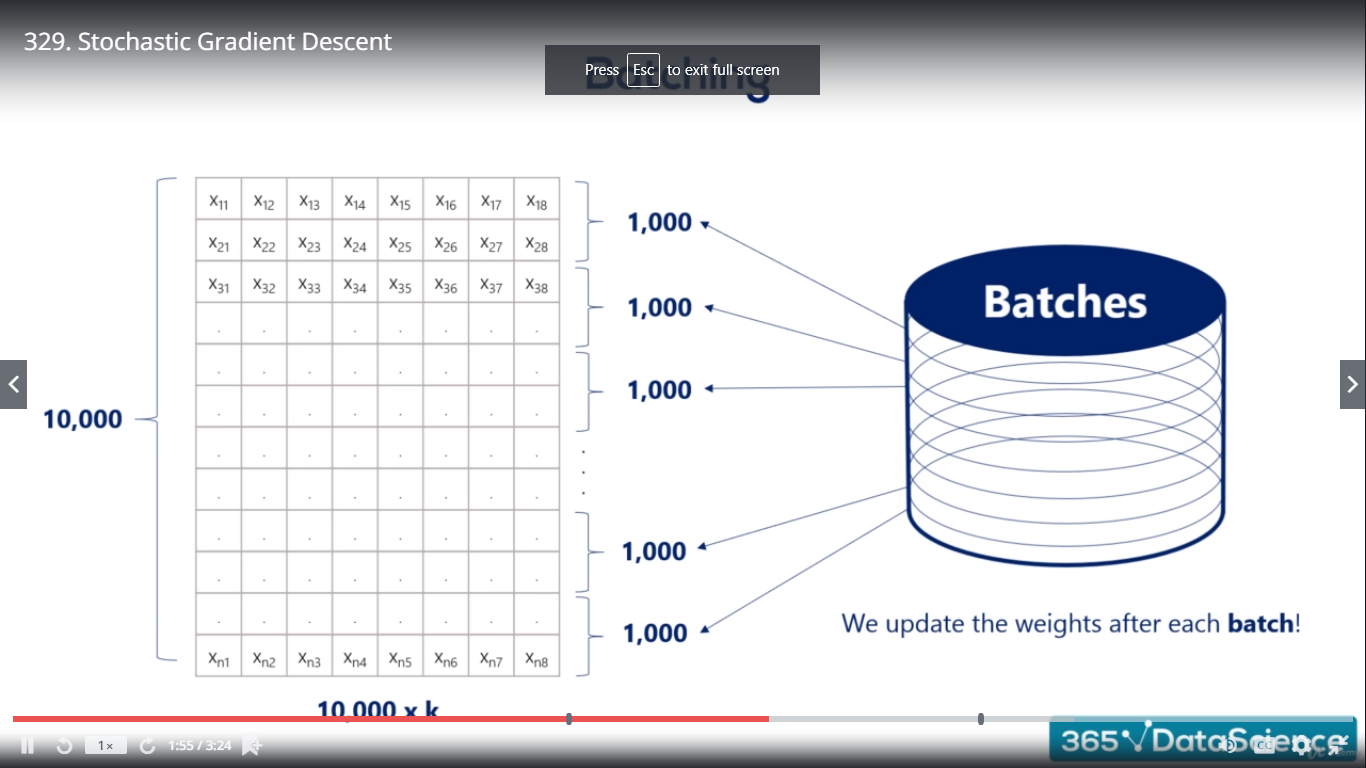
**Gradient Descent**

Stochastic Gradient Descent (SGD):- It works in the same way, but updates the weights many times inside a single epoch.

Batching: the process of splitting the dataset in n batches (mini-batches). We update the weights after each batch instead of every epoch.



CPU processor individually calculates the Batches at the same time so it saves a lot of time.

The SGT comes at a cost: it approximates things a bit

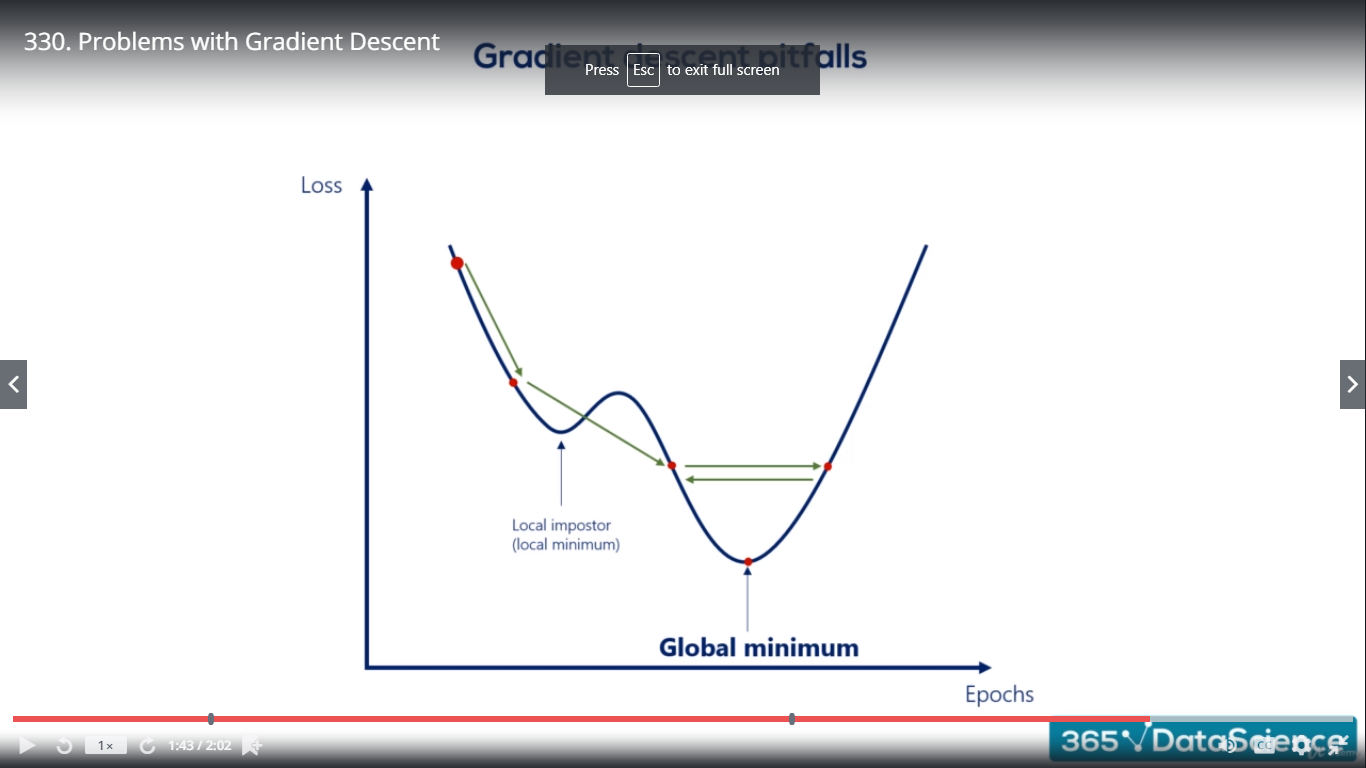
Practitioners refer to the mini-batch GT as SGD

**Gradient Descent Pitfalls**

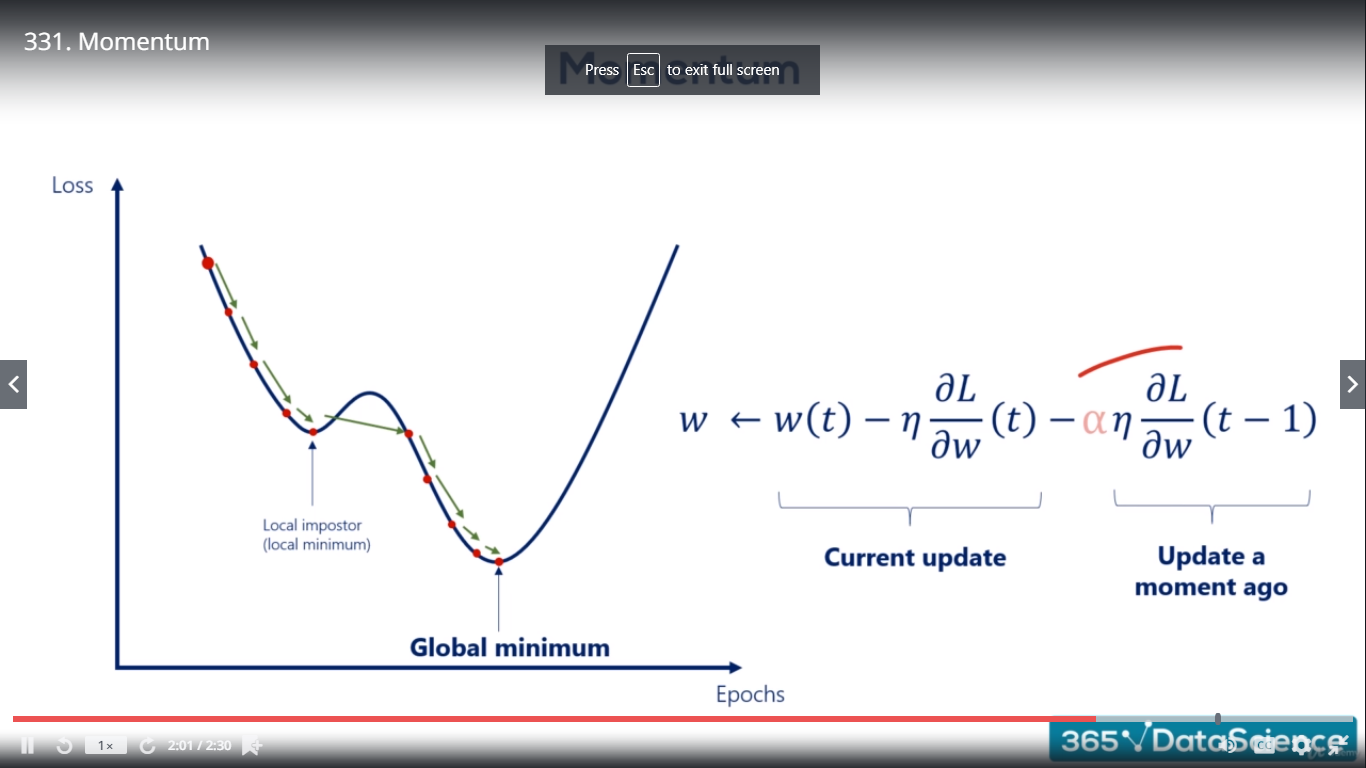
A single batch GD would be slow, but will eventually reach the minimum.

SGD would be much faster, but give us approximate answer.

Each local minimum is a suboptimal solution to the optimization problem.



**Momentum**



Α = 0.09 is conventional

**Learning Rate (ŋ) 🡪**

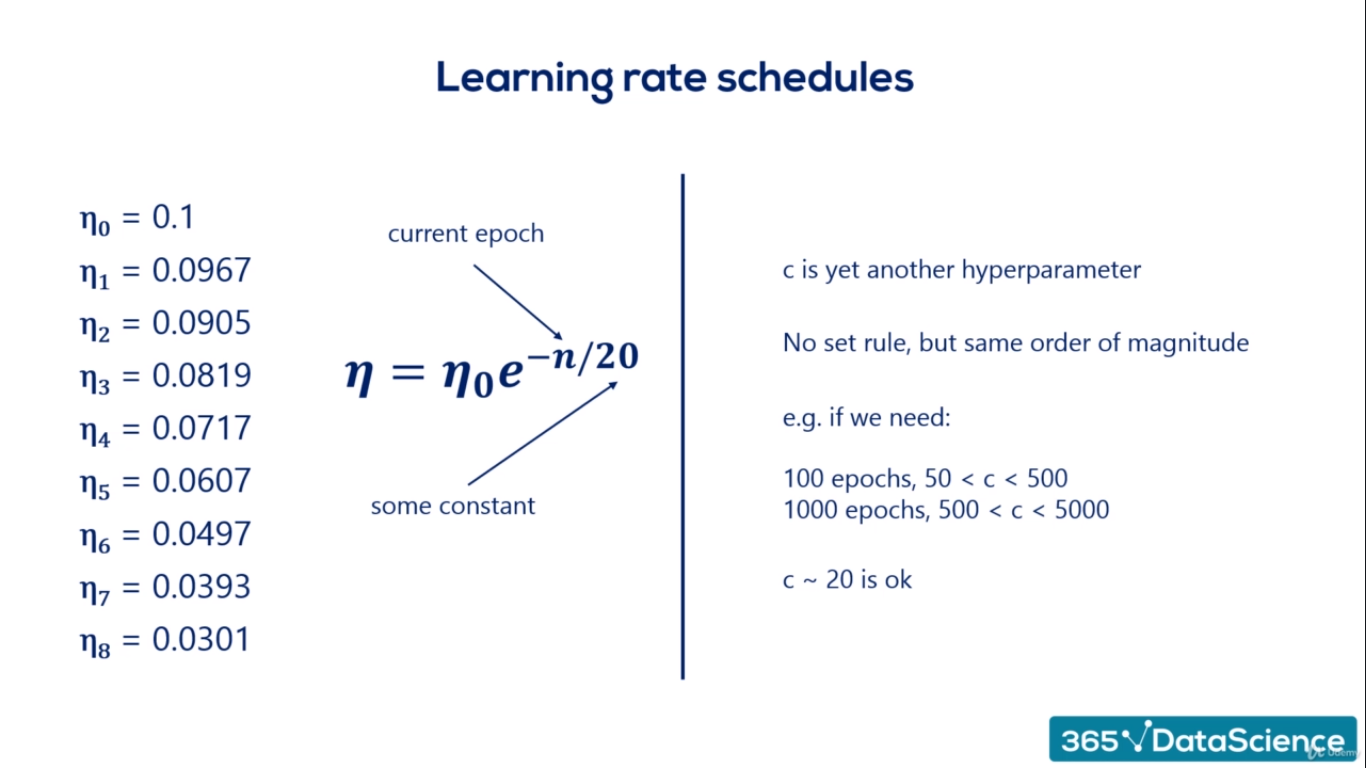
**small enough** so we gently descent, instead of oscillating or diverging

**Big enough** so we reach it in a rational amount of time

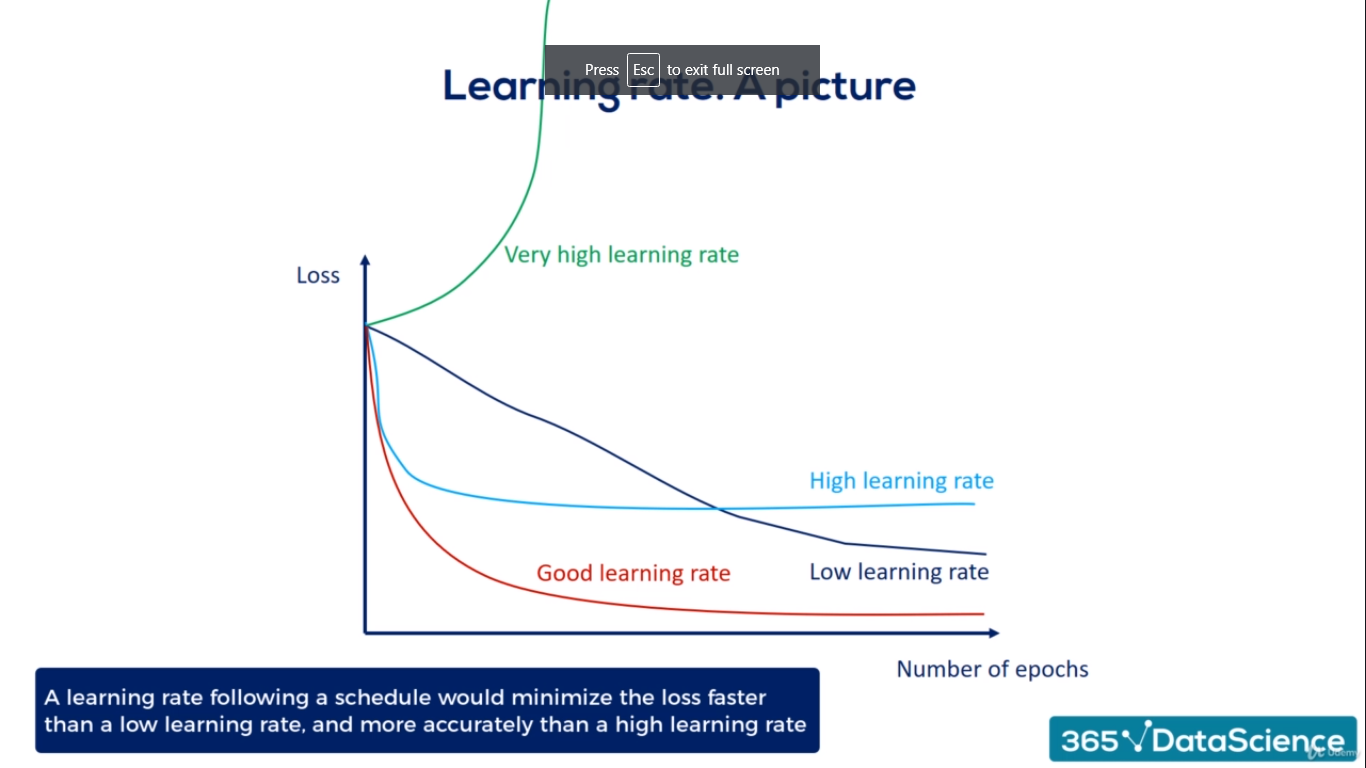
**Learning Rate Schedules: Small enough or Big enough**

|  |  |
| --- | --- |
| 1. We start from a high initial learning rate | First 5 epochs (ŋ = 0.1) |
| 1. At some point we lower the rate to avoid oscillation | Next 5 epochs (ŋ = 0.01) |
| 1. Around the end we pick a very small rate to get a precise answer | until the end (ŋ = 0.001) |

**Exponential Learning Rate:**



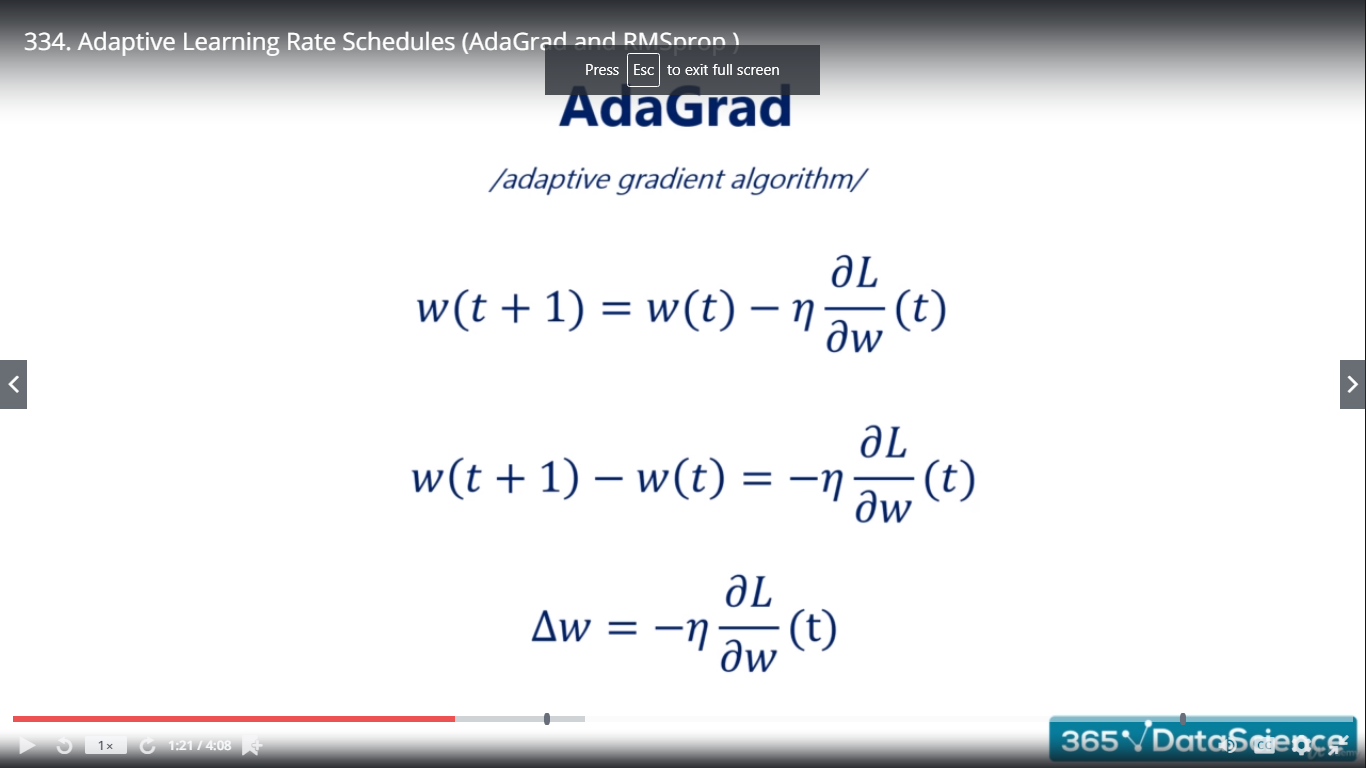
|  |  |
| --- | --- |
| **Hyperparameters (pre-set by us)** | **Parameters (found by optimizing)** |
| Width  Depth  Learning Rate (ŋ)  Batch Size  Momentum Coefficient (α)  Decay coefficient (c) | Weights (w)  Biases (b) |



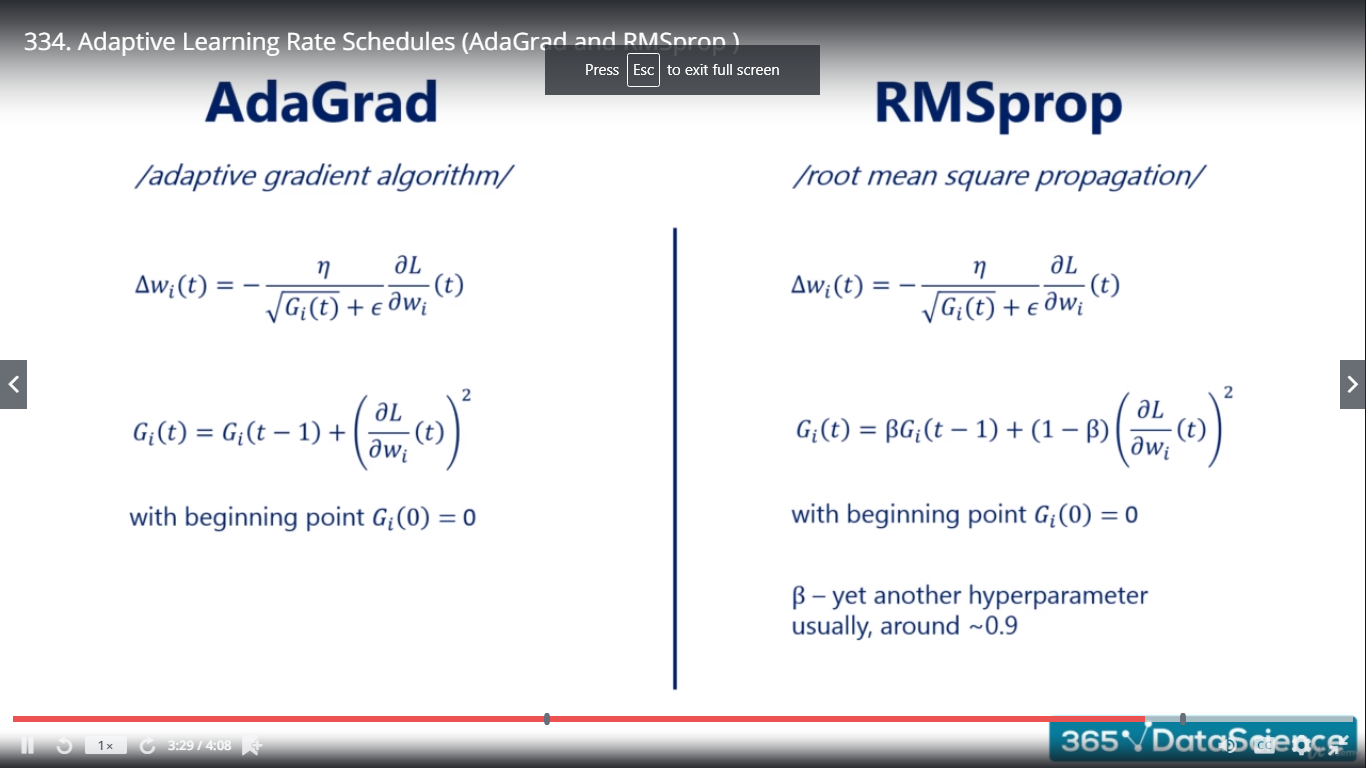
* A low learning rate would minimize the loss but quite slowly
* A high learning rate would minimize the loss faster but only to a certain extent then it starts oscillating
* A very high learning rate would not even minimize the loss
* A learning rate following a schedule would minimize the loss faster than a low learning rate, and more accurately than a high learning rate.
* High learning rate may not minimize the loss at all.
* Low learning rate eventually converges with the good learning rate

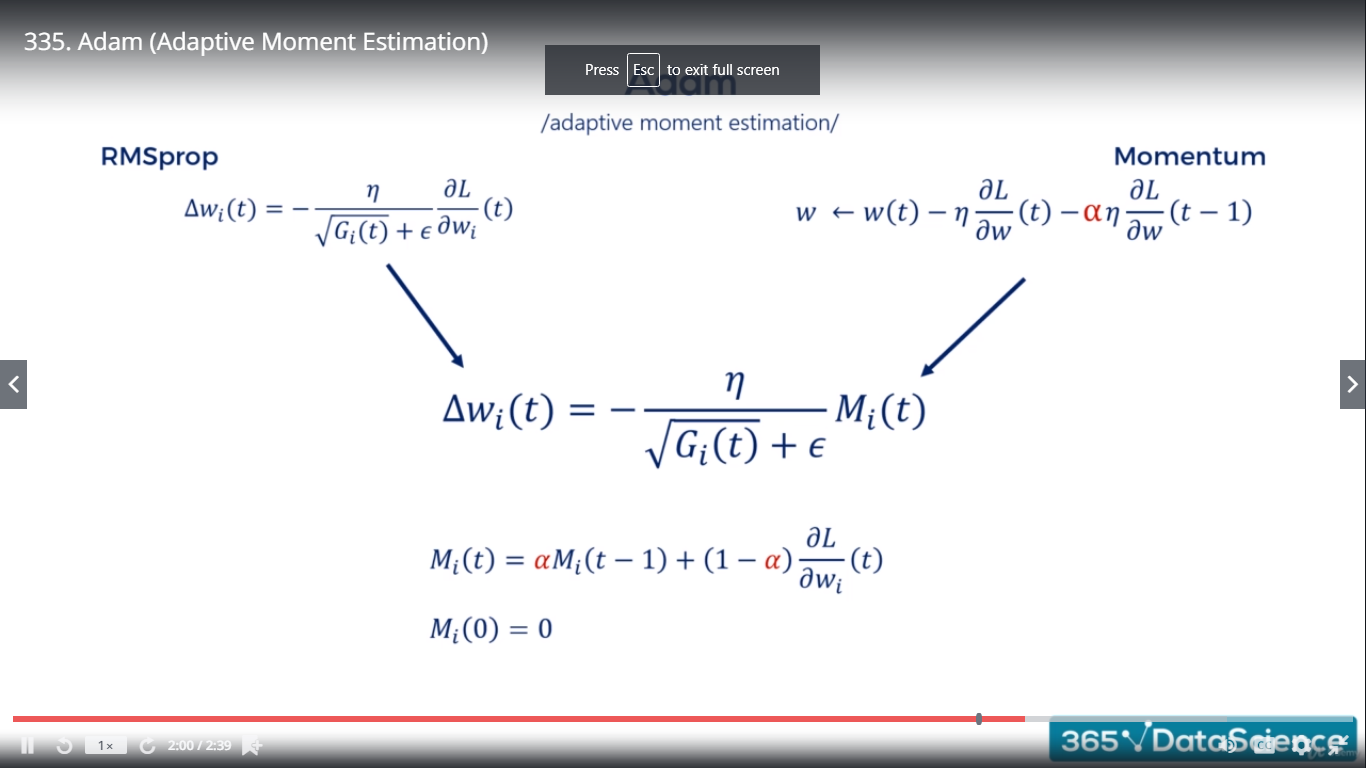
**Advanced Learning Rates Schedules**

**AdaGrad (Adaptive Gradient Algorithm) - 2011**: It dynamically varies the learning rate at each update and for each weight individually.









As with all science, data science is a long chain of academic research building on the top of each other.

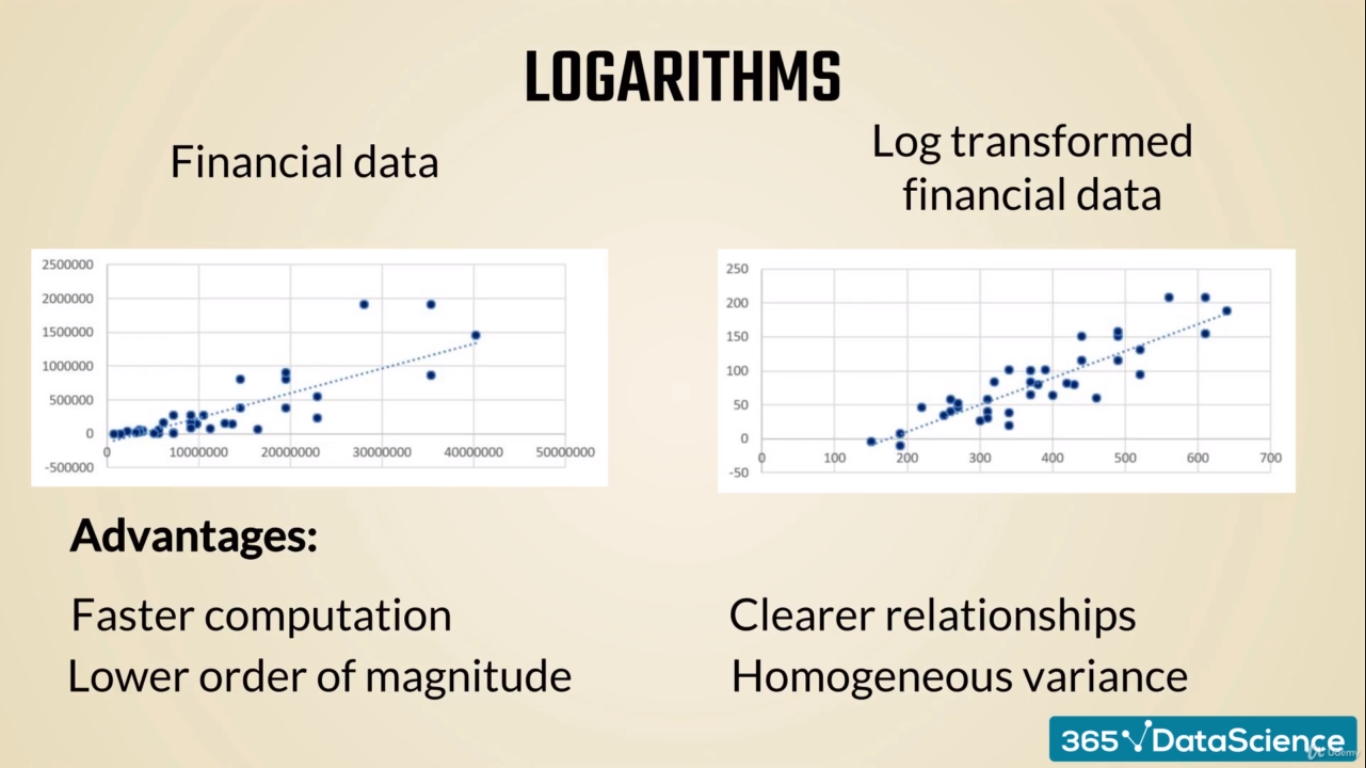
**Preprocessing**

Any manipulation of the dataset before running it through the model.

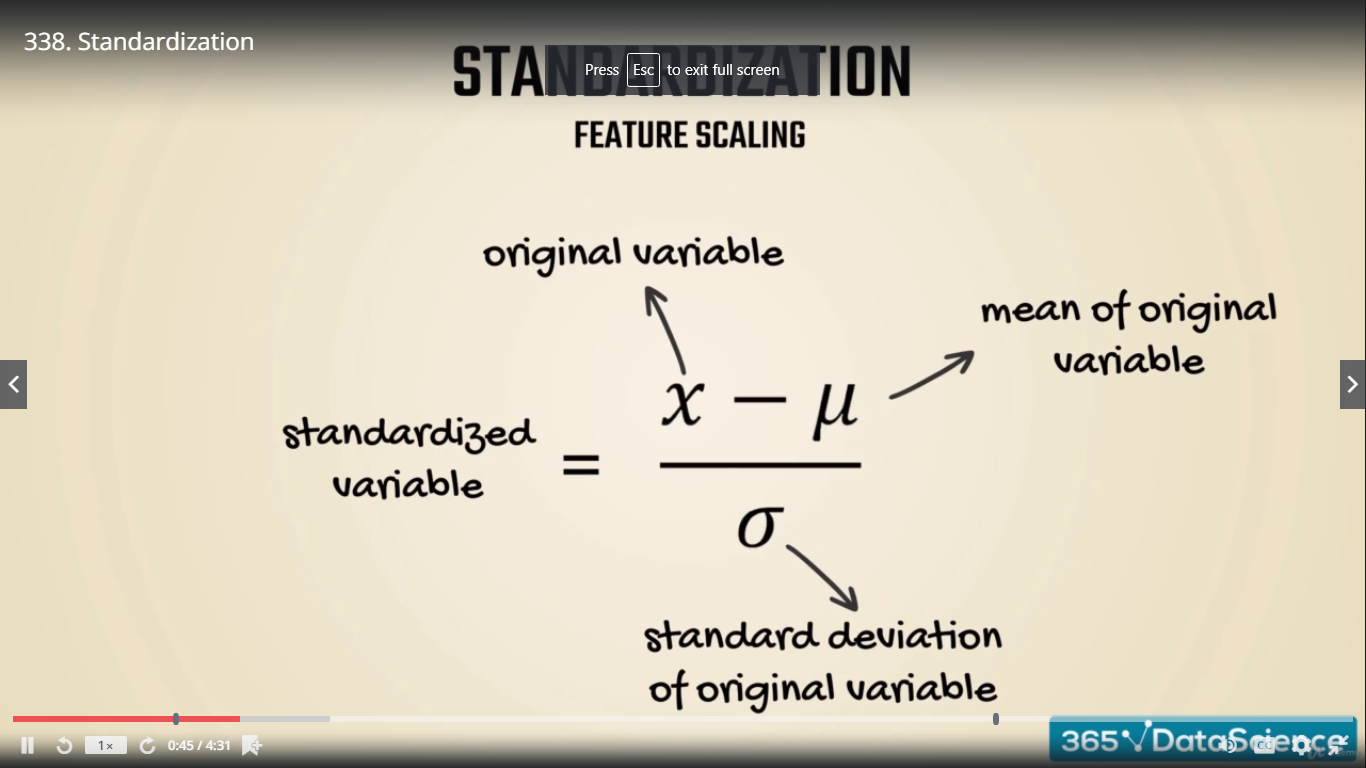
Motivation:-

* Compatibility (Tensorflow)
* Orders of magnitude
* Generalization

**Types of preprocessing**



**Standardization (Feature Scaling)** – The process of transforming data into a standard scale

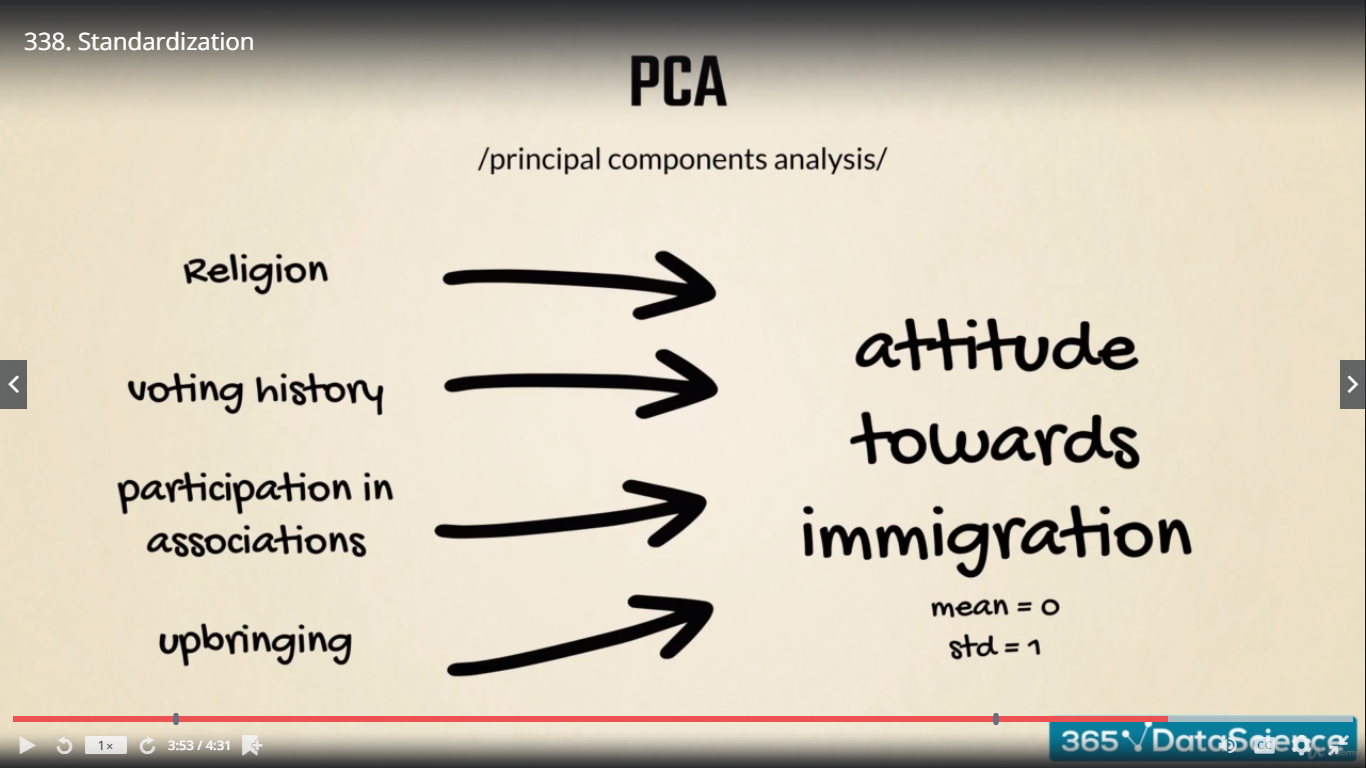




**Other Methods:**

**Normalization:** converts each sample to unit length vector using L1 or L2 norms

**PCA (Principal Components Analysis)-** Dimension reduction technique used to combine several variables into a bigger (latent) variable



**Whitening:** after performed PCA

Attitude towards immigration 🡺 Uncorrelated ‘attitude towards immigration’

**Preprocessing –**

**Categorical Data**

**Categories or group (non-numerical data)**

**How to encode categories in a way useful for ML**

1. One hot encoding (Few categories)
2. Binary encoding (Many categories)



