DATA MINING TECHNIQUES
Mining Time Series Data

Yijun Zhao

Northeastern University

Fall 2016

Time-Series Data

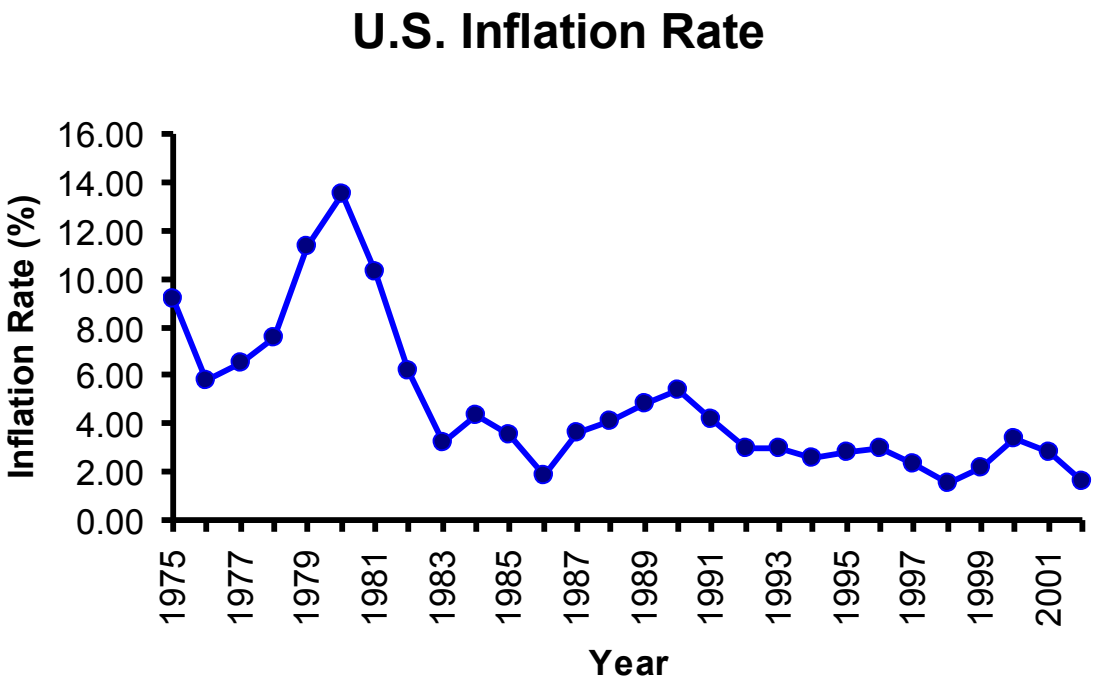
- Numerical data obtained at regular time intervals
- The time intervals can be annually, quarterly, daily, hourly, etc.
- Example:

Year:	1999	2000	2001	2002	2003
Sales:	75.3	74.2	78.5	79.7	80.2

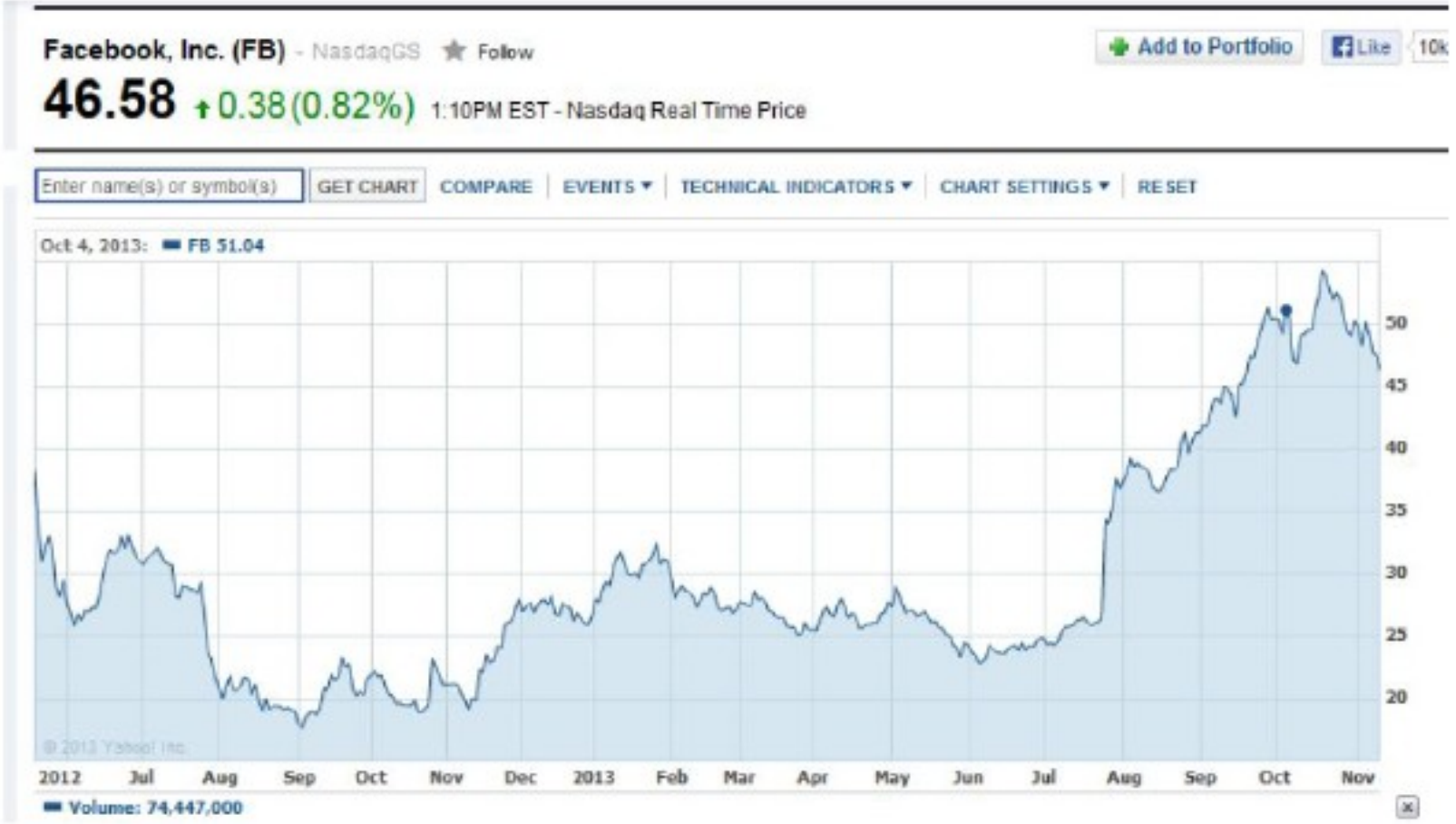
Time-Series Plot

A time-series plot is a two-dimensional plot of time series data

- **the vertical axis measures the variable of interest**
- **the horizontal axis corresponds to the time periods**



Time Series Example



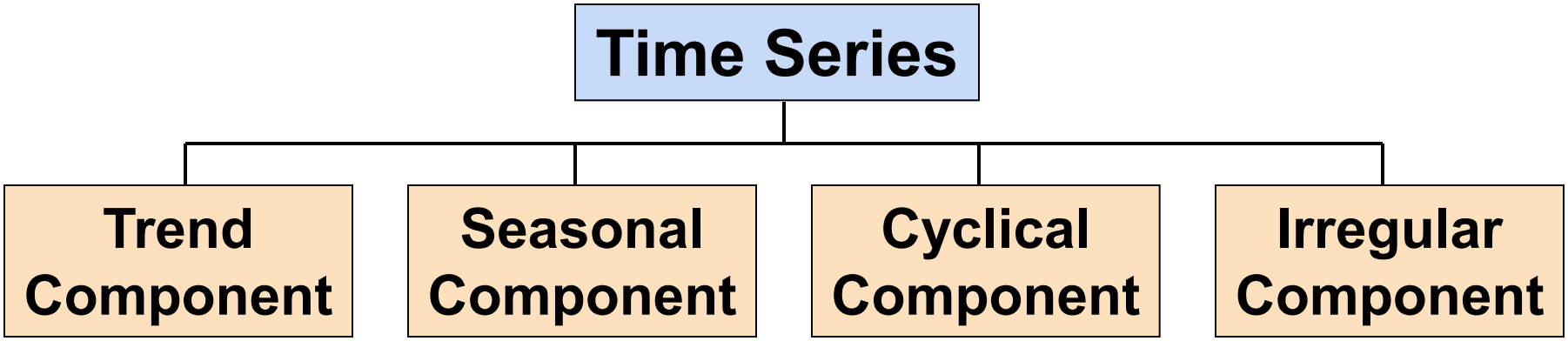
Mining Time Series Data

- **Prediction and Forecasting**
- **Similarity Search**

The Importance of Forecasting

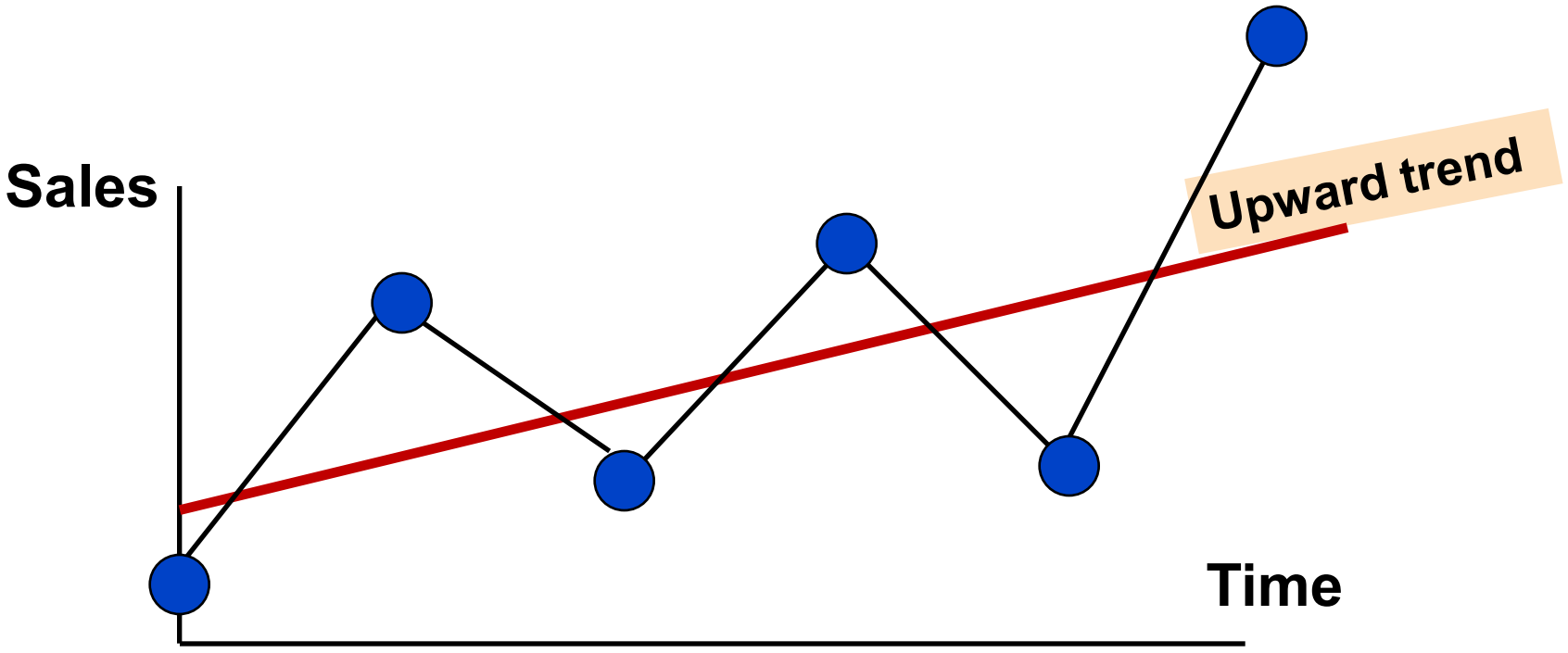
- Governments forecast unemployment, interest rates, and expected revenues from income taxes for policy purposes
- Marketing executives forecast demand, sales, and consumer preferences for strategic planning
- College administrators forecast enrollments to plan for facilities and for faculty recruitment
- Traders forecast stock prices, interest rates and volatilities to make profit

Time-Series Components



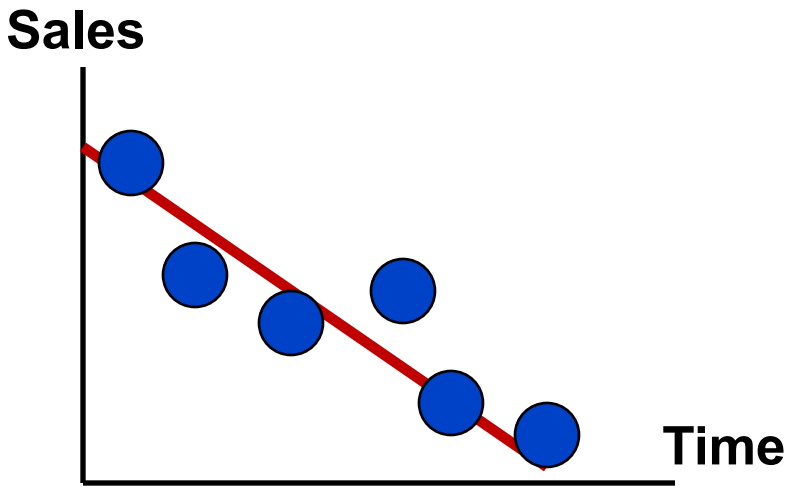
Trend Component

- **Long-run increase or decrease over time (overall upward or downward movement)**
- **Data taken over a long period of time**

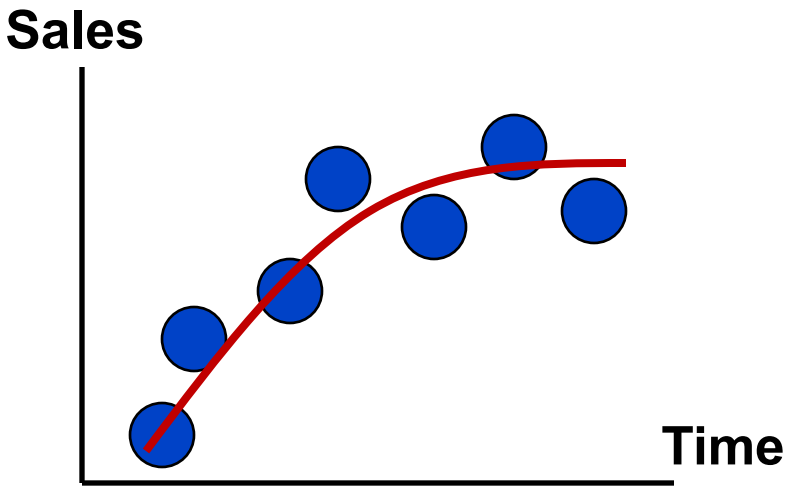


Trend Component

- Trend can be upward or downward
- Trend can be linear or non-linear



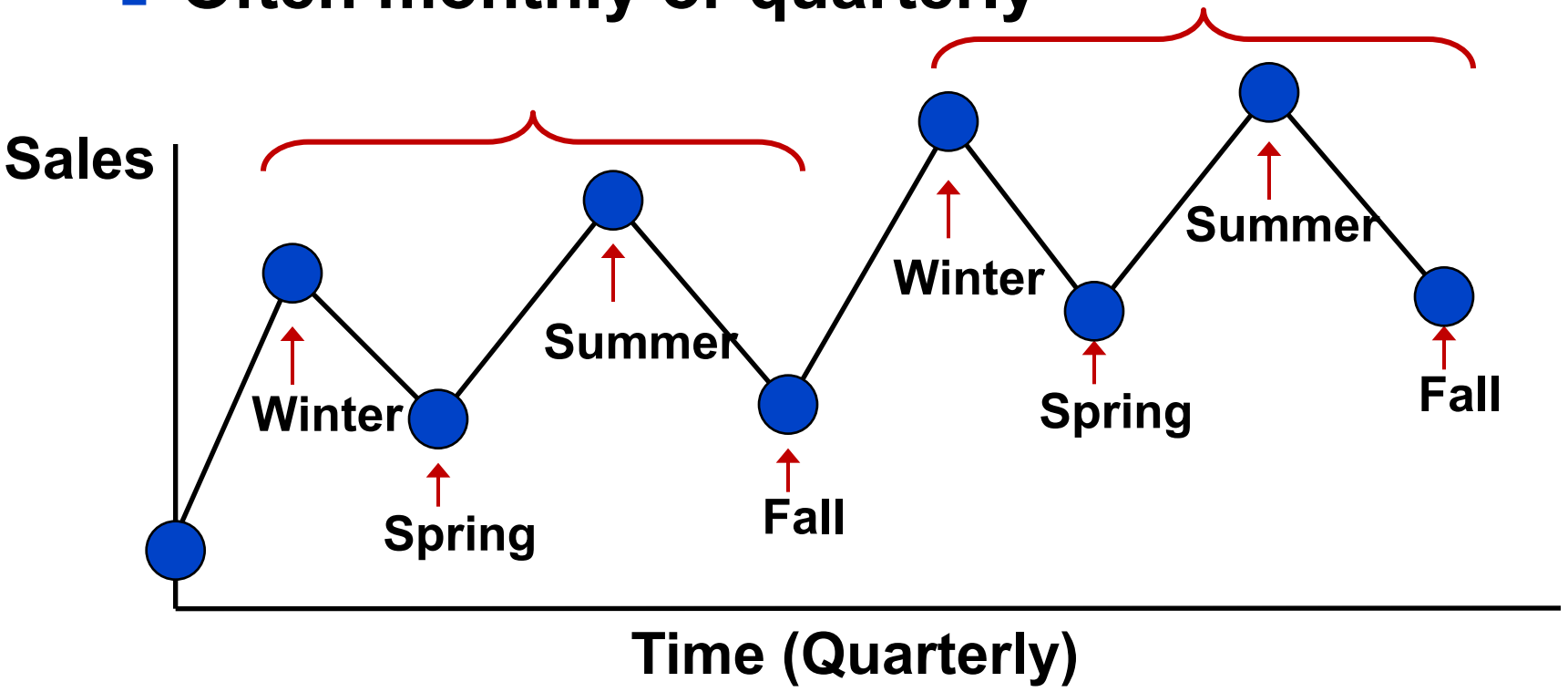
Downward linear trend



Upward nonlinear trend

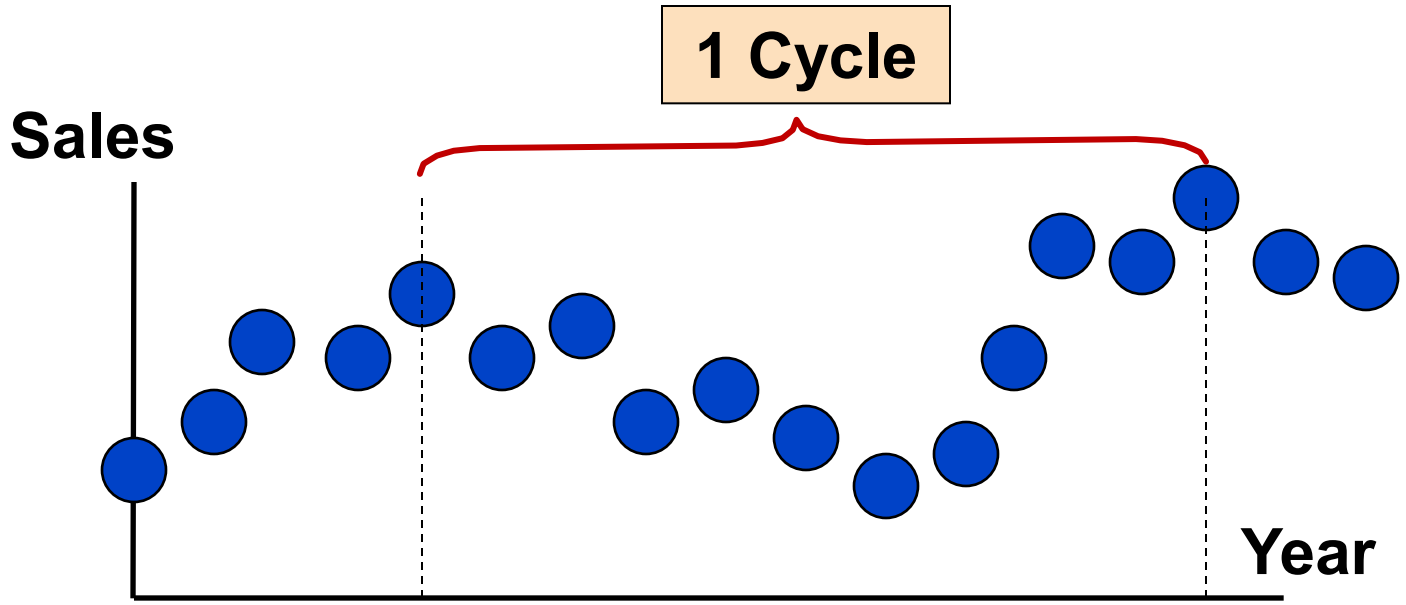
Seasonal Component

- Short-term regular wave-like patterns
- Observed within 1 year
- Often monthly or quarterly



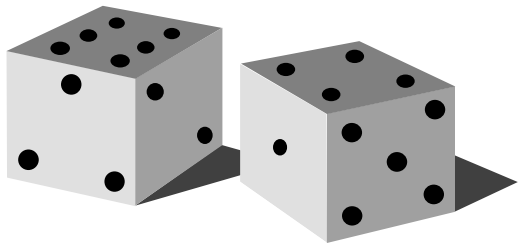
Cyclical Component

- Long-term wave-like patterns
- Regularly occur but may vary in length
- Often measured peak to peak/trough to trough



Irregular Component

- Unpredictable, random, “residual” fluctuations
- Due to random variations of
  - Nature
  - Accidents or unusual events
- “Noise” in the time series



Additive Time-Series Model

- Used primarily for forecasting

$$Y_i = T_i + S_i + C_i + I_i$$

where T_i = Trend value at time i

S_i = Seasonal value at time i

C_i = Cyclical value at time i

I_i = Irregular (random) value at time i

Multiplicative Time-Series Model

- Used primarily for forecasting

$$Y_i = T_i \times S_i \times C_i \times I_i$$

where T_i = Trend value at time i

S_i = Seasonal value at time i

C_i = Cyclical value at time i

I_i = Irregular (random) value at time i

Forecasting Time Series

- Smoothing-based forecasting (moving average)
- Trend based forecasting
- Autoregressive models
- Many alternative models exist

Moving Averages

- Calculate moving averages to get an overall impression of the pattern of movement over time

Moving Average: averages of consecutive time series values for a chosen period of length L

Moving Averages

- Used for smoothing
- A series of arithmetic means over time
- Result dependent upon choice of L (length of period for computing means)
- Examples:
  - For a 5 year moving average, L = 5
  - For a 7 year moving average, L = 7
  - Etc.

Moving Averages

- **Example:** Five-year moving average
  - First average:

$$MA(5) = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

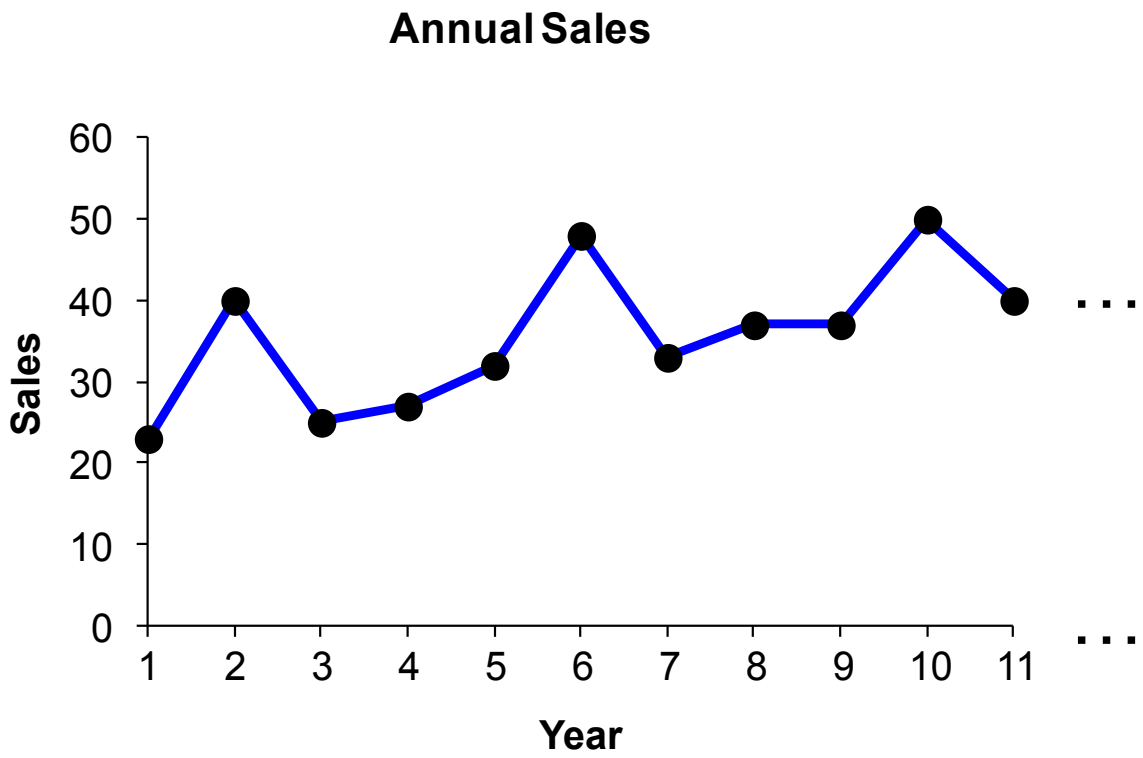
  - Second average:

$$MA(5) = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5}$$

  - etc.

Example: Annual Data

Year	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40
etc…	etc…



Calculating Moving Averages

Year	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40

Average Year	5-Year Moving Average
3	29.4
4	34.4
5	33.0
6	35.4
7	37.4
8	41.0
9	39.4
…	…

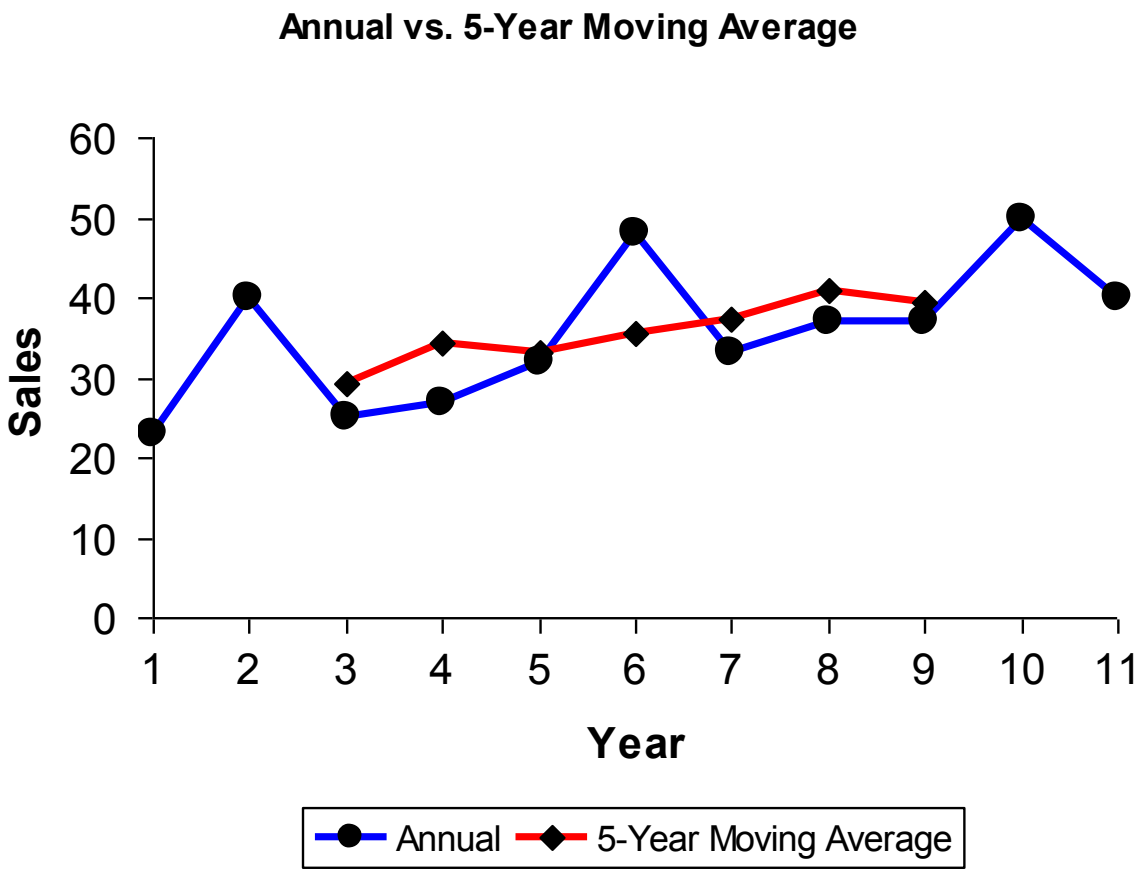
$$3 = \frac{1+2+3+4+5}{5}$$

$$29.4 = \frac{23+40+25+27+32}{5}$$

- Each moving average is for a consecutive block of 5 years

Annual vs. Moving Average

- **The 5-year moving average smoothes the data and shows the underlying trend**



Exponential Smoothing

- A weighted moving average
  - Weights decline exponentially
  - Most recent observation weighted most

- Used for smoothing and short term forecasting (often one period into the future)

Exponential Smoothing

- The weight (smoothing coefficient) is W
  - Subjectively chosen
  - Range from 0 to 1
  - Smaller W gives more smoothing, larger W gives less smoothing

Exponential Smoothing Model

$$E_1 = Y_1$$

For i = 2, 3, 4, …

$$E_i = WY_i + (1 - W)E_{i-1}$$

where:

E_i = exponentially smoothed value for period i

E_{i-1} = exponentially smoothed value already computed for period i - 1

Y_i = observed value in period i

W = weight (smoothing coefficient), $0 < W < 1$

Exponential Smoothing Example

- Suppose we use weight W = .2

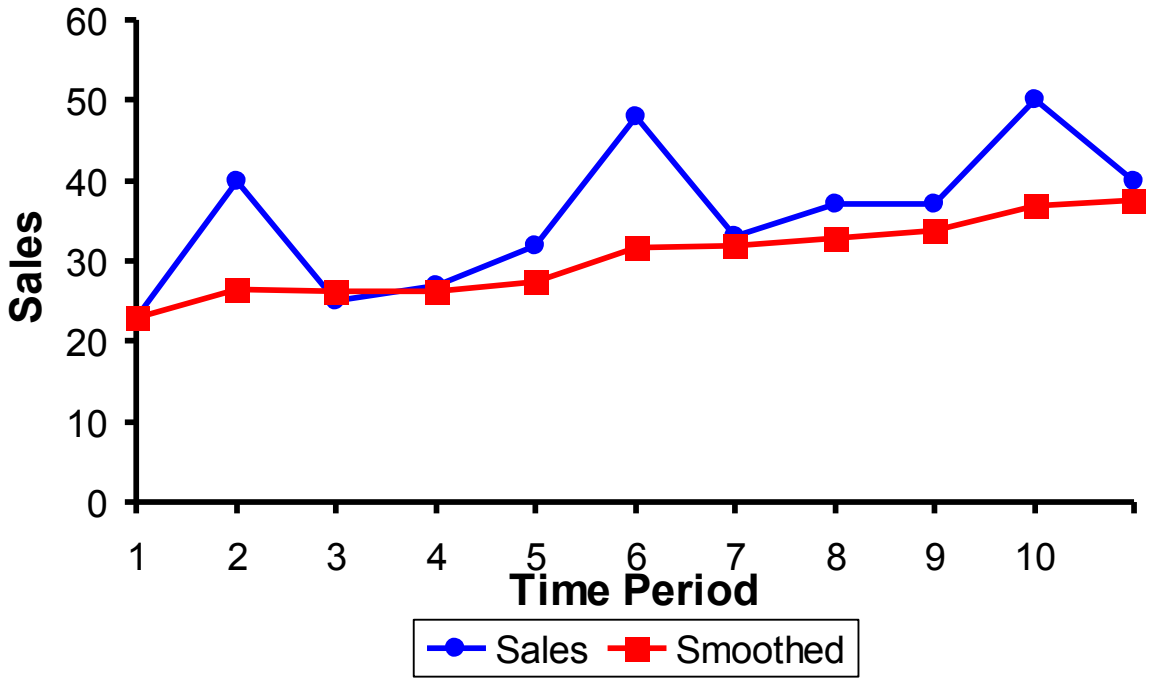
Time Period (i)	Sales (Y_i)	Forecast from prior period (E_{i-1})	Exponentially Smoothed Value for this period (E_i)
1	23	--	23
2	40	23	(.2)(40)+(.8)(23)=26.4
3	25	26.4	(.2)(25)+(.8)(26.4)=26.12
4	27	26.12	(.2)(27)+(.8)(26.12)=26.296
5	32	26.296	(.2)(32)+(.8)(26.296)=27.437
6	48	27.437	(.2)(48)+(.8)(27.437)=31.549
7	33	31.549	(.2)(48)+(.8)(31.549)=31.840
8	37	31.840	(.2)(33)+(.8)(31.840)=32.872
9	37	32.872	(.2)(37)+(.8)(32.872)=33.697
10	50	33.697	(.2)(50)+(.8)(33.697)=36.958
etc.	etc.	etc.	etc.

$E_1 = Y_1$ since no prior information exists

$$E_i = WY_i + (1-W)E_{i-1}$$

Sales vs. Smoothed Sales

- Fluctuations have been smoothed
- NOTE: the smoothed value in this case is generally a little low, since the trend is upward sloping and the weighting factor is only .2



Forecasting Time Period i + 1

- **The smoothed value in the current period (i) is used as the forecast value for next period (i + 1) :**

$$\hat{Y}_{i+1} = E_i$$

Trend-Based Forecasting

- Estimate a trend line using regression analysis

Year	Time Period (X)	Sales (Y)
1999	0	20
2000	1	40
2001	2	30
2002	3	50
2003	4	70
2004	5	65

- Use time(X) as the independent variable:

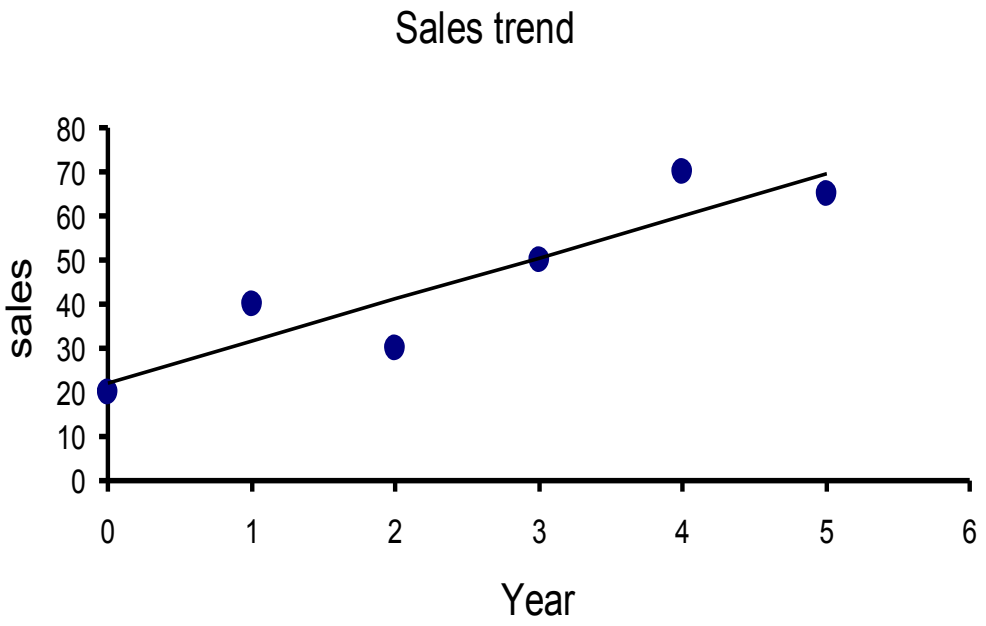
$$\hat{Y} = b_0 + b_1 X$$

Trend-Based Forecasting

Year	Time Period (X)	Sales (Y)
1999	0	20
2000	1	40
2001	2	30
2002	3	50
2003	4	70
2004	5	65

- **The linear trend forecasting equation is:**

$$\hat{Y}_i = 21.905 + 9.5714\,X_i$$



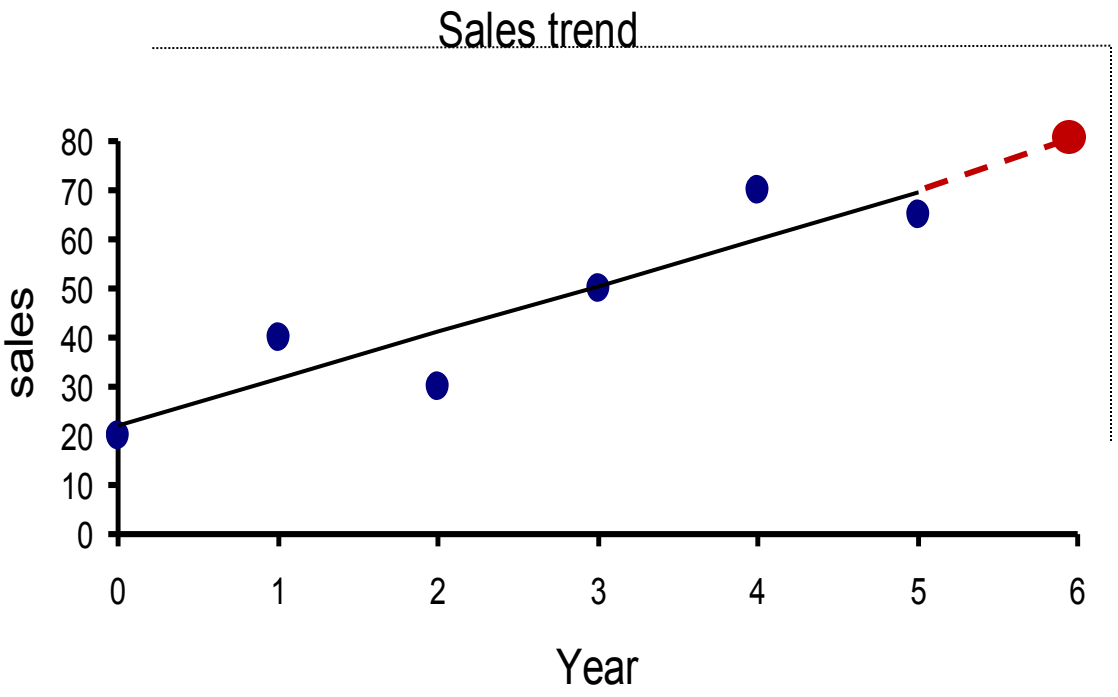
Trend-Based Forecasting

Year	Time Period (X)	Sales (Y)
1999	0	20
2000	1	40
2001	2	30
2002	3	50
2003	4	70
2004	5	65
2005	6	??

- **Forecast for time period 6:**

$$\hat{Y} = 21.905 + 9.5714(6)$$

$$= 79.33$$



Nonlinear Trend Forecasting

- A nonlinear regression model can be used when the time series exhibits a nonlinear trend
- Quadratic form is one type of a nonlinear model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

- Can try other functional forms to get best fit

Model Selection Using Differences

- Use a linear trend model if the first differences are approximately constant

$$(Y_2 - Y_1) = (Y_3 - Y_2) = \cdots = (Y_n - Y_{n-1})$$

- Use a quadratic trend model if the second differences are approximately constant

$$[(Y_3 - Y_2) - (Y_2 - Y_1)] = [(Y_4 - Y_3) - (Y_3 - Y_2)]$$
$$= \cdots = [(Y_n - Y_{n-1}) - (Y_{n-1} - Y_{n-2})]$$

Autoregressive Models

- Used for forecasting
- Takes advantage of autocorrelation
  - 1st order - correlation between consecutive values
  - 2nd order - correlation between values 2 periods apart
- p^{th} order Autoregressive models:

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \cdots + + A_p Y_{i-p} + \delta_i$$

Random Error

Autoregressive Model: Example

The Office Concept Corp. has acquired a number of office units (in thousands of square feet) over the last eight years. Develop the second order Autoregressive model.

Year	Units
97	4
98	3
99	2
00	3
01	2
02	2
03	4
04	6

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \delta_i$$

Autoregressive Model: Example Solution

- Develop the 2nd order table
- Build a regression model

Year	Y_i	Y_{i-1}	Y_{i-2}
97	4	--	--
98	3	4	--
99	2	3	4
00	3	2	3
01	2	3	2
02	2	2	3
03	4	2	2
04	6	4	2

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \delta_i$$

Model Output

	Coefficients
Intercept	3.5
X Variable 1	0.8125
X Variable 2	-0.9375

$$\hat{Y}_i = 3.5 + 0.8125 Y_{i-1} - 0.9375 Y_{i-2}$$

Autoregressive Model Example: Forecasting

Use the second-order equation to forecast number of units for 2005:

$$\hat{Y}_i = 3.5 + 0.8125Y_{i-1} - 0.9375Y_{i-2}$$

$$\hat{Y}_{2005} = 3.5 + 0.8125(Y_{2004}) - 0.9375(Y_{2003})$$

$$= 3.5 + 0.8125(6) - 0.9375(4)$$

$$= 4.625$$

Autoregressive Modeling Steps

1. Choose p

2. Form a series of "lagged predictor" variables

$$Y_{i-1}, Y_{i-2}, \ldots, Y_{i-p}$$

3. Build regression model using all p variables

Measuring Errors

- Choose the model that gives the smallest measuring errors

- Sum of Squared Errors (SSE)

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - Sensitive to outliers

- Mean Absolute Deviation (MAD)

$$MAD = \frac{\sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|}{n}$$

  - Less sensitive to extreme observations

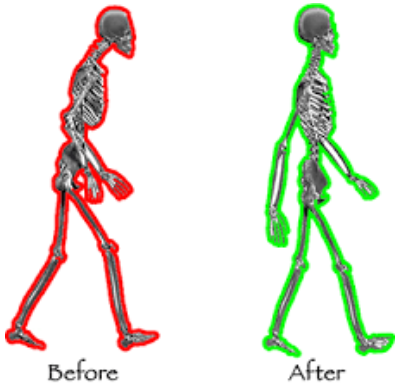
Principal of Parsimony

- Suppose two or more models provide a good fit for the data
- Select the simplest model
  - Simplest model types:
    - Least-squares linear
    - Least-squares quadratic
    - 1st order autoregressive
  - More complex types:
    - 2nd and 3rd order autoregressive
    - Least-squares exponential

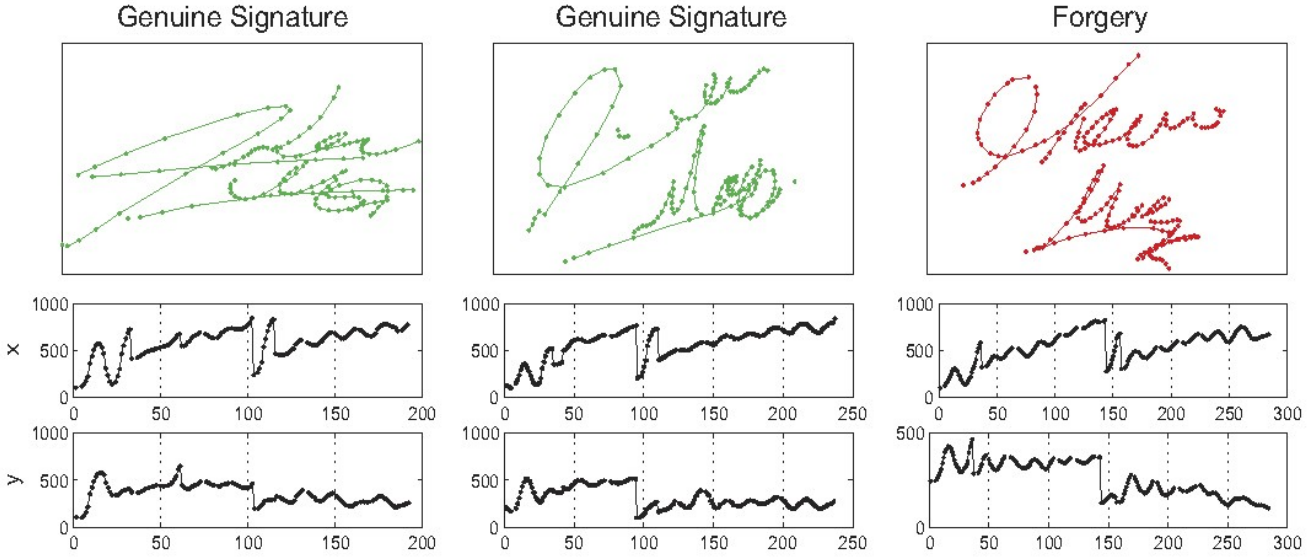
Time Series Similarity Search

Motivation

- **Identify the same person in different actions**

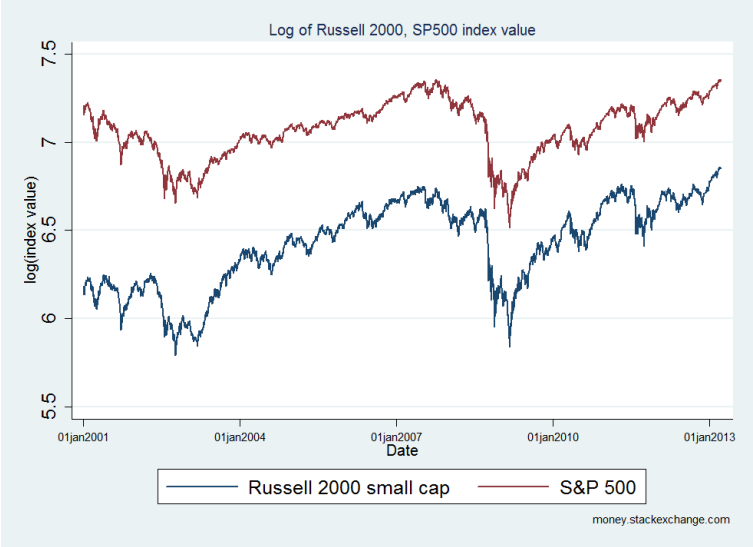


- **Detect forgery**



Motivation

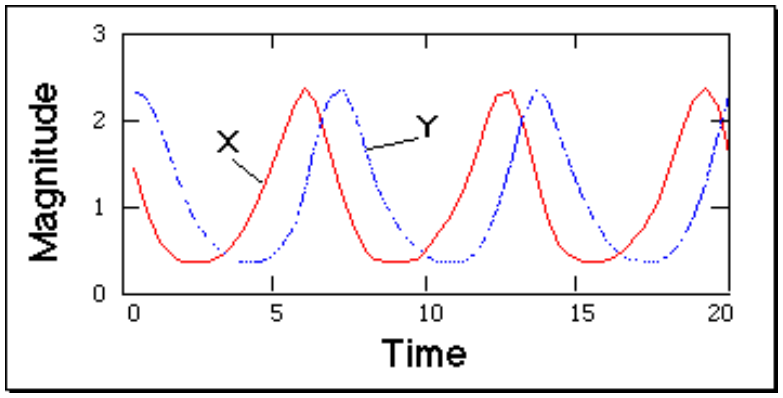
- **Find similar stocks or mutual funds**

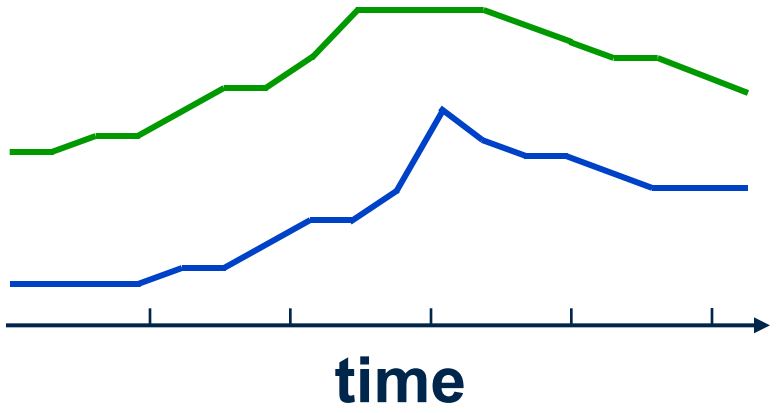


- **Find a time period with similar inflation rate and unemployment rate**

How to Measure Similarity

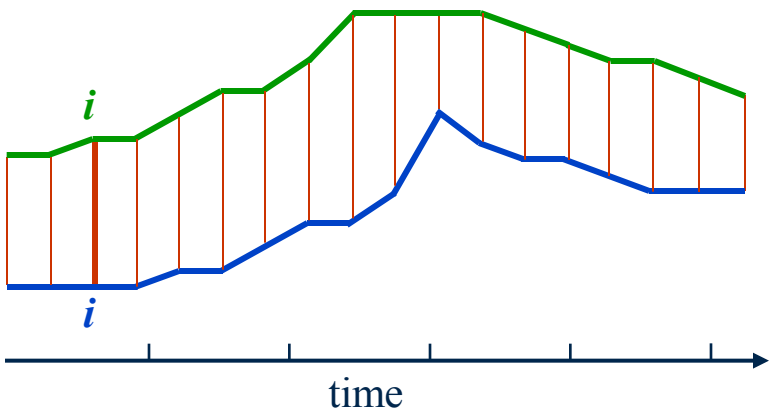
- **Euclidean? Manhattan? *Lp* norm?**

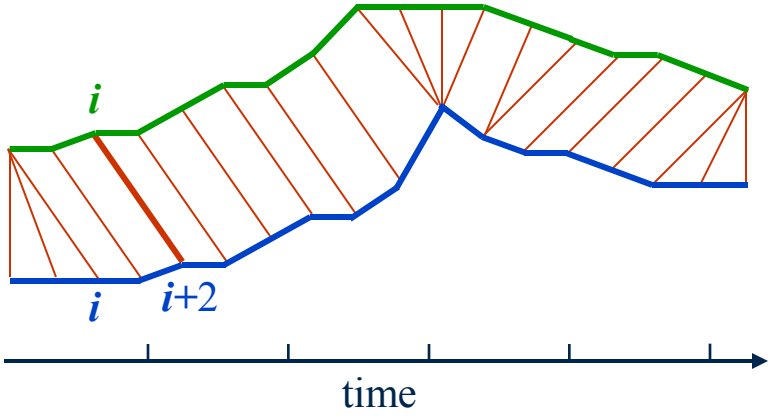




- **Need a method that allows elastic shifting of time axis to accommodate sequences that are similar but can be out of phase**

Why Dynamic Time Warping?

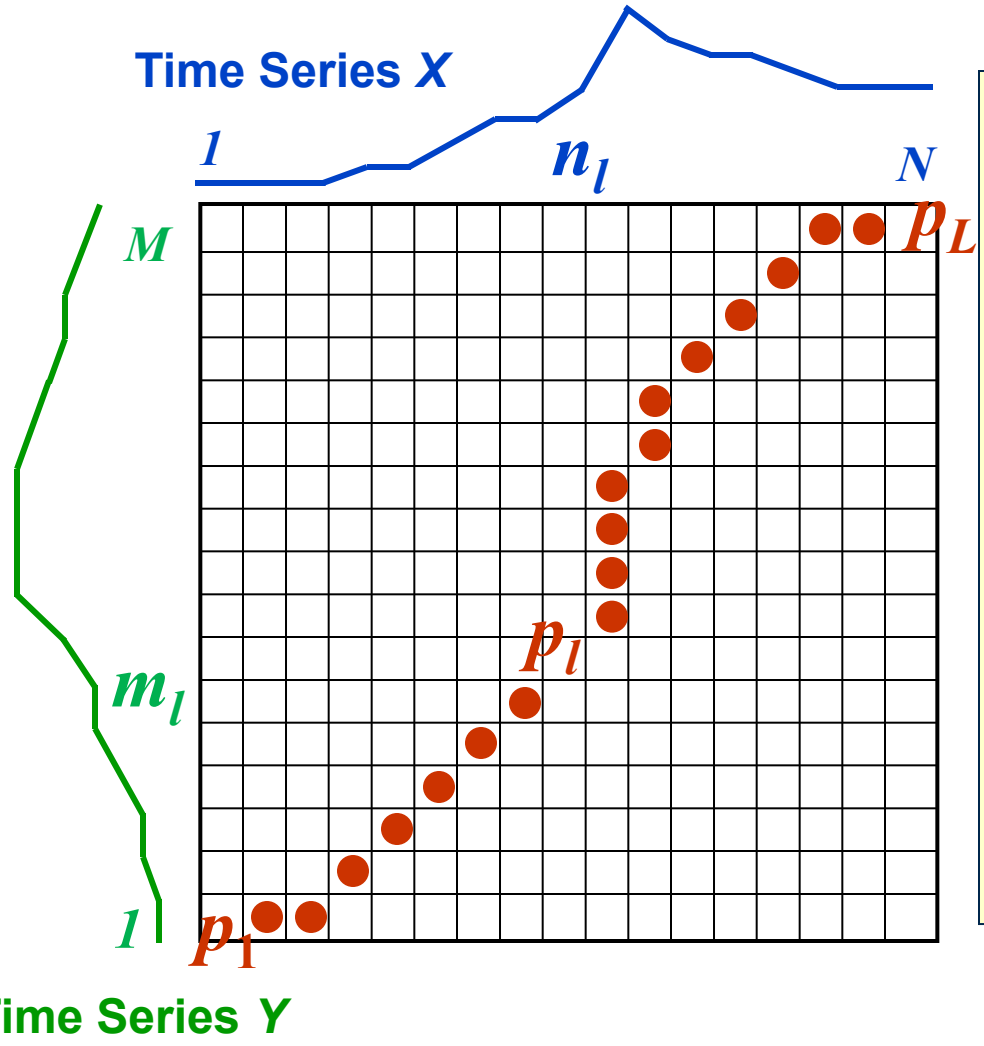




Any distance (Euclidean, Manhattan, …) which aligns the *i*-th point on one time series with the *i*-th point on the other will produce a poor similarity score.

A non-linear (elastic) alignment produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase in the time axis.

Warping Path



The *best alignment* between X and Y is the path through the grid

$$P = p_1, \dots, p_l, \dots, p_L$$

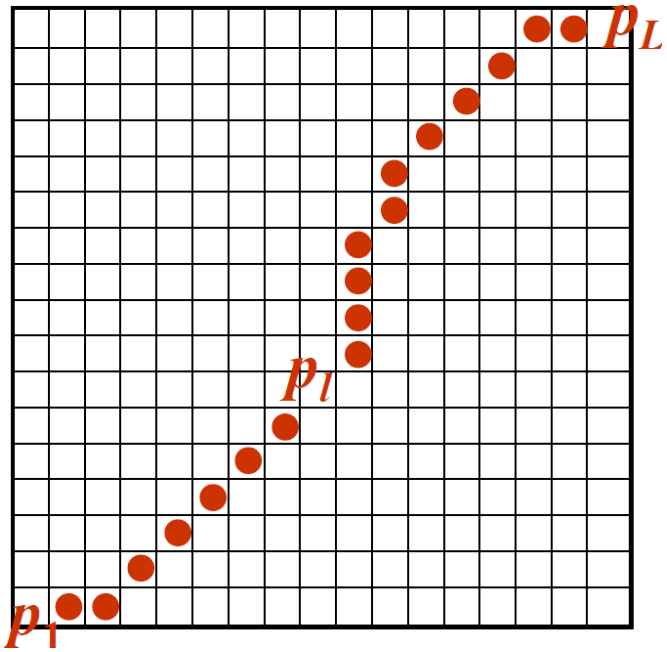
$$p_l = (n_l, m_l) \in [1:N] \times [1:M]$$

which *minimizes* the total distance between them.

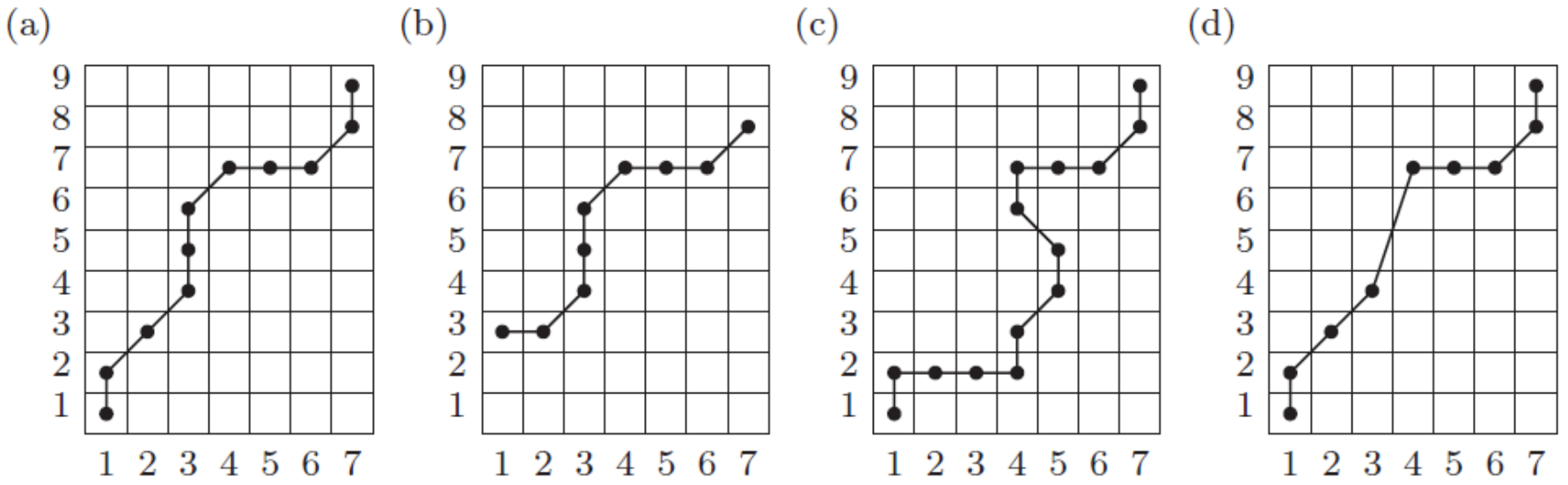
P is called a ***warping path***

Warping Path

- Given $X = \{x_1, x_2, \ldots x_N\}$ and $Y = \{y_1, y_2, \ldots y_M\}$
- An (N,M)-warping path is a sequence $P = (p_1, p_2, \ldots, p_L)$ satisfying the three conditions:
  - Boundary condition:
    - starts with (1,1), ends with (*N, M*)
  - Monotonicity condition:
    - never goes back in time
  - Step size condition:
    - one step at a time



Which one is a warping path?



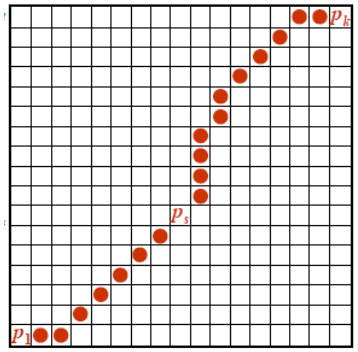
Cost Matrix

- Given $X = \{x_1, x_2, \ldots x_N\}$, $Y = \{y_1, y_2, \ldots y_M\}$ and a local distance (cost) measure

$$c(x_n, y_m) \quad n \in [1:N], \quad m \in [1:M]$$

- A *cost matrix* $C \in R^{N \times M}$ is defined as:

$$C(n, m) = c(x_n, y_m)$$



- DTW algorithm finds a *warping path* such that the overall cost is minimized

Cost Matrix Example

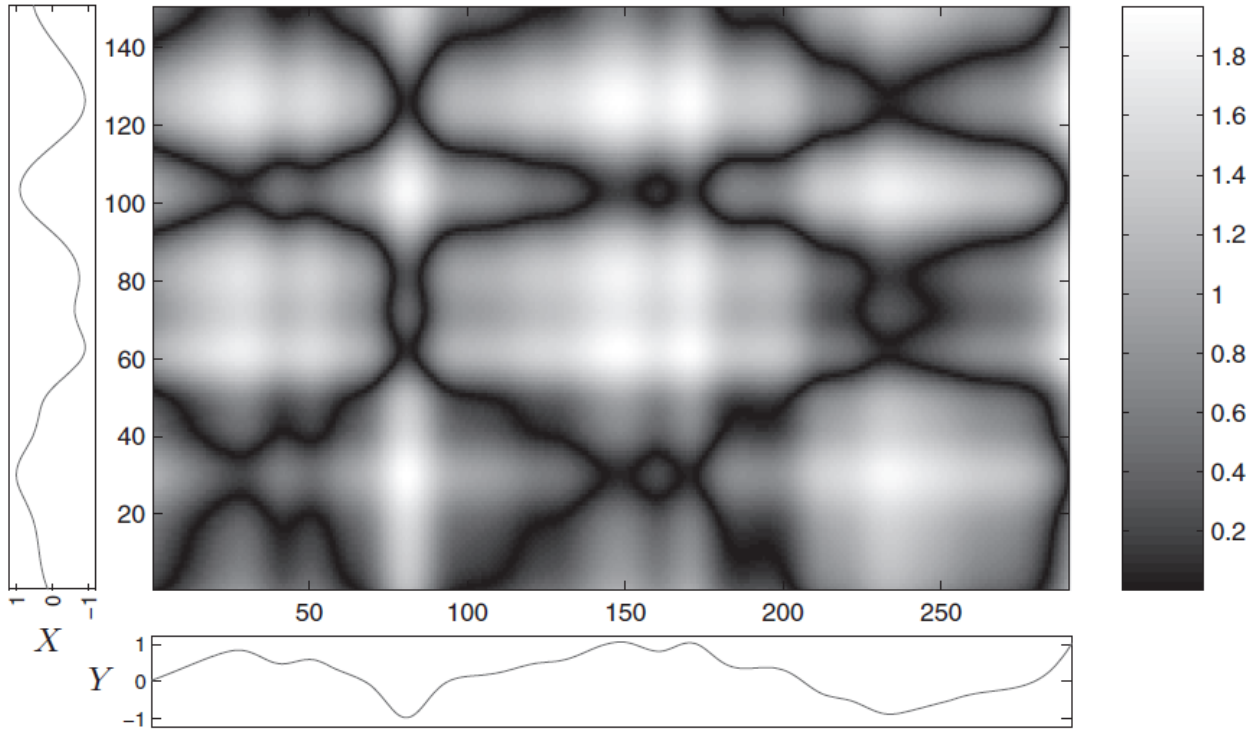
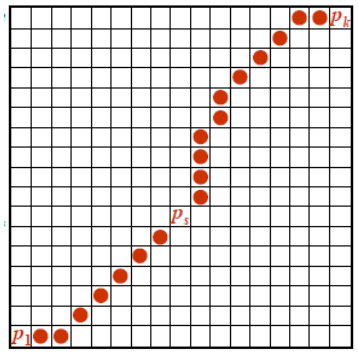


Fig. 4.2. Cost matrix of the two real-valued sequences X (*vertical axis*) and Y (*horizontal axis*) using the Manhattan distance (absolute value of the difference) as local cost measure c. Regions of low cost are indicated by *dark colors* and regions of high cost are indicated by *light colors*

DTW Distance

- Total cost of a warping path p

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



- Total cost of the optimal warping path p^*

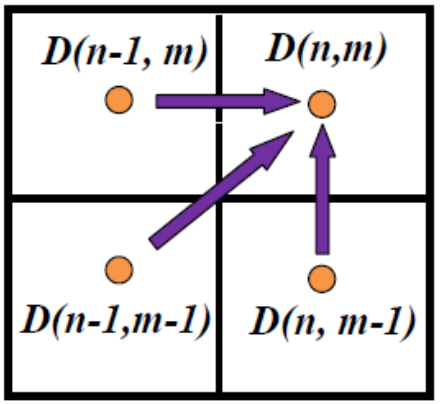
$$c_{p^*}(X,Y) = min\{c_p(X,Y) \mid p \; is \; a \; warping \; path\}$$

- The ***DTW distance*** btw *X* and *Y* is defined as

$$DTW(X,Y) = c_{p^*}(X,Y)$$

Dynamic Programming

- Let $D(n, m)$ denote the DTW distance btw prefix sequences $\{x_1, \ldots x_n\}$ and $\{y_1, \ldots y_n\}$
- D is called the *accumulated cost matrix*
  - Obviously, $D(N, M) = DTW(X, Y)$
- D satisfies the following identities



$$D(n, 1) = \sum_{k=1}^{n} c(x_k, y_1) \quad \forall\, n \in [1{:}\, N]$$

$$D(1, m) = \sum_{k=1}^{m} c(x_1, y_k) \quad \forall\, m \in [1{:}\, M]$$

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

$$\forall\, 1 < n \leq N \; and \; 1 < m \leq M$$

Classical DTW Algorithm

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,

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

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

Classical DTW Example

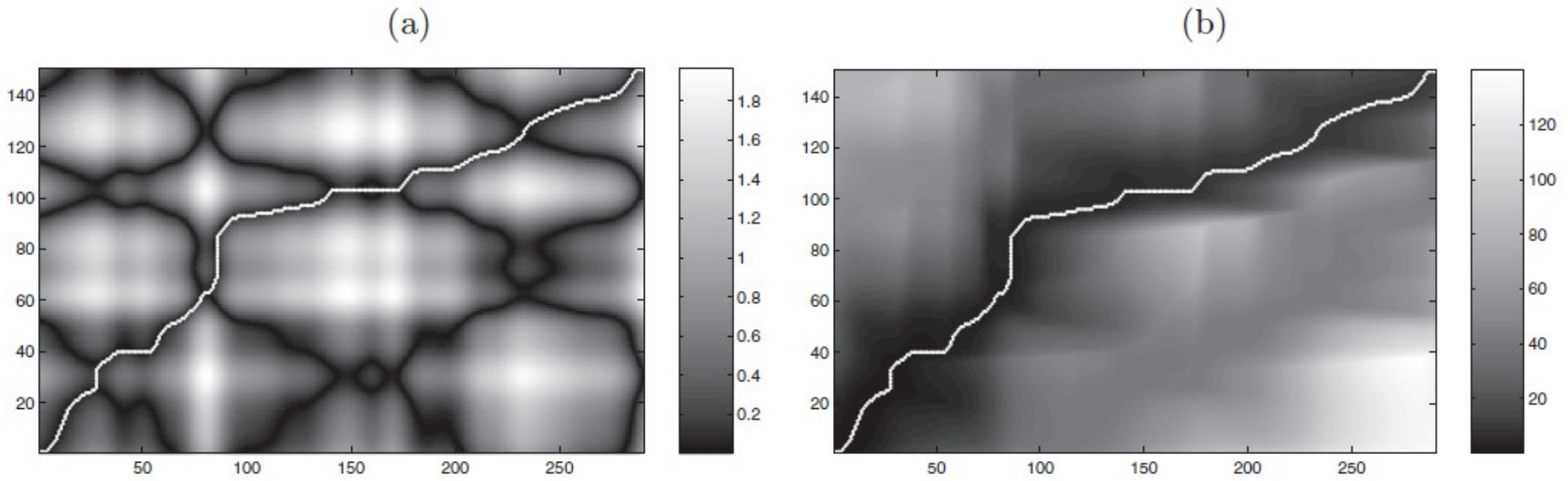


Fig. 4.4. **(a)** Cost matrix C as in Fig. 4.2 and **(b)** accumulated cost matrix D with optimal warping path p^* (*white line*)

Illustrating Example

$$X = \{1, 1, 2, 3, 5, 6\} \qquad Y = \{1, 2, 2, 4, 5\}$$

5	27	27	13	5	**1**	**2**
4	11	11	4	**1**	2	6
2	2	2	**0**	1	10	26
2	1	1	**0**	1	10	26
1	**0**	**0**	1	5	21	46
	1	1	2	3	5	6

DTW(X,Y) = 2

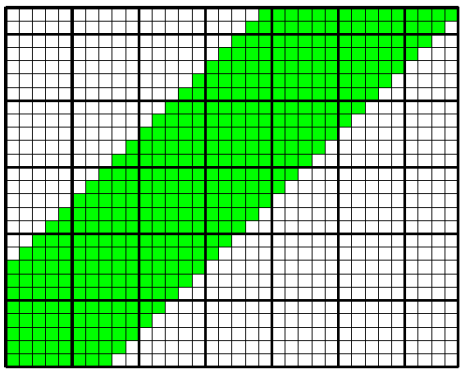
DTW Variations

Interest read; not required for this course

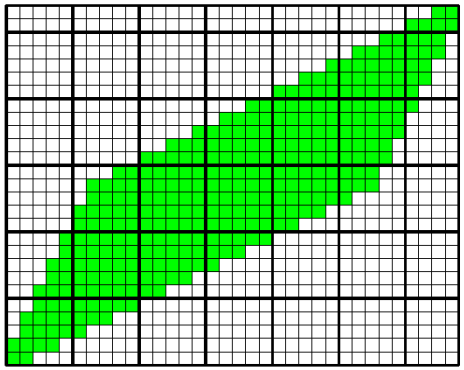
- **Local weights** $(w_d, w_h, w_v) \in \mathbb{R}^3$

$$D(n,m) = \min \begin{cases} D(n-1,m-1) + w_d \cdot c(x_n, y_m) \\ D(n-1,m) + w_h \cdot c(x_n, y_m) \\ D(n,m-1) + w_v \cdot c(x_n, y_m) \end{cases}$$

- **Global constraints**



- Sakoe-Chiba Band



- Itakura Parallelogram