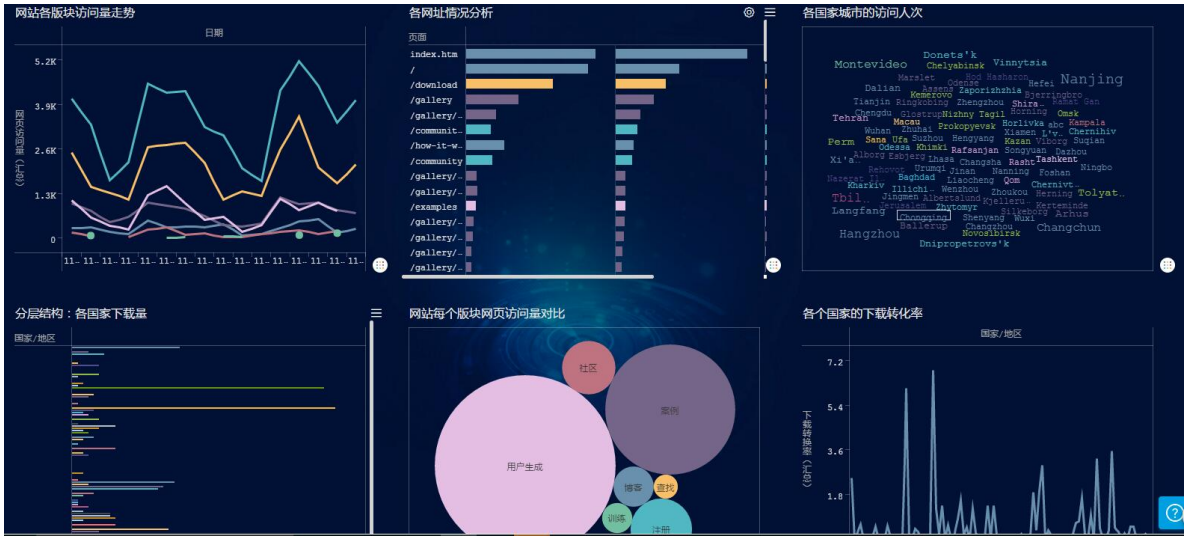
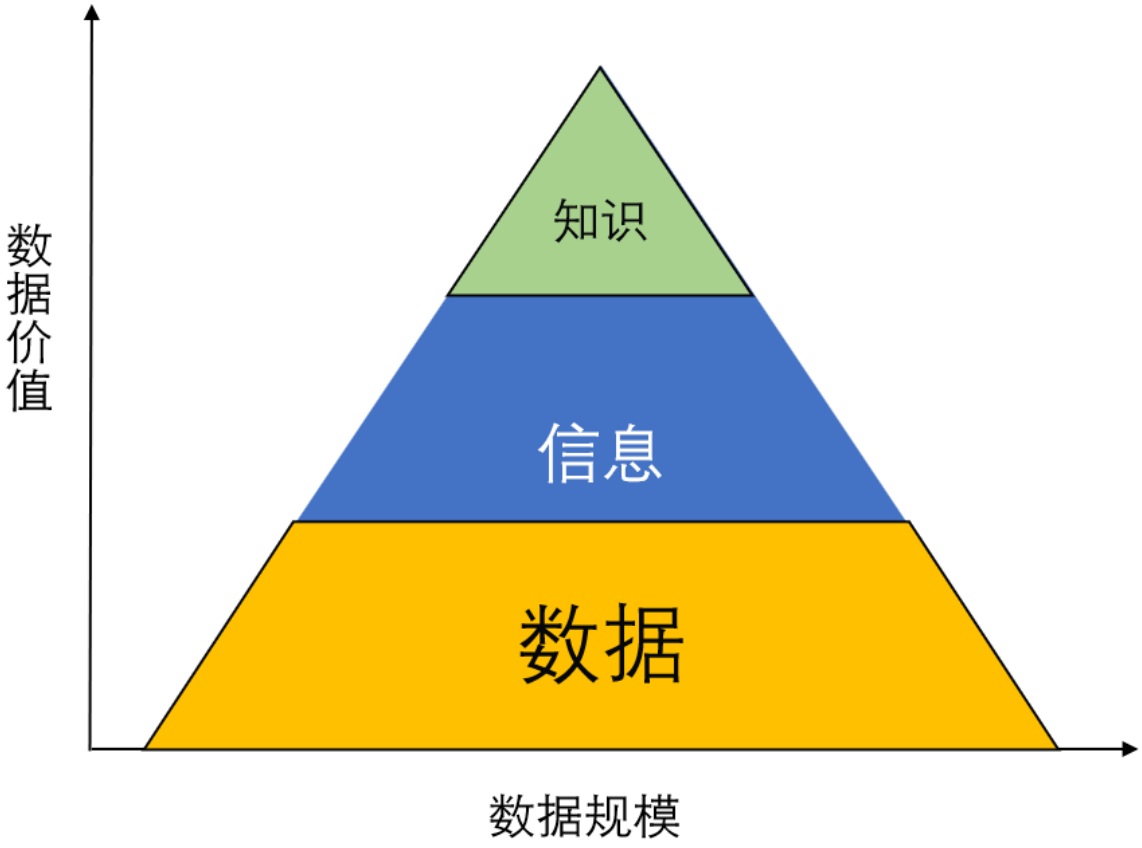
数据分析基础知识

一.什么是数据分析?



数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。



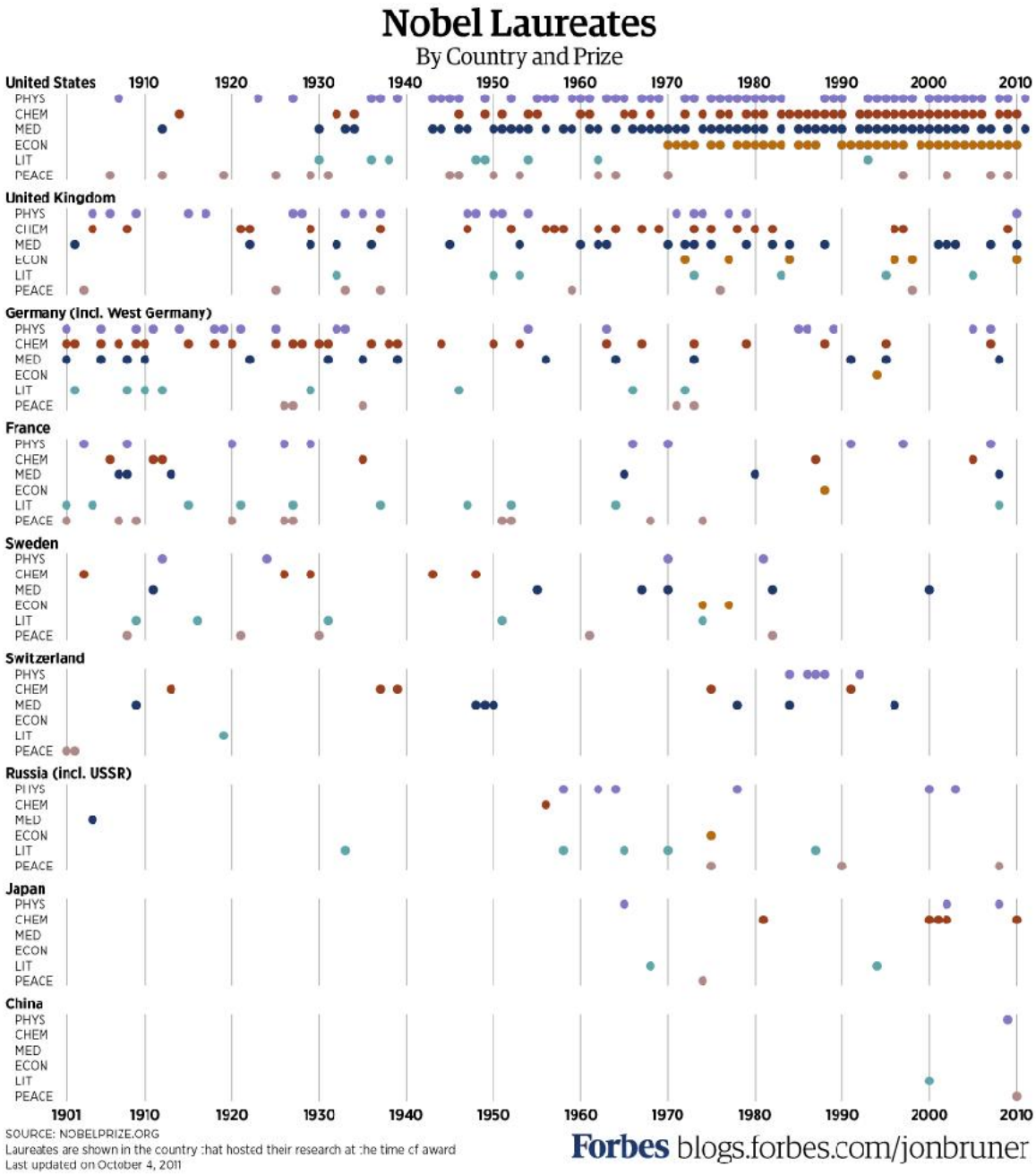
- 指导决策（以后吃火锅需要做好心理准备，并且注意不要吃太多）
- 知识:普遍的规律 (吃火锅会导致肚子疼)
- 信息:加入背景和联系的数据 (昨天吃了火锅，今天早上肚子疼)
- 数据:无背景的，无含义的 (有没有吃火锅，肚子有没有疼)

二.为什么数据分析?

数据分析的目的是把隐藏在一大批看来杂乱无章的数据中的信息集中和提炼出来，从而找出所研究对象的内在规律。并对决策者起到支持作用

2.1 一些例子

- 诺贝尔奖获得者国籍



上图以国家及奖项为横轴，时间为纵轴描绘了诺贝尔获奖者国籍的散点图。可以看到，1940年以前德国几乎是世界科技的中心，但是1940年之后，特别是1990年到2010年间，几乎所有诺贝尔奖得主都是美国人。
这就是一个描述性分析的例子，诺贝尔奖国籍的分布是一个比较明显的信息，但通过统计学的整理和描述，这些信息能够更加有效的表示进一步的结论。

- 啤酒和尿布的故事

沃尔玛的超市管理人员分析销售数据时发现了一个令人难于理解的现象：在某些特定的情况下，“啤酒”与“尿布”两件看上去毫无关系的商品会经常出现在同一个购物篮中。
实际上，在美国有婴儿的家庭中，一般是母亲在家中照看婴儿，年轻的父亲前去超市购买尿布。父亲在购买尿布的同时，往往会顺便为自己购买啤酒，这样就会出现啤酒与尿布这两件看上去不相干的商品经常会出现在同一个购物篮的现象。如果这个年轻的父亲在卖场只能买到两件商品之一，则他很有可能会放弃购物而到另一家商店， 直到可以一次同时买到啤酒与尿布为止。
沃尔玛发现了这一独特的现象，开始在卖场尝试将啤酒与尿布摆放在相同的区域，让年轻的父亲可以同时找到这两件商品，并很快地完成购物；而沃尔玛超市也可以让这些客户一次购买两件商品、而不是一件，从而获得了很好的商品销售收入。
这是一个探索性分析的例子，沃尔玛在长期的“购物篮分析”的过程中总结商品售卖规律，得到了大量的从销售数据表面无法得出的商品关联性结论。

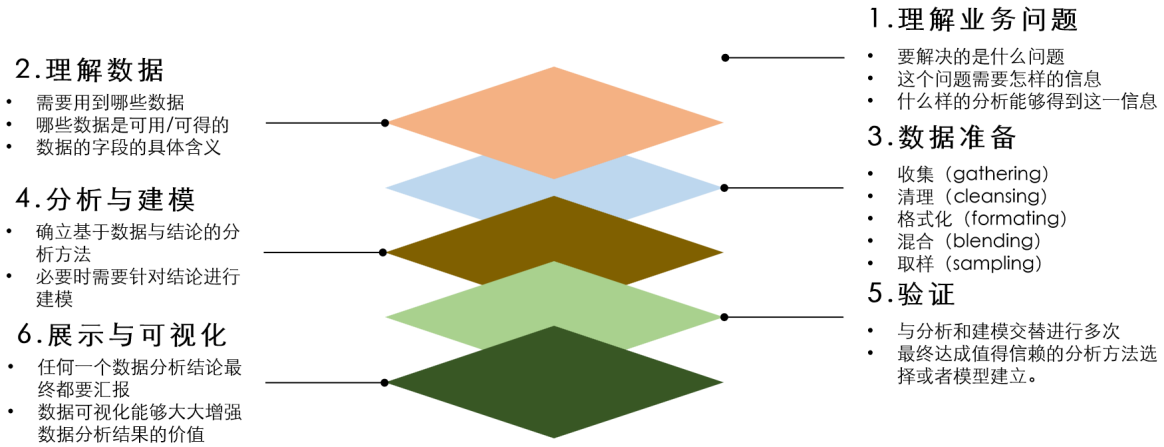
- 领导力现状的数据分析

我们已经通过模糊印象预期公司内存在新任经理较多，且集中在软件领域的情况；

通过对现有数据的分析，我们发现了确实软件互联网领域管理层的工作年限与年龄较小，且管理幅度较大；

这是一个验证性分析的例子，我们虽然通过直觉或者逻辑判断得到了某种结论，但是通过数据分析的方式我们才能更好更准确的验证这一结论是否合适和正确。

三.数据分析问题解决框架



四.数据分析的一些技巧

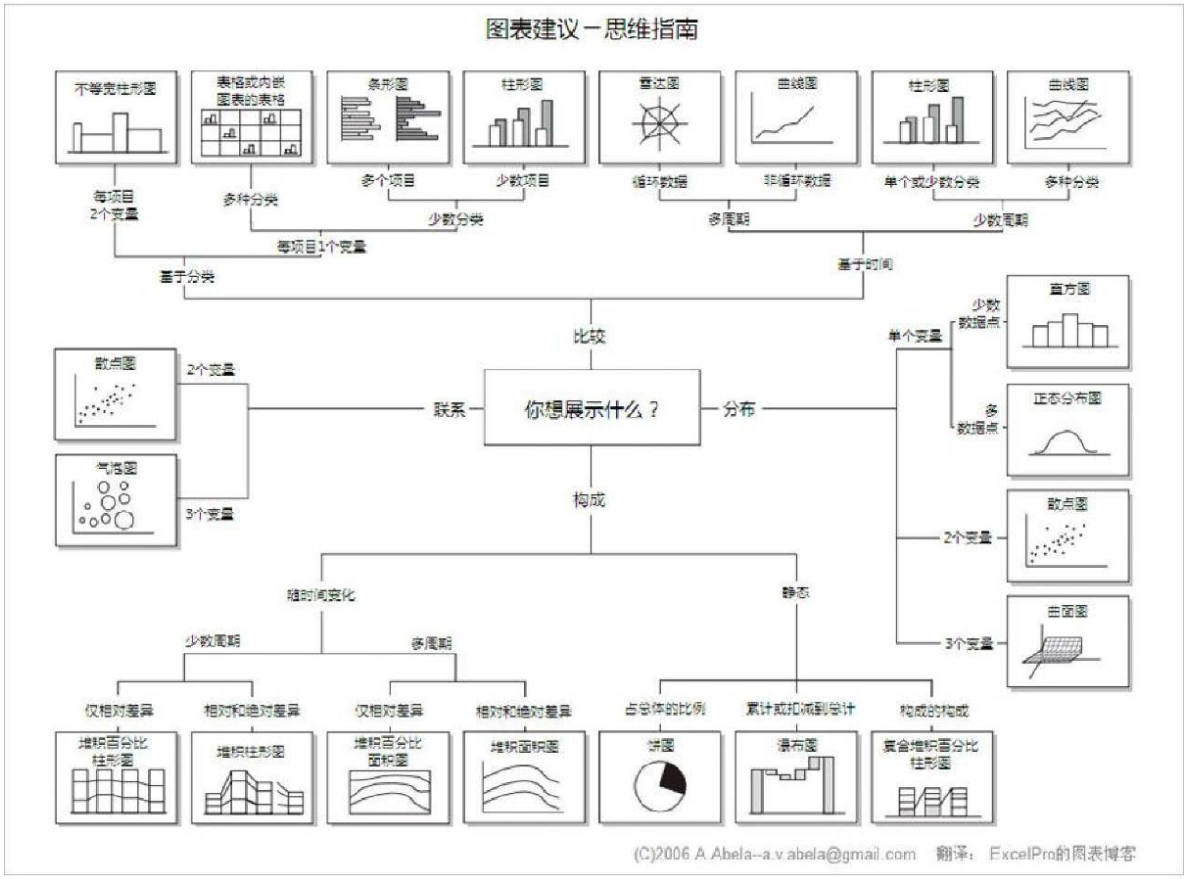
4.1 数据预处理Pandas

数据加载
数据清理
    1.处理缺失值
    2.处理重复数据
    3.数据类型转换
    4.处理异常值
数据分析与操作
    1.数据筛选
    2.分组聚合
    3.基本统计分析

4.2 数据可视化

图表是最常见的数据可视化方法，常用的图表有：

- 条形图
- 柱形图
- 直方图
- 折线图/曲线图
- 雷达图
- 饼图



通过数据分析，隐藏在数据内部的关系和规律就会逐渐浮现出来，那么通过什么方式展现出这些关系和规律，才能让别人一目了然呢？一般情况下，数据是通过表格和图形的方式来呈现的，即用图表说话。
多数情况下，人们更愿意接受图形这种数据展现方式，因为它能更加有效、直观地传递出分析师所要表达的观点。一般情况下，能用图说明问题的，就不用表格，能用表格说明问题的，就不用文字