

数据科学与工程导论

Introduction to Data Science and Engineering

- Person A is charged with the task of determining the ages of B's three children.
 - B tells A that the product of the children's ages is 36.
 - A replies that another clue is required.
 - B tells A the sum of the children's ages is 13.
 - A replies that another clue is needed.
 - B tells A that the oldest child plays the piano.
 - A tells B the ages of the three children.
- How old are the three children?

Ages of Children Problem

a. Triples whose product is 36

$$1+1+36=38$$
 $1+6+6=13$
 $1+2+18=21$ $2+2+9=13$
 $1+3+12=16$ $2+3+6=11$
 $1+4+9=14$ $3+3+4=10$

Ages of Children Problem

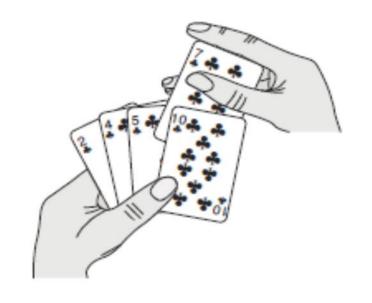


123

问题求解与思维方式

计算思维与数据思维

计算思维与数据思维实例

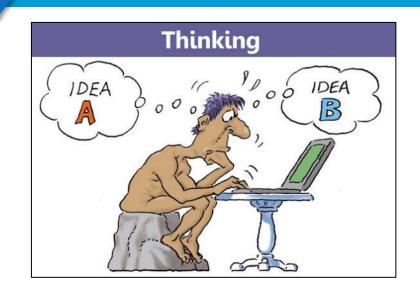


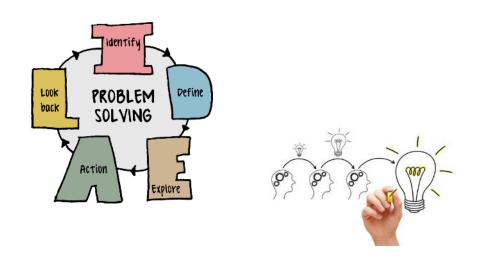
从玩牌的习惯性动作, 到计算机算法设计策略, 再到数据思维, 这条路并不遥远。

什么是思维

- 思维方式是人们大脑活动的内在程式,它对人们的言行起决定性作用。
- 思维方式决定行为结果,如果我们想在未来的学习工作上有所突破,一定要先在思维方式上有所提高。
- 思维方式决定了一个人的创新能力。

• 早在1972年,图灵奖得主Edsger Dijkstra就曾说: "我们所使用的工具影响着我们的思维方式和思维习惯,从而也深刻地影响着我们的思维能力",这就是著名的"工具影响思维"的论点。





思维方式

直线思维

一 人们最常用、最简单、最本能的思维模式 用直线模式去思考问题:从当前点出发,用刚刚发生过的事情建立方向,用直线去预测结果对于社会关系来说,运用直线思维可以形成简单的契约社会,人和人之间没有复杂的心理暗示 中国人很少采用直线思维思考社会问题,反而进行过度解读,形成了集体焦虑

逆向思维

对当前的状态进行反向思考。其目的是否定当前的状态,向相反的方向寻找目标

对于科学研究来说,逆向思维必不可少

逆向思维并不会持久存在,因为逆向思维的结果充满了风险

跳跃思维

- 不按部就班思考,间断性地向某个方向"跳起"的思维方式

 跳跃思维可以跨过鸿沟,到达新的起点,在科研难题的突破上是非常有用的

 跳跃思维要求人的大脑具有很强的活力、精力 正盛
 - 这种能力的外在表现就是"异想天开"

常见的思维模式

归纳思维

是人处理外界信息的一种手段 利用归纳思维,能够在短时间内对复杂的信息 建立各种模式,只要熟记几个简单的模式,就能掌握无穷多个可能的事实 很多人学习新知识很快,就是因为归纳思维运用得当。例如"举一反三"

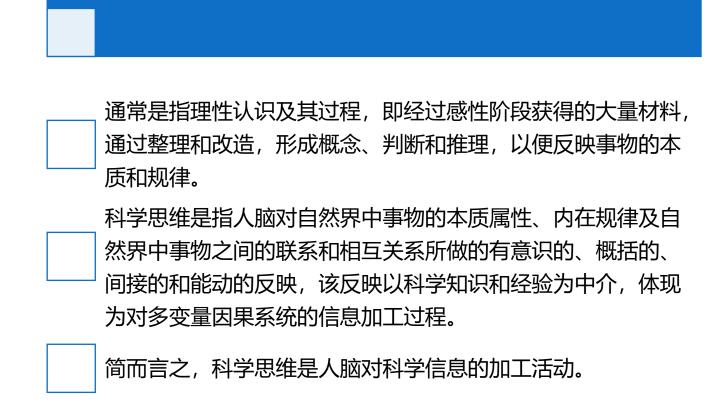
并行思维

人的思维方式一般从一件事情开始,完成后再
做另一件事情。这是自然选择,也是人的本能。
但是如果一个人负责一个团队,或者从事非常
复杂的系统性问题研究时,必须学会同时思考
几件事情,这就是并行思维。
我们并不是拥有多核的大脑,而是需要在不同

的课题之间进行快速切换。

常见的思维模式

科学思维

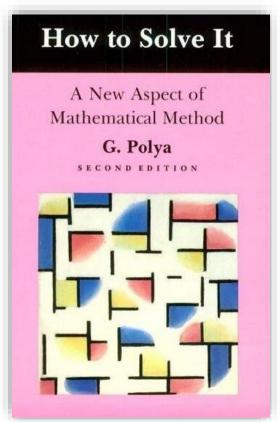




常见的思维模式

- Polya's Problem Solving Steps
 - 1. Understand the problem.
 - 2. Devise a plan for solving the problem.
 - 3. Carry out the plan.
 - 4. Evaluate the solution for accuracy and its potential as a tool for solving other problems.





问题求解



123

问题求解与思维方式

计算思维与数据思维

计算思维与数据思维实例

- 计算思维就是运用计算机科学的基本概念进行问题求解、系统设计以及人类行为理解。
- 计算思维是一种普适的思维,是每个人的基本技能,正如即制制出版促进了阅读、写作和算术(英文称为3R)的传播。
- 计算思维强调一切皆可计算, 从物理世界模拟到人类社会模 拟,从人类社会模拟再到智能 活动,都可认为是计算的某种 形式。



周以真 (Jeannette M. Wing) 教授提出计算思维的理念

什么是计算思维

- 计算思维的本质是抽象和自动化。它反映了计算的根本问题,即什么能被有效地自动进行。
- 计算是抽象的自动执行,自动化需要某种计算机去解释抽象。从操作层面上讲,计算就是如何寻找一台计算机去求解问题,隐含地说就是要确定合适的抽象,选择合适的计算机去解释执行该抽象,后者就是自动化。

计算思维与计算机科学

分而治之

把数据、过程或问题分解成更小的、易于管理或 解决的部分

模式识别

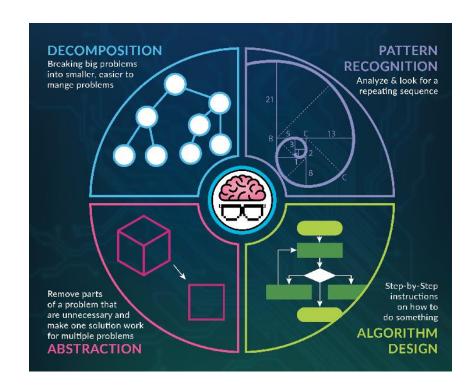
• 观察问题的模式、趋势和规律

抽象

• 识别模式形成背后的一般原理

算法设计

• 为解决某一类问题撰写一系列详细的步骤



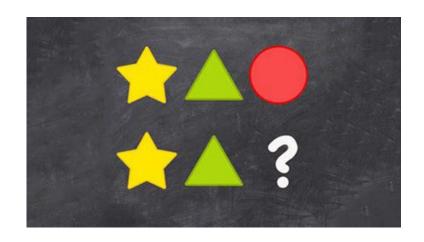
分而治之

- 分层思维帮助我们将复杂的问题拆解成小问题,把复杂的物体拆解成较轻易应付和理解的小物件,通过解决小问题而解决复杂的问题,使问题变得更加简单。
- 例:怎样制作汉堡包?我们可以将汉堡包分成几个部分:
 - Upper bun: 最上层的圆面包
 - Lettuce: 生菜
 - Tomato: 西红柿
 - Cheese: 奶酪
 - Beef patty: 牛肉馅饼
 - Lower bun: 下层的圆面包



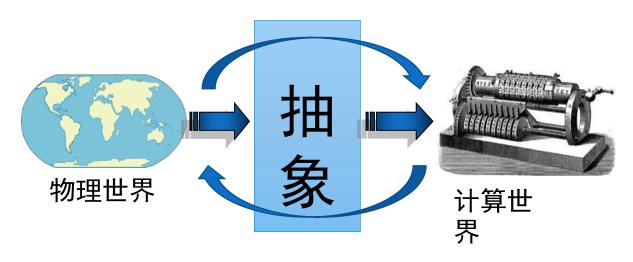
模式识别

- 任何事物都有相似性,模式识别教会我们寻找到事物之间的共同特点,利用这些相同的规律,去解决问题。当我们把复杂的问题分层到小问题时,我们经常会在小问题中找到模式,这些模式在小问题当中有相似点。
- 从以往的经验中得到规律并且举一反三将它 运用到其他的问题中,例如:
 - ✓顺序模式是按顺序排列项目(所有物品);
 - ✓分组模式是将相似的项目(物品)分成一个组。



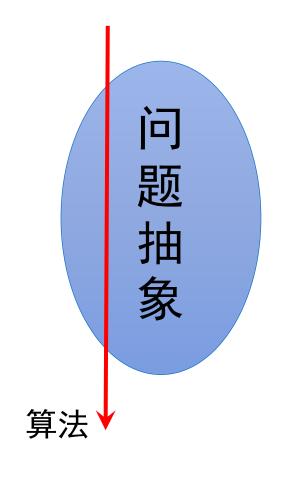
抽象

抽象化思维是将重要的信息提炼出来,去除次要信息的能力,掌握了抽象 化的能力,我们就可以将一个解决方案应用于其他事物中,制定出解决方 案的总体思路。



走向物理世界与计算世界的无缝连接

问题抽象



核心概念:

数学模型、表示、实现、转换

算法是计算思维的核心概念:

方法层: 算法 ~

表示层: 编程 🏲

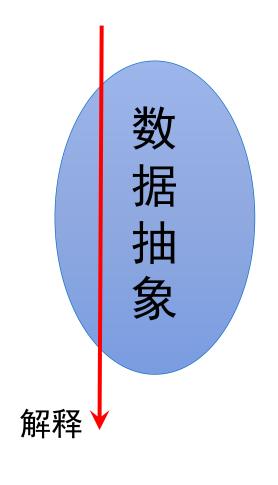
实现层: 机器 -

这差不多也就 是计算机科学 的主要内容了

问题抽象的分层映射



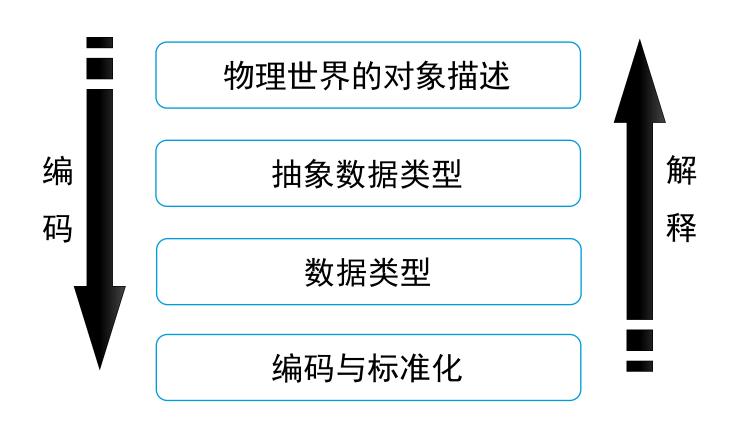
数据抽象



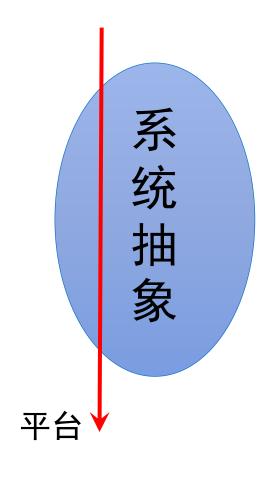
核心概念:

数据形态、数据组织、存储、检索与利用

数据抽象的分层映射



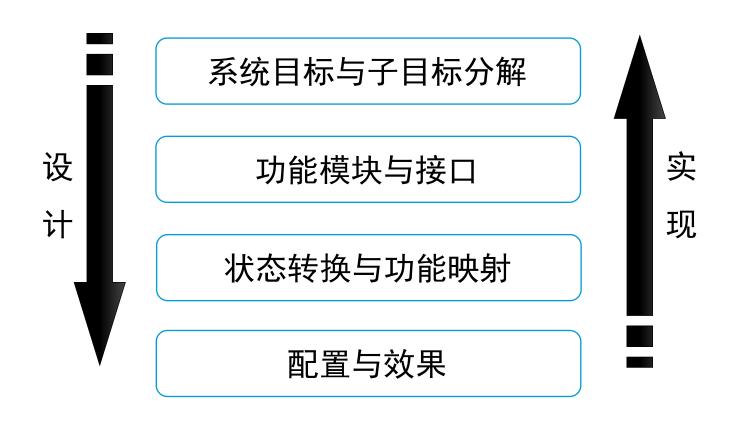
系统抽象



核心概念:

系统模型、功能逻辑、接口、实现

系统抽象的分层映射



数据抽象

• 编码 → 数据记录 → 数据库

问题抽象

- 处理单个员工的相关记录
- 处理任意有限多个员工的不同记录 (同样的处理方法)

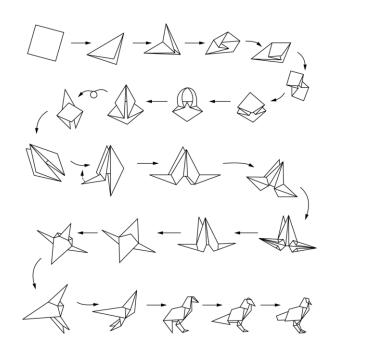
系统抽象

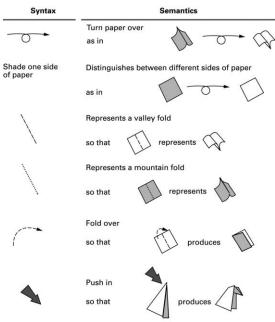
- 应用界面:某种形式的"窗口"
- 逻辑电路支撑 → 系统软件支撑 → 数据库 → 应用程序支撑

企业的工资表处理

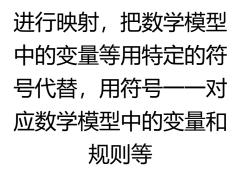
算法设计

- 一步一步解决问题的过程,按照的一定的顺序完成一个任务,同样的事情人人都会学习操作。
- An algorithm is an ordered set of unambiguous, executable steps that defines a terminating process.



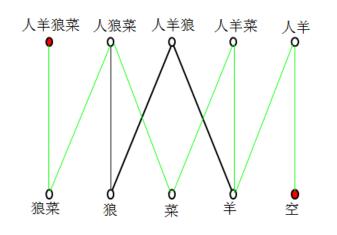


把实际问题抽象为数学 问题,并建模,将人对 问题的理解用数学语言 描述出来 通过编程把解决问题的 逻辑分析过程写成算法, 把解题思路变成计算机 指令,也就是算法



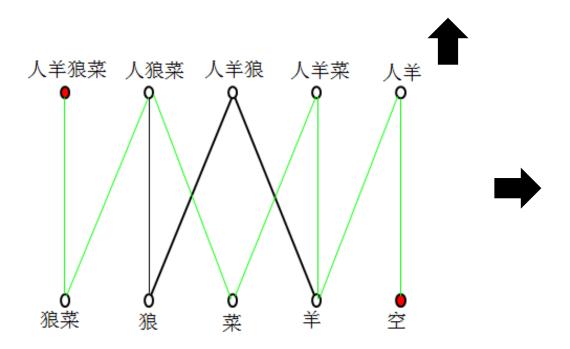
执行算法,进行求解, 计算机根据算法,一步 步完成相应指令,求出 结果

- 问题:人、狼、羊、菜用一条只能同时载两位的小船渡河,"狼羊"、"羊菜"不能在无人在场时共处,当然只有人能驾船。
- 图模型: 顶点表示"原岸的状态",两点之间有边当且仅当一次合理的渡河"操作"能够实现该状态的转变。
- 起始状态是"人狼羊菜",结束状态是"空","允许状态"可以有10个。
- 问题的解:找到一条从起始状态到结束状态的尽可能短的通路。





- 上述关系可以用一个布尔矩阵表示:
 - 它也可以表示成一个"数": 10000000011100000001010000000110......
 - 或者也可以表示成符号串: 16#28#2#6#3#768#384#320#112#32



```
      0
      0
      0
      0
      1
      0
      0
      0
      0

      0
      0
      0
      0
      0
      1
      1
      1
      0
      0

      0
      0
      0
      0
      0
      0
      1
      1
      0
      0

      0
      0
      0
      0
      0
      0
      0
      0
      1
      1

      1
      1
      0
      0
      0
      0
      0
      0
      0
      0

      0
      1
      1
      0
      0
      0
      0
      0
      0
      0

      0
      1
      0
      1
      0
      0
      0
      0
      0
      0
      0

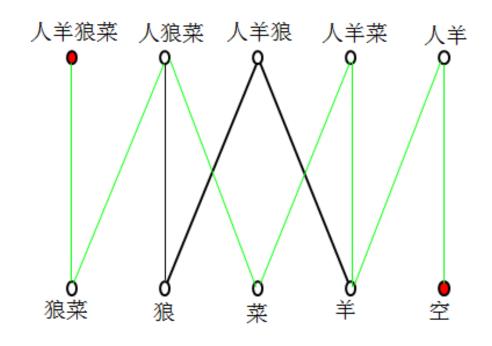
      0
      1
      0
      1
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
      0
```

方案1:

- 农夫带羊过河
- 农夫返回
- 农夫带狼过河
- 农夫带羊返回
- 农夫带菜过河
- 农夫返回
- 农夫带羊过河
- <结束>

方案2:

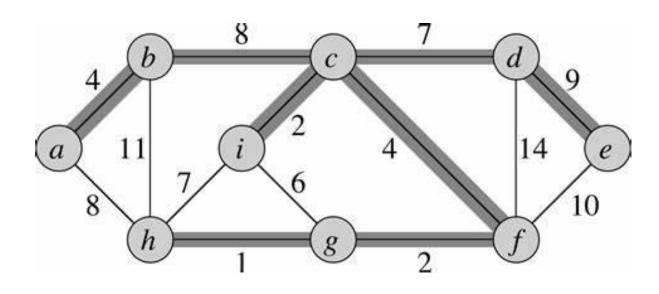
- 农夫带羊过河
- 农夫返回
- 农夫带菜过河
- 农夫带羊返回
- 农夫带狼过河
- 农夫返回
- 农夫带羊过河
- <结束>



- 数据思维是指获取数据、从数据中提取信息、论证结论可 靠性等过程中表现出来的一种思维模式,对于人类提高认 知起到巨大的作用,主要手段包括:
 - 计算方法: 利用计算机和计算思维来进行数据的处理;
 - 统计方法: 作为一种实证主义的方法, 目的是从偶然性中发现必然性, 对真理做出探究。
 - 统计学是检验理论与现实之间一致性的工作,统计学家设计收集数据的方法,提供数据特征的描述方法,并利用样本数据对总体特征做出估计、检验和预测。

什么是数据思维

• (例)行程设计:任意给出旅行的起点和终点,如何给出一个行程建议,使得在某些指标上"最短"?



Dijkstra算法或者动态规划算法来求解,算法 复杂度为O(n³)



记录物理世界人们旅行的选择,构建数据模型,对最受欢迎的路线进行排序

什么是数据思维

- 数据思维,这个概念虽然很早就有,但直到近几年,随着 大数据技术的飞速发展,重新又回到了思维认识的高度。
- 实际上,数据思维一直是人类的一种思维方式之一,而且 应该比科学思维形成得更早,也更朴实。
- 科学思维应该是在数据思维之上产生的,而且科学思维中也包括了数据思维,如一些基于统计的学科,其实就是数据思维的体现和应用。

数据思维的历史

"统计学不止是一种方法或技术,还含有<mark>世界观</mark>的成分,它是看待世界上万事万物的一种方法。我们常讲某事从统计观点看如何如何,指的就是这个意思。但统计思想也有一个发展过程。因此统计思想(或观点)的养成,不单需要学习一些具体的知识,还有能够从发展的眼光,把这些知识连缀成一个有机的、清晰的途径,获得一种历史的厚重感。"

——陈希孺,《数理统计学简史》

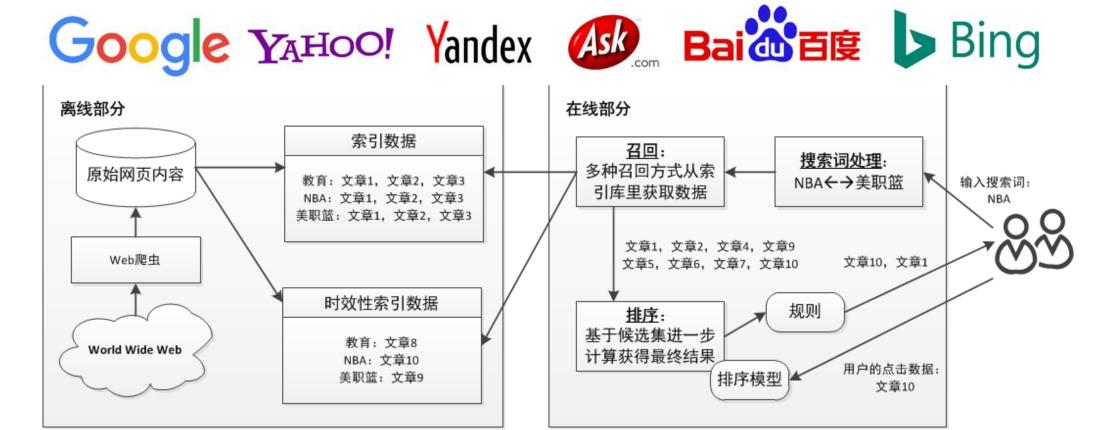
统计思维,就是用统计学的世界观去构建框架,然后去解释这个世界。

数据思维与统计学

- 计算环境的三大变革:
 - 海量数据资源 无尽的宝藏
 - 拓展的无线网络 无限的舞台
 - 智能化的设计与生产能力 无边的法力
- 不管你是从事那个领域的工作,数据爆炸和计算环境的革命为你提供了无限的创新机遇,就靠数据思维来发现了。
- 数据 + 数据思维 + 计算环境 = 无限的可能

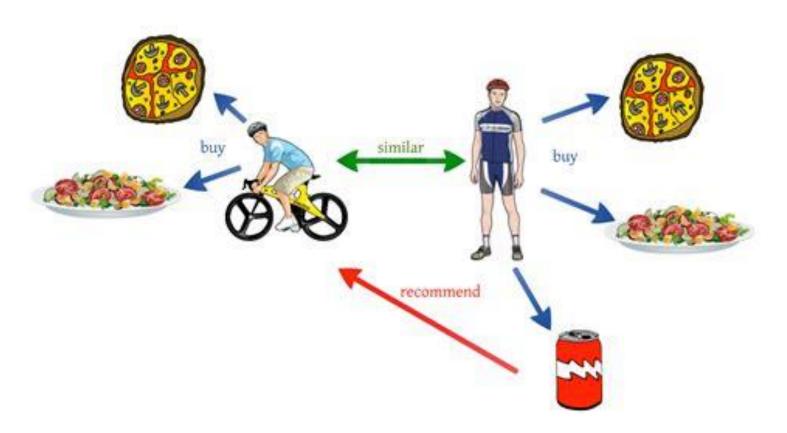
数据思维与计算环境

搜索引擎



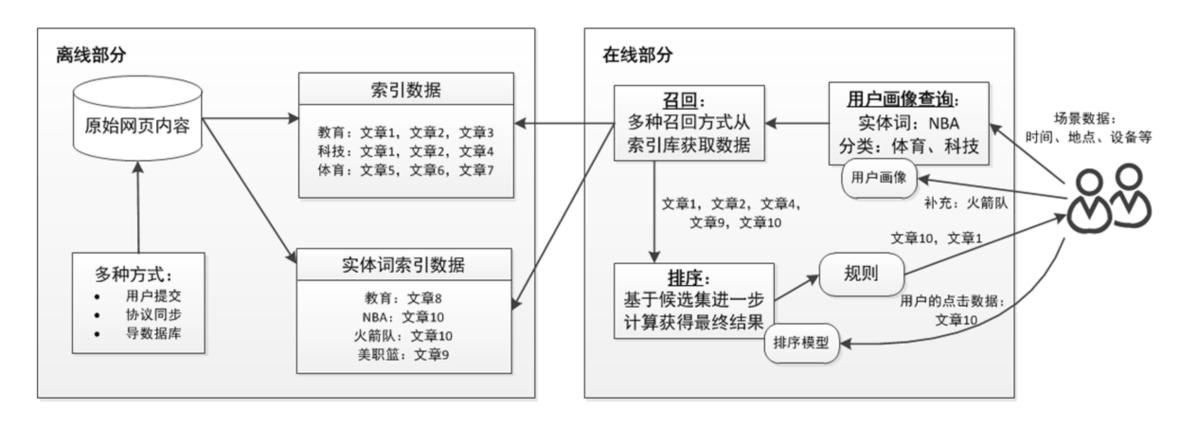
数据思维与计算环境

推荐系统



数据思维与计算环境

新闻推荐系统的架构



数据思维与计算环境



123

问题求解与思维方式

计算思维与数据思维

计算思维与数据思维实例

- 从前有座山,山里有座庙,庙里有个老和尚,正在给小和尚讲故事呢!故事是这样的:从前有座山,山里有座庙,庙里有个老和尚,正在给小和尚讲故事呢!故事是这样的:从前有座山,.....
- 递归(Recursion)是指在函数的定义中使用函数自身的方法。递归一词还较常用于描述以自相似方法重复事物的过程。



递归

• 斐波那契(Fibonacci)数列: [0,1,1,2,3,5,8,...]

```
def Fibonacci(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return Fibonacci(n - 1) + Fibonacci(n - 2)
```

•分治(Divide-and-Conquer)就是"分而治之"的意思,也就是说将一个复杂的大问题划分为若干相同或相似的、且相互独立的小问题,再把小问题继续划分为更小的问题,不断重复这个过程,直到"最小"的问题可以被直接地求解出来,然后将这些小问题的解层层合并,最终计算出原问题的解。

6 5 3 1 8 7 2 4

分治

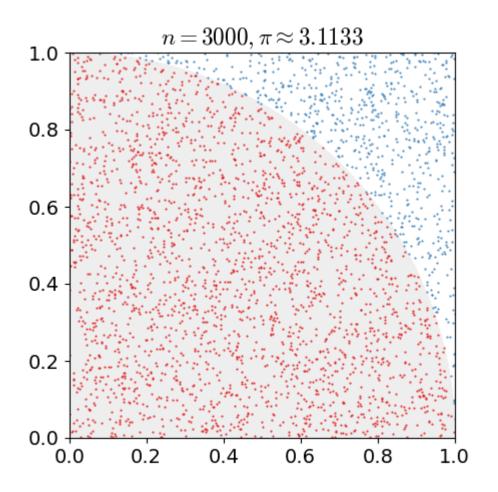
- 上面实例的精妙有趣之处,就是它对于"算法"的研究, 解决同一个问题可以设计出各种不同的算法,不是获得解 就结束了,而是要进一步分析不同算法之间对程序执行效 率的影响,然后选择最好的算法。
- "设计"就是算法研究中的最重要的问题,针对一个问题, 设计出高效的算法,而不单单是解决给定的一个问题。
- 这就是计算之美!

计算思维小结

• 蒙特卡罗方法(Monte Carlo method),也称统计模拟方法, 是1940年代中期由于科学技术的发展和电子计算机的发明, 而提出的一种以概率统计理论为指导的数值计算方法。是 指使用随机数(或更常见的伪随机数)来解决很多计算问 题的方法。

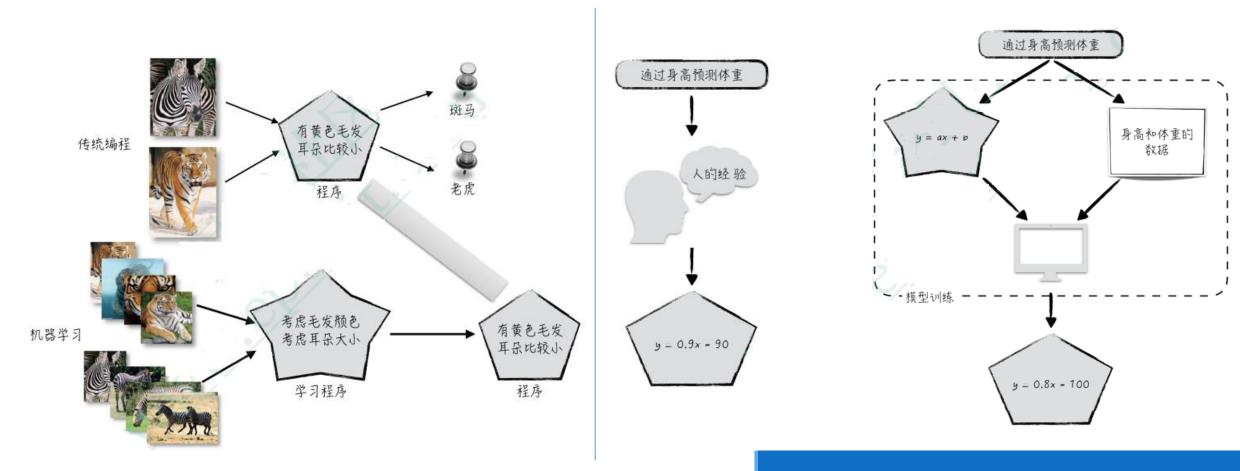
蒙特卡洛方法

```
# 蒙特卡洛法求Pi
    import random
3
    def Pi(times):
5
       sum = 0
       for i in range(times):
           x = random.random()
           y = random.random()
8
           d2 = x*x + y*y
9
10
           if d2 <= 1 :
11
              sum+=1
12
       return (sum/times*4)
                                        蒙特卡洛法求Pi
13
                                    .. import random
14
                                  >>> def Pi(times):
    times = 1000000000
15
                                           sum = 0
16
    x = Pi(times)
                                           for i in range(times):
    print ("Pi = \%.8f"%(x))
                                               x = random.random()
                                               y = random.random()
                                               d2 = x*x + y*y
                                               if d2 <= 1 :
                                                    sum+=1
                                           return (sum/times*4)
                                  >>> times = 100000000
                                  >>> x = Pi(times)
                                  >>> print ("Pi = %.8f"%(x))
              大约40秒以后
                                  Pi = 3.14172316
```

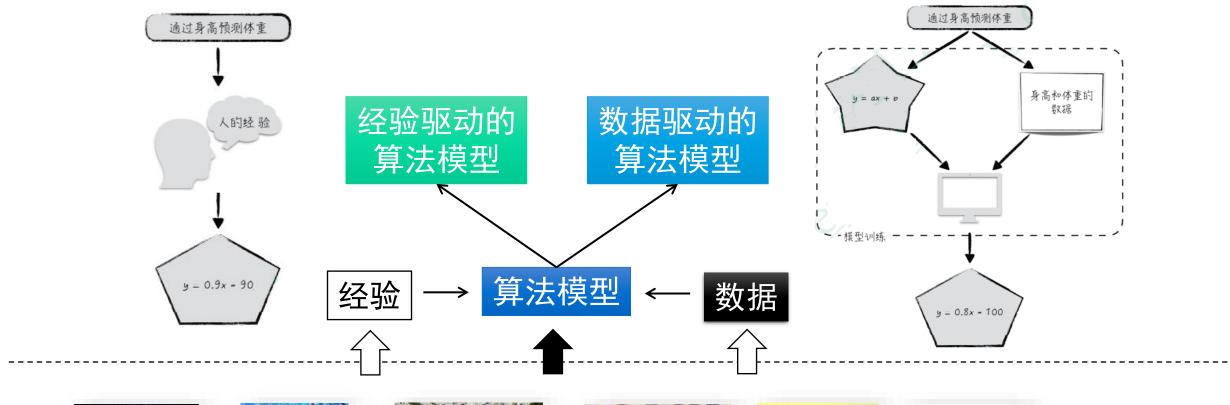


蒙特卡洛方法

• 从编程的角度来看, 机器学习是一种能自动生成程序的特殊程序。



机器学习方法















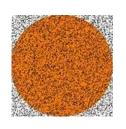
机器学习方法

- 上面实例的精妙有趣之处,就是在于以数据为中心或数据 驱动的问题求解过程,传统算法无法或很难解决的问题, 通过引入相关数据,往往就变得迎刃而解了。
- •这类问题除了需要考虑高效的算法之外,还要生产或采集 到与之相关的关键数据,才能保证问题的高质量求解;
- 这就是计算与数据之美!

数据思维小结













(数据为中心的) 问题求解









计算思维

数据思维

总结



123

问题求解与思维方式

计算思维与数据思维

计算思维与数据思维实例