

Machine Learning Zoomcamp
Session #1.5

Modelling step: Model Selection

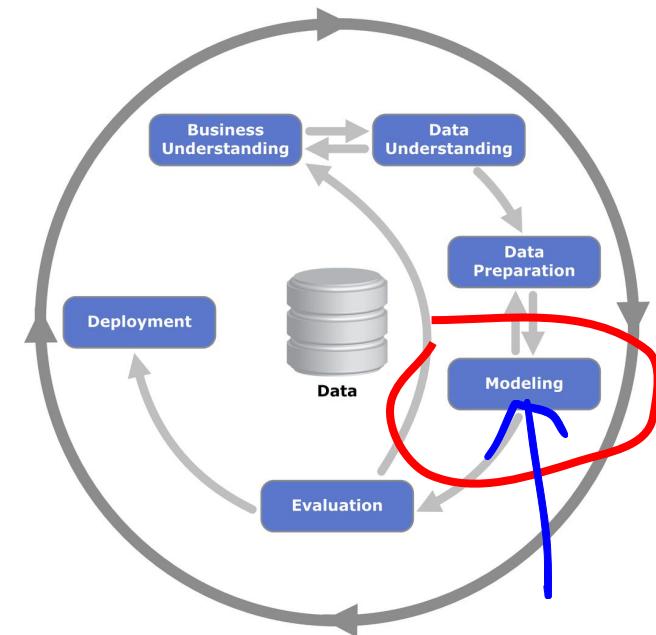
Session #1.5: Plan

- Process for selecting the best model
- Evaluating the model
- Multiple comparison problem

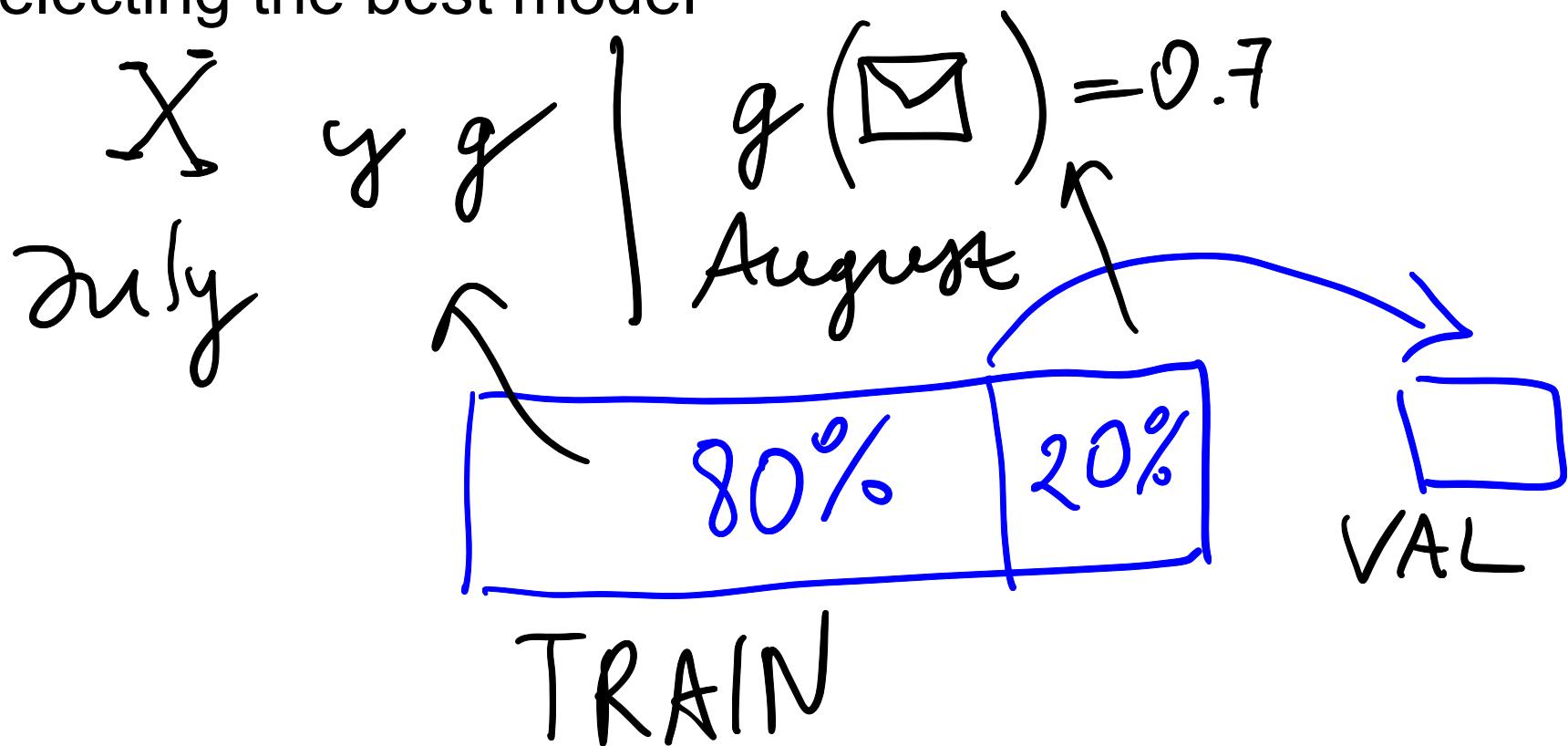
Modeling

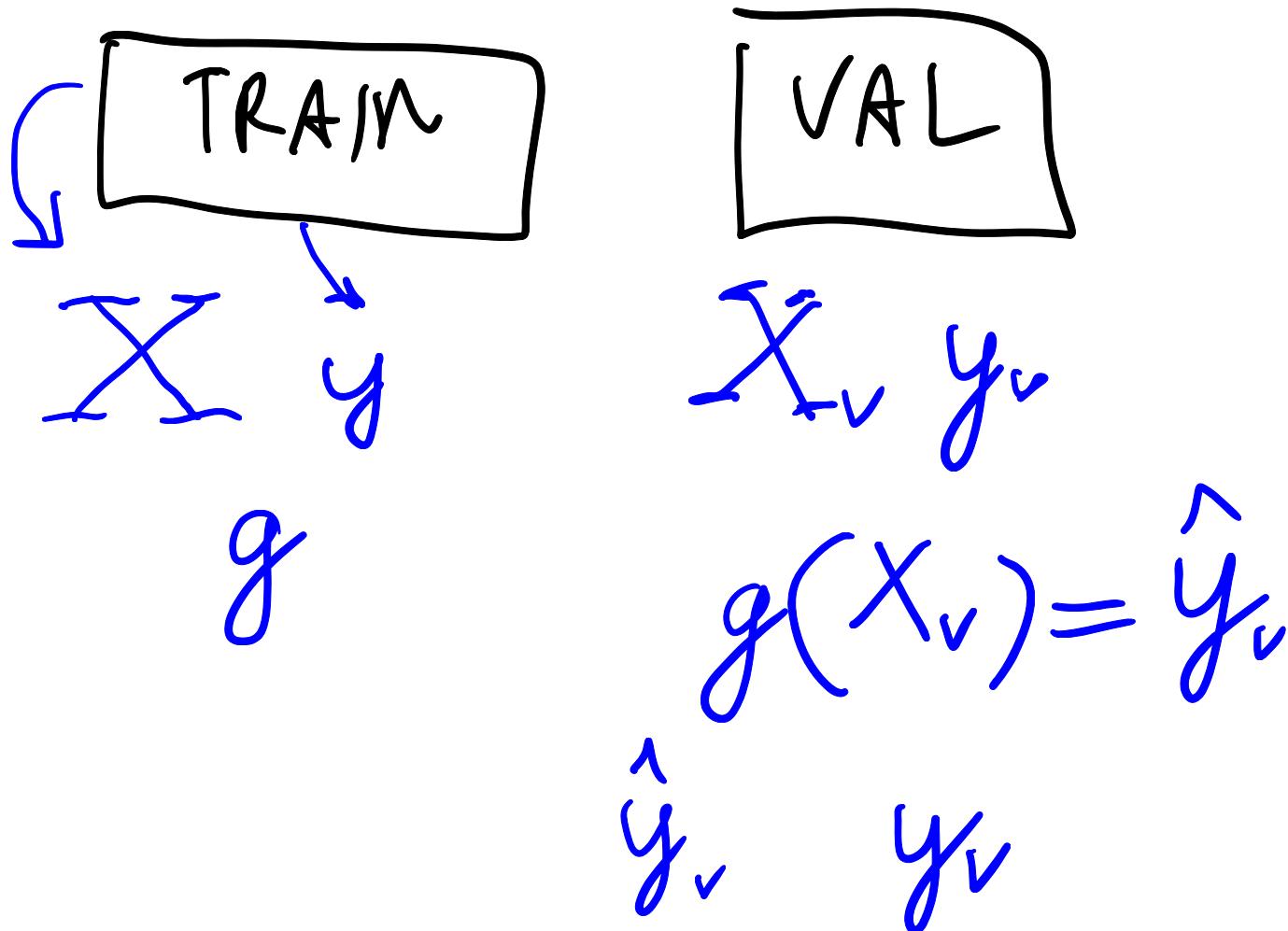
Which model to choose?

- Logistic regression
- Decision tree
- Neural network
- Or many others



Selecting the best model





$$\hat{y}_v - y_v$$
$$0.8 \quad \frac{1}{1} - \frac{1}{0}$$
$$0.7 \quad \frac{1}{1} \times \frac{1}{0}$$
$$0.6 \quad \frac{1}{1} - \frac{1}{1}$$
$$0.1 \quad \frac{0}{0} - \frac{0}{0}$$
$$0.9 \quad \frac{1}{1} - \frac{1}{0}$$
$$0.6 \quad \frac{1}{1} \times \frac{1}{0}$$

pred target

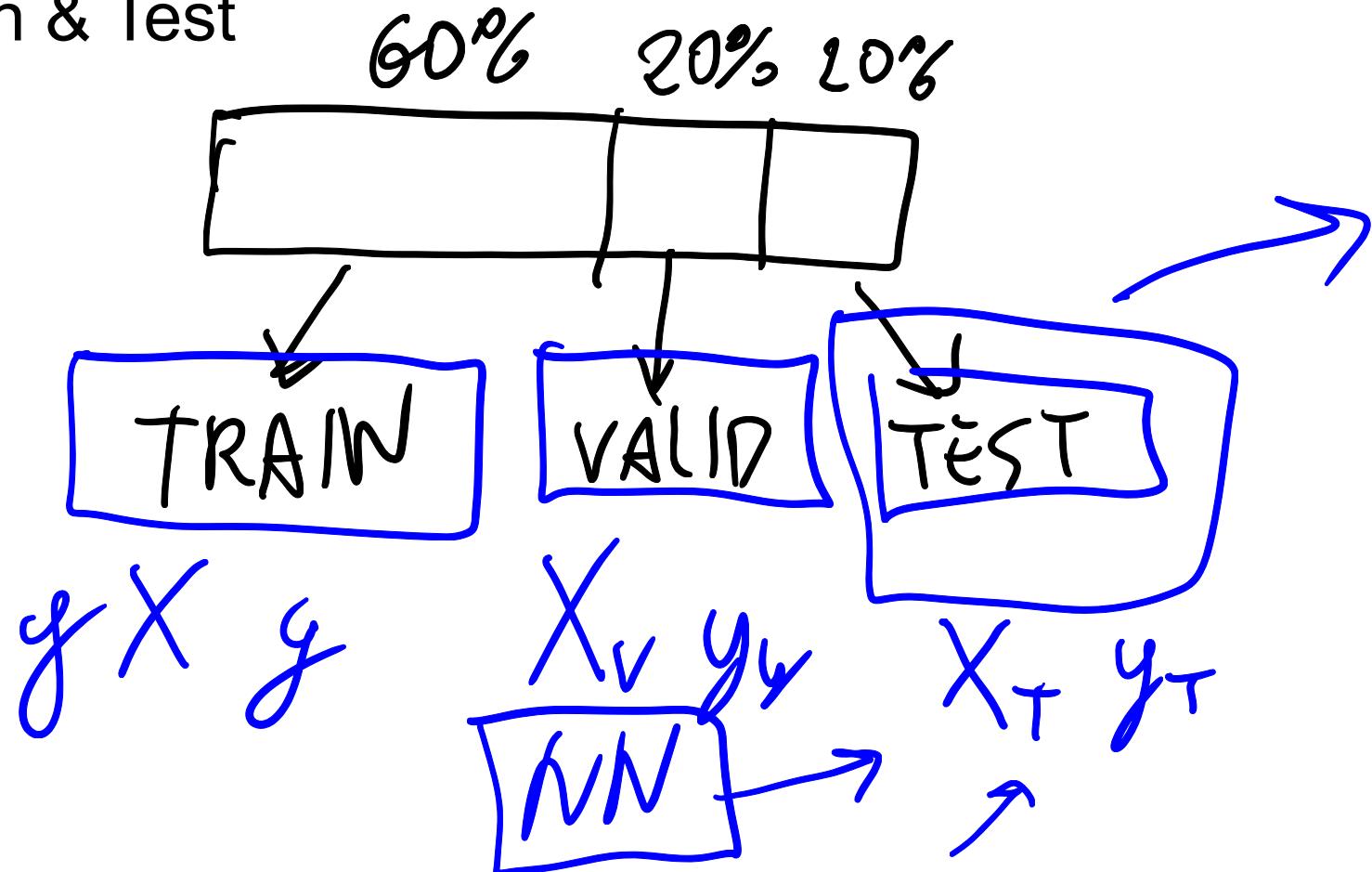
$$4/6 = \underline{\underline{66\%}}$$

g_1	LR	66 %
g_2	DT	60 %
	RT	67 %
NN	80 %	V

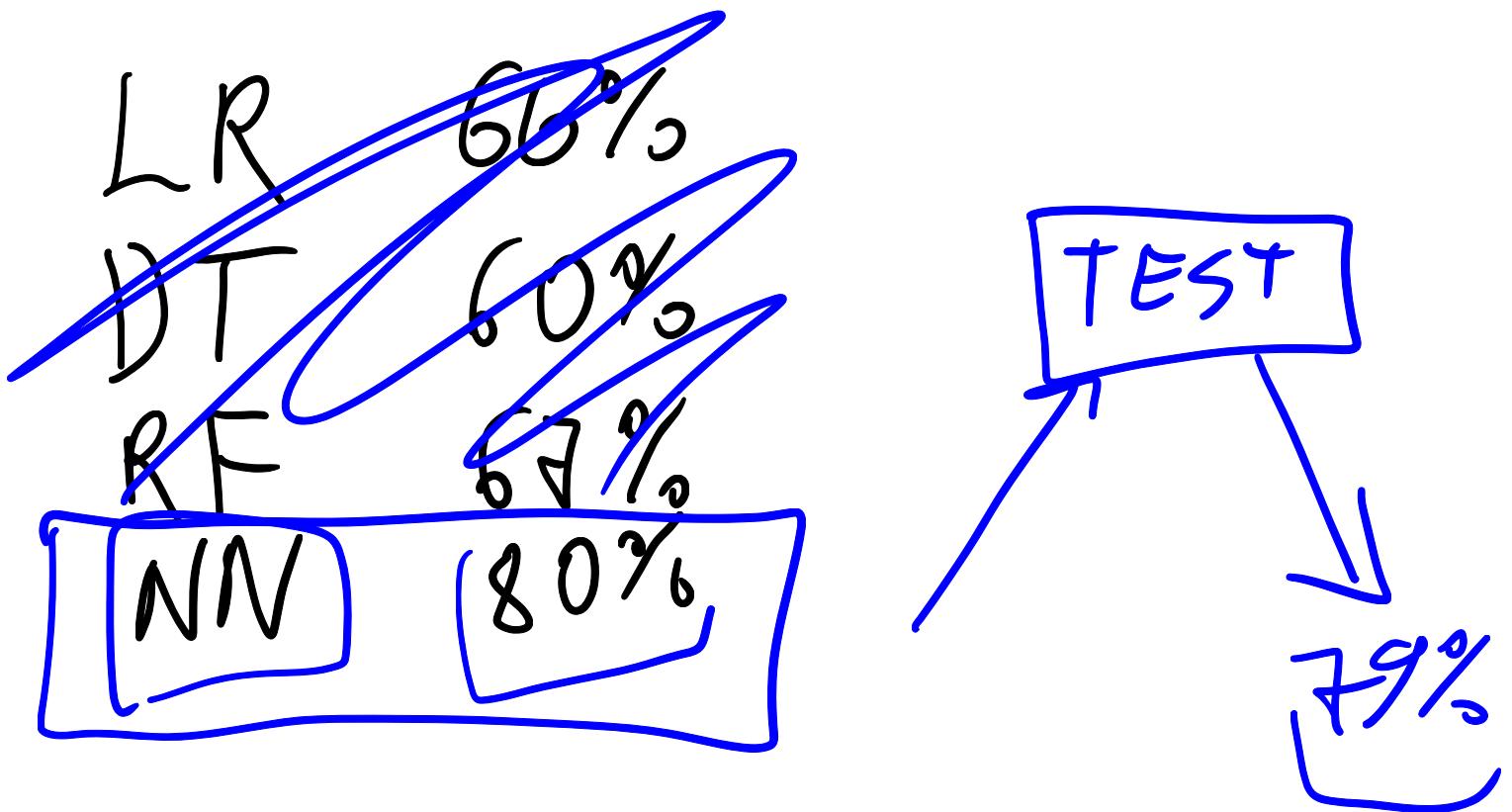
Multiple comparisons problem



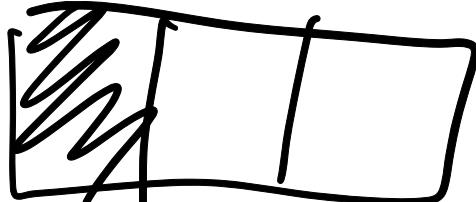
Validation & Test



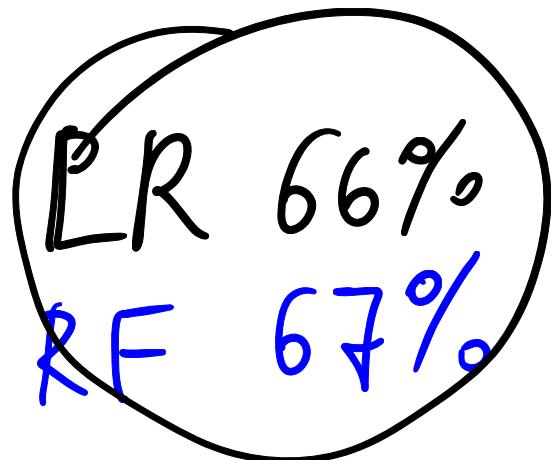
Train + validation + test



TRAIN TEST



← ① OSPLIT



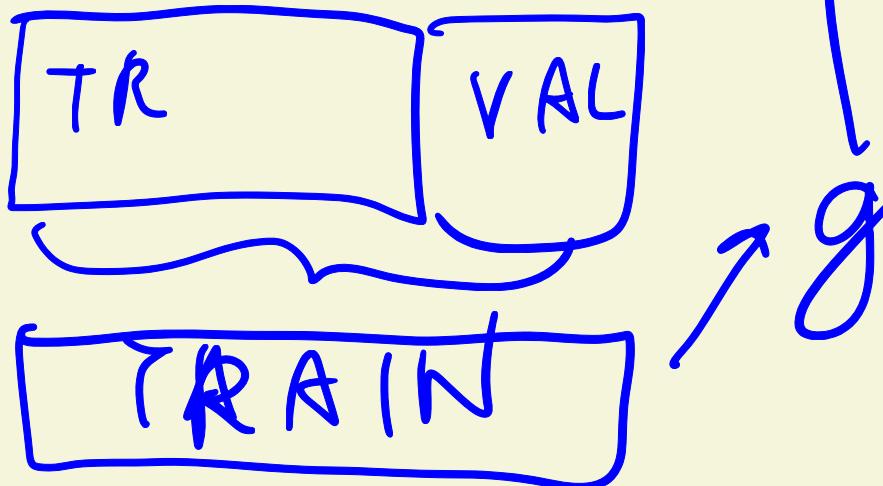
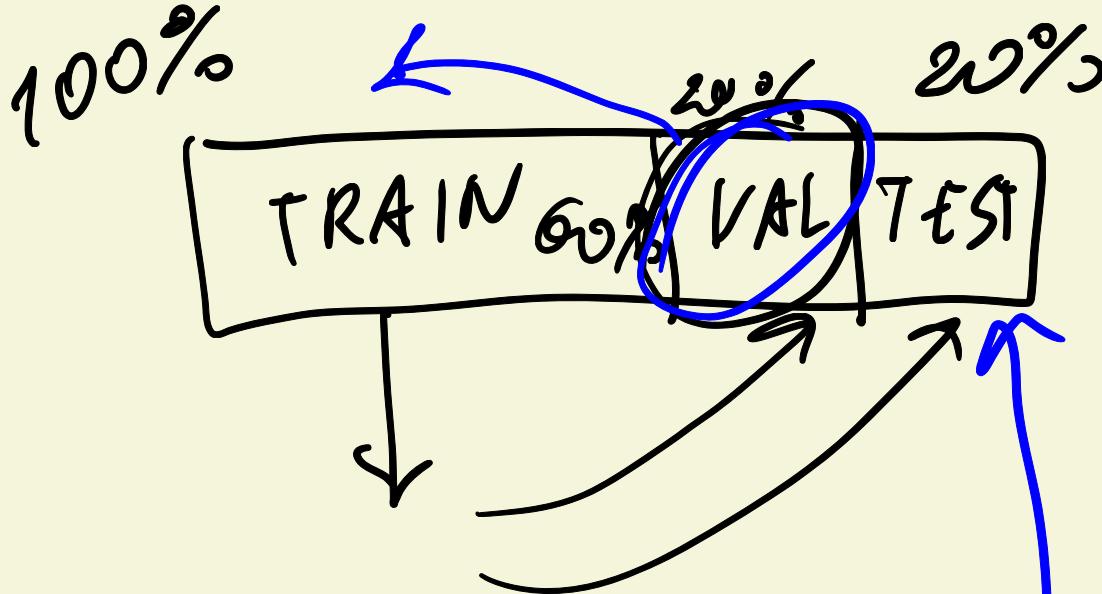
② TRAIN

③ VAL

④ SELECT THE REST

⑤ TEST
⑥ CHECK

||



Next

Introduction to NumPy

