Random Forest

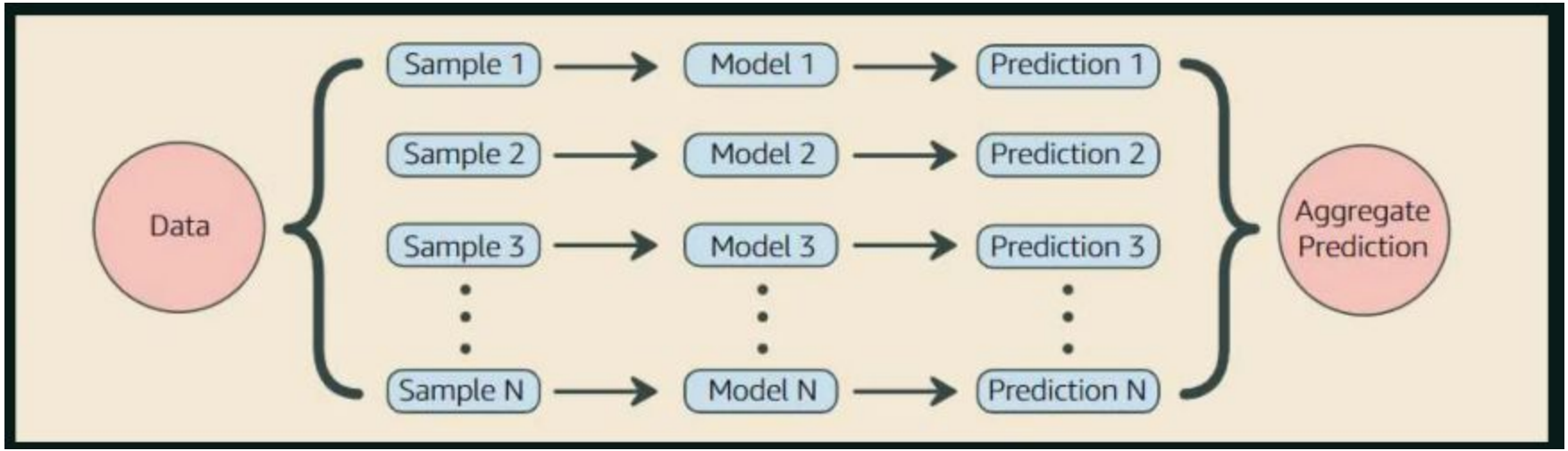
Agenda:

- What is random forest?
- How does random forest work?
- Advantages and disadvantages of random forest
- In our case…

What is a random forest?

A **Random Forest** is a model that builds **multiples decision trees,** each trained on slightly different data and features.-

- Each tree makes a prediction.
- Final prediction = combines all tree predictions (majority vote for classification and average for regression) .

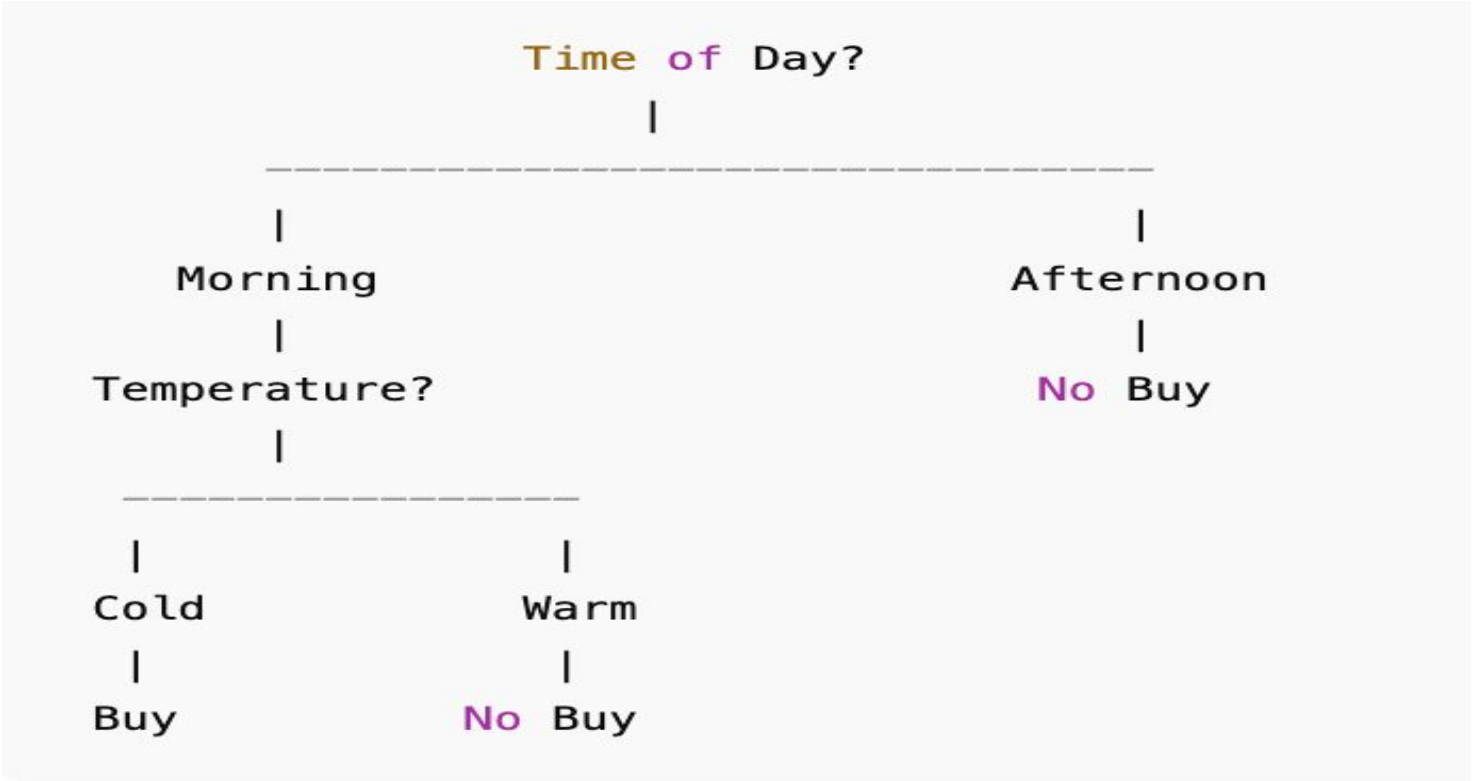


Example:

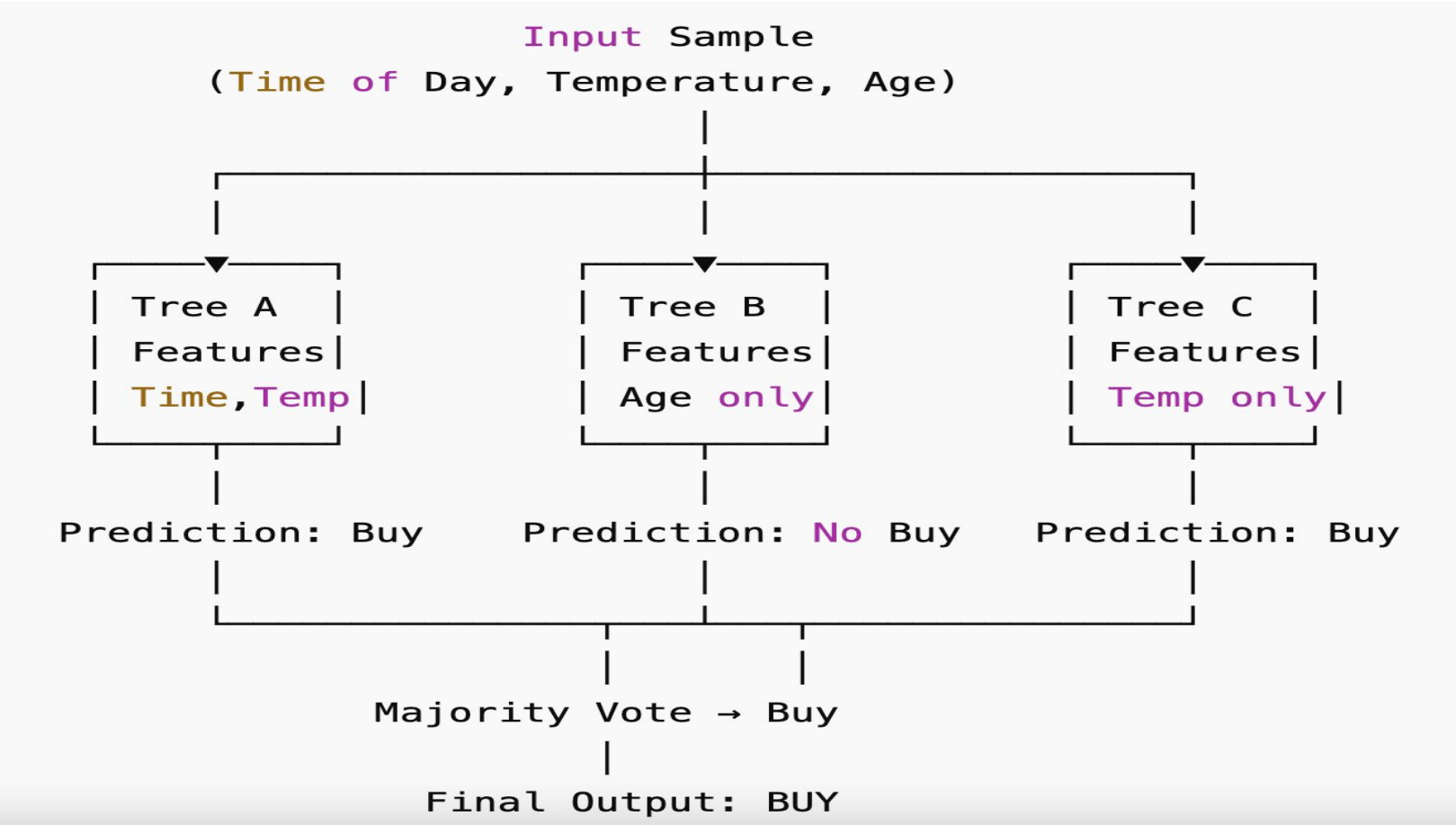
Predicting whether a customer will buy a product based on Time of Day, Temperature, and Age.

Time	Temp	Age	Buy?
Morning	Cold	22	Yes
Morning	Warm	45	No
Afternoon	Cold	19	Yes
Afternoon	Warm	35	No
Afternoon	Cold	28	Yes

Example: A single Decision Tree



Example: a Random Forest uses multiple Trees



Random Forest for Regression

- Random Forest can also predict **numerical values** instead of categories
- Each tree predicts a number, and the forest **averages** all predictions
- Example: Predicting house prices based on square footage, bedrooms, etc.
- **Process is the same as classification**, only the aggregation step changes (average instead of vote)

Advantages of Random Forest

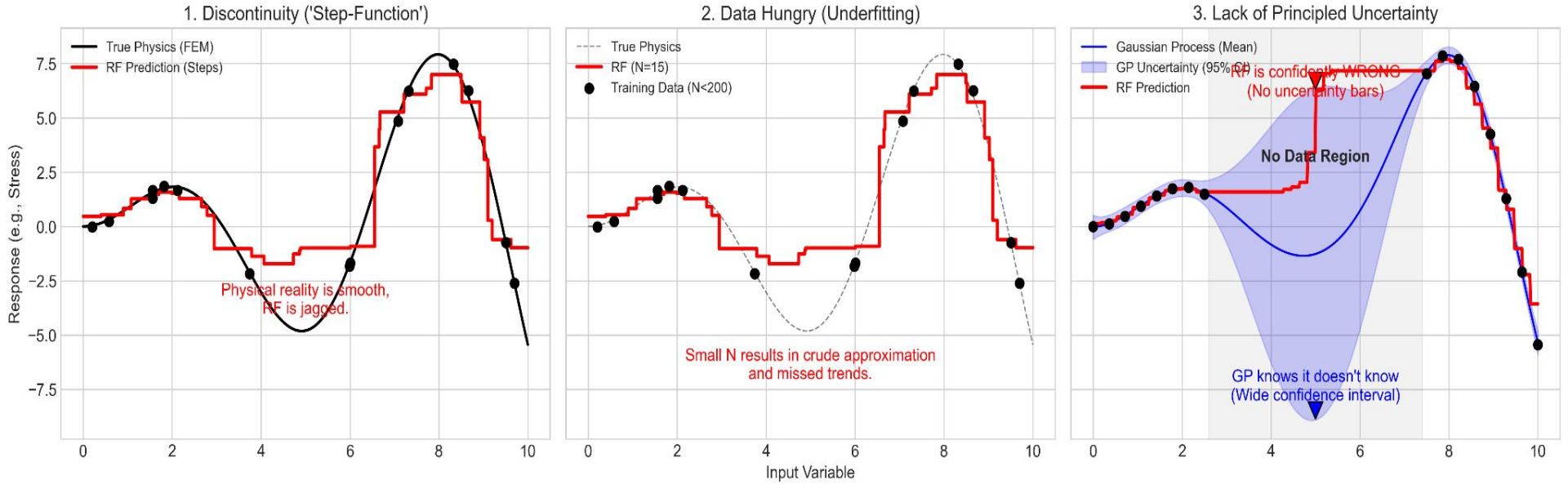
- **Robustness**: Random forests are resistant to overfitting, even with noisy datasets, because they aggregate many decision trees.
- **Versatility**: They can be used for both classification and regression problems.
- **Handles Missing Data**: Random forests can maintain good performance even when some data is missing.
- **Feature Importance**: Random forests can estimate the importance of each feature, making them valuable for feature selection.
- **Scalability**: Random forests are easily parallelizable, allowing them to scale to large datasets

Disadvantages of Random Forest

- **Interpretability**: While decision trees are easy to interpret, random forests are harder to interpret since they involve multiple trees. It can be difficult to understand the model as a whole.
- **Computationally Intensive**: Building a large number of trees can be computationally expensive, especially for large datasets.
- **Memory Usage**: Random forests require significant memory, as each tree needs to be stored.
- **Slow Predictions**: Making predictions with random forests can be slower than with a single decision tree, as it requires querying each tree in the forest

In Our Case…

- Random Forest can be used in this experiment for tasks such as **failure classification** and **variable importance analysis**. However, it is not suitable as the primary surrogate model for replacing the FEM simulation.
    - underfit in small-sample settings (n<200)
    - Its predictions are piecewise constant and not smooth
    - It lacks principled uncertainty quantification



Piecewise Constant
Physically unrealistic for smooth FEM fields.

High risk of underfitting when N < 200
Fails to capture complex trends poorly.

Cannot guide active learning effectively.