UCLA Computer Science

STAR AI RESEARCH LAB UCLA

Reasoning about Missing Data in Machine Learning

Guy Van den Broeck

Emerging Challenges in Databases and AI Research (DBAI) – Nov 12 2019

Outline

1. Missing data at prediction time
    a. Reasoning about expectations
    b. Applications: classification and explainability
    c. Tractable circuits for expectation
    d. Fairness of missing data

2. Missing data during learning

References and Acknowledgements

- Pasha Khosravi, Yitao Liang, YooJung Choi and Guy Van den Broeck. What to Expect of Classifiers? Reasoning about Logistic Regression with Missing Features, *In IJCAI*, 2019.

- Pasha Khosravi, YooJung Choi, Yitao Liang, Antonio Vergari and Guy Van den Broeck. On Tractable Computation of Expected Predictions, *In NeurIPS*, 2019.

- YooJung Choi, Golnoosh Farnadi, Behrouz Babaki and Guy Van den Broeck. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns, *In AAAI*, 2020.

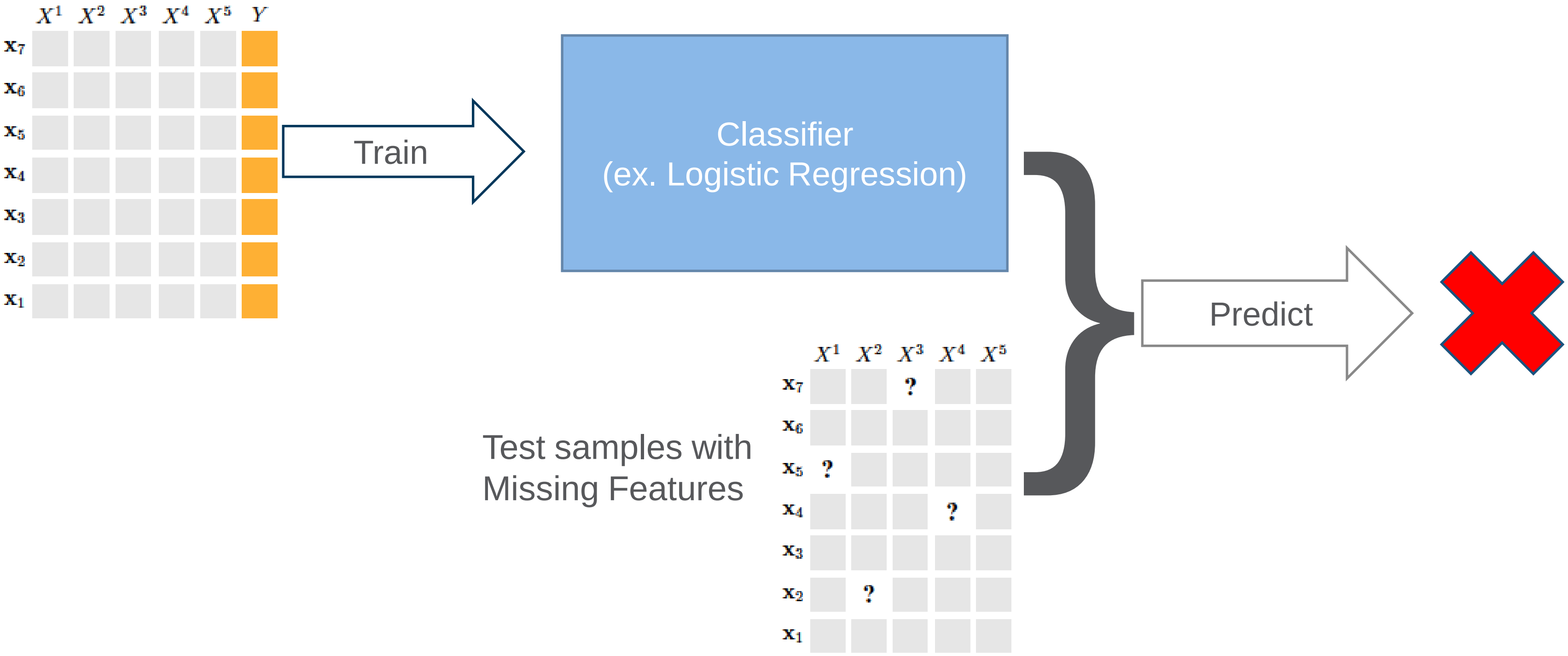
- Guy Van den Broeck, Karthika Mohan, Arthur Choi, Adnan Darwiche and Judea Pearl. Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data, *In UAI*, 2015.

Outline

1. **Missing data at prediction time**
   a. Reasoning about expectations
   b. Applications: classification and explainability
   c. Tractable circuits for expectation
   d. Fairness of missing data

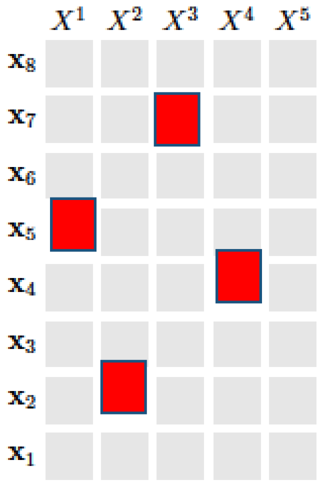
2. Missing data during learning

Missing data at prediction time



Common Approaches

- Fill out the missing features, i.e. doing imputation.
- Makes unrealistic assumptions (mean, median, etc).
- More sophisticated methods such as MICE don't scale to bigger problems (and also have assumptions).
- We want a more principled way of dealing with missing data while staying efficient.



Discriminative vs. Generative Models

Terminology:

- **Discriminative Model:** conditional probability distribution, $P(C\ |X)$**.**
  For example, Logistic Regression.

- **Generative Model:** joint features and class probability distribution, $P(C, X)$.
  For example, Naïve Bayes.

Suppose we only observe some features **y** in X, and we are missing **m:**

$$P(C|\boldsymbol{y}) = \sum_{\boldsymbol{m}} P(C, \boldsymbol{m}|\boldsymbol{y}) \propto \sum_{\boldsymbol{m}} P(C, \boldsymbol{m}, \boldsymbol{y})$$

We need a generative model!

Generative vs Discriminative Models

	Discriminative Models (ex. Logistic Regression)	**Generative Models (ex. Naïve Bayes)**
	$P(C \mid X)$	$P(C, X)$
Missing Features	😕	🙂
Classification Accuracy	🙂	😕

Outline

1. Missing data at prediction time
    a. **Reasoning about expectations**
    b. Applications: classification and explainability
    c. Tractable circuits for expectation
    d. Fairness of missing data

2. Missing data during learning

Generative Model Inference as Expectation

Let's revisit how generative models deal with missing data:

$$P(C|\boldsymbol{y}) = \sum_{\boldsymbol{m}} P(C, \boldsymbol{m}|\boldsymbol{y})$$

$$= \sum_{\boldsymbol{m}} P(C|\boldsymbol{m}, \boldsymbol{y})\, P(\boldsymbol{m}|\boldsymbol{y})$$

$$= \mathbb{E}_{\boldsymbol{m} \sim P(M|\boldsymbol{y})}\, P(C|\boldsymbol{m}, \boldsymbol{y})$$

It's an expectation of a classifier under the feature distribution

What to expect of classifiers?

What if we train both kinds of models:

1. Generative model for feature distribution $P(X)$.
2. Discriminative model for the classifier $F(X) = P(C \mid X)$.

"**Expected Prediction**" is a principled way to reason about outcome of classifier $F(X)$ under feature distribution $P(X)$.

$$E_{\mathcal{F},P}(\mathbf{y}) = \underset{\mathbf{m} \sim P(\mathbf{M}|\mathbf{y})}{\mathbb{E}} \left[\mathcal{F}(\mathbf{y}\mathbf{m}) \right]$$

Expected Predication Intuition

- **Imputation Techniques**: Replace the missing-ness uncertainty with *one* or *multiple* possible inputs, and evaluate the models.

- **Expected Prediction**: Considers *all possible inputs* and reason about expected behavior of the classifier.

$$E_{\mathcal{F},P}(\mathbf{y}) = \sum_{\mathbf{m}} P(\mathbf{m} \mid \mathbf{y}) \cdot \mathcal{F}(\mathbf{ym}) = \mathop{\mathbb{E}}_{\mathbf{m} \sim P(\mathbf{M} \mid \mathbf{y})} [\mathcal{F}(\mathbf{ym})]$$

Hardness of Taking Expectations

- How can we compute the expected prediction?

- In general, it is intractable for arbitrary pairs of discriminative and generative models. ☹

- Even when ☹
  - ✓ Classifier F is Logistic Regression and
  - ✓ Generative model P is Naïve Bayes,

  the task is NP-Hard.

Solution: Conformant learning

Given a classifier and a dataset, learn a generative model that

1. *Conforms* to the classifier: $F(X) = P(C \mid X)$.

2. Maximizes the likelihood of generative model: $P(X)$.

No missing features → Same quality of classification ☺
Has missing features → No problem, do inference ☺

<u>Example</u>: Naïve Bayes (NB) vs. Logistic Regression (LR):

- Given NB there is one LR that it conforms to
- Given LR there are many NB that conform to it

Naïve Conformant Learning (NaCL)

Logistic Regression Weights

NaCL

"Best" Conforming Naïve Bayes

Optimization task as a *Geometric Program*

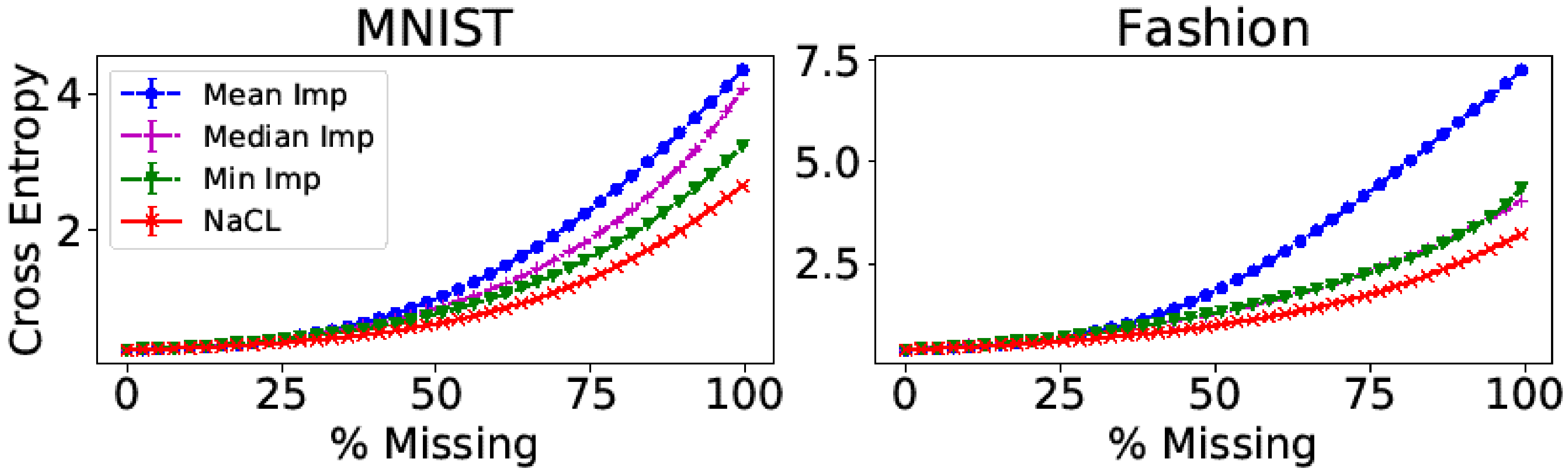
GitHub: github.com/UCLA-StarAI/NaCL

Outline

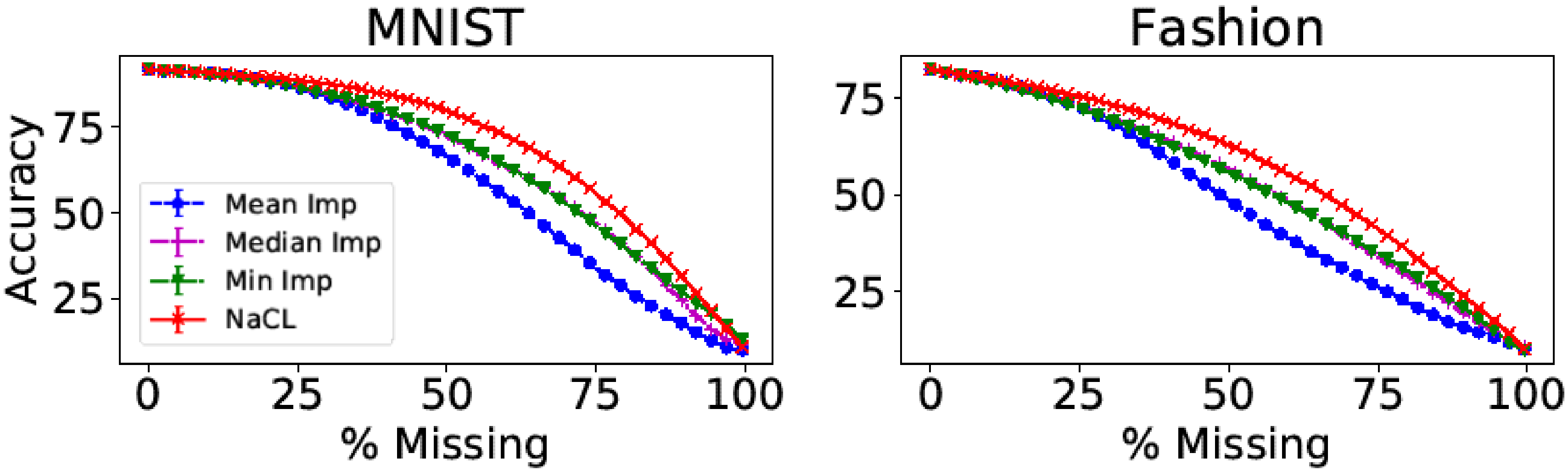
1. Missing data at prediction time
    a. Reasoning about expectations
    b. Applications: classification and explainability
    c. Tractable circuits for expectation
    d. Fairness of missing data

2. Missing data during learning

Experiments: Fidelity to Original Classifier



Experiments: Classification Accuracy

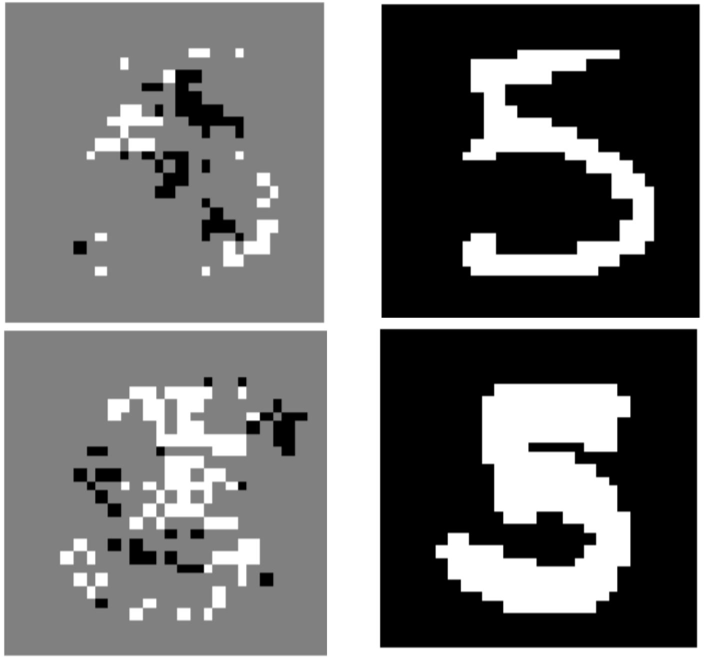


Sufficient Explanations of Classification

Goal:
To explain an instance of classification

Support Features:
Making them missing
→ probability goes down

Sufficient Explanation:
Smallest set of support features
that retains the expected classification

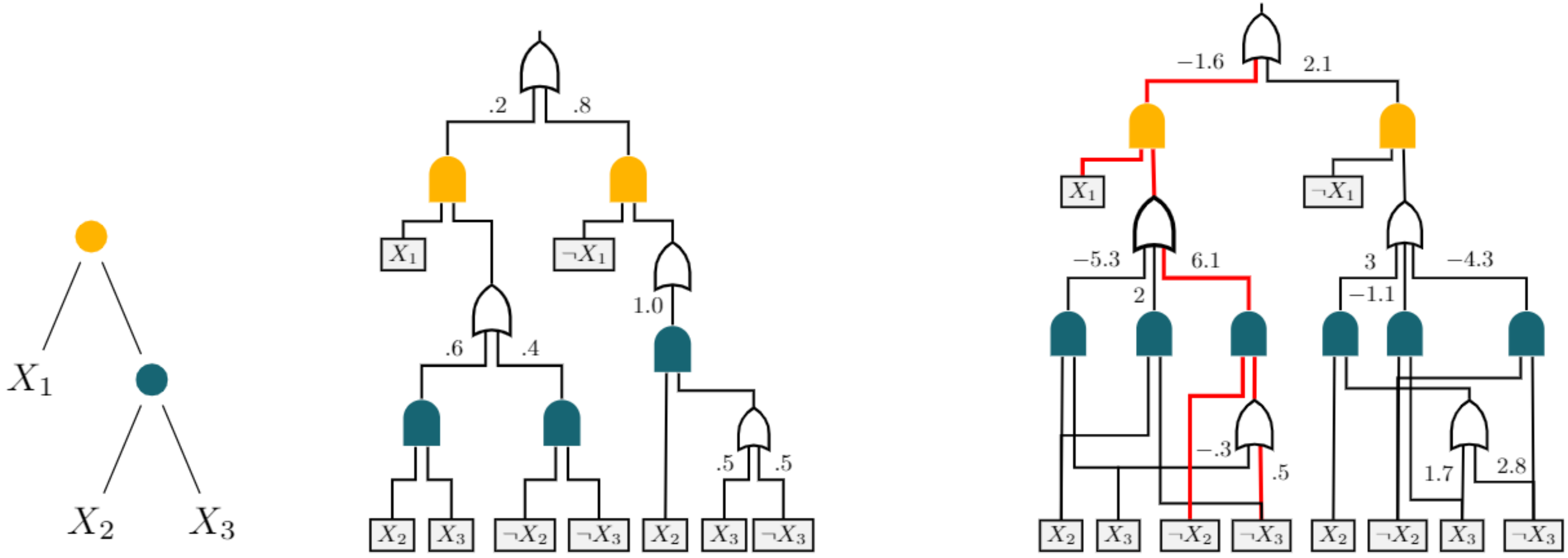


Outline

1. Missing data at prediction time
    a. Reasoning about expectations
    b. Applications: classification and explainability
    c. **Tractable circuits for expectation**
    d. Fairness of missing data

2. Missing data during learning

What about better distributions and classifiers?



(a) A vtree

(b) A Probabilistic Circuit

(c) A Logistic/Regression Circuit

Generative

Discriminative

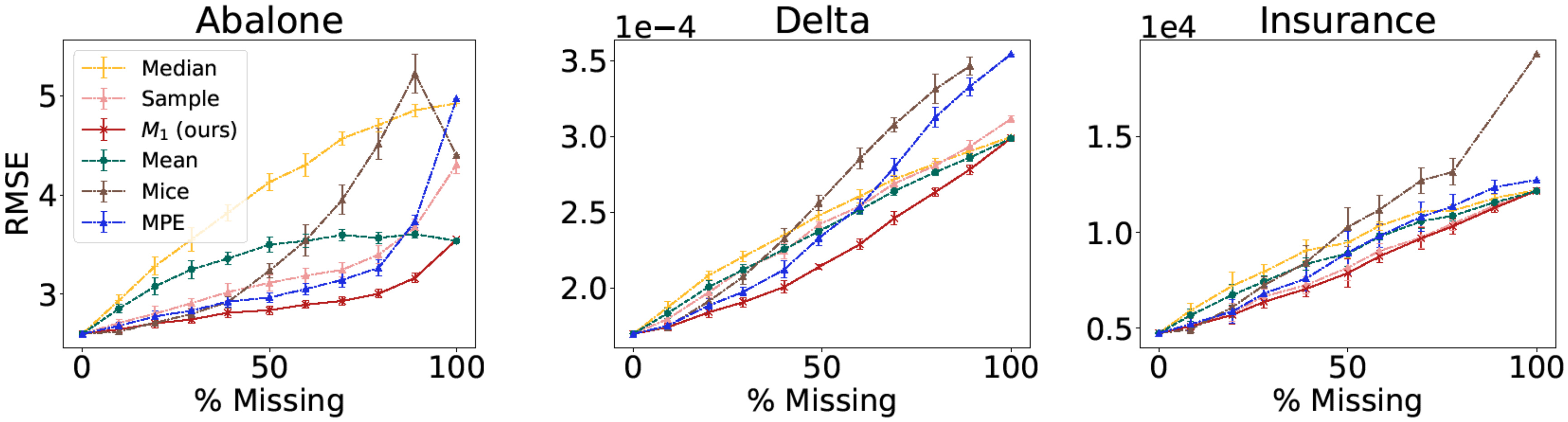
Hardness of Taking Expectations

If f is a regression circuit, and p is a generative circuit
with **different** vtree — Proved #P-Hard ☹

If f is a classification circuit, and p is a generative circuit
with **different** vtree — Proved NP-Hard ☹

If f is a regression circuit, and p is a generative circuit
with the **same** vtree — Polytime algorithm ☺

Regression Experiments



Approximate Expectations of Classification

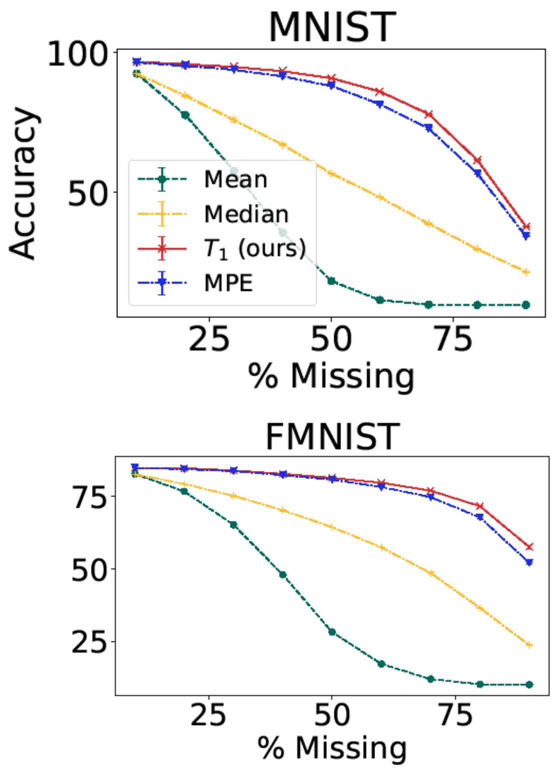
What to do for classification circuits?
(Even with same vtree, expectation was intractable.)

⇒ Approximate classification using Taylor series of the underlying regression circuit.

$$\mathbb{E}_{\mathbf{x}\sim p_n(\mathbf{x})}\left[\gamma \circ g_m(\mathbf{x})\right] \approx \sum_{i=0}^{d} \frac{\gamma^{(i)}(\alpha)}{i!} M_i(g_m - \alpha, p_n)$$

⇒ Requires higher order moments of regression circuit…

⇒ This is also efficient! ☺



Exploratory Classifier Analysis

Expected predictions enable reasoning about behavior of predictive models

We have learned an regression and a probabilistic circuit for
“Yearly health insurance costs of patients”

Q1: Difference of costs between smokers and non-smokers

$$M_1(f,\ p(.\mid \mathit{Smoker})) - M_1(f,\ p(.\mid \mathit{Non\ Smoker})) = 22,614$$

…or between female and male patients?

$$M_1(f,\ p(.\mid \mathit{Female})) - M_1(f,\ p(.\mid \mathit{Male})) = 974$$

Exploratory Classifier Analysis

Can also answer more complex queries like:

Q2: Average cost for female (F) smokers (S)
with one child (C) in the South East (SE)?

$$M_1(f, p(. \mid \mathsf{F}, \mathsf{S}, \mathsf{C}, \mathsf{SE})) = 30,974$$

Q3: Standard Deviation of the cost for the same sub-population?

$$\sqrt{M_2(.) - (M_1(.))^2} = 11,229$$

Outline

1. Missing data at prediction time
   a. Reasoning about expectations
   b. Applications: classification and explainability
   c. Tractable circuits for expectation
   d. Fairness of missing data

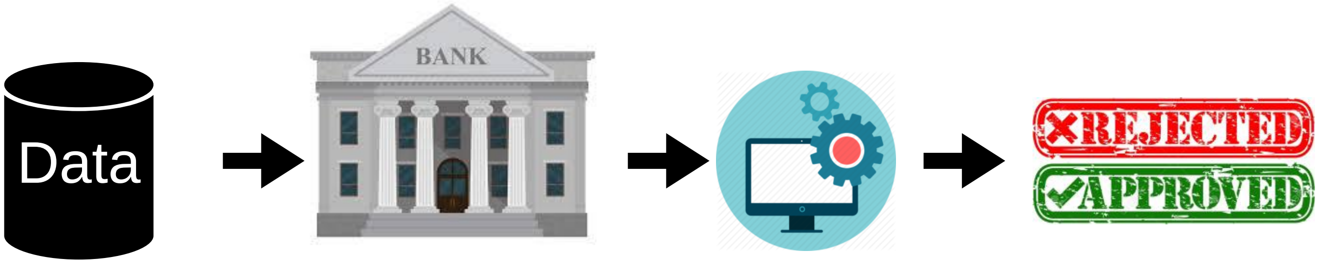
2. Missing data during learning

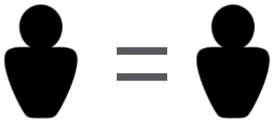
Algorithmic Fairness

Legally recognized 'protected classes'

Race (Civil Rights Act of 1964)
Color (Civil Rights Act of 1964)
Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
Religion (Civil Rights Act of 1964)
National origin (Civil Rights Act of 1964)
Citizenship (Immigration Reform and Control Act)
Age (Age Discrimination in Employment Act of 1967)
Pregnancy (Pregnancy Discrimination Act)
Familial status (Civil Rights Act of 1968)
Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act);
Genetic information (Genetic Information Nondiscrimination Act)

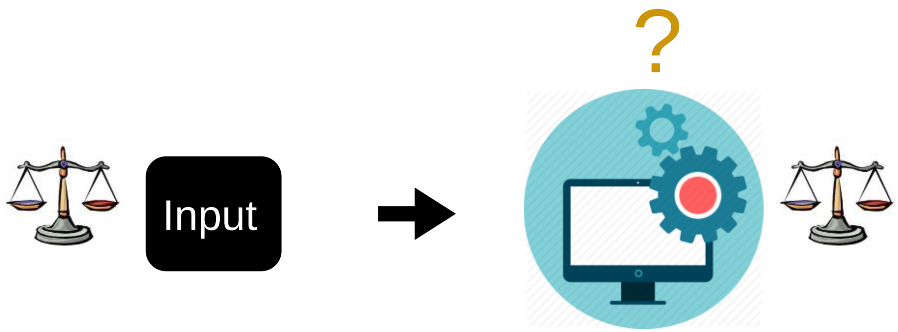
Individual Fairness



- Individual fairness: 
- Existing methods often define individuals as a **fixed set of observable features**
- Lack of discussion of certain features **not being observed at prediction time**

What about learning from fair data?

Model learned from repaired data can still be unfair!



Number of discrimination patterns:

Dataset	λ =0.5	λ =0.9	λ =0.95	λ =0.99	λ =1.0
COMPAS	11,512	7,862	8,872	8,926	0
Adult	>1e6	1,078	1,123	1,087	0
German	>1e6	1	9	0	0

Independent

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkata-subramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268.ACM, 2015

Individual Fairness with Partial Observations

- **Degree of discrimination**: $\Delta(\boldsymbol{x}, \boldsymbol{y}) = P(d|\boldsymbol{x}\boldsymbol{y}) - P(d|\boldsymbol{y})$

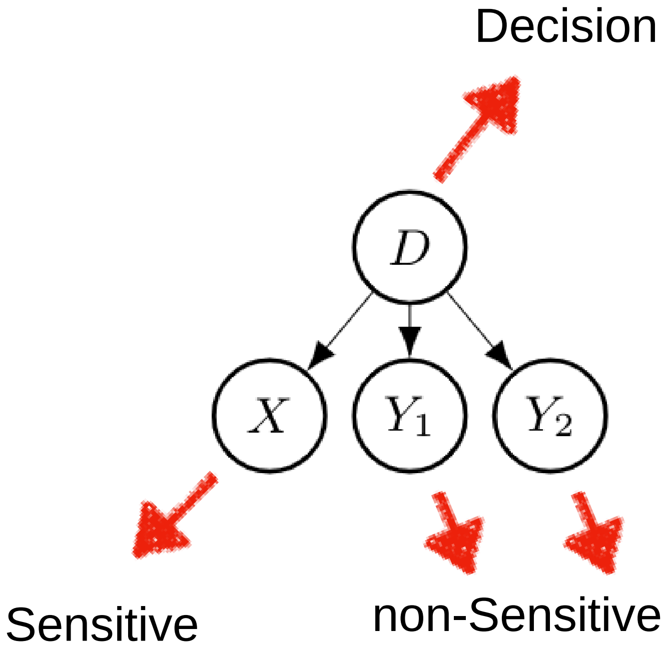
Decision given partial evidence

Decision without sensitive attributes

"What if the applicant had not disclosed their gender?"

- **$\boldsymbol{\delta}$-fairness**: $\Delta(\boldsymbol{x}, \boldsymbol{y}) \leq \delta,\ \forall \boldsymbol{x}, \boldsymbol{y}$
- A violation of δ-fairness is a **discrimination pattern** $\mathbf{x}, \mathbf{y}$.

Discovering and Eliminating Discrimination



1. **Verify** whether a Naive Bayes **classifier is δ-fair** by mining the classifier for discrimination patterns

2. **Parameter learning** algorithm for Naive Bayes classifier to **eliminate discrimination** patterns

Technique: Signomial Programming

$$\operatorname{argmax} P(C, X_1, X_2, \dots, X_m, Y_1, Y_2, , \dots, Y_n)$$

Max Likelihood Naive Bayes

$$s.t.$$

$$P(C|X_1, Y_1) - P(C|Y_1) \leq \delta$$

$$\dots$$

$$P(C|X_m, Y_1) - P(C|Y_1) \leq \delta$$

$$\dots$$

$$P(C|X_1, Y_n) - P(C|Y_n) \leq \delta$$

$$\dots$$

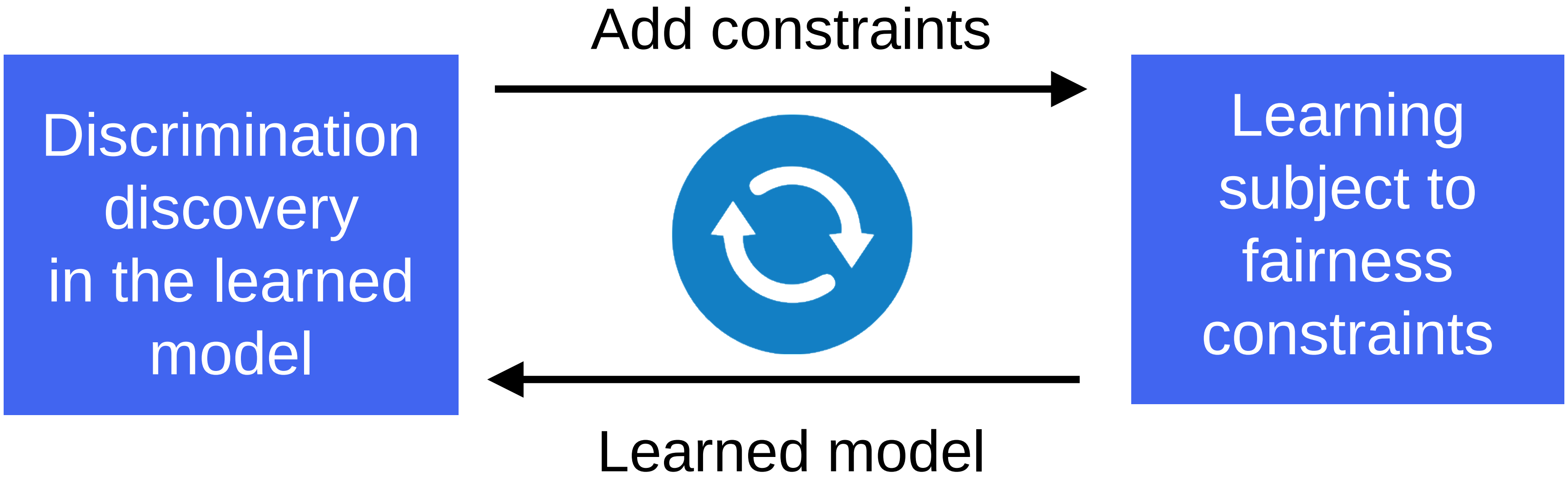
$$P(C|X_1, X_2, Y_1) - P(C|Y_1) \leq \delta$$

$$\dots$$

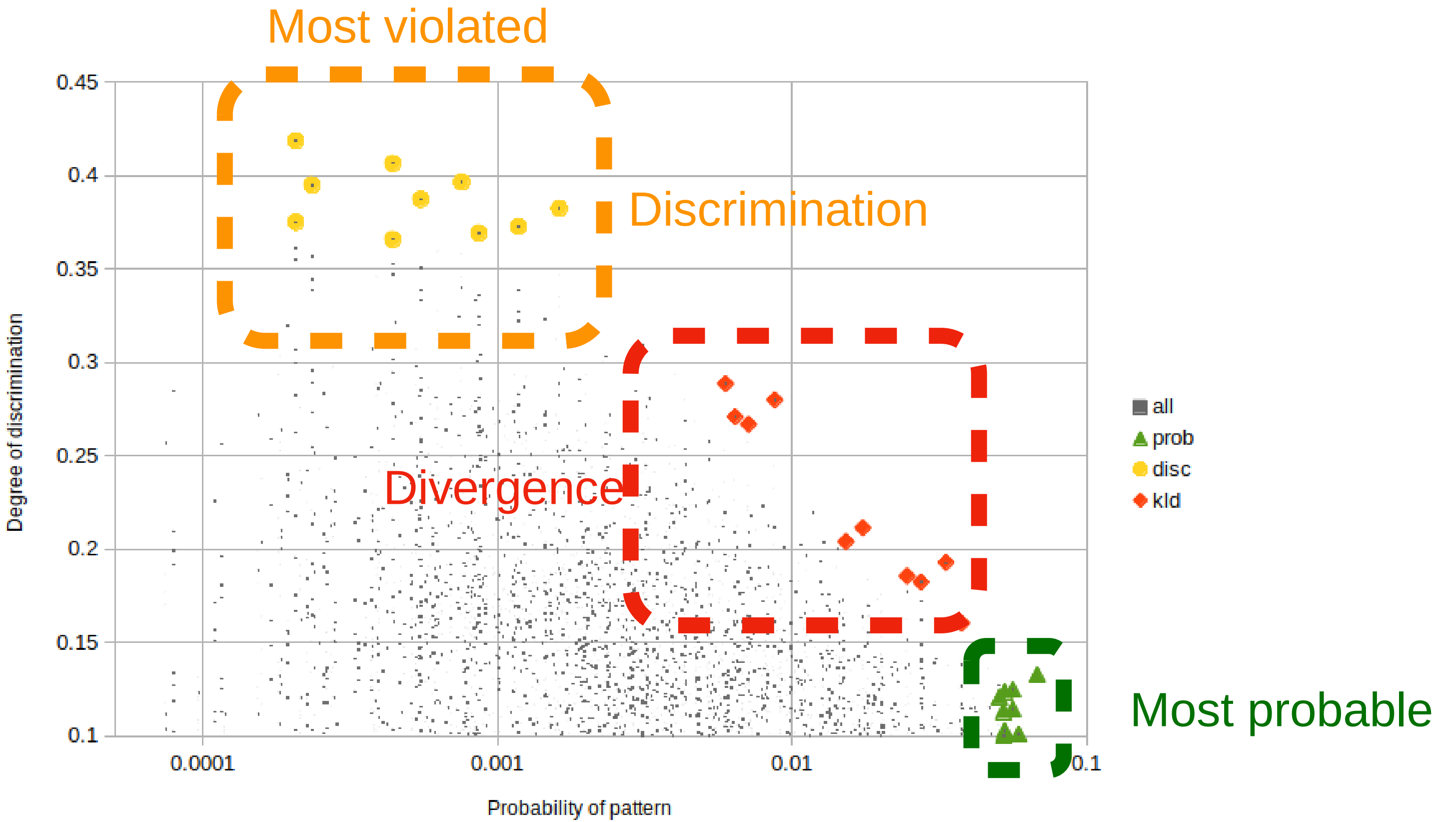
$$P(C|X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n) - P(C|Y_1, Y_2, \dots, Y_n) \leq \delta$$

δ-fair constraints

Cutting Plane Approach



Which constraints to add?



Quality of Learned Models?

Almost as good (likelihood) as unconstrained unfair model

Dataset	Unconstrained	δ-Fair	Independent
COMPAS	-207,055	-207,395	-208,639
Adult	-226,375	-228,763	-232,180
German	-12,630	-12,635	-12,649

Higher accuracy than other fairness approaches, while recognizing discrimination patterns involving missing data

dataset	Unconstrained	2NB	Repaired	δ-fair
COMPAS	0.880	0.875	0.878	0.879
Adult	0.811	0.759	0.325	0.827
German	0.690	0.679	0.688	0.696

Outline

1. Missing data at prediction time
    a. Reasoning about expectations
    b. Applications: classification and explainability
    c. Tractable circuits for expectation
    d. Fairness of missing data

2. **Missing data during learning**

Current learning approaches

	Likelihood Optimization
Inference-Free	✗
Consistent for MCAR	✔
Consistent for MAR	✔
Consistent for MNAR	✗
Maximum Likelihood	✔

Current learning approaches

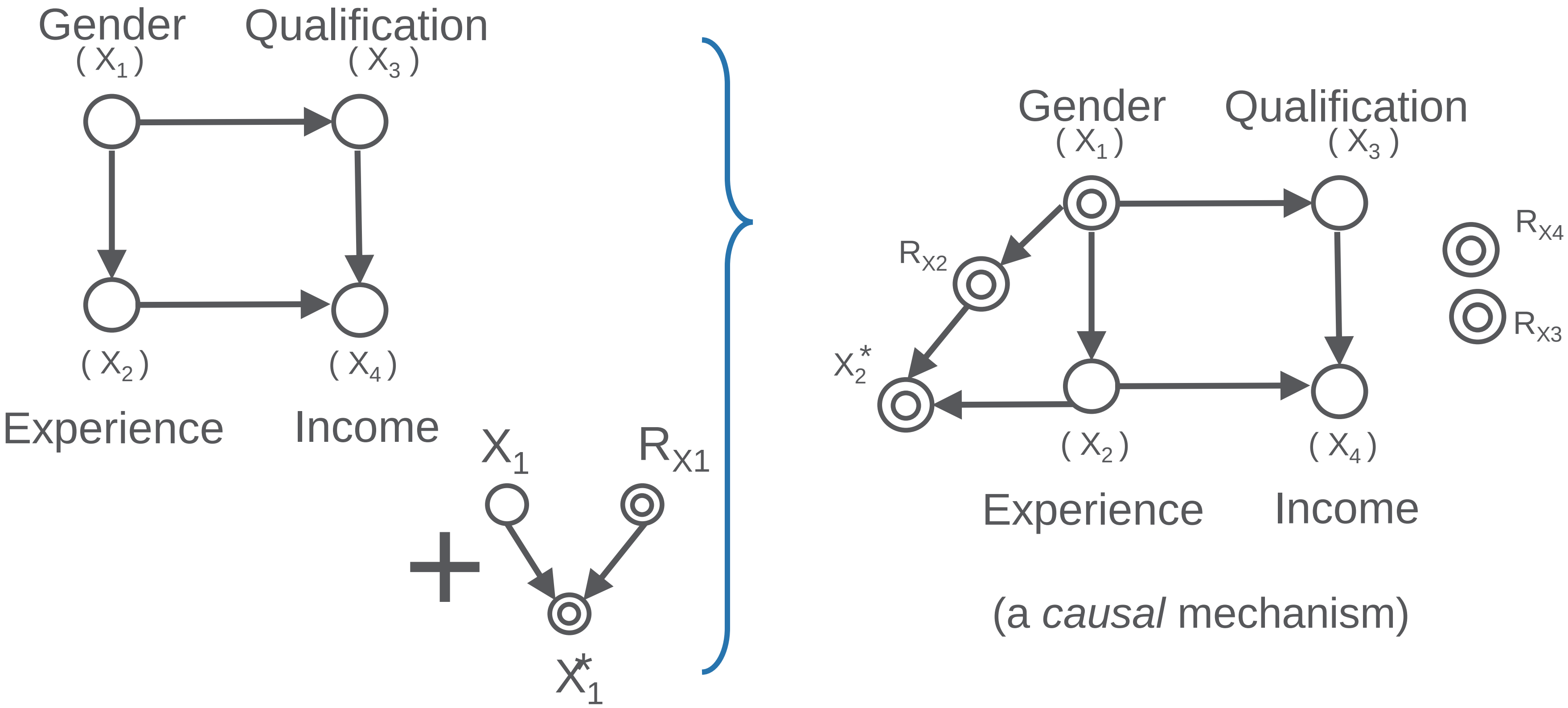
	Likelihood Optimization	Expectation Maximization
Inference-Free	✘	✘
Consistent for MCAR	✔	✔/✘
Consistent for MAR	✔	✔/✘
Consistent for MNAR	✘	✘
Maximum Likelihood	✔	✔/✘
Closed Form	n/a	✘
Passes over the data	n/a	?

Current learning approaches

	Likelihood Optimization	Expectation Maximization
Inference-Free	✘	✘
Consistent for MCAR	✔	✔/✘
Consistent for MAR	✔	✔/✘
Consistent for MNAR	✘	✘
Maximum Likelihood	✔	✔/✘
Closed Form	n/a	✘
Passes over the data	n/a	?

Conventional wisdom: downsides are inevitable!

Reasoning about Missingness Mechanisms

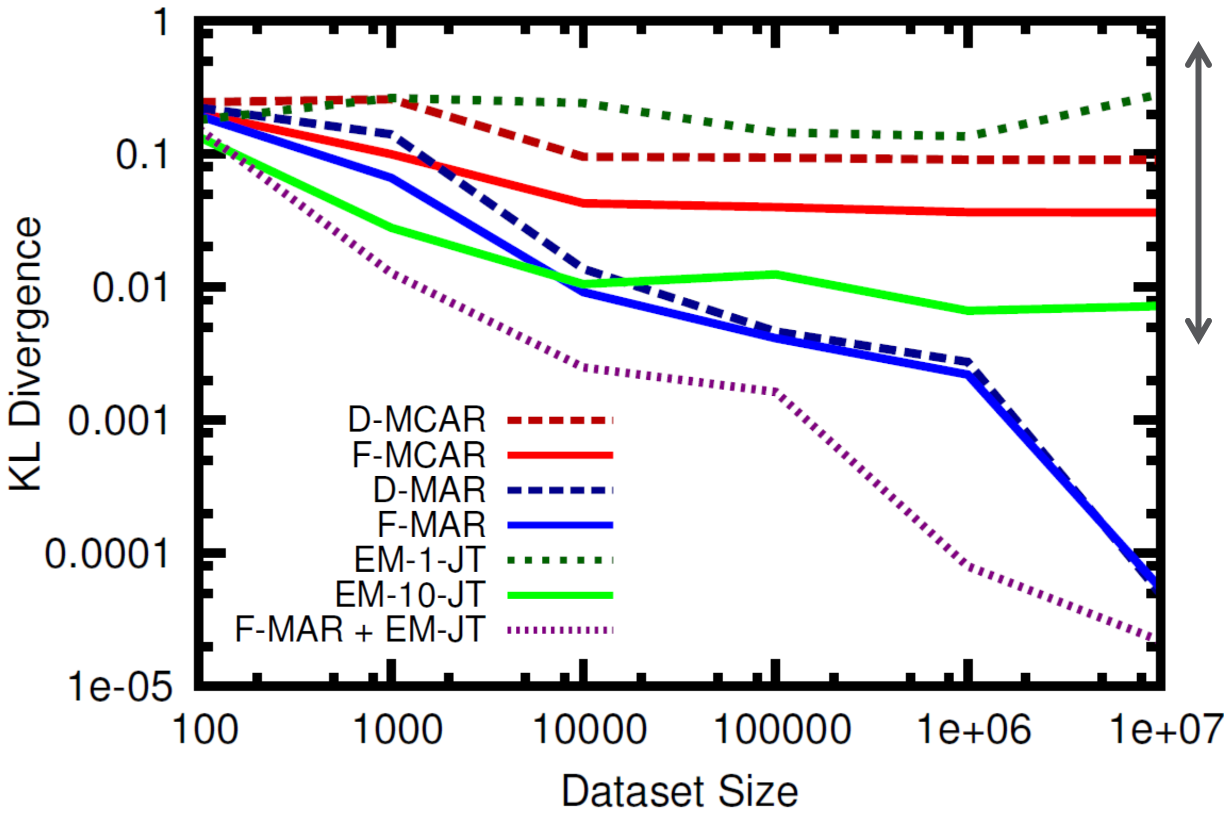


(a *causal* mechanism)

Deletion Algorithms for Missing Data Learning

	Likelihood Optimization	Expectation Maximization	Deletion [our work]
Inference-Free	✘	✘	✔
Consistent for MCAR	✔	✔/✘	✔
Consistent for MAR	✔	✔/✘	✔
Consistent for MNAR	✘	✘	✔/✘
Maximum Likelihood	✔	✔/✘	✘
Closed Form	n/a	✘	✔
Passes over the data	n/a	?	1

Benefits bear out in practice!



INCONSISTENT

Conclusions

- Missing data is a central problem in machine learning
- We can do better than classical tools from statistics
- By doing reasoning about the data distribution!
  - ➢ In a generative model that conforms to the classifier
  - ➢ Expectations using tractable circuits as new ML models
  - ➢ Using causal missingness mechanisms
- Important in addressing problems of robustness, fairness, and explainability

References and Acknowledgements

- Pasha Khosravi, Yitao Liang, YooJung Choi and Guy Van den Broeck. What to Expect of Classifiers? Reasoning about Logistic Regression with Missing Features, *In IJCAI*, 2019.

- Pasha Khosravi, YooJung Choi, Yitao Liang, Antonio Vergari and Guy Van den Broeck. On Tractable Computation of Expected Predictions, *In NeurIPS*, 2019.

- YooJung Choi, Golnoosh Farnadi, Behrouz Babaki and Guy Van den Broeck. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns, *In AAAI*, 2020.

- Guy Van den Broeck, Karthika Mohan, Arthur Choi, Adnan Darwiche and Judea Pearl. Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data, *In UAI*, 2015.

Thank You