数据探索

张冬松

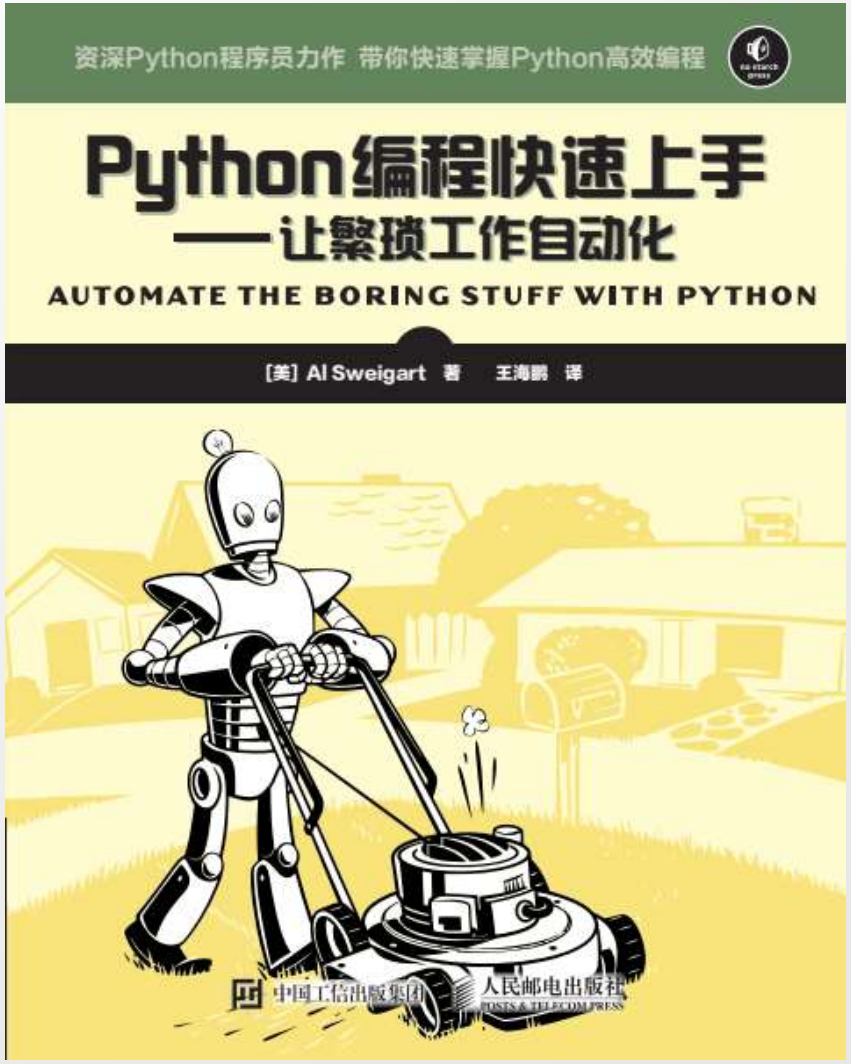
信阳学院
大数据与人工智能学院

dszhang@nudt.edu.cn

讲课前——Python

Python 黑魔法指南

作者：王炳明
版本：v2.0
发布时间：2020年05月12日
更新时间：2020年08月1日
微信公众号：Python编程时光
联系邮箱：wongbingming@163.com
Github：https://github.com/iswbm/magic-python



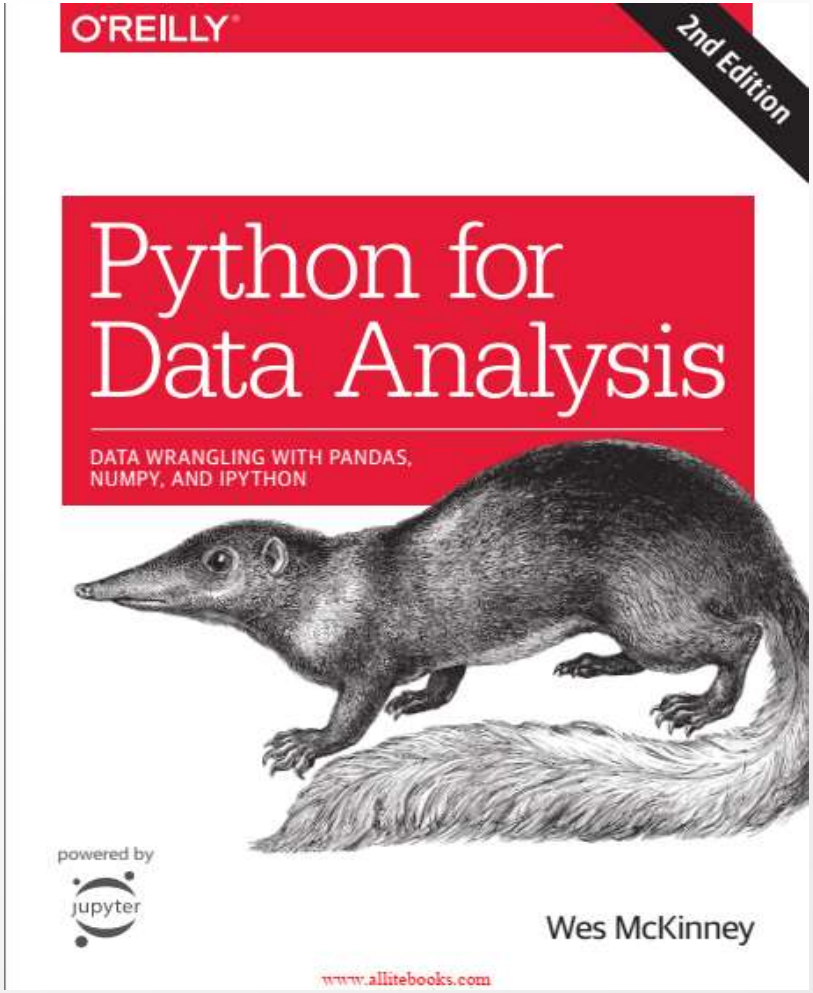
讲课前——PyCharm

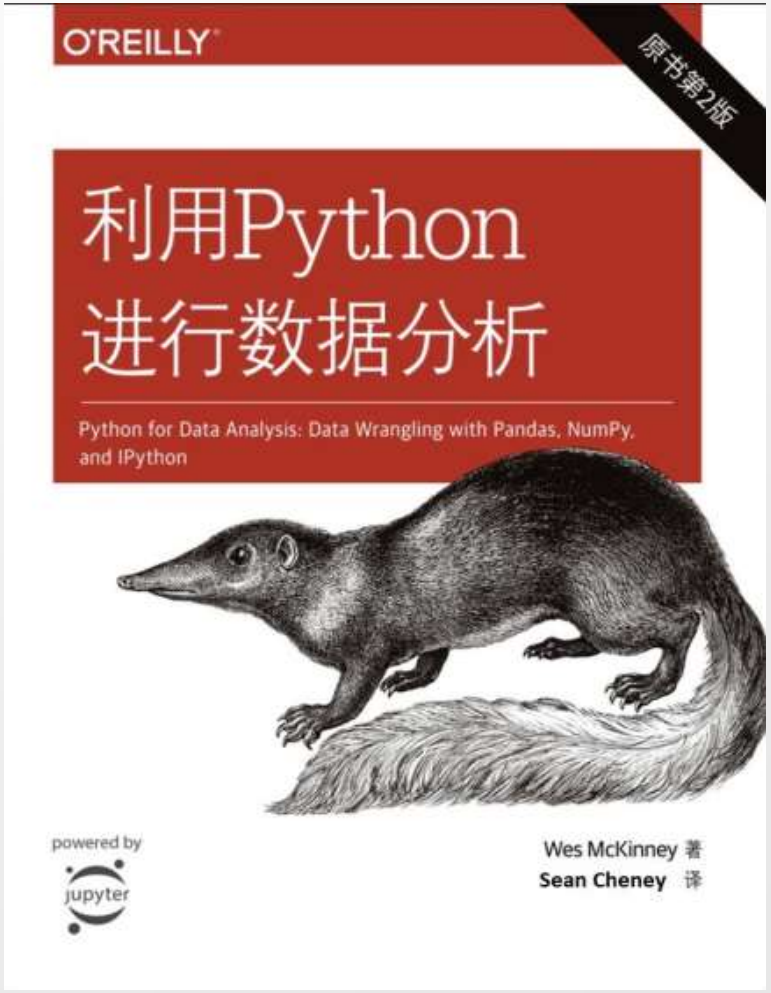
PyCharm 中文指南

作者：王炳明
版本：v1.0
发布时间：2020年08月30日
微信公众号：Python编程时光
联系邮箱：wongbingming@163.com
项目主页：http://pycharm.iswbm.com
Github：https://github.com/iswbm/pycharm-guide

回复"pycharm"，获取最新版 PDF

讲课前——Python for Data Analysis





目录 content

第一节 数据探索定义

第二节 数据汇总设计

第三节 数据可视化

第一节　数据探索定义

1.1 数据探索定义

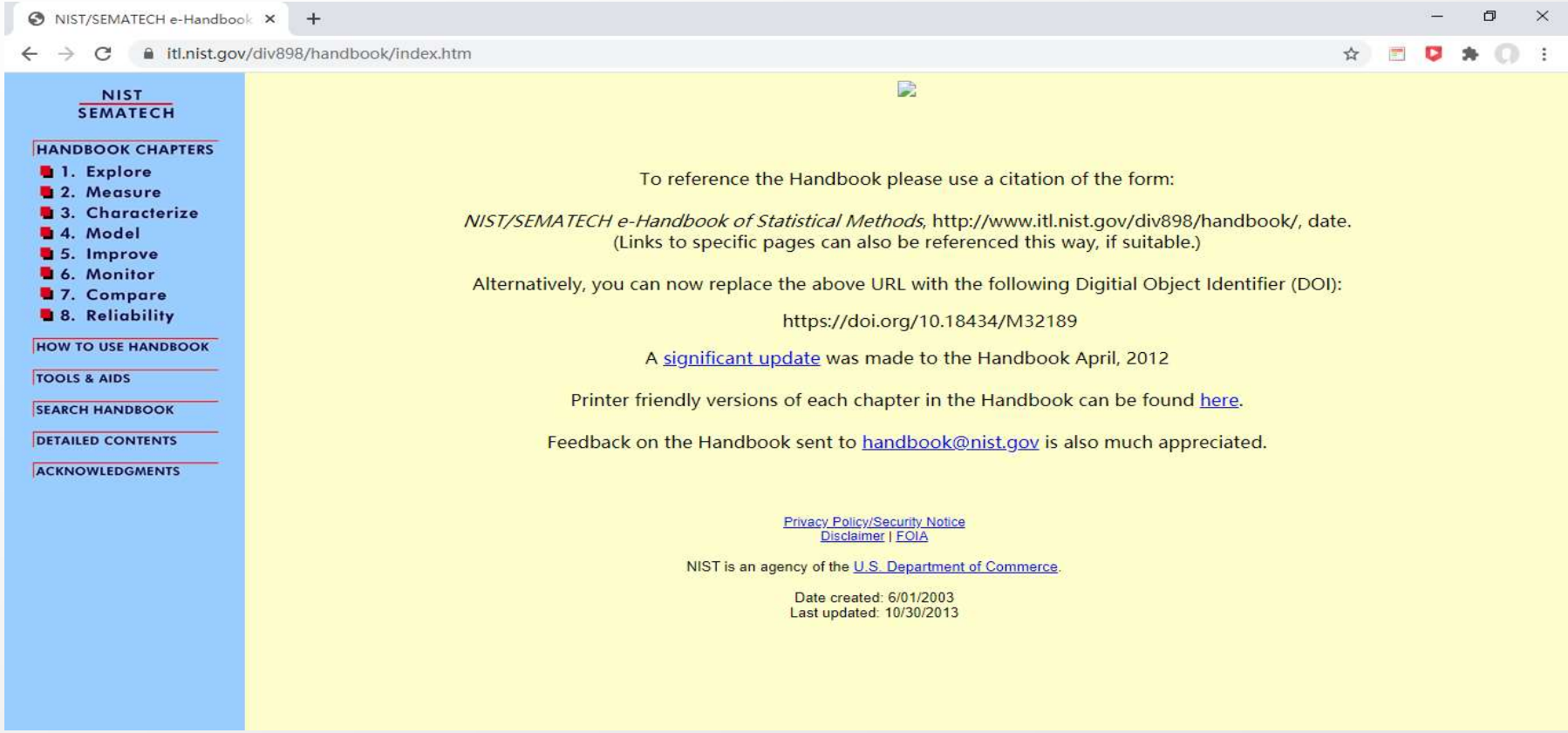
- 目的
  - 有助于选择合适的数据预处理和数据分析技术
  - 借助人的能力发现数据分析工具没有捕捉到的模式
- 探索性数据分析（Exploratory Data Analysis, EDA）
  - 统计学家 John Tukey于20世纪70年代创建

http://www.itl.nist.gov/div898/handbook/index.htm

1.1 数据探索定义

- 探索性数据分析（Exploratory Data Analysis, EDA）

http://www.itl.nist.gov/div898/handbook/index.htm



第二节　数据汇总设计

- 频率和众数
- 百分位数
- 位置度量：均值和中位数
- 散布度量：极差和方差

2 数据汇总统计

- 汇总统计是量化的，用单个数或数的小集合捕获可能很大的值集的各种特征
  - 家庭的平均收入、考试优秀率
  - 不同类型的属性值适用不同的汇总统计方法

2.1 频率与众数

- **频率：**给定一个在$\{v_1, \ldots, v_i, \ldots, v_k\}$上取值的分类属性$x$和$m$个对象的集合，值$v_i$的频率

$$frequency(v_i) = \frac{\text{具有属性值的}v_i\text{的对象数}}{m}$$

- **众数：**具有最高频率的值

2.1 频率与众数

例1. 考虑学生的集合。学生具有“年级”属性，取值{一年级、二年级、三年级、四年级}

年级	人数	频率
一年级	200	?
二年级	160	
三年级	130	
四年级	110	

属性“年级”不同取值的频率是？众数是？

2.1 频率与众数

例1. 考虑学生的集合。学生具有“年级”属性，取值{一年级、二年级、三年级、四年级}

年级	人数	频率
一年级	200	0.33
二年级	160	0.27
三年级	130	0.22
四年级	110	0.18

属性“年级”的众数是“一年级”

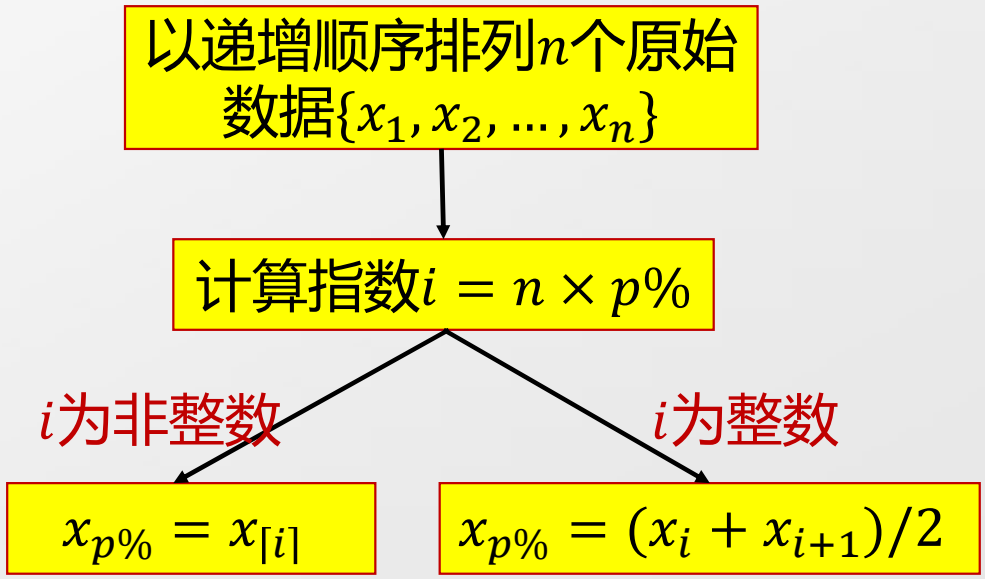
第二节　数据汇总设计

- 频率和众数
- 百分位数
- 位置度量：均值和中位数
- 散布度量：极差和方差

2.2 百分位数

- **定义：**给定一个有序的分类属性或连续属性x和$p(0 \le p \le 100)$，第p个百分位数$x_{p\%}$是一个值，使得x中$p\%$的观测值小于$x_{p\%}$

- **计算方法：**
  - $x_{0\%} = min(x)$
  - $x_{100\%} = max(x)$



2.2 百分位数

例2. 从1到10的整数的百分位数$x_{0\%}, x_{15\%}, \ x_{50\%}, \ x_{100\%}$分别为多少?

2.2 百分位数

例2. 从1到10的整数的百分位数$x_{0\%}, x_{15\%}, x_{50\%}, x_{100\%}$分别为多少?

$$x_{0\%} = \min(x) = 1$$

$$x_{15\%} = x_{[1.5]} = 2$$

$$x_{50\%} = \frac{x_5 + x_6}{2} = 5.5$$

$$x_{100\%} = \max(x) = 10$$

第二节　数据汇总设计

- 频率和众数
- 百分位数
- 位置度量：均值和中位数
- 散布度量：极差和方差

2.3 位置度量：均值和中位数

- 考虑m个对象的集合和属性x，设$\{x_{(1)},\ldots,x_{(m)}\}$是这$m$个对象以非递减序排序后的$x$属性值，则**均值**和**中位数**定义如下：

$$mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$median(x) = \begin{cases} x_{(r+1)} & \text{如果}m\text{是奇数，即}m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{如果}m\text{是偶数，即}m = 2r \end{cases}$$

2.3 位置度量：均值和中位数

- 均值和中位数的区别
  - 考虑值集{1，2，3，4，5，90}。其均值是17.5？中位数是3.5？

- 均值对离群点很敏感；对包含离群值的数据，中位数可以更稳健地提供值集中间的估计
- 当值以对称的方式分布式，均值可解释为值集的中间

2.3 位置度量：均值和中位数

- **截断均值**：指定$p(0 \le p \le 100)$，丢掉高端和低端$\left(\frac{p}{2}\right)\%$的数据，然后用常规的方法计算均值。
  - 克服传统均值对离群值敏感的问题
  - 中位数是$p = 100$时的截断均值
  - 标准均值是$p = 0$时的截断均值

考虑值集{1，2，3，4，5，90}，$p = 40$时的截断均值？3.5

第二节　数据汇总设计

- 频率和众数
- 百分位数
- 位置度量：均值和中位数
- 散布度量：极差和方差

2.4 散布度量：极差和方差

- **散布度量：**表明属性值是否散布很宽，或者是否相对集中在单个点（如均值）附近
- 给定属性x, 它具有m个值$\{x_1, x_2, \ldots, x_m\}$,x的**极差**定义为：

$$range(x) = \max(x) - \min(x)$$

- **方差**：

$$variance(x) = {s_x}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x - \bar{x})^2$$

2.4 散布度量：极差和方差

考虑值集{1，2，3，4，5，90}，极差? 方差?

极差：89 方差：1053 结论：数据分散

- 极差和方差都对离散值敏感
- 无法稳健地描述数据分布不均匀的情况下的散布度

2.4 散布度量：极差和方差

- **绝对平均偏差（absolute average deviation, AAD)**

$$AAD(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$$

- **中位数绝对偏差(median absolute deviation, MAD)**

$$MAD(x) = median(\{|x_1 - \bar{x}|, \ldots, |x_m - \bar{x}|\})$$

- **四分位数极差(interquartile range, IQR)**

$$interquartile\ range(x) = x_{75\%} - x_{25\%}$$

第三节 数据可视化

- 定义和作用
- 基本概念
- 技术
- 注意事项

3.1 定义和作用

- **定义：**以图形或表格的形式显示信息，以便能够借此分析或报告数据的特征和数据项或属性之间的关系
- **目标：**形成可视化信息的人工解释和信息的意境模型

3.1 定义和作用

- **作用：**
  - 快速吸收大量可视化信息，并发现其中的模式
  - 利用“锁在人脑袋中”的领域知识：将数据分析结果提供给领域专家

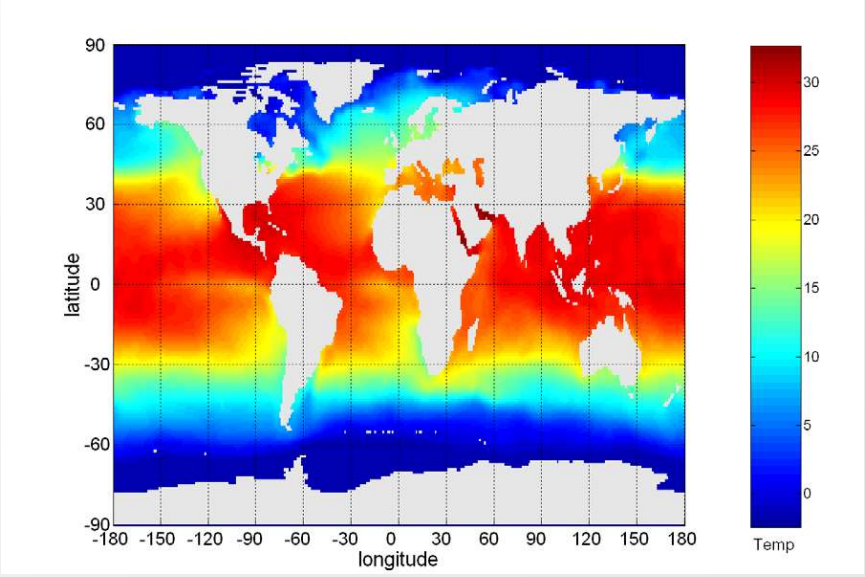


图 1982年7月的海洋表面温度（SST）

第三节　数据可视化

- 定义和作用
- 基本概念
- 技术
- 注意事项

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 数据对象
    - 单个属性：表的项或屏幕的区域
    - 表的一行（或列），或显示为图的一条线

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

图 信用卡欺诈数据

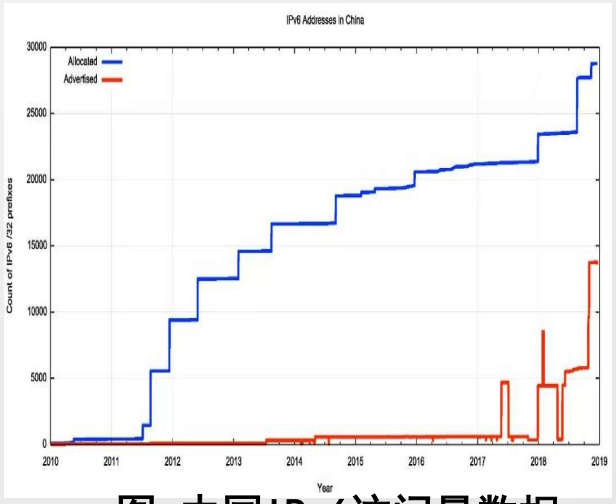


图 中国IPv6访问量数据

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 数据对象
    - 单个属性：表的项或屏幕的区域
    - 表的一行（或列），或显示为图的一条线
    - 二维或三维空间中的点

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 属性
    - 连续属性可以映射为连续的、有序的图形特征

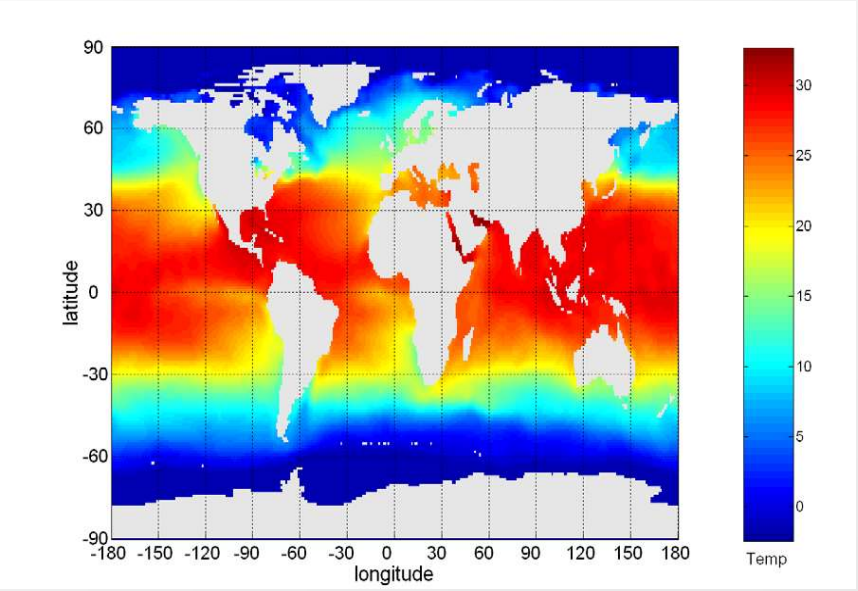


图 1982年7月的海洋表面温度（SST）

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 属性
    - 连续属性可以映射为连续的、有序的图形特征
    - 分类属性中每个类别可以映射到不同的位置、颜色、形状、方位

图　长沙天气预报图

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 属性
    - 连续属性可以映射为连续的、有序的图形特征
    - 分类属性中每个类别可以映射到不同的位置、颜色、形状、方位
    - 特别的，标称属性通常表示为无序的图形特征

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 联系
    - 显式的：点和点之间的连线

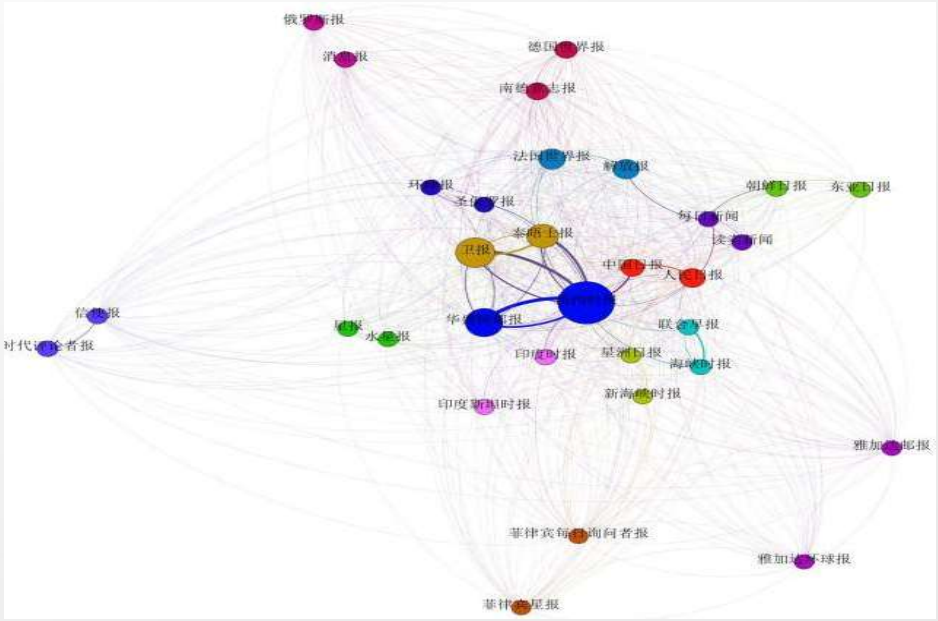


图　主流媒体关联数据

3.2 基本概念

- **表示：**将数据映射到图形元素
  - 联系
    - 显式的：点和点之间的连线
    - 隐式的：如对应数据对象的图形位置的相对位置趋向于保持对象的实际相对位置

第三节　数据可视化

- 定义和作用
- 基本概念
- 技术
- 注意事项

3.3 技术

- 少量属性的可视化
  - 显示单个属性观测值的分布
    - 直方图、盒装图、饼图
  - 显示两个属性之间的关系
    - 散布图

3.3 技术

- 少量属性的可视化
  - **直方图：**通过将可能的值分散到箱中，并显示落入每个箱中的对象数，显示属性的分布。
    - 分类属性: 每个值一个箱，值过多时可考虑合并
    - 连续属性：将值域划分为箱（通常等宽，但不必等宽）

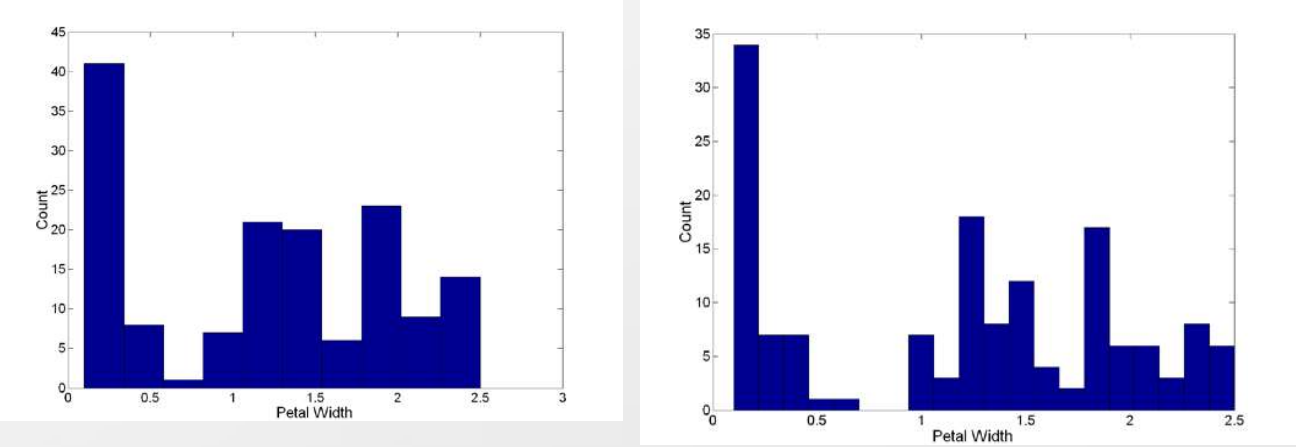
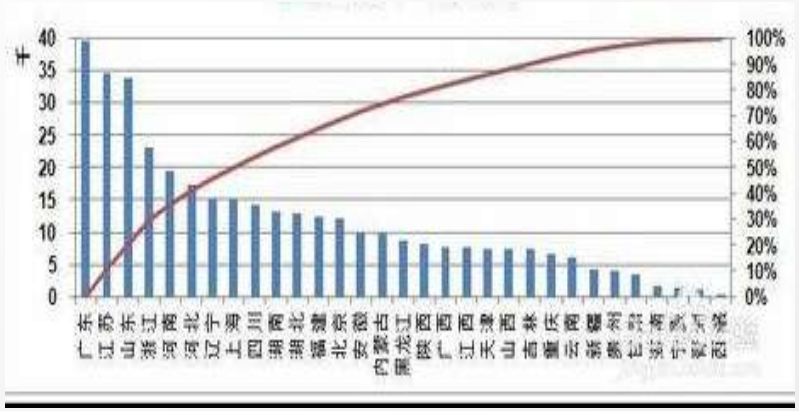


图　鸢尾花数据集花瓣宽度的直方图（10个箱 vs. 20个箱）

3.3 技术

- 直方图的变形
  - 相对频率直方图
  - Pareto直方图
  - 二维直方图



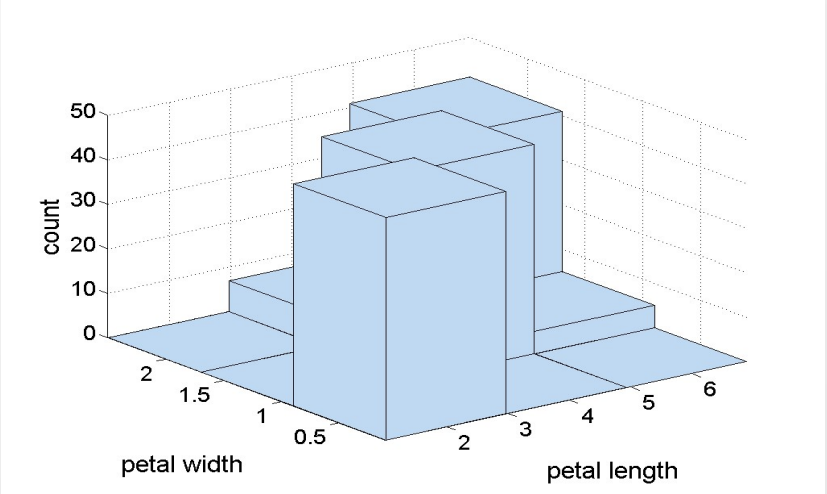
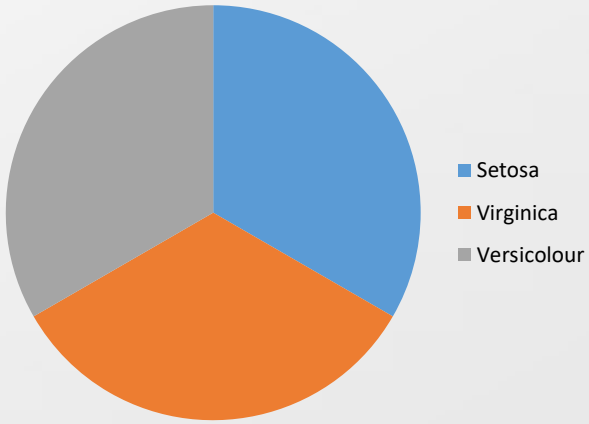


图 鸢尾花数据集花瓣长度和花瓣宽度的二维直方图

3.3 技术

- 饼图
  - 适用于具有较少值的分类属性
  - 使用圆的相对面积显示不同值得相对频率



3.3 技术

- 散布图
  - 图形化地显示两个属性之间的关系
  - 当类标号给出时，考察两个属性将类分开的程度
- 散布图矩阵
  - 同时考察多对属性

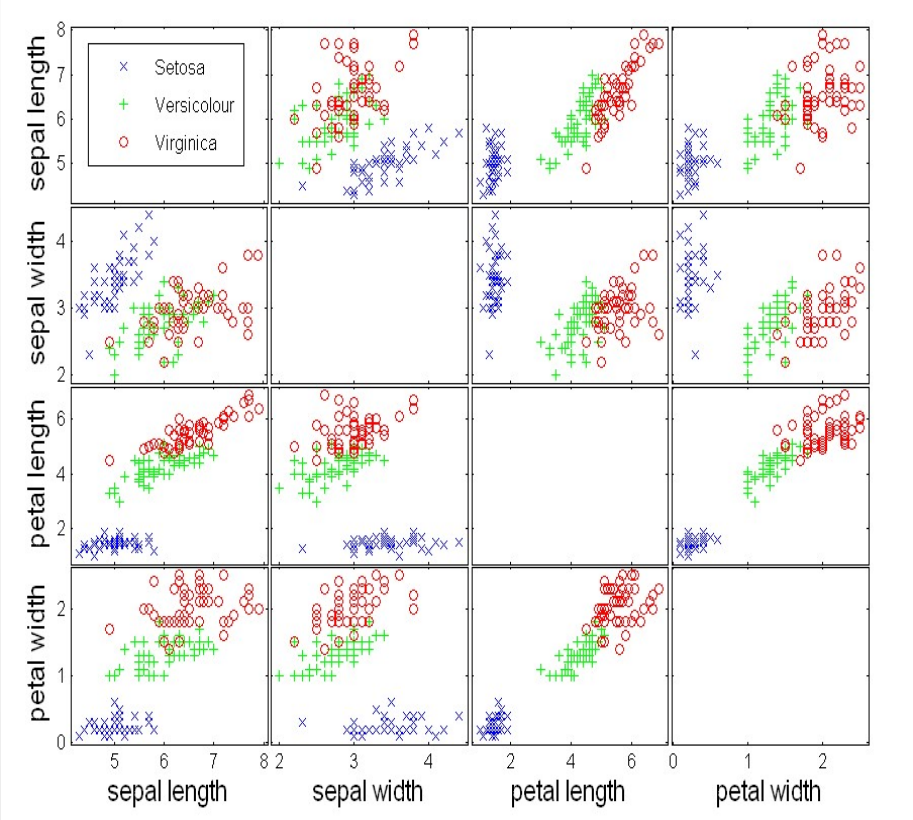


图　鸢尾花数据集的散布图矩阵

3.3 技术

- 可视化时间空间数据
  - 等高线图
    - 两个属性制定平面适当的位置，而第三个属性具有连续值

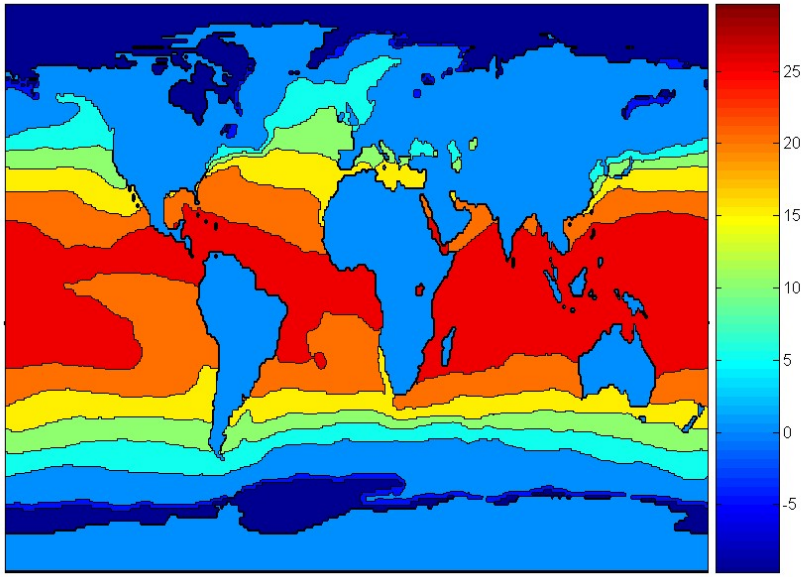


图 1998年12月份SST的等高线图

3.3 技术

- 可视化时间空间数据
  - 等高线图
  - 曲面图
  - 矢量场图
  - 低维切片图

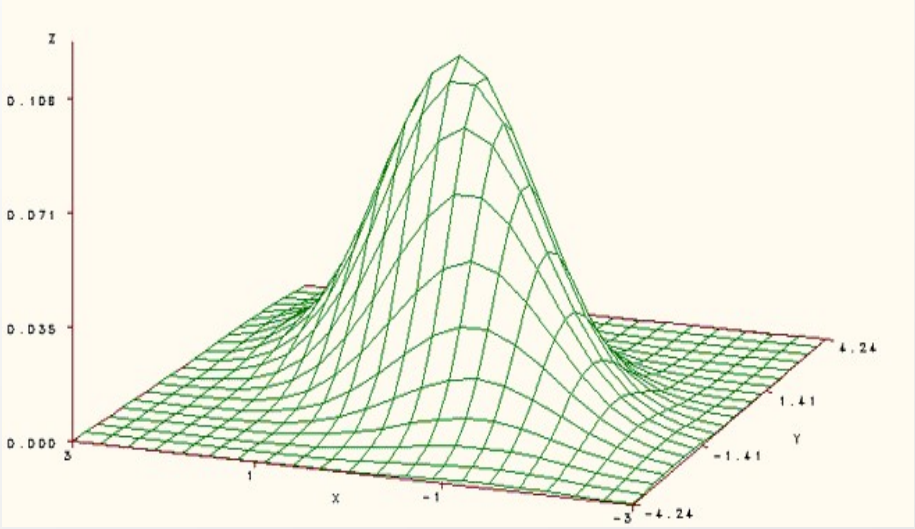
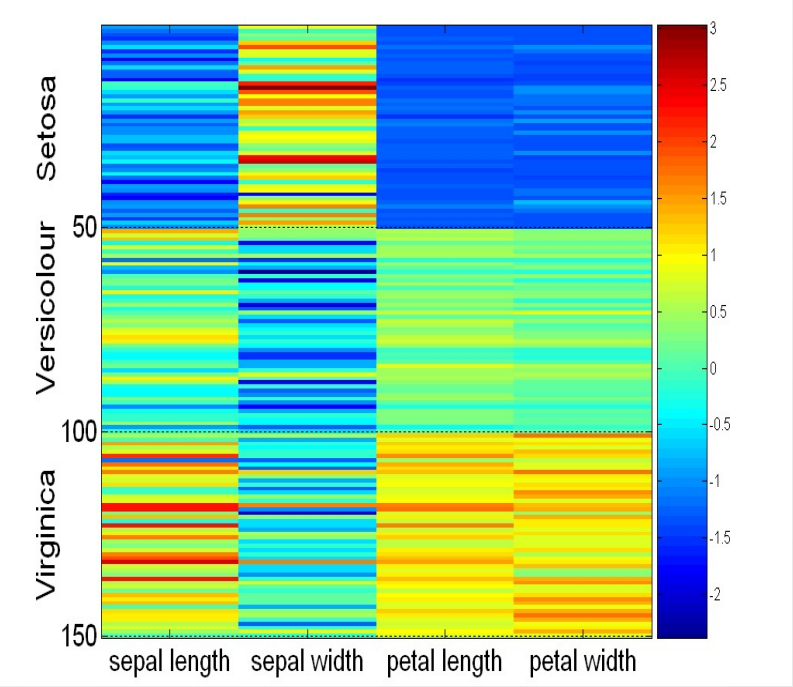


图 曲面图示例

3.3 技术

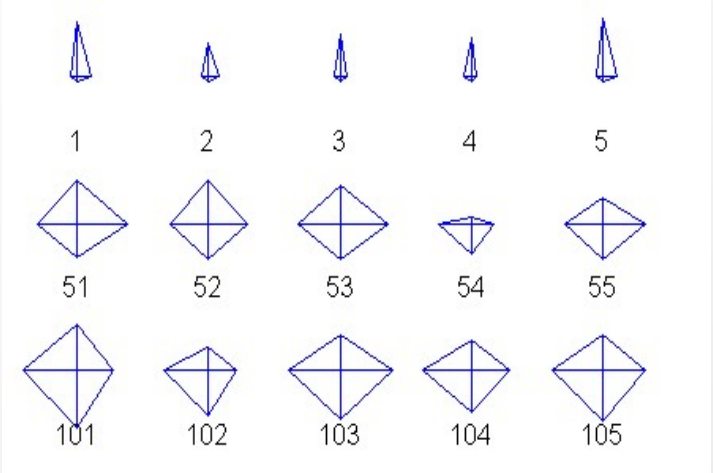
- 可视化高维数据
  - 数据矩阵
    - 根据类标将数据分类，不同类对应图像的不同区域
    - 用颜色和亮度衡量属性值的大小



鸢尾花数据矩阵图，其中列已标准化

3.3 技术

- 星形坐标和Chernoff脸
  - 每个数据对应一个图形
  - 每个属性对应一个坐标方向或一个脸部特征



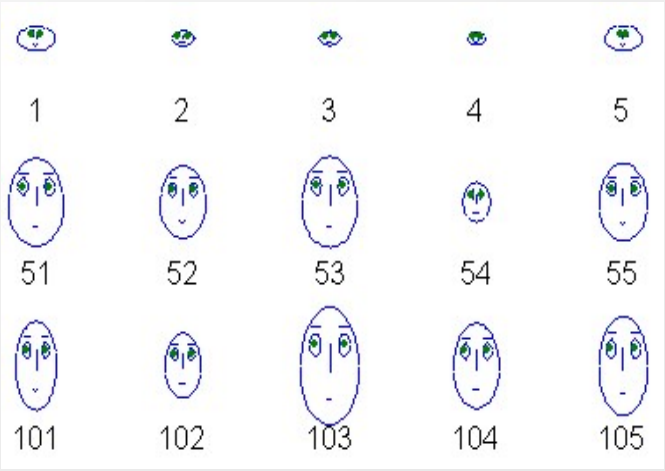


图 15种鸢尾花的星形坐标图形和Chernoff脸图形

第三节　数据可视化

- 定义和作用
- 基本概念
- 技术
- 注意事项

3.4 注意事项

- ACCENT原则
  - 理解（Apprehension）：最大化对变量之间的关系的理解
  - 清晰性（Clarity）：能识别图形中所有元素
  - 一致性（Consistency）：元素、符号形状和颜色与以前图形使用的一致
  - 有效性（Efficiency）：用尽可能简单的方法描绘复杂关系
  - 必要性（Necessity）:所有图形元素都是必要的
  - 真实性（Truthfulness）：图形元素可以准确定位和定标。

第四节　报告题目列表

- 报告题目7
- 报告题目8
- 报告题目9

4.1 报告题目7

7.XX大学历年来计算机专业研究生考试相关情况的数据分析研究

结合XX大学历年来计算机专业研究生考试分数线、报考情况、录取情况、考生分布、专业设置、招生人数等相关信息，利用数据挖掘技术对其进行数据分析，给出设计方案以及实现步骤。

4.2 报告题目8

8.可视化技术在数据分析系统中应用的综述

基于主流的可视化技术，并结合当前的数据分析系统，对可视化技术在数据分析系统中的应用进行分析与研究，并对未来的发展给出展望。

4.3 报告题目9

9.大数据云平台下数据湖和数据仓库的区别与联系研究

数据湖和数据仓库，是在今天大数据技术条件下构建分布式系统的两种数据架构设计取向，要看平衡的方向是更偏向灵活性还是成本、性能、安全、治理等企业级特性。但是数据湖和数据仓库的边界正在慢慢模糊，数据湖自身的治理能力、数据仓库延伸到外部存储的能力都在加强。结合未来的发展趋势，给出自己对两者的认识及判断。

张冬松
dszhang@nudt.edu.cn

谢谢！ Q&A

THANKS FOR YOUR ATTENTION