Speeding Up Data Science:

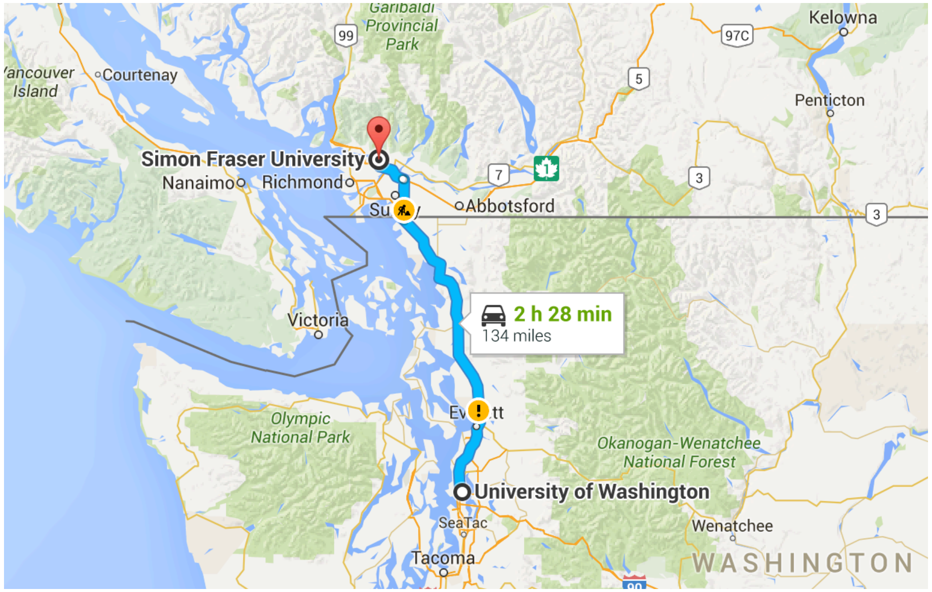
From a Data Management Perspective

Jiannan Wang

Database System Lab (DSL)
Simon Fraser University

NWDS Meeting, Jan 5, 2018

Simon Fraser University



SFU DB/DM Group

Ke Wang
(Joined SFU in 2000)

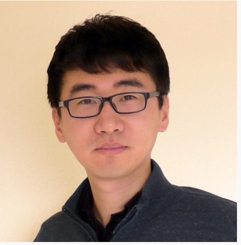
- Privacy-Preserving Data Publishing
- Secure Query Answering for Outsourced Databases

Martin Ester
(Joined SFU in 2001)

- Recommendation in Social Media
- Biological Data Mining

Jian Pei
(Joined SFU in 2004)

- Interpretable Machine Learning and Deep Learning
- Computational Fraud Investigation
- Robust AI models Against Adversarial Attacks



Jiannan Wang
(Joined SFU in 2016)

- Data Cleaning for Machine Learning
- Data Enrichment with Deep Web
- Interactive Analytics Over Big Data

My Lab's Mission

Speeding Up Data Science

Computer Science vs. Data Science

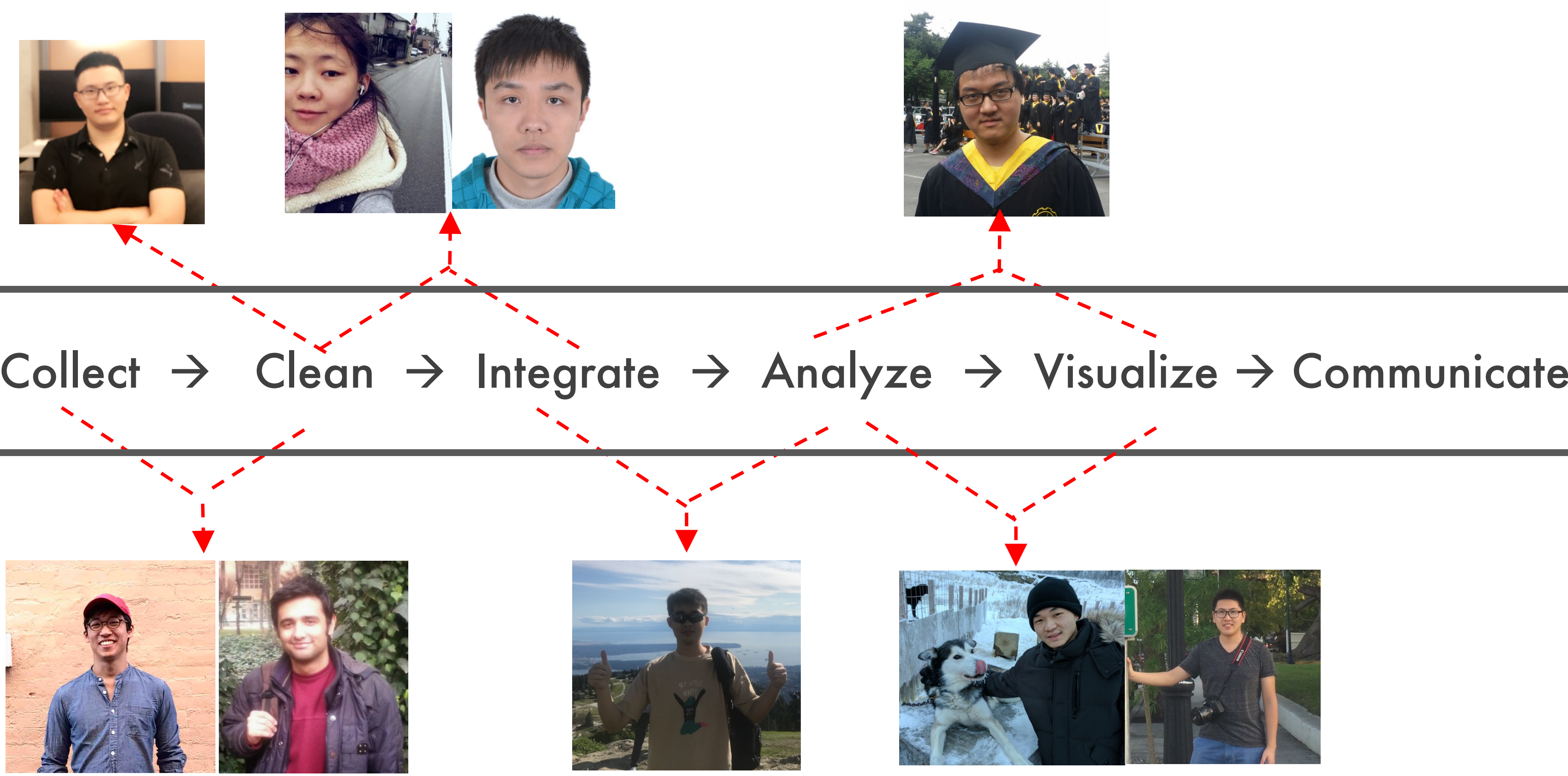
What	When	Who	Goal
Computer Science	1950-	Software Engineer	Write software to make computers work

Plan → Design → Develop → Test → Deploy → Maintain

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

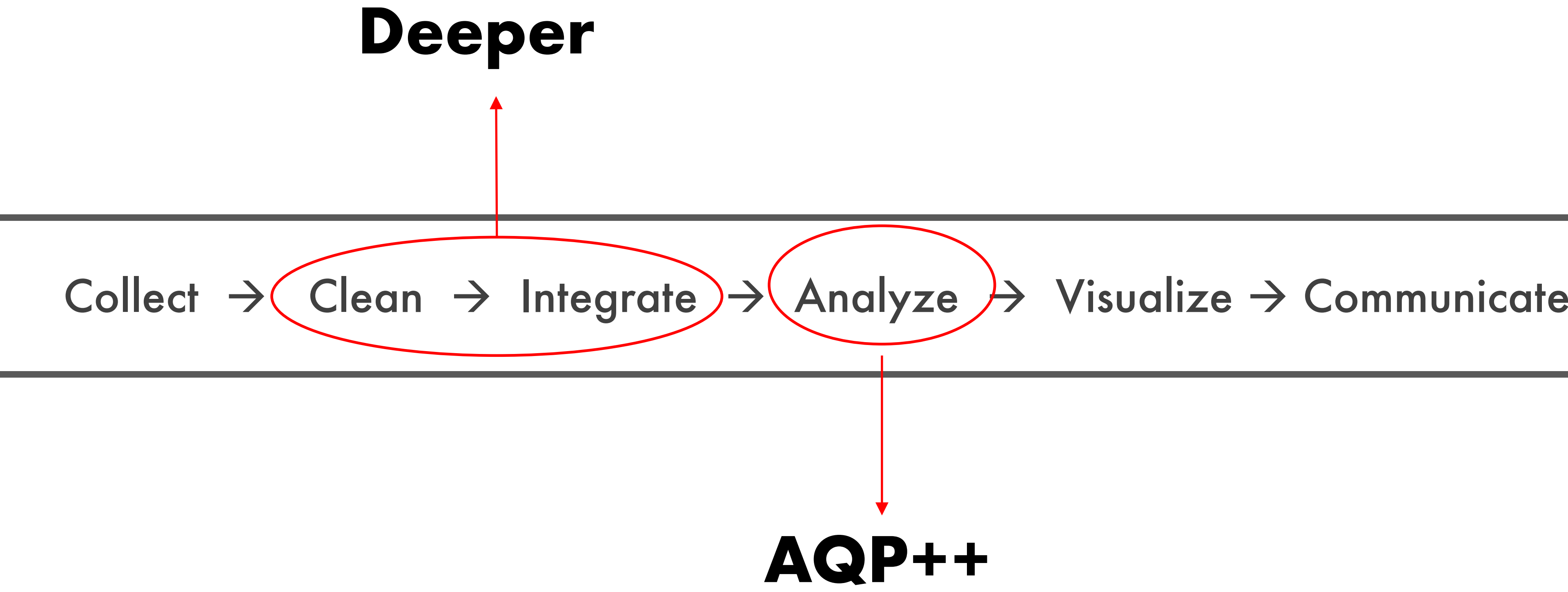
Collect → Clean → Integrate → Analyze → Visualize → Communicate

Lab Members



Collect → Clean → Integrate → Analyze → Visualize → Communicate

Today's Talk



Deeper

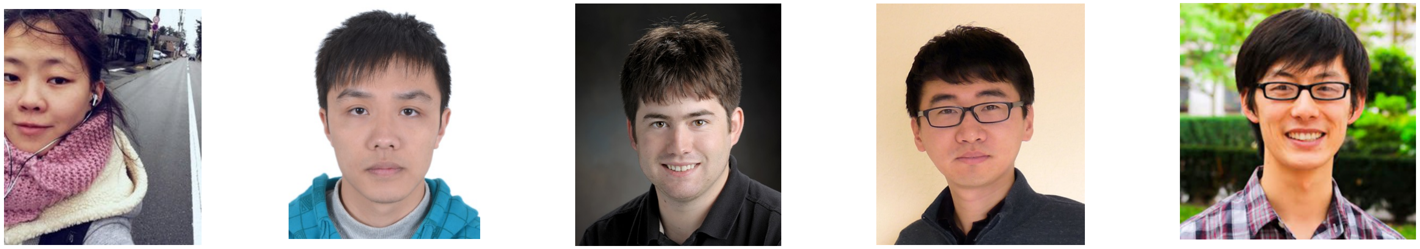
Collect → Clean → Integrate → Analyze → Visualize → Communicate

AQP++

Deeper (2016 -)

Leverage **Deep Web** To

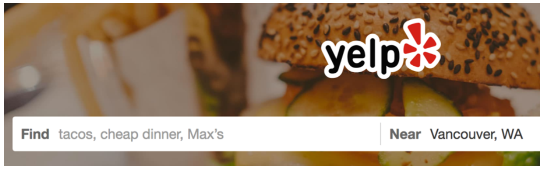
Speed Up **Data Enrichment & Cleaning**

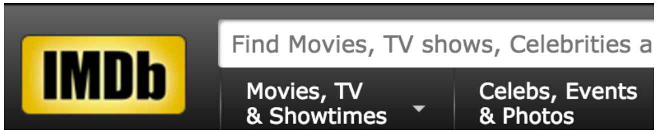


Pei Wang, Yongjun He, Ryan Shea, Jiannan Wang, Eugene Wu. Deeper: A Data Enrichment System Powered by Deep Web.
SIGMOD 2018 Demo (in submission)

Deep Web

Hidden Database





Invaluable External Resource

- Big: Consisting of a substantial number of entities
- Rich: Having rich Information about each entity
- High-quality. Being trustful and up-to-date

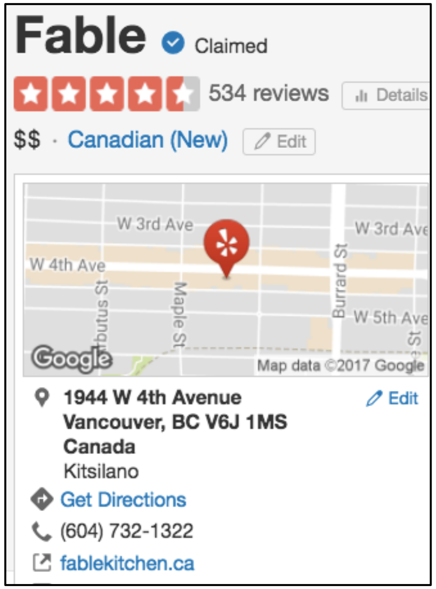
Data Enrichment & Cleaning

Name	City	Zip Code	Tel
Fable	Burnaby	V6J 1MS	(604)732-1322

How ?

Name	City	Zip Code	Tel	Category	Rating
Fable	Vancouver	V6J 1MS	(604)732-1322	Canadian (New)	4.5

Leverage Deep Web



Fable ✔ Claimed

★★★★★ 534 reviews Details

$$ · Canadian (New) Edit

1944 W 4th Avenue
Vancouver, BC V6J 1MS
Canada
Kitsilano

Edit

Get Directions

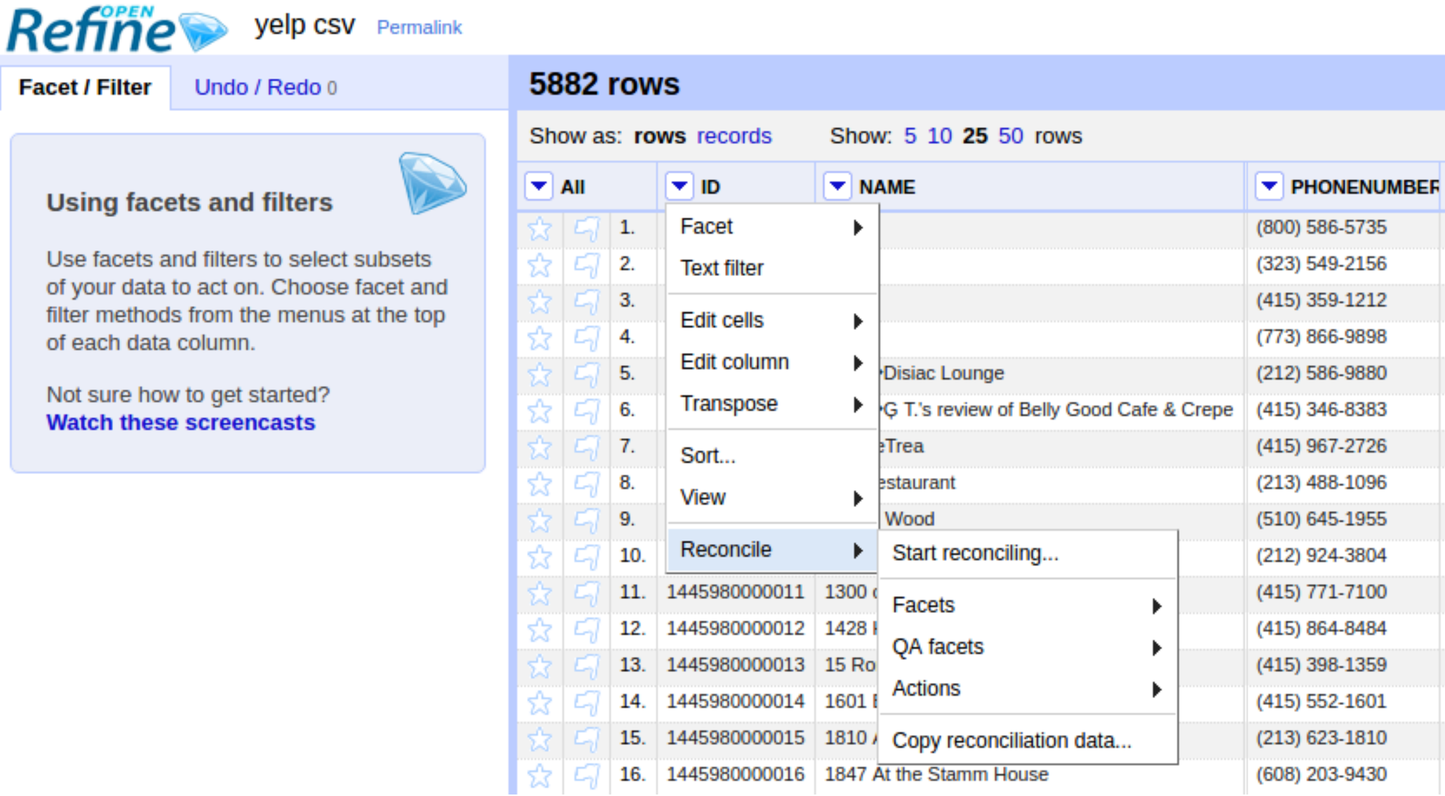
☎ (604) 732-1322

fablekitchen.ca

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

5882 rows

Show as: **rows** records Show: 5 10 **25** 50 rows

All	ID	NAME	PHONENUMBER
1.			(800) 586-5735
2.			(323) 549-2156
3.			(415) 359-1212
4.			(773) 866-9898
5.		·Disiac Lounge	(212) 586-9880
6.		·G T.'s review of Belly Good Cafe & Crepe	(415) 346-8383
7.		·Trea	(415) 967-2726
8.		·staurant	(213) 488-1096
9.		Wood	(510) 645-1955
10.			(212) 924-3804
11.	1445980000011	1300 (	(415) 771-7100
12.	1445980000012	1428 I	(415) 864-8484
13.	1445980000013	15 Ro	(415) 398-1359
14.	1445980000014	1601 E	(415) 552-1601
15.	1445980000015	1810 /	(213) 623-1810
16.	1445980000016	1847 At the Stamm House	(608) 203-9430

Limitations

Limited Query Budget

- Goolge Maps API allows 2,500 free requests per day

Dirty Data

- User's data is usually messy. Naïve queries will miss results

SmartCrawl

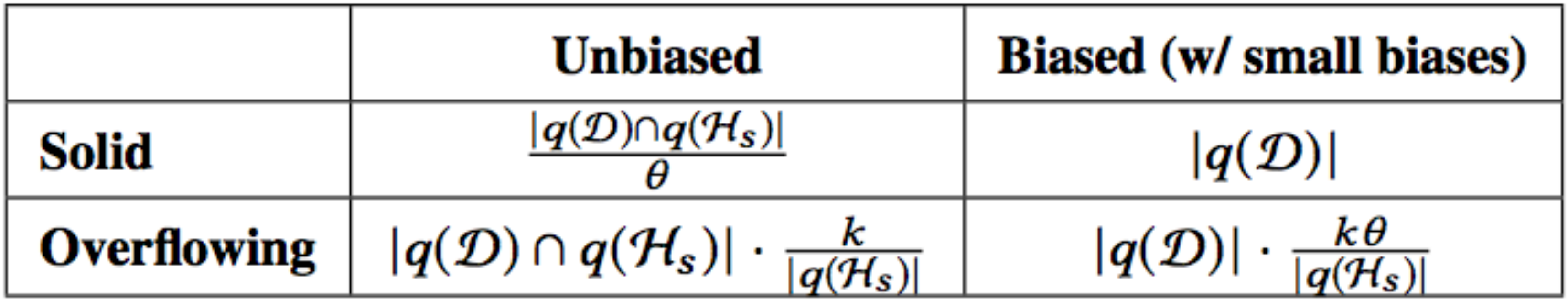
1. Generate a query pool Q

2. Select at most b queries from Q such that $|H_{crawled} \cap D|$ is maximized

3. Perform entity resolution between $H_{crawled}$ and D

Challenges

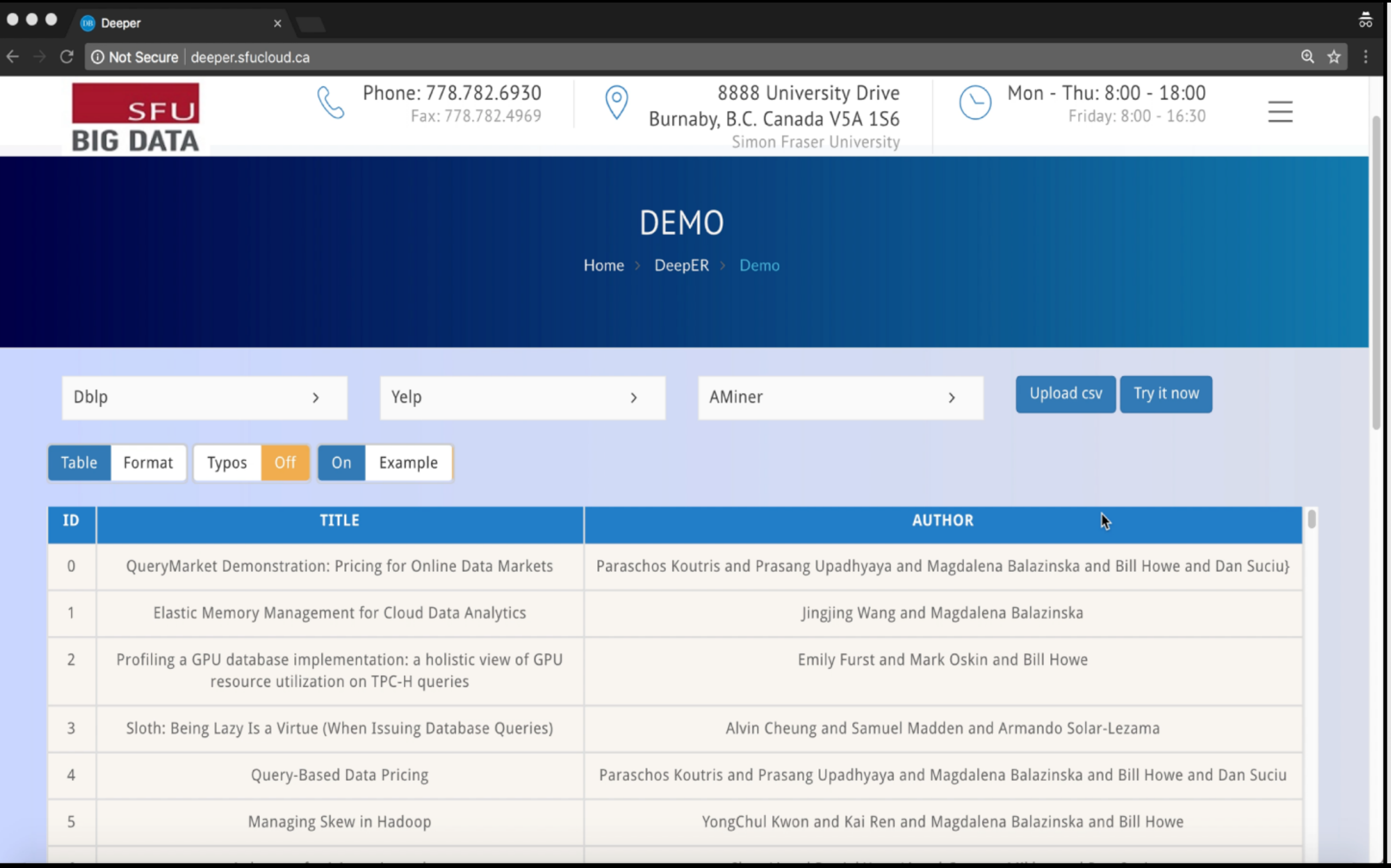
1. Query Benefit Estimation



	Unbiased	Biased (w/ small biases)
Solid	$\frac{\lvert q(\mathcal{D})\cap q(\mathcal{H}_s)\rvert}{\theta}$	$\lvert q(\mathcal{D})\rvert$
Overflowing	$\lvert q(\mathcal{D}) \cap q(\mathcal{H}_s)\rvert \cdot \frac{k}{\lvert q(\mathcal{H}_s)\rvert}$	$\lvert q(\mathcal{D})\rvert \cdot \frac{k\theta}{\lvert q(\mathcal{H}_s)\rvert}$

2. Efficient Implementations
3. Inadequate Sample Size
4. Fuzzy Matching

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



Phone: 778.782.6930
Fax: 778.782.4969

8888 University Drive
Burnaby, B.C. Canada V5A 1S6
Simon Fraser University

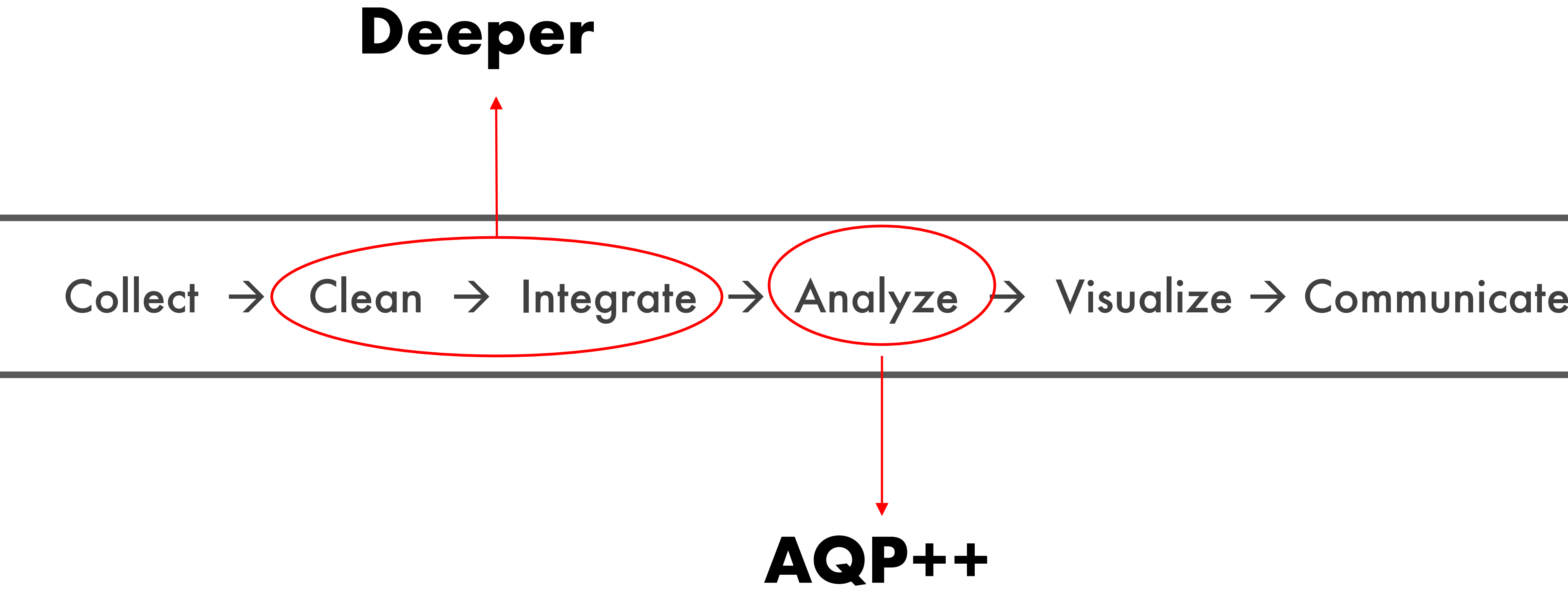
Mon - Thu: 8:00 - 18:00
Friday: 8:00 - 16:30

DEMO

Home › DeepER › Demo

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

Today's Talk



Deeper

Collect → Clean → Integrate → Analyze → Visualize → Communicate

AQP++

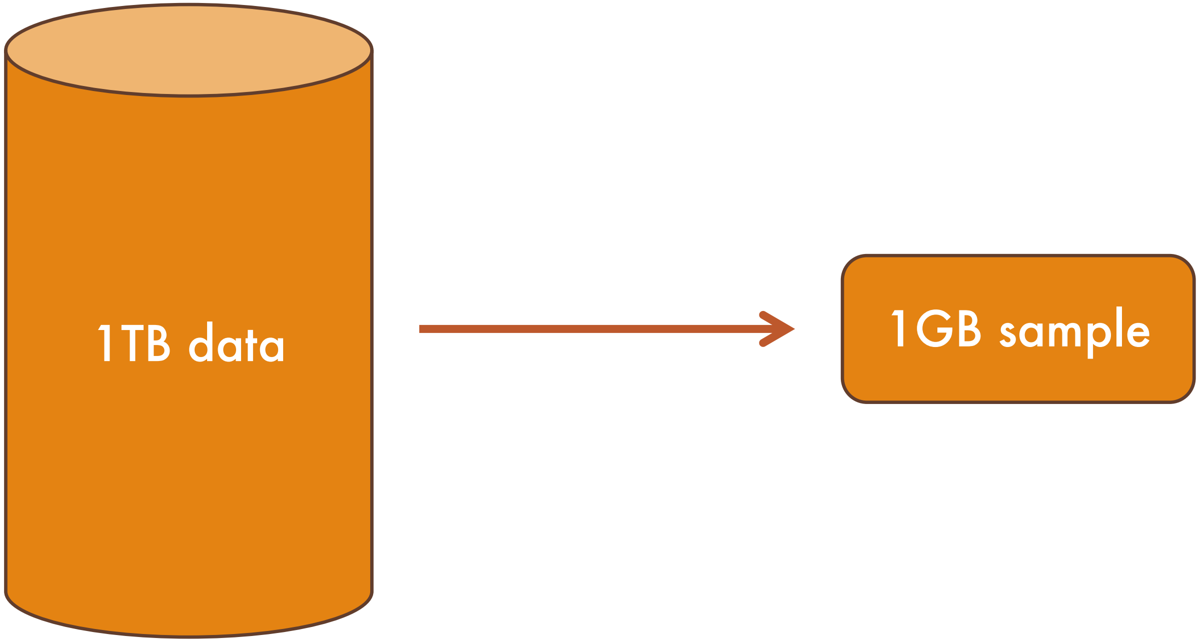
Interactive Analytics

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

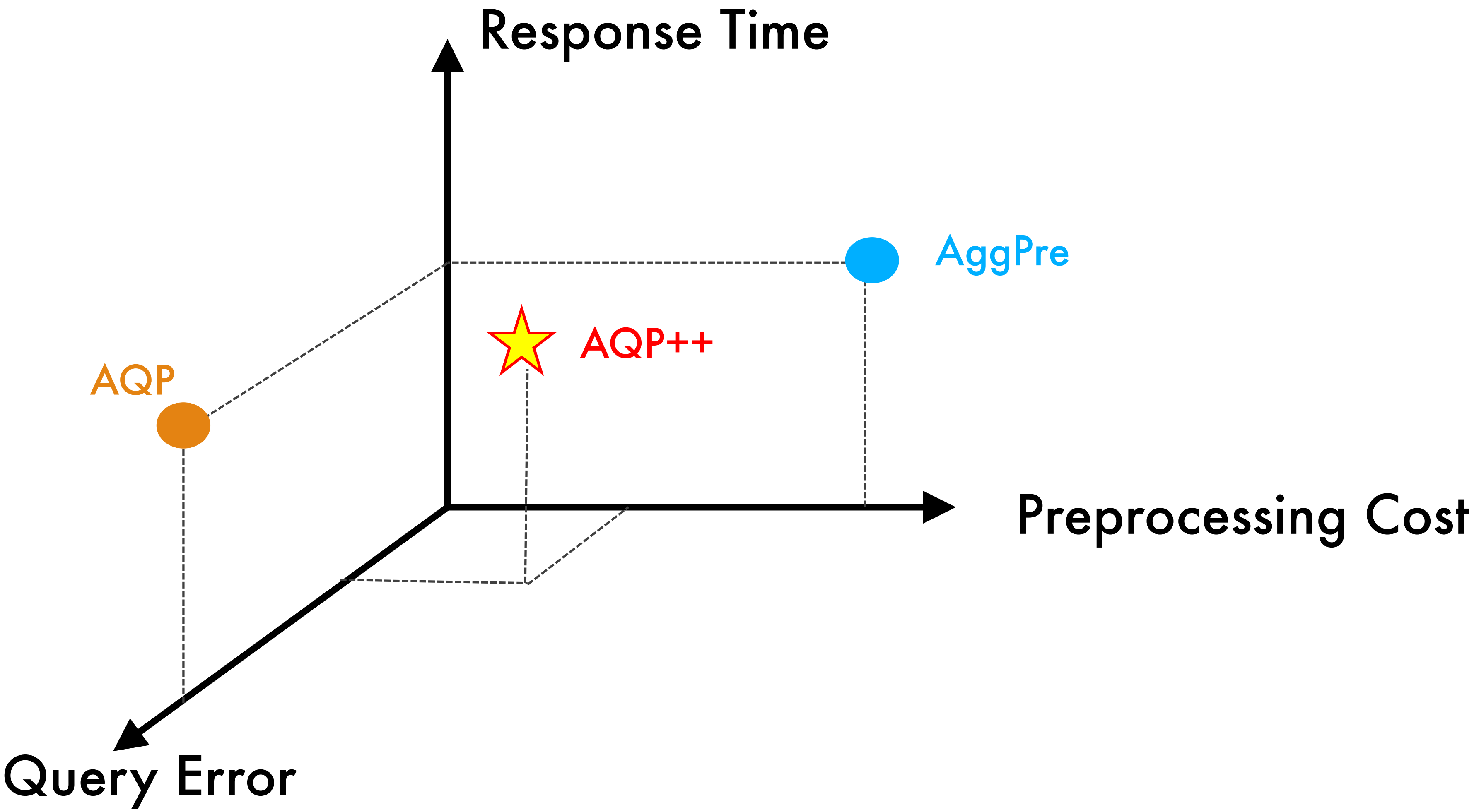
→

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

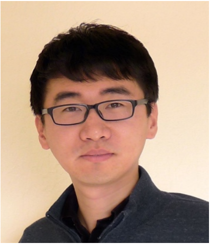
[1] Ho, Ching-Tien, et al. Range queries in OLAP data cubes. (1997)

Trade-Off


Response Time
AggPre
AQP++
AQP
Preprocessing Cost
Query Error

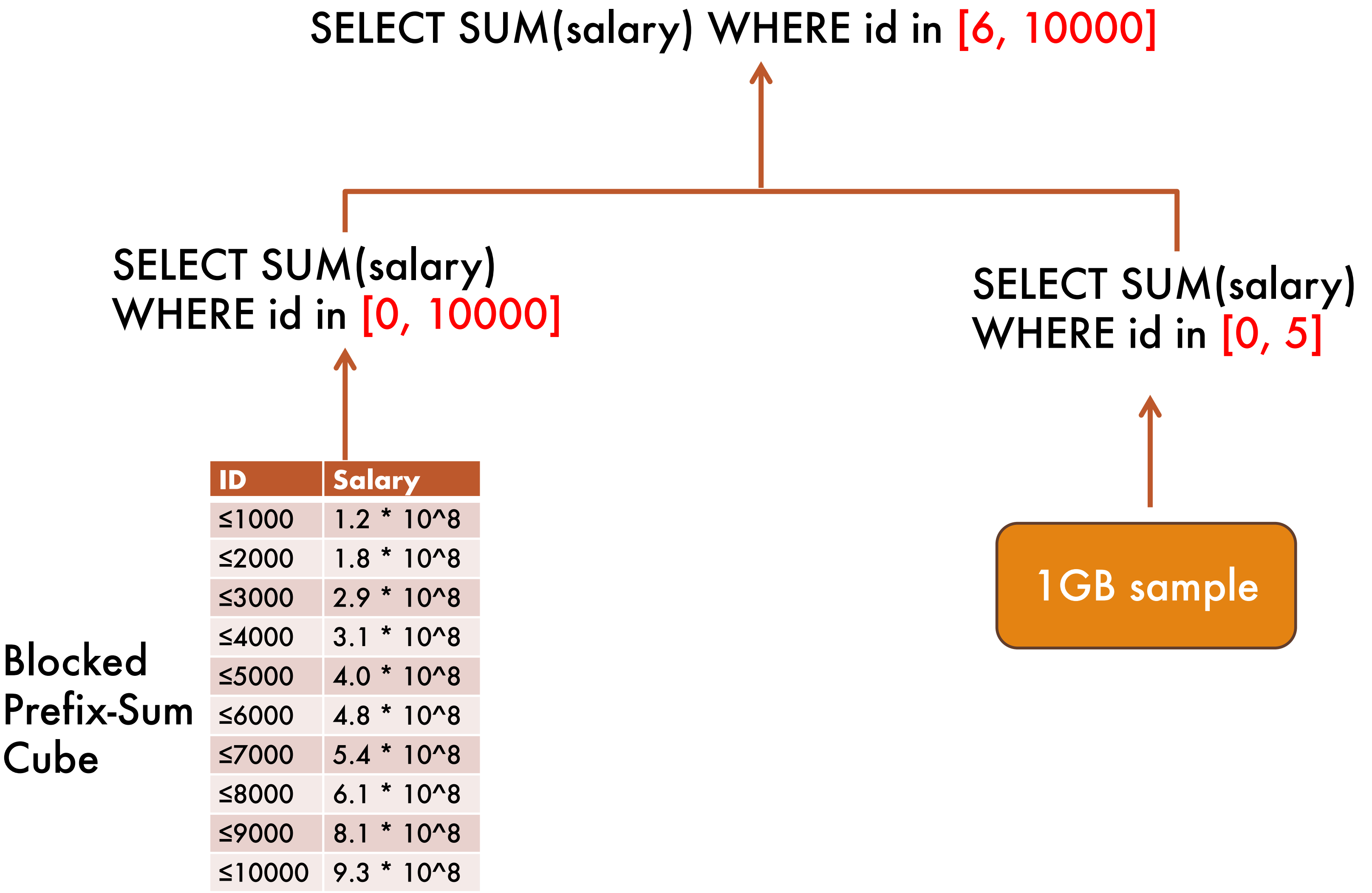
AQP++ (2016 -)

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 (to appear)**

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

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

1GB sample

Experimental Result

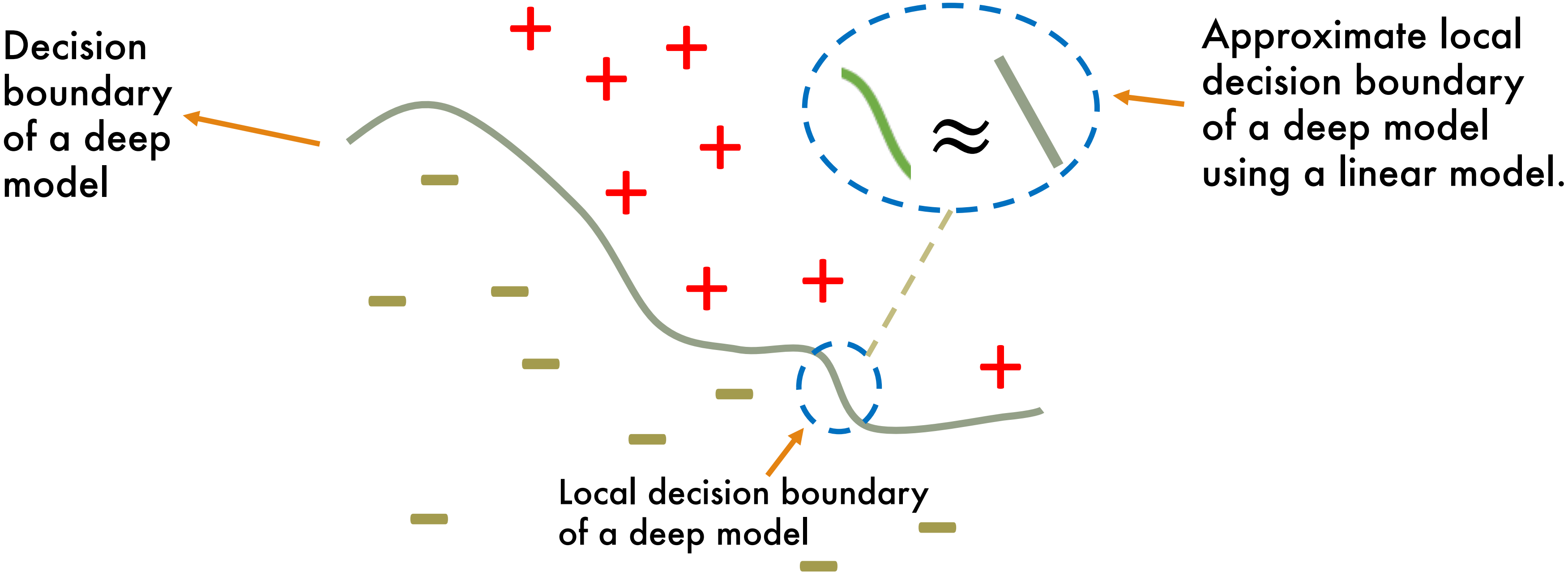
TPCD (Laptop,100GB)

- 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%

3 Posters From SFU

1. **Deeper** (Pei Wang)
2. **AQP++** (Jinglin Peng)
3. **DTLR: An Interpretation of Deep Neural Network** (Xia Hu)



Decision boundary of a deep model

Approximate local decision boundary of a deep model using a linear model.

Local decision boundary of a deep model

Take-away Messages

Our Mission

- Speeding Up Data Science

https://github.com/sfu-db

Thanks!

Deeper

- Leverage Deep Web to speed up data cleaning and enrichment

AQP++

- Connect AQP with AggPre to speed up data analysis