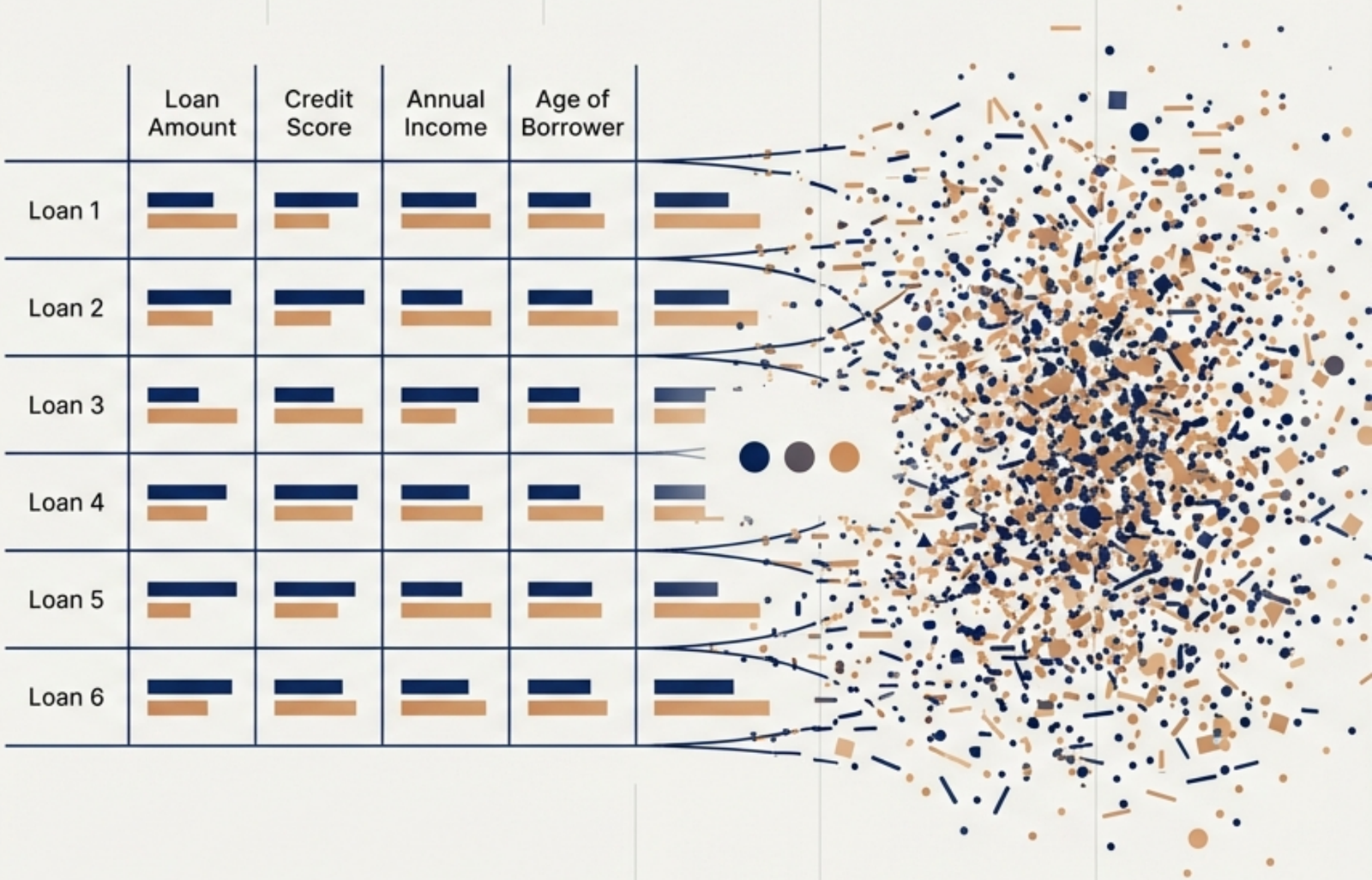
Tamising Complexity

A Guide to Principal Component Analysif Analysis

We're drowning in dimensions.

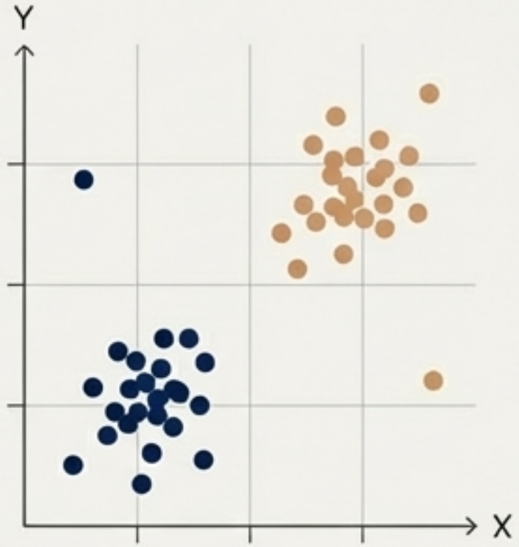
Consider a risk management scenario: we want to understand which loans have similarities to understand risk. A single loan can have hundreds of dimensions: loan amount, credit score, age of borrower, debt-to-income ratio, and many more. How do we find the pattern in all this noise?



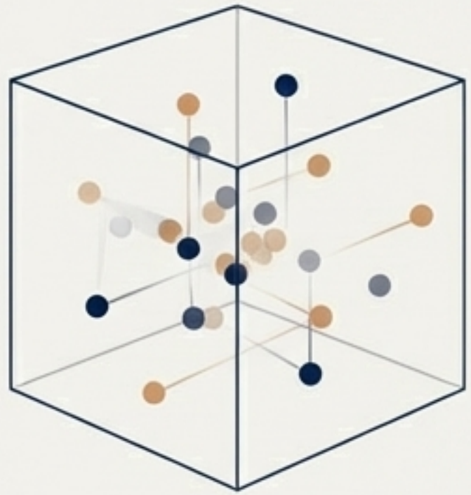
More data can lead to more problems: The Curse of Dimensionality.

Our intuition for space and distance fails as we add dimensions. Our ability to **visualize patterns**, train models, and find meaning plummets.

1D: Simple to interpret.



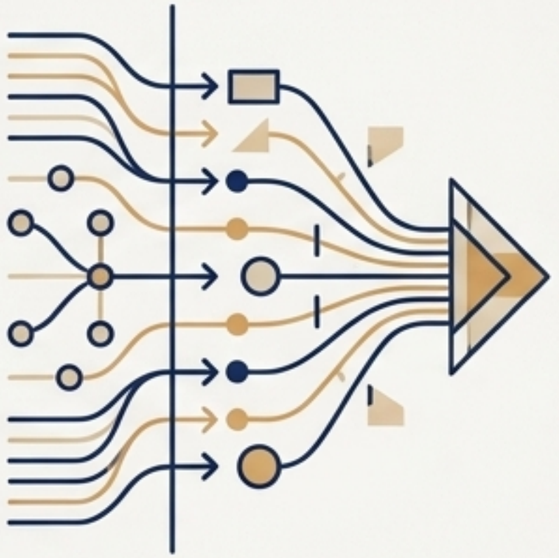
2D: Clear clusters emerge.



3D: Getting complex.

4D+: Visualization becomes impossible.

1901

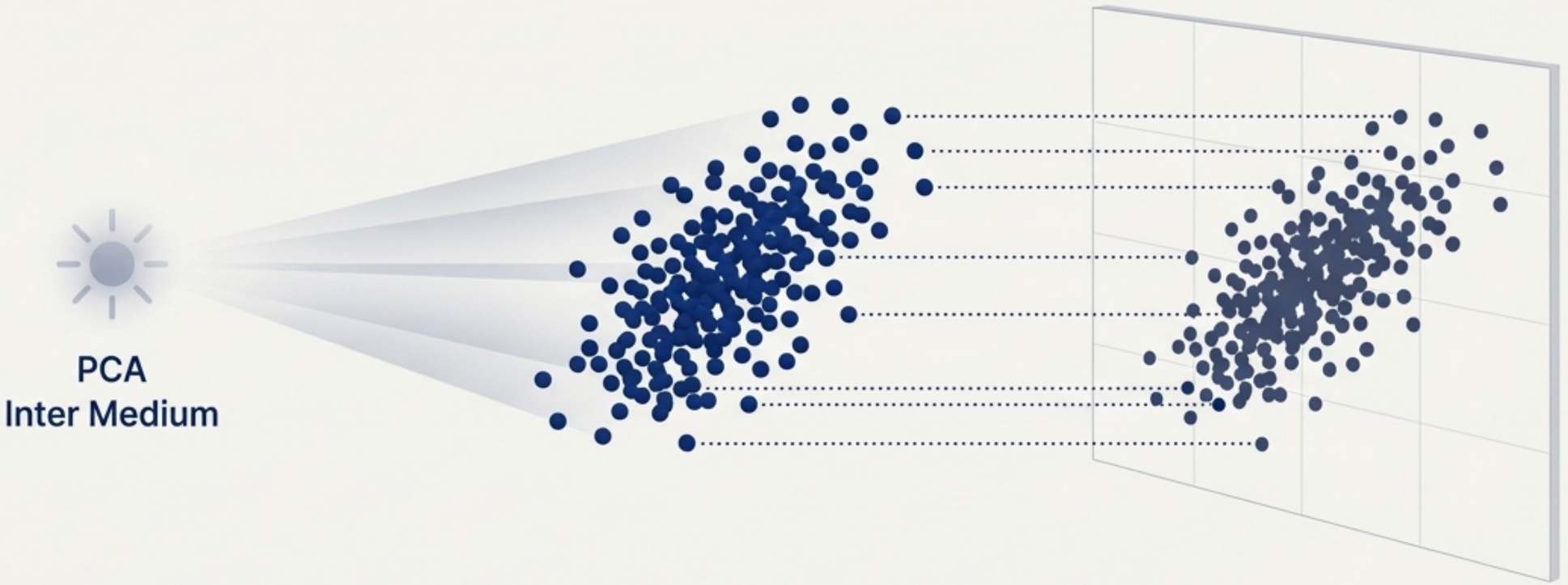


Today

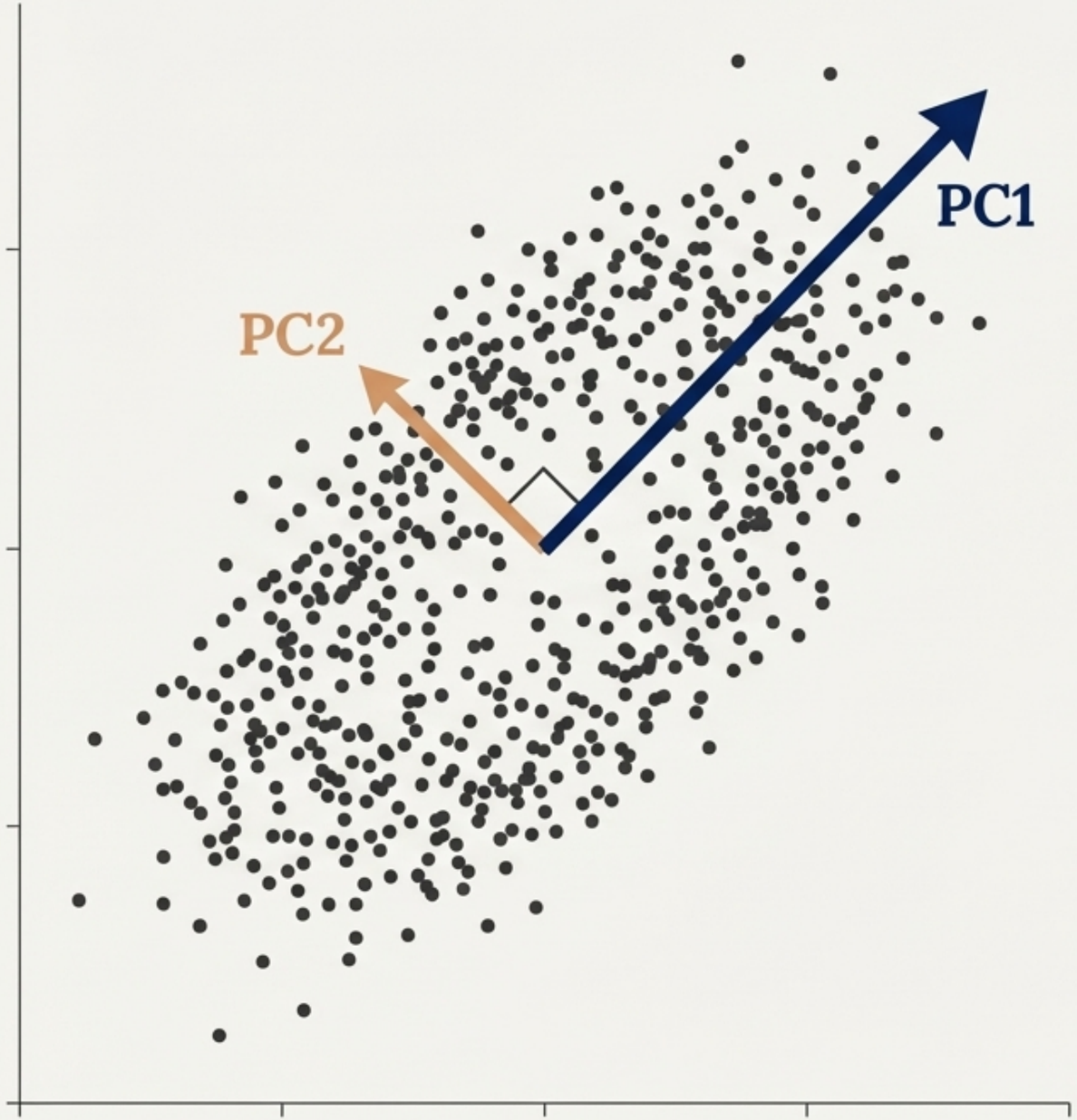
An idea for our time, over a century in the making.

Introducing **Principal Component Analysis (PCA)**. First developed by Karl Pearson in 1901, PCA is a technique to reduce the number of variables in a dataset while preserving as much of the original information as possible. It's a timeless idea, now supercharged by modern computing for today's machine learning challenges.

The core idea: finding the directions of maximum variance.



At a high level, PCA summarizes the information of a large dataset into a smaller set of new, uncorrelated variables. It effectively "squishes" hundreds of dimensions down into a few that tell the most interesting story about the data's shape and spread.



Meet the Principal Components.

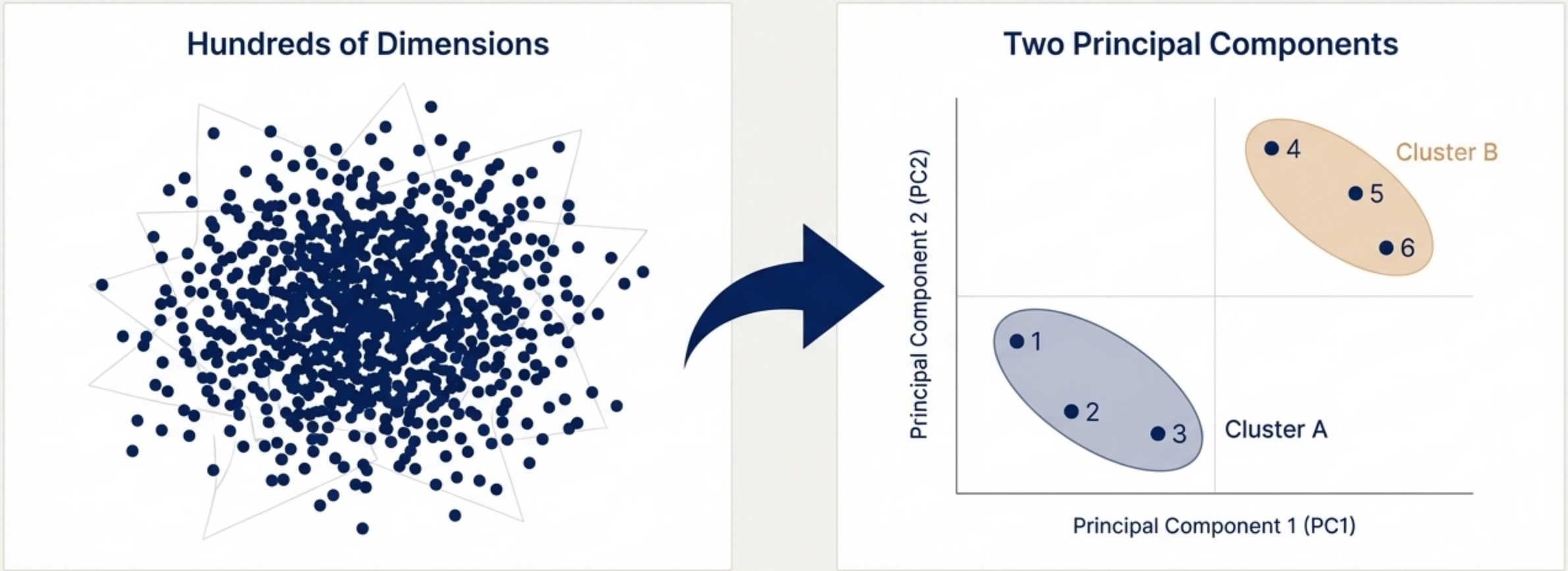
PCA creates new variables called Principal Components. The two major components are:

Principal Component 1 (PC1): The direction that captures the maximum possible variance in the data. It's the line that best represents the shape of the projected points. No other component can have higher variability.

Principal Component 2 (PC2): The direction that captures the next highest variance, but it must be uncorrelated with PC1 (i.e., perpendicular to it). The correlation between PC1 and PC2 is zero.

The loan data, reimagined.

By plotting the loan data against its first two principal components, the hidden structure becomes instantly clear. We've taken hundreds of potential dimensions and found the two that matter most for comparison.



The PCA Payoff: A simpler view leads to smarter results.

Reducing dimensions has four key advantages for data science and machine learning.

Accelerate ML Models

Faster training and inference with less data to process.

Visualize the Impossible

Visualize high-dimensional data by projecting it into a 2D or 3D space.

Fight Overfitting

Minimize the effects of overfitting, where models generalize poorly to new data.

Filter the Noise

Remove redundant information by focusing on components that capture the underlying patterns.

Use Case: Compressing images with a fraction of the data

PCA reduces image dimensionality while retaining essential information. By keeping only national. linaving only the most significant principal components, we can create compact representations of images, making them easier to store and transmit.

Original Image (100% of Components)

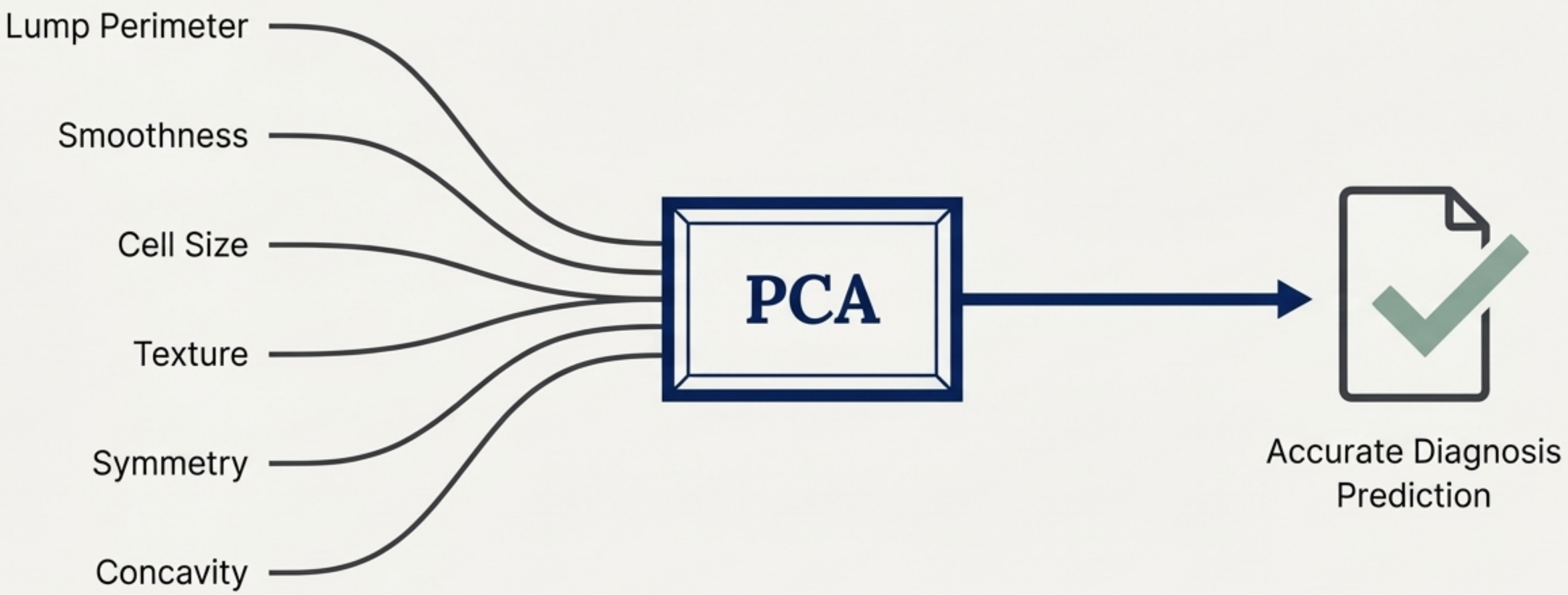
Reconstructed with 5% of components

Reconstructed with 25% of components

Reconstructed with 50% of components

Use Case: Finding life-saving signals in medical data

PCA can assist in diagnosing diseases earlier and more accurately. One study reduced six data attributes from a breast cancer dataset—like lump perimeter and smoothness—into a smaller feature space. A logistic regression model was then applied to this simplified data to more accurately predict whether cancer was present.



When should you use PCA?

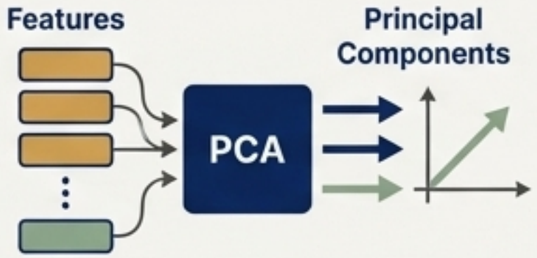
Consider using PCA when you need to:

- Visualize a dataset with more than 3 dimensions to understand its structure.
- Improve the speed and performance of a machine learning algorithm.
- Address issues of multicollinearity between your features.
- Perform noise reduction or extract the most informative features from your data.

Your toolkit for complex data.

The Problem

High-dimensional data is difficult to visualize, suffers from “The Curse of Dimensionality,” and can negatively impact model performance.



The Solution

PCA reduces dimensions by creating new, uncorrelated variables (Principal Components) that capture the maximum variance from the original data.

The Impact

This leads to faster models, clear visualizations of complex data, and powerful real-world applications in fields from finance to healthcare.

Find the simple story in your complex data.

What hidden patterns could you uncover?