重点实验室工作年度报告

（2022 年度）

实验室简介	上海市数据科学重点实验室是全国最早研究数据科学和大数据的科研机构之一，已成为国际数据科学研究的重要研究场所和数据科学人才培养基地，引领数据科学研究。实验室承担了一大批国家自然科学基金项目 、863 计划、973 计划、上海市科委项目、中国工程院咨询项目等,在大数据研究方面拥有丰富的经验,已经在金融证券、智能交通、智慧医疗等多领域构建了大数据挖掘平台。 实验室具备了良好的研究氛围和实验条件 ,凝聚了一大批从事大数据理论、技术和应用研究的专家学者,已形成多学科交叉的研究团队。建成了大规模计算环境和存储能力,储备了涵盖金融、生物医疗、交通、社会网络与舆情等多领域、跨学科的大数据资源。实验室已经成为上海大数据战略的技术研发和支持中心，经济社会发展的重要咨询机构，承担了上海市科委《大数据与云计算重大创新工程》、“大数据试验场”十三五战略规划、上海大数据试验场研发与转化功能型平台建设方案等编制工作，参与了科技部《大数据重大工程建议》、《面向 2030 重大科技项目：大数据重大项目实施方案》的编制工作。作为主要作者编著的中国工程院咨询研究报告《长三角大数据产业发展战略研究报告》上报国务院办公厅、上海市政府，为政府决策提供依据。原创提出的大数据试验场被列入 18 个上海市科创中心建设研发与转化功能型平台之一。大数据试验场的建设和使用，将在上海和国家大数据战略中发挥重要作用。作为副理事长单位，获批“大数据协同安全国家工程实验室”；作为副主任单位获批“大数据流通与交易技术国家工程实验室”。实验室在国际数据科学学术交流促进中具有重要的地位 ，创办了“International Workshop on

<table>
<tr><td></td><td colspan="3">Dataology and Data Science”、“International Conference on Data Science”和“数据科学家大会”。</td></tr>
<tr><td>主要研究方向</td><td>序号</td><td>研究方向</td><td>主要研究内容</td></tr>
<tr><td></td><td>1</td><td>数据科学基础理论</td><td>(a) 研究数据相似性理论;
(b) 研究数据测度和数据代数;
(c) 探索数据科学的研究方法</td></tr>
<tr><td></td><td>2</td><td>数据界探索</td><td>（a）数据基本规律研究;
（b）数据分类;
（c）数据界安全</td></tr>
<tr><td></td><td>3</td><td>数据技术及其应用</td><td>（a）科学研究的数据方法;
（b）领域数据学;
（c）大数据复杂性;
（d）大数据挖 掘技术;
（e）大数据应用</td></tr>
<tr><td>典型案例</td><td colspan="3">1.消防大数据：针对消防各类别警情高发区域、高发类型、高发时段分析难、统计难等
问题，开发了消防大数据分析应用平台，整合分析、展示各个街道消防警情、消防力量、以及物联网消防等数据，为消防指挥人员及物资调度等方面提供了有力保障。
2. 安全多方学习平台：与银联商务、蚂蚁集团、华为集团联合研发国内高校首个开源安全多方学习平台（https://github.com/FudanMPL），也是国内首个发表在安全四大顶会（S&P 和 CCS）上的用于多方训练的创新安全框架。实现基于 Shamir 秘密共享、加法秘密共享、向量空间秘密共享协议的多种安全多方计算算子，支持安全多方统计分析和线性回归、逻辑</td></tr>
</table>

	回归、BP 神经网络等主流机器学习模型的训练，支持决策树模型的安全推理，可用于金融风控、智慧医疗等隐私敏感场景。得到国家自然科学基金面上、上海市科委重点项目的直接支持。 3. 基于知识图谱的智能化技术体系：面向人工智能业界技术难题，与华为、阿里、字节、美团等头部企业携手攻关，在复杂知识表示、智能运维应用、多模态图谱构建、常识挖掘等方向取得突破性进展；面向我国国防与政务工作需求，与相关院校、研究院合作研究领域知识抽取、智能交互系统、辅助决策系统等原型系统，助力我国国防事业的自主可控发展；面向传统制造业企业，开展了基于工业知识图谱平台和工业认知的调研工作，并提出了一系列基于知识图谱的智能化解决方案。 4. CodeWisdom 软件智能化开发平台：针对软件开发中的知识问答需求，开发 CodeWisdom 软件智能化开发平台，基于 API、三方库、技术问答讨论等知识实现了软件开发问答机器人，支持面向开发人员的智能化问答，在多个企业中进行了应用，支撑了智能 IDE、软件开发资源推荐等智能化软件开发能力。 5. 软件开发大数据分析平台：针对企业代码质量和研发效能分析问题，持续提升软件开发大数据分析平台的系统能力，从平台稳定性、可扩展性等方面持续优化，形成代码自主可控分析、代码静态缺陷追踪、静态缺陷扫描规则定制推荐与修复优先级建议等多场景应用，在荣耀、亿通、中国电科 32 所等企业开展应用，取得了良好的效果。
代表性文章	1. BuildSheriff:Change-Aware Test Failure Triage for Continuous Integration Builds; 2. Learning Distinctive Margin toward Active Domain Adaptation; 3.E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning; 4. Language Models as Knowledge Embeddings;

	5. Gaviss：Boosting the Performance of GPU Accelerated NFV Systems via Data Sharing; 6. Characterizing Usages, Updates and Risks of ThirdParty Libraries in Java Projects; 7. A Progressive and Multi Prior Guided Network for Image Inpainting; 8. Alignment Work for Urban Accessibility: A Study of How Wheelchair Users Travel in Urban Spaces; 9. Efficient Consensus Motif Discovery of All Lengths in Multiple Time Series; 10.Online Summarizing Alerts through Semantic and Behavior Information;	
年终总结	研究成果	第一,承担了一批重大科研项目,包括国家重点研发计划、基金委重大或重点项目、上海市科委项目和企业委托项目共计128项,合同经费总额为 11473.68万元。 第二,110篇以研究人员为第一作者、通讯作者、署名为上海市数据科学重点实验室的学术研究论文发表在ICDM、KDD、WWW、IEEETrans. Knowl. Data Eng等具有影响力的国际会议和期刊上。 第三 ，出版专著和教材4本，新增36项专利。 第四，获得各种奖项共计8项。 第五,完成大数据试验场拟态试验条件扩容。
		第一，实验室培养在读硕士研究生594人、博士研究生133人，继续招收数据科学全日制研究生、专业博士研究生和硕士

	队伍建设与人才培养	研究生;此外,在站博士后 16 人。 第二，不仅培养本校学生,还积极吸引合作单位学生在大数据试验场平台上开展研究工作。
	开放交流与运行管理	第一,在开放课题方面,实验室继续资助 2020 年度申请的 2 项开放课题。 第二,在学术交流方面,实验室成员老师在各类学术会议以及论坛上做邀请报告 52 次。 第三,在仪器设备使用共享方面,实验室进一步丰富数据资源，拥有 2120 个 CPU 内核、271616 个 GPU 计算单元(CUDA cores)和 14605GB 内存总量，2035.1TB 存储能力，核心网络达 10Gb。 第三，在交流合作上，跟国家部委、市各委办、市区县和合作单位都建立了广泛联系和交流;为多家企业或相关项目组提供大数据技术与产品试验相关服务等。 第五，合理优化实验室管理制度，配备专职科研助理一名处理日常事务。
	依托单位支撑和保障情况	复旦大学作为实验室的依托单位,提供了良好的办公场地支持,给予了固定成员研究与办公场所，同时提供了相应的配套经费支持,以维护实验室日常运作。在引进人才方面,以开放联合方式,支持实验室发展。目前,实验室具备了良好的研究氛围和实验条件,凝聚了一大批从事大数据理论、技术和应用研究的专家学者,已形成多学科交叉的研究团队。