Lecture: Dimensionality Reduction in Geoscientific Data

Dimensionality reduction is an essential concept when working with complex geoscientific data, where the goal is to simplify the data while retaining the most important information. In this lecture, we'll address three key questions:

1. What is the dimension of geoscientific data?
2. Why do we want to reduce the dimensionality?
3. How do we typically perform dimensionality reduction?

1. What is the Dimension of Geoscientific Data?

In geosciences, we often deal with data that spans multiple dimensions. For example:

- **Geospatial data** may have **2D or 3D** dimensions representing latitude, longitude, and depth or altitude.
- **Time series data** adds a **temporal dimension**, such as hourly temperature or seismic waveforms over time.
- **Multispectral data** includes **spectral bands**, such as data from satellites with many channels capturing different wavelengths of light.

While these are the *obvious* spatial, temporal, or spectral dimensions, the true **dimensionality of the data** refers to the number of variables required to describe or reconstruct the system accurately. This means we are interested in the **intrinsic dimensionality**, which represents the underlying variables that control the observable data.

Example:

Consider earthquake data. While a seismic waveform is a high-dimensional time series, much of the information about the earthquake's size can be distilled down to just a few key variables:

- **Magnitude** (which controls the amplitude of the waves).
- **Distance from the earthquake** (which controls the energy attenuation).
- **Wave propagation effects** (like site effects or regional differences).

Thus, even though the waveform itself might have thousands of points, the **dimensionality of the earthquake data** could be significantly lower if we only need a handful of variables to capture most of the variation in the signal.

- **Remove Redundant Information**: High-dimensional datasets often contain redundant information. For instance, two variables might be highly correlated, so one might suffice to represent both.
- **Simplify Models**: Fewer dimensions allow for simpler models, which are easier to interpret and less prone to overfitting. This is especially important when using machine learning techniques in geoscience (e.g., predicting seismic hazard or climate trends).
- **Improve Computational Efficiency**: Reducing dimensionality leads to faster computations and more manageable data storage requirements. This is crucial for handling large-scale geospatial or temporal datasets such as satellite images or climate models.
- **Noise Reduction**: High-dimensional data is often noisy. Dimensionality reduction techniques help to focus on the most significant patterns, filtering out noise.
- **Visualize Data**: In high-dimensional datasets, it's often challenging to visualize the