Forecasting Using Machine Learning Tools

N. Gautam

January 11, 2025 at 6:00pm Indian Standard Time
(7:30am US Eastern)

Workshop is intended for 3 hours

INTRODUCTORY REMARKS

- Background
  - *Professor, Department of Electrical Engineering and Computer Science, Syracuse University*
  - *Amazon Scholar*

- Disclaimer: Opinions expressed are solely my own and do not express the views or opinions of my past or present employers

- Taught an NPTEL course "Decision Making Under Uncertainty" for 6 years

- Gave a lecture in an NPTEL Live Series titled, "Art and Science of Forecasting" in May 2020 (available on YouTube)

Poll 1

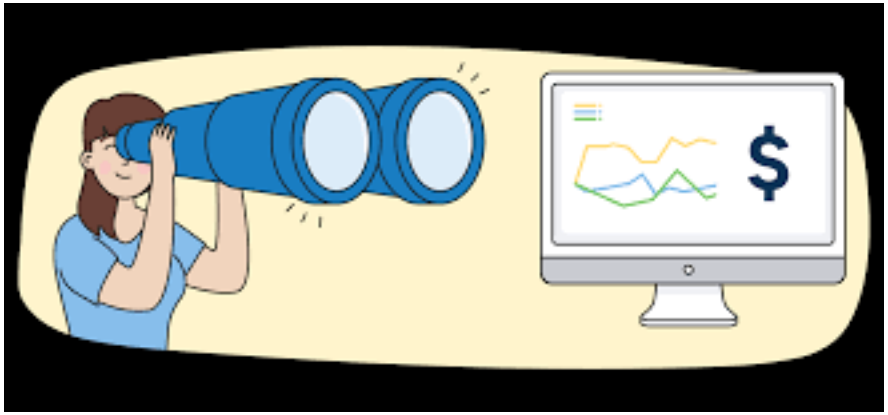
GROUND RULES

- There will be a lot of material presented
- Do not be distracted as one small miss would be very costly
- You are welcome to type up questions in the chat and after some time (about an hour) I will respond
- Do NOT unmute your mic to ask questions, unless I ask you to do so

WHAT IS FORECASTING?

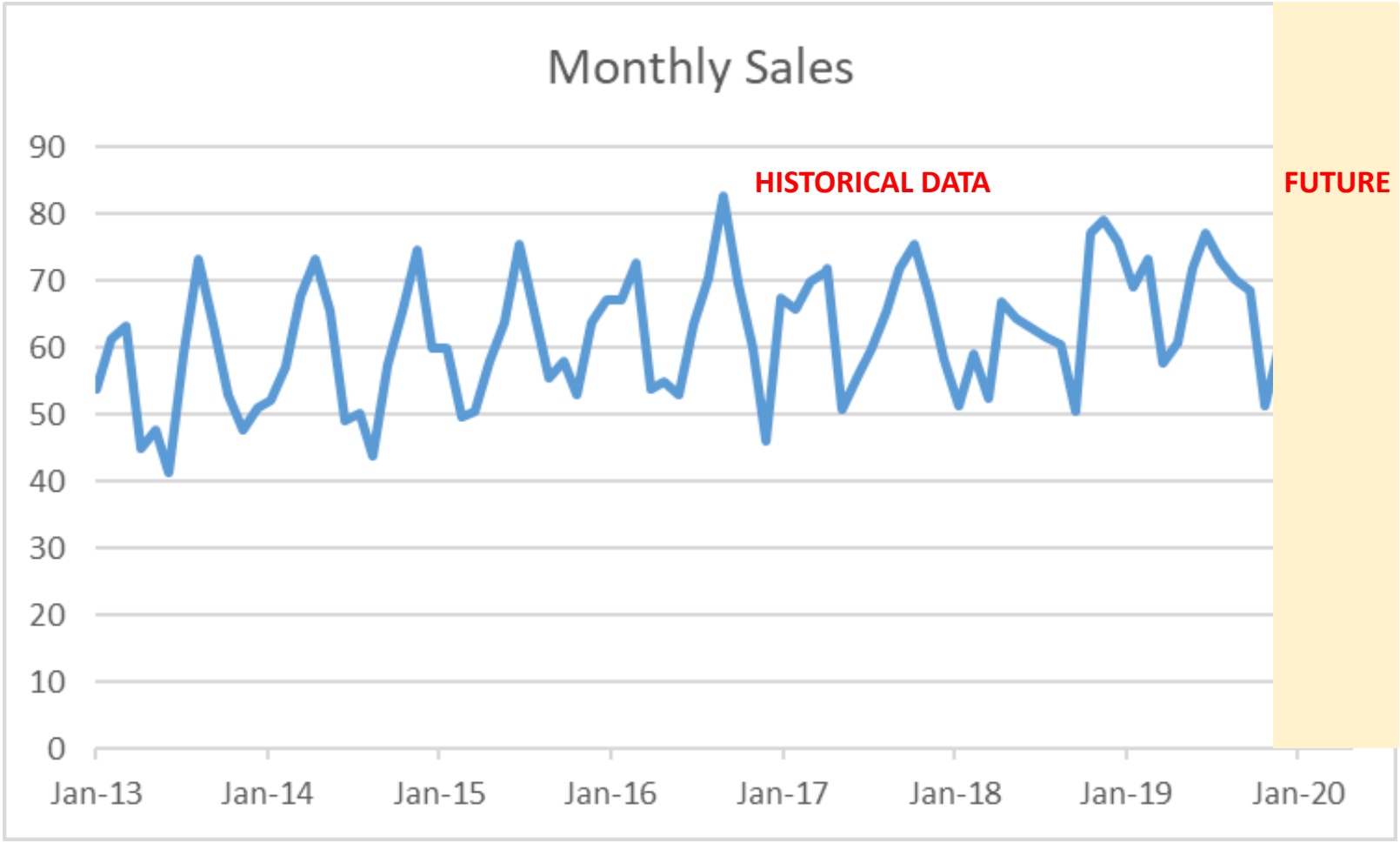
According to Google AI

- Forecasting is the process of
  - predicting future events or conditions
  - by analyzing past and current data
- It's a data-driven approach that helps businesses plan, budget, and make strategic decisions

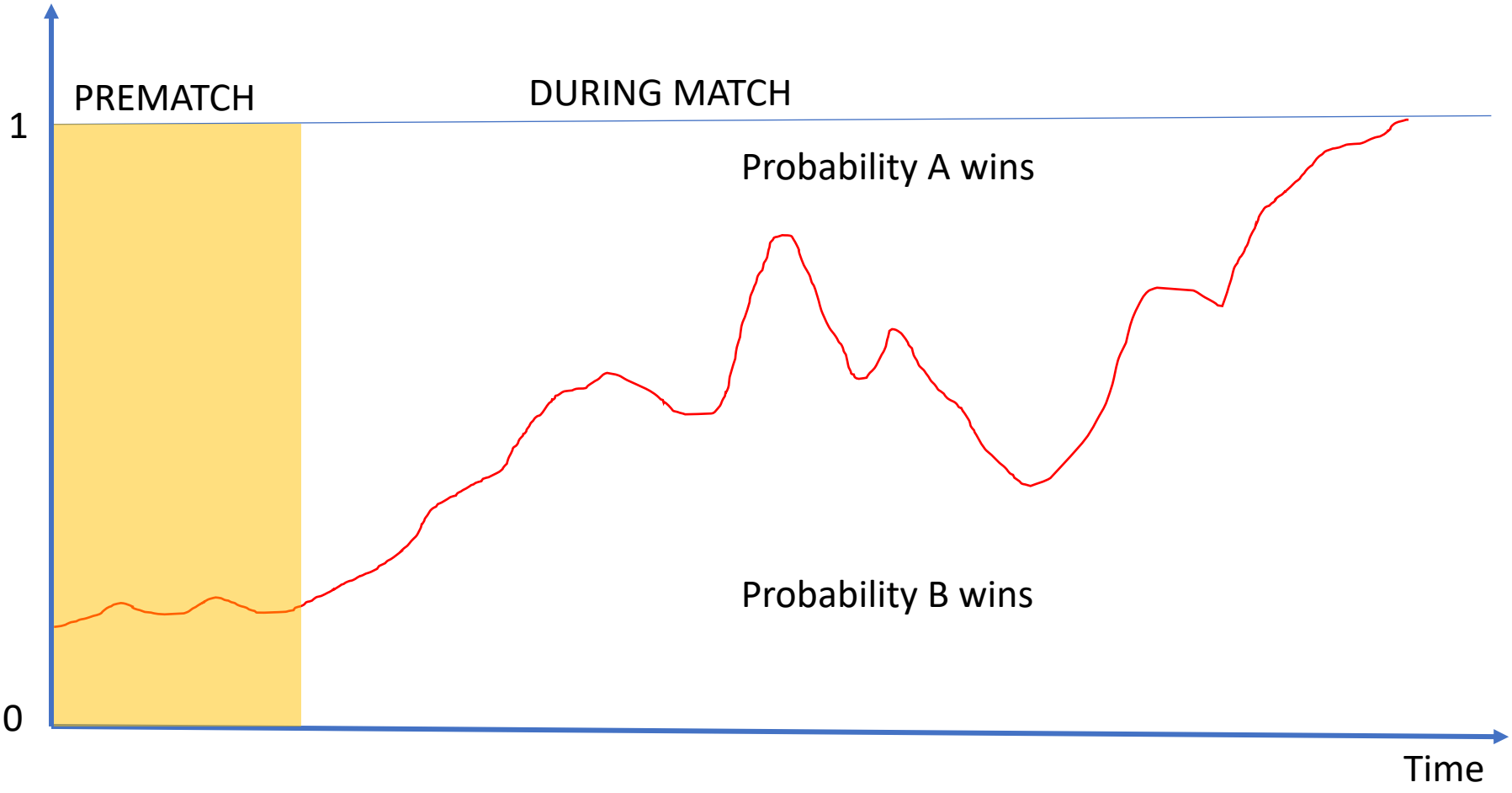


Poll 2

Sales Forecast

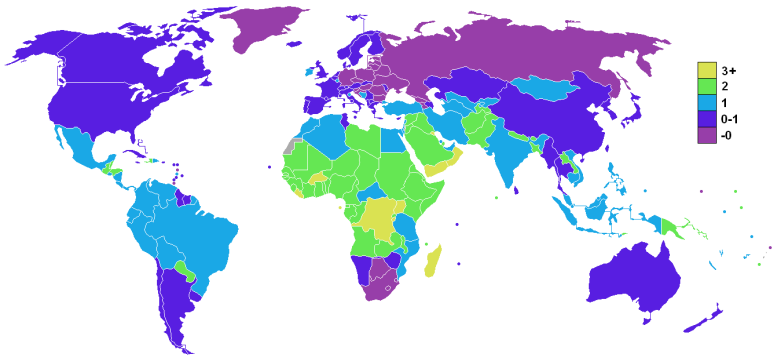


Predicting a Tennis Match's Winner

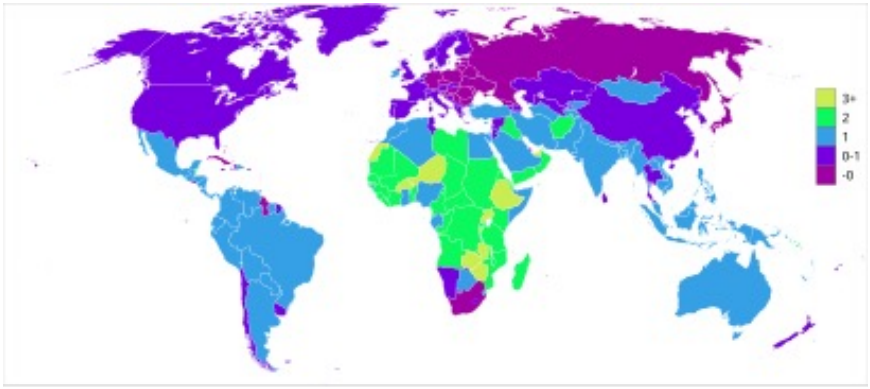


Population Growth Rate

2006



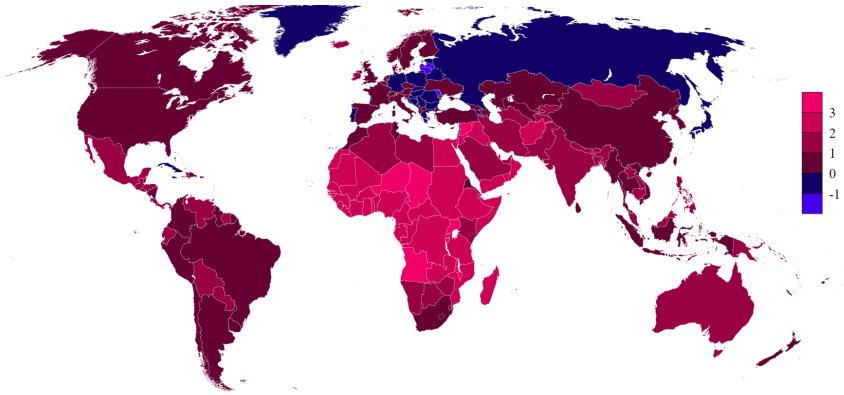
2011



Source: https://commons.wikimedia.org/wiki/

https://commons.wikimedia.org/wiki/File:Population_growth_rate_world_2018.svg

2019

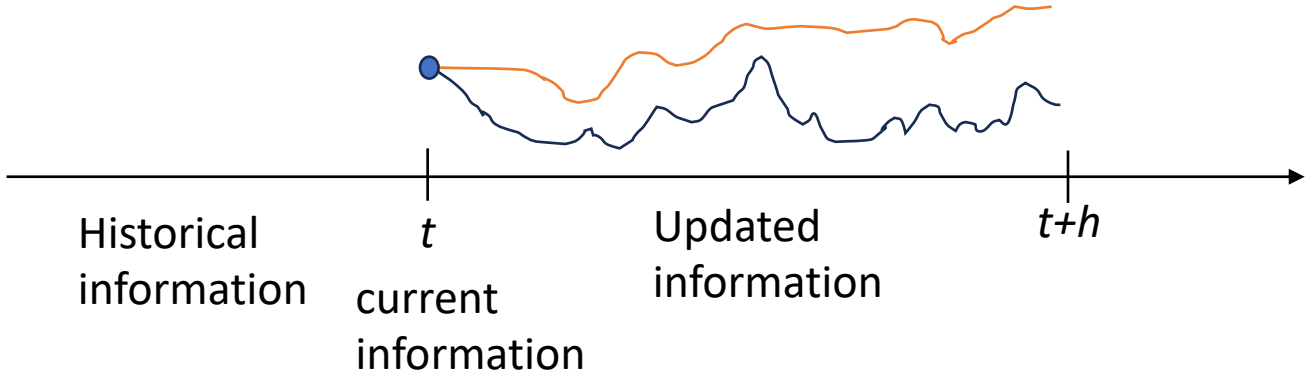


2050

FORECAST: What and When?

- Target forecasted:
  - Quantitative: Sales, temperature, stock price, wind power, arrival time
  - Categorical: Sports match, election, cyclone, movie award, super singer
- Forecast horizon:
  - Information known at time t
  - Forecast is to be made for time $t+h$, where h is the horizon
- Updates to forecast:
  - No (one-time forecast)
  - Allowed (as more information comes in, make updates)

Focus of this workshop



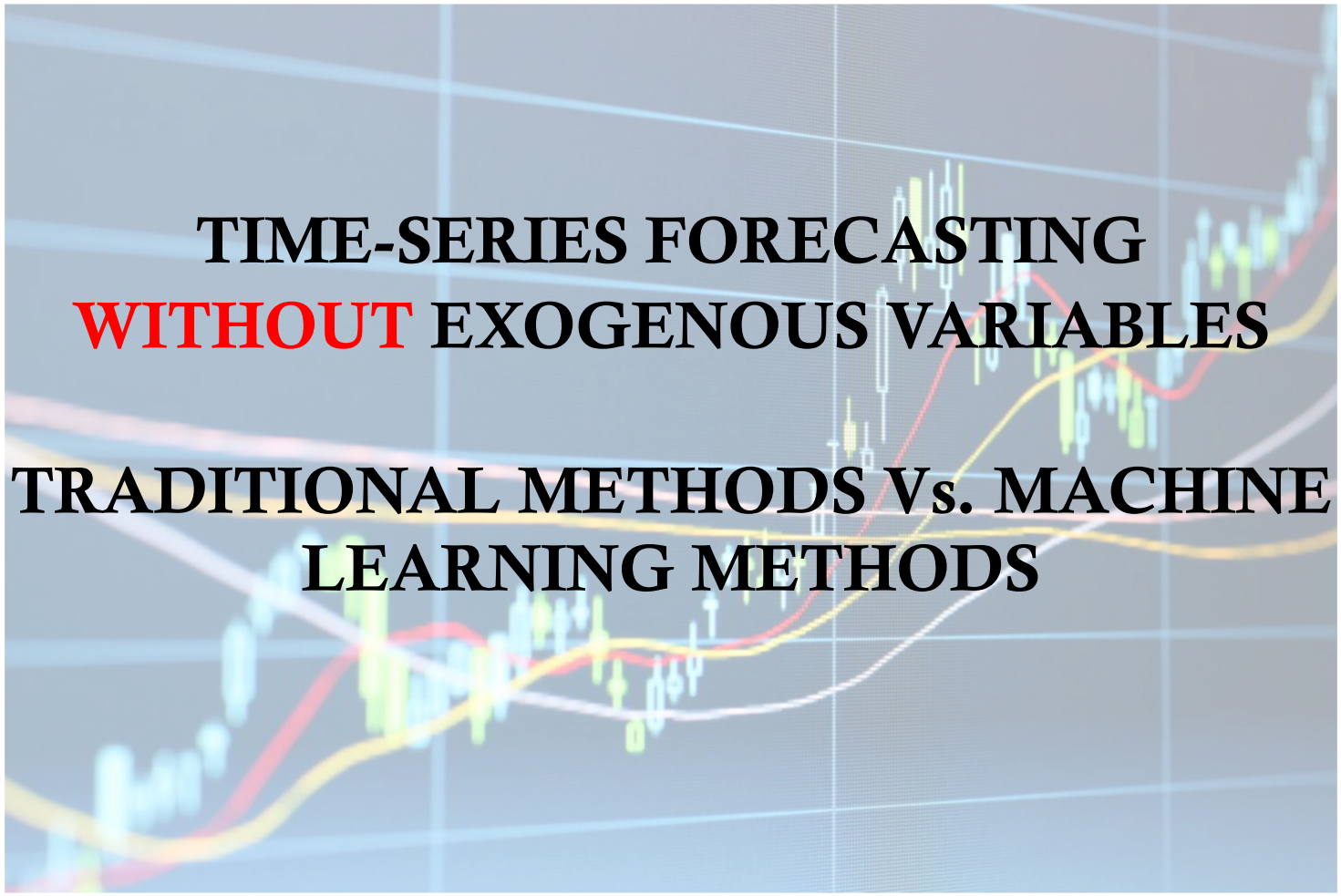
FORECASTS: With What Data?

- Historical data availability
  - As a time-series (sales data example)
  - Indirectly available (tennis match example)
- External factors
  - Significant for forecasting but not completely known or predictable
  - More significant than historical data, unknown, and unpredictable
- Data
  - Plenty available and forecast quality can be quickly checked
  - Sparingly available or verification is difficult or both

SCOPE OF THIS WORKSHOP

- Time-series target data with significant exogenous variables
- Target variable is quantitative with a one-time forecast
- Use of machine-learning models to make forecasts
- Execute code in python to implement models

Poll 3


TIME-SERIES FORECASTING
WITHOUT EXOGENOUS VARIABLES
TRADITIONAL METHODS Vs. MACHINE LEARNING METHODS

FORECASTING CALLS (SETTING)

- In a particular town there are records for the daily volume of emergency calls received
- This daily call volume is averaged over a month
- Data is available for a 40-year period (so, 480 data points)
- No other (exogenous) data is available
- Given historic daily average call volume including this month's, can we forecast what it would be next month?

Python Code for Forecasting Calls

- Open calls_data.csv using excel, check, and close
- Open Jupyter notebook provided: calls_data_forecast.ipynb
- Run the first to third cells
  - *You may need to pip install (or conda install) some libraries*
- Before running the fourth cell, check the following
  - *Is the CSV file in the same folder as the ipynb file?*
  - *If no, you will have to provide full PATH of where the CSV file is*
- As we go through the remaining cells, it may be useful to take notes at the cells by commenting out
  - *This way you will remember what each line in each cell does*

Poll 4

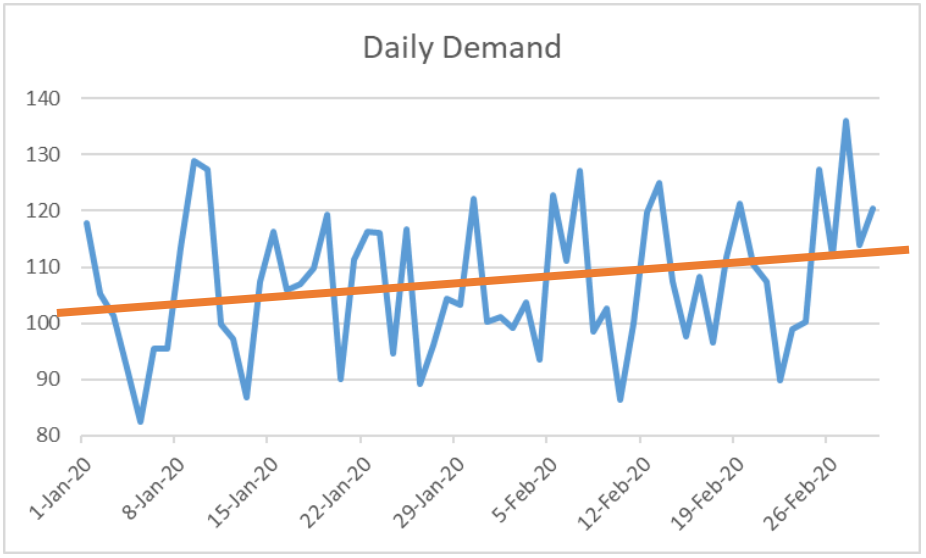
FORECASTING CALLS (PRELIM DATA ANALYSIS)

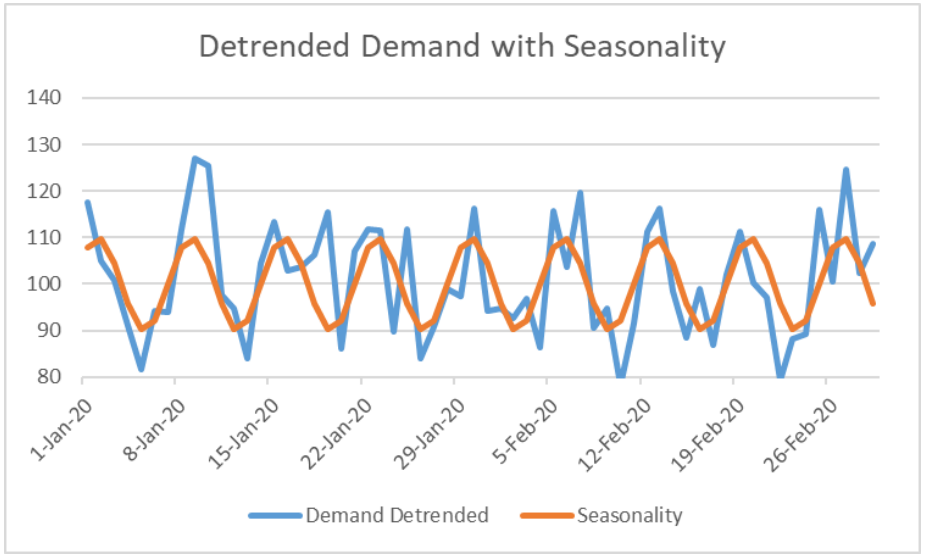
- Plot the data as a time-series to see how it looks like
  - Are there any trends?
  - Any seasonal patterns?
- Plot the autocorrelation and partial autocorrelation graphs
  - Are there correlations between lags in the time-series?
  - What is a good choice for seasonality?
  - What are significant features to be used in a forecasting model?
- No other (exogenous) data is available
  - Exogenous data will be the focus of future data sets

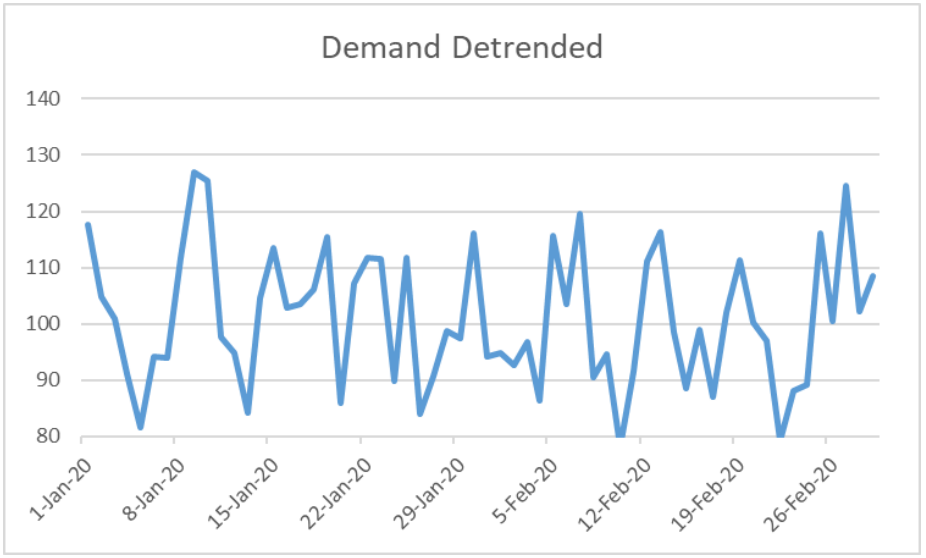
FORECASTING CALLS (MODEL CHOICES)

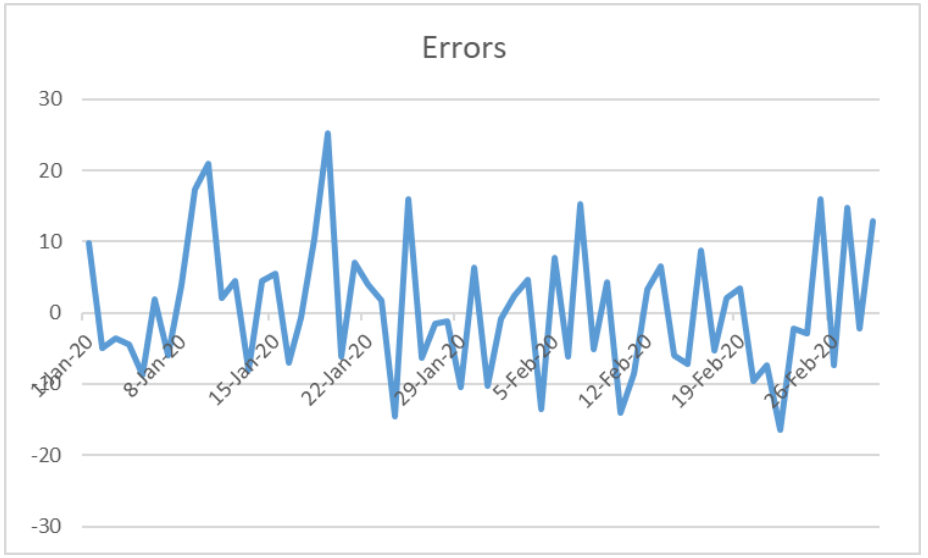
- This is ideal for traditional time-series analysis
  - *Exponential Smoothing (ES)*
  - *ARIMA*
- Typically, ES and ARIMA models are built when the training model size is less than 200 data points
  - So, create a subset data with the last 300 data points (we will use the first 200 of those for training, and the last 100 for testing)
- What about a machine-learning (ML) model?
  - *How would we build an ML model?*
  - *How would it compare against the traditional time-series models?*

Adjusting for Trend and Seasonality









FORECASTING CALLS (TREND & SEASONALITY)

- Detrending training data (and use in testing)
  - *Exponential Smoothing (ES): Trend can be set in function call*
  - *ARIMA: Use differencing to detrend and create stationarity*
  - *Machine Learning (ML): Use a regression line to obtain the trend*
- Account for seasonality in training (and use in testing)
  - *ES: Can specify seasonality in function call*
  - *ARIMA: Can use seasonal parameters in function call*
  - *ML: Can use features (based on partial autocorrelations) for seasonality*


ACCURACY

QUALITY OF FORECASTS

	1	2	3	4	5	6	7
Forecast sales	10	16	21	13	8	12	6
Actual sales	12	10	25	14	9	13	4

METRICS (Building block)

- Sum of squared errors

$(12-10)^2+(10-16)^2+(25-21)^2+(14-13)^2+(9-8)^2+(13-12)^2+(4-6)^2 = 63$

=> Mean square error = 9; root mean square error = 3

- Mean absolute error = $(2+6+4+1+1+1+2)/7 = 17/7 = 2.43$
- Median absolute error = 2
- 25^{th} percentile of absolute error = 1; 75^{th} percentile is nearly 4

QUALITY OF FORECASTS

	1	2	3	4	5	6	7
Forecast sales	10	16	21	13	8	12	6
Actual sales	12	10	25	14	9	13	4

METRICS (Relative)

- Mean absolute error ratio:
  - Mean actual sales = 87/7
  - Mean absolute error = (2+6+4+1+1+1+2)/7 = 17/7 = 2.43
  - Mean absolute error ratio = 17/87 = 19.5%
- Mean absolute % error = (2/12+6/10+4/25+1/14+1/9+1/13+2/4)/7 = 24.1%


TIME-SERIES FORECASTING
WITH EXOGENOUS VARIABLES

FORECASTING SALES (SETTING)

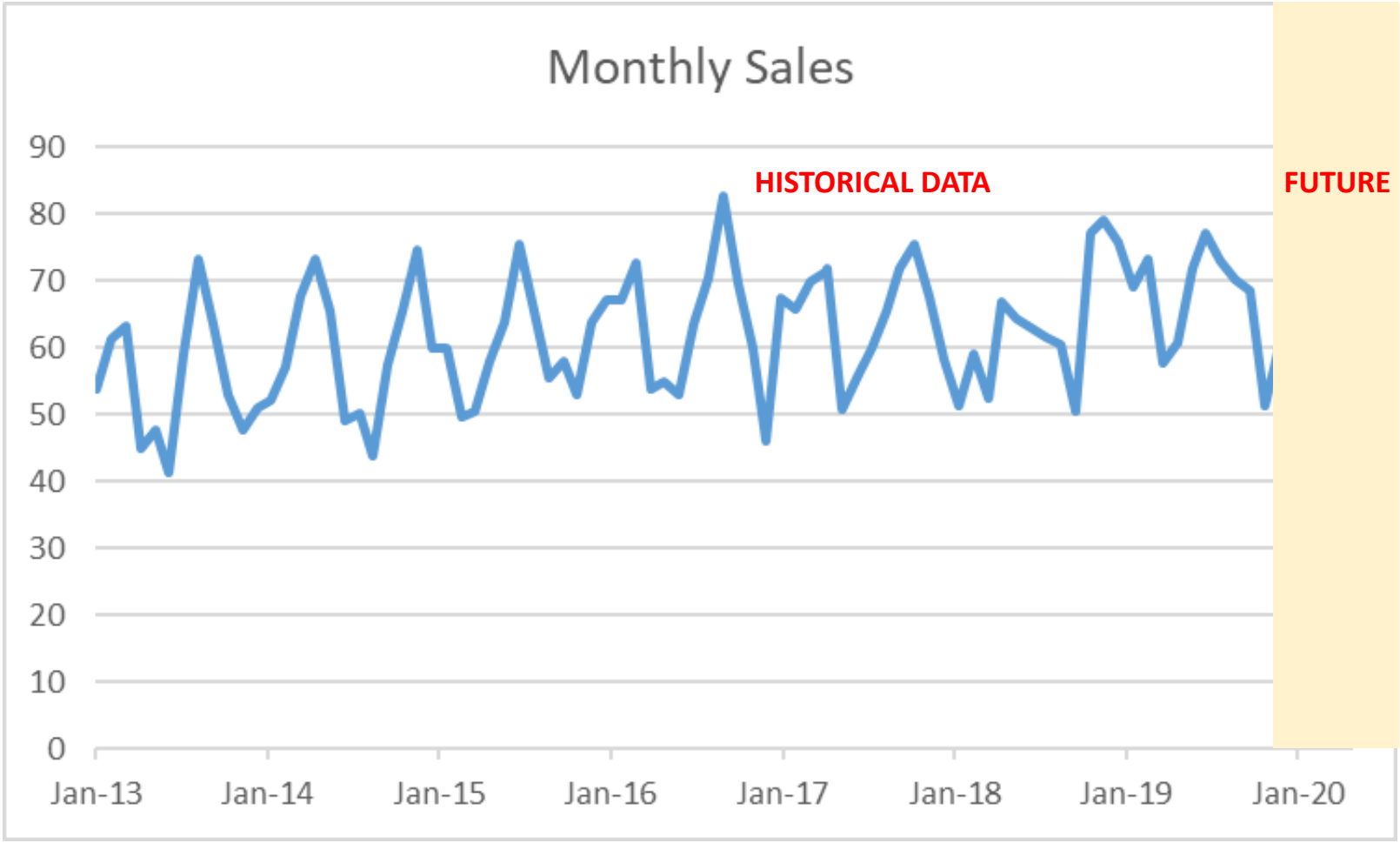
- For a particular product at a particular store, the number of "units sold" daily is available for a 2-year period
- Task: At the end of a day, what is the forecast of the "units sold" the next day?
- Exogenous data available (at end of a day) for next day:
  - Inventory level (number of products available at start of the day)
  - Demand forecast (number of products that would be demanded that day)
    - Not all demand would convert into sale
  - Price of the product
  - Discount available
  - Weather condition predicted
  - Holiday/promotion
  - Competitor pricing
- Given historic units sold including the day in question as well as the exogenous data for next day, can we forecast "units sold" the next day?

EXOGENOUS DATA

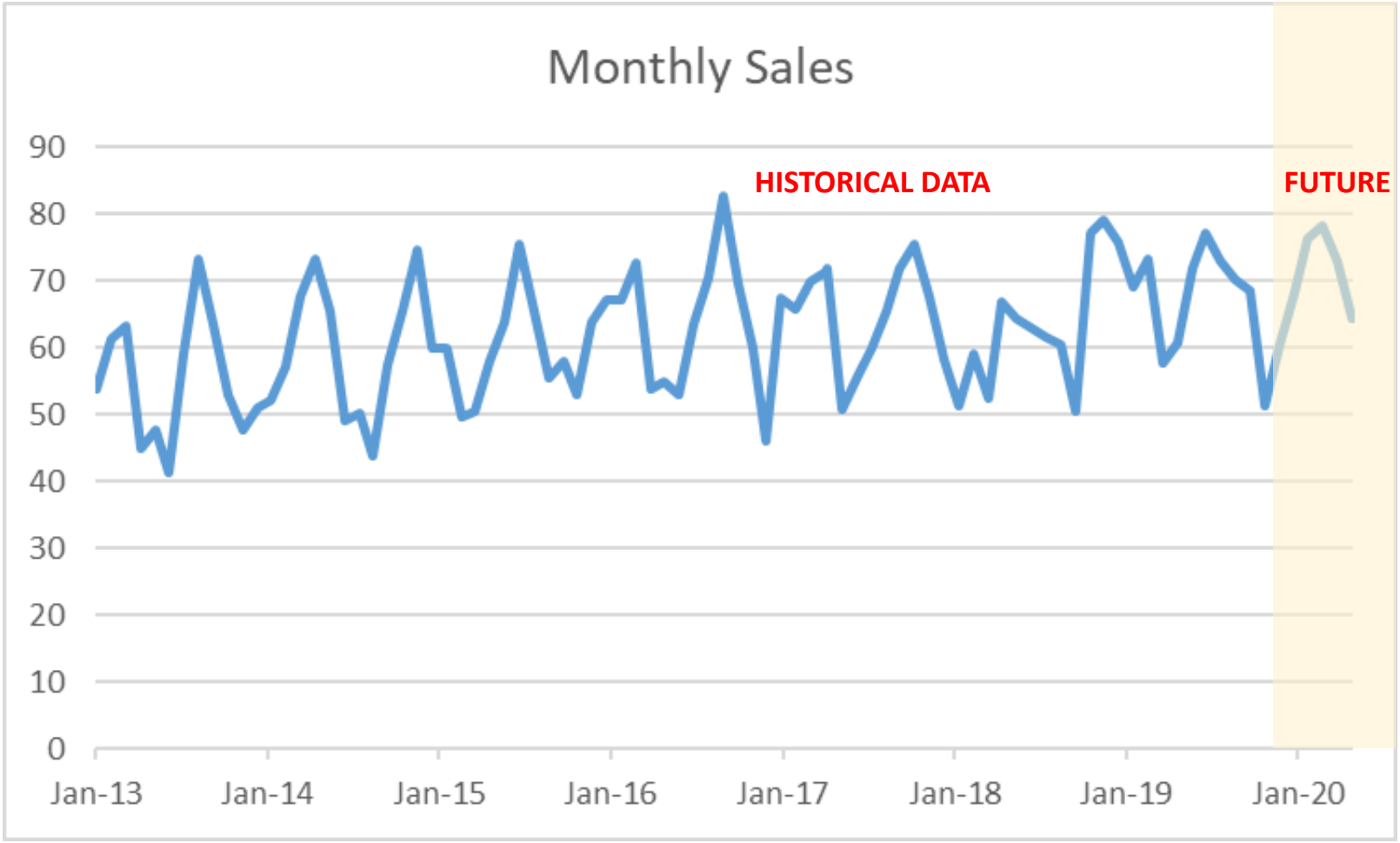
- Exogenous data available (at end of a day) for next day:
  - Inventory level, Demand forecast, Price of product, Discount available, Weather condition predicted, Holiday/promotion, Competitor pricing
- Notice how each of the above variables can affect the units sold
- Some of the variables may not be known precisely (esp. forecasts)
- There may be other exogenous factors that affect the number of units sold (but cannot be captured)

Poll 5

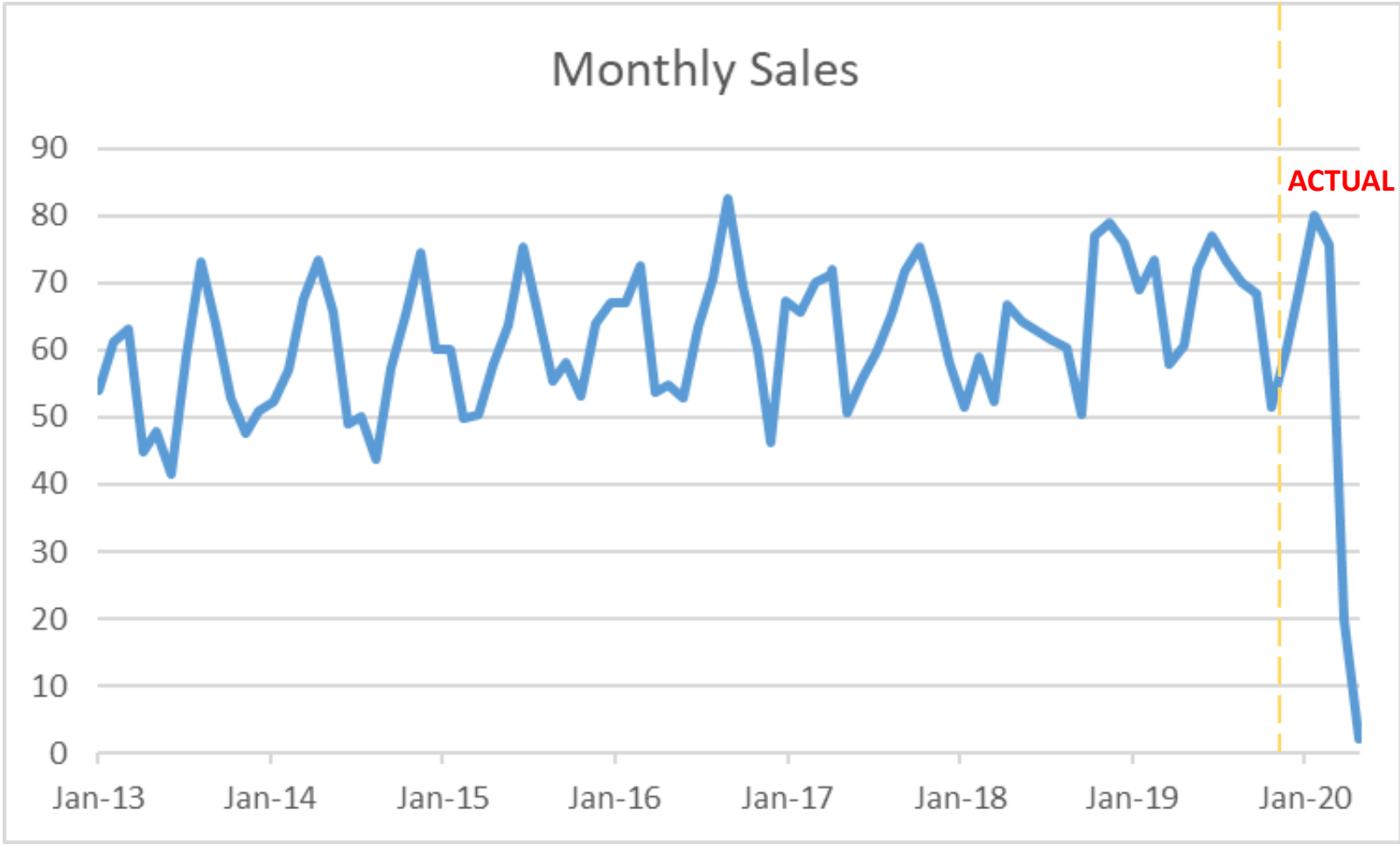
Sales Forecast



Prediction of Future Sales



Actual Sales



MIDWAY REMARKS

- It is a myth that if you throw a lot of data, you can make 100% accurate forecasts
- Yet, forecasts do improve with relevant exogenous data
- ML models are an excellent option for forecasting
- What is next?
  - *Go over python code for sales forecast*
  - *State and solve the solar forecast problem*
  - *Next steps in terms of related courses*
  - *Details about assessment*

FORECAST: BY WHAT METHOD?

- Time-series methods (Traditional, without exogenous variables)
  - Exponential smoothing (ES)
  - Autoregressive integrated moving average (ARIMA)
  - Ideal when data is small
- Time-series methods (with exogenous variables)
  - SARIMAX (i.e., ARIMA with Seasonality and eXogenous)
  - Prophet (by Meta/Facebook uses multiple seasons)
  - Ideal when data is small and exogenous features known with certainty
- Physics-based methods
  - Numerical weather forecasting
  - Ideal when all dependent variables can be accounted for
- Machine-learning models
  - Tree-based ensembles (random forest, gradient boosting, etc.)
  - Ideal for large data and exogenous variable are significant & uncertain

Poll 6

Python Code for Forecasting Units Sold

- Open prod2store3sales.csv using excel, check, and close
- Open Jupyter notebook provided: Sales_data_forecast.ipynb
- Run the first to third cells
- Run the fourth cell to read the CSV file

FORECASTING SALES (PRELIM DATA ANALYSIS)

- Use only the first 600 datapoints to answer the following
  - This is to not view the test data
- Plot the data as a time-series to see how it looks like
  - Are there any trends?
  - Any seasonal patterns?
- Plot the autocorrelation and partial autocorrelation graphs
  - Are there correlations between lags in the time-series?
  - What is a good choice for seasonality?
  - What are significant features to be used in a forecasting model?

FORECASTING SALES (MACHINE LEARNINGMODEL)

- What about a machine-learning (ML) model?
  - *What features would we use to build an ML model?*
  - *How would various ML models compare against each other?*

Poll 7

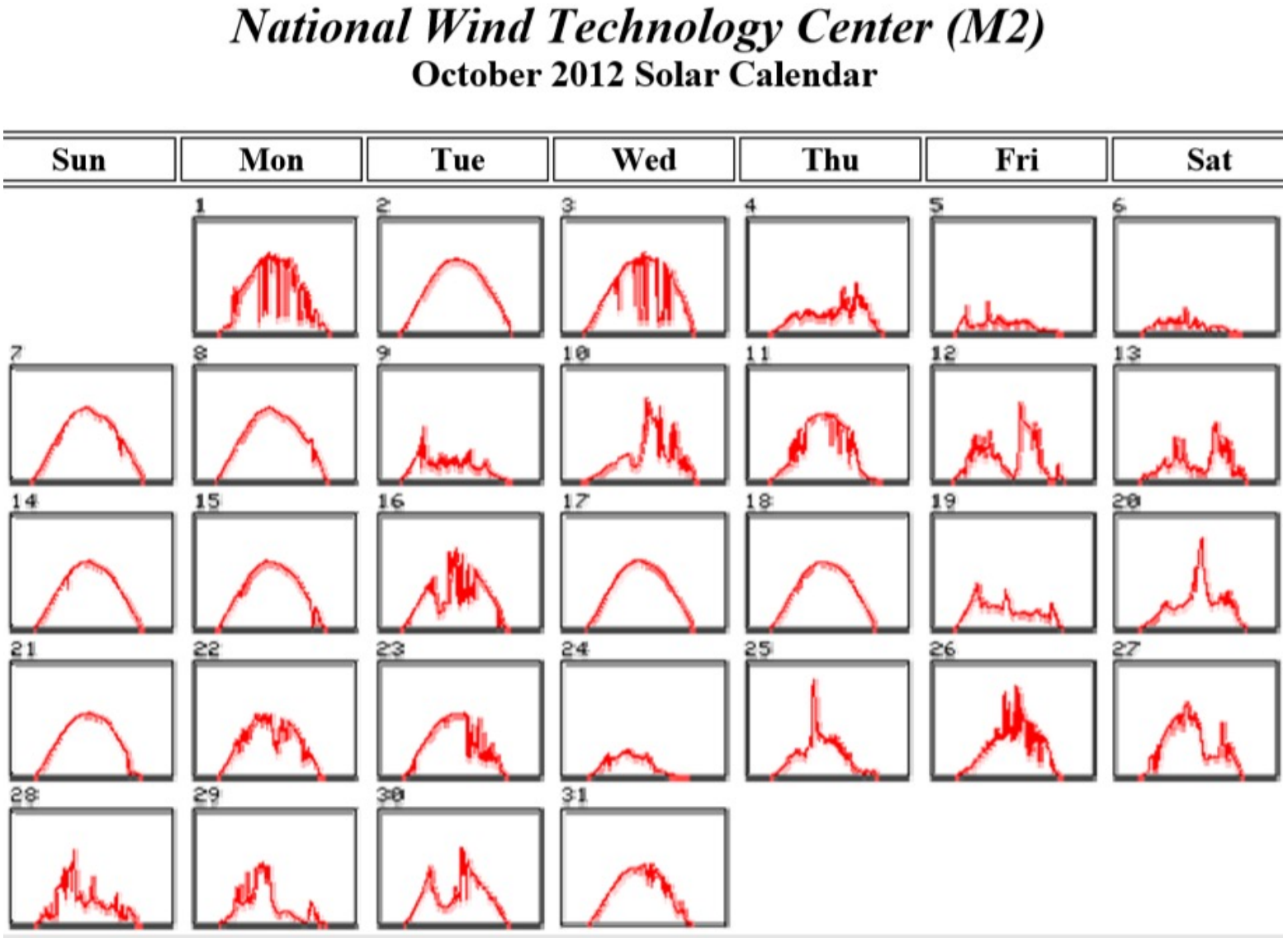
FORECASTING SOLAR POWER

- Renewable sources of power such as wind and solar have significant uncertainty and variability

- Ability to forecast power from renewable sources for the near future would result in effective utilization of non-renewable energy

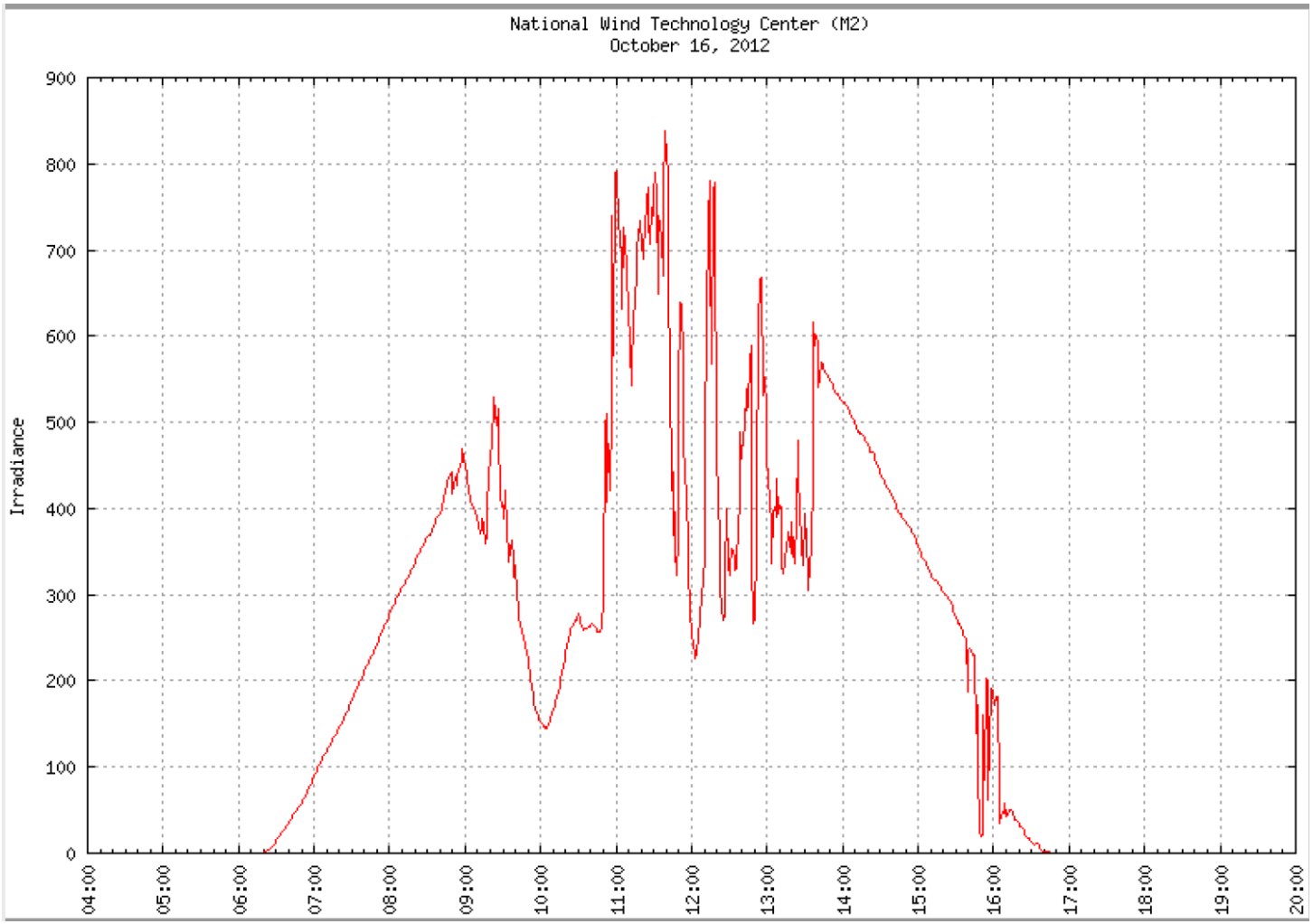
- Batteries are generally not very efficient

Solar Data and Forecasting Challenges



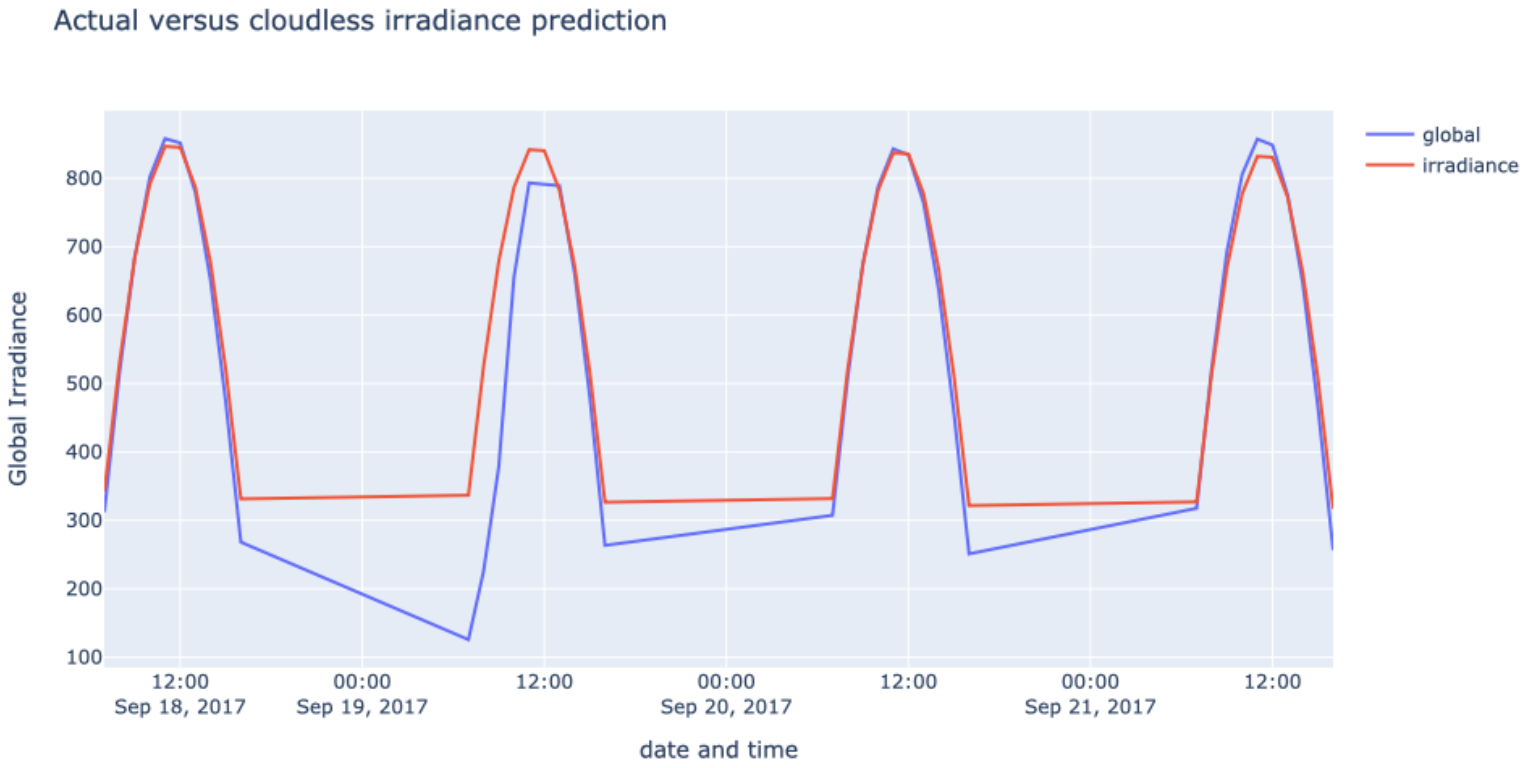
Poll 8

Solar Data for a Single Day



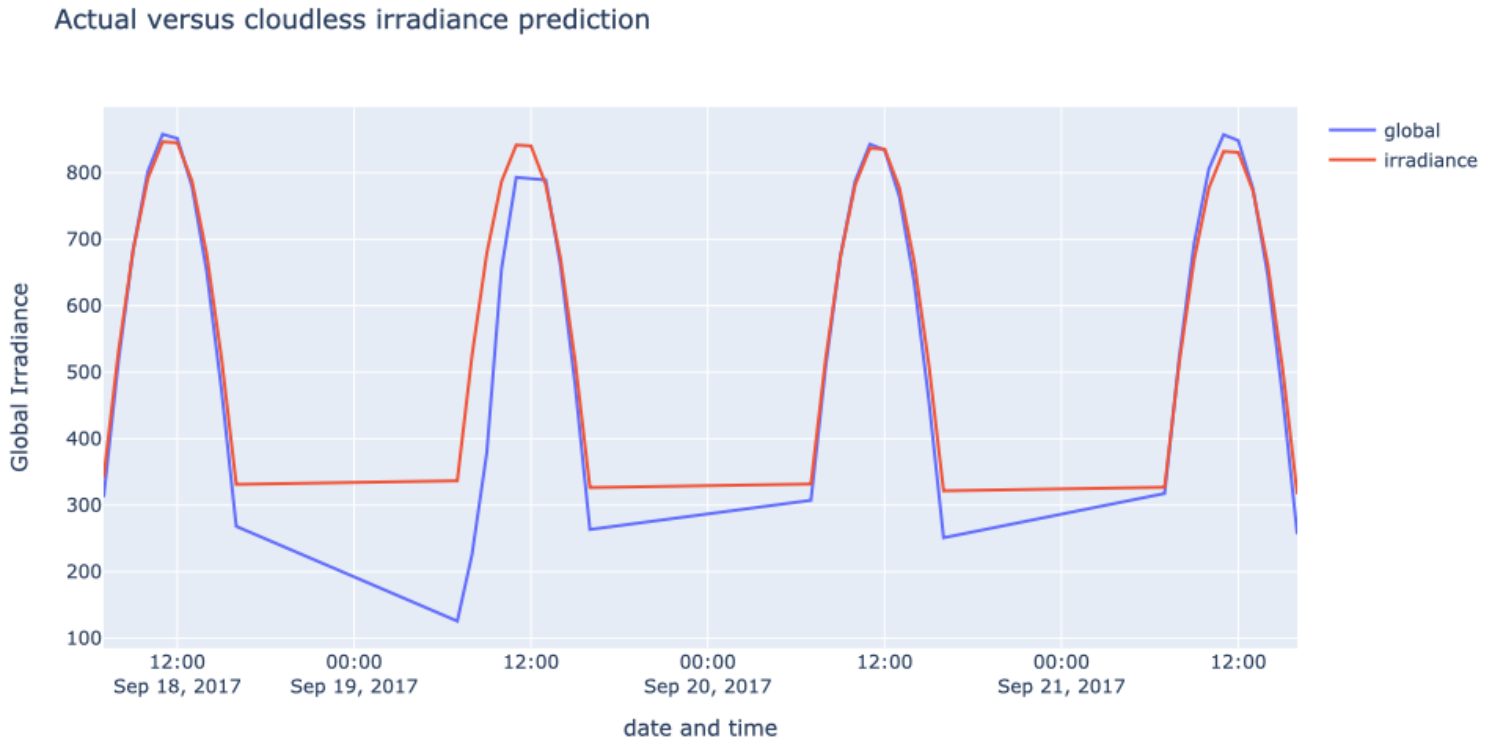
FORECASTING SOLAR POWER (PREPROCESS DATA)

- We only use 8 hours of day-time hourly data (8am to 3pm; denoting 8-9am and 3-4pm)
- We use astro-physics method to obtain cloudless irradiance



FORECASTING SOLAR POWER (PREPROCESS CONTINUED)

- Divide observed ‘global’ irradiance by cloudless irradiance
- Gives a scaled irradiance that adjusts for trend
- Need more information to perfectly adjust for season

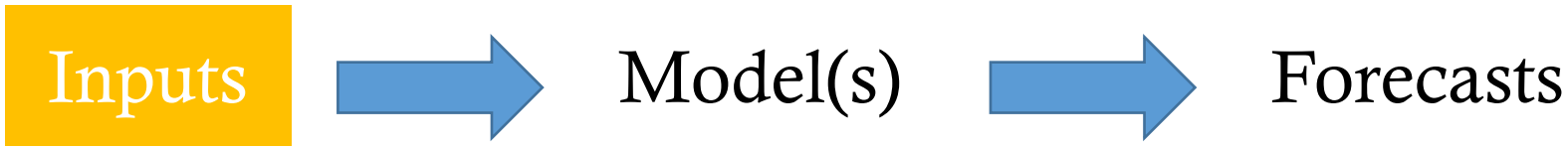


FORECASTING SOLAR POWER (PRELIM DATA ANALYSIS)

- Code NOT provided, only demo
- Plot unscaled data and scaled data to see the difference
- Study the impact on autocorrelation
- What features to use for lags?

Inputs

- Historical data (as a time series) for what you want to forecast
  - *Example: Solar power from a photovoltaic farm*
- Historical exogenous variables at the same time points
  - *Example: Weather data, cloud cover, temperature, date*
- Future predictions of exogenous variables
  - *Example: Weather forecasts*



Poll 9

Python Code for Forecasting Solar Data

- Save 3 CSV files
  - solarTAC_hourly_2016_4_months_scaled.csv
  - solarTAC_hourly_2017_4_months_scaled.csv
  - solarTAC_hourly_2016_4_months_unscaled.csv
- Open Jupyter notebook provided: Solar_data_forecast.ipynb

IMPORTANT COMMENTS

- Like any other Machine-Learning exercise, here too we must
  - *Be careful about feature selection*
  - *Be concerned about overfitting/underfitting*
  - *Experiment with multiple hyper-parameters*
  - *Try to predict by combining multiple models*
  - *Attempt to build hybrid models to improve forecasts*
- We will address some of the above
  - In the exercise for completion certificate
  - Provided in the end of this slide deck
- We have barely scratched the surface, what next?

TO LEARN MORE

Forecasting Using Statistical Models

- Online introductory-level textbook
- Forecasting: Principles and Practice (https://otexts.com/fpp3/) by Rob J Hyndman and George Athanasopoulos
- Uses time-series models implemented in R
  - *Python libraries available at other sites*
- Numerous real-life datasets to analyze
- Rob J Hyndman's website https://robjhyndman.com/ and also checkout the Hyndsight blog
- Does NOT do Machine Learning based forecasts

Forecasting Using Machine Learning Models

- Sign up with Kaggle if you have not done so already
- Take this course created by Ryan Holbrook
  - https://www.kaggle.com/learn/time-series
- Practice what you learned using this
  - https://www.kaggle.com/competitions/store-sales-time-series-forecasting

Syracuse University
College of Engineering
& Computer Science

Master of Science in Operations Research and System Analytics

(website)

Becoming an Excellent Forecaster

- Superforecasting: The Art and Science of Prediction by Philip Tetlock and Dan Gardner
- "How predictable something is, depends on what we are trying to predict, how far into the future, and under what circumstances."
- "Foresight is the product of particular ways of thinking, of gathering information, of updating beliefs."
- Hedgehogs versus foxes
- Appendix: Ten Commandments for Aspiring Superforecasters

CONCLUDING REMARKS

- Machine-learning models are ideal when we have
  - *Large amounts of data*
  - *Exogenous variable that are significant*
  - *Some uncertainty due to all factors not captured*
  - *Some variability even under similar features*
- Know when to not use ML models
  - *If a statistical time-series method would be better*
  - *If a physics-based model would be better*
- Try multiple models
  - *Sometimes hybrid models work better*
  - *Sometimes aggregation may be needed*
  - *There is no single model that is better than others*
- Limitation of Tree-based ML models
  - *Does not do extrapolation (so detrending is crucial)*
  - *Exponential smoothing model is a choice for trends*

ASSESSMENT

- You will receive a set of MCQ from NPTEL
- To answer them, you would have to run the three python codes we saw in the workshop
- They will be graded and NPTEL would send certificates to eligible candidates (who attend the sessions and complete the MCQ)
- All administrative questions will be answered by NPTEL