# Mathematics for Machine Learning (70015)

## Lecture 1.1 - Linear Regression Intro

Linear regression aims to provide a solution to the supervised learning problem; we are given a dataset of $N$ examples of inputs and expected outputs, with a goal of predicting the correct output for a new input. Examples of this include image classification (such as digit classification) and translation. A curve fitting problem in 1-dimension has an input space $\in \mathbb{R}$, and an output space $\in \mathbb{R}$.

To tackle this problem mathematically, we need to first describe the curve fitting problem mathematically. As each input is associated with a single output, this is equivalent to a function in mathematics. We are given a dataset of $N$ pairs, of inputs and outputs, where $\boldsymbol{x_n} \in \mathcal{X}$, which is usually $\mathbb{R}^D$ and $y_n \in \mathcal{Y}$ (usually $\mathbb{R}$ in this case); $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^N$. The goal is to find a function that maps from the input space to the output space **well**; $f : \mathcal{X} \to \mathcal{Y}$.

We need to first find candidates for functions that can perform the predictions. Functions need to be parameterised, such that some numbers $\boldsymbol{\theta}$ map to a function. From here, we need to pick the 'best' function, thus requiring us to define what good and bad functions are. A good function has the property $f(\boldsymbol{x_i}, \boldsymbol{\theta^*}) \approx y_i$; the ddoutput of the function closely matches the outputs of the training points. This can be defined with a **loss function**, for example;

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y_i - f(\boldsymbol{x_i}, \boldsymbol{\theta}))^2$$

Therefore, a good function is chosen by minimising the loss; $\boldsymbol{\theta^*} = \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$