Speeding Up Data Science: From a Data Management Perspective

Jiannan Wang

Simon Fraser University
SFU DB/DM Group

History

- Over 30 years of research experience in database and data mining
- Wrote a Data Mining Textbook widely used in the world
- Invented many famous data mining algorithms (e.g., FP-Growth, DBScan, Prefixspan)

```
Mining frequent patterns without candidate generation
J Han, J Pei, Y Yin
ACM sigmod record 29 (2), 1-12

Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth
J Pei, J Han, B Mortazavi-Asl, H Pinto, Q Chen, U Dayal, MC Hsu
icccn, 0215

A density-based algorithm for discovering clusters in large spatial databases with noise.
M Ester, HP Kriegel, J Sander, X Xu
Kdd 96 (34), 226-231
```
**SFU DB/DM Group**

- **Research Areas:** Machine Learning, Data Science, and Big Data Systems
- **Research Strengths:** Cloud Databases, Crowdsourced Data Management, Data Cleaning and Integration, Data Security and Privacy, Fraud Detection, Interpretable Machine Learning, Precision Medicine, Recommender Systems
- **Ranked 13\(^{th}\) in databases and data mining in North America** (source: csrankings.org)

<table>
<thead>
<tr>
<th>#</th>
<th>Institution</th>
<th>Count</th>
<th>Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Carnegie Mellon University</td>
<td>17.7</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>Univ. of Illinois at Urbana-Champaign</td>
<td>14.9</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Stanford University</td>
<td>13.0</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Georgia Institute of Technology</td>
<td>11.5</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>University of Michigan</td>
<td>11.5</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>Massachusetts Institute of Technology</td>
<td>10.3</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>Cornell University</td>
<td>10.2</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Purdue University</td>
<td>8.8</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>Pennsylvania State University</td>
<td>8.7</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>University of California - Los Angeles</td>
<td>8.6</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>University of Massachusetts Amherst</td>
<td>8.6</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>University of Illinois at Chicago</td>
<td>8.2</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>Simon Fraser University</td>
<td>8.1</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>University of Maryland - College Park</td>
<td>8.1</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>University of Waterloo</td>
<td>7.9</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>Duke University</td>
<td>7.6</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>University of California - Santa Barbara</td>
<td>7.6</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>University of California - Santa Cruz</td>
<td>7.5</td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>University of Wisconsin - Madison</td>
<td>7.4</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>Ohio State University</td>
<td>7.2</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>University of California - Riverside</td>
<td>7.0</td>
<td>10</td>
</tr>
<tr>
<td>22</td>
<td>University of California - San Diego</td>
<td>6.7</td>
<td>14</td>
</tr>
<tr>
<td>23</td>
<td>University at Buffalo</td>
<td>6.4</td>
<td>12</td>
</tr>
</tbody>
</table>

Ke Wang (Joined in 2000)  
Martin Ester (Joined in 2001)  
Jian Pei (Joined in 2004)  
Jiannan Wang (Joined in 2016)  
Tianzheng Wang (Joined in 2018)
My Lab’s Mission

Speeding Up Data Science
# Computer Science vs. Data Science

<table>
<thead>
<tr>
<th>What</th>
<th>When</th>
<th>Who</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>1950-</td>
<td>Software Engineer</td>
<td>Write software to make computers work</td>
</tr>
</tbody>
</table>

Plan → Design → Develop → Test → Deploy → Maintain

<table>
<thead>
<tr>
<th>What</th>
<th>When</th>
<th>Who</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science</td>
<td>2010-</td>
<td>Data Scientist</td>
<td>Extract insights from data to answer questions</td>
</tr>
</tbody>
</table>

Collect → Clean → Integrate → Analyze → Visualize → Communicate
Lab Members

Collect → Clean → Integrate → Analyze → Visualize → Communicate
Today’s Talk

**Deeper** [SIGMOD 2018 (Demo), SIGMOD 2019]
- Speed up data enrichment

**TARS** [VLDB 2019]
- Speed up data labeling

**AQP++** [SIGMOD 2018]
- Speed up data analysis
Deeper

Leverage Deep Web To
Speed Up Data Enrichment

Pei Wang
Yongjun He
Ryan Shea
Jiannan Wang
Eugene Wu

P. Wang et al. Deeper: A Data Enrichment System Powered by Deep Web. SIGMOD 2018 (demo)
P. Wang et al. Progressive Deep Web Crawling Through Keyword Queries For Data Enrichment. SIGMOD 2019
Why Data Enrichment?

Data Enrichment
- Extending a local DB with new attributes from external data sources

Various Use Cases
- For market analysis
- For ML model training
- For knowledge base construction

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Address</th>
<th>City</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boiling Point</td>
<td>4148 Main St</td>
<td>Vancouver</td>
<td>?</td>
</tr>
<tr>
<td>Flamingo Chinese</td>
<td>1652 SE Marine</td>
<td>Vancouver</td>
<td>?</td>
</tr>
<tr>
<td>Restaurant</td>
<td>Drive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun Sui Wah Restaurant</td>
<td>3888 Main Street</td>
<td>Vancouver</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The need for Deeper

Existing Solutions

◦ Leverage web tables [InfoGather SIGMOD 2012, InfoGather+ SIGMOD 2013]
◦ Purchase entire datasets [Data Pricing PVLDB 2011]

Deeper

◦ Leverage Deep Web (Hidden DB) for Data Enrichment
Why is it hard?

End-to-End system
- Schema Matching
- Deep Web Crawling
- Entity Resolution
- ...

Deep Web Crawling
- Prior Work [Zhang and Das, PVLDB 2011 tutorial]
- New challenge

Local DB
Hidden DB
Deeper System (v0.1)

https://deeper.sfucloud.ca
### Demo

**Video:** [https://youtu.be/QHYgllqqjWY](https://youtu.be/QHYgllqqjWY)

**Demo:** [https://deeper.sfucloud.ca](https://deeper.sfucloud.ca)

---

#### Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>QueryMarket Demonstration: Pricing for Online Data Markets</td>
<td>Paraschos Koutris and Prasang Upadhyaya and Magdalena Balazinska and Bill Howe and Dan Suciu</td>
</tr>
<tr>
<td>1</td>
<td>Elastic Memory Management for Cloud Data Analytics</td>
<td>Jingjing Wang and Magdalena Balazinska</td>
</tr>
<tr>
<td>2</td>
<td>Profiling a GPU database implementation: a holistic view of GPU resource utilization on TPC-H queries</td>
<td>Emily Furst and Mark Oskin and Bill Howe</td>
</tr>
<tr>
<td>3</td>
<td>Sloth: Being Lazy Is a Virtue (When Issuing Database Queries)</td>
<td>Alvin Cheung and Samuel Madden and Armando Solar-Lezama</td>
</tr>
<tr>
<td>4</td>
<td>Query-Based Data Pricing</td>
<td>Paraschos Koutris and Prasang Upadhyaya and Magdalena Balazinska and Bill Howe and Dan Suciu</td>
</tr>
<tr>
<td>5</td>
<td>Managing Skew in Hadoop</td>
<td>YongChul Kwon and Kai Ren and Magdalena Balazinska and Bill Howe</td>
</tr>
</tbody>
</table>

---

**SFU Big Data**

Phone: 778.782.6930
Fax: 778.782.4969
8888 University Drive
Burnaby, B.C. Canada V5A 1S6
Mon - Thu: 8:00 - 18:00
Friday: 8:00 - 16:30
Why is it hard?

End-to-End system
- Schema Matching
- Deep Web Crawling
- Entity Resolution
- ...

Deep Web Crawling
- Prior Work [Zhang and Das, PVLDB 2011 tutorial]
- New challenge
NaïveCrawl

Match one record at a time
OpenRefine is doing this!

Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started? Watch these screencasts
**SmartCrawl**

**Idea 1:** Generate a query pool with both specific and general queries (e.g., “transaction”, “query optimization”)
- Apply frequent pattern mining to local DB

**Idea 2:** Select the query with the highest estimated benefit
- Model it as a query-selectivity estimation problem

```sql
SELECT d, h FROM D, H
WHERE d satisfies q AND h satisfies q
AND match(d, h) = True
```
NaïveCrawl vs SmartCrawl

Local DB = 10,000, DBLP Hidden DB

4.5X more enriched records
Today’s Talk

Deeper  [SIGMOD 2018 (Demo), SIGMOD 2019]
◦ Speed up data enrichment

TARS  [VLDB 2019]
◦ Speed up data labeling

AQP++  [SIGMOD 2018]
◦ Speed up data analysis
Democratizing AI

Computing

- Amazon Web Services
- Azure
- Google Cloud Platform

Algorithms

- mxnet
- TensorFlow
- PyTorch

Training Data

The Bottleneck
A Promising Solution

Label Noise vs. Human Cost Trade-off

- Random
- Distance Supervision
- Crowdsourcing
- Domain Expert
Crowdsourced (Noisy) Labels

\[ X \] \[ Y \]
\hline
\( x_1 \) & ? \\
\( x_2 \) & ? \\
\( x_3 \) & ? \\
\( x_4 \) & ? \\
\( x_5 \) & ? \\

\[ X \] \[ W_1 \] \[ W_2 \] \[ W_3 \]
\hline
\( x_1 \) & +1 & +1 & +1 \\
\( x_2 \) & +1 & -1 & -1 \\
\( x_3 \) & -1 & -1 & -1 \\
\( x_4 \) & -1 & -1 & +1 \\
\( x_5 \) & +1 & +1 & +1 \\

Truth Inference

X | Y | P
\hline
\( x_1 \) & +1 & 0.9 \\
\( x_2 \) & -1 & 0.7 \\
\( x_3 \) & -1 & 0.8 \\
\( x_4 \) & -1 & 0.6 \\
\( x_5 \) & +1 & 0.9
Cleaning Noisy Label

Existing Work*
- No Cleaning
- Machine-based Cleaning

Our Solution
- Oracle-based Cleaning

TARS [named after an intelligent robot in the movie *Interstellar*]

**Label Cleaning Advisor for Crowdsourced Noisy Labels**

Mohamad Dolatshah, Mathew Teoh, Jiannan Wang, Jian Pei. Cleaning Crowdsourced Labels Using Oracles For Statistical Classification. *VLDB 2019*
Two Pieces of Advice

**Advice 1. Model Evaluation**

- How accurate is a model?

(1) Model
(2) Noisy Test Data

0.8±0.01

**Advice 2. Cleaning Strategy**

- Which label should be cleaned?

(1) Learning Algorithm
(2) Noisy Training Data

<instance_3, label_3>
Advice 2. Cleaning Strategy

Main Idea
○ Factor 1. Label Noise
○ Factor 2. Model Impact

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>+1</td>
<td>0.9</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-1</td>
<td>0.7</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-1</td>
<td>0.8</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-1</td>
<td>0.6</td>
</tr>
<tr>
<td>$x_5$</td>
<td>+1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Expected Model Improvement

$$(1 - P_i) \times [\text{Accuracy}(\text{model}_{\text{new}_i}) - \text{Accuracy}(\text{model}_{\text{current}})]$$

Practical Issues
○ Be careful to choose train/test data
○ Need to break the tie carefully
Experimental Result

Heart Dataset (noise rate: 0.6 ~ 1.0)
Today’s Talk

Deeper [SIGMOD 2018 (Demo), SIGMOD 2019]
- Speed up data enrichment

TARS [VLDB 2019 (under revision)]
- Speed up data labeling

AQP++ [SIGMOD 2018]
- Speed up data analysis
Interactive Analytics

How to enable interactive analytics over Big Data?
Two Separate Ideas

Idea 1. Approximate Query Processing (AQP)

SELECT SUM(salary) WHERE id in [6, 10000]
Two Separate Ideas

Idea 2. Aggregation Precomputation (AggPre)

SELECT SUM(salary) WHERE id in [6, 10000]

<table>
<thead>
<tr>
<th>ID</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50,000</td>
</tr>
<tr>
<td>2</td>
<td>62,492</td>
</tr>
<tr>
<td>3</td>
<td>78,212</td>
</tr>
<tr>
<td>4</td>
<td>120,242</td>
</tr>
<tr>
<td>5</td>
<td>98,341</td>
</tr>
<tr>
<td>6</td>
<td>75,453</td>
</tr>
<tr>
<td>7</td>
<td>60,000</td>
</tr>
<tr>
<td>8</td>
<td>72,492</td>
</tr>
<tr>
<td>9</td>
<td>88,212</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10000</td>
<td>86,798</td>
</tr>
</tbody>
</table>

Prefix-Sum Cube[1]

<table>
<thead>
<tr>
<th>ID</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤1</td>
<td>50,000</td>
</tr>
<tr>
<td>≤2</td>
<td>112,492</td>
</tr>
<tr>
<td>≤3</td>
<td>190,704</td>
</tr>
<tr>
<td>≤4</td>
<td>310,946</td>
</tr>
<tr>
<td>≤5</td>
<td>409,287</td>
</tr>
<tr>
<td>≤6</td>
<td>484,740</td>
</tr>
<tr>
<td>≤7</td>
<td>544,740</td>
</tr>
<tr>
<td>≤8</td>
<td>617,232</td>
</tr>
<tr>
<td>≤9</td>
<td>705,444</td>
</tr>
<tr>
<td>≤10000</td>
<td>9.3*10^8</td>
</tr>
</tbody>
</table>

Trade-Off

Response Time

Preprocessing Cost

Query Error

AQP

AQP++

AggPre
Connecting **Approximate Query Processing** With **Aggregate Precomputation**

Jinglin Peng, Dongxiang Zhang, Jiannan Wang, Jian Pei. AQP++: Connecting Approximate Query Processing with Aggregate Precomputation for Interactive Analytics. *SIGMOD 2018*
How AQP++ works?

### SELECT SUM(salary) WHERE id in [6, 10000]

### SELECT SUM(salary) WHERE id in [0, 10000]

### SELECT SUM(salary) WHERE id in [0, 5]

**Blocked Prefix-Sum Cube**

<table>
<thead>
<tr>
<th>ID</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤1000</td>
<td>1.2 * 10^8</td>
</tr>
<tr>
<td>≤2000</td>
<td>1.8 * 10^8</td>
</tr>
<tr>
<td>≤3000</td>
<td>2.9 * 10^8</td>
</tr>
<tr>
<td>≤4000</td>
<td>3.1 * 10^8</td>
</tr>
<tr>
<td>≤5000</td>
<td>4.0 * 10^8</td>
</tr>
<tr>
<td>≤6000</td>
<td>4.8 * 10^8</td>
</tr>
<tr>
<td>≤7000</td>
<td>5.4 * 10^8</td>
</tr>
<tr>
<td>≤8000</td>
<td>6.1 * 10^8</td>
</tr>
<tr>
<td>≤9000</td>
<td>8.1 * 10^8</td>
</tr>
<tr>
<td>≤10000</td>
<td>9.3 * 10^8</td>
</tr>
</tbody>
</table>

1GB sample
## Experimental Result

**TPCD (Laptop, 100GB)**
- 0.05% sample, skew = 2

<table>
<thead>
<tr>
<th></th>
<th>Preprocessing Cost</th>
<th>Response Time</th>
<th>Answer Quality (Avg Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Space</td>
<td>Time</td>
<td></td>
</tr>
<tr>
<td><strong>AQP</strong></td>
<td>51.2 MB</td>
<td>4.3 min</td>
<td>0.6 sec</td>
</tr>
<tr>
<td><strong>AggPre</strong></td>
<td>&gt; 10 TB</td>
<td>&gt; 1 day</td>
<td>&lt; 0.01 sec</td>
</tr>
<tr>
<td><strong>AQP++</strong></td>
<td>51.9 MB</td>
<td>9.8 min</td>
<td>0.64 sec</td>
</tr>
</tbody>
</table>
Take-away Messages

Our Mission
- Speeding Up Data Science

Deeper
- Leverage Deep Web to speed up data enrichment

TARS
- Build a Label Cleaning Advisor to speed up data labeling

AQP++
- Connect AQP with AggPre to speed up data analysis

https://github.com/sfu-db

Thanks!