# Are Attribute Inference Attacks Just Imputation?

**Bargav Jayaraman**
*bj4nq@virginia.edu*
University of Virginia

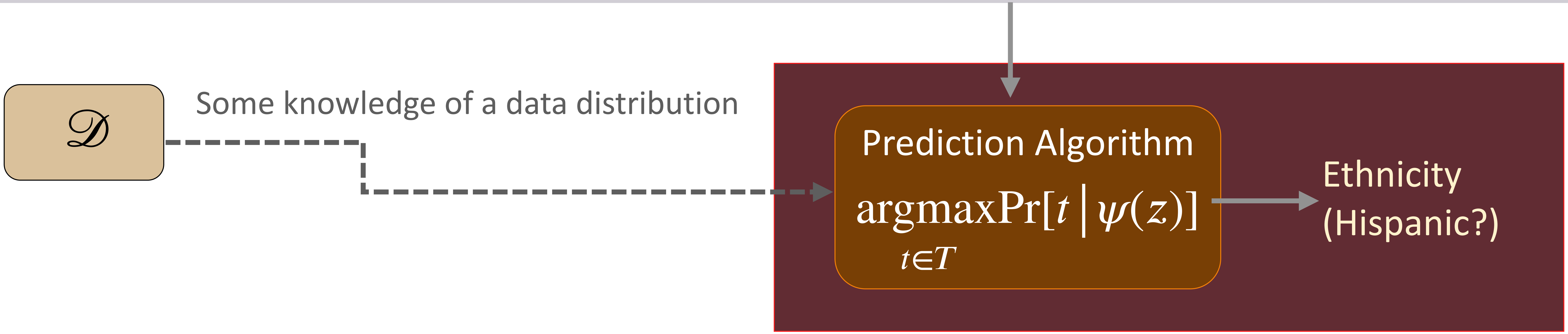
**David Evans**
*evans@virginia.edu*
University of Virginia

Texas-100X dataset (https://github.com/bargavj/Texas-100X), 925K hospital records from 441 Texas hospitals

# Imputation

*Infer missing attributes from available data*

| Hospital | Gender | Source | Stay Length | Patient Age | Ethnicity | Charges | ... | Procedure |
|---|---|---|---|---|---|---|---|---|
| 102 | 1 | 6 | 10 | 11 | ? | $34920.33 | ... | 34 |
| 102 | 0 | 2 | 3 | 8 | ? | $4062.46 | ... | 95 |
| 350 | 0 | 6 | 23 | 18 | ? | $105239.23 | ... | 62 |



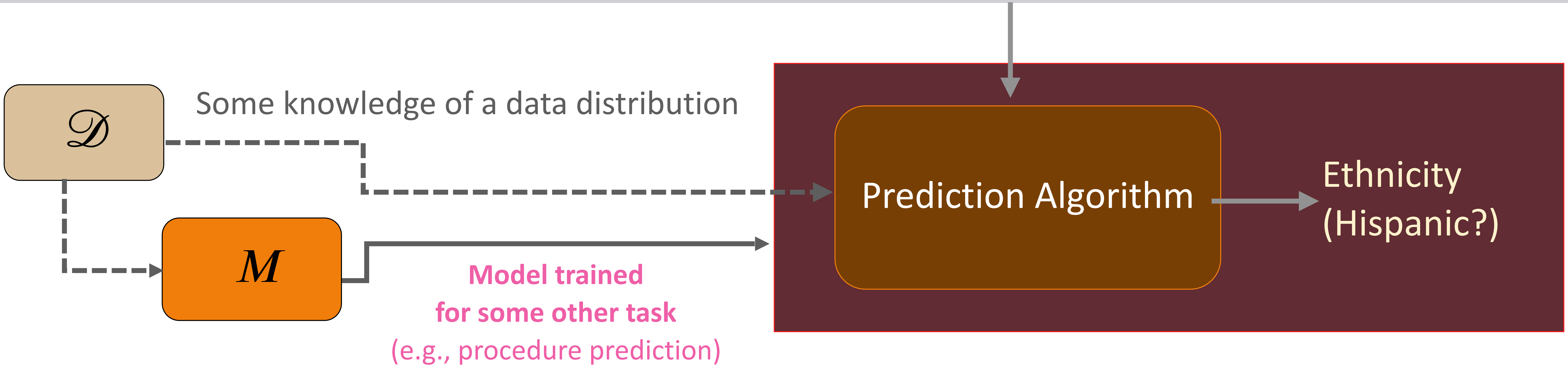

Traditionally, “imputation” is not considered an attack or privacy risk.

Texas-100X dataset (https://github.com/bargavj/Texas-100X), 925K hospital records from 441 Texas hospitals

# Attribute Inference

*Infer missing attributes from available data and **model***

| Hospital | Gender | Source | Stay Length | Patient Age | Ethnicity | Charges | … | Procedure |
|---|---|---|---|---|---|---|---|---|
| 102 | 1 | 6 | 10 | 11 | ? | $34920.33 | … | 34 |
| 102 | 0 | 2 | 3 | 8 | ? | $4062.46 | … | 95 |
| 350 | 0 | 6 | 23 | 18 | ? | $105239.23 | … | 62 |




“Attribute inference” is considered an attack and a privacy risk.

# Exploring Different Threat Settings

## Adversary Knows the Training Distribution

Adversary's Data Set Size

| | Yes | No |
|---|---|---|
| Large | **Prior AI Attacks**<br>Fredrikson et al. [USENIX Sec 14]<br>Yeom et al. [CSF 18]<br>Mehnaz et al. [USENIX Sec 22]<br>Imputation itself does well! | |
| Small | | |

# Average Prediction Accuracy

| Attribute Prediction Method | Texas-100X Gender | Texas-100X Ethnicity | Census19 Gender | Census19 Race |
|---|---|---|---|---|
| Naïve Most Common | 0.62 | 0.72 | 0.52 | 0.78 |
| Imputation | 0.66 | 0.72 | 0.59 | 0.82 |
| Yeom et al. [CSF 2018] | 0.62 | 0.64 | 0.63 | 0.06 |
| Mehnaz et al. [USENIX Sec 2022] | 0.59 | 0.60 | 0.63 | 0.06 |
| WCAI (our improvements to Yeom) | 0.68 | 0.74 | 0.64 | 0.83 |

Black-box attribute inference attacks do not meaningfully outperform imputation:
**no evidence that access to model helps**

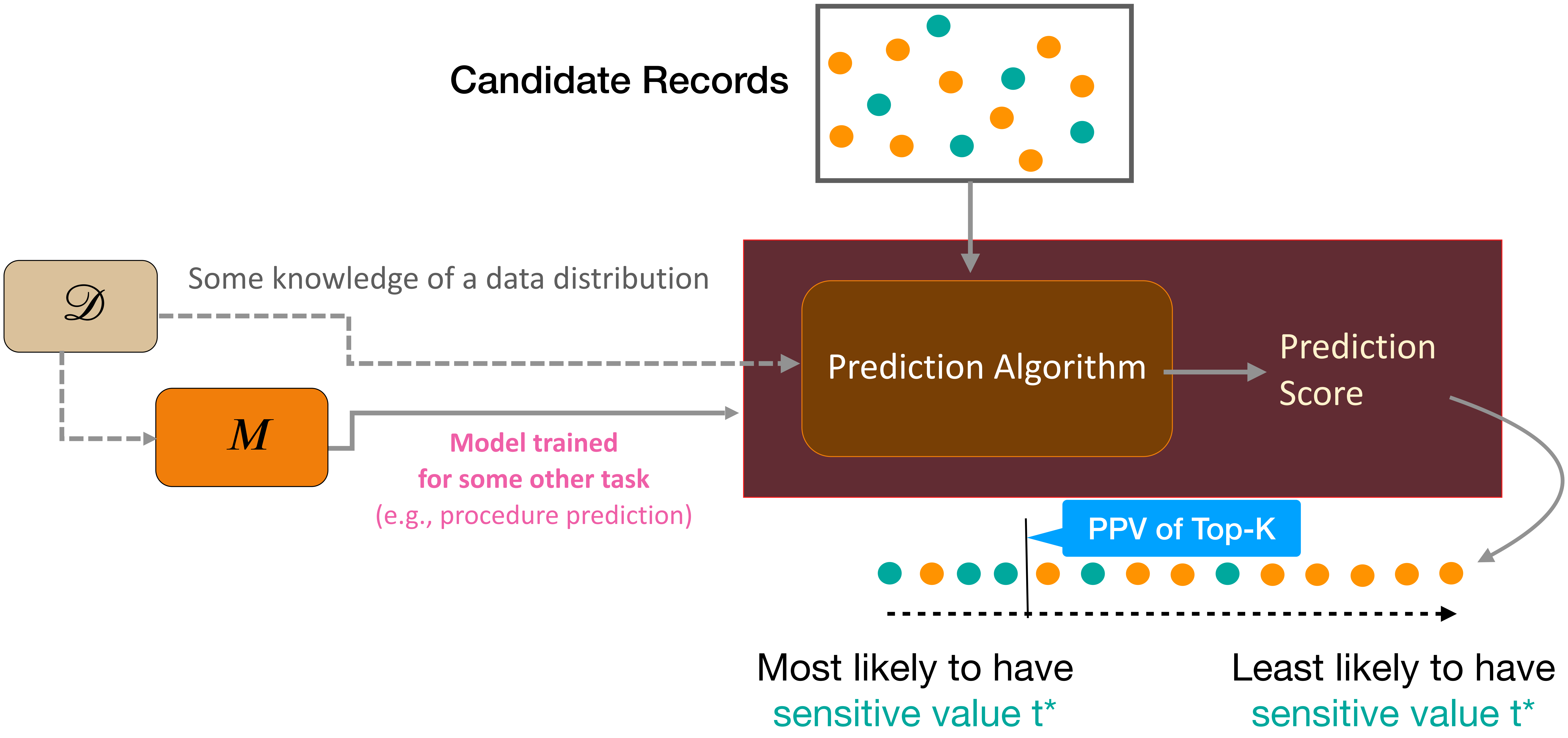
# Exploring Different Threat Settings

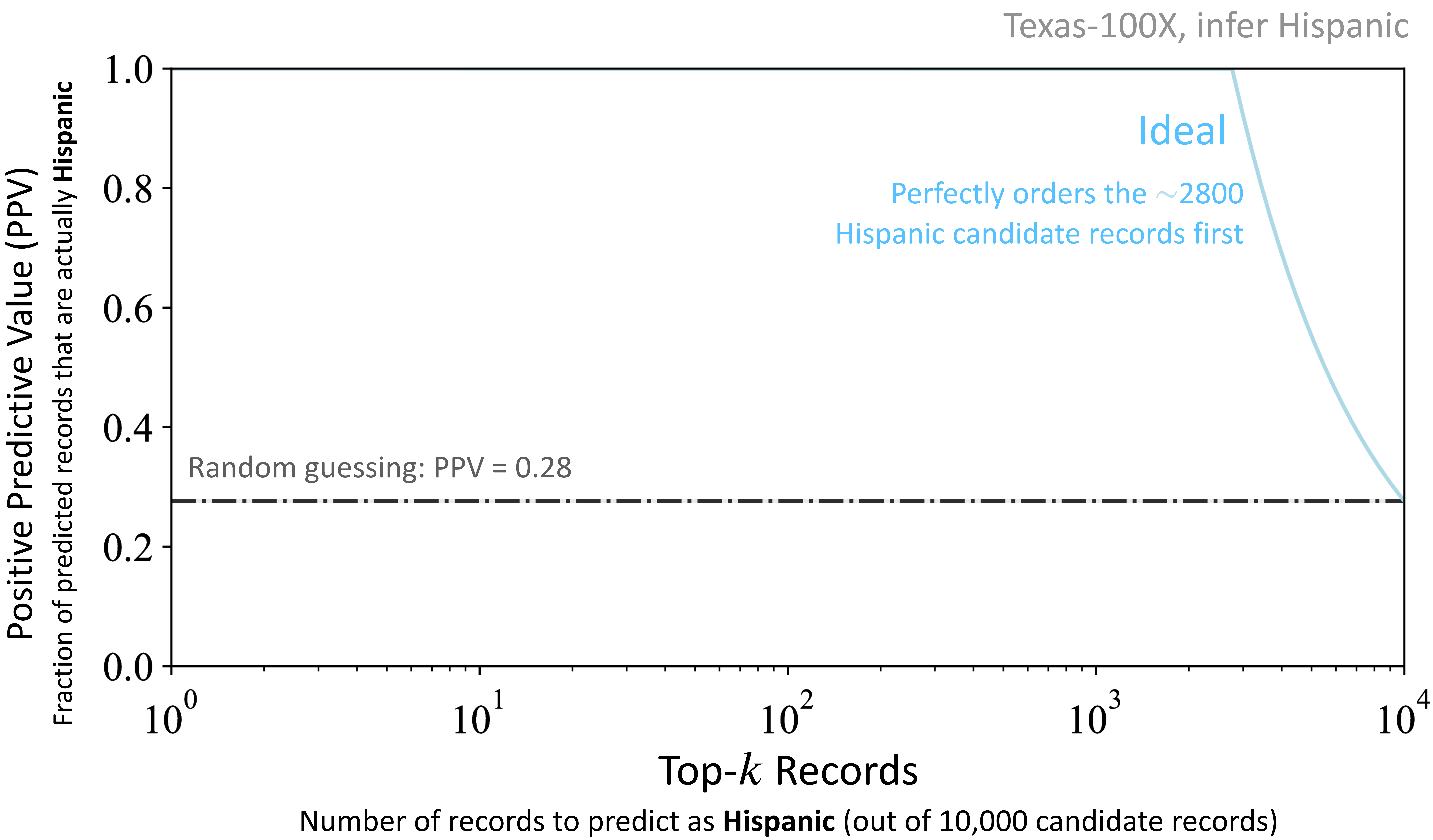
## Adversary Knows the Training Distribution

| Adversary's Data Set Size | Yes | No |
|---|---|---|
| Large | **Prior AI Attacks**<br>Fredrikson et al. [USENIX Sec 14]<br>Yeom et al. [CSF 18]<br>Mehnaz et al. [USENIX Sec 22]<br>Imputation itself does well! | ✗ |
| Small | ✗ | ✗ |

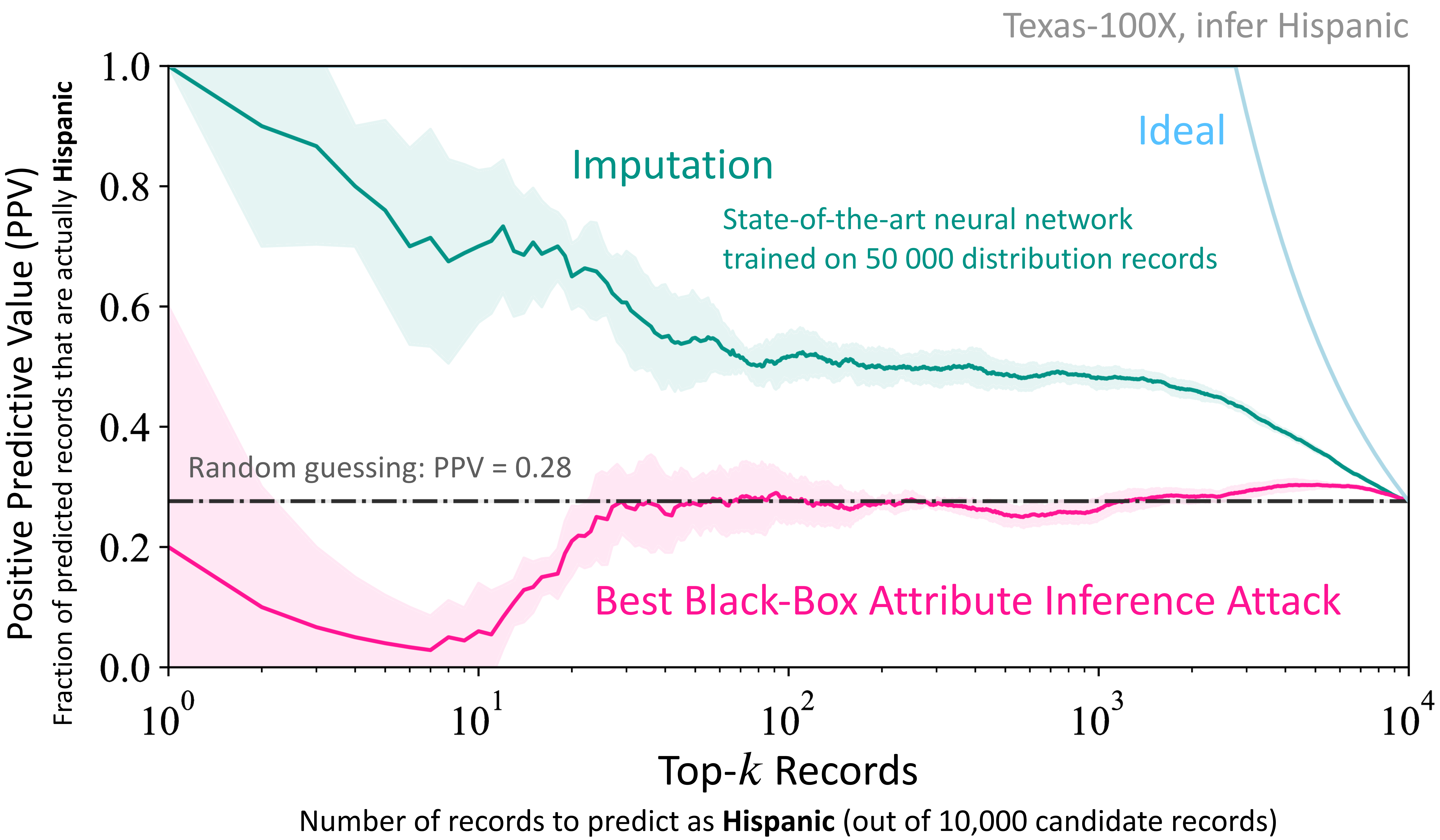
Cases where Attribute Inference > Imputation

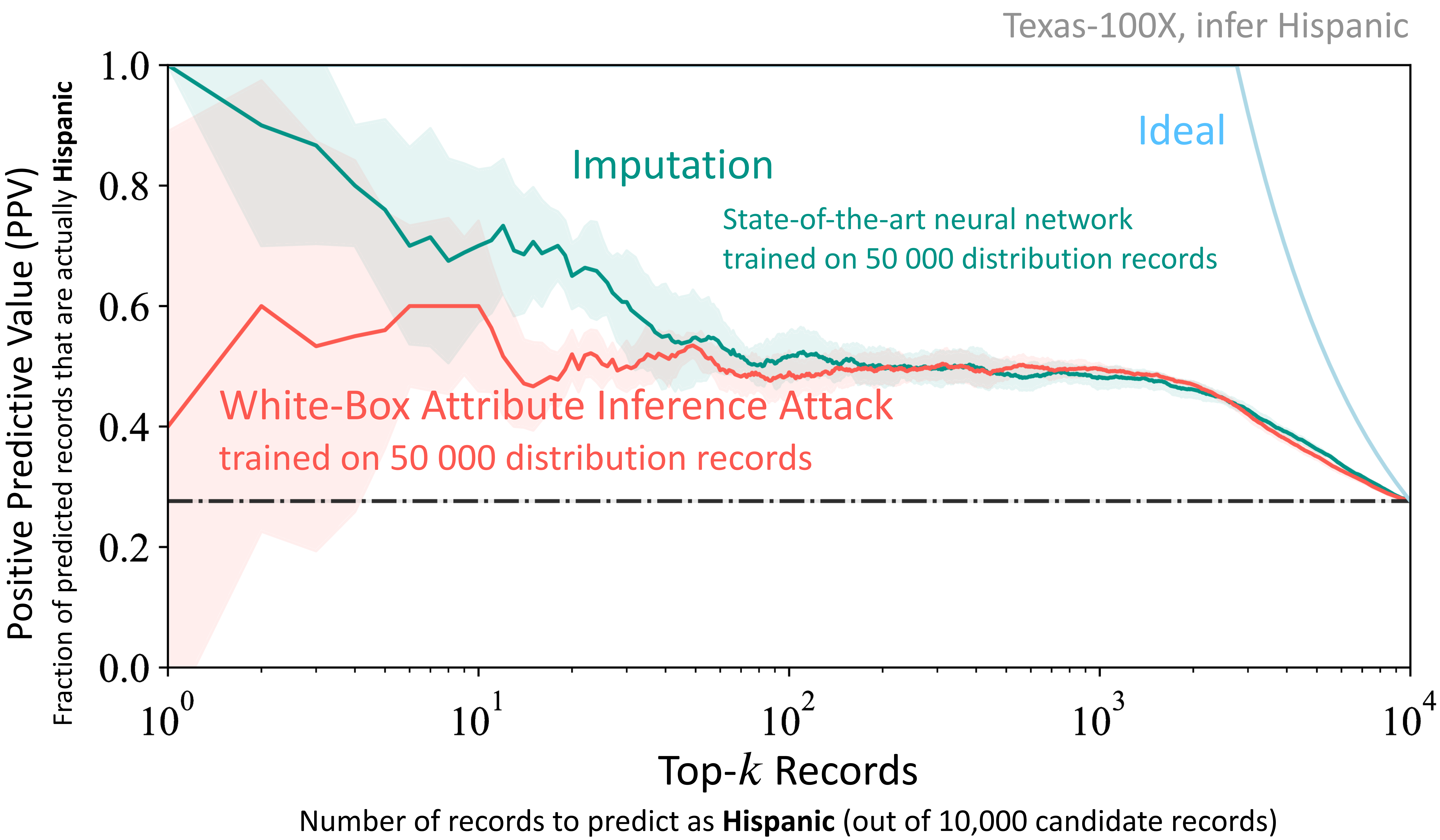
# Sensitive Value Inference Attack




Texas-100X, infer Hispanic
Ideal
Perfectly orders the ~2800
Hispanic candidate records first
Random guessing: PPV = 0.28
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)

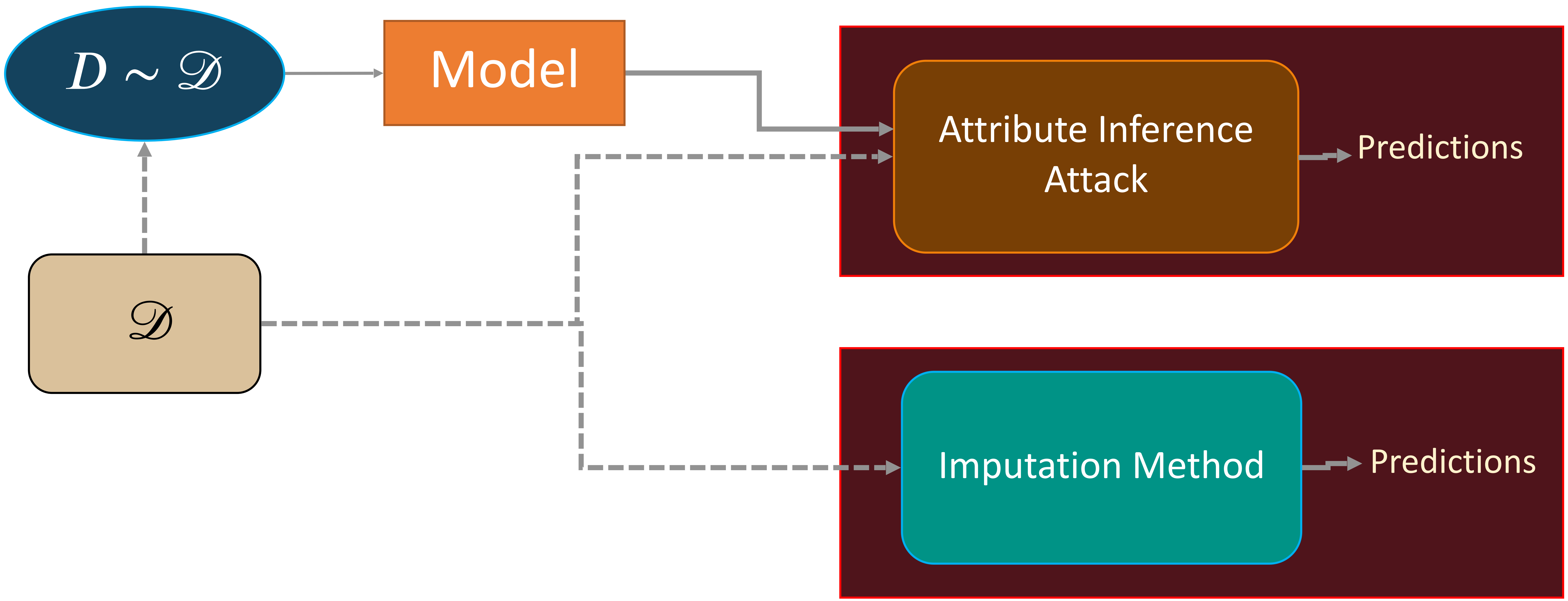

Texas-100X, infer Hispanic
Ideal
Imputation
State-of-the-art neural network
trained on 50 000 distribution records
Random guessing: PPV = 0.28
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)


Texas-100X, infer Hispanic
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
Ideal
Imputation
State-of-the-art neural network
trained on 50 000 distribution records
Random guessing: PPV = 0.28
Best Black-Box Attribute Inference Attack
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)



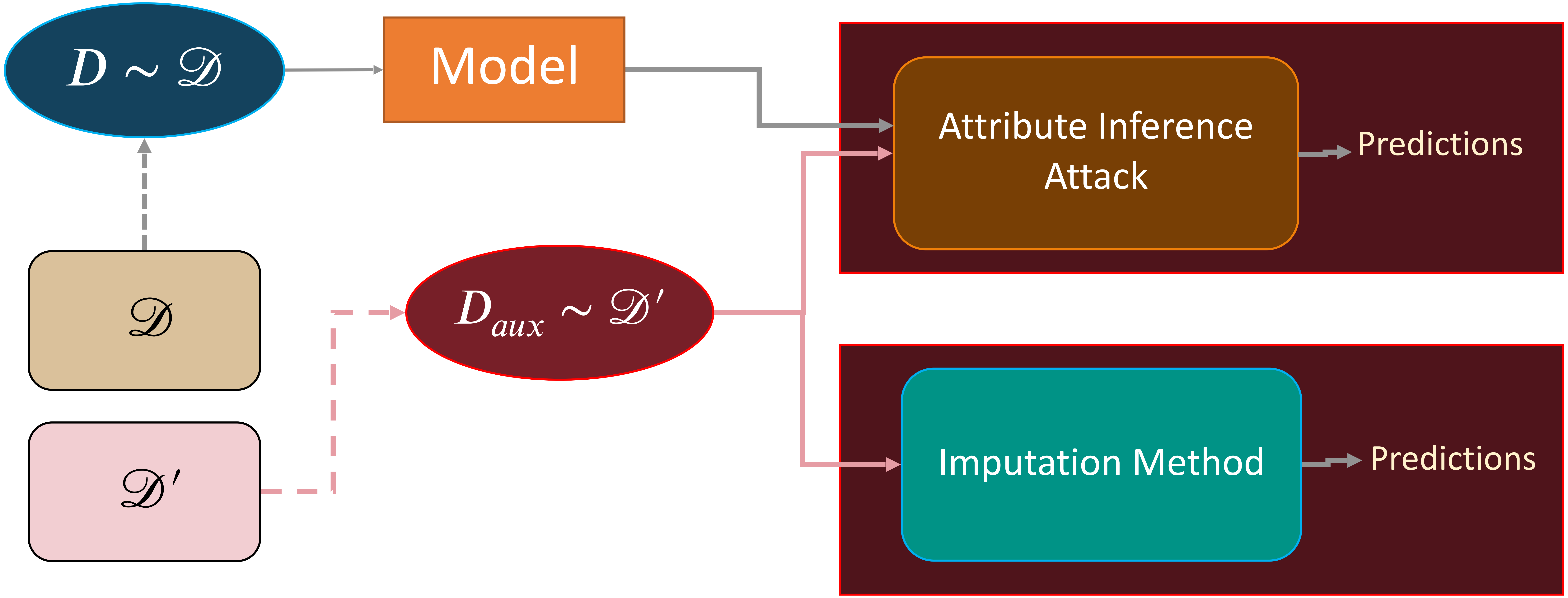
Texas-100X, infer Hispanic
Ideal
Imputation
State-of-the-art neural network
trained on 50 000 distribution records
White-Box Attribute Inference Attack
trained on 50 000 distribution records
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)

# Are there Attribute Inference attacks that matter?




If adversary has good prior knowledge of training distribution $\mathscr{D}$, unlikely that model improves ability to infer attributes

# Models Revealing Useful Information

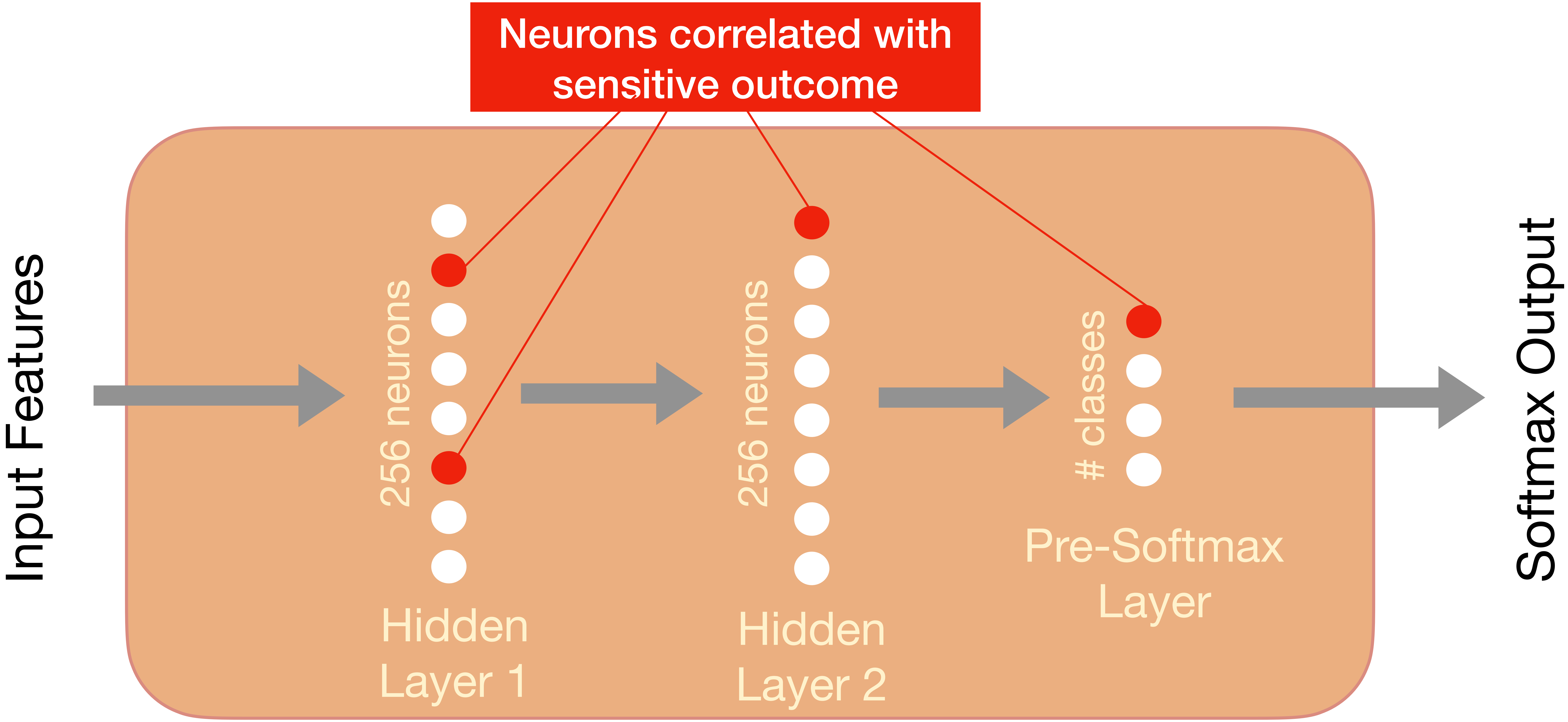



If adversary has ***limited*** prior knowledge of training distribution $\mathscr{D}$, model may improve ability to infer attributes


Texas-100X, infer Hispanic
Imputation: 50 000 distribution records
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)

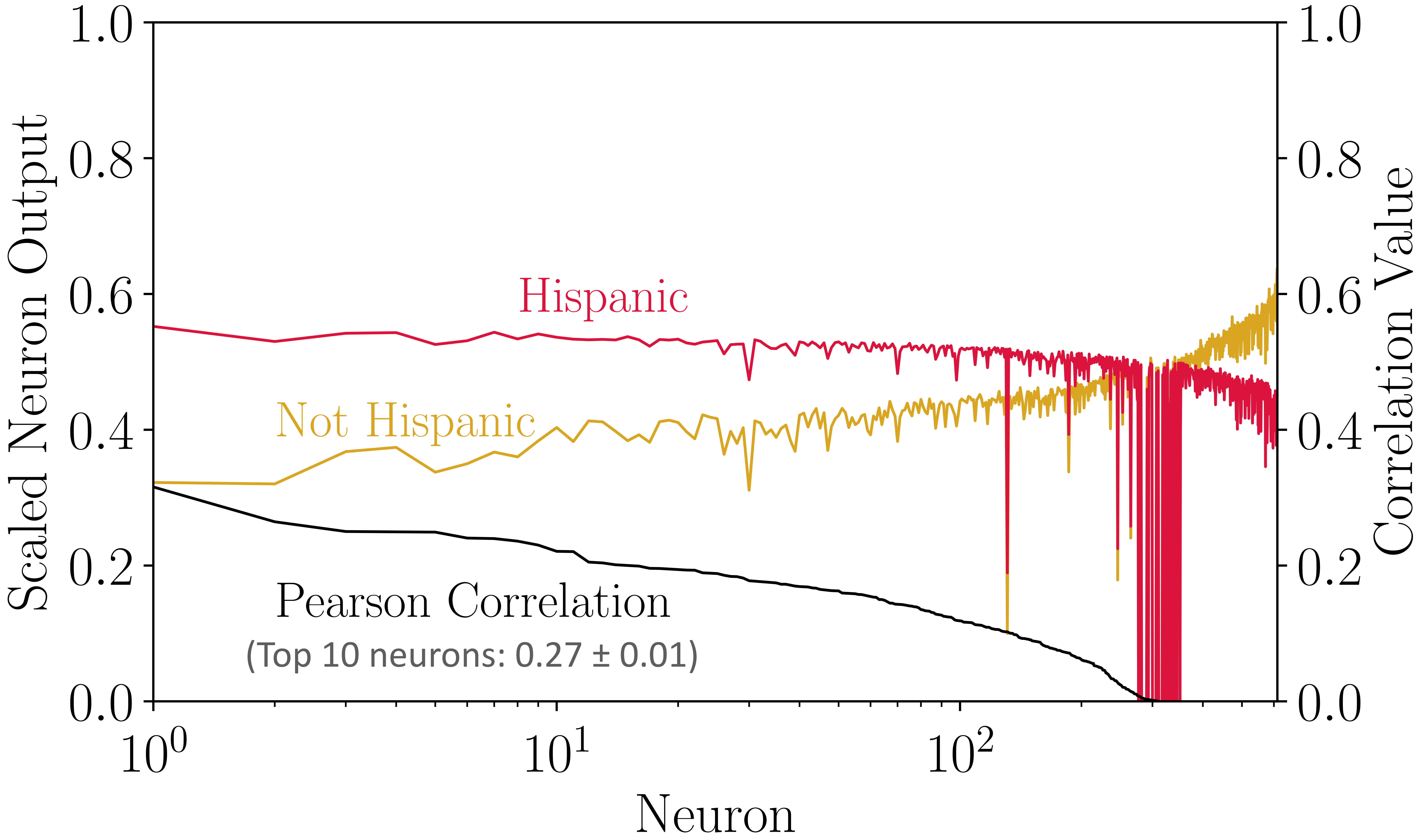
Texas-100X, infer Hispanic
Imputation: 50 000 distribution records
Imputation: 500 distribution records
Positive Predictive Value (PPV)
Fraction of predicted records that are actually **Hispanic**
1.0
0.8
0.6
0.4
0.2
0.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as **Hispanic** (out of 10,000 candidate records)

# Neuron Output Inference Attack




Neural Network Model Architecture

# Some Neurons Activate Differently



Texas-100X, infer Hispanic
Neuron Output Inference Attack
500 distribution records + Model
Imputation: 500 distribution records
Positive Predictive Value (PPV)
Fraction of predicted records that are actually Hispanic
0.0
0.2
0.4
0.6
0.8
1.0
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Top-$k$ Records
Number of records to predict as Hispanic (out of 10,000 candidate records)


| | Number of Records: | Same Distribution 5000 | Same Distribution 500 | Same Distribution 50 | Different Distribution 5000 | Different Distribution 500 | Different Distribution 50 |
|---|---|---|---|---|---|---|---|
| Texas-100X (Hispanic) | Imputation | 0.62 ± 0.05 | 0.39 ± 0.03 | 0.24 ± 0.01 | 0.44 ± 0.02 | 0.41 ± 0.05 | 0.37 ± 0.05 |
| | Neuron Output Attack | 0.49 ± 0.02 | 0.52 ± 0.03 | 0.47 ± 0.05 | 0.49 ± 0.02 | 0.49 ± 0.04 | 0.52 ± 0.07 |
| Census19 (Asian) | Imputation | 0.91 ± 0.03 | 0.25 ± 0.02 | 0.55 ± 0.06 | 0.90 ± 0.03 | 0.46 ± 0.04 | 0.42 ± 0.04 |
| | Neuron Output Attack | 0.85 ± 0.03 | 0.82 ± 0.05 | 0.67 ± 0.06 | 0.86 ± 0.05 | 0.85 ± 0.04 | 0.65 ± 0.12 |

**PPV at $k$=100 (1%)** (highlighted when model provides significant benefit)

# Differential Privacy does not Mitigate Attribute Inference Risk

Texas-100X, infer Hispanic

| | Without DP | With DP | Training | Non-Training |
|---|---|---|---|---|
| Imputation | 0.62 ± 0.05 | 0.62 ± 0.05 | 0.62 ± 0.05 | 0.63 ± 0.02 |
| Neuron Attack | 0.49 ± 0.02 | 0.49 ± 0.03 | 0.49 ± 0.03 | 0.48 ± 0.02 |

Results for models trained with $(\epsilon = 1, \delta = 10^{-5})$–DP

No significant differences between non-DP models, and no differences between predictions for candidates from training and non-training data

# Conclusion

**Attribute Inference Attacks *are doing* Imputation**

**Privacy Risk when the Distribution is not Public**

**Presenter:**

Bargav Jayaraman

*bj4nq@virginia.edu*

University of Virginia

Code Repository:

https://github.com/bargavj/EvaluatingDPML