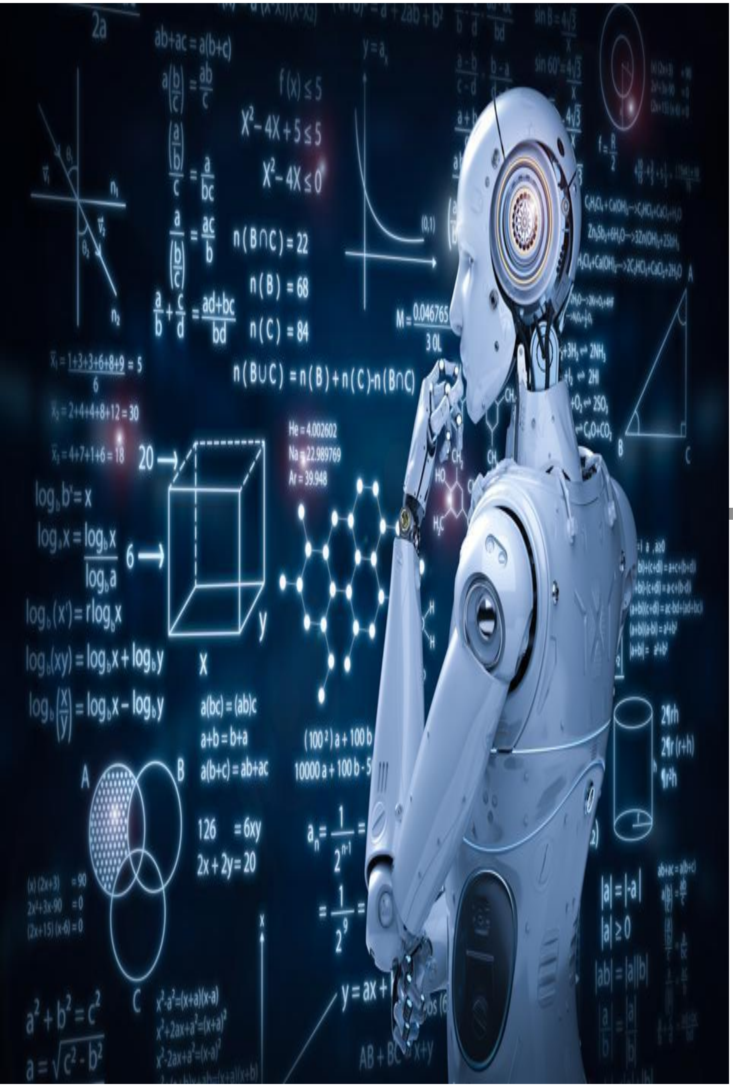


Practical Machine Learning

Day 16: Sep22 DBDA

Kiran Waghmare

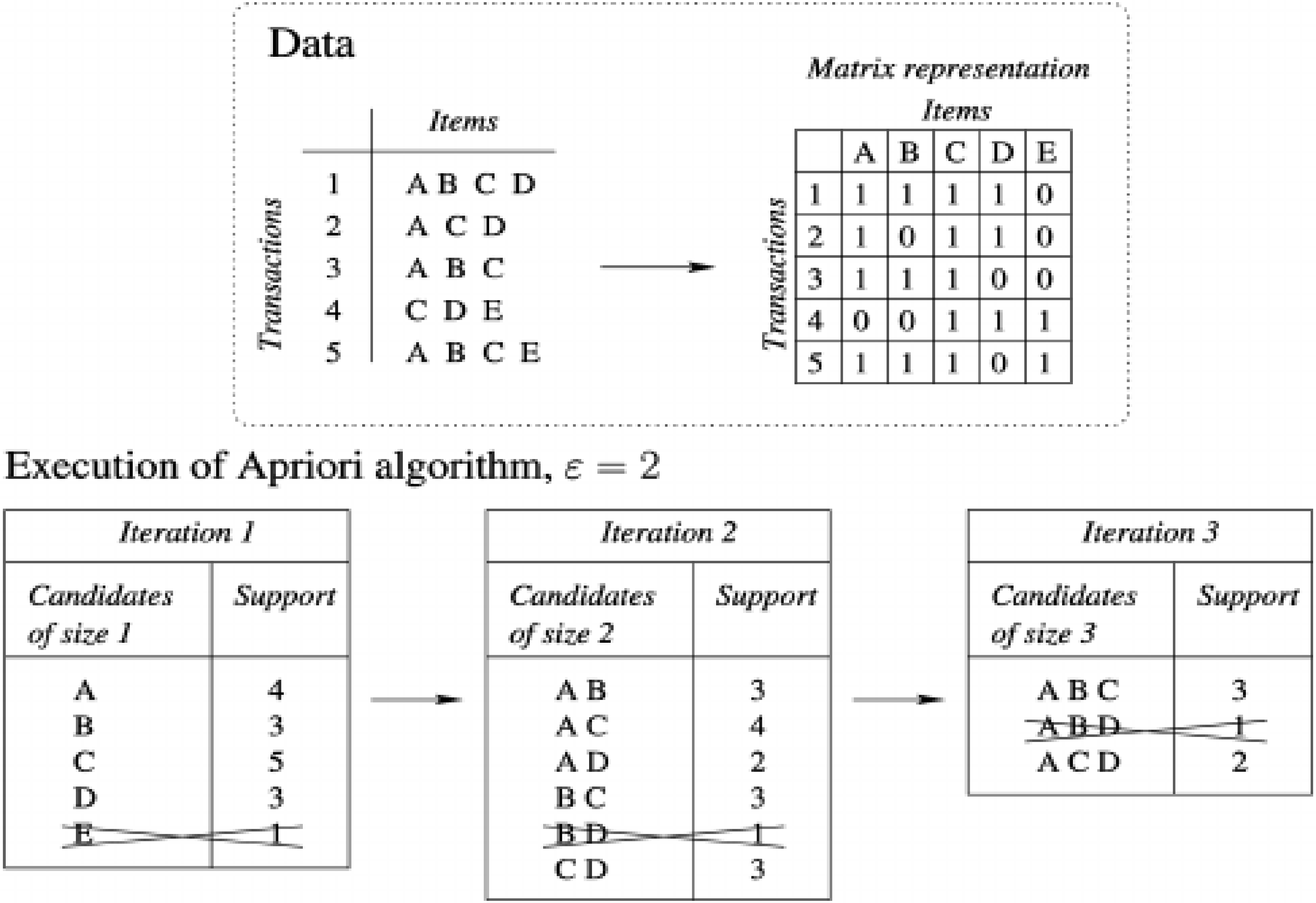
Agenda

- Association
  - Apriori
  - Market Basket Analysis

`Basket data’

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

ID	apples,	beer,	cheese,	dates,	eggs,	fish,	glue,	honey,	ice-cream
1	1	1		1			1	1	
2			1	1	1				
3		1	1			1			
4		1				1			1
5					1		1		
6						1			1
7	1			1				1	
8						1			1
9			1		1				
10		1					1		
11					1		1		
12	1								
13			1			1			
14			1			1			
15								1	1
16				1					
17	1					1			
18	1	1	1	1				1	
19	1	1		1			1	1	
20					1				



Data

Transactions	Items
1	A B C D
2	A C D
3	A B C
4	C D E
5	A B C E

Matrix representation

Transactions	A	B	C	D	E
1	1	1	1	1	0
2	1	0	1	1	0
3	1	1	1	0	0
4	0	0	1	1	1
5	1	1	1	0	1

Execution of Apriori algorithm, $\varepsilon = 2$

Iteration 1

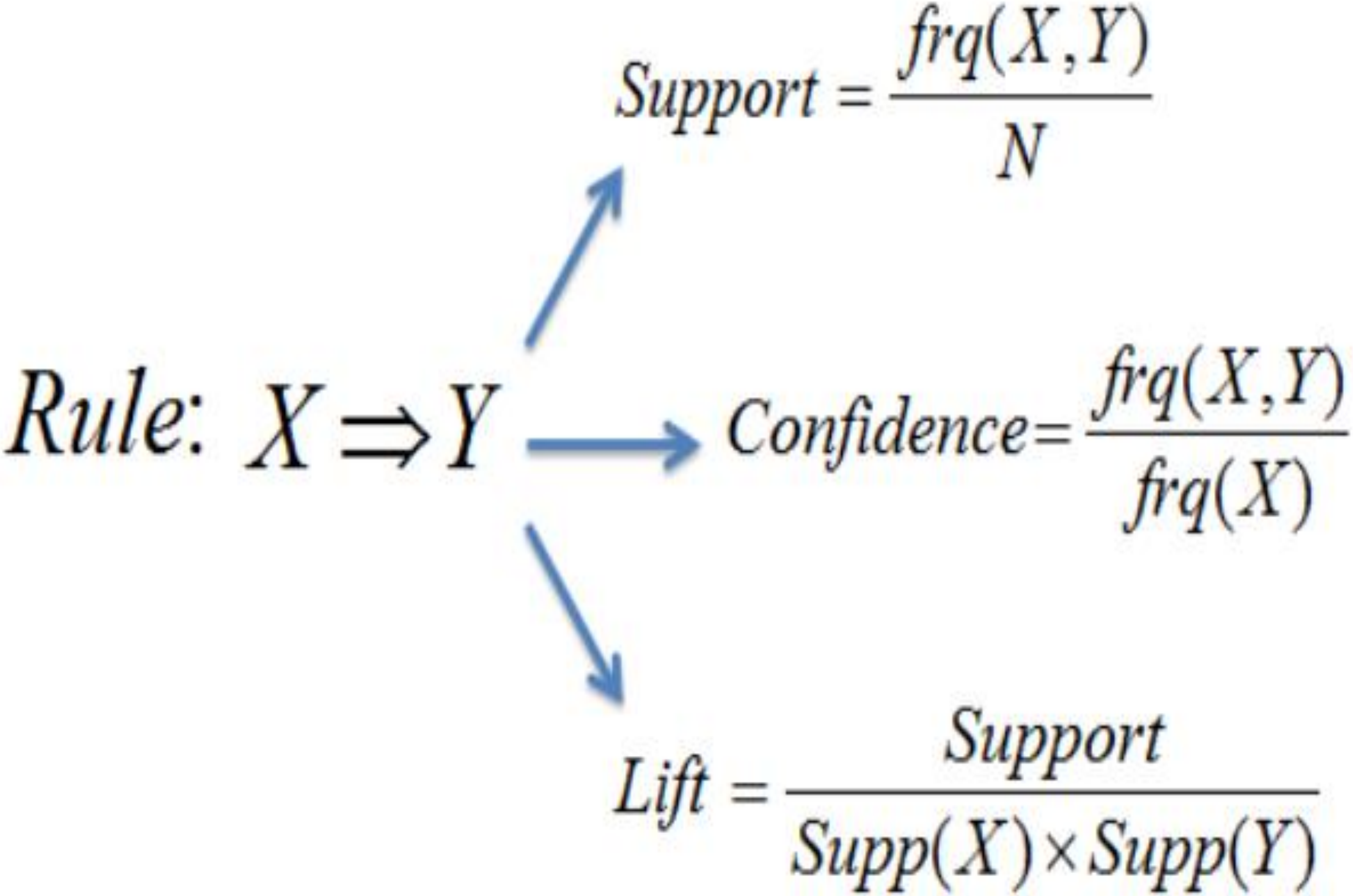
Candidates of size 1	*Support*
A	4
B	3
C	5
D	3
~~E~~	~~1~~

Iteration 2

Candidates of size 2	*Support*
A B	3
A C	4
A D	2
B C	3
~~B D~~	~~1~~
C D	3

Iteration 3

Candidates of size 3	*Support*
A B C	3
~~A B D~~	~~1~~
A C D	2



$$Rule: \; X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Discovering Rules

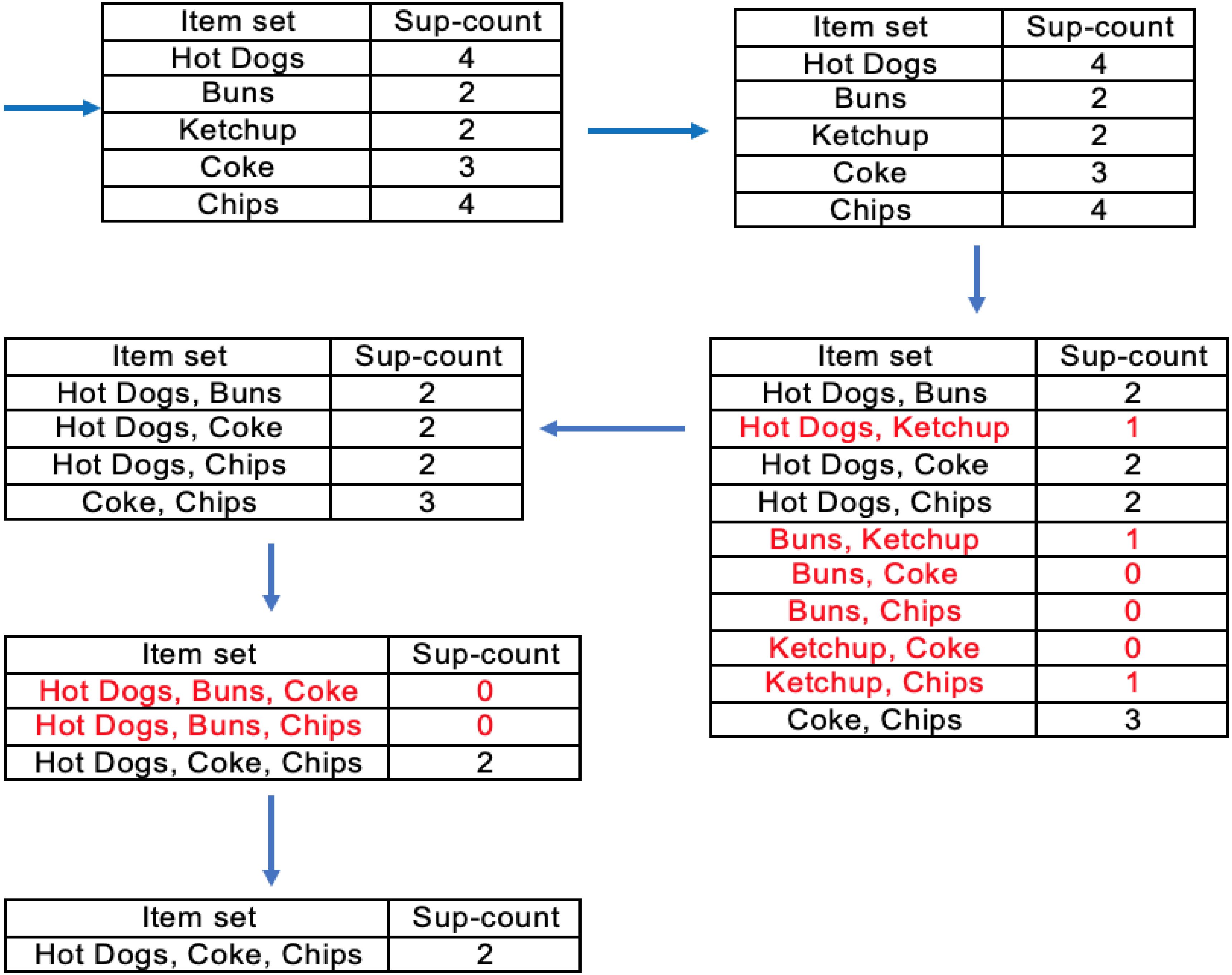
A common and useful application of data mining

A `rule' is something like this:

If a basket contains apples and cheese, then it also contains beer

Any such rule has two associated measures:

1. *confidence* – when the `if' part is true, how often is the `then' bit true? This is the same as *accuracy*.
2. *coverage* or *support* – how much of the database contains the `if' part?



Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

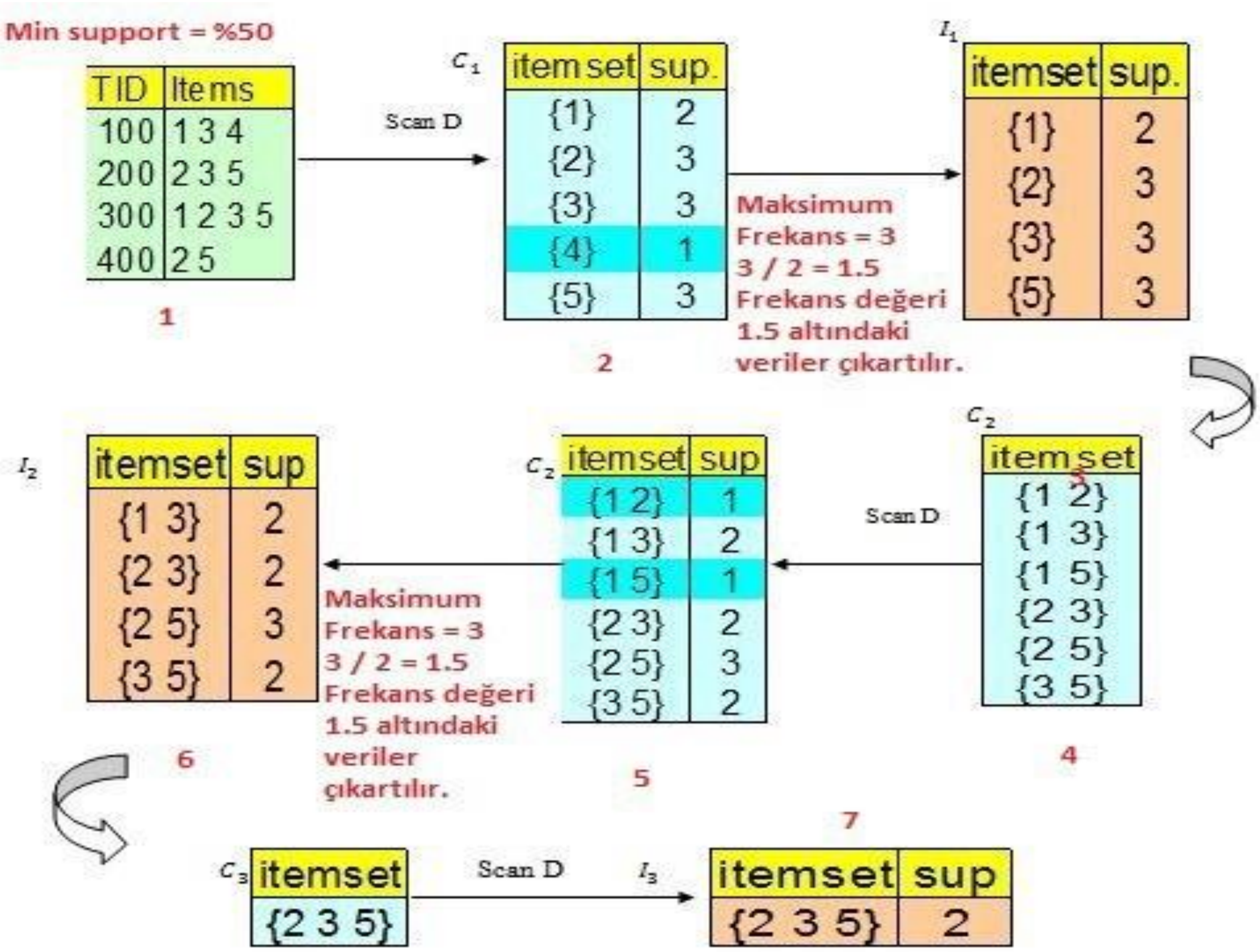
Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Ketchup	1
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Buns, Ketchup	1
Buns, Coke	0
Buns, Chips	0
Ketchup, Coke	0
Ketchup, Chips	1
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns, Coke	0
Hot Dogs, Buns, Chips	0
Hot Dogs, Coke, Chips	2

Item set	Sup-count
Hot Dogs, Coke, Chips	2



Min support = %50

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

1

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

2

Maksimum Frekans = 3
3 / 2 = 1.5
Frekans değeri 1.5 altındaki veriler çıkartılır.

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

3

4

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

5

Maksimum Frekans = 3
3 / 2 = 1.5
Frekans değeri 1.5 altındaki veriler çıkartılır.

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

6

C_3

itemset
{2 3 5}

Scan D

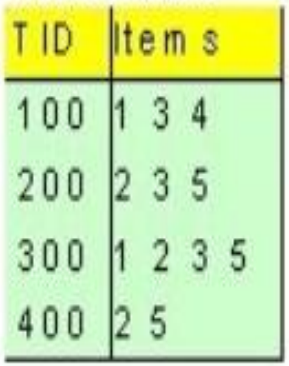
7

L_3

itemset	sup
{2 3 5}	2

The Apriori Algorithm -- Example

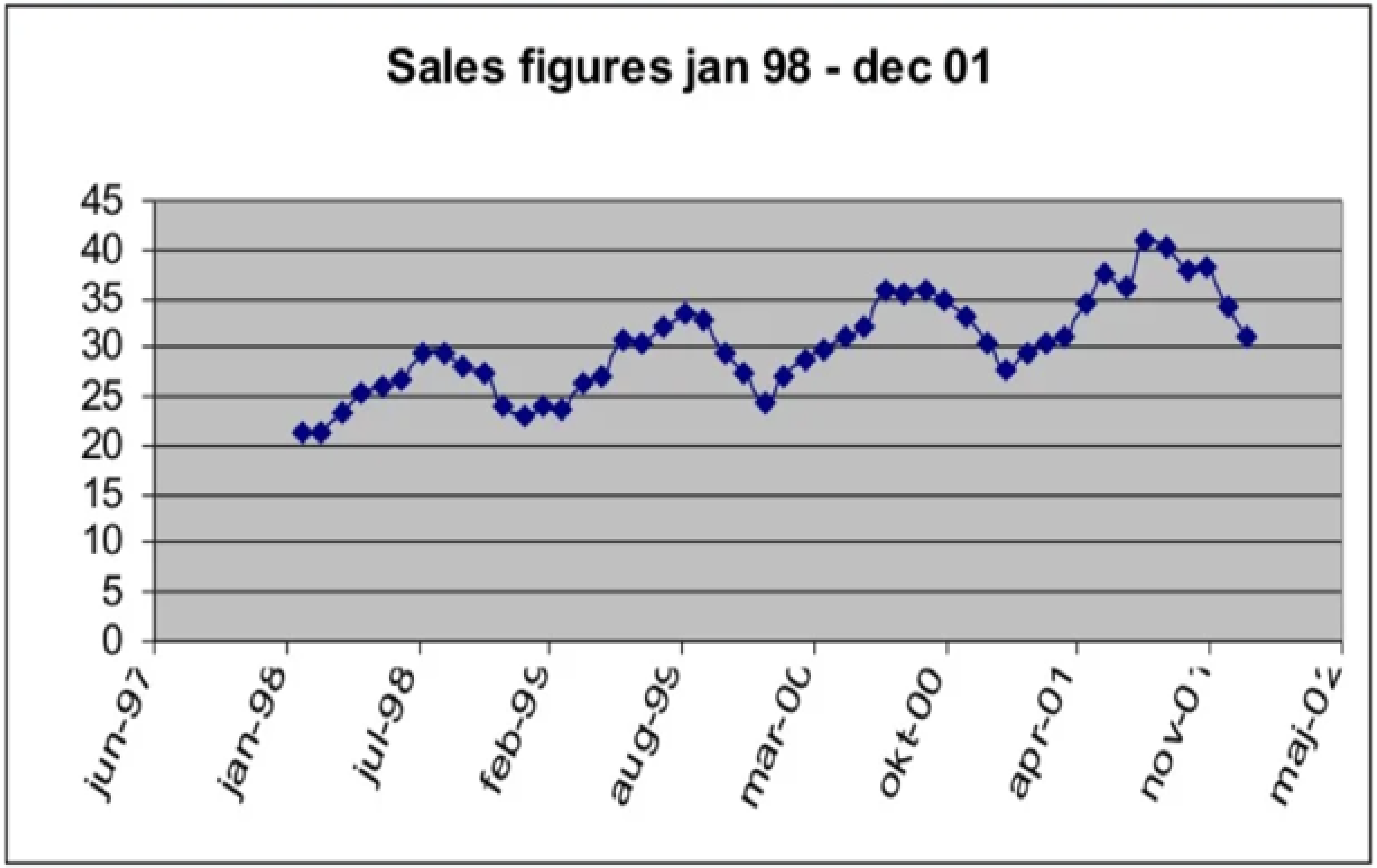
Database D



TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Time Series Analysis

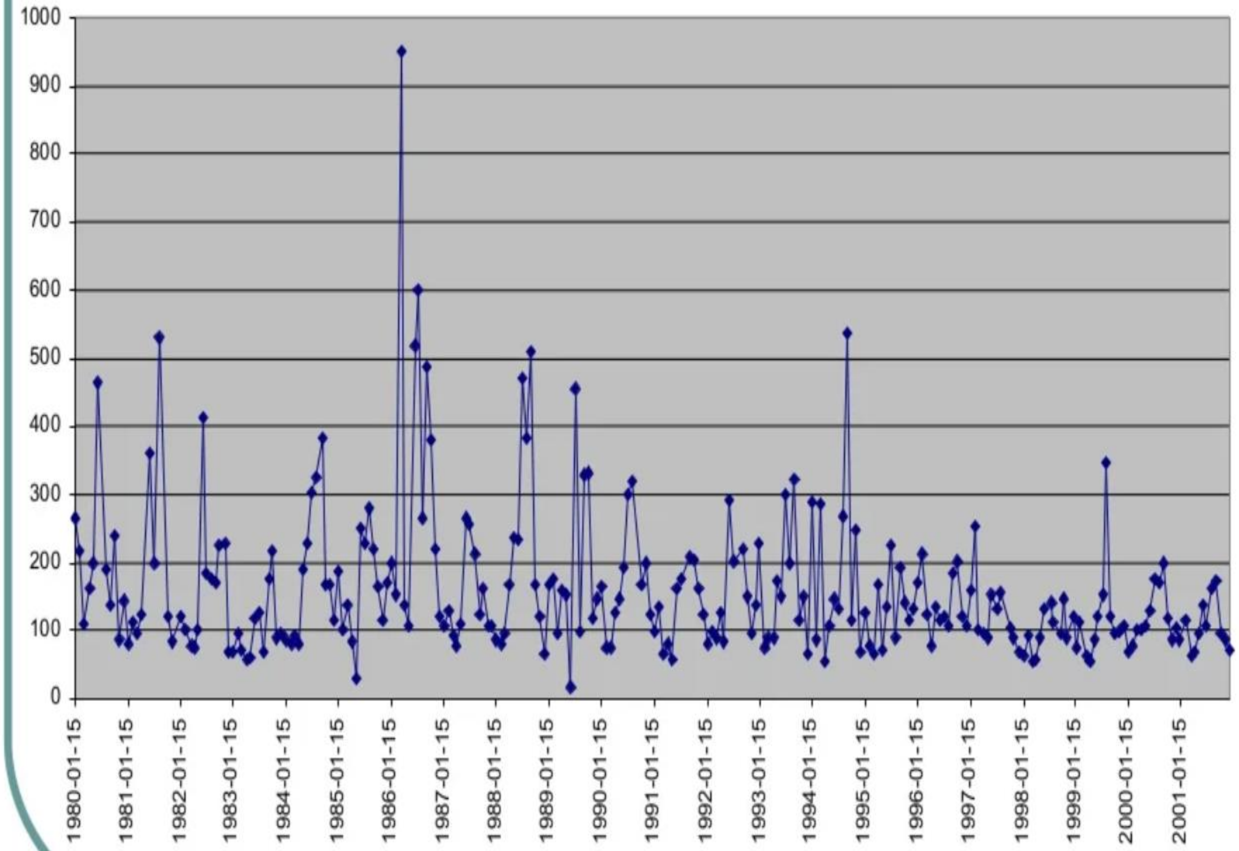
- What is Time Series Analysis?
  - The analysis of data organized across units of time.
- Time series is a basic research design
  - Data for one or more variables is collected for many observations at different time periods
  - Usually regularly spaced
  - May be either
    - univariate - one variable description
    - multivariate - causal explanation



A study on random sample of 4000 graphics from 15 of the world's news papers published between 1974 and 1989 found that more than 75% of all graphics were time series.

Cont…

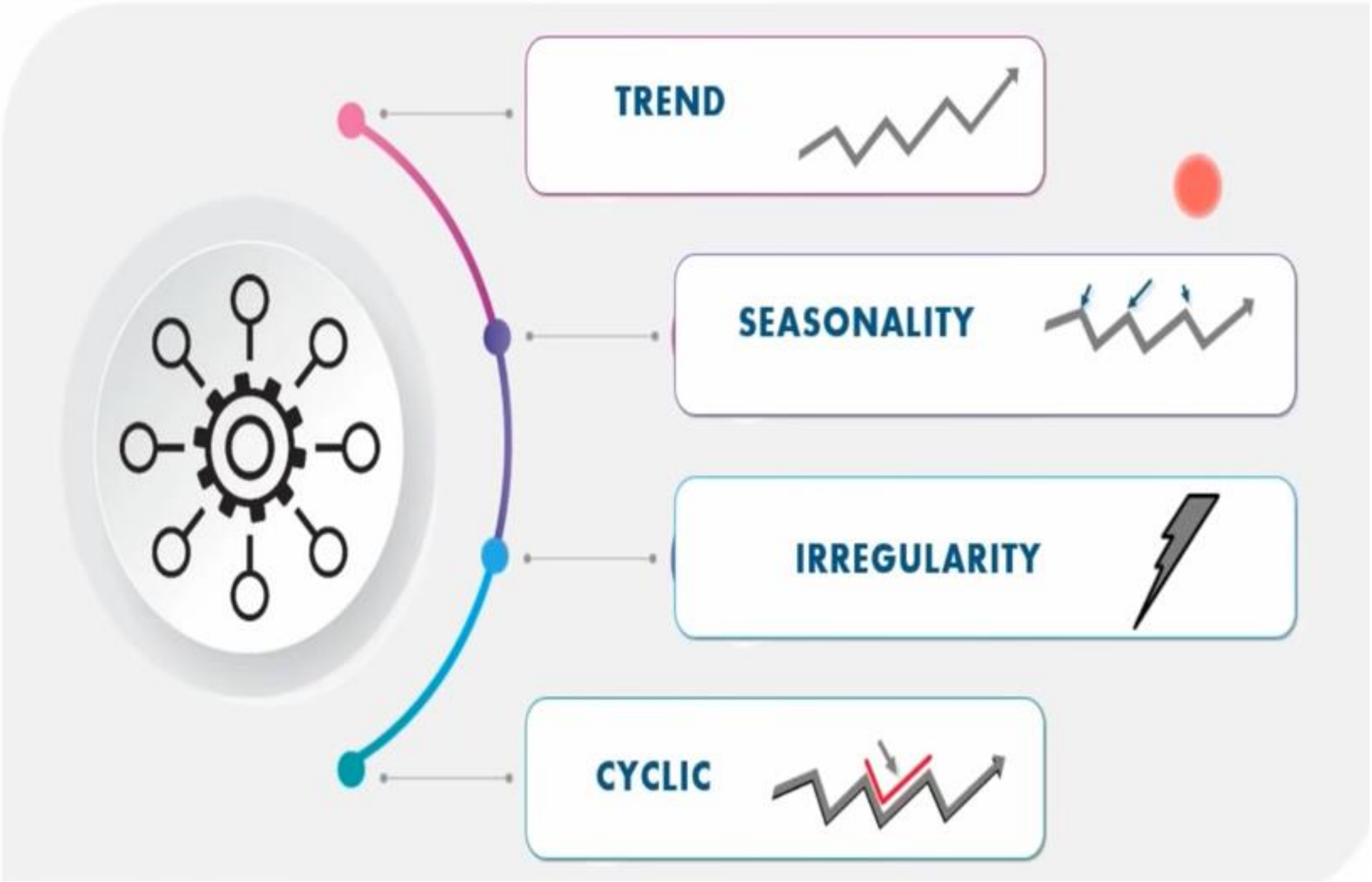
Tot-P ug/l, Rååan, Helsingborg 1980-2001



Time series examples

- Sales data
- Gross national product
- Share prices
- $A Exchange rate
- Unemployment rates
- Population
- Foreign debt
- Interest rates

Components Of Time Series



TREND

SEASONALITY

IRREGULARITY

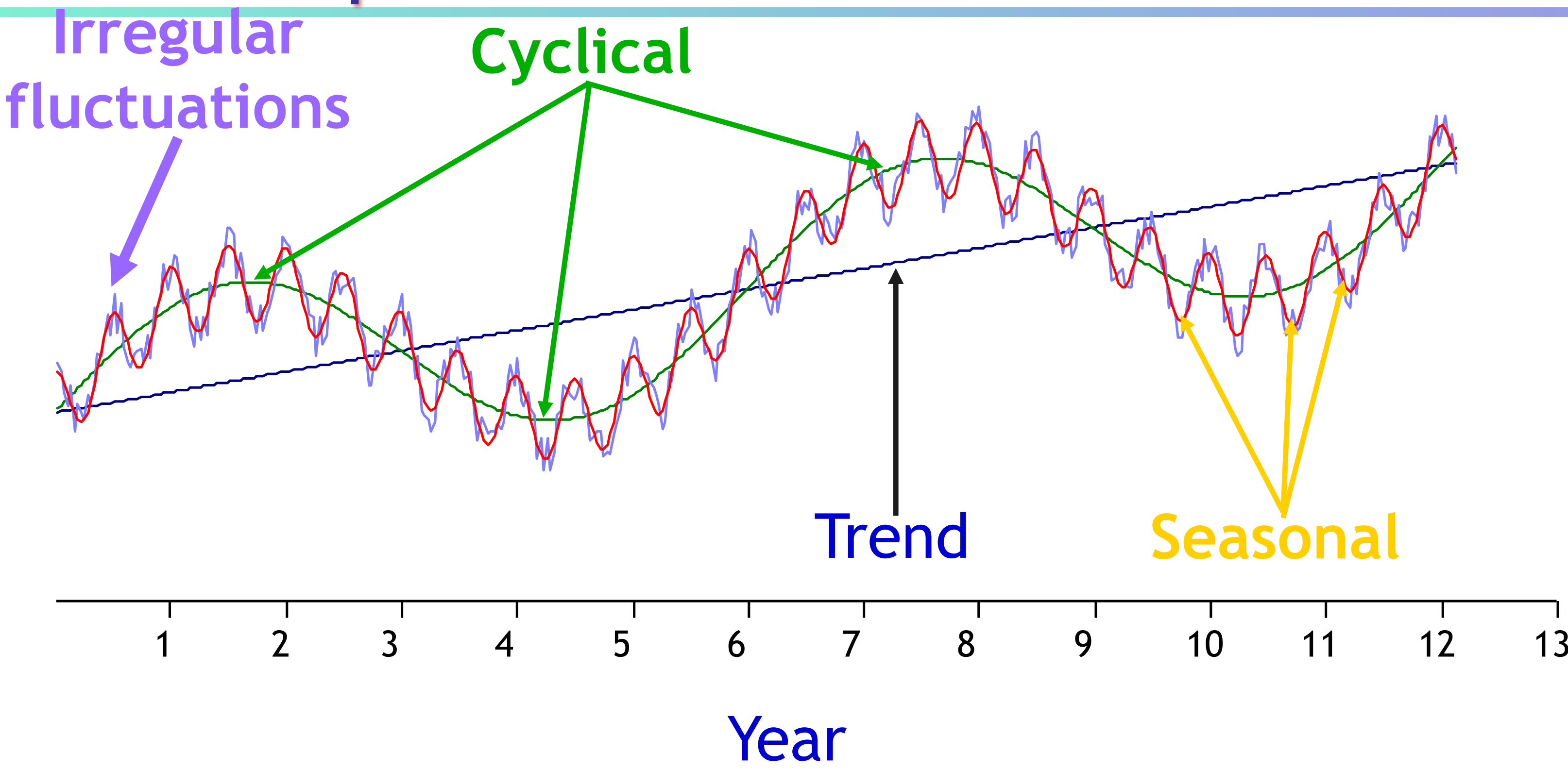
CYCLIC

Time series components

Time series data can be broken into these four components:

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular variation

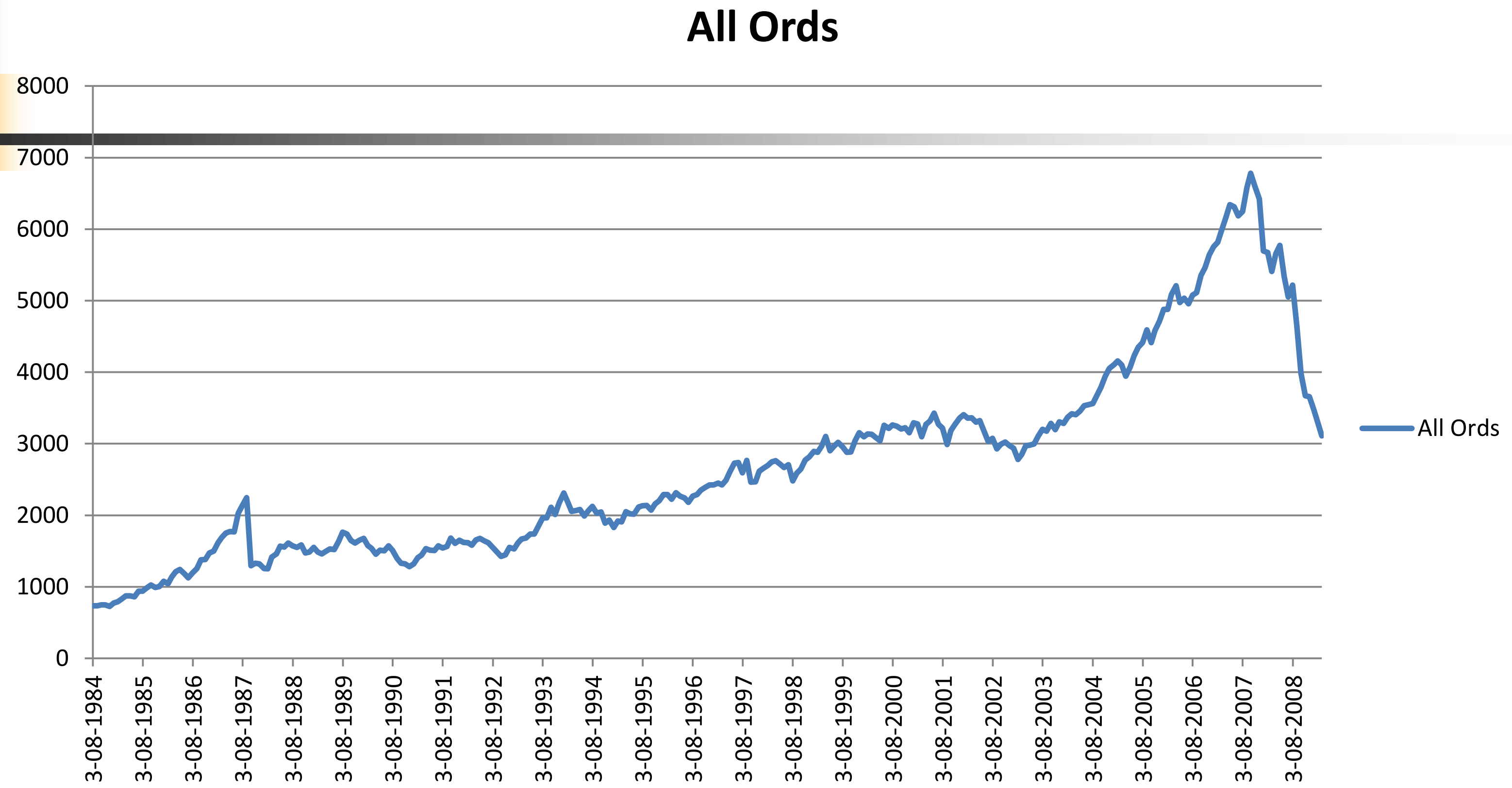
Components of Time-Series Data



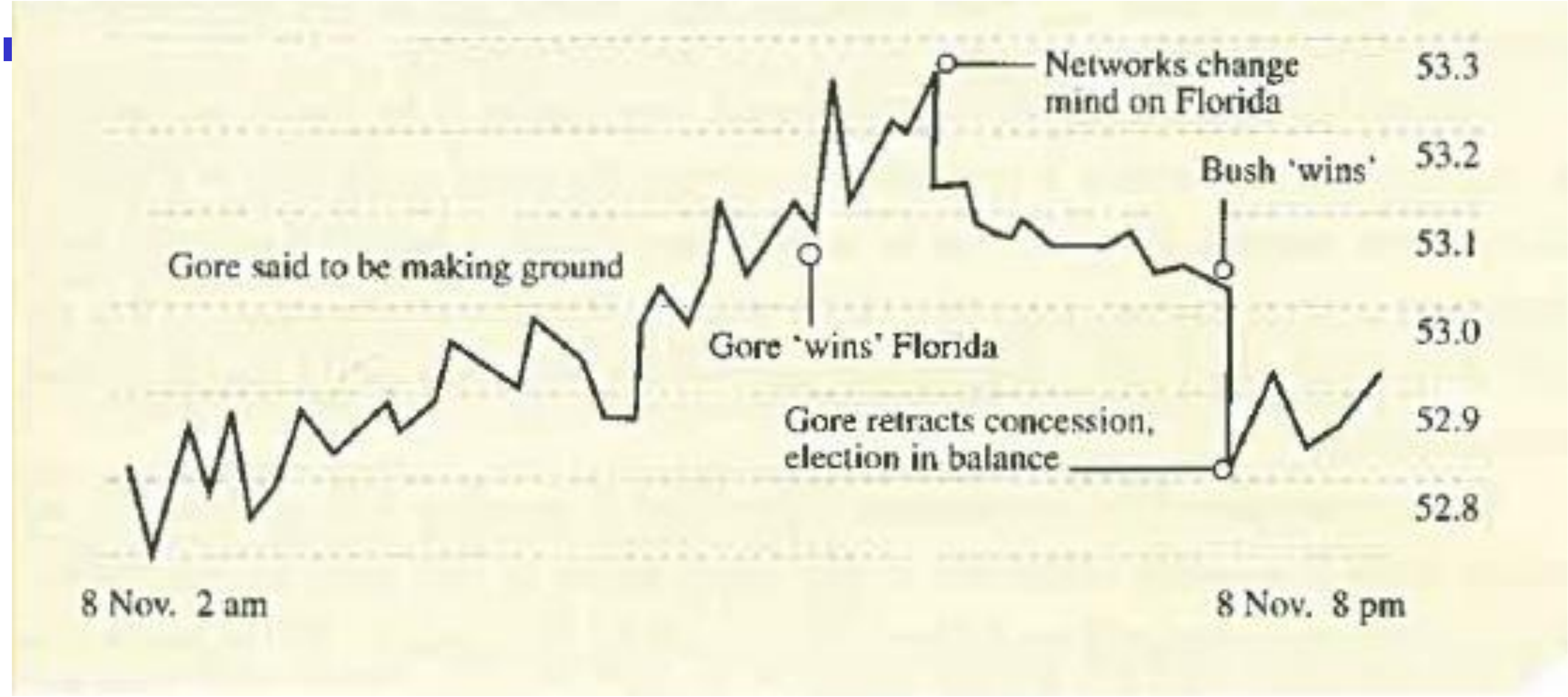
Predicting long term trends without smoothing?
What could go wrong?
Where do you commence your prediction from the bottom of a variation going up or the peak of a variation going down……….

Secular Trend

A secular trend identifies the underlying trend (direction) of the data: – increasing, decreasing or remaining constant. It is the long term direction of the data, usually described by the "line of best fit". And is deduced over a large number of periods. The following chart is a long term graph of the ASX200.

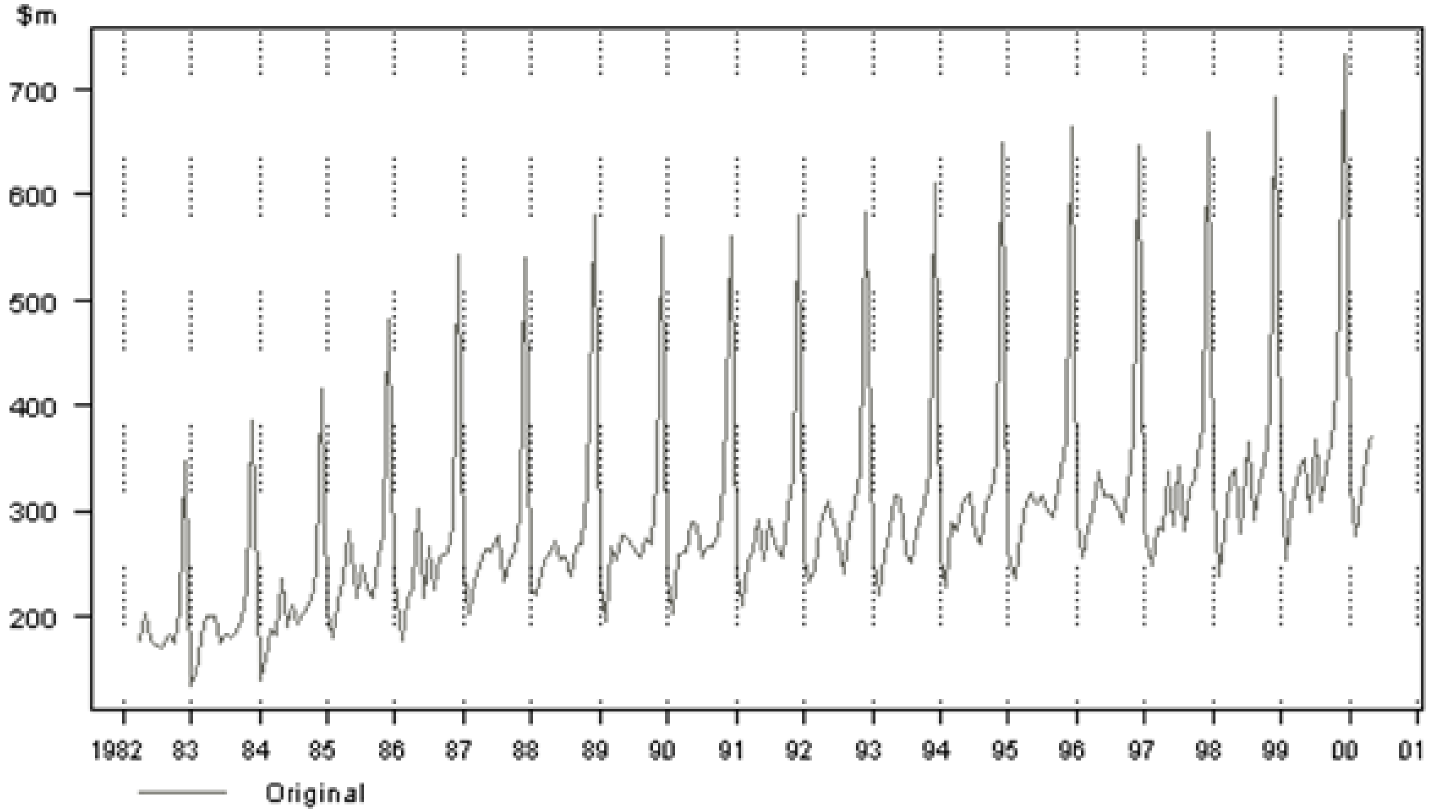


$A vs $US
during day 1 vote count 2000 US Presidential election

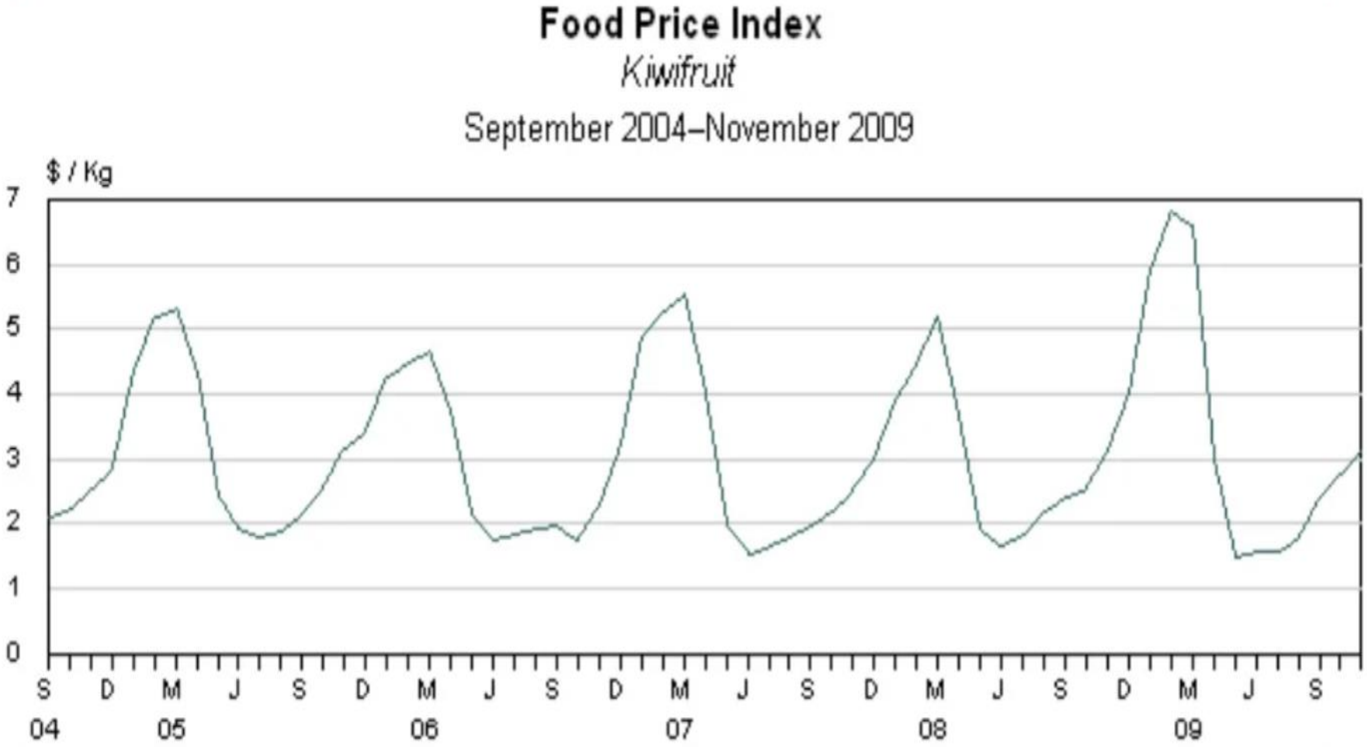


This graph shows the amazing trend of the $A vs $UA during an 18 hour period on November 8, 2000

Monthly Retail Sales in NSW Retail Department Stores

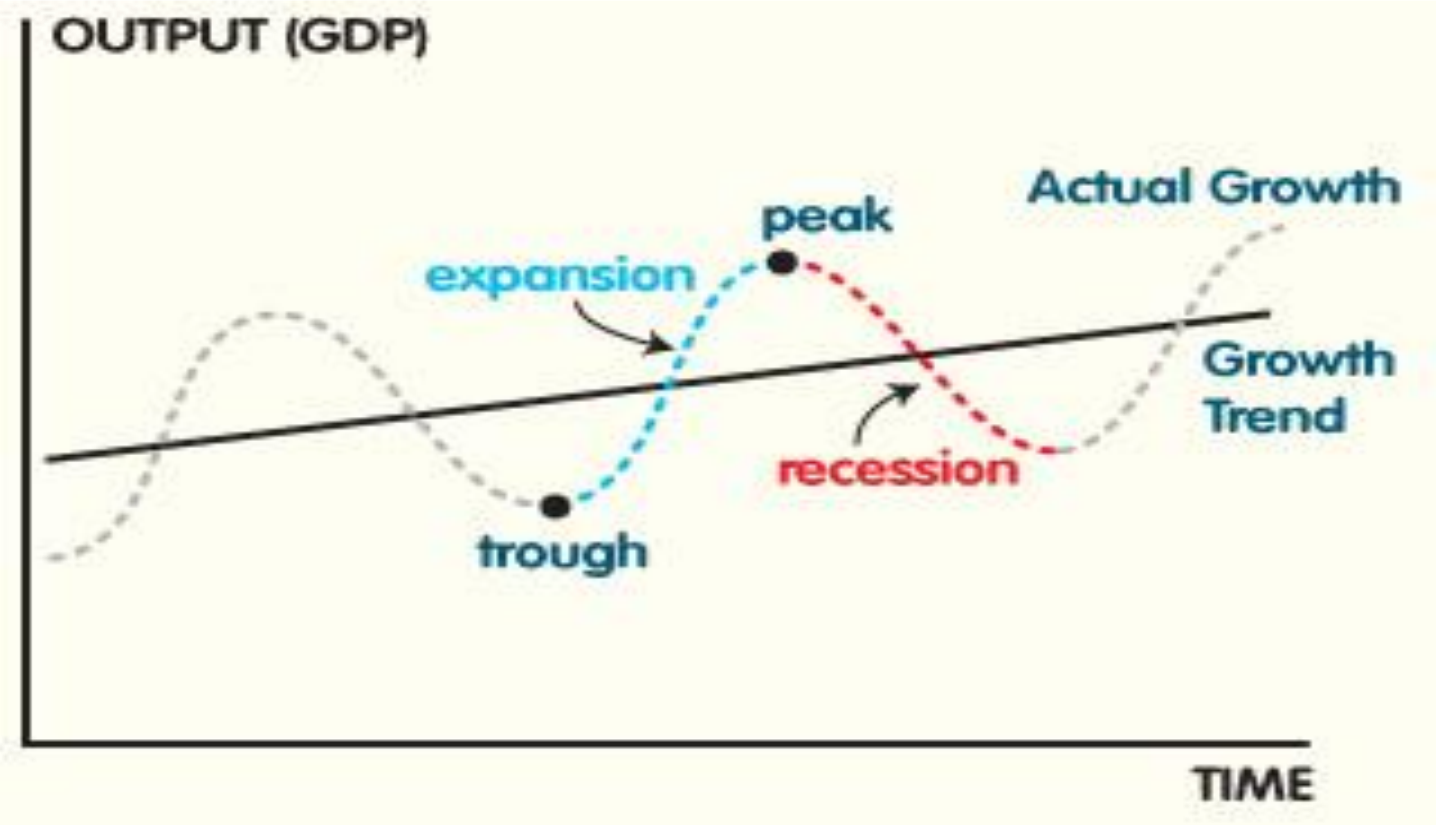


Cont…


Food Price Index
Kiwifruit
September 2004–November 2009
$ / Kg
7
6
5
4
3
2
1
0
S
D
M
J
S
D
M
J
S
D
M
J
S
D
M
J
S
D
M
J
S
04
05
06
07
08
09

Cyclical variation

This chart represents an economic cycle, but we know it doesn't always go like this. The timing and length of each phase is not predictable.



4. Irregular variation

An irregular (or random) variation in a time series occurs over varying (usually short) periods.

It follows no pattern and is by nature unpredictable.

It usually occurs randomly and may be linked to events that also occur randomly.

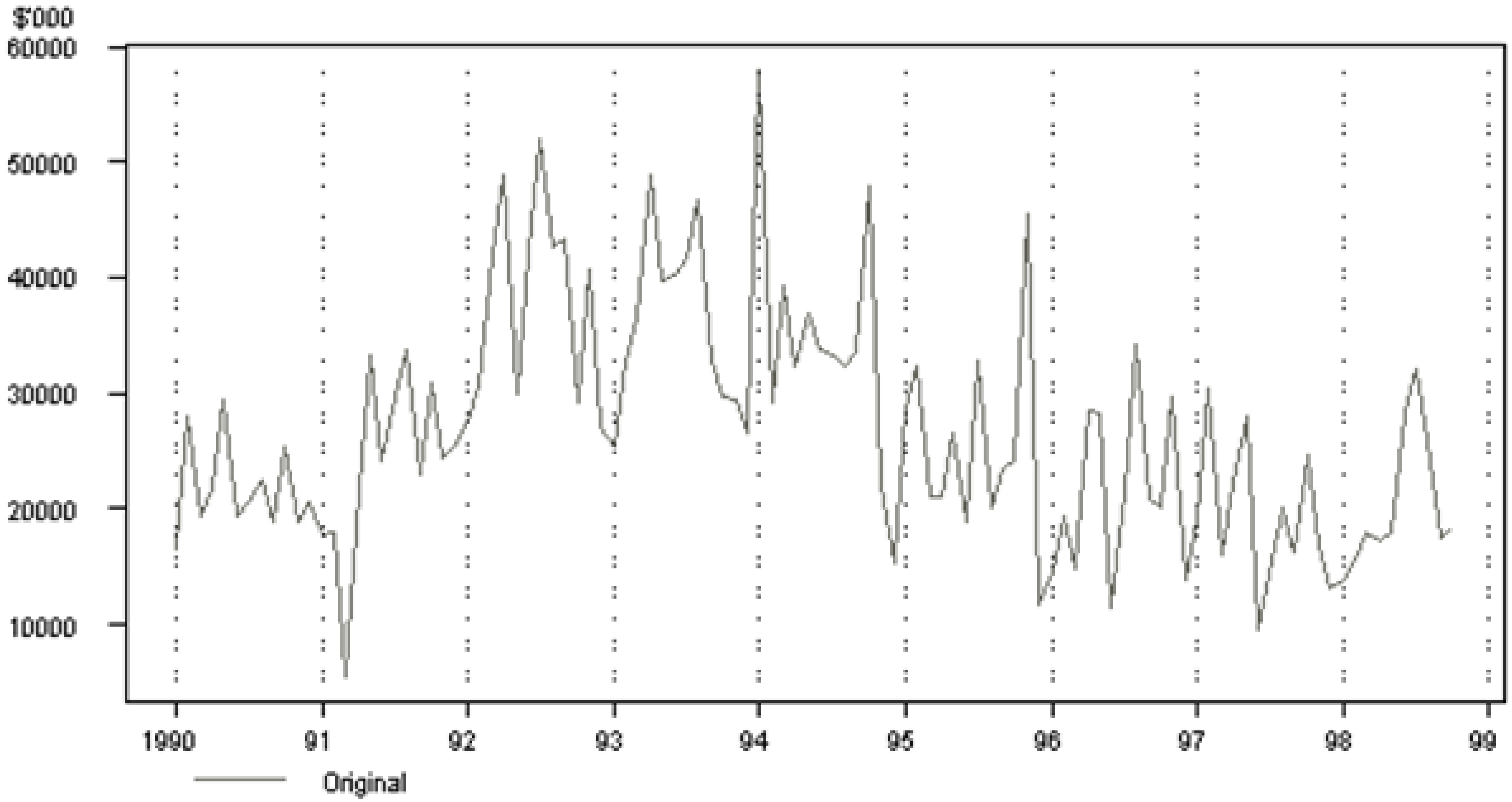
Irregular variation cannot be explained mathematically.

Irregular variation

If the variation cannot be accounted for by secular trend, season or cyclical variation, then it is usually attributed to irregular variation. Example include:

- Sudden changes in interest rates
- Collapse of companies
- Natural disasters
- Sudden shift s in government policy
- Dramatic changes to the stock market
- Effect of Middle East unrest on petrol prices

Monthly Value of Building Approvals ACT)


$000
Original