中国地质大学 计算机学院
China University of Geosciences

机器学习及其应用

陈伟涛
副教授，博士生导师
Email：wtchen@cug.edu.cn

2019.5.30

特朗普启动美国人工智能计划，五项原则欲维护AI领先地位

2月11日，美国总统特朗普签署了一项名为“维护美国人工智能领导地位”的行政命令，正式启动美国人工智能计划。同日，白宫发文称，美国联邦政府将集中资源发展人工智能，以“确保美国在人工智能方面的优势”。

迄今为止，已有18个国家启动了国家人工智能战略。

国务院关于印发
新一代人工智能发展规划的通知

国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

现将《新一代人工智能发展规划》印发给你们，请认真贯彻执行。

国务院

2017年7月8日

（此件公开发布）

教育部关于印发《高等学校人工智能创新行动计划》的通知- 中华人民 ...
www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html
2018年4月10日 - 高等学校人工智能创新行动计划. 人工智能的迅速发展将深刻改变人类社会生活、改变世界。为贯彻落实《国务院关于印发新一代人工智能发展规划的 ...

科技部关于发布科技创新2030—“新一代人工智能”重大项目2018年度 ...
www.most.gov.cn/mostinfo/xinxifenlei/fgzc/gfxwj/.../t20181012_142131.htm
2018年10月12日 - 为落实国务院印发的《新一代人工智能发展规划》的总体部署，现根据《新一代 ... 按照《关于深化中央财政科技计划（专项、基金等）管理改革的方案》（国 ...

机器学习在人工智能领域的重要地位

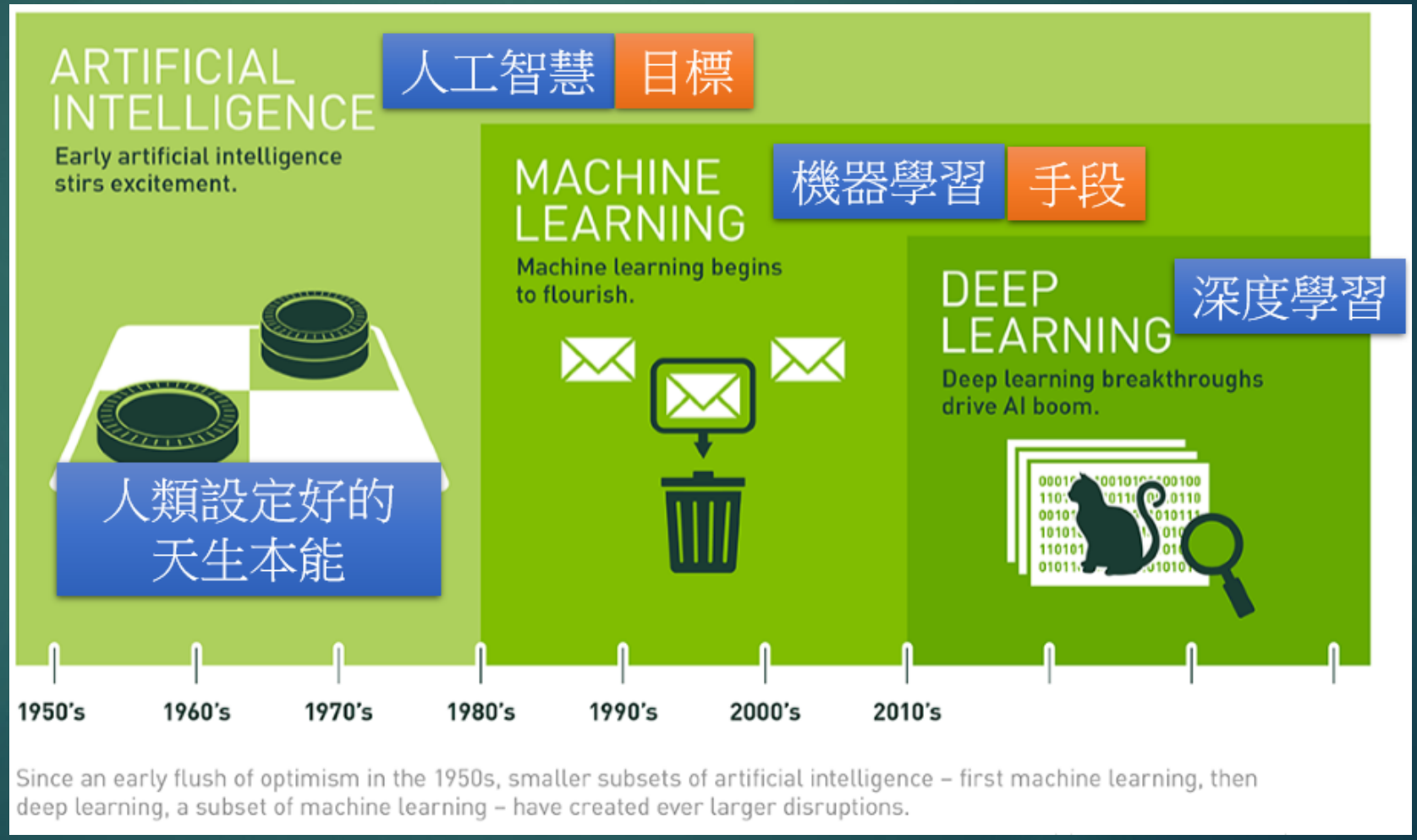
1956年，人工智能这门新兴学科在美国诞生，所谓人工智能就是机器模拟人类的思维完成等同于人类的工作。人脑可以通过思维进行学习是机器区别于人类最显著的特点，以适应不同环境的复杂工作，或者说人类可通过学习掌握之前所不具备的技能，进而随机应变，这是当前机器所远不能及的。

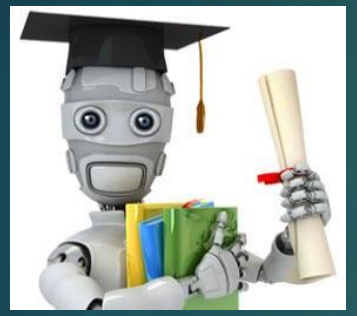
为了弥补这个缺憾，人工智能界的科学家自然想到了机器模拟人类学习的方式---机器学习。

机器学习是人工智能的组成部分，是人工智能的核心。

机器学习在人工智能领域的重要地位

人工智能是追求目标，机器学习是实现手段，深度学习是其中一种方法。





提纲

- 机器学习的定义
- 机器学习的发展历史和现状
- 机器学习的分类
- 机器学习的常见算法
- 机器学习的基本过程
- 机器学习的示例
- 机器学习的常见应用
- 流行的开源机器学习框架

机器学习的定义

在维基百科上，对机器学习提出以下几种定义：

- “机器学习是一门人工智能的科学，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”。

进一步理解机器学习

- 对于某给定的任务T，在合理的性能度量方案P的前提下，某计算机程序可以自主学习任务T的经验E；
- 随着提供合适、优质、大量的经验E，该程序对于任务T的性能逐步提高。
- 这里最重要的是机器学习的对象：
  - 任务Task, T，一个或者多个
  - 经验Experience, E
  - 性能Performance, P

  即：随着任务的不断执行，经验的累积会带来计算机性能的提升。

机器学习的发展历史

推理时期	1960s	赋予机器逻辑推理能力使机器获得智能；当时的AI程序证明力一些著名的数学定理，但由于缺乏知识，远不能实现真正的智能。
知识时期	1970s	将人类的知识总结出来教给机器使机器获得智能；即“专家系统”，在很多领域获得大量进展，但由于人类知识量巨大，故出现“知识工程瓶颈”。
机器学习时期	1980s	连接主义较为流行；代表方法为神经网络。
	1990s	统计学习占据舞台。代表方法包括支持向量机等。
	21Cent.	深度神经网络被提出，连接主义卷土重来。随着数据量和计算能力的不断提升，以深度学习为基础的诸多AI应用逐渐成熟。

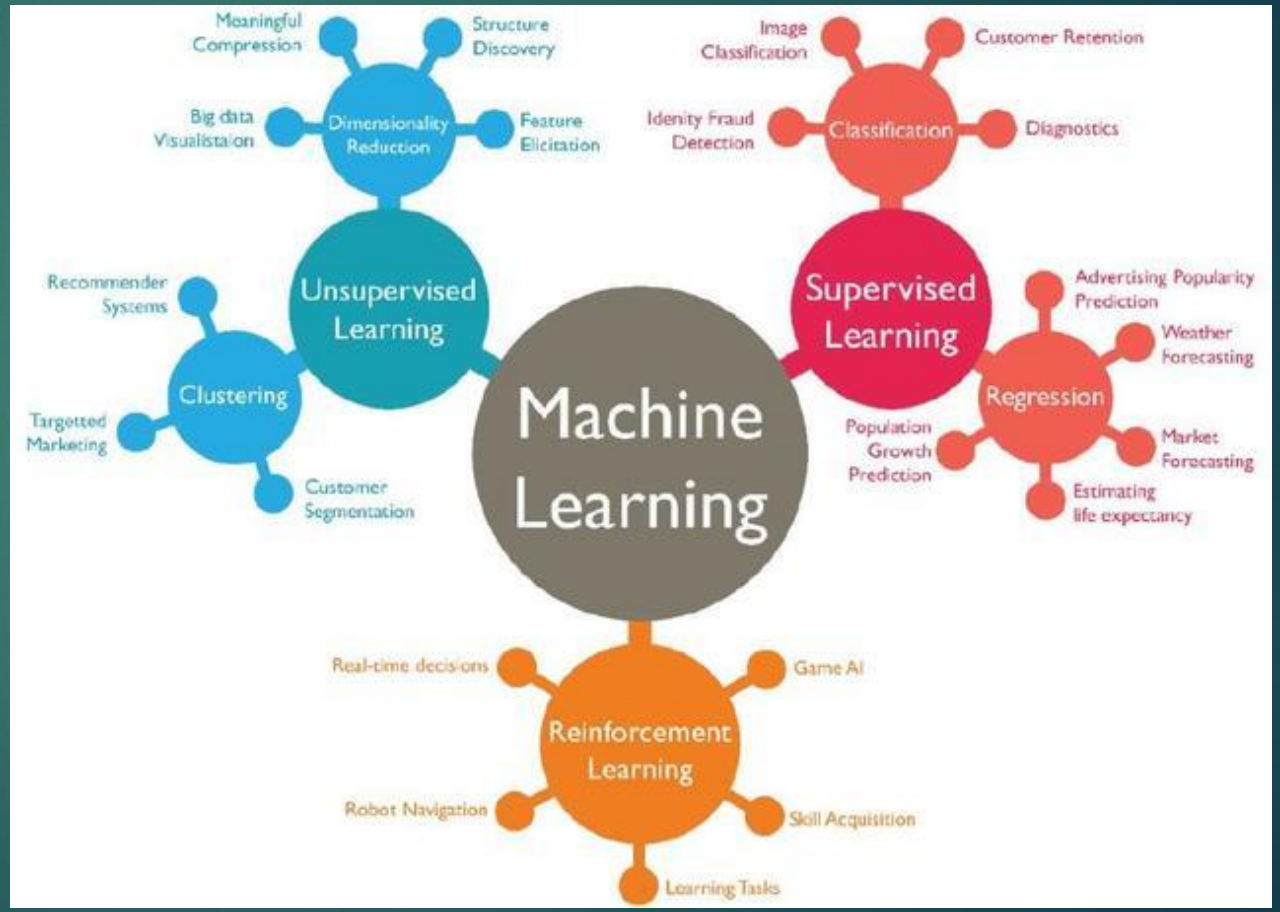
https://blog.csdn.net/hehe7zx

机器学习的发展现状

- 在搜索引擎方面Google的成功，使得Internet搜索引擎成为新兴产业。机器学习技术正在支撑着各类搜索引擎；
- DARPA(美国国防高级研究项目局)于2003年开始启动5年期PAL计划，这是一个以机器学习为核心的计划(涉及到AI的其他分支，如知识表示和推理、自然语言处理等)；
- 汽车自动驾驶。机器学习的主要任务是从立体视觉中学习如何行驶，根据观察人类的驾驶行为记录各种图像和操纵指令，并将它们进行正确分类；
- 在对天文物体进行分类、计算机系统性能预测、信用卡盗用检测、邮政服务属性识别、网络文档自动分类等方面，机器学习也在快速发展壮大。

机器学习的分类

- 监督学习
- 无监督学习
- 半监督学习
- 强化学习



机器学习的分类 -- 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，可以根据这个函数（模型）预测结果；
- 在监督式学习下，输入数据被称为“训练数据”，**每组训练数据有一个明确的标识或结果，**如，对防垃圾邮件系统中“垃圾邮件”、“非垃圾邮件”；
- 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。常见的监督学习算法包括回归分析和统计分类。

机器学习的分类 -- 无监督学习

- 在无监督式学习中，**数据并不被特别标识，学习模型是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。常见算法包括Apriori算法和k-Means算法。
- 监督学习和无监督学习的区别：训练集目标是否被标注。他们都有训练集，且都有输入和输出。

机器学习的分类 -- 半监督学习

- 半监督学习是介于监督学习与无监督学习之间一种机器学习方式，主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题；
- 应用场景包括分类和回归，算法包括一些对常用监督式学习算法的延伸，这些算法首先试图对未标识数据进行建模，在此基础上再对标识的数据进行预测，如图论推理算法（Graph Inference）或者拉普拉斯支持向量机（Laplacian SVM）等；
- 半监督学习从诞生以来，主要用于处理人工合成数据，无噪声干扰的样本数据是当前大部分半监督学习方法使用的数据，而在实际生活中用到的数据却大部分不是无干扰的，通常都比较难以得到纯样本数据。

机器学习的分类 -- 强化学习

- 强化学习通过观察来学习动作的完成，每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断；
- 在强化学习下，输入数据直接反馈到模型，模型必须对此立刻做出调整；
- 常见的应用场景包括动态系统以及机器人控制等。常见算法包括Q-Learning以及时间差学习（Temporal difference learning）。

机器学习的分类 -- 总结

- 在企业数据应用的场景下，人们最常用的可能就是监督式学习和无监督式学习的模型。
- 在图像识别等领域，由于存在大量的非标识的数据和少量的可标识数据，目前半监督式学习是一个很热的话题。
- 强化学习更多地应用在机器人控制及其他需要进行系统控制的领域。

机器学习的常见算法

- 回归算法（监督学习）
- 神经网络（监督学习）
- SVM支持向量机（监督学习）
- 聚类算法（无监督学习）
- 降维算法（无监督学习）
- 推荐算法（特殊）
- 其他算法

常见算法 -- 回归算法

- 回归算法有两个重要的子类：即线性回归和逻辑回归；
- 线性回归就是如何拟合出一条直线最佳匹配所有的数据，逻辑回归是一种与线性回归非常类似的算法；
- 线性回归处理的问题类型与逻辑回归不一致：

- **线性回归处理的是数值**问题，也就是最后预测出的结果是数字，例如房价。
- **逻辑回归属于分类算法**，也就是说，逻辑回归预测结果是离散的分类，例如判断这封邮件是否是垃圾邮件，以及用户是否会点击此广告等等。

常见算法 -- 神经网络

- 神经网络(也称之为人工神经网络，ANN)的诞生起源于对大脑工作机理的研究。早期生物界学者们使用神经网络来模拟大脑，后来，机器学习的学者们使用神经网络进行机器学习的实验，发现在视觉与语音的识别上效果都相当好。
- 神经网络算法是80年代机器学习界非常流行的算法。不过，进入90年代，神经网络的发展进入了一个瓶颈期。其主要原因是神经网络的训练过程很困难。
- 现在，携着“深度学习”之势，神经网络重装归来，重新成为最强大的机器学习算法之一。

常见算法 -- SVM支持向量机

- SVM算法是诞生于统计学习界，同时在机器学习界大放光彩的经典算法。
- 从某种意义上来说，支持向量机算法是逻辑回归算法的强化：通过给予逻辑回归算法更严格的优化条件，支持向量机算法可以获得比逻辑回归更好的分类界线。
- 支持向量机是一种数学成分很浓的机器学习算法（相对的，神经网络则有生物科学成分）。通过支持向量机算法，既可以保持计算效率，又可以获得非常好的分类效果。因此支持向量机在90年代后期一直占据着机器学习中最核心的地位，基本取代了神经网络算法。直到现在神经网络借着深度学习重新兴起，两者之间才又发生了微妙的平衡转变。

常见算法 -- 聚类算法

- 聚类算法是无监督学习算法中最典型的代表。
- 聚类算法就是计算种群中的距离，根据距离的远近将数据划分为多个族群。
- 聚类算法中最典型的代表就是K-Means算法。

常见算法 -- 降维算法

- 降维算法也是一种无监督学习算法，主要特征是将数据从高维降低到低维。维度表示数据的特征量的大小。例如，房价包含房子的长、宽、面积与房间数量四个特征，也就是维度为4维的数据。可以看出来，长与宽事实上与面积表示的信息重叠了，例如面积=长 × 宽。通过降维算法，可以去除冗余信息，将特征减少为面积与房间数量两个特征，即从4维的数据压缩到2维。这样，不仅利于表示，同时提高计算的性能。

- 降维算法的主要作用是压缩数据与提升机器学习的效率。通过降维算法，可以将具有几千个特征的数据压缩至若干个特征。另外，降维算法的另一个好处是数据的可视化，例如将5维的数据压缩至2维，然后可以用二维平面来可视。降维算法的主要代表是PCA算法(即主成分分析算法)。

常见算法 -- 推荐算法

- 推荐算法是目前业界非常火的一种算法，如亚马逊，天猫，京东等都在广泛地运用。推荐算法的主要特征就是可以自动向用户推荐他们最感兴趣的东西，从而增加购买率，提升效益。

常见算法 -- 推荐算法

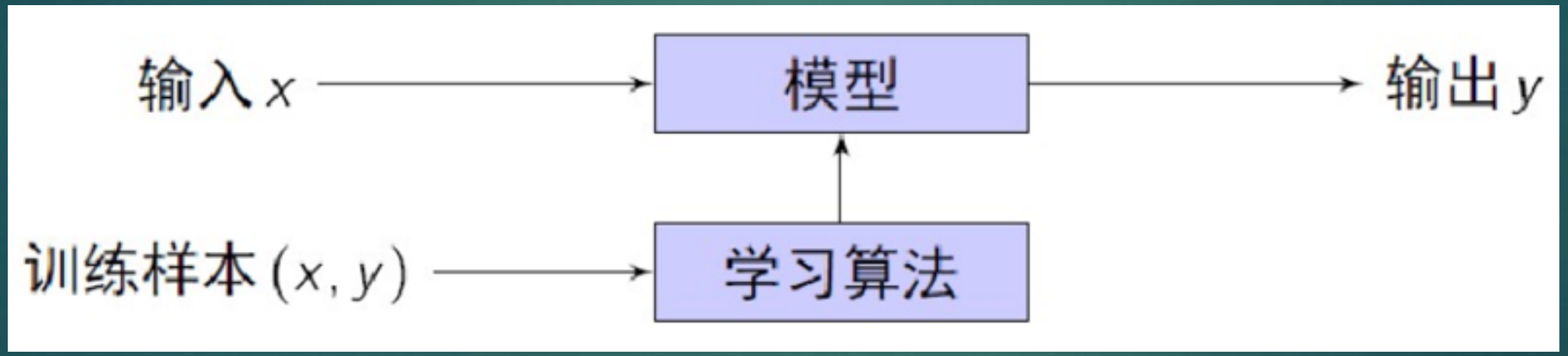
- 推荐算法有两个主要的类别：
  - 一类是基于物品内容的推荐，是将与用户购买的内容近似的物品推荐给用户，这样的前提是每个物品都得有若干个标签，因此才可以找出与用户购买物品类似的物品，这样推荐的好处是关联程度较大。
  - 另一类是基于用户相似度的推荐，则是将与目标用户兴趣（注：用户画像）相同的其他用户购买的东西推荐给目标用户，例如小A历史上买了物品B和C，经过算法分析，发现另一个与小A近似的用户小D购买了物品E，于是将物品E推荐给小A。
- 两类推荐都有各自的优缺点，在一般的电商应用中，一般是两类混合使用。推荐算法中最有名的算法就是协同过滤算法。

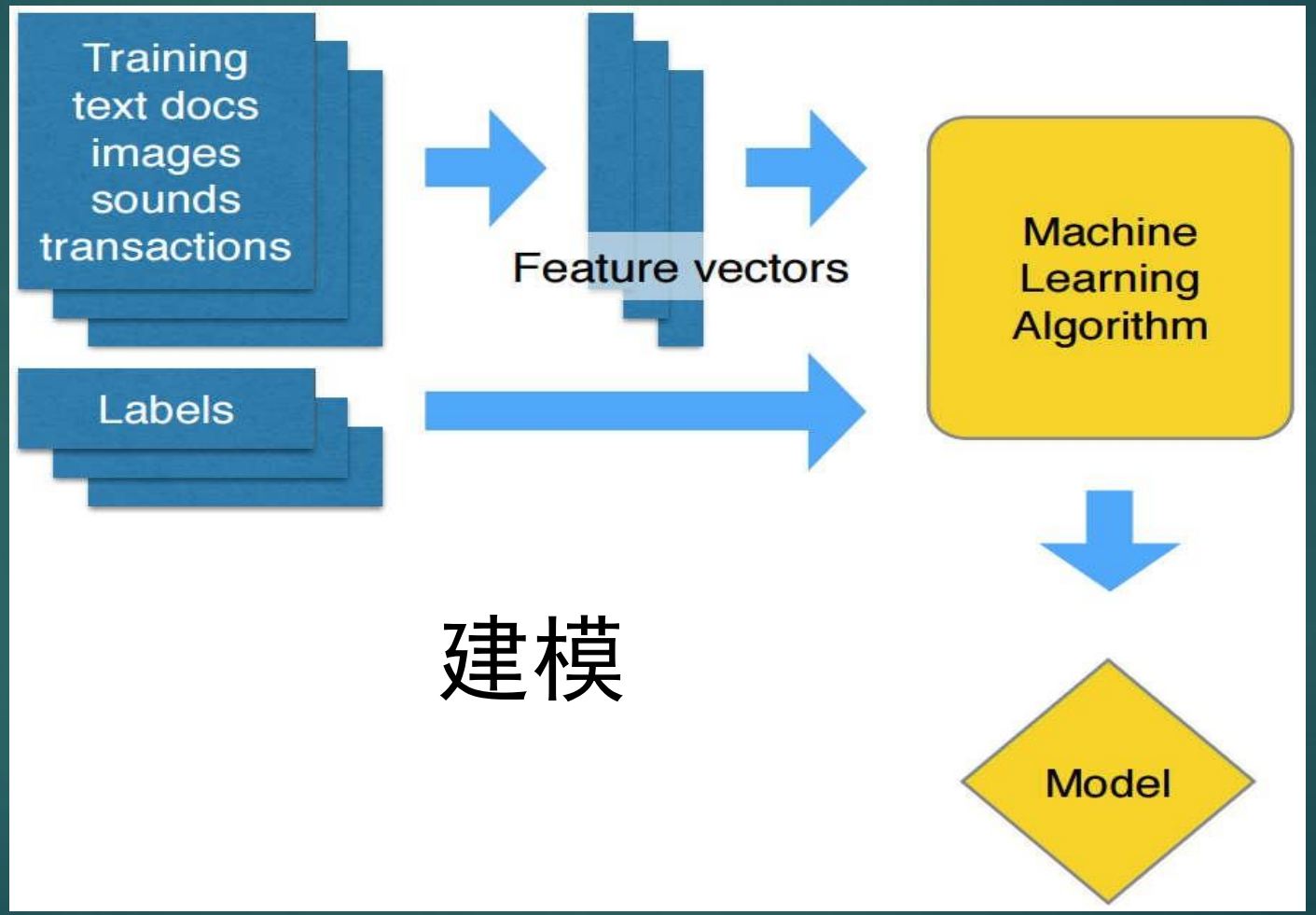
常见算法 -- 其他算法

- 除了以上算法之外，机器学习界还有其他的如高斯判别，朴素贝叶斯，决策树等等算法。机器学习界的一个特色就是算法众多，发展百花齐放。
- 除了这些算法以外，有一些算法的名字在机器学习领域中也经常出现。但他们本身并不算是一个机器学习算法，而是为了解决某个子问题而诞生的。可以理解他们为以上算法的子算法，用于大幅度提高训练过程。其中的代表有：梯度下降法，主要运用在线性回归，逻辑回归，神经网络，推荐算法中；牛顿法，主要运用在线性回归中；BP算法，主要运用在神经网络中；SMO算法，主要运用在SVM中。

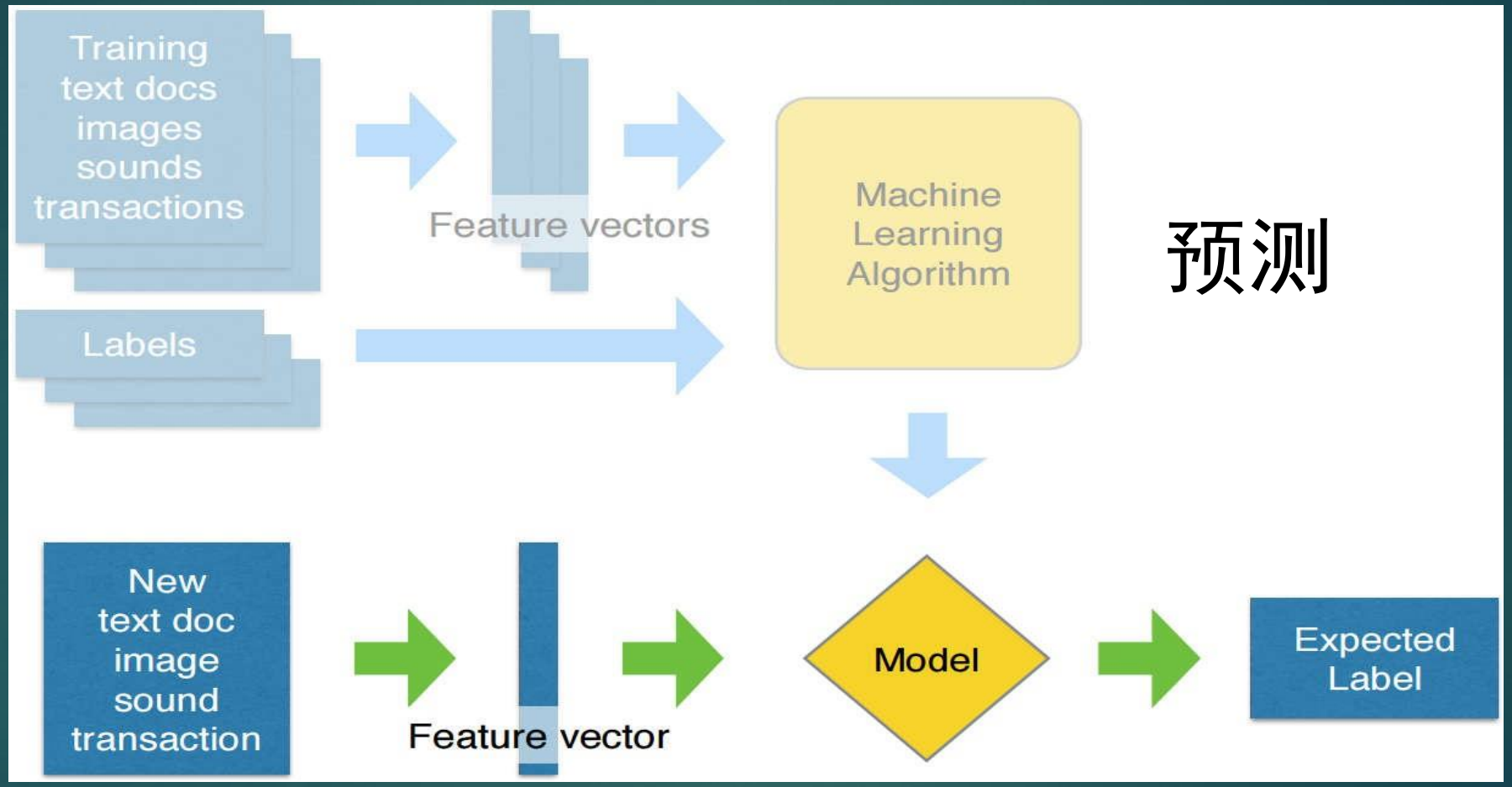
机器学习的基本过程

- 计算机从给定的数据中学习规律，即从观测数据（样本）中寻找规律、建立模型，并利用学习到的规律（模型）对未知或无法观测的数据进行预测。




Training
text docs
images
sounds
transactions
Labels
Feature vectors
Machine
Learning
Algorithm
Model

建模

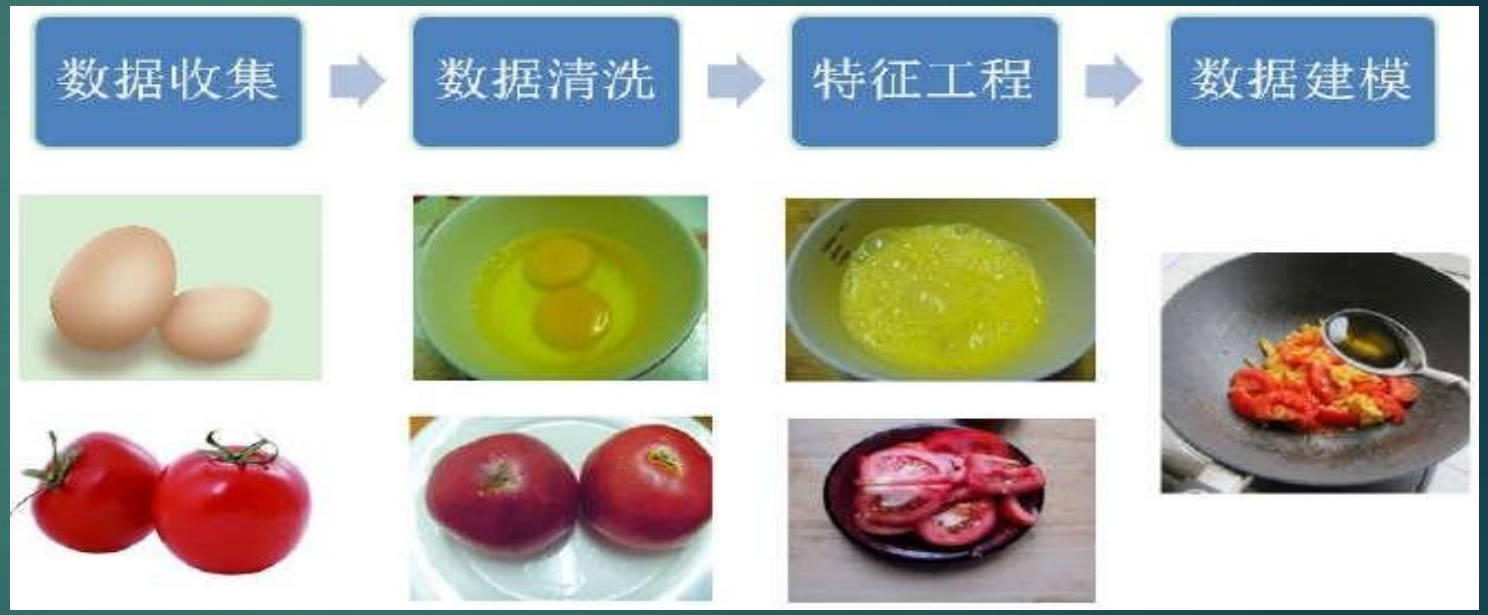

Training
text docs
images
sounds
transactions
Labels
Feature vectors
Machine
Learning
Algorithm
New
text doc
image
sound
transaction
Feature vector
Model
Expected
Label

预测

机器学习的操作流程

- 操作流程主要分6步:

A. 数据收集；
B. 数据预处理（清洗）；
C. 特征工程；
D. 数据建模；
E. 评估模型；
F. 预测新数据。



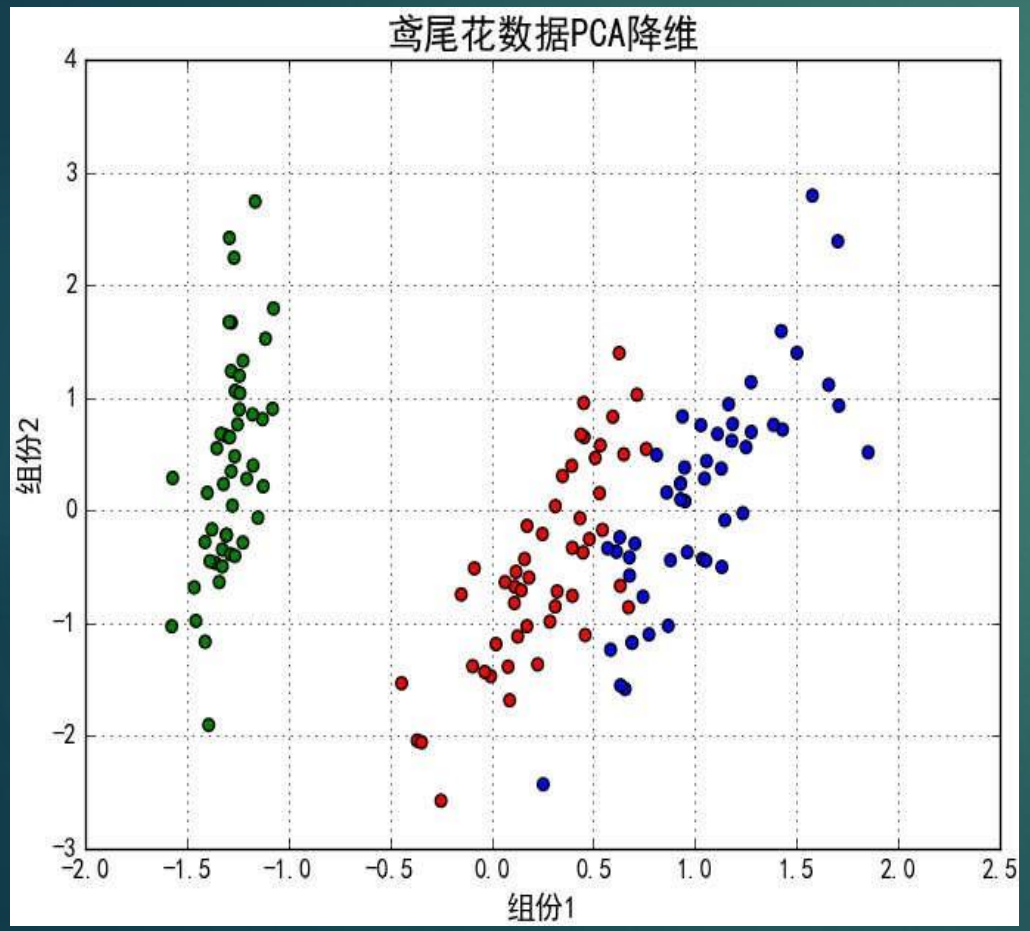
机器学习示例

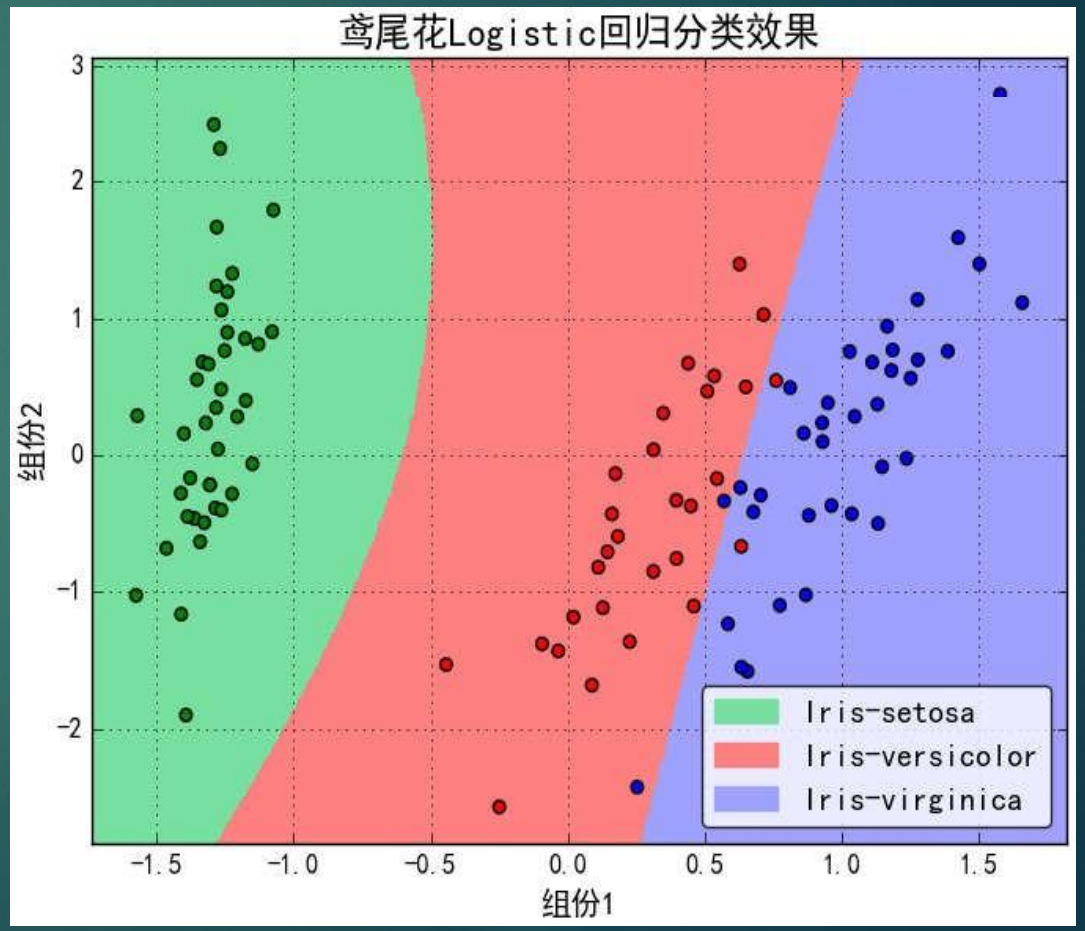
鸢尾花数据集分类

- 该数据集共150行，每行1个样本。每个样本有5个字段，分别是
  - 花萼长度（单位cm）
  - 花萼宽度（单位：cm）
  - 花瓣长度（单位：cm）
  - 花瓣宽度（单位：cm）
  - 类别（共3类）
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica


```
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
```

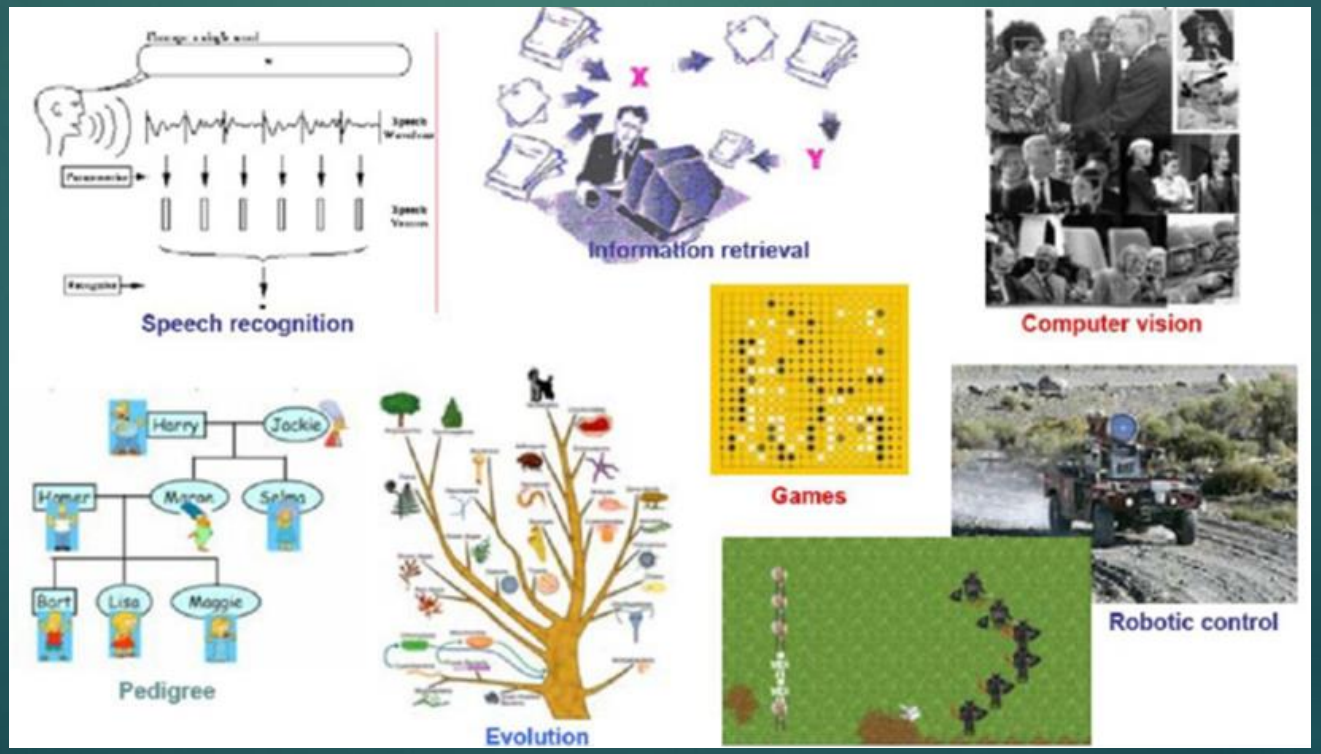
对鸢尾花数据集进行PCA降维，只保留两个主成分，数据可视化结果以及Logistic分类结果可视化如下图：





机器学习的常见应用

- 机器学习已广泛应用于数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别、战略游戏和机器人等领域。



流行的开源机器学习框架

TensorFlow是谷歌基于C++开发、发布的第二代机器学习系统。开发目的是用于进行机器学习和深度神经网络的研究。目前Google 的Google App 的语音识别、Gmail 的自动回复功能、Google Photos 的图片搜索等都在使用 TensorFlow 。
GitHub项目地址：
https://github.com/tensorflow/tensorflow

Scikit-Learn是用于机器学习的Python 模块，它建立在SciPy之上。基本功能主要被分为六个部分：分类、回归、聚类、数据降维、模型选择、数据预处理。
GitHub项目地址:
https://github.com/scikit-learn/scikit-learn

Caffe

Caffe 是由神经网络中的表达式、速度及模块化产生的深度学习框架。Caffe是一个基于C++/CUDA架构框架，开发者能够利用它自由的组织网络，目前支持卷积神经网络和全连接神经网络（人工神经网络）。在Linux上，C++可以通过命令行来操作接口，运算上支持CPU和GPU直接无缝切换。

GitHub项目地址:

https://github.com/BVLC/caffe

keras makes Deep Learning simple

Keras是基于Python开发的极其精简并高度模块化的神经网络库，在TensorFlow或 Theano 上都能够运行，是一个高度模块化的神经网络库，支持GPU和CPU运算。Keras侧重于开发快速实验，用可能最少延迟实现从理念到结果的转变，即为做好一项研究的关键。

GitHub项目地址:

https://github.com/fchollet/keras