PCA

Principal Component Analysis

Charles Parker

VP ML Algorithms, BigML

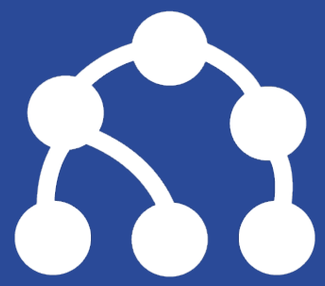
Issues with High Dimensionality

- Implicitly increases model complexity, prone to overfitting
- Requires more observations in order to generalize well
- Contains correlated or useless variables
- Data is difficult to visualize
- Takes a longer time to train models or make predictions

Principal Component Analysis addresses all of these issues

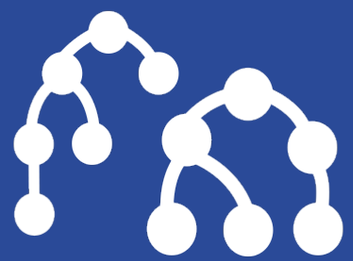
Other Approaches

MODEL



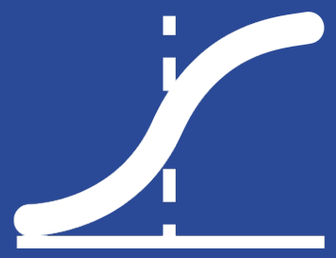
Pruning, Node threshold

ENSEMBLE



Bagging, Randomization

LOGISTIC REGRESSION



L1 and L2 penalties

DEEPNET

Dropout

Dimensionality Reduction

Manual Approach

Feature Selection

- Preserves the original variables and selects a subset
- Often uses recursive methods or statistical thresholds
- Examples: RFE, Chi-Squared Test, Boruta

Feature Extraction

- Transforms original variables into variables better suited for modeling
- Examples: word vectors, clustering
- **PCA falls into this category**

When to use PCA

1. You want to reduce the number of variables in your model, but it is not clear which should be eliminated
2. You want to generate variables that are not correlated
3. You are okay with sacrificing some amount of interpretability for potential downstream performance gains

How Does PCA Work?

Each PC is a ***linear combination*** of original variables

$$PC_1 = w_1F_1 + w_2F_2 + w_3F_3 + \ldots + w_NF_N$$

$$PC_2 = w_1F_1 + w_2F_2 + w_3F_3 + \ldots + w_NF_N$$

$$\vdots$$

$$PC_N = w_1F_1 + w_2F_2 + w_3F_3 + \ldots + w_NF_N$$

PCA Output

Original Data Matrix

	F_1	F_2	F_3	F_4	F_5	...	F_N
I_1							
I_2							
I_3							
I_4							
I_5							
...							
I_N							

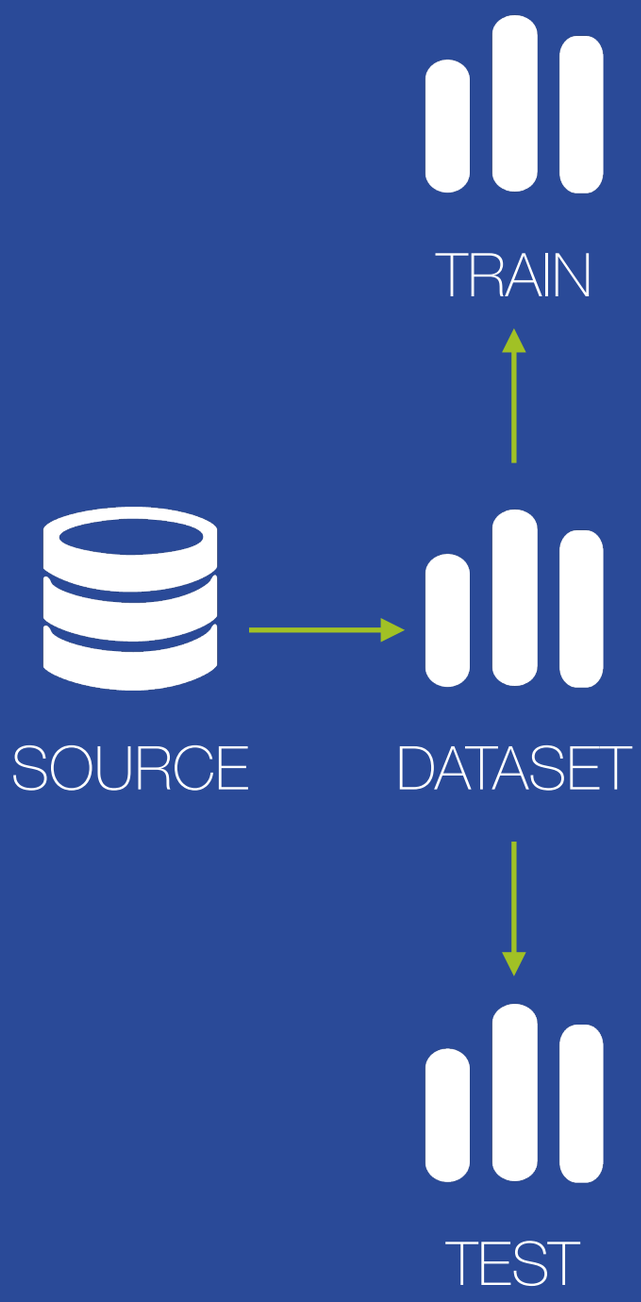
Transformed Data Matrix

	PC1	PC2	PC3	PC4	PC5	...	PC_N
I_1							
I_2							
I_3							
I_4							
I_5							
...							
I_N							

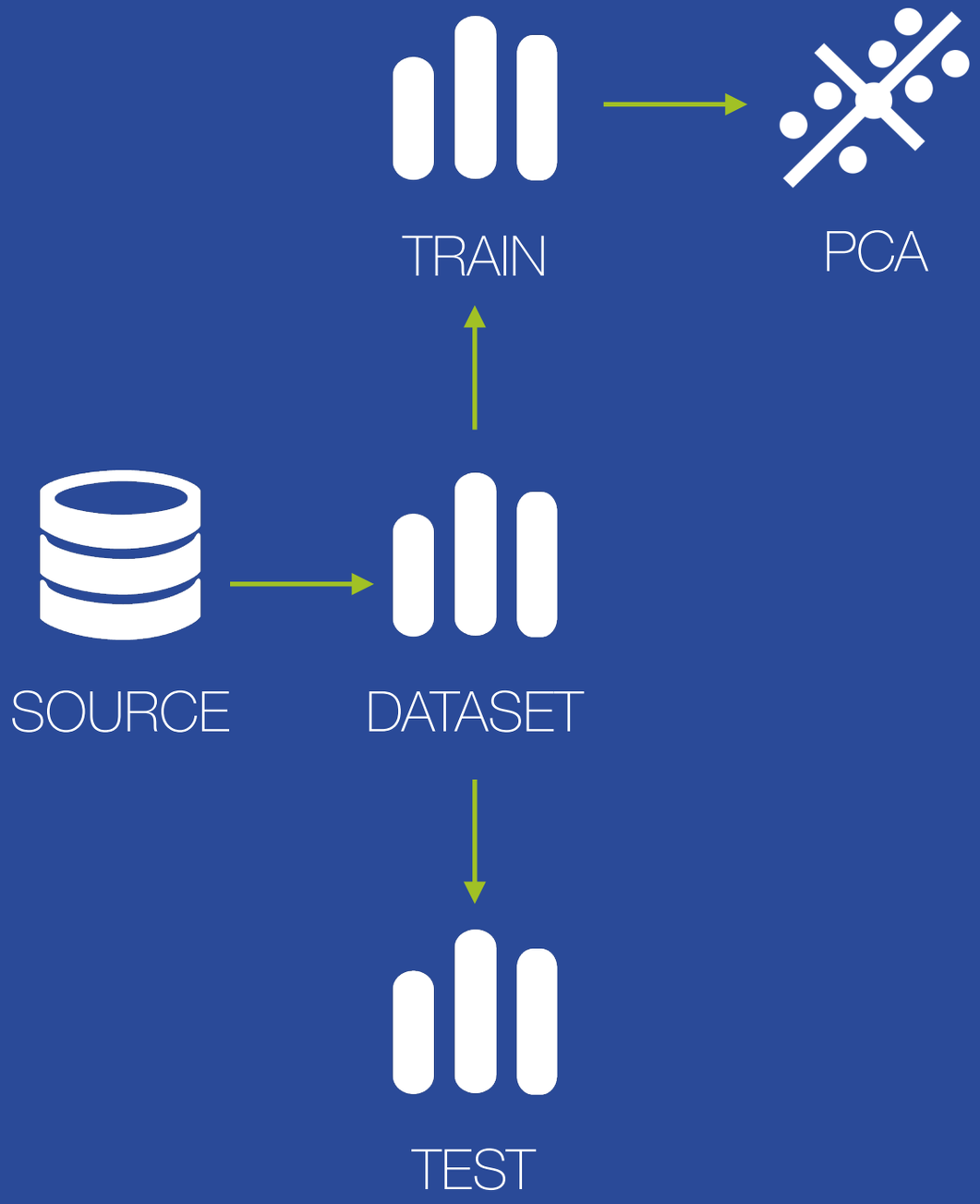
The new variables are the "principal components"

These principal components are not correlated

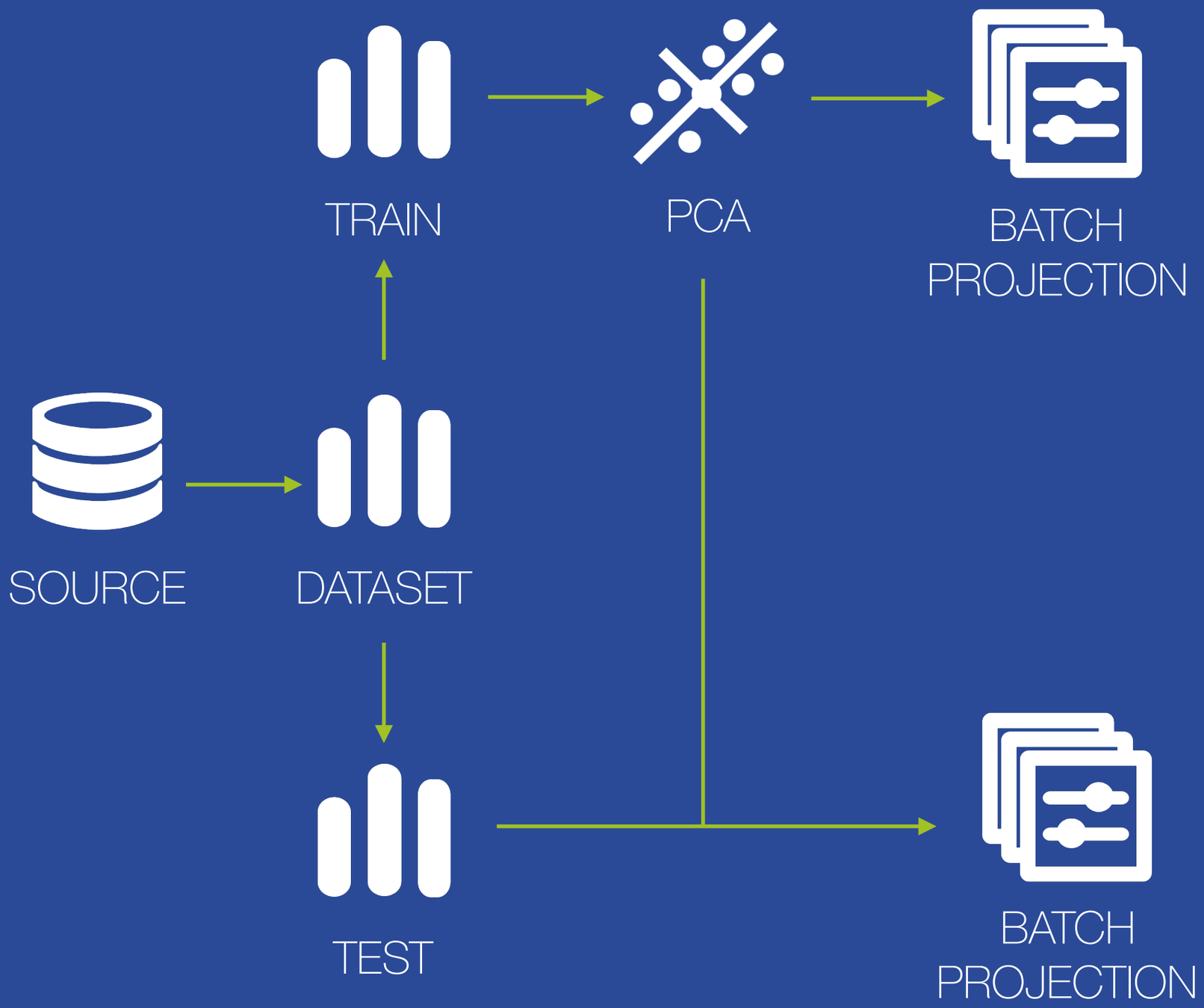
PCA Workflow



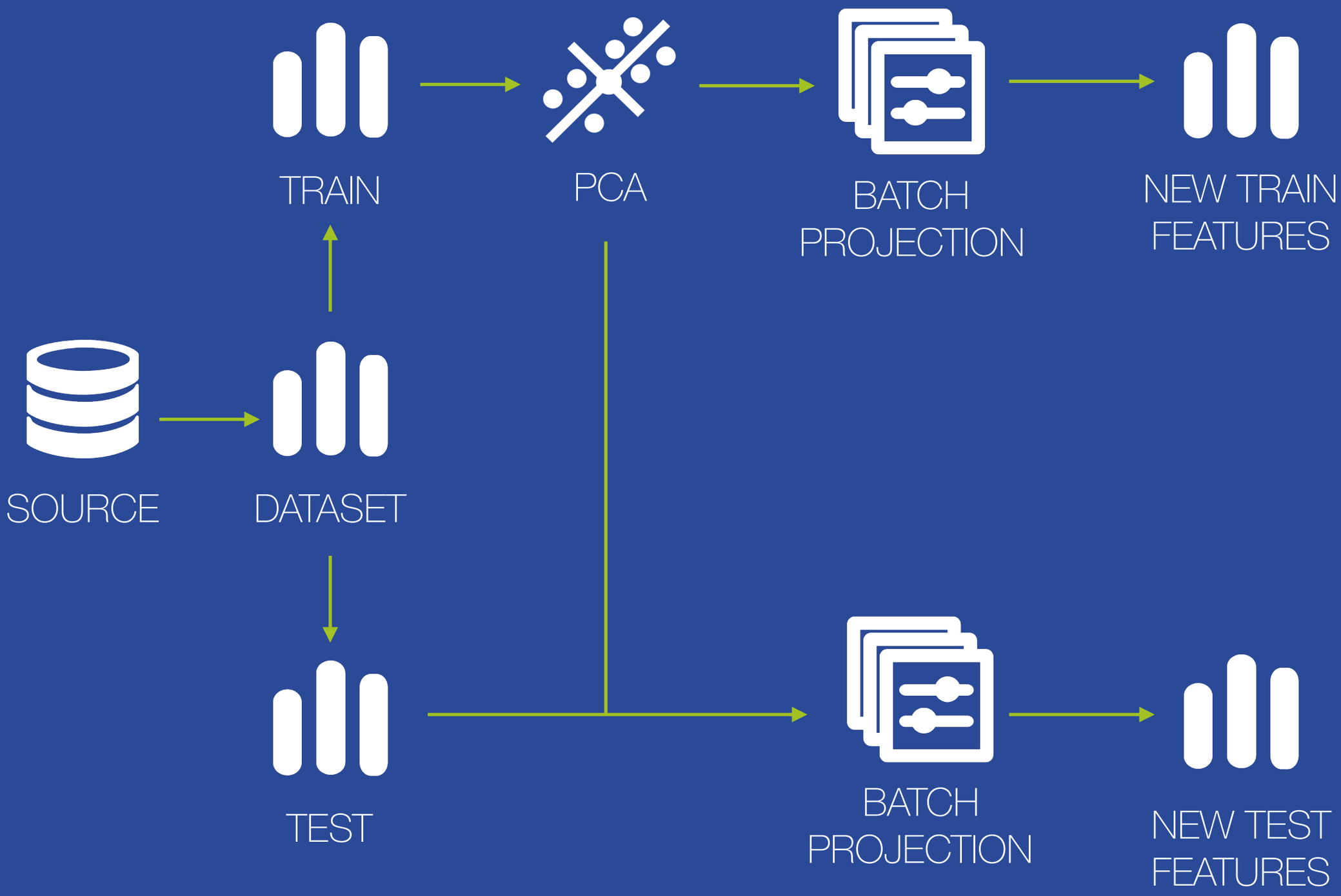
PCA Workflow



PCA Workflow



PCA Workflow



PCA Demo

Under the Hood: BigML PCA

- Standard PCA only applies to numerical data

- BigML uses three different data transformation methods in order to handle different data types
    - Numeric data: Principal Component Analysis (PCA)
    - Categorical data: Multiple Correspondence Analysis (MCA)
    - Mixed data: Factorial Analysis of Mixed Data (FAMD)

- BigML will automatically handle numeric, text, items, and categorical data without needing user input


bigml