（一）统计分析、数据仓库与可视化表达

1、 综述（大数据、人工智能、数据挖掘、机器学习：这些词的确切含义）
2、 假设检验：“小数据”时代是怎么玩的？
3、 “回归”是数据挖掘算法吗？
4、 度量、指标与维度
5、 星型模型与雪花模型
6、 下钻与上卷
7、 数据仓库的应用案例
8、 图表该怎么画才对？

（二）大数据相关技术综述

1、 hadoop：HDFS、Map-Reduce、Hbase、Hive、sqoop、pig、oozie 等
2、 spark：scala、spark-SQL、spark-Streaming 等
3、 搜索引擎：lucene（solr）、ES
4、 并发的机器学习工具：R-hadoop、spark-MLLIB、 spark-R、pyspark

（三）存储在 hbase 中的数据

1、 NoSQL（key-value）
2、 Hbase：安装
3、 行键与列簇
4、 如何利用 Hbase 的特点存储行业数据
5、 应用程序如何访问 Hbase 中的数据
6、 数据迁移工具：sqoop
7、 Hbase 的应用场景

（四）Hive：为 SQL 开发者留的活路

1、 Hive：安装（单用户与多用户）
2、 Hive：基本操作
3、 Hive：与典型的关系型数据库的区别
4、 存储业务数据时的注意点
5、 如果“想慢”，你还可以这样…（不恰当使用 hive 的案例介绍）
6、 Hive 的应用场景

（六）Spark 各组件的应用

1、 Hadoop 最大的特点是什么？
2、 Spark 概述与安装
3、 Scala：你可以一直“点”下去
4、 RDD：“映射”、“转换”解决一切
5、 spark-SQL
6、 spark-streaming
7、 spark-graphX
8、 spark-MLLIB
9、 应用场景

（七）机器学习-1

1、 数据挖掘、知识发现与机器学习
2、 工具：（早期）SPSS、SAS；
3、 目前流行的工具 R、Python 等
4、 决策树（熵、贪心法、连续的和离散的）
5、 聚类（k-means、k-medoid）
6、 监督学习、无监督学习的差异
7、 机器学习性能评价指标

（八）机器学习-2

1、 KNN
2、 关联规则（频繁项集、Apriori、支持度、置信度、提升度）
3、 神经网络（神经元、激励函数、前馈神经网络的 BP 算法）
4、 SVM（最大间隔、核函数、多分类的支持向量机）

（九）机器学习-3

1、 “概率派”与“贝叶斯派”
2、 朴素贝叶斯模型（皮马印第安人患糖尿病风险预测）
3、 极大似然估计与 EM 算法
4、 HMM（三个基本问题：评估、解码、学习）

（十）机器学习-4

1、 遗传算法 （交叉、选择、变异，“同宿舍”问题）
2、 无监督学习
3、 集成学习（adaboost、RF）
4、 强化学习

（十一）深度学习-1

1、 连接主义的兴衰
2、 地形要更陡：改进的目标函数
3、 0.9 的 100 次方等于几？克服梯度消散的方法（改进的激励函数、BN）
4、 利用“惯性”下山：改进的优化算法（Adagrad、RMSprop、Adam）
5、 防止“大锅饭”：dropout
6、 记忆的关键是“合理的忘记”：weight decay

（十二）深度学习-2

1、 让 AI 理解图像：典型 CNN
2、 各种 CNN
3、 让 AI 理解语言：RNN 与 LSTM、GRU
4、 左右互搏术：GAN
5、 电子游戏的新玩法：DQN