PSI:
Predictive Models

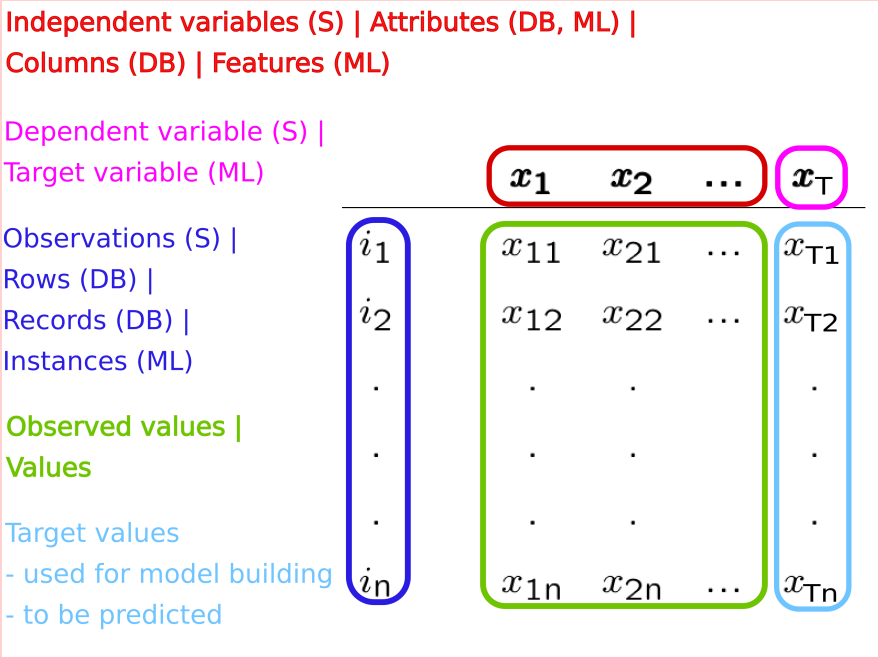
Technological University Dublin City Campus

School of Computer Science

Data

The data involved in predictive models:

- **dependent variables**: this is what is being 'predicted' (we only cover models with one dependent variable)
- **one or more independent variables**: these are used as the basis for the prediction



Predictive model types

By what it is being predicted, a model can be a **regression** or **classification**.

model type	**dependent (predicted) variable type**
regression	numeric
classification	categorical

The meaning of prediction

In the context of predictive data models the word **prediction**:

- is not used in the sense of forecasting
- means deriving values of the dependent variable from values of the independent variable(s)

Note that this could mean 'predicting' values that temporally preceded their predictors.

Model fitting

- A model is fitted (built) using known independent and dependent variable values.
- A fitted model is used for the derivation of unknown dependent variable values from known (but yet unseen by the model) independent variable values.
- Model fitting involves:
  - identifying the best set of predictor variables from a set of independent variables
  - creating a mechanism (function or algorithm) for the derivation of dependent variable values

EXAMPLE: Independent and dependent variables of prediction

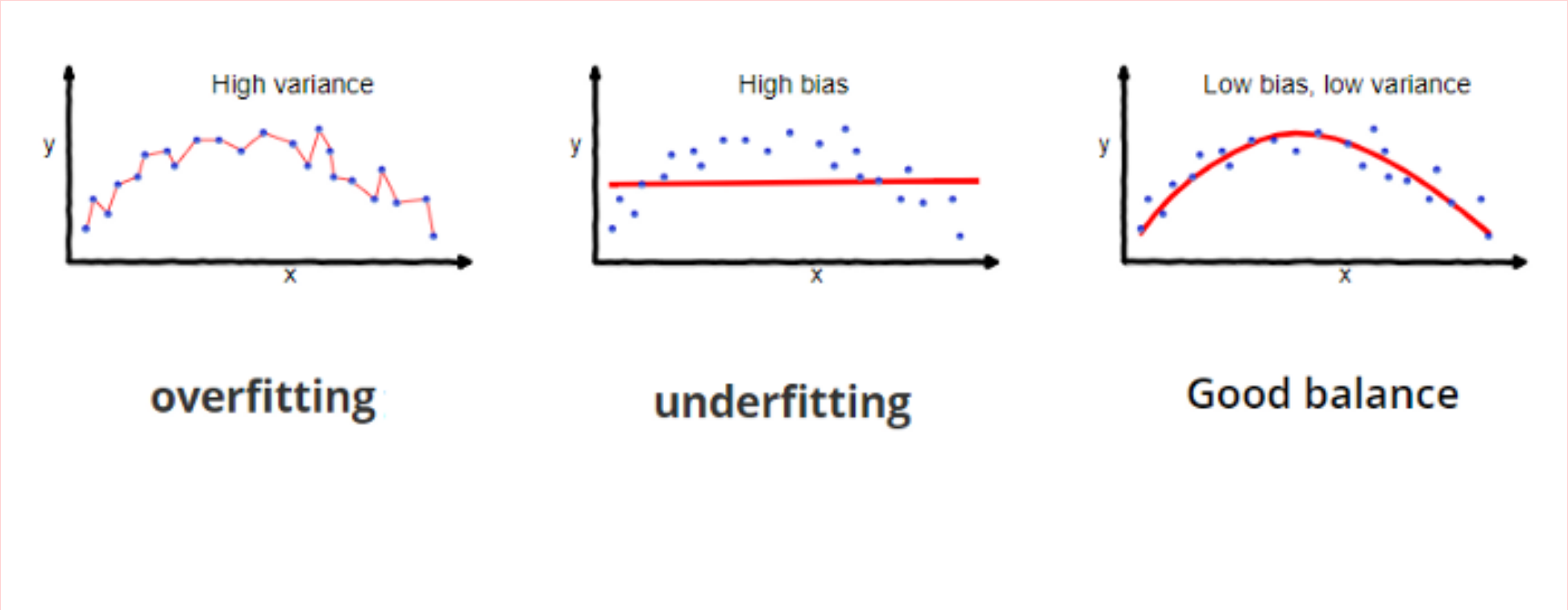
We want to see what the house prices in the country would be under certain conditions. We have historical data for various economic and population descriptors (e.g. population in different age groups, employment rates, average salary etc.) - these are the independent variables. We also have historical data for house prices - this is the dependent variable. All these data are used to fit a predictive model. Then some hypothetical scenario consisting of particular values for population, employment rate, average salary etc. is fed into the model and the model outputs a predicted average house price.

Overfitting and underfitting

- *Overfitting* is the term used for aproaches or instances of model fitting that cause the resulting model to reflect particular properties specific to the training data sample, in addition or opposition to general properties, pertinent to the entire population.
  - An overfitted model performs badly when predicting with yet unseen independent variable values. It performs very well with independent values used for fitting.
  - Overfitting can be avoided by using *holdout* data to test that the accurracy of a model when used on new data matches, or is close to, the accuracy of the model when used on training data.
  - An overfitted model is also referred to as a **high-variance** model.

- *Underfitting* is the term used for a model that does not represent the population well enough to make reasonable predictions.
  - An underfitted model does not predict well with any kind of independent variable values.
  - Underfitting is usually a result of insufficient quality data or failure to utilise available data correctly.
  - An underfitted model is also referred to as a **high-bias** model.

[Overfitting and underfitting]



Source: towardsdatascience.com