

CS145: INTRODUCTION TO DATA MINING

13: Time Series Data

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

November 2, 2021

Mid-term Exam

- Time
 - Thursday, 11/4, 10am-11:40am
- Location
 - In Class
- Policy
 - Closed book exam
 - You can take one “reference sheet” of A4 size
 - You can bring a **simple** calculator

Content to Cover

- All the content covered by this class

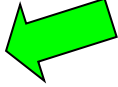
Type of Questions

- True or false / Multiple Choices / Short Answers
 - Conceptual questions
 - Computation questions

Methods to be Learnt

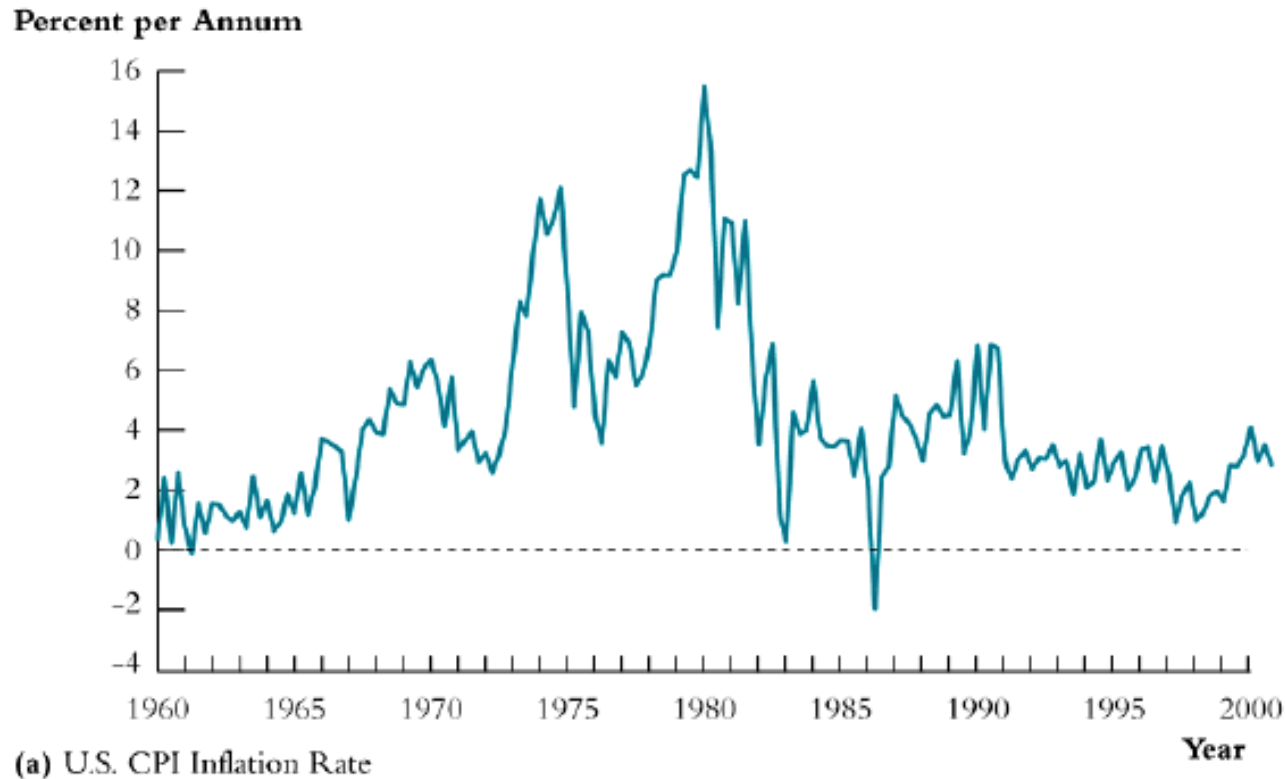
	Vector Data	Set Data	Sequence Data/Time Series	Text Data	Graph Data
Classification	Logistic Regression; Decision Tree; NN			Naïve Bayes for Text	Label Propagation
Clustering	K-means; Mixture Models			PLSA	Spectral Clustering
Prediction	Linear Regression; Regression Tree; NN GLM*		AR Model		
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan		
Similarity Search			DTW		P-PageRank
Ranking					PageRank

Time Series Data

- Basic Concepts 
- Time Series Prediction and Forecasting
- Time Series Similarity Search
- Summary

Example: Inflation Rate Time Series

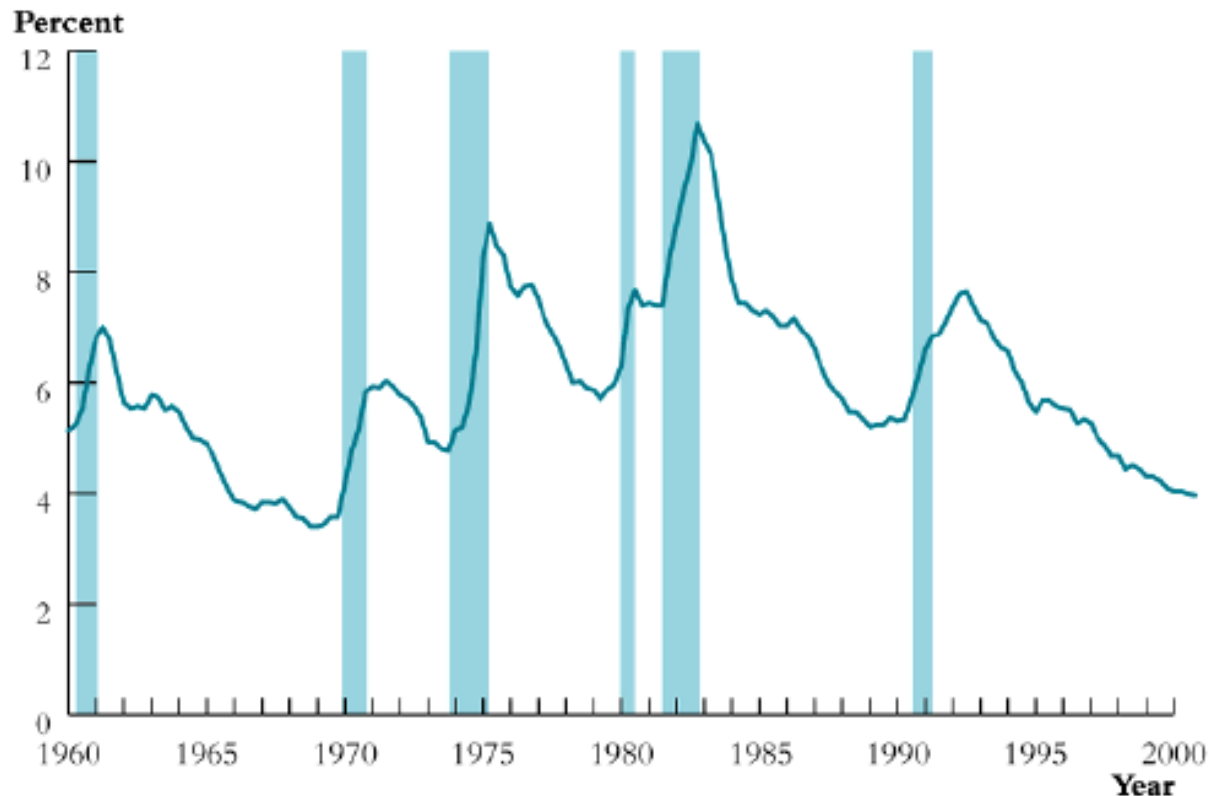
FIGURE 12.1 Inflation and Unemployment in the United States, 1960–1999



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

Example: Stock

Facebook, Inc. (FB) - NasdaqGS ★ Follow

[+ Add to Portfolio](#)

[f Like](#) < 10k

46.58 ↑ 0.38 (0.82%) 1:10PM EST - Nasdaq Real Time Price

Enter name(s) or symbol(s)

GET CHART

COMPARE

EVENTS ▾

TECHNICAL INDICATORS ▾

CHART SETTINGS ▾

RESET

Oct 4, 2013: ■ FB 51.04



© 2013 Yahoo! Inc.

■ Volume: 74,447,000



Market Summary > Facebook, Inc. Common Stock

NASDAQ: FB

[+ Follow](#)

135.92 USD **+4.19 (3.18%)** ↑

Nov 26, 2:15 PM EST · Disclaimer

1 day

5 days

1 month

6 months

YTD

1 year

5 years

Max

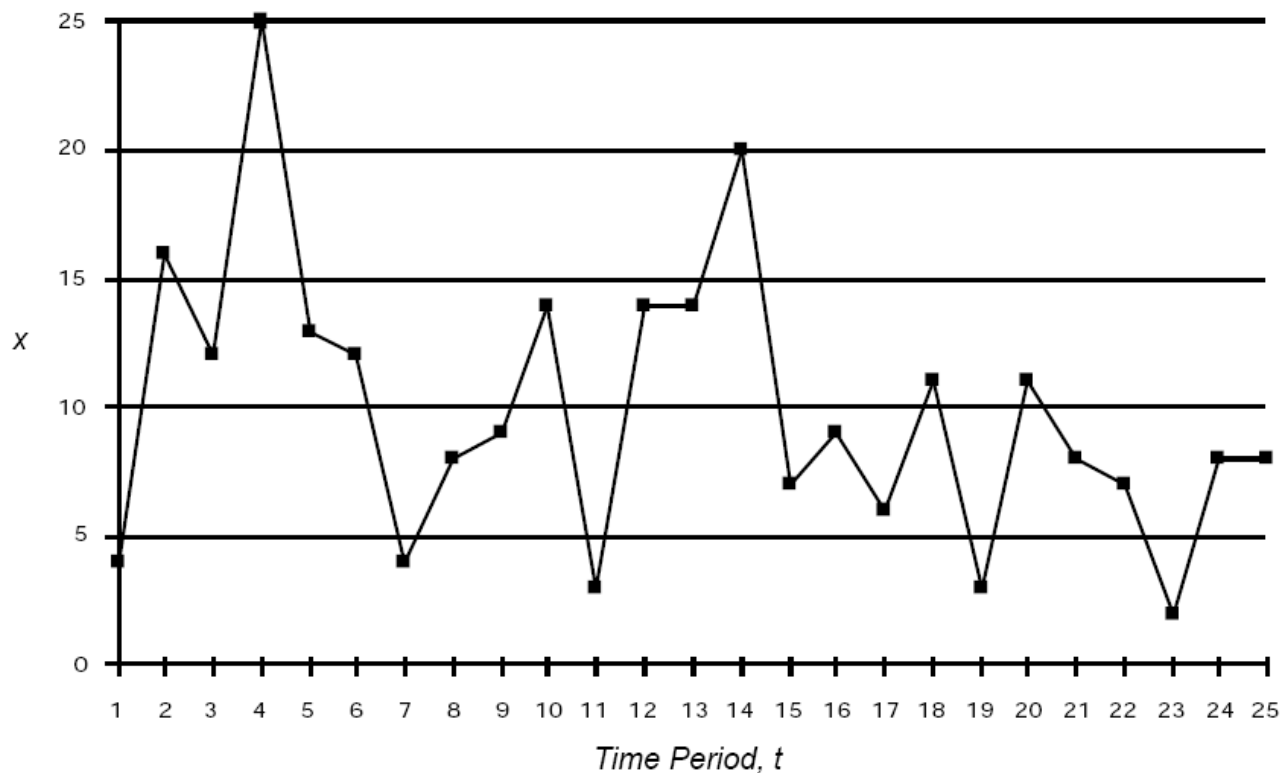


Open	133.00
High	136.16
Low	132.78
Mkt cap	390.61B
P/E ratio	20.47

Div yield	-
Prev close	131.73
52-wk high	218.62
52-wk low	126.85

Example: Product Sale

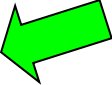
Time	Observations									
1 - 10	4	16	12	25	13	12	4	8	9	14
11 - 20	3	14	14	20	7	9	6	11	3	11
20 - 25	8	7	2	8	8	10	7	16	9	4



Time Series

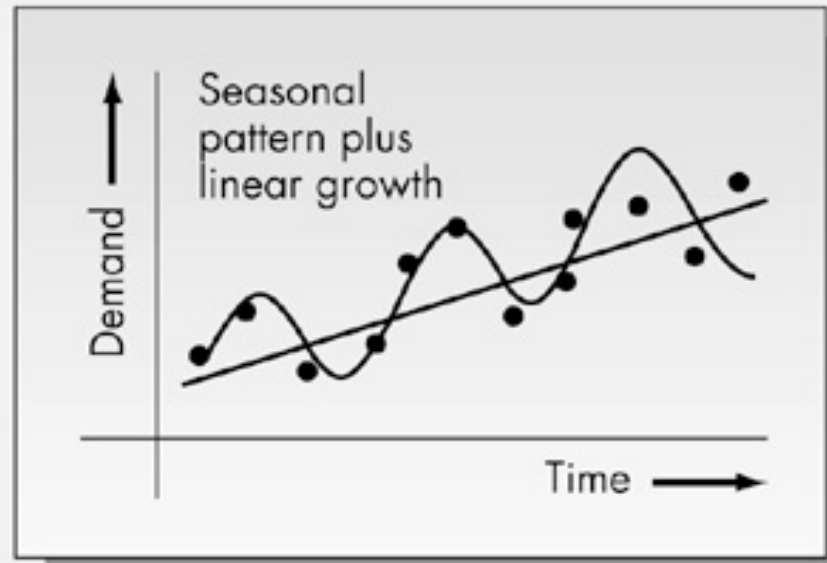
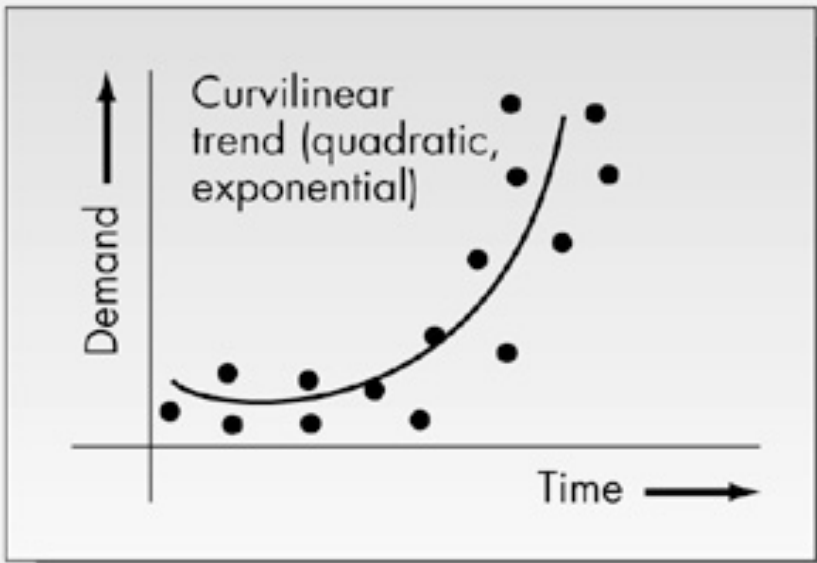
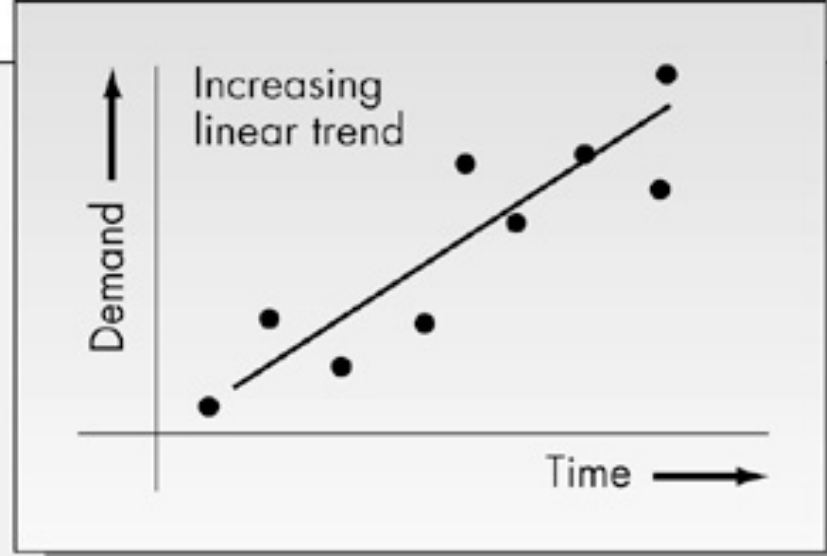
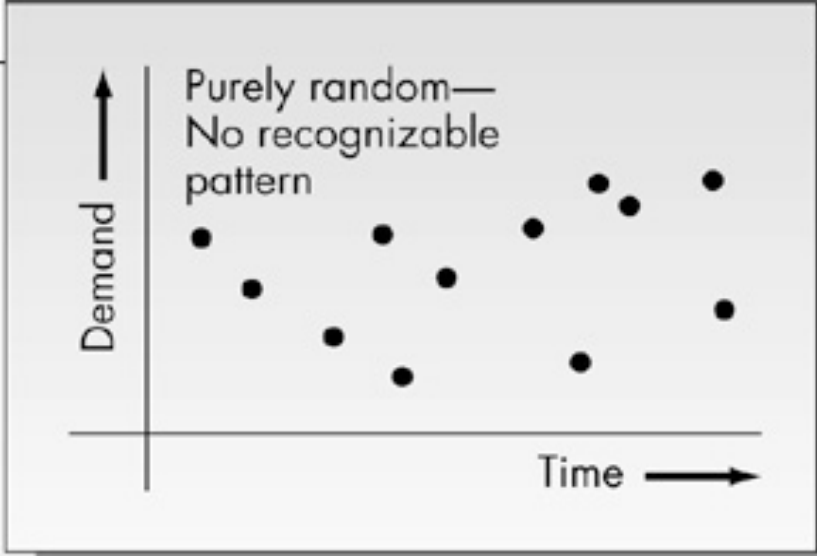
- A **time series** is a sequence of numerical data points, measured typically at successive times, spaced at (often uniform) time intervals
 - *Random variables for a time series are Represented as:*
 - $Y = \{Y_1, Y_2, \dots\}$, or
 - $Y = \{Y_t : t \in T\}$, where T is the index set
 - An observation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

Time Series Data

- Basic Concepts
- Time Series Prediction and Forecasting 
- Time Series Similarity Search
- Summary

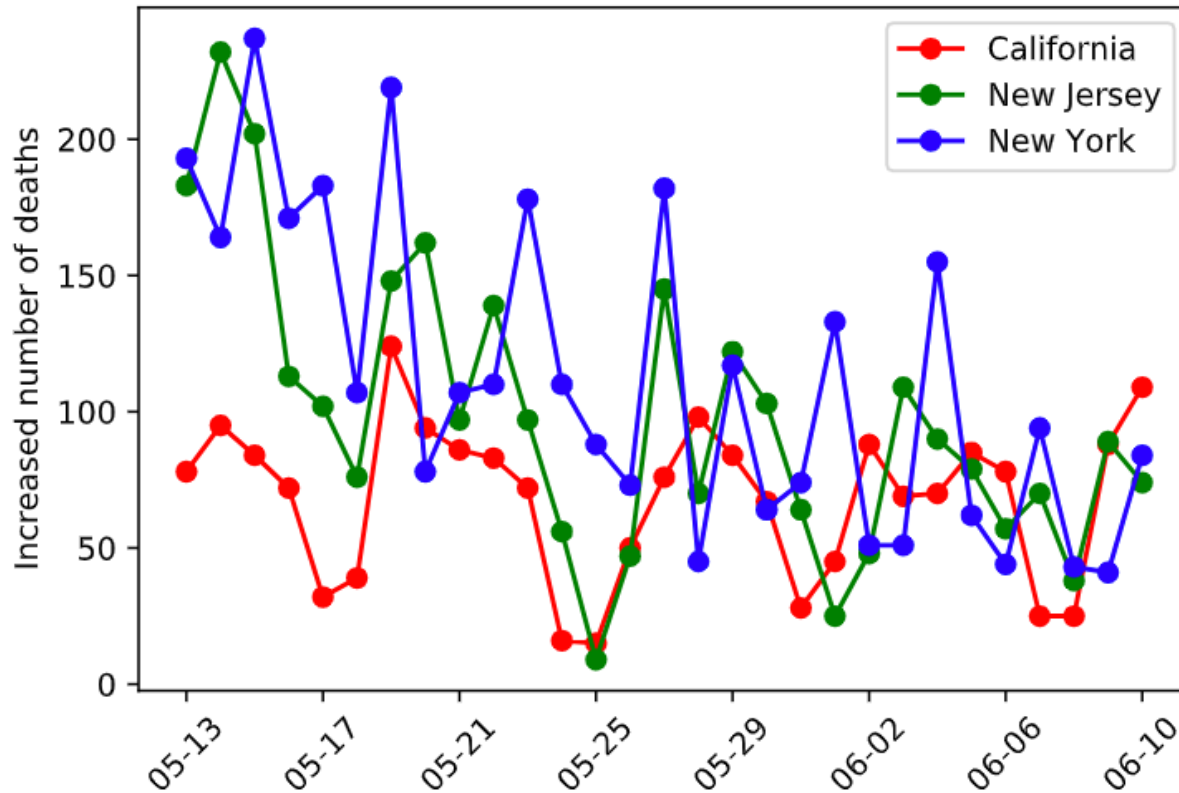
Categories of Time-Series Movements

- 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
 - E.g., almost identical patterns that a time series appears to follow during corresponding months of successive years.
 - Irregular or random movements



Example

- Example: the daily confirmed deaths/cases at a location



COVID-19 death count time series for 3 different states.

Lag, Difference

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

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

Quarter	U.S. CPI	Rate of Inflation at an Annual Rate (Inf_t)	First Lag (Inf_{t-1})	Change in Inflation (ΔInf_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

Autocorrelation

- Autocorrelation: the correlation between a time series and its lagged values
 - The first autocorrelation ρ_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 j th autocorrelation ρ_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 j th sample autocorrelation

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

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

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

Y_t	Y_{t-j}
y_{j+1}	y_1
y_{j+2}	y_2
\vdots	\vdots
y_{T-1}	y_{T-j-1}
y_T	y_{T-j}

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

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		
Lag	Autocorrelation of:	
	Inflation Rate (Inf_t)	Change of Inflation Rate (ΔInf_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

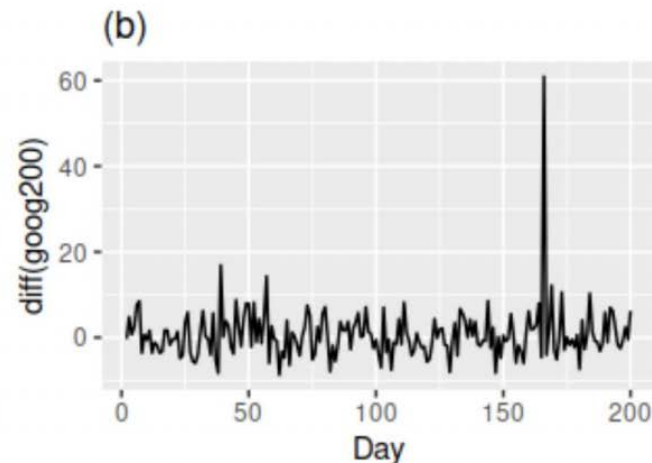
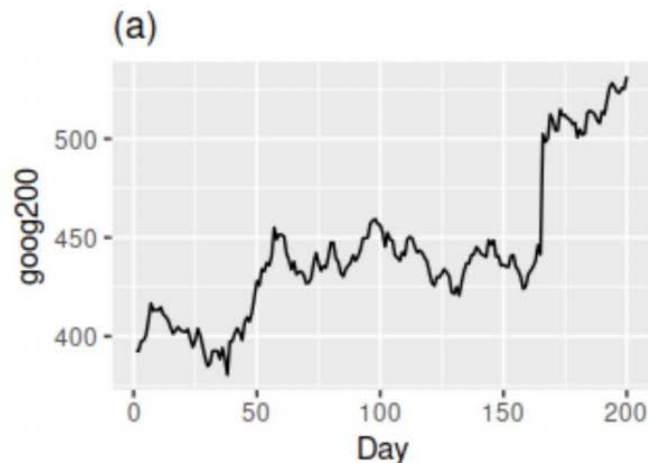
Focus on Stationary Time Series

- 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 $(Y_{s+1}, Y_{s+2}, \dots, Y_{s+T})$ 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 $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \dots, X_{s+T}, Y_{s+T})$ does not depend on s . Stationarity requires the future to be like the past, at least in a probabilistic sense.

General Assumption: Stationary

- Stationary timeseries tends to be fluctuating around some fixed level.
- In practice, a simple trick to achieve stationary is to **differ** the original timeseries.



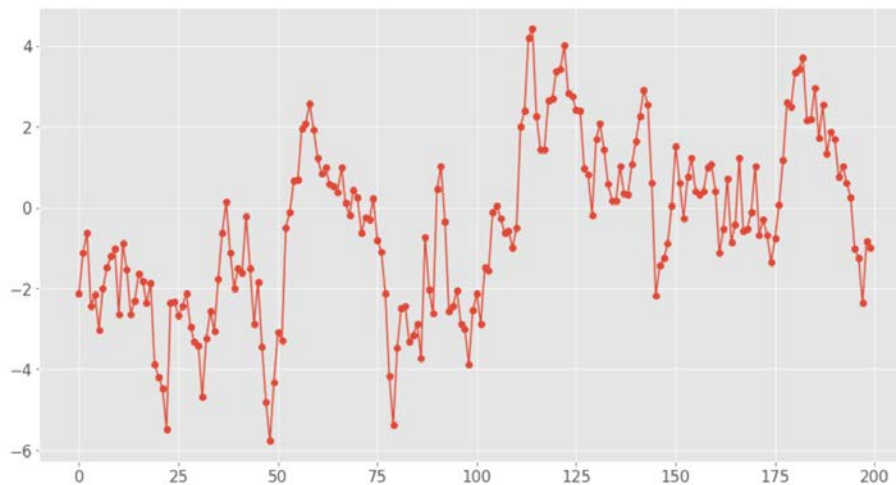
Autoregression

- Use past values Y_{t-1}, Y_{t-2}, \dots 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 **p th order autoregression**, Y_t is regressed against $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$

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}
 - w_t is the white noise: $u_t \sim N(0, \sigma^2)$



$$Y_t = 0.9Y_{t-1} + u_t$$

The p th Order Autoregression Model

AR(p)

- AR(p) model

- $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t = \sum_{i=1}^p \beta_i Y_{t-i} + \beta_0 + u_t$

- u_t is the white noise

- p is the number of lag observations included in the model, also called the lag order.

- Y_t is assumed to be stationary.

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

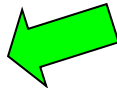
- So far we have considered forecasting models that use only past values of Y
- It makes sense to add other variables (X) that might be useful predictors of Y , above and beyond the predictive value of lagged values of Y :

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

A good reference for more

- <https://online.stat.psu.edu/stat510>

Time Series Data

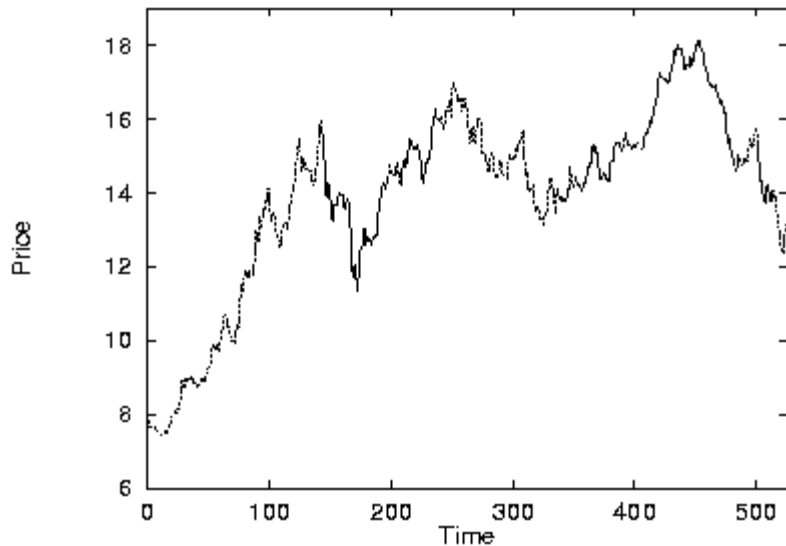
- Basic Concepts
- Time Series Prediction and Forecasting
- Time Series Similarity Search 
- Summary

Why Similarity Search?

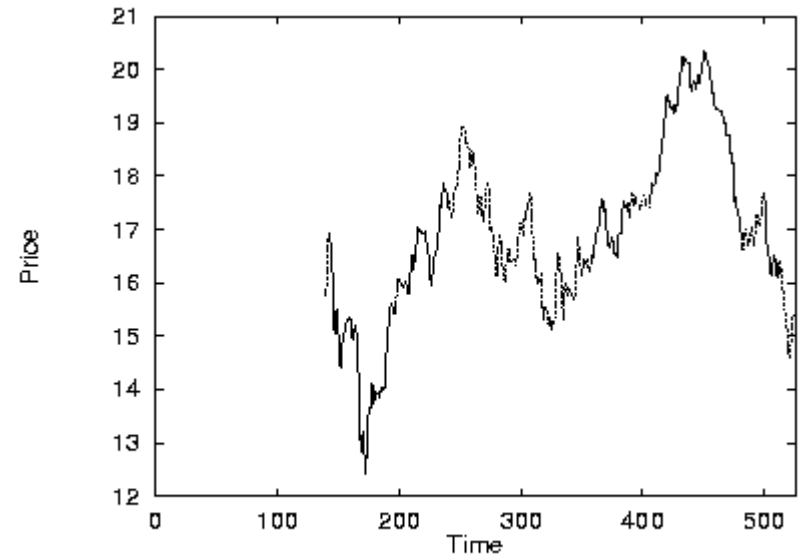
- Wide applications
 - Find a time period with similar inflation rate and unemployment time series?
 - Find a similar stock to Facebook?
 - Find a similar product to the query according to sales time series?
 - ...

Example

VanEck International Fund



Fidelity Selective Precious Metal and Mineral Fund



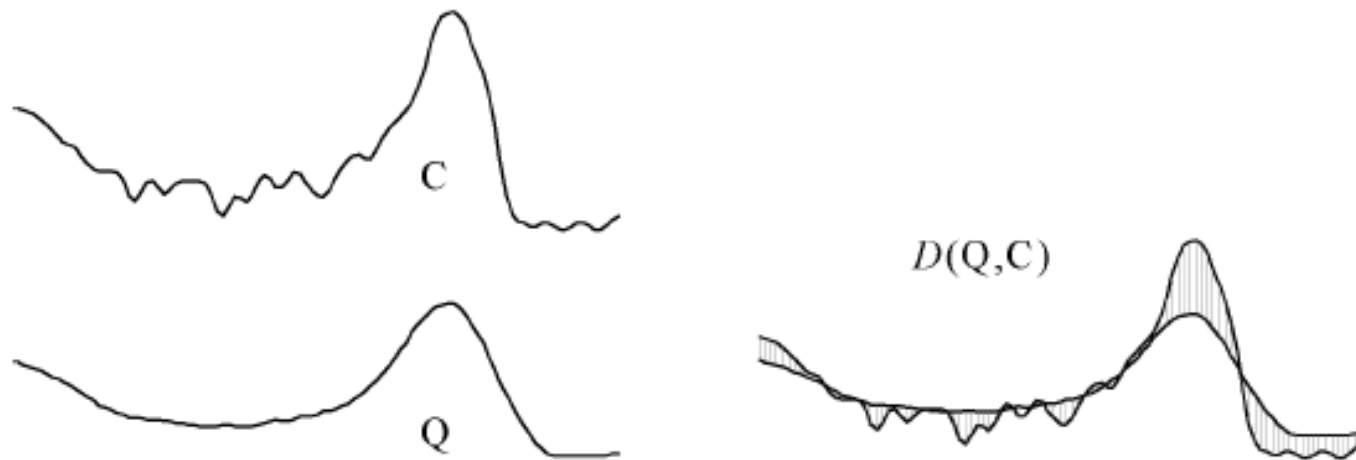
Two similar mutual funds in the different fund group

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

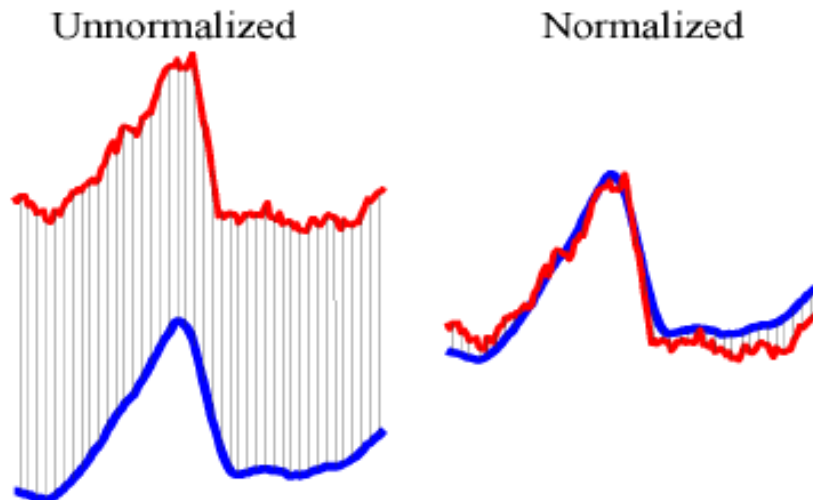
- Given two time series with **equal** length n
 - $C = \{c_1, c_2, \dots, c_n\}$
 - $Q = \{q_1, q_2, \dots, q_n\}$
 - $d(C, Q) = (\sum |c_i - q_i|^p)^{1/p}$
 - When $p=2$, it is Euclidean distance



Enhanced Lp Norm-based Distance

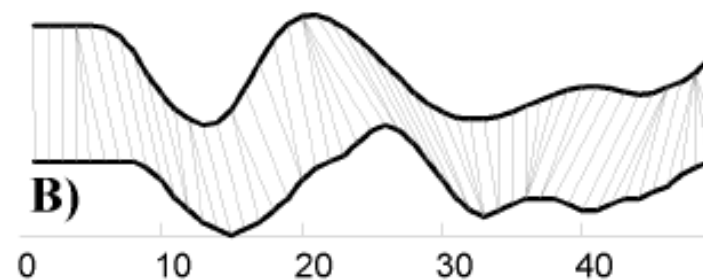
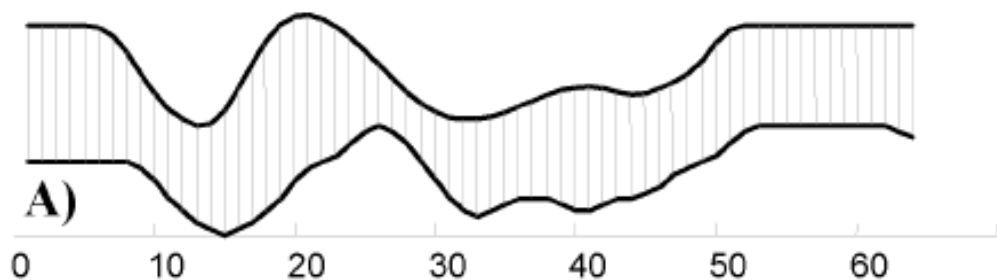
- Issues with Lp Norm: cannot deal with offset and scaling in the Y-axis
- Solution: normalizing the time series

- $$c'_i = \frac{c_i - \mu(C)}{\sigma(C)}$$



Dynamic Time Warping (DTW)

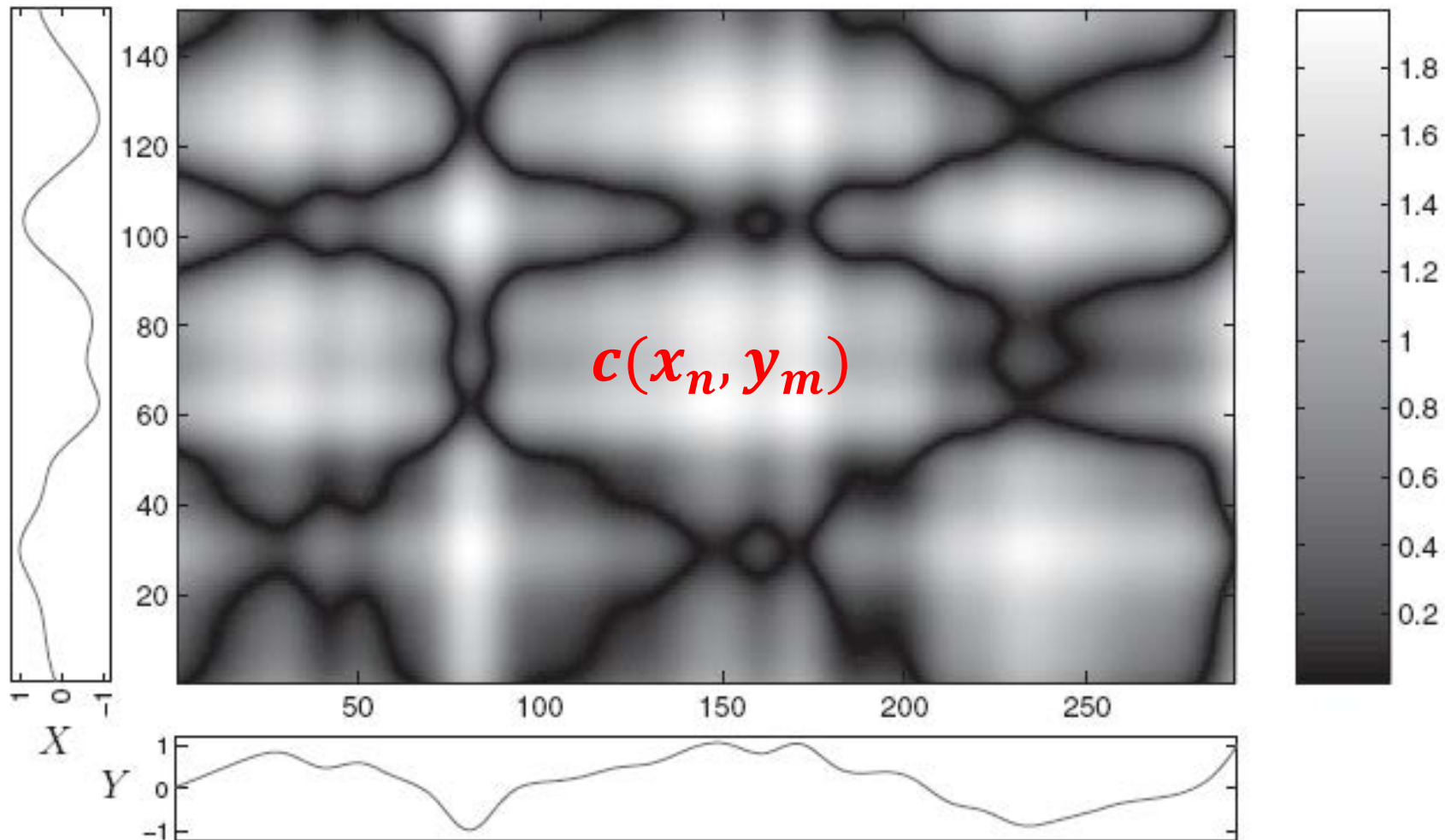
- For two sequences that do not line up well in X-axis, but share roughly similar shape
 - We need to warp the time axis to make better alignment



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 : $c(x_n, y_m)$
- Goal:
 - Find an alignment between X and Y, such that, the overall cost is minimized

Cost Matrix of Two Time Series

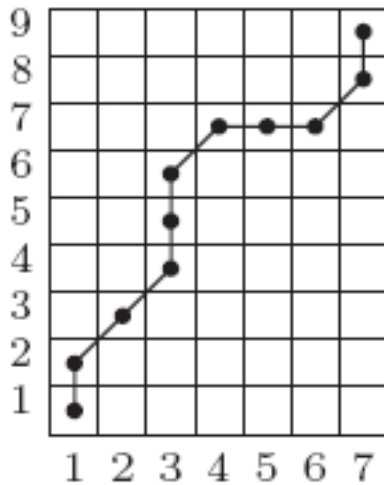


Represent an Alignment by Warping Path

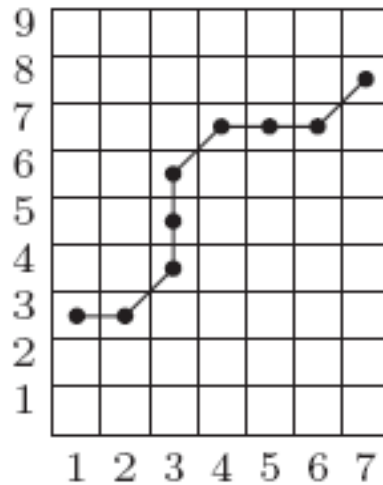
- An (N, M) -warping path is a sequence $p = (p_1, p_2, \dots, p_L)$ with $p_l = (n_l, m_l)$, satisfying the three conditions:
 - Boundary condition: $p_1 = (1, 1), p_L = (N, M)$
 - Starting from the first point and ending at last point
 - Monotonicity condition: n_l and m_l are non-decreasing with l
 - Step size condition:
 - $p_{l+1} - p_l \in \{(0, 1), (1, 0), (1, 1)\}$
 - Move one step up, right, or up-right

Q: Which Path is a Warping Path?

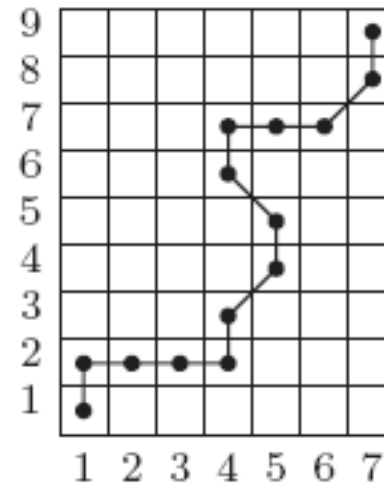
(a)



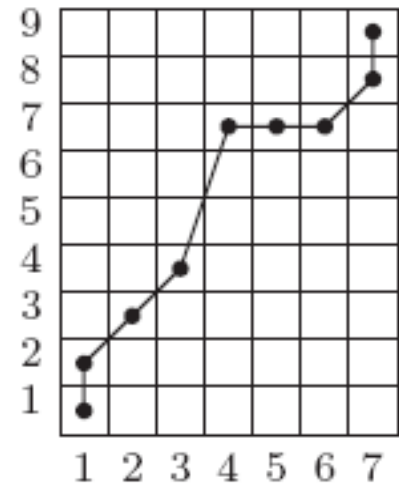
(b)



(c)



(d)



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^*
 - $c_{p^*}(X, Y) = \min\{c_p(X, Y) \mid p \text{ is an } (N, M) - \text{warping path}\}$
- 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 $X(1,\dots,n)$ and $Y(1,\dots,m)$
 - D is called accumulative cost matrix
 - Note $D(N,M) = \text{DTW}(X,Y)$
 - 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 $n = 1$
 - $D(n, 1) = \sum_{k=1:n} c(x_k, y_1);$
 - $D(1, m) = \sum_{k=1:m} c(x_1, y_k);$

Time complexity: $O(MN)$

Trace back to Get p^* from D

Algorithm: OPTIMALWARPINGPATH

Input: Accumulated cost matrix D .

Output: Optimal warping path p^* .

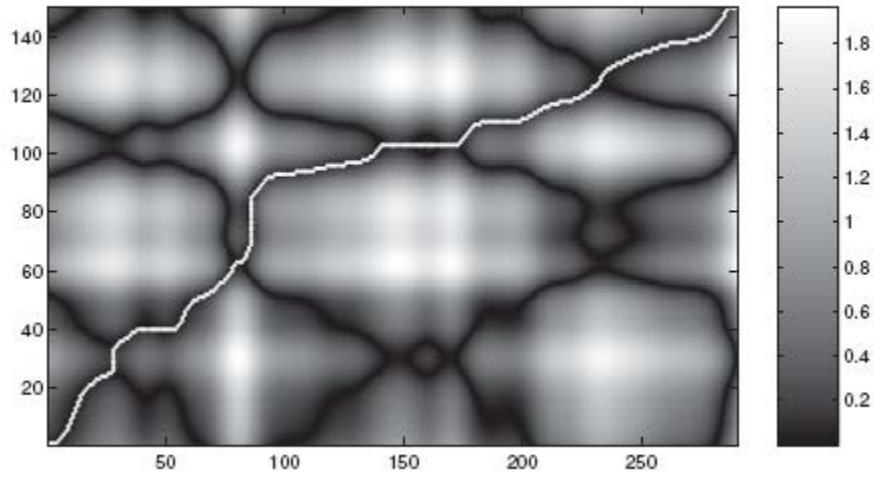
Procedure: The optimal path $p^* = (p_1, \dots, p_L)$ is computed in reverse order of the indices starting with $p_L = (N, M)$. Suppose $p_\ell = (n, m)$ has been computed. In case $(n, m) = (1, 1)$, one must have $\ell = 1$ and we are finished. Otherwise,

$$p_{\ell-1} := \begin{cases} (1, m-1), & \text{if } n = 1 \\ (n-1, 1), & \text{if } m = 1 \\ \operatorname{argmin}\{D(n-1, m-1), \\ \quad D(n-1, m), D(n, m-1)\}, & \text{otherwise,} \end{cases} \quad (4.6)$$

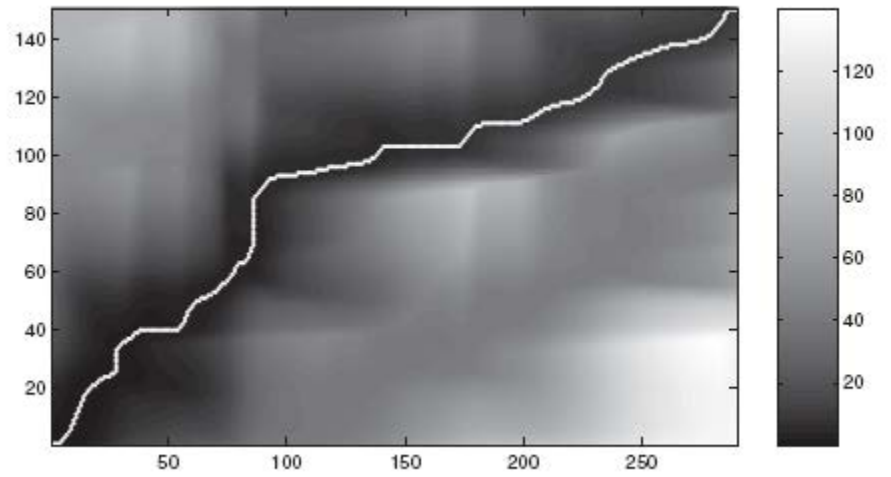
where we take the lexicographically smallest pair in case “argmin” is not unique.

Example

(a)



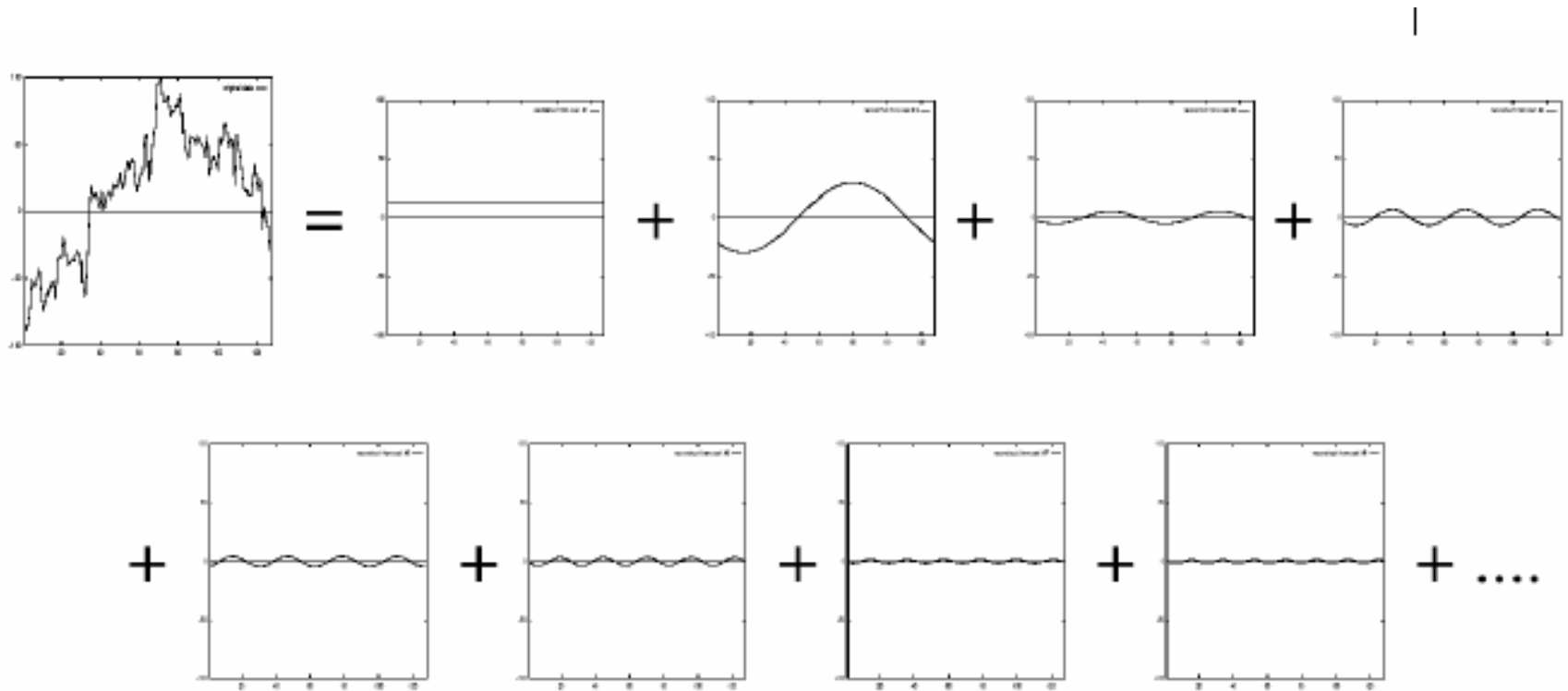
(b)



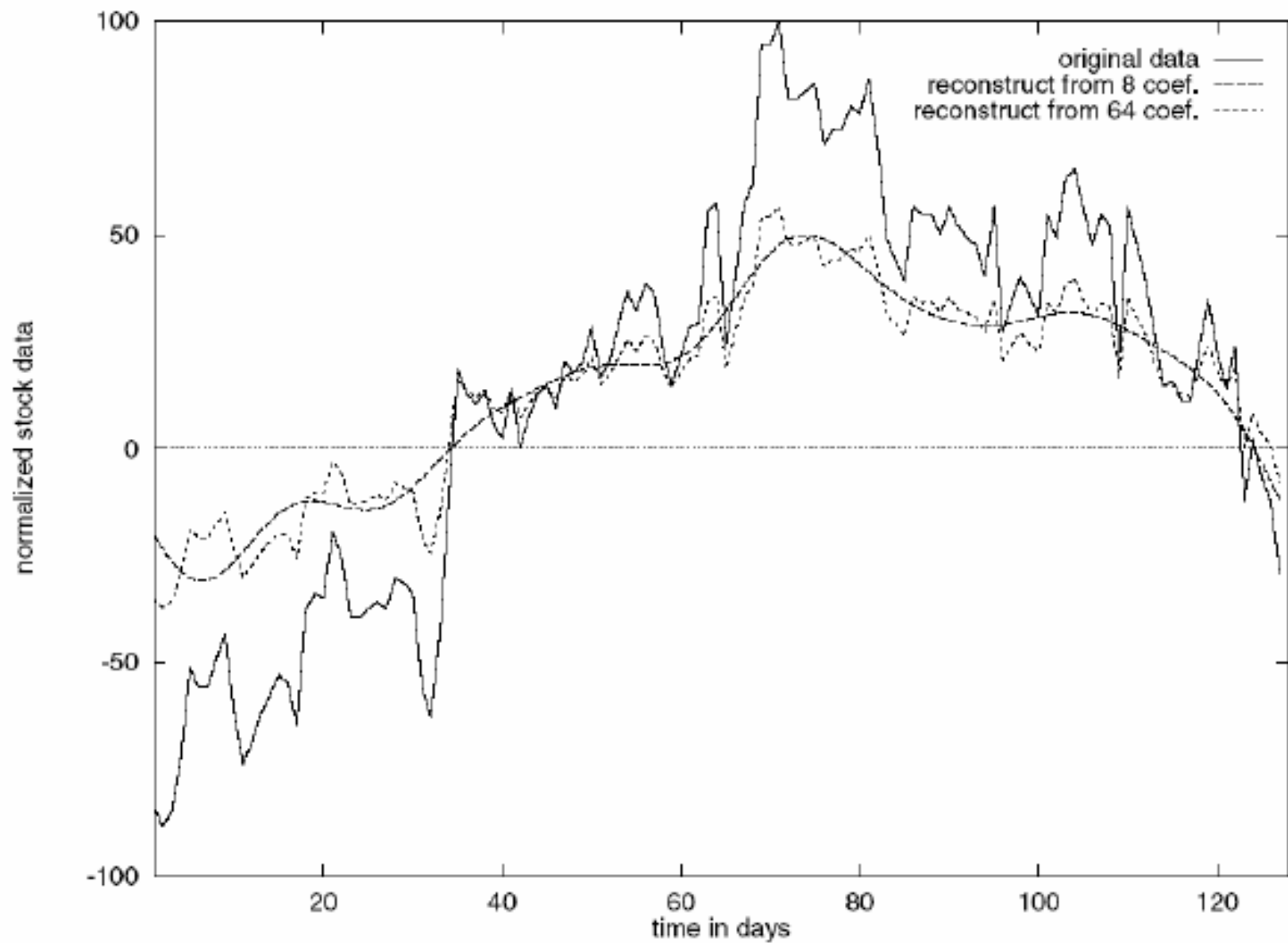
*Time Domain vs. Frequency Domain

- 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

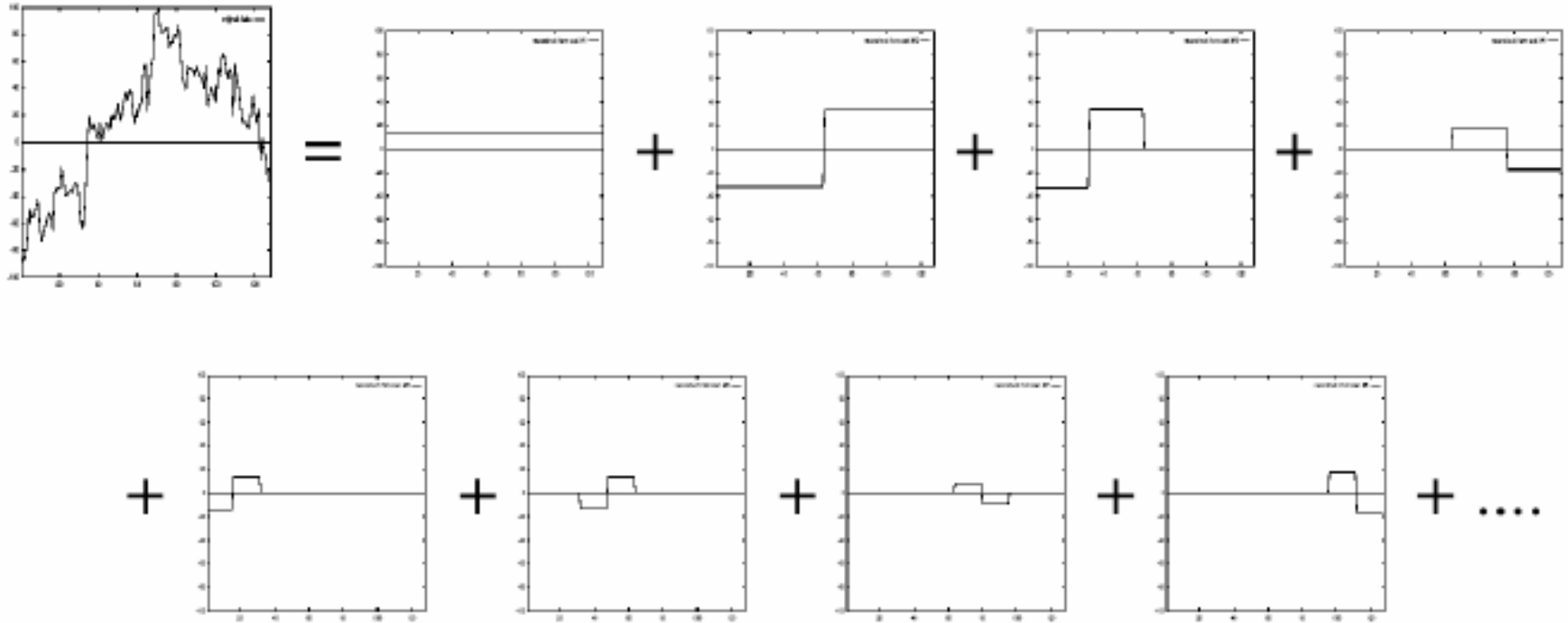
*Example of DFT

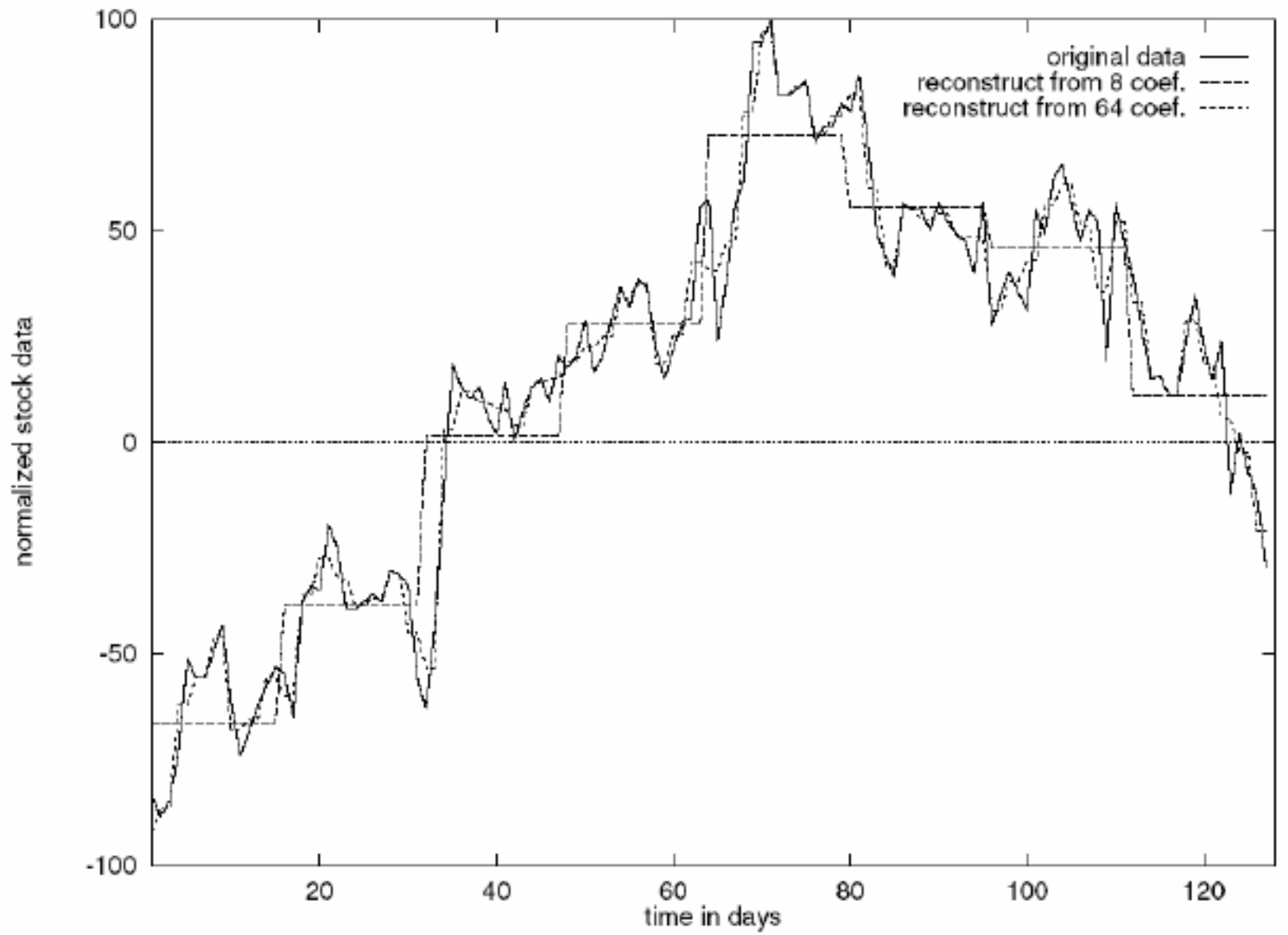


Figures taken from: "A comparison of DFT and DWT based similarity search in Time-series Databases" (Also figures on slide 9,17,18,24,25)



*Example of DWT (with Harr Wavelet)





*Discrete Fourier Transformation

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

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

- 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.)

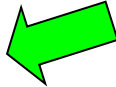
- Parseval's Theorem

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

- 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!

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

Time Series Data

- Basic Concepts
- Time Series Prediction and Forecasting
- Time Series Similarity Search
- Summary 

Summary

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