小象学院
ChinaHadoop.cn

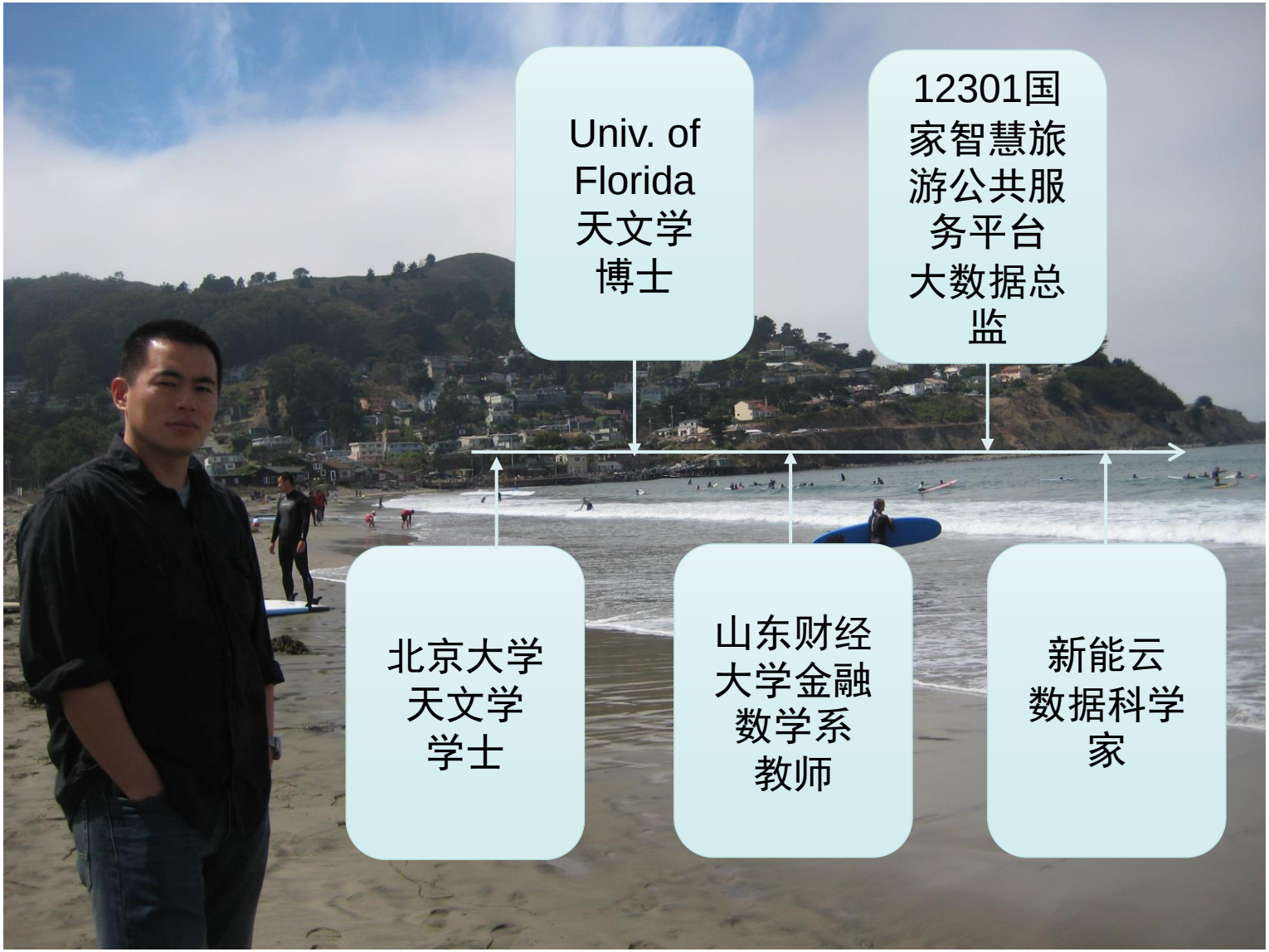
第一讲

理解数据分析和数据挖掘

- 课程介绍
- 什么是数据、数据分析和数据挖掘
- 数据行业职业规划

讲师简介：郭鹏程

小象学院 ChinaHadoop.cn



数据分析和数据挖掘

中国大数据在线教育领导者

小象学院
ChinaHadoop.cn

第一讲

理解数据分析和数据挖掘

- 课程介绍
- 什么是数据、数据分析和数据挖掘
- 数据行业职业规划

课程简介

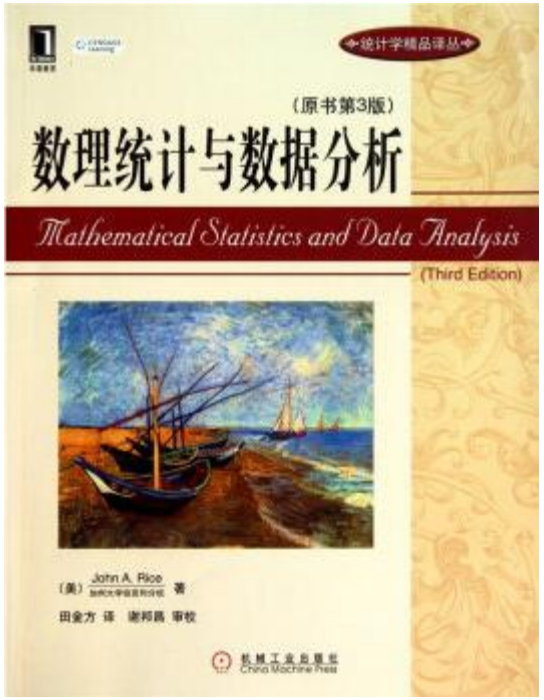
小象学院 ChinaHadoop.cn

- **面向对象**
  - 有短期工作经验，希望通过学习进入数据行业的；研究生或高年级本科生(背景调查)
  - 需要什么基础？ 高数，逻辑思维，热爱数据
- **课时安排：**
  - 预计36课时，每课时约一小时，每周六课时（暂定每周六、周日下午各三个课时，总共6周）；遇到节假日或有调整；
- **教学环境：**
  - 小象课程网站，QQ群（微信群）
  - 数据分析和挖掘的语言和环境： Python(x,y)，Anaconda

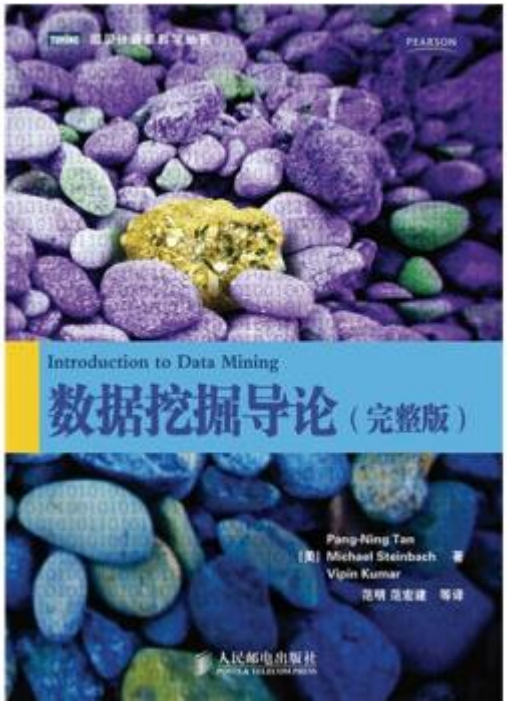
课程介绍

小象学院 ChinaHadoop.cn

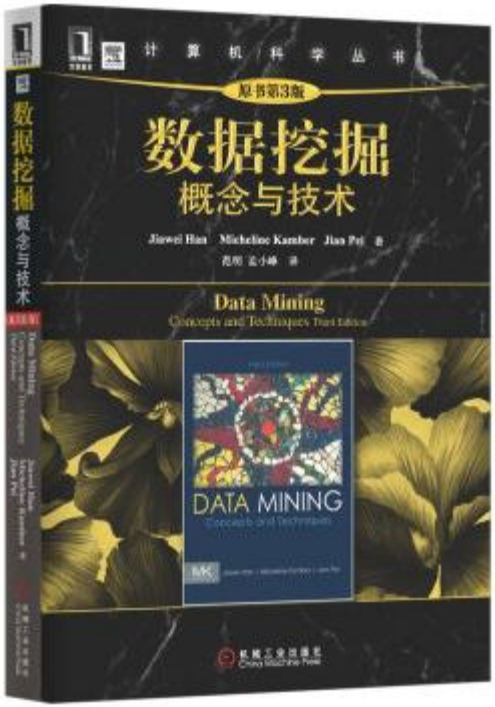
- **参考书：**



http://item.jd.com/10664952.html



http://item.jd.com/10380545.html



http://item.jd.com/11056660.html

课程设置和学习方法

- 理解数据，理解量化，理解模型
- 掌握概率统计
- 掌握一门工具
- 学会处理数据：收集、转换、载入
- 学会数据分析：分析思想和分析方法
- 学会基本的数据挖掘：

小象学院
ChinaHadoop.cn

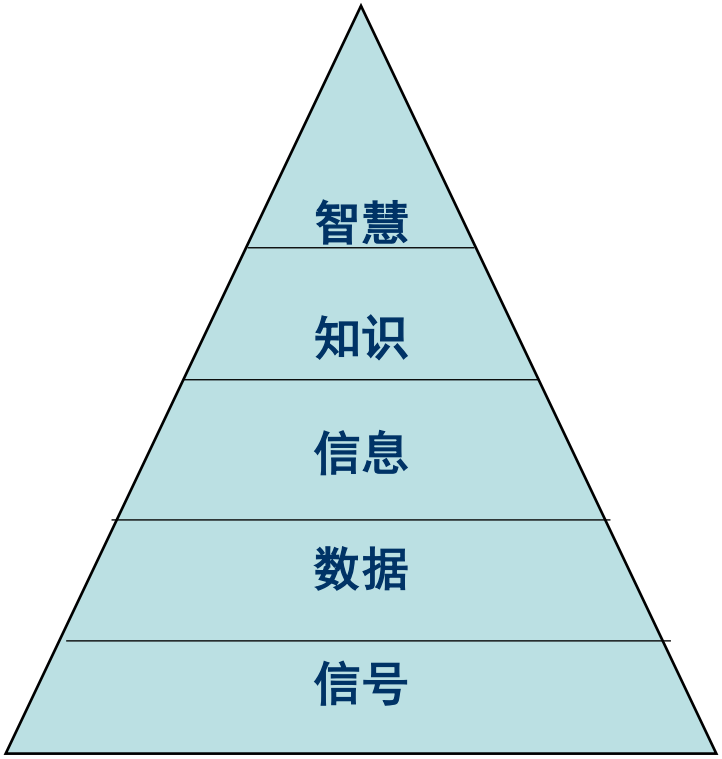
第一讲

理解数据分析和数据挖掘

- 课程介绍
- 什么是数据、数据分析和数据挖掘
- 数据行业职业规划

什么是数据？

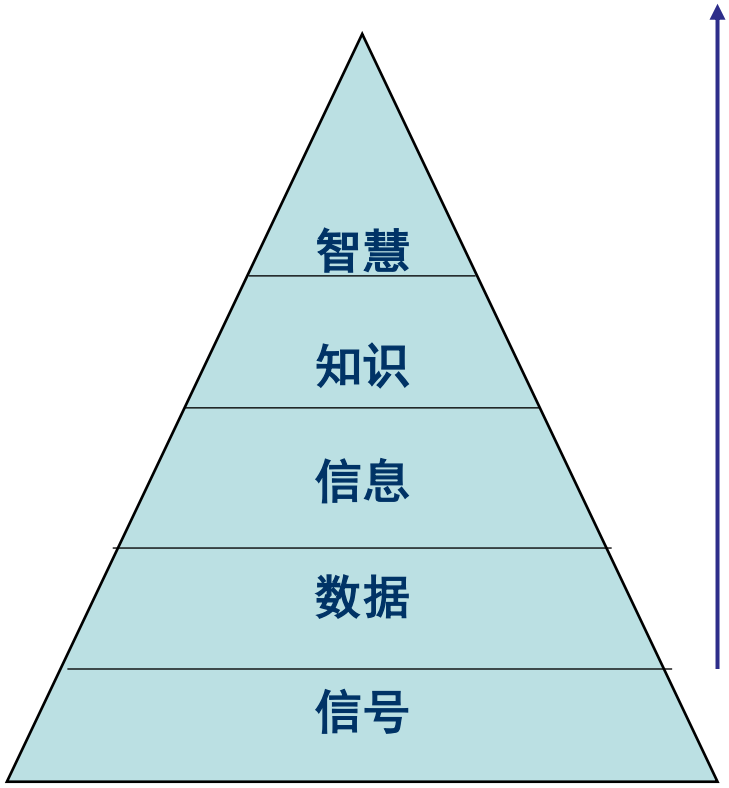
- 描述对象属性的
- 一切可以被记录的



什么是数据？

小象学院 ChinaHadoop.cn

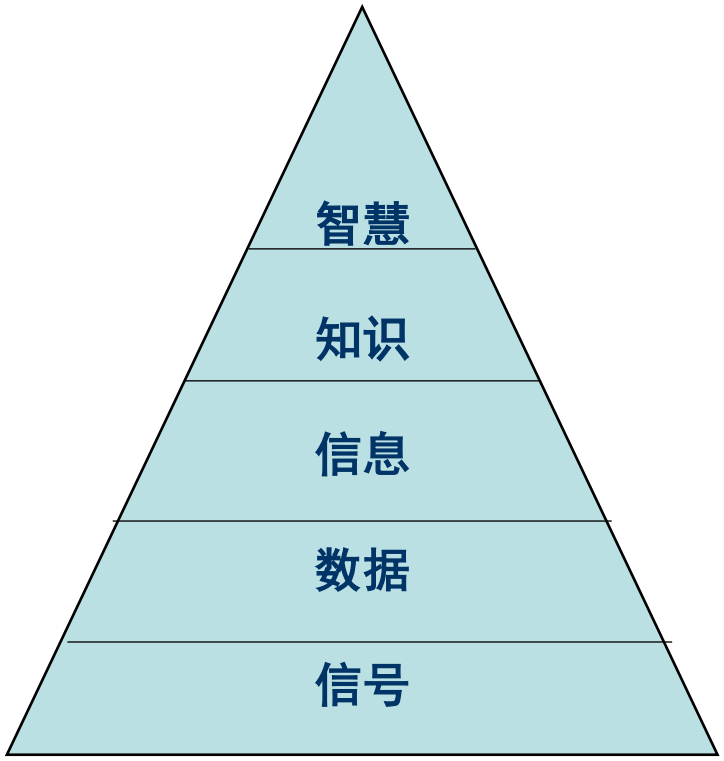
- 描述对象属性的
- 一切可以被记录的



你要怎么办？
02-14是情人节，通常伴随着比较高的表白成功率
02-14相应消费增高
对信号赋予意义：月-日 02-14
通过物理机制感知的：1,2,3,4

什么是数据？

- 描述对象属性的
- 一切可以被记录的



你要怎么办？
与数据行业高度相关….
统计后发现：“机器学习”高频
赋予意义：机器学习
音节：ji，xue，xi，ji

什么是数据

小象学院 ChinaHadoop.cn

- **一切可以被记录的**
  - 没有什么不可以被记录
  - 差别只在于记录的精度

- **描述对象属性的**
  - 描述对象的过程就是将对象抽象为若干个可以度量的属性
  - 是一个量化的过程，也是**数学建模**的第一步
  - 量化的属性，就是变量

数据的类型

小象学院 ChinaHadoop.cn

- 是否有序，数值/类别，连续/离散

属性类型	表述和允许的变换	例子	操作
Nominal 标称 （分类、定性）	与其他对象区别的名称 (=, ≠) 一对一变换	邮编、ID、姓名、性别	众数, 熵, 列联相关 χ^2 检验
Ordinal 序数 （分类、定性）	确定对象信息的序. (<, >) 保序变换	矿石硬度、成绩、街道号码	中值，百分位，秩相关
Interval 区间 （数值、定量）	区间属性，值之间的和差是有意义的 (+, -) 线性变换	日期，温度	均值、标准差、Pearson相关
Ratio 比例 （数值、定量）	比率变量，积和比有意义 (*, /) 线性变换（乘积）	绝对零度、货币量、计数、年龄	几何平均，调和平均，百分比变差

如何产生数据

小象学院 ChinaHadoop.cn

- **1. 用什么类型的数据来描述雨量？**
  - A. 离散无序变量
  - B. 离散有序变量
  - C. 连续变量
  - D. 顺序变量

- **2. 用什么类型的数据来描述客户的信用度？**
  - A. 离散无序变量（类别）
  - B. 离散有序变量（类别）
  - C. 连续变量
  - D. 顺序变量

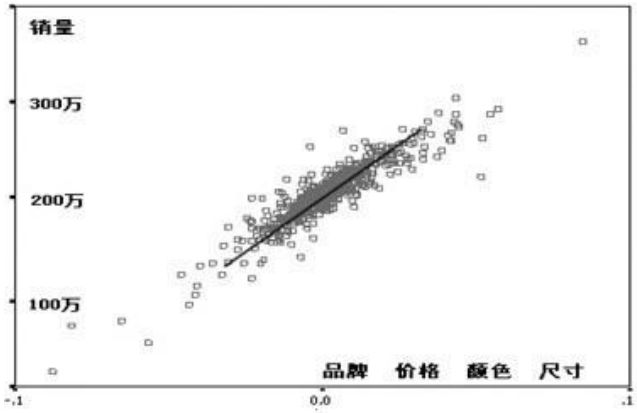
- **考查两点：信息量，获得方式**

什么是模型

小象学院 ChinaHadoop.cn

- 变量与变量之间的关系: x, y
- 定量：线性模型中的系数
- 定性：找到合适的模型描述x，y的关系

$$y = ax + b + \varepsilon$$



$$y = ax^2 + bx + c + \varepsilon$$

什么是数据分析

小象学院 ChinaHadoop.cn

- 根据变量类型和一定的假设（确定某类模型），来确定变量与变量之间的关系
- 当变量之间没有关系时？

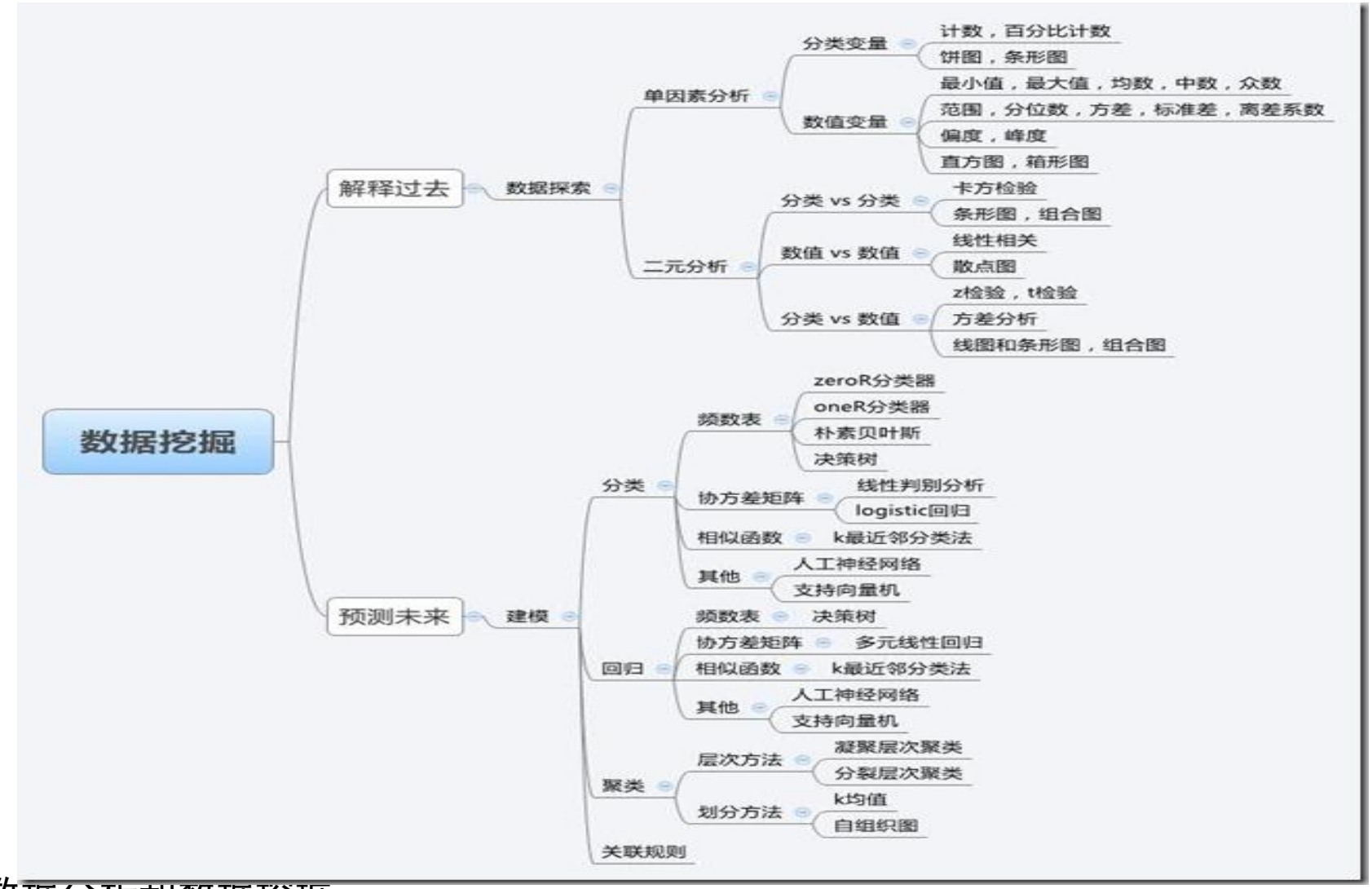
随机变量*　　随机变量*

$$y = b + \varepsilon$$

什么是数据挖掘

小象学院 ChinaHadoop.cn

- **发现新的模型（知识发现）**



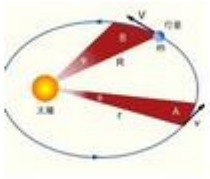
一个例子

小象学院 ChinaHadoop.cn

- **跟天文有关**

第谷

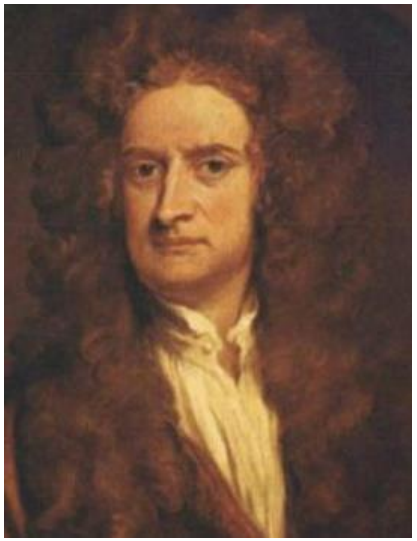
开普勒第一...



开普勒第二...

$$\frac{T^2}{a^3} = k$$

开普勒第三...



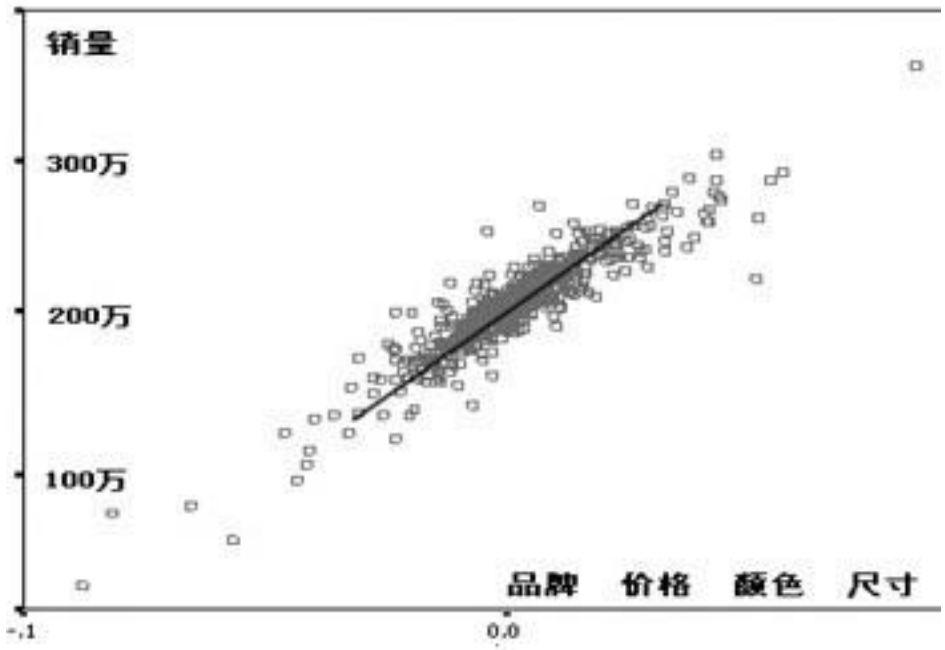
牛顿运动定律

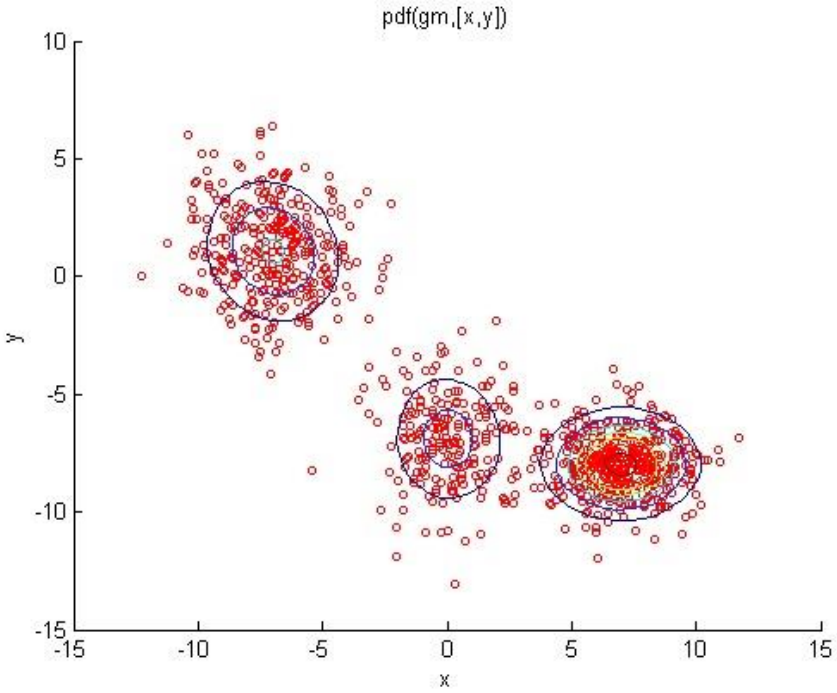
爱因斯坦
狭义相对论
广义相对论

再来一个例子

小象学院 ChinaHadoop.cn

- 商业应用





所有模型都是错的，但有些是有用的

All models are wrong, but some are useful.

数据分析和数据挖掘的关系

- **数据的用途：**
  - 记录、解释（理解）、预测、控制
- **数据分析：**
  - 统计、相关、回归；
  - 已知模式下的参数估计
- **数据挖掘：**
  - 发现模式（发现知识）
  - 分类、关联、聚类、回归
- **数据-信息-知识-智慧**

数据分析和数据挖掘的实现

- **离线分析和挖掘**
  - 静态数据，R，Python
- **在线分析和挖掘：**
  - OLAP，OLTP，大数据平台
- **实际开发时的建议**
  - 尽可能的链接各种数据源：使用统一身份认证取得关联依据
  - 保证数据质量
  - 冗余采集和存储：建立数据仓库
  - 业务人员与数据人员紧密合作建立分析模型：IT技术不是最重要的，但是不可或缺的
  - 充分利用第三方工具：降低成本

小象学院
ChinaHadoop.cn

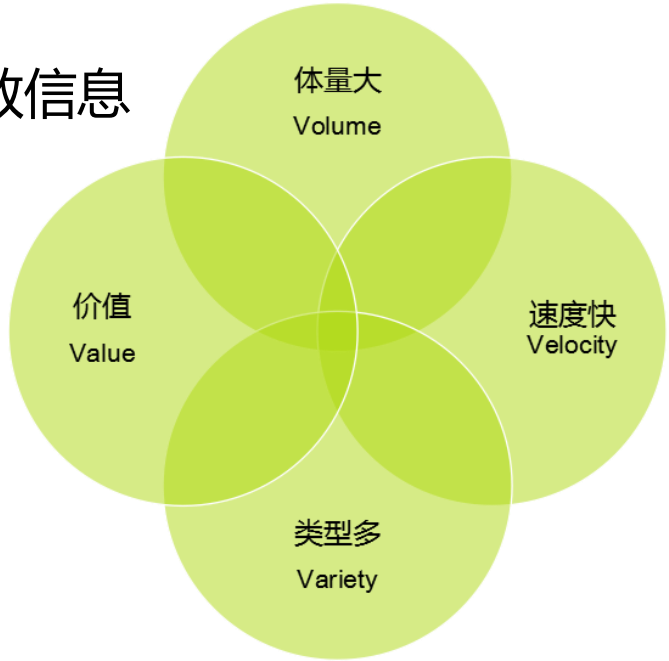
第一讲

理解数据分析和数据挖掘

- 课程介绍
- 什么是数据、数据分析和数据挖掘
- 数据行业职业规划

大数据分析和挖掘

- 数据获取更容易，快速产生大量数据
- 4V+1高维：多维数据分析
- 原始数据价值密度低
- 分析处理可提高价值密度
- 需要高效的分析、处理方案，快速提取有效信息



数据行业的职位类型

小象学院 ChinaHadoop.cn

- **数据库工程师**：维护数据，数据库开发
- **大数据工程师**：利用大数据平台维护、（简单）分析数据
- **模型工程师**：进行数学建模，将业务问题转化为数学问题
- **算法工程师**：算法优化等
- **数据工程师**：进行数据挖掘和数据分析
- **数据分析师**：对数据进行分析，并结合业务进行解读
- **数据运维媛**：检测数据质量，通过用户和产品运维保障数据（增长黑客）
- **精算师**：根据模型和数据计算（如定价），不只是在保险行业
- **运筹工程师**：根据业务和模型寻求最优解
- **可视化工程师**：将数据通过最好的UI方式呈现，便于理解和分析

Python数据分析环境

小象学院 ChinaHadoop.cn

SciPy.org Sponsored By ENTHOUGHT

Install | Getting Started | Documentation | Report Bugs | SciPy Central | Blogs

SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

NumPy
Base N-dimensional array package

SciPy library
Fundamental library for scientific computing

Matplotlib
Comprehensive 2D Plotting

IP[y]: IPython

IPython
Enhanced Interactive Console

Sympy
Symbolic mathematics

pandas
Data structures & analysis

More information...

http://www.scipy.org/install.html

SciPy.org Sponsored By ENTHOUGHT

SciPy.org

Installing the SciPy Stack

These are instructions for installing *the full SciPy stack*. For installing individual packages, such as NumPy and SciPy, see *Windows packages* below.

Scientific Python distributions

For most users, especially on Windows and Mac, the easiest way to install the packages of the SciPy stack is to download one of these Python distributions, which includes all the key packages:

- Anaconda: A free distribution for the SciPy stack. Supports Linux, Windows and Mac.
- Enthought Canopy: The free and commercial versions include the core SciPy stack packages. Supports Linux, Windows and Mac.
- Python(x,y): A free distribution including the SciPy stack, based around the Spyder IDE. Windows only.
- WinPython: A free distribution including the SciPy stack. Windows only.
- Pyzo: A free distribution based on Anaconda and the IEP interactive development environment. Supports Linux, Windows and Mac.

Linux packages

Users on Linux can quickly install the necessary packages from repositories.

Ubuntu & Debian

```
sudo apt-get install python-numpy python-scipy python-matplotlib ipython ipython-notebook python-pandas python-sympy python-nose
```

The versions in Ubuntu 12.10 or newer and Debian 7.0 or newer meet the current SciPy stack specification. Users might also want to add the NeuroDebian

小象学院
ChinaHadoop.cn

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn