

Foundations of Data Science

Lecture 4

Rumi Chunara, PhD
CS3943/9223

So Far...

- What is Data Science?
- Intro to R
- Data cleaning, sampling, processing
- Intro to ML – what is it
- Two Basic Algorithms
 - kNN
 - Linear Regression

Today

- Time-series Analyses
- Regression and lagged data in R

Time Series Discussions

- Overview
- Basic definitions
- Time domain
- Forecasting
- Frequency domain
- State space

Why Time Series Analysis?

- Sometimes the concept we want to learn is the relationship between points in time

What is a time series?

Time series:

**a sequence of measurements
over time**

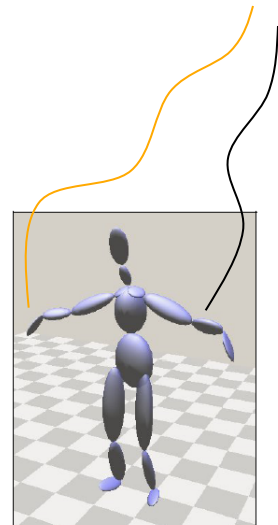
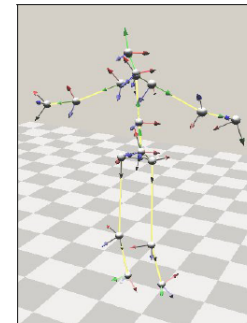
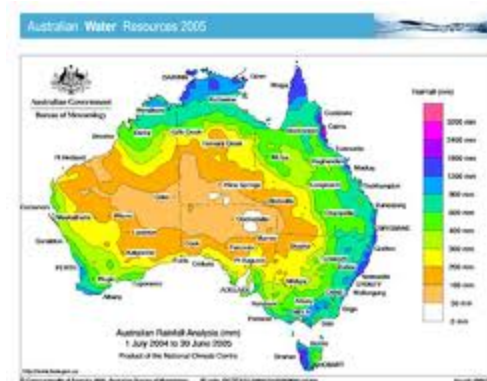
A sequence of random variables

X_1, X_2, X_3, \dots

Time Series Examples

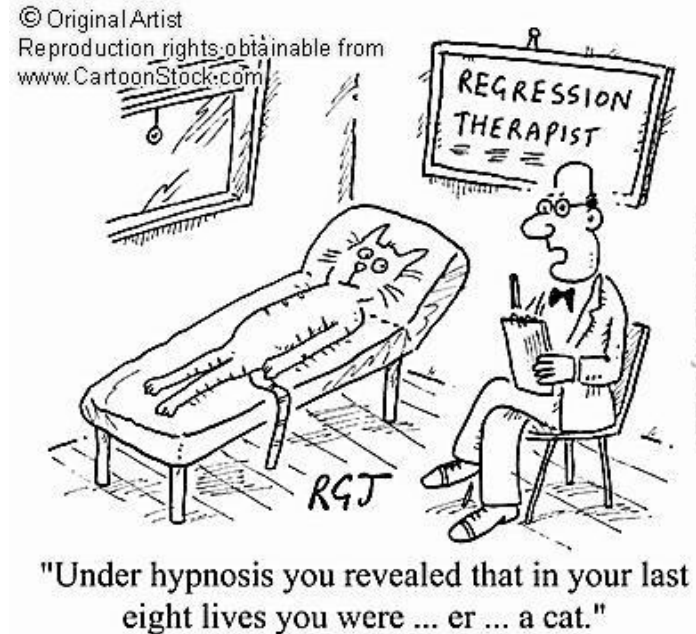
Definition: *A sequence of measurements over time*

- **Finance**
- **Social science**
- **Epidemiology**
- **Medicine**
- **Meteorology**
- **Speech**
- **Geophysics**
- **Seismology**
- **Robotics**

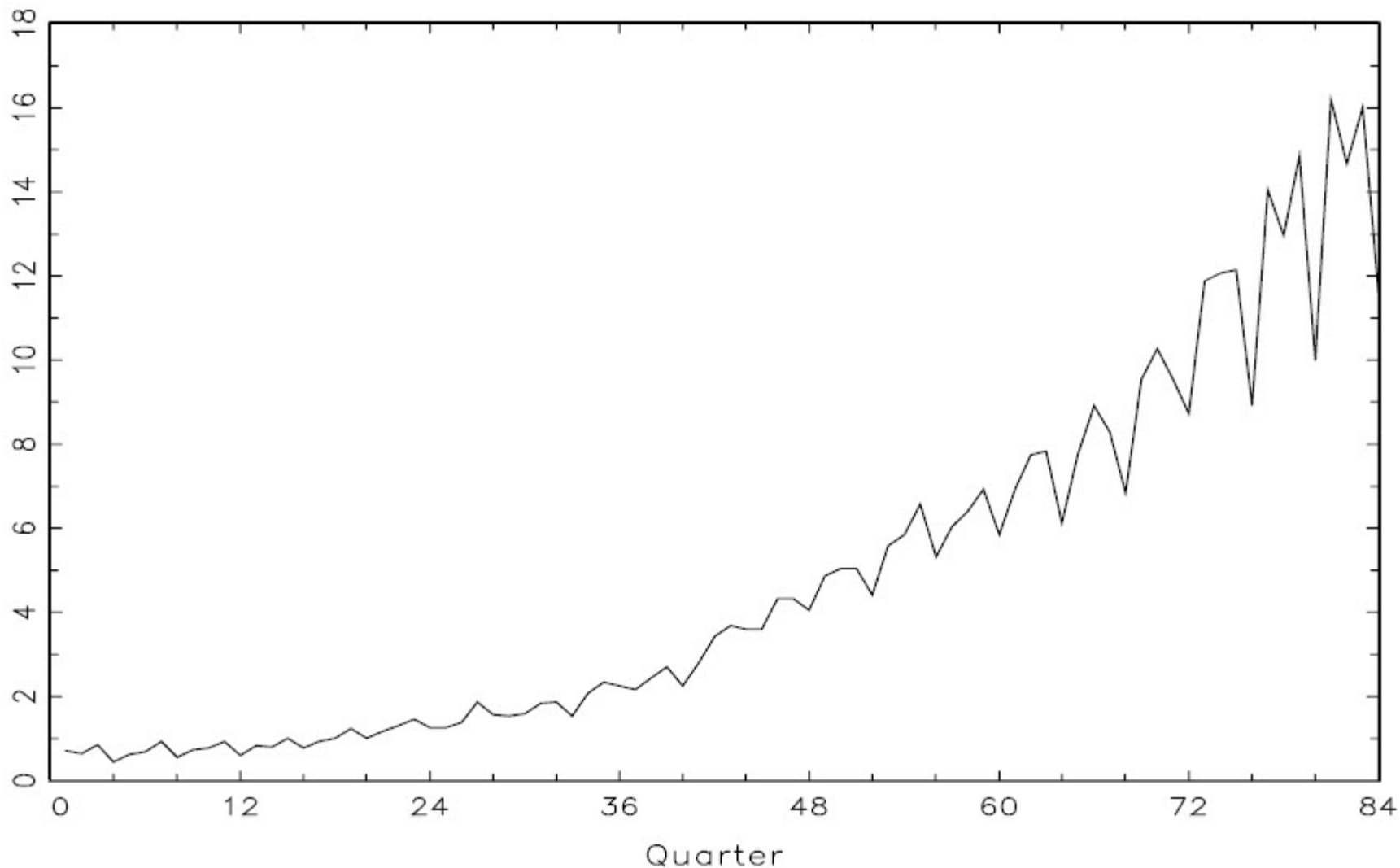


Three Approaches

- **Time domain approach**
 - Analyze dependence of current value on past values
- **Frequency domain approach**
 - Analyze periodic sinusoidal variation
- **State space models**
 - Represent state as collection of variable values
 - Model transition between states

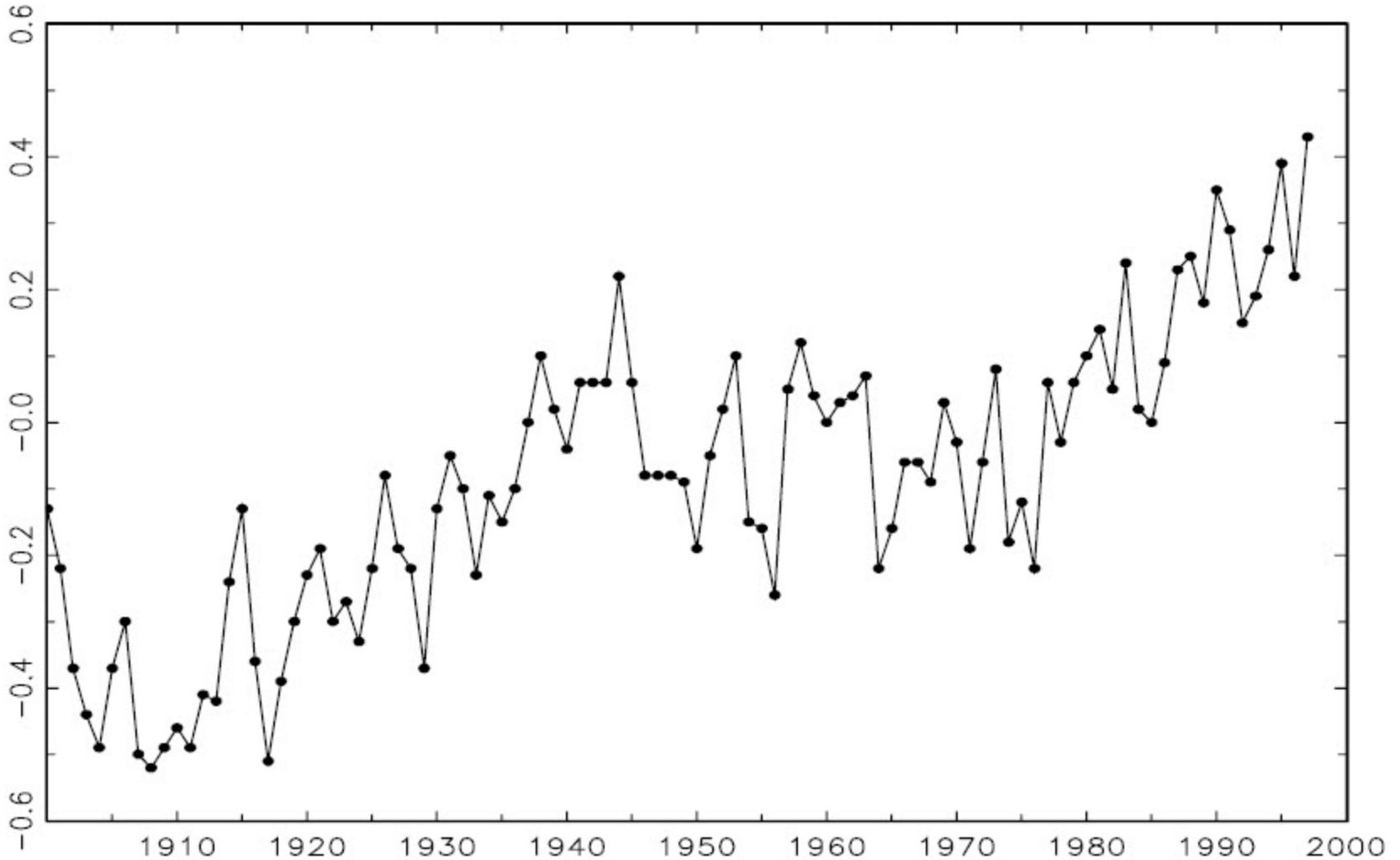


Sample Time Series Data



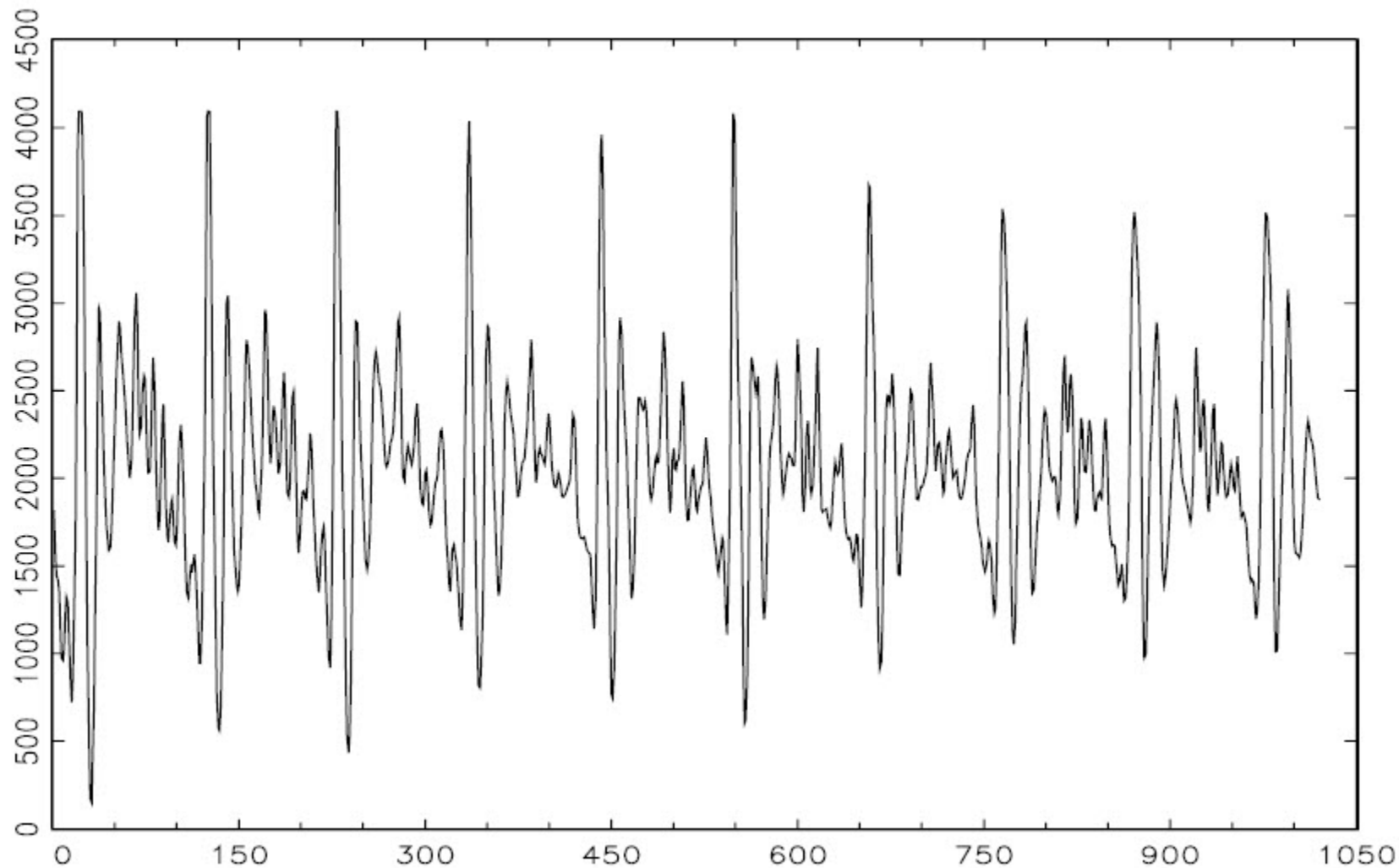
Johnson & Johnson quarterly earnings/share, 1960-1980

Sample Time Series Data



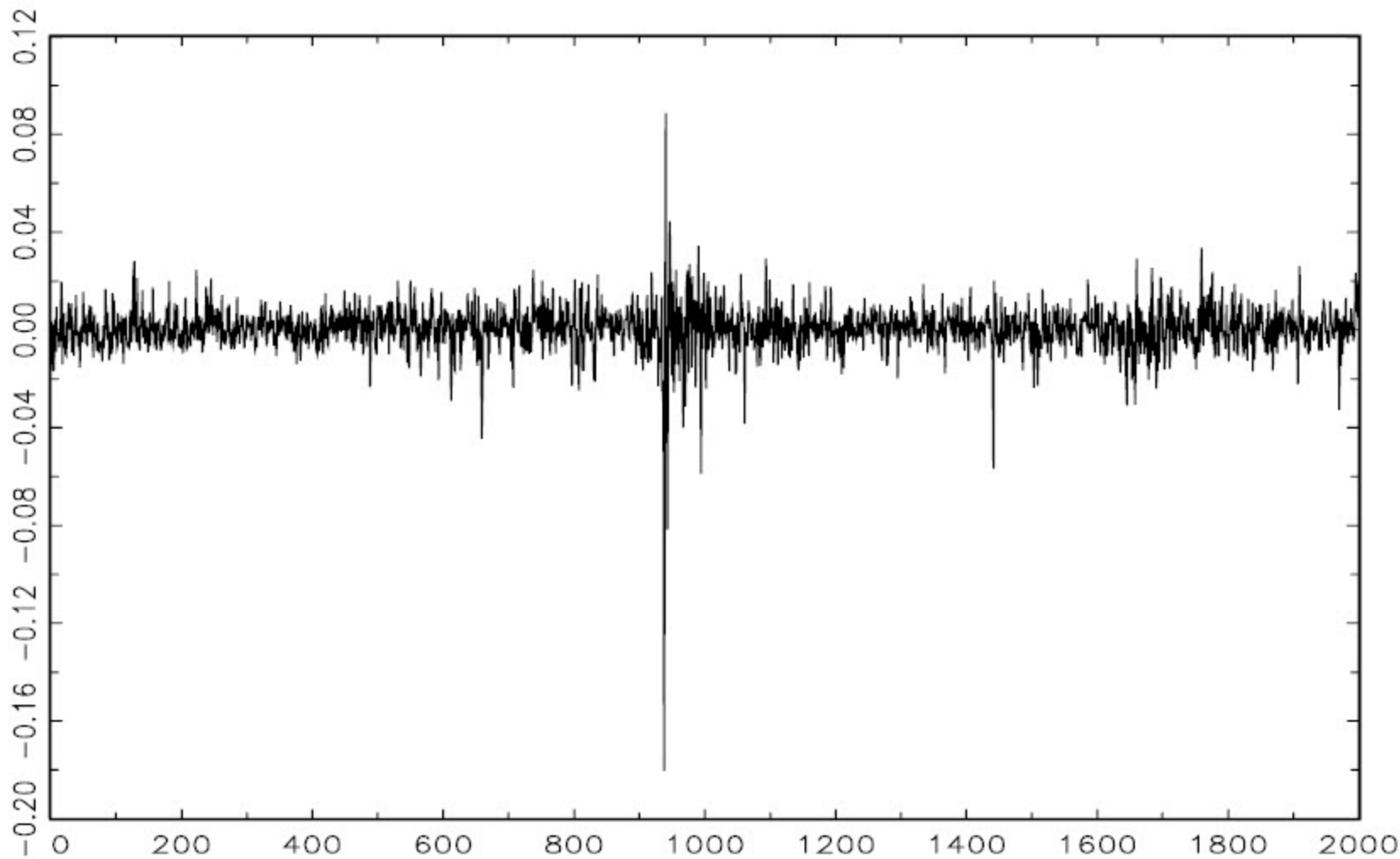
Yearly average global temperature deviations

Sample Time Series Data



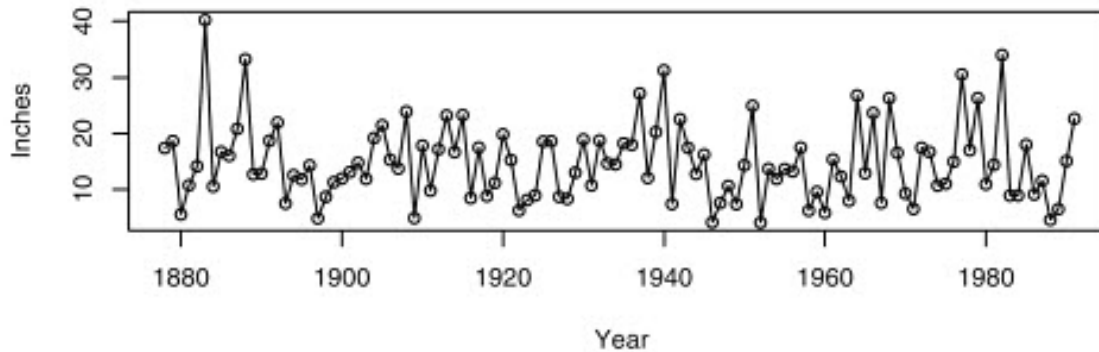
Speech recording of “aaa...hhh”, 10k pps

Sample Time Series Data

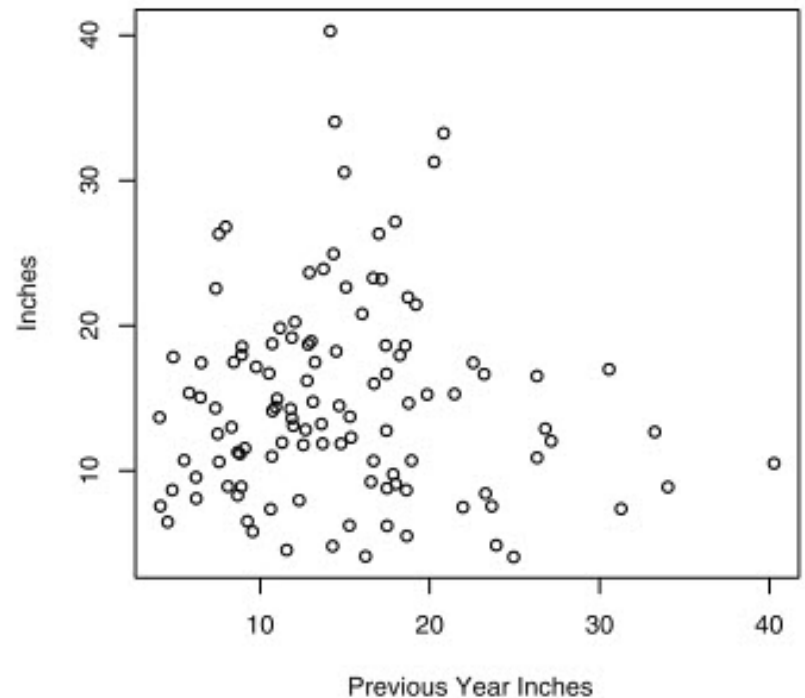


NYSE daily weighted market returns

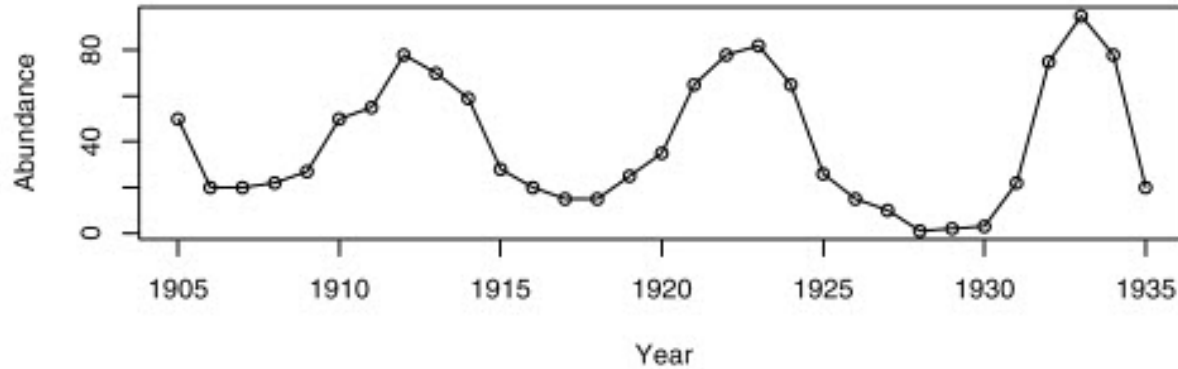
Not all time data will exhibit strong patterns...



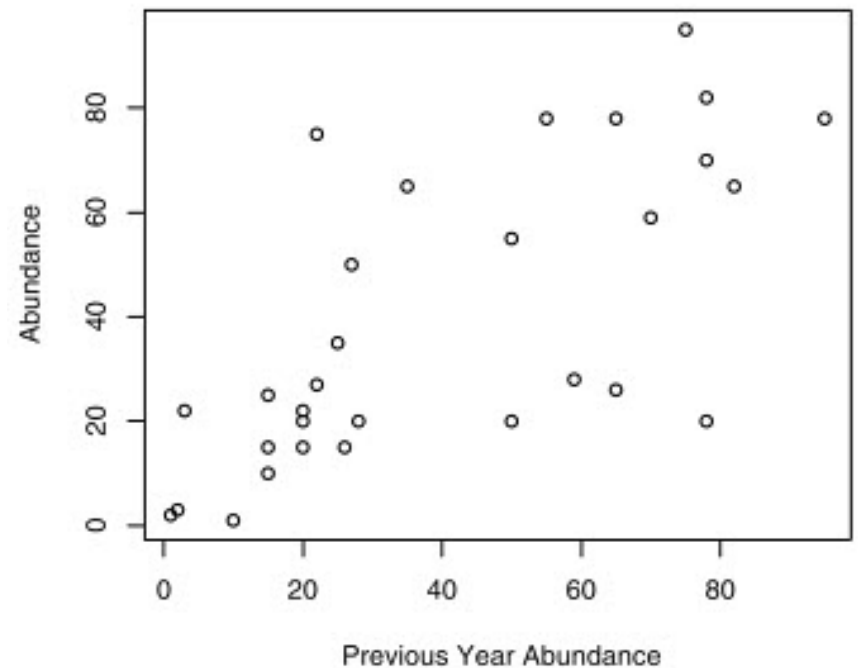
LA annual rainfall



...and others will be apparent



Canadian Hare counts

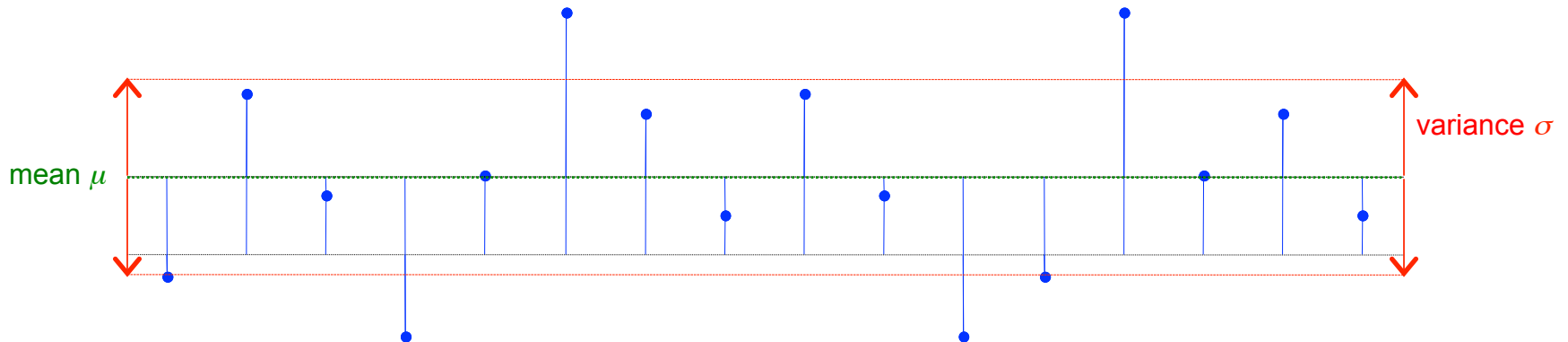


Time Series Discussions

- Overview
- Basic definitions
- Time domain
- Forecasting
- Frequency domain
- State space

Definitions

- Mean $\mu \equiv \mathbb{E}[x_t] := \frac{1}{N} \sum_{t=1}^N x_t$
- Variance $\sigma^2 \equiv \text{Var}[x_t] := \frac{1}{N} \sum_{t=1}^N (x_t - \mu)^2$



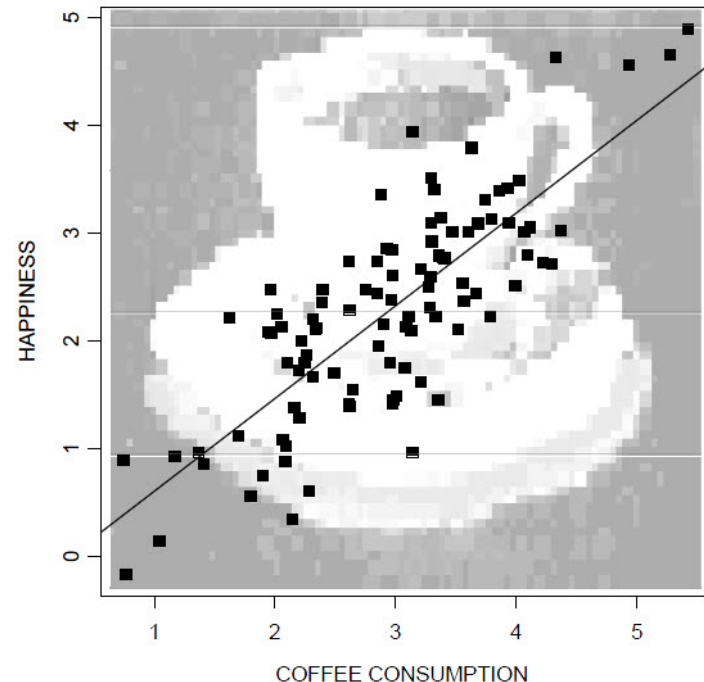
Definitions

- Covariance

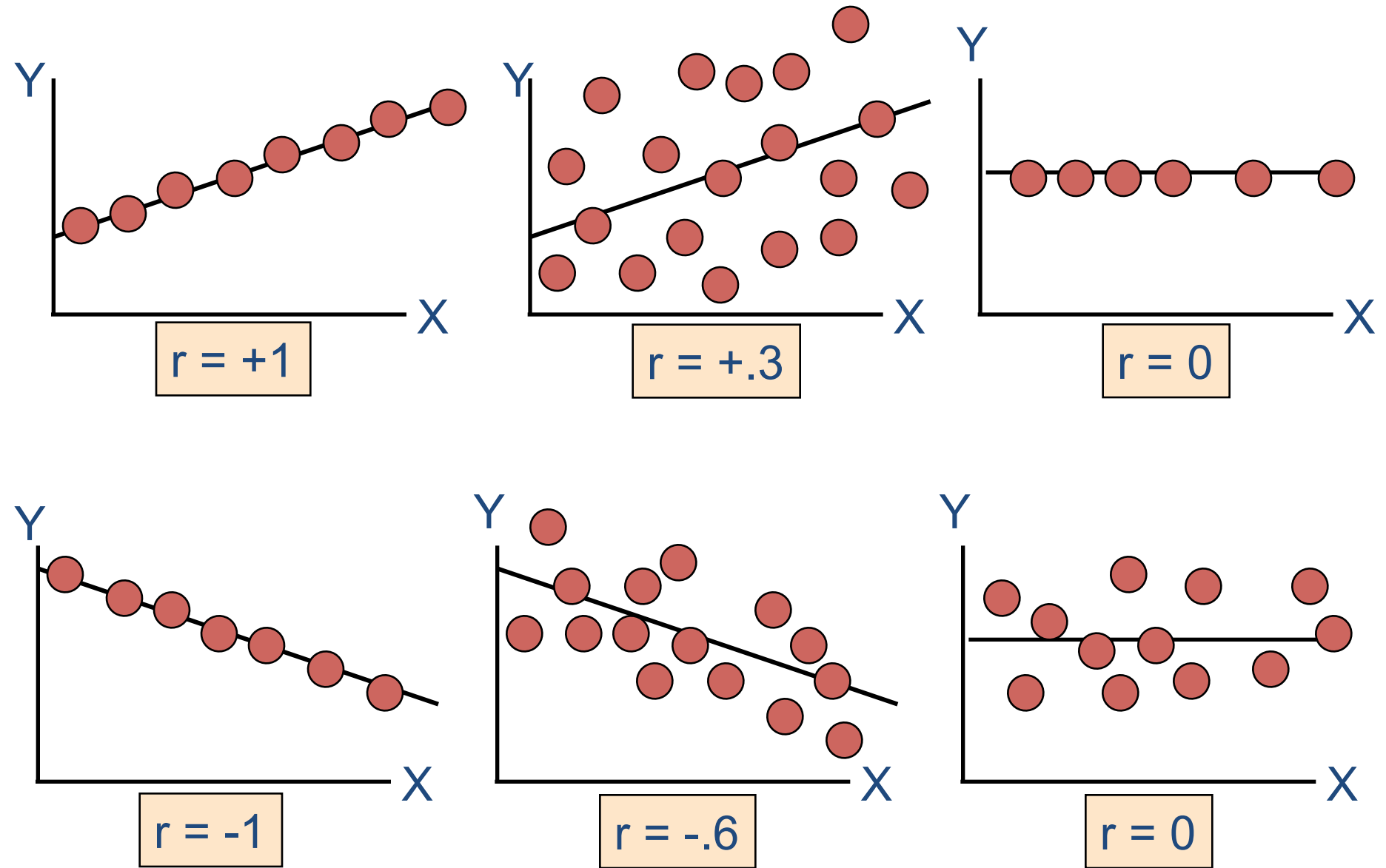
$$\text{Cov}(X, Y) = \sum_{i=1}^N \frac{(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Correlation

$$\text{Cor}(X, Y) = r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$



Correlation



Redefined for Time

- Mean function

$$\mu_X(t) = E(X_t) \quad \text{for } t = 0, \pm 1, \pm 2, \dots$$

Ergodic?

Autocovariance

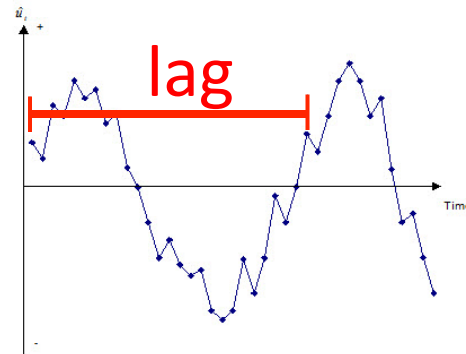
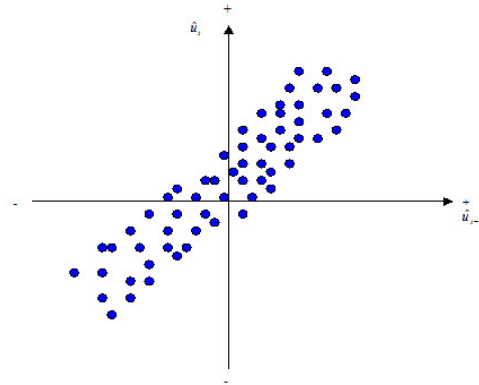
- $\gamma_X(h) = \text{Cov}(X_{t+h}, X_t)$

lag 

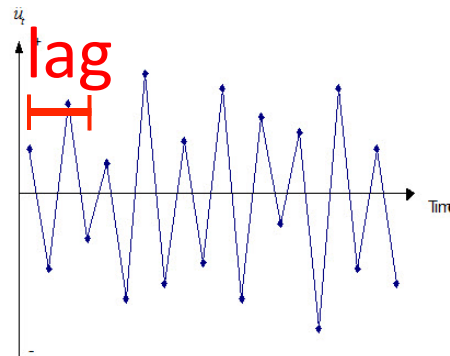
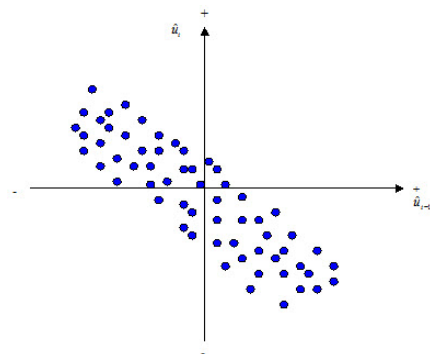
Autocorrelation

- $$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Cor}(X_{t+h}, X_t)$$

Autocorrelation Examples



Positive



Negative

Stationarity –

When there is no relationship

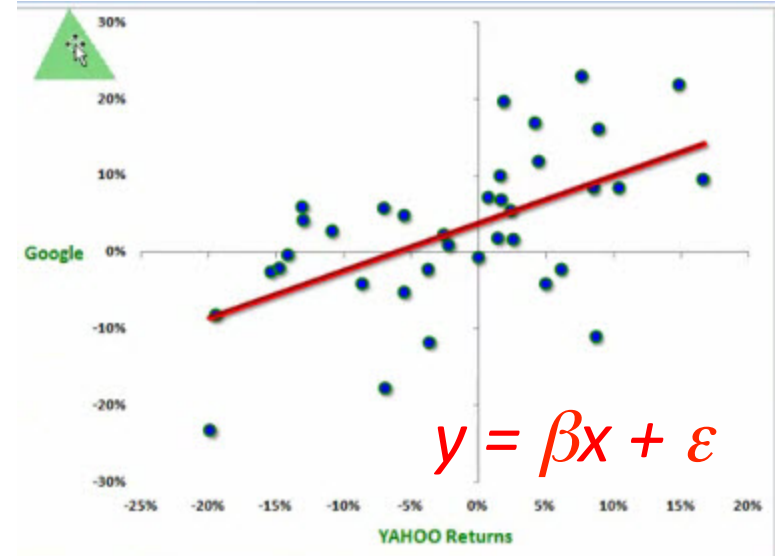
- $\{X_t\}$ is stationary if
 - $\mu_X(t)$ is independent of t
 - $\gamma_X(t+h, t)$ is independent of t for each h
- In other words, properties of each section are the same
- Special case: white noise

Time Series Discussions

- Overview
- Basic definitions
- Time domain
- Forecasting
- Frequency domain
- State space

Linear Regression

- Fit a line to the data
- Ordinary least squares
 - Minimize sum of squared distances between points and line



R^2 : Evaluating Goodness of Fit

- Least squares minimizes the combined residual

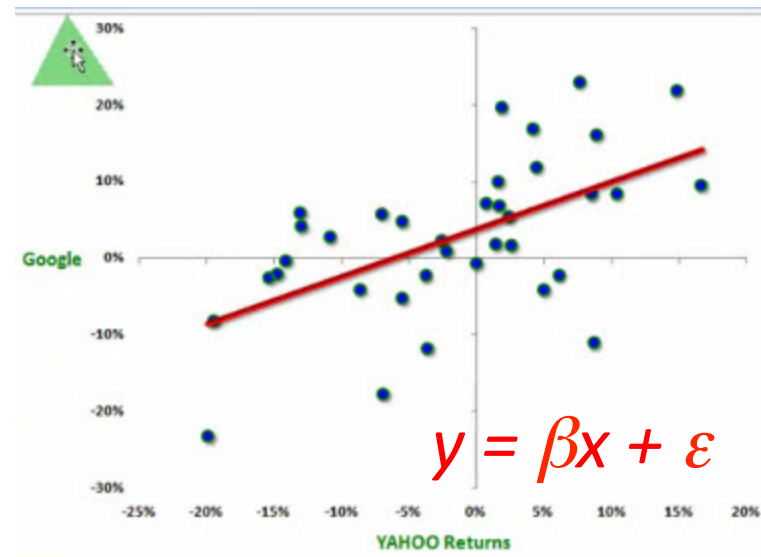
$$RSS = \sum_i (\hat{Y} - Y_i)^2$$

- Explained sum of squares is difference between line and mean

$$ESS = \sum_i (\hat{Y} - \bar{Y})^2$$

- Total sum of squares is the total of these two

$$TSS = ESS + RSS = \sum_i (\hat{Y} - Y_i)^2 + \sum_i (\hat{Y} - \bar{Y})^2$$

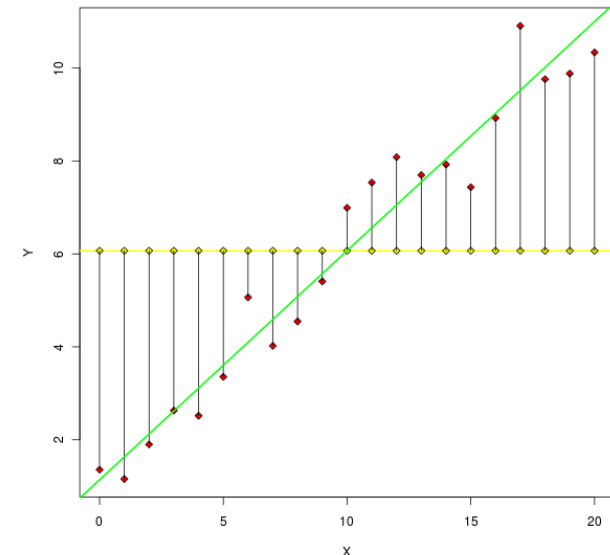
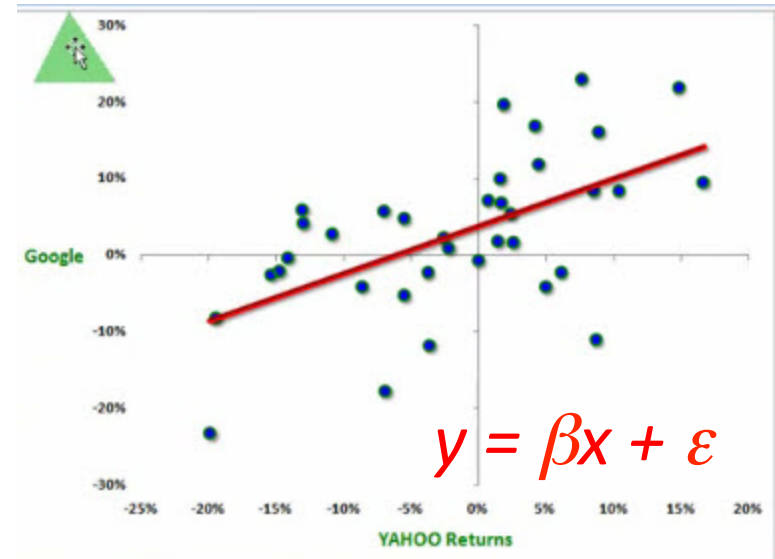


R^2 : Evaluating Goodness of Fit

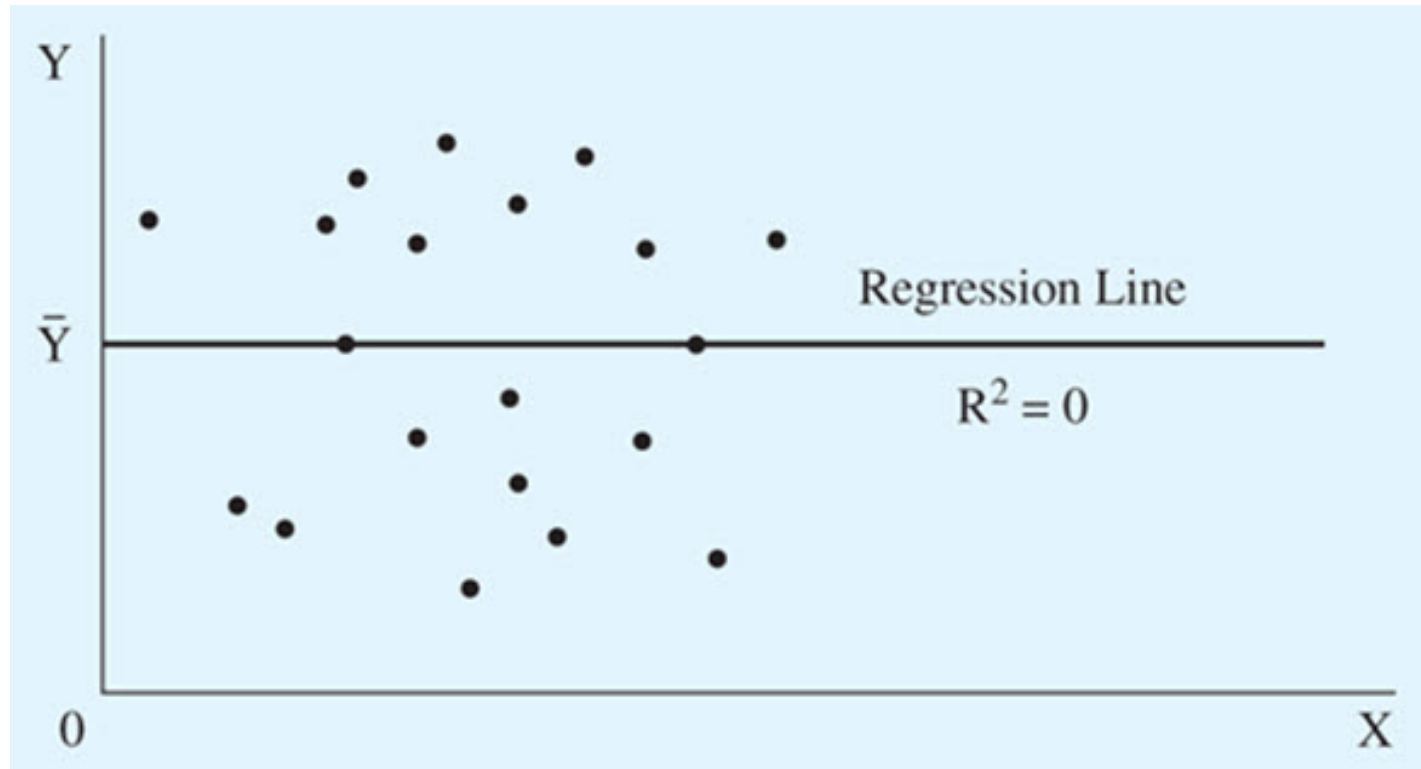
- R^2 , the coefficient of determination

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- $0 \leq R^2 \leq 1$
- Regression minimizes RSS and so maximizes R^2

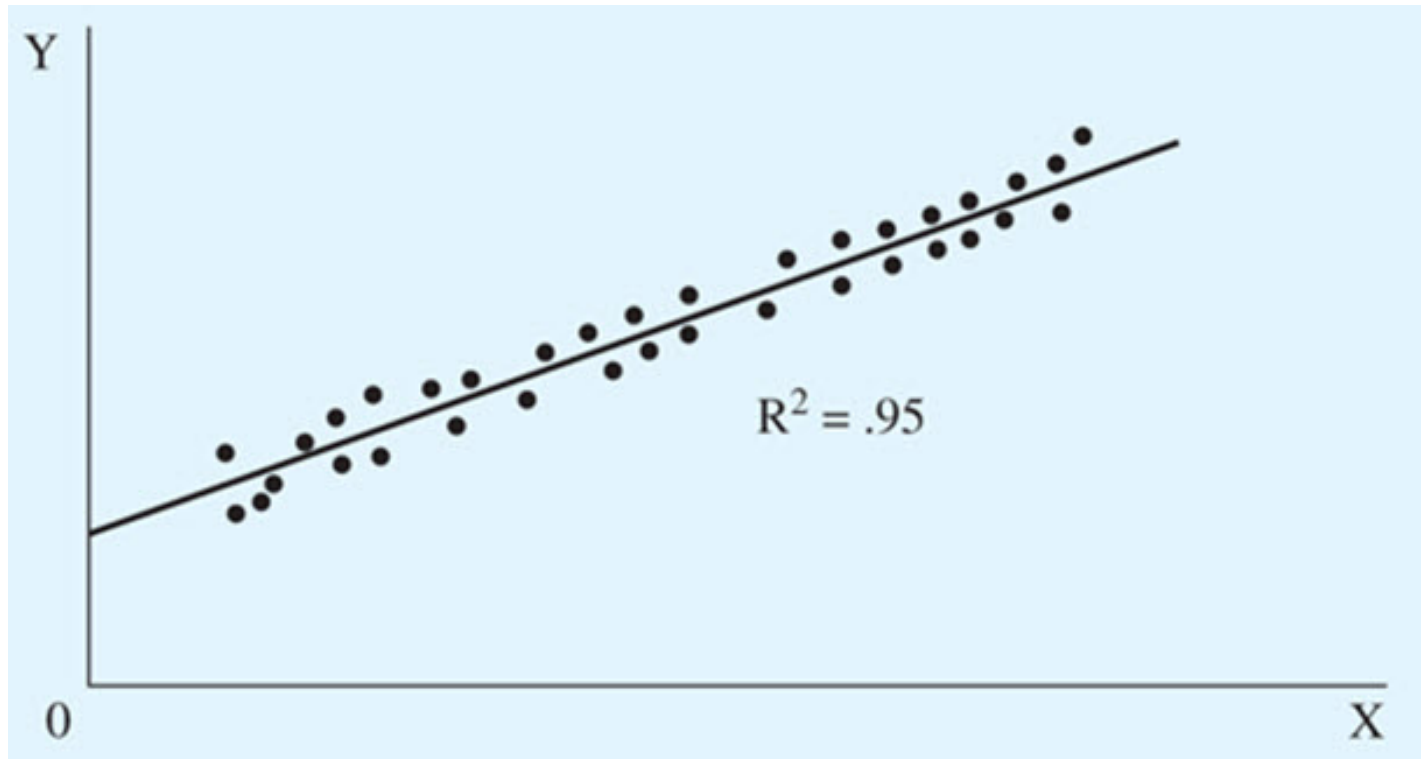


R^2 : Evaluating Goodness of Fit



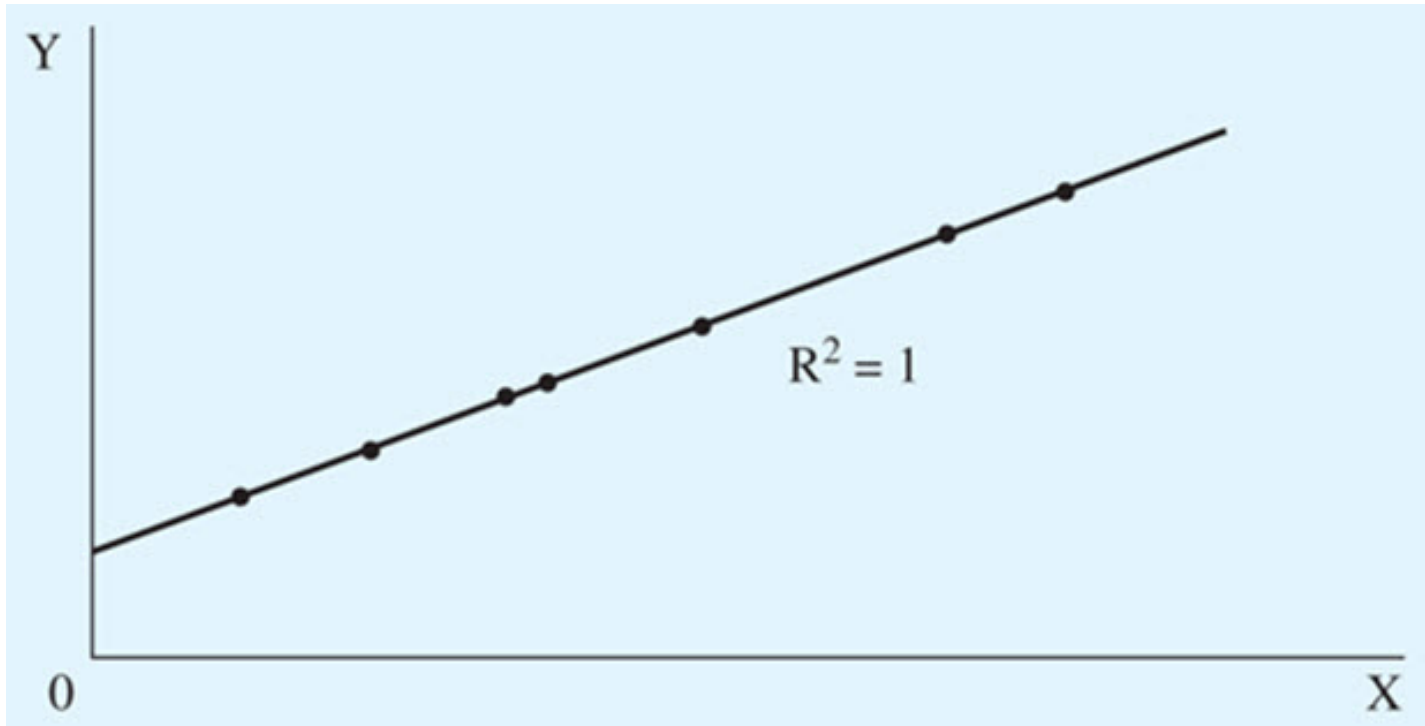
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 : Evaluating Goodness of Fit



$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

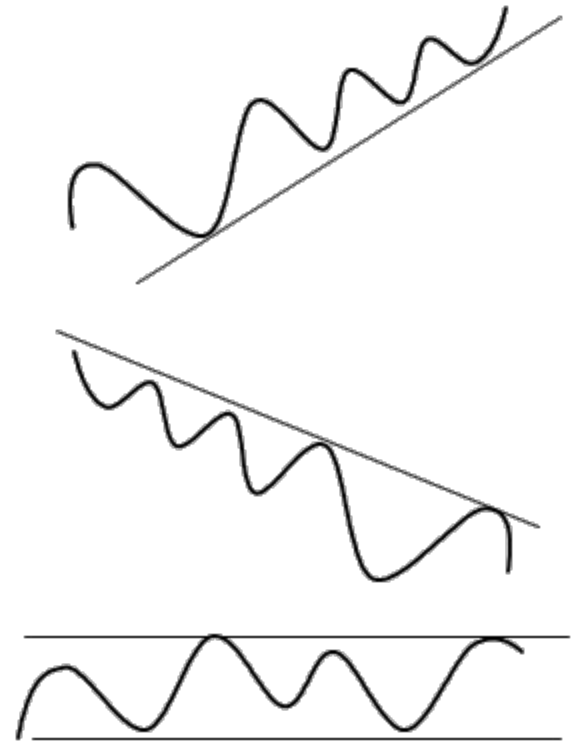
R^2 : Evaluating Goodness of Fit



$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

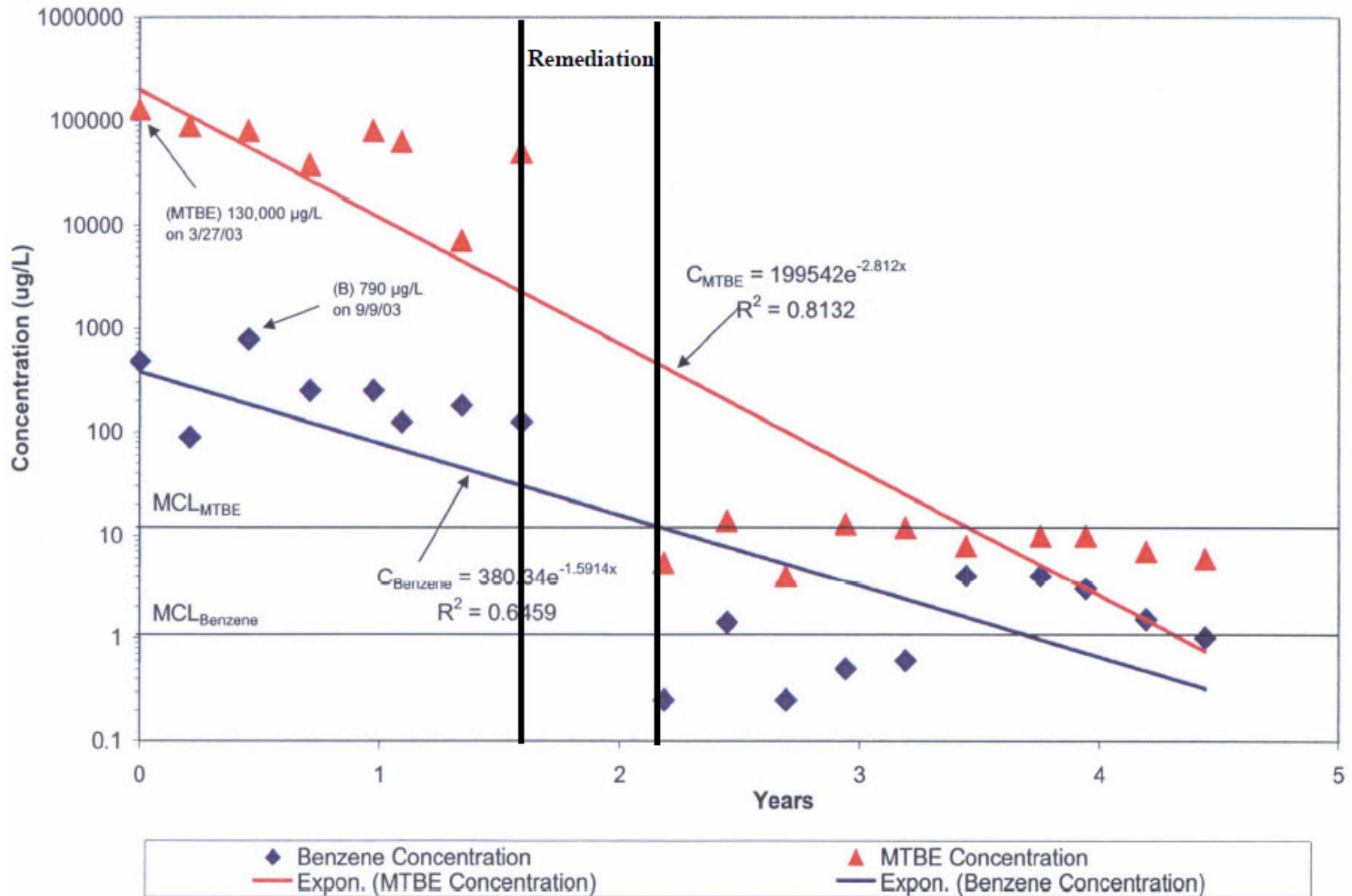
Linear Regression

- Can report:
 - Direction of trend (>0 , <0 , 0)
 - Steepness of trend (slope)
 - Goodness of fit to trend (R^2)



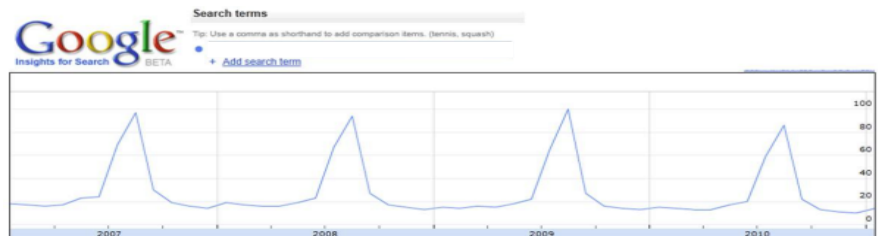
Examples

Natural Attenuation Trend Evaluation for



What if a linear trend does not fit my data well?

- Could be no relationship
- Could be too much local variation
 - Want to look at longer-term trend
 - Smooth the data
- Could have periodic or seasonality effects
 - Add seasonal components $X_t = a + b_1t + b_2Q_1 + b_3Q_2 + b_4Q_3 + b_5Q_4$
- Could be a nonlinear relationship



Moving Average

- Compute an average of the last m consecutive data points

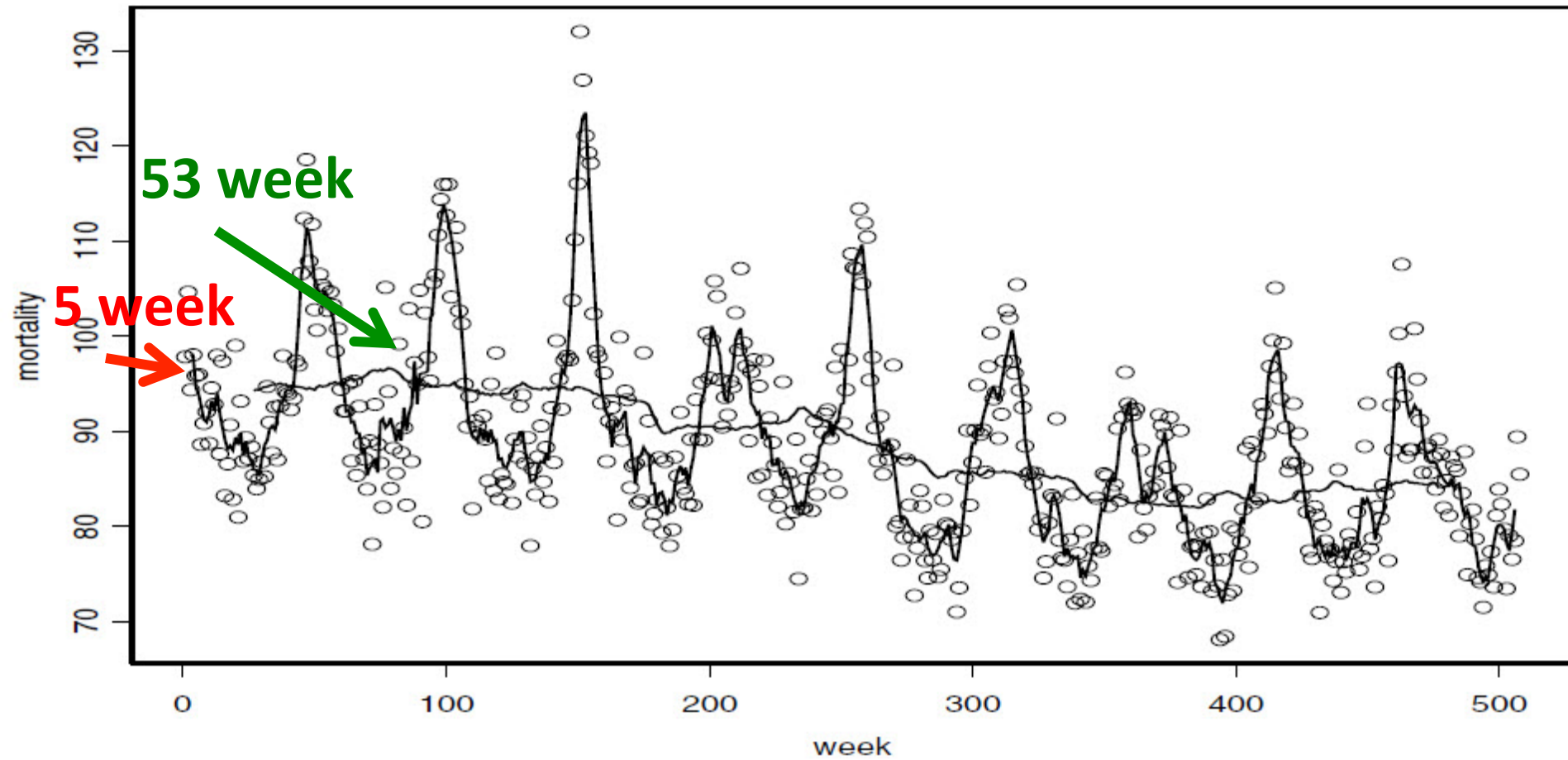
- 4-point moving average is

$$\bar{x}_{MA(4)} = \frac{(x_t + x_{t-1} + x_{t-2} + x_{t-3})}{4}$$

$$m_t = \sum_{j=-k}^k a_j x_{t-j}$$

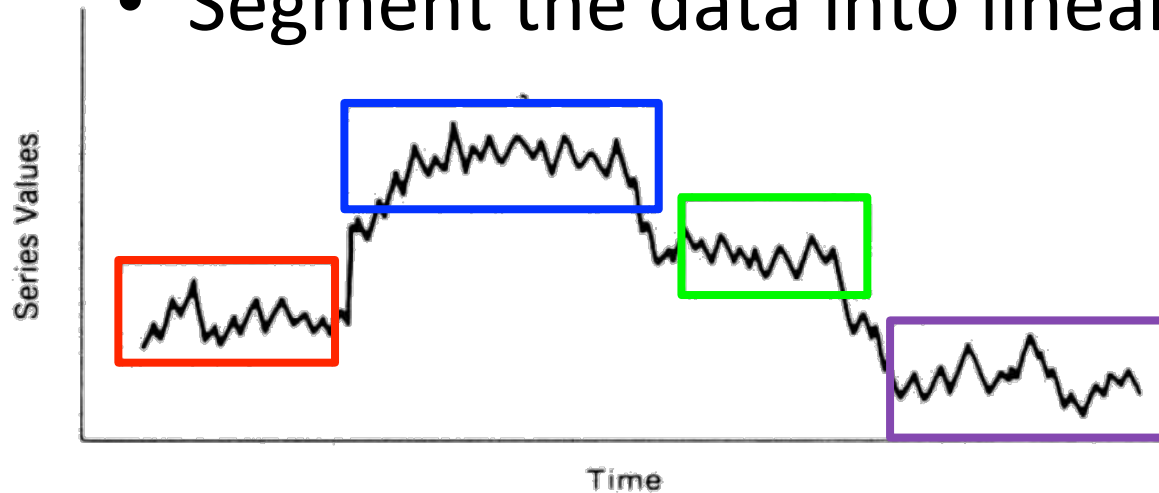
- Smooths white noise
 - Can apply higher-order MA
 - Exponential smoothing
 - Kernel smoothing

Power Load Data

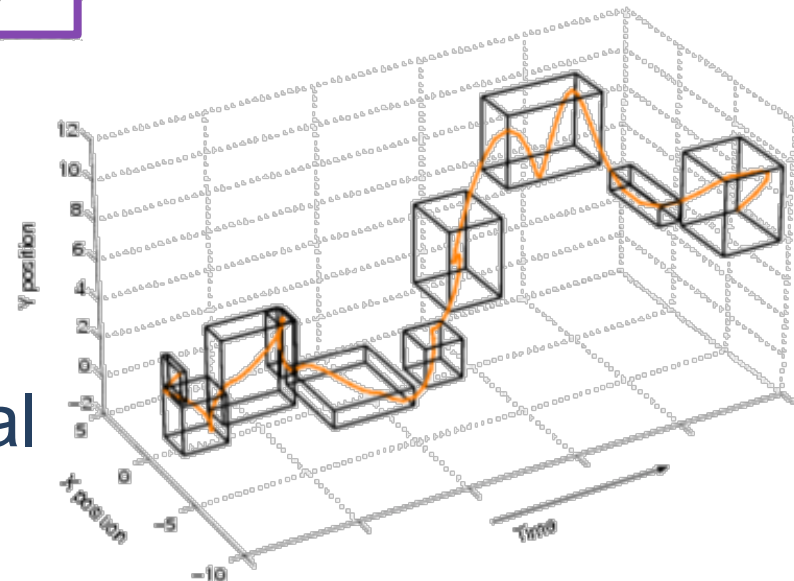


Piecewise Aggregate Approximation

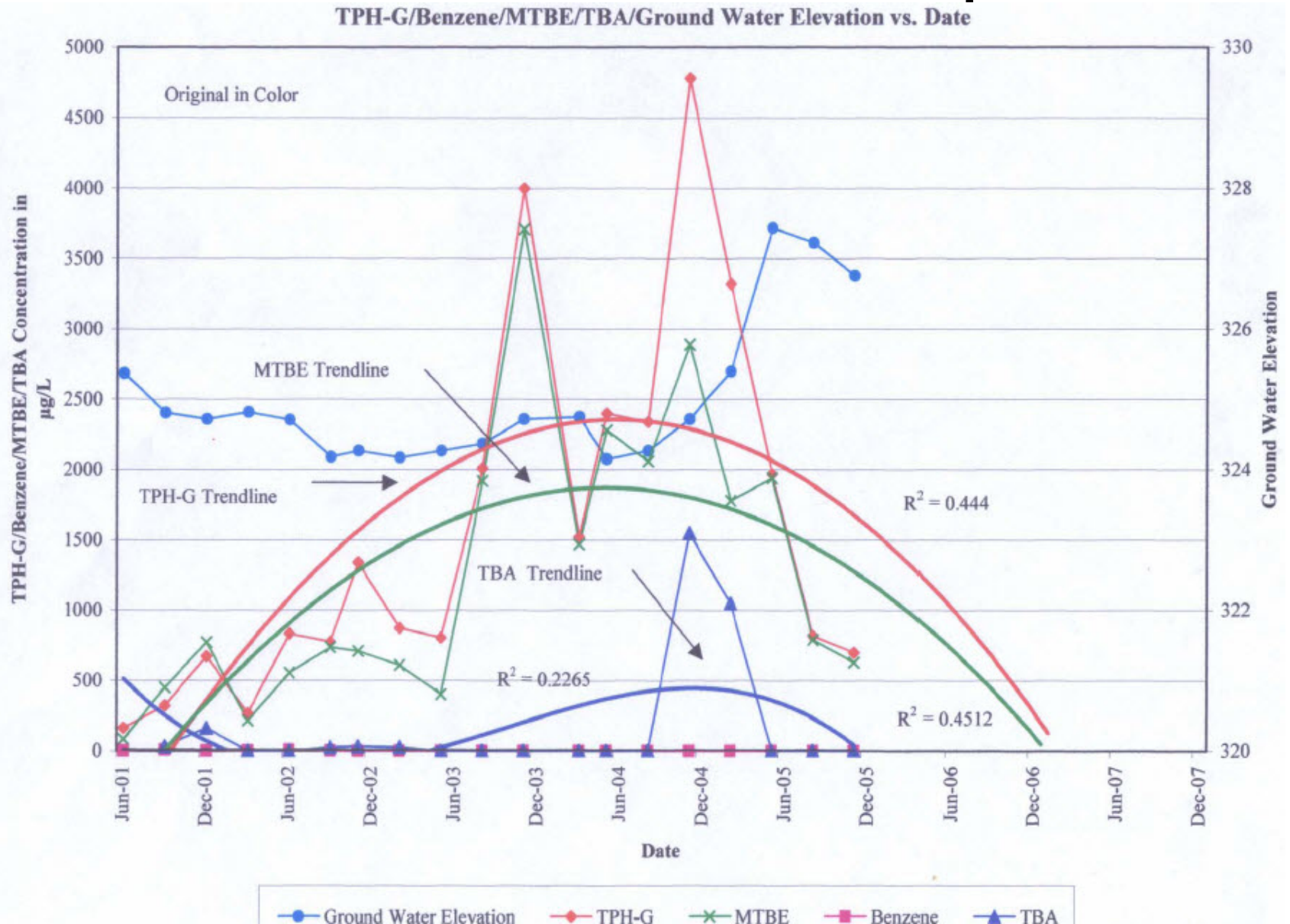
- Segment the data into linear pieces



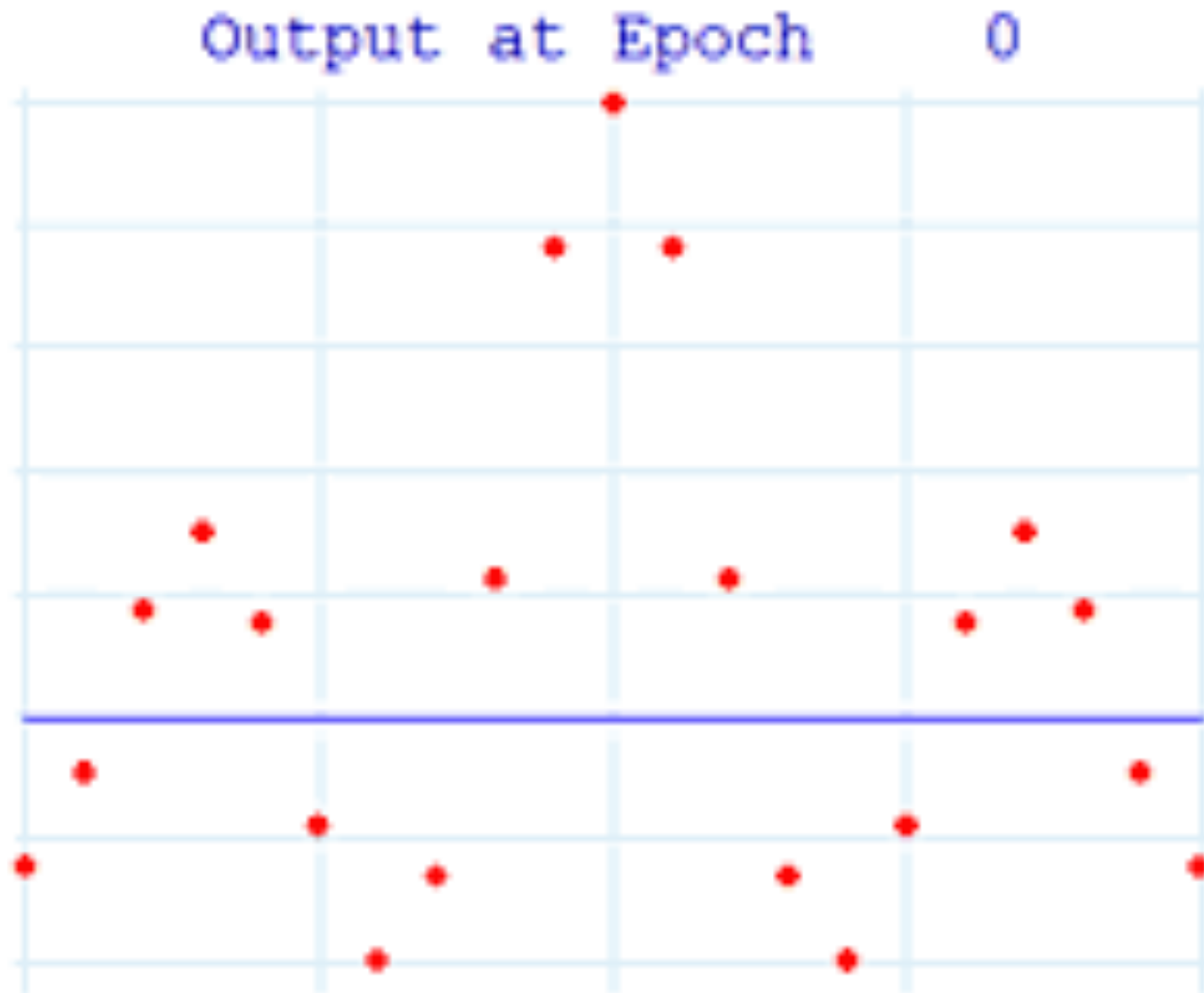
[Interesting paper](#): Hot Sax
Efficiently finding most unusual
time series subsequence



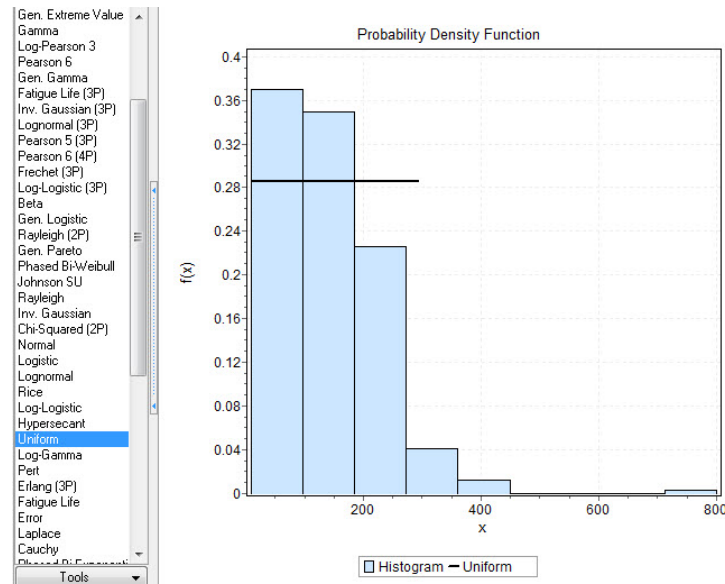
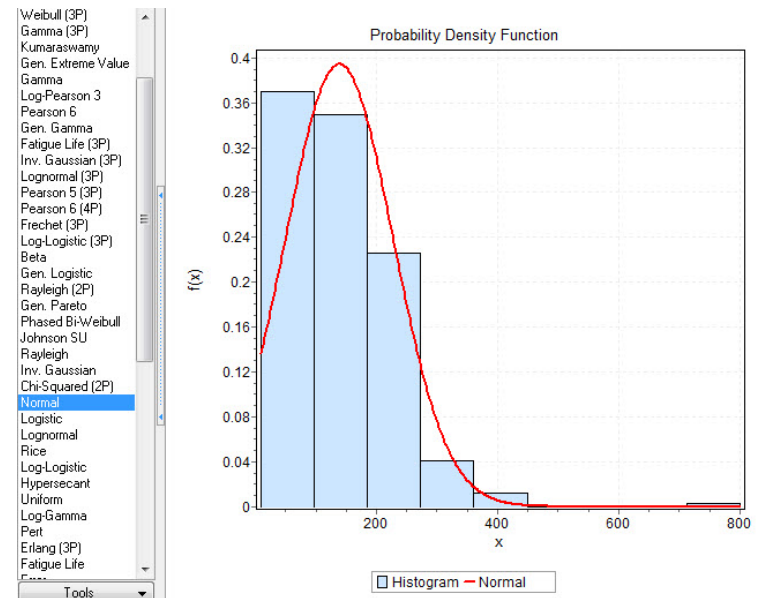
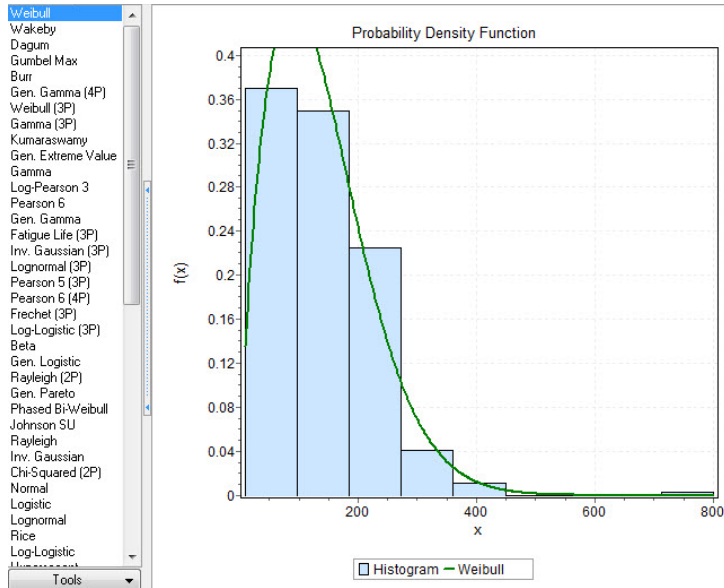
Nonlinear Trend Examples



Nonlinear Regression



Fit Known Distributions



ARIMA: Putting the pieces together

- Autoregressive model of order p : $AR(p)$
- Moving average model of order q : $MA(q)$
- $ARMA(p,q)$

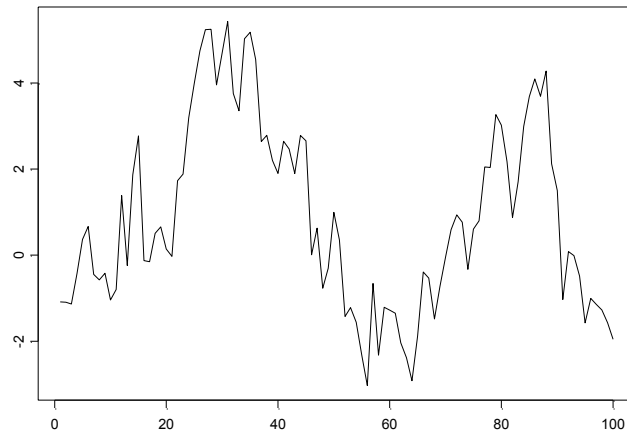
ARIMA: Putting the pieces together

- Autoregressive model of order p: AR(p)

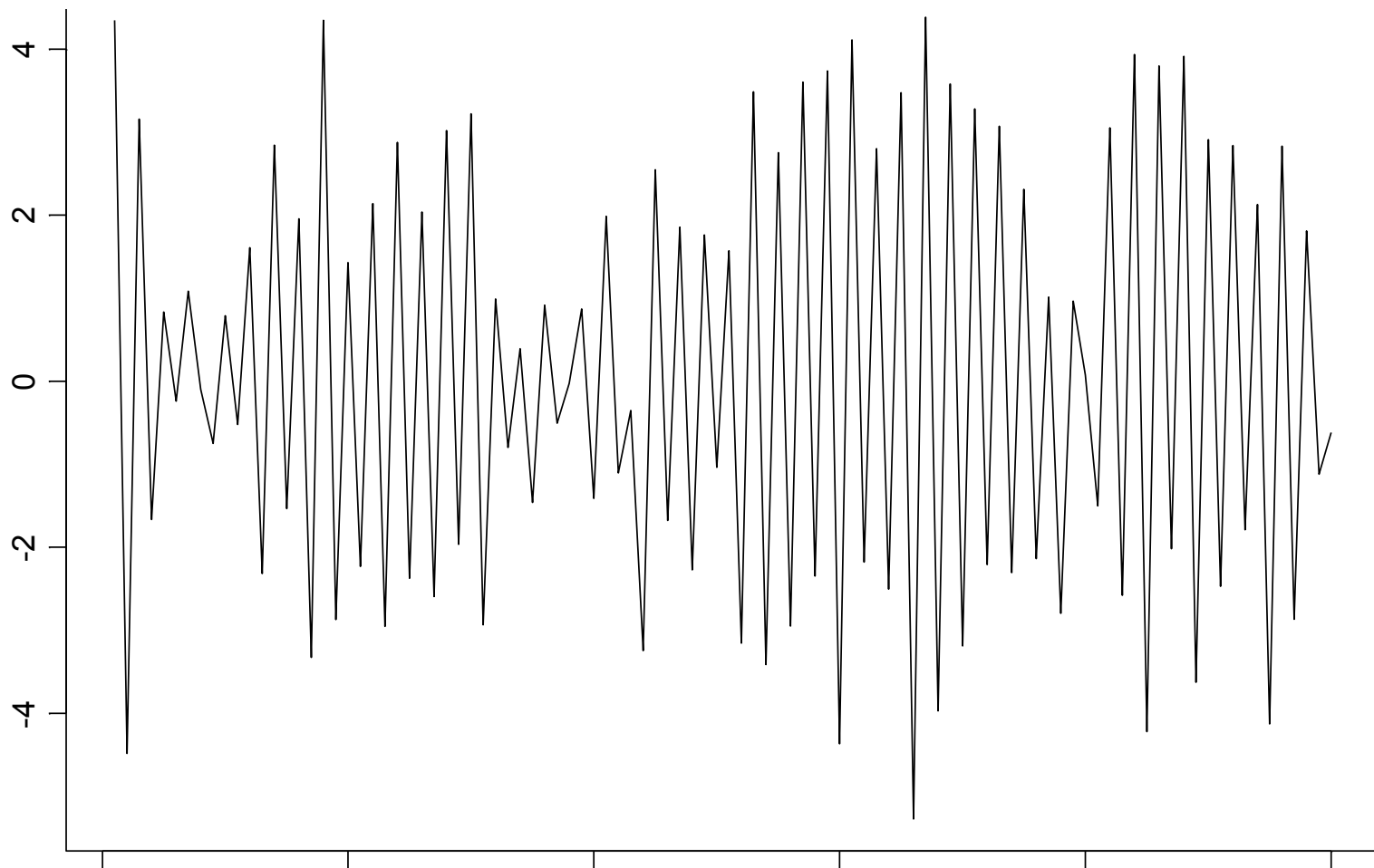
$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- Moving average model of order q: MA(q)
- ARMA(p,q)

$AR(1), \phi = 0.9$



$AR(1), \phi = -0.9$



ARIMA: Putting the pieces together

- Autoregressive model of order p: AR(p)

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- Moving average model of order q: MA(q)

$$x_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} + w_t$$

- ARMA(p,q)

ARIMA: Putting the pieces together

- Autoregressive model of order p: AR(p)

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- Moving average model of order q: MA(q)

$$x_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} + w_t$$

- ARMA(p,q)

– A time series is ARMA(p,q) if it is stationary and

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q}$$

ARMA

- Start with AR(1) sequence

$$\rho(h) - \phi\rho(h-1) = 0, \quad h = 1, 2, \dots$$

- This means

$$\rho(1) = \phi\rho(0)$$

$$\rho(2) = \phi\rho(1) = \phi^2\rho(0)$$

...

$$\rho(n) = \phi\rho(n-1) = \phi^n\rho(0)$$

- Which we can solve given roots z_i

$$\rho_n = \phi^n\rho(0) = (z_0^{-1})^n\rho(0)$$

ARIMA (AutoRegressive Integrated Moving Average)

- ARMA only applies to stationary process
- Apply differencing to obtain stationarity
 - Replace its value by incremental change from last value

Differenced	x1	x2	x3	x4
1 time		$x_2 - x_1'$	$x_3' - x_2'$	$x_4' - x_3'$
2 times			$x_3' - 2x_2' + x_1'$	$x_4' - 2x_3' + x_2'$

- A process x_t is ARIMA(p,d,q) if
 - AR(p)
 - MA(q)
 - Differenced d times
- Also known as **Box Jenkins**

Time Series Discussions

- Overview
- Basic definitions
- Time domain
- **Forecasting**
- Frequency domain
- State space

Objectives

- Give the fundamental rules of forecasting
- Calculate a forecast using a moving average, weighted moving average, and exponential smoothing
- Calculate the accuracy of a forecast

What is forecasting?

Forecasting is a tool used for predicting future demand based on past demand information.

Why is forecasting important?

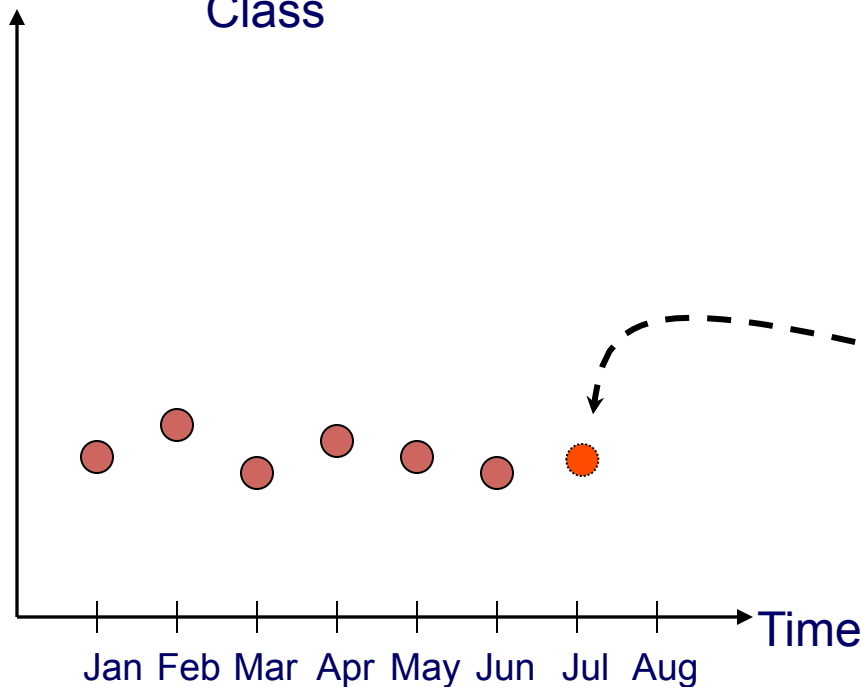
Demand for products and services is usually uncertain.

Forecasting can be used for...

- Strategic planning (long range planning)
- Finance and accounting (budgets and cost controls)
- Marketing (future sales, new products)
- Production and operations

What is forecasting all about?

Demand for Mercedes E Class



We try to predict the future by looking back at the past

Predicted demand looking back six months

- Actual demand (past sales)
- Predicted demand

What's Forecasting All About?

From the March 10, 2006 WSJ:

Ahead of the Oscars, an economics professor, at the request of Weekend Journal, processed data about this year's films nominated for best picture through his statistical model and predicted with 97.4% certainty that "Brokeback Mountain" would win. Oops. Last year, the professor tuned his model until it correctly predicted 18 of the previous 20 best-picture awards; then it predicted that "The Aviator" would win; "Million Dollar Baby" won instead.

Sometimes models tuned to prior results don't have great predictive powers.

Some general characteristics of forecasts

- Forecasts are always wrong
- Forecasts are more accurate for groups or families of items
- Forecasts are more accurate for shorter time periods
- Every forecast should include an error estimate
- Forecasts are no substitute for calculated demand.

Key issues in forecasting

1. A forecast is only as good as the information included in the forecast (past data)
2. History is not a perfect predictor of the future (i.e.: there is no such thing as a perfect forecast)

REMEMBER: Forecasting is based on the assumption that the past predicts the future! When forecasting, think carefully whether or not the past is strongly related to what you expect to see in the future...

Example: Mercedes E-class vs. M-class Sales

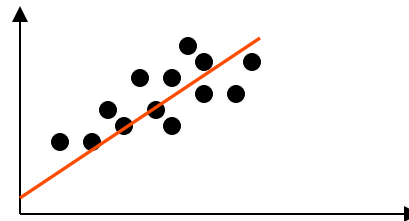
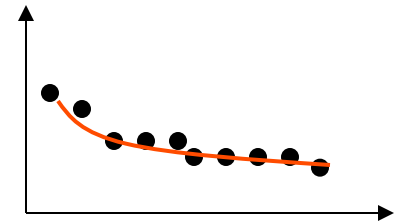
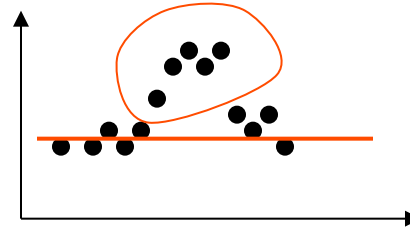
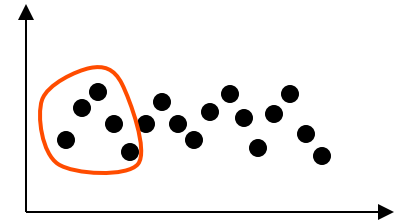
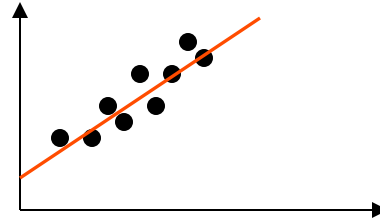
Month	E-class Sales	M-class Sales
<i>Jan</i>	23,345	-
<i>Feb</i>	22,034	-
<i>Mar</i>	21,453	-
<i>Apr</i>	24,897	-
<i>May</i>	23,561	-
<i>Jun</i>	22,684	-
<i>Jul</i>	?	?

Question: Can we predict the new model M-class sales based on the data in the the table?

Answer: Maybe... We need to consider how much the two markets have in common

What should we consider when looking at past demand data?

- Trends
- Seasonality
- Cyclical elements
- Autocorrelation
- Random variation



Some Important Questions

- What is the purpose of the forecast?
- Which systems will use the forecast?
- How important is the past in estimating the future?

Answers will help determine time horizons, techniques, and level of detail for the forecast.

Types of forecasting methods

Qualitative methods

Rely on subjective opinions from one or more experts.

Quantitative methods

Rely on data and analytical techniques.

Qualitative forecasting methods

Grass Roots: deriving future demand by asking the person closest to the customer.

Market Research: trying to identify customer habits; new product ideas.

Panel Consensus: deriving future estimations from the synergy of a panel of experts in the area.

Historical Analogy: identifying another similar market.

Delphi Method: similar to the panel consensus but with concealed identities.

Quantitative forecasting methods

Time Series: models that predict future demand based on past history trends

Causal Relationship: models that use statistical techniques to establish relationships between various items and demand

Simulation: models that can incorporate some randomness and non-linear effects

How should we pick our forecasting model?

1. Data availability
2. Time horizon for the forecast
3. Required accuracy
4. Required Resources

Time Series: Moving average

- The moving average model uses the last t periods in order to predict demand in period $t+1$.
- There can be two types of moving average models: simple moving average and weighted moving average
- The moving average model assumption is that the most accurate prediction of future demand is a simple (linear) combination of past demand.

Time series: simple moving average

In the simple moving average models the forecast value is

$$F_{t+1} = \frac{A_t + A_{t-1} + \dots + A_{t-n}}{n}$$

t is the current period.

F_{t+1} is the forecast for next period

n is the forecasting horizon (how far back we look),

Example: forecasting sales at Kroger

Kroger sells (among other stuff) bottled spring water

Month	Bottles
<i>Jan</i>	<i>1,325</i>
<i>Feb</i>	<i>1,353</i>
<i>Mar</i>	<i>1,305</i>
<i>Apr</i>	<i>1,275</i>
<i>May</i>	<i>1,210</i>
<i>Jun</i>	<i>1,195</i>
<i>Jul</i>	<i>?</i>

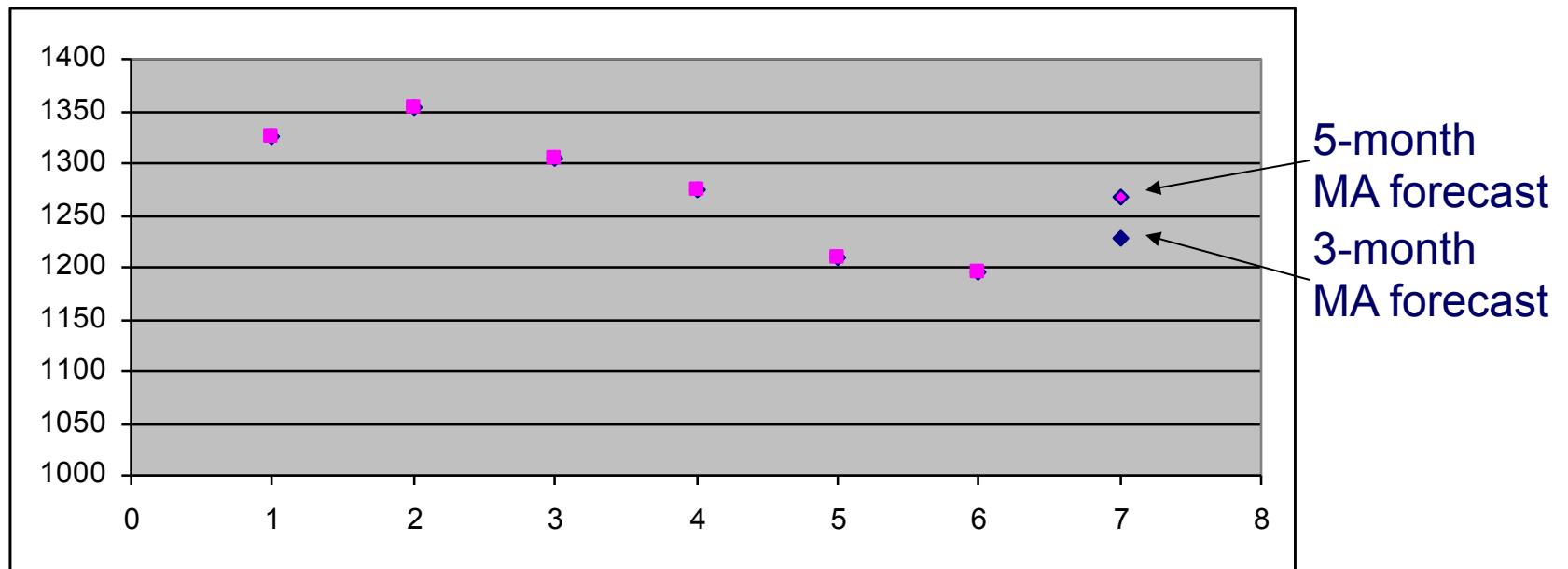


What if we use a 3-month simple moving average?

$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr}}{3} = 1,227$$

What if we use a 5-month simple moving average?

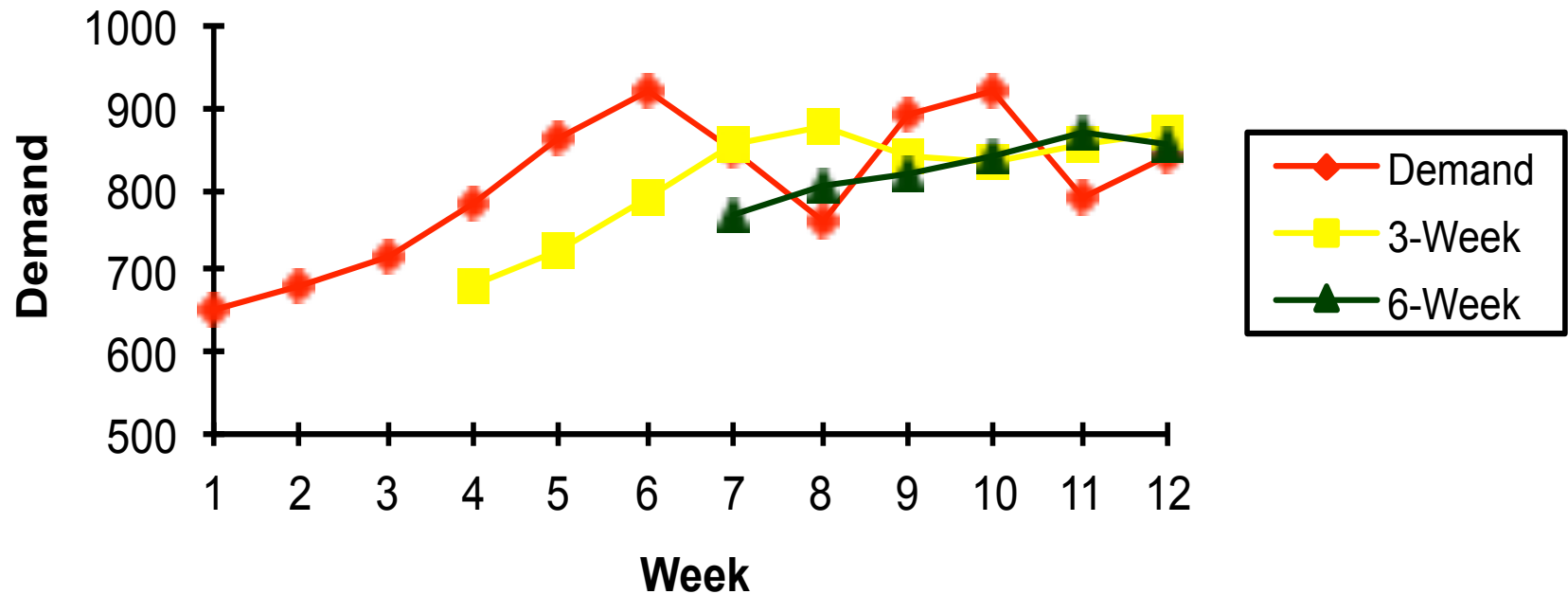
$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr} + A_{Mar} + A_{Feb}}{5} = 1,268$$



What do we observe?

5-month average smoothes data more;
3-month average more responsive

Stability versus responsiveness in moving averages



Time series: weighted moving average

We may want to give more importance to some of the data...

$$F_{t+1} = w_t A_t + w_{t-1} A_{t-1} + \dots + w_{t-n} A_{t-n}$$

$$w_t + w_{t-1} + \dots + w_{t-n} = 1$$

t is the current period.

F_{t+1} is the forecast for next period

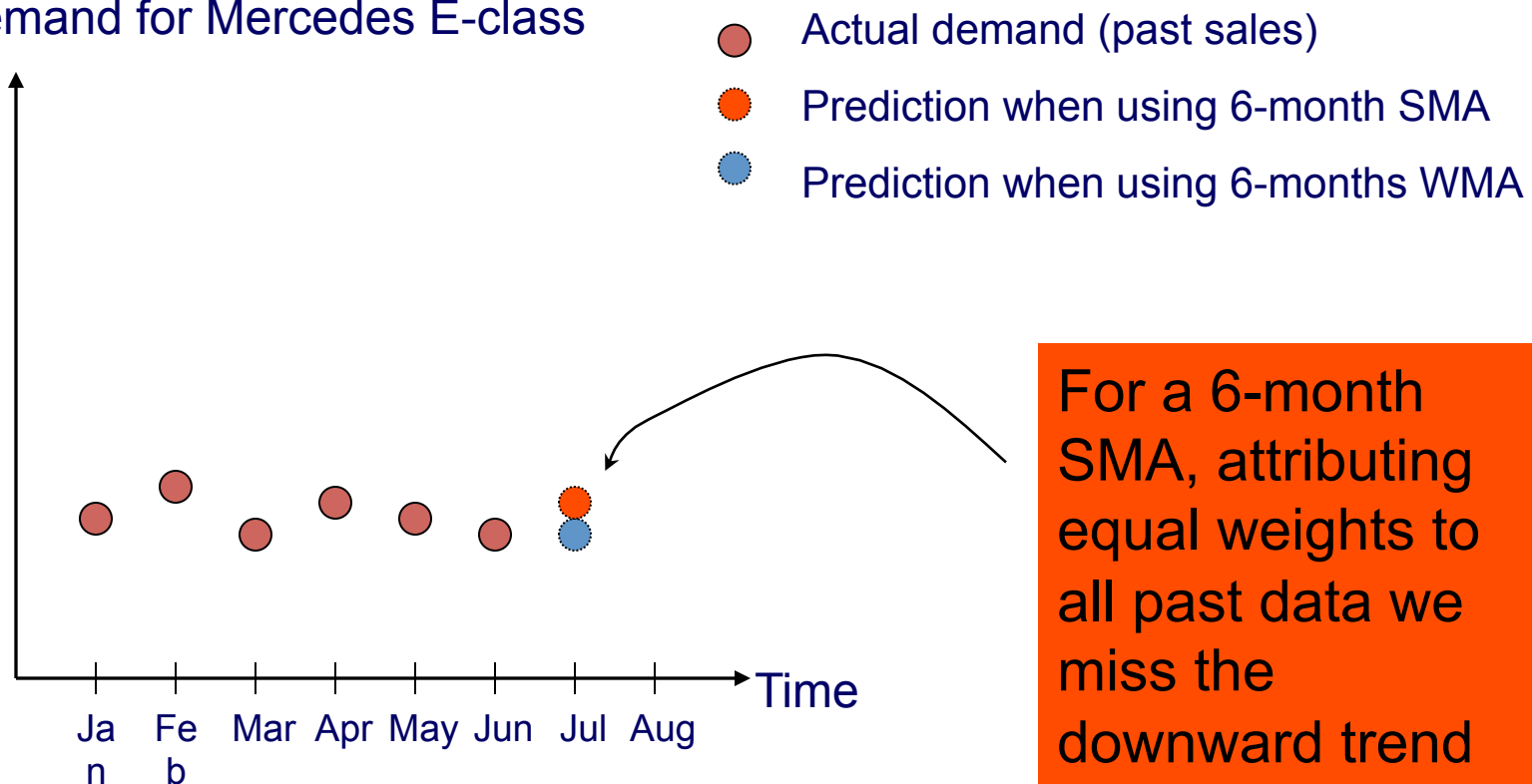
n is the forecasting horizon (how far back we look),

A is the actual sales figure from each period.

Why do we need the WMA models?

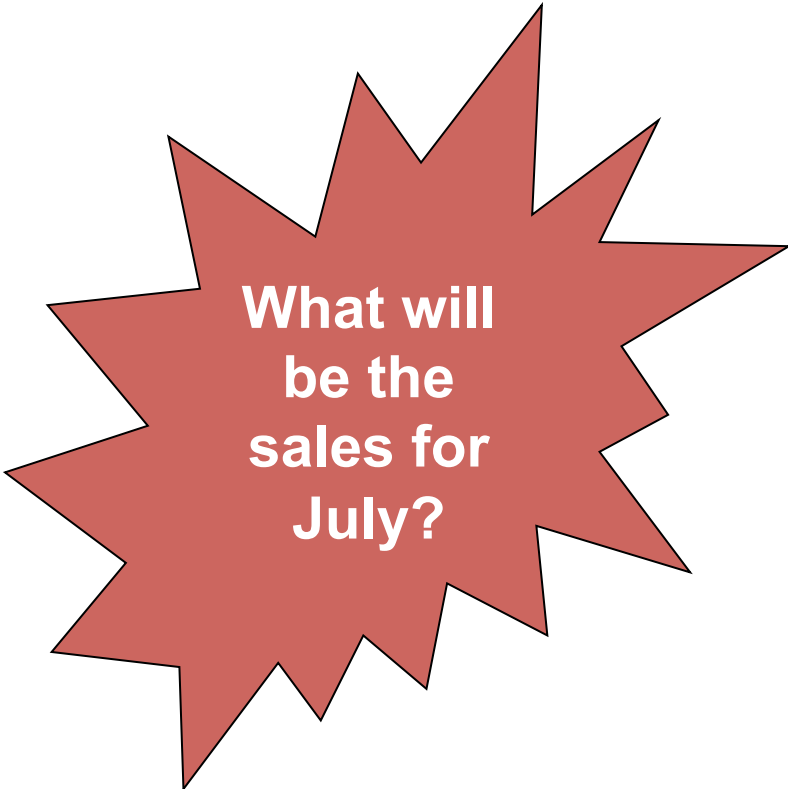
Because of the ability to give more importance to what happened recently, without losing the impact of the past.

Demand for Mercedes E-class



Example: Kroger sales of bottled water

Month	Bottles
<i>Jan</i>	<i>1,325</i>
<i>Feb</i>	<i>1,353</i>
<i>Mar</i>	<i>1,305</i>
<i>Apr</i>	<i>1,275</i>
<i>May</i>	<i>1,210</i>
<i>Jun</i>	<i>1,195</i>
<i>Jul</i>	<i>?</i>



What will
be the
sales for
July?

6-month simple moving average...

$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr} + A_{Mar} + A_{Feb} + A_{Jan}}{6} = 1,277$$

In other words, because we used equal weights, a slight downward trend that actually exists is not observed...

What if we use a weighted moving average?

Make the weights for the last three months more than the first three months...

	6-month SMA	WMA 40% / 60%	WMA 30% / 70%	WMA 20% / 80%
July Forecast	<i>1,277</i>	<i>1,267</i>	<i>1,257</i>	<i>1,247</i>

The higher the importance we give to recent data, the more we pick up the declining trend in our forecast.

How do we choose weights?

1. Depending on the importance that we feel past data has
2. Depending on known seasonality (weights of past data can also be zero).

**WMA is better than SMA
because of the ability to
vary the weights!**

Time Series: Exponential Smoothing (ES)

Main idea: The prediction of the future depends mostly on the most recent observation, and on the error for the latest

**Smoothing constant
alpha α**



Denotes the
importance of the past

Why use exponential smoothing?

1. Uses less storage space for data
2. Extremely accurate
3. Easy to understand
4. Little calculation complexity
5. There are simple accuracy tests

Exponential smoothing: the method

Assume that we are currently in period t . We calculated the forecast for the last period (F_{t-1}) and we know the actual demand last period (A_{t-1}) ...

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

The smoothing constant α expresses how much our forecast will react to observed differences...

If α is low: there is little reaction to differences.

If α is high: there is a lot of reaction to differences.

Example: bottled water at Kroger

Month	Actual	Forecasted
<i>Jan</i>	1,325	1,370
<i>Feb</i>	1,353	1,361
<i>Mar</i>	1,305	1,359
<i>Apr</i>	1,275	1,349
<i>May</i>	1,210	1,334
<i>Jun</i>	?	1,309

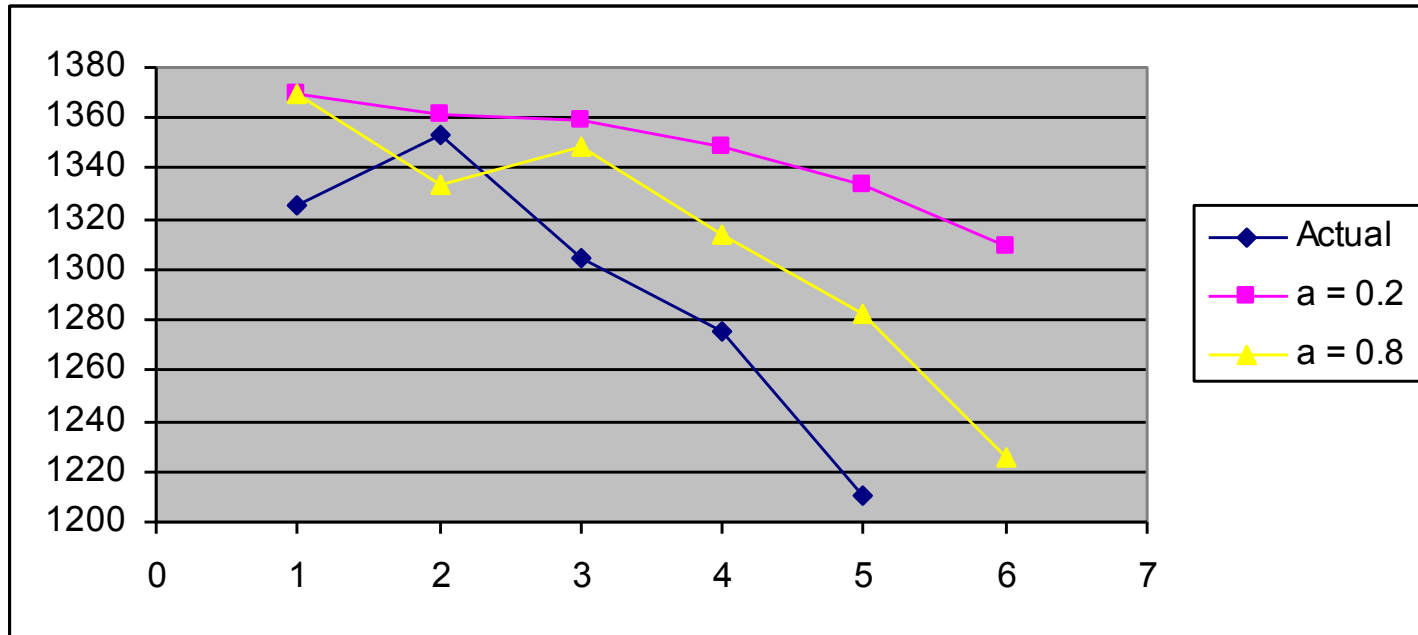
$$\alpha = 0.2$$

Example: bottled water at Kroger

Month	Actual	Forecasted
<i>Jan</i>	1,325	1,370
<i>Feb</i>	1,353	1,334
<i>Mar</i>	1,305	1,349
<i>Apr</i>	1,275	1,314
<i>May</i>	1,210	1,283
<i>Jun</i>	?	1,225

$$\alpha = 0.8$$

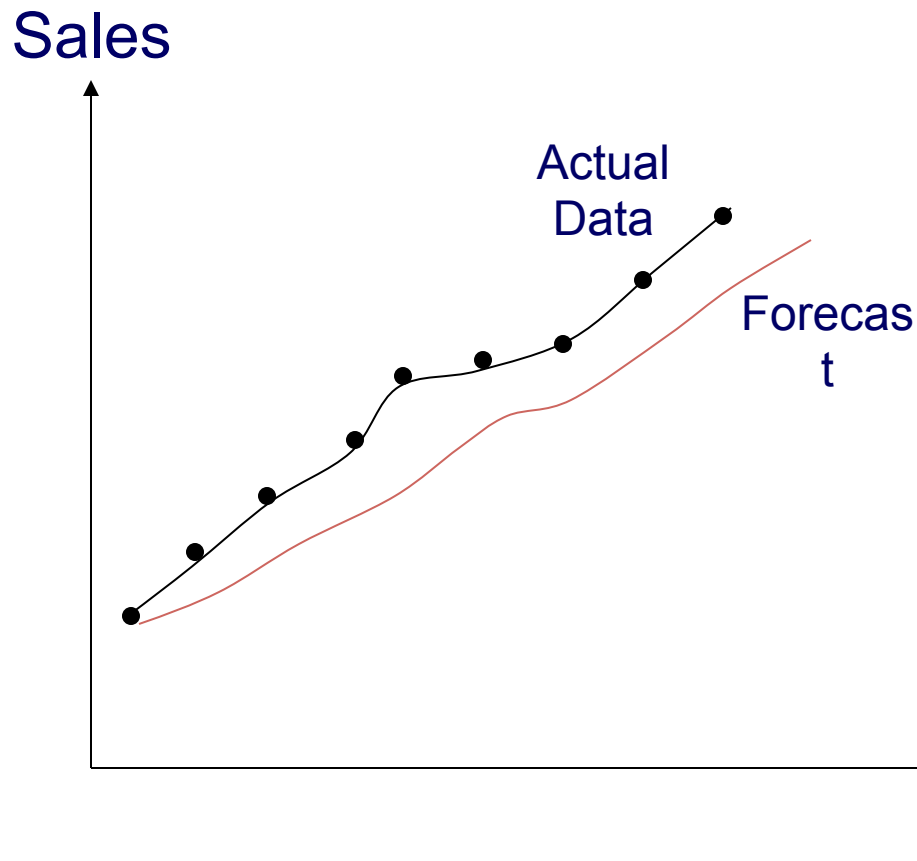
Impact of the smoothing constant



Trend..

What do you think will happen to a moving average or exponential smoothing model when there is a *trend* in the data?

Impact of trend



Regular exponential smoothing will always lag behind the trend.

Can we include trend analysis in exponential smoothing?

Exponential smoothing with trend

FIT: Forecast including trend

$$FIT_t = F_t + T_t$$

$$F_t = FIT_{t-1} + \alpha(A_{t-1} - FIT_{t-1})$$

$$T_t = T_{t-1} + \delta(F_t - FIT_{t-1})$$

The idea is that the two effects are decoupled,
(F is the forecast without trend and T is the trend)

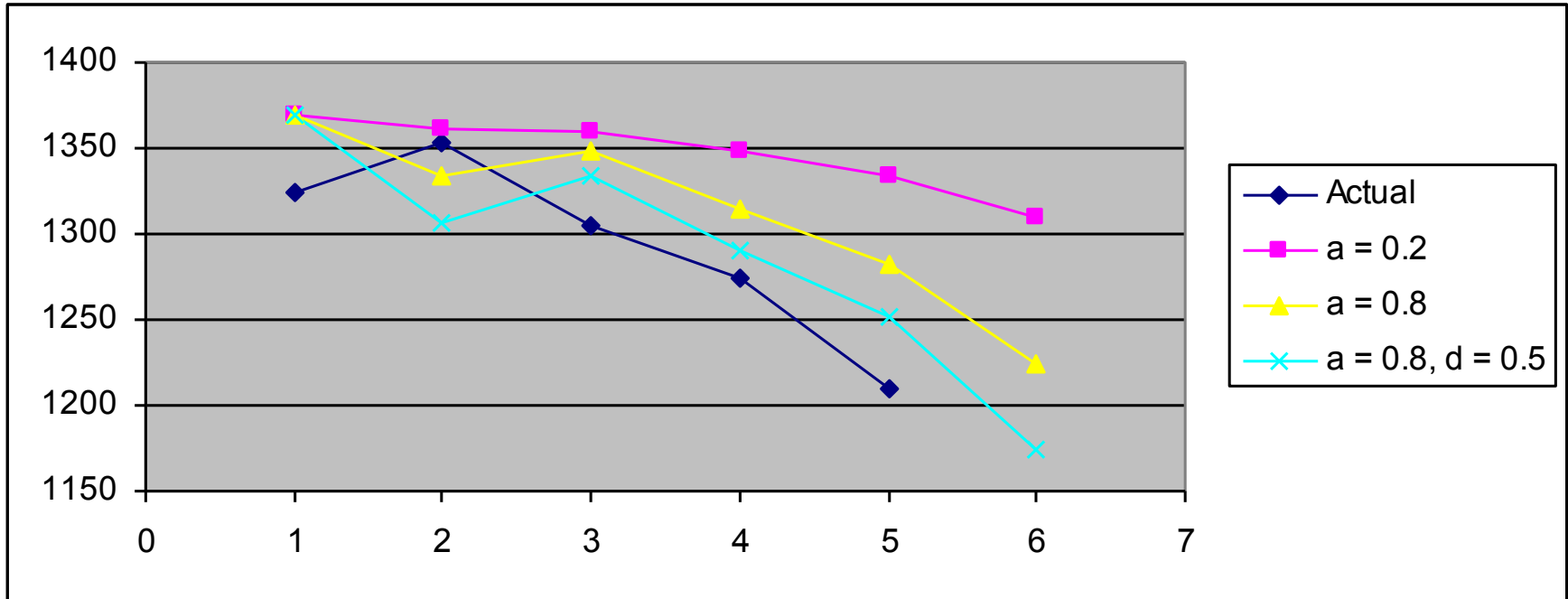
Example: bottled water at Kroger

	A_t	F_t	T_t	FIT_t
Jan	1325	1380	-10	1370
Feb	1353	1334	-28	1306
Mar	1305	1344	-9	1334
Apr	1275	1311	-21	1290
May	1210	1278	-27	1251
Jun		1218	-43	1175

$$\alpha = 0.8$$

$$\delta = 0.5$$

Exponential Smoothing with Trend



Linear regression in forecasting

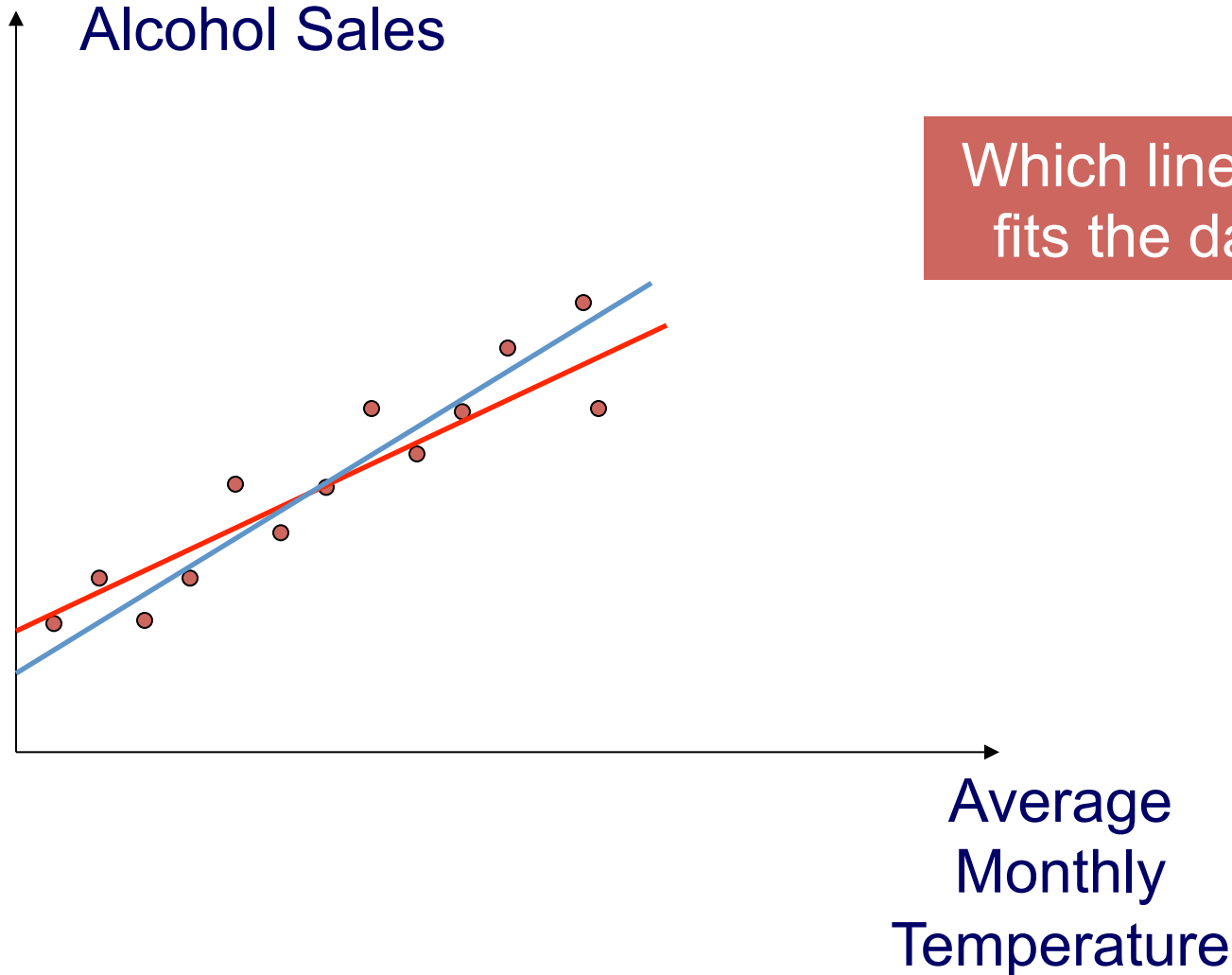
Linear regression is based on

1. Fitting a straight line to data
2. Explaining the change in one variable through changes in other variables.

$$\text{dependent variable} = a + b \times (\text{independent variable})$$

By using linear regression, we are trying to explore which independent variables affect the dependent variable

Example: do people drink more when it's cold?



Which line best fits the data?

The best line is the one that minimizes the error

The predicted line is ...

$$Y = a + bX$$

So, the error is ...

$$\varepsilon_i = y_i - Y_i$$

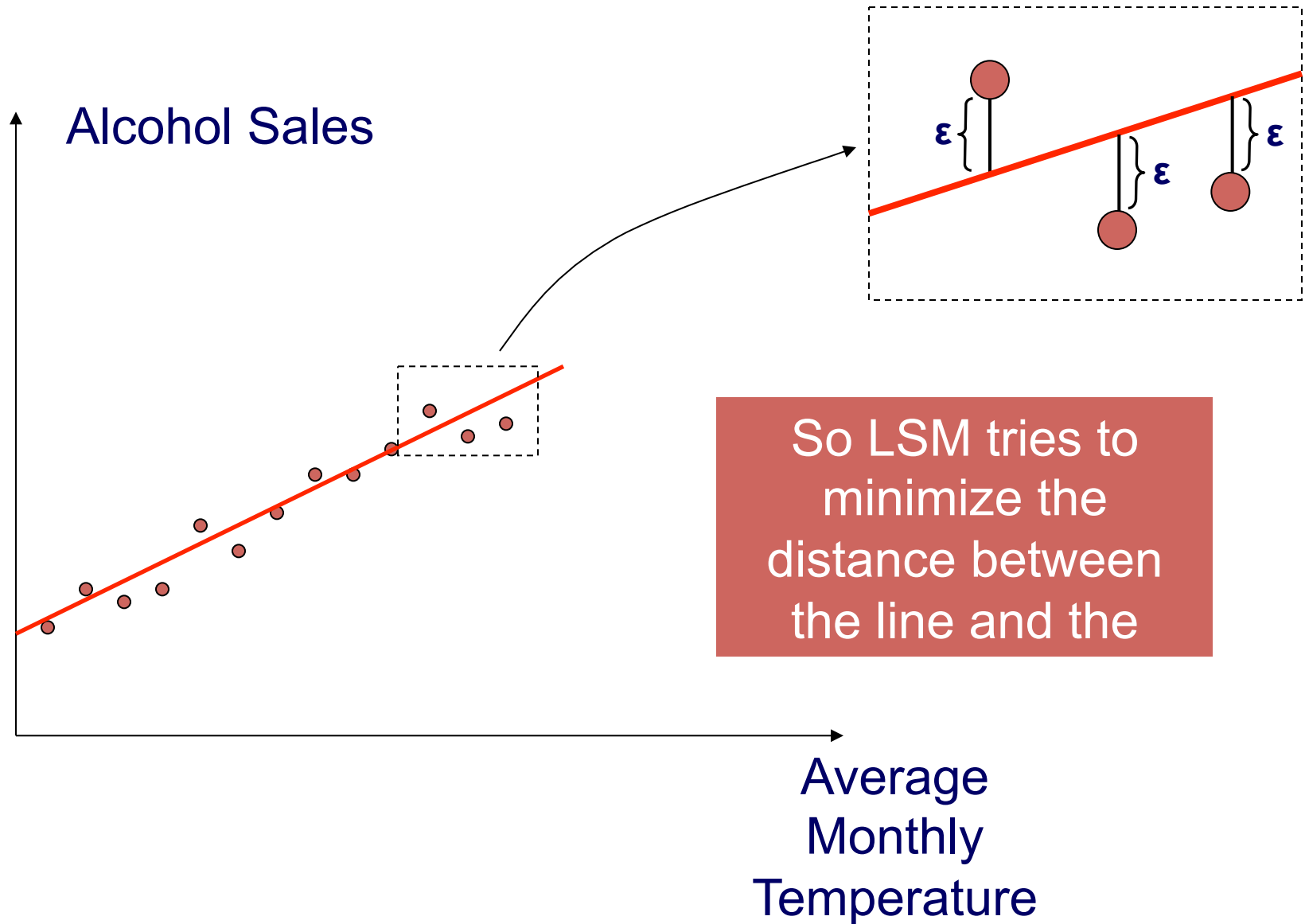
Where: ε is the error
y is the observed value
Y is the predicted value

Least Squares Method of Linear Regression

The goal of LSM is to minimize the sum of squared

$$\text{Min } \sum \varepsilon_i^2$$

What does that mean?



How can we compare across forecasting models?

We need a metric that provides estimation of accuracy

Forecast Error

Errors can be:

1. biased (consistent)
2. random

Forecast error = Difference between actual and forecasted value
(also known as *residual*)

Measuring Accuracy: MFE

MFE = Mean Forecast Error (Bias)

It is the average error in the observations

$$\text{MFE} = \frac{\sum_{i=1}^n A_t - F_t}{n}$$

1. A more positive or negative MFE implies worse performance; the forecast is biased.

Measuring Accuracy: MAD

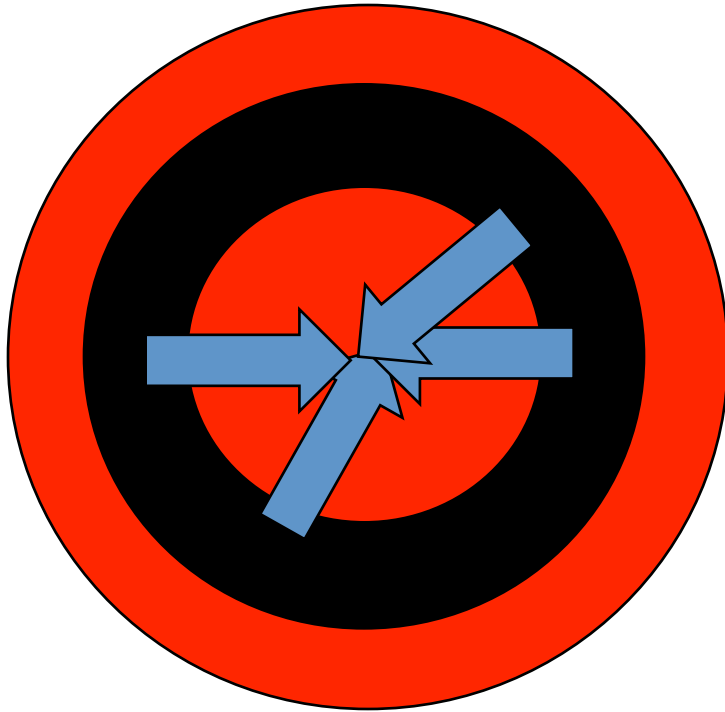
MAD = Mean Absolute Deviation

It is the average absolute error in the observations

$$\text{MAD} = \frac{\sum_{i=1}^n |A_t - F_t|}{n}$$

1. Higher MAD implies worse performance.
2. If errors are normally distributed, then $\sigma_{\varepsilon} = 1.25\text{MAD}$

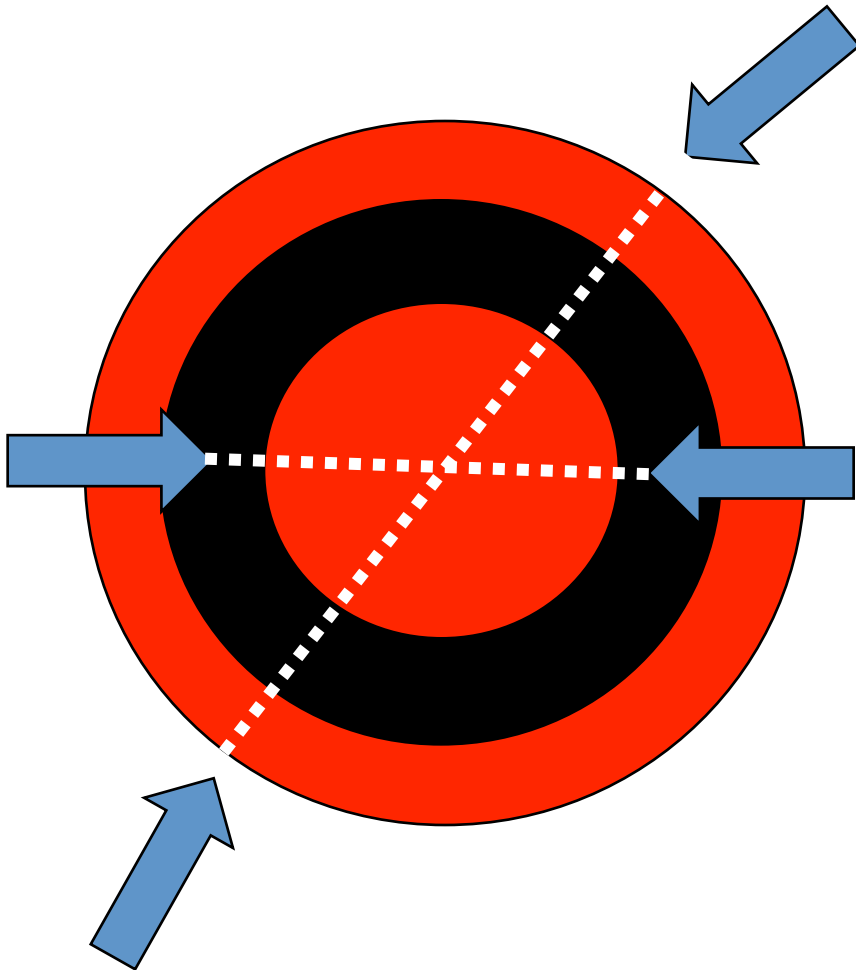
MFE & MAD: A Dartboard Analogy



Low MFE & MAD:

The forecast errors
are small &
unbiased

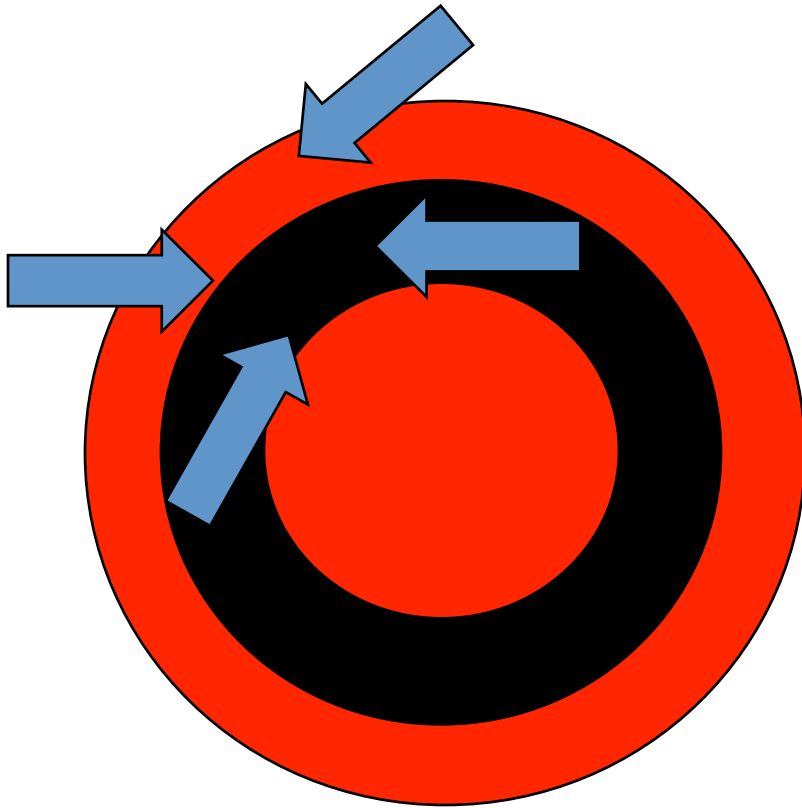
An Analogy (cont' d)



Low MFE but high
MAD:

On average, the
arrows hit the
bullseye (so much
for averages!)

MFE & MAD: An Analogy



High MFE & MAD:

The forecasts
are inaccurate &
biased

Key Point

Forecast must be measured for accuracy!

The most common means of doing so is by measuring the either the mean absolute deviation or the standard deviation of the forecast error

Measuring Accuracy: Tracking signal

The tracking signal is a measure of how often our estimations have been above or below the actual value. It is used to decide when to re-evaluate using a model.

$$\text{RSFE} = \sum_{i=1}^n (A_t - F_t) \qquad \text{TS} = \frac{\text{RSFE}}{\text{MAD}}$$

Positive tracking signal: most of the time actual values are above our forecasted values

Negative tracking signal: most of the time actual values are below our forecasted values

If $\text{TS} > 4$ or < -4 , *investigate!*

Example: bottled water at Kroger

Month	Actual	Forecast
<i>Jan</i>	<i>1,325</i>	<i>1,370</i>
<i>Feb</i>	<i>1,353</i>	<i>1,361</i>
<i>Mar</i>	<i>1,305</i>	<i>1,359</i>
<i>Apr</i>	<i>1,275</i>	<i>1,349</i>
<i>May</i>	<i>1,210</i>	<i>1,334</i>
<i>Jun</i>	<i>1,195</i>	<i>1,309</i>

Exponential Smoothing
($\alpha = 0.2$)

Month	Actual	Forecast
Jan	1,325	1370
Feb	1,353	1306
Mar	1,305	1334
Apr	1,275	1290
May	1,210	1251
Jun	1,195	1175

Forecasting with trend
($\alpha = 0.8$)
($\delta = 0.5$)

Question: Which one is better?

Bottled water at Kroger: compare MAD and TS

	MAD	TS
Exponential Smoothing	70	- 6.0
Forecast Including Trend	33	- 2.0

We observe that FIT performs a lot better than ES

Conclusion: Probably there is trend in the data which Exponential smoothing cannot capture

Which Forecasting Method Should You Use

- Gather the historical data of what you want to forecast
- Divide data into initiation set and evaluation set
- Use the first set to develop the models
- Use the second set to evaluate
- Compare the MADs and MFEs of each model