决战未来，胜在智能
哈尔滨工业大学基础学科拔尖人才培养
国际暑期学校实施方案

一、项目简介

为深入贯彻落实《关于加强基础学科人才培养的意见》的文件要求，走好基础学科人才自主培养之路，坚持面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康，全面贯彻党的教育方针，落实立德树人根本任务，大力培养造就一大批国家创新发展急需的基础研究人才，依托我学部在国内人工智能技术的优势和教学、科研的雄厚基础，承办“决战未来，胜在智能”国际暑期学校，拟举办时间为 2022 年 7 月 31 日-8 月 14 日，为期两周。旨在聚焦国际人工智能发展前沿知识以及基础学科重要研究方向，激发学生在人工智能领域的学习兴趣，加强国际和校际学生交流，帮助基础学科拔尖学生建立青年学术伙伴，为构建未来基础学科国际学术共同体奠定基础。

哈尔滨工业大学计算学部人工智能专业凝聚了计算机学科人 60 余年专业发展的结晶，承载着计算机应用技术数十年研究成果，依托其背后支撑的强大研究团队和教师队伍，汇集海外学成归来的新一批杰出学者代表，于 2019 年由教育部批准建立。该专业虽然最年轻，却规模大、实力强，有

着坚实的基础。人工智能专业独具特色，早在 1958 年便研制出了中国第一台“能说话会下棋”的数字计算机，被视为中国人工智能的起点，并逐渐形成了哈工大声图文特色的计算机学科。所依托的计算机科学与技术学科是国家首批重点学科一级学科，2011 年，进入 ESI 世界大学计算机学科排行榜前 1%，2012 年以来，多次在教育部学科评估中列全国计算机一级学科第 4 名，在最新的教育部计算机科学与技术学科评估中位列 A 档。人工智能专业拥有良好的教学和实践条件，拥有人机对话平台、无人机、智能感知等多种软硬件实践平台。计算学部已培养教育部部长、中科院院士怀进鹏，鹏程国家实验室主任、中国工程院院士高文，加拿大皇家科学院和工程院双院院士张大鹏等一大批人工智能人才，据脉脉数据统计，哈工大培养的人工智能人才数据国内高校之首。

本次暑期学校面向校内、C9/E9 高校以及与俄乌等合作高校本科生开设。在总结历届暑期学校经验的基础上丰富和凝练优秀传统，紧扣基础学科人才培养要求和目标，突出鲜明的人工智能主题，设计知识能力提升层层递进的教学环节，以人工智能领域课程讲授为引领，穿插拓宽国际视野的前沿讲座，组织理论联系实际的课题研究，全面提升学生实践能力。

本暑期学校将为学生学习了解人工智能提供各种机会：

与国际人工智能大师面对面交流机会。邀请图灵奖获得者 Joseph Sifakis 教授，IEEE Fellow、IEEE TIP 原主编、罗切斯特大学 Gaurav Sharma 教授，米兰理工大学 Francesco Amigoni 教授，莫斯科罗蒙诺索夫国立大学 Karabulatova Irina 教授四位国际知名学者及我学部四位中青年学者为学生授课和作报告，在传授人工智能领域前沿知识的同时，开拓学生国际视野，培养学生科学素养。

深度参与国际最前沿人工智能科学研究的机会。精心策划 9 个由我学部优势科研团队和骨干青年教师指导的创新实践项目。让学生在接触国际前沿知识的同时，提升项目研究的实践能力，增强团队合作意识，为校际学生交流切磋搭建良好的平台。

参与国际人工智能学科竞赛的机会。组织中俄大学生人工智能邀请赛，创造国际间学生切磋交流机会。

二、课程安排

1.总体安排

本次暑期学校分为四个环节：课程安排、前沿讲座、课题研究、学科竞赛和其他活动。

2.具体安排

具体安排	主讲人	职称/荣誉称号	单位	课程名称	学分	学时	简要介绍
课程安排（三选一）	Gaurav Sharma	IEEE FELLOW 教授 卓越科学数据科学领域的杰出研究员	罗切斯特大学	Graphical Models and Probabilistic Inference	1	16	图模型为机器学习提供了一个强大的框架，允许不同的因素有效地表示相互之间的依赖关系，使用有原则的概率方法进行推理和参数估计。这门课会提供图模型的介绍，这些模型已成为机器学习在计算机视觉、图像/视频处理、模式识别和分类、通信与纠错编码，生物信息学等多个不同领域的应用的标准工具。
	Francesco Amigoni	IEEE 高级会员 AAAI 会员 AI*IA 会员 教授	米兰理工大学	Multi-Agent and Multi-Robot Systems	1	16	智能体可被看作一个具有感知外部环境和自主 AI 能力的软件/硬件实体，而多智能体系统是一个在特定环境中进行交互的多个智能体所组成的计算系统，通过独立决策和彼此之间的协作完成特定任务。
	Karabulatova Irina	语言学博士 教授	莫斯科国立罗蒙诺索夫大学	Multimodal Communication Technologies of Manipulation in Mass Media and Mass Media and Issues of Information Security I	1	16	本课程将围绕网络舆论信息安全议题，介绍在大规模、自动化的网络信息监测、网络舆论交流引导场景中涉及的人工智能技术，包括多模态信息处理技术、自然语言处理技术、智能沟通技术等。

前沿讲座	Joseph Sifakis	图灵奖获得者教授		Machine Intelligence – Myths and Reality			基于对自主系统的研究，尝试对人类和机器智能进行多方面的比较，讨论比较人类和机器智能的现有标准的相关性和准确性，显示了科学知识和神经网络产生的知识之间的一些显著的类比和差异，讨论智能系统的社会影响，以及人类和机器之间协作的潜力，以推动知识开发和应用的前沿。
	张宏莉	哈工大计算学部网络空间安全学院院长，教授，博导，入选国家高层次人才计划、龙江学者特聘教授	哈工大计算学部	网络空间安全问题与挑战	1	4	围绕这个网络安全国家战略，介绍网络与信息安全国际形势，分析面临的网络空间安全主要威胁，介绍我国网络安全法律法规、网络安全内涵与技术发展现状、主要研究方向等。
	王忠杰	哈工大计算学部副主任、国家示范性软件学院院长，教授，博士生导师，中国计算机学会服务计算专委会副主任	哈工大计算学部	社会学视角的软件工程技术方案评估		4	本次讲座介绍软件技术之外的社会化因素，如何做全面的“社会化技术人”，如何打好社会化素质基础，特别是如何从社会学视角对软件工程技术解决方案对健康、安全、法律、隐私、文化、道德、伦理、就业、公平、环境、可持续发展等方面的影响进行全面评估。
	刘贤明	哈工大计算学部长聘教授，国家自然科学优秀青年基金获得者，哈工大青年科学家工作室学术带头人	哈工大计算学部	可信赖人工智能：理论与应用		4	如何构建值得信赖的人工智能系统成为学术界和工业界广泛关注的研究热点。本报告从不完备数据学习的角度，探讨深度学习在带噪声标签、不平衡数据分布、小样本、对抗样本和自监督条件下的鲁棒学习。
	车万翔	哈工大计算学部长聘教授，人工智能研究院副院长，社会计算与信息检索研究中心副主任，入选国家高层次青年人才计划	哈工大计算学部	自然语言处理新范式：基于预训练的方法		4	本次讲座将首先介绍预训练模型的演化过程，接着介绍预训练模型的最新研究进展，最后对自然语言处理领域今后的发展趋势进行了展望。
课题研究（九选一）	王亚东	哈工大计算学部教授、医学与健康学院执行院长、生物大数据教育部重点实验室主任，“中国十万人基因组计划”首席科学家，“十四五”国家重点研发计划“生物与信息融合（BT-IT 融合）”专项专家、“十三五”国家重点研发计划“生物安全”、“重大慢病”专项专家	哈工大计算学部	基于深度学习模型的中国人基因组大数据分析算法研究	1	1 周	针对中国参比人群全基因组测序数据（PB 量级），使用深度学习模型进行中国人基因组变异精确检测算法研究，创造快速准确的大规模基因组变异检测算法，并用于万分之一精度（世界最高精度）中国人基因组变异图谱的绘制。

	刘博	哈工大计算学部教授，生物大数据教育部重点实验室副主任，科技部“十四五”生物信息技术领域战略研究专家组成员	哈工大计算学部	DNA 数据存储高效编解码方法研究	1	1 周	本研究针对 DNA 数据存储中的核心技术—基于 DNA 序列的数据编解码方法为研究对象，重点研究在 DNA 存储框架下的数据高效表示、组织、索引相关数据结构与数据操作方法，建立高压缩比、高可靠性的 DNA 数据编解码方法。
	王宏志	哈工大计算学部长聘教授、博士生导师，英才学院副院长，海量数据计算研究中心主任，数据科学与大数据技术专业负责人，黑龙江省大数据科学与工程重点实验室主任	哈工大计算学部	Paxos 算法实现	1	1 周	Paxos 算法解决的问题是在一个可能发生上述异常的分布式系统中如何就某个值达成一致，保证不论发生以上任何异常，都不会破坏决议的共识。
	王宏志	哈工大计算学部长聘教授、博士生导师，英才学院副院长，海量数据计算研究中心主任，数据科学与大数据技术专业负责人，黑龙江省大数据科学与工程重点实验室主任	哈工大计算学部	机器学习算法自动推荐	1	1 周	随着目前数据的爆炸式增长，对于同一类任务有很多不同的算法被研究出来（例如：分类、回归），但是对于具有不同特点的数据集而言，并不是所有算法最终的性能都很好，算法之间最后的性能差异有很大的浮动。所以，如何根据数据集和算法的特点，为每个数据集尽可能地选择最适合它的算法成为目前的研究热点。
	王宏志	哈工大计算学部长聘教授、博士生导师，英才学院副院长，海量数据计算研究中心主任，数据科学与大数据技术专业负责人，黑龙江省大数据科学与工程重点实验室主任	哈工大计算学部	图数据的存储和索引	1	1 周	图数据有多种类型，比如 RDF 图和 native 图，结构上的不同造成了这两种图分别适合不同的存储方案和查询。比如 RDF 图上适合进行简单的连接查询，native 图上合适进行复杂的子图匹配操作。随着数据量的增加，我们越来越需要实现这些数据的高效存储。
	丁小欧	哈工大计算学部助理教授、师资博士后	哈工大计算学部	利用 AI 动态为大规模知识图谱建立索引	1	1 周	知识图谱的查询一直受到学术界和工业界的广泛关注，而其查询效率与知识图谱上建立的索引密切相关。而随着知识图谱数据的不断更新，以及其上查询工作负载的不断变化，最初建立的索引可能不能很好地满足更新后的知识图谱数据及其负载。因此，需要根据动态变化的数据及负载调整知识图谱索引。

	邹兆年	哈工大计算学部教授、博士生导师	哈工大计算学部	利用 AI 动态为大规模知识图谱选择存储结构	1	1 周	知识图谱的查询一直受到学术界和工业界的广泛关注，而其查询效率与知识图谱上建立的索引密切相关。而随着知识图谱数据的不断更新，以及其上查询工作负载的不断变化，最初建立的索引可能不能很好地满足更新后的知识图谱数据及其负载。因此，需要根据动态变化的数据及负载调整知识图谱索引。
	刘绍辉	哈工大计算学部副教授、博导	哈工大计算学部	视频目标检测与跟踪	1	1 周	初步学习 YOLOv5，DeepSort 算法的基本结构和运行环境配置，能够自己构建数据集，在已有的预训练模型上进行微调，然后能够应用已有算法构建基本的系统，在实际的场景中对目标能够检测和跟踪。
	刘绍辉	哈工大计算学部副教授、博导	哈工大计算学部	图像和视频信息隐藏算法设计	1	1 周	理解基本的图像和视频格式，能够操作图像和视频内容来隐藏相应的二进制信息，并能够具有一定的鲁棒性。

3.学科竞赛

组织中俄大学生人工智能创新邀请赛

本项比赛采用命题的方式：给定任务，学生可以使用百度提供的零门槛的 AI 开发平台 EasyDL 来进行开发，并对模型进行改进，最终根据任务的效果进行排名。使学生通过竞赛对人工智能相关的技术有更深入的了解。

4.其他活动

参观交流

参观哈工大计算机学院特色研究中心

哈尔滨工业大学校史博物馆、航天馆参观

哈尔滨城市规划展览馆、中央大街、太阳岛等著名景点参观

东北民俗体验

暑期学校成果交流会

三、课程与讲座详细信息

1.课程

（1）**Graphical Models and Probabilistic Inference**：图模型为机器学习提供了一个强大的框架，允许不同的因素有效地表示相互之间的依赖关系，使用有原则的概率方法进行推理和参数估计。这门课会提供图模型的介绍，这些模型已成为机器学习在计算机视觉、图像/视频处理、模式识别和分类、通信与纠错编码，生物信息学等多个不同领域的应用的标准工具。整个课程中使用其中一些应用程序以演示图模型在实践中的应用。本课程旨在为学生提供推理的基础以及使用概率模型进行估算的能力。

Gaurav Sharma　IEEE FELLOW，罗切斯特大学计算机科学、生物统计学、计算生物学、电气与计算机工程系教授，卓越科学数据科学领域的杰出研究员。 Gaurav Sharma 教授于 1992 年在印度科技学院获得通信工程硕士学位，1995 年在北卡罗莱纳州立大学获得应用数学硕士学位，1996 年在北卡罗莱纳州立大学获得电子计算机工程专业博士学位。于 2003 年 8 月开始在罗切斯特大学任职，2008 年 7 月担任电子成像系统中心主任，2009 年 7 月担任新兴和创新科学中心主任。主要的研究方向为彩色成像/图像处理、多媒体安全、生物信息学/基因组信号处理等。在这些领域中取得了丰厚的成果并得到世界各地同行的认可。曾担任施乐研发技术公司的首席科学家、彩色成像研发的项目负责人，拥有 52 项专

利。Sharma 教授是 IEEE 、SPIE 、 IS&T FELLOW。他也是 Sigma Xi，科学研究协会、Phi Kappa Phi 和 Pi Mu 荣誉学会的当选成员。

（2）Multi-Agent and Multi-Robot Systems：智能体可被看作一个具有感知外部环境和自主 AI 能力的软件/硬件实体，而多智能体系统是一个在特定环境中进行交互的多个智能体所组成的计算系统，通过独立决策和彼此之间的协作完成特定任务。为实现通用智能，AI 智能体必须学习如何在开放环境中与其他智能体进行互动。该类系统广泛存在于自动驾驶、机器人足球、智能家居和养老、智能仓储管理等典型场景中。本课程介绍多智能体和多机器人系统的定义和结构、多智能体与多机器人的基本互动形式（包括任务分配，协调等）、多智能体系统的分布式约束优化算法(包括问题定义、完全和近似算法设计等）、多智能体路径搜索算法（包括基于冲突的搜索、不完全算法/协作式 A*算法等），最后，介绍相应的编程框架和具体应用场景。

Francesco Amigoni 米兰理工大学教授，IEEE 高级会员，AAAI 会员，AI*IA 会员。主要研究领域为自主移动机器人、多主体系统等。2005 年获意大利人工智能协会青年研究人员“马尔科·索马维科”人工智能奖；2009 年获 IAT 最佳学生论文奖；2012 年获机器人世界杯救援模拟联赛虚拟机器人大赛冠军；2015 年获 Water Resources Planning and Management 期刊最佳研究论文奖；2016 年获 International

Workshop on Issues with Deployment of Emerging Agent-Based Systems 最佳论文奖；2017 年获 ISOCS/IEEEInternational Symposium on Olfaction and Electronic Nose 最佳论文奖；2018 年 International Conference on Intelligent Autonomous Systems 最佳论文提名。

（3）Multimodal Communication Technologies of Manipulation in Mass Media and Mass Media and Issues of Information Security：随着网络中大众媒体和社交媒体的快速发展，信息的传播自由度不断提高，往往能突破国界限制，造成大规模的舆论影响。近年来，各国网络空间中的群体舆情事件的发生频率变得越来越高，对大众媒体、社交媒体中的信息进行监控，对群体舆论加以引导，是当前重要的安全需求。本课程将围绕网络舆论信息安全议题，介绍在大规模、自动化的网络信息监测、网络舆论交流引导场景中涉及的人工智能技术，包括多模态信息处理技术、自然语言处理技术、智能沟通技术等。

Karabulatova Irina 语言学博士，教授，莫斯科罗蒙诺索夫国立大学人工智能和智能系统高级研究所机器学习和语义分析实验室首席研究员，曾荣获哈萨克斯坦共和国荣誉文化者称号，被称为“俄罗斯领先的心理语言学家”。她在语言理论、心理语言学和数字人文主义等领域都有着深入研究，著有 293 篇出版物，过去 20 余年中曾经主导或参与过多个以欧亚文化研究、移民语言演变为主题的项目，对于全

球化背景下的社会政治话语和民族文化演变都有着独特见解，荣获多项奖项。

2.前沿讲座

（1）Machine Intelligence-Myths and Reality：本报告基于对自主系统的研究，以及最近出版的分析信息、知识和智能之间联系的书籍，尝试对人类和机器智能进行多方面的比较。从什么是智能以及如何获得智能的问题出发，讨论人类智能和机器智能的相关性。此外，还比较了人类和机器所产生知识的准确性和有效性，揭示了科学知识和神经网络产生的知识之间显著的相似性和差异性。本报告强调自动化是从弱人工智能到通用人工智能的重要步骤，并展现了自主系统的特征，研讨自主系统与自动化系统的主要区别。展示了人类智力的某些方面是如何被心智系统所模拟的，即：一个具有解释智力行为的附加特征和机制的自主系统。此外，还确定了人类智能的两个特征功能，并强调了在与机器智能弥合差距方面的研究挑战。本报告最后展望未来，讨论智能系统的社会影响，以及人类和机器之间协作的潜力，以推动知识开发和应用的前沿。

Joseph Sifakis　中国科学院外籍院士、法国国家科研中心荣誉研究员、格勒诺布尔市 Verimag 实验室创始人、法国科学院院士、法国国家工程院院士、欧洲科学院院士、美国艺术与科学学院院士及美国国家工程院院士。研究领域主要包括系统设计的基本概念和应用，主要专注于系统设计的形

式化，即根据特定的要求实现可信赖、最优化且构造正确的系统。2007 年，Joseph Sifakis 被美国计算机协会(ACM)授予图灵奖，以表彰其在模型检查理论和应用方面做出的卓越贡献。模型检查是用数学算法来验证一个软件或硬件系统设计是否满足预设的需求，已应用于集成电路工业中，并且在嵌入式处理器和关键系统设计方面产生了重大影响。

（2）网络空间安全问题与挑战：围绕网络安全国家战略，介绍网络与信息安全国际形势，分析面临的网络空间安全主要威胁，介绍我国网络安全法律法规、网络安全内涵与技术发展现状、主要研究方向等。希望学生了解国际国内网络安全形势的严峻性，理解网络安全的基本概念，了解网络安全的技术体系和发展现状，建立正确的网络安全观。同时，提升学生网络安全素养和专业兴趣，突出课程思政的独特作用，筑牢学生的意志品格、政治素质和家国情怀。

（3）社会学视角的软件工程技术方案评估：身处当前“软件定义的社会”，软件工程师们面临着新规则、新挑战。在校计算机相关专业学生的未来职业道路如何铺就？你们的价值取向决定了未来社会、未来的软件行业的价值取向。本次讲座介绍软件技术之外的社会化因素，如何做全面的“社会化技术人”，如何打好社会化素质基础，特别是如何从社会学视角对软件工程技术解决方案对健康、安全、法律、隐私、文化、道德、伦理、就业、公平、环境、可持续发展等方面的影响进行全面评估，深入理解社会各因素，特别是

国家重大需求、政策法规对软件产业、软件企业、软件工程师的影响，从而能够敏锐地抓住软件创新的机会。

（4）可信赖人工智能：理论与应用：以数据驱动的深度学习为核心的人工智能技术飞速发展，在促进我们生活自动化方面发挥着越来越不可替代的作用。 如何构建值得信赖的人工智能系统成为学术界和工业界广泛关注的研究热点。本报告从不完备数据学习的角度，探讨深度学习在带噪声标签、不平衡数据分布、小样本、对抗样本和自监督条件下的鲁棒学习。

（5）自然语言处理新范式：基于预训练的方法：语言是人区别于动物的根本标志，具有无穷语义组合性、高度歧义性和持续进化性，准确处理自然语言是机器难以逾越的鸿沟，成为制约人工智能取得更大突破的主要瓶颈之一，也被誉为“人工智能皇冠上的明珠”。近年来以 BERT、GPT 为代表的、基于超大规模生语料库的预训练语言模型异军突起，充分利用大模型、大数据和大计算，使几乎所有自然语言处理任务性能都得到了显著提升，在若干公开数据集上宣称达到或超过了人类水平，成为了自然语言处理的新范式。本报告将首先介绍预训练模型的演化过程，接着介绍预训练模型的最新研究进展，最后对自然语言处理领域今后的发展趋势进行了展望。

四、课题研究详细信息

为充分发挥学部学科优势，让学生在接触前沿知识的基

础上提高实践能力，学部精心安排了生物信息、机器学习、大数据分析、图像处理等方向的实践项目，学生根据兴趣进行选择，通过理论与实践紧密结合的教学方式，力争使每名学生都有较大的收获。

项目 1：基于深度学习模型的中国人基因组大数据分析算法研究

分组方案：2-3 人/组，使用多种深度学习算法

题目来源：国家重点研发计划“精准医学研究 ”专项“中国十万人基因组计划”

“中国十万人基因组计划”是我国首个大规模人类基因组计划，旨在对中国参比人群开展基因组大数据采集和分析（PB 量级），绘制中国人基因组变异图谱，发现中国人特有变异，为我国精准医学、生物医药、生物安全等领域的未来发展奠定科学基础。

研究方案：针对中国参比人群全基因组测序数据（PB 量级），使用深度学习模型进行中国人基因组变异精确检测算法研究，创造快速准确的大规模基因组变异检测算法，并用于万分之一精度（世界最高精度）中国人基因组变异图谱的绘制。

项目 2：DNA 数据存储高效编解码方法研究

分组方案：2-3 人/组，使用不同 DNA 序列编解码方法

题目来源：国家重点研发计划“生物与信息融合（BT-IT 融合）”专项“基于 DNA 原理的高密度安全存储系统研发

与生物大数据应用示范”项目

“基于 DNA 原理的高密度安全存储系统研发与生物大数据应用示范”项目是我国“十四五”期间 BT-IT 融合科技发展的核心项目，其主要目标是研发基于 DNA 合成的高性能 DNA 数据存储技术，实现以 DNA 物质为基础的全新信息存储技术体系，创造未来信息科技发展的新基础。

研究方案：本研究针对 DNA 数据存储中的核心技术—基于 DNA 序列的数据编解码方法为研究对象，重点研究在 DNA 存储框架下的数据高效表示、组织、索引相关数据结构与数据操作方法，建立高压缩比、高可靠性的 DNA 数据编解码方法。

项目 3：Paxos 算法实现

分组方案：2-3 人/组

项目来源：国家自然科学基金重点项目，区块链的基础协议

研究方案：Paxos 算法是 1990 年由 Leslie Lamport 提出的一种基于消息传递且具有高度容错特性的共识算法。基于消息传递通信模型的分布式系统，不可避免地会发生以下错误：进程可能会慢、被杀死或者重启，消息可能会延迟、丢失、重复，在基础 Paxos 场景中，先不考虑可能出现消息篡改即拜占庭错误的情况。Paxos 算法解决的问题是在一个可能发生上述异常的分布式系统中如何就某个值达成一致，保证不论发生以上任何异常，都不会破坏决议的共识。一个典

型的场景是，在一个分布式数据库系统中，如果各节点的初始状态一致，每个节点都执行相同的操作序列，那么他们最后能得到一个一致的状态。为保证每个节点执行相同的命令序列，需要在每一条指令上执行一个“共识算法”以保证每个节点看到的指令一致。一个通用的共识算法可以应用在许多场景中，是分布式计算中的重要问题。因此从 20 世纪 80 年代起对于共识算法的研究就没有停止过。

要求：（1）设计一个小型伪分布式系统模拟 Paxos 算法进行实现并对你的系统进行压力测试；（2）系统的 proposer 应不少于 100，同时进行提案的 proposer 应不少于 20；（3）应尽可能提升并行性能与吞吐性能；（4）可以参考 https://en.wikipedia.org/wiki/Paxos_(computer_science)。

项目 4：机器学习算法自动推荐

分组方案：2-3 人/组

项目来源：国家自然科学基金重点项目，自动机器学习算法是人工智能平民化的必经之路。

研究方案：随着目前数据的爆炸式增长，对于同一类任务有很多不同的算法被研究出来（例如：分类、回归），但是对于具有不同特点的数据集而言，并不是所有算法最终的性能都很好，算法之间最后的性能差异有很大的浮动。所以，如何根据数据集和算法的特点，为每个数据集尽可能地选择最适合它的算法成为目前的研究热点。

构建一个模型（可以参考元学习的思想（meta-learning）），

使得其训练后可以：

A．针对分类任务，为新任务自动推荐最适合它的算法

B．针对回归任务，为新任务自动推荐最适合它的算法

C．尽可能提升整个模型的自动化程度，并在实验报告中说明详细。（例如：如果采用 meta-vector 来代表每个历史任务，那么设计一种方法可以自动从一个 candidate meta-feature list 中选择效果最好的 meta-features 来构成 meta-vector）

D．选择一个自己熟悉的主流大数据框架，使用这个框架实现上述功能，深入思考哪个部分可以采用并行处理，并提升其效率。不要为了使用框架而使用框架，如果使用框架效率降低，将实验结果写明，详细分析原因，写入实验报告。

说明：分类算法和回归算法的实现可以采用现有的算法库，例如：WEKA、sk-learn 等，A、B 两个任务每个任务待选择的算法不少于 15 种。实验数据集可以从 UCI 公开数据集下载 https://archive.ics.uci.edu/ml/index.php 自己进行适当的数据集格式转换。对于算法自动选择来说，不要走入“一个模型可以找到所有最优算法”的误区。我们设计的模型只能达到尽可能接近最优的算法。

除了工程文件和实验报告之外，你需要提交一份说明文档，描述你的算法设计、算法实现和实验结果。不必将思路局限于题目中提及的论文，可以考虑更合理的方式方法。

项目 5：图数据的存储和索引

分组方案：2-3 人/组

项目来源：国家自然科学基金重点项目，大图管理是当前大数据管理的核心内容之一。

研究方案：

（1）应用背景：图数据有多种类型，比如 RDF 图和 native 图，结构上的不同造成了这两种图分别适合不同的存储方案和查询。比如 RDF 图上适合进行简单的连接查询，native 图上合适进行复杂的子图匹配操作。随着数据量的增加，我们越来越需要实现这些数据的高效存储。

（2）数据来源：

http://swat.cse.lehigh.edu/projects/lubm/

http://github.com/facebook/linkbench

（3）实验要求：

1）熟悉图数据的基本概念，熟悉 RDF，RDF, RDFS, OWL，SPARQL 等基本概念。

2）请以上述两个数据集为基准，设计最佳的图存储方案来使得工作负载的执行效率最高。

3）考虑时下流行的分布式数据库 hugegraph、OrientDB、Cassandra 等（如果使用单机版数据库会影响你的成绩，可使用伪分布式进行实现）

4）请分析和对比究竟哪些因素影响了图存储，不同的工作负载类型都适合于怎样的存储方案。

项目 6：利用 AI 动态为大规模知识图谱建立索引

分组方案：2-3 人/组

项目来源：国家自然科学基金重点项目，知识图谱管理是人工智能的基础。

研究方案：

（1）项目背景：由于大数据的兴起，计算能力的升级，涌现出以知识图谱为代表的一批大数据时代的产物。知识图谱的查询一直受到学术界和工业界的广泛关注，而其查询效率与知识图谱上建立的索引密切相关。而随着知识图谱数据的不断更新，以及其上查询工作负载的不断变化，最初建立的索引可能不能很好地满足更新后的知识图谱数据及其负载。因此，需要根据动态变化的数据及负载调整知识图谱索引。

（2）数据来源：

DBPedia 知识图谱：https://wiki.dbpedia.org/develop/datasets

DBPedia 工作负载：http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/

（3）设计要求：利用所给 DBPedia 知识图谱及其负载，设计动态建立知识图谱索引的方法，要求：1）构造性能评估函数，以判断知识图谱及其负载对知识图谱查询性能的影响；2）利用 AI 方法动态感知知识图谱及其负载的变化是否对知识图谱查询性能产生影响；3）对于动态感知获得的结果，设计知识图谱索引调整算法，保证调整后知识图谱查询

性能的同时调整索引的代价不宜过大。

（4）实验要求：对所提出的方法进行合理实验评估，自行设计实验方法，记录 1）知识图谱的查询效率；2）索引存储空间；3）动态感知算法的效率。

项目 7：利用 AI 动态为大规模知识图谱选择存储结构

分组方案：2-3 人/组

项目来源：国家自然科学基金重点项目，知识图谱管理是人工智能的基础。

研究方案：

（1）题目背景：由于大数据的兴起，计算能力的升级，涌现出以知识图谱为代表的一批大数据时代的产物。知识图谱的查询一直受到学术界和工业界的广泛关注，而其查询效率与知识图谱的存储结构密切相关。而随着知识图谱数据的不断更新，以及其上查询工作负载的不断变化，最初选择的知识图谱存储结构可能不能很好地满足更新后的知识图谱数据及其负载。因此，需要根据动态变化的数据及负载调整知识图谱的存储结构。

（2）数据来源：

DBPedia 知识图谱：https://wiki.dbpedia.org/develop/datasets

DBPedia 工作负载：http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/

（3）设计要求：利用所给 DBPedia 知识图谱及其负载，

设计动态选择知识图谱存储结构的方法，要求：1）选用基于关系模型的知识图谱存储方式，参见参考文献（ http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=5841&journal_id=jos）；2）构造性能评估函数，以判断知识图谱及其负载对知识图谱查询性能的影响；3）利用 AI 方法动态感知知识图谱及其负载的变化是否对知识图谱查询性能产生影响；4）对于动态感知获得的结果，设计知识图谱存储结构调整算法，保证调整后知识图谱查询性能的同时调整存储结构的代价不宜过大。

（4）实验要求：对所提出的方法进行合理实验评估，自行设计实验方法，记录 1）知识图谱的查询效率；2）存储结构所占空间；3）动态感知算法的效率。

项目 8：视频目标检测与跟踪

分组方案：2-3 人/组，分别进行检测和跟踪系统的实现

题目来源：核九院科技发展中心。目标检测和跟踪系统在实际应用中有广泛的应用场景，例如机场低净空环境中需要对飞行物进行检测、跟踪和识别，并针对不同目标采取对应的措施来确保空域中无危险飞行物出现。这也是目前计算机视觉中应用最广泛的一项技术。

研究方案：初步学习 YOLOv5，DeepSort 算法的基本结构和运行环境配置，能够自己构建数据集，在已有的预训练模型上进行微调，然后能够应用已有算法构建基本的系统，在实际的场景中对目标能够检测和跟踪。

项目 9：图像和视频信息隐藏算法设计

分组方案：2-3 人/组，分别进行图像和视频信息隐藏算法的设计

题目来源：阿里云有限公司。针对目前数字媒体发行中所遇到的盗版、侵权等问题，设计相应的算法来进行对应的处理，包括隐藏信息进入载体中，对载体进行数字取证。

研究方案：理解基本的图像和视频格式，能够操作图像和视频内容来隐藏相应的二进制信息，并能够具有一定的鲁棒性。