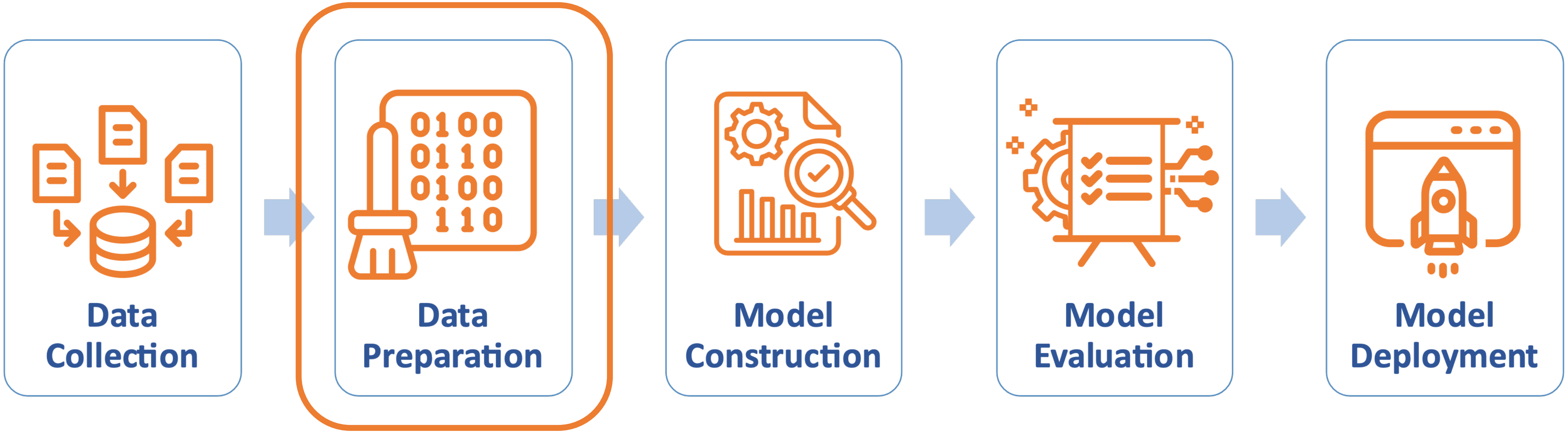
When Can We Ignore Missing Data in Model Training?

Cheng Zhen, Amandeep Singh Chabada, Arash Termehchy

Oregon State University

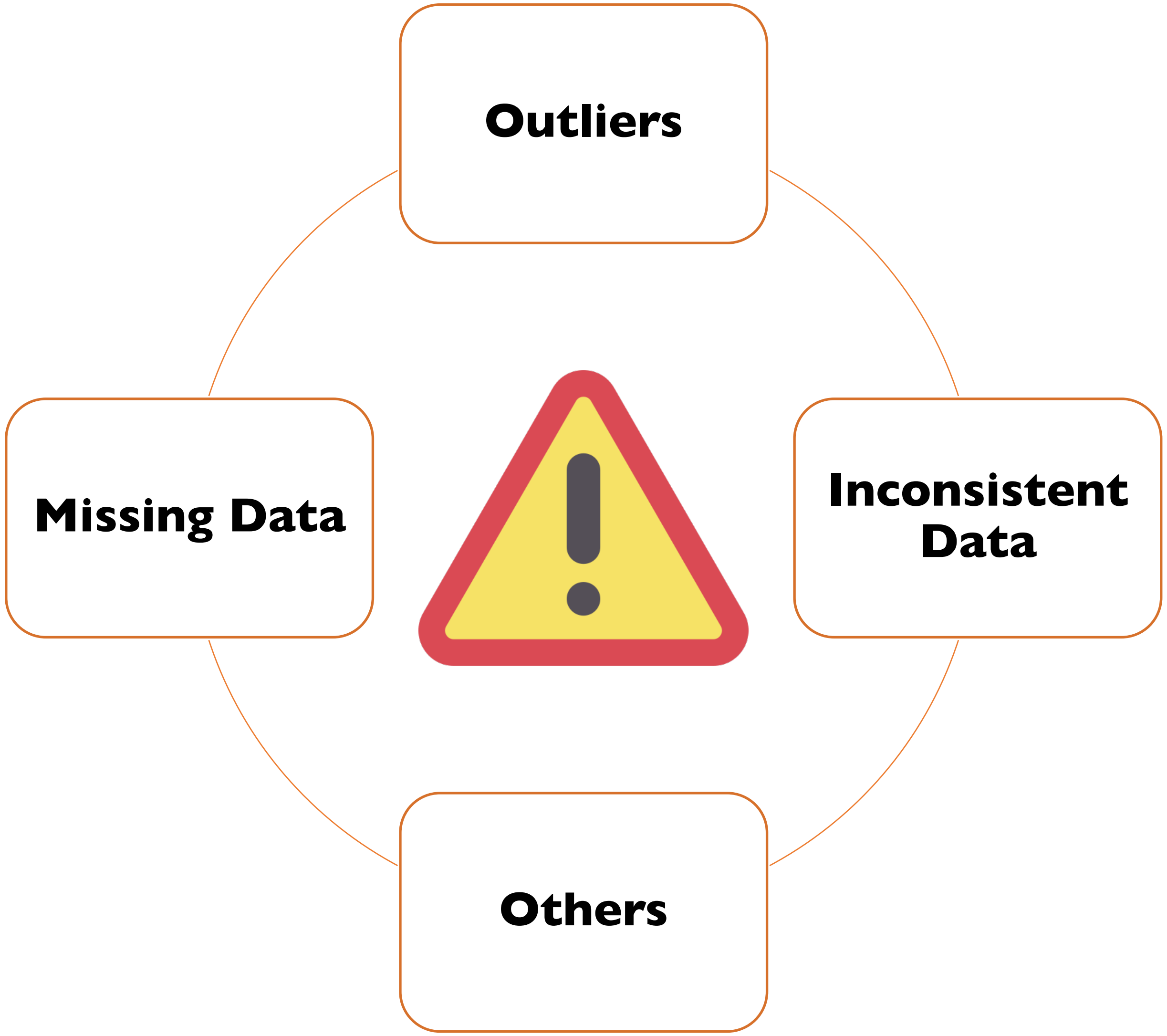
Machine Learning Pipeline



Most data scientists spend ~ 80% of their time preparing data for ML

Data Preparation

Clean the Problems in Raw Data



Example of Raw Data Problems

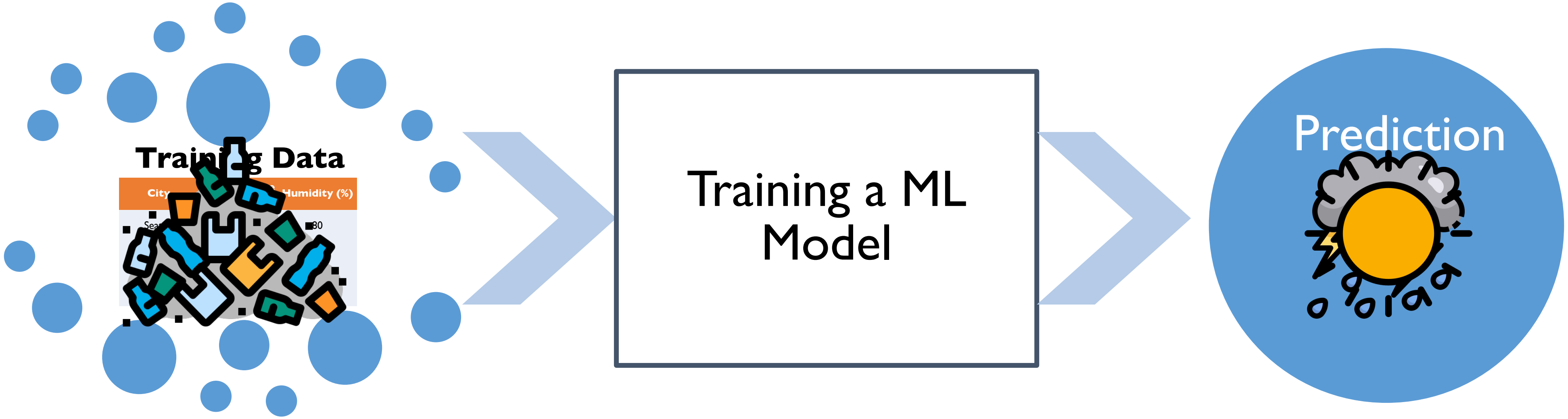
City	Temperature (F)	Humidity (%)	Rain (1) or no rain (-1)
Seattle	65	80	1
Portland	Null	30	-1
San Francisco	54	-9999	-1
San Diego	60	67	1
San Diego	70	67	1

Missing Data

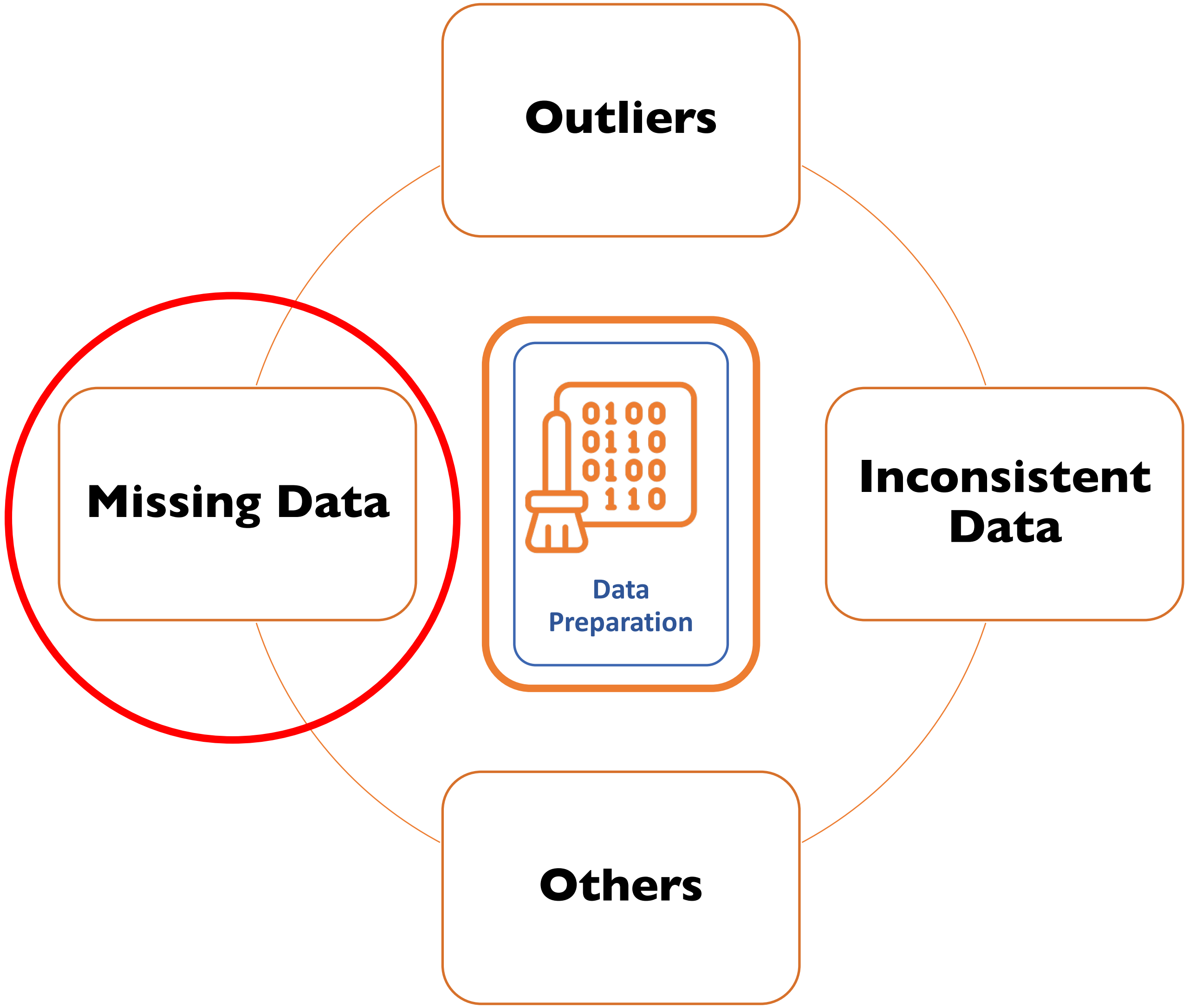
Inconsistent Data

Outliers

Wrong Result from Raw Data Problems



Our Research Focuses on Missing Data



Deleting records with missing values

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	**Null**	30
San Francisco	54	90

City	Temperature (F)	Humidity (%)
Seattle	65	80
San Francisco	54	90

- Loss of valuable information
- Might introduce bias

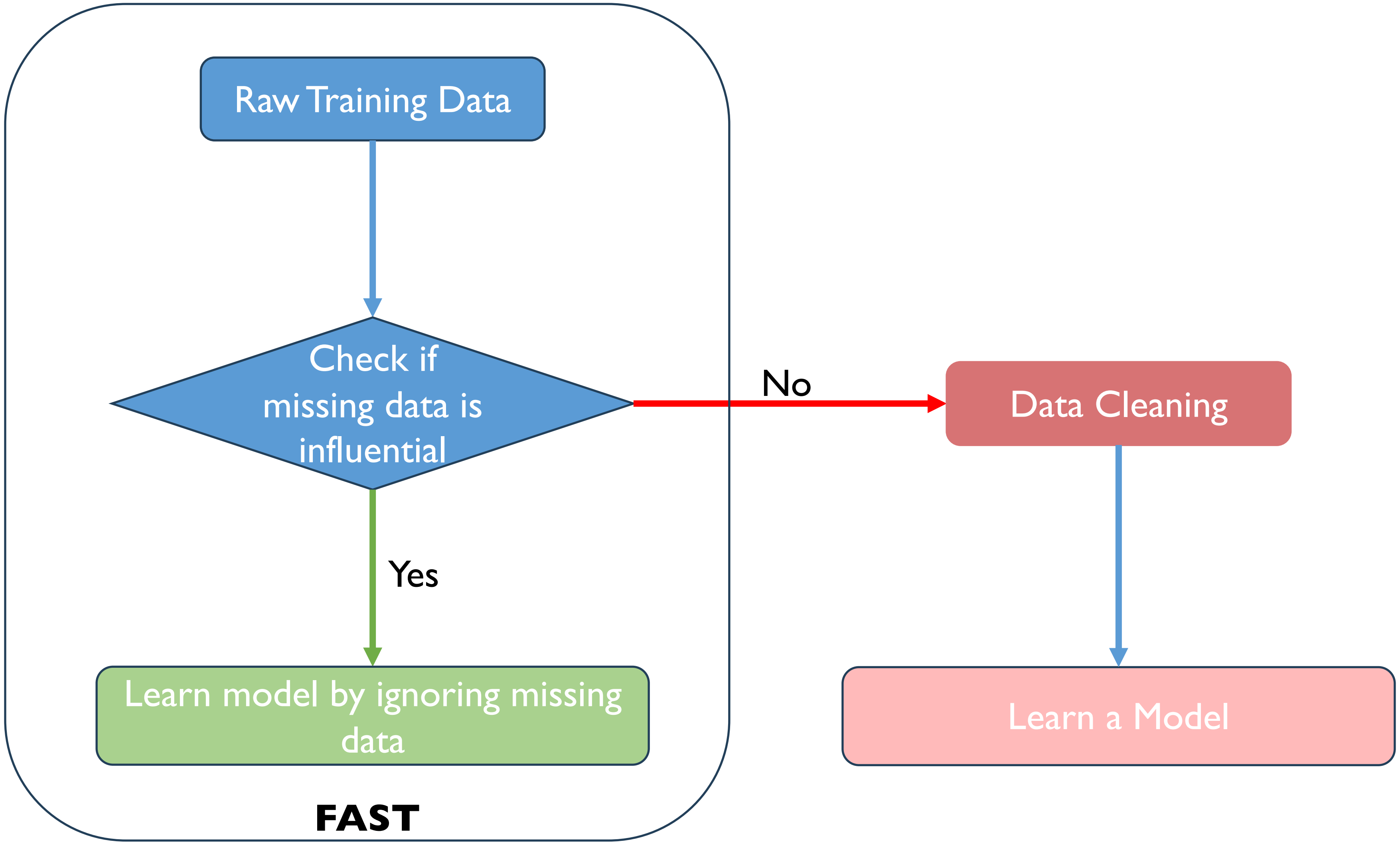
Data imputation

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	**Null**	30
San Francisco	54	90

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	60	30
San Francisco	54	90

- High Cost - Development & Time
- Not clear which imputation method is accurate

What if Missing Data is not Influential to Model?

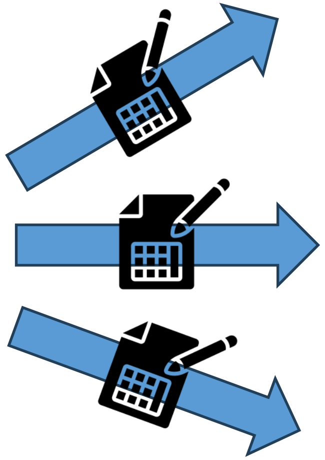


To Better Understand the Scenario

Define "repair" for missing data:

A complete data set that replaces "Null" values in raw data with specific values

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	**Null**	30
San Francisco	54	90



City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	60	30
San Francisco	54	90

Repair 1

•••

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	80	30
San Francisco	54	90

Repair ∞

When Imputation Makes No Difference on Models

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	60	30
San Francisco	54	90

Repair 1

Model training

…

City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	**80**	30
San Francisco	54	90

Repair ∞

Model training

They share the same model!

The same model is learned from all repairs → *Imputation is unnecessary*

Prior Work Detecting Unnecessary Data Cleaning

- DLearn *(Learning over dirty data without cleaning, SIGMOD 2020)*
  - Learn models that represent patterns over all possible clean repairs
  - 👎 Limited to relational models

- CPClean *(Nearest neighbor classifiers over incomplete information: from certain answers to certain predictions, VLDB 2021)*
  - Find models that predict the same result for all repairs in the validation set
  - 👎 Limited to KNN, and vulnerable to small/dirty validation set

OUR NEW APPROACH

GOAL

Develop a more generalizable method to determine the conditions where data cleaning is unnecessary for model training

Certain Models

A model that minimizes training loss for all repairs.

--- *"certain model is certainly optimal"*

Important Terms

- **Feature Input (X), and label output (y)**
- **Model (w):** Parameters that characterize the relationship between **X** and **y**
- **Loss Function:** Measures how much the model predictions deviate from the actual data

$$L(f(\mathbf{X}, \mathbf{w}), \mathbf{y})$$

- **Model Training (w*):** Finds the optimal model that minimizes training loss.

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in w} L(f(\mathbf{X}, \mathbf{w}), \mathbf{y})$$

Formally Defining Certain Models

A model $\boldsymbol{w}^*$ is a certain model if:

Loss Function

$$\underbrace{\forall \boldsymbol{X}^r \in \boldsymbol{X}^R, \forall \boldsymbol{y}^r \in \boldsymbol{y}^R}_{\text{Repairs}}, \underbrace{\boldsymbol{w}^* = \boldsymbol{arg}\min_{\boldsymbol{w} \in \boldsymbol{\mathcal{W}}} \boldsymbol{L}(\boldsymbol{f}(\boldsymbol{X}^r, \boldsymbol{w}), \boldsymbol{y}^r)}_{\text{Model Training}}$$

Repairs

Model Training

How to Check Certain Models

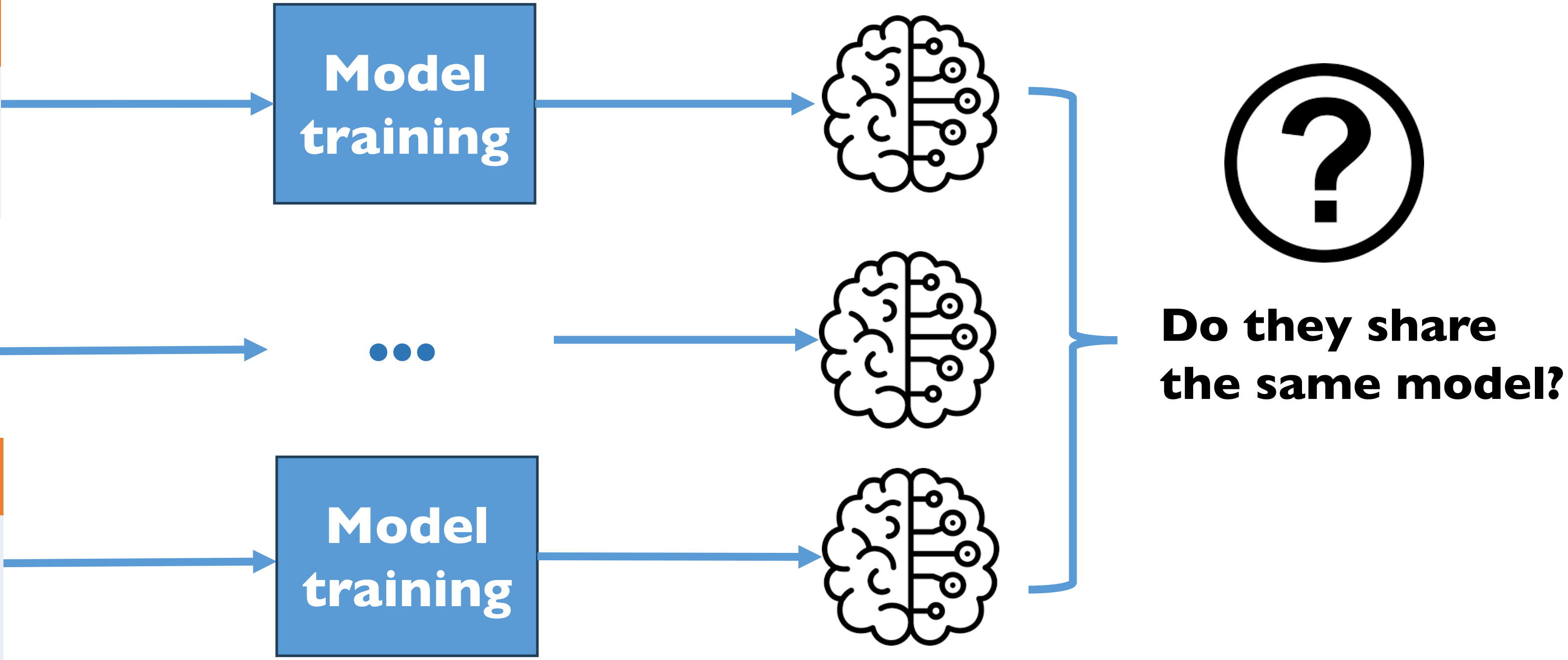
City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	60	30
San Francisco	54	90

Repair 1

…

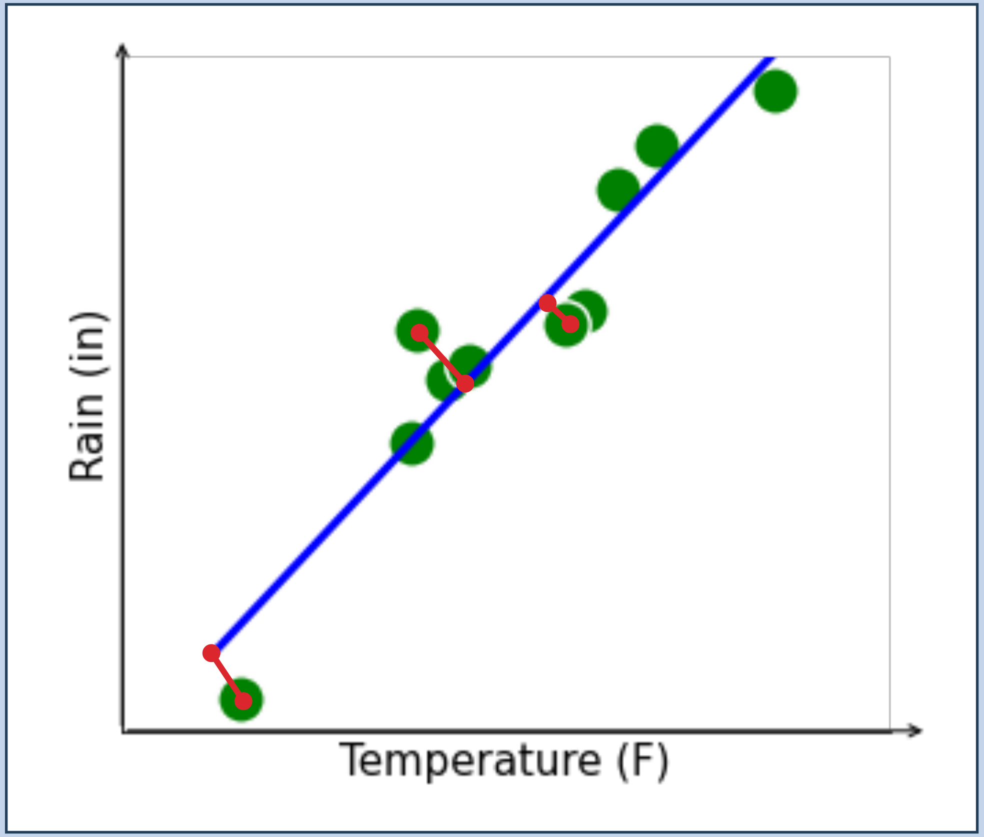
City	Temperature (F)	Humidity (%)
Seattle	65	80
Portland	**80**	30
San Francisco	54	90

Repair ∞



This is incredibly slow because there are often an infinite number of repairs

Certain Models for Linear Regression



Model Formulation

$$y = Xw + b$$

Loss function for Linear Regression

$$L(f(X, w), y) = \|Xw - y\|_2^2$$

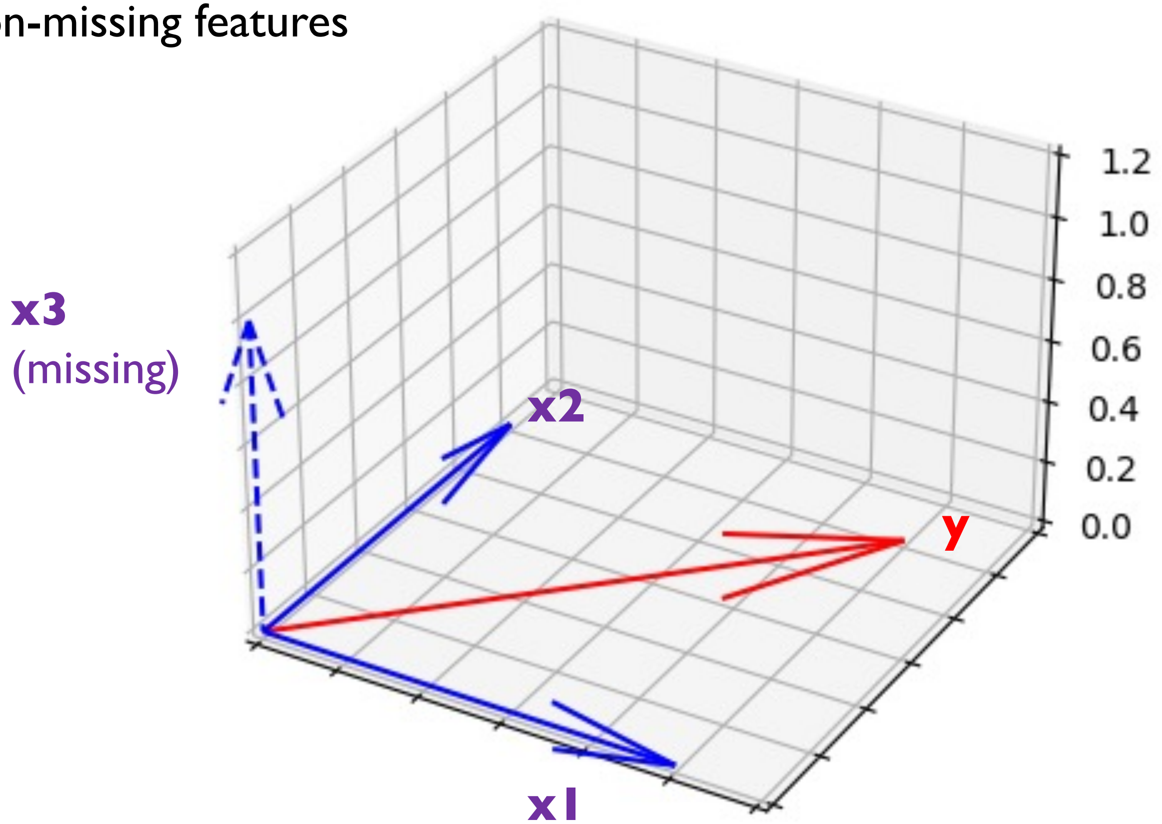
Certain Model

$$\forall \boldsymbol{X^r} \in \boldsymbol{X^R}, \boldsymbol{w}^* = \boldsymbol{arg} \min_{w \in \mathcal{W}} \|\boldsymbol{X^r w} - \boldsymbol{y}\|_2^2$$

Conditions for Certain Models Existing

x1	x2	x3	y
1	0	0	1
0	1	0	1
0	0	**Null**	0

x3 ⊥ the regression residue between the label and non-missing features



x3 does not contribute to loss minimization in any repair

Check the Orthogonality without Materializing all Repairs

$\boldsymbol{t}$: regression residue between the label and non-missing features

$\forall \boldsymbol{X}^r \in \boldsymbol{X}^R$, missing feature $\cdot\, \boldsymbol{t} = 0$ → Orthogonality Holds → Certain Model Exists

Theorem 1 ⇕ Checking two conditions

1) For null values, the corresponding inner product values are zeros

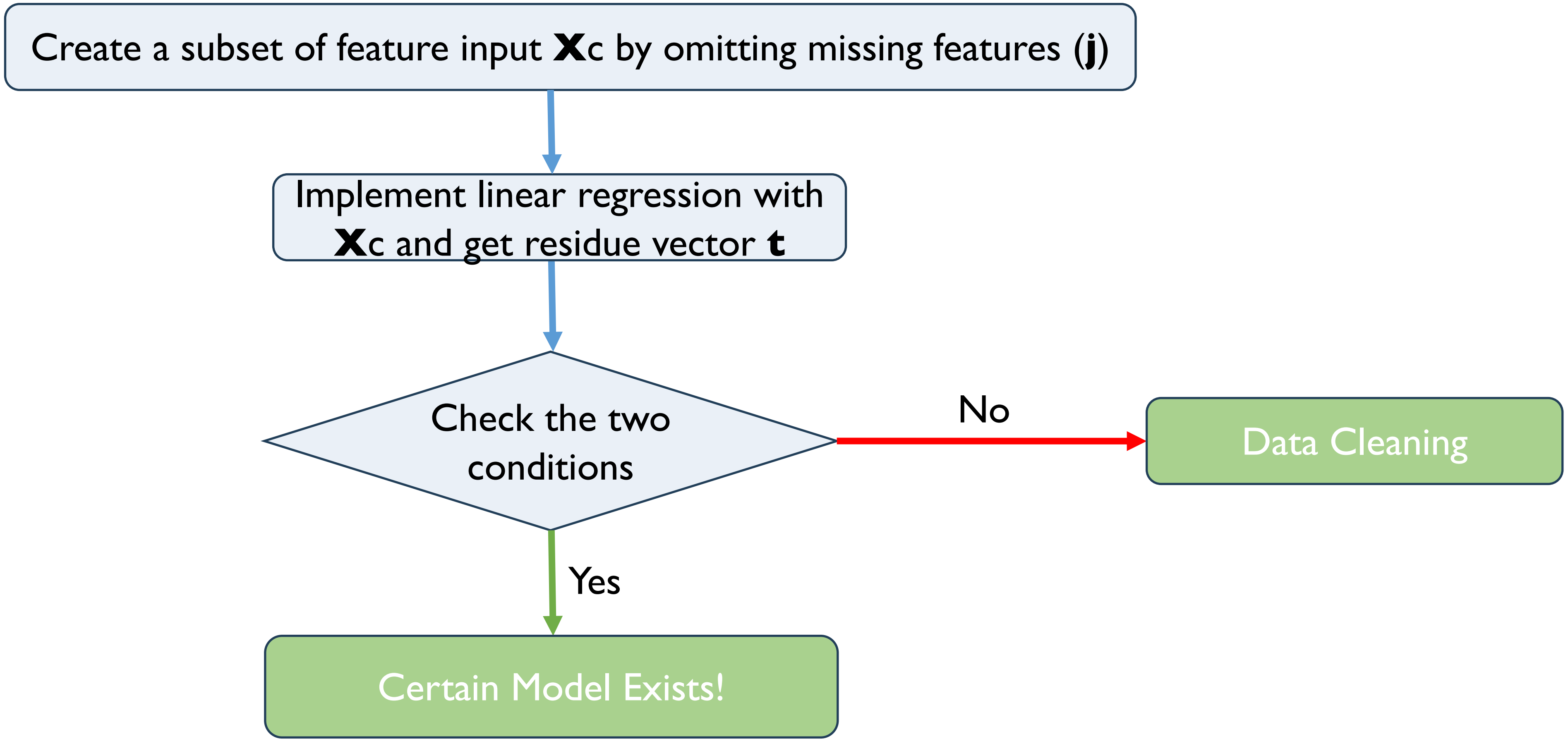
2) The sum of non−missing inner product components is zero

Example

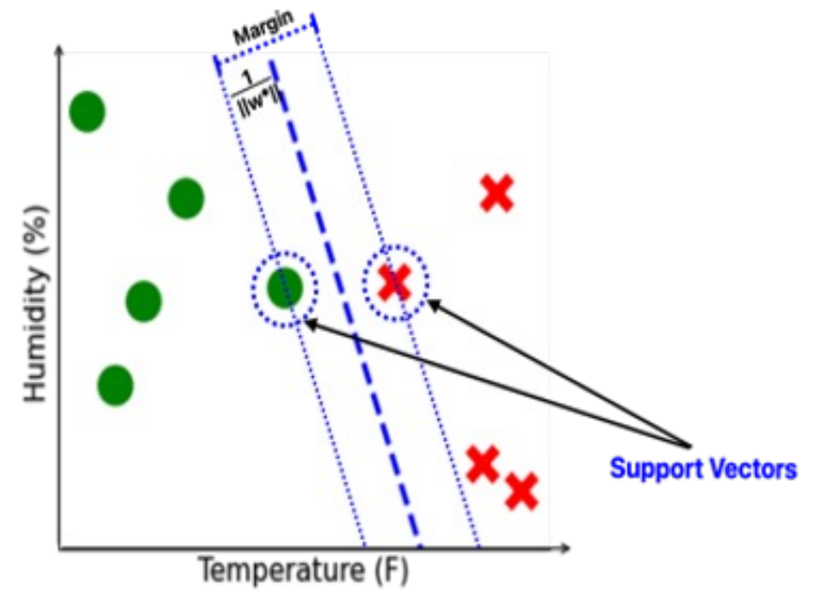
Missing Feature		t
1		-1
1	$\cdot$	1
Null		0

$$1 \cdot 1 + 1 \cdot -1 = 0$$

Efficient Algorithm to Check Certain Models



Defining Certain Models for Support Vector Machines(SVM)



Learn the decision boundary given by

$$w^T e = 0$$

Loss function for SVM

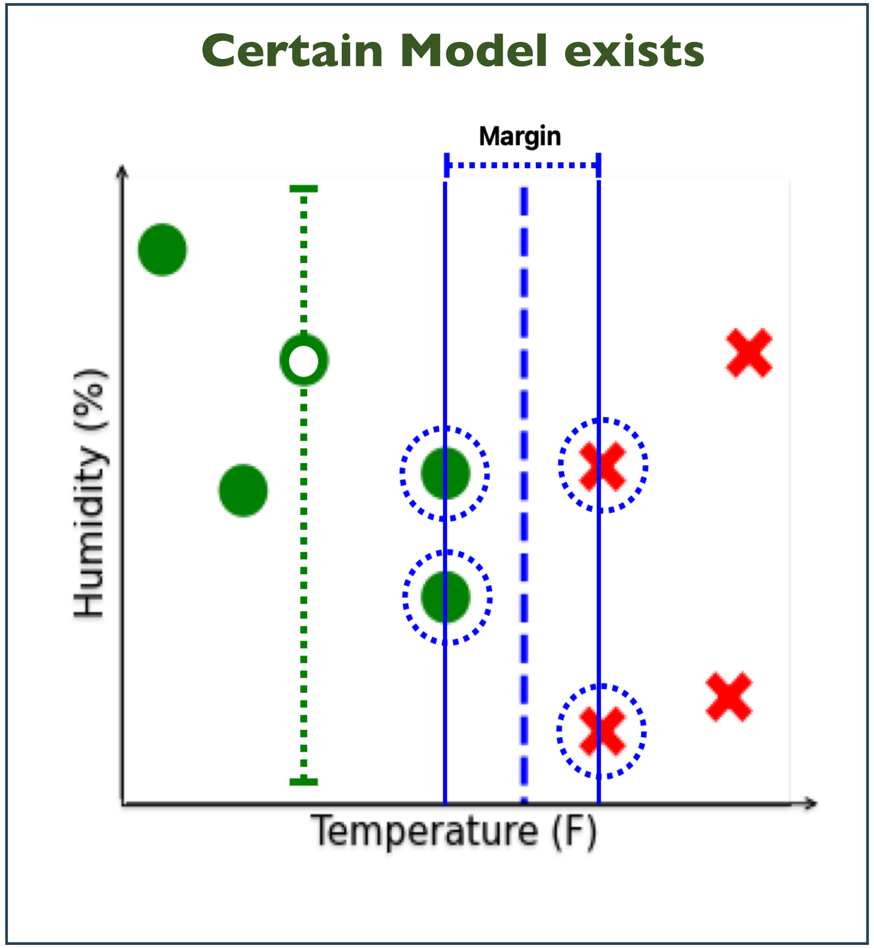
$$L(f(X,w),y) = \frac{1}{2}\|w\|_2^2 + \mathrm{C}\sum_{i=1}^{n} max\{0,1 - y_i w^T e_i\}$$

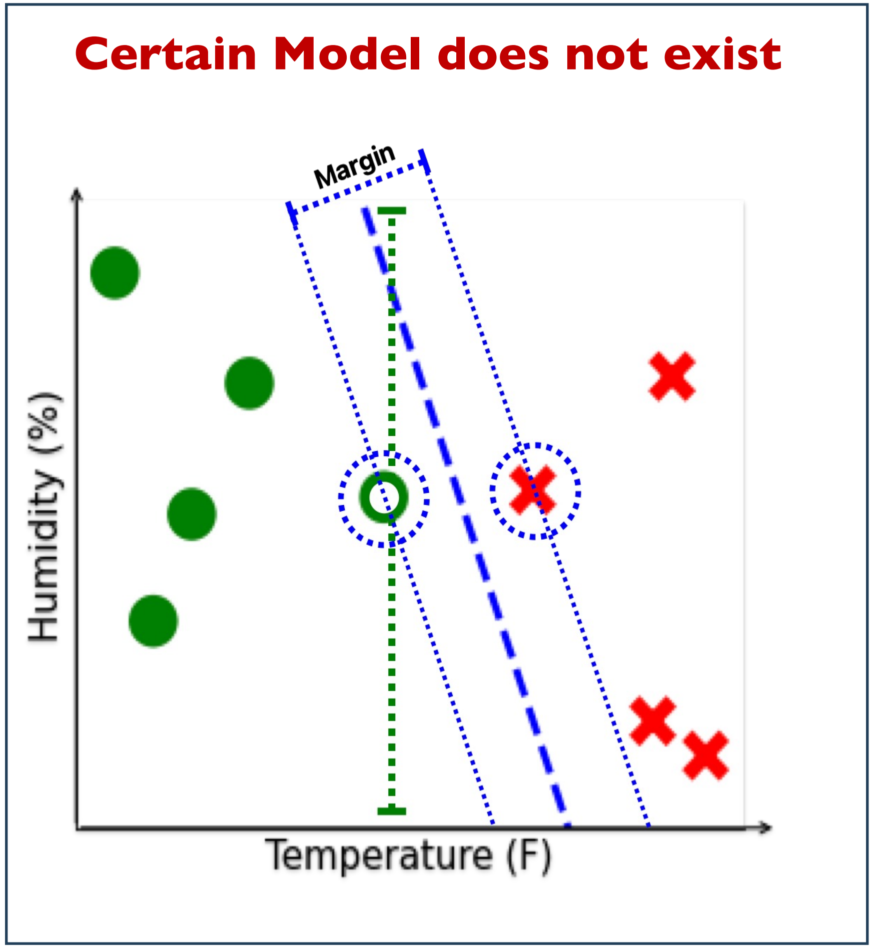
Certain Model

$$\forall \boldsymbol{X^r} \in \boldsymbol{X^R},$$

$$\boldsymbol{w}^* = \boldsymbol{arg}\min_{w \in \mathcal{W}} \frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{w}\|_2^2 + \mathbf{C}\sum_{i=1}^{n} \boldsymbol{max}\{\mathbf{0}, \mathbf{1} - \boldsymbol{y_i w^T e_i}\}$$

Conditions for Certain Models Existing



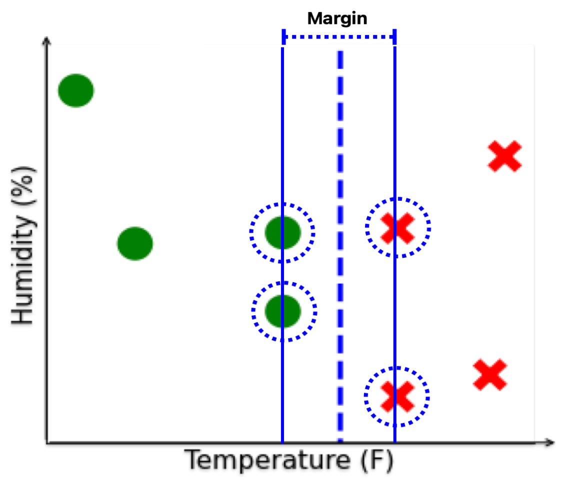


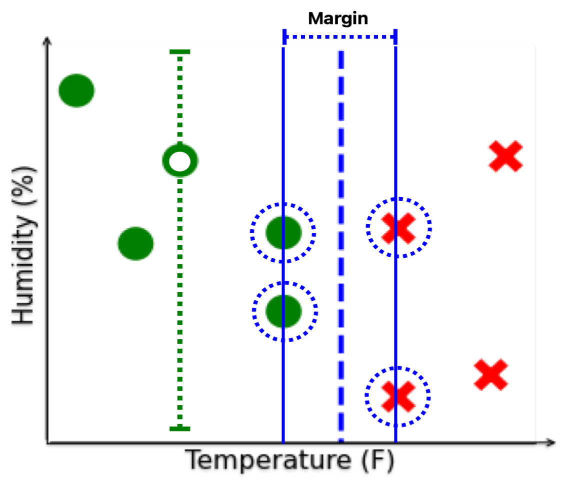
Missing training example is not a support vector in any repair => certain model exists

Check Support Vectors without Materializing all Repairs

Model **w'** trained without missing training examples

Check two conditions





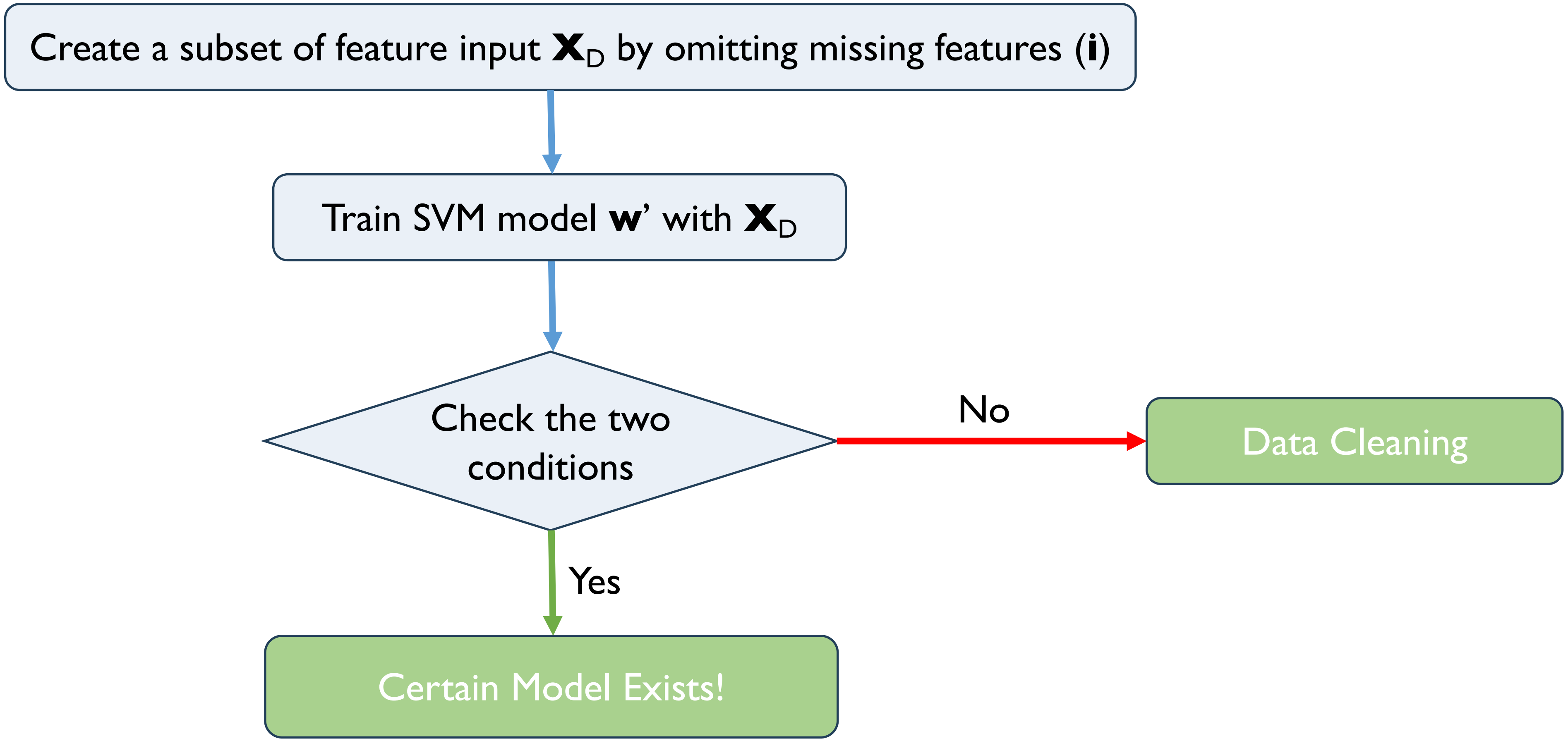
Theorem 2

1. The decision boundary in **w'** is parallel to the repair space

2. The missing example is outside of the maximum margin in **w'**

Certain Model Exists

Efficient Algorithm to Check Certain Models



EXPERIMENTAL RESULTS

Baseline: ActiveClean

(Activeclean: Interactive data cleaning for statistical modeling, VLDB 2016)

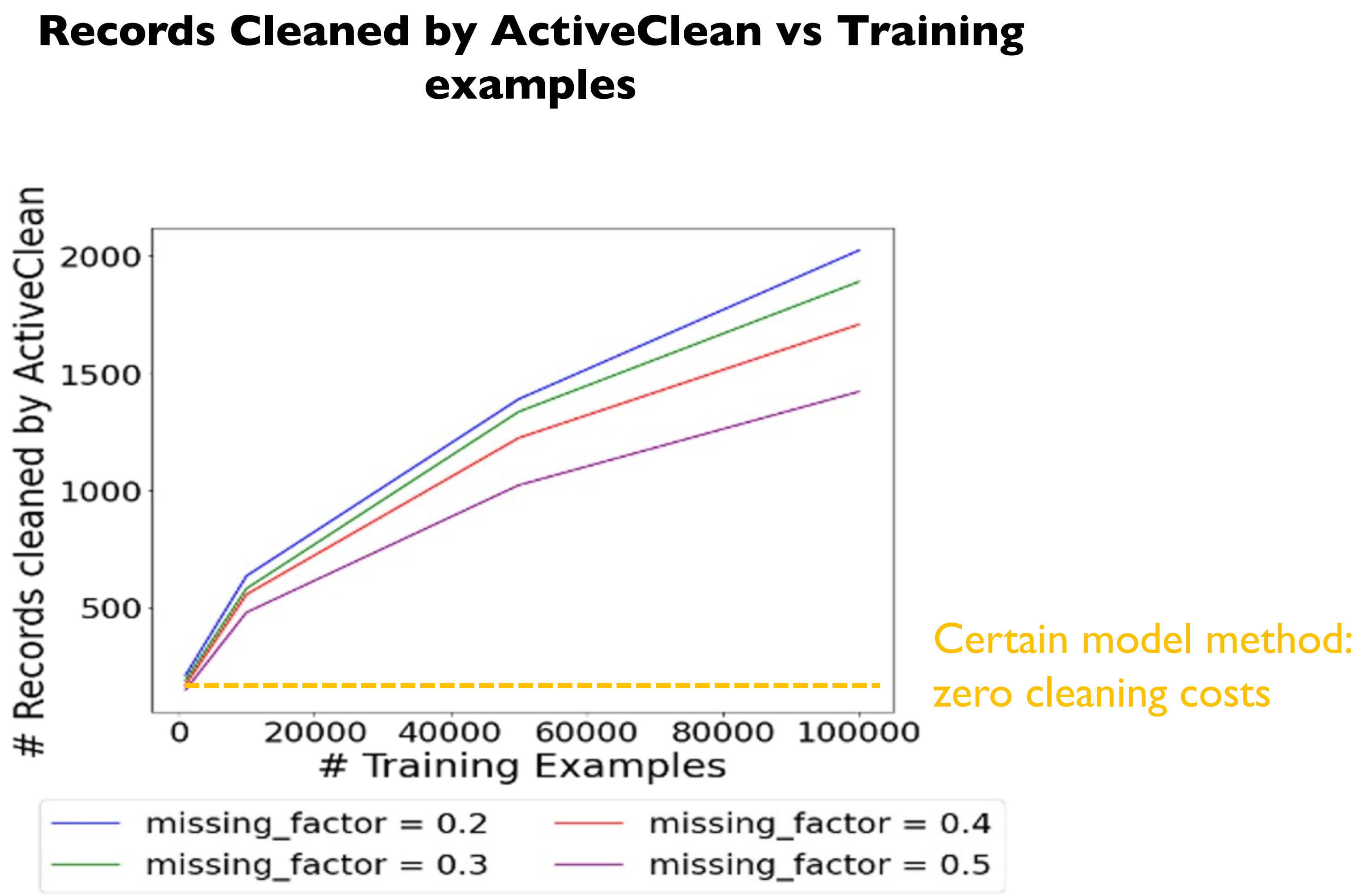
Reduce the effort of data cleaning for model training

- Prioritizes cleaning of training examples with large model gradients.
- Stops cleaning at the convergence of Stochastic Gradient Descent.

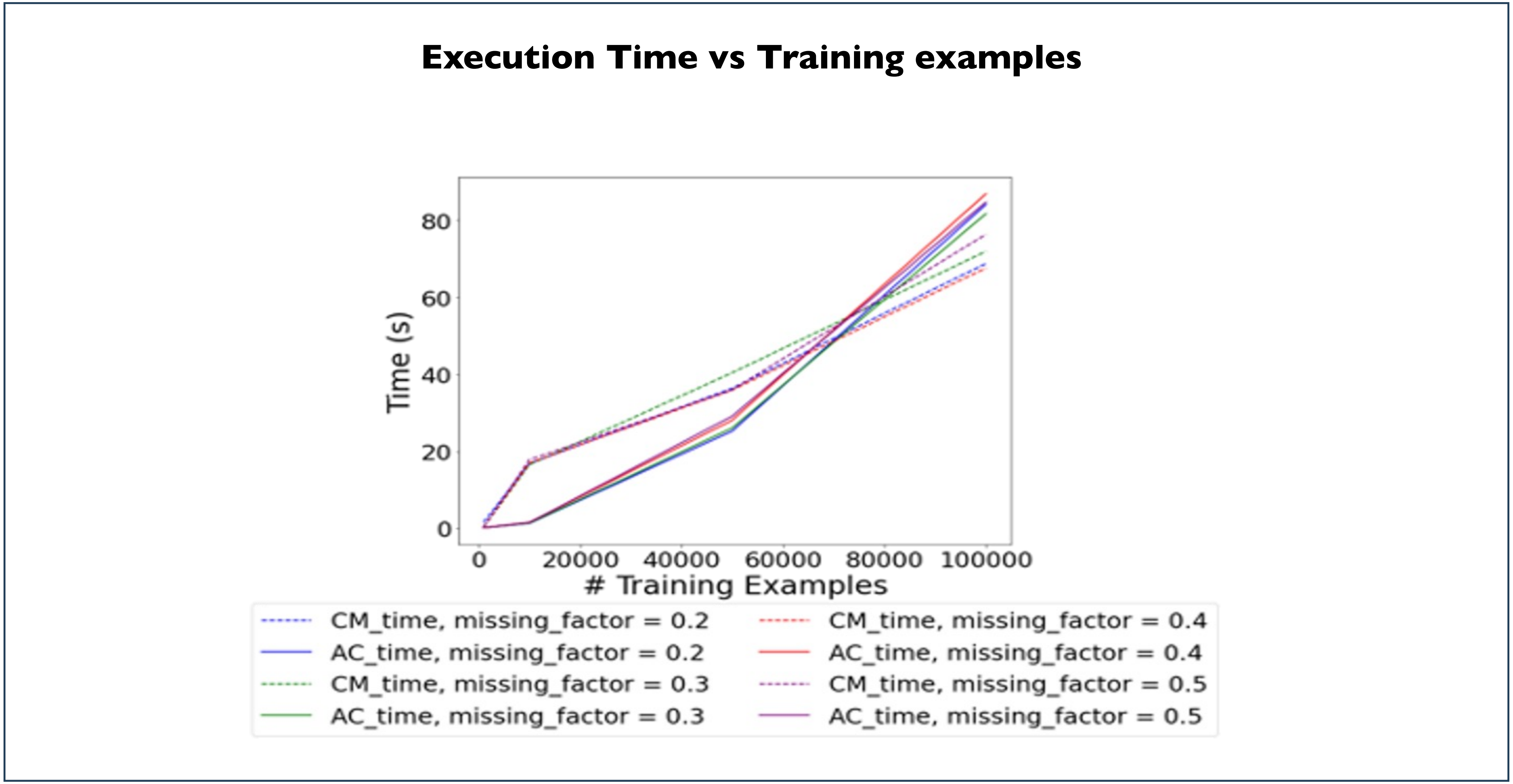
Experimental Setup - Certain Models

- **Dataset Details**
  - Synthetically generated
  - #Records: 1,000-100,000
  - #Features: 5,000
  - Missing Factor: 0.2-0.5
  - 80%-20% Train-Test split
  - Missingness introduced by random imputation

Cleaning Cost Savings for Linear Regression



Execution Time Comparison for Linear Regression



Certain Model vs ActiveClean

Reduced Cleaning Efforts

Comparable Accuracy Performance

Similar Computational Costs

CONCLUSION AND FUTURE WORK

Contributions

- Introduced a new condition of unnecessary data cleaning for statistical learning
- Offered efficient algorithms to check the condition for Linear Regression and SVM.
- Experimentally demonstrated the algorithms’ performance

Ongoing Work

- Extending efficient implementation to other ML models

  --- DNN, kernel methods, etc.

- Certain model may not exist in many data sets

  --- A more relaxed condition than the exact optimality.

THANK YOU

Oregon State University