同时属于多个类别
- 多标识分类问题
  - 图像、视频以及文档等可能同时属于多个类别
  - 基因可能控制多个生物功能

层次文本分类



http://lshtc.iit.demokritos.gr/

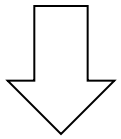
第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

中文分词

我	喜	欢	《	统	计
B	**B**	**I**	**O**	**B**	**I**
机	器	学	习	》	。
B	**I**	**I**	**I**	**O**	**O**

我喜欢《统计机器学习》。



我 喜欢 《 统计 机器学习 》 。

词性标注（POS）

The	postman	collected	letters	and	left	.
DET	**NN**	**VBD**	**NNS**	**CNJ**	**VBD**	**.**

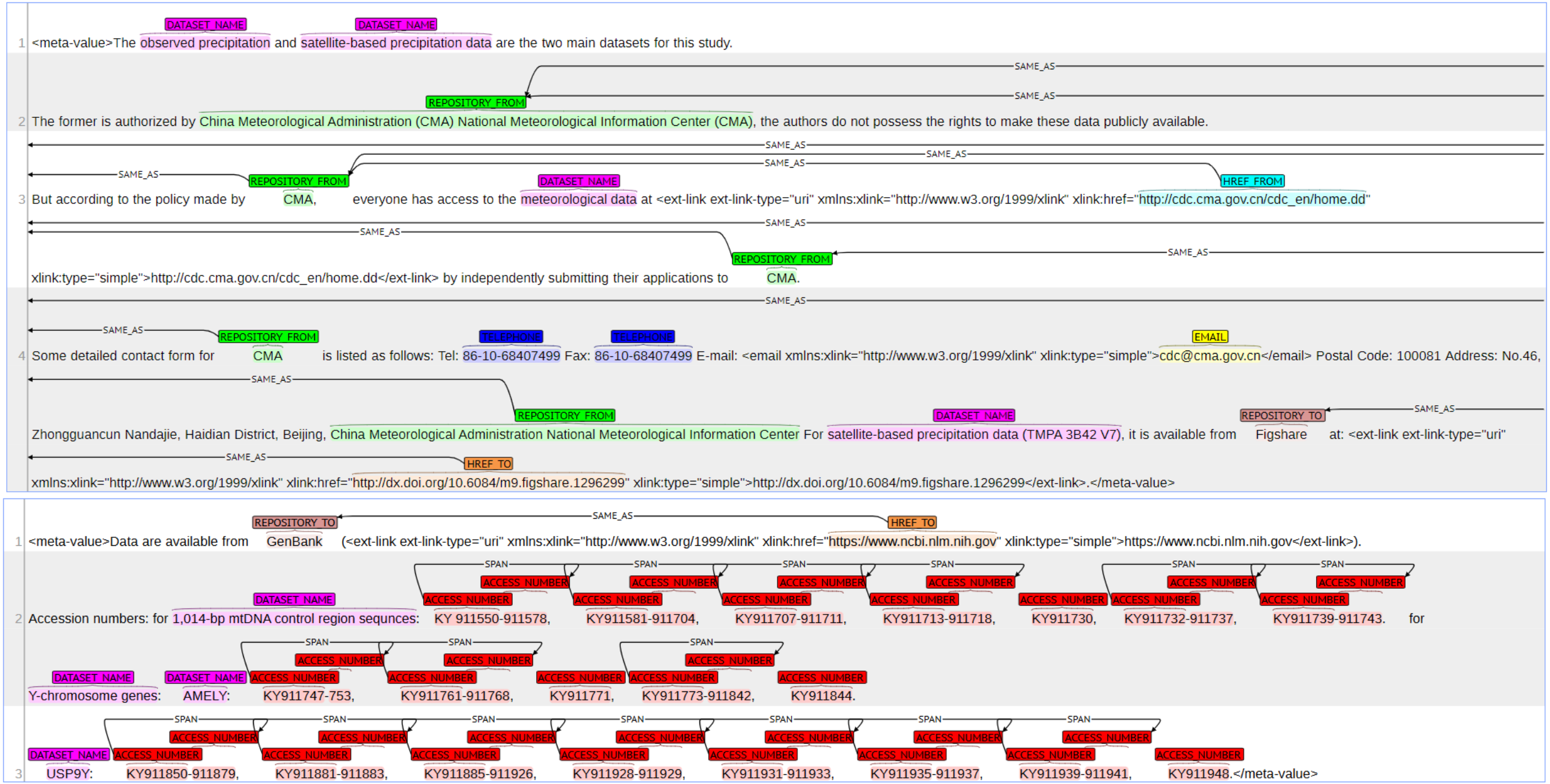
- DET：冠词
- NN：名词单数
- NNS：名词复数
- VBD：动词过去式
- CNJ：连词

命名实体抽取 (NER)

John	Smith	is	the	scientist
B-PER	**I-PER**	**O**	**O**	**O**
of	the	Hardcom	Corp	.
O	**O**	**B-ORG**	**I-ORG**	**O**

2-tag	**B, I**	**B, BI, BII, …**
3-tag	**B, I, O**	**O, BI, BII, …**
	B, M, E	**B, BE, BME, BMME, …**
4-tag	**B, M, E, S**	**S, BE, BME, BMME, …**
5-tag	**B, B_2, M, E, S**	**S, BE, BB_2E, BB_2ME, BB_2MME, …**
6-tag	**B, B_2, B_3, M, E, S**	**S, BE, BB_2E, BB_2 B_3E, BB_2 B_3ME, …**

可视化标注工具：brat



经典模型

- 隐马尔科夫模型 (HMM) (Rabiner 1989; Freitag & McCallum, 2000; Xu, 2007)
- 最大熵模型 (MaxEnt) (Berger, et. al., 1996; Ratnaparkhi, 1997)
- 最大熵马尔科夫模型 (MEMM) (McCallum, Freitag, & Pereira, 2000; Punyakanok & Roth, 2001)
- 条件随机场 (CRF) (Lafferty, McCallum, & Pereira, 2001; Lafferty, Zhu, & Liu, 2004)
- 感知器（Perceptron） (Collins, 2002; Li, Bontcheva, & Cunningham, 2005)
- BiLSTM-CRF (Huang et al., 2020)
- ……

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
- 本章小节

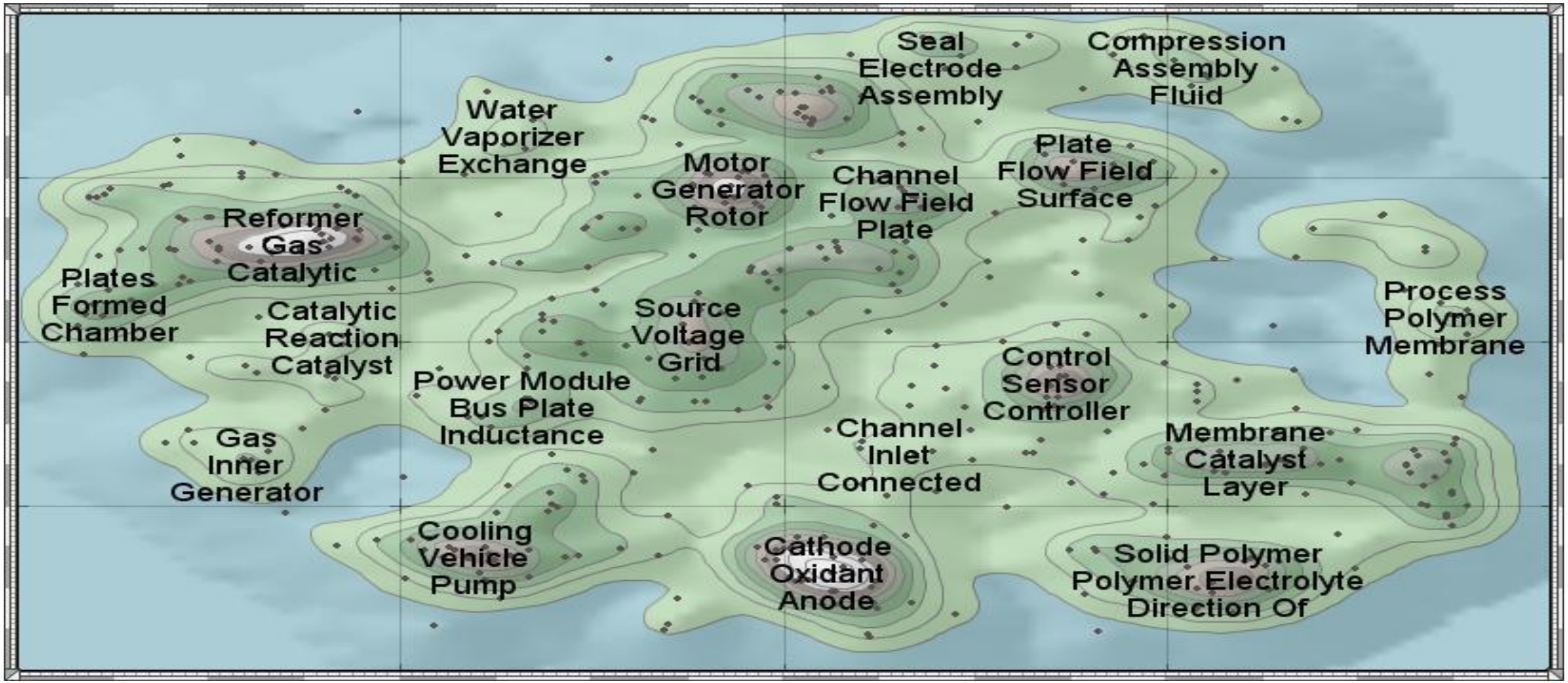
无监督学习

- 在无监督式学习中，数据并不被特别标识，学习模型为了推断出数据的一些内在结构；
- 常见的应用场景包括关联规则的学习以及聚类等；
- 监督学习与无监督学习的区别：训练集目标是否被标注。

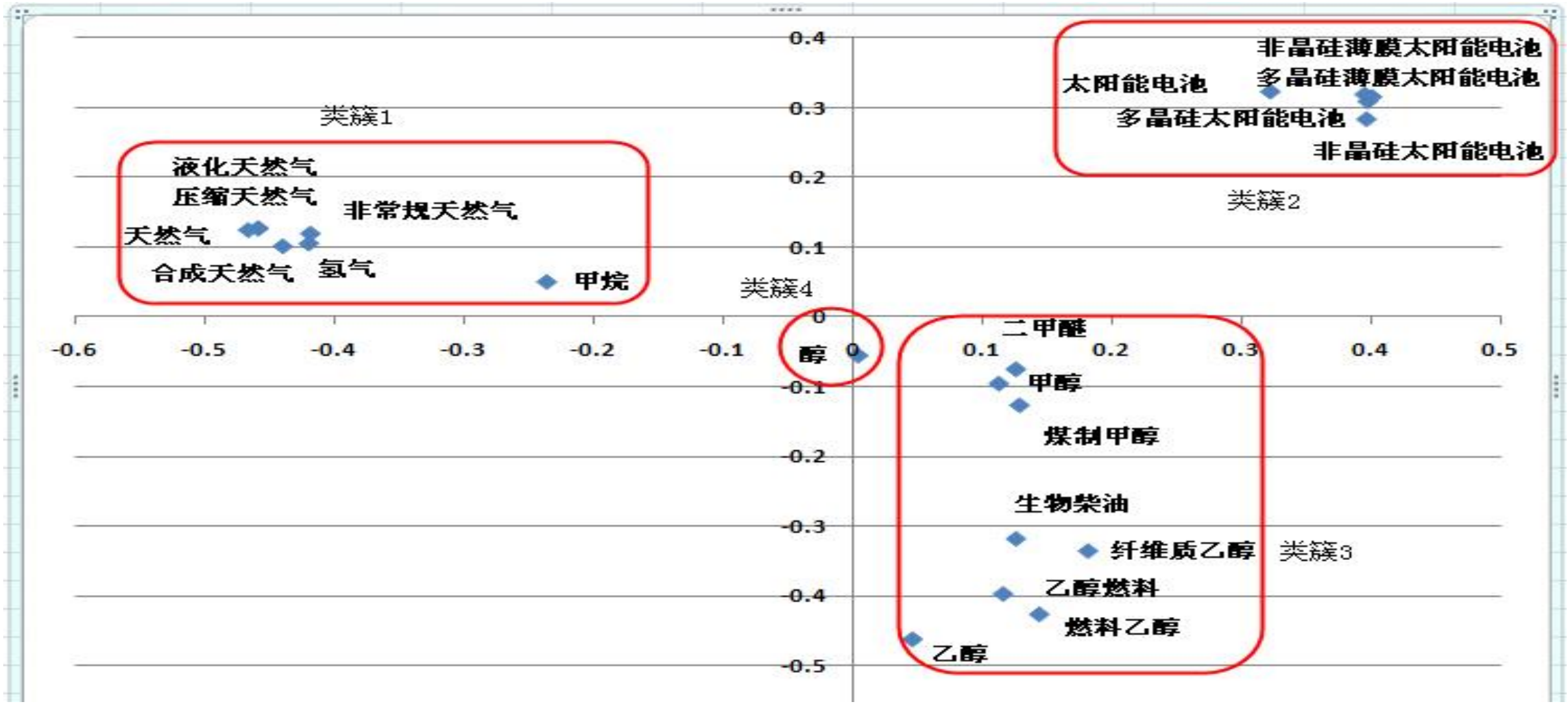
物以类聚，人以群分

- 人类倾向于跟志趣相投的人在一起，也就是“物以类聚，人以群分”
- 人们有足够的智力寻找重复的模式，不断将看到的、听到的、闻到和品尝到的与记忆中已经存在的东西相联系
- 聚类就是将一个给定集合中的相似项分成不同簇的过程
  - 这些簇可以看成一组集合，簇内相似，而簇间有别

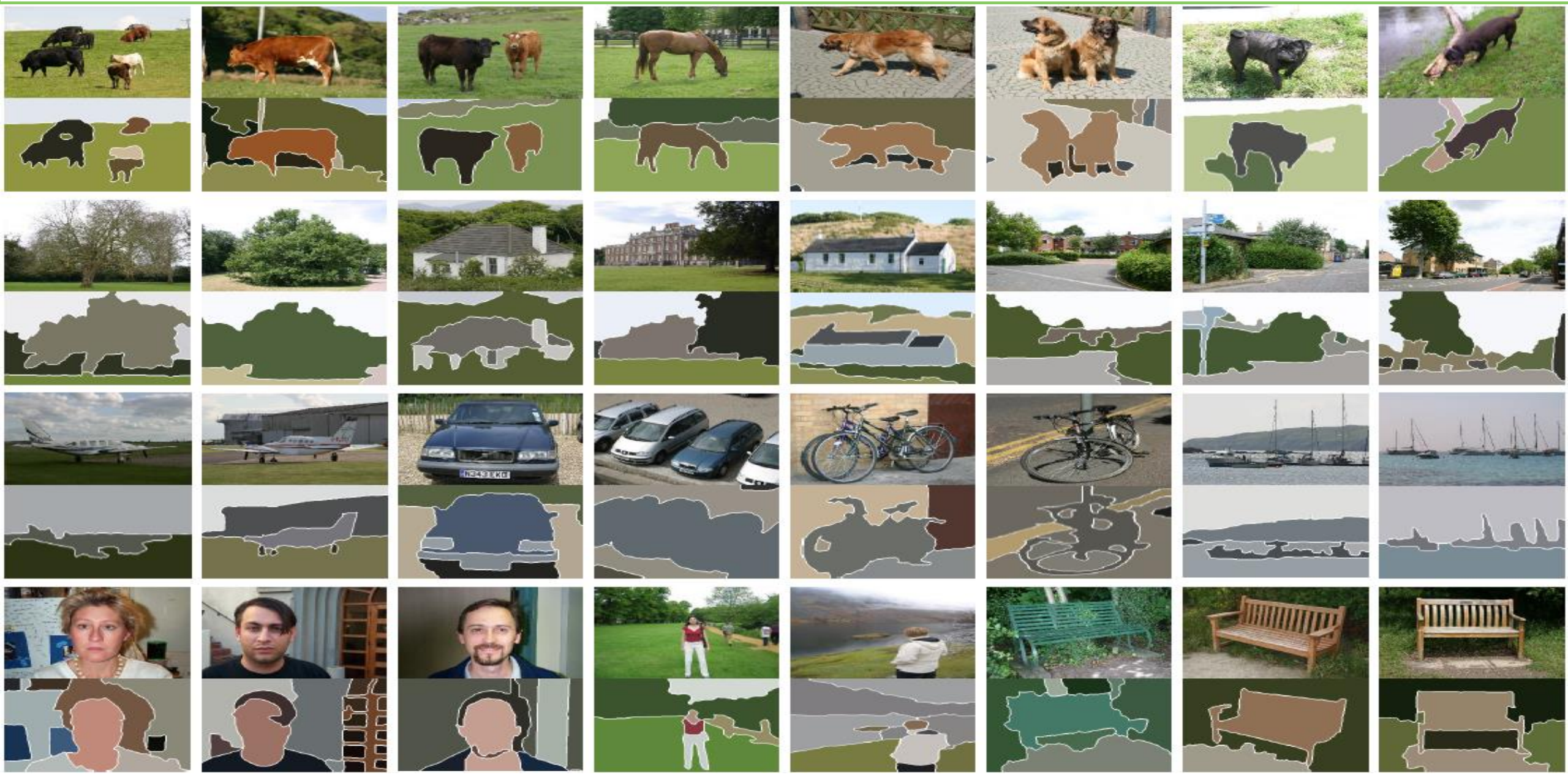
案例：专利地图


Seal
Electrode
Assembly
Compression
Assembly
Fluid
Water
Vaporizer
Exchange
Motor
Generator
Rotor
Channel
Flow Field
Plate
Plate
Flow Field
Surface
Reformer
Gas
Catalytic
Plates
Formed
Chamber
Catalytic
Reaction
Catalyst
Source
Voltage
Grid
Process
Polymer
Membrane
Power Module
Bus Plate
Inductance
Control
Sensor
Controller
Gas
Inner
Generator
Channel
Inlet
Connected
Membrane
Catalyst
Layer
Cooling
Vehicle
Pump
Cathode
Oxidant
Anode
Solid Polymer
Polymer Electrolyte
Direction Of

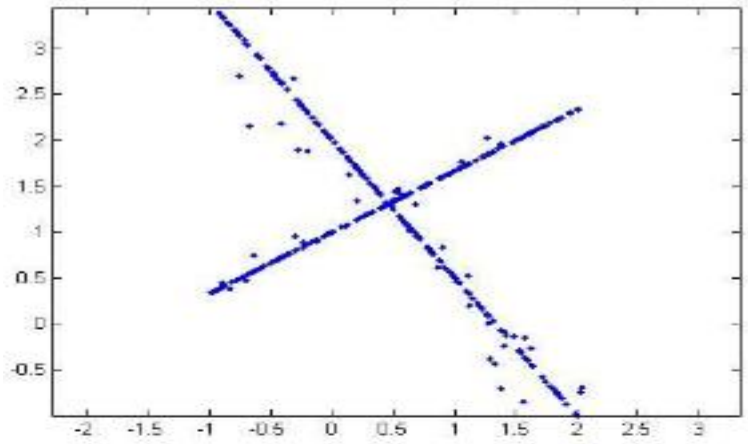
案例：词聚类


0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
非晶硅薄膜太阳能电池
多晶硅薄膜太阳能电池
太阳能电池
多晶硅太阳能电池
非晶硅太阳能电池
类簇2
类簇1
液化天然气
压缩天然气
非常规天然气
天然气
合成天然气
氢气
甲烷
类簇4
醇
二甲醚
甲醇
煤制甲醇
生物柴油
纤维质乙醇
类簇3
乙醇燃料
燃料乙醇
乙醇

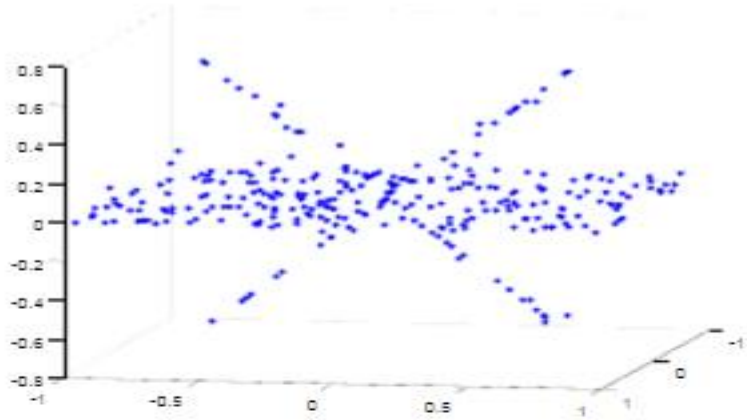
案例：图像分割



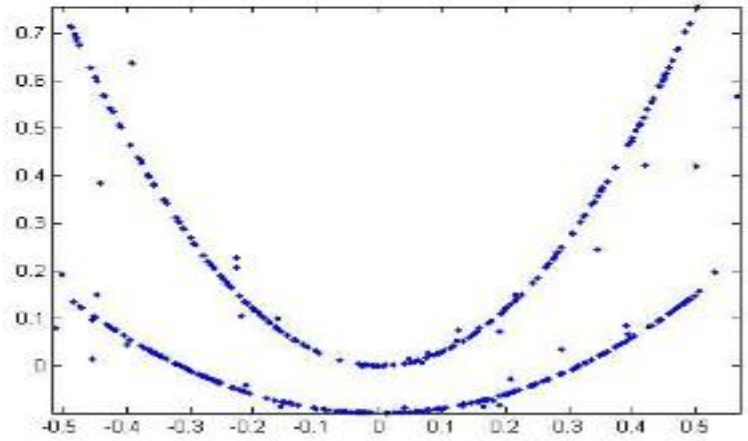
案例：全国研究生数学建模大赛



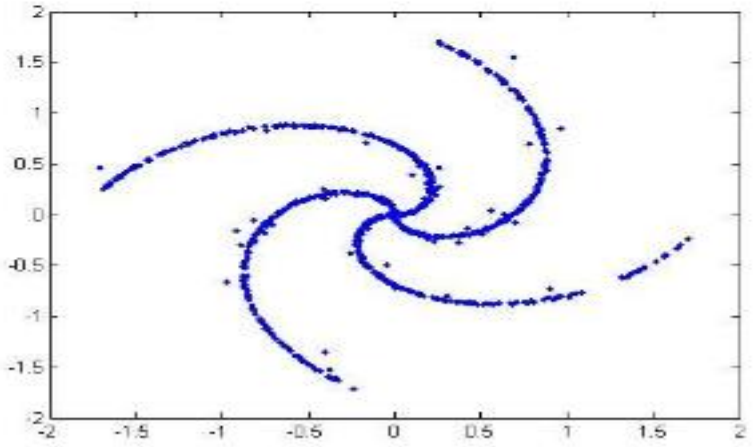
(a)



(b)

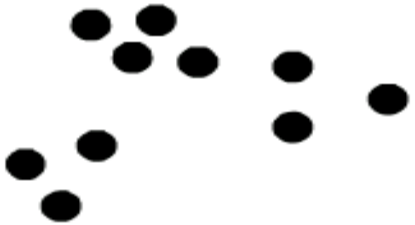


(c)

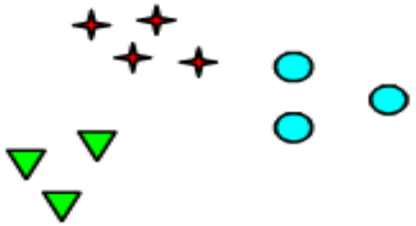


(d)

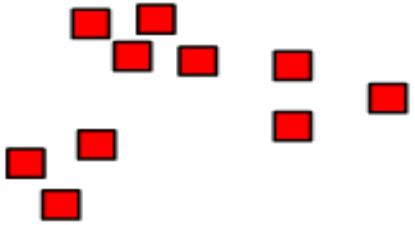
聚类结果的模糊性（1/2）



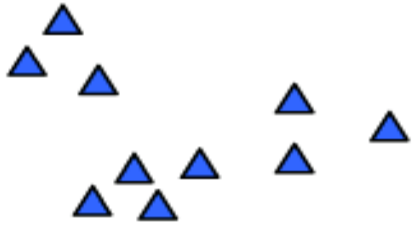
多少个聚类?



六个聚类

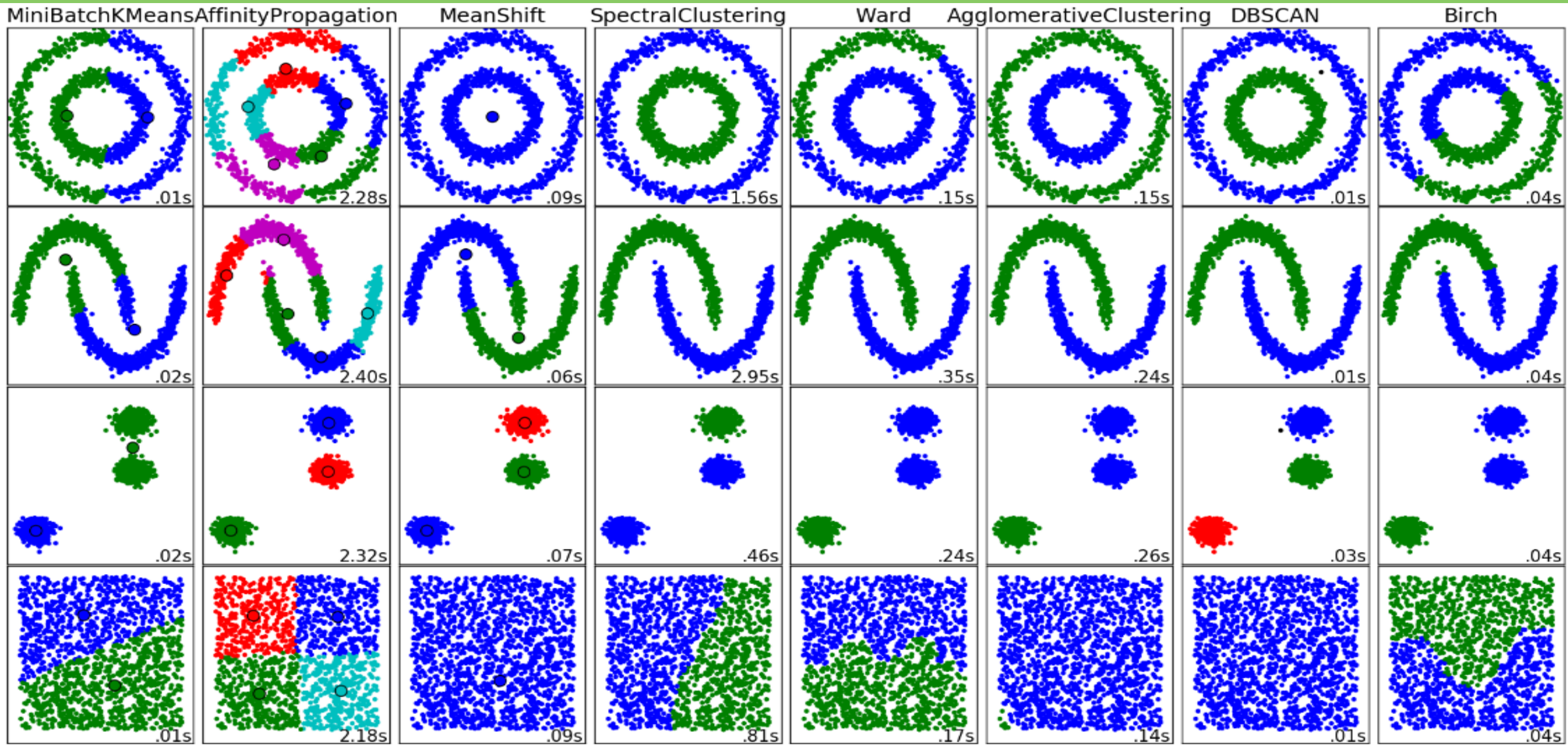


两个聚类



四个聚类

聚类结果的模糊性（2/2）


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.28s
.09s
1.56s
.15s
.15s
.01s
.04s
.02s
2.40s
.06s
2.95s
.35s
.24s
.01s
.04s
.02s
2.32s
.07s
.46s
.24s
.26s
.03s
.04s
.01s
2.18s
.09s
.81s
.17s
.14s
.01s
.04s

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

主题是什么 (1/3)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

法国	全国	教育	产品	卫生
欧洲	人大	学生	生产	下乡
德国	常委会	学校	质量	药
欧盟	人民	教师	企业	医疗
法	乔石	大学	工业	健康
德	委员长	学	技术	药品
巴黎	届	教学	名牌	农村
国	代表大会	高校	服装	医药
希拉克	委员会	大学生	开发	医院
瑞典	审议	学习	国内	保健
主题 1	主题 2	主题 3	主题 4	主题 5

图 1　人民日报语料在 LDA 模型上的训练结果(部分)

◆ David M. Blei, Andrew Y. Ng and Michael I. Jordan, 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, No. Jan, pp. 993-1022.

◆ 徐戈，王厚峰, 2011. 自然语言处理中主题模型的发展. *计算机学报*, Vol. 34, No. 8, pp. 1423-1436.

主题是什么 (2/3)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

◆ David M. Blei, Andrew Y. Ng and Michael I. Jordan, 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, No. Jan, pp. 993-1022.

主题是什么 (3/3)

- Usage of a theme:
  - Summarize topics/subtopics
  - Navigate documents
  - Retrieve documents
  - Segment documents
  - All other tasks involving unigram language models

SCIgen：CS论文自动生成器（1/2）

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the WMSCI 2005 website). There's also a list of known bogus conferences. Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See Examples for more details.

We went to WMSCI 2005. Check out the talks and video. You can find more details in our blog.

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:
Author 2:
Author 3:
Author 4:
Author 5:

Generate Reset

SCIgen currently supports Latin-1 characters, but not the full Unicode character set.

Vis

Xue

Abstract

Compact models and XML ha
interest from both cyberinform
tographers in the last severa
few futurists would disagree
tion of the Ethernet, which em
principles of disjoint algorithr
permutable tool for controllin
converters, which we call Slic
entirely an appropriate aim b
existing work in the field.

1 Introduction

Many end-users would agree
been for active networks, the
gestion control might never h
notion that leading analyst
"fuzzy" communication is re
The notion that physicists col
programming is rarely encou
metamorphic models and relia
operate in order to achieve th
courseware.

In this paper, we use encryp
demonstrate that the well-kn
algorithm for the investigatio
Zhao et al. [15] follows a
tion. In the opinions of m
conventional wisdom states t
is mostly addressed by the e

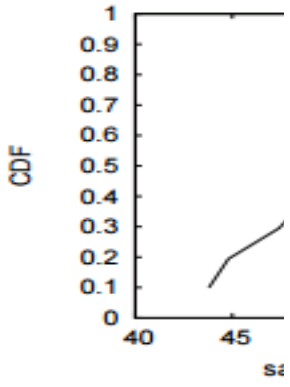


Figure 4: The expe
tic, as a function of

simulated instant
pared results to
(4) we measured
tion of RAM spa
We discarded the
ments, notably wl
database through
network.

Now for the cli
half of our experi
ities in the graph
introduced with
ond, the curve in
it is better known
Figure 5 should lo
as $f_*(n) = \log \log$

We have seen on
and 2; our other ex
paint a different
netic disturbances
experimental resul
ing the feedback l
method's USB key
otherwise. It mig

Bhabha fails to address several key issues that our approach does fix. In general, Slicer outperformed all existing heuristics in this area [9, 3, 1].

Several psychoacoustic and linear-time systems have been proposed in the literature [5]. The much-touted framework by Sato et al. does not create SMPs as well as our method. This is arguably ill-conceived. Brown et al. originally articulated the need for autonomous algorithms [14]. Noam Chomsky [13] suggested a scheme for visualizing robust algorithms, but did not fully realize the implications of XML at the time.

Our framework builds on related work in client-server information and complexity theory. Unlike many existing approaches, we do not attempt to learn or study authenticated archetypes. The choice of the Internet in [7] differs from ours in that we evaluate only technical methodologies in our algorithm [10]. Our solution to DHCP differs from that of Davis [11] as well.

6 Conclusion

In this work we showed that the acclaimed stochastic algorithm for the simulation of e-business by Thompson is NP-complete. One potentially tremendous shortcoming of our framework is that it can store extreme programming; we plan to address this in future work. We concentrated our efforts on demonstrating that the little-known linear-time algorithm for the construction of model checking by Williams [8] is recursively enumerable. On a similar note, our application can successfully prevent many information retrieval systems at once. We plan to explore more issues related to these issues in future work.

References

[1] Bachman, C. The influence of metamorphic theory on e-voting technology. *Journal of Distributed Models 4* (Apr. 2003), 57–61.

[2] Backus, J. The effect of cooperative communication on fuzzy e-voting technology. *Journal of Ubiquitous Modalities 55* (May 1994), 75–95.

[3] Bhabha, N., and Rangarajan, B. Decoupling von Neumann machines from congestion control in compilers. In *Proceedings of the Symposium on Pseudorandom Information* (Apr. 2003).

[4] Engelbart, D., Stearns, R., and Jacobson, V. Studying DHCP using pseudorandom modalities. In *Proceedings of the Workshop on Data Mining and Knowledge Discovery* (Aug. 2002).

[5] Floyd, R. The effect of constant-time technology on cryptoanalysis. In *Proceedings of the Symposium on Semantic, Semantic Algorithms* (Sept. 2001).

[6] Kumar, J., Milner, R., Daubechies, I., Tarjan, R., and Watanabe, M. Interactive, symbiotic symmetries for IPv6. *Journal of Pseudorandom, Real-Time Theory 79* (Mar. 2002), 43–53.

[7] Lakshminarayanan, K. Contrasting von Neumann machines and B-Trees. *OSR 7* (Feb. 1992), 86–107.

[8] Lamport, L., Agarwal, R., Feigenbaum, E., and Shamir, A. Synthesizing the location-identity split and suffix trees with *holypuet*. In *Proceedings of OOPSLA* (Feb. 1996).

[9] Li, B. O. On the construction of sensor networks. In *Proceedings of the Symposium on Interposable Symmetries* (July 2003).

[10] Li, I., and Zhou, H. Development of simulated annealing. In *Proceedings of SOSP* (Feb. 2005).

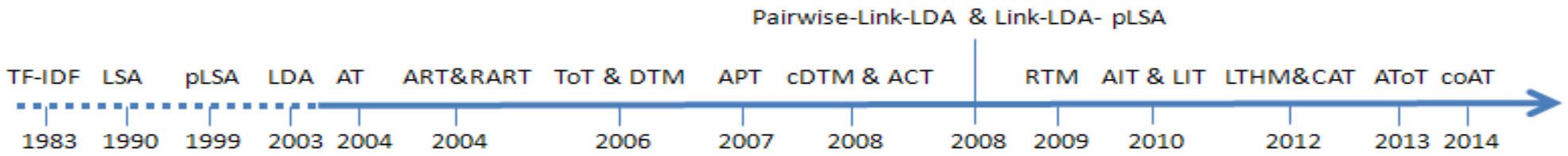
[11] Martinez, J., Martinez, C., Engelbart, D., Hennessy, J., Chomsky, N., and Garcia-Molina, H. Flip-flop gates considered harmful. *Journal of Heterogeneous Algorithms 23* (Aug. 2003), 55–64.

[12] Shenker, S. On the emulation of scatter/gather I/O. In *Proceedings of MICRO* (Oct. 2003).

[13] Stearns, R., and Newell, A. The impact of certifiable models on algorithms. *Journal of Embedded, Large-Scale Communication 7* (Jan. 2003), 57–65.

主题模型：实例

- LDA模型: Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003; Blei, Ng & Jordan, 2002; Mochihashi, 2004)
- AT模型: Author-Topic Model (Rosen-Zvi, et. al., 2004; Steyvers, et. al., 2004; Rosen-Zvi, et. Al., 2010)
- ACT模型: Author-Conference-Topic Model (Tang, et. al., 2008; Tang, et. al., 2008; Tang, et. al., 2010)
- AToT模型: Author-Topic over Time Model (Shi, et. al., 2013, Xu, et. al., 2014a, 2014b)
- coAT模型: coauthor Topic Model (An, et. al, 2014)
- …



◆ 张晗，徐硕，乔晓东, 2015. 融合科技文献内外部特征的主题模型发展综述. *情报学报*, Vol. 33, No. 10, pp. 1108-1120.

用户兴趣演化模型

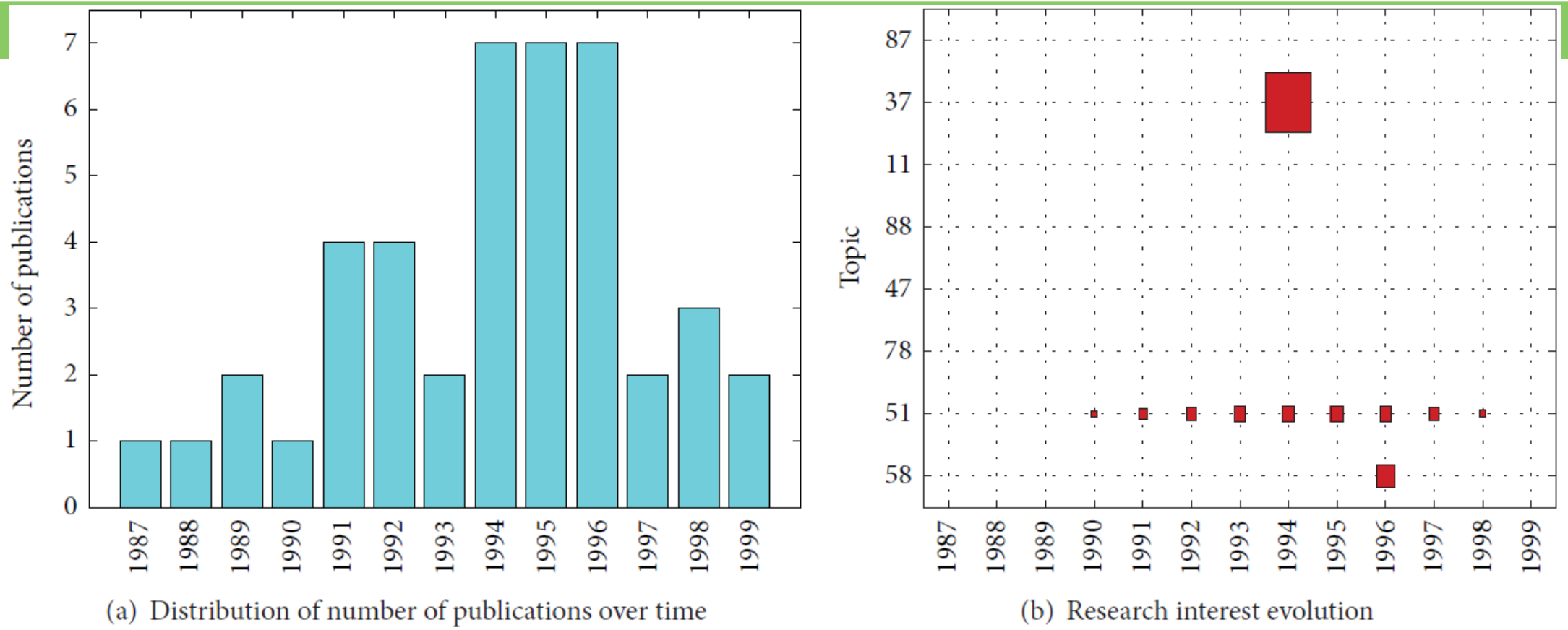


FIGURE 6: The distribution of number of publications and research interest evolution for Sejnowski_T.

◆ 史庆伟，乔晓东，徐硕，农国武，2013. 作者主题演化模型及其在研究兴趣演化分析中的应用. *情报学报*，Vol. 32, No. 9, pp. 912-919.

◆ Shuo Xu, Qingwei Shi, Xiaodong Qiao, Lijun Zhu, Han Zhang, Hanmin Jung, Seungwoo Lee, & Sung-Pil Choi, 2014. A Dynamic users' Interest Discovery Model with Distributed Inference Algorithm. *International Journal of Distributed Sensor Networks*, Vol. 2014, pp. 1-11.

主题N元语法模型（1/2）

Multi-task least-squares support vector machines

Shuo Xu · Xin An · Xiaodong Qiao · Lijun Zhu

Published online: 30 May 2013

Abstract There are often the underlying cross relatedness amongst multiple tasks, which is discarded directly by traditional single-task learning methods. Since multi-task learning can exploit these relatedness to further improve the performance, it has attracted extensive attention in many domains including multimedia. It has been shown through a meticulous empirical study that the generalization performance of Least-Squares Support Vector Machine (LS-SVM) is comparable to that of SVM. In order to generalize LS-SVM from single-task to multi-task learning, inspired by the regularized multi-task learning (RMTL), this study proposes a novel multi-task learning approach, multi-task LS-SVM (MTLS-SVM). Similar to LS-SVM, one only solves a convex linear system in the training phrase, too. What's more, we unify the classification and regression problems in an efficient training algorithm, which effectively employs the Krylow methods. Finally, experimental results on *school* and *dermatology* validate the effectiveness of the proposed approach.

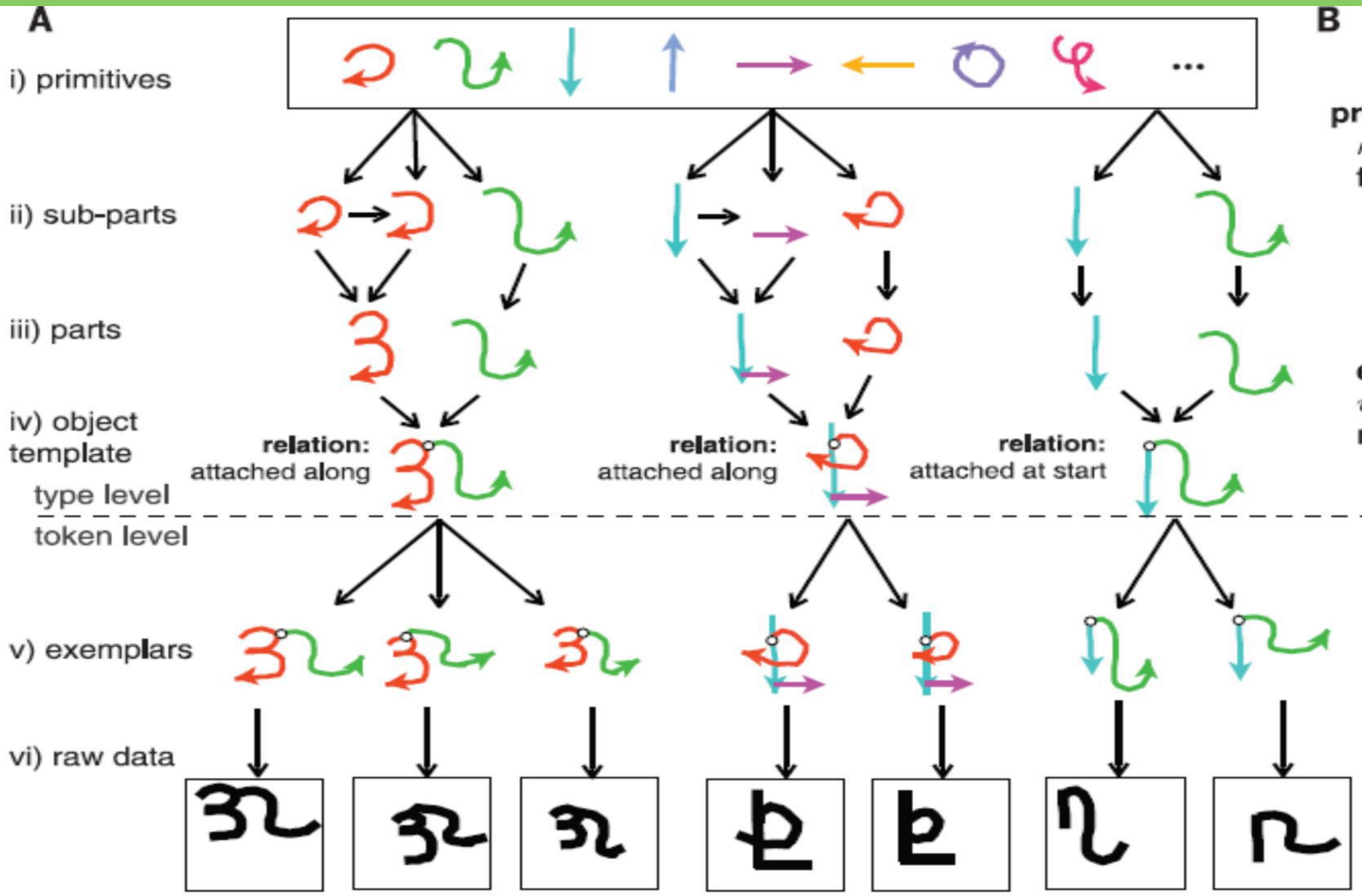
Keywords Multi-task learning · Least-Square Support Vector Machine (LS-SVM) · Multi-Task LS-SVM (MTLS-SVM) · Krylow methods

主题N元语法模型（2/2）

#2	#5	#7	#8	#9
synthetic biology	gene expression	synthetic biology	synthetic biology	polymerase chain reaction
life cycle	result suggest	dna synthesis	biological system	synthetic oligonucleotide
direct evolution	cell growth	homologous recombination	gene network	synthetic dna
shed light	gene regulation	gene synthesis	synthetic gene network	nucleic acid
developmental bias	synthetic promoter	de novo	genetic network	detection limit
synthetic genome	expression level	escherichia coli	system biology	time pcr
generative bias	signaling pathway	gene assembly	synthetic gene	situ hybridization
sexual reproduction	cell proliferation	dna sequencing	gene expression	result show
gene sequence	synthetic oligonucleotide	chemical synthesis	experimental datum	dna microarray
dna array	dependent manner	dna assembly	mathematical model	single nucleotide polymorphism
synthetic association	stress response	high efficiency	result show	high sensitivity
mutation operator	fluorescence microscopy	error rate	gene regulatory network	quantum dot
genetic structure	promoter activity	high yield	large number	flow cytometry
wide association	dna damage	genome engineering	petri net	molecular beacon
synthetic alphoid	dna binding	error correction	biochemical network	dna sample
#11	#12	#14	#21	#23
synthetic biology	genetic interaction	crystal structure	escherichia coli	synthetic biology
building block	saccharomyces cerevisiae	high affinity	synthetic biology	genetic circuit
biological part	synthetic genetic array	binding affinity	metabolic engineering	gene expression
genetic part	synthetic genetic interaction	synthetic biology	metabolic pathway	logic gate
genetic circuit	synthetic genetic	conformational change	carbon source	escherichia coli
dna nanostructure	protein interaction	de novo	gene cluster	biological system
wide range	dna damage	minor groove	lactic acid	cell communication
escherichia coli	synthetic lethal	binding specificity	secondary metabolite	gene circuit
large scale	dna replication	arabinogalactan polysaccharide	growth rate	synthetic gene circuit
synthetic system	gene function	dna sequence	cell growth	quorum sensing
dna assembly	synthetic biology	helix bundle	metabolic network	positive feedback
biological circuit	bud yeast	dna interaction	bacillus subtili	synthetic gene network
biological organism	synthetic lethality	dna aptamer	bacillus subtilis	synthetic circuit
design process	result suggest	im polyamide	high yield	mathematical model
design strategy	biological process	synthetic dna	pseudomonas aeruginosa	quorum sense

◆ Shuo Xu, Liyuan Hao, Guancan Yang, Kun Lu, and Xin An, 2021. A Topic Models based Framework for Detecting and Forecasting Emerging Technologies. *Technology Forecasting and Social Change*, Vol. 162, pp. 120366.

Bayesian学习（1/2）



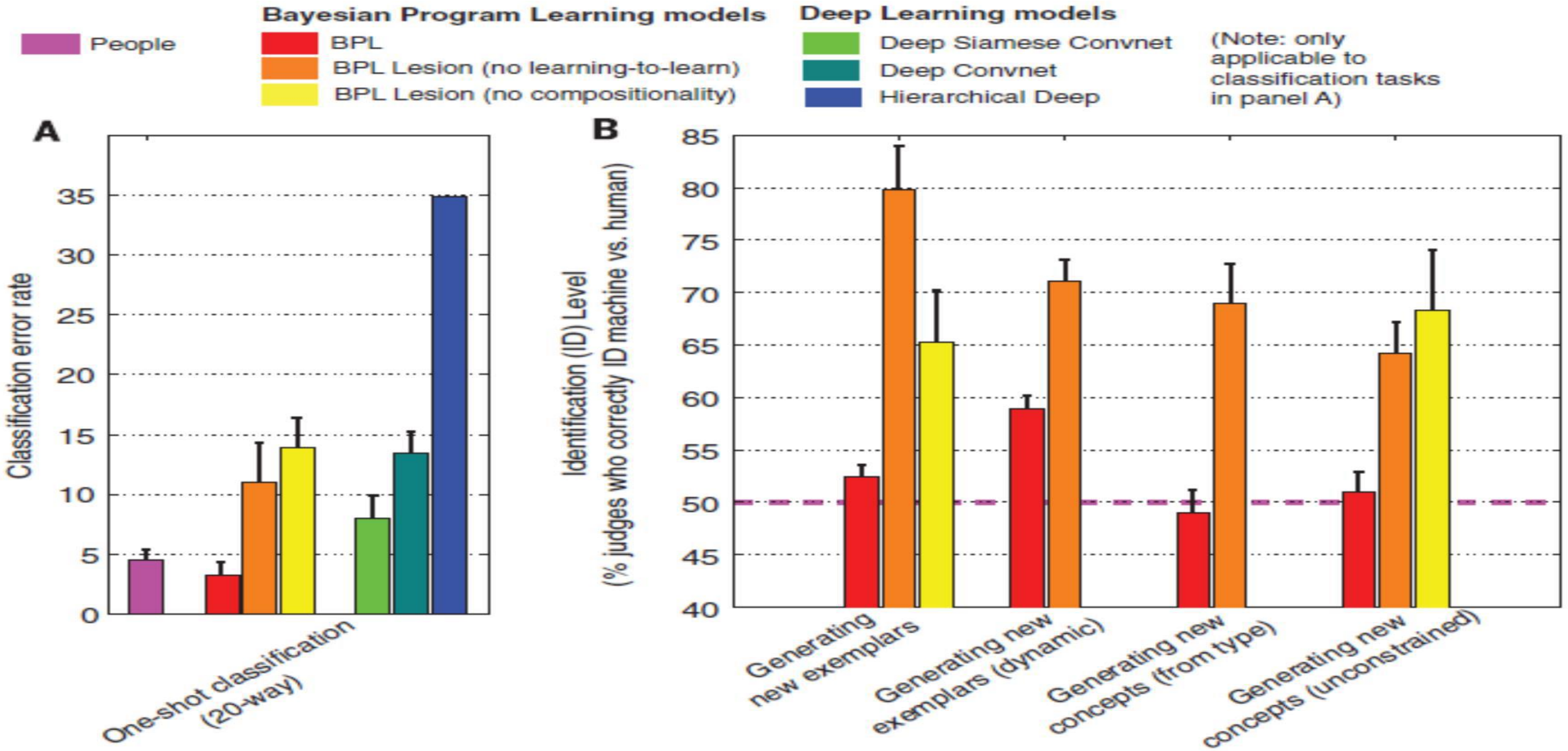
B

procedure GENERATETYPE
- $\kappa \leftarrow P(\kappa)$ ▷ Sample number of parts
- **for** $i = 1 \ldots \kappa$ **do**
  - $n_i \leftarrow P(n_i|\kappa)$ ▷ Sample number of sub-parts
  - **for** $j = 1 \ldots n_i$ **do**
    - $s_{ij} \leftarrow P(s_{ij}|s_{i(j-1)})$ ▷ Sample sub-part sequence
  - **end for**
  - $R_i \leftarrow P(R_i|S_1, \ldots, S_{i-1})$ ▷ Sample relation
- **end for**
- $\psi \leftarrow \{\kappa, R, S\}$
- **return** @GENERATETOKEN(ψ) ▷ Return program

procedure GENERATETOKEN(ψ)
- **for** $i = 1 \ldots \kappa$ **do**
  - $S_i^{(m)} \leftarrow P(S_i^{(m)}|S_i)$ ▷ Add motor variance
  - $L_i^{(m)} \leftarrow P(L_i^{(m)}|R_i, T_1^{(m)}, \ldots, T_{i-1}^{(m)})$ ▷ Sample part's start location
  - $T_i^{(m)} \leftarrow f(L_i^{(m)}, S_i^{(m)})$ ▷ Compose a part's trajectory
- **end for**
- $A^{(m)} \leftarrow P(A^{(m)})$ ▷ Sample affine transform
- $I^{(m)} \leftarrow P(I^{(m)}|T^{(m)}, A^{(m)})$ ▷ Sample image
- **return** $I^{(m)}$

◆Brenden M. Lake, Ruslan Salakhutdinov and Joshua B. Tenenbaum, 2015. Human-Level Concept Learning through Probabilistic Program Induction. *Science*, Vol. 350, No. 6266, pp. 1332-1338.

Bayesian学习（2/2）

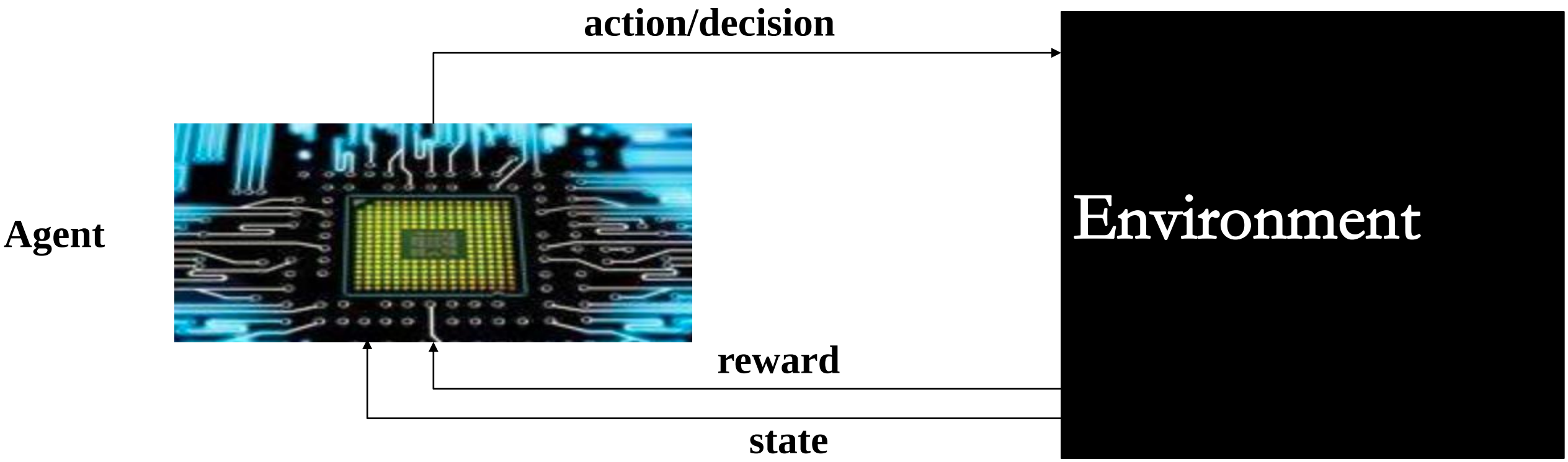


第一章：绪论

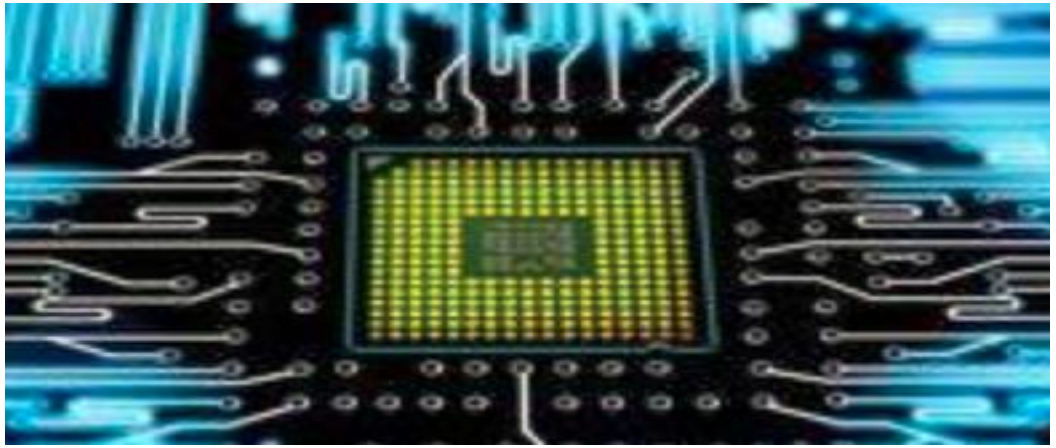
- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

如何训练狗?

什么是强化学习

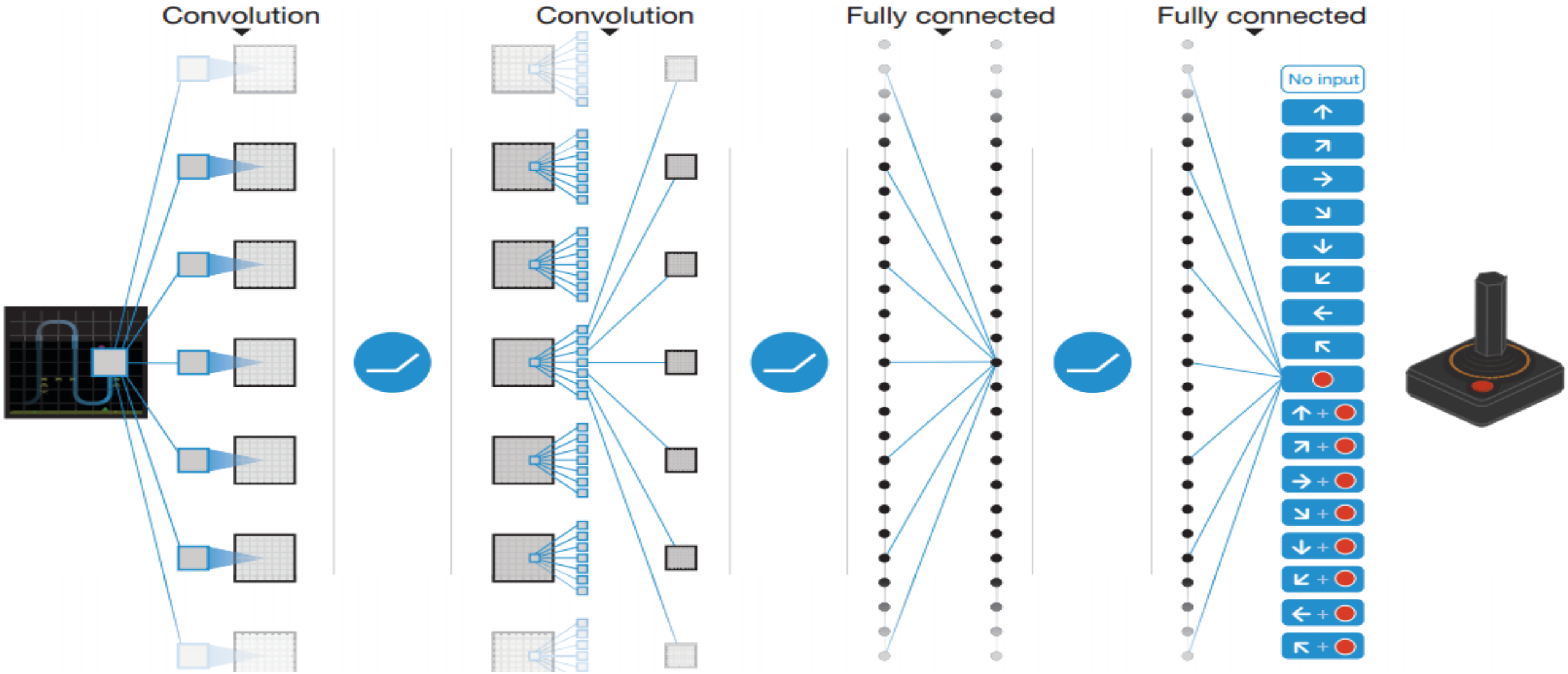


Atari Breakout游戏（1/2）

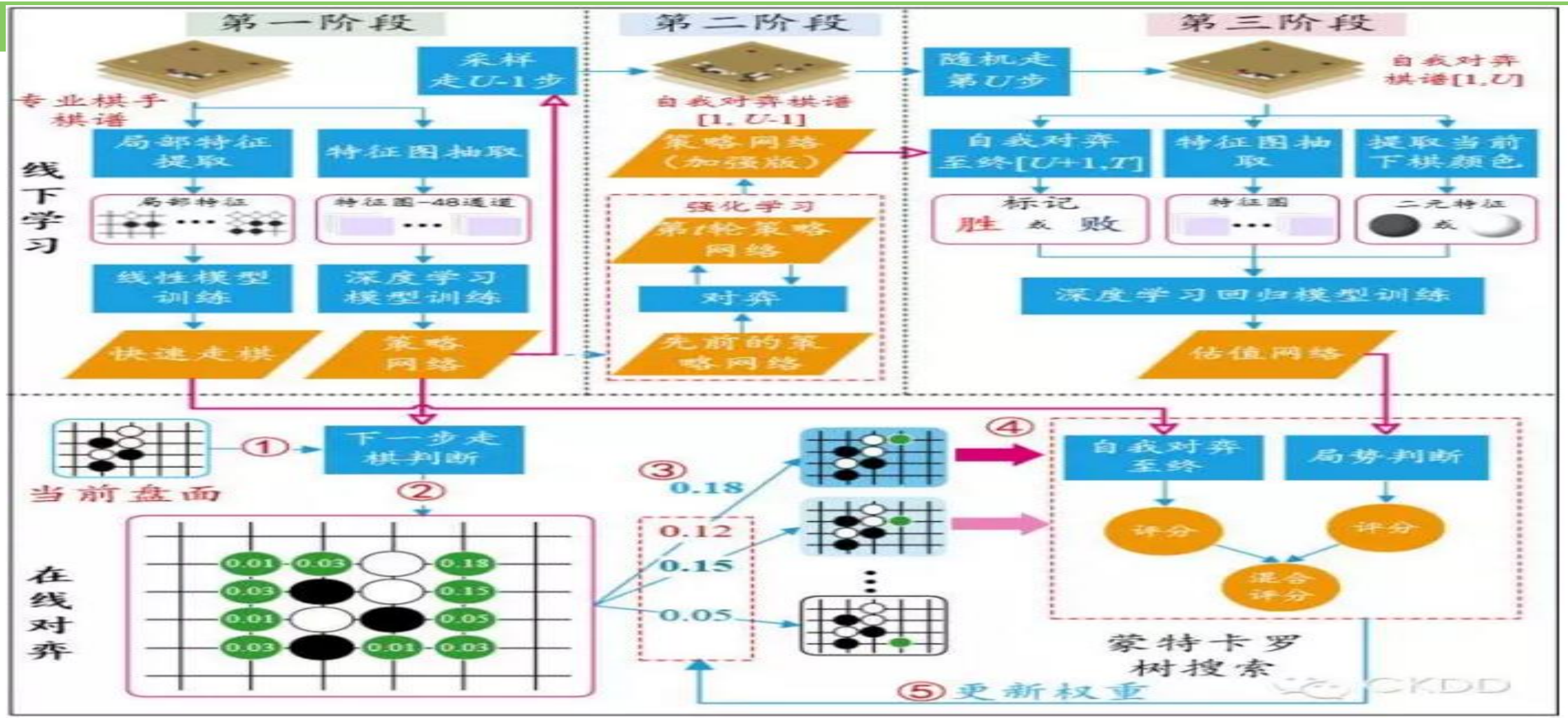


◆Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al., 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*, Vol. 518, No. 7540, pp. 529-533.

Atari Breakout游戏（2/2）



AlphaGo



◆David Silver, Aja Huang, Chris J. Maddison, et al., 2016. Mastering the game of Go with Deep Neural Networks and Tree Search. *Nature*, Vol. 529, No. 7587, pp. 484-489.

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
- 本章小节

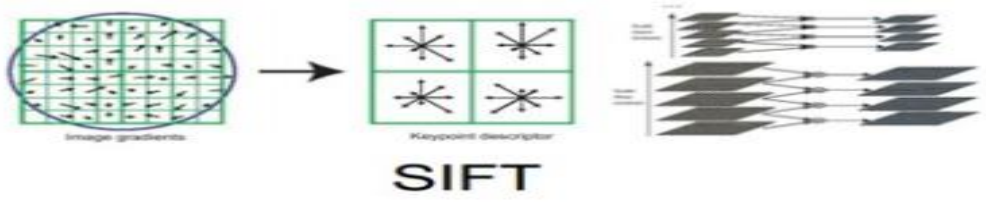
动机（1/2）

传统的模式识别方法：

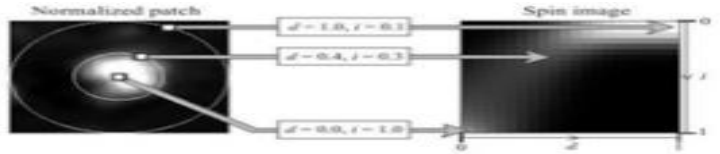
- 良好的特征表达，对最终算法的准确性起了非常关键的作用
- 识别系统主要的计算和测试工作耗时主要集中在特征提取部分
- 特征的样式目前一般都是人工设计的，靠人工提取特征

动机（2/2）

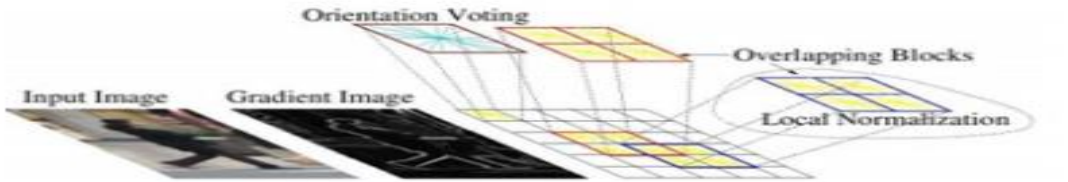
- 机器学习中，获得好的特征是识别成功的关键
- 目前存在大量人工设计的特征，不同研究对象特征不同，特征具有多样性，如：SIFT, HOG, LBP等
- 手工选取特征费时费力，需要启发式专业知识，很大程度上靠经验和运气



SIFT

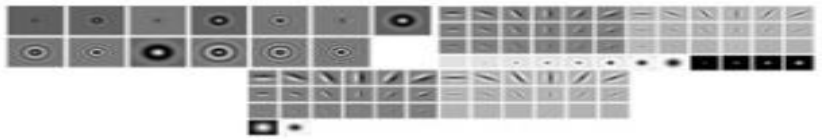


Spin image

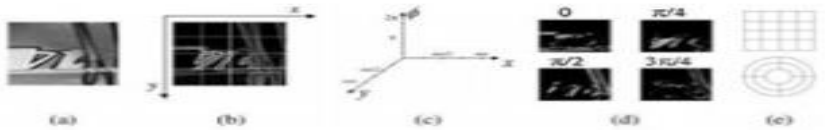


HoG

RIFT

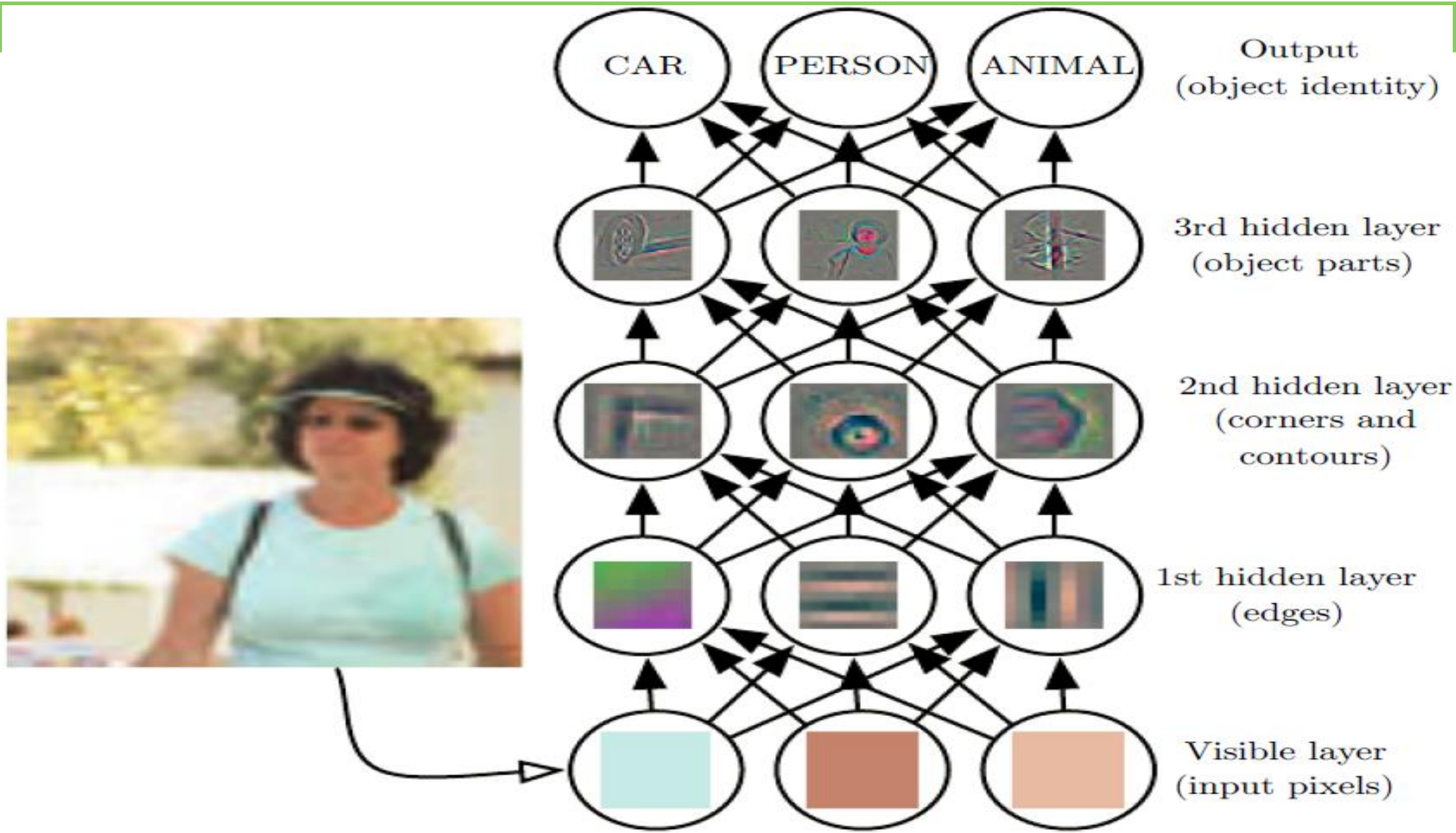


Textons

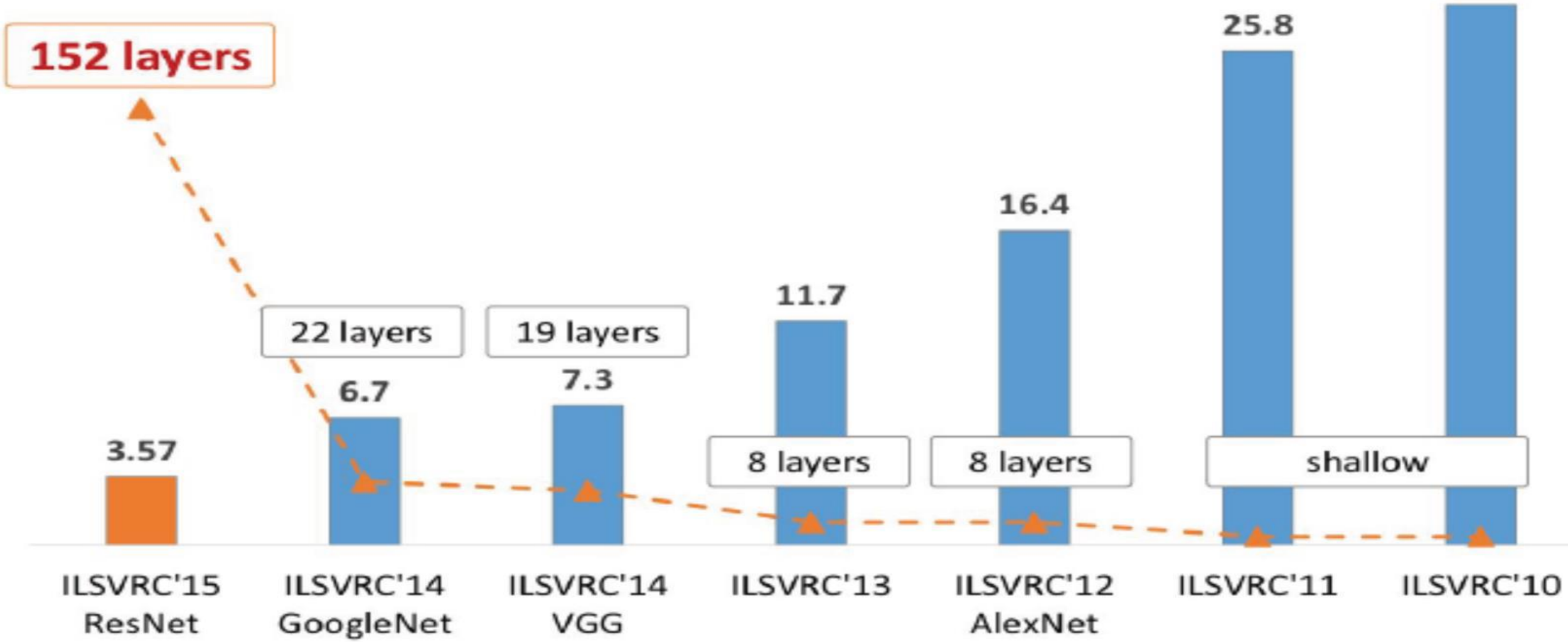


GLOH

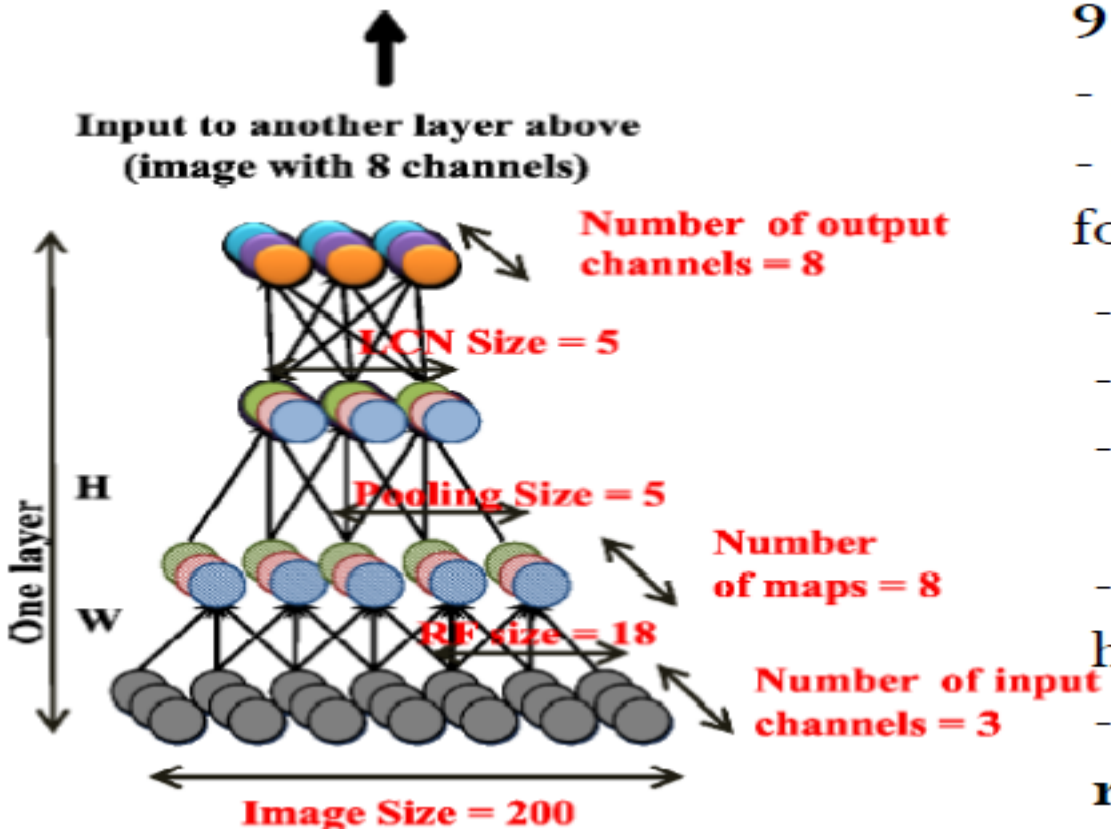
图像识别（1/2）


CAR
PERSON
ANIMAL
Output
(object identity)
3rd hidden layer
(object parts)
2nd hidden layer
(corners and
contours)
1st hidden layer
(edges)
Visible layer
(input pixels)

图像识别（2/2）


152 layers
22 layers
19 layers
8 layers
8 layers
shallow
3.57
6.7
7.3
11.7
16.4
25.8
ILSVRC'15
ResNet
ILSVRC'14
GoogleNet
ILSVRC'14
VGG
ILSVRC'13
ILSVRC'12
AlexNet
ILSVRC'11
ILSVRC'10

Big Model + Big Data + Big/Super Cluster



9 layers sparse autoencoder with:
- local receptive fields to scale up;
- local L2 pooling and local contrast normalization for invariant features
  - **1B parameters** (connections)
  - **10M** 200x200 images
  - train with **1K machines (16K cores)** for **3 days**

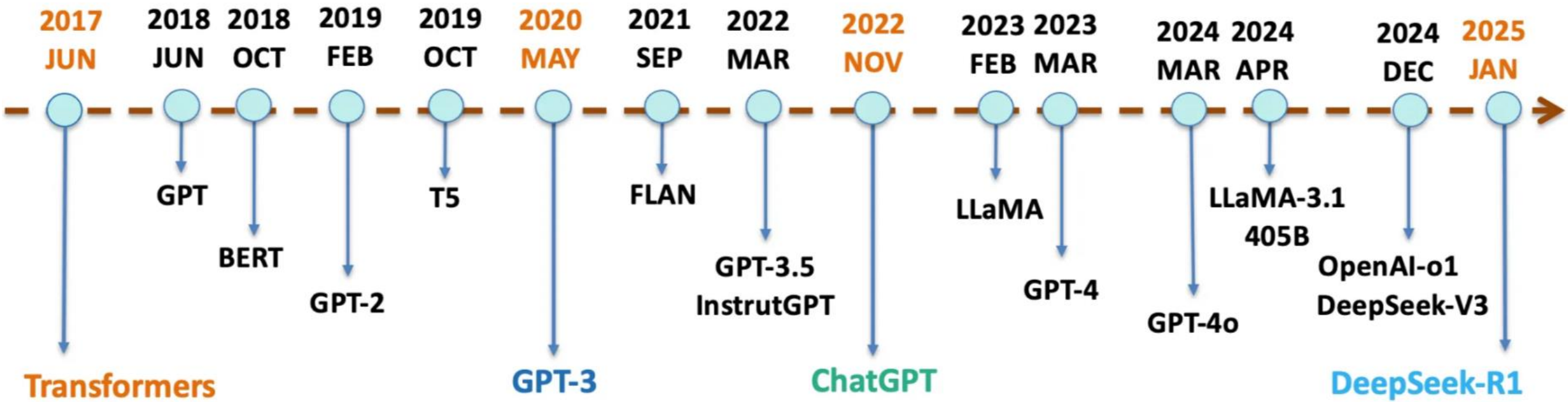
-able to build high-level concepts, e.g., cat faces and human bodies

-**15.8% accuracy** in recognizing 22K objects (**70% relative improvements**)

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

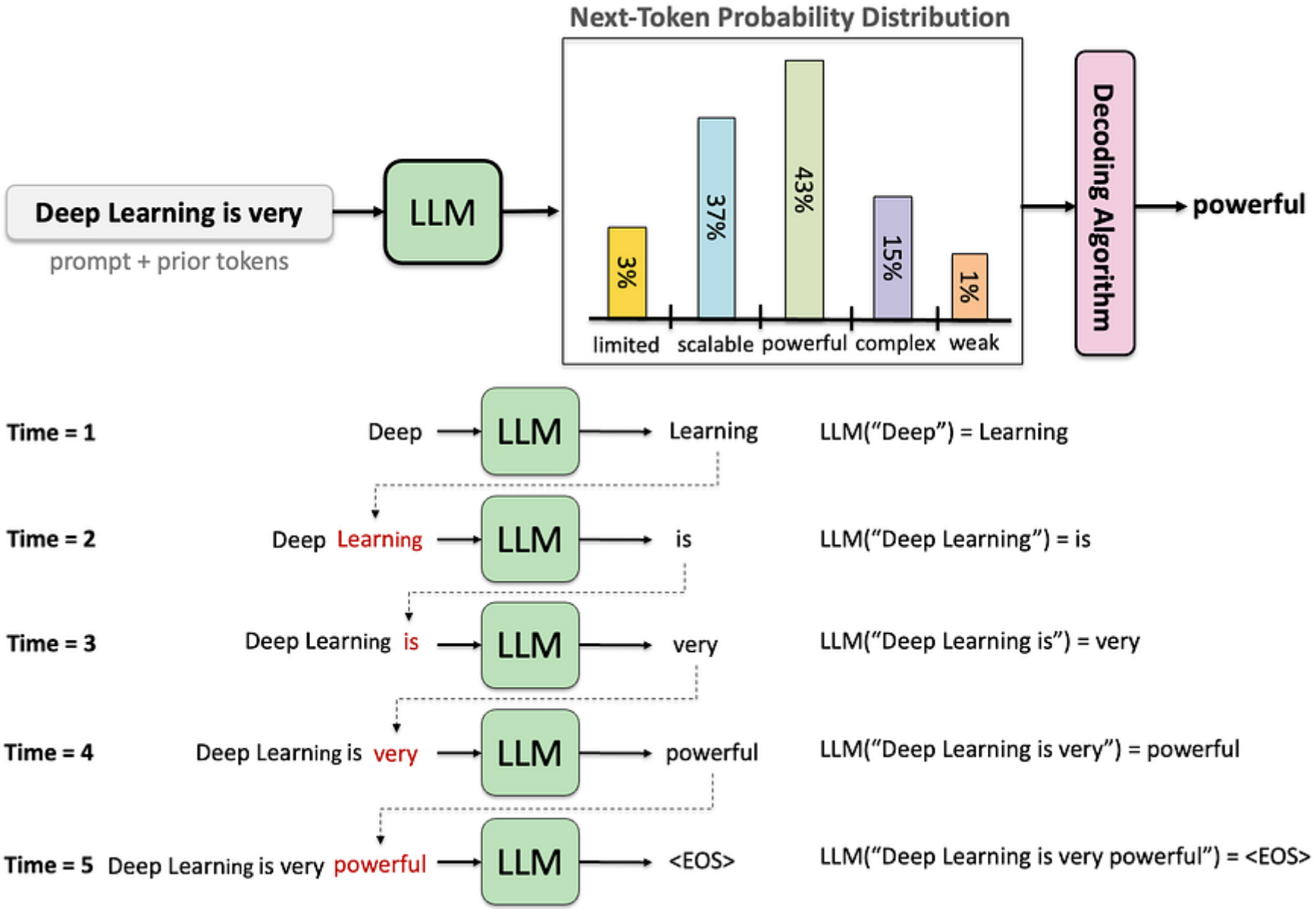
LLM简史


2017 JUN
Transformers
2018 JUN
GPT
2018 OCT
BERT
2019 FEB
GPT-2
2019 OCT
T5
2020 MAY
GPT-3
2021 SEP
FLAN
2022 MAR
GPT-3.5
InstrutGPT
2022 NOV
ChatGPT
2023 FEB
LLaMA
2023 MAR
GPT-4
2024 MAR
GPT-4o
2024 APR
LLaMA-3.1
405B
2024 DEC
OpenAI-o1
DeepSeek-V3
2025 JAN
DeepSeek-R1

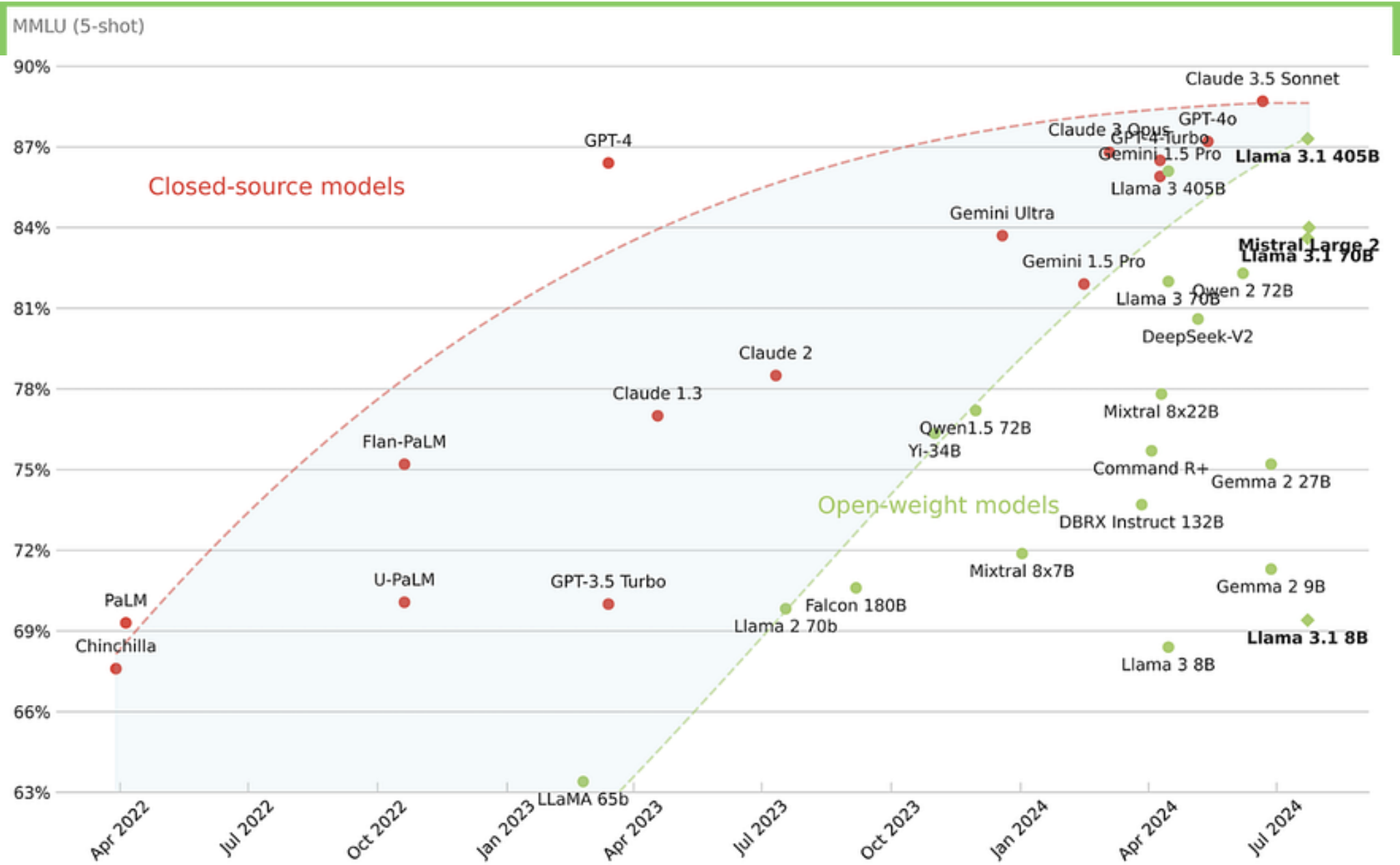
语言模型与大语言模型

- 语言模型是一种「人工智能系统」，旨在处理、理解和生成类似人类的语言。它们从大型数据集中学习模式和结构，使得能够产生连贯且上下文相关的文本，应用于翻译、摘要、聊天机器人和内容生成等领域。
- LLM 专指那些包含数百万甚至数十亿参数的语言模型。
  - 在 2018 年至 2019 年间开始流行，伴随着 BERT 和 GPT-2 等模型的发布。其中，BERT 拥有 3.4 亿参数，而 GPT-2 的参数规模达到了 15 亿。
  - LLM真正引起广泛关注是在 2020 年 GPT-3 发布之后。GPT-3 凭借其惊人的 1750 亿参数规模，展示了通过扩展模型参数量所能带来的革命性能力提升。

自回归语言模型


Next-Token Probability Distribution
Deep Learning is very
prompt + prior tokens
LLM
3%
37%
43%
15%
1%
limited
scalable
powerful
complex
weak
Decoding Algorithm
powerful
Time = 1
Deep
LLM
Learning
LLM("Deep") = Learning
Time = 2
Deep Learning
LLM
is
LLM("Deep Learning") = is
Time = 3
Deep Learning is
LLM
very
LLM("Deep Learning is") = very
Time = 4
Deep Learning is very
LLM
powerful
LLM("Deep Learning is very") = powerful
Time = 5
Deep Learning is very powerful
LLM
<EOS>
LLM("Deep Learning is very powerful") = <EOS>

开源和开放权重模型


MMLU (5-shot)
90%
87%
84%
81%
78%
75%
72%
69%
66%
63%
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023
Jul 2023
Oct 2023
Jan 2024
Apr 2024
Jul 2024
Closed-source models
Open-weight models
Claude 3.5 Sonnet
GPT-4o
Claude 3 Opus
GPT-4-Turbo
Gemini 1.5 Pro
Llama 3.1 405B
Llama 3 405B
GPT-4
Gemini Ultra
Mistral Large 2
Llama 3.1 70B
Gemini 1.5 Pro
Qwen 2 72B
Llama 3 70B
DeepSeek-V2
Claude 2
Claude 1.3
Mixtral 8x22B
Qwen1.5 72B
Yi-34B
Flan-PaLM
Command R+
Gemma 2 27B
DBRX Instruct 132B
Mixtral 8x7B
U-PaLM
GPT-3.5 Turbo
Gemma 2 9B
PaLM
Falcon 180B
Llama 2 70b
Llama 3.1 8B
Chinchilla
Llama 3 8B
LLaMA 65b

第一章：绪论

- 认识大数据
- 机器学习及其发展历史
- 主要机器学习任务
  - 监督学习（Supervised Learning）
  - 序列标注方法（Sequence Labeling）
  - 无监督学习（Unsupervised Learning）
  - 概率主题模型（Probabilistic Topic Modeling）
  - 强化学习（Reinforcement Learning）
  - 深度学习（Deep Learning）
  - 大语言模型（Large Language Model）
- 本章小节

本章小节

- 大数据的定义
- 机器学习的定义
- 机器学习的发展历史
- 机器学习的相关资源
- 机器学习的主要任务