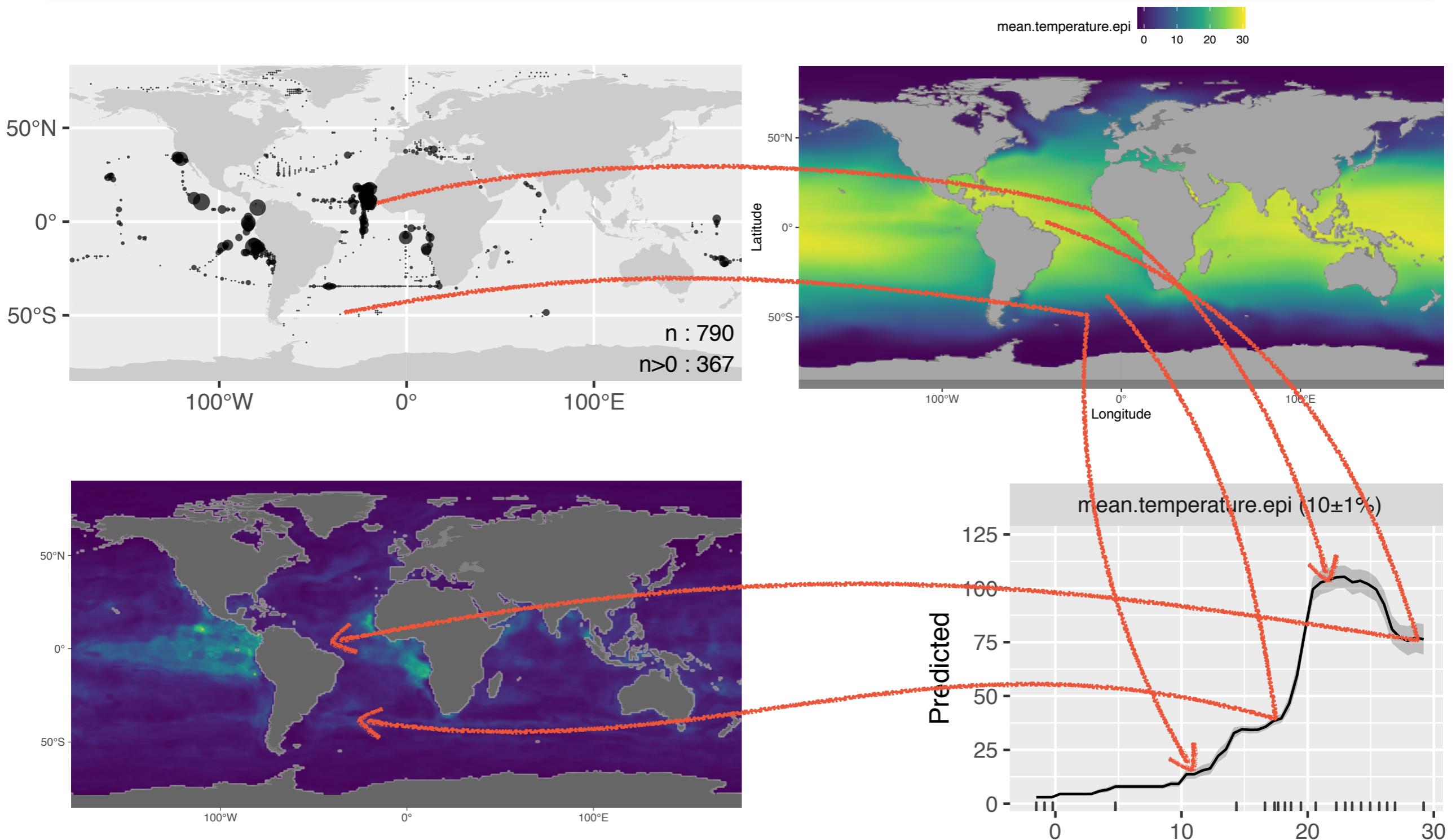


Habitat modelling

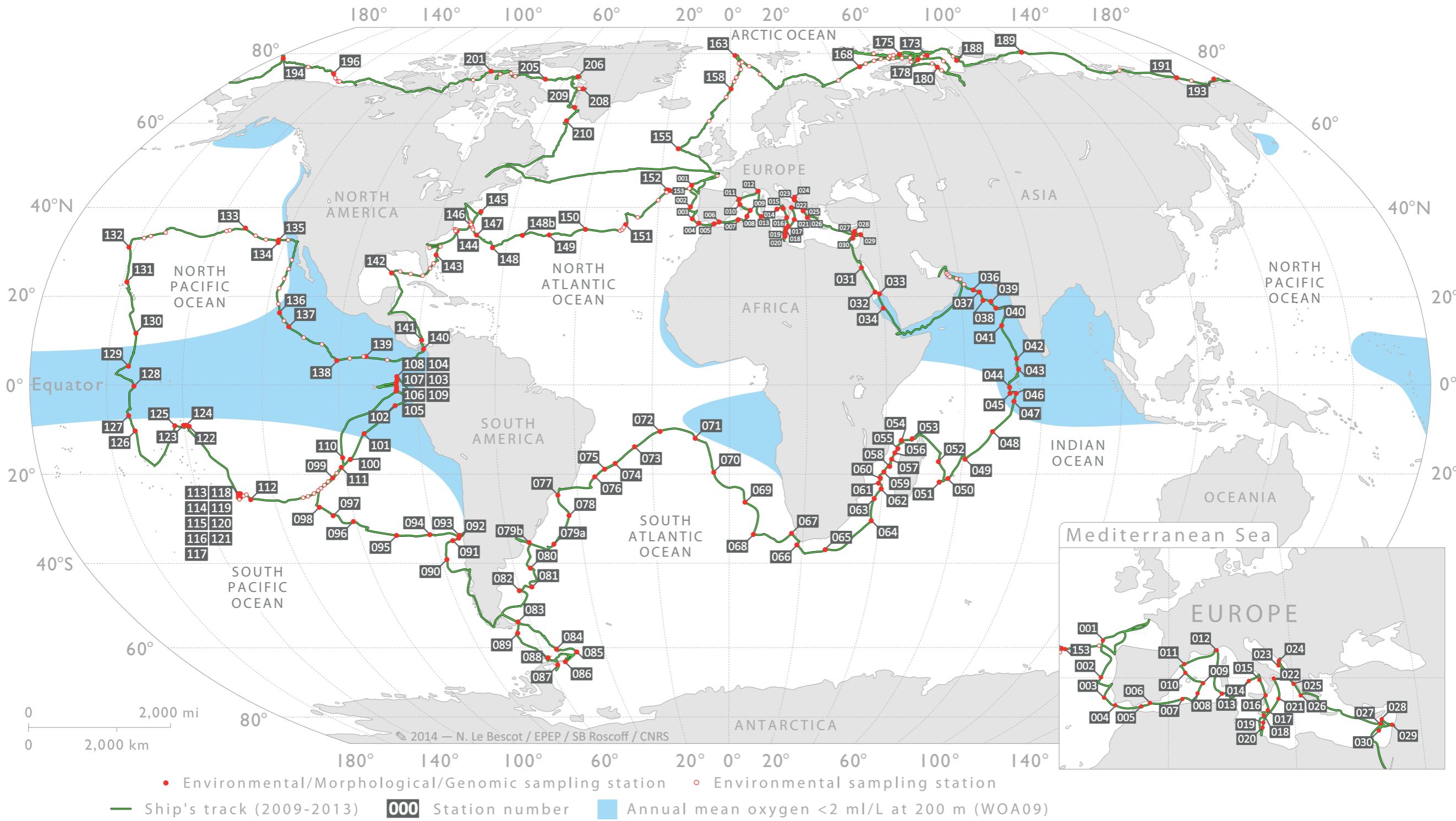
For plankton genomics



The principle of habitat modelling



Application to Tara Oceans metagenomic signal



How to construct the response curve?

= Regression trees

We need

a response variable (e.g. presence/absence)

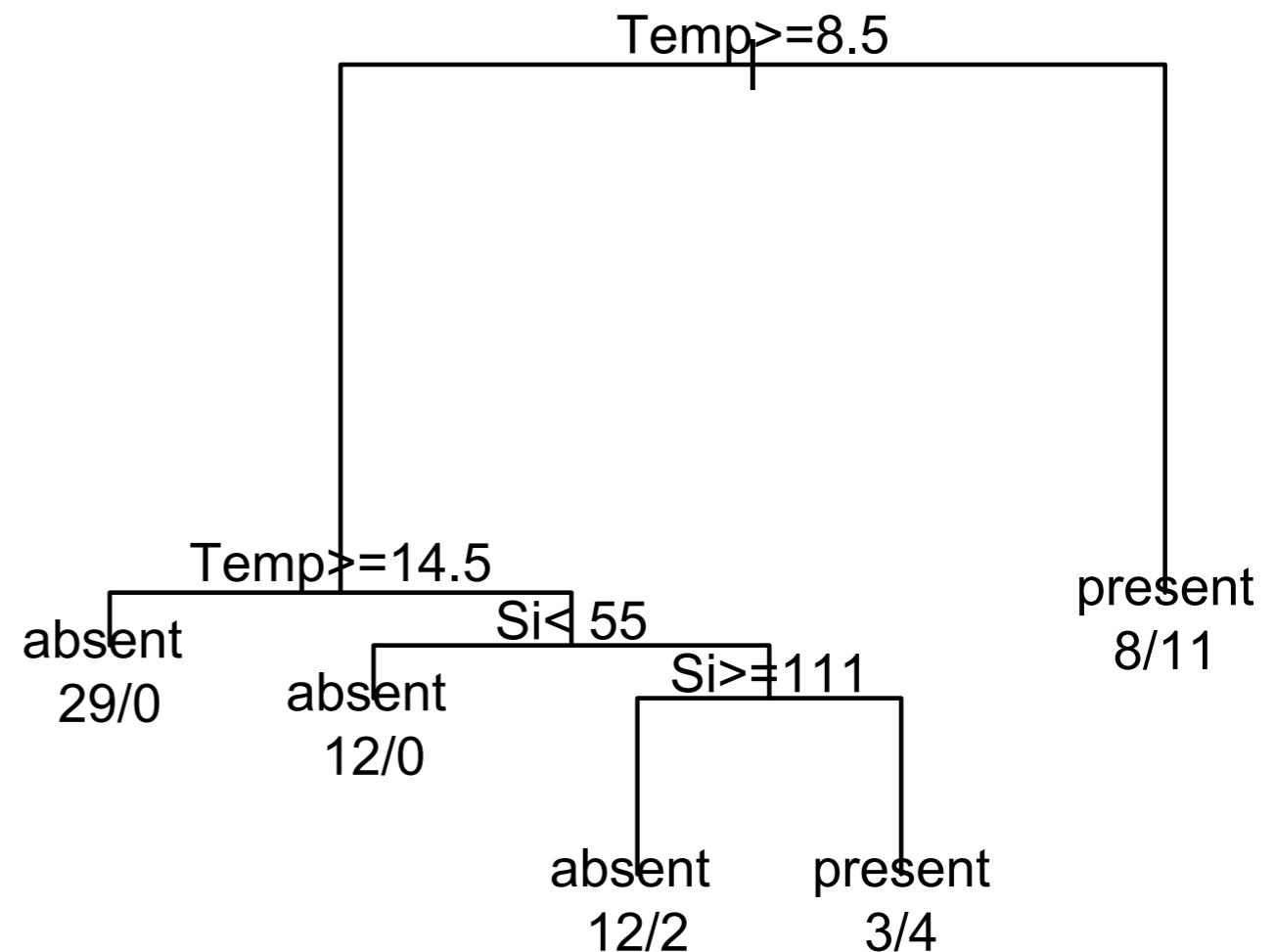
a metric to evaluate (e.g. % or correct predictions)

predictor variables (e.g. environment variables)

Then we **separate** our observations in “bags” according to values of environmental variables.

How? Try everything and see what's best!

That's why it's called machine learning!



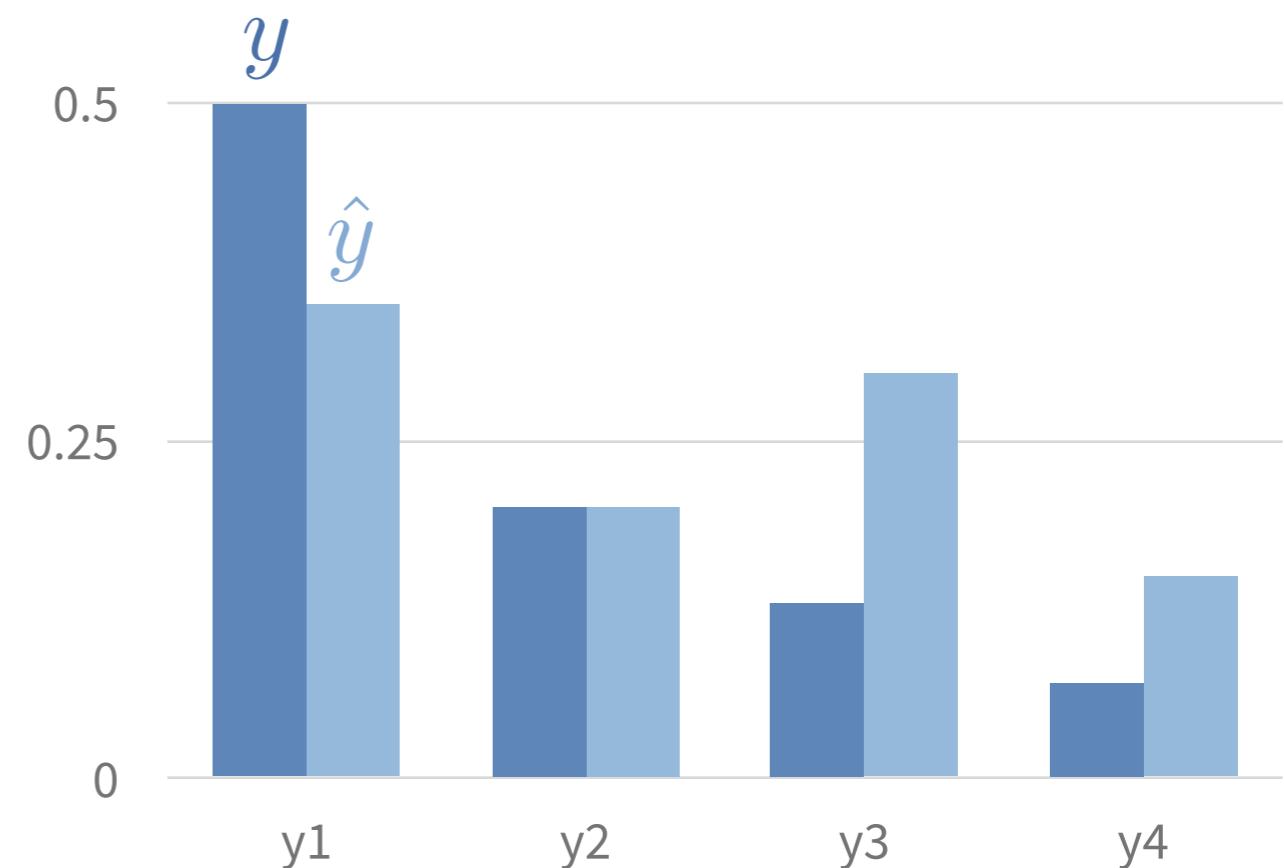
Regression trees for *multiple concentrations*

The response variable is a **vector** of concentrations

The metric is the **multiple Root Mean Square Error**

$$\sqrt{(y - \hat{y})^2}$$

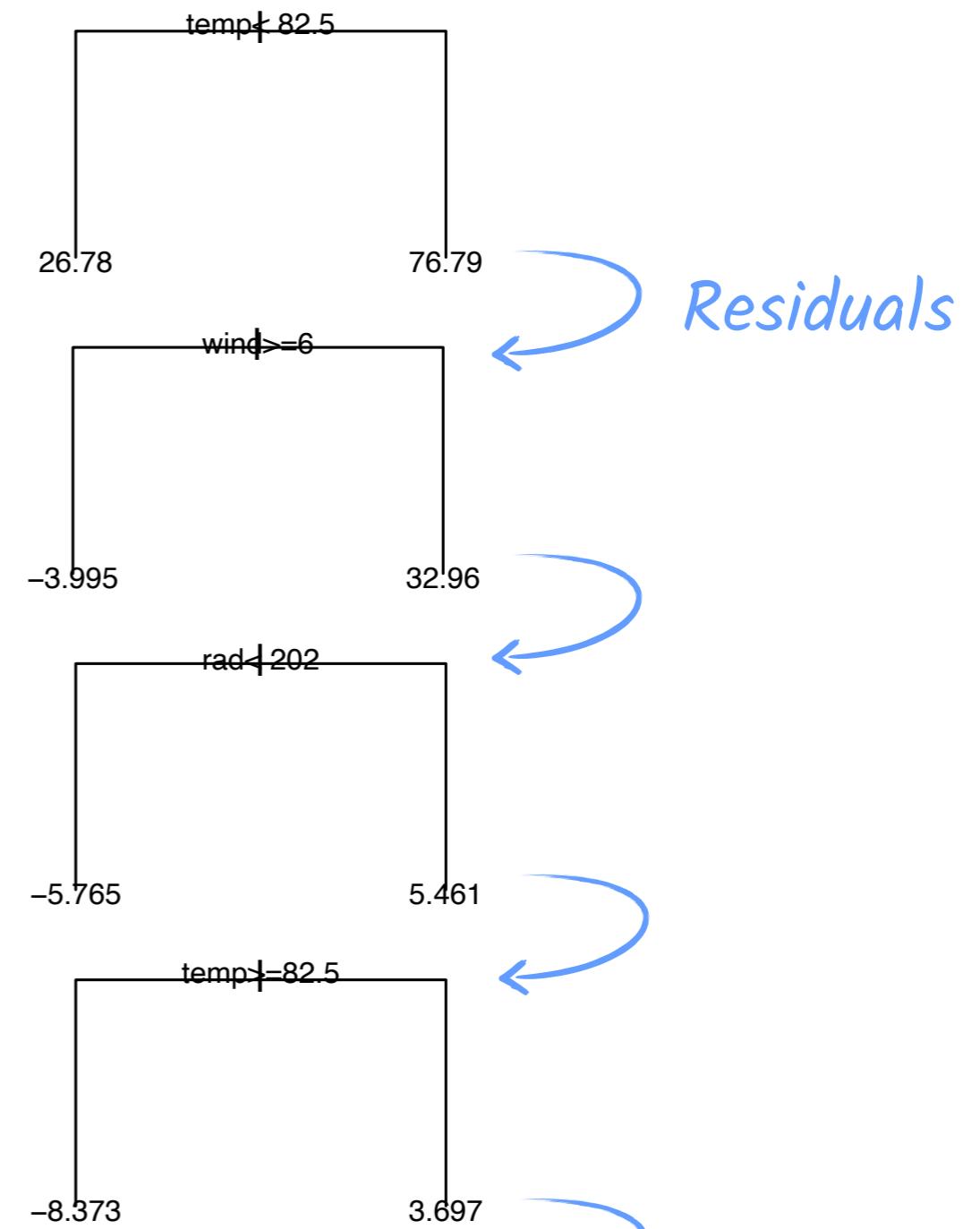
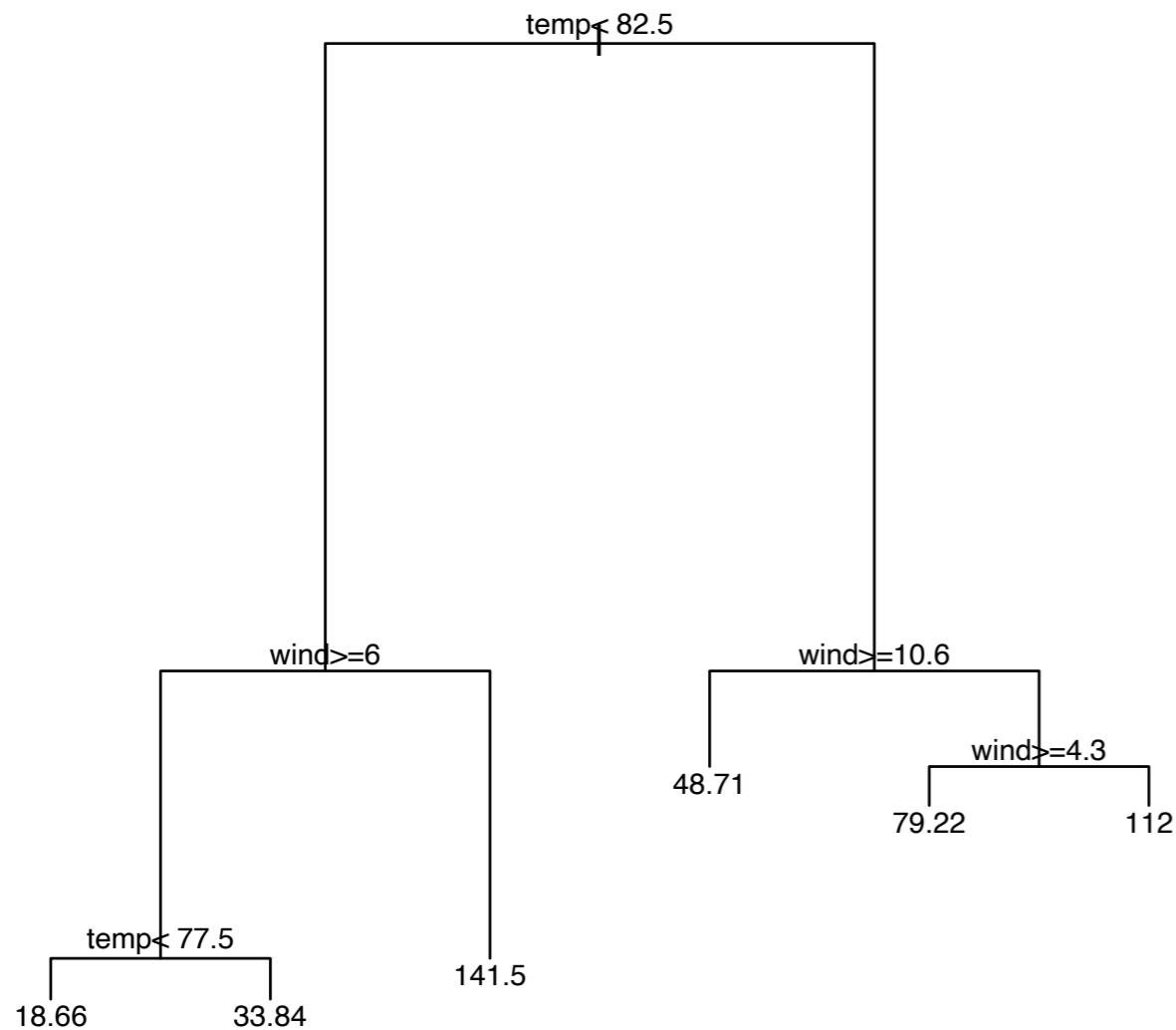
The explanatory variables are environmental **climatologies** (temperature, Chl a, mixed layer depth, etc.)



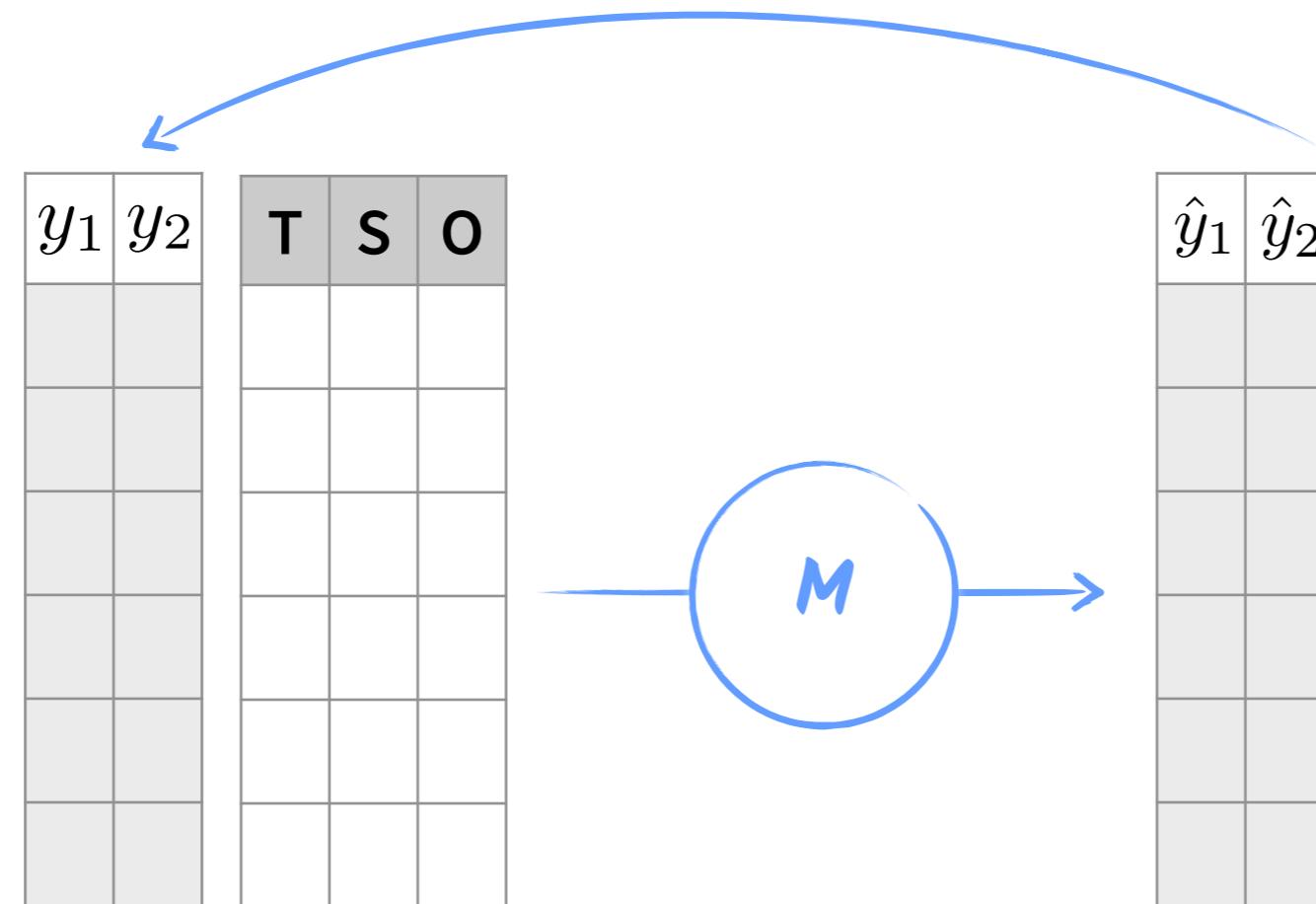
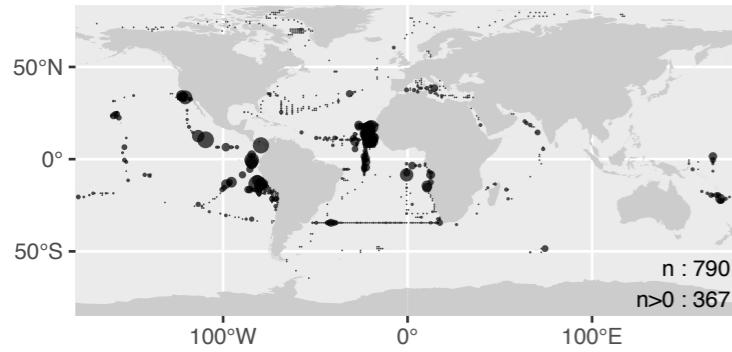
Boosted regression trees

Replace one deep tree

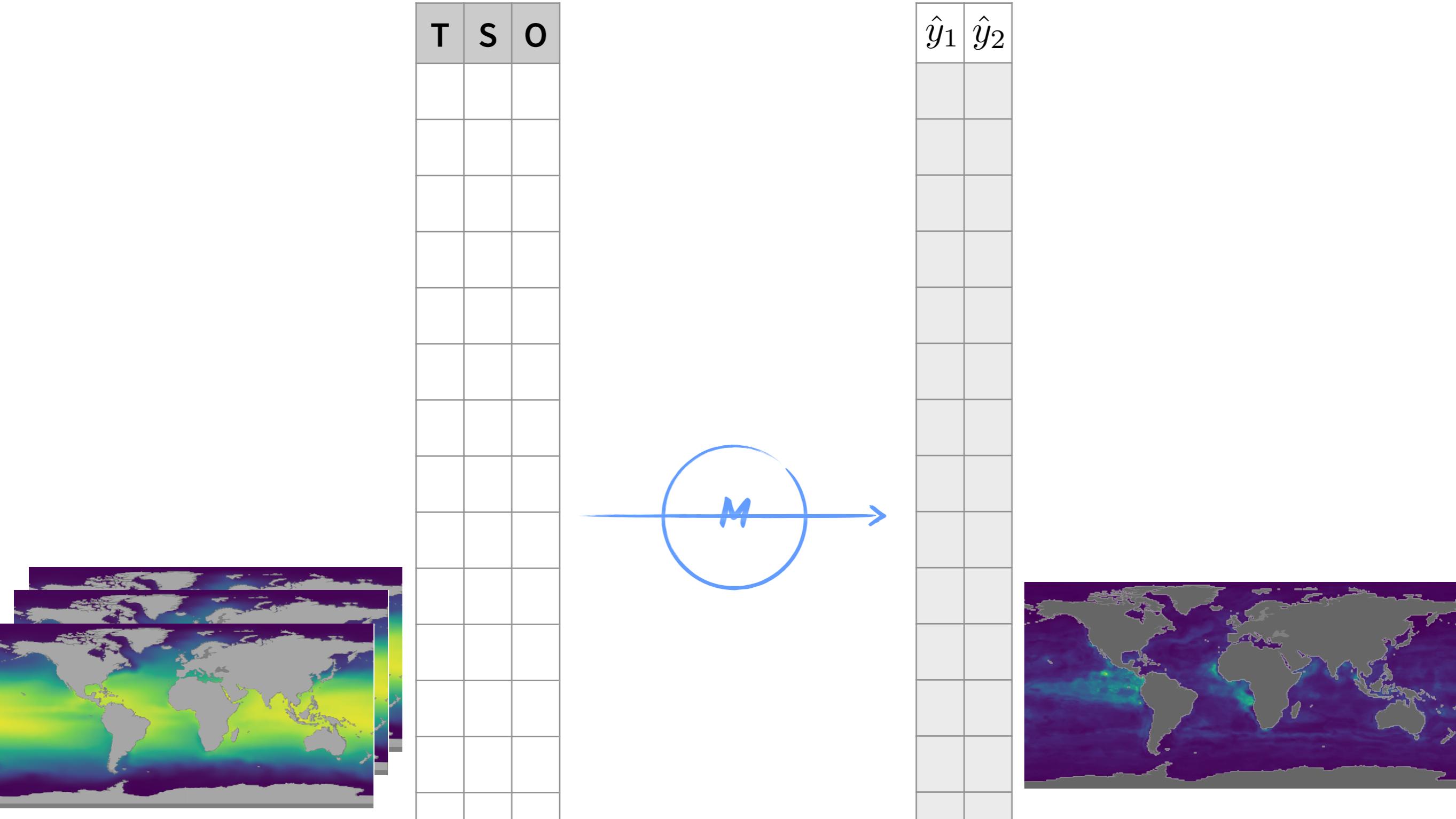
with **many** shallow ones, applied in sequence



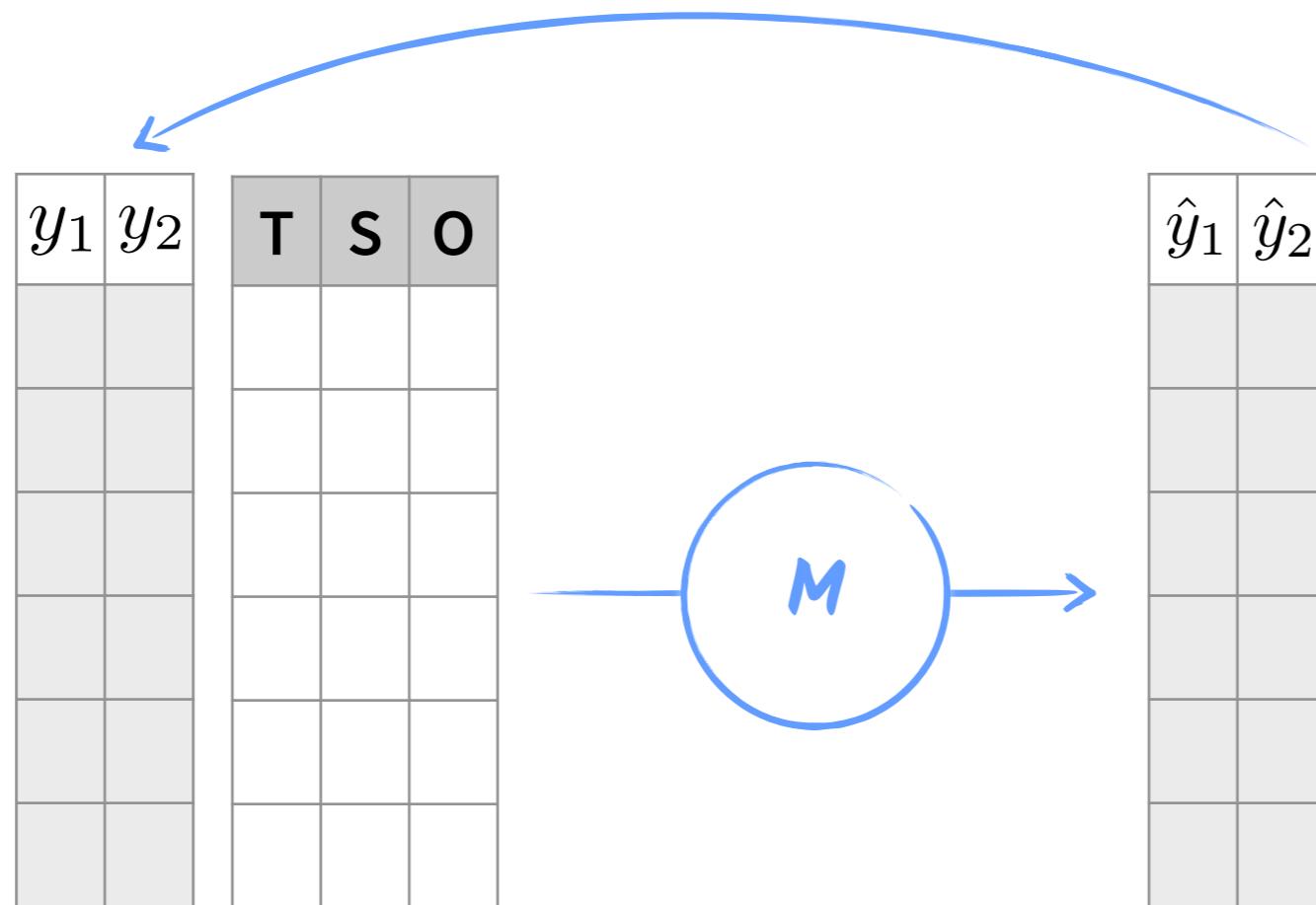
Training



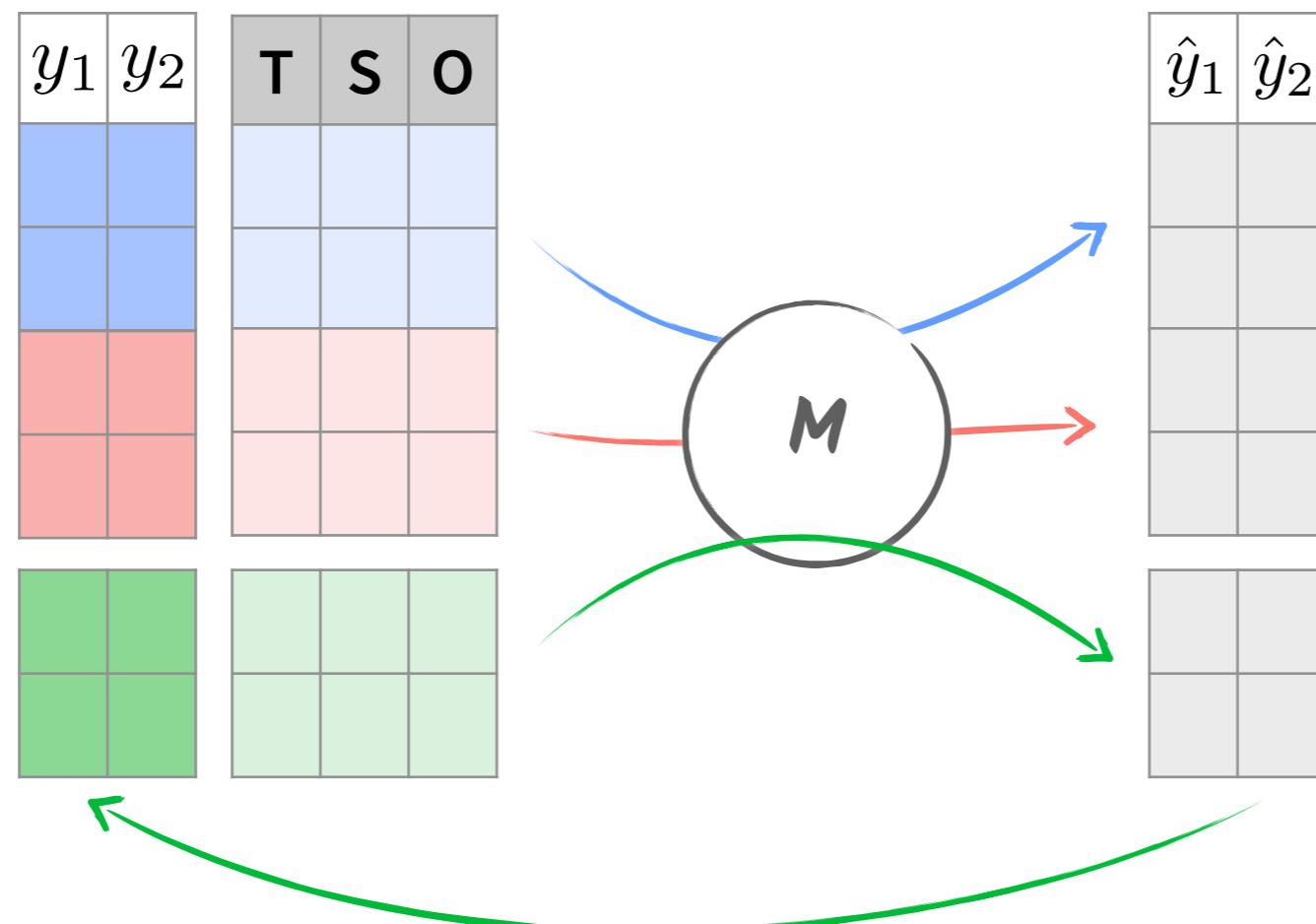
Prediction



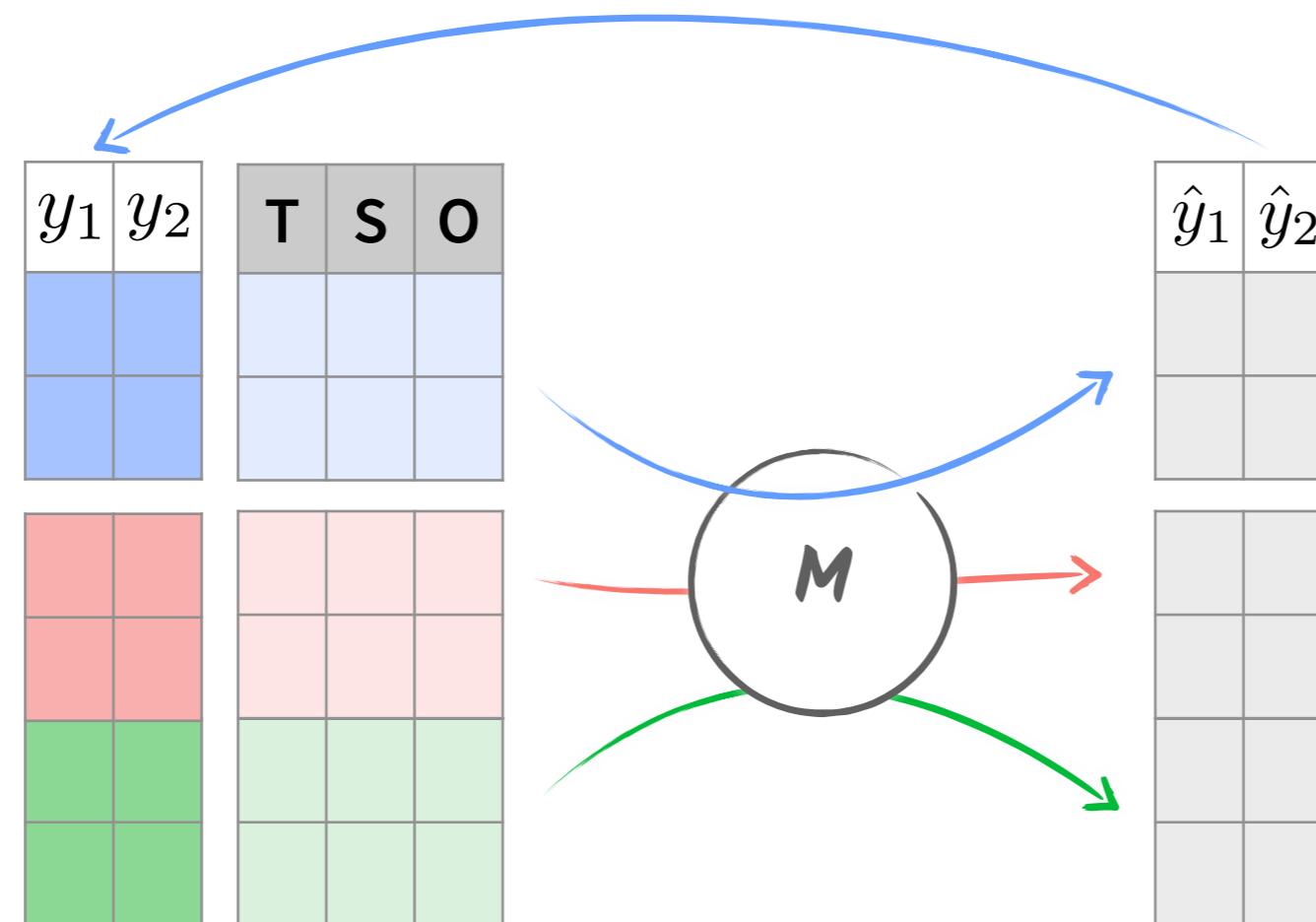
Cross-validation during training



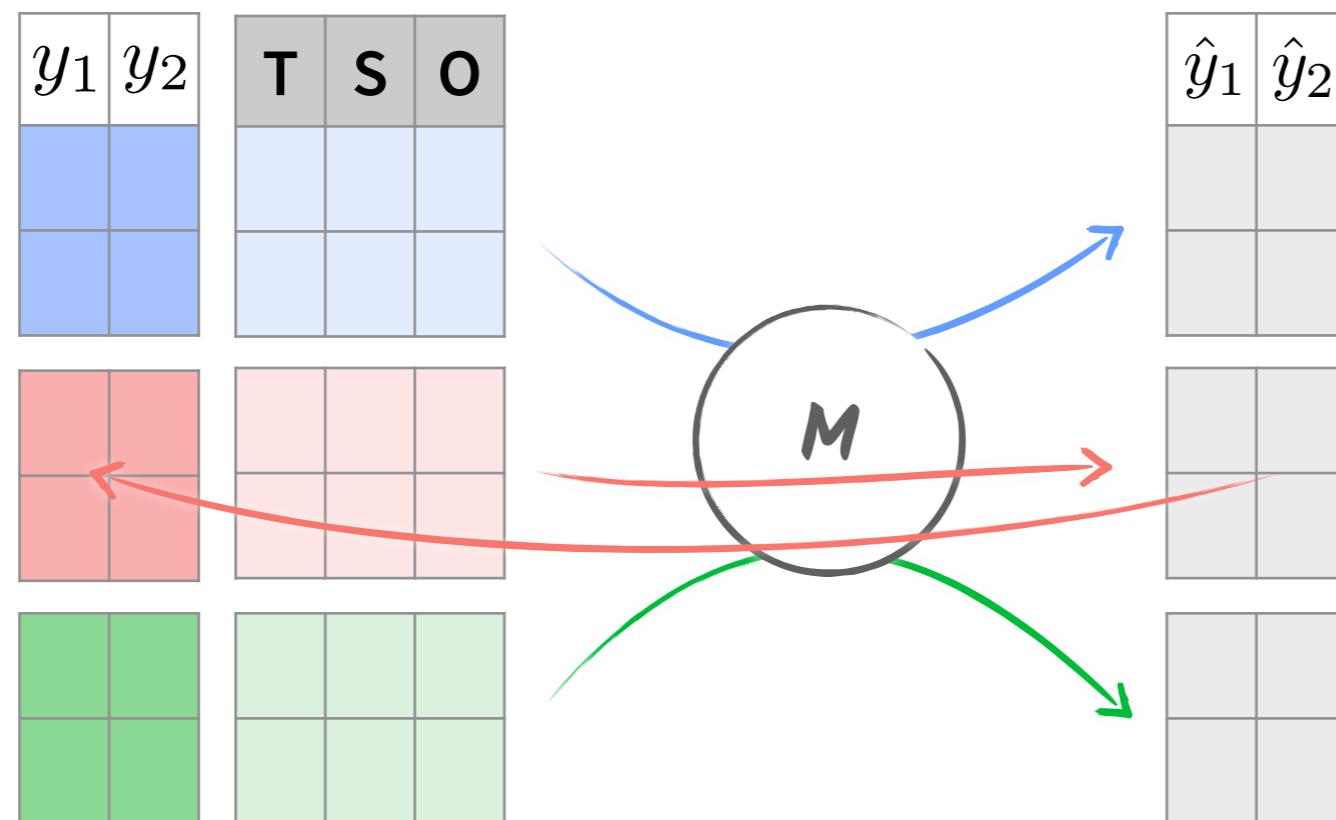
Cross-validation during training



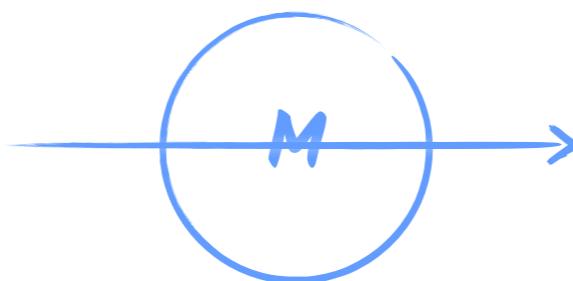
Cross-validation during training



Cross-validation during training

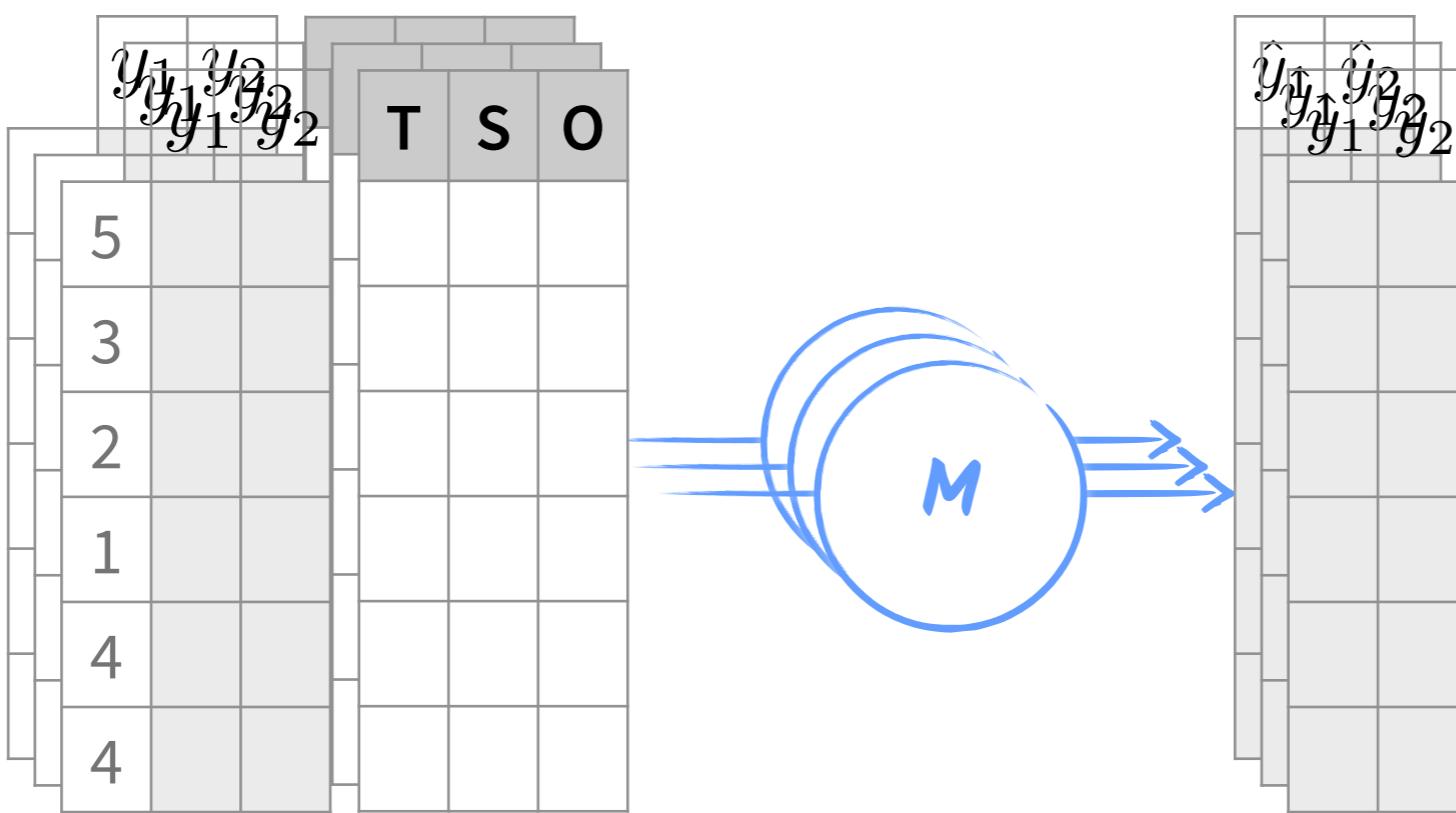


Bootstrap for prediction uncertainty

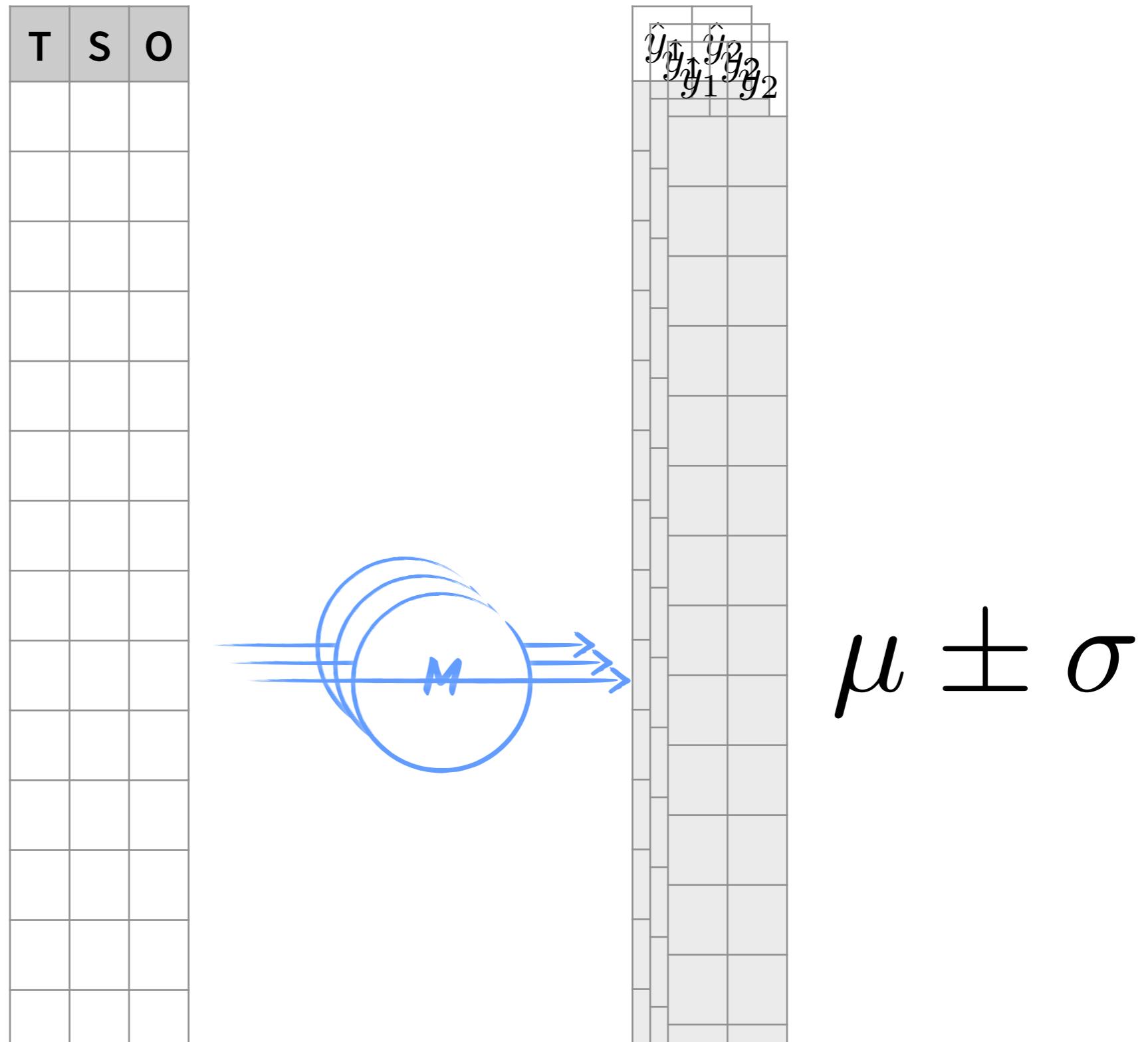


$\sigma^?$

Bootstrap for prediction uncertainty



Bootstrap for prediction uncertainty



The background of the slide is a high-angle aerial photograph of a coastal landscape. The upper portion shows a mix of dark blue water and white, textured clouds. Below, a rocky shoreline is visible, followed by a strip of green vegetation and some buildings, suggesting a small town or resort area.

Thank you



Blue-cloud

Piloting innovative services for Marine Research & the Blue Economy