面向数据的思维模式与R语言的数据项目开发

张丹
青萌数海CTO

背景

- 很多公司已经完成了数据的原始积累，如何让沉睡的数据发挥价值，是急需要功课的难关！
- 数据项目和软件项目、互联网项目都有非常大的不同，不确定性、跨学科知识点、工程落地，都是影响数据项目成功与失败的重要因素。
- 掌握数据思维，科学的方法论，专业的团队，便利的工具，才能让数据项目走向成功。

目录

面向数据的思维模式

- 数据很重要，要想把业务做好，需要有更高质量和更好来源的数据。
- 目标：不是如何做出系统，而是如何提升业务价值？

要怎么去写代码？？

角色对比

开发工程师，5年工作经验，熟练掌握一种编程语言。

能力要求：

- 理解需求
- 快速写代码完成开发任务
- 没有bug
- 性能很好
- 设计系统架构（非功能需求）
- 解决技术难题

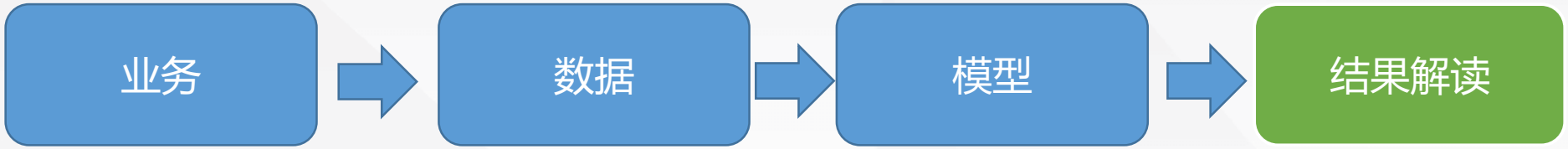
数据分析师，5年工作经验，熟练掌握一种数据分析编程语言。

能力要求：

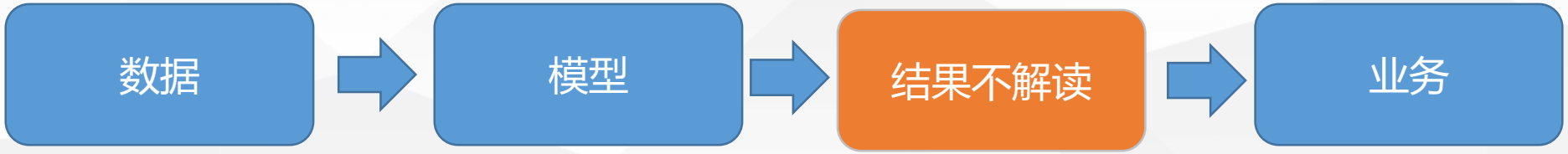
- 提出需求
- 形成数据分析方法体系
- 洞察规律的本质
- 用数据论证规律背后的逻辑
- 设计数据产品

建模思维对比

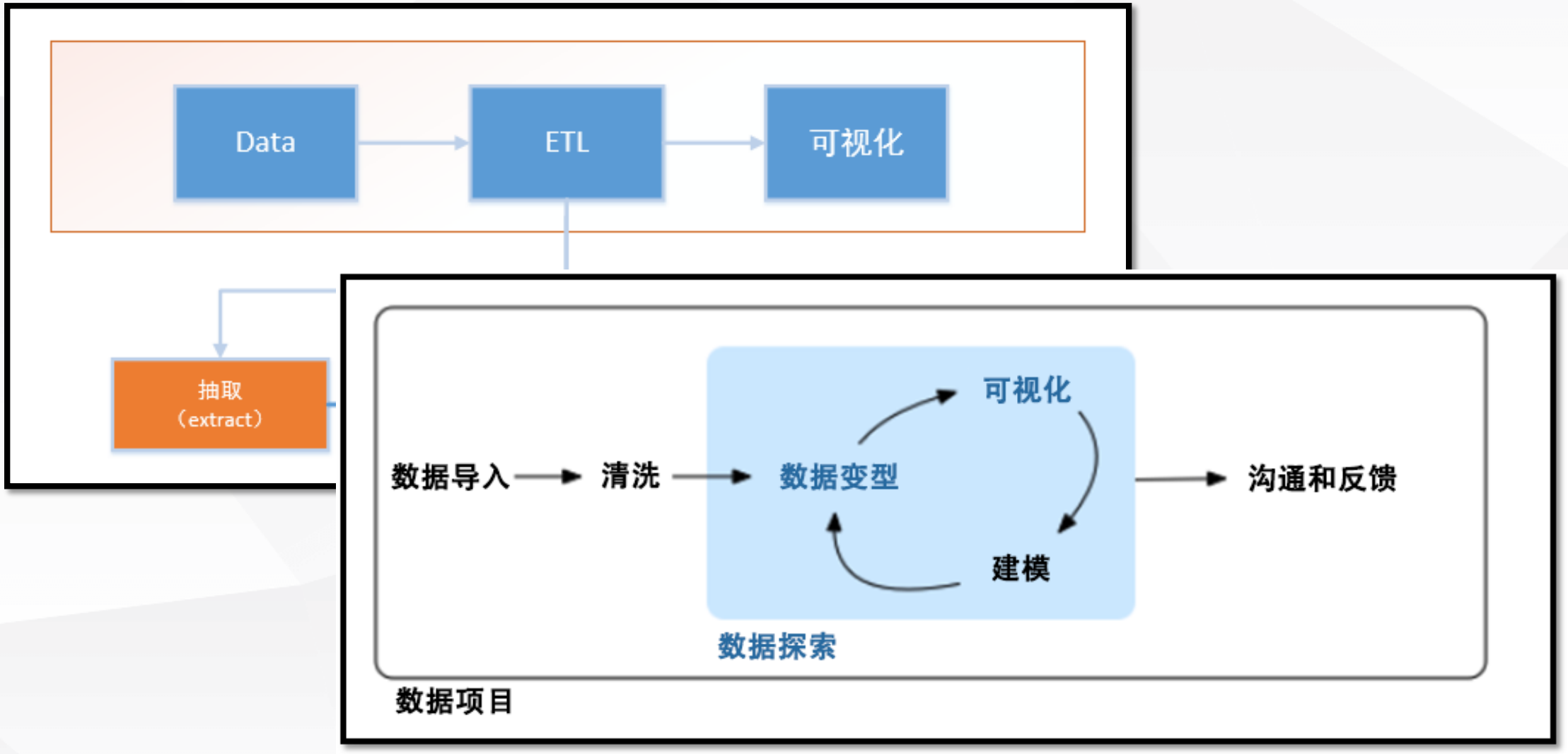
数据科学建模思维



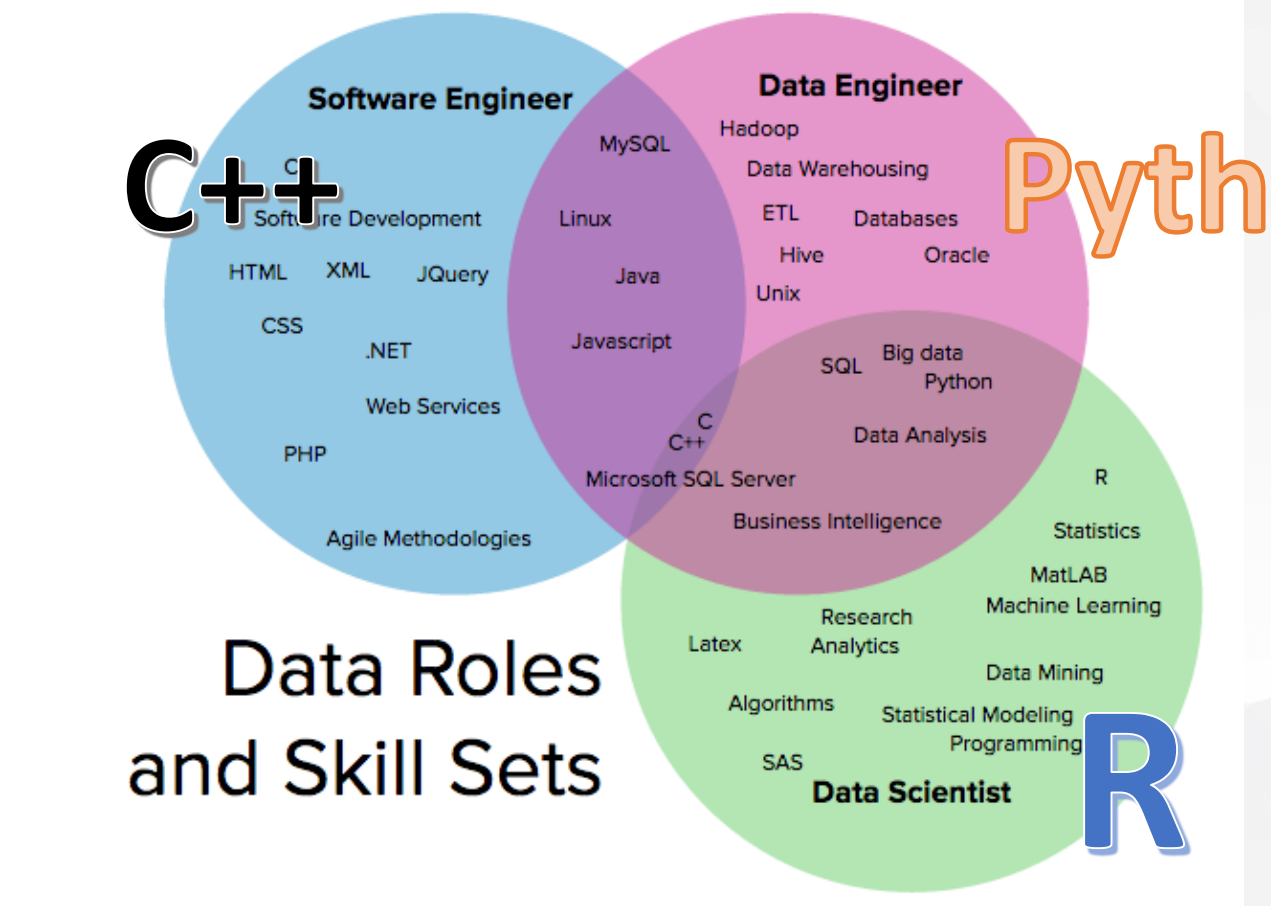
IT建模思维



BI项目与数据项目



编程语言对比


Software Engineer
C++
C
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
Data Engineer
Hadoop
MySQL
Data Warehousing
Linux
ETL
Databases
Hive
Oracle
Java
Unix
Javascript
SQL
Big data
Python
C
C++
Data Analysis
Microsoft SQL Server
Business Intelligence
Python
R
Statistics
MatLAB
Machine Learning
Research
Latex
Analytics
Data Mining
Algorithms
Statistical Modeling
Programming
SAS
Data Scientist
R
Data Roles
and Skill Sets

复合型知识体系

IT技术

编程能力，操作系统，网络，大数据技术，挖掘算法，爬虫，数据库，内存控制，高性能计算，软件使用等

基础学科

初等数学，高等数学，线性代数，离散数学，概率论，统计学，计量经济学、投资学、金融学

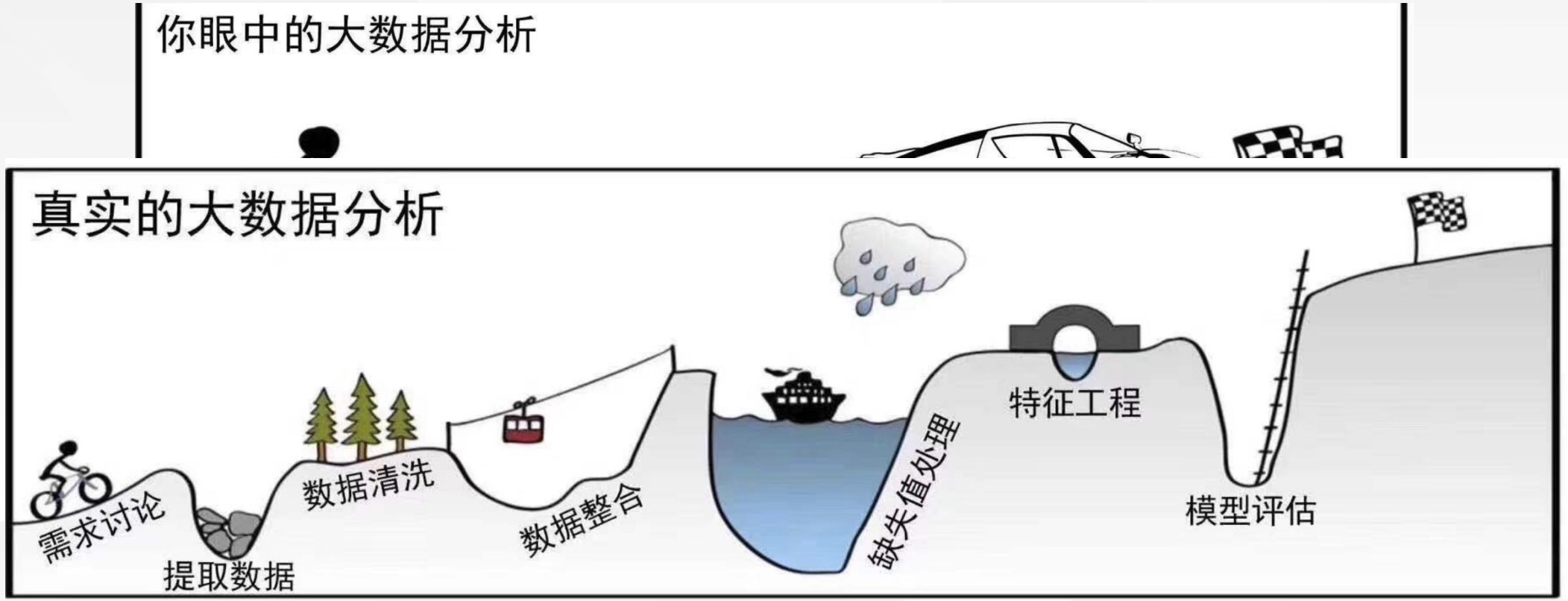
业务理解

金融机构：银行，证券，保险，信托、基金等
金融市场：股票、债券，期货，基金，外汇，P2P，理财、利率、汇率等
互联网，广告，生物，医药，进出口，制造业等

目录

如何开展一个数据项目

项目建设步骤：**需求讨论、数据提取、数据整合、数据清洗、特征工程、模型搭建和模型评估**



需求讨论

数据项目的开展过程中同样需要数据与业务的不断讨论：

1. 业务出发，确认目标
2. 认识数据，观察数据
3. 统计维度分析，发现数据异常值
4. 与业务沟通，确认数据逻辑
5. 无监督学习，分组
6. 与业务沟通，打标签
7. 有监督学习，模型训练
8. 与业务沟通，进行业务解读
9. 形成数据产品

切记不能各说各的。

数据提取

业务数据的来源和存储格式各异，且通常数量庞大，需要快速精确地从中提取建模分析所需的数据。

数据整合

把不同数据源和不同内容的数据，根据彼此间的关系整合到一个完整的数据集。

宽表

数据清洗

对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

01 补全缺失数据

03 去除、修改逻辑错误的数据

商品规格
花王 | Merries系列 | M | 48片/包 | 1.8kg
花王moony | Merries | L | 36片/包 | 1.9kg
Moony | Merries纸尿裤系列 | L | 48片/包 | 1.85kg
花王merries | 纸尿裤系列 | L | 36片/包、48片/包、2包/箱 | 1.9kg、2.1kg

语义分割

part1	part2	part3	part4	part5
花王	Merries系列	M	48 片/包	1.8 kg
花王moony	Merries	L	36 片/包	1.9 kg
Moony	Merries纸尿裤系列	L	48 片/包	1.85 kg
花王merries	纸尿裤系列	L	36 片/包 48 片/包 2 包/箱	1.9 kg 2.1 kg

自然语言处理

品牌	系列	型号	规格	重量
花王	Merries	M	48	1.8 kg
花王	Merries	L	36	1.9 kg
花王	Merries	L	48	1.85 kg
花王	Merries	L	36	1.9 kg
花王	Merries	L	48	2.1 kg

特征工程

特征工程是利用数据领域的相关知识来创建能够使机器学习算法达到最佳性能的特征的过程。

数据标准化

- **单位不同或量级不同的数据，。**

数据离散化

- **连续数据的范围很广，等级划分**

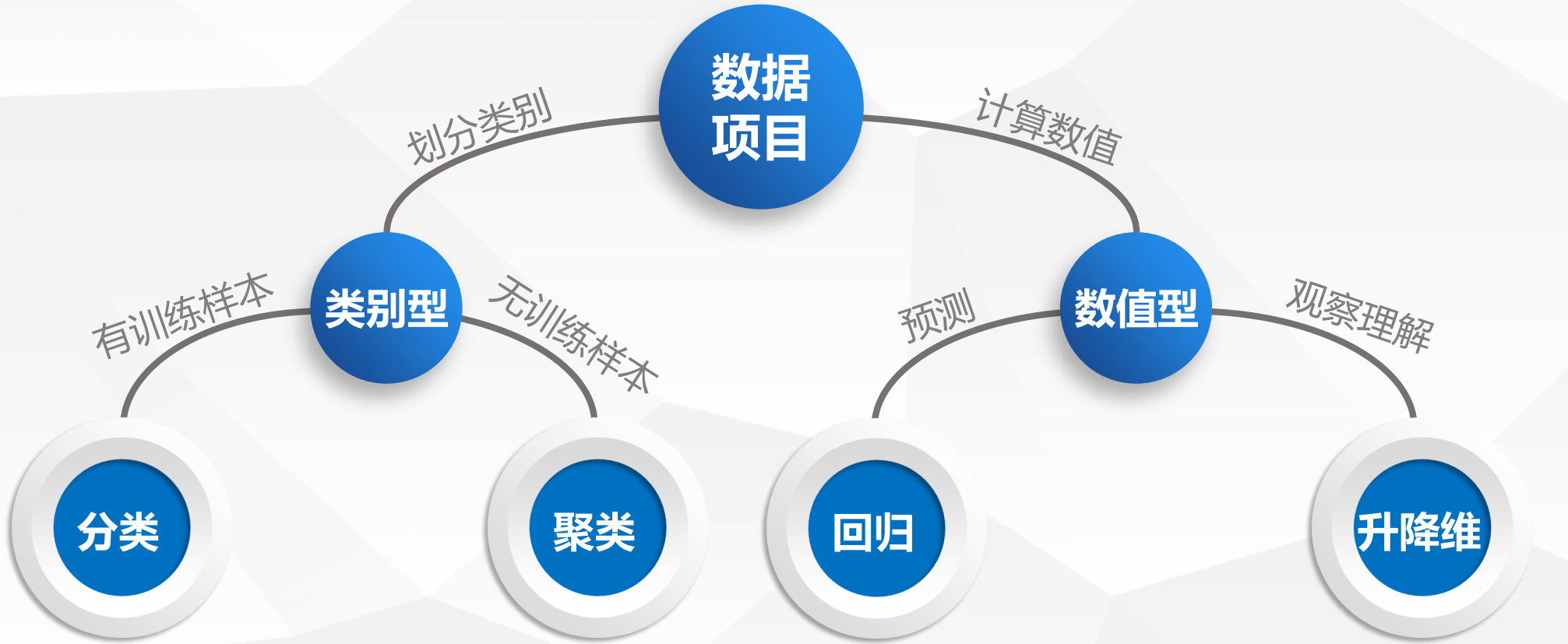
统计变换

- **数据分布倾斜有很多负面影响，**

分类特征编码

- **编码处理**
- **特征矩阵**

模型搭建


数据项目
划分类别
计算数值
类别型
数值型
有训练样本
无训练样本
预测
观察理解
分类
聚类
回归
升降维
决策树
神经网络
SVM
逻辑回归
K-means
层次聚类
DBSCAN
线性回归
广义回归
自回归
主成分分析
相关特征构建

模型评估

基于项目需求和数据情况，选择合适的模型进行训练，以及与之匹配的模型评估准则。

分类

✓ **混淆矩阵**

真实情况	预测结果	
	正例	反例
正例	*TP* (真正例)	*FN* (假反例)
反例	*FP* (假正例)	*TN* (真反例)

聚类

✓ **有基准数据：**

与基准数据的符合程度

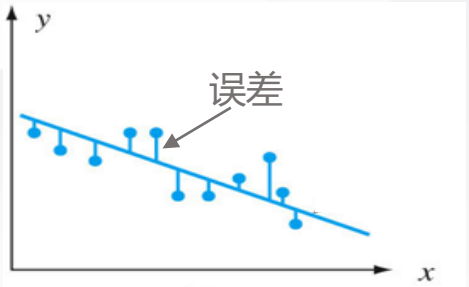
✓ **无基准数据：**

相似程度，差异程度。

回归

✓ **预测值与真实值的误差：**

均方根误差、平均绝对误差、R-Squared



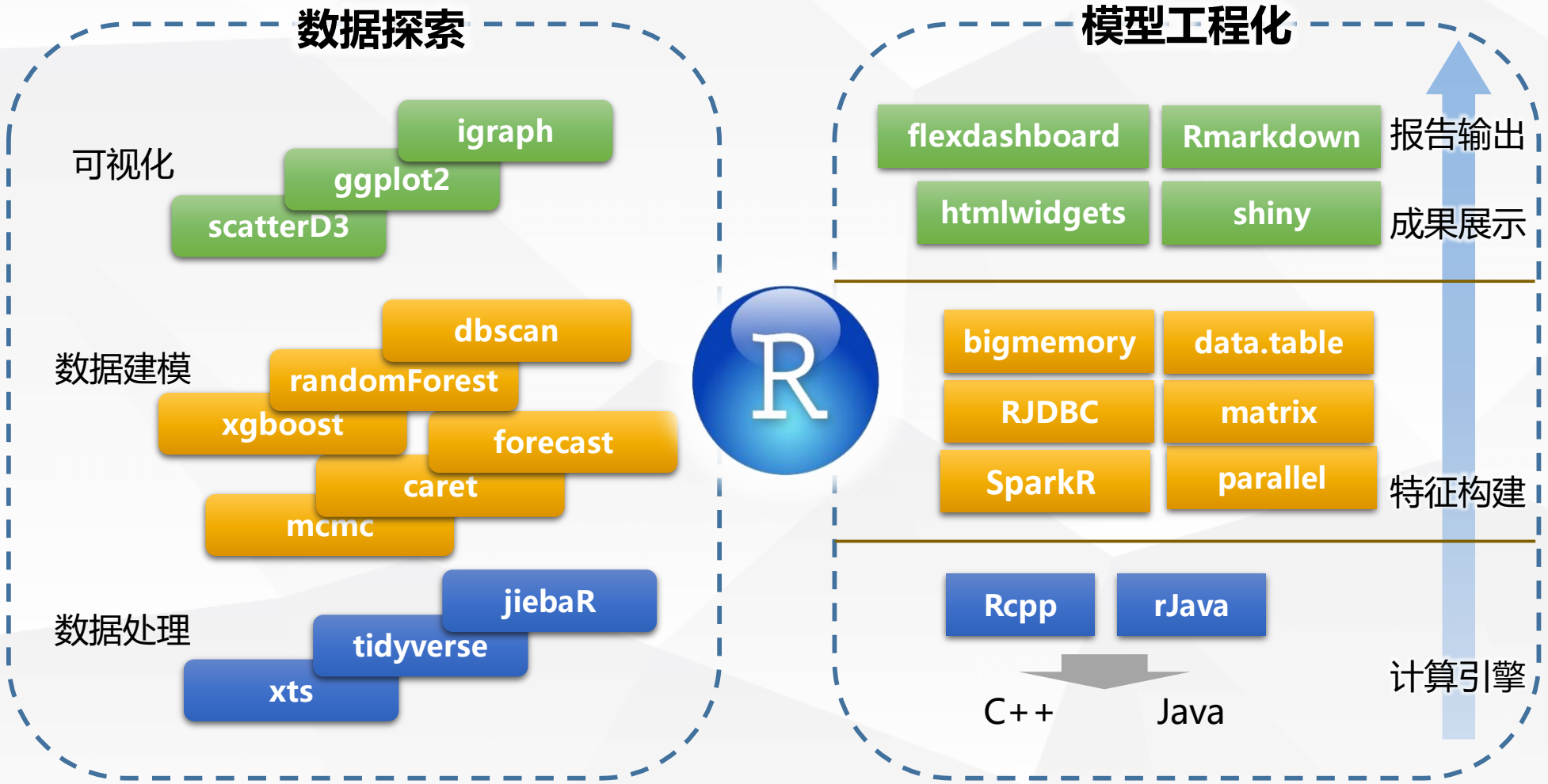
升降维

✓ **解释能力：**

方差贡献率

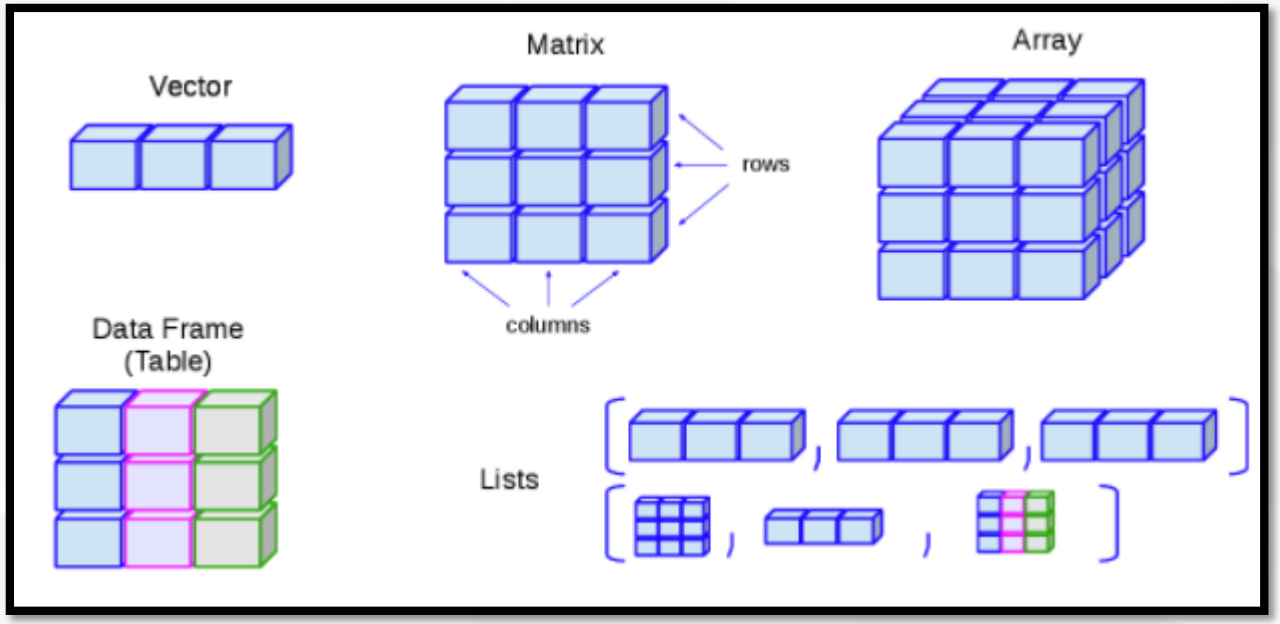
目录

基于R语言的数据工程解决方案



用R语言进行数据处理

合并、分组、排序、筛选、转置、差分、填充、移动、清洗、回归、分布检验、高数计算。

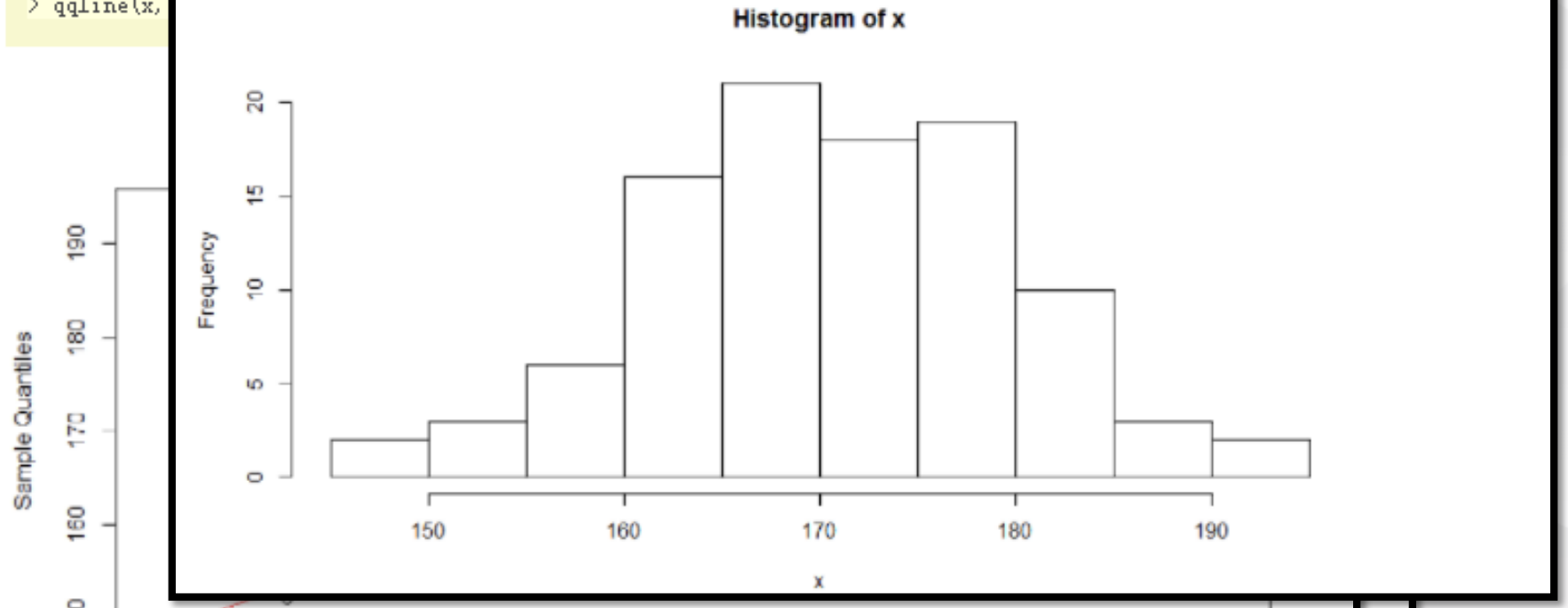


正态分布

```
# 生成身高数据
> set.seed(1)
> x<-round(rnor
> head(x,20)
 [1] 164 172 16

# 画出散点图
> plot(x)
```

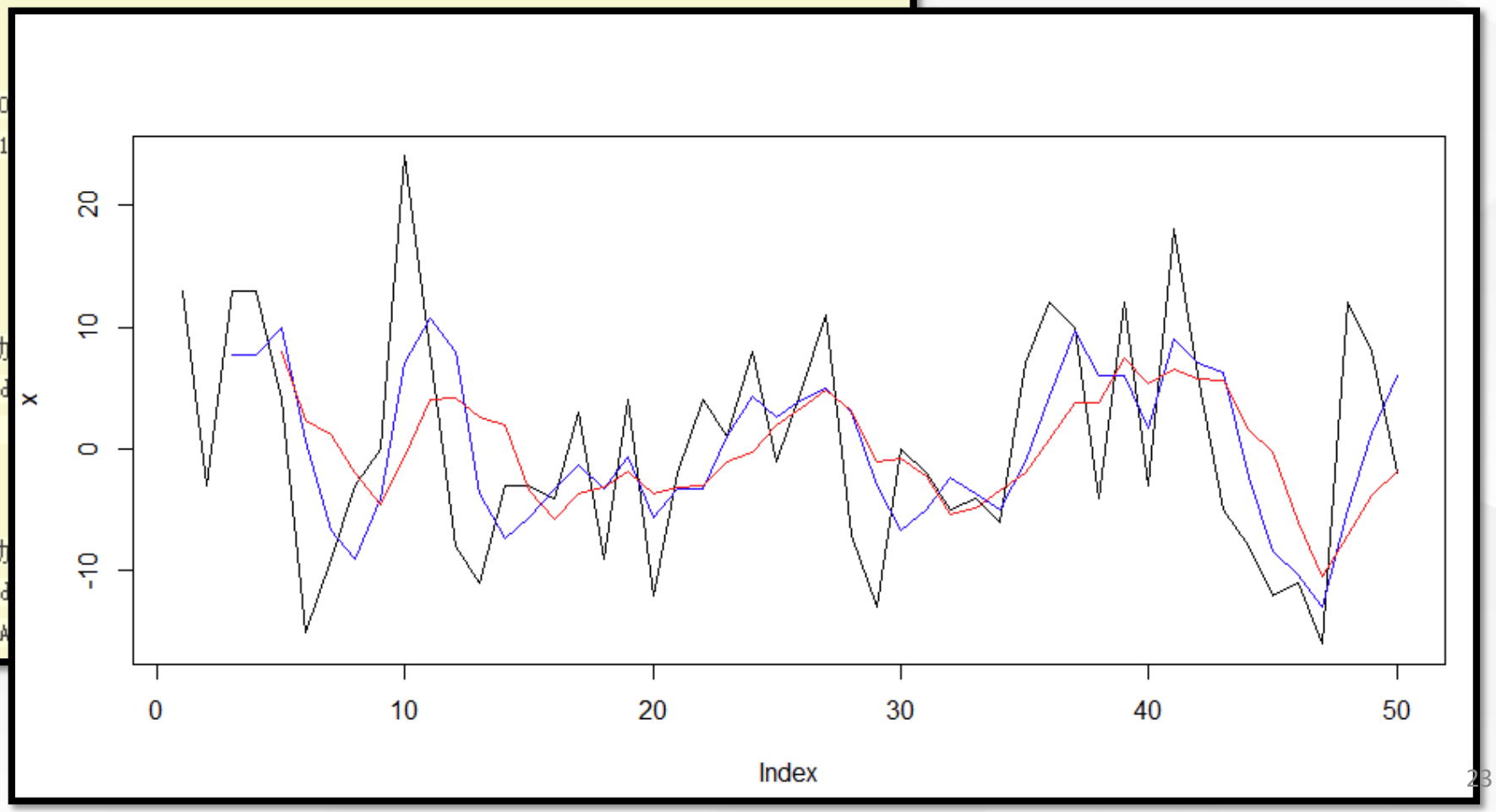
```
> shapiro.test(x)
```

```
> qqnorm(x)
> qqline(x,
```



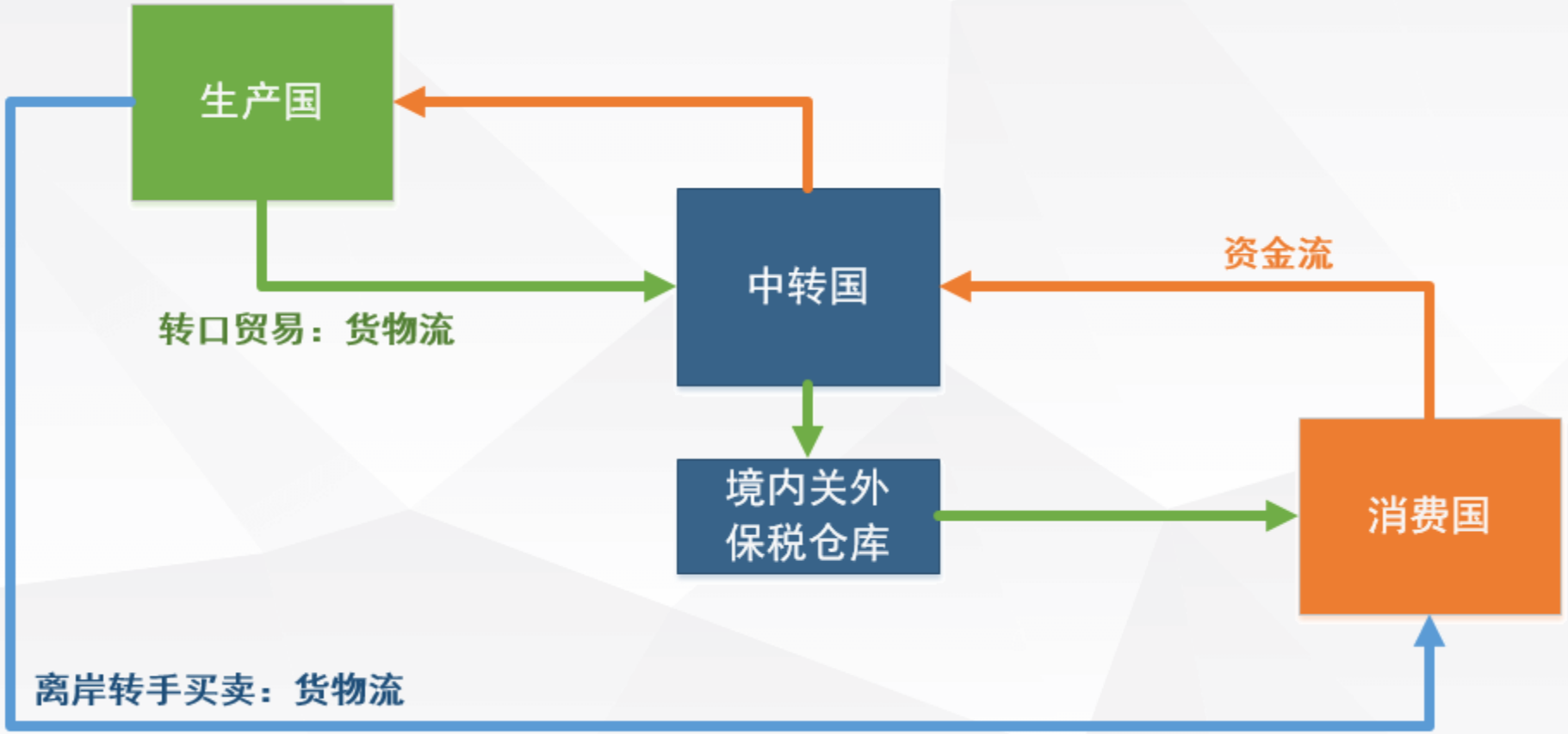
移动平均(MA)

```
# 生成50个随机数
> set.seed(0)
> x<-round(rnorm(50
 [1]  13  -3  13  1

# 加载TTR包
> library(TTR)

# 计算周期为3的移动
> m3<-SMA(x,3);head
 [1]        NA
[10]  7.0000000

# 计算周期为5的移动
> m5<-SMA(x,5);head
 [1]   NA   NA   NA
```



模型案例：转口贸易

在转口贸易中，货物流和资金流分离，利用货物贸易的资金通道，进行资金套利的虚假贸易企业。



指标体系

指标层

特征层


业务规则度量衍生指标
贸易指标
交易行为
商品指标
总体特征
统计+业务衍生指标
基本特征
收支指标
贸易信贷
国别指标
结售汇指标
异常分类
数据平台
数据整合、清洗、过滤、填充、变形、转换、分布检查、标准化、离散化
原始数据
+
业务特点
基本信息维度
国际收支维度
海关维度
财务登记维度
汇率维度
贸易政策
异常登记维度
美元兑人民币中间价
各个币种对美元
利率
关税

模型结果和解读

异常特征：离岸转手买卖

“铜”进出口，交易量巨大且高频，经营范围不匹配，高频交易，对手复杂，1年稳定贸易逆差。

操作：期限套利

在现货市场持有现货铜，在期货市场反向操作，进行套利。

总结

- 本文从数据思维开始，对比了IT思维和数据思维的区别，介绍如何开展一个数据项目的，到R语言的代码操作细节过程，最后到结合数据项目案例。

- 以数据思维做数据，发现数据价值！

张丹，青萌数海CTO，微软MVP，数据科学家。

10年以上互联网应用架构经验，在R、Java、NodeJS、大数据、数据挖掘等方面有深厚的积累。

精通量化投资交易策略，熟悉中国金融二级市场、交易规则和投研体系。 熟悉数据学科方法论，在外汇、海关、区块链等领域均有落地的尝试。

著有《R的极客理想：量化投资篇》、《R的极客理想：工具篇》、《R的极客理想：高级开发篇》，英文版图书被CRC出版集团引进，在美国发行。个人博客：http://fens.me 。

北京青萌数海科技有限公司
为企业、政府等客户提供数据分析服务的高科技公司。

谢谢您的聆听

如果你和我一样，欢迎加入我的团队！