博士生科研入门辅导

骆昱宇
数据科学与分析学域
香港科技大学（广州）

大纲

- 浅谈科研入门
- 做研究需要哪些素质
- 学术论文撰写

浅谈科研入门

从论文选题到论文发表

研究方向 → 具体问题 → 技术创新 → 实验评估 → 论文写作

从论文选题到论文发表

研究方向 → 具体问题 → 技术创新 → 实验评估 → 论文写作

从研究方向到具体问题

研究对象	研究层面	主要问题	现有挑战	核心思路
结构化数据可视化系统	方法层	现有系统高度依赖用户主动理解数据集和可视化，对用户要求高	可视分析高门槛	自动可视化
	用户层	现有系统难以支持普通用户准确表达可视化意图，容易答非所问	用户意图难表达	
	数据层	现有系统容易忽视数据错误对可视化结果的影响，容易误导用户	分析结果不精准	

如何选择研究方向？

问题背景与现有挑战

数据

起飞时刻表	航班号	航空公司	目的地	起飞延迟（分）	到达延迟（分）	乘客数
01/01 00:05	CA4118	中国国航	北京	-1	0	204
01/01 04:00	ZH9163	深圳航空	巴黎	0	-2	193
01/01 07:05	3U8886	四川航空	广州	0	3	188
01/01 10:44	3U8888	四川航空	上海	11	9	112
01/01 12:32	CA4110	中国国航	北京	24	26	142
01/01 16:12	CA9916	南方航空	北京	13	7	152
01/01 20:11	MF1816	厦门航空	上海	10	4	301
...	...	...	...	...	...	...

CSV MySQL

中国科技创新2030“新一代人工智能”和“大数据”专项均将可视化和可视化分析列为大数据智能亟需突破的关键共性技术。

《“十四五”规划和2035年远景规划纲要》将可视化列为数字经济产业和大数据领域需要重点推进的技术。

现有挑战

可视化系统

方法	可视分析高门槛	领域知识 分析技能
用户	用户意图难表达	编程实现 复杂交互
数据	数据质量较敏感	数据错误 误导用户

4

如何选择研究方向？

- 问老板、师兄师姐
- 自己多读论文、多交流
  - 读顶会会议论文
  - 多听报告
  - 多参加会议
- 交流访问
  - 参加国际会议、做学术报告
  - 国际交流放学、有一定科研基础再出去交流
  - 了解科研动态和领域牛人

好的研究方向应具备的条件

- 要有实际需求 – **重要**
- 能吸引人 - **新颖**
- 有较大的未知空间 – **非显而易见**
- 实验室有好的相关积累 – **基础**

从论文选题到论文发表

研究方向 → 具体问题 → 技术创新 → 实验评估 → 论文写作

具体研究问题举例

结构化数据自动可视化 -- 研究框架

研究概览

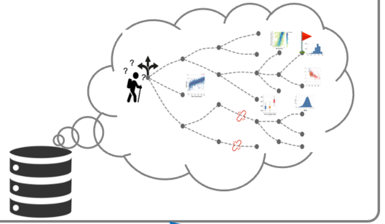
DEEPEYE 智能可视化系统 = 自动可视化 + 用户意图 + 干净数据

领域知识指导的全自动可视化
[SIGMOD'23, TKDE'22, ICDE'18]

自然语言驱动的问答式可视化
[SIGMOD'21, IEEE VIS'21]

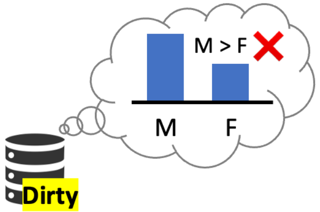
数据质量感知的渐进式可视化
[ICDE'20, VLDB'20]

可视分析高门槛



用户意图难表达

数据质量较敏感



传统可视化和可视分析系统 = 可视分析方法 + 用户 + 数据

如何选择具体研究问题？

- 一般来说，解决一个具体的研究问题，就能发一篇Paper。
- 确定一个具体研究问题的方法：
  - 找老板要Idea
  - 自己读论文，从研究前沿去推敲可以做的题目

- 具体研究问题的特点：
  - 因**新需求**、**新技术**的出现，而产生的**新问题**
    - 例如，如何**在元宇宙中**做**大数据**的**实时分析**？
  - 已经被前人定义好的**老问题**
    - 例如，人机对弈围棋、轨迹的相似度查询

如何选择具体研究问题？

- 新的研究问题：
  - 需要明确定义这个问题，为什么重要？
  - 需要有较好的技术贡献。
  - 这类问题的研究讨论是需要有一个完整的故事链条和恰当的技术贡献
    - 是有什么新需求导致我有了这么个具体的问题？
    - 针对这个问题，我做了哪些技术贡献？
  - **研究者的贡献侧重于：清晰地定义一个好的新问题**
- 老的研究问题：
  - 问题已经很清晰明确了，例如轨迹的相似度查询。
  - 这类问题的研究套路基本就是在已有的Benchmark上打榜！
  - **研究者的贡献侧重于：新技术**

如何判断一个研究问题的价值

- **老问题**
  - 问题已经明确定义
  - 目标：更好的方法

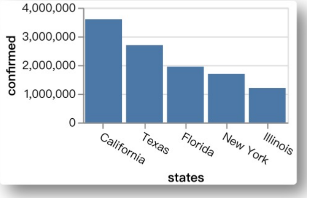
- **新问题**
  - 目标：解决没人研究过的问题
  - 条件
    - 有真实的应用背景
    - 能获得真实的数据集（或者自己构建真实数据集）
    - 结果好衡量，如果需要衡量效果（quality），最好有benchmark

举例：老问题

Natural Language to Visualization (NL2VIS)

Create a bar chart showing the top 5 states with the most confirmed cases until 2021-03-08

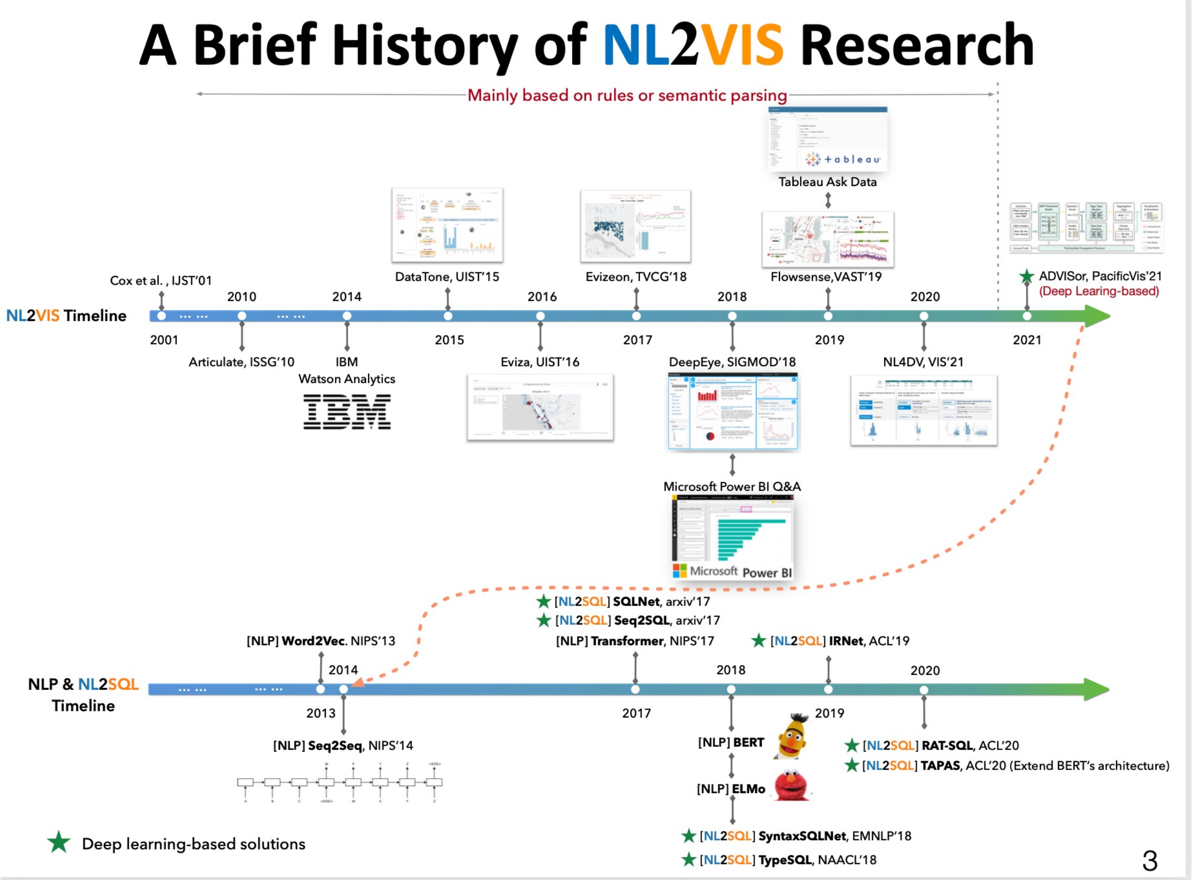
date	states	cases	number
2021-03-08	California	confirmed	3599250
2021-03-08	California	deaths	54220
2021-03-08	New York	confirmed	1694651
2021-03-08	New York	deaths	48335



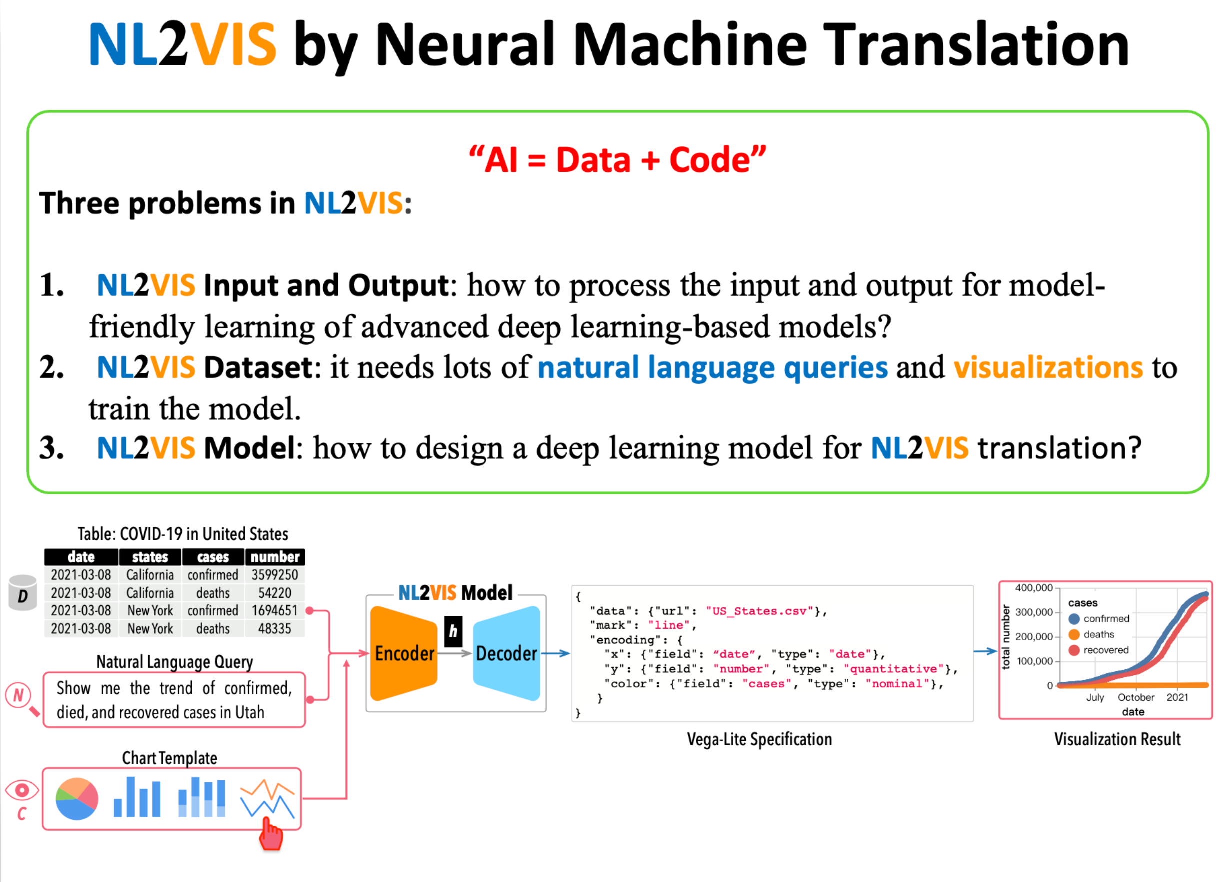
Tabular Data (e.g., COVID-19) | Natural Language Query | Visualization

NL2VIS Advantages:

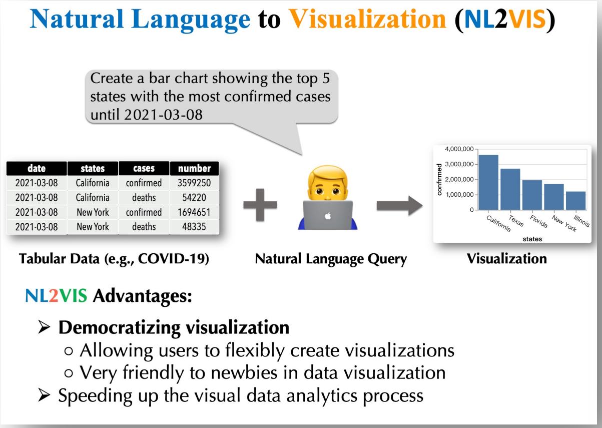
- **Democratizing visualization**
  - Allowing users to flexibly create visualizations
  - Very friendly to newbies in data visualization
- Speeding up the visual data analytics process



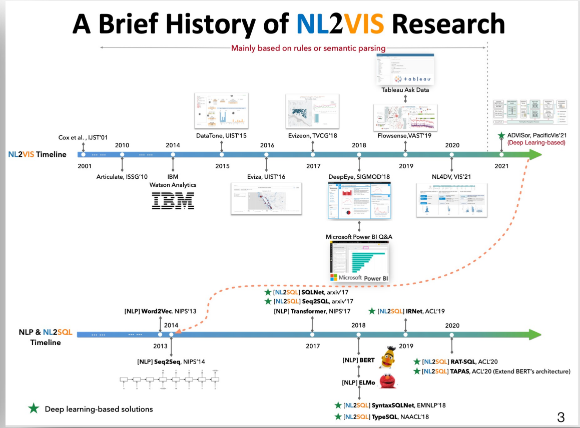
举例：老问题



NL2VIS by Neural Machine Translation
"AI = Data + Code"
Three problems in NL2VIS:
1. NL2VIS Input and Output: how to process the input and output for model-friendly learning of advanced deep learning-based models?
2. NL2VIS Dataset: it needs lots of natural language queries and visualizations to train the model.
3. NL2VIS Model: how to design a deep learning model for NL2VIS translation?
Table: COVID-19 in United States
date | states | cases | number
2021-03-08 | California | confirmed | 3599250
2021-03-08 | California | deaths | 54220
2021-03-08 | New York | confirmed | 1694651
2021-03-08 | New York | deaths | 48335
D
Natural Language Query
N
Show me the trend of confirmed, died, and recovered cases in Utah
Chart Template
C
NL2VIS Model
Encoder
h
Decoder
{
"data": {"url": "US_States.csv"},
"mark": "line",
"encoding": {
"x": {"field": "date", "type": "date"},
"y": {"field": "number", "type": "quantitative"},
"color": {"field": "cases", "type": "nominal"},
}
}
Vega-Lite Specification
cases
confirmed
deaths
recovered
total number
date
July October 2021
Visualization Result



Natural Language to Visualization (NL2VIS)
Create a bar chart showing the top 5 states with the most confirmed cases until 2021-03-08
date | states | cases | number
2021-03-08 | California | confirmed | 3599250
2021-03-08 | California | deaths | 54220
2021-03-08 | New York | confirmed | 1694651
2021-03-08 | New York | deaths | 48335
Tabular Data (e.g., COVID-19)
Natural Language Query
Visualization
NL2VIS Advantages:
➢ Democratizing visualization
○ Allowing users to flexibly create visualizations
○ Very friendly to newbies in data visualization
➢ Speeding up the visual data analytics process



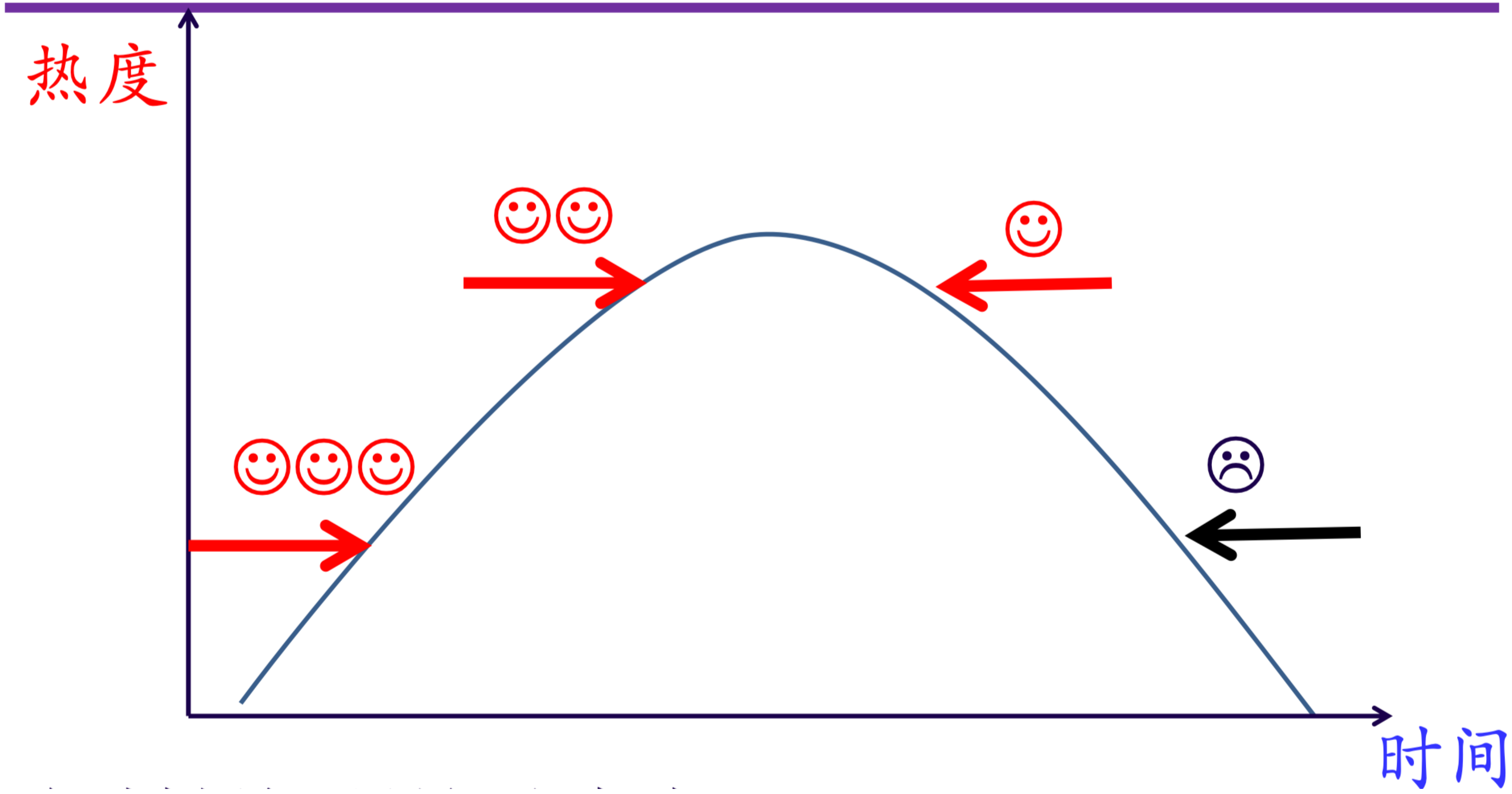
A Brief History of NL2VIS Research
Mainly based on rules or semantic parsing
NL2VIS Timeline
NLP & NL2SQL Timeline
IBM
Deep learning-based solutions
3

如何做老问题？

- 更快、更高、更强（Faster, Higher, Stronger）
  - 提高性能
  - 提高准确率
  - 提高通用性，例如去掉参数或者假设
  - 充足的实验比较

- 应用到新的环境、满足新需求
  - 新硬件、新场景（元宇宙、大模型即服务）
  - 隐私要求（如何解决可视化中的隐私保护要求？）
  - 算力不足（如何用低成本的硬件训练大模型？）

老问题的生命周期



如何判断一个研究问题的价值

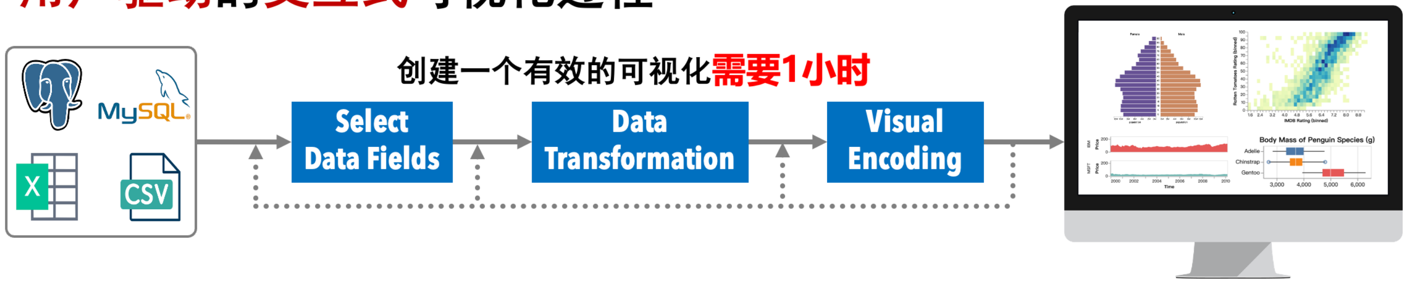
- 老问题
  - 问题已经明确定义
  - 目标：更好的方法

- **新问题**
  - 目标：解决没人研究过的问题
  - 条件
    - 有真实的应用背景
    - 能获得真实的数据集（或者自己构建真实数据集）
    - 结果好衡量，如果需要衡量效果（quality），最好有benchmark

举例：新问题

结构化数据自动可视化（AutoVIS）

用户驱动的交互式可视化过程



算法驱动的全自动可视化过程

降低门槛，提高效率

如何找新问题？

- 多想、多联系实际
- 读交叉学科论文
- 新问题需要的条件
  - 问题意义 — 令人信服
  - 应用背景
  - 可以解决
- 当前新问题例子
  - 大模型与可视分析的交叉研究
  - 元宇宙下的数据分析

如何研究新问题？

- 多想多借鉴
- 创新 – 不能重复别人工作
- 发散性思维
- 抽象能力（Problem definition）
- 深入分析
- 有毅力

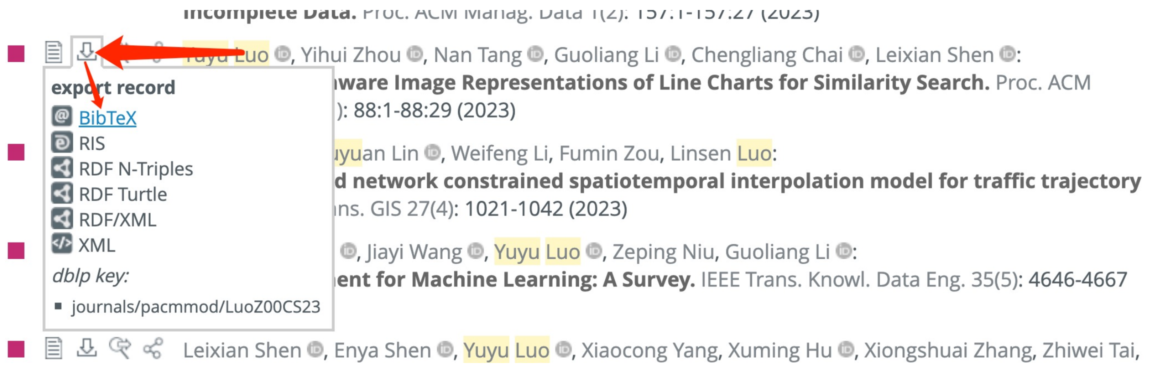
如何找论文？

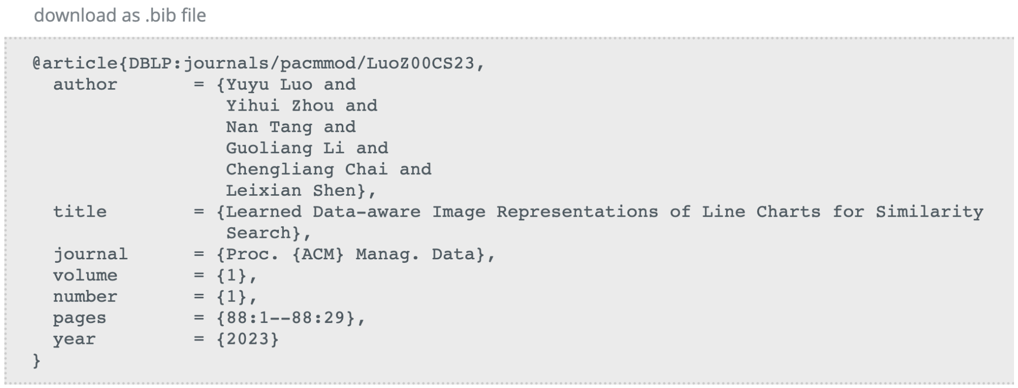
- 找一个好的问题，离不开大量阅读相关文献！
- **如何找文献？**
  - **90%的精力都集中在领域的顶会顶刊**
    - **数据库：SIGMOD、VLDB、ICDE、KDD、TKDE、VLDB Journal**
    - **可视化：IEEE VIS、TVCG、CHI、IEEE VR、EuroVIS**
  - **10%的精力放在其他交叉领域（即对你的研究有帮助的）**
    - **AI：ICML、ICLR、ACL、CVPR**

- 参考CCF A类列表或者清华推荐的顶会顶刊列表
  - https://www.ccf.org.cn/Academic_Evaluation/By_category/
  - https://numbda.cs.tsinghua.edu.cn/~yuwj/TH-CPL.pdf

常见的文献调研网站

- 谷歌学术
  - https://scholar.google.com/
- https://www.connectedpapers.com/
- DBLP（这里的论文比较规整）
  - http://dblp.org/
  - 用Latex写论文的时候，优先从DBLP导出论文的引用！






```
@article{DBLP:journals/pacmmod/LuoZ00CS23,
  author       = {Yuyu Luo and
                  Yihui Zhou and
                  Nan Tang and
                  Guoliang Li and
                  Chengliang Chai and
                  Leixian Shen},
  title        = {Learned Data-aware Image Representations of Line Charts for Similarity
                  Search},
  journal      = {Proc. {ACM} Manag. Data},
  volume       = {1},
  number       = {1},
  pages        = {88:1--88:29},
  year         = {2023}
}
```

如何读论文？

- 找方向时
  - 广度优先
  - 看idea，不用注重细节
- 找到感兴趣方向时
  - 深度优先
  - 看细节
  - 看实验、数据集
  - 能够实现别人代码
  - 抓住论文的要害
  - 批判式阅读

做研究需要哪些素质？

如何做研究？

- 脚踏实地（努力）
  - 不能熊瞎子掰棒子（掰一个，吃一个，夹一个，丢一个。最后到了地头，才发现辛辛苦苦忙乎了一场，除了吃到肚里的，结果还是丢得多，捡得少）
- 规划能力（远见）
- 了解相关工作，做好实验（知彼）
- 认清论文质量（知己）
- 多总结，多借助于工具（总结）
  - 总结经验教训

做研究的大忌

- 没有兴趣
- 不了解前人的工作
- 浮躁、急于求成 、不求甚解
- 马虎
- 懒惰

一些经验分享

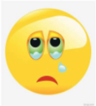
找准科研方向、沟通交流很重要
撸起袖子去干、快速迭代不拖延

脚踏实地、写好每一行代码
仰望星空、打开想象的空间

三四月做的事情
八九月自有答案

一些经验分享

- 一定要保证科研时间

		
努力	Always	?
躺平	Seldom	Always

✓ 至少每天保证5-6小时的有效科研（课题）时间

✗ 玩手机

✗ ✗ ✗ 打游戏

喝一碗科研鸡汤

- 勤奋努力
- 创新钻研
- 自强不息

兴趣 or 心态

建立和谐的师生关系

- 目标一致
- 相互欣赏
- 共同进步
- 契约精神

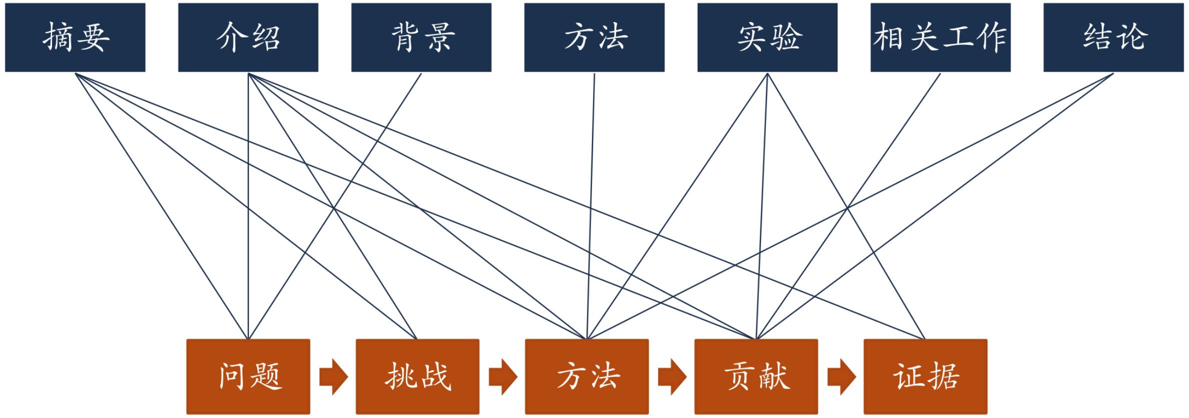
学术论文撰写

如何衡量一个工作的价值？

- 问题新颖程度
- 方法创新程度
- 技术理论深度
- 实验结果好坏（与现有工作对比）
- 论文撰写功底（逻辑结构、英语）

如何写论文

- 八股文
  - 模仿（可以参考课题组以往的论文，或者相近研究方向的论文）
- 摘要
- 引言
- 相关工作
- 整体框架
- 技术细节（3个左右的创新点）
- 实验
- 总结



摘要：论文的高度总结（精华）

摘要一般包含：

- What（解决/研究什么问题？）
  - 交代背景
- Why（为什么写这篇论文？）
  - 新东西：提出一个新问题、发现一个新规律
  - 旧东西，新方法：比现有的更快、更好、成本节约等等
- How（大概怎么做？Key Idea）
- 你做了什么贡献?
- 实验效果如何?

摘要：论文的高度总结（精华）

What：
研究什么问题？（交代背景）

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.*, a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

摘要：论文的高度总结（精华）

Why：
旧东西，新方法：我们对于这个问题有怎样新的观察?
“我的提出的方法可能‘效果更好’”

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.*, a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

摘要：论文的高度总结（精华）

How：
我们大概是怎么做的？Key Idea

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.*, a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

摘要：论文的高度总结（精华）

实验效果

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.*, a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

摘要：论文的高度总结（精华）

DeepEye: Towards Automatic Data Visualization

Yuyu Luo[†] Xuedi Qin[†] Nan Tang[‡] Guoliang Li[†]
[†]*Department of Computer Science, Tsinghua University, China* [‡]*Qatar Computing Research Institute, HBKU, Qatar*
{luoyuyu@mail., qxd17@mails., liguoliang@}tsinghua.edu.cn, ntang@hbku.edu.qa

What

Why

How

So What

***Abstract*—Data visualization is invaluable for explaining the significance of data to people who are visually oriented. The central task of automatic data visualization is, given a dataset, to visualize its compelling stories by transforming the data (e.g., selecting attributes, grouping and binning values) and deciding the right type of visualization (e.g., bar or line charts).**

We present DEEPEYE, a novel system for automatic data visualization that tackles three problems: (1) *Visualization recognition:* given a visualization, is it "good" or "bad"? (2) *Visualization ranking:* given two visualizations, which one is "better"? And (3) *Visualization selection:* given a dataset, how to find top-k visualizations? DEEPEYE addresses (1) by training a binary classifier to decide whether a particular visualization is good or bad. It solves (2) from two perspectives: (i) *Machine learning:* it uses a supervised *learning-to-rank model* to rank visualizations; and (ii) *Expert rules:* it relies on experts' knowledge to specify partial orders as rules. Moreover, a "boring" dataset may become interesting after data transformations (e.g., binning and grouping), which forms a large search space. We also discuss optimizations to efficiently compute top-k visualizations, for approaching (3). Extensive experiments verify the effectiveness of DEEPEYE.

A. scheduled	B. carrier	C. destination city name	D. departure delay (min)	E. arrival delay (min)	F. passengers
01-Jan 00:05	UA	New York	-4	1	193
01-Jan 04:00	AA	Los Angeles	0	-2	204
01-Jan 06:13	MQ	San Francisco	7	-11	96
01-Jan 07:33	OO	Atlanta	11	-2	112
...	...	...	...	...	...

Table I
AN EXCERPT OF FLIGHT DELAY STATISTICS

Although (1) and (2) can be quantified formally, by statistical deviations and correlations, respectively, our 55 minutes thought is to study (3) despite the hardness of quantifying human perception, because one fundamental request from users is just to find eye-catching and informative charts. The bad news is that users have poor choices for (3).

Example 1: Consider a real-world table about *flight delay statistics of Chicago O'Hare International (Jan – Dec, 2015)*, with an excerpt in Table I (https://www.bts.gov). Naturally, the Bureau of Transportation Statistics wants to visualize some valuable insights/stories of the data.

Figure 1 shows sample visualizations DEEPEYE considers

引言：摘要的扩充版，论文的总结版

- 引言一般包含
  - Why：研究的动机（讲故事）
    - 问题背景/应用前景
  - How：基本的Idea，即你对这个问题的观察是什么？
  - What：论文解决什么问题？
  - So What：创新点（技术创新+实验结果）
  - 文章结构（Roadmap/Paper Organization）非必要，通常和创新点融合

吸引审稿人，几乎决定文章能否被录取

引言：摘要的扩充版，论文的总结版

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.,* NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.,* a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

1 INTRODUCTION

Natural language interface is a promising interaction paradigm for simplifying the creation of visualizations [32,43,52]. If successful, even novices can generate visualizations simply like a Google search. Not surprisingly, both commercial vendors (*e.g.,* Tableau's Ask Data [46], Power BI [2], ThoughtSpot [3], and Amazon's QuickSight [1]) and academic researchers [7,13,20,33,34,40,42,45,49,50,57] have investigated to support the translation from NL queries to visualizations (NL2VIS).

引言：摘要的扩充版，论文的总结版

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware opti … ocess, and visualization-aware rendering to produce better visuali … natural language queries, **ncNet** is also designed to take an opti … al input, where the chart template will be served as a constraint to … ation and user study, showing that **ncNet** achieves good accuracy …

NL2VIS needs both natural language understanding that uses machines to comprehend natural language queries, and translation algorithms to generate targeted visualization using a visualization language. Natural language understanding is considered an AI-hard problem [56], with many intrinsic difficulties such as ambiguity and underspecification. Many tools from the NLP community, especially based on statistical phrase-based translation [26] and neural machine translation [4, 10], have been used to tackle NL2VIS.

The state-of-the-art NL2VIS methods (for example, NL4DV [40] and FlowSense [57]) are statistical phrase-based translation, which treats natural language understanding and machine translation as two steps. They first employ NLP toolkits (for example, NLTK [5], Stanford CoreNLP [37], and NER [12]) to parse an NL query and produce a variety of linguistic annotations (for example, parts of speech, named entities, etc), based on which they then devise algorithms to generate target visualizations. They are good choices when there are not many training datasets to train deep learning models.

引言：摘要的扩充版，论文的总结版

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.,* NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.,* a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could ... and user study, showing that **ncNet** achieves good accuracy in the **nvBench** ...

We present **ncNet**[1], an end-to-end solution using a Transformer-based sequence-to-sequence (seq2seq) model, which translates an NL query to a visualization. It adopts self-attention to generate a rich repre-

sentation (high dimensional vectors) of the input, **ncNet** enables smart visualization inference (*e.g.,* guessing the missing column, selecting a chart type, etc).

Besides making smarter inferences, a system can obtain more information (or "hint") from the user, by either obtaining a one-shot hint from the user or iteratively requiring more information (*a.k.a.* conversational systems) [6]. The hint can be of various formats, such as NL queries, tables, chart templates, with one main criterion to be easy-to-use. We propose to use chart templates as additional hints, where a user can specify the output to be a pie chart or a scatter plot with a simple click. In practice, chart templates have been widely used in all commercial products, including Tableau, Excel, Google Sheets, and so on. Due to the flexibility of the seq2seq model, we just treat the selected chart template C as another sequence, together with the NL query N and the dataset D as the input $\mathbf{X}$.

引言：摘要的扩充版，论文的总结版

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.*, NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.*, a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

Contributions. In this work, we make several contributions, including:

- proposing **ncNet**, a Transformer-based [53] seq2seq model for supporting NL2VIS;
- presenting a novel visualization-grammar, namely Vega-Zero, with the main purpose to simplify the NL2VIS translation using neural machine translation techniques. Moreover, transforming it to other visualization languages are straightforward;
- enhancing **ncNet** by allowing the user to select a chart template, which will be used to improve the translation accuracy;
- devising two optimization techniques: **attention forcing** for incorporating pre-defined domain knowledge and **visualization-aware translation** for better final visualization generation; and
- demonstrating that **ncNet** can well support NL2VIS with several use cases, as well as conducting a quantitative study.

引言：摘要的扩充版，论文的总结版

- 注意事项
  - 突出创新点（新问题？新技术？效果好？）
  - 突出技术深度/新颖性
  - 明确说出和现有方法的区别（老问题）

可以慢慢形成自己讲故事的风格。

文章结构比英语更重要

- 结构清晰
  - 总分结构
  - Leading text, Highlight
  - 贯穿全文的形象化的例子（尤其是将复杂理论或者方法时）
- 图文并茂
- 言简意赅
  - 避免重复的文本，尽量少用复合句

目标：站在读者的角度，把故事讲清楚、讲好。

文章结构比英语更重要

大纲

Learned Data-aware Image Representations of Line Charts for Similarity Search

YUYU LUO, Tsinghua University, China
YIHUI ZHOU, Tsinghua University, China
NAN TANG, QCRI, Qatar / HKUST (GZ), China
GUOLIANG LI*, Tsinghua University, China
CHENGLIANG CHAI, Beijing Institute of Technology, China
LEIXIAN SHEN, Tsinghua University, China

Finding line-chart images similar to a given line-chart image query is a common task in data exploration and image query systems, *e.g.*, finding similar trends in stock markets or medical Electroencephalography images. The state-of-the-art approaches consider either data-level similarity (when the underlying data is present) or image-level similarity (when the underlying data is absent).

In this paper, we study the scenario that during query time, only line-chart images are available. Our goal is to train a neural network that can turn these line-chart images into representations that are aware of the data used to generate these line charts, so as to learn *better* representations. Our key idea is that we can collect both data and line-chart images to learn such a neural network (at training step), while during query (or inference) time, we support the case that only line-chart images are provided. To this end, we present LineNet, a Vision Transformer-based Triplet Autoencoder model to learn data-aware image representations of line charts for similarity search. We design a novel pseudo labels selection mechanism to guide LineNet to capture both data-aware and image-level similarity of line charts. We further propose a diversified training samples selection strategy to optimize the learning process and improve the performance. We conduct both quantitative evaluation and case studies, showing that LineNet significantly outperforms the state-of-the-art methods for searching similar line-chart images.

解决方案概览（Overview / Framework）

- 一般都是总分架构，每章都是总分结构
- 第一段概述本章结构，下面概述整个framework，首先凝练整个论文贡献，论文整体怎么做的，细节ref到后面技术点1-3。
  - 介绍整个流程（workflow）和architecture
  - 概述和凝练每个技术点/module/component的用途、目的、解决了什么问题（细节ref到后面）
  - 论文除了framework最好有2-3个创新技术点
- 注意事项
  - 要图文并茂，通过图来解释架构
  - 读者读了这章基本了解了论文做法，不看细节也比基本明白了
  - 一定要自包含，不依赖于其他内容能看懂
  - 要有leading text，概述本小节解决什么问题，通过一章的leading text就了解了脉络

解决方案概览（Overview / Framework）

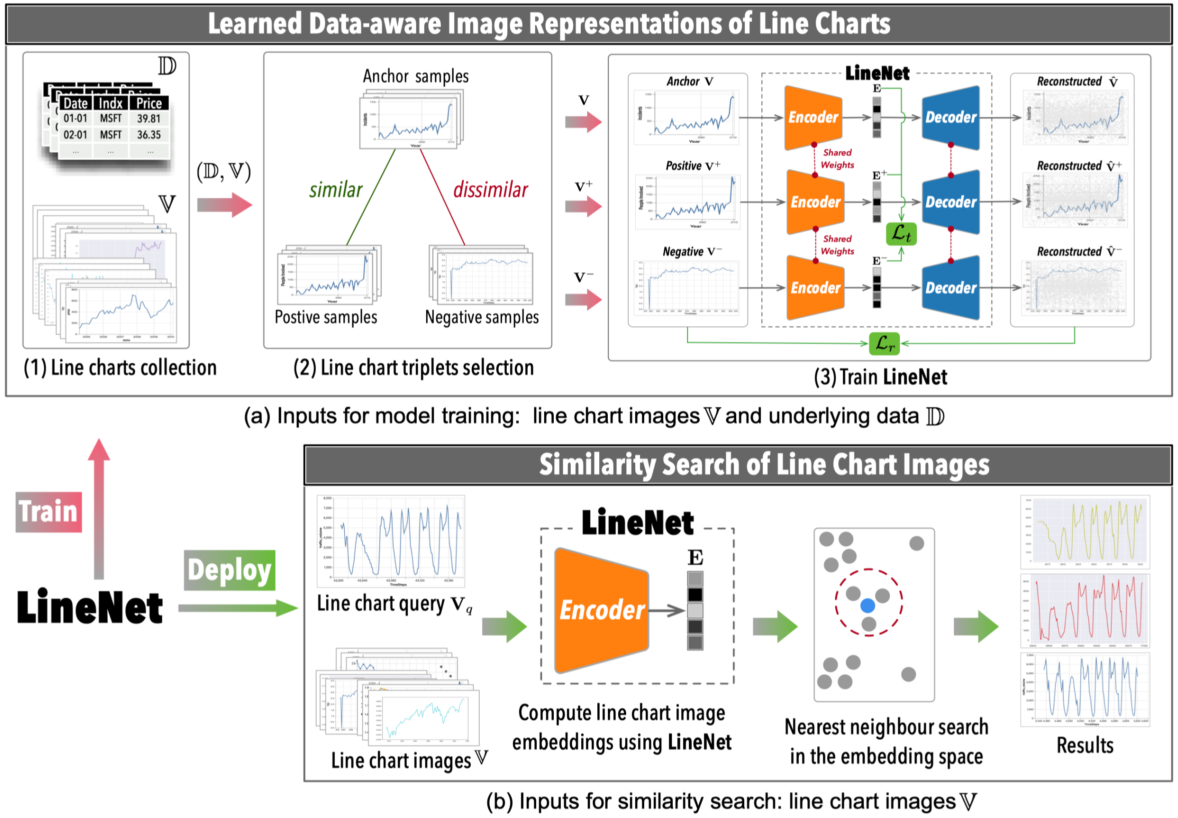


Fig. 2. The solution overview of LineNet.

3 AN OVERVIEW OF LINENET

3.1 Design and Train LineNet

Key Idea. To learn visualization representations of line charts, we need enough labeled similar/dissimilar images. However, *human-annotated examples* are far from being enough. To this end, we propose to generate "*similar/dissimilar*" pseudo labels *as a proxy for* the similarity of the corresponding line chart images **V** by computing **dist**$(\cdot,\cdot)$ based on their underlying data **D**. We then design a Vision Transformer-based Triplet Autoencoder framework, *i.e.*, LineNet. The key idea behind the framework is that we can learn the inherent image-level similarity by leveraging the Vision Transformer-based Autoencoder while capturing the data-level similarity based on the Triplet Network with the deep metric learning technique.

For training Triplet Autoencoder-based framework, each training sample is a triplet consisting of an anchor line chart image, a similar and a dissimilar one to the anchor (see Figure 2(a)-(2)).

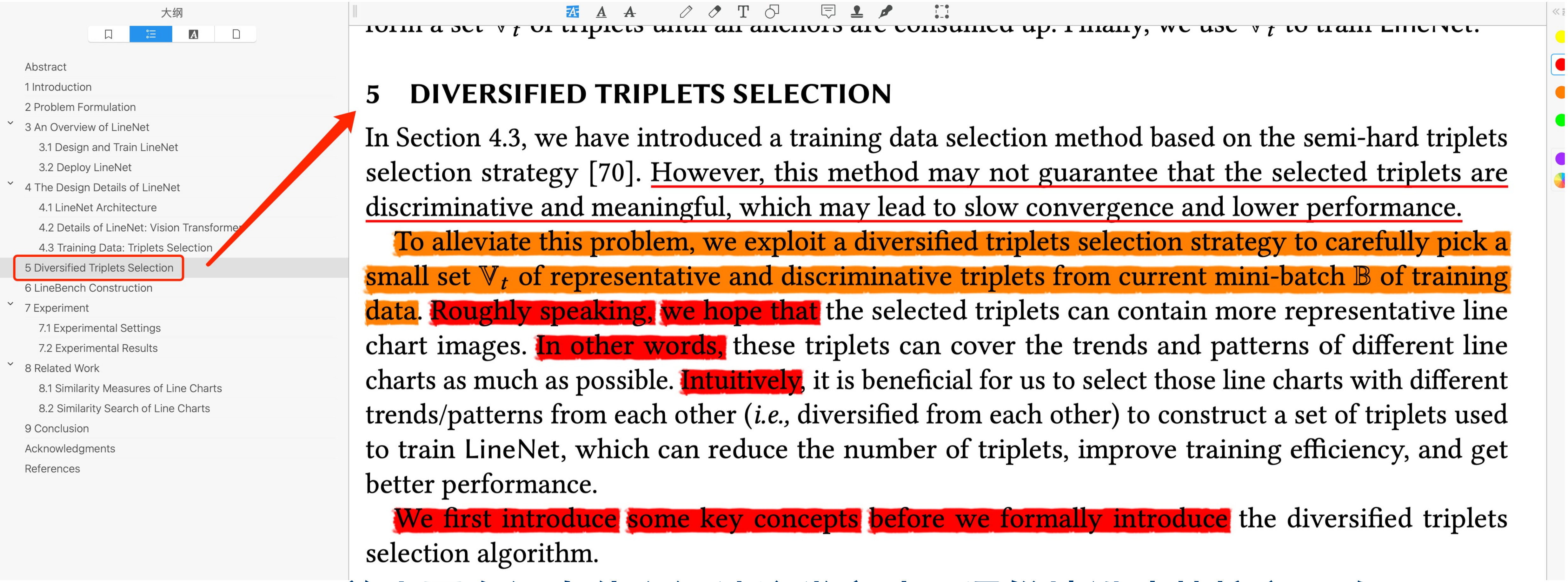
LineNet Overview. As shown in Figure 2(a)-(3), LineNet is built using a Triplet Autoencoder architecture, which consists of three identical autoencoders. For each autoencoder, it consists of an Encoder and a symmetric Decoder. At the training phase, LineNet takes as input a line chart image triplet $(\mathbf{V}, \mathbf{V}^+, \mathbf{V}^-)$, which is comprised of an anchor, a positive (*i.e.*, similar), and a negative (*i.e.*, dissimilar) line chart image. We compute **dist**$(\cdot,\cdot)$ based on their underlying data **D** to generate triplets, which indicates their *data similarity*. We will discuss how to select triplets in Section 4.3.

For learning data-aware image representations of line charts, we first use line chart image triplets to enforce LineNet to capture the "data similarity" between line chart images by leveraging the deep metric learning techniques [29, 70]. For example, as shown in Figure 3, the optimization objective of metric learning is to minimize the distance between the anchor (*i.e.*, blue point) and its positive samples (*i.e.*, green stars) while maximizing the distance between the anchor and negative samples

“技术创新”部分的写作

- 不要指望读者能很容易理解到你的技术点，如果很容易的话，那么这篇论文多半是没有多少技术创新/贡献。
- 要把读者当“傻子”，可以先介绍简单的，循序渐进展开。
- 合理使用图、表和公式。
- 尽量要用例子来解释难懂的技术部分。
- 逻辑一定要清晰连贯，否则读者很容易get lost。

“技术创新”部分的写作



大纲

Abstract
1 Introduction
2 Problem Formulation
3 An Overview of LineNet
3.1 Design and Train LineNet
3.2 Deploy LineNet
4 The Design Details of LineNet
4.1 LineNet Architecture
4.2 Details of LineNet: Vision Transformer
4.3 Training Data: Triplets Selection
5 Diversified Triplets Selection
6 LineBench Construction
7 Experiment
7.1 Experimental Settings
7.2 Experimental Results
8 Related Work
8.1 Similarity Measures of Line Charts
8.2 Similarity Search of Line Charts
9 Conclusion
Acknowledgments
References

form a set $\mathbb{V}_t$ of triplets until all anchors are consumed up. Finally, we use $\mathbb{V}_t$ to train LineNet.

5 DIVERSIFIED TRIPLETS SELECTION

In Section 4.3, we have introduced a training data selection method based on the semi-hard triplets selection strategy [70]. However, this method may not guarantee that the selected triplets are discriminative and meaningful, which may lead to slow convergence and lower performance.

To alleviate this problem, we exploit a diversified triplets selection strategy to carefully pick a small set $\mathbb{V}_t$ of representative and discriminative triplets from current mini-batch $\mathbb{B}$ of training data. Roughly speaking, we hope that the selected triplets can contain more representative line chart images. In other words, these triplets can cover the trends and patterns of different line charts as much as possible. Intuitively, it is beneficial for us to select those line charts with different trends/patterns from each other (*i.e.*, diversified from each other) to construct a set of triplets used to train LineNet, which can reduce the number of triplets, improve training efficiency, and get better performance.

We first introduce some key concepts before we formally introduce the diversified triplets selection algorithm.

首先要介绍为什么设计这类方法、通俗地讲清楚核心思路

“技术创新”部分的写作

也可以适当的炫技

negative) samples selection problem aims to pick a set of positive (*resp.* negative) samples $\mathbb{V}'_p \subseteq \mathbb{V}_p$ (*resp.* $\mathbb{V}'_n \subseteq \mathbb{V}_n$) with size k such that:

$$\mathbb{V}'_p = \mathop{\arg\max}_{\widetilde{\mathbb{V}}_p \subseteq \mathbb{V}_p, |\widetilde{\mathbb{V}}_p|=k} \mathcal{P}(\mathbf{V}_a, \widetilde{\mathbb{V}}_p) \tag{6}$$

$$\mathbb{V}'_n = \mathop{\arg\max}_{\widetilde{\mathbb{V}}_n \subseteq \mathbb{V}_n, |\widetilde{\mathbb{V}}_n|=k} \mathcal{N}(\mathbf{V}_a, \widetilde{\mathbb{V}}_n) \tag{7}$$

$$\mathcal{P}(\mathbf{V}_a, \widetilde{\mathbb{V}}_p) = \lambda \sum_{\mathbf{V}_i \in \widetilde{\mathbb{V}}_p} [R(\mathbf{V}_i, \mathbf{V}_a) + D(\mathbf{V}_i, \mathbf{V}_a)] + \frac{2(1-\lambda)}{k-1} \sum_{\mathbf{V}_i, \mathbf{V}_j \in \widetilde{\mathbb{V}}_p} D(\mathbf{V}_i, \mathbf{V}_j) \tag{8}$$

$$\mathcal{N}(\mathbf{V}_a, \widetilde{\mathbb{V}}_n) = \lambda \sum_{\mathbf{V}_i \in \widetilde{\mathbb{V}}_n} [(1 - R(\mathbf{V}_i, \mathbf{V}_a)) + (1 - D(\mathbf{V}_i, \mathbf{V}_a))] + \frac{2(1-\lambda)}{k-1} \sum_{\mathbf{V}_i, \mathbf{V}_j \in \widetilde{\mathbb{V}}_n} D(\mathbf{V}_i, \mathbf{V}_j) \tag{9}$$

where $\lambda \in [0, 1)$ is a parameter controlling the *trade-off* between *relevance* and *diversity*, which can be adjusted. The intuition behind this definition is that we aim to maximize $\mathcal{P}(\mathbf{V}_a, \widetilde{\mathbb{V}}_p)$ (*resp.* $\mathcal{N}(\mathbf{V}_a, \widetilde{\mathbb{V}}_n)$) so that we can derive a set of relevant and diversified line chart images with relatively high relevance to the anchor line chart image as well as high diversity. More concretely, the first two terms of $\mathcal{P}(\mathbf{V}_a, \widetilde{\mathbb{V}}_p)$ (*resp.* $\mathcal{N}(\mathbf{V}_a, \widetilde{\mathbb{V}}_n)$) aim to select the semi-hard triplets [70], while the third term adjusts the selection process by considering the *diversity* among positive (*resp.* negative) line charts. We scale down the diversity part by $\frac{2(1-\lambda)}{k-1}$ because there are k items in the first two terms while $\frac{k(k-1)}{2}$ items in the diversity terms.

Unfortunately, the diversified top-k positive/negative samples selection problem is NP-hard. Because our problem can be reduced to the NP-hard *max-sum dispersion* problem [26] when $\lambda = 0$.

Therefore, we devise a greedy algorithm to select top-k positive/negative samples effectively and efficiently, as shown in Algorithm 2 (Lines 9-15). The key idea behind the scenes is that it first greedily constructs a set $\mathbb{V}_a$ of diversified anchors. Next, the algorithm incrementally and greedily selects a positive (*resp.* negative) sample to diversified results $\mathbb{V}'_p$ (*resp.* $\mathbb{V}'_n$).

Diversified Triplets Selection Strategy. Based on the above discussion, we now introduce our diversified triplets selection algorithm. The pseudo code is shown in Algorithm 2. It takes as input a mini-batch of line charts $\mathbb{B}$ (training data), k for the number of anchors (triplets). First, it clusters the $\mathbb{B}$ into k clusters and pick one random point from each cluster to form the diversified anchor set $\mathbb{V}_a$ (Lines 2-5). Next, it iterates each anchor $\mathbf{V}_a$ to select a set of diversified positive/negative samples to anchor $\mathbf{V}_a$ (Lines 6-15). In each loop, it first generates a set of candidate positive/negative samples based on the precomputed pseudo labels, which is straightforward to perform (Line 7). Next, it aims at maximizing the Eq. (10) to select a diversified positive sample and to build the positive samples set $\mathbb{V}'_p$ incrementally (Lines 9-11). This process will be repeated until k positive samples are derived. The selection of diversified negative samples is similar, we omit the discussion due to the space constraint. Finally, it constructs the triplets based on the $\mathbf{V}_a$ and the selected diversified positive (*resp.* negative) set $\mathbb{V}'_p$ (*resp.* $\mathbb{V}'_n$) (Line 15). After iterating all anchors in $\mathbb{V}_a$, the algorithm generates a set $\mathbb{V}_t$ of diversified triplets to drive the training of LineNet.

$$\mathcal{P}'(\mathbf{V}_i, \mathbf{V}_a) = \lambda[R(\mathbf{V}_i, \mathbf{V}_a) + D(\mathbf{V}_i, \mathbf{V}_a)] + \frac{1-\lambda}{k} \sum_{\mathbf{V}_j \in \mathbb{V}'_p} D(\mathbf{V}_i, \mathbf{V}_j) \tag{10}$$

$$\mathcal{N}'(\mathbf{V}_i, \mathbf{V}_a) = \lambda[1 - R(\mathbf{V}_i, \mathbf{V}_a) + 1 - D(\mathbf{V}_i, \mathbf{V}_a)] + \frac{1-\lambda}{k} \sum_{\mathbf{V}_j \in \mathbb{V}'_n} D(\mathbf{V}_i, \mathbf{V}_j) \tag{11}$$

Algorithm 2: DiversifiedTripletsSelection

Input: A mini-batch of line chart $\mathbb{B}$, k;
Output: A mini-batch of diversified triplets $\mathbb{V}_t$;
// 1. Diversified k anchors selection
1 $\mathbb{V}_t, \mathbb{V}_a \leftarrow \emptyset, \emptyset$; // Initialize the triplet and anchor set
2 $\{C_1, C_2, \ldots, C_k\} \leftarrow$ KmeansClustering($\mathbb{B}, k$);
3 **for** C_i *in* $\{C_1, C_2, \ldots, C_k\}$ **do**
4     $\mathbf{V}_a \leftarrow$ RandomPick(C_i);
5     $\mathbb{V}_a \leftarrow \mathbb{V}_a \cup \{\mathbf{V}_a\}$;
// 2. Diversified top-k triplets selection
6 **for** $V_a \in \mathbb{V}_a$ **do**
    // candidate $\mathbb{V}_p$,$\mathbb{V}_n$ generation based on $\mathbf{V}_a$
7     $\mathbb{V}_p, \mathbb{V}_n \leftarrow$ CandidateSampleGeneration($\mathbf{V}_a$);
8     $\mathbb{V}'_p, \mathbb{V}'_n \leftarrow \emptyset, \emptyset$;
    // diversified positive samples selection
9   **while** $|\mathbb{V}'_p| \le k$ **do**
10         $\mathbf{V}'_p \leftarrow \arg\max_{\mathbf{V}_i \in \mathbb{V}_p} \mathcal{P}'(\mathbf{V}_i, \mathbf{V}_a)$;
11         $\mathbb{V}'_p \leftarrow \mathbb{V}'_p \cup \{\mathbf{V}'_p\}, \mathbb{V}_p \leftarrow \mathbb{V}_p \setminus \{\mathbf{V}'_p\}$;
    // diversified negative samples selection
12   **while** $|\mathbb{V}'_n| \le k$ **do**
13         $\mathbf{V}'_n \leftarrow \arg\max_{\mathbf{V}_i \in \mathbb{V}_n} \mathcal{N}'(\mathbf{V}_i, \mathbf{V}_a)$;
14         $\mathbb{V}'_n \leftarrow \mathbb{V}'_n \cup \{\mathbf{V}'_n\}, \mathbb{V}_n \leftarrow \mathbb{V}_n \setminus \{\mathbf{V}'_n\}$;
    // Construct diversified triplets
15     $\mathbb{V}_t \leftarrow \mathbb{V}_t \cup$ ConstructTriplets($\mathbf{V}_a, \mathbb{V}'_p, \mathbb{V}'_n$);
16 **return** $\mathbb{V}_t$;

“技术创新”部分的写作

Diversified Triplets Selection Strategy. Based on the above discussion, we now introduce our diversified triplets selection algorithm. The pseudo code is shown in Algorithm 2. It takes as input a mini-batch of line charts $\mathbb{B}$ (training data), k for the number of anchors (triplets). First, it clusters the $\mathbb{B}$ into k clusters and pick one random point from each cluster to form the diversified anchor set $\mathbb{V}_a$ (Lines 2-5). Next, it iterates each anchor $\mathbf{V}_a$ to select a set of diversified positive/negative samples to anchor $\mathbf{V}_a$ (Lines 6-15). In each loop, it first generates a set of candidate positive/negative samples based on the precomputed pseudo labels, which is straightforward to perform (Line 7). Next, it aims at maximizing the Eq. (10) to select a diversified positive sample and to build the positive samples set $\mathbb{V}'_p$ incrementally (Lines 9-11). This process will be repeated until k positive samples are derived. The selection of diversified negative samples is similar, we omit the discussion due to the space constraint. Finally, it constructs the triplets based on the $\mathbf{V}_a$ and the selected diversified positive (*resp.* negative) set $\mathbb{V}'_p$ (*resp.* $\mathbb{V}'_n$) (Line 15). After iterating all anchors in $\mathbb{V}_a$, the algorithm generates a set $\mathbb{V}_t$ of diversified triplets to drive the training of LineNet.

Algorithm 2: DiversifiedTripletsSelection

Input: A mini-batch of line chart $\mathbb{B}$, k;
Output: A mini-batch of diversified triplets $\mathbb{V}_t$;
// 1. Diversified k anchors selection
1 $\mathbb{V}_t, \mathbb{V}_a \leftarrow \emptyset, \emptyset$; // Initialize the triplet and anchor set
2 $\{C_1, C_2, ..., C_k\} \leftarrow$ KmeansClustering($\mathbb{B}, k$);
3 **for** C_i *in* $\{C_1, C_2, ..., C_k\}$ **do**
4 $\quad \mathbf{V}_a \leftarrow$ RandomPick(C_i);
5 $\quad \mathbb{V}_a \leftarrow \mathbb{V}_a \cup \{\mathbf{V}_a\}$;
// 2. Diversified top-k triplets selection
6 **for** $\mathbf{V}_a \in \mathbb{V}_a$ **do**
$\quad$ // candidate $\mathbb{V}_p$,$\mathbb{V}_n$ generation based on $\mathbf{V}_a$
7 $\quad \mathbb{V}_p, \mathbb{V}_n \leftarrow$ CandidateSampleGeneration($\mathbf{V}_a$);
8 $\quad \mathbb{V}'_p, \mathbb{V}'_n \leftarrow \emptyset, \emptyset$;
$\quad$ // diversified positive samples selection
9 $\quad$ **while** $|\mathbb{V}'_p| \leq k$ **do**
10 $\quad\quad \mathbf{V}'_p \leftarrow \arg\max_{\mathbf{V}_i \in \mathbb{V}_p} \mathcal{P}'(\mathbf{V}_i, \mathbf{V}_a)$;
11 $\quad\quad \mathbb{V}'_p \leftarrow \mathbb{V}'_p \cup \{\mathbf{V}'_p\}, \mathbb{V}_p \leftarrow \mathbb{V}_p \setminus \{\mathbf{V}'_p\}$;
$\quad$ // diversified negative samples selection
12 $\quad$ **while** $|\mathbb{V}'_n| \leq k$ **do**
13 $\quad\quad \mathbf{V}'_n \leftarrow \arg\max_{\mathbf{V}_i \in \mathbb{V}_n} \mathcal{N}'(\mathbf{V}_i, \mathbf{V}_a)$;
14 $\quad\quad \mathbb{V}'_n \leftarrow \mathbb{V}'_n \cup \{\mathbf{V}'_n\}, \mathbb{V}_n \leftarrow \mathbb{V}_n \setminus \{\mathbf{V}'_n\}$;
$\quad$ // Construct diversified triplets
15 $\quad \mathbb{V}_t \leftarrow \mathbb{V}_t \cup$ ConstructTriplets($\mathbf{V}_a, \mathbb{V}'_p, \mathbb{V}'_n$);
16 **return** $\mathbb{V}_t$;

如果有算法，最好提供伪代码并进行对应解释

“技术创新”部分的写作

Algorithm 2: GSS: GREEDYSUBGRAPHSELECTION

Input: An ERG $\mathcal{G}(\mathcal{V},\mathcal{E},w)$, k $(2 < k < |\mathcal{V}|)$;
Output: Subgraph $G(V,E)$;
$\mathcal{B}_{max} \leftarrow 0$; Collection of vertex set $\mathcal{C} \leftarrow null$;
for each v in $\mathcal{V}$ **do** $m[v] \leftarrow null$;
$\mathcal{G}(\mathcal{V},\mathcal{E},b) \leftarrow$ *EstimatedBenefit*$(\mathcal{G}(\mathcal{V},\mathcal{E},w))$;
Sort $\mathcal{E}$ by b in descending order;
for *each edge* (v,v') *in* $\mathcal{E}$ **do**
  if $m[v] = m[v'] = null$ **then**
    Add $\{v,v'\}$ to $\mathcal{C}$; *// Case 1*
    $m[v] \leftarrow \{v,v'\}$; $m[v'] \leftarrow \{v,v'\}$;
    continue ;
  if $m[v] = null$ **then**
    $v_f \leftarrow v$; $v_t \leftarrow v'$; *// Case 2*
  else
    $v_f \leftarrow v'$; $v_t \leftarrow v$; *// Case 3*
  Add v_f into $m[v_t]$;
  $m[v_f] \leftarrow m[v_t]$;
  if $|m[v_t]| = k$ **then**
    Get the $G'(V',E')$ induced by vertices in $m[v_t]$;
    $\mathcal{B} \leftarrow$ sum of the benefit of all edges in $G'(V',E')$;
    if $\mathcal{B} > \mathcal{B}_{max}$ **then**
      $G(V,E) \leftarrow G'(V',E')$;
      $\mathcal{B}_{max} \leftarrow \mathcal{B}$;
    for each vertex u in $G'(V',E')$ **do** $m[u] \leftarrow null$;
return $G(V,E)$;

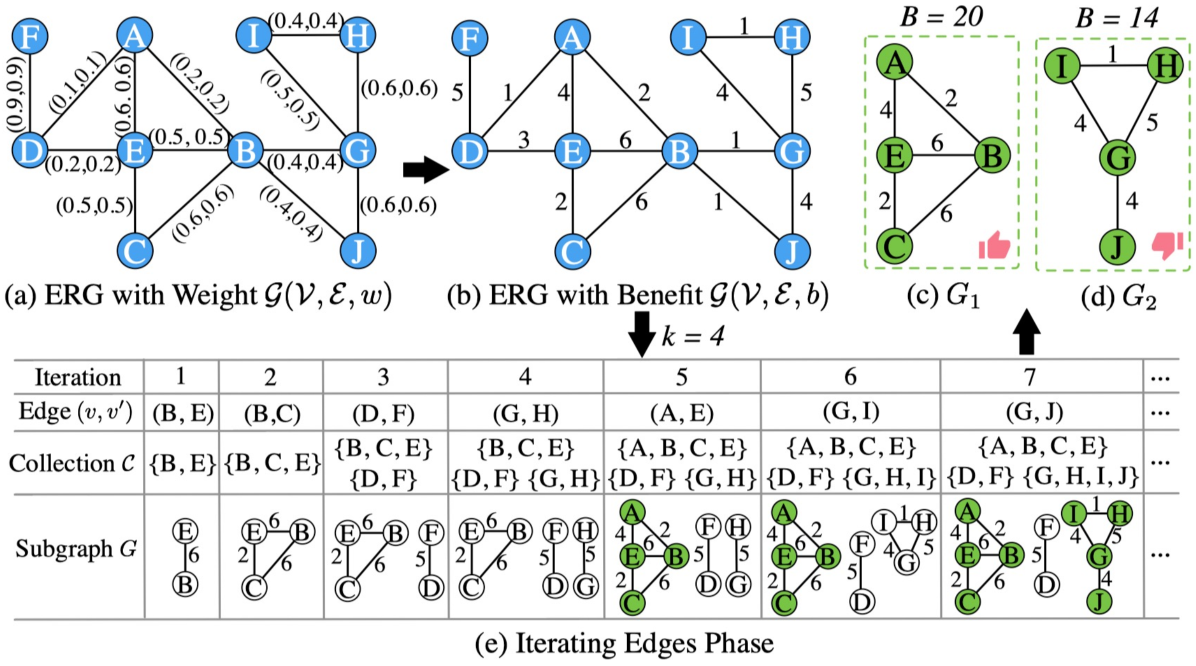


Fig. 7. Example of CQG Selection

复杂的算法可以增加例子，用于阐述理解

相关工作 Related Work

- 引用相关论文，说出具体区别
- 多夸奖别人、承认别人的贡献
  - 这是科学精神的体现
  - 你的审稿人可能就是该文的作者
- 保证语言正确，不要曲解相关工作的真实意思
- 严谨在相关工作描述部分，直接复制它文的内容

相关工作 Related Work

8 Related Works

There are many works on exploiting heterogeneous annotation data to improve various NLP tasks. Jiang et al. (2009) proposed a stacking-based model which could train a model for one specific desired annotation criterion by utilizing knowledge from corpora with other heterogeneous annotations. Sun and Wan (2012) proposed a structure-based stacking model to reduce the approximation error, which makes use of structured features such as sub-words. These models are unidirectional aid and also suffer from error propagation problem.

承认开创性的工作

不足之处

Qiu et al. (2013) used multi-tasks learning framework to improve the performance of POS tagging on two heterogeneous datasets. Li et al. (2015) proposed a coupled sequence labeling model which could directly learn and infer two heterogeneous annotations. Chao et al. (2015) also utilize multiple corpora using coupled sequence labeling model. These methods adopt the shallow classifiers, therefore suffering from the problem of defining shared features.

改进的工作

不足之处

Our proposed models use deep neural networks, which can easily share information with hidden shared layers. Chen et al. (2016) also adopted neural network models for exploiting heterogeneous annotations based on neural multi-view model, which can be regarded as a simplified version of our proposed models by removing private hidden layers.

和直接竞争模型的区别

Unlike the above models, we design three sharing-private architectures and keep shared layer to extract criterion-invariance features by introducing adversarial training. Moreover, we fully exploit eight corpora with heterogeneous segmentation criteria to model the underlying shared information.

我们工作的共享和优点

结论

- 总结文章贡献
- 不要与引言重复
- 和引言不同的语调
  - 引言：读者还不知道技术细节
  - 结论：读者已经看完了论文

如何取标题？

➢ 用一句话概括你所做的工作

- 模板：问题+方法+贡献
- 可以适当地“别出心裁”

Interactive Cleaning for Progressive Visualization through Composite Questions

Natural Language to Visualization by Neural Machine Translation

RW-tree : A Learned Workload-aware Framework for R-tree Construction

DeepEye: Towards **Automatic Data Visualization**

Learned Data-aware Image Representations of Line Charts for Similarity Search

论文撰写的一些建议

1. 组织全文的叙事逻辑（禁忌想到什么写什么）
   - 组织故事线（全文结构）
   - 确定每一个部分写什么内容
   - 不同部分之间的逻辑联系和起承转合
2. 下笔填充内容
   - 写完一段之后，请读一遍（发现低级错误）
   - 多使用简单句，把事情讲清楚
   - 注意一定要用好连接词。
   - 用ChatGPT、Grammarly检查语法。

初稿严禁直接使用用ChatGPT生成！！！
初稿严禁直接使用用ChatGPT生成！！！
初稿严禁直接使用用ChatGPT生成！！！
初稿严禁直接使用用ChatGPT生成！！！
初稿严禁直接使用用ChatGPT生成！！！
初稿严禁直接使用用ChatGPT生成！！！

写论文之前

- 做好PPT与导师充分讨论，如果导师不直接上的话
- 列好提纲（可以让导师帮忙），尽量做填空题
- 做好规划，万（每）事（章/段）开头难

一些常见的问题

- 第三人称单数、冠词、一句话不能有俩动词，不要用复杂句子
- 用词要地道合理，不要随便查字典就放上去，可以保守一点
- 符号简洁清晰，定义了就要用，别用着用着就乱了
- 画图要用心、美观，一个图想两天，画一天很正常很正常
- 两句话之间加不上连词的时候，就要小心了！
- 要用心体会老师/师兄的指导，不要经常重复犯错误
- 逻辑还是第一位的，千万不要让人去猜逻辑、猜内容，be specific

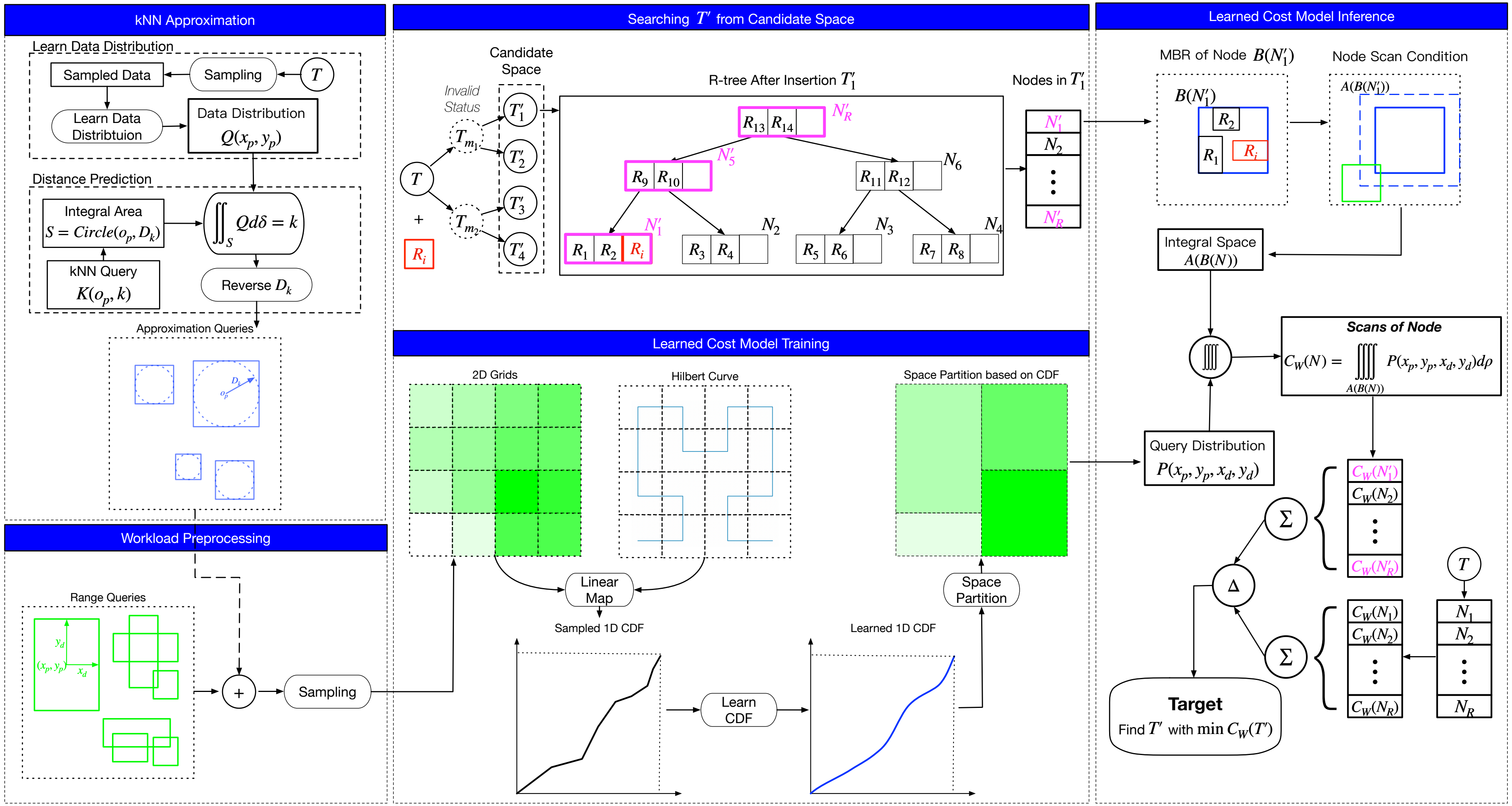
论文排版的一些建议

- 用对相关学术论文的投稿模板
- 尽量学习使用LaTeX进行论文排版
- 在线协作进行LaTeX编辑（Overleaf）
- **一定要重视画图！画图！画图！**

https://color.adobe.com/zh/explore

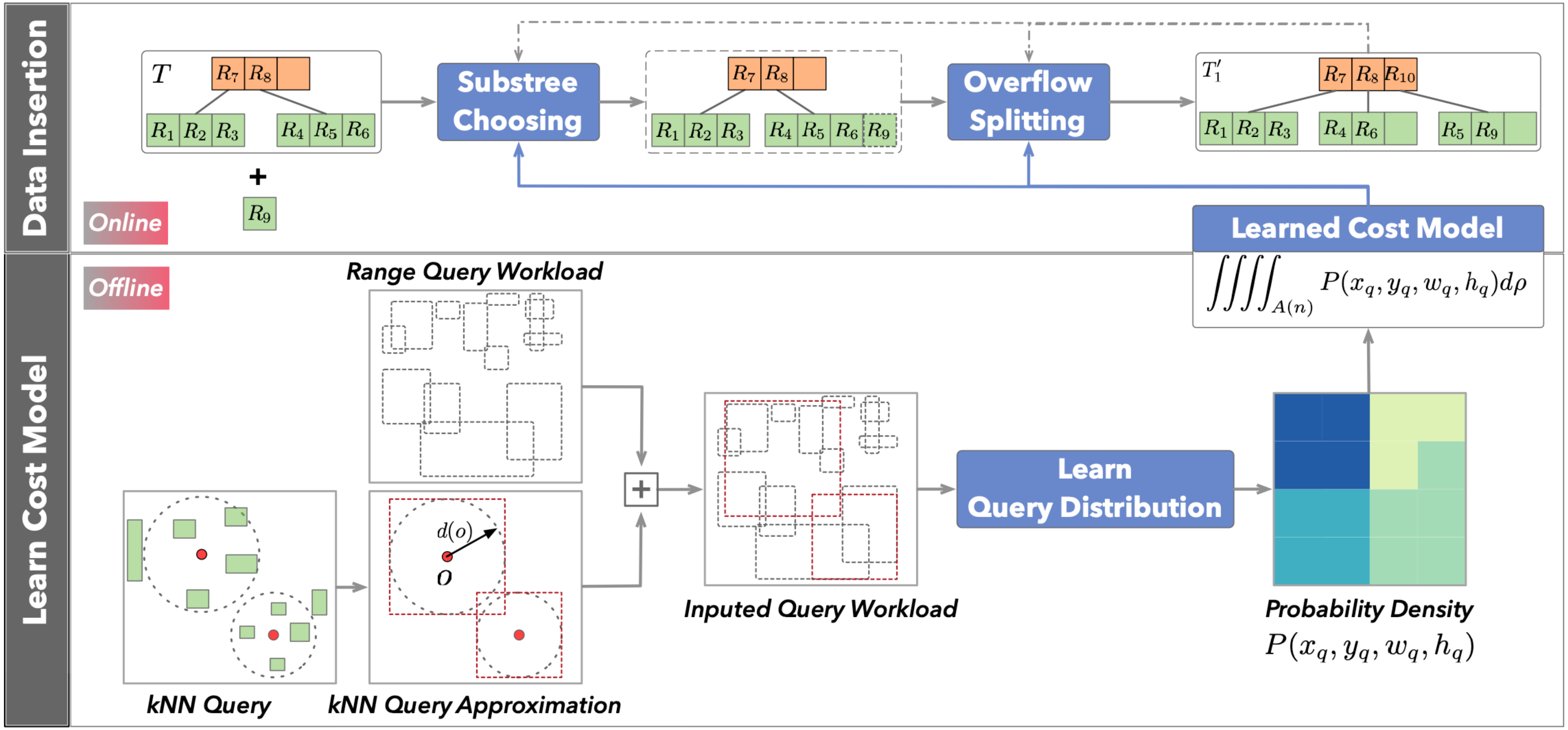
https://www.omnigroup.com/omnigraffle

要重视论文中的插图！



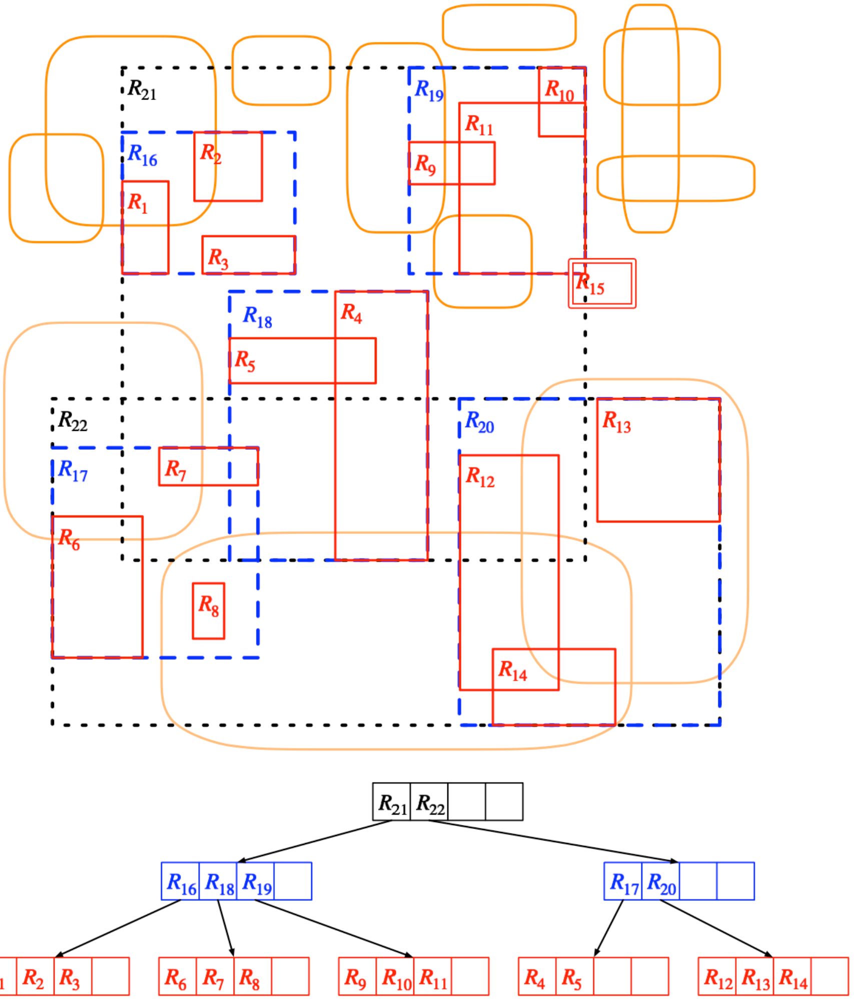
kNN Approximation
Learn Data Distribution
Sampled Data
Sampling
Learn Data Distribtuion
Data Distribution
Distance Prediction
Integral Area
kNN Query
Reverse D_k
Approximation Queries
Workload Preprocessing
Range Queries
Sampling
Searching T' from Candidate Space
Candidate Space
Invalid Status
R-tree After Insertion T'_1
Nodes in T'_1
Learned Cost Model Training
2D Grids
Hilbert Curve
Space Partition based on CDF
Linear Map
Space Partition
Sampled 1D CDF
Learn CDF
Learned 1D CDF
Learned Cost Model Inference
MBR of Node $B(N'_1)$
Node Scan Condition
Integral Space
Scans of Node
Query Distribution
Target
Find T' with min $C_W(T')$

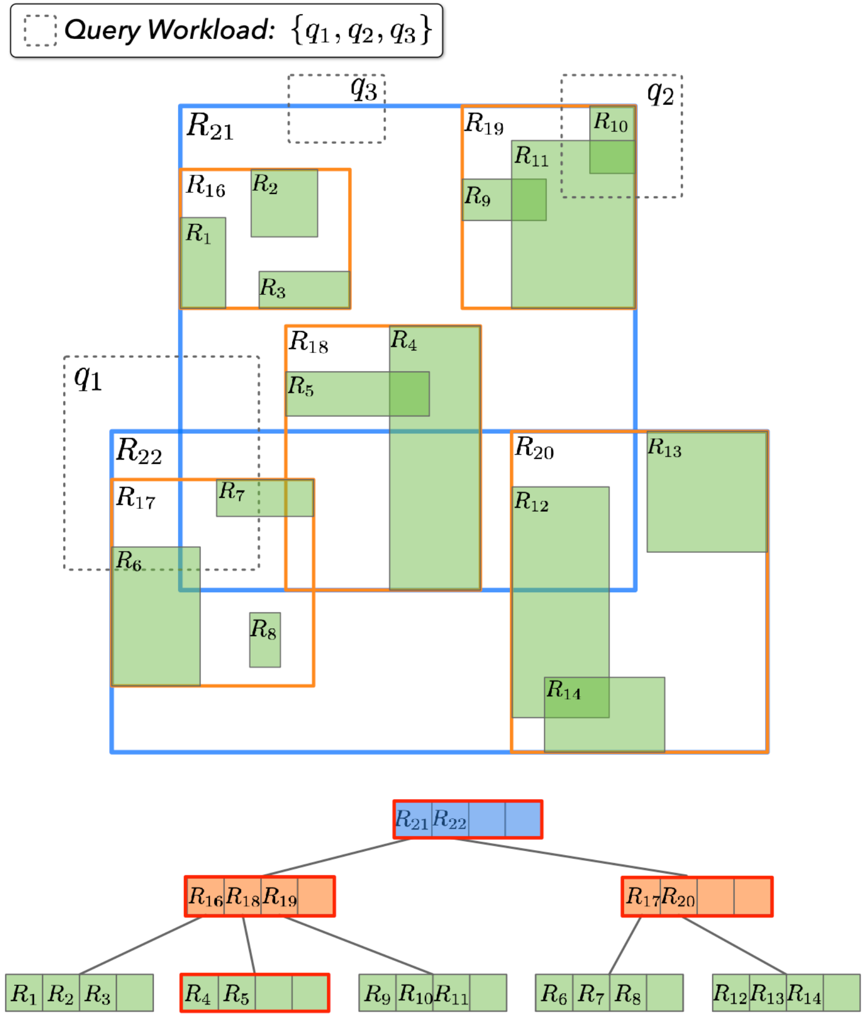
要重视论文中的插图！



Data Insertion
Online
T
R_7 R_8
R_1 R_2 R_3 R_4 R_5 R_6
+
R_9
Substree Choosing
R_7 R_8
R_1 R_2 R_3 R_4 R_5 R_6 R_9
Overflow Splitting
T_1'
R_7 R_8 R_{10}
R_1 R_2 R_3 R_4 R_6 R_5 R_9
Learned Cost Model
$\iiiint_{A(n)} P(x_q, y_q, w_q, h_q)d\rho$
Learn Cost Model
Offline
Range Query Workload
$d(o)$
o
kNN Query
kNN Query Approximation
Inputed Query Workload
Learn Query Distribution
Probability Density
$P(x_q, y_q, w_q, h_q)$

要重视论文中的插图！


R21
R22
R16 R18 R19
R17 R20
R1 R2 R3
R6 R7 R8
R9 R10 R11
R4 R5
R12 R13 R14
R15


Query Workload: {q1, q2, q3}
q1
q2
q3
R21 R22
R16 R18 R19
R17 R20
R1 R2 R3
R4 R5
R9 R10 R11
R6 R7 R8
R12 R13 R14

审稿人的角度

- **我希望看到:**
  1. Novel Problem: 新问题，尤其是定义一个有用的新问题
  2. Novel Method：解决旧/新问题的新颖有效的方法
  3. Nice Story: 好的写作、吸引人的故事、逻辑性强
  4. Nice Presentation：排版美观、插图漂亮（避免低级的语法/排版/拼写错误）
- **我讨厌看到:**
  1. Old school problem, with simple combination of existing methods
  2. Worse Presentation: writing, figures
  3. Experiment： without strong baselines, inappropriate settings

论文投稿与被拒！

- 伤心一两天就让他过去吧！
- 不要气馁：
  - 任何人的论文都有可能被拒！
  - 顶会顶刊接受率一般只有20%左右！
- 找原因，补短板，适当根据Reviewer的意见修改打磨！

Thanks!

Have fun, learn stuff

Dr. Yuyu LUO
Data Science and Analytics Thrust
Information Hub, HKUST(GZ)
yuyuluo@hkust-gz.edu.cn
http://luoyuyu.vip