天善智能
TIANSHAN SOFT

数据分析与建模的基础知识

讲师：安迪生

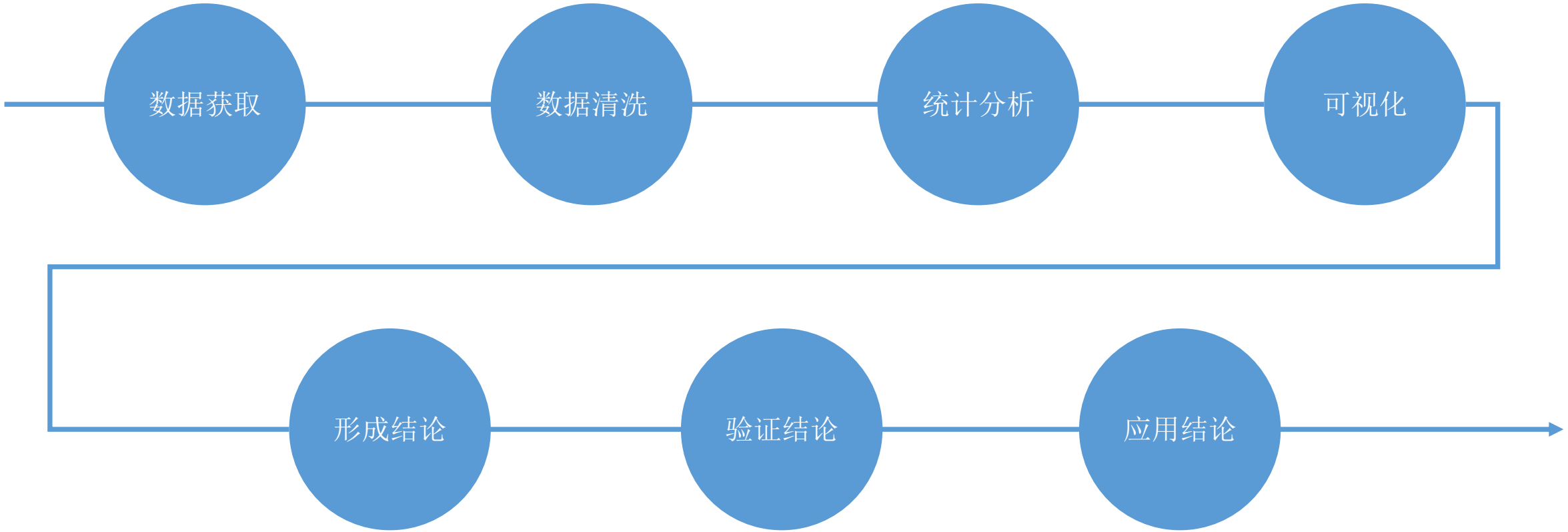
课程简介

- 随着大数据、人工智能算法和机器学习算法的兴起，越来越多的金融风控人员将量化模型引入到风控业务当中去。这意味着数据分析技术在金融风控工作中起到一个非常重要的角色。本节课将给大家简单介绍一下数据分析的基础知识。

目录

数据分析的基本概念

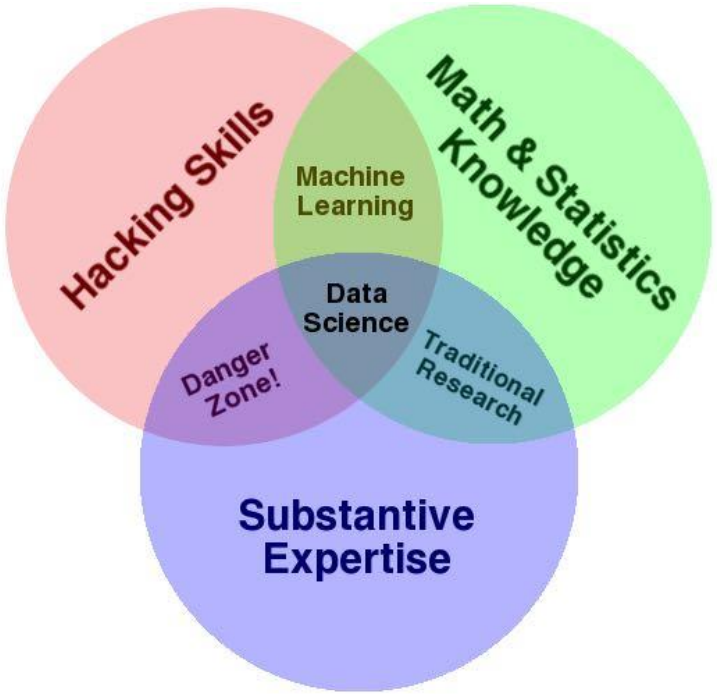
- 数据分析是一项从自然环境、社会环境、网络环境中提取数据，实施分析，得出结论并验证的工作。



数据分析的基本概念

- 不是为了分析而做分析

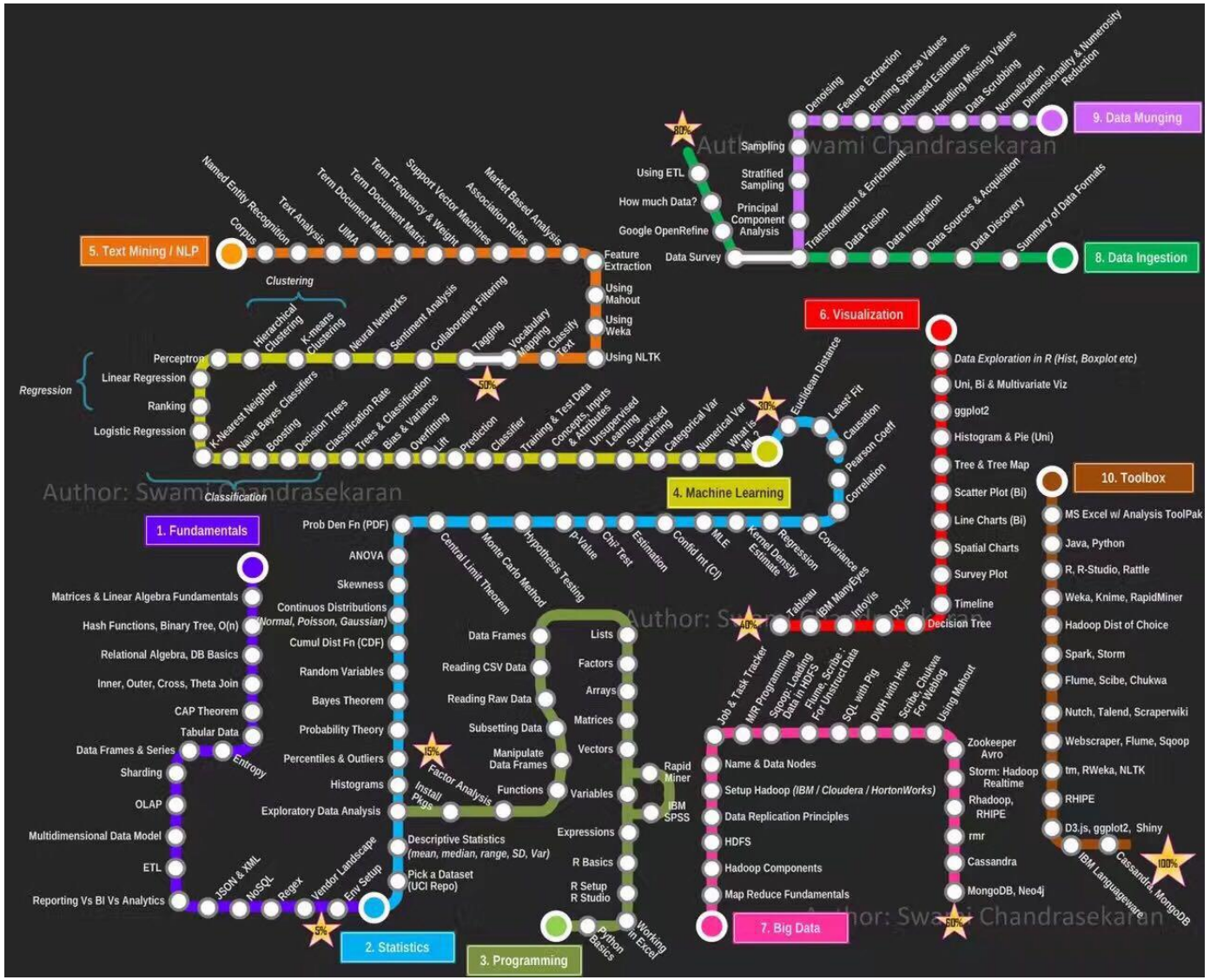
针对特定的问题，用适当的学科知识从数据中提炼信息，形成结论



- 数理知识基础
- 数据获取、加工能力
- 行业知识

数据分析的基本概念

- 数据分析从业人员的修炼途径



数据分析的基本概念

- 数据获取的途径

公共数据库

- 多大免费
  权威
- 粒度粗
  更新慢
  覆盖度低

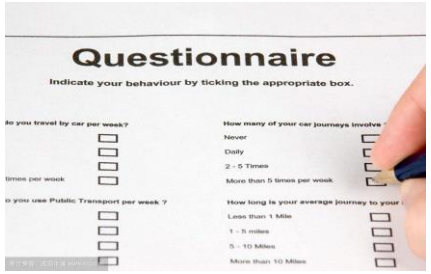
私有数据库

- 更新及时
  粒度细
- 价格高
  访问权限受限

网络爬虫

- 免费
  数据来源广
- 技术要求高
  数据脏
  数据可靠性低

问卷调查



- 有针对性
  可靠性较高
- 搜集量较少
  使用范围受限

设备采集

- 准确度高
- 成本高
  使用范围受限

数据分析的基本概念

- 数据清洗

做数据清洗的原因：脏数据；不满足分析要求

数据质量要求和清洗办法

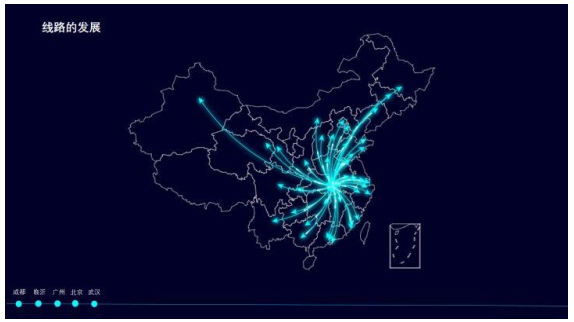
完整性	唯一性	权威性	一致性	合法性
•通过其他信息补全 •通过前后数据补全 • 剔除	• 按主键去重 • 合并同一主键下的数据	• 最权威的那个渠道的数据	• 建立数据体系，包含但不限于指标体系、维度、单位、频度、数据	• 设定强制合法规则 • 字段内容合法规则 • 字段格式合法规则 • 离群值人工特殊处理

数据可视化

- 可视化的意义

可视化数据包含的信息不会超过数据本身，但是能让使用者更加容易发掘数据的信息。在数据可视化下，信息的获取、加工、输出会变得更加简洁。

- 数据可视化的场景

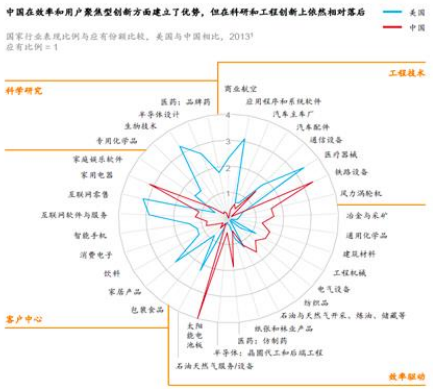


交通数据（旅游、物流、航空、海运）

地理信息（区域对比）

数量对比（占比、计数等）

时间序列（股价、人口数等）

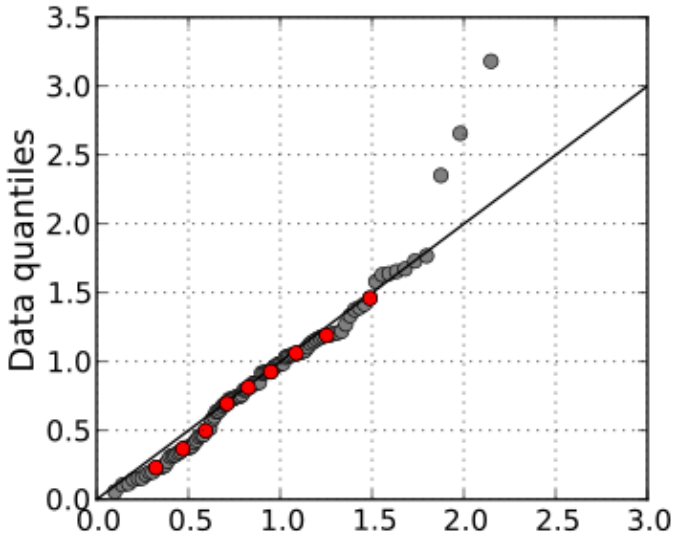


多维度展示（客群画像等）

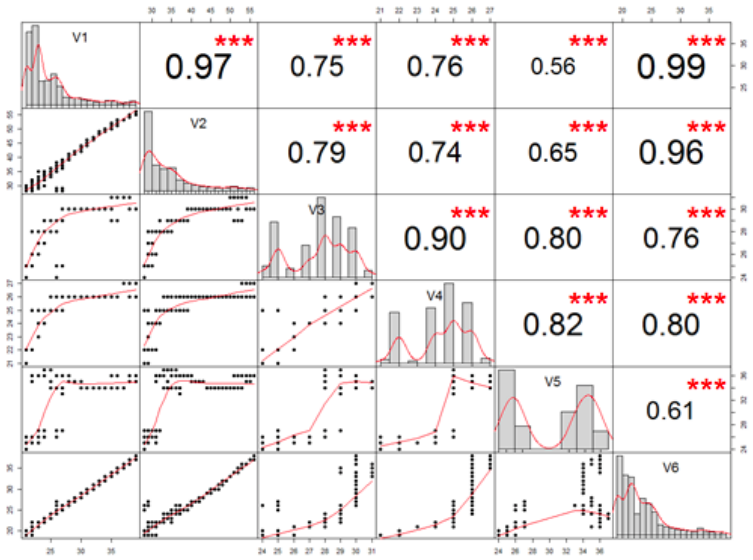
数据可视化

- 在统计分析里的可视化案例

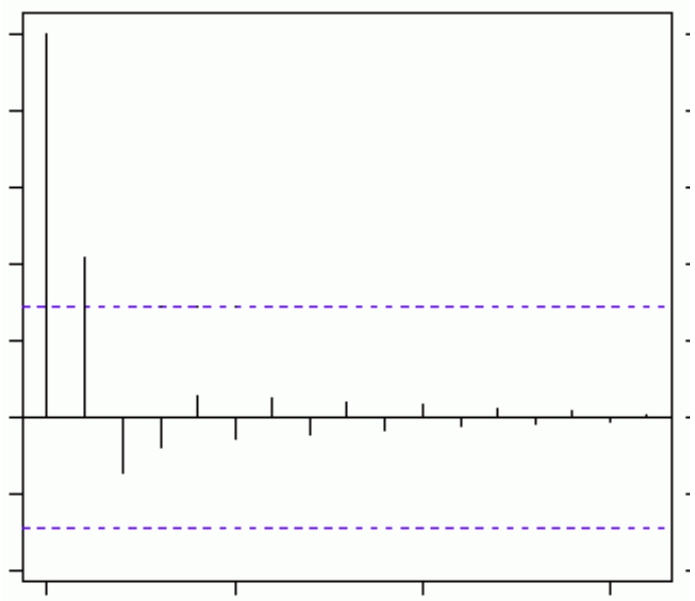
正态性检验：QQ-plot



相关性检验：scatter matrix



时间序列：ACF



数据可视化

- 数据可视化常用的工具

专业工具

- Tableau：优秀的数据可视化展示工具，数据图表制作能力强，操作简单，上手快不需要写代码，数据的导入和加载都是向导式，内置美观的可视化图表，不用考虑配色，表格处理好格式即可。
- DataV：阿里云出品，付费（5元/月），拥有极其丰富的图表选择，编程简易，支持丰富的数据接入方式（其中有API接口），拥有动画效果

通用工具

- Excel：entry level的“菜鸟”到骨灰级专家都能玩转
- R和Python：丰富的内嵌图表和海量的三方库，灵活性高

数据分析的常用模型

- 描述性统计量
- **常用的单变量统计**：对于X的观测值$\{X_1, \dots, X_n\}$
- 均值

$$\mu = \frac{1}{n}\sum X_i$$

- 方差/标准差

$$var = \frac{1}{n-1}\sum (X_i - \mu)^2 , \sigma = var^{1/2}$$

- 分位点、中位数

$$p_i = \tilde{X}_{\lfloor n*i/100 \rfloor}$$, 其中$\tilde{X}$是X升序排列

- **常用的多变量统计**：对于X和Y的观测值$\{X_1, \dots, X_n\}$， $\{Y_1, \dots, Y_n\}$
- 协方差，相关系数

$$Cov(X,Y) = \frac{1}{n-1}\sum (X_i - \mu_X)(Y_i - \mu_Y) , corr(X,Y) = \frac{Cov(X,y)}{\sqrt{var_X \times var_y}}$$

数据分析的常用模型

- 有监督模型

在分析过程中，存在一个或多个“目标”变量，使得我们需要去研究其他变量（称为独立变量，或者特征）如何影响这（些）个目标变量。

例如下面的2个案例

1. 研究新生入学成绩、性别、第一学期平均学习时长是如何影响期末考试成绩

2. 研究竞选中，选民的学历、收入、民族、职业等因素如何影响候选人竞选成功

单一目标变量占了绝大多数的场景。

- 回归和分类

当目标变量是连续型数值变量时，是回归模型，如案例1

当目标变量是取值为2或更多的类别型变量时，是分类模型，如案例2

数据分析的常用模型

- 有监督模型

回归：线性回归，部分广义线性回归，神经网络/深度学习模型等

分类：SVM，分类树，朴素贝叶斯，逻辑回归，kNN，神经网络/深度学习模型

排序：page rank

- 有监督模型的损失函数

$$loss\ function = error\ cost + complexity\ cost$$

说明：

✓ 回归和分类，并没有本质的区别

✓ 部分模型同时适用于二者，如ANN，DL，CART等

✓ 除了上述的单一模型外，还有各种集成模型。例如基于bagging 的随机森林，基于boosting的AdaBoost，GBDT，xgboost。又： GBDT，xgboost仅仅是集成框架，不表示具体的回归或者分类模型

数据分析的常用模型

- 无监督模型

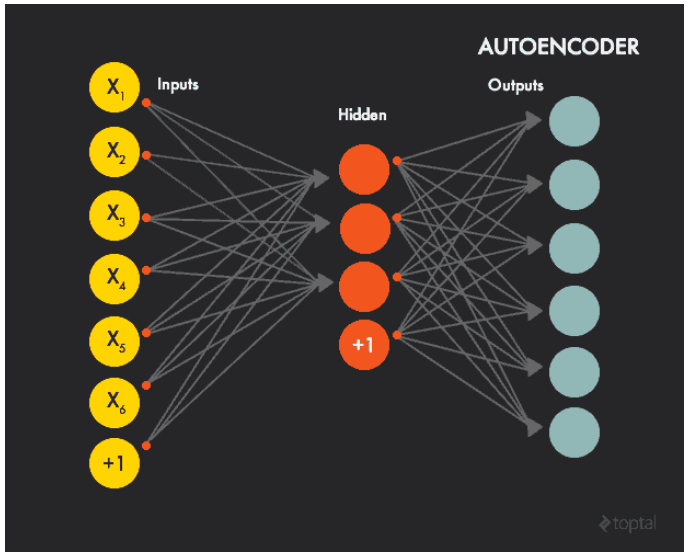
对特征：主成分分析、因子分析等

对样本：关联分析、部分聚类分析、复杂网络、生成模型（如自动编码机、GAN等）

说明

✓ 除了有/无监督外，还有半监督模型

✓ 增强学习不认为是有/无监督模型



无监督：自动编码器

数据分析的常用工具

- 数据分析的利器

R

- 面向统计分析的编程语言，丰富的作图功能，开源
- CRAN
- Rstudio
- install.packages(), library()

Python

- 自有软件，胶水语言，免费的MATLAB
- pip install yourPackage
- import yourPackage as pkg
- From yourPackage import yourFunction

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI

天善智能
TIANSHAN SOFT