Knowledge Discovery & Data Mining

– Data Preprocessing –

Data compression & Sampling

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Based on the original version by Professor Jiliang Tang

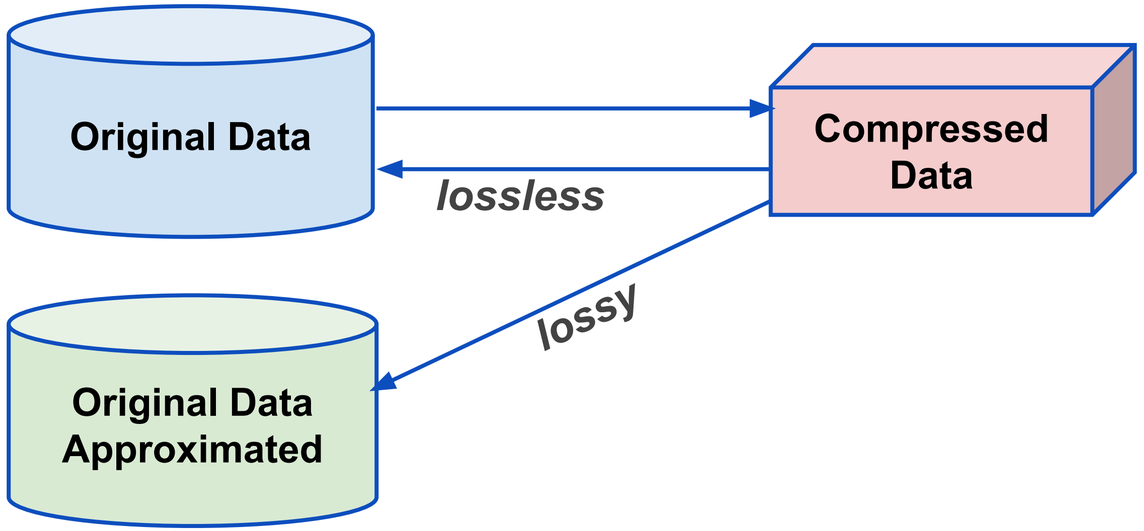
Data Reduction

- Data compression
  - Discrete wavelet transform (DWT)
- Sampling
  - Sampling without replacement
  - Sampling with replacement
  - Cluster or Stratified Sampling

Data Compression

In **data compression**, transformations are applied so as to obtain a reduced or "compressed" representation of the original data.

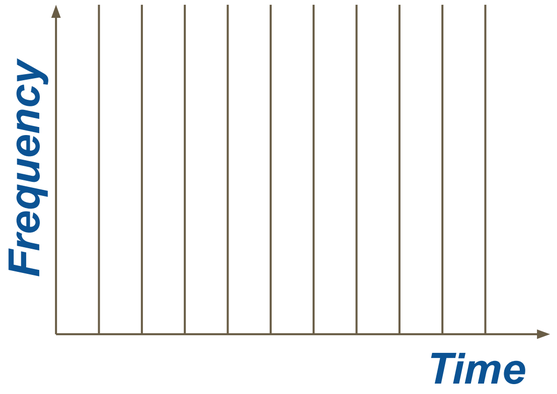
- If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called **lossless**.
- If,we can reconstruct only an approximation of the original data, then the data reduction is called **lossy**



Discrete wavelet transform (DWT)

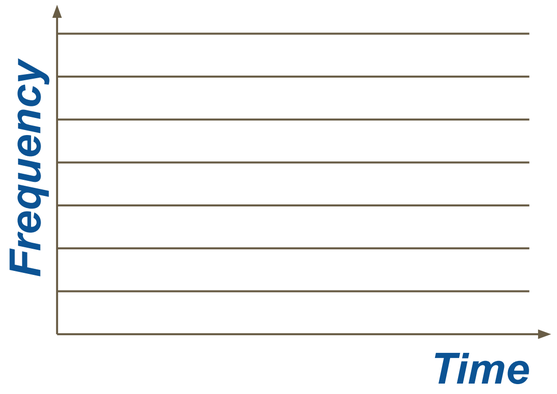
The **discrete wavelet transform (DWT)** is a linear signal processing technique that, when applied to a data vector x, transforms it to a numerically different vector, x’ , of wavelet coefficients.

Time series



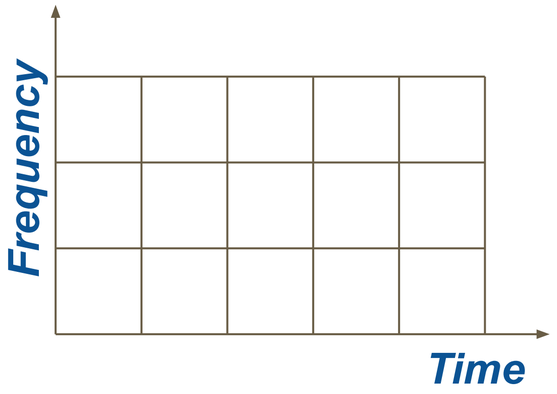
focuses on the raw temporal evolution of data

Fourier transform



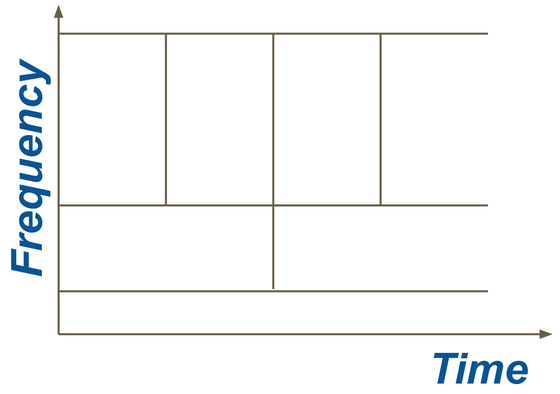
breaks down a signal into its constituent frequencies, but lose the exact timing of these frequencies.

Spectrogram



provides a time-frequency representation, showing when and with what intensity each frequency is present.

Wavelet transform

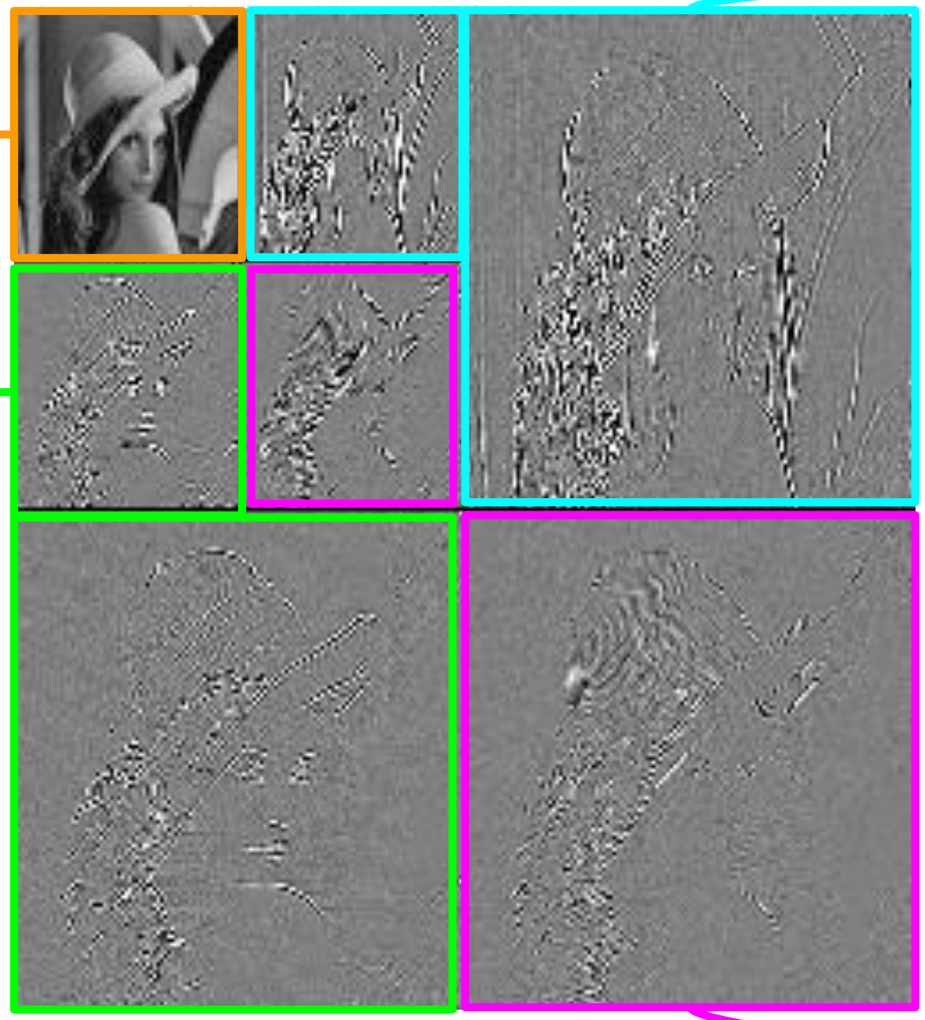


offers a multi-resolution perspective, allowing for variable time and frequency resolution, ideal for signals that have non-stationary or evolving characteristics.

Discrete wavelet transform (DWT)

Used for image compression

A low-resolution of the original image.



Horizontal details of the image. It captures the difference between rows in the image.

Vertical details of the image. It captures the difference between columns in the image.

Diagonal details of the image. It captures the diagonal changes in the image.

Discrete wavelet transform (DWT)

- **Data Preparation**: Ensure the data length, L, is a power of 2. If necessary, pad the data with zeros to satisfy this condition.
- **Transform Functions:** Utilize two main functions during the transformation, a smoothing(e.g. averaging) operation and a differencing operation. These functions are applied to pairs of consecutive data points.
- **Transform Application:** For each pair of consecutive data points, produce one average value and one difference value. This results in a dataset where the first half contains the average values, and the second half contains the difference values, but the overall length remains unchanged.
- **Recursive Application for Multi-resolution Analysis:** For multi-level transformations, recursively apply the transform to the average values from the previous level. This process can be repeated multiple times to achieve different levels of detail.

Discrete wavelet transform (DWT)

Example. Suppose we have a time series S = [2, 2, 0, 2, 3, 5, 4, 4], using DWT, then it can be transformed to S'=

- **Data Preparation**: Ensure the data length, L, is a power of 2. If necessary, pad the data with zeros to satisfy this condition.
- **Transform Functions:** Utilize two main functions during the transformation, a smoothing(e.g. averaging) operation and a differencing operation. These functions are applied to pairs of consecutive data points.
- **Transform Application:** For each pair of consecutive data points, produce one average value and one difference value. This results in a dataset where the first half contains the average values, and the second half contains the difference values, but the overall length remains unchanged.
- **Recursive Application for Multi-resolution Analysis:** For multi-level transformations, recursively apply the transform to the average values from the previous level. This process can be repeated multiple times to achieve different levels of detail.

Discrete wavelet transform (DWT)

Example. Suppose we have a time series S = [2, 2, 0, 2, 3, 5, 4, 4], using DWT, then it can be transformed to S'=

Resolution	Averages	Detail Coefficients(difference)
8	[2, 2, 0, 2, 3, 5, 4, 4]	
4	[2 , 1 , 4 , 4]	[0 , -1 , -1 , 0]

Suppose we aim to represent S using only L/2 numbers; which four values could best approximate S?

to reconstruct the original list of values

Discrete wavelet transform (DWT)

Example. Suppose we have a time series S = [2, 2, 0, 2, 3, 5, 4, 4], using DWT, then it can be transformed to S'=[**2.75**, **-1.25**, **0.5**, **0**, **0**, **-1**, **-1**, **0**]

Resolution	Averages	Detail Coefficients(difference)
8	[2, 2, 0, 2, 3, 5, 4, 4]	
4	[**2** , **1** , **4** , **4**]	[**0** , **-1** , **-1** , **0**]
2	[**1.5** , **4**]	[**0.5** , **0**]
1	[**2.75**]	[**-1.25**]

Discrete wavelet transform (DWT)

Example. Suppose we have a time series S = [2, 2, 0, 2, 3, 5, 4, 4], using DWT, then it can be transformed to S'=[**2.75**, **-1.25**, **0.5**, **0, 0**, **-1**, **-1**, **0**]

Resolution	Averages	Detail Coefficients(difference)
8	[2, 2, 0, 2, 3, 5, 4, 4]	
4	[**2** , **1** , **4** , **4**]	[**0** , **-1** , **-1** , **0**]
2	[**1.5** , **4**]	[**0.5** , **0**]
1	[**2.75**]	[**-1.25**]

lossless or lossy?

Discrete wavelet transform (DWT)

Example. Suppose we have a time series S = [2, 2, 0, 2, 3, 5, 4, 4], using DWT, then it can be transformed to S'=[**2.75**, **-1.25**, **0.5**, **0, 0**, **-1**, **-1**, **0**]

Resolution	Averages	Detail Coefficients(difference)
8	[2, 2, 0, 2, 3, 5, 4, 4]	
4	[**2** , **1** , **4** , **4**]	[**0** , **-1** , **-1** , **0**]
2	[**1.5** , **4**]	[**0.5** , **0**]
1	[**2.75**]	[**-1.25**]

Where is the Data Compression?

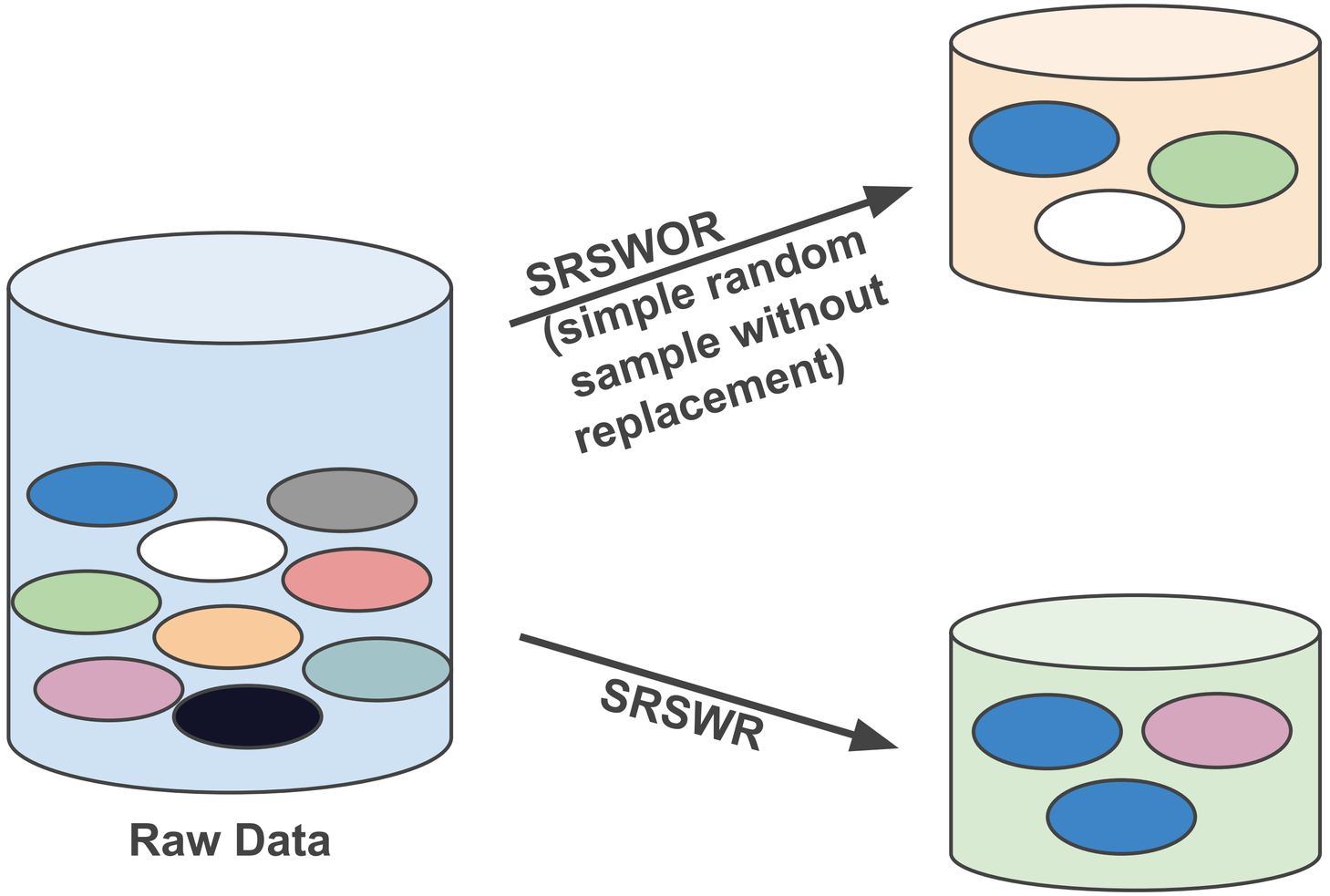
Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling

Types of Sampling

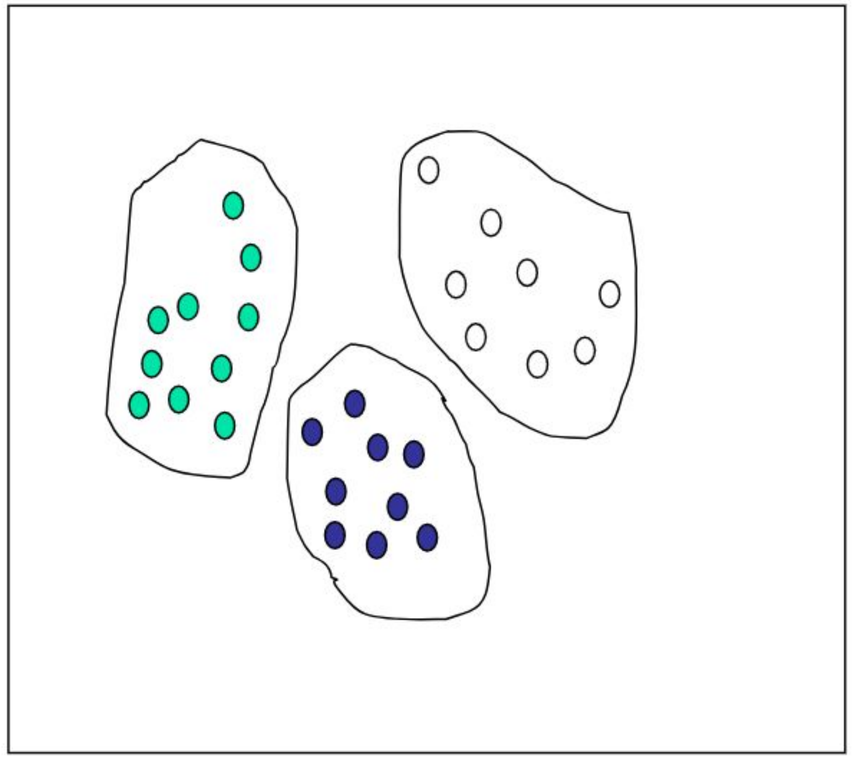
- Simple random sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement (SRSWOR)
  - Once an object is selected, it is removed from the population
- Sampling with replacement (SRSWR)
  - A selected object is not removed from the population
- Stratified sampling:
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Sampling: With or Without Replacement

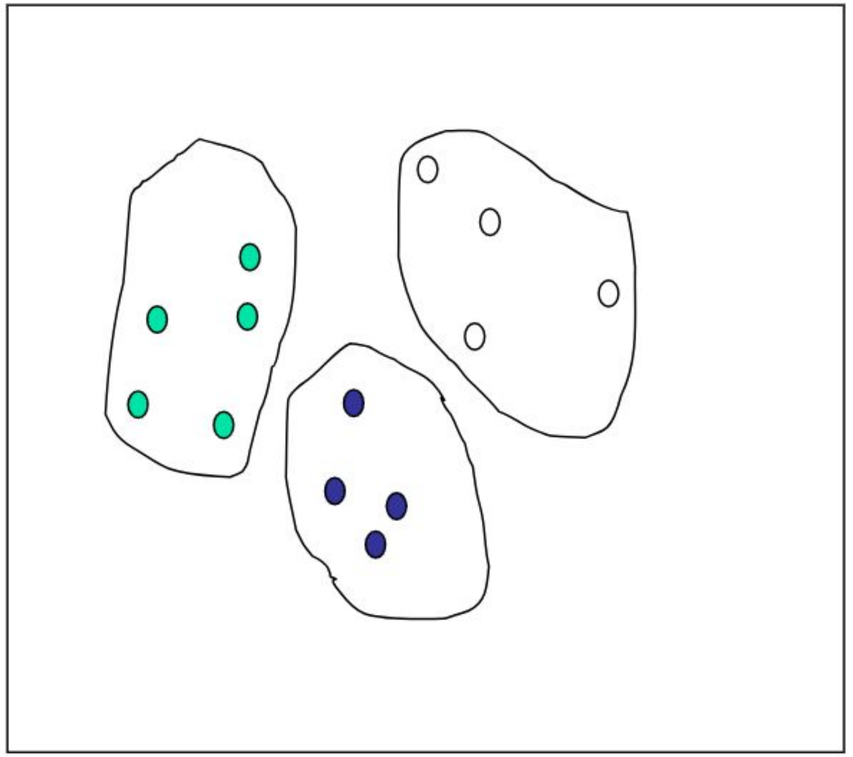


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Summary

- Data compression
  - Discrete wavelet transform (DWT)
- Sampling
  - Sampling without replacement
  - Sampling with replacement
  - Cluster or Stratified Sampling