Time Series for Data Scientists

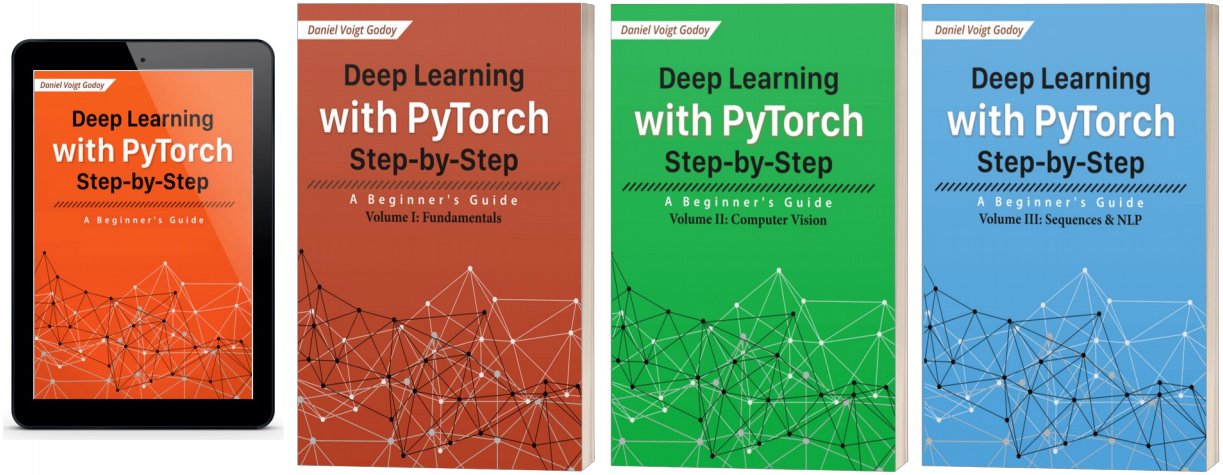
Daniel Voigt Godoy

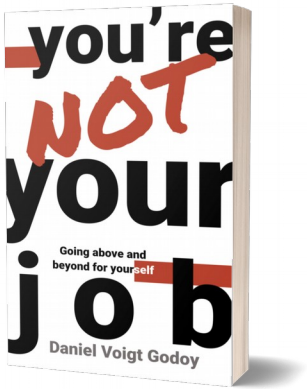
Data Science Retreat

July 2024

About Me

- Programmer
- Data Scientist
- Teacher
- Author

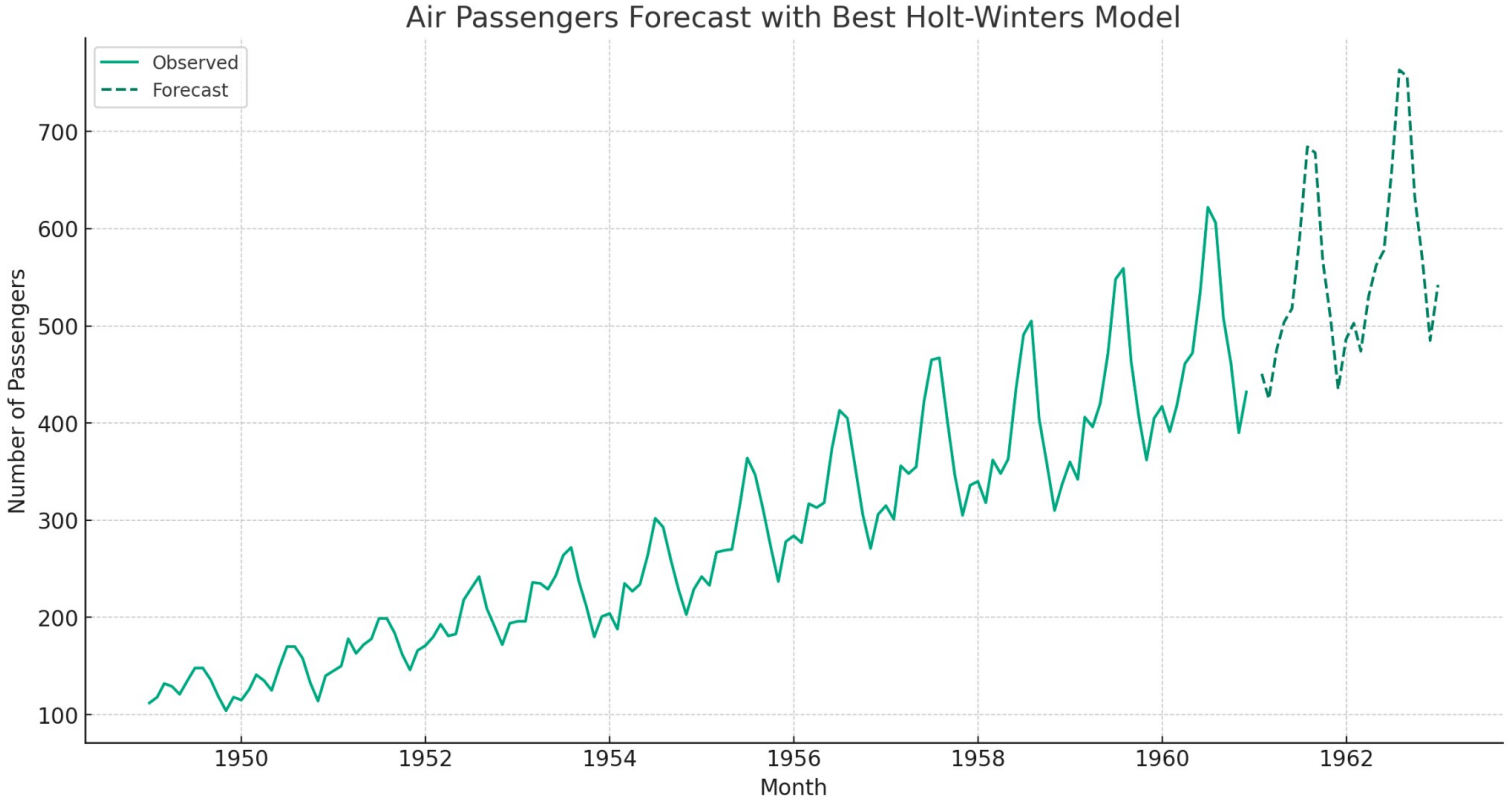




Agenda

- Tasks in Time Series Analysis
- Why working with time series is different?
- Properties of time series data: trend, seasonality and noise
- Autocorrelation and Partial Autocorrelation
- Using SkTime and Darts
- Moving averages and exponential smoothing
- Decomposition
- Baseline techniques
  - Next day prediction, Moving Averages, Exponential moving averages
- Classical techniques
  - ARIMA, Holt-Winters (Exponential smoothing)
- Using ML for time series analysis

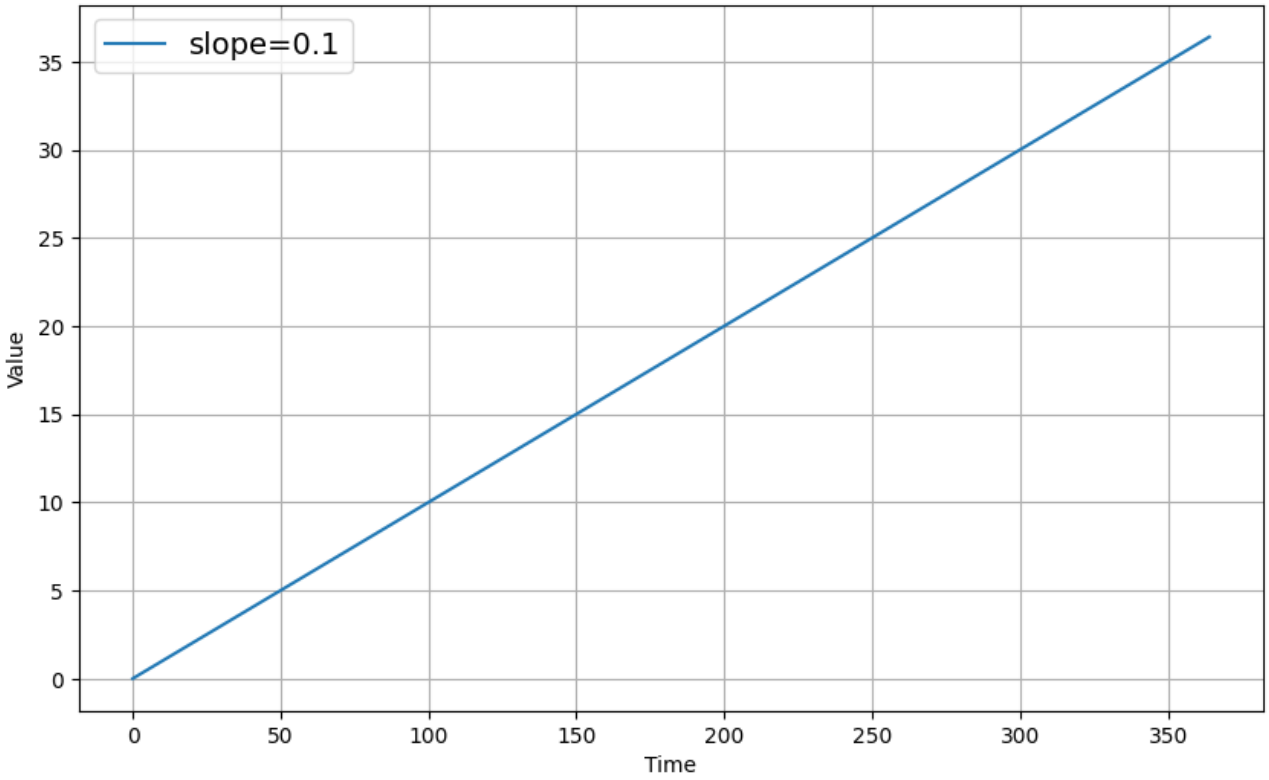
What is a time series?



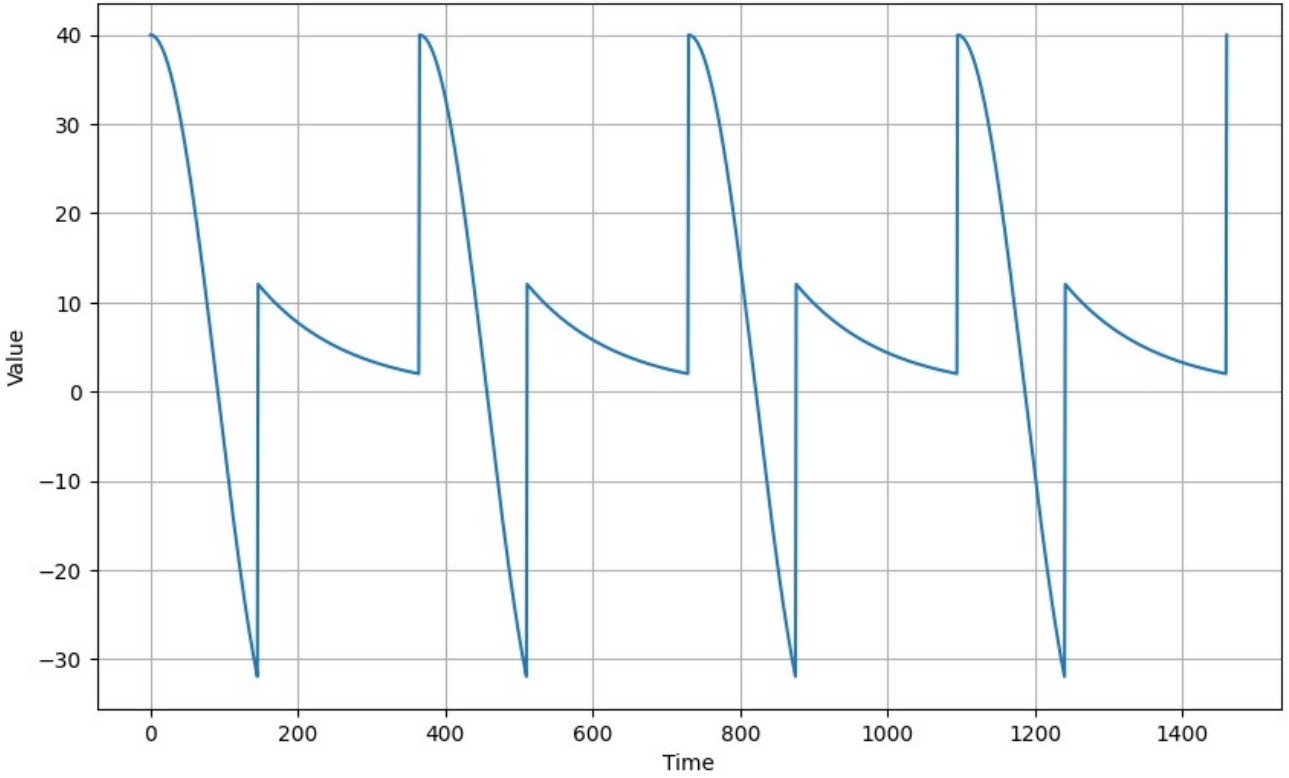
Knowing the jargon

- Trend
- Seasonality
- Residual aka Noise
- Stationarity
- Autoregressive
- Autocorrelation and Partial Autocorrelation
- Differencing
- Backtesting
- Exogenous variable
- Look-ahead problem
- Multivariate vs univariate
- Recursive forecasting
- Exponential moving average
- Exponential smoothing
- ARIMA

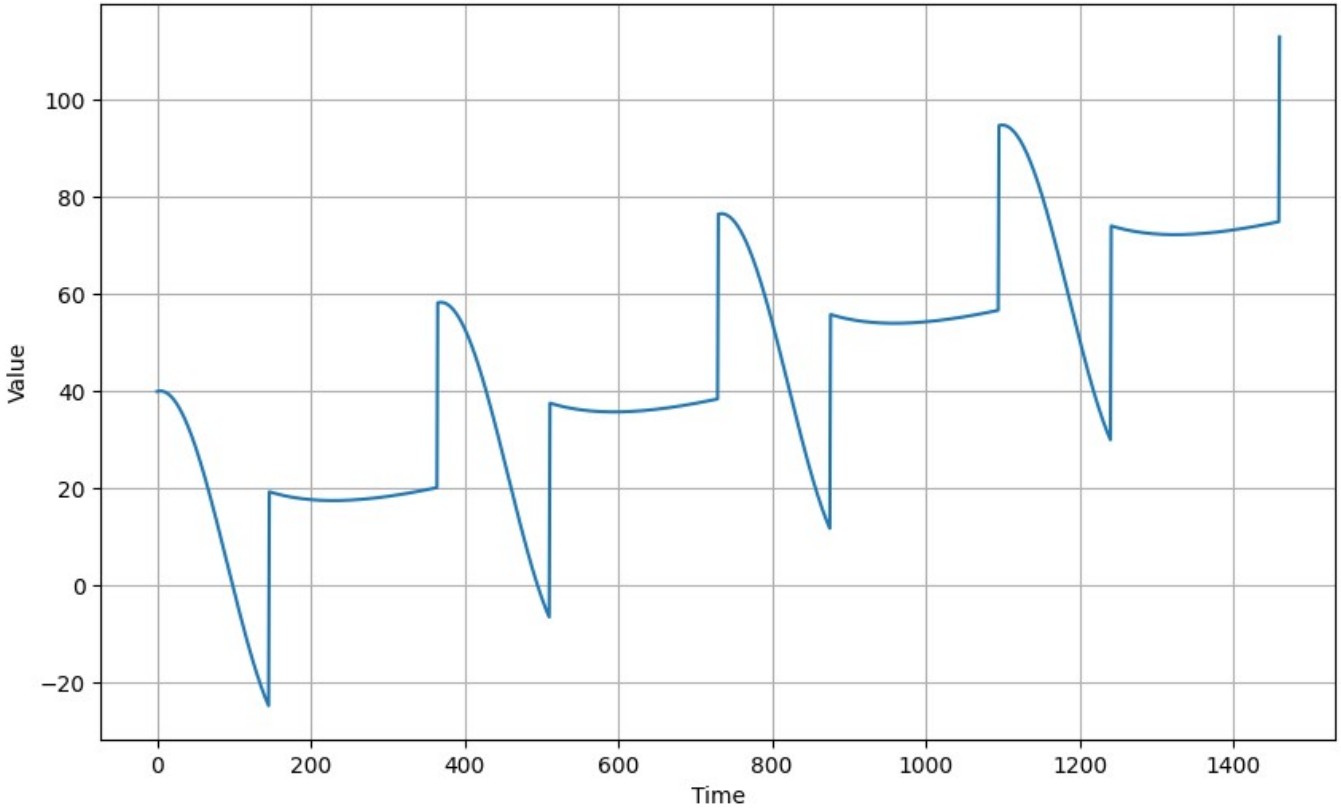
Trend


slope=0.1
35
30
25
20
15
10
5
0
Value
0
50
100
150
200
250
300
350
Time

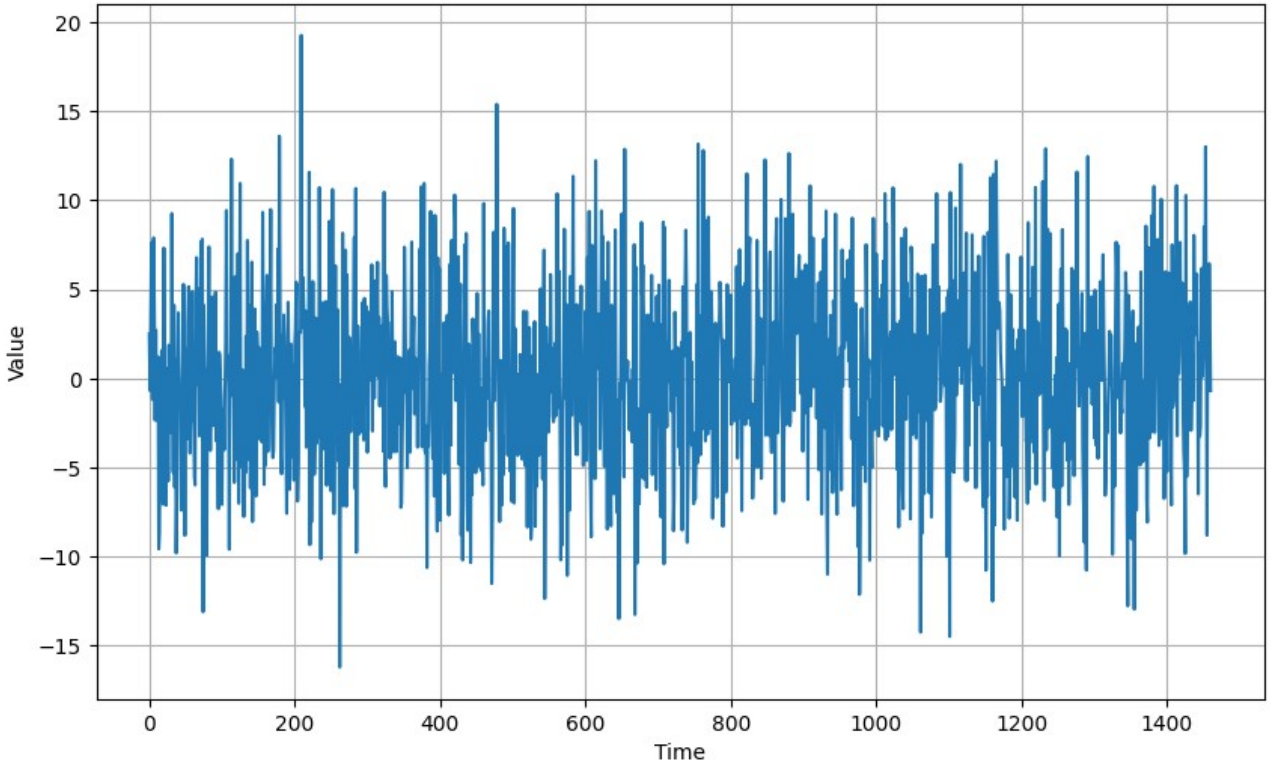
Seasonality



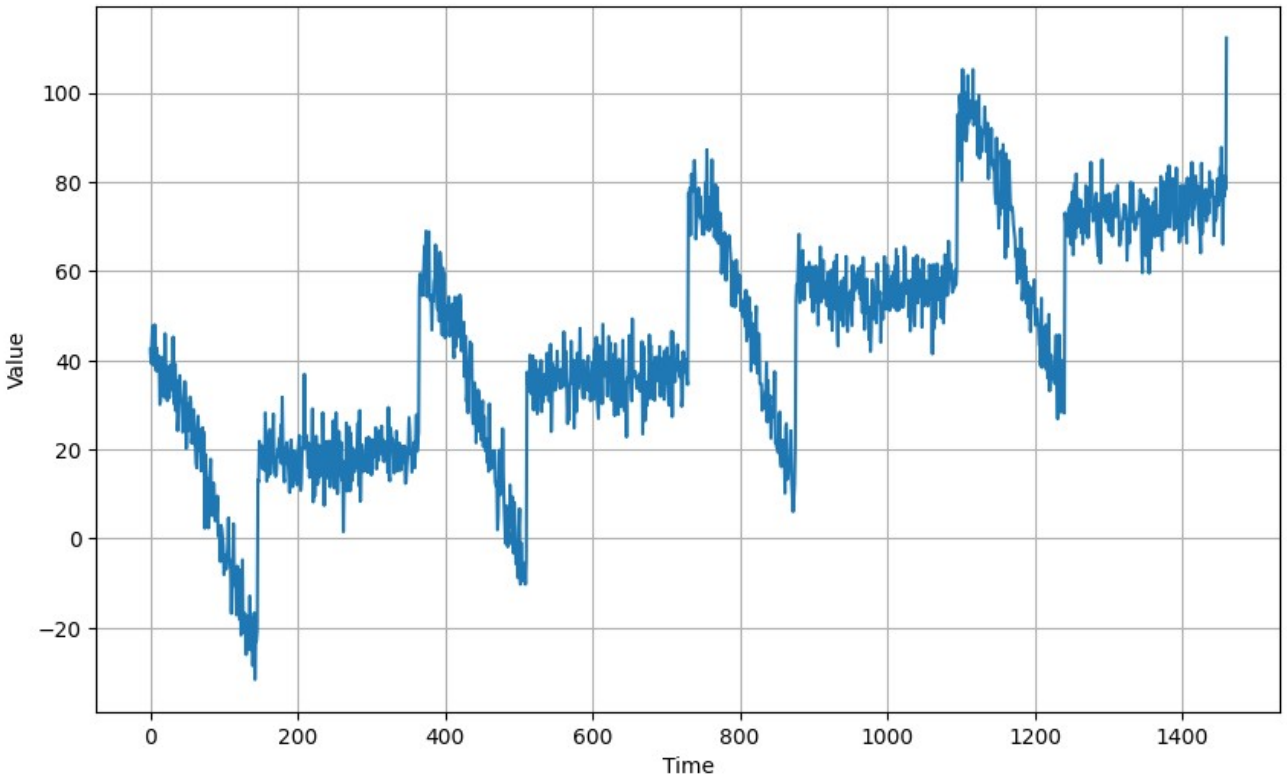
Trend and Seasonality



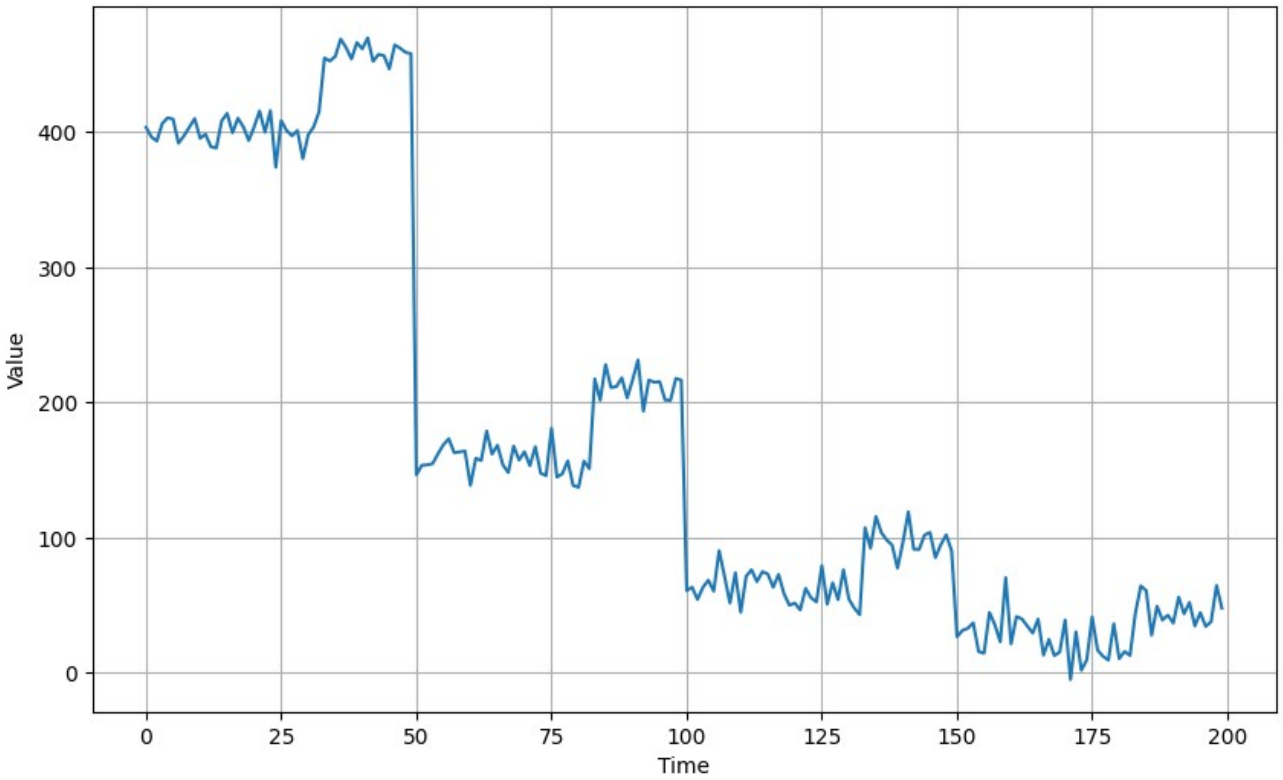
Noise



Trend + Seasonality + Noise

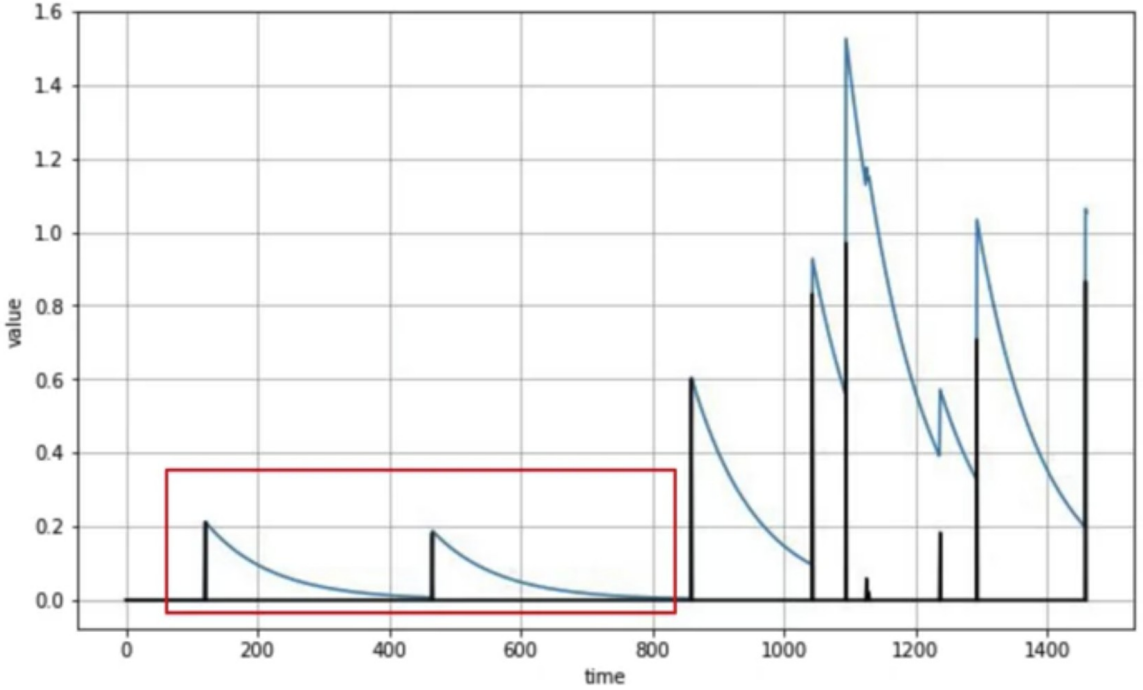


Autocorrelation

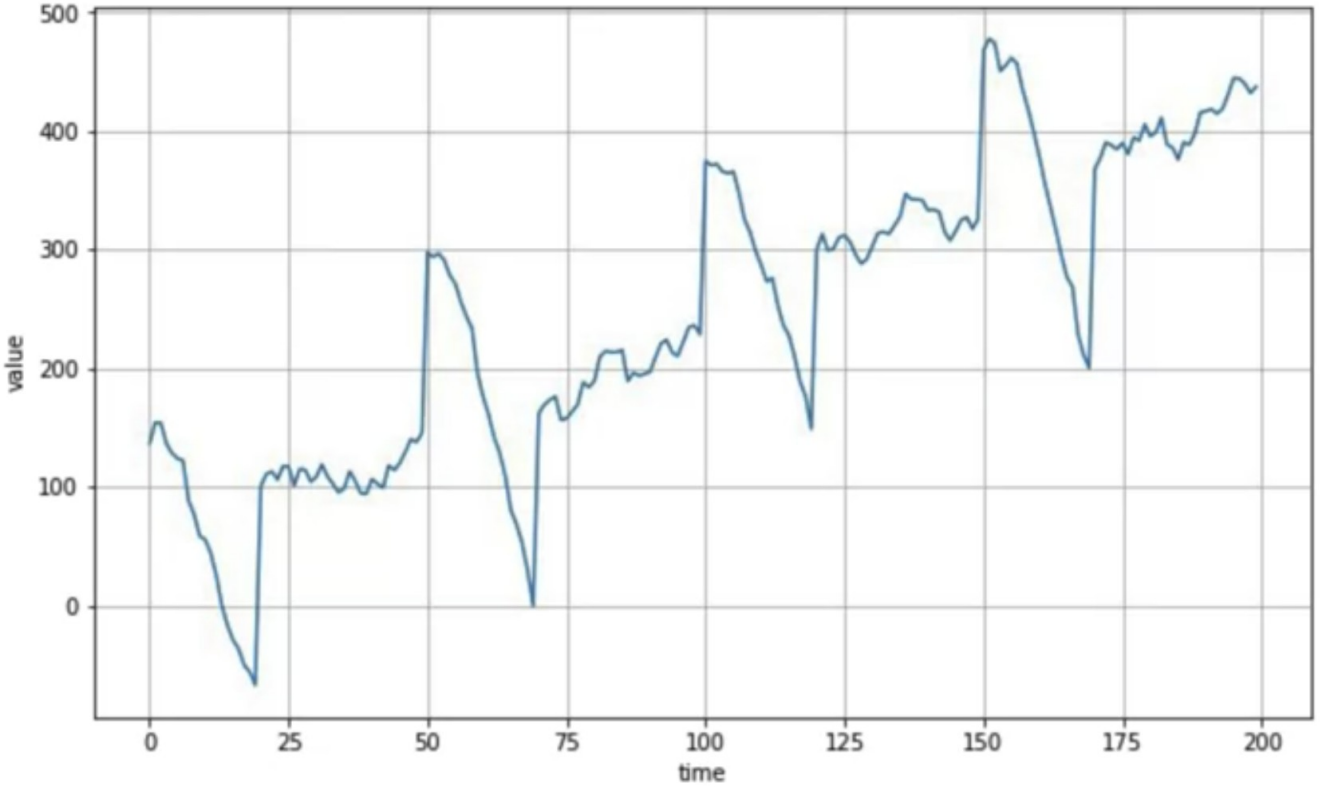


Memory and “innovations” in time series

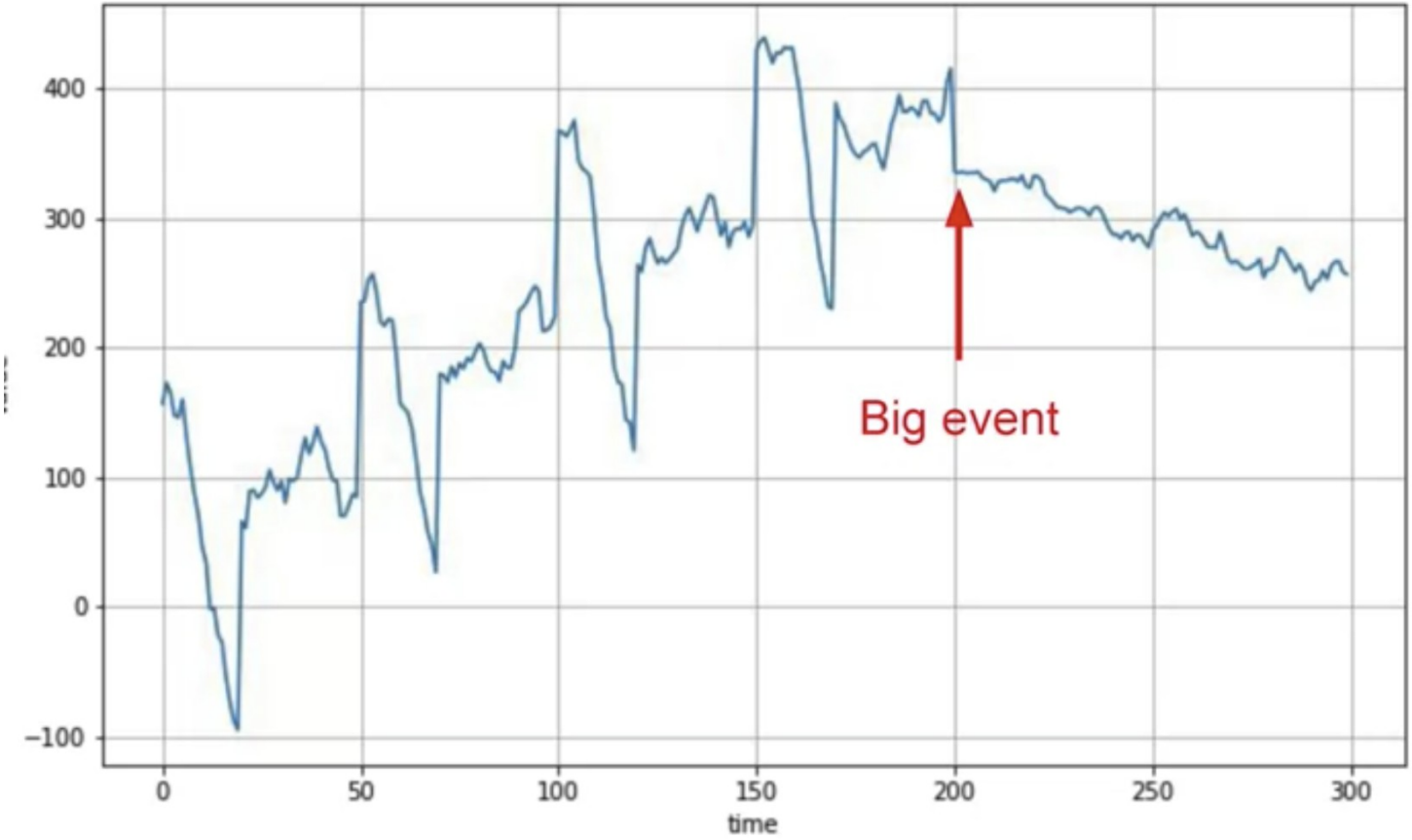
v(t) = 0.99 × v(t–1) + occasional spike



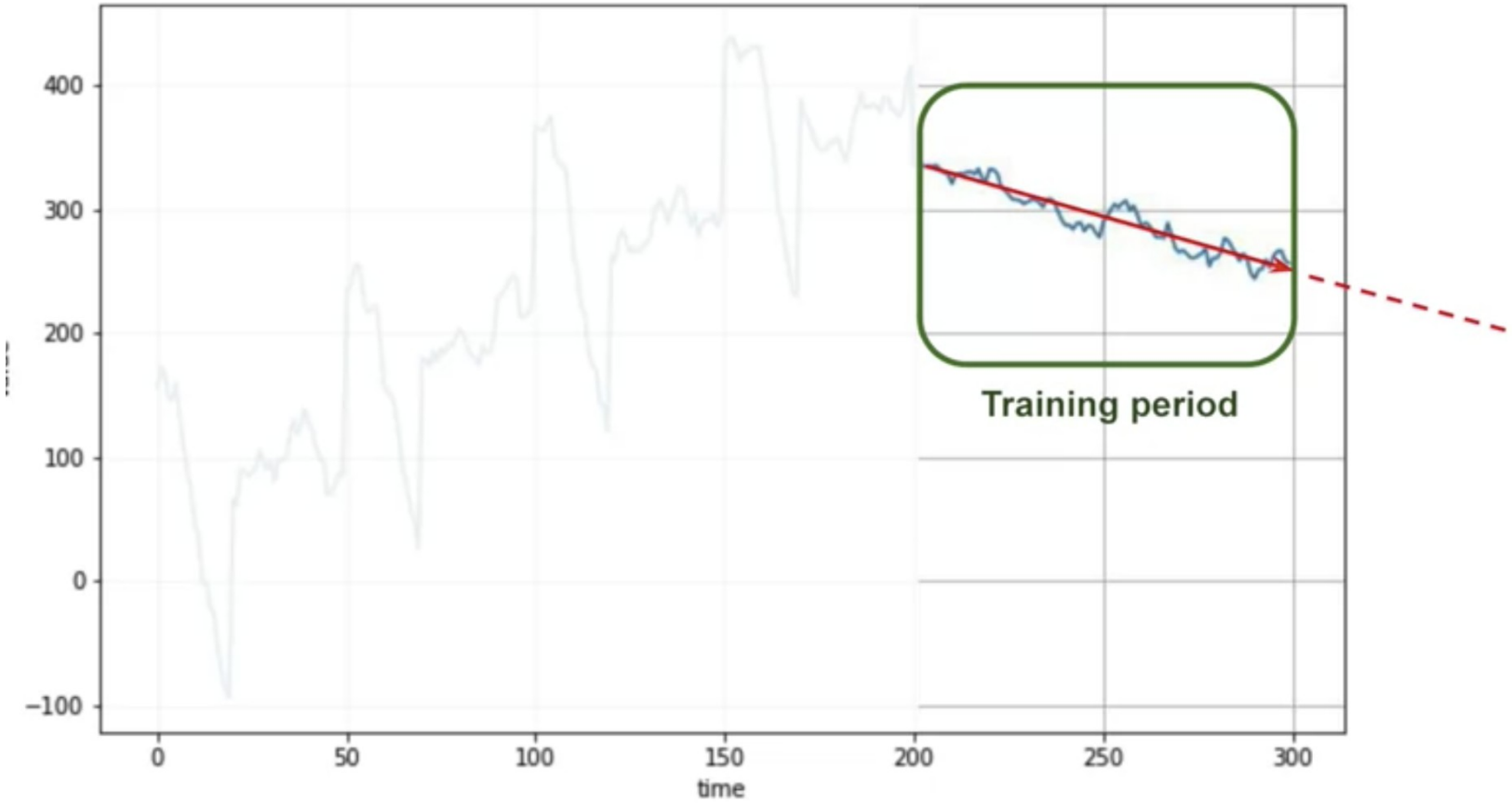
Trend + Seasonality + Autocorrelation + Noise



Non-Stationary Time Series



Non-Stationary Time Series



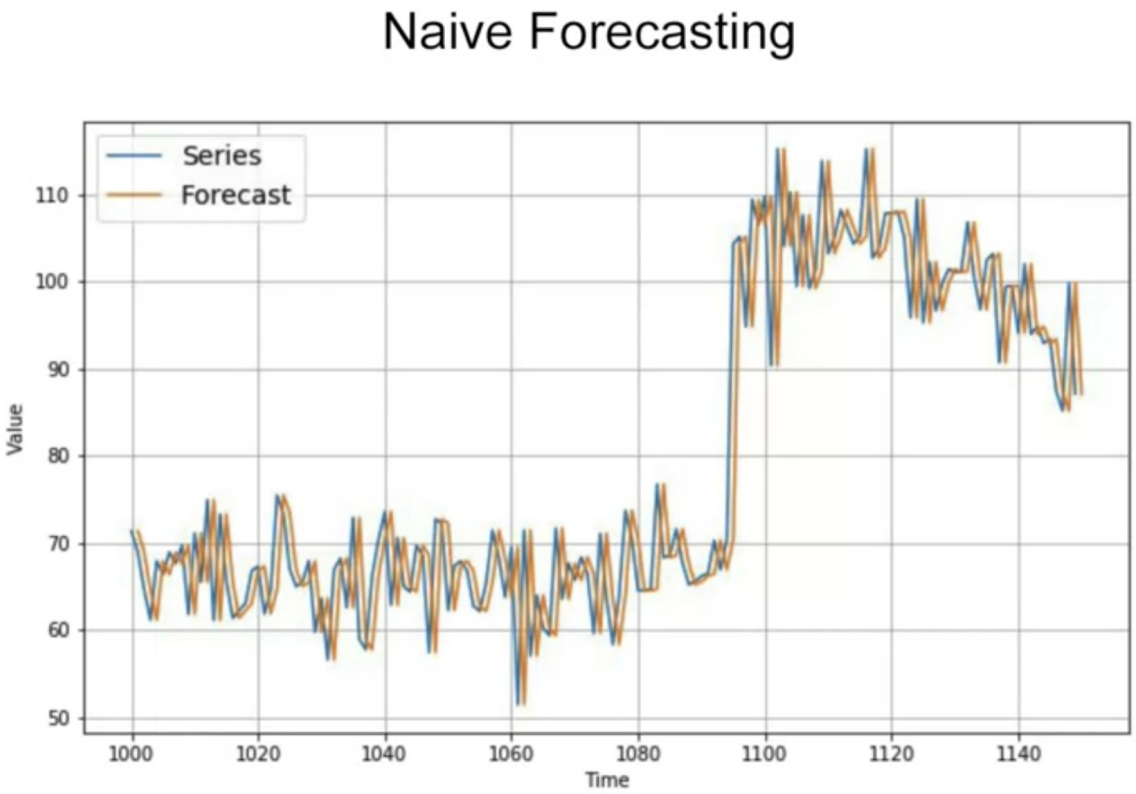
Explore time series properties with synthetic data

- N0 - Components

Tasks in Time Series Analysis

- Forecasting (‘predicting the future’ - our focus today)
- Anomaly Detection (identifying weird events)
- Classification (detecting patterns, e.g. cardiac arrest, speech recognition)

Baselines for Forecasting



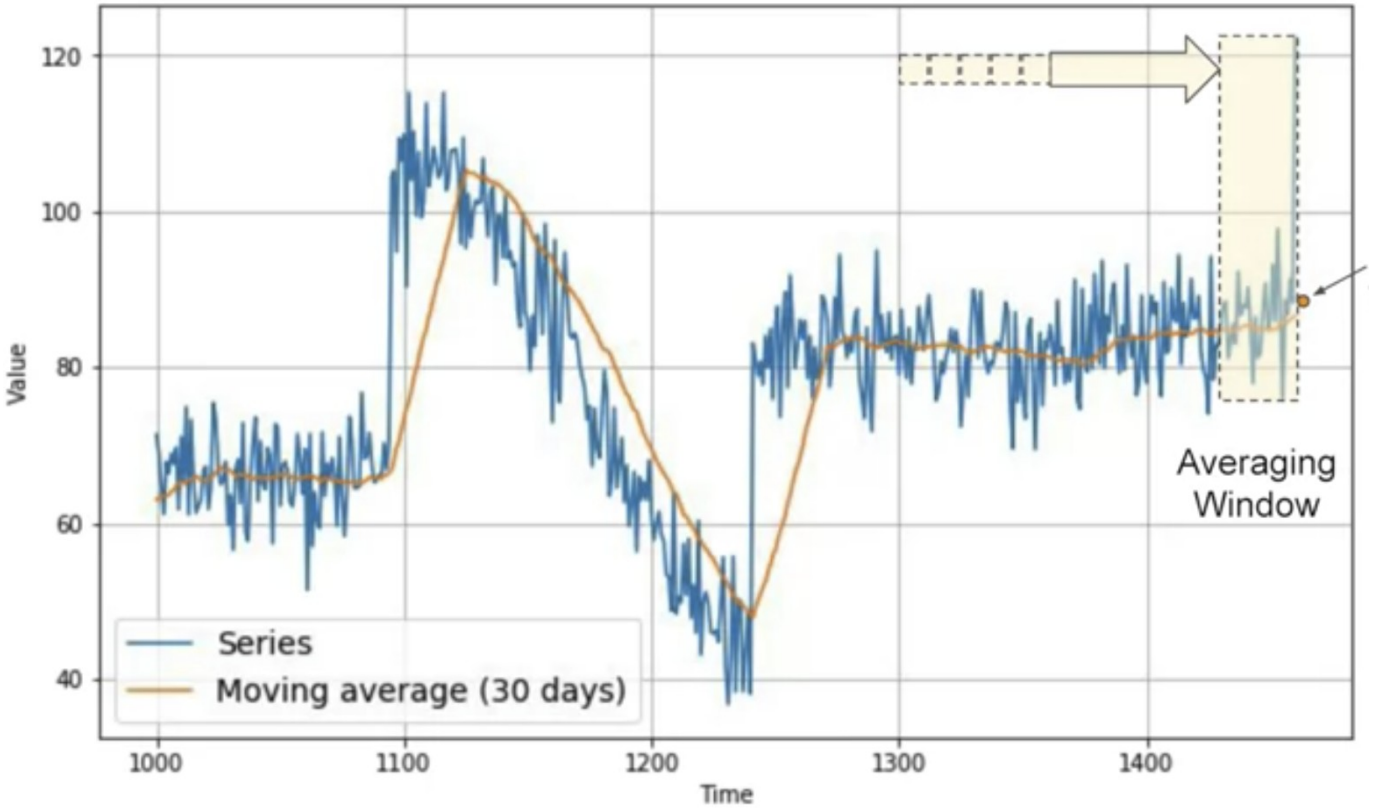
Moving averages

Pandas was invented specifically for time series analysis

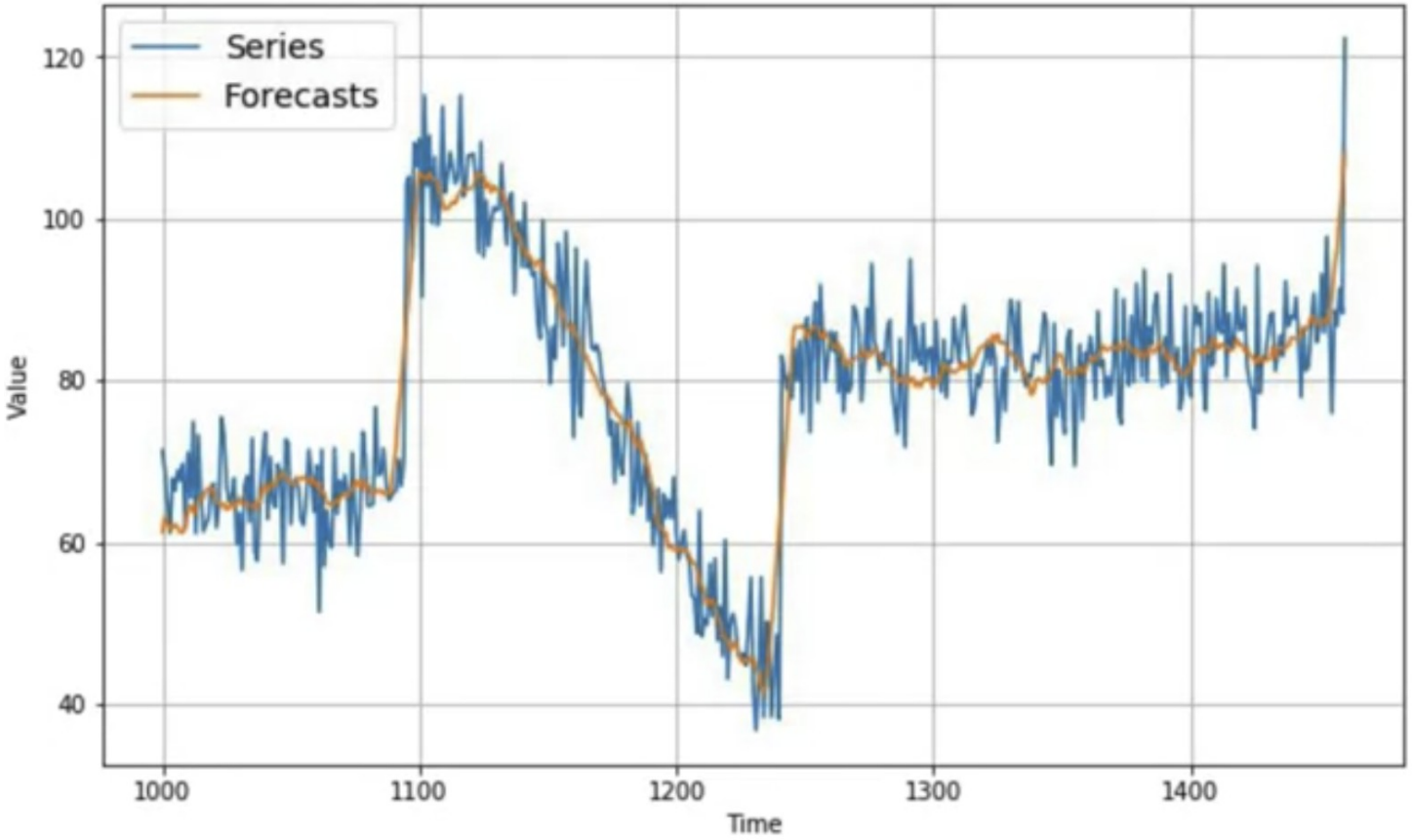
$$\bar{a}_{\mathrm{SM}} = \frac{x_n + x_{n-1} + \ldots + x_{M-(n-1)}}{M}$$

$$\bar{a}_{\mathrm{SM}} = \frac{1}{M} \sum_{i=0}^{n-1} x_{M-i}$$

Moving Average



Smoothing Both Past and Present Values



Metrics to Measure Accuracy

MAE - targets median of the true value distribution

RMSE - targets mean of the true value distribution, sensitive to outliers

MAPE - percentage of error, popular but weird when actual = 0 or forecast > actual

SMAPE - same error when forecast < actual and vice-versa [read]

Notebook 1

Notebook 2

Metrics

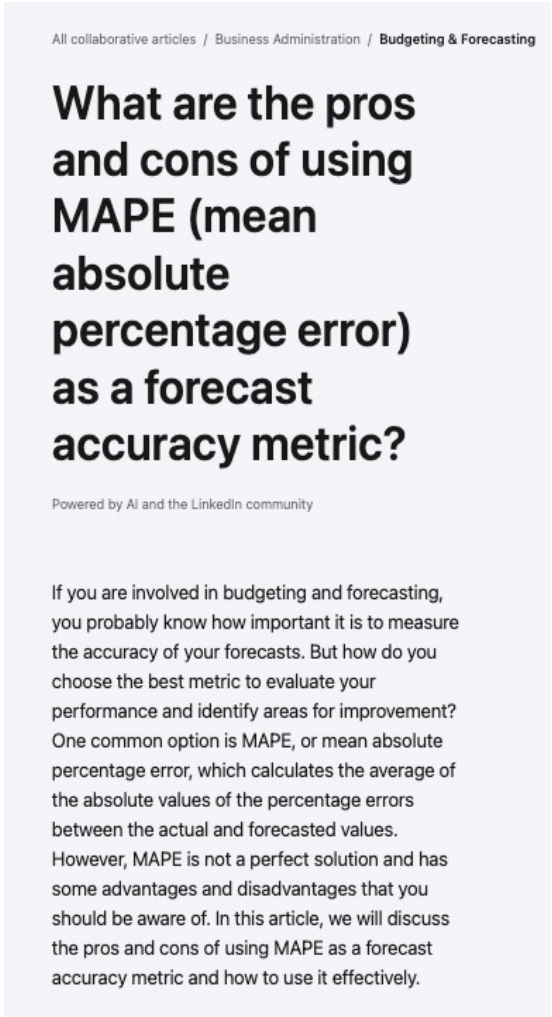
```
errors = forecasts - actual

mse = np.square(errors).mean()

rmse = np.sqrt(mse)

mae = np.abs(errors).mean()

mape = np.abs(errors / x_valid).mean()
```



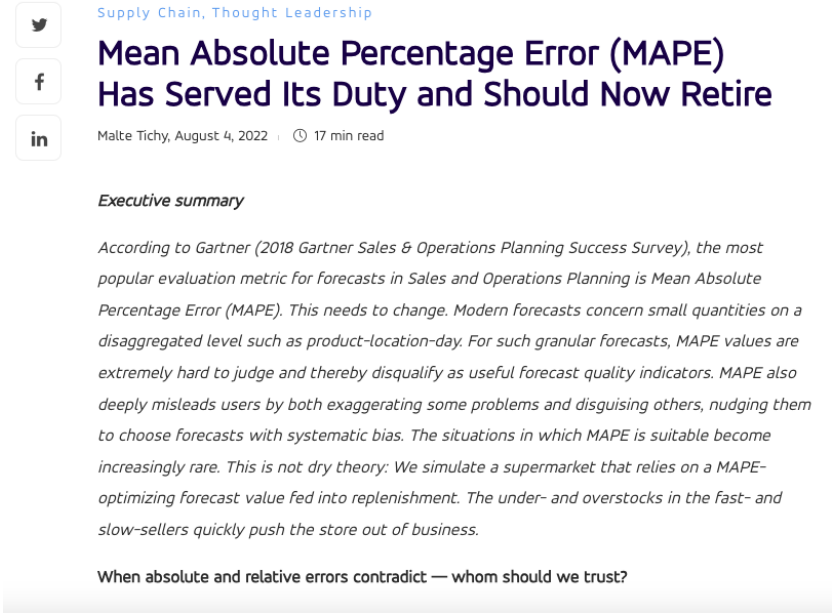
All collaborative articles / Business Administration / **Budgeting & Forecasting**

What are the pros and cons of using MAPE (mean absolute percentage error) as a forecast accuracy metric?

Powered by AI and the LinkedIn community

If you are involved in budgeting and forecasting, you probably know how important it is to measure the accuracy of your forecasts. But how do you choose the best metric to evaluate your performance and identify areas for improvement? One common option is MAPE, or mean absolute percentage error, which calculates the average of the absolute values of the percentage errors between the actual and forecasted values. However, MAPE is not a perfect solution and has some advantages and disadvantages that you should be aware of. In this article, we will discuss the pros and cons of using MAPE as a forecast accuracy metric and how to use it effectively.

[What are the pros and cons of using MAPE (mean absolute percentage error) as a forecast accuracy metric?]



Supply Chain, Thought Leadership

Mean Absolute Percentage Error (MAPE) Has Served Its Duty and Should Now Retire

Malte Tichy, August 4, 2022 | 17 min read

Executive summary

According to Gartner (2018 Gartner Sales & Operations Planning Success Survey), the most popular evaluation metric for forecasts in Sales and Operations Planning is Mean Absolute Percentage Error (MAPE). This needs to change. Modern forecasts concern small quantities on a disaggregated level such as product-location-day. For such granular forecasts, MAPE values are extremely hard to judge and thereby disqualify as useful forecast quality indicators. MAPE also deeply misleads users by both exaggerating some problems and disguising others, nudging them to choose forecasts with systematic bias. The situations in which MAPE is suitable become increasingly rare. This is not dry theory: We simulate a supermarket that relies on a MAPE-optimizing forecast value fed into replenishment. The under- and overstocks in the fast- and slow-sellers quickly push the store out of business.

When absolute and relative errors contradict — whom should we trust?

[MAPE has served its duty]

The mean absolute percentage error (MAPE) is one of the most popular measures of the forecast accuracy. It is recommended in most textbooks (e.g., Bowerman et al., 2004, Hanke and Reitsch, 1995), and was used as the primary measure in the M-competition (Makridakis et al., 1982). MAPE is the average of absolute percentage errors (APE). Let A_t and F_t denote the actual and forecast values at data point t, respectively. Then, MAPE is defined as:

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_t - F_t}{A_t}\right|, \tag{1.1}$$

where N is the number of data points. To be more rigorous, Eq. (1.1) should be multiplied by 100, but this is omitted in this paper for ease of presentation without loss of generality. MAPE is scale-independent and easy to interpret, which makes it popular with industry practitioners (Byrne, 2012).

However, MAPE has a significant disadvantage: it produces infinite or undefined values when the actual values are zero or close to zero, which is a common occurrence in some fields. If the actual values are very small (usually less than one), MAPE yields extremely large percentage errors (outliers), while zero actual values result in infinite MAPEs. In practice, data with numerous zero values are observed in various areas, such as retailing, biology, and finance, among others. For the area of retailing, Fig. 1 (Makridakis, Wheelwright, & Hyndman, 1998) illustrates typical intermittent sales data. Many zero sales occur during the time periods considered, and this leads to infinite or undefined MAPEs.

https://www.sciencedirect.com/science/article/pii/S0169207016000121

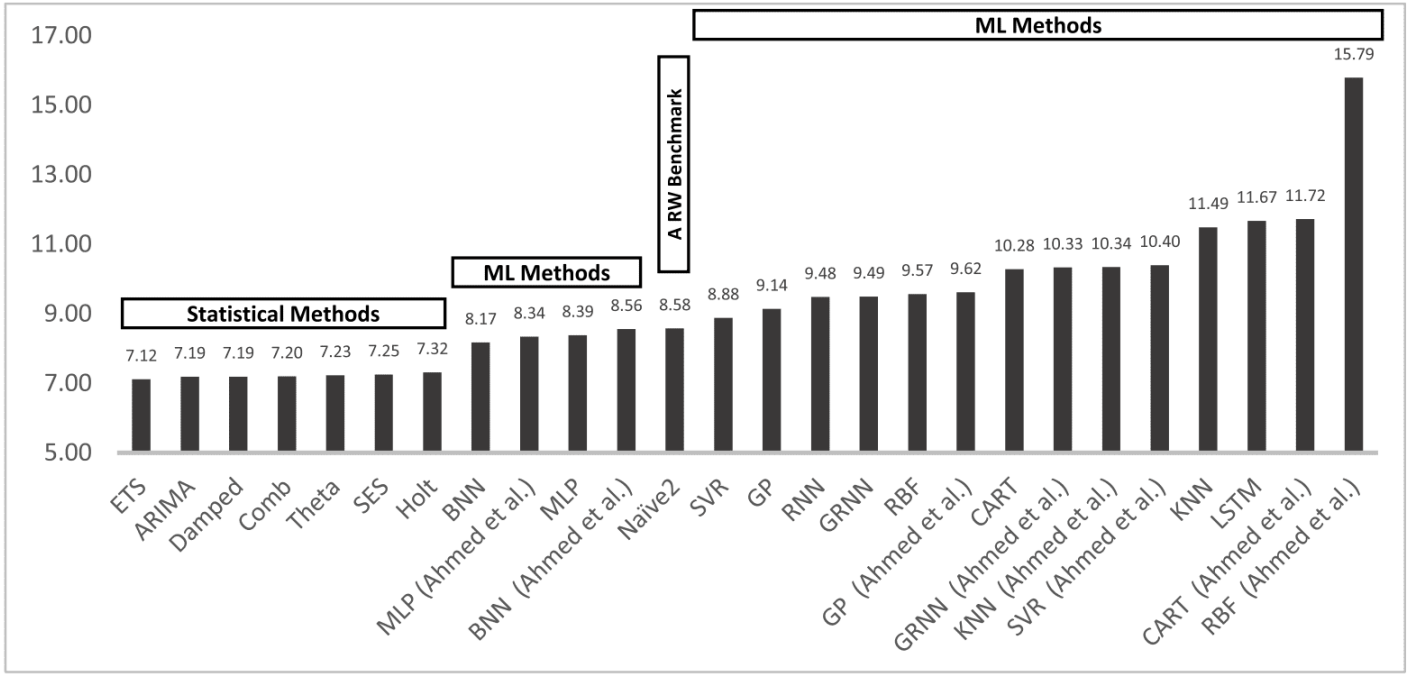
Exercise Part #0

Why use moving averages?

- Remove outliers
- Smoothen out short-term fluctuations
- Highlight trends or cycles
- N1 – Smoothing
- N2 – Theta Model

Exercise Part #1

Statistical and ML Forecasting Methods



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889

The Makridakis Competitions

Spyros Makridakis

Time series is a field of data science that has ran competitions to benchmark forecast models, **20 years before Kaggle**

https://en.wikipedia.org/wiki/Makridakis_Competitions

The Makridakis Competitions

No.	Informal name for competition	Year of publication of results	Number of time series used	Number of methods tested	Other features
1	M Competition or M-Competition[1][5]	1982	1001 (used a subsample of 111 for the methods where it was too difficult to run all 1001)	15 (plus 9 variations)	Not real-time
2	M-2 Competition or M2-Competition[1][6]	1993	29 (23 from collaborating companies, 6 from macroeconomic indicators)	16 (including 5 human forecasters and 11 automatic trend-based methods) plus 2 combined forecasts and 1 overall average	Real-time, many collaborating organizations, competition announced in advance
3	M-3 Competition or M3-Competition[1]	2000	3003	24	
4	M-4 Competition or M4 Competition	2020[7]	100,000	All major ML and statistical methods have been tested	First winner Slawek Smyl, Uber Technologies
5	M-5 Competition or M5 Competition	Initial results 2021, Final 2022	Around 42,000 hierarchical timeseries provided by Walmart	All major forecasting methods, including Machine and Deep Learning, and Statistical ones will be tested	First winner Accuracy Challenge: YeonJun In. First winners uncertainty Challenge: Russ Wolfinger and David Lander
6	M-6 Competition or M6 Competition	Initial results 2022, Final 2024	Real time financial forecasting competition consisting of 50 S&P500 US stocks and of 50 international ETFs	All major forecasting methods, including Machine and Deep Learning, and Statistical ones will be tested	

https://en.wikipedia.org/wiki/Makridakis_Competitions

M3 Competition

The time series included yearly, quarterly, monthly, daily, and other time series. In order to ensure that enough data was available to develop an accurate forecasting model, minimum thresholds were set for the number of observations: 14 for yearly series, 16 for quarterly series, 48 for monthly series, and 60 for other series.

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	146	102	83	58	245	11	645
Quarterly	204	83	336	76	57	0	756
Monthly	474	334	312	145	111	52	1428
Other	4	0	0	29	0	141	174
Total	**828**	**519**	**731**	**308**	**413**	**204**	**3003**

The five measures used to evaluate the accuracy of different forecasts were: symmetric mean absolute percentage error (also known as symmetric MAPE), average ranking, median symmetric absolute percentage error (also known as median symmetric APE), percentage better, and median RAE.

M4 Competition

To get precise and compelling answers, the M4 Competition utilized 100,000 real-life series, and incorporates all major forecasting methods, including those based on Artificial Intelligence (ML), as well as traditional statistical ones. In order to ensure that enough data are available to develop an accurate forecasting model, minimum thresholds were set for the number of observations: 13 for yearly, 16 for quarterly, 42 for monthly, 80 for weekly, 93 for daily and 700 for hourly series.

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	6538	3716	3903	6519	1088	1236	23000
Quarterly	6020	4637	5315	5305	1858	865	24000
Monthly	10975	10017	10016	10987	5728	277	48000
Weekly	112	6	41	164	24	12	359
Daily	1476	422	127	1559	10	633	4227
Hourly	0	0	0	0	0	414	414
Total	**25121**	**18798**	**19402**	**24534**	**8708**	**3437**	**100000**

M4 Competition (cont.)

Overall Weighted Average (OWA): This metric is calculated by obtaining the average of the symmetric mean absolute percentage error (sMAPE) and the mean absolute scaled error (MASE) for all the time series of the model and also calculating it for the Naive2 predictions. Both sMAPE and MASE are scale independent.

$$sMAPE = \frac{200}{H} \sum_{i=1}^{H} \frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}| + |\hat{y}_{T+i}|} \qquad MASE = \frac{1}{H} \sum_{i=1}^{H} \frac{|y_{T+i} - \hat{y}_{T+i}|}{\frac{1}{T+H-m} \sum_{j=m+q}^{T+H} |y_j - y_{j-m}|}$$

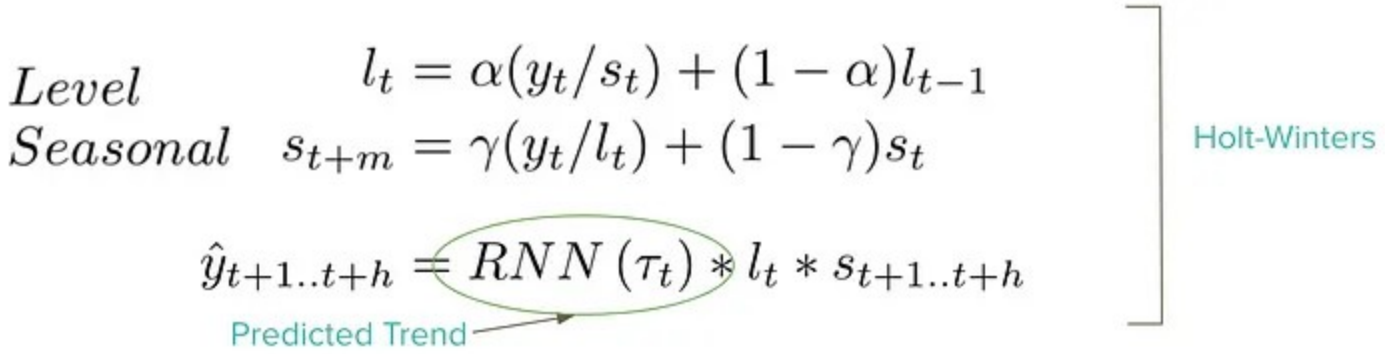
$$OWA = \frac{1}{2} \left[\frac{sMAPE}{sMAPE_{Naive2}} + \frac{MASE}{MASE_{Naive2}} \right]$$

M4 Competition (cont.)

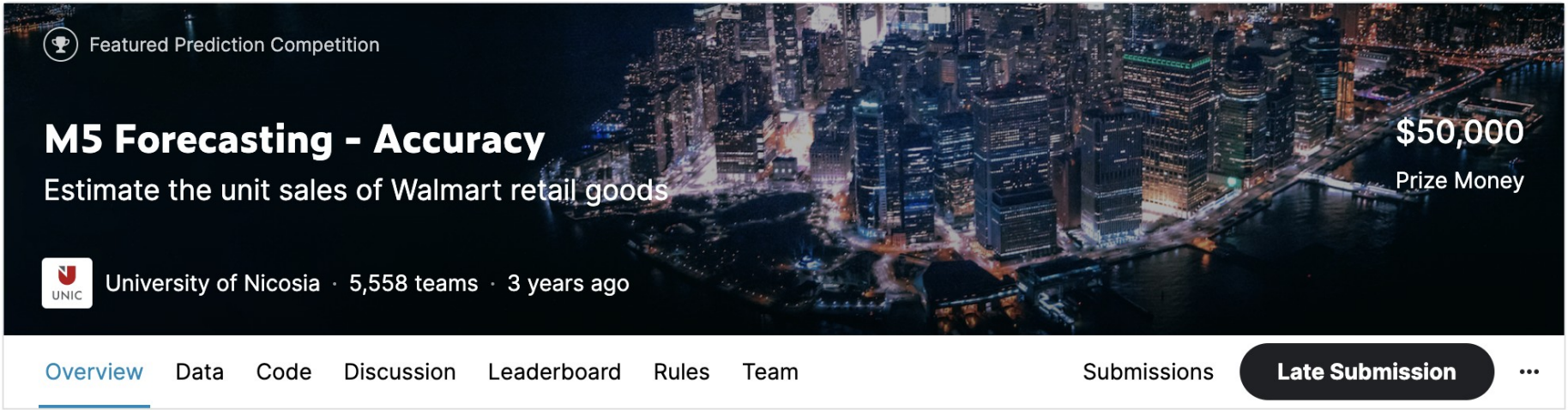
- The combination of methods was the king of the M4. Out of the 17 most accurate methods, 12 were "combinations" of mostly statistical approaches.
- The biggest surprise, however, was a "hybrid" approach utilizing both Statistical and ML features. This method, produced the most accurate forecasts as well as the most precise PIs and was submitted by Slawek Smyl, Data Scientist at Uber Technologies. According to sMAPE, it was close to 10% (a huge improvement) more accurate than the Combination (Comb) benchmark of the Competition (see below). It is noted that in the M3 Competition (Makridakis & Hibon, 2000) the best method was 4% more accurate than the same Combination.

M4 Competition (cont.)

- ES-RNN



The Makridakis 5 Competition (M5 - Accuracy)



https://www.kaggle.com/competitions/m5-forecasting-accuracy

M5 Competition

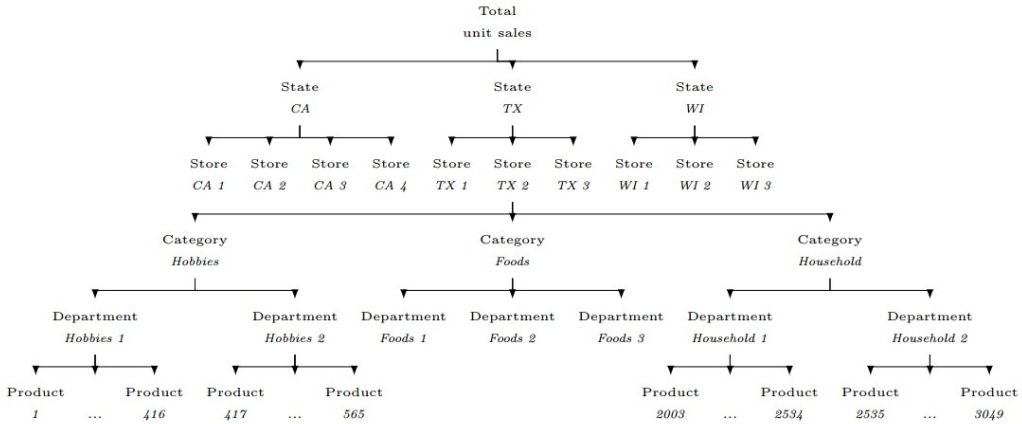
The data was provided by Walmart and consisted of around 42,000 hierarchical daily time series, starting at the level of SKUs and ending with the total demand of some large geographical area. In addition to the sales data, there was also information about prices, advertising/promotional activity and inventory levels as well as the day of the week the data refers to. This competition uses a Weighted Root Mean Squared Scaled Error (RMSSE).

This competition was of the "M" competitions to feature primarily machine learning methods at the top of its leaderboard. All of the top-performing were, "pure ML approaches and better than all statistical benchmarks and their combinations." The LightGBM model, as well as deep neural networks, featured prominently in top submissions.

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t - \hat{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})^2}},$$

M5 Competition

- The series of the dataset were grouped and highly correlated, thus enabling the utilization of multivariate and “cross-learning" methods.
- The dataset involved daily data which requires accounting for multiple seasonal patterns, special days, and holidays.
- The dataset included exogenous/ explanatory variables, such as product prices, promotions, and special events.



M5 Competition

Training Code of the Winning Solutions

Preprocessing Code of the Winning Solutions

M5 accuracy competition: Results, findings, and conclusions - ScienceDirect

M5 Forecasting- Accuracy. Forecasting is done using Xgboost… | by Jaswanth Badvelu | Towards Data Science

M6 Competition

https://mofc.unic.ac.cy/ten-hypotheses

The M6 Financial Forecasting Competition

The efficient market hypothesis (EMH) posits that share prices reflect all relevant information, which implies that consistent outperformance of the market is not feasible. The EMH is supported by empirical evidence, including the yearly "Active/Passive Barometer" Morningstar study which regularly finds that active, professional investment managers do not beat, on average, random stock selections. On the other hand, legendary investors like Warren Buffett, Peter Lynch and George Soros, among others, as well as celebrated firms including Blackstone, Bridgewater Associates, Renaissance Technologies, DE Shaw and many others have achieved phenomenal results over long periods of time, amassing returns impossible to justify by mere chance, and casting doubts about the validity of the EMH. It is the express purpose of the M6 competition to empirically investigate this paradox and to shed new light on the EMH by explaining the poor performance of active funds, as well as the exceptional performance of the likes of Warren Buffet, whose fund has achieved an average annual return of 20.0% since 1965, almost double that of S&P500's 10.2% annual gain during that period.

https://mofc.unic.ac.cy/the-m6-competition/

M6 Competition

The investment universe will consist of two classes of assets:

- 50 stocks from the Standard and Poor's (S&P) 500 index, and
- 50 international exchange-traded funds (ETFs).

The 50 stocks and 50 ETFs will be selected such that they are broadly representative of the market. We will announce the names of the 100 assets closer to the commencement of the M6 competition.

The forecasting performance for a particular submission point will be measured by the Ranked Probability Score (RPS).

Decomposing a Time Series

N3 – Decomposition

N4 – STL Decomposition

Exercise Part #2

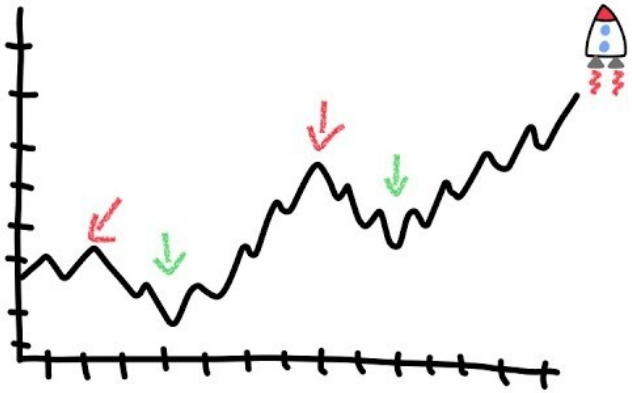
Why working with time series is different?

- We usually have as target the same feature that we use for training
- Missing values require extra care
- This creates problems with (usually unintentional) data leakage between training and testing
  - Lookahead problem - often appears in imputation and smoothing
- The “simple and traditional” methods (ARIMA) depend on assumptions that most real data doesn’t have
  - **We torture the data into having these features through differencing**
- Working with time series is one of most challenging tasks in data science

Why working with time series is so difficult?

- Missing and noisy data is very common
- Low interpretability of methods
- Difficult (for humans) to spot and describe patterns in squiggly lines
- Exogenous variables
  - How can we predict events like the COVID outbreak? (Black swans)
  - How many unmeasured events contribute to our output?
- Communicating to and keeping trust from stakeholders and teammates who often have conflicting or vested interests **is the biggest challenge**
- **It's easy to give bad predictions and it's easy to lose trust**

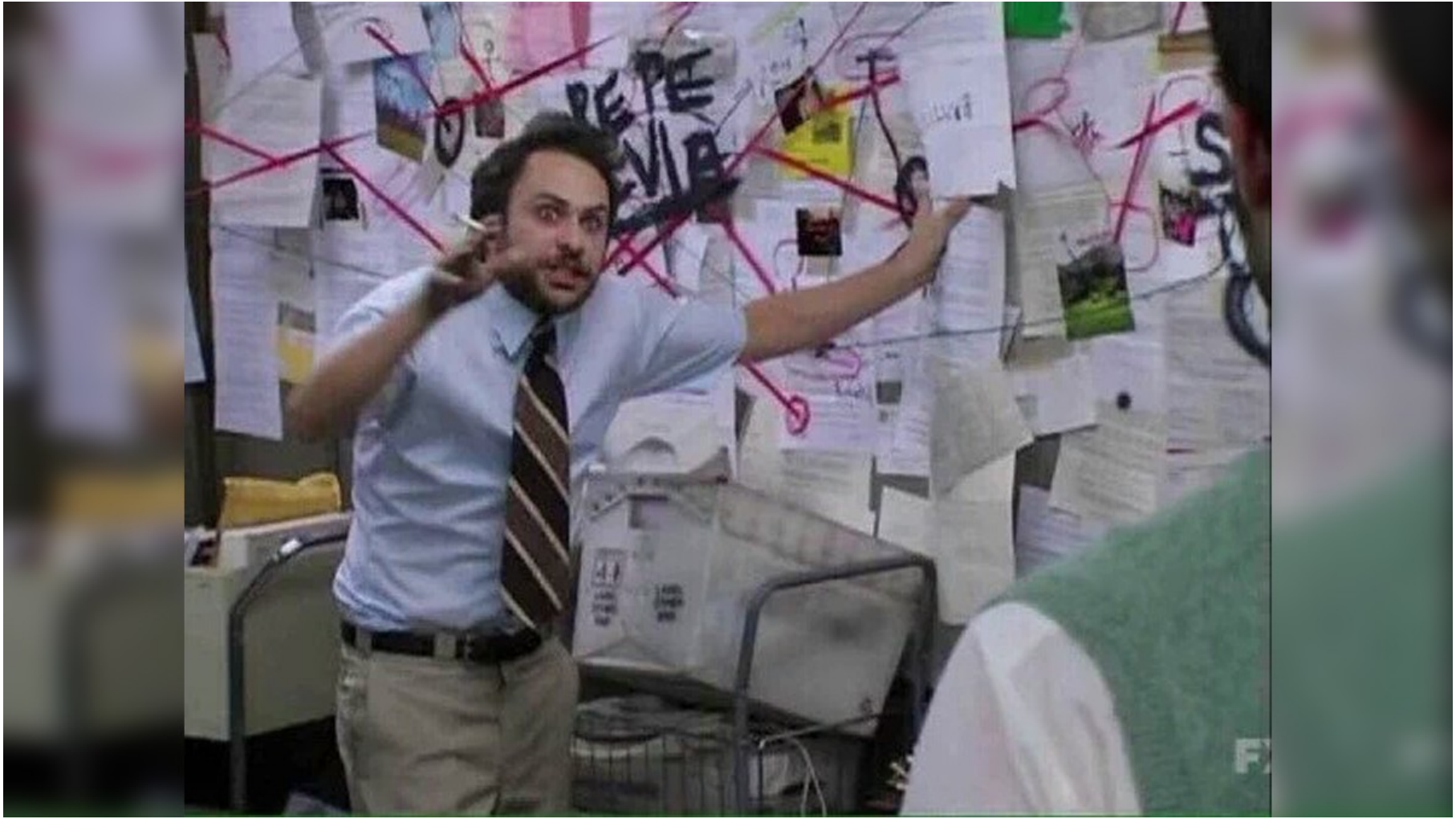
What is Apophenia?



Apophenia Explained | Pareidolia, Confirmation Bias, & Other Pattern Errors

Micro Analysis

Another point worth mentioning is that the structure of the final leg in a corrective pattern (the C wave) is always a 5 wave drop. Furthermore, these patterns are fractal, so a small 5 wave patterns will morph into a larger one, which turns into an even bigger one. So, if we analyze the structure of the initial drop, we can get clues on what is playing out.

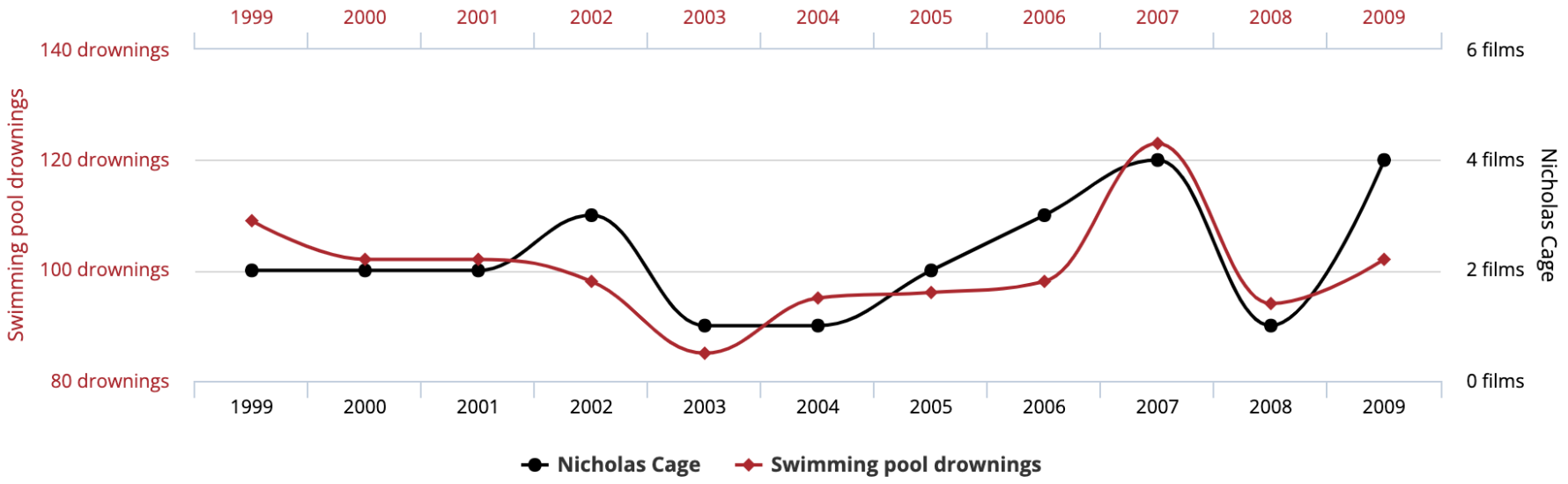

PEPE

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

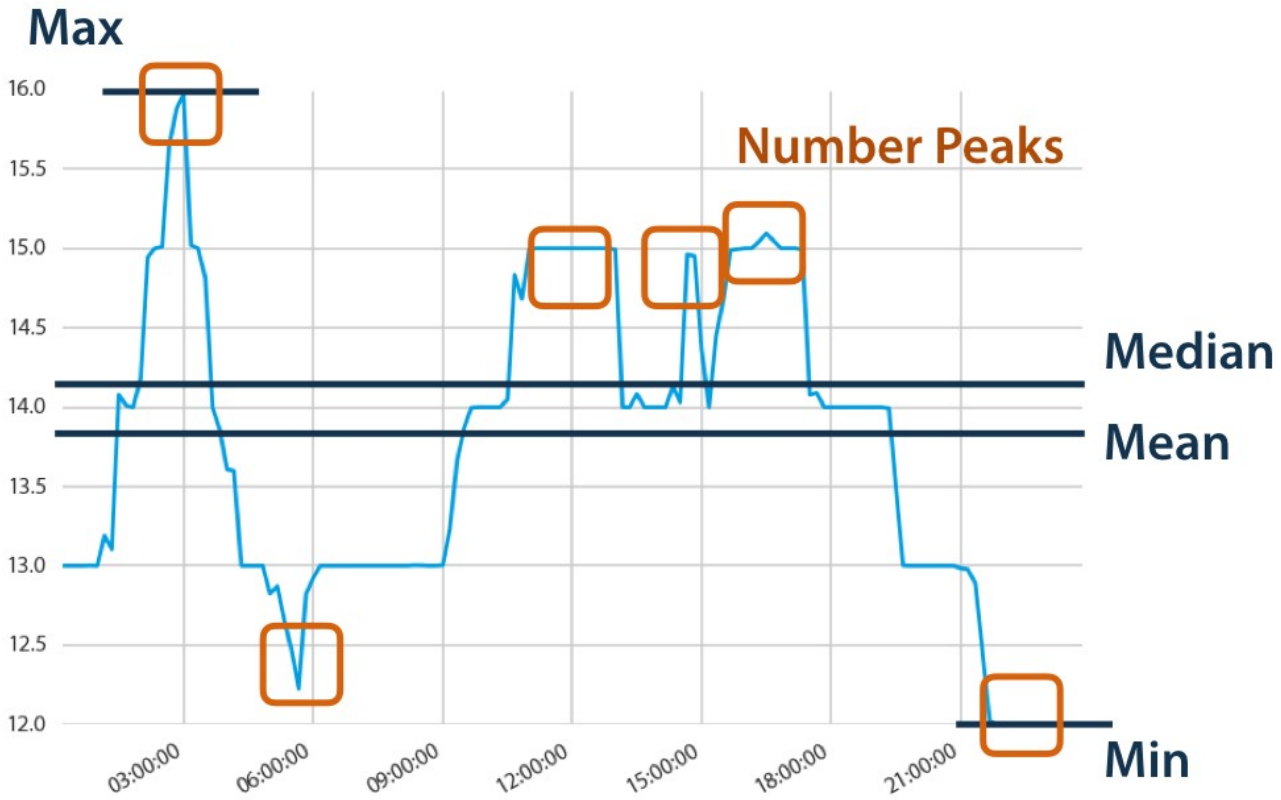
Correlation: 66.6% (r=0.666004)



tylervigen.com

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Time Series Feature Extraction



hctsa feature name	Description
Distribution	
DN_HistogramMode_5	Mode of z-scored distribution (5-bin histogram)
DN_HistogramMode_10	Mode of z-scored distribution (10-bin histogram)
Simple temporal statistics	
SB_BinaryStats_mean_longstretch1	Longest period of consecutive values above the mean
DN_OutlierInclude_p_001_mdrmd	Time intervals between successive extreme events above the mean
DN_OutlierInclude_n_001_mdrmd	Time intervals between successive extreme events below the mean
Linear autocorrelation	
CO_f1ecac	First $1/e$ crossing of autocorrelation function
CO_FirstMin_ac	First minimum of autocorrelation function
SP_Summaries_welch_rect_area_5_1	Total power in lowest fifth of frequencies in the Fourier power spectrum
SP_Summaries_welch_rect_centroid	Centroid of the Fourier power spectrum
FC_LocalSimple_mean3_stderr	Mean error from a rolling 3-sample mean forecasting
Nonlinear autocorrelation	
CO_trev_1_num	Time-reversibility statistic, $\langle(x_{t+1} - x_t)^3\rangle_t$
CO_HistogramAMI_even_2_5	Automutual information, $m = 2, \tau = 5$
IN_AutoMutualInfoStats_40_gaussian_fmmi	First minimum of the automutual information function
Successive differences	
MD_hrv_classic_pnn40	Proportion of successive differences exceeding 0.04σ [20]
SB_BinaryStats_diff_longstretch0	Longest period of successive incremental decreases
SB_MotifThree_quantile_hh	Shannon entropy of two successive letters in equiprobable 3-letter symbolization
FC_LocalSimple_mean1_tauresrat	Change in correlation length after iterative differencing
CO_Embed2_Dist_tau_d_expfit_meandiff	Exponential fit to successive distances in 2-d embedding space
Fluctuation Analysis	
SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	Proportion of slower timescale fluctuations that scale with DFA (50% sampling)
SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	Proportion of slower timescale fluctuations that scale with linearly rescaled range fits
Others	
SB_TransitionMatrix_3ac_sumdiagcov	Trace of covariance of transition matrix between symbols in 3-letter alphabet
PD_PeriodicityWang_th0_01	Periodicity measure of [31]

Table 1 The *catch22* feature set spans a diverse range of time-series characteristics representative of the diversity of interdisciplinary methods for time-series analysis. Features in *catch22* capture time-series properties of the distribution of values in the time series, linear and nonlinear temporal autocorrelation properties, scaling of fluctuations, and others.

https://arxiv.org/pdf/1901.10200.pdf

Consequences

Our predictions will always have a margin of error. *“**How much will they be wrong?”*** is something that every stakeholder will ask us and we will never be able to have a definite answer **other than *the methodology that we used to generate our forecast.***

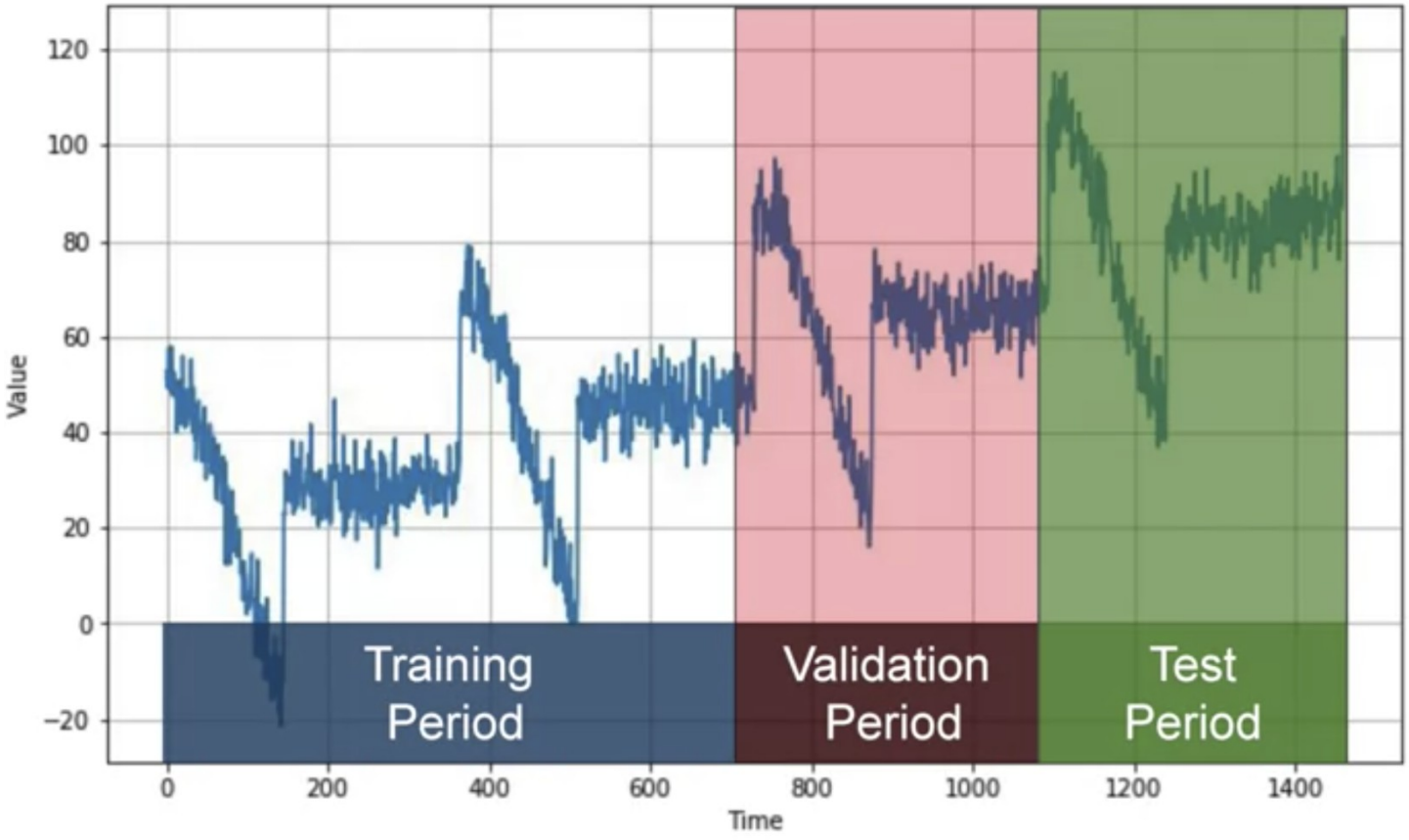
ML and Feature Engineering

N5 – Pitfalls

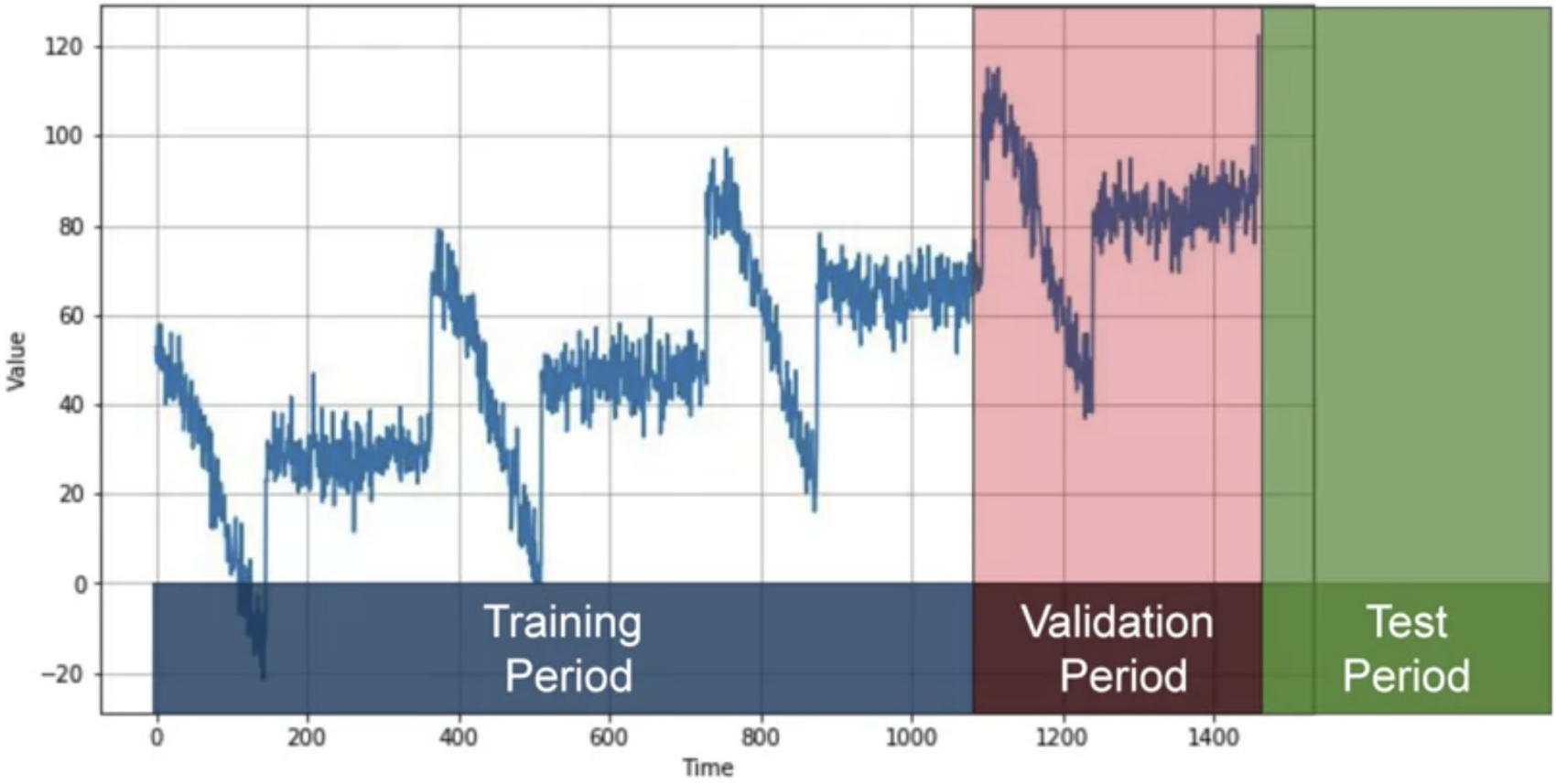
N6 – Feature Engineering

Exercise Part #3

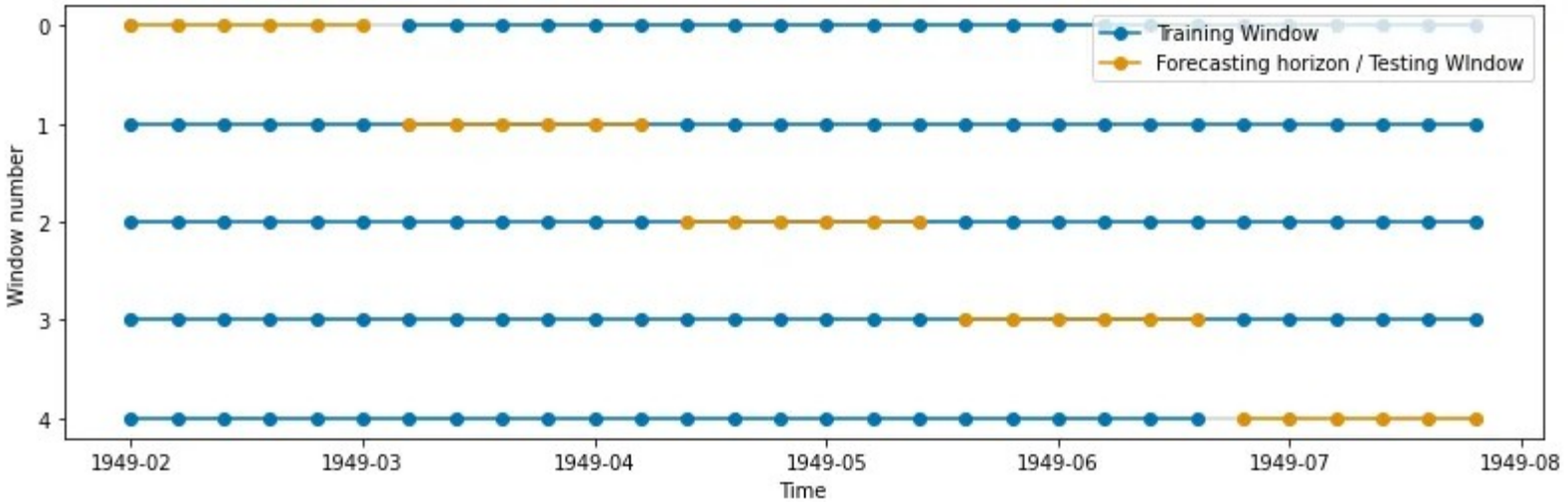
Fixed Partitioning



Fixed Partitioning

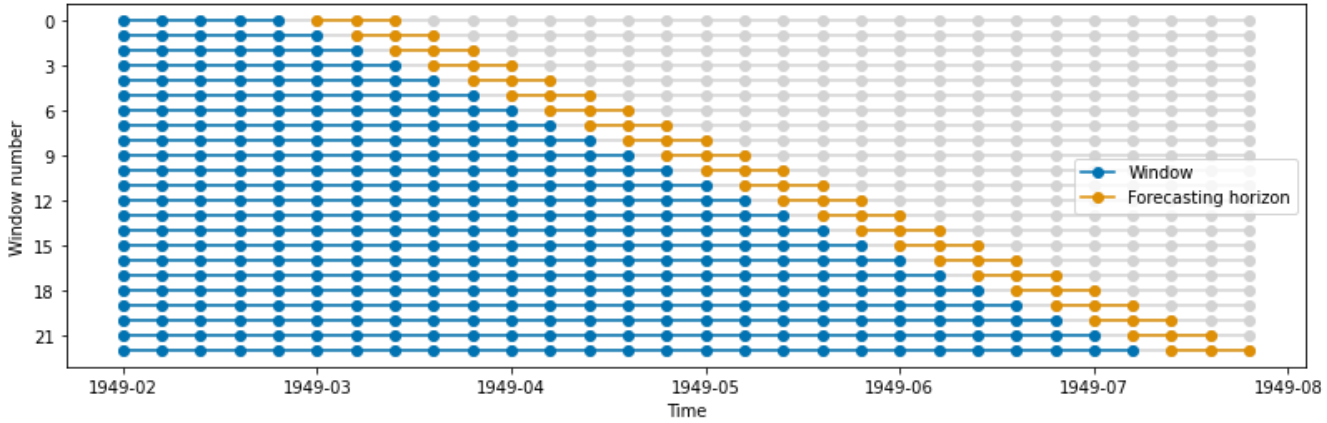


Cross-validation in time series



Does this look correct?

Cross-validation in time series



Train with [A]	*test with [B]*
Train with [A B]	*test with [C]*
Train with [A B C]	*test with [D]*
Train with [A B C D]	*test with [E]*
Train with [A B C D E]	*test with [F]*

[Don't Use Simple K-Fold Cross Validation on Time Series Data]

[N7 - Splitters]

Exercise Part #4

BONUS

N8 – Structural Breaks
N9 - TBATS
N10 – Intermittend Demand

Exercise Part #5

Review questions

- What is seasonality?
- What is trend?
- What is autocorrelation? What is partial autocorrelation?
- What is noise?
- What is a moving average? What is the difference between a centered MA and a trailing MA?
- What is a naive (aka “null”) prediction in the context of time series?
- What is a non-stationary time series?
- What are the practical considerations of using a trailing moving average vs a centered moving average to smoothen a time series?
- What is a window size?

Review questions

- What is differencing?
- What does it mean to detrend a time series?
- What does it mean for a time series to be stationary?
- What are the weaknesses of the ARIMA model?
- What is a theta model?
- How does exponential smoothing work?
- How does the Holt-Winters model work?
- Which model was used by most of the winners of the M5 competition?
- How does the expanding window approach differ from regular k-fold cross validation?
- What are the drawbacks of using MAPE as an evaluation metric?

Resources

- Time Series Analysis
- Modern Time Series Analysis - 2019 SciPy Tutorial
- Time Series with Konrad (discussion and tutorial from two Kaggle Grandmasters)
- Gradient boosting review