

智能数据可视分析技术综述^{*}

骆昱宇¹, 秦雪迪¹, 谢宇鹏², 李国良¹

¹(清华大学 计算机科学与技术系, 北京 100084)

²(青海大学 计算机技术与应用系, 青海 西宁 810016)

通讯作者: 李国良, E-mail: liguoliang@tsinghua.edu.cn

摘要: 如何从海量数据中快速有效地挖掘出有价值的信息以更好地指导决策, 是大数据分析的重要目标。可视化分析是一种重要的大数据分析方法, 它利用人类视觉感知特性, 使用可视化图表直观呈现复杂数据中蕴含的规律, 并支持以人为本的交互式数据分析。然而, 可视分析仍然面临着许多挑战, 例如数据准备代价高、交互响应高延迟、可视分析高门槛和交互模式效率低。为应对这些挑战, 研究者从数据管理、人工智能等视角出发, 提出了一系列方法以优化可视分析系统的人机协作模式和提高系统的智能化程度。本文系统性地梳理、分析和总结了这些方法, 提出了智能数据可视分析的基本概念和关键技术框架。然后, 在该框架下, 本文综述和分析了国内外面向可视分析的数据准备、智能数据可视化、高效可视分析和智能可视分析接口的研究进展。最后, 本文展望了智能数据可视分析的未来发展趋势。

关键词: 数据可视化; 可视分析; 智能数据可视分析; 数据管理; 人工智能

中图法分类号: TP311

中文引用格式: 骆昱宇, 秦雪迪, 谢宇鹏, 李国良. 智能数据可视分析技术综述. 软件学报, 2023.

英文引用格式: Luo YY, Qin XD, Xie YP, LI GL. Intelligent Data Visual Analytics Techniques: A Survey. *Ruan Jian Xue Bao/Journal of Software*, 2023 (in Chinese). <http://www.jos.org.cn/1000-9825/0000.htm>

Intelligent Data Visual Analytics Techniques: A Survey

LUO Yu-Yu¹, QIN Xue-Di¹, XIE Yu-Peng², LI Guo-Liang¹

¹(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

²(Department of Computer Science, Qinghai University, Xining 810016, China)

Abstract: How to quickly and effectively mine valuable insights from massive data to better guide decision-making is an important goal of big data analytics. Visual analytics takes advantage of the characteristics of human visual perception, utilizes visualization charts to encode complex data intuitively, and supports human-centered interactive data analytics, which is one of the important ways of big data analytics. However, visual analytics still faces several challenges, such as the high cost of data preparation, the high latency

* 基金项目: 国家自然科学基金(61925205, 62232009, 62072261); 国家重点研发计划(2020AAA0104500); 北京国家信息研究中心

of interaction response, the steep learning curve of visual analytics, and the low efficiency of interaction mode. To address the above challenges, recent studies propose to optimize the human-computer interaction mode and improve the visual analytics system's intelligence by leveraging data management and artificial intelligence techniques. In this paper, we systematically survey recent advances in this field and present the basic concepts and key technical framework of intelligent data visual analytics. Secondly, we summarize the overall architecture of intelligent data visual analytics, which comprises effective data preparation for visual analytics, intelligent data visualization, efficient visual analytics, and intelligent visual analytics interfaces. Next, we go through the prior art, paying particular attention to problems that may attract interest from data management, visualization, and machine learning communities. Finally, we discuss the research opportunities of intelligent data visual analytics.

Key words: data visualization; visual analytics; intelligent data visual analytics; data management; artificial intelligent

1 引言

随着计算机硬件和大数据处理技术的高速发展，海量数据智能分析的瓶颈已经从“如何快速地处理海量数据”转变为“如何从海量数据中快速有效地挖掘出有价值的信息”。可视化和可视分析基于人类的视觉感知特性，结合数据分析和人机交互等技术，利用可视化图表去解构复杂数据中蕴含的知识和规律。这种技术贯穿于数据科学的全生命周期，被誉为大数据智能领域的最后一公里，已在许多大数据应用分析场景取得令人瞩目的效果。因此，中国科技创新 2030 “新一代人工智能”和“大数据”专项都将可视化和可视分析列为广大数据智能的关键技术^[1,2]。然而，如图 2 所示，传统的可视分析极度依赖用户频繁主动地参与可视分析的全生命周期^[3,4]，包括数据准备、数据转换、可视化映射、可视化渲染绘图、用户交互、可视分析等阶段，对用户的专业技能要求较高，系统的智能化程度较低。因此，传统的可视分析模式和系统存在可视分析高门槛、数据准备代价高、交互响应高延迟、交互模式效率低等挑战。

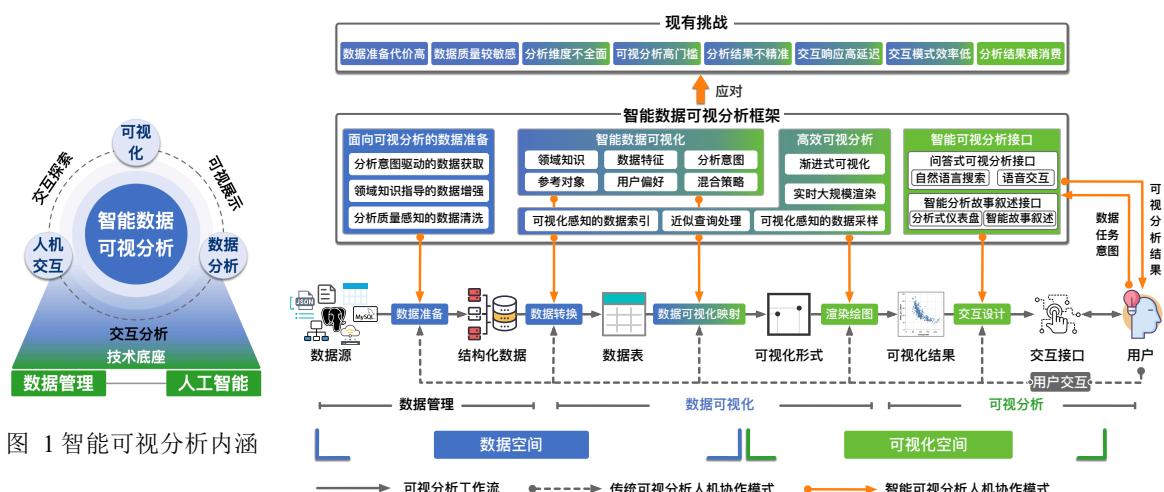


图 2 可视分析工作流和智能数据可视分析技术框架

为了提高可视分析系统的整体效能，研究者们^[4-8]从人工智能和数据管理的视角出发，将人工智能和数据管理技术赋能可视化和可视分析系统，提高系统的智能化程度，进而帮助用户高效地参与可视分析全生命周期的数据准备、可视化、可视分析交互等环节，优化可视分析的人机协作模式，提高可视分析的质量和效率。基于此，**智能数据可视分析** (Intelligent Data Visual Analytics) 的概念应运而生，其核心思想是“算法赋能”和“以简驭繁”，通过数据管理和人工智能技术赋能可视分析的工作流，将传统可视分析工作流中的用户的主动探索和分析变为机器算法的智能辅助探索和分析，降低可视化和可视分析的生产和消费成本，协同优化可视分析全生命周期的数据管理、可视化和可视分析的人机协作模式，致力于辅助用户高效地进行可

视分析。从学科关系的视角出发,如图1所示,智能数据可视分析是以数据管理和人工智能技术为支撑,通过人机交互手段进行交互式数据分析,通过可视化手段进行数据的信息解构和分析结果的直观呈现,帮助用户快速地从海量数据中挖掘出有价值的信息。从可视化工作流的视角出发,如图2所示,智能数据可视分析技术可以优化传统可视分析工作流的人机协作模式,提高可视分析的效能。具体而言,智能数据可视分析技术可以优化传统可视分析工作流中的数据准备、可视化生成、大数据高效可视分析和可视分析人机交互接口4个模块。接下来,本文将围绕上述4个模块,展开介绍智能数据可视分析技术:

(1) **面向可视分析的数据准备:** 传统可视化和可视分析工作流中的数据准备工作没有针对可视化和可视分析的特点进行优化,存在数据准备代价高、数据质量较敏感和分析维度不全面的挑战。首先,在数据发现阶段,传统方法没有根据用户的分析任务进行相关数据集/数据元组的发现,从而导致在数据准备阶段融合了大量对可视分析无关或者没有蕴含足够洞察的数据集,加重后续可视分析的负担。其次,在数据清洗阶段,传统方法力求找到数据集中的所有错误并进行清洗,以为后续的可视分析提供高质量的数据集。然而,这种数据清洗方式的代价通常很高。如果在数据准备阶段提前考虑可视分析的意图,即清洗与可视分析查询相关的数据子集,则在降低数据清洗代价的同时还能提高可视分析的质量。此外,如果获取的数据集属性过于单一,通常会导致分析的维度过于局限。因此,可以通过关联相关数据源进行数据增强,丰富可视分析的维度。面向可视分析的数据准备技术旨在运用数据管理和人工智能技术,结合可视化和可视分析的特性,优化可视分析工作流中数据准备阶段的人机协作模式,为用户以低成本的方式准备高质量和语义丰富的数据,以支持高质量的可视化和可视分析。

(2) **智能数据可视化:** 数据可视化通过可视化图表来解构复杂数据中蕴含的知识和规律。在可视化阶段,概括来说需要解决两大核心的任务为“需要可视化哪些数据(What data is needed?)”和“以什么样的方式进行数据的可视化(How to visualize the data?)”。传统的可视化方式需要用户在对数据集理解的基础上,选择和过滤出用于生成可视化结果的数据子集,挑选合适的数据维度并进行一系列的数据转换操作(如聚集操作等),最后通过可视化工具将该数据表映射到可视化空间中,渲染生成可视化结果。如果生成的可视化结果不满足可视分析中用户的任务需求,则需要重复上述的若干步骤直到找到用户满意的可视化结果。不难看出,传统的可视化过程通常是循环迭代的,需要用户参与到数据选择、转换和可视化映射等环节,存在可视分析高门槛、交互模式效率低、分析结果不精准和分析维度不全面的挑战。为了解决上述挑战,智能数据可视化技术需要结合用户的分析意图、数据特征、领域知识等,自动地生成和推荐给定数据集中有价值的可视化结果,帮助用户高效地进行可视化和可视分析。

(3) **高效可视分析:** 在数据量急剧增长的情况下,受计算能力可扩展性和显示设备局限性的约束,会导致可视分析的交互响应延迟较高。一方面,这是由于可视分析系统的数据处理和分析时间较长;另一方面,大规模的数据点难以高效渲染并在有限的显示设备上进行呈现和实时交互。为了解决上述挑战,研究人员从硬件和计算框架、数据管理、人工智能和可视化技术出发研究高效可视分析技术,协同优化可视分析中的数据管理和可视化交互的效率。例如,基于可视化感知的数据索引技术和近似查询处理技术,高效地进行数据组织和处理;利用人工智能技术进行用户交互行为的预测,进行用户分析查询的高效重写和数据预取;基于视觉感知的采样、渐进可视化和实时渲染技术,进行大规模数据的高效渲染和实时交互。

(4) **智能可视分析接口:** 可视分析接口是用户与系统交互的媒介,一方面,系统需要通过交互接口获得用户可视分析的意图和操作指令。传统的交互方法需要用户根据可视分析系统的交互设计规则,学习特定系统的交互方式(如编程指令或图形化界面操作方式等),对用户的专业要求技能较高,交互接口的学习成本也较大,存在可视分析门槛高和交互模式效率低的挑战;另一方面,可视分析的结果需要通过交互接口呈现给用户,传统的方法仅仅是将可视分析的碎片化发现直接呈现给用户,需要用户进一步挖掘这些碎片化可视分析结论的内在逻辑关系和因果关系,并进一步整理成可在组织内传播的可视分析报告,存在可视分析结果难消费的挑战。基于上述讨论,一方面,智能可视分析接口需要为用户提供简单的交互接口(例如基于自然语言查询的接口),并通过智能算法进行用户分析意图的理解和可视分析结果的生成和推

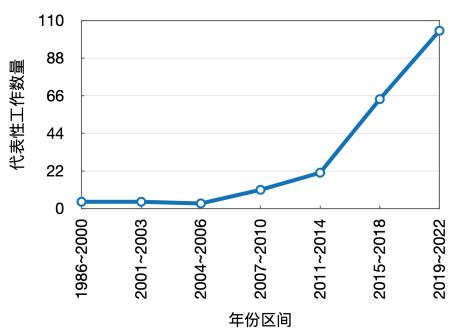
荐,降低可视分析系统的使用门槛和优化系统的人机协作模式。另一方面,智能可视分析接口还需要基于人工智能技术,自动挖掘可视分析结果之间的内在联系,通过关系挖掘、信息补全、文本生成等技术,基于用户可视分析得到的碎片化结果智能地生成分析式仪表盘和可视分析故事叙述,提高用户整理和共享可视分析结果的效率,从而缓解可视分析结果难消费的挑战。

综上所述,智能数据可视分析以人工智能和数据管理技术为支撑,结合可视化和可视分析、人机交互等技术,对可视分析工作流的数据准备、可视化生成、大数据高效可视分析和可视分析人机交互接口4个模块进行协同优化:优化可视分析中数据准备阶段的人机协作模式,以支持用户以低成本的方式准备高质量的分析数据;通过智能可视化手段,自动地生成和推荐数据集中有意义的可视化和可视分析结果给用户,优化可视化的生产模式;基于数据管理和可视化技术提高分析数据的处理效率,以支持海量数据的实时分析和交互;基于数据挖掘、自然语言处理和可视化技术为用户提供问答式可视分析接口,并根据可视分析的结果智能地生成分析式仪表盘和可视分析故事叙述,降低用户利用可视分析结果的代价。

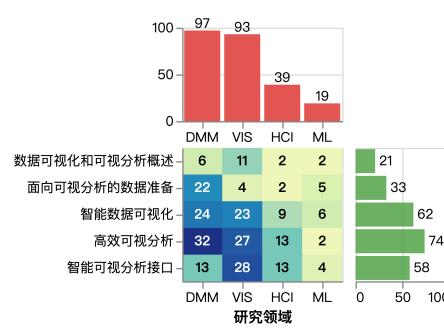
综述调查范围:为了更好地对智能数据可视分析的研究进展进行梳理、总结和分析,本文对三十多年来(1986~2022)近200篇论文进行了系统性地梳理、总结和分析。如表1所示,本文主要调查了可视化、数据挖掘和数据管理、人机交互和机器学习领域的相关国际会议和期刊论文。通过对图3(a)论文发表年份变化的观察,会发现所有会议或者期刊随着时间推移论文的总数呈现出稳步上升的趋势,这也说明了智能数据可视分析在未来将会受到持续的关注。图3(b)展示了本文主要章节所调查论文的分布情况,其中可视化与数据挖掘和数据管理领域的相关论文数量最多,由此反映出本文所调查的论文与本文题目的相关性较高,还可以看出各章节对各领域的分析都有所涉及且重点突出。

表1 本文综述调查范围

研究领域	会议/期刊
可视化和图形学 Visualization (VIS)	IEEE VIS (InfoVis, VAST, SciVis), TVCG, EuroVis, PacificVis TOG, SIGGRAPH, CGF
数据挖掘和数据管理 Data Mining and Management (DMM)	KDD, SIGMOD, VLDB, ICDE, TKDE, The VLDB Journal TODS, CIDR, DASFAA, EDBT, IEEE BigData
人机交互 Human-Computer Interaction (HCI)	CHI, UIST, IUI, HILDA, AVI, HCI
机器学习 Machine Learning (ML)	ICML, NeurIPS, CVPR, ACL, EMNLP, IJCAI, AAAI



(a) 论文发表年份分布



(b) 本文主要章节调查论文情况

图3 本文综述调查的研究论文分布情况

与相关综述性文章的区别: Qin等人^[4]主要从数据库的视角出发,梳理和总结了可用于提高可视化创建效率的相关技术,主要包括数据自动可视化技术和支持海量数据高效可视化的技术。Battle等人^[9]也从数据管理的视角出发,总结了可以支持海量数据交互式可视化的数据管理技术。Zhu等人^[10]梳理了基于机器学习和专家规则的数据自动可视化技术。Shen等人^[8]对可视化和可视分析的自然语言接口的相关研究和

应用进行了梳理总结。Wang 等人^[11] 基于 85 篇代表性论文, 梳理了可用于提高可视化效率的机器学习技术。Wu 等人^[12] 提将可视化结果作为一种特殊的数据, 并梳理了如何使用人工智能技术管理和利用这类数据。然而, 上述综述往往从单一的学科视角出发^[4], 或只涵盖了智能数据可视分析的个别细分领域^[10-13], 或未能包括最新进展^[4,10]。因此, 本文从可视化、数据管理和人工智能的视角出发对智能数据可视分析的研究进展进行梳理, 希望能够帮助研究者和从业者增强对智能数据可视分析最新进展的了解。

本文的主要贡献: 首先, 本文通过调查近 200 篇智能可视分析领域的研究工作, 总结出智能数据可视分析的 4 个核心模块, 凝练出智能数据可视分析的基本概念和关键技术, 揭示了智能数据可视分析和多研究领域的交叉关系。其次, 本文系统性地梳理和分析了智能数据可视分析的代表性工作, 深入浅出地对各细分领域的研究进行分析讨论, 研究者们可以快速地掌握各细分领域的研究进展和机遇。最后, 本文还探讨了智能可视分析的发展趋势并为研究者们提供了未来可能的探索方向。

本文的组织结构: 本文主要讨论关系型数据, 基于图 2 所示的智能数据可视分析框架。本文在第 2 章介绍可视化的基本概念和可视分析的基本流程。第 3 章介绍面向可视分析的数据准备技术, 具体包括分析意图驱动的数据获取、领域知识指导的数据增强和分析质量感知的数据清洗技术; 第 4 章梳理智能数据可视化技术, 包括基于融合分析意图的可视化推荐和基于用户偏好的可视化推荐等技术; 第 5 章分析了高效可视分析技术, 包括基于高效数据管理的高效可视分析、可视化感知的高效可视分析、人工智能驱动的高效可视分析和基于硬件和计算框架加速的高效可视分析技术; 第 6 章展开介绍智能可视分析接口技术, 具体包括问答式可视分析接口和智能分析故事叙述接口; 第 7 章讨论智能数据可视分析的未来发展趋势和研究机会, 并在第 8 章总结全文。

2 可视化和可视分析概述

为了方便无相关知识背景的读者阅读本文, 本章首先介绍可视化的基本概念, 并由此引出用于表示和查询可视化的查询语言。最后, 本章会阐述以可视化为媒介的可视分析的基本流程, 以及其与自动数据分析的联系和区别。

2.1 可视化

可视化可以将复杂的数据以可视化图表的信息进行呈现, 利用人类视觉感知特性, 帮助用户解构复杂数据蕴含的信息。一个好的可视化结果具有化繁为简和一目了然等特性, 可以解构数据中蕴含的复杂信息, 高效传达数据知识, 引导读者关注到数据的重要特性和赋予读者洞察不同数据侧面的能力。

图 4 形象化地展示了可视化是如何将已经准备好的关系型数据转换成可视化结果。概括来说, 可视化是指从给定数据集中, 选择合适的数据属性, 进行必要的数据转换操作, 选择恰当的可视化编码方式, 最后渲染绘图将可视化结果呈现出来。



图 4 可视化的基本概念示意图

如图 4 所示, 本文根据可视化流程中的数据形式, 将可视化流程分为数据空间和可视化空间。

数据空间对原始数据进行一系列的数据转换操作, 如过滤、选择、聚集等, 提炼原始数据的主要特性并将原始数据转换成易于可视化的形式。例如, 图 4 中使用了 SQL 查询完成数据转换, 该 SQL 查询选择了 Cars 数据表的 Origin 属性, 并对该属性进行分组和计数操作, 得到转换数据以进行可视化表示。

可视化空间将数据映射为合适的可视化形式，以帮助用户直观理解数据蕴含的知识和规律。例如，图 4 的可视化映射将转换数据映射为柱状图，并将转换数据的 Origin 和 COUNT(Origin)分别映射到柱状图的 X 轴和 Y 轴。最后通过渲染绘图，将数据转换为屏幕上显示的可视化结果。

2.2 可视化查询语言

正如 SQL (Structured Query Language, 结构化查询语言) 可以方便数据库用户对库内数据进行增、删、改、查操作，可视化的过程也需要相应的查询语言声明式地表示可视化过程中的数据转换和可视化映射等一系列工作。本文根据可视化查询语言 (VQL, Visualization Query Language) 的表达程度 (Expressiveness) 和易用程度 (Ease-of-Use) 两个维度对主流可视化查询语言进行总结。



图 5 可视化查询语言/系统概览

如图 5 所示，表达性最高的是图形化编程接口，例如 OpenGL^[14]、Java2D^[15]、HTML Canvas^[16] 等，这类图形化编程接口可以直接将数据映射到像素空间并渲染显示到屏幕上，是上层声明式可视化查询语言的基础。然而，图形化编程接口也对用户的编程和相关技能有着极高的要求，既需要用户明确指明需要对哪部分数据子集进行何种可视化操作，又需要用户在掌握这些底层编程语言的语法的前提下指明这些操作具体是如何实现的。为了简化可视化的编程难度，陆续出现了一些声明式的可视化查询语言。这些声明式的可视化查询语言封装了底层图形编程接口的具体实现方式，只给用户暴露出了如图 4 所示的数据转换和可视化映射的编程接口。

对于声明式的可视化查询语言，本文根据易用程度可以细分为高级语言和低级语言。其中，高级语言有 Vega-Lite^[17]、CompassQL^[18]、VizQL^[19]、Altair^[20]、Echarts^[21]、ggplot2^[22]、ZQL^[23] 等。用户需要根据语法指定如何进行可视化，即主要指定用于可视化的数据属性和必要的数据转换操作等，具体如何将数据元素映射到可视化空间则由查询语言依据预定义的规则进行执行。相较于高级语言，低级语言如 Vega^[24] 则提供给用户更多渲染绘图的参数，例如指定柱状图柱子的宽度等。图 6 (a) 给出了使用 Vega 可视化查询语言从图 4 中的 Cars 数据集生成对应柱状图的示例。其中，蓝色区域部分大致对应图 4 中的数据空间的操作，主要负责进行数据集的选择、数据属性的选择以及相应的数据转换操作；绿色部分大致对应图 4 中的可视化空间的操作，负责可视化映射和渲染绘图的相关参数指定。不难发现，低级可视化查询语言需要用户指定较多的参数，学习和使用的门槛较高。高级可视化查询语言对可视化过程做了较多的封装，用户需要指定的参数较少，易用性较高。图 6 (b) 给出了 Vega-Lite 可视化查询语言的示例，可以发现 Vega-Lite 的语法规则比低级语言 Vega 更加简洁，用更少的参数可以实现同样的可视化。Vega-Lite 查询语言设计的核心是将可视化过程描述为从数据到图形标记的映射，这些映射方式由用户指定。CompassQL^[18] 的语法规则与 Vega-Lite 相似，但 CompassQL 是面向可视化推荐的查询语言，其核心是将可视化过程看成是推荐过程，即允许用户不指定具体的映射方式，由 CompassQL 根据推理策略去推荐合适的映射方式。本文将在 4.2 节详细介绍 CompassQL。

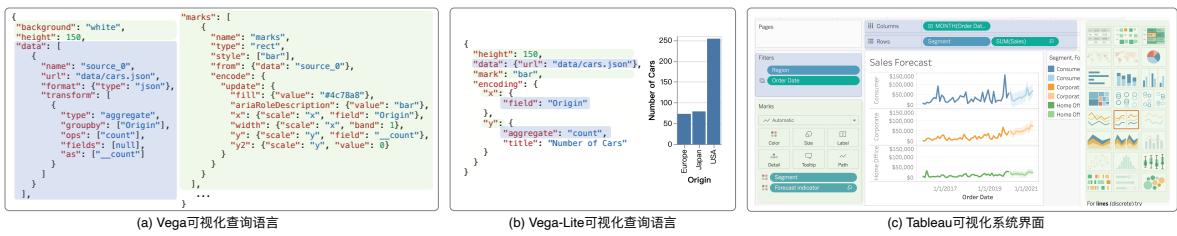


图 6 可视化查询语言示例

为了方便普通用户进行数据的可视化分析，研究人员开发了基于图形化界面的可视化系统。可视化系统可以图形化交互组件来帮助用户通过交互式或者系统推荐的方式生成可视化结果，以辅助用户从数据中获取数据洞察（Insights）。如图 6（c）所示，展示了 Tableau^[25] 可视化系统的操作界面，用户在蓝色区域进行数据空间的操作，在绿色区域指定可视化空间的参数，Tableau 为用户提供了基于点击和拖拽的交互界面，让用户可以不用写代码就能完成可视化。根据用户的参与程度和系统的智能化程度，图 5 将可视化系统分为用户创建类和智能推荐类。对于用户创建类的系统，常见的有 GoogleSheets^[26]、Excel^[27]、ManyEyes^[28]、DataIllustrator^[29]、Lyra^[30,31]、APT^[32] 等。以 Excel^[27] 为例，这类系统在用户选择用于可视化的数据行和列之后，需要用户选择系统提供的可视化模板，生成对应的可视化结果。如果用户还需要对数据进行如分组和聚集操作，则需要用户在可视化之前就完成好相应的数据处理操作。对于智能推荐类的系统，常见的有 Tableau^[25]、QuickSight^[33]、Qlik^[34]、Zenvisage^[23]、DeepEye^[5,35,36]、Voyager2^[37] 等，这些系统的智能化程度比较高，可以基于当前数据的特征，自动地完成数据转换（如聚集）和可视化映射。概括来说，这类系统可以从数据特征、领域知识、用户意图等多个方面综合分析，自动地推荐给定数据集的有意义的可视化结果。本文将这类技术概括为智能数据可视化技术，并在第 4 章展开介绍智能数据可视化技术。

2.3 可视分析与自动数据分析

可视分析的核心目标就是通过人机协作，使用自动化或者交互式的数据分析手段，以可视化为人机交互媒介，从数据集中进行知识发现，从而指导用户进行科学决策。在过去 30 余年间，学术界涌现出了许多经典的可视化和可视分析流程参考模型^[38-45]，这些模型从不同角度建模了可视化和可视分析的人机交互模式。Haber 等人^[38] 在 1990 年提出了可视化的流程概念模型，该线性模型通过三个步骤（数据浓缩/增强、可视化映射和渲染）来概括从数据到可视化的关键过程。Pirolli 等人^[39] 提出了信息觅食（Information Foraging）理论，该理论可以解释可视分析中人机交互的机理。基于上述理论，Card 等人^[40] 在 1999 年提出了一个循环模型来表示可视化和可视分析生命周期中用户与可视化循环迭代的关键步骤。对于创建具体的可视化对象，Munzner 等人^[41] 提出了一个用于设计和验证可视化的嵌套模型。

步入 21 世纪，可视分析与机器学习等自动数据分析技术的结合越来越密切，Keim 等人^[44] 提出了一个可视分析的交互模型，该模型重点概括了可视分析工作流中机器算法和用户各自扮演的角色以及协同模式。因为智能数据可视分析高度依赖机器算法以及用户交互，为了更好地诠释数据可视分析与自动数据分析直接的关联与区别，本文基于 Keim 等人^[44] 提出的可视分析的工作流程，给出从数据源到知识指导的科学决策全周期的可视分析流程示意图，包含了可视分析流程的每个步骤以及可视分析与基于模型的数据分析的内在关联，如图 7 所示，其起点是数据源，终点是基于分析结果的科学决策。从数据到决策有两条路径：即数据可视分析和自动数据分析。数据可视分析从数据到可视化的过程基本遵循了图 2 所示的可视化工作流，用户通过与可视化结果进行交互，从中挖掘出有用的知识。自动数据分析主要是依赖模型自动地挖掘数据中蕴含的知识和规律，以指导用户进行科学决策。基于模型的自动数据分析的经典案例是“啤酒与尿布的故事”，即对用户购物的历史数据进行分析，发现大量用户经常同时购买尿布和啤酒这两个看似不相关的商品，进而指导商场将尿布和啤酒两个货架摆放在较近的位置。

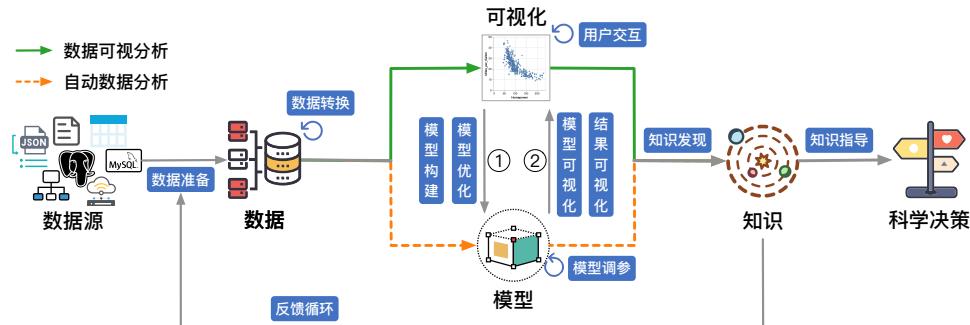


图 7 可视分析和自动数据分析的联系

数据可视分析和自动数据分析并不是对立的。实际上，在现实的数据分析工作流中，用户经常需要在数据可视分析和自动数据分析两个工作流中进行切换和迭代循环。如图 7 中的①和②所示，对于两个分析工作流的中间结果，从可视化的视角出发（①），通过对自动数据分析工作流构建的数据模型进行可视化，以帮助用户通过可视化结果交互式地去改进模型的参数，优化自动数据分析的质量。从模型的视角出发（②），可以将模型的分析结果进行可视化，用户通过对可视化结果进行分析，一方面，可以从中挖掘出有价值的知识；另一方面，也可以提前发现自动数据分析工作流的错误结果。因此，现代数据分析工具通常需要将数据可视分析和自动数据分析两个工作流进行有机融合，用户可以依据当前数据分析的需求，从两个工作流中进行无缝切换衔接。随着人工智能技术的迅猛发展，虽然可以通过深度学习等技术进行部分任务的自动数据分析，但是在很多复杂的场景下，人对复杂知识的理解和建模依然是优于机器算法。因此，在数据分析中，人依然是数据分析工作流的核心要素，即以人为本的数据分析，一些机器智能算法和工具的作用是增强人的能力而不是完全取代人。

3 面向可视分析的数据准备

数据准备是数据分析流程中的重要阶段，数据准备包括数据获取、数据融合、数据清洗等步骤。数据准备的过程是非常繁杂的，需要耗费大量的人力和时间。研究和实践表明，数据科学家通常需要花约 80% 的时间来准备数据^[46]。这表明数据准备代价高的特点。此外，数据准备还涉及到对原数据集进行数据增强和数据清洗，可以缓解分析维度不全面和分析结果对数据质量较敏感的挑战，如图 8 所示。针对上述讨论，研究者们通过考虑可视分析流程的特性，研究面向可视分析的数据准备算法^[47-52]和系统^[53-59]，以帮助用户更加高效地准备数据，从而提高可视分析的质量和效率。

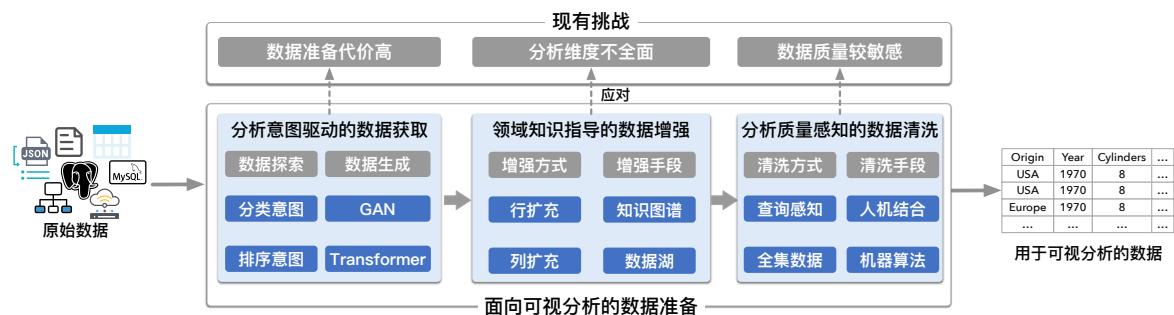


图 8 面向可视分析的数据准备

在这些数据准备工作中，有一部分是专门面向下游数据可视分析任务进行数据准备的。这些工作结合数据可视分析任务的特点来进行有针对性的数据准备，从而达到更加高效、有效的数据可视分析。概括来说，

面向数据可视分析的数据准备的基本思想可以分为：(1) 分析意图驱动的数据获取：用户的原始数据集可能是很大的，而用户可能只对其中的一小部分数据感兴趣。在这种情况下，对所有的数据进行可视分析不仅会耗费大量的时间，还会获得不准确的分析结果。因此，针对用户可视分析意图对数据进行筛选，然后对筛选后得到的满足一定条件的数据进行数据准备可以大大地提高可视分析的效率；针对用户可视分析数据获取难的数据生成：进行数据可视分析的第一步是获取数据。然而很多公司和机构产生的数据涉及大量的用户隐私信息，比如用户姓名、电话号码、电子邮箱、家庭住址等。为了保护用户隐私，这些公司和机构不会将真实数据公之于众。这为用户进行数据可视分析带来了极大的困难。因此，生成可供用户进行数据可视分析的数据是一个非常重要的问题。(2) 领域知识指导的数据增强：用户在数据可视分析中，试图发现一些有趣的现象，并对其进行解释。数据扩充可以为原始数据增加更多的列（即属性）或行（即记录），从而丰富可视分析的维度，可获得更多有意义的分析结论。例如，如果为国家数据集添加人口数和面积属性，则可以对国家总人口数和平均面积人口数等进行分析；如果为论文数据集添加论文发表机构的世界排名，则可以对论文的引用数和机构世界排名进行一些相关性分析；如果为一个航班延误数据集添加天气数据，则可以进一步挖掘航班延误和天气之间的关系；(3) 分析质量感知的数据清洗：可视化和可视分析结果的质量与数据密切相关，数据错误可能会导致可视化结果出现偏差，从而误导用户得出错误的结论。相较于直接使用通用的数据清洗算法检测和修复数据全集的错误，分析质量感知的数据清洗旨在考虑用户的可视分析查询，只检测和清洗与可视化和可视分析结果高度相关的数据子集，从而实现降低数据清洗的代价并提高可视分析结果的质量。

接下来，本节将详细介绍面向数据可视分析的数据准备如何基于上述思想进行数据获取（第 3.1 节）、数据增强（第 3.2 节）和数据清洗（第 3.3 节），如图 8 所示。

3.1 分析意图驱动的数据获取

在进行数据可视分析之前，首先要获取数据。相关工作可以分为两类：数据探索，即获取满足用户可视分析意图的数据^[53,55-57]，以及数据生成，即生成由于隐私保护策略难以获取的可视化数据^[47,50-52,60]。

AIDE^[53] 可以帮助用户找到满足可视分析意图的、感兴趣的数据。在对原始数据按照用户可视分析意图进行筛选后，用户可以专注于可视分析自己感兴趣的数据。AIDE 认为用户的分析意图可以用一个 SQL 查询 Q 来表示， Q 的查询结果即为满足用户分析意图的数据。AIDE 根据用户标注的数据来预测 Q ，然后将 Q 的查询结果返回给用户，用户可以在 AIDE 返回的数据上进行进一步的可视分析。AIDE 为用户提供原始数据中的若干元组供用户标注，用户可以将这些元组标注为“真”或“假”：如果该元组满足用户的分析意图（即用户对该元组感兴趣），用户将该元组标注为“真”；如果该元组不满足用户的分析意图（即用户对该元组不感兴趣），用户将该元组标注为“假”。基于用户的标注数据，AIDE 训练一个二分类（决策树）模型，原始数据中被该模型预测为真的元组即为满足用户分析意图的数据。需要注意的是，用户标注的数据是由系统提供的，且用户标注的数据量是很小的，因为标注数据往往需要耗费较大的人力代价。因此，选择哪些元组供用户标注以获得最好的标注效益（即在标注少量数据的情况下就可以准确预测用户的分析意图）是一个重要的问题。主动学习（Active Learning）^[61] 技术可以用来选择训练数据，从而快速训练得到性能较好的机器学习模型。主动学习中经常选择具有高多样性^[53,61-63] 和高模型不确定性的^[64-67] 数据供用户标注，以最小化用户标注代价。AIDE 选择数据空间中具有代表性的数据点供用户标注。

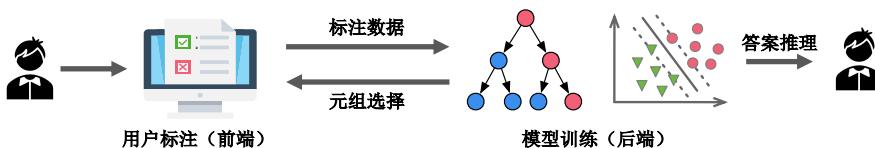


图 9 DExPlorer：分析意图驱动的数据获取

与 AIDE^[53] 类似, DExPlorer^[55-57]也可以帮助用户获取满足用户可视分析意图的、感兴趣的数据, 同时, DExPlorer 还可以将这些数据按照用户感兴趣程度的高低进行排序。例如, 考虑一个二手车数据集, 用户想要获取满足一定品牌、型号、里程数的汽车, 并将满足条件的二手车按照价格和马力的某种组合进行排序, 比如按照线性函数 $-0.018 \times \text{price} + 0.982 \times \text{powerPS}$ 对二手车进行排序, 即用户想要价格较低、马力较高的二手车。DExPlorer 可以帮助用户获取满足上述条件的二手车, 并将这些二手车按照用户心中隐含的排序意图进行排序。与 AIDE 类似, 在 DExPlorer 中, 用户也需要对数据进行标注, 然后 DExPlorer 根据用户的标注数据预测用户的分析意图和排序意图。如图 9 所示, DExPlorer 系统由前端和后端两部分组成:

(1) 前端: DExPlorer 的前端为用户展示一个包含 k 个元组的列表, 用户对列表中的元组进行标注。由于 DExPlorer 不仅要找到满足用户分析意图的数据, 还要将这些数据进行排序, 因此 DExPlorer 不仅要求用户提供真假标注, 还需要用户提供偏序标注。(1) 真假标注: 用户需要标注列表中元组的“真假”, 即将满足用户分析意图的元组标注为真, 将不满足用户分析意图的元组标注为假; (2) 偏序标注: 用户需要对标注为真的元组按照感兴趣程度的高低进行排序, 即将更感兴趣的元组排在靠前的位置。用户的真假标注和偏序标注结果将会传至后端。

(2) 后端: DExPlorer 的后端接收来自前端的标注数据, 根据这些标注数据训练机器学习模型, 从而进行答案推理(即预测满足用户可视分析意图的数据, 并将这些数据按照用户感兴趣程度排序)和元组选择(即选择下一轮要标注的元组)。(1) 答案推理: DExPlorer 训练二分类(随机森林)模型来预测满足用户可视分析意图的数据, 以及 RankingSVM 模型来对这些数据进行排序; (2) 元组选择: 不同于传统的仅考虑不确定性和多样性的主动学习方法, DExPlorer 既考虑了分类模型和排序模型的不确定性, 又考虑了标注列表中包含的 k 个元组之间的多样性。然而, 最大化元组不确定性和多样性的元组选择问题是一个 NP 难问题, 因此论文提出了基于动态规划的启发式算法来快速解决此问题。

在很多情况下, 为了保护用户的隐私信息, 相关公司和机构不会公布其所拥有的数据, 这导致数据分析师无法获取数据进行相应的分析。因而, 许多工作^[47,50-52,60,68]致力于生成与真实数据相似的数据, 生成的数据可以用于一系列下游应用, 比如, 可视化、机器学习、SQL 查询等。近年来, 有一些工作^[47,50,51,60]尝试使用深度生成对抗网络(Generative Adversarial Network, GAN)来生成关系型数据。这些工作首先训练 GAN 模型来学习真实数据的分布, 然后使用 GAN 的生成器来生成与真实数据分布相似的新数据。GAN 模型由两部分组成: 生成器 G 和判别器 D。生成器 G 接收一个随机噪音, 然后神经网络(比如全连接神经网络、卷积神经网络、长短期记忆网络等)将该噪音转换为一个元组, 神经网络的不同神经元的输出对应元组的不同属性; 判别器 D 是一个二分类模型, 可以判断一个元组是真实存在的还是由生成器生成的。G 企图生成被 D 认为是真实元组的元组, D 则试图正确区分真实元组和由 G 生成的元组, 因此 G 和 D 的训练过程是一个不断对抗的过程。当 G 和 D 达到均衡状态时, GAN 模型训练完毕, 此时 GAN 的生成器 G 可以用来生成新的数据。

基于 GAN 的关系数据生成方法^[47,50,51,60]可以生成与真实数据分布非常相似的数据, 但由于其训练数据是真实数据, 所以很可能泄露真实数据的隐私信息。SERD^[52]则可以在保护真实数据隐私信息的前提下生成与真实数据相似的数据, 因此数据分析师可以在生成的数据上进行相应的数据分析, 而不用担心隐私问题。SERD 可以生成与真实实体解析数据集的实体对的相似度向量分布相似的实体解析数据集, 因而使用生成实体解析数据集和真实实体解析数据集训练的实体解析模型相似(即在同一个测试集上有相似的性能)^[47,51], 因而使用生成实体解析数据集训练的实体解析模型可以直接用在真实实体解析数据集上。同时, 生成的实体与真实实体非常相似, 人们很难辨别一个实体是生成的还是真实存在的。

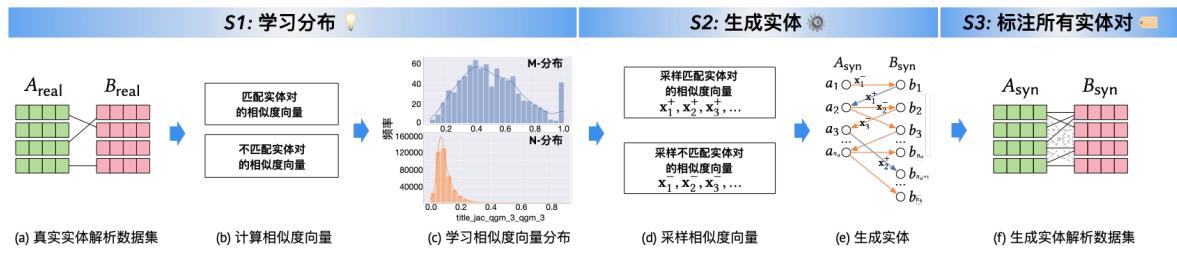


图 10 SERD: 隐私保护的数据生成流程

图 10 展示了 SERD 生成实体解析数据集的过程，共包含三步：(1) 学习分布：SERD 首先计算真实实体解析数据集实体对的相似度向量，然后分别计算匹配和不匹配实体对相似度向量的混合高斯分布；(2) 生成实体：SERD 从已经生成的实体中采样实体 (SERD 会首先自动地生成第一个实体以进行冷启动)，从学习到的相似度向量中采样相似度向量，然后根据采样得到的实体和相似度向量生成新的实体——新实体和采样的实体的相似度向量接近采样的相似度向量。SERD 使用满足差分隐私的 Transformer 模型生成具有语义信息且满足相似度向量的新实体，使用 GAN 模型来拒绝与真实实体不相似的生成实体。从匹配相似度向量分布中采样的相似度向量对应的实体对被标记为匹配，从不匹配相似度向量分布中采样的相似度向量对应的实体对被标记为不匹配；(3) 标注所有实体对：在生成所有实体后，对于没有匹配或不匹配标签的实体对，按照实体对对应相似度向量属于匹配或不匹配相似度向量分布的概率对其进行标记。SERD 的生成算法满足差分隐私，因而攻击者无法通过生成实体解析数据集获取真实实体的隐私信息。

3.2 领域知识指导的数据增强

在对数据进行可视分析时，可以对数据进行扩充，包括为原始数据增加更多的列（即属性）以丰富可视分析的维度，或者为原始数据增加更多的行以获得更多的分析样本。相关工作包括基于知识图谱的数据增强^[58,69,70]和基于数据湖的数据增强^[59,71]。

CAVA^[58] 是一个使用知识图谱来对数据进行列增强的可视分析系统。很多可视分析系统都假设用户在进行数据分析时，数据已经完全准备完毕，比如数据的列（即属性）已经固定。在这些系统中，数据增强与数据分析是相互分离，互不影响的。但是 CAVA 可以支持用户在数据分析的同时对数据的列进行扩充，即 CAVA 中的数据增强和数据分析是紧密结合的：只有有助于后续分析任务的属性才会被扩充到原始数据集上。CAVA 首先将原始数据集中的每一行映射到知识图谱中的一个实体节点，然后检查这些实体节点在知识图谱中的可以作为该实体属性的邻居节点。被多个实体节点作为邻居节点的属性可能会被扩充为原始数据集的新属性。例如，用户想要可视化一个包含各个国家基本信息的数据集，CAVA 会将该数据集中的每一行映射到知识图谱中的一个国家节点，比如德国、法国等。而这些国家节点在知识图谱上都存在一个表示国土面积的相邻节点（这些节点和相邻节点的边表示属性“面积”），则 CAVA 可以为原始数据集扩充属性“面积”，并将国家节点相邻节点的面积属性的取值填充到对应行的“面积”属性。CAVA 会为原始数据集推荐多个可扩充的属性，比如对于上述国家数据集，CAVA 可能会推荐“面积”、“人口数”、“GDP”、“相邻国家”等属性。CAVA 会将这些推荐的扩充属性的分布、数据质量、知识图谱结构等以可视化的形式展示给用户，用户可以从中选择需要扩充的属性。用户研究表明，CAVA 推荐的扩充属性可以有效地提升用户数据可视分析的结果。

Juneau^[59] 可以从数据湖中发现与当前分析数据表相关的表，并将此搜索功能集成到数据分析环境（比如 Jupyter Notebook）中，从而帮助数据科学家们随时发现与当前分析任务有用的表。在用户进行交互式数据可视分析的过程中，用户可以随时使用 Juneau 通过关键字、数据表等搜索发现有用的表，比如为机器学习任务发现更多的训练数据或训练特征、为当前数据连接更多的属性以进行可视化、为数据清洗任务发现缺失的数据等。Juneau 定义了一系列衡量表的相关性的评估指标，包括行和列的重合程度、数据模式匹配程度、

数据源的相似程度等。同时, Juneau 使用了剪枝、近似查询等技术来加速搜索相关表的过程。

3.3 分析质量感知的数据清洗

现实世界中的数据往往都会包含一定的数据错误, 比如重复记录、别名、缺失值、异常值等, 直接在脏数据上进行数据分析可能会误导用户得出错误的分析结论^[72]。因此, 数据清洗是数据管理和分析中重要的一个环节。然而, 现有的研究和实践表明数据清洗往往占据了数据科学家 60% 的精力^[71-73]。虽然目前有许多代表性工作研究基于机器算法的全自动数据清洗或者基于人机结合的交互式数据清洗模型^[74,75], 但是这些工作大多考虑检测和清洗整个数据集中的数据错误, 尽可能将数据集中的所有错误检测出来并做相应的修复, 这也导致了数据清洗的代价过高。

在可视化和可视分析领域, 目前常见数据清洗工具有 Tableau Prep (<https://tableau.com/products/prep>)、OpenRefine (<http://openrefine.org/>) 和 Data Wrangler^[76]。Tableau Prep 支持在 Tableau 中进行数据集的整合和清洗, 提供了过滤、拆分和重命名等操作以清理和转换数据集。OpenRefine 支持清洗重复记录和简单的字符串转换。Data Wrangler^[76] 则提供了一些简单的字符串转换, 例如替换、字符串拆分等。

在可视化和可视分析领域, 除了直接应用通用的数据清洗算法进行数据清洗以提高数据分析的质量, 还可以结合可视化和可视分析的特性, 进行任务驱动的数据清洗, 以降低数据清洗的代价, 提高可视化和可视分析的效率和质量。一方面, 根据可视化的特性, 用户可以通过可视化结果主动发现数据集中存在的数据错误(如异常趋势和异常值); 另一方面, 数据分析往往只查询和处理数据集的子集, 因此可以只清洗对分析查询结果影响较大的数据错误^[77,78]。下面本文将给出一个阐述性例子, 如图 11 所示, 数据集 D1 包含了重复记录、别名、异常值和缺失值四类数据错误。如果直接在 D1 上进行可视分析, 可能会得出错误的分析结果, 例如用户通过可视化查询 Q1 可视化柱状图来展示各个会议论文引用值总和的排名, 可以看出由于异常值、重复记录和同名等错误的影响, 所得柱状图 V1 是不准确且容易误导用户得出错误结论。如果将相同的可视化查询 Q1 在清洗干净的数据集 D2 上进行可视化, 得出相应的柱状图 V3, 可以看出用户从 V3 得到的结论将和从 V1 得到的结论不同。然而, 并不是数据错误对所有的可视化结果都有显著的影响, 例如可视化查询 Q2 无论在 D1 或 D2 上, 其可视化结果都是一样的, 可以避免用户得出不同的分析结论。

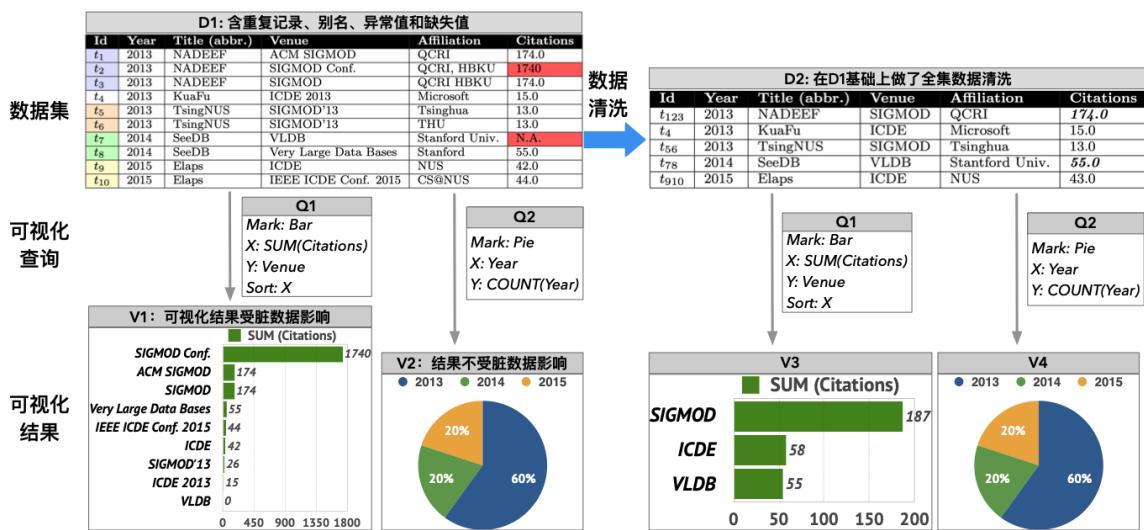


图 11 脏数据对可视化结果的影响

VisClean^[77,78]是基于上述思想的分析质量感知的数据清洗系统。VisClean 考虑用户的分析查询, 结合交互式数据清洗的渐进式可视化, 可以在包含错误的数据集中通过渐进式可视化来最终生成高质量的可视化结果, 辅助用户进行可视化分析, 避免错误的可视化结果误导用户得出错误的分析结果。如图 12 所示, 原始

的脏数据为 D_1 , 对于用户的一个可视化查询 Q_1 , 在未进行数据清洗时, 查询结果包含较多的数据错误, 也导致了可视化结果不精准; 然后, VisClean 会自动检测数据集中可能包含的与当前可视化结果有关的错误, 并构建一个 Error-and-Repairing (ERG) 图来组织所有检测出来的错误, 同时通过收益模型从 ERG 中选出一个潜在收益最大的子图, 该子图包含需要用户回答的问题, 然后根据用户的回答结果来对数据进行清洗 (清洗后的数据为 D_2), 同时更新 Q_1 的查询结果, 得到一个较好的可视化结果; 用户会不停地迭代执行上述步骤, 直到清洗得到了好的数据 D_n 和好的可视化结果。

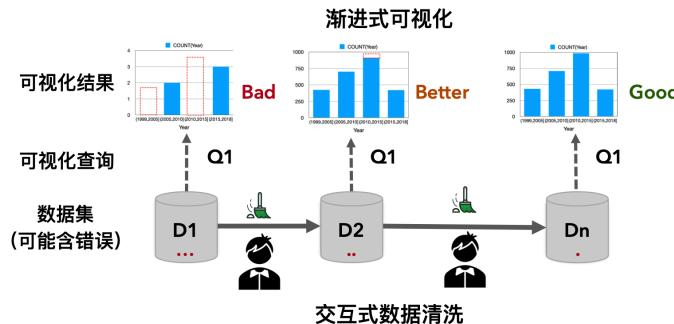


图 12 分析质量感知的数据清洗

VisClean 主要解决了两个问题: (1) 如何在脏数据中通过交互式的数据清洗来高效地提高可视化的质量; 以及 (2) 如何减少用户的交互代价, 即让用户少回答问题。针对第一个问题, VisClean 系统首先根据后台的数据清洗模型来自动地检测数据集可能的错误并产生相应的修复建议; 针对第二个问题, VisClean 提出了一个收益模型来估计每次与用户交互带来的收益, 即对可视化结果质量的提升。然后根据该收益模型, 提出了一个高效的算法来选取一组潜在收益最大的数据清洗问题与用户进行交互。

与现有的数据清洗算法相比, VisClean 只对用户指定的可视化查询所涉及到的数据自己进行清洗, 即只对可能影响可视化结果的数据错误进行清洗, 而不是对整个数据集的所有数据和所有错误进行清洗, 因此可以大大节省用户可视化分析过程中清洗数据的代价, 同时提高可视化和可视分析的质量和效率。

4 智能数据可视化

数据可视化是可视分析工作流中最为关键的一个环节, 它将原始数据映射成可视化图表, 帮助用户快速地捕捉到可视化图表传递的数据信息。然而, 从数据表中创建出能辅助用户理解数据或呈现数据中蕴含的知识的可视化结果却是一项富有挑战的任务。首先, 可视化是一个繁琐且需要相关专业技能的任务, 需要用户掌握特定的编程语言或者数据可视化系统, 具有门槛高的特点。其次, 普通用户甚至是许多分析师都需要用大量的时间来理解数据集的语义, 以选择用于可视化的数据属性, 执行恰当的数据转换操作, 最后选择合适的可视化图表进行渲染呈现, 整个过程是循环迭代, 存在交互模式效率低的特点。此外, 因为数据可视化高度依赖于用户的专业技能, 对于复杂的数据或者复杂的数据分析任务, 可能会存在分析结果不精准和分析维度不全面的情况。概括来说, 数据可视化存在可视分析高门槛、交互模式效率低、分析结果不精准、分析维度不全面等挑战^[79-83]。

为了解决上述挑战, 智能数据可视化技术应运而生, 其目标是根据当前分析数据集的数据特征、分析任务、用户意图和领域知识等, 自动地推荐若干有意义的可视化结果给用户, 帮助用户快速理解数据集及其蕴含的知识和规律^[4,5,10-12,35,36,84-86]。

本章首先概述智能数据可视化的基本概念和技术要点, 然后梳理智能数据可视化的代表性工作。

4.1 智能数据可视化概述

本节将围绕图 13 概述智能数据可视化的推荐流程、实现方法和推荐维度。

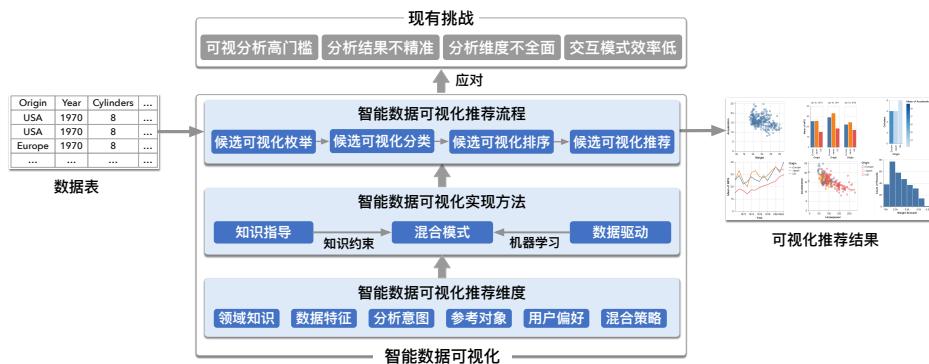


图 13 智能数据可视化推荐技术要点

4.1.1 智能数据可视化推荐流程

为了从数据集中推荐若干有意义的可视化结果，以帮助用户更好地理解数据集和解构数据中蕴含的知识和规律，智能数据可视化的推荐流程通常包括候选可视化枚举、分类、排序和推荐 4 个步骤：

(1) **候选可视化枚举**：候选可视化枚举的主要目的是为了确定潜在有意义的可视化结果，用于后续的可视化分类、排序和推荐环节。在枚举候选可视化之前，必须要明确定义候选可视化枚举的搜索空间。如图 4 所示，数据可视化创建的过程分别需要在数据空间和可视化空间执行特定操作。因此，数据可视化的搜索空间可以看成是数据空间和可视化空间所有可能操作的枚举组合。例如，对于给定的数据表，数据空间的操作包括对于数据表的过滤、选择、聚集等操作；对于可视化空间的操作包括可视化图表 X/Y 轴的映射、可视化图表类型的映射和可视化颜色等。因此，现有的研究工作主要从可视化查询的枚举组合^[5,23,35,87,88]或者基于可视化领域知识/规则约束^[7,18,89]来候选可视化枚举的搜索空间，进而支持候选可视化的高效枚举。

Luo 等人^[5,35,36]根据可视化过程中可视化查询在数据空间和可视化空间的枚举组合，确定了一个 m 列（即 m 个属性）的数据表，其二维可视化的搜索空间，如图 14 (a) 所示。对于 m 列的数据表，当 DeepEye 选择其中 2 列数据用于可视化时，先后分别需要进行数据选择 (SELECT)、数据转换 (Transform)、数据排序 (Order) 和可视化映射 (Visualize) 4 个操作。(1) 对于数据选择 (SELECT) 操作，共有 $m(m-1)$ 种枚举组合。(2) 对于数据转换操作，对 X 轴数据，可以执行分桶 (Binning)、分组 (Group By) 或者不做数据转换操作。其中，对于分桶操作，可以对日期类型的数据按分钟、小时、天等 7 种时间尺度进行分桶；对于数值类型数据，可以按特定步长进行分桶；或者可以基于用户自定义函数进行分桶操作，因此共有 9 种分桶操作。对于 Y 轴数据，可以执行 3 种聚集操作或者不做聚集操作，该步骤共有 $(1+9+1) \times 4 = 44$ 种枚举组合。(3) 对于排序操作，可以选择对 X 轴或 Y 轴数据进行排序，或者不执行排序操作，则共有 3 种可能性。(4) 对于可视化映射操作，可以根据预定义规则映射为柱状图、饼状图、折线图和散点图四类可视化图表，有 4 种可能性。因此，上述 4 个操作步骤的枚举组合为 $m(m-1) \times 44 \times 3 \times 4 = 528m(m-1)$ ，即 DeepEye 的可视化搜索空间为 $528m(m-1)$ 。

Draco^[89] 则基于答案集编程 (Answer Set Programming, ASP) 将可视化查询表达为逻辑事实 (Logical Facts)，这些的逻辑事实可用于检查可视化查询是否满足可视化设计的领域知识和相关限制。如图 14 (b) 所示，灰色方框表示可视化搜索空间，橙色方框为 Draco 的可视化搜索空间。Draco 的可视化搜索空间是基于一系列的可视化领域知识和相关限制来确定的，主要包含两类规则和限制：(1) Draco 首先定义了聚集规则 (aggregate rules) 去限制可视化查询每一个分项可以填充的具体值。例如可视化映射分项 Mark 可以填充 Bar、Pie、Point 等值用于表示可视化图表类型；(2) Draco 通过定义完整性约束 (Integrity Constraints) 来解决可视化查询分项之间的完整性和冲突。

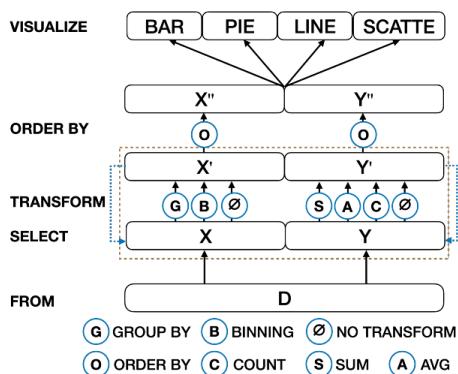
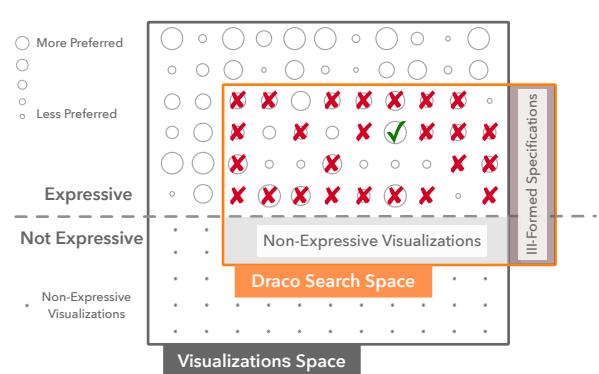
(a) DeepEye^[5] 的可视化搜索空间 (二维可视化)(b) Draco^[89] 的可视化搜索空间

图 14 两类可视化搜索空间定义方式

由于数据可视化的搜索空间巨大，对于候选可视化的枚举是十分耗时的，因此现有的方法几乎都会基于可视分析的任务场景、领域知识和专家规则对可视化的搜索空间进行剪枝，以提高可视化枚举的效率。

(2) **候选可视化分类**: 对于可视化搜索空间的任一候选可视化，智能数据可视化系统需要判断该可视化是否是有意义的，即是否选择该可视化结果用于后续的推荐环节。对于候选可视化的分类，现有工作通常从数据特征^[5,35,36]、美学指标^[90]、领域知识^[91]和分析任务^[7,92]等维度来评判候选可视化的“好/坏”。因为候选可视化分类环节（任务）本质上是评价可视化结果的分类任务，因此现有工作通常采用基于专家规则的启发式算法、机器学习分类模型或深度学习模型来实现。对于基于启发式算法的工作（如 CompassQL^[18]），通常是采用算分函数来给候选可视化赋分；对于基于机器学习分类模型的工作（如 DeepEye^[5,35,36]），通常是基于标注数据训练一个分类器并使用该分类器对候选可视化进行分类（通常是二分类任务）；对于基于深度学习模型的工作（如 VizML^[90]），也是基于标注数据训练神经网络并用其预测相应的可视化查询参数。

(3) **候选可视化排序**: 给定两个候选可视化，候选可视化排序可基于不同的排序策略（如用户偏好）来评估哪一个候选可视化“更好”，即反映其“价值”。候选可视化分类和排序是候选可视化推荐的基础，即前者反映了单个可视化的“单一价值”，后者则反映了多个可视化的“组合价值”，可用于可视分析仪表盘推荐或者可视化结果列表推荐等任务。

(4) **候选可视化推荐**: 给定一个数据集和 n 个候选可视化，候选可视化推荐任务旨在选择 k 个可视化结果返回给用户 ($k \leq n$)。如果考虑这 k 个可视化结果内部的先后顺序，则可以看成是 top- k 可视化结果列表推荐任务；如果不考虑这 k 个可视化结果内部的先后顺序，则可以看成是 k 个可视化结果集合推荐任务，可支持自动构建可视化分析仪表盘等具体应用场景。

值得注意的是，并不是所有的智能可视化工作都严格遵循上述 4 个步骤来实现端到端的可视化结果生成和推荐，如图 16 中介绍，部分工作不考虑多个可视化之间的排序^[18,93,95]；而另一部分工作则采用深度学习模型并基于用户自然语言查询意图以推荐相应的可视化^[6-7,96]，这类工作将可视化推荐工作看成是查询翻译任务，尽管没有严格遵循上述 4 个步骤，但这类工作所采用的深度学习模型本质上也是做分类和排序任务。

4.1.2 智能数据可视化实现方法

基于上述的 4 个智能数据可视化推荐流程，本文基于现有研究工作，总结了数据可视化推荐的 3 种实现方法，分别是知识指导、数据驱动和混合模式：

(1) **知识指导**: 知识指导的智能数据可视化推荐方法主要是通过构建可视化领域的知识图谱或者通过对可视化领域知识进行编码，融合相关的领域知识进行可视化结果的推荐。知识指导的方法首先限定了候选可视化的搜索空间，其次为候选可视化的分类、排序和推荐提供了依据。常见的知识指导的智能可视化有 ShowMe^[97]、Voyager^[93]、CompassQL^[18]、Voyager2^[37]、QuickInsights^[91]、NL4DV^[99]、TaskVis^[92]、

Lux^[98]、SeeDB^[87]、Zenvisage^[23,88]、VisPilot^[100]、VISER^[101]和BDVR^[102]。

(2) **数据驱动:** 数据驱动的智能数据可视化推荐方法主要是指通过机器学习和深度学习等手段, 基于大量的可视化语料库学习可视化结果的推荐。常见的数据驱动的智能数据可视化有 Data2VIS^[103]、VizML^[90]、Table2Charts^[104]、VizRec^[105]、PVisRec^[106]、RGVisNet^[107]、VisGNN^[109]和VisGuide^[110]。

(3) **混合模式:** 混合模式是指系统融合了知识指导和数据驱动的方法, 共同进行可视化结果的推荐。例如, 通过知识指导的方式获得一系列可视化领域知识的约束, 用来限定可视化的推荐空间, 并结合数据驱动的方式, 通过机器学习排序模型进行可视化结果的推荐。常见的混合推荐的智能可视化有 SEQ2VIS^[6]、ncNet^[7]、Sevi^[96]、DeepEye^[5,35,36]、Draco-Learn^[89]、VizDeck^[111]、Dziban^[112]和KG4VIS^[113]。



图 15 在不同推荐维度和实现方法视角下的智能数据可视化相关工作

4.1.3 智能数据可视化推荐维度

本节基于现有研究工作, 总结了智能数据可视化推荐的 6 个维度, 分别是领域知识、数据特征、混合策略、分析意图、参考对象和用户偏好, 如图 15 所示。

(1) **领域知识:** 领域知识是可视化和可视分析中必不可少的先验知识, 可以指导分析师挖掘数据中蕴含的信息。以可视化映射为例, 常见的领域知识有, 两列数值类型的数据适于使用散点图进行呈现。领域知识指导的可视化推荐可以将可视化领域的知识融合到可视化推荐过程中的候选可视化枚举、候选可视化排序和推荐环节。常见的领域知识指导的可视化推荐方法有 ShowMe^[97]、Voyager^[93]、CompassQL^[18]、Voyager2^[37] 和 QuickInsights^[91]。

(2) **数据特征:** 数据可视化是基于数据的, 从数据特征的维度考虑, 可以挖掘数据集蕴含的知识和规律。例如, 如果算法检测到一列数值数据包含异常值, 可以使用箱型图进行可视化, 以给用户进一步进行评估。数据特征驱动的可视化推荐考虑数据中特殊的分布、趋势和值, 以可视化的形式推荐给用户。常见的数据特征驱动的可视化推荐方法有 Data2VIS^[103]、VizML^[90] 和 Table2Charts^[104]。

(3) **分析意图:** 用户的分析意图可视化和可视分析指明了方向和目标。例如, 用户的分析意图是找到若干个能反应不同地区 GDP 产值比较的可视化结果, 这时系统可能需要推荐展示不同地区 GDP 产值的柱状图给用户。从分析意图维度出发的智能可视化推荐, 需要对分析意图进行解析, 进行相应的数据属性选择、数据转换操作和可视化编码操作。常见的分析意图融合的可视化推荐方法有 NL4DV^[99]、TaskVis^[92]、SEQ2VIS^[6]、ncNet^[7]、Lux^[98]、Sevi^[96] 和 RGVisNet^[107]。

(4) **参考对象:** 参考对象是指系统根据指定的参考可视化, 找到在某种维度上与之相似或者不相似的其他可视化结果。例如, 用户指定了某地 GDP 产值的折线图, 要求系统找到其他所有地区趋势相似的折线图。因此, 从参考对象维度出发的智能数据可视化推荐, 需要自动地进行数据属性的选择、数据转换和可视化编码操作, 推荐给用户符合参考对象特性的可视化结果。常见的基于参考对象的可视化推荐方法有 SeeDB^[87]、Zenvisage^[23,88]、VisPilot^[100]、Dziban^[112] 和 VISER^[101]。

(5) **用户偏好:** 用户是数据可视化和可视分析的中心, 不同用户可能会对不同的数据属性和可视化编

码方式有不同的偏好。因此，从用户偏好维度出发的智能数据可视化推荐，需要融合用户对数据属性和可视化编码方式的偏好进行数据可视化结果的推荐。常见的考虑用户偏好的可视化推荐方法有 BDVR^[102]、VisGuide^[110]、VizRec^[105]、PVisRec^[106] 和 VisGNN^[109]。

(6) **混合策略：**混合策略是指融合了上述 5 种推荐维度的若干种。常见的混合策略是将领域知识与数据特征相结合，例如首先基于数据特征自动地推荐若干从数据角度有意义的可视化结果，然后结合领域知识仅保留符合可视化审美的结果。常见的基于混合策略的可视化推荐方法有 VizDeck^[111]、DeepEye^[5,35,36]、Draco-Learn^[89]、KG4VIS^[113] 和 Lux^[98]。

图 15 总结了在不同的推荐维度和实现方法视角下的智能数据可视化相关工作。例如，基于数据驱动的实现方法中，Data2VIS^[103]、VizML^[90] 和 Table2Charts^[104] 考虑基于数据特征进行可视化结果的推荐；VizRec^[105]、PVisRec^[106] 和 VisGNN^[109] 则考虑从用户偏好维度进行可视化结果的推荐。

通过图 15 的比较分析，可以看出主流的工作分别从领域知识、用户偏好和混合策略三个维度出发设计智能数据可视化方法。在每个维度方法的具体实现方面，基于领域知识和混合策略的方法几乎采用基于知识指导（如专家规则和知识图谱）的实现方法进行实现。本文认为，未来智能数据可视化方法在一定程度上会从混合策略出发，考虑可视化推荐流程的用户、任务和数据等多方目标，结合知识指导和数据驱动方式，实现一个能支持多场景、多任务、个性化和自适应的智能数据可视化系统。

	系统	年份	核心方法	用户主要输入	可视化主要类型	推荐模式		推荐空间		优点	不足	基于学习	主要模型
						单个结果	多个结果	数据空间	可视化空间				
领域知识 数据特征	Show Me	2007	基于可视化规则系统，进行可视化结果推荐	数据表	B/L/P/S/H	✓			✓	基于专家知识，可以满足基本需求	高度依赖专家知识，手工配置耗时耗力，灵活性不高	X	X
	Voyager	2015	基于Compass推荐引擎，通过一系列启发式规则，自动从数据表属性选择，数据转换和可视化映射三个阶段进行可视化结果的推荐	数据表 部分可视化查询	B/L/S/H		✓	✓	✓	形式化定义了可视化推荐的三个核心阶段，开发了一个完整的可视化推荐系统	高度依赖Compass编码的规则	X	基于启发式规则
	CompassQL	2016	CompassQL是一种可视化推荐语言，给定一个部分完整的CompassQL，它可生成一系列启发式规则以解决缺失部分的语义以进行可视化的结果推荐	数据表 部分可视化查询	B/L/S/H		✓	✓	✓	通过可视化推荐语言的形式限定了可视化推荐的推荐空间	收敛候选的可视化推荐结果代价过高	X	X
	Voyager2	2017	基于CompassQL可视化推荐语言，在给定部分完整的CompassQL作为输入的基础上生成并推荐相关的可视化结果	数据表	B/L/S/H		✓	✓	✓	基于CompassQL设计并开发了一个完整的可视化推荐系统，具有良好的用户体验	仍然需要用户通过交互界面操作完成一个部分完整的可视化推荐构造，还未做到全自动化地可视化推荐	X	CompassQL
	QuickInsights	2019	通过定义的洞察挖掘框架，设计高效算法，自动计算给定数据的若干洞察结果，以可视化的方式返回给用户	数据表	B/L/P/S		✓	✓		所设计的算法高效。能基于已定义的洞察挖掘框架快速地进行数据集的洞察挖掘	所提出的洞察挖掘框架存在大量的启发式规则，自适应性不强	X	基于启发式规则
数据特征 分析意图	Data2Vis	2018	基于数据表，可视化数据库库进行序列到序列的数据挖掘和翻译模型训练，基于训练的模型在新数据集上自动推荐可视化结果	数据表	B/L/P/S		✓	✓	✓	方法简单，易于部署	只针对Vega-Lite可视化语法进行自动可视化推荐，支持的数据较少	✓	Seq2Seq with RNN
	VizML	2019	基于收集的数据，可视化语料库，通过神经网络从语料库中学习可视化设计	数据表	B/L/S/P/H		✓		✓	从大量的真实语料库中进行学习	专注于可视化设计方式	✓	全连接前馈神经网络
	Table2Charts	2021	将可视化推荐形式（表格到可视化的列例生成问题，基于Excel）可视化语料库，训练编译正确的SQL模型进行可视化结果的推荐	数据表	B/L/P/S/A/R		✓	✓	✓	从大量的真实语料库中进行学习	没有考虑分组、聚集等数据转换操作	✓	CopyNet as deep Q-network
分析意图 参考对象	NL4DV	2020	NL4DV基于CoreNL工具将用户输入的自然语言查询语句进行解析，并使用一系列启发式规则将映射到Vega-Lite可视化查询元素	数据表 自然语言查询	B/L/S/H		✓	✓	✓	支持端到端的基于自然语言查询的用户分析意图的可视化推荐	基于规则的自然语言查询理解可能行性和适配性不强	X	X
	TaskVis	2021	TaskVis将可视化分析中的分析任务通过可视化推荐的过程进行	数据表 用户指定的数据表 用户指定的分析任务	B/L/S/P/G/H		✓	✓	✓	总结了一系列可视化分析中的分析任务	基于规则进行结果的推荐，适应性较弱	X	X
	SEQ2VIS	2021	SEQ2VIS通过自然语言查询的可视化语料库和BERT进行序列到序列模型的训练，基于训练的SQL2Vis模型支持基于自然语言查询的用户分析意图驱动的可视化结果推荐	数据表 自然语言查询	B/L/P/S		✓	✓	✓	通过大量真实数据进行学习，支持端到端的基于自然语言查询的用户分析意图的可视化推荐	可视化推荐的可解释差	✓	Seq2Seq with LSTM
	ncNet	2021	ncNet基于Transformer模型在nvBench语料库上训练自然语言查询到数据可视化查询的翻译模型	数据表 自然语言查询 可视化模板（可选）	B/L/P/S		✓		✓	通过大量真实数据进行学习，通过最先进的Transformer模型学习基于自然语言查询的翻译模型并生成自然语言翻译模型，同时允许用户选择一个额外的可视化模板作为输入，提高鲁棒性	可视化推荐的可解释差	✓	Transformer-based ncNet
	Sevi	2022	Sevi是基于ncNet模型实现的语音查询到数据可视化查询的系统	数据表 自然语言查询	B/L/P/S		✓		✓	通过大量真实数据进行学习，构建了一个端到端的支持语音和自然语言交互进行大数据驱动的可视化推荐系统	可视化推荐的可解释差	✓	ncNet
参考对象 用户偏好	RGVisNet	2022	RGVisNet是一个检索生成混合式的基于自然语言查询的自动可视化模型	数据表 自然语言查询 可视化查询模版库	B/L/P/S		✓		✓	通过结合大量的可视化查询模版库，提高端到端的模型性能	高度依赖可视化模版库	✓	GNN
	SeeDB	2015	基于用户指定的可视化结果，推荐与之不相似的其他可视化结果	数据表 可视化	B		✓	✓		根据用户指定的参考可视化推荐最不相似的结果	支持的可视化推荐场景有限	X	启发式算法
	Zenvisage	2017	基于用户指定的折线图，推荐趋势与之相似的其他折线图	数据表 锚点可视化	L		✓	✓		根据用户指定参考可视化推荐相似的结果	支持的可视化推荐场景有限	X	启发式算法
	VisPilot	2019	根据用户指定的参考可视化，基于评估函数自动推荐基于执行操作后的可视化结果	数据表 锚点可视化	B		✓	✓		根据用户指定参考可视化，推荐最近支持的可视化结果	仅支持POLA中钻取（Drill-down）操作的可视化结果，帮助用户进行数据探索	X	启发式算法
	Dziban	2020	Dziban 使用 Draco 可视化知识库自动完成部分可视化推荐的全自动，并通过相似性算法支持推荐与提供的起点进行可视化相关的结果	数据表 锚点可视化 部分可视化查询	B/L/S		✓		✓	Dziban优化的算法自动推荐的可视化结果与用户给定的锚点可视化结果可能保持一致	支持的可视化推荐场景有限	✓	利用Draco进行快选可视化推荐，根据与给定的锚点可视化结果的相似度进行最近邻的推荐
混合策略	VISER	2020	基于按例可视化（Visualization by Example）的思想，根据用户提供的数据表和部分可视化行为样例输入，通过代码生成的方式，在数据表上推荐其相对应可视化形式的结果	数据表 锚点可视化	B/L/S		✓	✓	✓	可以根据用户提供的样例进行相应的可视化推荐，可以支持按例可视化推荐	难以支持通用的可视化推荐场景	X	基于代码合成的推荐算法
	BDVR	2009	基于用户交互数据进行个性化可视化结果推荐	数据表 用户实时交互数据	B/L/S		✓			基于用户交互记录，考虑用户对于可视化的个人喜好	用户偏好场景检测不够，算法适应性弱	X	基于启发式算法
	VizRec	2016	基于协同过滤算法进行个性化可视化结果推荐	数据表 用户信息	B/L/T/G		✓	✓	✓	考虑进行个性化推荐	可生成化的算法简单，且基于协同过滤算法也简单	X	基于规则进行可视化推荐，基于协同过滤进行个性化推荐
	PVisRec	2021	基于用户交互记录的训练数据，训练个性化可视化推荐模型	数据表 用户信息	B/L/S		✓	✓	✓	考虑进行个性化推荐	对每一个用户都需要维护相对应的知识图谱，成本较大	✓	神经网络
	VisGNN	2022	将用户与可视化的交互数据表示成图，基于神经网络进行个性化可视化推荐，学习用户的偏好，用于后续的可视化推荐	数据表 用户信息	B/L/S		✓	✓	✓	考虑进行个性化推荐	没有考虑数据转换操作	✓	图神经网络
混合策略 参考对象	VizGuide	2022	VizGuide是一个半自动的数据可视化推荐系统，基于用户的输入的可视化数据表示成图，通过数据表与可视化的交互记录，学习用户的偏好，用于后续的可视化推荐	数据表 用户交互数据	B/L/P		✓		✓	考虑进行个性化推荐	没有考虑数据转换操作	✓	浅层回归
	VizDeck	2012	基于数据特征训练排序模型，结合启发式规则进行可视化结果的推荐	数据表	B/L/S		✓	✓		具有完整的系统实现	推荐模型迁移能力弱	✓	Linear Model
	DeepEye	2018	基于数据特征和领域知识协同，通过领域知识挖掘选取可视化推荐结果，通过训练分类模型“择好”的可视化推荐，通过训练排序模型和高效可视化推荐法推断有意义的可视化结果	数据表	B/L/P/S		✓	✓	✓	结合领域知识和数据特征协同进行可视化结果的推荐，推荐效果好且适应性高	专家规则的更新比较繁琐	✓	Rules-Decision Tree-RankNet
	Draco-Learn	2018	Draco-Learn的可视化规则表示成一系列的规则，并将可视化结果用可视化规则进行表示。基于此，Draco-Learn以系统地枚举所有可能的规则来挑选规则“好”的可视化推荐，通过训练排序模型和高效可视化推荐法推断有意义的可视化结果	数据表 部分可视化查询	B/L/P/S		✓		✓	将可视化规则表示成可编程的一系列规则，可扩展性强	没有很好地考虑数据集的数据特征与可视化推荐的关系	✓	RankSVM用于可视化结果排序
	KG4VIS	2021	基于真实的数据库可视化语料库构建知识图谱，基于知识图谱进行表示学习。基于此，KG4VIS通过知识图谱进行可视化推荐以及相关推荐规则的生成	数据表 可视化知识图谱	B/L/S/H/BP		✓	✓	✓	基于知识图谱进行可视化结果推荐，具有较好的解释性	所构建的知识图谱仅包含有限的可视化设计领域知识	✓	TransE-adv进行知识图谱的表示学习，并基于启发式算法进行可视化结果的推荐
	Lux	2022	基于Show Me类似的可视化规则，开发了一个良好支持Jupyter Notebook环境使用的可视化推荐工具包	数据表 部分可视化查询	B/L/S/H/G		✓	✓	✓	提供基于Jupyter Notebook环境的自动数据可视化推荐工具，方便在交互式数据科学的场景下进行实时数据分析	可视化推荐搜索空间较小	X	基于启发式规则

图 16 智能数据可视化代表性工作概览。可视化类型：B (Bar Chart,柱状图)，L (Line Chart,折线图), P (Pie Chart, 饼状图), S (Scatter Chart, 散点图), H (Heatmap,热力图), T (Timeline,时间轴图), G (Geographic Chart,地图), BP (Box Plot,箱型图), A (Area Chart,面积图), R (Radar Plot,雷达图)

基于上述讨论，本文梳理和总结了近些年智能数据可视化领域代表性工作的类别、发表年份、核心方法（思想）、允许的用户输入、支持的可视化类型、推荐模式和推荐空间、主要采用的算法/模型、以及系统的优缺点，如图 16 所示。在讨论各项代表性工作之前，本文首先基于图 16 的数据进一步分析这些代表性工作的特点，梳理这些代表性工作主要采用的算法/模型与年份的关系，推荐模式和推荐空间占比，如图 17 所示。从图 17 (a) 可以看出，在 2010 年之前，所有的智能数据可视化工作都是采用启发式算法进行实现。2010 到 2018 年，开始有部分工作采用基于学习的算法（如 Learning-to-Rank）实现，但采用启发式算法的工作仍占多数。得益于人工智能算法在可视化领域的应用，在 2018 年之后，基于学习的智能数据可视化代表性工作数量迅猛增长，并大幅度超过了基于启发式算法的工作。从图 17 (a) 的趋势可以看出，未来将会有越来越多的工作结合人工智能实现智能数据可视化推荐。从图 17 (b) 可以看出，70% 以上的工作都会同时考虑推荐数据空间和可视化空间的设计，仅有少数工作只考虑可视化空间的设计推荐。图 17 (c) 则反映绝大部分工作都会考虑推荐多个可视化结果，仅有 24% 的工作只考虑推荐单个可视化结果。

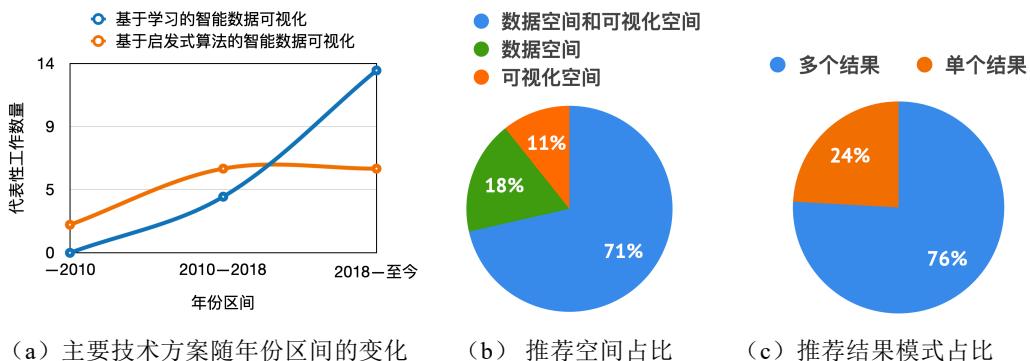


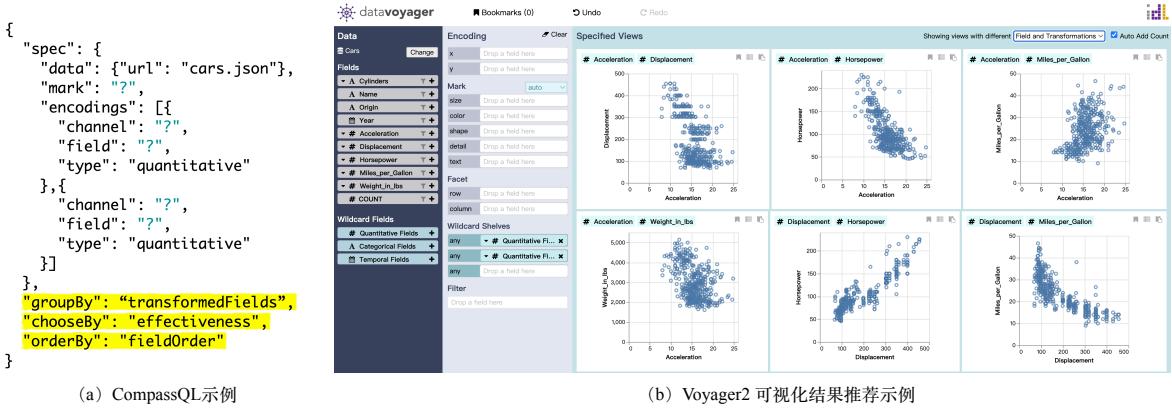
图 17 智能数据可视化技术特点分析

4.2 领域知识指导的数据可视化推荐

领域知识指导的数据可视化推荐方法旨在利用数据可视化和可视分析领域沉淀下来的领域知识，通过对领域知识进行编程表示，自动地完成数据可视化推荐的三个流程，即候选可视化的枚举、排序和推荐。常见的可视化领域知识例如一列数据值分布不均，可以使用散点图或者箱型图进行可视化。因为领域知识指导的方法实现简单，推荐效果在知识和规则考虑的范围内有保障，因此早期的智能数据可视化方法大多基于领域知识进行可视化结果的推荐。图 16 列举了近年来领域知识指导的数据可视化推荐的 5 个代表性系统。

早期的数据可视化推荐系统 ShowMe^[97] 基于一系列数据可视化映射的领域知识，自动地完成将数据属性映射到不同可视化元素的步骤（Automatic Marks）。例如，如果用户选定图 4 中 Cars 的分类型数据属性 Origin 并对数值型属性 Cylinders 做求平均的聚集操作，则 ShowMe 会将这两个数据属性自动映射为柱状图，其中柱状图的 X 轴是 Origin，Y 轴是 Avg(Cylinders)。基于 Automatic Marks 功能，ShowMe 可以支持用户通过交互界面增量式地选择想要进行数据可视化的数据属性，ShowMe 会自动地推荐合适的数据可视化结果。然而，ShowMe 高度依赖专家的领域知识，对于没有被领域知识覆盖的情况，其推荐效果较差。

与 ShowMe^[97] 类似，Voyager^[93] 也基于一系列领域知识（启发式规则）进行可视化结果的推荐。用户上传数据集到 Voyager 系统后，Voyager 会自动地推荐能反应该数据集每一列数据分布特征的单变量可视化结果（如直方图）；接下来，用户可以选择感兴趣的数据列，Voyager 系统会自动地推荐若干个与该数据列相关的可视化结果。与 ShowMe^[97] 主要推荐单个可视化结果不同，Voyager 可以推荐多个可视化结果。

图 18 CompassQL^[18] 查询语言和 Voyager2^[37] 可视化推荐结果示例

上述两个系统都是基于领域知识，通过启发式算法进行可视化结果的推荐，通常是将可视化推荐过程中的数据空间的推荐和可视化空间的推荐耦合在一起。这存在着编程困难、算法优化难和领域知识难扩展等问题。为了更好地形式化表达可视化的推荐过程，Wongsuphasawat 等人提出了面向数据可视化推荐的查询语言 CompassQL^[18]。CompassQL 的语法与 Vega-Lite 相似，如图 18 (a) 所示，CompassQL 包含了与 Vega-Lite 相似的 data、mark 和 encodings 字段，分别用于指定数据集、可视化类型和可视化映射。此外，CompassQL 还包括了为可视化推荐而设计 3 种的语法：1) 通配符；2) 分组；3) 选择和排序。如图 18 (a) 所示，通配符使用 “?” 表示，这表明用户可以选择部分感兴趣的数据或者可视化属性，使用通配符的地方表示需要系统进行推荐。基于上述通配符，CompassQL 会自动推荐一些候选的可视化结果，这些可视化结果可能彼此之间在数据属性/可视化编码方面比较相似，分组 (groupBy) 的作用是根据某种策略对这些结果进行分组展示，以减少推荐结果的冗余度，如图 18 (a) 的分组策略是按照数据列 (transformedFields) 进行分组展示。选择 (chooseBy) 和排序 (orderBy) 分别定义了两个评分函数，分别用于候选可视化结果的生成和排序。图 18 (b) 展示了 (a) 所示的 CompassQL 在 Cars 数据集上的可视化推荐结果。

基于 CompassQL 可视化推荐语言，Wongsuphasawat 等人^[37] 在 Voyager 的基础上开发了 Voyager2。Voyager2 是一个可以同时支持自动数据可视化推荐和用户交互式可视化的系统。图 18 (b) 展示 Voyager2 系统的前端，用户可以以拖拽的方式从 Data 区域选择若干感兴趣数据列到 Encoding 区域。用户的交互会被转换成 CompassQL 语言，如果 Voyager2 检测到 CompassQL 中还存在通配符，则自动进行可视化结果的推荐；反之，则根据用户指定的可视化属性和映射方式进行相应的可视化结果推荐。

QuickInsights^[91] 可以自动从数据集中挖掘数据洞察 (Data Insights) 并以可视化结果的方式展示给用户。QuickInsights 依据领域知识定义了 3 类共 12 种洞察类型，如趋势、相关性、季节周期等。基于此，其设计了一套基于搜索-评分的洞察挖掘框架，以自动地挖掘和推荐给定数据集中蕴含的数据洞察。

上述领域知识指导的数据可视化推荐方法可以依赖特定的领域知识，进行相关的可视化结果推荐，对于被领域专家知识覆盖的情况，其推荐的效果是非常不错的。然而，这类方法大多没有考虑数据集本身的数据特征，因此忽视了一部分可视化推荐的搜索空间。

4.3 数据特征驱动的数据可视化推荐

图 16 对比了近年来数据特征驱动的数据可视化推荐的 3 个代表性工作。这些工作的共性特点是基于数据集的数据特征，通过训练机器学习模型，来自动地推荐可视化结果。

Data2Vis^[103] 将数据可视化推荐任务看成是序列到序列的翻译任务，即从数据序列到可视化查询语言序列。基于这个假设，Data2Vis^[103] 基于双向循环神经网络 (BiRNN) 构建序列到序列的 Data2Vis 自动数据可视化模型。Data2Vis 的输入是经过预处理的数据序列，输出是用 Vega-Lite 查询语言表示的可视化序列。

Data2Vis 基于从 11 个不同数据集生成的 4300 个（数据序列，可视化序列）训练样本进行学习。虽然 Data2Vis 可以基于数据特征通过训练数据序列到可视化序列的模型进行可视化结果的推荐，但是 Data2Vis 存在可解释性差、数据转换操作少、泛化性差和仅支持 Vega-Lite 可视化语法等缺点。

相较于 Data2Vis 的小规模训练数据，VizML^[90] 考虑了 100 余个与可视化设计相关的数据特征并从百万量级的真实可视化数据中进行神经网络模型的训练，以学习自动数据可视化中可视化空间的自动推荐。相较于前面提到的数据可视化推荐工作，同时考虑自动推荐可视化过程中的数据空间和可视化空间的操作，VizML 仅考虑了两类共五种可视化空间的自动化设计推荐任务，没有考虑数据空间的推荐任务。例如，给定用于可视化的数据列，VizML 仅考虑这些数据列应该映射为何种可视化类型以及与可视化类型 X/Y 轴的映射关系。

与 Data2Vis 类似，Table2Charts^[104] 也将数据可视化的推荐看成是数据序列到可视化序列的学习任务。前文提到 Data2Vis 的可视化序列是用 Vega-Lite 语言表示，而 Table2Charts 使用一个通用的序列化可视化图表模板表示，该图表模板仅包含必要的可视化元素（如可视化类型、数据列等），因此，Table2Charts 可以支持使用多种可视化语言进行最终结果的渲染。相较于 Data2Vis 直接将数据序列作为直接输入，Table2Charts 对数据表中数据列和统计信息进行了表示学习。Table2Charts 通过具有复制机制和启发式搜索的深度 Q 学习进行数据序列到可视化模板序列的生成。

4.4 融合分析意图的数据可视化推荐

基于领域知识或者数据特征的数据可视化推荐系统可以推荐在领域知识或者数据特征视角下有意义的可视化结果给用户，但这些可视化结果不一样符合用户特定的分析意图。因此，融合分析意图的数据可视化推荐系统应运而生。概括而言，这类系统可以允许用户通过系统交互组件来直接指定或者间接指定用户的分析意图，系统根据用户的分析意图提示，推荐相关的可视化结果。

图 16 总结了近年来融合分析意图的数据可视化推荐的 5 个代表性方法。其中，这类方法又可以细分为基于交互组件直接指定分析意图（TaskVis^[92]）和基于自然语言间接指定分析意图（NL4DV^[99]、SEQ2VIS^[6]、ncNet^[7]、Sevi^[96] 和 RGVisNet^[107]）。注意，本文将 Lux^[98] 归类为混合策略下，因为 Lux 除了基于分析意图进行推荐外，还考虑了其它推荐维度（如基于用户交互记录等）。

TaskVis^[92] 首先总结并维护了 18 个在学术界和工业界常用的分析任务。当用户使用 TaskVis 进行可视化的时候，可以选择若干个其想要进行可视化和可视化分析的分析任务，TaskVis 会基于用户的选择，使用一套基于规则的可视化枚举和排序方法，产生若干个可视化结果推荐给用户。

NL4DV^[99]、SEQ2VIS^[6]、ncNet^[7] 和 Sevi^[96] 通过自然语言接口的方式获得用户的分析意图。概括而言，这些系统实现融合分析意图的数据可视化推荐的核心技术是理解用户的自然语言查询，并基于此推荐相关的可视化结果给用户，即其核心任务是将自然语言查询映射为对应的可视化结果（Natural Language to Visualization, NL2VIS）。

NL4DV^[99] 集成了数据可视化和可视分析的 5 种常见分析任务（相关性、分布和趋势等）。用户使用 NL4DV 时，首先通过自然语言接口输入对数据集可视化的意图（如“Visualize car weight and number of cylinders over the years.”），NL4DV 首先使用 Stanford CoreNLP^[114] 对该自然语言查询进行依存句法分析，然后基于一系列启发式规则推理出用户感兴趣的数据属性和潜在的数据分析任务。最后，NL4DV 基于解析出来的数据属性和分析任务，基于类似 ShowMe^[97] 和 CompassQL^[18] 的可视化规则进行可视化的枚举和排序。NL4DV 使用 Stanford CoreNLP^[114] 提供的自然语言解析和标注接口进行用户分析意图的推理，为用户提供了一个端到端的基于自然语言查询表达的用户分析意图的可视化结果推荐。但是，NL4DV 的基于一系列规则的分析意图理解方法存在可扩展性、鲁棒性和适应性不强等缺点。

近些年来，随着深度学习技术在自然语言处理领域的广泛应用，出现了许多先进的基于深度学习的自然语言处理模型，这些模型在诸如机器翻译任务上表现出媲美人类的能力^[115,116]。基于上述讨论，研究者们开始探讨如何使用深度学习技术完成 NL2VIS 任务。为了能够训练和评测基于深度学习的 NL2VIS 模型，

SEQ2VIS^[6] 提出了首个 NL2VIS 数据可视化基准数据集 nvBench，该数据集包含了 153 个数据库、780 个数据表和 25,750 个自然语言查询和可视化结果样本对 ((NL,VIS)Pairs) 并覆盖了 105 个领域（如医疗和体育等）。基于所提出的 NL2VIS 基准数据集 nvBench，SEQ2VIS 提出了一个基于序列到序列模型的 NL2VIS 模型，模型的输入是自然语言查询序列，模型的输出是数据可视化查询语言。通过该模型，SEQ2VIS 提供给用户一个端到端的基于深度学习技术的融合用户分析意图的可视化推荐系统。

除了通过用户提供的自然语言查询来理解用户的分析意图，ncNet^[7] 还支持用户额外指定数据可视化中常见的可视化模板（如 Excel 中提供的柱状图模板）。因此，ncNet 需要同时对用户提供的自然语言查询和可视化模板进行处理和理解，以推荐最优的可视化结果给用户。为了将可视化模板融合到 NL2VIS 的翻译过程，ncNet 设计了一个序列化的可视化语言 Vega-Zero，Vega-Zero 可以理解为是序列化之后的 Vega-Lite。因此，ncNet 可以将可视化模板使用 Vega-Zero 进行序列化表示，并将该序列拼接到自然语言查询之后，作为模型的输入。ncNet 提出了基于 Transformer^[115] 的序列到序列的 NL2VIS 模型，其输出是一个序列化的 Vega-Zero。Sevi^[96] 则是在 ncNet 的基础上，开发了一个语音转文字（Speech-to-Text）的交互层，可以支持用户通过语音的方式指定用户的分析意图，大大地方便了在移动设备等键盘输入不方便的交互设备上进行问答式的数据可视化和可视分析。

上述基于自然语言查询的数据可视化推荐工作的核心是理解用户的自然语言查询意图，并将其翻译成对应的可视化查询语句。这类端到端的“翻译”方法因自然语言的歧义性原因，在准确性方面仍然有很大的提升空间。基于这考量，RGVisNet^[107] 提出了利用现有的可视化查询模板库，在此基础上提出了检索-生成（Retrieval-Generation）混合模型，实现基于用户自然语言查询的数据可视化结果推荐。给定用户自然语言查询，RGVisNet 首先从可视化查询模板库中检索出与之最相关的模板；然后，RGVisNet 基于该查询模板，根据当前的自然语言查询进行部分修正并输出最终的可视化查询结果。

4.5 基于参考对象的数据可视化推荐

基于参考对象的数据可视化推荐是指系统根据用户指定的参考对象（如给定的参考可视化或者数据转换方式）推荐在某个维度上与之相似或者不相似的其他可视化结果。图 16 总结了近年来基于参考对象的数据可视化推荐的 5 个代表性工作。

SeeDB^[87] 通过计算候选可视化与用户给定的参考可视化之间的差异来推荐可视化结果，SeeDB 认为差异越大的可视化结果能蕴含更多的数据洞察，因此差异较大的候选可视化会优先推荐给用户。与 SeeDB 相反，Zenvisage^[23,88]主要是推荐与用户指定的参考可视化相近的结果。为了便于用户指定给定数据集的参考可视化，Zenvisage 提出了一种类似 SQL 语言的可视化语言 ZQL。用户可以通过 ZQL 指定用于可视化的 X/Y 数据属性和数据转换方式（如分桶聚集等），Zenvisage 通过基于用户输入的 ZQL 在可视化空间中枚举候选可视化结果，并返回给用户相似的可视化结果。

VisPilot^[100] 则专注于交互式数据分析的下钻（Drill-down）操作过程中的数据可视化推荐。VisPilot 发现用户在交互式数据分析的下钻操作过程中，通过添加单个过滤条件以获得新的可视化结果，分析师可能会错误地将可视化结果的变化归因于“局部差异”，而忽略了导致这一可视化结果变化的根本原因。为了解决这些问题，VisPilot 自动选择一小组信息丰富且有意义的可视化来传达数据集中的关键分布。VisPilot 的一个重点是通过考虑数据统计和两个可视化之间的向下钻取关系来避免向下钻取的谬误。

在连续性的可视化推荐场景，现有的可视化推荐系统在推荐新的可视化结果时，没有考虑之前推荐的可视化结果（锚点可视化结果）的特性，因此可能会出现当前推荐的可视化结果与历史上的推荐结果具有不同的可视化映射和数据转换方式的情况，加大用户对当前推荐结果的理解难度。为了解决上述挑战，Dziban^[112] 考虑了在推荐当前可视化结果时锚点可视化结果的特性。概括而言，Dziban 在用户给定的数据集和部分数据可视化查询的约束下，推荐在感知上与锚点可视化结果相似的其他可视化。Dziban 使用 Draco^[89] 可视化知识库自动完成部分可视化查询的自动补全，并通过相似性算法推荐与提供的锚点可视化相似的结果。ViSER^[101] 提出按例可视化（Visualization by Example）的思想进行可视化结果的推荐。ViSER 根据用

户提供的数据表和部分可视化作为样例输入，通过自动合成可视化查询语言的方式，在数据表上推荐具有相同可视化形式的候选可视化结果。

4.6 考虑用户偏好的数据可视化推荐

为了使推荐的可视化结果更能满足不同用户的偏好，研究者们基于用户与可视化系统的交互记录，通过启发式规则或者训练机器学习模型以实现个性化数据可视化推荐。图 16 对比了近年来考虑用户偏好的数据可视化推荐的 5 个代表性工作：BDVR^[102]、VizRec^[105]、PVisRec^[106]、VisGNN^[109] 和 VisGuide^[110]。

BDVR^[102] 通过监测用户与可视化系统的交互操作（如标记某一可视化）作为依据来推测用户的数据可视化偏好。具体而言，BDVR 首先通过用户在系统上的查询、过滤和标记交互操作，将用户的交互操作映射到 4 个预定义的交互模式（如下钻），并基于一系列启发式规则，基于检测的交互模式进行后续的可视化结果的推荐。VizRec^[105] 首先通过可视化领域知识（启发式规则）对给定数据集枚举候选可视化，然后使用了基于协同过滤（Collaborative Filtering）的个性化数据可视化推荐方法。对于每个数据集，VizRec 构建了一个用户-可视化矩阵，该矩阵维护了不同用户对不同可视化结果的评分。VizRec 的基本思想是通过该矩阵找到与目标用户相似的近邻用户，通过近邻用户的评价对目标用户产生推荐。具体而言，VizRec 通过计算目标用户与矩阵中其他所有用户的相似性，根据相似性排序从大到小依次选择前 K 个最相似的用户作为目标用户的近邻集合。VizRec 认为具有相似喜好的用户对于同一个可视化结果会给出相似的评分。因此，在目标用户的近邻集合生成后，VizRec 根据近邻集合中用户对可视化结果的评分，来预测目标用户对于这些可视化结果的评分，以此作为可视化推荐的依据。VizRec 假设使用单个数据集，并且仅适用于有大量用户对同一数据集有足够的可视化评分的情况，这在现实情况中是很难满足的。PVisRec^[106] 则是对每一个用户都提供一个个性化的数据可视化推荐模型。具体而言，对于若干个用户和有用户可视化结果的数据集，PVisRec 通过构建数据属性、用户和可视化之间关系图的方式进行用户偏好的学习。类似地，VisGNN^[109] 也是基于类似的思想，通过图神经网络学习数据可视化中的用户偏好。

4.7 基于混合策略的数据可视化推荐

前文提到的数据可视化推荐的代表性工作，大多重点考虑从单个维度进行可视化结果的推荐，但在现实场景中，需要权衡多个维度，例如领域知识和数据特征等，从而协同推荐最优的可视化结果给用户。本节将上述这类工作称为基于混合策略的数据可视化推荐。图 16 对比了近年来的 5 个代表性工作：VizDeck^[111]、DeepEye^[5,35,36]、Draco-Learn^[89]、KG4VIS^[113] 和 Lux^[98]。上述部分系统考虑了从领域知识和数据特征的角度协同推荐可视化结果，有些系统还考虑了分析意图、数据特征和交互历史等推荐维度。这些工作的共同特点是考虑融合多种可视化推荐策略；而不同点主要体现在推荐维度和具体的实现方法，下面将详细介绍。

VizDeck^[111] 结合启发式规则和可视化质量模型进行可视化结果的推荐。该可视化质量模型考虑了用于可视化 X/Y 轴的数据列之间的一些统计特性（例如离散系数和熵等）。此外，VizDeck 还为用户提供了一个投票机制，用户可以通过该机制调整可视化结果的排名；并为用户提供了基于关键词的可视化结果搜索功能，用户可以通过关键词检索相关的可视化结果。

DeepEye^[5,35,36] 系统主要解决了如何自动地从“数据驱动”的角度推荐给用户“好的”可视化结果。首先，DeepEye 基于线下收集的一些数据可视化案例作为训练数据，基于机器学习，分别训练了分类模型和排序模型。在系统运行的时候，用户指定需要进行数据可视化分析的数据集，DeepEye 首先枚举大量可能“有意义”的数据可视化结果作为候选集，然后通过预先训练好的分类模型筛选出“有意义”的可视化结果。在排序环节，DeepEye 基于 Learning-to-Rank 排序模型对“有意义”的可视化结果进行排序，最后将排序之后的结果推荐给用户。当给定的数据集属于特定领域或预先训练的排序模型表现不佳的情况下，DeepEye 系统也支持引入领域的规则，进行数据可视化结果的推荐。领域专家可以为系统指定若干专家规则，DeepEye 系统通过将这些专家规则编码成偏序关系（Partial Order），最后基于偏序的有向图模型进行可视化结果的选择和推荐工作。最后，DeepEye 还提出了一个线性模型来融合上述两种方法的推荐结果。

Draco-Learn^[89] 通过对领域知识表示成在一系列逻辑规则的硬约束（Hard Constraints）和软约束（Soft Constraints），并通过可视化推荐训练数据学习排序模型用于可视化结果的推荐。具体而言，Draco-Learn 首先基于 Draco^[89] 将可视化结果用逻辑规则进行表示，并将现有的可视化领域知识表示成一系列硬约束和软约束。因此，Draco-Learn 可以将可视化结果的推荐看成是在可视化空间中的一系列硬约束和软约束的组合优化问题，这些约束的权重可以通过 RankSVM^[117] 从成对的可视化结果数据中进行学习。最后，Draco-Learn 可以系统地枚举符合硬约束的候选可视化结果，并基于软约束的得分去进行最优可视化结果的推荐。相较于前文提到的领域知识指导的数据可视化推荐方法，例如 CompassQL^[18]，Draco-Learn 的约束权重是从可视化推荐训练数据中进行学习的，CompassQL 的权重则是专家预设的。相较于 Draco^[89] 将可视化领域知识表示成一系列基于逻辑规则的软约束和硬约束，KG4VIS^[113] 则将可视化领域知识表示成知识图谱。首先，KG4VIS 使用了 VizML^[90] 提供的可视化语料库，从中抽取了 81 个数据特征，并基于这些数据特征以及可视化语料库提供的每个可视化结果涉及到的数据属性和可视化映射方式构建了知识图谱。该知识图谱由数据特征、数据属性和可视化映射方式三类实体及其之间的关系组成。然后，KG4VIS 使用 TransE^[118] 对知识图谱中的实体和关系进行表示学习。基于此，给定一个新的数据集，KG4VIS 可以从具有语义意义规则的知识图谱中推断出有意义的可视化结果推荐给用户。

近年来，越来越多的用户使用 Python 进行交互式数据分析和其他数据科学任务。在此背景下，面向 PandasDataFrame 表格型数据结构的自动可视化推荐框架—Lux^[98] 应运而生。Lux 为用户提供了三类数据可视化推荐的应用程序接口：基于分析意图、基于 DataFrame 结构和基于用户的操作历史。由于后两者使用的场景有限，本章只介绍 Lux 中基于分析意图的推荐方法。Lux^[98] 的思路与 TaskVis 比较类似，它提供了 3 种用户意图：增强（Enhance）、过滤（Filter）和概括（Generalize）。其中，增强是指 Lux 在当前可视化的基础上增加一个数据属性，以可视化更多数据属性之间的关系；过滤是指在当前可视化的基础上，保持 X/Y 轴属性不变，增加一个额外的过滤属性；概括是指去掉当前可视化的某一个属性，并可选择性地对另一个属性进行分桶/计数等操作，以展示该属性基本的数据特征。

5 高效可视分析

智能数据可视化和可视化分析的研究重点主要是如何从数据集中根据用户的分析意图和领域知识等维度，精准地选择用户感兴趣的数据子集和推荐可视化和可视分析结果。然而，随着数据量的急剧增长，如何对大规模数据进行可视分析和实时交互已经吸引了学术界和工业界的广泛关注。

受限于计算能力可扩展性和显示设备局限性，大规模数据的高效可视分析主要面临两大挑战：交互响应高延迟和分析结果难呈现。交互响应高延迟主要体现在可视分析系统难以在用户可接受的响应时间内（通常是 500 毫秒^[119]）返回能满足用户查询条件的可视化结果。分析结果难呈现主要体现在可视分析系统的显示设备（例如笔记本电脑屏幕）尺寸大小和分辨率是有局限的，如果直接将大规模的数据点直接在屏幕上渲染，一方面，其渲染结果过于稠密，用户难以捕捉到有用的信息；另一方面，其渲染时间较长，难以满足用户实时交互的需求。

基于上述讨论的高效可视分析的两大挑战，本章从硬件和计算框架、数据管理、可视化和人工智能的视角讨论这些技术如何协同工作，以解决由计算能力可扩展性和显示设备局限性导致的高效可视分析难题。



图 19 支持高效可视分析的主要技术概览

图 19 概览了支持高效可视分析的主要技术。蓝色方框所涉及的技术主要是解决计算能力可扩展性导致的分析效率问题；蓝绿色方框所涉及的技术在不同的场景下，分别可以解决计算能力可扩展性和显示设备局限性带来的挑战。

表 2 高效可视分析代表性工作

研究领域	主要技术	代表性论文
数据管理	数据索引	KD-Box ^[120] 、Falcon ^[121] 、FlashView ^[122] 、Nanocubes ^[123] 、Smartcube ^[124] 、Hashedcubes ^[125] 、Kyrix ^[126,127] 、Kyrix-S ^[128] 、RSATree ^[129] 、BigIN4 ^[130] 、IDEA ^[131] 、QDS ^[132]
	数据聚集	imMens ^[133] 、Tabula ^[134] 、Nanocubes ^[123] 、Smartcube ^[124] 、Hashedcubes ^[125] 、Time Lattice ^[135] 、Profiler ^[136] 、DICE ^[137] 、STASH ^[138] 、QDS ^[132] 、Stolte et al. ^[139] 、Falcon ^[121] 、M4 ^[140]
	预取缓存	Falcon ^[121] 、SW ^[141] 、Dosh et al. ^[142] 、ATLAS ^[143] 、ForeCache ^[144] 、Kyrix ^[126,127] 、IDEA ^[131] 、Guo et al. ^[145]
	物化视图	Marviq ^[146] 、Tabula ^[134] 、Kyrix ^[126,127] 、QDS ^[132] 、ForeCache ^[144]
	近似查询处理	Sample+Seek ^[152] 、Pangloss ^[153] 、SampleAction ^[154] 、IncVisAge ^[155] 、Heatflip ^[156] 、IFocus ^[157] 、PFunk-H ^[158]
	列式存储	SeeDB ^[87] 、TDE ^[159] 、BigIN4 ^[130]
可视化	近似可视化	Sample+Seek ^[152] 、Pangloss ^[153] 、IncVisAge ^[155] 、PFunk-H ^[158] 、GeoExpo ^[160] 、Chen et al. ^[161] 、Vizdom ^[162] 、DenseLines ^[163] 、PDD/PDK ^[164] 、IFocus ^[157]
	渐进式可视化	ProgressiveInsights ^[165] 、Heatflip ^[156] 、Drum ^[166] 、IncVisAge ^[155] 、VisReduce ^[167] 、SampleAction ^[154] 、IFocus ^[157] 、imMens ^[133]
人工智能	分析意图预测	GeoExpo ^[160] 、SW ^[141] 、Brown et al. ^[168] 、ForeCache ^[144]
	AI4DB	Maliva ^[169] 、NeuralCubes ^[170]
硬件和计算框架	GPU 加速	MapD ^[171] 、imMens ^[133] 、IDEA ^[131] 、McDonnel et al. ^[172]
	分布式计算	HadoopViz ^[173] 、SHAHED ^[174] 、ATLAS ^[143] 、DICE ^[137] 、VisReduce ^[167]
	内存计算	GeoSparkViz ^[175] 、Tabula ^[134] 、GeoExpo ^[160] 、STASH ^[138]

本章将从数据管理、可视化、人工智能和硬件和计算框架四个维度综述近年来具有代表性的高效可视分析技术，如表 2 所示。除了本章综述的相关技术，Qin 等人^[4] 和 Battle 等人^[176] 从数据管理和可视化的视角也综述了高效可视分析的相关技术。

5.1 基于高效数据管理的高效可视分析

在现有的硬件条件下，可视分析系统可以依赖底层的数据管理系统（数据库系统）进行高效的数据管理以提高数据处理的效率，加速可视化查询。典型的技术包括数据索引、数据聚集、预取与缓存、物化视图、近似查询处理、列式存储等，如表 2 所示。（1）数据索引^[120,122,123,125-128,133]可以帮助数据库系统快速获取用户想要可视化的数据，并对其进行可视计算。树形索引被广泛应用于为空间数据、数据立方体（Data Cube）等构建索引，以快速获取满足可视化查询条件的数据；（2）数据聚集^[121,123-125,133,134]将一组数据的值汇总（比如求平均值、计数等）在一起，为用户提供聚集后的结果。不同层级的聚集结果为用户提供不同层级（比如年、月、日）的数据抽象。索引可以帮助数据库系统快速获取需要可视化的不同组的数据，并计算聚集结果；（3）预取缓存^[121,141-144]基于用户当前或历史查询的可视化预测用户下一步要浏览的数据，并将这些数据提前加载或缓存至内存；（4）物化视图^[126,127,132,134,144,146]将某些高频可视化查询的结果存储在数据库中，如果之后的某些查询涉及这些已经物化的查询，可以直接从数据库中读取查询结果，而无需再次计算。比如，数据立方体中可以存储每个子立方的聚集结果；（5）近似查询处理^[152-158]从较大的原始数据中采样一部分具有代表性的数据，并使用采样数据计算的可视化作为原始数据的近似可视化；（6）列式存储^[87,130,159]可以为以在线分析处理（On-line Analytical Processing, OLAP）为主的可视化查询提供更好地查询性能。接

下来，本节将介绍数据索引和近似查询处理技术的代表性工作，更多基于数据管理的高效可视分析技术请参考 Qin 等人^[4] 和 Battle 等人^[176]。

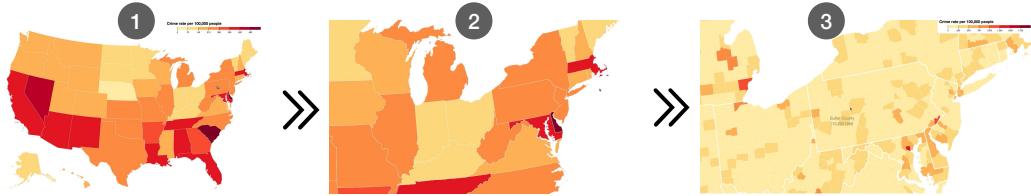


图 20 Kyrix 对可视化结果的交互式放大和缩小操作示例

数据索引。 Kyrix^[126,127] 使用 R 树、B 树或者哈希索引为数据构建索引，以支持在可视化上的高效放大（Zoom in）、缩小（Zoom out）操作。放大、缩小是可视化中一种常见的交互方式，通过放大、缩小操作，用户可以查看不同数据层级（或粒度）聚集的结果：放大允许用户查看可视化结果的更多细节，缩小允许用户查看更高层级的聚集结果。图 20 显示了使用 Kyrix 绘制的美国各地区犯罪率，其中①是原始数据绘制的可视化，②和③是对①进行放大操作后看到的更小区域内的犯罪率。Kyrix 首先执行可视化查询对原始数据进行数据转换操作，然后对于查询结果中的每一个数据点，计算其在画布上的坐标，并为数据点的坐标列构建索引。用户的缩放操作会在画布上选择不同的数据区域，Kyrix 可以根据索引快速查询所选区域的数据，并计算其可视化结果。Kyrix-S^[128] 则是 Kyrix 一个新的扩展版本。Kyrix-S 专门处理大规模散点图（或空间数据），使用 R 树对数据坐标建立索引。KD-Box^[120]、Falcon^[121]、FlashView^[122] 构建树形索引来支持大规模数据的交互式访问；imMens^[133]、Nanocubes^[123]、Hashedcubes^[125] 则构建数据立方体，预计算并缓存不同的数据分片的结果，从而提高数据查询的效率。

近似查询处理。 当要进行可视化的数据量较大时，数据库可能无法快速计算可视化查询的结果。为了在交互时间内返回可视化，一些工作^[152-158]使用近似查询处理（Approximate Query Processing, AQP）技术来高效地计算可视化查询的近似结果。这些工作可以分为三类：(1) 一次性近似查询处理^[152,153]根据用户希望的最大误差或最大运行时间从原始数据中采样，并将采样数据的计算结果作为真实结果的近似，同时还会计算该近似结果的误差范围和置信度。Sample+Seek^[152] 为聚集查询提供近似结果，可以保证近似可视化结果与真实可视化结果中的数据分布误差在一定范围内。Pangloss^[153] 在为用户提供近似可视化结果的同时还提供精确的可视化结果，从而使得用户可以验证近似可视化的质量：Pangloss 首先使用 AQP 技术快速地为用户提供近似可视化，然后，在用户去查看其他可视化时，Pangloss 会在后台继续计算该可视化的精确结果，并将其提供给用户进行校验；(2) 渐进式近似查询处理^[154-156]随着时间的推移不断增加采样数据，从而不断提高近似可视化质量，并将计算得到的不断变化的近似可视化展示给用户（一次性近似查询处理只为用户提供一个最终估计得到的近似可视化结果）。SampleAction^[154] 每秒更新一次展示给用户的近似可视化。随着时间的推移，这些可视化变得越来越接近真实的可视化。IncVisAge^[155] 则解决了 SampleAction 中相邻近似可视化可能发生突变的问题；(3) 可视化感知的近似查询处理^[157,158]根据人类对可视化的感知进行采样。当采样数据达到一定数量后，继续采样只会给近似可视化带来微小的变化，而这种变化是人眼无法分辨的，因此 PFunk-H^[158] 会停止采样。对于柱状图，不同柱子的数据之间的大小关系是一个非常重要的感知信息，IFocus^[157] 可以快速地采样并保证柱子之间的相对关系在一定误差范围内。

5.2 可视化感知的高效可视分析

随着数据量的急剧增大，除了采用高效数据管理技术来提高可视分析系统在数据存储、处理和查询等环节的效率，现有许多系统还结合了用户在可视化和可视分析中的交互习惯以及人类视觉感知特性来提高系统的整体效率。具体而言，这些系统采用了近似可视化的思想并辅以近似查询过程来权衡系统的查询处理效率和可视化结果的精准率。例如，给定一个一亿条记录的数据集，计算图 21 (a) 所示的近似可视化结果可能

只需要 3 秒，而计算图 21 (b) 所示的精准可视化结果则可能需要 10 分钟，然而普通用户却很难通过肉眼观察上述两个可视化的细微差别。基于近似可视化的高效可视分析的关键是要其确保近似可视化结果所传递出的分析结论（如 VLDB 会议论文引用之和是最多的）和精准可视化结果所传递的分析结论是一致的。

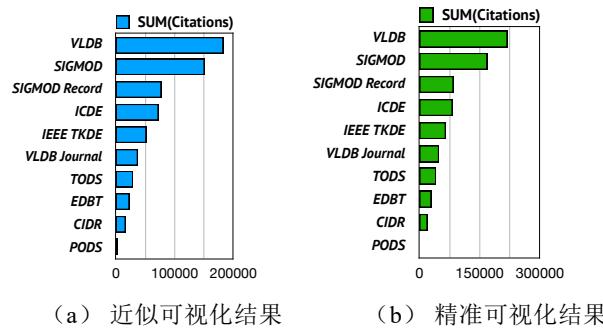


图 21 近似可视化结果示例

近似可视化一方面可以结合近似查询处理等技术，通过少量数据反映总体数据的可视化特性以缓解计算能力可扩展性带来的挑战；另一方面可以基于人类视觉感知特性，进行大规模数据点的视觉感知驱动的数据约简和聚类，以缓解显示设备局限性带来的挑战。

如表 2 所示，常见的基于近似可视化的高效可视分析代表性工作有：Sample+Seek^[152]、Pangloss^[153]、IncVisAge^[155]、PFunk-H^[158]、IFocus^[157]、GeoExpo^[160]、Chen et al.^[161]、Vizdom^[162]、DenseLines^[163] 和 PDD/PDK^[164]。根据上述代表性工作的主要技术特点，本文做以下分类讨论。

(1) 一类工作是基于近似查询处理技术，通过选择和查询样本数据，以在较短的响应时间内返回给用户一个可接受精度的可视化结果。这类的代表性工作有 Sample+Seek^[152]、Pangloss^[153]、PFunk-H^[158]、IFocus^[157]、GeoExpo^[160] 和 Chen et al.^[161]。例如，Sample+Seek^[152] 是面向单表的近似查询处理技术，提出了度量有偏的采样，支持逐步增加样本规模以满足用户设置的置信度区间或者响应时间上限。Pangloss^[153] 则是基于 Sample+Seek 的近似查询处理技术的近似可视化系统，该系统支持基于 Sample+Seek 在用户要求的响应时间内计算出近似可视化结果以满足用户的即时可视分析任务，与此同时，其后端会持续计算精准的可视化结果。值得注意的是，Sample+Seek^[152]、Pangloss^[153] 和 IFocus^[157] 还支持提供查询结果的置信度区间，以辅助用户更好地对近似可视化结果进行判断。概括而言，上述方法是通过对可视化的数据查询过程进行近似查询，以支持海量数据的高效可视分析。

(2) 除了近似查询数据，另一类工作是结合人类的视觉感知特性，对用于渲染可视化的数据进行进一步的约简处理，实现近似可视化，以缓解显示设备局限性引起的挑战。例如，使用折线图可视化大规模线段时，线段彼此之间会有重合，使得用户很难捕捉到有用的信息；为此，DenseLines^[163] 提出了基于密度的算法，针对大规模的线段，高效地计算其在 X 轴和 Y 轴的密度范围并保留相应的极值，依次约简用于渲染的数据并保留了每个线段的最主要的特征，使得用户可以高效地捕捉到折线图的趋势和每个线段的主要特征。PDD/PDK^[164] 则提出了基于感知的线性降维技术，以最大程度保持数据约简后渲染结果在视觉感知上与原始渲染结果尽可能一致和有效。

渐进式可视化一方面可以与数据聚合、数据索引和近似查询处理技术协同优化可视化的数据管理；另一方面也可以通过渐进式分析不同层次的数据，以减少单次用于渲染的数据点。如表 2 所示，常见的基于渐进式可视化的高效可视分析代表性工作有：ProgressiveInsights^[165]、Heatflip^[156]、Drum^[166]、IncVisAge^[155]、IFocus^[157]、SampleAction^[154]、VisReduce^[167] 和 imMens^[133]。

与近似可视化相比，渐进式可视化有以下两点显著特征。

(1) 渐进式可视化强调可通过增量查询数据以逐步提高可视化结果的精准程度。这类的代表性工作有 ProgressiveInsights^[165]、Drum^[166]、IncVisAge^[155]、SampleAction^[154]、Heatflip^[156] 和 VisReduce

^[167]。ProgressiveInsights^[165]提出了一个渐进式可视分析的工作流，支持用户首先在局部可视化结果上进行交互并提供反馈，以引导算法更好地计算用户感兴趣的数据，从而减小计算开销。此外，Drum^[166]提出了将单个查询分解成等价的多个查询（即这些查询分别查询部分数据），这些等价的多个查询按“节奏”依次返回查询结果，以渐进式可视化数据，为用户提供了良好的交互体验。

除了在交互模式上的创新，更多的工作关注数据采样技术和索引设计。例如，IncVisAge^[155]和SampleAction^[154]的核心思想都是通过实时采样技术来逐步提高可视化的精度。Heatflip^[156]则是提出了支持渐进式可视化热力图数据的中间件技术，其核心思想是通过自适应索引技术来高效管理样本数据以实现热力图的渐进式可视化。VisReduce^[167]则是基于MapReduce计算框架实现渐进式可视化。

(2) 渐进式可视化可以通过可视化设计来渐进式地分析不同层次的数据。例如，imMens^[133]提出使用数据立方体和分桶操作来支持高效可视化大规模数据。imMens^[133]可以通过分桶大小来支持用户逐层次地对可视化进行交互（如放大等操作），缓解了渲染大规模数据点带来的问题；此外，imMens^[133]还提出了通过数据立方体来管理预算的多维数据，以支持用户的实时交互。

5.3 人工智能驱动的高效可视分析

人工智能技术可以用于提高数据管理效率和人机协作效率以支持高效可视分析。

一方面，以智能索引和智能查询重写为代表的智能数据管理(AI4DB)^[177]技术可以用来提高可视分析系统的数据管理效率。例如，智能查询重写可以针对交互式可视分析中用户给定的分析查询进行优化，以提高数据处理的效率，进而提升系统的交互性。如表2所示，代表性的工作有：Maliva^[169]和NeuralCubes^[170]。Maliva^[169]使用机器学习技术来重写可视化查询Q，使得数据库可以在某个时间约束（比如小于500ms）下返回Q的计算结果。Maliva通过为原始查询Q添加提示(Hint)或者进行将Q重写为近似查询来加速Q的计算过程。比如对于一个查询Q“计算推特上2020年11月26日某地区发表的关于新冠肺炎的讨论数”，Maliva可能会为其添加“在时间属性上使用B+树索引”的提示来指示数据库引擎如何生成更高效的物理计划，从而加速Q的执行过程；或者在原数据表采样得到的更小的表上执行Q。Maliva由三部分组成：前端、中间件和数据库。用户在前端的可视化查询请求Q传至中间件，中间件将Q重写为Q'，然后将Q'传至数据库，数据库将Q'的查询结果返回给前端。其中后端的数据库是一个黑盒，只接收查询，并返回结果，具体的查询重写都由中间件完成。中间件将重写Q的任务视为一个马尔科夫决策过程(Markov Decision Process, MDP)^[178]，通过最大化收益(最小化Q执行时间)的方式来训练MDP模型。NeuralCubes^[170]通过训练神经网络来预测可视分析中的聚集查询结果，类似于数据立方体。NeuralCubes的核心思想是通过神经网络来拟合可视分析中用户查询和可视化结果之间的关系，因此当用户给定查询后，可以避开数据库的数据查询操作而是使用神经网络直接输出查询结果。

另一方面，通过预测用户的可视分析行为以进行数据预取和查询预测等，可以降低系统的响应时间。例如，在时空大数据可视分析的场景下，预测用户可能感兴趣的时间段，以提前处理和加载这部分数据子集。此外，使用人工智能技术对用户的分析意图进行预测，以动态调度不用的数据侧面用于可视化，可以缓解显示设备局限性带来的挑战。例如，在时空大数据可视分析的场景下，预测用户感兴趣的分析层次（如省级），以减少县市级等不必要的数据点的渲染。如表2所示，代表性的工作有：GeoExpo^[160]、SW^[141]、Brown et al.^[168]和ForeCache^[144]。

5.4 基于硬件和计算框架加速的高效可视分析

随着数据规模的急剧增长，为可视分析系统在用户可以容忍的时间范围内（例如500毫秒）返回查询处理结果带来了新的挑战。通常而言，对于单机的可视分析系统，可以通过GPU加速和并行计算框架等方式提高数据处理效率。如果单机服务器的硬件资源难以满足大规模数据处理的需求，则可以通过硬件横向扩展（如分布式计算）的方式来提高可视分析系统的数据处理效率。如表2所示，基于硬件和计算框架加速的高效可视分析可以分为GPU加速、分布式和内存计算。

GPU (Graphics Processing Unit, 图形处理器) 是一种多核心、高内存带宽的可用于诸如科学计算和深度学习等数据密集型计算任务的硬件。近年来, 研究人员开始利用 GPU 的大规模并行处理能力来提高可视化分析系统的数据处理效率。如表 2 所示, 基于 GPU 硬件进行并行加速的可视化分析系统有 MapD^[171]、imMens^[133]、IDEA^[131] 和 McDonnel et al.^[172]。以 MapD^[171] 为例, 它利用 CPU/GPU 协同并行计算以提高可视化分析系统的 SQL 查询效率。首先, MapD 是单机部署的, 可以避免了因为分布式部署导致的网络开销问题。对于一个 SQL 查询, MapD 首先将其编译成 CPU/GPU 的机器码, 然后同时在 CPU/GPU 上执行, 同时, MapD 尽可能向量化执行 SQL 查询以同时处理更多的数据。对于 SQL 查询结果, MapD 直接使用 CUDA/OpenGL 的编程接口将查询结果映射到 OpenGL 顶点缓冲区以获得最优的实时后端渲染。

分而治之的分布式计算思想也被研究人员应用到可视化分析系统的海量数据处理中, 尤其是在时空大数据可视化领域, 其数据处理的特性十分符合分而治之的计算思想。如表 2 所示, 基于分布式计算技术的高效可视化分析技术有 HadoopViz^[173]、SHAHED^[174]、ATLAS^[143]、DICE^[137] 和 VisReduce^[167]。上述工作中比较具有代表性的是 HadoopViz^[173], 它是基于 Apache Hadoop MapReduce^[179] 计算框架。此外, 也有研究人员使用基于 Apache Spark^[180] 的大数据内存计算框架提高可视化分析系统的数据处理效率, 如表 2 所示, 代表性的作品有: GeoSparkViz^[175]、Tabula^[134]、GeoExpo^[160] 和 STASH^[138]。

6 智能可视分析接口

可视化分析接口是用户与系统交互的媒介, 可视化系统中常见的交互接口是窗口、图标、菜单和指标 (Windows、Icon、Menu、Pointer, WIMP) 交互接口^[8], 用户需要将自己的可视化意图转化成一系列的逻辑操作与系统进行交互。基于 WIMP 交互接口的系统, 通常需要用户根据可视化分析系统的交互设计规则, 学习系统的特定交互方式 (如编程指令或图形化界面操作方式等), 对用户的专业能力要求较高, 存在可视化分析门槛高和交互模式效率低的挑战^[181-183]。另一方面, 可视分析的结果需要通过交互接口呈现给用户, 传统的方法仅仅将可视分析的碎片化发现直接呈现给用户, 需要用户进一步组织这些碎片化分析结论的内在逻辑和因果关系, 存在可视化分析结果难消费的挑战^[184,185]。

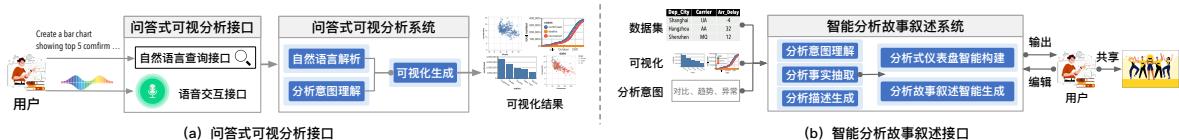


图 22 智能可视分析接口: (a) 问答式可视化分析接口; (b) 智能分析故事叙述接口

本章将会介绍应对上述挑战的问答式可视化分析接口和智能分析故事叙述接口。

(1) **问答式可视化分析接口** (图 22 (a)) 可以接收以自然语言查询或语音查询为载体的用户分析意图作为系统的输入, 系统基于自然语言处理技术进一步处理和分析用户的可视化意图, 生成和推荐符合用户意图的可视化/可视化结果。问答式可视化分析接口可以降低可视化分析系统的用户交互门槛, 提高可视化/可视化分析的生产效率, 打破人与结构化数据间的壁垒^[7,8,96,181,186]。

(2) **智能分析故事叙述接口** (图 22 (b)) 可以根据用户在可视化过程中的用户输入和对单一可视化结果的碎片化发现, 通过分析意图理解、分析事实抽取、分析描述生成等过程, 自动关联碎片化可视化分析结果之间的内在逻辑和外在延展, 最后以智能构建分析式仪表盘或智能生成分析故事叙述的形式帮助用户进行分析结果的故事叙述 (Storytelling), 促进对可视化/可视化分析结果的消费^[184,185]。

表 3 总结了近年来智能可视化分析接口领域的代表性工作, 本文将在第 6.1 节综述问答式可视化分析接口 (自然语言查询和语音交互接口), 在第 6.2 节综述智能分析故事叙述接口 (分析式仪表盘智能构建和分析故事叙述智能生成)。

表 3 智能可视分析接口的代表性工作

研究领域	代表性论文
基于自然语言查询的问答式可视分析	Articulate [187]、DataTone [188]、Eviza [189]、Evizeon [190]、DeepEye [5,35,36]、FlowSense [191]、DataBreeze [192]、SneakPique [193]、Sentifiers [194]、NL4DV [99]、GeoSneakPique [195]、Snowy [196]、QRec-NLI [197]、ADVISor [198]、SEQ2VIS [6]、ncNet [7]
基于语音交互的问答式可视分析	Valletto [199]、Orko [200]、DataBreeze [192]、Data@Hand [201]、Sevi [96]
分析式仪表盘智能构建	VizDeck [111]、TheRecDashboard [202]、Voder [203]、LADV [204]、MultiVision [205]、Dashbot [95]
可视故事叙述智能生成	Supporting [206]、iStoryline [207]、DataShot [208]、StoryAnalyzer [209]、Calliope [210]、ChartStory [211]

6.1 问答式可视分析接口

传统的可视分析系统通常需要用户学习系统的交互方式。在此基础上，用户将自己的可视分析意图转化成一系列与系统交互方式相关的逻辑操作。例如，在图 18 (b) 的 Voyager2 可视分析系统中，如果用户想可视化 Cars 数据集的 Origin 属性列的数据分布情况，则需要选中 Data 面板的 Origin 数据按钮，然后将其拖到 Encoding 面板中的 X 数据框，并在 Mark 面板中做相应的操作。上述过程不但需要用户了解可视分析系统的交互设计理念和规则，而且交互过程繁琐，是典型的“人适应系统”的交互模式。

为了解决上述挑战，问答式可视分析接口应运而生。问答式可视分析接口可以接收以自然语言查询或者语音查询为载体的用户分析意图作为输入，系统基于智能算法进行用户分析意图的理解，并作出符合预期的响应。由于基于自然语言查询或语音查询的交互方式广泛存在，如通过搜索引擎检索与用户自然语言查询相关的网页，这类交互方式已被用户熟练掌握。因此，问答式可视分析接口大大降低了用户与系统的交互门槛，优化可视分析的人机协作模式，是“系统适应人”的交互模式。

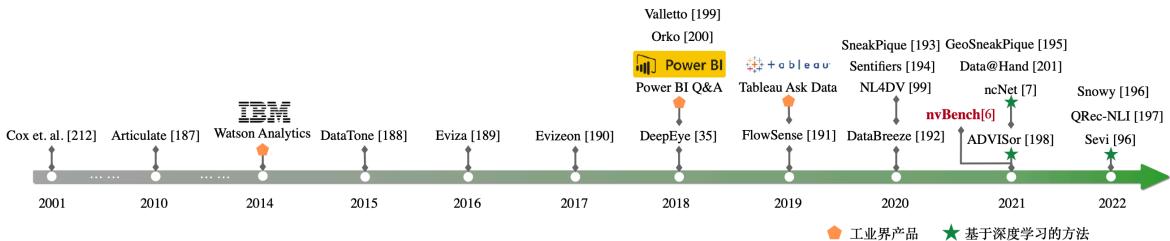


图 23 问答式可视分析接口的发展历程

如图 23 所示，问答式可视分析的研究可以追溯到 2001 年 Cox 等人^[212] 提出的基于简单语法规则解析的基于自然语言查询的数据可视化方法。近年来，工业界和学术界都越来越关注如何将自然语言查询接口与数据可视化和可视分析相结合，从而促进非专业用户对数据的探索，推动数据可视分析的全民化。

本节将综述自然语言查询接口（第 6.1.1 节）和语音交互接口（第 6.1.2 节）两种形式的问答式可视分析接口。

6.1.1 基于自然语言查询的问答式可视分析

表 4 分析总结了近年来基于自然语言查询的问答式可视分析的代表性工作。本文按照这些工作的主要实现方法，将其分为基于启发式算法和基于深度学习两类，并梳理和总结了各项工作的核心方法、系统的输入和输出，以及对比了各类方法的优缺点。

表 4 基于自然语言查询的问答式可视分析的代表性工作

类型	系统	年份	方法	输入	输出	优点	不足
基于启发式算法	Articulate ^[187]	2010	基于自然语言分析器和图生成算法	数据集和自然语言查询	对应的可视化	支持模糊的自然语言查询到合适可视化的转换	仅支持预定义规则以内的分析任务
	DataTone ^[188]	2015	通过自然语言分析器将自然语言转换为数据和可视化特征,然后基于模板生成对应的可视化	数据集和自然语言查询(支持语音查询)	对应的可视化	用户可通过系统提供的交互组件对自然语言查询中的歧义进行修正	不支持多表分析和可视化结果的分面搜索
	Eviza ^[189]	2016	基于概率语法和有限状态机	数据集和自然语言查询	对应的可视化	与现有的知识库相结合增强了语义分析能力	不支持部分具有复杂语法结构的自然语言查询
	Evizeon ^[190]	2017	该系统在 Eviza 的基础上额外加入了语用学原则	数据集和自然语言查询	对应的可视化和解决歧义的交互组件	可识别同义词和具有复杂语法结构的自然语言查询	解析自然语言查询的能力受制于人工制定的语法规则
	DeepEye ^[5,35,36]	2018	在数据特征和领域知识的协同下通过 LambdaMART 算法推荐出可视化结果	数据集和自然语言查询	对应的可视化	数据特征与领域知识相结合所推荐出的可视化效果好且易于理解	制定专家规则的代价较高
	FlowSense ^[191]	2019	通过语义分析器解析自然语言查询然后将其分类成对应的可视化分析任务	自然语言查询	对应的可视化	为用户提供完整的交互式数据探索体验而不只是回答某个特定查询或者展示单一可视化结果	仅适用于数据流图的编辑,不支持回答分析问题和复杂的数据转换
	DataBreeze ^[192]	2020	基于统计的语音识别模型和规则实现多模态输入的识别	数据集、自然语言查询和用户实时交互数据	对应的可视化	将多模态交互和单元可视化相结合,给用户带来新颖的可视化创作体验	不支持基于现有可视化结果进行后续的分析操作
	SneakPique ^[193]	2020	基于自然语言分析器和预定义规则	自然语言查询	对应的可视化	在用户输入自然语言查询语句时,以交互组件的形式为用户提供直观的数据预览功能	数据预览没有针对用户交互记录和数据的语义进行优化
	Sentifiers ^[194]	2020	基于词共现和情感分析将自然语言查询中模糊的修饰词与数据特征相关联	数据集和自然语言查询	对应的可视化	通过分析语义,为用户提供了与自然语言查询相关的数据字段和数据范围过滤器	不支持含有专业术语或多个模糊修饰词的自然语言查询
	NL4DV ^[99]	2020	基于自然语言分析器和预定义规则	数据集和自然语言查询	对应的可视化	NL4DV 是一个与接口无关的工具包,它可以将自然语言查询中分析出的属性、任务和可视化形式化为 JSON 对象供开发人员使用	不支持复杂的数据转换和针对可视化结果进行后续的筛选操作
基于深度学习	GeoSneakPique ^[195]	2021	基于自然语言分析器和预定义规则	自然语言查询和用户实时交互数据	对应的可视化	用户可通过直观的地图交互组件了解指定区域的数据情况	不支持领域知识和个性化推荐
	Snowy ^[196]	2021	基于数据特征和语言学给用户推荐可视化查询建议	数据集和自然语言查询	对应的可视化和后续的自然语言查询建议	系统会基于数据特征自动推荐出一系列查询语句供用户参考,并为用户后续输入的查询提供建议	没有考虑数据的语义,且基于规则的方法只能识别一部分已知的用户意图

			基于日志的推荐模型，该模型通过探索相似主题下其他用户的查询操作，向当前用户推荐对应数据集的自然语言查询	数据集和自然语言查询	对应的可视化	系统会自动推荐具有洞察力的自然语言查询供用户选择，并协助用户进行下一步数据探索	目前仅支持数据选择、分组与聚合三种基本数据分析操作，不支持算术运算和数据过滤
基于深度学习	QRec-NLI ^[197]	2022					
	ADVISOR ^[198]	2021	使用基于 BERT ^[116] 的聚合网络和数据网络生成查询语句，然后使用基于规则的方法生成可视化结果	数据集的表头和自然语言查询	带注释的可视化	通过深度学习为用户提供带注释的可视化，使用户能更好地理解可视化结果	不支持复杂的数据转换（例如排序、分组和过滤操作）
	SEQ2VIS ^[6]	2021	基于 LSTM 的序列到序列模型	数据集和自然语言查询	对应的可视化	在真实数据集上训练，能较好分析用户意图然后推荐出合适的可视化	可视化推荐缺乏可解释性
	ncNet ^[7]	2021	基于 Transformer 的序列到序列模型	数据集、自然语言查询和图表模板（可选）	对应的可视化	提出了面向序列模型的可视化语言 Vega-Zero 并实现了端到端的 NL2VIS	不支持问答式可视化分析

基于启发式算法的系统。基于启发式规则的系统，主要是使用 Stanford CoreNLP^[114] 对用户输入的自然语言查询进行分词、词性标注、依存句法分析等，并基于上述结果，通过启发式规则将用户的自然语言查询映射到一系列预定义的可视分析任务和可视化元素上，最后返回可视化结果给用户。基于该方法的代表性工作有：Articulate^[187]、DataTone^[188]、Eviza^[189]、Evizeon^[190]、DeepEye^[5,35,36]、FlowSense^[191]、DataBreeze^[192]、SneakPique^[193]、Sentifiers^[194]、NL4DV^[99]、GeoSneakPique^[195]、Snowy^[196]、QRec-NLI^[197] 和 Tableau Ask Data^[25]。

以工业界的系统为例，Tableau Ask Data^[25] 允许用户使用自然语言与数据进行交互，当用户输入部分查询语句时，自动补全的内容将呈现在搜索框下方，为用户提供即时查询推荐（如图 24 所示），然后基于用户的自然语言查询推荐相应的可视化结果。Tableau Ask Data 不但可以为用户提供基于自然语言查询的可视化功能，还可以通过查询自动补全的方式帮助用户更好地表达可视分析的意图，使得非专业用户也能够从数据中快速获得他们想要的关键信息，从而提高数据到知识的转换效率。

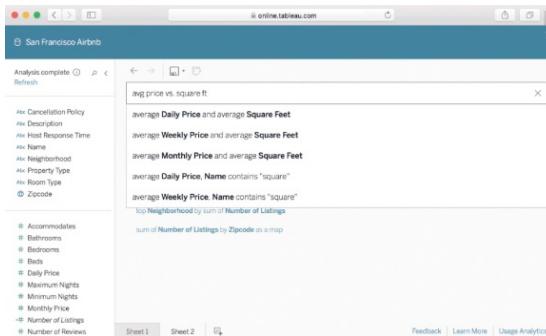


图 24 Tableau Ask Data^[25] 的用户界面

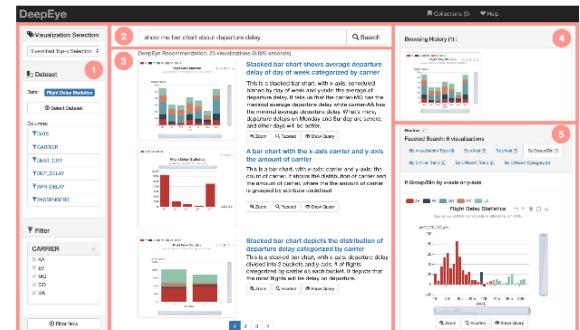


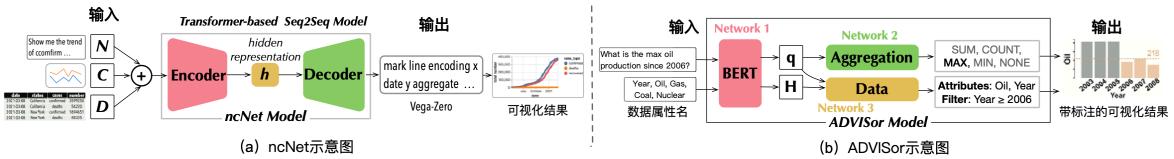
图 25 DeepEye^[5,35,36]的用户界面

学术界的 DeepEye^[5,35,36]也提供了类似的功能，用户只需给定一个数据集和关键字查询语句，DeepEye 就会基于关键字匹配算法推荐出与用户输入信息相关的可视化。当用户选中感兴趣的可视化时，还可以进行分面搜索等交互操作。如图 25 所示，该图为 DeepEye 的界面截图，用户可以在区域①中选择用于可视化/可视分析的数据集，并可以进行数据过滤或选择感兴趣的属性列。当用户选定数据集后，DeepEye 会基于数据集的数据特征和领域知识推荐出有意义的可视化结果，并显示在区域③中。除了系统直接推荐的可视化结

果, 用户还可以通过区域②中的搜索框输入自然语言查询语句获得相应的可视化结果。此外, DeepEye 还支持分面搜索功能。例如, 若用户选择区域③中的某一可视化结果进行分面搜索, 其搜索结果会在区域⑤中显示, 同时浏览记录会在区域④中展示。

由于自然语言具有模糊性, 导致自然语言查询接口难以从带有歧义和意图不明确的自然语言中提取出有效信息。针对此问题 Articulate^[187] 使用自然语言处理技术和机器学习方法将模糊不清的句子转换为明确的表达, 然后应用启发式图生成算法来创建可视化图表。而 DataBreeze^[192] 则是通过引入多模态交互的方式(包括文本、语音和触控), 借助各式交互方式之间可以优势互补的优点, 加深了系统对用户意图的理解程度。在另两项研究工作 Eviza^[189] 和 DataTone^[188] 中, 用户都可以在系统的指导下, 通过解决歧义的交互组件对自然语言查询中模糊的部分进行修正。其中 Eviza^[189] 基于概率语法和有限状态机实现, DataTone^[188] 基于自然语言分析器和可视化模板实现。后续的研究工作 Evizeon^[190] 和 NL4DV^[99] 都对 DataTone^[188] 中解决歧义的交互组件进行了扩展, 提出了更丰富的解决歧义的交互组件形式(例如地图形式)。其中 Evizeon^[190] 是 Eviza^[189] 的扩展, 它额外加入了语用学原则, 使系统支持识别冗长的复合语句、同义词和近义词。NL4DV^[99] 是一个自然语言驱动的数据可视化工具包, 它可以很容易地集成到现有的可视化系统中, 以提供面向可视化的自然语言查询接口。在解决了用户输入的自然语言中的歧义后, 自然语言查询接口可以进一步为用户提供直观的数据预览服务。例如, SneakPique^[193] 相较于 Tableau Ask Data^[25] 和 Sentifiers^[194] 所提供的查询语句补全功能, 引入了一种更直观的数据预览功能, 该功能可以协助用户快速制定或者改良他们输入的自然语言查询语句。在应用层面上, 自然语言查询接口还可以结合到地理位置分析上。例如, GeoSneakPique^[195] 提出了一种利用地图交互组件协助解析自然语言查询的方法, 用户可通过自然语言查询结合操作地图交互组件的方式对感兴趣的地理区域进行数据分析工作。然而, 上述的自然语言查询接口都是一次性提供可视化结果, 未考虑与用户的持续互动, 以通过上下文信息进一步优化可视化结果。最近的研究 Snowy^[196] 和 QRec-NLI^[197] 就对此进行了改进。其中 Snowy^[196] 通过基于数据特征和语言学的方法, 为用户输入的查询提供进一步的自然语言查询建议。类似的, QRec-NLI^[197] 带有渐进式查询推荐模块, 可协助用户进行下一步数据探索。自然语言查询接口除了用于数据探索以外, 还可以辅助用户构建数据流图。例如, FlowSense^[191] 通过使用带有特殊话语标记和占位符的语义分析器来泛化到形式各异的数据集和数据流图, 用户可以通过简单的自然语言轻松地扩展和调整数据流图。

基于深度学习的系统。概括来说, 基于启发式规则实现的系统具有灵活度低、可扩展性差和系统维护难等特点。另一方, 近年来基于深度学习的自然语言处理技术取得了迅猛的发展, 并在许多应用领域取得了令人瞩目的效果, 这也为基于自然语言查询的数据可视化带来了机遇。为了推动深度学习技术驱动的基于自然语言查询的数据可视化(Natural Language to Visualization, NL2VIS)的发展, 除了相应的深度学习技术, 还需要足够多的高质量训练数据。因此, 有研究人员通过人工标注等方式, 构建了若干个可以支持 NL2VIS 任务的语料库, 分别是: VizNet^[213]、Quda^[214] 和 NLV^[215]。其中, VizNet^[213] 提供了一个超 3100 万个数据集的大规模语料库, 它可以用于比较可视化技术的有效性、训练自动可视化模型并为算法提供公共基准。Quda^[214] 包含了 14035 个自然语言查询, 每个查询都带有一个或以上的分析任务描述。它可以帮助面向可视化的自然语言查询接口通过深度学习的模型训练划分任务的类别。但它们并不是针对自然语言到可视化转换而设计的。另外, NLV^[215] 整理了包含 893 个自然语言查询的数据集, 并且基于措辞(例如现实中的用户使用怎样的数据类型)和所含信息(例如图表类型)来描述这些自然语言查询, 因此该数据集可用于评估自然语言到可视化转换的系统性能, 然而 NLV^[215] 的规模太小, 无法满足深度学习模型训练的要求。为了解决此难题, Luo 等人^[6] 提出了一种新颖的 NL2SQL-to-NL2VIS 合成器, 他们从现有的 NL2SQL 基准数据集中合成出第一个 NL2VIS 基准数据集(nvBench), 该基准数据集具有高精度、较全的覆盖范围, 以及多样的可视化类型。SEQ2VIS^[6]、ADVISor^[198] 和 ncNet^[7] 是后来发展出的基于深度学习的 NL2VIS 系统。具体而言, SEQ2VIS^[6] 是基于 LSTM 的序列到序列的模型, 它使用了大型数据集进行模型训练, 能较好地分析用户意图然后推荐出合适的可视化。

图 26 ncNet^[7] 和 ADVISor^[198] 网络结构和学习任务概览

ncNet^[7] 和 ADVISor^[198] 的实现框架如图 26 所示。ncNet^[7] 是一个基于 Transformer 的序列到序列模型，该模型将自然语言查询转换为合适的可视化，使用了几种可视化优化方法以提高效果，包括了使用注意力机制来优化模型训练的过程和使用可视化感知渲染来生成出更好的可视化结果。ncNet 使用了基准数据集 nvBench^[6] 进行训练，并取得了良好的准确率。与 ncNet^[7] 类似的是 ADVISor^[198]，它首先使用 BERT^[116] 进行用户输入的自然语言查询和对应数据表的属性名的表示，然后通过神经网络分别学习对应的聚集函数与数据选择和过滤操作，最后基于启发式规则组合输出最终的可视化结果。虽然同样是基于深度学习的实现，它们间的差异还是存在的，具体为以下几点：第一，对于模型训练而言，ADVISor 使用（自然语言查询，SQL）样本对进行训练，而 ncNet 使用的是（自然语言查询，可视化查询）样本对进行训练，并且输出 Vega-Zero 语句。由于它们的训练集不同，所以是不同的深度学习任务。第二，从输入的角度而言，ADVISor 只允许用户输入自然语言查询，而 ncNet 还可以让用户额外选择一个可视化模板。最后，对于数据转换操作而言，ADVISor 不支持 ncNet 的复杂数据转换操作，例如排序、分组和过滤操作。

尽管研究人员已经开始对自然语言转化为可视化图表进行研究，但目前研究的智能化程度还有待提高。上述介绍的自然语言转化为可视化图表的研究工作，基本都是通过提前定义好的规则进行解析、基于自然语言处理中的语法分析技术等方法将得到的自然语言解析成计算机能读懂的编码。这样的实现不仅准确率低，而且会使得用户输入内容的灵活性受到限制。虽然近年来也出现了基于学习的方法，但由于缺乏类似 VizNet^[213] 和 nvBench^[6] 这种可以用于深度学习模型训练的大型公共基准数据集，对于推进基于学习的研究工作也受到了限制。并且进一步的改良工作，还需要聚焦在如何提升用户体验上，在用户不知道如何开始查询或者下一步怎么探索时给予更多的推荐以及需要明确何时、如何提出推荐。

6.1.2 基于语音交互的问答式可视分析

相较于以自然语言查询为媒介的问答式可视分析，本节介绍的基于语音交互的问答式可视分析会给用户带来更为自然和便捷的用户交互体验，尤其是在移动端设备。作为一种新型的人机交互方式，基于语音交互的问答式可视化分析存在着两大挑战。第一，如何使得机器算法能正确且高效地理解语音传达的真实意图，尤其是在语音输入存在口音偏差等情况下。第二，如何使得机器算法正确地将用户的分析意图转换为目标可视化查询语言或者可视分析操作。表 5 梳理了应对上述挑战的代表性工作，并总结了各项工作核心方法、系统的输入和输出，以及对比了各类方法的优缺点。

在可视分析系统中，研究人员通常将语音交互接口与其他交互接口协同使用，以此为用户带来更好的交互体验。多模态交互可以通过一种模态的优点来补充其他模态的不足，从而保证了数据探索的流畅性。例如，Valletto^[199] 可以通过与数据“聊天”的形式进行数据可视化分析，它允许用户在传统的图形用户界面上使用基于语音的交谈界面和多点触控的手势指定和生成可视化图表。Orko^[200] 是一款网络数据可视化系统，它结合了基于触控和自然语言的交互方式，为用户提供了新颖的网络数据探索体验。DataBreeze^[192] 提出了一种新颖的多模态交互（文本、语音和触控）与单元可视化相结合的数据探索方式。它通过协同使用多种交互模式，以实现不同交互方式的优势互补，从而达到提升用户体验的目的。与 Orko^[200] 相比，DataBreeze^[192] 在融合引擎的支持下，还允许用户同时执行语音和触控操作。

表 5 基于语音交互的问答式可视分析的代表性工作

系统	年份	方法	输入	输出	优点	不足
Valletto ^[199]	2018	基于规则实现多模态输入到用户意图的映射	数据集、语音查询和用户交互记录	对应的可视化	多模态交互有利于更好地理解用户的真实意图 同时支持语音查询和基于触控的直接操作方式	没有考虑数据的语义并且语音转文本的过程中存在误差
Orko ^[200]	2018	通过基于语法和词典的方法解析语音查询	数据集、语音查询和用户实时交互数据	对应的可视化	仅支持与单个可视化进行交互且可视化类型单一	
DataBreeze ^[192]	2020	基于统计的语音识别模型和规则实现多模态输入的识别	数据集、自然语言查询和用户实时交互记录	对应的可视化	将多模态交互和单元可视化相结合, 给用户带来新颖的可视化创作体验	不支持基于现有可视化结果进行后续的分析操作
Data@Hand ^[201]	2021	利用基于语法分析器的语音识别框架把语音转换为文本	数据集、语音查询和用户实时交互数据	对应的可视化	首个通过语音和触控协同交互进行数据探索的移动应用	语音的识别和解释没有结合用户交互记录进行分析
Sevi ^[96]	2022	基于 ncNet 模型	数据集、语音查询和可选的图表模板	对应的可视化	支持普通用户通过语音查询创建可视化结果	需要先将语音转为文本, 再翻译成对应的可视化查询, 期中存在一定的误差

随着智能手机的普及, 在移动端也出现了类似的多模态交互式的数据分析软件, 例如 Data@Hand^[201], 它是一款移动应用程序, 用户可以利用语音和触控相协同的方式, 方便且快速地对个人数据进行可视化探索。近年来, 深度学习的蓬勃发展也给研究人员带来了启发, Sevi^[96] 是一个基于 ncNet^[7] 并且通过 nvBench^[6] 进行模型训练的端到端的可视化系统。该系统通过融合自然语言、语音交互和深度学习技术, 更好地指导了非专业用户通过语音的方式创建可视化。为了进一步增强探索性, 给用户带来更好的交互体验, Srinivasan 等人^[216] 还提出了一种在多模态交互界面中决定何时推荐以及推荐哪些自然语言命令示例的方法。

文本和语音输入都是自然语言查询常用的方式, 然而在语音交互的过程中, 由于语音到文本识别失败导致的其他类型的错误和歧义是需要特别注意的, 这会影响可视化分析的效果。此外, 上述实现仅仅是简单地对语音进行转换, 而没有对语音进行深入的分析, 例如可以通过语调、口音, 语言文化等方式进一步洞察用户的情绪和兴趣。最后, 本文还发现一部分可视化问答是通过扩展现有的问答引擎来回答可视化相关问题的, 但由于它们不是专门为可视化而开发的, 所以解析能力受到了限制。可替代的方法是使用基于可视化数据集训练出的模型。当然基于数据集开发的模型也有局限性, 它可以处理问题的范围被数据集的多样性所限制。因此, 逐步扩展数据集和尝试模型转移是进一步值得研究的事情。未来的问答式可视分析接口可以结合更加先进的自然语言处理技术以及可视分析的领域知识, 支持多领域、多任务、多模态的对话式问答式可视分析, 进一步推动可视化和可视分析的全民化。

6.2 智能分析故事叙述接口

如图 22 (b) 所示, 智能分析故事叙述系统通过对用户提供的数据集、可视化和分析意图等信息, 进行意图理解、事实抽取和描述生成等操作, 从而将碎片化的可视分析结果组织成条理清晰、通俗易懂的分析式仪表盘或者故事叙述。用户可以基于此进行分析结果的分享或进一步的交互式数据分析, 这种交互模式加速了数据可视化和可视分析结果的消费。

在大多数情况下, 智能可视分析系统把数据集作为输入并输出对应的可视化。随着数据可视化的技术发展, 如今输出的可视化已经不局限于传统图表, 还涉及一些更丰富的表达形式 (例如网络图、时间线和信息图)。叙事可视化已经发展为可视化的重要分支之一, 然而对于普通用户而言, 将数据转换为有意义的叙述可视化是十分耗时的。因为用户不仅要探索数据以获取感兴趣的信息, 还需要具备可视化的专业知识以合适

的可视化类型呈现信息。因此研究人员提出了自动生成叙事可视化的方法，本节将围绕智能分析故事叙述接口的两个重要领域进行讨论，包括了第 6.2.1 节分析式仪表盘智能构建和第 6.2.2 节分析故事叙述智能生成。



图 27 分析式仪表盘和可视分析故事叙述的主要生成过程

6.2.1 分析式仪表盘智能构建

分析式仪表盘由一组可视化和若干交互组件构成，可以简明地呈现出数据洞察，并通过交互组件以支持用户的交互式可视分析。其优点在于把重要信息都汇聚在一个界面中，使用户可以快速找到有价值的信息，从而提升提高数据分析的效率^[217]。不过，生成一个令用户满意度较高的分析式仪表盘并非易事，正如图 27 (a) 所示，只有在理解用户偏好和数据潜在特征的基础上，才能生成出合适的可视化图表和文本描述信息，从而构建出分析式仪表盘。受机器学习巨大成功的启发，研究人员将机器学习技术应用到分析式仪表盘的自动生成上，使得生成步骤更为智能，以实现更好的可视化效果。目前具有代表性的研究成果如表 6 所示。

表 6 分析式仪表盘智能构建的代表性工作

系统	年份	方法	输入	输出	优点	不足
VizDeck ^[111]	2012	基于数据特征和用户历史行为推荐可视化结果	数据集	可视分析仪表盘	可通过用户与系统交互的历史行为获得用户对可视化的偏好	仅支持与 SQLShare 交互且可视化类型单一
TheRec Dashboard ^[202]	2015	基于内容并结合用户配置文件的探索性仪表盘推荐系统	数据集和用户配置文件	可视分析仪表盘	系统允许用户将感兴趣的可视化加入到配置文件中，提高了推荐的相关性	用户配置文件需要手动扩展，并且系统仅支持三种基础可视化类型
Voder ^[203]	2017	基于预定义的可视化映射规则和启发式算法	数据集	可视分析仪表盘	系统为用户提供交互式的数据事实，能帮助用户更好地理解可视化结果	推荐可视化时没有考虑数据的语义且不支持个性化推荐
LADV ^[204]	2021	基于 Faster R-CNN 的深度学习模型	仪表盘的图像或者草图	可视分析仪表盘	可实现可视化仪表盘的高效创建、定制和分享	通过默认数据集填充而成的仪表盘并没有考虑具体数据和可视化的特点
MultiVision ^[205]	2021	基于孪生神经网络对单个可视化进行评分，然后通过自定义指标组合出仪表盘	数据集和用户实时交互数据	可视分析仪表盘	将自动推荐集成到多视图可视化和数据探索的过程中	不支持个性化推荐并且深度学习模型缺少评测基准
DashBot ^[95]	2022	基于深度强化学习生成可视分析仪表盘	数据集	可视分析仪表盘	无需大规模的训练集即可推荐出可视化结果	没有考虑可视化组合的有效性和数据的语义信息

VizDeck^[111] 是首个通过训练模型得出可视化推荐结果的基于 Web 的工具，该工具根据数据的统计信息生成可视化结果，然后用户从可视化结果中挑选出感兴趣的可视化，在此交互过程中 VizDeck 会根据用户的行为优化推荐策略，在无需编程的情况下几秒内就能推荐出可视分析仪表盘。它核心的推荐方式为图 27 (a) ③中的基于规则推荐，由于这些规则是人工提前设计的，所以可能会产生出真实用户不感兴趣或质量较低的可视化结果。为解决这一问题，MultiVision^[205] 结合了图 27 (a) ②中的操作日志约束和③中的深度

学习推荐，它利用深度学习模型向用户展示数据列的候选选择，并且当用户与推荐结果交互时，所记录的操作日志将以离线的方式输入到模型当中，从而提高可视化推荐的质量和用户设计分析式仪表盘时的效率。MultiVision^[205] 的操作界面如图 28 所示，用户导入数据集后区域 A 中会展示对应的数据列，点击区域 B 中的 MV Recommender 后区域 D 中会列出当前系统推荐的图表，用户点击感兴趣的图表后区域 E 中会以仪表盘的形式展示用户所感兴趣的图表，并且用户还可以通过区域 C 中的图表编辑器调整图表的参数，区域 F 则为系统设置面板。由于缺乏大规模的多视图可视化数据集，所以 MultiVision^[205] 所训练的深度学习模型只针对单一可视化进行评分操作，其展示的仪表盘是利用单一评分结合自定义规则所间接生成出来的，这就可能降低推荐结果的有效性。

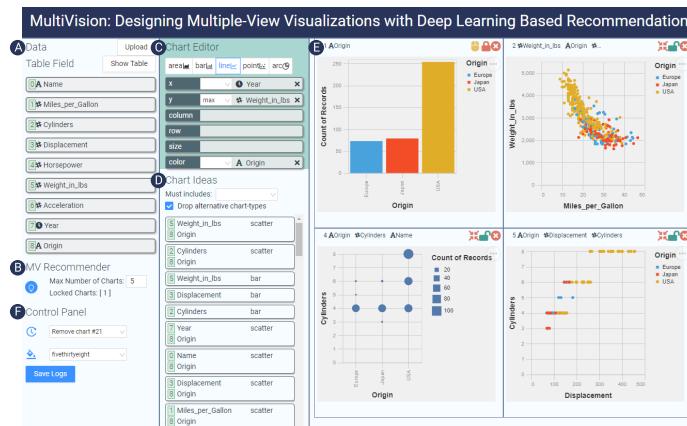


图 28 MultiVision^[205] 用户交互界面截图

受 MultiVision 启发，DashBot^[95] 采用了图 27 (a) ③中的深度强化学习推荐方式，将生成分析式仪表盘的问题建模为具有奖励函数和有限行动空间的马尔科夫决策过程。该系统以已有的可视化知识为基础，构建出训练环境，并设计了一个深度神经网络作为智能体，模仿人类的探索行为。做到了在无需大规模标记数据作为训练集的情况下，即可为用户推荐出合适的可视化结果。而它的不足之处在于没有给可视化加以描述信息，使普通用户难以理解所推荐的可视化。针对此问题，Voder^[203] 加入了如图 27 (a) 所示的步骤④描述生成，它结合了自然语言生成技术和统计函数，从数据集中提取数据事实，并将其转换为描述信息与可视化一起呈现给用户，该系统还允许用户与自动生成的数据事实进行交互以促进数据探索。在交互方式上，LADV^[204] 提供了更灵活的输入方式，它允许用户以仪表盘的图像或草图作为输入，然后利用如图 27 (a) 步骤①中的深度学习提取信息的技术与可视化规则相结合，生成出对应的分析式仪表盘。最后，TheRecDashboard^[202] 还尝试了基于内容的方法来支持图 27 (a) 所示步骤②中的用户自定义约束。它允许用户把自己感兴趣的推荐加入到集合中，然后系统会通过检索文化、科学和教育的多个来源，向用户推荐出具有个性化的仪表盘。

虽然上述的研究工作都颇有成效，但是展示的可视化图表类型较为普通，若能引入新颖的更具有表现力的复合图表和三维图表，将会进一步提高分析式仪表盘的可视化质量和有效性。另外由于深度学习模型缺乏可解释性，通常被视为黑匣子进行运作，用户很难理解推荐出可视化的原因，导致对可视化推荐结果产生质疑。若能把可解释性与用户自定义的规则相结合，将会更容易取得用户的信任，同时又能进一步提高推荐结果的质量。

6.2.2 分析故事叙述智能生成

数据可视化通过使用复杂的算法和人机交互技术对原始数据进行分析。虽然上一节所提及的分析式仪表盘已经为用户提供了便捷的具有探索性的可视化服务。而对于没有相关数据可视化分析经验的用户来说，还

是难以发现数据中蕴含的知识与本质规律。这就需要系统以通俗易懂的方式把分析结果传达给用户。“可视化分析故事叙述”^[218] 的概念结合了发现数据中有意义的信息，使用可视化技术对其进行表示，以及将这些碎片化可视化分析结果智能地排序后叙述给用户，它达到了提高用户对数据理解的效果并且促进了可视化和可视分析的快速消费。

然而，基于分析结果创建一个优秀的可视化分析故事叙述对用户的专业技能要求高，即需要用户同时具备领域知识、数据分析和可视化等能力，存在高门槛的特点。对于普通用户而言，他们期待的是能直接从输入数据集中自动生成高质量可视化分析故事叙述并且支持交互式故事编辑功能的系统。

表 7 总结和对比了近年来分析故事叙述智能生成领域的代表性工作。这些系统的主要工作流程如图 27 (b) 所示，系统首先会从用户输入的数据集中提取故事片段，然后通过排序算法选择出最佳的故事片段，最后使用可视化映射、描述生成和布局排版技术向用户推荐出可视化分析故事叙述。值得注意的是，有部分系统还加入了如图 27 (b) ②中的用户偏好分析，从而为用户提供更具个性化的可视化分析故事叙述。

表 7 分析故事叙述智能生成的代表性工作

系统	年份	方法	输入	输出	优点	不足
Supporting ^[206]	2018	基于数据维度生成故事片段，并根据这些故事片段的特征排列出有意义的故事叙述	数据集	可视化分析故事叙述	定义了合成故事过程中的基本任务，并提出了利用数据维度来组织故事叙述的方法	设计故事叙述的过程中没有考虑用户的个性化需求
iStoryline ^[207]	2018	基于真实用户对故事叙述的创建过程确定了一组设计规则，由此作为约束为用户自动生成故事叙述	数据集	可视化分析故事叙述	提出了创建故事叙述的设计空间并确定了用户创建手绘风格的故事叙述时所遵循的设计流程	需要预先准备好故事数据，并且不允许用户从零开始创建故事叙述
DataShot ^[208]	2019	基于规则的方式从数据集中提取数据事实，然后将其映射到用示例训练的决策树中获取可视化结果，最后组合成故事叙述	数据集	可视化分析故事叙述	DataShot 将数据探索和呈现深度融合使得用户可一键生成初步的故事叙述结果，然后再按需修改	没有考虑数据的语义和依赖关系，容易忽略一部分数据特征
StoryAnalyzer ^[209]	2020	基于自然语言分析器和数据可视化库生成故事叙述	数据集	可视化分析故事叙述	将自然语言处理库和数据可视化库相结合，帮助用户快速生成和理解故事叙述	没有考虑领域知识和数据的语义
Calliope ^[210]	2020	以基于信息论计算得出的信息量为度量指标进行蒙特卡洛树搜索生成故事片段，然后按照逻辑顺序将其组合成故事叙述	数据集	可视化分析故事叙述	通过一种面向逻辑的蒙特卡洛树搜索算法对数据空间进行探索来生成故事叙述并支持故事的编辑操作	没有考虑数据的语义并且信息量的计算耗时较长从而限制了搜索次数降低了故事叙述的质量
ChartStory ^[211]	2021	基于先前研究总结出一组设计要求，以此为基础生成数据驱动的故事叙述	数据集和用户实时交互数据	可视化分析故事叙述	提出了一种支持自动生成故事片段、布局和描述信息的分析方法，且支持交互式故事编辑	处理过量的可视化图表时会产生冗余的故事片段

可视化分析故事叙述使数据更具表现力，这有助于有效地传达有意义的信息。对于数据可视化领域中故事叙述的生成，创建各类图表和对应的文本描述是至关重要的。为了达到这一目的，Supporting^[206] 提出了一个具有指导性意义的框架，它包含了合成故事过程中的基本任务和常用方法。此框架提出后，陆续出现了各种故事叙述生成的方法，其中最常用的是基于模板的方法。例如，Calliope^[210] 是一个可以从数据集中智能

推荐出可视分析故事叙述的系统。类似的, DataShot^[208] 可以直接通过数据集自动生成指定类型的可视分析故事叙述, 例如“可视分析报表”。由于它们生成过程中只经过了图 27 (b) 的①和③两步, 所以生成出的可视分析故事叙述不一定是用户所感兴趣的, 后续还需要提供编辑功能交由用户进一步调整故事叙述的形式。也因如此它们的主要受众为普通用户, 对于希望完全控制分析过程的专业分析师来说是不适用的。而 ChartStory^[211] 就解决了这一识别和表示故事叙述的挑战, 它可以根据数据探索中生成的故事片段作为第一步(如图 27 (b) ①所示), 然后协助分析师在此基础上设计出具有漫画风格的可视分析故事叙述。另外, 想生成出用户感兴趣的故事叙述, 与用户进行交互是必不可少的, 为此 iStoryline^[207] 加入了如图 27 (b) ②中的用户偏好模块。它将用户交互集成到优化算法中, 允许用户修改自动生成的布局, 根据自己的兴趣创建新颖的可视分析故事叙述。同样, StoryAnalyzer^[209] 使用自然语言处理库和可视化库生成相互联系并且可以与用户交互的操作界面, 操作界面上的每个可视化会相互影响, 当鼠标停留在列名时系统会使相关的元素和描述信息都高亮显示。上述介绍的系统主要通过基于模板或规则的方式生成故事叙述, 此类做法缺乏灵活性和可扩展性, 只能应对部分的可视化任务, 难以处理现实中繁杂的数据。如今, 深度学习的进步也带动了数据可视分析的发展, 使用基于学习的方法(例如, 生成图形布局^[219])不需要预先定义规则, 也能得到相应的可视化结果。不过由于训练集不足的原因, 所达到的性能往往比基于规则的方法要差。

7 研究展望与未来趋势

智能数据可视化和可视分析基于数据管理、可视化、人机交互、人工智能等技术, 协同优化可视化和可视分析的人机协作模式, 提高可视分析系统的智能化程度和降低可视分析的门槛, 可推动可视分析的全民化。基于上述愿景, 本章将从面向可视分析的数据准备、智能数据可视化、高效可视分析和智能可视分析接口等多个方面展望智能数据可视分析的发展趋势和研究机会。

7.1 面向可视分析的数据准备

支持隐私保护的多方协同可视分析。在现实场景中, 数据通常被不同的数据提供方所持有, 尽管工业界和学术界都关注如何高效地分析这些被多个数据提供方所持有的数据, 但仍然面临以下的挑战。首先, 现有的可视分析系统通常假设数据源都存储在分析端, 如果直接将多方数据直接汇集以进行可视分析, 一方面, 这可能会存在诸如数据安全、数据版权和法律合规等问题; 另一方面, 在可视分析交互中的多方数据的传输和存储还存在着额外的开销。其次, 多方协同可视分析还会面临着数据版本控制、分析交互协作等方面的挑战。因此, 如何在多方原始数据不出库的情况下, 基于安全多方计算、机密计算、联邦学习和差分隐私等关键技术, 探索新型的多方协作可视分析的交互模式, 实现隐私保护的多方协同可视分析是值得研究的方向。隐私保护的多方协同可视分析的实现, 有利于缓解许多分析场景存在的数据孤岛问题; 有利于促进跨领域多维度的数据增强, 丰富可视分析的维度; 有利于发挥不同数据提供方的数据优势, 提高可视分析的质量和效率。

面向可视分析的自动数据增强。面向可视分析的数据增强可以丰富数据可视化和可视分析的维度。例如, 给定一个航班延误的数据表, 分析师除了对该表分析之外, 还可以关联天气数据以进一步挖掘航班延误和天气之间的关系。给定一个可视分析的基础数据集和外部数据源(如数据湖), 面向可视分析的自动数据增强的核心任务是基于可视分析的任务目标和数据特征等, 自动地从外部数据源中选择合适的数据属性和数据记录, 以丰富可视分析的维度。尽管机器学习任务通常也采用数据增强的手段以丰富训练数据从而提高模型的效能, 但在可视分析的场景中进行数据增强, 还会面临额外的挑战。首先, 扩充基础数据表的属性是需要考虑数据语义, 即不能简单地通过连接和合并操作扩充数据属性, 如何自动选择合适属性扩充则是第一个需要应对的挑战。其次, 在机器学习任务中, 可以通过在数据增强后的数据集上训练模型, 以评估模型性能的提升, 从而自动地评价数据增强的收益, 但在可视分析场景中, 数据增强的收益与分析任务、用户目标和领域知识等高度相关, 如何自动高效地评价收益也是一大挑战。概括而言, 该研究需要解决“什么样的场景需要进行数据增强? ”、“如何自动且高效地进行数据增强? ”、“如何防止数据增强带来的数据冗余问题? ”

等挑战。

7.2 智能数据可视化

多维度自适应的智能数据可视化。一方面，目前的智能数据可视化系统大多着重于从单一维度来实现智能数据可视化，例如依据当前数据的特征推荐合适的可视化结果。然而，当面对不同的数据领域、分析目标和用户群体时，可视化和可视分析的目标和导向也是不同的。另一方面，人工智能在商品等推荐系统中已经发挥了不可替代的作用。因此，如何利用人工智能技术，进一步提高现有智能可视化推荐系统的效能是一个值得研究的方向。智能数据可视化系统需要根据不同的用户群体、分析场景、分析意图、数据领域等，自适应地从多个维度协同推荐最有价值的可视化结果给用户，以帮助用户更快地发现数据中最有价值的信息。

可解释性智能可视分析。现有的可视化和可视化分析系统主要聚焦于如何帮助用户生成有意义的数据可视化结果，使得用户可以通过可视化结果进行分析推理。然而，这些可视化结果往往不具备较好的“解释性”，这主要体现在用户可能仅是通过当前的可视化结果得出一个结论，但不能知道得出该结论的根本原因。因此，可视化和可视分析系统需要对结果提供一定的解释性，以帮助用户更好地做出准确的推理判断，避免得出错误性的推论。概括来说，可解释性可视分析的目标是让用户“知其然且知其所以然”。可解释性可视分析的研究可以结合数据血缘（Data Lineage）和根因分析（Root Cause Analysis）等技术开展。

7.3 高效可视分析

人工智能赋能的高效可视分析。在大数据时代，海量数据的高效可视分析面临着交互响应高延迟的挑战。现有的研究大多针对计算能力可扩展性的约束，从硬件和计算框架、数据管理、人工智能和可视化四个角度进行优化。然而，现有的研究未能很好地考虑可视化和可视分析任务的特性，来协同优化计算能力可扩展性和显示设备局限性两个约束带来的挑战。未来的研究可从人工智能的角度进行切入，基于人工智能技术协同优化可视分析工作流中的数据管理、用户交互行为建模、用户分析意图预测、可视化表示等，以支持海量数据可视分析的实时交互。例如，通过对用户历史交互行为进行分析学习，以预测用户的可视分析查询；也可以利用智能数据管理技术进行可视分析查询的重写和针对分析场景进行精准的数据预取。针对海量数据带来的庞大存储开销，还可以使用深度学习技术进行数据和可视化结果的压缩表达，减小数据的通信代价和显示渲染代价。

7.4 智能可视分析接口

问答式可视分析接口。目前基于自然语言查询和语音交互的问答式可视分析已经有效地降低了可视分析的门槛，优化了可视分析的交互模式，为非专业用户提供了一个友好且智能的可视分析接口。然而，现有的问答式可视分析接口依然存在诸如准确率低、鲁棒性不强和会话领域有限等问题。未来的问答式可视分析接口可以结合更加先进的自然语言处理技术以及可视分析的领域知识，支持多领域、多任务、多模态、多轮会话的问答式可视分析，进一步推动可视化和可视分析的全民化。

智能分析故事叙述生成。可视分析故事叙述的核心是基于挖掘可视分析碎片化结果之间的内在联系，并基于数据特征、领域知识和分析任务进一步延展数据故事的外在表现，以分析式仪表盘和可视分析图文报告的形式帮助用户快速地理解数据并捕捉数据传递的知识和规律。现有的工作大多是基于领域专家的经验，通过故事叙述模板和简单的数据挖掘算法生成分析式仪表盘和可视分析图文报告，这种方法难以应对现实世界中的复杂需求和场景，灵活性和可扩展性较差。未来，智能分析故事叙述需要紧密结合人工智能、可视化、数据挖掘、自然语言处理等技术，通过人工智能技术增强知识发现和逻辑关联的能力，通过可视化技术智能设计和优化故事叙述的模板和风格，通过数据挖掘技术进一步挖掘数据中隐含的模式和规律，并通过自然语言处理技术进行相关描述的生成和优化。

7.5 智能可视分析的评测基准

正如 ImageNet^[220] 的构建加速了深度学习驱动的计算机视觉技术的研究和应用，智能可视分析领域也需

要若干公开的基准数据集以促进机器学习/深度学习技术驱动的智能可视分析技术的研究和落地。目前大多数智能可视分析系统都是基于自行构建的数据集进行算法设计/模型训练，这些数据集大多是基于自有数据或者网络爬虫进行构建，并进行手动标注。然而，这些数据集因为隐私或者版权等问题而没有被公开，从而导致在相同的可视分析任务（如可视化推荐任务）上，不同的推荐算法使用不同的数据集进行评测，彼此之间缺乏公平的比较，不利于领域的研究。尽管目前已有部分智能可视分析的基准数据集，例如 IDEBench^[22] 和 nvBench^[6]，但这些基准数据集依然存在任务单一和数据规模小等缺点。因此，智能可视分析的基准数据集的研究迫在眉睫。针对智能可视分析的基准数据集的研究，首先需要抽象出智能可视分析中若干基准的任务，例如可视化结果推荐、用户交互建模和交互可视分析效能等，并针对这些基准的任务通过专家和众包标注/测评的方式构建跨领域的多用户基准数据集。

7.6 智能可视分析的应用生态

目前大多数的可视分析系统是由研究者根据不同的可视分析任务，采用不同的技术路线和框架搭建的，由此不难发现智能数据可视分析的应用生态存在两大问题：（1）缺少公共的可复用的智能可视分析模块；（2）不同可视分析系统之间协作和兼容难，且难与其他数据科学系统进行无缝地衔接。深度学习/机器学习能在各行各业产生重要影响的其中一个因素是构建了良好的应用生态，例如开发和维护共有的基础计算模型（如 scikit-learn）和构建多方共享共建的机器学习开发框架（如 PyTorch）。因此，学术界和工业界需要共同努力，共同凝练智能可视分析的标准计算和开发框架，共同构建智能可视分析的基础计算和分析模型，促进智能可视分析应用生态的构建和发展。

8 总结

本文从数据管理、可视化和可视分析、人工智能等视角出发，通过对智能数据可视分析领域的代表性工作进行梳理、总结和分析，凝练了智能数据可视分析核心概念和关键技术框架，总结了四个核心研究领域：面向可视分析的数据准备、智能数据可视化、高效可视分析和智能可视分析接口。基于此，本文总结和分析了近年来上述四个研究领域面临的研究挑战及取得的最新进展。智能可视分析是一个新兴的交叉研究领域，在大数据智能时代发挥着重要的作用，目前正处于高速发展的阶段，但仍存在大量的研究问题亟需解决。基于此，本文讨论了该领域存在的开放性问题，并展望了智能数据可视分析的发展趋势。

References:

- [1] China Artificial Intelligence 2.0 Development Strategy Research Group. Strategic Research on Artificial Intelligence 2.0 in China. 1th ed., China: Zhejiang University press, 2018. (in Chinese)
- [2] Chen Wei, Shen Zeqian, Tao Yubo. Data Visualization. 2th ed., China: Publishing House of Electronics Industry, 2013. (in Chinese)
- [3] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. Readings in information visualization - using vision to think. 1th ed., San Francisco: Academic Press, 1999. 579-581.
- [4] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. Making data visualization more efficient and effective: a survey. The VLDB Journal, 2020,29(1):93-117.
- [5] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. Deepeye: Towards automatic data visualization. In 34th IEEE International Conference on Data Engineering. Paris, France: IEEE Computer Society, 2018. 101-112.
- [6] Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. SIGMOD 2021: International Conference on Management of Data. China: ACM, 2021.1235-1247.
- [7] Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. Natural language to visualization by neural machine translation. IEEE Trans. on Visualization and Computer Graphics. 2022,28(1):217-226

- [8] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xionghuai Zhang, Zhiwei Tai, and Jianmin Wang. Towards natural language interfaces for data visualization: A survey. *IEEE TVCG*, 2022. doi: 10.1109/TVCG.2022.3148007.
- [9] Battle, Leilani, and Carlos Scheidegger. "A structured review of data management technology for interactive visualization and analysis." *IEEE Transactions on Visualization and Computer Graphics*, 2020, 27(2): 1128-1138.
- [10] Sujia Zhu, Guodao Sun, Qi Jiang, Meng Zha, and Ronghua Liang. "A survey on automatic infographics and visualization recommendations." *Visual Informatics*, 2020, 4(3): 24-40.
- [11] Wang, Qianwen, Zhutian Chen, Yong Wang, and Huamin Qu. "Applying machine learning advances to data visualization: A survey on ml4vis." *arXiv preprint arXiv:2012.00467* ,2020.
- [12] Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. 2021. AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2021.
- [13] Jiazhi Xia, Jie Li, Siming Chen, Hongxing Qin, and Shixia Liu. A Survey on Interdisciplinary Research of Visualization and Artificial Intelligence. *SCIENTIA SINICA Informationis*. 2021. (in Chinese)
- [14] Open gl. <https://en.wikipedia.org/wiki/OpenGL>. Accessed: 2022-05-01.
- [15] Java 2d. <https://zetcode.com/gfx/java2d/>. Accessed: 2022-05-01.
- [16] Html canvas. https://en.wikipedia.org/wiki/Canvas_element. Accessed: 2022-05-01.
- [17] Arvind Satyanarayanan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. on Visualization and Computer Graphics*.2017, 23(1):341-350
- [18] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock D. Mackinlay, Bill Howe, and Jeffrey Heer. Towards a general-purpose query language for visualization recommendation. *HILDA@SIGMOD* 2016. San Francisco, CA, USA: ACM, 2016. 4.
- [19] Pat Hanrahan. Vizql: a language for query, analysis and visualization. *ACM SIGMOD* 2006.721.
- [20] Altair. <https://altair-viz.github.io/index.html>. Accessed: 2022-05-01.
- [21] Apache echarts. <https://echarts.apache.org/en/index.html>. Accessed: 2022-05-01.
- [22] ggplot2. <https://ggplot2.tidyverse.org/>. Accessed: 2022-05-01.
- [23] Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Xiaofo Yu, Sean Zou, Lijin Guo, Changfeng Liu, Chaoran Wang, Karrie Karahalios, and Aditya G. Parameswaran. Fast-forwarding to desired visualizations with zenvisable. In *CIDR* 2017. Chaminade, CA, USA: CIDR, 2017.
- [24] Arvind Satyanarayanan, Kanit Wongsuphasawat, and Jeffrey Heer. Declarative interaction design for data visualization. *UIST* 2014. 669-678.
- [25] Tableau. <https://www.tableau.com/>. Accessed: 2022-05-01.
- [26] Google sheets. <https://www.google.com/sheets/about/>. Accessed: 2022-05-01.
- [27] Excel. <https://www.microsoft.com/en-us/microsoft-365/excel>. Accessed: 2022-05-01.
- [28] Many eyes. <https://www.ibm.com/support/pages/many-eyes-and-visualization-data>. Accessed: 2022-05-01.
- [29] Data illustrator. <https://data-illustrator.cs.umd.edu/>. Accessed: 2022-05-01.
- [30] Lyra. <http://idl.cs.washington.edu/projects/lyra/>. Accessed: 2022-05-01.
- [31] Arvind Satyanarayanan and Jeffrey Heer. Lyra: An interactive visualization design environment. *Computer Graphics Forum*, 2014.
- [32] Jock D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graphics*, 1986, 5(2):110 -141.
- [33] Quicksight. <https://aws.amazon.com/cn/quicksight/>. Accessed: 2022-05-01.
- [34] Qlik. <https://www.qlik.com/us/>. Accessed: 2022-05-01.
- [35] Yuyu Luo, Xuedi Qin, Nan Tang, Guoliang Li, and Xinran Wang. Deepeye: Creating good data visualizations by keyword search. In: Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, eds. *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD Conference 2018. Houston, TX, USA:ACM,2018.1733-1736.
- [36] Yuyu Luo, Xuedi Qin, Chengliang Chai, Nan Tang, Guoliang Li, and Wenbo Li. Steerable self-driving data visualization. *IEEE Trans. on Knowledge and Data Engineering*. 2022, 34(1):475-490.

- [37] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock D. Mackinlay, Bill Howe, and Jeffrey Heer. Voyager 2: Augmenting visual analysis with partial view specifications. CHI 2017. 2648-2659.
- [38] Haber, Robert B., and David A. McNabb. "Visualization idioms: A conceptual model for scientific visualization systems." Visualization in scientific computing, 1990, 74: 93.
- [39] Pirolli, Peter, and Stuart Card. "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis." In Proceedings of international conference on intelligence analysis, 2005, 5: 2-4.
- [40] Card, Mackinlay. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
- [41] T. Munzner, "A nested model for visualization design and validation," IEEE TVCG, 2009, 15(6): 921-928.
- [42] Lamba, Manika, and Margam Madhusudhan. "Information Visualization." In Text Mining for Information Professionals, 2022.243-293.
- [43] Ren L, Du Y, Ma S, Zhang XL, Dai GZ. Visual analytics towards big data. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):1909-1936 (in Chinese). <http://www.jos.org.cn/1000-9825/4645.htm>
- [44] Daniel A. Keim, Gennady L. Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In: Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, eds. Information Visualization-Human-Centered Issues and Perspectives, volume 4950 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008. 154-175.
- [45] J. J. Van Wijk, "The value of visualization," in Proceedings of the IEEE Visualization Conference 2005:79-86.
- [46] Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=666536e96f63>. Accessed: 2022-05-01.
- [47] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. Relational data synthesis using generative adversarial networks: A design space exploration. VLDB, 2020, 13(11):1962-1975.
- [48] Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, and Jianhua Feng. Cost-effective crowdsourced entity resolution: A partial-order approach. In SIGMOD, 2016, 969-984.
- [49] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. TKDE, 2007, 19(1):1-16.
- [50] Haipeng Chen, Sushil Jajodia, Jing Liu, Noseong Park, Vadim Sokolov, and VS Subrahmanian. Faketables: Using gans to generate functional dependency preserving tables with bounded real data. In IJCAI,2019. 2074-2080.
- [51] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. VLDB, 2018, 11(10):1071-1083.
- [52] Xuedi Qin, Chengliang Chai, Nan Tang, Jian Li, Yuyu Luo, Guoliang Li, and Yaoyu Zhu. Synthesizing privacy preserving entity resolution datasets. In ICDE, 2022. 2359-2371.
- [53] Kyriaki Dimitriadou, Olga Papaemmanoil, and Yanlei Diao. Explore-by-example: An automatic query steering framework for interactive data exploration. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014. 517-528.
- [54] Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, Adel Ardalan, Jeffrey R. Ballard, Han Li, Fatemah Panahi, Haojun Zhang, Jeffrey F. Naughton, Shishir Prasad, Ganesh Krishnan, Rohit Deep, and Vijay Raghavendra. Magellan: Toward building entity matching management systems. PVLDB, 2016, 9(12):1197-1208.
- [55] Xuedi Qin, Chengliang Chai, Yuyu Luo, Tianyu Zhao, Nan Tang, Guoliang Li, Jianhua Feng, Xiang Yu, and Mourad Ouzzani. Interactively discovering and ranking desired tuples by data exploration. The VLDB Journal, 2022. 1-25.
- [56] Xuedi Qin, Chengliang Chai, Yuyu Luo, Nan Tang, and Guoliang Li. Interactively discovering and ranking desired tuples without writing sql queries. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pages, 2020. 2745-2748.
- [57] Xuedi Qin, Chengliang Chai, Yuyu Luo, Tianyu Zhao, Nan Tang, Guoliang Li, Jianhua Feng, Xiang Yu, and Mourad Ouzzani. Ranking desired tuples by database exploration. In 2021 IEEE 37th International Conference on Data Engineering (ICDE). 2021.1973-1978.

- [58] Dylan Cashman, Shenyu Xu, Subhajit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Endert, and Remco Chang. Cava: A visual analytics system for exploratory columnar data augmentation using knowledge graphs. *IEEE TVCG* 2020, 27(2).
- [59] Yi Zhang and Zachary G Ives. Finding related tables in data lakes for interactive data science. *ACM SIGMOD* 2020. 1951-1966.
- [60] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In: Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, eds. New York: NeurIPS, 2019.
- [61] Burr Settles. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, 2009.
- [62] Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *ICML*, 2003. 59-66.
- [63] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *ICML*, 2004. 74.
- [64] Yanyan Shen, Kaushik Chakrabarti, Surajit Chaudhuri, Bolin Ding, and Lev Novik. Discovering queries based on example tuples. In *SIGMOD*, 2014. 493-504.
- [65] Anna Fariha and Alexandra Meliou. Example-driven query intent discovery: Abductive reasoning using semantic similarity. *PVLDB*, 2019,12(11):1262-1275.
- [66] Enhui Huang, Liping Peng, Luciano Di Palma, Ahmed Abdelkaft, Anna Liu, and Yanlei Diao. Optimization for active learning-based interactive database exploration. *PVLDB*, 2018,12(1):71-84.
- [67] Min Xie, Tianwen Chen, and Raymond Chi-Wing Wong. FindYourFavorite: An interactive system for finding the user's favorite tuple in the database. In *SIGMOD*, 2019. 2017-2020.
- [68] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 2017, 42(4):25:1-25:41.
- [69] Bryan Chan, Leslie Wu, Justin Talbot, Mike Cammarano, and Pat Hanrahan. Vispedia: Interactive visual exploration of wikipedia data via search-based integration. *IEEE Transactions on Visualization and Computer Graphics*, 2008, 14(6):1213-1220.
- [70] Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda, and William Ribarsky. Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *IEEE transactions on visualization and computer graphics*, 2015,22(1):210-219.
- [71] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K Elmagarmid, Ihab F Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. The data civilizer system. In *CIDR*, 2017.
- [72] Shuang Hao, Guoliang Li, Jianhua Feng, Ning Wnag. Survey of structured data cleaning methods. *Journal of Tsinghua Univ (Sci& Technol)*. 2018, Vol. 58, No. 12. (in Chinese)
- [73] Morton, Kristi, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. "Support the data enthusiast: Challenges for next-generation data-analysis systems." *Proceedings of the VLDB Endowment*, 2014, 7(6): 453-456.
- [74] Khayyat, Zuhair, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. "Bigdansing: A system for big data cleansing." In *ACM SIGMOD* 2015: 1215-1230.
- [75] Abedjan, Ziawasch, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. "Detecting data errors: Where are we and what needs to be done?" *PVLDB* 2016, 9(12): 993-10049, no. 12 (2016): 993-1004.
- [76] Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive visual specification of data transformation scripts." In *Proceedings of the sigchi conference on human factors in computing systems*, 2011: 3363-3372.
- [77] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. "Visclean: Interactive cleaning for progressive visualization." *Proceedings of the VLDB Endowment*, 2020, 13(12): 2821-2824.
- [78] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. Interactive cleaning for progressive visualization through composite questions. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020. 733-744.
- [79] Wang, Chenglong, Yu Feng, Rastislav Bodik, Isil Dillig, Alvin Cheung, and Amy J. Ko. "Falx: Synthesis-powered visualization authoring." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.1-15.

- [80] Wu, Aoyu, Liwenhan Xie, Bongshin Lee, Yun Wang, Weiwei Cui, and Huamin Qu. "Learning to automate chart layout configurations using crowdsourced paired comparison." In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021. 1-13.
- [81] Qian, Xin, Ryan A. Rossi, Fan Du, Sungchul Kim, Eunyee Koh, Sana Malik, Tak Yeon Lee, and Joel Chan. "Learning to Recommend Visualizations from Data." In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.
- [82] Lee, Doris Jung-Lin, Vidya Setlur, Melanie Tory, Karrie G. Karahalios, and Aditya Parameswaran. "Deconstructing categorization in visualization recommendation: A taxonomy and comparative study." IEEE Transactions on Visualization and Computer Graphics, 2021.
- [83] Vartak, Manasi, Silu Huang, Tarique Siddiqui, Samuel Madden, and Aditya Parameswaran. "Towards visualization recommendation systems." ACM SIGMOD Record, 2017, 45(4): 34-39.
- [84] Zhou, Mengyu, Wang Tao, Ji Pengxin, Han Shi, and Zhang Dongmei. "Table2analysis: Modeling and recommendation of common analysis patterns for multi-dimensional data." In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(01): 320-328.
- [85] Ma, Pingchuan, Rui Ding, Shi Han, and Dongmei Zhang. "MetaInsight: Automatic Discovery of Structured Knowledge for Exploratory Data Analysis." In Proceedings of the 2021 International Conference on Management of Data, 2021. 1262-1274.
- [86] Leixian Shen, Enya Shen, Zhiwei Tai, Yihao Xu, Jiaxiang Dong, and Jianmin Wang. "Visual Data Analysis with Task-based Recommendations." Data Science and Engineering, 2022.
- [87] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya G. Parameswaran, and Neoklis Polyzotis. SEEDB: efficient data-driven visualization recommendations to support visual analytics. Proc. VLDB Endow, 2015, 8(13):2182-2193.
- [88] Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya G. Parameswaran. Effortless data exploration with zenvisage: An expressive and interactive visual analytics system. Proc. VLDB Endow, 2016, 10(4):457-468.
- [89] Dominik Moritz, Chenglong Wang, Greg L. Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. IEEE TVCG, 2019, 25(1):438-448.
- [90] Kevin Zeng Hu, Michiel A. Bakker, Stephen Li, Tim Kraska, and César A. Hidalgo. Vizml: A machine learning approach to visualization recommendation. CHI 2019. Glasgow, Scotland, UK: ACM, 2019.128.
- [91] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quickinsights: Quick and automatic discovery of insights from multi- dimensional data. SIGMOD Conference 2019. Amsterdam, The Netherlands: ACM, 2019. 317-332.
- [92] Leixian Shen, Enya Shen, Zhiwei Tai, Yiran Song, and Jianmin Wang. TaskVis: Task-oriented Visualization Recommendation. In: Marco Agus, Christoph Garth, and Andreas Kerren, eds. EuroVis 2021- Short Papers. Netherlands: The Eurographics Association, 2021.
- [93] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1):649-658.
- [94] Patel, Hima, Shanmukha Guttula, Ruhi Sharma Mittal, Naresh Manwani, Laure Berti-Equille, and Abhijit Manatkar. "Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems." ACM SIGKDD 2022. 4814-4815.
- [95] Deng, Dazhen, Aoyu Wu, Huamin Qu, and Yingcai Wu. "DashBot: Insight-Driven Dashboard Generation Based on Deep Reinforcement Learning." IEEE Transactions on Visualization and Computer Graphics (Proc. of VIS), 2022.
- [96] Jiawei Tang, Yuyu Luo, and Guoliang Li Mourad Ouzzani, and Hongyang. Sevi: Speech-to-visualization through neural machine translation. In SIGMOD '22: International Conference on Management of Data, 2022. 2353-2356.
- [97] Jock D. Mackinlay, Pat Hanrahan, and Chris Stolte. Show me: Automatic presentation for visual analysis. IEEE TVCG, 2007, 13(6).
- [98] Doris Jung Lin Lee, Dixin Tang, Kunal Agarwal, Thyne Boonmark, Caitlyn Chen, Jake Kang, Ujjaini Mukhopadhyay, Jerry Song, Micah Yong, Marti A. Hearst, and Aditya G. Parameswaran. Lux: Always-on visualization recommendations for exploratory dataframe workflows. Proc. VLDB Endow, 2021, 15(3):727-738.
- [99] Arpit Narechania, Arjun Srinivasan, and John T. Stasko. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. IEEE Trans. on Visualization and Computer Graphics, 2021, 27(2):369-379.

- [100] Doris Jung Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya G. Parameswaran. Avoiding drill-down fallacies with VisPilot: assisted exploration of data subsets. IUI 2019. Marina del Ray, CA, USA: ACM, 2019. 186-196.
- [101] Chenglong Wang, Yu Feng, Rastislav Bodík, Alvin Cheung, and Isil Dillig. Visualization by example. POPL 2020:49:1-49:28.
- [102] David Gotz and Zhen Wen. Behavior-driven visualization recommendation. IUI 2009. 315-324.
- [103] Victor Dibia and Çagatay Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. IEEE Computer Graphics and Applications, 2019, 39(5):33-46.
- [104] Mengyu Zhou, Qingtao Li, Xinyi He, Yuejiang Li, Yibo Liu, Wei Ji, Shi Han, Yining Chen, Dixin Jiang, and Dongmei Zhang. Table2charts: Recommending charts by learning shared table representations. ACM SIGKDD 2021. 2389-2399.
- [105] Belgin Mutlu, Eduardo Veas, and Christoph Trattner. Vizrec: Recommending personalized visualizations. ACM TIIS, 2016, 6(4): 1-39.
- [106] Xin Qian, Ryan A. Rossi, Fan Du, Sungchul Kim, Eunyee Koh, Sana Malik, Tak Yeon Lee, and Nesreen K. Ahmed. Personalized visualization recommendation. ACM Transactions on the Web (TWEB), 2021.
- [107] Song, Yuanfeng, Xuefang Zhao, Raymond Chi-Wing Wong, and Di Jiang. "RGVisNet: A Hybrid Retrieval-Generation Neural Framework Towards Automatic Data Visualization Generation." ACM SIGKDD 2022. 1646-1655.
- [108] Zeng, Zehua, Phoebe Moh, Fan Du, Jane Hoffswell, Tak Yeon Lee, Sana Malik, Eunyee Koh, and Leilani Battle. "An evaluation-focused framework for visualization recommendation algorithms." IEEE TVCG, 2021, 28(1): 346-356.
- [109] Fayokemi Ojo, Ryan A. Rossi, Jane Hoffswell, Shunan Guo, Fan Du, Sungchul Kim, Chang Xiao, and Eunyee Koh. Visgnn: Personalized visualization recommendation via graph neural networks. WWW '22: The ACM Web Conference 2022.
- [110] Yu-Rong Cao, Xiao-Han Li, Jia-Yu Pan, and Wen-Chieh Lin. Visguide: User-oriented recommendations for data event extraction. CHI '22: CHI Conference on Human Factors in Computing Systems. New Orleans, LA, USA: ACM, 2022. 412:1-412:13.
- [111] Alicia Key, Bill Howe, Daniel Perry, and Cecilia R. Aragon. Vizdeck: self-organizing dashboards for visual analytics. SIGMOD 2012.
- [112] Halden Lin, Dominik Moritz, and Jeffrey Heer. Dziban: Balancing agency & automation in visualization design via anchored recommendations. CHI '20: CHI Conference on Human Factors in Computing Systems. Honolulu, HI, USA: ACM, 2020. 1-12.
- [113] Haotian Li, Yong Wang, Songheng Zhang, Yangqiu Song, and Huamin Qu. Kg4vis: A knowledge graph-based approach for visualization recommendation. IEEE Trans. on Visualization and Computer Graphics, 2022, 28(1):195-205.
- [114] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. ACL 2014, System Demonstrations. Baltimore, MD, USA. 2014. 55-60.
- [115] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 2017. 5998-6008.
- [116] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. 4171-4186.
- [117] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. ICANN '99, 1999. 97-102.
- [118] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. 27th Annual Conference on Neural Information Processing Systems 2013. 2787-2795.
- [119] Zhicheng Liu and Jeffrey Heer. The effects of interactive latency on exploratory visual analysis. IEEE TVCG 2014, 20(12):2122-2131.
- [120] Yue Zhao, Yunhai Wang, Jian Zhang, Chi-Wing Fu, Mingliang Xu, and Dominik Moritz. Kd-box: Line-segment-based kd-tree for interactive exploration of large-scale time-series data. IEEE Trans. on Visualization and Computer Graphics, 2022, 28(1):890-900
- [121] Dominik Moritz, Bill Howe, and Jeffrey Heer. Falcon: Balancing interactive latency and resolution sensitivity for scalable linked visualizations. CHI 2019. Glasgow, Scotland, UK: ACM, 2019. 694.
- [122] Zhifei Pang, Sai Wu, Gang Chen, Ke Chen, and Lidan Shou. Flashview: an interactive visual explorer for raw data. PVLDB 2017.
- [123] Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. IEEE Trans- actions on Visualization and Computer Graphics, 2013, 19(12):2456-2465.

- [124] Can Liu, Cong Wu, Hanning Shao, and Xiaoru Yuan. Smartcube: An adaptive data management architecture for the real-time visualization of spatiotemporal datasets. *IEEE transactions on visualization and computer graphics*, 2019, 26(1):790-799.
- [125] Cícero A. L. Pahins, Sean A. Stephens, Carlos Eduardo Scheidegger, and João Luiz Dihl Comba. Hashedcubes: Simple, low memory, real-time visual exploration of big data. *IEEE Transactions on Visualization and Computer Graphics*, 2017, 23:671-680.
- [126] Wenbo Tao, Xiaoyu Liu, Çagatay Demiralp, Remco Chang, and Michael Stonebraker. Kyrix: Interactive visual data exploration at scale. In: 9th Biennial Conference on Innovative Data Systems Research, CIDR 2019. Asilomar, CA, USA: Online Proceeding www.cidrdb.org, 2019.
- [127] Wenbo Tao, Xiaoyu Liu, Yedi Wang, Leilani Battle, Çagatay Demiralp, Remco Chang, and Michael Stonebraker. Kyrix: Interactive pan/zoom visualizations at scale. *Computer Graphics Forum*, 2019, 38(3):529-540.
- [128] Wenbo Tao, Xinli Hou, Adam Sah, Leilani Battle, Remco Chang, and Michael Stonebraker. Kyrix-s: Authoring scalable scatterplot visualizations of big data. *IEEE Trans. on Visualization and Computer Graphics*, 2021, 27(2):401-411.
- [129] Honghui Mei, Wei Chen, Yating Wei, Yuanzhe Hu, Shuyue Zhou, Bingru Lin, Ying Zhao, and Jiazhixia. Rsatree: Distribution-aware data representation of large-scale tabular datasets for flexible visual query. *IEEE TVCG*, 2019, 26(1):1161-1171.
- [130] Qingwei Lin, Weichen Ke, Jian-Guang Lou, Hongyu Zhang, Kaixin Sui, Yong Xu, Ziyi Zhou, Bo Qiao, and Dongmei Zhang. Bigin4: Instant, interactive insight identification for multi-dimensional big data. In: Yike Guo, Faisal Farooq. ACM SIGKDD 2018. 547-555.
- [131] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. The case for interactive data exploration accelerators (ideas). In: Carsten Binnig, Alan Fekete, Arnab Nandi. Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA'16.
- [132] Cícero A. L. Pahins, Nivan Ferreira, and João Luiz Dihl Comba. Real-time exploration of large spatiotemporal datasets based on order statistics. *IEEE Trans. on Visualization and Computer Graphics*, 2020, 26(11):3314-3326.
- [133] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. imMens: Real-time visual querying of big data. *Computer Graphics Forum*, 2013, 32(3).
- [134] Jia Yu and Mohamed Sarwat. Turbocharging geospatial visualization dashboards via a materialized sampling cube approach. In 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020. 1165-1176.
- [135] Fabio Miranda, Marcos Lage, Harish Doraiswamy, Charlie Mydlarz, Justin Salamon, Yitzchak Lockerman, Juliana Freire, and Claudio T. Silva. Time Lattice: A data structure for the interactive visual analysis of large time series. *Computer Graphics Forum*, 2018, 37(3):23-35.
- [136] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. Proftler: Integrated statistical analysis and visualization for data quality assessment. In Proceedings of the International Working Conference on Advanced Visual Interfaces, 2012. 547-554.
- [137] Niranjan Kamat, Prasanth Jayachandran, Karthik Tunga, and Arnab Nandi. Distributed and interactive cube exploration. In 2014 IEEE 30th International Conference on Data Engineering, 2014. 472-483.
- [138] Saptashwa Mitra, Paahuni Khandelwal, Shrudeep Pallickara, and Sangmi Lee Pallickara. Stash: Fast hierarchical aggregation queries for effective visual spatiotemporal explorations. In 2019 IEEE International Conference on Cluster Computing (CLUSTER), 2019. 1-11.
- [139] Chris Stolte, Diane Tang, and Pat Hanrahan. Multiscale visualization using data cubes. *IEEE TVCG*, 2003, 9(2):176-187.
- [140] Uwe Jugel, Zbigniew Jerzak, Gregor Hackenbroich, and Volker Markl. M4: A visualization-oriented time series data aggregation. *Proc. VLDB Endow.*, 2014, 7(10):797-808.
- [141] Alexander Kalinin, Ugur Cetintemel, and Stan Zdonik. Interactive data exploration using semantic windows. In: Curtis Dyreson, Feifei Li, M. Tamer Özsu. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD 2014. 505-516.
- [142] Punit R. Doshi, Elke A. Rundensteiner, and Matthew O. Ward. Prefetching for visual data exploration. In Proceedings of the Eighth International Conference on Database Systems for Advanced Applications, DASFAA '03. USA: IEEE, 2003. 195-202.
- [143] Sye-Min Chan, Ling Xiao, John Gerth, and Pat Hanrahan. Maintaining interactivity while exploring massive time series. In 2008 IEEE Symposium on Visual Analytics Science and Technology, 2008. 59-66.

- [144] Leilani Battle, Remco Chang, and Michael Stonebraker. Dynamic prefetching of data tiles for interactive visualization. In: Fatma Özcan, Georgia Koutrika, Sam Madden. Proceedings of the 2016 International Conference on Management of Data, SIGMOD 2016.
- [145] Hanqi Guo, Jiang Zhang, Richen Liu, Lu Liu, Xiaoru Yuan, Jian Huang, Xiangfei Meng, and Jingshan Pan. Advection-based sparse data management for visualizing unsteady flow. *IEEE Trans. on Visualization and Computer Graphics*, 2014, 20(12):2555-2564.
- [146] Liming Dong, Qiushi Bai, Taewoo Kim, Taiji Chen, Weidong Liu, and Chen Li. Marviq: Quality-aware geospatial visualization of range-selection queries using materialization. *SIGMOD* 2020. 67-82.
- [147] J. C. Baird. Psychophysical analysis of visual space. Pergamon Press Oxford, New York, [1st ed.] edition, 1970.
- [148] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387): pp. 531–554, 1984.
- [149] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE TVCG*. 19(12), 2013.
- [150] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *SIGCHI* 2010.
- [151] D. Kersten, P. Mamassian, and A. Yuille. Object Perception as Bayesian Inference. Department of Statistics, UCLA, pages 1–34, 2011.
- [152] Bolin Ding, Silu Huang, Surajit Chaudhuri, Kaushik Chakrabarti, and Chi Wang. Sample + seek: Approximating aggregates with distribution precision guarantee. In *SIGMOD*, 2016. 679-694.
- [153] Dominik Moritz, Danyel Fisher, Bolin Ding, and Chi Wang. Trust, but verify: Optimistic visualizations of approximate queries for exploring big data. In *CHI*, 2017. 2904-2915.
- [154] Danyel Fisher, Igor Popov, Steven Drucker, and m.c. schraefel. Trust me, i'm partially right: Incremental visualization lets analysts explore large datasets faster. In *CHI*, 2012. 1673-1682.
- [155] Sajjadur Rahman, Maryam Aliakbarpour, Ha Kyung Kong, Eric Blais, Karrie Karahalios, Aditya Parameswaran, Ronitt Rubinfeld, Sajjadur Rahman, Maryam Aliakbarpour, and Ha Kyung Kong. I've seen" enough": incrementally improving visualizations to support rapid decision making. *PVLDB*, 2017, 10(11):1262-1273.
- [156] Niklas Stoehr, Johannes Meyer, Volker Markl, Qiushi Bai, Taewoo Kim, De-Yu Chen, and Chen Li. Heatflip: Temporal-spatial sampling for progressive heat maps on social media data. *IEEE International Conference on Big Data*. Seattle, WA, USA: IEEE, 2018. 3723-3732.
- [157] Albert Kim, Eric Blais, Aditya G. Parameswaran, Piotr Indyk, Samuel Madden, and Ronitt Rubinfeld. Rapid sampling for visualizations with ordering guarantees. *PVLDB*, 2015, 8(5):521-532.
- [158] Daniel Alabi and Eugene Wu. Pfunk-h: approximate query processing using perceptual models. In *HILDA@SIGMOD*, 2016. 10-16.
- [159] Richard Wesley, Matthew Eldridge, and Paweł T. Terlecki. An analytic data engine for visualization in tableau. In: Timos Sellis, Renée J. Miller, Anastasios Kementsietsidis, Yannis Velegrakis. *Proceedings of the 2011 ACM SIGMOD*. New York: ACM, 2011. 1185-1194.
- [160] Mohamed Sarwat. Interactive and scalable exploration of big spatial data - a data management perspective. In *2015 16th IEEE International Conference on Mobile Data Management*, 2015, 1:263-270.
- [161] Xin Chen, Tong Ge, Jian Zhang, Baoquan Chen, Chi-Wing Fu, Oliver Deussen, and Yunhai Wang. A recursive subdivision technique for sampling multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 2020, 26(1):729-738.
- [162] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. Vizdom: Interactive analytics through pen and touch. *Proc. VLDB Endow.*, 2015, 8(12):2024-2027.
- [163] Dominik Moritz and Danyel Fisher. Visualizing a million time series with the density line chart. *CoRR*, 2018, abs/1808.06019.
- [164] Yunhai Wang, Kang Feng, Xiaowei Chu, Jian Zhang, Chi-Wing Fu, Michael Sedlmair, Xiaohui Yu, and Baoquan Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Trans. on Visualization and Computer Graphics*, 2018, 24(5).
- [165] Charles D. Stolper, Adam Perer, and David Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12):1653-1662.

- [166] Jianfeng Jia, Chen Li, and Michael J. Carey. Drum: A rhythmic approach to interactive analytics on large data. IEEE BigData. 2017.
- [167] Jean-Francois Im, Felix Giguere Villegas, and Michael J. McGuffin. Visreduce: Fast and responsive incremental information visualization of large datasets. 2013 IEEE International Conference on Big Data (IEEE BigData 2013). Santa Clara: IEEE, 2013. 25-32.
- [168] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. Finding waldo: Learning about users from their interactions. IEEE Transactions on visualization and computer graphics, 2014, 20(12):1663-1672.
- [169] Qiushi Bai, Sadeem Alsudais, Chen Li, and Shuang Zhao. Maliva: Using machine learning to rewrite visualization queries under time constraints. EDBT 2023. 157-170.
- [170] Zhe Wang, Dylan Cashman, Mingwei Li, Jixian Li, Matthew Berger, Joshua A. Levine, Remco Chang, and Carlos Scheidegger. Neural- cubes: Deep representations for visual data exploration. 2021 IEEE International Conference on Big Data (Big Data).
- [171] Christopher Root and Todd Mostak. Mapd: a gpu-powered big data analytics and visualization platform. ACM SIGGRAPH 2016 Talks.
- [172] Bryan McDonnel and Niklas Elmquist. Towards utilizing gpus in information visualization: A model and implementation of image-space operations. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1105-1112.
- [173] Ahmed Eldawy, Mohamed F. Mokbel, and Christopher Jonathan. Hadoopviz: A mapreduce framework for extensible visualization of big spatial data. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), 2016. 601-612.
- [174] Ahmed Eldawy, Mohamed F. Mokbel, Saif Alharthi, Abdulhadi Alzaidy, Kareem Tarek, and Sohaib Ghani. Shahed: A mapreduce-based system for querying and visualizing spatio-temporal satellite data. In 2015 IEEE 31st International Conference on Data Engineering, 2015.
- [175] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Geosparkviz: A scalable geospatial data visualization framework in the apache spark ecosystem. Proceedings of the 30th International Conference on Scientific and Statistical Database Management, SSDBM '18, New York: ACM, 2018.
- [176] Leilani Battle and Carlos Scheidegger. A structured review of data management technology for interactive visualization and analysis. IEEE Trans. on Visualization and Computer Graphics, 2021, 27(2):1128-1138.
- [177] Guoliang Li, Xuanhe Zhou, and Lei Cao. AI meets database: AI4DB and DB4AI. In: Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, eds. SIGMOD '21: International Conference on Management of Data, Virtual Event. China: ACM, 2021. 2859-2866.
- [178] Chelsea C White III and Douglas J White. Markov decision processes. European Journal of Operational Research, 1989, 39(1):1-16.
- [179] Apache hadoop homepage. <https://hadoop.apache.org/>. Accessed: 2022-05-01.
- [180] Apache spark homepage. <http://spark.apache.org/>. Accessed: 2022-05-01.
- [181] Lee, Bongshin, Arjun Srinivasan, Petra Isenberg, and John Stasko. "Post-wimp interaction for information visualization." Foundations and Trends® in Human–Computer Interaction, 2021, 14(1): 1-95.
- [182] Chen, Yiru, and Eugene Wu. "PI2: End-to-end Interactive Visualization Interface Generation from Queries." In Proceedings of the 2022 International Conference on Management of Data, 2022. 1711-1725.
- [183] Siddiqui, Tarique, Paul Luh, Zesheng Wang, Karrie Karahalios, and Aditya G. Parameswaran. "From Sketching to Natural Language: Expressive Visual Querying for Accelerating Insight." ACM SIGMOD Record, 2021, 50(1): 51-58.
- [184] Kim, Jun Ho. "Interactive interface for data analysis and report generation." U.S. Patent 11,397,746, 2022-07-26.
- [185] Pandey, Aditeya, Arjun Srinivasan, and Vidya Setlur. "MEDLEY: Intent-based Recommendations to Support Dashboard Composition." arXiv preprint arXiv:2208.03175, 2022.
- [186] Wang, Yun, Zhitao Hou, Leixian Shen, Tongshuang Wu, Jiaqi Wang, He Huang, Haidong Zhang, and Dongmei Zhang. "Towards Natural Language-Based Visualization Authoring." IEEE Trans. on Visualization and Computer Graphics, 2022.
- [187] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoong Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In International Symposium on Smart Graphics, Springer, 2010. 184-195.

- [188] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In Proceedings of the 28th annual acm symposium on user interface software & technology, 2015. 489-500.
- [189] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. Eviza: A natural language interface for visual analysis. In Proceedings of the 29th annual symposium on user interface software and technology, 2016. 365-377.
- [190] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. Applying pragmatics principles for interaction with visual analytics. IEEE transactions on visualization and computer graphics, 2017, 24(1):309-318.
- [191] Bowen Yu and Cláudio T Silva. Flowsense: A natural language interface for visual data exploration within a dataflow system. IEEE transactions on visualization and computer graphics, 2019, 26(1):1-11.
- [192] Arjun Srinivasan, Bongshin Lee, and John Stasko. Interweaving multimodal interaction with flexible unit visualizations for data exploration. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(8):3519-3533.
- [193] Vidya Setlur, Enamul Hoque, Dae Hyun Kim, and Angel X Chang. Sneak pique: Exploring autocompletion as a data discovery scaffold for supporting visual analysis. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, 2020. 966-978.
- [194] Vidya Setlur and Arathi Kumar. Sentifiers: Interpreting vague intent modifiers in visual analysis using word co-occurrence and sentiment analysis. In 2020 IEEE Visualization Conference (VIS), 2020. 216-220.
- [195] Vidya Setlur, Sarah Battersby, and Tracy Wong. Geosneakpique: Visual autocompletion for geospatial queries. In 2021 IEEE VIS.
- [196] Arjun Srinivasan and Vidya Setlur. Snowy: Recommending utterances for conversational visual analysis. In The 34th Annual ACM Symposium on User Interface Software and Technology, 2021. 864-880.
- [197] Xingbo Wang, Furui Cheng, Yong Wang, Ke Xu, Jiang Long, Hong Lu, and Huamin Qu. Interactive data analysis with next-step natural language query recommendation. arXiv preprint arXiv:2201.04868, 2022.
- [198] Can Liu, Yun Han, Ruike Jiang, and Xiaoru Yuan. Advisor: Automatic visualization answer for natural-language question on tabular data. In 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), 2021. 11-20.
- [199] Jan-Frederik Kassel and Michael Rohs. Valletto: A multimodal interface for ubiquitous visual analytics. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 2018. 1-6.
- [200] Arjun Srinivasan and John Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. IEEE transactions on visualization and computer graphics, 2017, 24(1):511-521.
- [201] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. Data@ hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In CHI 2021.
- [202] Gerwald Tschinkel, Cecilia Di Sciascio, Belgin Mutlu, and Vedran Sabol. The recommendation dashboard: A system to visualise and organise recommendations. In 2015 19th International Conference on Information Visualisation, 2015. 241-244.
- [203] Arjun Srinivasan, Steven M Drucker, Alex Endert, and John Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. IEEE transactions on visualization and computer graphics, 2018, 25(1):672-681.
- [204] Ruixian Ma, Honghui Mei, Huihua Guan, Wei Huang, Fan Zhang, Chengye Xin, Wenzhuo Dai, Xiao Wen, and Wei Chen. Ladv: Deep learning assisted authoring of dashboard visualizations from images and sketches. IEEE TVCG, 2020, 27(9):3717-3732.
- [205] Aoyu Wu, Yun Wang, Mengyu Zhou, Xinyi He, Haidong Zhang, Huamin Qu, and Dongmei Zhang. Multivision: Designing analytical dashboards with deep learning based recommendation. IEEE Transactions on Visualization and Computer Graphics, 2021, 28(1):162-172.
- [206] Siming Chen, Jie Li, Gennady Andrienko, Natalia Andrienko, Yun Wang, Phong H Nguyen, and Cagatay Turkay. Supporting story synthesis: Bridging the gap between visual analytics and storytelling. IEEE transactions on visualization and computer graphics, 2018, 26(7).
- [207] Tan Tang, Sadia Rubab, Jiewen Lai, Weiwei Cui, Lingyun Yu, and Yingcui Wu. istoryline: Effective convergence to hand-drawn storylines. IEEE transactions on visualization and computer graphics, 2018, 25(1):769-778.
- [208] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. Datashot: Automatic generation of fact sheets from tabular data. IEEE Trans. on Visualization and Computer Graphics, 2020, 26(1):895-905.
- [209] Michel Mitri. Story analysis using natural language processing and interactive dashboards. JCIS, 62(2):216-226, 2022.

- [210] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao. Calliope: Automatic visual data story generation from a spreadsheet. IEEE TVCG, 2021.
- [211] Jian Zhao, Shenyu Xu, Senthil Chandrasegaran, Christopher James Bryan, Fan Du, Aditi Mishra, Xin Qian, Yiran Li, and Kwan-Liu Ma. Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. IEEE TVCG, 2021.
- [212] Kenneth C. Cox, Rebecca E. Grinter, Stacie Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. A multi-modal natural language interface to an information visualization environment. Int. J. Speech Technol., 2001, 4(3-4):297-314.
- [213] Kevin Hu, Snehal Kumar Neil's Gaikwad, Madelon Hulsebos, Michiel A Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayanan, and Çağatay Demiralp. Viznet: Towards a large-scale visualization learning and benchmarking repository. CHI 2019.
- [214] Siwei Fu, Kai Xiong, Xiaodong Ge, Siliang Tang, Wei Chen, and Yingcai Wu. Quda: Natural language queries for visual data analytics. arXiv preprint arXiv:2005.03257, 2020.
- [215] Arjun Srinivasan, Nikhila Nyapathy, Bongshin Lee, Steven M Drucker, and John Stasko. Collecting and characterizing natural language utterances for specifying data visualizations. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.
- [216] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. Discovering natural language commands in multimodal interfaces. In Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019. 661-672.
- [217] Nik Nur Ayuni Nik Alwi, Noor Haftzah Hassan, Ahmad Fazreen Baharudin, Nur Azaliah Abu Bakar, and Nurazean Maarop. Data visualization of supplier selection using business intelligence dashboard. In International Visual Informatics Conference, 2019. 71-81.
- [218] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. More than telling a story: Transforming data into visually shared stories. IEEE computer graphics and applications, 2015, 35(5):84-90.
- [219] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. ACM Transactions on Graphics, 2019, 38(4):1-15.
- [220] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248-255.
- [221] Philipp Eichmann, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. Idebench: A benchmark for interactive data exploration. SIGMOD Conference 2020, online conference. Portland, OR, USA:ACM, 2020. 1555-1569.

附中文参考文献:

- [1] 中国人工智能 2.0 发展战略研究项目组.《中国人工智能 2.0 发展战略研究》.第一版,中国:浙江大学出版社,2018.
- [2] 陈为, 沈则潜, 陶煜波.《数据可视化》.第二版.中国:电子工业出版社, 2013.
- [13] 夏佳志, 李杰, 陈思明, 秦红星, 刘世霞. 可视化与人工智能交叉研究综述. 中国科学: 信息科学. 2021.
- [43] 任磊, 杜一, 马帅, 张小龙, 戴国忠. 大数据可视分析综述. 软件学报, 2014, 25(9):1909-1936. <http://www.jos.org.cn/1000-9825/4645.htm>
- [72] 郝爽, 李国良, 冯建华, 王宁. 结构化数据清洗技术综述. 2018.



骆昱宇 (1996-), 男, 博士生, 主要研究领域为智能数据可视分析, 支持高效可视分析的数据管理。



秦雪迪 (1995-), 女, 博士生, 主要研究领域为数据生成, 数据探索, 数据可视化。



谢宇鹏 (1996-), 男, 硕士生, 主要研究领域为数据可视化。



李国良 (1980-), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为数据库, 大数据分析与挖掘, 群体计算。