大数据分析实践

Introduction

Qiong Zeng (曾琼)

qiong.zn@sdu.edu.cn

胜利日将举行盛大阅兵

胜利日 1945-2025

天安门阅兵是纪念活动重要组成部分

阅兵式里的大数据分析系统

- **数字力量**：基于大数据分析的毫米级精度方阵（实现多单位协同指挥与精准行动）、北斗定位系统（提供实时精准定位与数据支撑）、量子通信车（保障数据传输效率与安全）、东风-51导弹（依托数据链实现智能目标锁定与打击）、智能防空体系（通过大数据实时处理威胁信息）、无人机集群控制系统（基于算法协同执行复杂任务）。
- **装备安全保障**：高空高速无人机可依托数据分析精准识别与压制航母战斗群；智能水雷通过大数据识别技术准确判别敌我舰船，实现自主攻防决策。
- **军民融合**：民企参与研发的北斗三号芯片，显著提升定位精度与数据处理能力，为系统提供强大底层算力支持。

https://baijiahao.baidu.com/s?id=1841881729958096787&wfr=spider&for=pc
https://baijiahao.baidu.com/s?id=1841663536417320849&wfr=spider&for=pc

目录

学习目标

知识目标

能够阐述清楚大数据、大数据分析实践的必要性，了解本门课程的考核方式及要求

重点

能力目标

可根据课程要求规划本学期项目实施计划

重点

素质目标

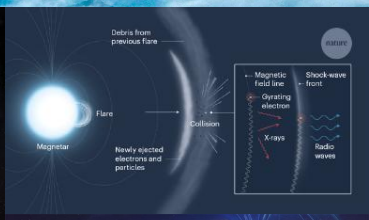
能够以批判性思维以及发展的眼光看待大数据分析

难点

世界最大：中国天眼FAST

Five-hundred-meter Aperture Spherical radio Telescope

- 4450块反射面板单元
- 突破了射电望远镜的百米极限
- 拥有30个足球场大的接收面积
- 与号称“地面最大的机器”的德国波恩100米望远镜相比灵敏度提高约10倍



中国天眼FAST

如何探索和处理海量数据以挖掘有效信息?

主观题 10分

设置

请联系阅兵式中可能有的大数据分析技术，提2-5个关键词，描述什么数据是大数据?

作答

Definition of Big Data

Nature 2008

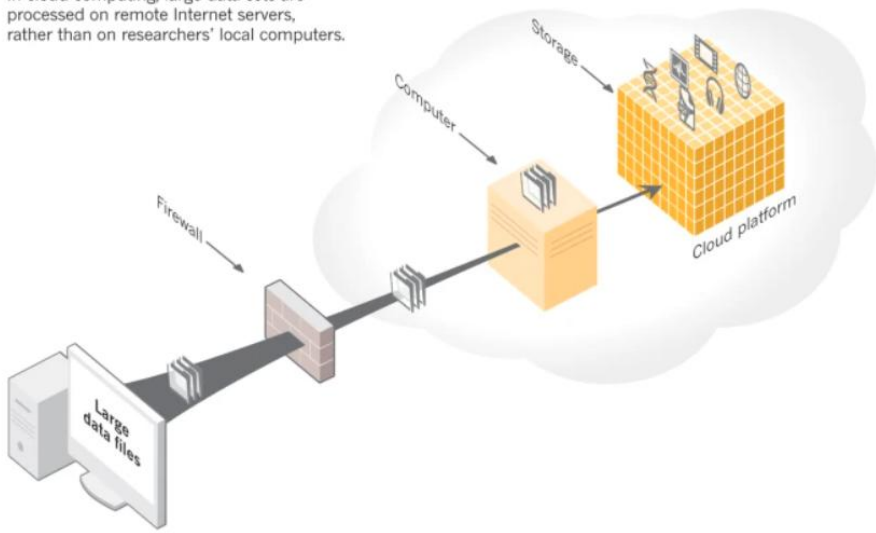
Science in the Petabyte Era

Science 2011

Dealing with data

Nature, 2013

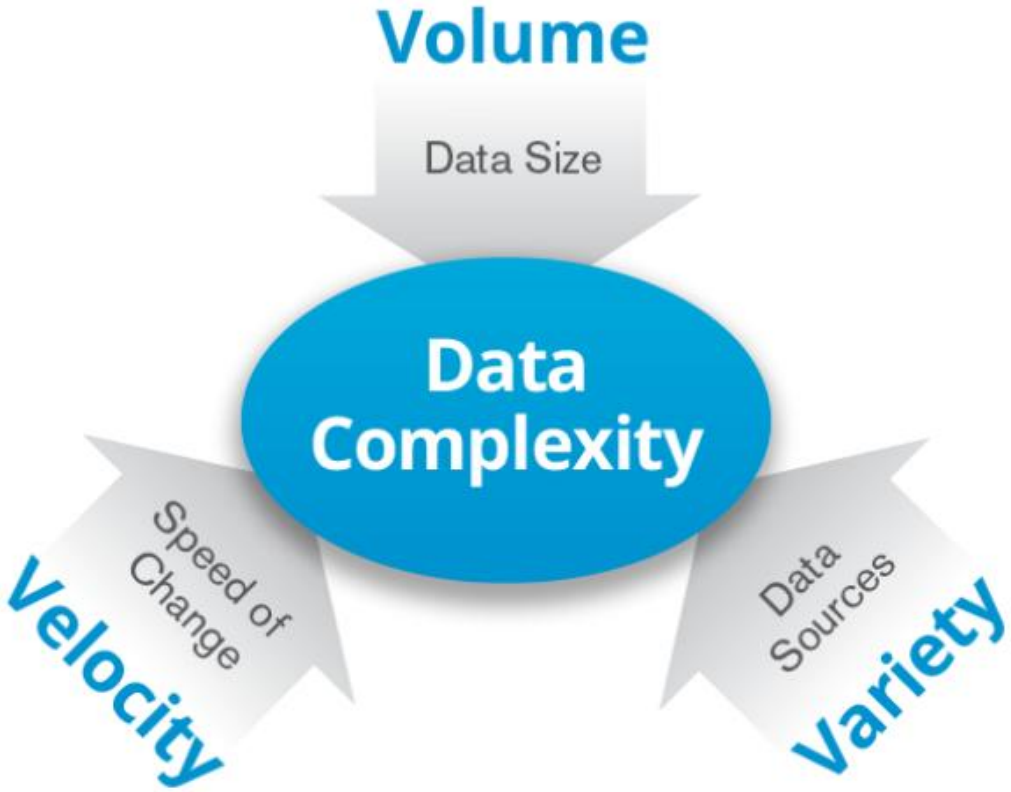
The big challenges of big data



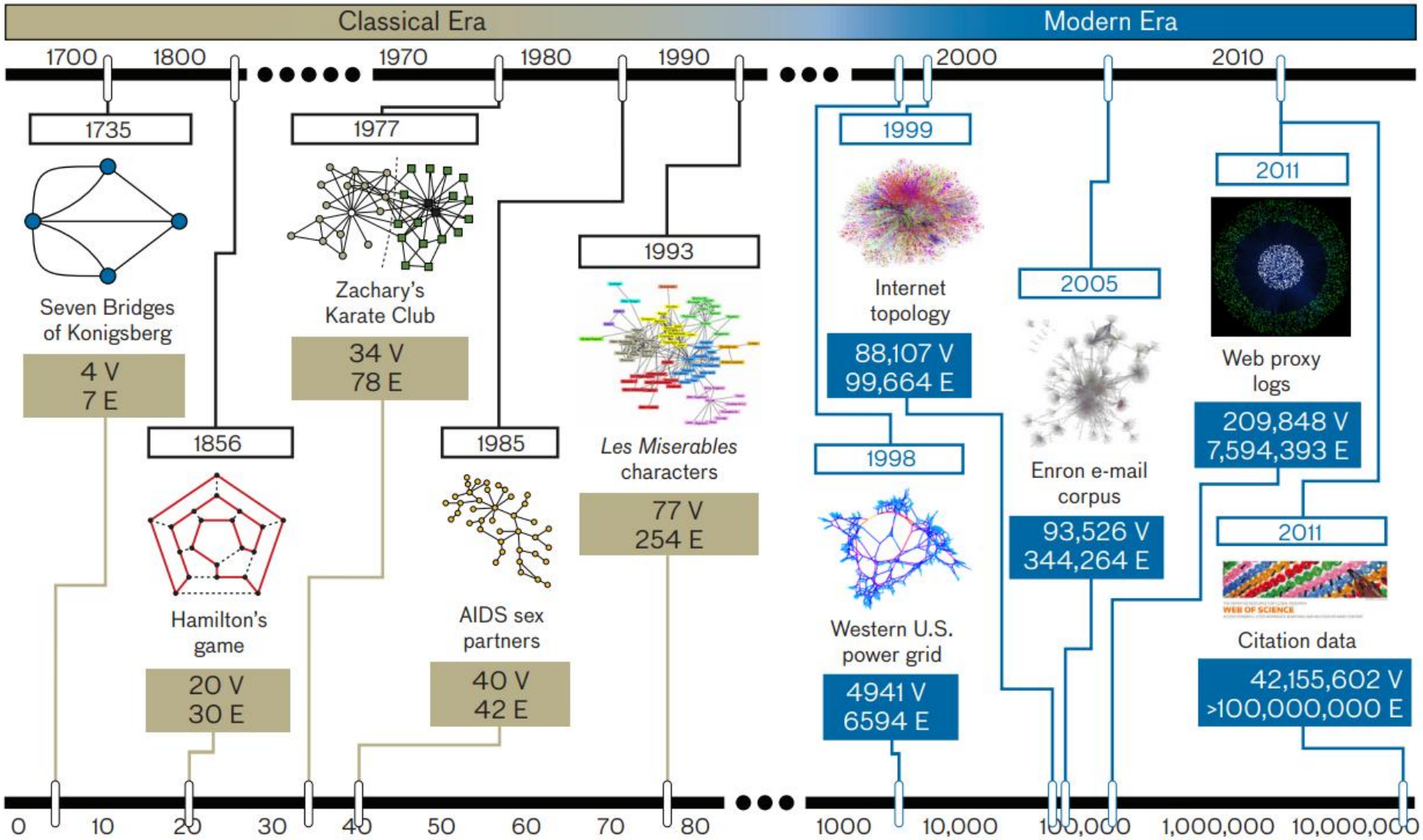
Definition of Big Data

"*Big data is* ***high-volume, high-velocity and high-variety*** *information assets that demand* ***cost-effective, innovative*** *forms of information processing for* ***enhanced insight and decision making***."

Gartner



High Volume


Classical Era
Modern Era
1700
1800
1970
1980
1990
2000
2010
1735
Seven Bridges of Konigsberg
4 V
7 E
1856
Hamilton's game
20 V
30 E
1977
Zachary's Karate Club
34 V
78 E
1985
AIDS sex partners
40 V
42 E
1993
Les Miserables characters
77 V
254 E
1999
Internet topology
88,107 V
99,664 E
1998
Western U.S. power grid
4941 V
6594 E
2005
Enron e-mail corpus
93,526 V
344,264 E
2011
Web proxy logs
209,848 V
7,594,393 E
2011
Citation data
42,155,602 V
>100,000,000 E
0
10
20
30
40
50
60
70
80
1000
10,000
100,000
1,000,000
10,000,000

High Velocity

High Variety

多源异构

海量高维

动态模糊

全局稀疏

局部冗余

时空断裂

Why Big Data?

- More data are being collected
- Open source code
- Commodity hardware / Cloud
- Domain requirements

- **High-Volume**
- **High-Velocity**
- **High-Variety**

Data Analytics

Improving human intelligence with machine intelligence

Big Data Analytics

IMDb: Movie Data

IMDb | Find Movies, TV shows, Celebrities and more... | All | Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist

More at IMDbPro »

It's a Wonderful Life (1946) Top 5000

Approved 130 min - Drama | Family | Fantasy - 7 January 1947 (USA)

Your rating: ★★★★★★★★★★ -/10

8.7

Ratings: **8.7**/10 from 202,743 users

Reviews: 632 user | 187 critic

An angel helps a compassionate but despairingly frustrated businessman by showing what life would have been like if he never existed.

Director: Frank Capra

Writers: Frances Goodrich (screenplay), Albert Hackett (screenplay), 4 more credits »

Stars: James Stewart, Donna Reed, Lionel Barrymore | See full cast and crew »

+ Watchlist | Watch Trailer | Share...

Details

Country: USA

Language: English

Release Date: 7 January 1947 (USA) See more »

Also Known As: The Greatest Gift See more »

Filming Locations: California, USA See more »

Box Office

Budget: $3,180,000 (estimated)

Opening Weekend: £49,845 (UK) (19 December 2008)

Gross: £682,222 (UK) (24 December 2010)

See more »

Company Credits

Production Co: Liberty Films (II) See more »

Show detailed company contact information on IMDbPro »

Technical Specs

Runtime: 130 min | 118 min (DVD edition)

Sound Mix: Mono (RCA Sound System)

Color: Color (colorized) | Black and White

Aspect Ratio: 1.37 : 1

See full technical specs »

主观题 10分

设置

假如你是一位数据分析师，请根据这些电影票房数据，提出一个你感兴趣的问题。

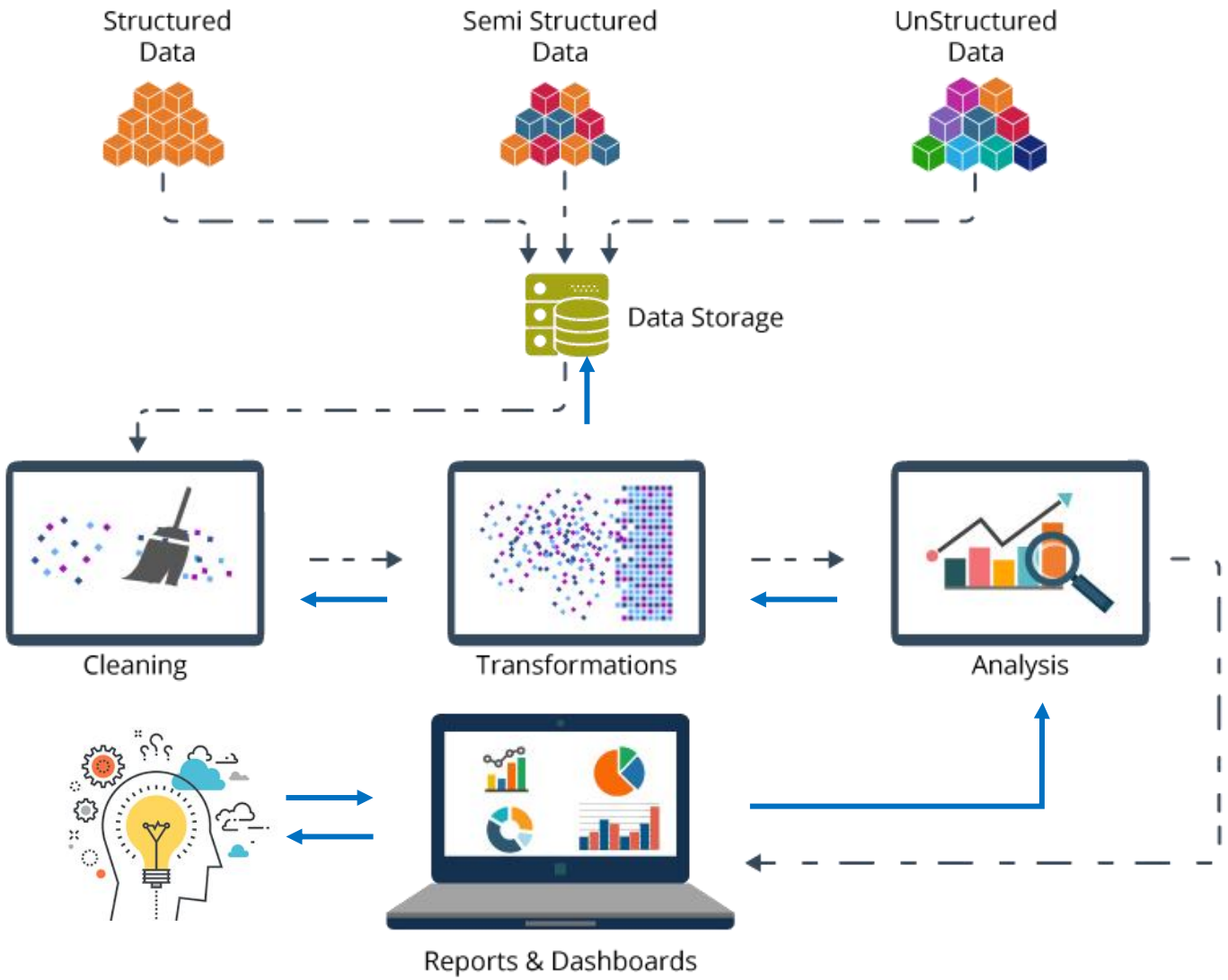
作答

Big Data Analytics

IMDb: Movie Data

Techniques to implement the analytics

Big Data Analytics



- Huge Data Volumes Storage
- High-Speed Networks
- High Performance Computing
- Massive Parallelism
- Data Management

- Data Wrangling
- Data Mining and Analytics
- Machine Learning

- Data Visualization and interaction

Techniques Towards Big Data

- Huge Data Volumes Storage
- High-Speed Networks
- High Performance Computing
- Massive Parallelism
- Data Management

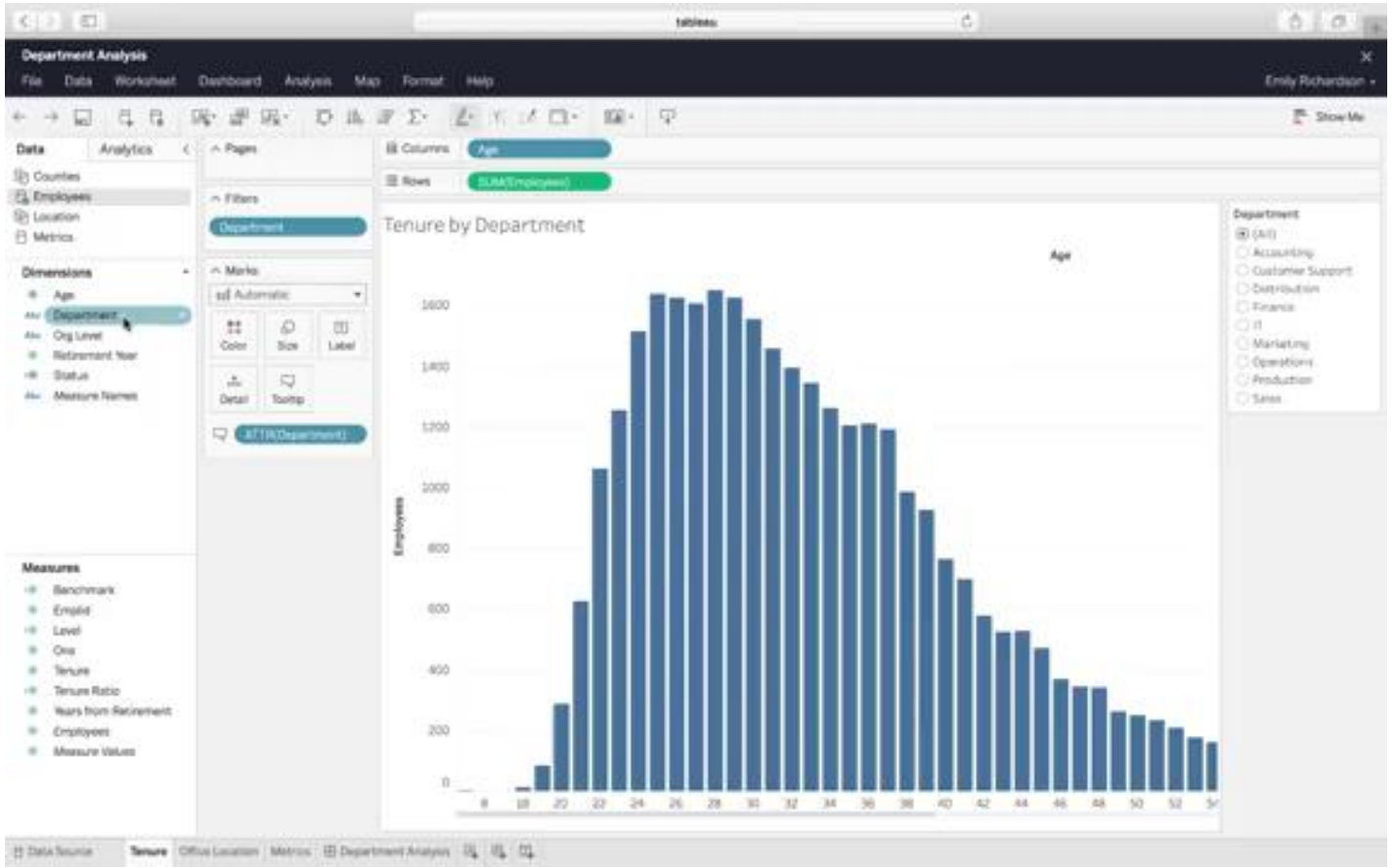
Interactive Data Exploration

- Data Wrangling
- Data Mining and Analytics
- Machine Learning

- Data Visualization and interaction

Data Analytics

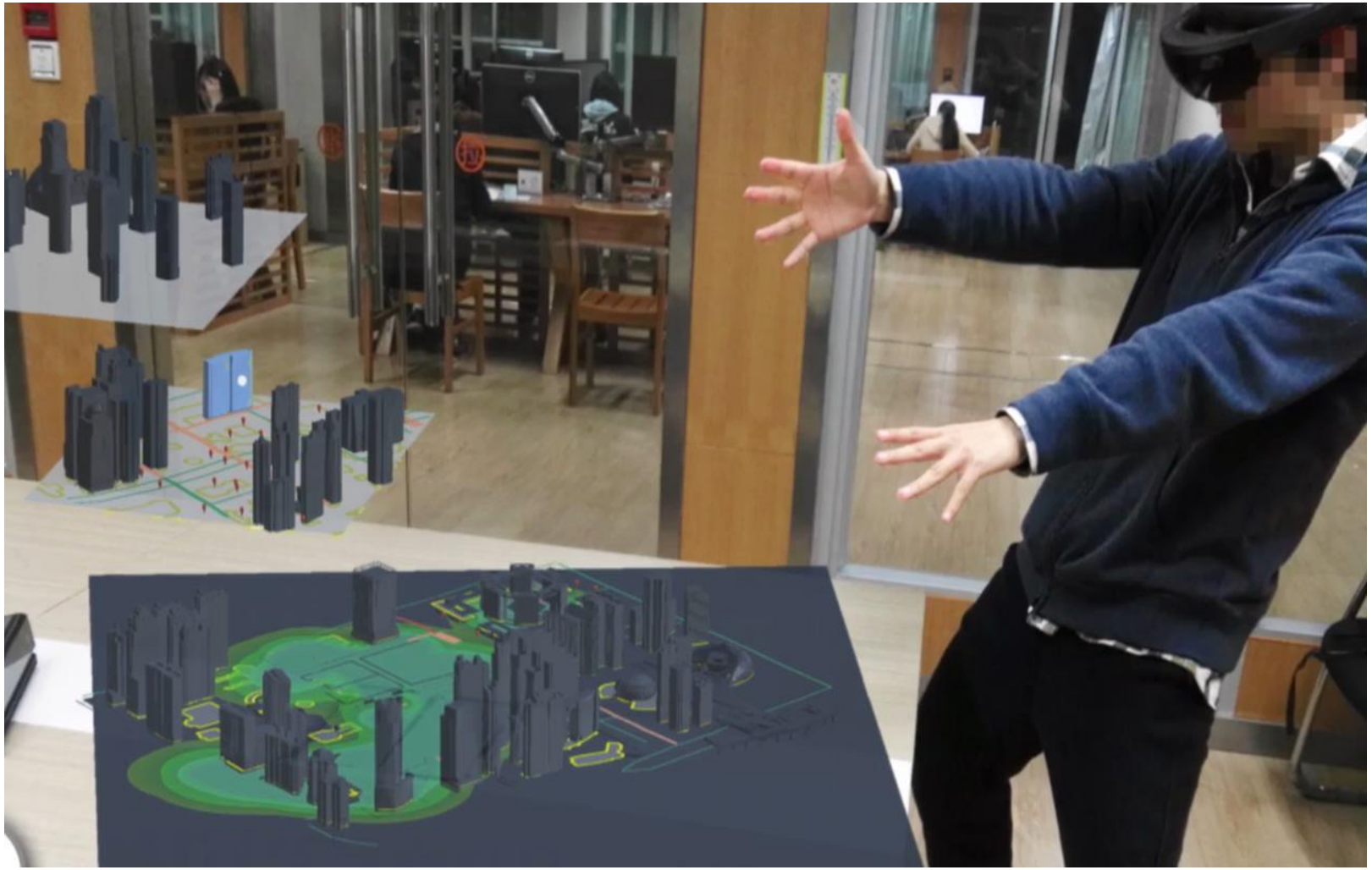
Visual Analytics



Harness the power of your data**. Unleash the potential of your people.**
Choose the analytics platform that disrupted the world of business intelligence.

Data Analytics

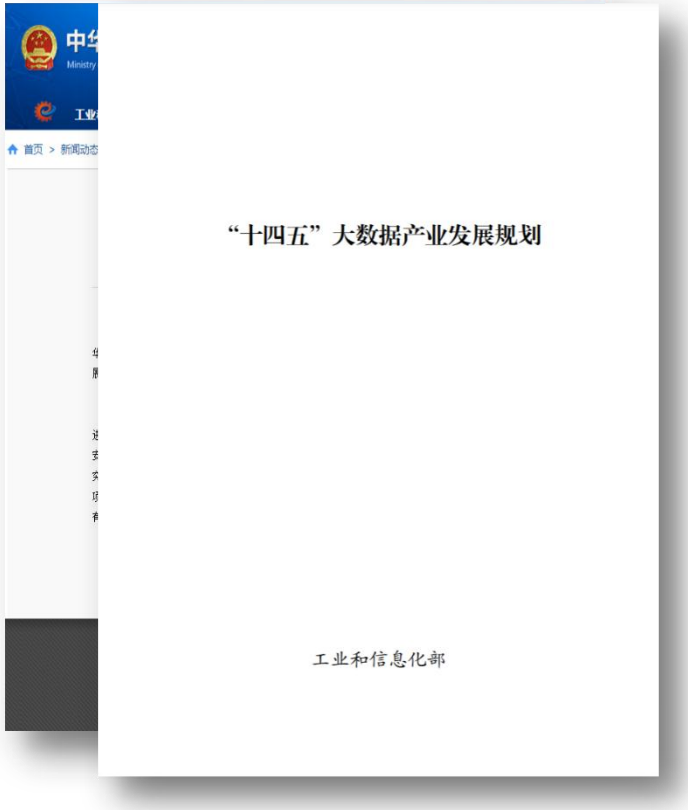
Immersive Environment



Project-focused Course

课程背景

- 人机智能融合的大数据分析是**国家安全情报决策、武器研究、智慧城市决策**等**国家战略**领域及**社会民生**需求的基础技术手段
- 国家《“十四五”大数据产业发展规划》强调了数据分析对于构建稳定**高效产业链**的作用
- 然而我国大数据产业存在技术支撑不强，**基础软硬件、开源框架与国际先进水平存在差距**



史上最严技术出口管制

2018年11月，美国最新14类技术出口管制之第六条：数据分析（可视化、自动分析算法）；

(i) Systems-on-Chip (SoC); or
(ii) Stacked Memory on Chip.
(5) Advanced computing technology, such as:
(i) Memory-centric logic.
(6) Data analytics technology, such as:
(i) Visualization;
(ii) Automated analysis algorithms; or
(iii) Context-aware computing.
(7) Quantum information and sensing technology, such as
(i) Quantum computing;
(ii) Quantum encryption; or
(iii) Quantum sensing.
(8) Logistics technology, such as:
(i) Mobile electric power;
(ii) Modeling and simulation;
(iii) Total asset visibility; or
(iv) Distribution-based Logistics

我国与美国之间的差距

大规模数据分析平台几乎被美国国家实验室所垄断

SQL Server
ORACLE DATABASE
VTK
PostgreSQL
hadoop
VISIT
VisTrails
VolView
Scibian
ParaView

课程历史

“人工智能+”引发**产业界大数据人才需求**上涨

四名具有**国家级项目**经验专业教师实施课程教学改革

党中央“国家大数据战略”

2022 大数据分析实践**国际课程**教改项目

2021 推出低代码声明反馈式**教学平台**

2020 数据科学与大数据相关**省级**教改项目

2019 **重点教改**培育本课程，引入**国家级项目**作为课程实践项目

2018 建设**数据科学与大数据技术**新工科创新试验班

2017 教育部“新工科”建设

2016 培养数据科学方向学生

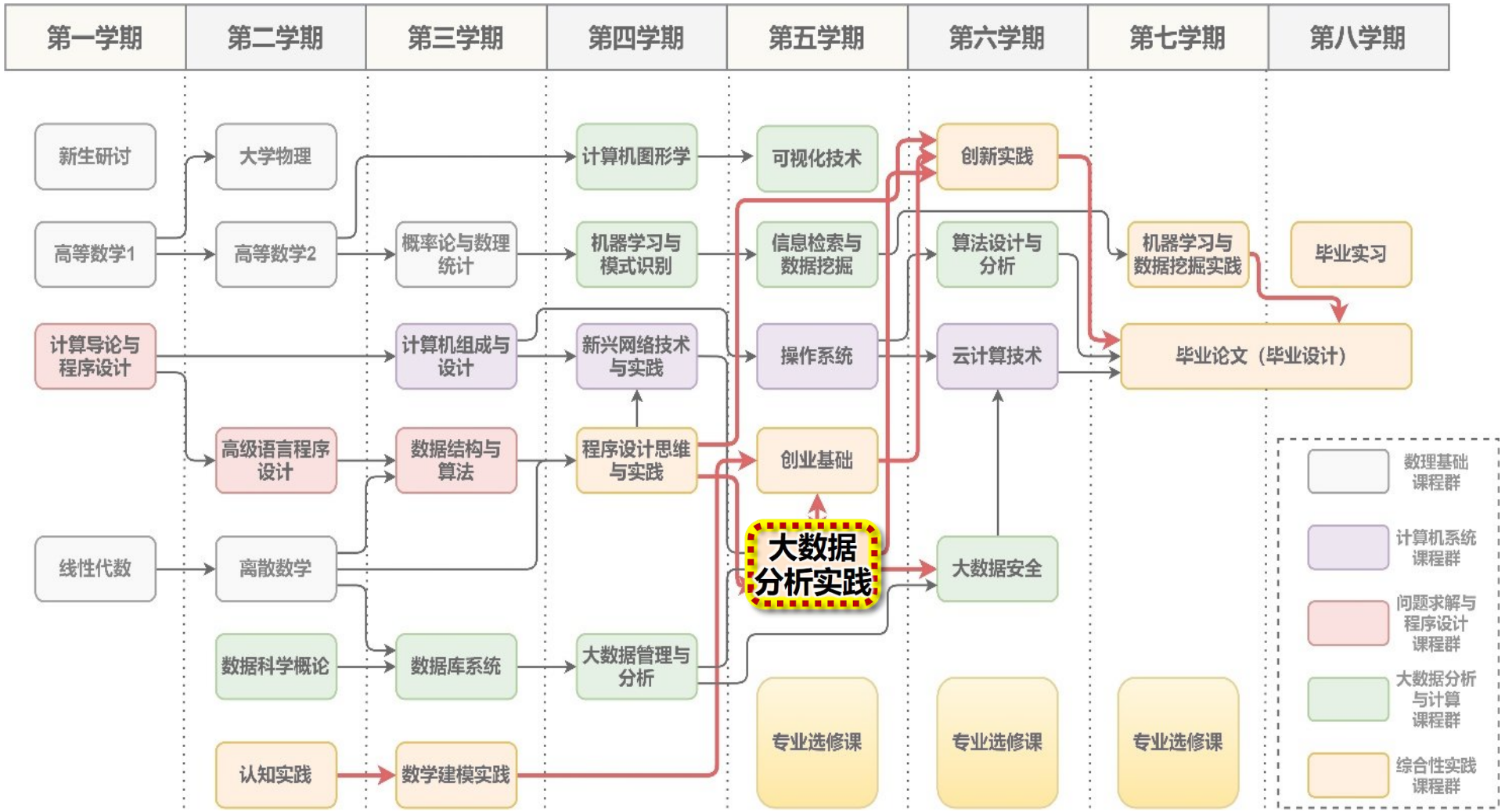
2001 《数据分析技术》《数据仓库数据挖掘》等大数据分析前身课程

课程地位

专业培养目标

培养数据科学与大数据技术**拔尖**人才，学生具有坚实的数理基础和宽广扎实的计算机科学知识，具有**独立的研究能力**，熟练的**沟通能力**，具有逻辑推理、计算分析、算法优化、随机运用等方面的能力，具备理论思维、计算思维、数据思维、并行思维、实验思维等**科学素养**，实现**科学与工程**的紧密联系，具有高度的社会责任感和良好的职业道德，具有过硬的社会竞争力和国际化视野，具有不断**学习的能力和开拓创新**精神，具有良好的**团队合作和组织管理**能力，能够在业界工作中发挥领袖作用。

课程地位



课程类型

专业必修课

学时安排

课堂32学时
实验32学时

课程学分

3学分

面向专业

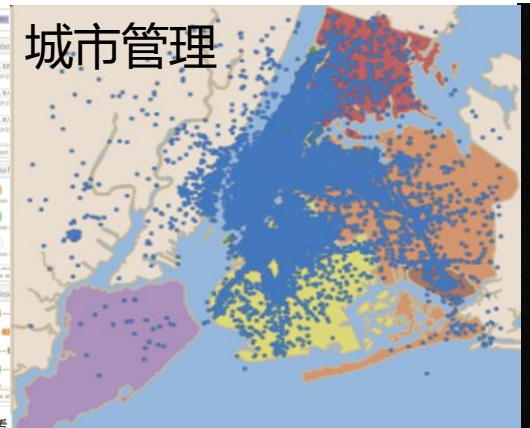
数据科学与大数据技术专业

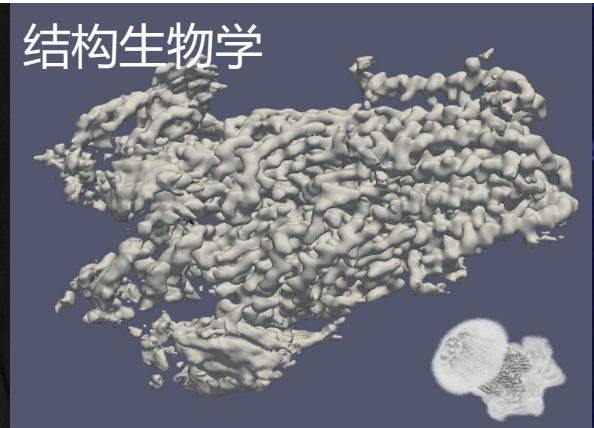
授课对象

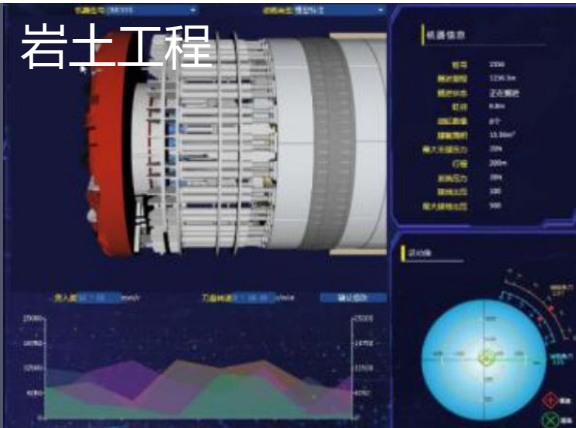
本科三年级

课程职责

课程职责：培养学生构建人机智能融合的**大数据探索分析系统**能力







课程目标

课程职责：培养学生构建人机智能融合的**大数据探索分析系统**能力

素养目标

- 树立为社会民生和国家战略研发国产大数据分析软件系统的自主创新意识
- 具备系统研发团队合作精神以及职业素养

- 关注社会民生与国家战略
- 自主创新
- 团结合作

能力目标

- 自主研发面向社会民生以及国家需求的人机智能融合大数据分析系统
- 具备需求分析、前沿技术调研、系统开发、系统验证以及持续改进的数据系统工程研究及实践能力

- 系统实践能力
- 系统研究能力

知识目标

- 掌握数据管理、数据分析、数据可视化、人机交互等开发大数据探索式分析系统所需的算法设计思想、基础方法、核心技术及实现流程

- 数据处理
- 统计学习
- 大规模数据系统
- 数据可视化
- 人机交互
- 普适计算

目标细化

课程大纲

教学设计

以**真实项目问题**为导向的教与学

教学实施

项目引领的**分工式课堂教学**

考核评价

项目全生命周期综合评价

能力目标
能够开发大数据探索式分析系统

目标达成标志
是否按计划完成大数据探索式分析系统

教与学
大数据分析系统项目团队分工合作教与学

考评
大数据分析系统、课堂讨论、创新报告、以及学生互评

递进式教学内容重组

模块一	基础	数据处理	统计学习	• 数理基础 • 基础编程技巧
模块二	分析	数据可视化	人机交互	• 数据分析工具 • 多源数据呈现能力
模块三	提高	普适计算	大规模数据系统	• 大规模数据处理进阶 • 团队协作熟练度

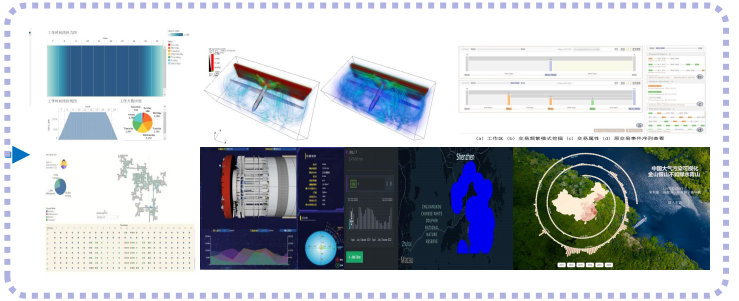
数据分析实践平台

- 面向海量冷冻电镜数据的高分辨率原位结构智能解析软件平台
- 云端融合的多模态数据交互意图理解
- 面向移动认知网络的频谱大数据理论与关键技术研究
- 感知驱动的信息可视化研究
- 盾构地质精细探测与智能掘进关键技术及产业化
- 渐进式大数据可视分析
- 十亿亿次计算原位可视分析并行扩展绘制算法

国家级科研项目

- 冷冻电镜数据三维可视分析系统
- 唐宋诗词可视分析系统；文本语义情感分析系统；人体姿态估计系统
- 基于GPS的城市人口数据分析系统；基于GPS的绑架事故分析系统；基于GPS的城市规划系统
- 资金流与涉网诈骗可视分析系统疫情数据可视分析系统
- 盾构掘进数字孪生可视分析系统；城市数字孪生可视分析系统
- 大气污染时空态势分析系统；期刊图像及文本分析系统
- 电磁场自动可视化系统

学生实践项目



学生实践创新

课程内容

目标：构建大数据探索式分析系统

社会民生和国家战略为载体的真实项目解构

项目来源	国家项目名称	实践项目名称
国家重点研发计划	面向海量冷冻电镜数据的高分辨率原位结构智能解析软件平台	冷冻电镜数据三维可视分析系统
国家重点研发计划	云端融合的多模态数据交互意图理解	唐宋诗词可视分析系统
		文本语义情感分析系统
		人体姿态估计系统
国家重点研发计划	面向移动认知网络的频谱大数据理论与关键技术研究	基于GPS的城市人口数据分析系统
		基于GPS的绑架事故分析系统
		基于GPS的城市规划系统
国家自然科学基金	感知驱动的信息可视化研究	资金流与涉网诈骗可视分析系统
		疫情数据可视分析系统
山东省重点研发计划项目	盾构地质精细探测与智能掘进关键技术及产业化	盾构掘进数字孪生可视分析系统
		城市数字孪生可视分析系统
科技部高端外国专家引进计划	渐进式大数据可视分析	大气污染时空态势分析系统
		期刊图像及文本分析系统
国防科工局国防基础科研核科学挑战专题	千亿亿次计算原位可视分析并行扩展绘制算法	电磁场自动可视化系统

大气污染实时态势分析系统

基于真实项目执行过程的理论教学知识重构

理解项目需求

分析任务：
- 大气污染源分析
- **大气污染时空态势分析**
- 大气污染传输模式分析
- 大气污染预测

收集及处理数据

中国大气污染源数据集
- 时间：2013-2018
- 空间分辨率：15 kilometers
- 空间坐标：经度、纬度
- 污染源变量：PM2.5、PM10、O_3、CO、SO_2、NO_2
- 大气变量：风向U、风向V、温度、相对湿度、压力
- 40000行数据、13个维度

建立数据模型

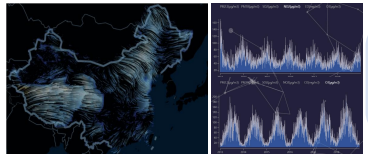
LSTM + K-Means

- Initial preparation
  - As the daily data fluctuates frequently, 28 days is selected as the window size
- Training set: test set, 8:2
- Data processing: separate the month and day and convert them to int format

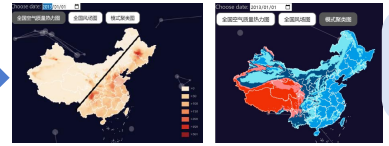
```
data_cs.head()
```

	AQI	PM2_5	PM10	SO2	NO2	CO	O3	Time	Year	Month	Day
0	75.250000	55.200001	65.410004	27.07	25.530001	1.11	27.469999	2013-01-01	2013	1	1
1	117.224998	88.779999	101.500000	30.91	32.830002	1.61	23.780001	2013-01-02	2013	1	2
2	68.900002	48.520000	55.770000	22.74	24.000000	1.63	26.129999	2013-01-03	2013	1	3
3	78.375000	57.700001	55.860001	20.66	21.320000	1.56	26.020000	2013-01-04	2013	1	4
4	125.600006	95.480003	81.220001	27.82	28.280001	1.42	19.900000	2013-01-05	2013	1	5

呈现数据



探索交互分析



基础

大数据分析系统 | 实践项目导论 | 科研实践入门 | 系统构建经验谈

分析

统计分析 | 机器学习与深度学习工具 | 数据采样与降维 | 数据质量管理 | 众包 | 电子表格

普适计算 — 提高 — 大规模计算

理论配套小实验

- SPARK实践
- 手机移动数据采集与分析
- 人机交互实践
- BERT实践
- BERT环境配置
- 可视化设计
- 统计方法实践
- 电子表格实践
- 基于Pandas的数据清洗
- 基于Pandas的数据采样
- 项目规划
- 团队分组

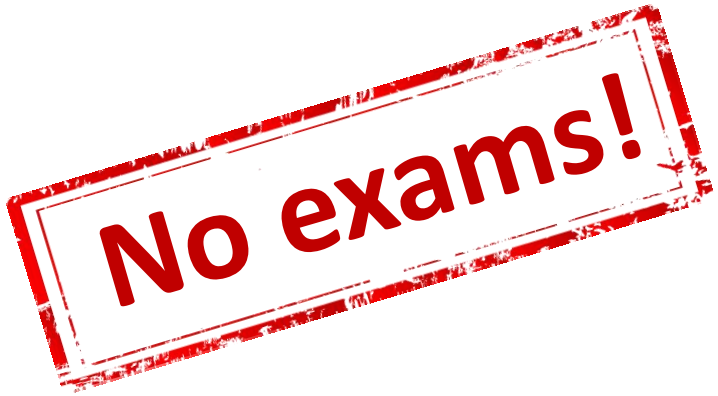
课程大纲

6 Personal assignments

6 Group assignments

上课日期	授课内容	实验内容	周次
20250903	课程入门、大数据探索式分析	课程实践项目介绍、项目组队测试	第一周
20250910	项目经验谈、科研实践入门	项目管理工具制定项目计划	第二周
20250917	数据采样与降维	**数据采样实践**	第三周
20250924	**数据质量管理**	数据质量实践	**第四周**
20251001	众包与电子表格	电子表格实践	第五周
20251008	统计分析方法与工具	统计方法实践	第六周
20251015	可视化设计	可视化设计实践	第七周
20251022	中期汇报（论文+项目进展）1	中期进展报告	第八周
20251029	中期汇报（论文+项目进展）2	BERT实践环境配置	第九周
20251105	机器学习方法与工具	BERT实践	第十周
20251112	人机交互方法与工具	Canis/Cast/Libra实践	第十一周
20251119	普适计算	手机移动数据采集与分析	第十二周
20251126	大规模数据分析系统	SPARK实践	第十三周
20251202	如何撰写项目论文	大项目收尾	第十四周
20251209	项目结题报告1		第十五周
20251216	项目结题报告2	大项目验收	第十六周

Basic Factors



Faculty Qiong Zeng (曾琼)

Teaching Assistant Xi Duan (段曦)

Zhiyuan Meng (孟致远)

Time Wednesday 10:10-12:00 (振声苑N202)

Friday 16:10-18:00 (N3)

Office Hour Friday 10:00-12:00 (N3-412)

Course Grading

Final Project (55%)

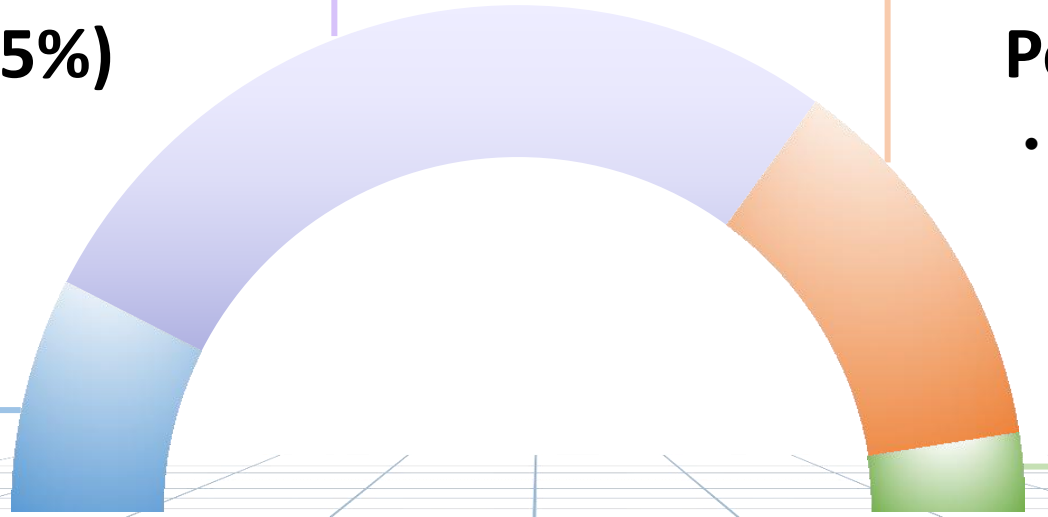
- Paper **Writing** (<= 4 pages + 1 page reference, 20%)
- Final Project **Presentation** (Oral + Demo, 20%)
- Source Code Quality (10%)
- Middle Term Checkpoint (5%)

Task (15%)

- Middle-Term **Presentation** (15%): *choose a most relevant reference to your project*

Class Participation (25%)

- In-class Discussion (5%)
- Assignments **(20%): online and offline [6 personal assignments + 6 group assignments]**



Peer Assessment (5%)

- **Notes**: each group leader will get a **5%** bonus.

Course Grading

一级指标		二级指标		优秀	良好	中等	及格	不及格
指标	比例	指标	比例	≥9分	8-9分	7-8分	6-7分	<6分
大数据分析系统演示	20%	系统完整性	40%	系统功能完整，涵盖数据清洗、数据分析挖掘、数据可视化、数据交互等模块，能够满足任务分析的所有需求。	系统功能基本完整，基本涵盖数据清洗、数据分析挖掘、数据可视化、数据交互等模块，，能够满足任务分析的大部分需求。	系统功能不够完整，涵盖数据清洗、数据分析挖掘、数据可视化、数据交互等模块有缺失，但能够满足任务分析的大部分需求。	系统功能不完整，涵盖数据清洗、数据分析挖掘、数据可视化、数据交互等模块有较大缺失，无法满足任务分析的多数需求，存在一些系统问题。	系统功能严重不完整，未涵盖数据清洗、数据分析挖掘、数据可视化、数据交互等模块，无法满足任务分析的基本需求。
		系统创新性	20%	系统采用了新的技术或方法，具有较强的创新性，并取得了显著的效果。	观点基本明确，论述基本清晰，基本易于理解；术语使用基本规范，解释基本到位；概念定义基本准确，无重大歧义；句子结构基本合理，语句通顺。	观点不够明确，论述部分清晰；术语使用不够规范，解释存在较大不到位；概念定义不够准确，有较多歧义；句子结构不够合理，阅读困难。	观点模糊，论述基本清晰，较难理解；术语使用基本规范；概念定义不准确，存在重大歧义；句子结构表述混乱。	观点不明确，论述无逻辑，术语使用错误，缺少解释，语句不通顺

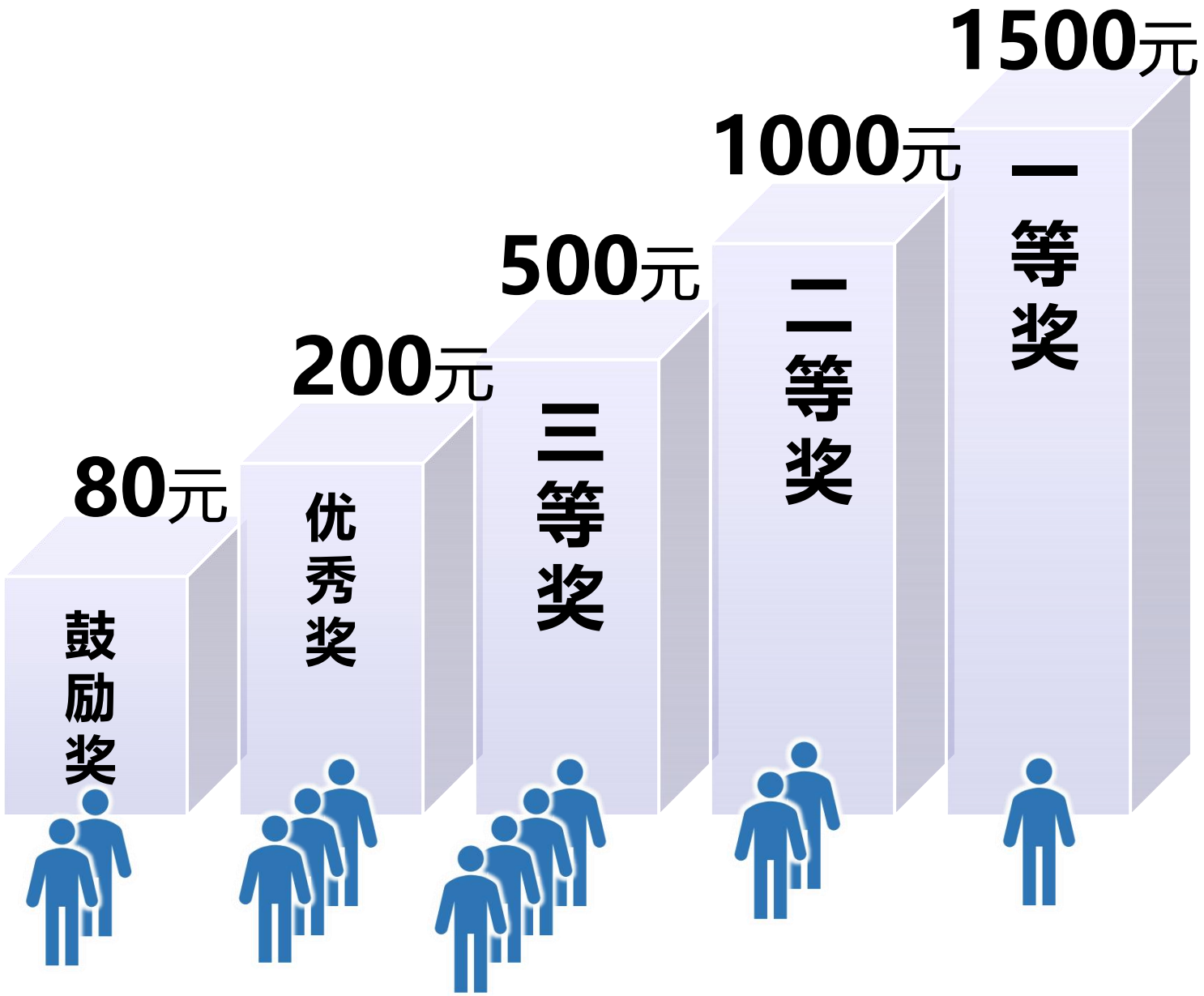
Course Grading

一级指标		二级指标		优秀	良好	中等	及格	不及格
指标	比例	指标	比例	≥9分	8-9分	7-8分	6-7分	<6分
项目论文	20%	写作规范	15%	论文结构完整，包含引言、相关工作、技术细节、实验分析、总结、参考文献等核心模块，各模块内容充实完整。语言表达准确、流畅，无语法错误；图表清晰美观，图标题描述自洽。	论文结构论文结构完整，包含引言、相关工作、技术细节、实验分析、总结、参考文献等模块，各模块均有涉及。语言表达基本准确，少量语法错误；图表清晰美观。	论文结构论文结构基本完整，参考文献、结论等模块存在缺失。语言表达基本准确，语法错误较多。图表描述不符合规范。	论文结构论文结构缺失参考文献部分。语言表达基本准确，语法错误较多。图表描述不清晰，排版存在一定混乱。	论文结构存在严重缺失，比如无相关文献调研。语言表达存在较大逻辑错误以及逻辑不通顺。无图表或图表混乱。
		语言表达	20%	观点明确，论述清晰，易于理解；术语使用规范，解释清楚到位；概念定义准确，无歧义；句子结构合理，语句通顺。	观点基本明确，论述基本清晰，基本易于理解；术语使用基本规范，解释基本到位；概念定义基本准确，无重大歧义；句子结构基本合理，语句通顺。	观点不够明确，论述部分清晰；术语使用不够规范，解释存在较大不到位；概念定义不够准确，有较多歧义；句子结构不够合理，阅读困难。	观点模糊，论述基本清晰，较难理解；术语使用基本规范；概念定义不准确，存在重大歧义；句子结构表述混乱。	观点不明确，论述无逻辑，术语使用错误，缺少解释，语句不通顺

Course Grading

一级指标		二级指标		优秀	良好	中等	及格	不及格
指标	比例	指标	比例	≥9分	8-9分	7-8分	6-7分	<6分
项目论文	20%	写作规范	15%	论文结构完整，包含引言、相关工作、技术细节、实验分析、总结、参考文献等核心模块，各模块内容充实完整。语言表达准确、流畅，无语法错误；图表清晰美观，图标题描述自洽。	论文结构论文结构完整，包含引言、相关工作、技术细节、实验分析、总结、参考文献等模块，各模块均有涉及。语言表达基本准确，少量语法错误；图表清晰美观。	论文结构论文结构基本完整，参考文献、结论等模块存在缺失。语言表达基本准确，语法错误较多。图表描述不符合规范。	论文结构论文结构缺失参考文献部分。语言表达基本准确，语法错误较多。图表描述不清晰，排版存在一定混乱。	论文结构存在严重缺失，比如无相关文献调研。语言表达存在较大逻辑错误以及逻辑不通顺。无图表或图表混乱。
		语言表达	20%	观点明确，论述清晰，易于理解；术语使用规范，解释清楚到位；概念定义准确，无歧义；句子结构合理，语句通顺。	观点基本明确，论述基本清晰，基本易于理解；术语使用基本规范，解释基本到位；概念定义基本准确，无重大歧义；句子结构基本合理，语句通顺。	观点不够明确，论述部分清晰；术语使用不够规范，解释存在较大不到位；概念定义不够准确，有较多歧义；句子结构不够合理，阅读困难。	观点模糊，论述基本清晰，较难理解；术语使用基本规范；概念定义不准确，存在重大歧义；句子结构表述混乱。	观点不明确，论述无逻辑，术语使用错误，缺少解释，语句不通顺

Course BONUS



Expectations from You

- Be comfortable **asking questions, identifying a problem**
- Be comfortable **reading research papers**
- Be comfortable **coding**: data processing, machine learning, visualizations, interaction
- **Active participate** in discussions and our class
- Be comfortable summarizing your **progress**
- Be comfortable **writing up** experimental results in a **research paper format**

Project Topics

	Topics	潜在产出	合作教师
1	IEEE SciVis Contest 海洋大气气候数据可视分析系统【冠军奖：1000刀+IEEE CG&A论文】	**系统**、论文、专利或软著	曾琼、段曦
2	基于华为昇腾算力与计图框架的冷冻电镜数据分析大模型构建与效能优化【华为云】		曾琼、王飞宇
3	融合知识图谱与时序分析的海洋新闻态势感知系统【国家深海基地合作】		曾琼、孟致远
4	基于实时位置数据的传染病密接个体识别系统【国自然项目】		滕德军、曾琼
5	大规模空间数据交叉匹配数据处理系统【国自然项目】		滕德军、曾琼
6	人类肠道微生物数据资源库与多维智能分析平台【微生物国重合作】	**系统**、专利、软著、	曾琼、蒋荷
7	《大数据分析实践》智慧课程平台【国家一流课程建设经费资助】		曾琼、栾峻峰

Project Groups

竞标：队长 → 队员 → 项目

基本规则：每组获得相同数量的初始金币（10），金币将用于队员的选取以及项目题目

随堂测试选队长 → **双方互联** → **项目选取**

- 3题，满分100，30min
- **得分=平均百分制绩点*0.5+随堂测试得分*0.5**
- 每个人的金币数量为0.1*（选课总人数+1-得分排名）

- 队员总身价**不可超过**队伍初始金币

- **同时**选项目
- **出现多**个队伍选择同样项目时，根据团队**总身价先后**顺序排名
- 每个项目**少于等于2个**队伍选择

Project Groups

竞标：队长 → 队员 → 项目

基本规则：每组获得相同数量的初始金币（10），金币将用于队员的选取以及项目题目

随堂测试选队长

- 3题，满分100，30min
- **得分=上学期百分制绩点*0.5+随堂测试得分*0.5**
- 每个人的金币数量为0.1*（选课总人数+1-得分排名）

双方互联

- 队员总身价不可超过队伍预算金币

项目选取

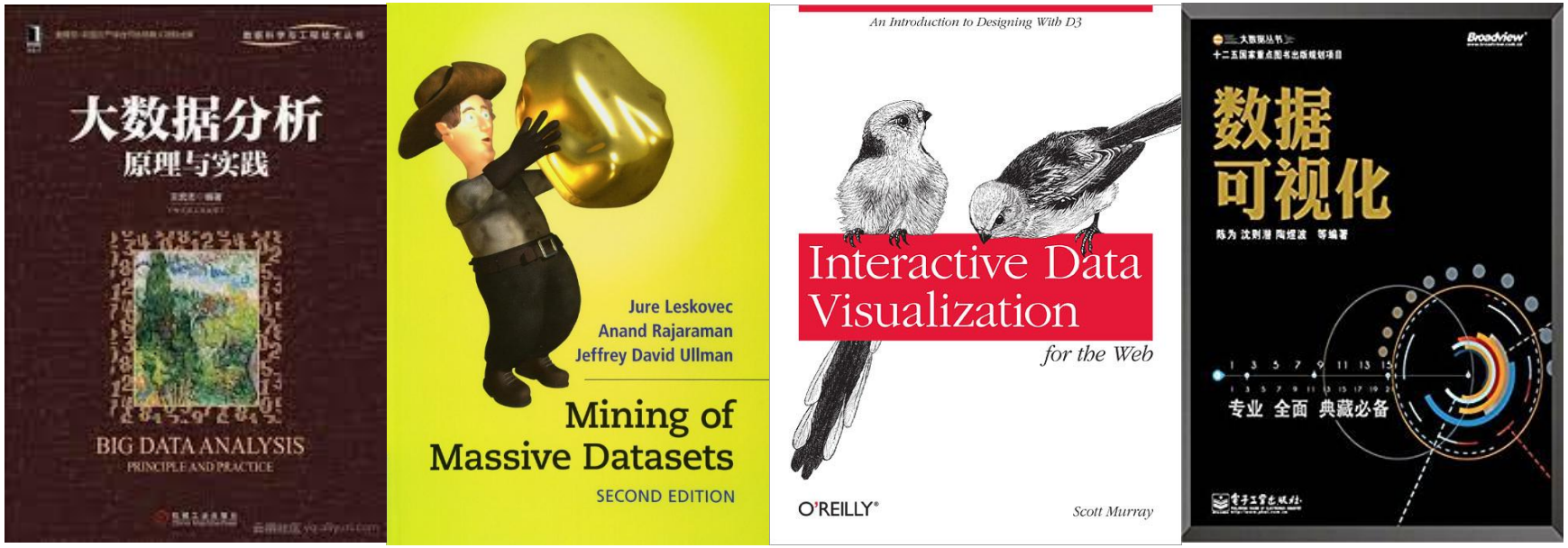
- 同时选项目
- 总身价先后顺序排名
- 每个项目不超过3个队伍选择

周五下午赠提神版随堂测试！！！！

	Week														
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
G1	ZQ	DX	DX	ZY	ZY	ZY	DX	ZY	ZQ	DX	DX	ZY	ZY	ZQ	DX
G2	ZQ	DX	DX	ZY	ZY	ZY	DX	ZY	ZQ	DX	DX	ZY	ZY	ZQ	DX
G3	DX	ZQ	DX	ZY	ZY	ZY	DX	ZQ	DX	ZQ	DX	ZY	DX	ZY	ZY
G4	DX	ZQ	DX	ZY	ZY	ZY	DX	ZQ	DX	ZQ	DX	ZY	DX	ZY	ZY
G5	DX	ZY	ZQ	DX	DX	DX	ZQ	ZY	DX	ZY	ZQ	DX	ZY	ZY	ZY
G6	DX	ZY	ZQ	DX	DX	DX	ZQ	ZY	DX	ZY	ZQ	DX	ZY	ZY	ZY
G7	ZY	ZY	ZY	ZQ	DX	ZQ	ZY	DX	ZY	ZY	ZY	ZQ	DX	DX	DX
G8	ZY	ZY	ZY	ZQ	DX	ZQ	ZY	DX	ZY	ZY	ZY	ZQ	DX	DX	DX
G9	ZY	DX	ZY	DX	ZQ	DX	ZY	DX	ZY	DX	ZY	DX	ZQ	ZY	ZQ
G10	ZY	DX	ZY	DX	ZQ	DX	ZY	DX	ZY	DX	ZY	DX	ZQ	ZY	ZQ
G11	ZY	ZY	ZY	ZQ	DX	ZQ	ZY	DX	ZY	CC	ZY	ZQ	DX	DX	DX
G12	ZY	DX	ZY	DX	ZQ	DX	ZY	DX	ZY	DX	ZY	DX	ZQ	ZY	ZQ

Course Information

课程网站： https://sdubigdatacourse.github.io/#/



- **Conference**: SIGMOD, KDD, VIS, CHI, SIGIR, MM, ……
- **Researchers** (acknowledgements):
  Daniel Keim: *Mastering the information age*
  Stratos Idreos: *Overview of Data Exploration Techniques*

学生成果

2019-2020-1	2020-2021-1	2021-2022-1	2022-2023-1	2023-2024-1
AR空间多类散点图采样与显示 语义情感分析系统 多人姿态估计与跟踪系统	面向珍惜鸟类保护的环境监测系统 工厂安全监测与分析 基于GPS绑架事故分析 公园内游客行为模式及异常行为分析	工业环境监测可视分析 城市监视器数据的公民救援分析 保护区鸟类模式分析 水道污染环境监测系统 大气污染可视分析系统	城市数据聚类分析 城市数字孪生可视分析 GIS数据城市绑架事故分析 城市数据安全监测系统 时空数据可视分析系统	**大规模冷冻电镜数据可视分析系统** **海洋非法捕捞识别系统** **大气污染时空经济效益可视分析系统** **全球温度趋势分析系统** **计算机科学家社交网络分析** **中药药理可视分析系统** **全场景生鲜超市库存决策系统** **运营商大数据栅格时序图预测系统**
多类蓝噪声采样、AR/VR技术、LSTM、RAM、DenseNet	Tableau、DBSCAN聚类算法、数据库、用户交互式可视化界面、CNN、五折交叉验证	D3、Tableau、无监督聚类算法、参数化聚类异常值检测、ARIMA模型、Dynamic Time Warping、LSTM	Tableau、K均值、DBSCAN、TOPSIS、word2Vec、GPT-2、LSTM、WMD、ArcGIS、PageRank、高斯和计算密度估计、KM算法、协同过滤算法、Vue3、Echarts、D3	**线性回归、XGBoost、LSTM、D3、LightGBM、贪心算法、SVM、GCN、BiLSTM、逆地理编码、t-SNE、孤立森林、Streamlit、Transformer、随机森林方法、ECharts、GAN、Redis、Nginx、Docker、Vue.js、Three.js、LLAMA大模型、GPT-4、ChatGPT、熵权法、KNN、SIREN、ONNX.js**

学生成果

大气污染时空经济效益可视分析系统

项目来源：生物气溶胶多组分高灵敏度在线监测技术（科技部）

后端：

编程语言：Python
Web 框架：FastAPI
数据库 ORM：SQLAlchemy
数据库： MySQL
GPT-4

测试：

接口测试工具：Hoppscotch

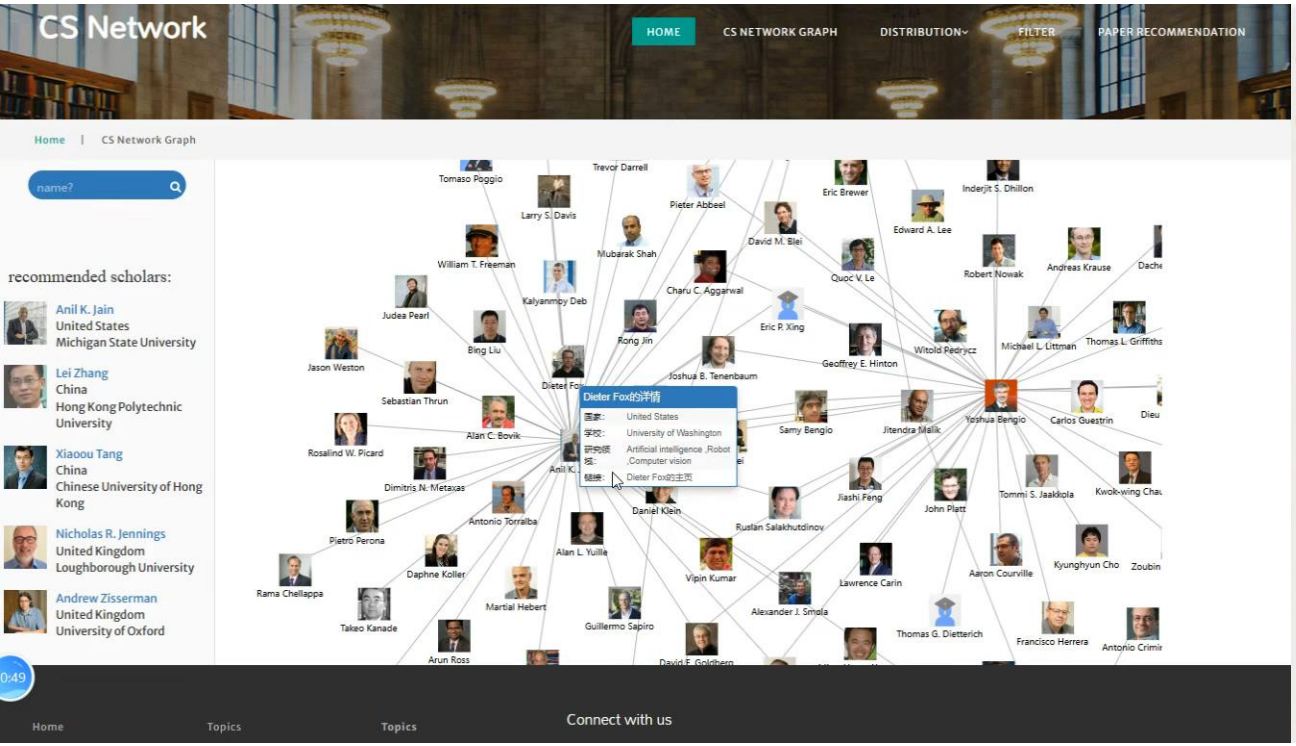
前端：

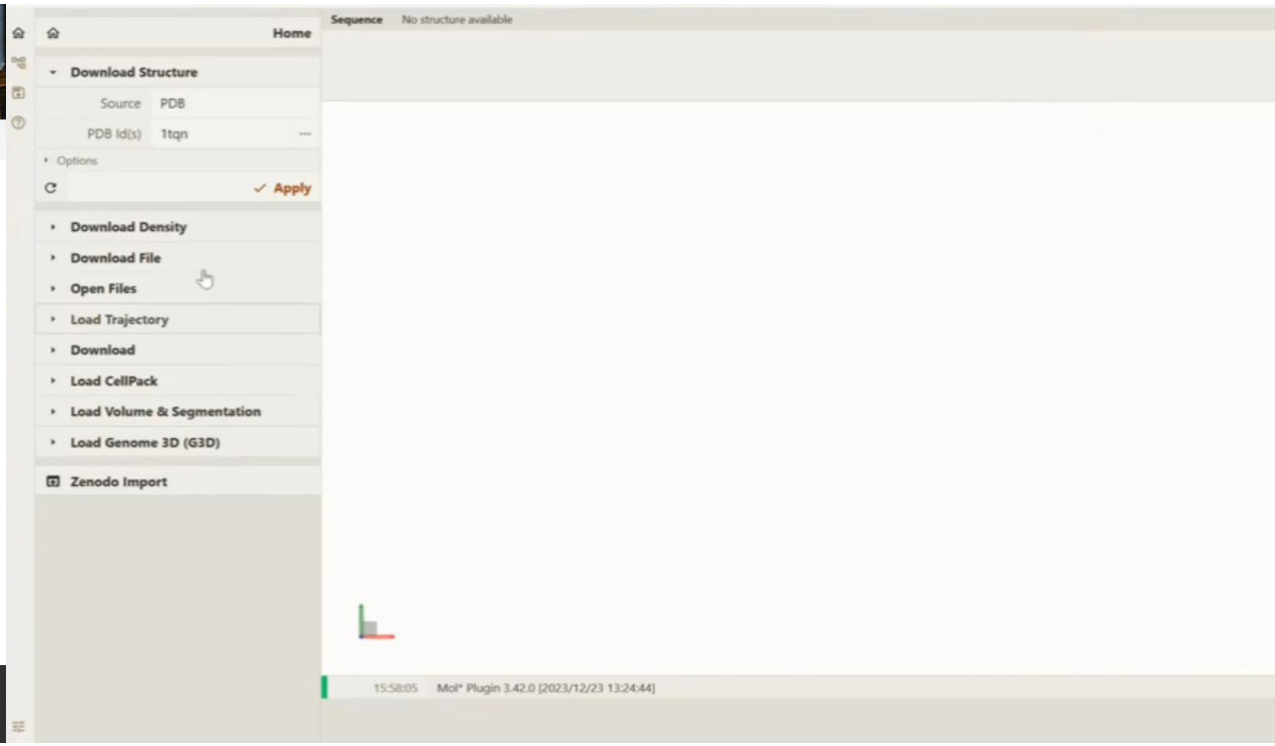
JavaScript 框架：Vue.js
可视化库：ECharts
地图 API：百度地图 API

部署：

主从式架构
分布式文件系统：Minio
缓存与消息队列：Redis
负载均衡与反向代理：Nginx
容器虚拟化技术：Docker

学生成果





计算机科学社交网络分析

来源：渐进式大数据可视分析（科技部）

大规模冷冻电镜数据分析

来源：面向海量冷冻电镜数据的
高分辨率原位结构智能解析平台（科技部）

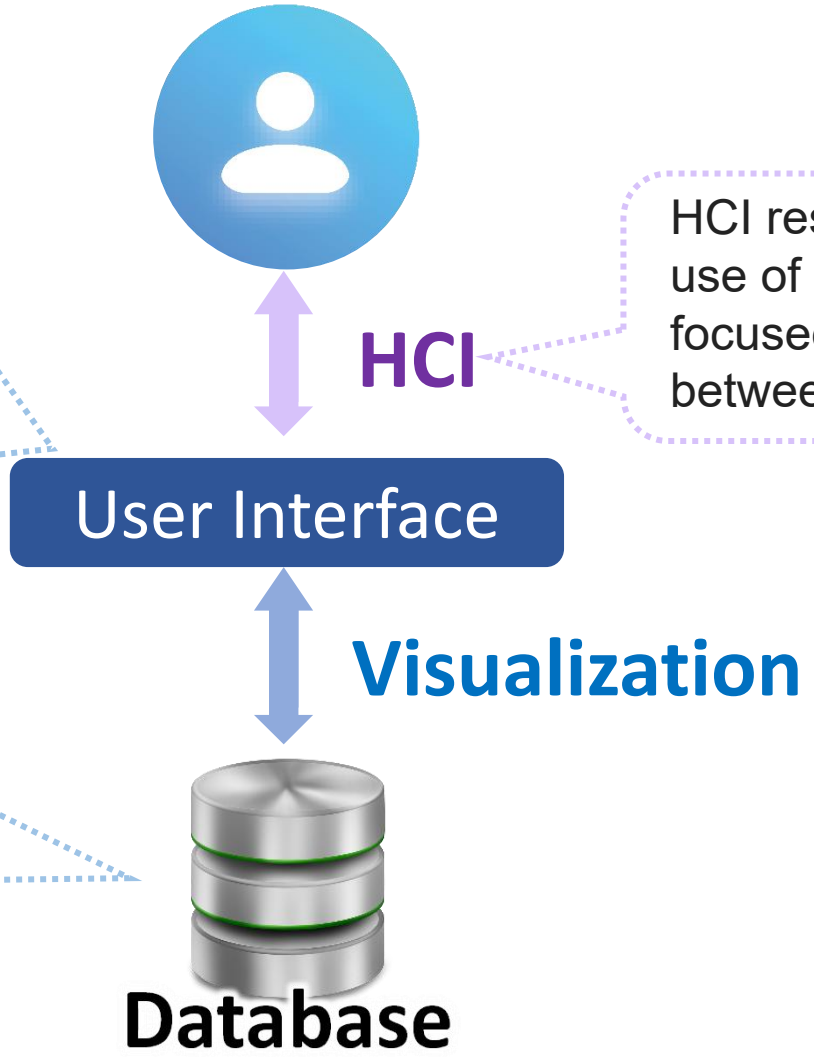
Practices on Big Data Analytics

Interactive Data Exploration

Data Exploration

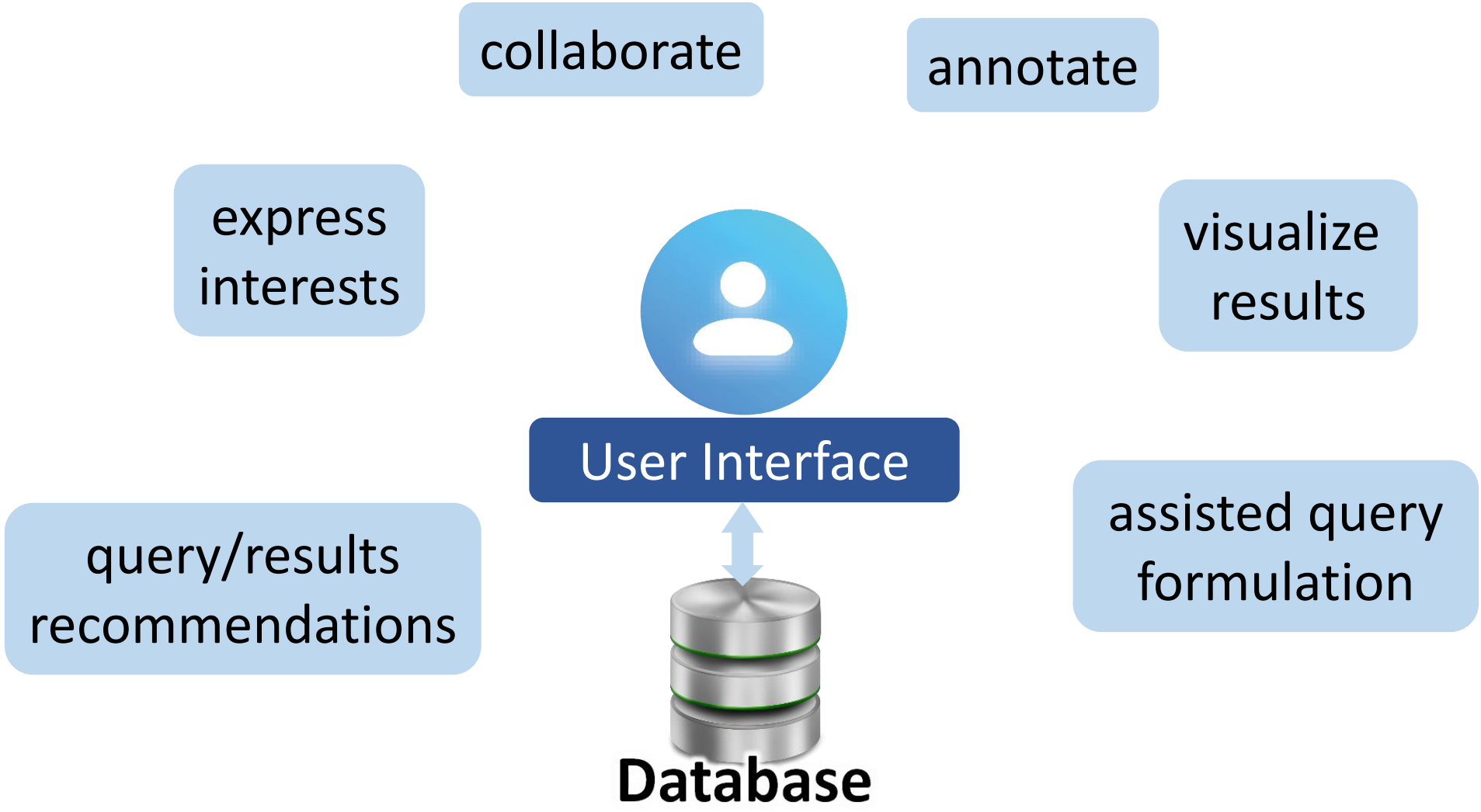
Data visualization refers to the techniques used to communicate data or information by encoding it as **visual objects** (e.g., points, lines or bars) contained in graphics. The goal is to communicate information **clearly and efficiently** to users.

A database is an **organized collection** of data generally stored and accessed electronically from a computer system.

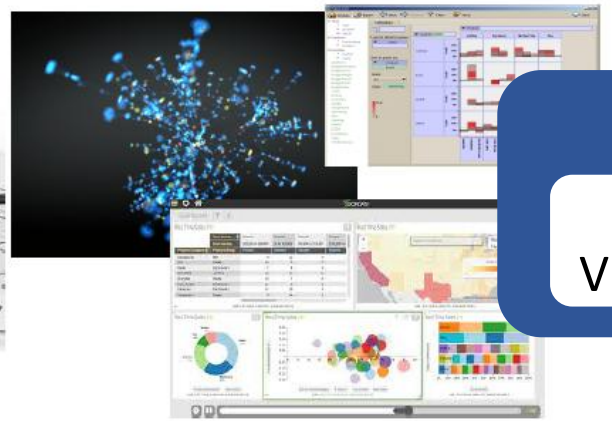


HCI researches the design and use of computer technology, focused on the **interfaces** between people and computers.

Interactive Data Exploration



Interactive Data Exploration



User Interface

Data Visualization

Exploration Interface

Middleware Techniques

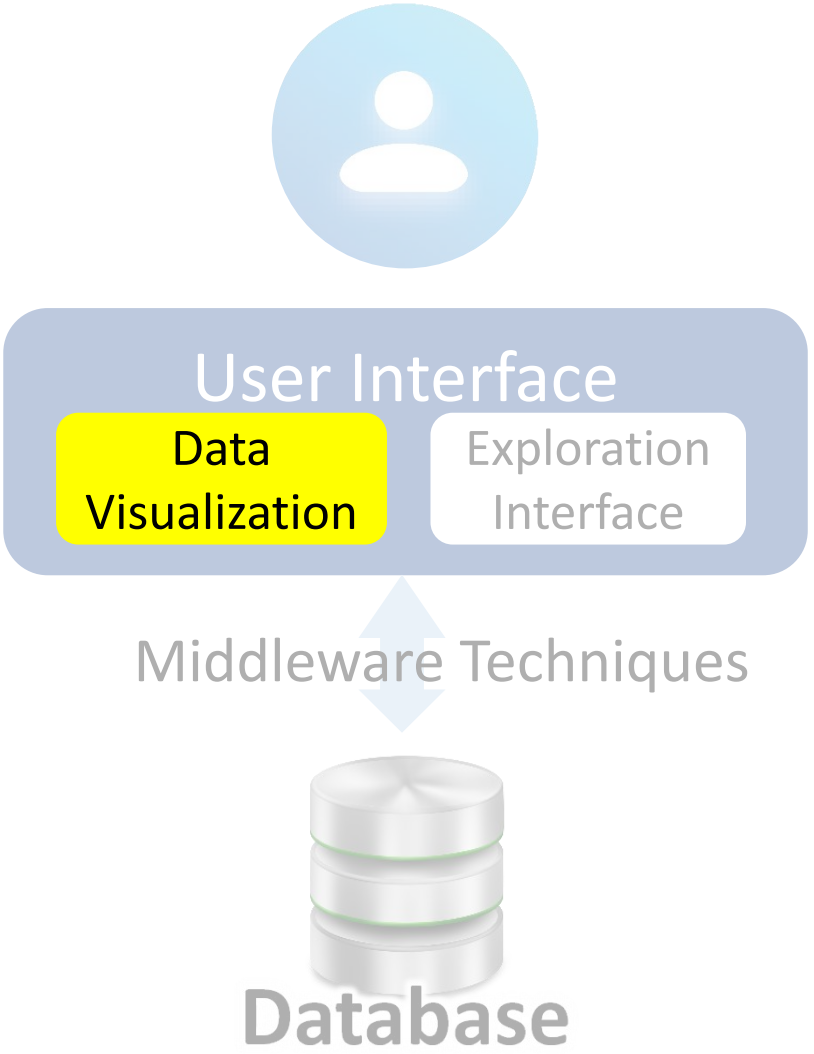
Database

IDE: Data Visualization

Visualization Tools

Visual Optimizations

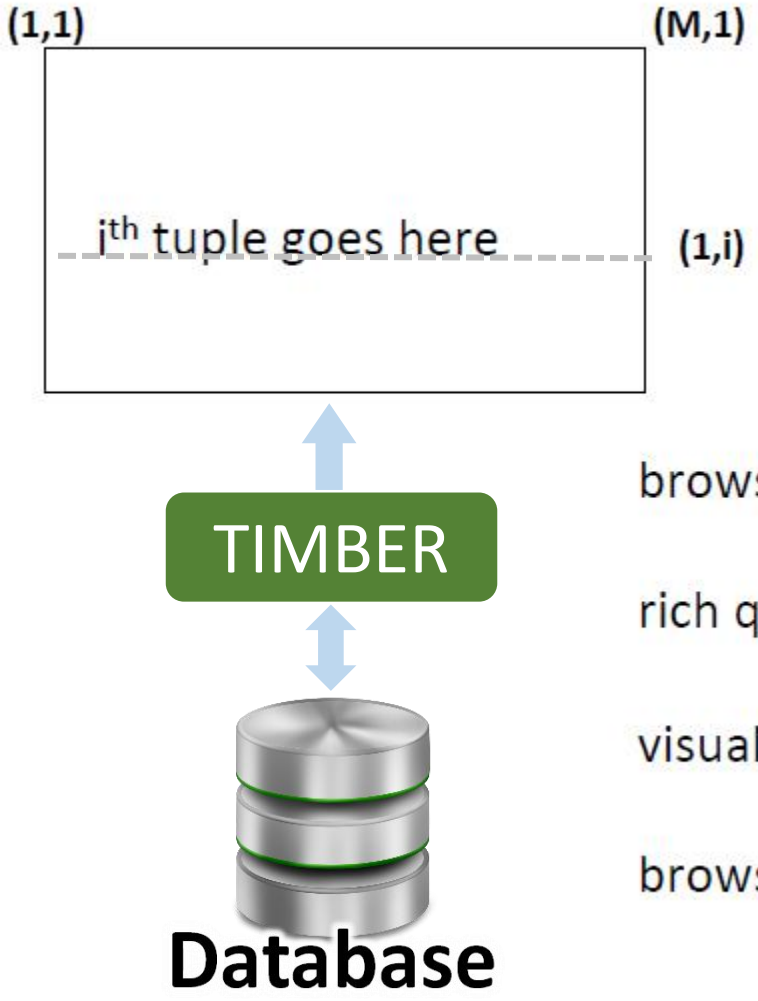
Automatic Visualization



VISUALIZATION TOOLS

TIMBER **POLARIS** **AstroShelf**

TIMBER



Back in 1982...

window-based "sophisticated" browser for relational DBs

browser for multiple relations/tuples

rich query language for icon-oriented DBs

visual editor of text objects

browser for geographical data

TIMBER: a sophisticated relation browser. Michael Stonebraker, Joseph Kalash. VLDB 1982.

TIMBER

For example, suppose an EMP relation is declared as follows:

CREATE EMP (name = c10, salary = i2, visual = icon)

Suppose the graphical token for an employee has been specified as a rectangle. It might be represented as some coding for the following information:

lines:

(0,0) -> (24,0)
(24,0) -> (24,4)
(24,4) -> (0,4)
(0,4) -> (0,0)

text:

"name =" at (1,3)
"salary =" at (2,3)

fields:

name at (1,10)
salary at (2,10)

scale-x = 1/24
scale-y = 1/24

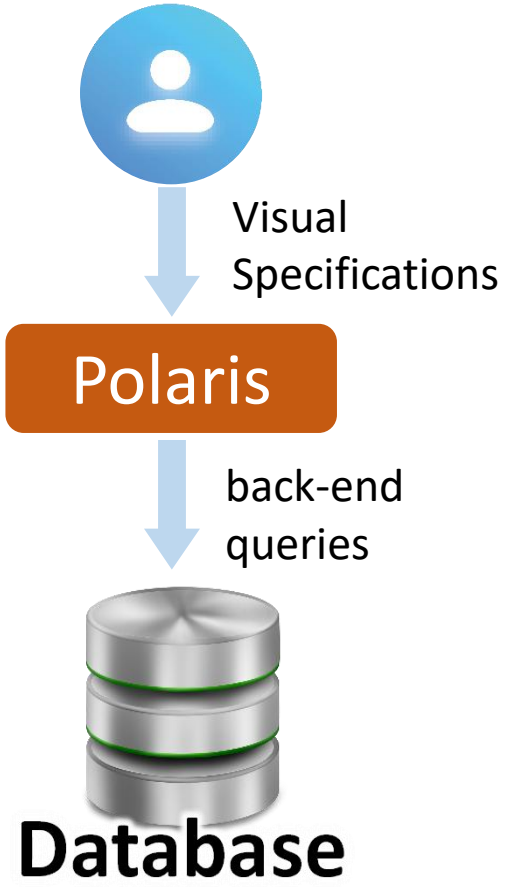
and have a screen image of:

name = Kalash
salary = 10000

TIMBER: a sophisticated relation browser. Michael Stonebraker, Joseph Kalash. VLDB 1982.

Polaris

Explore large multi-dimensional databases (e.g., business, science)



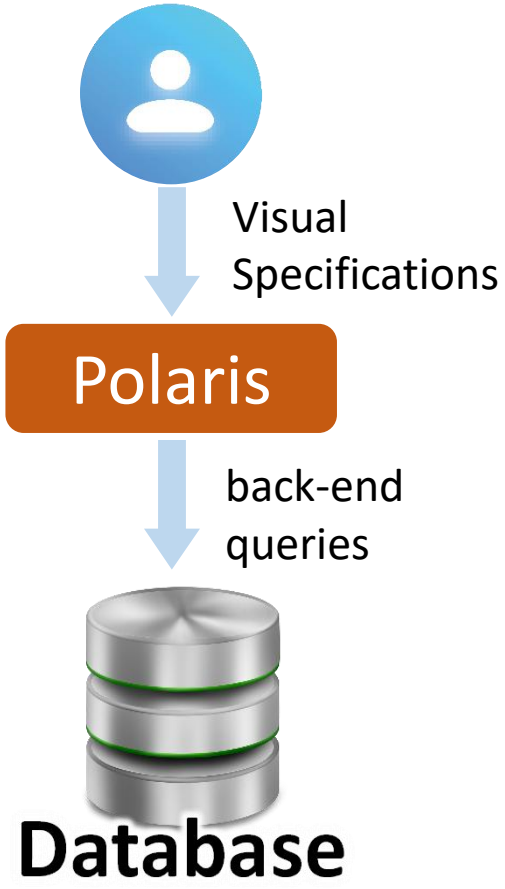
- Unify tables and graphs
- Expressiveness
- Interface simplicity
- Code simplicity

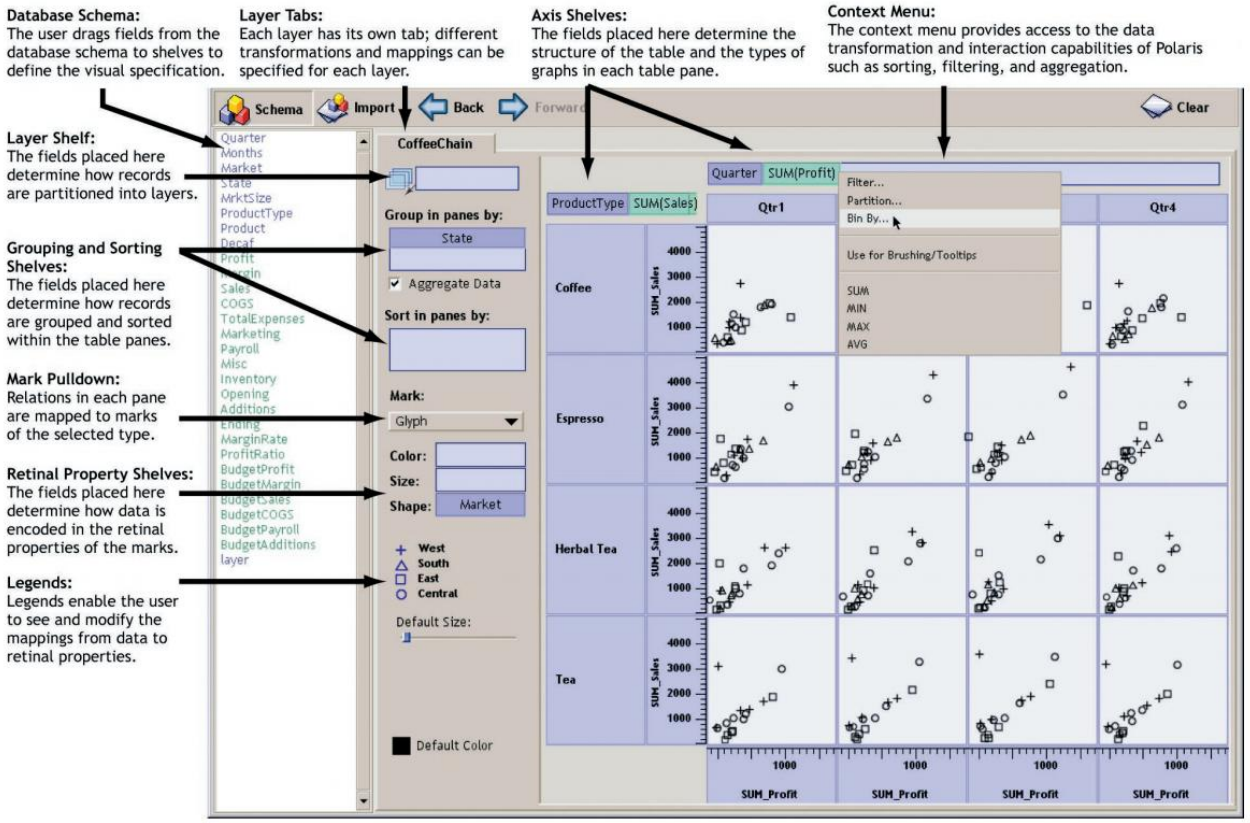
	A	B	C	D	E
1	Region	East			
2					
3	Sum of Sales		Product		
4	Month	Salesperson	Dairy	Meat	Grand Total
5	May	Buchanan		$17,578	$17,578
6		Davolio	$22,977		$22,977
7	May Total		$22,977	$17,578	$40,555
8	Jun	Buchanan	$10,017	$7,711	$17,728
9		Davolio	$6,805	$5,575	$12,380
10	Jun Total		$16,822	$13,286	$30,108
11	Grand Total		$39,799	$30,864	$70,663

Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris

Explore large multi-dimensional databases





Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: Operands

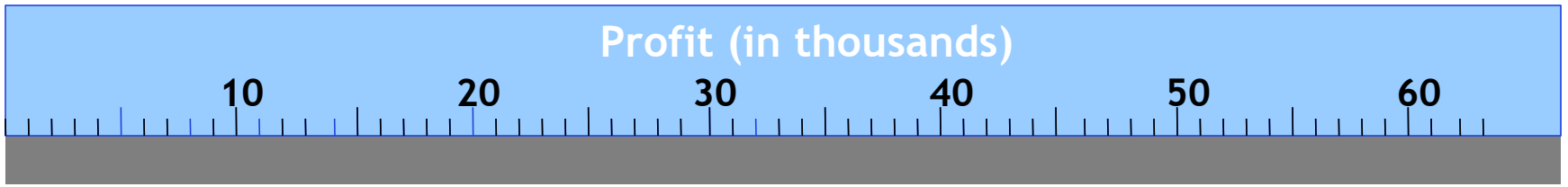
Ordinal fields - interpret domain as a set that partitions table into rows and columns:

QUARTER = {Quarter1, Quarter2, Quarter3, Quarter4} →

Quarter 1	Quarter 2	Quarter 3	Quarter 4
31,400	35,600	37,120	30,900

Quantitative fields – treat domain as single element set and encode spatially as axes:

PROFIT = {P[0 - 65,000]} →



Polaris: Concatenation Operands

Ordered union of set interpretations:

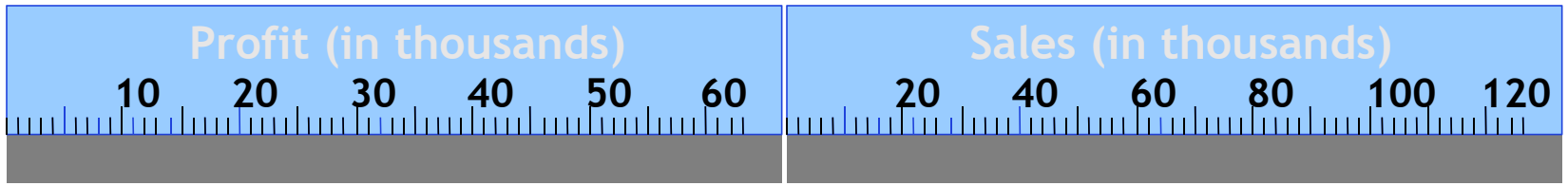
QUARTER + PRODUCT_TYPE

= {QTR1,QTR2,QTR3,QTR4} + {Coffee, Tea}

= {QTR1,QTR2,QTR3,QTR4, Coffee, Tea}

Quarter 1	Quarter 2	Quarter 3	Quarter 4	Coffee	Tea
31,400	35,600	37,120	30,900	37,120	30,900

PROFIT + SALES = {P[0-65,000], S[0-125,000]}

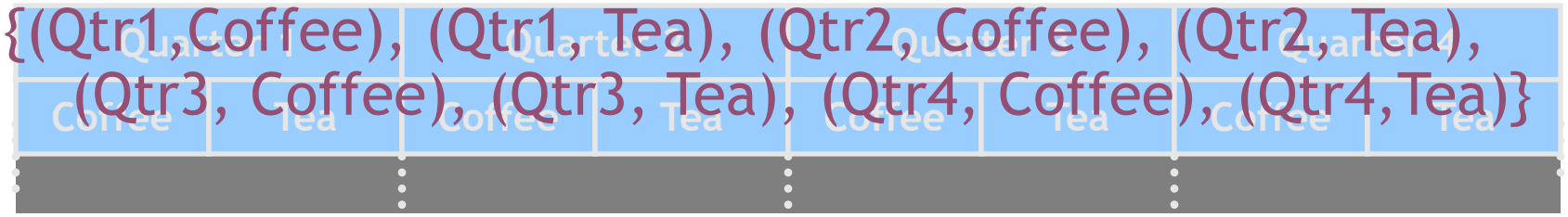


Polaris: Cross Operands

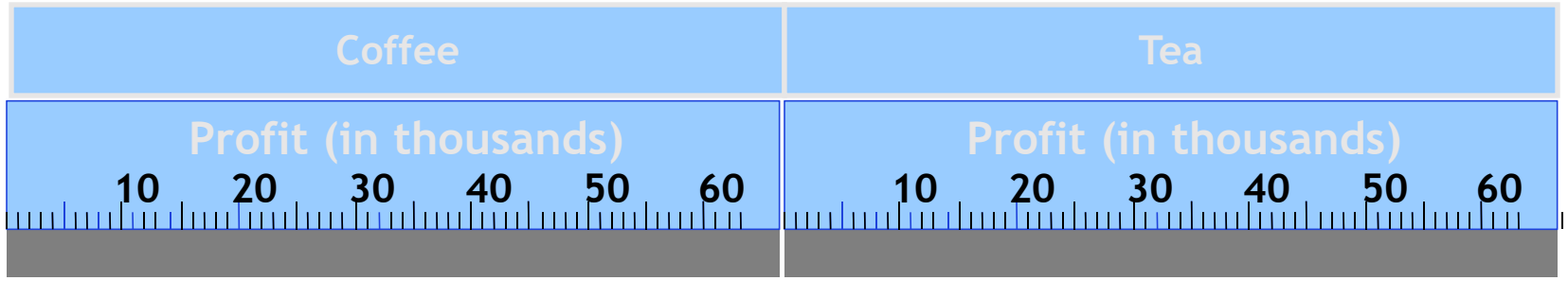
Cross-product of set interpretations:

QUARTER X PRODUCT_TYPE =

{(Qtr1, Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea),
(Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4, Tea)}



PRODUCT_TYPE X PROFIT =



Polaris: Nest Operands

QUARTER X MONTH

would create entry twelve entries for each quarter i.e. (Qtr1, December)

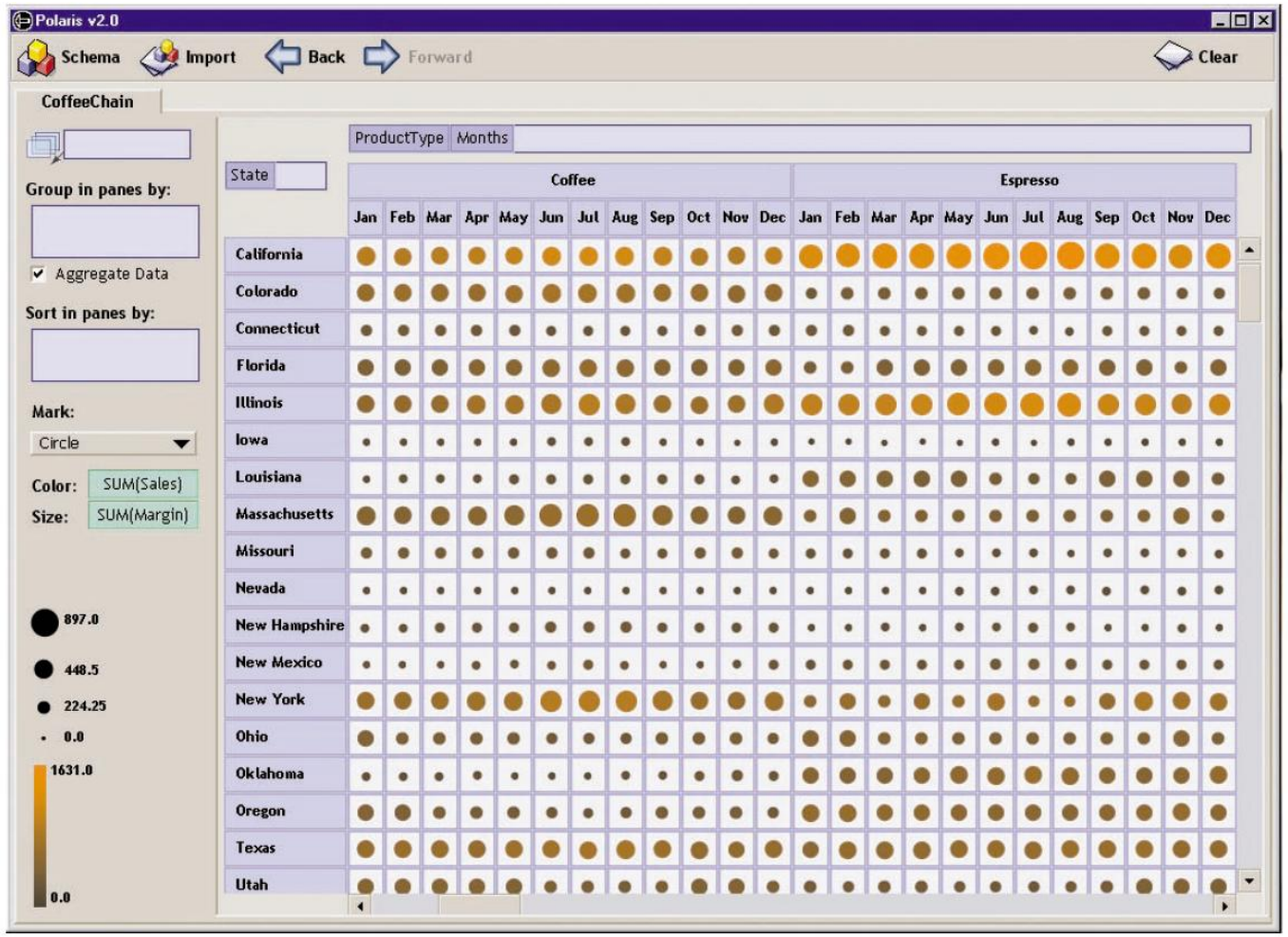
QUARTER / MONTH

would only create three entries per quarter

Qtr1			Qtr2			Qtr3			Qtr4		
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec

Polaris: formalism

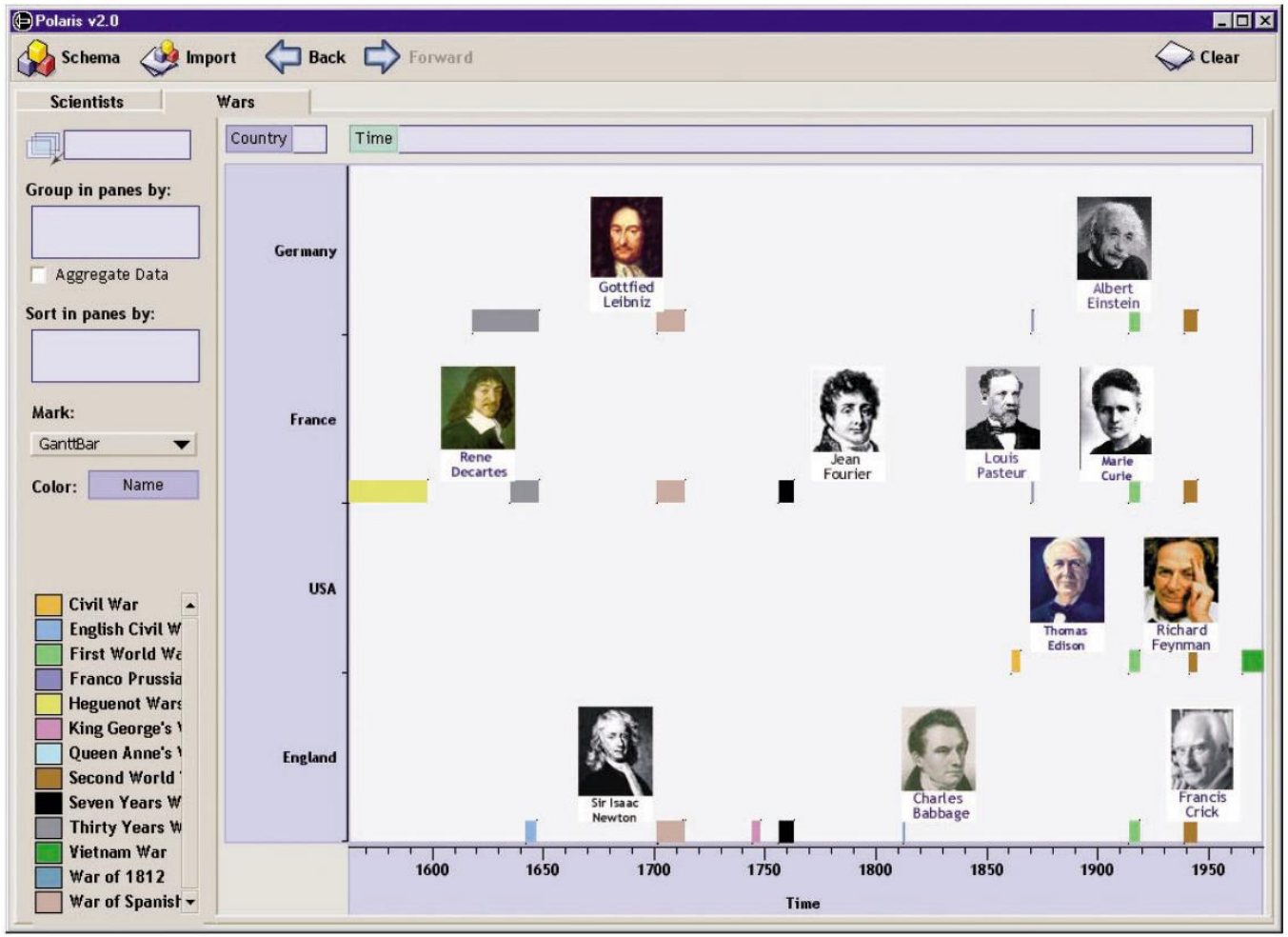
Specification of different graph types



Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: formalism

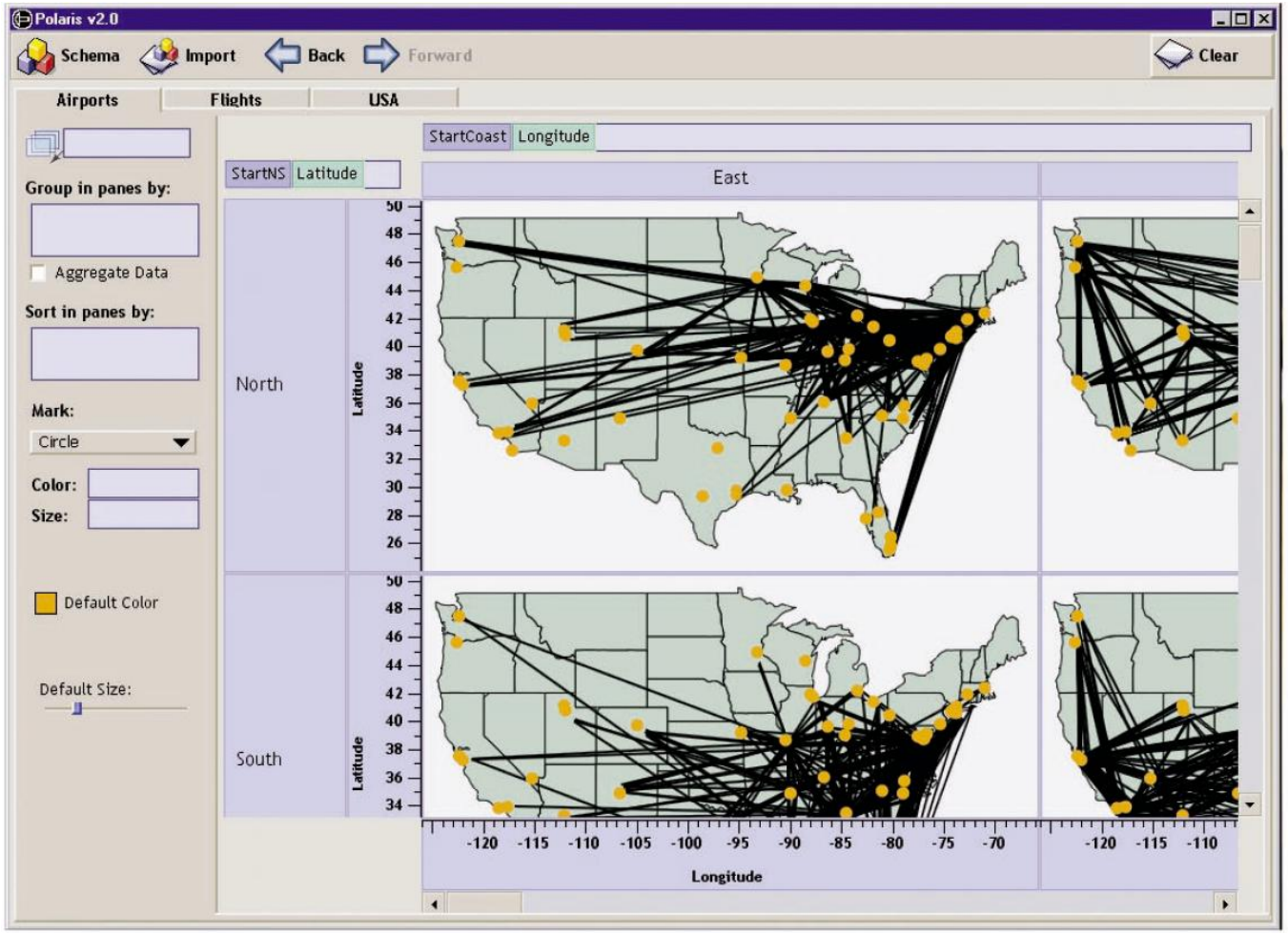
Specification of different graph types



Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: formalism

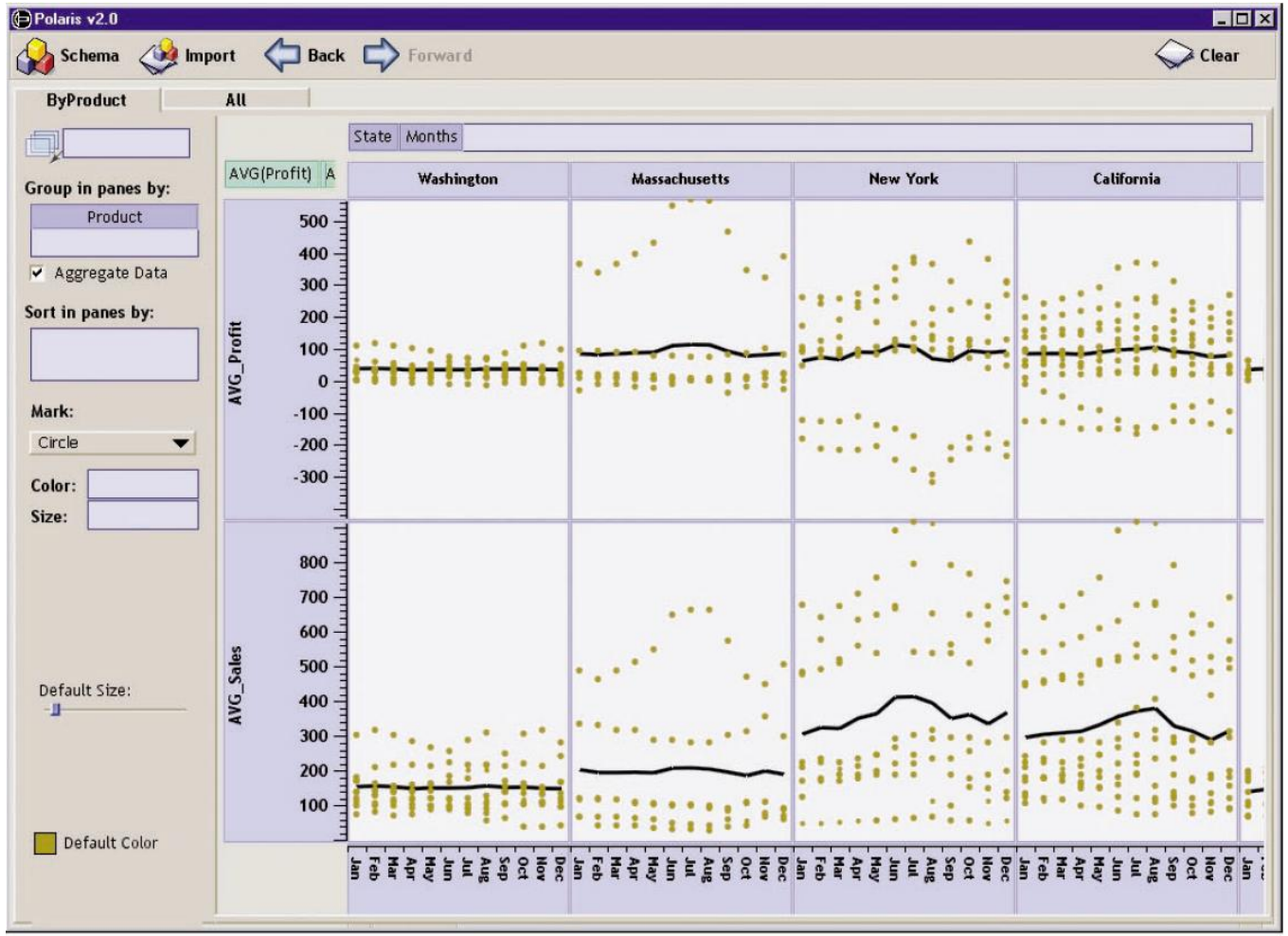
Specification of different graph types



Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: formalism

Specification of different graph types



Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: formalism

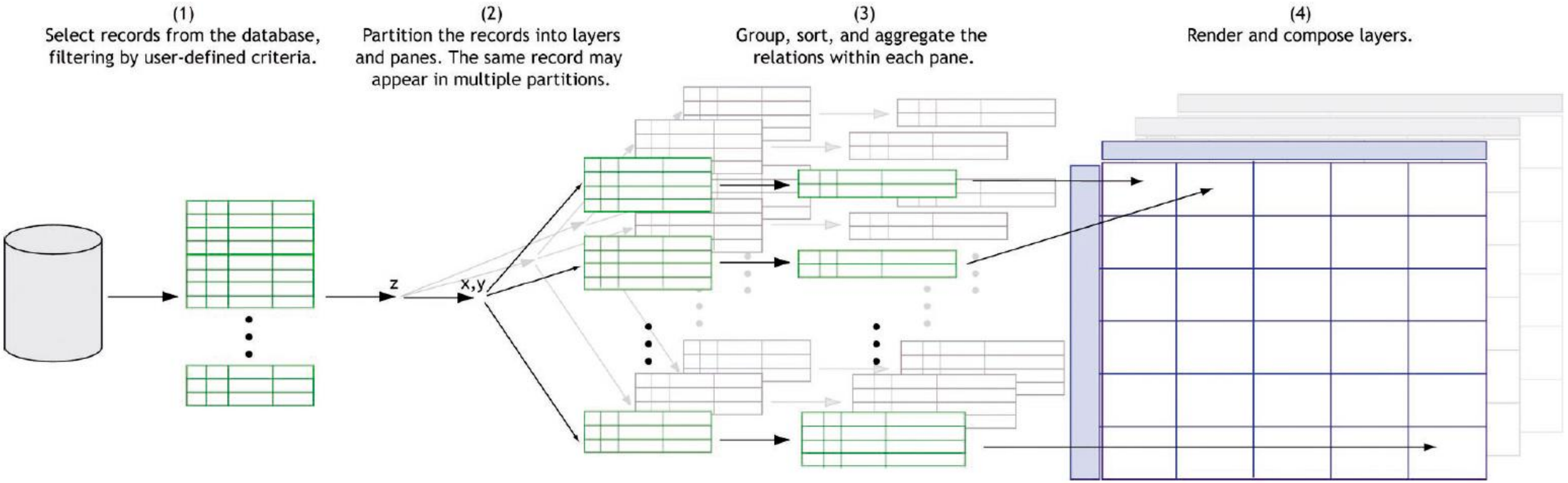
Encoding of data as retinal properties of marks in graphs

property	marks	ordinal/nominal mapping	quantitative mapping
shape	glyph	○ □ + △ S U	
size	rectangle, circle, glyph, text		
orientation	rectangle, line, text		
color	rectangle, circle, line, glyph, y-bar, x-bar, text, gantt bar		min … max

Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

Polaris: formalism

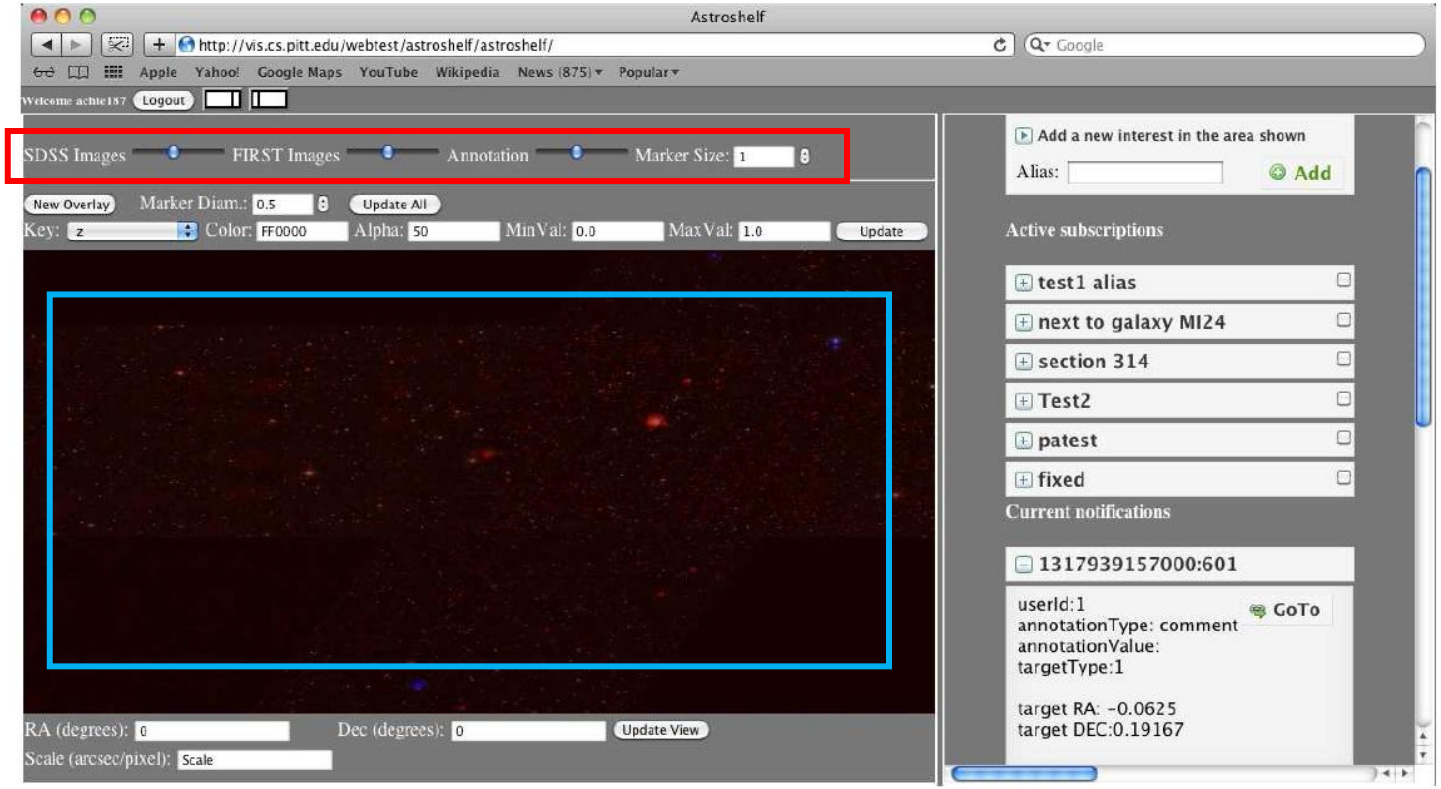
Translation of visual specification into SQL queries



Polaris: a system for query, analysis and visualization of multidimensional relational databses. Christo Stolte, Diane Tang, Pat Hanrahan. TVCG 2002.

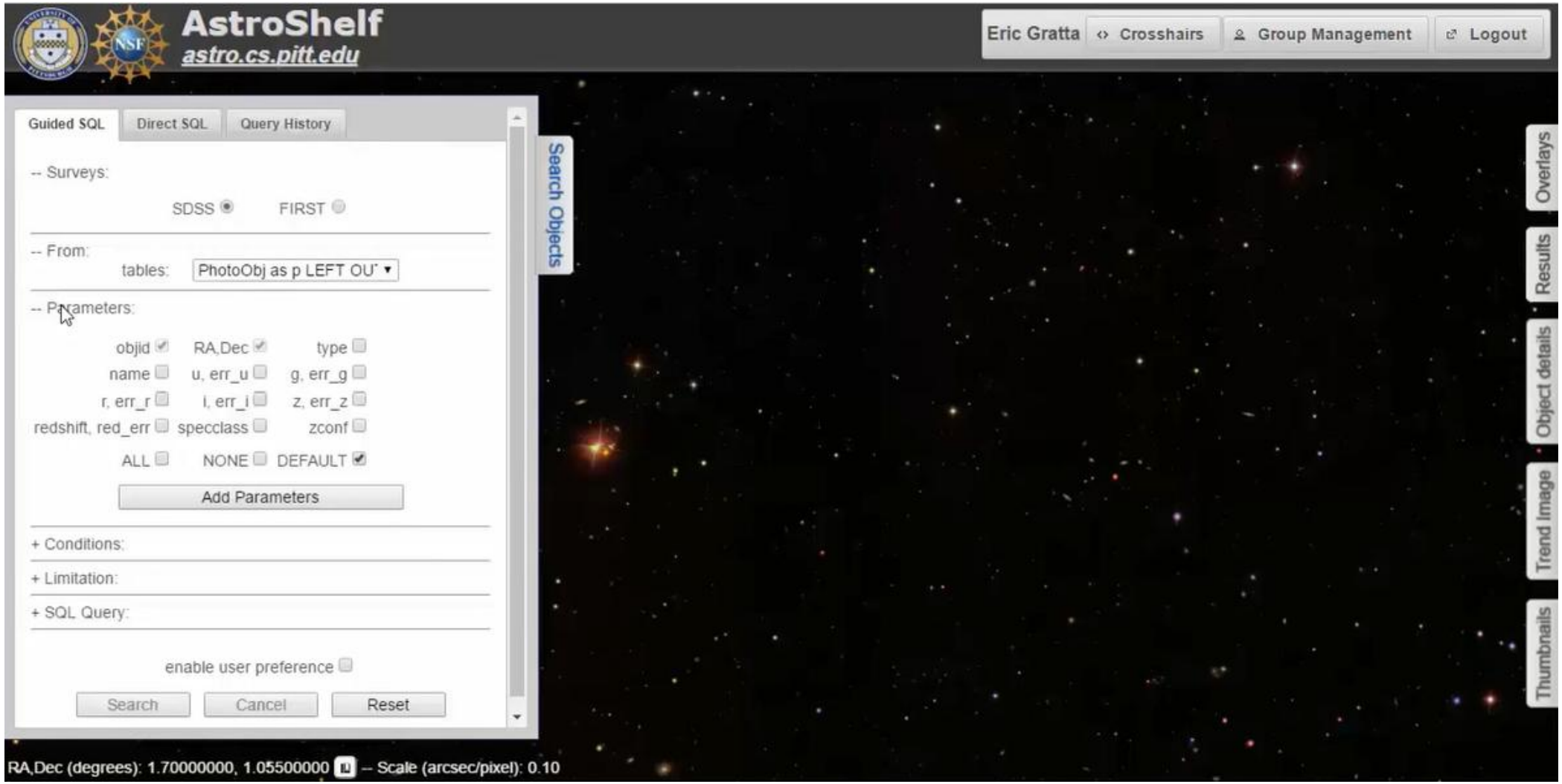
AstroShelf: goals

- **Scalable annotation framework**
- **Continuous workflow enactment system**



AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

AstroShelf

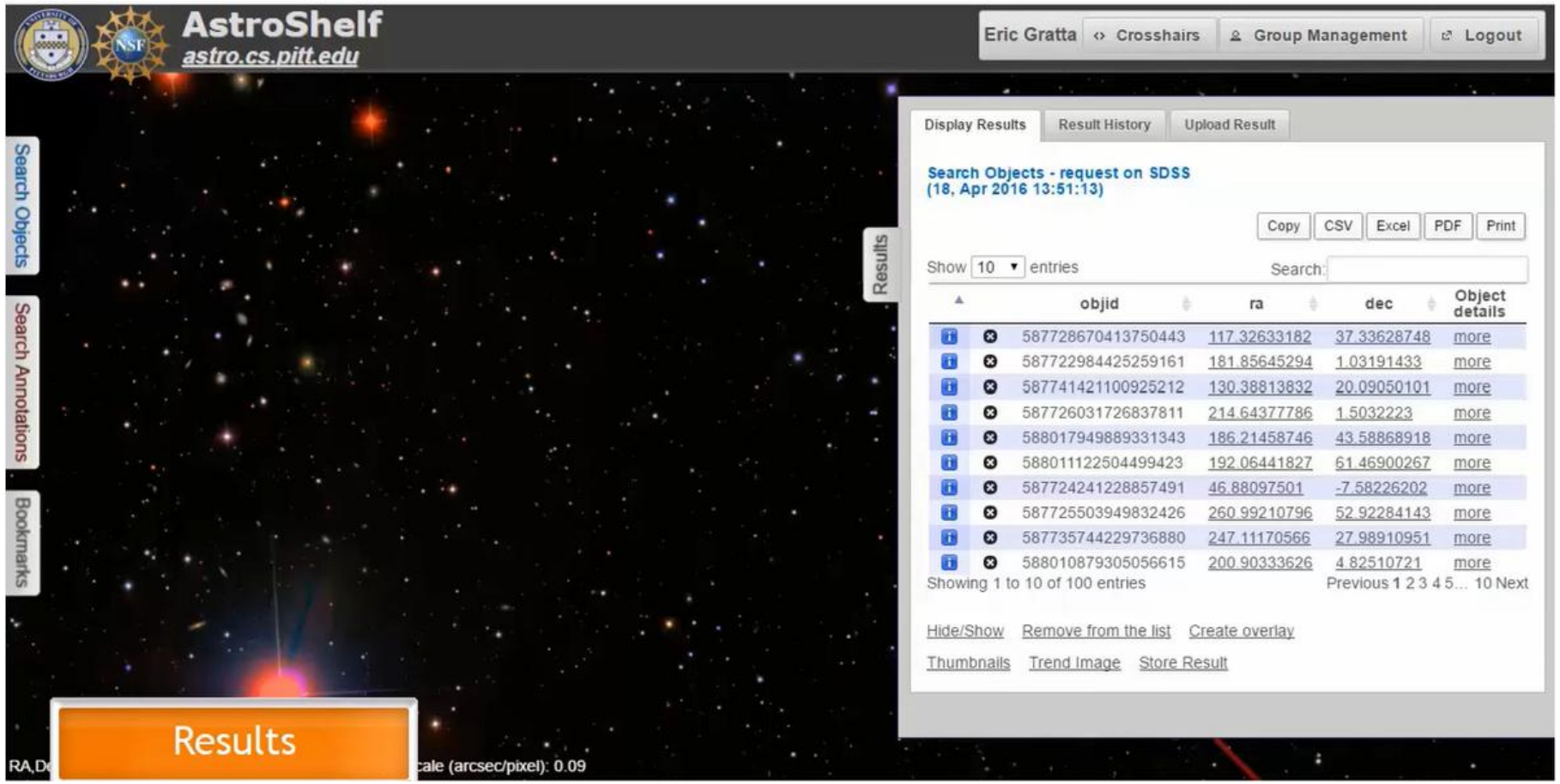


AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

AstroShelf

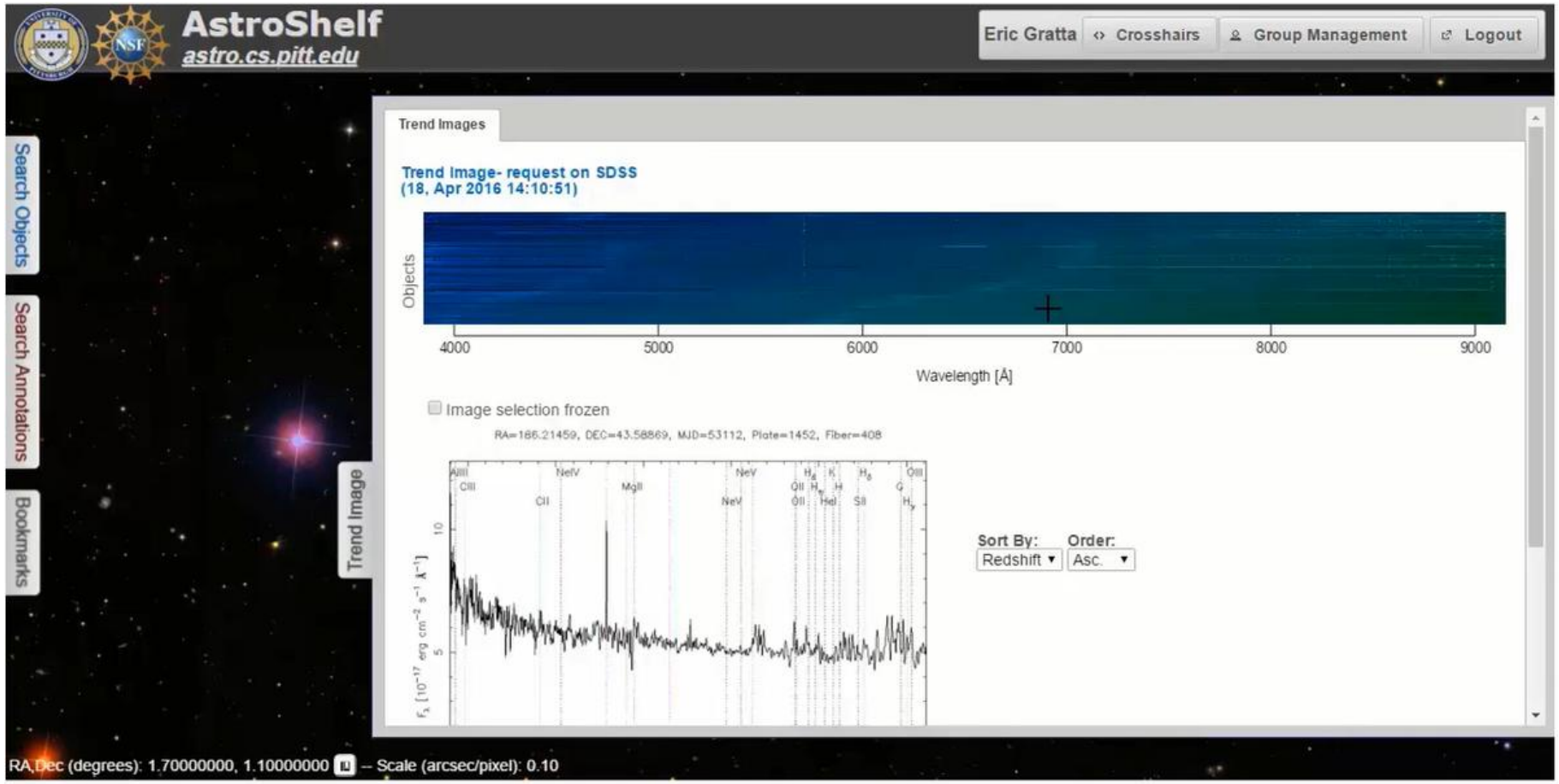
AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

AstroShelf



AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

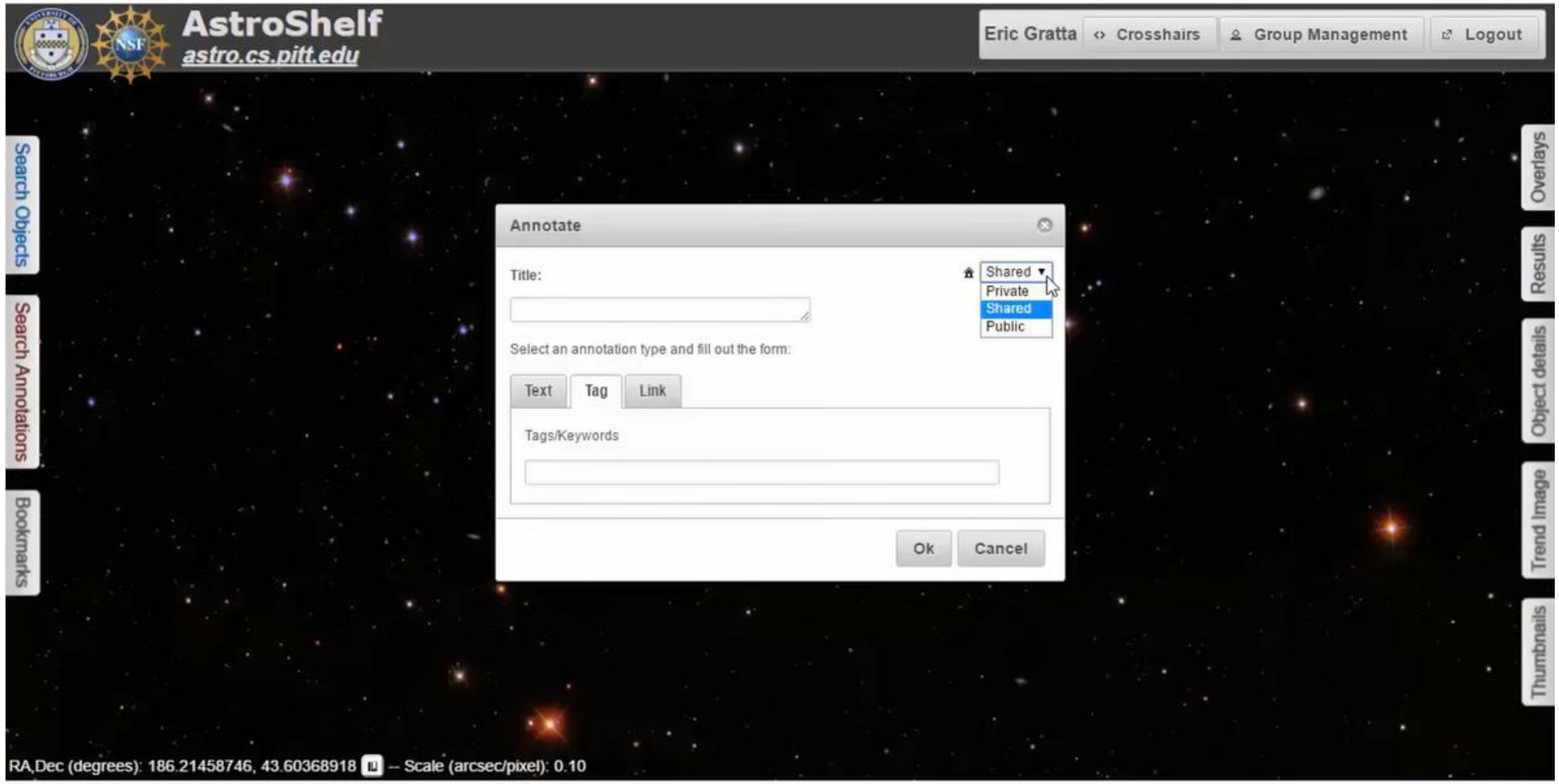
AstroShelf



Trend images are used to visualize overall patterns in a data set.

AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

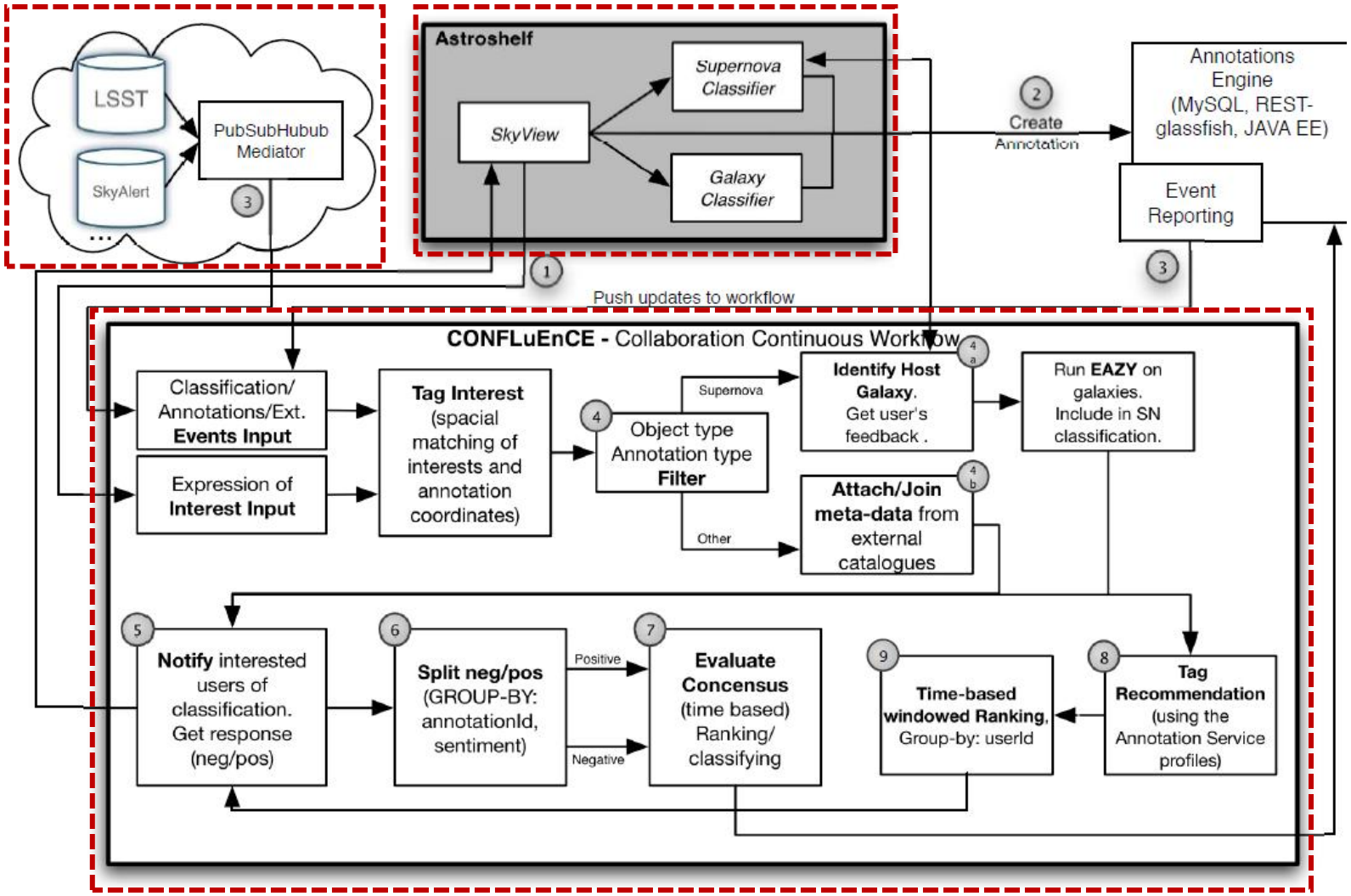
AstroShelf



Annotations: *<annotation type, annotation value>*.

AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

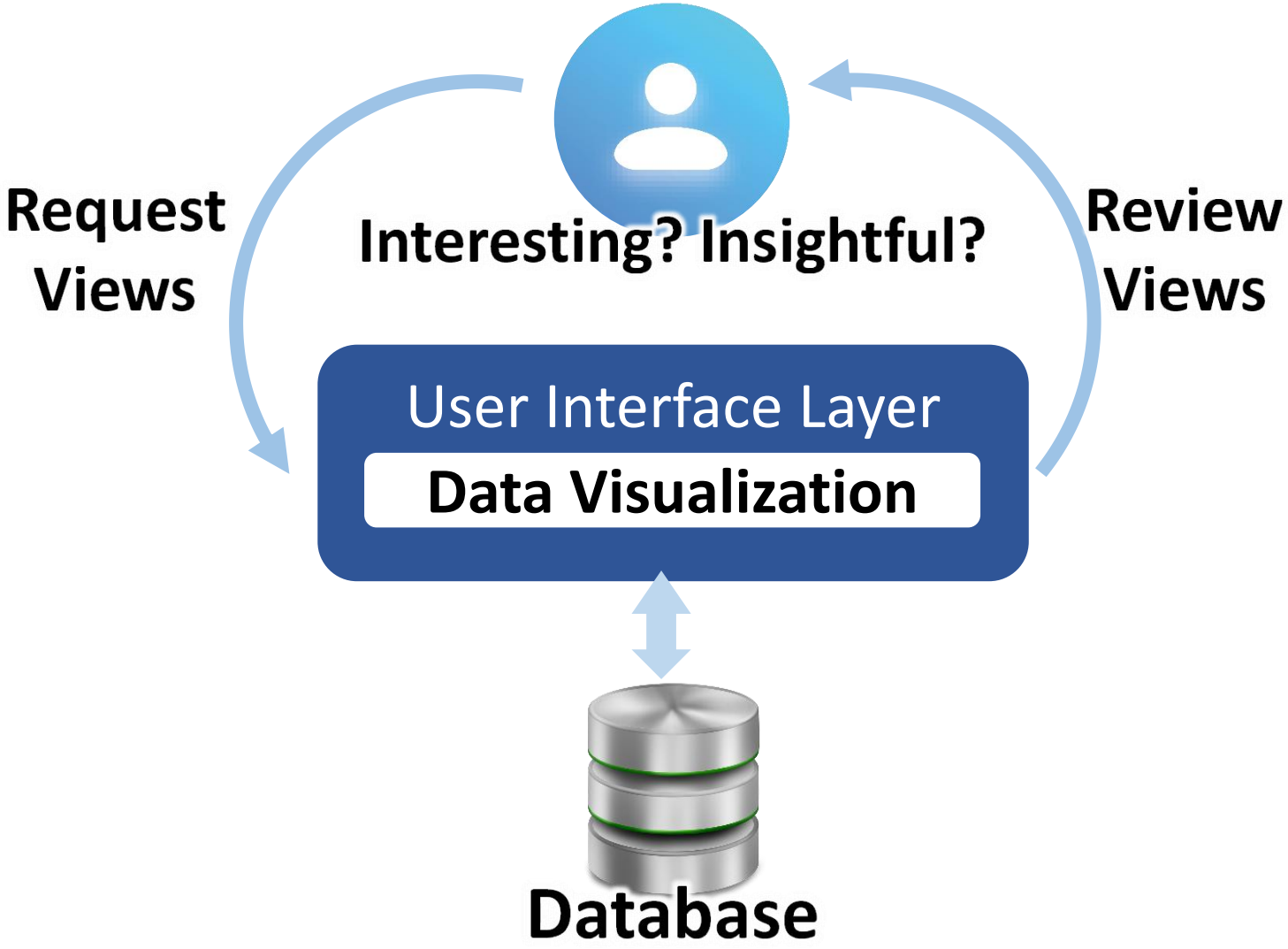
AstroShelf



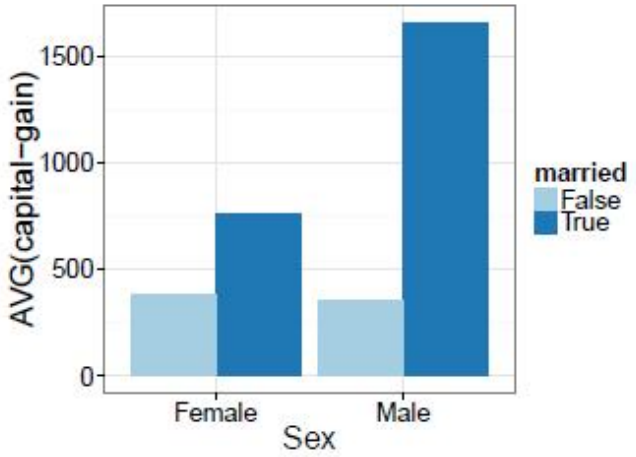
AstroShelf: understanding the universe through scalable navigation of galaxy of annotations. Neophytou et al. SIGMOD 2012.

Automatic Visualization

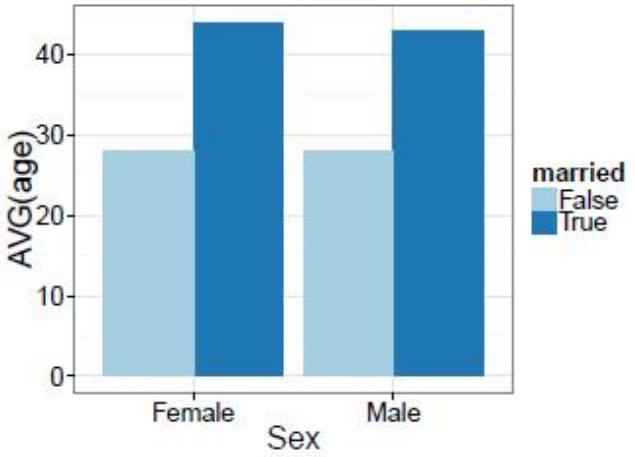
manual, repetitive explorationfor best visualization(s)



SeeDB



(a) Interesting Visualization



(b) Uninteresting Visualization

Sex	Avg Capital Gain
Unmarried Adults	
Female	380
Male	356
Married Adults	
Female	758
Male	1657

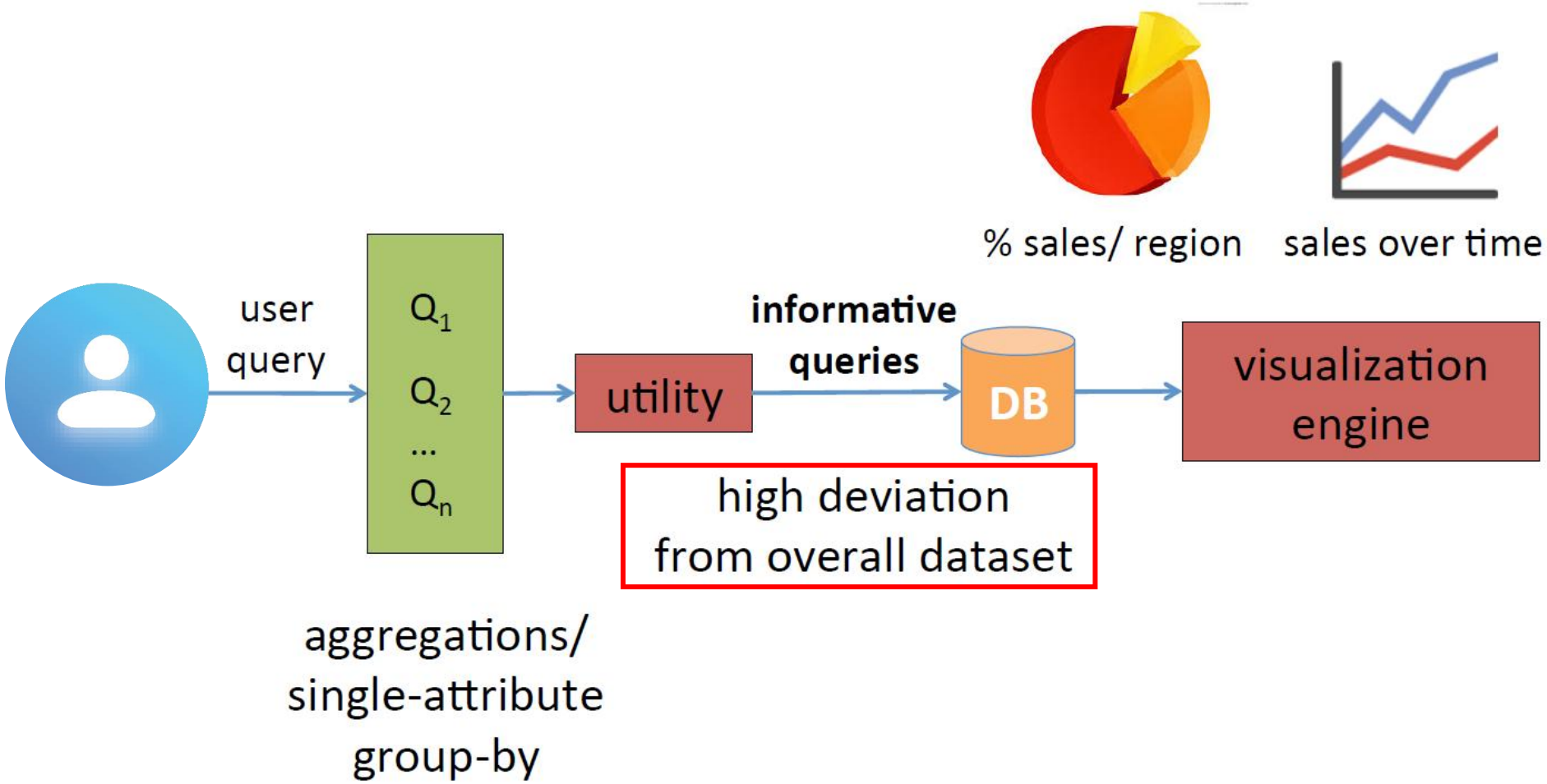
(c) Data: Avg Capital Gain vs. Sex

Sex	Avg Age
Unmarried Adults	
Female	28
Male	28
Married Adults	
Female	44
Male	43

(d) Data: Avg Age vs. Sex

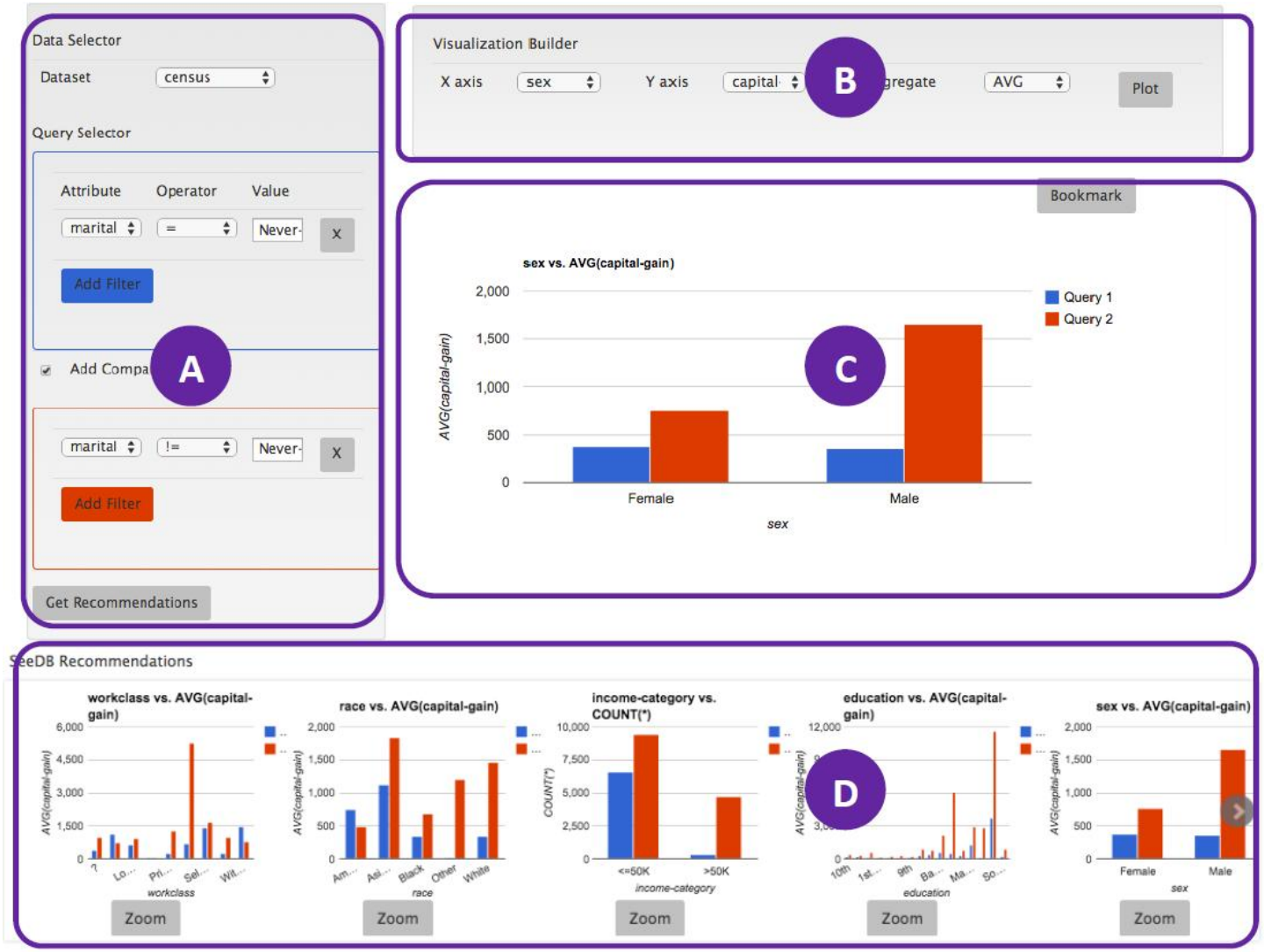
SEEDB: efficient data-driven visualization recommendations to support visual analytics. Vartak et al. VLDB 2015.

SeeDB



SEEDB: efficient data-driven visualization recommendations to support visual analytics. Vartak et al. VLDB 2015.

SeeDB



SEEDB: efficient data-driven visualization recommendations to support visual analytics. Vartak et al. VLDB 2015.

SeeDB

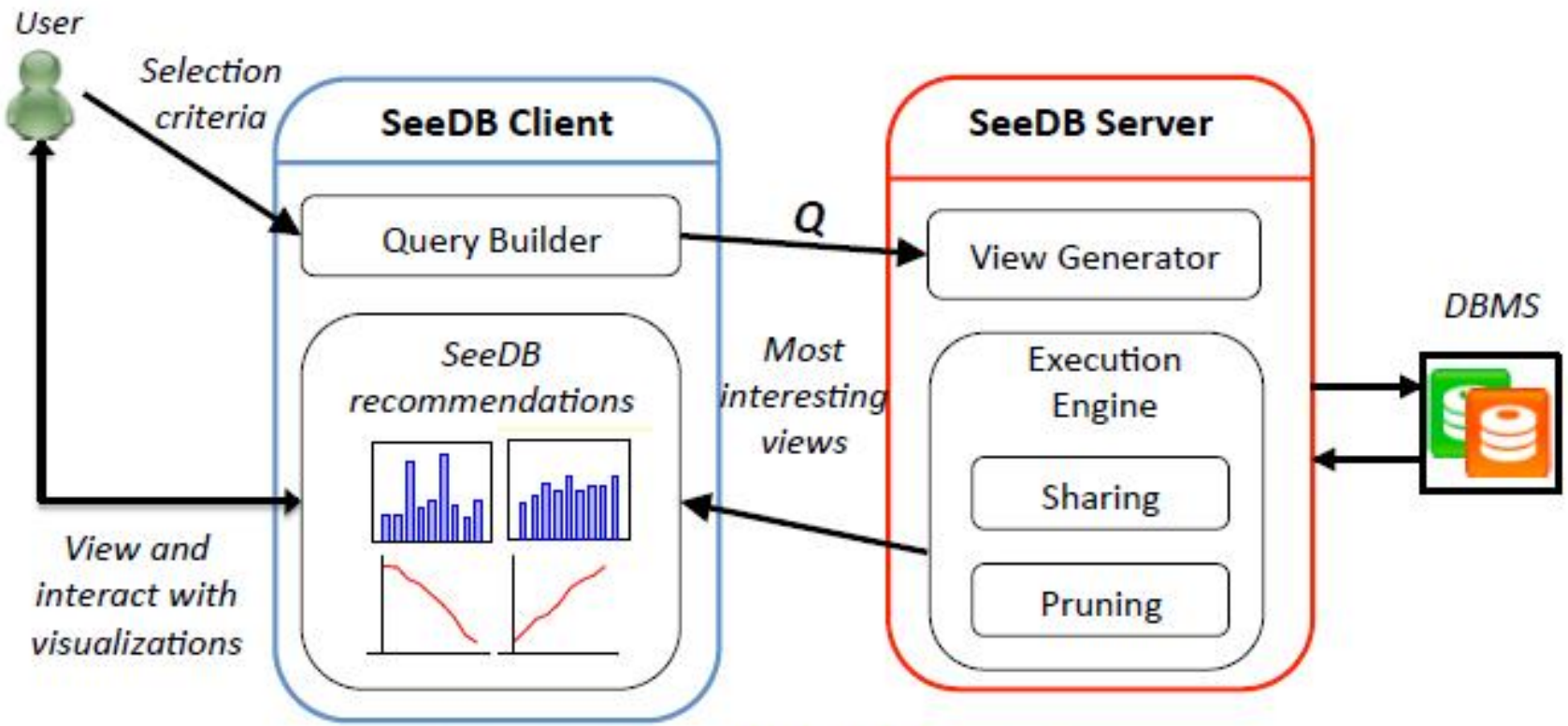


Figure 3: SeeDB Architecture

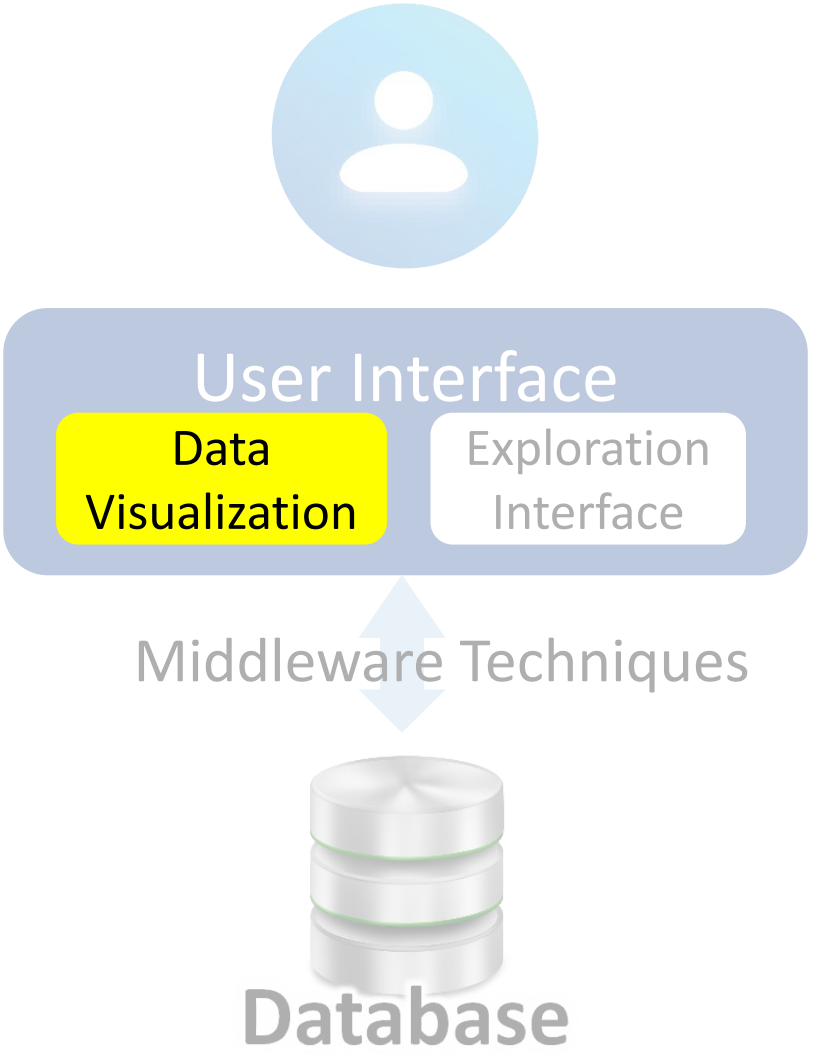
SEEDB: efficient data-driven visualization recommendations to support visual analytics. Vartak et al. VLDB 2015.

IDE: Data Visualization

Visualization Tools

Visual Optimizations

Automatic Visualization



单选题 1分

设置

探索式数据分析跟传统数据分析相比有什么不同?

A 探索式数据分析不涉及数据库操作

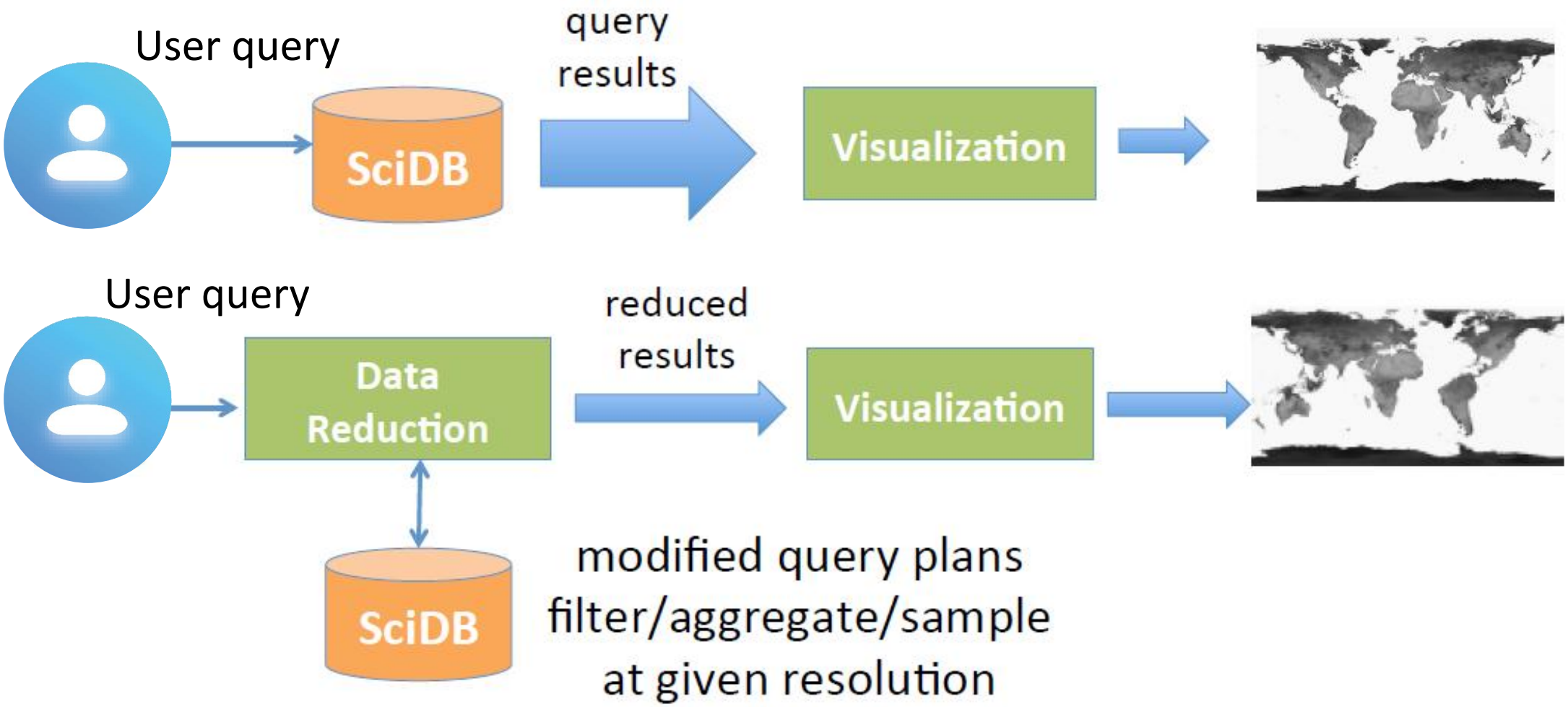
B 传统数据分析不涉及可视化

C 探索式数据分析中用户难以通直接通过查询获得想要分析的信息

提交

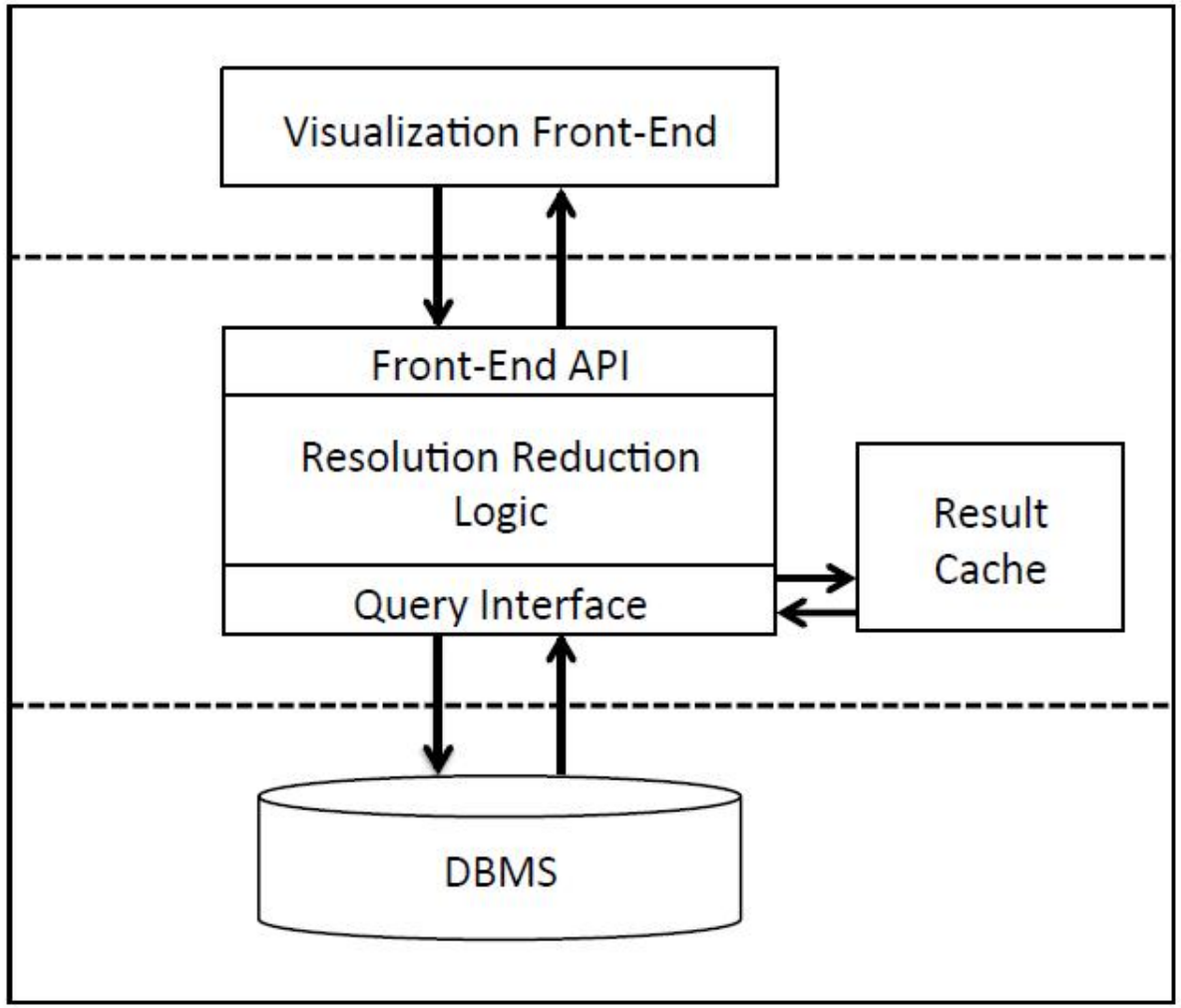
Visual Optimization

ScalaR: Resolution Reduction



Dynamic Reduction of Query Result Sets for Interactive Visualizations. Leilani Battle, Michael Stonebraker, Remco Chang. VIS 2013.

ScalaR: resolution reduction



Dynamic Reduction of Query Result Sets for Interactive Visualizations. Leilani Battle, Michael Stonebraker, Remco Chang. VIS 2013.

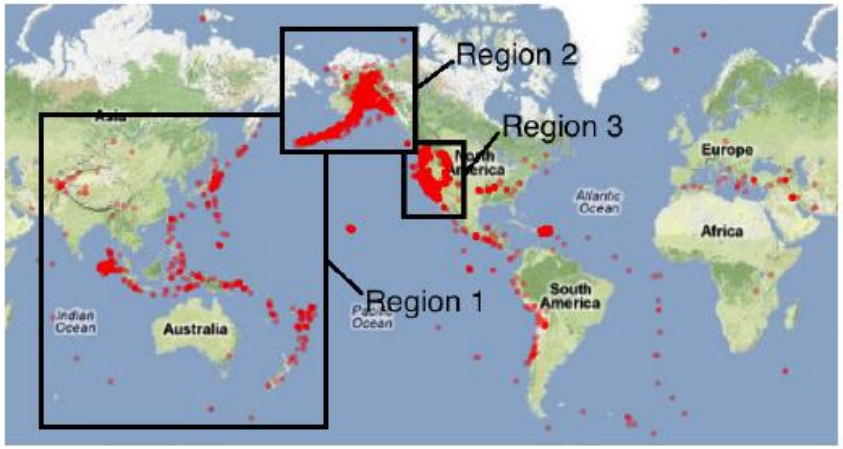
ScalaR: resolution reduction

Resolution Reduction Techniques:

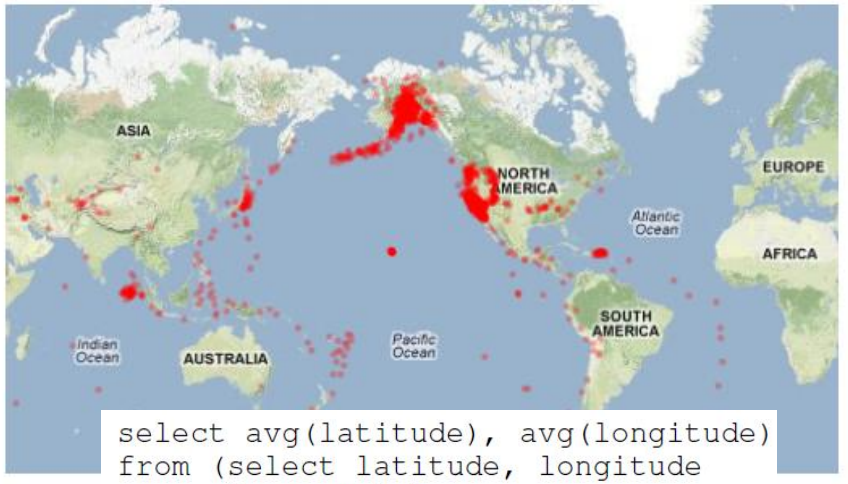
- **Aggregation**: group the data into n sub-matrices and return summaries over the sub-matrices
- **Sampling**: given a probability value p, return roughly that fraction of data as the result, where p*|data| = n
- **Filtering**: given a set of filters over the data, return the elements that pass these filters

Dynamic Reduction of Query Result Sets for Interactive Visualizations. Leilani Battle, Michael Stonebraker, Remco Chang. VIS 2013.

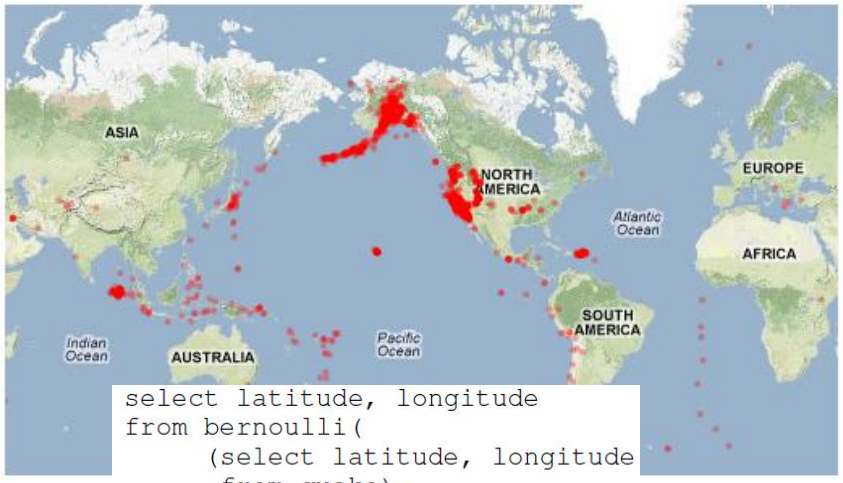
ScalaR



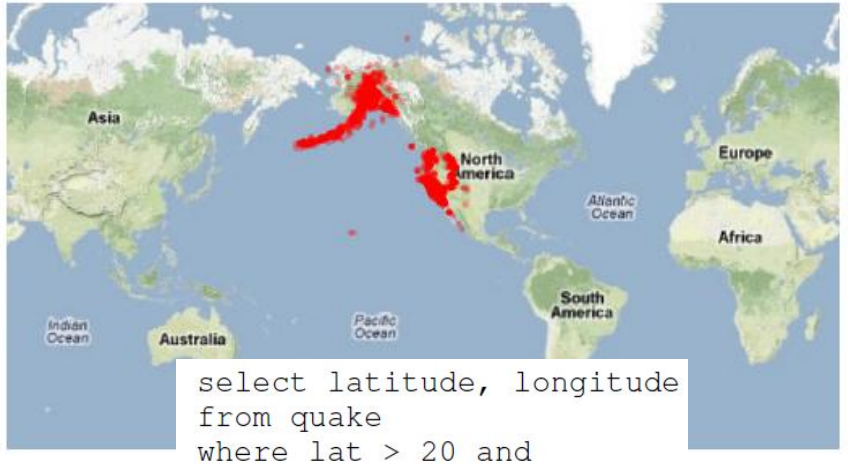
6381*6543 (a) Original query




```
select avg(latitude), avg(longitude)
from (select latitude, longitude
      from quake)
regrid 32, 33
```



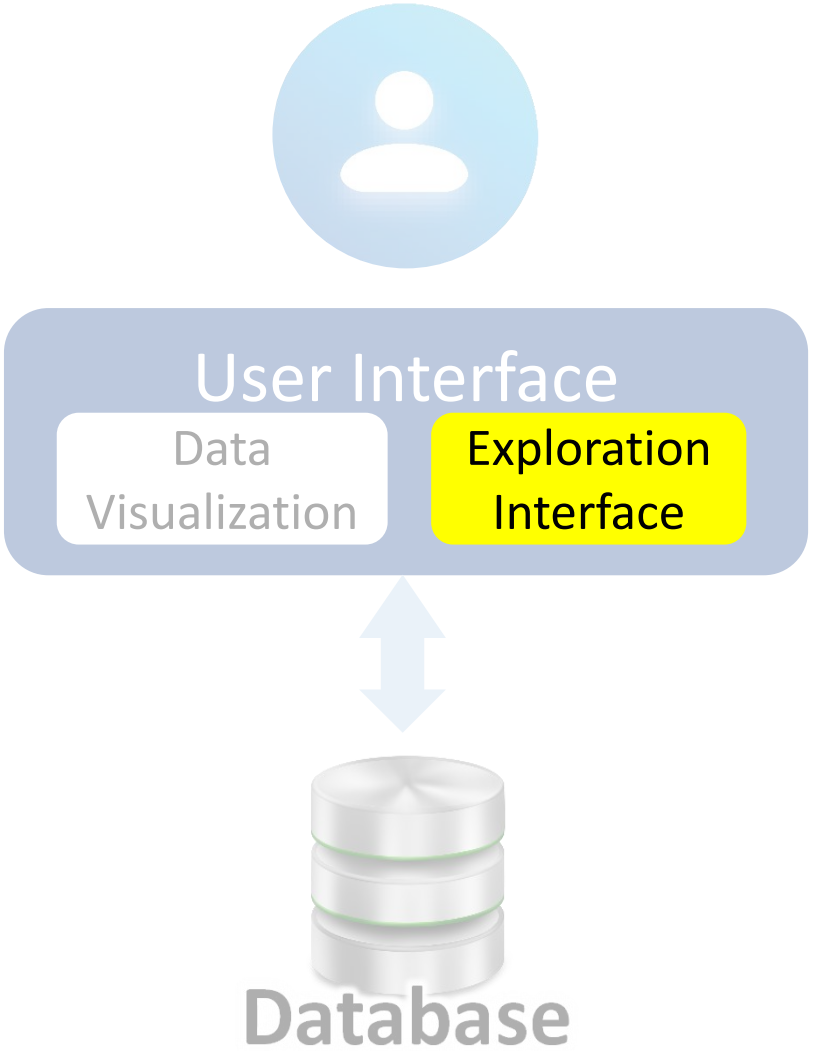

```
select latitude, longitude
from bernoulli(
     (select latitude, longitude
      from quake),
     0.327,
     1)
```




```
select latitude, longitude
from quake
where lat > 20 and
(lon < -100 or lon > 170)
```

Dynamic Reduction of Query Result Sets for Interactive Visualizations. Leilani Battle, Michael Stonebraker, Remco Chang. VIS 2013.

IDE: Exploration Interface

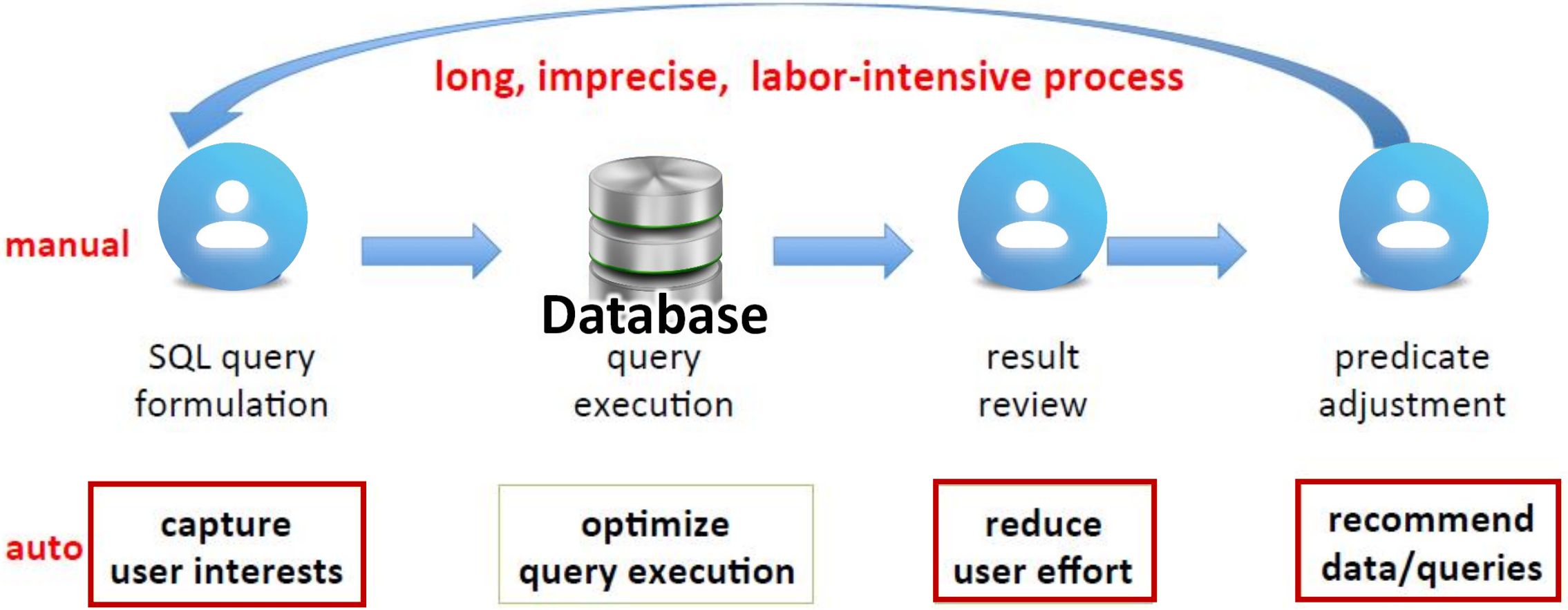


Automatic Exploration

Novel Query Interfaces

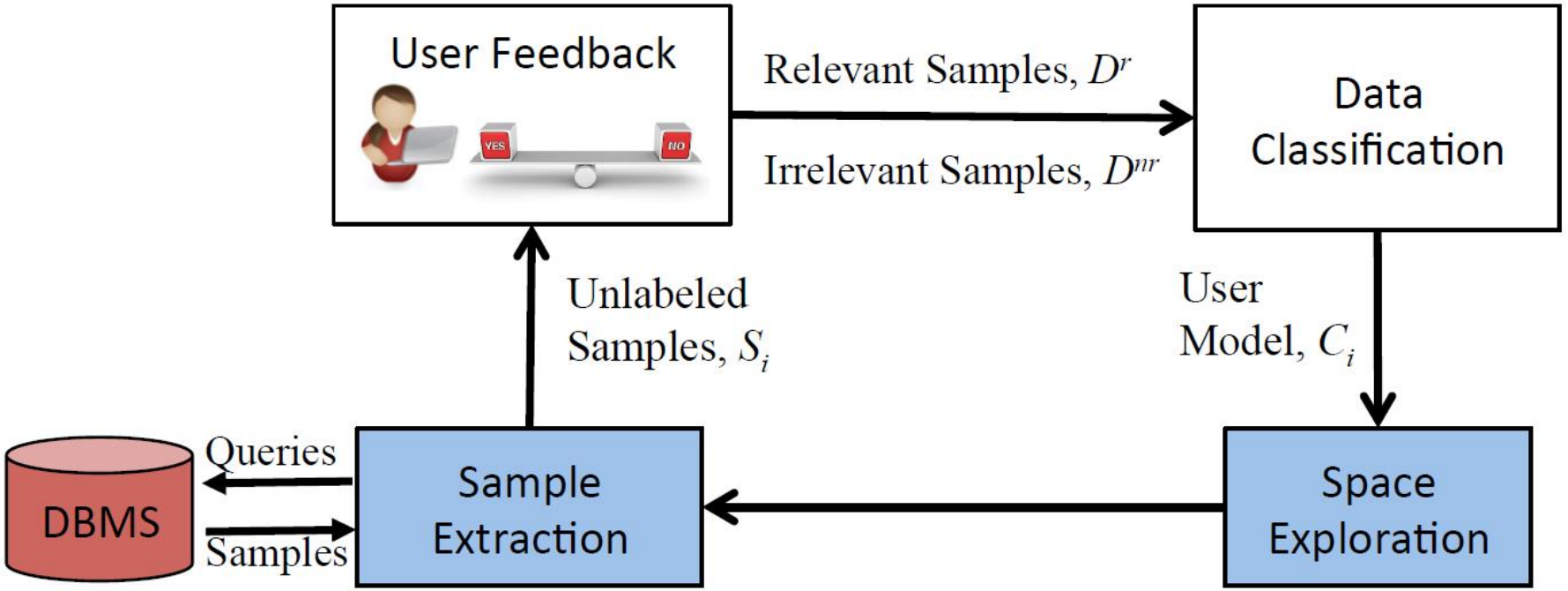
Automatic Exploration

Manual vs Automatic Exploration



AIDE

Explore by Example

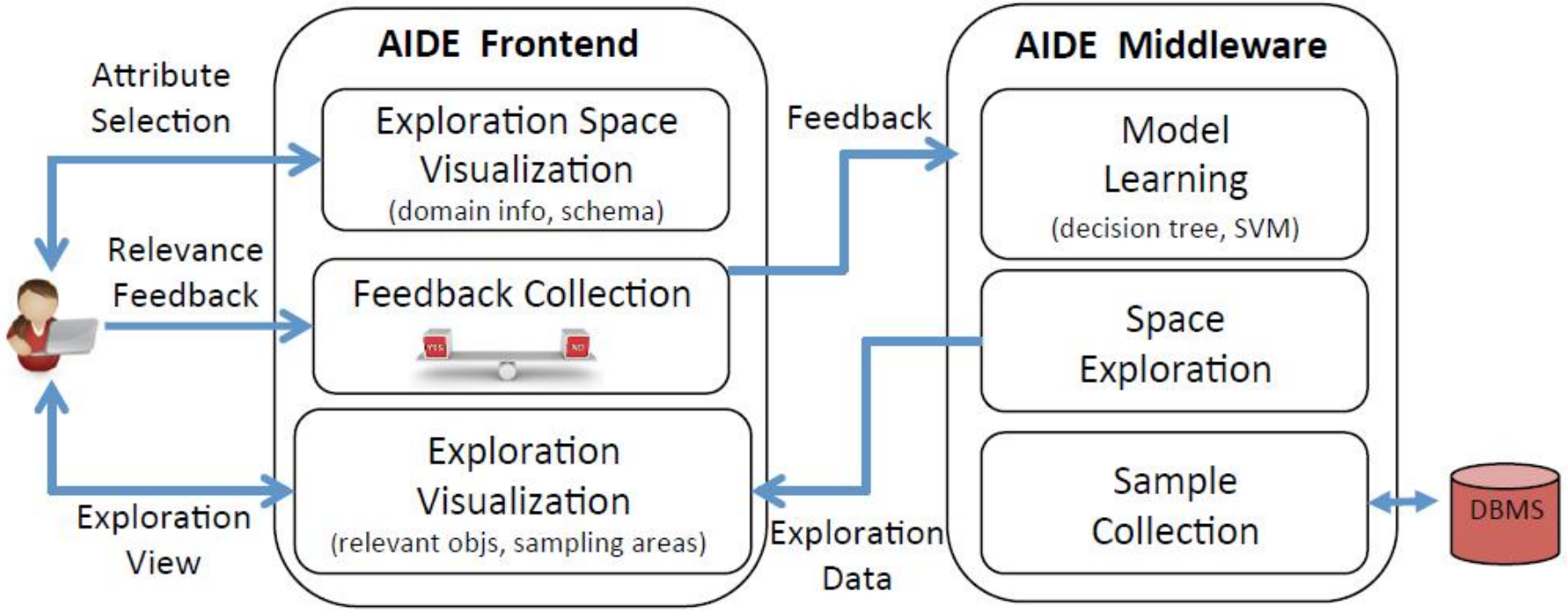


AIDE: an automatic user navigation system for interactive data exploration. Diao et al. VLDB 2015.

AIDE: an active learning-based approach for interactive data exploration. K. Dimitriadou, O. Papaemmanouil, Y. Diao. TKDE 2016.

AIDE

Explore by Example



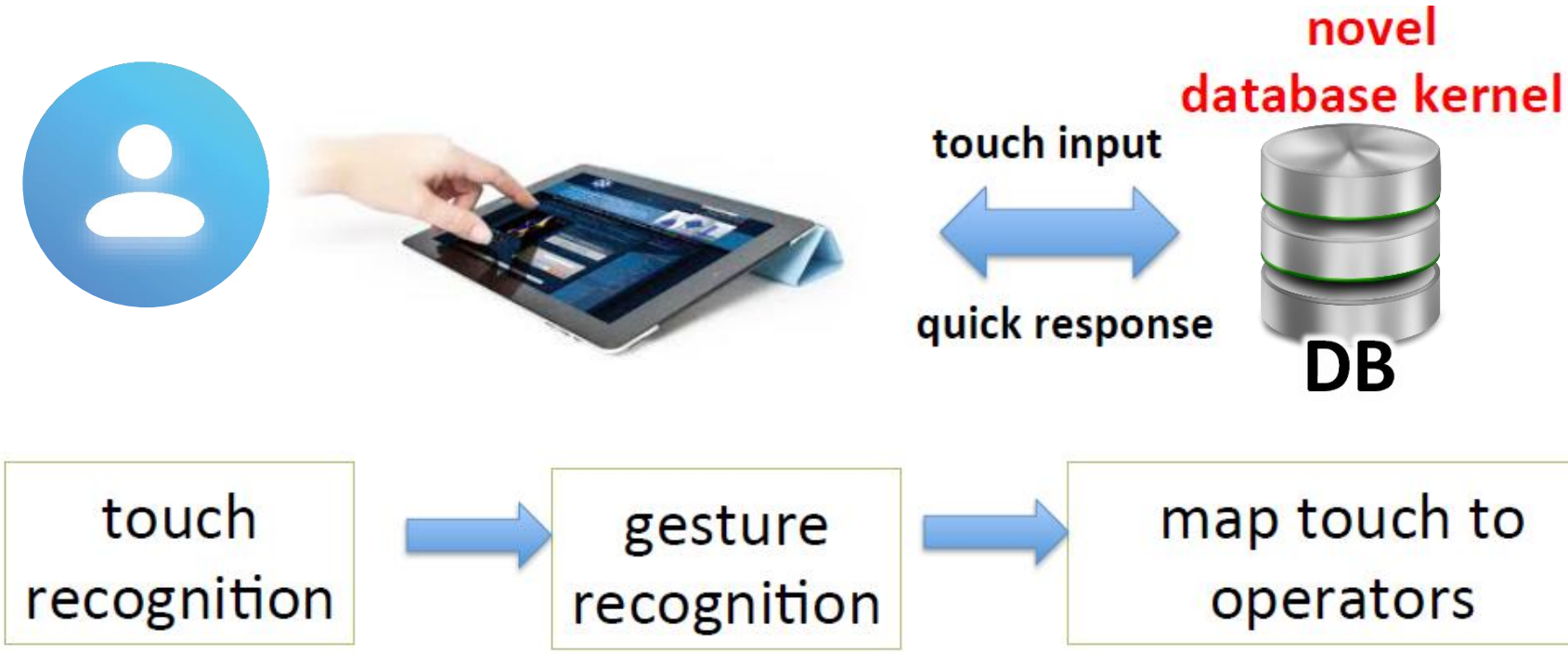
AIDE: an automatic user navigation system for interactive data exploration. Diao et al. VLDB 2015.

AIDE: an active learning-based approach for interactive data exploration. K. Dimitriadou, O. Papaemmanouil, Y. Diao. TKDE 2016.

Query Interfaces

dbTouch

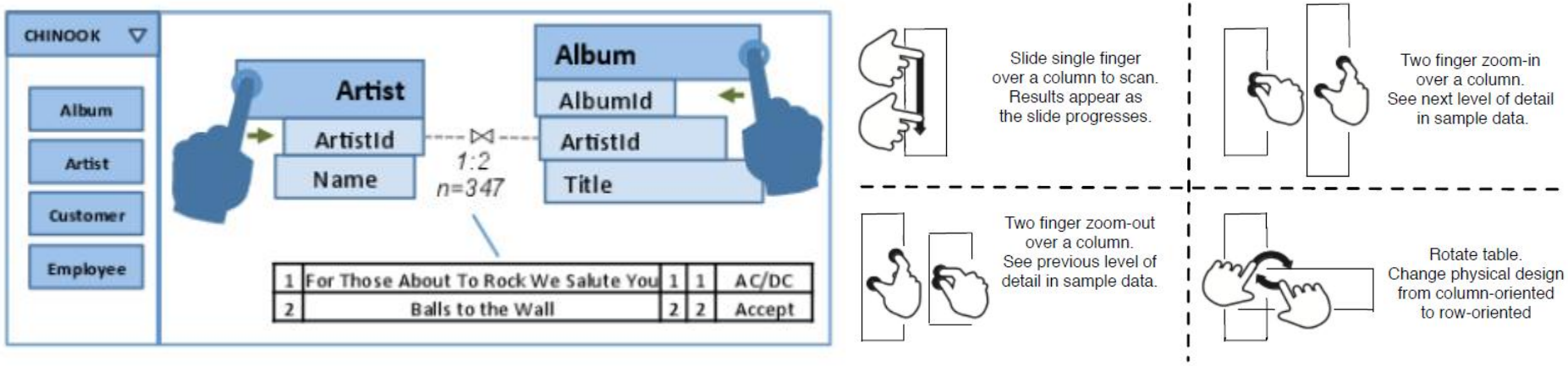
No-keyboard Interfaces



dbTouch: Analytics at your fingertips. Stratos Idreos, Erietta Liarou. ICDE 2014.

dbTouch

No-keyboard Interfaces



query context + gesture recognition = query intend

query space + search pattern = query template

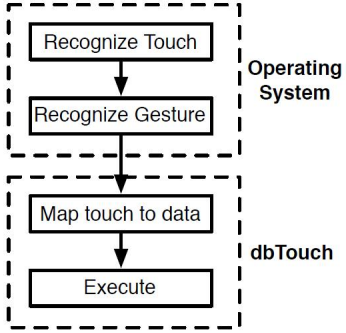
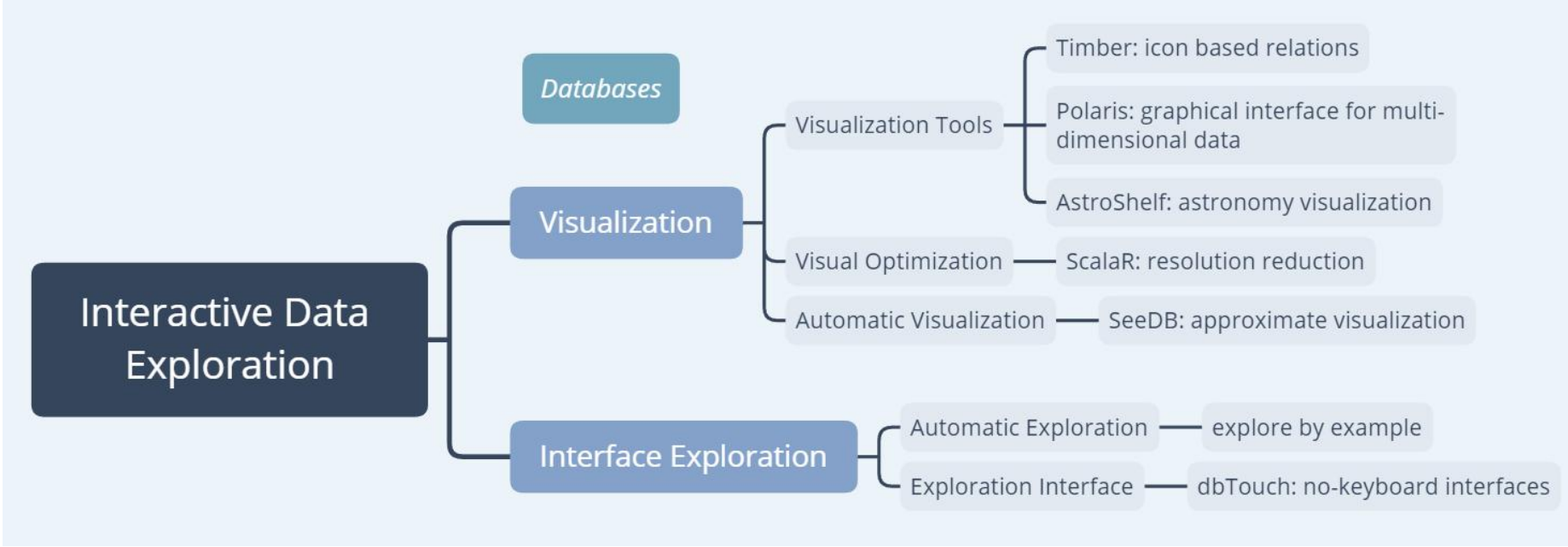


Figure 3: dbTouch system layers.

dbTouch: Analytics at your fingertips. Stratos Idreos, Erietta Liarou. ICDE 2014.

Summary

- Big data is significant for artificial intelligence and human intelligence
- Project-driven course
- Interactive data exploration systems



Thank You

QQ群：1057859502