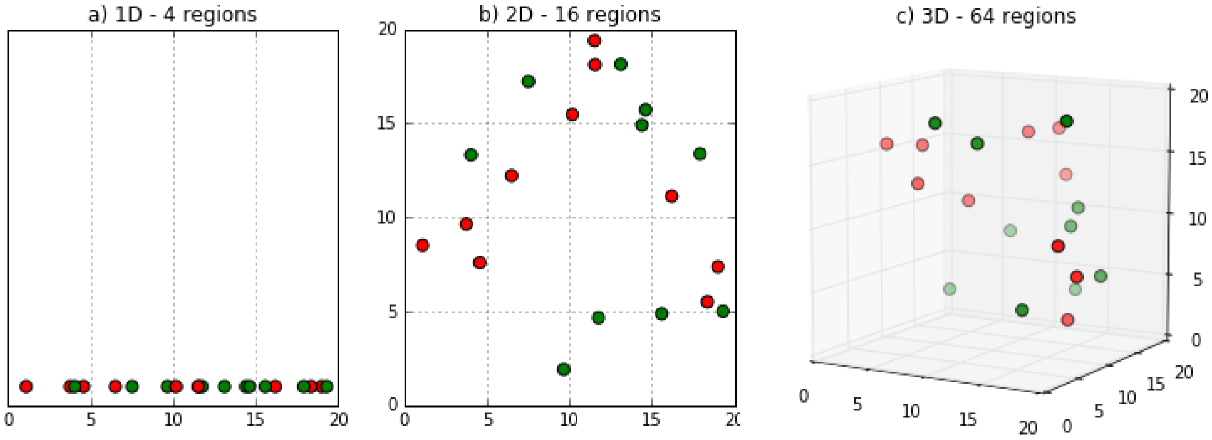
Dimensionality Reduction

CS 418. Introduction to Data Science

Dimensionality Reduction
Why Does it Matter? (I)

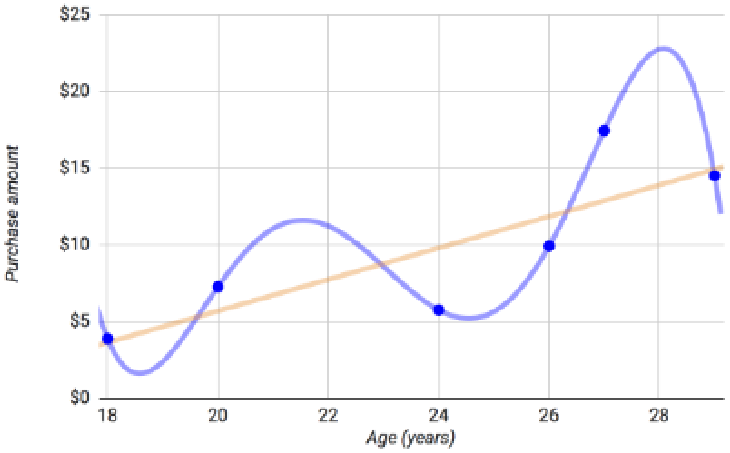
- Datasets can have a large number of variables.
- Data analysis becomes significantly harder as the number of variables (**dimensionality**) increases.
- **Curse of dimensionality.**
  - As the **dimensionality increases**, the data becomes **increasingly sparse** in the space that it occupies. Thus, the observations in the dataset are **not a representative sample** of all possible observations.



Source: KDnuggets, *Must-Know: What is the curse of dimensionality?* (2017)

Dimensionality Reduction
Why Does it Matter? (II)

- Datasets can have a large number of variables.
- Data analysis becomes significantly harder as the number of variables (**dimensionality**) increases.
- **Overfitting.**
  - An **overly complex** model (with too many variables or too many parameters) has a tendency to **overfit**; that is, to learn specific patterns from the training data that **do not generalize well to unseen data**.

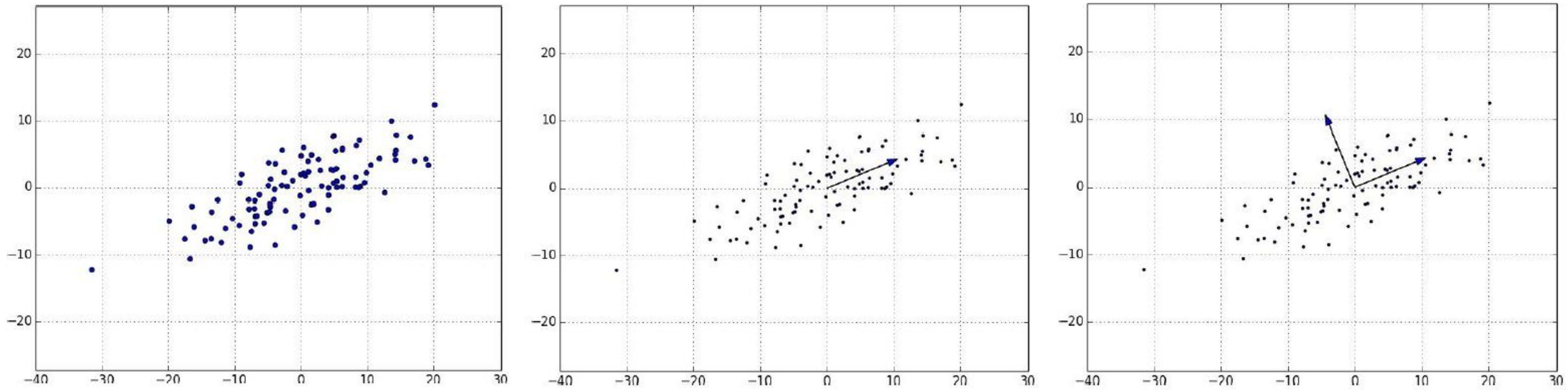


Dimensionality Reduction
Introduction

- **Dimensionality reduction** is the **reduction of the number of variables** in the dataset.
  - **Feature creation** or **feature extraction** is the reduction of the number of variables by **creating new variables** that are a combination of the original variables.
    - **Principal component analysis** (**PCA**).
  - **Feature selection** is the reduction of the number of variables by **selecting a subset of the original variables**.
    - **Univariate feature selection:** select the "best" k variables.
    - **Multivariate feature selection:** select the "best" **subset** of variables.
      - **Embedded approaches**: feature selection is part of the model building process (e.g., **LASSO regression**).
      - **Filter approaches:** variables are selected **before building the model** using some approach that is independent of the model building process.
      - **Wrapper approaches:** the results of the model are used to select the "best" subset of variables.

Dimensionality Reduction
Principal Component Analysis (I)

- **Principal component analysis** (**PCA**) is a **feature extraction** technique for **continuous** variables that creates new variables that are:
  - **Linear combinations** of the original variables.
  - **Orthogonal** to each other.
  - Capture the **maximum amount of variation** in the data.
- ***Example***:



Source: Joel Grus. *Data Science from Scratch* (2015)

Dimensionality Reduction
Principal Component Analysis (II)

- Mathematically, **PCA** is an **orthogonal linear transformation** that transforms the data from the original $\boldsymbol{p}$-dimensional space to a new $\boldsymbol{k}$-dimensional space such that the **variance is maximized**.
  - Since our goal is to **reduce the dimensionality** of the data, we generally choose $\boldsymbol{k < p}$.
- This transformation is defined by a set of **coefficients** that map each observation to the new space.
- Each new variable is given by:

$$\boldsymbol{Z_j = w_{j1}X_1 + w_{j2}X_2 + \cdots + w_{jp}X_p}$$

where:

  - Z_j with $j = 1, \ldots, k$ are the **new variables**.
  - $X_1, X_2, \ldots, X_p$ are the **original variables**.
  - $w_{j1}, w_{j2}, \ldots, w_{jp}$ are the **coefficients** for the jth new variable.

Principal Component Analysis
Advantages and Disadvantages

- What are some of the **advantages** of **PCA**?
  - Reduces the dimensionality of the data without eliminating any of the original variables.
  - New variables are **independent of one another**.
    - This satisfies the assumption of **no collinearity** made by multiple linear regression models.
- What are some of the **disadvantages** of **PCA**?
  - New variables are **hard to interpret** since they are linear combinations of the original variables.
  - **Sensitive to the scale** of the original variables.
    - We can address this limitation by **standardizing** the data so that every variable has a **mean of 0** and a **standard deviation of 1**.

Dimensionality Reduction
References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).
- Cathy O’Neil and Rachel Schutt. *Doing Data Science* (2013).
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. *Introduction to Data Mining* (2019).