第1章 机器学习引论

好好学习，天天向上。

——毛泽东，1951年题词

北京交通大学《机器学习》课程组

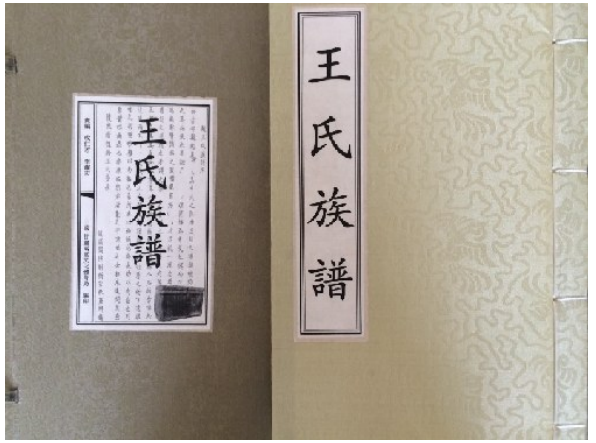
提要

1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

数字化发轫

人类有文字之后，开始数字化时代。 我们用史书来记载一个**民族和国家的宏观生活**，以及部分细节。

- ✓ 中国有确切记录始于公元前841年
- ✓ 之后出现较为准确的大事记，即历史（包括国家历史、地方志、以及族谱等）
- ✓ 记事简约，统计粗略。



大数据时代

- **以前的个人数据**：族谱有部分粗略信息，国家掌握部分粗略信息，其他组织掌握部分粗略个人信息

- **现在的个人数据**：已经量化到秒甚至微秒，如遍布各处的摄像头、无时不开的智能手机定位系统、以及一些其他社交媒体、电子商务等。对于个人的数据记录从衣食住行到工作购物娱乐教育运动医疗，已经无所不包。

大数据时代

Facebook：

- 月活跃用户接近8.5亿
- 每天上传的照片总量为2.5亿张
- 4.25亿移动用户
- 1000亿个关系链接

腾讯：

- QQ：月活跃用户超8亿，人际关系链超1000亿
- 微信：月活跃用户超3.5亿；日均消息量超50亿
- 空间：月活跃用户超6亿；日均相册上传超4亿
- 游戏：腾讯月活跃用户4.5亿；手机游戏月活跃用户近2亿

大数据时代的特征

- 数据粒度细化（分辨率很高）
- 数据传播模式日新
- 数据生产门槛极低、单位极多
- 数据存储单位极多（分布式存储）
- 数据深度利用门槛高

数据粒度细化程度高

- 数据**采集技术**的进步，导致数据采集的**分辨率越来越高**
- 数据**存储能力**越来越高，海量**实时数据**的存储也变得越来越可行
- 数据**采集手段**的多样化，导致**数据描述的多视角**

数据传播模式日新

■**网络技术成熟以前**：传播方式、范围和速度均有限，传播主体门槛高

■**网络技术成熟以后**：**新传播方式不断出现**（网站、微信、微博、Facebook、Twitter, YouTube、TikTok等），传播范围扩大，速度快，传播主体门槛极低，对于数据管理和安全提出了极大的挑战，数据安全的理念和技术面临更新

数据生产门槛低

- **大数据时代以前**：数据生产的要求较高，要求生产者具有较高的知识文化修养

- **大数据时代**：智能手机和各种应用软件的普及，使得人人都是数据生产者，每个企事业也是数据生产者，**对于数据生产者已经没有了要求**，除了设备。

数据存储单位多

- **网络技术成熟以前**，数据存储于特定的地方。
- **网络技术成熟以后**，数据生产者不一定存储，数据浏览者也可能存储，数据传输者也同样可能存储。

数据深度利用门槛高

- **较为成熟技术**

  部分免费提供：搜索，主要培养用户习惯，收集用户使用数据。

  部分付费提供：广告

- **亟待解决的问题**

  如何从数据中提取潜在知识，甚至过滤有害数据，都是巨大的挑战

如何面对大数据时代的问题

智能化是唯一出路。

现在，小到智慧学校，智慧工厂，大到智慧交通、智慧城市、甚至智慧地球，今天的智慧农业也已经提上了研究甚至应用日程。

但是，其面对的主要挑战是什么呢？

从大数据中期望得到什么?

- 希望大数据能够帮助加深对世界的科学认识、科学管理，**提高生产效率和生活质量**。
- 这需要从大数据中**提取出相关的知识**，特别是具有泛化能力的知识。
- 大数据的**处理能力**超过了人的能力，**不得不让位**于机器学习。
- 因此，大数据时代，机器学习任务持续吃重。

提要

1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

机器学习的起源

20世纪30~50年代

- 1936：Alan Turing，自动机模型理论
- 1943：Warren McCulloch和Walter Pitts，MP神经元模型
- 1950：Claude Shannon，逻辑主义
- 1951：John von Neumann，符号演算
- 1952：Arthur Samuel（IBM），西洋跳棋程序
- 1956：达特茅斯会议，人工智能
- 1958：Frank Rosenblatt，感知器

机器学习学科的诞生

20世纪80年代

- 1980年诞生了第一届机器学习学术会议：国际机器学习大会（ICML）源于1980年在CMU举办的机器学习研讨会
- 1983年出版了第一本机器学习著作《机器学习：一种人工智能途径》
- 1986年诞生了第一个正式的机器学习学术期刊：《Machine Learning》于1986年创刊

机器学习的成熟与蓬勃发展

20世纪90年代以后

- 1997年Tom Mitchell的教材《Machine Learning》：标志着机器学习的成熟
- 1990~2010期间，诞生了众多的理论和算法，并走向实用
- 2012年之后，深度学习极速发展

提要

1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

学习的第一个定义

最常见的学习定义：**聚焦学习效果**

计算机系统能够利用经验提高自身的性能。

- 这个定义最早可追溯到人工智能发明者之一：西蒙。
- 大多数机器学习教科书采用这个定义，比如周志华《机器学习》，Tom Mitchell《Machine Learning》等。

学习的第二个定义

学习的可操作定义：**聚焦学习的可操作性**

学习就是一个基于经验数据的函数估计问题。

---Vapnik，统计学习理论的本质，清华大学出版社，2000

- 机器学习算法的分类是根据这个定义给出的，如无监督学习，有监督学习，弱监督学习等。
- 现今文献中常见的学习理论也是以这个定义为学习的默认定义。

学习的第三个定义

学习的可理解定义：**聚焦学习的可理解性**

提取重要模式、趋势，并理解数据，即从数据中学习。

——Hastie T, Tibshirini R, Friedman J. The elements of statistical learning, Springer, 2003

■ 统计学出身的机器学习研究者对于这个定义情有独钟。

三个定义与知识的关系

- 三个定义都强调从经验或者数据中**提取知识**
- 从人的角度：**数据是知识的外在指称，知识是数据的内蕴表示**，本书的初衷正是基于这一点。

如何构建一个机器学习任务的基本框架呢？

提要

1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

机器学习基本框架

- **数据表示：** 给定n个对象的特性表示
- **学习判据：** 判断学习结果好坏
- **学习算法：** 搜寻学习结果
- **学习结果评估：** 评估学习效果

对象特性表示

- 对象特性**输入**表示

  观测得到的对象特性描述。

- 对象特性**输出**表示

  学习得到的对象特性描述。

丑小鸭定理：不存在独立于问题而普遍适用的特征表示，特征的有效与否是问题依赖的。

对象的特性输入表示X

对象集合：$O=\{o_1, o_2, \ldots, o_k, o_N\}$

- 特征矩阵：$[x_{\tau k}]_{n\times p}$

$$x_k=[x_{1k}, x_{2k}, \ldots, x_{pk}]^{\mathrm{T}}$$

- 相异性矩阵：$[d_{kl}]_{n\times n}$
- 相似性矩阵：$[s_{kl}]_{n\times n}$

对象可以是文本、图像、语音等

抽样对象特性输出表示Y

对象集合：$O=\{o_1, o_2, \ldots, o_k, o_N\}$

■特征矩阵：$Y=[y_{\tau k}]_{n\times d}$

对象也可以是文本、图像、语音等

学习判据

用来评估学习的知识空间中各元素对于具体数据集的拟合程度，又称**目标函数**。

$$\boldsymbol{F(\hat{x}) = \omega^T \hat{x} + b}$$

$$L = D\big(f(X), F(X)\big) = \sum_{k=1}^{N} (f(\hat{x}_k - F(\hat{x}_k))^2 = \sum_{k=1}^{N} (f(\hat{x}_k) - \omega^T \hat{x}_k - b)^2$$

学习算法

- 依据目标函数，根据优化算法**找出最优化知识表示的过程**。
- 是否存在不依赖于具体问题的最优学习算法呢？

——否，任何一个算法都只是适合特定问题特定数据的，具体可见**没有免费午餐定理**。

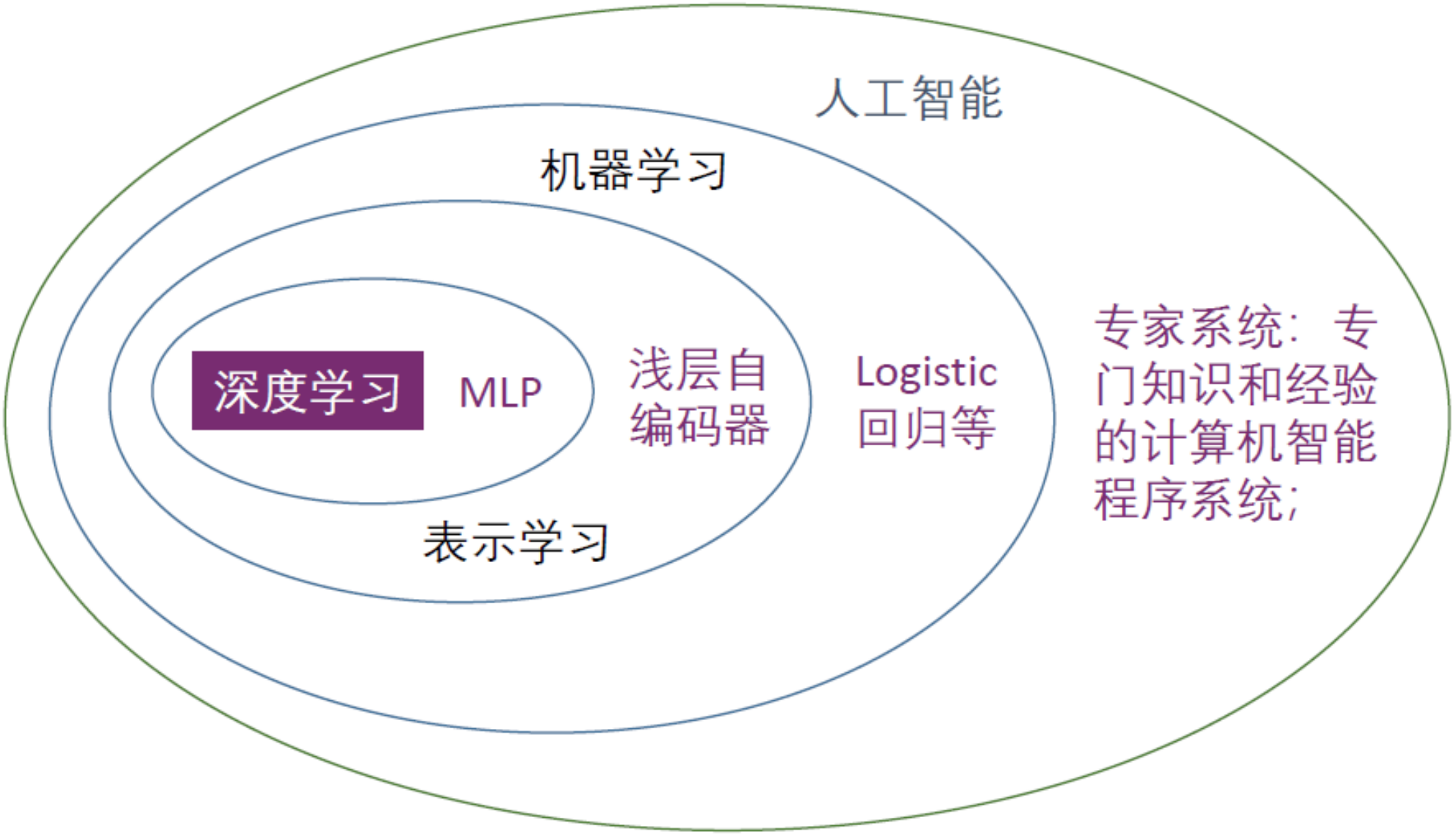
学习结果评估

- 很多时候，学习算法一般不能得到最优结果，只能局部最优或者满意解。
- 而且，学习模型本身是问题的一个简化，**学习到的结果**是否满足实际需求，需要**进行评价**。

机器学习的相关概念

- 人工智能
- 机器学习
- 表示学习
- 深度学习
- 数据挖掘

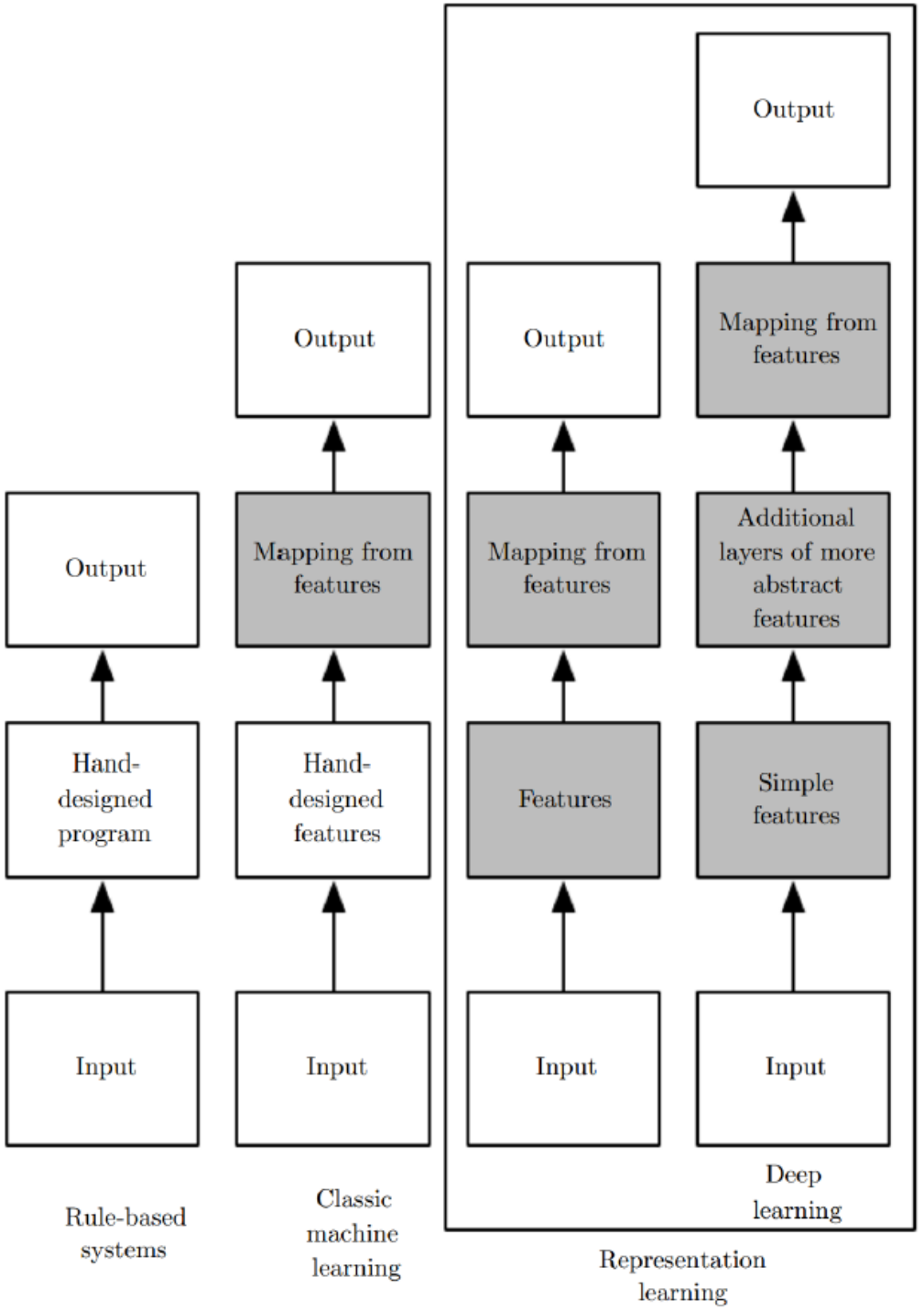
机器学习的相关概念


人工智能
机器学习
深度学习
MLP
浅层自编码器
表示学习
Logistic回归等
专家系统：专门知识和经验的计算机智能程序系统；

机器学习的相关概念

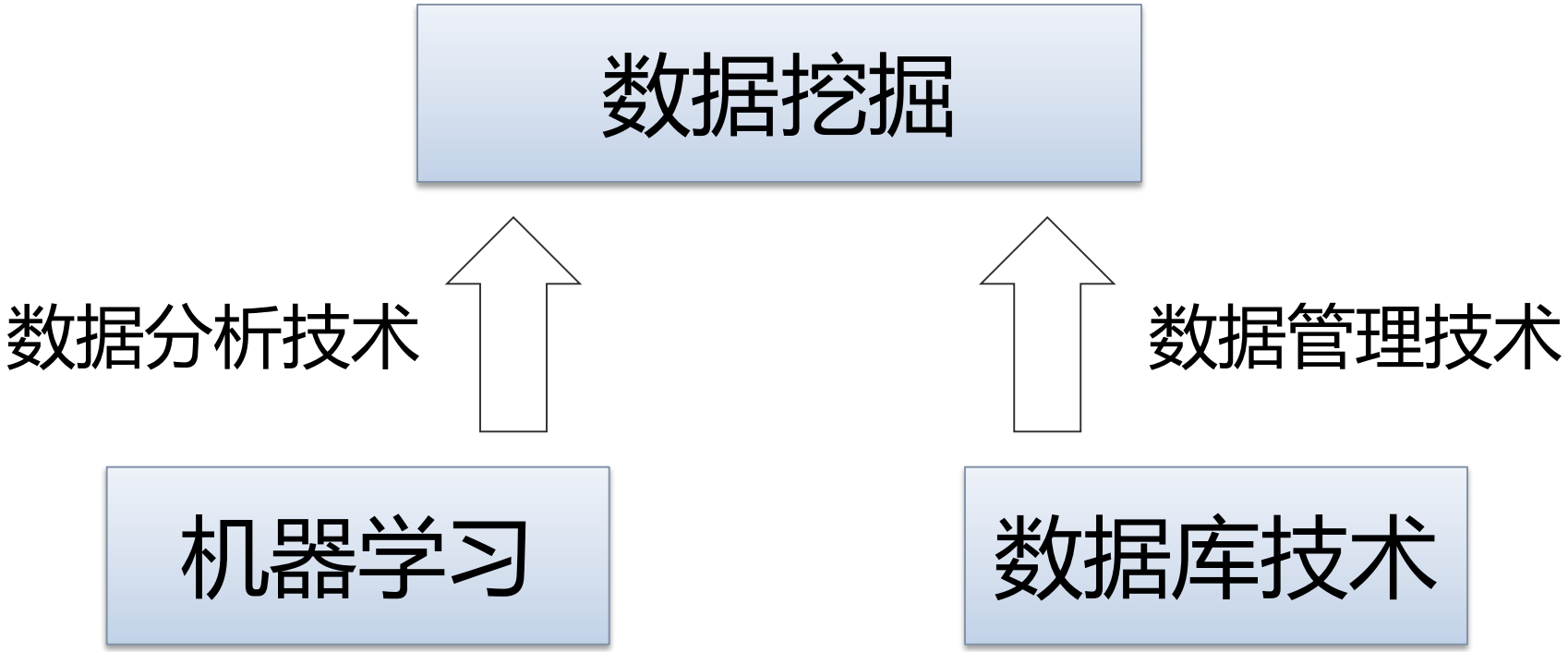
- 深度学习及其他人工智能方法

阴影：可学习部分

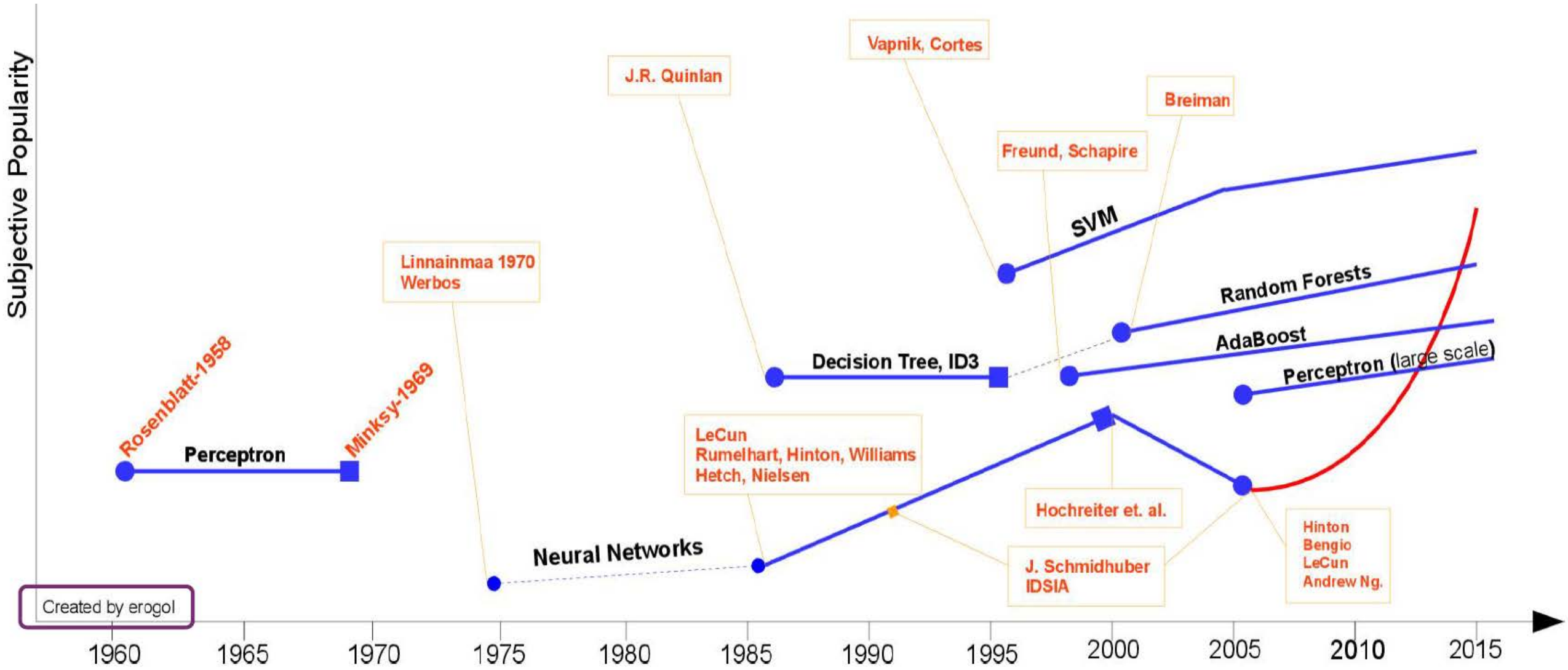


机器学习的相关概念

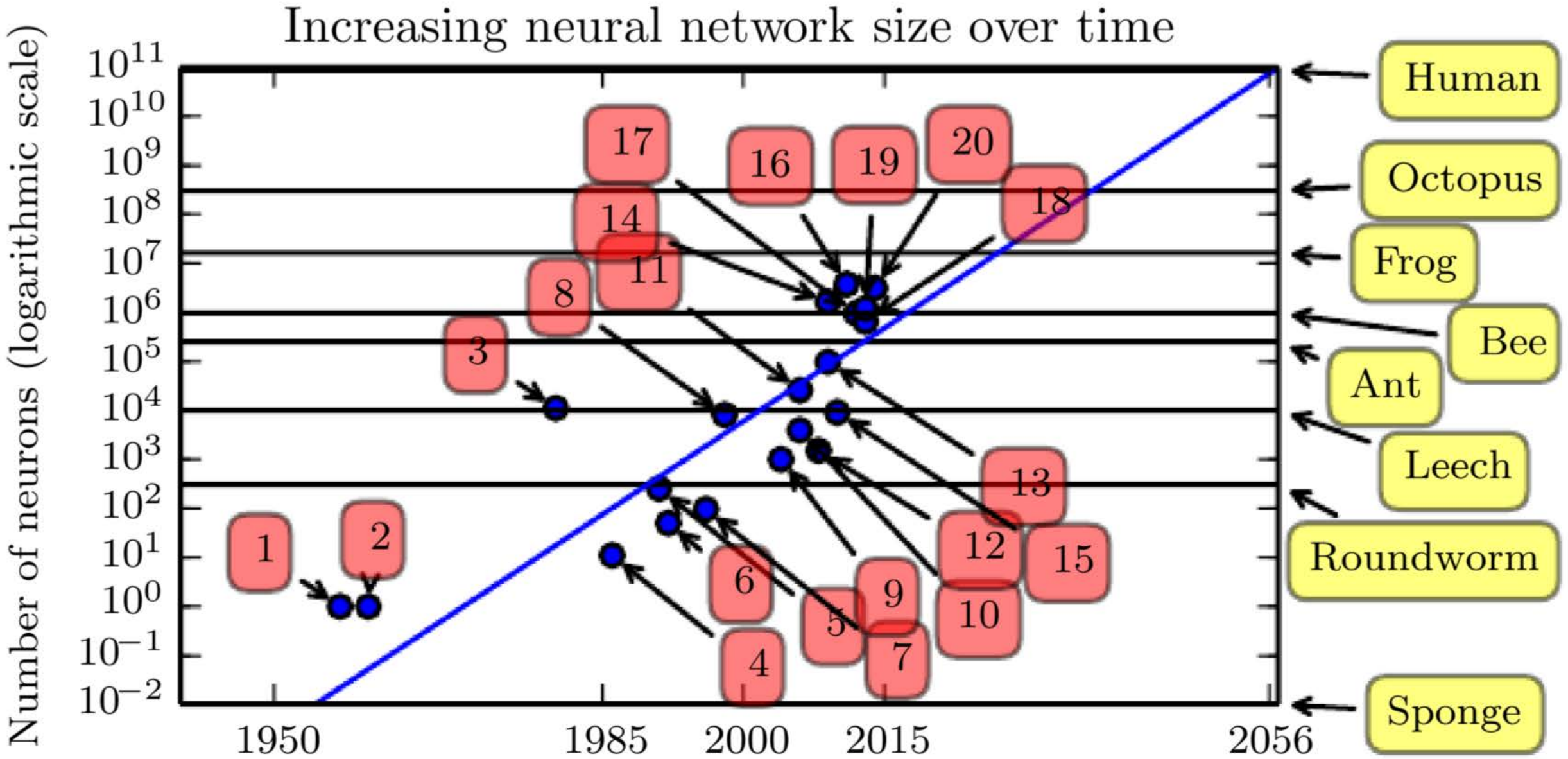
- 机器学习与数据挖掘



机器学习经典算法发展历程



神经网络的规模


Increasing neural network size over time
Number of neurons (logarithmic scale)
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
1950
1985
2000
2015
2056
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Human
Octopus
Frog
Bee
Ant
Leech
Roundworm
Sponge

提要

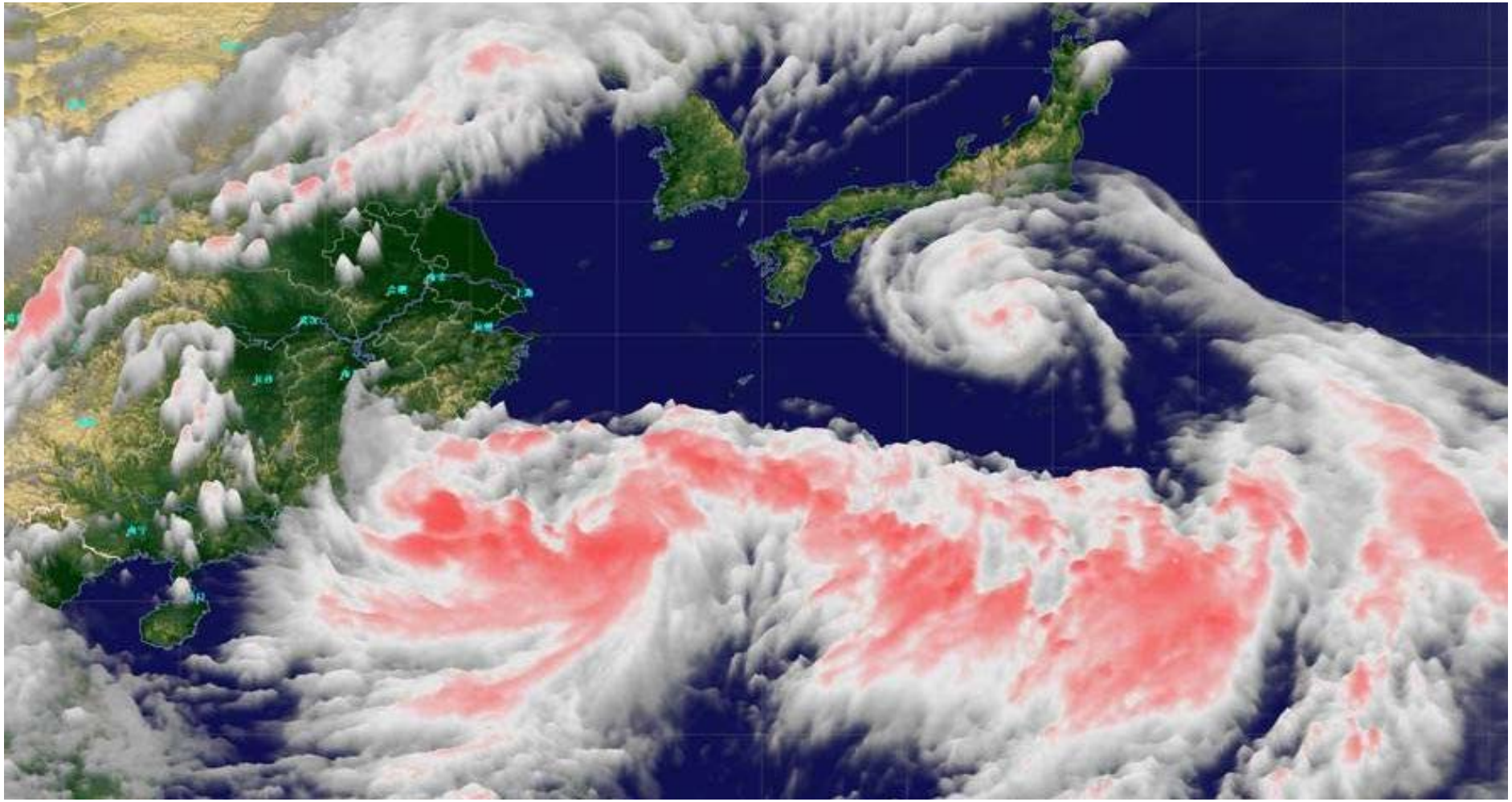
1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

搜索引擎

- Google、Baidu、 搜狗

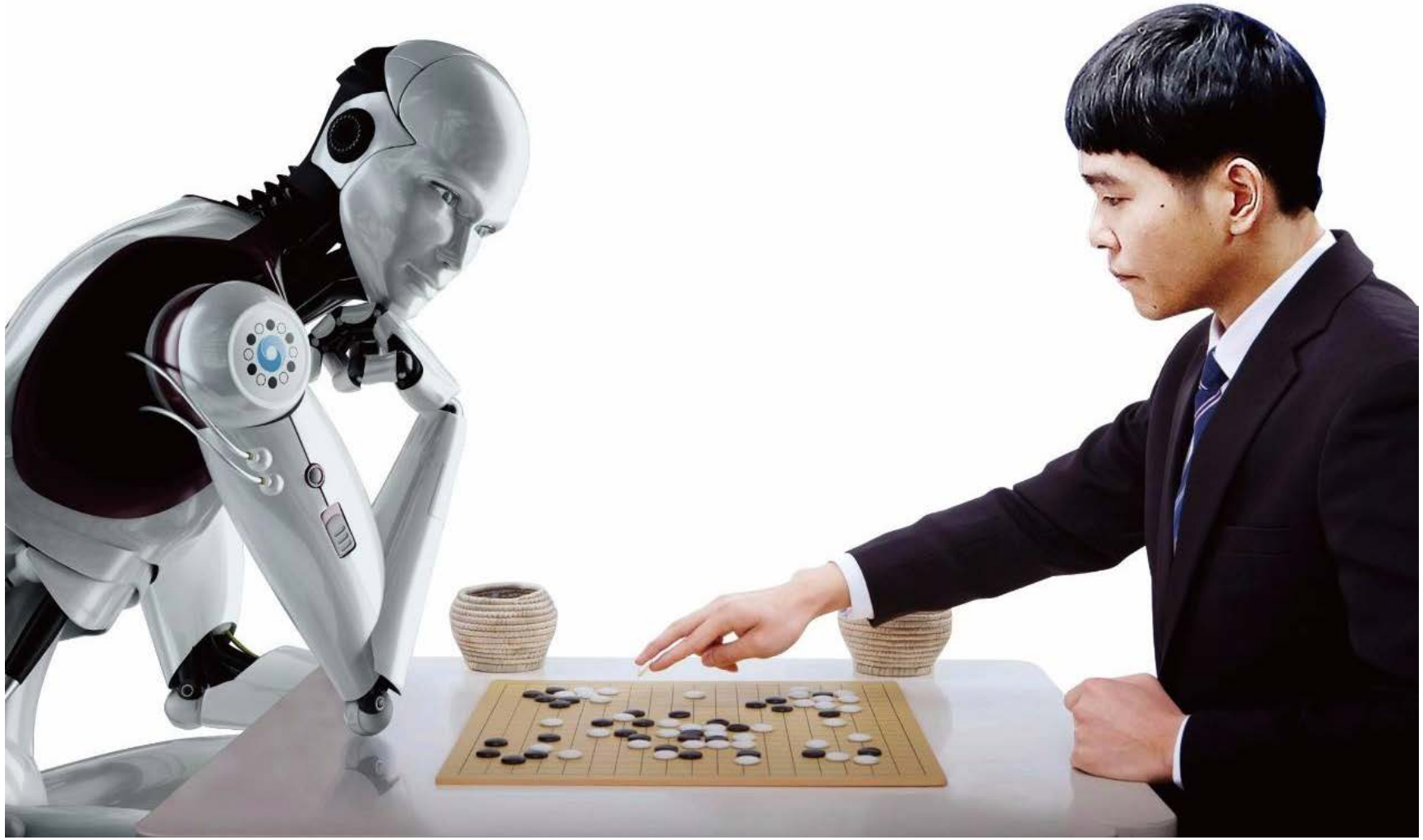
预测

例如：天气预报

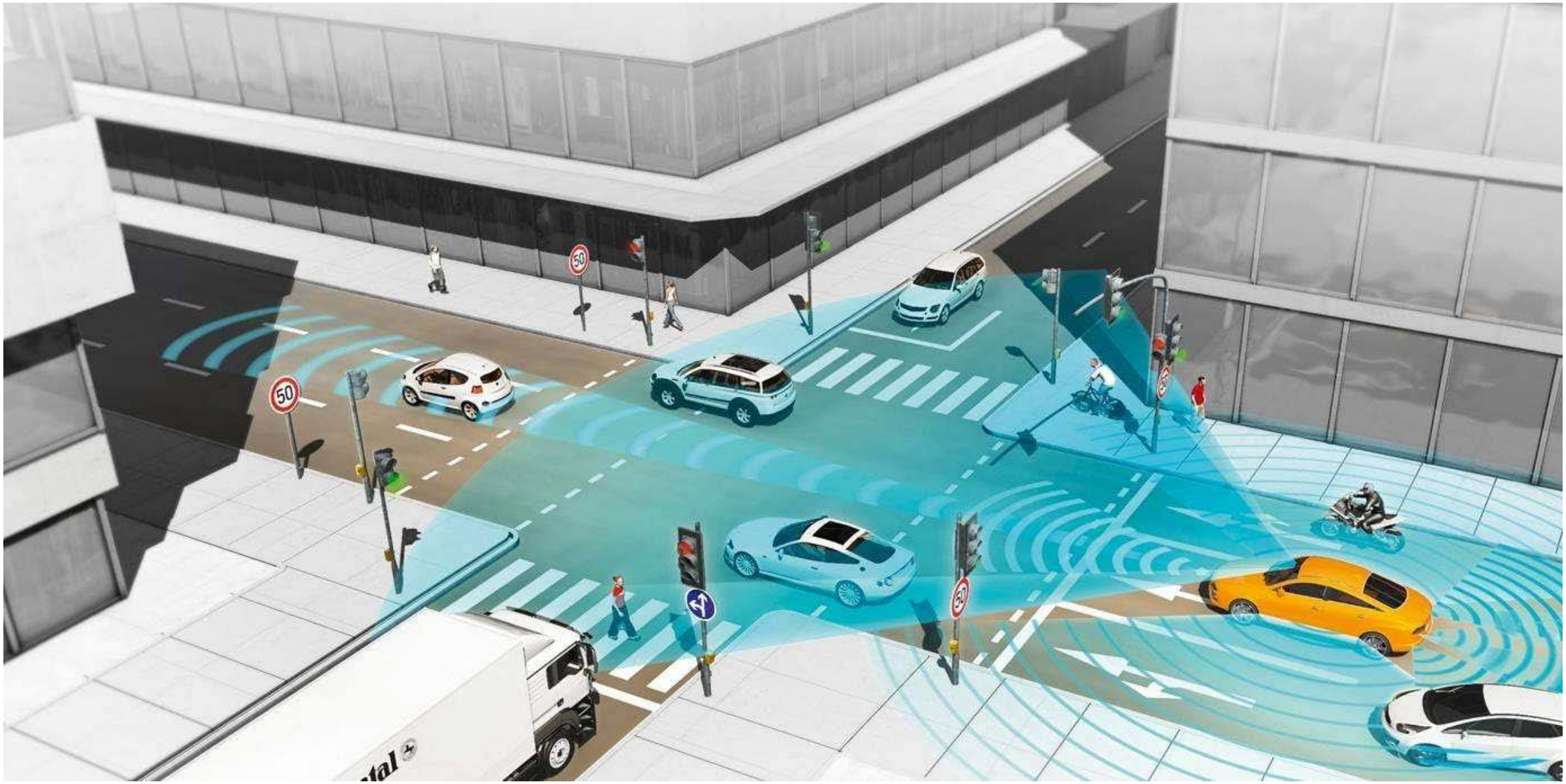


游戏

- AlphaGo、AlphaZero

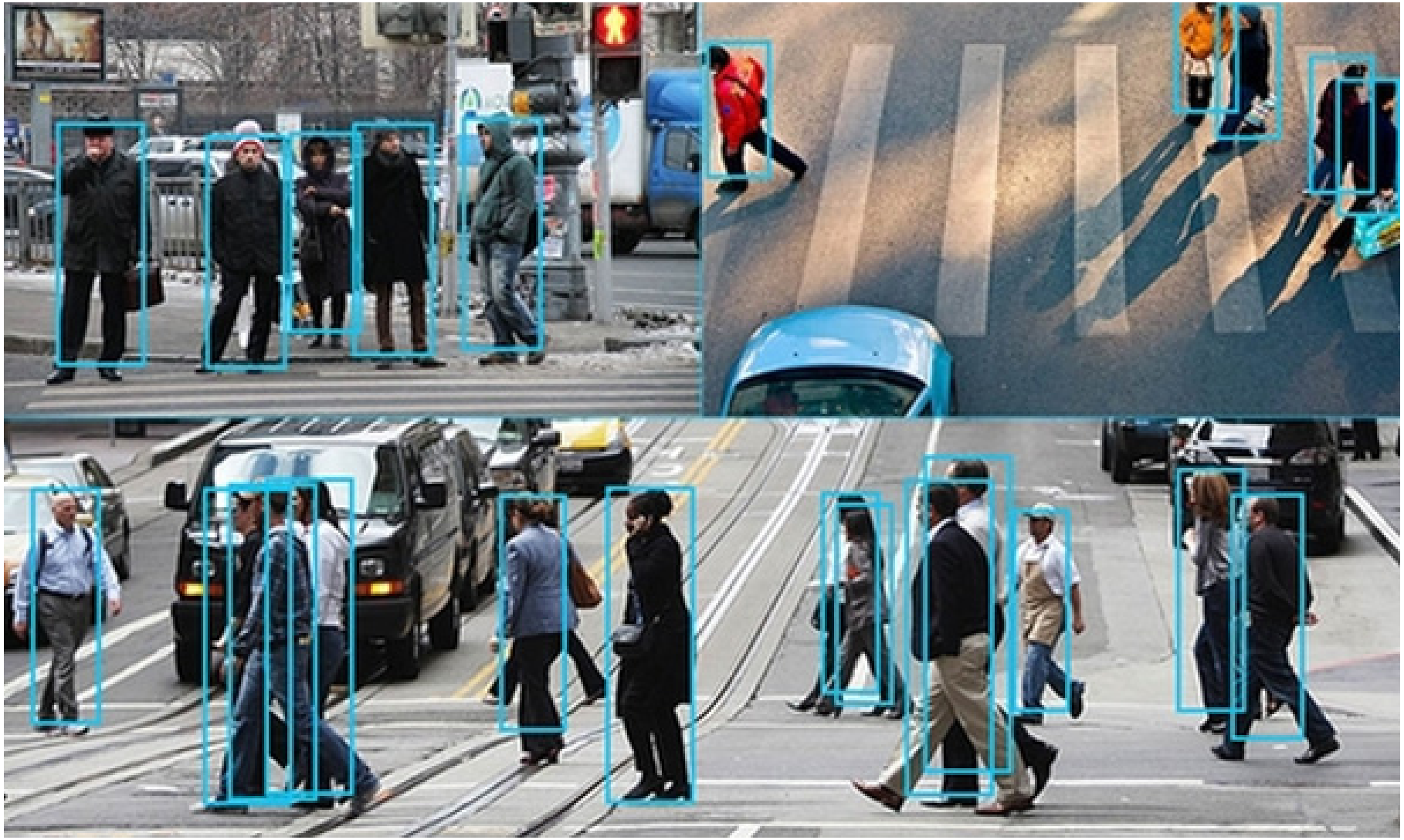


无人驾驶



计算机视觉

- 目标跟踪



ImageNet Challenge

- 2015年：微软的神经网络系统错误率为4.94%，低于人类测试者的5.1%。

	layers	error(top 5)	
AlexNet	8	15%	2012
VGGNet	19	7.32%	2014
GooleNet	22	6.66%	2014
MSRA	152	3.57%	2016

新兴交叉学科

- 计算广告学
- 计算社会学
- 计算金融学
- 计算历史学
- 计算生物学
- ……

提要

1. 机器学习的缘起：大数据
2. 机器学习的出现
3. 机器学习的定义
4. 学习模型的基本组成
5. 机器学习的用途
6. 机器学习的学习资源

数据资源

- UCI数据库：https://archive.ics.uci.edu/ml/datasets.html
- Kaggle数据集：https://www.kaggle.com/datasets
- Amazon数据集：https://registry.opendata.aws/
- Microsoft数据集：https://msropendata.com/
- Awesome数据集：https://github.com/awesomedata/awesome-public-datasets
- 计算机视觉数据集：https://www.visualdata.io/
- Google数据集搜索服务：https://toolbox.google.com/datasetsearch

算力平台资源

- PC机
- 服务器（CPU，GPU）
- 分布式集群
- 云计算平台
- 超算中心

算法平台资源

- Scikit-learning
- Spark MLlib
- Apache Mahout
- TensorFlow、PyTorch、Caffe、Keras、MXNet、Theano

中文书籍资源

- 于剑，《机器学习：从公理到算法》，清华大学出版社，2017
- 李航，《统计学习方法》（第2版），清华大学出版社，2019
- 周志华，《机器学习》，清华大学出版社，2016

会议资源

- **ICML** (International Conference on Machine Learning)
- **ICLR** (International Conference on Learning Representations)
- **NeurIPS** (Annual Conference on Neural Information Processing Systems)
- **KDD** (Knowledge Discovery and Data Mining Conference)
- **AAAI** (the Association for the Advance of Artificial Intelligence)
- **IJCAI** (International Joint Conference on Artificial Intelligence)
- **COLT** (Conference On Learning Theory)
- **ECML** (European Conference on Machine Learning)
- **TheWebConf** (The Web Conference) (原WWW)

期刊资源

- **TPAMI** (IEEE Transactions on Pattern Analysis and Machine Intelligence)
- **JMLR** (The Journal of Machine Learning Research)
- **Artificial Intelligence**

CCF推荐会议期刊排名

http://www.ccf.org.cn/xspj/rgzn/

中国计算机学会推荐国际学术刊物
(人工智能)

A类

序号	刊物名称	刊物全称	出版社	地址
1	AI	Artificial Intelligence	Elsevier	http://dblp.uni-trier.de/db/journals/ai/
2	TPAMI	IEEE Trans on Pattern Analysis and Machine Intelligence	IEEE	http://dblp.uni-trier.de/db/journals/pami/
3	IJCV	International Journal of Computer Vision	Springer	http://dblp.uni-trier.de/db/journals/ijcv/
4	JMLR	Journal of Machine Learning Research	MIT Press	http://dblp.uni-trier.de/db/journals/jmlr/

B类

序号	刊物名称	刊物全称	出版社	地址
1	TAP	ACM Transactions on Applied Perception	ACM	http://dblp.uni-trier.de/db/journals/tap/
2	TSLP	ACM Transactions on Speech and Language Processing	ACM	http://dblp.uni-trier.de/db/journals/tslp/
3		Computational Linguistics	MIT Press	http://dblp.uni-trier.de/db/journals/coling/
4	CVIU	Computer Vision and Image Understanding	Elsevier	http://dblp.uni-trier.de/db/journals/cviu/
5	DKE	Data and Knowledge Engineering	Elsevier	http://dblp.uni-trier.de/db/journals/dke/index.html
6		Evolutionary Computation	MIT Press	http://dblp.uni-trier.de/db/journals/ec/
7	TAC	IEEE Transactions on Affective Computing	IEEE	http://dblp.uni-trier.de/db/journals/taffco/

北京交通大学《机器学习》课程组

于　剑：jianyu@bjtu.edu.cn；　　李晓龙：hlli@bjtu.edu.cn；

景丽萍：lpjing@bjtu.edu.cn；　　吴　丹：wudan@bjtu.edu.cn；

田丽霞：lxtian@bjtu.edu.cn；　　万怀宇：hywan@bjtu.edu.cn；

黄惠芳：hfhuang@bjtu.edu.cn；　　王　晶：wj@bjtu.edu.cn.