第13章 可视化效果评测与用户实验

挑战

- 用户评测实验，对可视化方法和技术的进一步应用提供有说服力的证据
- 用户实验对于可视化研究也至关重要
  - 可视化研究者需要比较新技术与已有技术的优劣
- 可视化方法、技术和系统的用户实验面临诸多挑战和难题。实证性研究所共有的
  - 如何定义研究目的和问题并选择适当的方法，如何设计实验，如何保证严谨的数据采集和分析过程
- 可视化技术的评测与人机交互技术的评测有诸多相通之处，特别是在交互界面的可用性研究方面
- 可视化评测的另一个关注点是视觉表达中采用的编码及其可读性
- 用户实验最终需要回答的问题是：可视化技术是否能够更好地帮助用户解读某些数据
  - 某些时候，由于用于评测的数据集太小、参与用户不是目标人群、实验任务设计不当等因素，用户实验并不能有效地回答研究所要解决的问题。这些挑战意味着要完成一个严谨有效的可视化用户实验并不是一件容易的事

1 评测流程

评测流程

- 明确研究目的并定义研究问题
  - 在进行评测之前，研究者首先需要明确的是评测的目的
  - 其次，研究者需要围绕研究目的进一步清晰地定义研究所要解决的具体问题
- 提出研究假设
  - 在给出研究假设的时候，应尽量避免使用宽泛的命题
  - 对于可视化技术来说，相对更好的命题是“用户在使用可视化系统甲时，能比使用可视化系统乙时更高效地对某类特定数据进行聚类分析”
- 设计研究方案和具体方法
  - 着手设计研究的具体方案并且选择合适的方法
- 收集和分析数据
  - 对参与的用户进行必要的指导，安排必要的练习，以及提供适当的反馈
  - 在比较多种技术或系统时，细节方面需尽量保持一致
  - 此外，现有技术已经能够很好地保证某些用户数据采集的实时性和客观性，比如任务的完成时间和正确率等，应当充分利用
  - 在分析数据时，重要的是保证针对不同类型的数据选择正确的方法
- 验证研究假设并得出结论
  - 得到实验结果之后，需要判断研究假设是否成立，或者是否有足够的证据来支持或推翻研究假设，进而得到研究的主要结论

2 评测方法

2 评测方法

- 2.1 用户实验（User Studies）
- 2.2 专家评估（Expert Review/Heuristic Evaluation）
- 2.3 案例研究（Case Studies and Use Cases）
- 2.4 指标评估（Metrics）
- 2.5 众包（Crowdsourcing）
- 2.6 标注（Labeling）

2.1 用户实验（User Studies）

- 用户实验提供了一种科学可靠的方法来评估可视化的效果。实现一个可视化系统的目的是为了辅助用户完成数据的理解、分析等任务
  - 最终实现的系统能否满足用户的需求？
  - 用户在使用时是否容易上手？
  - 完成任务的过程是否有了效率或准确率的提升？

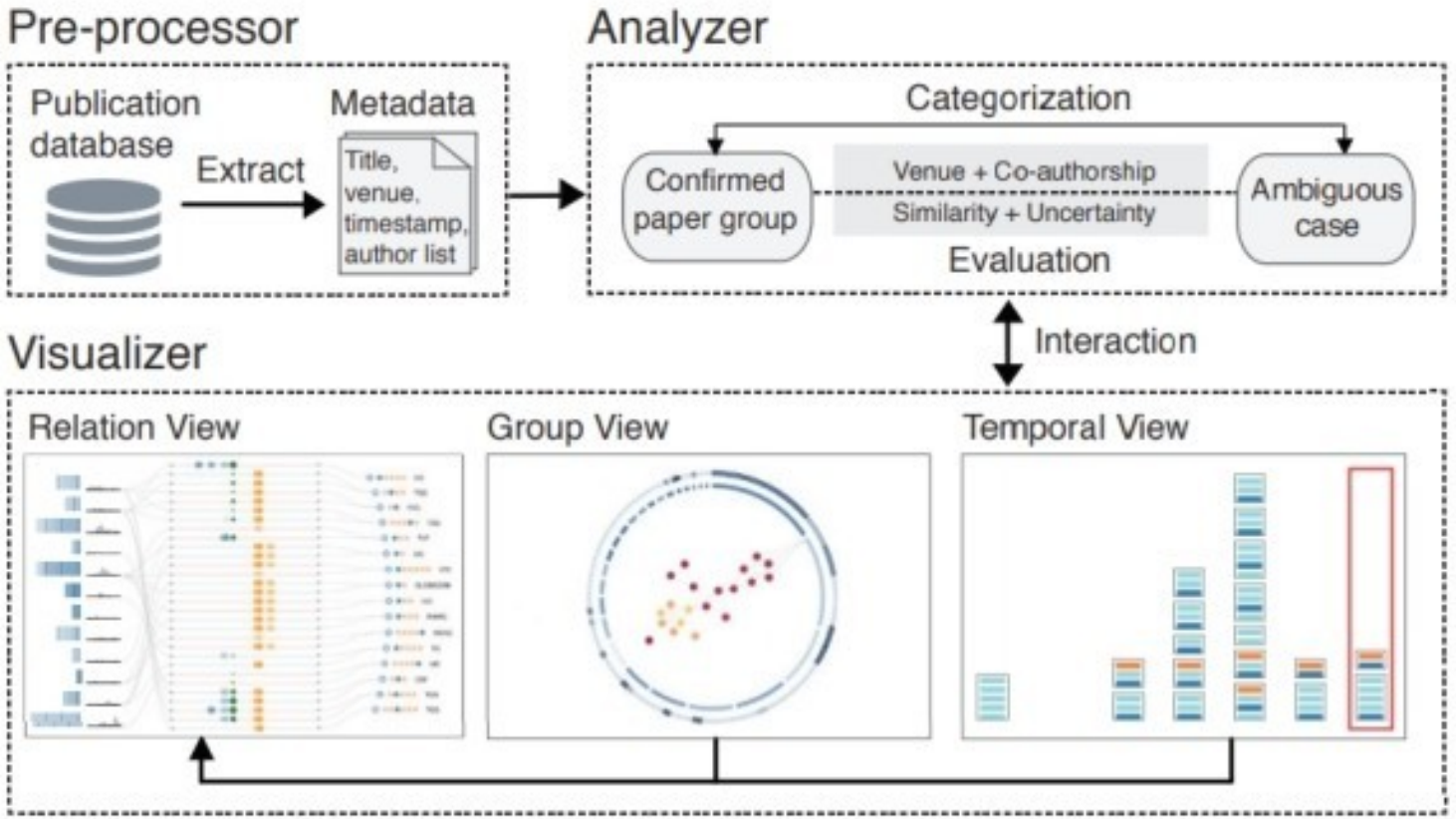
- 这一系列问题的答案都是评估一个可视化系统优劣的标准。通过收集用户使用数据来进行评估的方法被称为用户实验

2.2 专家评估（Expert Review/Heuristic Evaluation）

- 专家评估通常需要符合条件的专家级用户参与 [Tory2005]，从而避免了招募大量用户参与评测的麻烦
  - 这些评估者是领域的专家，他们对所使用的数据和需要完成的目标任务非常了解，能够对可视化技术在多大程度上适用于这样的数据和任务做出比较准确的判断。
  - 可视化技术评测的参与者也可以包含可视化专家，他们对可视化设计有丰富的知识，并具有可视化工具开发经验。可视化专家对可视化的有效性有自己的一套评判标准，并在评测中依据这些标准做出自己的判断

NameClarifier系统

- 通过交互解决出版物中作者名字歧义问题的可视分析系统
- 为了评估NameClarifier的有效性，研究者与两位负责维护大学出版物记录的专家分别进行了一对一的访谈

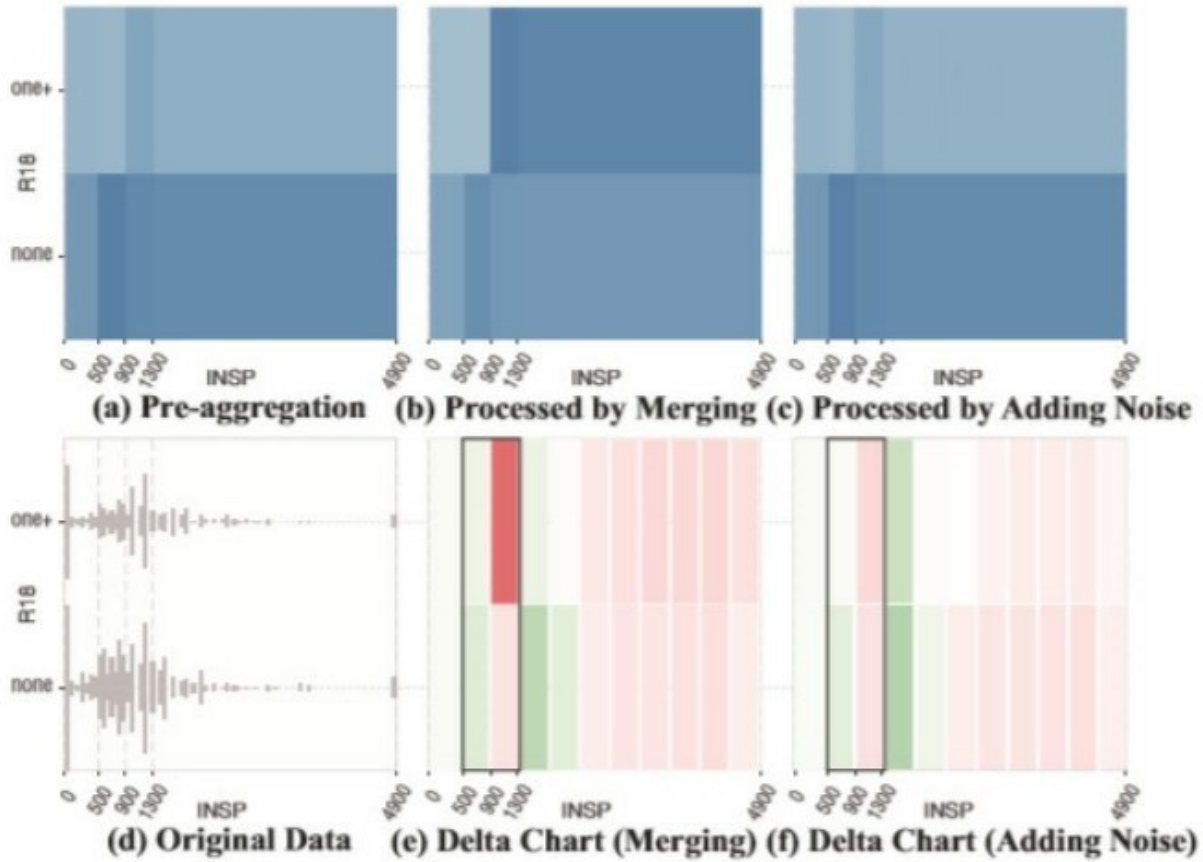


2.3 案例研究（Case Studies and Use Cases）

- 试图通过描述可视化技术和系统如何帮助解决一个现实的问题并完成目标任务来证明其有效性
  - 这样的案例研究关键在于，案例必须是真实的和有切实需求的。这样才能对有类似需求的用户具有说服力，使他们有信心尝试使用该技术去解决实际问题

隐私保护方法

- 通过可视分析方法实现了一种兼顾实用性的隐私保护方法，并通过一个案例来证明该方法的实用性



2.4 指标评估（Metrics）

- 对于可视化的子模块，如布局和交互等，可以通过一些指标来对它们的部分特性进行评估
  - 以图的布局算法为例，算法的时间复杂度、生成结果的易读性或美观程度（节点重叠量、边交叉数、临边最小夹角等）都可以被用来检验生成的结果

- 指标只能客观地从某个角度进行量化评估
  - 实际上，人的主观认知十分复杂，且具有多样性。对于喜好程度等依赖认知的评估条目来说，根据经验得出的一个或一组指标无法全面地模拟出主观认知的过程，因此也不能完全取代用户实验得出的真实实验结果

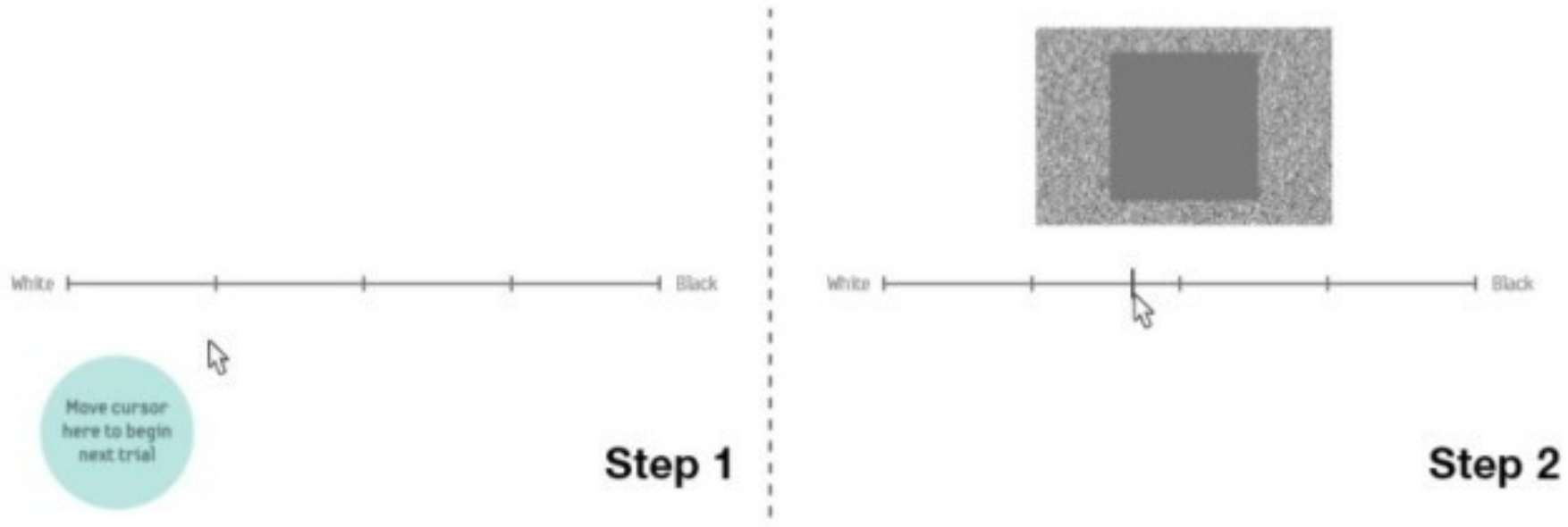
- 为了提高指标评估的效果，在评估时，可以通过允许用户交互地调节指标计算函数中的参数来定制精确的评估指标

2.5 众包（Crowdsourcing）

- 研究者设置一些条件来筛选参与用户，并将任务发布在网上
  - 必须考虑到参与用户使用的设备、实验环境的差异，提前准备好应对方法，加上必要的说明和详细的指导以顺利地得到数据

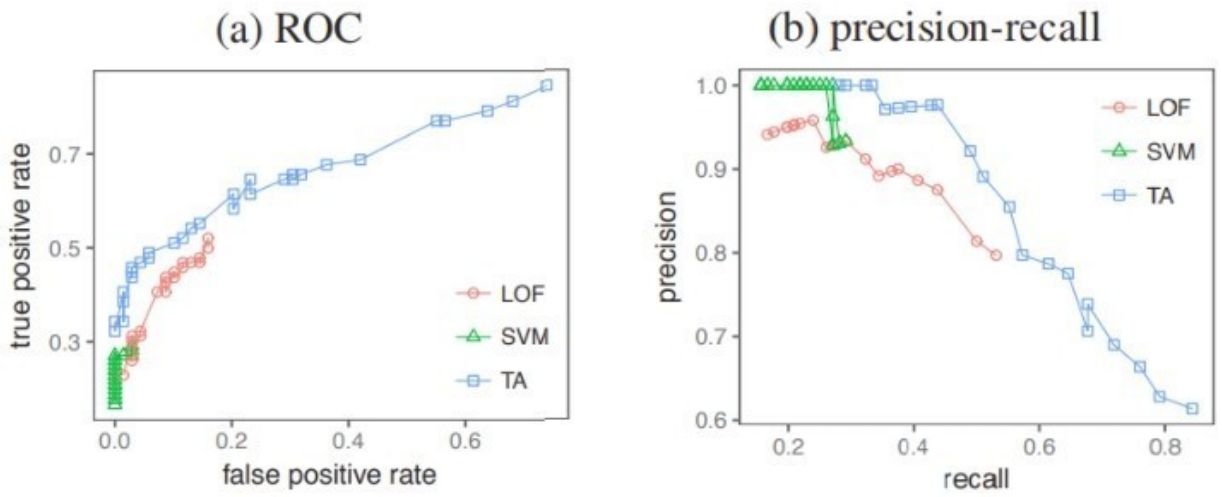
众包评测方法

- 研究调研对象的评分行为可能受到的影响



2.6 标注（Labeling）

- 评估结果的准确性需要基于标准答案。在一些情况下，标准答案来自人工标注的结果
- 在评估Voila时，研究者使用了人工标注来验证方法的有效性



3 用户试验

3 用户实验

- 3.1 确定实验目标
- 3.2 准备实验
- 3.3 进行实验
- 3.4 分析结果并讨论
- 3.5 评测案例分析

3.1 确定实验目标

- 用户实验的目标是探索未知内容、验证猜想、评估对象，或者对一组对象进行比较

- 探索未知内容
  - 在用户使用的过程中暴露缺点、发现亮点，是用户实验的重要作用
- 验证猜想
  - 猜想正确与否需要实践，也就是需要用户实验的结果来检验
- 评估对象
  - 从用户的角度证明实验对象的优劣
- 对一组对象进行比较
  - 从多种方案中选出最合适的方案时，可以依据用户实验的结果对它们进行比较

3.1.1 研究对象

- 研究对象可以是系统，也可以是具体的可视化方法或交互设计
- 如果实验目标为对比多个对象，那么研究者需要对相关工作进行调研，考虑不同的特点，尽可能全面地选择具有代表性的对象进行实验
  - 实验对象不是越多越好
  - 每个对象的样本数据不应过少，这就意味着对象数量的增加会增大对样本的需求，造成一定的负担

3.1.2 任务

- 定义合适的测试任务的前提是了解可视化技术所支持的用户任务，对测试任务的选择也决定了用户实验及其结论所适用的范围
- Keller等人在[Keller1993]中总结出了9大任务
  - 鉴定（Identify）
    - 基于可视化中显示出来的特性鉴别特定物体。例如，从CT医学影像中找到肿瘤
  - 定位（Locate）
    - 确定物体的位置。例如，在气象数据中找到风暴的中心和移动的路线
  - 区别（Distinguish）
    - 区分一个物体。例如，区分高度超过某个阈值的物体和其余物体
  - 分类（Categorize）
    - 对物体分类。例如，按照物体不同的材料质地或形状进行分类
  - 聚类（Cluster）
    - 将相似的物体按照彼此关系归类。例如，在社交网络中按照朋友关系将人群分成不同的社区。一个相似的操作是分割（Segmentation），也就是区分不同的物体
  - 排名（Rank）
    - 将一组物体按照一定的规则排序。例如，按照数值或时间顺序排列
  - 比较（Compare）
    - 查看两个或更多物体之间的相似之处和不同之处
  - 联系（Associate）
    - 表现两个或更多物体之间的关系。例如，通过气象数据可视化，将温度与地理位置联系起来
  - 关联（Correlate）
    - 找到两个或更多物体之间的因果或互动关系。例如，发现贷款利率与经济增长之间的关系

3.1.3 指标

- 需要确定判断任务完成的效率和准确率的指标
  - 如，是否需要准确地鉴定某个物体每一次的出现？在聚类或排序时，多大的误差是可以接受的？当定位地图上的一个物体时，需要精确到国家、城市还是经纬度坐标？所选择的测试任务和测量指标会影响研究结果的内部效度和外部效度
- 内部效度和外部效度是判断实证性研究有效性的基本指标 [Mark2012]
  - 内部效度指研究者实际测量的和想要测量的指标之间的贴切程度。二者越接近，研究的内部效度越高；二者差异越大，研究的内部效度越低。内部效度越低，研究的有效性也就越低
  - 外部效度指研究结论有效范围的大小。研究结论的适用范围越大，研究的外部效度越高；反之，则外部效度越低
  - 研究者应当在保证研究的内部效度的前提下，找到内部效度和外部效度之间的平衡点
- 在设计评测方案时，必须全面、严谨地考虑到各种可能影响评测效度的因素，如自变量和因变量的定义、目标用户和任务的选择、可视化技术所指向的数据及其特性和测量指标的选择

3.2 准备实验

- 3.2.1 数据
- 3.2.2 用户
- 3.2.3 实验设计

3.2.1 数据

- 可视化技术通常是针对某一类或者某些类数据而设计和实现的。数据类型和用于用户测试的数据大小往往会影响可视化技术的效果
  - 在理想情况下，可视化技术的用户测试中使用的数据应该首先适用于测试的可视化技术
  - 其次，数据应该具有代表性并且包含不同属性的数据集

数据的属性

- 数据类型（Type）
  - 了解数据属性对于特定的可视化技术的效果是否有影响
- 数据量（Size）
  - 测中使用的数据集应当包括常见大小的数据集以及某些极端尺寸的数据集
- 数据的维度（Dimensionality）
  - 评测中非常重要的一项是对高维度数据的可扩展性
- 数据的多元性（Number of Parameters）
  - 了解可视化技术能有效处理的最大变量数
- 数据结构（Structure）
  - 评测需要包括所适用的各种结构的数据
- 数据的范围（Range）
  - 要重点测试极值情况下可视化的性能
- 数据的分布（Distribution）
  - 数据值和数据属性（如时间和空间属性）
- 公共数据集
  - 加州大学埃尔文分校的机器学习数据集 [UCI2012]、卡耐基梅隆大学的StatLib数据集 [StatLib2012]、大规模网络数据集

3.2.2 用户

- 在设计一个可视化评测时，能准确地描述并选择目标用户是至关重要的。选择参与用户时需要考虑的主要因素
  - 对应用领域的熟悉程度
    - 对于可视化技术所面向的数据和专业领域的熟悉程度
  - 对测试任务的熟悉程度
    - 对于所要完成的任务的熟悉程度
  - 对数据的熟悉程度
    - 用户是否曾经接触过同类型或者相似的数据？用户是否已经对这样的数据有一个合理的认知模型
  - 对可视化技术的熟悉程度
    - 对可视化技术的熟悉程度决定了评测中用户是否需要一个学习的过程
  - 对可视化环境的熟悉程度
    - 同样的技术在不同环境下的实现对用户将造成不同的体验
- 研究者可以选择尽量多的群体参与评测，从而更好地了解可视化技术对哪些用户更有效，以及背后的人因学（Human Factors）上的原因

3.2.3 实验设计

- 实验设计指的是对整个实验流程进行规划。Ronald Fisher提出的三条原则
  - 重复（Replication）：实验通常会受到不确定性的影响。重复实验有利于降低不确定性
  - 随机（Randomization）：通过随机方法，如随机数表、抽签等方式，将参与用户分配到不同的组中，以随机顺序实验。这样可以降低实验之外的因素带来的影响。
  - 局部控制（Local Control）：将可能影响实验的因素分开讨论。例如，如果对原有的可视化方法分别做了交互和布局上的改进，那么为了具体观察这两部分改进带来的影响，应当分别实验原有方法——只做交互改进的方法、只做布局改进的方法和做这两种改进的方法。

多实验对象，多次实验

- 当实验的对象不止一个时，为了保证每个对象的参与用户群体相同，每个参与用户需要对不同的对象分别实验
- 在多次实验中，参与用户可能会累积经验或产生疲劳感。为了降低类似因素带来的影响，研究者需要对每个参与用户使用对象的顺序进行规划
  - 拉丁方是一种常见的解决方法

实验顺序	第 1 次	第 2 次	…	第 n 次
参与用户 1	实验对象 1	实验对象 2	…	实验对象 n
参与用户 2	实验对象 2	实验对象 3	…	实验对象 1
…	…	…	…	…
参与用户 n	实验对象 n	实验对象 1	…	实验对象 $n-1$

3.3 进行实验

- 在正式实验开始之前，通过预先进行的小规模试点实验（Pilot Experiment）
  - 检查实验设计、环境是否存在缺陷，了解用户在实际操作时可能会遇到的问题，及时进行实验优化，避免浪费
- 研究者需要提供一个良好的实验环境
- 在正式实验中，研究者需要先向参与用户全面地介绍整个实验
  - 包括实验目的、实验流程、实验中可能用到的设备或系统的使用方法、需要完成的任务和评价标准等
- 为了确保参与用户掌握需要了解的内容，做任务之前，参与用户通过完成一些类似的练习来熟悉实验
- 在做任务的过程中，研究者应尽可能提供自动化的辅助
  - 包括流程上的指示、数据的收集（用时、操作记录）等，以避免失误造成样本数据浪费
  - 同时，在整个过程中需要至少有一个对实验完全了解的人在参与用户遇到问题的时候能及时做出解答
  - 当实验不能在短时间内完成时，应将整个实验过程分段，在每段之间为参与用户留出休息的时间
- 实验的最后，对参与用户进行采访，具体问题可以从以下两个角度来设计
  - 参与实验的感受：有助于改进实验
  - 完成任务时的思考：在讨论实验结果时，解释某些现象需要对这部分信息进行总结

3.4 分析结果并讨论

- 用户实验的最后一步是对收集到的数据进行分析和讨论
  - 在实验中，收集到的数据包括用户的个人信息、完成任务过程中的行为数据、评价数据，以及最后的采访记录等
- 对于可被量化的数据，可以利用统计方法进行假设检验
  - 建立假设。给出一个命题，即零假设（Null Hypothesis），记为H0。与H0对立的另一个命题，被称为备择假设（Alternative Hypothesis），记为H1。当确认H0为假时，研究者将接受H1（H0与H1不一定互补）。一般H1反映了研究者的假设
  - 构造检验统计量。假设H0为真，基于样本数据，通过构造统计量（如）来判断是否正确
  - 确定拒绝域和接受域。将样本空间分成两部分，分别对应接受H0和接受H1（拒绝H0）。这里需要计算两个域的临界点
  - 计算临界点。在判断H0是否为真时，有一定概率会出现错误。在分析数据时，往往会指定出现错误的概率不超过α（显著性水平），根据α计算临界点c
  - 给出判断。在H0为真的前提下，观察样本数据是落在拒绝域还是接受域

统计检验方法

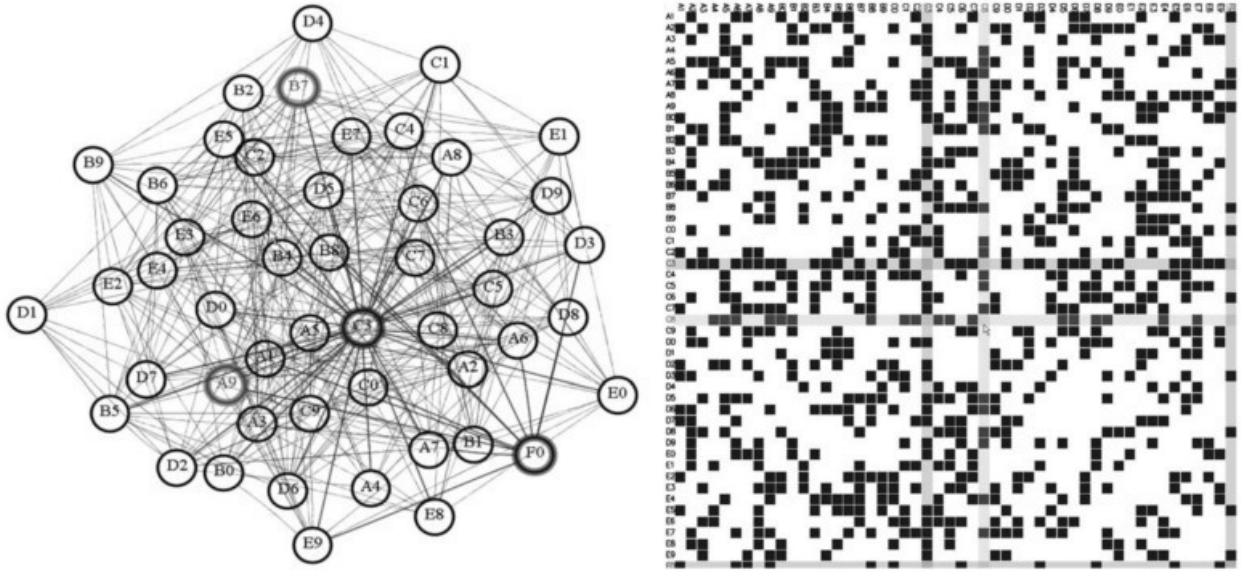
- 卡方检验（Chi-squared Test）
  - 用于确定在一个或多个类别中观察到的频率和期望频率之间是否有显著性差异。常被用于独立性检验
- P值（P-value）
  - 检验假设H0成立或表现更为严重的可能性
- F检验（F-test）
  - 也被称为联合假设检验（Joint Hypotheses Test），可以在H0下检验是否具有F分布。常被用于分析多个相关因素对因变量的影响

3.5 评测案例分析

- 3.5.1 案例一
- 3.5.2 案例二

3.5.1 案例一

- 网络数据通常可以用点线图（Node-Link Diagram）和邻接矩阵（Adjacency Matrix）来可视化
- 两种方法有各自的优点和局限性。Ghoniem等人在[Ghoniem2005]中对这两种可视化的可读性进行了全面的评测



第一步：确定实验目标

- 研究者关注两种可视化的可读性，且希望所做的分析有一定的通用性，与所用数据的来源和领域无关
  - 具体到网络数据，用户最关心的是与网络连接结构相关的信息
- 设计了7个任务
  - （1）估计网络中节点的数量
  - （2）估计网络中链接的数量
  - （3）找到网络中链接最多的节点
  - （4）按照名字在可视化中找到对应的节点
  - （5）找到两个节点之间的直接链接
  - （6）找到两个节点之间的共同邻节点
  - （7）找到两个节点之间的路径
- 如果用户能够在短时间内正确地完成这些任务，对应的可视化就有较好的可读性
  - 因此，可对用户完成这些任务的时间和正确率进行测量和统计，并以此作为评测可读性的指标。这两个指标是测量用户行为的基本指标，在最后的结果分析中都起到了重要的作用

第二步：准备实验

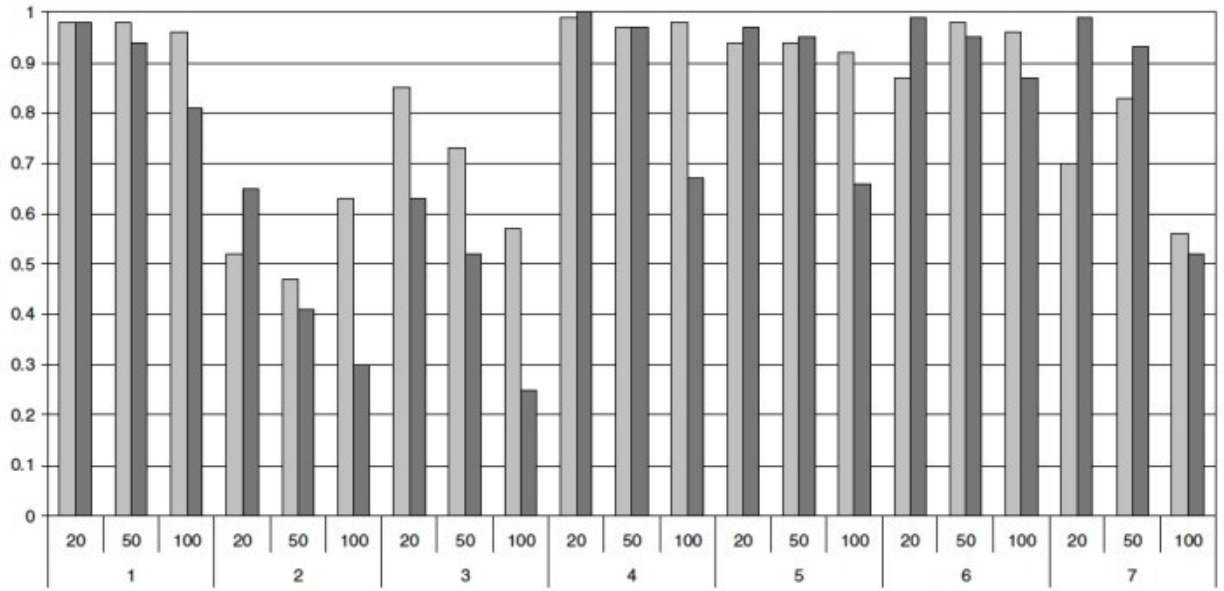
- 研究者选择使用随机生成的网络数据进行测试
  - 研究者在数据的选取时不但从评测的数据特性出发，也认真地考虑了各个目标任务的需求
- 在评测中，研究者采用了招募志愿参与者进行可用性评测的方法
- 两种可视化的实现和优化程度对于评测最终的结果是否有效也非常重要
  - 对于点线图，研究者采用了可视化开源工具GraphViz[Graphviz2012]，使用的布局程序是neato
  - 矩阵可视化由自己开发的可视化程序实现，其中节点在横轴和纵轴上按照名字的字母表顺序排列

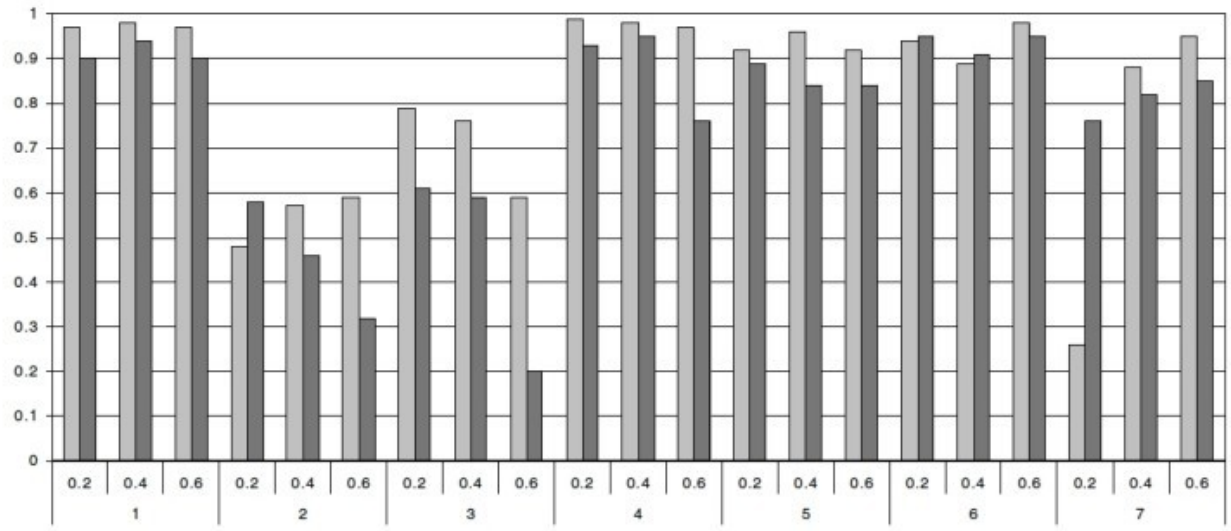
第三步：进行实验

- 在开始用户测试之前，研究者通过演示向测试者介绍如何正确地解读这两种可视化，并如何完成目标任务
- 其后，用户在研究者的帮助下尝试完成一些示范的任务，以确保他们对可视化方法、系统的交互和要完成的任务有准确的理解。如果还有疑问，研究者会再次演示，直到确认测试者掌握了这两种可视化
- 最后，对测试者提出如下三点要求
  - （1）必须尽快完成任务
  - （2）必须尽量正确地回答问题
  - （3）如果觉得某个任务无法完成，则可以跳过它进入下一个任务
- 在实验中，研究者对用户完成任务的时间进行了有效的安排和控制。用户一共需要完成126个任务（2个可视化×9个网络图×7个任务），每个任务限时45秒

第四步：分析结果并讨论

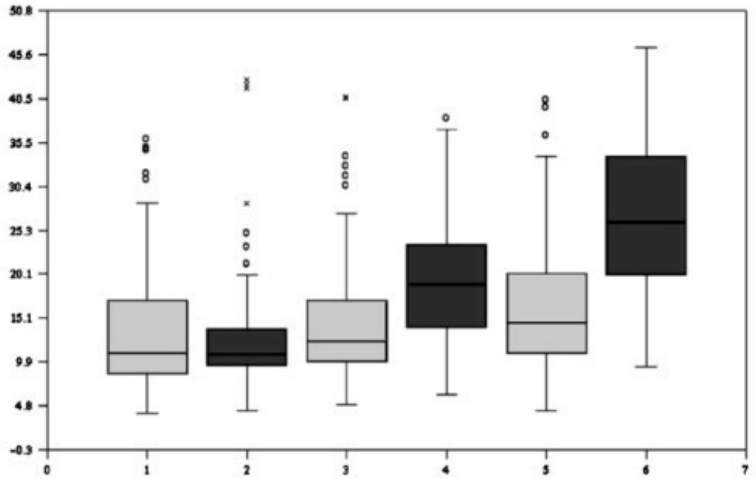
- 实验所得到的结果主要包括完成任务的时间和正确率。研究者的目标不仅是总体的表现对比，还希望了解网络大小和密度对可视化可读性的影响



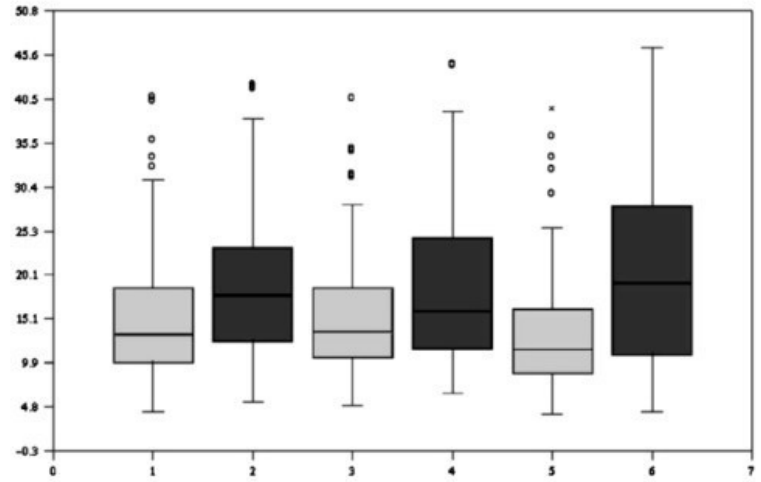


任务1的结果分析

- 估计网络中节点的总数
- 用盒须图表示了任务完成时间的分布。随着网络图变大，用户通过矩阵可视化完成任务的时间变化不大
- 而用户采用点线图的完成时间的中位值和方差都大大增加
- 网络密度对两种可视化所对应的完成时间都影响不大



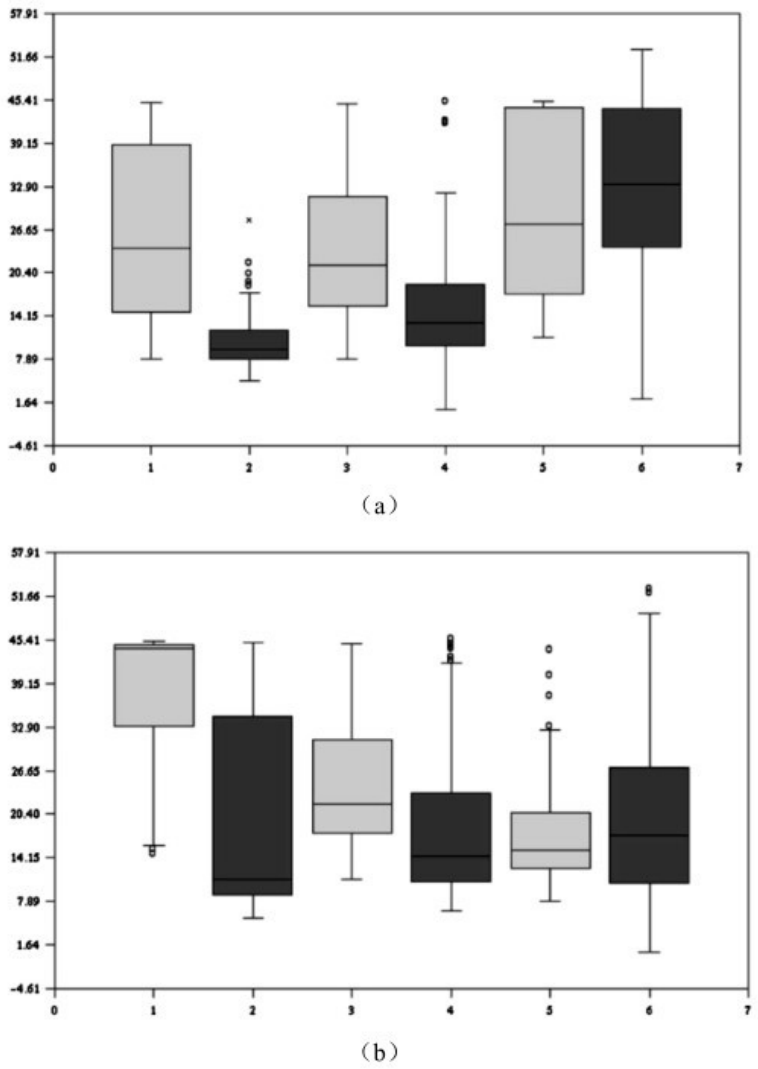
(a)



(b)

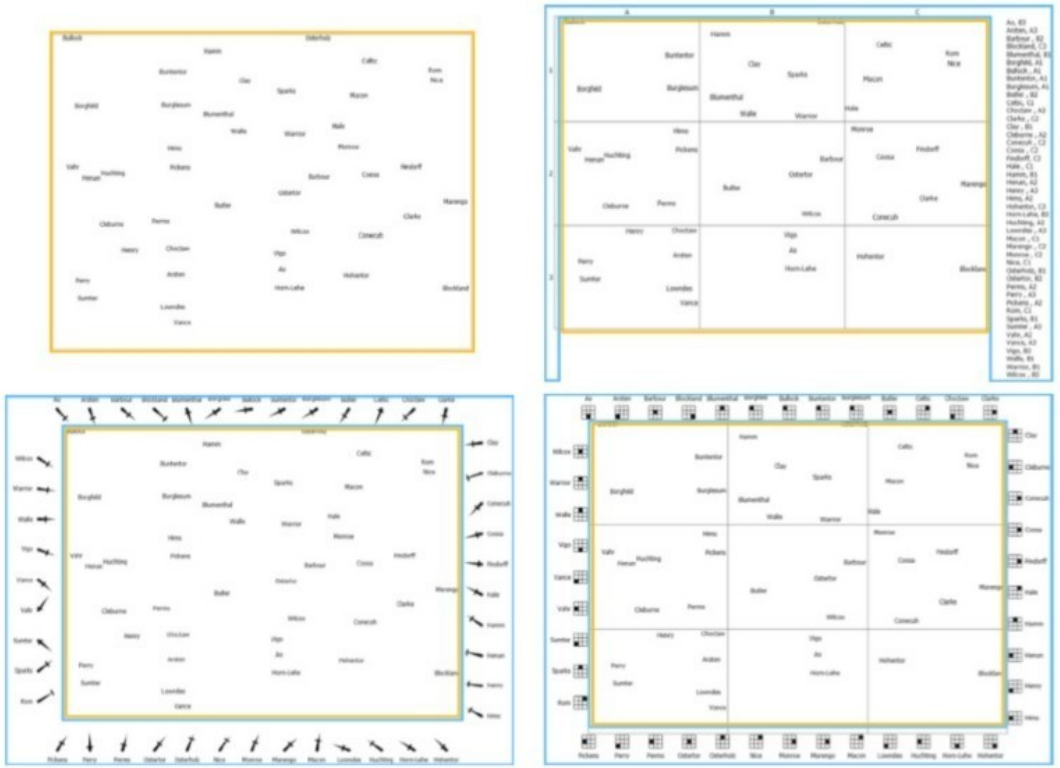
任务7的结果分析

- 找到两个节点之间的路径，即找到一系列连接两个节点的链接
- 对于点线图，网络图越大，所需要的时间越多，网络密度对完成时间影响不大
- 矩阵表示，网络密度越大，所需要的完成时间越短
- 对于节点较少且密度小的网络，矩阵可视化在寻找路径上并不如点线图有效
- 对于高密度的大网络，矩阵可视化反而更有效
  - 研究者对此的推测是，当网络密度增大时，任意两个节点更有可能直接相连，因此寻找路径简化为寻找两点之间直接链接的任务，矩阵可视化变得更为有效



3.5.1 案例二

- 游客来到陌生的地方，想要在地图上找到某个感兴趣的区域（AOI），需要花费很多时间来浏览地图。标注和注释可以有效地缩短他们的搜索过程
  - 什么样的标注才是最有效的呢？Netzel等人 [Netzel2017]使用眼动仪对4种在地图上标注的方法进行了比较



第一步：确定实验目标

- 常见的标注方法是首先对多个区域的名字在图外排序，并在名字旁边给出关于相应位置的提示。研究者希望对一种无标注的方法（基准方法）和三种提供不同标注提示的方法进行比较
  - 图内标注（WA）：仅在图内相应位置直接标出区域名，不添加标注
  - 网格参考标注（GA）：通过2D笛卡儿坐标系将地图分成多个小单元，并基于行列对小单元编号。在地图外，在每个区域名字旁边会给出它所在位置的单元编号，如“A3”
  - 方向标注（DA）：除了坐标，通过方向和距离也可以确定一个区域的位置。该方法用箭头指出区域的方向，并在箭头上通过标注的位置表示距离的大小
  - 缩略图标注（MA）：它同样是基于网格单元划分的。不同之处在于，每个单元的位置不是通过编号来表示的，而是通过缩略图的相应位置高亮来表示的
- 在实验开始之前，基于理论分析，研究者提出了5点假设
  - （1）不使用标注的方法在地图上找到AOI，需要花费比使用标注的方法更长的时间，即：WA＞GA,WA＞DA,WA＞MA
  - （2）使用三种有标注的方法所花费的时间也存在差异：GA＞DA＞MA
  - （3）使用有标注的方法时，参与用户的扫视长度将会大于不使用标注的方法，因为有了标注的提示，用户的扫视可以有更长的跳转
  - （4）使用DA、GA和MA方法时，可视搜索从外部区域开始，然后转向内部，最后结束于目标标签。此外，参与用户注视外部区域的平均时间应该比注视内部区域的平均时间要短。这是因为标签的实际搜索比利用视觉辅助来估计标签的位置需要更多的时间
  - （5）在注视内部区域时，使用DA方法的扫视运动模式不同于GA和MA方法。使用DA方法时，参与用户的目光会沿着一条线定向搜索；而使用GA和MA方法时，参与者会在某个格子中搜索。因此，在后续扫视中，比起GA和MA方法，DA方法将会有更小的角度偏差
- 为了验证这些假设，参与用户将被要求戴着眼动仪，尽可能快地完成在地图上准确找到特定标签的任务。每次任务的完成时间将被记录下来。除此之外，眼动仪将记录参与用户的视线移动情况

第二步：准备实验

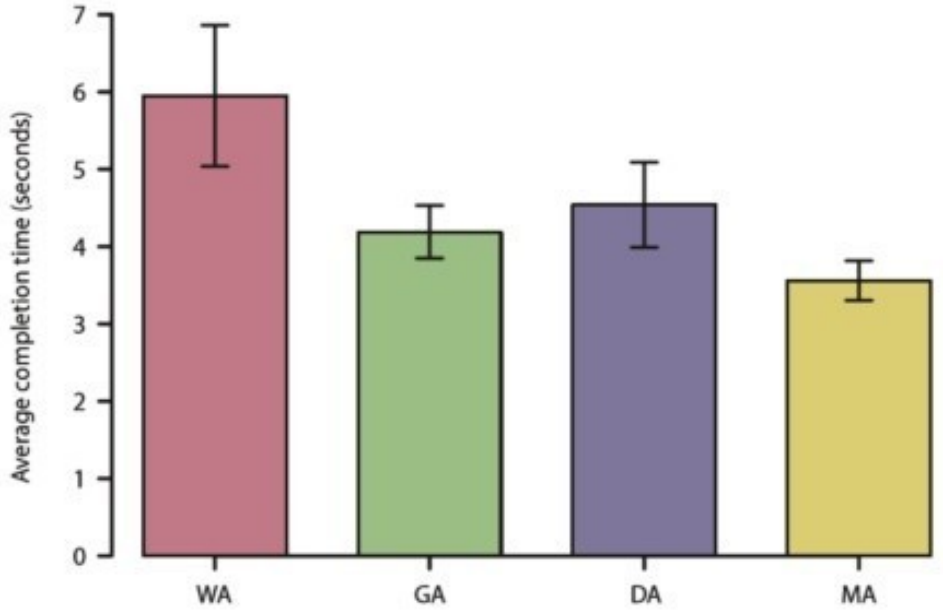
- 研究者生成了一些人造地图数据以供实验使用。具体生成方法为，将美国、法国、德国和英国的主要城市名字分散到3×3网格中，并在每格内先随机放置5个不同的标签（共45个）
- 招募了32位大学生，在完成时间不超过60分钟的实验后，每位参与用户将得到10欧元的报酬。每个参与用户需要基于80张地图完成任务（每种方法各20张）
  - 因此，每4个参与用户可以给出一份全数据集的测试结果
  - 实验使用两阶段平衡以补偿学习成本和疲劳效应，即：将80张地图分成4组，每组内每种标注方法各5张，随机排序
- 该研究在研究者所在的实验室进行
  - 在实验过程中，除参与用户以外，房间内只有一位实验操作员，因此，实验环境比较安静
  - 在实验过程中，参与用户坐在屏幕前方约60厘米处，以保证眼动仪的良好校准
  - 眼动追踪软件的标准滤波器参数为最小覆盖范围是10像素；最短固定间隔为30ms。因为参与者的头部并未固定，所以其到屏幕的距离并不恒定。不过鉴于头部运动的影响很小，视角1° 可以对应约35像素

第三步：进行实验

- 请参与用户签署同意书，通过Snellen图表完成视力测试并提供一些信息
  - 经测试，全部参与用户的视力为正常或矫正后正常。统计结果表明，在32位参与用户中，有27位男士、5位女士；年龄在20～32岁之间，平均年龄为22.8岁；有29位专业为计算机科学或者软件工程。
- 向参与者讲解任务并引导他们完成教程
  - 教程中包括每种标注方法的解释和示例任务。
- 在任务执行过程中，整个图像分两部分向用户呈现：先显示目标标签名字；按键后显示地图图像
  - 在定位标签过程中，参与用户不允许使用任何辅助手段，包括鼠标、手指灯，以避免对方向标注的影响。当参与用户找到目标后，再次按键，使用鼠标选择找到的目标，结束对该图像的实验并开始下一个图像的任务。
- 填写调查问卷
  - 参与用户需要回答一些主观问题，比如：你是否使用了搜索技巧？MA和GA方法哪个在定位时更方便？
- 在本次实验中，由于技术问题导致眼动追踪记录错误，两位参与用户的测试数据被排除。

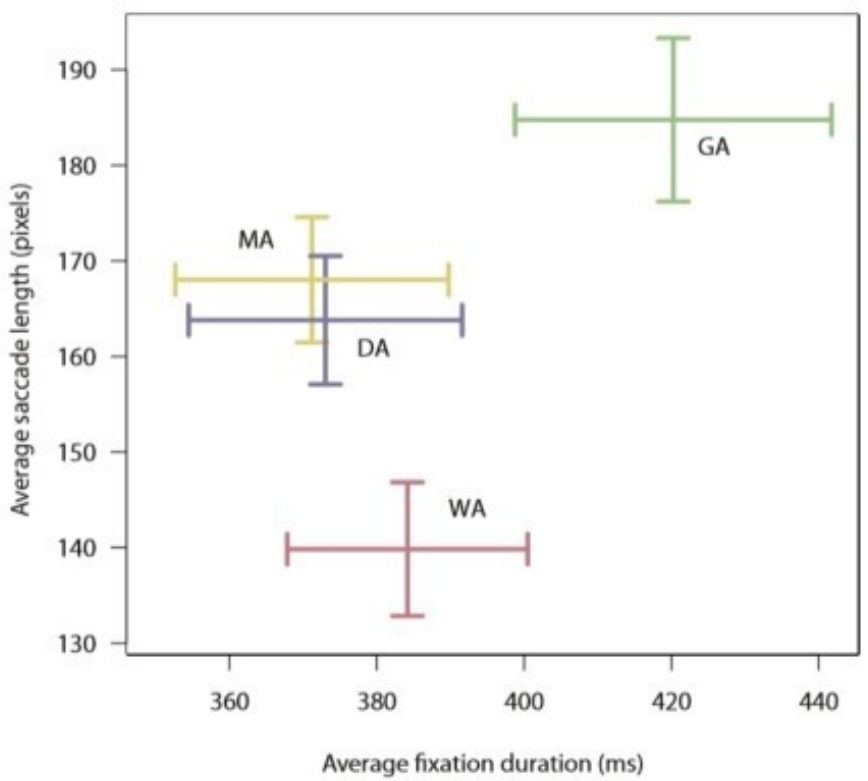
第四步：分析结果并讨论

- 1.任务执行分析
- 研究者根据任务的完成时间对标注方法进行了评估
  - 参与者平均需要3.56s（MA）、4.19s（GA）、4.54s（DA）和5.95s（WA）完成任务
  - 与基准方法WA相比，MA快40.2%,GA快29.6%,DA快23.7%。



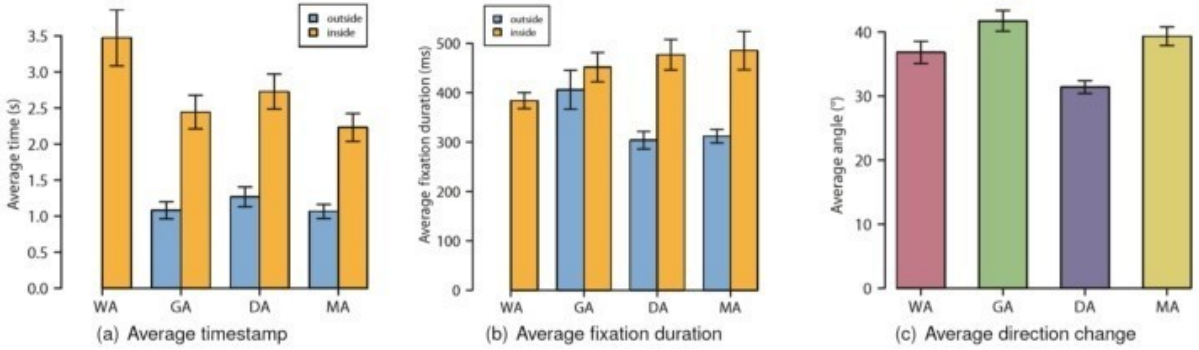
2.眼动数据分析

- 在分析眼动数据时，研究者将平均注视时间和平均扫视长度作为相关量进行了评估
- 由于设计相似，DA和MA方法形成了一个聚类，另外两种方法则分散在其他地方
- 对注视时间的事后分析表明，GA方法和所有其他方法之间存在差异（$p < 0.05$）；对扫视长度的分析表明，除MA和DA方法之外，所有成对组合都有显著性差异（$p < 0.007$）。这个结论证明了假设3



眼动数据分析-2

- 对GA、DA和MA三种标注方法的两个视觉搜索阶段进行了分析
  - 在第一阶段中，参与用户将注意力集中在地图的外部区域
  - 在第二阶段中，他们将注意力切换到地图内搜索标签。为了提取这两个阶段，研究者分别查看了每次注视的时间戳。
  - 这两个阶段是明显分开的，第一阶段发生在第二阶段之前。使用除WA方法以外的任意方法时，参与用户都会先在外部区域搜索，而WA方法没有第一阶段。这证明了假设4的第一部分
- 为了验证假设4的第二部分，研究者分析了两部分的注视时长
  - 第二阶段的平均时间比较长。在第一阶段中，与MA和DA方法相比，GA方法的注视时间更长。使用GA方法时，参与用户不仅必须找到标签所在格子的标注，还必须记住它的坐标。因此，其必须花费更长的时间。而使用MA和DA方法时，参与用户可以较快速地确定粗略位置，并进入第二阶段，搜索标签
- 研究者继续分析在第二阶段中扫视路径方向变化的平均角度，以讨论假设5
  - 为了处理搜索方向的反转（180°转弯），研究者研究了较小的方向变化角度。统计检验结果显示标注方法对角度确实有显著性影响



3.主观评估

- 图内标注（WA）
  - 19位参与用户使用了搜索技巧。最常见的是水平或垂直扫视，或者从屏幕中间开始，以螺旋形向外扫视
- 网格参考标注（GA）
  - 25位参与用户使用了搜索技巧。他们先从标注中找到目标标签和相应的网格坐标，然后再到相应的单元格中寻找
- 方向标注（DA）
  - 26位参与用户使用了搜索技巧。他们先在图像周围找到方向注释，然后跟着箭头寻找标签
- 缩略图标注（MA）
  - 25位参与用户使用了一种常见的搜索技巧。与DA方法类似，他们首先在标注中搜索，然后再跳转到相应的格子中进行寻找

总结

- 可视化方法的系统性科学评测是可视化应用发展、研究深入的重要驱动力之一
- 只有具备了完善的评测体系，可视化研究才能向正确的方向前进
- 由于种种原因，可视化评测方面的研究还很欠缺