



机器学习 Python实践

中山大学物理与天文学院

李霄栋

2019年秋季学期

课程宗旨

- 低门槛，快速入门（理工科福利课）



- 强调 概念的理解+动手实践（实用主义，拿来主义，专门搞事情）

- 不强调算法底层实现（想当砖家的可以走了）



学点撩汉技巧搞定那个帅哥。
还没想成为这方面理论砖家。

本课程内容



混完这学期，我也是会
机器学习的人了。

- Python 语言与常用库（1-3周）
- （基于 sklearn库） 人工智能基本概念与常用术语；经典机器学习算法（knn，线性算法，支持向量机，决策树，朴素贝叶斯，随机森林，主成分分析，简单全连接神经网络，...）
- （基于 tensorflow+keras） 深度学习（卷积神经网络，循环神经网络）（最后3-4周）

考核



公选课，除非你做得太过分，原则上不当杀手。
以鼓励为主，目的让大家学到东西，对机器学习产生兴趣。

在物天我是公认的天使/佛系
如果挂了你，只能是你太过分

- 平时成绩（80%）

- 期中考查+期末作业
- 期末作业方式多样

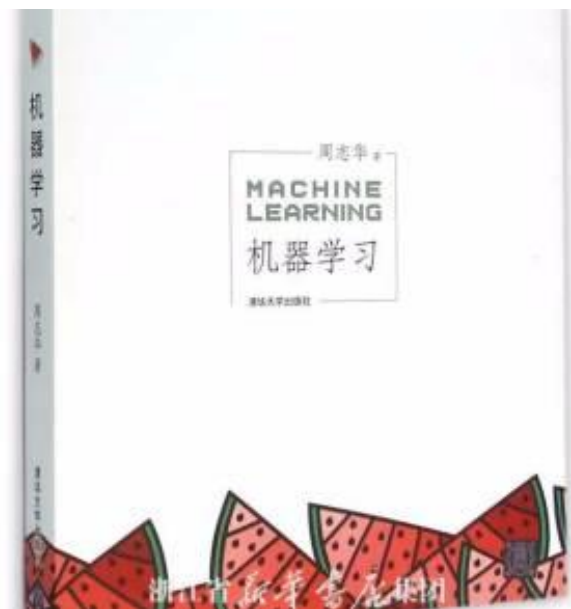
- 平时布置的练习题；利用 ML 完成一个小 project；一篇 4-5 页小论文或学习心得报告；...

- 期末考试（20%）

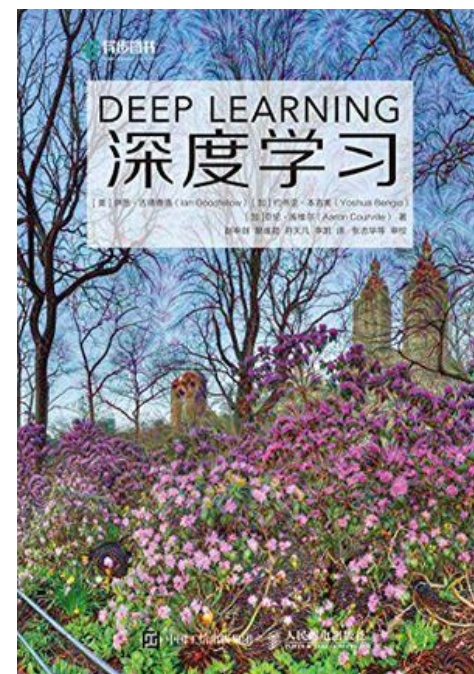
- 差异性、人性考核

- 不以绝对实力为唯一考核标准；更看重你到底进步了多少
 - 期末请写 ~**100**字报告，说明自己“学到了什么”；每人都给自己打个分

参考资料



周志华《机器学习》（西瓜书）



Deep Learning 深度学习
(AI 圣经)

参考资料



大部头的书不好看，
老师我特意把内容精简
做成课件了

- 为方便大家阅读，本课件刻意做得，很罗嗦。
- 如果对某个概念不懂，建议直接 谷歌/必应/百度

★ 收藏 | 967 | 分享

人工神经网络 🔒 锁定

 本词条由“科普中国”科学百科词条编写与应用工作项目 审核。

人工神经网络（Artificial Neural Network，即ANN），是20世纪80年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

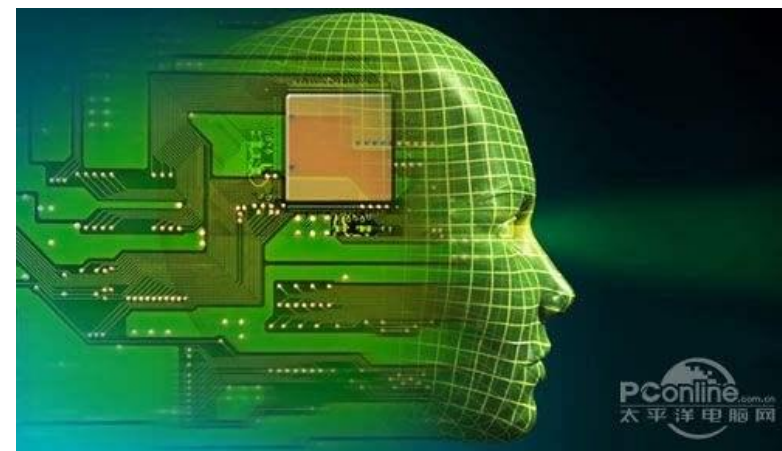


机器学习 概论



机器学习的需要

- 人类第一次构思可编程计算机时，已经在思考计算机能否变得智能。
- 在人工智能的早期，通过一系列形式化的数学规则来描述的问题迅速得到解决。
- 而那些对人来说很容易执行、但很难形式化描述的任务（如，识别人们所说的话或图像中的脸），对人来说很简单，对计算机却构成巨大的挑战。

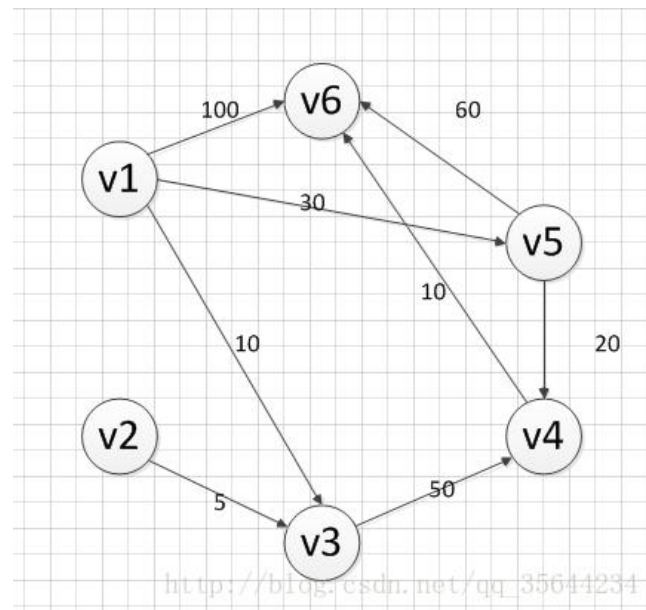


举例：形式化（explicit）的问题

问题1：寻找 1,000,000 以内的所有质数。

问题2：在右边有向图中，寻找 **v1** 到其他顶点的最短路径。

问题3：进行一场国际象棋比赛。



上述问题的解决方式可以用一系列形式化的数学规则来描述。

解决问题的方式为：人向计算机提供**形式化 (explicit)** 的算法或指令，计算机按照人的指令工作。

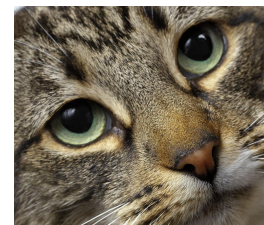
这些抽象而形式化的任务对人类而言是困难的脑力劳动，对计算机却是最容易的。

这样的问题不需要机器学习就能很好地解决。



上世纪末，IBM公司的软件“深蓝”击败了俄罗斯国际象棋棋王卡斯帕罗夫，震惊世界。

举例：非形式化的问题



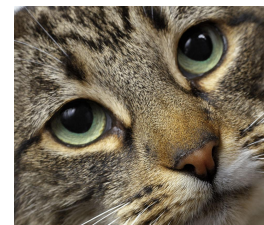
问题：辨认右图中的猫和狗。

问题的解决无法用形式化的规则来描述。

人们无法找到给出一个严格、万能的定义，向机算计说明，满足什么样特征的图片是猫或者狗。

（动物会处于不同的姿态；只有部分身体可见；具有不同的明亮光线；动物与背景混在一起；...）

举例：非形式化的问题



问题：辨认右图中的猫和狗。

这些问题对计算机很困难，但人却可以凭借着自己在日常生活中获得的巨量知识，使用直觉轻而易举的解决。解决过程中并不涉及到形式化的数学推理（你们辨认图片时，有没在大脑进行推理：因为**、**条件满足，所以推断，这是猫...）。

计算机需要获得同样的知识才能表现出智能。

把这些非形式化的知识传递给计算机，是机器学习需要解决的问题。

举例：硬编码知识库遇到的困难

- 一些人工智能项目力求将世界的知识用**形式化的语言**进行**硬编码（hard-code）**。计算机可以使用**逻辑推理规则**来自动地**理解**这些形式化语言中的声明。这就是众所周知的**知识库（knowledge base）**方法。
- 这其中最著名的项目是 Cyc (Lenat and Guha, 1989)。Cyc 包括一个**推断引擎**和一个使用 CycL 语言描述的**声明数据库**，定义了 320 万条人类定义的断言，涉及 30 万个概念。这些声明是由人类监督者输入的。



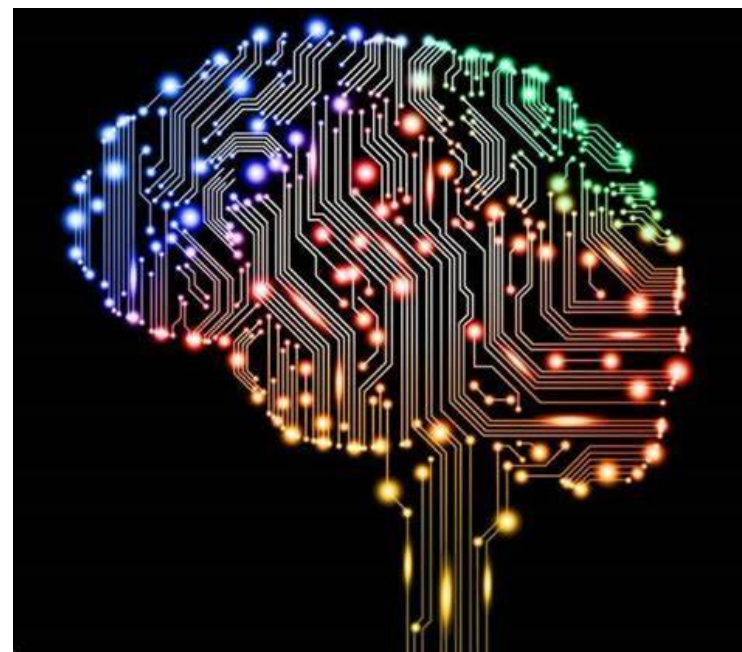
举例：硬编码知识库遇到的困难

- 这是一个笨拙的过程。Cyc 在很多问题的理解上出现问题。例如，Cyc 不能理解一个关于“名为 **Fred** 的人在早上剃须”的故事。
- Cyc 的推理引擎检测到故事中的不一致：它知道人体的构成不包含电器零件，但由于 Fred 正拿着一个电动剃须刀，它认为实体——“正在剃须的 **Fred**”含有电器零件。
- 因此，它产生了这样的疑问——**Fred**在刮胡子的时候是否仍然是一个人。



机器学习的需要

- 依靠硬编码的知识体系面临的困难表明，AI系统需要具备**自己获取知识的能力**，即从原始数据中提取模式的能力。
- 这种能力称为**机器学习**（machine learning）。引入机器学习使计算机能够解决涉及现实世界知识的问题，并能做出看似主观的决策（如，决定是否进行剖腹产；区分垃圾邮件和合法电子邮件）。



机器学习的定义 (from wiki)

- Machine learning (ML) is the scientific study of **algorithms** and **statistical models** that computer systems use to perform a specific task **without** using **explicit instructions**, relying on **patterns** and **inference** instead.
- Machine learning algorithms build a **mathematical model** based on sample data, known as "**training data**", in order to make predictions or decisions without being **explicitly programmed** to perform the task.

表示/特征

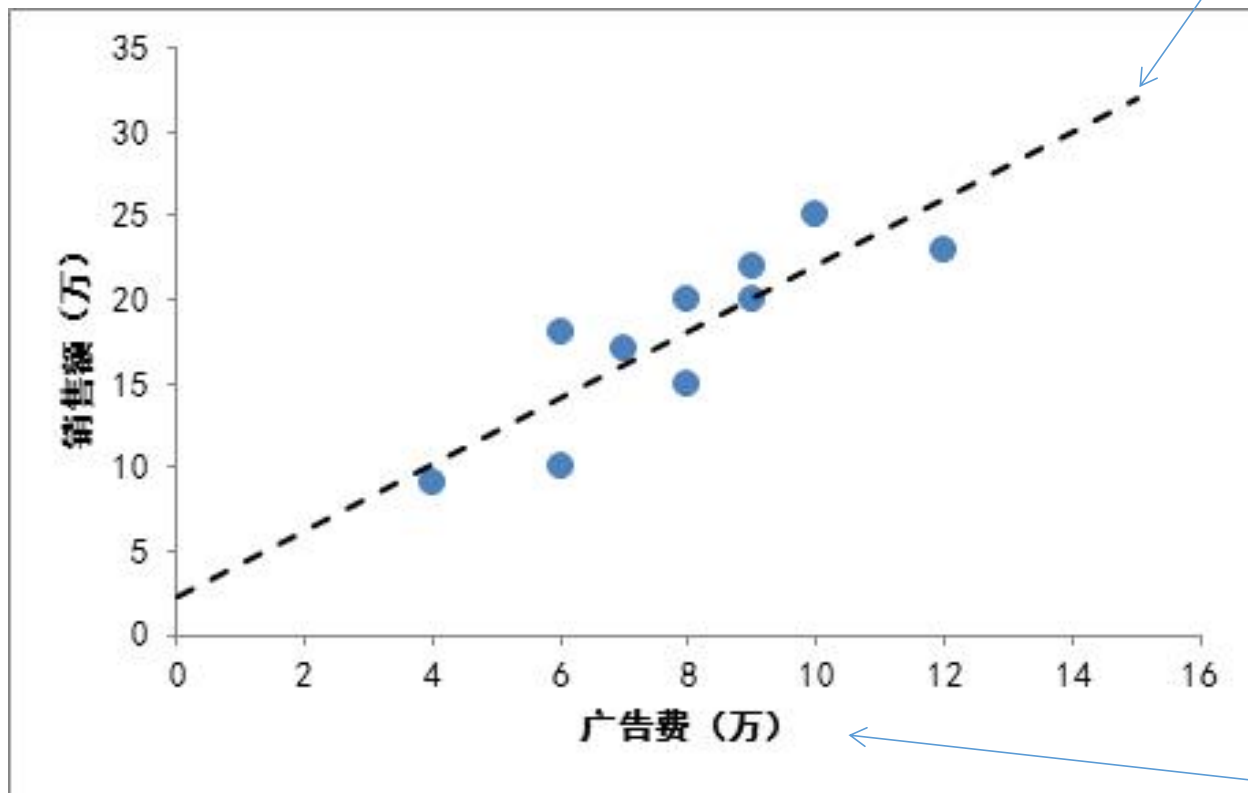
- 机器学习算法的性能很大程度上依赖于给定数据的表示（**representation**）/ 特征（**feature**）。很多任务通过以下方式解决：先提取一个合适的特征集，然后将这些特征提供给算法。
- 例如，对于研究生导师选取学生的任务来说，一个有用的特征是其学习成绩。这个特征对判断该学生是学霸还是学渣提供了有力线索。
- 对于简单的问题，数据的特征集完全可以由人工准备好并输入给机器学习算法。



通过眼神这个特征，我很容易知道
你们现在有没有认真听课

最简单的机器学习：线性回归

某公司产品的广告费与销售额关系图



从特征映射到结果

机器学习旧数据，
建立模型，预测新数据

基于的特征十分简单，
可以由人工准备

建立的模型也十分简单，
可以由人工定义好

特征、表示

基于已有的数据，机器建立起线性的模型，
对新的数据点作出预测

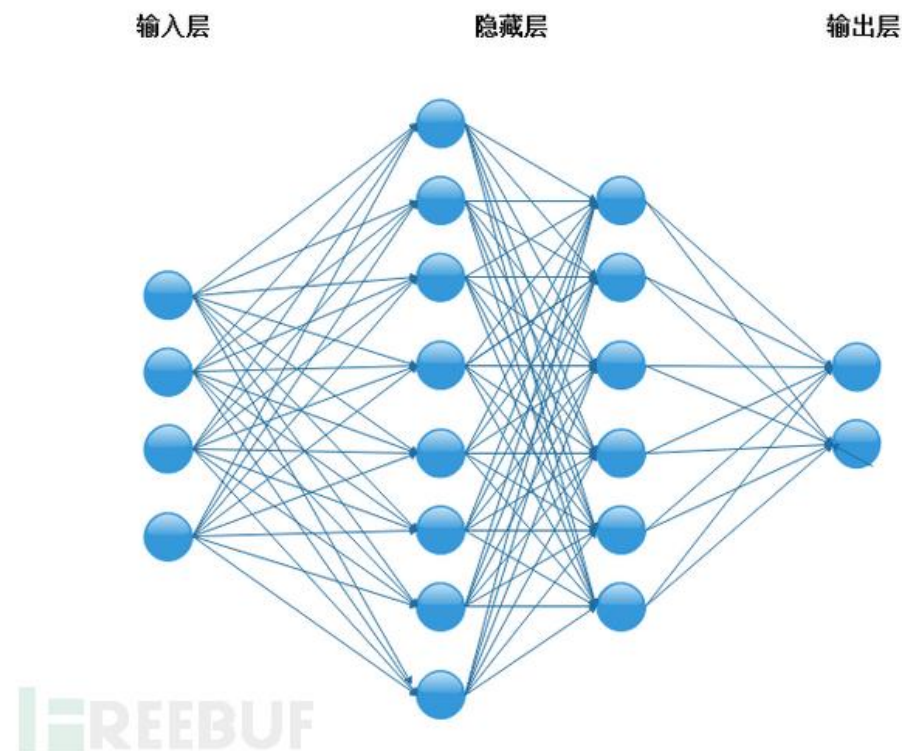
特征学习的困难

- 例如，我们编写一个程序来监测照片中的车。我们可能会想用轮子的存在与否来作为一个特征。但是我们难以准确地根据像素值来描述轮子：
 - 车轮的图像会因场景而异，如落在车轮上的阴影、太阳照亮的车轮的金属零件、汽车挡泥板对车轮的遮挡等。
- 从原始数据提取出如此高层次、抽象的特征是非常困难的。如此复杂的特征设计需要耗费大量人工、时间和精力，甚至花费整个研究团队几十年时间。
- 解决这个问题的途径是使用机器学习来挖掘特征本身——这催生了深度学习（deep learning）。

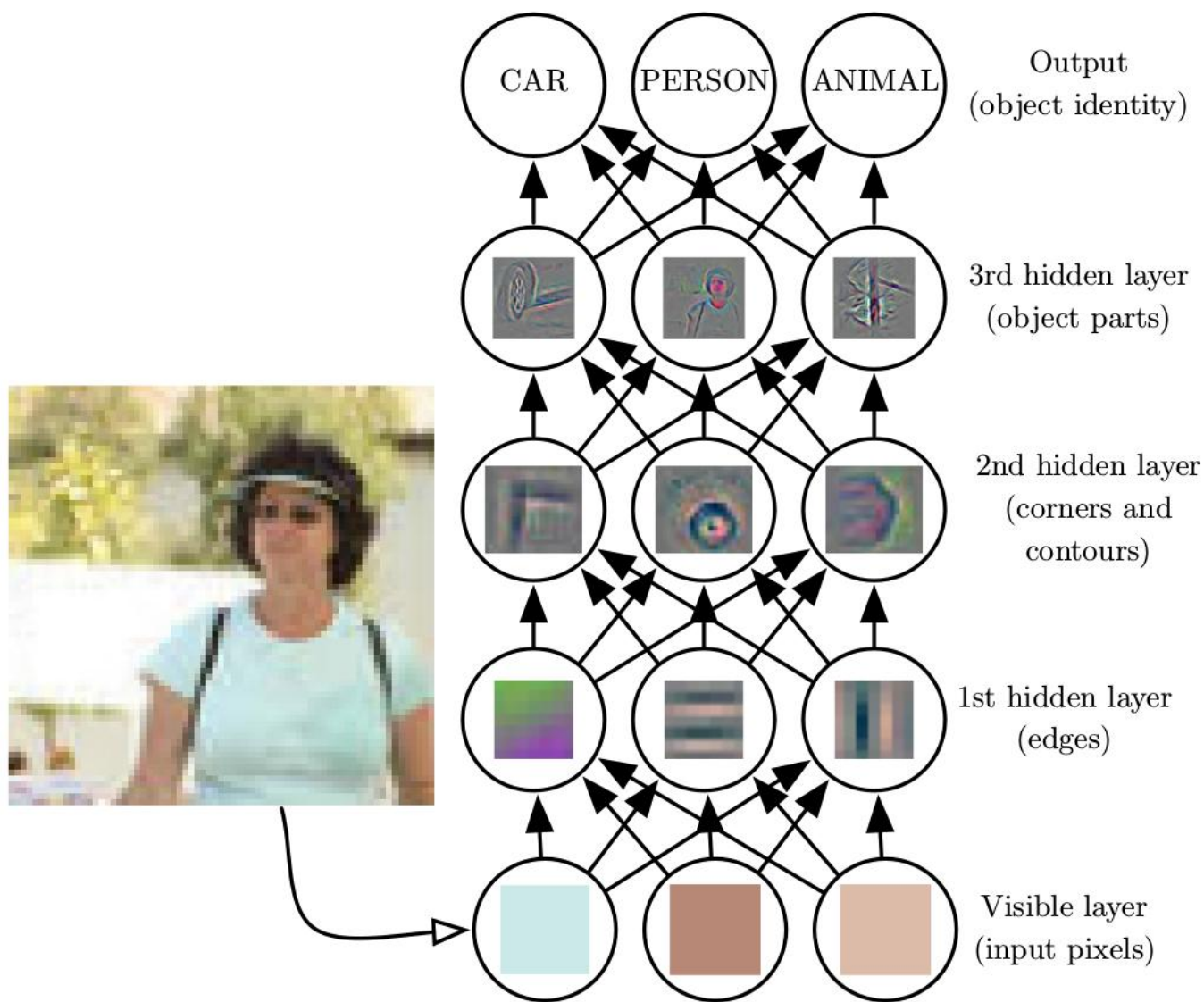


深度学习：特征的特征

- 深度学习（**deep learning**）通过其他较简单的特征来表达复杂特征，解决了特征学习中的核心问题。
- 深度学习让计算机通过较简单的概念构建复杂的概念。其典型例子是神经网络（**neural network**）或多层感知机（**multilayer perceptron, MLP**）。
- 数学上说，多层感知机是将一组输入值映射到输出值的复杂数学函数。该函数由许多较简单的函数复合而成，往往具有分层的函数复合结构。



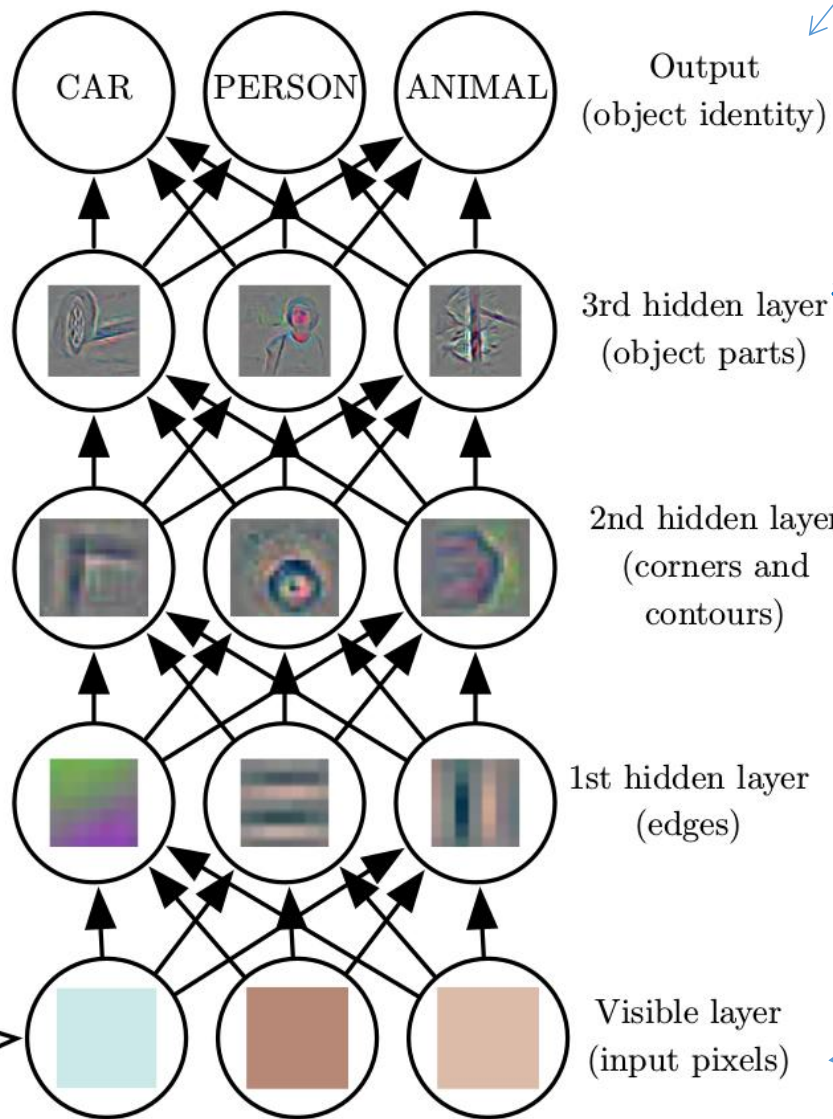
深度学习示例



左图：深度学习系统通过组合较简单的概念（例如，折线、角、轮廓）来表示图像中人的概念：

- 直接输入的特征为最简单的像素。这些被成为可见层（**visible layer**），包含我们直接观察到的变量。
- 深度学习从图像中提取出越来越多具有抽象特征，构成一个或多个隐藏层（**hidden layer**）。它们的值不在原始数据中给出。
- 第1层可以轻易地通过比较相邻像素的亮度识别边缘。第2层可以容易地将边缘识别为角和轮廓。第3层通过组合角和轮廓，检测出整体的物体。

深度学习示例



从特征映射到结果

不可见的
特征、表示

可见的
特征、表示

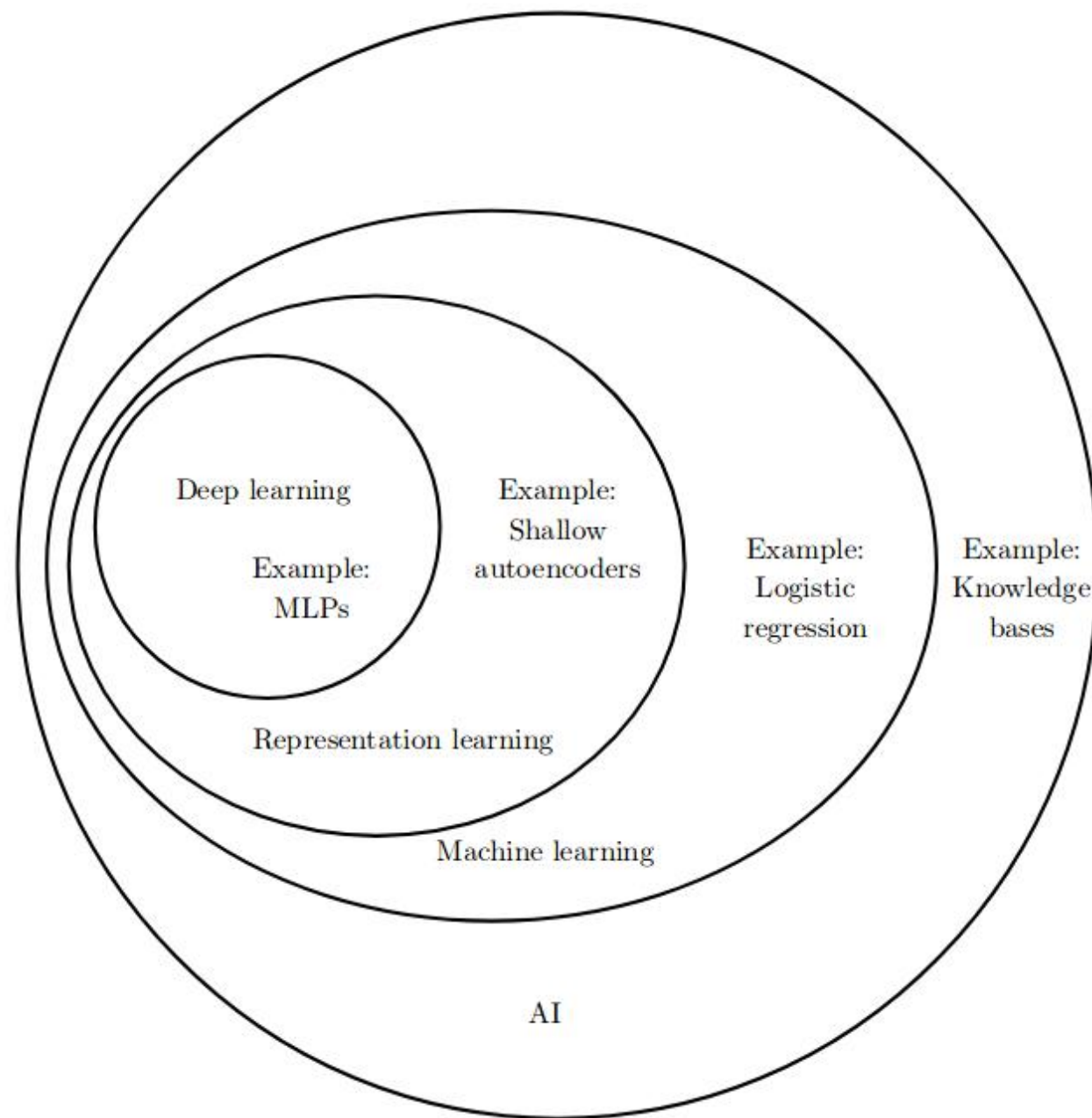
机器学习旧数据，
建立模型，预测新数据

隐藏的特征十分复杂，
难以由人工准备

建立的模型也十分复杂，
无法由人工定义好，
甚至在建立好之后，
人工也无法理解

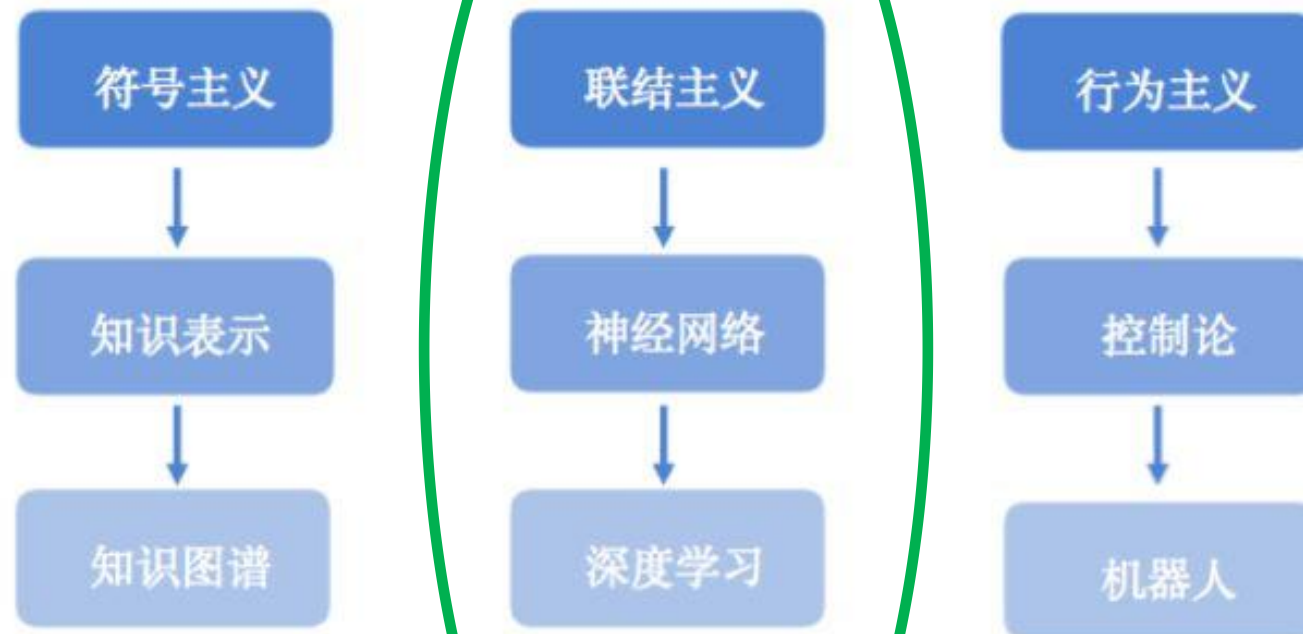
外圈到里圈，

- 人工智能（例如：知识库硬编码）
- 机器学习（例如：逻辑回归）
- 表示学习（例如：浅层的自编码器（自编码指算法能将输入数据转换成不同的表示。自编码器通常也通过神经网络实现））
- 深度学习（例如：多层感知机）



人工智能的三大学派

本课程内容



风控利器

2018年第二届金融科技风控大会
2018(2nd)
Financial Risk Management Summit

主办方：



从左到右，

- 基于形式化规则的系统

- 输入 -> 手工设计的程序 -> 输出

- 经典机器学习

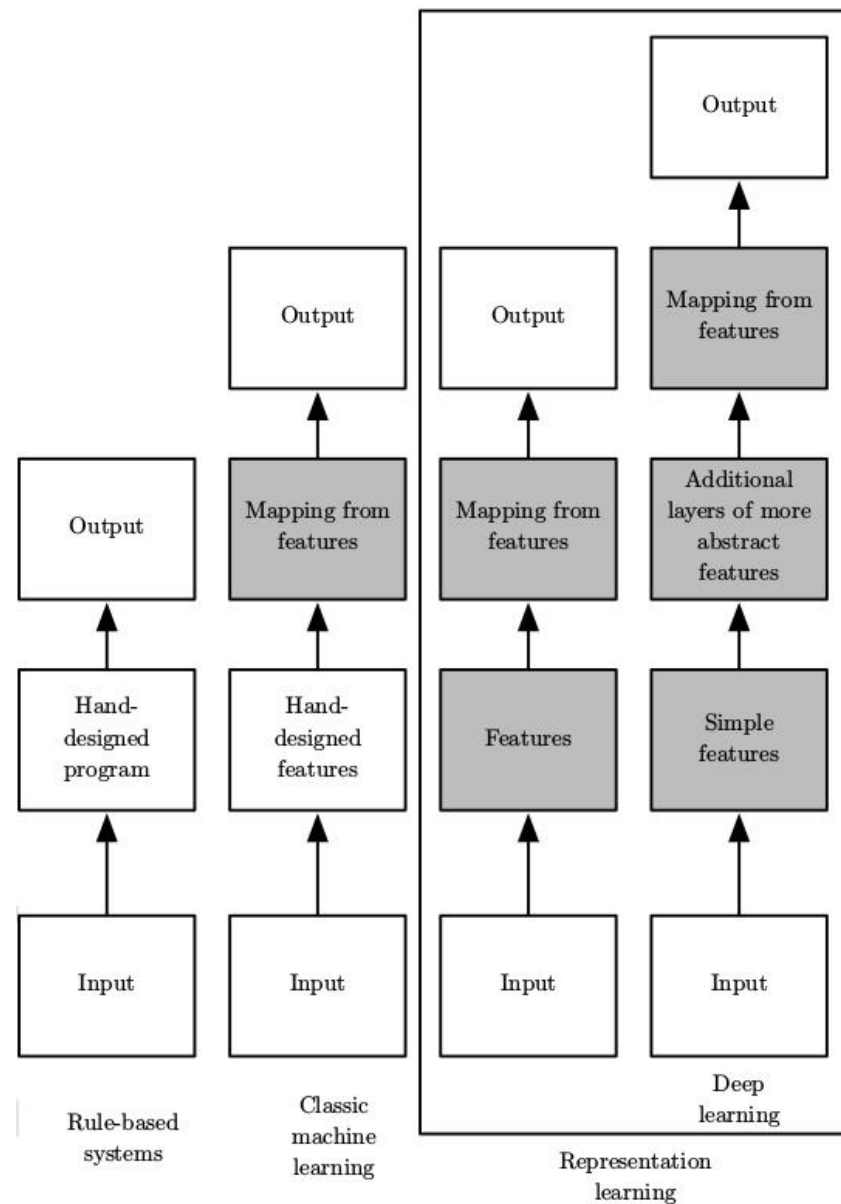
- 输入 -> 手工设计的特征 -> 从特征映射 -> 映射到输出

- 表示学习

- 输入 -> 机器自己生成特征 -> 从特征映射 -> 映射到输出

- 深度学习

- 输入 -> 机器提取简单特征 -> 组合为更加抽象特征 -> 从特征映射 -> 映射到输出



本课程内容

- Python 语言与常用库（1-3周）
- （基于 sklearn） 人工智能基本概念与常用术语，各种相对简单的算法（knn，线性算法，支持向量机，决策树，朴素贝叶斯，随机森林，主成分分析，简单全连接神经网络，...）（属于 经典机器学习）
- （基于 tensorflow+keras） 较复杂的架构（卷积神经网络，循环神经网络），实现深度学习（最后3-4周）（属于 深度学习）

人工智能的发展历史

- 迄今为止深度学习已经经历了3次发展浪潮：
 - 上世纪40-60年代，深度学习的雏形出现在控制论（cybernetics）；
 - 80-90年代，深度学习表现为联结主义（**connectionism**）；
 - 2006年，才真正以深度学习之名复兴。



2017年，深度学习程序AlphaGo 3:0 击败
世界排名第一的围棋选手柯洁，深度学习红遍世界

围棋虽然有形式化的规则，但问题本身太过复杂。对这样的问题，传统的形式化算法仍然无能为力



AI诞生

1956

达特茅斯会议

低谷

1970-1980

大规模数据和复杂任务不能完成，计算能力无法突破

专家系统

1982后

神经网络+5代计算机

低谷

1990-2000

DARPA无法实现，政府投入缩减

深度学习

2006-至今

突破性进展，进入发展热潮



1957

Perception 感知
神经网络的发明



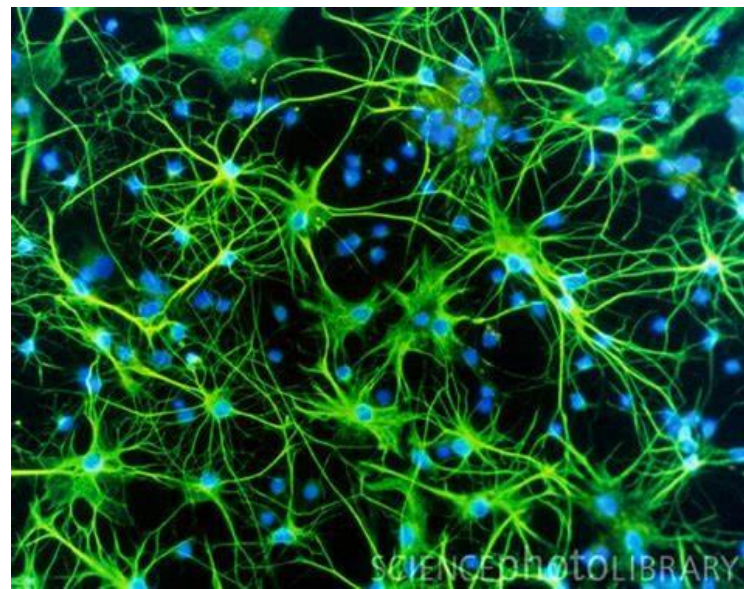
1986

BP 神经网络算法
的发明和应用



联结主义大行其道

- 联结主义的中心思想为，当网络将大量简单的计算单元连接在一起时，可以实现智能行为。
- 当动物的许多神经元一起工作时会变得聪明。单独神经元或小集合的神经元不是特别有用。



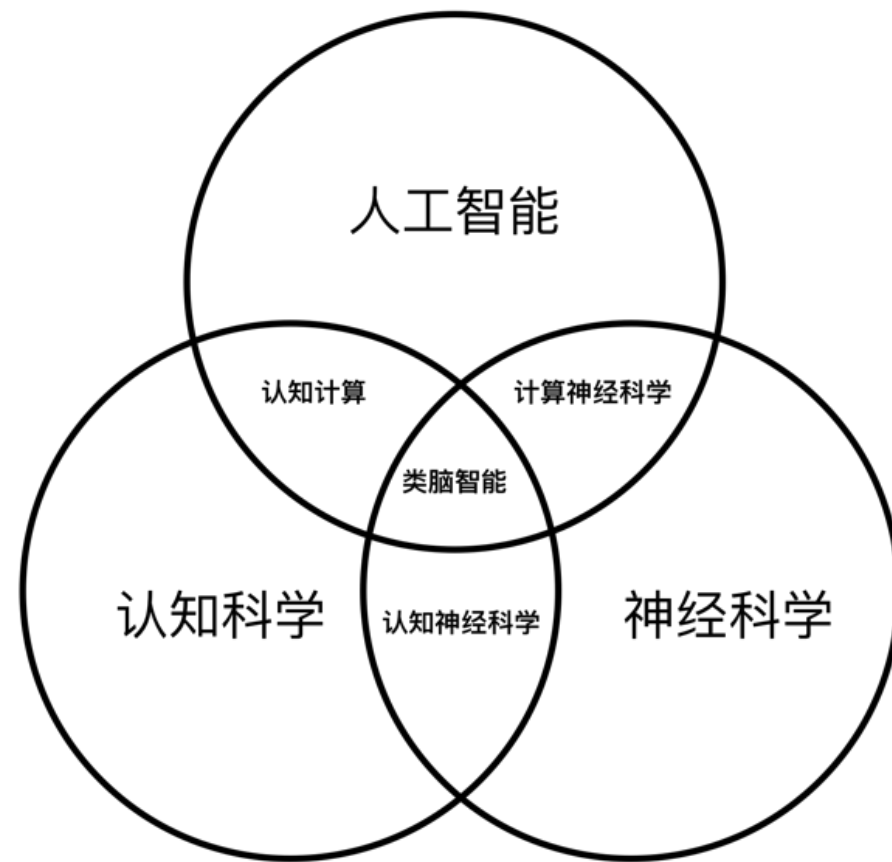
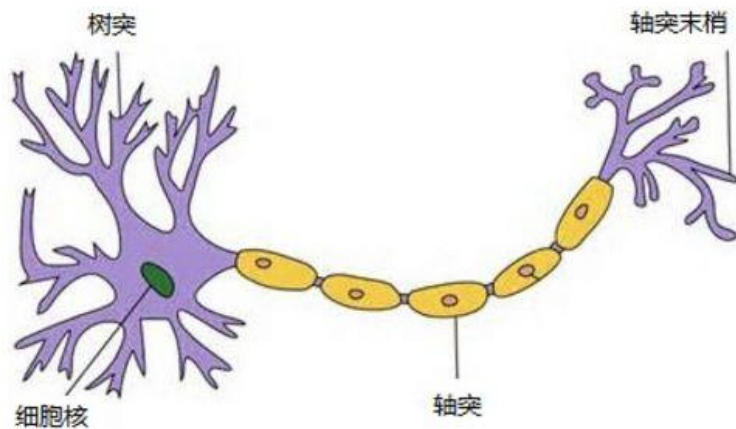
联结主义大行其道

- 人工网络神经元规模大概每2.4年扩大一倍，目前的规模还是比原始的脊椎动物（如青蛙）的神经系统要小。直到21世纪50年代，人工神经网络才能具备与人脑相同数量级的神经元。
- 生物神经元的功能可能比目前人工神经元更强大，因此人工智能达到人脑的水平还有相当距离。



与神经科学的关系

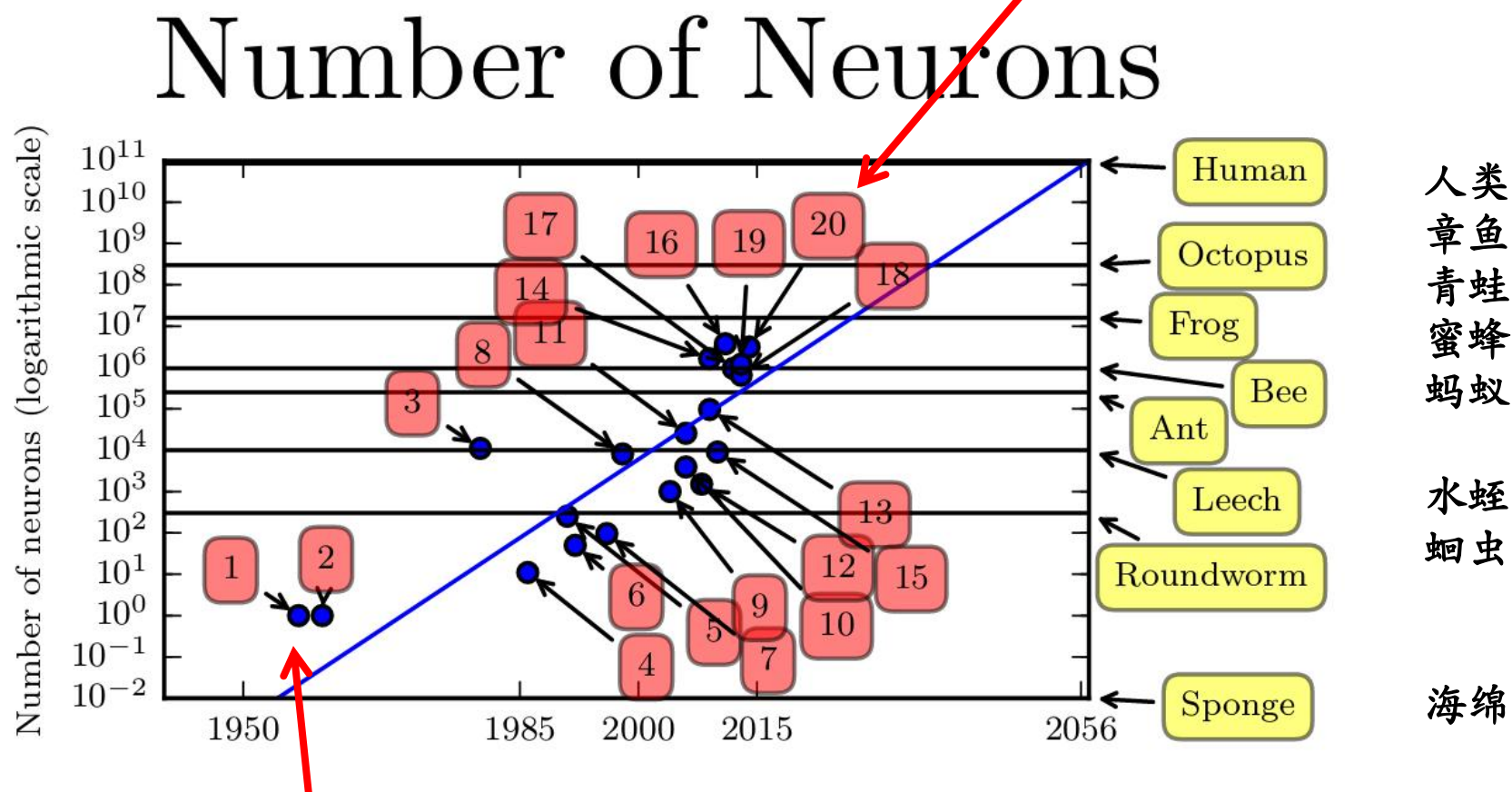
- 媒体经常强调深度学习与**大脑**的相似性。的确，神经网络的发展受到很多来自**神经科学**的启发。虽然现代的深度学习从更多领域获取灵感。



- 也有一些研究人员尝试利用深度学习研究大脑的工作原理，称为“**计算神经科学**”。

神经网络增长规模

20: GoogLeNet, Szegedy et al. 2014



1: 感知机, Rosenblatt, 1958, 1962

本次人工智能的兴起主要有以下几个特点





人工智能投资分析

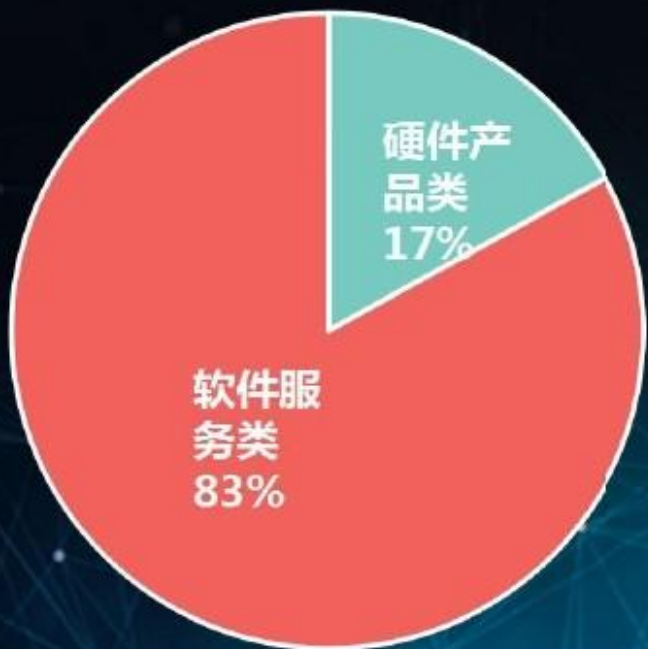
2012-2015年人工智能行业投资额及投资次数



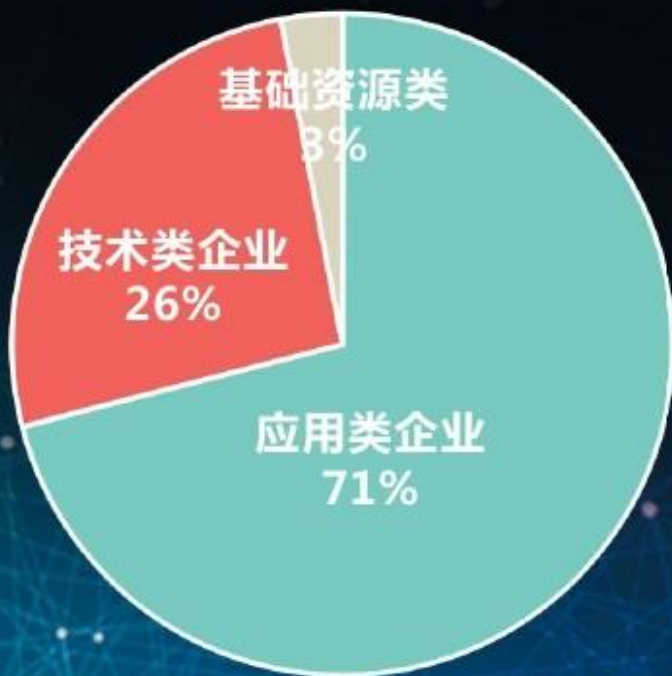
2012-2015年投资人工智能机构数量



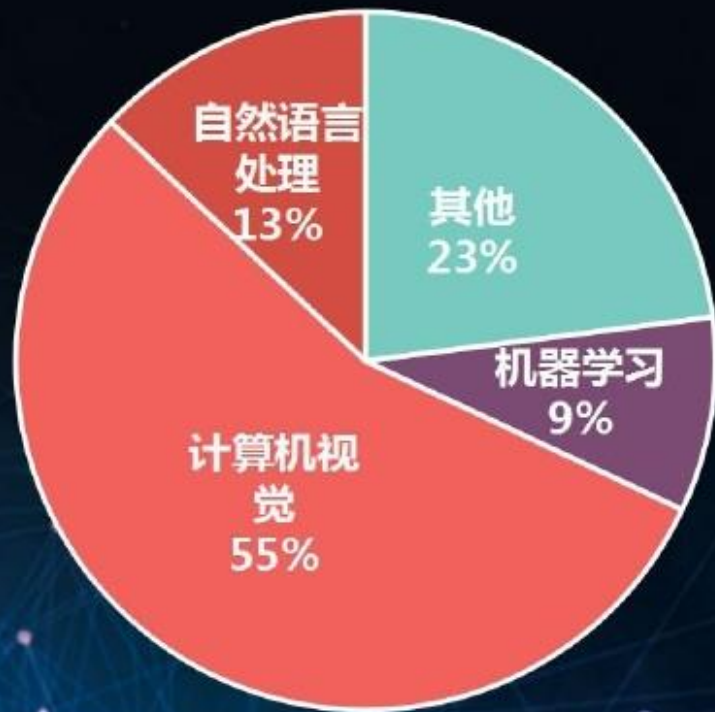
人工智能投资分析



软件，硬件比例



应用，技术比例



不同技术方向比例

2018人工智能全球高校排名

(仅基于论文数量)

排名前三的分别为：卡内基梅隆大学、清华大学、牛津大学

第4到第10名的学校分别为：北京大学、香港科技大学、新南威尔士大学、南洋理工大学、阿尔伯塔大学、南京大学、浙江大学。

Rank	Institution	Count	Faculty
1	▶ Carnegie Mellon University	74.2	40
2	▶ Tsinghua University	70.9	45
3	▶ University of Oxford	51.5	21
4	▶ Peking University	50.0	42
5	▶ HKUST	45.3	10
6	▶ UNSW	42.4	20
7	▶ Nanyang Technological University	42.1	22
8	▶ University of Alberta	41.7	17
9	▶ Nanjing University	40.5	20
10	▶ Zhejiang University	39.6	32
11	▶ University of California - Los Angeles	36.4	11
12	▶ Ben-Gurion University of the Negev	36.2	11
13	▶ Bar-Ilan University	27.3	11
14	▶ University of Southern California	26.6	16
15	▶ University of Liverpool	26.2	14
15	▶ University of Massachusetts Amherst	26.2	17
17	▶ University of Texas at Austin	26.0	14
18	▶ University of Washington	25.2	21
19	▶ Imperial College London	24.4	6
20	▶ Cornell University	24.3	12

学习本课程的建议

多运动



建议

- 使用自己的电脑，而非机房的电脑

- 方便 课下练习 + 结课后继续

- 自行安装 Anaconda3 （自带python以及各种模块；自带sklearn） + tensorflow + keras
 - 鼓励使用 Linux 系统



建议

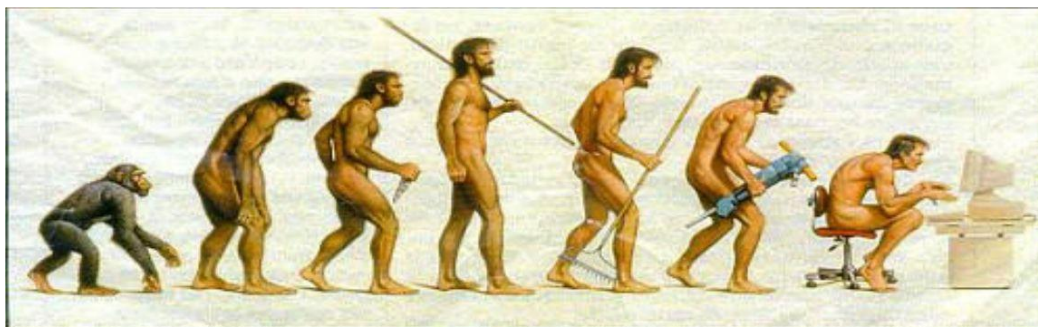
- 开放！灵活！放手折腾！
 - 没有绝对的对错，能解决问题的就是老大
 - 任何问题，首先百度、biying、Google（非常重要）



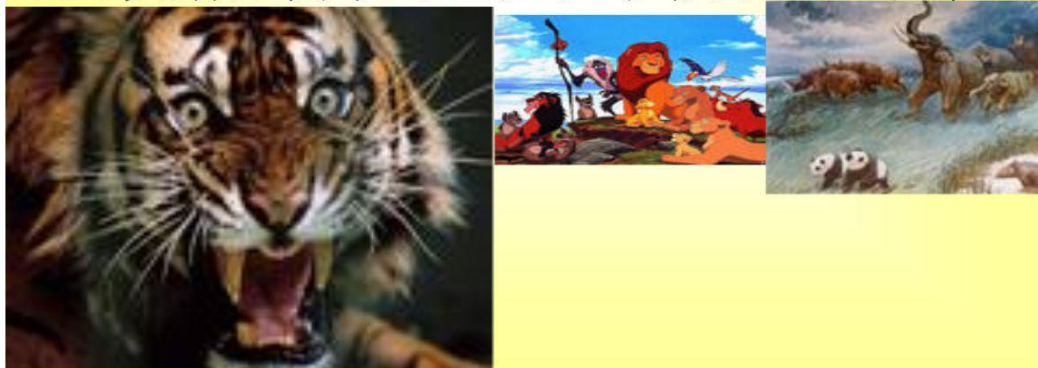
- 自己多折腾，做调参侠积累经验

建议

- 充分发挥主观能动性
 - 针对自己的专业方向，思考和寻找应用的场景，有针对性地实践



主观能动性是人与物的区别



建议

- 结合我国人工智能的战略发展规划，为我国成为真正的科技强国作出
一分贡献！

