Introduction & Organization

Systems for Interactive Data Exploration

Carsten Binnig
Brown University
Important Facts

Faculty: Tim Kraska & Carsten Binnig

Time: Wed 3:00-5:20pm (today: 3:00-4:00pm)

Type: 2000-level, Seminar-style (Paper Reading/Writing + Systems-Building)

Max. Participants: ~30 Students (2000-level)

Pre-requisites:
• CSCI 0320 (Software Engineering)
• CSCI 0330 (Computer Systems)
• One of CSCI 1270, CSCI 1951-A, CSCI 1670 (DB, DS or OS)
Data Exploration Vision
Data Exploration Vision

Avatar (2009)
Today’s User Interfaces
Today’s Big Data Backends

Current Generation:

- Flink
- Myria
- Spark
- Hadoop

...
Data Exploration “Bottlenecks”

- Text-based Input & Output
- Slow & Batched Execution
- Long Data Loading
Focus of this Seminar

Visual Interactive Data Exploration

Next Generation Big Data Systems?

Data Ingest

Progressive High-speed Execution

Avoid Data Loading Cost

Domain Expert Data Scientist
Our System:
Vizdom / IDEA
Challenge 1:
Interactive Latencies

The Effects of Interactive Latency on Exploratory Visual Analysis
Zhicheng Liu and Jeffrey Heer

In this research, we have found that interactive latency can play an important role in shaping user behavior and impacts the outcomes of exploratory visual analysis. Delays of 500ms incurred significant costs, decreasing user activity and data set coverage while reducing rates of observation, generalization and hypothesis. Moreover, initial exposure to higher latency interactions resulted in reduced rates of observation and generalization during subsequent analysis sessions in which full system performance was restored.

Techniques to achieve interactive response times
• **Offline Preparation**: Pre-compute, Indexing, Offline Sampling
• **Online Approximation**: Online Aggregation + Online Joins, Online Sampling, ...
Online Aggregation

SELECT major, AVG(final_grade) FROM grades GROUP BY major

<table>
<thead>
<tr>
<th>id</th>
<th>major</th>
<th>final_grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2.34</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3.45</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2.79</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>3.01</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>9876345</td>
<td>2</td>
<td>2.66</td>
</tr>
</tbody>
</table>

Challenges:
- Other operations such as joins, nested queries
- Rare events will not be visible in interactive time
Challenge 2: Rare Events & Tail of Distributions

Solution: Stratified Sampling?
- workload must be known
- requires much memory

Idea: Online Outlier Index

```
50k  t_x
100k...
150k...
200k...
250k...
500k...
1m  t_z
```
Challenge 3: Connect & Explore

Challenges:

• **Offline techniques**: no time for data loading + preparation (sampling, indexing, ...)
• **Online techniques**: Require a randomized stream from data source
Challenge 4: Interactive and Steerable ML

Machine learning (ML) algorithms are particularly time consuming.

Online learning algorithms: work in mini batches and can incrementally update a model using individual data items.

However, online learning are not optimized for user interactions:
- Are not optimized for quick response times but for maximal benefit.
- Do not allow user input / steering while running (e.g., manually change centers of clustering algorithms).

Challenge: Need for interactive ML techniques that can be steered by users.
Challenge 5: Quantifying Risk

False Discoveries:

Other Pitfalls (Simpson Paradox):

Berkeley admissions (Fall 1973)

Visual Interactive Tools allow users to test many more hypothesis at a short amount of time (even worse: visual recommendation engines)

Challenges:

- Traditional methods for multi-comparison adjustment are too pessimistic. How to adapt for visual data exploration?
- How to efficiently detect statistical pitfalls online?
Opportunity 1: Human Perception

**Intuition:** Do not compute what the user can not perceive

**Opportunity:** avoid unnecessary computation to **speed-up** visualizations / not waste resources
Opportunity 2: Think Time

Response time after interaction <500ms ... BUT Think time between interactions > 5-7s

Opportunity: use think time in between interactions to prepare for next step (e.g., for linking-and-brushing)
Opportunity 3: Visual Prediction

Prediction of what the user does next is a hard problem if it should be solved in a general way on the query level (e.g., SQL)

```
SELECT major, AVG(final_grade)
FROM grades
GROUP BY major
```

Next Query?

Opportunity: Visual interface restricts / implies certain interactions (e.g., user can only link visible visualizations)
Opportunity 4: Users Sessions & Reuse

**Queries** in interactive data exploration sessions **build on each other** (e.g., the modify filter of previous query)

**Opportunity:** Redesign data structures to **maximize effect of reuse of intermediate results** (e.g., compute only delta)
Opportunity 5: Modern Hardware

Most big data problems are not that big (< 1TB) + high-end hardware is becoming affordable

Large Commodity Clusters

Small High-Performance Clusters

Opportunity: Redesign systems for modern hardware + avoid unnecessary overhead (e.g., fault-tolerance)
Organization

Part 0: Introductions (09/07-09/14)

Part 1: Paper Reading & Presentations & Reviews (09/21-10/12)

Part 2: Systems Building & Research (10/19-12/07)

Presentations & Projects will be in groups of 2
Paper Reading & Presentation

Each group has to present one paper: 15 min presentation + 15 min discussion

Each student has to write a review for each paper (except for the one he/she is presenting) and defend review

The group which presents a paper must meta-review all reviews and moderate the discussion

More details about reviews and discussions next week
Systems Building & Research

Work in parallel on the next-generation system (A-Ware) for interactive data exploration

Pick a research challenge to be integrated into the system (e.g., online clustering, data ingest from HDSF, online indexing, ...) 

We provide an intro to the code & do regular 1-on-1 meetings

Each group has to write a short research paper (6 pages) and do two systems demonstrations (mid-and final-demo)

More details about project ideas during paper presentations
Grading

Projects 50%

Presentations 15%

Reviews 15%

Participation 20%
Next Steps

Register yourself using the form on the course website as soon as you are sure that you want to take the class.

We will hand out overwrite codes each Friday (max. participants is 30).

Please find a partner + select three papers from reading list.

Enter your partner + paper wishes in group registration form on class website (latest by 09/20) or better before.

First paper presentations start 09/21 -> we need to assign first papers by begin/mid of next week!
Questions?