加州理工学院远程科研项目 2021 暑假




项目背景

金融工程与数据分析、天体物理与人工智能两个主题的远程科研项目由世界排名第 4 的加州理工学院（Caltech）知名教授讲授精选课程，并有博士生助教解析课程内容、协助学员科研报告的撰写、提供报告修改意见。顺利完成项目后，学员将获得结业证书与科研报告，以及相应的学员推荐证明信与优秀学员证明，提升个人学术背景。

项目主题

编号	课程主题	开课日期	结课日期	时长	项目费用	课程信息
CTO1	商业金融与数据分析	7 月 17 日	8 月 22 日	6 周	9980 元	附件 1
CTO2	天体物理与人工智能	7 月 17 日	8 月 22 日	6 周	9980 元	附件 2

大学简介



加州理工学院

加州理工学院（California Institute of Technology），简称“加州理工”（Caltech），创立于 1891 年，位于美国加利福尼亚州洛杉矶东北郊的帕萨迪纳（Pasadena），是世界顶尖私立研究型大学、公认的最为典型的精英学府之一。加州理工学院的规模很小，全校学生仅 2000 人左右；截止 2019 年 10 月，该校共有 74 位校友、教授及研究人员曾获得诺贝尔奖（世界第八），平均每千人毕业生就有一人获奖（22 位校友），为世界大学诺贝尔奖得主密度之冠。加州理工的校友、教授和研究人员中还产生了 6 位图灵奖得主（世界第九）以及 4 位菲尔兹奖得主（世界第十六）。

2021 年 QS 世界大学排名：世界第 4 名

项目收获

顺利完成在线科研学术项目的学员，将收获科研报告与主课教授个人签发的项目结业证书、学员推荐证明信。除此之外，科研报告最佳小组的每一位成员还将获得额外的优秀学员证明。

项目结业证书

顺利完成课程学习的学员，将获得项目结业证书，作为此次课程学习的证明。

学员推荐证明信

授课教师将根据学员的课堂表现和科研报告，为每位学员出具项目学员推荐证明信。

优秀学员证明

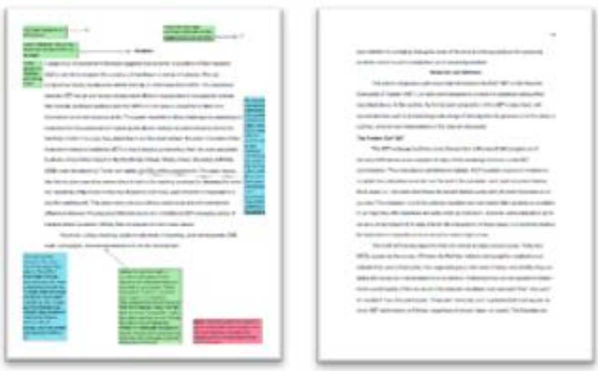
根据科研报告各小组的撰写情况，评选最佳小组，并为最佳小组成员颁发优秀学员证明信。

科研报告（示例）

学员将以小组为单位完成科研报告的撰写，为自己的学术生涯打下坚实的基础。

结业证书（示例）

学员推荐证明信(示例)



科研报告（示例）

附件 1：商业金融与数据分析

课程概览

本课程将涵盖计量金融的最新研究进展，机器学习和数据挖掘中的常用方法，并重点介绍从数据中学习的数学和统计方法，用以研究交易与投资。项目学员将通过课程与研究项目了解有哪些统计学习方法存在，并了解这些统计方法的工作方式与工作导向，并区分这些统计方法无法达到最有效果的原因。

在实践项目中，导师将带领项目学员使用 python 来完成投资组合配置，并使用累积收益、平均日收益和夏普比率等指标对投资组合的表现进行统计分析。项目将分析 Facebook, Netflix 和 Twitter 在过去 7 年的业绩。

对于想要正确管理项目组合、可视化数据集、找到有用的模式以及获得有价值的见解(如股票的每日收益和风险)的投资者来说，了解这个课程所教授的知识是至关重要的。该课程运用的方法可在将来实际应用于分析公司股票、指数或货币以及投资组合的绩效。

学习目标

课程结束后，学生将达到以下学习目标：

- 执行探索性数据分析和财务数据可视化
- 投资组合配置和计算投资组合统计指标
- 使用 Plotly Express 完成交互式数据可视化

课程师资

康斯坦丁·祖伊夫（Professor K. Zuev)
加州理工学院，计算机科学与数学系，副教授

他拥有应用数学与土木工程双博士，其主要研究兴趣包括我的研究兴趣包括应用概率、统计计算和数据科学，特别是数据科学在小概率事件建模、网络科学和量化金融上的应用等。相关研究包括：金融资产的因果网络、标普 500、富时 100、斯托克斯 50 指数的动态分析、基于可见性图的标普 500 时间序列分析，市场崩溃时期的投资人行为，金融市场的隐形互动，复杂系统的因果互动，障碍期权的高效定价策略。

他曾获得如下奖项：ASCIT 学术与研究委员优秀教学奖、美国物理联合协会最有影响力研究论文利物浦大学工程学院荣誉主管、伯努利数理统计与概率学会委员会主席、出版《数据科学对统计推断的有效性》、诺斯罗普·格鲁曼杰出教学奖等。

项目日程

周数	内容
第一周	**欢迎致辞&课程导览** **专业课（1）：**

- 统计学的总体背景、统计学习的三种重要类型、推到回归函数的统计决策理论要素

  we discuss the big picture of statistical learning with bits of history, three main types of statistical learning, and the elements of statistical decision theory that helps us to derive the regression function, which is the cornerstone of many regression methods.

- 最近邻法和经典线性回归、“维度灾难”现象

  we will learn two popular methods for solving the regression problems: the nearest-neighbor method and classical linear regression. We will also discuss a phenomenon called "the curse of dimensionality," and see how it affects both methods.

辅导课（1）：实践项目和 python 编程简介：基本算法、控制流与函数定义

Introduction to course project and python programming: basic syntax, control flows, function definitions.

第二周

专业课（2）：

- 贝叶斯分析、K-NN 和线性回归、权衡偏差与方差

  we will discuss several high-level fundamental questions. We will see how the statistical decision theory can be used to derive the optimal classification prediction rule, called the Bayes classifier, and how the k-NN and linear regression method can be extended for solving classification problems. We will also discuss the bias-variance trade-off, which describes two competing statistical forces that make the problem of selecting of the best method for a given problem a very challenging task.

- 线性回归模型的线性代数性质、如何通过最小化残差平方和来估计回归参数

  we will focus on linear algebraic properties of the linear regression model. In particular, we will discuss how the regression parameters can be estimated by minimizing the residual sum of squares, how this minimization boils down to the normal equation, and how solving normal equation leads to ordinary least squares.

辅导课（2）：机器学习流水线（Machine Learning Pipelines）实践导论

Introduction to machine learning pipeline in real practice.

第三周

专业课（3）：

- 模型选择：校验集方法、留一法和一般的 K-fold 交叉验证

  we will discuss a very interesting, important, and cool topic: model selection. In particular, we will learn three sampling-based methods for performing model selection and model assessment: the validation set approach, the leave-one-out cross-validation, and general K-fold cross-validation.

We will also talk about their relative advantages and drawbacks.

- 收缩法和脊回归

  we will discuss shrinkage methods in general and ridge regression - one of the two most popular shrinkage methods - in particular. We will also learn how ridge regression works together with cross-validation and will describe its geometric interpretation that will help us to better understand ridge regression intuitively and visually.

- 标准普尔 500 指数案例分析：危机对美国股票价格的影响

  we will share the research Time Series Analysis of S&P 500 Index using Visibility Graphs: The behavior of stock prices has been thoroughly studied throughout the last century, and contradictory results have been reported in the corresponding literature. In this paper, a network theoretical approach is provided to investigate how crises affected the behavior of US stock prices. We analyze high frequency data from S&P500 via the Horizontal Visibility Graph method, and find that all major crises that took place worldwide in the last twenty years, affected significantly the behavior of the price-index. Nevertheless, we observe that each of those crises impacted the index in a different way and magnitude. Interestingly, our results suggest that the predictability of the price-index series increases during the periods of crises.

辅导课（3）：使用 Pandas（Python 数据分析库）进行时间序列数据分析

Data science using Python lib pandas for time series data

第四周

专业课（4）：

- 套索算法、最优子集选择、岭回归

  we will see that ridge regression fails to improve the model interpretation, since it does not exclude any predictors from the regression model even when the tuning parameter lambda is large. This will motivate us to define and study the LASSO, another very popular shrinkage method. Finally, we will discuss the relative advantages and disadvantages of best subset selection, ridge regression, and the LASSO.

- 支持向量机（SVM）分类器的介绍和相关重要概念

  we will prepare ourselves to the introduction of the support vector machine (SVM) classifier, one on the most popular classification methods introduced by the computer science community in the 90s. We will discuss several important notions and concepts the SVM is based on: separating hyperplanes, the maximal margin hyperplane and classifier, support vectors, and Lagrange duality.

	辅导课（4）：使用 Python 机器学习库 Scikit-learn 进行数据分类 *Classifications using python library scikit-learn*
第五周	**专业课（5）：** 案例分析：金融资产因果网络的研究 *we will share the research Causality Networks of Financial Assets: Through financial network analysis we ascertain the existence of important causal behavior among certain financial assets, as inferred by eight different causality methods. Our results contradict the Efficient Market Hypothesis and open new horizons for further investigation and possible arbitrage opportunities. Moreover, we find some evidence that two of the causality methods used, at least to some extent, could warn us about the financial crisis of 2007-2009. Furthermore, we test the similarity percentage of the eight causality methods and we find that the most similar pair of causality-induced networks is on average less than 50% similar throughout the time period examined, rendering thus the comparability and substitutability among those causality methods rather dubious. We also rank both the causal relationships and the assets in terms of overall causality exertion and we find that there is an underlying bonds regime almost monopolizing in some cases the realm of causality. Finally, we observe a recurring pattern of Oil's rising role as the financial network faces the Chinese stock market crash.*
	辅导课（5）：使用 Python 机器学习库 Scikit-learn 进行数据聚类分析 *Hands-on exercise: Clustering using python library scikit-learn*
第六周	**专业课（6）：** • 树型算法:回归树和分类树；递归二元分割和代价复杂度剪枝 *we will discuss two tree-based methods: regression trees and classification trees. We will also learn two very important algorithms that are used for constructing these trees: recursive binary splitting and cost complexity pruning.* • 装袋算法和随机森林 *we will discuss two very popular techniques, bagging and random forest, that can be used for aggregating several individuals trees and combining them into a single more accurate model. We will also learn how to estimate the test errors of the resulting models using the so-called "out-of-bag" error estimates.*
	辅导课（6）：使用 Python 图形库 Plotly 进行数据可视化 *Hands-on exercise: Data visualization using python library, plotly* **科研报告撰写辅导**

备注：以上课程时间安排，根据实际情况，可能会略有调整。

附件 2：天体物理与人工智能

课程概览

机器学习(Machine Learning)是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。

机器学习正在迅速地成为物理学家和化学家新的强大的工具，使他们能够不论在实验还是模拟过程中从大量的数据中提取必要的信息。通过采用、发展和应用机器学习的方法，可以以一种前所未有的方式分析高维复杂数据，促进物理科学每个分支上取得重大进展。

到目前为止，机器学习在物理科学上的大多数应用都局限于“低垂的果实”（ 可轻易实现的目标），因为它们大多集中于将已有的物理模型与数据进行拟合以及发现强信号。但我们相信机器学习也能够提供令人兴奋的机会，去学习模型本身。也就是说，去学习数据背后的物理原理和结构。并且在现实情况下，机器学习也将能够产生和设计复杂新奇的物理结构与目标。最终使物理学家不仅能够拟合他们的数据，而且获得有物理意义的模型，例如通过维持输入微观物理量预测的关系，以及通过有物理意义的约束，如守恒定律或对称关系。

不同领域的交流是相互的。从一开始，机器学习就受到了统计物理学方法的启发。许多现代机器学习工具，如变分推理和最大熵，都是物理学家发明的技术改进。物理学、信息理论和统计学密切相关，它们的目标都是从杂乱的数据中提取有效的信息。我们希望在数据中发现物理原理的特定背景下进一步推动学科交叉应用。

学习目标

课程结束后，学生将达到以下学习目标：

- 掌握数学建模、经典物理基础
- 理解黑洞成像理论
- 学会数据收集、数据分析理论，掌握贝叶斯概率的统计入门
- 机器学习入门，掌握 R/Python 等编程语言基础
- 掌握电磁学基础、原子核物理学基础、广义相对论基础
- 学会计对现有加速器与探测器公开数据的收集与整理
- 应用数据挖掘和机器学习相关知识对数据进行处理、分析、汇总展示
- 使用 LaTeX 整合研究成果、完成学术论文的写作

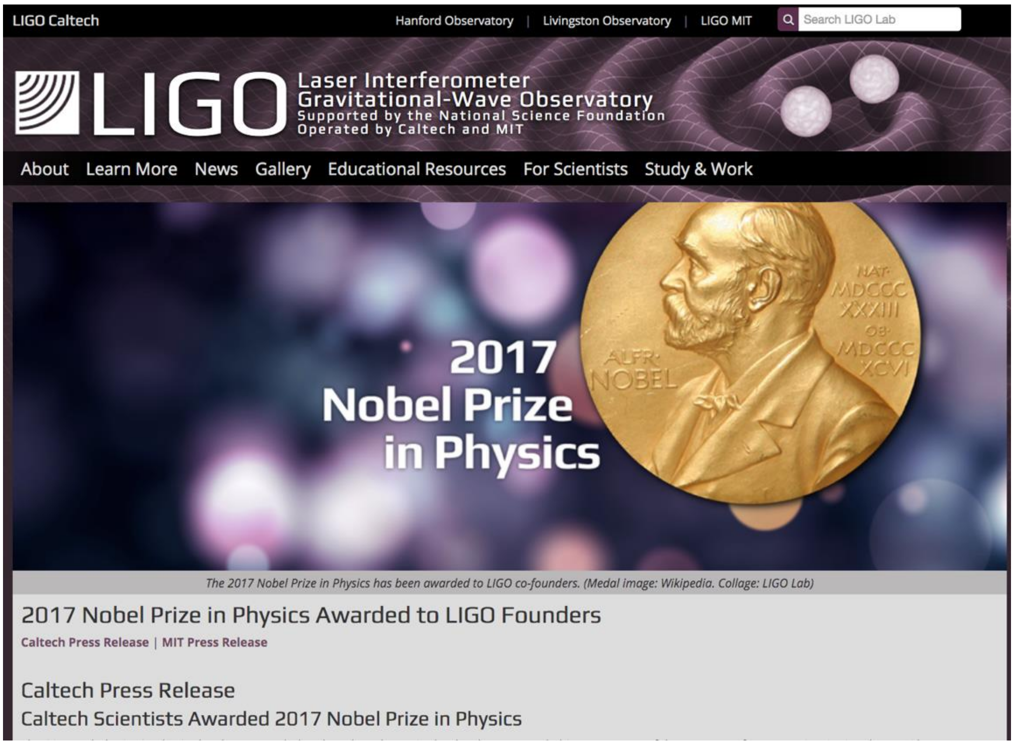
课程师资

陈教授（Y. Chen）
加州理工学院，终身教授，诺贝尔物理奖--引力波论文作者之一

他于 1999 年本科毕业于北大物理系，2003 年在加州理工学院获得博士学位。目前是加州理工学院物理学终身教授，美国物理协会〔American Physical Society, APS〕会士，LIGO 科学联盟核心成

员。他曾于马克斯·普朗克引力物理研究所（也称为**爱因斯坦研究所**）担任科研组组长；曾获得洪堡基金会最高荣誉奖，以表彰陈教授对引力波研究的特殊贡献；其研究获得国家科学基金会（National Science Foundation）支持。

2016 年 2 月 11 日，科研人员宣布激光干涉引力波观测站（LIGO）于 2015 年 9 月首次探测到引力波。该消息震惊世界，堪称科学史上的里程碑，学界甚至公众舆论都因此掀起一股热潮。引力波被称为“时空的涟漪”，探测引力波是一种全新的天文观测手段，是对**广义相对论和黑洞理论的直接验证**。爱因斯坦在广义相对论预言了引力波的存在：这是一种以光速传播的时空波动，是时空曲率的扰动以行进波的形式向外传递的一种方式。自从引力波被提出，人类探索它的努力就从未停止。20 世纪年代，对于脉冲星（或称波霎）双星系统 PSR1913+16 的观测间接证实了引力波的存在；90 年代起，美国启动 LIGO 计划；到 2015 年，第二代 LIGO 干涉仪成功观测到两个黑洞合并的引力波，也是首个距离地球 13 亿光年的引力波源 GW150914。



(2017 年诺贝物理学奖颁给引力波探测 LIGO 科学联盟)

项目日程

周数	内容
第一周	**专业课（1）：** • 学习狭义相对论的基础知识，并简要介绍微分。 *Study the basics of special relativity, and brief introduction to differential.* • 在牛顿引力中学习解析力学和粒子运动 *Study analytical mechanics and particle motion in Newtonian gravity.*
	辅导课（1）

第二周	**专业课（2）：** • 复习多元微积分和线性代数的基础知识 *Review the basics of multivariate calculus and linear algebra.* • 学习如何使用 Mathematica 绘图和做简单的计算 *Learn how to use Mathematica to make plots and do simple calculations.*
	辅导课（2）
第三周	**专业课（3）：** • 学习广义相对论的基础知识:度规，测地线，史瓦西时空，以及广义相对论在太阳系中的影响 *Study the basics of general relativity: metric, geodesics, the Schwarzschild spacetime, and effects of general relativity in the solar system* • 复习常微分方程的基础知识 *Review the basics of ordinary differential equations* • 学习如何使用 Mathematica 来解决常微分方程 *Learn how to use Mathematica to solve ordinary differential equations.*
	辅导课（3）
第四周	**专业课（4）：** • 在广义相对论中研究引力透镜和黑洞 *Study gravitational lensing and black holes in general relativity* • 构造 Mathematica 代码，追踪围绕史瓦西黑洞的光线 *Construct Mathematica code that trace light rays around a Schwarzschild black hole*
	辅导课（4）
第五周	**专业课（5）：** • 学会使用 LaTeX 来写科学论文。开始写一份研究报告，并在项目的其余部分不断更新 *Learn to use LaTeX to write scientific papers. Start writing a research report that will be updated throughout the rest of the project* • 追踪施瓦兹希尔德黑洞周围的光线，从 Gralla, Holz 和 Wald 的纸上的黑洞图像中恢复结果 *Trace light rays around Schwazshild black hole, recover results from paper by Gralla, Holz and Wald on the image of a black hole*
	辅导课（5）

第六周	**专业课（6）：** • 假设时空是史瓦西解的修正，找出黑洞的图像会有什么不同 *Suppose space-time is modified from the Schwarzschild solution, find out how the image of the black hole would differ* • 假设黑洞被暗物质的光环包围，并找出黑洞的图像会有什么不同 *Suppose black hole is surrounded by a halo of dark matter, and find out how the image of the black hole would differ*
	辅导课（6）及科研报告撰写

备注：以上课程时间安排，根据实际情况，可能会略有调整。