AI 前沿论文最新进展

2018.11.01 方建勇

1，研究强调了从细菌、动植物到城市等多种生物和社会有机体的规模增长与规模增长之间的关系。然而，迄今为止，在国家一级确定类似关系已证明具有挑战性。原因之一是，与前者不同的是，各国都有预先确定的边界，这限制了它们“有机”增长的能力。本文件通过确定和验证国家一级有机增长的一项有效措施来解决这一问题：夜间轻排放，这是进行更多生产性活动的能源分配的代表。将这一指标与人口规模进行比较，说明夜间的光排放量与超线性增长有关，而国家一级的人口规模则与次线性增长有关。然后利用过去三十年的高分辨率地理空间数据集探讨这些关系及其对经济不平等的影响。

2，我们将在 B.D，A.Pitkin，“随机波动率跳跃模型中期权定价的高阶紧致有限差分格式”中发展的方案推广到由 Duffie，Panand 和 Singleton 导出的所谓的带有同期跳跃的随机波动率(SVCJ)模型。通过大量的数值实验，通过与标准二阶中心差分格式的比较，对该格式的性能进行了评估。我们观察到，新的高阶紧致格式在提高效率和计算时间的同时，实现了三阶收敛。

3，这项工作适用于社区微型电网，在这种情况下，一个社区的实体可以彼此之间交换能源和服务，而不需要通过公共电网的通常渠道。

我们引入并分析了一个运行社区微电网的框架，并在参与单位之间分享由此产生的收入和成本。拟议的方法确保社区内每个实体所达成的解决办法不会比它单独采取行动所能达到的解决办法更差。因此，每个实体自然被激励在自愿的基础上参与社区活动。社区微电网运营商作为一个仁慈的规划者，根据具体的分享政策，在各实体之间重新分配收入和成本。通过这种方式，每个实体都可以直接从社区中受益，这要归功于资源的更有效分配、所支付的高峰电力的减少以及储备金总额的增加。一个执行边际定价方案的内部本地市场旨在确定社区价格。该框架以双层模型的形式制定，下层问题进行市场清算，而上层问题执行社区各实体之间的共享政策。数值计算结果证明了该方法的有效性。

4，本文提出了一种新的金融网络环境下的系统性风险度量方法。为此，我们给出了系统风险的定义，该定义基于在不同层次上发展起来的网络节点周围聚集邻居的结构。所提出的测度包含了有序聚类系数的广义概念。L 节点的我.在 Cerqueti 等人中介绍。(2018 年)。还从系统风险评估的角度对其性质进行了探讨。关于时变全球银行网络的实证实验表明，所提出的系统性风险度量方法的有效性，并提供了关于系统风险在过去几年中是如何变化的洞见，同时也考虑到最近的金融危机以及随后对具有全球系统重要性的银行实施更严格的监管。

5，本文对网格化的欧洲降水资料进行了分析。我们将简单线性回归与聚类等数据挖掘工具相结合，并利用现代自助方法对结果的强度进行了评价。65 年来，我们一直使用 0.5 级的日温度网格，这是由欧洲气候评估(EuropeanClimate 估测)创建的。我们已经通过改变线性回归的起点来检验结果的稳定性–这种方法可能对气候学家找到评估欧洲全球变暖的“最佳”起点很有价值。对不同的自举方法进行了比较，结果表明，相关加权自举法是检验估计量的最优方法。

6，河流水质监测越来越多地使用自动化的现场传感器，从而能够更及时地识别意想不到的值。然而，由于技术问题造成的异常使这些数据混淆，而数据的数量和速度阻碍了人工检测。提出了一种利用浊度、电导率和河流水位数据对现场传感器高频水质数据进行异常自动检测的框架。在识别最终用户需求和定义异常后，我们对它们的重要性进行了排序，并选择了合适的检测方法.高优先级异常包括突发性孤立尖峰和水平位移，其中大部分是用自回归综合移动平均模型等基于回归的方法正确分类的。然而，由于变量之间的复杂关系，使用其他水质变量作为协变量会降低性能.当我们用预报代替异常测量时，漂移和异常低或高变率周期的分类有所改善，但这种虚高的假阳性率。基于特征的方法在高优先级异常方面也表现良好，但在检测低优先级异常方面也不太熟练，从而导致了较高的假阴性率。与基于回归的方法不同，所有基于特征的方法都产生了较低的假阳性率，但不需要进行培训或优化。基于规则的方法成功地检测

出不可能的值和丢失的观测值。因此，我们建议使用多种方法来提高异常检测性能，同时最小化错误检测率。此外，我们的框架强调了终端用户和分析人员之间的通信对于检测性能和最终用户需求的最佳结果的重要性。该框架适用于其他类型的高频时间序列数据和异常检测应用。

7，临床医生和研究人员对如何更好地个性化干预越来越感兴趣。动态治疗方案(DTR)是一系列预先指定的决策规则，可用于指导针对个人不断变化的需求而制定的一系列治疗或干预措施的交付。序贯多分配随机试验(SMART)是一种允许构建有效 DTRs 的研究工具。我们推导出计算三种常见的两阶段智能设计的总样本大小的易于使用的公式，其中主要目的是使用在整个研究中收集的连续重复测量结果来比较两个嵌入式 DTRs。我们表明，SMART 的样本大小公式可以写成标准的两臂随机试验的样本大小公式的乘积，一个通货紧缩因素解释了重复测量分析导致的统计效率的提高，另一个通货膨胀因素解释了 SMART 的设计。智能设计膨胀系数通常是对第一阶段治疗的预期响应概率的函数。我们回顾了使用 SMART 的重复测量结果进行 DTR 效果分析的建模和估计，以及标准误差的估计。我们还给出了各种常用工作相关结构的重复测度协方差矩阵的估计量。该研究旨在开发一种 DTR，以提高酒精和可卡因依赖患者参加治疗的动机。

8，本文档提供了关于 MDL 原理在复杂图形分析中的使用的教程描述。本文以图中最大团的大小分析为例，对初步主题作了简要的总结，并描述了基本原理。我们还讨论了如何解释这种分析的结果，并指出了几个常见的缺陷。

9，杠杆分数抽样为大型矩阵的近似计算提供了一种有吸引力的方法。实际上，它允许推导出适合于当前问题的复杂程度的忠实近似。然而，进行杠杆分数抽样本身就是一个挑战，需要进一步的近似。本文研究了核定义的正定矩阵的杠杆评分抽样问题。我们的贡献是双重的。首先，我们提出了一种新的利用评分抽样的算法，其次，通过推导出一种新的核岭回归求解器，开发了统计学习中的一种新方法。我们的主要技术贡献是表明所提出的算法是目前解决这些问题的最有效和最准确的算法。

10，我们将自适应重要性抽样(AIS)作为一个在线学习问题进行了研究，并论证了在这一适应过程中，探索与开发之间的权衡的重要性。借鉴土匪文献的思想，提出了一种基于分割的 AIS 算法–Daisee.我们进一步引入了 AIS 的遗憾概念，并证明了 Daisee□(T 型¬√(原木 T 型)34)累积伪后悔 T 型迭代次数。然后，我们将 Daisee 扩展到自适应地学习样本空间的分层划分，以便进行更有效的采样，并通过经验验证这两种算法的性能。

11，本文介绍了一种新的非监督的时间序列数据降维和信号分解方法–对比多元奇异谱分析方法。通过使用适当的背景数据集，该方法以一种方式转换目标时间序列数据集，以提取目标数据集中增强的子信号，而不是仅反映最大方差的子信号。这就把目标从找到解释方差最大的信号转移到了对分析师最重要的信号上。我们在一个示例上演示了我们的方法，并展示了该方法在公共 mHealth 数据集的心电图信号的下游聚类中的应用。

12，随着物联网(物联网)设备数量的不断增加，通过物联网设备产生的数据也越来越多。据预测，到 2025 年，物联网设备的数量将超过地球上的人类数量。因此，通过这些物联网设备产生的数据将是巨大的。这使得储存量激增。最有希望的解决方案之一是在云上存储数据。市场对物联网云平台的数量已无计可施。尽管有各种各样的物联网云平台可供使用，但在对其进行分类或比较以供开发的应用程序方面，在整个文献数据库中很少有尝试。本文将物联网平台分为四大类：公开交易、开源、开发人员友好和端到端连接。根据给定的一般物联网体系结构，识别并比较了每个类别中的一些流行平台。本研究对物联网的新手和应用开发人员根据构建应用的需求选择最合适的平台具有一定的参考价值。

13，快速的城市化和日益增长的交通对世界大都市地区产生了严重的社会、经济和环境影响。了解道路网络与交通条件之间复杂的相

互作用，具有十分重要的意义。提出了一种利用 GPS 跟踪估计大城市交通状况的新框架。通过求解一个基于交通流理论的凸优化规划，对网络行程时间进行了初步估计。然后，迭代地细化估计的网络旅行时间和车辆通过路径。最后，作者采用双层优化方法对 GPS 数据不覆盖的路段的交通状况进行了估计。作者对两种最先进方法的评价和比较表明，相对改进达 96.57%。作者通过将旧金山和北京的道路网络与实际地理信息系统(GIS)数据相结合，进一步进行了现场试验，这些数据涉及 128，701 个节点，148，899 个路段，以及超过 2，600 万条 GPS 轨道。

14，为了了解内容选择是如何进行的，我们在新闻、个人故事、会议和医学文章等领域进行了深度学习模型的实验。我们发现，与简单的模型相比，最先进的提炼总结器的许多复杂特性并不能提高性能。这些结果表明，为一个新领域创建一个摘要器比以前的工作更容易，并使那些拥有大量数据集(即新闻)的领域的深度学习模型的好处受到质疑。同时，他们也提出了新的总结研究的重要问题，即需要新的句子表达形式或更适合于摘要任务的外部知识来源。

15，我们在这里研究政党成员的行为，目的是确定意识形态团体是如何在不同的(支离破碎和非支离破碎的)政党系统中随着时间的推移而产生和演变的。利用巴西和美国的公共投票数据，我们提出了一种方法来识别和描述意识形态团体、他们的成员两极分化，以及

这些群体是如何在 15 年的时间内演变的。我们的结果揭示了两个案例研究的非常不同的模式，无论是在结构和动力学性质上。

16，我们分析了自动帐户或机器人对社交网络中的意见的影响。我们使用著名的 DeGroot 模型的一个变体对意见进行建模，该模型将意见与网络结构联系起来。我们发现，基于这一网络模式的观点与 Twitter 用户讨论 2016 年希拉里·克林顿(Hillary Clinton)和唐纳德·特朗普(Donald Trump)总统大选的推文之间存在着很强的相关性，为这一模式的有效性提供了证据。然后，我们利用网络模型来预测，如果网络不包含任何可能试图操纵意见的机器人，那么这些意见会是什么。

利用一种 BOT 检测算法，我们识别了占网络不到 1%的 BOT 帐户。通过分析机器人帖子，我们发现支持唐纳德·特朗普的机器人数量是支持希拉里·克林顿的两倍。我们从网络中删除机器人，并使用网络模型重新计算意见。我们发现，机器人的观点发生了重大变化，克林顿机器人的变化几乎是特朗普机器人的两倍，尽管数量较少。对机器人行为的分析表明，这种巨大的变化是由于机器人的移动频率是人类的一百倍。这种观点转变的不对称性是因为克林顿的机器人比特朗普的机器人高出 50%。我们的结果表明，社交网络中少数高度活跃的机器人会对意见产生不成比例的影响。

17，世界正见证着网络物理系统(Cps)的空前增长，这将为我们的世

界带来革命性的变革，在环境监测、移动卫生系统、智能交通系统等多个领域创造新的服务和应用。{信息和通信技术}(信通技术)部门由于智能手机、平板电脑和视频流的广泛使用以及预期在不久的将来部署传感器的显著增长，{数据}流量正在显著增长。{it}预计将显著提高原始感知数据的增长率。本文介绍了 CPS 分类法{VIA}，对数据的收集、存储、访问、处理和分析作了概述。与其他调查文件相比，这是第一次对 CPS 的大数据进行全景调查，我们的目标是对 CPS 的各个方面进行全景总结。此外，CPS{需要}网络安全来保护{他们}免受恶意攻击和未经授权的入侵，这将成为网络中不断生成大量数据的一个挑战。{因此，我们还}概述了为 CPS 大数据存储、访问和分析而提出的各种安全解决方案。我们还讨论了在 CPS 环境下迎接绿色挑战的大数据。

18，将数据汇总为文本有助于人们理解它。它还改进了数据发现，因为搜索算法可以将文本与关键字查询匹配。在本文中，我们探讨了数据的文本摘要的特点，以了解有意义的摘要是什么样子。我们提出了两项补充性研究：一项有 69 名学生参加的数据搜索日记研究，该研究提供了对搜索数据的人的信息需求的洞察；以及一项摘要研究，包括一个实验室和一个众包组件，共有 80 名了解数据的参与者，该研究为 25 个数据集编写了摘要。在每一项研究中，我们都进行了定性分析，以确定关键主题和常见的数据集属性，人们在搜索和理解数据时都会考虑这些问题。这些结果帮助我们设计了一个模板，

以创建更有意义的数据文本表示，以及改进总体数据搜索体验的指导方针。

19，神经网络已经成功地应用于激光雷达数据的处理，特别是在自动驾驶场景中。然而，现有的方法在很大程度上依赖于对激光雷达传感器产生的脉冲信号进行预处理，从而导致计算开销大、延迟大。本文提出了一种利用尖峰神经网络(SNN)直接解决原始时间脉冲的目标识别问题的方法。为了帮助评估和基准制定，建立了一个综合的时间脉冲数据集，以模拟激光雷达反射在不同的道路场景。对不同噪声条件下的识别精度和时间效率进行了测试，结果表明，该方法具有较好的识别性能，推理准确率可达 99.83%(含 10%噪声)，平均识别延迟为 265 ns。重点介绍了 SNN 在自主驾驶中的应用前景及相关应用。特别是，据我们所知，这是第一次尝试使用 SNN 直接对原始 Lidar 时间脉冲进行目标识别。

20，PDF 格式的流行和 PDF 查看器提供的丰富 JavaScript 环境使 PDF 文档成为恶意软件开发人员的一个有吸引力的攻击矢量。PDF 文档对组织的安全构成严重威胁，因为大多数用户都不怀疑它们，因此很可能从不受信任的来源打开文档。我们建议使用保守的抽象解释来识别恶意 PDF，从而静态地解释嵌入式 JavaScript 代码的行为。目前，最先进的工具有：(1)基于已知恶意示例的结构相似性静态识别 PDF 恶意软件；或(2)动态执行代码以检测恶意行为。这两种

方法都会受到规避攻击，这些攻击模仿良性文档的结构，或者在动态分析时不显示它们的恶意行为。相反，抽象的解释忽略了两种类型的回避。与两种最先进的 PDF 恶意软件检测工具的比较表明，我们保守的抽象解释方法实现了类似的准确性，同时对规避攻击具有更强的抗御能力。

21，最近发生的数据泄露事件要求各组织主动识别对其系统的网络攻击。Darweb/Deepweb(D2web)论坛和市场提供了黑客匿名讨论现有漏洞并将恶意软件商业化以利用这些漏洞的环境。这些平台为安全从业者提供了一个威胁情报环境，允许挖掘与组织定向网络攻击相关的模式。在本文中，我们描述了一个系统(称为 DARKMENTION)学习关联规则的攻击指标从 D2web 到现实世界的网络事件。使用所学的规则，DARKMENTION 在攻击之前生成并向安全操作中心(SOC)提交警告。我们的目标是设计一个系统，自动生成具有企业针对性的、及时、可操作、准确和透明的警告。我们证明 DARKMENTION 符合我们的目标。特别是，我们表明，它的性能优于基线系统，这些系统试图生成与两家企业相关的网络攻击警告，F1 平均得分分别提高了 45%和 57%。此外，DARKMENTION 是作为一个更大的系统的一部分来部署的，该系统是根据 IARPA 网络攻击自动非常规传感器环境(原因)程序建立的。它正在积极地发出攻击前平均 3 天的警告。

22，在一个给定的场景中，同时准确地预测交通参与者的每一个可能的交互是自动车辆的一个重要能力。目前的研究大多集中在单一实体的预测上，而不包含环境信息。虽然有些方法的目的是预测多辆车，但它们要么是独立地预测每一辆车，而不考虑与周围实体的可能交互，要么是产生离散的联合运动，无法直接用于自主车辆的决策和运动规划。本文提出了一种概率框架，能够联合预测多个相互作用的道路参与者在任何驾驶场景下的连续运动，并能够预测每次相互作用的持续时间，从而提高预测的性能和效率。所提出的交通场景预测框架包括两个层次模块：上模块和下模块。上面模块预测车辆的意图，下模块预测交互场景实体的运动。使用一个示例现实世界场景来实现和检查所提议的框架。

23，在危险危机期间的应急管理领域，拥有足够的情况意识信息是至关重要的。它需要从卫星图像、当地传感器和当地居民生成的社交媒体内容等来源获取和整合信息。捕获、表示和集成这些异构和多样化信息的一个大胆障碍是缺乏适当的本体来正确地对该领域进行概念化、聚合和统一数据集。因此，本文引入了移情本体论，对危机应急管理和规划领域的核心概念进行了概念化。虽然 Pempathi 有一个粗粒度的视图，但它认为必要的概念和关系在这个领域是必不可少的。此本体可在 https://w3id.org/empathi/.上使用。

24，类不平衡是包括深度学习在内的分类模型中普遍存在的问题，

其提取任务特征的能力在不平衡的环境中受到影响。然而，在以往的研究中，处理大量班级之间的不平衡的挑战，通常都是通过深度学习来解决的，并没有得到足够的重视。本文提出了一种深度过采样框架的扩展，利用自动生成的抽象标签，即弱标记学习中使用的一种侧信息，增强了针对班级不平衡的深度表示法学习。我们试图利用标签将实例的深度表示引导到不同的子空间，从而导致分类问题固有子任务的软分离。我们的实证研究表明，该框架在大类和小类之间不平衡的情况下，对图像分类基准有了很大的改进。

25，我们介绍了 DeepGRU，一种基于深度学习的手势和动作识别器。我们的方法看似简单，但适用于各种应用场景。DeepGRU 只使用原始的姿态和矢量数据，无论数据集的大小、训练样本的数量或输入设备的选择，都能获得较高的识别精度。我们的方法的核心是一组堆叠的 Grus，两个完全相连的层和一个全局注意模型。我们证明，在缺乏强大的硬件和只使用 CPU 的情况下，我们的方法仍然可以在很短的时间内被训练，这使得它适合于快速原型和开发场景。我们在 7 个公开的数据集上评估我们提出的方法，这些数据集涵盖了广泛的交互以及数据集的大小。在大多数情况下，我们的表现优于先进的基于姿态的方法。对 NTU RGB+D 数据集的交叉测试和交叉视测的识别准确率分别为 84.9%和 92.3%，在 UT-Kinect 数据集上的识别准确率为 100%。

26，目前最先进的视频理解方法采用时态抖动来模拟分析不同帧速率的视频。然而，对于多速率视频来说，这并不是很好的工作，因为在多速率视频中，动作或子动作以不同的速度发生。帧采样率应根据不同的运动速度而变化。在这项工作中，我们提出了一个简单而有效的策略，称为随机时间跳过，以解决这种情况。该策略通过在训练过程中随机化采样率来有效地处理多速率视频。这是一个详尽的方法，它可以潜在地涵盖所有的运动速度变化。此外，由于大量的时间跳过，我们的网络可以看到原来超过 100 帧的视频剪辑。这样的时间范围足以分析大多数动作/事件。我们还介绍了一种基于遮挡感知的光流学习方法，该方法可以生成改进的运动地图，用于人类的动作识别。我们的框架是端到端的培训，实时运行，并在六个广泛采用的视频基准上实现了最先进的性能。

27，集成聚类一直是数据挖掘和机器学习领域的一个热门研究课题。尽管近年来取得了很大的进展，但目前的集成聚类研究仍存在两个具有挑战性的问题。首先，现有的算法大多倾向于在对象级对集成信息进行研究，但往往缺乏在较高粒度层次上挖掘丰富信息的能力。二是多基地聚类中的直接联系(如直接相交或成对共现)，而忽略了其中隐含的多尺度间接关系。针对这两个问题，本文提出了一种基于随机游走聚类相似性快速传播的集成聚类方法。首先构造一个聚类相似度图，将基簇作为图节点，利用聚类 Jaccard 系数计算初始边权值。在构造的图上定义了一个转移概率矩阵，在此矩阵的基础

上进行随机游动过程来传播图的结构信息。具体来说，通过研究从不同节点开始的传播轨迹，通过考虑轨迹关系，可以得到一个新的聚类相似矩阵。然后将新获得的聚类相似性矩阵从聚类级映射到对象级，实现增强的共关联矩阵，既能同时捕获对象间的共现关系，又能在集群中捕获多尺度的聚类关系。最后，提出了两种新的共识函数来获得一致性聚类结果。对各种真实世界数据集的广泛实验证明了我们的方法的有效性和效率。

28，机器翻译正在转向一种基于深度神经网络的端到端方法.在流行的语言对，如英、法、英、汉等方面，最先进的技术都取得了令人印象深刻的成绩。然而，对于英语和越南语来说，平行语料库的短缺和昂贵的超参数搜索给基于神经的方法带来了实际挑战。本文着重从两个方面对英语–越南语翻译进行了改进：(1)建立了迄今为止最大的开放越南语–英语语料库；(2)利用最新的神经模型进行了广泛的实验，获得了最高的 BLEU 分数。我们的实验提供了有效使用低资源语言对的不同神经机器翻译模型的实例。

29，声学到单词模型是端到端语音识别器，它使用单词作为目标，而不依赖于发音词典或字素。由于缺乏语言知识，这些模型难以训练。还不清楚训练数据的数量如何影响这些模型的优化和概括。在这项工作中，我们研究了在不同数量的训练数据下声学到单词模型的优化和推广。此外，我们研究了三种类型的归纳偏差，利用发音

字典，字边界注释和对字持续时间的约束。我们发现限制单词持续时间会带来最大的改进。最后，我们分析了模型学习的嵌入空间这个词，发现空间有一个由单词发音支配的结构。这表明，词语的语境，而不是语音结构，应该成为声学 – 词汇模型中归纳偏差的未来焦点。

30，噪声和伪像是低剂量 CT（LDCT）数据采集所固有的，并且将显着影响成像性能。由于统计和技术的不确定性，完美的噪声去除和图像恢复在 LDCT 的背景下是难以处理的。在本文中，我们将生成对抗网络（GAN）框架与视觉注意机制一起应用，以数据驱动/机器学习的方式处理这个问题。我们的主要想法是将视觉注意力知识注入到 GAN 的学习过程中，以提供强大的噪声分布先验。通过这样做，发生器和鉴别器网络都被赋予视觉注意信息，因此它们不仅要特别注意噪声区域和周围结构，而且还要明确地评估恢复区域的局部一致性。我们的实验定性和定量地证明了所提出的方法与临床 CT 图像的有效性。

31，物联网正在迅速发展，现在消费者可以使用许多连接设备。随着这种增长，从智能手机管理设备的物联网应用程序引发了重大的安全问题。通常，这些应用程序通过敏感凭据（如电子邮件和密码）进行保护，这些凭据需要通过特定服务器进行验证，因此需要访问 Internet 的权限。不幸的是，即使开发人员用心良苦，这样的应用程

序也可以保证安全，以保证用户的凭据不会泄漏到 Internet 上的未授权服务器。例如，如果应用程序依赖于第三方库，那么这些库可能会捕获并泄漏敏感凭据。应用程序中的错误还可能导致可泄漏凭据的可利用漏洞。本文介绍了我们正在进行的原型工作，该原型使开发人员能够控制应用程序中的信息如何从敏感 UI 数据流向特定服务器。我们扩展 FlowFence 以对敏感的 UI 数据实施细粒度的信息流策略。

32，关于妇女在学术界的作用的辩论集中在可能成为许多国家所见性别差距根源的各种现象。然而，尽管科学研究的协作性越来越强，但研究合作中的性别问题问题却得到了一定程度的处理。在本文中，我们采用创新的文献计量方法，基于个别学者的合作倾向，允许通过领域，学科和合作形式来衡量性别差异：校内，校外国内和国际。对意大利学者科学生产的分析表明，除了国际合作之外，女性研究人员在所有分析形式中都有更大的合作能力，与男性同事相比仍然存在差距。

33，本研究通过使用从附加扩展实时数据的智能对象（SO–ERD：例如智能容器，智能托盘等）获取的链可追溯性数据来提出应用功能的含义，以改善物流层面的风险管理链。最近使用可追溯性数据的应用程序和可追溯性系统中的主要问题已经被学术文献探索过。信息根据当前可追溯性数据的使用情况进行分类，以支持风险检测

和运营，战术和战略层面的决策。研究发现，实时数据对所有决策层面的运输活动的使用产生了重大影响，如食品质量控制和合作伙伴之间的协作计划。但是，各种合作伙伴捕获的基于事件的可追溯性数据的汇总存在一些不确定性，这些数据阻止了链的数据使用。在工业 4.0 和物联网（IoT）的环境下，SO–ERD 可以实时地通过链进行独立的数据跟踪。其数据有可能克服当前问题并改善供应链风险管理。因此，基于文献综述，提出了 SO–ERD 数据的使用，提出了风险管理的含义，揭示了供应链中当前对决策职能的关注。其影响可能是对域需求的影响。

34，急性卒中病变分割和预测任务具有重要的临床意义，因为它们可以帮助医生做出更明智的时间关键治疗决策。由于它们的异质外观，动态进化和患者间差异，这些病变的自动分割是一项复杂的任务。通常，用针对慢性中风或其他脑损伤开发的方法来处理急性中风病变任务。然而，急性卒中的病理生理学和解剖学确立了一个需要特别考虑的固有不同的问题。在这项工作中，我们提出了一种新颖的深度学习架构，专门用于急性脑卒中任务，包括用减少的数据逼近复杂的非线性函数。在我们的策略中，使用基于最先进的列车采样策略的混合策略来解决类别不平衡，该策略针对其他脑部病变相关任务而设计，更适合于急性卒中病变的解剖学和病理生理学。所提出的方法在三个不相关的公共国际挑战数据集（ISLES）上进行评估，而没有任何特定于数据集的超参数调整。这些涉及亚急性

卒中病变分割，急性卒中半影估计和急性 MR 图像的慢性范围预测的任务。针对慢性卒中和相关生物医学任务的类似深度学习架构，以及针对每个挑战提交分段测试图像进　　行盲目在线评估，分析了所提出的架构的性能。与其他提交的策略相比，我们的方法在所有三个挑战中的最佳提交条目中实现了最高级别的性能，显示了在没有超参数调整的情况下处理不同的无关任务的能力。为了提高我们结果的可重复性，已经发布了该方法的公开版本。

35，机器学习技术深深植根于我们的日常生活中。然而，由于追求良好的学习成绩是知识和劳动密集型的，因此人类专家大量参与机器学习的各个方面。为了使机器学习技术更容易应用并减少对经验丰富的人类专家的需求，自动机器学习～（AutoML）已成为工业和学术界的热门话题。在本文中，我们提供了对现有 AutoML 工作的调查。首先，我们介绍和定义 AutoML 问题，灵感来自自动化和机器学习领域。然后，我们提出了一个通用的 AutoML 框架，它不仅涵盖了几乎所有现有方法，还指导了新方法的设计。之后，我们从两个方面对现有作品进行分类和审查，即问题设置和使用的技术。最后，我们提供了 AutoML 方法的详细分析，并解释了其成功应用程序的原因。我们希望这项调查不仅可以作为 AutoML 初学者的深刻见解，也可以作为未来研究的灵感来源。

36，使用自然语料库对语义的深度学习模型的标准评估在他们可以

告诉我们关于学习表示的保真度方面是有限的，因为语料库很少具有良好的语义复杂度。为了克服这个限制，我们提出了一种生成多重量化自然语言推理（NLI）数据集的方法，其中可以精确地表征语义复杂性，并且我们使用这种方法来表明 NLI 的各种常见体系结构不可避免地会失败编码重要信息;只有具有强制词汇对齐的模型才能避免这种破坏性的信息丢失。

37，功能数据通常包含振幅和相位变化。在许多数据情况下，相位变化被视为一种干扰效应，并在预处理过程中被消除，尽管它可能包含有价值的信息。本文主要研究振幅和相位变化的主成分联合分析(PCA)。

由于翘曲函数的空间具有复杂的几何结构，分析的一个关键要素是将翘曲函数转换为 L2(□)。我们介绍了不同的转换方法，并展示了它们如何适合于一般的转换类。这使我们能够比较它们的长处和局限性。在 PCA 的背景下，我们的结果提供了支持中心对数比变换的论据.

我们进一步嵌入了 Hadjipantelis 等人的现有方法。(2015)和 Lee 和 Jung(2017)将振幅和相位变化的联合主元分析引入多元泛函主成分分析框架，在此基础上，我们研究了基于适当度量的估计量的性质。通过地震学的应用说明了这种方法。

38，我们在基于比较的设置中考虑分类问题：给定一组对象，我们

只能访问表单“对象”的三重奏比较。X 我.更接近对象 XJ 而不是反对 X 钾“.”本文介绍了一种新的分类器学习方法 TripletBoost。其主要思想是将三叉树信息聚合为弱分类器，然后将其提升为强分类器。该方法有两个主要优点：(1)适用于任何度量空间的数据；(2)只使用被动获取的和有噪声的三重态来处理大规模问题。我们得到了理论推广的保证和所需三重子数的下界，并在经验中证明了我们的方法既能与最先进的方法竞争，又能抵抗噪声。

39，当在人群中进行流行病学研究时，选择偏差会威胁到因果推理的有效性。这种偏见可能会发生，不管选择的人群是否是目标人群，甚至在没有暴露–结果混淆的情况下也可能发生这种偏差。然而，往往很难量化选择偏差的程度，而敏感性分析很难进行和理解。在本文中，我们证明了由于选择而产生的偏差的大小可以用由参数定义的简单表达式来限制，这些参数描述了对偏差负责的未测量因子(S)与被测变量之间的关系。对于这些未测量的因素，不需要函数形式的假设。利用关于选择机制的知识，研究人员可以通过指定边界中参数的大小来解释选择偏差的可能程度。我们还表明，根据目标人群的不同，边界不同，因此可以使用摘要度量来计算将风险比移到空值所需的参数的最小幅度。摘要措施可以用来确定选择的总体实力，这是解释一个结果所必需的。然后，我们证明了在某些情况下或在某些假设下，可以简化边界和摘要度量。利用不同选择机制的例子，我们还演示了研究人员如何实现这些简单的灵敏度分析。

40，理解复杂机器学习模型功能的技术正变得越来越流行，这不仅是为了改进验证过程，而且是为了通过探索性分析来提取有关数据的新见解。尽管目前存在大量这样的工具，但大多数假设预测是点估计，并使用对这些估计的灵敏度分析来解释模型。利用轻量级概率网络，我们展示了在灵敏度分析中如何包含预测不确定性导致：(I)更加健壮和可推广的模型；(Ii)通过不确定性分解实现模型解释的一种新方法。在……里面特别地，我们引入了一种新的正则化方法，它既考虑了预测的均值，又考虑了预测的方差，并证明了得到的网络对未见数据提供了更好的推广。此外，我们还提出了一种通过输入域的不确定性来解释预测不确定性的新方法，从而为验证和解释深度学习模型提供了新的方法。

41，无监督维数选择是一个重要的问题，它试图降低数据的维数，同时保持最有用的特征。虽然降维通常用于构造低维嵌入，但它们产生了难以解释的特征空间。此外，在传感器设计等应用中，需要直接在输入域中进行缩减，而不是构造转换后的空间。因此，尺寸选择(DS)的目的是解决组合问题的识别顶部钾尺寸，这是有效的实验设计，减少数据，同时保持它的解释，并设计更好的传感机制所需的尺寸。本文提出了一种基于图形信号分析的 DS 特征影响度量方法。通过对合成图信号进行蓝色噪声谱的分析，表明我们可以测量每个维数的重要性。通过在监督学习和图像掩蔽方面的实验，证明

了该方法在高维空间的关键特征提取方面的优越性，只使用了原始特征的一小部分。

42，隐马尔可夫模型的 MCMC 算法由于数据本身所固有的时间依赖性，往往依赖于前向后向采样器，具有较大的样本量。近年来，利用隐马尔可夫过程的混合，利用数据的小块来逼近完全后验的后验推理方法已经发展起来。然而，在存在由稀有潜在状态引起的不平衡数据的情况下，所提出的小批估计往往会排除稀有状态数据，从而导致对相关发射参数的推断不佳以及对罕见事件的预测或检测不准确。在这里，我们建议使用一个初步的聚类来对稀有聚类进行过采样，并在随机梯度 MCMC 中降低梯度估计中的方差。我们在真实的和综合的例子上证明了预测和推理的准确性有很大的提高。

43，在本文中，我们仅观察问题输入数据和决策者在多轮中的相应决策，演示了如何学习决策者的目标函数。我们的方法是基于在线学习和工作的线性目标上的任意可行集，我们有一个线性优化预言。因此，它推广了以往基于 KKT 的方法–系统分解和对偶。我们提出的两种精确算法–分别基于乘法权更新和在线梯度下降–以 O(1/sqrt(T)的速度收敛，从而允许作出基本上与观察到的决策者在相对较少的观察后作出的决定一样好的决策。我们还讨论了几个有用的推广，如非线性目标函数的近似学习和次优观测的情况。最后，我们证明了我们的方法在广泛的计算研究中的有效性和可能的应用。

44，本文从标记时间点过程和随机微分方程随机最优控制(SDES)的角度，探讨了感病易感(SIS)流行过程的模型和控制策略的发展。与以前的工作相比，这种新的视角特别适合于利用关于疾病爆发的细粒度数据，并使我们能够克服现有控制策略的缺点。我们的控制策略依靠治疗强度来决定治疗的对象和时间，以减少感染人数。对合成数据的初步实验表明，我们的控制策略始终优于其他几种控制策略。展望未来，我们相信，我们的方法为制定实际数据驱动的流行病过程控制策略提供了一个有希望的步骤。

45，我们提出了 Laplacian K-模用于联合聚类和密度模式发现，并提出了该问题的凹-凸松弛，并给出了一种扩展到大数据集和高维数的并行算法。我们优化了松弛的紧界(辅助函数)，在每次迭代中，计算每个簇分配变量的独立更新，保证收敛性。因此，我们的绑定优化器可以用于大规模数据集的分配.此外，我们还证明，通过简单的最大值运算，可以得到密度模式作为赋值变量的副产品，其额外的计算成本与数据点数成线性关系。我们的公式不需要存储一个完全亲和矩阵并计算其特征值分解，也不需要对每个点的单纯形约束执行昂贵的投影步骤和拉格朗日-对偶内迭代。此外，与均值漂移不同，我们的密度模估计不需要内环梯度上升迭代.它的复杂性与特征空间维数无关，产生的模式是输入集中有效的数据点，适用于离散域和任意核。我们对各种数据集进行了综合实验，结果表明，该算法在

优化质量(即离散变量收敛目标值)和聚类精度方面具有很强的竞争力。

46，攻击者可能使用各种技术来欺骗自动说话人验证系统，使其成为真正的用户。同时，反欺骗方法的目的是使系统对此类攻击具有较强的鲁棒性。ASVspoof 2017 挑战的重点是重播攻击，其目的是测量重放攻击检测的极限，并制定针对重播攻击的对策。在本文中，我们提出了一种重播攻击检测系统–注意过滤网络，该系统由一种基于注意的过滤机制和一个基于 ResNet 的分类器组成，该机制增强了频率和时间域的特征表示。我们证明，网络使我们能够可视化自动获取的特征表示，这有助于欺骗检测。注意力过滤网络达到 8.99 的评价 EER%在 ASVspoof 2017 版本 2.0 数据集上。通过系统融合，我们的最佳系统进一步获得了 30。%相对于 2017 年 ASVspoof 增强基线系统的相对改进。

47，本文研究了有限长度销售季节下的动态分类优化问题。T 型。在每个时间段，卖方在基数约束下向到达的客户提供各种可替换的产品，客户根据离散的选择模型在所提供的产品之间进行购买。大多数现有的工作将每个产品与一个实值固定平均效用相关联，并假设一个多项式 Logit 选择(MNL)模型。在许多实际应用中，产品的特征/上下文信息是现成的。本文通过假设平均效用与特征之间的线性关系，综合了特征信息。此外，我们允许产品的特征信息随着时间

的变化而变化，这样潜在的选择模型也可以是非平稳的。为了解决在这种变化的上下文 MNL 模型下的动态分类优化问题，我们需要同时学习潜在的未知系数，并对分类进行决策。为此，我们制定了一项基于上信心范围的政策，并根据以下顺序建立了后悔界限。奥˜(丁 T 型¯√)，在哪里丁是特征的维数，并且奥˜抑制对数依赖。我们进一步建立了下界。Ω(丁 T 型¯√/钾)哪里钾是提供的分类的基数约束，通常是很小的。什么时候钾是一个常数，我们的政策是最优的，直到对数因素。在 UCB 算法的开发阶段，我们需要解决一种基于学习信息的组合优化问题。我们进一步发展了一种近似算法和一种有效的贪婪启发式算法。我们的数值研究进一步证明了拟议政策的有效性。

48，尽管卷积神经网络(CNNs)近年来在各种图像处理和计算机视觉任务中得到了广泛的应用，但如何降低资源有限平台参数的存储成本仍然是一个具有挑战性的问题。在以往的研究中，张量分解(Td)通过将卷积层的核嵌入到一个低秩子空间中，取得了良好的压缩性能。然而，TD 的使用在内核或其指定的变体上是天真的。与传统的方法不同，本文证明了核可以嵌入到更一般的甚至随机的低秩子空间中。我们通过随机改组张量分解(RsTD)压缩卷积层来证明这一点，并使用 CIFAR–10 对标准分类任务进行了压缩。此外，我们还分析了训练数据的空间相似性对核的低秩结构的影响。实验结果表明，即使核被随机地洗牌，CNN 也能被显著压缩。此外，在较大的压缩比

范围内，基于 RsTD 的分类方法比传统的基于 TD 的方法具有更稳定的分类精度。

49，DBSCAN 是一种经典的基于密度的聚类方法，具有极大的实用价值.然而，它隐含地需要计算每个样本点的经验密度，从而导致二次最坏情况的时间复杂度，这在大型数据集中可能太慢。
我们提出 DBSCAN+，这是 DBSCAN 的一个简单的修改，它只需要计算一个点的子集的密度。实验表明，与传统的 DBSCAN 相比，DBSCAN+不仅可以提供竞争性能，而且可以在占用运行时的一小部分时间内增加带宽超参数的鲁棒性。
我们还提出了统计一致性保证，显示了计算费用和估计率之间的权衡。令人惊讶的是，在某种程度上，我们可以在降低计算成本的同时享受相同的估计率，这表明 DBSCAN+是一种次二次算法，它为水平集估计获得了极大的最优速率，这一性质可能是独立的。

50，近年来，随着卷积神经网络在许多具有挑战性的机器学习领域取得重大成就，人工神经网络已经不能满足我们的需求，网络的设计成本越来越高，自动生成的体系结构也越来越受到人们的关注和关注。一些关于自动生成网络的研究已经取得了很好的效果。然而，它们主要针对的是一系列的单层，如卷积或逐层汇聚等。在精心制作的神经网络中，有许多优雅而富有创造性的设计，如 Google 网中的初始块、残留网络中的剩余块和密集卷积网络中的密集块。在强

化学习的基础上，利用这些网络的优势，提出了一种新的多块神经网络的自动设计过程，该网络的结构包含了上述多个类型的块，目的是对深层神经网络进行结构学习，并探讨是否可以将不同的块组合在一起形成性能良好的神经网络。最优网络是由 Q-学习代理创建的，该 Agent 被训练成顺序地选择不同类型的块。为了验证该方法的有效性，我们利用自动生成的多块神经网络对具有有限计算资源的图像基准数据集 MNIST、SVHN 和 CIFAR-10 图像分类任务进行了实验研究。结果表明，该方法是非常有效的，与人工神经网络和先进的自生成神经网络相比，其性能相当或更好。

51，近年来，深度神经网络(DNN)的发展开创了一个全新的人工智能时代。在解决非常复杂的问题，如视觉识别和文本理解方面，DNNs 被证明是非常出色的，足以与人竞争，甚至超越人。尽管 DNNS 取得了令人鼓舞的成功，但深入的理论分析仍未解开其神奇之谜。DNN 结构的设计以网络深度、神经元数目和激活等经验结果为主导。为了解释 DNN，最近发表了一些杰出的著作，其中一些已经建立了对其内部机制的第一次窥见。尽管如此，关于 DNNS 如何运作的研究仍处于初级阶段，还有很大的改进余地。本文扩展了关于线性区域界的分段线性激活神经网络(PLNN)的先例研究。我们给出了(I)单层 PLNN 的精确线性区域的最大数目；(Ii)多层 PLNN 的上界；(Iii)整流器网络上最大线性区域数的一个更紧的上界。导出的界也间接解释了为什么深层模型比浅层模型更强大，以及激活函数的

非线性如何影响网络的表现力。

52，我们探讨如何利用深度学习(DL)来预测急性髓系白血病(AML)的预后。在 TCGA(癌症基因组图谱)数据库中，本研究使用了 94 例 AML 患者。输入数据包括年龄，10个常见的细胞遗传学和23个最常见的突变结果，输出为预后(诊断死亡，DTD)。在我们的 DL 网络中，自动编码器被堆叠成一个分层的 DL 模型，从该模型中压缩和组织原始数据，并提取高级特征。该网络是用 R 语言编写的，旨在预测特定病例(DTD≥730 天)的 AML 预后。DL 网络预测预后的准确率为83%。作为概念研究的证明，我们的初步结果证明了 DL 在应用下一代测序(NGS)数据预测预后的未来实践中的实际应用。

53，典型相关分析(CCA)是一种在多视图数据中寻找最大相关变量的线性表示学习方法。非线性 CCA 将这一概念扩展到了更广泛的变换家族，这对于许多现实世界的应用来说更加强大。在给定联合概率的情况下，交替条件期望(ACE)为非线性 CCA 问题提供了一个最优解.然而，在只有有限的观测量的情况下，它的性能有限，计算负担也越来越大。本文介绍了非线性 CCA 问题(ITCCA)的信息论框架，扩展了经典 ACE 方法。我们建议的框架寻求压缩的数据表示，允许最大程度的相关性。通过这种方式，我们控制了表示的灵活性和复杂性之间的权衡。在有限样本范围内，与非线性选择相比，我们的方法在减少计算负担时表现出良好的性能。此外，ITCCA 还提供了

理论上的界限和最优性条件，因为我们建立了与速率失真理论、信息瓶颈和远程信源编码的基本联系。此外，它还意味着一种“软”降维，因为压缩水平是由原始噪声数据和我们提取的信号之间的相互信息来测量(和控制)的。

54，机器学习技术已经深深地植根于我们的日常生活中。然而，由于追求良好的学习成绩是知识密集型和劳动密集型的，所以人类专家在机器学习的各个方面都投入了大量的精力。为了使机器学习技术更容易应用，减少对有经验的人类专家的需求，自动机器学习~(AutoML)已成为工业界和学术界的研究热点。本文对现有的AutoML作品进行了综述。首先，我们介绍并定义了AutoML问题，并从自动化和机器学习两方面给出了启示。然后，我们提出了一个通用的AutoML框架，它不仅涵盖了几乎所有现有的方法，而且指导了新方法的设计。然后，从问题设置和所采用的技术两个方面对现有的工作进行了分类和回顾。最后，我们对AutoML方法进行了详细的分析，并解释了它们成功应用的原因。我们希望这一调查不仅能为AutoML初学者提供一个有洞察力的指导方针，而且对今后的研究也有一定的启示。

55，提出了一种预测ICD–10诊断码的多模态机器学习模型.我们开发了独立的机器学习模型，可以处理来自不同模式的数据，包括非结构化文本、半结构化文本和结构化表格数据。我们进一步采用集

成的方法来集成所有特定于模态的模型来生成ICD-10码.还提取了关键证据，使我们的预测更具说服力和可解释性。我们使用医疗信息集市为重症监护III(模拟-III)数据集，以验证我们的方法。对于ICD码的预测，我们的最佳模型(micro-F1=0.7633，micro-aUC=0.9541)明显优于其他基线模型，包括 TF-f1=0.6721，micro-aUC=0.7879和 text-cnn 模型(micro-f1=0.6569，micro-aUC=0.9235)。在可解释性方面，我们的方法实现了文本数据的 Jaccard 相似系数(JSC)为0.1806，对表格数据的相似度系数(JSC)为 0.3105，其中训练有素的医生分别达到 0.2780 和 0.5002。

56，拉格朗日资料同化是海洋和大气模拟中的一个复杂问题。大规模地球物理流中的漂移跟踪涉及到漂移位置的不确定性、复杂的惯性效应等因素，使其与数值模型模拟的拉格朗日轨迹进行比较具有极大的挑战性。时间和空间离散化是模拟大规模流动所必需的因素，也有助于真实和模拟漂流轨迹的分离。这些湍流中固有的混沌平流倾向于分离甚至紧密间隔的示踪粒子，使得仅基于漂移位移的误差度量不适合于模型参数的估计。我们建议用相干结构着色(CSC)领域的误差来评估模型的技巧。CSC 场提供了流中潜在的相干模式的空间表示，并且我们证明了它是评估模型精度的一个更稳健的度量。通过在粒子初始化过程中考虑空间不确定性和沿弹道随机误差的影响以及时间离散的两个测试案例，证明了相干结构着色场中的误差可以精确地确定单个或多个同时未知的模型参数，而基于漂移位移

误差的传统误差度量则不能实现。由于 CSC 场增强了模型参数之间的误差差异，使得模型参数扫描中的误差最小化变得更加明显。该方法用于分析流中的单参数和多参数估计的有效性和鲁棒性表明，实际海洋和大气模式的拉格朗日数据同化将受益于类似的方法。