دورهٔ آموزشی «علم داده»

Data Science Course

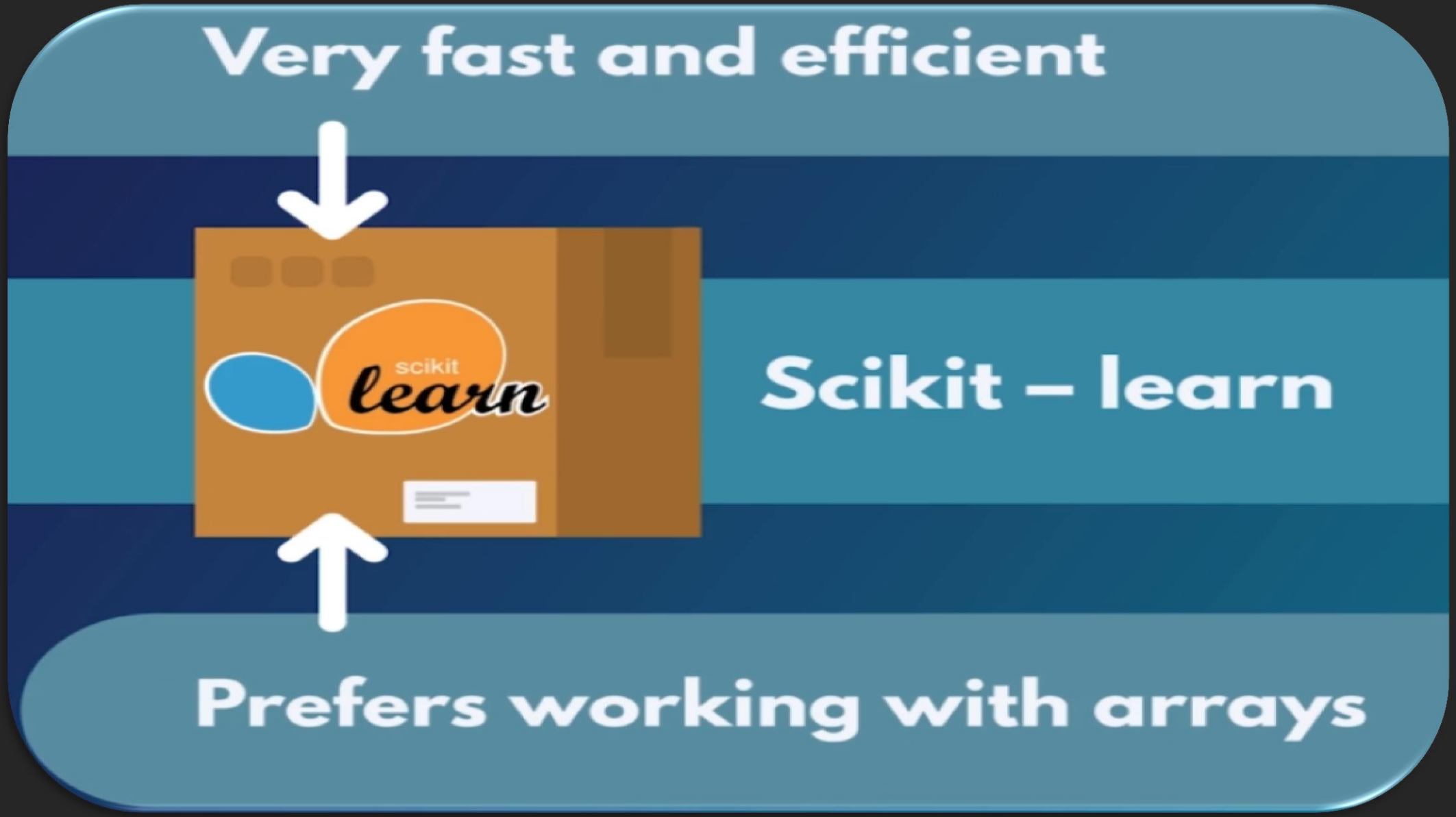
جلسهٔ پانزدهم- (بخش اول)

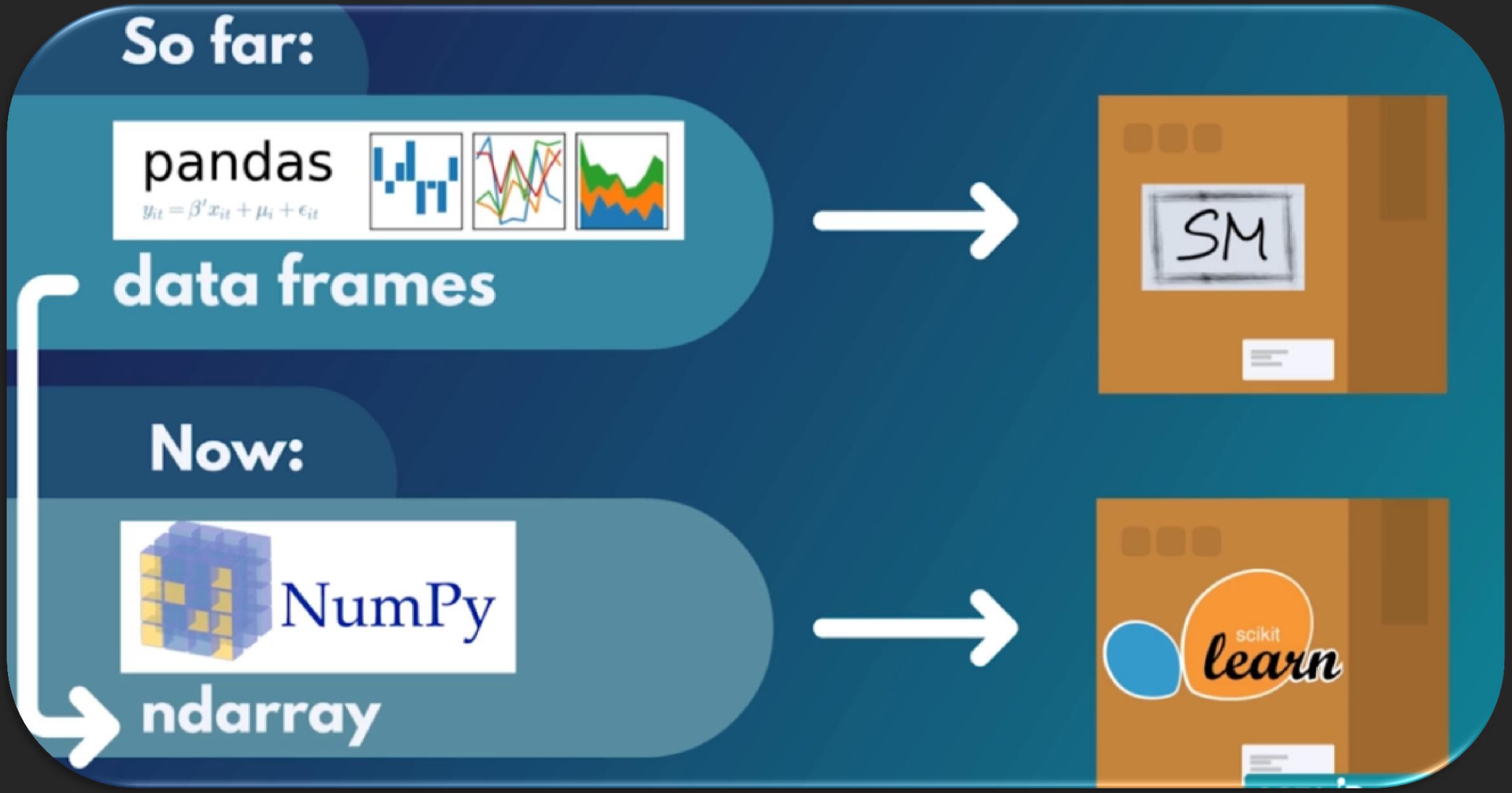
ورود به دنیای سای‌کیت لِرن

Into the realm of Scikit Learn

مدرس: محمد فزونی

عضو هیات علمی دانشگاه گنبدکاووس

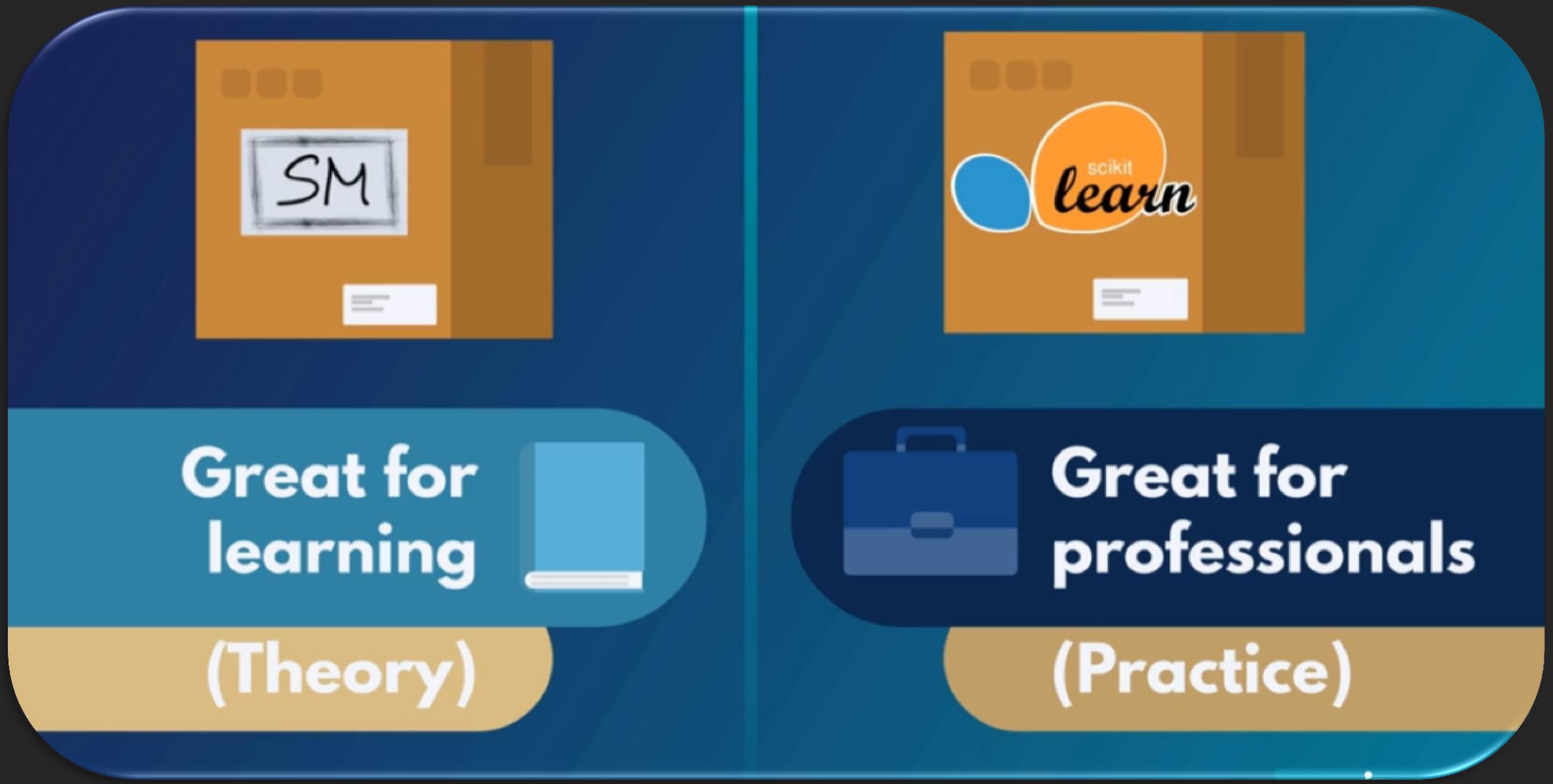

Very fast and efficient
scikit
learn
Scikit – learn
Prefers working with arrays


So far:
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
data frames
SM
Now:
NumPy
ndarray
scikit
learn

Advantages of Scikit - learn

- Incredible documentation
- Variety
  - Regression
  - Classification
  - Clustering
  - Support vector machines
  - Dimensionality reduction

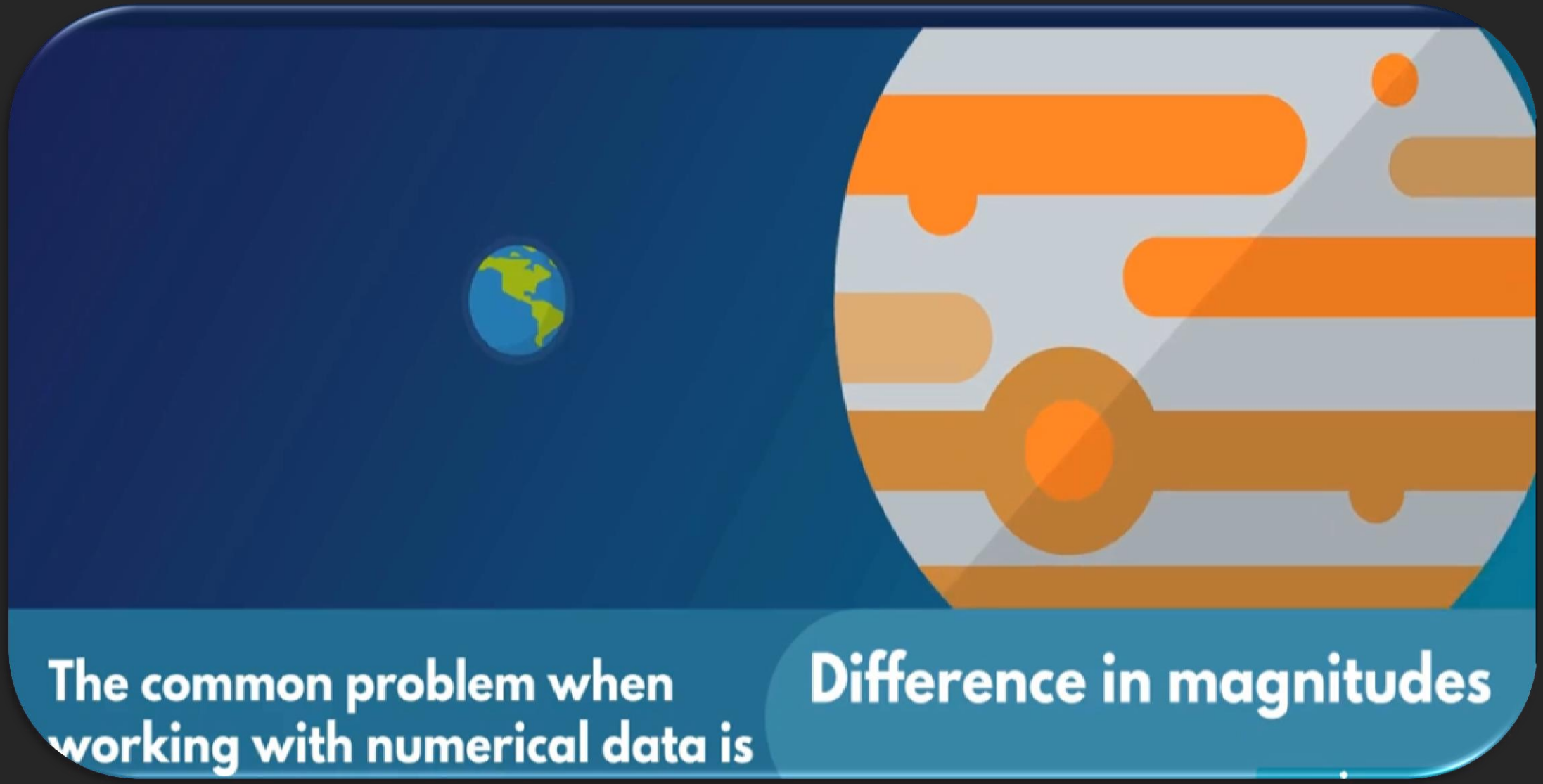
Shortcomings


SM
Great for learning
(Theory)
scikit learn
Great for professionals
(Practice)

- رگرسیون، رگرسیون لجستیک، خوشه‌بندی (آمار پیشرفته)، شبکه‌های عصبی (یادگیری ماشین). در عوض اولی را یادگیری ماشین بگوئیم و دومی رو یادگیری عمیق.

- رگرسیونی که از سای‌کیت لرن مدل میشه اطلاعات کمتری نسبت به نمونه استتز مدل میده اما یکسری قابلیتهایی داره که در ادامه خواهیم فهمید بی‌نظیرند.

- چون سای کیت لرن خودش رو یک پکیج یادگیری ماشین میدونه، برای خیلی از موارد آماری در این پکیج، دستور مستقیم نداریم


The common problem when working with numerical data is
Difference in magnitudes

Standardization

Feature scaling

The process of transforming data into a standard scale

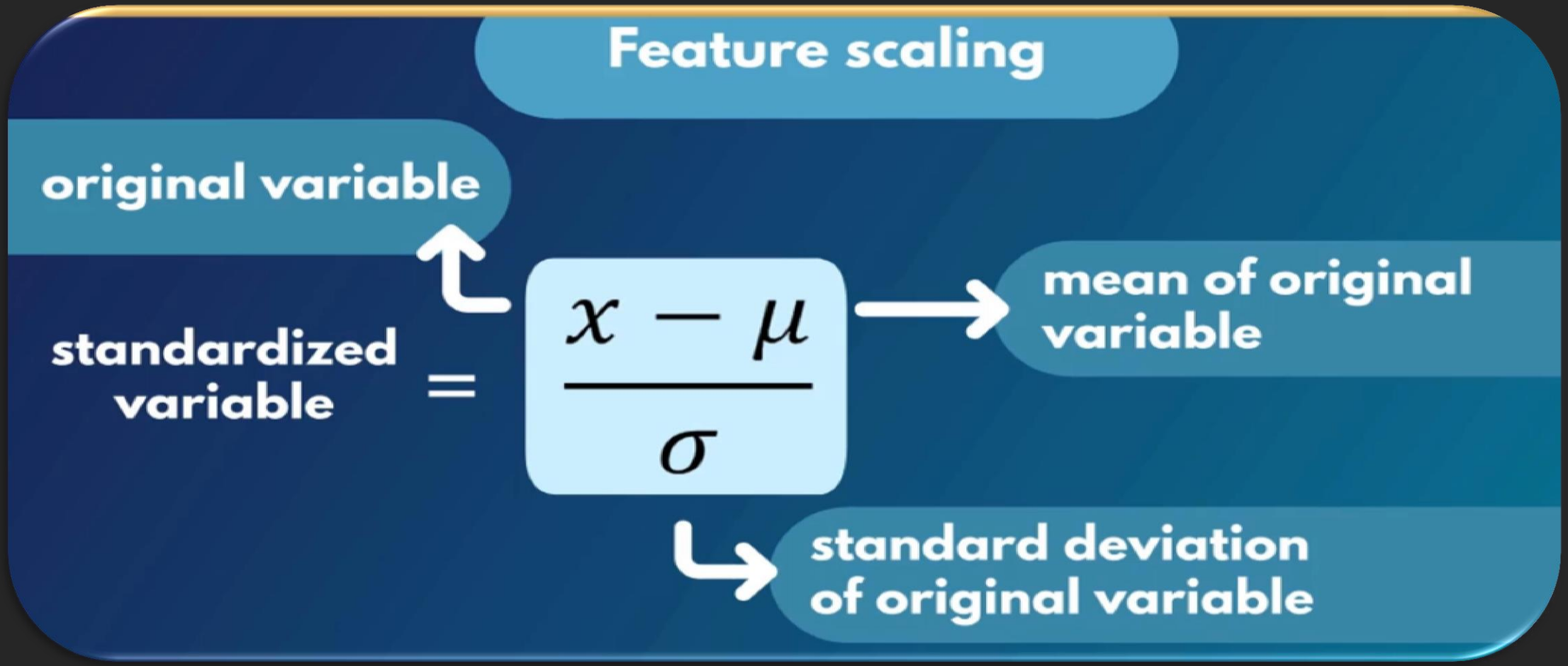
STANDARDIZATION: the process of subtracting the mean and dividing by the standard deviation (a type of normalization)

NORMALIZATION: has different meaning depending on the case; here - we subtract the mean but divide by the L2-norm of the inputs

$$L^2 - \text{Norm}$$

$$\|x\|_2^2 = \sum_{i=1}^{n} |x_i|^2 , \qquad x = (x_1, \dots, x_n) \in \mathbb{R}^n$$

Standardization



StatsModels summary for the same regression

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.407
Model:	OLS	**Adj. R-squared:**	0.392
Method:	Least Squares	**F-statistic:**	27.76
Date:	Thu, 14 Mar 2019	**Prob (F-statistic):**	6.58e-10
Time:	11:36:31	**Log-Likelihood:**	12.720
No. Observations:	84	**AIC:**	-19.44
Df Residuals:	81	**BIC:**	-12.15
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2960	0.417	0.710	0.480	-0.533	1.125
SAT	0.0017	0.000	7.432	0.000	0.001	0.002
Rand 1,2,3	-0.0083	0.027	-0.304	0.762	-0.062	0.046

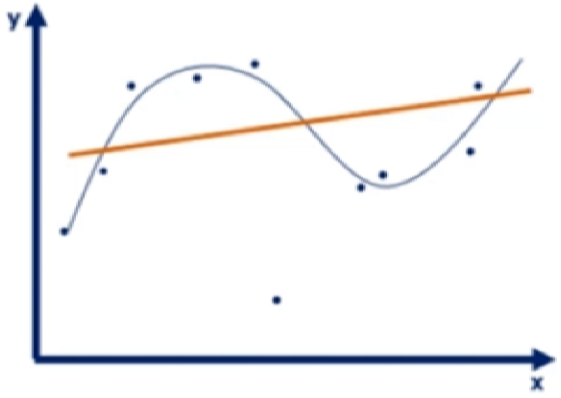
Omnibus:	12.992	**Durbin-Watson:**	0.948
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	16.364

If a variable has a p-value > 0.05, we can disregard it

UNDERFITTING	OVERFITTING
The model has not captured the underlying logic of the data	Our training has focused on the particular training set so much, it has "missed the point"

Underfitting and overfitting

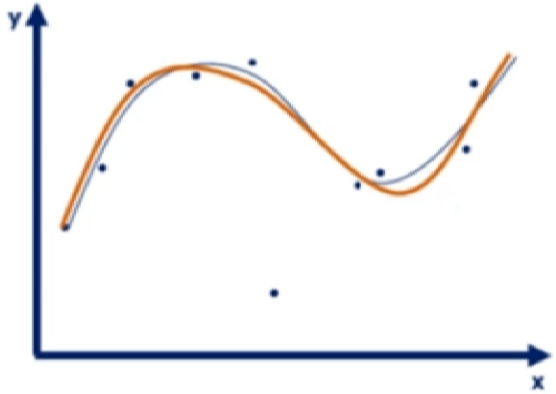
An **underfitted** model



Doesn't capture any logic

- Low train accuracy

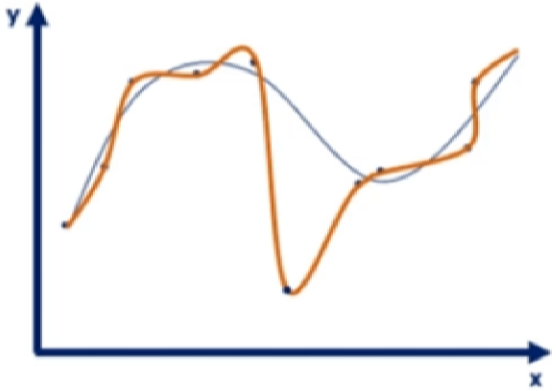
A **good** model



Captures the underlying logic of the dataset

- High train accuracy

An **overfitted** model



Captures all the noise, thus "missed the point"

- High train accuracy

Thanks for watching AMIGOS

Stay in touch via

elmedadeir@gmail.com

Also thanks to Data Science Course 2020 by Udemy and Data Science 365 team. Almost all the slides have been duplicated from this wonderful course.