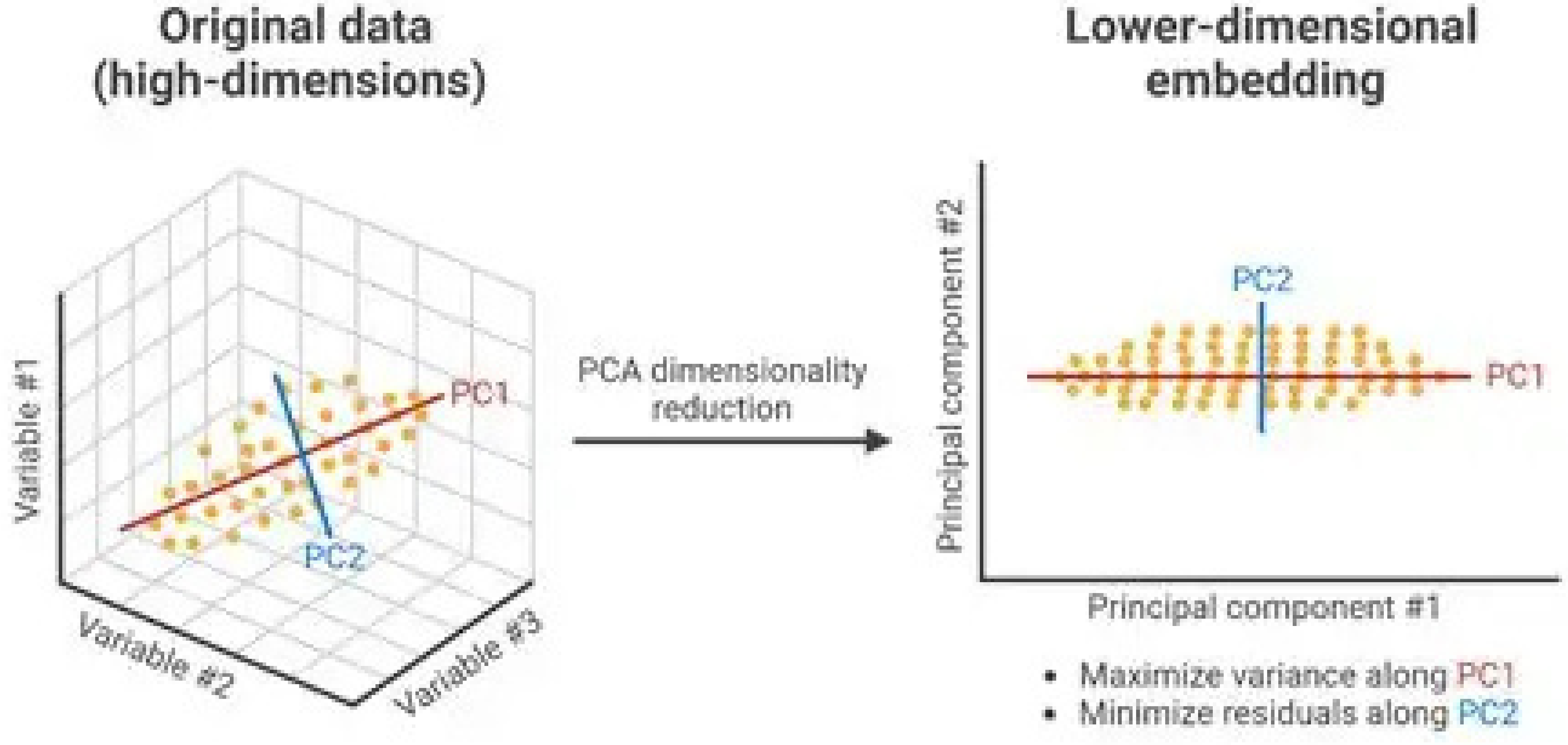
Principal Component Analysis (PCA) Transformation



England make history

The Game

12-PAGE PULLOUT

Henry Winter

Tony Cascarino

Matt Lawton

IN THE NEWS

NDEPENDENT

The history-makers

Extra-time victory puts England into first final since 196

What is Dimensionality Reduction?
Dimensionality reduction means reducing the number of input features (columns) while keeping important information.

Feature selction

Feature extraction

🎯 Simple Intuition

Feature Selection = Removing bad players from a team

Feature Extraction = Mixing all players to create super players example

🏠 Step 1: Original Dataset

Suppose we have this housing data:

House	No_of_Rooms	No_of_Grocery_Stores	Price (in $1000)
H1	2	5	50
H2	3	2	65
H3	4	8	80
H4	2	1	48
H5	5	6	95

- Correlation(Rooms, Price) = 0.90 ✅ (strong)
- Correlation(Grocery, Price) = 0.10 ❌ (very weak)

So:

👉 Grocery stores do not affect price much in this dataset.

New dataset:

House	No_of_Rooms	Price
H1	2	50
H2	3	65
H3	4	80
H4	2	48
H5	5	95

✅ We removed **No_of_Grocery_Stores**

🔥 What Happened?

- We did NOT transform data
- We did NOT combine features
- We simply removed one column

That is **Feature Selection**.

When:

- Features >> Observations
- Example: 10,000 features, 500 samples

Problems:

- Overfitting
- Sparse data
- Distance measures become meaningless

Feature extraction reduces dimension while preserving information.

PCA

Suppose we have 5 students and we record **their marks in 5 subjects:**

Student	Math	Physics	Chemistry	Biology	English
S1	85	80	78	70	90
S2	88	82	80	72	85
S3	60	65	70	68	75
S4	90	85	88	80	95
S5	55	60	58	65	70

Student	PC1 (Science)	PC2 (Non-Science)
S1	2.5	0.3
S2	2.7	0.1
S3	-0.5	0.2
S4	3.0	0.5
S5	-2.0	-1.0

◆ Step 5: Benefits

1. **Dimensionality Reduction** → simpler model
2. **Remove correlation** → Science marks combined in PC1
3. **Visualization** → 2D plot possible
4. **Noise reduction** → minor differences in individual subjects ignored

🏠 Example: House Dataset

Suppose we have 5 houses with 3 features:

House	Price ($1000)	Number_of_Rooms	Bathroom_Size (sq ft)
H1	50	2	30
H2	65	3	35
H3	80	4	40
H4	48	2	28
H5	95	5	50

- **3 features** → Price, Rooms, Bathroom size
- Rooms and Bathroom size are **correlated** (more rooms → bigger bathroom usually)
- We want to reduce features → **keep max info**

Suppose we reduce **3 features → 2 principal components (PC1, PC2)**

- **PC1**: Combination of Rooms and Bathroom_Size (main variance direction)
- **PC2**: Combination of Price (minor variance direction)

House	PC1	PC2
H1	-1.80	0.05
H2	-0.45	-0.12
H3	0.94	0.20
H4	-1.97	-0.05
H5	3.28	-0.08

Rooms	Original	Combined in PC1
Bathroom_Size	Original	Combined in PC1
Price	Original	Partially in PC2
# Features	3	2
Correlation	Rooms & Bathroom correlated	Components uncorrelated

House	Price	Rooms	Bathroom Size
H1	50	2	30
H2	65	3	35
H3	80	4	40
H4	48	2	28
H5	95	5	50

Step 1: Standardize the Data

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

- Compute mean & std:

Feature	Mean	Std
Price	67.6	18.86
Rooms	3.2	1.30
Bathroom	36.6	8.17

- Standardized matrix (combined for all features):

$$X_{\text{scaled}} = \begin{bmatrix} -0.92 & -0.92 & -0.81 \\ -0.13 & -0.15 & -0.20 \\ 0.66 & 0.62 & 0.42 \\ -1.04 & -0.92 & -1.06 \\ 1.43 & 1.54 & 1.65 \end{bmatrix}$$

Step 2: Compute Covariance / Correlation Matrix

$$C = \frac{1}{n-1} X_{\text{scaled}}^T X_{\text{scaled}}$$

Approx covariance matrix:

$$C = \begin{bmatrix} 1 & 0.99 & 0.98 \\ 0.99 & 1 & 0.97 \\ 0.98 & 0.97 & 1 \end{bmatrix}$$

> This shows features are highly correlated.

Step 3: Calculate Eigenvalues (c_{λ_i})

- Solve:

$$\det(C - \lambda I) = 0$$

- Approximate eigenvalues:

$$c_{\lambda_1} = 2.89, \quad c_{\lambda_2} = 0.09, \quad c_{\lambda_3} = 0.02$$

Step 3a: Select Top Eigenvalues

- We want to **reduce 3 features → 2 principal components**
- Select **top 2 eigenvalues**:

$$c_{\lambda_1} = 2.89, \quad c_{\lambda_2} = 0.09$$

Reason for selection:

- Largest eigenvalues capture **most variance** in the dataset
- Smallest eigenvalue (0.02) corresponds to **direction with negligible variance**, can be ignored

Step 4: Calculate Eigenvectors for Selected Eigenvalues

- For each selected eigenvalue, solve:

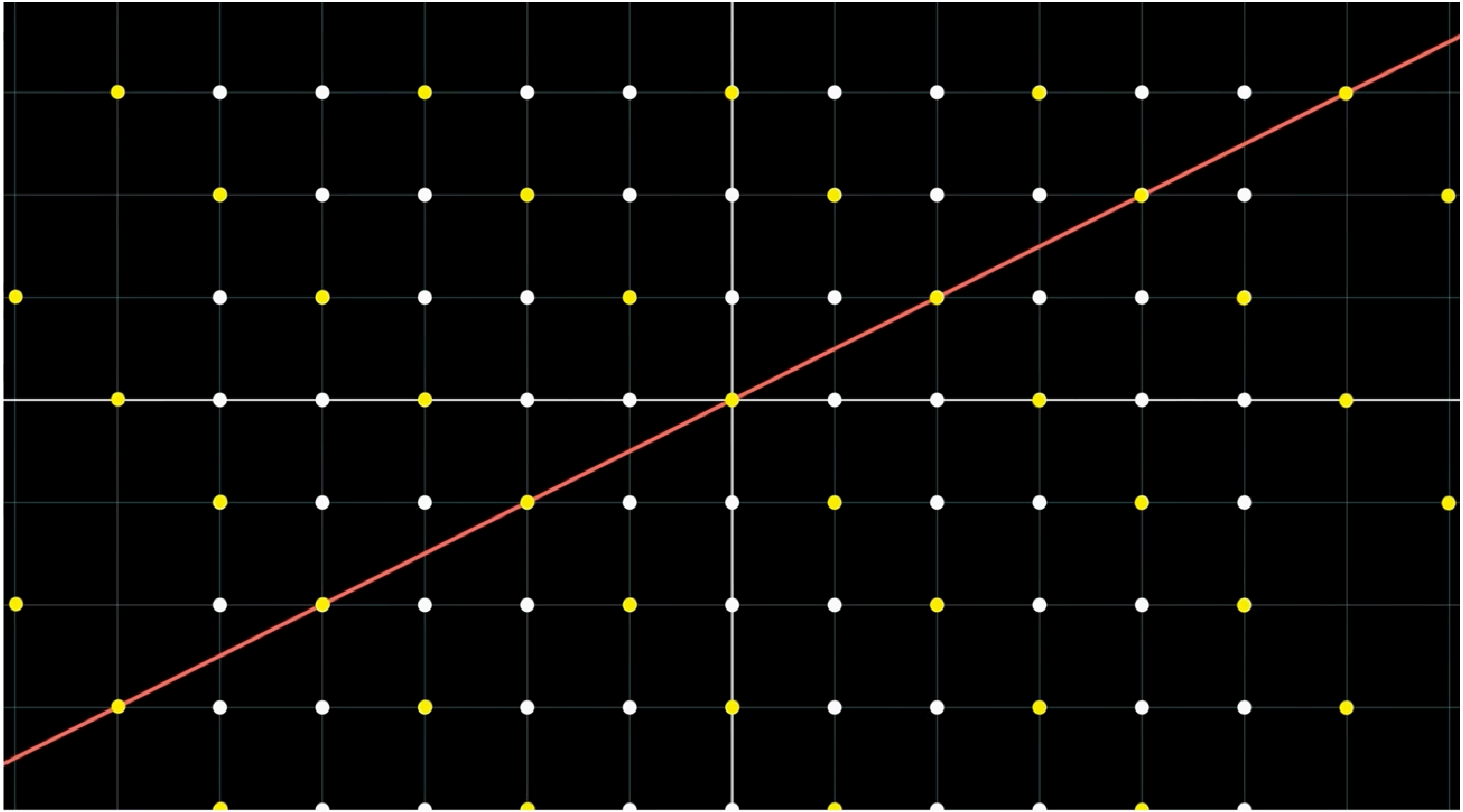
$$(C - \lambda_i I)v_i = 0$$

- Eigenvectors corresponding to top 2 eigenvalues:

$$v_1 = [0.58, 0.57, 0.58] \quad (\text{for } c_{\lambda_1})$$

$$v_2 = [-0.3, -0.7, 0.64] \quad (\text{for } c_{\lambda_2})$$

> Each eigenvector shows a **direction in original feature space** along which data varies.



Step 5: Form Projection Matrix

- Combine eigenvectors of selected eigenvalues:

$$W = [v_1 \ v_2] = \begin{bmatrix} 0.58 & -0.30 \\ 0.57 & -0.70 \\ 0.58 & 0.64 \end{bmatrix}$$

Step 6: Project Data onto Principal Components

$$Z = X_{\text{scaled}} \cdot W$$

scaled

Resulting 2D data (5×2):

House	PC1	PC2
H1	-1.43	-0.05
H2	-0.17	0.10
H3	0.90	0.08
H4	-1.56	-0.03
H5	2.26	-0.10

✅ **3 features reduced → 2 principal components**

✅ PC1 captures **largest variance** from c_{λ_1}

✅ PC2 captures residual variance from c_{λ_2}