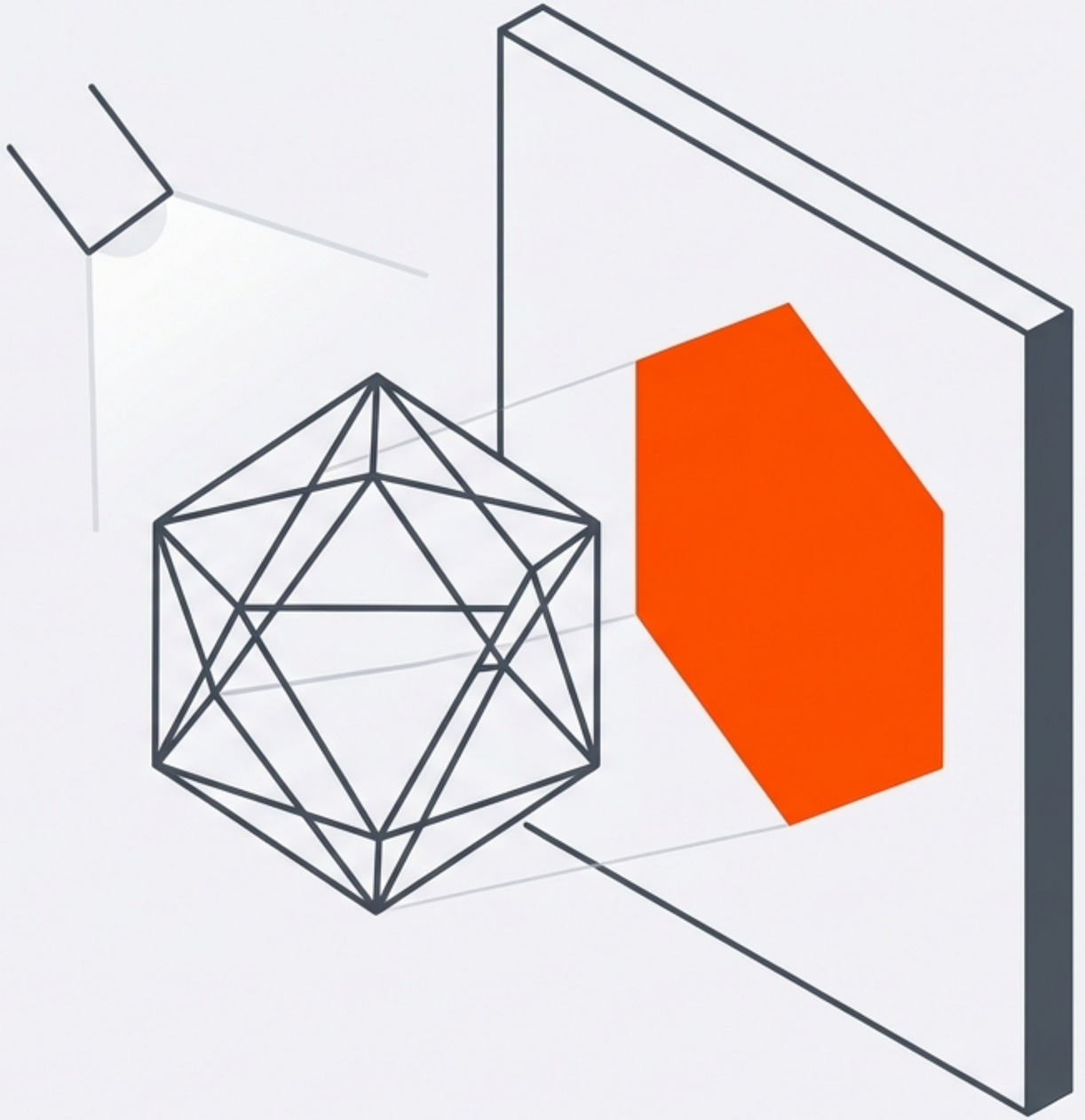
Principal Component Analysis (PCA)

The Art of Data Simplification



A Concept-to-Code Guide on Dimensionality Reduction

The Curse of Dimensionality

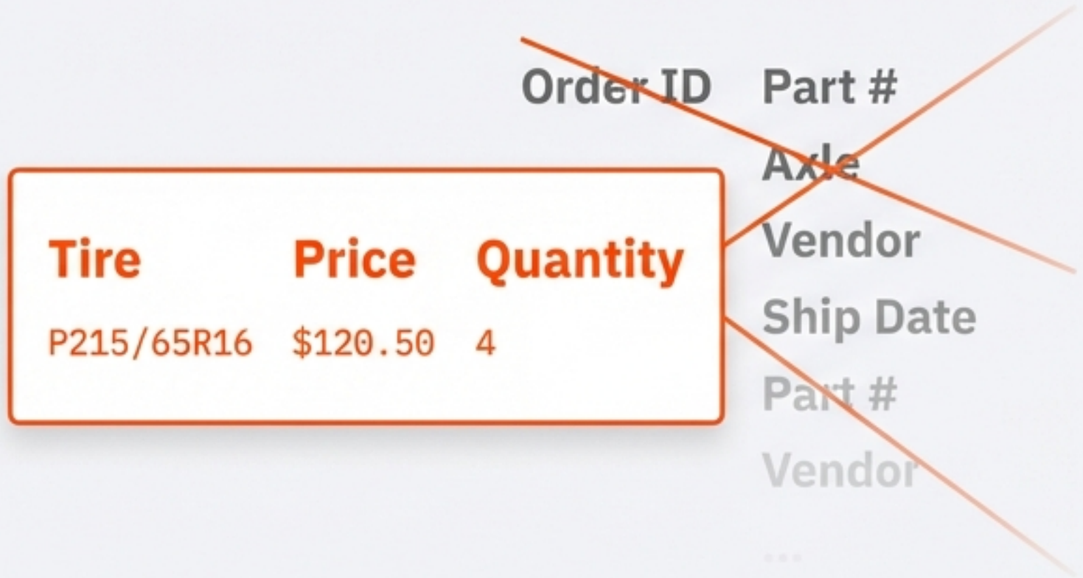
The Problem

Messy Spreadsheet

Tire	Axle	Order ID	Price	Quantity	Part #	Vendor	Ship Date
P215/65R16	Rear	1002345	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002346	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002347	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002348	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002345	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002346	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002347	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002348	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002349	$120.50	3	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002340	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002341	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002342	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002343	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002344	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002345	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002346	$120.50	4	A12-998	Vendor X	2023-10-25
P215/65R16	Rear	1002347	$99.00	3	A12-998	Vendor X	2023-10-25

Noise & Redundancy

The Insight



Correlation analysis reveals we **only need 3 attributes** to predict sales. We discard the noise to **find the signal**.

Why We Reduce Dimensions

Four critical advantages of pre-processing data.

Computational Efficiency

Drastically cuts training time. For large datasets (7GB+), this saves days of model iteration.

Redundancy Removal

Prevents models from over-weighting duplicated information.

Storage Optimization

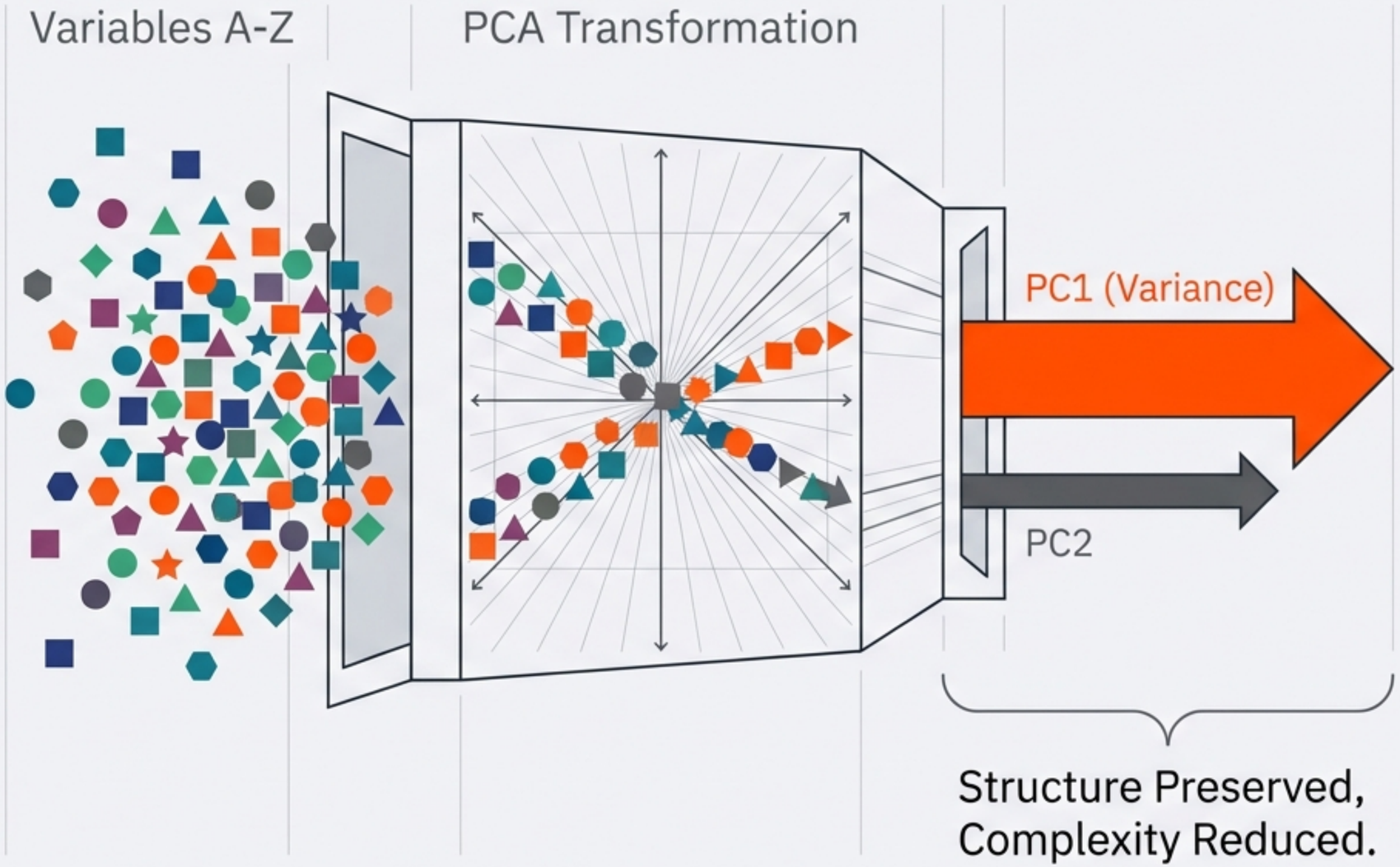
Eliminates the need to store irrelevant columns in big data pools.

Visualization

Enables the plotting of complex high-dimensional data in 2D or 3D for stakeholder presentation.

What is Principal Component Analysis?

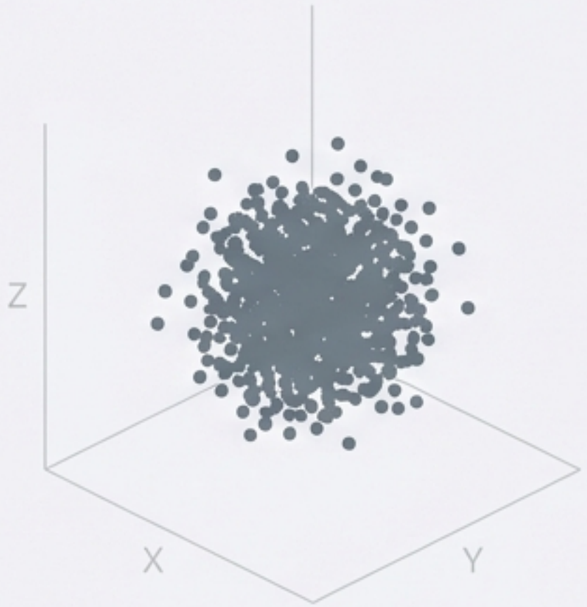
PCA is a technique for reducing the dimensionality of datasets while minimizing information loss. It constructs new variables—Principal Components—that are linear combinations of the initial variables.



The Intuition: The Photographer's Dilemma

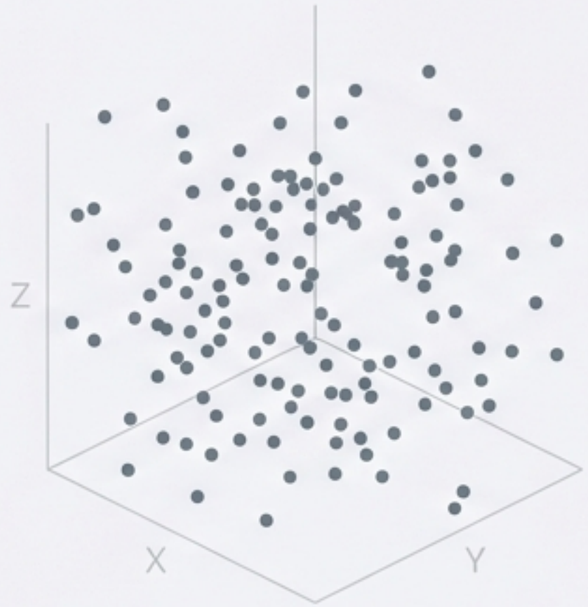
PCA is finding the camera angle that reveals the most information.

Angle A (Poor)



Low Variance / Overlap

Angle B (Poor)



High Noise / No Structure

Angle C (The Principal Component)

Maximum Variance / distinct Data

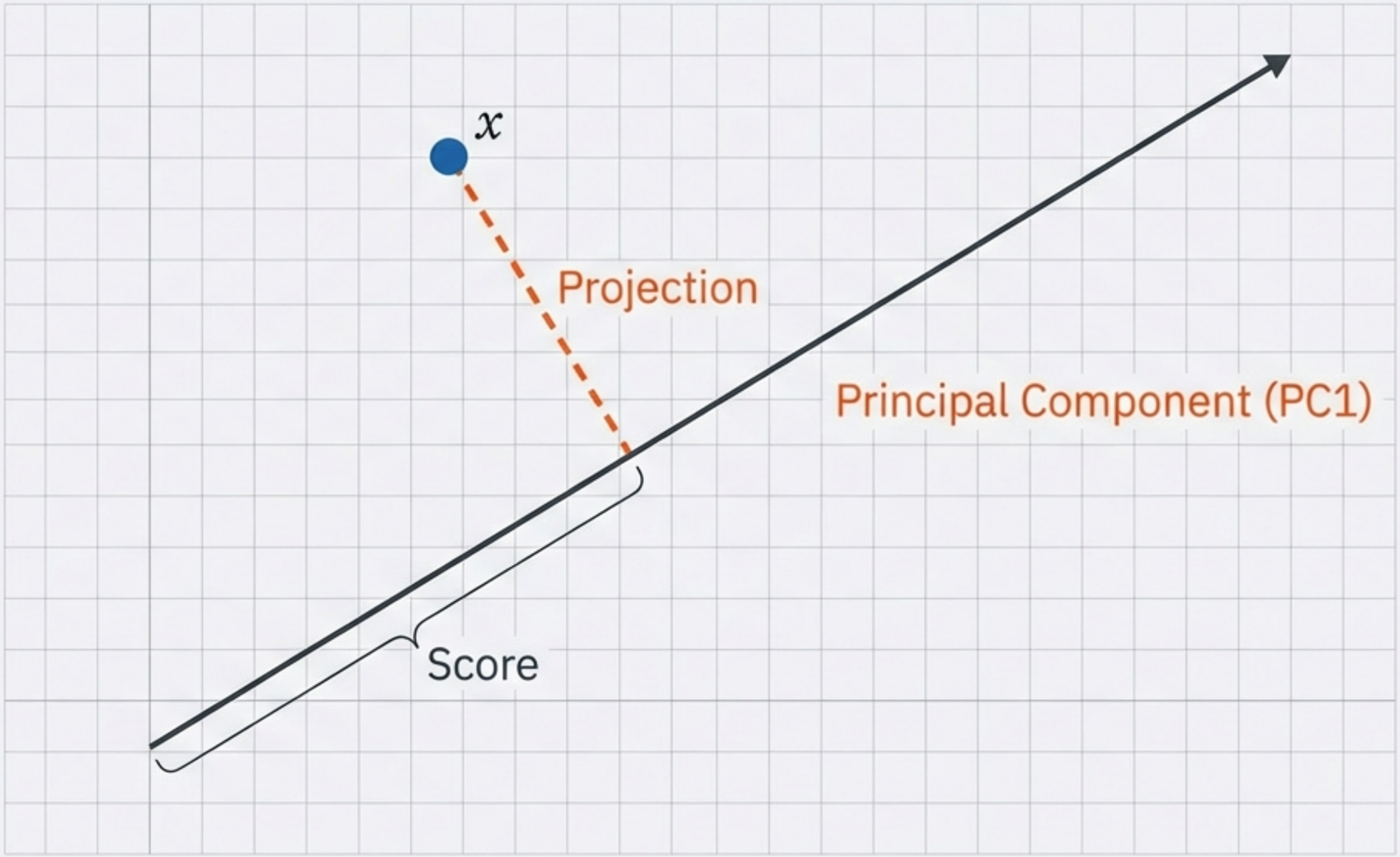
Just as a photographer moves to avoid heads blocking faces, PCA rotates the data to find the projection where points are most spread out.

The Vocabulary of PCA

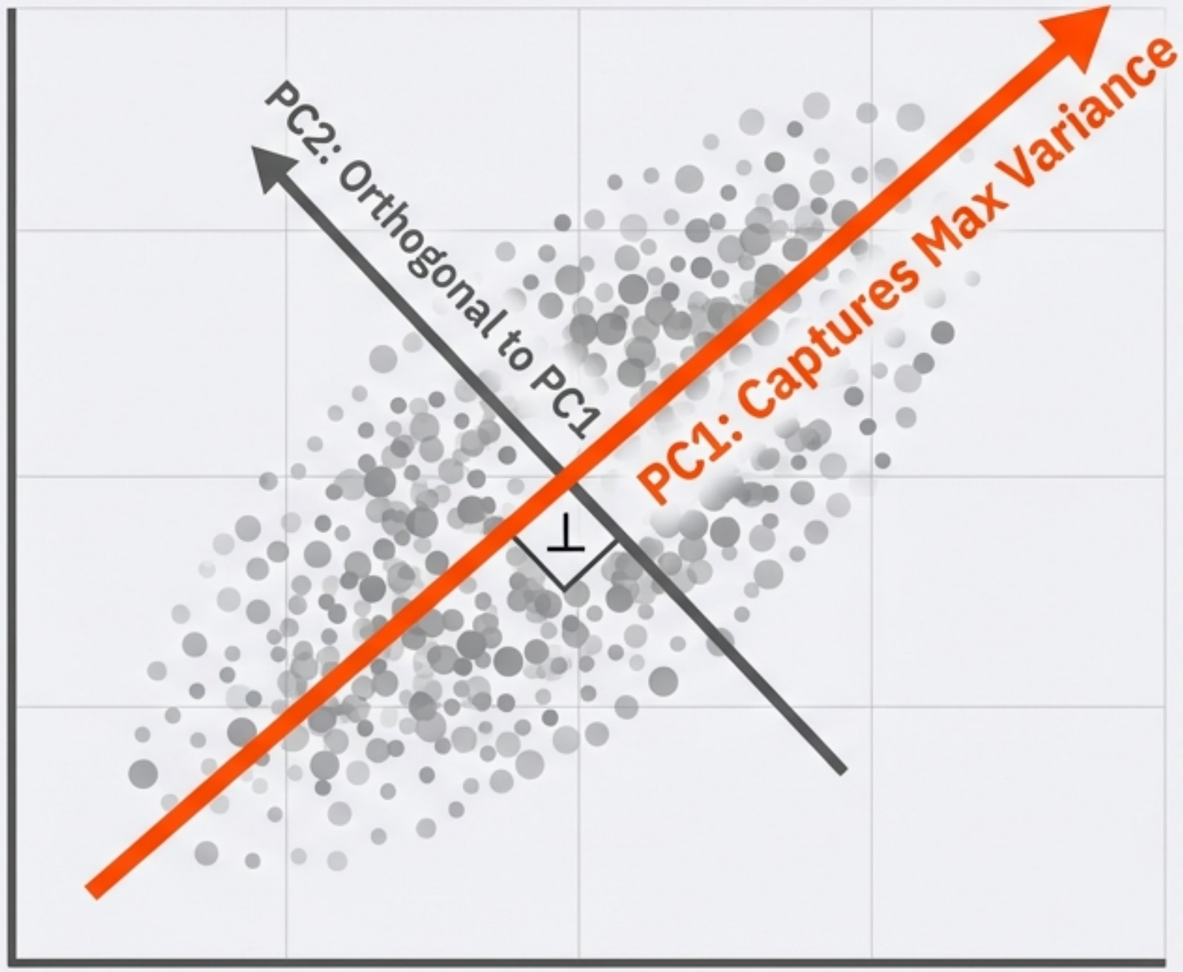
Dimensions
Number of columns/ features in the dataset.

Principal Component
New variable created by combining initial variables.

Projection
The perpendicular distance between a data point and the PC line.



Properties of Principal Components



Priority: PC1 always captures the most information. PC2 captures the second most.

Orthogonality: PC2 is always 90° to PC1. They are statistically independent.

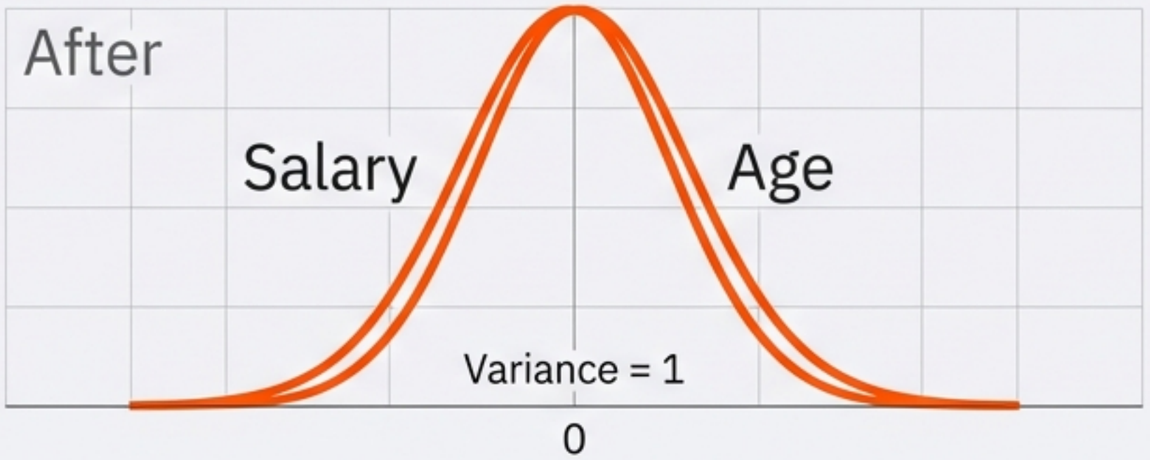
Constraint: Number of PCs ≤ Original Attributes.

How PCA Works: The Workflow

Step 1: Standardization

We must scale attributes so they lie within similar boundaries. Without this, a variable with a large range (like “Salary”) dominates a variable with a small range (like “Age”).

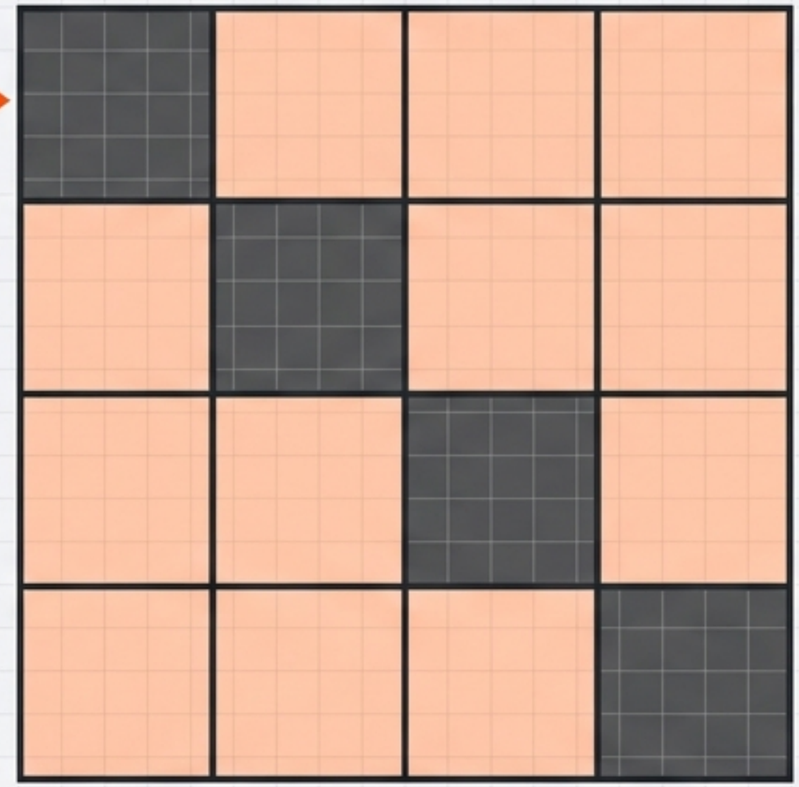
$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$



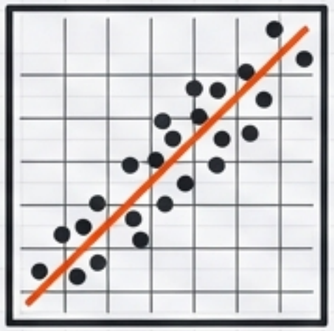
Step 2: Covariance Matrix Computation

The matrix expresses the correlation between attributes.

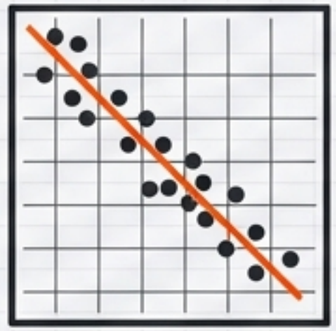
Variance (x to x)



Covariance (x to y)



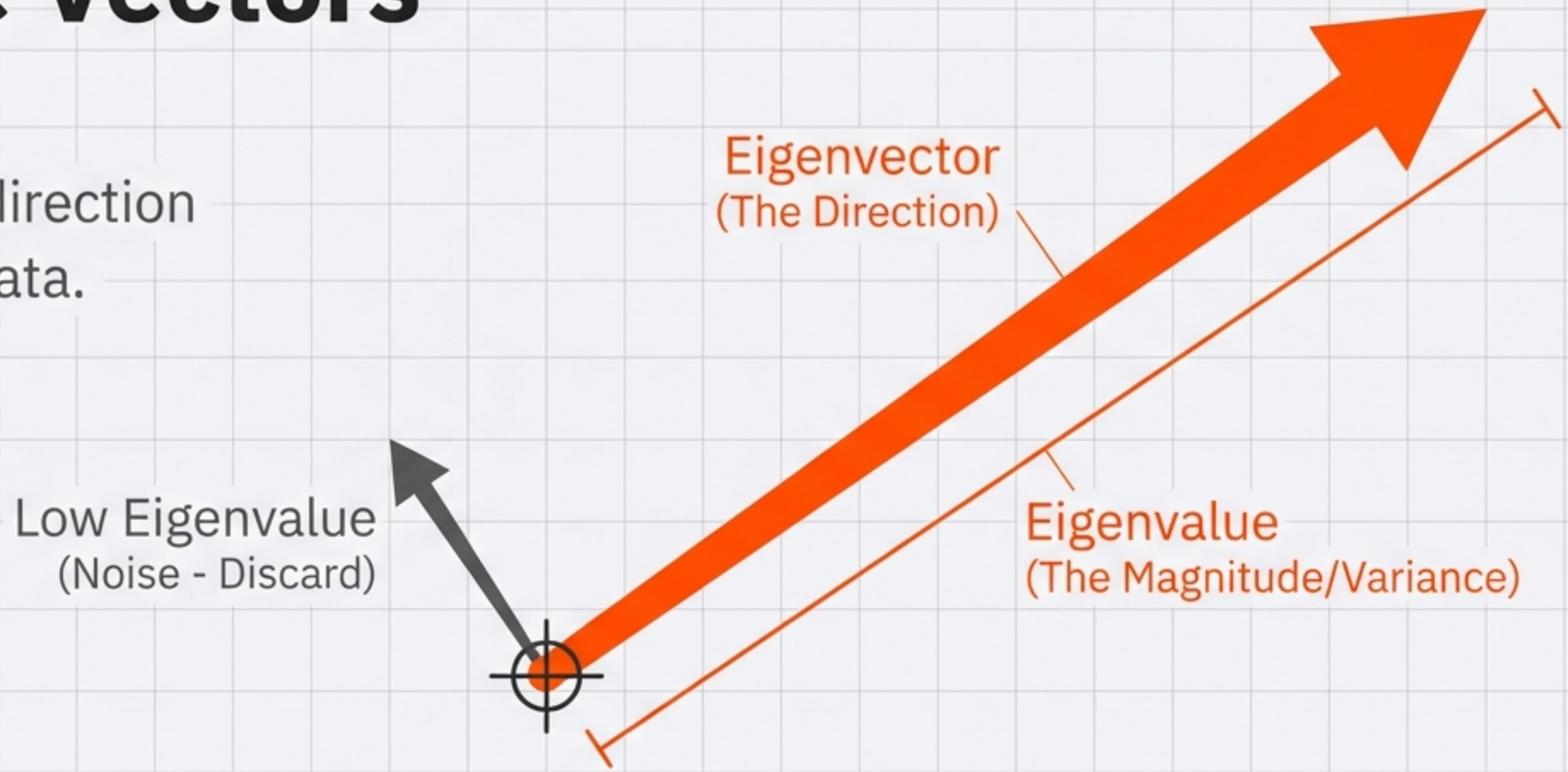
Positive Covariance
(Directly Proportional)



Negative Covariance
(Inversely Proportional)

Step 3 & 4: Eigenvectors and Feature Vectors

The math extracts the direction and magnitude of the data.



Feature Vector = Keeping the components with high Eigenvalues.

Case Study: Breast Cancer Detection

Using Scikit-Learn to classify Malignant (0) vs. Benign (1) tumors.

```
['mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean
 smoothness', 'mean compactness', 'mean concavity', 'mean concave points',
'mean symmetry', 'mean fractal dimension', 'radius error', 'texture
'mear', error', 'primeter' e                                          ror',
'perimeter error', 'area err                                          ,
'compactness error', 'concav                                          rror',
'concavity error', 'concave                                           rror',
'fractal dimension error', '                                           area',
'worst radius', 'fractal dim                                          ror',
'worst radius', 'worst textu                                          rea',
'worst texture', 'worst perimeter', 'worst area', 'worst smoothness', 'worst
'compactness', 'worst concavity', 'worst concave points', 'worst
'symmetry', 'worst fractal dimension']
```

The Challenge: 30 Dimensions.
Impossible to visualize.
Impossible to find patterns with the naked eye.

Implementation in Python

1. **Scale:**
   Normalize the 30 features.

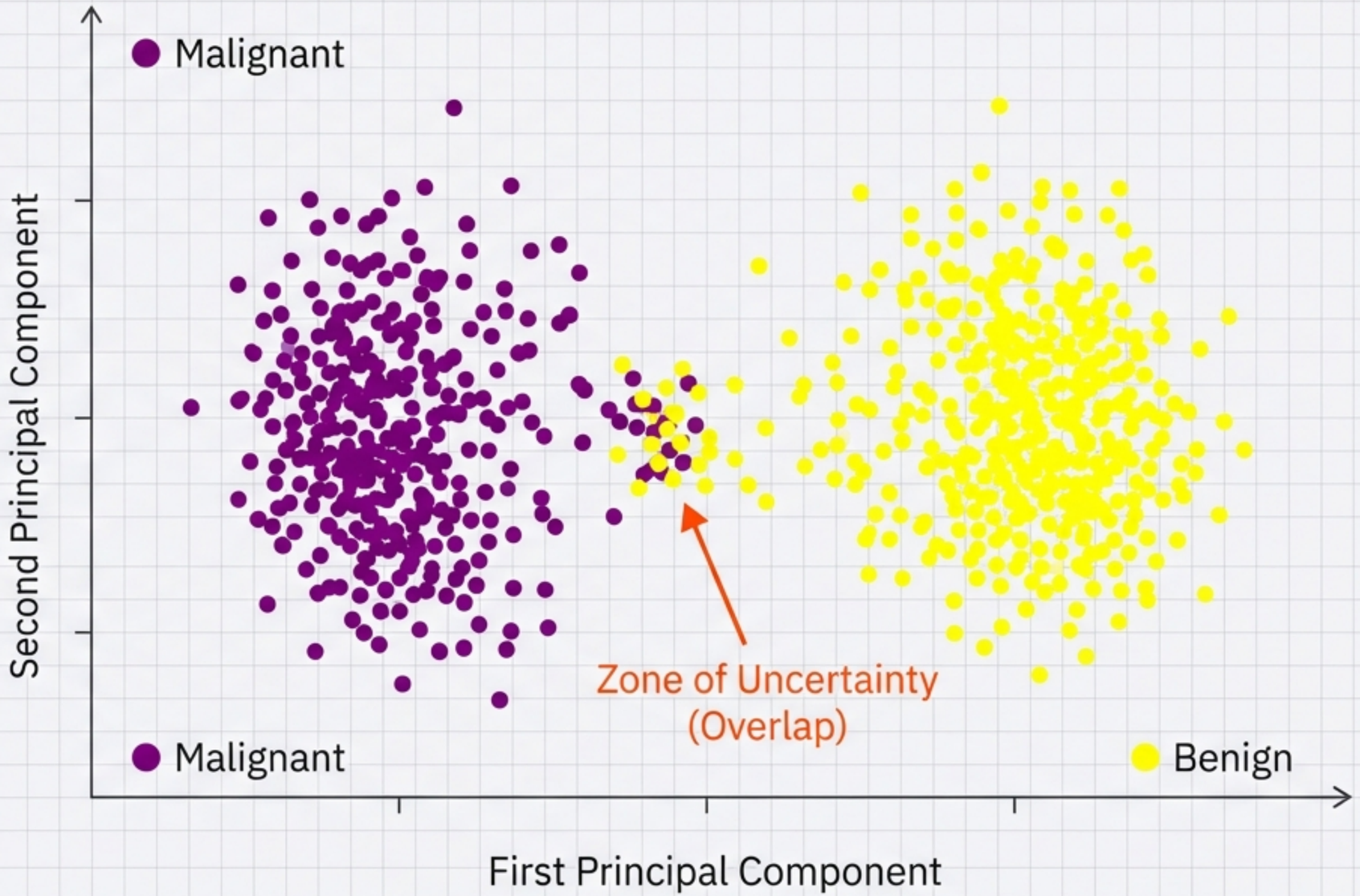
2. **Initialize:**
   Set PCA to find 2 components.

3. **Fit & Transform:**
   Compress 30 dims → 2.

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# 1. Scale the data
scaler = StandardScaler()
scaler.fit(df)
scaled_data = scaler.transform(df)

# 2. Initialize PCA
pca = PCA(n_components=2)

# 3. Fit and Transform
pca.fit(scaled_data)
x_pca = pca.transform(scaled_data)
```

Result: From 30 Dimensions to 2

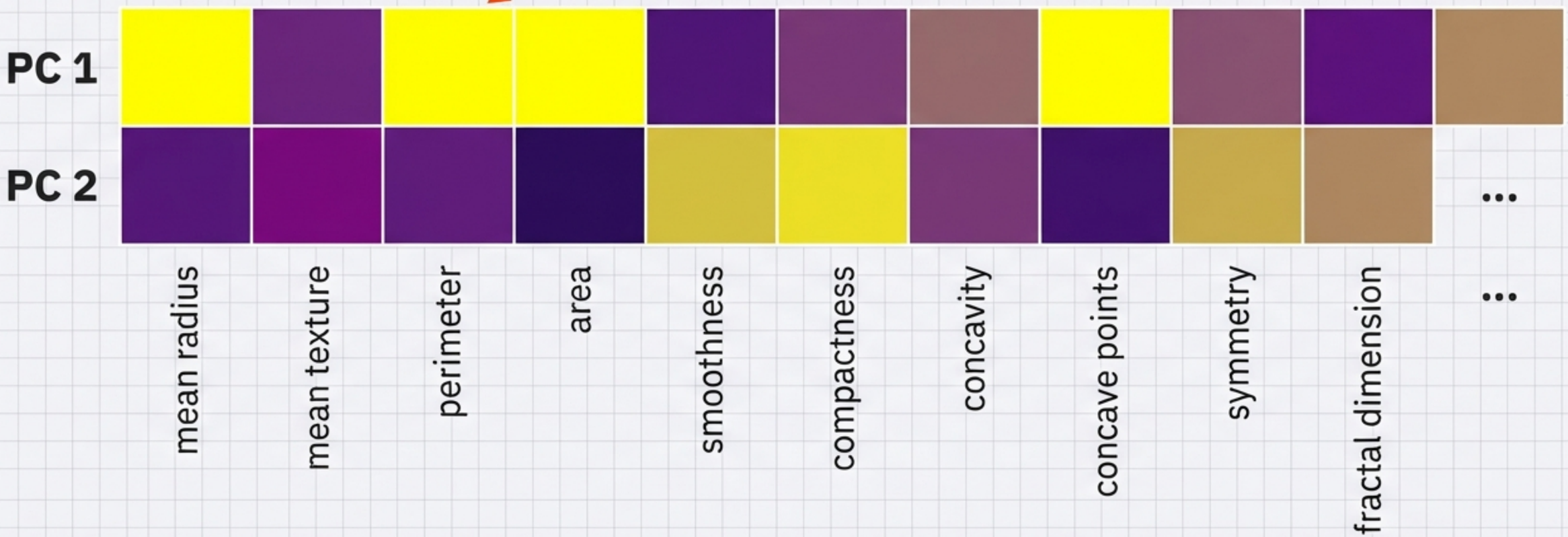
We can now see a clear separation between Benign and Malignant classes that was invisible before.



Interpreting the Black Box

Since ‘PC1’ is a new variable, we use a heatmap to see which original features contributed to it.

Radius and Perimeter are the strongest drivers of Principal Component 1.



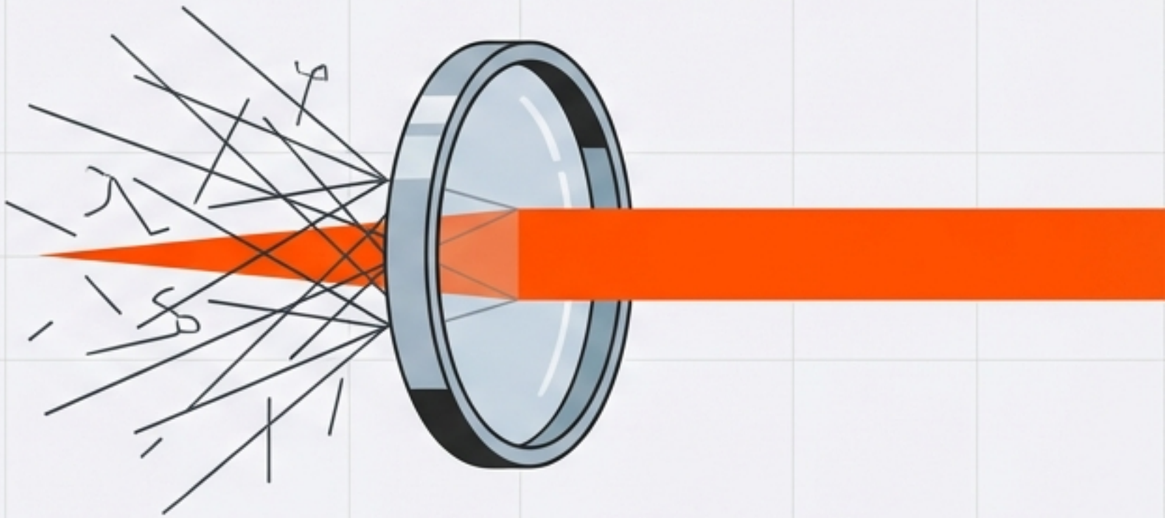
The PCA Trade-off

What we give up:

X

Direct interpretability of variables.

What we gain:



Processing speed, storage efficiency, and the ability to visualize the story hidden in the noise.

PCA changes the perspective to reveal the structure.