# CS6220: DATA MINING TECHNIQUES

## Mining Time Series Data

**Instructor: Yizhou Sun**
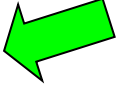
yzsun@ccs.neu.edu

November 18, 2014

# Methods to Learn

| | Matrix Data | Set Data | Sequence Data | Time Series | Graph & Network |
|---|---|---|---|---|---|
| **Classification** | Decision Tree; Naïve Bayes; Logistic Regression SVM; kNN | | HMM | | Label Propagation |
| **Clustering** | K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means | | | | SCAN; Spectral Clustering |
| **Frequent Pattern Mining** | | Apriori; FP-growth | GSP; PrefixSpan | | |
| **Prediction** | Linear Regression | | | **Autoregression** | |
| **Similarity Search** | | | | **DTW** | P-PageRank |
| **Ranking** | | | | | PageRank |

# Mining Time Series Data

- Basic Concepts

- Time Series Prediction and Forecasting

- Time Series Similarity Search

- Summary

# Example: Inflation Rate Time Series



FIGURE 12.1 Inflation and Unemployment in the United States, 1960–1999
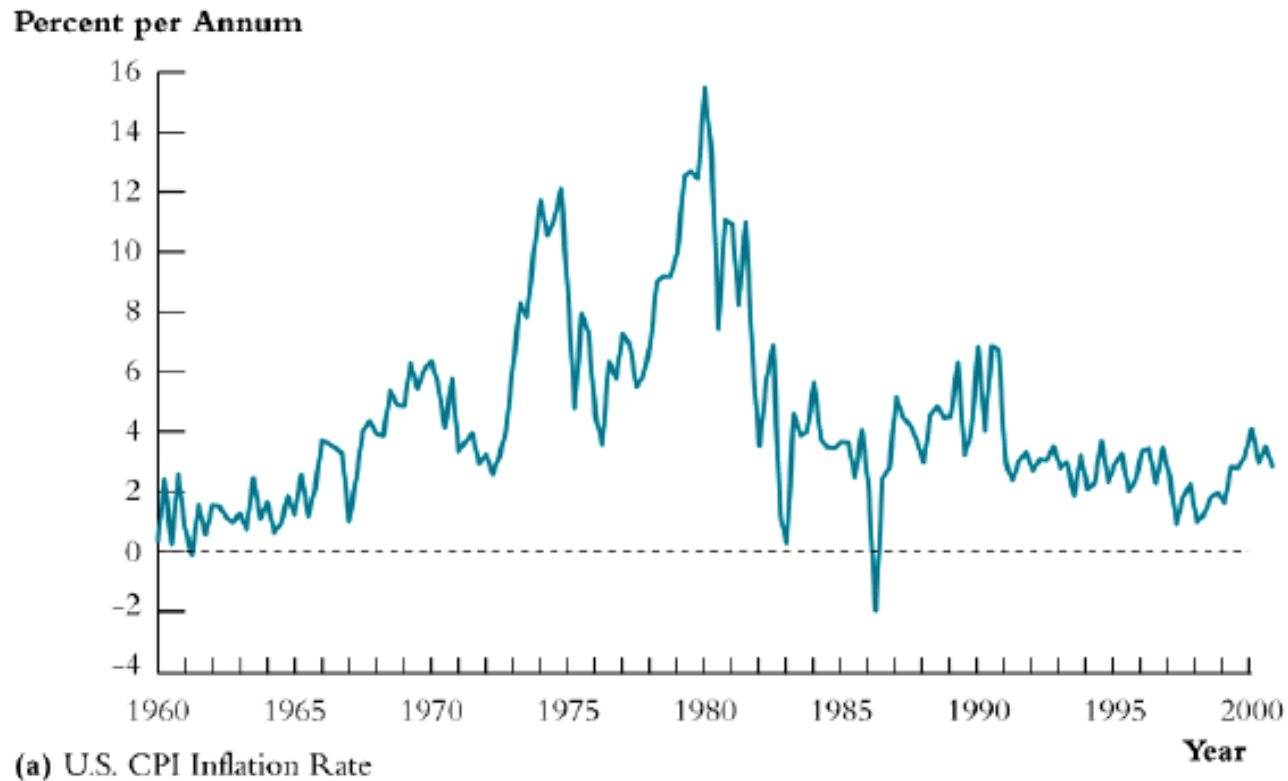
(a) U.S. CPI Inflation Rate

Price inflation in the United States (Figure 12.1a) drifted upwards from 1960 until 1980, and then fell sharply during the early 1980s. The unemployment rate in the United States (Figure 12.1b) rises during recessions (the shaded episodes) and falls during expansions.

# Example: Unemployment Rate Time Series

**FIGURE 12.1** Inflation and Unemployment in the United States, 1960–1999
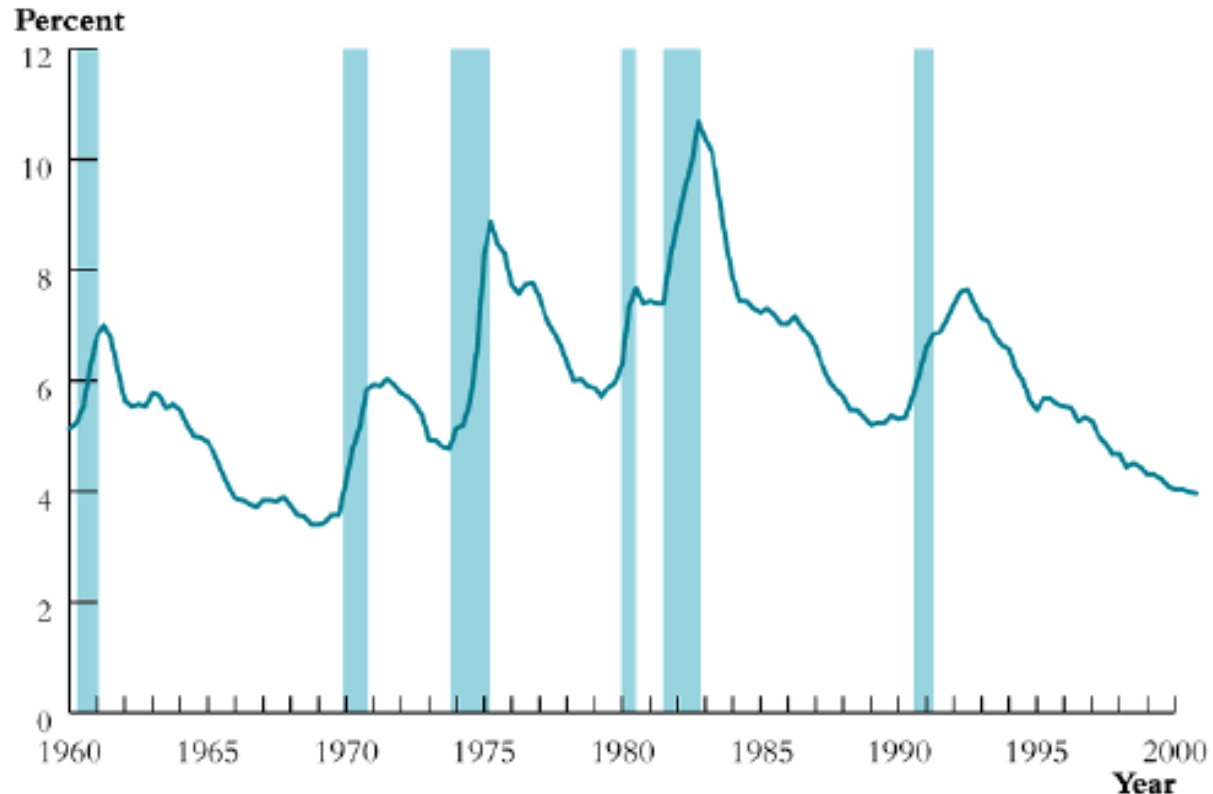


**(b)** U.S. Unemployment Rate

Price inflation in the United States (Figure 12.1a) drifted upwards from 1960 until 1980, and then fell sharply during the early 1980s. The unemployment rate in the United States (Figure 12.1b) rises during recessions (the shaded episodes) and falls during expansions.

# Example: Stock

# Example: Product Sale
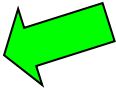
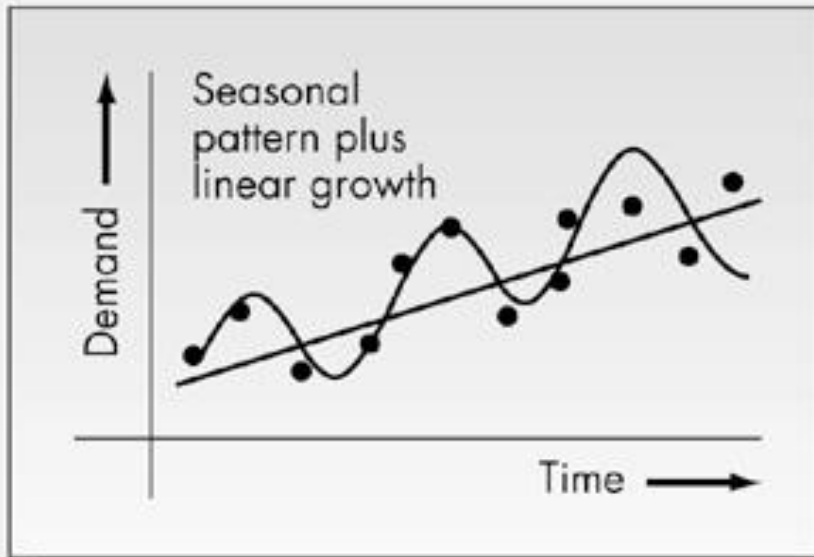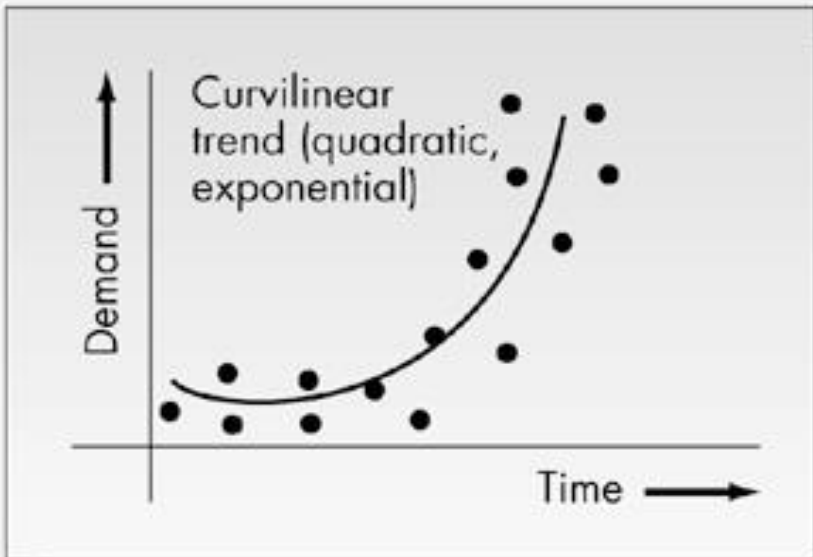| Time | Observations | | | | | | | | | |
|------|----|----|----|----|----|----|----|----|----|----|
| 1 - 10 | 4 | 16 | 12 | 25 | 13 | 12 | 4 | 8 | 9 | 14 |
| 11 - 20 | 3 | 14 | 14 | 20 | 7 | 9 | 6 | 11 | 3 | 11 |
| 20 - 25 | 8 | 7 | 2 | 8 | 8 | 10 | 7 | 16 | 9 | 4 |
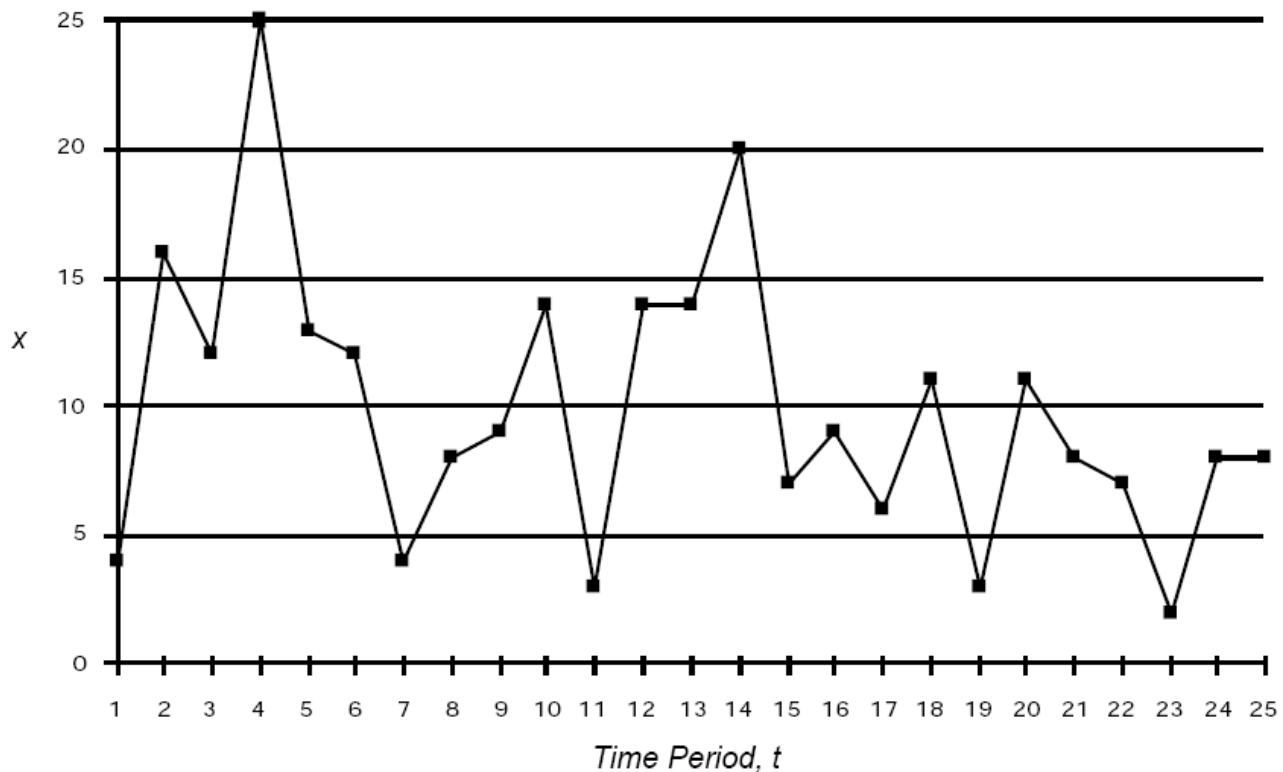
# Time Series

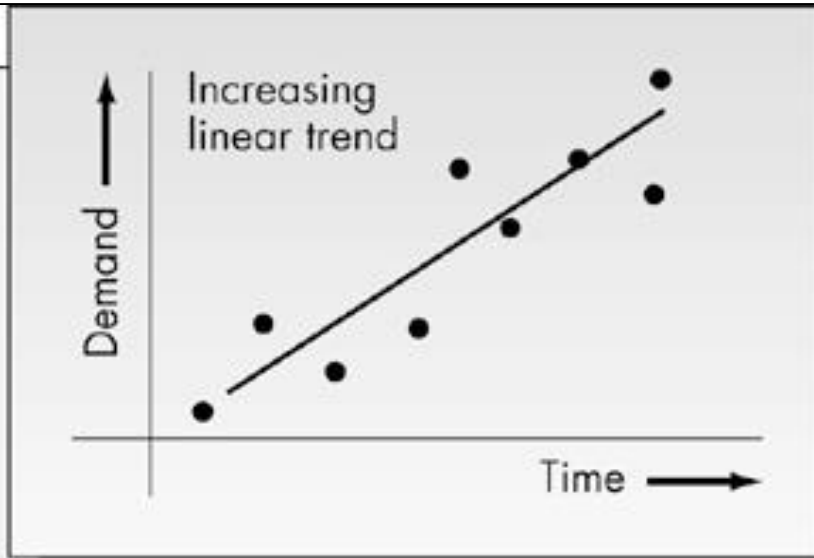- A **time series** is a sequence of numerical data points, measured typically at successive times, spaced at (often uniform) time intervals
  - *Random variables for a time series are Represented as:*
    - $Y = \{Y_1, Y_2, \dots\}, or$
    - $Y = \{Y_t : t \in T\}, where\ T\ is\ the\ index\ set$
  - An observation of a time series with length $N$ is represent as:
    - $Y = \{y_1, y_2, \dots, y_N\}$

# Mining Time Series Data

- Basic Concepts

- Time Series Prediction and Forecasting ⬅

- Time Series Similarity Search

- Summary

# Categories of Time-Series Movements

- Categories of Time-Series Movements (T, C, S, I)

  - Long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time

  - Cyclic movements or cycle variations: long term oscillations about a trend line or curve

    - e.g., business cycles, may or may not be periodic

  - Seasonal movements or seasonal variations

    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.

  - Irregular or random movements

Purely random—No recognizable pattern

Increasing linear trend

Curvilinear trend (quadratic, exponential)

Seasonal pattern plus linear growth
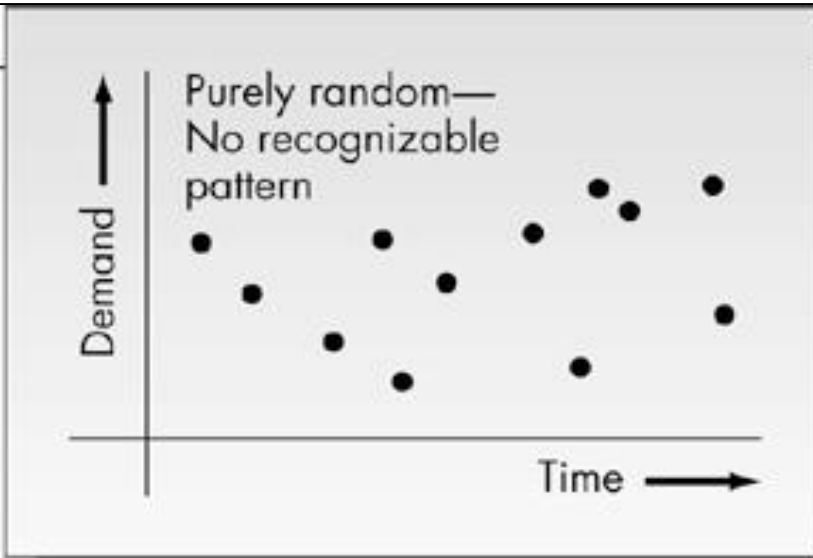
# Lag, Difference

- The first lag of $Y_t$ is $Y_{t-1}$; the jth lag of $Y_t$ is $Y_{t-j}$

- The first difference of a time series, $\Delta Y_t = Y_t - Y_{t-1}$

  - Sometimes difference in logarithm is used $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$

# Example: First Lag and First Difference

| Quarter | U.S. CPI | Rate of Inflation at an Annual Rate ($Inf_t$) | First Lag ($Inf_{t-1}$) | Change in Inflation ($\Delta Inf_t$) |
|---------|----------|-----------------------------------------------|-------------------------|--------------------------------------|
| 1999:I | 164.87 | 1.6 | 2.0 | −0.4 |
| 1999:II | 166.03 | 2.8 | 1.6 | 1.2 |
| 1999:III | 167.20 | 2.8 | 2.8 | 0.0 |
| 1999:IV | 168.53 | 3.2 | 2.8 | 0.4 |
| 2000:I | 170.27 | 4.1 | 3.2 | 0.9 |

**TABLE 12.1** Inflation in the United States in 1999 and the First Quarter of 2000

13

# Autocorrelation

- Autocorrelation: the correlation between a time series and its lagged values

  - The first autocorrelation $\rho_1$

  $$\text{corr}(Y_t, Y_{t-1}) = \frac{\text{cov}(Y_t, Y_{t-1})}{\sqrt{\text{var}(Y_t)\,\text{var}(Y_{t-1})}}$$

  - The jth autocorrelation $\rho_j$

  $$\text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\,\text{var}(Y_{t-j})}}.$$

**Autocovariance**

# Sample Autocorrelation Calculation

- The jth sample autocorrelation

  - $\hat{\rho}_j = \dfrac{\widehat{cov}(Y_t, Y_{t-j})}{\widehat{var}(Y_t)}$

  - Where $\widehat{cov}(Y_t, Y_{t-j})$ is calculated as:

  $$\frac{1}{T-j-1}\sum_{t=j+1}^{T}(Y_t - \bar{Y}_{j+1,T})(Y_{t-j} - \bar{Y}_{1,T-j})$$

  - i.e., considering two time series: Y(1,...,T-j) and Y(j+1,...,T)

# Example of Autocorrelation

- For inflation and its change

**TABLE 12.2** First Four Sample Autocorrelations of the U.S. Inflation Rate and Its Change, 1960:I–1999:IV

| | Autocorrelation of: | |
| Lag | Inflation Rate ($Inf_t$) | Change of Inflation Rate ($\Delta Inf_t$) |
| --- | --- | --- |
| 1 | 0.85 | −0.24 |
| 2 | 0.77 | −0.27 |
| 3 | 0.77 | 0.32 |
| 4 | 0.68 | −0.06 |

$\rho_1 = 0.85$, **very high: Last quarter's inflation rate contains much information about this quarter's inflation rate**

# Focus on Stationary Time Series

- Stationary is key for time series regression: Future is similar to the past in terms of distribution

A time series $Y_t$ is **stationary** if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on $s$; otherwise, $Y_t$ is said to be **nonstationary**. A pair of time series, $X_t$ and $Y_t$, are said to be **jointly stationary** if the joint distribution of $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \ldots, X_{s+T}, Y_{s+T})$ does not depend on $s$. Stationarity requires the future to be like the past, at least in a probabilistic sense.

# Autoregression

- Use past values $Y_{t-1}, Y_{t-2}, \ldots$ to predict $Y_t$

  - An *autoregression* is a regression model in which $Y_t$ is regressed against its own lagged values.

  - The number of lags used as regressors is called the *order* of the autoregression.

    - In a **first order autoregression**, $Y_t$ is regressed against $Y_{t-1}$

    - In a **pth order autoregression**, $Y_t$ is regressed against $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$

# The First Order Autoregression Model AR(1)

- AR(1) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

- The AR(1) model can be estimated by OLS regression of $Y_t$ against $Y_{t-1}$

- Testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ provides a test of the hypothesis that $Y_{t-1}$ is not useful for forecasting $Y_t$

# Prediction vs. Forecast
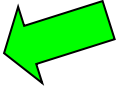
- A *predicted value* refers to the value of *Y* predicted (using a regression) for an observation in the sample used to estimate the regression – this is the usual definition

  - Predicted values are "in sample"

- A *forecast* refers to the value of *Y* forecasted for an observation *not* in the sample used to estimate the regression.

  - Forecasts are forecasts of the future – which cannot have been used to estimate the regression.

# Time Series Regression with Additional Predictors

- So far we have considered forecasting models that use only past values of *Y*

- It makes sense to add other variables (*X*) that might be useful predictors of *Y*, above and beyond the predictive value of lagged values of *Y*:

- $$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p}$$
$$+ \delta_1 X_{t-1} + \ldots + \delta_r X_{t-r} + u_t$$

# Mining Time Series Data

- Basic Concepts

- Time Series Prediction and Forecasting

- Time Series Similarity Search

- Summary

# Why Similarity Search?

- Wide applications
  - Find a time period with similar inflation rate and unemployment time series?
  - Find a similar stock to Facebook?
  - Find a similar product to a query one according to sale time series?
  - ...

# Example

VanEck International Fund
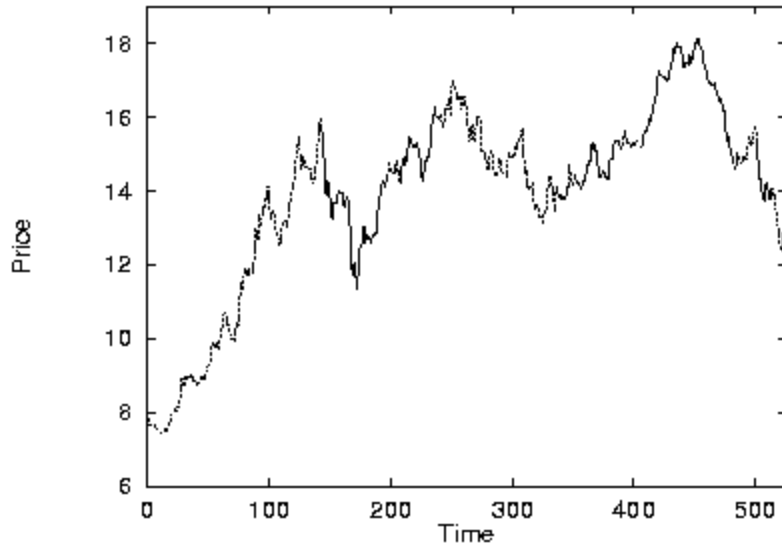
Fidelity Selective Precious Metal and Mineral Fund
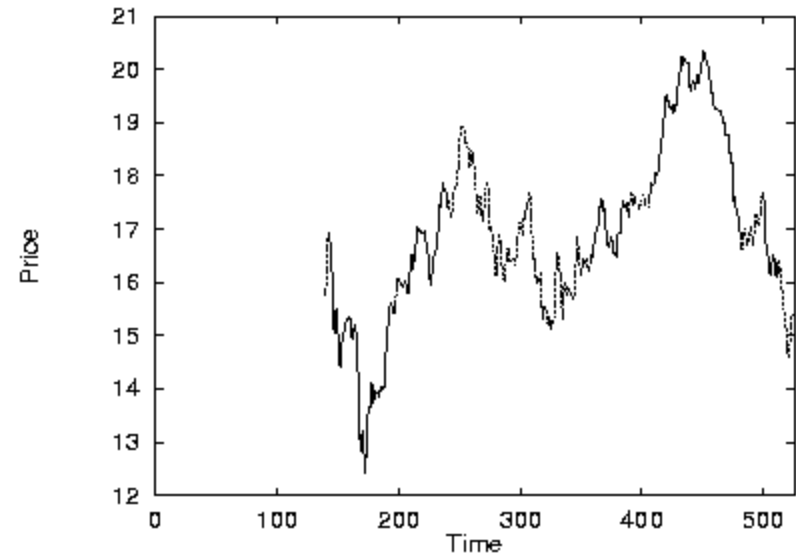


Two similar mutual funds in the different fund group

# Similarity Search for Time Series Data

- Time Series Similarity Search
  - Euclidean distances and $L_p$ norms
  - Dynamic Time Warping (DTW)
  - Time Domain vs. Frequency Domain

# Euclidean Distance and Lp Norms

- Given two time series with equal length n
  - $C = \{c_1, c_2, \ldots, c_n\}$
  - $Q = \{q_1, q_2, \ldots, q_n\}$
  - $d(C, Q) = (\sum |c_i - q_i|^p)^{1/p}$
    - When p=2, it is Euclidean distance
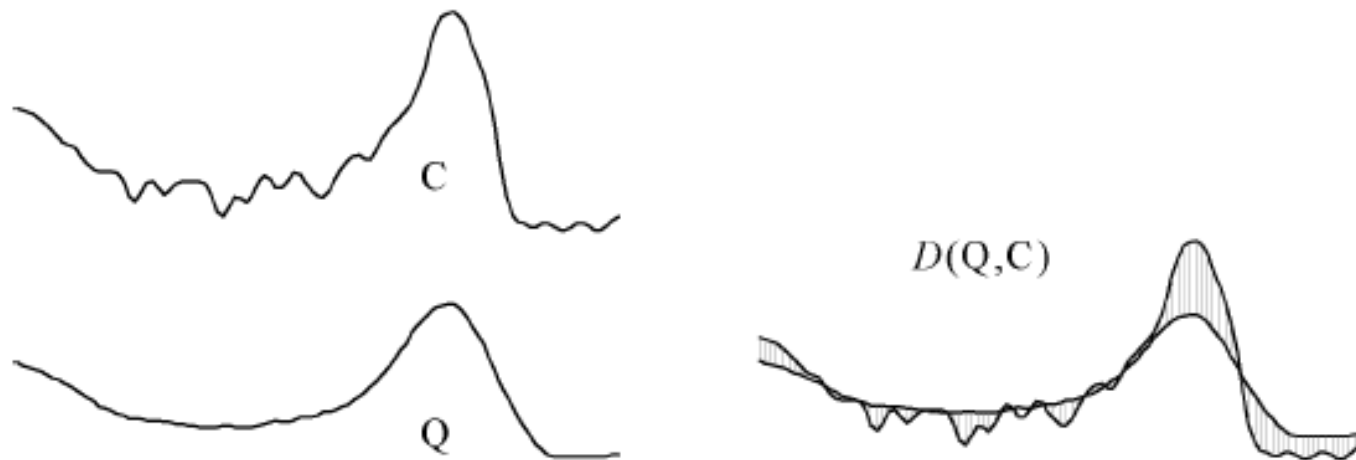
# Enhanced Lp Norm-based Distance

- Issues with Lp Norm: cannot deal with offset and scaling in the Y-axis

- Solution: normalizing the time series

  - $c_i' = \dfrac{c_i - \mu(C)}{\sigma(C)}$

Unnormalized       Normalized
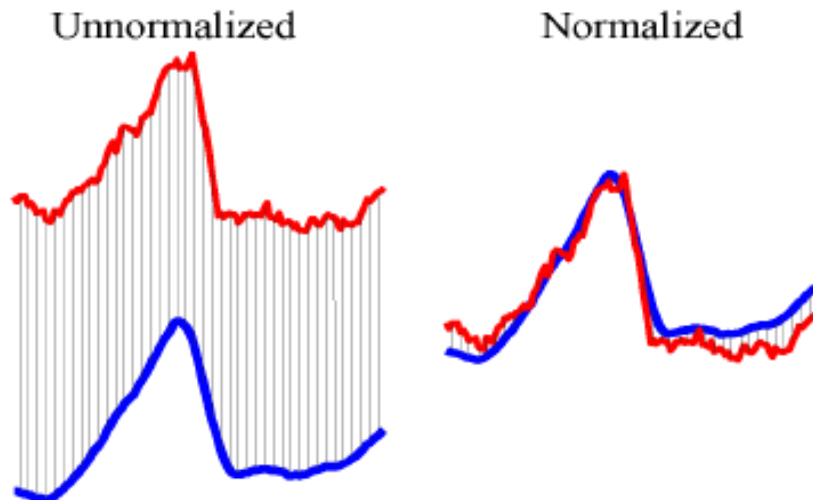
# Dynamic Time Warping (DTW)

- For two sequences that do not line up well in X-axis, but share roughly similar shape

  - We need to warp the time axis to make better alignment
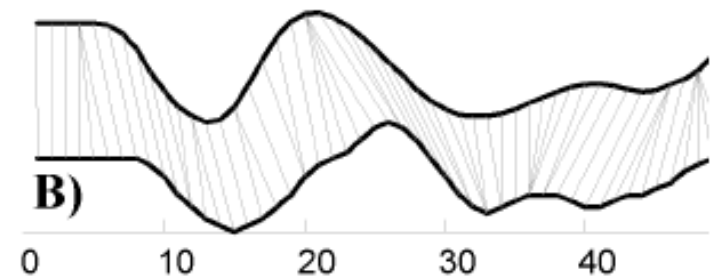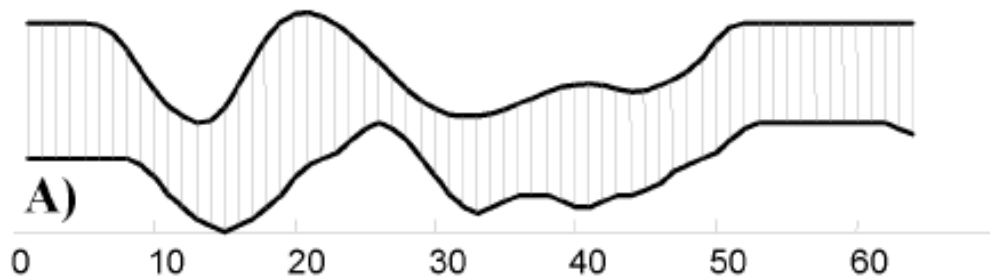
# Goal of DTW

- Given
  - Two sequences (with possible different lengths):
    - $X = \{x_1, x_2, \dots, x_N\}$
    - $Y = \{y_1, y_2, \dots, y_M\}$
  - A local distance (cost) measure between $x_n$ and $y_m$
- Goal:
  - Find an alignment between X and Y, such that, the overall cost is minimized

# Cost Matrix of Two Time Series
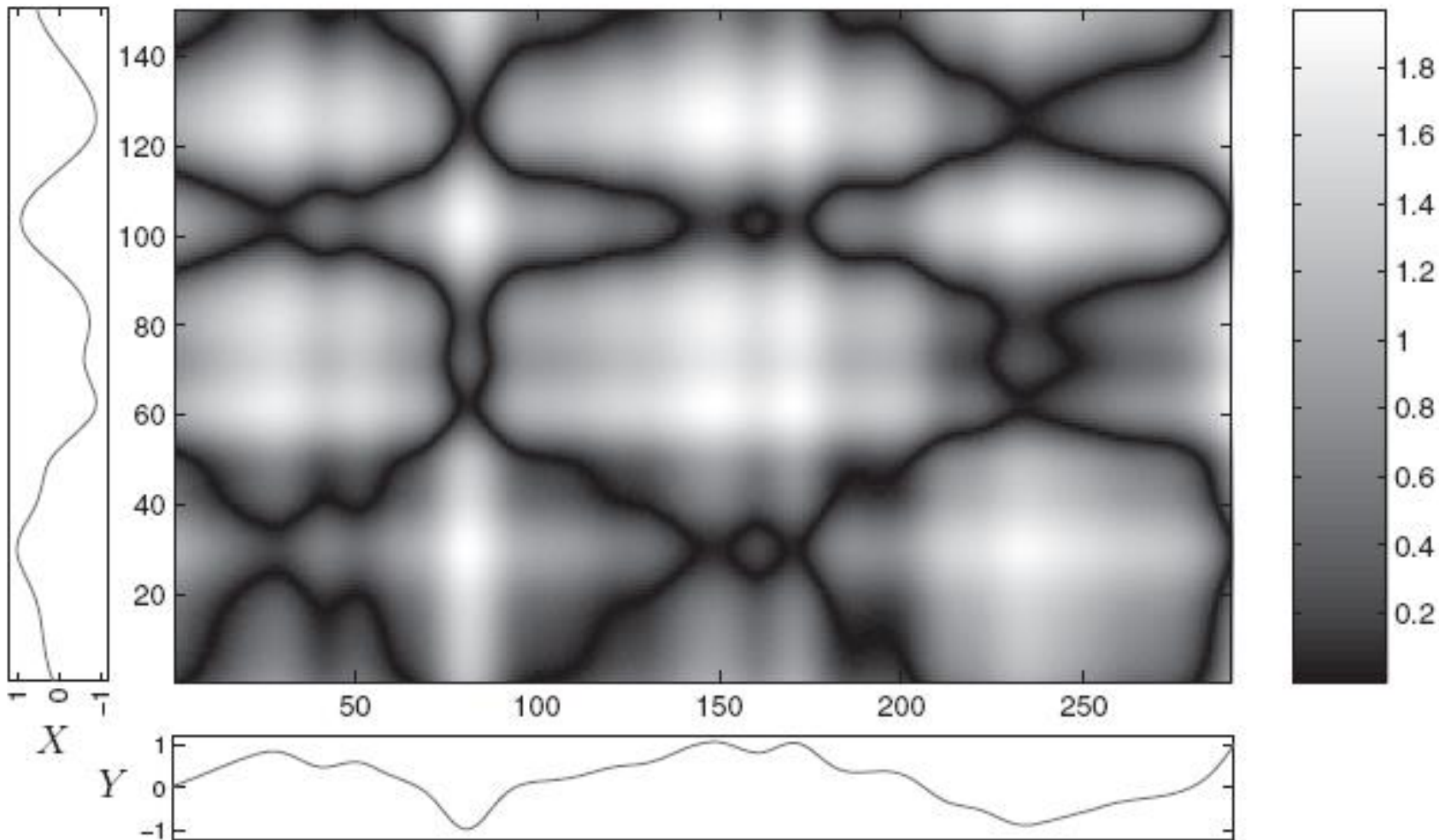
# Represent an Alignment by Warping Path

- An (N,M)-warping path is a sequence $p = (p_1, p_2, \ldots, p_L)$ with $p_l = (n_l, m_l)$, satisfying the three conditions:

  - Boundary condition: $p_1 = (1,1), p_L = (N, M)$
    - Starting from the first point and ending at last point

  - Monotonicity condition: $n_l$ and $m_l$ are non-decreasing with $l$

  - Step size condition: $p_{l+1} - p_l \in \{(0,1), (1,0), (1,1)\}$
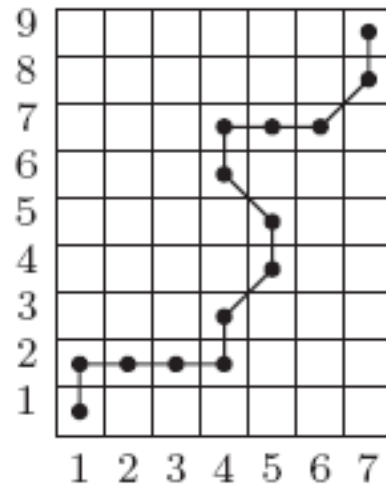    - Move one step right, up, or up-right

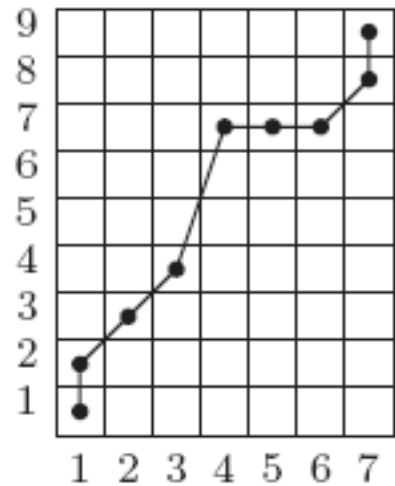# Q: Which Path is a Warping Path?
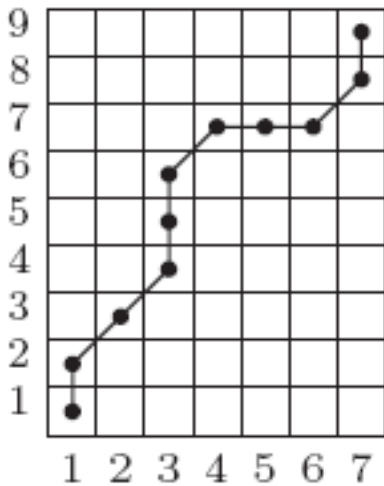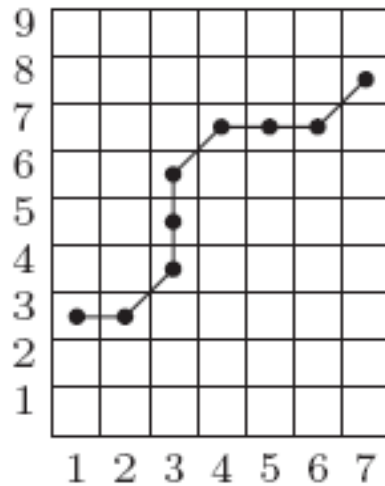
# Optimal Warping Path

- The total cost given a warping path p

  - $c_p(X, Y) = \sum_l c(x_{n_l}, y_{m_l})$

- The optimal warping path p*

  - $c_{p^*}(X, Y) =$
    $\min\{c_p(X, Y) | p \text{ is an } (N, M) - warping\ path\}$

- DTW distance between X and Y is defined as:

  - the optimal cost $c_{p^*}(X, Y)$

# How to Find p*?

- Naïve solution:

  - Enumerate all the possible warping path

  - Exponential in N and M!

# Dynamic Programming for DTW

- Dynamic programming:

  - Let $D$(n,m) denote the DTW distance between $X(1,\ldots,n)$ and $Y(1,\ldots,m)$

  - $D$ is called accumulative cost matrix

    - Note D(N,M) = DTW(X,Y)

  - Recursively calculate $D$(n,m)

    - $D(n,m) = \min\{D(n-1,m), D(n,m-1), D(n-1,m-1)\} + c(x_n, y_m)$

    - When m or n = 1

      - $D(n,1) = \sum_{k=1:n} c(x_k, 1);$        **Time complexity: O(MN)**

      - $D(1,m) = \sum_{k=1:m} c(1, y_k);$

# Trace back to Get p* from D

**Algorithm:** OPTIMALWARPINGPATH

**Input:**    Accumulated cost matrix $D$.
**Output:**  Optimal warping path $p^*$.

**Procedure:** The optimal path $p^* = (p_1, \ldots, p_L)$ is computed in reverse order of the indices starting with $p_L = (N, M)$. Suppose $p_\ell = (n, m)$ has been computed. In case $(n, m) = (1, 1)$, one must have $\ell = 1$ and we are finished. Otherwise,

$$p_{\ell-1} := \begin{cases} (1, m - 1), & \text{if } n = 1 \\ (n - 1, 1), & \text{if } m = 1 \\ \text{argmin}\{D(n - 1, m - 1), \\ \qquad D(n - 1, m), D(n, m - 1)\}, \text{otherwise}, \end{cases} \qquad (4.6)$$
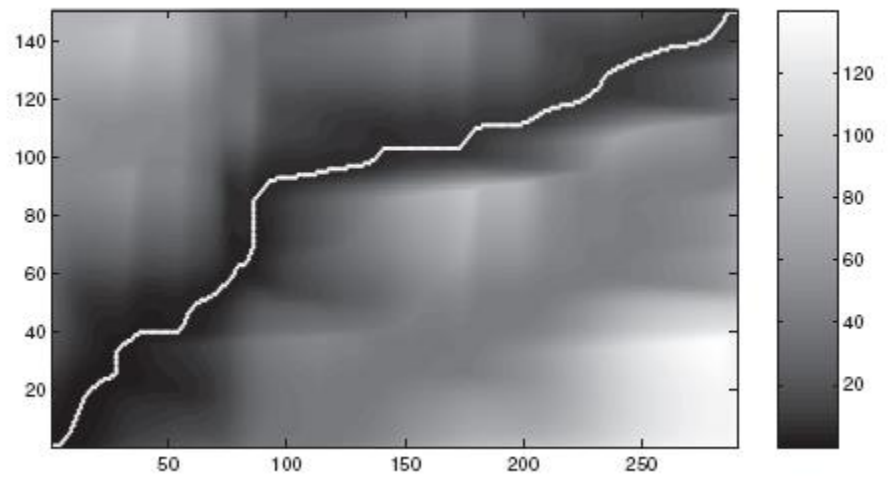
where we take the lexicographically smallest pair in case "argmin" is not unique.

# Example
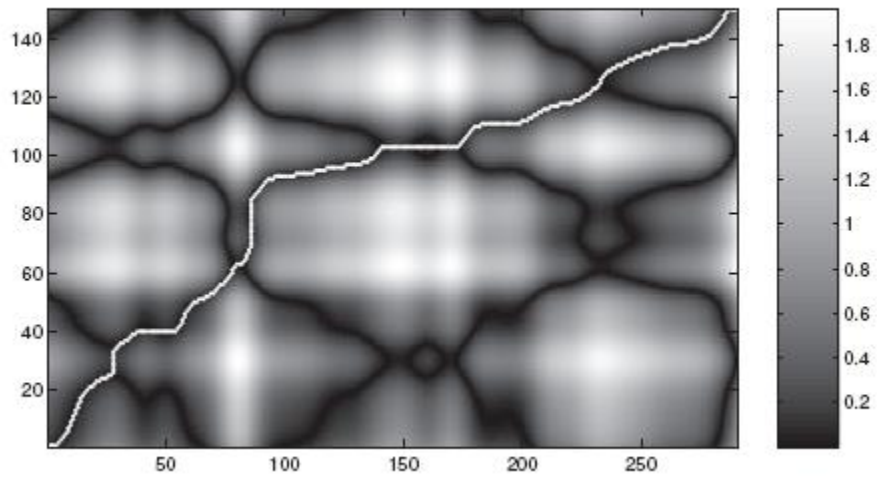


(a)                                           (b)

# Time Domain vs. Frequency Domain

- Many techniques for signal analysis require the data to be in the frequency domain

- Usually data-independent transformations are used

  - The transformation matrix is determined a priori

    - discrete Fourier transform (DFT)

    - discrete wavelet transform (DWT)

- The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain
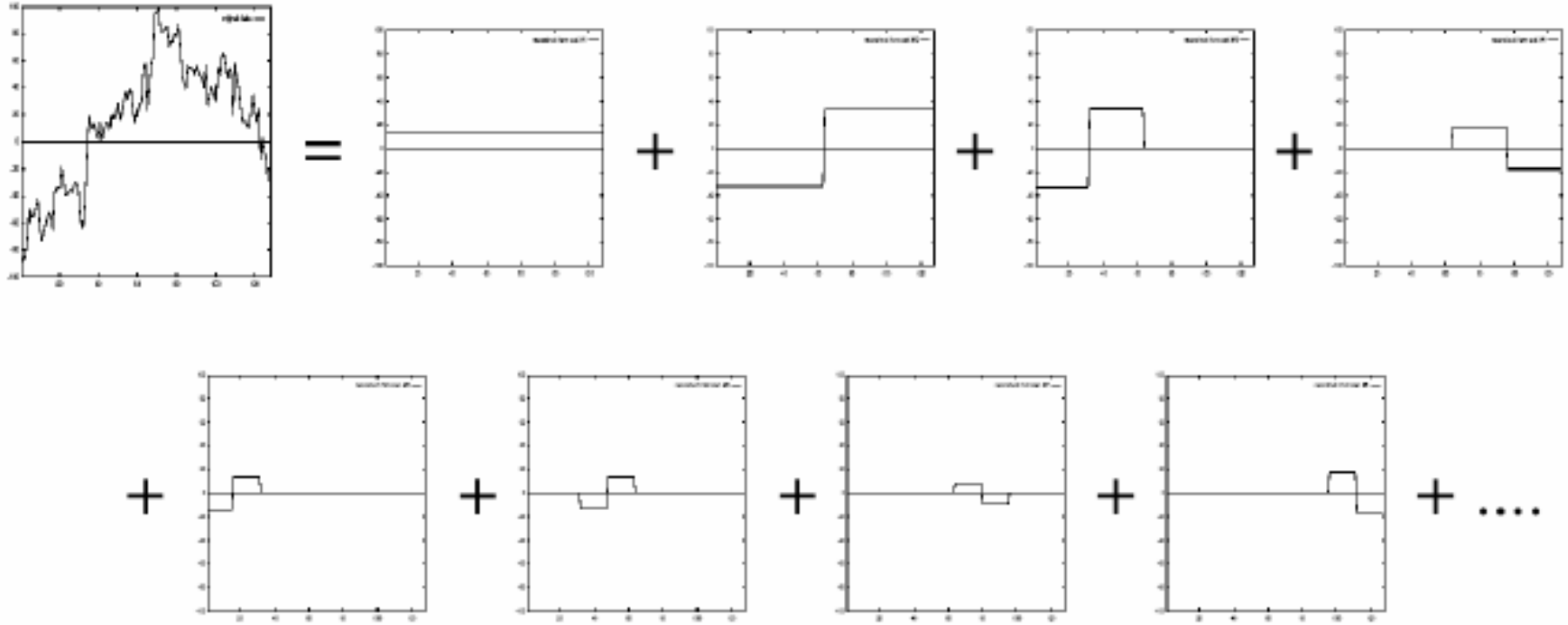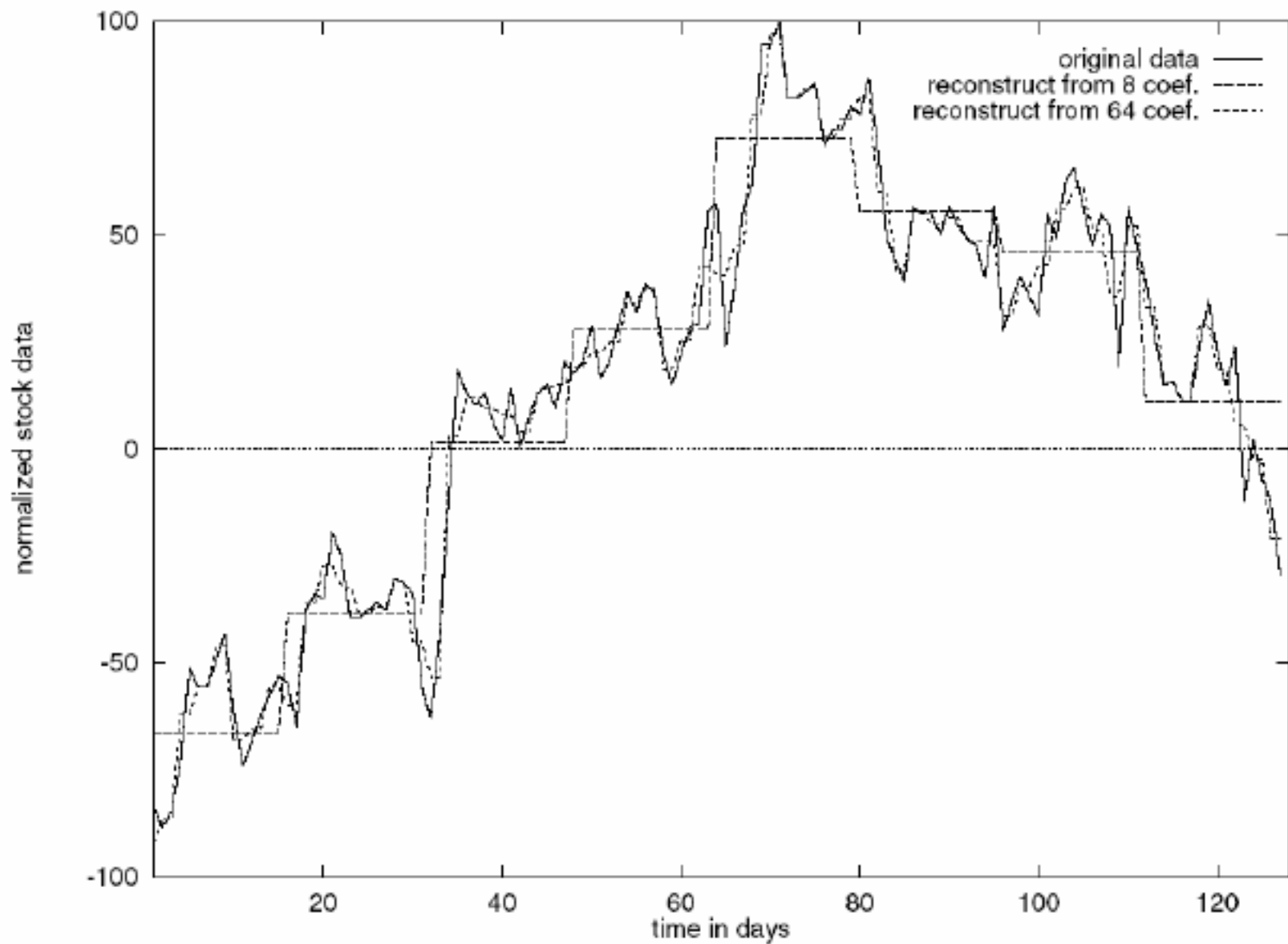
# Example of DFT
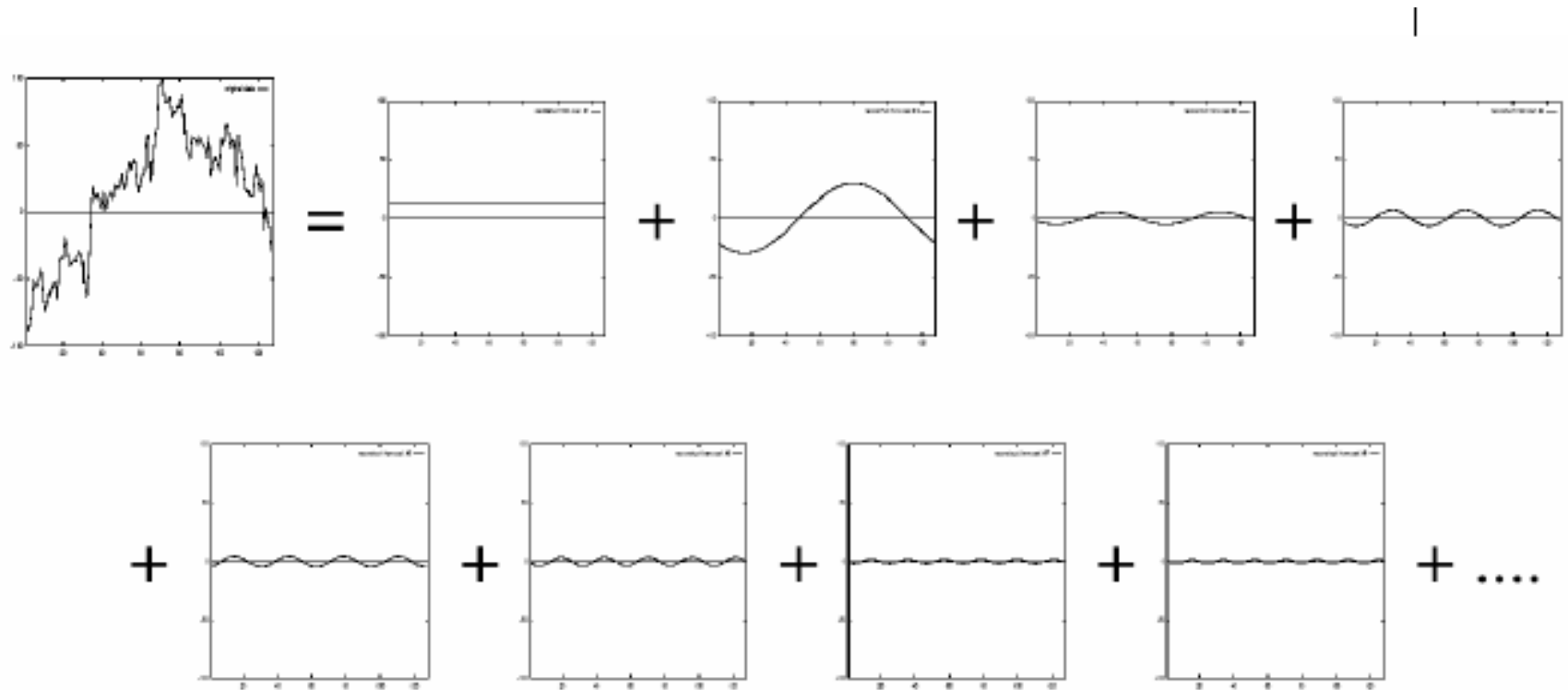


Figures taken from: "A comparison of DFT and DWT based similarity search in Time-series Databases" (Also figures on slide 9,17,18,24,25)
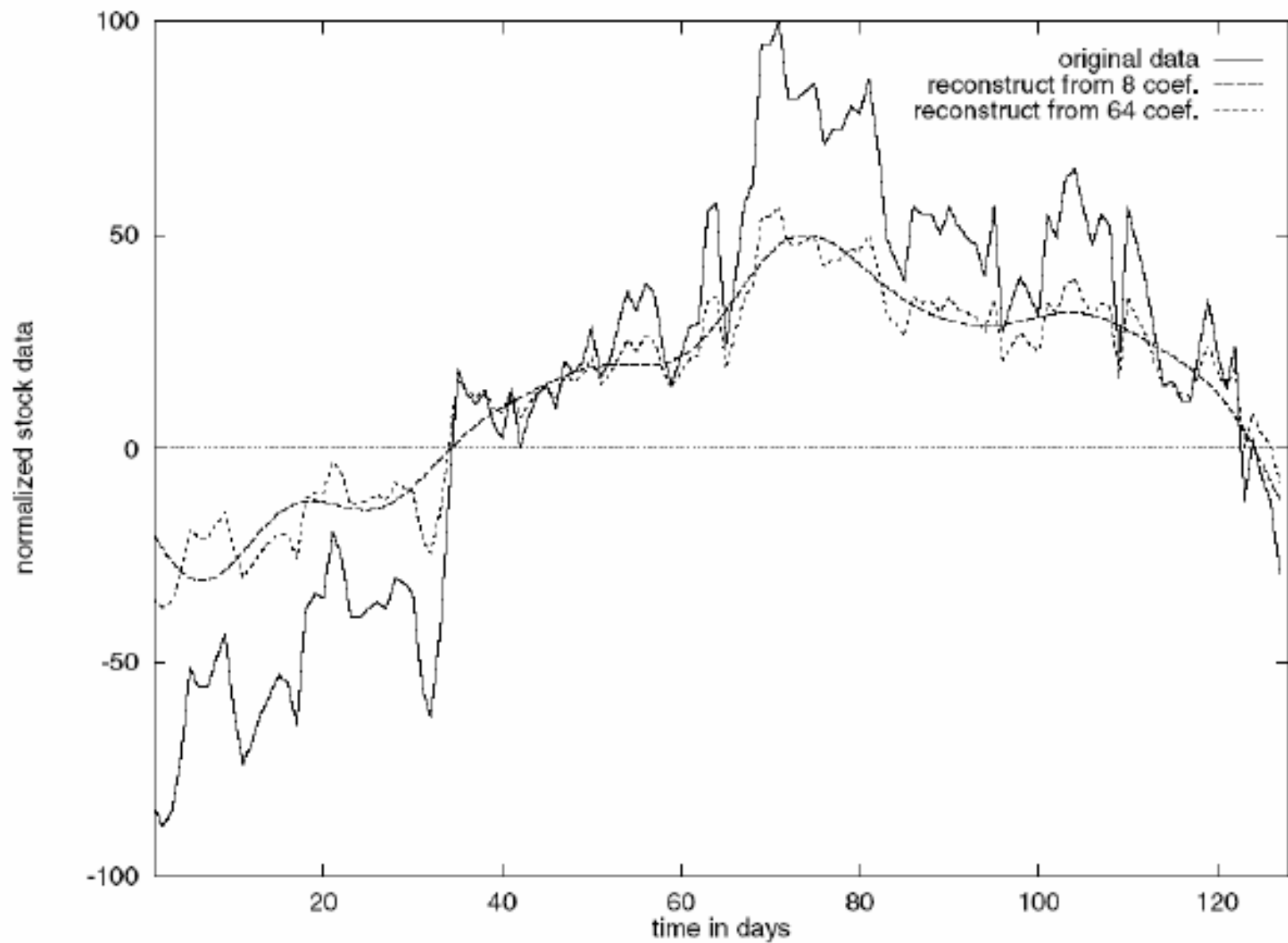
# Example of DWT (with Harr Wavelet)

# *Discrete Fourier Transformation

from $\vec{x} = [x_t], t = 0, \ldots, n-1$ to $\vec{X} = [X_f], f = 0, \ldots, n-1$:

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp(-j2\pi ft/n), \ f = 0, 1, \ldots, n-1$$

- DFT does a good job of concentrating energy in the first few coefficients

- If we keep only first a few coefficients in DFT, we can compute the lower bounds of the actual distance

- Feature extraction: keep the first few coefficients (F-index) as representative of the sequence
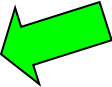
# *DFT (Cont.)

- Parseval's Theorem

$$\sum_{t=0}^{n-1} | \ x_t \ |^2 = \sum_{f=0}^{n-1} | \ X_f \ |^2$$

- The Euclidean distance between two signals in the time domain is the same as their distance in the frequency domain

- Keep the first few (say, 3) coefficients underestimates the distance and there will be no false dismissals!

$$\sum_{t=0}^{n} | \ S[t] - Q[t] \ |^2 \leq \varepsilon \Rightarrow \sum_{f=0}^{3} | \ F(S)[f] - F(Q)[f] \ |^2 \leq \varepsilon$$

# Mining Time Series Data

- Basic Concepts

- Time Series Prediction and Forecasting

- Time Series Similarity Search

- Summary

# Summary

- Time Series Prediction and Forecasting
  - Autocorrelation; autoregression
- Time series similarity search
  - Euclidean distance and Lp norm
  - Dynamic time warping
  - Time domain vs. frequency domain