



数据科学与工程导论

Introduction to Data Science and Engineering

数据科学与工程导论

—— 数据与计算之美



王伟

wwang@dase.ecnu.edu.cn

华东师范大学



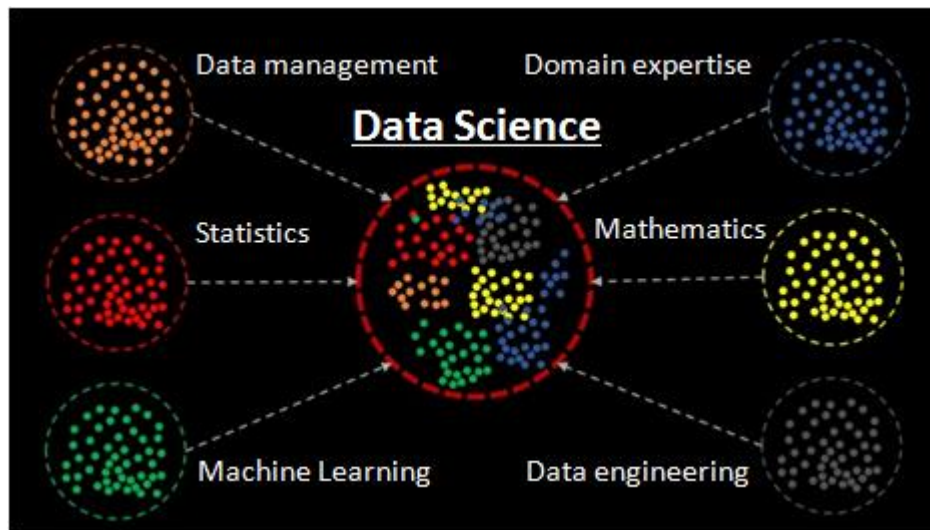
Bio

- **王伟**，华东师范大学数据学院，研究员
- ACM/IEEE会员、CCF高级会员，CCF教育专委会委员、CCF大数据专委会委员、开源社理事
- University of Florida, Visiting Research Scholar
- University of Wisconsin- Madison, Honorary Fellow, USA
- 研究方向：容器云、开源数字社会学、大规模在线学习系统
- E-mail: wwang@dase.ecnu.edu.cn

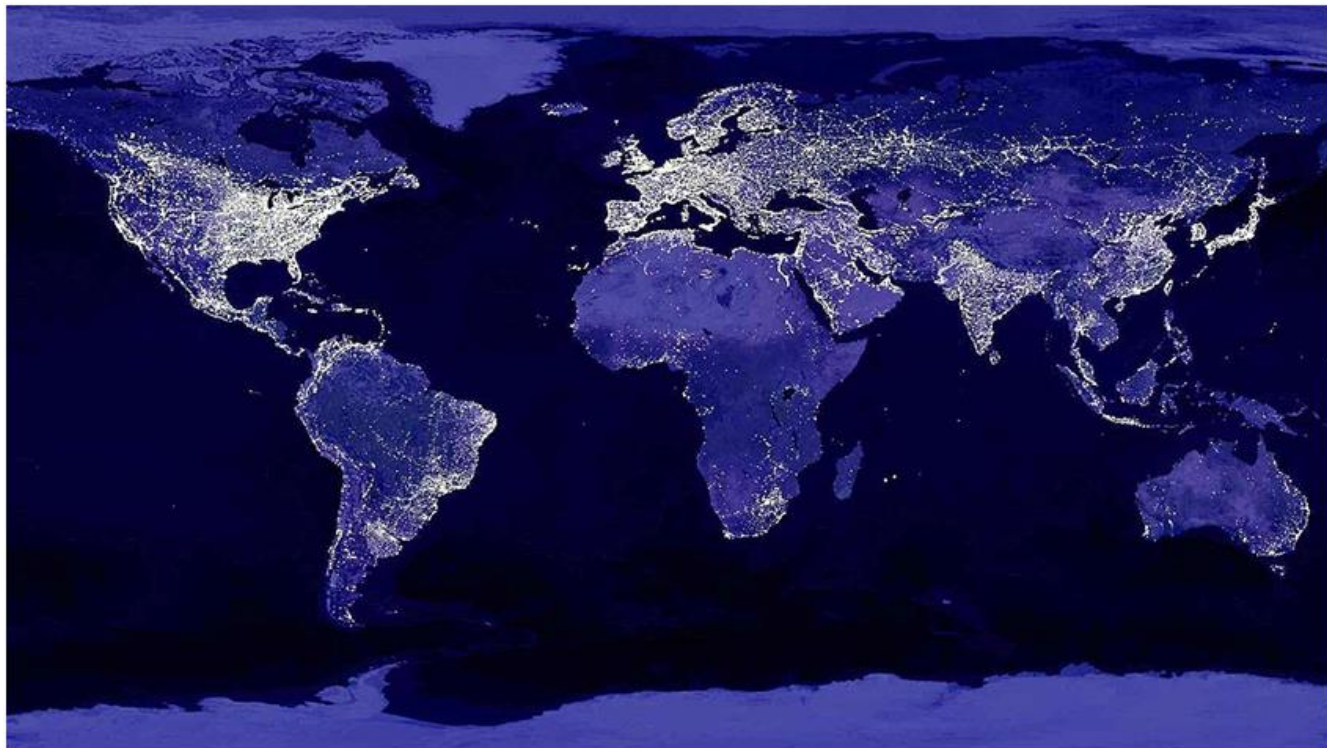


Outline

- 课程背景
- 课程简介
- 课程模式
- 知识体系



开篇实例：NASA从太空中拍摄城市夜间亮度



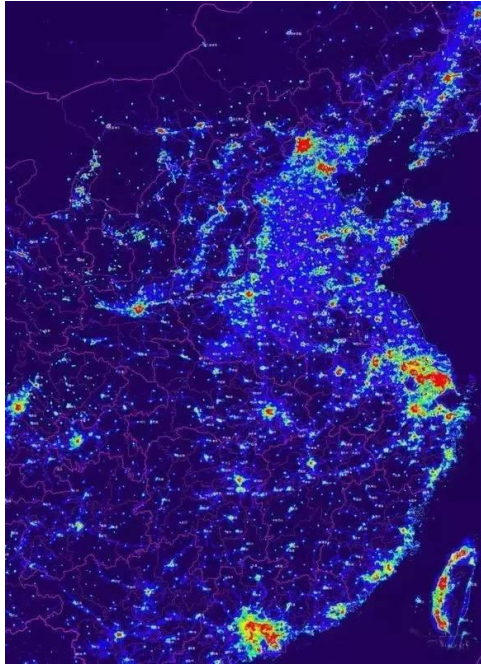
Satellite Reveals New Views of China



VIIRS light data
by NASA



Satellite Reveals New Views of China



Thermodynamic chart



Railway network

A row of five orange and silver stationary bikes parked in a line. The bikes are arranged diagonally, showing their front wheels, handlebars, and seats. They have a modern, sleek design with orange frames and silver accents. The background is a plain, light-colored surface.

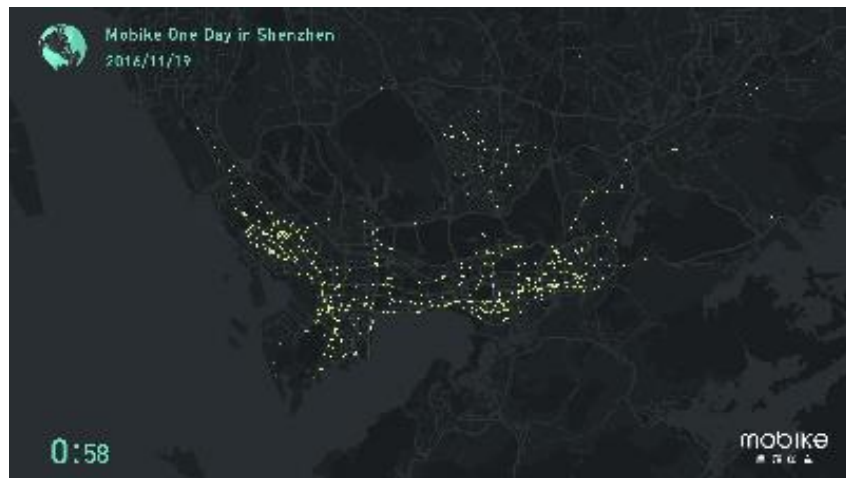


曾经的共享单车大战



数据点亮城市

- 深圳市摩拜单车的一天



摩拜算法大赛：让摩拜出行更智能！

- **任务：**根据摩拜提供的数据，预测骑行的目的地所在区块。



时代的呼唤

国家



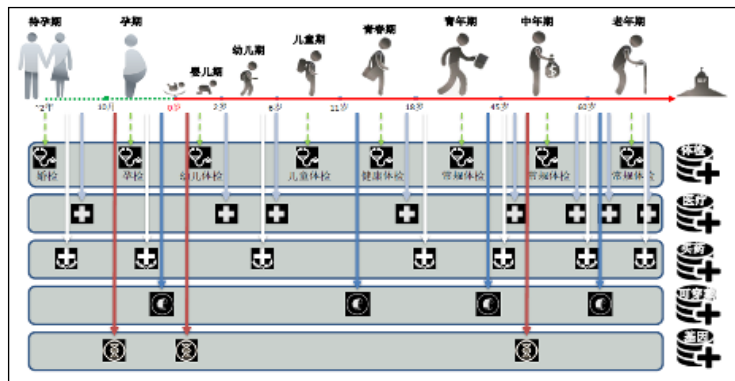
企业



机器



人类



背景1：智能时代



ALL Systems Go

At last — a computer program that can beat a champion Go player

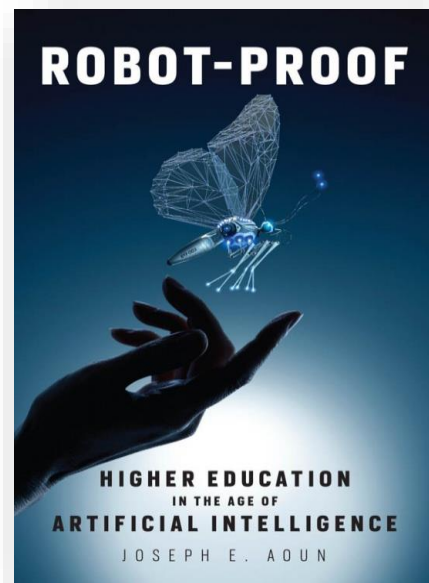
Nature 2016.01



DARK FACTORY

The robotics revolution is changing what machines can do. Where do humans fit in?

The New Yorker 2017.10

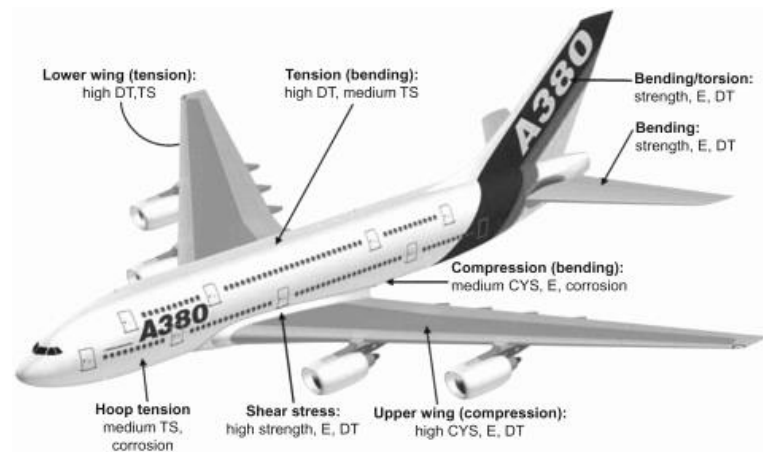


Robot-Proof

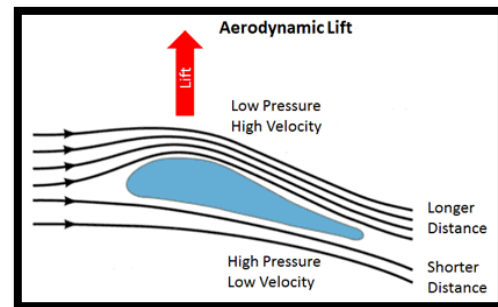
Higher Education in the Age of Artificial Intelligence

MIT Press 2017.08

航天学的启示



今天，我们并没有把航空技术看成是“人工飞行”，它就是飞行。同样，我们也不应该将**技术智能**视为“人工的”东西，而应该就把它看成是**增强人类能力的智能**。



如何赋予增强智能？ 只能靠教育！

人类增强智能 = 人脑智能 + 技术智能

从面向“知识”到面向“能力”的转变

- 基本素养的提升
 - 数字素养（Digital literacy）、数据素养（Data literacy）、人文素养（Human literacy）
- 核心能力的提升
 - 学习能力、问题求解能力、信息获取能力、分析推理能力、决策能力、.....
- 综合认知的提升
 - 系统性思维（System thinking）、数据思维（Data thinking）
 - 设计思维（Design thinking）、批判性思维（Critical thinking）
 - 认知敏捷性（Cognitive agility）、创业精神（Entrepreneurship）

背景2：科学范式与数据思维



实验思维-科学归纳

1000年前



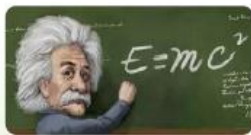
- 对自然现象的描述论证
- 对自然现象进行系统归类

牛顿三大定律提出



逻辑思维-模型推演

数百年前



- 采用建模方式
- 由特殊到一般进行推演

爱因斯坦相对论提出



计算思维-仿真模拟

几十年前



- 用计算方式模拟复杂现象
- 科学数据可以用模拟的方法获得

阿波罗登月计划成功



数据思维-数据密集型科学

2007年以后



- 与大数据密切相关
- 采用IT技术获取、处理、存储、统计分析数据，从中获取知识

AI进入高速发展期

大数据

- 大数据作为继云计算、物联网之后IT行业又一颠覆性的技术，备受关注已是毋庸置疑的事实。它好比是21世纪的石油和金矿，是一个国家提升综合竞争力的又一关键资源。
- 大数据既是一类数据，也是一项技术，还是一种理念。



大数据催生教育改革

- 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：华东师范大学、中国人民大学、复旦大学等32所
- 截至2019年4月，教育部公布新专业审批结果，总共477所高校获批“数据科学与大数据技术”专业，682所职业院校获批“大数据技术与应用”专业；
- 全国高校大数据教育联盟、大数据教育联盟、中原大数据教育联盟、数据中国“百校工程”等。

课程简介

为什么学习本课程？

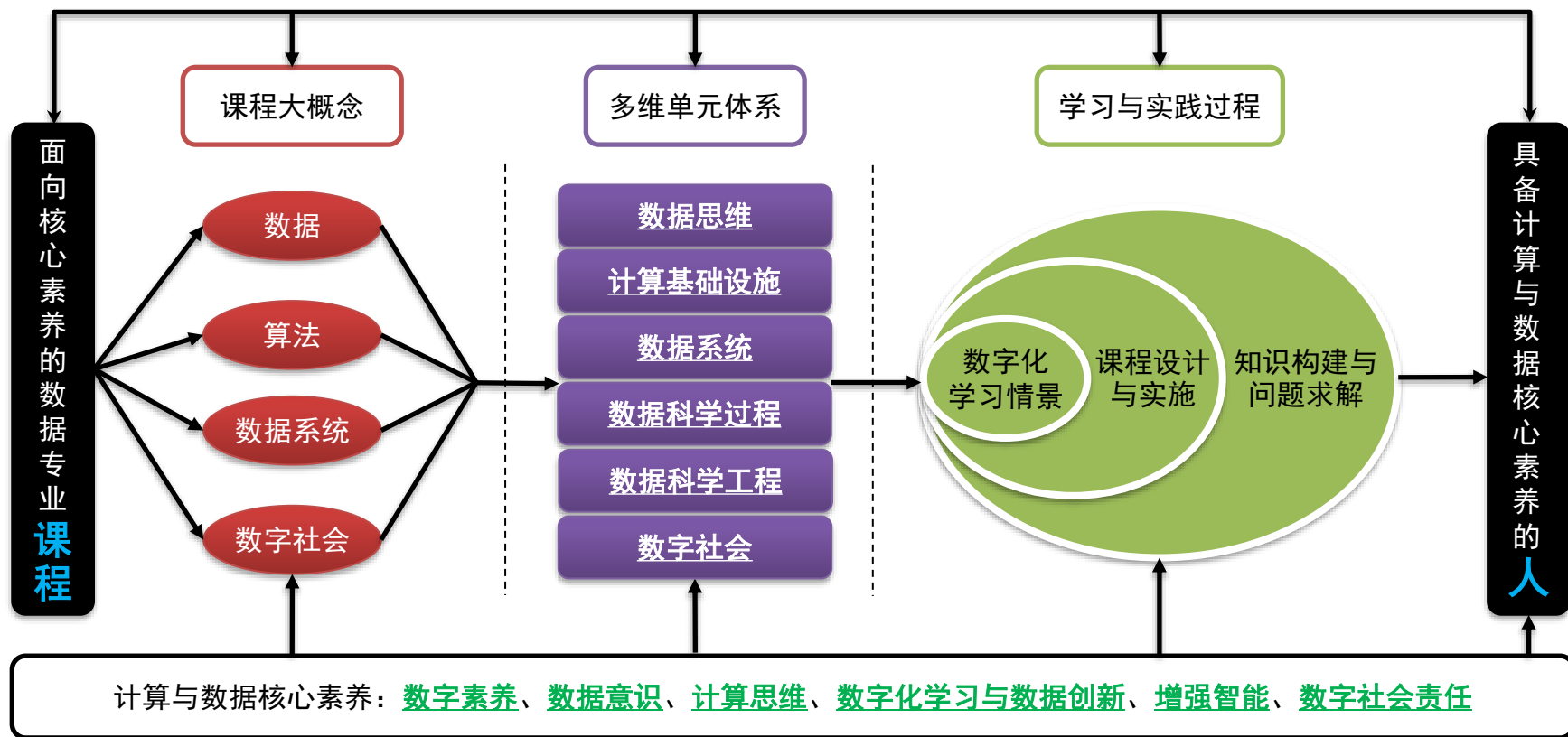
- 数据科学和大数据的理念和思维方式已经成为人们应该具备的基本常识。
- 拥有这种理念，才能够掌握数据和运用数据的人，才能在“一切都被记录，一切都被分析”的数据化时代生存和发展。
- 数据专业基础课！！！！



课程目标

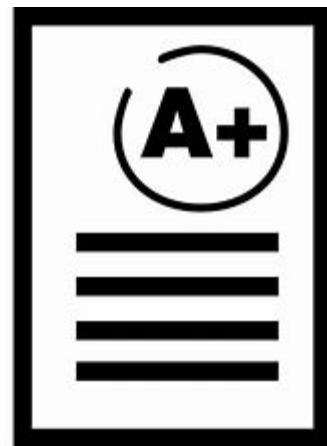
- 了解数据专业全貌，建立数据思维的意识；
- 掌握数据科学与工程的基本内涵和应用模式；
- 培养以数据为中心的问题求解能力，系统性的学习数据科学与工程的核心原理与关键技术；
- 培养开源开放的精神，建立基于开源工具的数据分析与处理意识，并做到初步的数据编程训练；
- 让大家感受到数据与计算的美，数据与计算的愉悦；
- 点燃大家对数据专业的热情与兴趣！

面向核心素养的《数据科学与工程导论》架构



课程成绩

- 平时出勤： 10%
- 实训与作业： 20%
- 期中测试： 30%
- 课程设计： 40%



课程设计

- 组队规则：1个人/组
- 完成一个完整的数据作品
 - 涉及完整的数据科学过程
 - 真实数据、有趣的问题
 - 一个数据作品报告
- 时间节点：
 - 第10周：开始选题
 - 第11~16周：完成项目作品
 - 第17周：Lightning talk（每人8分钟的时间）

数据科学与工程学院

《基于交通信息参数的探索和分析》

数据科学与工程学院

- 学校名称：华东师范大学
- 撰稿人：陈源凯
- 邮箱：jokermh@qq.com

基于空间地理位置数据和房产信息
对链家上海在售二手房房价进行分析与评估

黄振杰
2019.1.10

上海 文化空间分析

REPORTER: 张双华
TIME: 2019.1.10

基于房屋价格
数据的分析与建模

张若男

基于小米电视评论文
本数据的情感分析

10172100347
史奕林
2019年1月

爬取《我不是药神》影评
进行可视化展示



金庸的武侠世界

帕金森症的预测

演示者 王康超



用k-means改进knn在cifar-10
上的预测时间

SymbolNet : First Step of Handwriting to Latex

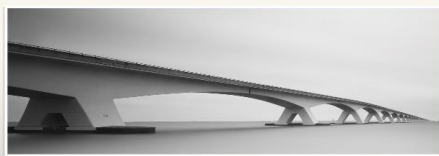
Tao Cheng

East China Normal University
taocheng01@gmail.com

January 10, 2019

图像文本识别模型探索

王子刚 10175501108



华东师范大学 200062

基于线性回归和岭回
归的台风路径预测

魏加波

数据科学与工程学院

机器学习实现王者荣耀阵容胜负预测与阵容推荐

数据科学与大数据技术
徐志雄

基于上海市二手房数据的分析

© 2019 2017 版本制作者 ID : 10175501117

基于Python工作的
数据分析与可视化

王子刚

1. 数据清洗
2. 数据探索
3. 数据建模
4. 数据可视化
5. 数据部署
6. 数据评估
7. 数据反馈
8. 数据迭代

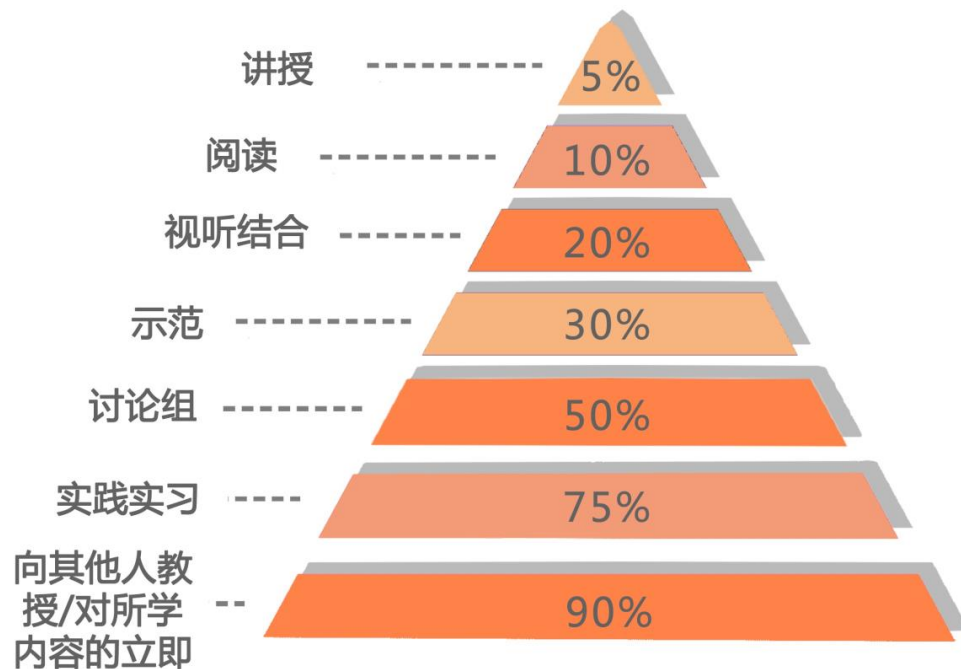
基于北美进口影片在中国票
房表现数据的分析与建模

课程模式

主体思路



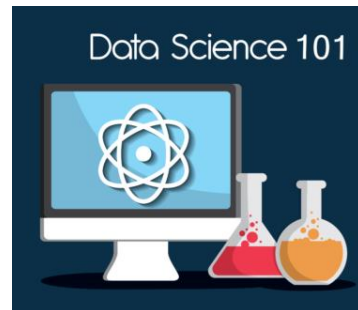
激发思辨，建立意识



动手实践，训练思维

教学模式


- 线下教学
 - 基本概念、思维方式、原理技术、应用案例
- 线上实训平台
 - 课件同步、重点解释、补充阅读、动手实践
- 综合实践与创新
 - Project based learning



课程微信公众号：嘉数汇



动手实践，训练思维

水杉在线

首页水杉学堂水杉工坊水杉校场水杉码园


“水杉在线”是华东师范大学教师科学与工程学院推出的新一代数字化全场景在线学习平台，是一个面向学生“学”、“练”、“赛”、“创”一体的综合性学习社区。目前，“水杉在线”正式上线的业务模块包括“水杉学堂”、“水杉工坊”、“水杉校场”、“水杉码园”四个子系统。同时，基于平台上积累的学习行为数据，将持续探索个性化导学、自适应学习、AI助教等智能教育应用。通过云计算、大数据、人工智能等技术手段，实现“有教无类、因材施教、寓教于乐、教学相长”的中国教育智慧。

“水杉学堂”：一站式全民计算机科学教育课堂。

“水杉工坊”：交互式在线实训服务的沉浸体验。


“水杉校场”：个性化在线编程学习的自动规划。

“水杉码园”：社交化软件项目开发的合作分享。




扫一扫手机看

00:02 / 04:00




水杉学堂
超名师资源
超酷实训体验
贴心学习服务
让学习无处不在!

进入




水杉工坊
交互式实训
极速不卡顿
大数据实训
AI实训

进入



水杉校场
个性化考试
自适应学习
精准定位
自我提升

进入



水杉码园
课程作业
课外协作
刷题刷题
数据分享

进入

交互式学习界面 (workbench)

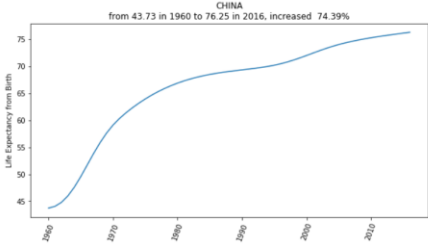
数据科学与工程导论

- Git实训教程
 - git实训行
- Python实训—自学是门手艺
 - 自学是一门手艺

GitCourse

全球范围内都一样，在过去的五十年里，人们的平均寿命预期增长得非常惊人.....

拿中国地区做例子，根据世界银行的数据统计，中国人在出生时的寿命预期，从 1960 年的 43.73 岁，增长到了 2016 年的 76.25 岁，56 年间的增幅竟然有 74.39% 之多! (执行右边的代码可以查看下图曲线的绘制过程)



如此发展下去，虽然人类不大可能永生不死，但平均寿命依然在持续延长是个不争的事实。与上一代不同，现在的千禧一代，需要面对的是百岁人生——毫无疑问，不容置疑。

这么长的人生，比默认的想象中可能要多出近一倍的人生，再叠加上另外一个因素——这是个变化越来越快的世界——会是什么样子？

我是 1972 年出生的。从交通工具来看，我经历过出门只能靠步行，大街上都是牛车马车，机动车顶多见过头拖拉机，到有自行车，到见过摩托车，到坐小汽车，到自己开车，到开有自动辅

Jupyter 1.1

可信的 Python 3

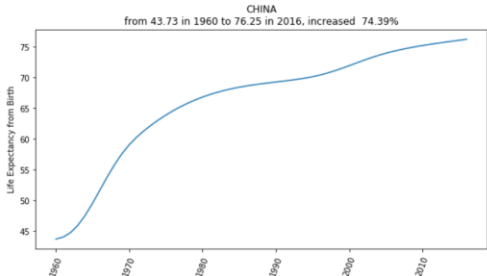
File Edit View Insert Cell Kernel Widgets Help

运行 标记

```
plt.tick_params(axis='x', rotation=70)
plt.title('CHINA\n' + note)

# plt.savefig('life-expectancy-china-1960-2016.png', transparen
plt.show()

# data from:
# https://databank.worldbank.org/data/reports.aspx?source=2&ser
```



知识体系

本课程的知识框架

四条线贯穿起来：

1. 数据思维：以数据为中心的问题求解

- 计算思维 + 统计思维

2. 基础设施：数据管理的全生命周期技术

- 采集、存储、计算、分析、展示

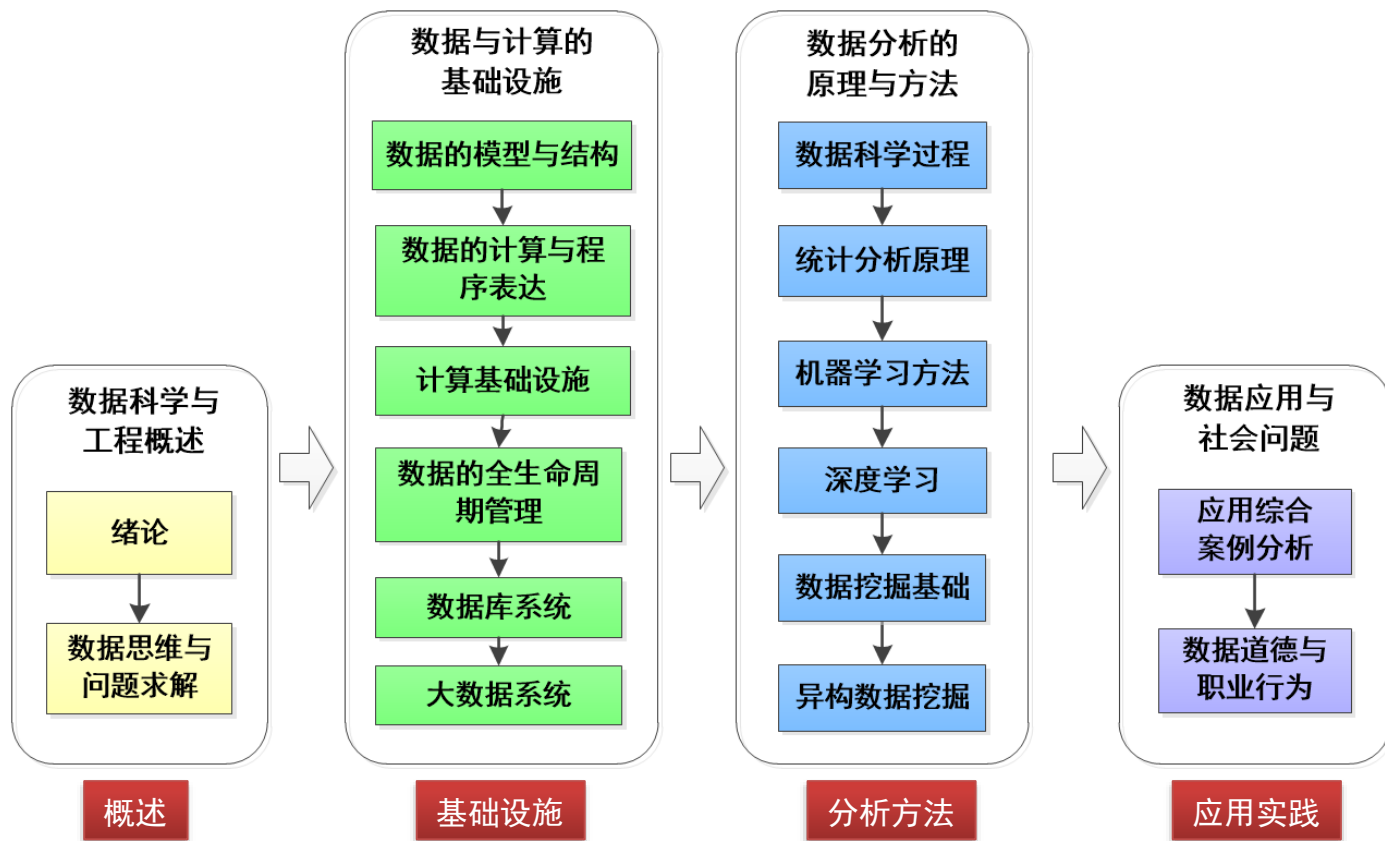
3. 分析方法：统计与算法重新定义世界

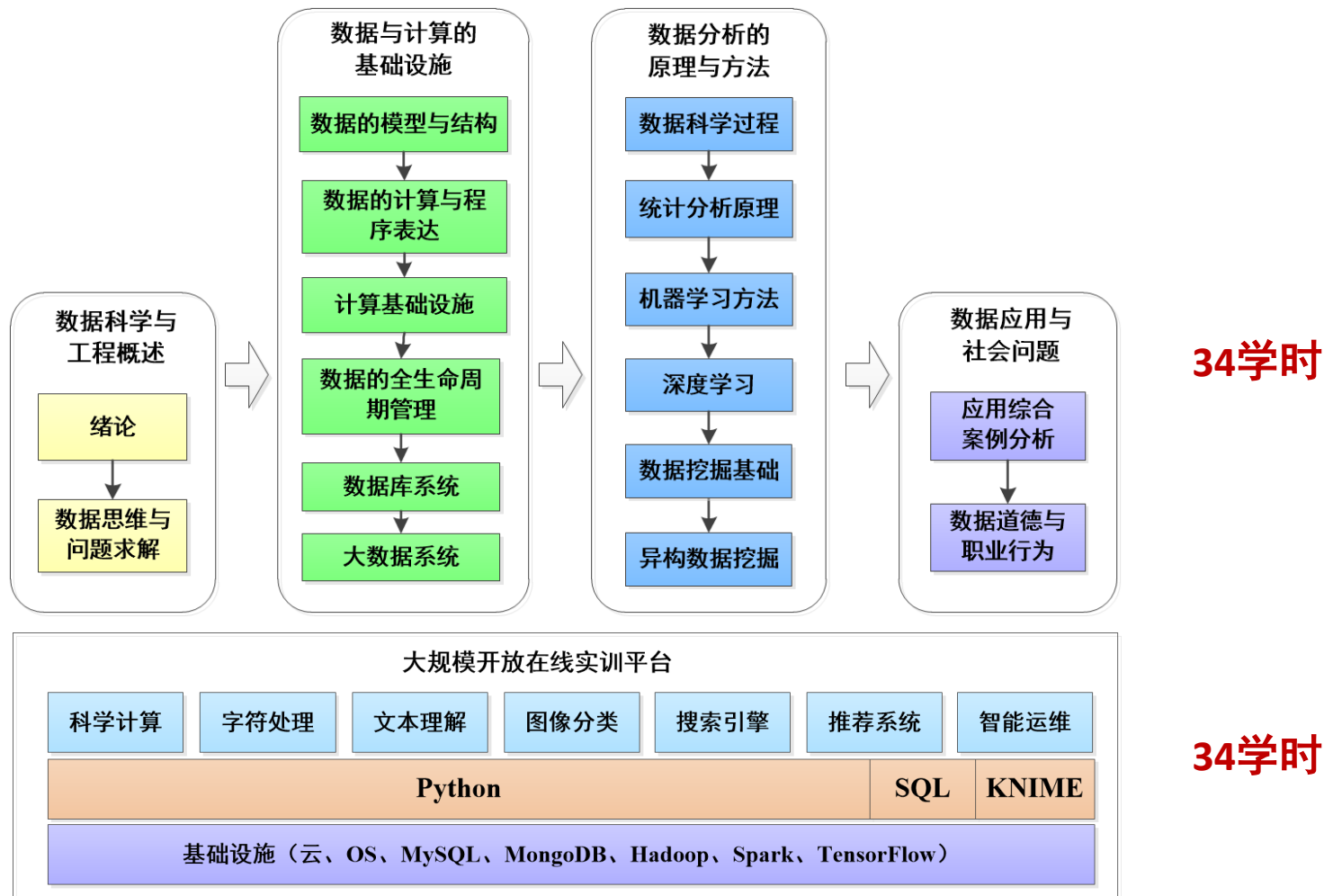
- 基本分析方法：算法分析、统计模型
- 进阶工具平台：数据科学过程、数据工作流、数据工程平台

4. 开源实践：Git, Python, MySQL, KNIME, Hadoop, Spark, TensorFlow



课程大纲（34理论 + 34实践）





	一级知识点	实践内容	学时分配
1	数据科学与工程的基本概念	Git与Python基础	2+2
2	数据思维与问题求解	Python问题求解	2+2
3	数据的模型与结构	Python数据表示与数据结构基础	2+2
4	数据的计算与程序表达	Python算法	2+2
5	计算基础设施	Python程序性能评测	2+2
6	数据的全生命周期管理（系统角度）	Python数据采集与存储	2+2
7	数据库系统	SQL数据处理与分析 NoSQL数据处理与分析	4+4
8	大数据系统	基于Python的MapReduce数据处理	2+2
9	数据科学过程（工程角度）	Python与KNIME的数据科学过程	2+2
10	统计数据分析的原理	Python统计分析	2+2
11	机器学习方法	Python机器学习	2+2
12	异构数据挖掘	Python结构化数据挖掘 Python非结构化数据挖掘	4+4
13	数据隐私与社会问题	Python数据安全与隐私	2+2
14	应用综合案例分析	智慧城市、人工智能（TensorFlow）、教育大数据	4+4
	复习、答辩	-	4

多维数据、图形图像、
自然语言、Web页面...

大问题、大体量、
快速度、高并发...

统计算法、ML算法、
算法加速、参数优化...

搜索、电商、
生物、教育...

博

大

精

深

数据科学与工程

广 (泛)

开 (源)

思 (维)

路 (数)

数据模型、程序表达
串、链、树、表、图...

Hadoop、Hbase、
Hive、Spark...

计算思维、数据思维、
系统思维、设计思维...

数据科学过程、
在线协作实训...

THANK

YOU



DaSE
Data Science
& Engineering