Foundation of Data Science
Lecture 8, Module 1
Fall 2022

Rumi Chunara, PhD

<u>Fine Print</u>: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Today

- Time-series Analyses
- Regression and lagged data

Time Series Discussions

- Overview
- Basic definitions
- Time domain

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

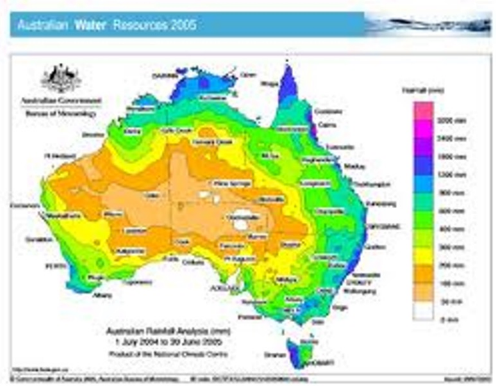
A sequence of random variables

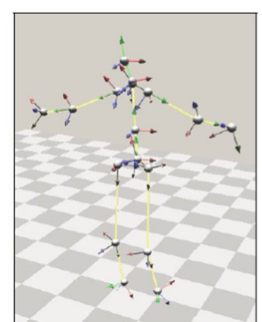
$x_1, x_2, x_3, \ldots$

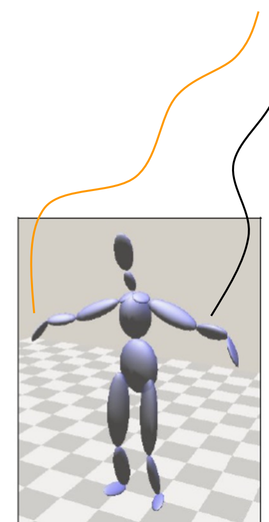
Time Series Examples

Definition: *A sequence of measurements over time*

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



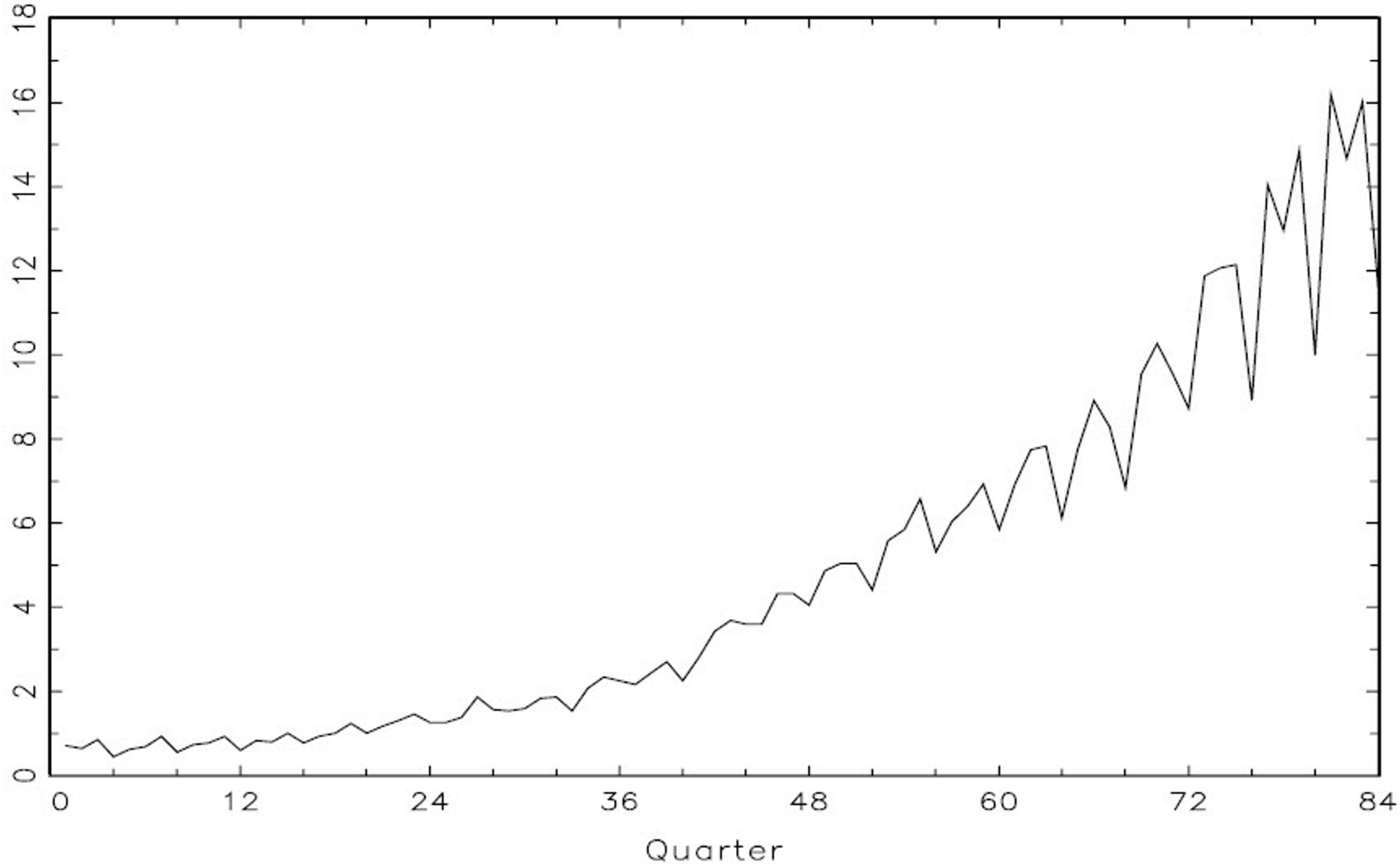




Three Approaches

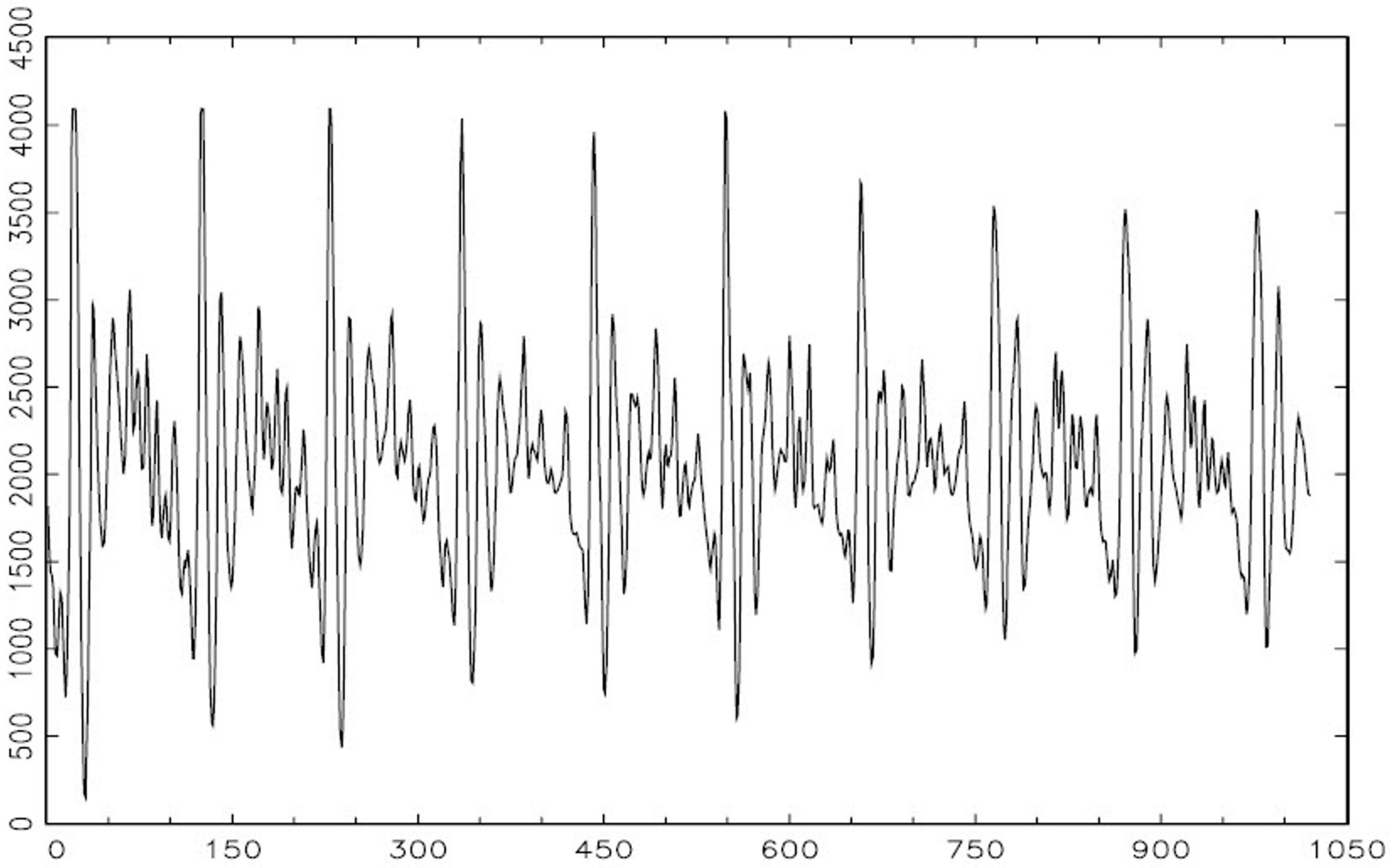
- Time domain approach
  - Analyze dependence of current value on past values
- Frequency domain approach
  - Analyze periodic sinusoidal variation (sine wave)
- State space models
  - Represent state as collection of variable values
  - Model transition between states

Sample Time Series Data



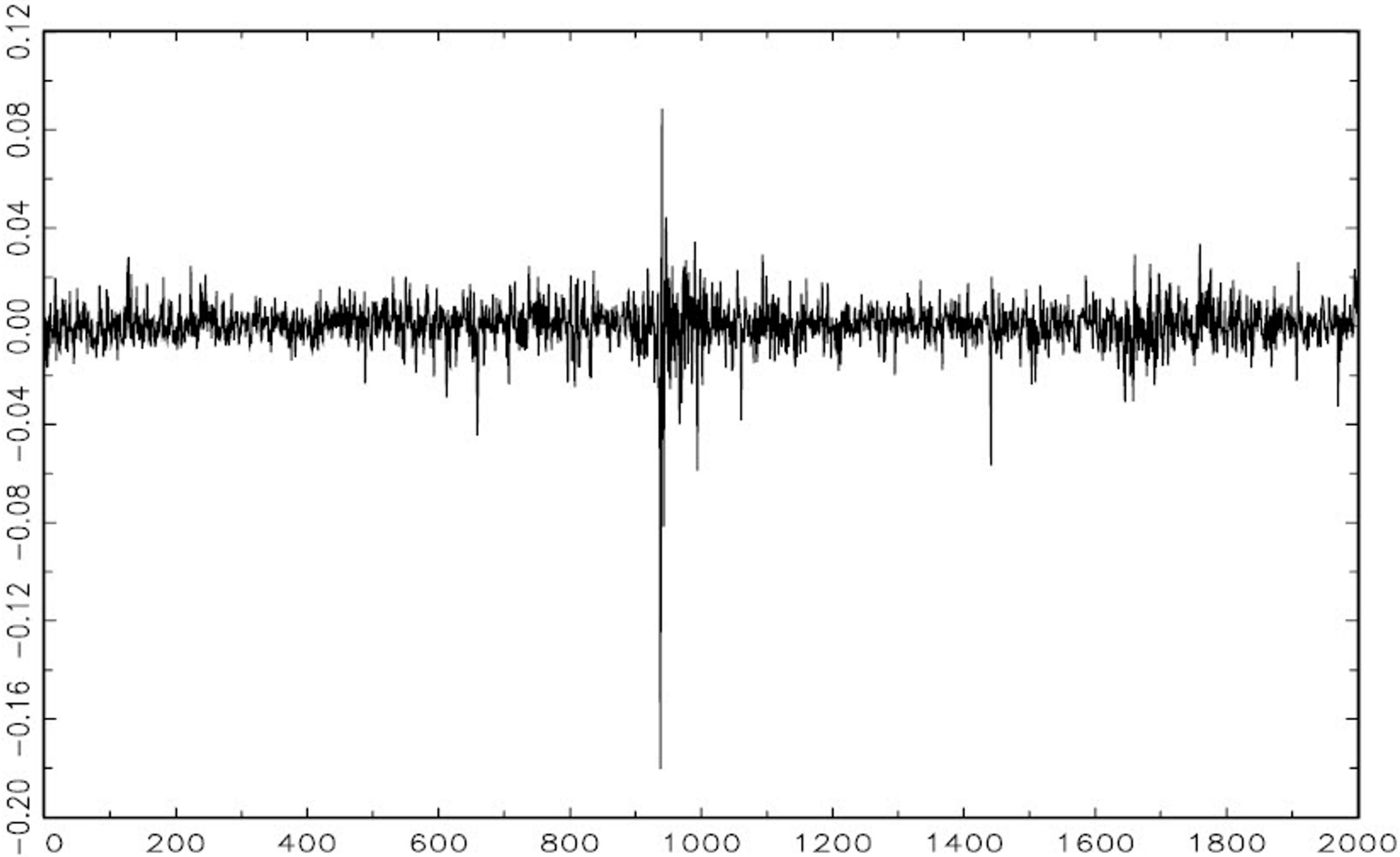
Johnson & Johnson quarterly earnings/share, 1960-1980

Sample Time Series Data



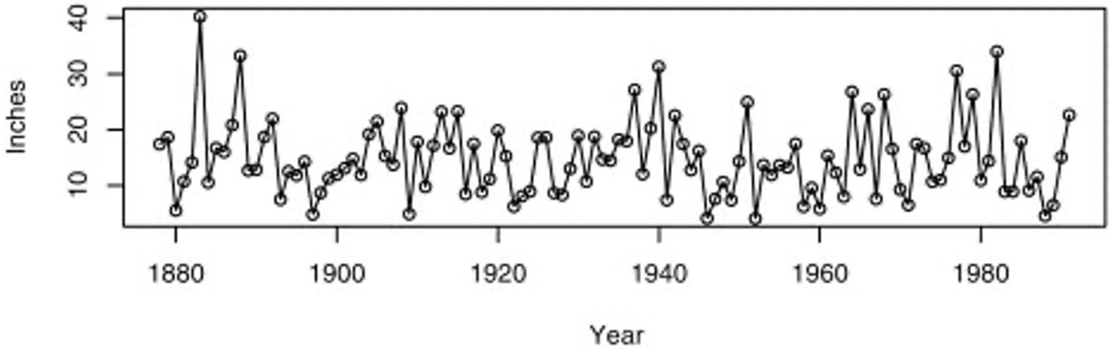
Speech recording of “aaa…hhh”, 10k pps

Sample Time Series Data

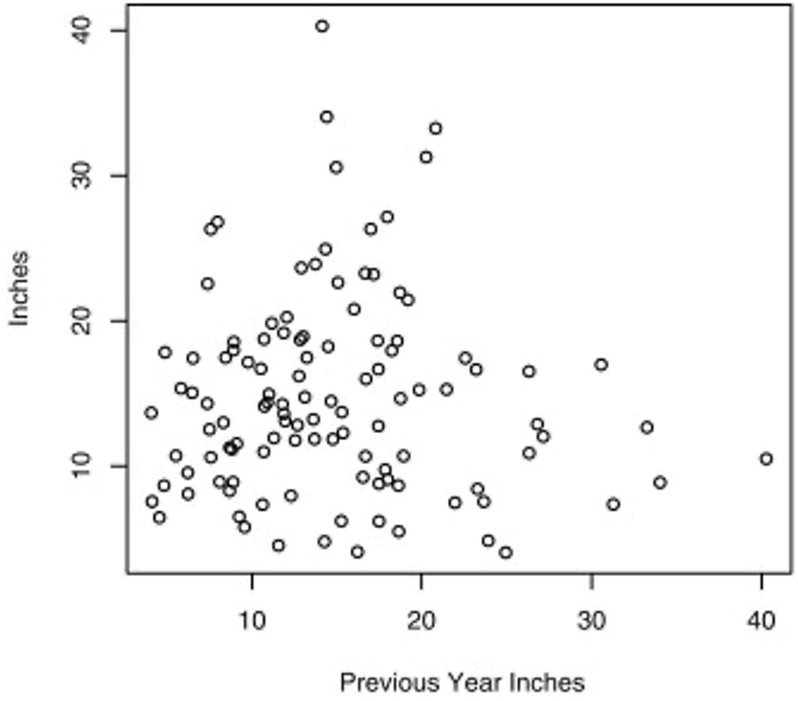


NYSE daily weighted market returns 2/2/84 - 12/31/92

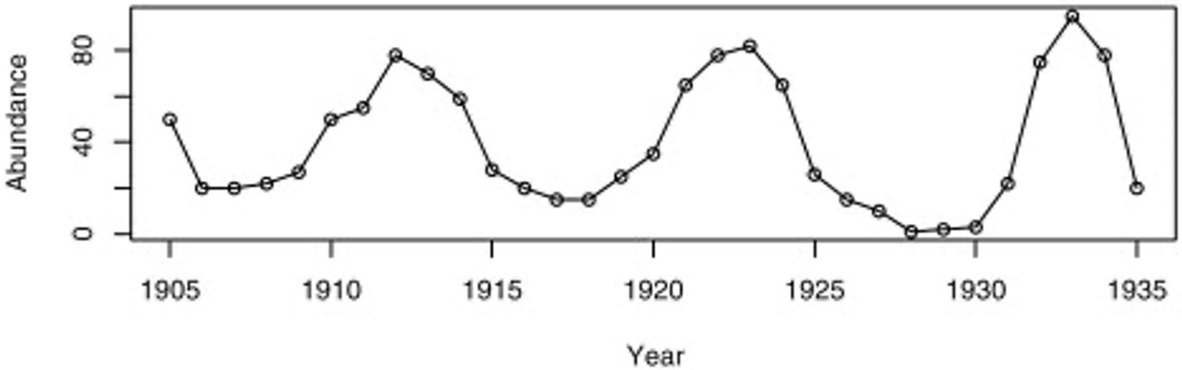
Not all time data will exhibit strong patterns



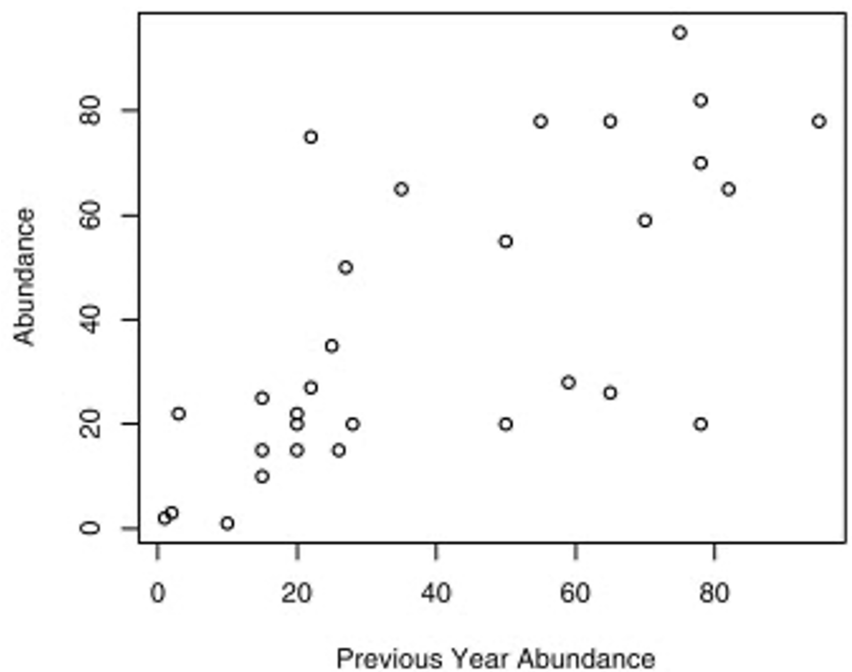
LA annual rainfall



…and others will be apparent



Canadian Hare counts



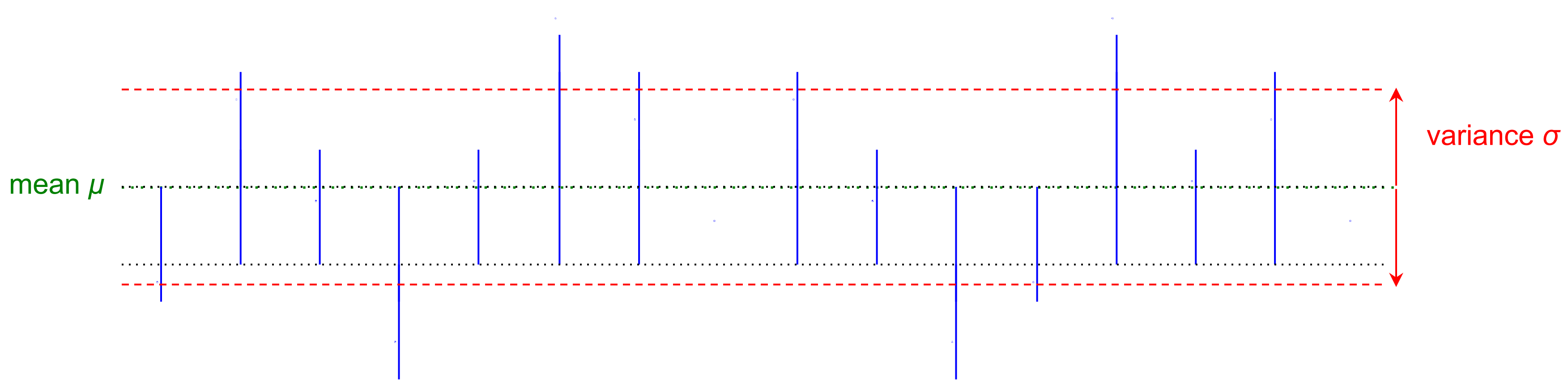
Time Series Discussions

- Overview
- Basic definitions
- Time domain

Original Definitions - Random Variables

- Mean $\mu \equiv \mathsf{E}[x_t] := \frac{1}{N}\sum_{t=1}^{N} x_t$

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



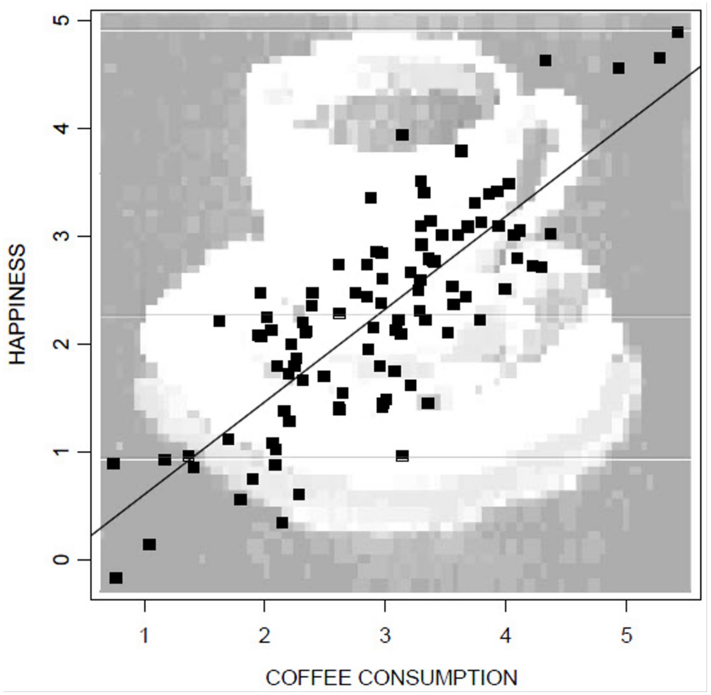
Original Definitions - Random Variables

- Covariance

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

- Correlation

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



Redefined for Time

Mean function

$$\mu_X(t) = E(X_t) \quad for\, t = 0, \pm 1, \pm 2, \ldots$$

- Ergodic process: if the mean computed over any time *t* is the same as the ensemble mean, then the process is ergodic
  - Example: tossing a fair coin multiple times (random variable X_t can be 0 or 1)

Redefined for Time

- Autocovariance - how X_t relates to its previous values

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

lag

- Autocorrelation

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

Durbin-Watson test provides measures of significance for autocorrelation.

Metrics for Unknown Distributions

Sample Mean function

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

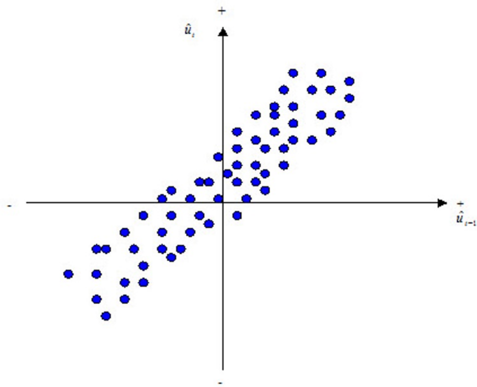
Sample Autocovariance

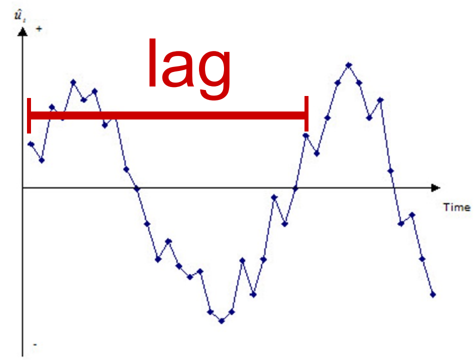
$$\hat{\gamma}(h) := n^{-1}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n$$

Sample Autocorrelation

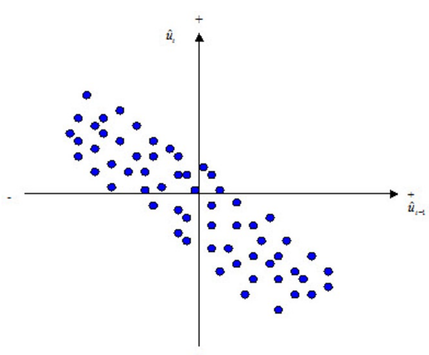
$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n$$

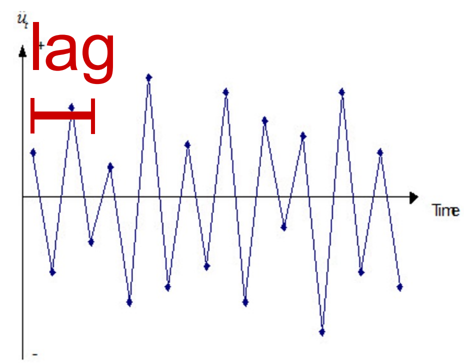
Autocorrelation Examples





Positive



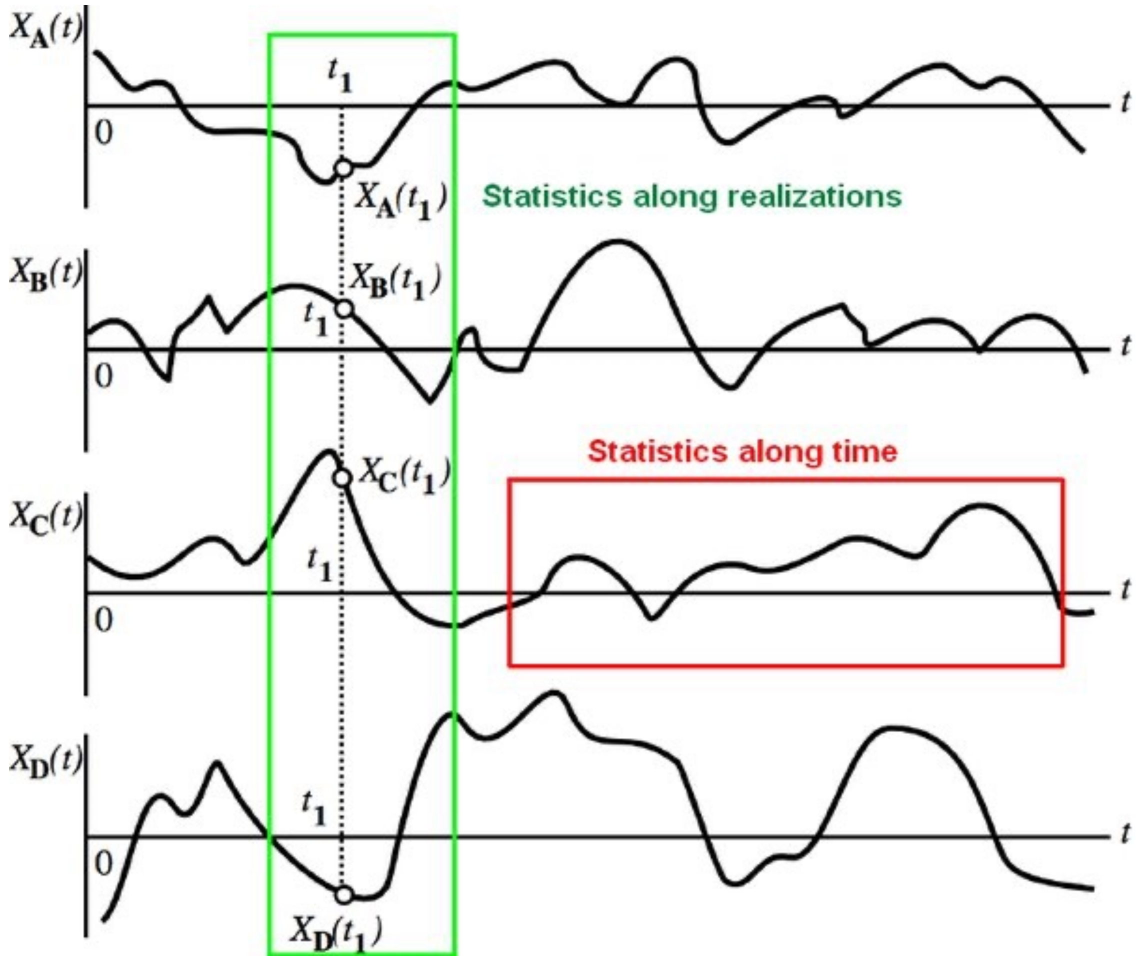


Negative

Stationarity Time Series

- $\{X_t\}$ is stationary if
  - $\mu_X(t)$ is independent of t
  - $\gamma_X(t+h,t)$ is independent of t for each h
- Special case: white noise
  - $\{X_t\}$ is a sequence of **uncorrelated** random variables, each with constant mean and variance
- Stationary series are much easier to forecast with
  - Much of time series analysis involves trying to reduce a complicated series to a stationary one

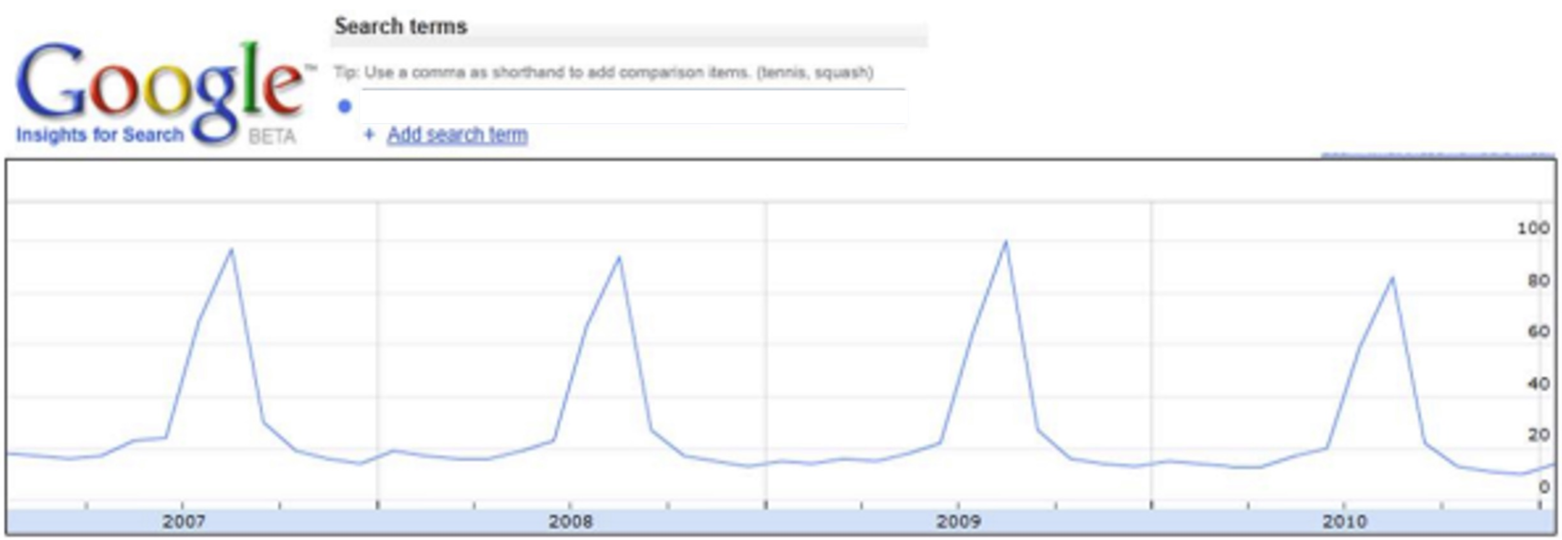
Stationary vs. Ergodic



https://www.quora.com/What-is-the-difference-between-a-stationary-ergodic-and-a-stationary-non-ergodic-process#:~:text=Stationary%20non%2Dergodic%20processes%20do,between%20temporal%20and%20ensemble%20statistics.&text=For%20an%20ergodic%20process%2C%20you,representative%20of%20the%20random%20process.

What if a linear trend does not fit my data?

- Could be no relationship
- Could be too much local variation. In that case:
  - Look at longer-term trends
  - Smooth the data
  - Check for nonlinear relationships



Moving Average

- Compute an average of the last *m* consecutive data points
  - 4-point moving average is

$$\bar{x}_{MA(4)} = \frac{(x_t + x_{t-1} + x_{t-2} + x_{t-3})}{4} \qquad m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

- Smooths white noise
- Exponential smoothing
  - Higher weights to more recent times

Power Load Data