Preventing Data Leakage

⚠ What is Data Leakage?

Information from test set influences training, causing overly optimistic performance estimates

✗ Wrong Workflow

1. Entire Dataset
2. Normalize all data
   ⚠ Uses test statistics!
3. Split into Train/Test
4. Train Model

✓ Correct Workflow

1. Entire Dataset
2. Split into Train/Test
   ✓ Split FIRST!
3. Normalize using train stats only
4. Train Model

🚨 Common Leakage Sources

Using test set statistics

Features with future info

Duplicate samples

Improper CV folds

⚡ Consequence: Overly optimistic performance → Model fails in production