Machine Learning

机器学习

大纲

- 课程简介
- 人工智能简史
- 序章

个人简介

课程简介

课程内容

- 机器学习中的经典算法
- 机器学习实战
- 详情见教学大纲和进度表

- 加强利用机器学习方法解决实际问题的能力
- 为后续课程打好基础

课堂形式

- 理论教学
- 上机实践

考核方式

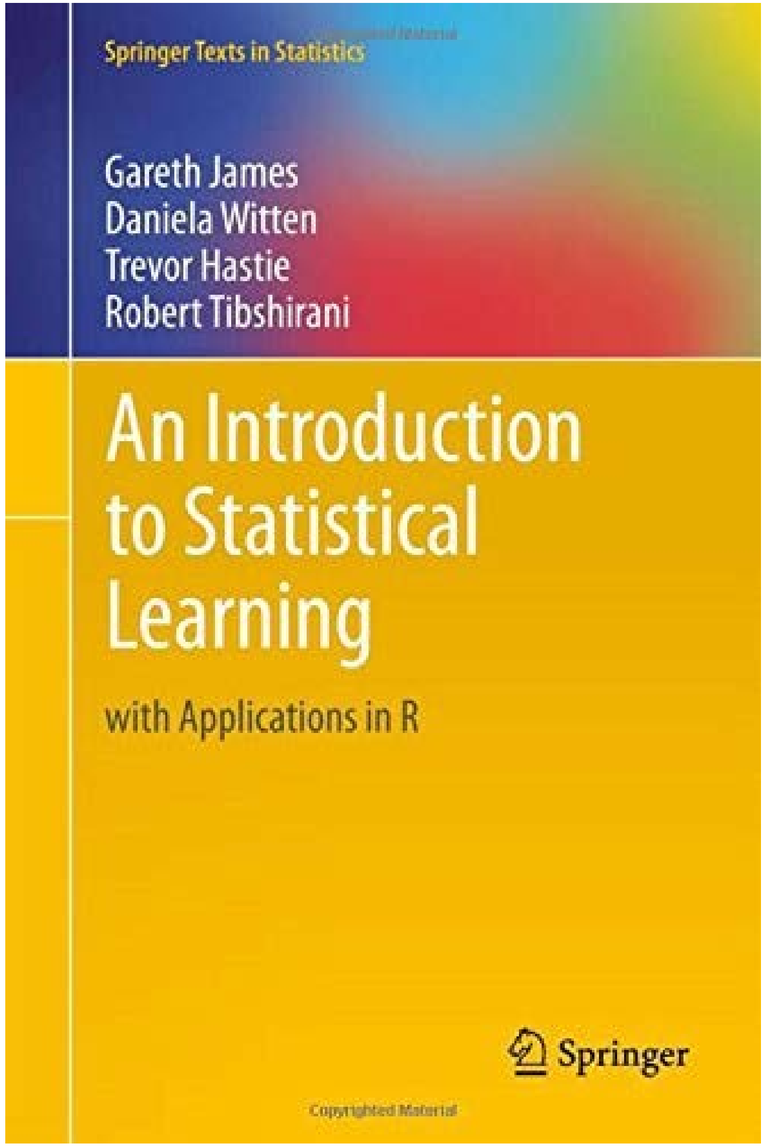
- 考勤——10%
- 作业——30%（包含一次小组作业，每组人数小于等于3）
- 期末闭卷考试——60%

- 考试题型：选择题、填空题、判断题、简答题、计算题、证明题……
- 考试范围：上课内容

参考书目

- An Introduction to Statistical Learning, with Applications in R
- James, Witten, Hastie, Tibshirani著
- 下载链接：

https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf



课外读物

- Andrew Ng 吴恩达
  - CS229
  - DeepLearning.AI

Andrew Ng

- 周志华——西瓜书

- 感受计算机和统计对机器学习理解的差异性与一致性

History of AI

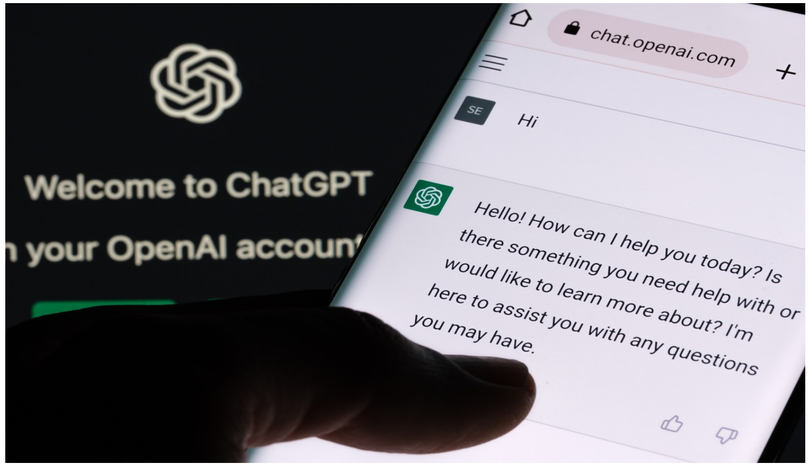
人工智能简史

影视作品中的AI

实际生活中的AI


ALPHAGO
00:05:30
Google DeepMind
Challenge Match
LEE SEDOL
00:28:34
知乎 @shamesdi


Hey Siri


Welcome to ChatGPT
your OpenAI account
chat.openai.com
Hi
Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any questions you may have.

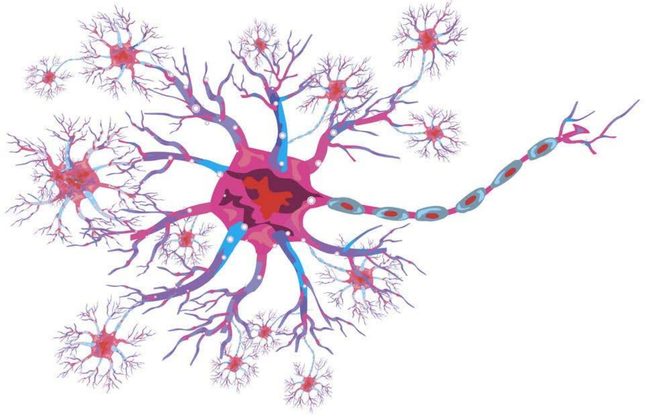
谷歌2022开发者大会

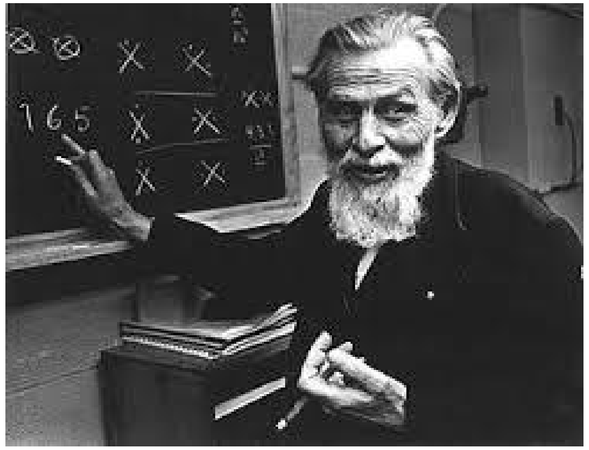
- 我们来看看世界前沿的AI技术是如何落地的：
- https://www.bilibili.com/video/BV1pA4y1Z7Kg?p=1

- 人工智能已经从方方面面深入影响着我们的生活
  - 网约车平台派单
  - 个性化推荐
  - 医疗影像诊断
  - AIGC
  - ……

起源

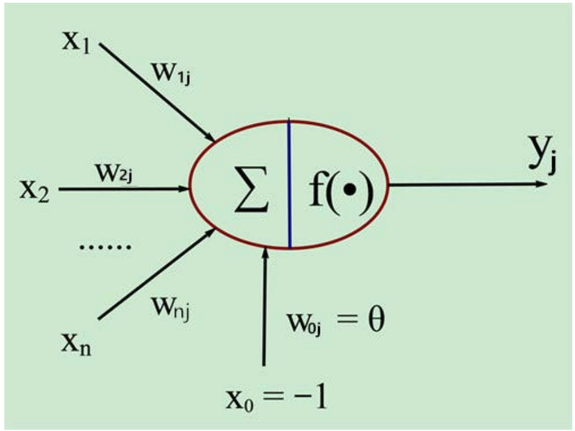
- 1943年，美国神经科学家McCilloch和数学家Pitts在《数学生物物理学公告》上发表论文《神经活动中内在思想的逻辑演算》。
- 按照生物神经元的结构和工作原理，他们构造了一个抽象和简化的模型，即MP模型。
- 人工神经网络的大门由此开启。





Warren McCulloch (1898-1969)

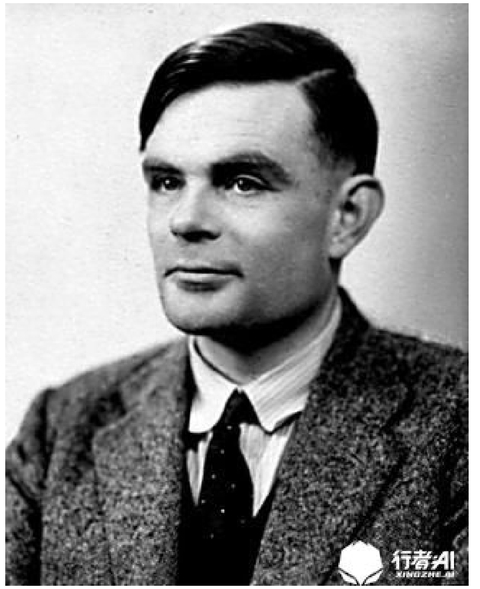
Walter Pitts (1923-1969)



MP模型

图灵测试(Turing Testing)

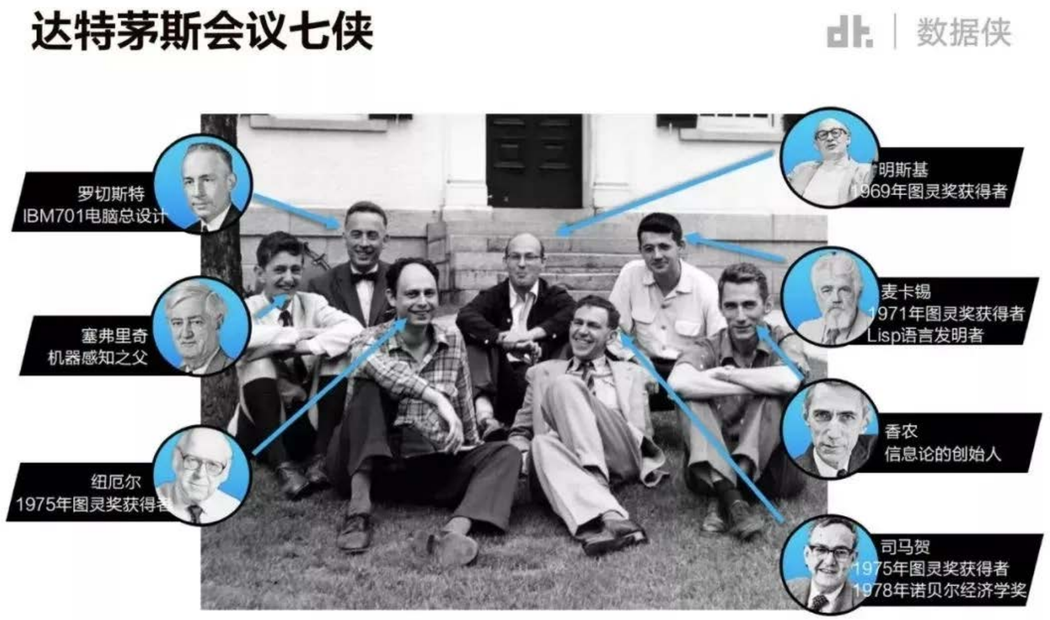
- 1950年，图灵发表论文《计算机器与智能》，预言了创造出**具有真正智能的机器**的可能性。
- 由于“智能”这一概念难以确切定义，提出著名的**图灵测试：**如果一台机器能够与人类展开对话而不能被辨别出其机器身份，那么称这台机器具有智能。
- 1952年，图灵提出一个更具体的想法：让计算机来冒充人，如果判断正确的人不足70%，那么可以判断计算机具有人类智能。
- 直到2014年才有聊天机器人勉强通过图灵测试。
- 图灵测试自诞生来产生了巨大影响，图灵奖被称为“计算机界的诺贝尔奖”，图灵也被冠以“人工智能之父”的称号。



Alan Mathison Turing，1912-1954

人工智能(Artificial Intelligence, AI)诞生

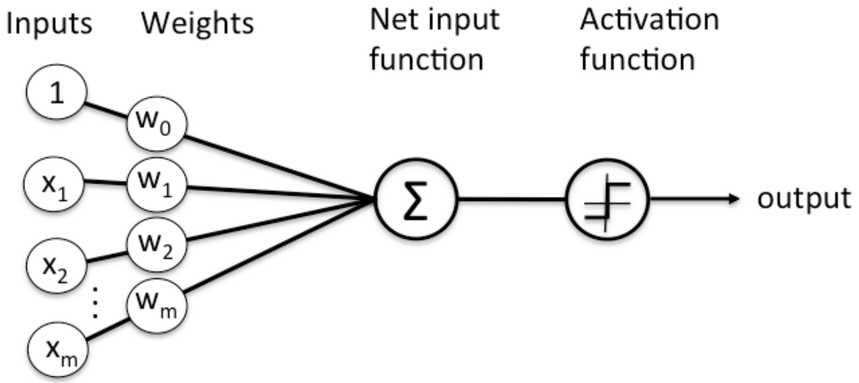
- 人工智能的起源被公认为1956年的达特茅斯会议：人工智能夏季研讨会(Summer Research Project on Artificial Intelligence)
- 会议主要议题：自动计算机、自然语言处理、神经网络、计算规模理论、自我改造、抽象、随机性与创造性。
- 与当今的人工智能议题并无太大差别！
- 会议足足开了两个月的时间，虽然没有达成普遍的共识，却起了一个名字：人工智能。

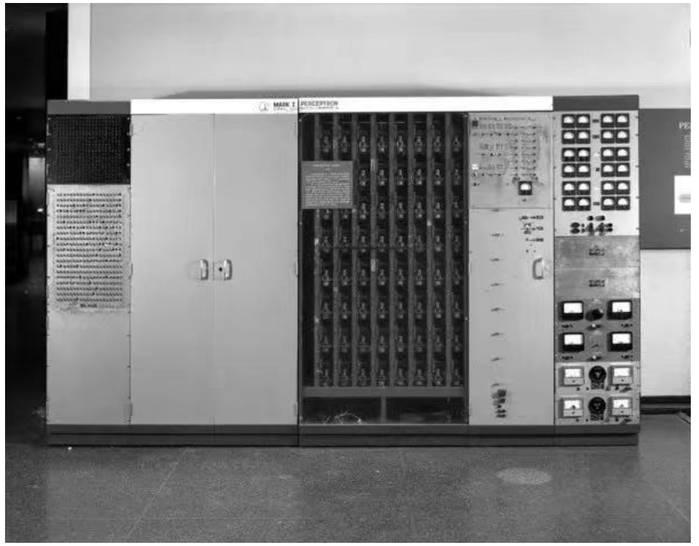


达特茅斯学院旧址

感知机(Perceptron)

- Frank Rosenblatt (1958)《The perceptron: a probabilistic model for information storage and organization in the brain》提出了可以模拟人类感知能力的机器。
- 相当于一个线性分类器：$y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$。
- 和MP模型的最大区别：使用梯度下降法实现了对参数$\boldsymbol{w}$的学习。
- 1963年证明了对于二分类问题，如果训练数据集是线性可分的，那么感知机可以在有限次迭代后收敛。
- 理论实践效果引发了第一次人工神经网络的浪潮。





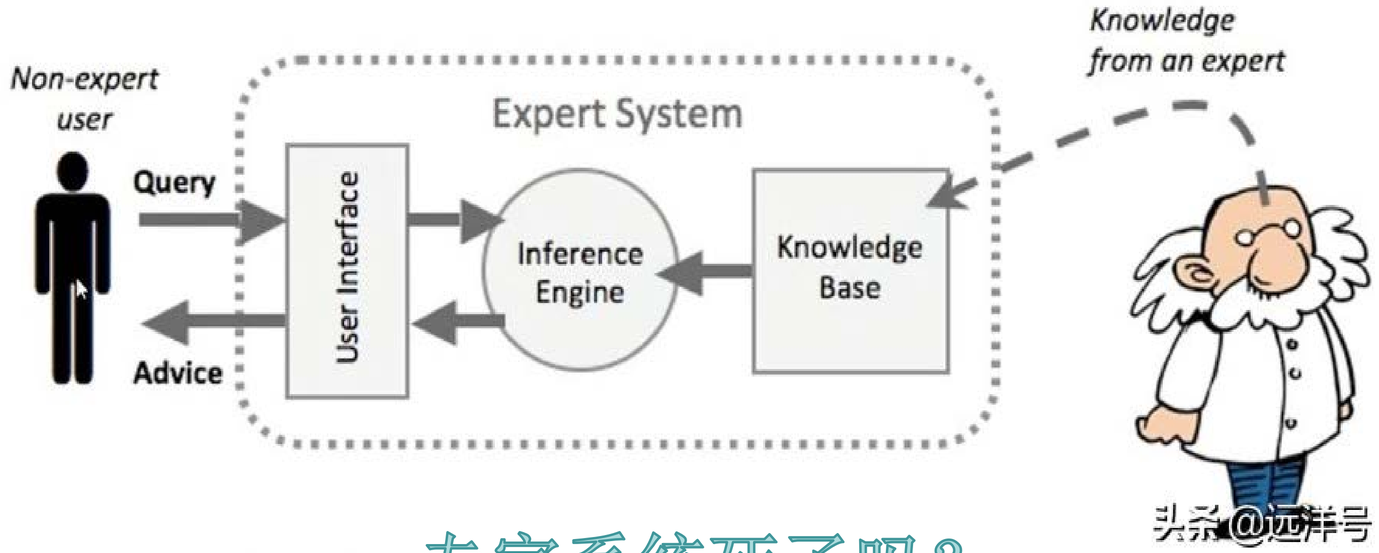
Mark I 感知机

质疑与第一次寒冬

- 1969年Minsky在《Perceptrons》一书中，仔细分析了以感知机为代表的单层神经网络系统的功能及局限，证明感知机不能解决**异或（XOR）**等线性不可分问题。
- Rosenblatt和Minsky等人在当时已经意识到：多层感知机能够解决线性不可分的问题。遗憾的是，没有人能够及时推广感知机到多层神经网络。
- 随着数字计算机的发展、经费的枯竭、以及人们对《Perceptrons》一书的误解等多种因素，人工智能迎来了第一次寒冬，神经网络的研究也陷入了将近20年的停滞。

另一股风潮——专家系统

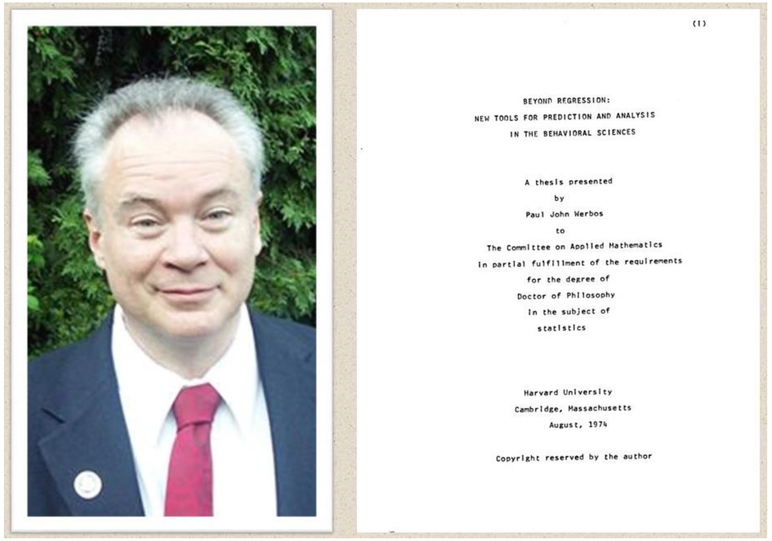
- 20世纪60年代，数字计算机的飞速发展，人们误以为数字计算机可以解决人工智能问题，感知机被暂时抛弃。**专家系统**主导了人工智能的开发。
- 由 *if-then* 场景构建的大型网络来模仿人类思想和决策，过滤查询以实现一些预编程的最终结果。
- **专家系统的缺点：**
  - 需要设计大量的规则
  - 需要领域专家
  - 可移植性差
  - 学习能力差
  - 人能考虑的范围有限



专家系统死了吗？

转机——反向传播算法(Backpropagation Algorithm)

- 1974年Paul Werbos在其博士论文中提出反向传播算法，来解决多层神经网络的学习问题，但并未受到重视。
- 1986年，BP被重新发现：深度学习教父Geoffrey Hinton发明了适用于多层感知机的BP算法，并采用Sigmoid函数进行非线性映射，有效解决了非线性分类和学习的问题。
- “在我之前，很多人提出了不同版本的反向传播。其中大部分是独立提出的，我觉得我承受了过多的赞誉。我看到媒体说我提出了反向传播，这是完全错误的。科研人员认为他因为某事获得了过多赞誉，这样的情况不常见，但这就是其中之一！我的主要贡献是展示如何使用 BP 算法学习分布式表征，因此我要做出澄清。”——Hinton



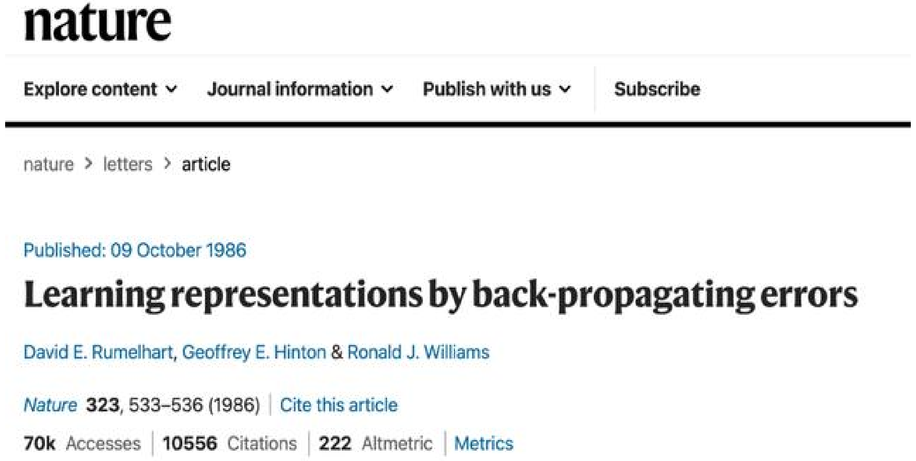
(1)

BEYOND REGRESSION:
NEW TOOLS FOR PREDICTION AND ANALYSIS
IN THE BEHAVIORAL SCIENCES

A thesis presented
by
Paul John Werbos
to
The Committee on Applied Mathematics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
statistics

Harvard University
Cambridge, Massachusetts
August, 1974

Copyright reserved by the author



nature

Explore content ⌄ Journal information ⌄ Publish with us ⌄ Subscribe

nature › letters › article

Published: 09 October 1986

Learning representations by back-propagating errors

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams

Nature **323**, 533–536 (1986) | Cite this article

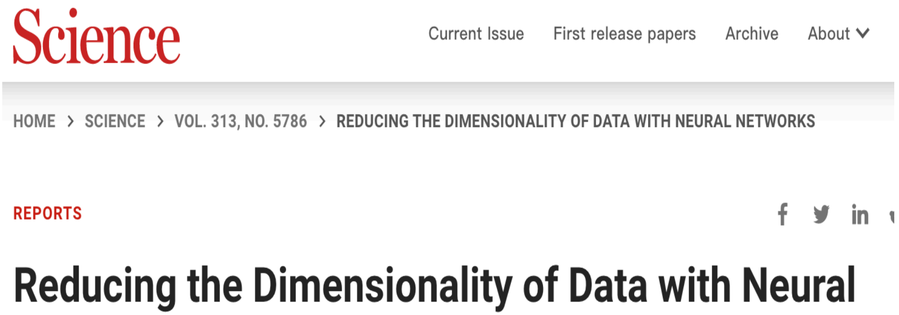
70k Accesses | **10556** Citations | **222** Altmetric | Metrics

第二次寒冬&百花齐放

- 多层神经网络解释性差，计算速度慢，并在1991年被指出BP算法存在梯度消失问题
- 1986年，著名的**决策树**算法诞生
- 1995年，**AdaBoost**（Adaptive Boosting）算法诞生，集成学习兴起，相继诞生了**Random Forest**（2001）、**XGBoost**（2015）、**LightGBM**（2017）等经典算法
- 同样在1995年，机器学习领域中一个最重要的突破**支持向量机**（Support Vector Machines，SVM）被提出。从此将机器学习社区分为神经网络社区和支持向量机社区。
- SVM以统计学为基础，和神经网络有明显的差异，在以前许多神经网络模型不能解决的任务中取得了很好的效果。
- 神经网络再次陷入寒冬…

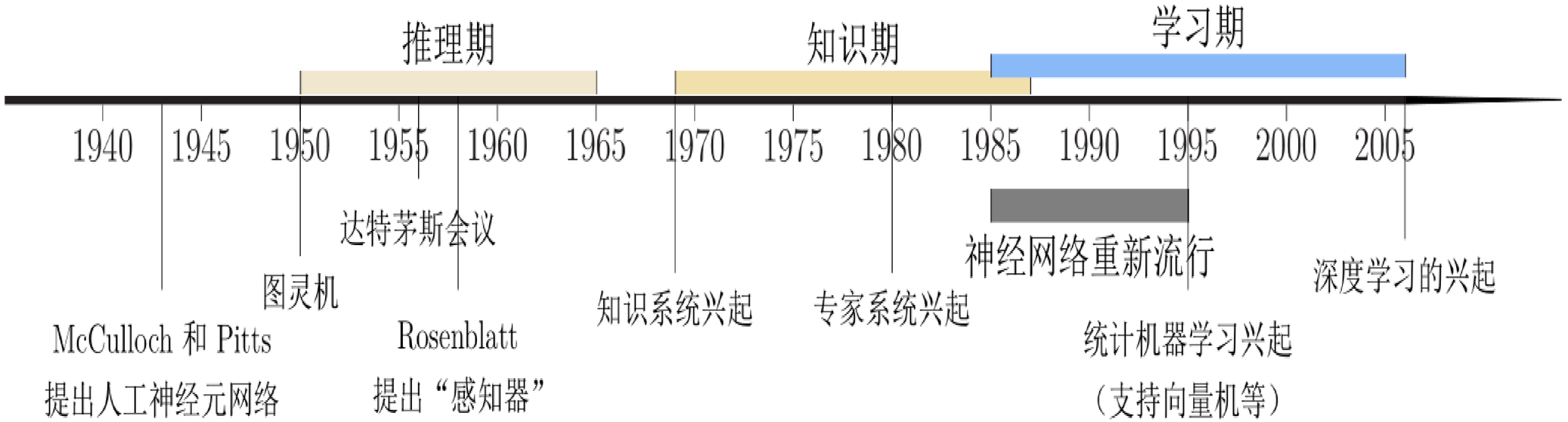
深度学习(Deep Learning)

- 2006年，Hinton在《Science》上发表了一篇文章，开启了深度学习在学界和业界的浪潮。
- 主要观点：1. 有很多隐含层的人工神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类；2. 深度神经网络在训练上的难度，可以通过“逐层初始化”来有效克服
- 海量数据的积累
- GPU等尖端硬件设备的普及
- Tensorflow等软件极大降低入门门槛
- 神经网络研究者的大力坚持



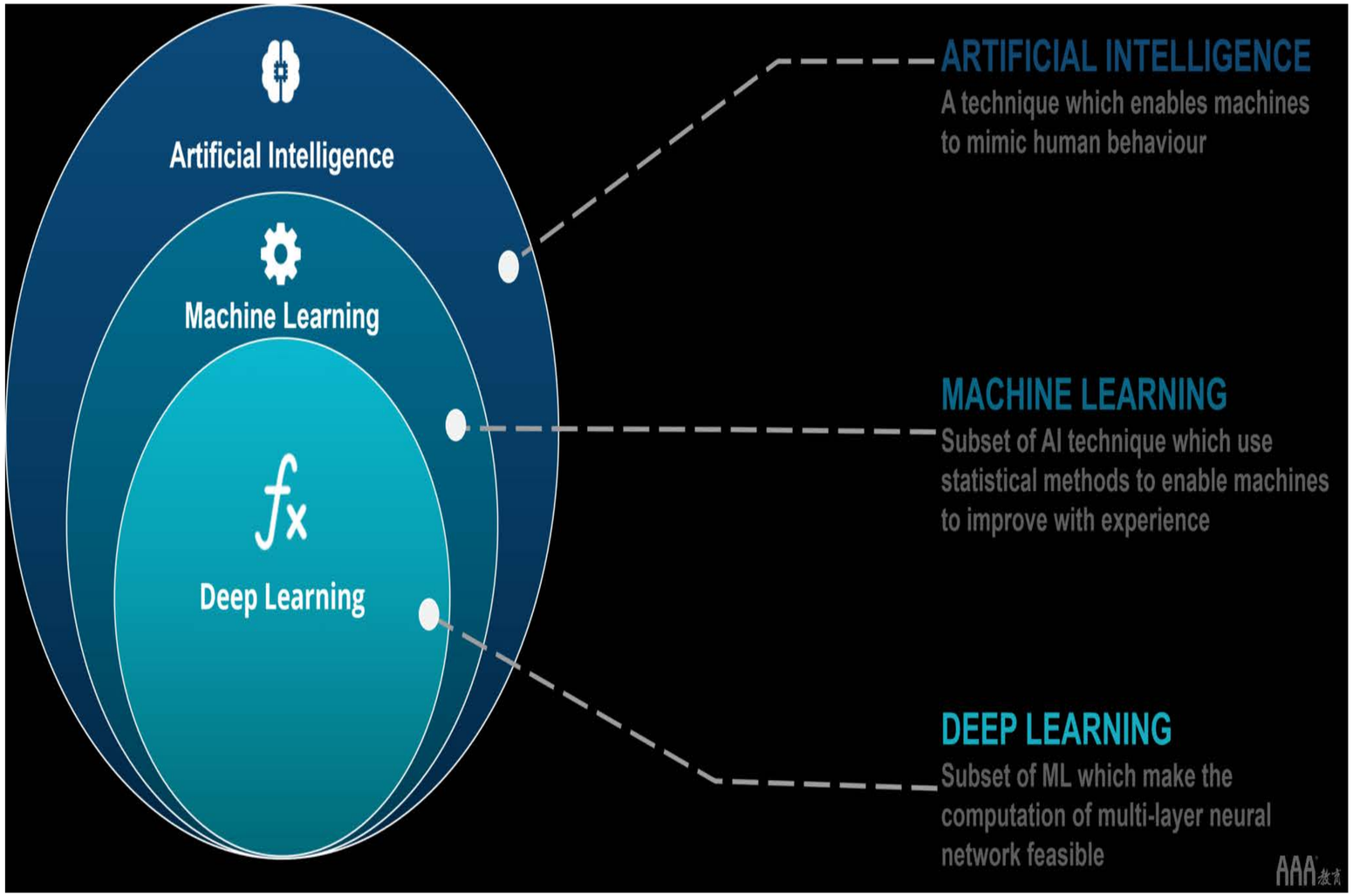
2018年图灵奖得主

一段波澜壮阔的历史


推理期
知识期
学习期
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
达特茅斯会议
图灵机
神经网络重新流行
深度学习的兴起
知识系统兴起
专家系统兴起
统计机器学习兴起
（支持向量机等）
McCulloch 和 Pitts
提出人工神经元网络
Rosenblatt
提出“感知器”

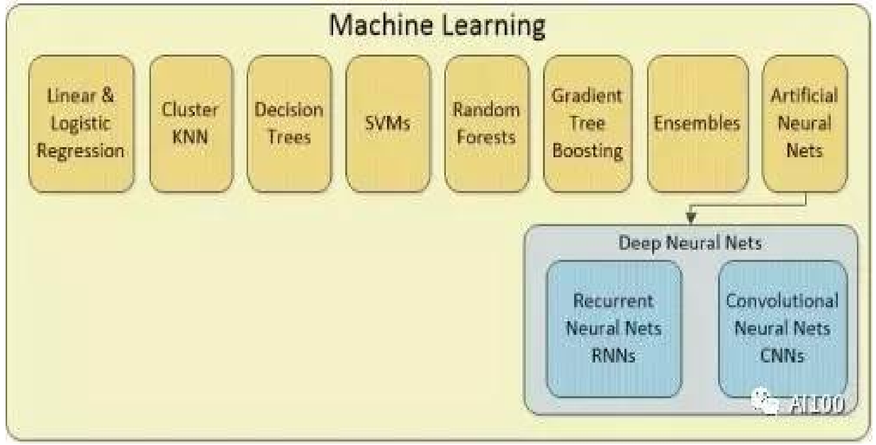
AI & Machine Learning & Deep Learning

- 人工智能是使机器模仿人类行为的技术。
- 机器学习是一类人工智能技术，应用于数据集中以寻找某种数据模式的所有电脑算法的统称。
- 深度学习是一类机器学习方法，使得深层神经网络的计算变得可行。
- **深度学习是否已经让传统的机器学习变得无关紧要？**



Machine Learning VS. Deep Learning

- 首先，深度学习并不是独立于机器学习的，而是其中的一个分支。
  - 如果你想成为图像和语言处理方面的专家，你可以试着在深度学习方面深入探索，尤其是在循环神经网络和卷积神经网络这两个分支。但是，这并不意味着其他方面的研究没有价值。
- 深度学习在数据分析竞赛网站Kaggle的压倒性优势
  - 非结构化的数据占大多数，而这是深度学习的强项
  - 排名靠前的方案成败在毫厘之间。但是，这与你在商业中遇到的大多数数据科学问题毫不相关。倾注大量心血以求得准确性的少量提高，对于提升商业经济的发展来说是不现实的。
- 深度学习的问题
  - 很难训练，有时甚至不能被训练。
  - 如果你正在建立一个深层神经网络，那么你的程序调试时间很可能会花费数周甚至是数月。
  - 需要极其大量的标记数据来实现其训练过程，这对很多公司来说非常困难或者成本太高。



真实的数据科学市场

- 第一类：想要并且需要最前沿的数据科学技术，将其与用户结合起来，从而使得自己与竞争对手区分开来。
  - 代表企业：谷歌、OpenAI、阿里达摩院、腾讯AI lab……
  - 如果你想专门研究深度学习、非结构化文本和图像数据，那么你就需要到这些地方去学习。
- 但是，当下超过80%的数据科学市场仍是对消费者行为的预测
  - 消费者为什么会来、为什么会停留、为什么会离开、下次会买什么或者下次很可能会买什么。
  - 它存在于所有的面向客户的系统中，为客户推荐要购买的商品、解决问题的方法。
  - 上述所有的应用程序，都需要在传统的机器学习方法的协调配合下，才能有效快速地运行。
- 所以，深度学习不会使传统的机器学习方法过时、无用。
- 相应地，要想成为一名数据科学家，你必须首先掌握传统机器学习的方法。

Overview

序章

什么是机器学习？

- Arthur Samuel (1959)：机器学习是一个研究领域，它赋予计算机无需明确编程就可以学习的能力。
- Tom Mitchell (1998)：如果计算机程序在任务 T 中的性能（由 P 测量）随着经验 E 的引入而提高，则可以说计算机程序可以从关于任务 T 和性能测度 P 的经验 E 中学习。
- 机器学习可以看作是一个构造算法的过程：通过探索训练数据集来构造统计模型，并试图找到一个最优的函数用于映射从输入到输出的过程。
- 实际中的机器学习问题：“Is this cancer?”，“Which of these people are good friends with each other?”，“Will this person like this movie?”

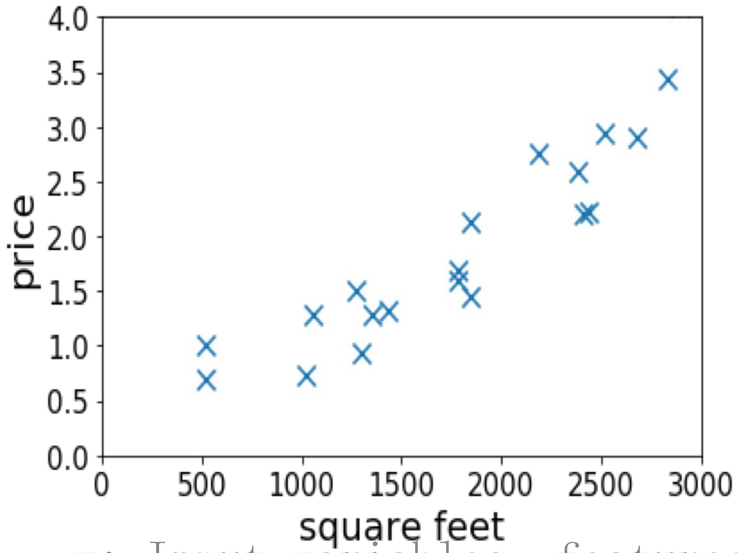
机器学习问题分类

- 有监督学习 (Supervised learning)
  - 既给予“特征信息”又反馈“结果信息”的机器学习问题
- 无监督学习 (Unsupervised learning)
  - 只给予“特征信息” 的机器学习问题
- *强化学习 (Reinforcement learning)
- *半监督学习 (Semi-Supervised learning)

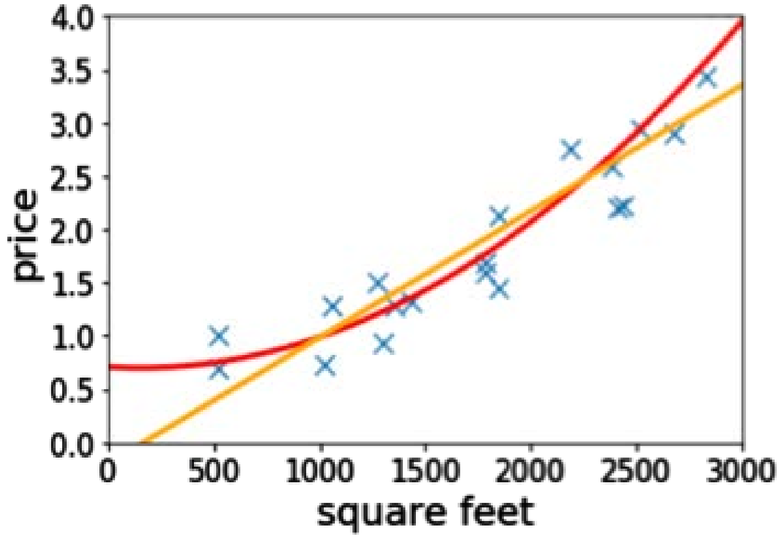
有监督学习案例：房价预测

- 数据：n个房屋的价格与套内面积信息$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$
- 目标：建立模型，基于套内面积预测房屋价格

y: Output variable, response variable



x: Input variables, features, covariates

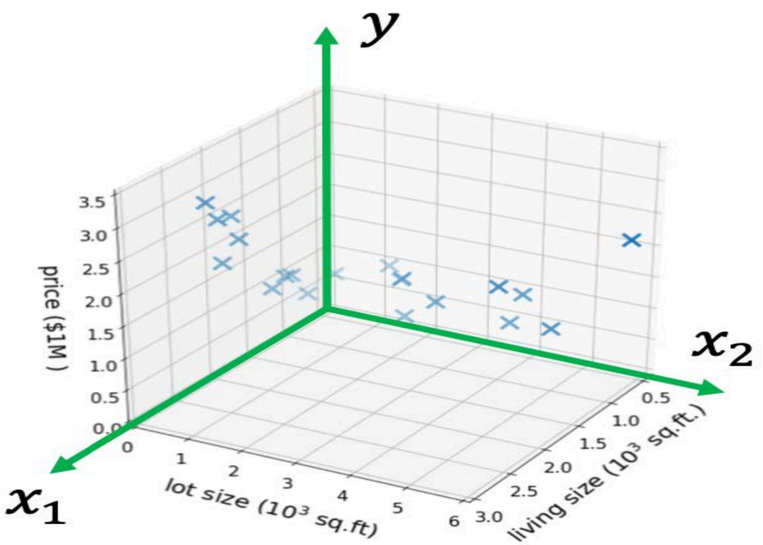


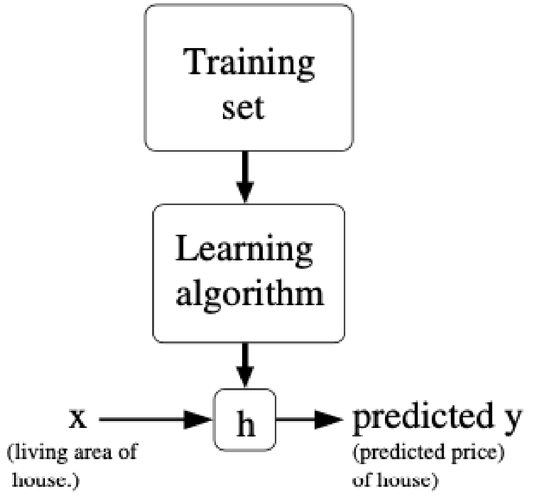
房价预测：更多特征

- 除了房屋套内面积外，我们也获取了占地面积
- 数据：$\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$，其中 $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)^T$
- 目标：找到一个映射函数

  （套内面积，占地面积）→（价格）

- 实际应用中，特征的维数可能会很高

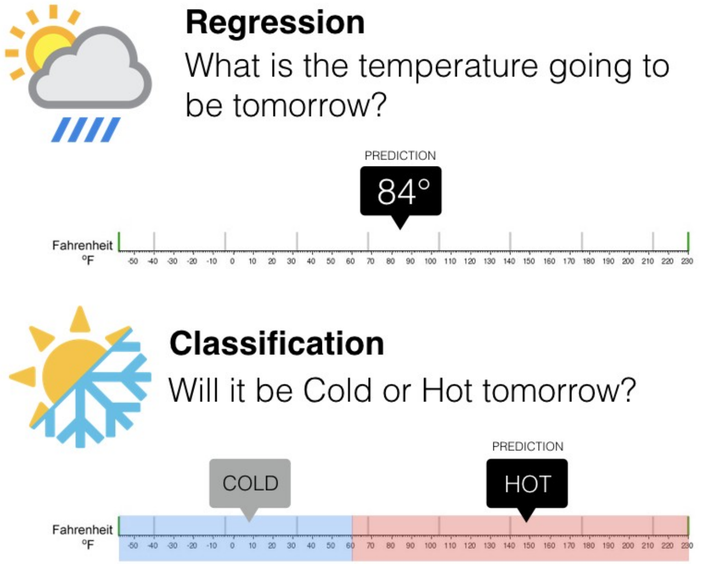


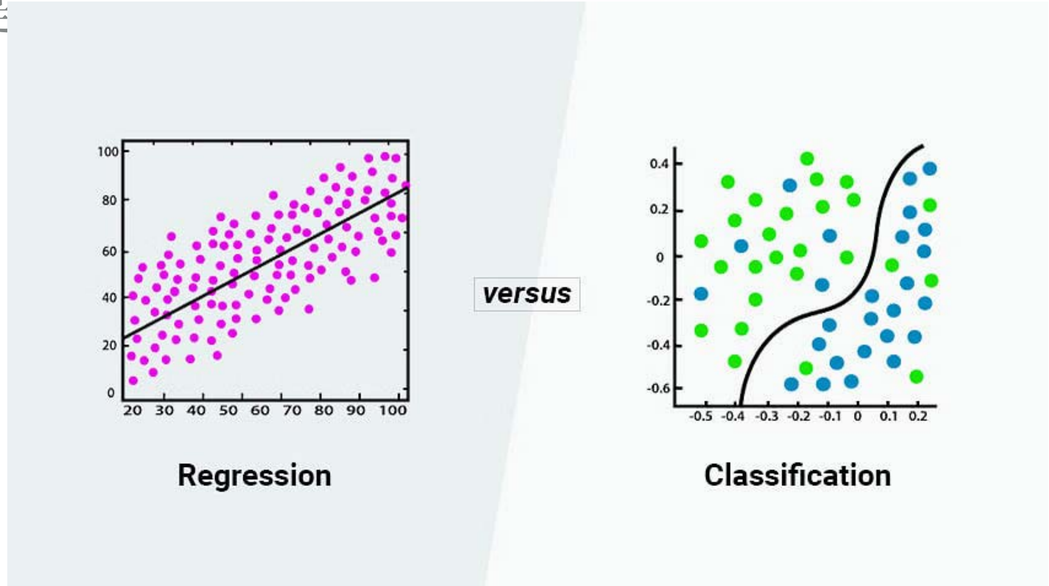


有监督学习：回归与分类

- 回归（Regression）：　　响应变量y是连续变量

- 分类（Classification）：响应变量y是离散变量
  - 例：　给定(套内面积，占地面积)，判断房屋类型是独栋别墅还是联排别墅
  - 此时的特征不变，响应变量y取值为 0/1
  - 多分类问题：响应变量y取值为 0/1/2…



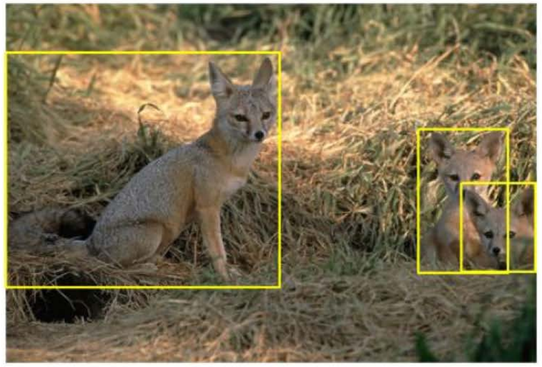


有监督学习——计算机视觉

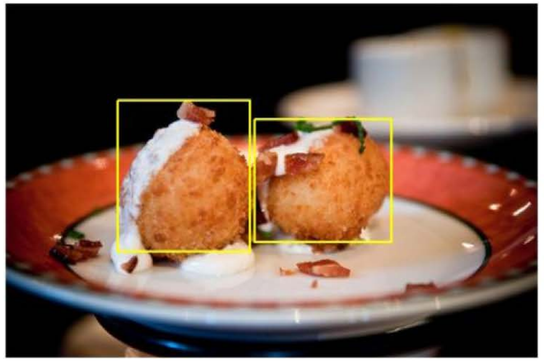
- 图像分类
- x =图像的原始像素，y =图像中的主要目标物

有监督学习——计算机视觉

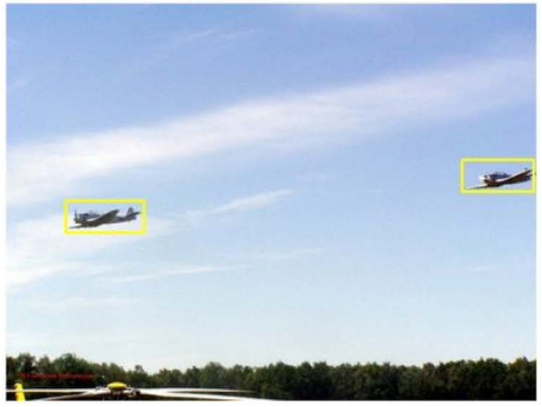
- 目标物检测
- x =图像的原始像素，y =目标物方框



kit fox



croquette

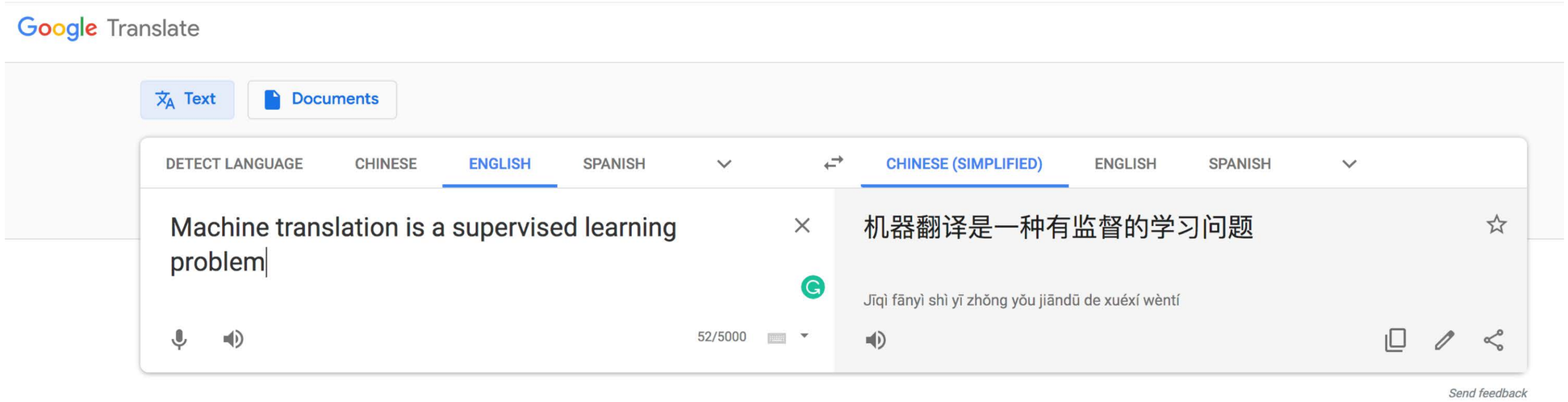


airplane

frog

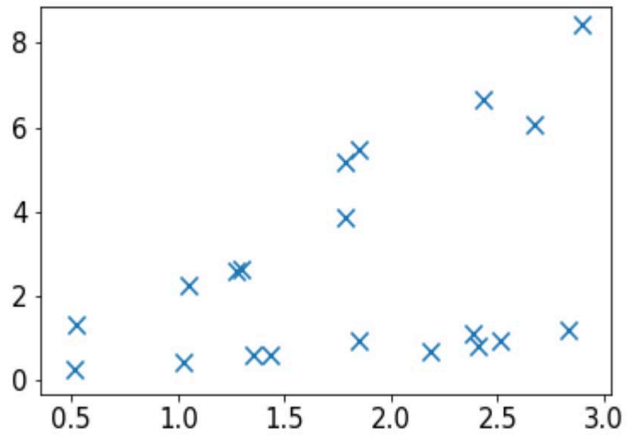
有监督学习——自然语言处理

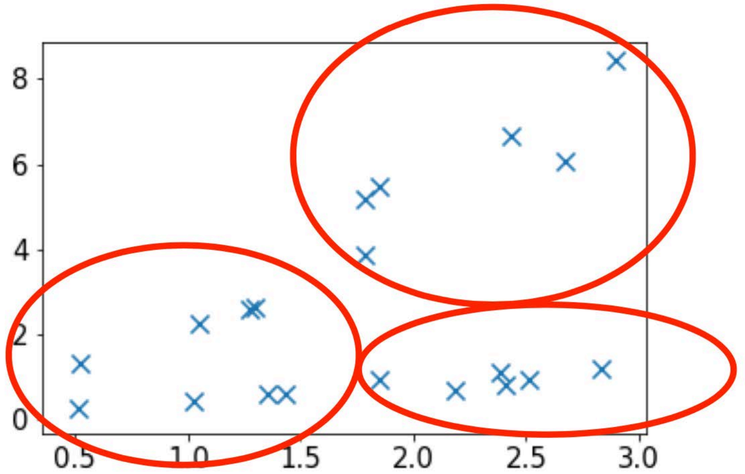
- 机器翻译
- x =英文，y =中文

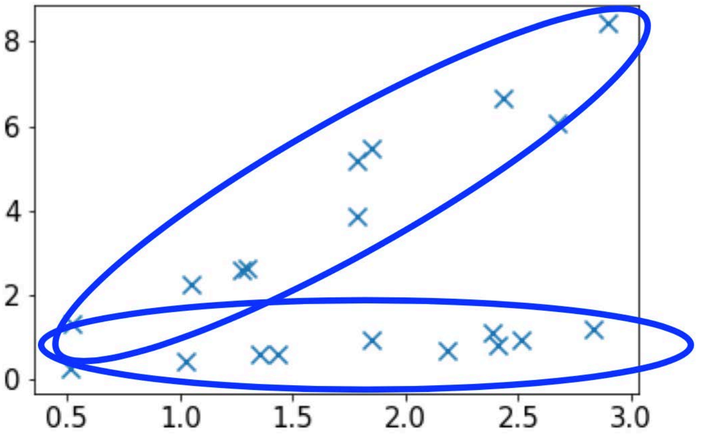


无监督学习

- 有监督学习：训练数据为$\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$
- 无监督学习：训练数据为$\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$
- 目标： find interesting structures in the data
- 聚类（Clustering）：基于训练数据确认观测是否落在相对有区分的子组里。





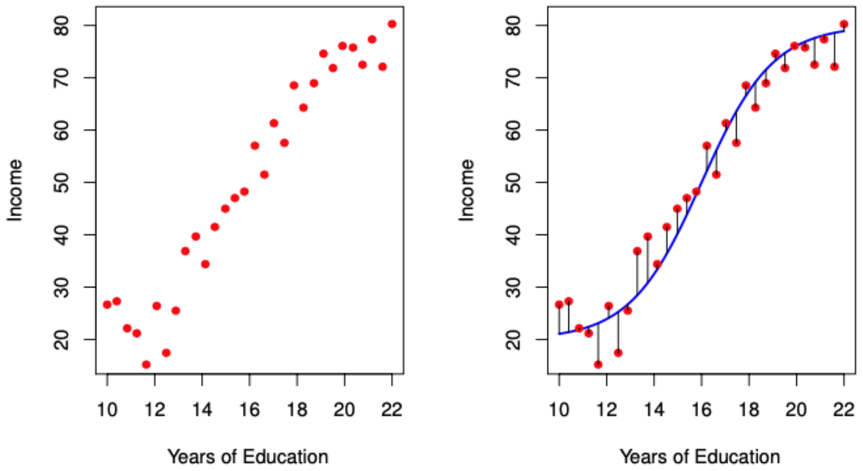


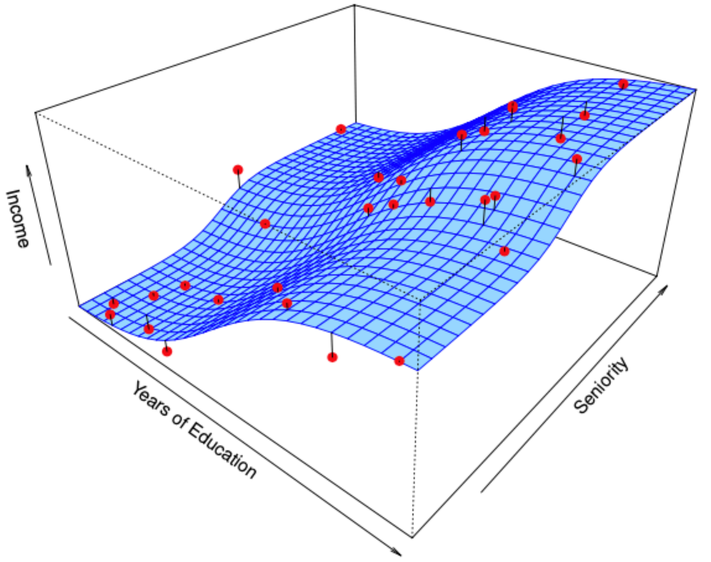
有监督学习：回归

- 假设响应变量Y和d个协变量$\boldsymbol{X} = (X_1, \ldots, X_d)$之间满足如下关系：

$$Y = f(\boldsymbol{X}) + \epsilon.$$

- f未知，ϵ为随机误差项。
- <u>核心目标：基于观测值，估计$\boldsymbol{f}$</u>





为什么估计f

- **预测（Prediction）**：给定$\boldsymbol{X}$预测Y。
- $\hat{Y} = \hat{f}(\boldsymbol{X})$，$\hat{f}$为$f$的估计，$\hat{Y}$即为$Y$的预测。
- 预测任务不在意$\hat{f}$的具体形式（black box），只要能得到准确的预测。
- 预测准确度取决于两方面：可约误差（reducible error）与不可约误差（irreducible error）。

$$E\left(Y-\hat{Y}\right)^2 = E\left[f(\boldsymbol{X})+\epsilon-\hat{f}(\boldsymbol{X})\right]^2 = E\underbrace{\left[f(\boldsymbol{X})-\hat{f}(\boldsymbol{X})\right]^2}_{\text{reducible}} + \underbrace{Var(\epsilon)}_{\text{irreducible}}$$

为什么估计f

- **推断**（Inference）：理解Y与$\boldsymbol{X}$之间的关系。
- 想要估计f，但目标却不是对Y做预测。
  - 哪一个特征与响应变量Y有关系？
  - 每个特征与响应变量Y的关系是什么？
  - Y与$\boldsymbol{X}$的关系可以用一个线性方程准确描述吗？还是需要更加复杂的模型？
- $\hat{f}$此时不能被当做黑箱，模型的可解释性变得重要。

- **预测+推断**：预测Y的同时理解Y与$\boldsymbol{X}$的关系
- 例：建立房屋价格与犯罪率、房屋面积、房屋是否带泳池、社区平均收入等因素的联系

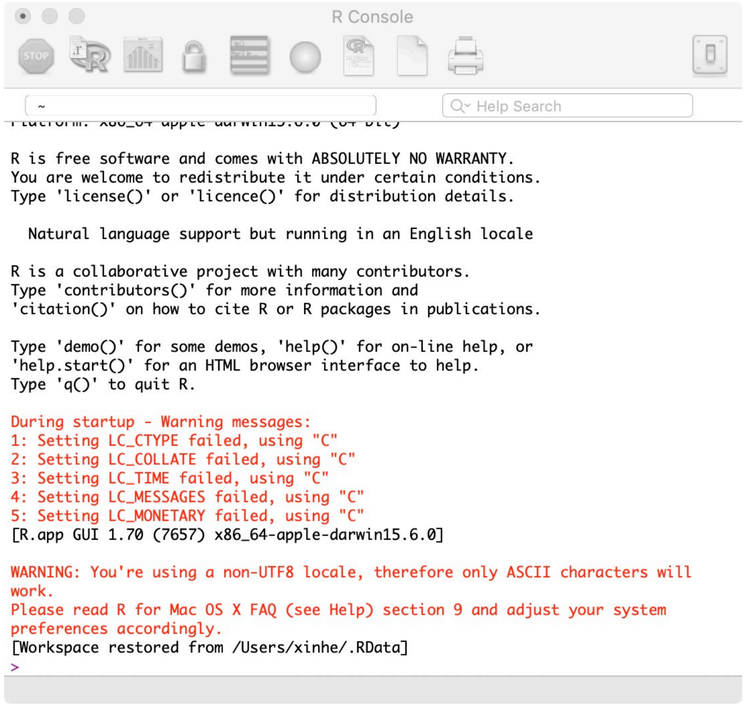
如何估计f

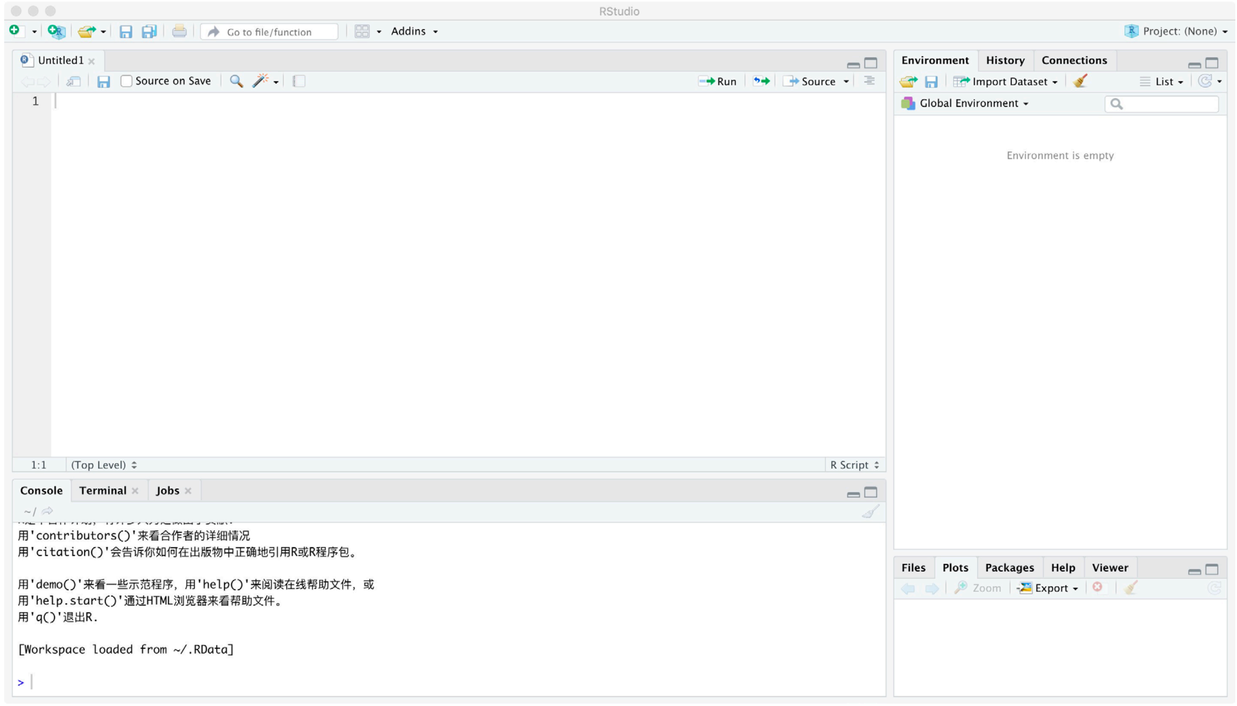
- 训练数据（training data）：事先收集到的n个观测。
- $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$，其中$\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)^T$。
- 目标：利用训练数据训练模型，从而估计f。

- 估计方法：参数方法（parametric）与非参数方法（non-parametric）

统计软件

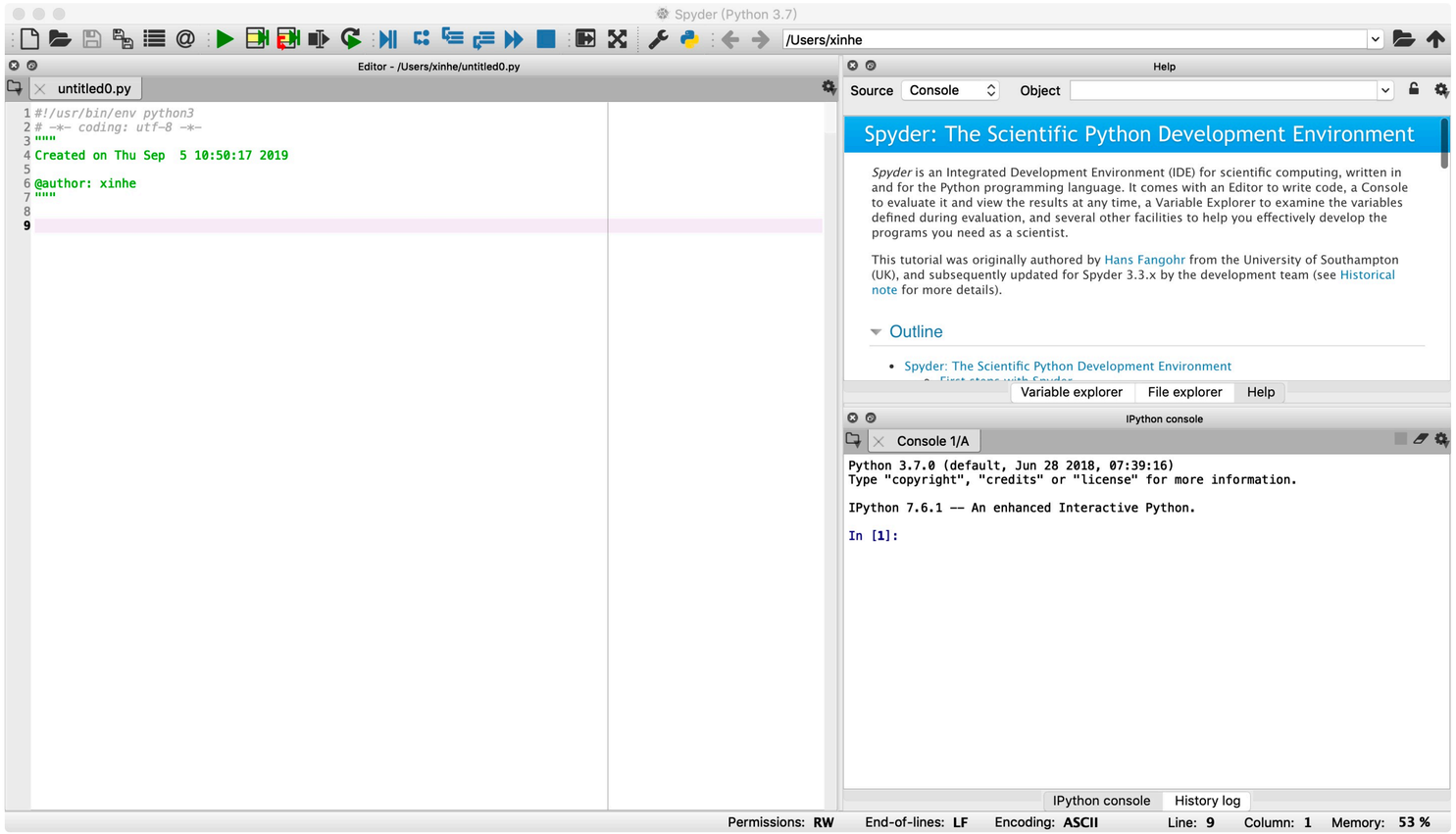
- R与R studio





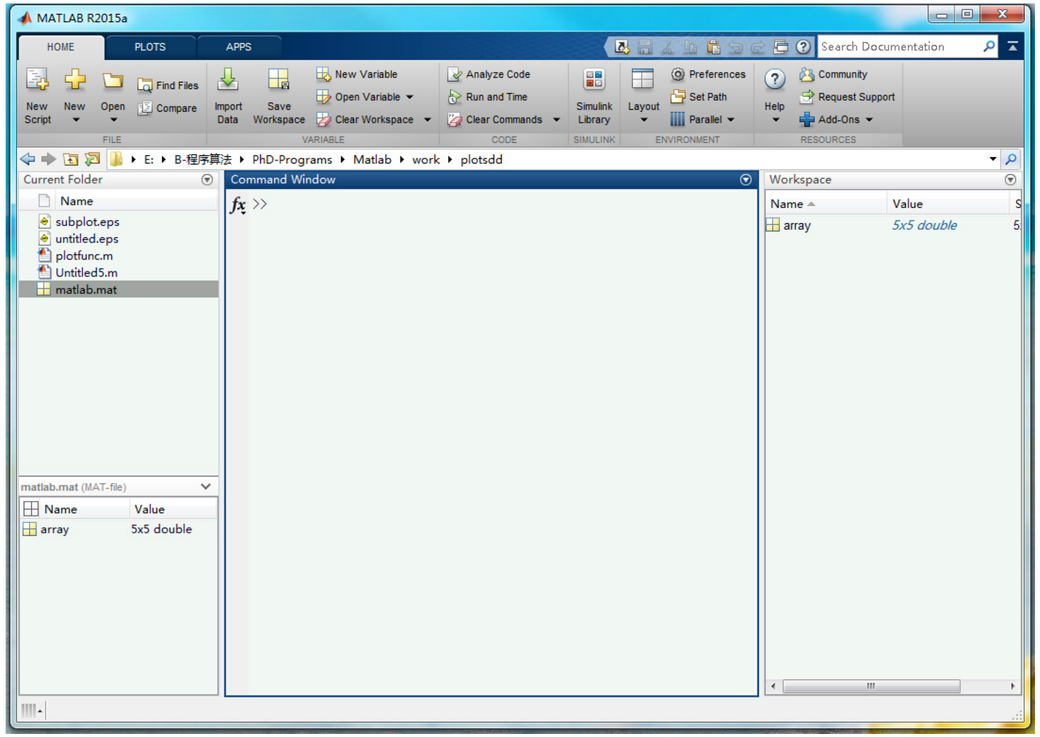
统计软件

- Python与Spyder

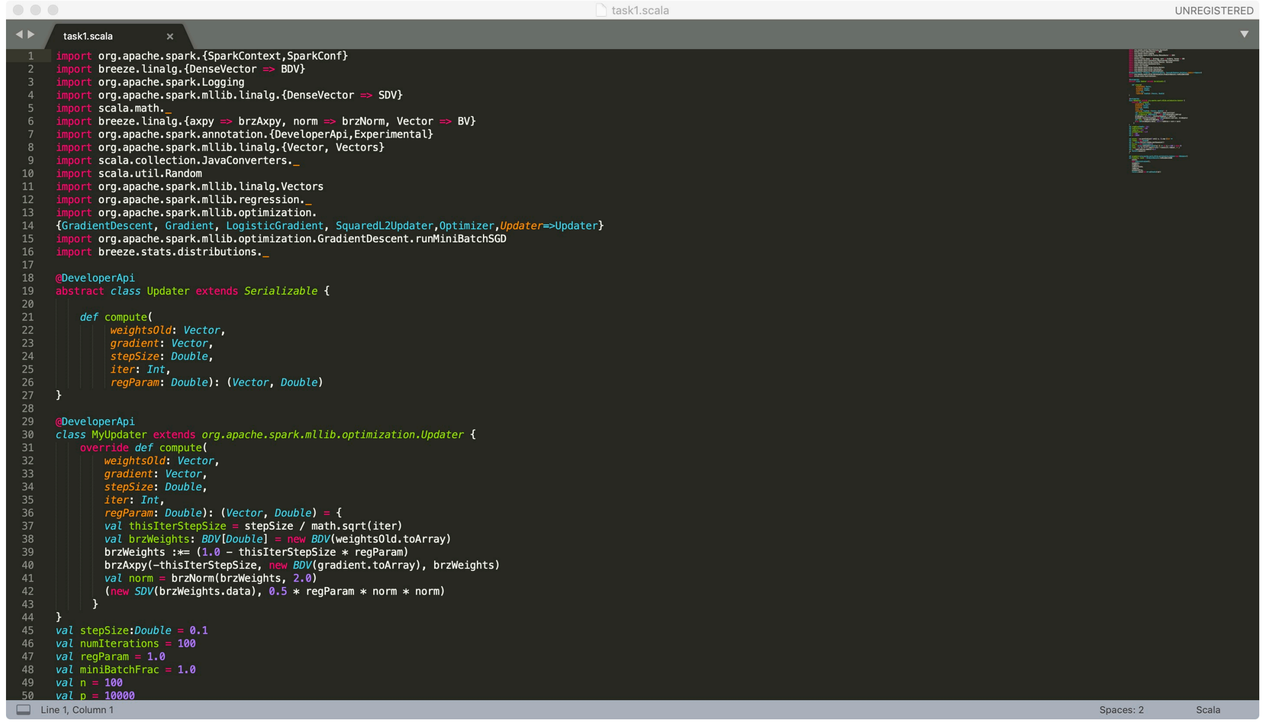


统计软件

- Matlab



- Sublime：一款功能较强的编译器



总结

- 人工智能的发展历史
- 机器学习的定义、分类
- 模型评价
- 统计软件