第1届全国高校大数据教学研讨会（BDTS2017）
大会特邀报告

2017年5月12日至13日，第1届全国高校大数据教学研讨会（BDTS2017）在厦门大学科艺中心音乐厅隆重举行。本届研讨会由教育部高等学校计算机类专业教育指导委员会主办，厦门大学、厦门理工学院、贵州师范大学、人民邮电出版社联合承办，旨在搭建专业的大数据教学交流平台，汇聚全国高校大数据教学精英力量，共同探讨大数据专业和课程体系建设，为加快推进全国高校大数据教学发展贡献力量。来自全国300多所院校的400余名教师参加了本次研讨会。

厦门大学谭绍滨校长助理、人民邮电出版社教育中心营销部肖稳副主任，北京大学、中国科学院、厦门大学、华东师范大学、同济大学等重点院校的6位大数据教学知名专家，以及来自国内知名大数据企业的3名业界专家出席会议并做特邀大会报告。厦门大学林子雨助理教授主持会议。

更多内容请访问大会官网：http://dblab.xmu.edu.cn/post/bigdata2017/

同济大学 王伟 副教授 在做大会特邀报告

如何建设大数据与数据科学通识实践类课程？

王 伟

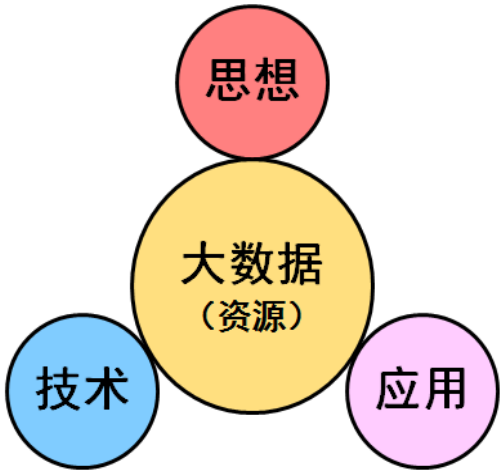
同济大学，计算机科学与技术

2017年5月于厦门

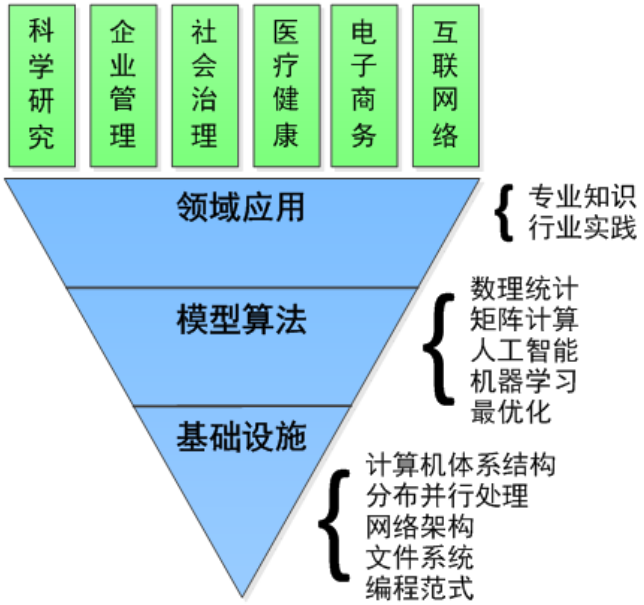
Outline

- 对大数据与数据科学的思考
- 导论类课程建设与教学实践
- 数据科学与大数据实践平台
- 总结：未来的挑战

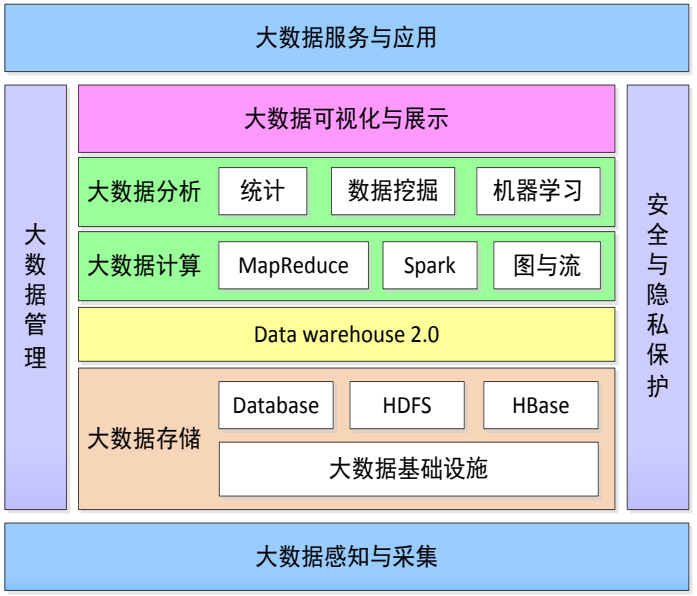
大数据的冲击



思维模式



课程教育



技术体系

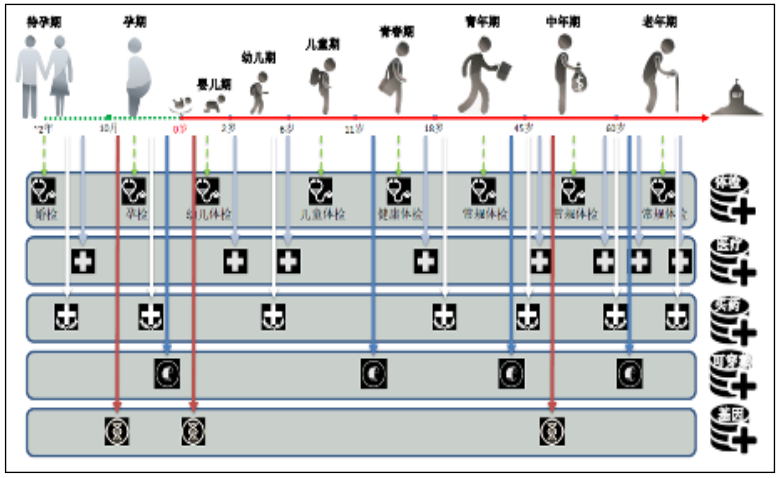
时代的呼唤

国家

企业

机器

人类



教育的变革

COMPUTATIONAL THINKING

在马来西亚召开的2016国际工程联盟会议上，中国被接纳为《华盛顿协议》正式签约成员。

教　育　部　文　件

教高[2017]2号

教育部关于公布2016年度普通高等学校本科
专业备案和审批结果的通知

教育部公布第二批”数据科学与大数据技术专业”获批高校名单

大数据教育改革

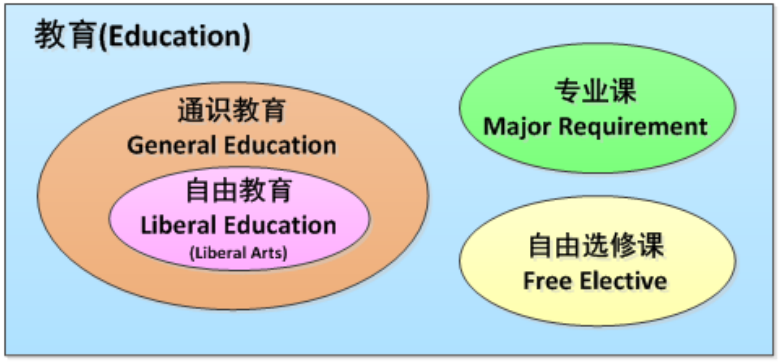
- 2016年大数据与数据科学的教育改革
  - 《数据科学与大数据技术》本科专业（专业代码：080910T）
  - 《大数据技术与应用》高职专业（专业代码：610215）
- 2017年3月，教育部公布第二批“大数据专业”获批高校，两批共35所
  - 第一批：北京大学、对外经济贸易大学、中南大学3所
  - 第二批：中国人民大学、复旦大学等32所
- 全国高校大数据教育联盟、数据中国“百校工程”等等

思考1：数据科学的通识属性

通识教育的属性

- 通识教育，既是为常识服务的教育。
- 通识教育的精神支柱和思想来源： Liberal arts（自由博雅）
  - 旨在培养一流的头脑、一流的心灵；
  - 一种间接的思维训练，传授的是“软实力”；
  - 关注审辨式思维、交流能力和解决问题的能力。

中国通识教育的意义：为了解决中国社会的实际问题，发展具有中国特色的通识教育。



谈谈我心目中的Liberal Arts, 嘉数汇, 2017-02-05.

数据科学作为一门通识课

- 数据科学有利于培养信息时代一个健全的人；
- 数据科学有利于培养跨学科视野；
- 数据科学有利于培养表达自我所必备的技能；
- 数据科学有利于培养个人的科学思维方式；
- 数据科学有利于围绕数据开展实践。

A liberal arts background gives data science its soul.

>> 专　点

大数据适合通识教育实践

■ 王　伟

国内外的许多著名大学都非常重视本科教育，而本科教育的风格是通识教育。

什么是通识教育？

所谓通识教育，旨在为学生带来完整的知识结构，养成触类旁通的通用智慧。在耶鲁，本科生院和研究生院之外，医学院、法学院、商学院等都叫作职业学院。

通识教育起源于古希腊，指公民应当具备的知识与能力，要参与公共生活才是完整的人，也就是说，要会打仗，会辩论，能打官司，在法庭上为自己辩护，所以得懂哲学、逻辑学、语言、演讲、音乐、天文、数学，等等。

现代社会分工细密，古希腊的通识教育当然只能从精神上去效仿，现代公民应当具备的知识，一般来说就是部分自然科学和人文学科。从大学毕业，应当具备现代公民的完整智识能力，而不仅仅是被打造成用于解决特定问题的工具。

设施，学生们在此食宿，娱乐，college不管教学，主要是一种生活的组织，旨在使学生们增加交流，培养集体荣誉感，或者说，在现代社会重现一种雅典学园之风。

所有老师，不管多么大牌，都有为本科生开课之责任，并且课上非常重视和学生交流，课外也会推荐很多经典著作。阅读经典著作是通识课的一大特点。因此学生的任务也相对比较繁重。

再谈大数据课程与通识实践, 嘉数汇, 2016-11-08.

Liberal Arts理念向数据科学注入“博雅”之心

- 问正确问题的能力
- 科学方法观
- 团队协作的精神
- 沟通交流的能力
- 三观正确的决策

Ultimately lead to a better understanding of the natural world.

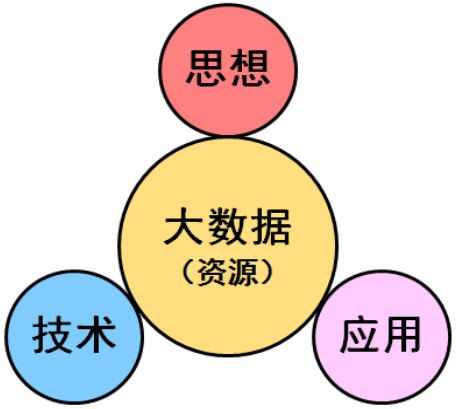
向“数据科学”注入“博雅”之心, 嘉数汇, 2017-03-02.

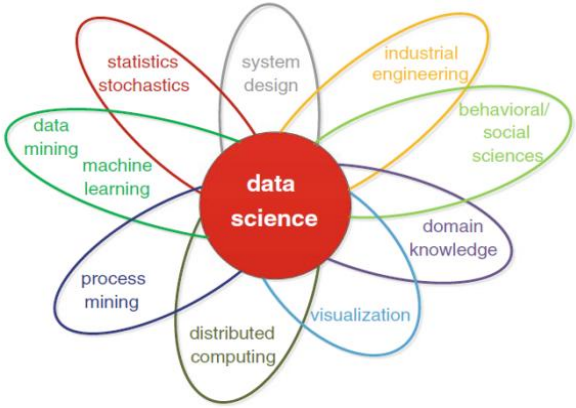
思考2：数据科学的知识体系

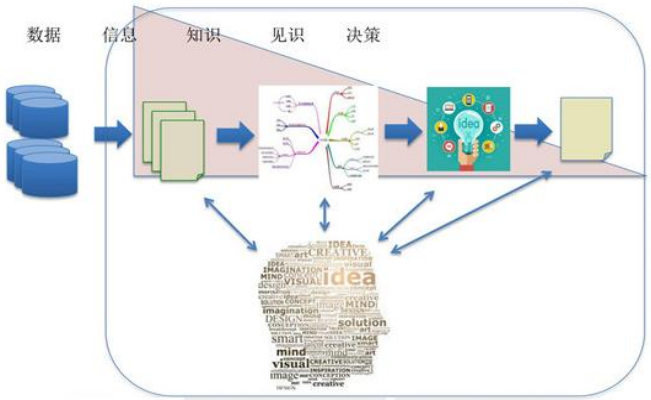
数据科学：从大数据到行动

数据科学 ≈ 思维 + 计算机科学 + 统计 + 应用

- 首先，建立数据思维方式，学习怎样利用数据；
- 其次，应该了解数据清理、集成、探索等相关技术；
- 最后，洞见和商业意识也至关重要。



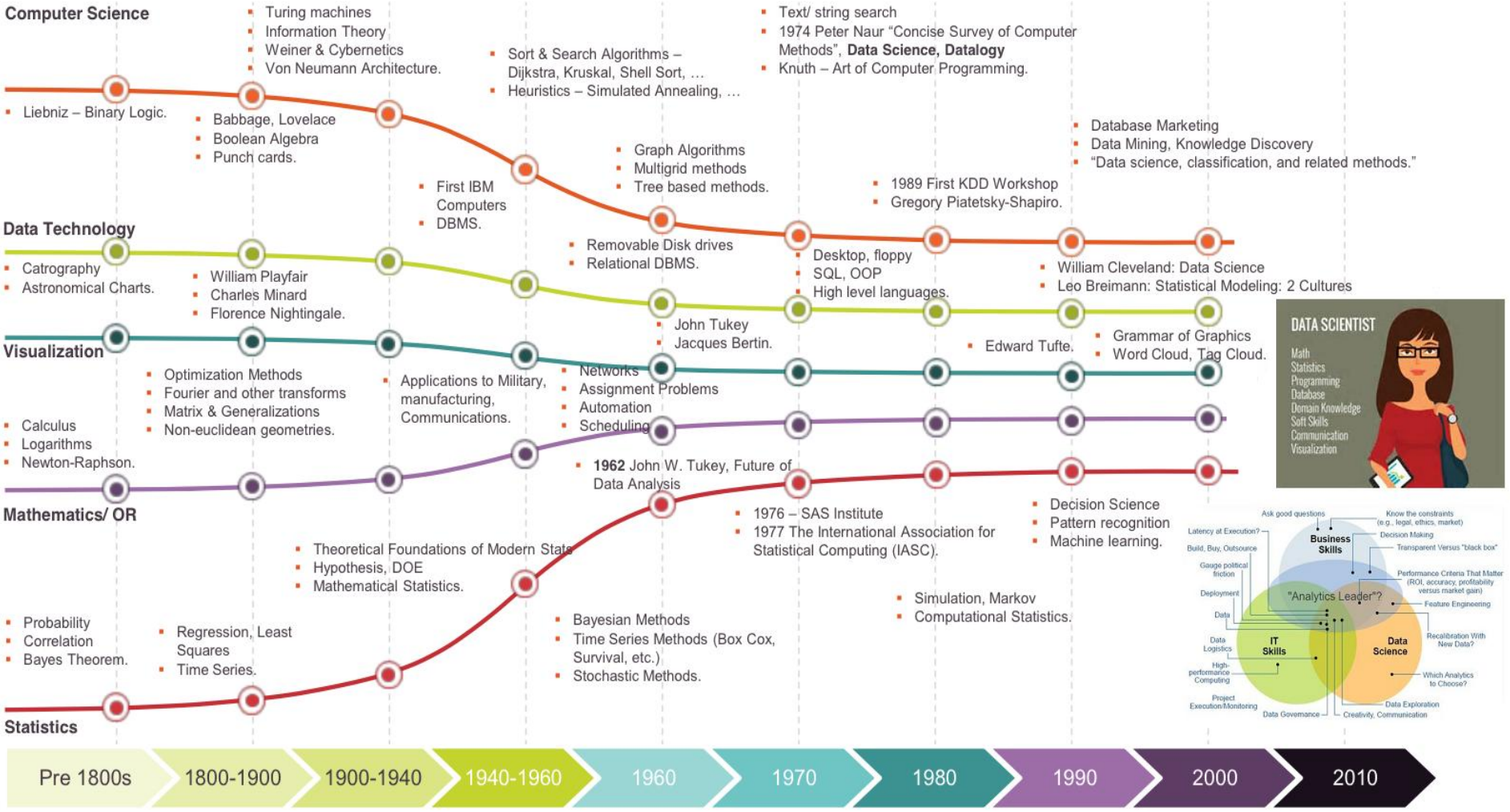




数据科学的三大学科支柱

- 数据科学天生就是一个交叉学科，和数据科学最为密切的一些学科包括：计算机科学与技术、数学、统计学、信息管理、情报学等。数据科学的三大主要支柱为：
  - Datalogy (数据学)：对应数据管理 (Data management)
  - Analytics (分析学)：对应统计方法 (Statistical method)
  - Algorithmics (算法学)：对应算法方法 (Algorithmic method)

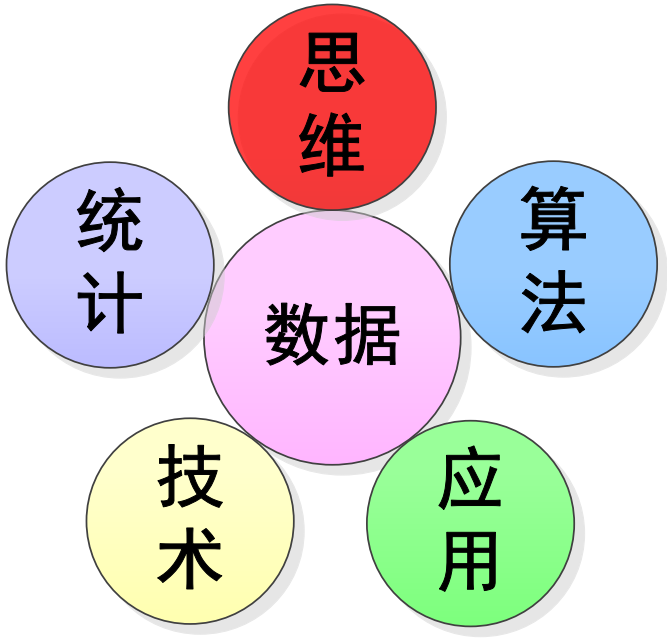
数据科学发展历程



Created by *Mamatha Upadhyaya*, modified by *Mr.Wang*

数据科学的五大要素

- A-SATA模型
  - 分析思维 (Analytical Thinking)
  - 统计模型 (Statistical Model)
  - 算法计算 (Algorithmic Computing)
  - 数据技术 (Data Technology)
  - 综合应用 (Application)



数据科学的核心知识点

- 分析思维 (Analytical thinking)：包括计算思维 (Computational thinking)和统计思维 (Statistical thinking)；
- 数学基础：微积分、线性代数、概率统计、离散数学等；
- 数据建模与评估：统计模型、实验设计、模型评估等；
- 算法实现：问题求解能力和算法设计；
- 数据管理：设计数据的整个生命周期，包括感知、存储、计算、分析、可视化等；
- 知识转化：沟通交流，道德规范等。

思考3：数据科学的实践方法

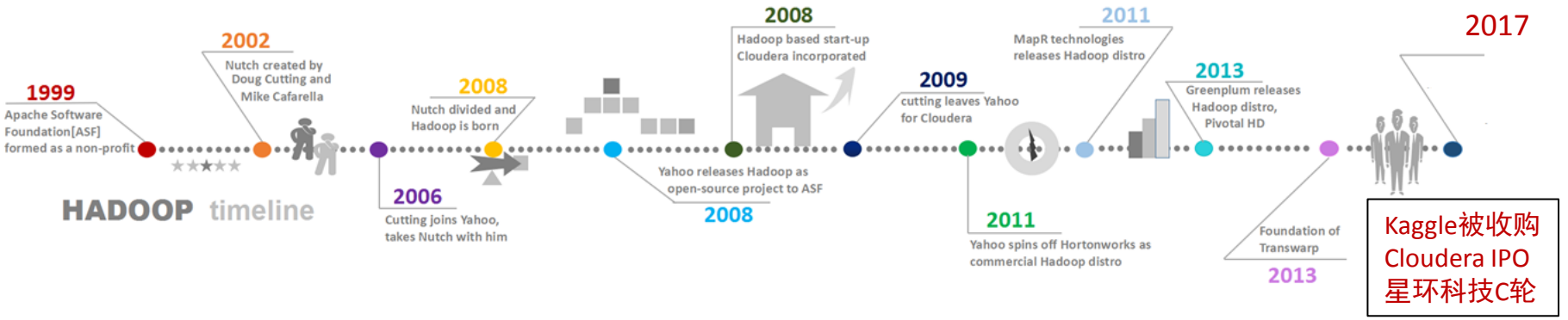
信息化、大数据与教育

习近平：

“当今世界，科技进步日新月异，互联网、云计算、大数据等现代信息技术深刻改变着人类的思维、生产、生活、学习方式，深刻展示了世界发展的前景。”

“因应信息技术的发展，推动教育变革和创新，构建网络化、数字化、个性化、终身化的教育体系，建设“人人皆学、处处能学、时时可学”的学习型社会，培养大批创新人才，是人类共同面临的重大课题。”

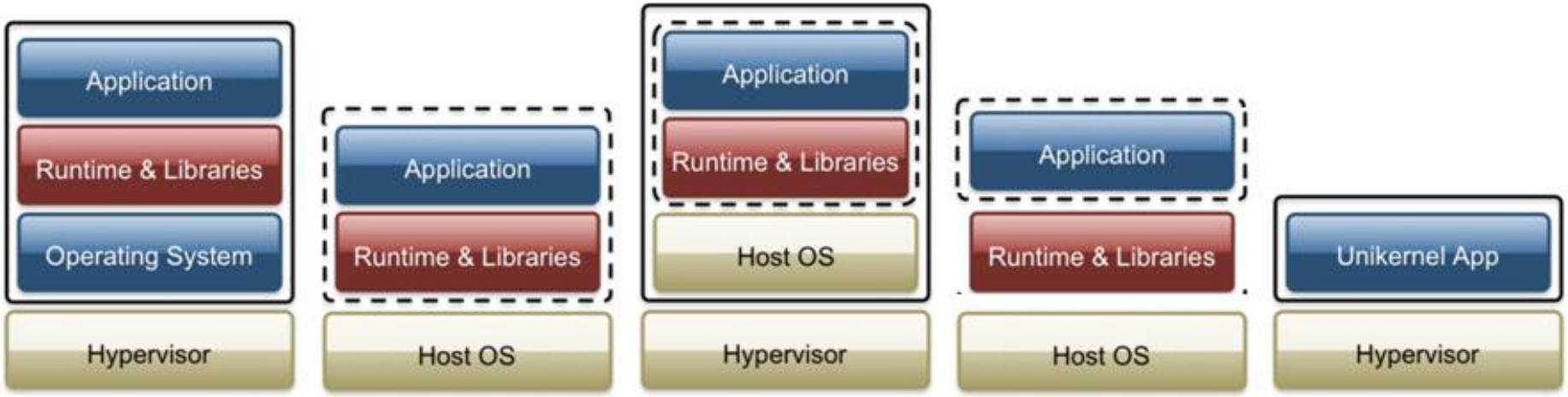
大数据技术的演变



ETL 数据装载工具	Workflow 工作流开发工具	数据质量 管理工具	可视化 报表工具	机器学习 建模工具	统计挖掘 开发工具	资源 管理工具	分析管理工具
SQL批处理 Batch Processing	交互式分析 OLAP Analysis	实时数据库 OLTP Transactional Processing	数据挖掘 机器学习 算法库 / 框架 Machine Learning	深度学习 Deep Learning	图分析引擎 Graph Analysis	流处理引擎 Streaming Processing	应用级引擎
批处理框架 Map/Reduce2， Tez		高性能处理框架 Spark		向量处理框架 TensorFlow			通用计算引擎
短时任务资源管理框架 YARN		长时任务资源管理框架 Mesos		资源隔离 / 调度 / 管理框架 Kubernetes			资源管理框架
分布式文件系统 HDFS	分布式大表 HBase	搜索引擎 Elastic Search	分布式缓存 Redis	消息队列 Kafka	分布式协作服务 Zookeeper		分布式存储引擎

云计算技术的演变


1979
chroot
一套UNIX操作系统，为每个进程提供一套隔离化磁盘空间。
2000
FreeBSD Jails
与chroot定位类似，包含其它进程沙箱机制对资源进行隔离。
2001
Linux VServer
每个分区被称为一套安全上下文，其中的虚拟化系统被称为一套虚拟私有服务器。
2004
Solaris 容器
将系统资源控制与由分区提供的边界加以结合，各分区在单一操作系统实例之内。
2005
OpenVZ
安装有补丁的Linux内核实现虚拟化、隔离能力、资源管理以及检查点交付。
2006
Process容 器
对一整套进程集合中的资源使用量加以限制、分配与隔离。
2007
Control Groups
即谷歌实现的 cgroups，于 2007 年被添加至 Linux 内核当中。
2008
LXC
通过 liblxc 库交付，提供可与 Python3、Python2、Lua、Go、Ruby 以及 Haskell 等语言相对接的 API。
2011
Warden
不与 Linux 紧密耦合，以后台进程方式运行并提供 API 以实现容器管理。
2013
Docker
最具人气且应用最为广泛的容器管理系统，引入了一整套与容器管理相关的生态系统。
2014
Rocket
由 CoreOS 开发，专门用于解决部分 Docker 当中存在的缺陷。
2016
Windows容器
Docker 能够以原生方式在 Windows 平台上运行容器。


Application
Runtime & Libraries
Operating System
Hypervisor
虚拟机
Application
Runtime & Libraries
Host OS
容器
Application
Runtime & Libraries
Host OS
Hypervisor
在虚拟机中运行的容器
Application
Runtime & Libraries
Host OS
Serverless
Unikernel App
Hypervisor
Unikernel

数据科学与大数据实践平台

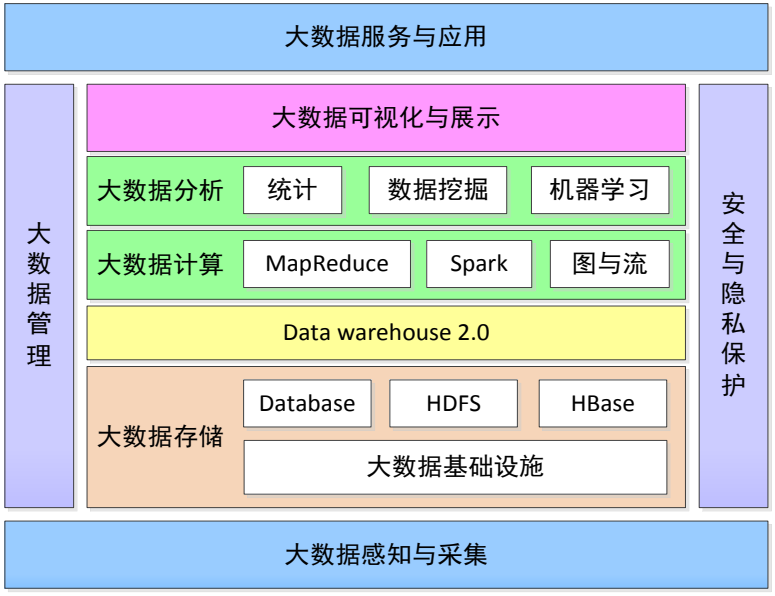
- 围绕大数据和数据科学提供全在线课程教学、习题练习、动手实践、视化案例库、数据竞赛等服务。
- 通过校企合作，对接广泛的大数据基础软件；
- 通过自主研发的云件技术，在线学习使用任意数据分析工具。
- 形成一个学、做、测、竞、评、奖的大数据在线学习闭环。

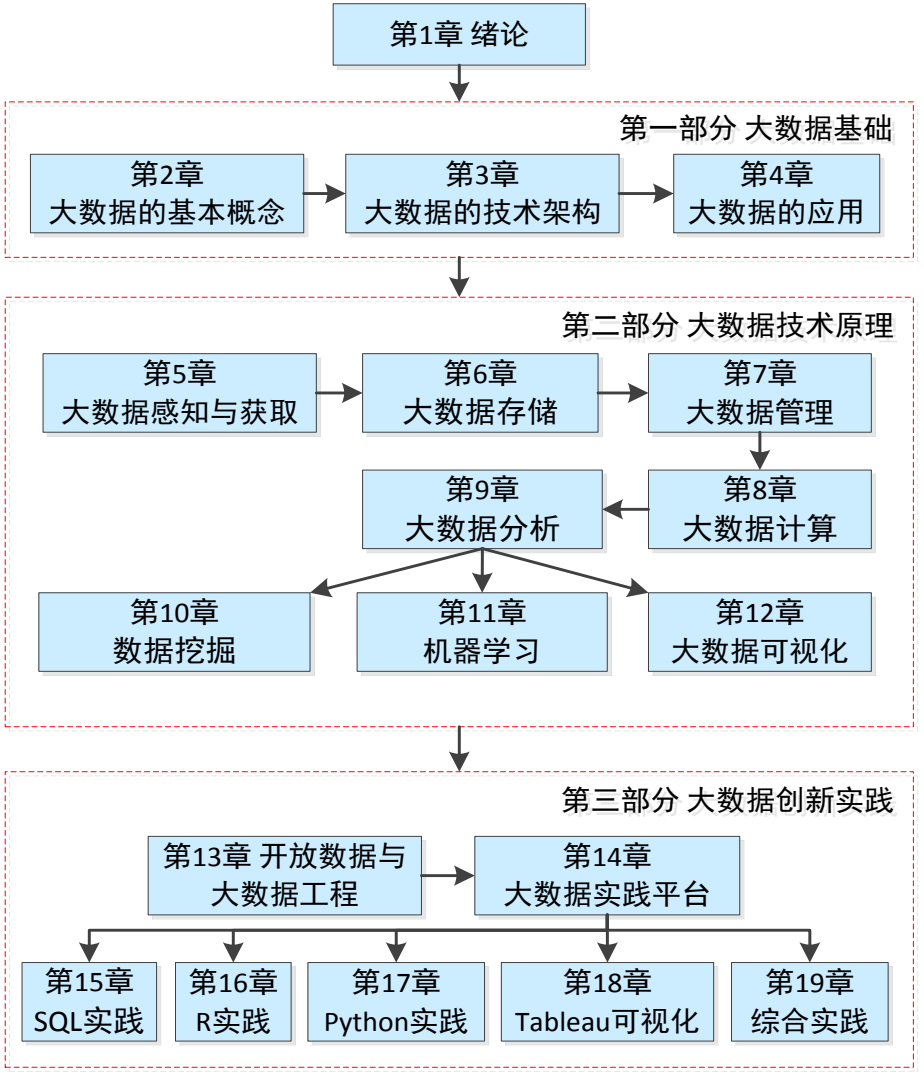
导论类课程建设与教学实践

2016年

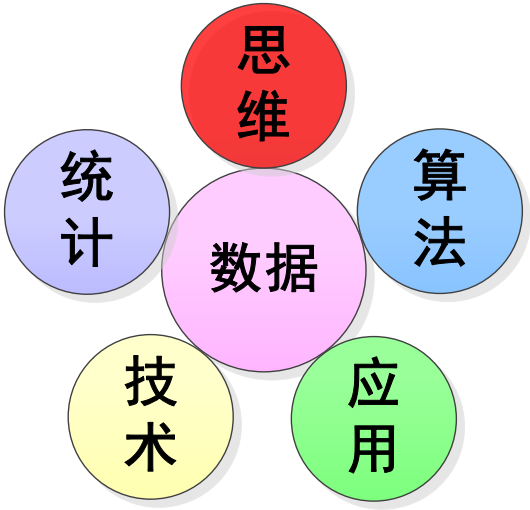
2017年

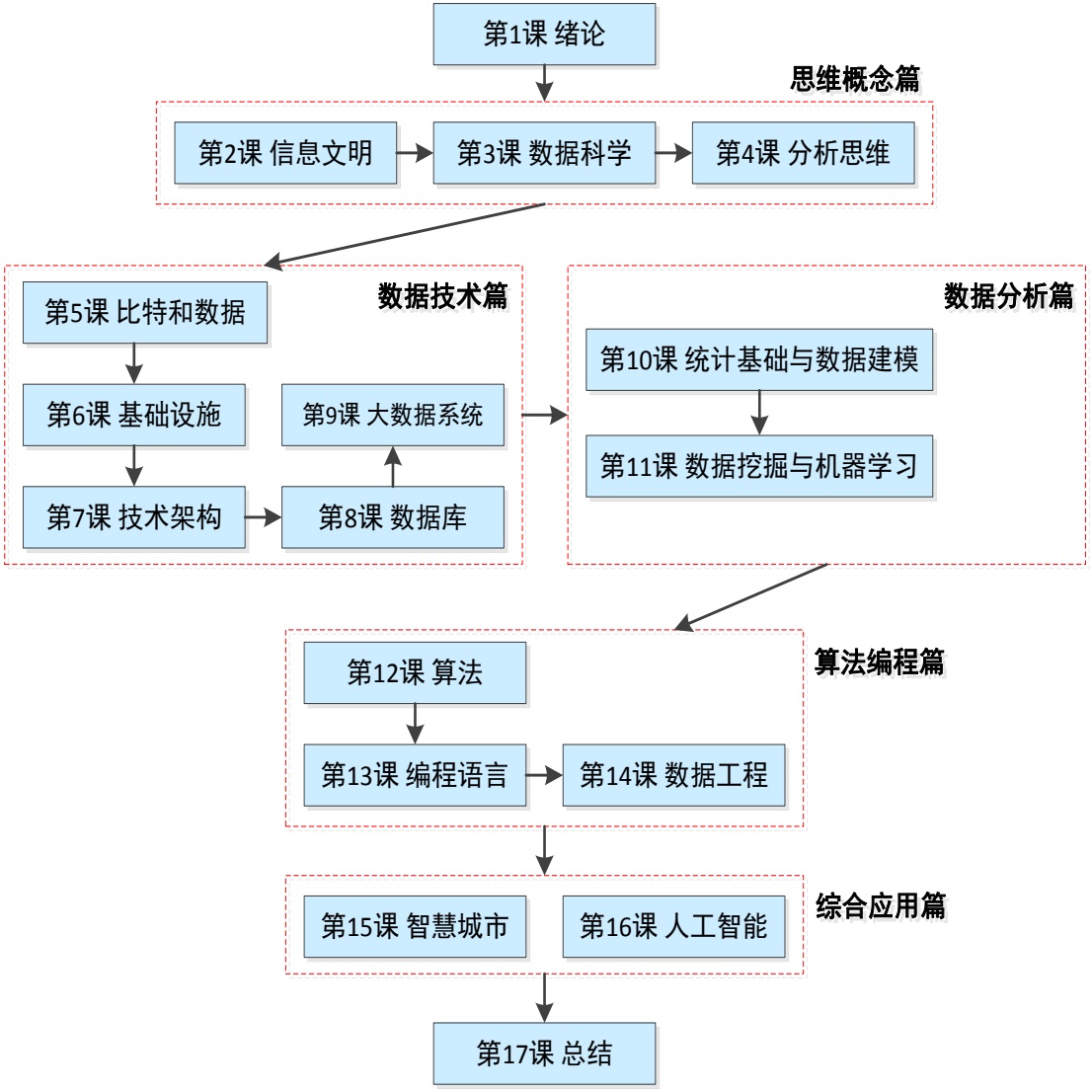
《大数据原理与实践》课程总览





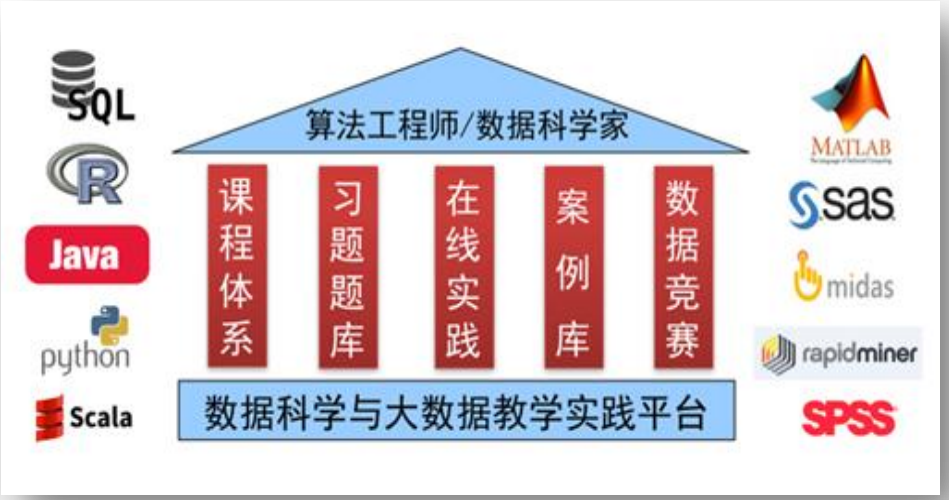
《数据科学通识导论》课程总览





“自由博雅”的16个字实践

- 建立对话；激发思辨；
- 协作交流；动手实践。



1. 建立对话、激发思辨

课件

文章

互动

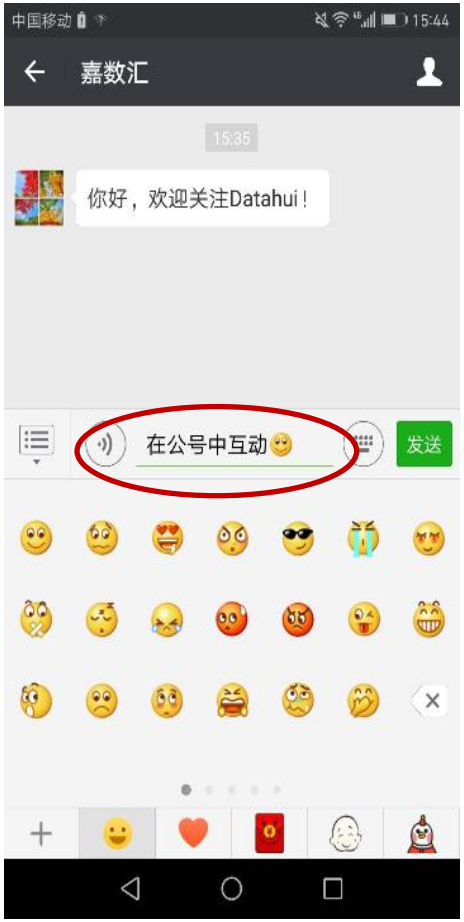
嘉数汇

微信公众号每周更新模式

- 周二：发布本周课件初稿，授课；
- 周三、周四：互动、点评与问答；
- 周五：发布最终版课件以及相应文本注释；
- 周末：课外阅读文章。

如何回答与互动

在公号中输入：姓名+学号



在公号中互动

在文章中留言

互动交流

《数据科学通识导论》互动交流 - 201701期

2017年第01期互动交流。

《数据科学通识导论》互动交流 - 201703期

每周我们会选取和汇总几条比较有代表性的学员问答进行反馈，希望通过这种方式达到交流的目的，启发大家进一步的思考。

《数据科学通识导论》互动交流 - 201704期

《数据科学通识导论》互动交流 - 201705

编者按：每周我们会选取和汇总几条比较有代表性的学员问答进行反馈，希望通过这种方式达到交流的目的，启发大家进一步的思考。

《数据科学通识导论》互动交流 - 201705-06期

编者按：每周我们会选取和汇总几条比较有代表性的学员问答进行反馈，希望通过这种方式达到交流的目的，启发大家进一步的思考。

《数据科学通识导论》互动交流 - 201708期

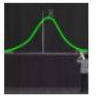
编者按：每周我们会选取和汇总几条比较有代表性的学员问答进行反馈，希望通过这种方式达到交流的目的，启发大家进一步的思考。

《数据科学通识导论》互动交流 - 201709期

编者按：每周我们会选取和汇总几条比较有代表性的学员问答进行反馈，希望通过这种方式达到交流的目的，启发大家进一步的思考。

更新于 星期一 12:03

钟形曲线的陷阱：我们能够冲破阶层固化的束缚吗？



建立对话、激发思辨

“ZZZZZB”:

“数据科学”
分析，就其本
直观地展现出
边可见的例子
到积累，等积累足够多的

张翔 1452229 (Ice) 昨天 23:16

第一周讨论内容: 1.通识课，我认为就是以科普为目的的课程，用于拓宽知识
面
听
数

“suyi”:

我认为通识教育是培养一个人如何理解世界、分析世界和表达自我的能力，培养一个人在满足
温饱之后想要达到更高
识教育是为了达到不断学

张云笛+
1，通识
业，或
一起上
课我觉
五六节
成一些
参与度，
修课，
些概念
时候讲
个课程。
老师就
会很好。 然后“数据科学“嘛，还是希望老师能够讲一些关于数据的算法之类
的。 ps.老师，我想问一问，我现在有一点r和python的编程基础，但是直接接
触kaggle又会有些困难，这个时候，我应该怎么学习啊？

朱涵林+1352356 (朱涵林) 星期四 10:10

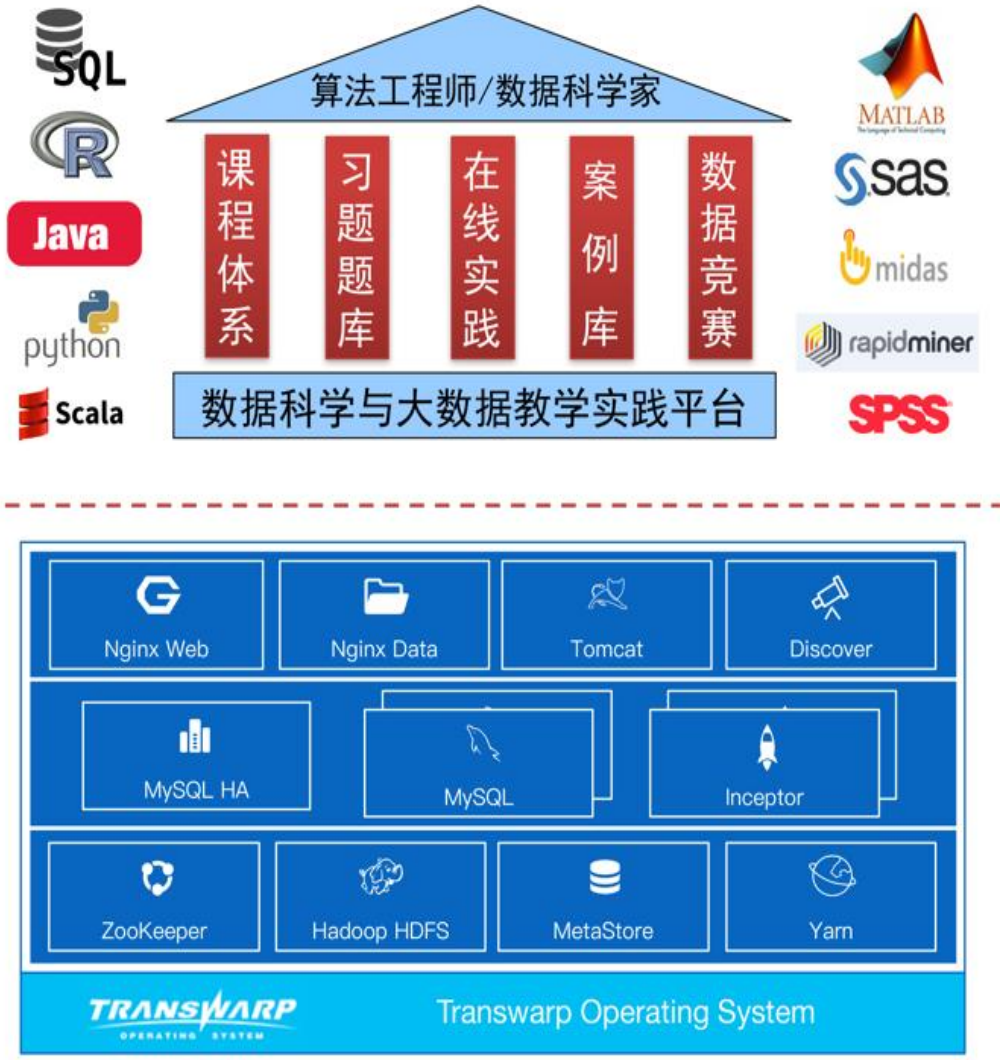
数据科学:以数据为中心的科学。其涵盖面也非常广，存储,索引,科学计算,可视化等方面都是数据科学的范畴。 学习数据科学，首先应该知道数据的分类，结构化数据，半结构化数据，非结构化数据。再根据数据的特点，找到适合它的存储方式，是传统关系数据库还是分布式存储系统。针对不同的系统，制定一定的策略进行索引和信息检索。为了从数据中得到我们所需的信息，需要我们对数据进行清洗,使用机器学习的方法对其进行分类，特征提取，或是建立相关模型进行预测，为了有更高的效率也许需要分布式计算。当有了数据及其分析结果，为了将其展示出来，向别人更好的表达自己的成果，数据可视化也必不可少。。。。。数据科学四个字，其实是计算机领域和统计学领域众多学科相互协作的集合体。想要好好学习，既要有概念性的认识和总体的认识，比如数据的分类，什么样的数据适合什么样的策略，针对现有数据和算法我们能干什么，不能干什么，不同架构不同模块有什么特点和优劣等。有了对数据科学总体的认识后，可以挑选自己感兴趣的方向进行进一步的学习和研究，在算法效率，准确度等方面进行严谨的思索和探索。在“通”和“专”上不断的往复学习，才能学好数据科学。

2. 协作交流、动手实践


数据科学与大数据教学实践平台
在这里拥抱大数据
教学
告别传统的“古板”教学方法，创造全新模式。在这里会发布与课堂内容同步的课后教学任务，学生们可要及时查看哦！
实践
我们提供各种数据，利用课堂上学到的数据分析的方法，在平台的在线编辑器中亲自实践，免去各种软件工具的安装。
竞赛
有竞争才有进步，我们会不定期的举行各种数据集下的大数据竞赛，让大家通过比赛发现蕴藏在数据之后的真相。
案例
所谓“教学相长”，我们会定期更新有关大数据的各种案例，看看别人的思路对你是否有所启发。通过学习案例实现进步。

数据科学与大数据实践平台

- 围绕大数据和数据科学提供课程教学、习题练习、动手实践、视化案例库、数据竞赛等服务。
- 通过校企合作，对接广泛的大数据基础软件；
- 通过自主研发的云件技术，在线学习使用任意数据分析工具。
- 形成一个学、做、测、竞、评、奖的大数据在线学习闭环。

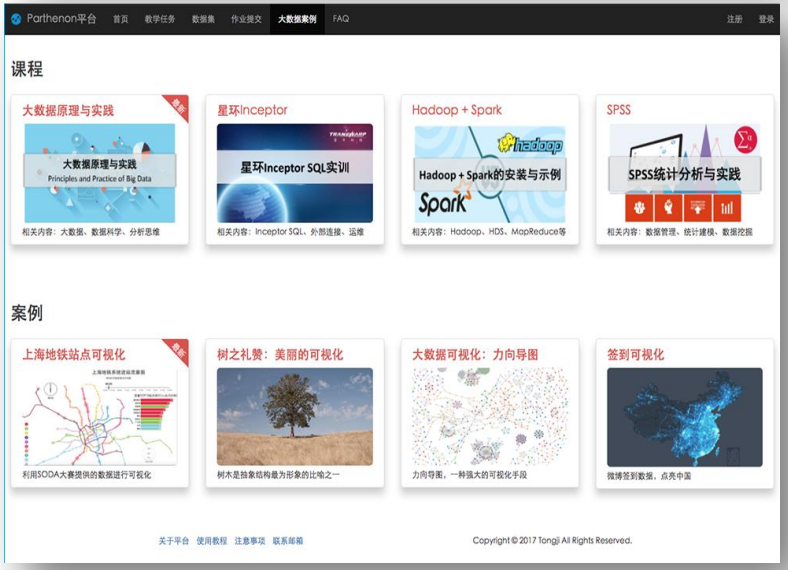


课程

- 通识基础
- SQL实践
- 平台搭建
- 数据分析
- 工具使用

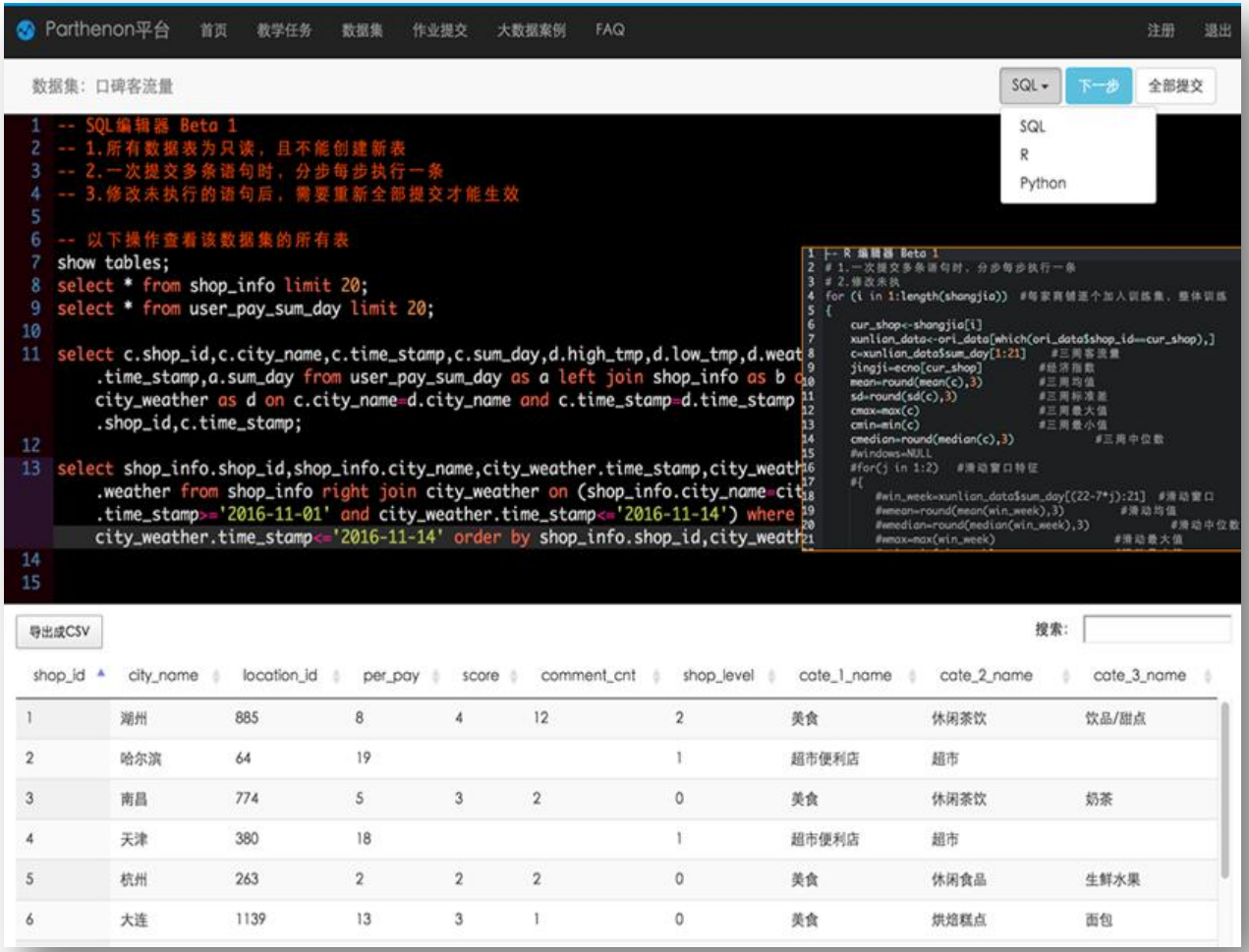
案例

- 经典案例
- 学生作品
- 创新创意



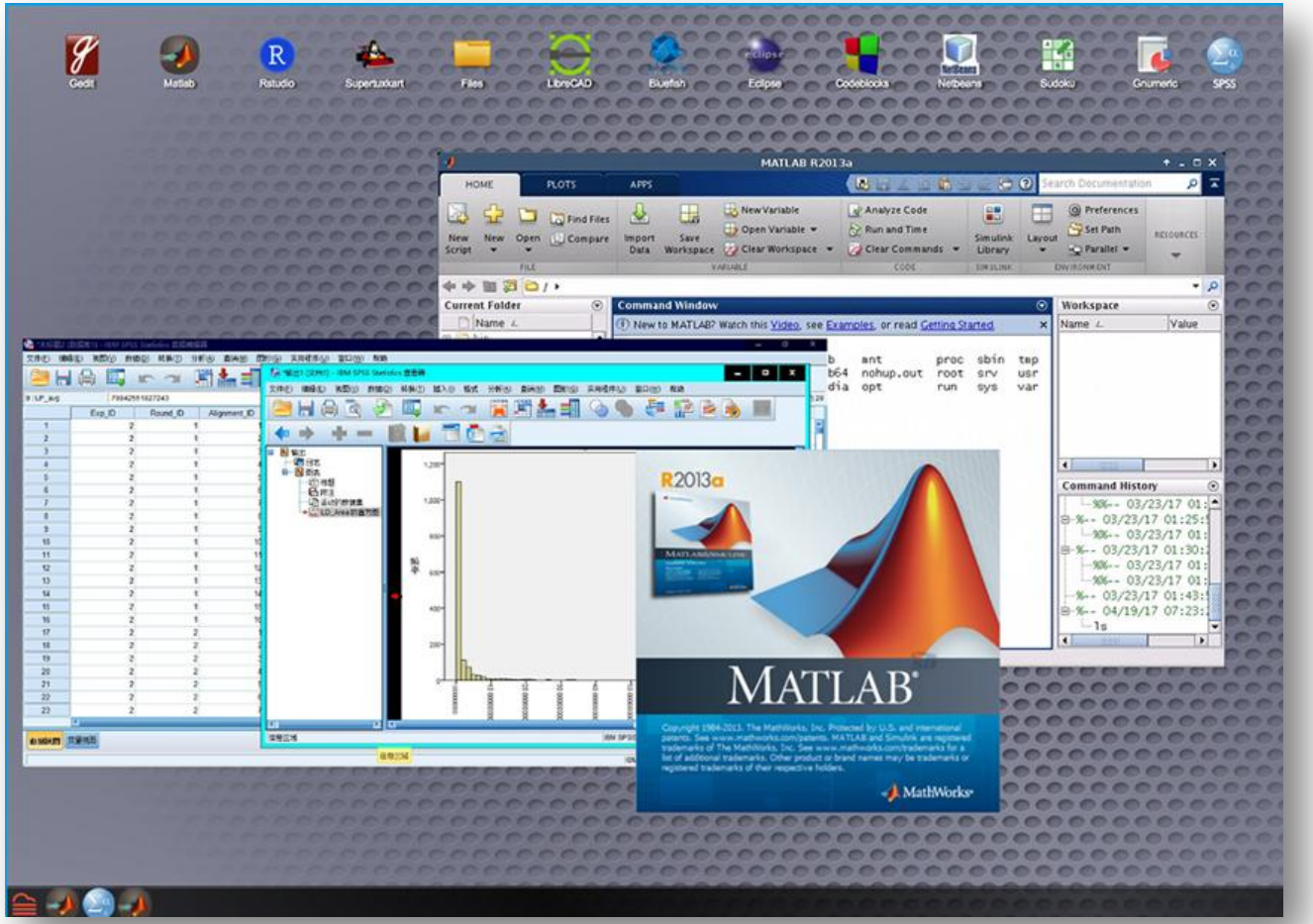
在线实践

- 课程联动
- 方便直观
- 多语言支持
- 学练测闭环
- 在线竞赛
- 数据马拉松



云件工具

- 简洁方便
- 桌面工具
- 数据工具
- 存储服务
- 在线指导
- 团队协作



Dong Guo, Wei Wang*, et al., Cloudware: An Emerging Software Paradigm for Cloud Computing, In Proceedings of the Internetware 2016, Beijing, China, September 18, 2016.

技术架构

Pantheon - Data Science and Big Data Education and Practice Platform

Hadoop/Spark ecosys environment	Tupllace (SQL)	OperaR (R)	Pyramithon (Python)	Math founddation	Edenbase (Data)	DaVinci (Case)	Renaissance (Cloudware)	Datathon
	Big IDE				Big work book		Colosseum	

PantheonOS

Transwarp OS

Inceptor	Discover	Hyperbase	Stream
Midas	Hue	Zeppelin	Oozie

Cloudware OS

Rancher	Mensa	Pualsar	Ara	Pysix
Prism	Charon	Cygnus	Norma	Carina

容器服务（Container Services）

用户管理	调度与编排	性能监控	运营管理	性能调优

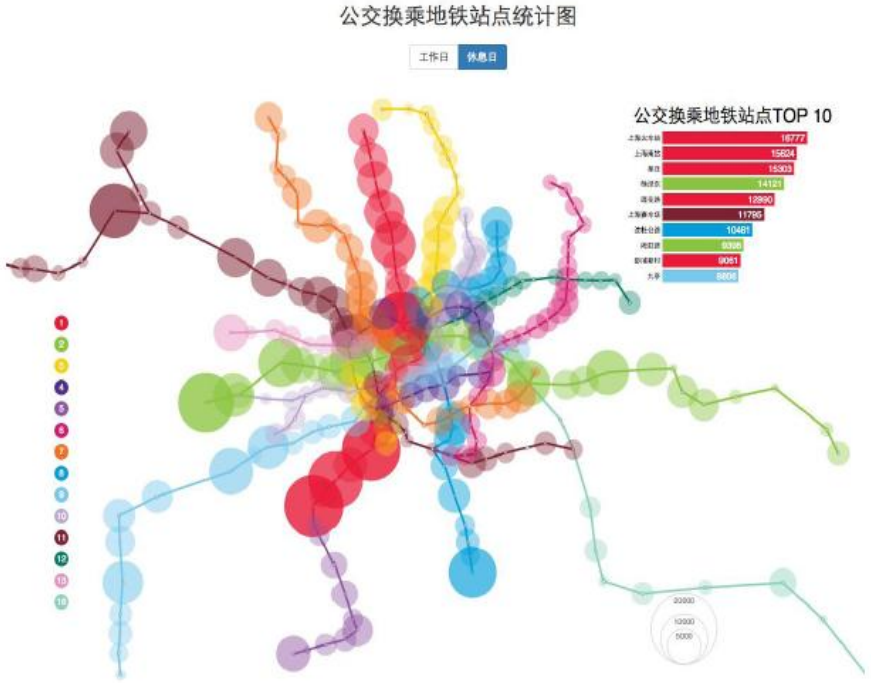
基础设施服务（Infrastructure Services）

存储 Storage	计算 Computing	网络 Networking	安全 Security	负载均衡 LB

作品展示：上海地铁系统进站流量图

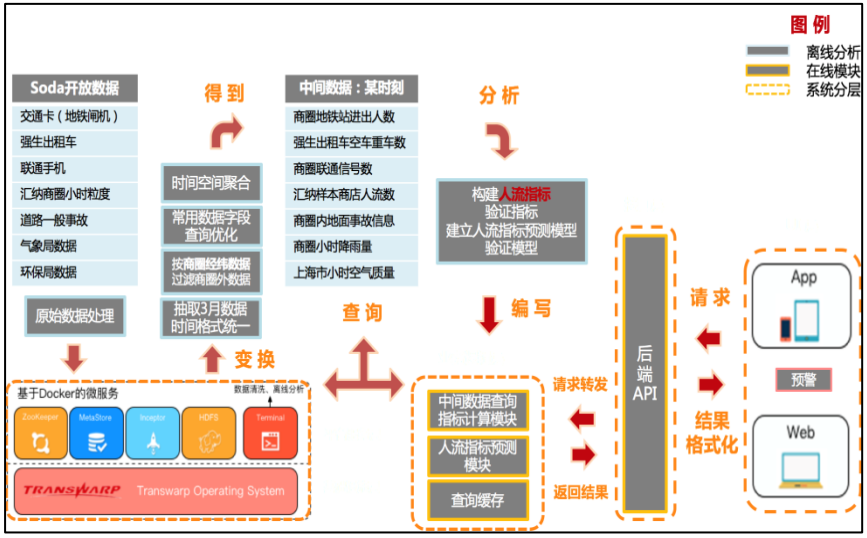
工作日

休息日



SODA 2015作品

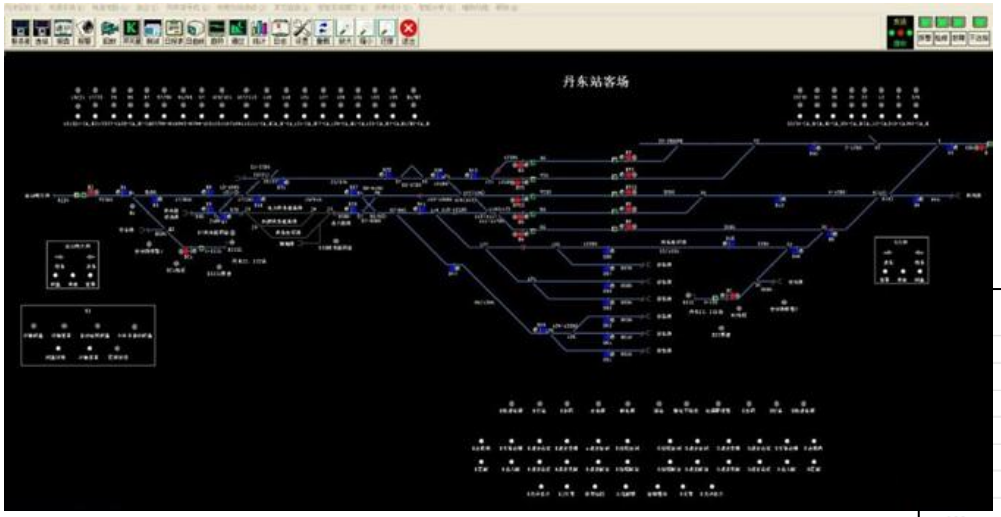
作品展示：基于人流指数预测的商圈公共安全预警系统

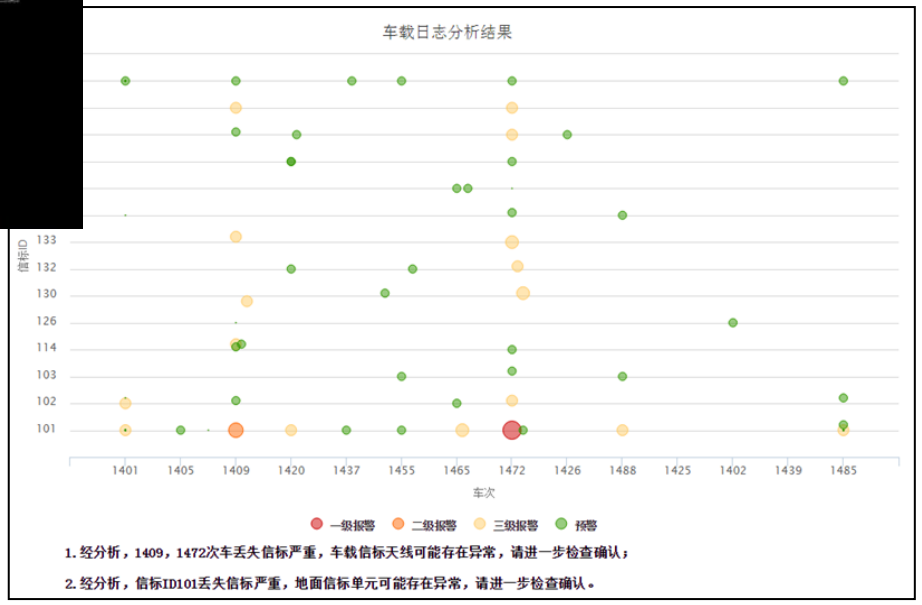


基于人流指数预测的商圈公共安全预警系统, 嘉数汇, 2016-12-06.

SODA 2016作品

作品展示：轨道交通运维大数据分析



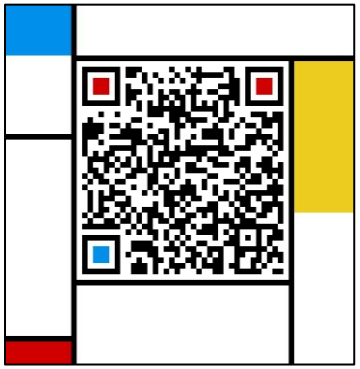


2016课程设计作品

总结：未来的挑战

- 大数据与数据科学知识体系的沉淀
- 通识类课程的改革（取代计算机基础？）
- 新工科背景下大数据与数据科学实践平台
- 真正做到以学生为本

Thanks!

课程公众号   个人微信

一个好的教育，是一个灵魂对另一个灵魂的呼唤；
一门好的课程，是一个生命与另一个生命的碰撞。