Three core types of ML problems

CEMS
Chemical Engineering & Materials Science

Supervised → data is labeled ($\vec{y}$)

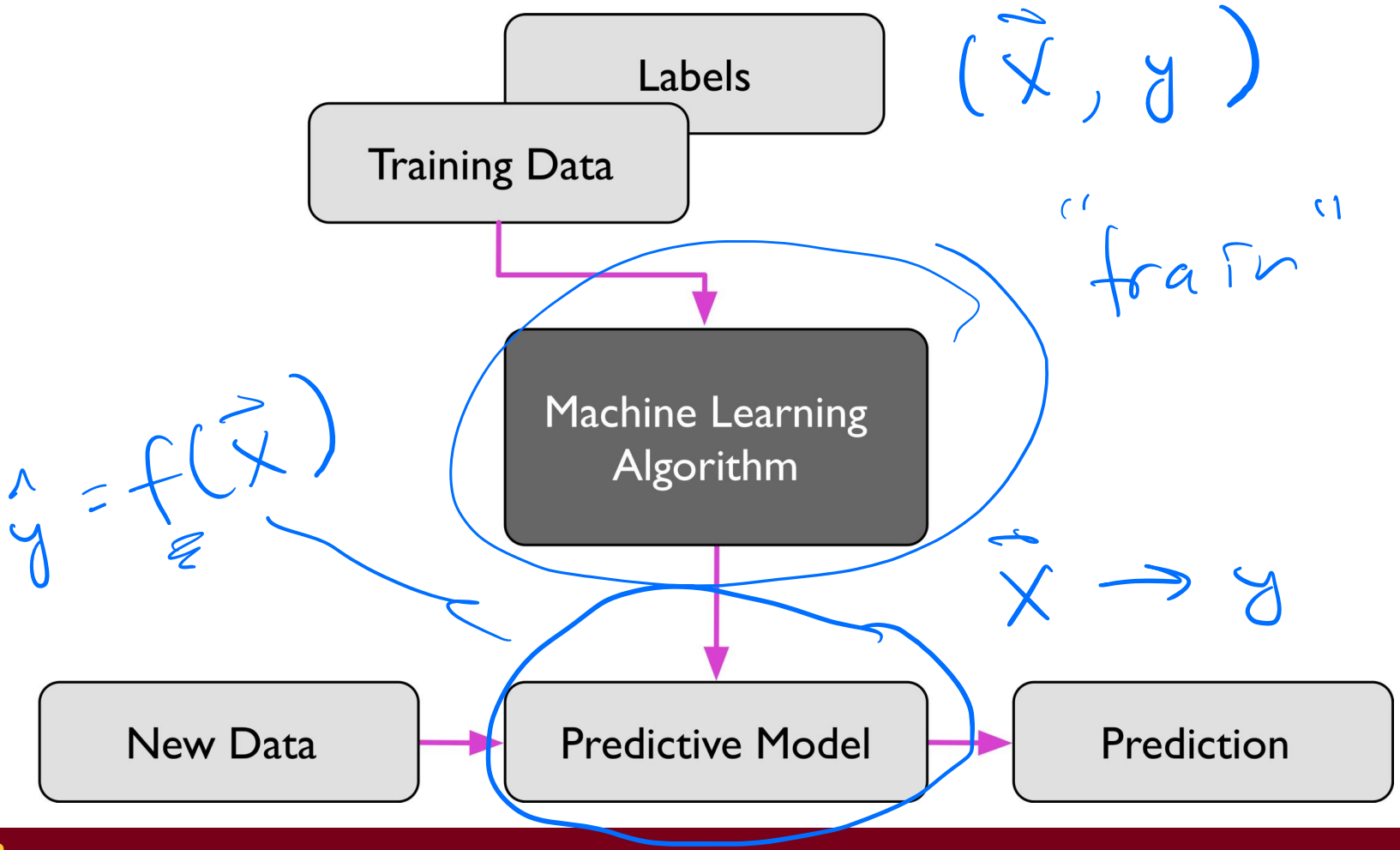
↳ $\vec{y} = f(\vec{x})$

Unsupervised → data is unlabeled

↳ find structure in $\vec{X}$

Reinforcement → learn "actions" (1) max reward.

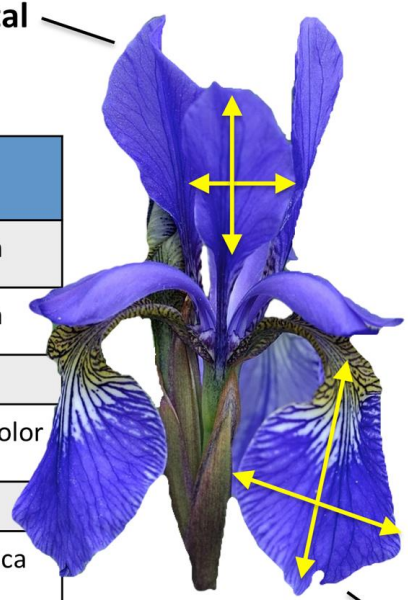
Generic supervised learning problem



Generic supervised learning problem

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



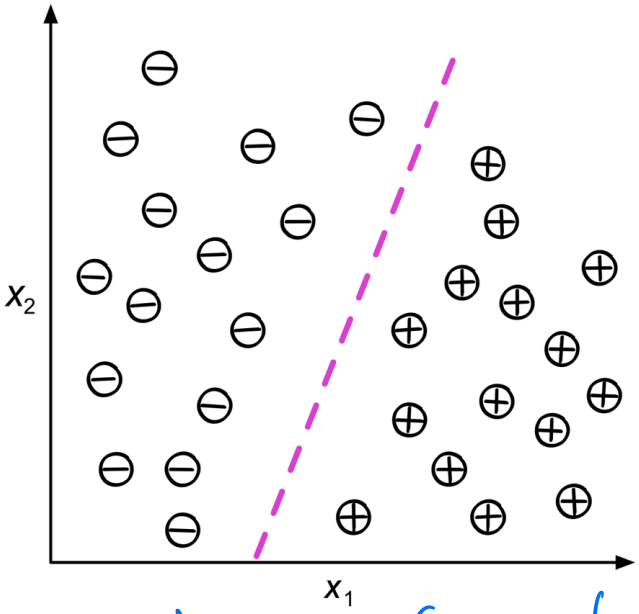
Features
(attributes, measurements, dimensions)

Class labels
(targets)

$x^{(i)}, y^{(i)} \rightarrow$

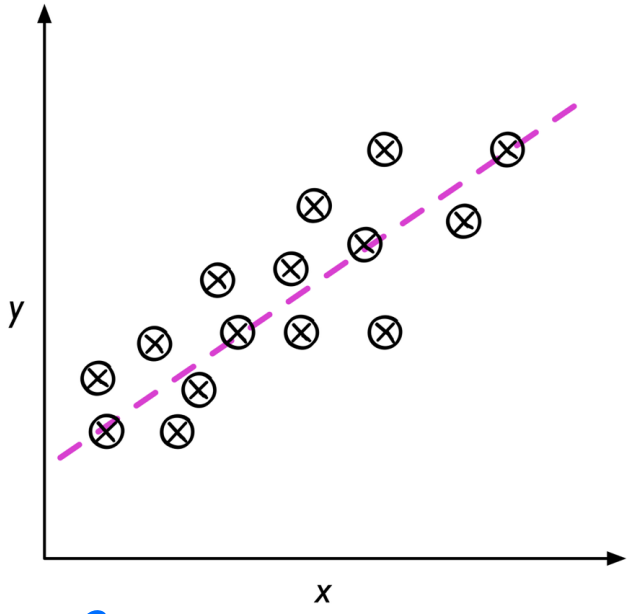
$\hat{y} = f(\vec{x})$

Classification vs regression



Classification

(discrete) ← y → (continuous)



Regression

Why ML models fail

$\hat{y} = f(\vec{x}) \approx y$, Failure = $\hat{y}$ is not close enough y

Causes

- $\vec{x}$ is not expressive enough (bad features)
- y is not long enough (not enough data)
- f can't map $\vec{x} \rightarrow y$ (wrong model)
- $\vec{x}$ and/or y are biased, noisy, incomplete (bad data)

How much data is enough?

CEMS Chemical Engineering & Materials Science

#data needed $\propto$ 1) how weak $\vec{X}$ is
2) how close we need $\hat{y}$ and y to be
3) how broad $(\vec{X}, y)$ are

very rough guide

↳ well-structured data + good features
↳ $\sim 10^2$ +
↳ weak structure, basic features, complex problem
↳ $\sim 10^4$ +

What about features?

CEMS
Chemical Engineering
& Materials Science

What are attributes of good/bad features?

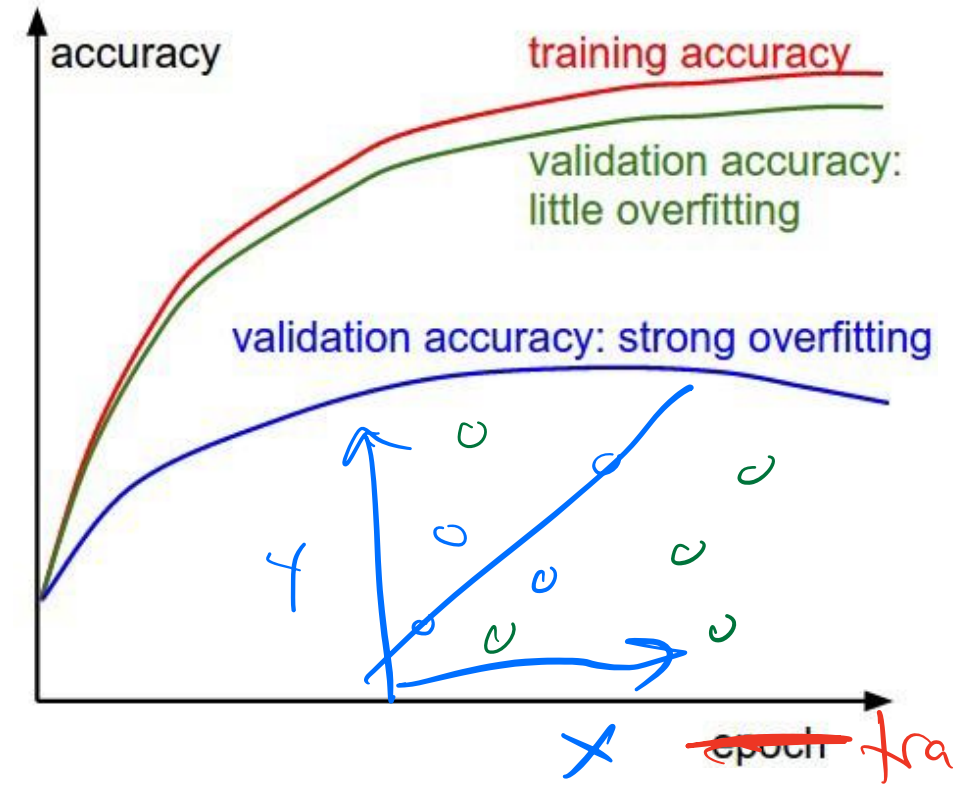
Good

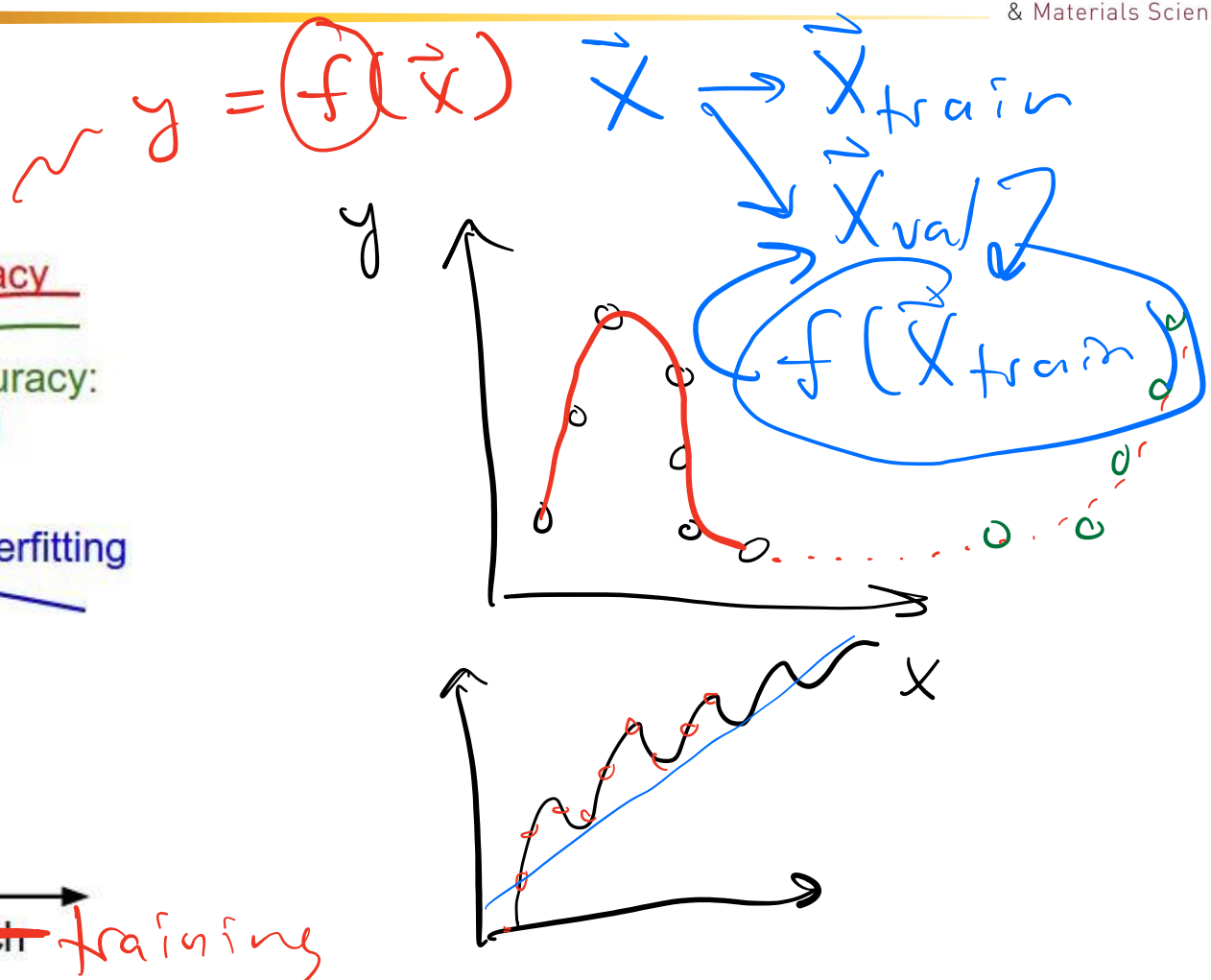
- Strong covariance w/ target.
- Cheap to calculate
- Complementary.
- Interpretable
- Well defined.

Bad

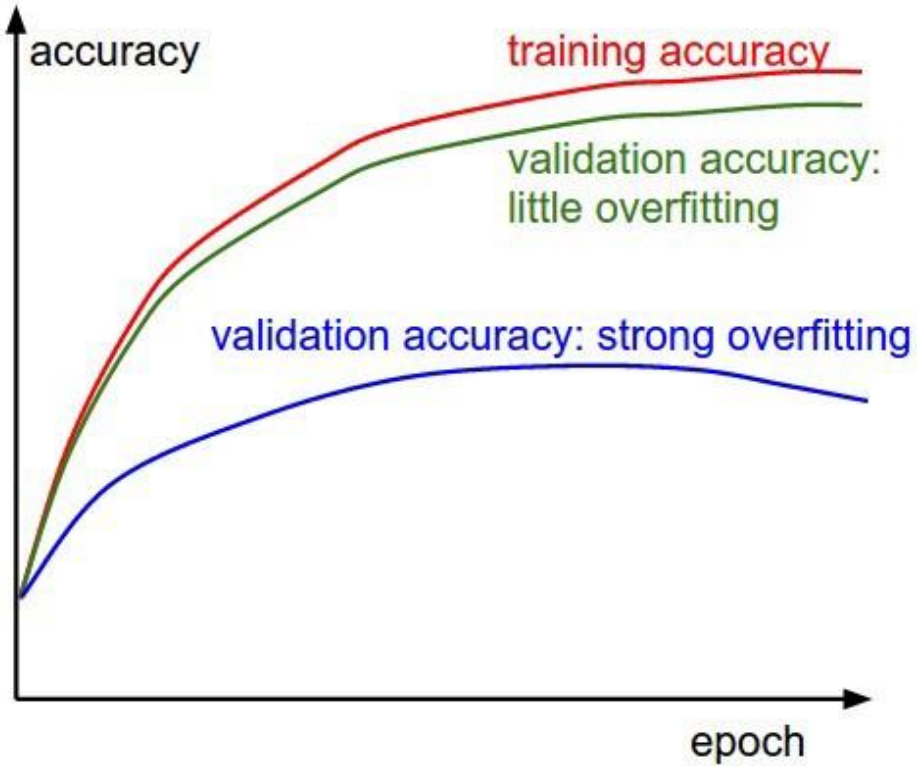
(- Irrelevant)
- Expensive
- Correlated
- Sparse
- Opaque

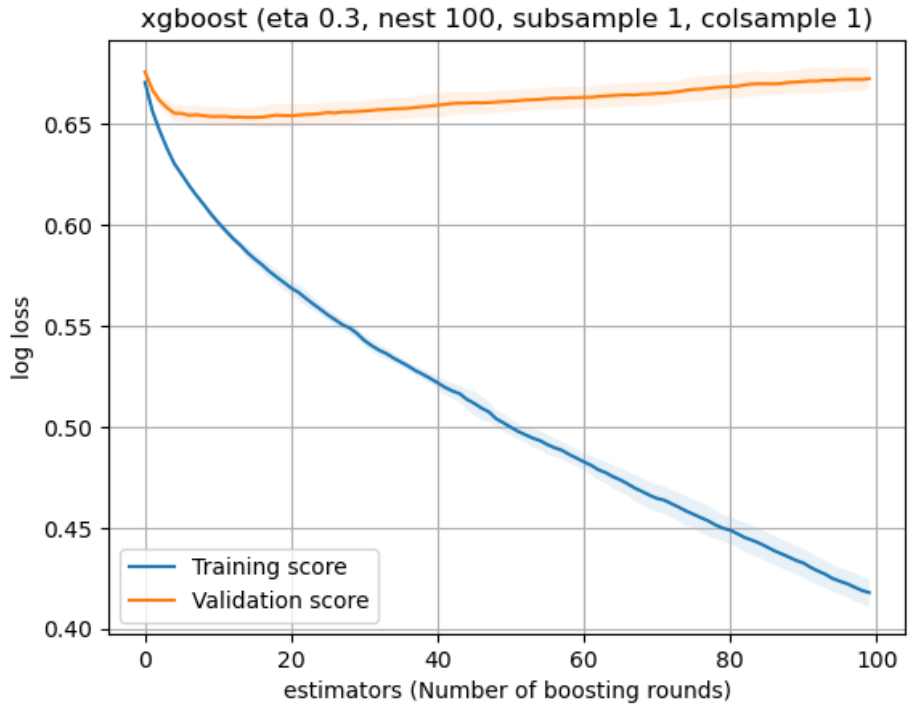
Validation is essential!





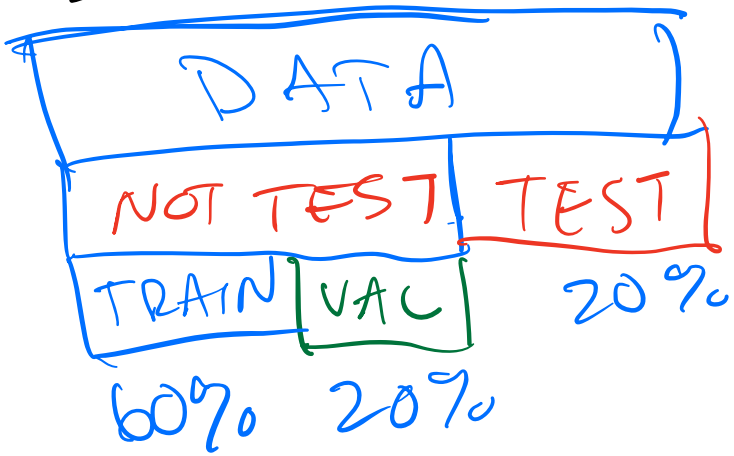
Validation is essential!





How to approach an ML problem

1) Research question
2) Hypothesis
3) Data collection
4) Analysis



4a) Preprocess data
4b) Develop training & validation protocol.
4c) Train models
4d) Evaluate
4e) Test

Domain knowledge (your background) is key!

CEMS
Chemical Engineering
& Materials Science

What is the right property to predict?

What is the right way to assess performance?

Which data points are compatible?

Which features might be informative?

…