Data Mining

Lecture 4

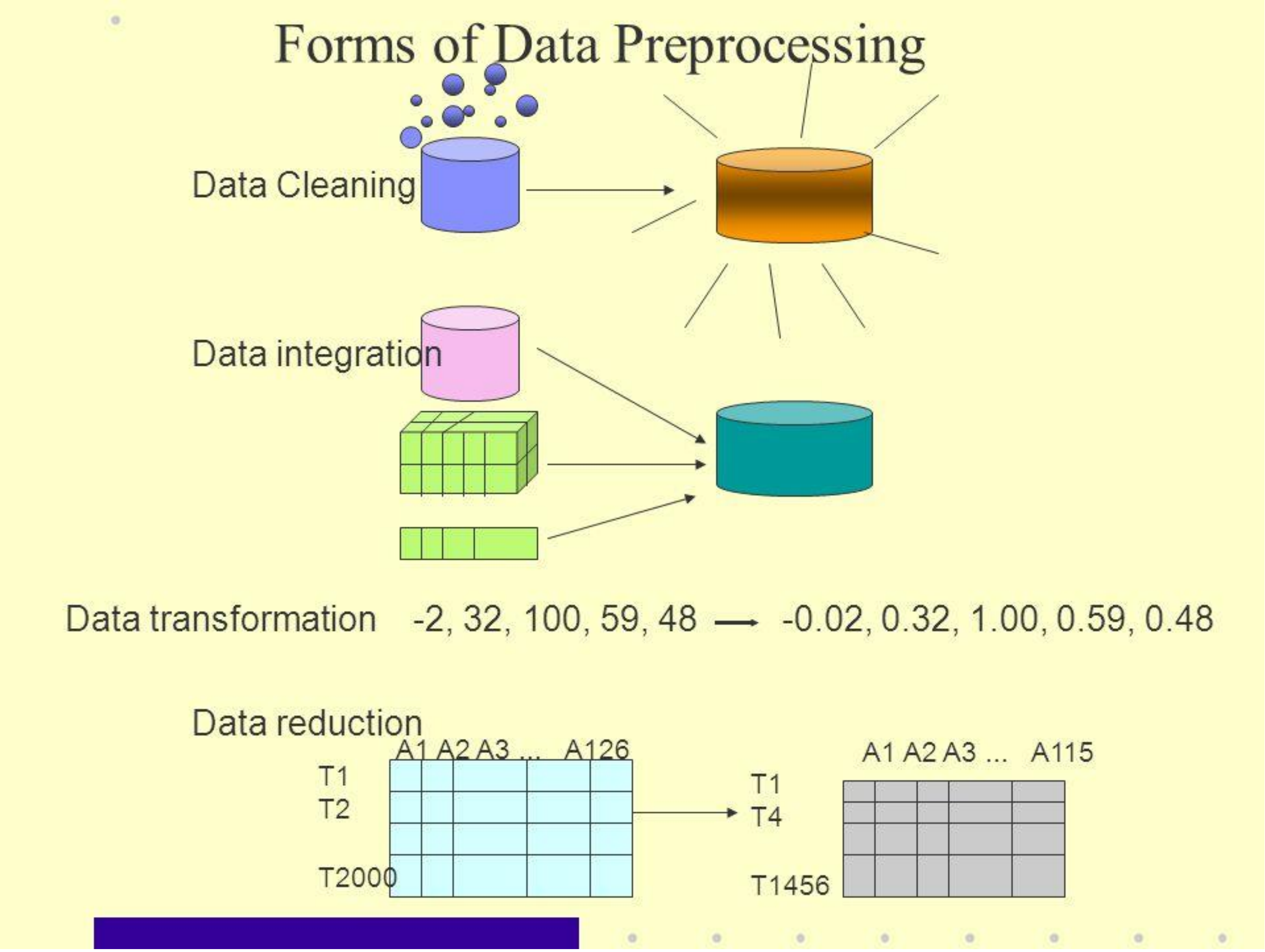
Data Preprocessing

Dr. Salem Othman

Summer 2023

Outline

- Data Preprocessing



Data Preprocessing

- Aggregation
- Sampling
- Discretization and Binarization
- Attribute Transformation
- Dimensionality Reduction
- Feature subset selection
- Feature creation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction - reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More “stable” data - aggregated data tends to have less variability

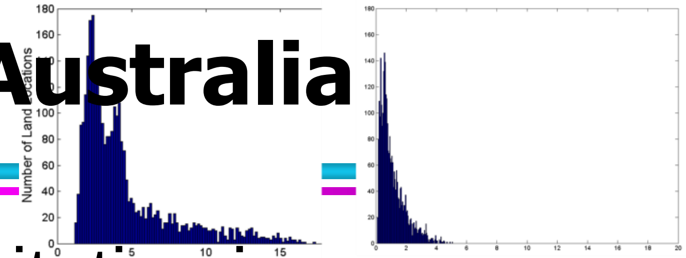
Table 2.4. Data set containing information about customer purchases.

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Example: Customer Purchases

- An obvious issue is how an aggregate transaction is created; i.e., how the values of each attribute are combined across all the records corresponding to a particular location to create the aggregate transaction that represents the sales of a single store or date.
- **Quantitative** attributes, such as **price**, are typically aggregated by taking a **sum** or an **average**.
- A **qualitative** attribute, such as item, can either be omitted or summarized in terms of a higher level category, e.g., televisions versus electronics.

Example: Precipitation in Australia



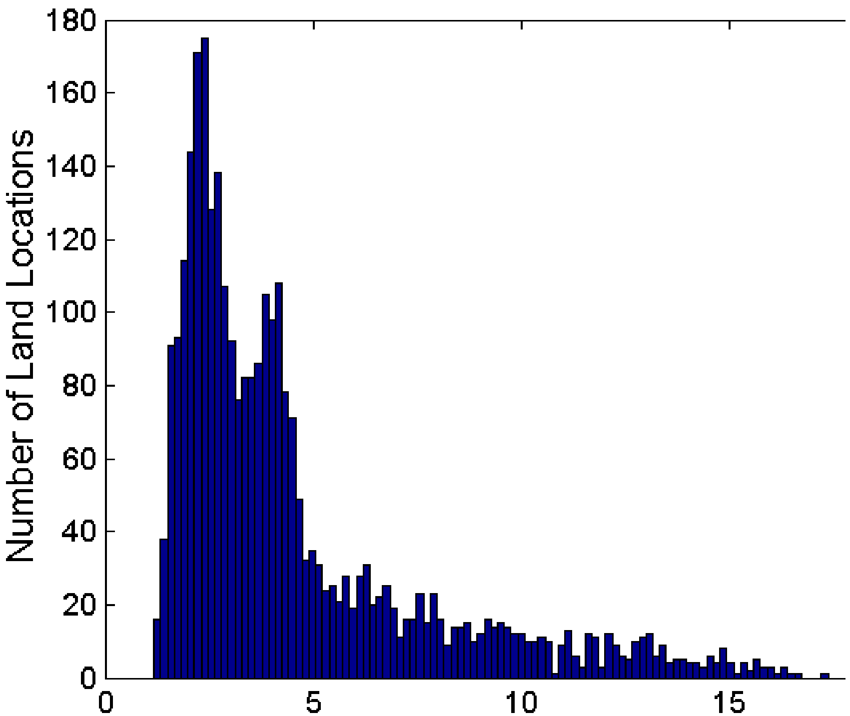
- This example is based on precipitation in Australia from the period 1982 to 1993.

  The next slide shows
  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5◦ by 0.5◦ grid cells in Australia, and

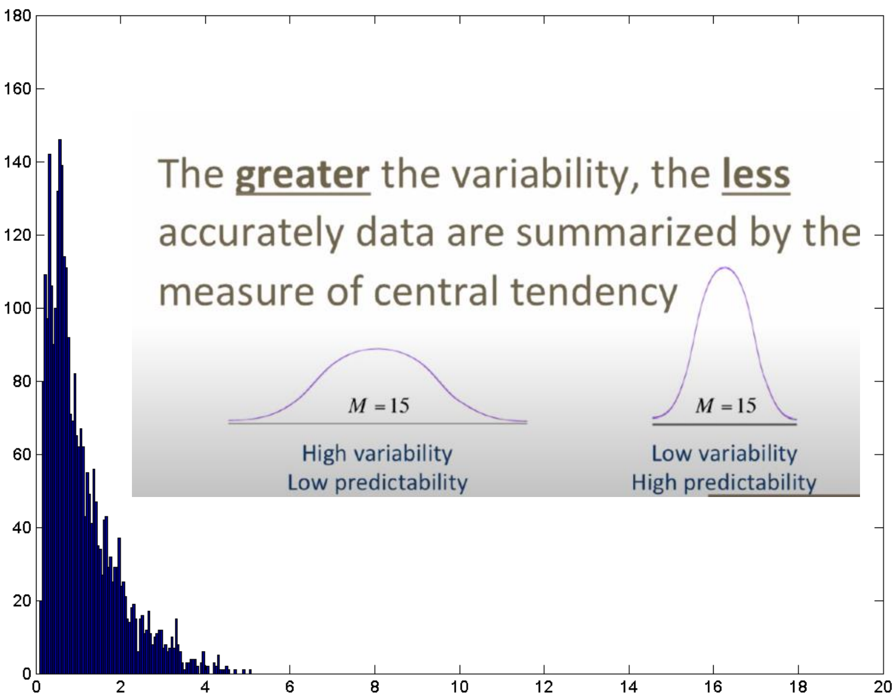
    How To Calculate The Standard Deviation - YouTube
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia …

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

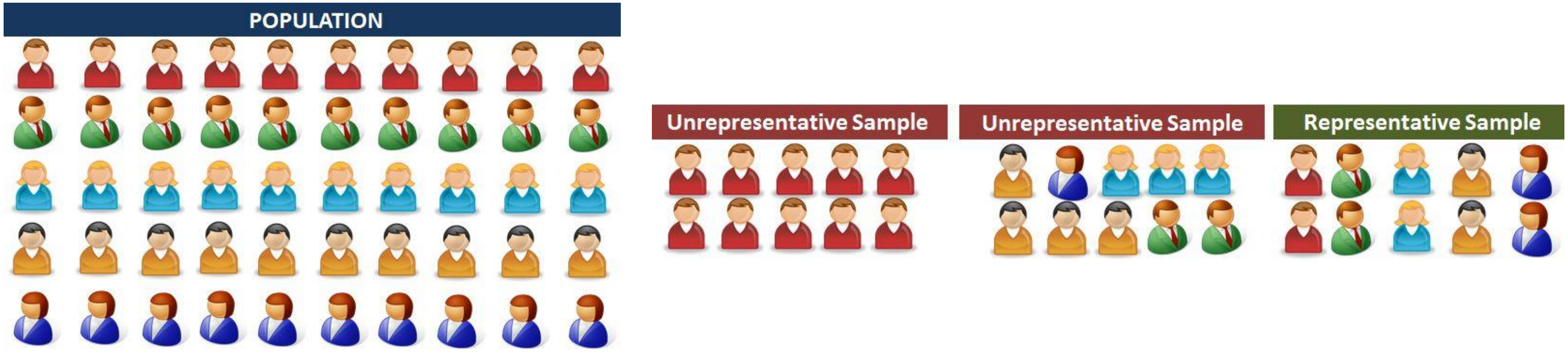
- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

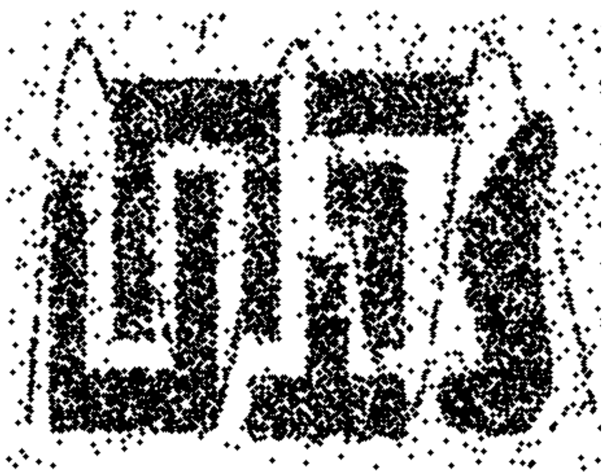
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Sampling …

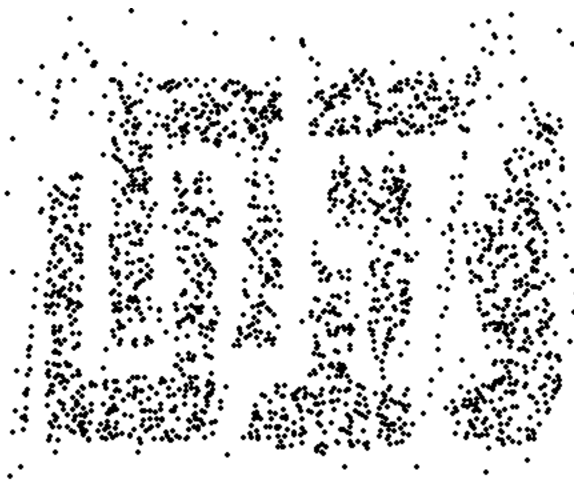
- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data



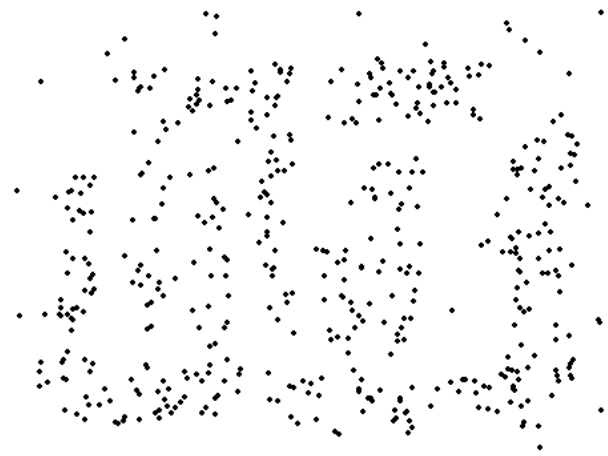
Sample Size



8000 points



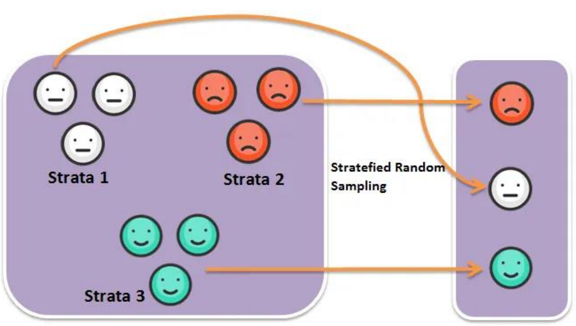
2000 Points



500 Points

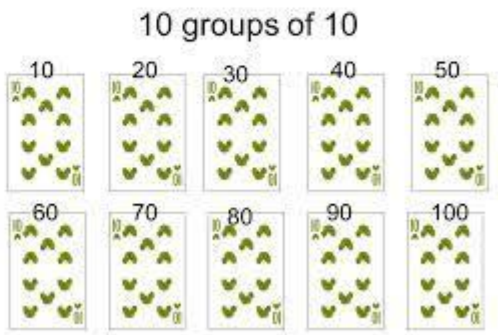
Types of Sampling

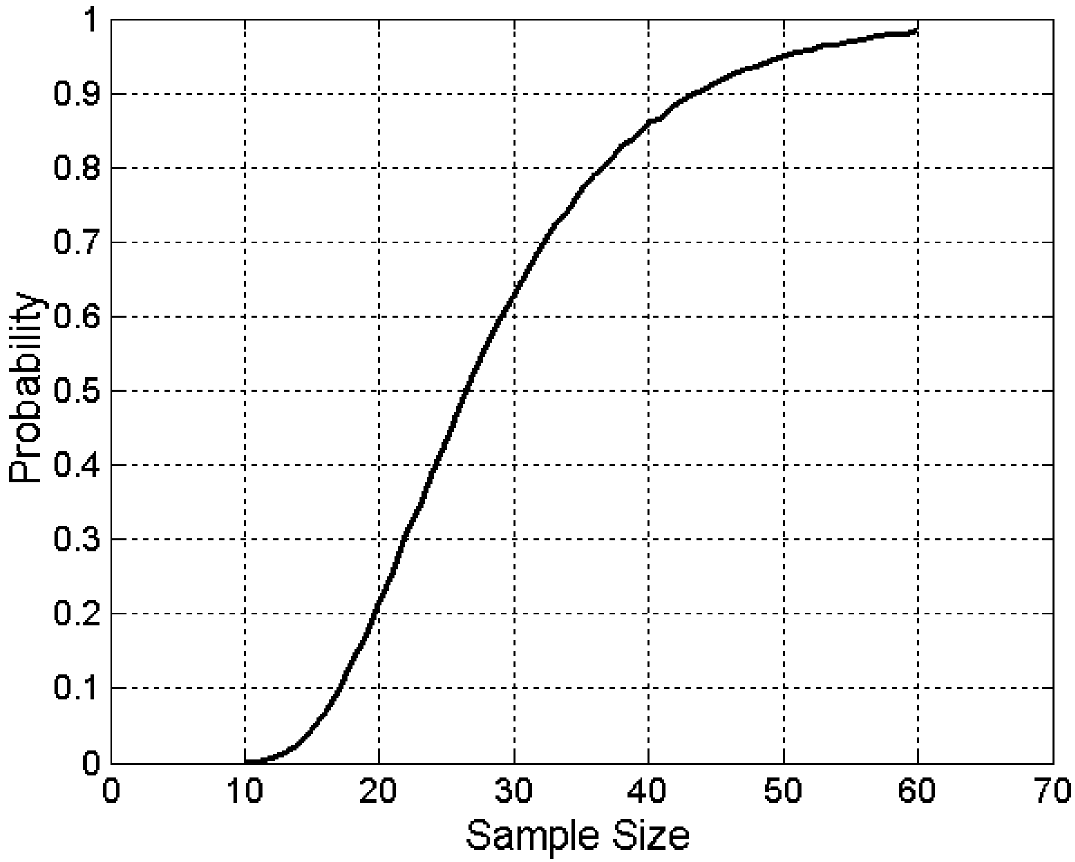
- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition



Sample Size

- **What sample size is necessary to get at least one object from each of 10 equal-sized groups.**

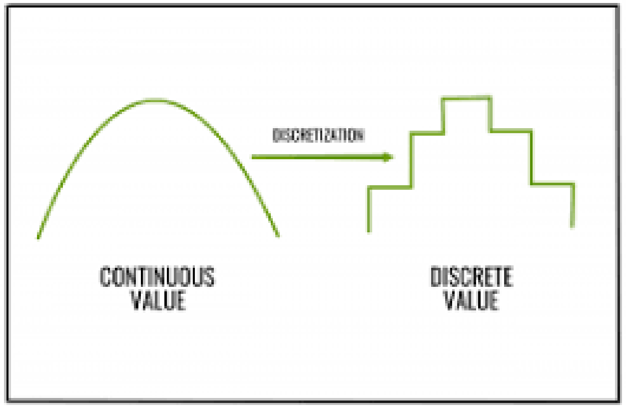




Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is used in both **unsupervised** and **supervised** settings

S.No	Age Group	Replaced Value
1	[17,30]	Young
2	[31,44]	Middle
3	[45,58]	Old



Discretization algorithms

Discretization algorithms can be divided into:

- unsupervised vs. supervised – unsupervised algorithms do not use class information

- static vs. dynamic

  Discretization of continuous attributes is most often performed *one attribute at a time, independent of other attributes – this is known as static attribute discretization.*

  Dynamic algorithm searches for all possible intervals for all features simultaneously.

Discretization Process

Any discretization process consists of two steps:

- 1st, the number of discrete intervals needs to be decided

Often it is done by the user, although a few discretization algorithms are able to do it on their own.

- 2nd, the width (boundary) of each interval must be determined

Often it is done by a discretization algorithm itself.

Discretization Problems

- Deciding the number of discretization intervals:

  large number – more of the original information is retained

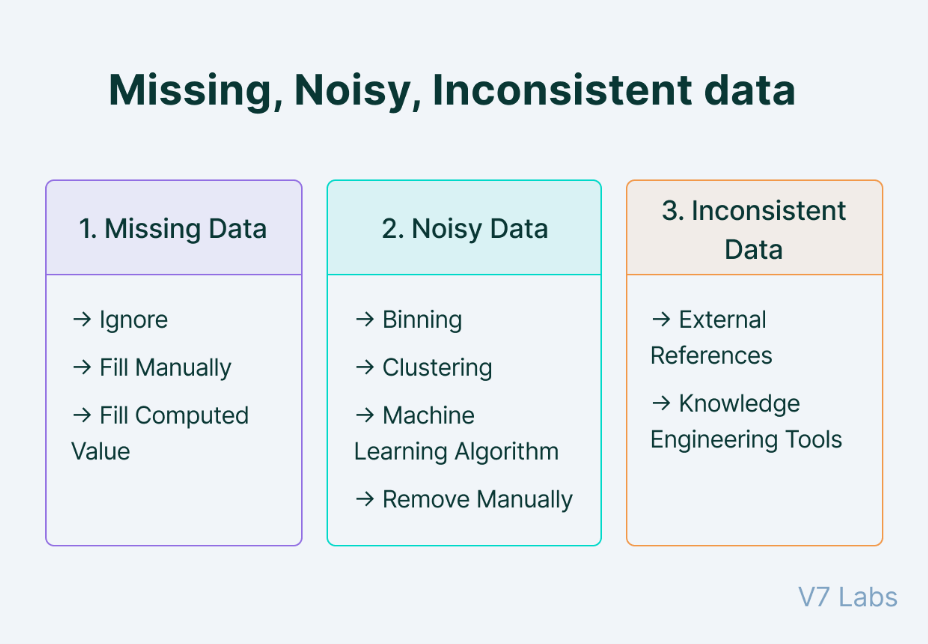
  small number – the new feature is “easier” for subsequently used learning algorithms

- Computational complexity of discretization should be low since this is only a preprocessing step

Discretization Algorithms

- Unsupervised Discretization Algorithms
  - Equal Width
  - Equal Frequency

- Supervised Discretization Algorithms
  - Information Theoretic Algorithms
    - CAIM
    - χ^2 Discretization
    - Maximum Entropy Discretization
    - CAIR Discretization
  - Other Discretization Methods
    - K-means clustering
    - One-level Decision Tree
    - Dynamic Attribute
    - Paterson and Niblett



Equal Width Binning

$$X = [10, 15, 16, 18, 20, 30, 35, 42, 48, 50, 52, 55]$$

$$w = \left\lfloor \frac{max - min}{x} \right\rfloor$$

Categories : $[min, min + w - 1], [min + w, min + 2 * w - 1], [min + 2 * w, min + 3 * w - 1] \cdots [min + (x - 1) * w, max]$

```
Notations,
x = number of categories
w = width of a category
max, min = Maximum and Minimun of the list
```

x = 4	
w = (55-10)/4 = 12	
[min, min+w-1]	[10, 21]
[min+w, min+2*w-1]	[22, 33]
[min+2*w, min+3*w-1]	[34, 45]
[min+3*w, max]	[46, 55]

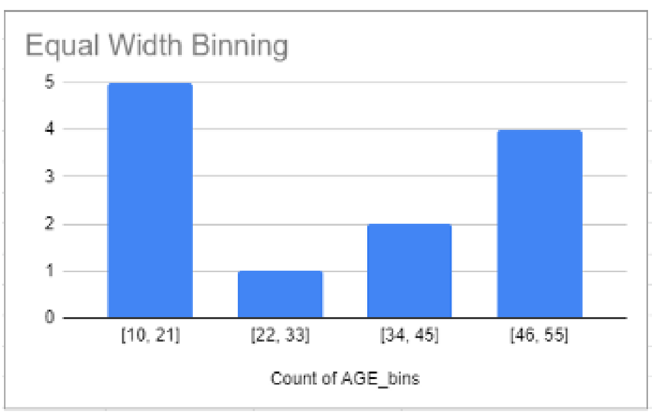
AGE	AGE_bins
10	[10, 21]
15	[10, 21]
16	[10, 21]
18	[10, 21]
20	[10, 21]
30	[22, 33]
35	[34, 45]
42	[34, 45]
48	[46, 55]
50	[46, 55]
52	[46, 55]
55	[46, 55]

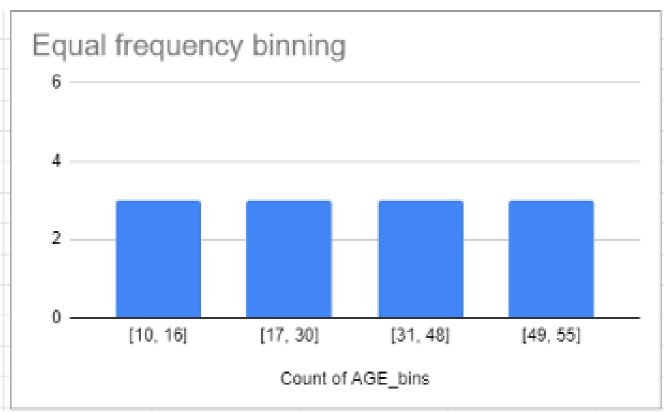
[Feature Engineering — deep dive into Encoding and Binning techniques | by Satyam Kumar | Towards Data Science]

Equal frequency binning

$$freq = \frac{n}{x}$$

AGE	AGE_bins
10	[10, 16]
15	[10, 16]
16	[10, 16]
18	[17, 30]
20	[17, 30]
30	[17, 30]
35	[31, 48]
42	[31, 48]
48	[31, 48]
50	[49, 55]
52	[49, 55]
55	[49, 55]





Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

- Typically used for association analysis

- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Applications

- Binarization in Natural Language Processing
- Binarization in digital image processing

Handling Categorical Variables

1. **Label Encoding**: This is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.
2. **Ordinal Encoding**: Similar to label encoding, but here the order matters. We encode the labels based on the order.
3. **Binary Encoding**: This method first converts the categories into numeric labels, then those numbers are converted into binary code, hence binary encoding.
4. **Hashing Encoding**: In this technique, the categorical variable is first converted into a string and then hashed. The hashed value is used as the representation of the category.
5. **Target Encoding**: In target encoding, the categories are replaced with the mean target value for that category. For example, if we are predicting whether a customer will default on a loan or not (1 - default, 0 - no default), then we can replace a categorical variable like occupation with the mean default rate for each occupation.
6. **Frequency or Count Encoding**: In frequency encoding, we replace the category with the count of the category in the data set.
7. **Embedding Encoding or Entity Embedding**: This is a method that uses a neural network to learn the representation for the categorical variables. This method can capture more complex patterns compared to other encoding methods.
8. **Leave One Out Encoding**: This is a similar technique to target encoding but it avoids the target leakage problem. In this method, we calculate the mean target of each category using all the rows excluding the current row.
9. **James-Stein Encoding**: This method shrinks the estimates of the mean towards the overall mean, which can be useful when dealing with categories with low counts.
10. **M-estimator Encoding**: This method is a robust version of target encoding against outliers.

Embedding Techniques

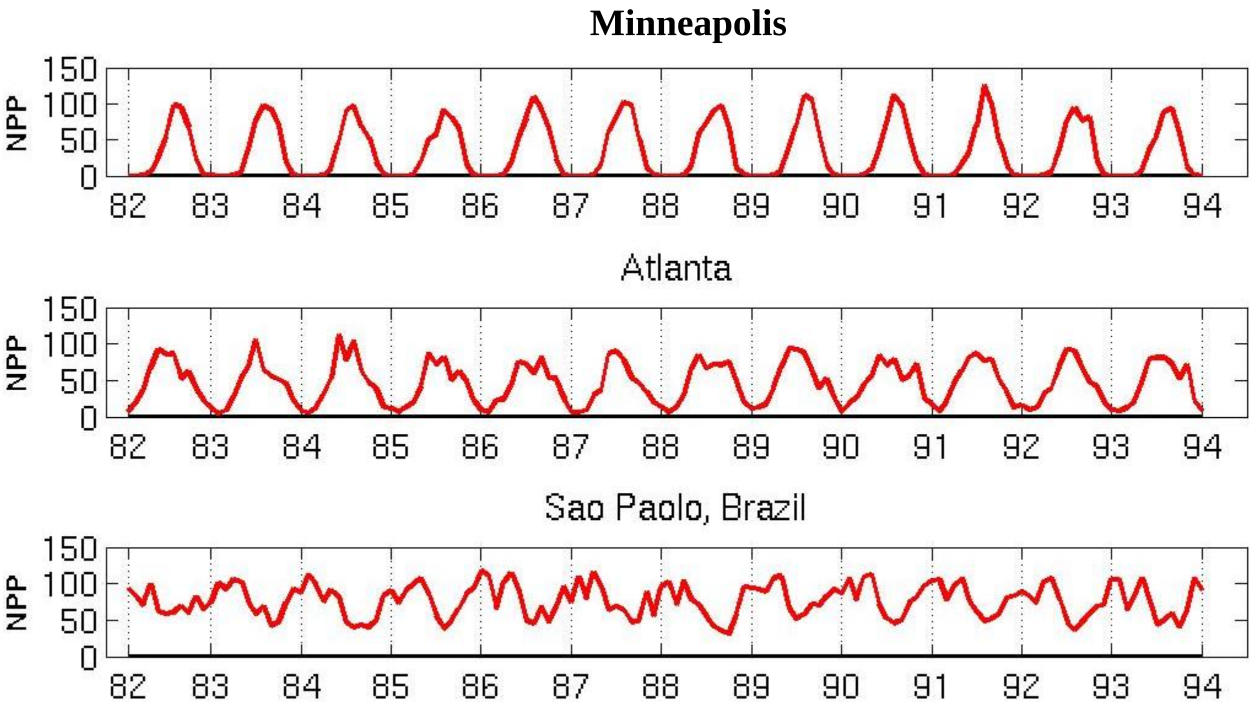
- **Entity Embeddings**: This technique is used to convert categorical variables into a form that keeps the semantic properties of the categories. Entity embeddings not only reduce memory usage but also speed up neural networks compared to one-hot encoding. They were first used in the third-place result in the Kaggle competition of Rossmann Store Sales forecasting.
- **Word Embeddings**: This technique is most commonly used in Natural Language Processing (NLP). Word2Vec, GloVe, and FastText are some of the popular word embedding methods. They transform a word into a dense vector that represents the semantic meaning of the word.
- **BERT Embeddings**: BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based machine learning technique for NLP pre-training. BERT creates word embeddings that are more nuanced and context-aware compared to older methods like Word2Vec or GloVe.
- **Graph Embeddings**: This technique is used for representing nodes, edges, and their features in a graph in a dense vector form. They are used in graph neural networks. Examples include Node2Vec, GraphSAGE, etc.
- **Knowledge Graph Embeddings**: These are a subset of Graph embeddings and are used for representing entities and relations in knowledge graphs. TransE, TransH, TransR, and TransD are some of the popular knowledge graph embedding methods.
- **Image Embeddings**: Techniques such as CNNs are used to convert images into dense vector representations.
- **Sequence Embeddings**: Techniques like Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTM), and Transformers are used to convert sequences (like sentences or time series) into dense vector representations.

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \dfrac{x - \text{mean}(x)}{\text{standard deviation }(x)}$	$x_{\text{norm}} = \dfrac{x - \min(x)}{\max(x) - \min(x)}$

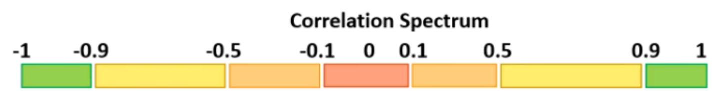
Example: Sample Time Series of Plant Growth



Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlation Strength	Positive	Negative
Perfect	r = 0.9 to 1	r = -0.9 to -1
Strong	r = 0.5 to 0.9	r = -0.5 to -0.9
Weak	r = 0.1 to 0.5	r = -0.1 to -0.5
Uncorrelated	r = 0 to 0.1	r = 0 to -0.1

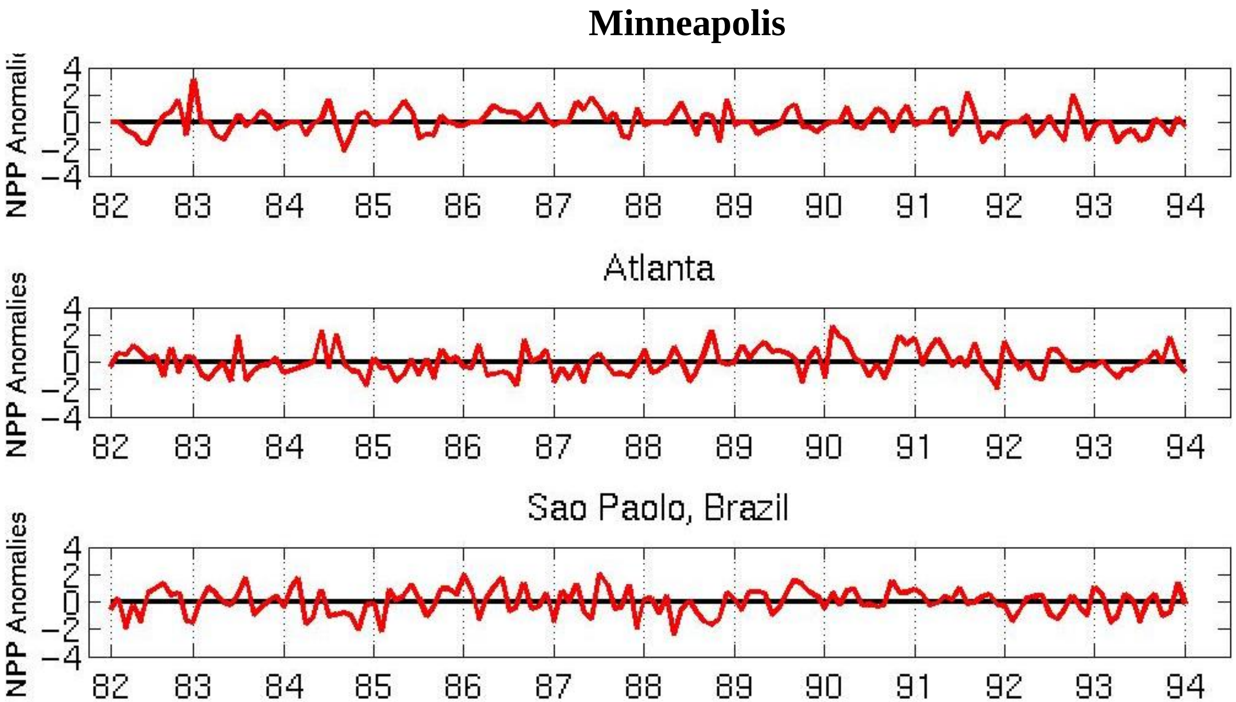
Note :- Any correlation above 0.3 and below -0.3 is considered significant.



Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	$\boxed{-0.7581}$
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Seasonality Accounts for Much Correlation



Normalized using monthly Z Score:

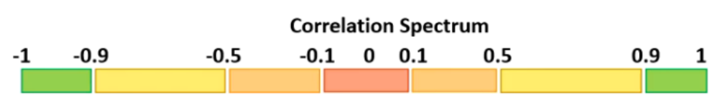
Subtract off monthly mean and divide by monthly standard deviation

Correlation Strength	Positive	Negative
Perfect	r = 0.9 to 1	r = -0.9 to -1
Strong	r = 0.5 to 0.9	r = -0.5 to -0.9
Weak	r = 0.1 to 0.5	r = -0.1 to -0.5
Uncorrelated	r = 0 to 0.1	r = 0 to -0.1

Note :- Any correlation above 0.3 and below -0.3 is considered significant.

Correlations between time series

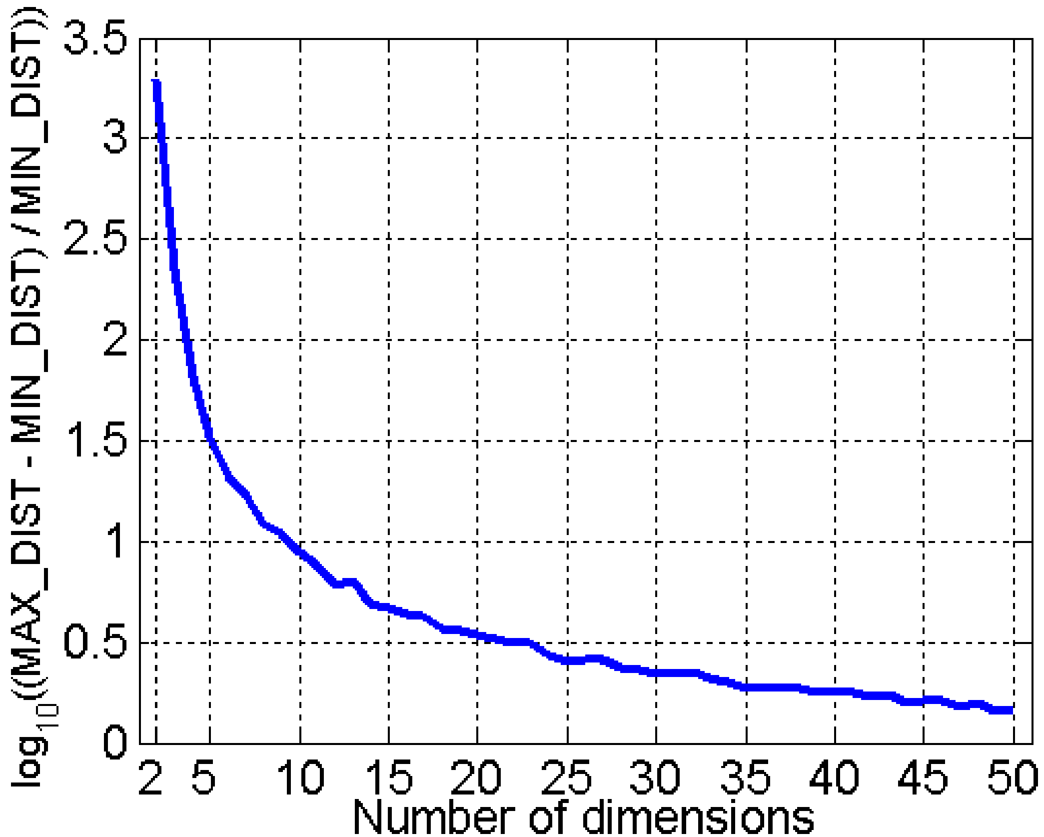
	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	$\boxed{0.0906}$
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000



Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

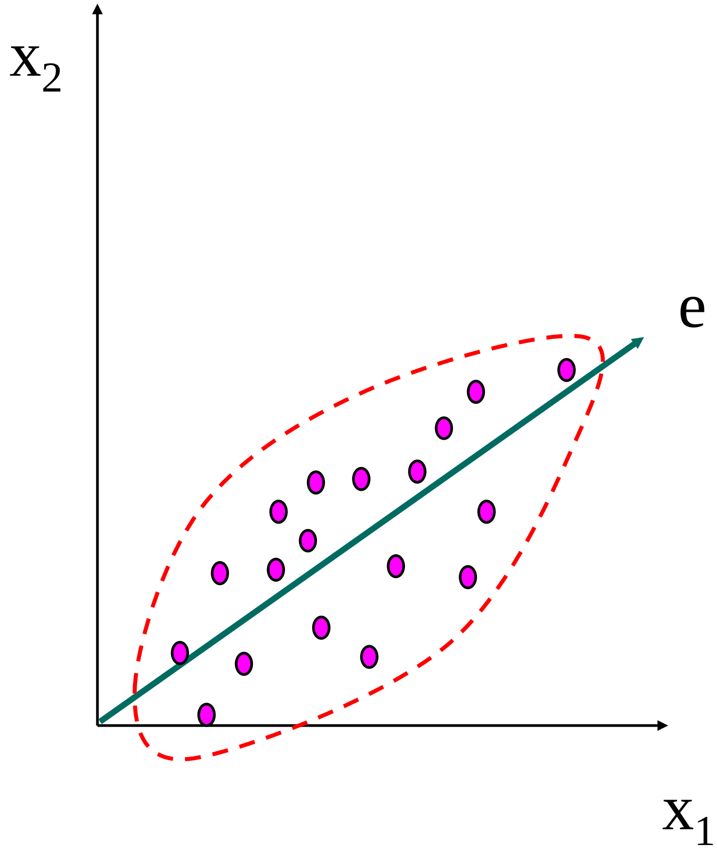
Dimensionality Reduction

- Purpose:
  – Avoid curse of dimensionality
  – Reduce amount of time and memory required by data mining algorithms
  – Allow data to be more easily visualized
  – May help to eliminate irrelevant features or reduce noise

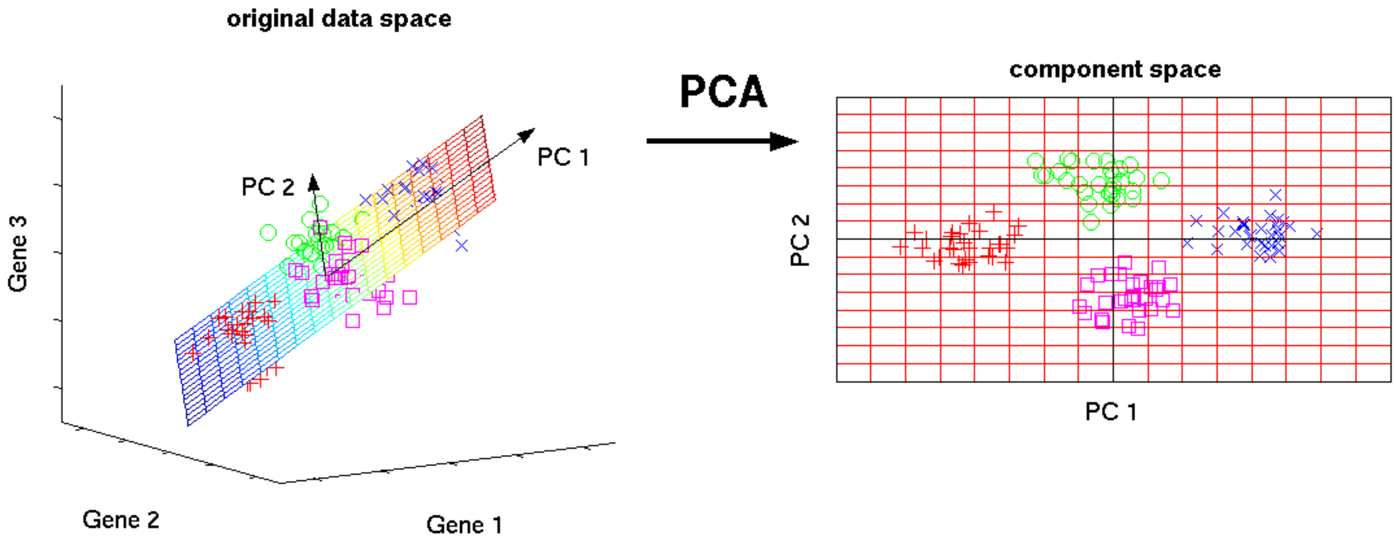
- Techniques
  – Principal Components Analysis (PCA)
  – Singular Value Decomposition (SVD)
  – Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA



$$Var(y_1) = Var(X \cdot b_1) = E[X \cdot b_1]^2 = E[(X \cdot b_1)^T (X \cdot b_1)] = \ldots$$

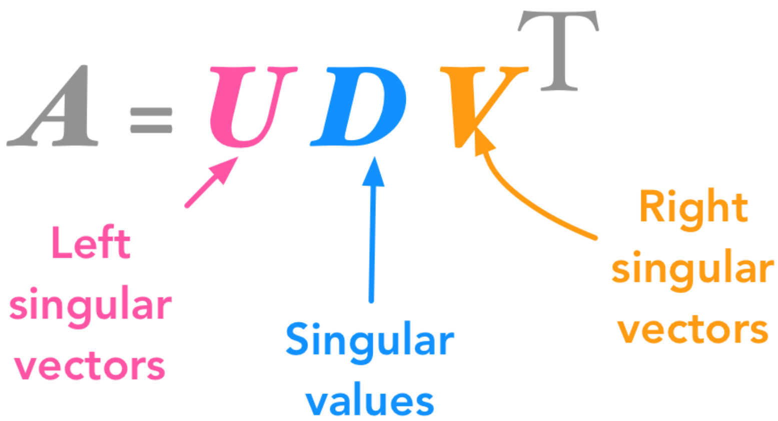
$$\ldots = \frac{1}{n}(Xb_1)^T(Xb_1) = \frac{1}{n} b_1^T X^T (Xb_1) = b_1^T \frac{X^T X}{n} b_1 = b_1^T C_X b_1$$

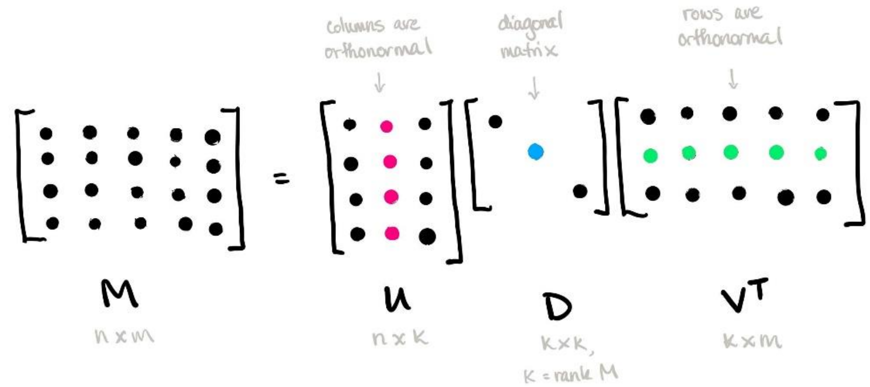
[Principal Component Analysis. Step by step intuition, mathematical… | by Andrea Grianti | Towards Data Science]

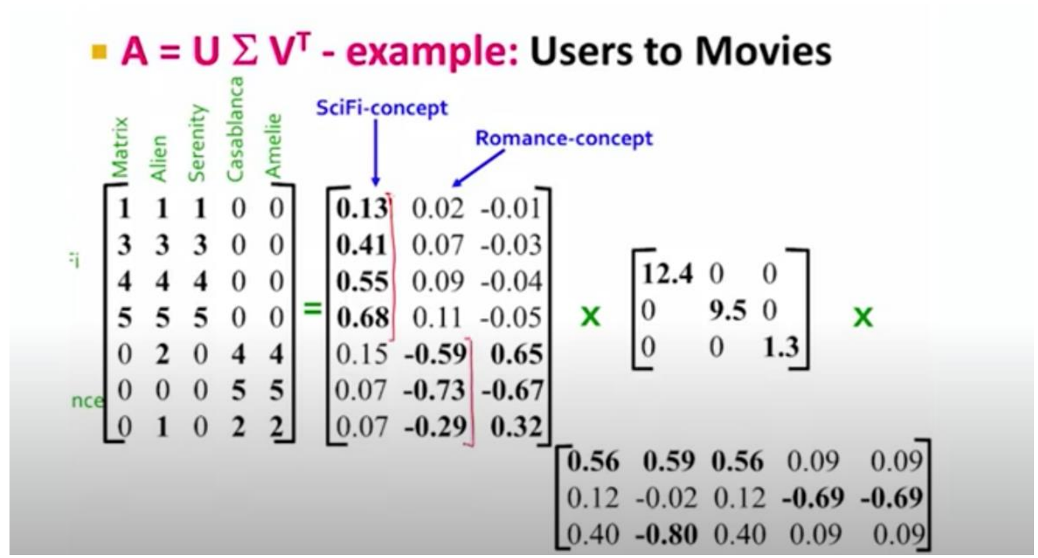
Dimensionality Reduction: PCA

256

Singular Value Decomposition







Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

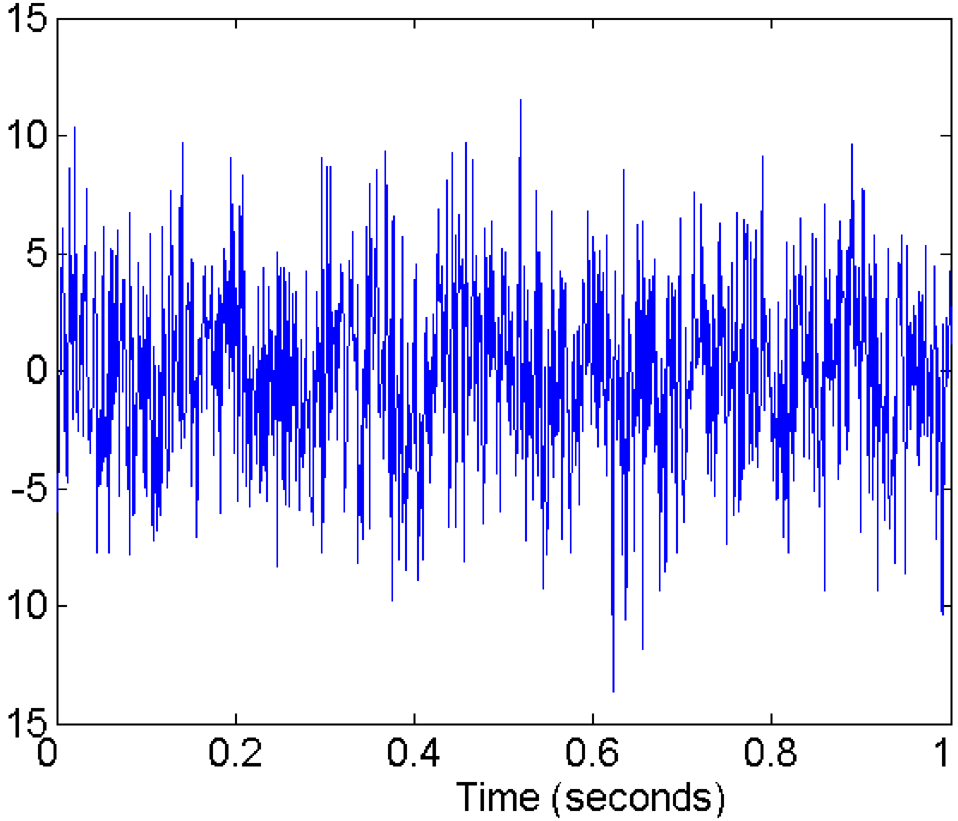
Feature Creation

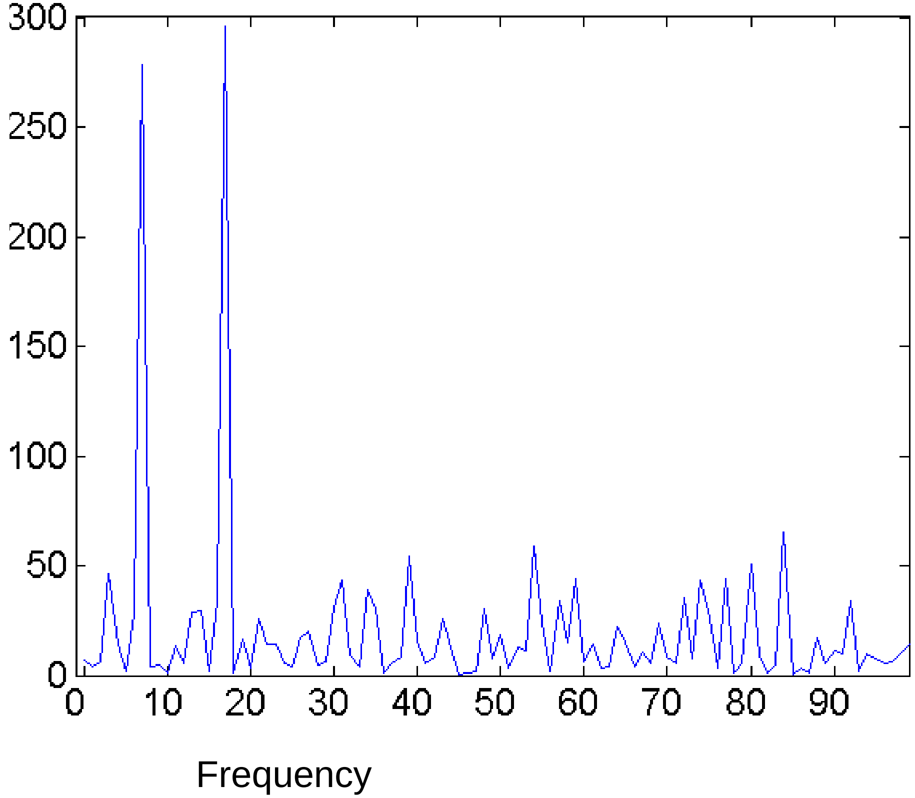
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier and wavelet transform





Two Sine Waves + Noise

Frequency