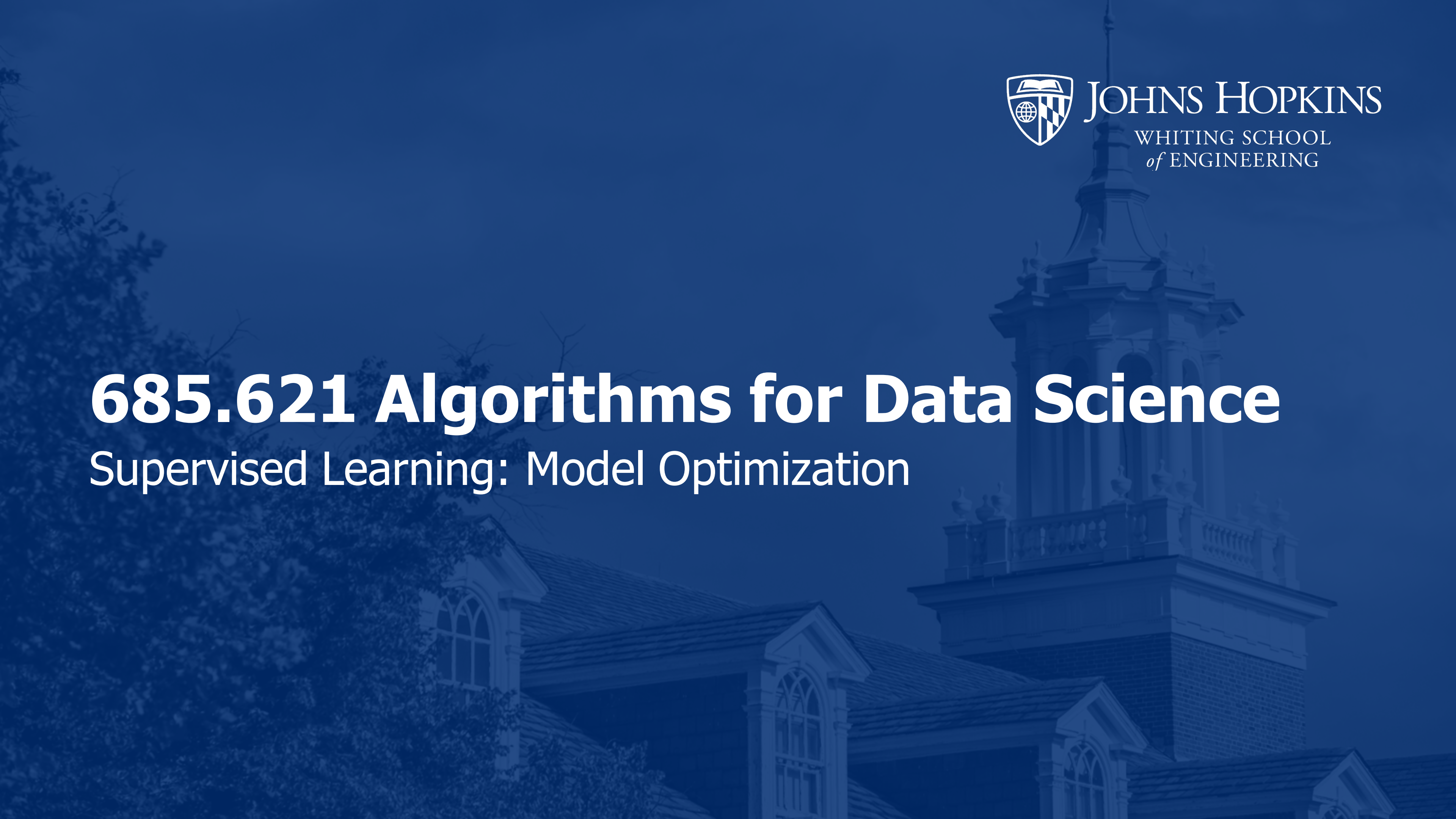
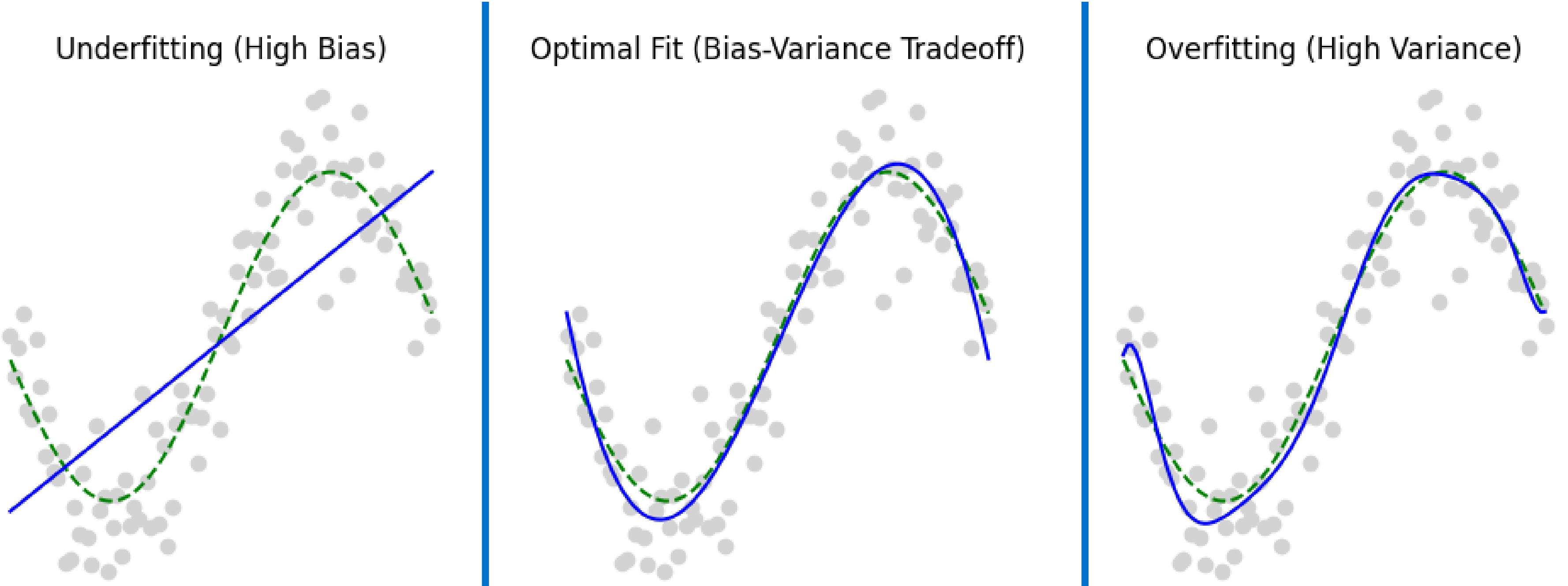


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

685.621 Algorithms for Data Science

Supervised Learning: Model Optimization

The Bias-Variance Tradeoff



Underfitting (High Bias)

Optimal Fit (Bias-Variance Tradeoff)

Overfitting (High Variance)

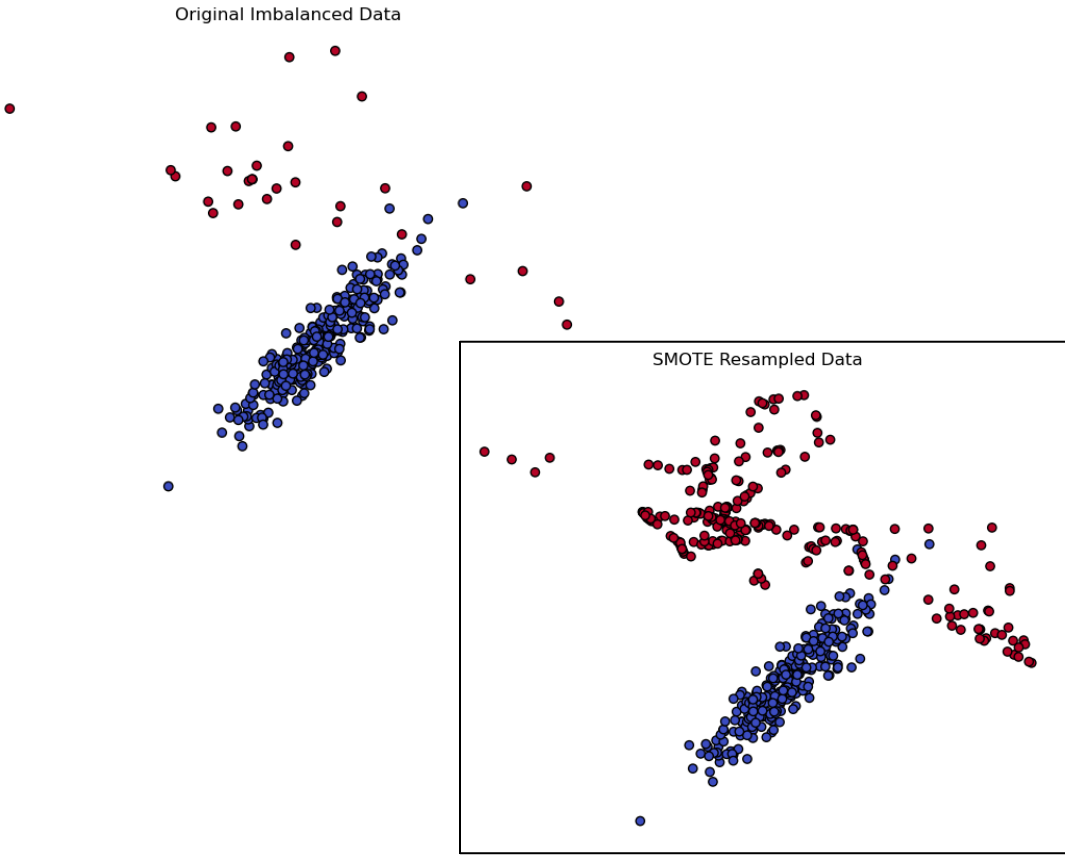
Feature Selection and Imbalanced Data

- **Why Feature Selection?**
  - Reduces overfitting by removing redundant/noisy features
  - Improves model interpretability
  - Speeds up training time
- **Methods**:
  - Fisher's Linear Discriminant Ratio
  - Decision Tree

- **Problem:** When one class dominates, models may favor the majority class.
- **Solutions:**
  - **Resampling**
    - Oversampling the minority class (SMOTE)
    - Under-sampling the majority class
  - **Class Weighting:** Assign higher penalty to minority misclassifications
  - **Synthetic Data Generation:** Generate new examples for the minority class

SMOTE: Synthetic Minority Over-sampling Technique

- **How SMOTE Works:**
  - Identify k-nearest neighbors of a minority class sample.
  - Generate synthetic points between the sample and its neighbors
  - Repeat until class distribution is balanced


Original Imbalanced Data

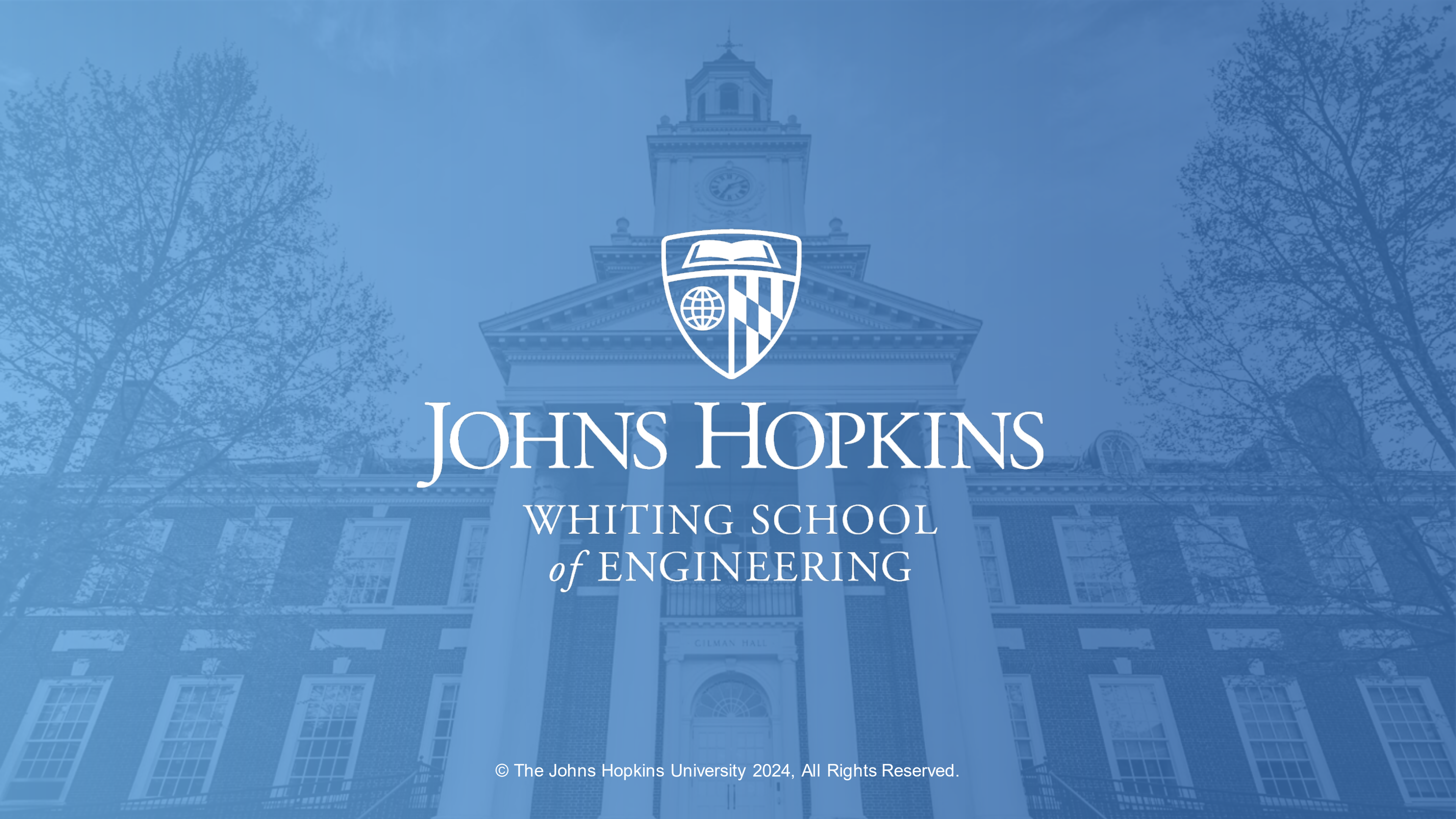
SMOTE Resampled Data

Why Data Leakage Can Ruin Your Model

- **What is data leakage?**
  - Occurs when a model has access to information that wouldn't be available during real-world predictions
- **Types of Data Leakage:**
  - **Target Leakage:** Features contain information about the label that wouldn't be available at prediction time.
  - **Cross-Validation Leakage:** Training data is inadvertently exposed to test data.
- **Example of Target Leakage:**
  - **Bad Feature**: A credit risk model includes "Number of late payments in the next 3 months"
  - **Why It's Bad**: The model would "cheat" by using future information

How to Avoid Data Leakage

- **Best Practices to Prevent Leakage:**
  - Remove future-dependent features
  - Perform preprocessing (e.g., scaling, encoding) only on training data.
  - Use proper cross-validation techniques
    - **Mistake** – splitting after preprocessing (introduces leakage)
    - **Mistake** – not stratifying data in imbalanced classification
    - **Fix** – Use Stratified K-Fold Cross-Validation for class balance
    - **Fix** – Apply transformations inside the cross-validation loop.



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING