数萃培训课程(中级系列)

2017年6月

#### 培训对象：

* 高校数据科学及相关专业教师
* 企业数据分析员
* 在校学生(计算机，应用数学)

### **数据科学与大数据技术专业（代码080910T）**

2016年2月，教育部公布新增“数据科学与大数据技术”本科专业

“数据科学与大数据技术”专业（专业代码080910T）强调培养具有多学科交叉能力的大数据人才。该专业重点培养具有以下三方面素质的人才：一是理论性的，主要是对数据科学中模型的理解和运用；二是实践性的，主要是处理实际数据的能力。三是应用性的，主要是利用大数据的方法解决具体行业应用问题的能力。

该专业包括基础课程、核心课程及选修课程三大模块。其中专业基础课程涵盖了数学、统计学、计算机科学等理论知识。专业核心课程侧重数据分析、计算、计算机技术以及大数据应用等方面的内容，包括了《数据科学导引》、《数据采集》及《大数据应用导论》等课程。专业选修课程在专业核心课程的基础上深入学习，包括《深度学习》、《大数据分析的算法》、《云计算与大数据平台》等，同时也在应用方面开设了新兴的行业课程，例如《金融大数据应用》、《健康医疗大数据：理论与应用》、《交通大数据：理论与应用》等。

培养目标：  
  
“数据科学与大数据技术”专业，培养德、智、体、美全面发展，掌握数据科学的基础知识、理论、及技术，包括面向大数据应用的数学、统计，计算机等学科基础知识，数据建模、高效分析与处理， 统计学推断的基本理论、基本方法和基本技能。对自然科学和社会科学等应用领域中大数据的了解，具有较强的专业能力和良好外语运用能力，能胜任数据分析与挖掘算法研究和大数据系统开发的研究型和技术型人才。

### **大数据技术与应用（代码610215）**

2016年9月，教育部公布新增“大数据技术与应用”专科专业

“大数据技术与应用”专业（专业代码610215）强调培养具有大数据实践能力的大数据人才。该专业重点培养具有以下两方面素质的人才：一是工具的掌握，掌握数据采集和数据分析的基本工具；二是数据分析能力，掌握实用数据分析和初步数据建模能力。

该专业包括基础课程、核心课程及选修课程三大模块。其中专业基础部分侧重为语言和专业基础方面的课程，包括《大数据的Python基础》、《Linux系统基础》、《大数据的统计基础》等。专业核心部分涵盖了数据采集、存储与处理方面的内容，包括《数据存储（MySQL）》及《数据清洗》等课程。专业选修部分以大数据分析、开发及应用为重点，开设包括《数据分析导论》、《大数据行业应用导论》、《数据可视化》、《Hadoop大数据平台基础》等课程。

培养目标：  
  
“大数据技术与应用”专业，培养掌握数据科学的基础知识及大数据相关技术，掌握大数据清洗和分析常用工具的使用，具有卓越的实践能力，能胜任数据清洗、数据存储、数据分析与挖掘、大数据系统开发与构建等工作的专业应用型人才。

### **思考**

#### 数据从业人员定位？

* 后端(数据库，数据清洗)
* 中端(建模)
* 前端(Web展示, 可视化)

#### 如何学/教？

* 初级阶段:
  + 一门主流语言(R/Python)
  + 常用统计方法(回归，聚类等)
  + 数据可视化(R/python基本图形库与扩展库)
* 中级阶段:
  + 统计计算与编程
  + 高级统计分析(时间序列，变量选择与模型选择/评估，贝叶斯分析) ->
  + 数据挖掘/机器学习
* 高级阶段：
  + 深度学习 (文本挖掘，自然语言处理，社交网络及图模型，人工智能)
  + 大数据与高性能计算：并行计算 -> 分布式计算(Hadoop/Spark) Scala, sparkR, Microsoft R
  + 大数据平台与开发（javascript, node.js, gpu编程，docker技术）

#### 主要教/学什么？

在校(统计学)学生应该主要掌握：

1. 数据库技术(熟练MySQL及不同数据的转换)
2. 一门大数据分析语言(精通能编程)
3. 大数据常用统计建模方法(熟练并解释)
4. 大数据常用算法建模方法（思想与使用）
5. 大数据高性能计算方法（基本）
6. 大数据分析技术(了解)

#### 培养/考核技能：

1. R/python语言基础
2. 统计基础知识
3. ETL基础
4. 数据收集(爬虫)
5. 数据可视化
6. 自动化报告
7. R/python编程与开发
8. 高性能计算
9. 大数据平台使用与高性能计算

### **高校大数据课程设置**

#### 基础核心课程:

1. 大数据导论
2. 统计分析基础
3. 数据库与数据处理
4. 编程基础:R/python

#### 大数据分析必修课:

1. （基于R）数据可视化
2. （基于R）统计机器学习
3. （基于R）编程与高性能计算
4. （基于R）大数据平台使用

#### 大数据分析选修课:

1. 基于python的大数据统计分析
2. 贝叶斯分析与应用
3. 文本数据处理与爬虫技术
4. 深度学习与应用

### **众创数萃中级大数据分析师培训大纲**

#### **基础培训课程(常年开放)**

A-1. 大数据导论(陈鹏程，Schubert, 林子雨, 周宁奕, 周扬, 黄志敏, 林祯舜，牟刚)  
A-2. 统计分析基础(丁辉，张日权)  
A-3. 数据库与数据处理(郞大为，李浩)  
A-4. R语言基础(汤银才，张东，练勇强)  
A-5. python语言基础(肖凯，靳军)

### A-1. 大数据导论

* 课程类型: 大数据分析普及课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 企业高层管理者
  + 数据业务主管
* 要求: 无
* 讲师: 企业数据科学家
  + Schubert, 林子雨, 周宁奕, 周扬, 黄志敏, 林祯舜，牟刚
* 时间: 1天/6小时
* 价格: 5800/2800

References:  
  
\* Data.Science.For.Dummies(2nd Ed)  
\* [dummies.com](http://www.dummies.com/programming/big-data/data-science/data-science-for-dummies-cheat-sheet/)

1. 大数据概念
2. 大数据行业动态
3. 大数据案例欣赏
4. 大数据关键技术
5. 大数据分析知识库
6. 大数据人才需求分析

### A-2. 统计分析基础

* 课程类型: 大数据分析师初级课程
* 课程简介:
* - 数据是米，模型是水，想做出好吃的饭，还得用统计思想这把火。  
  - 烧一道好菜，你需要三个要素：上好的食材，好用的炊具，以及好的烹饪方法。研究大数据也是同样的道理。想做好大数据的工作，你的食材是大数据，你的炊具是计算机，而最重要的烹饪方法则是统计思维。值得一提的是，食材和炊具只要肯花钱都可以买得到，唯有烹饪方法是你的——你必须自己学。(林共进)  
  - 统计学是数据分析的灵魂，是互联网+信息化时代大数据科学的核心。在21世纪，人们将广泛认识到：统计学是科学思想的中心成分之一。美国知名调查机构PayScale及CareerCast网站显示2016年和2017年最佳工作均与统计学相关。C.R.Rao教授：人类活动内的一切领域都要求统计学的专业知识和技术，统计思维总有一天会像读与写一样成为一个有效率公民的必备能力。  
  - 本课程共二大模块七个章节，分别从一维和二维数据角度讲述数据分析背后深刻的统计思想、基本原理和分析方法，让学员从理论高度了解数据呈现规律的统计描述、量化方法和推断技巧。  
  - 本课程是数据分析的基础课程，参与者可以通过学习获得对数据的敏感性、统计模型的熟练性和数据分析结果解释的正确性。
* 客户获益:
* - 从数据的分布与相互关系上掌握统计建模的诀窍  
  - 诊断数据处理与分析存在的问题  
  - 获取进一步学习统计、用好统计的
* 适合人群：
  + 数据分析人员
  + 数据有关管理者
* 要求: 高等数学
* 讲师: 丁辉，张日权
* 课时: 3天/18小时
* 价格: 3000/2400

M1：一维数据的统计分析

1. 数据分析与统计思想

* 概论论与统计学
* 测量/误差与随机/分布
* 概率与计算
* 数据/随机变量与分布
* 独立与相关性

1. 描述性统计分析

* 总体与样本
* 中心趋势的度量
* 离散程度的度量
* 其他特征量: 偏度、峰度、极差、异常值
* 数据的图表展示

1. 统计推断

* 常用的统计分布
* 统计推断的基本问题
* 大数定律与中心极限定理
* 数据、参数与似然函数
* 统计检验、p值与功效
* 估计的精度与置信区间

1. 常用分布的统计推断

* 正态分布的推断：估计与拟合
* 区间估计与样本量的确定
* t分布与t检验: 单样本、两样本
* 二项分布的推断：估计与检验
* 泊松分布的推断：估计与检验

M2： 二维数据的统计分析

1. 二维数据的统计推断

* 二维数据的图形比较
* 相关性度量与计算
* 二维正态分布与特征量
* 二个正态总体的比较
* 二个比例比较

1. 线性回归

* 模型假设
* 简单线性回归
* 多元性回归
* 数据变换
* 多项式回归
* 回归预测与变量选择

1. 方差分析

* 单因素方差分析
* 单因素协方差分析
* 双因素方差分析
* 重复测量方差分析
* 用回归做方差分析

### A-3. 数据库与数据处理

* 课程类型: 大数据分析普及课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
* 要求: Excel
* 讲师: 郞大为，李浩
* 课时: 2天/12小时
* 价格: 2400/2000

References:  
  
\* 2008. Data Manipulation with R  
\* Using SQL in R  
\* 2016. Advanced R - Data Programming and the Cloud  
\* 2016. Data Wrangling with R

M1. 数据与数据库

1. 数据基础

* 数据的类型
* 数据的运算
* 结构化与非结构化数据
* 数据库简介
* 数据库操作语言
* 案例：火车时刻表的分析

1. [数据库基础](http://www.w3school.com.cn/sql/sql_intro.asp)

* SQL简介
* 常用SQL语法
* 数据表之间的关联
* SQL函数与Group
* 常用数据库简介
* 案例: 电商销售数据, SQL汇总分析

M2. 数据的获取

1. R语言数据获取

* 读取文件中的数据
* 读取Excel数据表
  + RODBC
  + xlsx
  + readr
* 读取其他类型的数据
* 读取其他统计软件格式的数据(SAS, Stata, SPSS)

1. R与数据库的连接

* 使用RMySQL访问MySQL关系型数据库
  + 安装MySQL, RMySQL
  + 从MySQL读取数据
* 基于DBI包访问数据库
  + RJDBC
  + ROracle
  + RPostgreSQL
  + RSQLite
* 数据库操作: sqldf包简介
* 从非关系型数据库NoSQL读取数据：RMongodb

1. 数据的爬取

* 爬虫基础
* 爬虫软件
* 基于R的爬虫
* 基于python爬虫

M3. 基于R数据整理

1. 常规的R包

* reshape2：数据的变形与汇总
  + melt()
  + cast()
  + 数据汇总
* plyr：数据的拆分、应用、合并
  + adply()
  + ddply()
  + mdply()
  + 按组运算

1. 更强易用的R包

* data.table：更快地完成数据集的数据处理
* dplyr
  + dplyr的基本函数
  + 数据汇总,数据连接
  + dplyr连接数据库
  + 案例: 使用dplyr整理汽车经销商数据
* lubridate：处理时间数据

### A-4. R语言基础

* 课程类型: 大数据分析普及课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
* 要求: Office/EXCEL
* 讲师: 汤银才，张东，练勇强
* 课时: 4天/24小时
* 价格: 4800/3800

M1：R入门

1. 大数据与数据科学

* 数据科学与分析工具
* R及其优势
* R安装与配置
* R包安装与使用
* R资源与帮助

1. R快速入门教程

* R中的基本语法
* R中的数据对象及其属性
* R的工作空间与管理
* R编程基础
* R程序调试

1. R编辑器与RStudio

* R常用编辑器
* Rstudio功能与使用技巧
* Rstudio进阶
* 项目管理
* Rmarkdown与报告生成

M2：R数据集创建与管理

1. 数据集的创建

* 常用数据对象与创建
  + 标量
  + 向量
  + 因子
  + 矩阵
  + 数据框(集)
  + 数组
  + 列表
* 数据的读取
* 数据的存储
* 数据集的合并与子集提取
* 缺失值的处理
* 数据表数据的切片、切块与组合

1. 数据操作: 对于向量的处理

* 合并数据框的行(向量)与列(向量)
* apply系列函数
* 数据分组并调用函数
* 数据拆分与合并
* 数据排序
* 访问数据中的列
* 查找符合条件的数据索引
* 分组运算

1. 数据对象的操作

* 赋值与常用运算
* 基本的数学运算
* 用于矩阵的运算
* 与统计分布相关的函数
  + 概率
  + 密度/质量
  + 分位数
  + 随机数产生

M3：R绘图初步

1. 基本的绘图命令

* 大趋势：信息可视化
* R绘图基础: 低级与高级绘图命令
* 基本绘图函数: plot, points, lines, curve
* 绘图三要素设置详解(颜色，点型，线型)
* 绘图信息补充(title,text,legend,axis)

1. 一维数据的可视化

* 常用统计分布与4类函数
* 一维离散变量的分布图示
* 一维连续变量的分布图示
* 一维连续分布诊断图
* 非参数密度估计与展示

1. 二维数据的可视化

* 二个离散变量的分布图示
* 二个混合变量的分布图示
* 二个连续变量的分布图示
* 多变量的可视化

M4：R数据探索与比较分析

1. 数据的描述性统计分析

* 常用描述性统计量及其计算
* 单个连续型变量描述性统计量的获取
* 分组计算描述性统计量

1. 相关性度量

* 变量的类型与转换
* 两个定性变量之间的关联性
* 两个有序变量之间的关联性
* 两个定量变量之间的关联性
* 定性变量与定量变量之间的关联性

1. 相关性检验

* 组间差异比较
  + 独立样本的t检验
  + 非独立样本的t检验
  + 组间差异的非参数检验
* 分类变量比较
  + 列联表的生成
  + 联合分布、边际分布与条件分布
  + 独立性检验(卡方检验，Fisher精确检验，McNemar检验，Cochran-Mantel-Haenszel检验)

M5：R统计建模

1. 回归模型

* lm()函数中的公式表示
* 一元线性回归
* 多元线性回归
* 回归预测
* 分位数回归

1. 广义线性模型

* 广义线性模型概述
* glm()函数介绍
* logistic回归
* Poisson回归

1. 模型的检验与比较

* 回归模型诊断
* 变量选择
* 模型比较
* 异常值判断
* 预测与交叉检证

### A-5. Python基础

* 课程类型: 大数据分析普及课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 企业高层管理者
  + 数据业务主管
* 要求: Office/EXCEL
* 讲师: 肖凯，靳军
* 课时: 4天/24小时
* 价格: 4800/3000

- Reference  
  
\* 2014. Python Data Analysis  
\* 2014. matplotlib Plotting Cookbook  
\* 2015. Python Data Science Essentials  
\* Regression Analysis with Python

* 第1讲：数据分析方法概述及相关工具
  + 认识数据
  + 数据分析的步骤和原则（确定/分解/评估/决策）
  + 相关工具概述及对比（excel/spss/R/matlab/python/Java）
* 第2讲：python环境和基础语法
  + 安装anaconda套件
  + 基础环境jupyter/ipython
  + 基本数据结构
  + 基本语法
  + 迭代器
  + 函数
  + python在业界的应用案例分享
* 第3讲：数据操作与计算
  + numpy数组与操作
  + numpy统计函数
  + numpy线性代数
  + 科学计算和最优化(scipy)
  + 强大灵活的数据结构 pandas DataFrame
* 第4讲：绘图与可视化
  + matplotlib中绘图
  + pandas中绘图
  + Charts中绘图
  + ggplot, seaborn
  + 交互式可视化bokeh
  + 在线工具plot.ly
  + 高级应用案例分享
  + 绘制地图
  + 绘制3D图形
* 第5讲：统计分析库
  + 概率和统计分析
  + 时间序列分析
  + 简单回归分析(statsmodels,scikit-learn)
  + 多元回归分析(statsmodels,scikit-learn)
  + 多项式回归
  + logistic回归(statsmodels)

#### **中级培训课程(定期开放)**

B-1. 数据可视化(谢佳标, 王旭，魏鹏)  
B-2. 高级统计分析(丁辉，徐安察，李洪成)  
B-3. 贝叶斯分析与应用(徐安察，汤银才，张东)  
B-4. 统计机器学习(谢佳标，尹志，王旭)  
B-5. 网络爬虫与文本挖掘(尹志，靳军)

### B-1. 数据可视化

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 基于R大数据分析与开发的用户
* 要求: R基础
* 讲师: 谢佳标, 王旭，魏鹏
* 课时: 6天/36小时
* 价格: 8000/6400
* References:
  1. Business Intelligence with R
* R数据可视化手册(R Graphics Cookbook)
  1. Exploratory Multivariate Analysis by Example - Using R
* [2014. Displaying Time Series，Spatial and Space－Time Data with R](https://pakillo.github.io/R-GIS-tutorial/)
* Interactive and Dynamic Graphics for Data Analysis with R and GGobi

M1: 绘图品质的提升

1. R绘图系统

* R graphics基本绘图命令(复习)
* 区域分割与绘图
* 图形输出常见问题处理
* 常用统计在不同系统中的比较

1. 图形的渲染

* rainbow函数
* 高质量图形渲染库Cairo
* RColorBrewer扩展包
* scales: brewer.pal

M2: 常用绘图系统介绍

1. lattice绘图系统

* 一个简单的lattice例子
* 图形参数设置
* 面板函数
* 图形的叠加：条件变量与条件变量设置
* 其他常用统计图

1. ggplot2绘图系统

* 最简单的绘图函数：qplot
* ggplot2的语法：以散点图为例
* 图层与统计图展示
* ggthemes主题包介绍

M3: 高维数据可视化

1. 3D可视化

* 3D静态可视化基础
* 3D静态可视化提高
* 常用动态3D可视化包介绍
* 动态3D散点图
* 基于RGL 3D可视化

1. 交互式绘图包

* rCharts（nPlot, hPlot, mPlot）
* recharts
* plotly
* bokeh/rbokeh
* REmap

1. R中的仪表盘

* flexdashboard
* shinydashboard
* shiny仪表盘应用: 美国大选数据可视化
* PowerBI中使用R制作仪表盘

M4. 统计机器学习中的可视化

1. 多元统计中的可视化

* PCA可视化(FactoMineR)
* CA与MCA可视化(FactoMineR)
* 基于Factoshiny函数的可视化
* 时间序列绘图(时间为条件变量，分组变量，补充变量)

1. 基于GIS的动态可视化

* 能用地图
* googleis
* RgoogleMaps
* 空间数据绘图(热图，参考图，物理图)
* 时空数据绘图
* 案例：R语言天气可视化应用(张丹)

### B-2. 高级统计分析

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 基于R大数据分析与开发的用户
* 要求: 统计分析基础
* 讲师: 丁辉，徐安察, 李洪成
* 课时: 3天/18小时
* 价格: 4600/3000

M1. 时间序列分析

1. 随机过程与时间序列
   * 时间序列的特征
   * 时间序列的描述性分析
   * 白噪声与平稳性
2. 平稳时间序列模型
   * AR模型
   * MA模型
   * ARMA模型
3. 非平稳序列模型
   * ARIMA模型
   * 季节效应
4. 异方差模型
   * ARCH模型
   * GARCH模型
   * EGARCH模型
   * SV模型

M2. 多元统计分析

1. 多变量回归
   * 多变量回归分析
   * 协方差分析
2. 多变量降维
   * 岭回归与Lasso
   * 主成分分析
3. 多变量分类
   * 分类
   * 判别分析
   * 聚类分析

M3. 重抽样方法

1. 随机数据的产生
   * 常用随机数据的产生
   * 重要性抽样
2. 蒙特卡罗方法
   * 蒙特卡罗积分与方差减少技术
   * 自助法
   * jacknife

M4. 最优化方法

1. 极值问题
   * 极大似然估计
   * 一维最优化问题及其求解
   * 多维最优化问题及其求解
   * Laplace近似
   * EM算法
2. 其他优化算法
   * 线性规划
   * 遗传算法
   * 图优化

M5. 空间统计分析(2012. Spatial Data Analysis in Ecology and Agriculture Using R)

* 空间数据的自相关性
* 空间自相关性度量
* 空间自相关数据的收集与整理
* 空间自相关数据的多元分析(PCA, 回归树，随机森林)
* 空间自相关数据的多元分析
* 空间自相关数据的回归分析

### B-3. 贝叶斯分析与应用

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 基于R大数据分析与开发的用户
* 要求: R基础，统计分析基础
* 讲师：徐安察，汤银才，张东
* 课时: 6天/36小时
* 价格: 6800/5500

M1. 贝叶斯分析入门

1. 贝叶斯分析概述
   * 先验信息与提取
   * 贝叶斯公式
   * 后验推断
   * 估计
   * 检验
   * 预测
2. 单参数贝叶斯模型
   * 二项分布: 成功率
   * 正态分布: 均值/方差
   * 指数分布
   * 泊淞分布

M2. 贝叶斯分析进阶

1. 多参数贝叶斯模型
   * 正态分布
   * 多项分布
2. MCMC方法
   * 贝叶斯计算
   * 抽样方法(复习)
   * M-H算法
   * Gibbs抽样
   * 常用的M-H算法
3. MCMC软件:
   * BUGS, WinBUGS, OpenBUGS及在R中的实现
   * JAGS与rjags, runjags
   * Stan与rstan

M3. 实用贝叶斯模型

1. 回归模型(Bayesian essentials with R)
   * 线性模型
   * 最小二乘估计
   * 基于Jeffreys先验分析
   * 基于G-先验分析
2. 广义线性模型(Bayesian essentials with R)
   * logit模型
   * Probit模型
   * 对数线性模型
3. 分层贝叶斯模型
   * 分层贝斯模型的构建
   * 正态分布
   * logistic回归模型
4. 混合模型(Bayesian essentials with R)
   * 有限混合
   * EM解决方法
   * MCMC解决方法
   * 未知混合个体

M4. 贝叶斯方法的应用

1. 时间序列分析(Bayesian essentials with R)
   * AR模型
   * MA模型
   * ARMA模型
   * 隐马尔可夫模型
2. Capture-Recapture模型
   * C-R模型(Bayesian essentials with R)
   * 空间Capture-Recapture模型
3. 贝叶斯网络(Learning Probability Graphical Models in R)
   * 有向图
   * 贝叶斯网络
4. 贝叶斯时空统计分析
   * WinBUGS/R2WinBUGS
   * INLA

### B-4. 统计机器学习

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 基于R大数据分析与开发的用户
* 要求: R基础，高维统计分析
* 讲师：谢佳标，尹志，王旭
* 课时: 6天/36小时
* 价格: 5800/4500

M1：认识数据挖掘

1. 数据挖掘概述
   * 数据挖掘过程
   * 数据挖掘对象
   * 数据挖掘方法
   * 数据挖掘应用
   * 无监督与有监督学习
   * 机器学习常用R包
     + e1071
     + caret
     + H2O Lagrange
2. 模型评估与选择
   * 分类的性能评价
   * 混淆矩阵
   * 风险图
   * ROC曲线及相关图表(plotroC, pRoC, ROCR)
   * 利用caret包比较ROC曲线
   * 交叉验证
   * K折交叉验证基本原理
   * 利用e1071包完成交叉验证
   * 利用caret包完成交叉验证

M2: 聚类分析

1. 几类常用的聚类方法
   * K-means
   * K-medoids
   * 系谱聚类
   * 密度聚类
   * 期望最大化聚类
   * 相关R包: cluster
   * 实例:
2. 隐变量模型
   * 概述
   * 混合模型
   * 隐马尔柯夫模型
   * 聚类分析
   * 实例：

M3: 线性分类器

1. 判别分析
   * 判别函数
     + 线性判别函数
     + 二次判别函数
   * 朴素贝叶斯分析
   * rocchio算法
   * 相关的R包: caret
   * 实例:
2. 支持向量机
   * 基本原理
   * 相关的R包
   * 可视化分析

M4: 非线性分类器

1. kNN
2. 决策树
   * 树的构建
   * 分类回归树CART
   * C4.5/5.0
   * 相关的R包: rpart
   * 实例:
3. 随机森林
   * 基本原理
   * 相关的R包：
     + randomForest, randomForestSRC, ggRandomForests,
     + gbm, glmnet, ranger

M5. 集成学习

* Bagging
* AdaBoost
* xgboost

### B-5. 网络爬虫与文本挖掘

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R/python基础
* 讲师: 尹志，靳军
* 课时: 4天/24小时
* 价格: 4800/3000

- Reference  
  
\* 2014. XML and Web Technologies for Data Sciences with R(BookZZ.org)  
\* 2015. Automated Data Collection with R

M1. 网络爬虫基础

1. 技术准备
   * 网络通信基础
   * HTTP协议简介
   * Web开发知识介绍
   * 网站分析知识介绍
2. 开发环境与语言
   * 开发环境安装与使用（Anaconda套件与PyCharm）
   * Python基础数据结构（元组/列表/字符串/字典）
   * Python基础语法（条件/循环/函数/类/模块）
   * 常用Python库使用案例分享
   * Python技巧与实践分享

M2. 数据爬取与存贮

1. 网络爬虫工具库
   * 基础Python爬虫库（urllib/Requests）
   * 认识正则表达式
   * “漂亮”的爬虫库-Beautiful Soup
   * 静态网页爬取案例分享
   * Selenium与“幻影”浏览器- PhantomJS
   * Ajax和DHTML网站爬取
   * 动态网页爬取案例分享
   * 利用API进行数据采集
2. 网络爬虫存储
   * 文件读取与保存
   * 关系数据库存储-MySQL
   * 爬虫配合MySQL存储案例分享
   * 分布式存储-NoSQL数据库
   * 爬虫配合MongoDB存储案例分享
   * HDFS简介

M3. 网络爬虫提升

1. 分布式爬虫
   * 多线程爬虫
   * 多进程爬虫
   * 爬虫队列设计
   * 集群化爬取
2. 网络爬虫框架
   * Python网络爬虫框架介绍
   * Scrapy基本使用
   * Scrapy进阶使用
   * 爬虫框架使用案例分享
3. 网络爬虫突破
   * 模拟登录
   * 常见验证码突破
   * 爬虫代理池
   * 各类网页内容处理
   * 爬取移动端APP技巧
   * 设计健壮的网络爬虫

M4. 文本挖掘

1. 文本挖掘技术基础
   * 文本挖掘全流程概述
   * 自然语言处理库（NLTK）
   * TextBlob文本处理库介绍
   * 中文分词介绍（jieba）
   * 词云介绍
2. 文本挖掘技术进阶
   * 文本挖掘预处理技术
   * 文本特征处理
   * 文本聚类
   * 主题模型
   * 基于深度学习的文本挖掘
   * 文本挖掘案例分享

#### **拓展培训课程(不定期开放)**

C-1. 数据治理(郞大为，唐力，李浩)  
C-2. 深度学习(尹志，魏鹏，王旭)  
C-3. R语言编程与开发(谢佳标，王旭)  
C-4. 大数据平台技术与应用(尹志，谢佳标, 刘逸铭)  
C-5. Python大数据分析(肖凯，李浩)

### C-1. 数据治理

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R/python基础
* 讲师: 郞大为，唐力，李浩
* 课时: 3天/18小时
* 价格: 3800/3000

### C-2. 深度学习

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R/python基础，统计机器学习
* 讲师: 尹志，魏鹏，王旭
* 课时: 4天/24小时
* 价格: 4800/4000

-------≈

M1: 基础知识与准备

1. 深度学习背景
   * 什么是深度学习
   * 传统机器学习局限性
   * 深度学习反思
2. 深度学习框架
   * 八大深度学习框架概述
   * tensorflow,caffe环境的准备
3. 人工神经网络
   * 感知器
   * 激活函数（sigmoid, ReLu等对比）
   * 梯度下降算法
   * 反向传播算法介绍与详细推导

M2: 深度神经网络与应用

1. 计算机视觉与卷积神经网络
   * 计算机视觉背景
   * 卷积神经网络训练细节(卷积操作和池化操作)
   * 图像检测与分割
   * caffe，MXNet使用
   * 案例(卷积神经网络): 手写数字识别
   * AlexNet，GoogLeNet解读
2. 自然语言处理与循环神经网络
   * 自然语言处理背景
   * 循环神经网络训练细节
   * 不同语言翻译（Translation）
   * caffe，TensorFlow使用
   * 案例(循环神经网络): 生成手写字符
   * 长短时记忆(LSTM)训练原理

M3. 无监督学习网络结构

1. 限制性波尔茨曼机
   * 波尔茨曼机网络结构
   * 限制性波尔茨曼机网络结构
   * CD算法
   * 限制性波尔茨曼机协同过滤上的应用
2. 深度信念网络
   * 贝叶斯网络
     + 表示
     + 推理（精确推理，近似推理）
     + 学习（参数估计与结构学习）
   * 深度信念网络
     + 基本网络结构
     + 参数学习
     + 应用
3. 自动编码器
   * 自动编码器结构与原理
   * 堆叠自动编码器结构与原理
   * 系数自动编码器结构与原理
4. 案例(无监督学习)：图片数据

### C-3. R语言编程与开发

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R基础，数据可视化
* 讲师: 谢佳标，王旭
* 课时: 6天/36小时
* 价格: 6800/5200

- References:  
\* Programming Graphical User Interfaces in R  
\* 2016. FasteR！\_HigheR！\_StrongeR！  
\* 2016. Efficient R programming  
\* 2015. R High Performance Programming  
\* 2013. Seamless R and C++ Integration with Rcpp  
\* Parall programming  
\* - 2012. Parallel R  
\* - 2009. PR---Parallel R  
\* 2009. State-of-the-art in Parallel Computing with R  
\* 2011. Hands-on tutorial for parallel computing with R. Computational Statistics

M1: 编程篇

1. R编程基础
   * R中的基本数据结构
   * R中的控制语句：if, for, while, repeat
   * 函数的构建与调用
   * 基本的debugging方法/函数
   * Scoping rules
2. R性能监控
   * 监控时间: system.time
   * 本地缓存工具memoise
   * 性能监控工具Rprof
   * summaryRprof
   * 性能可视化工具lineprof

M2. R编程进阶

M3. 高性能计算

1. 提升R的计算性能
   * R的缺陷与克服
   * 向量化编程/函数
   * apply系列函数： apply, lapply, sapply, tapply, mapply
   * 其他Apply方式
2. 并行计算
   * 并行化介绍
     + 并行程序化目的与设计
     + Flynn并行计算分类(SISD,SIMD,MIMD)
     + 并行问题的分类: embarrassingly-parallel, Data-parallel
     + 并行化线程沟通协议(库): PVM, MPI，NWS,Socket
     + 并行场景:多核心(CPU),多计算机(Cluster), Grid计算
     + 常用的R并行计算包与4种协议下的速度比较
   * 常用并行R包
     + Rmpi: 基于MPI
     + snow (支持4种沟通协议)
     + snowfall (snow的简单化封装)
     + multicore (单计算机)
     + Parallel: multicore+snow (多计算机)
     + apply类函数1(parApply,parLapply,parSapply)
     + apply类函数2(mclapply, mcmapply,需要fork进程支持)
     + nws(支持NWS-NetWorkSpace)
     + foreach: 代替apply系列函数, for
     + doMC: multicore+foreach (需要fork进程支持)
     + doParallel: multicore+foreach+snow
     + 案例1: K-means算法在平行实现
     + 案例2: 交叉验证
3. 其他并行问题
   * 在Hadoop/Spark上并行
   * 并行MCMC介绍
   * cudaBayesreg：CUDA上贝叶斯分析
   * RScaLAPACK: 封装线性代数库ScaLAPACK
   * pRapply: lapply并行化

M4: 开发篇

1. R包开发
   * R包的构成
   * R包重新编译与安装
   * Rstudio中开发R包
   * 标准化R包开发流程
   * 案例: R语言天气可视化
   * 案例: 每日中国天气
2. R GUI开发
   * 早期开发平台: GWidgets, GTK/RGtk2, Tcl/Tk, Qt
   * 示例: 基于gWidgets和gWidgetsRGtk2计算软件开发
   * Rpanel
3. [基于HTMLWidgets应用与widget开发](https://www.htmlwidgets.org)
   * HTMLWidgets介绍
   * 成熟的R Widget
   * Widget创建
4. 跨平台通信
   * Rserver与Java
   * Rsession与Java
   * rJava
   * Node.js与R
   * 案例: R中的BI实现
     + R + Tableau
     + R + PowerBI

M5. 基于shiny应用开发(Ref: Web Application Development with R Using Shiny)

1. shiny基础

* shiny简介
* 构建APP
* 部件设计
* 部署分享shinyapp

1. shiny高级开发

* 高级shiny技巧
* 开发案例

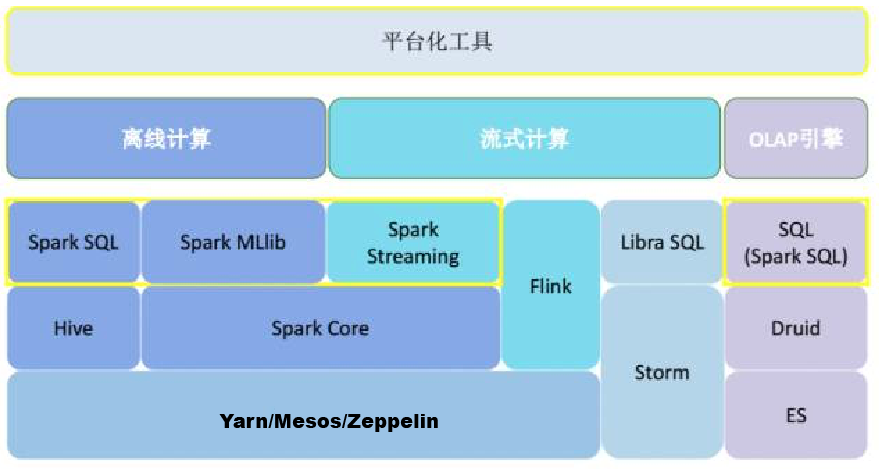
1. shinydashboard简介

M5. 基于R成熟应用介绍

1. 统计分析平台
   * R commander(Ref: Using the R commander)
   * R Commander插件
   * R Deducer
2. 数据挖掘平台
   * Rattle
3. R学习平台
   * swirl
4. 可视化编程平台
   * Red R
   * doodleBUGS

### C-4. 大数据平台技术与应用

* 课程类型: 大数据分析师中级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R基础，数据库基础
* 讲师：尹志，谢佳标，刘逸铭
* 课时: 6天/36小时
* 价格: 6800/5200



M1. Hadoop

1. Linux操作基础
   * 大数据分析操作系统
   * Unbuntu安装
   * Linux下常用的命令
2. Hadoop基础
   * Hadoop介绍
   * hadoop集群搭建(centos/unbuntu/linux/mac)
   * HDFS原理
   * HDFS Shell操作实战
   * YARN介绍
3. Haddop下的常用工具
   * Sqoop
     + Sqoop介绍与安装
     + Sqoop基础
     + Sqoop导入实战
     + Sqoop增量导入
     + Sqoop导出实战
     + Sqoop job
   * Hive
     + Hive架构
     + Hive环境搭建
     + Hive实战
     + Hive工作原理
     + 基于Hive的日志分析
     + 日志分析的ETL自动调度
   * HBase
     + HBase架构
     + HBase安装
     + HBase实战
     + Hive与HBase集成实战
   * kylin
     + kylin架构
     + kylin搭建
     + kylin实战
     + kylin性能调优
4. Rhadoop

M2. Scala

1. Scala语言基础
   * Scala语言概述
   * Scala基础：类、对象、继承、特质、模式匹配
2. Scala基本操作
   * 函数定义和高阶函数
   * 针对集合的操作
   * 遍历操作、map操作和flatMap操作、filter操作、reduce操作、fold操作
   * 函数式编程实例WordCount
3. scalaR

M3. Spark基础

1. Spark入门

* Spark简介
* Spark运行架构
* Spark工作原理
* Ubuntu环境的准备
* Spark集群搭建

1. Spark使用

* Spark开发环境
* 第一个Spark应用程序：WordCount
* 在集群上运行Spark应用程序
* Spark 监控管理
* Spark 应用程序部署
* Mapreduce和Spark模型的比较

M4. Spark提高

1. Spark SQL原理和实践

* Spark SQL简介
* Spark SQL 工具使用
* Spark SQL实例与编程
* DataFrame的创建
* 读取和保存数据（读写Parquet、通过JDBC连接数据库、连接Hive读写数据）

1. Spark Streaming原理和实践
   * Spark Streaming简介
   * Spark Streaming 流数据架构
   * Spark Streaming 原理与优化
   * DStream操作概述
   * 输入源
   * 转换操作
   * 输出操作
   * Storm 和Spark 的区别与比较
2. Spark MLlib实践
   * Spark MLlib简介与入门
   * 机器学习工作流
   * Spark MLlib 矩阵向量
   * Spark MLlib 回归与分类算法(逻辑斯蒂回归分类器、决策树分类器)
   * Spark MLlib 聚类算法(KMeans聚类算法、高斯混合模型(GMM)聚类算法)
   * Spark MLlib 关联规则算法
   * Spark MLlib 协同过滤推荐算法
   * Spark MLlib 特征抽取和处理
   * 案例：Spark 在线广告CTR 预测应用案例详解与分析
3. Spark 生态
   * Spark/Yarn
   * Spark/Mesos
   * Spark/Zeppelin
4. SparkR

M5. supR

### C-5. Python大数据分析

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 数据业务主管
  + 大数据分析与开发的用户
* 要求: R/python基础，统计机器学习
* 讲师: 肖凯，李浩
* 课时: 4天/24小时
* 价格: 5800/4800
* References
  1. Python Data Science Essentials
  2. Python Data Analysis
  3. Mastering Machine Learning with scikit－learn
* 第1讲：python编程与提高(复习)
  1. 基本语法（条件/循环/函数/类/模块）
  2. python语言编程的最佳实践经验
  3. Python爬虫
  4. 最简单的制作python包的方法
* 第2讲：数据挖掘概览
  1. 数据挖掘与模型
  2. 建模流程和步骤
  3. 常见算法
  4. 高级应用案例解析
* 第3讲：典型数据挖掘项目示范
  1. 基于titanic数据集预测生存概率
  2. 演示如何进行绘图探索
  3. 特征处理和建模
* 第4讲: 降维与异常值处理
  + PCA
  + LFA, LDA, ICA
  + RBM
  + 异常值判别
  + 正则化方法
  + 特征工程
* 第5讲：常用机器学习算法
  + scikit-learn
  + 聚类: k-Means
  + k-近邻
  + 二分类: 基于logistic回归
  + 朴素贝叶斯分类器
  + 感知机到SVM分类器
  + 感知机到ANN
  + 决策树
* 第6讲：机器学习进阶
  + 参数调优
  + 集成学习
  + 神经网络
  + 深度学习
* 第7讲：数据挖掘案例讲解
  + 基于MNIST数据集识别数字
  + 展示如何进行特征构造
  + 常规机器学习算法和深度学习算法的效果差异
  + 机器学习的结果评估
  + 深度学习的最新进展分享
* 第8讲：现代分析技术的应用
  + 文本挖掘(NLTK)
  + 社交网络分析(NLTK)
  + 空间数据的展示与地理信息分析
  + 图像分析技术简介

### **众创数萃大数据实战训练营**

D-1. 大数据统计分析实战训练营(R/python)(谢佳标，李洪成，张东)  
D-2. 大数据分析师实战训练营(R)(李舰，郞大为，练勇强)  
D-3. 大数据分析师实战训练营(python)(肖凯，尹志，李浩)  
D-4. web可视化工程师实战训练营(周宁奕，刘逸铭)  
D-5. 量化金融分析师实战训练营(张家齐，李孟育，靳军)

### D-1. 大数据统计分析实战训练营(R/python)

参考:

* [纽约数据科学研究院Data Science Bootcamps](http://nycdatascience.com)
* [约翰霍普金斯大学Data Science专项课程](https://www.coursera.org/specializations/jhu-data-science/1?utm_medium=courseDescripTop)
* [datasociety](https://datasociety.co)
* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 大数据分析与开发的用户
* 要求: R/python基础，统计基础
* 讲师: 谢佳标，李洪成，张东
* 课时: 10天/60小时
* 价格: 10800/8800

- Reference:  
  
\* 2011. Using R for data management, statistical analysis, and graphics

M1: 数据分析师工具

* 数据处理的常规工具(数据库，软件，展示)
* 训练营主要工具
  + R, RStudio
  + markdown, Rmarkdown
  + python, jupyternotebook
  + git, GitHub

M2: 收集与清洗数据

* 数据存贮系统
* 寻找与获取来源的数据
  + 网络（web）
  + 社交网络(微信，博客)
  + API
  + 数据库(MySQL)
* 数据清洗与处理

M3: 数据的探索性分析

* 特征量的提取
* 探索数据的分布
* 探索数据之间的关系
* 探索高维数据的信息
* 数据的动态可视化
* 探索数据的空间分布

M4: 数据的建模、预测与验证

* 模型的类型
* 预测流程
* 交叉验证
* 模型的评估
  + 回归模型评估
  + 分类模型评估

M5: 数据的相关性探索

* 相关分析
  + 简单相关关系
  + 自相关分析
  + 偏相关分析
  + 互相关分析
  + 典型相关分析
* 聚类分析
* 关联分析
  + 关联规则挖掘
  + 序列模式挖掘
* 主成分分析

M6: 数据的回归分析与优化

* 简单线性回归
* 多元线性回归
* 回归模型的诊断
  + 残差分析
  + 变量选择
  + 模型比较
  + 多重共线性
  + 离群值检测
* 回归的改进: 正则化
* Logistic回归

M7: 复杂回归分析

* 梯度提升回归树
* 神经网络
* 支持向量基
* 决策树
  + ID3算法
  + C4.5/5.0算法
  + CART算法
* 集成与随机森林

M8. 时间序列分析

1. 时间序列概述
   * 时间序列的特征
   * 时间序列的描述性分析
   * 白噪声与平稳性
2. 平稳时间序列模型: ARMA模型
3. 非平稳序列模型：ARIMA模型
4. 异方差模型: GARCH类模型

### D-2. 大数据分析师实战训练营(R)

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 大数据分析与开发的用户
* 要求: R基础，统计基础
* 讲师: 李舰，郞大为，练勇强
* 课时: 10天/60小时
* 价格: 10800/8800

李老师，毕业于北京大学，浙江大学软件学院兼职教授、华东师范大学硕士研究生导师，台北商业大学业界专业教师，曾任Mango Solutions（全球最大的R技术开发公司） 中国区数据总监，2015上海R语言大会主席，专注于数据科学在行业里的应用。2005年从事R语言相关研究，擅长R语言的工程开发与分析建模，是 Rweibo、Rwordseg、tmcn 等 R 包的作者，著有《数据科学中的R语言》，翻译了《R语言核心技术手册》、《机器学习与R语言》，撰写过大量R语言的基础和高级应用类文章，从事R语言数据挖掘分析培训多年，探索出一套以案例讲解带动理论理解和软件操作熟悉的方法。

* References:
* 2012.Customer and Business Analytics---Applied Data Mining for Business Decision Making Using R
  1. R FOR DATA SCIENCE （Dan Toomey）
* 课程目标:

1. 大数据分析
   * 大数据的基本特点
   * 大数据的分析技术
   * 大数据的存储与管理
   * Hadoop、Spark 等工具的应用
2. R语言应用
   * R 语言统计分析
   * R 语言可视化应用
   * R 语言数据挖掘与机器学习
   * R 语言的高性能运算

* M1. 数据基础与R入门
  + 认识数据
  + 数据分析方法概述
  + 大数据与数据科学概述
  + 常见分析工具概述
  + R漫谈
  + R工作环境介绍
  + Rstudio简介
  + R常用操作
  + R在业界应用案例分享
* M2. R语言数据操作与编程基础
  + R基础数据结构
  + 函数操作与函数式编程
  + 数据的读入与写出
  + 文件的操作
  + R与数据库操作
  + 常用函数介绍(数据操作、字符处理、日期处理)
  + 编程与控制语句(条件、循环、apply操作)
  + dplyr与数据处理
  + 案例1：编写一个模拟排队的函数
  + 案例2：制作一个R包
  + 案例3：电商数据的清洗
  + 案例4：网站文本数据的清洗
* M3. 统计建模
  + 线性回归与预测
  + 模型诊断
  + 回归扩展(非线性驾照、logistic驾照、lasso)
  + 主成分分析和因子分析
  + 聚类分析和判别分析
  + 多维变量的探索
  + 时间序列分析简介
  + 蒙特卡罗方法简介
  + 案例5：足球比赛数据分析
  + 案例6：销量数据的预测
* M4. 数据挖掘与机器学习
  + 无监督学习介绍
  + 关联规则
  + 案例7：零售数据的关联规则挖掘
  + 分类算法: 从logistic回归说起
  + 案例8：信用卡违约预测分析
  + 机器学习结果评估
  + 多重交叉验证
  + 常用分类方法（决策树、随机森林、支持向量基）
  + 案例9：
  + 神经网络与深度学习
  + 案例10：足球比赛数据分析的机器学习
* M5. 数据可视化
  + 描述性统计与统计图形介绍
  + R中图形设备与作图方式
  + 常用图形参数介绍
  + ggplot2介绍
  + 动态可视化示例与业界进展
  + 数据分布的研究(直方图、QQ图、热图)
  + 数据关系的探索(散点图与相关分析，箱线图与因子分析，马赛克图与残联表分析)
  + 统计图应用案例及常见误区
  + 案例9：地理数据的可视化
  + shiny介绍
  + 案例10：使用shiny开发一个小型动态分析系统
* M6. 现代分析技术的应用
  + 自然语言处理与文本挖掘
  + 案例11：网络舆情的文本挖掘
  + 社交网络分析
  + 案例12：诗人的社会关系
  + 空间数据的展示与地理信息分析
  + 图像分析技术简介
  + 最优化方法与运筹学简介

### D-3. 大数据分析师实战训练营(python)

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 大数据分析与开发的用户
* 要求: R基础，统计基础
* 讲师: 肖凯， 尹志， 李浩
* 课时: 10天/60小时
* 价格: 10800/8800
* 课程目标: 掌握python语言
  + 统计分析
  + 可视化应用
  + 数据挖掘与机器学习
  + 文本挖掘

M1. python基础

1. 数据分析方法概述及相关工具
   * 认识数据
   * 数据分析的步骤和原则（确定/分解/评估/决策）
   * 相关工具概述及对比（excel/spss/R/matlab/python/Java）
2. python环境和基础语法
   * 安装anaconda套件
   * 基本数据结构（列表/字符串/字典）
   * 基本语法（条件/循环/函数/类/模块）
   * 基础环境linux-shell/IDLE/notebook
   * python与ipython
   * python语言编程的最佳实践经验
   * python在业界的典型应用

M2. 常用工具库

1. 基础工具库
   * 数值计算 numpy
   * 绘图与可视化 matplotlib与 Chart
   * 数据操作 pandas
   * 高级应用案例分享
2. 统计分析库
   * 概率和统计分析(statsmodels)
   * 科学计算和最优化(scipy)
   * 线性回归和logistic回归(statsmodels)

M3. python数据挖掘

1. 数据挖掘初步
   * 数据挖掘与模型
   * 建模流程和步骤
   * 常见数据挖掘算法
   * 特征工程
   * 正则化方法
   * 主成分分析(PCA)
   * 高级应用案例解析
2. 数据挖掘算法详解
   * 决策树
   * 集成学习
   * 参数调优
   * 感知机
   * 神经网络
   * 深度学习

M4. 数据挖掘案例讲解

1. titanic数据集
   * 绘图探索: 可视化
   * 特征处理和建模
   * 生存概率预测
2. MNIST数据集: 数字识别
   * 特征构造
   * 常规机器学习算法和深度学习算法的效果差异
   * 机器学习的结果评估
   * 深度学习的最新进展分享

M5. python爬虫与文本挖掘

1. Python爬虫
   * 数据采集（HTML解析，API使用）
   * 存储数据（MySQL）
   * 基础爬虫库（urllib，request）
   * 认识正则表达式（re模块）
   * 漂亮的爬虫库（BeautifulSoup）
   * 方便的现成框架(Scrapy)
2. 文本挖掘技术基础
   * 文本挖掘全流程概述
   * 自然语言处理库（NLTK）
   * TextBlob文本处理库介绍
   * 中文分词介绍（jieba）
   * 词云介绍(wordcloud)
3. 文本挖掘技术进阶
   * 文本挖掘预处理技术
   * 文本特征处理
   * 文本聚类
   * 主题模型
   * 基于深度学习的文本挖掘
   * 文本挖掘案例分享

M6. 现代分析技术 - 社交网络分析 - 空间数据的展示与地理信息分析 - 图像分析技术简介

### D-4. web可视化工程师实战训练营

参考:   
- 可视化工程师修炼手册：跟着我爬虫+数据库+可视化八个快动作, 大数据文摘，2017-06-04

* 课程类型: 大数据分析师高级课程
* 课程简介:

在大数据时代，最为火爆的技术型岗位当属数据分析师(科学家)、算法工程师和全栈(前端)工程师。全栈(前端)工程师站在了时代的风口，他们既精通前端又能胜任后端(及移动端)，极大减少企业沟通成本。现在的调查分析表明，他们的就业趋向更好、需求更多，薪资也更胜一筹。现在web开发群体5成以上为全栈开发者，精通前后端的全栈工程师成了IT界新一代宠儿。然而其中懂得数据科学中数据动态可视化工程师成为大数据产业链中的稀缺人才。  
  
本课程由五大模板组成，逐步展示前后端web技术栈，  
通过对爬虫、数据库、数据清洗、可视化、部署上线等技术链路的技能学习和对地理可视化、graph可视化等专类可视化的实操经验积累，让学员学会在互联网+时代如何从容地获取数据、分析数据、通过动态可视化技术展示数据蕴含的价值。  
  
本课程是大数据分析可视化实战训练营，学员将在课程案例中学习与体验，在随堂作业与结业项目中进行网站数据爬取、存储、可视化，全面复习并实践教授的技能，真正做到学以致用！

* 客户获益:
* 适合人群：
  + 数据从业人员
  + 大数据分析与开发的用户
* 要求: 数据库基础, javascript基础
* 讲师: 周宁奕，刘铭逸
* 课时: 10天/60小时
* 价格: 10800/6800
* 课程目标:

1. 全面理解前后端web技术栈
2. 全程学习爬虫、数据库、数据清洗、可视化、部署上线等技术链路
3. 实操体验地理可视化、graph可视化等专门工具

M1: Web数据可视化概览

1. 可视化的纵与横

* 数据可视化基础
* 数据产业链路
* 技术预备知识
  + Nodejs
  + HTML5基础
  + SQL
* 案例：
  + 基于地图的房价分析系统，mapbox
  + 基于图与文本搜索的邮件门分析工具，palantir与本拉登

1. 服务器与web基础

* web生态简介
* 服务器的原理
* 服务器简介与操作
* HTML5基础

M2: 爬虫与数据处理

1. Node.js爬虫技术

* 如何获取数据，几种途径与特点
* Node.js爬虫的基本原理、实现方法、应用方式
* 实战（队列、请求池、模拟登录、ip代理）
* 扩展与延伸
* 案例：房价数据爬取

1. 关系型数据库Postgres SQL

* 数据库选型
* 导入数据、查询数据、更新数据、表结构与索引
* SQL实战
* 延伸: ORM、客户端与工具
* 案例：
  + 上海房价分析
  + 上海轨道交通分析

M3：图表可视化

1. D3.js图表基础
   * 使用D3.js开发3种基础图表
   * 开发实时更新的动态基础图表
   * 作业：自身图表库的构建、
   * 基于echart的可视化
   * 基于d3的通用图表可视化
2. D3.js图表进阶

* 延伸色彩的算法实现

M4: webGIS数据可视化

* 地理数据可视化分类
* 基于D3.js和leaflet的地理可视化
* 用dat.gui实现参数控制
* 更多延伸地理投影、色彩搭配
* 案例：
  + 莆田黑医院可视化
  + 房价项目可视化

M5: 图可视化(graph layout)

* 最简单的乒乓球游戏（JS/CSS动画）
* 物理系统的横向扩展：
  + canvas
  + konva.js
  + 盗梦空间
  + 爱舍尔的画与脑洞打开的游戏扩展（双曲面空间与投影）
* 物理引擎与图布局计算：
  + 泡泡图circle packing问题
  + 用Gephi优化图可视化算法
* 基于D3.js实现力引导图可视化
* 案例：
  + 希拉里邮件门分析
  + 社交网络分析插件

### D-5. 量化金融分析师实战训练营

* 课程类型: 大数据分析师高级课程
* 课程简介:
* 客户获益:
* 适合人群：
  + 数据从业人员
  + 大数据分析与开发的用户
* 要求: R/python基础，统计基础
* 讲师: 张家齐，李孟育，靳军
* 课时: 10天/60小时
* 价格: 10800/8800