Speeding Up Data Science:
From a Data Management Perspective

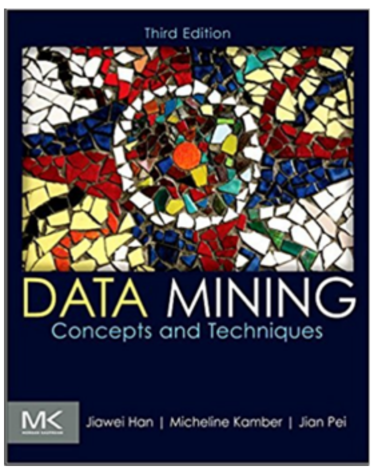
Jiannan Wang

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	7978	2000
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	2662	2001
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	12578	1996

SFU DB/DM Group

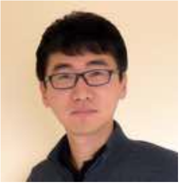
- **Research Areas:** Machine Learning, Data Science, and Big Data Systems
- **Research Strengths:** Cloud Databases, Crowdsourced Data Management, Data Cleaning and Integration, Data Security and Privacy, Fraud Detection, Interpretable Machine Learning, Precision Medicine, Recommender Systems
- **Ranked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

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

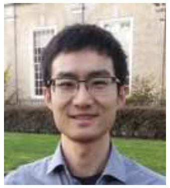
Ke Wang (Joined in 2000)

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

What	When	Who	Goal
Computer Science	1950-	Software Engineer	Write software to make computers work

Plan → Design → Develop → Test → Deploy → Maintain

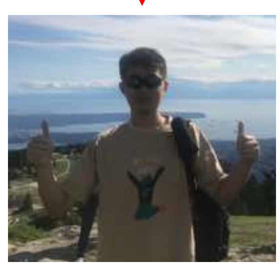
What	When	Who	Goal
Data Science	2010-	Data Scientist	Extract insights from data to answer questions

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

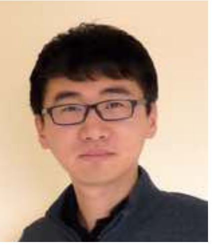
Leverage **Deep Web** To

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

Restaurant	Address	City	Category
Boiling Point	4148 Main St	Vancouver	?
Flamingo Chinese Restaurant	1652 SE Marine Drive	Vancouver	?
Sun Sui Wah Restaurant	3888 Main Street	Vancouver	?

Various Use Cases

- For market analysis
- For ML model training
- For knowledge base construction

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

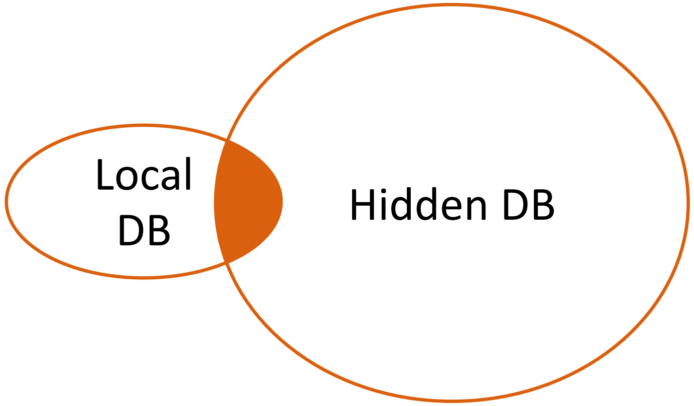
Why is it hard?

End-to-End system

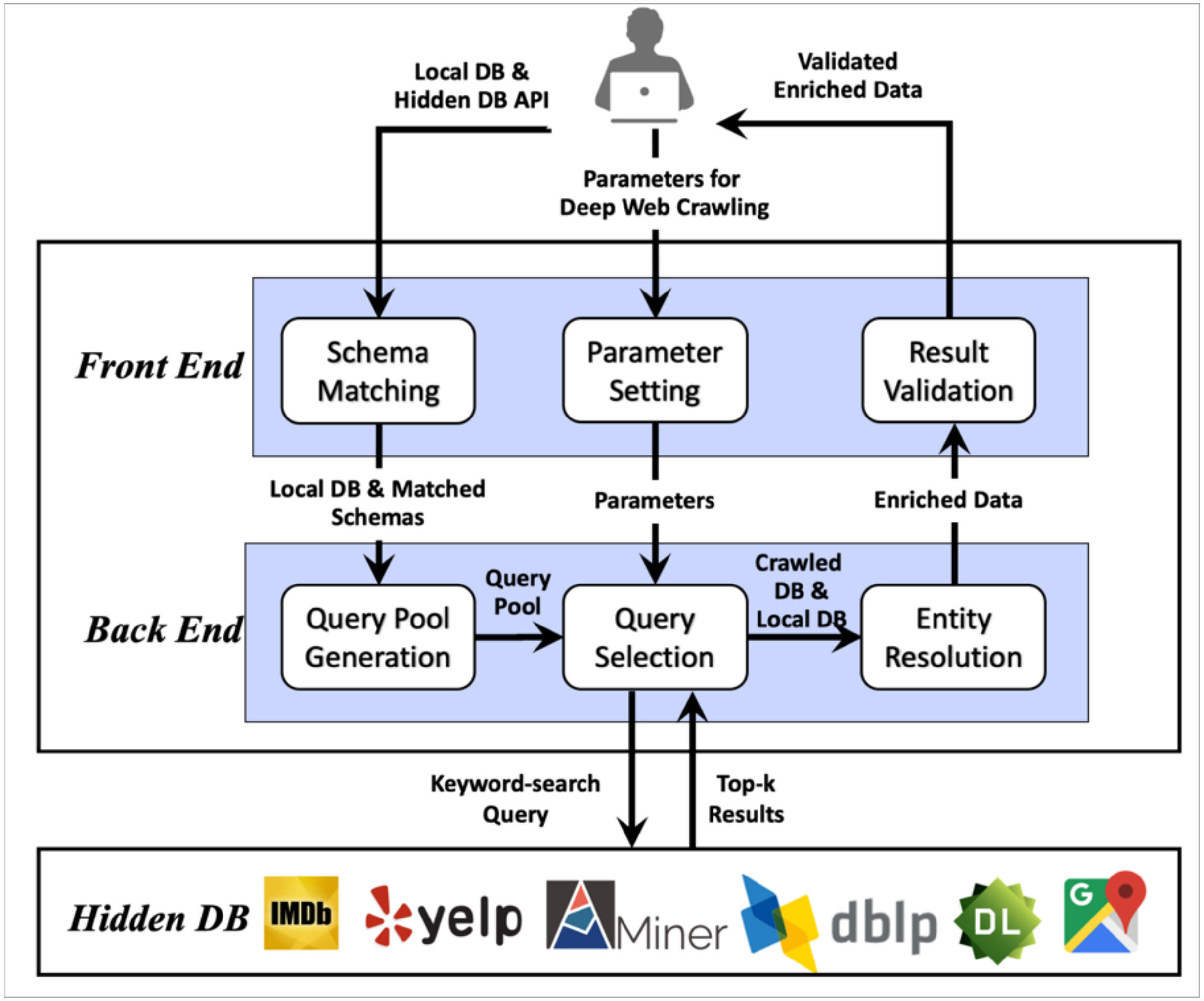
- Schema Matching
- Deep Web Crawling
- Entity Resolution
- …

Deep Web Crawling

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

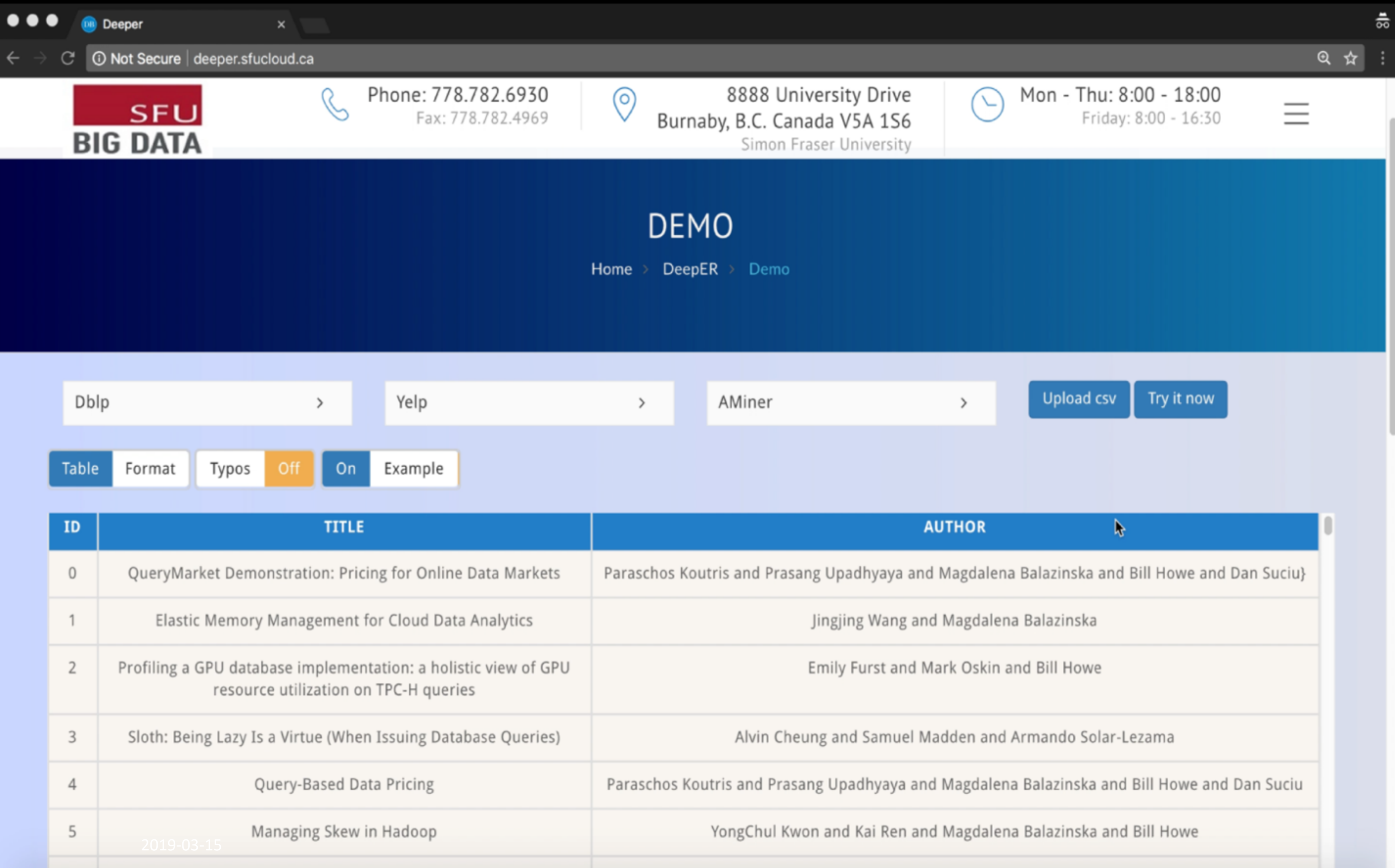


Deeper System (v0.1)



https://deeper.sfucloud.ca

Demo: https://deeper.sfucloud.ca
Video: https://youtu.be/QHYgLIqqjWY



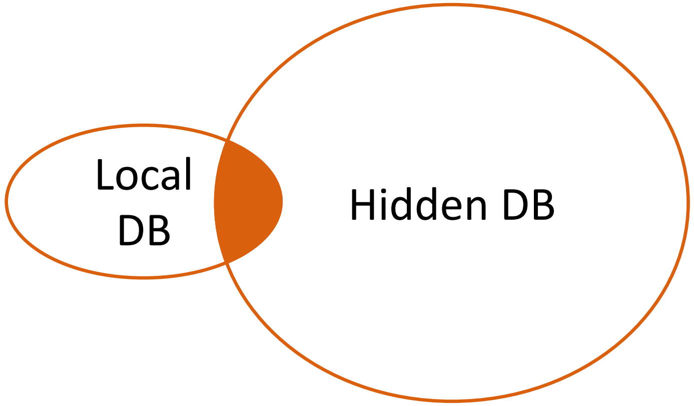
Why is it hard?

End-to-End system

- Schema Matching
- Deep Web Crawling
- Entity Resolution
- …

Deep Web Crawling

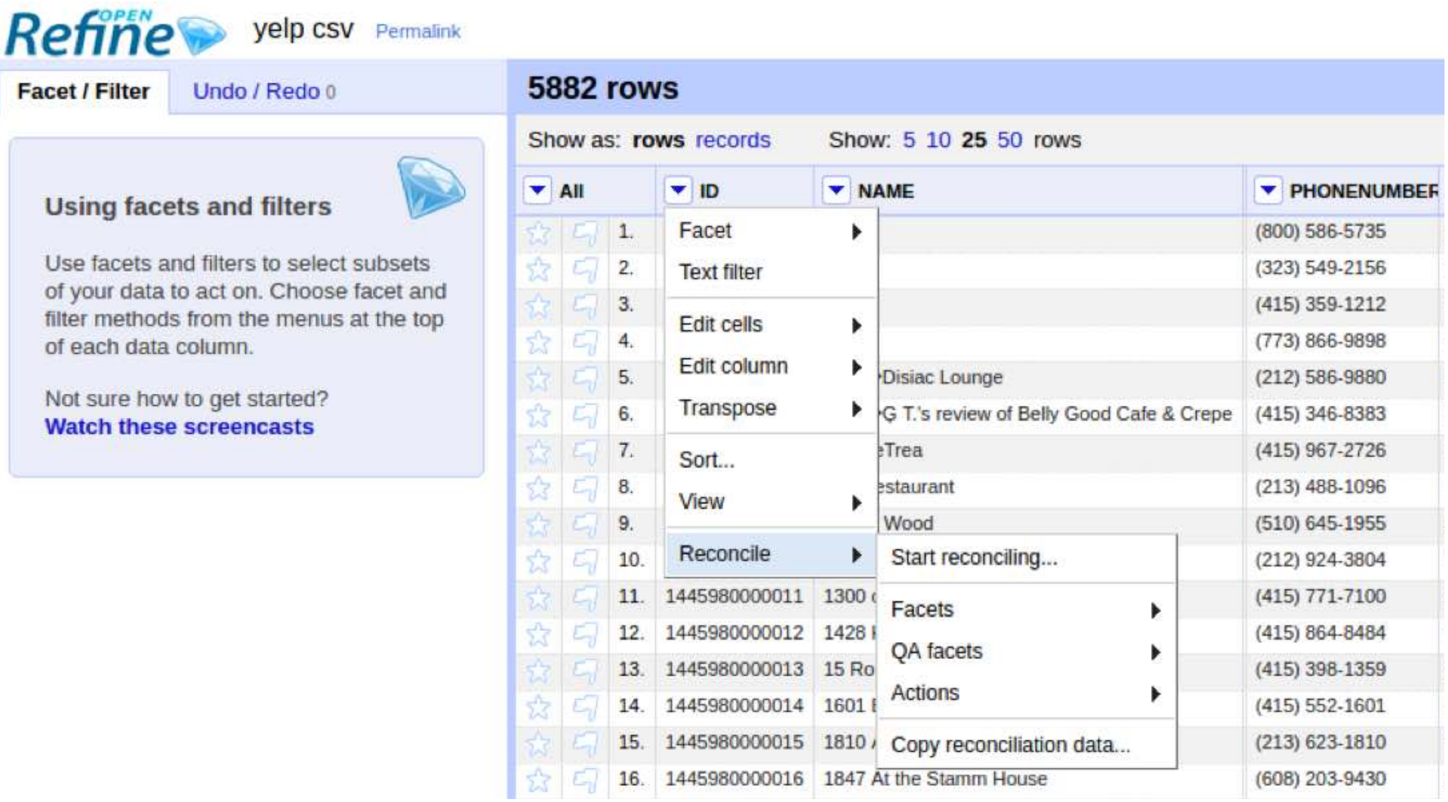
- Prior Work [Zhang and Das, PVLDB 2011 tutorial]
- New challenge



NaïveCrawl

Match one record at a time

OpenRefine is doing this!



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

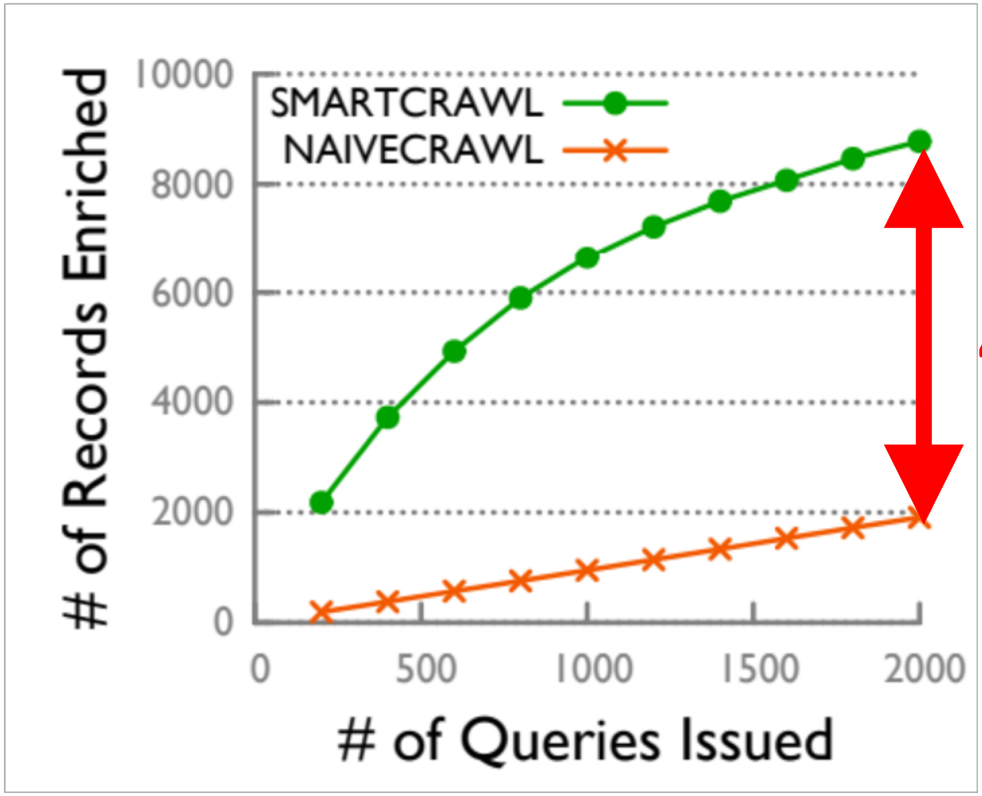
Idea 2: Select the query with the highest estimated benefit

- Model it as a query-selectivity estimation problem

```
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

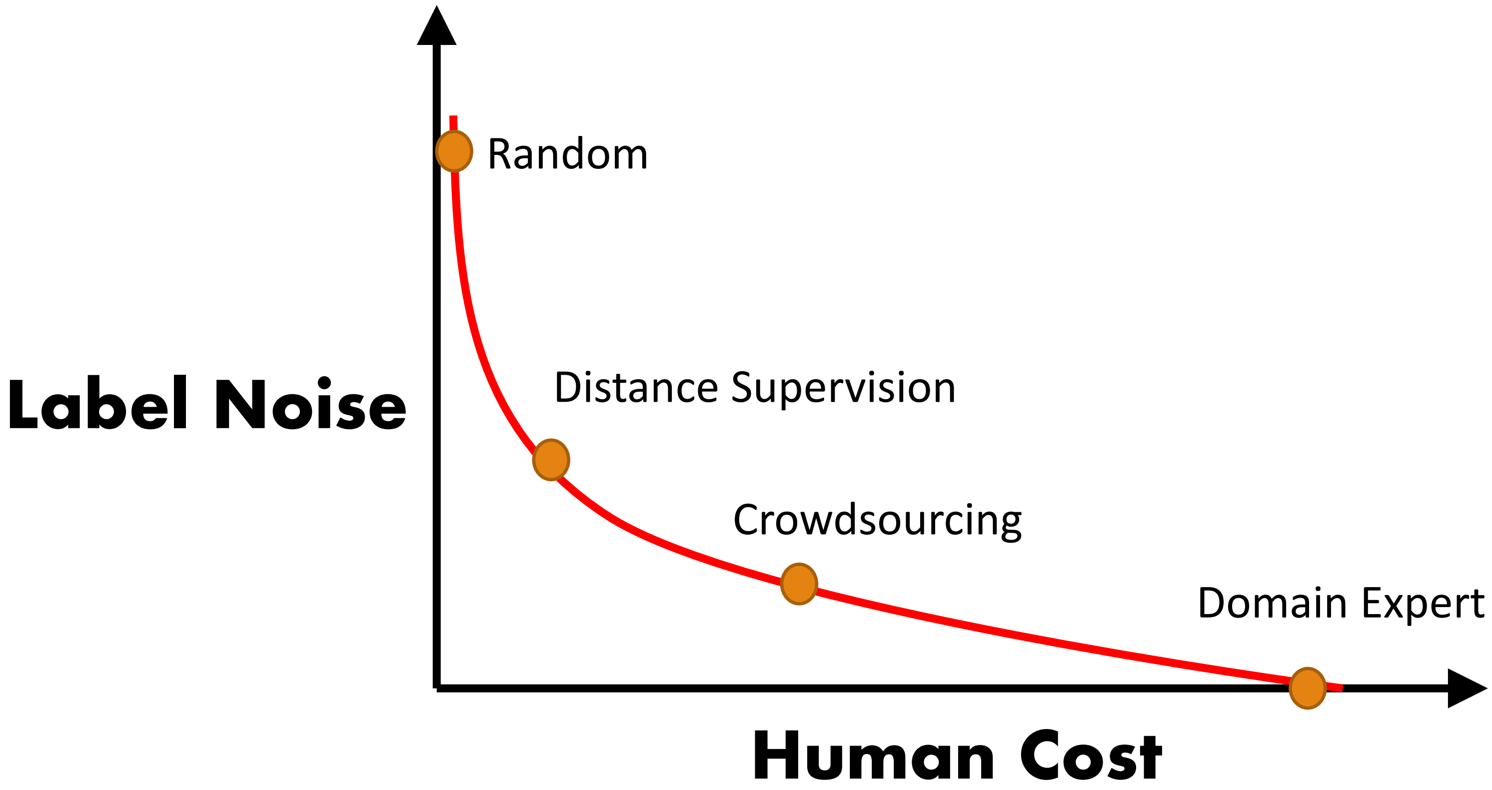
Computing

Algorithms

Training Data

The Bottleneck

A Promising Solution



Label Noise vs. Human Cost Trade-off

Crowdsourced (Noisy) Labels

X	Y
x_1	?
x_2	?
x_3	?
x_4	?
x_5	?

amazon mechanical turk beta

X	W_1	W_2	W_3
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
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

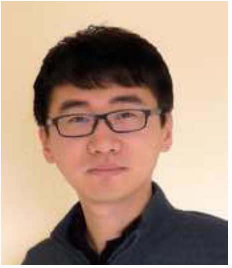
* Benoît Frénay, Michel Verleysen: Classification in the Presence of Label Noise: A Survey. IEEE Trans. Neural Netw. Learning Syst. 25(5): 845-869 (2014)

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

Label Cleaning **Advisor** for

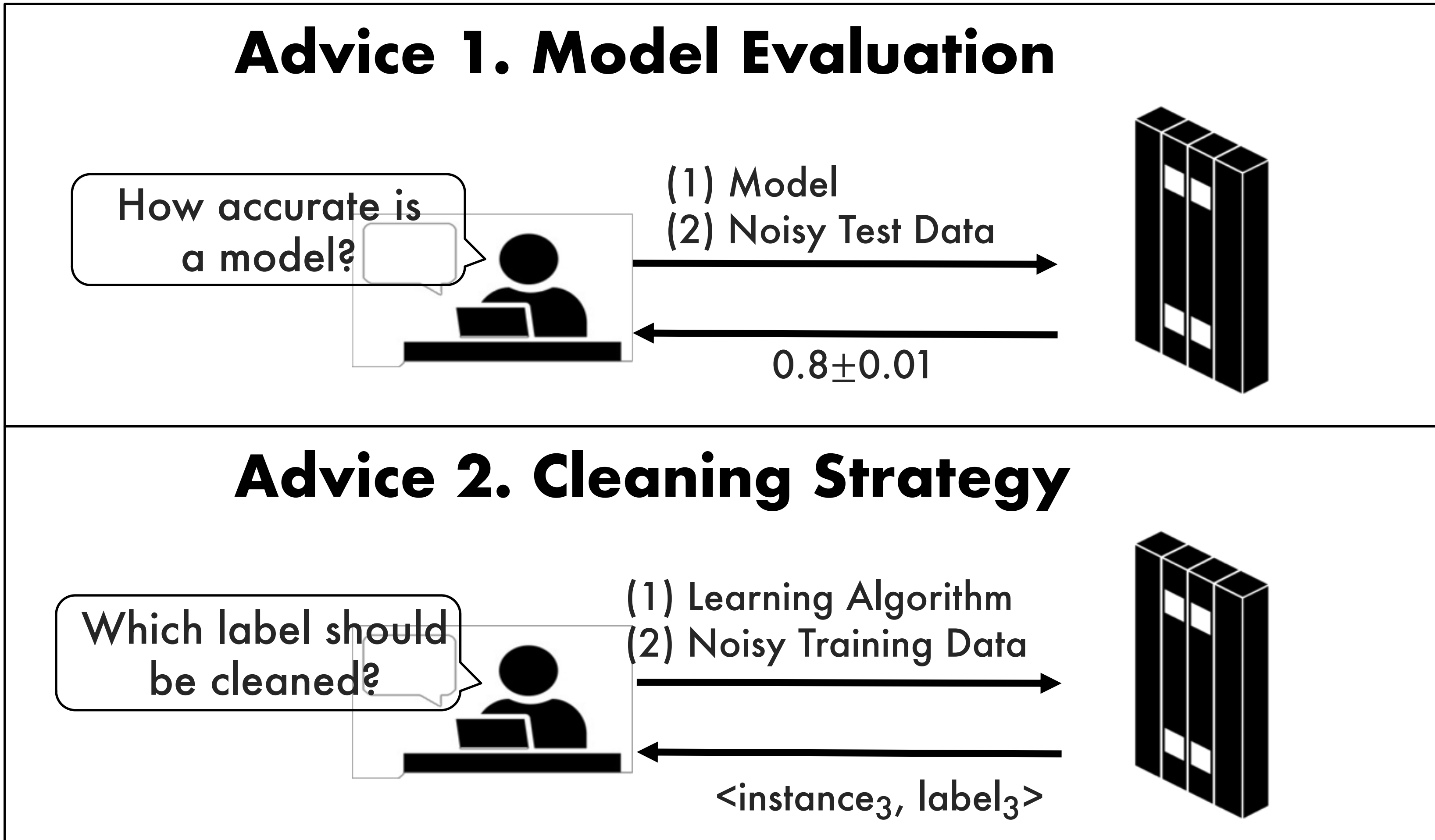
Crowdsourced Noisy Labels






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

Two Pieces of Advice



Advice 2. Cleaning Strategy

Main Idea

- Factor 1. Label Noise
- Factor 2. Model Impact

X	Y	P
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

Expected Model Improvement

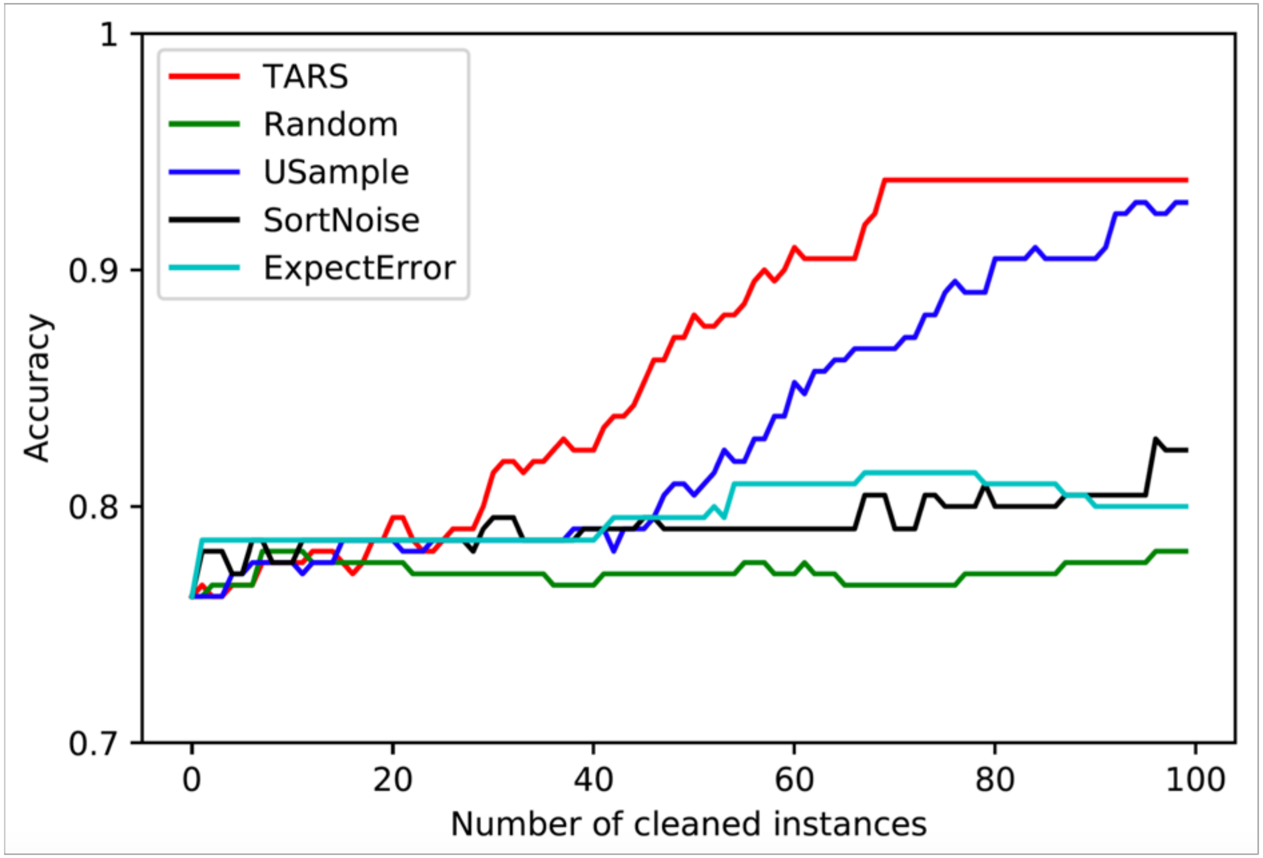
$$(1 - P_{\mathrm{i}}) \times [\mathbf{Accuracy}(\mathrm{model}_{\mathrm{new}_i}) - \mathbf{Accuracy}(\mathrm{model}_{\mathrm{current}})]$$

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

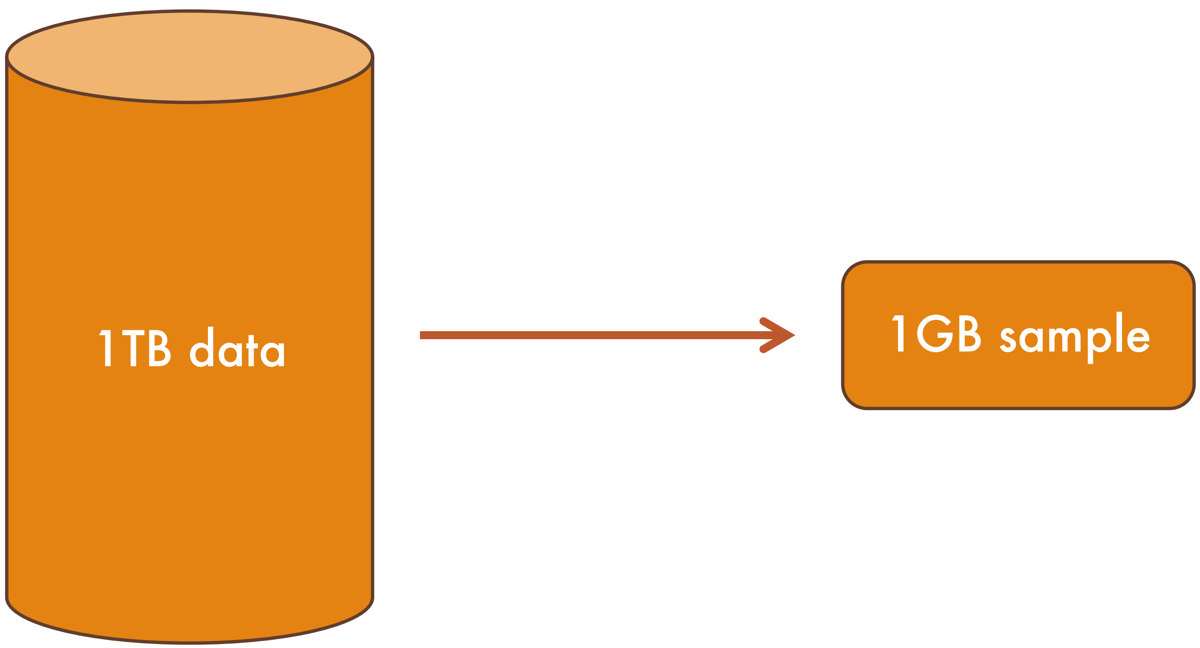
Interactive Analytics

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]

Base Table

ID	Salary
1	50,000
2	62,492
3	78,212
4	120,242
5	98,341
6	75,453
7	60,000
8	72,492
9	88,212
…	
10000	86,798

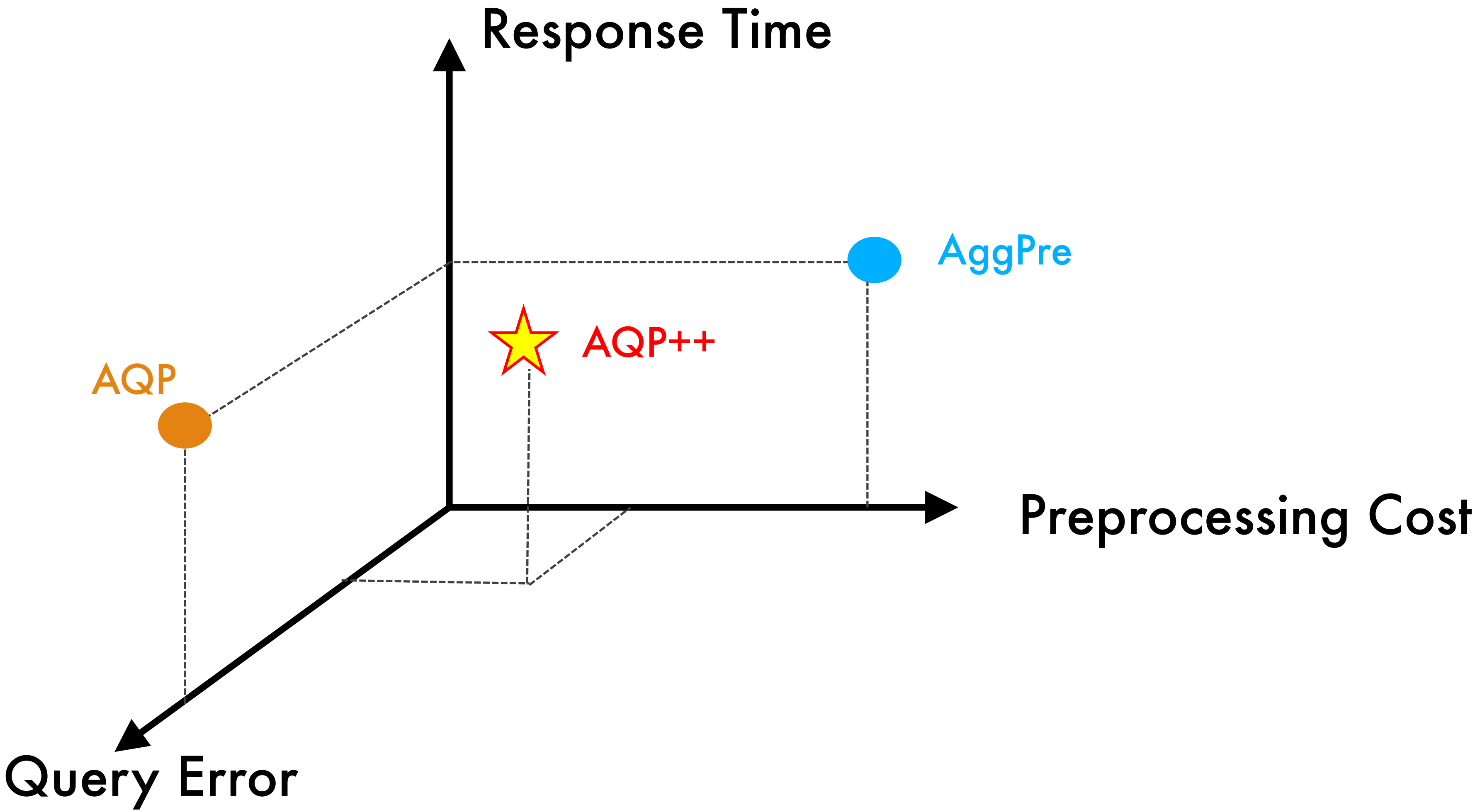
→

Prefix-Sum Cube[1]

ID	Salary
≤1	50,000
≤2	112,492
≤3	190,704
≤4	310,946
≤5	409,287
≤6	484,740
≤7	544,740
≤8	617,232
≤9	705,444
…	
≤10000	9.3*10^8

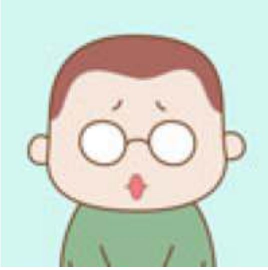
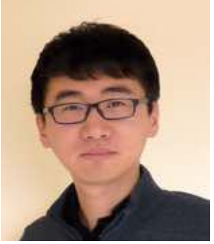
[1] Ho, Ching-Tien, et al. Range queries in OLAP data cubes. (1997)

Trade-Off


Response Time
AggPre
AQP++
AQP
Preprocessing Cost
Query Error

AQP++

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]

1GB sample

Blocked Prefix-Sum Cube

ID	Salary
≤1000	1.2 * 10^8
≤2000	1.8 * 10^8
≤3000	2.9 * 10^8
≤4000	3.1 * 10^8
≤5000	4.0 * 10^8
≤6000	4.8 * 10^8
≤7000	5.4 * 10^8
≤8000	6.1 * 10^8
≤9000	8.1 * 10^8
≤10000	9.3 * 10^8

Experimental Result

TPCD (Laptop,100GB)

- 0.05% sample, skew = 2

	Preprocessing Cost		Response Time	Answer Quality (Avg Err.)
	Space	Time		
AQP	51.2 MB	4.3 min	0.6 sec	2.67%
AggPre	> 10 TB	> 1 day	< 0.01 sec	0.00%
AQP++	51.9 MB	9.8 min	0.64 sec	0.28%

Take-away Messages

Our Mission

- Speeding Up Data Science

https://github.com/sfu-db

Thanks!

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