数据诱导的分组及其发现方法

Yilan Jiang、Cindy Xiong Bearfield、Steven Franconeri、Eugene Wu

arXiv:2601.08256v1 [cs.HC] 13 Jan 2026

摘要——**理解可视化需要读者同时考量可视化设计和底层数据值。可视化领域的既有研究主要关注由可视化设计元素（如颜色或图表类型）驱动的感知特性，但视觉设计如何与数据值相互作用，进而影响解读和推理，仍未得到充分探讨。点图和条形图常用于帮助用户识别构成趋势和簇的数据点组，但也容易呈现出由空间排列而非数据本身固有模式所形成的分组。这些“数据诱导分组”可能导致次优的数据比较，并可能使用户得出错误结论。我们以点图为案例研究，开展了两项用户研究，以了解数据诱导分组的普遍程度。我们发现，尽管基于趋势的分组与名义数据无关，用户在两种条件下仍都会依赖数据诱导的分组。基于研究结果，我们建立了一个模型，用于预测用户是否可能将给定的一组点图数据点感知为一组。我们还讨论了两个应用案例，说明该模型如何帮助可视化设计师诊断点图中潜在的用户感知分组，并通过数据重排提供重新设计方案，以更好地突出期望的分组。**

索引词—**知觉分组、知觉机制、设计工具。**

I. 引言

可视化使读者能够快速提取关键信息并比较数据，而糟糕的设计则可能误导读者或令其不知所措。可视化领域已有研究提出了一些指导原则，帮助人们根据读者意图（例如分析目标）、数据类型（例如类别型、连续型）或数据分布（例如数据点总数、属性之间的相关性）选择合适的设计[1]–[4]。例如，散点图适用于聚类和相关性任务 [5], [6]，但不太适合用于比较两类数据平均值等摘要任务，尤其是在数据点数量增加时 [2]。折线图有助于利用定量数据检测趋势，甚至可以缓解因果偏差 [7], [8]，但用于可视化类别数据时会产生误导 [9]。

我们认为，可视化设计师还应考虑设计选择与底层数据值之间的相互作用。例如，数据值的空间排列可能使用户看到一些分组或趋势，而这些其实是由我们的感知编码机制造成的，并非数据本身所具有的。无论用户的目标是什么，这些人为形成的分组都可能使用户更倾向于看到某些关系或进行某些视觉比较，从而得出无效结论。

本文旨在以点图为研究场景，了解用户如何对数据进行分组。尽管条形图是汇总数据的传统选择，但越来越多的研究表明，条形图会引入自身的系统性偏差，例如条形内部效应和条形末端限制效应 [10], [11]。因此，近期研究日益倡导使用基于点的表示，并常辅以变异性指标，作为一种更清晰、偏差更小的替代方案 [12]。这一转变使得专门考察点图中的分组行为变得十分重要。我们创建了一个示例点图，如图1所示，其中呈现了七名学生考试成绩的名义数据；不同的x轴排序会凸显不同的分组和模式。读者可能会将成绩最高的三名学生归为一组，或将左侧的五个数据点与其余数据点区分开来。或者，读者可能会注意到所有学生的成绩呈“上升趋势”，但由于x轴表示的是名义数据，这种趋势具有误导性。

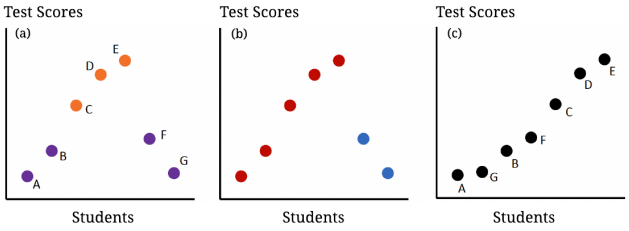


图1. 展示七名学生考试成绩的名义数据。人们可以将成绩最高的三名学生视为一组，将其余学生视为另一组（a）；也可以将左侧五个数据点看作呈现“增长趋势”，并将右侧两个数据点视为异常值（b）。然而，名义型x轴上的排列顺序并无意义。重新排列x轴，可能会让观察者认为全部七个数据点呈现“增长趋势”（c）。

在这个例子中，感知到的分组发生了变化，而底层数据保持不变。Scheidegger 和 Kinneman 将这些分组称为幻觉模式 [13]，因为视觉表示（以及由此感知到的模式）发生了变化，而数据并未改变。这些幻觉分组在底层数据中隐含地引入了一个“组 ID”通道。例如，图1（左）引入了一个隐含的分组属性，其中 CDE 构成组 1，ABFG 构成组 2。我们将其称为数据诱导的分组。

数据诱导的分组与人类知觉文献中的知觉分组有关。尽管这一概念类似于<stijl id='1'>幻觉模式，后者指用户感知到的非预期模式，但数据诱导的分组专门关注空间配置的影响，不涉及更广泛的认知偏差。在数据解读过程中，读者会通过视觉分组将可视化拆分为若干单元，从而对数据点 [3]进行分析。这个过程使读者能够进行比较并

Yilan Jiang 任职于美国伊利诺伊大学厄巴纳-香槟分校，地址为美国伊利诺伊州厄巴纳，邮编 61801（电子邮箱：yilanj2@illinois.edu）。Cindy Xiong Bearfield 任职于美国佐治亚理工学院，地址为美国佐治亚州亚特兰大，邮编 30332（电子邮箱：cxiong@gatech.edu）。Steven Franconeri 任职于美国西北大学，地址为美国伊利诺伊州埃文斯顿，邮编 60208（电子邮箱：franconeri@northwestern.edu）。Eugene Wu 任职于美国哥伦比亚大学，地址为美国纽约州纽约市，邮编 10027（电子邮箱：ewu@cs.columbia.edu）。

计算生成观看者结论所需的数据统计。关键在于，一旦读者在知觉上将某些数据点归为一组，他们就需要付出极大努力才能“打破”这些组、形成新的组，并比较这些新组 [14], [15]。因此，设计师应当经过深思熟虑地构建可视化，以促进“正确”的分组行为。已有研究表明，可视化设计（例如颜色和空间排列的选择）会影响读者如何在视觉上对数据进行分组以及比较数据 [14], [16]。然而，研究人员在开展用户研究时，通过随机化数据值来控制其潜在影响；因此，底层数据值对分组行为的影响，以及数据值可能如何与设计选择交互作用、进而影响观看者结论，仍未得到探索。在本文中，我们构建数据驱动的分类体系和预测模型，以推进对可视化中数据诱发的分组行为的理解。遵循先前可视化研究的做法，该研究表明，在解读可视化时，即使只有六个数据点的简单可视化，也能帮助研究人员了解潜在的知觉分组机制 [17],，因此我们采用类似的设计来考察包含六个数据点的点图。研究这些更简单的案例，为在扩展到更大、更复杂的数据集之前揭示潜在的知觉机制提供了一种可行的方法 [18]。这项研究使我们能够提供一种自动化方法，揭示人类分组行为中更广泛的趋势，从而让实践者在可视化设计中考虑数据值的影响；同时，通过赋予设计师在设计过程早期打破幻觉模式的能力，减轻数据诱导分组所带来的有害影响。

本文有三项主要贡献。

- 首先，我们开展了两项用户研究，以了解数据诱导的分组通道的普遍程度和特征，以及指示抵消视觉分组效应的能力。两项研究都展示了由 6 个随机数据点构成的点图，其中 x轴为名义型，y 轴为数值型。第一项研究要求参与者识别他们感知到的分组；第二项研究则额外提醒参与者，x轴为名义型。我们研究参与者是否会依据 x 轴上的趋势或聚类进行判断，即使已明确提醒他们该轴的名义型性质。

- 其次，我们利用用户研究的结果构建预测模型。给定点图中的一组数据点，这些模型会预测用户将这些数据点感知为一组的可能性。模型根据这组数据点的共线性和聚类特征，提取可解释的特征。在留出数据集上，我们的模型取得了 97 – 99% 的 F1 分数（精确率和召回率的调和平均数）。

- 第三，我们介绍该模型的两种应用：一是作为诊断工具，预测点图中的数据诱导的分组；二是作为重新设计工具，帮助可视化创作者诱导或打破数据诱导的分组。我们将这些功能整合到可视化设计工具中，帮助设计师预判潜在的数据诱导的分组，并从数百种可能的重新设计方案中进行搜索，从而找到能够突出有助于实现预期比较的组的设计。

II. 相关工作

识别一种视觉构型不仅涉及辨认各个组成部分，还涉及理解它们之间的关系。这一过程称为知觉组织，指的是识别场景中不同元素之间的关联方式 [19]。通过这一机制，视觉元素被分组，并形成结构，从而构成高层次的视觉对象。

知觉组织在我们解读可视化时发挥着至关重要的作用。观看者可以运用自上而下控制来影响可视化的组织方式，而知觉组织也会自然地从低层级组件构建出分层的视觉表征。关于知觉组织的理论大多基于格式塔分组原则 [20]。例如，接近原则指出，彼此位置接近的元素通常会被归为一组；相似原则则认为，具有相似特征（如颜色或形状）的物体会被感知为一组 [21]。本文聚焦于连续性原则：当元素形成连续轮廓时，它们便会被归为一组 [22], ，例如点图中排列整齐的数据点。

A. 可视化中的比较

比较是一项重要的视觉分析任务 [23]。例如，在可视分析中，常见的用户目标包括将观测数据与假设结果进行比较，或比较代码更改前后的模式，或比较不同操作条件下的模式 [24], [25]。进行比较时，人们总是需要先对视觉元素进行分组，因为分组决定了可供比较的单位。视觉系统往往优先关注“整体而非细节”，因此比起单个物体，人们更容易感知由物体构成的组 [26], [27]。这是因为辨别单个物体需要更多时间和注意 [28]。在可视化的上下文中，既有研究表明，空间线索是最强大的分组机制之一 [14]。因此，我们可以推断，最容易进行的比较会涉及空间上归为一组的数据点；数据点之间的邻近性有助于人们更快速、更直观地进行比较。

可视化研究群体开发了支持数据视觉比较的工具。例如， [29]提出了一种概念模型，允许人们总结数据并计算数据标记、趋势及其他图表元素之间的差异，从而在数据探索过程中进行临时比较。Hearst 等人 [30] 根据数据分布，探索了针对单数和复数最高级（例如“最高价格”和“最高价格们”）以及数值型等级形容词（例如“更高”）的恰当可视化响应。

Gaba 等人 [31] 采用了不同的方法。在识别出与特定比较类型相关的自然语言后，作者建立了自然语言描述的比较意图与最适合实现这些比较的可视化设计之间的映射。研究人员还构建了系统和比较可视化工具，以满足更具体的比较需求 [32], ，例如用于比较社交网络的系统 [33], ，以及支持虚拟现实中的数据比较的工具

，还有用于揭示两个多项数据流或序列之间相似点和差异的可视化 [35]。

与此同时，在这些旨在支持对可视化数据进行比较的研究中，人类参与者的实证研究结果也推动了人们对视觉比较所涉及的感知与认知机制的理解。例如，某些视觉特征很容易比较，例如物体平移 [36] 和数据密度 [37]。其他视觉编码通道则可能增加比较任务的难度 [38]–[40]，，例如方向、纹理、颜色和缩放。这些视觉特征会带来不同的比较可供性。例如，人们比较的数据值组合会因其空间排列方式而异（例如，垂直对齐与水平对齐） [41]–[43]，，进而得出不同的结论[17]。选择不同的可视化类型也会影响人们比较哪些数据，以及从中得出什么结论。例如，人们阅读条形图时，更喜欢比较成对数据量值（例如，这根条比那根条高）；而阅读折线图时，则更喜欢比较随时间发生的变化或数据间的关系（例如，x 随时间增加） [5], [8], [9], [44], [45]。呈现的数据值数量也会影响比较的速度和准确性 [46], [47]。需要比较的项目越多，比较效率就越低。

B. 可视化中的知觉与分组

要进行数据比较，查看者必须先提取数据组，然后对其进行比较，以识别差异或相似之处 [48]–[50]。数据组的提取通常通过将可视化中的数据点划分为视觉单元来完成 [51]。由于我们的视觉系统倾向于“先见森林，后见树木” [26]，，对数据元素进行分组比感知单个数据项更直观 [28], [52]。可视化领域的既有研究已经生成了相关模型，用于预测人们倾向于将哪些数据元素分为一组 [14], [16]。例如，当人们阅读描绘两组、每组包含两个条形的条形图时，他们倾向于将空间位置相近的条形归为一组，并比较这些组的平均高度 [14]。

数据点的知觉特征会影响人们形成视觉组的方式。这些特征通常分为“空间”特征或“特征”特征 [53]–[55]。特征线索可以包括颜色、形状或尺寸编码。

有些特征比其他特征更容易、更直观地被分组。例如，颜色相似性比形状相似性更能作为分组线索，因此，与形状各异的簇相比，颜色各异的散点图中的簇更容易被准确区分 [56], [57]。在形状编码中，与正方形和圆形等封闭形状相比，十字形和星号形等开放形状更难与其他开放形状区分开来 [58]。空间线索包括邻近性（彼此靠近的物体）、共线性（沿同一条直线排列的物体）和区域共享（位于同一区域内的物体）。一般来说，数据元素之间的距离很重要。增大两个单元之间的间距，有助于将它们分离 [59], [60]。在条形图中，以不规则的间距将条形分组，可以使有关这些组的信息比采用均匀间距的条形图更加突出 [16]。

现有研究表明，在影响知觉方面，空间线索的作用往往超过特征线索 [54], [61]–[63]。即使观察者知道按空间位置对数据单元分组会损害任务表现，空间线索也很难被抑制 [20], [60]。数据可视化往往同时利用空间线索和特征线索来编码数据，但在可视化知觉中，空间分组线索似乎会优先受到关注 [14]。观察者一旦选定一个元素作为注意焦点，就会对附近的元素更加敏感 [64], [65]。因此，他们更可能选择空间位置更近的元素进行比较，而不考虑这些元素是否具有相似的特征 [20]。例如，即使两根相邻的条形颜色不同，观察者也会优先比较它们，而不是优先比较颜色相同但彼此距离较远的条形。

分组线索会显著影响用户在可视化中比较哪些元素。如果分组线索误导观察者形成“错误”的视觉组，他们最终会进行无效或低效的比较。然而，由于分组知觉特征之间可能相互竞争，因此要确保观察者选择“正确”的组并不容易。可视化领域目前还不全面了解设计特征如何驱动视觉数据解读中的分组行为 [16]。这可能会带来危险，因为一旦形成视觉组，就极难打破这些组并形成新的组（这一过程称为知觉重组） [14]。对视觉智力测试的分析表明，知觉重组任务是测试中最困难的问题之一，因为这类任务需要付出大量努力进行知觉计算 [15]。这表明，图表观察者在进行知觉重组时会遇到困难。因此，设计可视化时，至关重要的是立即且有力地支持“正确”组的形成。这促使我们开发一种模型，用于预测并约束图表观察者在前注意阶段形成的视觉组。

III. 数据诱导的分组

我们在 Prolific [66] 上开展了两项研究，以了解参与者如何在感知上对数据进行分组，以及这些分组在何时可能导致误解。两项研究都使用了带有名义型 x 轴和数值型 y 轴的点图。第一项研究要求参与者识别组，第二项则阐明 x 轴的名义型性质。我们假设，明确的指示可以减少参与者对 x 轴排序或分组的依赖。

随后，我们训练了可解释机器学习模型，以预测参与者是否会将图表中的给定数据点子集识别为一个组。借助这些模型，我们评估了假设，并提出了一种考虑数据诱导的分组的可视化设计工具。

A. 用户研究

1）一般方法：在两项研究中，参与者都使用图 2 所示的界面，观看一系列点图。每张点图都包含六个均匀分布在 x 轴上的数据点，其 y 轴取值在 0 到 100 之间随机生成。

选择仅包含六个数据点的主要原因，是在认知负荷与参与者参与度之间取得平衡[67]。每项研究都包含 40 张随机点图，其 y 轴取值按固定纵横比（5–95）缩放 [68], [69]。

研究1：星座（N = 20）：我们请用户“在给定的图表中，突出显示所有看起来‘属于一起’的数据点组。”总体而言，参与者从40张图表中识别出2380个组。我们发现，共线性是群组形成的有力预测因素，因为参与者会将趋势视为组。即使x轴是名义型且无关紧要，x轴和y轴上的分离也会影响聚类。由于这项研究运用知觉分组原则来识别夜空中的星座组，与之相似，因此我们将其命名为 *Constellation* 研究 [70]。

研究2：情境（N = 20）：这项研究通过将六个数据点与六个肥料品牌的绩效分数（y 轴）相关联，提供了真实世界的数据背景（x轴）。这种设定旨在让参与者了解数据背景，鼓励他们根据数据中感知到的关系进行分组，而不只是依据基于图像的模式。指示通过提及肥料，强调了x轴的名义型属性。我们请参与者假设哪些数据点可能共享某个将它们联系起来的隐藏变量。我们也不希望鼓励或阻碍任何特定类型的分组。因此，我们给出的指示如下：“查看每张图表时，请突出显示那些看起来‘属于一起’的数据点组，以帮助确定哪些肥料可能具有相似的活性成分”。参与者从40张图表中识别出2426个组。我们请他们为每张图表 5^{th} 写下观察结果，以便更深入地了解参与者如何形成各自的组。我们将这项研究称为语境图研究。

参与者使用界面（图 2）来高亮或套索选中他们认为属于同一组的数据点。他们可以清除选择、将一组添加到右侧的图库中，也可以从图库中移除一组。当他们对已添加的组感到满意时，可以点击提交。

为了让参与者熟悉界面，他们必须完成一个教程，学习如何在示例任务中高亮、删除和清除一组数据点；教程同时采用文字指示和动画演示。每位参与者都必须按照指示在示例图表上进行操作，才能继续。研究结束时，参与者需要填写一份简短的人口统计调查（年龄组和教育水平）。所有参与者均按每小时 12 美元的标准获得报酬。

为尽量减少众包平台 [71], [72], 上的恶意行为，我们采取了多项措施。我们在 Prolific 上预先筛选参与者，要求其英语流利度达标，且过往通过率至少为 97%。我们记录了任务完成时间，并剔除了异常值。最后，我们在启动正式实验前开展了一系列试点研究。此外，21 名参与者使用与正式研究相同的刺激材料和界面参加了试点研究。试点结果为教程部分的设计提供了依据，以确保参与者能够顺利完成实验。

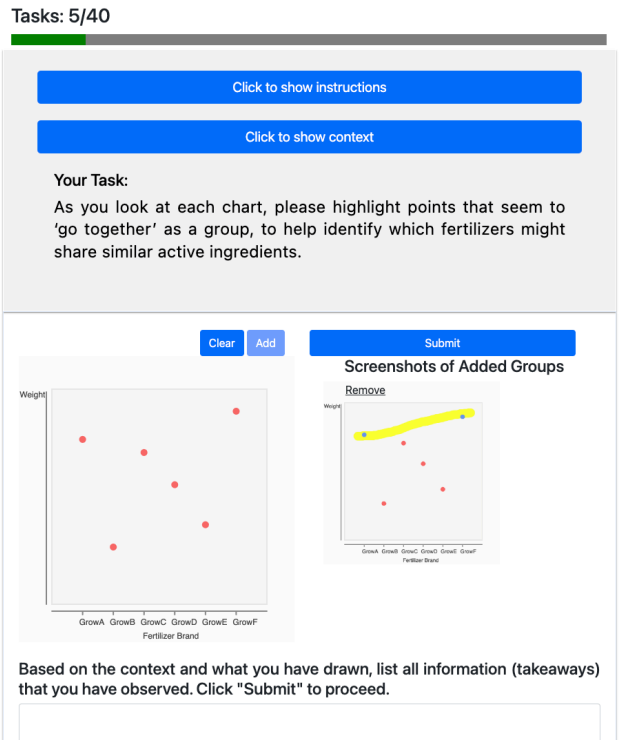


图 2. 任务界面。参与者可以高亮或套索选中数据点组、清除选择，并将组添加到右侧的图库中或从中移除。每完成 5 个图表，我们还会添加一个文本框，让参与者描述他们在图表中看到的内容。

在招募的 61 名参与者中（研究 1 有 20 名，研究 2 有 20 名，试点研究有 21 名），47 人填写了完整的人口统计问卷。他们的平均年龄为 26.4 岁（标准差 = 5.9，范围为 19-41 岁）。略低于一半的人拥有至少学士学位（21/47 ≈ 45%），其余人则报告拥有副学士学位、读过部分大学课程或持有高中毕业证书（55%）。在固定点李克特量表上对过往数据可视化经验进行评分（1=完全没有， 5 =经常创作），平均分为 2.77 ± 1.34，表明他们对此的熟悉程度为低至中等。参与者对任务投入程度的自我报告采用三点量表（1 =非常少， 3 = 很高），平均分为 2.45 ± 0.62，这表明参与者总体上积极参与了教程和分组任务。

2) 假设：我们为该研究提出了两个假设。虽然二者相似，但第一个旨在验证：当 x轴 已知为名义型时，参与者对趋势形状的依赖会减少；第二个则旨在确定参与者会依赖哪些因素：

H1：在 Contextual 研究中，与 Constellation 研究相比，参与者对共线性的依赖会减少。

H2：在 Contextual 研究中，与 Constellation 研究相比，参与者会更加关注 y轴数值。

B. 用户研究结果概述

在深入探讨模型开发之前，我们首先概述用户选择的组。我们先将这些组整理为关于组的基本统计量，然后介绍关于组的语义特征。

1) 基本统计量：首先，我们报告图 3 中两项研究的总体组数和组大小，以及按图表和参与者分别统计的组数与组大小，

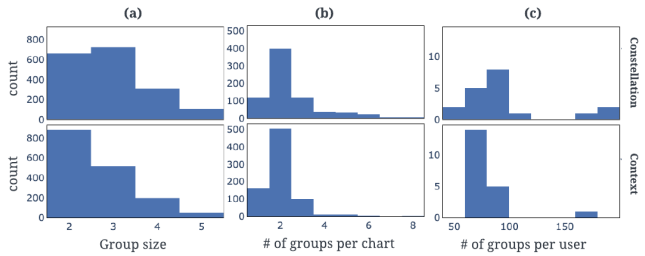


图 3. (a) 群组大小的分布，(b) 每张图表中组数的分布，以及 (c) 每位参与者所对应组数的分布。

Constellation 研究的群组大小更大，每张图表中的组数更多，每位参与者划分的组也更多。在两项研究中，尤其是 Contextual 研究中，每张图表的组数主要集中在两个左右。这是合理的，因为总共只有六个数据点，而且大多数组由两到三个数据点组成，因此参与者很可能只依据 y 轴来划分组。

2) 语义统计：为了研究这些组的性质，我们手动检查了大量组，并根据观察到的常见属性设计了两类特征：共线性和聚类。共线性特征通过对输入组拟合线性模型，并计算其误差（残差和）及斜率得出。聚类特征基于这样一种直觉：参与者会根据数据点与图表中其他数据点之间的距离来识别簇。形成簇是一种直观的感知过程，在视觉分析中很常见，尤其是在点图中 [70], [73], [74]。空间位置相近或具有相似视觉特征（例如颜色相同）的数据点往往会被归为一组。一个挑战是，评估目标数据点组 g与图表中其余数据点 r之间簇分离程度的方法有很多。最终，我们设计了若干常见特征，详述如下，并汇总于图 4：

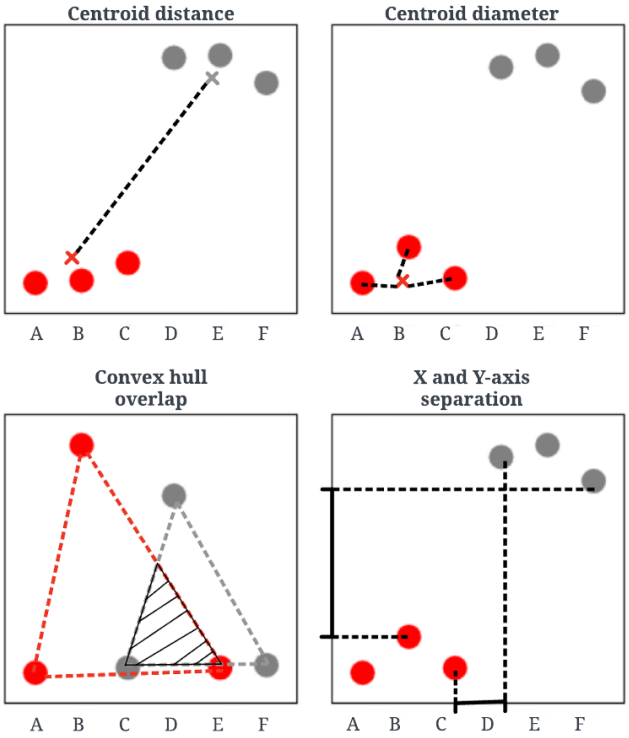


图 4. 聚类特征示例，不包括簇比；簇比是质心距离与簇直径的比值。红色圆点表示候选组g。

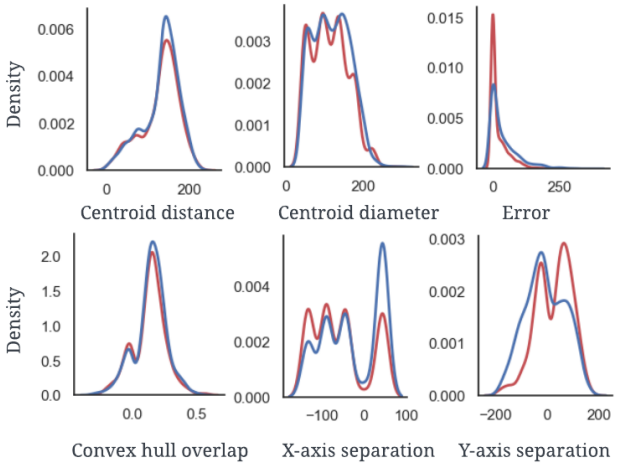


图 5. 质心距离、质心直径、误差、凸包重叠、x 轴和 y 轴分离的分布。Constellation 研究用蓝色表示，而Contextual 研究用红色表示。

- **质心距离**：我们计算g 和r的质心之间的欧几里得距离，并除以图表的对角线长度进行归一化。
- **质心直径**：我们计算g 中的每个点到其质心的平均欧几里得距离。
- **质心比率**：我们计算质心距离与质心直径的比值。
- **凸包重叠**：给定g 的凸包h_g （以及r的凸包h_r ），我们计算它们重叠部分的周长与其面积并集的比值。对于构成线段的凸包，我们假设该线段的宽度为 1 像素：$(h_g \cap h_r)/(h_g \cup h_r)$。
- **X 轴和 Y 轴间距**：当 x 轴为名义型时，沿 x 轴的间距不应作为识别组的依据。因此，我们分别计算每个轴的分离度量。对于 y 轴，特征f_y 用于衡量g 中的数据点与r中的数据点之间的最小距离。类似地，f_x 定义为沿 x 轴的分离度量。我们期望 f_y 成为主导特征；当 x 轴为名义型时，共线性和f_x 不应影响视觉分组。

图 5 展示了各项特征的整体密度分布。在两项研究中，质心距离、质心直径和凸包重叠等特征呈现出相似的分布。值得注意的是，在 Contextual 研究中，误差更集中在 0 附近，这与两点组出现频率较高的情况相符，如图 3 所示。Constellation 研究中的组也更倾向于沿 x轴 分布在距离其他数据点较远的位置，而 Contextual

表 I 语境图因子的固定效应估计（Constellation 与语境图）。

因变量	β_1	SE	z	p
误差	15.630	5.523	2.830	**0.005**
质心直径	2.407	6.649	0.362	0.717
质心距离	2.821	3.959	0.713	0.476
凸包重叠	0.013	0.012	1.085	0.278
X轴间距	29.578	5.990	4.938	$<$ **0.001**
Y轴间距	−33.049	8.481	−3.897	$<$ **0.001**

研究中的组则更倾向于分布在 y 轴 上距离较远的位置。

为了评估六项语义指标是否因研究条件而异，我们分别拟合了线性混合效应模型，将研究条件（*Constellation* 与语境图）设为固定效应，并为参与者设置随机截距（每位参与者绘制 40 张图表）。表 I 报告了固定效应系数（β_1），用于量化相对于语境图基线的平均变化。与语境图相比，Constellation 图表的误差显著更高，且x轴和 y 轴间距更大（均$p < .01$）；质心直径、质心距离和凸包重叠则未显示出显著差异。这些推断性结果进一步印证了图 5 中的描述性趋势。。

C. 模型开发

为了分析每项研究的结果，我们设计了一个预测模型，以一组数据点为输入，并输出参与者将其选为一组的可能性。一个挑战是，组中的数据点数量可能各不相同，因此无法直接将其用作模型输入。此外，我们旨在开发一个采用简单特征的可解释模型 .

模型与特征。 基于这些原因，我们使用了第 III-B 节中的语义特征，并通过逻辑回归和决策树模型对其进行评估。我们测试了卷积神经网络 (CNNs)，但发现其性能与更简单的决策树相当（F1：96–97%）。鉴于数据集规模较小且准确率相近，我们优先选择了轻量、透明的模型。

训练数据。 拟合模型的一项挑战是，研究数据可能存在类别不平衡，因为参与者并不总是会选择非组。我们无法简单地将所有未被选中数据点的子集标记为负例。例如，参与者可能会选择数据点 AB 作为一组，也可能会将 CDE 视为另一个潜在组，但选择不对其进行标记。

因此，我们试图识别每个图表中明显不属于组的数据点子集。具体而言，设g 为一个候选数据点子集，随后我们依据三个标准对其进行评估：1）g 未被任何参与者选中；2）相较于参与者选中的组，其线性拟合效果较差；3）g的凸包与图表中其余数据点形成的凸包不重叠。第二个标准通过线性拟合误差（残差和）进行衡量，并检查g的误差是否大于平均误差。第三个标准只需检查凸包特征是否为零。

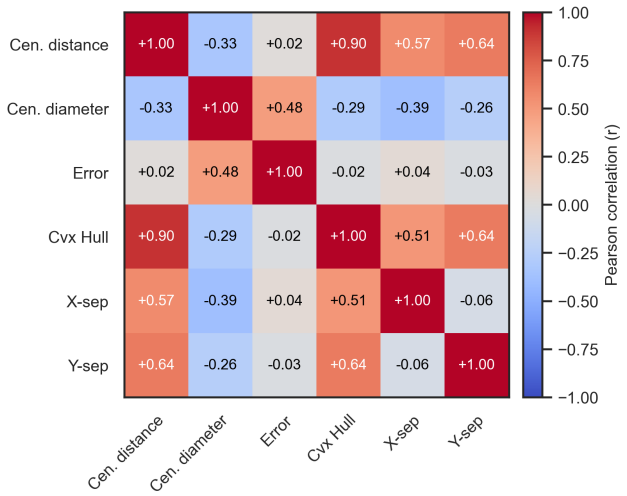


图 6. Constellation 研究中六个语义特征的成对皮尔逊相关矩阵。单元格颜色表示相关性的强度和方向（红色 = 为正，蓝色 = 为负）；数值标签表示四舍五入到小数点后两位的r。缩写——Cen. distance：质心距离；Cen. diameter：质心直径；Error：分组误差；Cvx Hull：凸包重叠；X-sep：x轴间距；Y-sep：y 轴间距。

基于这些要求，我们生成了1272个示例来验证负例。其中一位作者手动标注了这些示例，以识别其中的“假”负例。如图 10所示，在这一过程中，我们将所有红色数据点确实 构成一个组的实例归类为“假”负例。在1271个实例中，我们只识别出42个假负例，这表明我们的负例生成过程是有效的。

由于每张图表仅包含六个数据点，而只有一个数据点的组没有意义，因此我们将训练集限制为包含 2 至 5 个数据点的组。总而言之，Constellation 研究包含 1502 个正例和 1308 个负例，而 Contextual 研究包含 1639 个正例和 1396 个负例。对于每项研究，我们使用 70% 的数据进行训练和超参数调优，20% 用于测试，剩余 10% 作为留出数据来报告最终结果。

D. 模型训练与评估

我们使用各项用户研究的数据和所有可用特征，分别训练了逻辑回归模型和决策树模型，并将决策树的最大深度限制为 3。对于逻辑回归，我们使用方差膨胀因子（VIF）剔除高度相关的特征。值得注意的是，质心直径和质心距离与凸包重叠和 x 轴分离度高度相关，因此我们剔除了前两者。图 6 展示了 Constellation 研究中六个语义特征的皮尔逊相关矩阵，突显了它们之间的两两依存关系。

为提高可读性，我们在表格中使用缩写形式的指标名称：**P** 表示精确率，**R** 表示召回率。全文始终使用这些缩写。表 II 报告了五折交叉验证（交叉验证）测试分数的均值和标准差，以及独立留出（留出）集上的结果。对于留出数据，表 II 显示所有模型的精确率均为 100%，这意味着它识别出的任何组都被

表 II 五折交叉验证（交叉验证）和独立留出（留出）性能。每个条目表示为“交叉验证均值 ± 标准差 / 留出”。粗体 F1 标示每项研究中的最佳模型。

	逻辑回归			决策树		
	P	R	F1	P	R	F1
Constellation	0.94 ± 0.02 / 1.00	0.88 ± 0.02 / 0.87	0.91 ± 0.01 / 0.93	0.97 ± 0.03 / 1.00	0.88 ± 0.03 / 0.95	**0.92 ± 0.03 / 0.97**
上下文	0.94 ± 0.01 / 1.00	0.91 ± 0.02 / 0.86	0.92 ± 0.01 / 0.92	0.92 ± 0.01 / 1.00	0.94 ± 0.01 / 0.91	**0.93 ± < 0.01 / 0.95**

表 III 逻辑回归（LR）和决策树（DT）模型的特征重要性（SHAP 值绝对值均值）。缩写：X-SEP = x 轴分离度，Y-SEP = y 轴分离，凸包 = 凸包重叠，质心比 = 质心比率。

		特征重要性（|SHAP|）					
		斜率	误差	X 分离	Y 分离	凸包	质心比
Constellation	LR	0.01	0.00	0.06	0.03	**0.31**	**0.22**
Constellation	DT	0.00	**0.19**	0.00	0.00	**0.15**	**0.23**
上下文	LR	0.01	**0.09**	0.02	**0.10**	**0.14**	**0.18**
上下文	DT	0.00	**0.15**	0.00	**0.13**	**0.16**	**0.11**

参与者识别为组。这表明其正向预测具有可靠性。与逻辑回归相比，决策树模型的召回率提高了 $5-9\%$，这可能是因为特征与目标变量之间存在非线性关系。最后，在两项研究中，决策树的 F1 分数介于 96% 至 97% 之间，凸显了即便是简单的决策树模型，也能有效识别数据诱导的分组。

随后，我们使用 SHapley 加性解释(SHAP) [75]来估算每个特征对最终结果的贡献。如 Eq.1 所示，其中 F 是所有特征的集合，SHAP 通过对所有可能组合中的加权贡献取平均，计算特征 i 的边际贡献，记为 ϕ_i。由于 SHAP 是模型无关的，SHAP 值可作为“特征重要性”的指标，从而能够比较逻辑回归与决策树模型中特征重要性的变化。

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \quad (1)$$

表 III 展示了两个模型和两项研究中的特征重要性绝对值。例如，斜率的 SHAP 值表明，较大的斜率对模型最终结果的边际贡献平均为 $0-0.01$。SHAP 值也可以按单个样本计算；不过，本研究报告的是所有样本的平均值。

1) 分析：我们发现，凸包重叠、质心比率和共线性（以线性拟合<rs tyle id='3'>误差衡量）对群组识别的影响最大。结合模型的高准确率来看，这些因素的显著贡献支持了我们的直觉：簇分离和共线性在预测数据诱导的分组时发挥着重要作用。

直线的斜率并未影响群组识别。相比之下，即使在指示强调 x轴为名义型的 Contextual 研究中，线性拟合误差仍表现出很高的贡献。这表明，数据诱导的分组

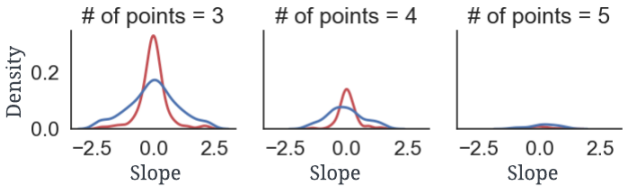


图 7. 按群组大小划分的线性拟合紧密（误差较低）时斜率的分布。Constellation 研究以蓝色表示，Contextual 研究以红色表示。误差越低，线性拟合越紧密。当一个组的线性拟合较紧密（误差较低，左图）时，Contextual 研究的分布集中在零附近，范围较窄。这意味着，与 Constellation 研究相比，其斜率接近水平。

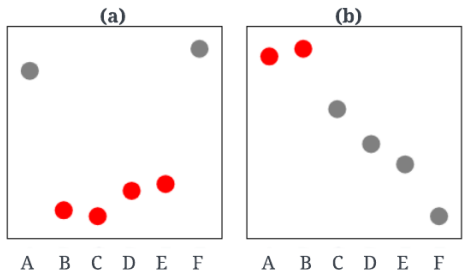


图 8. 低斜率示例。 选定的组 g 为红色，其余的 数据点 r 为灰色。g 几乎是水平的，并且与 r相距很远，因此可以将其理解为一个簇，而不是一条直线。

可能会压过用户对图表已有的背景知识。由于斜率在我们的实验中没有任何影响，因此在本文余下部分，我们将术语共线性 与其误差特征视为同义词。

比较单维特征时，我们发现，在两项研究中，用户都不只是依靠 x轴分离来识别组。同时，在 Constellation 研究中，y 轴分离的贡献几乎为零；但在 Contextual 研究中，其贡献增加到 $0.1-0.13$。这一贡献几乎与共线性和聚类特征的贡献一样高，这表明用户试图忽略 x轴的影响，转而关注沿 y 轴的分离。

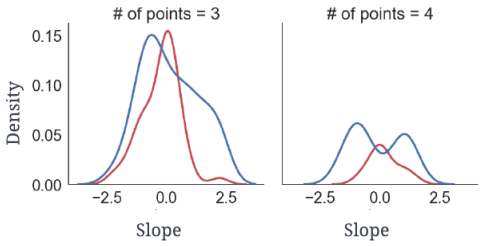


图 9. 仅聚类模型误判的各组斜率分布。 Constellation 研究以蓝色表示，而 Contextual 研究以红色表示。误差越小，线性拟合越紧密。请注意，群组大小为 5 时没有斜率分布，因为这些组没有被仅聚类模型误判。

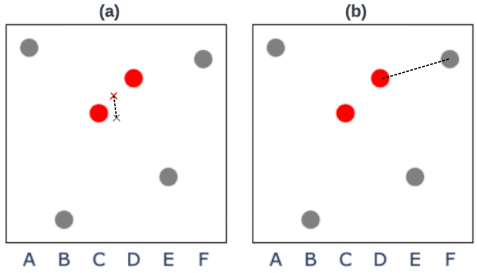


图 10. Contextual 研究中的假阴性组示例。所选组 g 以红色表示，而图表中的其余数据点r 以灰色表示。（a）g 位于由r形成的凸包内，且两者质心之间的距离（x标记）非常小。（b）不计算质心比率，而是计算g与r中的数据点之间的最小欧几里得距离，并以虚线表示。

2）评估假设 *1*：假设 1 探讨的是，与 Constellation 条件相比，用户在 Contextual 条件下是否较少依赖共线性。乍看之下（表 III），SHAP 结果似乎表明并非如此：在 Contextual 条件下，共线性的贡献高于 Constellation 条件，这说明用户在 Contextual 条件下似乎更加依赖共线性。

我们观察到，在表 III 中 Constellation 条件下的逻辑回归模型中，共线性竟然没有贡献，而凸包的贡献大致等于 Contextual 条件下误差与凸包贡献之和。这种差异不太可能是自然现象，因为从线性模型类切换到决策树模型类时，共线性通常会有较高的贡献。

这表明，共线性和凸包重叠编码了部分冗余信息。由于抽样变异，决策树恰好更依赖凸包重叠，进而使 SHAP 归因产生偏差。我们通过为每项研究训练两个决策树模型对此进行了调查：一个仅使用共线性特征，另一个仅使用聚类特征。我们发现，使用这两种模型进行预测时，Constellation 研究中 77% 的测试样本和 Contextual 研究中 81% 的测试样本都得到了准确预测。两种模型预测结果的高度重叠表明，在使用两组特征进行训练时，模型可能会选择其中任意一组。

我们进一步绘制了所有线性拟合误差较低的组的斜率分布，并按组的大小分面（图 7）。我们排除了两点组，因为任意一对数据点都显然可以构成一条直线。一个鲜明的模式立即显现：Contextual 条件下近水平斜率的分布密度明显高于 Constellation 条件。这些组大多由同时能被仅共线性模型和仅聚类模型准确预测的集合构成。图 8 清楚地展示了这一模式，其中绘制了两个斜率较低的组作为示例。这两个组都可以合理地归类为簇，而非趋势，因为它们沿 y 轴与其余数据点分离（y 轴分离程度较高）。

为了验证这一假设，我们采用了模型级联：首先将训练样本输入一个仅使用聚类特征的决策树模型。只有预测错误的样本才会被输入到另一个基于共线性特征训练的决策树模型中。这种方法使我们能够将共线性的影响与聚类的影响区分开来。

表 IV 模型级联第二阶段的评估。决策树仅使用共线性特征，并应用于被仅聚类模型错误分类的样本。

	P	R	F1	支持度
Constellation	0.91	0.92	0.91	151
上下文	0.81	0.89	0.85	112

图 9 重新绘制了仅针对被仅聚类模型错误预测的组的斜率分布，并按群组大小分面显示。所有大小为 5 的组均被仅聚类模型正确预测，因此我们没有绘制其分布。我们观察到，Contextual 条件下的峰值消失了，其分布变得与 Constellation 条件下的分布相当或更低。这表明，大多数斜率接近水平的共线组都被识别为簇。

此外，Constellation 研究的斜率分布宽得多：Constellation 研究中有一个较大的 3 点组子集，其斜率为 $0.25 - 2.5$；而 4 点组呈现双峰分布，其中大多数组的斜率为 -1 或 1。相比之下，Contextual 条件的分布仍以 0 为中心，并呈现陡峭的下降。

最后，表 IV 报告了第二个模型仅应用于聚类特征无法预测的测试样本时的准确率。F1 分数越高，表示共线性对非聚类组的预测能力越强。我们发现，Constellation 条件下的 F1 分数（0.91）远高于 Contextual 条件下的 F1 分数（0.85），这表明当一组数据点并非明显的聚类时，用户对共线性的依赖程度较低。需要注意的是，使用 F1 分数进行推断时，应将其视为反映总体趋势的指标，而非经过统计验证的证据。

基于这项分析，我们认为，尽管用户在 Contextual 条件下仍在很大程度上依赖共线性，但有较弱的证据表明，他们对共线性的依赖程度略低。

3）评估假设 *2*：假设 2 探讨用户是否倾向于更多依赖 y 轴分离，而非基于 Constellation 的分析。为检验这一假设，我们

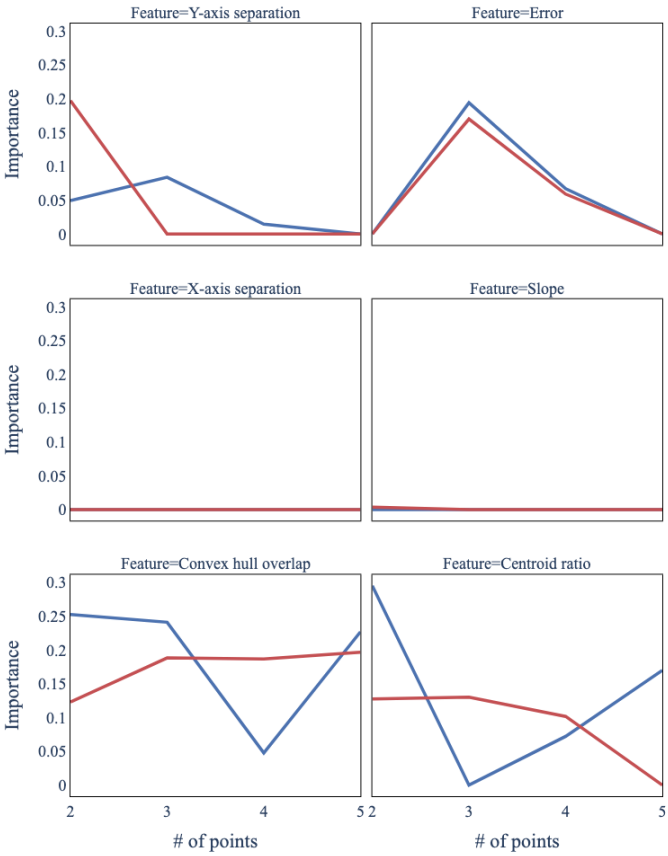


图 11. 决策树模型上的组大小与 SHAP 特征重要性。 Constellation 研究以蓝色表示， 而Contextual 研究以红色表示。y 轴显示 SHAP 绝对值的平均值，表示目标特征对模型预测的边际贡献。

分别训练了仅使用 y 轴分离作为输入特征的决策树模型。F1 分数越高，说明仅凭 y 值越能预测一组数据点是否会被视为一个组。我们发现，从 Constellation 研究切换到上下文研究后，留出数据集上的 F1 分数从 83% 提高到了 92%。这支持了我们的假设。

E. 进一步的模型改进

先前的研究表明，组大小可能在模型中发挥重要作用。现在，我们分析模型对组大小的敏感度，并处理一个与质心比率定义相关的失败案例。我们的结果将模型在留出集上的 F1 分数提高至 $0.97-0.99$。

1) 群组大小： 我们首先针对每种群组大小训练了一个决策树模型，并绘制了各项特征重要性随群组大小变化的图表（图 11）。有趣的是，对小群组（例如，y 轴分离）很重要的特征，对较大的群组却无关紧要；而共线性对最小（2）和最大（5）群组大小无关紧要，但对中间群组大小则变得重要。同时，簇分离特征在最小和最大群组大小时很重要，但在中间群组大小时相对不那么重要。斜率和 x 轴分离度在所有情况下都无关紧要。

表 V 各研究留出集上的最终决策树（DT）性能。 → 左侧的分数来自初始模型；→ 右侧的分数来自最终模型。所有情况下的精确率均为 1.00，因为留出集不包含假阳性。

	P	R	F1
Constellation	1.00 → 1.00	0.95 → 0.97	0.97 → 0.99
上下文	1.00 → 1.00	0.92 → 0.94	0.96 → 0.97

总体而言，“边缘”群组（群组大小最小和最大）的特征重要性始终不同于“中间”群组（包含 3 个和 4 个数据点的群组）。例如，对于任意两个构成一条线的数据点，共线性在语义上并无意义；而一个包含 5 个数据点的群组涵盖了图表中除一个数据点外的所有数据点，因此它与剩余数据点之间的簇分离更为显著。基于这些发现，我们训练了两个独立模型：一个用于边缘群组，另一个用于中间群组。在推断过程中，我们根据群组大小选择相应的模型。

2) 质心比率： 为了深入探究导致边缘群组召回率较低的因素，我们通过目视检查测试数据集中的所有假阴性样本进行了误差分析。通过这项分析，我们发现了当前特征无法捕捉的一种模式。具体而言，如果组 g 中的数据点完全被其余数据点r包围，模型就难以将其识别为一个组。图 10展示了这一情况，其中组g 中的数据点以红色标出。

我们发现，这种情况会使质心比率特征发生足够大的偏差，从而导致误分类。具体而言，g 和r 的质心非常接近，而由于质心比率的分子是它们之间的质心距离，因此该比率最终接近于零。

为了解决这一问题，我们修改了质心比率，将分子替换为g 中的数据点与r中的数据点之间的最小距离。当g 和r 互不重叠时，这一度量与质心距离一致；而当g 包含在r中时，该度量仍具有意义。我们在图10(b)中展示了这一情况。

3) 综述： 这些方法的结合带来了近乎完美的模型。表 V 总结了最终的模型质量，其中 Constellation 和 Contextual 研究的总体 F1 分数分别为 0.99 和 0.97。对于 Contextual 研究，性能提升主要源于将边缘条件与中间条件分开。通过这种方式，边缘（中间）模型从 0.93 提升至 0.95（0.99）。原始 Constellation 模型的 F1 分数已经达到很高的 0.98，因此边缘（中间）模型仅有小幅提升，增至 0.98（0.99）。

IV. 潜在应用

除了理论应用之外，我们还展示了可视化设计如何将诊断工具和重新设计工具纳入工作流程，从识别有问题的分组到评估不同的布局。尽管当x轴为名义型时，数据诱导的分组往往是非预期的产物（Kindlemann等人称之为幻觉

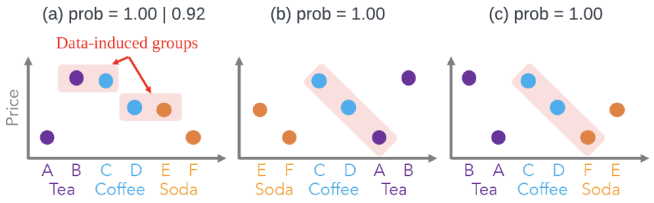


图 12. 图表展示了不同饮料类别中各项商品的价格。图表 (a) 显示了跨越类别边界的数据诱导分组，以及训练后的模型对这些分组给出的预测概率。简单地重新排列类别 (b) 或每个类别中的商品 (c)，可能会破坏这些分组，但也可能诱导出新的分组（例如，共线性）。

等人。 [13])，但它们本身并无好坏之分。希望突出一组期望分组的设计师，可以重新排列名义轴，使这些期望分组在视觉上更加醒目。同样，不理想的数据诱导分组也可能需要通过重新排列x轴来“打破”。请看以下示例：

示例 1. 图*12*展示了不同饮料类别中各项商品的价格，其中饮料类别构成了一种层级结构，自然地对底层数据进行分组。设计师希望设计一张图表，尽量减少跨饮料层级的比较。

在实际工作流程中，创建定价仪表板的设计师可以使用诊断工具，快速检查是否存在容易造成误导的跨类别分组，然后再定稿并交给利益相关者。例如，左侧图表标出了两个跨类别的数据诱导分组：*BC* 将茶类和咖啡类商品分为一组，而 *DE* 将咖啡类和苏打饮料类商品分为一组。

设计师可能会尝试通过交换茶和苏打饮料类别的位置（中间图），或者在各自类别内交换 *B* 与 *A*、*F* 与 *E* 的位置（右侧图），来解决这些违规项。尽管这些调整消除了上述违规项，却因共线性而诱导出新的跨类别分组。若能看到这些新组，设计师便能更好地进行权衡。此外，重新设计工具可以自动评估 *x*轴的不同排列，并推荐能够尽量减少违规项且遵循层级约束的排列。

本节展示预测模型如何帮助设计师实现这些目标。我们首先介绍一种诊断工具，当可视化容易产生数据诱导的分组（例如基于共线性的分组）时，该工具会向设计师发出提醒。随后，我们介绍一种自动重新设计工具，该工具在名义顺序重排的空间中进行搜索，以突出期望分组并减少非期望分组的出现。

A. 诊断应用程序

图 13（1-3）展示了诊断应用程序的设计。主界面垂直分为两部分，左侧（1）显示原始图表（请注意，这与图 12 中的饮料数据集相同），并圈出了期望分组。右侧边栏（2）显示预测模型识别出的数据诱导分组图库。每个组都以单独的图表呈现，组内的数据点以红色（不期望的组，也称为违规项）或蓝色（期望的组）突出显示。图表按群组大小排列

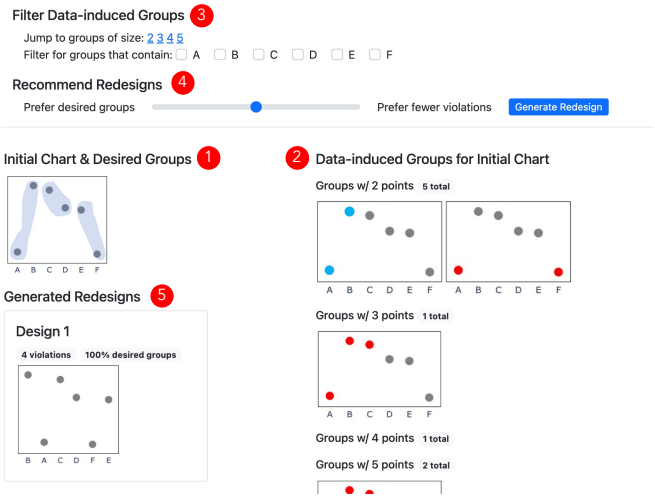


图 13. 诊断工具（1–3）和重新设计工具（4–5）的界面。参与者在初始图表（1）中绘制分组，系统则列出数据诱导分组（2），其中期望分组以蓝色显示，违规项以红色显示。参与者可以筛选分组（3），或请求系统提出重新设计方案（4、5）。

，每种大小的图表都以密集网格的形式展示，便于快速浏览。界面顶部（3）设有控件，可跳转到特定的群组大小。用户还可以筛选组，只显示包含特定数据点集合的组。

为识别数据诱导分组，我们开发了一个过程 $Diagnose(P, G_D) \rightarrow G_O$。该过程的输入为$P$，即图表中渲染的数据点的（$x$,$y$）像素坐标集合，以及$G_D \subseteq P$，即期望分组的集合。该过程输出用户可能感知到的分组集合$G_O \subseteq P$。每个元素$g_o \in G_O$ 都会标注模型估计的其不确定性$g_o.prob$，以及它是否属于违规项$g_o.violation$，具体定义如下。

我们使用上下文模型评估每个包含两个或更多数据点的子组 $g \subseteq P$。如果 g 不是期望分组，且模型以不确定性 ≥ 0.9 [1] 将其标记为正例，我们就将其保留为候选项。我们发现，仅此一项就会产生大量冗余的共线组。例如，考虑一条包含四个数据点的直线。直线上的所有子集也都是共线的，模型会将它们标记为正例。因此，如果 g 与另一个共线组相比是真子集，我们就不将其视为数据诱导分组。

示例 2. 图 *13* 使用图 *12*（左）中的数据点作为输入，并将三个类别指定为期望分组 $G_D = \{AB, CD, EF\}$。模型识别出 *9* 个用户可能感知到的数据诱导分组，其中只有 *2* 个期望分组包含在内。用户可以筛选包含“*A*”的组，侧边栏会显示相应的 *4* 个组。包含蓝色数据点的组对应期望分组，而红色数据点对应的组则是潜在违规项。用户随后认为违规项过多，计划更改 *x* 轴排列。

B. 重新设计应用程序

可视化图表阅读者一旦形成视觉分组，就很难再打破这些分组并形成

[1]我们通过使用训练数据进行超参数调优来选择此阈值。

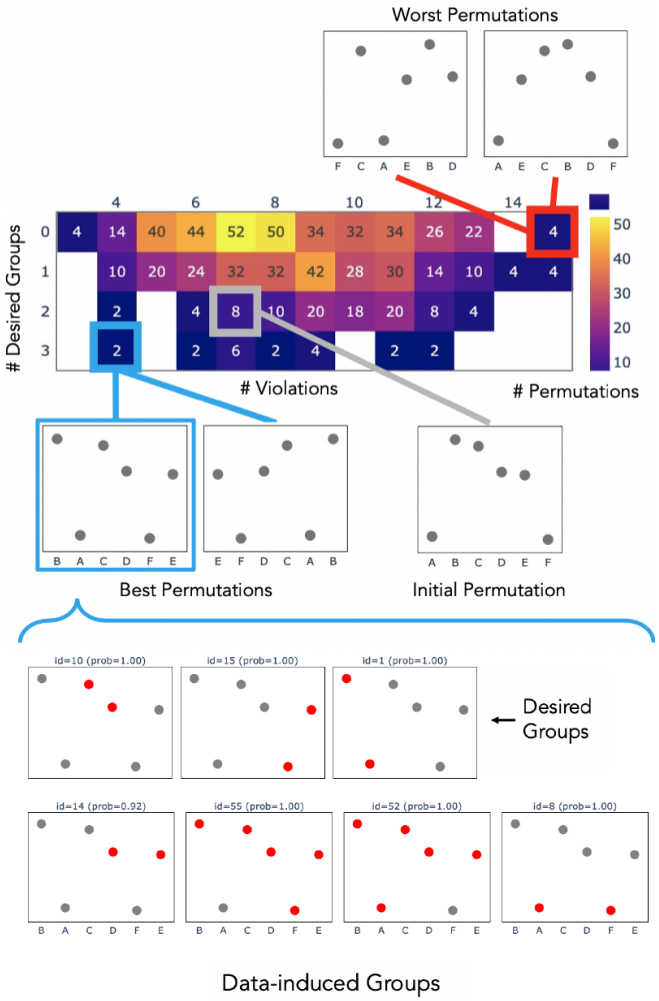


图 14。图 12 中的图表共有 720 种可能的 x 轴排列。热力图根据用户可能看到的违规项数量（x轴）和期望分组数量（y 轴）展示排列的分布。每个单元格都标注了排列数量。我们以蓝色突出显示最优设计的两个示例 optimal design in blue （所有期望分组，4 个违规项）、最差设计（红色）（没有期望分组，15 个违规项），以及灰色的初始排列（2 个期望分组，7 个违规项）。底部列出了突出显示的最优设计中的所有数据诱导分组，以及训练后的模型对这些分组的预测概率。

新的分组 [14]。这可能导致数据误读。因此，设计师能够轻松重新设计可视化，以减少违反预期的分组，这一点至关重要。为此，我们扩展了该应用，使其能够通过重新排列 x轴来提出自动重新设计方案，从而突出期望分组，同时打破非预期分组（图 13 (4-5)）。设计师在主图表 (1) 中绘制出期望分组后，可以使用滑块指定优先突出期望分组，还是尽量减少非期望分组 (4)。随后，他们点击“生成重新设计”按钮，系统便会求解一个优化问题，并推荐排名靠前的排列 (5)。开发者可以点击任意推荐排列，以更新侧边栏中的数据诱导的分组列表。

优化定义如下。对于给定的排列 $P_i \in$排列(P)，我们首先计算 $G_O =$诊断$(P_i,\ G_D)$，然后将两个因素的加权和作为其得分 s_i 。$s_d(P_i)$计算 G_D 中期望分组的总概率，这些分组已在 G_O中识别出来。然后，$s_v(P_i)$会对G_O 中属于违规项的组数进行惩罚：

$$s(P_i) = \alpha s_d(P_i) - (1-\alpha)s_v(P_i) \qquad \alpha \in [0,1]$$
$$s_d(P_i) = \sum_{g_o \in G_D \wedge g_o \in G_O} g_o.prob$$
$$s_v(P_i) = |\{g_o \in g_O | g_o.violation\}|$$

此外，设计师还可以将搜索空间限制为自定义的排列集合。例如，数据可能包含应在任何推荐排列中保留的层级结构H ： 图 12 按饮料类别呈现产品，设计师希望同一类别中的饮料在排列中保持相邻。因此，交换 B（茶）和 C（咖啡）等操作会违反层级约束。在这种情况下，只能在类别内部重新排列数据点（例如交换 A 和 B），或重新排列整个类别（例如交换茶和咖啡）。具体做法是提供一个方法 *isValid*（P），对于有效排列，该方法必须返回 true。

形式上，我们通过穷举x轴的所有排列来求解以下问题。

问题1 (重新设计问题)。 给定数据点 P 和期望分组 G_D，返回 $P^* = \text{argmax}_{P_i \in \text{排列}(P)} s(P_i)$，其中 $isValid(P^*)$。

示例3。 设计师希望系统推荐一种重新设计方案，并使用图 *13 (4)* 中的滑块指定他们偏好的排列，即确保能够感知到 $G_D = \{AB、CD、EF\}$ 中的所有期望分组。随后，用户点击“生成重新设计”。系统会评估所有 $6! = 720$个排列。最优排列列于 *(5)*，用户可以逐一点击，查看侧边栏中的数据诱导分组列表。在此情形下，最优排列仅有 *4* 项违规（而非 *7* 项），并确保所有期望分组都能被观察到。这个场景展示了实践者如何反复使用该工具来探索权衡，明确原本隐含的决策：哪些视觉偏差可以容忍。

为更好地了解优化景观，图14按违规项数量（x轴）和满足的期望分组数量（y 轴）总结了所有排列的分布。值得注意的是，没有任何排列能够在不产生其他数据诱导分组的情况下保证三个期望分组（左下角）。只有两种排列能够保证所有期望分组，同时出现4项违规。相比之下，356种排列（接近排列总数的50%）无法满足任何期望分组，最坏情况下还会出现多达15项违规。

该灰色的初始排列ABCDEF 仅展示了两个期望分组，并且存在 7 项违规。两个蓝色的最优排列互为镜像，我们在其下方紧接着列出高亮图表中的所有数据诱导分组。我们还列出了两个红色的最差排列示例，粗略观察便可发现一些难以忽视的虚假模式（例如抛物线形状）。归根结底，庞大的搜索空间以及劣质排列的普遍存在，促使我们将自动化重新设计与诊断工具纳入设计师的工具箱，也证明了这样做的合理性。

V. 讨论、局限性与未来方向

可视化能够增强数据解读，但受知觉系统局限的影响，也可能导致错误的假设 [3], [76]。例如，Correll 等人证明，当参与者看到一张展示某个数值在 12个月内如何变化的折线图，并被要求找出平均值最高的月份时，他们凭直觉选择了数值最高的月份，而不是平均值最高的平均值 [77]。当直觉知觉与准确的数据解读相冲突时，就会出现这种差异。为此，研究人员基于实证研究制定了设计指南。我们的工作以此为基础，研究点图中的错觉群组，并提出改进可视化设计的自动化方案。从可用性的角度来看，我们的工作流示例展示了这些工具如何融入常见的可视化任务，使设计师能够在创建图表时预判分组效果，并通过迭代不断加以调整。

讨论。 可视化图表阅读者会依据经典的格式塔原则——邻近性、相似性、良好延续性与共线性、共同区域以及连通性——迅速分割并归组标记，从而理解可视化。这一过程将可视化组织成可供比较的感知单元 [78]–[83]。这种组织大多在早期自动完成，并由前注意机制引导，将注意力指向结构化单元 [84], [85]。这些初始分组随后成为注意锚点：标记一旦绑定为单元，注意与比较便会围绕这些单元进行，因此，拆散已经形成的组以创建新组可能极其费力 [14], [86], [87]。在此基础上，我们的模型将与分组相关的特征——空间邻近性、群组大小、共线性和共同区域——操作化，以预测哪些数据点可能被归为一组。我们展示了以可能无意中凸显错觉群组的方式排列数据点所带来的风险，并提出重新设计建议，以促进有意义的视觉分组。

目前，可视化设计会将数据属性映射到视觉特征，但表格内容很少会影响这一过程。我们的研究表明，仅凭可视化设计不足以理解用户会如何感知可视化中的模式。相反，数据值和分布发挥着至关重要的作用，有时甚至会压过用户对数据的逻辑理解（例如，对数据类型的理解）。这意味着，尽管可视化设计V_1 总体上可能比V_2, 更有效，但对于特定数据集而言，后者可能更合适。换言之，数据集的属性可能使得可视化设计必须彻底改变。

局限性与未来工作。 首先，我们的模型会分别预测每一组数据点的情况，并在某些聚类和分离特征中隐式考虑组外的数据点。然而，可视化中的数据点除了聚类和分离之外，还会彼此交互，从而对知觉产生叠加影响。例如，感知到的数据形状（如对称性和正交性）会影响知觉和记忆 [88], ，而折线图中的极值可能影响人们感知到的整体线形，并扭曲读者对该线条平均位置的感知 [89]。未来研究的一个重要方向，是在预测模型中纳入数据组与数据点之间更多的感知交互。此外，我们的特征定义是针对点图构建的，并假定类别轴可以重新排序；将类似特征扩展到其他图表类型（例如条形图和折线图），并支持固定顺序，则留待未来工作。

其次，我们的模型是在用户识别的组和合成生成的负例上训练的。生成过程采用一种简单的启发式方法，识别那些“显然”不属于任何组的数据点子集。然而，这种启发式方法可能会给训练数据集带来偏差。未来的工作可以探索其他启发式方法，并招募更多用户来识别负例、创建更大的数据集。此外，空间能力等个体差异可能会影响数据模式检测，未来也可对此展开研究 [90]。此外，当前工作采用了预测建模的方法。未来的工作可以通过定性反馈（例如出声思维研究）来识别人们的分组倾向，并系统地检验和验证人们使用特定分组策略的条件。例如，研究人员可以利用大语言模型，在各种分组行为假设下生成可视化 [91] ，并评估人们对这些可视化的反应 [92]。

第三，我们的工作聚焦于恰有 6 个数据点的点图。虽然我们的模型旨在不受组内数据点数量的影响，而且我们认为研究结果具有普遍性，但仍需开展更多工作，以评估包含更多标记、其他数据类型、视觉编码和设计的情形。例如，我们所有的图表都将名义属性映射到 x 轴。已有研究表明，人们对水平面和垂直面的感知可能存在不对称性（例如，与一条水平线长度相同的垂直线会被认为更长 [93]），因此未来的工作可以检验我们的发现是否适用于名义 y 轴或多维平面的情形。此外，未来研究可以在用户研究中使用超过六个数据点来构建训练数据集，以检验研究结果的普遍性。另外，我们的用户研究明确要求参与者寻找分组，而这可能只反映了有限范围的现实世界中的可视化任务。尽管这种设定使我们能够在受控环境下分离并考察分组行为，未来的工作仍应研究当用户执行更开放式或领域特定任务时，数据诱导的分组会如何产生。

此外，我们的研究凸显了当前可视化研究中的一个空白：在非常规图表（例如没有明确网格线的点图）中，空间配置与用户解读之间的交集。未来的研究可以探讨类似的分组行为是否也会出现在更复杂的可视化中，例如包含重叠类别值的条形图；或者出现在用户未被明确要求识别分组的任务中。此外，将视觉可分性指标中的概念（例如 [94],提出的指标）与我们的预测建模框架相结合，或许能提供一种更全面的方法，以理解并缓解感知偏差。

VI. 结论

我们开展了两项用户研究，以了解数据诱导的分组的影响：这种知觉分组依据底层数据集中的数值，而非视觉规范。第一项研究要求用户识别点图中的分组，第二项则提醒用户数据属于名义型。这些研究揭示了两个关键发现：（1）共线性是视觉分组的重要决定因素，往往会凌驾于有关数据名义型属性的明确指示之上；（2）当共线性线索较弱或不存在时，邻近性和基于质心的分组等聚类特征占主导地位。我们构建了准确的模型，并展示了这些模型如何有望用于辅助可视化设计。

致谢

本研究部分由美国国家科学基金会资助，项目编号为 IIS-2237585、IIS-2311575、III-2453462、IIS-2107490、IIS-1901485、1845638、1740305、2008295、2106197、2103794 和 2312991；此外还获得了 Amazon、Google、Adobe、人工智能与技术中心（CAIT）、Columbia SIRS 和 Intellect Design 的支持。

参考文献

[1] J. Mackinlay，《关系信息图形表示的自动化设计》，《*ACM*图形学汇刊》，第5卷，第2期，第110–141页，1986年。[2] Y. Kim 和 J. Heer，《评估任务和数据分布对视觉编码有效性的影响》，载于《计算机图形学论坛》，第37卷，第3期。Wiley在线图书馆，2018年，第157–167页。[3] S. Franconeri、L. Padilla、P. Shah、J. Zacks 和 J. Hullman，《视觉数据传播的科学：哪些方法有效》，《公共利益中的心理科学》，第22卷，第3期，第110–161页，2021年。[4] G. J. Quadri 和 P. Rosen，《按任务划分的基于知觉的可视化研究综述》，*IEEE*可视化与计算机图形学汇刊，第28卷，第12期，第5026–5048页，2021年。[5] B. Saket、A. Endert 和 C. Demiralp，《基础可视化的任务有效性》，*IEEE*可视化与计算机图形学汇刊，第25卷，第7期，第2505–2512页，2018年。[6] F. Yang、L. T. Harrison、R. A. Rensink、S. L. Franconeri 和 R. Chang，《相关性判断与可视化特征：一项比较研究》，*IEEE*可视化与计算机图形学汇刊，第25卷，第3期，第1474–1488页，2018年。[7] L. A. Best、L. D. Smith 和 D. A. Stubbs，《线性与非线性趋势的知觉：利用斜率和曲率信息辨别趋势》，《知觉与运动技能》，第104卷，第3期，第707–721页，2007年。[8] C. Xiong、J. Shapiro、J. Hullman 和 S. Franconeri，《可视化数据中的因果关系错觉》，*IEEE*可视化与计算机图形学汇刊，第26卷，第1期，第853–862页，2019年。[9] J. Zacks 和 B. Tversky，《条形与折线：图形传播研究》，《记忆与认知》，第27卷，第6期，第1073–1079页，1999年。[10] G. E. Newman 和 B. J. Scholl，《描绘平均值的条形图在知觉上会被误读：柱内偏差》，《心理学通报与评论》，第19卷，第4期，第601–607页，2012年。[11] S. H. Kerns 和 J. B. Wilmer，《两张图表走进酒吧：基于读数的测量揭示了柱顶极限误差——对均值柱状图的一种常见、类别性误读》，《视觉学杂志》，第21卷，第12期，第17–17页，2021年。[12] E. Holder 和 C. Xiong，《离散程度与差异：可视化社会结果时，隐藏变异性可能助长刻板印象化》，*IEEE*可视化与计算机图形学汇刊，第29卷，第1期，第624–634页，2022年。[13] G. Kindlmann 和 C. Scheidegger，《可视化设计的代数过程》，*IEEE*可视化与计算机图形学汇刊，第20卷，第12期，第2181–2190页，2014年。[14] C. X. Bearfield、C. Stokes、A. Lovett 和 S. Franconeri，《图表传达了什么？分组线索引导观众在条形图中进行比较并得出结论》，*IEEE*可视化与计算机图形学汇刊，2023年。

[15] A. Lovett 和 K. Forbus，《将视觉问题求解建模为类比推理》。《心理学评论》，第124卷，第1期，第60页，2017年。[16] R. Fygenson、S. Franconeri 和 E. Bertini，《标记的排列会影响其传达的信息：条形图中的排序、分区、间距和着色》，IEEE可视化与计算机图形学汇刊，2023年。[17] C. Xiong、V. Setlur、B. Bach、K. Lin、E. Koh 和 S. Franconeri，《条形图的视觉排列会影响观看者的比较结论》，IEEE可视化与计算机图形学汇刊，第28卷，第1期，第955–965页，2021年。[18] R. A. Rensink，《关于可视化科学的发展前景》，载于《以人为中心的可视化手册》。Springer，2013年，第147–175页。[19] M. Wertheimer，《知觉形式的组织法则》。1938年。[20] J. Wagemans、J. H. Elder、M. Kubovy、S. E. Palmer、M. A. Peterson、M. Singh 和 R. von der Heydt，《视觉知觉中的格式塔心理学百年：I. 知觉分组与图形—背景组织》。《心理学公报》，第138卷，第6期，第1172页，2012年。[21] S. E. Palmer，《知觉表征中的层级结构》，《认知心理学》，第9卷，第4期，第441–474页，1977年。[22] W. Prinzmetal 和 W. P. Banks，《良好延续性影响视觉检测》，《知觉与心理物理学》，第21卷，395，1977 5，389期，页码 – 。[23] M. Gleicher，《比较可视化的考量》，IEEE可视化与计算机图形学汇刊，第24卷，第1期，第413–423页，2017年。[24] I. Rae、F. Zhou、M. Bilsing 和 P. Bunge，《通过数字“田野”观察理解视觉调查模式》，载于《2022年 CHI 计算机系统中的人为因素会议论文集》，2022年，第1–16页。[25] L. Battle 和 J. Heer，《探索性视觉分析的特征：文献综述及 Tableau 中分析溯源的评估》，载于《计算机图形学论坛》，第 38 卷，第3期。Wiley在线图书馆，2019年，第145–159页。[26] D. Navon，《先见森林，再见树木：全局特征在视觉知觉中的优先性》，《认知心理学》，第9卷，383，1977 3，353期，页码 – 。[27] R. Kimchi，《整体加工与全局/局部范式的优先性：批判性综述》。《心理学公报》，第112卷，第1期，第24页，1992年。[28] M. Ahissar 和 S. Hochstein，《视觉知觉学习的反向层级理论》，《认知科学趋势》，第8卷，第10期，第457–464页，2004年。[29] E. Wu，《用于临时比较的视图组合代数》，IEEE可视化与计算机图形学汇刊，第28卷，第6期，第2470–2485页，2022年。[30] M. Hearst、M. Tory 和 V. Setlur，《面向视觉分析自然语言界面中模糊修饰语的界面默认设置》，载于《2019年 IEEE 可视化会议（VIS）》，2019年，第21–25页。[31] A. Gaba、V. Setlur、A. Srinivasan、J. Hoffswell 和 C. Xiong，《比较难题与可视化之室：探究语言如何影响视觉设计》，IEEE可视化与计算机图形学汇刊，第29卷，第1期，第1211–1221页，2022年。[32] M. Gleicher、D. Albers、R. Walker、I. Jusufi、C. D. Hansen 和 J. C. Roberts，《信息可视化中的视觉比较》，《信息可视化》，第10卷，第4期，第289–309页，2011年10月。 [在线]。可访问：https://doi.org/10 1177 1473871611416549 。 / A. Pister、C. Prieur 和 J.-D. Fekete，《Combinet：二部多变量动态社交网络的视觉查询与比较》，载于《计算机图形学论坛》，2023年。[34] L. Joos、S. Jaeger-Honz、F. Schreiber、D. A. Keim 和 K. Klein，《VR 中的网络视觉比较》，IEEE可视化与计算机图形学汇刊，第28卷，第11期，第3651–3661页，2022年。[35] S. Chen、N. Andrienko、G. Andrienko、J. Li 和 X. Yuan，《Co-bridges：多项目数据流的成对视觉连接与比较》，IEEE可视化与计算机图形学汇刊，第27卷，第2期，第1612–1622页，2020年。[36] S. L. Franconeri，《理解可视化数据的三种知觉工具》，《心理科学当前方向》，第30卷，第5期，第367–375页，2021年。[37] A. V. Pandey、J. Krause、C. Felix、J. Boy 和 E. Bertini，《探究人类在分析大量散点图时的相似性知觉》，载于《2016年 CHI 计算机系统中的人为因素会议论文集》，2016年，第3659–3669页。[38] S. L. Franconeri，《视觉资源的性质与现状》。2013年。[39] A. Larsen 和 C. Bundesen，《空间分离对视觉模式匹配的影响：心理转换作用的证据》，《实验心理学杂志：人类知觉与表现》，第24卷，第719–31页 , 07 1998年。

[40] W. S. Cleveland 和 R. McGill，《图形感知：理论、实验及其在图形方法开发中的应用》，《美国统计学会会刊》，第 79 卷，第 387 期，第 531–554 页，1984年。 [41] S. LYi、J. Jo 和 J. Seo，《比较布局再探：设计空间、指南与未来方向》，《IEEE可视化与计算机图形学汇刊》，第 27 卷，第 2 期，第 1525–1535 页，2020年。 [42] N. Jardine、B. D. Ondov、N. Elmqvist 和 S. Franconeri，《视觉比较的感知代理》，《IEEE可视化与计算机图形学汇刊》，第 26 卷，第 1 期，第 1012–1021 页，2019年。 [43] B. Ondov、N. Jardine、N. Elmqvist 和 S. Franconeri，《面对面：评估视觉比较》，《IEEE可视化与计算机图形学汇刊》，第 25 卷，第 1 期，第 861–871 页，2018年。 [44] C. Xiong、C. Stokes 和 S. Franconeri，《视觉数据展示中的视觉显著性与分组线索引导关系感知》，《视觉学杂志》，第 21 卷，第 9 期，第 2095–2095 页，2021年。 [45] S. Lee、S.-H. Kim 和 B. C. Kwon，《 Vlat：可视化素养评估测试的开发》，《IEEE可视化与计算机图形学汇刊》，第 23 卷，第 1 期，第 551–560 页，2016年。 [46] C. Nothelfer 和 S. Franconeri，《直接编码数据差值对数据对关系感知的益处之测量》，《IEEE可视化与计算机图形学汇刊》，第 26 卷，320，2019 1，311第。页 – 。 [47] C. M. McColeman、F. Yang、T. F. Brady 和 S. Franconeri，《重新思考视觉通道的等级》，《IEEE可视化与计算机图形学汇刊》，第 28 卷，第 1 期，第 707–717 页，2021年。 [48] P. Shah 和 E. G. Freedman，《条形图和折线图理解：自上而下过程与自下而上过程的交互》，《认知科学专题》，第 3 卷，第 3 期，第 560–578 页，2011年。 [49] A. L. Michal 和 S. L. Franconeri，《视觉程序与特定图表解读相关》，《认知研究：原理与启示》，第 2 卷，第 1 期，第 1–10 页，2017年。 [50] A. L. Michal、D. Uttal、P. Shah 和 S. L. Franconeri，《用于提取大小关系的视觉程序》，《心理学通报与评论》，第 23 卷，第 6 期，第 1802–1809 页，2016年。 [51] S. Pinker，《图表理解理论》，《人工智能与测试的未来》，第 73–126 页，1990年。 [52] B. C. Love、J. N. Rouder 和 E. J. Wisniewski，《整体与局部加工的结构性解释》，《认知心理学》，第 38 卷，第 2 期，第 291–316 页，1999年。 [53] M.-S. Kim 和 K. R. Cave，《分组对视觉搜索中空间注意的影响》，《普通心理学杂志》，第 126 卷，第 4 期，第 326–352 页，1999年。 [54] D. Yu、D. Tam 和 S. Franconeri，《相似性分组是序列进行的，不受间距或群组大小影响》，《视觉学杂志》，第 14 卷，第 10 期，第 808–808 页，2014年。 [55] D. Yu、X. Xiao、D. K. Bemis 和 S. L. Franconeri，《作为基于特征选择的相似性分组》，《心理科学》，第 30 卷，第 3 期，第 376–385 页，2019年。 [56] J. Duncan 和 G. W. Humphreys，《视觉搜索与刺激相似性》，《心理学评论》，第 96 卷，第 3 期，第 433 页，1989年。 [57] M. Gleicher、M. Correll、C. Nothelfer 和 S. Franconeri，《多类别散点图中的平均值知觉》，《IEEE可视化与计算机图形学汇刊》，第 19 卷，2325，2013 12，2316第。页 – 。 [58] D. Burlinson、K. Subramanian 和 P. Goolkasian，《开放形状与封闭形状：新的知觉类别？》，《IEEE可视化与计算机图形学汇刊》，第 24 卷，第 1 期，第 574–583 页，2017年。 [59] P. Shah 、 R. E. Mayer 和 M. Hegarty，《作为知识建构辅助工具的图表：引导图表理解过程的信号提示技术》，《教育心理学杂志》，第 91 卷，第 4 期，第 690 页，1999年。 [60] J. L. Brooks，《知觉分组的传统原则与新原则》，2015。 [61] S. L. Franconeri、D. K. Bemis 和 G. A. Alvarez，《数量估计依赖于一组分割对象》，《认知》，第 113 卷，第 1 期，第 1–13 页，2009年。 [62] D. Yu、D. Tam 和 S. L. Franconeri，《格式塔相似性分组并非并行构建》，《认知》，第 182 卷，第 8–13 页，2019年。 [63] M. Zhao、H. Qu 和 M. Sedlmair，《条形图中的邻近感知》，载于《2019年CHI计算机系统中的人为因素会议论文集》，2019年，第 1–12 页。 [64] M. I. Posner，《注意定向》，《实验心理学季刊》，第 32 卷，第 1 期，第 3–25 页，1980年。 [65] C. L. Folk 和 B. A. Anderson，《注意捕获中的目标不确定性效应：颜色单例集合还是多种注意控制设置？》，《心理学通报与评论》，第 17 卷，第 3 期，第 421–426 页，2010年。

[66] S. Palan 和 C. Schitter，“Prolific.ac——一个用于在线实验的被试库”，《行为与实验金融杂志》，第17卷，页码22–27，2018年。 [67] A. Treisman，“在视觉搜索特征和物体时的知觉分组与注意”，《实验心理学杂志：人类知觉与表现》，第8卷，第2期，第194页，1982年。 [68] J. Heer 和 M. Agrawala，“多尺度倾斜至45度”，IEEE可视化与计算机图形学汇刊，第12卷，第5期，页码701–708，2006年。 [69] Y. Wang、Z. Wang、L. Zhu、J. Zhang、C.-W. Fu、Z. Cheng、C. Tu 和 B. Chen，“有没有一种稳健的技术来选择折线图的纵横比？”，IEEE可视化与计算机图形学汇刊，第24卷，第12期，页码3096–3110，2017年。 [70] S. Yantis，“多元素视觉追踪：注意与知觉组织”，《认知心理学》，第24卷，第3期，页码295–340，1992年。 [71] R. Borgo、B. Lee、B. Bach、S. Fabrikant、R. Jianu、A. Kerren、S. Kobourov、F. McGee、L. Micallef、T. von Landesberger 等，“信息可视化众包：机遇与陷阱”，载于《群众中的评估：众包与以人为中心的实验：Dagstuhl研讨会15481，德国达格施图尔城堡，2015年11月22日至27日，修订版论文集》。Springer，2017年，页码96–138。 [72] R. Borgo、L. Micallef、B. Bach、F. McGee 和 B. Lee，“利用众包评估信息可视化”，载于《计算机图形学论坛》，第37卷，第3期。Wiley在线图书馆，2018年，页码573–595。 [73] R. Amar 和 J. Stasko，“用于信息可视化设计与评估的基于知识任务的框架”，载于IEEE信息可视化研讨会，2004年，页码143–150。 [74] B. Julesz，“不同知觉层级上的簇形成”，载于《模式识别方法论》。爱思唯尔，1969年，页码297–315。 [75] S. M. Lundberg 和 S.-I. Lee，“解释模型预测的统一方法”，《神经信息处理系统进展》，第30卷，2017年。 [76] C. R. Ceja、C. M. McColeman、C. Xiong 和 S. L. Franconeri，“《真相还是方形》：纵横比会影响对位置编码的回忆”，IEEE可视化与计算机图形学汇刊，第27卷，第2期，页码1054–1062，2020年。 [77] M. Correll、D. Albers、S. Franconeri 和 M. Gleicher，“比较时间序列数据中的平均值”，载于SIGCHI计算机系统中的人为因素会议论文集，2012年，页码1095–1104。 [78] M. Wertheimer，“知觉形式的组织法则”，载于《格式塔心理学资料集》，W. D. Ellis 编。伦敦：劳特利奇与基根·保罗出版社，1938年，页码71–88，原作发表于1923年。 [79] J. Wagemans、J. H. Elder、M. Kubovy、S. E. Palmer、M. A. Peterson、M. Singh 和 R. von der Heydt，“视觉知觉中的格式塔心理学百年：I. 知觉分组与图形—背景组织”，《心理学公报》，第138卷，第6期，页码1172–1217，2012年。 [80] J. Wagemans、J. Feldman、S. Gepshtein、R. Kimchi、J. R. Pomerantz、P. A. van der Helm 和 C. van Leeuwen，“视觉知觉中的格式塔心理学百年：II. 概念与理论基础”，《心理学公报》，第138卷，第6期，页码1218–1252，2012年。 [81] S. E. Palmer，“共同区域：知觉分组的一项新原则”，《认知心理学》，第24卷，第3期，页码436–447，1992年。 [82] S. E. Palmer 和 I. Rock，“重新思考知觉组织：均匀连通性的作用”，《心理学评论》，第101卷，第4期，页码725–753，1994年。 [83] D. J. Field、A. Hayes 和 R. F. Hess，“人类视觉系统的轮廓整合：局部‘关联场’的证据”，《视觉研究》，第33卷，第2期，页码173–193，1993年。 [84] A. M. Treisman 和 G. Gelade，“注意的特征整合理论”，《认知心理学》，第12卷，第1期，页码97–136，1980年。 [85] J. M. Wolfe 和 T. S. Horowitz，“哪些属性会引导视觉注意的部署？它们又是如何发挥作用的？”，《自然综述：神经科学》，第5卷，第6期，页码495–501，2004年。 [86] R. Egly、J. Driver 和 R. D. Rafal，“在物体与位置之间转移视觉注意：来自正常受试者和顶叶损伤受试者的证据”，《实验心理学杂志：总论》，第123卷，第2期，页码161–177，1994年。 [87] B. J. Scholl，“物体与注意：研究现状”，《认知》，第80卷，第1–2期，页码1–46，2001年。 [88] K. Marriott、H. Purchase、M. Wybrow 和 C. Goncu，“网络图中视觉特征的可记忆性”，IEEE可视化与计算机图形学汇刊，第18卷，第12期，页码2477–2485，2012年。 [89] T. Savalia、C. Ceja、R. Cowell 和 C. Xiong，“数据形状和响应方式可能导致对可视化中数据平均位置的估计产生偏差”，视觉学杂志，第22卷，第14期，页码3413–3413，2022年。

[90] T. M. Green 和 B. Fisher，“迈向交互的个人方程：人格因素对可视分析界面交互的影响”，载于*2010*年*IEEE*可视分析科学与技术研讨会。IEEE，2010年，页码203–210。[91] Y. Tian、W. Cui、D. Deng、X. Yi、Y. Yang、H. Zhang 和 Y. Wu，“ChartGPT：利用大语言模型从抽象自然语言生成图表”，*IEEE*可视化与计算机图形学汇刊，2024年。[92] H. W. Wang、J. Hoffswell、V. S. Bursztyn、C. X. Bearfield 等，“人类对图表的要点总结与大语言模型预测的吻合程度如何？一项关于不同布局条形图的案例研究”，*IEEE*可视化与计算机图形学汇刊，2024年。[93] J. O. Robinson，《视觉错觉心理学》。Courier Corp，1998。[94] M. Aupetit 和 M. Sedlmair，“Sepme：2002年新视觉分离度量”，载于*2016*年*IEEE*太平洋可视化研讨会（*PacificVis*）。IEEE，2016年，页码1–8。

七、人物简介

Yilan Jiang 获得了数据科学理学硕士学位，目前正在伊利诺伊大学厄巴纳-香槟分校攻读工业与企业系统工程博士学位。她的研究主要关注数据驱动的产品设计、优化以及机器学习的应用。

Cindy Xiong Bearfield 是佐治亚理工学院交互计算学院的助理教授。她获得了西北大学认知心理学博士学位和统计学理学硕士学位。她致力于打通心理学与数据可视化领域，探究视觉数据解读与传播背后的认知与感知过程。

Steven Franconeri 是西北大学心理学教授，也是西北大学认知科学项目主任。他研究心理学、教育学和信息可视化领域中的视觉思维与视觉传播。

Eugene Wu 是哥伦比亚大学副教授，也是数据、智能体与流程实验室（DAPLab）的联合主任。他的研究兴趣是人类与数据交互系统。他在数据系统、可视化、数据清洗与提取、机器学习、解释、众包和智能体系统等方面均有所贡献。