DS3Lab@
ETH ZURICH

@DS3Lab
www.DS3Lab.com

Towards Understanding End-to-end Learning in the Context of Data:

Machine Learning Dancing over Semirings and Codd's Table

Ce Zhang

ce.zhang@inf.ethz.ch

Systems@ETH Zürich

Rapid Progress of ML Systems

DS3Lab@ ETH ZURICH

Training becomes faster

Training becomes more automatic/systematic

Azure Machine Learning | Cloud AutoML Vision | Amazon SageMaker

Apache SystemDS™ | Cerebro | Angel

snorkel | ZeroER | mlflow™

ModelDB | NorthStar | ...

Systems@ETH Zürich

Rapid Progress of ML Systems

DS3Lab@ ETH ZURICH

Training becomes faster

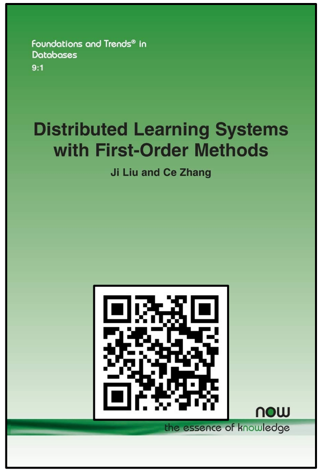
Training becomes more automatic/systematic

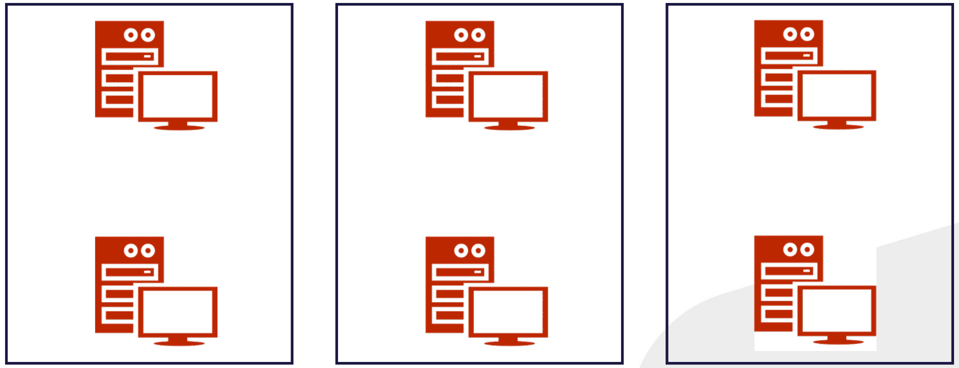
Azure Machine Learning · Cloud AutoML Vision · Amazon SageMaker

Apache SystemDS™ · Cerebro · Angel

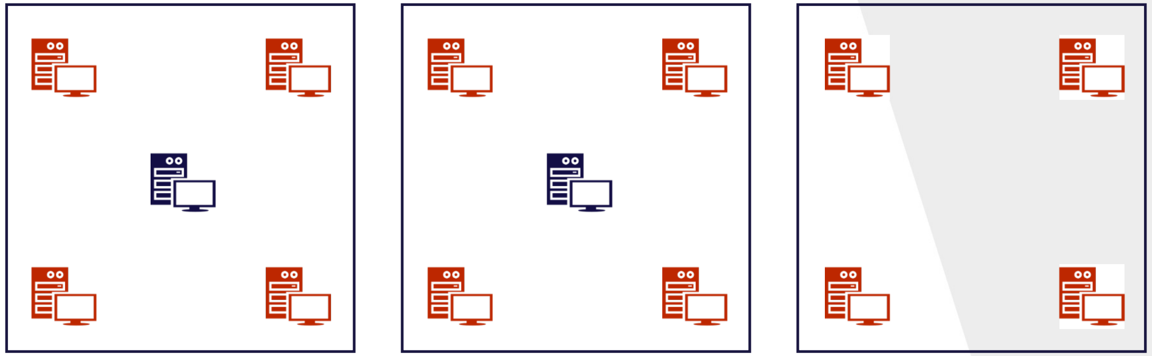
snorkel · ZeroER · mlflow™

ModelDB · NorthStar · ...





Lossless · Sparsification · Quantization



Centralized · Asynchronous · Decentralized

+ Co-design algorithms, systems and hardware accelerators (FPGAs) to enable deep integration of ML into DB, Spark, Serverless, MPI

Rapid Progress of ML Systems

DS3Lab@ ETH ZURICH

Training becomes faster

Training becomes more automatic/systematic

Azure Machine Learning | Cloud AutoML Vision | Amazon SageMaker

Apache SystemDS™ | Cerebro | Angel

snorkel | ZeroER | mlflow™

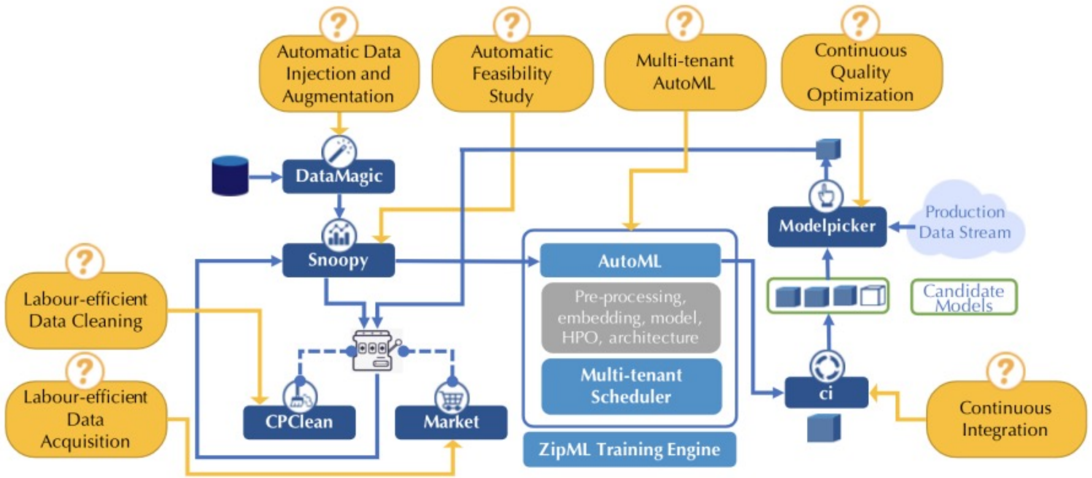
ModelDB | NorthStar | ...

Distributed ML in DB/Serverless/Spark/MPI? Check out

http://zip.ml

MLOps & MLDev made easy? Check out

http://ease.ml



+ streamline and manage the end-to-end process of MLDev and MLOps: feasibility study, CI/CD, active model selection and adaptation, data debugging…

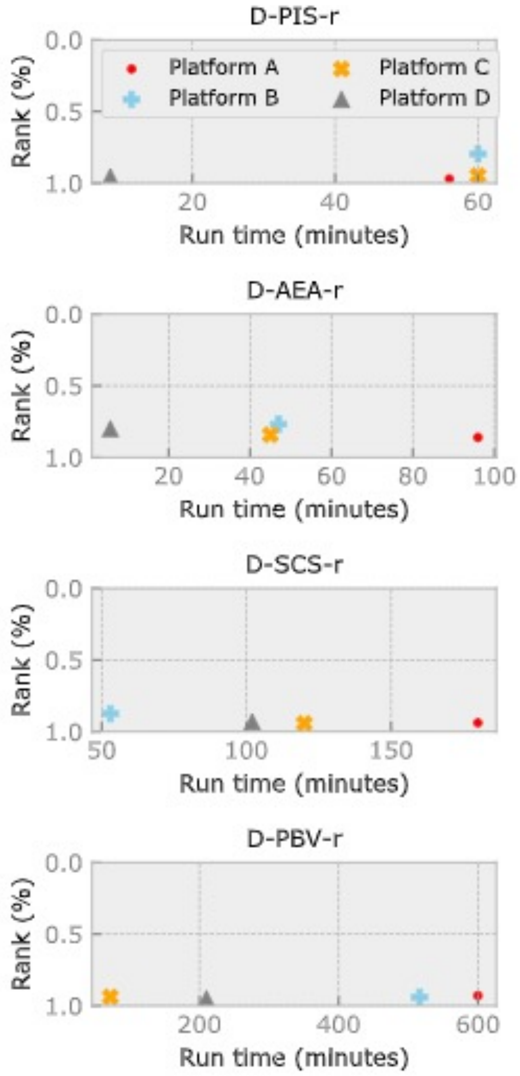
While getting *some* ML models has never been easier,

Understanding has never been more important

DS3Lab@
ETH ZURICH

ML today is now a Data Problem

- Given the raw features from Kaggle, most AutoML platforms rank in the bottom 50%.
- It is the data that we need to improve, and knowledge that we need to integrate, to build better ML applications.
- ***To improve data, we need to first understand them.***



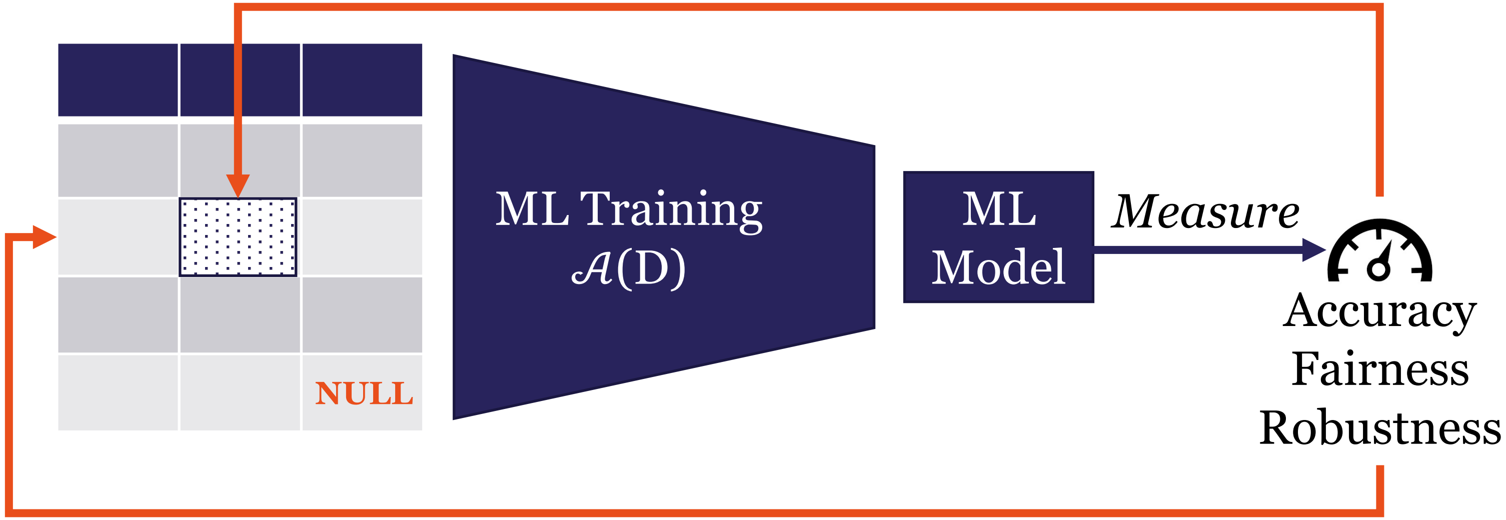
MLBench

VLDB (2018)
http://www.vldb.org/pvldb/vol11/p1220-liu.pdf

Updated version (2021):
https://arxiv.org/abs/1707.09562

Two Examples of Understanding

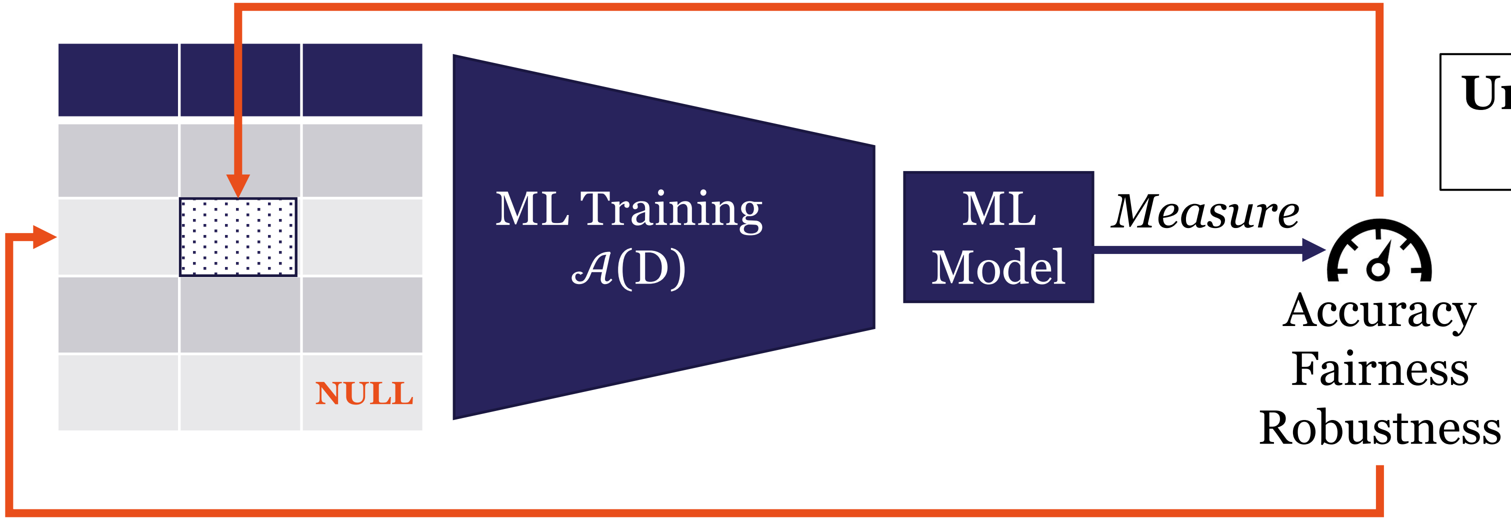
__Q1__. Which data problem should I fix first to improve my accuracy/fairness/robustness?



__Q2__. Which training example is to blame or is most important for my accuracy/fairness/robustness?

Two Examples of Understanding

Q1. Which data problem should I fix first to improve my accuracy/fairness/robustness?



Q2. Which training example is to blame or is most important for my accuracy/fairness/robustness?

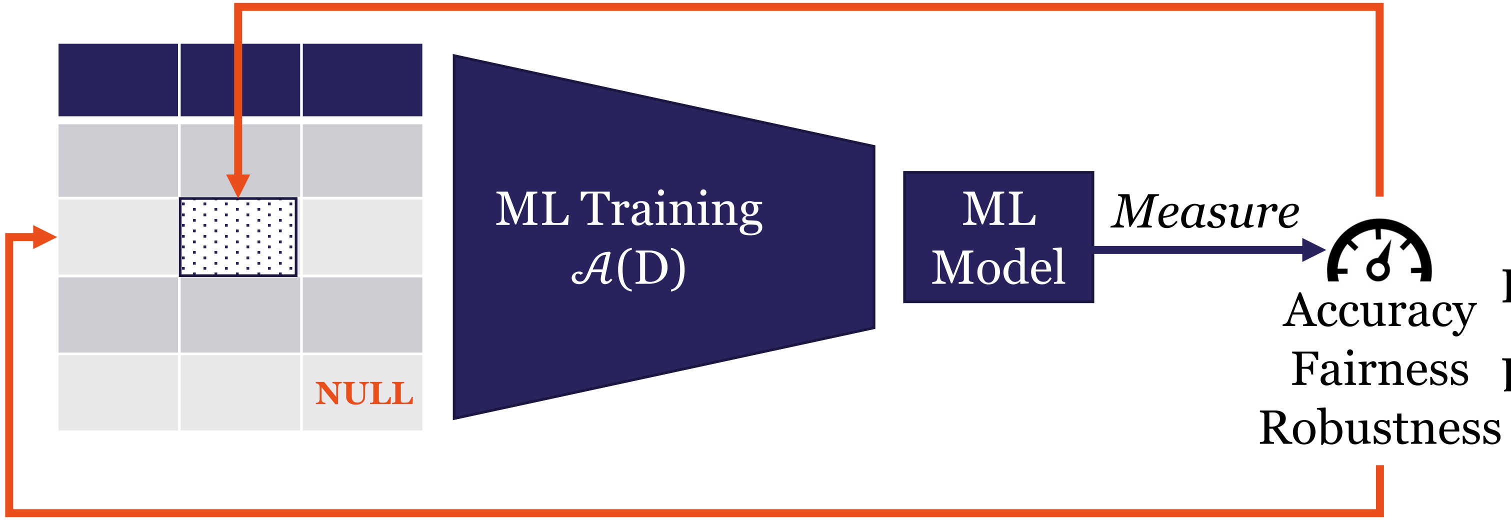
Understanding these two questions has a profound impact on ML DevOps.

- Data *cleaning* for ML
- Data *debugging* for ML
- Data *acquisition* for ML
- Data *valuation* for ML

Two Examples of Understanding

DS3Lab@ ETH ZURICH

__Q1__. Which data problem should I fix first to improve my accuracy/fairness/robustness?



__Q2__. Which training example is to blame or is most important for my accuracy/fairness/robustness?

Understanding these two questions has a profound impact on the ML DevOps.

ActiveClean

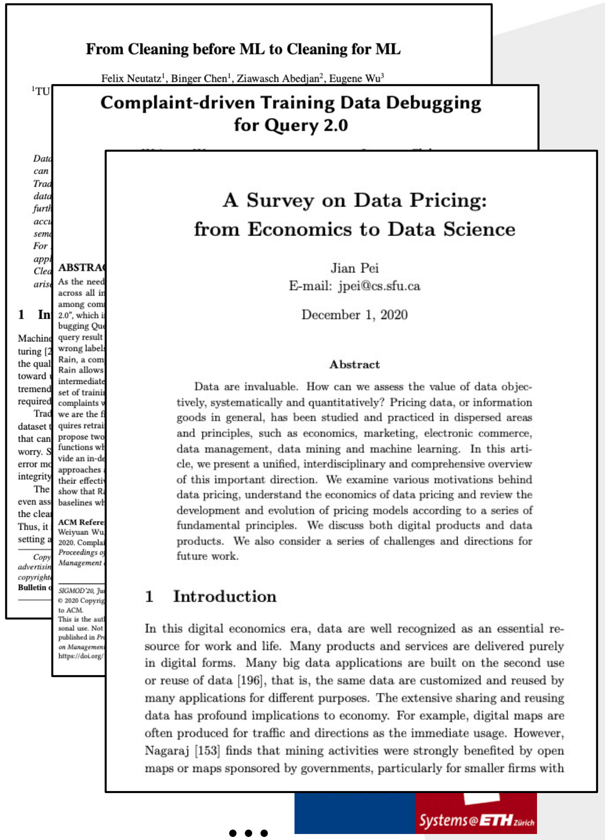
BoostClean

AlphaClean

Influence Func.

Data Shapley

…



…

So, what's the *problem*?

Two Examples of Understanding

DS3Lab@ ETH ZURICH

Real-world ML applications don't often look like this

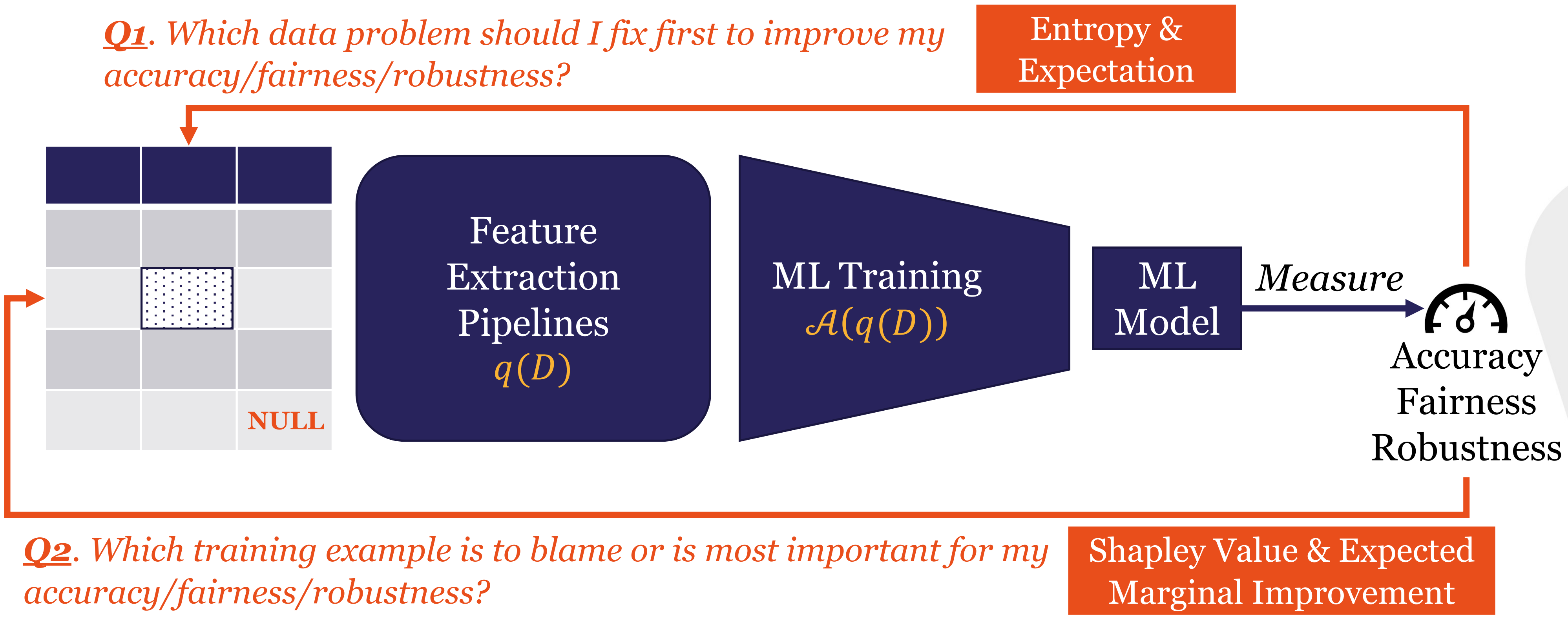
They look something like this

These data transformations really make things hard!

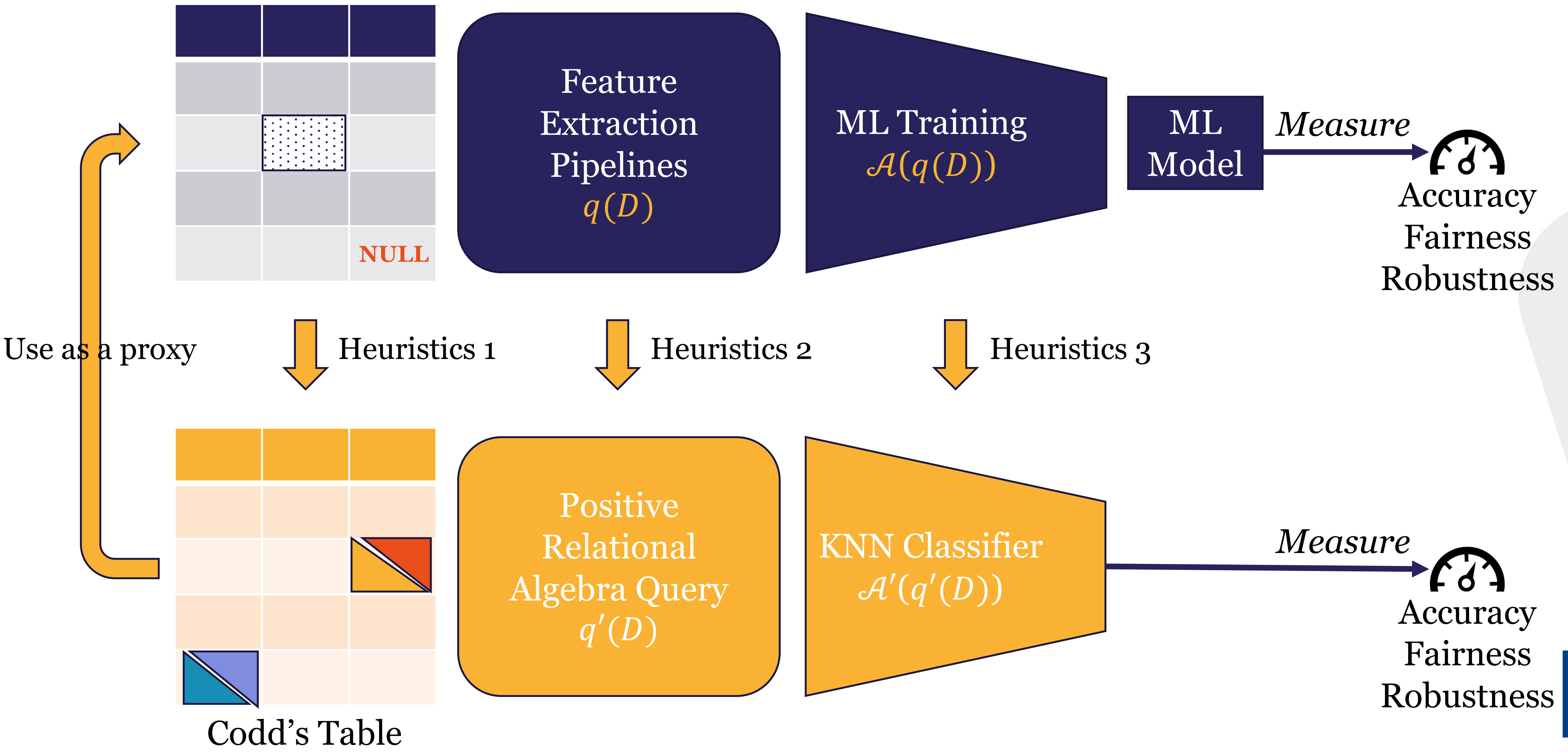
If the DB community don't fix this, who else are going to deal with data transformations?

ETH Systems@ETH Zürich

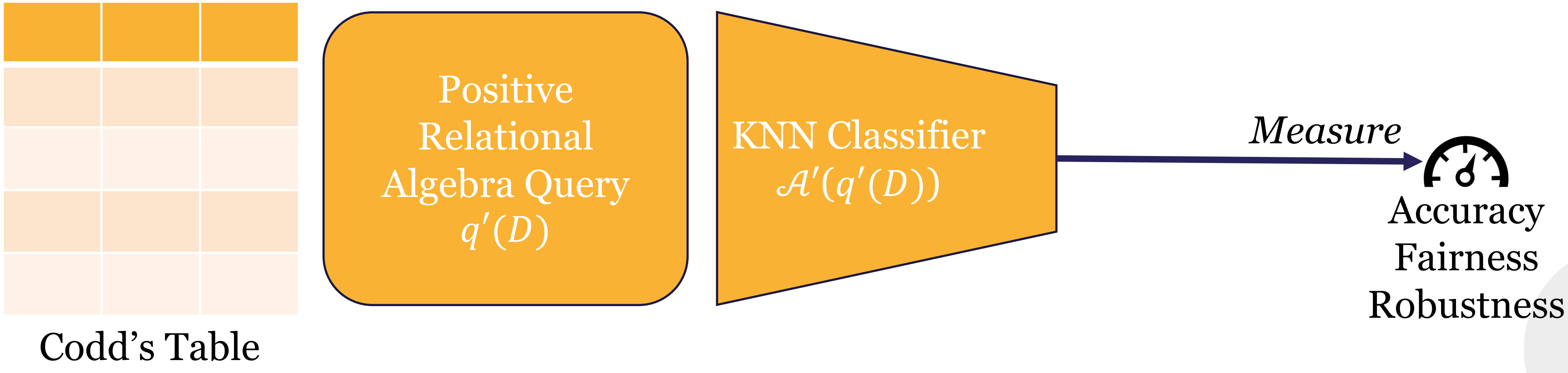
Our Goal


Q1. Which data problem should I fix first to improve my accuracy/fairness/robustness?
Entropy & Expectation
NULL
Feature Extraction Pipelines
$q(D)$
ML Training
$\mathcal{A}(q(D))$
ML Model
Measure
Accuracy
Fairness
Robustness
Q2. Which training example is to blame or is most important for my accuracy/fairness/robustness?
Shapley Value & Expected Marginal Improvement

Our Approach



Takeaways



Takeaway 1

For a dominating number of realistic pipelines, for their proxy pipeline, we can compute exactly the Entropy, Expected Prediction, and Shapley Value, in PTIME.

Takeaway 2

These fundemental quantities, i.e., Entropy, Expected Prediction, Shapley Value, provide principled frameworks for many applications: Data *cleaning, debugging, valuation* for ML, *Defense* against backdoor adversarial attacks

Takeaway 3

Use these proxy pipelines as a proxy work well in many scenarios.

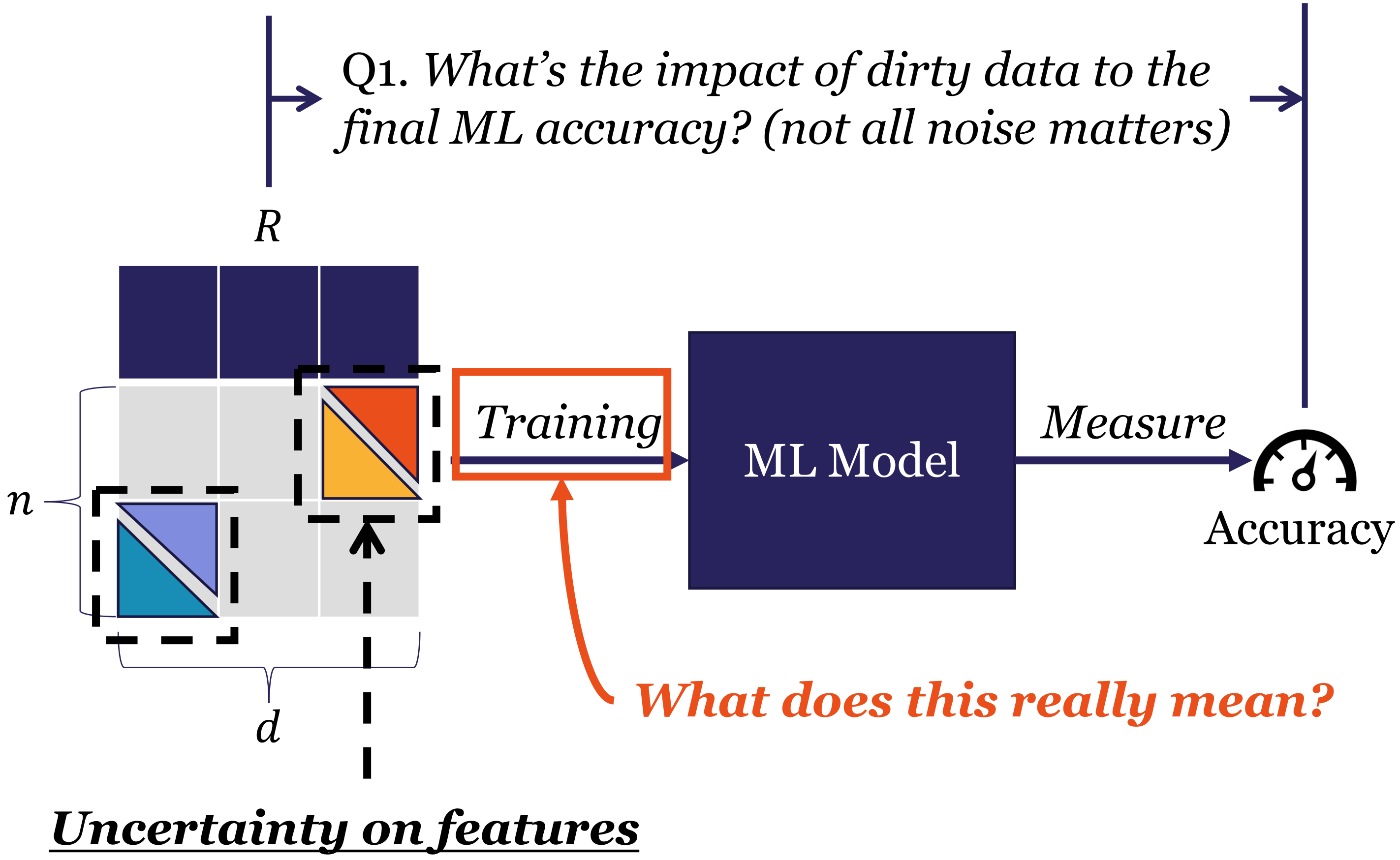
I will also show you cases when they don't.

Overview

- *Theoretical Result*
  - *Entropy/Expectation* of ML Training over Incomplete Information and Uncertainty
  - *Shapley value* of ML Training over Feature Extraction Pipelines
- *Applications*
  - Data *Cleaning* for ML
  - *Defense* against backdoor adversarial attacks
  - Data *Debugging* for ML
  - Data *Market* for ML
- *Heuristics of Approximating Real-world Pipelines*
- *Failure Cases & "Call for Help"*

DS3Lab@ ETH ZURICH

Entropy: Setting



Unknowns (Incomplete information) in DB

DS3Lab@
ETH ZURICH

- *What is Unknown in DB?*
- *How does DB answer queries under Unknowns?*
- *… the nightmare of many undergrads in DB class…*

pid	name	age
1	John	32
2	Anna	29
3	Kevin	NULL
4	NULL	30

```
SELECT *
FROM Person
WHERE age < 30;
```

pid	name	age
2	Anna	29

Unknown

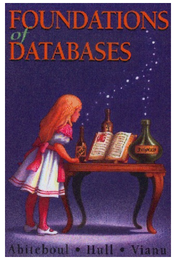
```
Train ML MODEL
```

????

Unknowns (Incomplete information) in DB

DS3Lab@ ETH ZURICH

- What is NULL?
  - *Sure/Certain Answer*



Codd Table

pid	name	age
1	John	32
2	Anna	29
3	Kevin	NULL

```
SELECT *
FROM Person
WHERE age < 30;
```

pid	name	age
2	Anna	29

"Sure answer"

Possible Worlds

pid	name	age
1	John	32
2	Anna	29
3	Kevin	1

```
SELECT *
FROM Person
WHERE age < 30;
```

pid	name	age
2	Anna	29
3	Kevin	1

pid	name	age
1	John	32
2	Anna	29
3	Kevin	2

```
SELECT *
FROM Person
WHERE age < 30;
```

pid	name	age
2	Anna	29
3	Kevin	2

pid	name	age
1	John	32
2	Anna	29
3	Kevin	30

```
SELECT *
FROM Person
WHERE age < 30;
```

pid	name	age
2	Anna	29

…

If you take this view, there is nothing stop us from applying the same semantics to ML

ML over Unknowns: Certain Prediction (CP)

DS3Lab@ ETH ZURICH

Relation with NULL

pid	name	age
1	John	32
2	Anna	29
3	Kevin	{1,2,30}

1. Train ML model
2. Predict on a new tuple t

Possible Worlds without NULL

pid	name	age
1	John	32
2	Anna	29
3	Kevin	1

pid	name	age
1	John	32
2	Anna	29
3	Kevin	2

pid	name	age
1	John	32
2	Anna	29
3	Kevin	30

Each possible world:

1. Train ML model
2. Predict on a new tuple t

Result on relation with NULL		World 1	World 2	World 3
YES	*"Sure answer"*	YES	YES	YES
NO	*"Sure answer"*	NO	NO	NO
NULL	*"Sure answer"*	YES	NO	NO
33% Yes	*Expectation*	YES	NO	NO
0.915	*Entropy*	YES	NO	NO

ETH Systems@ETH Zürich

ML over Unknowns: Certain Prediction (CP)

DS3Lab@ ETH ZURICH

- Given an input relation R with a known noise model …
- … it defines a probability distribution over many possible worlds $\{R_1 \ldots R_n\}$
- Given a test example x, the *expected prediction* on x is

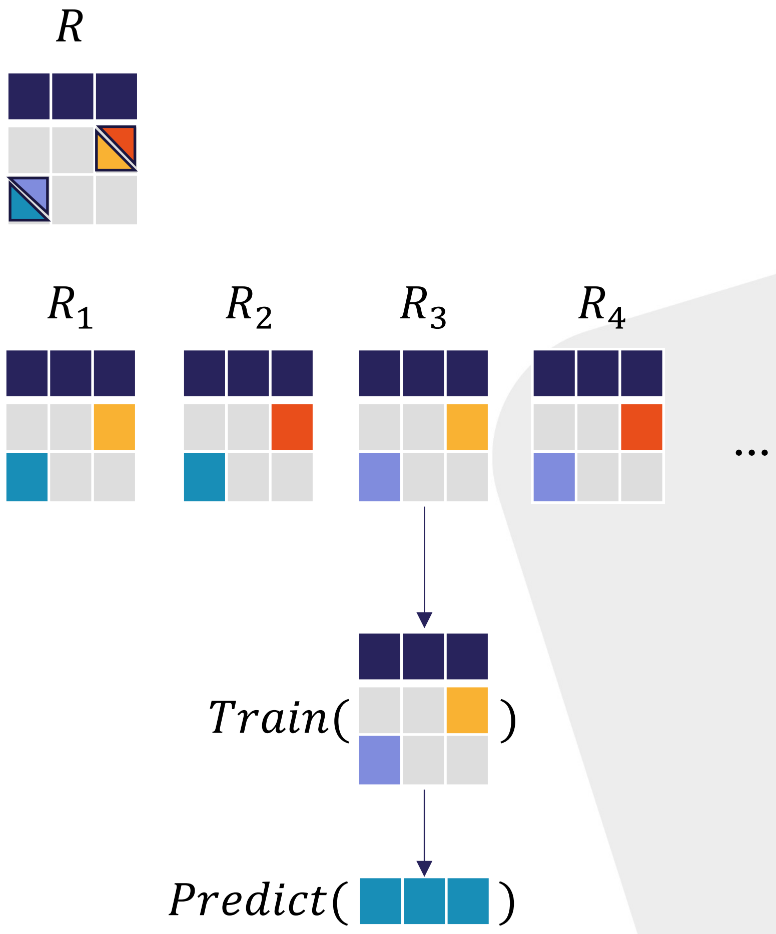
$$CP(x) = \mathbb{E}\mathcal{A}(R)[x]$$

where $\mathcal{A}(R)$ is a classifier trained on one possible R.

- The *entropy* is:

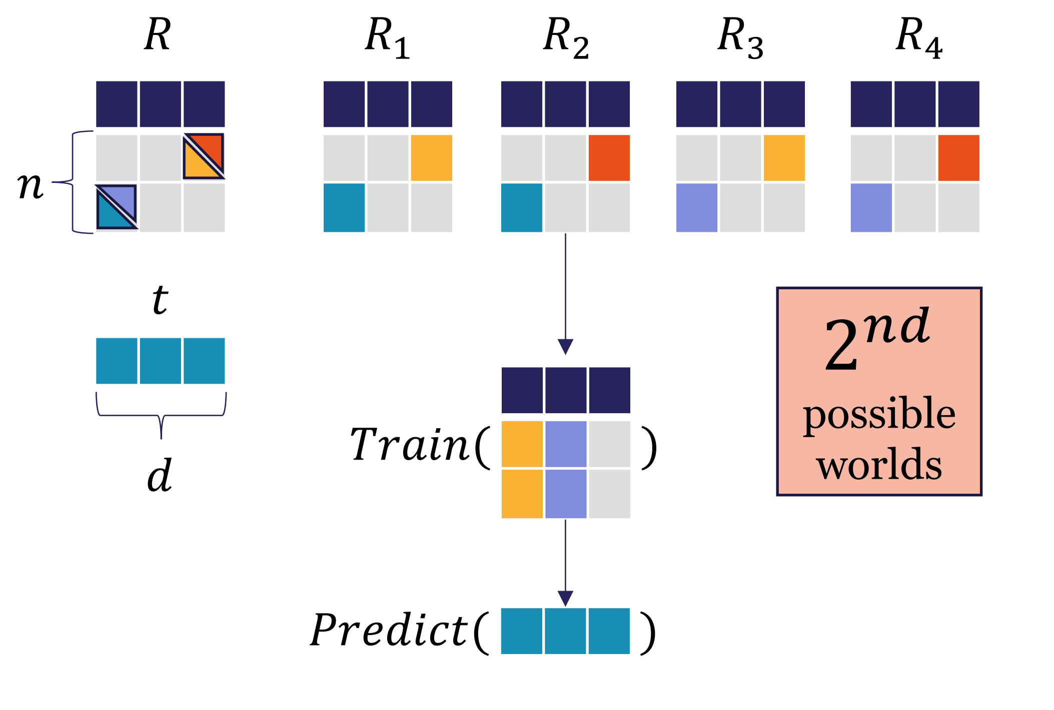
$$H(\mathcal{A}(R)[x])$$

- Intuition of why this is useful:
  - *Why Entropy goes to 0, noises in training set do not matter at all. (It won't change the result)*
  - *If we take this view,* ***data cleaning for ML becomes the process of minimizing Entropy****.*



Challenge and Result

Challenges



Results

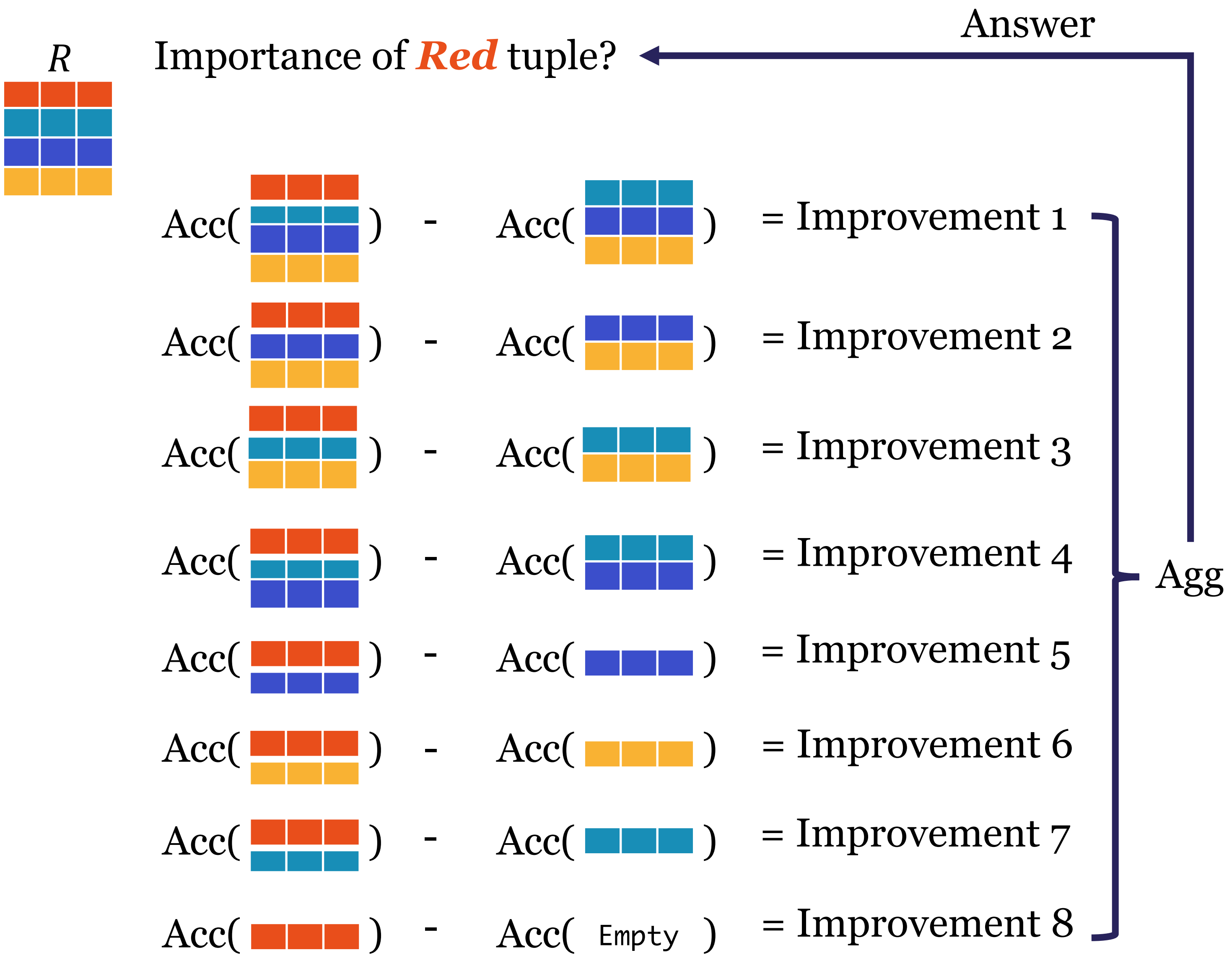
1. **General Classifier**
   - #P-Hard
   - Approximation with MCMC
2. **K-Nearest Neighbor Classifier**
   - $O(nd)$ – Linear in # Unknowns to explore $O(2^{nd})$ many worlds!
   - http://vldb.org/pvldb/vol14/p255-karlas.pdf Trust me ☺
   - Also apply to pipelines with only **map** operators

Shapley Value: Setting

DS3Lab@ ETH ZURICH

Q2. *What's the importance of each training example to the final ML accuracy? (not all input example matters)*

Expected Marginal Improvements

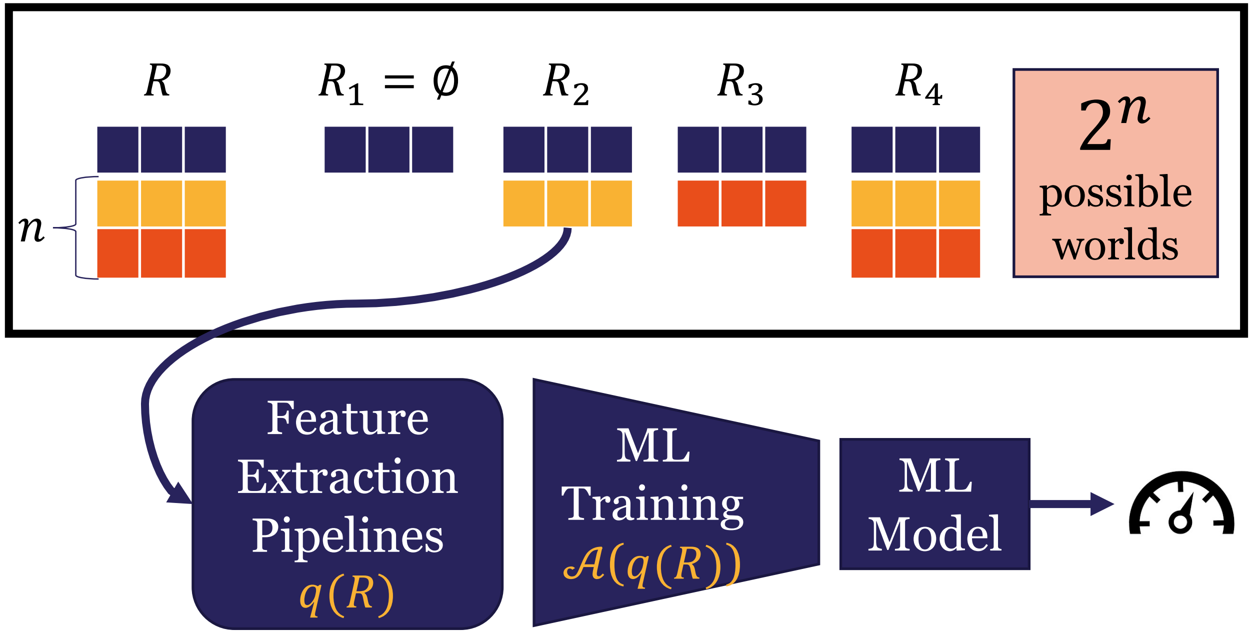


Shapley Value

$$v(t) = \frac{1}{n!}\sum_{R_i \subseteq R\setminus\{t\}} \binom{|R|}{|R_i|}\left[U\left(R_i \bigcup t\right) - U(R_i)\right]$$

- *Sound game-theoretical foundation*
- *Many good properties*
- *Has been shown to work well for data debugging (w/o feature extraction pipelines)*

Challenge



Results

1. **General Classifier**
   - #P-Hard
   - Approximation with MCMC
2. **K-Nearest Neighbor Classifier**

Pipeline Type	1NN	KNN
map	$O(N \log N)$	$O(N^2)$
map + fork	$O(N \log N)$	$O(N^2M^2)$
1-to-many joins	$O(N^4)$	$O(N^4)$
Polymonial with PTIME counting oracle	PTIME	PTIME

- https://ds3lab.inf.ethz.ch/datascope.html Trust me ☺

Goal: Illustrate how locality of KNN can help. Similar ideas apply to both Entropy and Shapley Value.

Data Valuation

All can be generalized; but it makes the analysis more engaged so let's only talk about these simplified assumptions ☺

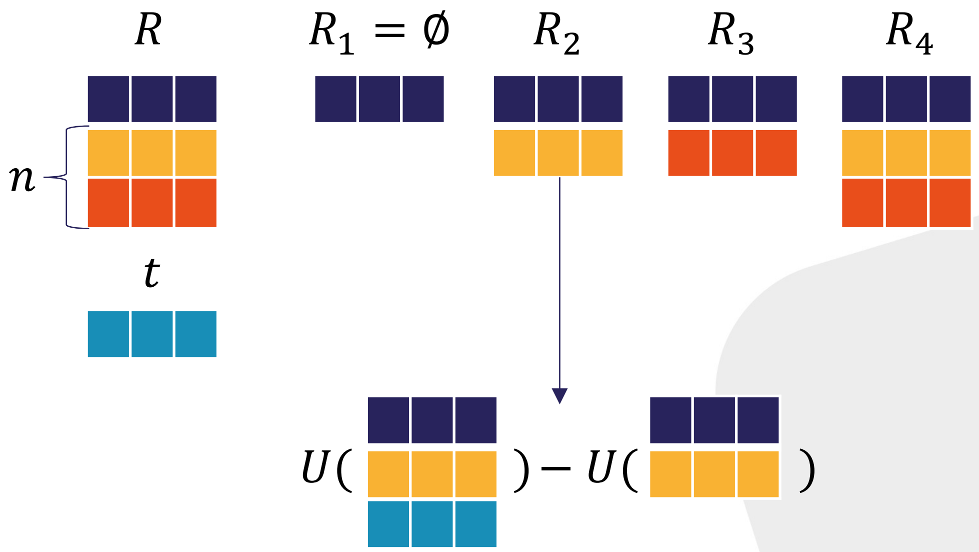
Simplified Version (For Presentation)

- **Input**:
  - Training Relation R, $|R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$, s.y is the label of s
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R,s): r.y = s.y\}|/K$$

- **Output**: the ***value*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t \right) - U(R_i) \right]$$



Goal: $v(t) = \frac{1}{n!} \sum_i \binom{n}{|R_i|} \times [U(R_i \cup t) - U(R_i)]$

Data Valuation: 1. A Trivial $O(\binom{n}{K} + n \log n)$ Algorithm

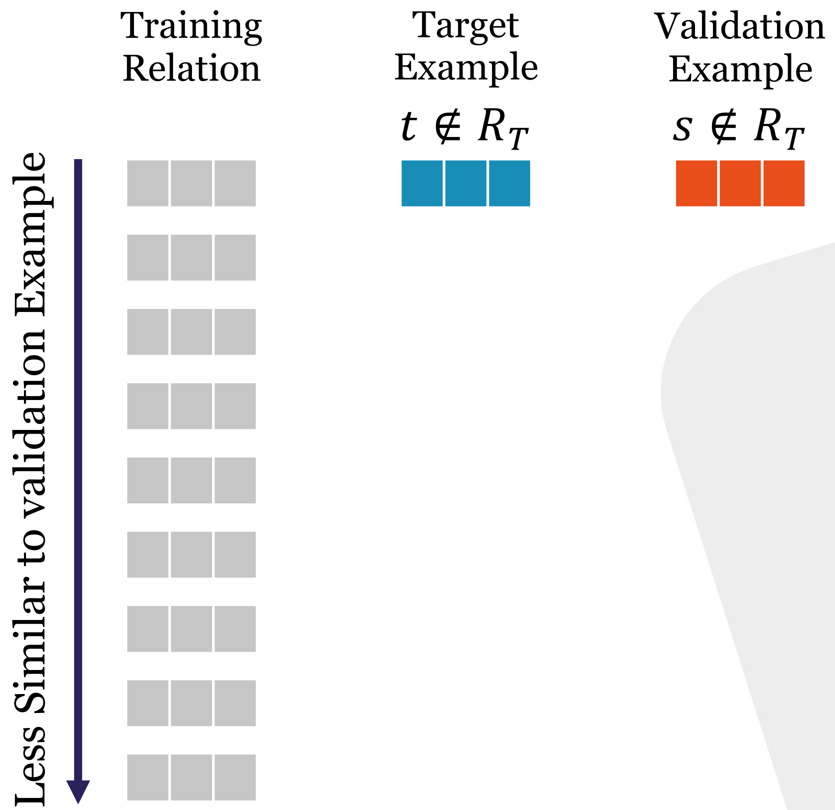
Simplified Version (For Presentation)

- **Input**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R, s): r.y = s.y\}|/K$$

- **Output**: the ***value*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t\right) - U(R_i) \right]$$



Data Valuation: 1. A Trivial $O(\binom{n}{K} + n \log n)$ Algorithm

DS3Lab@ ETH ZURICH

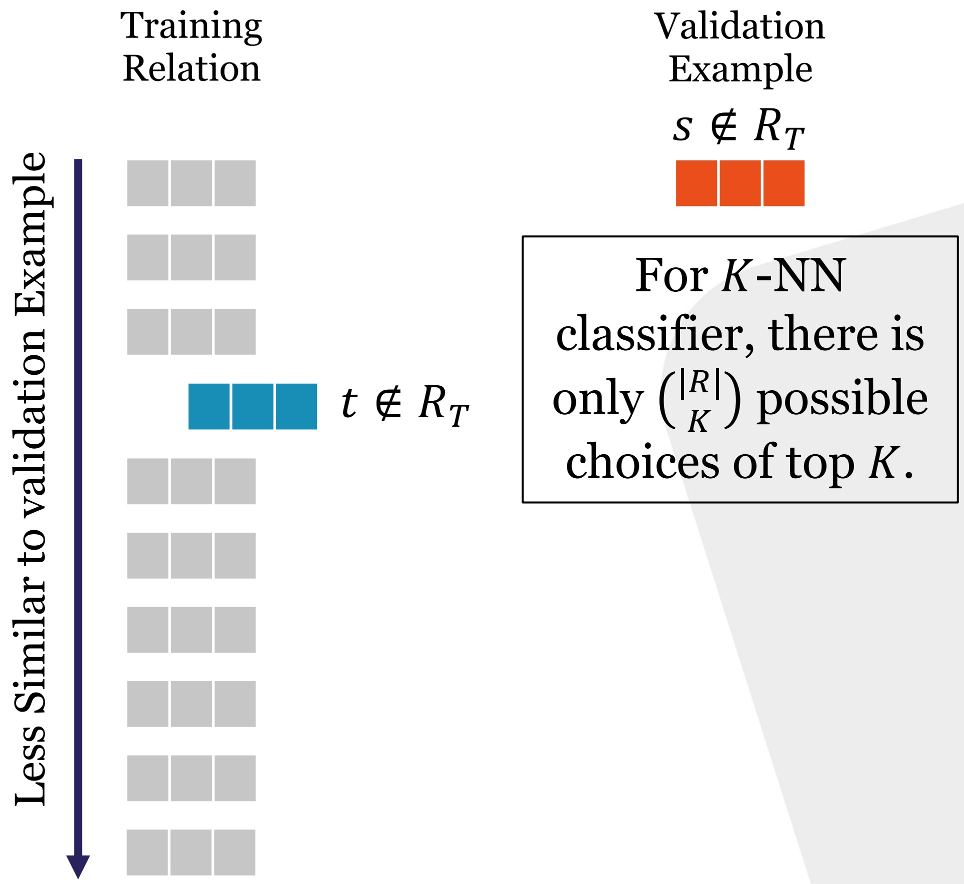
Simplified Version (For Presentation)

- **<u>Input</u>**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R, s): r.y = s.y\}|/K$$

- **<u>Output</u>**: the ***<u>value</u>*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t\right) - U(R_i) \right]$$



ETH Systems@ETH Zürich

Data Valuation: 1. A Trivial $O(\binom{n}{K} + n \log n)$ Algorithm

DS3Lab@ ETH ZURICH

Simplified Version (For Presentation)

- **<u>Input</u>**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R, s) : r.y = s.y\}| / K$$

- **<u>Output</u>**: the ***<u>value</u>*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t\right) - U(R_i) \right]$$

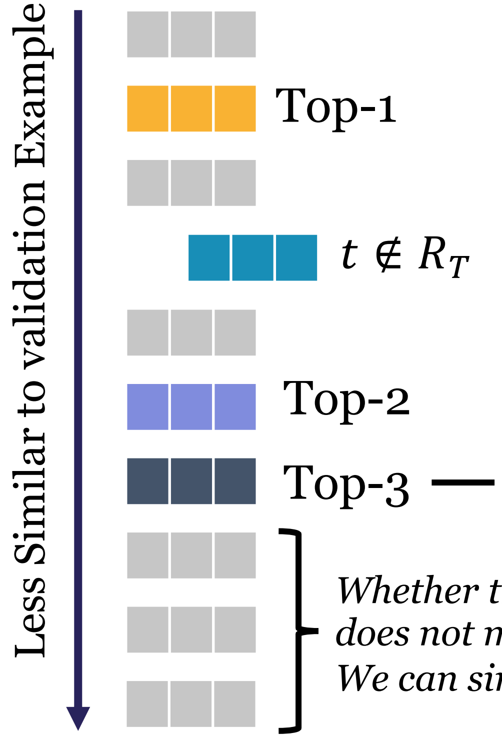
Can we do better?

(Yes, this trivial algorithm does not use all the properties that KNN has)

Training Relation

Validation Example

$s \notin R_T$



For K-NN classifier, there is only $\binom{|R|}{K}$ possible choices of top K.

With t added into R, this tuple will be out of Top-3, Top-1/Top-2 will still be in Top-3

Whether these examples in possible world does not matter – They won't be Top-3: We can simply calculate # worlds.

=> For each of $\binom{|R|}{K}$ possible top-K's, one only needs to compare the label of t and the label of the last in top-K.

ETH Systems@ETH zürich

Data Valuation: 2. An $O(n \log n)$ Algorithm for $K = 1$

DS3Lab@ ETH ZURICH

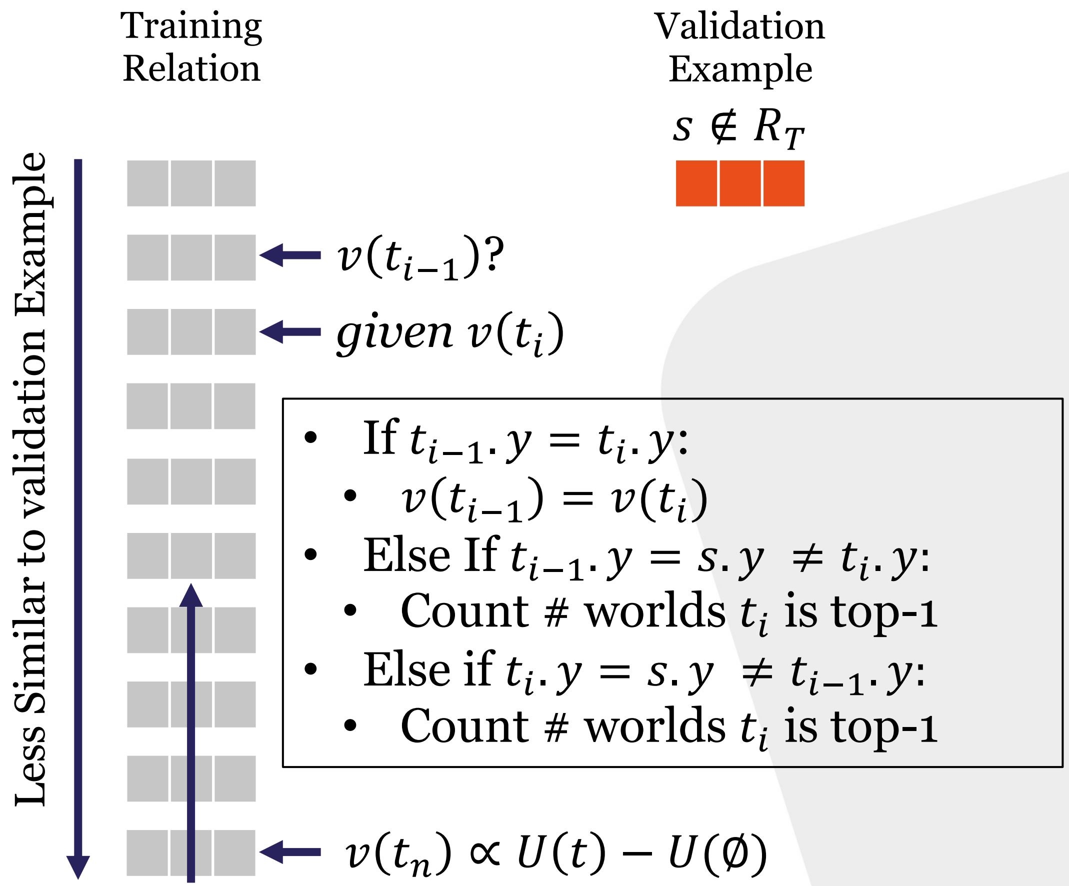
Simplified Version (For Presentation)

- **Input**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R, s) : r.y = s.y\}|/K$$

- **Output**: the ***value*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t \right) - U(R_i) \right]$$



ETH Systems@ETH Zürich

Data Valuation: 3. An $O(n \log n)$ Algorithm

DS3Lab@ ETH ZURICH

Simplified Version (For Presentation)

- **Input**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R,s): r.y = s.y\}|/K$$

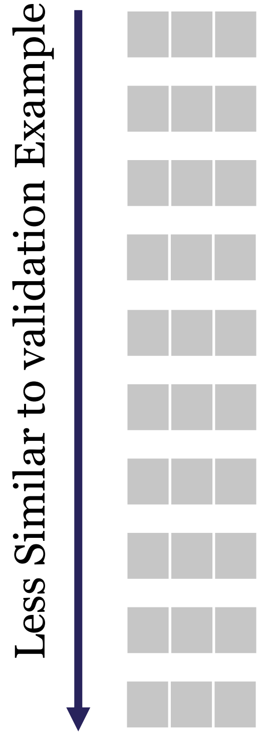
- **Output**: the ***value*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t\right) - U(R_i) \right]$$

(Challenge: how to efficiently, incrementally maintain one combinatorial term.)

$$\sum_{k=0}^{N-2} \frac{1}{\binom{N-2}{k}} \sum_{m=0}^{\min\{K-1,k\}} \binom{i-1}{m}\binom{N-i-1}{k-m} = \sum_{m=0}^{\min\{K-1,i-1\}} \sum_{k'=0}^{N-i-1} \frac{\binom{i-1}{m}\binom{N-i-1}{k'}}{\binom{N-2}{m+k'}} = \frac{\min\{K,i\}(N-1)}{i}$$

Training Relation



Validation Example

$s \notin R_T$

$$v(t_n) = \frac{\mathbb{I}[t_n.y = s.y]}{n}$$

$$v(t_i) = v(t_{i+1}) + \frac{\mathbb{I}[t_i.y=s.y] - \mathbb{I}[t_{i+1}.y=s.y]}{K} \frac{\min\{K,i\}}{i}$$

(1) Sort;
(2) Single Pass Scan

ETH Systems@ETH Zürich

A single pass of sorting is still too slow for systems (we have could have billions data points to sort)

Data Valuation: (ε, δ)-Approximation with LSH

DS3Lab@ ETH ZURICH

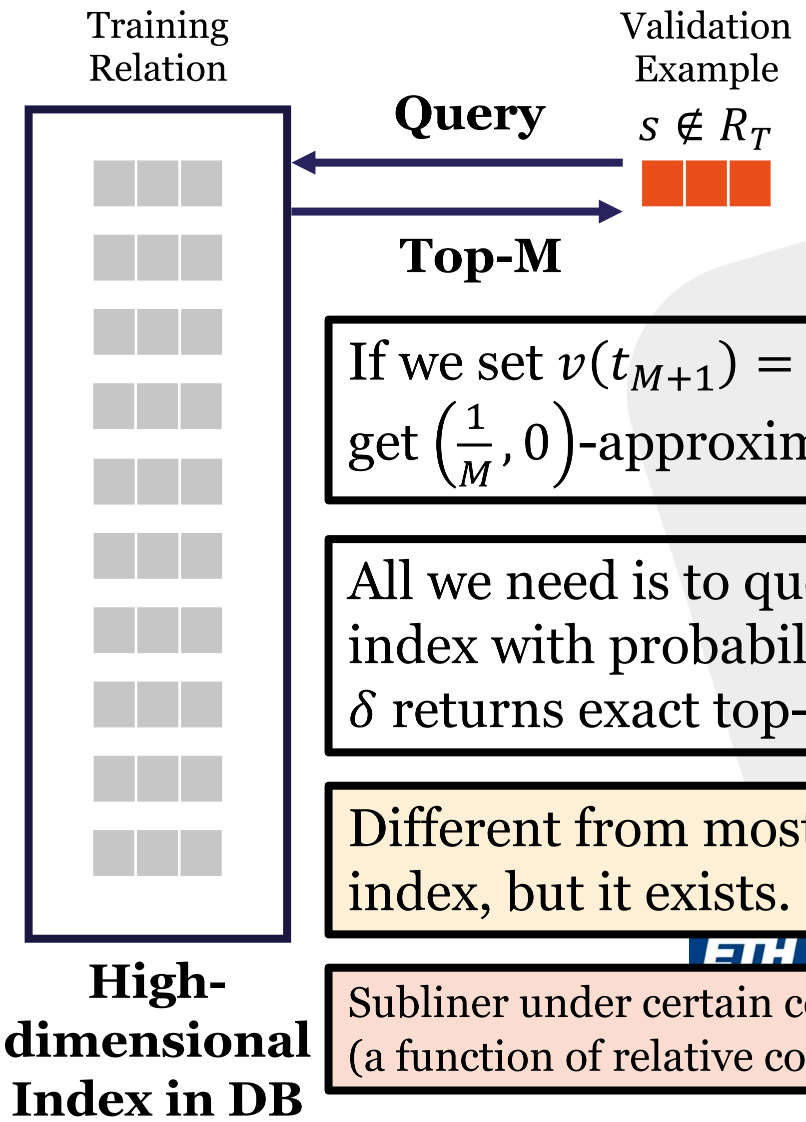
Simplified Version (For Presentation)

- **Input**:
  - Training Relation $R, |R| = n$
  - Target Training Example $t \notin R$
  - Validation Example $s \notin R$
  - Classification task
  - Utility:

$$U(R) = |\{r \in KNN(R, s): r.y = s.y\}|/K$$

- **Output**: the ***value*** of t over R

$$v(t) = \frac{1}{n!} \sum_{R_i \subseteq R_T} \binom{|R|}{|R_i|} \left[U\left(R_i \bigcup t \right) - U(R_i) \right]$$



If we set $v(t_{M+1}) = 0$, we get $\left(\frac{1}{M}, 0\right)$-approximation.

All we need is to query an index with probability $1 - \delta$ returns exact top-M.

Different from most DB index, but it exists.

Subliner under certain conditions (a function of relative contrast)

Enough Theory!
How can we use these?

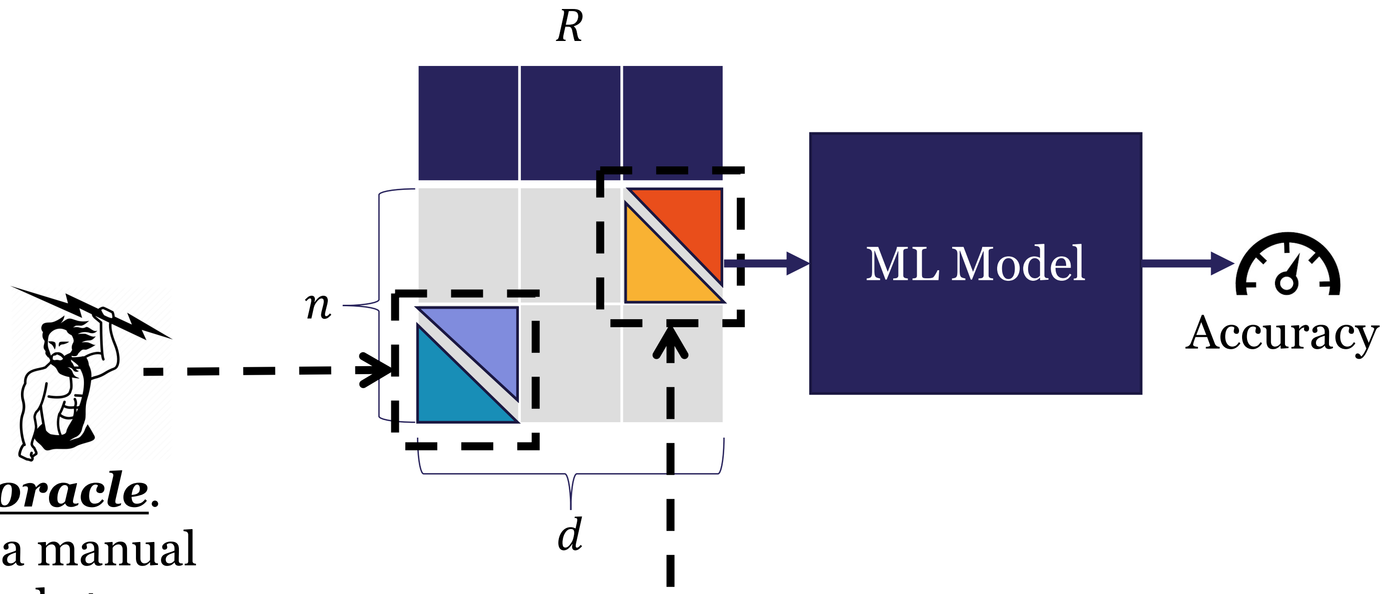
Overview

- *Theoretical Result*
  - *Entropy/Expectation* of ML Training over Incomplete Information and Uncertainty
  - *Shapley value* of ML Training over Feature Extraction Pipelines
- *Applications*
  - Data *Cleaning* for ML
  - *Defense* against backdoor adversarial attacks
  - Data *Debugging* for ML
  - Data *Market*
- *Heuristics of Approximating Real-world Pipelines*
- *Failure Cases & "Call for Help"*

DS3Lab@ ETH ZURICH

ETH Systems@ETH Zürich

Data Cleaning for ML



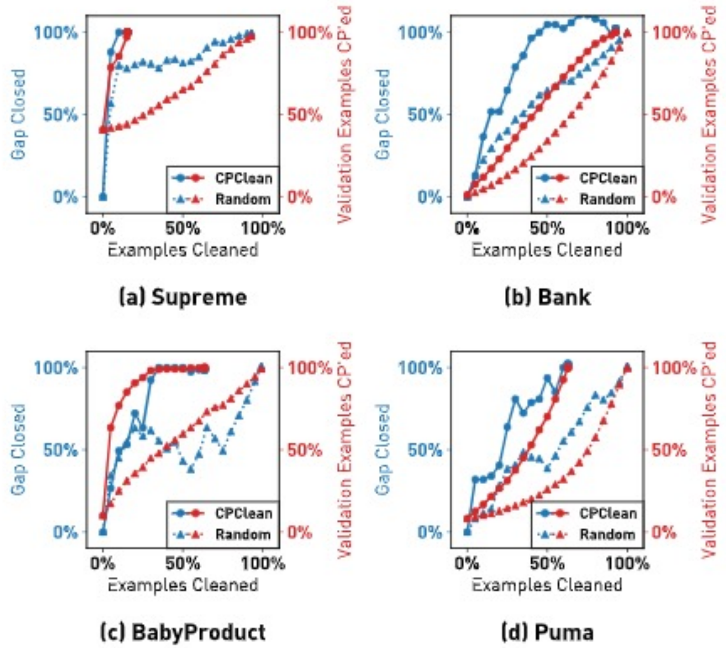
Cleaning oracle.
There exist a manual cleaning oracle to decide the ground truth value for each feature.

Uncertainty on features.
Think about it as outputs of state-of-the-art data cleaning methods – each method output a candidate value.

How should we prioritize which cell to clean?

Idea: Clean examples that minimize the *expected entropy*.

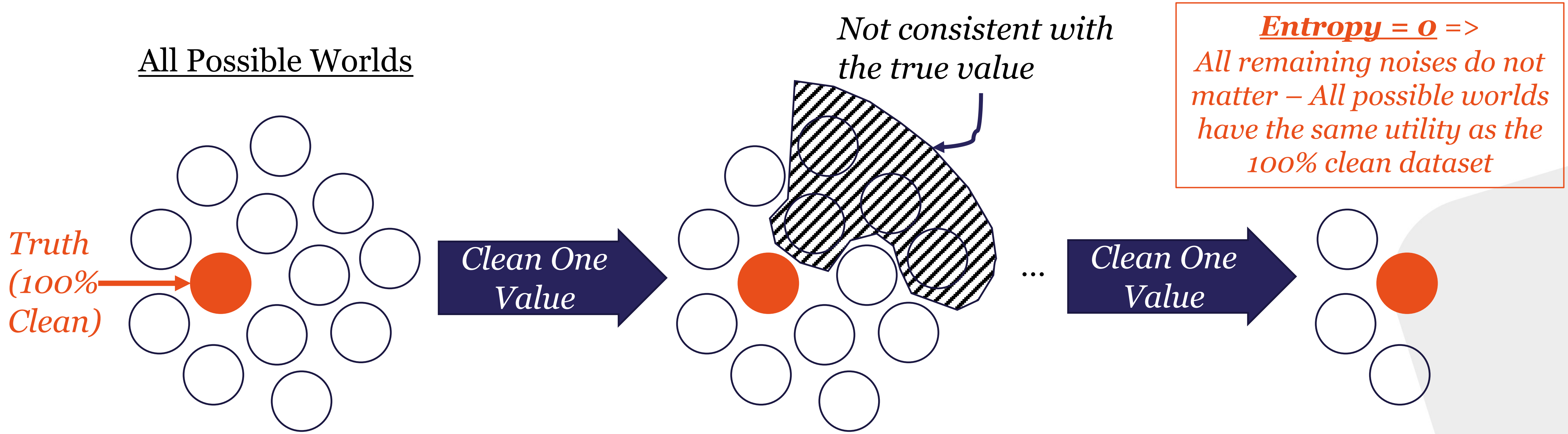
Use the results of decades of research on Sequential Information Maximization for entropy minimization



More results in:
http://vldb.org/pvldb/vol14/p255-karlas.pdf

Data Cleaning for ML - Entropy Minimization

DS3Lab@
ETH ZURICH



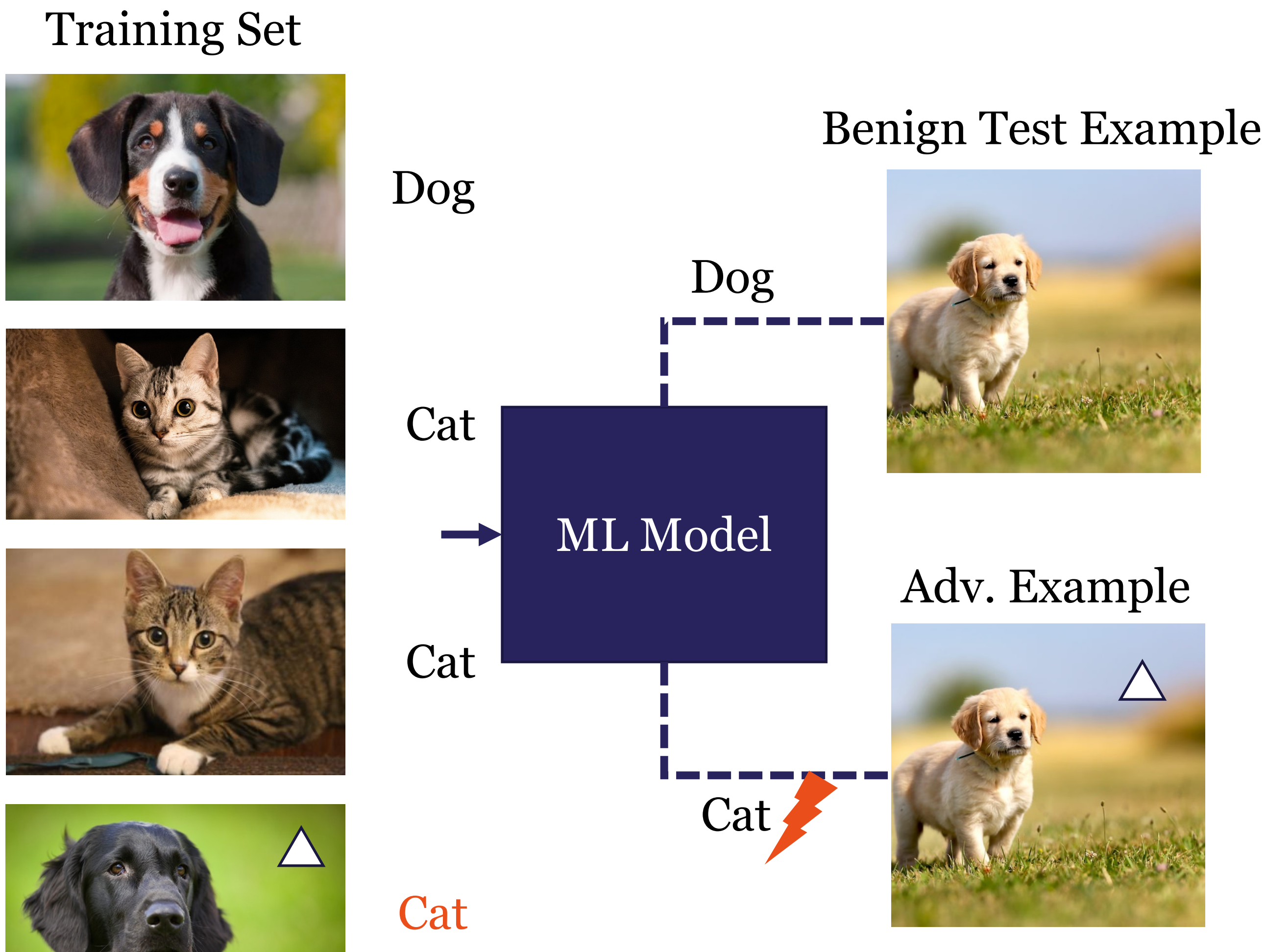
Under this view, Data Cleaning for ML becomes:

How to find cleaning opeations $o_1 \ldots o_n$ such that we decrease final entropy as much as possible?

Under mild technical conditions, we can greedily pick the next cleaning operations o_{i+1} that decreases our expected entropy.

Lot of interesting studies (decades) about Sequential Information Maximization

Defend against backdoor adversarial attacks



Can we make our training process robust to this type of attacks?

Idea: Inject Gaussian noise to the training set and compute the *expection.*

This allows us to provide certifications on robustness, similar how people are doing randomized smoothing for evasion attacks.

To our best knowledge, this gives us the first robustness certification for backdoor attacks https://arxiv.org/abs/2003.08904

Robustness of Expectations

DS3Lab@ ETH ZURICH

Inspired by the seminal work of randmized smoothing, but significantly generalizes it.

Let's look at $g(x) = \mathbb{E}_{z \sim Z} f(x, z)$

What can we say about $g(x)$ and $g(x + \delta)$?

Informal Theorem

If $g(x) \geq p$ and $z \sim Z$ with probability density function μ, then

$$g(x+\delta) \geq \mathbb{P}_{z \sim Z}\left[\frac{\mu(\delta + z)}{\mu(z)} \leq Z^{-1}(1-p)\right]$$

https://arxiv.org/abs/2003.08904

Application 1. Defend about Backdoor

$$g(\boxed{x}) = \mathbb{E}_{z \sim \boxed{Z}} f(x, z)$$

Training Set

Injected Gaussian Noise

Training over a new training set $x + z$, then do inference

Get a certificate looks like: As long as the attacker introduces a perturbations δ with 2-norm smaller than C, the inference result will not change by Δ.

When we have a Deep neural networks, sample some training sets, train a model for each.

When we have a KNN classifier, we can calculate this term exactly in PTIME.

ETH Systems@ETH Zürich

Robustness of Expectations - Quantum Systems

DS3Lab@
ETH ZURICH

Inspired by the seminal work of randmized smoothing, but significantly generalizes it.

Let's look at $g(x) = \mathbb{E}_{z\sim Z} f(x, z)$

What can we say about $g(x)$ and $g(x + \delta)$?

Informal Theorem

If $g(x) \geq p$ and $z \sim Z$ with probability density function μ, then

$$g(x+\delta) \geq \mathbb{P}_{z\sim Z}\left[\frac{\mu(\delta+z)}{\mu(z)} \leq Z^{-1}(1-p)\right]$$

https://arxiv.org/abs/2003.08904

Application 2. Robustness of Quantum Systems

$$g(\boxed{x}) = \mathbb{E}_x \boxed{f(x)}$$

Input states, N qbits

Quantum Circuit

https://www.nature.com/articles/s41534-021-00410-5

Get a certificate looks like: As long as the attacker introduces a perturbated state ρ with $fidelity(\rho, x) > C$, the result of this quantum system will not change by Δ.

We don't need to inject noises as in the previous case. We get robustness "for free" because of the probabilistic semantics of quantum systems.

ETH Systems@ETH Zürich

Robustness of Expectations - PDB & Joint Inference

DS3Lab@ ETH ZURICH

Inspired by the seminal work of randmized smoothing, but significantly generalizes it.

Let's look at $g(x) = \mathbb{E}_{z\sim Z} f(x, z)$

What can we say about $g(x)$ and $g(x+\delta)$?

Informal Theorem

If $g(x) \geq p$ and $z \sim Z$ with probability density function μ, then

$$g(x+\delta) \geq \mathbb{P}_{z\sim Z}\left[\frac{\mu(\delta+z)}{\mu(z)} \leq Z^{-1}(1-p)\right]$$

https://arxiv.org/abs/2003.08904

Application 3. Robustness of PDB & Joint Inference

$$g(\boxed{x}) = \mathbb{E}_x \boxed{q(x)}$$

Output of NNs, or PDB

Probabilistic DB queries: e.g., SQL
Probabilistic Inference queries: MLN, CRF

https://arxiv.org/abs/2106.06235
https://arxiv.org/abs/2003.00120

Get a certificate looks like: As long as the attacker introduces a perturbations δ on x with 2-norm smaller than C, the result of this quantum system will not change by Δ.

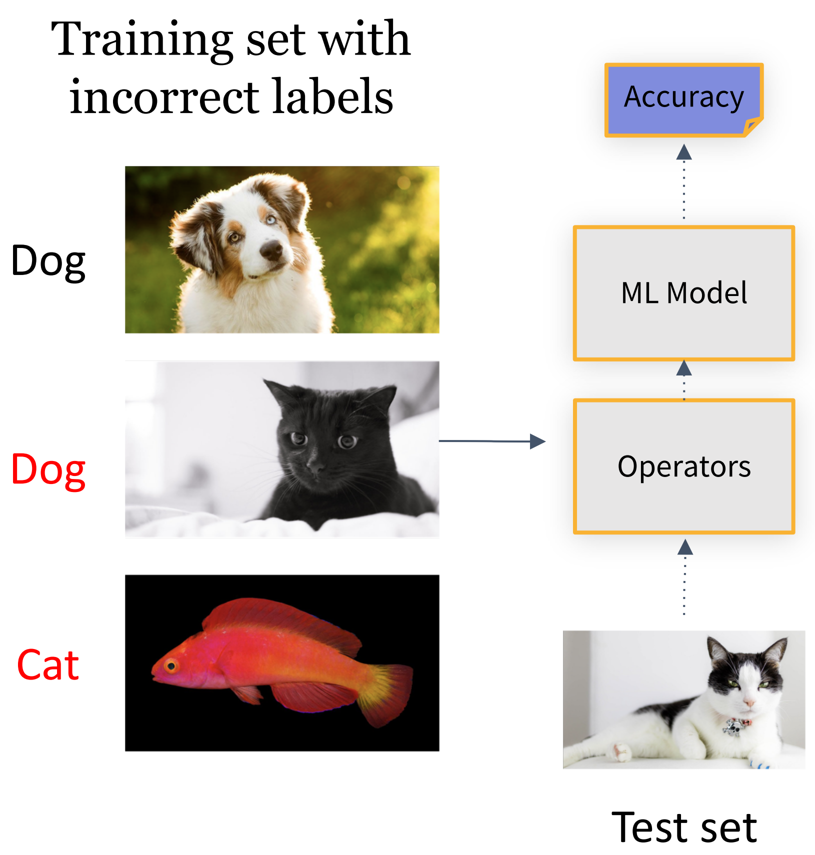
We don't need to inject noises as in the previous case. We get robustness "for free" because of the probabilistic semantics of these probabilistic inference systems.

ETH Systems@ETH Zürich

Robustness & data cleaning are all naturally connected to ***Entropy*** and ***Expectations***. An ***efficient proxy*** for computing these could go a long way.

Data Debugging

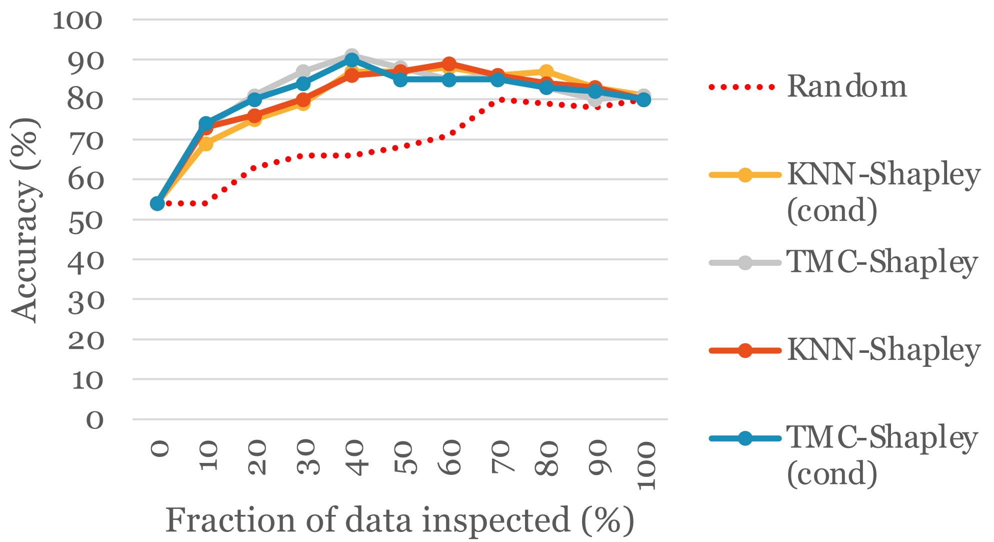
DS3Lab@ ETH ZURICH



How can we find out those training examples with wrong labels?

Idea: Data examples with incorrect labels should have a small (often negative) Shapley value.

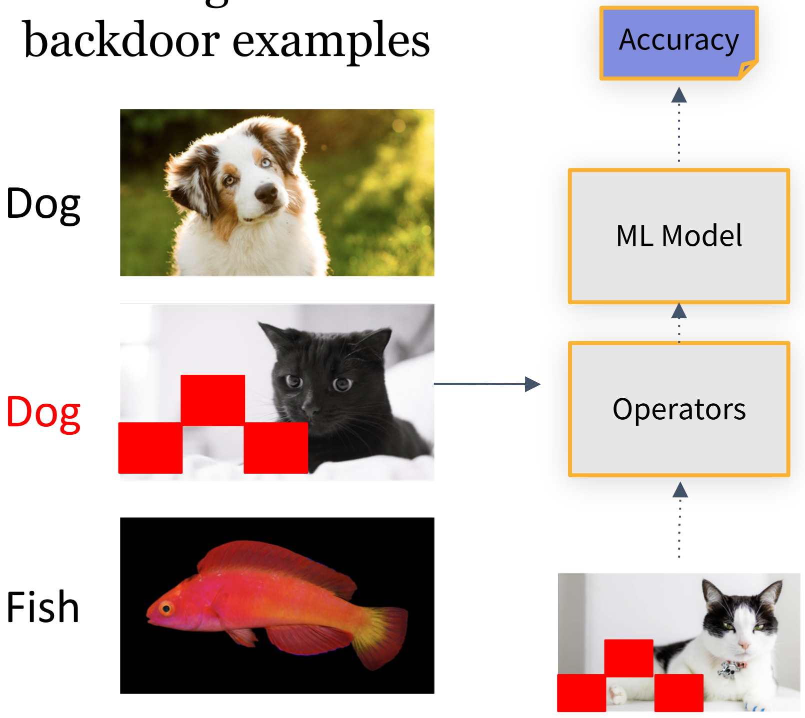
O(N log N) if we use KNN and if a pipeline can be conditioned as map-fork.
Orders of magnitude faster than MCMC: < 1 second for reasonable dataset.



Data Debugging

DS3Lab@
ETH ZURICH

Training set with backdoor examples

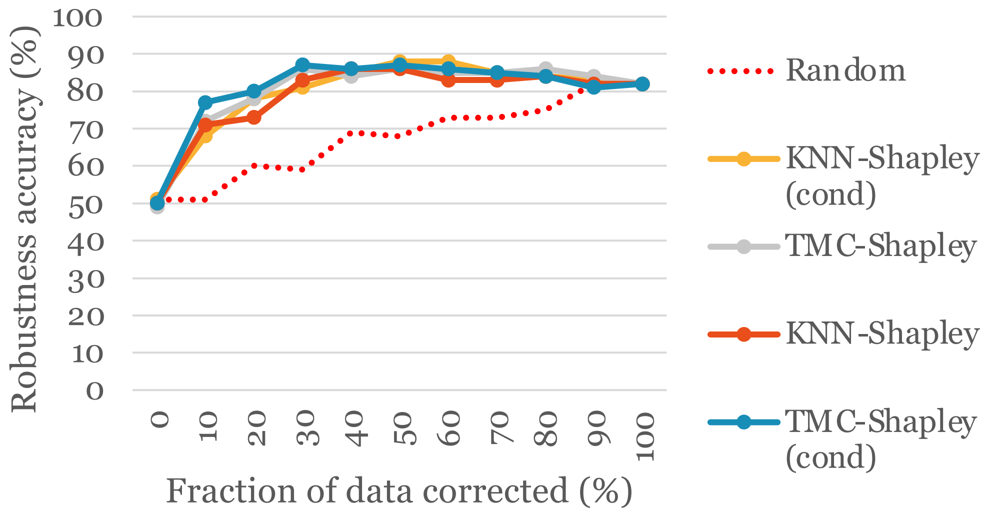


Test set

How can we find out those training examples with backdoor patterns?

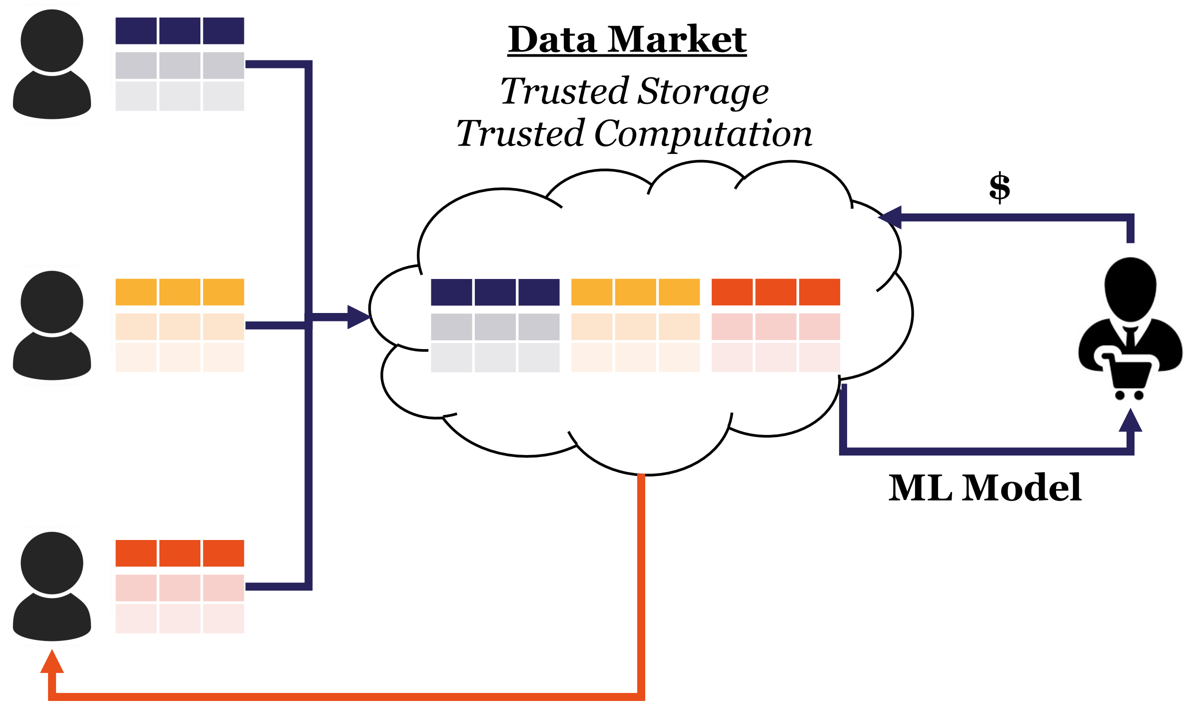
<u>Idea</u>: Data examples with backdoor labels should have a small (often negative) Shapley value.

O(N log N) if we use KNN and if a pipeline can be conditioned as map-fork.
Orders of magnitude faster than MCMC: < 1 second for reasonable dataset.



Expected marginal improvement / Shapley value provides a principled framework for data debugging. An ***efficient proxy*** for computing these could go a long way.

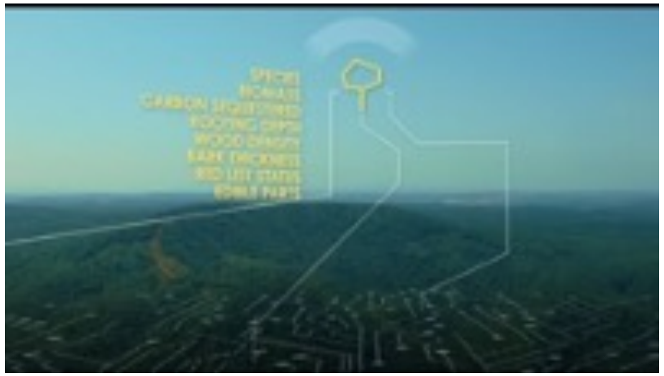
Data Market

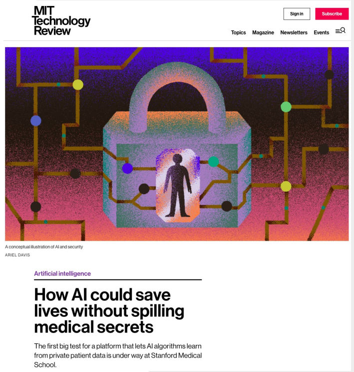


How should we fairly distribute $ to each data contributor?

Idea: Use Shapley value (which is actually why we originally looked at this)

O(N log N) algorithm if we use KNN and if a pipeline can be conditioned as map-fork.




MIT Technology Review

Artificial intelligence

How AI could save lives without spilling medical secrets

The first big test for a platform that lets AI algorithms learn from private patient data is under way at Stanford Medical School.

https://kara.cloud

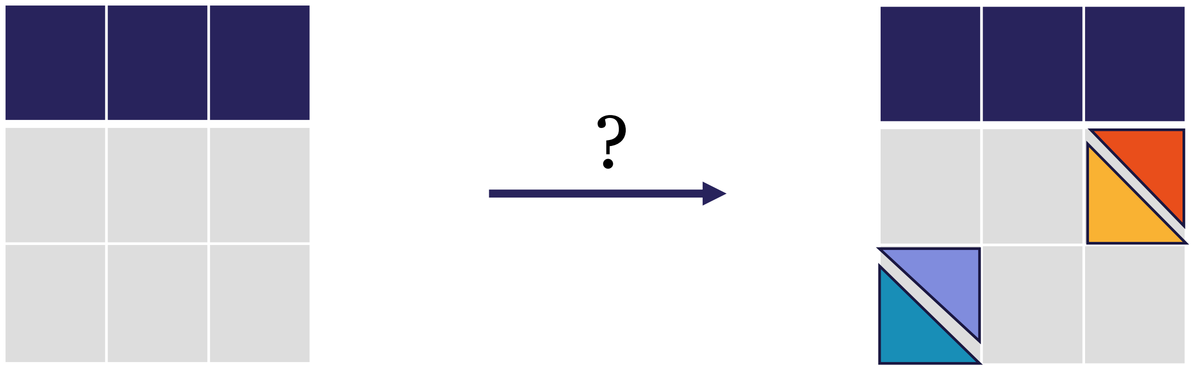
Overview

- *Theoretical Result*
  - *Entropy/Expectation* of ML Training over Incomplete Information and Uncertainty
  - *Shapley value* of ML Training over Feature Extraction Pipelines
- *Applications*
  - Data *Cleaning* for ML
  - *Defense* against backdoor adversarial attacks
  - Data *Debugging* for ML
  - Data *Market*
- *Heuristics of Approximating Real-world Pipelines*
- *Failure Cases & "Call for Help"*

DS3Lab@ ETH ZURICH

ETH Systems@ETH Zürich

Heuristics 1: Unknown Noise to Codd's Table



Input Data without Noise Model

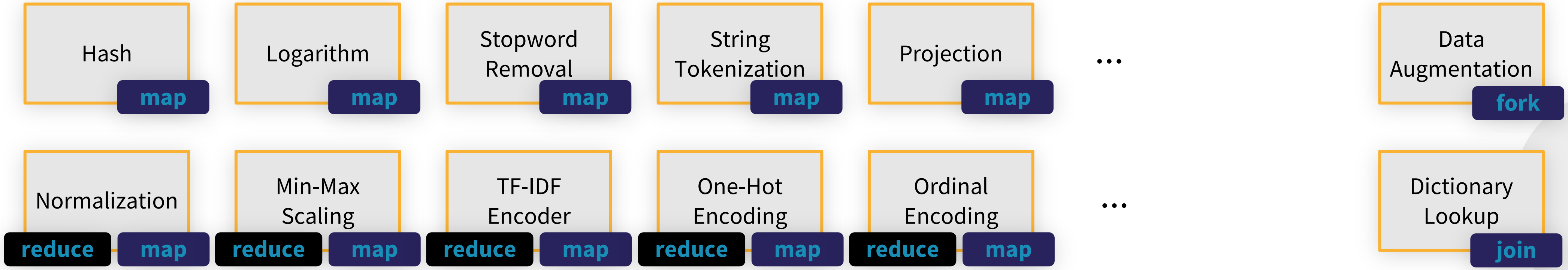
Input Data with Noise Model

1. Run a collection of SOTA *automatic data cleaning tools*
2. Use their cleaning result for each cell as a candidate value
3. Give them a uniform prior

Heuristics 2: ML Pipelines to Positive RA

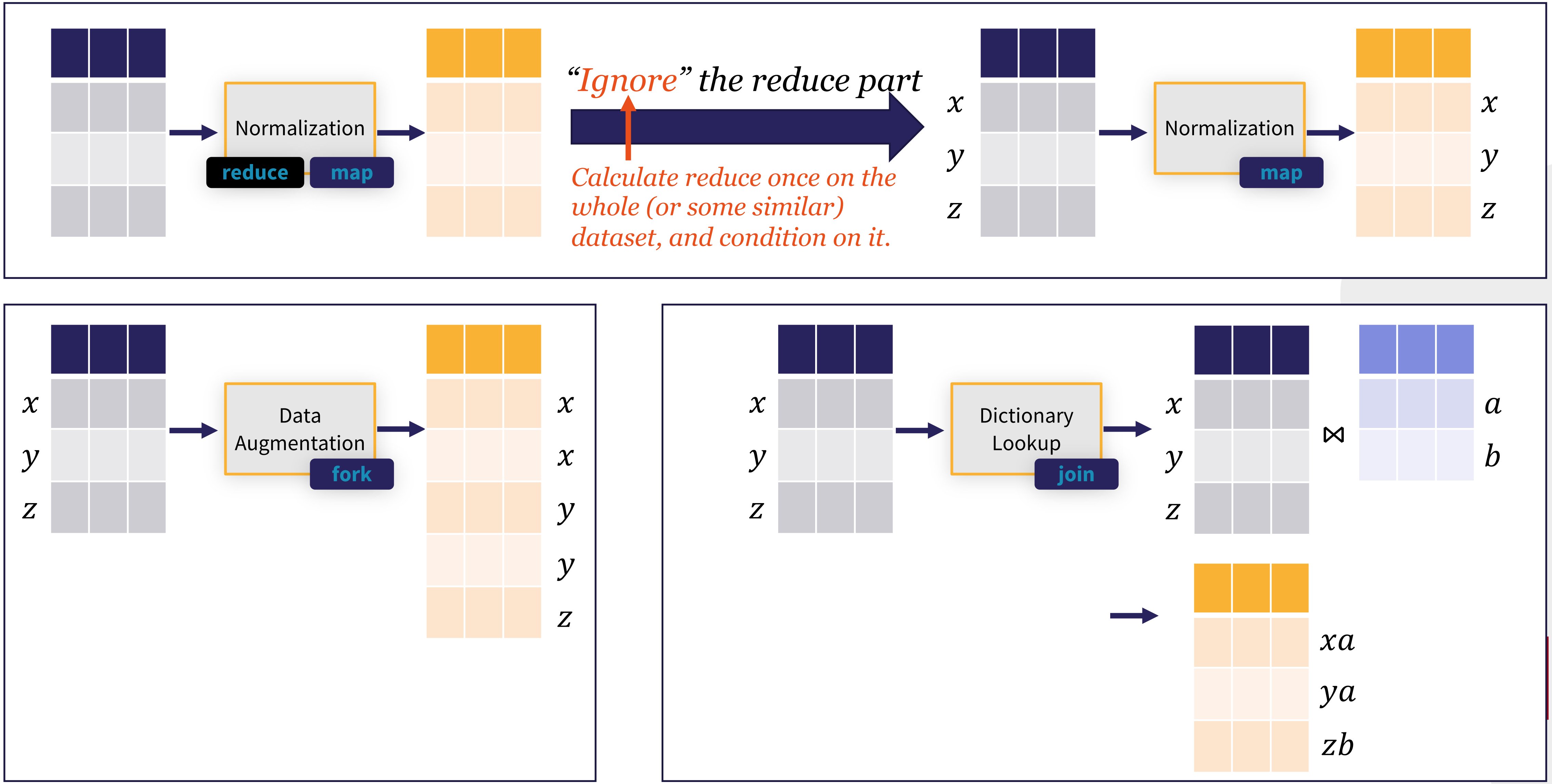
DS3Lab@
ETH ZURICH

- Dataflow using a diverse set of transformations, but different operators largely fall into two types:



- Many reduce functions are relatively stable with respect to the removal of data examples
- <u>**Heuristics**</u>: If we <u>*fix the reduce part to be over the whole dataset*</u>, conditioning on this, we can approximate a majority of pipelines.

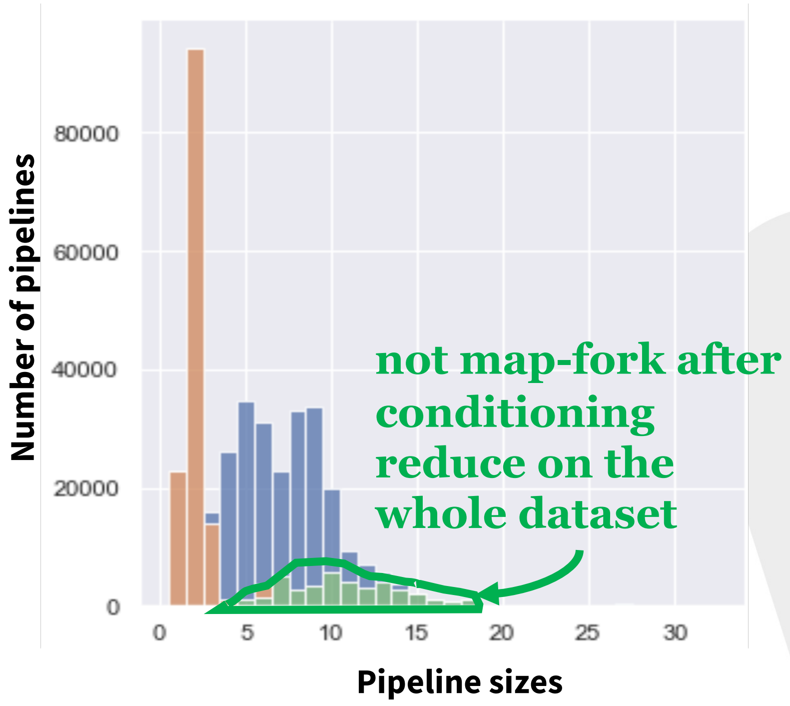
Heuristics 2: ML Pipelines to Positive RA



Heuristics 2: Characteristics of Pipelines

DS3Lab@
ETH ZURICH

- We take ~500K real-world pipelines
  - https://arxiv.org/pdf/1912.09536.pdf
- A majority of which fits into the map-fork pattern, after the conditioning heuristics.
- In additional to these pipelines:
  - Data federation/market (different subsets, union) introduces fork
  - Data augmentatin introduces fork



ETH
Systems@ETH zürich

Overview

- *Theoretical Result*
  - *Entropy/Expectation* of ML Training over Incomplete Information and Uncertainty
  - *Shapley value* of ML Training over Feature Extraction Pipelines
- *Applications*
  - Data *Cleaning* for ML
  - *Defense* against backdoor adversarial attacks
  - Data *Debugging* for ML
  - Data *Market*
- *Heuristics of Approximating Real-world Pipelines*
- *Failure Cases & "Call for Help"*

DS3Lab@
ETH ZURICH

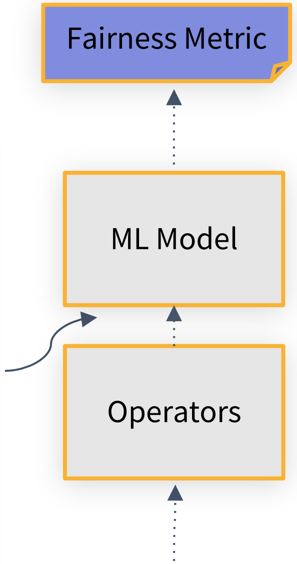
ETH
Systems@ETH Zürich

Failure Case (Shapley)

DS3Lab@ ETH ZURICH

Unbalanced training set with sensitive feature

race	sex	capital_gain	capital_loss	hours_per_week	country	target
White	Male	2174	0	40	United-States	<=50K
White	Male	0	0	13	United-States	<=50K
White	Male	0	0	40	United-States	<=50K
Black	Male	0	0	40	United-States	<=50K
Black	Female	0	0	40	Cuba	<=50K
White	Female	0	0	40	United-States	<=50K
Black	Female	0	0	16	Jamaica	<=50K

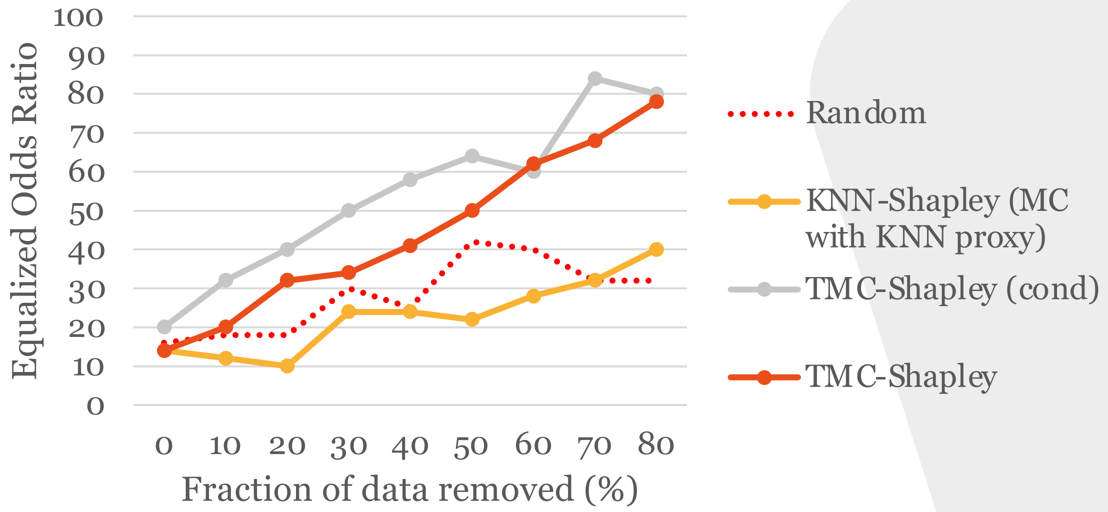


Balanced test set

How can we identify over-represented examples to remove?

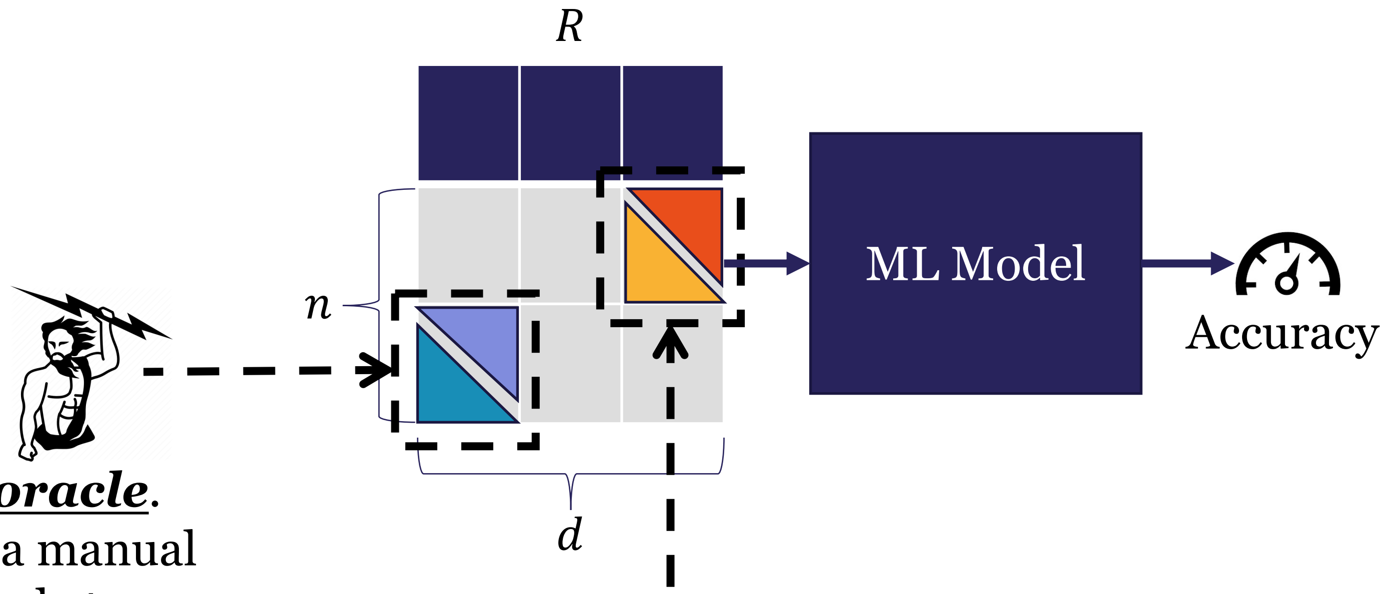
Idea: Data examples that are over-represented should have a small Shapley value.

KNN proxy does not work! (Conditioning still works)



Why? KNN captures more of a local structure, not the populational structure. (HELP!)

Failure Case (Entropy/Expectation)



Cleaning oracle. There exist a manual cleaning oracle to decide the ground truth value for each feature.

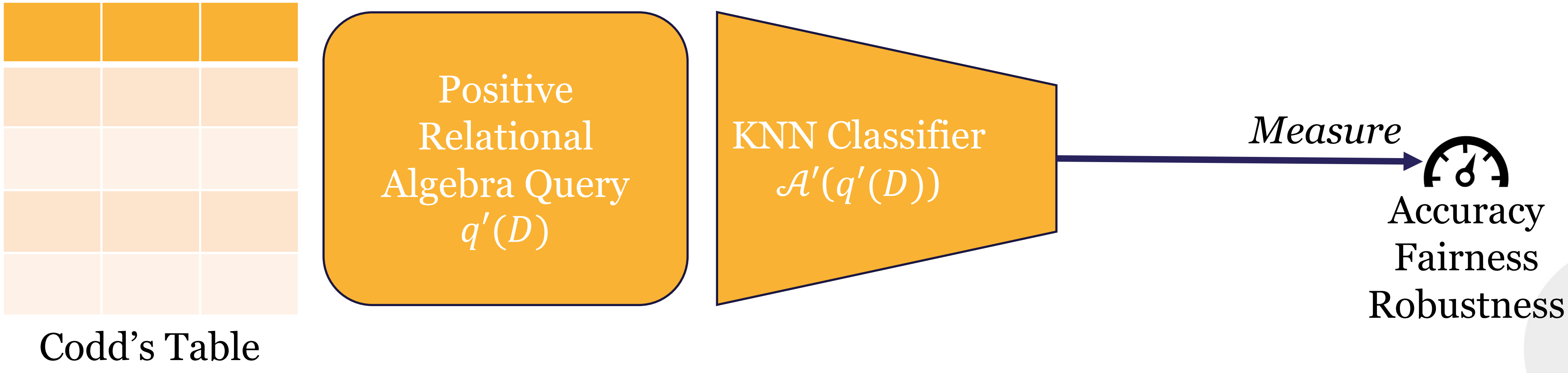
Uncertainty on features. Think about it as outputs of state-of-the-art data cleaning methods – each method output a candidate value.

What if the ML model is robust to noise? Or trained in a differentially private way?

KNN will probably also not work here. Different sensitivity to noises. (HELP!)

What’s Next?

DS3Lab@
ETH ZURICH



Takeaway 1

For a dominating number of realistic pipelines, for their proxy pipeline, we can compute exactly the Entropy, Expected Prediction, and Shapley Value, in PTIME.

Takeaway 2

These fundemental quantities, i.e., Entropy, Expected Prediction, Shapley Value, provide principled frameworks for many applications: Data *cleaning, debugging, valuation* for ML, *Defense* against backdoor adversarial attacks

Takeaway 3

Use these proxy pipelines as a proxy work well in many scenarios.

I will also show you cases when they don’t.

What is a pipeline, really?

Entropy & Expectation

Shapley Value & Expected Marginal Improvement

We don't know how to talk about it.

What is a FE pipeline?

What is NOT a FE pipeline?

? *Can we say anything about this beyond that it is a dataflow graph?*

? *Are all dataflow graphs feature extraction pipelines that we care about?*

? *What are the axioms describing "care about"?*

? *What is the corresponding algebraic structure?*

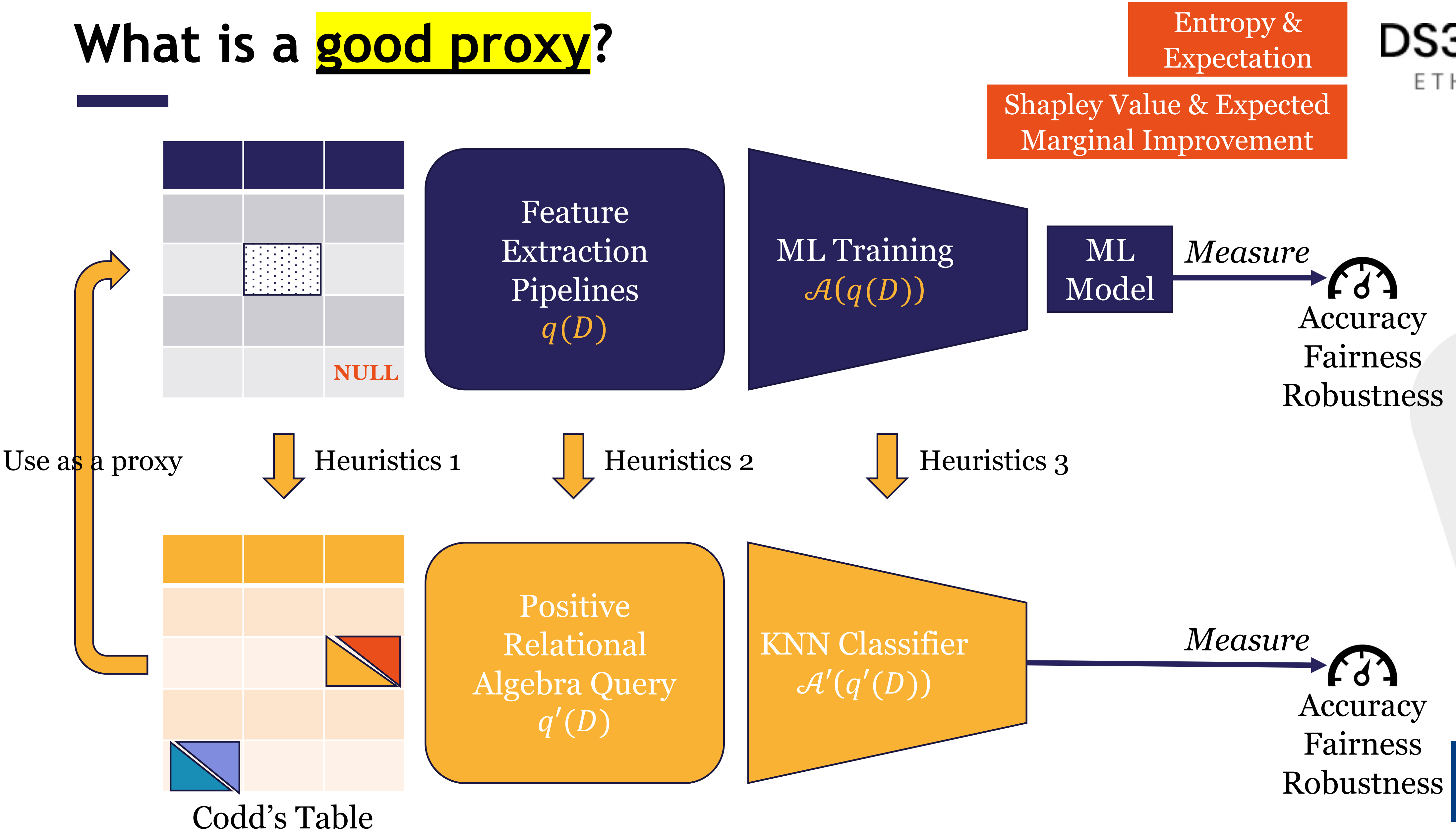
In good-old days, this is a semiring.

Why? From axioms of positive relational algebra.

It gets complex when it comes to negation.

But there are still axioms and differnet formulations, e.g., m-semirings.

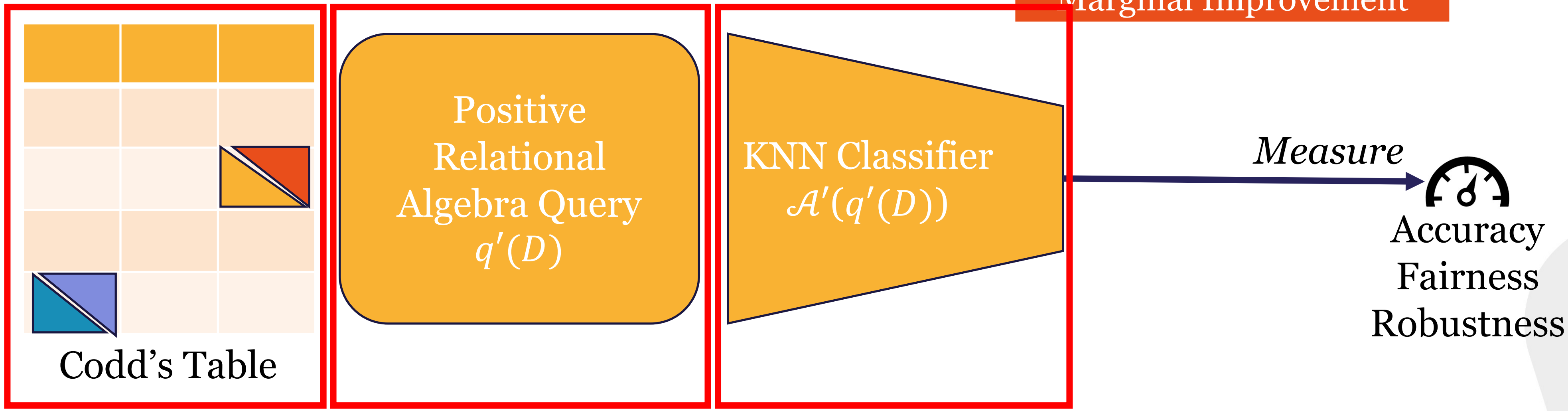
What is a good proxy?



What is a good proxy?

Entropy & Expectation

Shapley Value & Expected Marginal Improvement



Lose dependencies and interactions of noises

c-table, pc-table?

How to reason/approximate about the noise model?

Lose computation

Think about max but you remove the largest item

How to model the computation components?

Lose global, populational structure

What is a better proxy to use? Capture more global structure, yet being efficient.

What is a good metric?

Entropy & Expectation

Shapley Value & Expected Marginal Improvement

Alternatives?

Rényi Entropy?

Efficient estimator of Entropy?

Gradient-based Perturbation?

Leave-one-out?

Influence Functions?

What are their relationships?

1. Bojan Karlaš*, Peng Li*, ... Xu Chu, Wentao Wu ... . *VLDB* **2021**.
2. Ruoxi Jia, Fan Wu, ... Bo Li, Dawn Song. *CVPR* **2021**.
3. Linyi Li*, Maurice Weber*, ... *CCS* 2021.
4. Maurice Weber, Nana Liu ... Zhikuan Zhao. *npj Quantum Info.* 2021.
5. Ruoxi Jia*, David Dao* ... . *AISTATS* **2019**.
6. Ruoxi Jia, David Dao ... . *VLDB* **2019**.
7. Maurice Weber*, Xiaojun Xue* ... *Arxiv:2003.08904*
8. Fotis Psallidas ... Matteo Interlandi...Carlo Curino, Markus Weimer. *Arxiv:1912.09536*

https://ds3lab.inf.ethz.ch/
datascope.html

MLOps & MLDev made easy? Check out

http://ease.ml

Distributed ML in DB/Serverless/Spark/MPI? Check out

http://zip.ml

DS3Lab@
ETH ZURICH

@DS3Lab
www.DS3Lab.com

Acknowledgement

FNSNF Swiss National Science Foundation
Horizon 2020 European Union Funding for Research & Innovation
FNSNF BRIDGE
SDSC
BRC CH Botnar Research Centre for Child Health
Microsoft
Google
ORACLE
Alibaba Group
Kwai
ebay
cisco
Zürcher Kantonalbank
ZURICH
inovalon
MeteoSwiss
Mercedes-Benz
swisscom
ETH Systems@ETH zürich