華東師範大學
EAST CHINA NORMAL UNIVERSITY

$$+a)^n=\sum_{k=0}^{n}\binom{n}{k}x^k a^{n-k}$$

数据中台

Data Power Platform

绪论 课程介绍

$$(1+x)^n=1+\frac{nx}{1!}+\frac{n(n-1)x^2}{2!}+\cdots$$

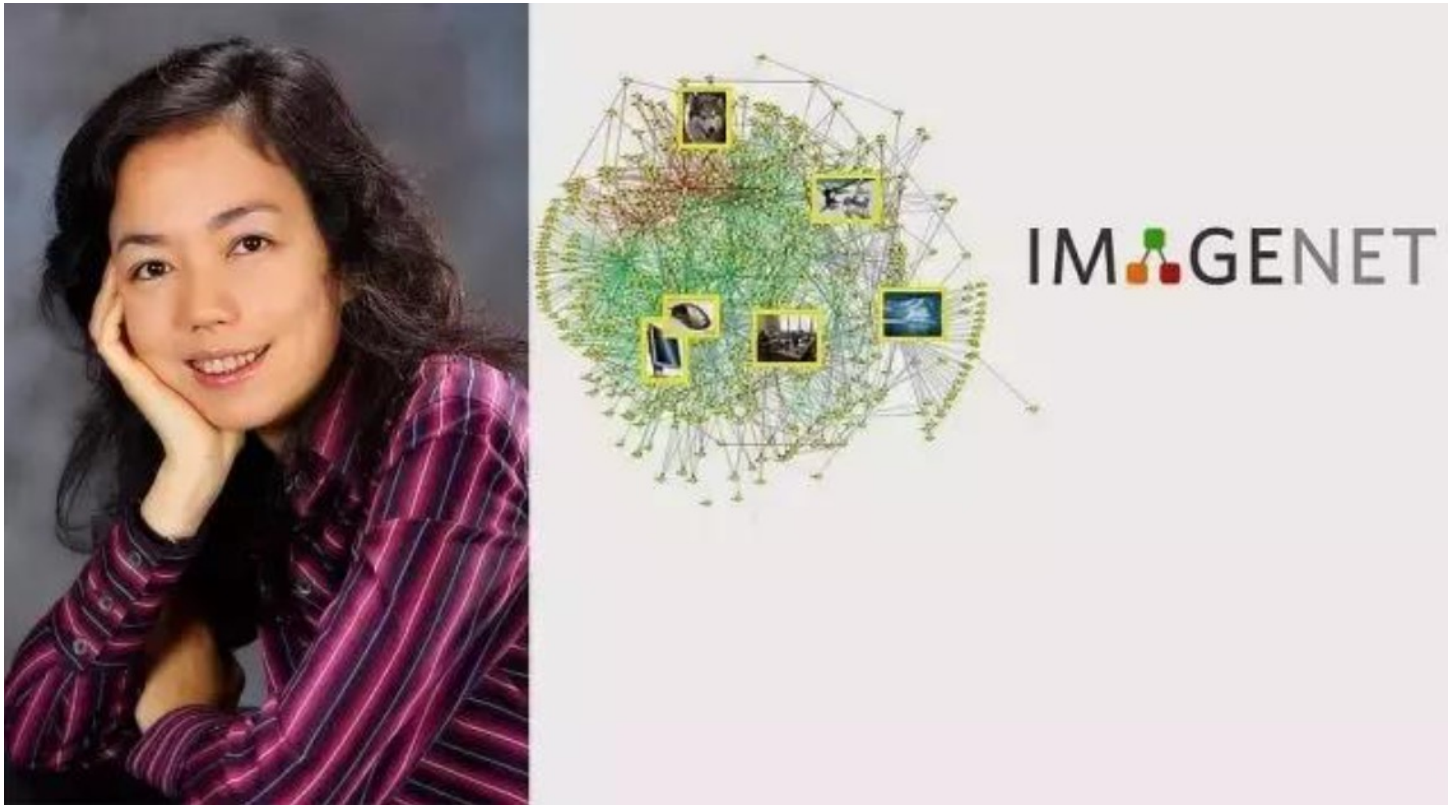
课程提纲

Content

课程提纲

Content

ImageNet 挑战赛

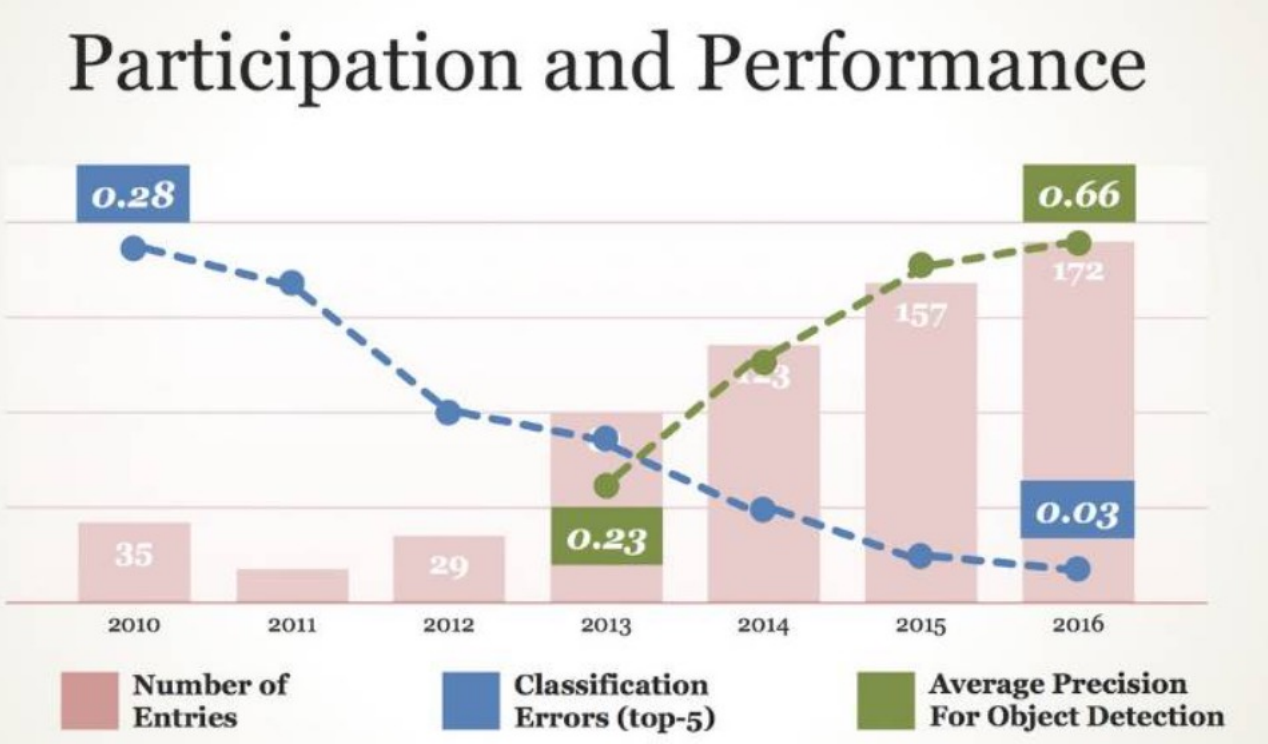


- 算法 VS. 数据
  - 2006年之前，大家更看重算法
  - 数据需要能够反映真实世界：需要建设更好的数据集

ImageNet 挑战赛

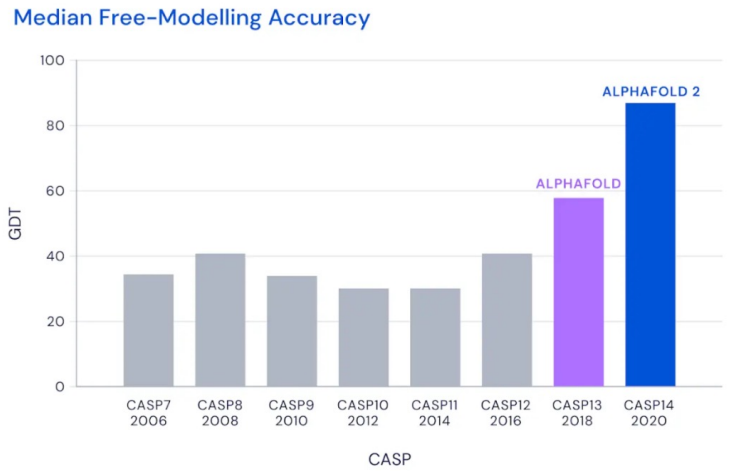
- ImageNet数据集的构建
  - 本科生收集图片每小时10美元，需要90年才能完成
  - 算法获取 + 人工确认，未来算法也会受限
  - 众包平台
    - ✓ 亚马逊Mechanical Turk 可以聘用世界各地的人帮忙标注数据，费用低
      - 如何保证标注质量？
      - 如何避免系统被欺骗？
  - 两年半时间完成了数据集标注
    - ✓320万张标记图片，共分成5,247类
  - 目前拥有15 million的图像数据集，大约有22,000类
- 标注数据是最苦最累的活

ImageNet 挑战赛结果



- **ILSVRC**：ImageNet Large-Scale Visual Recognition Challenge
- ILSVRC从2010年开始举办，到2017年是最后一届
  - ImageNet 改变了人们的思维方式
  - 数据改变了AI 和 世界

Alphafold



- CASP 竞赛由 John Moult 和 Krzysztof Fidelis 两位教授于 1994 年创立
- CASP (The Critical Assessment of protein Structure Prediction)旨在对蛋白质结构预测进行评估，被誉为蛋白质结构预测的奥林匹克竞赛
- 11 月 30 日，谷歌旗下 AI 技术公司 DeepMind 提出的深度学习算法‘Alphafold’ 破解了出现 50 年之久的蛋白质分子折叠问题

“神探”抓小偷

a 正常出行者

b 旅游者

c 购物者

d 扒手

- 北京智能交通卡：在2014年4-6月共有600w用户的16亿条记录
- 目标：根据公交卡出行记录识别小偷。

http://www.kdd.org/kdd2016/papers/files/adf0629-duA.pdf

随申码

- 随申码2020年2月17日正式上线，是上海市民的生活服务码，方便市民工作、生活、出行等需要
- 随申码背后是数据的问题，汇集了交通出行、卫计委、运营商和公安的数据
- 除了防疫，现在随申码已扩展用于公交、地铁和医保等应用

数据是新能源 (Data is Power)

新零售、新制造、新金融、新技术、新能源

数据是新能源 (Data is New Power)

“蒸汽能”
(Steam Power)
第一次技术革命
英国，机械系

“电能”
(Electric Power)
第二次技术革命
美国，电机系

“数据能”
(Data Power)
第三次技术革命
中国，数据学院

大变局时代：“未来已来，一切重构”

Future is coming, the world is being reshaped

[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

数据是新能源 (Data is Power)

新能源(New Power)的出现促进经济的变革

农业经济 → 工业经济 → 数字经济

[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

数字经济的特征 (Digital Economy)

数字经济的基本特征

- 以**数据**资源为重要生产要素，以现代信息**网络**为主要载体
- 以信息通信技术**融合**应用，全要素数字化**转型**为推动力
- 促进公平与效率更加统一

重大的时代转型

生产方式变革，**生产关系再造**
经济结构重组，生活方式巨变

数字经济背景下，数据成为第五大生产要素

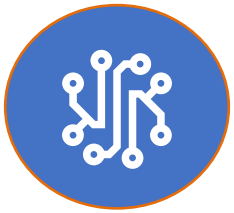
[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

数据是一种新的生产要素

- 数据成为独立的一种生产要素
  - 2017年，互联网经济时代，**数据是新的生产要素**，是基础性资源和战略性资源
  - 2020年4月9日，数据和**土地、劳动力、资本、技术**等一样是一种生产要素（中央文件）
  - 数据是数字经济腾飞的基础性资源，面临着数据孤岛、数字鸿沟、数据隐私和数据安全等诸多挑战

数据的新特征

数据最大限度的重复使用（数据复用）可以提升企业的效能



非消耗品

- 数据可以无限次使用
- 数据可以复用

相对复杂

- 数据使用过程产生新的数据
- 富含价值：1 + 1 > 2

规模急剧增长

- 2018 年达到 33 ZB
- 数据以指数级增长

$$1\ ZB = 2^{10}\ EB = 2^{20}\ PB = 2^{30}\ TB$$

数据应用与开发存在的问题

效率问题

- 需求交付周期长，平均一周左右
- 企业有哪些数据？找到所需要的数据费时费力
- 大量的数据很少被访问，但是却占用了大量的存储资源
- 实时数据处理需求难以满足

协助问题

- 数据孤岛问题突出，存在数据质量问题
- 部门间系统的重复建设
- 业务逻辑混乱

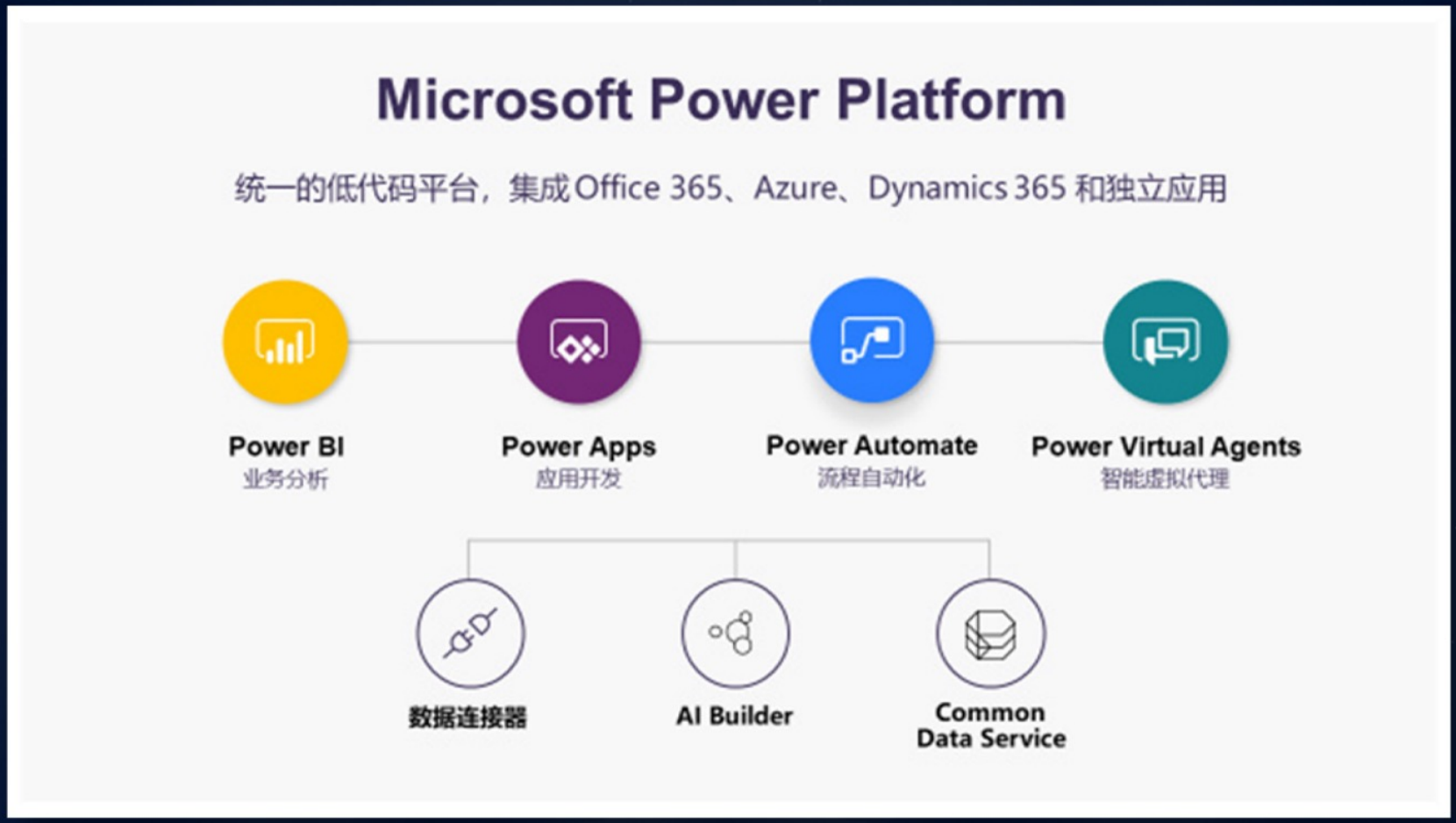
能力问题

- 开发人员多
- 既懂业务又了解技术的人少
- 业务创新能力有待提高

数据中台诞生：Supercell的启发

- 2015年，马云访问芬兰移动游戏公司 Supercell
  - 公司团队不到200名
  - Supercell 经过6年沉淀下来的游戏开发过程中那些公共的、通用的游戏素材和算法，让团队可以像搭积木一样快速研发一款新游戏
  - 一款游戏平均负责团队平均2-5人，不超过7人
  - 年税前利润15亿美金，2016年以86亿美元被腾讯收购
- 2015年底，阿里巴巴启动中台战略
  - “大中台、小前台” 的组织机制和业务机制
  - 集合整个阿里集团的运营数据能力、产品技术能力，对各前台业务形成强力支撑
- 随后，腾讯、百度等头部互联网企业纷纷推进数据中台建设

Microsoft Power Platform



[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

Microsoft Power Platform

面对数字化转型，每一家公司都将成为软件公司

— Satya Nadella

助力企业数字化转型的低代码平台

- **数据大众化：**通过连接器和通用数据服务（Common Data Service）整合业务数据，提升数据洞察能力
- **开发大众化：**低代码、低门槛，“全民低代码开发”灌注企业强大创新力
- **AI大众化：**利用AI Builder，根据数据和需求量度身定制，使APPs和流程更加智能，创建一些神奇的AI认知服务功能

[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

课程提纲

Content

1 为何需要数据中台？

2 课程概述

3 授课信息

数据中台 (DPP)： 发挥数据要素作用的平台

数据中台

以打通**部门或数据孤岛**的统一数据平台为基础，构建**统一数据资产体系**，并以API服务方式为**全渠道业务（分析 + 应用）**提供即时交付能力的企业级数据架构。

- **统一数据平台：**它不会取代原来的系统，而是把原来组织中分散在各系统中的数据汇聚到统一平台之中。
- **数据资产体系：**数据资产体系规划。对数据打标签，组织目录和结构，便于发现和使用。
- **数据服务 （Data as a Service，DaaS）：**以API的标准接口方式向前端的业务场景或分析场景提供服务。

[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

数据中台 (DPP)：企业级的能力复用平台

数据化创新平台

- 支撑企业数字业务应用的标准化及快速定制化
- 实现数据驱动的精细化运营，沉淀企业的数据资产
- 解决企业业务在面向产业互联、生态发展过程中所遇到的应变与响应能力问题

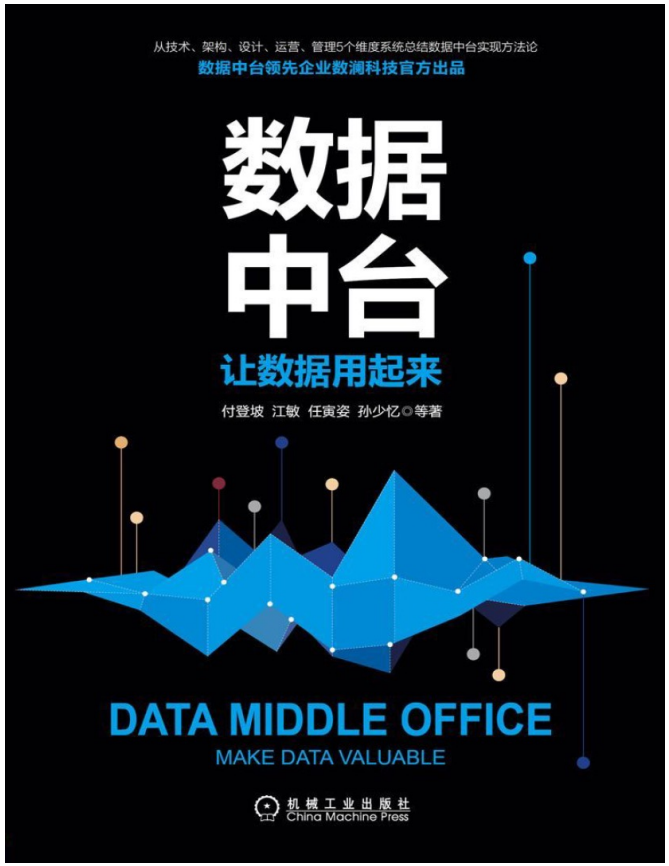
数字化转型的基础设施

- 构建数据驱动的增长体系
- **低代码敏捷**应用开发
- 科技企业想利用数据中台改造传统产业
- 传统企业也希望能利用数据中台完成数字化转型
- 前提是需要先打破企业数据的孤岛状态

[周傲英. 历史照亮未来: 数据化时代的数据库. 华东师范大学第八届数据科学与工程暑期学校. 2021]

主要参考书

- 付登波 等.《**数据中台：让数据用起来**》机械工业出版社. 2020.



理论课内容

- 导论
  - 第一章 数据应用发展历史
  - 第二章 什么是数据中台
- 方法和实现
  - 第三章 数据中台建设与架构
  - 第四章 数据中台建设的评估与选择
  - 第五章 数据汇聚联通：打破企业数据孤岛
  - 第六章 数据开发：数据价值提炼工厂
  - 第七章 数据体系建设
  - 第八章 数据资产管理
  - 第九章 数据服务体系建设
- 运营与安全
  - 第十章 数据中台运行机制
  - 第十一章 数据安全管理

实践课内容

- 动手做一个数据中台，实现四大核心能力
  - 汇聚整合（第5章）：数据获取和存储
    - ✓爬虫采集
    - ✓格式转换
  - 提纯加工（第6、7章）：数据处理和数据体系建设
    - ✓预处理、离线计算/实时分析引擎、可视化/Notebook建模
    - ✓贴源数据ODS、统一数仓DW、标签数据TDM、应用数据ADS
  - 服务可视化（第8、9章）：数据资产化和服务化
    - ✓数据资产门户
    - ✓查询服务、分析服务、推荐服务、圈人服务
    - ✓数据大屏、数据报表、智能应用
  - 价值变现（第10、11章）：中台运营
    - ✓可阅读、易理解、好使用、有价值
    - ✓价值挖掘和模式创新

课程提纲

Content

课堂文明 Class Civility

- 原则：**不要影响其他人**（老师、同学）
- 不迟到、不早退
- 不交头接耳、不大声喧哗
- 手机关机或静音

学习建议

- 不是一门读读背背的课程
- 实践内容贯穿课程始终
- 很多问题没有标准答案，学会思考很重要

考核办法和评分规则

- 采用考查方式和百分制
- 评分：
  - 项目报告（描述，代码，结果）60%
    - ✓ 子任务1：汇聚整合
    - ✓ 子任务2：提纯加工
    - ✓ 子任务3：服务可视化
    - ✓ 子任务4：价值变现
  - 项目演讲 20%
  - 出勤和课堂表现 20%
- 要求：
  - 项目报告需要包括完整的项目介绍、设计和实现的描述、算法、代码、测试、具体的执行过程和结果等。
  - 项目演讲每人30分钟（20分钟自述、10分钟提问）
- 禁忌：
  - 发现项目报告内容互相抄袭1次即作零分处理

课程安排

周次	讲课内容	课时	实践内容	课时
W1	绪论和课程介绍、 项目布置和SPEC介绍	2		
W2-W11	第一章-第十一章	2*10	项目报告	2*12
W12-W13	实际应用案例	2*2		
W14-W18	项目演讲 & 总结	3*2		

项目报告提交截止时间：W14上课前，未按时交将影响成绩。

- 具体准确截止时间后续课程会通知。

课程联络

- 讲师
  - 陈岑 cenchen@dase.ecnu.edu.cn
- 助教
  - 王嘉宁 52245903002@stu.ecnu.edu.cn
- 微信群