DS501 Capstone 项目说明

请选择你想做的Track（6/28周三中午12点问卷关闭）：
https://docs.google.com/forms/d/e/1FAIpQLSdXE6apHBWHPKIlSOtqui7PRJDSbftG6nUPIDPvLXc5NY4DEA/viewform?usp=sf_link

Capstone 工业级实战项目日程安排		
周五分项目个性化实战指导	**周六专属内容拓展提高**	**周中分项目小组讨论**
按track分教室上课，以zoom meeting的形式进行，课程负责人会在课前提供课程链接。每位同学都可以发言和老师进行沟通	*全班统一听课，以传统的zoom webinar的形式进行，会收到zoom的邮件提醒。以guest speaker的讲解为主，同学们可以通过Q&A向老师提问*	*按track分小组讨论，由组长建meeting，将meeting链接发给组员*
7:00pm-9:00pm PST	7:00pm-9:00pm PST	时间自选

Capstone项目都有哪些Track？

- **[Track 1] Kaggle - breast cancer**
  - 数据内容：
    - Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Label: Diagnosis (M = malignant, B = benign).
    - a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)
  - 项目目标：
    - Utilize various tools to clean and explore data.
    - Apply different models in this binary classification problems and improve performance by designed metrics.

- **[Track 2] 神测数据**
  - 公司介绍：国内领先的用户行为分析产品
  - 数据内容: 该公司官网访问约一周的数据，包括用户访问时产生的点击按钮、申请账号、提交验证码、观看视频、离开页面等行为记录。详细的日志描述并配以官方技术文档以及API手册说明。

    - 项目目标:
        - Clean dirty log data and transform it for analytics.
        - Exploratory data analysis, e.g. find user activity levels for different events, and user interaction with web components.
        - Find the conversion rate of users, identify key factors that bottleneck the conversion rate.
        - Propose any hypothesis and set up experiments for testing.
        - Build machine learning models to predict user behaviors, including but not limited to signup, churn, etc.
        - Discover interesting insights in the dataset and suggest how to improve the user signup rate.

- **[Track 3] 某知名音乐播放盒数据挖掘**
    - 公司介绍：某知名音乐播放平台
    - 数据内容：刚出炉的新鲜数据：）每日260K新增用户的3 million+的歌曲播放记录（不断更新中），包括用户uid, 用户os, 播放歌曲的rid, 歌曲的类型, 歌曲名称, 歌手名称, 播歌时长, 歌曲时长等信息。
    - 项目目标：
        - Validate dataset, identify missing values and find inconsistencies in the dataset.
        - Perform data cleaning and transformation, feature engineering
        - Exploratory data analysis, e.g. find most popular songs, most active users
        - Clustering users based on their listening behaviors, find latent features
        - Build music recommendation system based on user listening history, including: popularity-based recommender, item-item based recommender, matrix factorization-based recommender.

Capstone评价标准是什么？

- 项目完整程度：同学们完成项目的流程的完整度（data processing, explorotary，建立模型，performance等）
- 项目复杂程度：同学们完成的项目需要利用起课程中学到的技术和模型
- 项目展示水平：同学们在Demo Day上项目展示
- 项目商业价值：同学们在项目完成以后对business创造多少价值
- 项目新颖程度：同学们的项目能解决多少未被解决的问题

项目代码要放在哪里？

项目的代码统一放在同学们自己的Github Repo上。

我该如何选择自己的Capstone项目？

请同学们按照自己的志愿，选择两个自己最想做的Track，并在问卷上标明顺序。我们将按照同学们的志愿情况，为大家分配track。

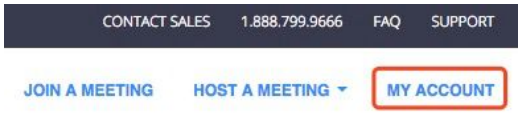
周中分小组讨论的目标是什么？

我们希望同学们通过互相合作与支持，共同完成Capstone项目。每次分小组讨论会结束后，请每组在每周五实战指导前，请各小组向老师提交周中分小组讨论报告。报告形式可以是你们讨论的会议记录，也可以是Github上的代码汇总。

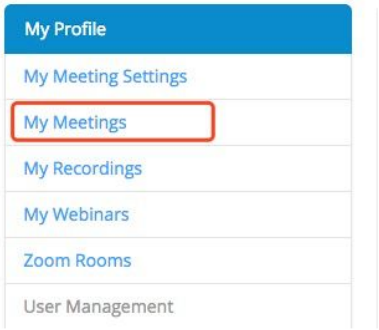
周中分小组讨论的形式是什么？

由组长建meeting，把链接发给组员，推荐使用zoom或者google hangout。以下是使用zoom建meeting的流程参考：

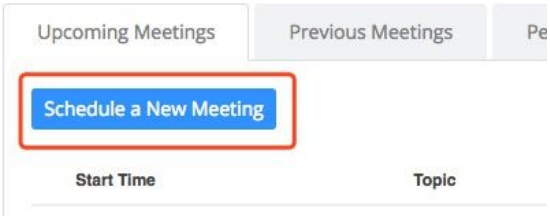
- 在zoom.us上注册一个账号
- 点击右上角my account



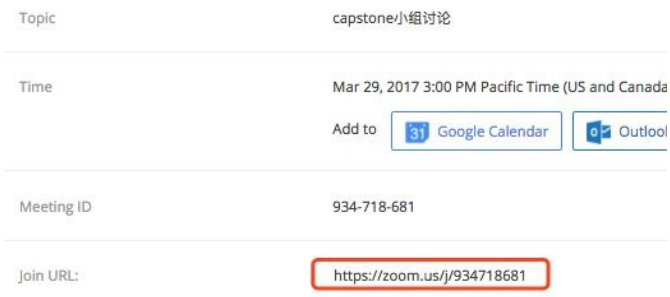
点击左边菜单栏的“my meeting”



- 点击“schedule a new meeting”设置一个meeting



- 填入小组组会的时间，其他使用default
- 复制join URL，发送给组员即可



Demo Day 的形式是什么？

Demo Day各小组同学对自己的项目成果进行展示，课程组将参与同学们的项目展示，并对项目进行评审。