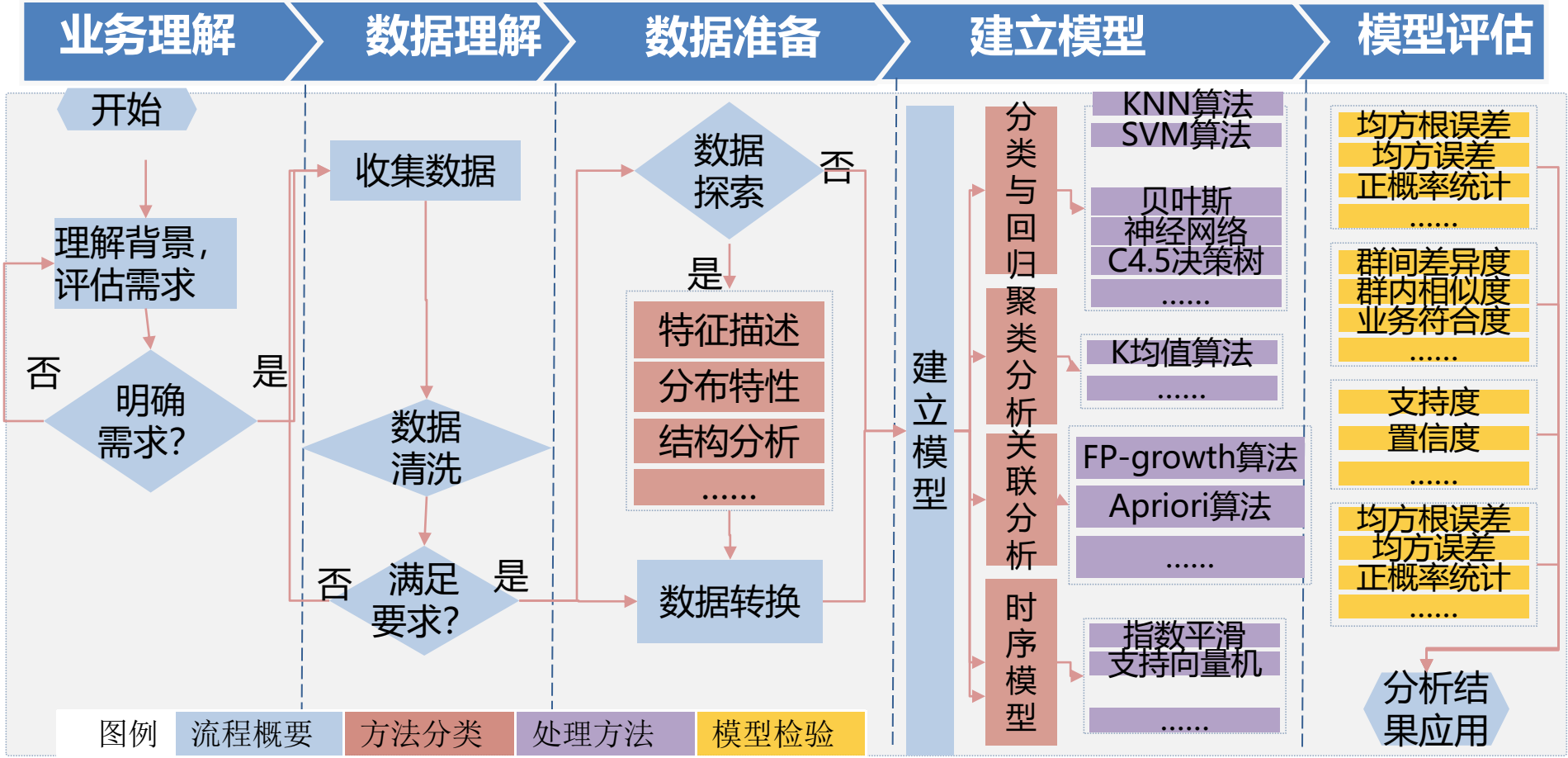
大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

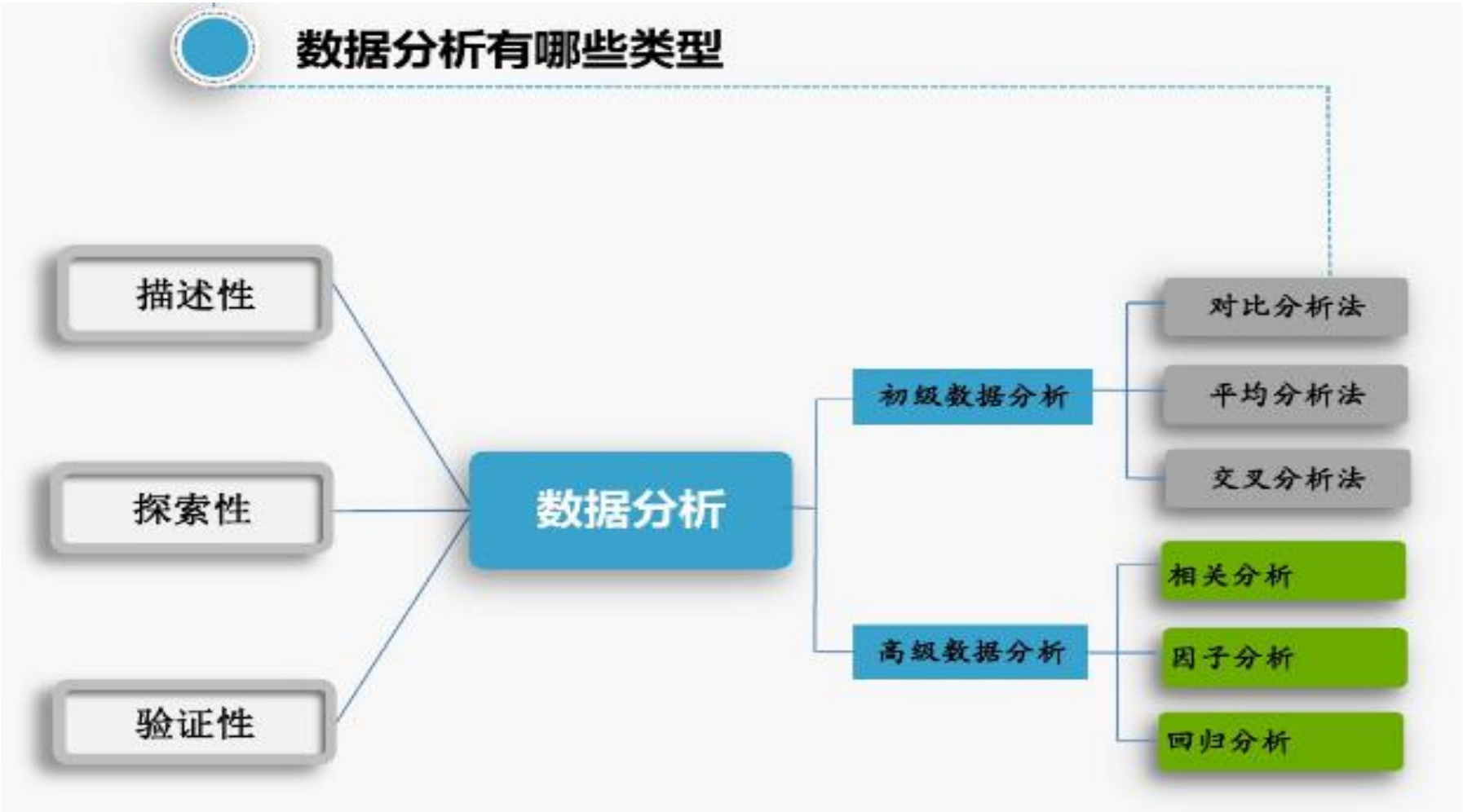
北京理工大学计算机学院　孙新

2019年1月

数据分析框架


业务理解
数据理解
数据准备
建立模型
模型评估
开始
理解背景，评估需求
明确需求?
否
是
收集数据
数据清洗
满足要求?
否
是
数据探索
否
是
特征描述
分布特性
结构分析
……
数据转换
建立模型
分类与回归
聚类分析
关联分析
时序模型
KNN算法
SVM算法
贝叶斯
神经网络
C4.5决策树
……
K均值算法
……
FP-growth算法
Apriori算法
……
指数平滑
支持向量机
……
均方根误差
均方误差
正概率统计
……
群间差异度
群内相似度
业务符合度
……
支持度
置信度
……
均方根误差
均方误差
正概率统计
……
分析结果应用
图例
流程概要
方法分类
处理方法
模型检验

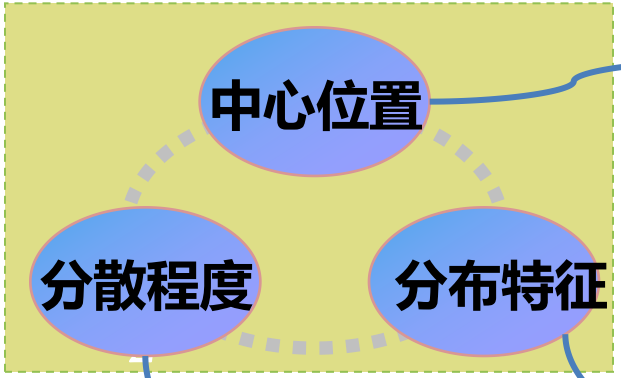
数据分析有哪些类型


描述性
探索性
验证性
数据分析
初级数据分析
高级数据分析
对比分析法
平均分析法
交叉分析法
相关分析
因子分析
回归分析

数据分析的方法

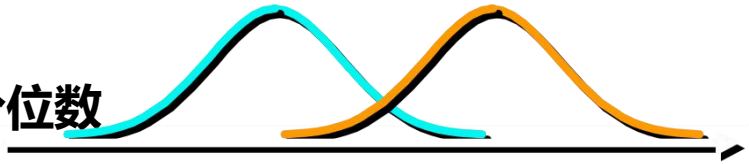
- **数据分析主要包括分析数据和建模，**一般涉及：
  - 统计数据分析方法
  - 基于机器学习的数据分析方法
- 统计数据分析的方法可以帮助我们了解数据，为了形象直观地了解关注群体的年龄、性别比例、职业等，需要对数据进行**数据描述性分析。**

数据描述性分析



中心位置

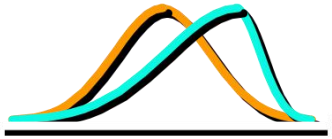
- 众数
- 中位数/四分位数
- 均值



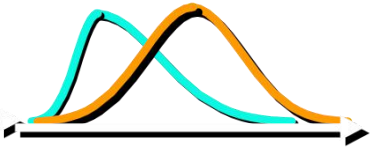
分布特征

- 偏度

数据分布偏斜程度的测度



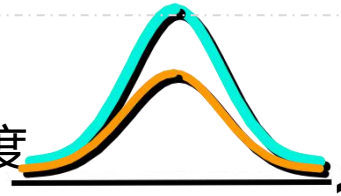
左偏分布



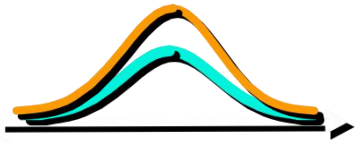
右偏分布

- 峰度

数据分布扁平程度的测度



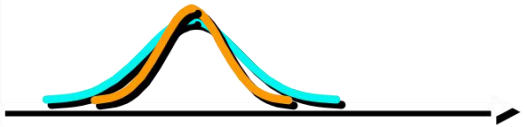
尖峰分布



扁平分布

分散程度

- 方差和标准差
- 极差、四分位差



社交网络数据分析示例

- 如果希望对大数据进行更深层次地探索，总结出规律和模型，则需要更加智能的**基于机器学习的数据分析方法。**

- 例如：如何利用社交网络上的数据，分析大众的情绪或者心态呢？
  - 简单地讲，这是一个自动分类的问题，即把人的情绪分为若干类，然后把网络数据根据内容确定为其中的一类（或者几类）。
- 实现的方法大致有两种，
  - 第一种是**有监督的机器学习**，
  - 第二种是**无监督的机器学习**，

社交网络数据分析示例

- **有监督的机器学习**，大致步骤如下：
- 首先，从网络数据中选取一些**样本**（比如帖子，也称为**训练样本**），人工地对每个帖子打上一个情绪的标签（高兴、愤怒、焦虑等等），这就将它们各自分到所属的类别。
- 其次，**根据每一类情绪对应的帖子，找到相应的特征**，这些特征可以是简单的关键词、关键词的组合、表情符号，甚至是一些标点符号（比如问号和感叹号），也可以是表述时用的句式、语法结构等等，每一种情绪对应的特征是不同的。
- 最后，把大量收集来的网络数据拿来（**测试样本**），从中抽取特征，和每一类情绪的特征做比对，就能大致确定大众在网络数据中所反映的情绪。

社交网络数据分析示例

- 这种方法主要的缺点是手工标注出每一个样本所对应的情绪，工作量很大，为了克服这个问题，可以采用一种**无监督的机器学习方法**，
- 具体方法：
  - 也就是说，一开始随机地给样本设定一种情绪，当然这种情绪的初始设定大部分是不正确的，不过没有关系。
  - 接下来，采用自适应的机器学习方法，通过多次迭代来修改最初的错误，直到计算机找不到更多的错误为止。
  - 这样，前面有监督的机器学习的第一步就自动完成了，以后的步骤则相同。
  - 其次，根据每一类情绪对应的帖子，找到相应的特征。
  - 最后，把大量收集来的网络数据拿来，和每一类情绪的特征做比对，确定所反映的情绪。
- 这种方法的好处是减少人工，缺点是计算量非常大，而且有时机器学习的算法找不出错误，不等于错误不存在。因此，两种方法各有千秋

社交网络数据分析示例

- 通过大数据分析大众情绪的好处是，大家在社交网络或者网络媒体上发言时，通常不会刻意隐瞒自己的观点，因此分析的准确性要比问卷调查好很多。

- 总之，基于机器学习的数据分析方法，可以从海量数据中找到人们未知的、可能有用的、隐藏的规则，
  - 具体采用的方法包括：分类和预测，关联分析、聚类分析等，可以发现一些无法通过观察图表得出的深层次原因。

谢 谢

Thank you for your attention!