Galaxy: Towards Scalable and Interpretable Explanation on High-dimensional and Spatio-Temporal Correlated Climate Data

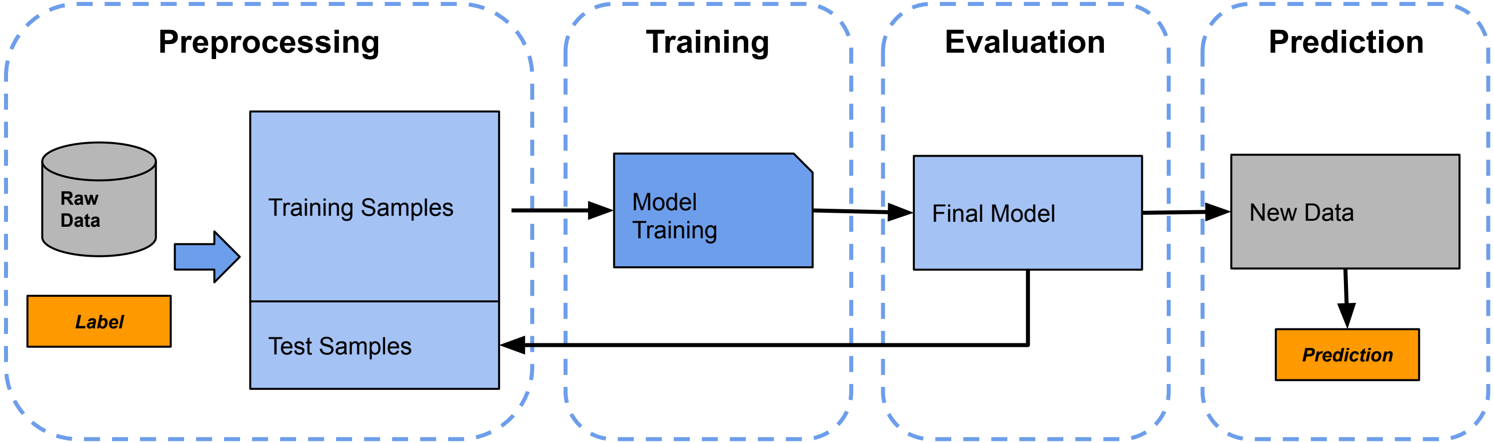
Yong Zhuang[1], David Small[2], Xin Shu[1], Kuiyu[3], Shafiqul Islam[2], Wei Ding[1]

[1]Department of Computer Science, University of Massachusetts Boston
[2]Department of Civil and Environmental Engineering, Tufts University
[3]School of Computer and Information, Hefei University of Technology

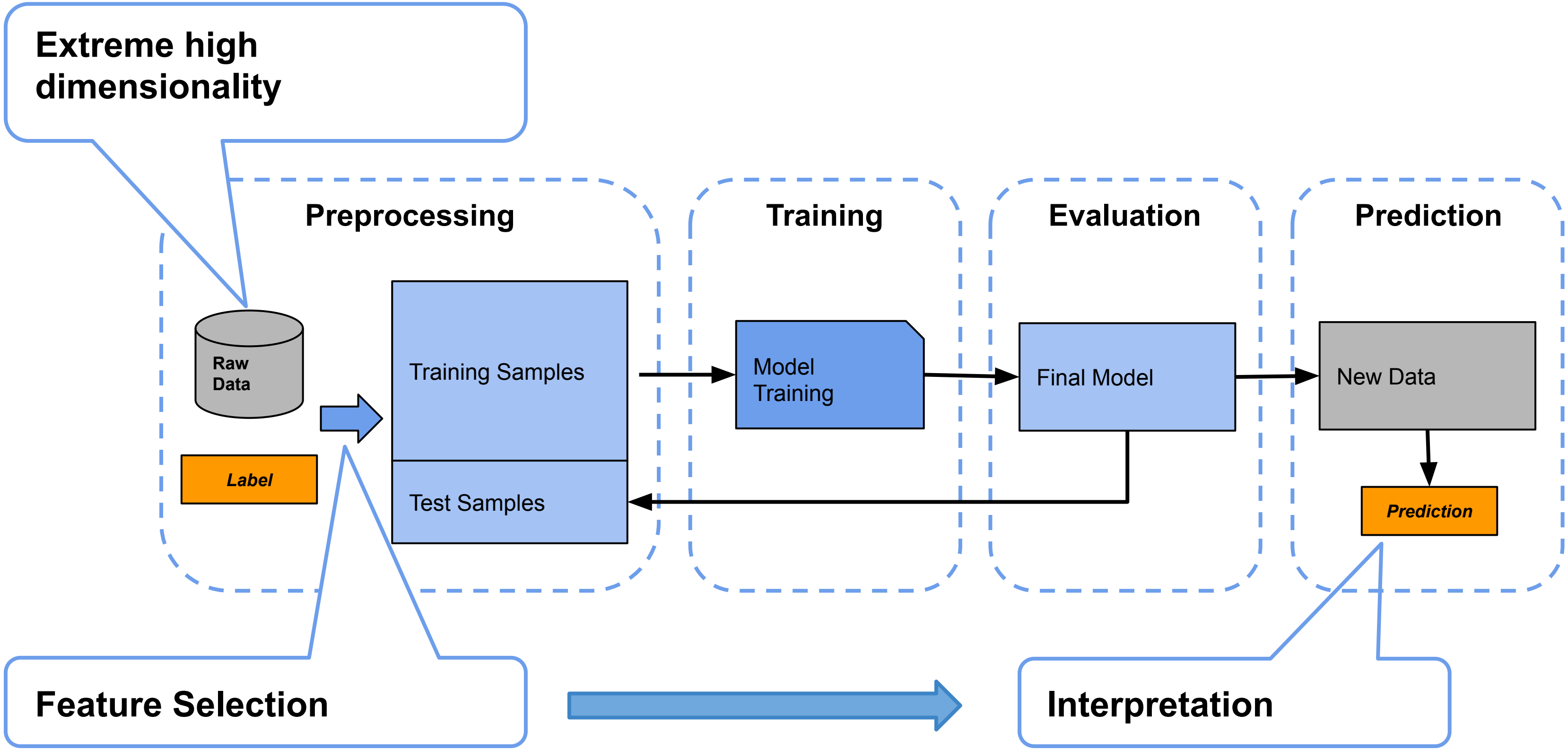
Agenda

- ❖ Characteristics of Climate Data
- ❖ Research Goal
- ❖ Minimal Target Explanation
- ❖ Galaxy Space
- ❖ Experiments

Roadmap for building machine learning systems on climate data



Roadmap for building machine learning systems on climate data



Extreme high dimensionality

Feature Selection

Interpretation

Characteristics for Climate Data

- **Extreme high dimensionality:** Climate science is one of the largest sources of data for data-driven research. Research on the climate phenomenon requires analyzing data under a large space-time window, which involves an enormous amount of features. And this will lead to “the curse of dimensionality”.

- **Scale amplification:** Weather systems are very sensitive to changes in initial conditions. So many small perturbations in air motion could compound to result in large changes over longer time frames.

- **Error magnification:** Because the system is so sensitive, measurement error in monitoring devices can lead to errors in analysis.

- **Local interpretability :** climate phenomena are the result of the interactions and operations of atmospheric physical effects on multiple spatial-temporal scales. This means the climate events occurred in different regions or different times may have different explanatory patterns.

Research Goal

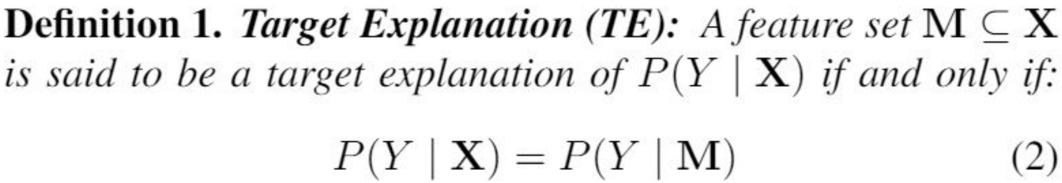
- **Feature Selection**:

  Selecting the relevant subset features from highly dimensional data and thus reducing learning complexity.

- **Global Interpretability**:

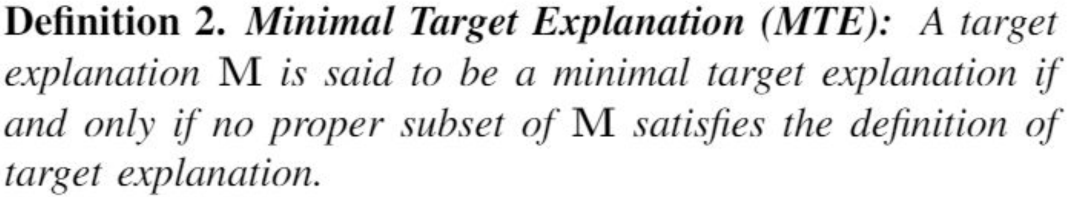
  Selected features should be explainable for samples.

Minimal Target Explanation



Definition 1. ***Target Explanation (TE):*** *A feature set* $\mathbf{M} \subseteq \mathbf{X}$ *is said to be a target explanation of* $P(Y \mid \mathbf{X})$ *if and only if:*

$$P(Y \mid \mathbf{X}) = P(Y \mid \mathbf{M}) \tag{2}$$

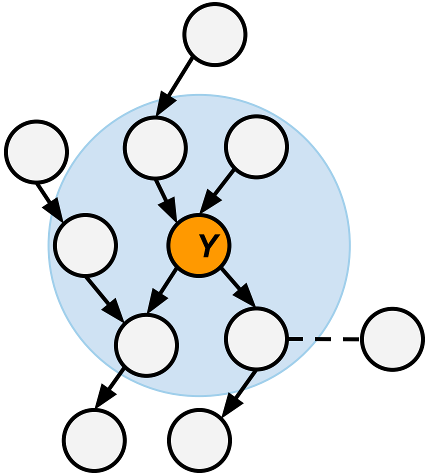


Definition 2. ***Minimal Target Explanation (MTE):*** *A target explanation* $\mathbf{M}$ *is said to be a minimal target explanation if and only if no proper subset of* $\mathbf{M}$ *satisfies the definition of target explanation.*

Minimal Target Explanation in Causal Analysis Theory

In faithful distributions, the Markov boundary of a node contains all the variables that shield the node from the rest of the Bayesian network. This means that the Markov boundary of a node is the only knowledge needed to predict the behavior of that node.

--- Judea Pearl, 1988.



In faithful distributions, Markov boundary corresponds to a local causal neighborhood of the that variable and consists of all its direct causes, effects, and causes of the direct effects.

--- Neapolitan, 2004; Tsamardinos and Aliferis, 2003.

Non Faithfulness and Multiple Markov boundaries

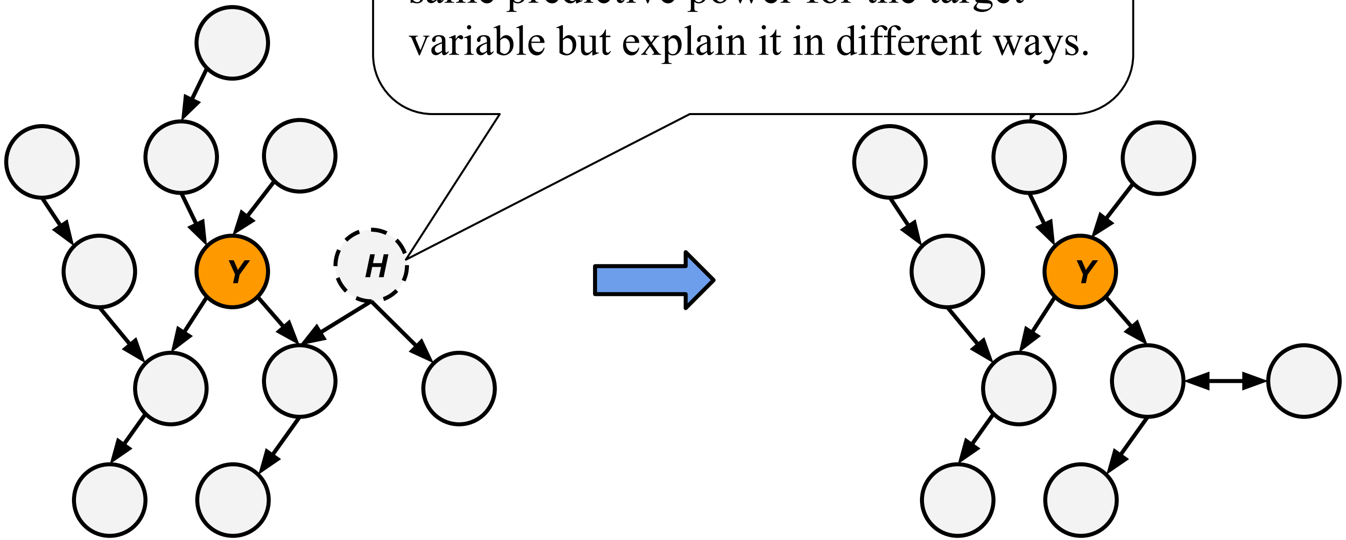
Markov boundary = all explanatory variables of the target variable?

If the joint probability distribution P is faithful to G (the graph of the Bayesian network), then there is a unique Markov boundary of the target variable (Judea Pearl, 1988). However, in real-world data, the faithfulness condition is always violated. This makes the Markov boundaries of the target variable are not unique in the common case (Alexander Statnikov, 2013).

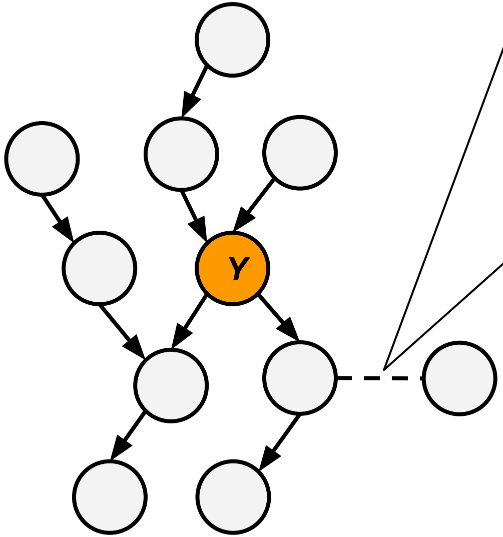
No

2 Reasons for Seeking Multiple Markov Boundaries

1. Hidden Common response variables. Two variables caused by the same hidden variable may have the same predictive power for the target variable but explain it in different ways.

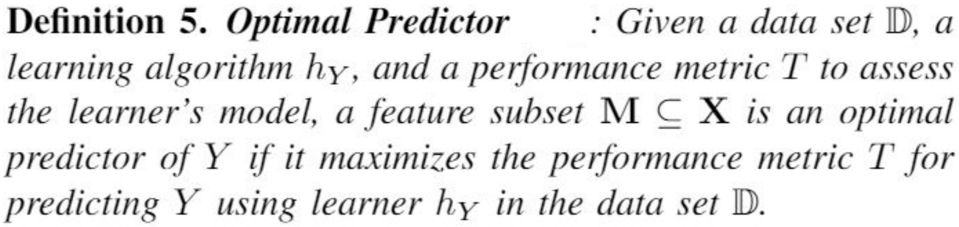


2 Reasons for Seeking Multiple Markov Boundaries

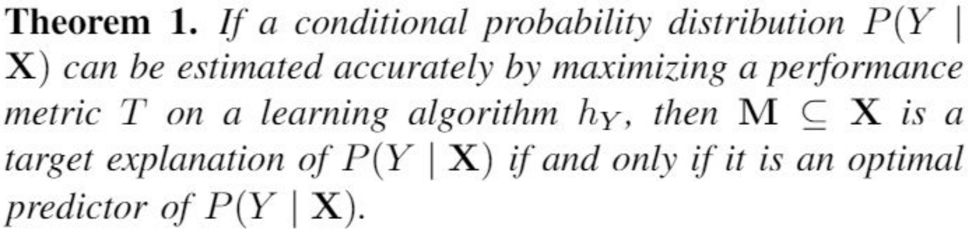


2. Hypothesis test error or Pure chance (Freedman's simulation experiment conducted in 1989) may lead the “Spurious Correlation”, Spurious Correlation make the misinterpretation of cause and effect.

Minimal Target Explanation (MTE) Detection

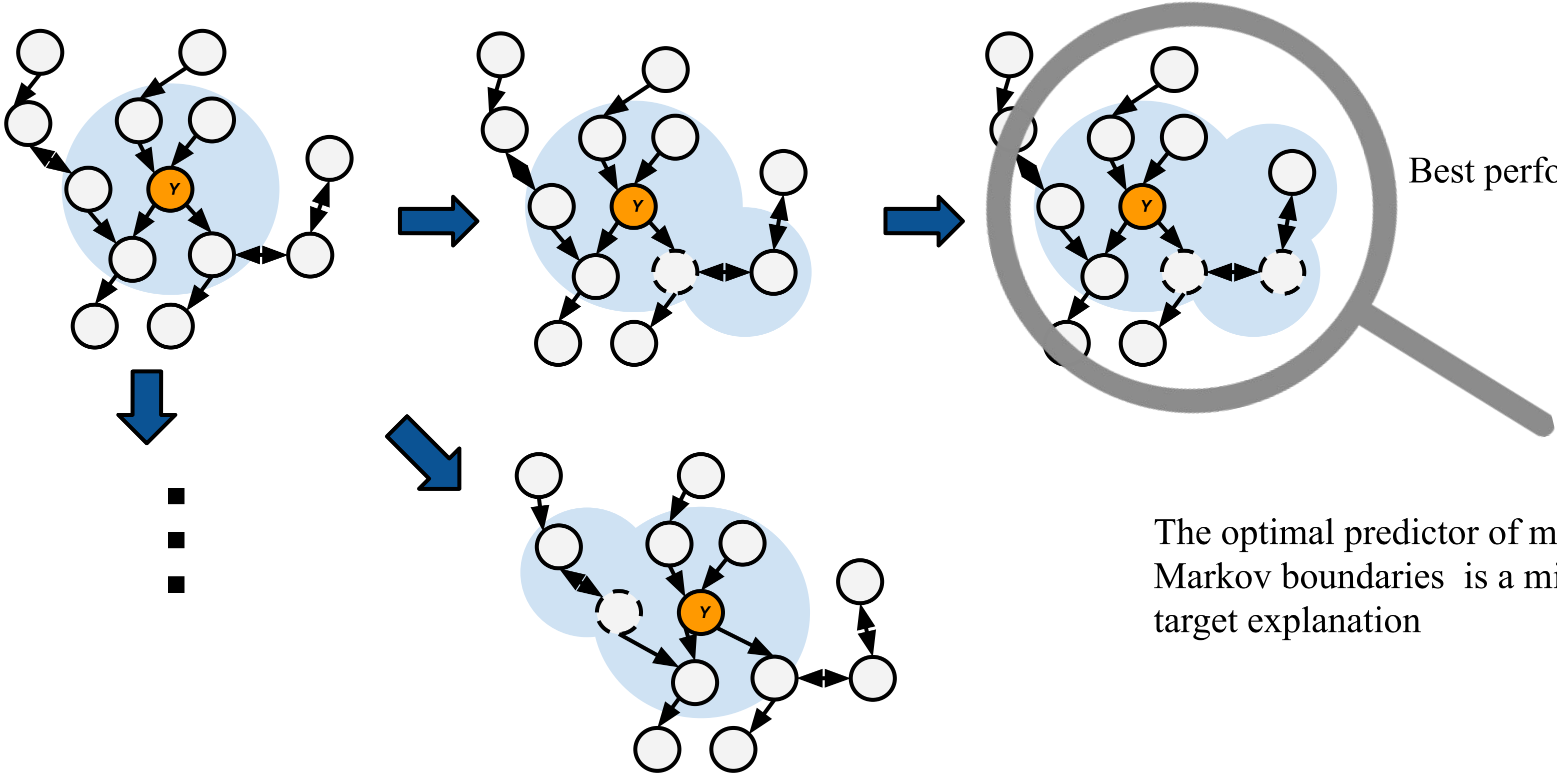


Definition 5. ***Optimal Predictor*** *: Given a data set $\mathbb{D}$, a learning algorithm h_Y, and a performance metric T to assess the learner's model, a feature subset $\mathbf{M} \subseteq \mathbf{X}$ is an optimal predictor of Y if it maximizes the performance metric T for predicting Y using learner h_Y in the data set $\mathbb{D}$.*



Theorem 1. *If a conditional probability distribution $P(Y \mid \mathbf{X})$ can be estimated accurately by maximizing a performance metric T on a learning algorithm h_Y, then $\mathbf{M} \subseteq \mathbf{X}$ is a target explanation of $P(Y \mid \mathbf{X})$ if and only if it is an optimal predictor of $P(Y \mid \mathbf{X})$.*

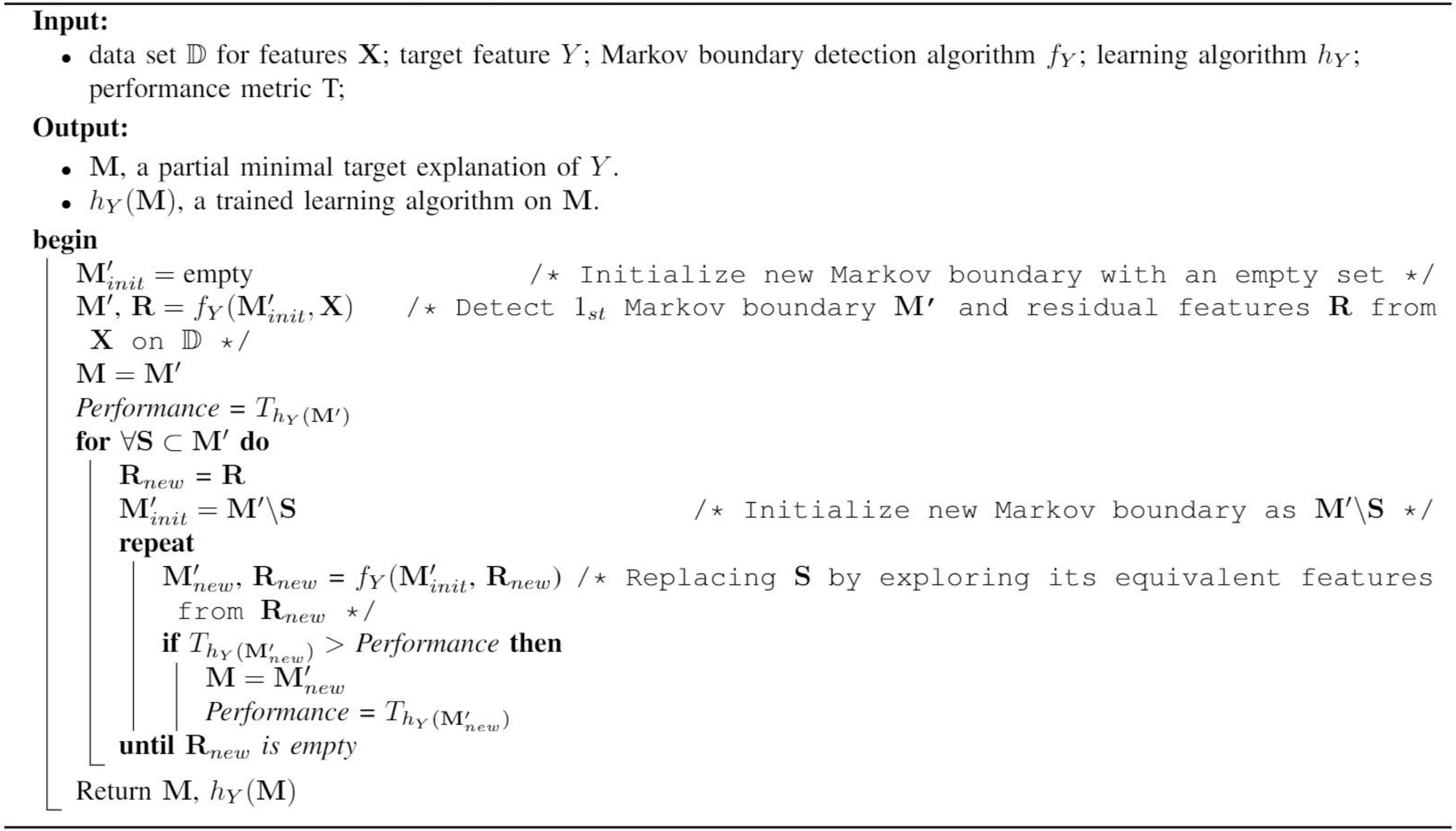
Recursive Search for Minimal Target Explanation (MTE)



Best performance

The optimal predictor of multiple Markov boundaries is a minimal target explanation

Minimal Target Explanation (MTE) Detection



Input:

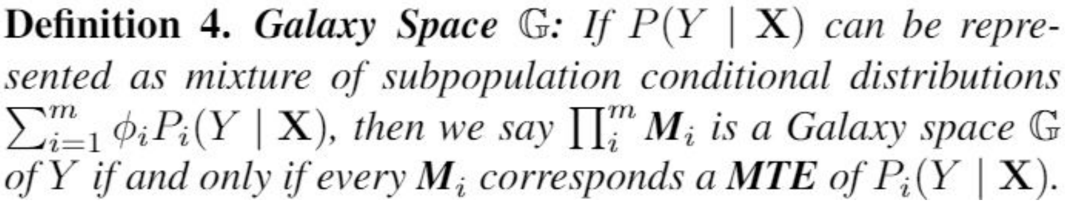
- data set $\mathbb{D}$ for features $\mathbf{X}$; target feature Y; Markov boundary detection algorithm f_Y; learning algorithm h_Y; performance metric T;

Output:

- $\mathbf{M}$, a partial minimal target explanation of Y.
- $h_Y(\mathbf{M})$, a trained learning algorithm on $\mathbf{M}$.

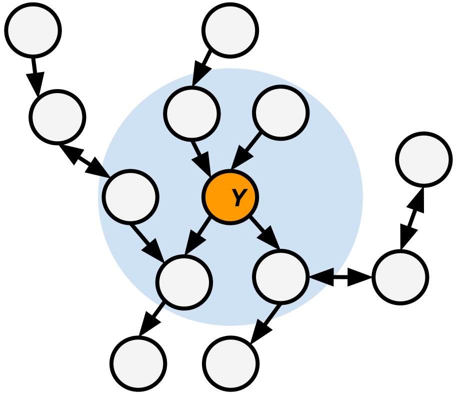
```
begin
    $\mathbf{M}'_{init}$ = empty    /* Initialize new Markov boundary with an empty set */
    $\mathbf{M}', \mathbf{R} = f_Y(\mathbf{M}'_{init}, \mathbf{X})$    /* Detect $1_{st}$ Markov boundary $\mathbf{M}'$ and residual features $\mathbf{R}$ from $\mathbf{X}$ on $\mathbb{D}$ */
    $\mathbf{M} = \mathbf{M}'$
    $Performance = T_{h_Y(\mathbf{M}')}$
    for $\forall \mathbf{S} \subset \mathbf{M}'$ do
        $\mathbf{R}_{new} = \mathbf{R}$
        $\mathbf{M}'_{init} = \mathbf{M}' \backslash \mathbf{S}$    /* Initialize new Markov boundary as $\mathbf{M}' \backslash \mathbf{S}$ */
        repeat
            $\mathbf{M}'_{new}, \mathbf{R}_{new} = f_Y(\mathbf{M}'_{init}, \mathbf{R}_{new})$ /* Replacing $\mathbf{S}$ by exploring its equivalent features from $\mathbf{R}_{new}$ */
            if $T_{h_Y(\mathbf{M}'_{new})} > Performance$ then
                $\mathbf{M} = \mathbf{M}'_{new}$
                $Performance = T_{h_Y(\mathbf{M}'_{new})}$
        until $\mathbf{R}_{new}$ is empty
    Return $\mathbf{M}$, $h_Y(\mathbf{M})$
```

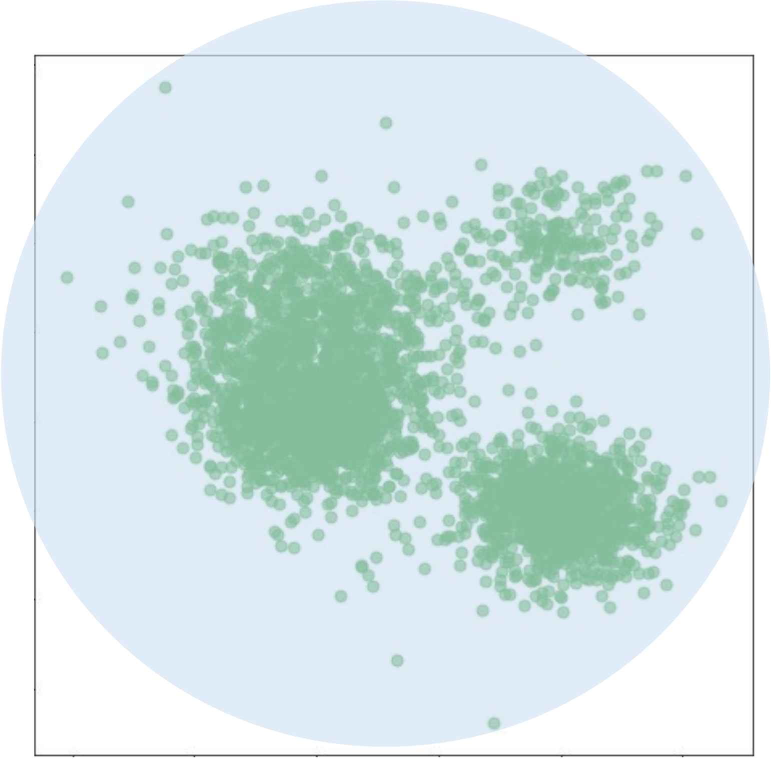
Galaxy Space



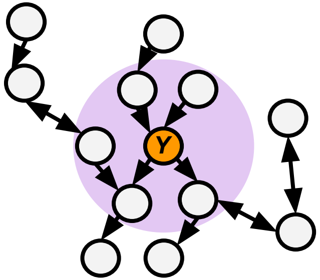
Definition 4. ***Galaxy Space*** $\mathbb{G}$***:*** *If* $P(Y \mid \mathbf{X})$ *can be represented as mixture of subpopulation conditional distributions* $\sum_{i=1}^{m} \phi_i P_i(Y \mid \mathbf{X})$*, then we say* $\prod_{i}^{m} \boldsymbol{M}_i$ *is a Galaxy space* $\mathbb{G}$ *of* Y *if and only if every* $\boldsymbol{M}_i$ *corresponds a* ***MTE*** *of* $P_i(Y \mid \mathbf{X})$.

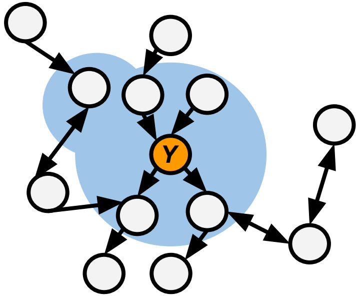
Minimal Target Explanation for the overall population of the samples

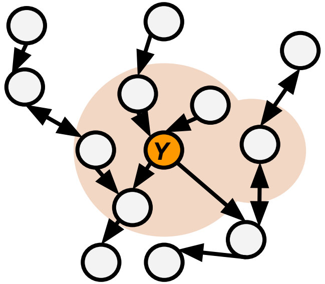


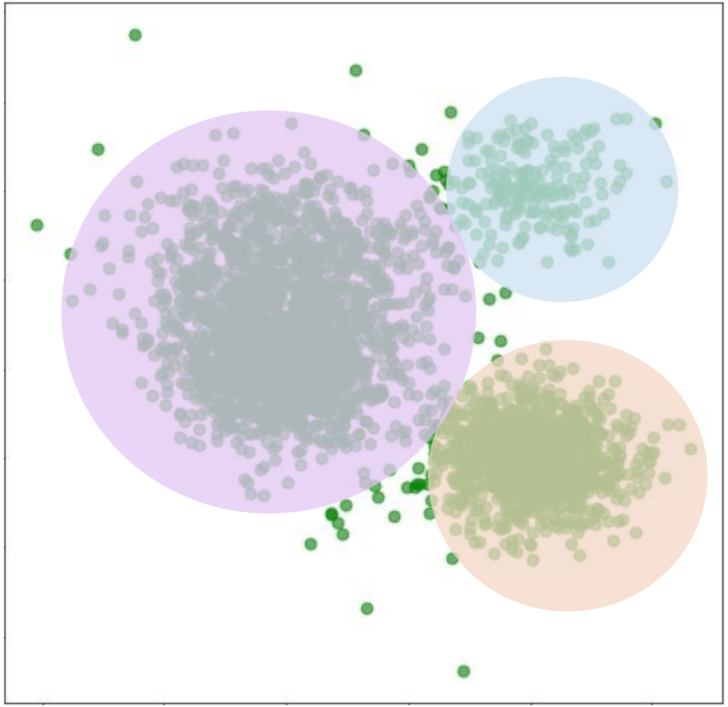


Galaxy space for all sub-populations of the samples

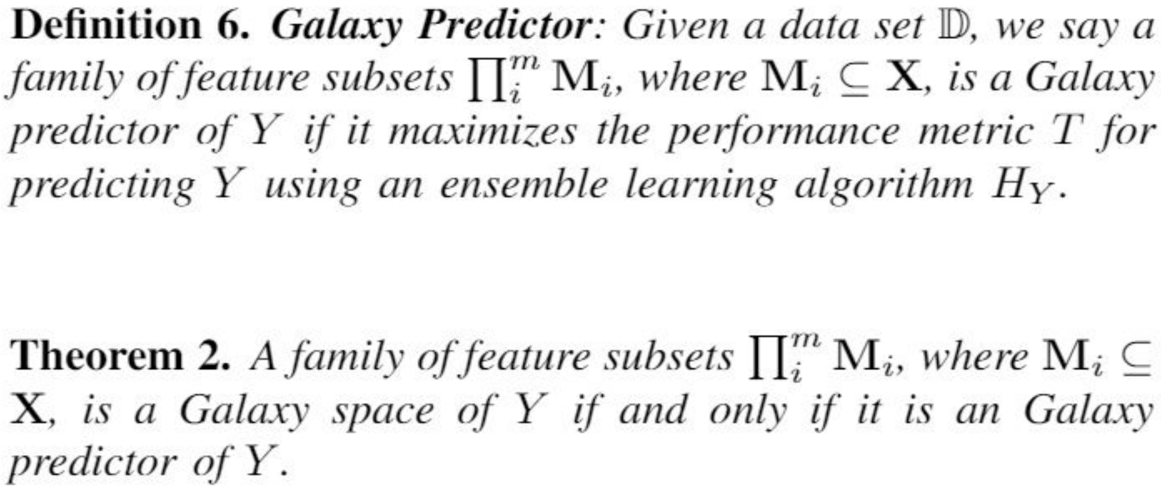








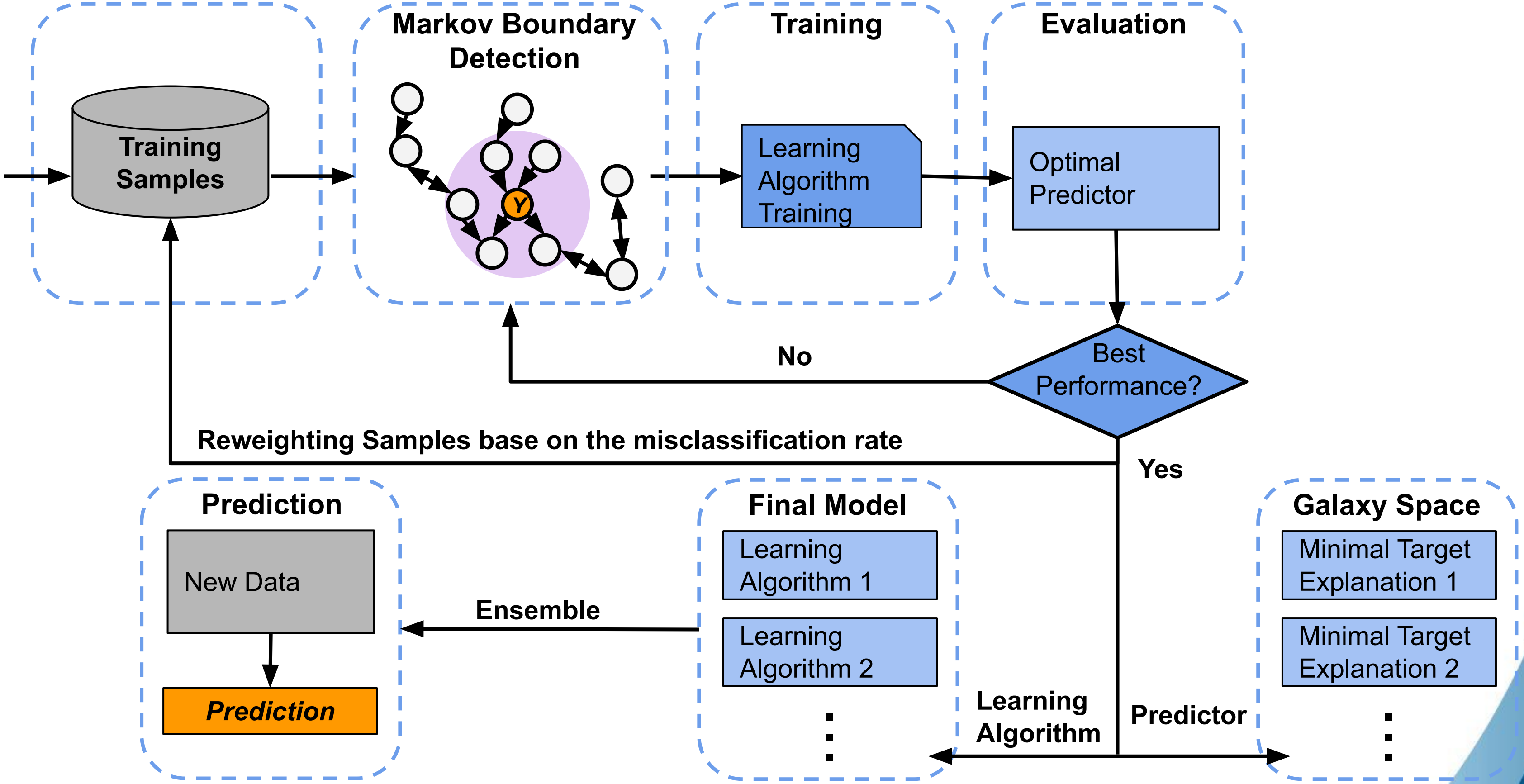
Galaxy Space Detection



Definition 6. ***Galaxy Predictor***: *Given a data set $\mathbb{D}$, we say a family of feature subsets $\prod_i^m \mathbf{M}_i$, where $\mathbf{M}_i \subseteq \mathbf{X}$, is a Galaxy predictor of Y if it maximizes the performance metric T for predicting Y using an ensemble learning algorithm H_Y.*

Theorem 2. *A family of feature subsets $\prod_i^m \mathbf{M}_i$, where $\mathbf{M}_i \subseteq \mathbf{X}$, is a Galaxy space of Y if and only if it is an Galaxy predictor of Y.*

Galaxy Space Detection



Training Samples

Markov Boundary Detection

Training

Learning Algorithm Training

Evaluation

Optimal Predictor

Best Performance?

No

Reweighting Samples base on the misclassification rate

Yes

Prediction

New Data

Prediction

Ensemble

Final Model

Learning Algorithm 1

Learning Algorithm 2

⋮

Learning Algorithm

Predictor

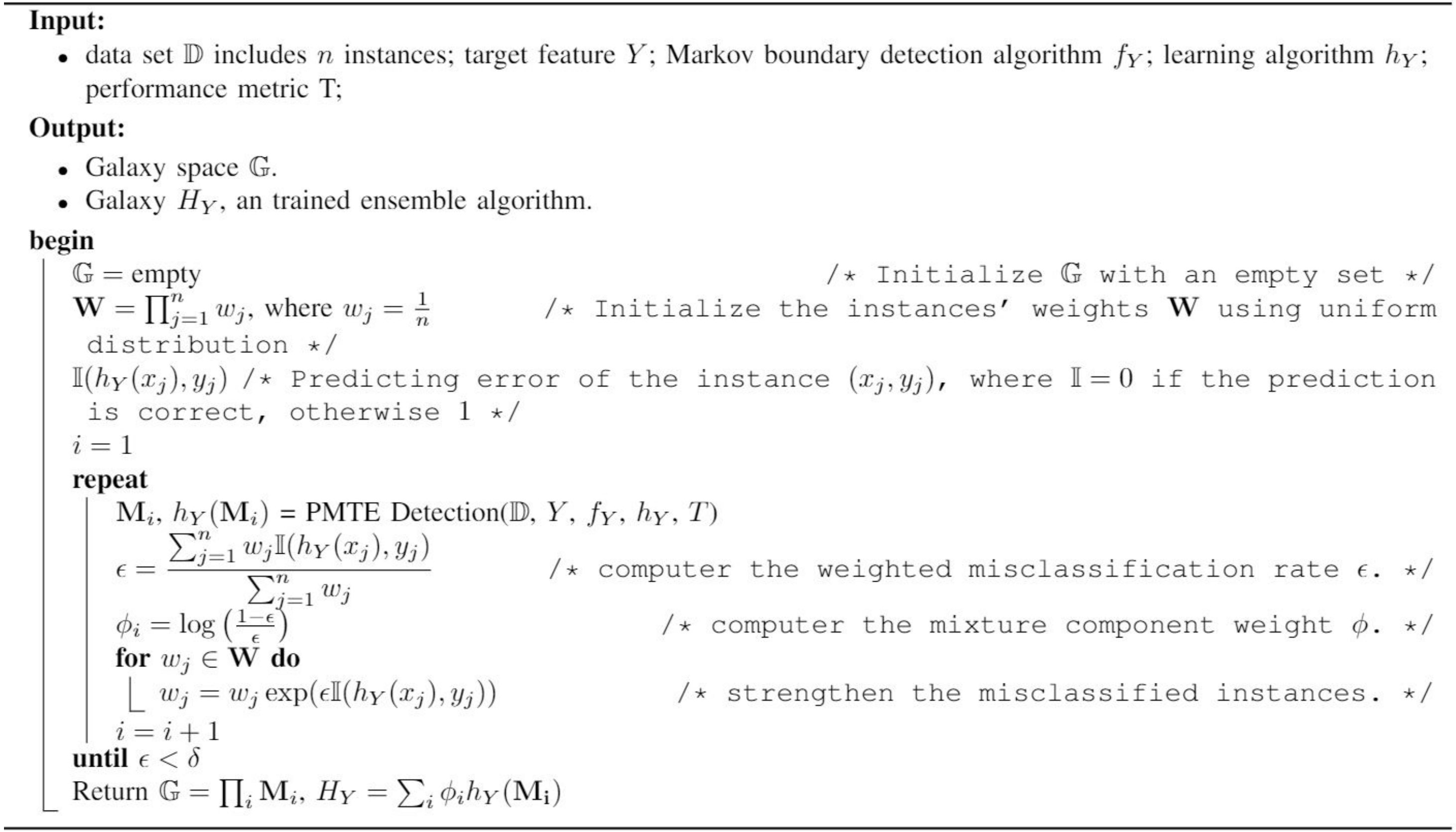
Galaxy Space

Minimal Target Explanation 1

Minimal Target Explanation 2

⋮

Galaxy Space Detection



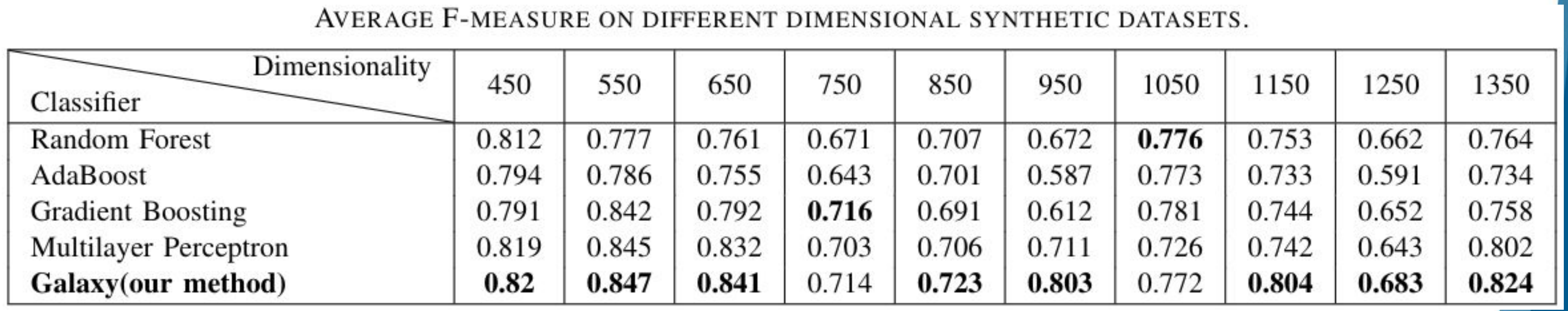
```
Input:
  • data set $\mathbb{D}$ includes $n$ instances; target feature $Y$; Markov boundary detection algorithm $f_Y$; learning algorithm $h_Y$;
    performance metric T;
Output:
  • Galaxy space $\mathbb{G}$.
  • Galaxy $H_Y$, an trained ensemble algorithm.
begin
    $\mathbb{G} = $ empty                                    /* Initialize $\mathbb{G}$ with an empty set */
    $\mathbf{W} = \prod_{j=1}^{n} w_j$, where $w_j = \frac{1}{n}$    /* Initialize the instances' weights $\mathbf{W}$ using uniform
     distribution */
    $\mathbb{I}(h_Y(x_j), y_j)$ /* Predicting error of the instance $(x_j, y_j)$, where $\mathbb{I} = 0$ if the prediction
     is correct, otherwise 1 */
    $i = 1$
    repeat
        $\mathbf{M}_i$, $h_Y(\mathbf{M}_i)$ = PMTE Detection($\mathbb{D}$, $Y$, $f_Y$, $h_Y$, $T$)
        $\epsilon = \frac{\sum_{j=1}^{n} w_j \mathbb{I}(h_Y(x_j), y_j)}{\sum_{j=1}^{n} w_j}$    /* computer the weighted misclassification rate $\epsilon$. */
        $\phi_i = \log\left(\frac{1-\epsilon}{\epsilon}\right)$         /* computer the mixture component weight $\phi$. */
        for $w_j \in \mathbf{W}$ do
            $w_j = w_j \exp(\epsilon \mathbb{I}(h_Y(x_j), y_j))$    /* strengthen the misclassified instances. */
        $i = i + 1$
    until $\epsilon < \delta$
    Return $\mathbb{G} = \prod_i \mathbf{M}_i$, $H_Y = \sum_i \phi_i h_Y(\mathbf{M_i})$
```

Experiments on Synthetic Data

- **Synthetic data generation:**
  In order to simulate highly correlated data that represent data collected in real-world climate applications, we generate a ***d***-dimensional synthetic data set using classification data generator in Python package scikit-learn with high redundancy and noise.

- **Experiment settings:**
  We demonstrate the effectiveness of Galaxy by comparing its F-measure against a bench of candidate methods: Random Forest, AdaBoost, Gradient Boosting, and multilayer perceptron. Simulated data were generated with feature counts ***d*** ranging from 450 to 1,350 in increments of 100 features. Each of the comparison methods was run against each of the subsets of the overall dataset after feature reduction. All computations were performed on the same hardware and datasets.

Results



AVERAGE F-MEASURE ON DIFFERENT DIMENSIONAL SYNTHETIC DATASETS.

Dimensionality Classifier	450	550	650	750	850	950	1050	1150	1250	1350
Random Forest	0.812	0.777	0.761	0.671	0.707	0.672	**0.776**	0.753	0.662	0.764
AdaBoost	0.794	0.786	0.755	0.643	0.701	0.587	0.773	0.733	0.591	0.734
Gradient Boosting	0.791	0.842	0.792	**0.716**	0.691	0.612	0.781	0.744	0.652	0.758
Multilayer Perceptron	0.819	0.845	0.832	0.703	0.706	0.711	0.726	0.742	0.643	0.802
Galaxy(our method)	**0.82**	**0.847**	**0.841**	0.714	**0.723**	**0.803**	0.772	**0.804**	**0.683**	**0.824**

Explanations for Precipitation Forecasting

The Precipitation Data Set:

The total number of variables in all levels is 18. All these meteorological variables are sampled at the spatial domain of 0°E to 375.5°E and 90°N to 20°S with a resolution of 2.5°×2.5° (totally 5,904 locations) and a daily temporal resolution. We pick the samples collected during the rainy season (March to November) during the years 1951-2017. The target feature is the historical spatial average precipitation (the mean of daily precipitation totals from 23 stations) of the Des Moines River basin in Iowa from the same time period. In the experiments, We set the lead time as 5 days, “look ahead” period as 10 days. Finally, each sample has 5,313,600 features (18 variables×5,904 locations×10 days).

Variable	Pressure level
Zonal Wind	850hPa, 500hPa, 200hPa
Meridional Wind	
Geopotential Height	
Atmospheric Temperature	
Specific humidity	850hPa
Relative humidity	700hPa, 925hPa
Vertical velocity	700hPa
The precipitable water	
Sea level pressure	

Explanations for Precipitation Forecasting

The Precipitation Data Set:

Lead time: 5 days.
Look ahead: 10 days.
Number of meteorology variable: 18
Locations:
0^{0}E to 375.5^{0}E and 90^{0}N to 20^{0}S with a resolution of 2.5^{0} $\times 2.5^{0}$ (5904 locations).
Total number of features:
5904 * 5 * 10 * 18 = 5,313,600
Target: The historical spatial average precipitation data of the Des Moines River basin.

Variable	Pressure level
Zonal Wind	850hPa, 500hPa, 200hPa
Meridional Wind	
Geopotential Height	
Atmospheric Temperature	
Specific humidity	850hPa
Relative humidity	700hPa，925hPa
Vertical velocity	700hPa
The precipitable water	
Sea level pressure	

Galaxy Space

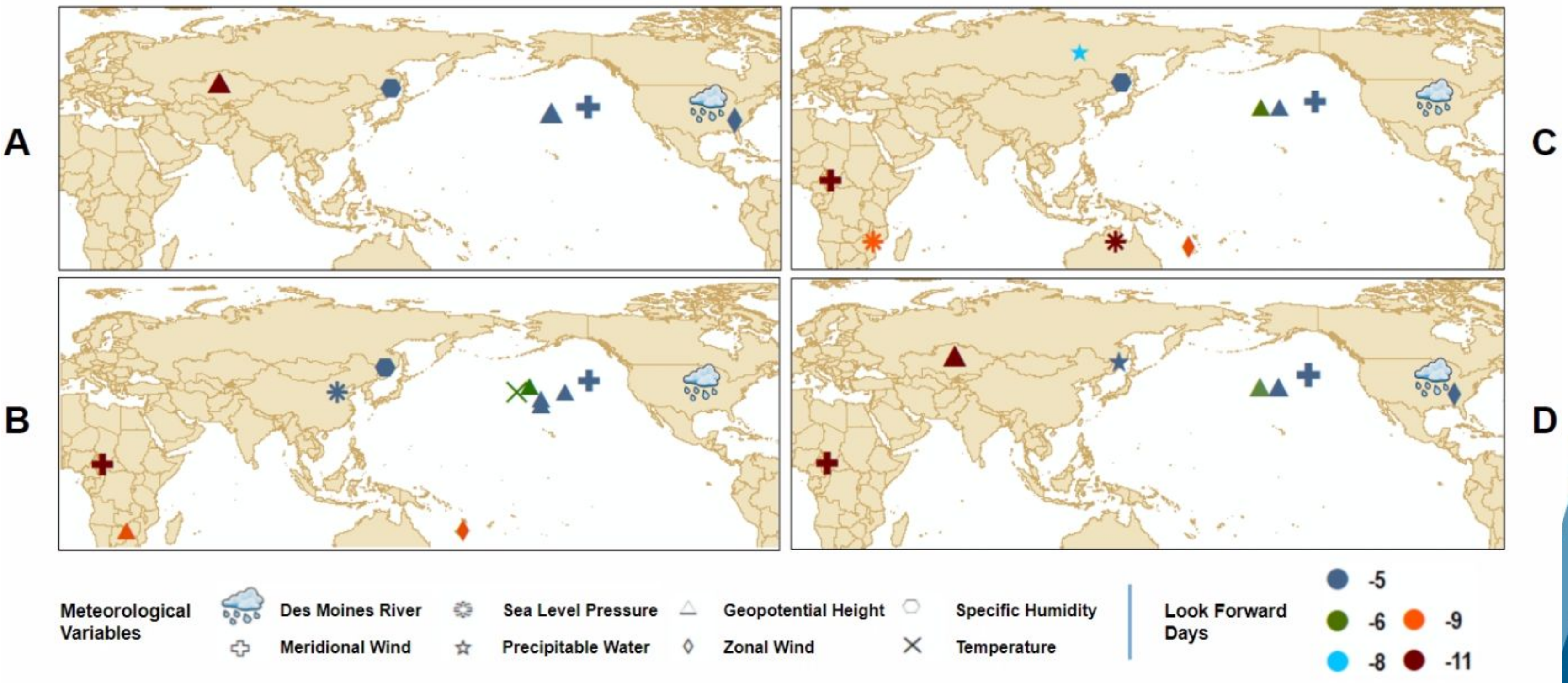


Fig. 3. Four MTEs of the precipitation at Des Moines river basin detected by Galaxy.

Thank You!

UMASS
BOSTON