Data Acquisition

Xiaohui Yu

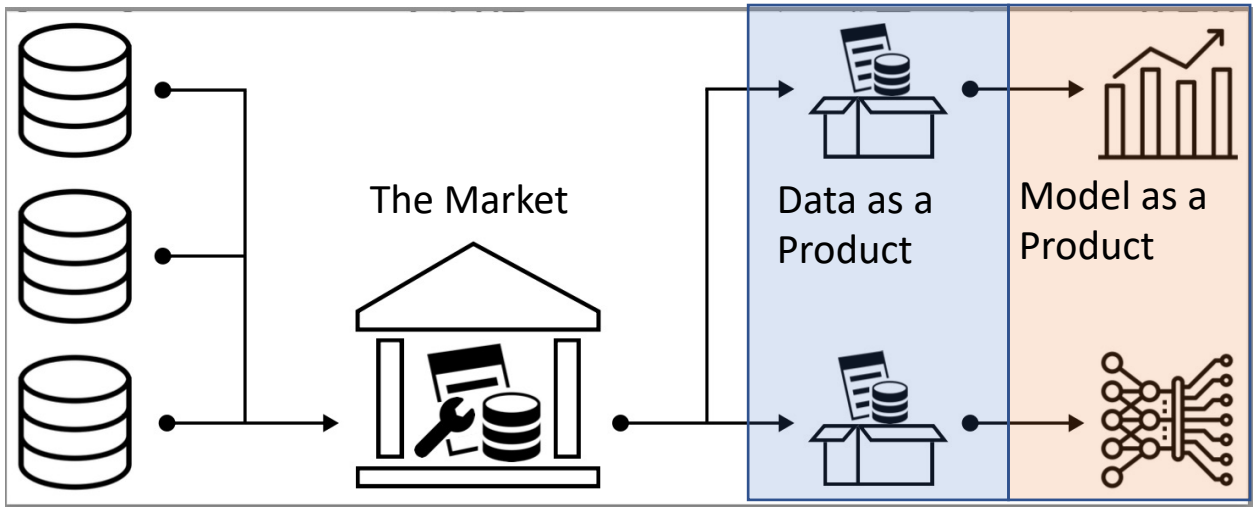
York University

xhyu@yorku.ca

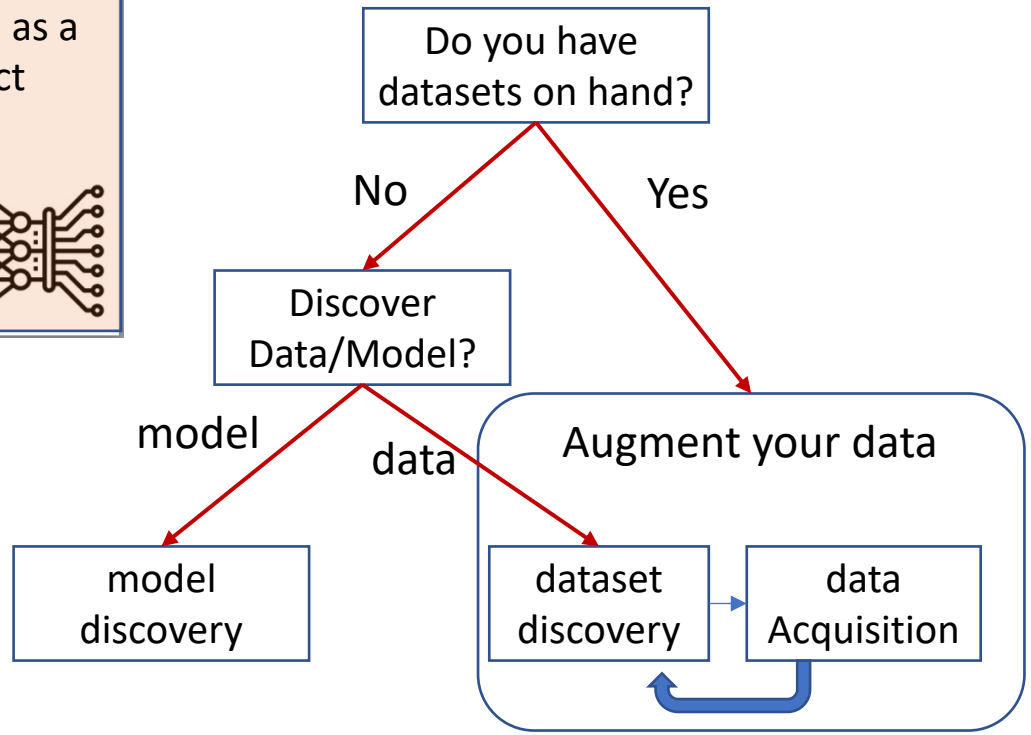
Outline

- Overview
- Instance Acquisition
- Feature Acquisition
- Value Acquisition

Overview

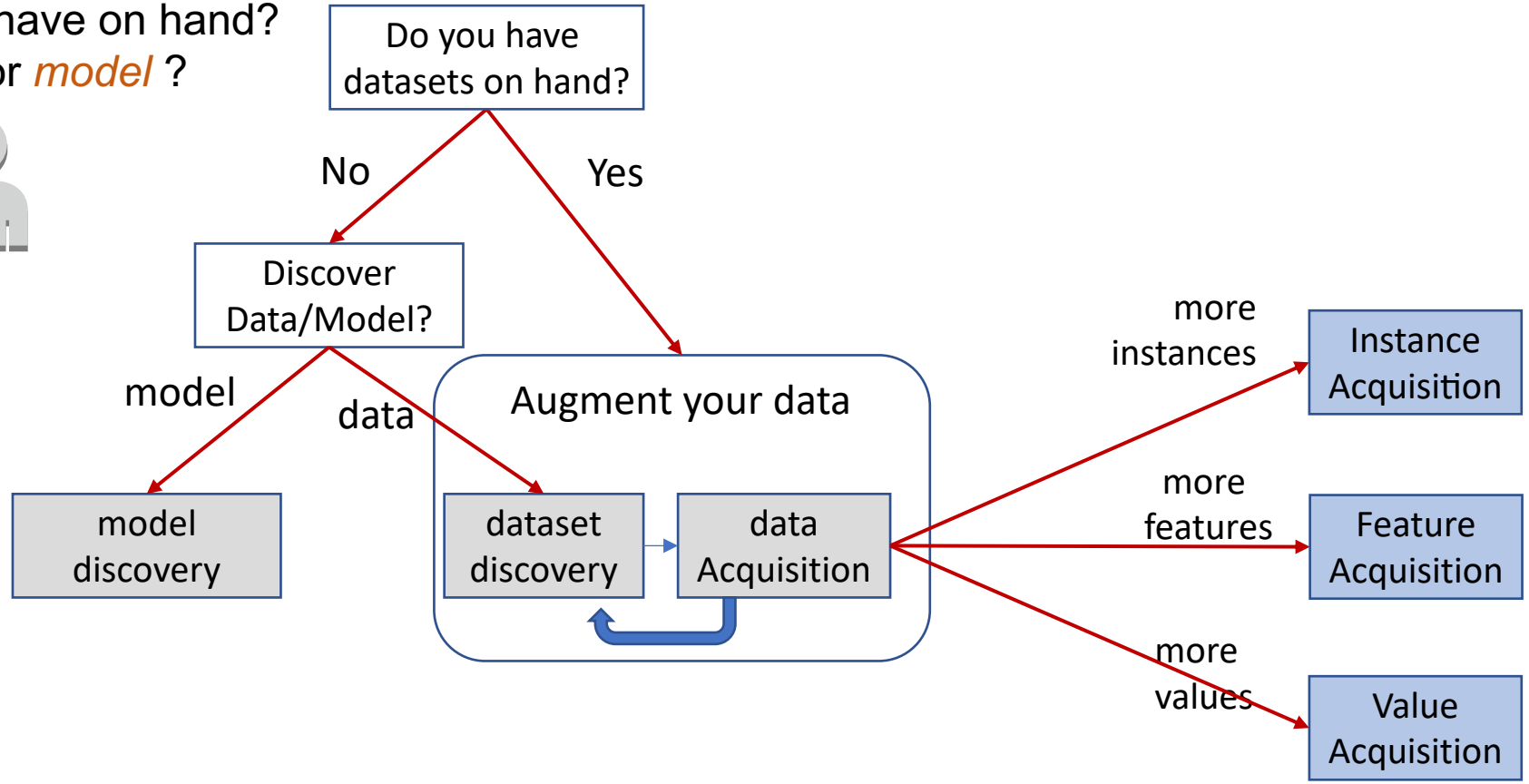


What do you have on hand?
Pay for *data* or *model* ?

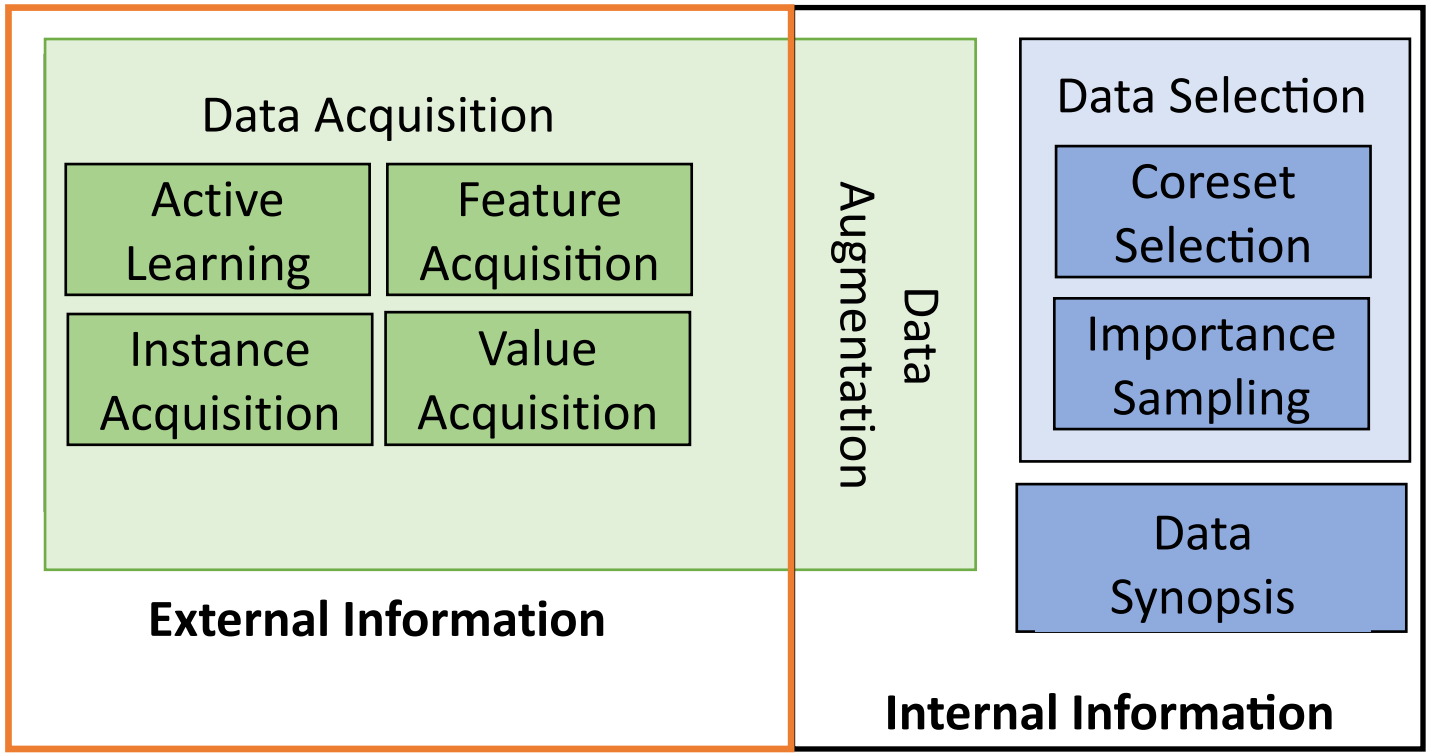


Overview

What do you have on hand?
Pay for *data* or *model* ?



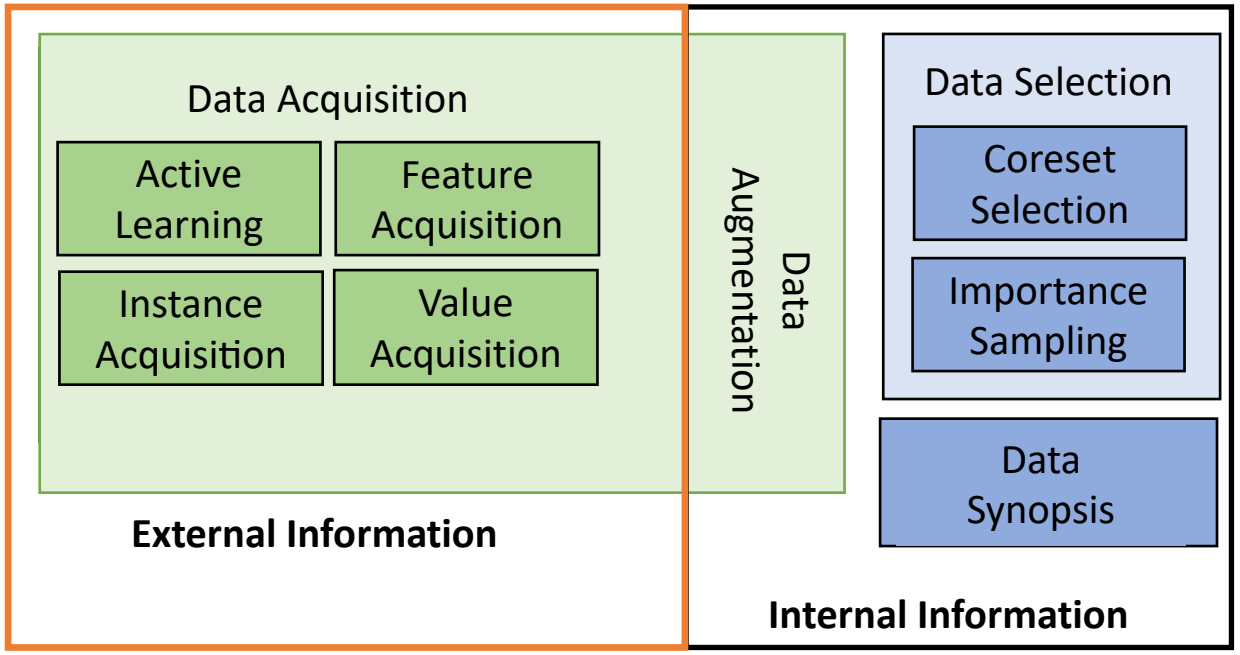
Overview



Mapping Relevant Concepts

Overview

- Relevant concepts:
  - Data Augmentation
  - Coreset Selection
  - Importance-sampling for ML

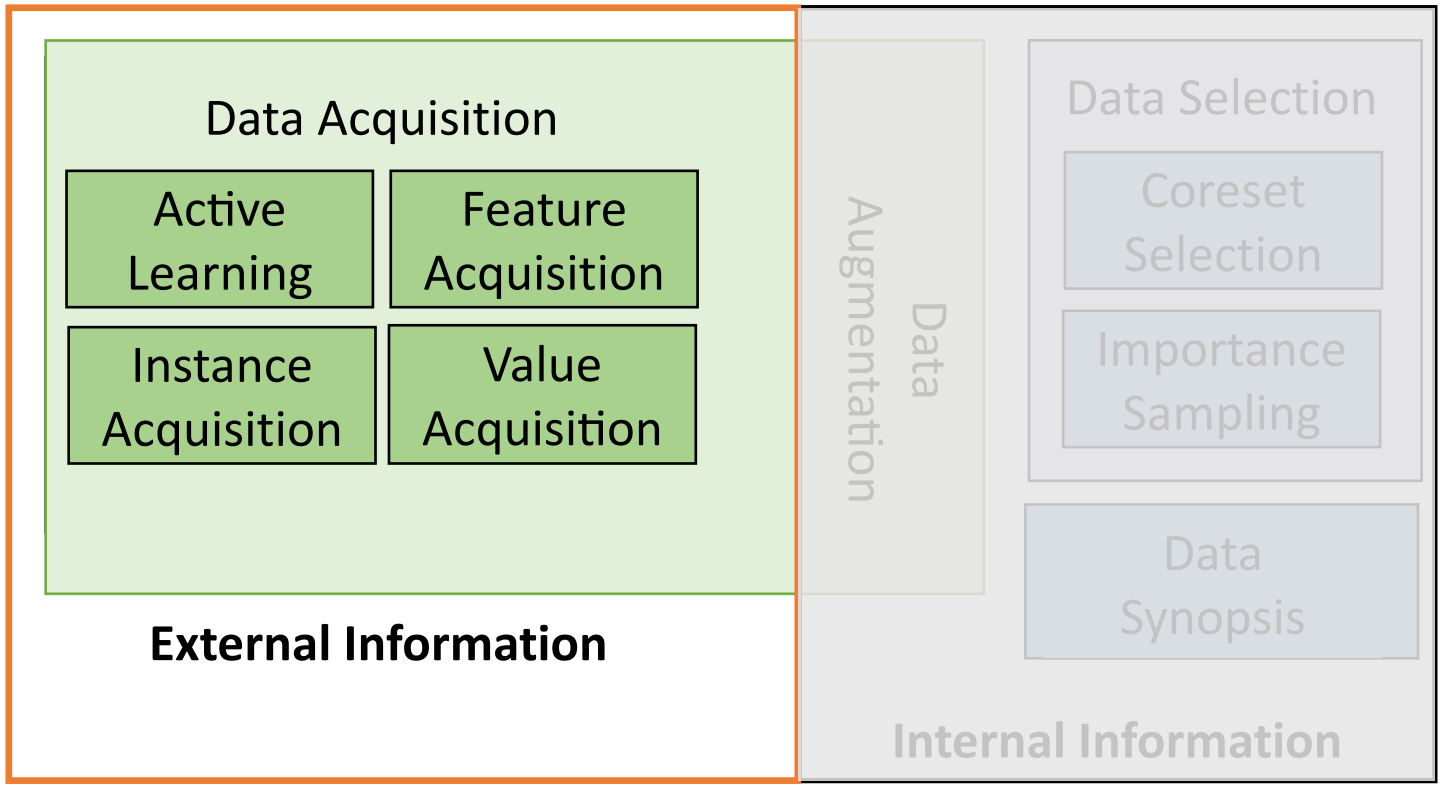


Mapping Relevant Concepts

Overview of Data Acquisition

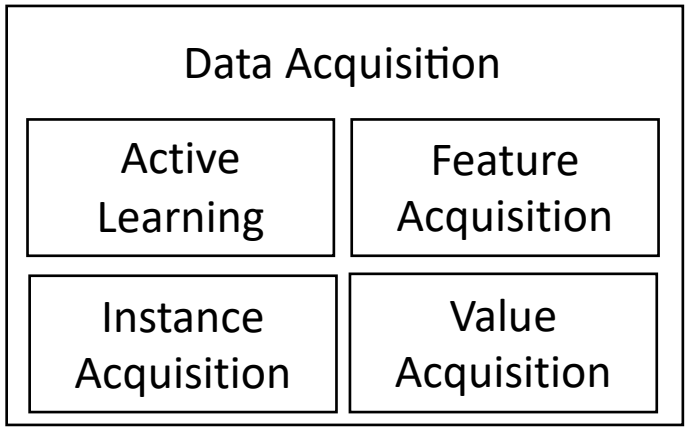
Relevant concepts:

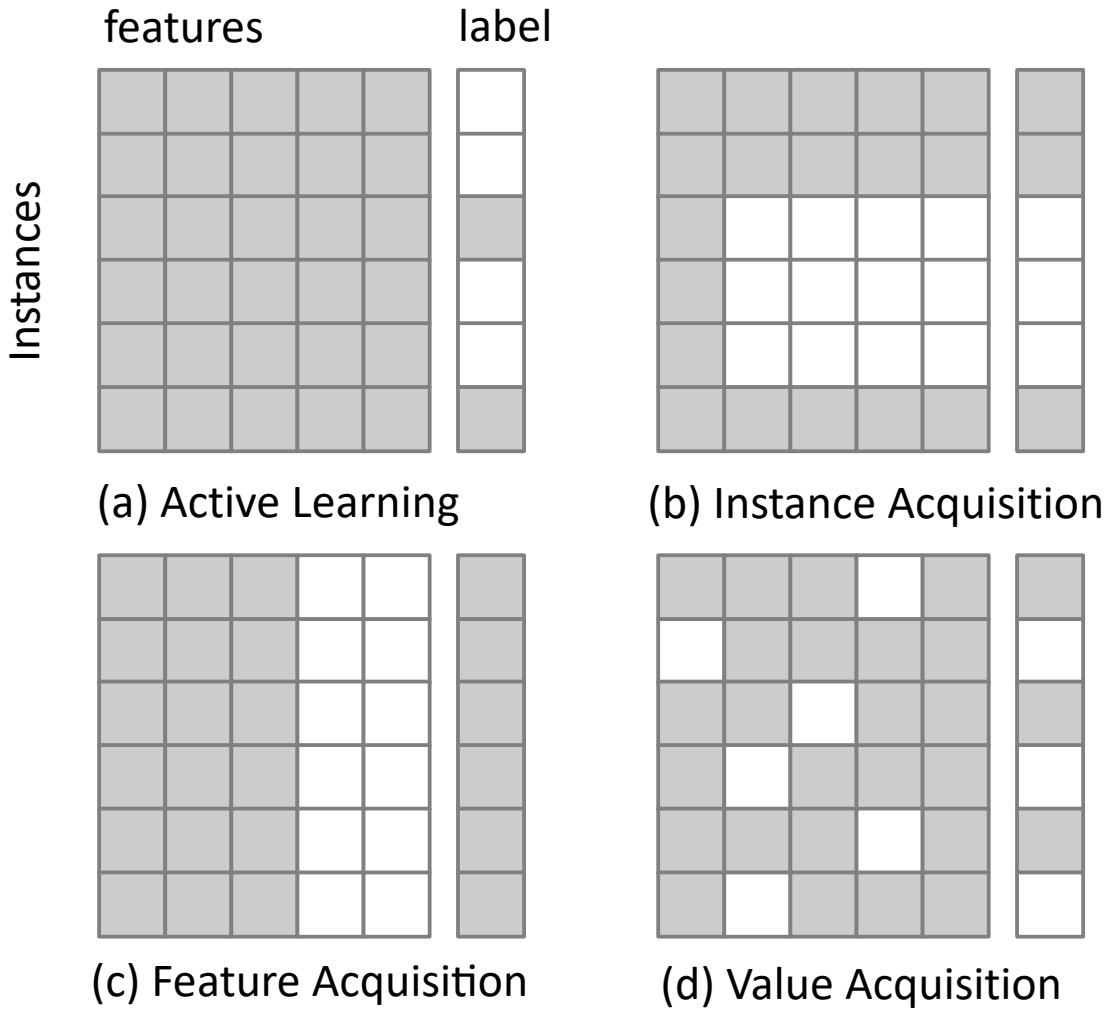
- Data Augmentation
- Coreset
- Importance-sampling for ML
- Crowdsourcing



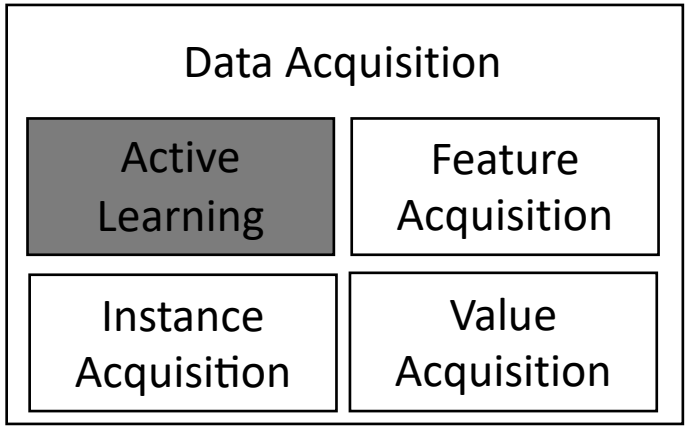
Mapping Relevant Concepts

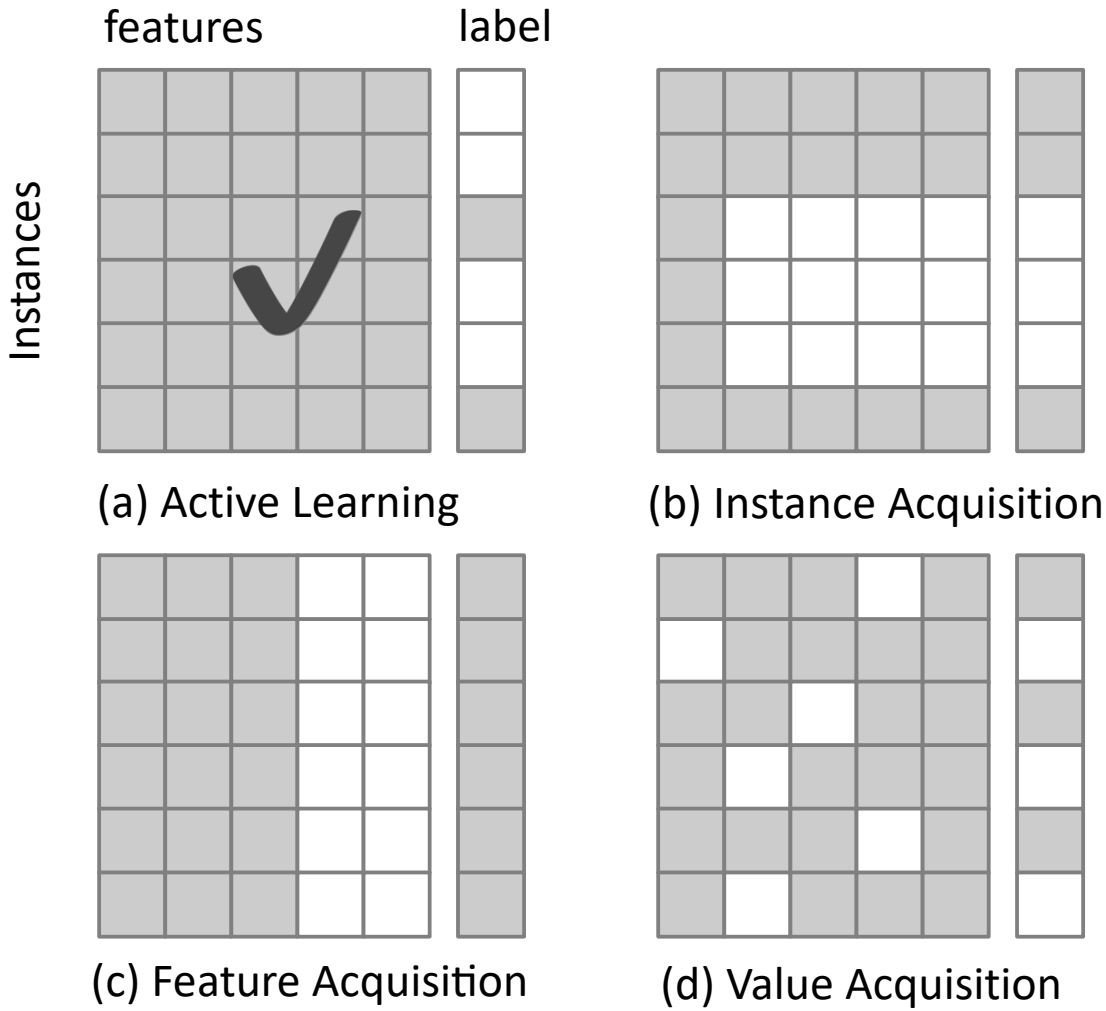
Overview of Data Acquisition



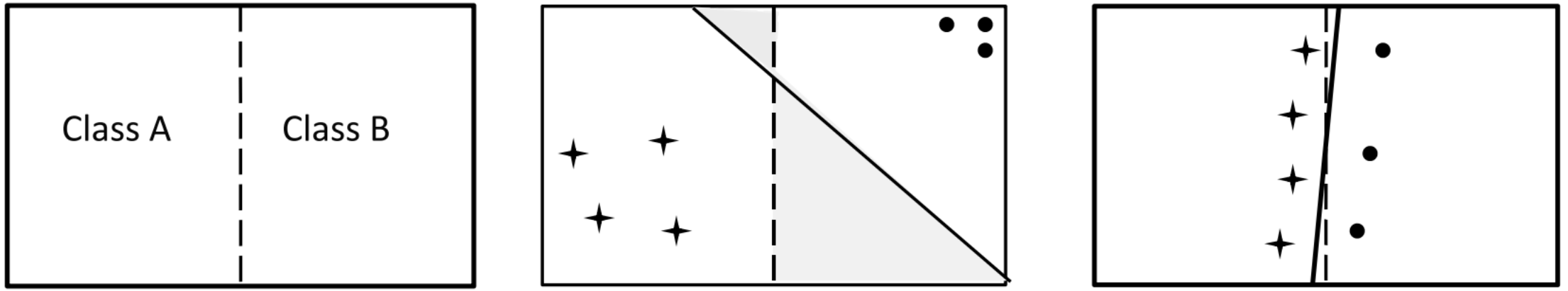


Overview of Data Acquisition





Active Learning



(a) Class Separation

(b) Random Sample with SVM Classifier

(c) Active Sample with SVM Classifier

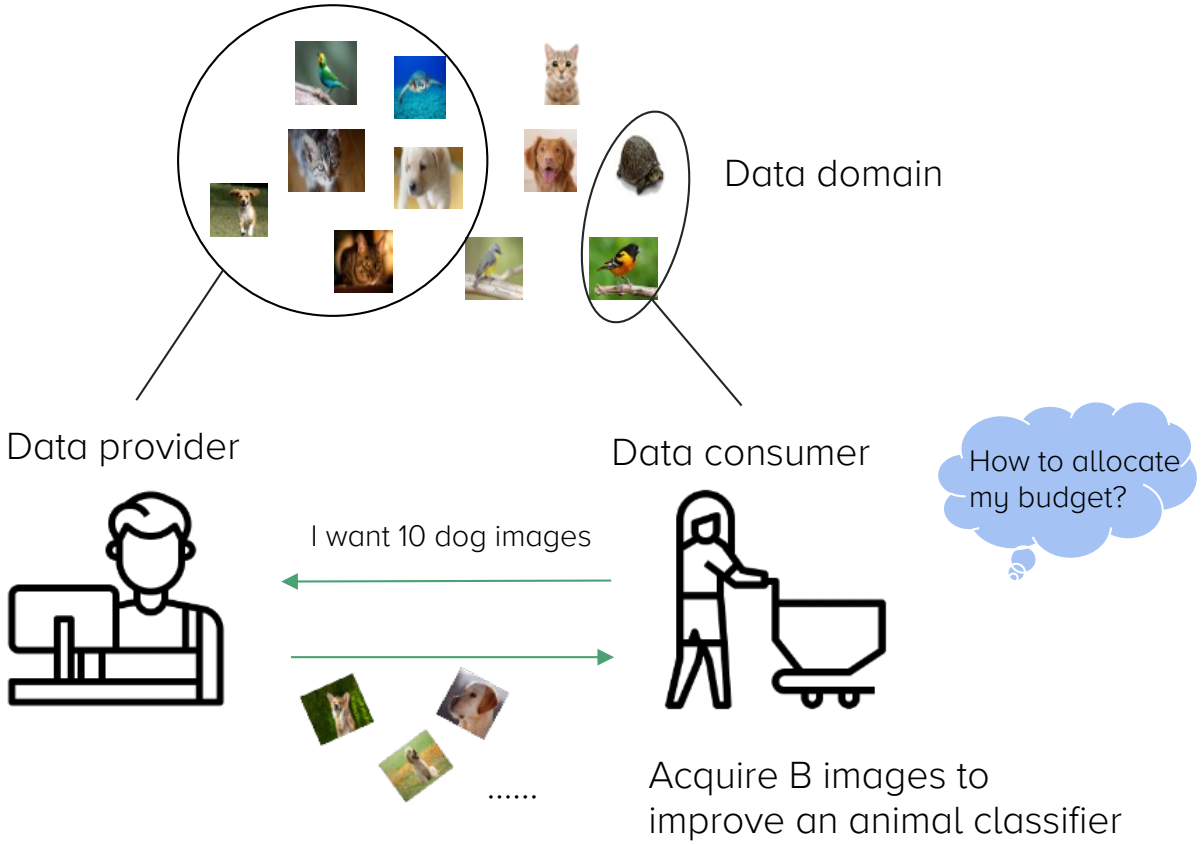
Aggarwal, C. C., Kong, X., Gu, Q., Han, J., & Yu, P. S. Active learning: A survey. In *Data Classification: Algorithms and Applications* (pp. 571-605), 2014.

Active Learning

- Aggarwal, C. C., Kong, X., Gu, Q., Han, J., & Yu, P. S. Active learning: A survey. In *Data Classification: Algorithms and Applications* (pp. 571-605), 2014.
- Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. Journal of artificial intelligence research 4 (1996): 129-145.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1183–1192. 2017.
- Kirsch, Andreas, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems 32 (2019).

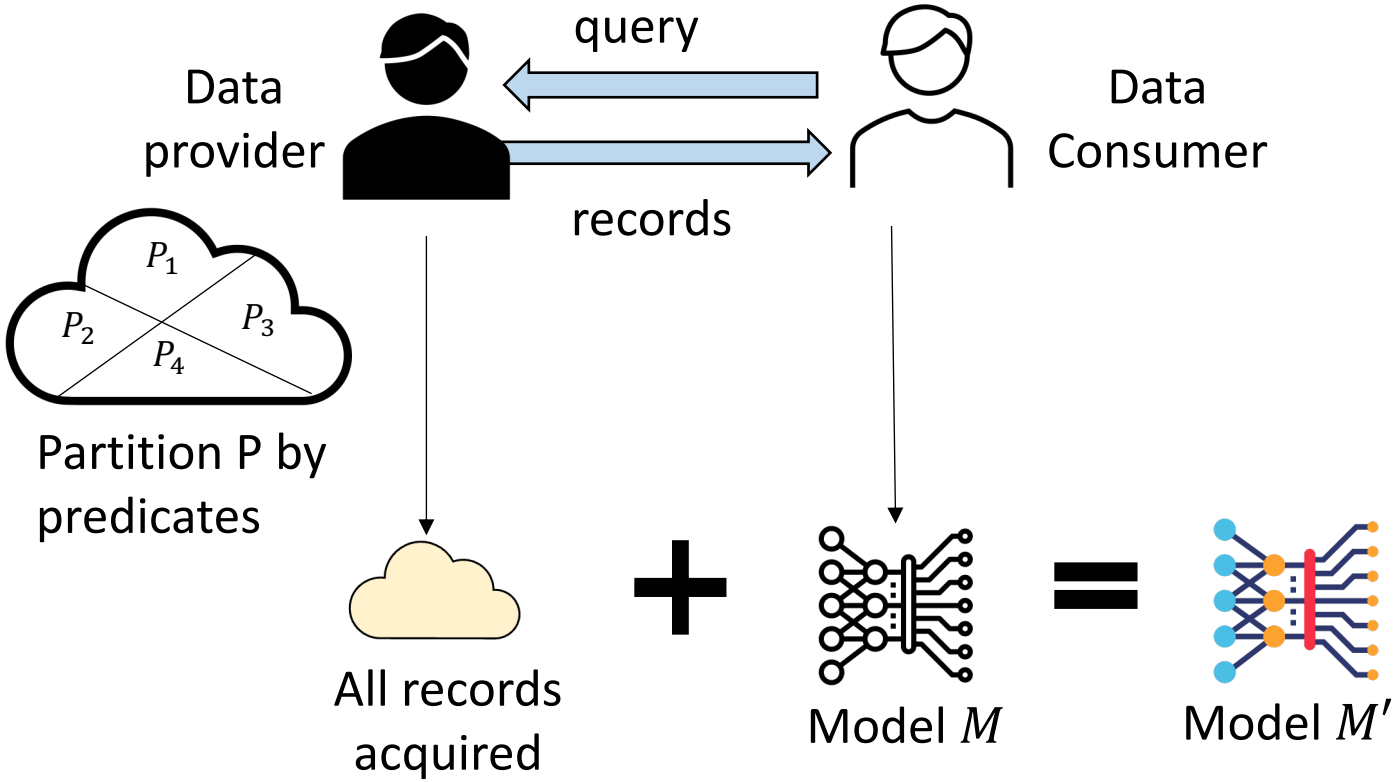
Utility-based Instance Acquisition – Part 1

Data Acquisition for Improving Machine Learning Models [VLDB 2021]



Objective: acquire instances with the highest utilities

Utility-based Instance Acquisition – Part 1



Which records are more beneficial (with higher utility)?
New objective: Acquire more records with high utility

Yifan Li, Xiaohui Yu, Nick Koudas: Data Acquisition for Improving Machine Learning Models. Proc. VLDB Endow. 14(10): 1832-1844 (2021)

Utility-based Instance Acquisition – Part 1

Method 1: An Estimation-and-Allocation (EA) solution

1) *Estimation stage* : estimate the utility of each predicate by acquiring some records

Predicate	True utility	Estimated utility
P_1	U_1	U_1'
P_2	U_2	U_2'
……	……	……
P_4	U_4	U_4'

Statistical test

$\epsilon - \delta$ approximation

$\Pr(|U_i - U_i'| > \epsilon) < \delta$

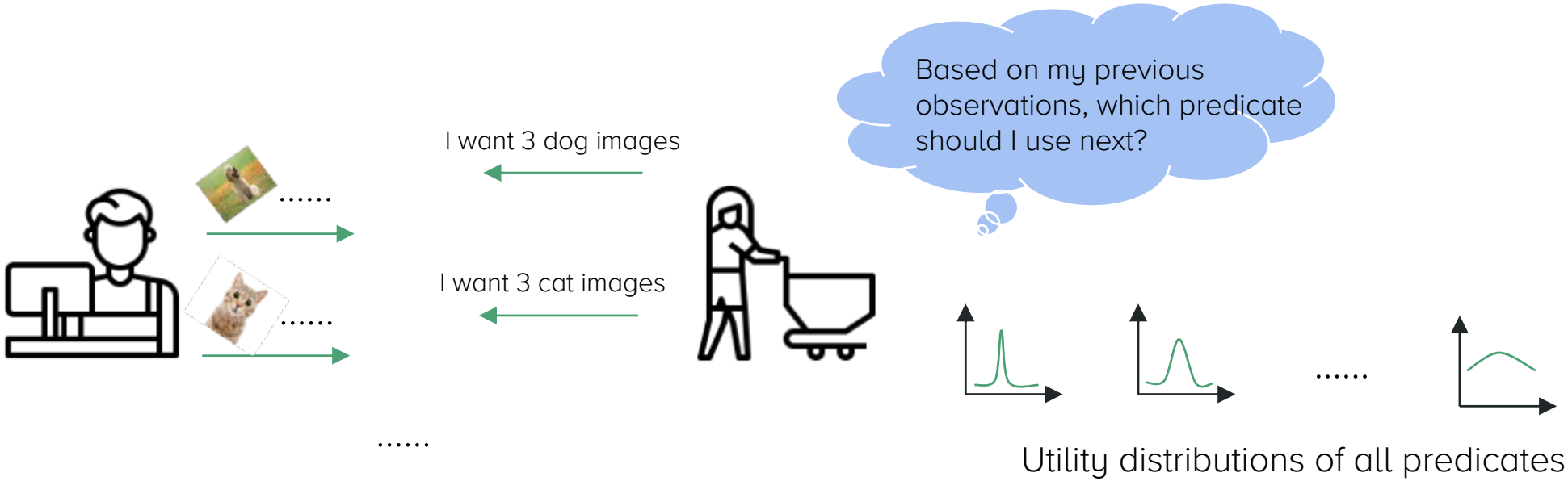
More records for estimation → More likely to pass the tests for small ϵ, δ

2) *Allocation stage* : allocate the remaining budget based on the estimated utilities

Strategy	Formula	Property
Linear allocation	$B_r U_i' / (U_1' + U_2' + \cdots + U_n')$	More biased to predicates with high utilities
Square-root allocation	$B_r \sqrt{U_i'} / (\sqrt{U_1'} + \sqrt{U_2'} + \cdots + \sqrt{U_n'})$	More uniform

Utility-based Instance Acquisition – Part 1

Method 2: A Sequential Predicate Selection (SPS) solution



Thompson Sampling Selector

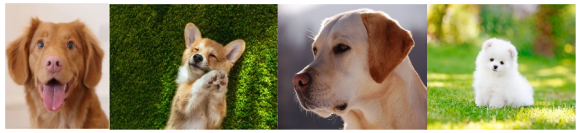
Exploitation: whose expected utility is higher?

Exploration: whose utility estimation is more uncertain?

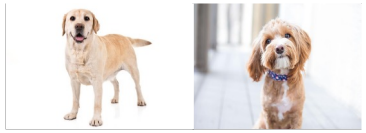
Utility-based Instance Acquisition – Part 1

- Utility measure: which predicate (images) are more useful (have high utility) to the consumer?
- One possible utility measure: novelty

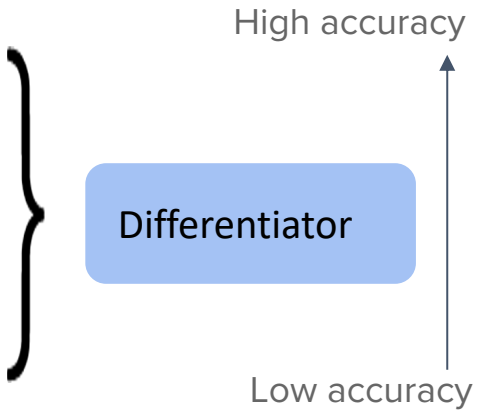
  Assumption: high utility ⇔ more new information



Provider's dog images



Consumer's dog images



Novelty: requires no model re-training!

Utility-based Instance Acquisition – Part 2

Data Acquisition for Improving Model Confidence [SIGMOD 2024]

Confidence metric

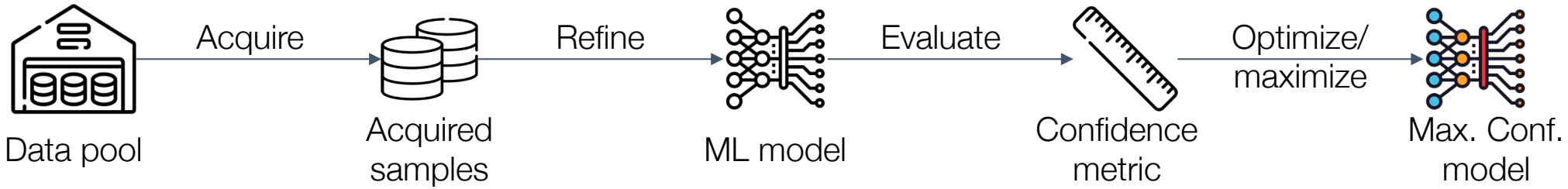
$$\arg\max_{\mathcal{D}_B \in \mathcal{D}} \text{conf}[\mathcal{M}(\mathcal{T} \cup \mathcal{D}_B)]$$

Acquired samples of size B

Data pool

ML model

Original training data



Yifan Li, Xiaohui Yu, Nick Koudas: Data Acquisition for Improving Model Confidence, in SIGMOD 2024.

Utility-based Instance Acquisition – Part 2

- General format of existing confidence metrics:

The minimal distance between *e* and samples in *T* with different labels

$$\mathrm{conf}(\mathcal{M}) = \sum_{e \in \mathcal{E}} \mathrm{conf}(e) = \sum_{e \in \mathcal{E}} \mathrm{F}(d(e, \mathcal{T}), \bar{d}(e, \mathcal{T}))$$

Evaluation dataset

The minimal distance between *e* and samples in *T* with the same label

- Anticipated Confidence Improvement (ACI)
  - ACI of sample s: the model confidence improvement resulted from acquiring s
  - Acquiring s may change the ACI values of all other samples in the data pool!

Utility-based Instance Acquisition – Part 2

Brute-force Approach

- Enumerate all subset of the data pool with *B* samples and find the one leads to the maximal confidence improvement
- Complexity of the Brute-force approach:

$$O(\underbrace{\frac{\mathcal{D}!}{B! * (\mathcal{D}-B)!}}_{\text{Possible number of solutions}} * B * \overbrace{\underbrace{|\mathcal{E}|}_{\text{Cardinality of the evaluation set}} * (|\mathcal{T}| + B) * \underbrace{D}_{\text{Data space dimensionality}}}^{\text{Cost of ACI calculation for each sample}})$$

Under what condition, can we solve the problem in poly-time?

Utility-based Instance Acquisition – Part 2

Optimization Opportunity - 1

- Do we need to exhaustively search all possible subsets of size B?

$$O(\frac{\mathcal{D}!}{B! * (\mathcal{D} - B)!} * B * |\mathcal{E}| * (|\mathcal{T}| + B) * D)$$

Utility-based Instance Acquisition – Part 2

- Condition: Top-B independence
  - Let S consists of the B samples with the highest ACIs, if acquiring any sample in S does not change ACI of other samples in S, top-B independence is satisfied
- Bulk Acquisition (BA)
  - Acquire the B samples with the highest ACIs
  - Optimal solution under Top-B independence:
  - Complexity:

$$O(|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D)$$

*Please check our paper for the formal definition of Top-B independence and how to verify the property

Utility-based Instance Acquisition – Part 2

- Dominance
  - Acquiring sample s leads to higher confidence improvement w.r.t each evaluation sample than acquiring another sample t: ***s dominates t***
- *Progressive dominance*
  - A scenario under which we can find a sample s dominating |D|-B samples at each round of the acquisition process
- Sequential Acquisition (SA)
  - Acquire the sample that dominates |D|-B samples each round, and update the ACI of remaining samples
  - Optimal solution under progressive dominance
  - Complexity:

$$O(B * \max\{|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D, |\mathcal{D}|^2 * |\mathcal{E}|\})$$

*Please check our paper for the formal definition of Progressive Dominance and how to verify the property

Utility-based Instance Acquisition – Part 2

Optimization Opportunity - 2

- Do we need to search the entire data pool?

$$O(|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D)$$

$$O(B * \max\{|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D, |\mathcal{D}|^2 * |\mathcal{E}|\})$$

Utility-based Instance Acquisition – Part 2

Neighbor-based Acquisition

- Observation: Only samples in the data pool that are close to samples in E will improve the model confidence
- Neighbor-based pruning
  - For each sample in E, find its kNNs in the data pool, forming the candidate set
- Heuristic solution: conduct BA or SA on the candidate set
  - Called **kNN-BA** and **kNN-SA**
- Faster solution with lower ultimate confidence improvement; but the suboptimality can be bounded! (check paper for details)

Utility-based Instance Acquisition – Part 2

Optimization Opportunity - 3

- Do we need to actually compute the ACI?

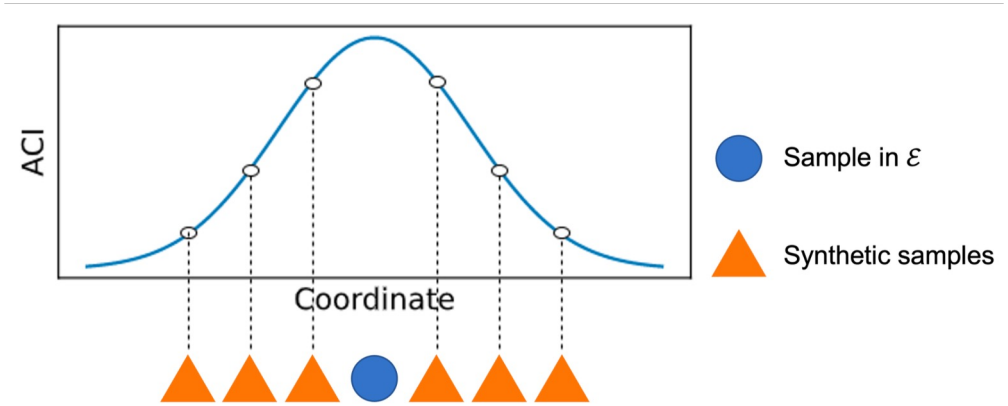
$$O(|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D)$$

$$O(B * \max\{|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D, |\mathcal{D}|^2 * |\mathcal{E}|\})$$

Utility-based Instance Acquisition – Part 2

Distribution-based Acquisition (DA)

- Overall design
  - Train a regression model offline for ACI prediction
  - Use the regression model to estimate the ACI of new samples
- Illustration:



- Fastest solution, support data acquisition in streaming settings

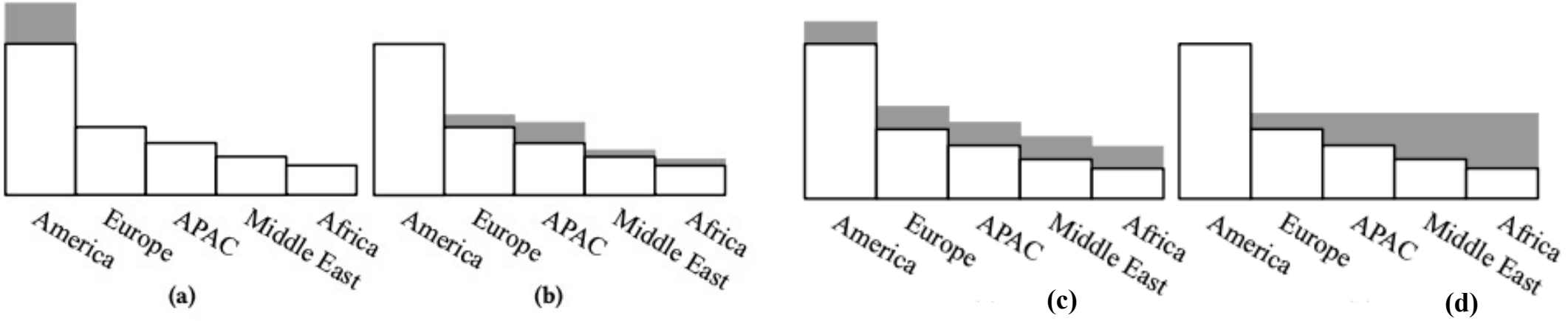
Utility-based Instance Acquisition – Part 2

Summary

- The problem of **data acquisition for improving ML confidence given a budget**
- Acquisition strategies: **BA** and **SA**, each leading to the optimal solution under certain assumptions
- Optimization: **neighbor-based pruning**, greatly reducing the acquisition overhead with slight sacrifice of confidence improvement
- Optimization: **Distribution-based acquisition** to support data acquisition in streaming settings

Utility-based Instance Acquisition – Part 3

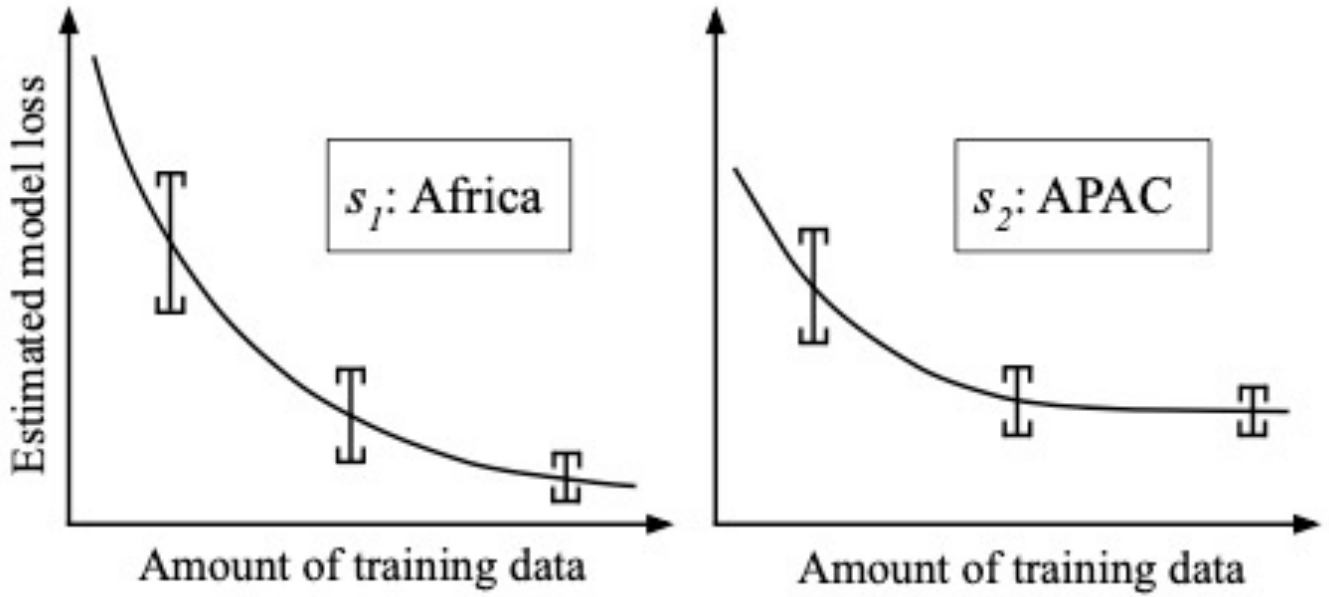
Slice Tuner [SIGMOD 2021]: Considers both model performance and fairness



Customer slices in different regions, where the boxes represent the initial slices (height = slice size) while the gray bars on top indicate the amounts of acquired data per slice.

Adapted from Tae, Ki Hyun, and Steven Euijong Whang. "Slice tuner: A selective data acquisition framework for accurate and fair machine learning models." In *SIGMOD*, pp. 1771-1783. 2021.

Utility-based Instance Acquisition – Part 3



Hypothetical learning curves of two slices.

Tae, Ki Hyun, and Steven Euijong Whang. "Slice Tuner: A selective data acquisition framework for accurate and fair machine learning models." In *SIGMOD*, pp. 1771-1783. 2021.

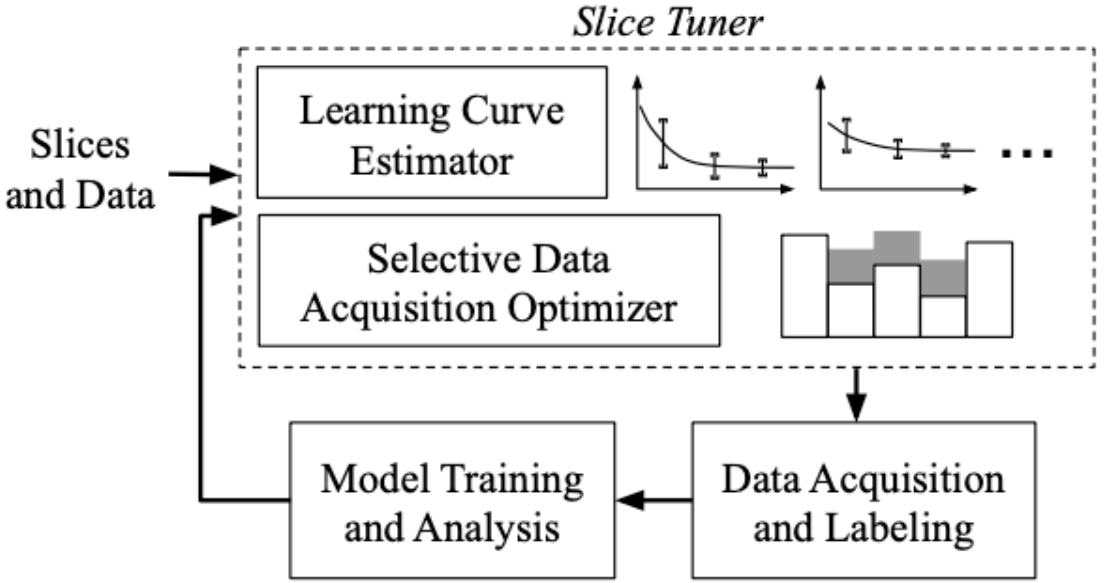
At the core is the ability to estimate the learning curves of slices, which reflect the cost benefits of data acquisition.

Utility-based Instance Acquisition – Part 3

- Slice Tuner Problem definition:

Given a set of examples D, its slices $S = \{s_i\}_{i=1}^{n}$, a trained model M, a data acquisition cost function C, and a data acquisition budget B, the selective data acquisition problem is to acquire d_i examples for each slice $s_i \in S$ such that the following are all satisfied:

(1) The average loss $\psi(D,M)$ is minimized,
(2) The unfairness $avg_{s_i \in S} |\psi(s_i,M) - \psi(D,M)|$ is minimized, and
(3) The total data acquisition cost $\sum_i C(s_i) \times d_i = B$.



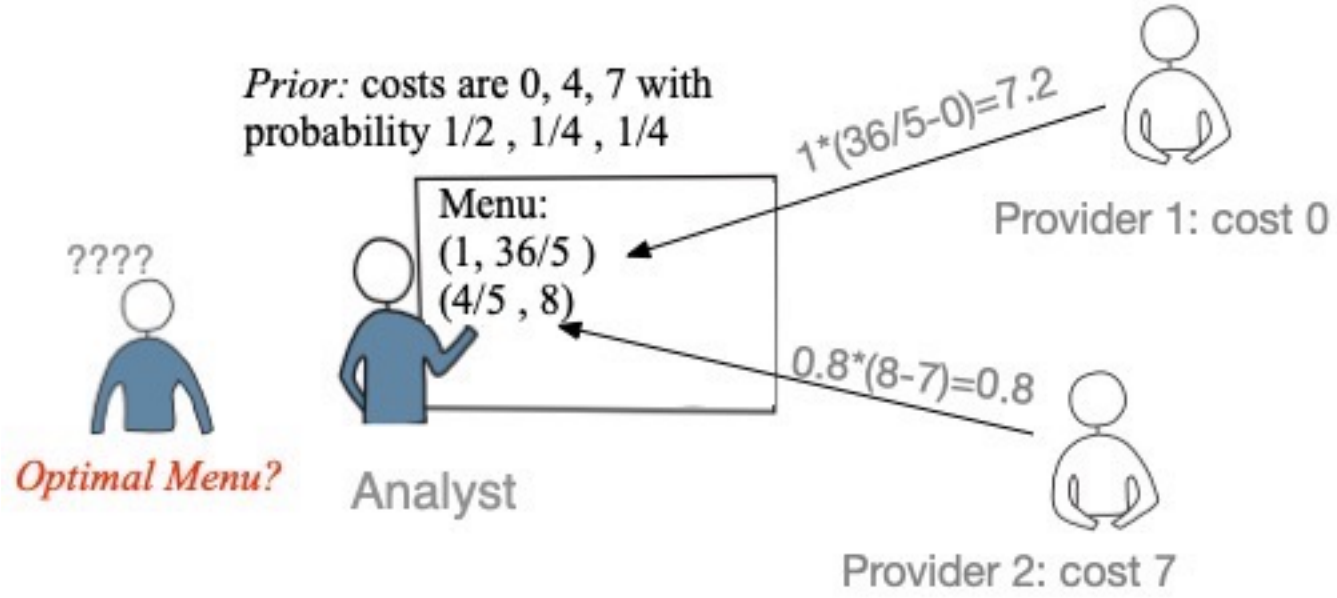
Workflow of Slice Tuner.

Tae, Ki Hyun, and Steven Euijong Whang. "Slice tuner: A selective data acquisition framework for accurate and fair machine learning models." In *SIGMOD*, pp. 1771-1783. 2021.

Mechanism Design for Instance Acquisition

Data Acquisition for Statistical Estimation [EC 2018]

- focusing on designing optimal incentive compatible mechanisms to maximize both data providers' and consumers' payoff.
- E.g., buying verifiable data from a population in order to estimate a statistic of interest, such as the expected value of some function of the underlying data.

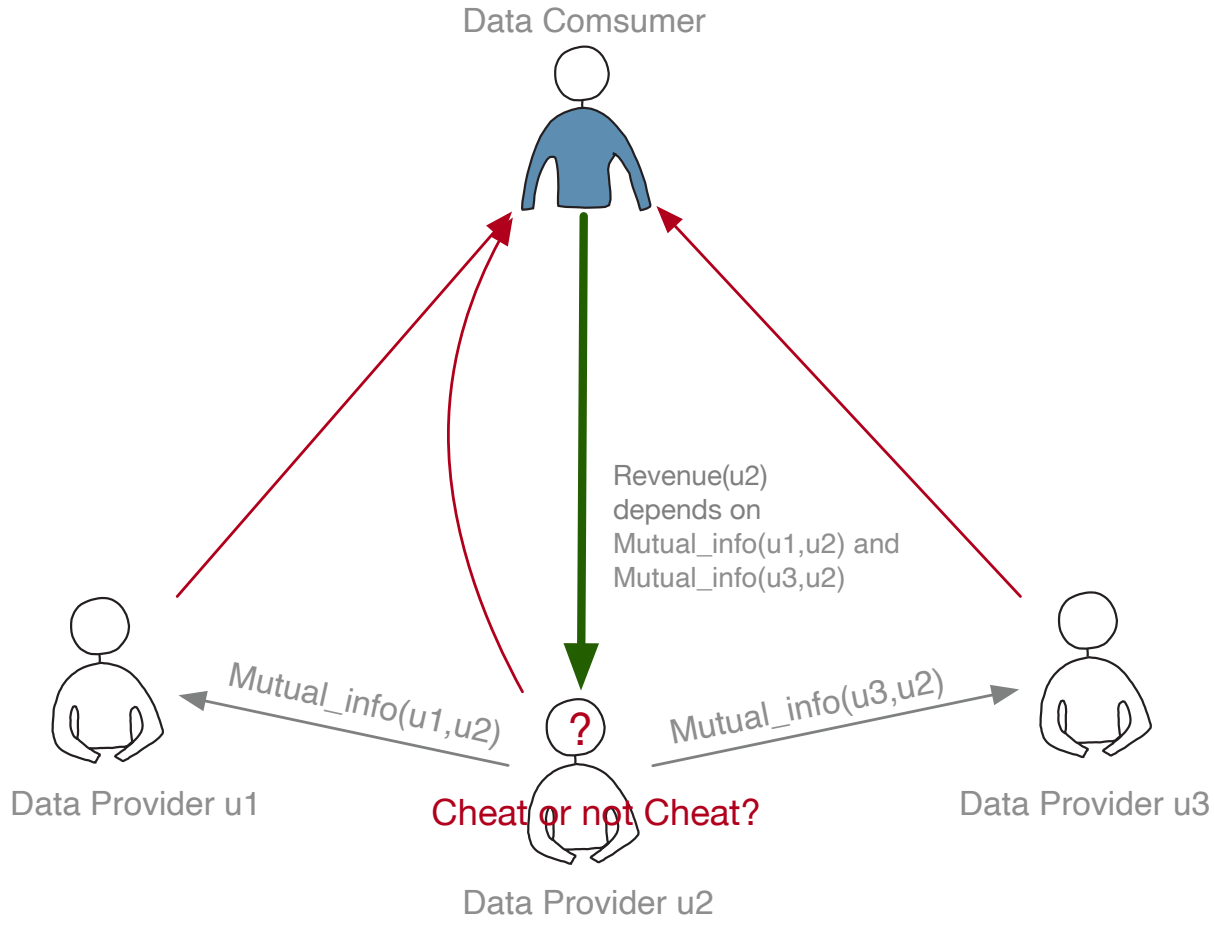


Chen, Yiling, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. "Optimal data acquisition for statistical estimation." In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 2018.

Mechanism Design for Instance Acquisition

Truthful data acquisition [NeurIPS 2020]

- The data analyst has a budget to purchase datasets from multiple data providers. She does not have any test data that can be used to evaluate the collected data and can assign payments to data providers solely based on the collected datasets.
- a game-theoretic mechanism to motivate the data providers to report data truthfully.

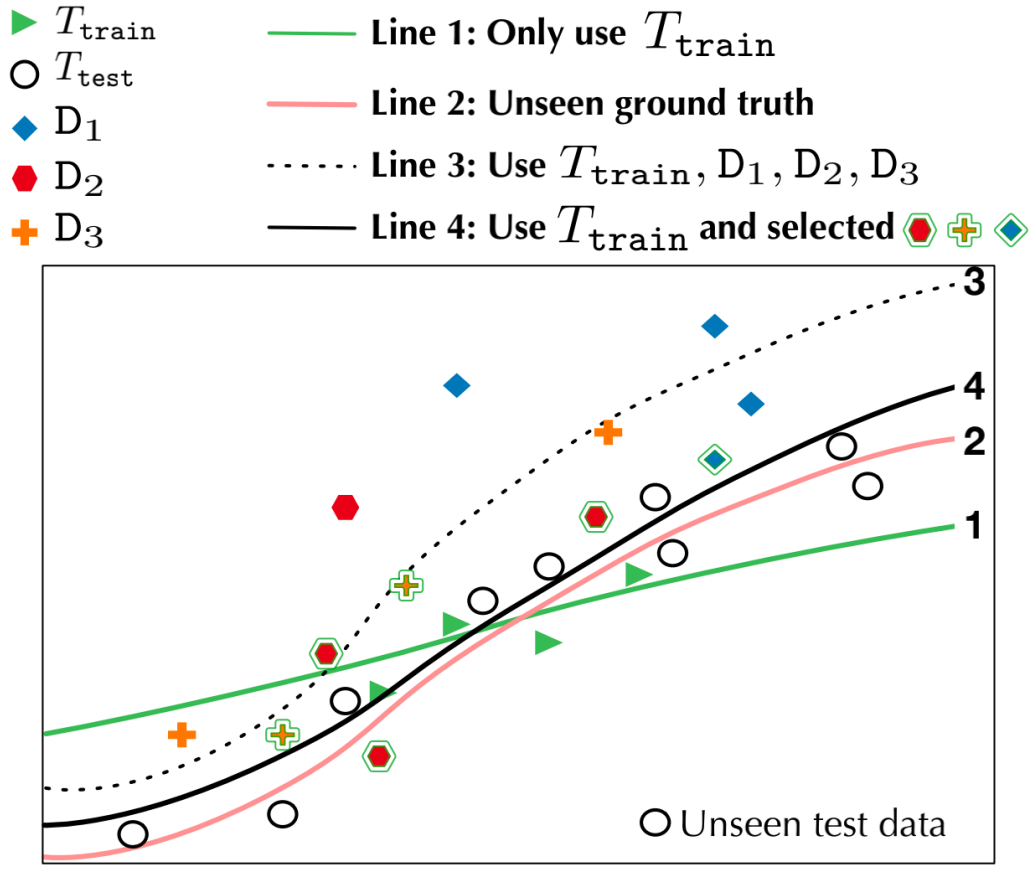


Chen, Yiling, Yiheng Shen, and Shuran Zheng. "Truthful data acquisition via peer prediction." In *NeurIPS* 33 (2020): 18194-18204.

Mechanism Design for Instance Acquisition

- Chen, Yiling, and Shuran Zheng. "Prior-free data acquisition for accurate statistical estimation." In Proceedings of the 2019 ACM Conference on Economics and Computation, pp. 659-677. 2019.
- Chen, Yiling, and Shuran Zheng. "Prior-free data acquisition for accurate statistical estimation." Proceedings of the 2019 ACM Conference on Economics and Computation. 2019.
- Zheng, Shuran, et al. "Active information acquisition for linear optimization." arXiv preprint arXiv:1709.10061 (2017).
- Roth, Aaron, and Grant Schoenebeck. "Conducting truthful surveys, cheaply." Proceedings of the 13th ACM Conference on Electronic Commerce. 2012.
- Cai, Yang, Constantinos Daskalakis, and Christos Papadimitriou. "Optimum statistical estimation with strategic data sources." Conference on Learning Theory. PMLR, 2015.
- Abernethy, Jacob, et al. "Low-cost learning via active data procurement." Proceedings of the Sixteenth ACM Conference on Economics and Computation. 2015.

Methods without Explicit Consideration to Budget - AutoData

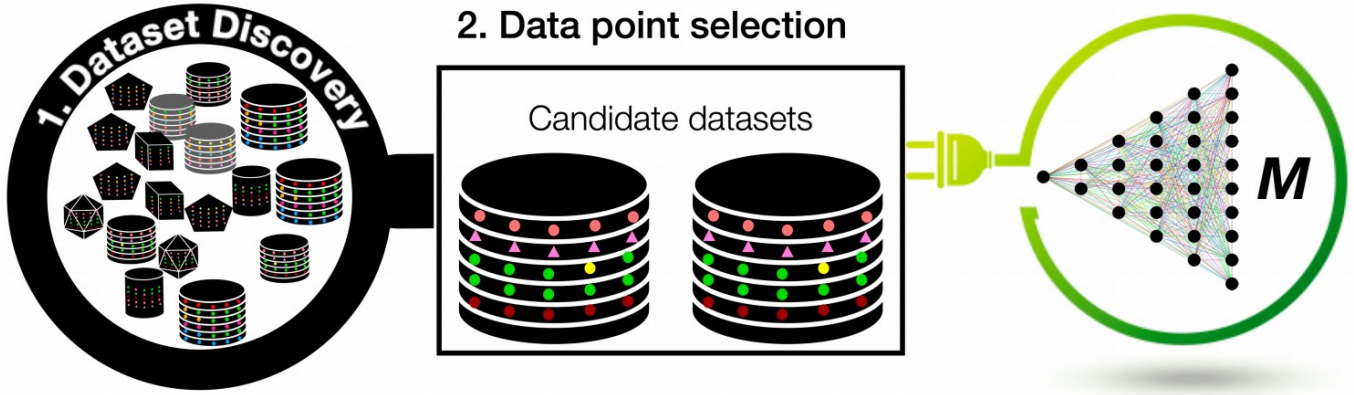


- T_{train}: small, lack of training data
- Model Line 1: trained on T_{train}, low performance
- How to improve? → *Data acquisition*

Chai, C., Liu, J., Tang, N., Li, G., & Luo, Y. (2022). Selective data acquisition in the wild for model charging. In *VLDB 2022*.

Methods without Explicit Consideration to Budget - AutoData

- Observations:
  - More data is needed
  - Not all data is useful
- Challenges:
  - Heterogeneous candidate datasets
  - Effectively select useful data points

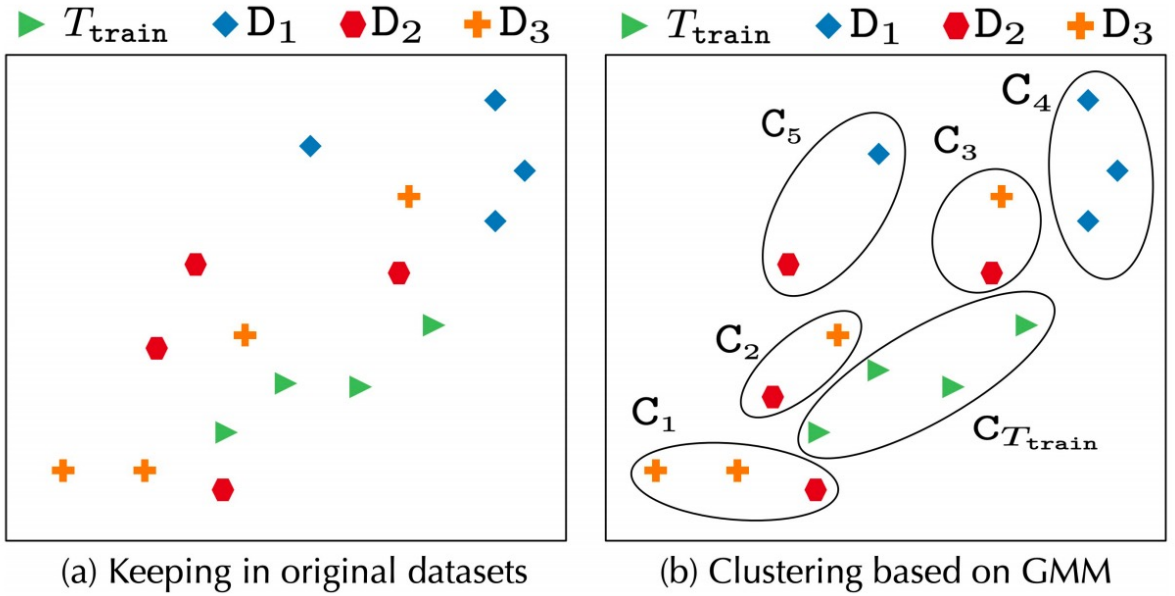


Framework of *Autodata*

Chai, C., Liu, J., Tang, N., Li, G., & Luo, Y. (2022). Selective data acquisition in the wild for model charging. *Proceedings of the VLDB Endowment*, *15*(7), 1466-1478.

Methods without Explicit Consideration to Budget - AutoData

- Modeling heterogeneous datasets
  - Goal: finding useful data points → fine-grained modeling
  - Clustering → partition all data into distinct groups
  - Data pool P → clusters $\mathbf{C} = \{C_1, \ldots, C_n\}$
  - C_i has its own distribution $p_i = (\mu_i, \Sigma_i)$
- Iterative data selection
  - Select
    - MAB-Based
    - DQN-Based
  - Retrain & Evaluate
  - Update
  - Repeat above process



(a) Keeping in original datasets

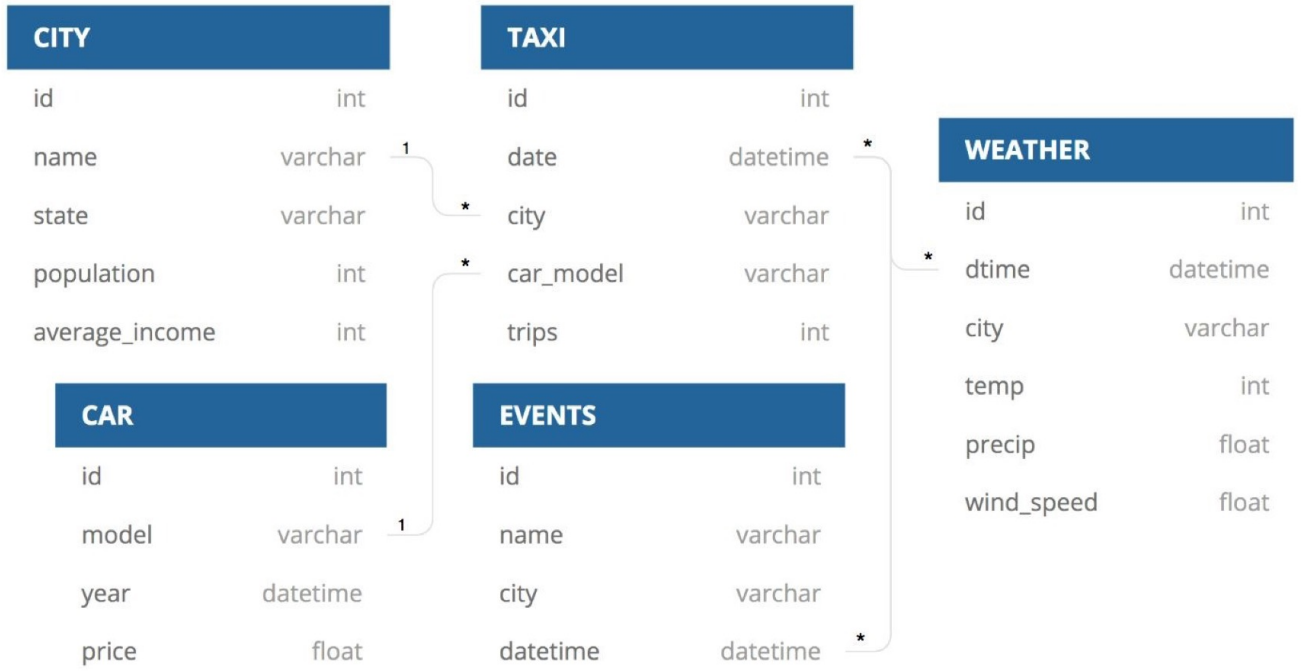
(b) Clustering based on GMM

Data Acquisition

- Instance Acquisition
- **Feature Acquisition**
- Value Acquisition

Feature Acquisition: ARDA

- What if the original dataset provided by the user does not contain enough signal (predictive features) to create an accurate model? We need to find features in joinable tables.



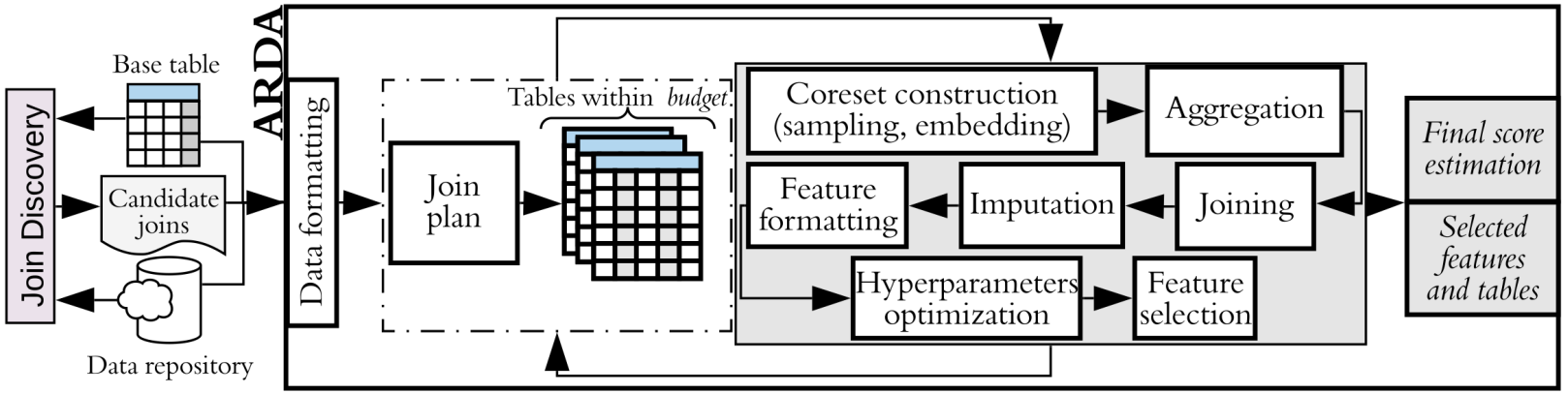
Example Schema:

- Initially a user has a base table TAXI
- She finds a pool of joinable tables to see if some of them can help improve prediction error for taxi demand for a specific date given in target column trips.

Chepurko, Nadiia, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. "ARDA: automatic relational data augmentation for machine learning." In *Proc. of the VLDB Endow.,* 13(9), 1373–1387, 2020

Feature Acquisition: ARDA

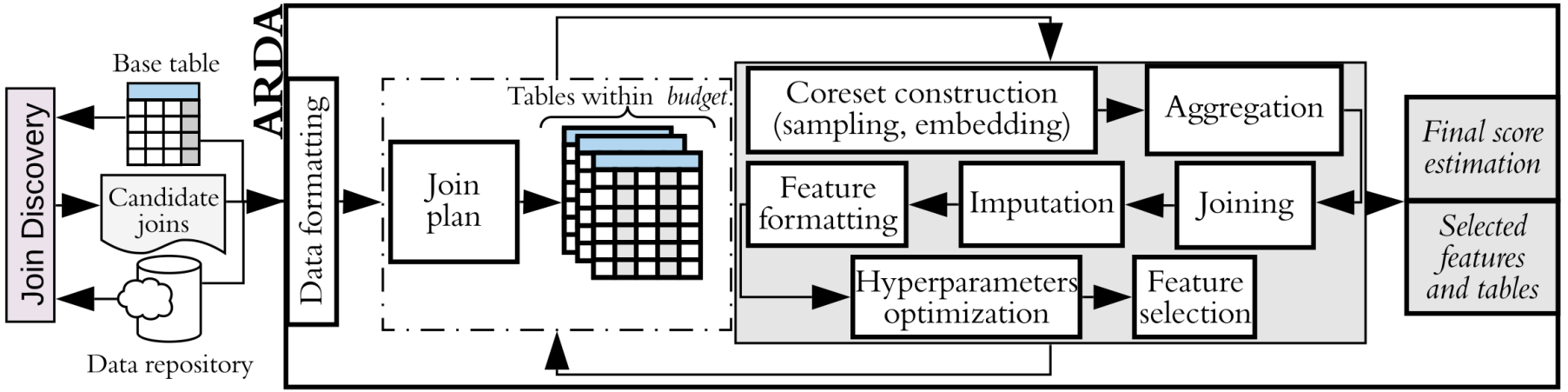
- Input to ARDA: a reference to a database and a collection of candidate joins from a data discovery system: a description of the columns in the base table that can be used as foreign keys into other tables.
- Often, data discovery systems provide a ranking of the candidate joins based on expected relevancy (usually determined by simple heuristics).



Workflow of ARDA, adapted from Chepurko, Nadiia, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. "ARDA: automatic relational data augmentation for machine learning." In *Proceedings of the VLDB Endowment,* 13(9), 1373–1387, 2020

Feature Acquisition: ARDA

1) Input to ARDA: a reference to a database and a collection of candidate joins
2) Coreset construction: using sampling or matrix sketching
3) Join plan & Join execution
4) Feature Selection: considering whether a particular join is a useful augmentation
5) Final estimate

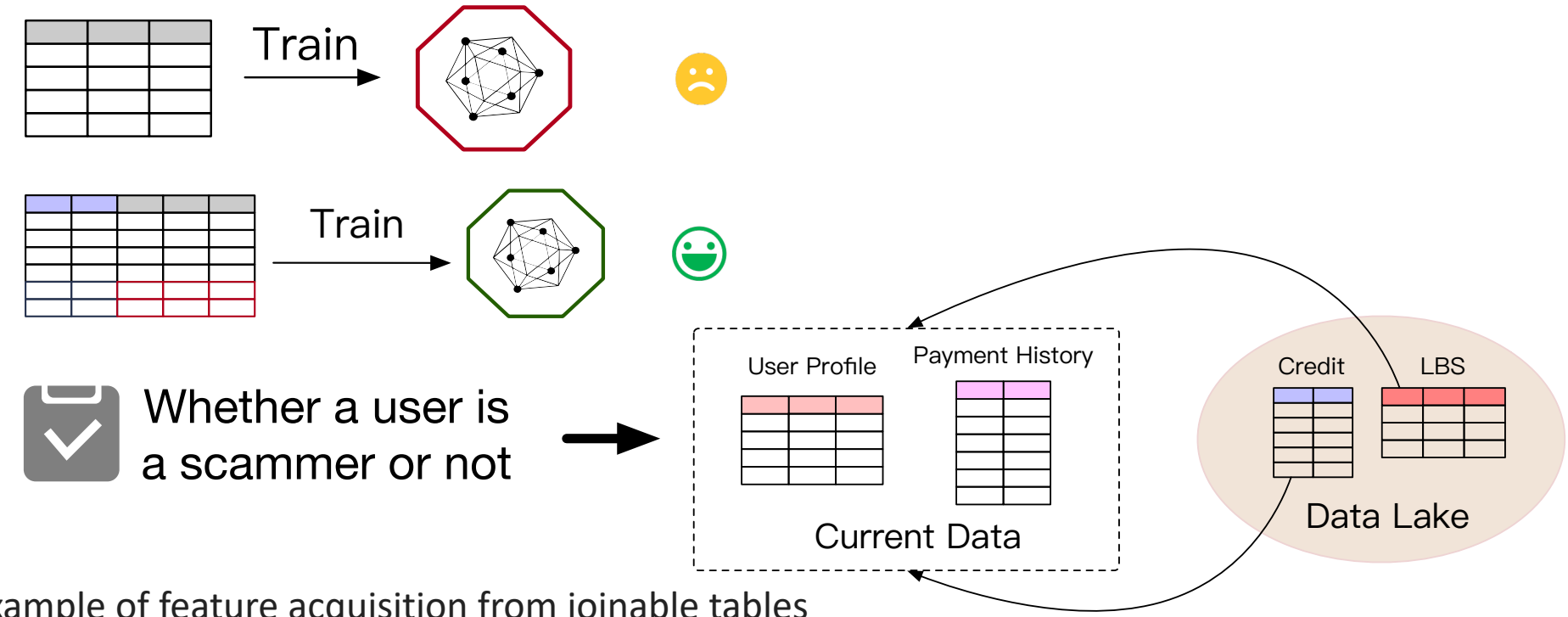


Workflow of ARDA

Chepurko, Nadiia, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. "ARDA: automatic relational data augmentation for machine learning." In *Proc. of the VLDB Endow.,* 13(9), 1373–1387, 2020

Feature Acquisition: AutoFeature

Feature Augmentation with Reinforcement Learning [ICDE 2022]

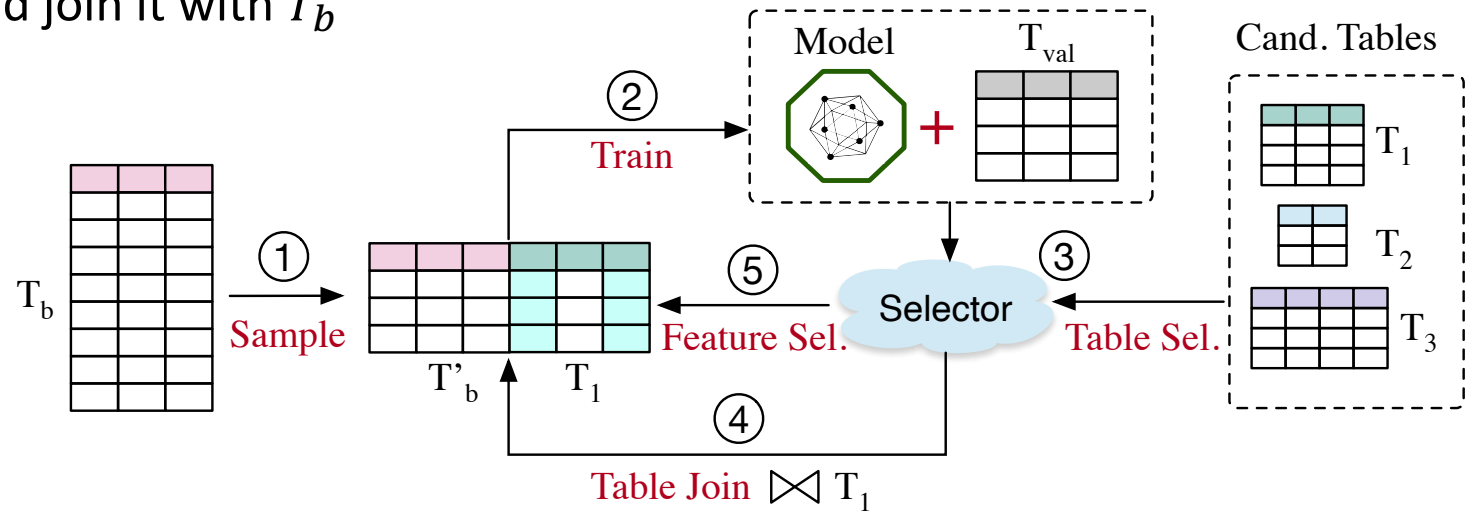


An example of feature acquisition from joinable tables

Liu, Jiabin, Chengliang Chai, Yuyu Luo, Yin Lou, Jianhua Feng, and Nan Tang. "Feature augmentation with reinforcement learning." In *ICDE*, pp. 3360-3372, 2022.

Feature Acquisition: AutoFeature

- Iterative framework:
  - Sample data from T_b
  - Train M and test the performance
  - Select a candidate table and join it with T_b
  - Choose features



The structure of feature selector

Liu, Jiabin, Chengliang Chai, Yuyu Luo, Yin Lou, Jianhua Feng, and Nan Tang. "Feature augmentation with reinforcement learning." In *ICDE*, pp. 3360-3372, 2022.

Feature Acquisition

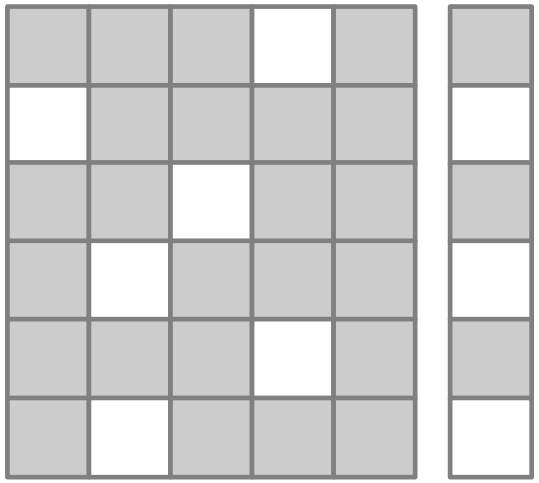
- Kumar, Arun, Jeffrey Naughton, Jignesh M. Patel, and Xiaojin Zhu. "To join or not to join? thinking twice about joins before feature selection." In SIGMOD, pp. 19-34. 2016.
- Roh, Yuji, Geon Heo, and Steven Euijong Whang. "A survey on data collection for machine learning: a big data-ai integration perspective." IEEE Transactions on Knowledge and Data Engineering 33(4): 1328-1347, 2019.
- R. Battiti. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks, 5(4):537–550, 1994.
- A. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artif. Intell., 97(1-2):245–271, 1997.
- R. Kohavi and G. H. John. Wrappers for feature subset selection. Artif. Intell, 97(1-2):273–324, 1997.

Data Acquisition

- Instance Acquisition
- Feature Acquisition
- **Value Acquisition**

Value Acquisition

- Challenges:
  - Estimation of value utility
  - Computational Complexity
  - Granularity



Value Acquisition

- fill(r, A, v): Fill in an empty column A in row r to have value v.
- upvote(r): Upvote a complete row r .
- downvote(r): Downvote a partial row r.

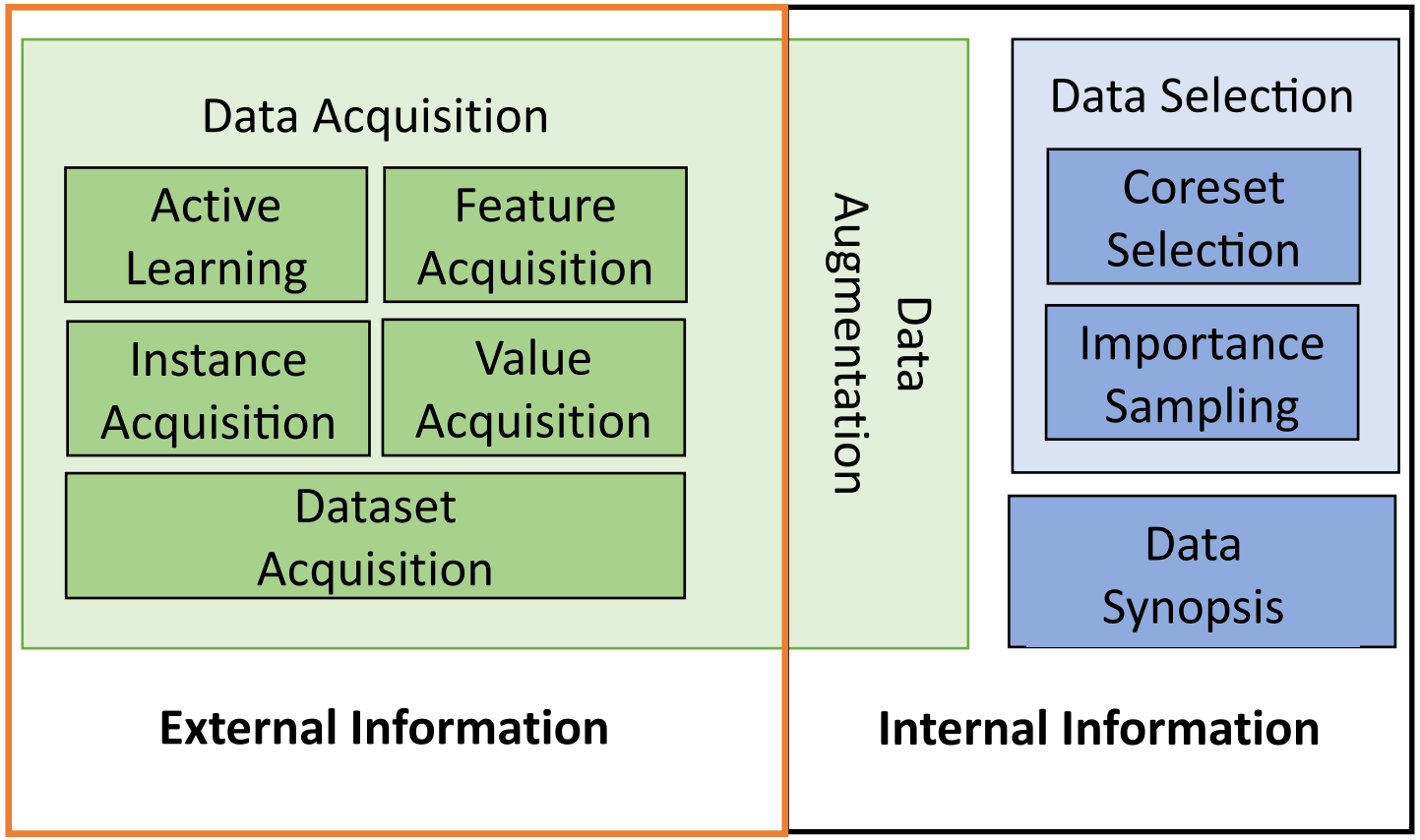
name	nationality	position	caps	goals	↑	↓
Lionel Messi	Argentina	FW	83	37	2	0
Ronaldinho	Brazil	MF	97	33	3	0
Ronaldinho	Brazil	FW	97	33	2	1
Iker Casillas	Spain	GK	150	0	2	0
David Beckham	England	MF	115	17	1	0
Neymar	Brazil	FW			0	1
Zinedine Zidane					0	0
	France	DF			0	0
					0	0
					0	0

Example of CrowdFill, adapted from Park, Hyunjung, and Jennifer Widom. "Crowdfill: collecting structured data from the crowd." In *SIGMOD* 2014.

Value Acquisition

- Dan Lizotte, Omid Madani, and Russell Greiner. Budgeted learning of naive-Bayes classifiers. In UAI, 2003.
- Aloak Kapoor and Russell Greiner. Learning and classifying under hard budgets. In ECML, pages 170–181, 2005.
- Prem Melville, Foster J. Provost, and Raymond J. Mooney. An expected utility approach to active feature-value acquisition. In ICDM, pages 745–748, 2005.
- Prem Melville, Maytal Saar-Tsechansky, Foster J. Provost, and Raymond J. Mooney. Active feature-value acquisition for classifier induction. In ICDM, pages 483–486, 2004.
- Z. Zheng and B. Padmanabhan. On active learning for data acquisition. In ICDM, 2002.
- Maytal Saar-Tsechansky, Prem Melville, and Foster J. Provost. Active feature-value acquisition. Management Science, 55(4):664–684, 2009.

Other Related Areas



Mapping Relevant Concepts

Relevant concepts:

- Data Augmentation
- Coreset
- Importance-sampling for ML
- Crowdsourcing
- Data marketplace

References

- Gershtein, Shay, Tova Milo, Slava Novgorodov, and Kathy Razmadze. "Classifier Construction Under Budget Constraints." In SIGMOD, pp. 1160-1174. 2022.
- Deshpande, Amol, Carlos Guestrin, Samuel R. Madden, Joseph M. Hellerstein, and Wei Hong. "Model-driven data acquisition in sensor networks." In VLDB, pp. 588-599. 2004.
- Zhao, Zixuan, and Raul Castro Fernandez. "Leva: Boosting Machine Learning Performance with Relational Embedding Data Augmentation." In SIGMOD, 2022.
- Roh, Yuji, Geon Heo, and Steven Euijong Whang. "A survey on data collection for machine learning: a big data-ai integration perspective." IEEE Transactions on Knowledge and Data Engineering 33.4 (2019): 1328-1347.
- Johnson, Tyler B., and Carlos Guestrin. "Training deep models faster with robust, approximate importance sampling." Advances in Neural Information Processing Systems 31 (2018).

References

- Lomasky, Rachel, et al. "Active class selection." In ECML, 2007.
- Katharopoulos, Angelos, and François Fleuret. "Not all samples are created equal: Deep learning with importance sampling.” In ICML, 2018.
- Wang, Jiayi, et al. "Coresets over multiple tables for feature-rich and data-efficient machine learning." Proceedings of the VLDB Endowment 16.1 (2022): 64-76.
- Coleman, Cody, et al. "Selection via proxy: Efficient data selection for deep learning." arXiv preprint arXiv:1906.11829 (2019).
- Li, Guoliang, et al. "Crowdsourced data management: Overview and challenges.” In SIGMOD 2017.
- Li, Kaiyu, et al. "Crowdrl: An end-to-end reinforcement learning framework for data labelling." In ICDE 2021.
- Cormode, Graham, et al. "Synopses for massive data: Samples, histograms, wavelets, sketches." Foundations and Trends® in Databases 4.1–3 (2011): 1-294.

References

- Li, Feifei, et al. "Wander join: Online aggregation via random walks." In SIGMOD 2016.
- Phillips, Jeff M. "Coresets and sketches." Handbook of discrete and computational geometry. Chapman and Hall/CRC, 2017. 1269-1288.
- Li, Kaiyu, and Guoliang Li. "Approximate query processing: What is new and where to go? A survey on approximate query processing." Data Science and Engineering 3 (2018): 379-397.
- Shan, Caihua, et al. "A crowdsourcing framework for collecting tabular data." IEEE Transactions on Knowledge and Data Engineering 32.11 (2019): 2060-2074.