EPFL

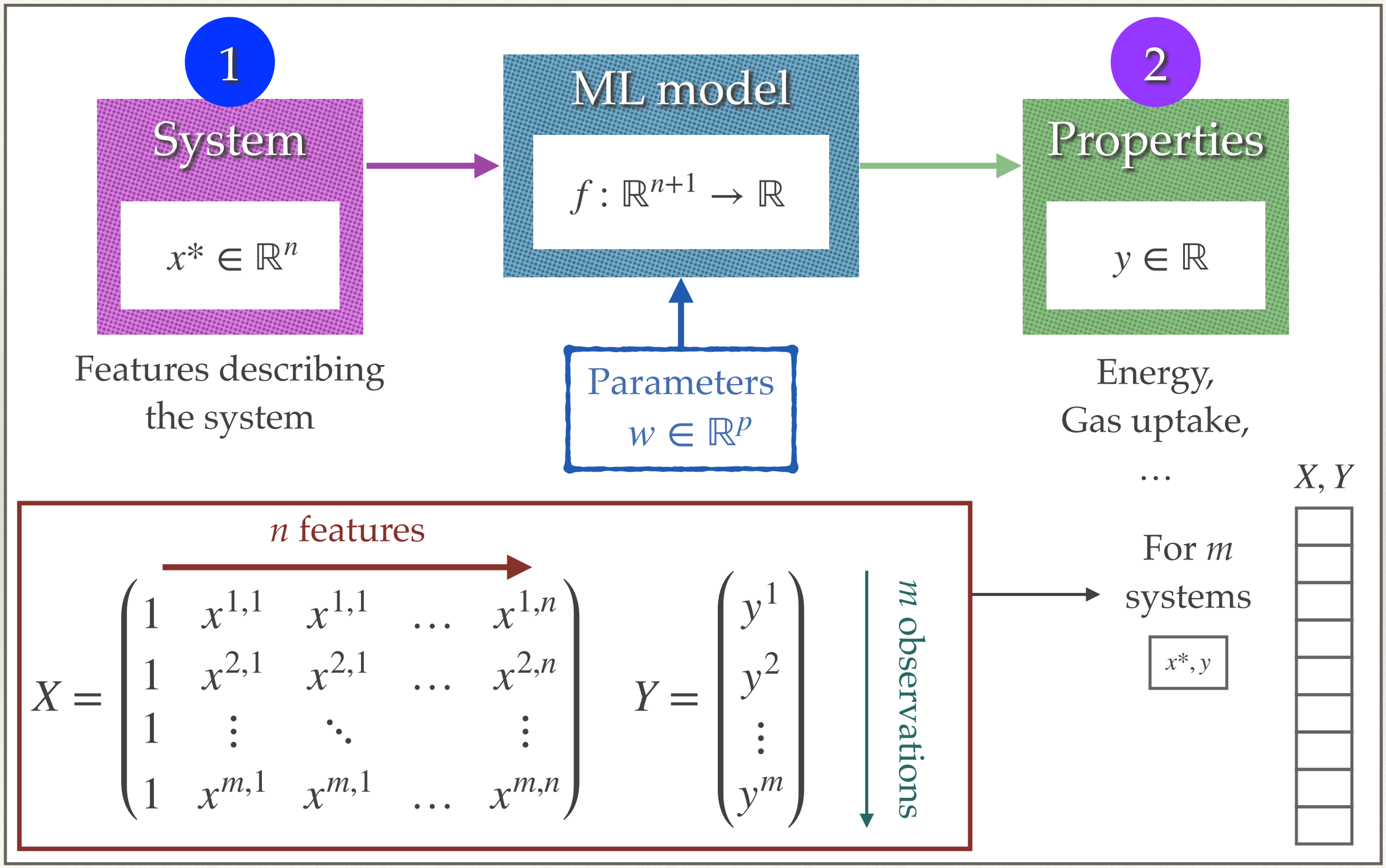
ChE-609

Seyed Mohamad Moosavi, Kevin Maik Jablonka & Berend Smit

Basics of machine learning for chemistry and materials science

seyedmohamad.moosavi@epfl.ch @SMohMoosavi

Collect data: features and labels



Training the model (finding parameters w)

1

System

$x^* \in \mathbb{R}^{n+1}$

Features describing the system

ML model

$f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$

3

Parameters

$w \in \mathbb{R}^p$

2

Properties

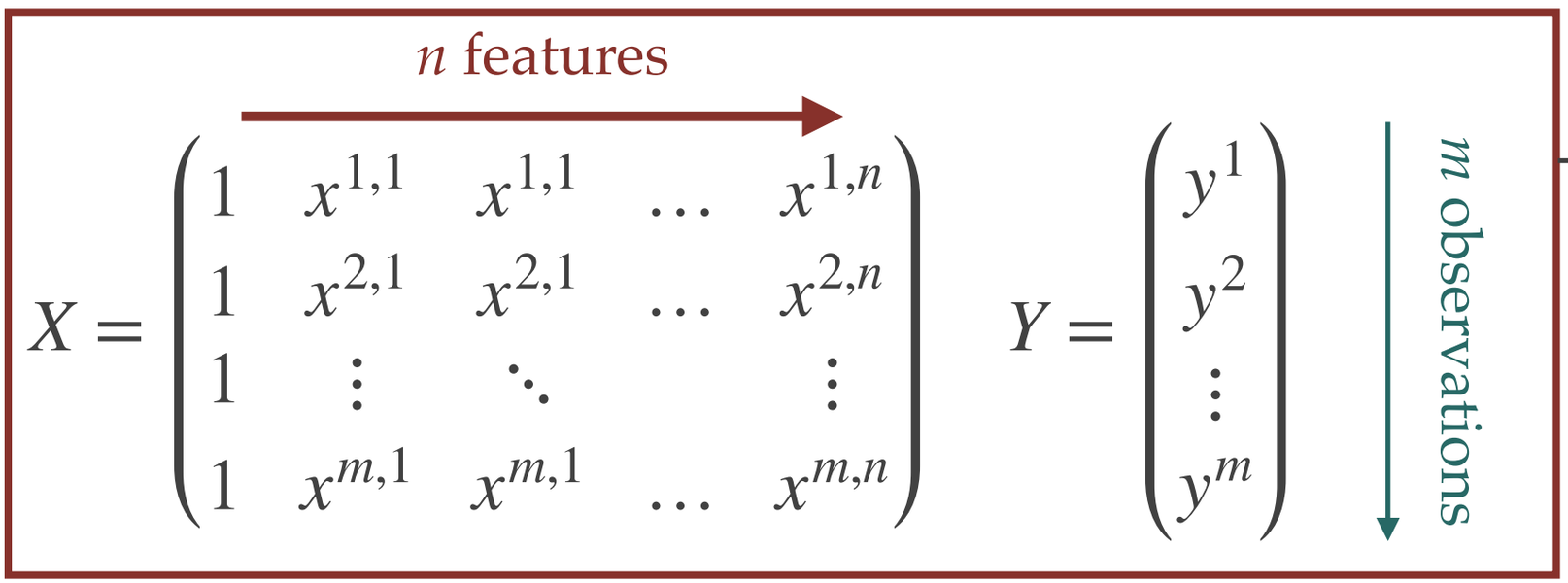
$y \in \mathbb{R}$

Energy, Gas uptake, …

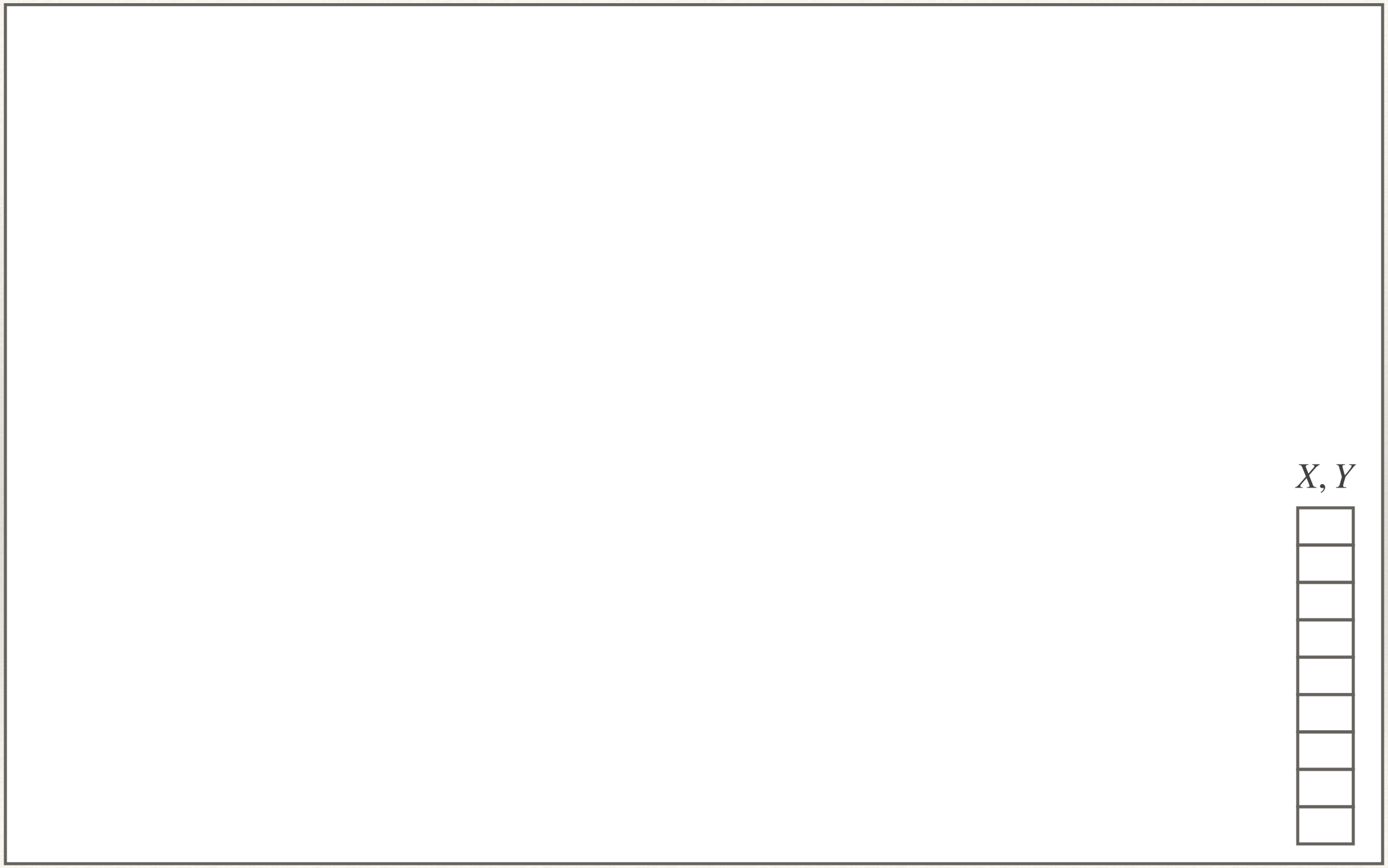
For m systems

x^*, y

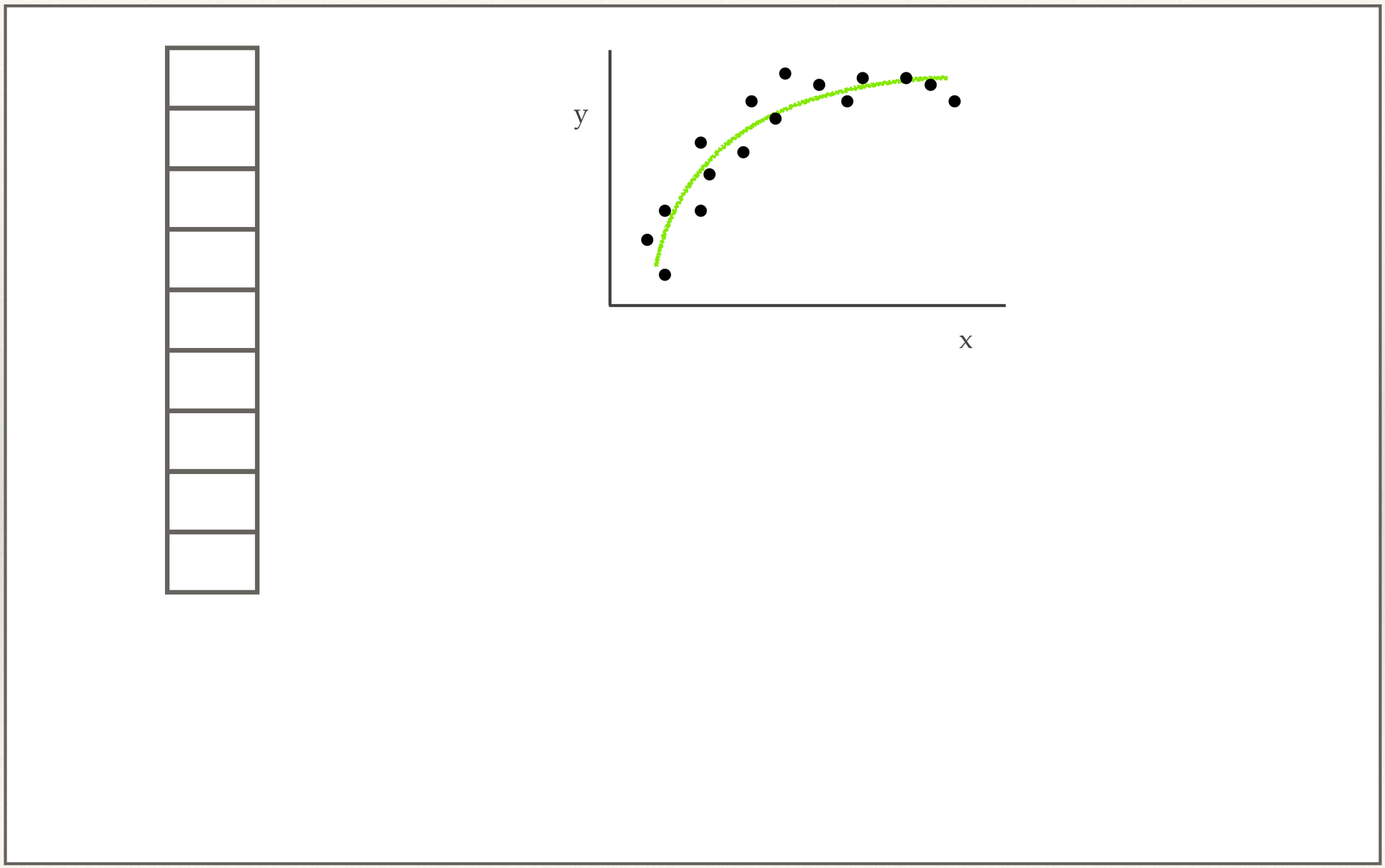
X, Y



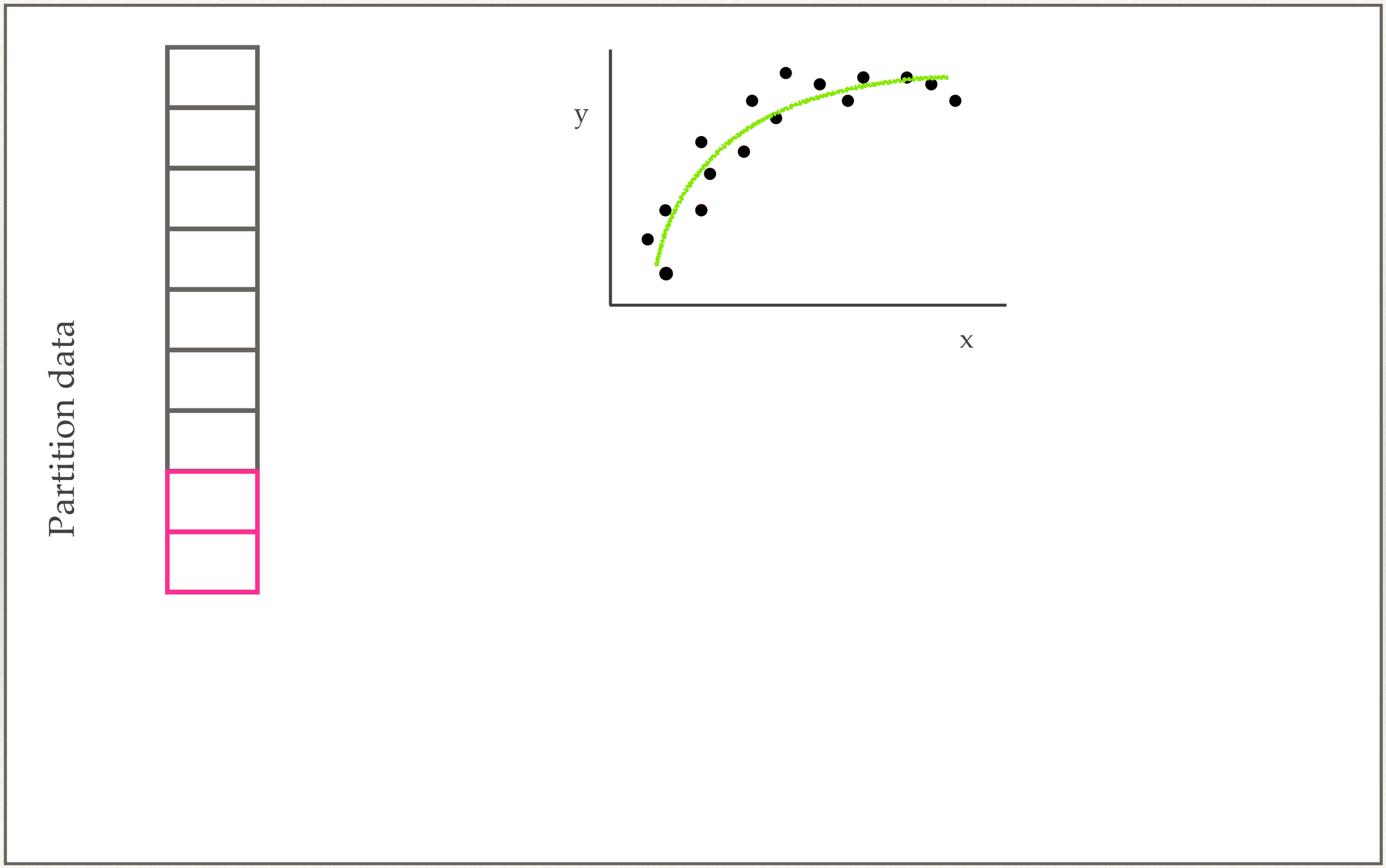
Training the model (finding parameters w)



Training the model (finding parameters *w*)



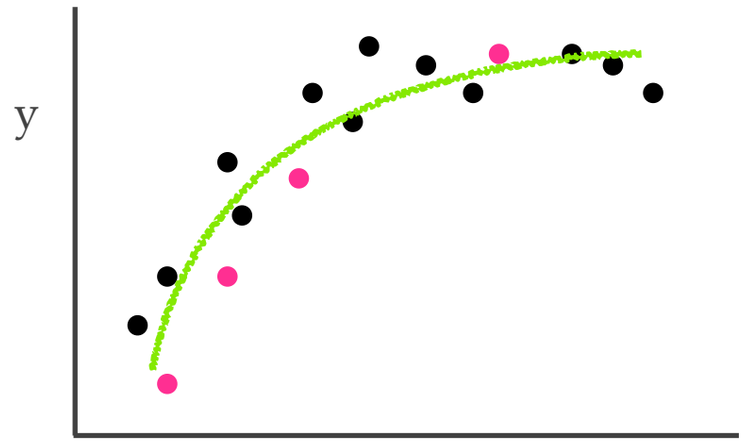
Training the model: data partitioning → keep some data for evaluation



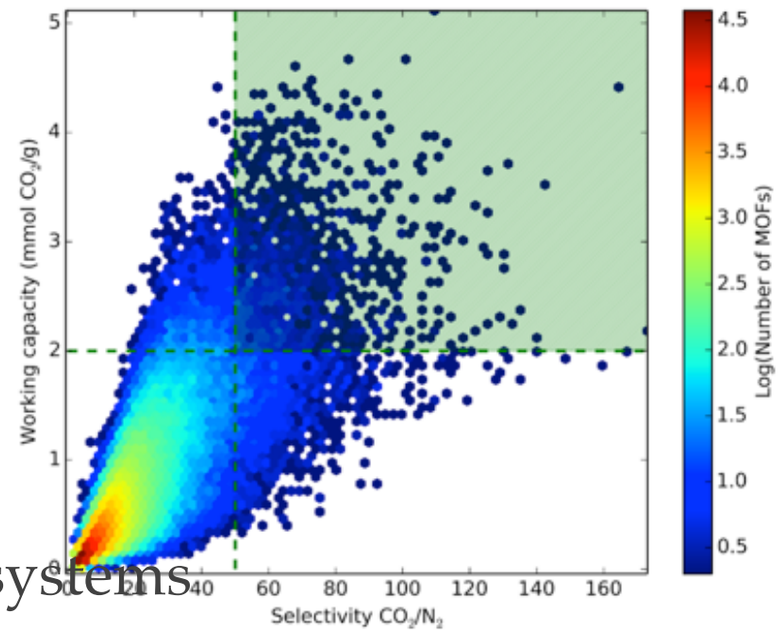
Training the model: data partitioning

Partition data

Train set



Boyd et al. Nature 2019
~300,000 hypothetical MOFs
Search for ~9,000 top performing



The train-test split is not trivial!

Test set

Only for model evaluation!

Note: the aim is to eventually use the model on the systems that the model has never seen