Time Series Analysis

INFO 523 - Lecture 12

Dr. Greg Chism

Lesson 2:

Processing Timeseries Data

Lagging Values

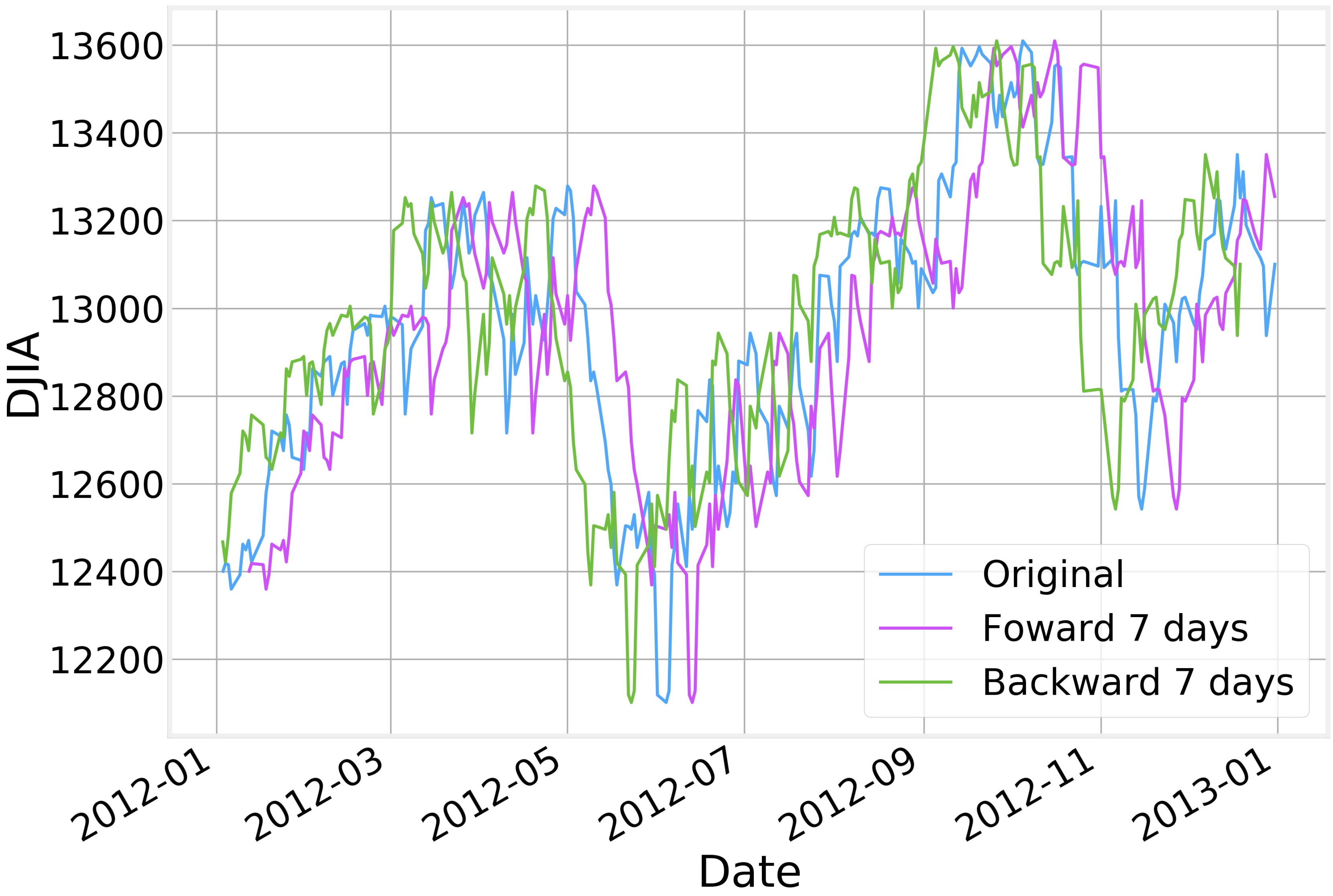
- While analyzing time series, we often refer to values that our time series took **1**, **2**, **3**, etc., time steps in the past
- These are known as lagged values and denoted:

$$X_{t-l}$$

- where l is the value of the lag we are considering.

Lagging Values

Lagging Values



Differences

- Perhaps the most common use case for lagged values is for the calculation of differences of the form:

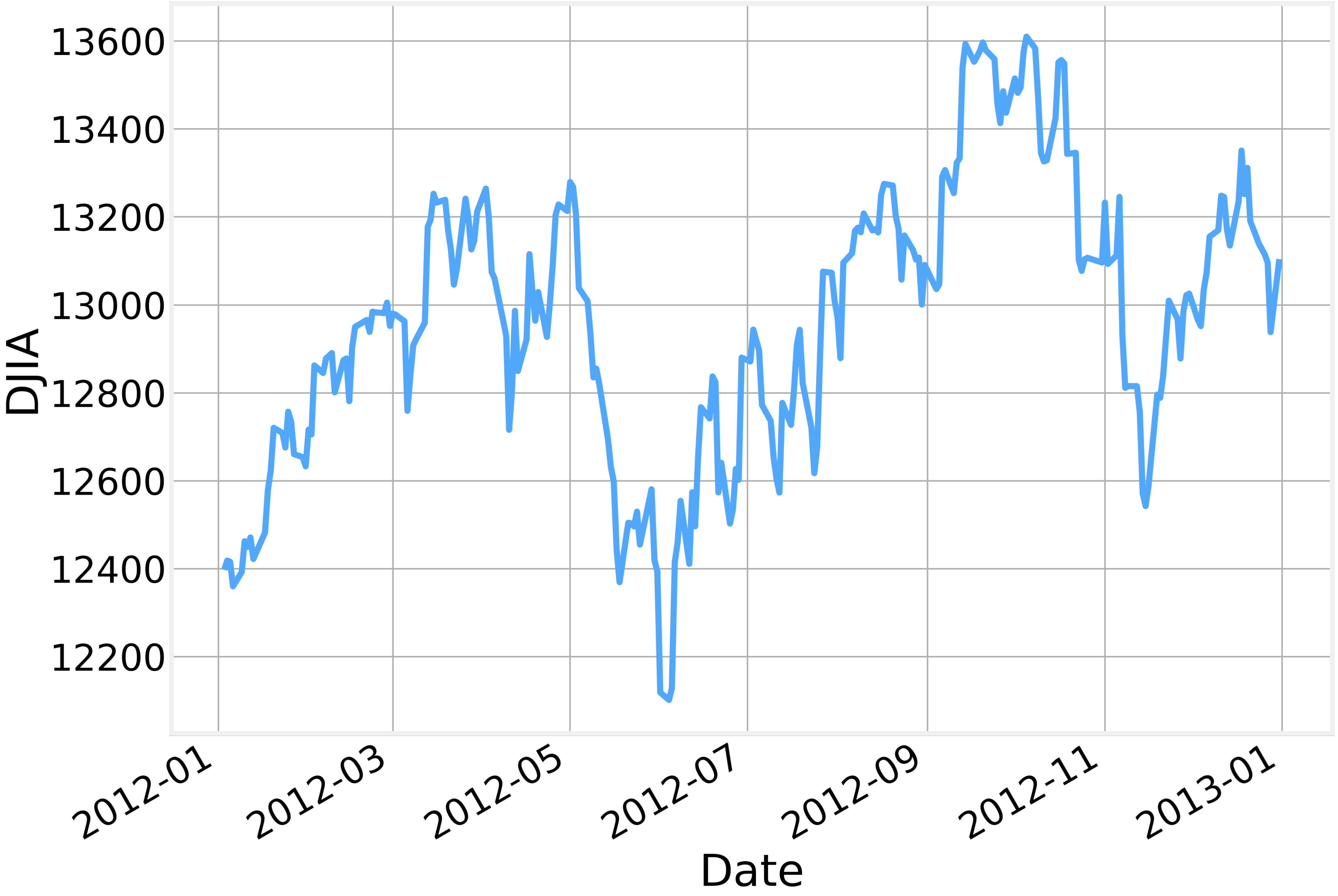
$$X_t - X_{t-l}$$

- Where $l \geq 1$ is the value of the lag we are interested in.
- Naturally, higher order differences can also be used, in which case, the difference of the difference is calculated:

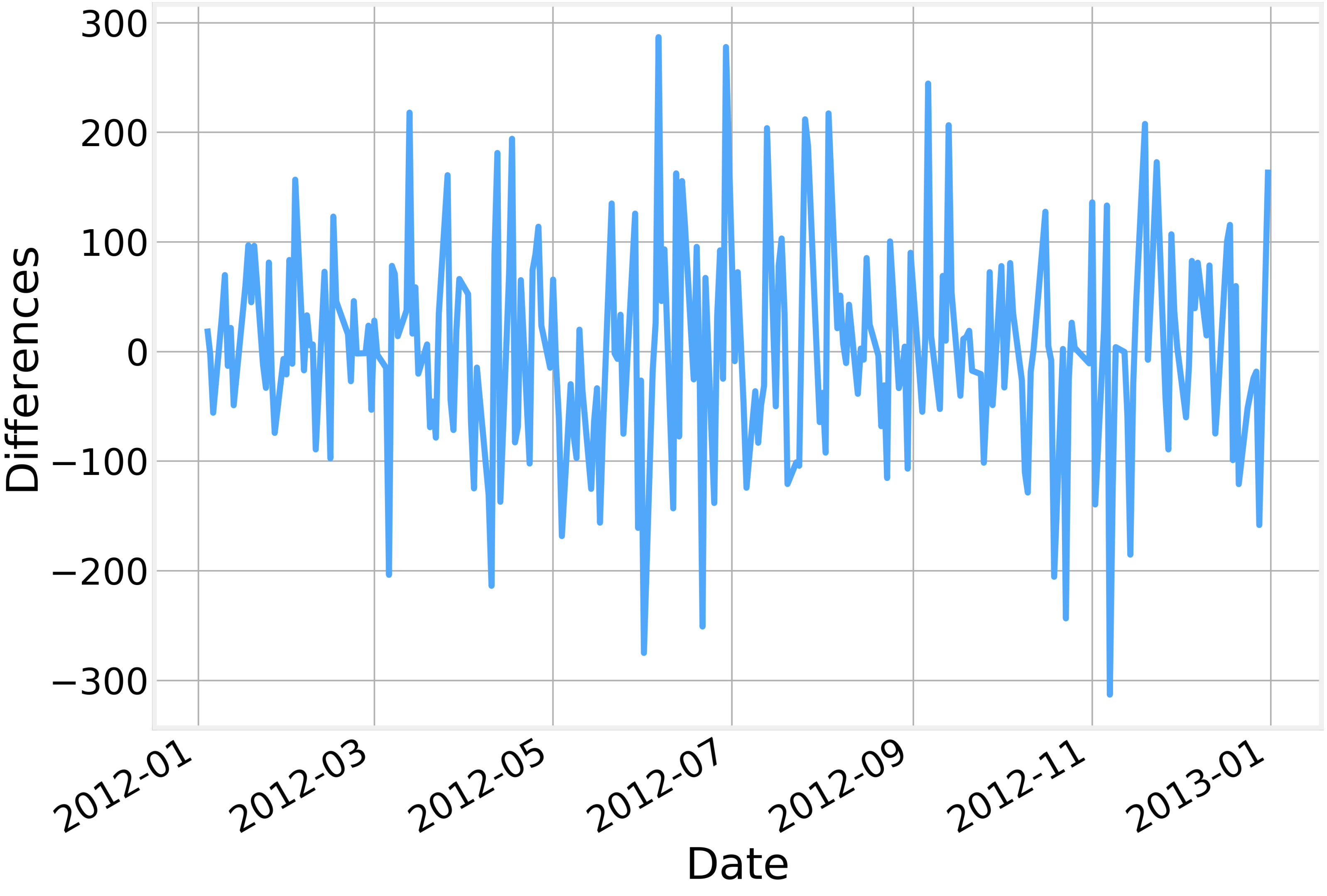
$$y_t = x_t - x_{t-l}$$
$$z_t = y_t - y_{t-l} \equiv x_t - 2x_{t-l} + x_t - 2l$$

- This can be thought of as a discrete version of the usual derivative of a function.
- Differences are also a particularly simple way to detrend a time series

Differences



Differences



Windowing

- When analyzing the temporal behavior of of a signal, we often need to evaluate if specific quantities are time varying or not

Windowing

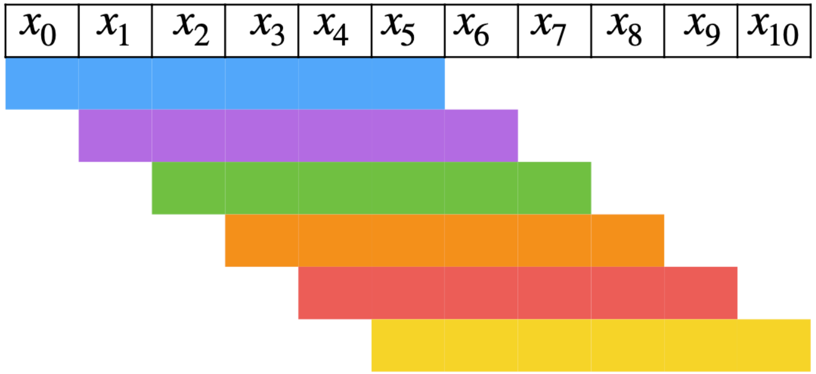
- When analyzing the temporal behavior of of a signal, we often need to evaluate if specific quantities are time varying or not
- A common approach is to use sliding windows of a given length to evaluate the required values

Windowing

- When analyzing the temporal behavior of of a signal, we often need to evaluate if specific quantities are time varying or not
- A common approach is to use sliding windows of a given length to evaluate the required values
- So, a sliding window of width **6** on a series of length **11** would look like:

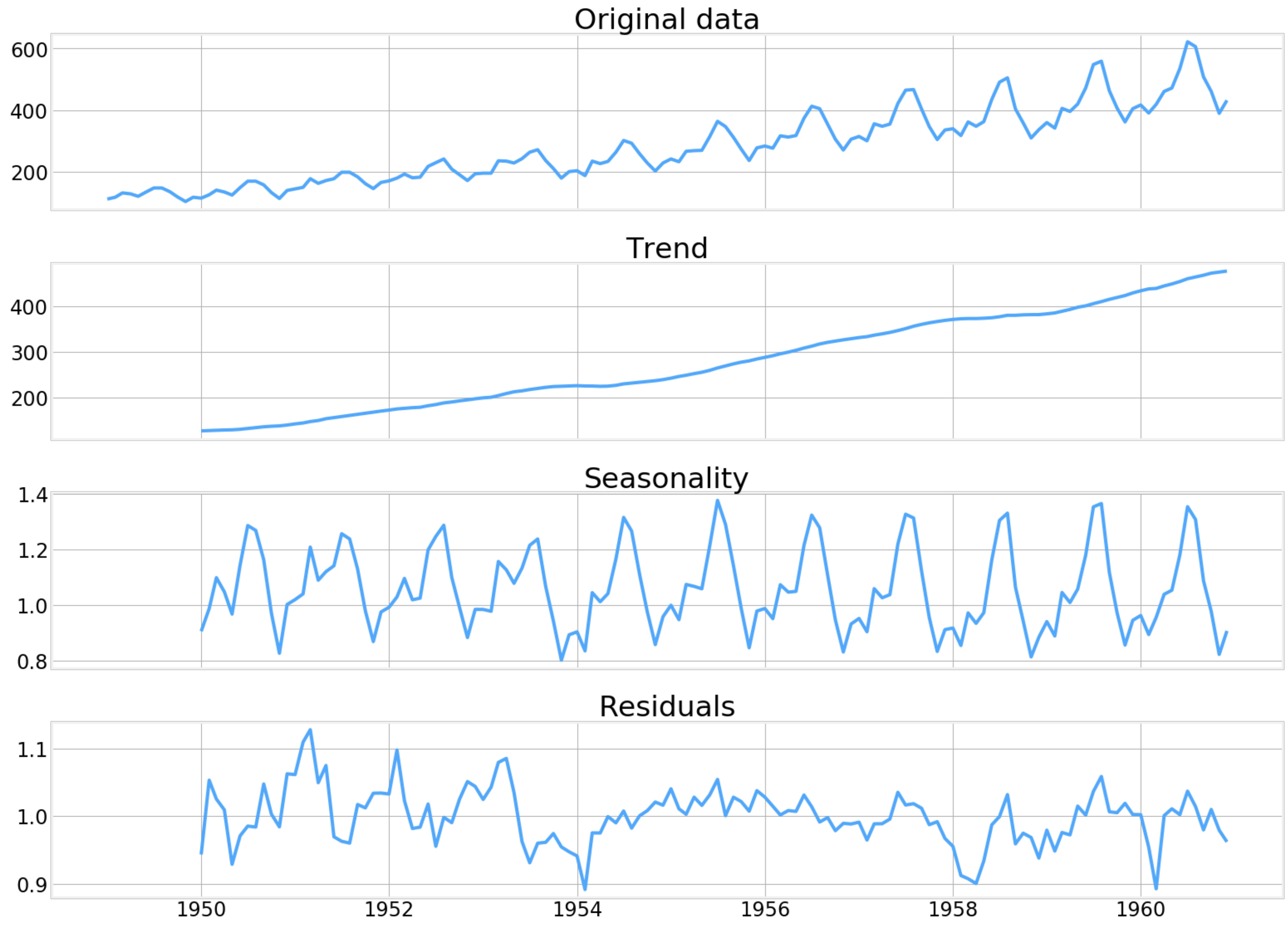
Windowing

- When analyzing the temporal behavior of of a signal, we often need to evaluate if specific quantities are time varying or not
- A common approach is to use sliding windows of a given length to evaluate the required values
- So, a sliding window of width **6** on a series of length **11** would look like:



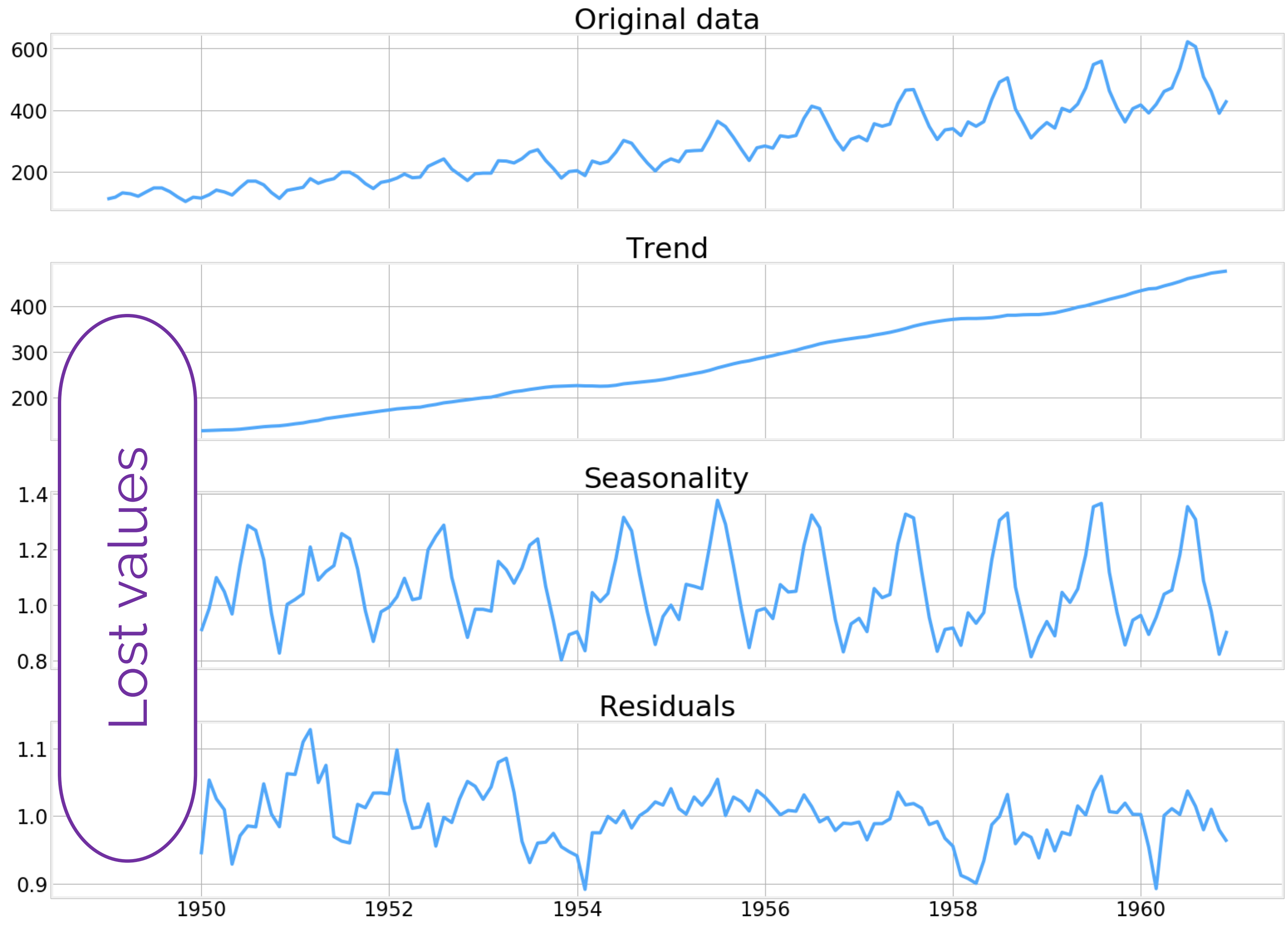
$$= \begin{matrix} x_0 & x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \\ x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} \end{matrix}$$

Windowing



Original data
600
400
200
Trend
400
300
200
Seasonality
1.4
1.2
1.0
0.8
Residuals
1.1
1.0
0.9
1950
1952
1954
1956
1958
1960

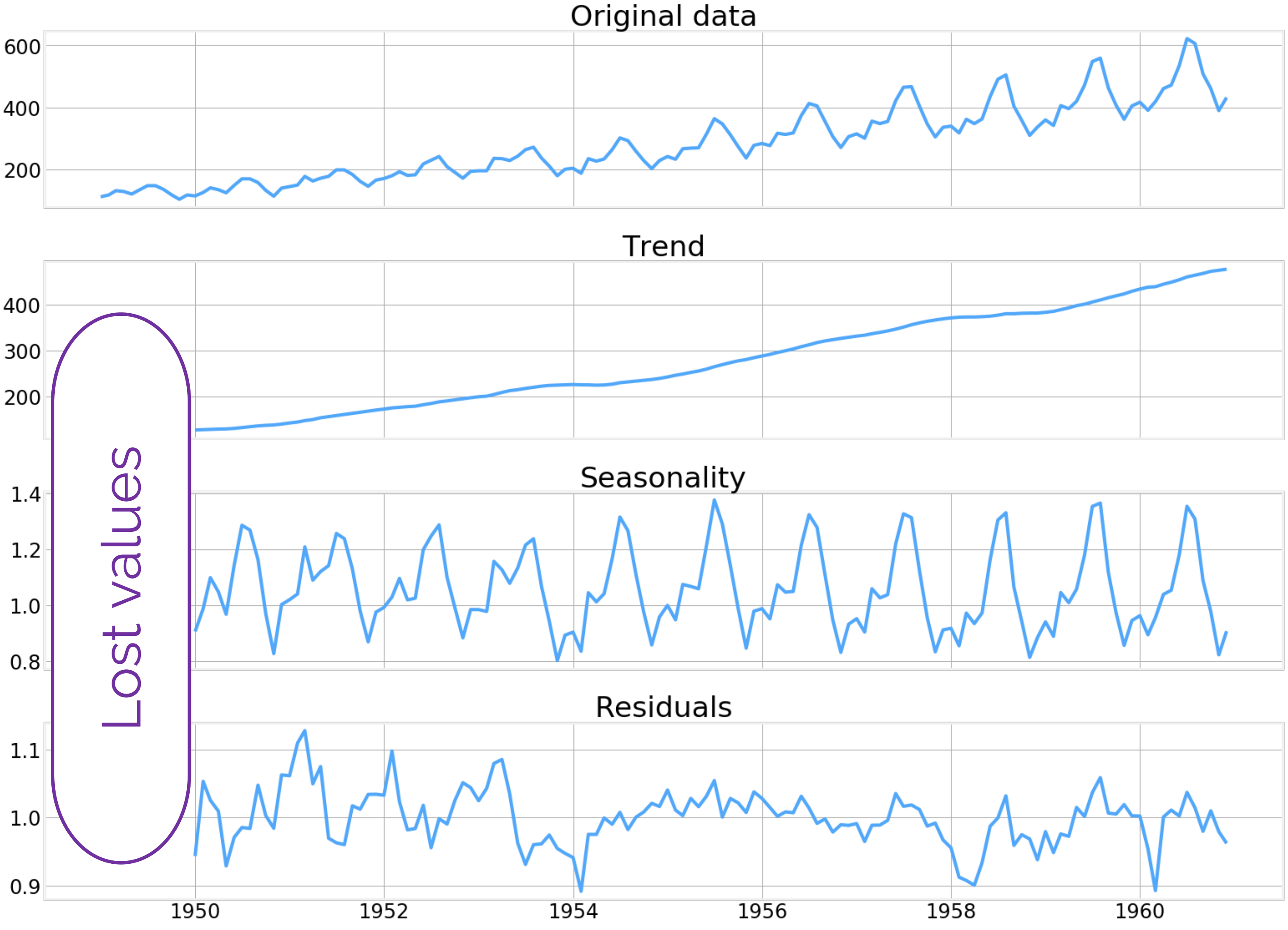
Windowing



Original data
600
400
200
Trend
400
300
200
Seasonality
1.4
1.2
1.0
0.8
Residuals
1.1
1.0
0.9
1950
1952
1954
1956
1958
1960
Lost values

Windowing

One common approach is to place all "lost values" at the beginning as it avoids "future leaking" when splitting the dataset



Running Values

- In the first part of the lecture, we already used running averages to detrend a time series

Running Values

- In the first part of the lecture, we already used running averages to detrend a time series
- Other common metrics are:
  - Variance
  - Maximum value
  - Minimum value
  - etc...

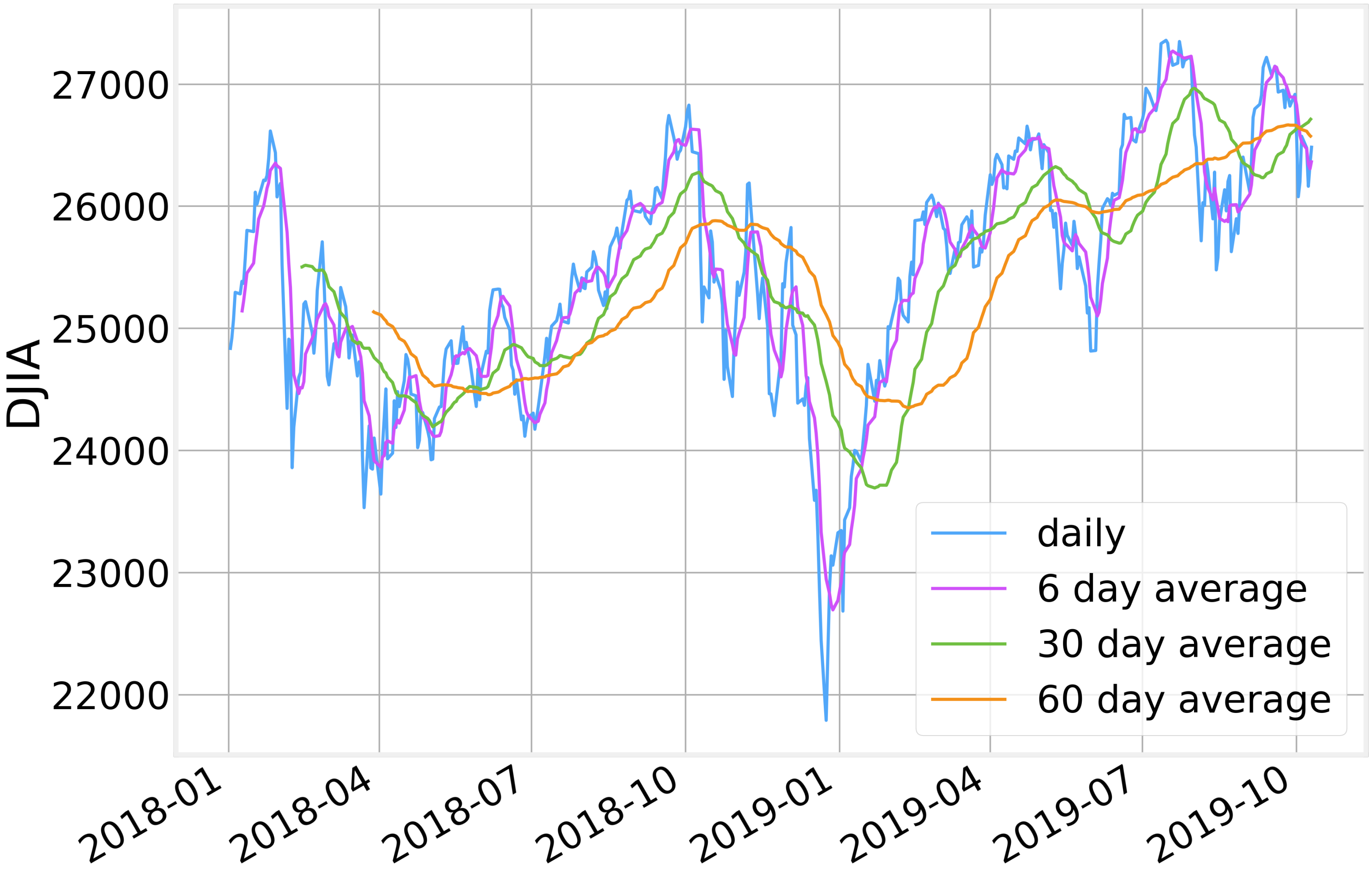
Running Values

- In the first part of the lecture, we already used running averages to detrend a time series
- Other common metrics are:
  - Variance
  - Maximum value
  - Minimum value
  - etc…
- One important detail to note is that while using windowing to calculate running values we “lose” a number of points equal to the width of the window

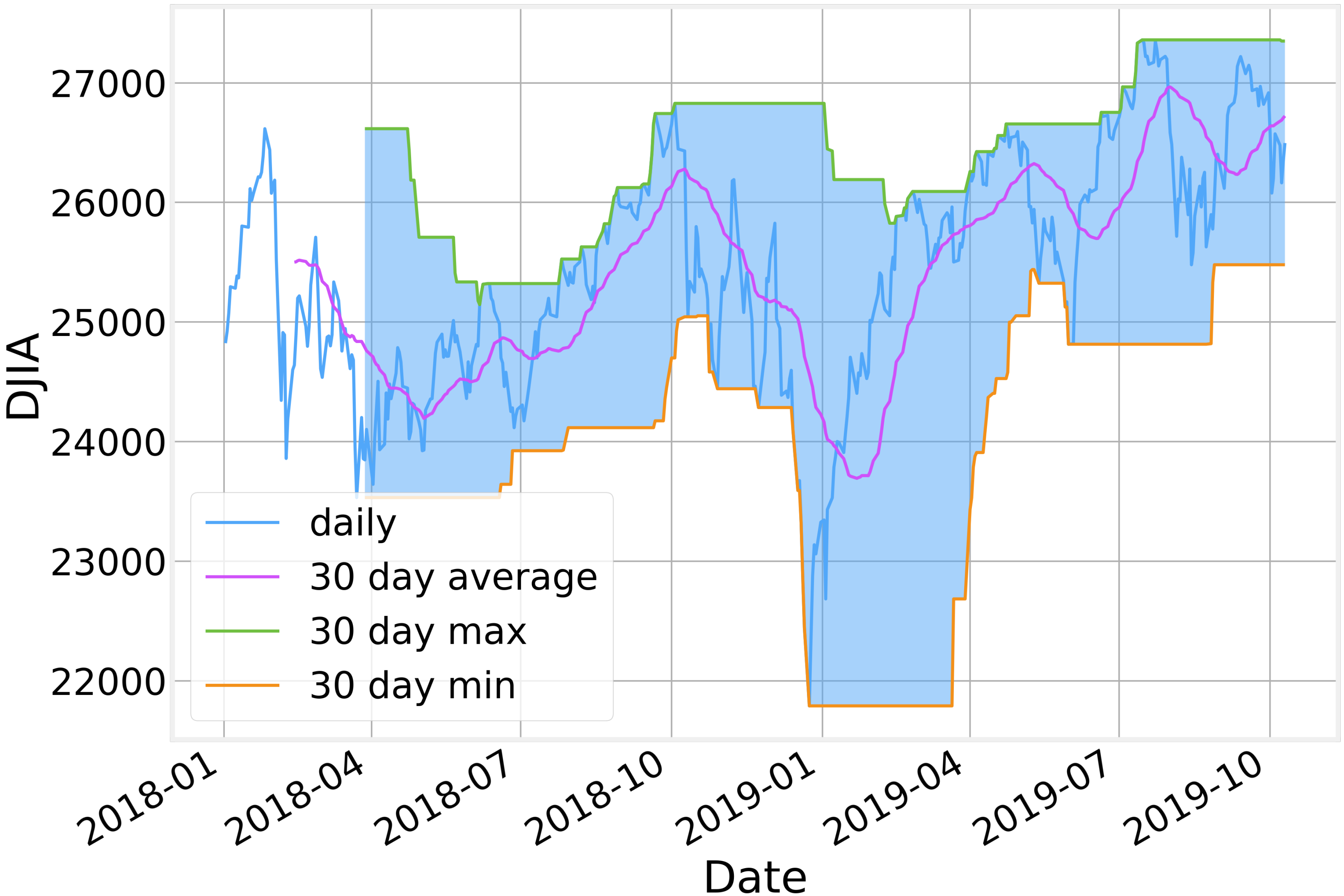
Running Values

- In the first part of the lecture, we already used running averages to detrend a time series
- Other common metrics are:
  - Variance
  - Maximum value
  - Minimum value
  - etc…
- One important detail to note is that while using windowing to calculate running values we “lose” a number of points equal to the width of the window
- Depending on the application we can choose to place the missing values in either or (or even both) extremes of the time interval

Running Values



Envelopes

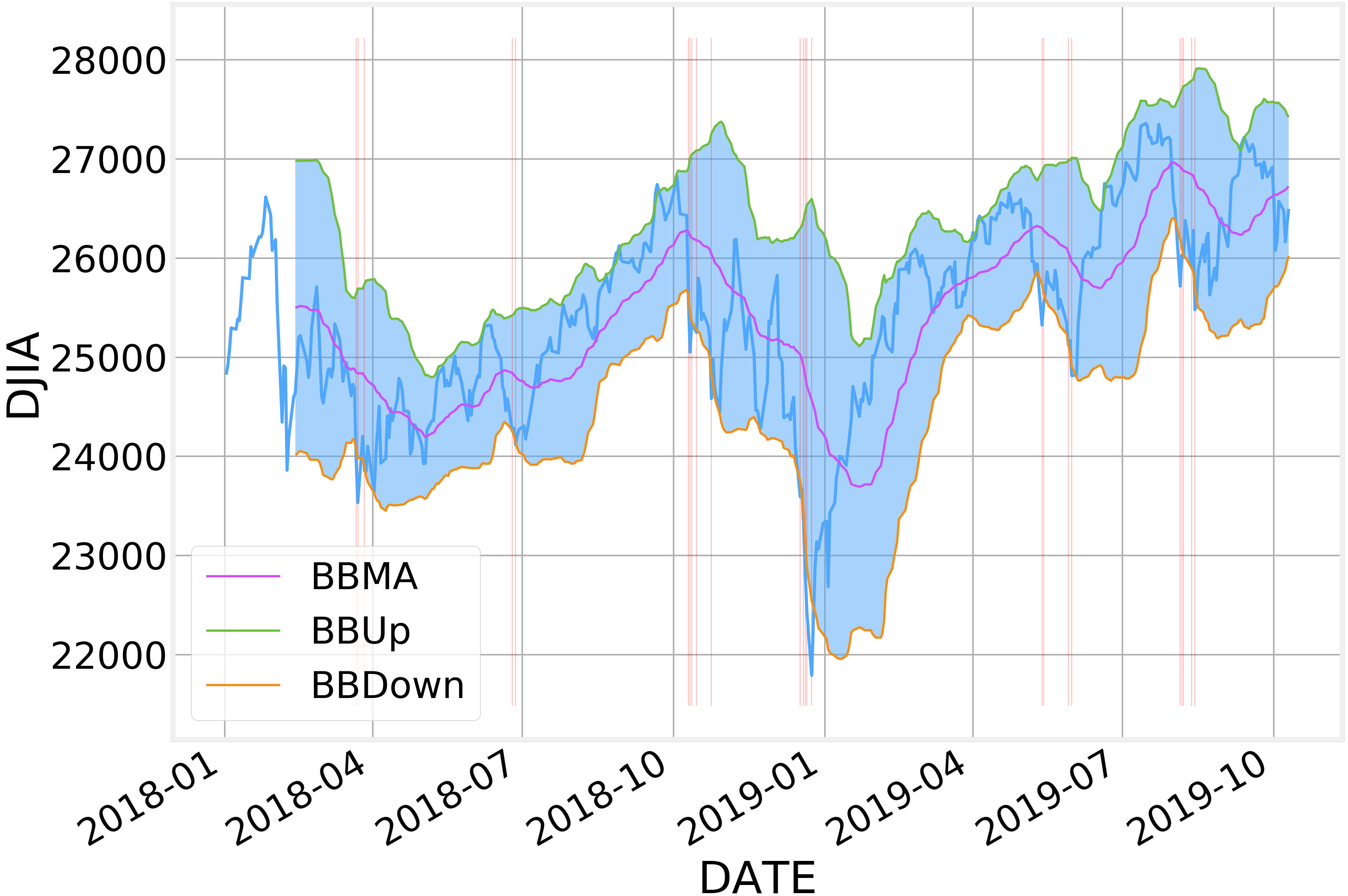


27000
26000
25000
24000
23000
22000
DJIA
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
2019-07
2019-10
Date
daily
30 day average
30 day max
30 day min

Bollinger Bands

- A common use for application for running values is the calculation of Bollinger Bands.
- Introduced by John Bollinger in the 1980s as a complement to more traditional time series technical analysis techniques.
- Bollinger Bands are defined by two components:
  - A $\boldsymbol{N}$ period moving average, μ_N
  - The area $\boldsymbol{K}$ standard deviations above and below the moving average $\mu_N \pm K\sigma_N$
- Both μ_N and σ_N are computed on a running window of size $\boldsymbol{N}$
- The values $\boldsymbol{N}$ and $\boldsymbol{K}$ are application specific. For stock trading, $\boldsymbol{N = 20}$ and $\boldsymbol{K = 2}$
- Whenever the time series steps out of the Bollinger Band that's a clear indication of a change in the temporal behavior.

Bollinger Bands



Exponential Running Average

- One alternative to a simple running average is Exponential Smoothing

Exponential Running Average

- One alternative to a simple running average is Exponential Smoothing
- The exponentially “smooth” version of a time series is given by:

$$z_t = \alpha x_t + (1 - \alpha) z_{t-1}$$

Exponential Running Average

- One alternative to a simple running average is Exponential Smoothing
- The exponentially "smooth" version of a time series is given by:

$$z_t = \alpha x_t + (1 - \alpha) z_{t-1}$$

- The smaller the value of the weight α, the less influence each point has on the transformed time series.

Exponential Running Average

- One alternative to a simple running average is Exponential Smoothing
- The exponentially "smooth" version of a time series is given by:

$$z_t = \alpha x_t + (1 - \alpha) z_{t-1}$$

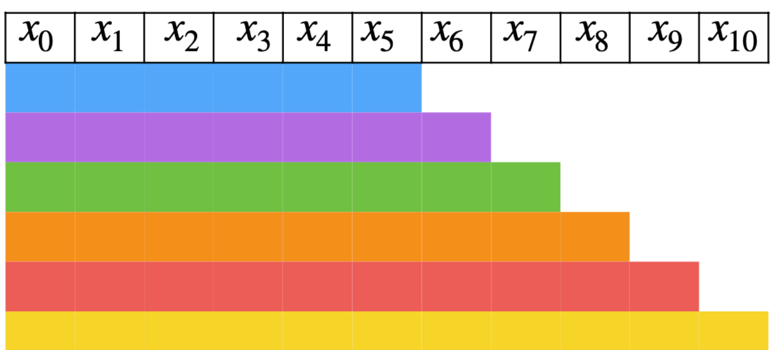
- The smaller the value of the weight α, the less influence each point has on the transformed time series.
- Each point depends implicitly on all previous points

Exponential Running Average

- One alternative to a simple running average is Exponential Smoothing
- The exponentially "smooth" version of a time series is given by:

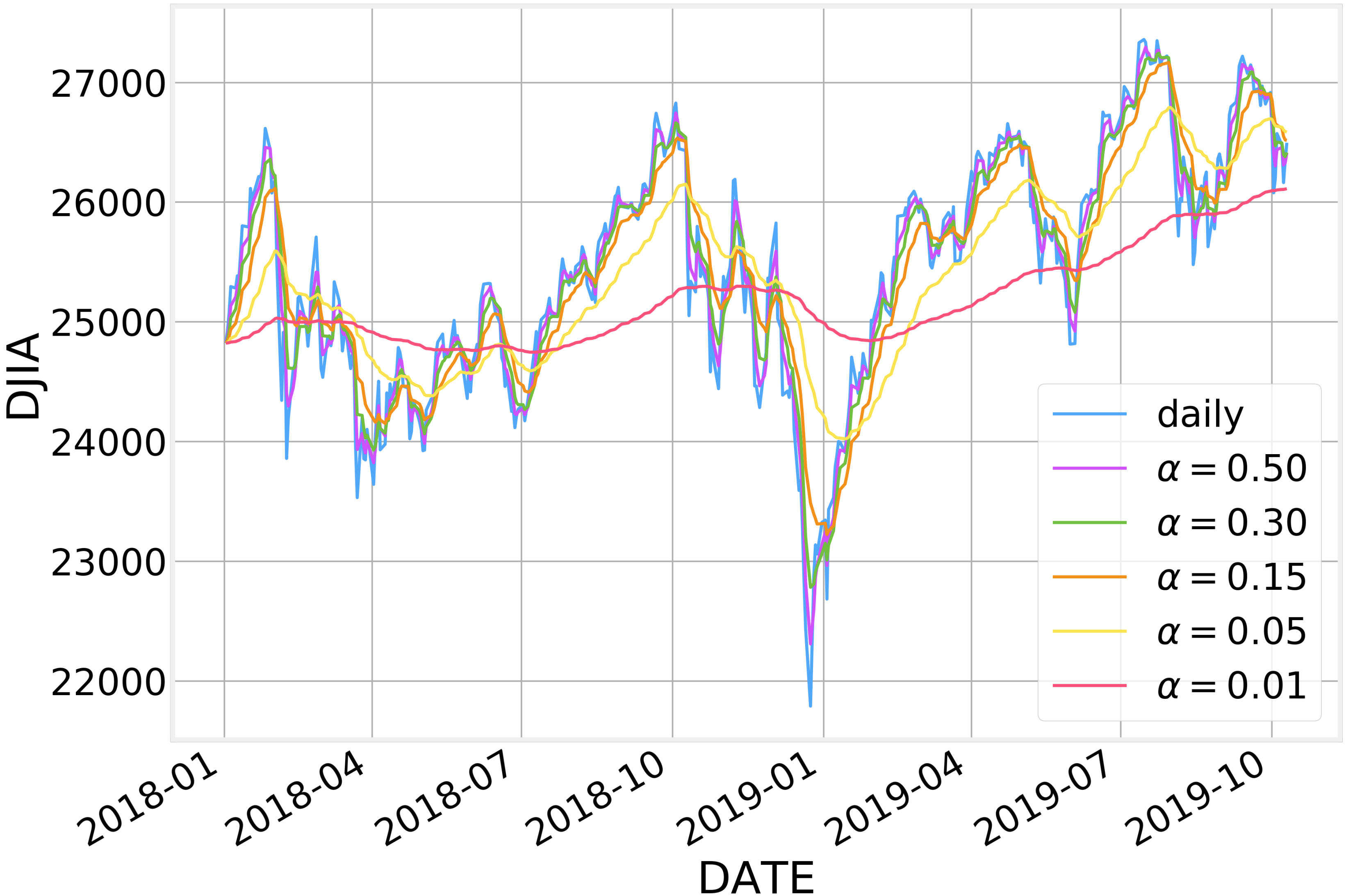
$$z_t = \alpha x_t + (1 - \alpha) z_{t-1}$$

- The smaller the value of the weight α, the less influence each point has on the transformed time series.
- Each point depends implicitly on all previous points

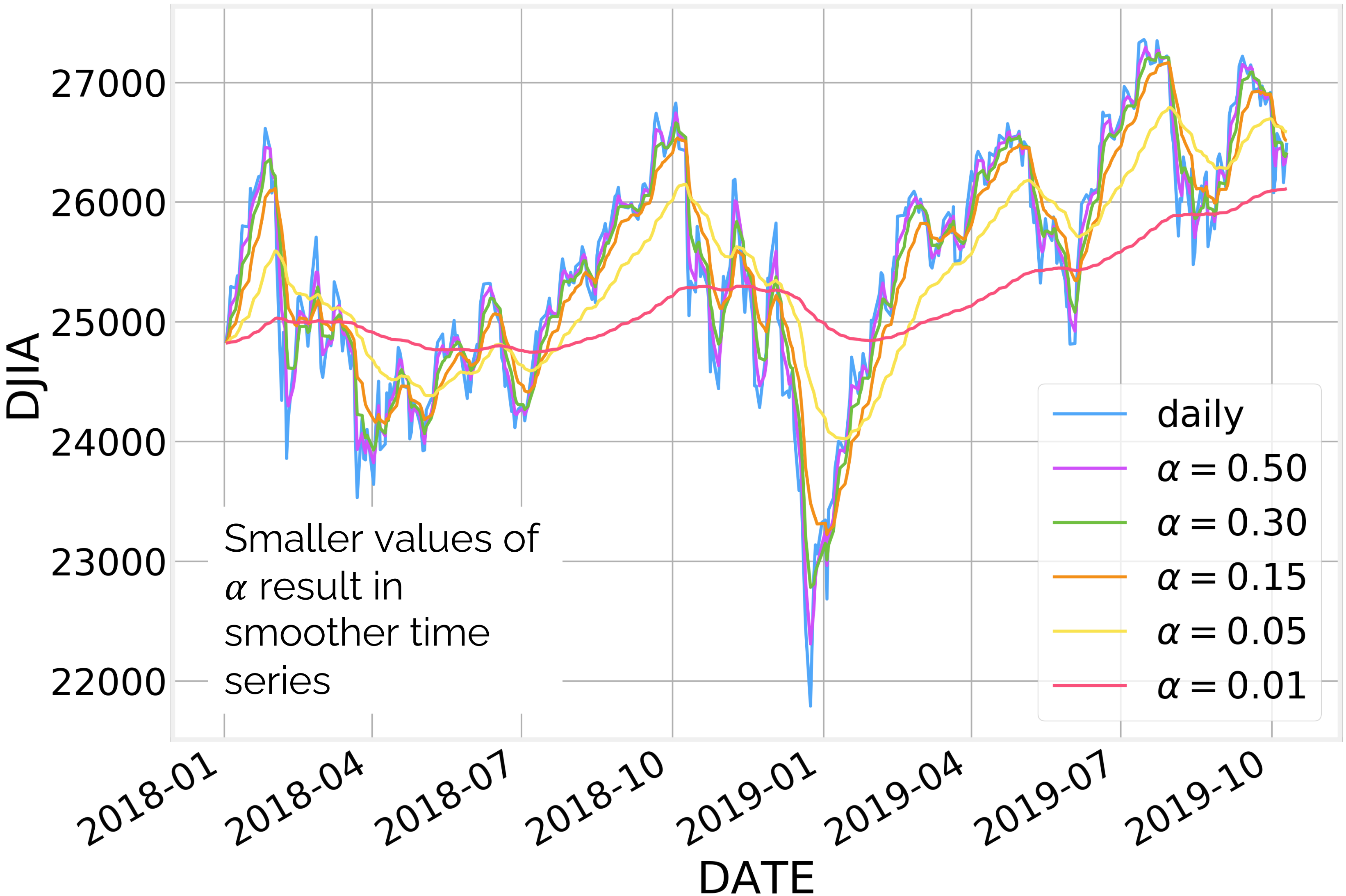


$$\begin{pmatrix} \alpha & & & & \\ \alpha(1-\alpha)^1 & \alpha & & & \\ \alpha(1-\alpha)^2 & \alpha(1-\alpha)^1 & \alpha & & \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha(1-\alpha)^{n-1} & \alpha(1-\alpha)^{n-2} & \alpha(1-\alpha)^{n-3} & & \alpha \end{pmatrix}$$

Exponential Running Average



Exponential Running Average



Forecasting

- We can also use the Exponential Moving Averages as a simple forecasting tool.

Forecasting

- We can also use the Exponential Moving Averages as a simple forecasting tool.
- The value at time $t + 1$ is given by:

$$z_{t+1} = \alpha x_t + (1 - \alpha) z_{t-1}$$

Forecasting

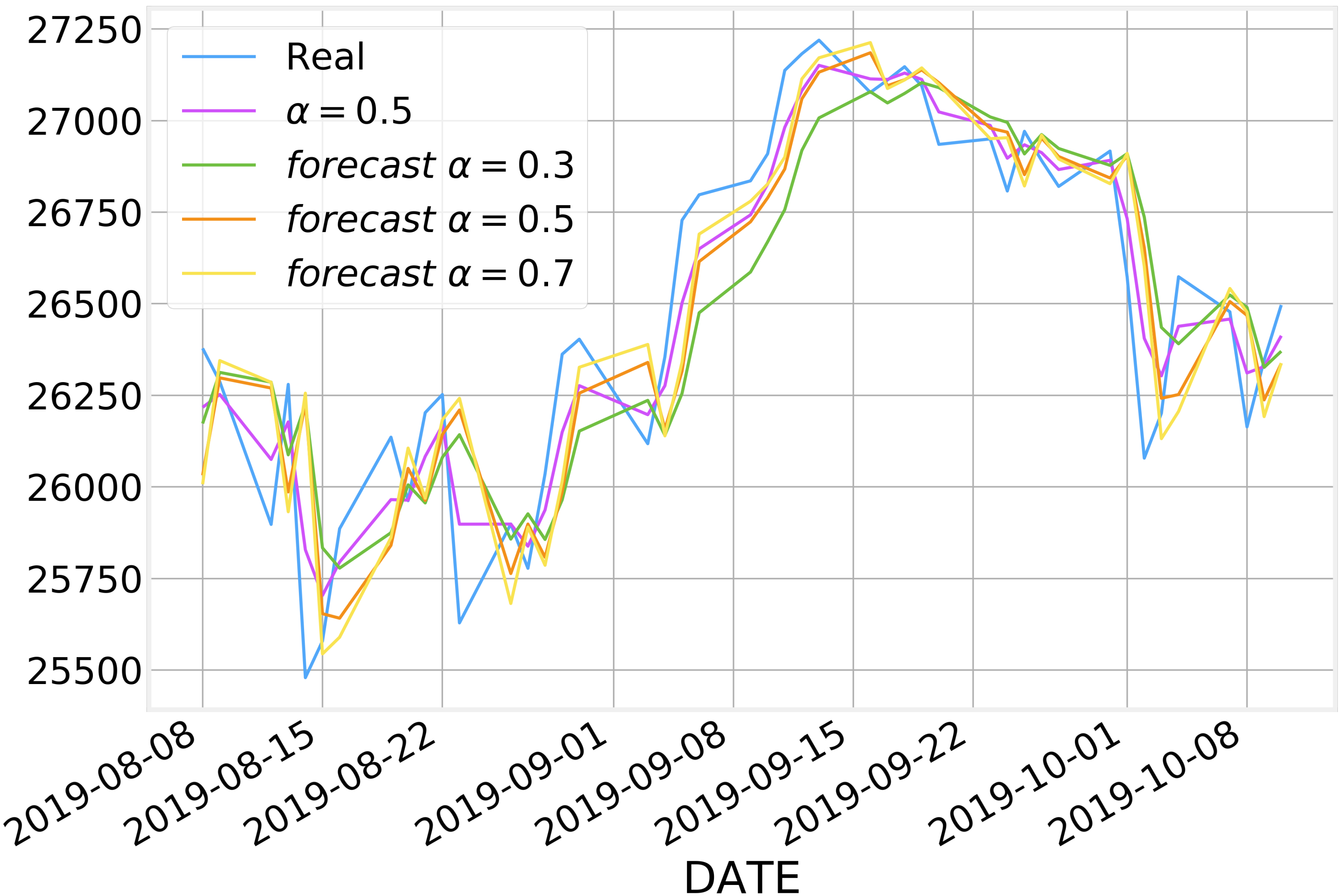
- We can also use the Exponential Moving Averages as a simple forecasting tool.
- The value at time $t + 1$ is given by:

$$z_{t+1} = \alpha x_t + (1 - \alpha) z_{t-1}$$

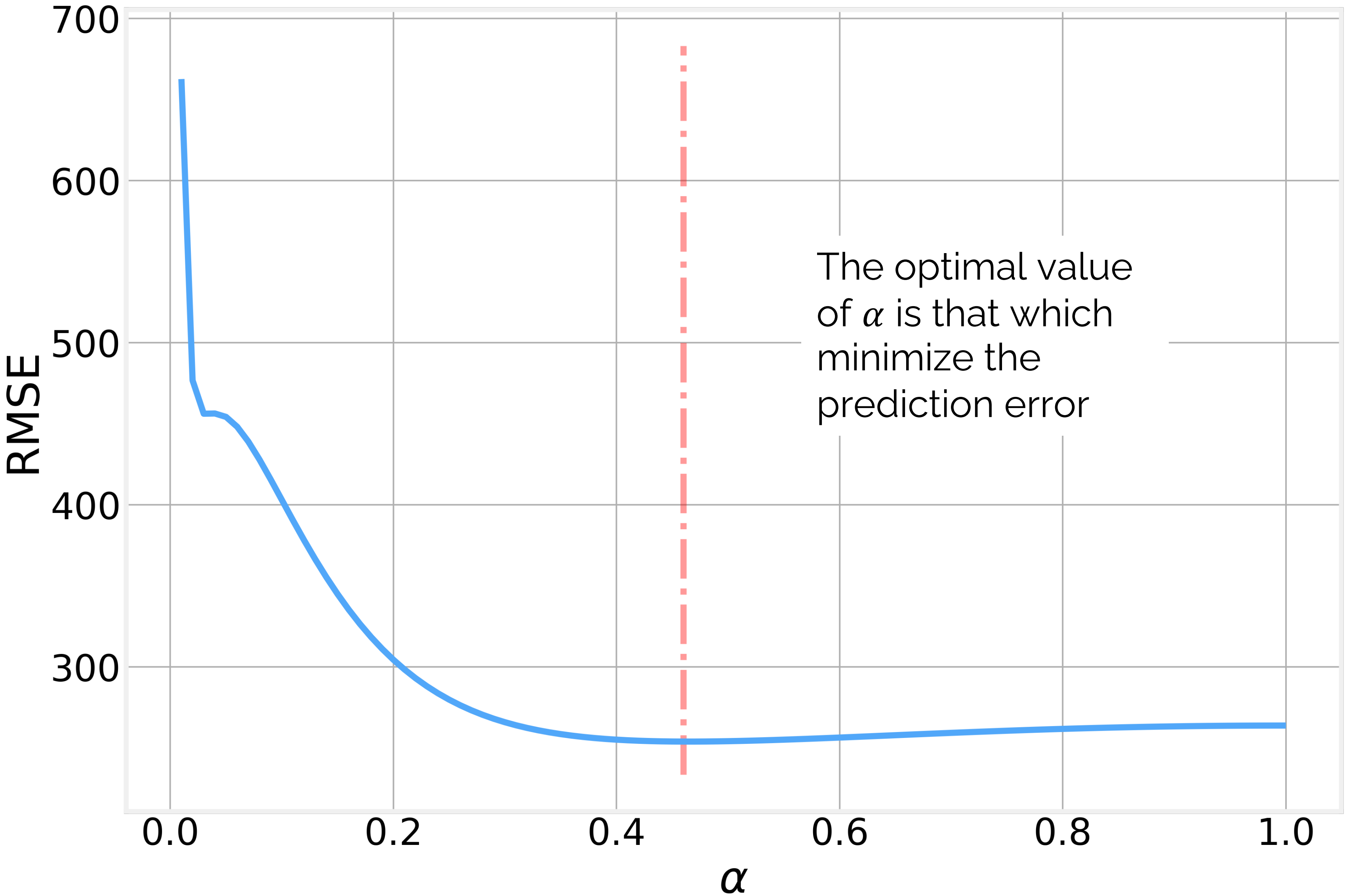
- Which we can consider to be a prediction on the value of x_{t+1}, based on the current value of z_t and some factor of out current error value $x_t - z_t$:

$$z_{t+1} = z_t + \alpha(x_t - z_t)$$

Forecasting

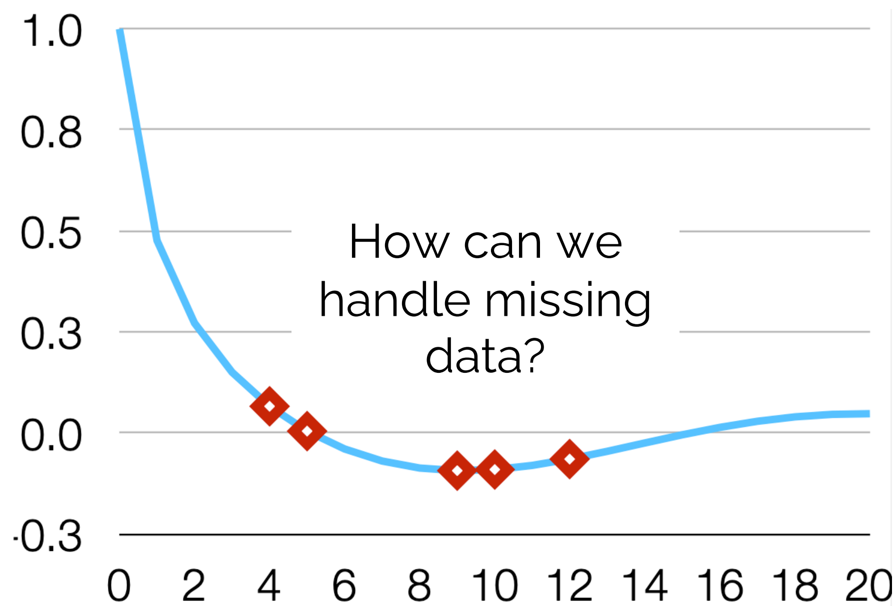


Forecasting



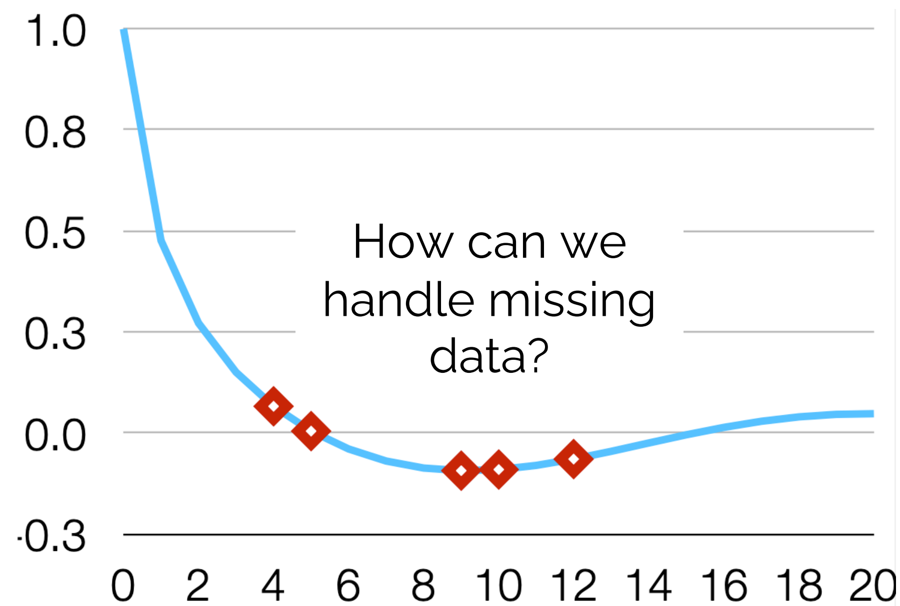
Fill Methods

- Sometimes the time series is incomplete



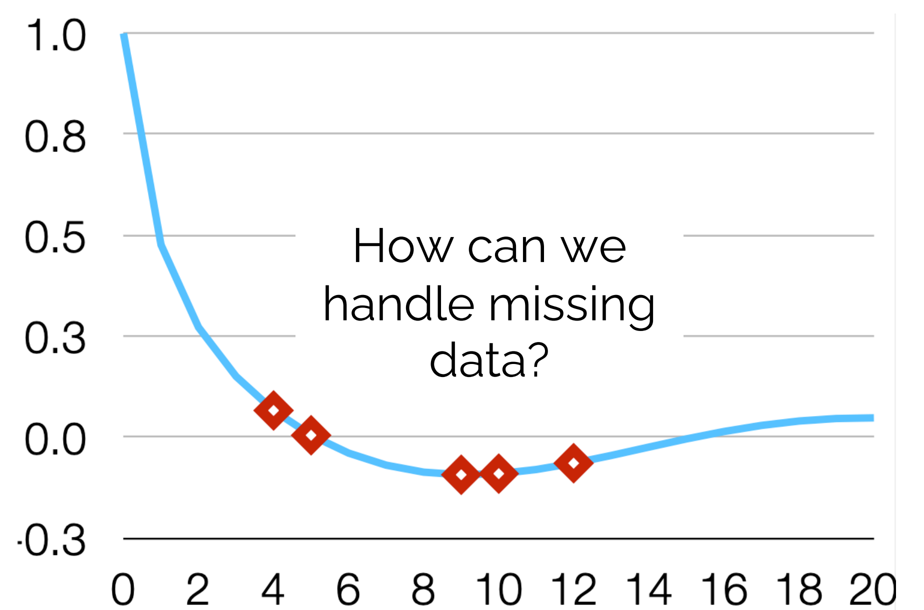
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.



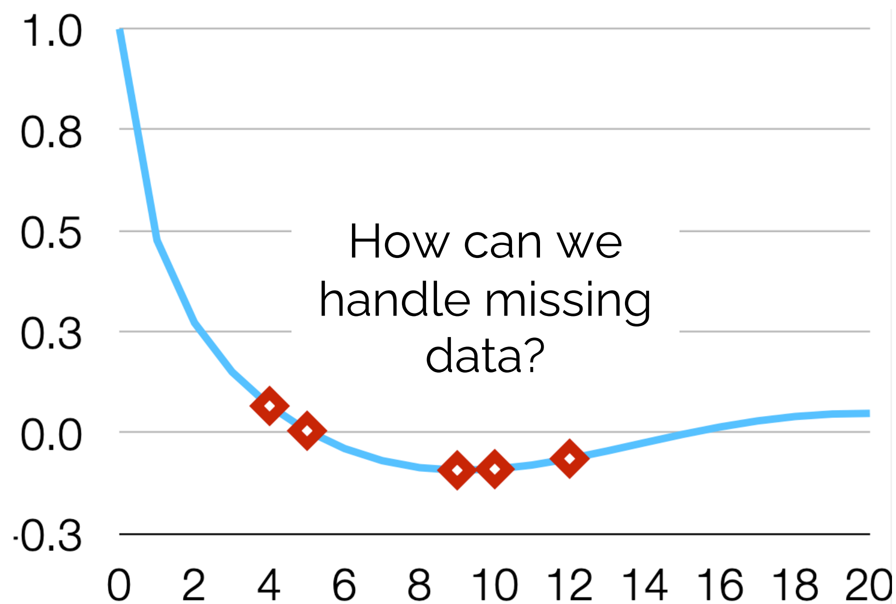
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan



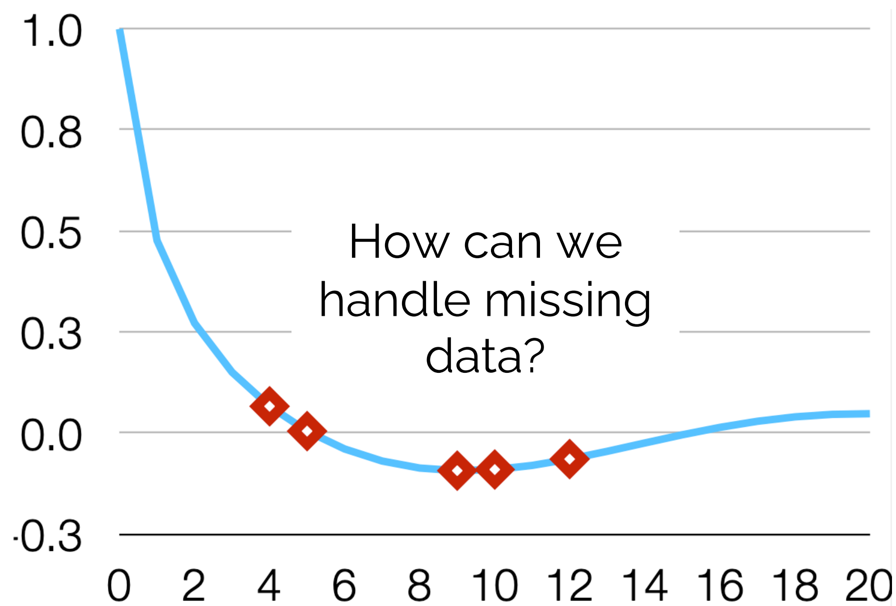
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan
- Several techniques have been developed to handle this case



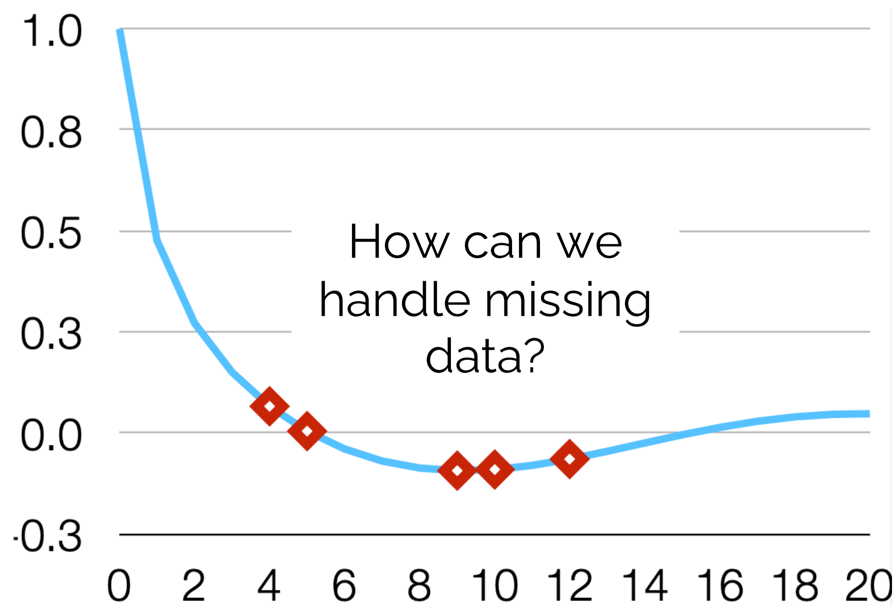
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan
- Several techniques have been developed to handle this case
  - forward fill – keep the last valid value



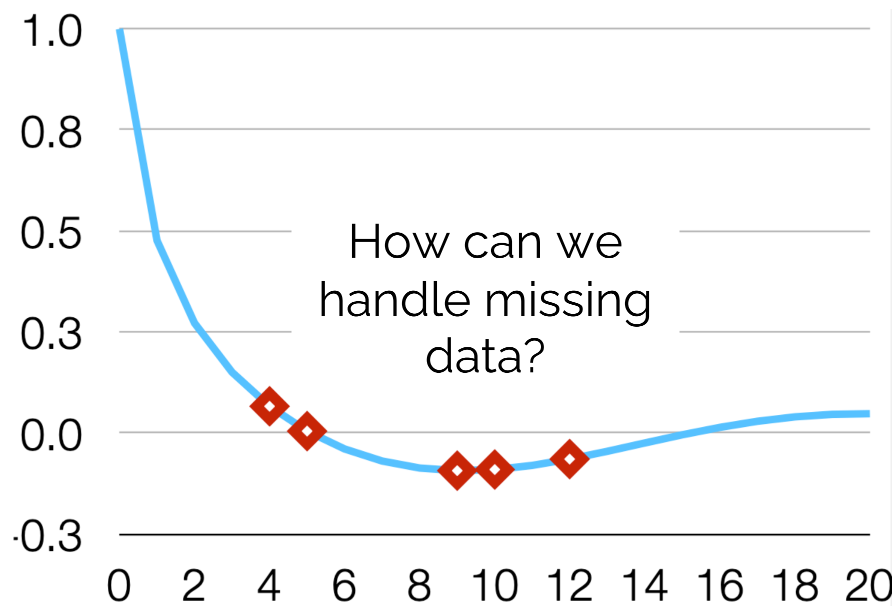
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan
- Several techniques have been developed to handle this case
  - forward fill – keep the last valid value
  - back fill – keep the next valid value



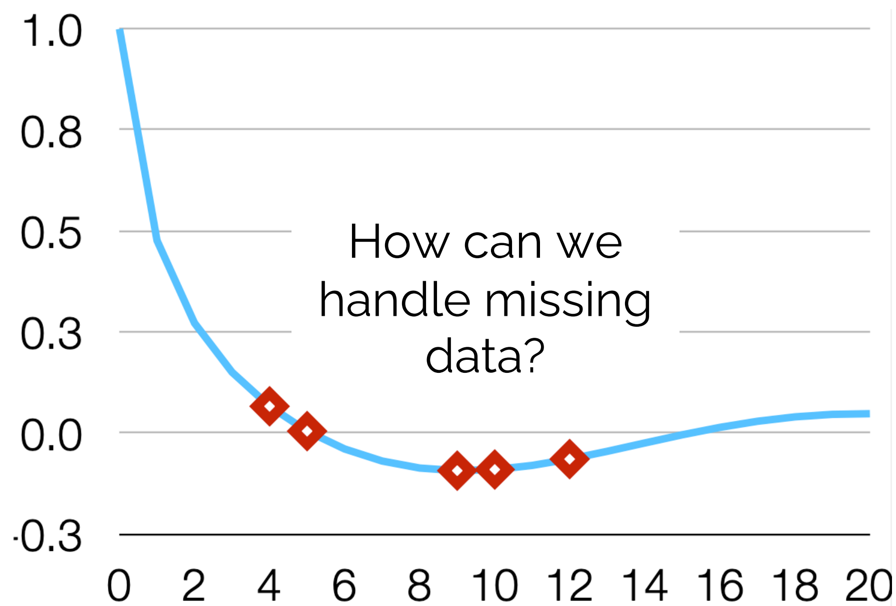
Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan
- Several techniques have been developed to handle this case
  - forward fill – keep the last valid value
  - back fill – keep the next valid value
  - interpolate – add values by interpolating between the previous and the next value

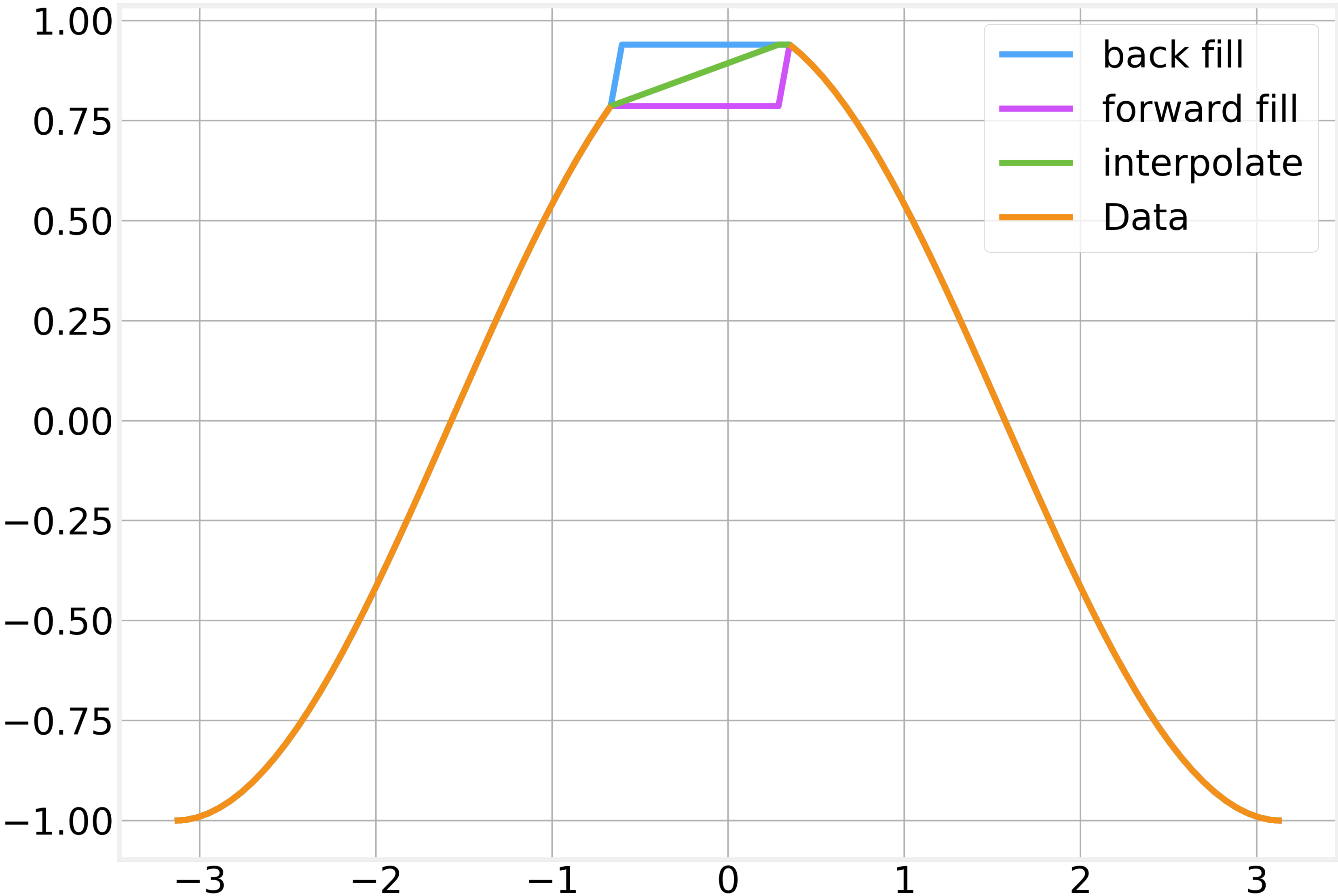


Fill Methods

- Sometimes the time series is incomplete
- Missing data points can be due to data corruption, data collection issues, etc.
- Missing values are represented as nan
- Several techniques have been developed to handle this case
  - forward fill – keep the last valid value
  - back fill – keep the next valid value
  - interpolate – add values by interpolating between the previous and the next value
  - imputation – add values based on what we expect the missing values to be



Fill Methods



Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.

Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.
- In many cases, our analysis requires that we resample the data to a different time scale

Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.
- In many cases, our analysis requires that we resample the data to a different time scale
- Resampling to a longer timescale is relatively simple and similar to aggregation:

Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.
- In many cases, our analysis requires that we resample the data to a different time scale
- Resampling to a longer timescale is relatively simple and similar to aggregation:
  - Transforming from daily to weekly frequency requires simply aggregating by week

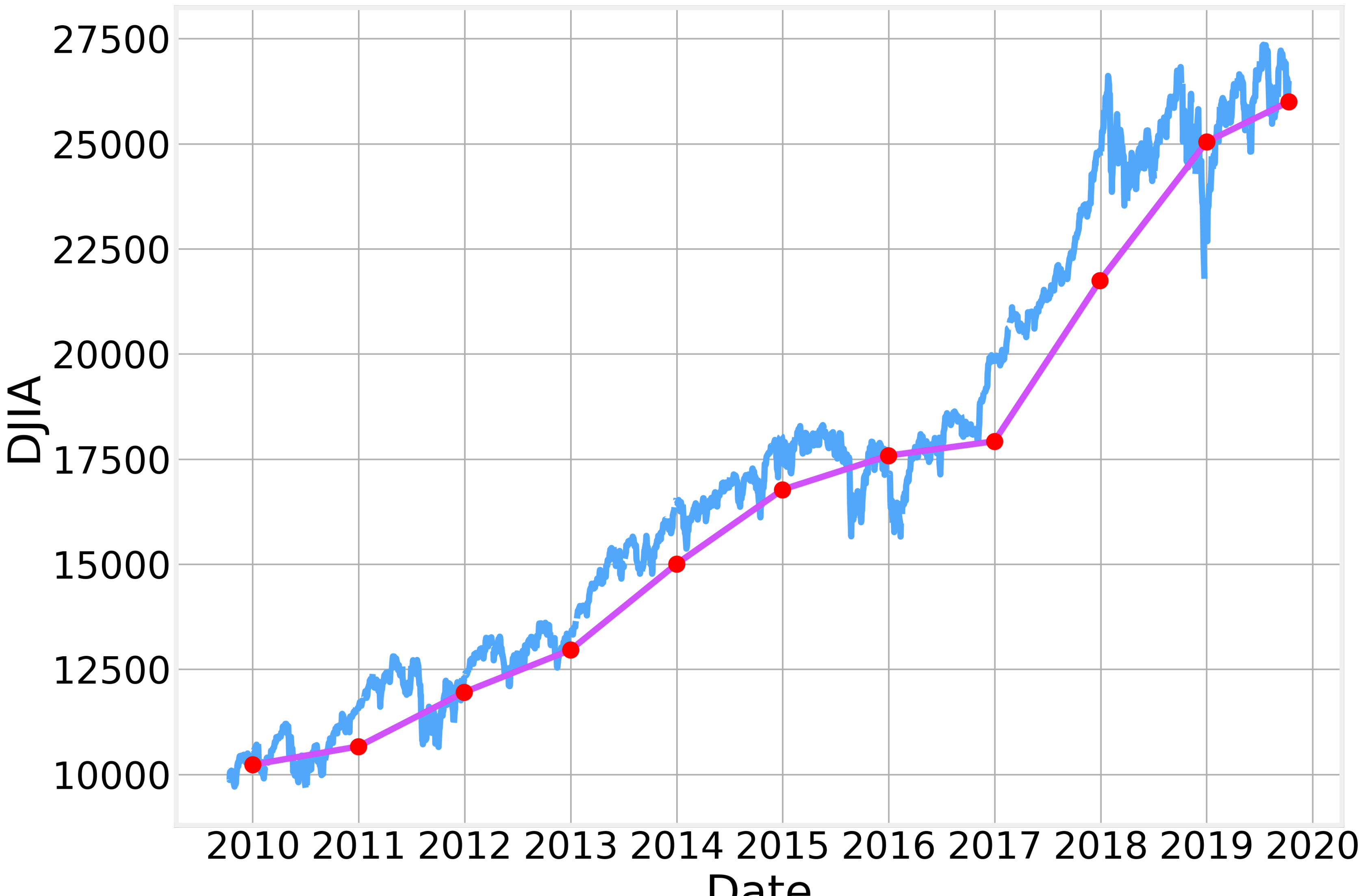
Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.
- In many cases, our analysis requires that we resample the data to a different time scale
- Resampling to a longer timescale is relatively simple and similar to aggregation:
  - Transforming from daily to weekly frequency requires simply aggregating by week
- Resampling to shorter timescales requires interpolation or imputation to make up for the missing values

Resampling

- Time series typically have an intrinsic time scale at which the data was collected: ticks, seconds, days, months, etc.
- In many cases, our analysis requires that we resample the data to a different time scale
- Resampling to a longer timescale is relatively simple and similar to aggregation:
  - Transforming from daily to weekly frequency requires simply aggregating by week
- Resampling to shorter timescales requires interpolation or imputation to make up for the missing values
  - Going from weekly to daily frequency requires specifying how to allocate the values for each day of the week

Resampling



27500
25000
22500
20000
17500
15000
12500
10000
DJIA
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Date

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance

- Bootstrapping measures the desired property in a large number of samples (with replacement), of the observed dataset.

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance
- Bootstrapping measures the desired property in a large number of samples (with replacement), of the observed dataset.
- Each sample has equal size to the observed dataset.

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance
- Bootstrapping measures the desired property in a large number of samples (with replacement), of the observed dataset.
- Each sample has equal size to the observed dataset.
- From the entire population of samples, the empirical bootstrap distribution of the expected values can be obtained to provide information about the distribution in the total population

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance
- Bootstrapping measures the desired property in a large number of samples (with replacement), of the observed dataset.
- Each sample has equal size to the observed dataset.
- From the entire population of samples, the empirical bootstrap distribution of the expected values can be obtained to provide information about the distribution in the total population

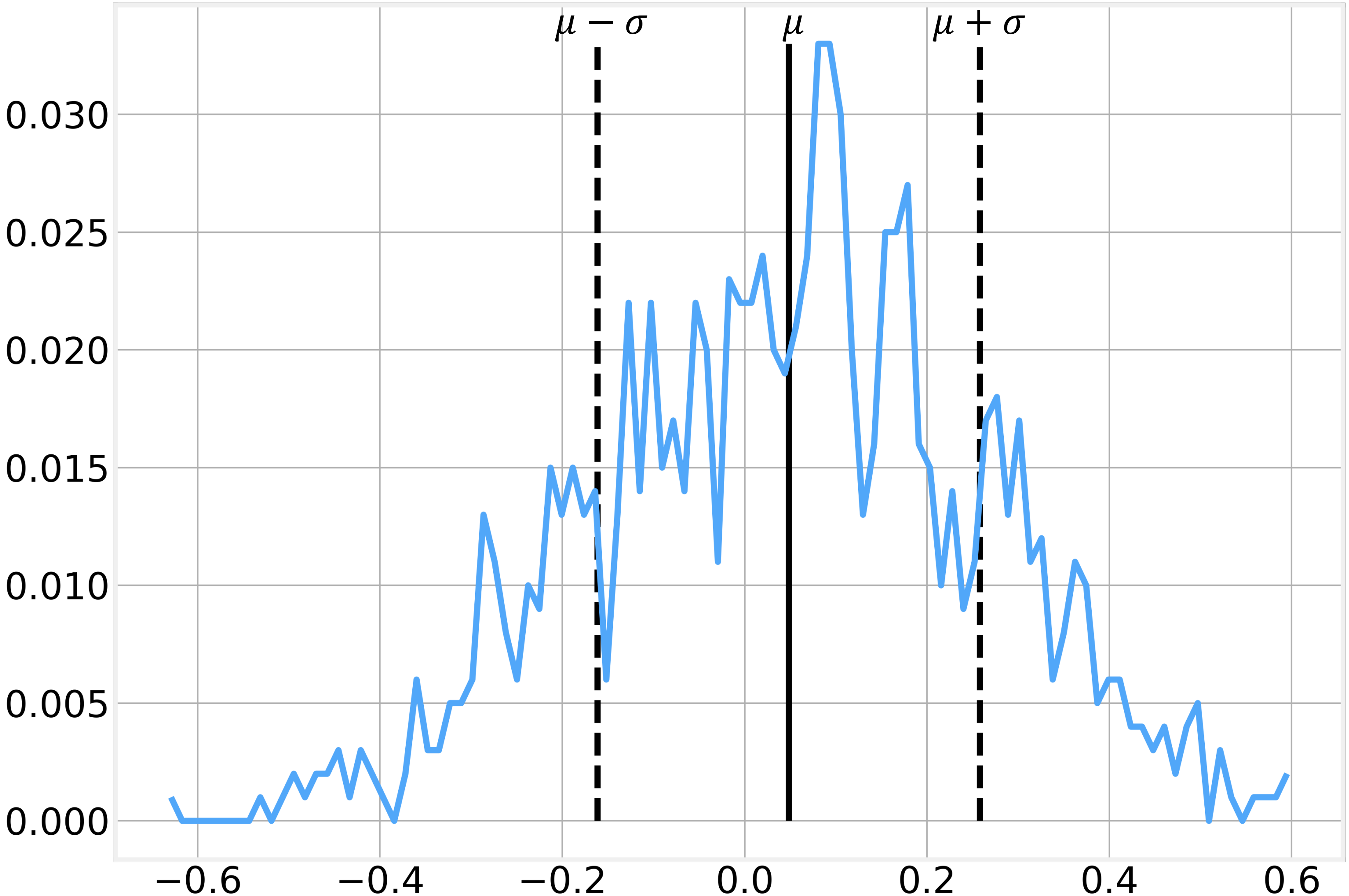
Also see [Jackknife Estimation]

Bootstrapping

- Bootstrapping is another alternative to estimate statistical properties such as mean and variance
- Bootstrapping measures the desired property in a large number of samples (with replacement), of the observed dataset.
- Each sample has equal size to the observed dataset.
- From the entire population of samples, the empirical bootstrap distribution of the expected values can be obtained to provide information about the distribution in the total population

Also see [Jackknife Estimation]

Bootstrapping



Code – Transformations