LOVELY PROFESSIONAL UNIVERSITY

NAAC GRADE A++

INT395- SUPERVISED ML

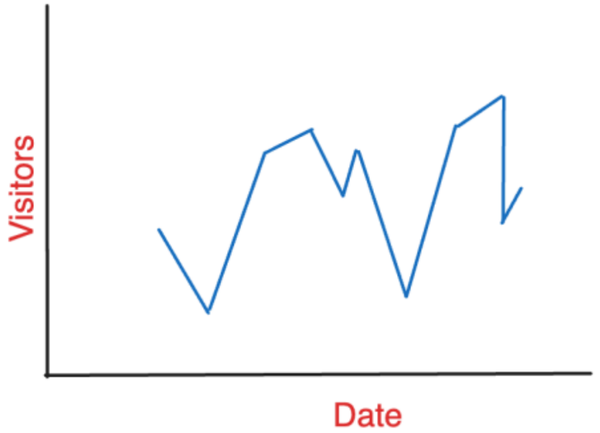
Unit 5: Time Series Regression

Presented By: Blossom Kaler
Assistant Professor
SCAI, LPU

WHAT IS TIME SERIES DATA?

- Data collected sequentially over time
- Each observation is associated with a timestamp
- Order of observations is crucial
- Past values influence future values
- Cannot be randomly shuffled like tabular data
- Widely used in forecasting problems
- Common in finance, healthcare, IoT, and weather
- Time index can be seconds, days, months, or years

Date	Visitors
01/02/23	100
02/02/23	80
03/02/23	150
.	.
.	.
.	.

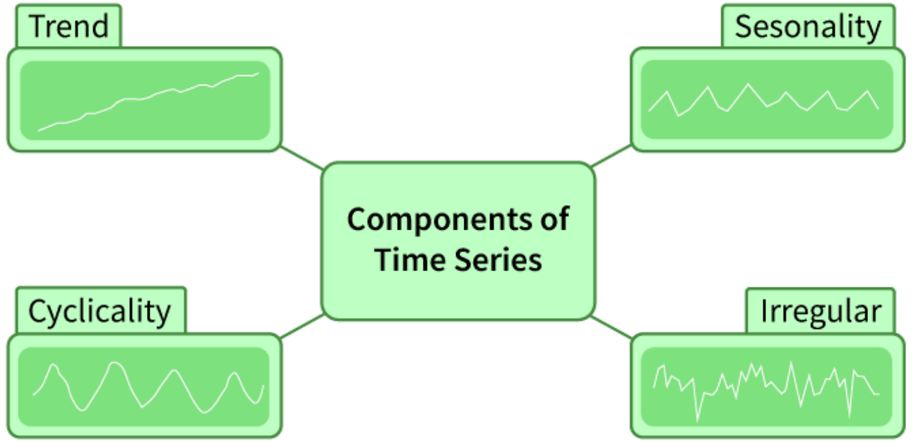


KEY CHARACTERISTICS

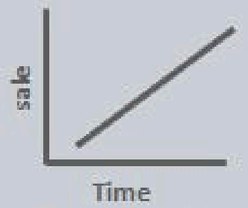
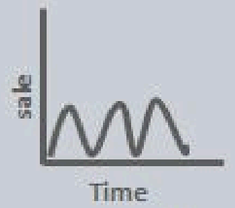
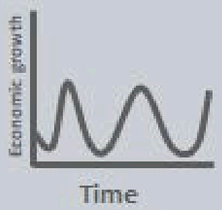
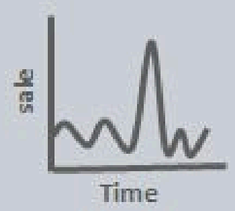
- **Chronological order**: Data is arranged strictly by time
- **Sequential order**: Each value depends on previous values
- **Temporal components**: Includes trend, seasonality, and noise
- **Constant frequency**: Observations recorded at fixed intervals
- **Dynamic nature**: Patterns change over time
- **Time dependency**: Past values influence future values

TEMPORAL COMPONENTS

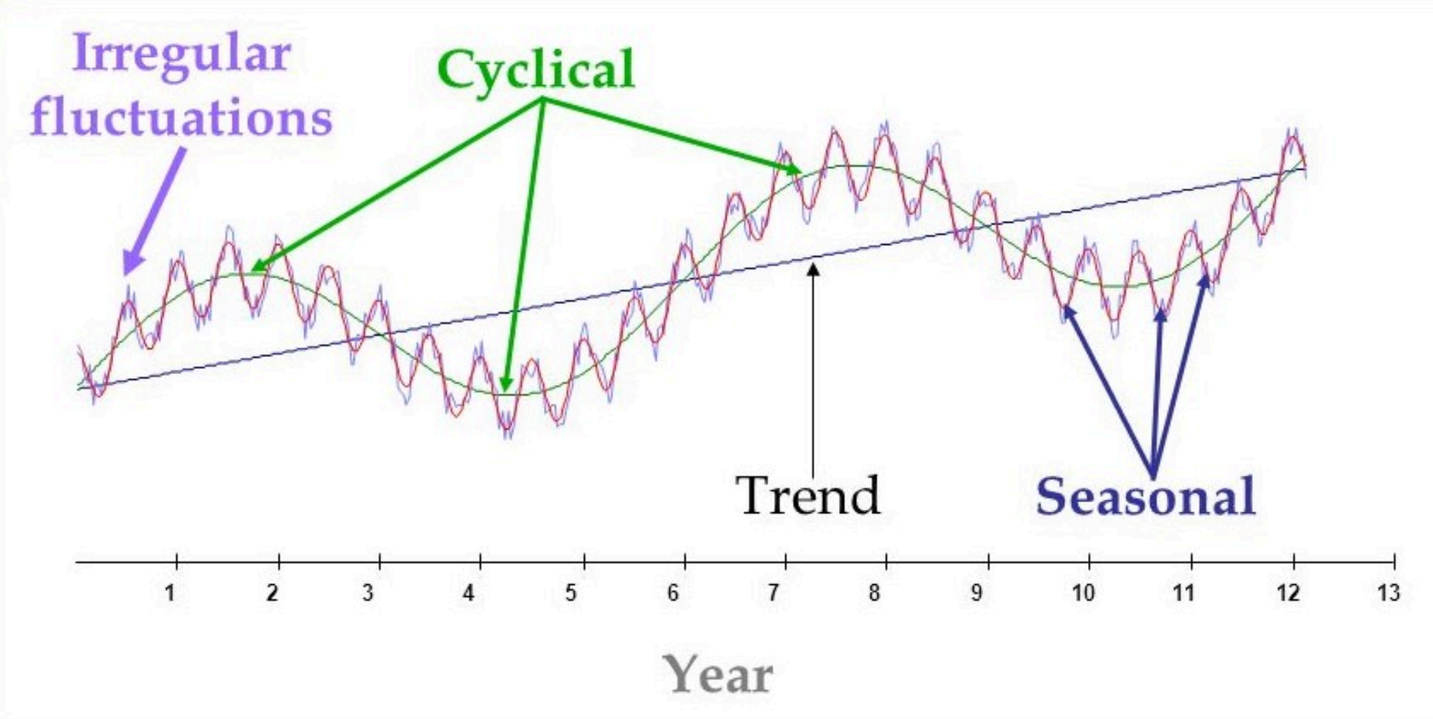
- Temporal components in time series analysis are the underlying patterns that make up data over time
- **Trend:** long-term increase or decrease
- **Seasonality**: repeating patterns at fixed intervals
- **Cyclic behavior:** long-term oscillations (non-fixed)
- **Stationarity**: constant mean and variance over time
- **Irregular component**: random or unpredictable variation
- **Noise**: measurement errors or randomness
- Components can exist together
- Visualization is key for component analysis



TEMPORAL COMPONENTS

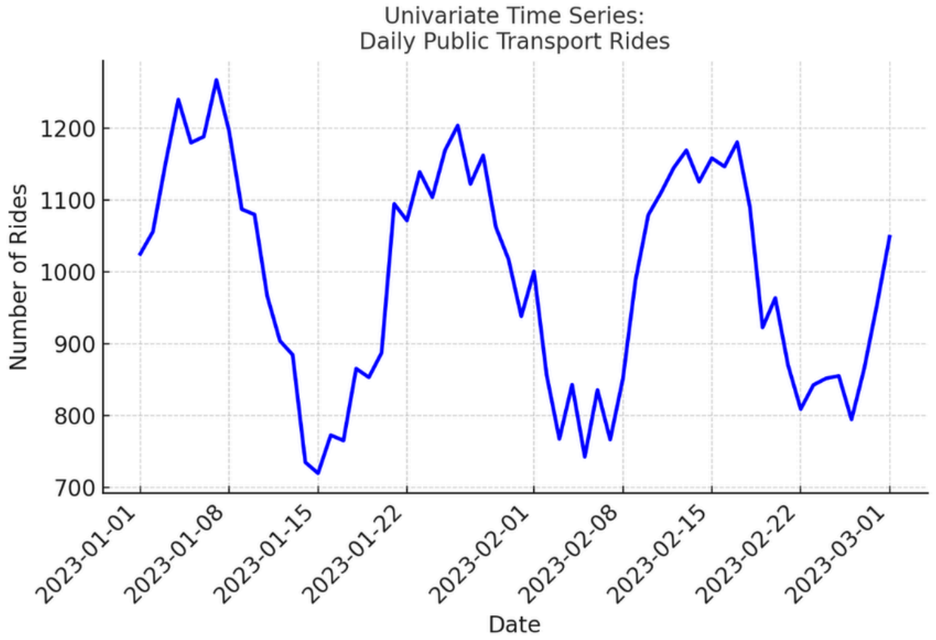
Topic	Trend	Seasonality	Cyclic	Irregularity
Fixed Time Interval	Yes	Yes	No	No
Movement Type	Long/Short Term	Short term	Long/Short Term	Random/Irregular
Graph	 Positive Trend	 Seasonality	 Cyclic	 Irregular
Example	Growth in sales over time	Growth of ice-cream sales every summer	Fluctuation of economic growth	Price of the stock over time

TIME SERIES ANALYSIS



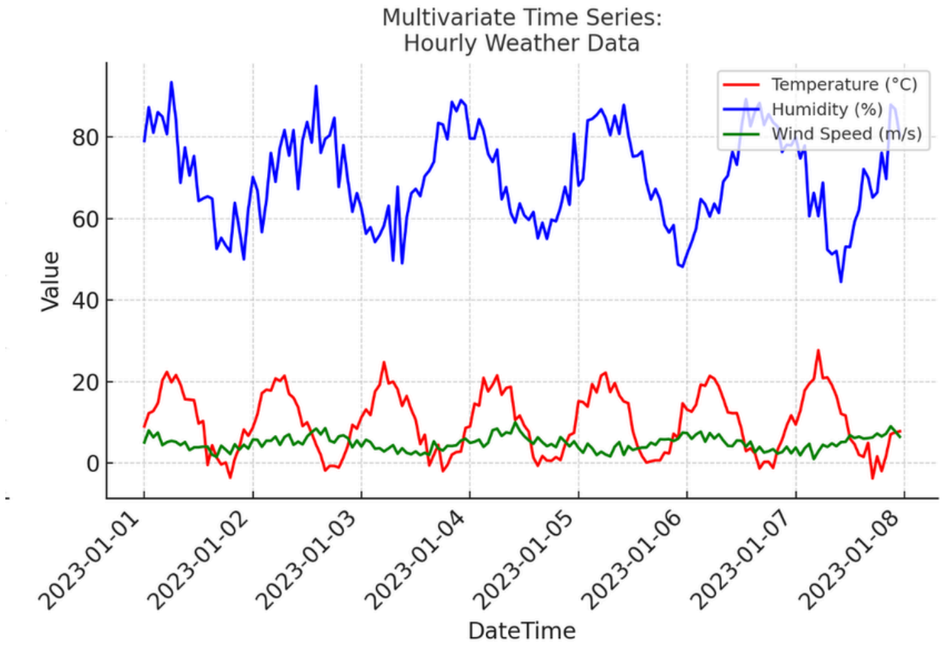
UNIVARIATE TIME SERIES

- Contains only one variable
- Example: daily temperature, monthly sales
- Easier to analyze and model
- Uses past values of the same variable
- Common in basic forecasting tasks
- **Limitations**:
- Cannot capture variable interactions
- Lower predictive power in complex systems
- Performance depends heavily on history length
- Sensitive to noise

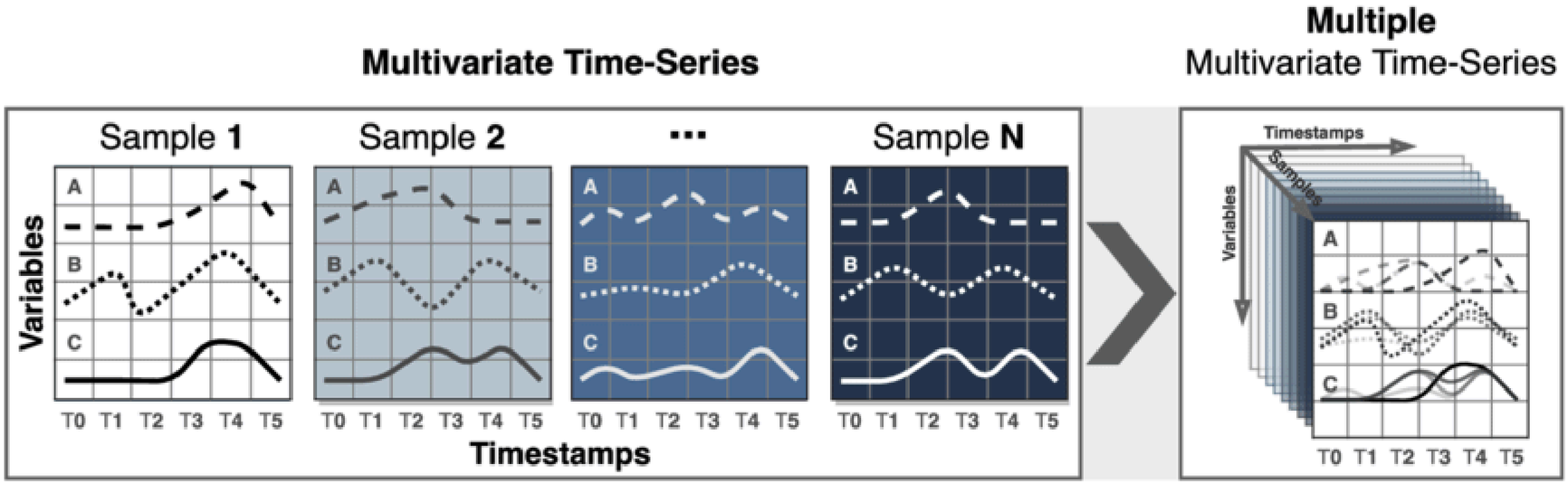


MULTIVARIATE TIME SERIES

- Multiple variables observed over time
- Variables influence each other
- Example: sales, marketing spend, price
- More realistic representation
- Can capture complex dependencies
- Requires advanced models
- Higher computational cost
- Difficult interpretation

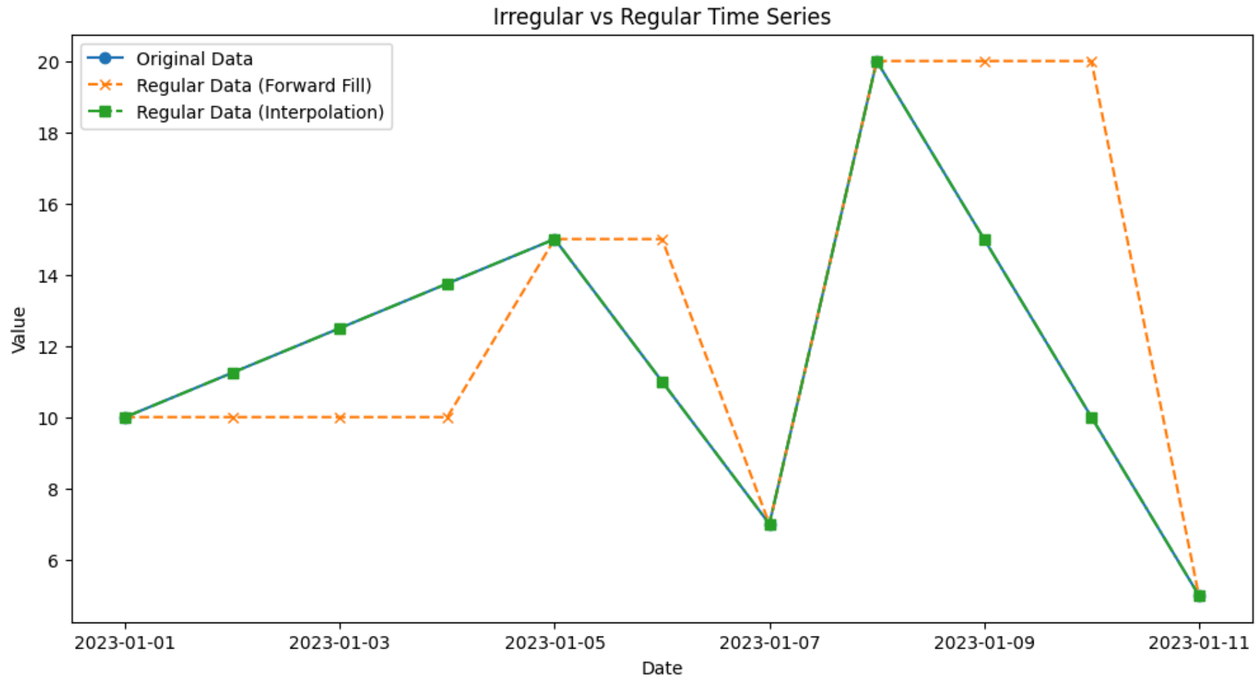


MULTIVARIATE TIME SERIES



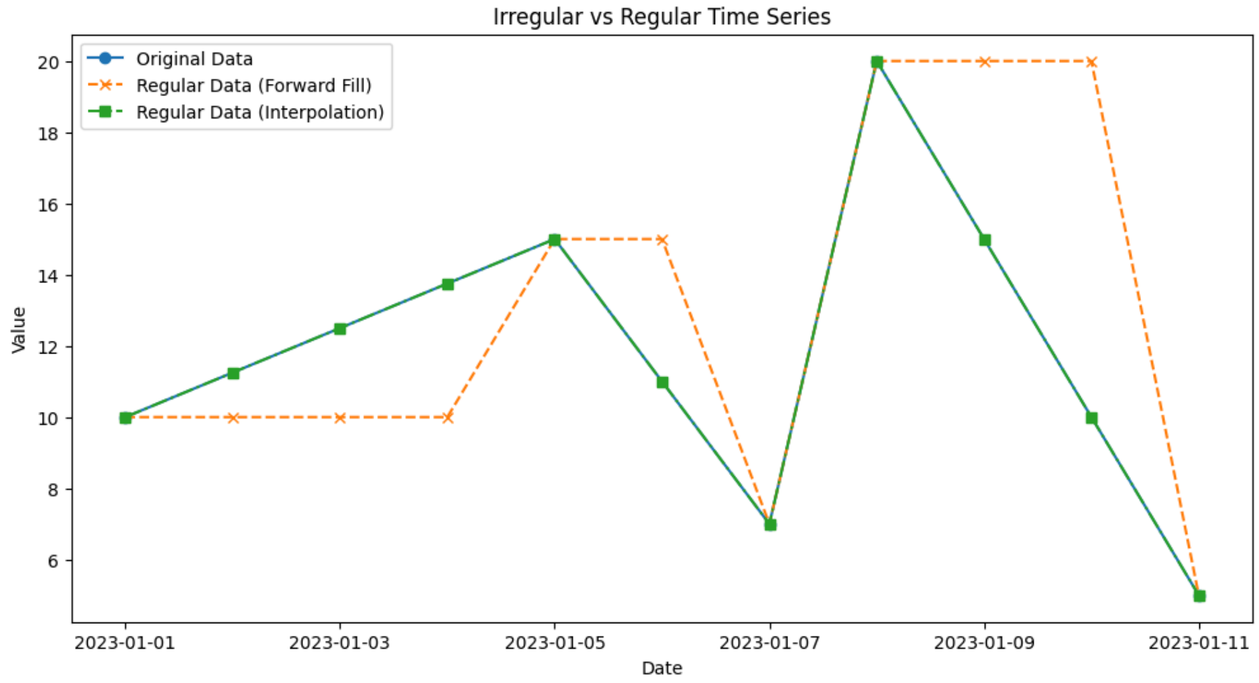
TIME SERIES-REGULAR INTERVALS

- Observations recorded at equal time gaps
- Examples: hourly temperature, monthly revenue
- Simple indexing and alignment
- Required for classical models
- Easy to plot and analyze
- Supports lag-based features
- Simplifies forecasting
- Preferred data format



TIME SERIES-IRREGULAR INTERVALS

- Time gaps between observations vary
- Example: medical event logs
- Missing or inconsistent timestamps
- Cannot directly apply ARIMA
- Requires preprocessing
- Common in real-world data
- Needs careful handling
- Time gaps affect modeling



PREPARING TIME SERIES DATA

- Time series data is ordered and time-dependent
- Raw data often contains **missing timestamps**
- May have **irregular time intervals**
- **Noise and outliers** affect forecasting accuracy
- Many models assume stationarity
- Proper preparation prevents data leakage
- Improves model performance and reliability

UNDERSTANDING TIME INDEX

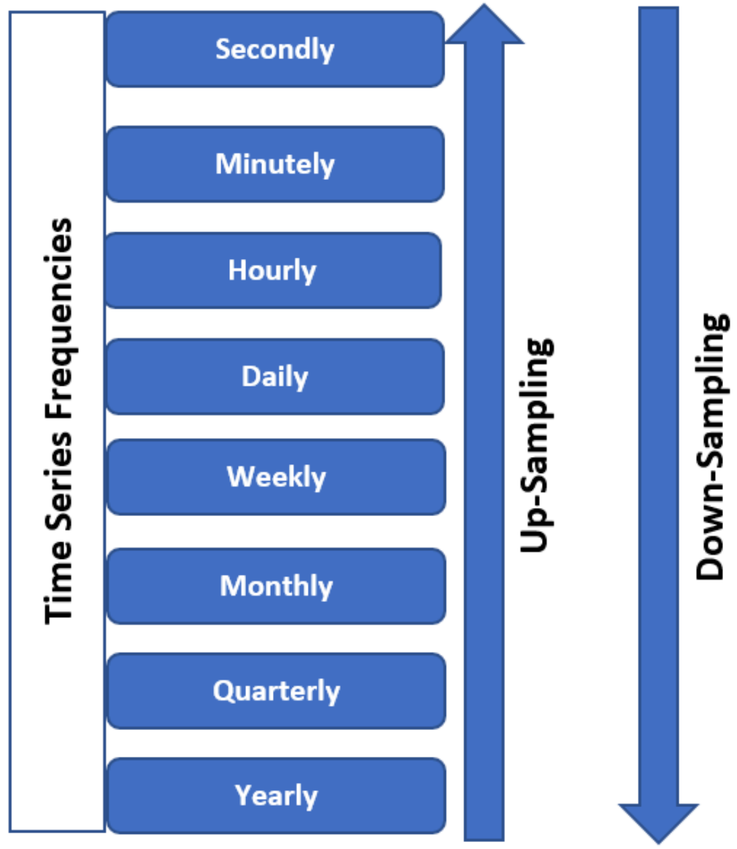
- Convert time column to datetime format
- Set datetime as the index
- Ensure chronological order
- Check frequency (daily, monthly, hourly)
- Detect gaps in timestamps
- Verify timezone consistency
- Unique timestamp per observation
- Foundation for all further steps

	Month	Passengers
1	1949-01	112
2	1949-02	118
3	1949-03	132
4	1949-04	129
5	1949-05	121
6	1949-06	135
7	1949-07	148
8	1949-08	148
9	1949-09	136
10	1949-10	119
11	1949-11	104
12	1949-12	118

RESAMPLING TO FIXED FREQUENCY

- Resampling means converting an irregular time series into a regularly spaced time series.
- Many time series models require data at constant time intervals
- Can be done using **aggregation** or **interpolation**
- May introduce missing values after resampling
- Essential preprocessing step for irregular time series

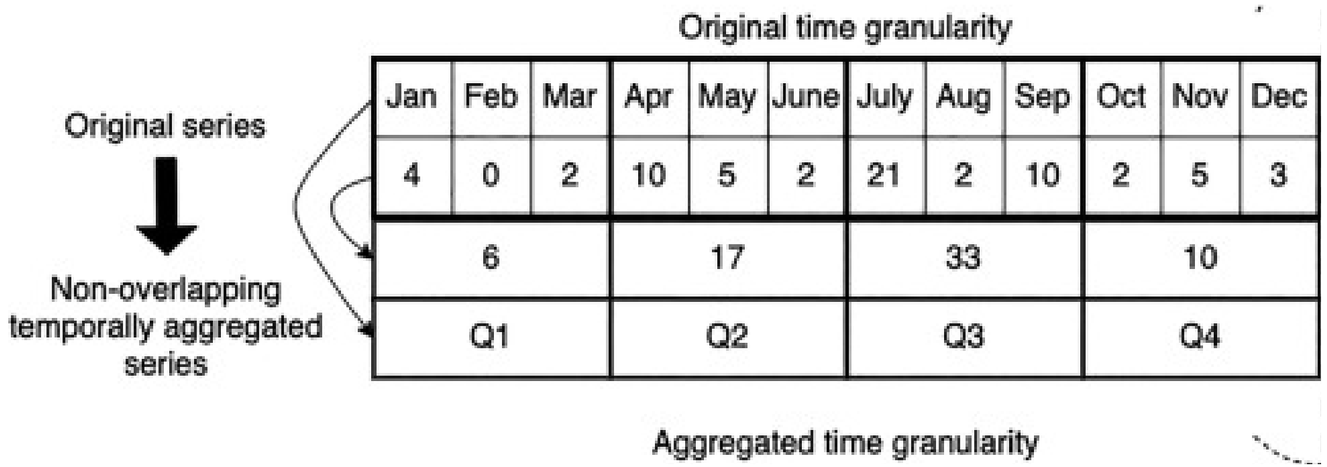
- **Example**:
- Customer purchases occurring at random times → resampled to daily sales



AGGREGATION (MEAN, SUM, COUNT)

- Aggregation summarizes multiple observations into one value per time interval.
- **Mean**: average value in the interval (used for temperature, stock price)
- **Sum**: total value in the interval (used for sales, rainfall)
- **Count**: number of events in the interval (used for transactions)
- Smoothens short-term fluctuations
- Useful when high-frequency data is noisy
- Choice depends on data meaning

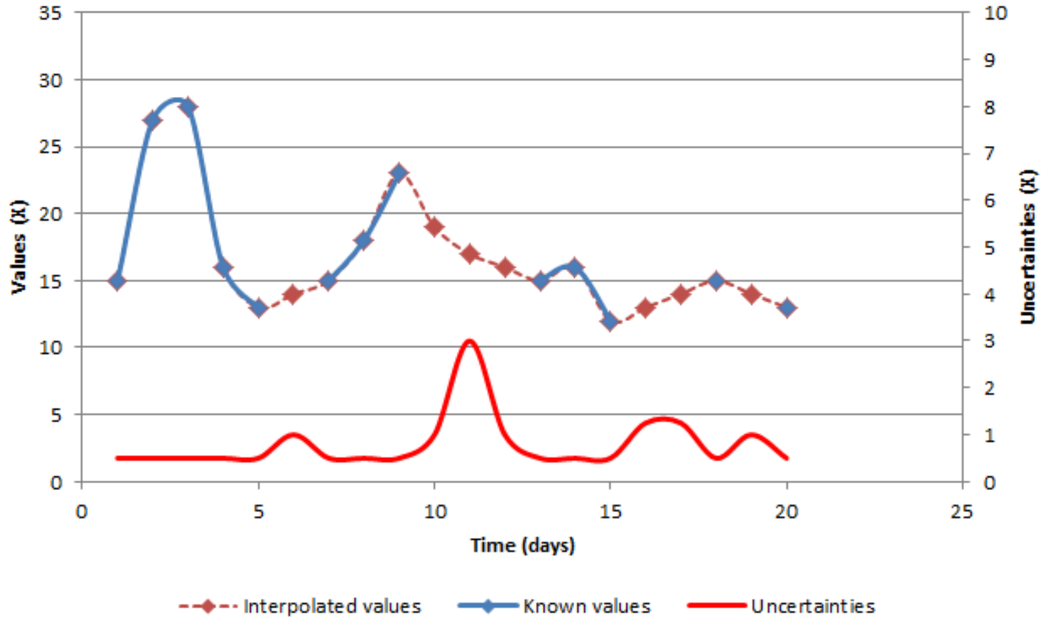
- **Example**:
- Hourly sales → daily total sales using sum



INTERPOLATION METHODS

- **Interpolation estimates missing values using surrounding data points.**
- Used when data points are missing at certain timestamps
- Assumes smooth change between observations
- **Common methods: linear, polynomial**
- Preserves continuity of data
- Works well for gradual changes
- Not suitable for sudden jumps or spikes

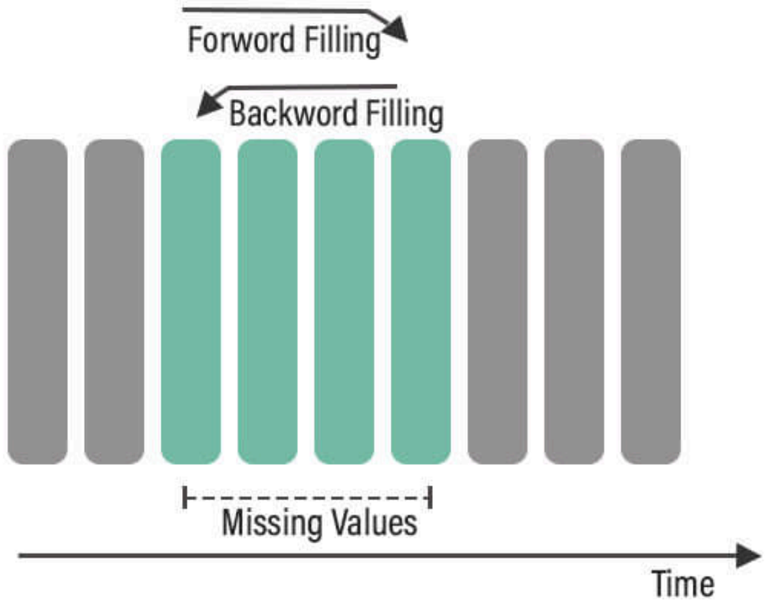
- **Example**:
- Missing temperature for a day → estimated using nearby days



FORWARD AND BACKWARD FILLING

- These methods fill missing values using nearby known values.
- **Forward fill**: uses the last observed value
- **Backward fill**: uses the next observed value
- Simple and fast methods
- Assumes value remains constant for some time
- Useful in sensor and financial data
- Can distort trends if overused
- Best for short missing gaps

- **Example**:
- Missing stock price on holiday → filled with previous day's price



DATA SPLITTING STRATEGIES

- You cannot randomly split a time series dataset.
- Because future cannot be used to predict the past.

Common strategies:

A) Train–Test Split

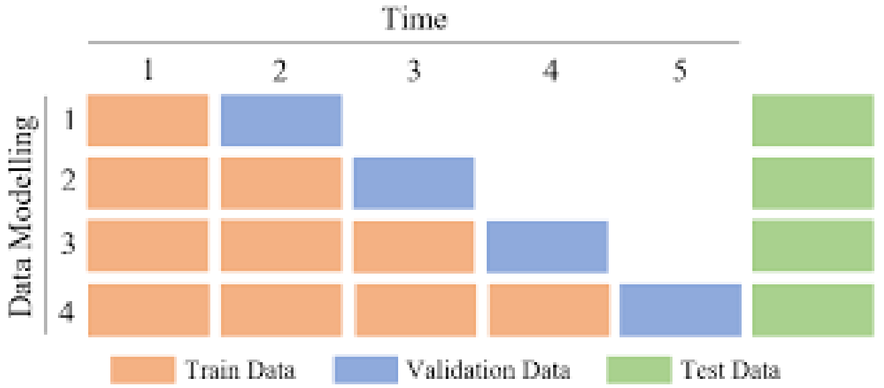
- Keep the earlier part for training, later part for testing.
- Train: 2018 → 2021
- Test : 2022

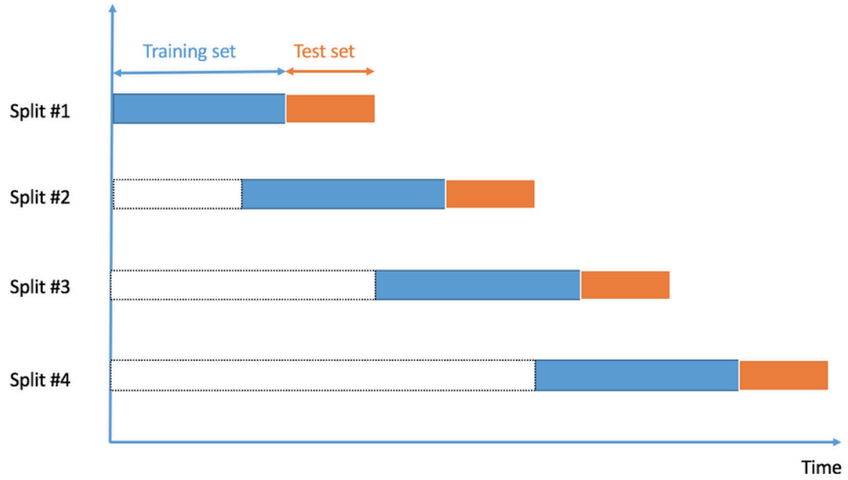
B) Rolling (Walk-Forward) Validation

- You train the model on increasing chunks of time:
- Step 1: Train: Jan–Mar → Test: Apr
- Step 2: Train: Jan–Apr → Test: May
- Step 3: Train: Jan–May → Test: Jun

C) Sliding Window Validation

- Training window size is fixed and moves forward.
- Use last k observations only
- Old data is dropped as new data arrives
- Example timeline (window size = 3):
- Train: [1 2 3] → Test: [4]
- Train: [2 3 4] → Test: [5]
- Train: [3 4 5] → Test: [6]





AUTOREGRESSIVE (AR) MODEL

- AR models predict a value using its own past values
- **“Auto” = self, “Regression” = prediction using past data**
- Similar to linear regression with lag features
- Assumes past behavior influences present
- The error term is added to capture random, unpredictable variations in the time series that cannot be explained by past values.
- **Lag**: a past value of the same time series
- **y (t-1)→ 1st lag, y (t-2)→ 2nd lag**

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

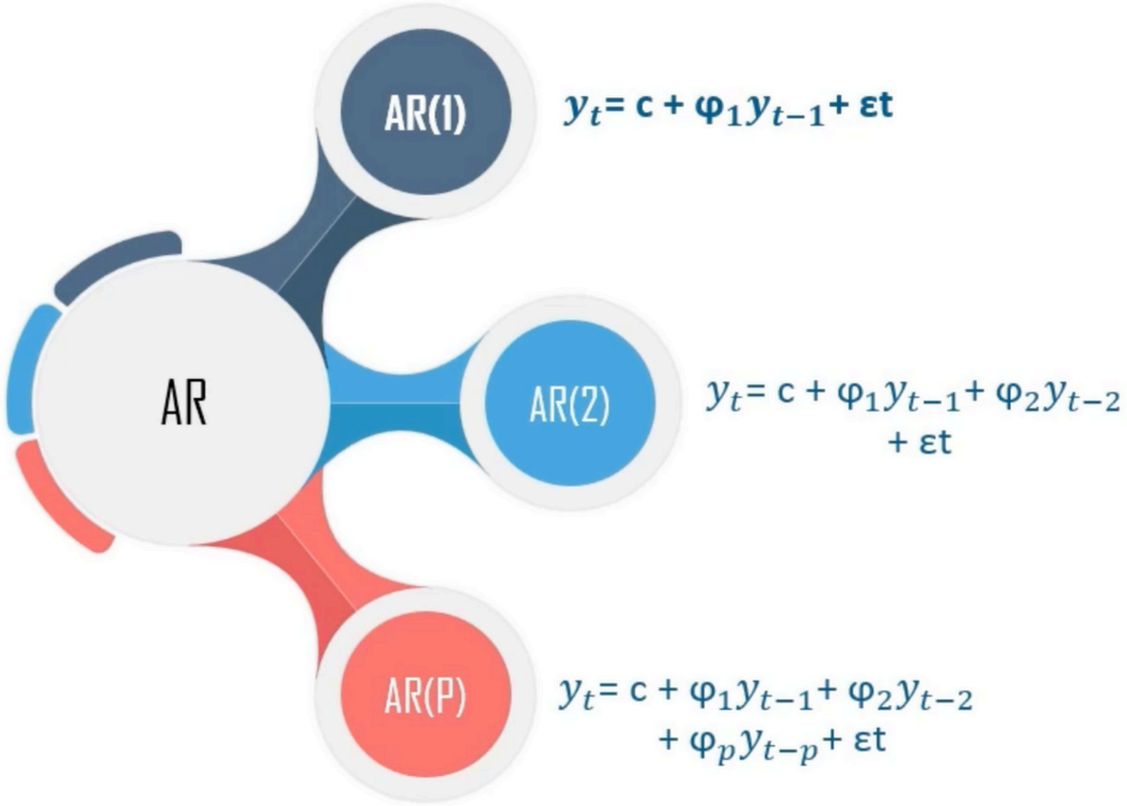
Where:

- X_t is the value at time t.
- c is a constant.
- $\phi_1, \phi_2, \ldots, \phi_p$ are the model parameters.
- $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ are the lagged values.
- ε_t represents white noise or random error at time t.

AR(1), AR(2) … AR(P)

- **AR(1)** uses only one past value:
- Meaning:
- If yesterday's sales were high → today's likely high
- But slightly adjusted by randomness

- **AR(2):** Current value depends on the two immediately previous values of the same time series plus a random error term.
- **AR(p):** Current value depends on the previous p lagged values of the same time series plus a random error term.



AUTOREGRESSIVE MODEL

- **Assumptions of AR Models:**
- Time series is stationary (constant mean & variance)
- Relationship between current and past values is linear
- Error term is white noise (zero mean, constant variance, no autocorrelation)
- Correct lag order p is selected (using PACF)

- **Partial Autocorrelation Function** (PACF) measures the direct relationship between a time series and its lagged values, after removing the effect of intermediate lags.

- **Real-Life Analogy**
- Your sleep today depends on:
- How you slept yesterday
- How tired you were the day before

- **You don't need:**
- Weather
- News
- Friends' sleep
- Just your own past behavior.
- That is Autoregression.

MOVING AVERAGE (MA) MODEL

- Moving Average (MA) is a time series model that uses past error terms
- Current value depends on past random shocks, not past values
- Smooths out random fluctuations in the series
- Useful when data shows short-term irregular patterns
- Works well for stationary time series

be represented as:

$$\mathbf{X_t} = c + \epsilon_t + \theta_1.\epsilon_{t-1} + \theta_2.\epsilon_{t-2} + \ldots + \theta_q.\epsilon_{t-q}$$

Here,

- $\mathbf{X_t}$ is the value of time series at time t
- $\mathbf{c}$ is a constant or the mean of the time series
- $\epsilon_\mathbf{t}, \epsilon_{\mathbf{t-1}}, \epsilon_{\mathbf{t-2}}, \ldots, \epsilon_{\mathbf{t-q}}$ are the white noise terms associated with the time series at time t, t-1, t-2, … , t-q.
- $\theta_\mathbf{1}, \theta_\mathbf{2}, \ldots, \theta_\mathbf{q}$ are the moving average constants.

EXAMPLE - MA(2) MODEL

Scenario: Daily Travel Time to Office
- Normal travel time: 30 minutes

Disruptions (Error Terms):
- Today (ε t): Heavy rain → +5 minutes
- Yesterday (ε t-1): Accident → +10 minutes
- Day before yesterday (ε t-2): Road repair → +6 minutes

MA(2) Explanation:
- Today's travel time depends on:
- Today's disruption
- Yesterday's disruption
- Day-before-yesterday's disruption
- **But not by your travel time last week.**
- Disruptions older than 2 days do not affect today

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

q = number of past error terms used

ε_t = white noise (random shock)

θ_i = impact of past shocks

NEED FOR ARIMA

- Before ARIMA, we had:
- AR models → depend on past values
- MA models → depend on past errors

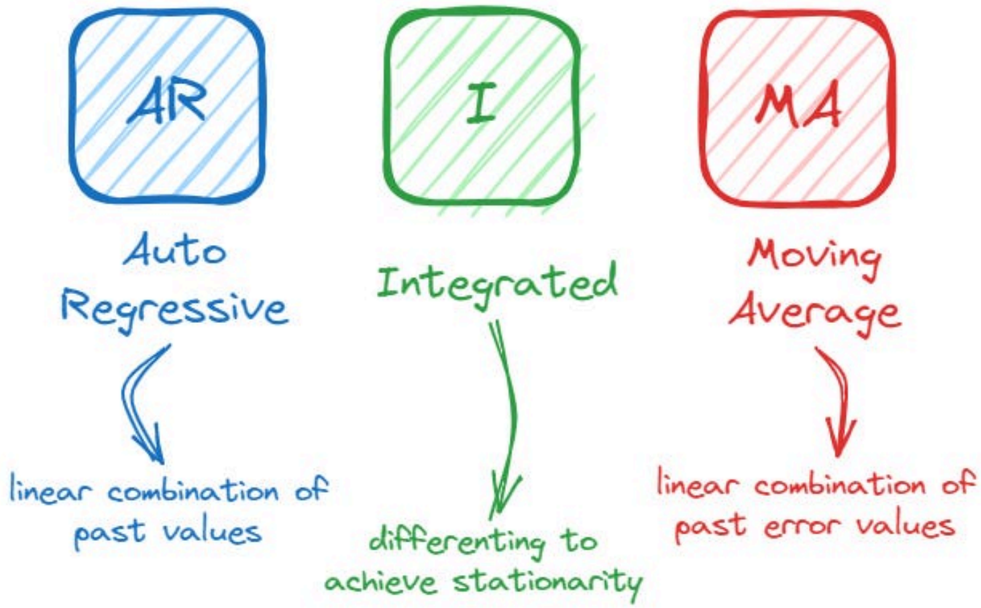
- But real-world time series data usually has:
- A trend (sales, population, demand)
- Noise
- Non-stationarity

- AR and MA fail when the data is not stationary.

- ARIMA was designed to answer this:
- “How do we forecast data that has trend + memory + randomness?”

ARIMA MODEL

- ARIMA = Autoregressive Integrated Moving Average
- Combines AR and MA models with differencing
- Suitable for non-stationary time series
- **Autoregressive (AR)**: Current value depends on past values (lags) of the series
- **Integrated (I)**: Series is differenced d times to make it stationary
- **Moving Average (MA):** Current value depends on past error terms (random shocks)



ARIMA-EXAMPLE

STEP 1: Check if Data Has a Trend (Stationarity)

- What does "trend" mean?
- If values keep increasing or decreasing → trend exists.

Example data (monthly sales):

- **Month: 1 2 3 4 5**
- **Sales: 10 12 14 16 18**

- Clearly:
- Values are increasing
- **This is not stationary**

STEP 2: Remove Trend using Differencing (This is "I")

What is differencing? Subtract yesterday's value from today's value.

Formula:

$$\text{Difference} = y_t - y_{t-1}$$

Apply differencing to sales:

Original: 10, 12, 14, 16, 18

First difference: 2, 2, 2, 2

Now: Data is stationary

So here: d = 1

Real-life analogy

Instead of tracking salary, track salary increase.

ARIMA-EXAMPLE

STEP 3: Decide How Much Past to Use (AR and MA)

- Now that data is stable, ARIMA asks:
- “How many past values should I remember?”

- **AR (p) — Autoregressive Part**

- In our differenced data:
- 2, 2, 2, 2

- We can say:
- Today’s change ≈ yesterday’s change
- **So: p = 1**

STEP 3: Decide How Much Past to Use (AR and MA)

- **MA (q) — Moving Average Part**
- Uses past errors.

- Simple idea:
- “If I was wrong yesterday, I’ll adjust today.”
- For simplicity:
- **q = 0** (ignore errors for now)

- **So our ARIMA model becomes:**
- **ARIMA(1,1,0)**

ARIMA-EXAMPLE

FINAL STEP: STEP 4: Make Prediction (Putting It All Together)

- Now the model:
- Uses differenced data
- Learns relationship
- Predicts next difference
- Converts back to original scale

$$y_t^{(d)} = c + \varepsilon_t + \underbrace{\phi_1 y_{t-1}^{(d)} + \phi_2 y_{t-2}^{(d)} + \ldots + \phi_p y_{t-p}^{(d)}}_{\text{Auto-Regressive}} + \underbrace{\theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_p \varepsilon_{t-q}}_{\text{Moving Average}}$$

Integrated

SEASONAL ARIMA (SARIMA)

- SARIMA is an extension of ARIMA for seasonal time series
- It models: Seasonal patterns (repeating cycles)
- Used when data shows regular repetition over fixed intervals
- Examples:
- Monthly sales (yearly seasonality)
- Daily traffic (weekly seasonality)
- Hourly electricity load (daily seasonality)

Data	Season length (m)
Monthly sales	12
Daily data with weekly pattern	7
Hourly data with daily pattern	24

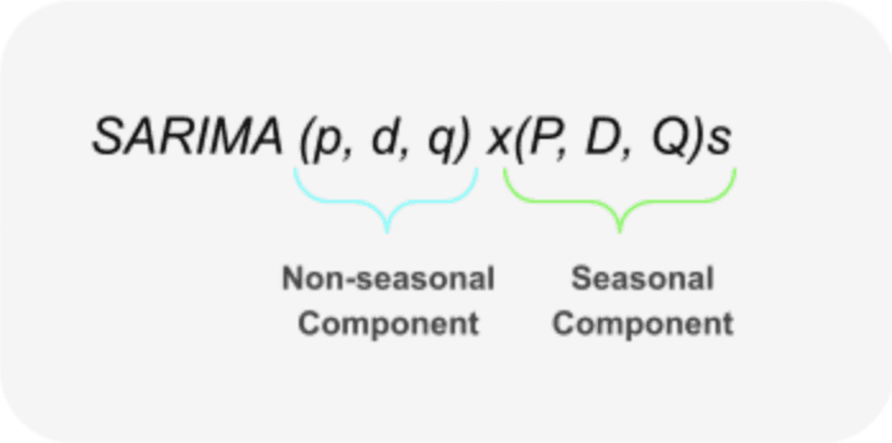
SARIMA

Non-seasonal parameters

- p → Autoregressive order (recent past values)
- d → Differencing (removes trend)
- q → Moving Average (past errors)

Seasonal parameters

- P → Seasonal AR (same season last cycle)
- D → Seasonal differencing
- Q → Seasonal MA
- s → Length of season (period)



SARIMA NUMERICAL

Month	Sales
1	100
2	120
3	140
4	160
5	105
6	125
7	145
8	165

- Month 1 ≈ Month 5
- Month 2 ≈ Month 6
- Month 3 ≈ Month 7
- Month 4 ≈ Month 8

✓ Clear seasonal repetition

✓ Fixed interval pattern

Observations

- Pattern repeats every **4 months**
- Same shape appears again from Month 5 onward

Season length (s) = 4

SARIMA-SEASONAL DIFFERENCING

$$y'_t = y_t - y_{t-s}$$

Apply (s = 4)

Month	Calculation	Result
5	105 – 100	5
6	125 – 120	5
7	145 – 140	5
8	165 – 160	5

- **After seasonal differencing:**
- Values are stable
- No increasing or decreasing trend
- No further differencing required
- **d = 0**

- **Based on behavior:**
- Current value depends on:
- Previous time step → **AR(1)**
- Same season last cycle → Seasonal **AR(1)**
- Assume no MA components for simplicity

FINAL SARIMA MODEL

$$\mathbf{SARIMA}(1,0,0)(1,1,0)_4$$

Meaning:

- **AR(1):** depends on last month
- **Seasonal AR(1):** depends on same month last season
- **Seasonal differencing:** removes repeating cycle of 4 months

LOVELY PROFESSIONAL UNIVERSITY

NAAC GRADE A++

THANK YOU