数据科学与人工智能简介

童世炜
中国科学技术大学 BDAA实验室

2017.06.16

基于大数据 + 机器学习的智能服务已应用于多个行业

金融

医疗
卫生

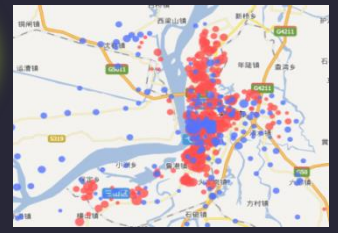
电子
商务

教育

BIG
DATA

军事

移动
商务



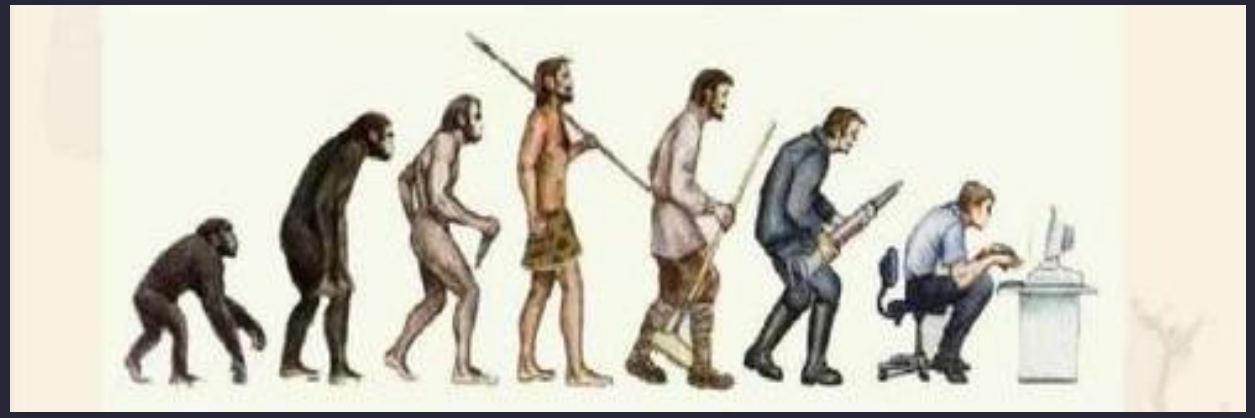
气象

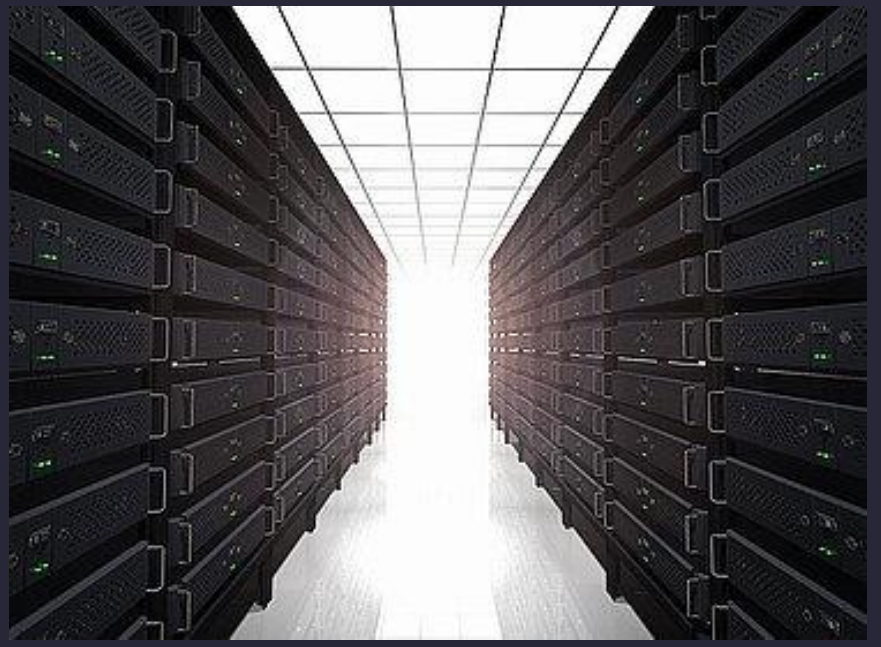
什么是大数据？

人工智能是什么？

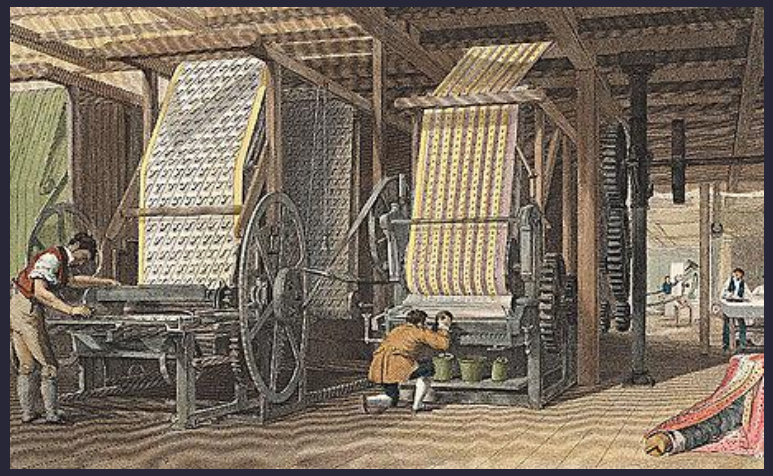
机器要学习的是什么？

从石器到计算机





工业革命与信息化浪潮



计算机程序

```
randolph@Randolph ~/Desktop python3 TrafficLight.py
```

01

红绿灯

TRAFFIC LIGHT

示例代码

02

操作系统

OPERATING SYSTEM

用户界面将需要能接受用户输入，比如键盘、鼠标等

程序需要监听键盘和鼠标

根据输入做出对应响应，例如用户双击图标时打开程序

计算机程序

输入：0个或多个

$f(x)$

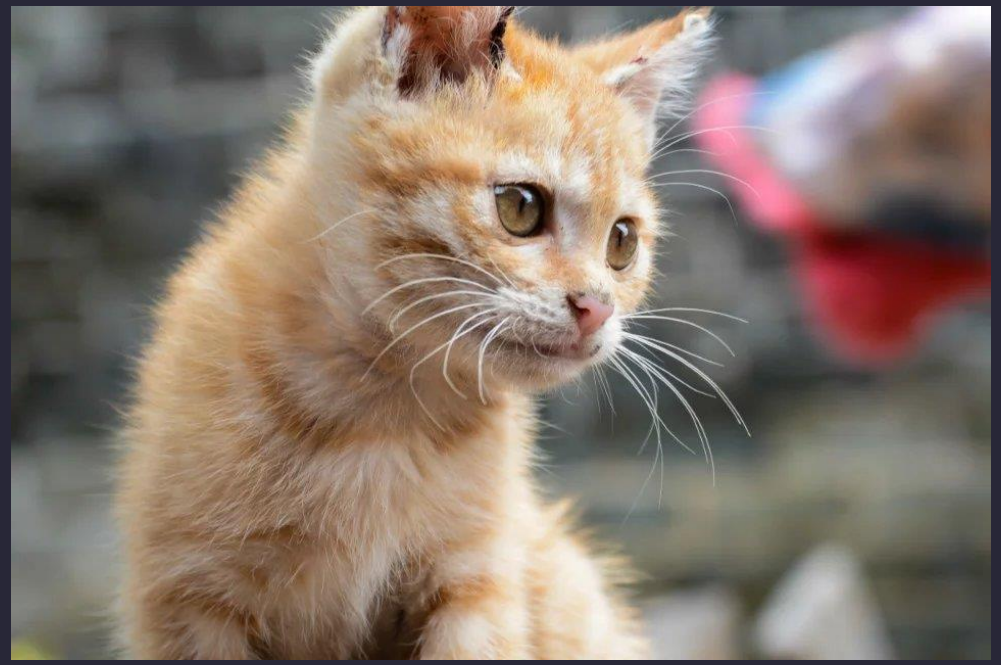
规则：响应输入，针对不同的输入做出对应的操作

计算机程序

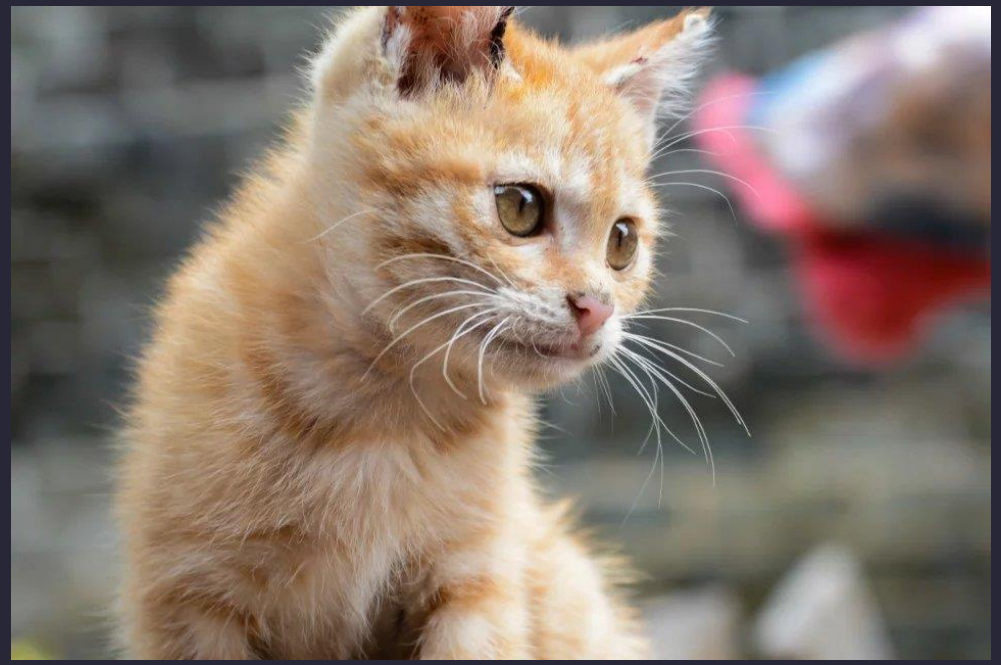
一个人在教会电脑之前，别说他真正理解这个东西了。

——Donald Knuth

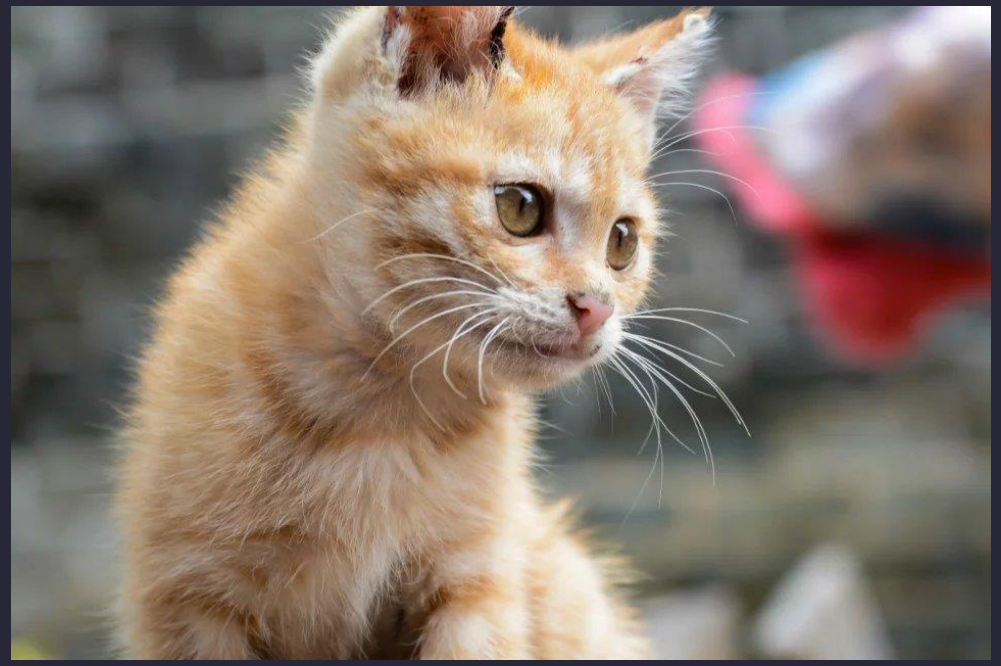
图灵的猫



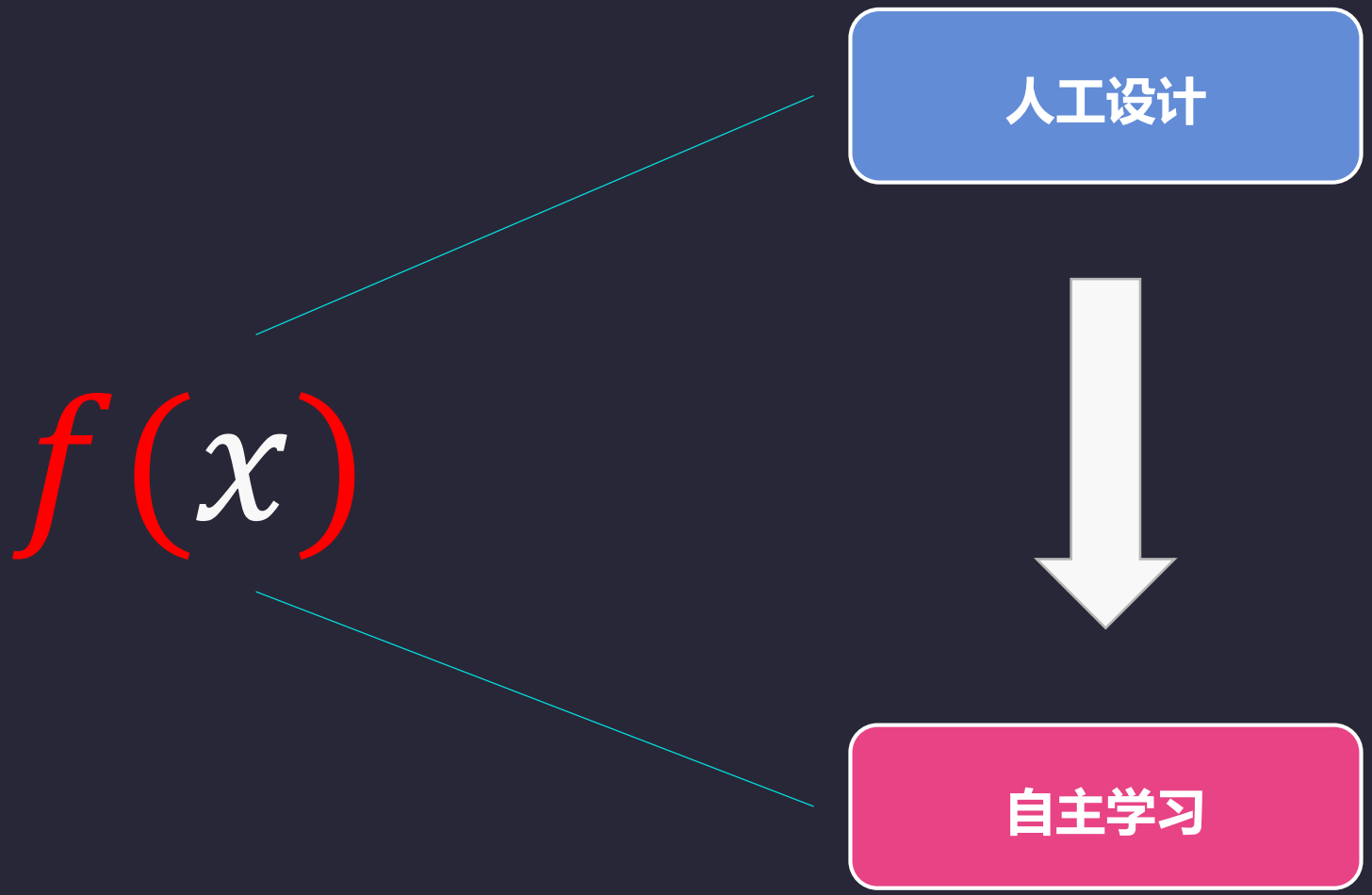
图灵的猫



图灵的猫



从人工建模到数据智能



图灵的猫

图灵的猫

数据




1




0

$f(x)$

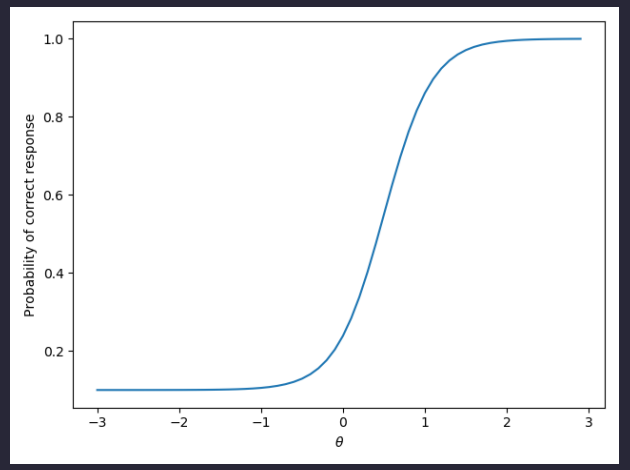
1

从数据到规律

$$D \overset{?}{\mapsto} f$$

数据　　规律

从数据中来到数据中去



中世纪

英尺的诞生

1英尺 = 男子的平均脚长

16名男子的平均脚长被用来估计男子的平均脚长

应对特异形状的脚：最长和最短的脚不计入

1951

Item Response Theory

观察、构建数据模型、采用定量分析方法来分析考试成绩或者问卷调查数据，确定潜在心理特征是否可以通过测试题被反应出来，以及测试题和被测试者之间的互动关系

2017

AlphaGo系列

AlphaGo通过学习棋谱+自我训练击败击败李世石

AlphaGo Zero通过自我训练完胜AlphaGo

科学范式

几千年前

经验科学

第一范式

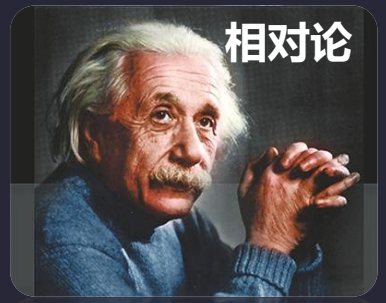
- 以归纳法为主，带有盲目性的观测和实验
- 科学实验

几百年前

理论科学

第二范式

- 以演绎法为主，关注理论总结和理性概括
- 数学模型



几十年前

计算科学

第三范式

- 重视数据模型构建、定量分析方法，利用计算机来分析和解决
- 科学计算

今天

数据密集型科学

第四范式

- 先有了大量的已知数据，然后通过计算得出之前未知的理论
- 机器学习

困难与挑战

- 针对一类问题的特征，如何设计合适的学习算法
- 如何使用数据来有效地获取函数参数具体值

- 如何验证所得规律
- 如何保证所得规律的正确性

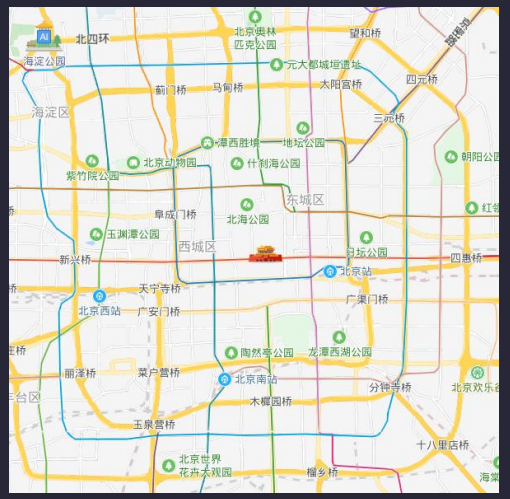
数据时代

- **数据**： 从计算机科学的角度，所有能够输入到计算机并被计算机程序处理的符号的总称
- **数据时代**：我们生活在数据中，所有人都在制造和分享数据。
- **数据时代的数据特点**
  - 数据量大：数据量已到ZB (2 ^ 60KB) 级别
    - 1 ZB = 地球上沙粒的总量
    - KB → MB → GB → TB → PB → EB → ZB → YB → NB → DB
    - PB以上级别的数据，最有效的传输方式是空运，而不是网络
  - 价值密度低：挖掘大数据中的价值类似沙里淘金，需要从海量数据中挖掘稀疏但珍贵的信息
  - 数据类型多
    - 形式多样：图片、文本、视频……
    - 来源多样：互联网、物联网……
- **数据时代信息处理的要求**
  - 高实时性：1秒定律

Google

当文字成为数据

当沟通成为数据



当方位成为数据

美团 淘

当生活成为数据

困难与挑战

- 针对一类问题的特征，如何设计合适的学习算法
- 如何使用数据来有效地获取函数参数具体值

- 数据类型多样，包括各种格式和形态的数据
- 数据量大，但存在噪声，有效数据少
- 数据处理要求高实时性

- 如何验证所得规律
- 如何保证所得规律的正确性

数据智能

人工智能

构建一套具有学习、归纳能力的自动化系统，它可以完成对数据的特征抽取及数据分布模式的辨别与学习。

机器学习

讨论各式各样的适用于不同问题的函数形式，以及如何使用数据来有效地获取函数参数具体值。

数据科学

用科学的方法研究、处理和应用数据，挖掘令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识，并据此更好的服务人们的生活。

深度学习

深度学习是机器学习中的一类函数，它们的形式通常为多层神经网络。近年来，已逐渐成为处理图像、文本语料和声音信号等复杂高维度数据的主要方法。

数据是一种原材料，数据库、数据挖掘、云计算、高性能计算、机器学习等都可以看作是对这种原材料进行存储烹饪加工等的手段和技术，目的就是做出各种美食

大纲
TOC

01

数据科学基础

数据的收集、处理、分析、加工
潜在模式的挖掘

02

机器学习基础

基本的学习算法
如何有效对算法结果进行评估
如何有效地获取函数参数具体值

03

深度学习基础

神经网络的原理
几个典型的神经网络
深度网络的问题

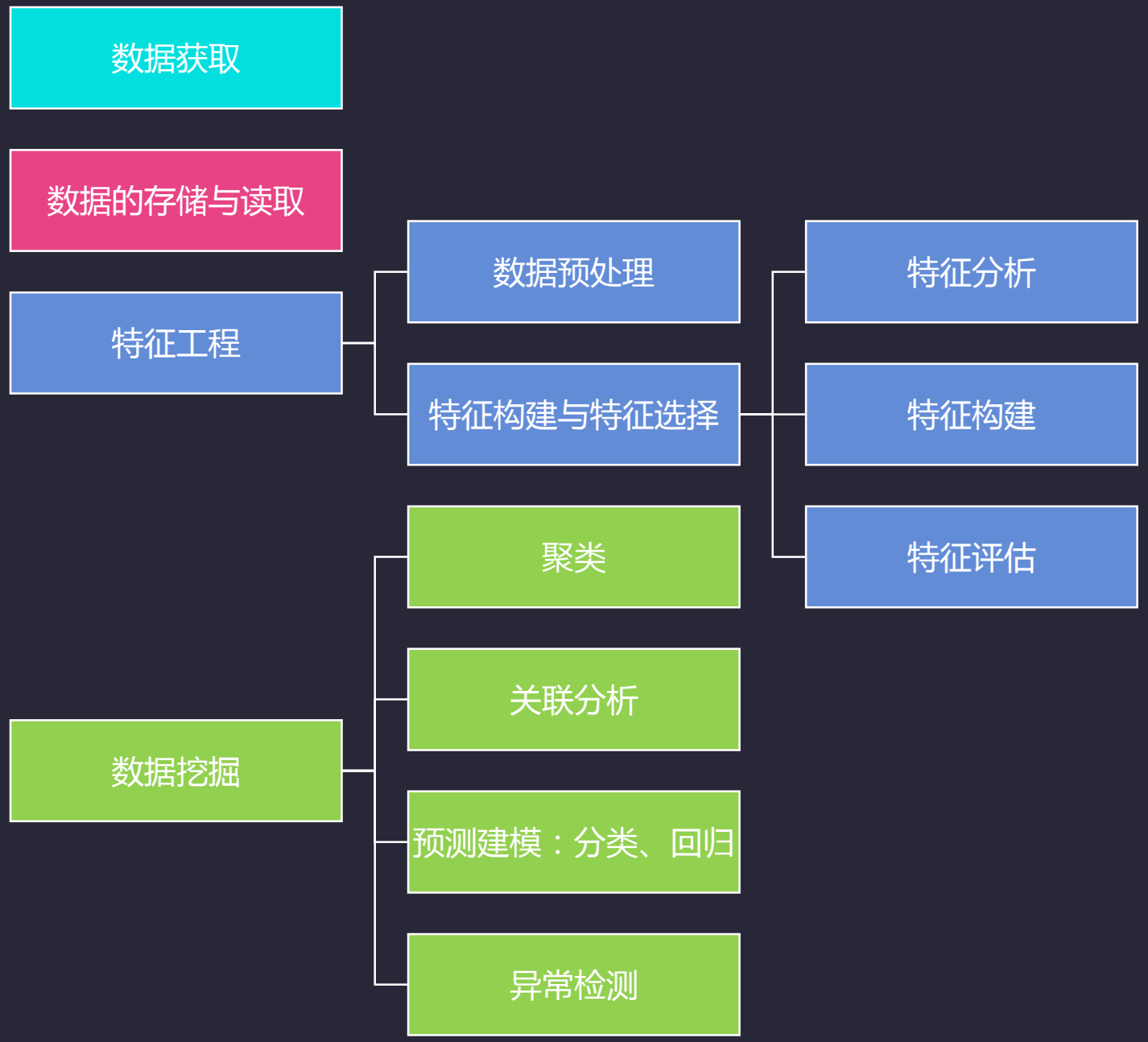
04

Less is More

更少的人工，更多的自动化
无监督学习、迁移学习、强化学习、元学习、自动机器学习.....

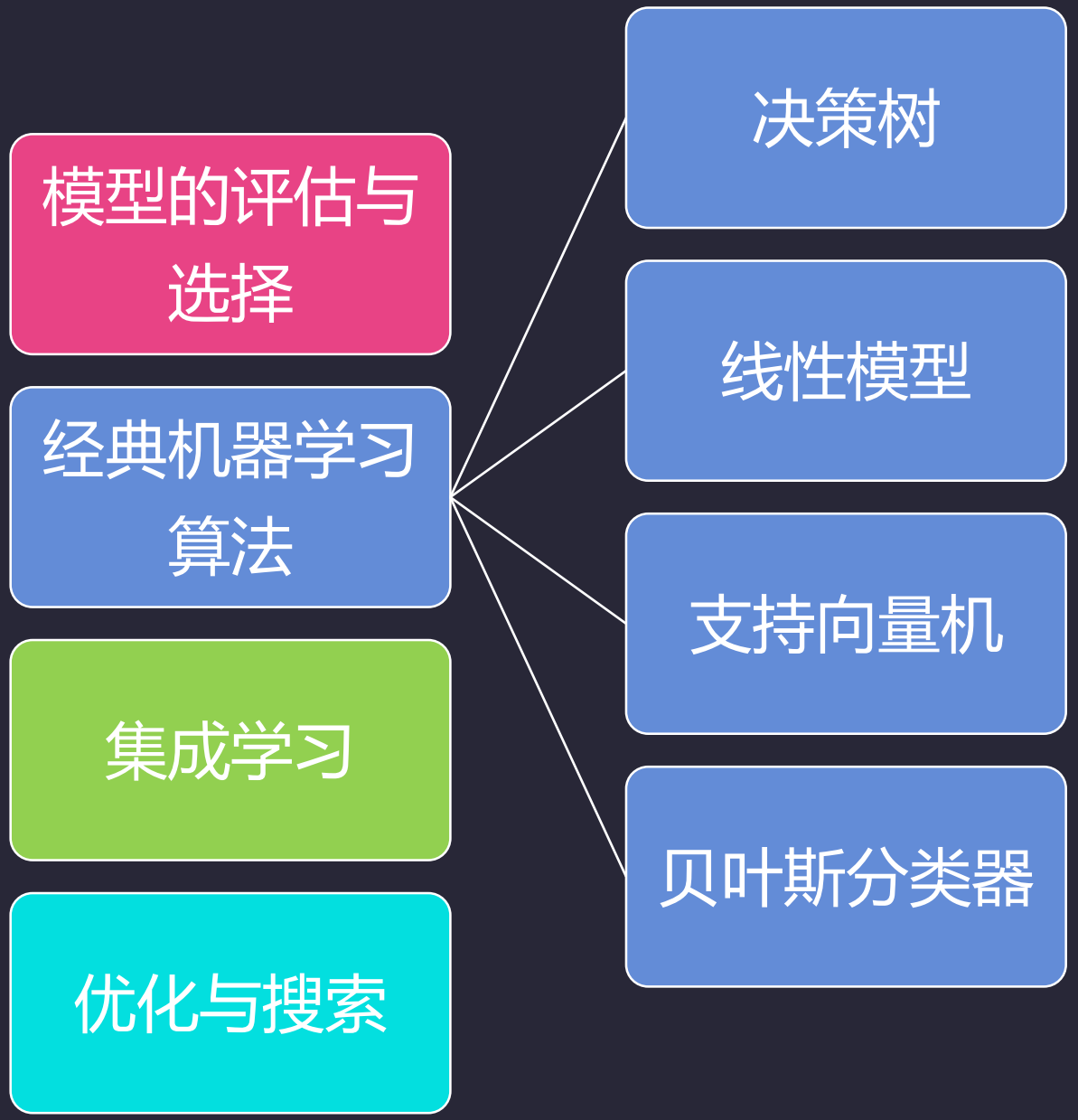
数据科学基础

- 数据获取
- 数据的存储与读取
- 特征工程
  - 数据预处理
  - 特征构建与特征选择
    - 特征分析
    - 特征构建
    - 特征评估
- 数据挖掘
  - 聚类
  - 关联分析
  - 预测建模：分类、回归
  - 异常检测



机器学习基础

- 模型的评估与选择
- 经典机器学习算法
  - 决策树
  - 线性模型
  - 支持向量机
  - 贝叶斯分类器
- 集成学习
- 优化与搜索



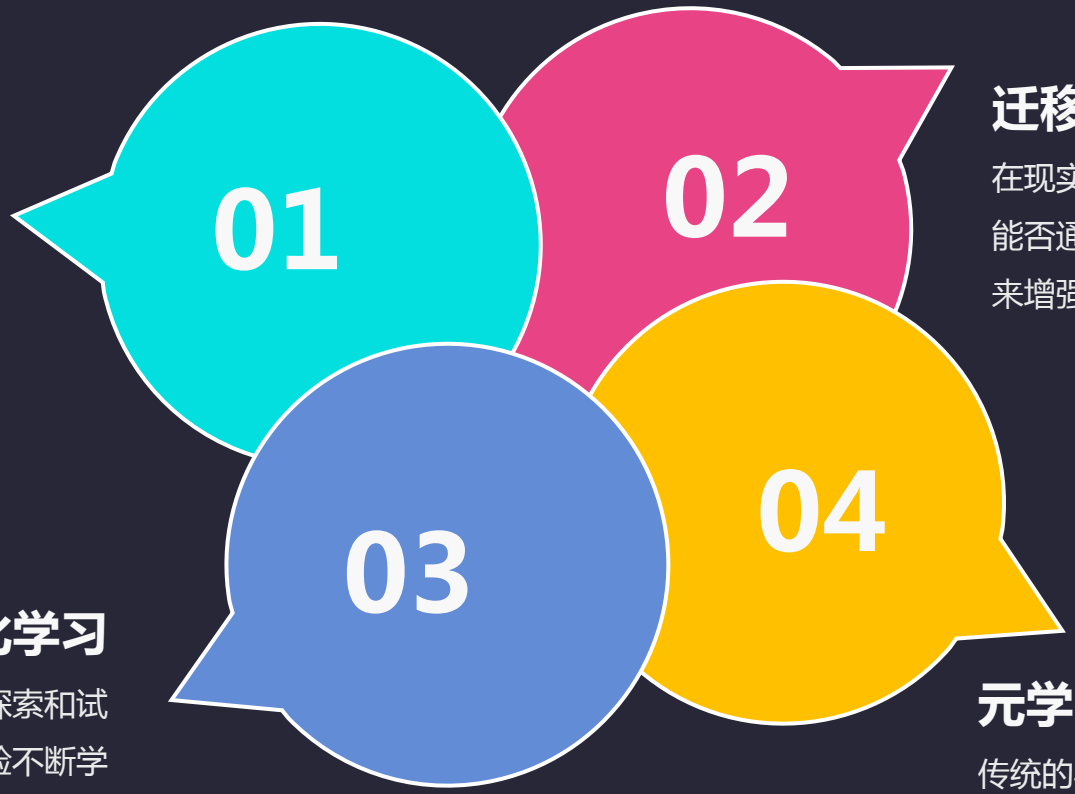
深度学习基础

- BP神经网络
  - 梯度
  - 计算图
- 变种
  - 序列分析
    - 循环神经网络 RNN
  - 结构分析
    - 卷积神经网络 CNN
    - 图神经网络 GNN
- 深度网络的两朵乌云
  - 梯度消失与梯度爆炸
  - 解释性

Less is More

无监督学习

现实生活中，常常因为缺乏足够的先验知识或人工标注成本过高，因此难以人工标注类别，那么是否有办法让机器在这种情况下进行学习呢



迁移学习

在现实生活中，常常存在数据不均衡的情况，能否通过某些方式，利用数据量丰富的数据集来增强训练数据量较少的模型的效果

强化学习

在许多问题中，我们需要通过在环境中进行探索和试错来积累经验，并在此过程中利用积累的经验不断学习，是否有办法让机器具有这种探索-利用的能力呢

元学习

传统的学习算法需要人工分析问题的特点以选定特定的函数，是否可以将函数的选择也交由机器来完成呢

参考文献

[1] 数据科学导论，刘淇，中国科学技术大学

[2] 动手学深度学习，Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola

tswsxk.github.io/handbook