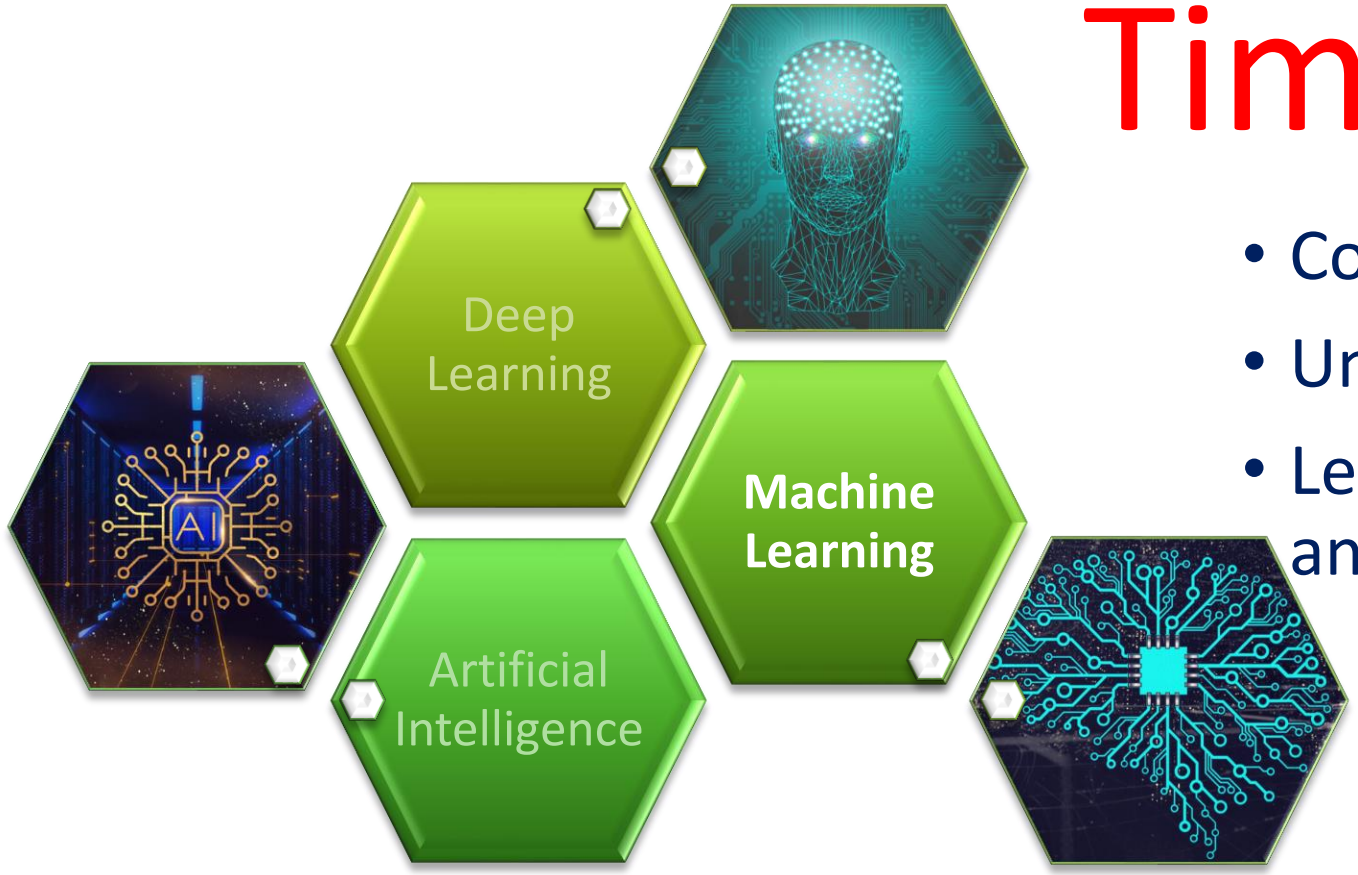


Time Series

- Course Code: CSE471
- Unit 1: **Time Series: An Introduction**
- Lecture 2: Finding time series data and retrofitting.

Origin of Statistical Time Series Analysis,

1. John Graunt, in 17th-century London haberdasher. Graunt was interested in studying the population of London, and he used time series data to track the number of births, deaths, and marriages over time. He published his findings in a book called "Natural and Political Observations Made upon the Bills of Mortality"
2. In the 19th century, the work of Adolphe Quetelet and Francis Galton helped to lay the foundations of modern time series analysis. Quetelet studied the distribution of human traits, such as height and weight. He developed the concept of the "average man," which is a hypothetical person who represents the average of a population
3. Galton studied the inheritance of traits. He developed the concept of the "correlation coefficient," which is a measure of the association between two variables.

Origin of Statistical Time Series Analysis,

- In the 20th century, the development of computers and statistical software made it possible to analyze time series data more effectively. This led to the development of new time series analysis methods, such as autoregressive moving average (ARMA) models and autoregressive integrated moving average (ARIMA) models. These models are used to forecast future values of time series data.
- Today, time series analysis is a widely used tool in a variety of fields, including economics, finance, meteorology, and medicine. It is used to track trends, forecast future values, and identify patterns in time series data.

Origin of Machine Learning Time Series Analysis

- NASA, “Weather Forecasting Through the Ages,” Nasa.gov, February 22, 2002.

NASA gives a history of how weather forecasting came to be, with emphasis on specific research challenges and successes in the 20th century.

- Richard C. Cornes, “Early Meteorological Data from London and Paris: Extending the North Atlantic Oscillation Series” May 2010.

This doctoral thesis offers a fascinating account of the kinds of weather information available for two of Europe’s most important cities, complete with extensive listings of the locations and nature of historic weather in time series format.

- Dan Mayer, “A Brief History of Medicine and Statistics,”, 2004,

This chapter of Mayer’s book highlights how the relationship between medicine and statistics depended greatly on social and political factors that made data and statistical training available for medical practitioners.

- Simon Vaughan, “Random Time Series in Astronomy”, (2013)

Vaughan summarizes the many ways time series analysis is relevant to astronomy and warns about the danger of astronomers rediscovering time series principles or missing out on extremely promising collaborations with statisticians.

Finding and wrangling the time series data

- Finding time series data from online repositories
- Discovering and preparing time series data from sources not originally intended for time series
- Addressing the common problems that one encounter with time series data, especially the difficulties that arises from timestamps

Where to find time series data

- Finding the time series data depends on one of these goals:
  - Finding an appropriate data set for learning or experimentation purposes
  - Creating a time series dataset, out of existing data that is not stored in an explicitly time- oriented form

Prepared Datasets

UC Irvine Machine Learning Repository

Datasets | Contribute Dataset | About Us | Search datasets...

Filters

time series

- Keywords
- Data Type
- Subject Area
- Task
- # Attributes
- # Instances
- Attribute Type

Browse Datasets

SORT BY # VIEWS, DESC | EXPAND ALL

Bach Chorales

Time-series data based on chorales; challenge is to learn generative grammar; data in Lisp

100 Instances

Hierarchical Sales Data

This dataset contains hierarchical sales data gathered from an Italian grocery store

Clustering, Other | Time-Series | 1.8K Instances

Multivariate Gait Data

Bilateral (left, right) joint angle (ankle, knee, hip) times series data collected from 10 healthy subjects under 3 w

Classification, Regression, C... | Sequential, Multivariate, Ti... | 181.8K Instances

Physical Therapy Exercises Dataset

This dataset contains wearable inertial and magnetic sensor data during the execution of physical therapy exer

Classification, Clustering, Ot... | Time-Series | 268.88K Instances

Prepared Datasets

Time Series Classification | Home | Datasets | Algorithms | Results | Researchers | Code | Bibliography | UEA Papers | About Us

Train Size

Less than 100 (49)
100 to 500 (87)
Greater than 500 (47)

Test Size

Less than 300 (86)
300 to 1000 (52)
Greater than 1000 (45)

Length

Less than 300 (94)
300 to 700 (41)
Greater than 700 (48)

Classes

Less than 10 (139)
10 to 30 (34)
Greater than 30 (10)

Type

Device (10)

Dataset	Train Size	Test Size	Length	No. of Classes	Type
AbnormalHeartbeat	303	303	3053	5	AUDIO
ACSF1	100	100	1460	10	DEVICE
Adiac	390	391	176	37	IMAGE
AllGestureWiimoteX	300	700	0	10	MOTION
AllGestureWiimoteY	300	700	0	10	MOTION
AllGestureWiimoteZ	300	700	0	10	MOTION
ArrowHead	36	175	251	3	IMAGE
ArticularyWordRecognition	275	300	144	25	MOTION
AsphaltObstacles	390	391	0	4	MOTION
AsphaltObstaclesCoordinates	390	391	0	4	MOTION
AsphaltPavementType	1055	1056	0	3	MOTION
AsphaltPavementTypeCoordinates	1055	1056	0	3	MOTION
AsphaltRegularity	751	751	0	2	MOTION
AsphaltRegularityCoordinates	751	751	0	2	MOTION

Prepared Datasets

IEEE DataPort™ DATASETS | SUBMIT A DATASET | COMPETITIONS | SEARCH IEEE

Datasets

Search Terms: Search title, author, abstract, category, data format, DOI

Category: - Any -

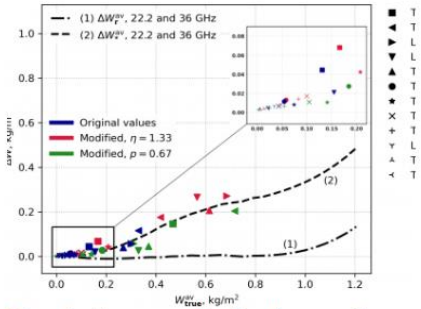
Dataset Type: - Any -

Search

DATASET CATEGORIES

- Artificial Intelligence (1,012)
- Astronomy (13)
- Biomedical and Health Sciences (328)
- Biophysiological Signals (107)
- Climate Change/Environmental (62)

Brain-computer interface-based



Simulation results for "smooth water surface -...

Data and Scripts for A Climate Resilience Assessment...

Tara Oceans Sample Kinase Dataset

Prepared Datasets

kaggle

Create

Home

Competitions

Datasets

Models

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Web traffic Forecast...

Research Paper Lab...

Multi Label Classifie...

Multi-Label Classific...

postpartum-depres...

Datasets

New Dataset | Your Work

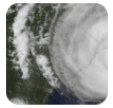
Time series

Filters

All datasets | Computer Science | Education | Classification | Computer Vision | NLP | Data Visualization | Pre-Trained Model

3,224 Datasets

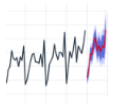
Hotness



Daily Climate time series data
sumanthvrao · Updated 4 years ago
Usability **10.0** · 2 Files (CSV) · 23 kB
278
Silver

House Property Sales Time Series
HtAG Holdings · Updated 4 years ago
Usability **10.0** · 2 Files (CSV) · 165 kB
115
Bronze

DJIA 30 Stock Time Series
szrlee · Updated 6 years ago
Usability 7.1 · 33 Files (CSV) · 3 MB
295
Gold



Time Series starter dataset
podsyp · Updated 3 years ago
Usability **10.0** · 1 File (CSV) · 2 kB
50
Bronze

Microsoft Stock- Time Series Analysis
Vijay V Venkitesh · Updated 2 years ago
64

Prepared Datasets

data.gov.in Open Government Data (OGD) Platform India

HOME CATALOG APIs SECTORs CDOs METRICs | LOGIN | REGISTER

/ Search

The advanced search feature allows you to search Catalog and Resources with title, description, keywords, domain, sector, state/ministry/department/organization, and asset jurisdiction.

Filter By

time series

All 10 Exact Match

- Domain(1)
- Sector
- Ministry/Department
- State/Department
- Asset Jurisdiction
- Type
- APIs for Consumption

Selected Filters :

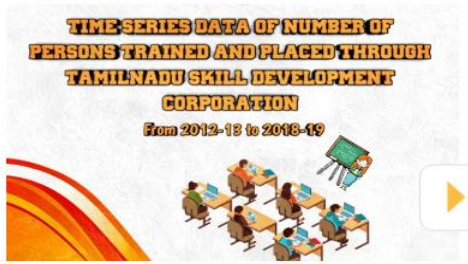
Domain: data.gov.in

Show More

Visualizations (10) | Blogs (9) | Infographics (1)

Domestic Tourists Arrival in Tamil Nadu 2018

Time series data of Rainfall in Tamil Nadu as per Statistical Hand Book 2020



Time series data Number of Persons trained and placed through Tamil Nadu Skill . . .

Time *Series* Data Foreign Trade

Views Downloads CSV Export in Other Format

Other Useful resources

- Government Time series datasets
  - NOAA National Centers for Environmental information publishes a variety of time series data relating to temperatures at every 15 minutes for all US whether stations
  - Bureau of Labor Statistics publishes a monthly index of the national unemployment rate.

Finding time series data in structured data

- ***Timestamped recordings of events***
  - If there is a timestamp on your data, you have the potential to construct a time series. Even if all you do is record the time a file was accessed with no other information, you have a time series.
- ***"Timeless" measurements where another measurement substitutes for time***
  - In some cases, time is not explicit in data but is accounted for in the underlying logic of the data set.
- ***Physical traces***
  - Many scientific disciplines record physical traces, be it for medicine, audiology, or weather.

Retrofitting a time series data collection from a collection of tables

- **To create a more complete time series**: If the tables contain different but overlapping time periods, then combining them can create a more complete time series.
- **To improve the quality of the time series:** If the tables contain different levels of noise or accuracy, then combining them can improve the overall quality of the time series.
- **To make the time series more accessible:** If the tables are stored in different locations or formats, then combining them can make the time series more accessible to users.

How to retrofit a time series data collection from a collection of tables:

- Step 1: Identify the tables that contain time series data.
- Step 2: Identify the columns in each table that contain the time series data.
- Step 3: Combine the columns from each table into a single column.
- Step 4: Normalize the data in the combined column so that it is all on the same scale.
- Step 5: Save the combined column as a new table.

Data Cleaning

Unwanted Observations	• Remove Duplicate observation • Remove redundant/irrelevant values
Missing Data	• Fix unknown/ missing values using imputation • Remove incomplete entries
Noisy Data	• Fix problems with mislabelled data • Fix problems such as same attribute with different name
Inconsistent Data	• Unwanted Values that do not fit the dataset • Outliers

Data Cleaning: Example

Solution: *Ignore/ Fill Manually/ Use mathematical tools to fill/replace values*

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values
Remove entry

Invalid values
Fill Manually

Misfielded values
Fill Manually/ Remove entry

Misspellings
Fill Manually

Uniqueness
Remove repeated data

Formats
Modify to 90's, 70's… (Data Binning)

Attribute dependencies
Modify to 90's, 70's… (Data Binning)

https://quantdare.com/data-cleansing-and-transformation/

Cleaning Time series Data.

- Missing data
- Changing the frequency of a time series (that is, upsampling and downsampling)
- Smoothing data
- Addressing seasonality in data
- Preventing unintentional lookaheads

Handling missing data

The most common methods to address missing data in time series are:

- **Imputation**

  When we fill in missing data based on observations about the entire data set.

- **Interpolation**

  When we use neighboring data points to estimate the missing value. Interpolation can also be a form of imputation.

- **Deletion of affected time periods**

  When we choose not to use time periods that have missing data at all.

Deletion of affected time periods

- Deleting time periods with missing data will result in less data for your model.
- Whether to preserve the data or throw out problematic time periods will depend on your use case
- Whether you can afford to sacrifice the time periods in question given the data needs of your model.

Filling the missing values

- Forward fill
- Backward fill
- Moving average
- Interpolation

Forward Fill

- One of the simplest ways to fill in missing values is to carry forward the last known value prior to the missing one, an approach known as *forward fill*

Forward Fill

	Units sold
2021-01-03	5.0
2021-01-10	4.0
2021-01-17	NaN
2021-01-24	NaN
2021-01-31	1.0
2021-02-07	NaN
2021-02-14	3.0
2021-02-21	6.0
2021-02-28	NaN
2021-03-07	2.0

Original data

	Units sold
2021-01-03	5.0
2021-01-10	4.0
2021-01-17	4.0
2021-01-24	4.0
2021-01-31	1.0
2021-02-07	1.0
2021-02-14	3.0
2021-02-21	6.0
2021-02-28	6.0
2021-03-07	2.0

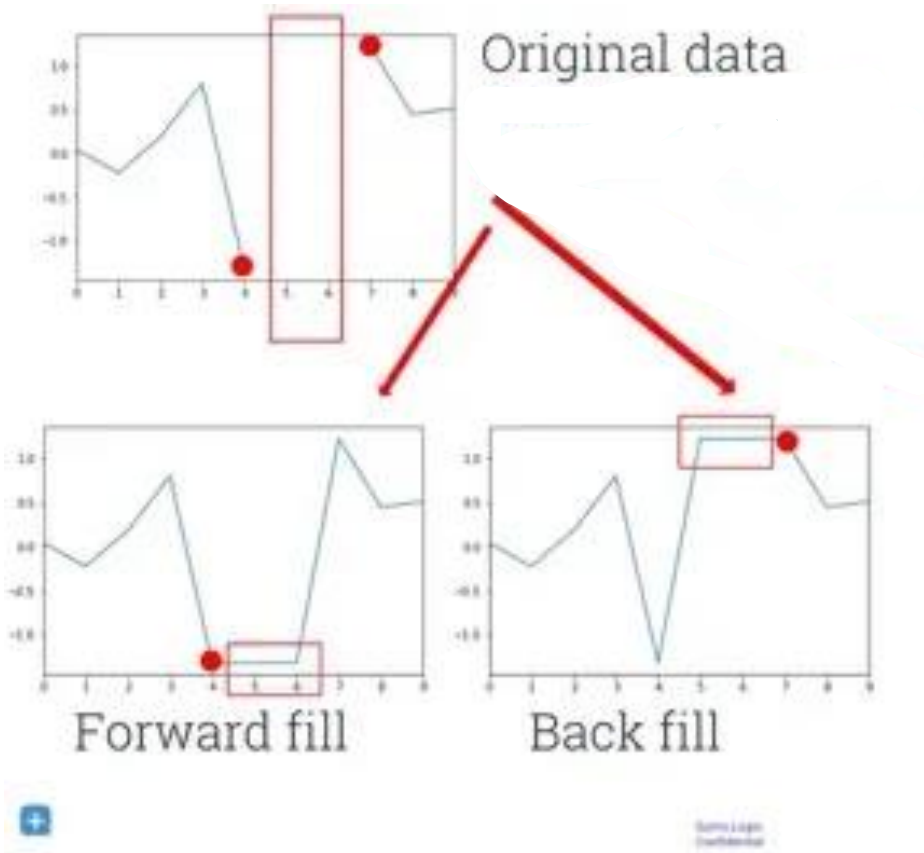
ForwardFill data

Python code:

```
forward_filled=df.fillna(method='ffill')
print(forward_filled)
```

Backward Fill

- Backward filling means fill missing values with next known data point.



Backward Fill

	Units sold
2021-01-03	5.0
2021-01-10	4.0
2021-01-17	NaN
2021-01-24	NaN
2021-01-31	1.0
2021-02-07	NaN
2021-02-14	3.0
2021-02-21	6.0
2021-02-28	NaN
2021-03-07	2.0

Original data

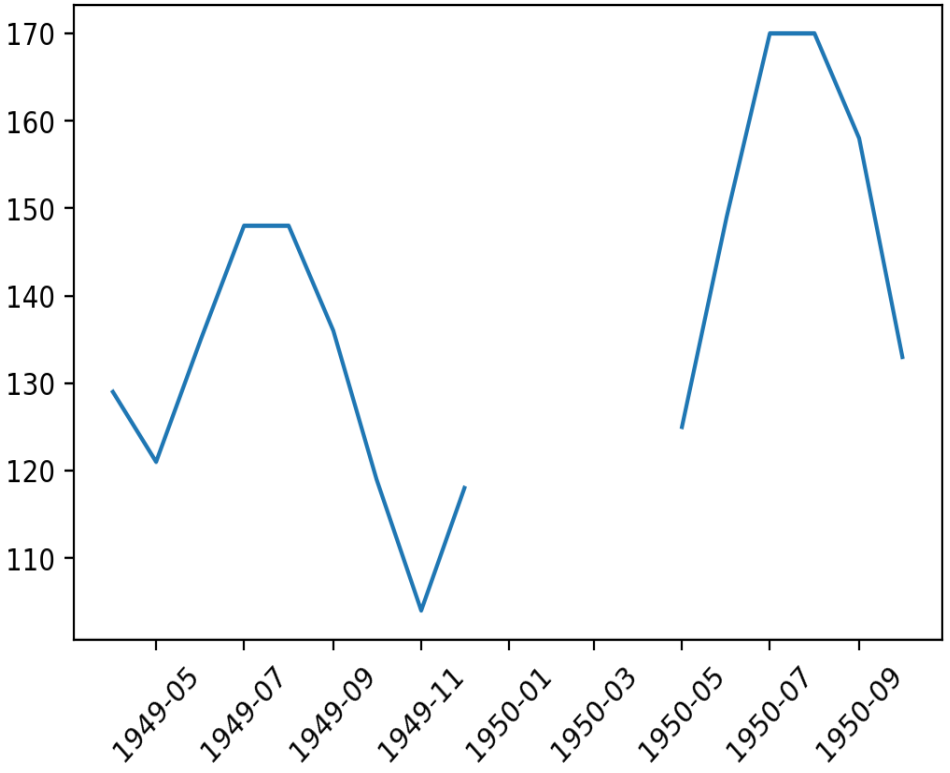
	Units sold
2021-01-03	5.0
2021-01-10	4.0
2021-01-17	1.0
2021-01-24	1.0
2021-01-31	1.0
2021-02-07	3.0
2021-02-14	3.0
2021-02-21	6.0
2021-02-28	2.0
2021-03-07	2.0

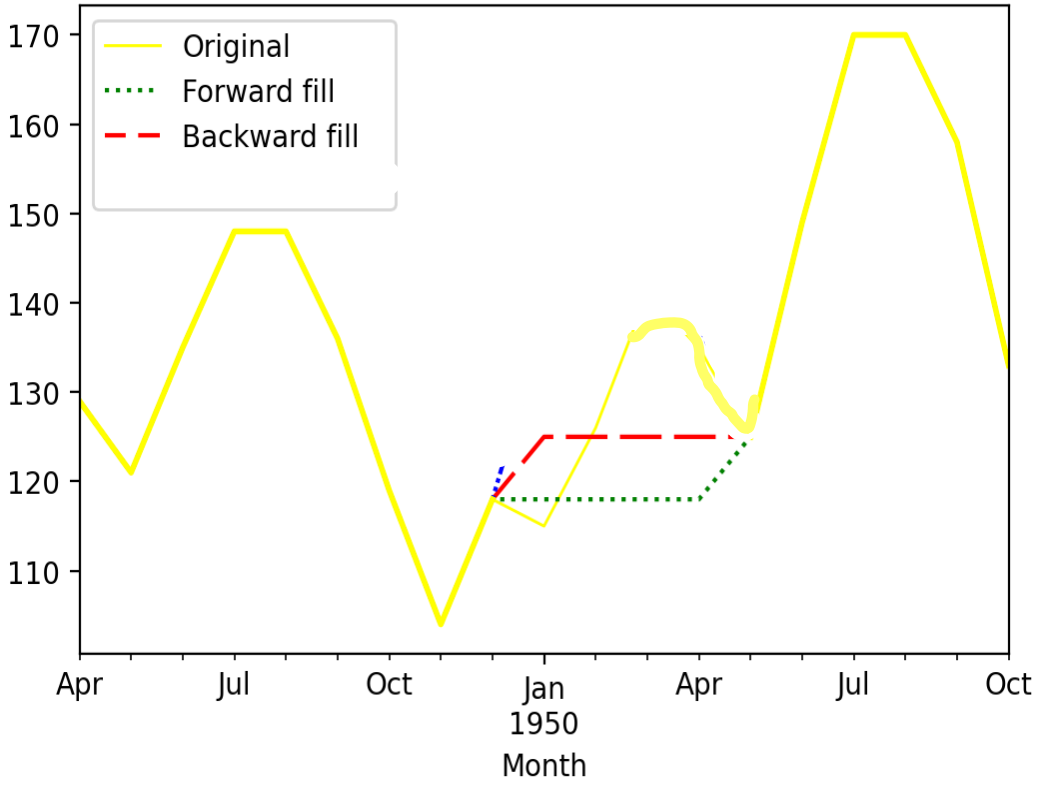
BackwardFill data

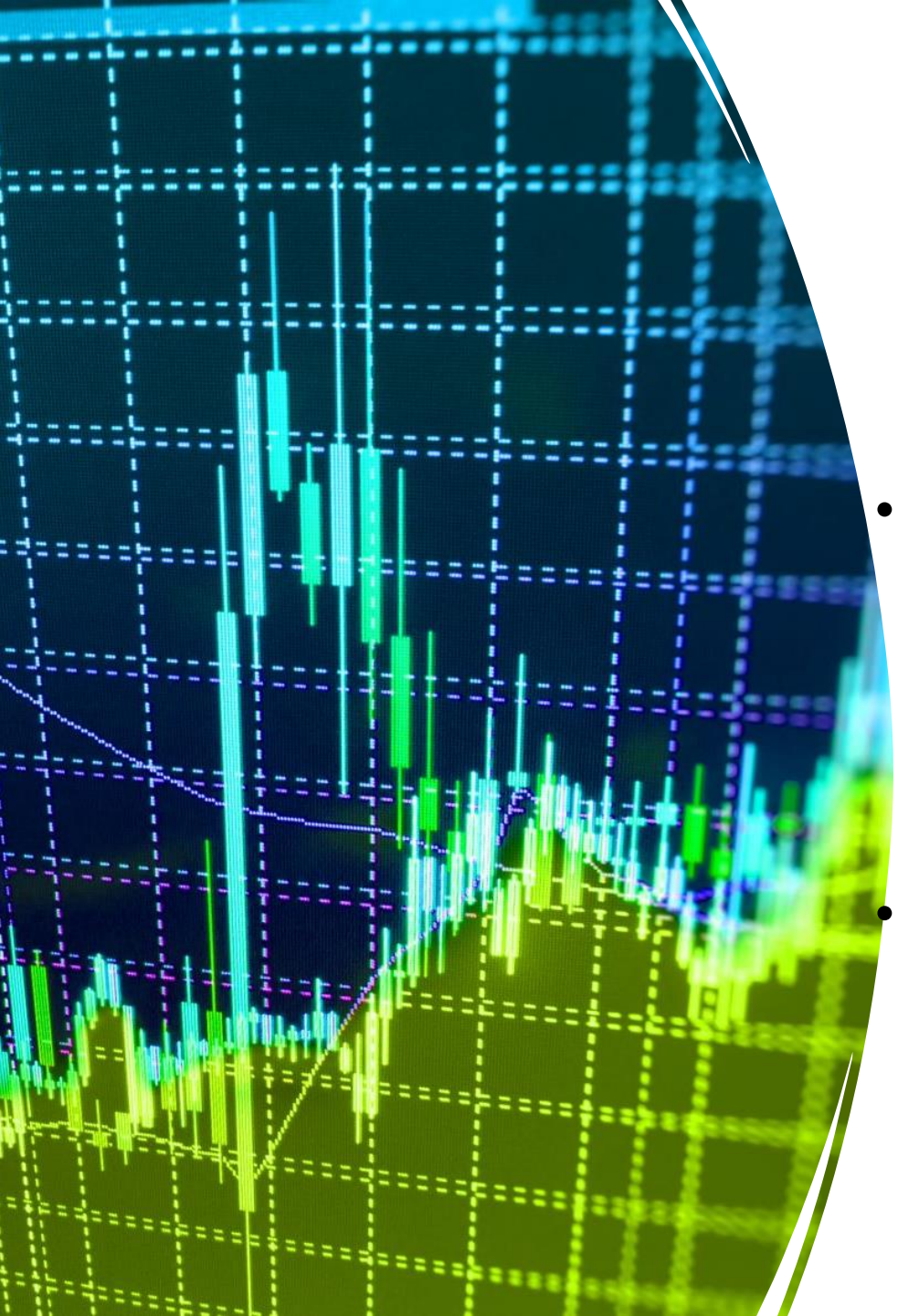
Python code:

```
backward_filled=df.fillna(method='bfill')
print(backward_filled)
```

Comparison


170
160
150
140
130
120
110
1949-05
1949-07
1949-09
1949-11
1950-01
1950-03
1950-05
1950-07
1950-09

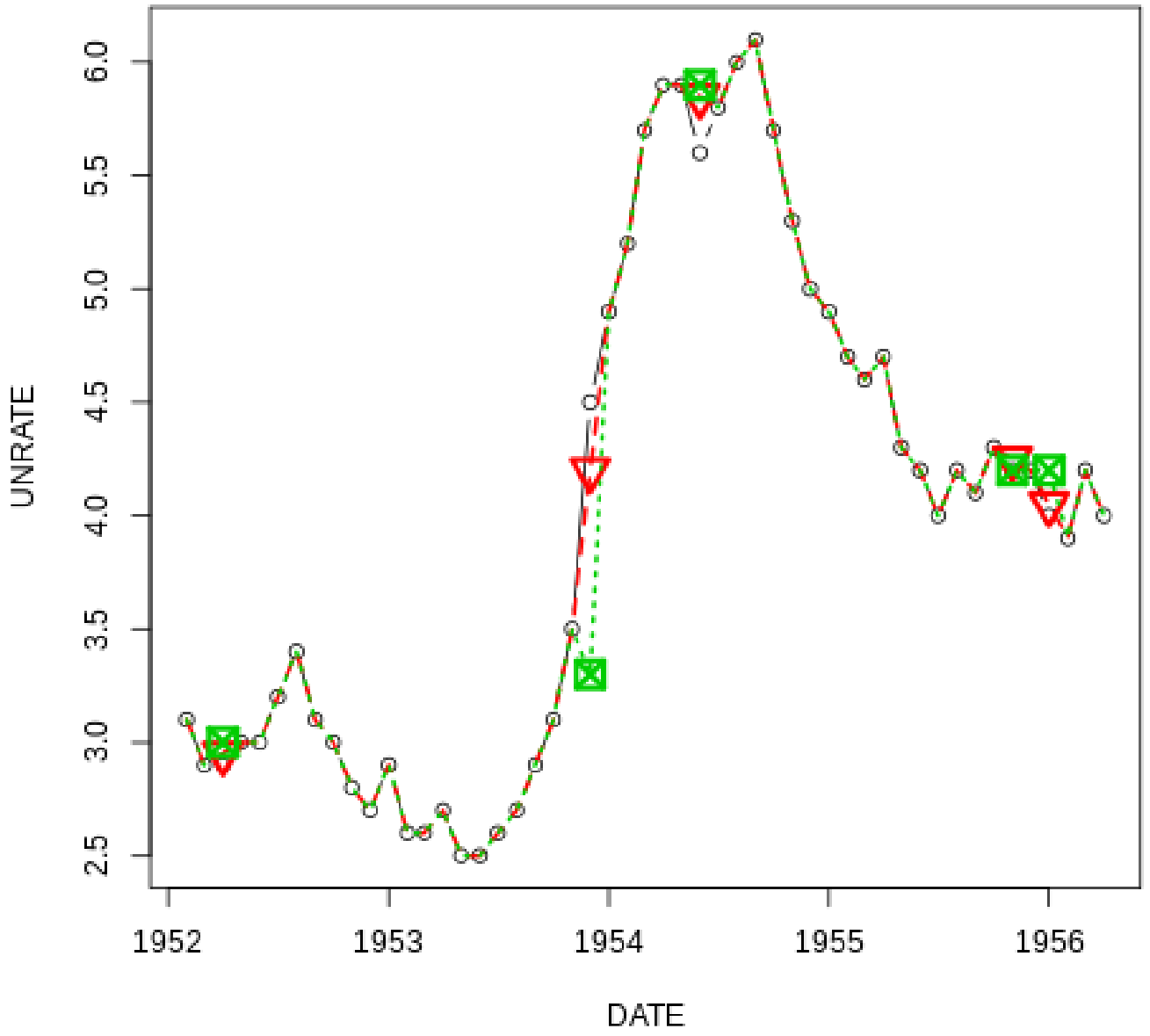

Original
Forward fill
Backward fill
170
160
150
140
130
120
110
Apr
Jul
Oct
Jan
1950
Apr
Jul
Oct
Month

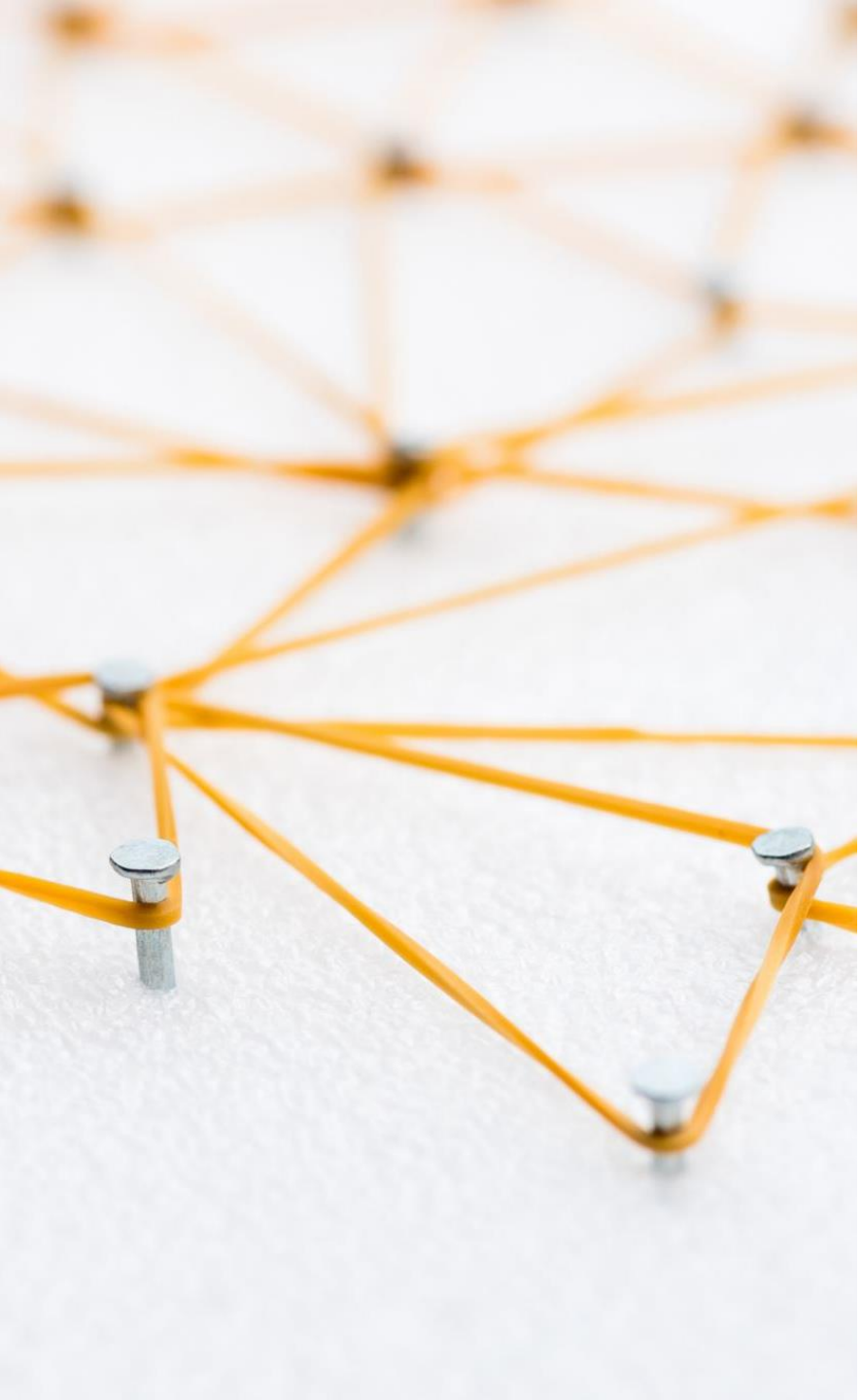


Moving average

- We can also impute data with either a rolling mean or median. Known as a *moving average*, this is similar to a forward fill in that you are using past values to “predict” missing future values (imputation can be a form of prediction). With a moving average, however, you are using input from *multiple* recent times in the past.
- A moving average doesn’t have to be an arithmetic average. For example, exponentially weighted moving averages would give more weight to recent data than to past data. Alternately, a geometric mean can be helpful for time series that exhibit strong serial correlation and in cases where values compound over time.

- The results for both a forward-looking moving average and a moving average computed with future and past data

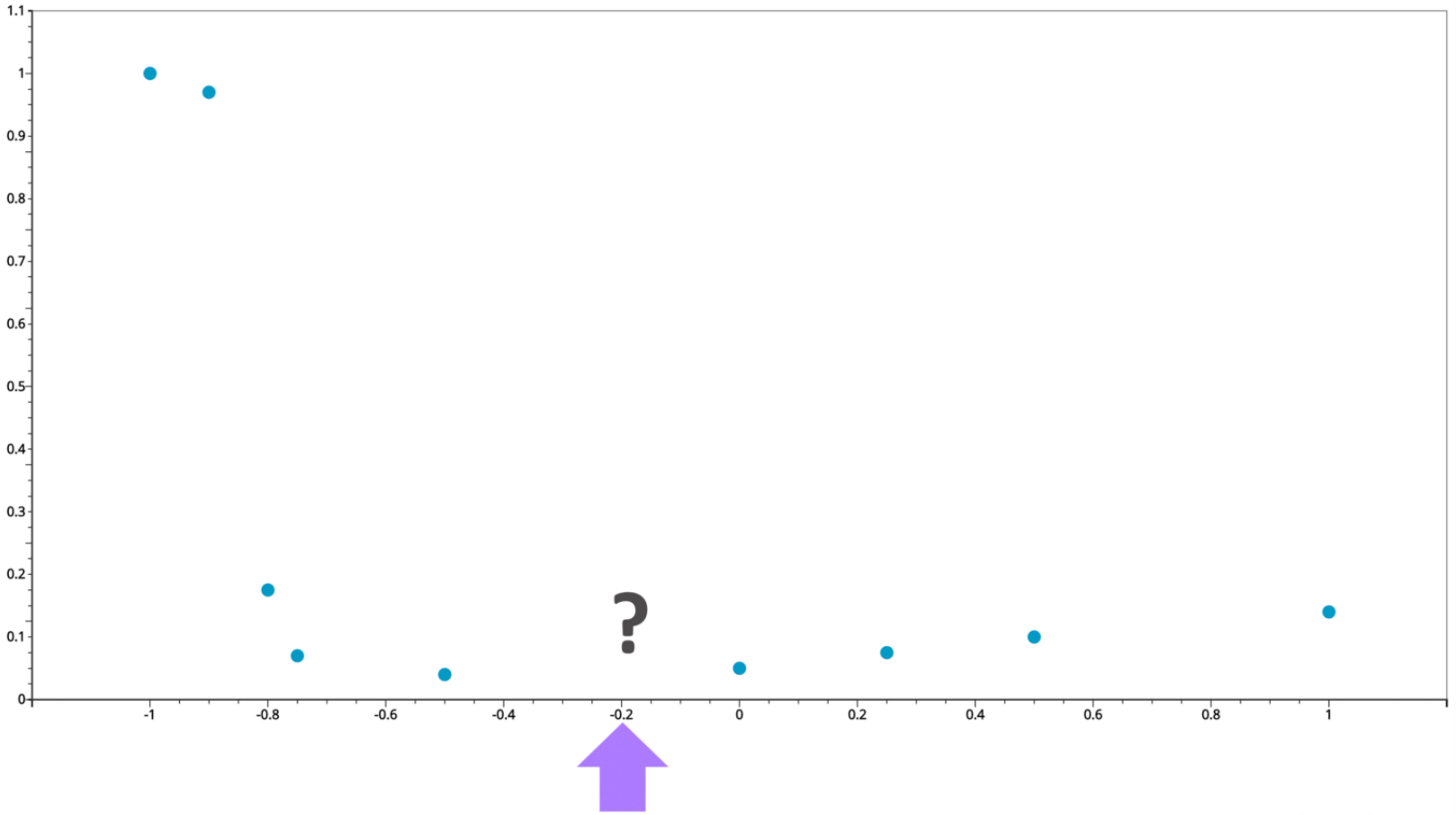




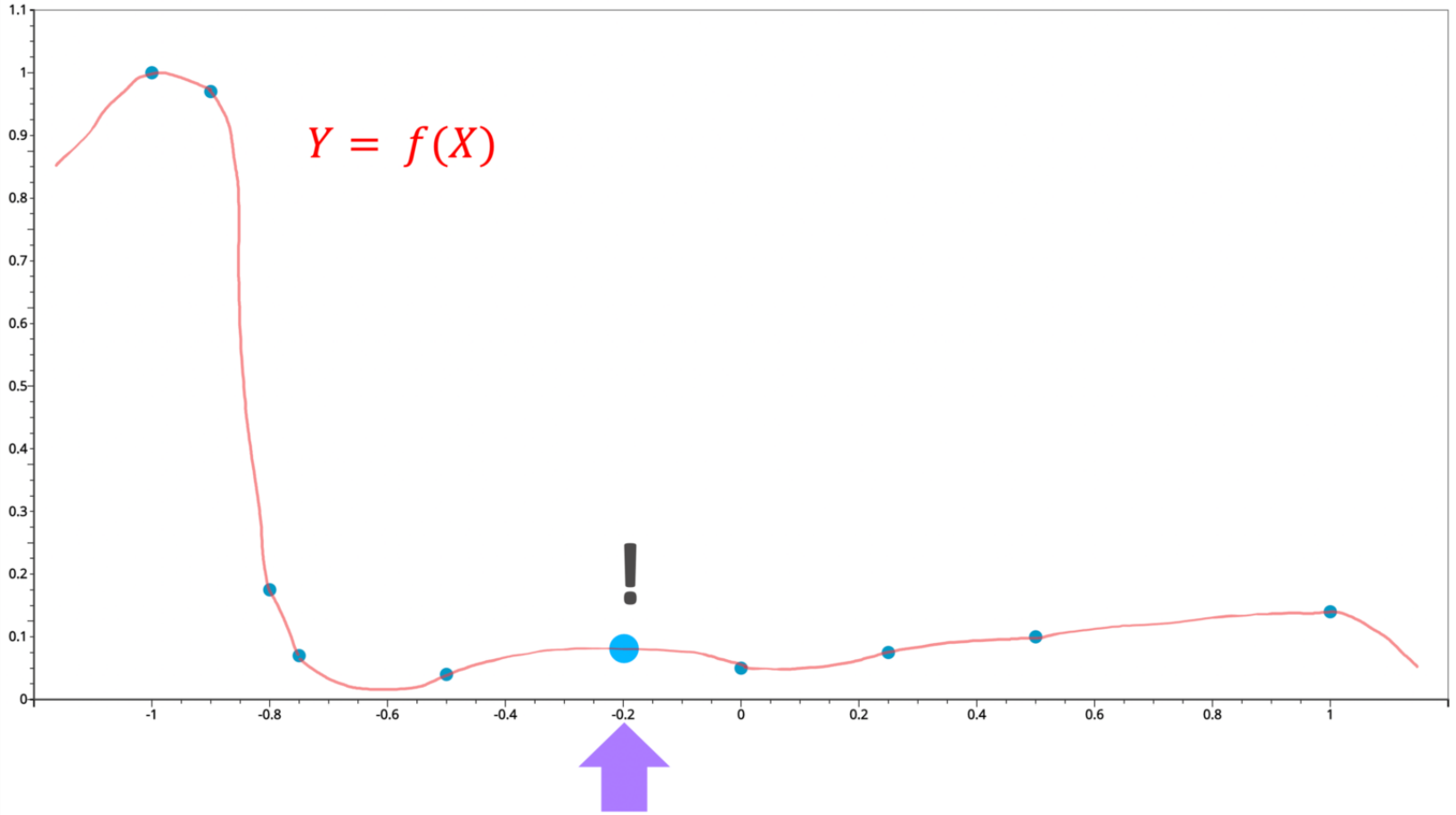
Interpolation

- Interpolation is a method of determining the values of missing data points based on geometric constraints regarding how we want the overall data to behave. For example, a linear interpolation constrains the missing data to a linear fit consistent with known neighboring points.

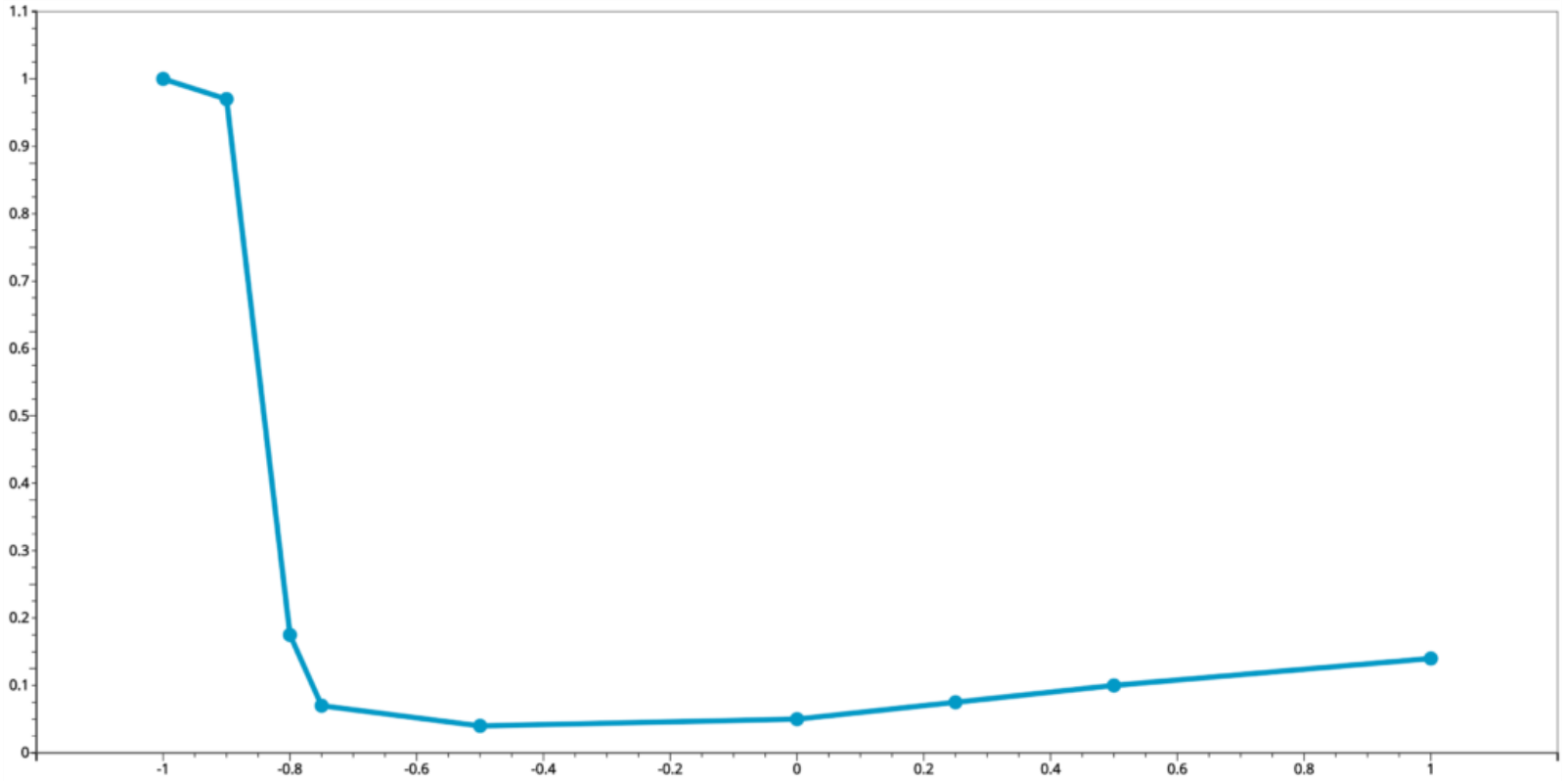
Linear Interpolation



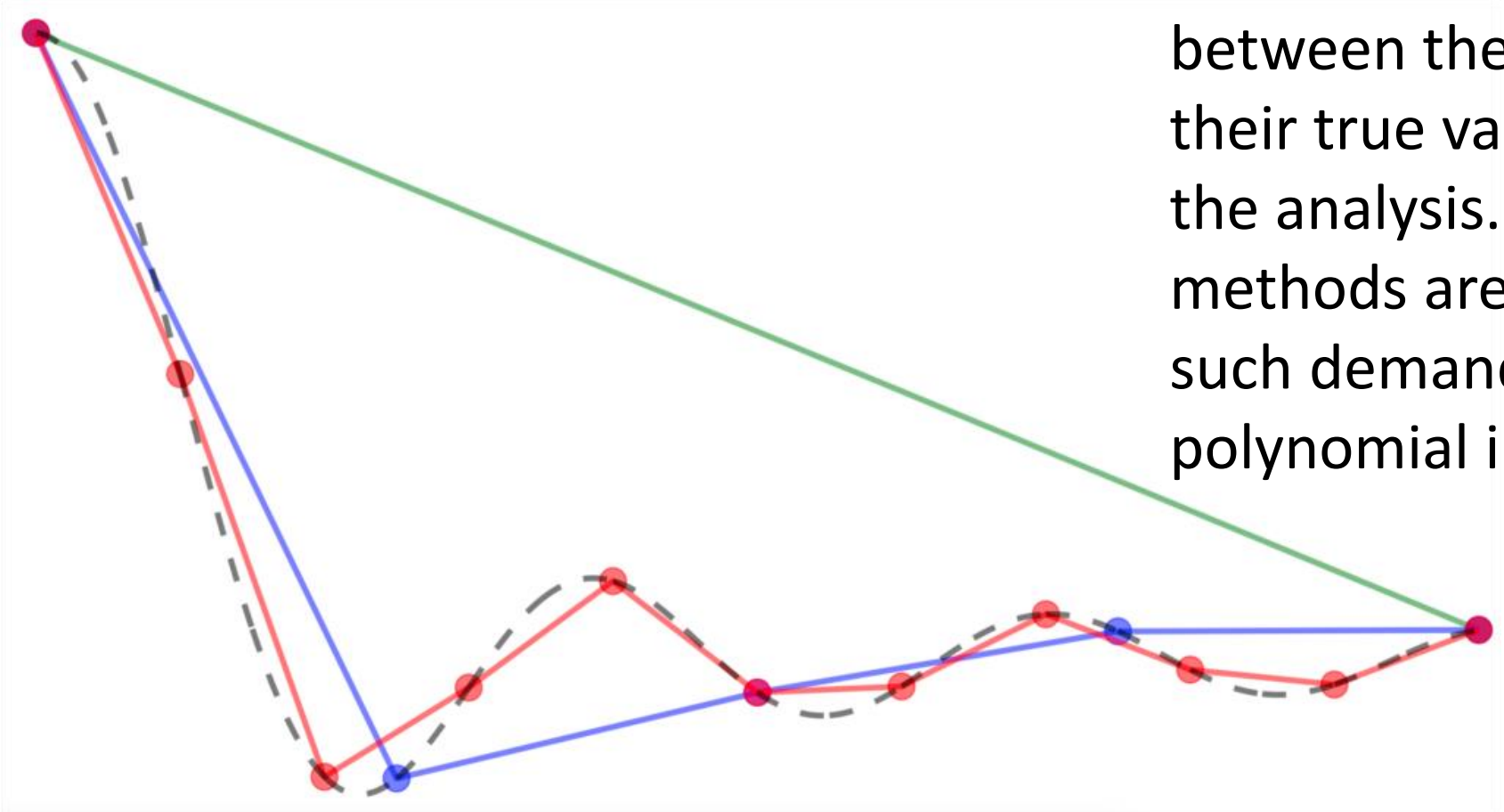
Linear Interpolation



Linear Interpolation

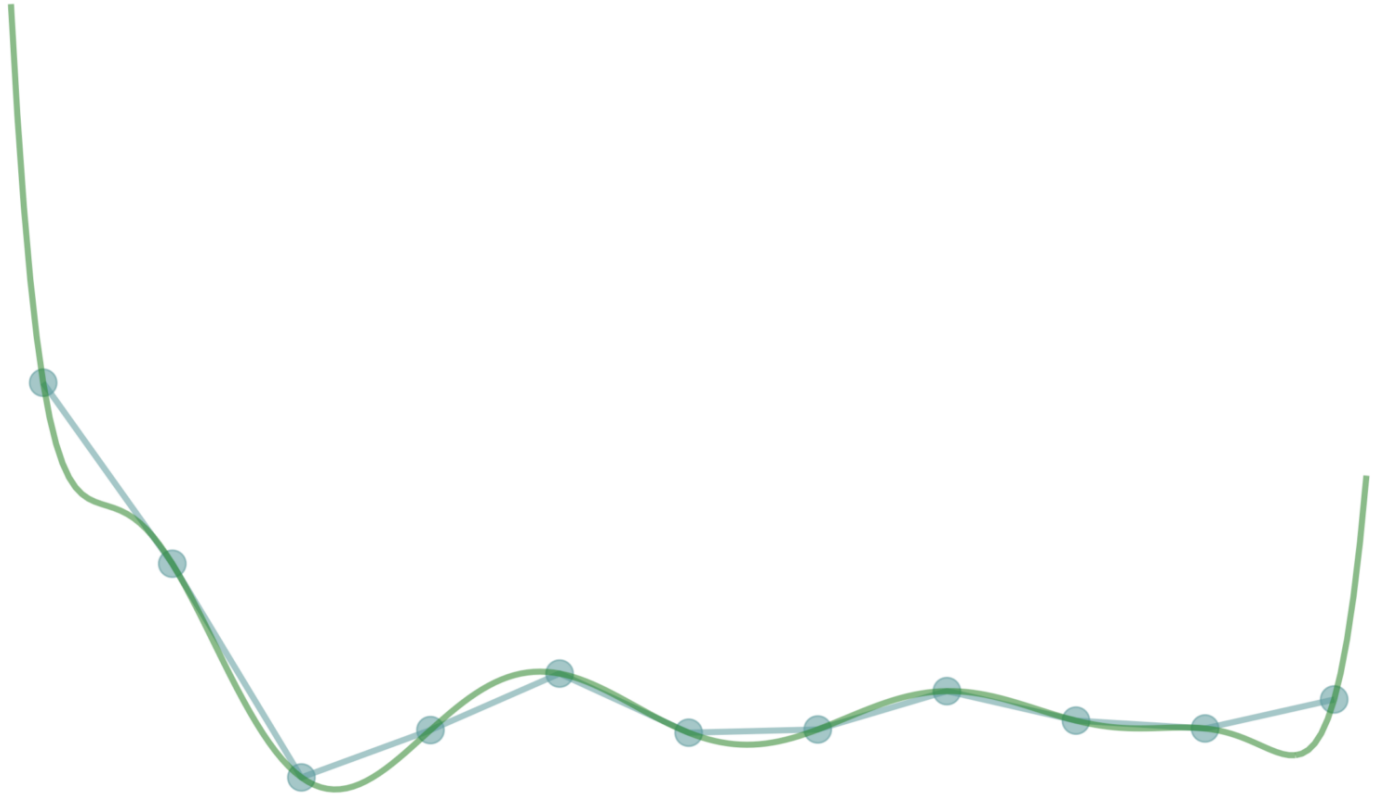


linear Interpolation



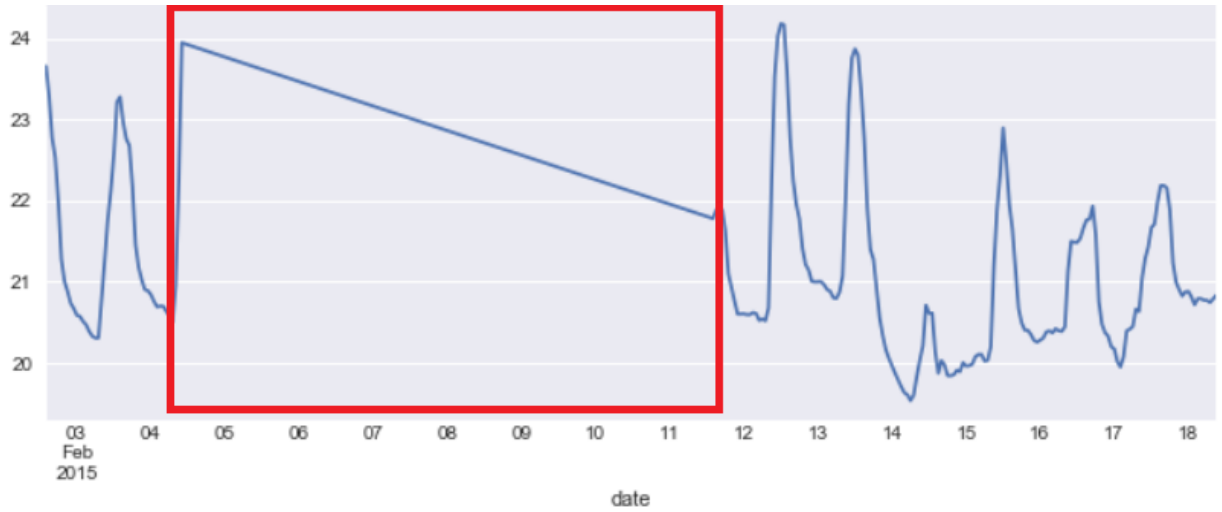
When we have fewer points, the errors between the interpolated values and their true values have a larger impact on the analysis. Thus, more interpolation methods are created in order to meet such demands. One of the examples is polynomial interpolation

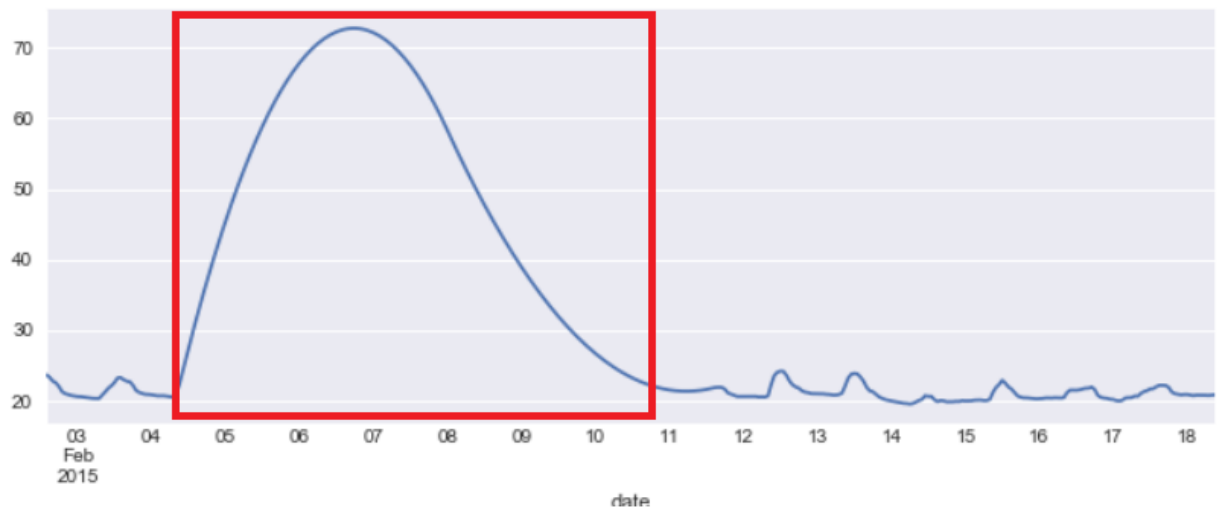
Polynomial Interpolation



$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

Linear and polynomial Interpolation





Python code:

```
df3['Temperature'].interpolate('linear').plot()
```

Python code:

```
df3['Temperature'].interpolate('Polynomial', order=2).plot()
```

Interpolation

	date	fruit	price
0	2021-01-01	apple	0.8
1	2021-01-02	apple	NaN
2	2021-01-03	apple	NaN
3	2021-01-04	apple	1.2
4	2021-01-01	mango	NaN
5	2021-01-02	mango	3.1
6	2021-01-03	mango	NaN
7	2021-01-04	mango	2.8

interpolate

$$\frac{1.2 - 0.8}{3} = 0.133$$

interpolate

$$\frac{3.1 - 1.2}{2} = 0.95$$

interpolate

$$\frac{2.8 - 3.1}{2} = -0.15$$

	date	fruit	price
0	2021-01-01	apple	0.800
1	2021-01-02	apple	0.933
2	2021-01-03	apple	1.067
3	2021-01-04	apple	1.200
4	2021-01-01	mango	2.150
5	2021-01-02	mango	3.100
6	2021-01-03	mango	2.950
7	2021-01-04	mango	2.800

+0.133

+0.133

+0.95

-0.15

DownSampling and Upsampling

Change the frequency of the timestamps in your data collection. This is called *upsampling* and *downsampling*, for increasing or decreasing the timestamp frequency, respectively.

When to perform DownSampling?

Downsampling is subsetting data such that the timestamps occur at a lower frequency than in the original time series.This is most often done in the following cases.

- The original resolution of the data isn’t sensible.

When to perform DownSampling?

Downsampling is subsetting data such that the timestamps occur at a lower frequency than in the original time series.This is most often done in the following cases.

- The original resolution of the data isn’t sensible.
- Focus on a particular portion of a seasonal cycle.

When to perform DownSampling?

Downsampling is subsetting data such that the timestamps occur at a lower frequency than in the original time series.This is most often done in the following cases.

- The original resolution of the data isn’t sensible.
- Focus on a particular portion of a seasonal cycle.
- Match against data at a lower frequency.

Upsampling

- Upsampling is representing data as if it were collected more frequently than was actually the case.

Where to perform Upsampling?

- Irregular time series.
- Inputs sampled at different frequencies.
- Knowledge of time series dynamics