eleks®

HOW TO BUILD ADVANCED PREDICTION WITH ADDING EXTERNAL DATA

Taras Firman, Ph.D.,
Senior Data Scientist

AI & Big Data CONFERENCE

Monthly VS weekly VS daily

M:

Advantages **– Fast to compute, easier to model, easier to identify changes in trends, better for strategic long term forecasting.**
Disadvantages **– If you need to plan as the daily level for capacity, people and spoilage of product then higher levels of forecasting won't help understand the demand on a daily basis as a 1/30th ratio estimate is clearly insufficient.**

W:

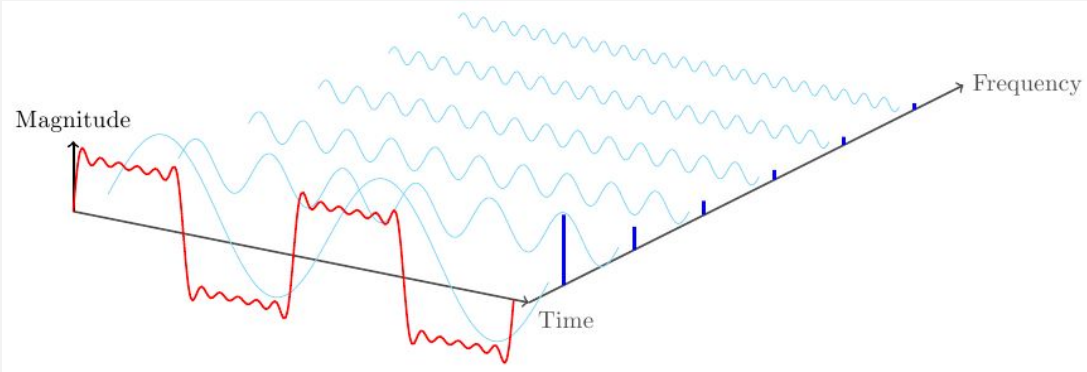
Advantages **– When you can't handle the modeling process at a daily level you "settle" for this. When you have very systematic cyclical cycles like "articice extents" that follow a rigid curve and not need for day of the week variations.**
Disadvantages **– Floating Holidays like Thanksgiving, Easter, Ramadan, Chinese New Year change every year and disrupt the estimate for the coefficients for the week of the year impact which can be handled by creating a variable for each.**

D:

Advantages **– Weekly data can't deal with holidays and their lead/lag relationships. If a holiday has days 1,2,3 before the holiday as very large volume a daily model can forecast that while the weekly won't be able year in and year out model and forecast that impact as the day of the week that the holiday occurs changes every year.**
Disadvantages **– Slower to process, but this can be mitigated by reusing models.**

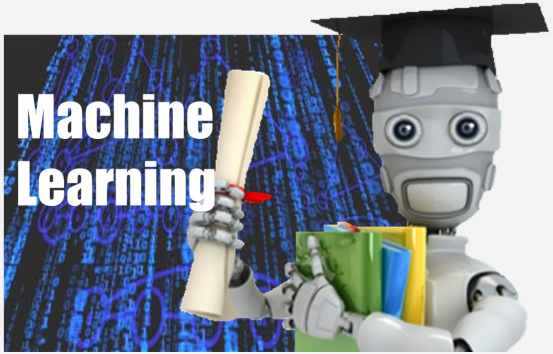
Prediction approaches

- **Time domain**



- **Frequency domain**

- **Machine learning**



Forecasting’s short history

- **Generic models**
  (Moving Average Process MA(q), Exp Smoothing, Autoregressive Process AR(p), Autoregressive Moving Average ARMA(p, q), Autoregressive Integrated Moving Average ARIMA (p, d, q))

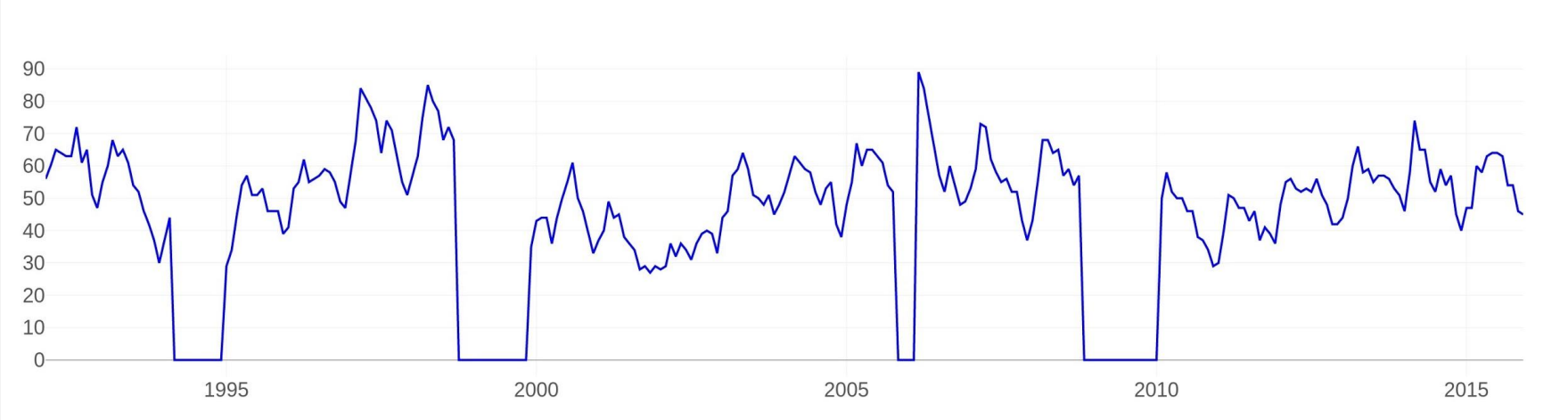
- **State Space models and Kalman Filter**

- **Multivariate vector models**

- **Feature extraction & ML**

- **DL approaches**
  (LSTM Recurrent Neural Networks)

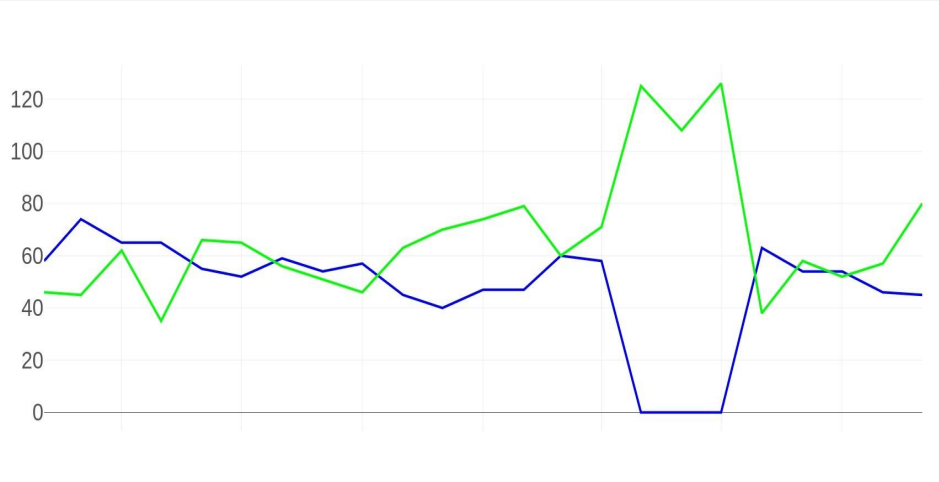
Interpolation and extrapolation



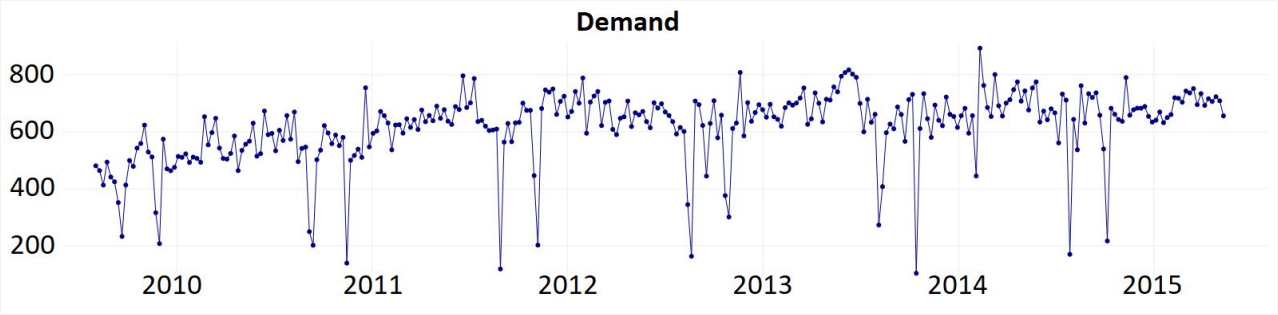
$$\texttt{range}(seasonality) \rightarrow \max$$

$$y_i = f_i(y_{-i}) \text{ for all } i \in I \subset \{j : y_j = 0\}$$

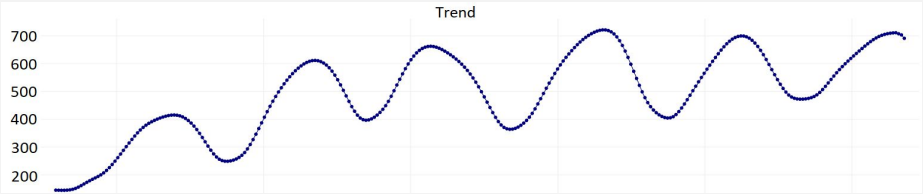
Retail data. Substitutes. Categories

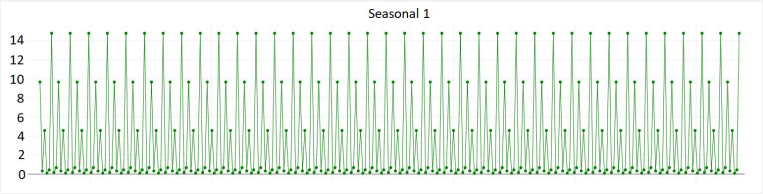


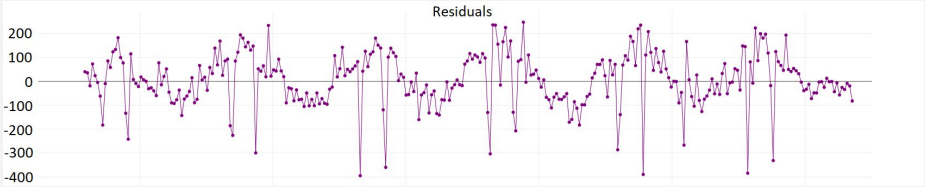
Decomposition tactic

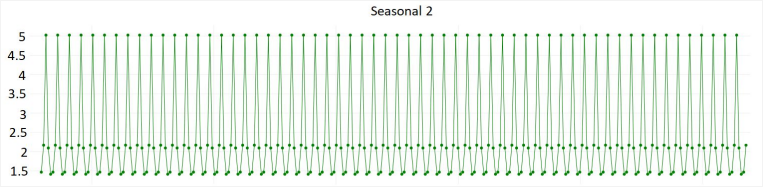


$$\begin{cases} \texttt{range}(seasonality) \rightarrow \max \\ \texttt{range}(residuals) \rightarrow \min \end{cases}$$







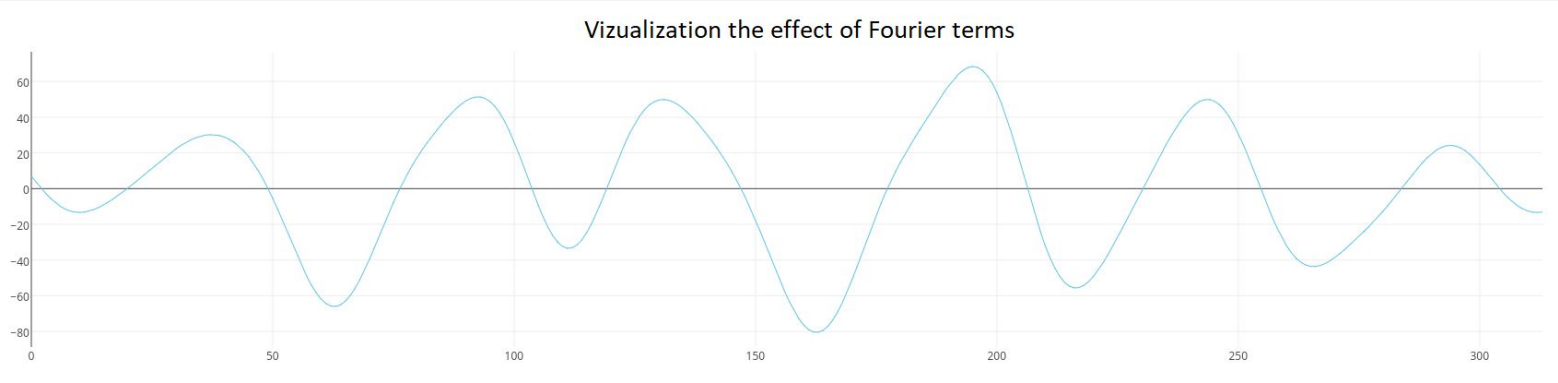


Trend aproximation

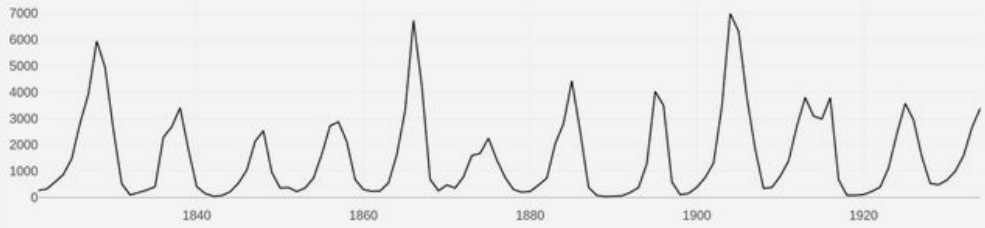
The approximation of the trend can be found from the formula below

$$y_{trend} = P_n(t) + \sum_{\alpha_m \in A_k} \left(a \sin\left(\frac{2\pi\alpha_m t}{T}\right) + b \cos\left(\frac{2\pi\alpha_m t}{T}\right)\right)$$

where $P_n(t)$ is a degree polynomial and A_k is a set of indexes, including the first k indexes with highest amplitudes.

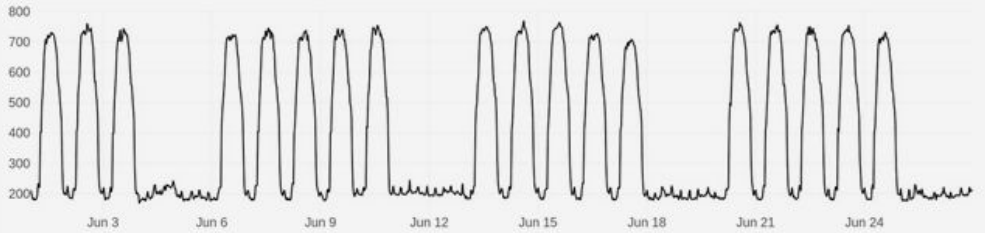


Seasonality VS Cycles



Canadian lynx data
Aperiodic population cycles of approximately 10 years

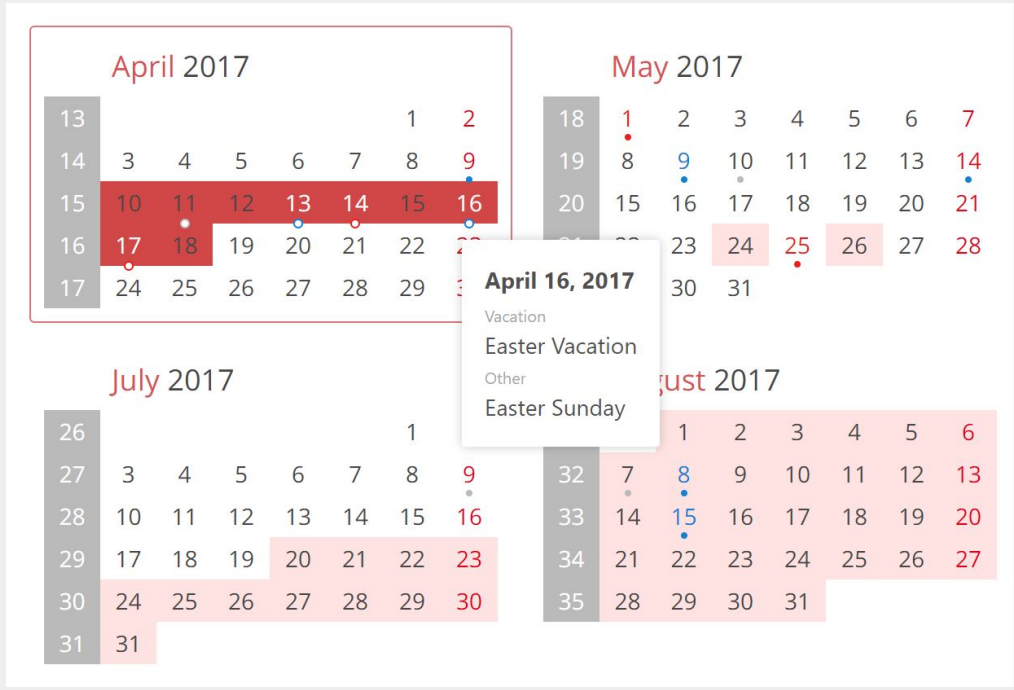
Monthly **sales of new one-family houses sold in USA**
Strong seasonality within each year and strong cycles with period 6-10 years



Half-hourly electricity demand in England
Multi-seasonality with daily and weekly patterns

Calendar

- Holidays
- Vacation
- School vacation
- Fasting and Abstinence
- Festivals
- Shopping holiday



External sources. APIs

OfficeHolidays
http://www.officeholidays.com/

OfficeHolidays

HolidayCalendar
https://holidaycalendar.com/

10Times
https://10times.com

10times

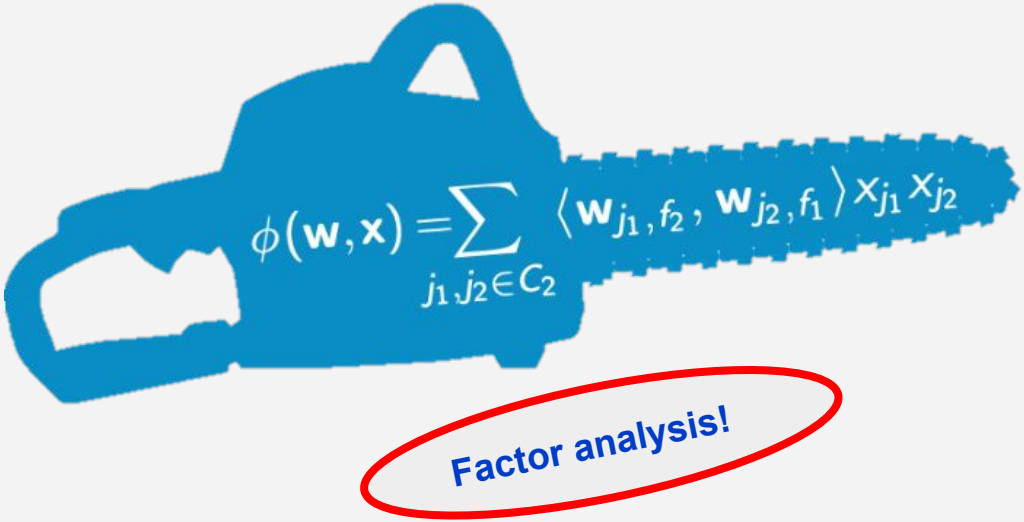
Wunderground API
https://www.wunderground.com/weather/api/

WEATHER UNDERGROUND

Google trends
https://trends.google.com/

Feature engineering for residuals

- One-hot encoding
- Counting
- Statistical moments
- Percentiles
- Lags
- Logs
- Peaks
- Least-squares spectral analysis
- Nonlinear transformations



Correlation types

- **Pearson correlation** is statistic to measure the degree of the relationship between linearly related variables.

Assumptions: both variables should be normally distributed and have linearity and homoscedasticity relationship (normally distributed about the regression line)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- **Spearman rank correlation** is non-parametric test that is used to measure the degree of association between two variables.

Assumptions: it doesn't make any assumptions about the distribution.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{where } d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$$

- **Kendall tau** is a statistic used to measure the ordinal association between two measured quantities.

Assumptions: data must be at least ordinal and scores on one variable must be montonically related to the other variable.

$$\tau = \frac{s_1 - s_2}{\frac{1}{2}n(n-1)}$$

where s_1/s_2 is number of concordant/discordant pairs

How to work with a short history?

Predicting the Past and Predicting the Future

Stochastic Simulation (Monte-Carlo)

Error measuring

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2$$

is scale-dependent

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

is scale-dependent

$$\mathrm{MPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{A_t - F_t}{A_t}$$

is the computed average of percentage errors. The formula can be used as a measure of the bias in the forecasts

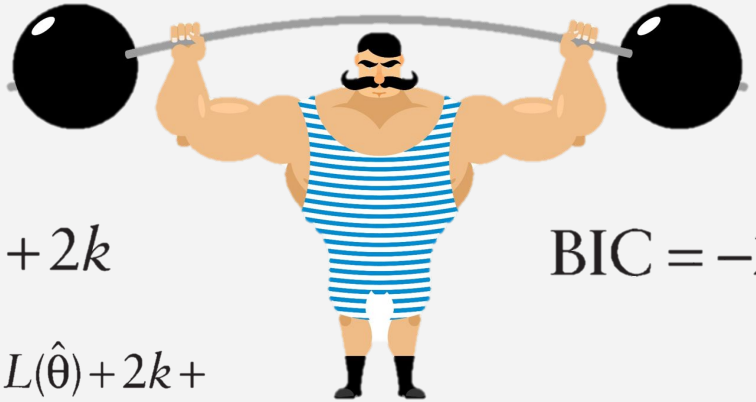
$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

usually expresses accuracy as a percentage. It puts a heavier penalty on negative errors, than on positive errors.

$$\mathrm{SMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

is an accuracy measure based on percentage (or relative) errors. One supposed problem with SMAPE is that it is not symmetric since over- and under-forecasts are not treated equally.

Robustness. Model selection



$$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2k$$

$$\mathrm{BIC} = -2\log L(\hat{\theta}) + k\log n$$

If n/k < 40: $\mathrm{AIC}_c = -2\log L(\hat{\theta}) + 2k + \\ + (2k+1)/(n-k-1)$

where

- θ - the set of model parameters;
- $L(\hat{\theta})$ - the likelihood of the candidate model given the data;
- k - the number of estimated parameters in the candidate model;
- n - the number of observations.

Existing solutions with Python

Pandas

Statsmodels

Scikit-learn

XGBoost

PyFlux

Prophet

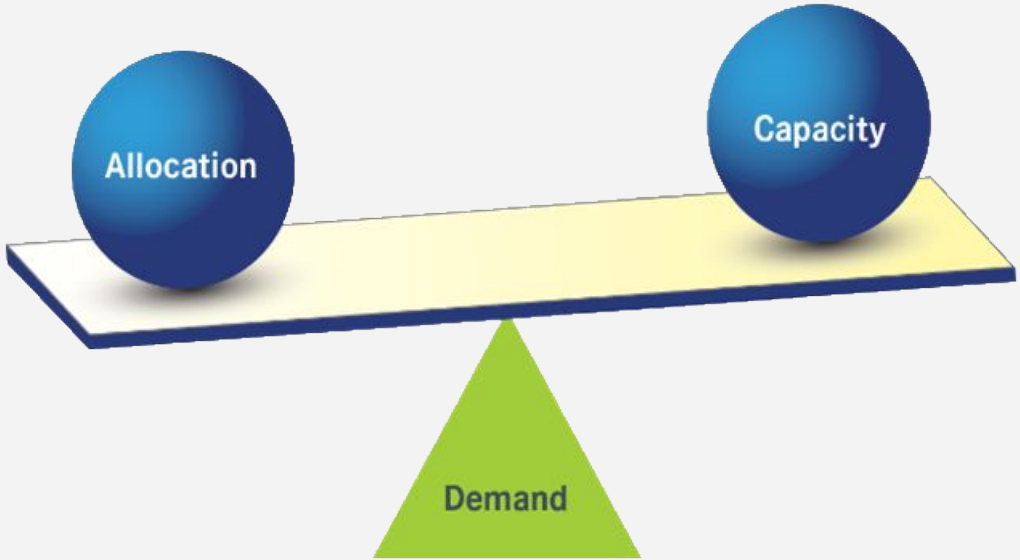
PyAF

TensorFlow

Cesium

Usage

- Capacity planning
- Utilization maximization
- Cost minimization
- Dynamic pricing
- Supply chain management



Inspired by Technology.
Driven by Value.

Taras Firman
email: taras.firman@eleks.com
skype: tarasinho_318
AI&BigData 2017
4 November, Lviv

Have a question? Write to eleksinfo@eleks.com

Find us at eleks.com