
数据科学导论
Introduction to Data Science

数据科学导论

—— 数据与计算之美

王 伟
wwang@dase.ecnu.edu.cn
華東師範大學

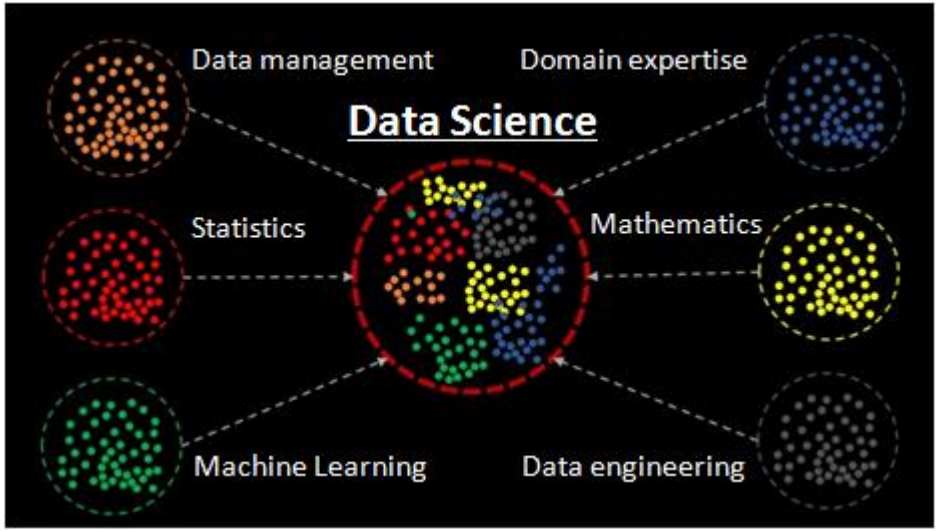
DaSE
Data Science
& Engineering

Bio

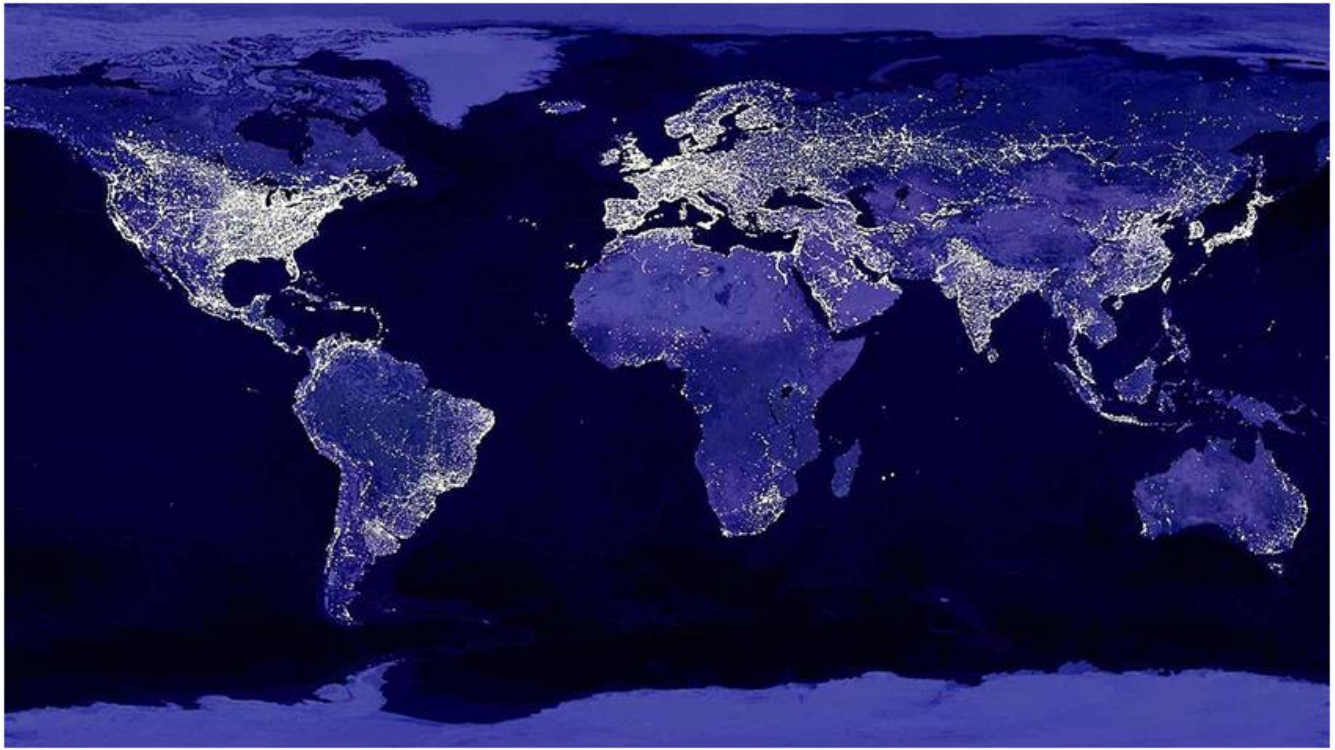
- 王伟，华东师范大学数据学院，教授
- ACM/IEEE会员、CCF高级会员，CCF 开源专委会委员、CCF大数据专委会委员、开源社副理事长
- University of Florida, Visiting Research Scholar
- University of Wisconsin- Madison, Honorary Fellow, USA
- 研究方向：容器云、开源数字生态学、计算教育学
- E-mail: wwang@dase.ecnu.edu.cn


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

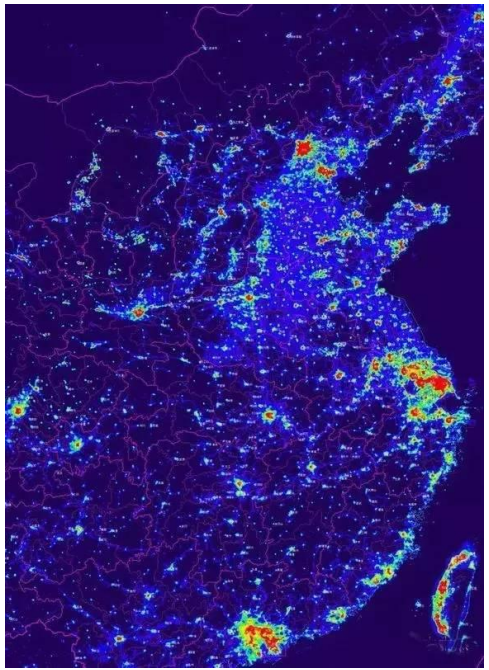


来源：https://www.nasa.gov/

Satellite Reveals New Views of China

VIIRS light data
by NASA

Satellite Reveals New Views of China



Thermodynamic chart

Railway network

摩拜单车


mobike 摩拜单车
351192
总骑行距离（公里）
42563
总减排量（千克）
1917/2843
被骑行 / 总运营车辆
153/5163
实时 / 总骑行次数
昨天
今天
骑行次数
1500
1000
500
0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
性别比例
年龄比例
骑行比例
长宁区
静安区
卢湾区
东华大学
徐汇区
上海体育馆
漕河泾
上海植物园
华东理工大学
三林镇
北殷家宅
S20
复旦大学
附属儿科医院
外环高速
华泾镇
乔里康
莘庄立交
为民村
闵行区

曾经的共享单车大战


智享单车
公共自行车
骑点
奇奇出行
CCbike
小蓝单车
7号电单车
小鸣单车
黑鸟单车
哈罗单车
何在
酷骑单车
1步单车
由你单车
踏踏
funbike单车
骑呗单车
熊猫单车
闪电单车
云单车
优拜单车
电电Go单车
ofo共享单车
摩拜单车


Tencent 腾讯
摩拜单车
VS
ofo


摩拜
UBIKE
QIDEI
ofo
bicycle
小鸣
bluegogo

数据点亮城市

- 深圳市摩拜单车的一天

来源：http://youku.com/

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。

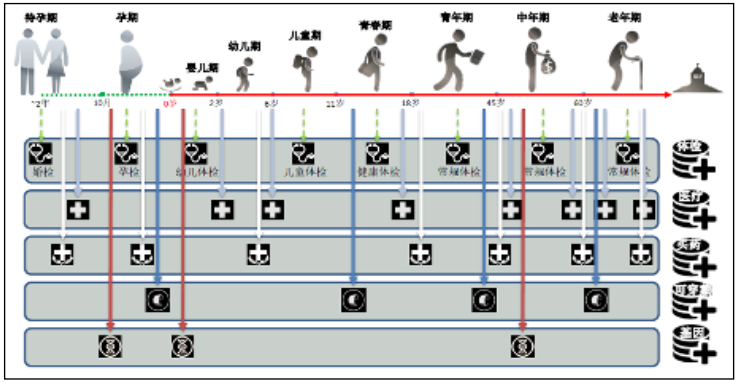
时代的呼唤

国家

企业

机器

人类



智能时代

ALL Systems Go

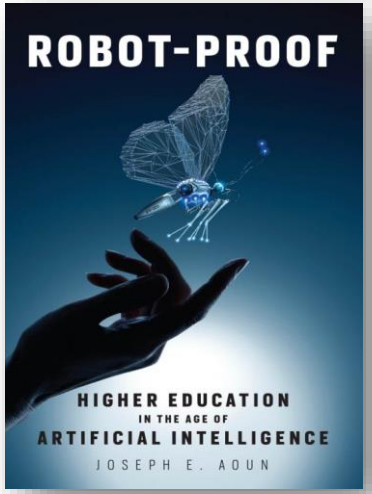
At last — a computer program that can beat a champion Go player

***Nature* 2016.01**

DARK FACTORY

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

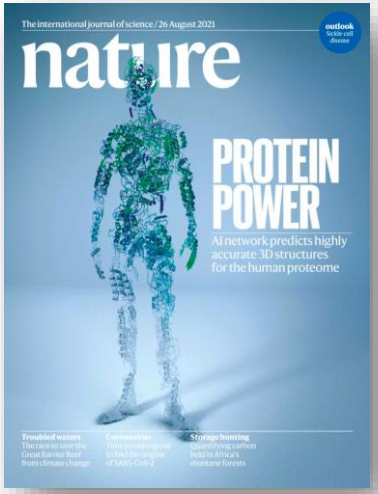
***The New Yorker* 2017.10**



Robot-Proof

Higher Education in the Age of Artificial Intelligence

***MIT Press* 2017.08**

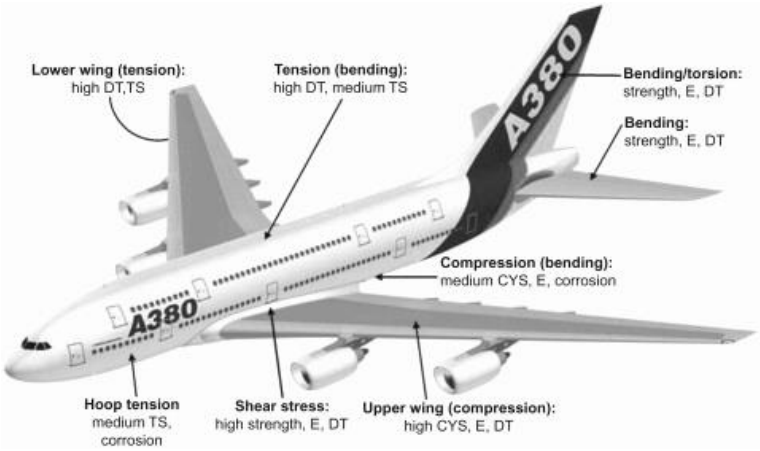


AlphaFold

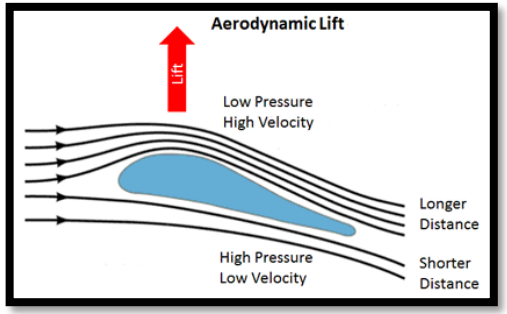
AlphaFold, software that can predict the 3D shape of proteins, is already changing biology.

***Nature* 2021.08**

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。



群聊: 2023数据科学导论课程

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

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
- 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

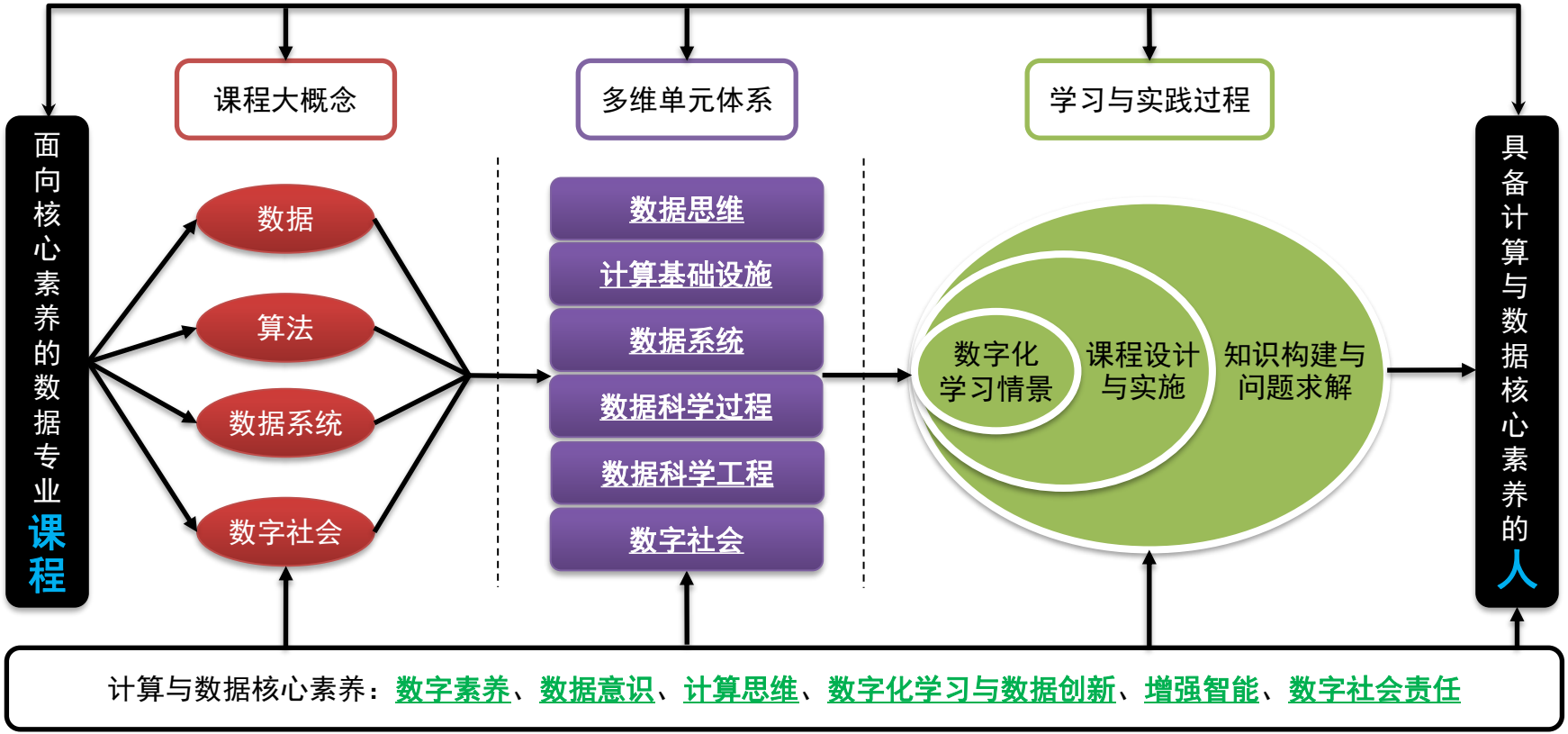
为什么学习本课程？

- 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《数据科学导论》架构



课程成绩

- 平时出勤：10%
- 社区活跃：10%
- 平时作业：30%
- 期末大作业：50%

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8周：开始选题
  - 第9~12周：完成项目作品
  - 讲解与演示（每人 10 分钟以内的视频）

数据科学与工程学院

《基于交通信息参数的探索和分析》

数据科学与工程导论

- 学校名称：华东师范大学
- 报告人：马源航
- 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

莫根杰
2019.1.10

上海
文化空间分析

REPORTER：施双宇
TIME：2019.1.10

基于房屋价格数据的分析与建模

基于小米电视评论文本数据的情感分析

10172100347
史继林
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 王原昭

华东师范大学200062

基于线性回归和岭回归的台风路径预测

魏如诚

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

王子明 10175501108

用k-means改进knn在cifar-10上的预测时间

数据科学与工程导论

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
徐志楠

基于上海市二手房数据的分析

基于Python工作的数据分析与可视化

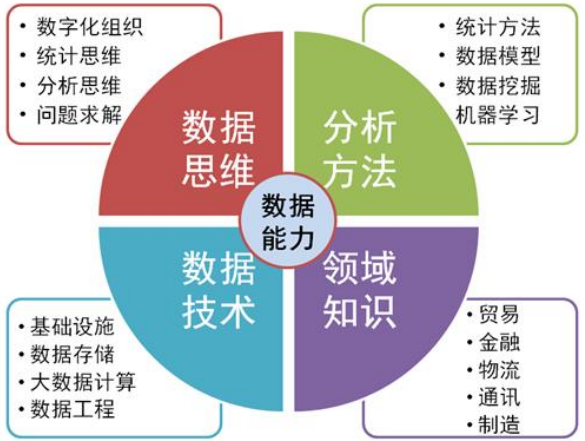
基于北美进口影片在中国票房表现数据的分析与建模

知识体系

本课程的知识框架

四条线贯穿起来：

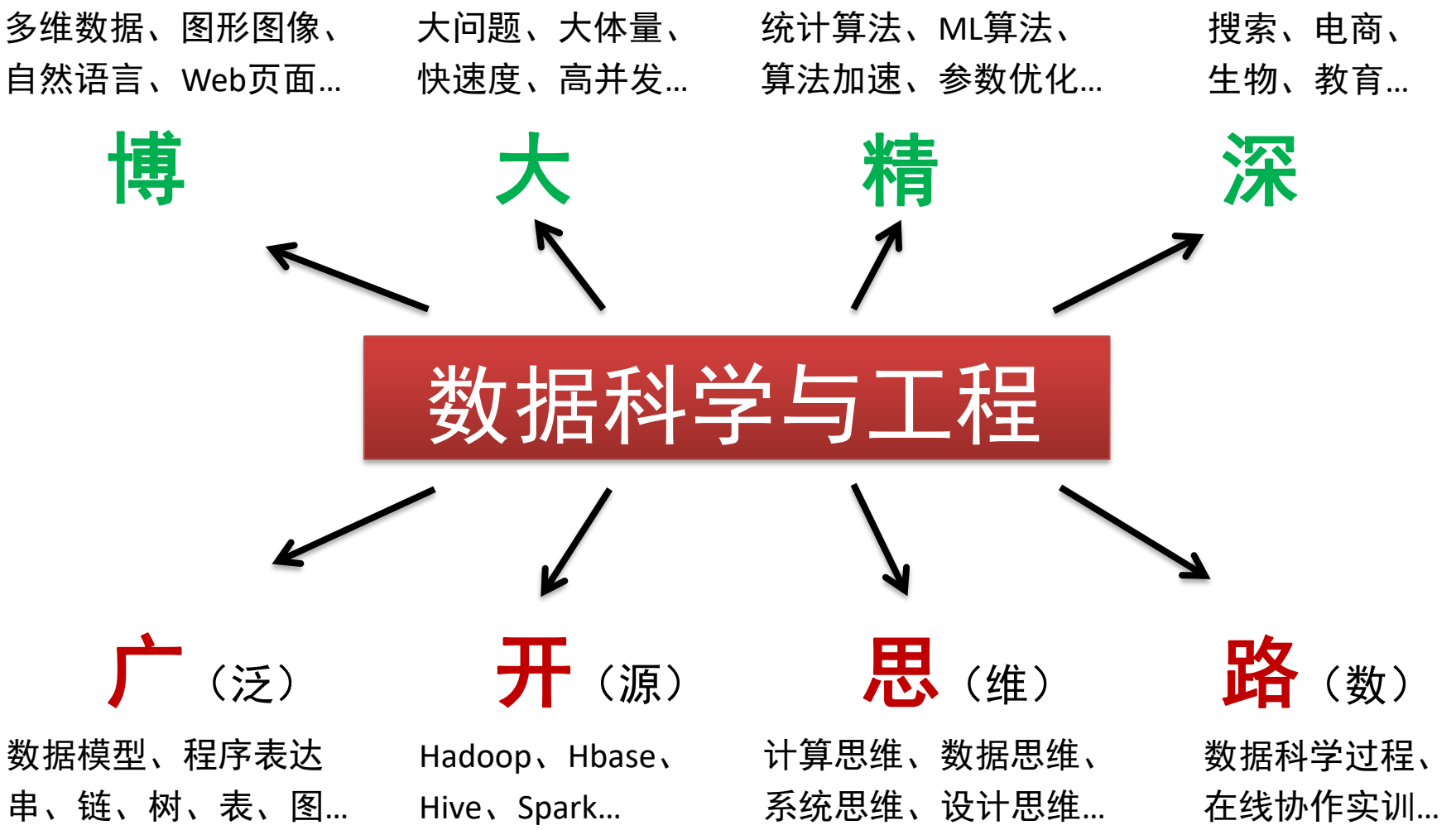
1. 数据思维：以数据为中心的问题求解
   - 计算思维 + 统计思维
2. 基础设施：数据管理的全生命周期技
   - 采集、存储、计算、分析、展示
3. 分析方法：统计与算法重新定义世界
   - 基本分析方法：算法分析、统计模型
   - 进阶工具平台：数据科学过程、数据工作流、数据工程平台



课程安排（Tentative）

https://github.com/X-lab2017/ds-2023-autumn

周数	日期	内容	主讲	本周任务	课件	开放资源
01	09-11	数据科学概述	@will-ww	任务01		
02	09-18	Python 语言	@will-ww			
03	09-25	计算系统与基础设施	@will-ww			
04	10-02	数据全生命周期管理	@will-ww			
05	10-09	数据分析方法	@will-ww			
06	10-02	机器学习				
07	10-16	数据挖掘				
08	10-23	数据科学综合实践	@will-ww			
09	10-30	课程大作业	@will-ww			


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…

EAST CHINA NORMAL UNIVERSITY
华东师范大学

DaSE
Data Science
& Engineering