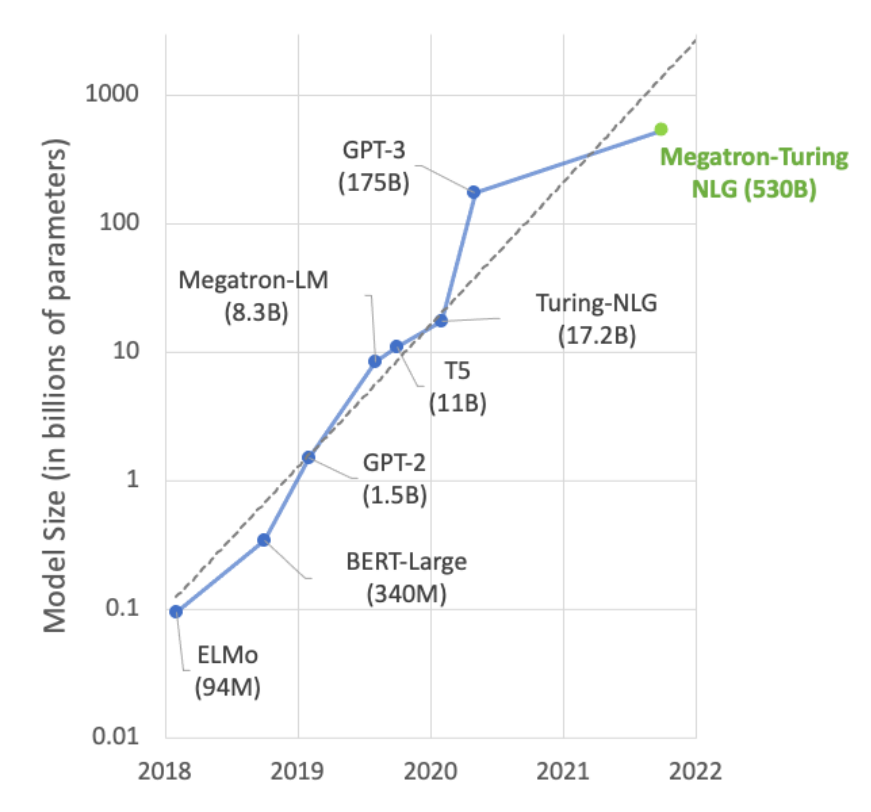
MATHEMATICAL TOPICS IN MACHINE LEARNING

(LECTURE 2 – EMPIRICAL RISK MINIMIZATION)

Professor Gavin Brown

OUR QUESTION

"Are bigger models always better models?"



TODAY:

A mathematical framework to understand the nature of **overfitting**.

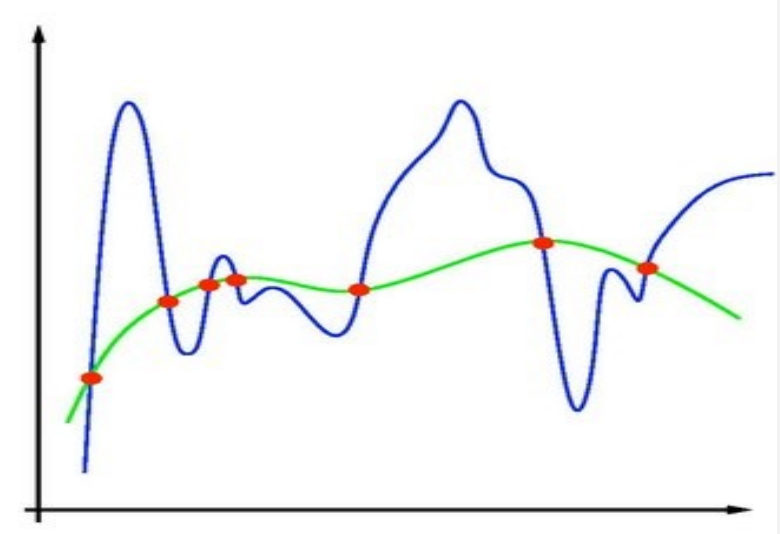
A way to express the error of a model in two parts – separating **size of model** from the **size of data**.

STATISTICAL LEARNING THEORY

Mathematical framework for understanding/deriving learning algorithms.

Pioneered by Vladmir Vapnik in 1960s USSR.
Re-discovered and popularized in 1990s.

Empirical risk minimization is part of this.



EMPIRICAL RISK MINIMIZATION

Basically, this means *using a training set and minimizing a loss function*.

But, phrasing it formally brings other benefits, as we will discover….

$$\ell_{train}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\Big(y_i - f(\mathbf{x}_i;\mathbf{w})\Big)^2.$$

model

true label

features

parameters

A MODEL IS A **FUNCTION** $f : \mathcal{X} \to \mathcal{Y}$

$$\mathcal{X} \longrightarrow \mathcal{Y}$$

$$[0, 255]^{128 \times 128}$$

$$\mathbb{R}^k \ : \ y_j \geq 0, \ \sum_{j=1}^{k} y_j = 1$$

(i.e. a probability distribution over ***k*** classes)

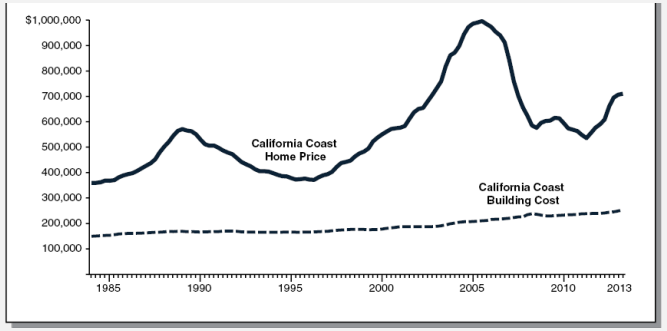
This is a **classification** problem.

A MODEL IS A **FUNCTION** $f : \mathcal{X} \to \mathcal{Y}$

$$\mathcal{X} \longrightarrow \mathcal{Y}$$

$$\mathbb{R}^d \qquad \mathbb{R}$$

	A	B	C	D
1	RowID	Variable_1	Variable_2	Variable_3
2	1	12.34	aa	12
3	2	34	aa	33
4	3	44	cc	#N/A
5	4	-433	ff	44
6	5		gg	66
7	6	43	dd	33
8	7	34	#NAME?	66



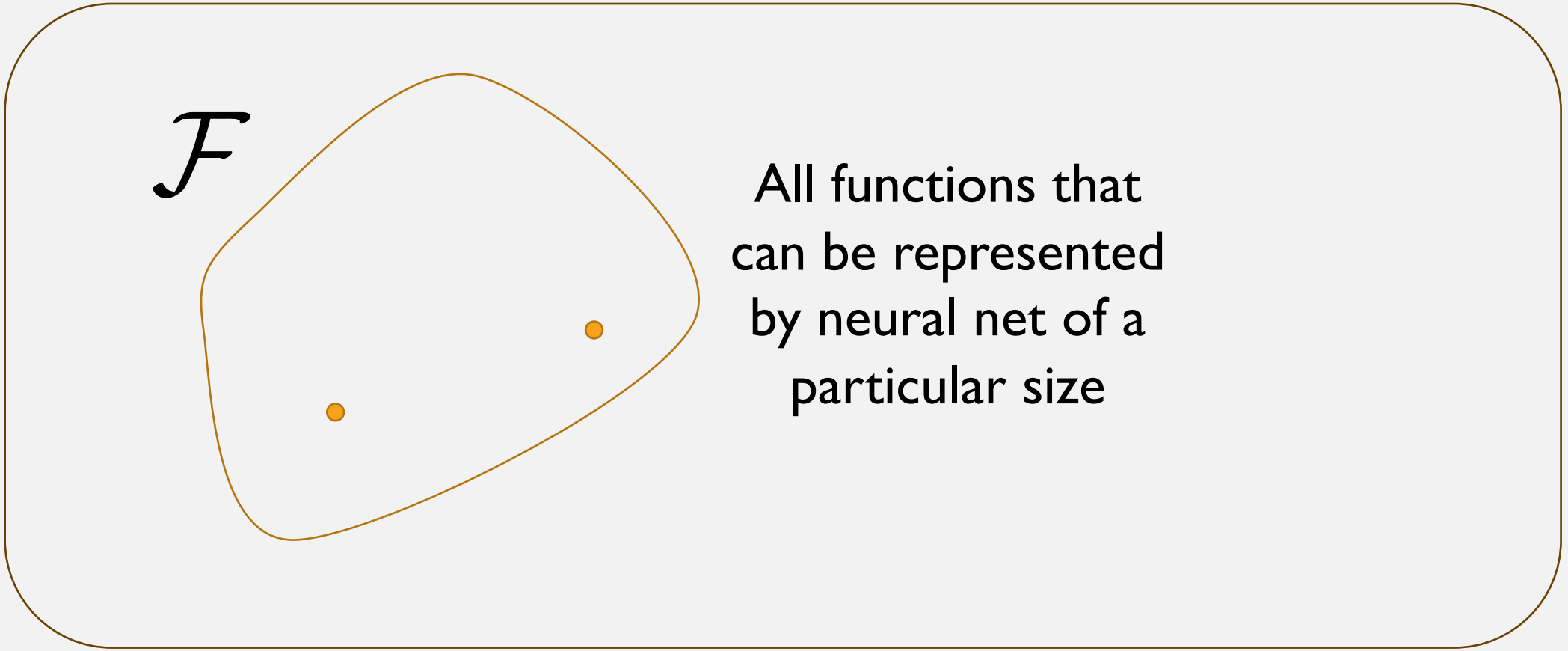
This is a **regression** problem.

A MODEL IS A **FUNCTION** $f : \mathcal{X} \to \mathcal{Y}$

$$\mathcal{X} \longrightarrow \mathcal{Y}$$

$$\mathbb{R}^d \qquad\qquad \mathbb{R}^k$$

A MODEL "FAMILY"



Ω All possible functions from X to Y

THE I.I.D. ASSUMPTION

We assume datapoints $(\mathbf{x}, \mathbf{y})$ are sampled from a distribution.

$$(\mathbf{x}, \mathbf{y}) \sim P(\mathbf{x}, \mathbf{y})$$

Our dataset is then assumed to be *n* independent samples from $P(\mathbf{x}, \mathbf{y})$

$$\mathcal{S}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$$

… **I**ndependent and **I**dentically **D**istributed.

THE EMPIRICAL RISK

The "hat" denotes that it's an **estimate**.

$$\hat{R}(f, \mathcal{S}_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i)).$$

THE EMPIRICAL RISK **MINIMIZER**

$$\hat{R}(f, \mathcal{S}_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i)).$$

$$f_{erm} := \operatorname*{arginf}_{f \in \mathcal{F}} \hat{R}(f, \mathcal{S}_n).$$

The model obtained if we found the **global minimum** of the empirical risk.
Note that we are working within the constraints of our model family here.

EMPIRICAL RISK **VS** POPULATION RISK

$$\hat{R}(f, \mathcal{S}_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i)).$$

$$f_{erm} := \operatorname*{arginf}_{f \in \mathcal{F}} \hat{R}(f, \mathcal{S}_n).$$

The **estimated** risk, using an IID sample of size *n*.

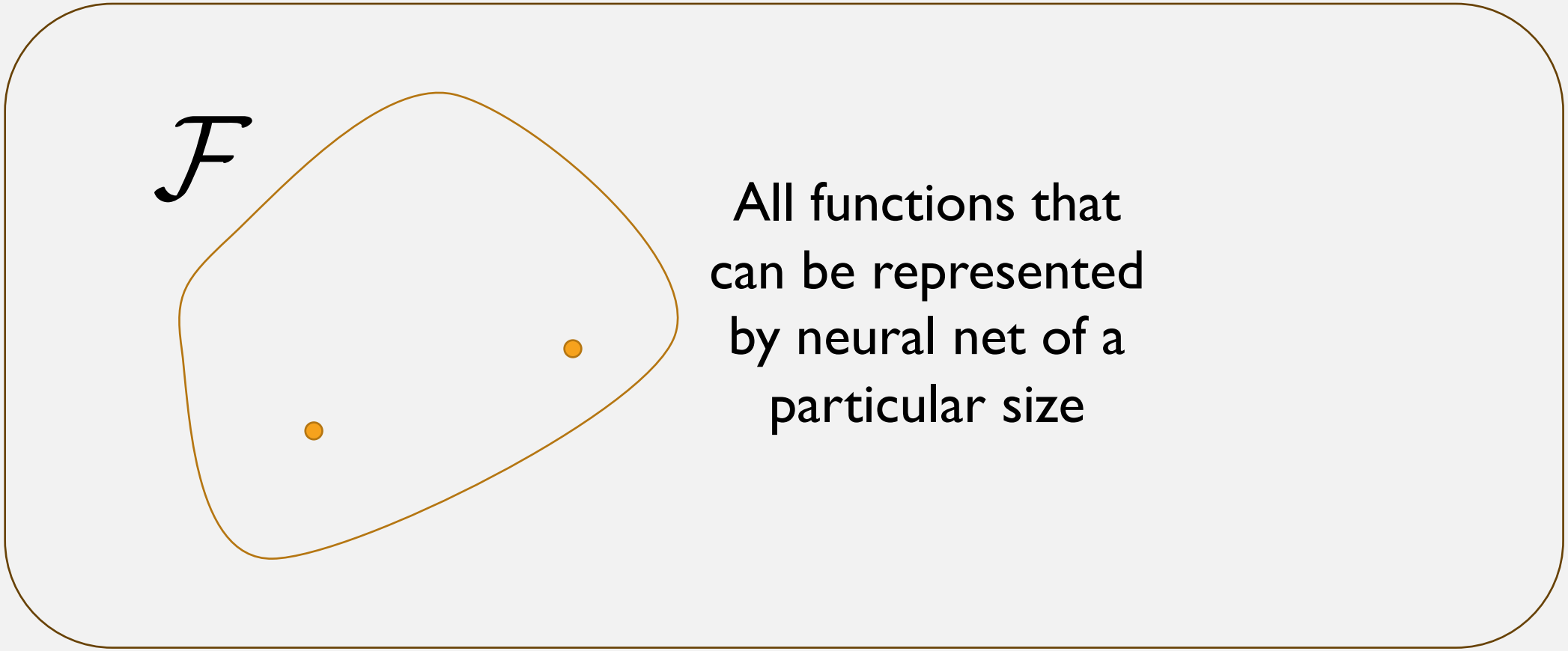
$$R(f) := \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \Big[\ell(\boldsymbol{y}, f(\boldsymbol{x})) \Big]$$

$$f^* := \operatorname*{arginf}_{f \in \mathcal{F}} R(f).$$

The **true** risk, using all possible data we may ever encounter.

*(**NOT** the same thing as testing data!)*

WHAT IF WE HAD NO RESTRICTED FAMILY?



Ω All possible functions from X to Y

THE BAYES MODEL

What is the optimal prediction overall?

(named after Thomas Bayes, of Bayes' theorem)

For squared loss, this is…

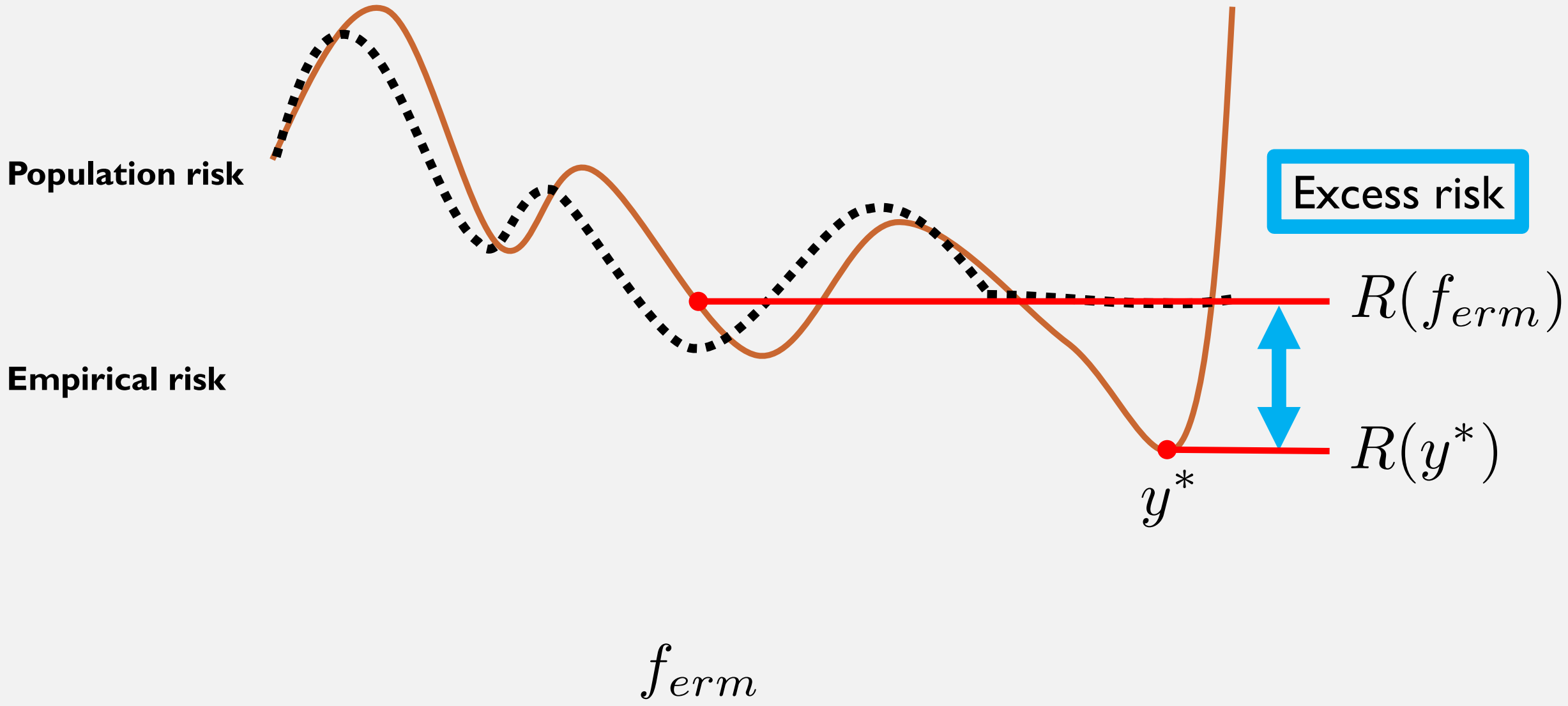
$$y^* := \arg \inf_{f \in \Omega} R(f)$$

$$:= \arg \inf_{f \in \Omega} \mathbb{E}_{\mathbf{xy}}[(y - f(\mathbf{x}))^2]$$

$$:= \mathbb{E}_{y|\mathbf{x}}[y]$$

Proof is detailed in notes.

VISUALIZING **EMPIRICAL** VS **POPULATION** RISK



EXCESS RISK

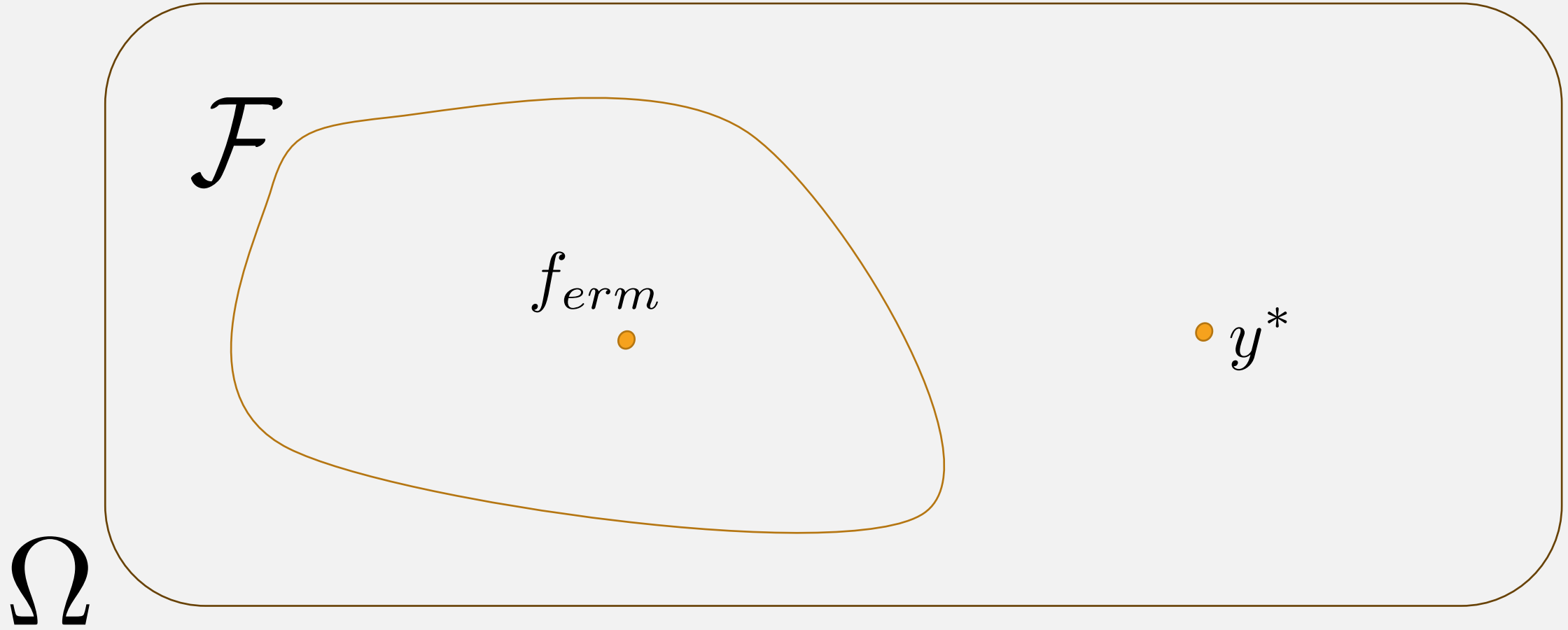
$$R(f_{erm}) - R(y^*)$$

Difference between population risk of ERM, and the population risk of the Bayes model.

Assuming we can optimize our model perfectly to the training data, this measures how much extra risk we have, above that of the "perfect" model.

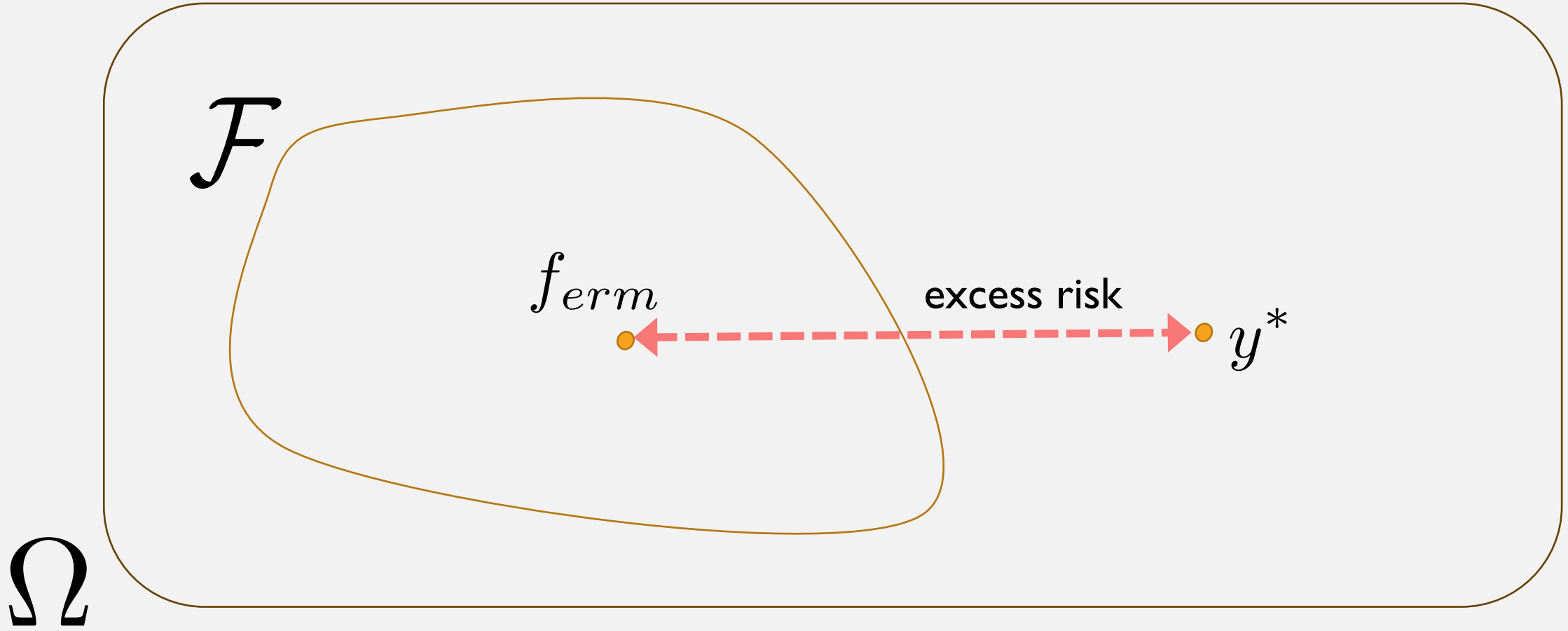
DECOMPOSING THE EXCESS RISK

$$R(f_{erm}) - R(y^*)$$



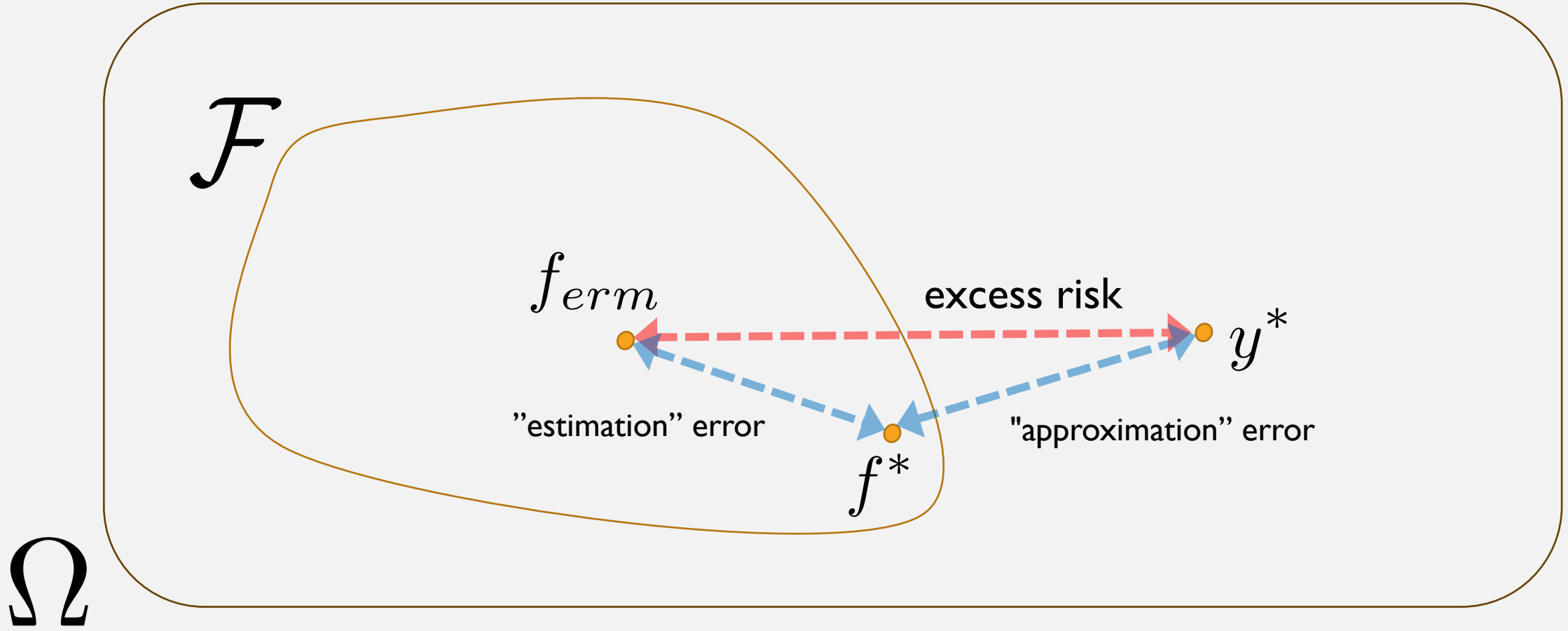
DECOMPOSING THE EXCESS RISK

$$R(f_{erm}) - R(y^*)$$



DECOMPOSING THE EXCESS RISK

$$R(f_{erm}) - R(y^*)$$



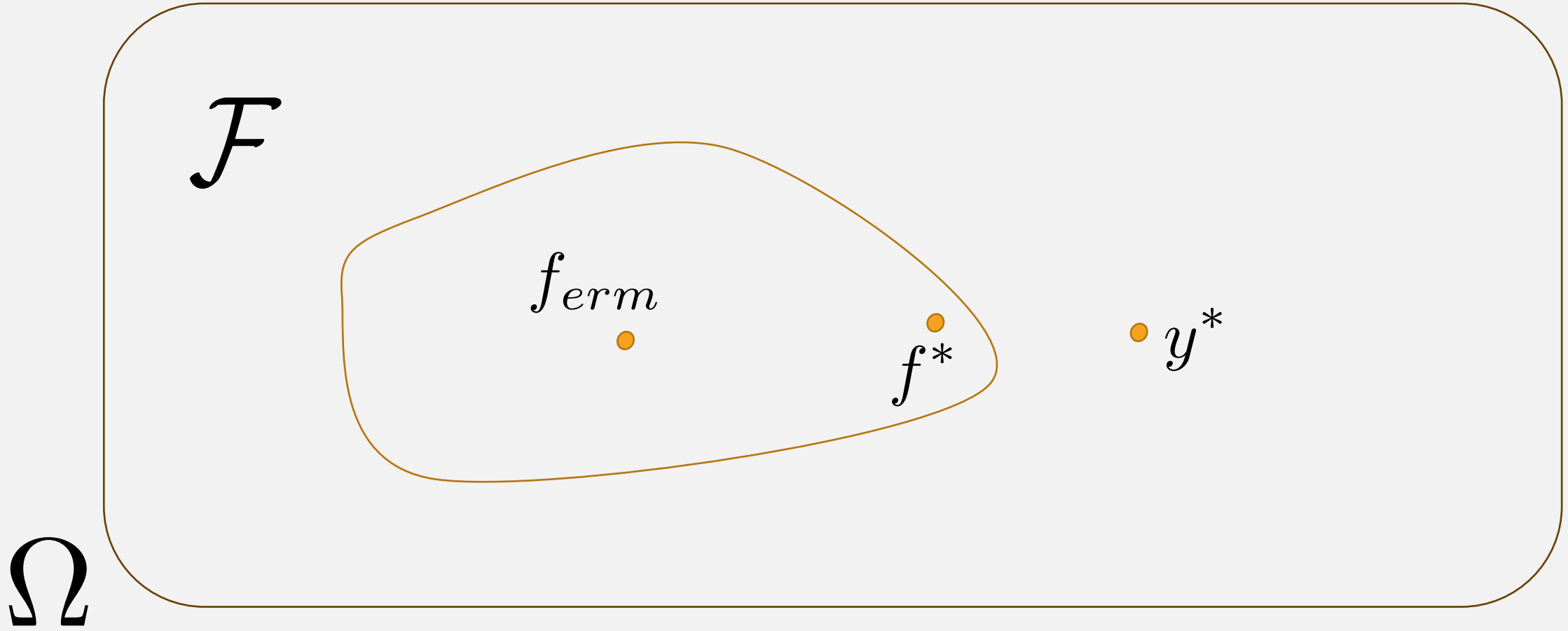
APPROXIMATION VS ESTIMATION ERROR

$$R(f_{erm}) - R(y^*) = \underbrace{R(f_{erm}) - R(f^*)}_{\text{Estimation}} + \underbrace{R(f^*) - R(y^*)}_{\text{Approximation}}$$

Approximation error … error due to having a restricted model family, unable to represent the Bayes model.

Estimation error … error due to having a small sample, where empirical risk is a poor estimate of population risk.

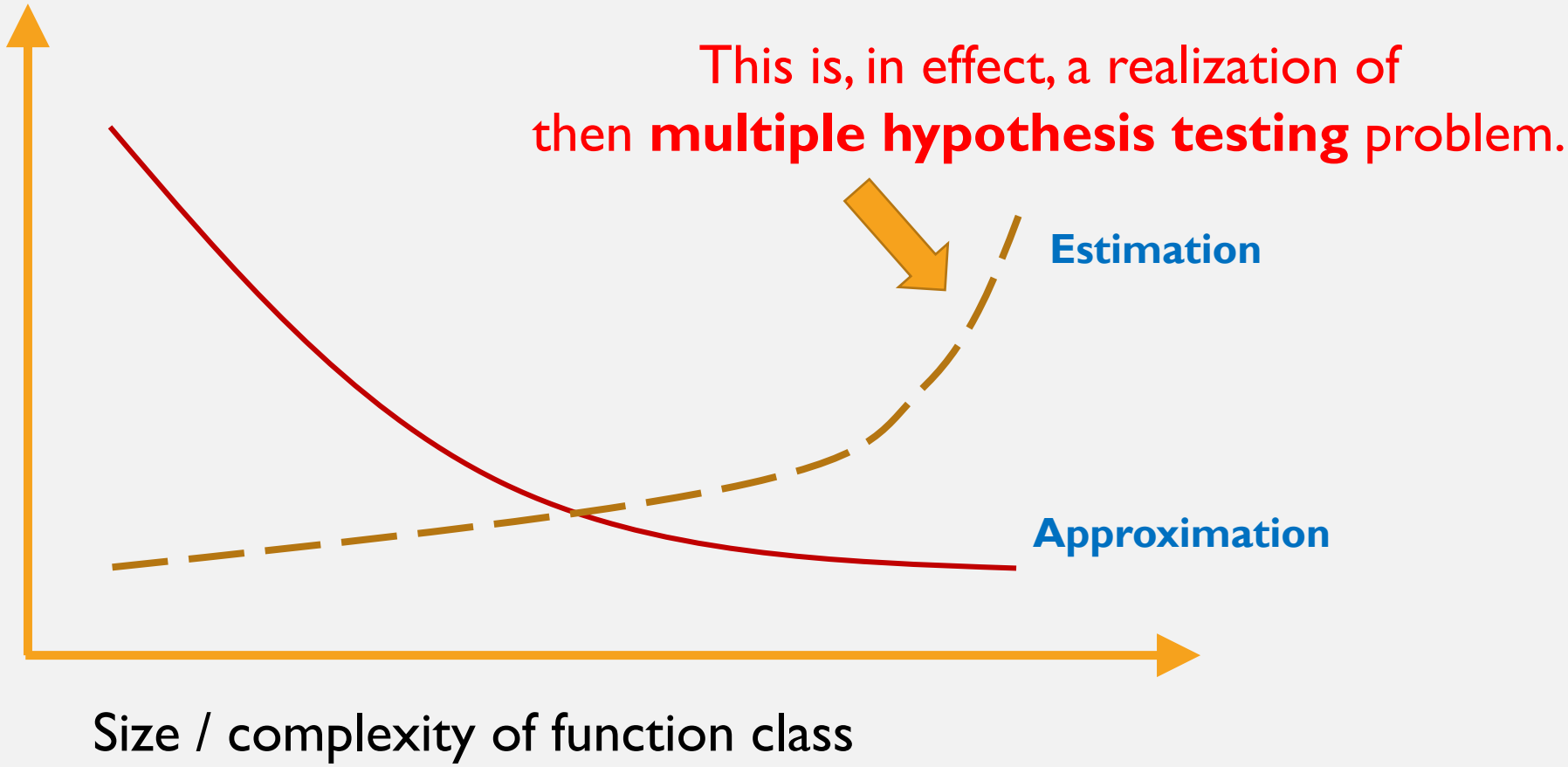
WHAT IF I ALLOW THE FUNCTION CLASS TO GROW?



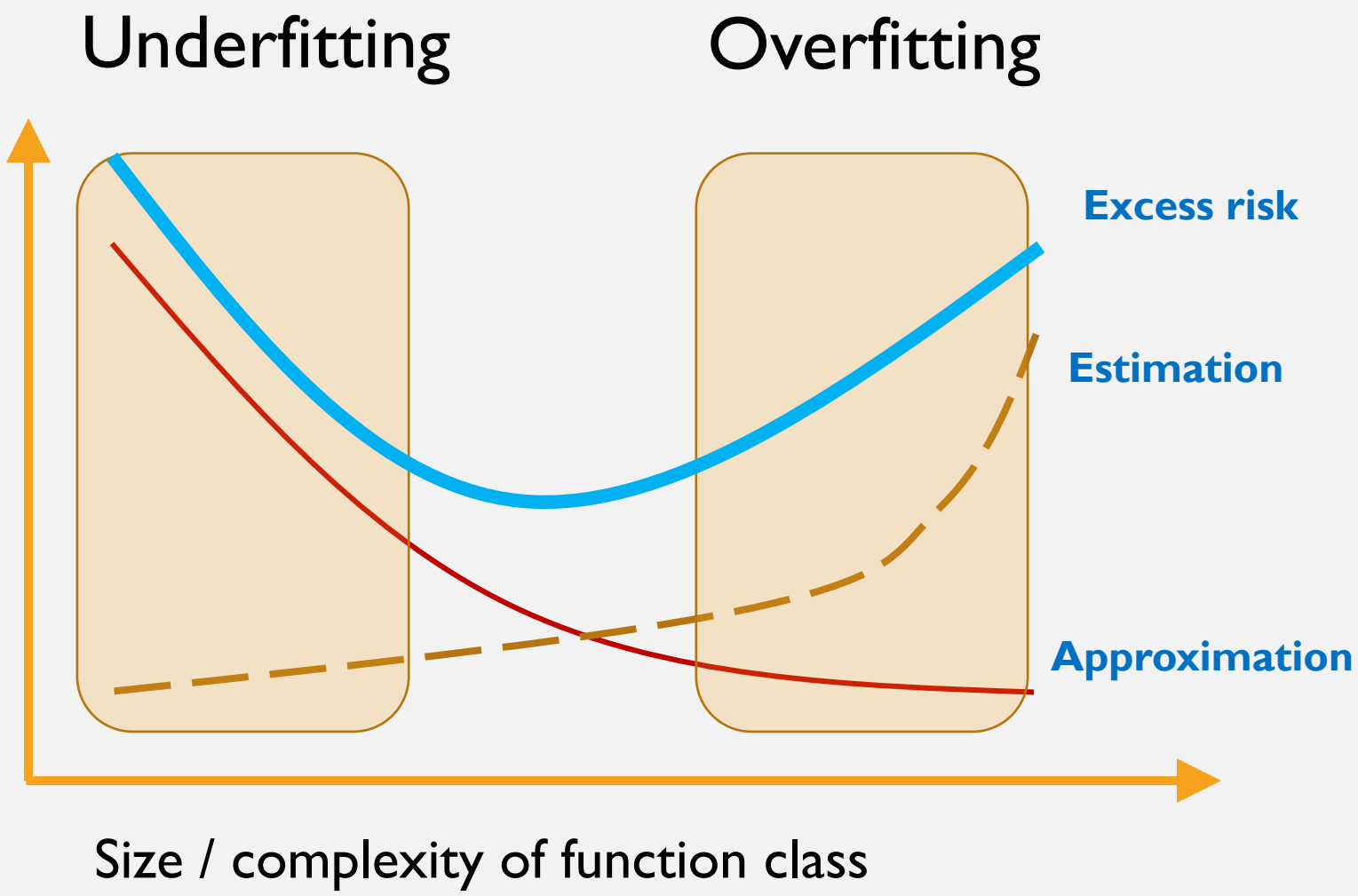
e.g. making the neural network BIGGER, more layers, more non-linearities.

APPROXIMATION VS ESTIMATION ERROR

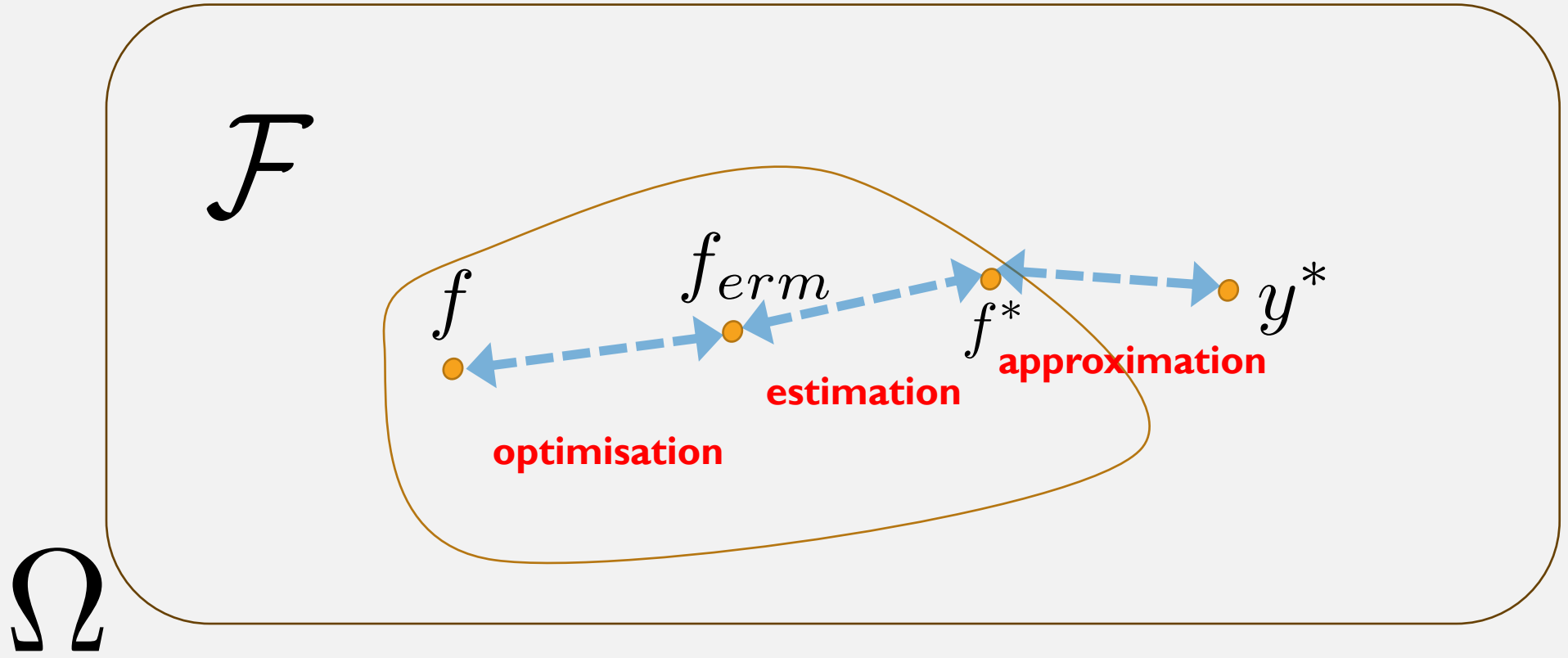
Really big neural nets have effectively zero approximation error.
But it's a tradeoff. As one goes down, the other goes up.



APPROXIMATION VS ESTIMATION ERROR



WE ASSUMED WE COULD FIND THE EMPIRICAL RISK MINIMIZER..



$$\underbrace{R(f) - R(y^*)}_{\text{excess risk of } f} = \underbrace{R(f) - R(f_{erm})}_{\text{optimisation error}} + \underbrace{R(f_{erm}) - R(f^*)}_{\text{estimation error}} + \underbrace{R(f^*) - R(y^*).}_{\text{approximation error}}$$

TODAY WE DID…

By the end of this week you should be able to define,
and use the following concepts correctly in a conversation.

- the loss, empirical risk, and population risk of a model
- the empirical/population risk minimizer
- the Bayes model
- the optimal model within a model family/class
- the approximation error
- the estimation error
- the optimisation error

SEE YOU IN THE STUDY SESSIONS (OR NEXT WEEK).