UTS 2019

MACHINE LEARNING – TRAINING AND TEST

SHORT COURSE

- Models learned from data WILL work if you use an appropriate model family and have a lot of samples in your training dataset.

- In virtually all practical cases, you still need to double-check your learned model on test data to verify.

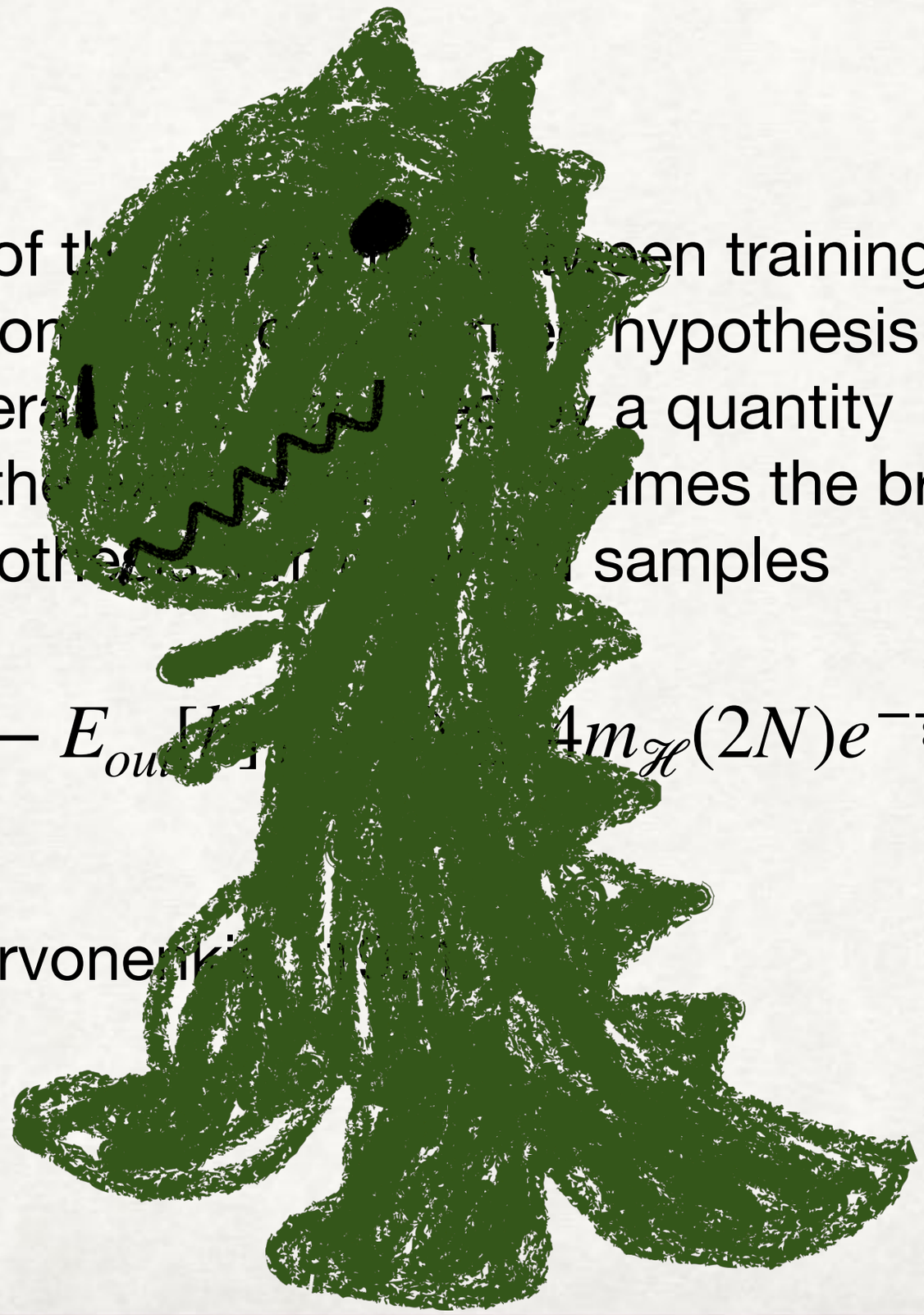
GUARANTEED BOUND OF RISK

- The probability of the difference between training error and generalisation error for a learned hypothesis exceeding a tolerance is bounded by a quantity proportional to the exponential of -N times the break point of the hypothesis family for 2N samples

- $P(\sup_{h \in \mathcal{H}} |E_{in}[h] - E_{out}[h]| > \epsilon) \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}$

- Vapnik and Chervonenkis, 1971

GUARANTEED BOUND OF RISK



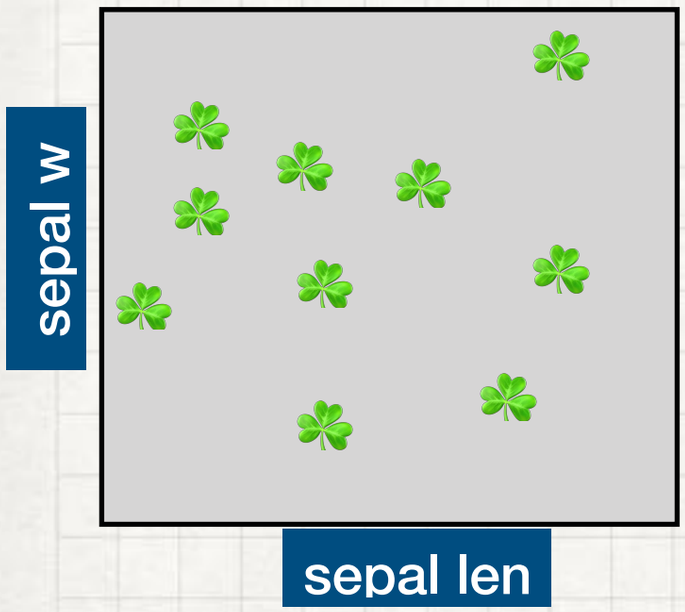
- The probability of th[illegible]een training error and generalisation [illegible] hypothesis exceeding a toler[illegible] a quantity proportional to th[illegible]imes the break point of the hypoth[illegible] samples

- $P(\sup_{h \in \mathcal{H}} |E_{in}[h] - E_{out}[h]| [illegible] 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$

- Vapnik and Chervonenki[illegible]

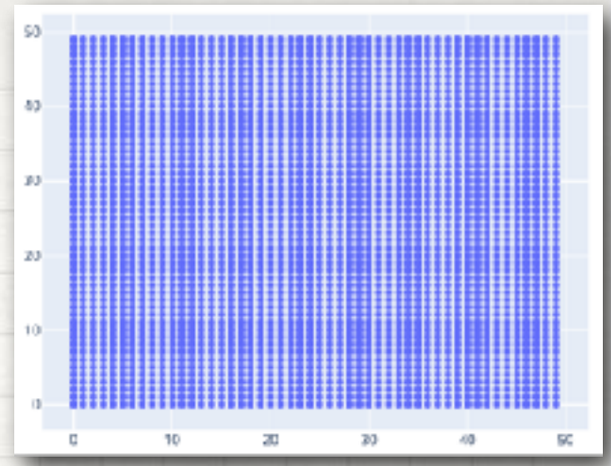
REVIEW

Binary Classification
Identify Target Concept



A complete table of all possible hypotheses

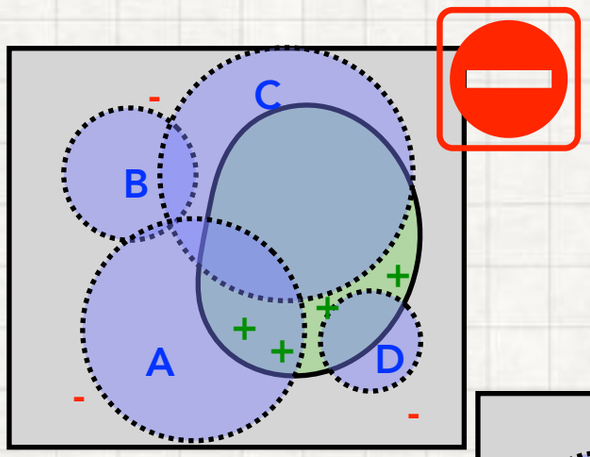
	X1	L	L	L	M	M	M	H	H	H
Y|X	X2	L	M	H	L	M	H	L	M	H
q0		0	0	0	0	0	0	0	0	0
q1		0	0	0	0	0	0	0	0	1
q2		0	0	0	0	0	0	0	1	0
q3		0	0	0	0	0	0	0	1	1
q4		0	0	0	0	0	0	1	0	0
q5		0	0	0	0	0	0	1	0	1
q6		0	0	0	0	0	0	1	1	0
q7		0	0	0	0	0	0	1	1	1
...										



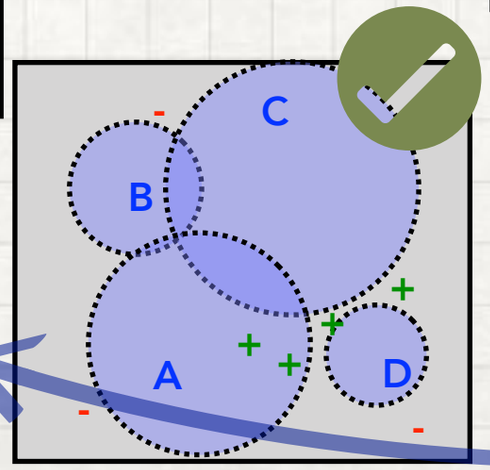
$2^{2500}=$
375828023454801203683362418972386504867736551759258677056523839782231681498337708535732725752658844333702457497526057760309227891351617765651907310968780236464694043316236562146724416478591131832593729111221580180531749232777515579969899075142213969117994877343802049421624954402214529390781647563339535024772584901607666862982567918622849636160208877365834950163790188523026247440507390382032188892386109905869706753143243921198482212075444024333665547868565593896895856381265823772240377217022399914414660261857526515029364722809110185003203754963367499515695215418504417479258440662952796718726052857925526601307020479982183347493563216774695296825517658582675027158940078877272500707803502629523772120288422974862635978797921763382209326174489509376

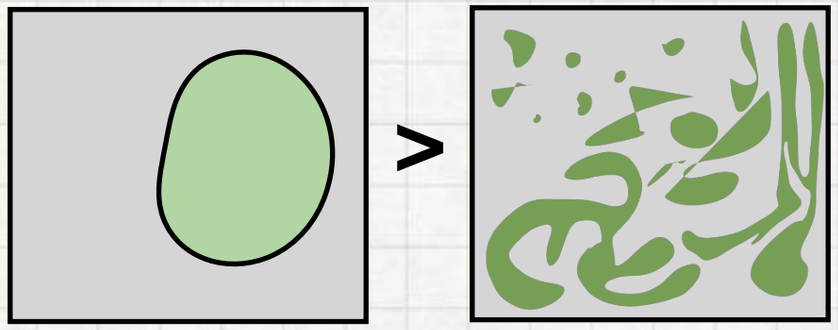
Is large

And doesn't work

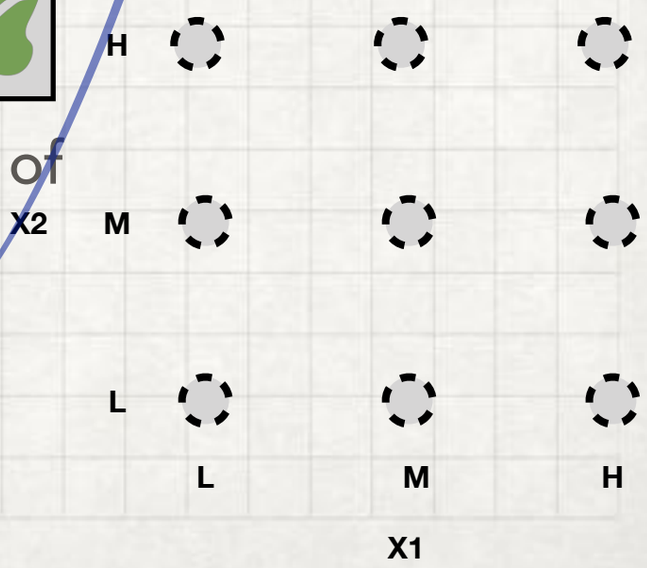


Train: use data to evaluate hypothesis





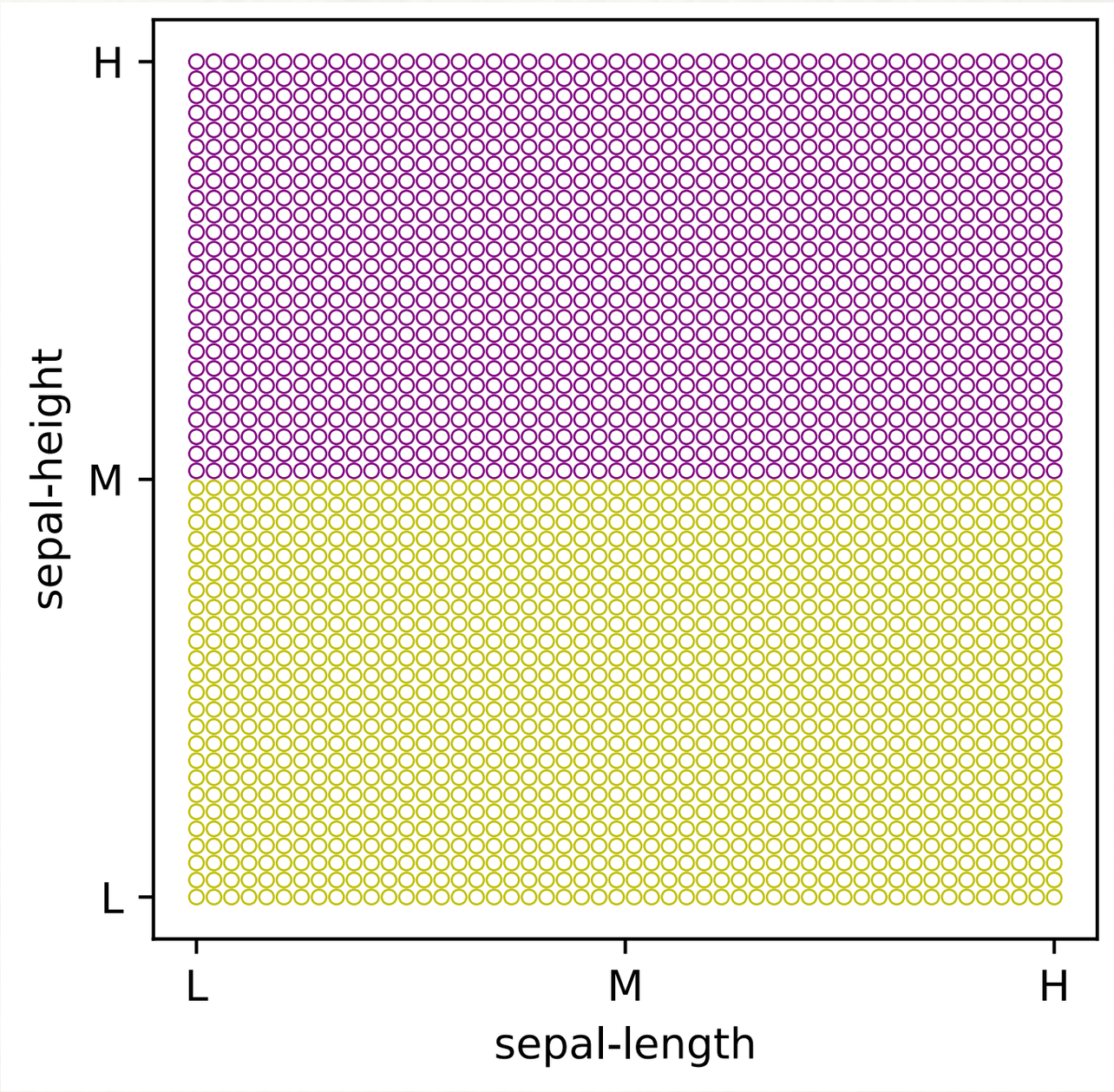
Preference of hypotheses of "regular" data-target relationship



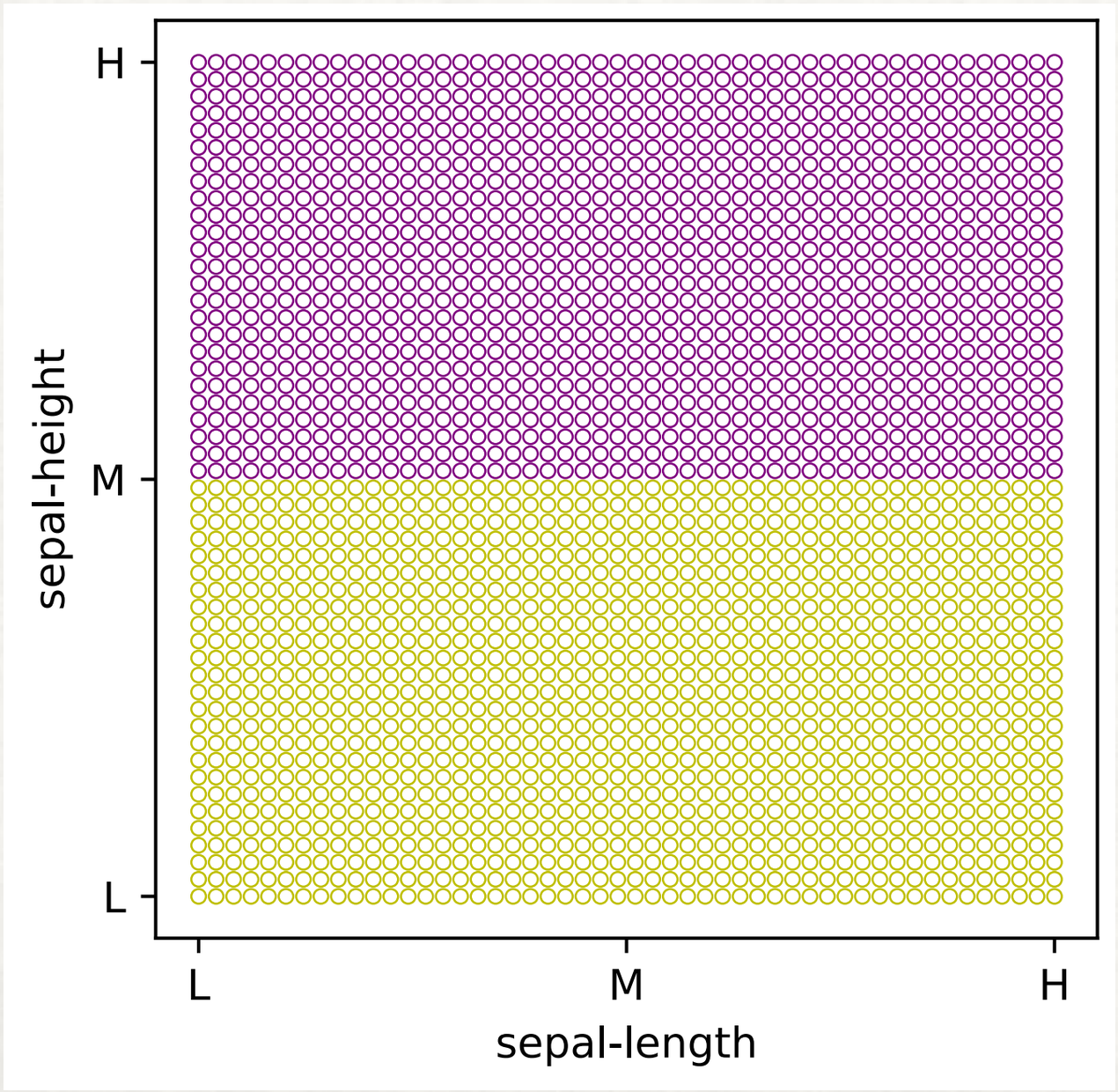
MOD1
EVALUATION HYPOTHESIS VIA DATA
HOEFFDING'S INEQUALITY

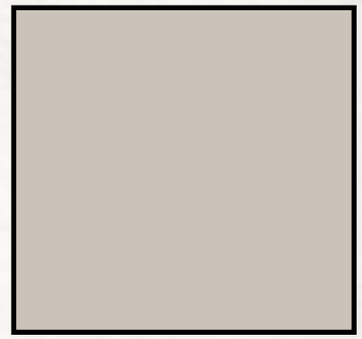
AN EXAMPLE OF ASSESSING A HYPOTHESIS

- Consider the hypothesis shown on right: “up: class-0; down: class-1”.
- We are interested in its overall performance.



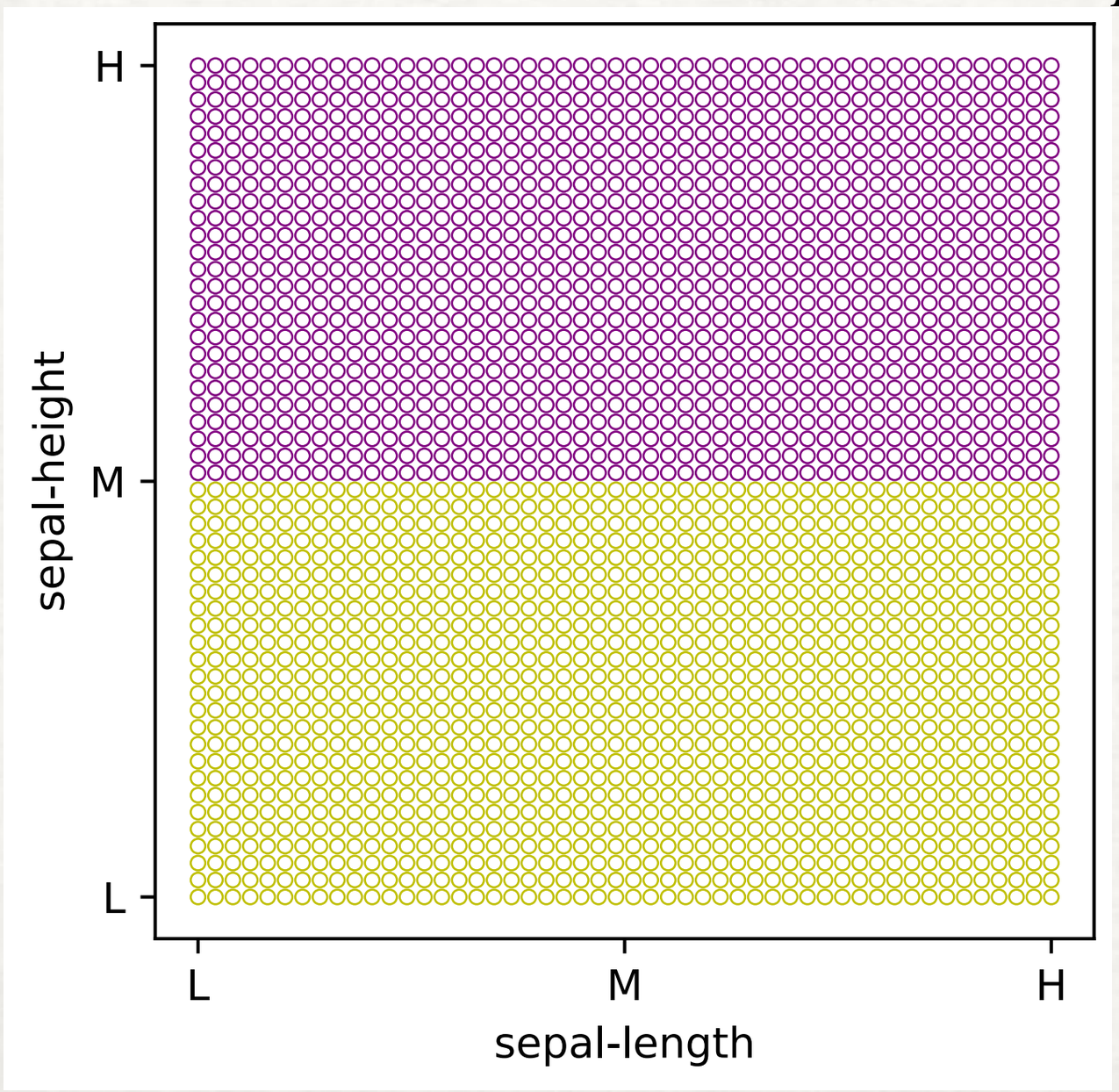
GOLDEN STANDARD OF ASSESSMENT





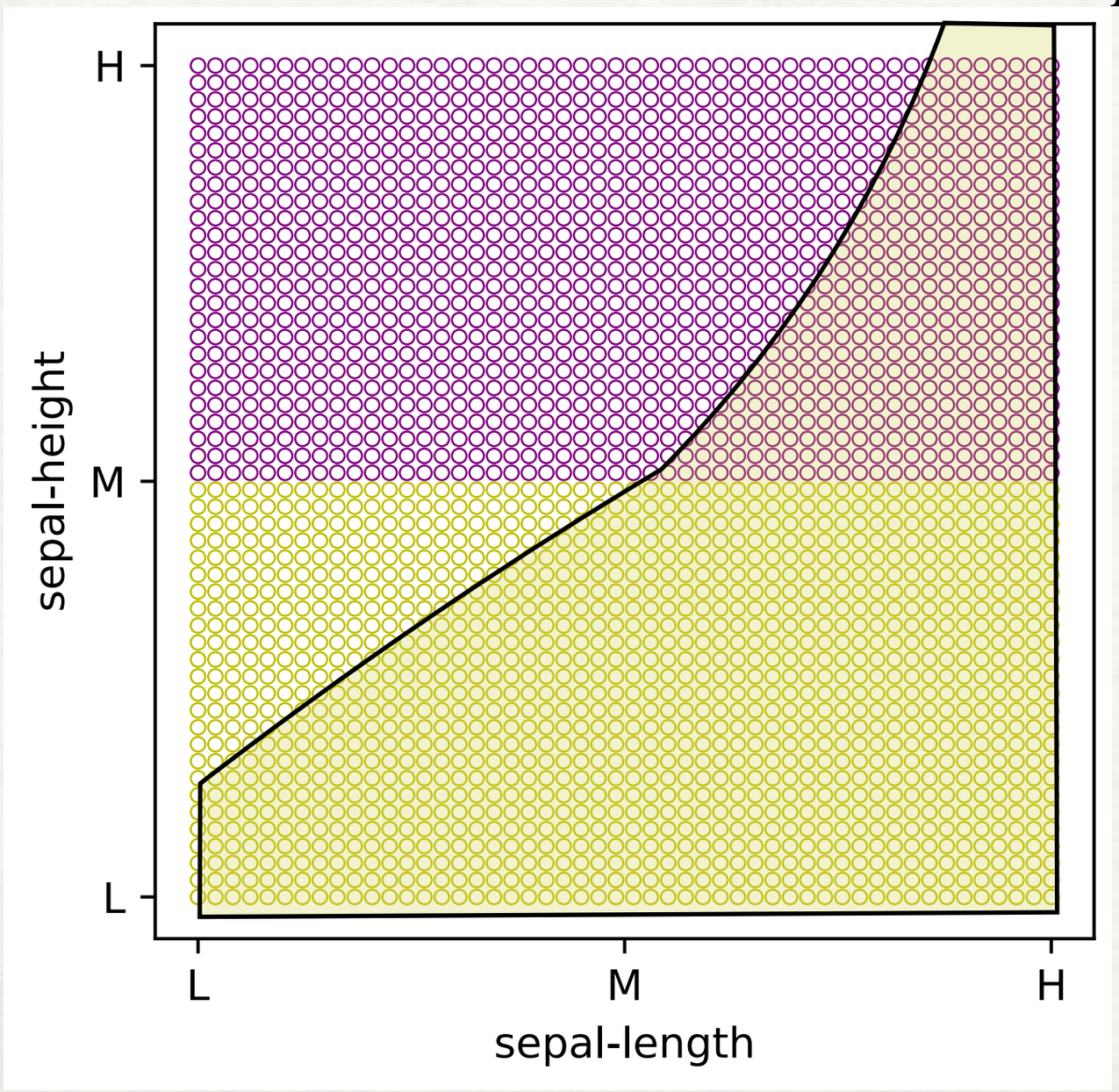
GOLDEN STANDARD OF ASSESSMENT

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

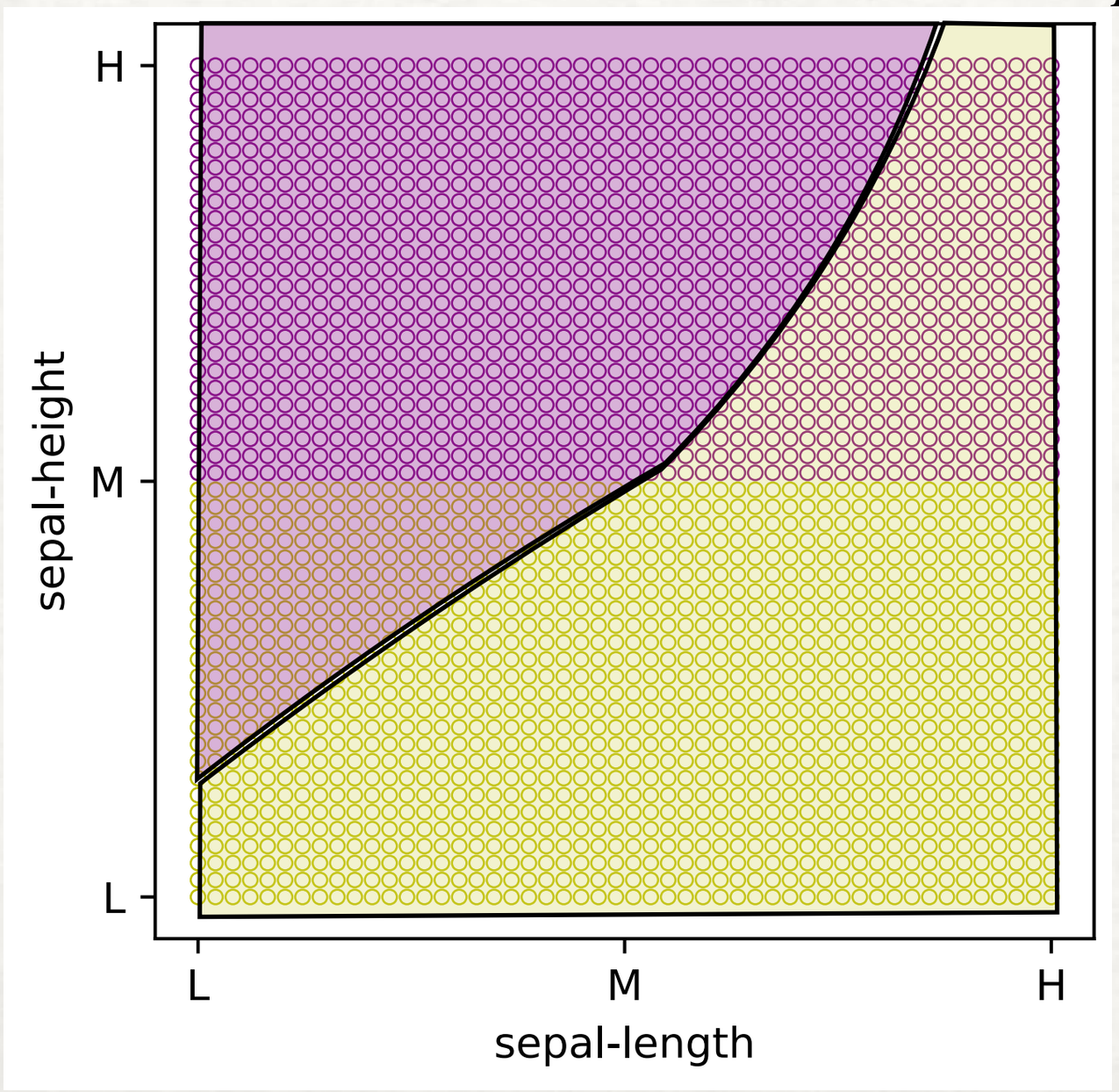


GOLDEN STANDARD OF ASSESSMENT

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

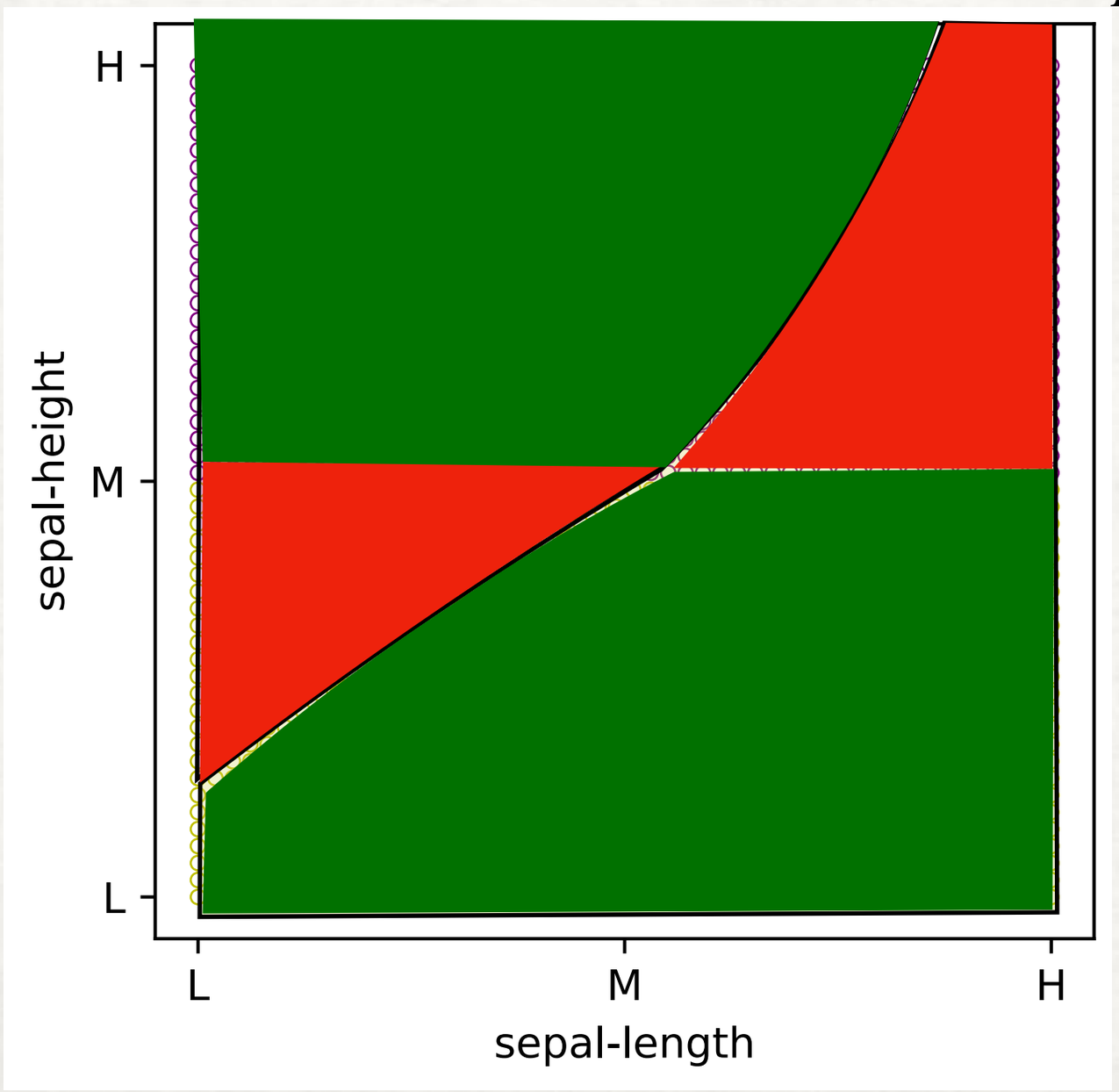


GOLDEN STANDARD OF ASSESSMENT



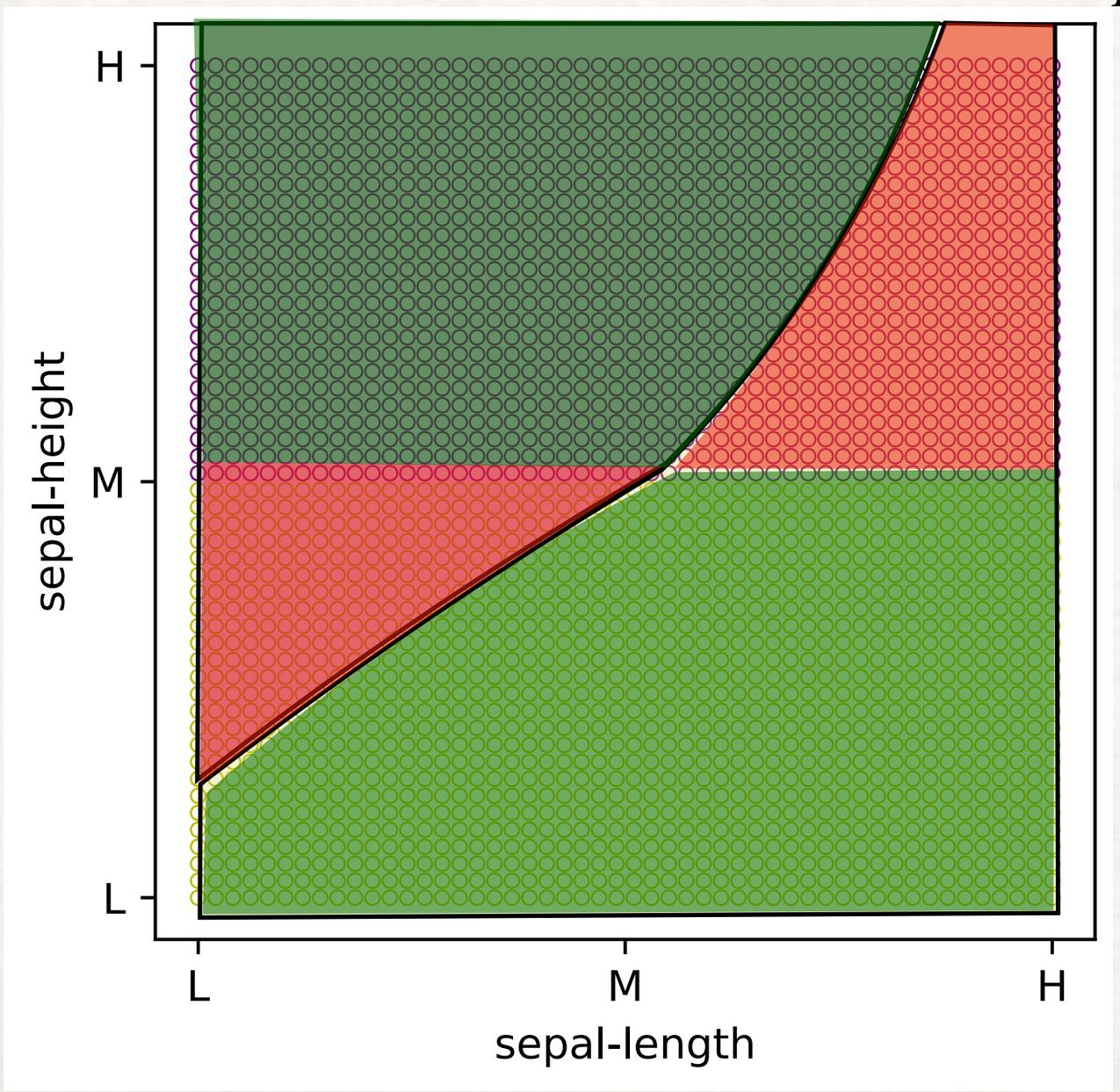
$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

GOLDEN STANDARD OF ASSESSMENT



$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

GOLDEN STANDARD OF ASSESSMENT



$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

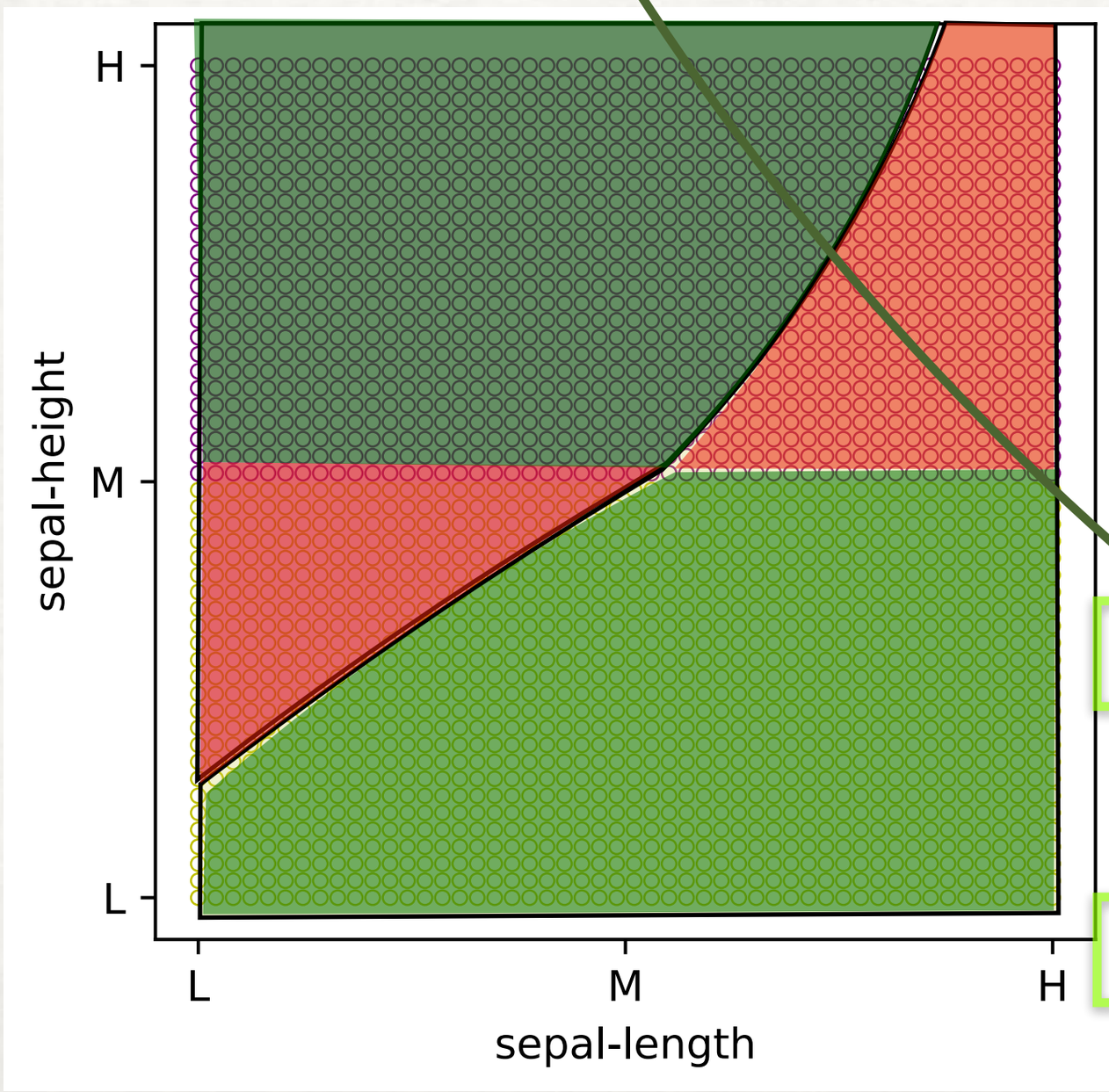
B. Yes, we want.

Can we use E_{out}?

A. Yes

B. No

GOLDEN STANDARD OF ASSESSMENT



$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

B. Yes, we want.

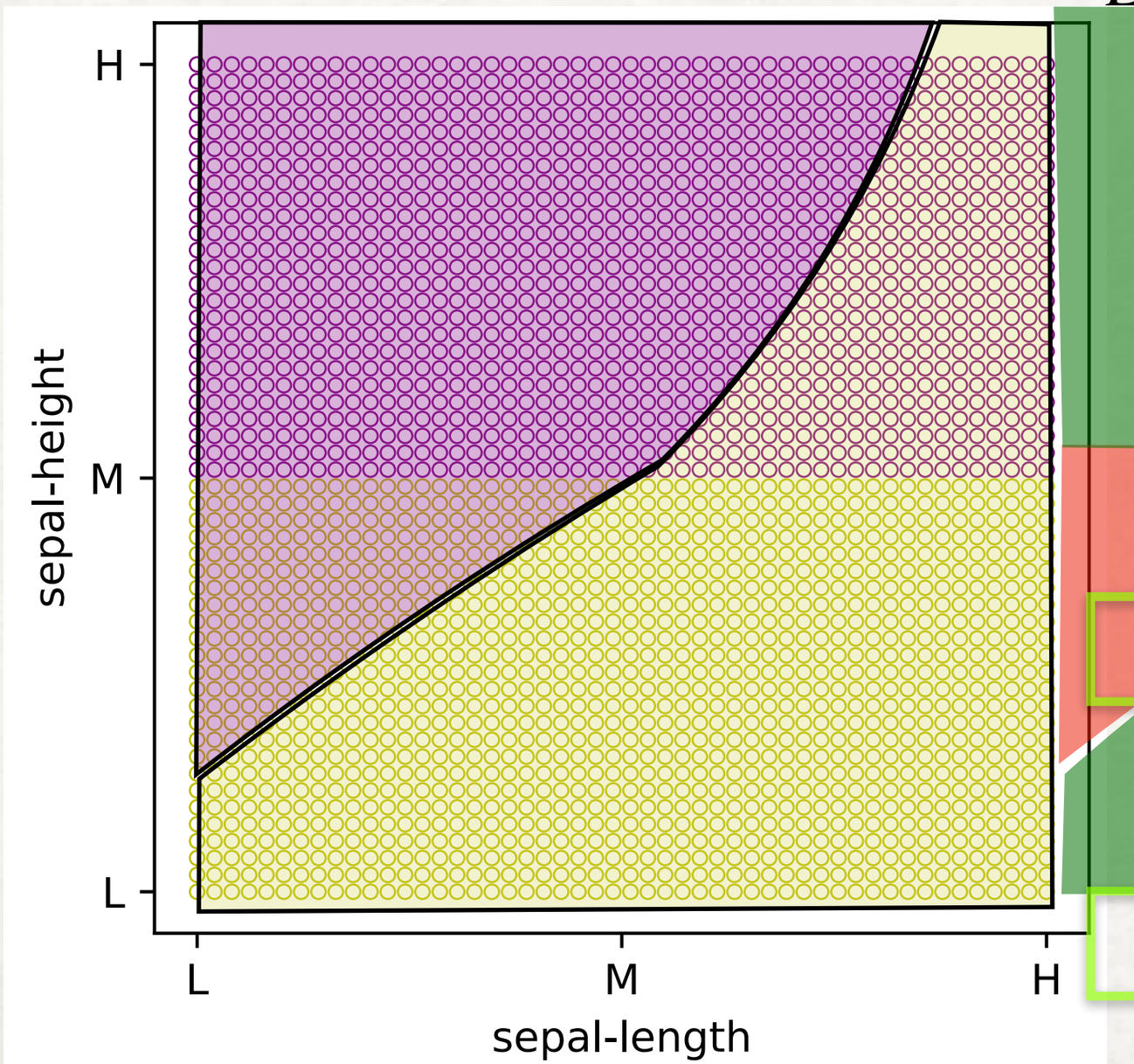
Can we use E_{out}?

A. Yes

B. No

GOLDEN STANDARD OF ASSESSMENT

$$E_{out}[h] = \frac{\int_{x\in\mathcal{X}} [h(x) \neq y(x)]}{\int_{x\in\mathcal{X}}}$$



Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

B. Yes, we want.

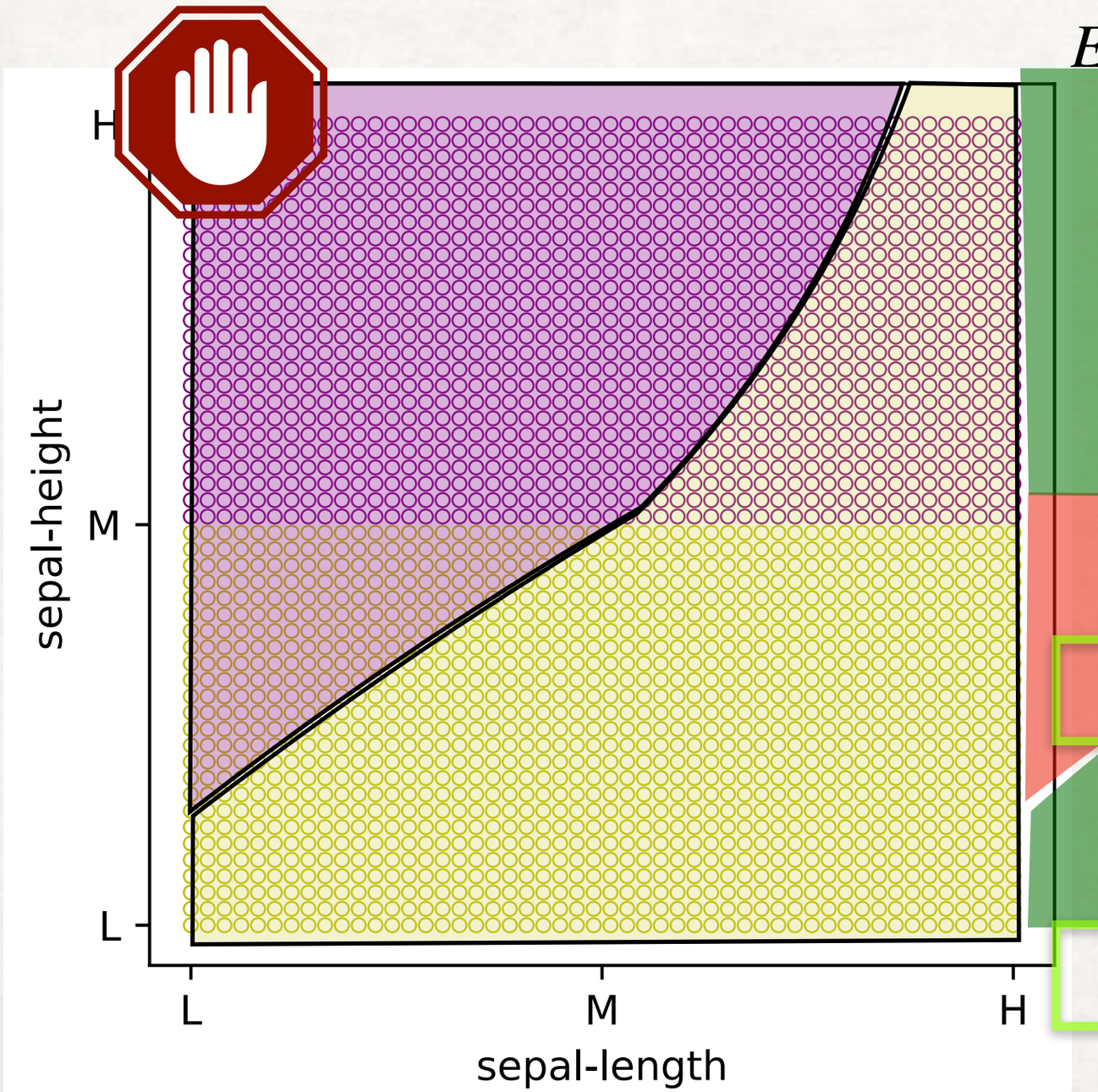
Can we use h?

A. Yes

B. No

GOLDEN STANDARD OF ASSESSMENT

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$



Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

B. Yes, we want.

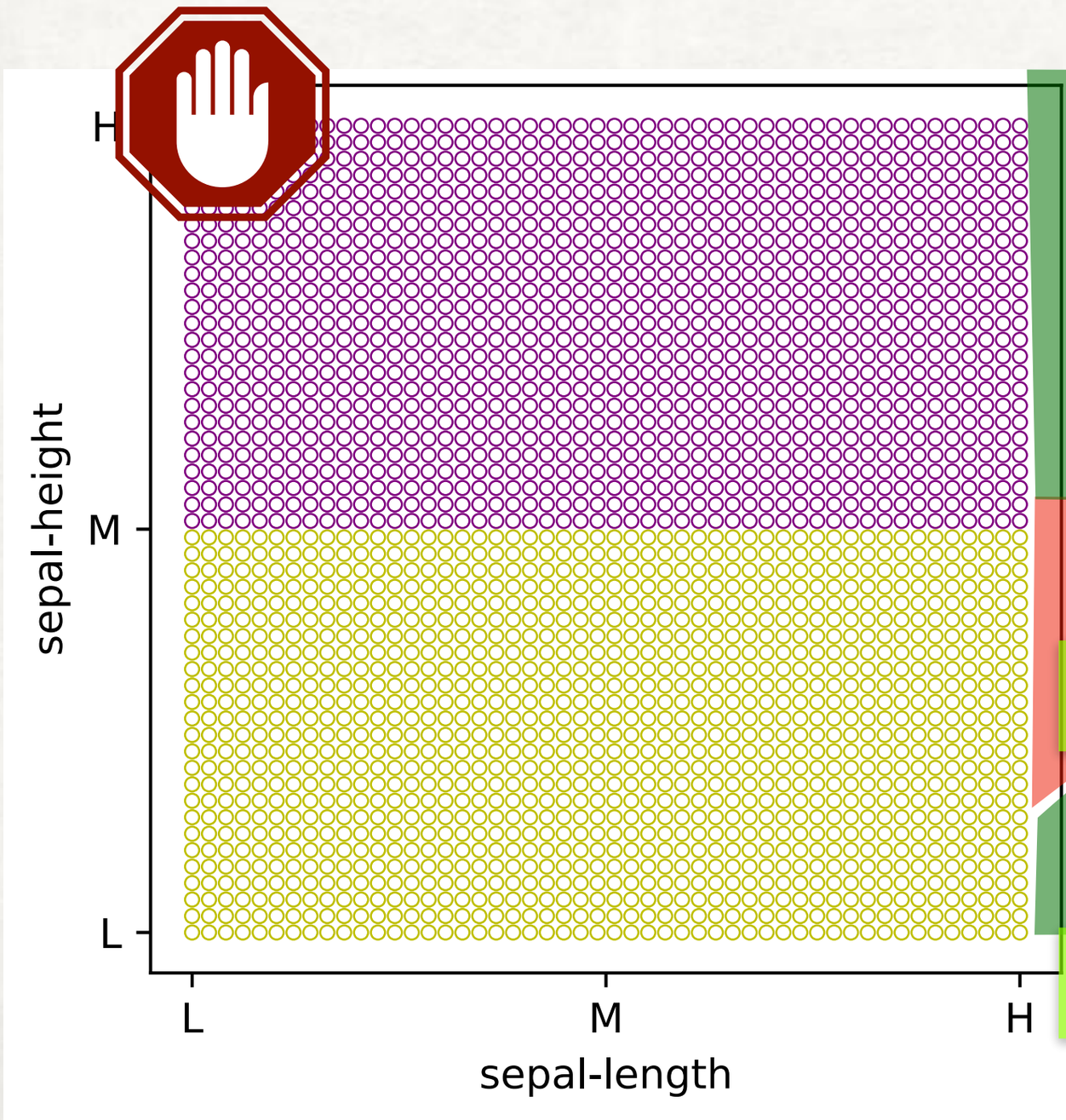
Can we use h?

A. Yes

B. No

GOLDEN STANDARD OF ASSESSMENT

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$



Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

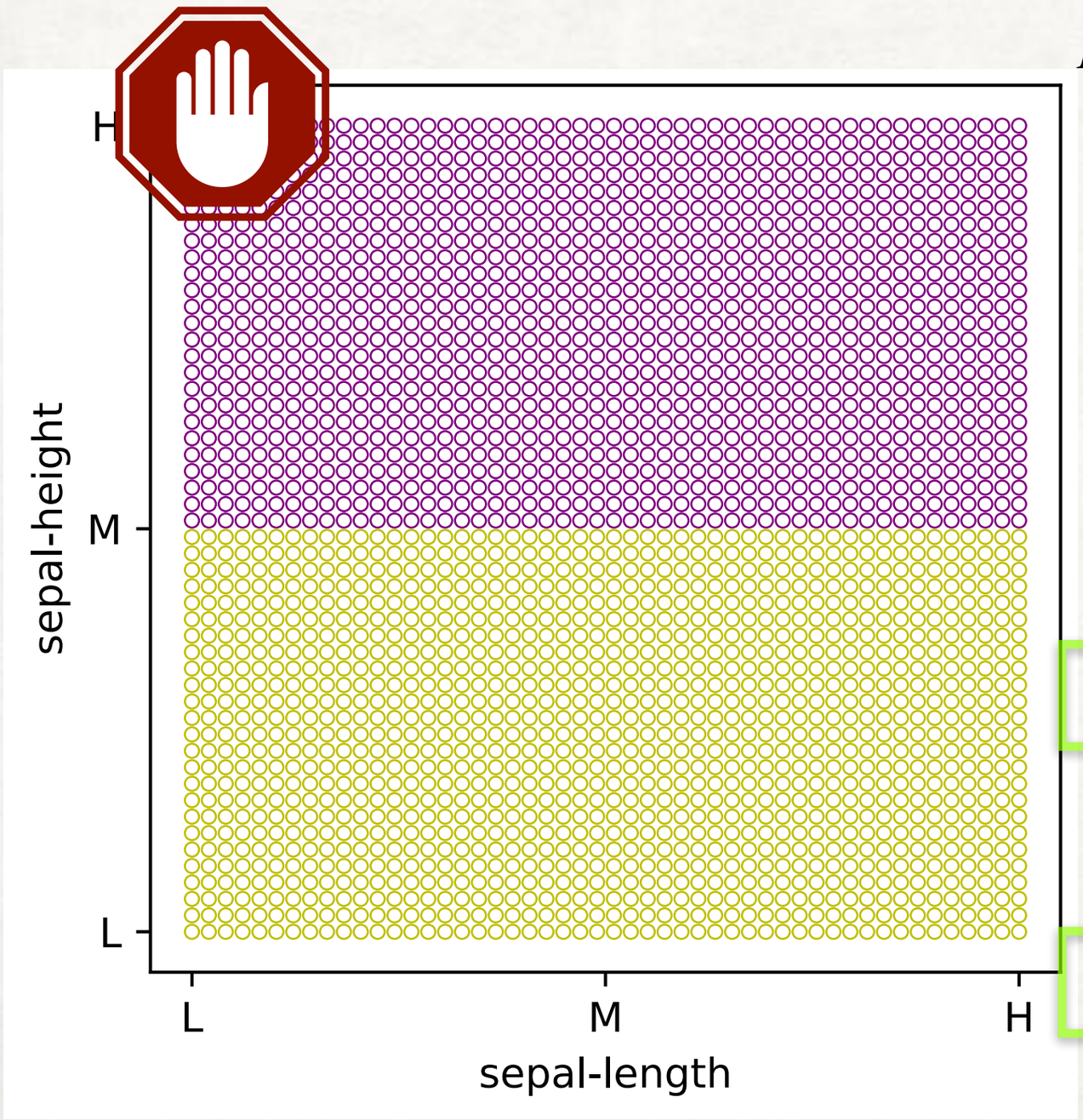
B. Yes, we want.

Can we use h?

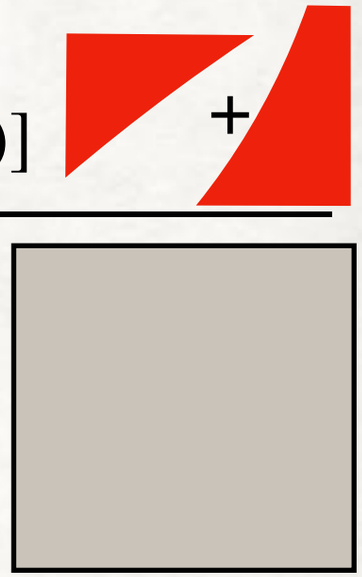
A. Yes

B. No

GOLDEN STANDARD OF ASSESSMENT



$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$



Q: Given an h, do we want to use E_{out} to evaluate h?

A. No, the criterion doesn't consider observation (data).

B. Yes, we want.

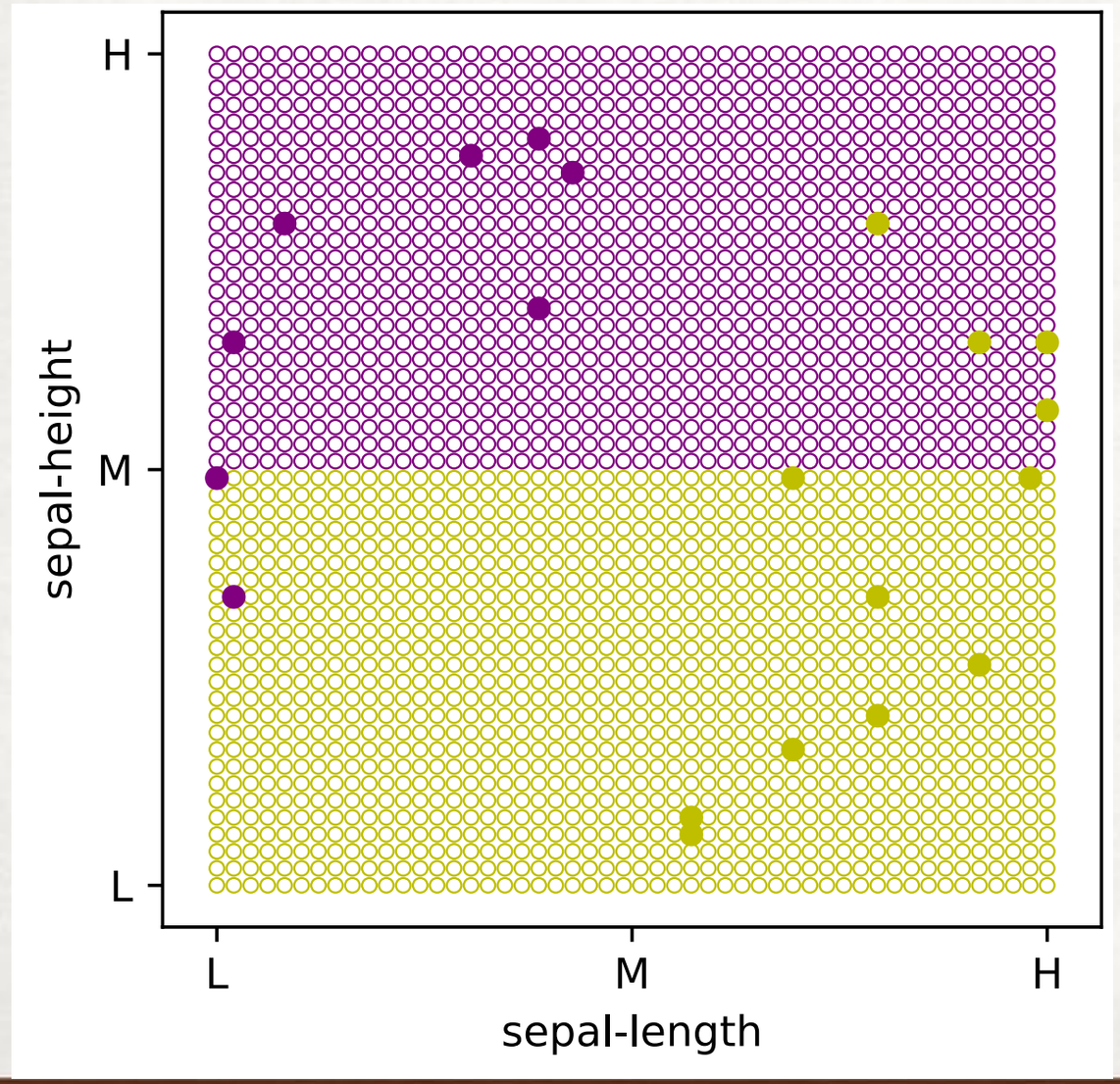
Can we use h?

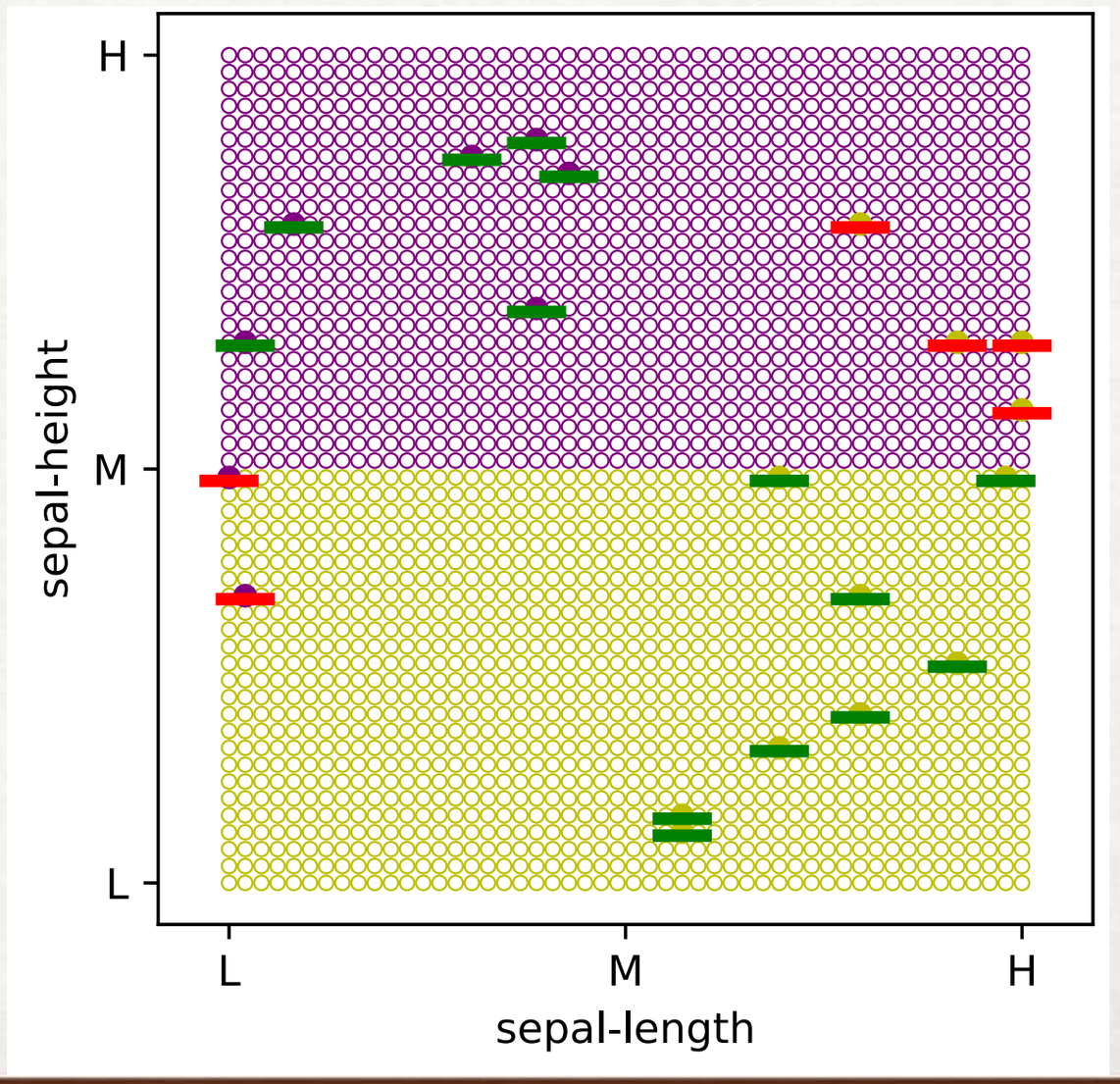
A. Yes

B. No

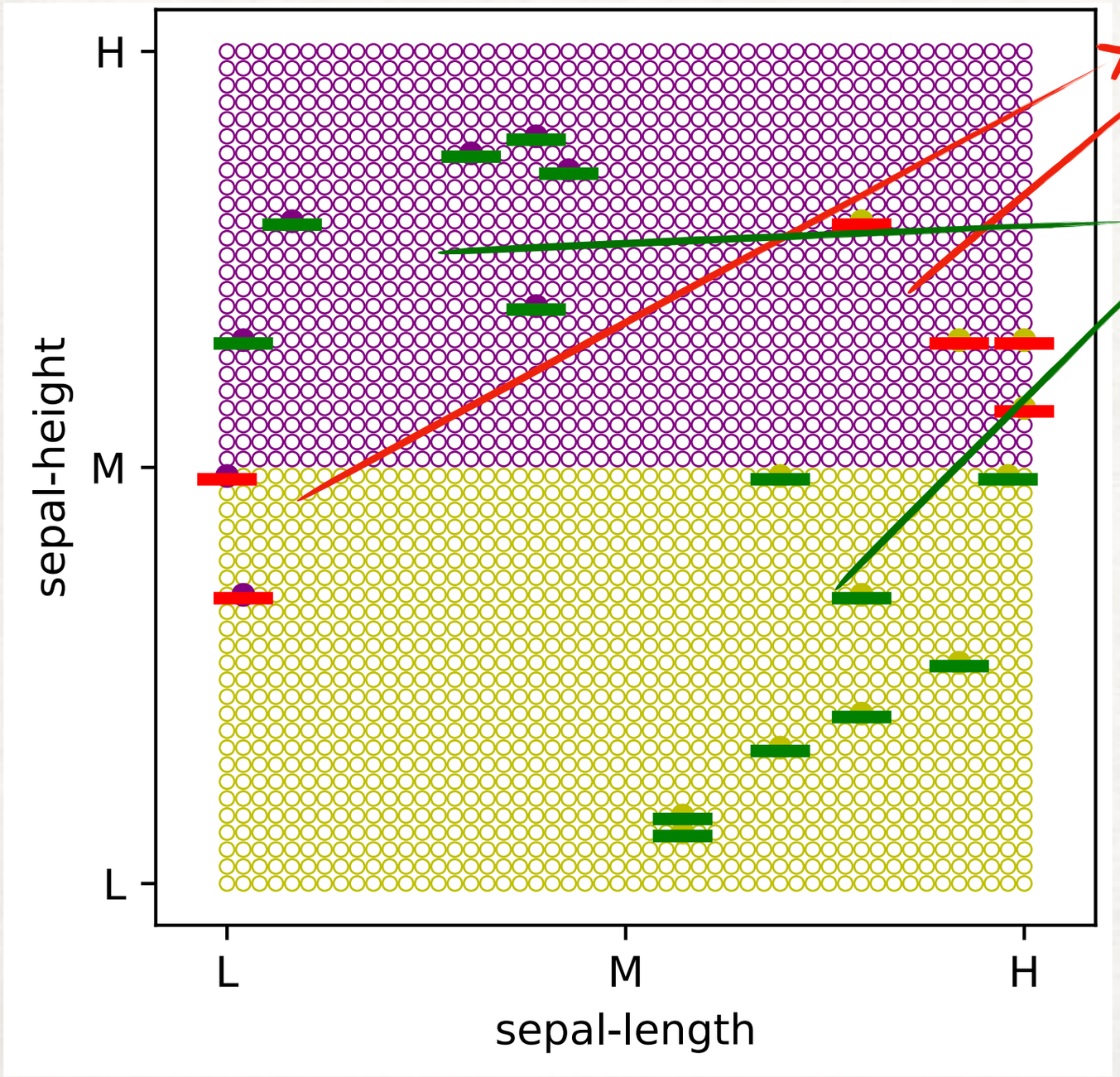
"EMPIRICAL" ASSESSMENT OF A HYPOTHESIS

- We answer the "how well" question empirically — compare the predictions made by the hypothesis at data samples where we have access to the ground-truth answers.





EMPIRICAL ASSESSMENT
IN-SAMPLE VS OUT-OF-SAMPLE ERRORS

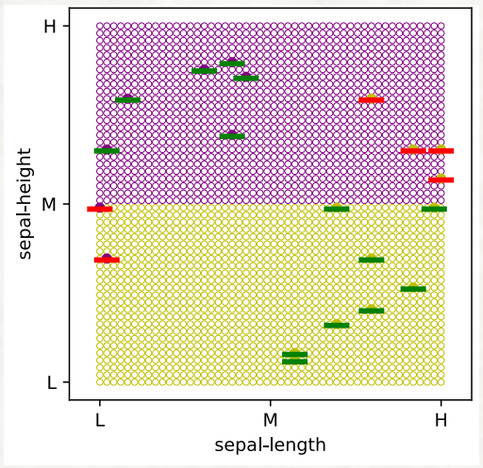


#.ERR = 6

#.CORR = 14

$$E_{in} = 0.3$$

E-IN AND E-OUT



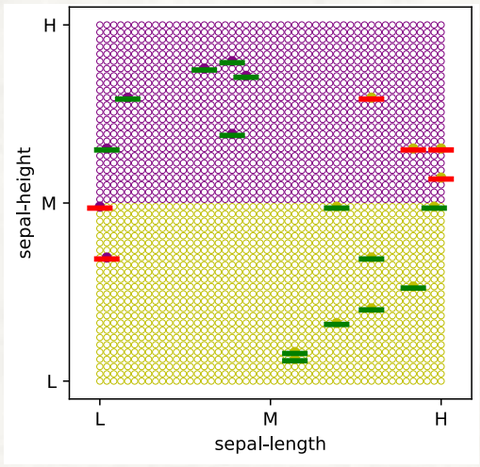
$$E_{in}[h; D_{train}] = \frac{\sum_i h(x_i) \neq y_i}{\sum_i 1 = N}$$

#.ERR = 6

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq y(x)]}{\int_{x \in \mathcal{X}}}$$

+

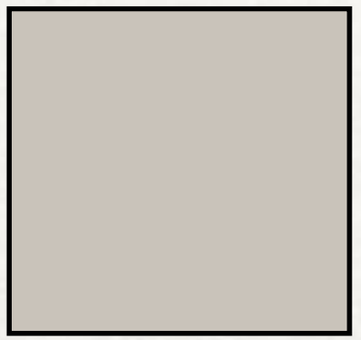
E-IN AND E-OUT



$$E_{in}[h; D_{train}] = \frac{\sum_i h(x_i) \neq y_i}{\sum_i 1 = N}$$

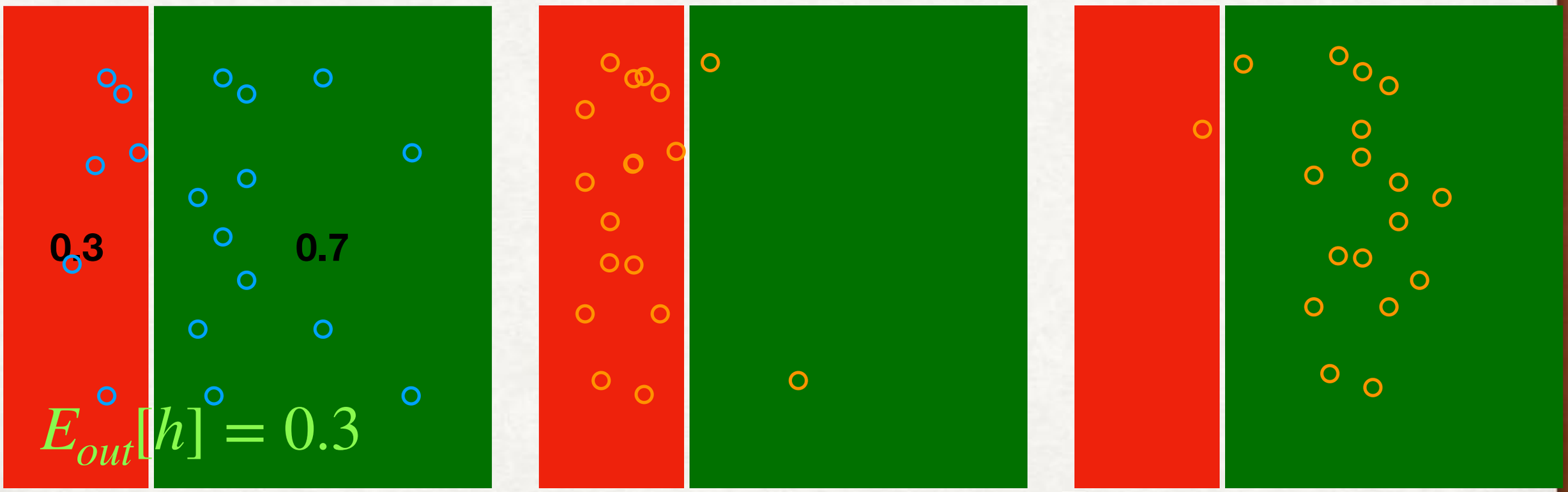
#.ERR = 6

$$E_{out}[h] = \frac{\int_{x \in \mathcal{X}} [h(x) \neq$$  $$]$$  $$+$$

$$\int_{x \in \mathcal{X}}$$ 

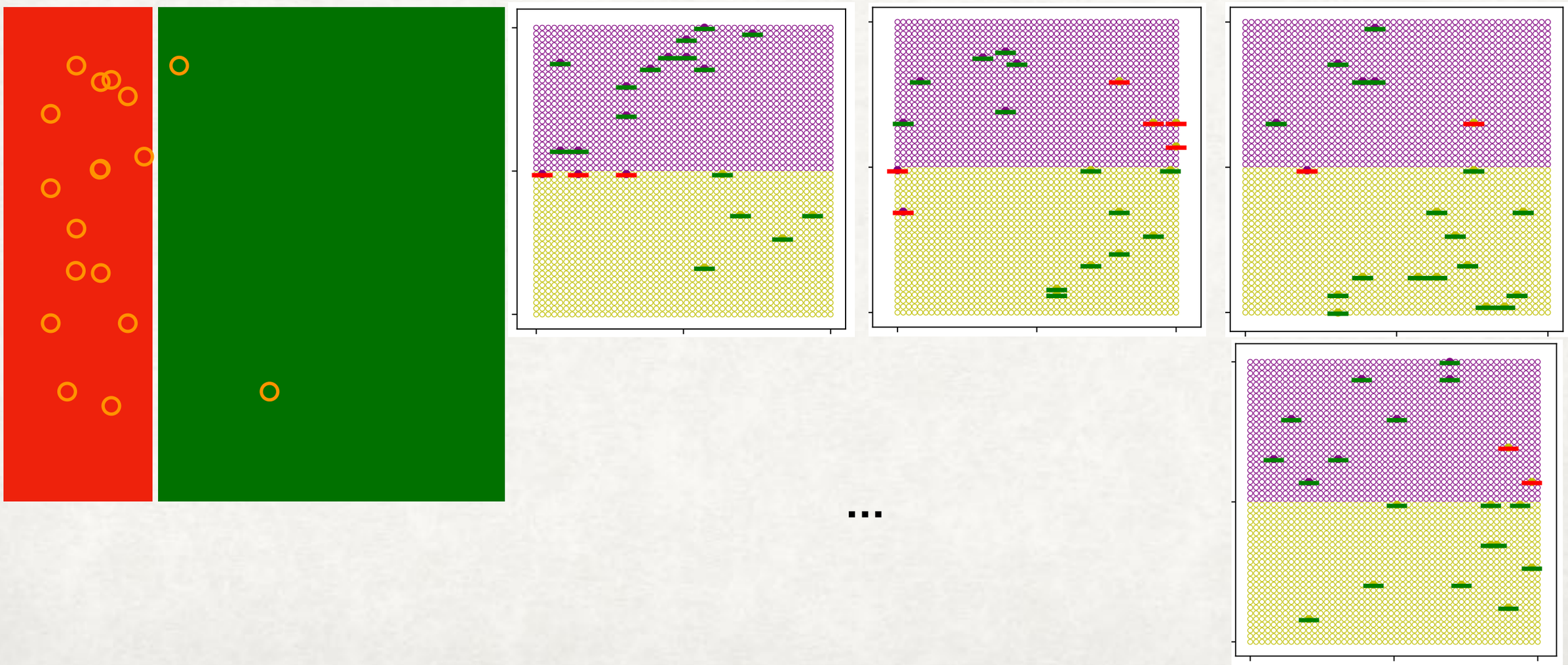
RATIONALE OF EMPIRICAL ASSESSMENT

- Estimate the generalisation error of a hypothesis (red/green ratio) via looking *through* samples in the training dataset.
- For a "typical" training dataset D, $E_{in}[h; D]$ is likely to be close to E_{out}.
- It becomes extremely unlikely to draw such a D^{Bad}, so $E_{in}[h; D^{Bad}]$ is very different from E_{out}.



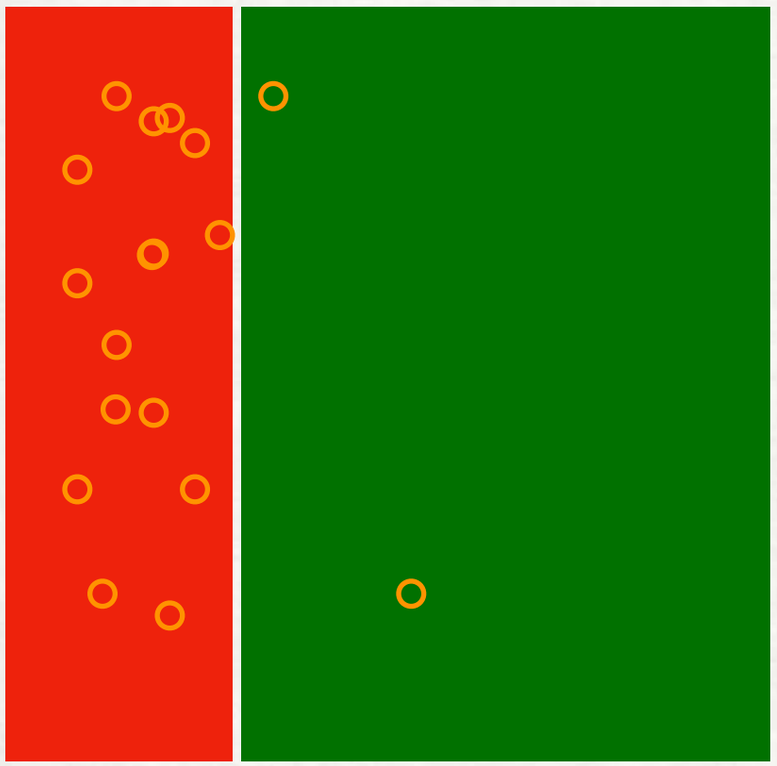
TYPICAL TRAINING DATASETS

- The chance you get a rare D is really rare — when the data set size is big.



HOEFFDING'S INEQUALITY

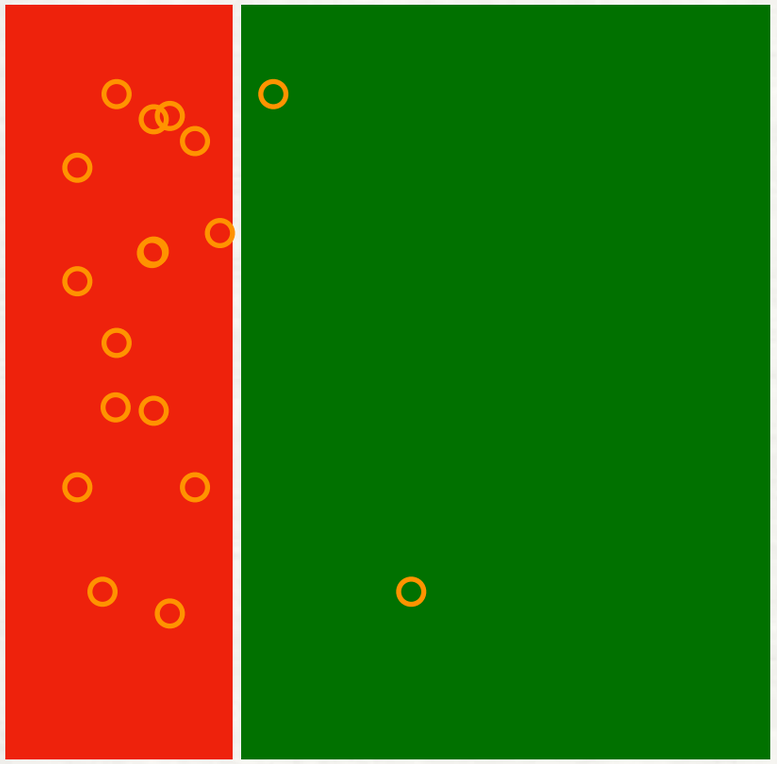
- The chance you get a rare D is really rare — when the dataset size is big.



$$P(|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

HOEFFDING'S INEQUALITY EXPLAINED

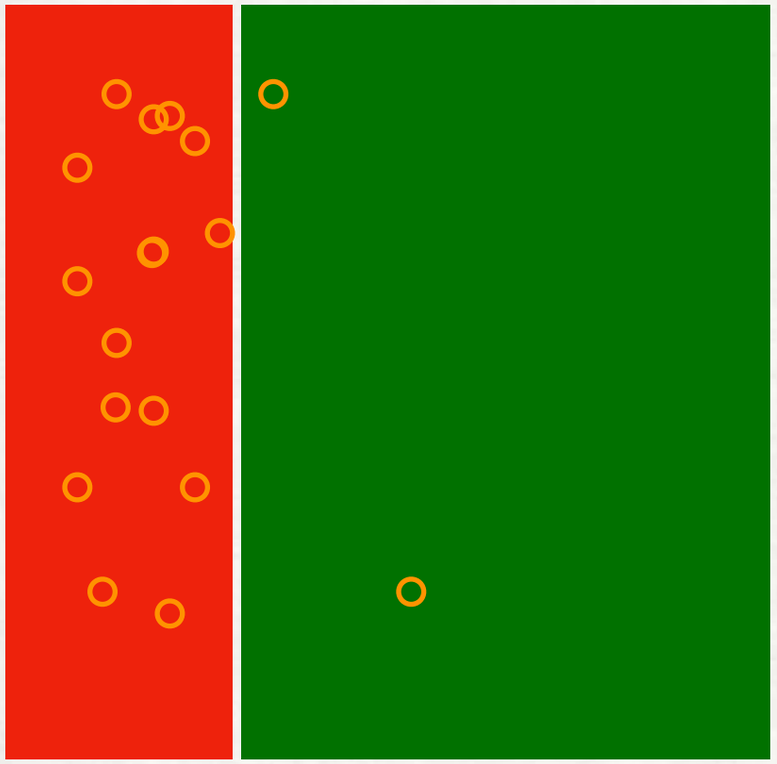
- The chance you get a rare D is really rare — when the data set size is big.



$$P(|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

HOEFFDING'S INEQUALITY EXPLAINED

- The chance you get a rare D is really rare — when the data set size is big.

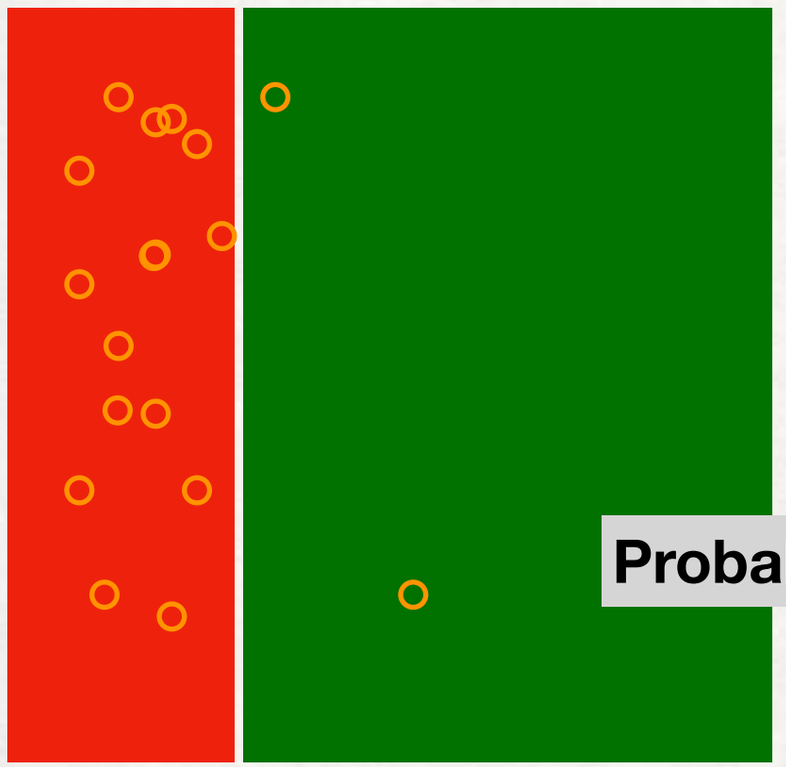


$$P(\,|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Bad Thing Happens:
Evaluation on training data is not reliable

HOEFFDING'S INEQUALITY EXPLAINED

- The chance you get a rare D is really rare — when the data set size is big.



$$P(|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

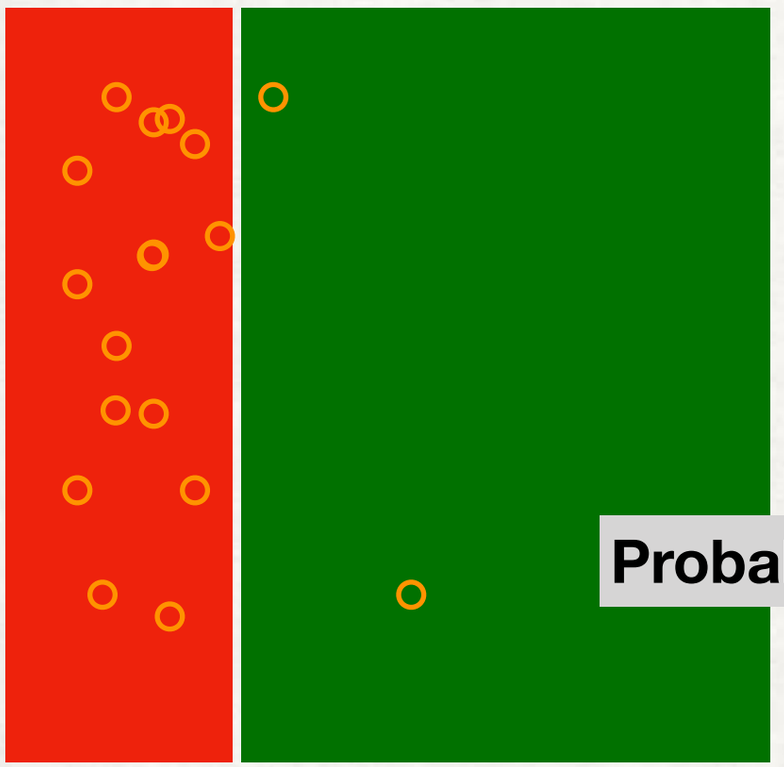
Probability of

Bad Thing Happens: Evaluation on training data is not reliable

drops exponentially

HOEFFDING'S INEQUALITY EXPLAINED

- The chance you get a rare D is really rare — when the data set size is big.



$$P(|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Probability of

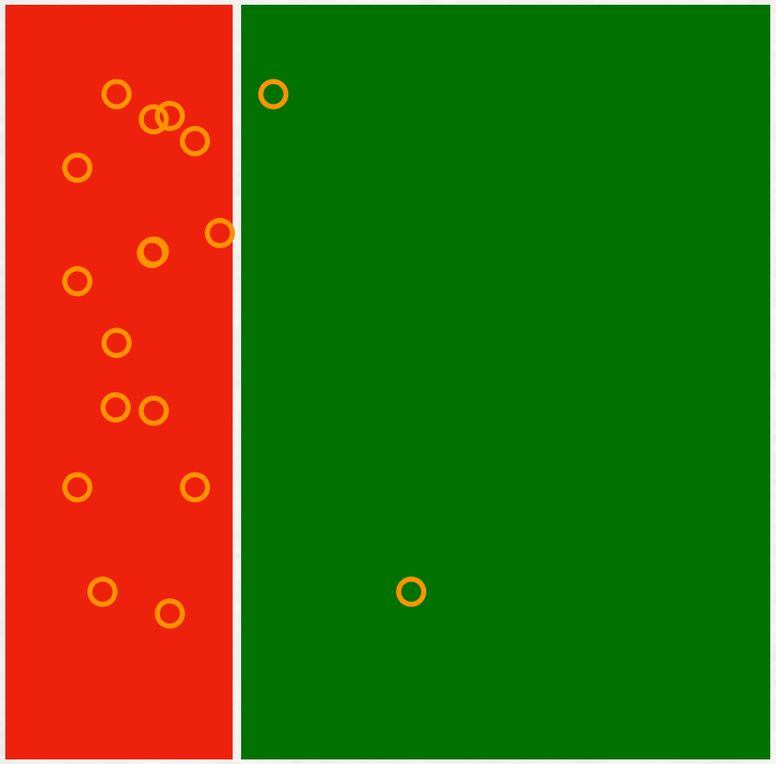
Bad Thing Happens: Evaluation on training data is not reliable

drops exponentially

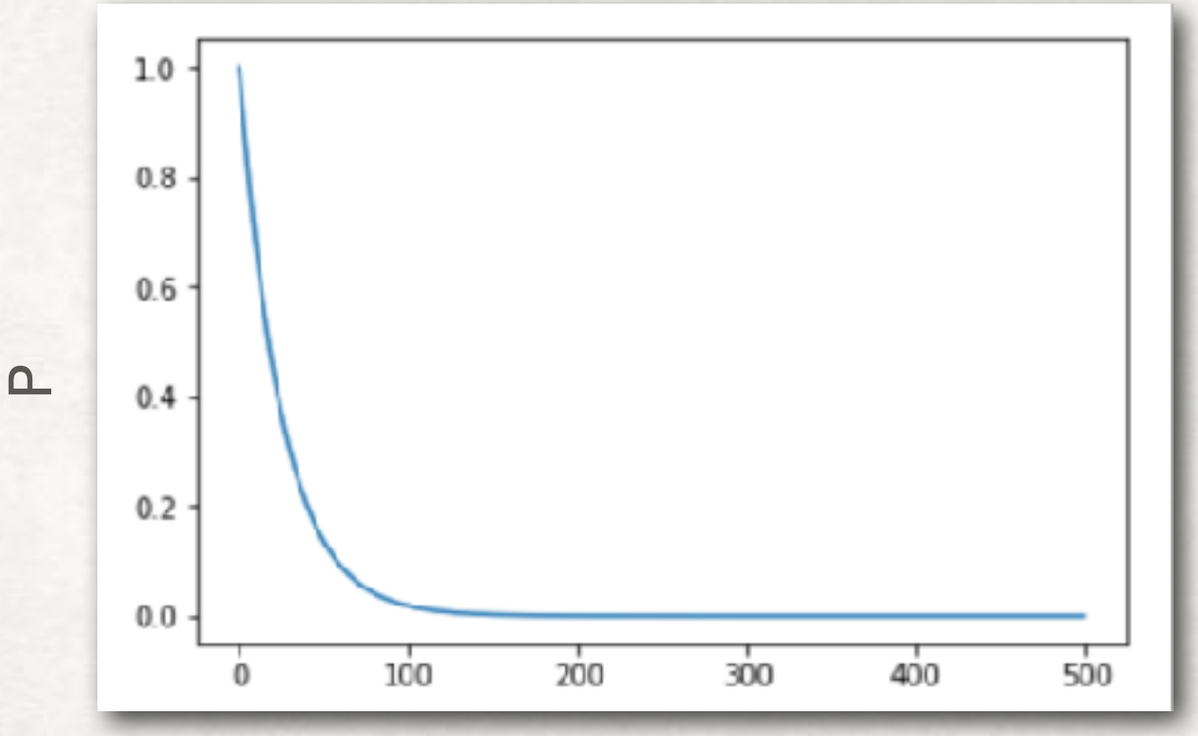
with #. data samples, but modulated by your tolerance of "how bad".

HOW FAST IT DROPS

- The chance you get a rare D is really rare — when the data set size is big.



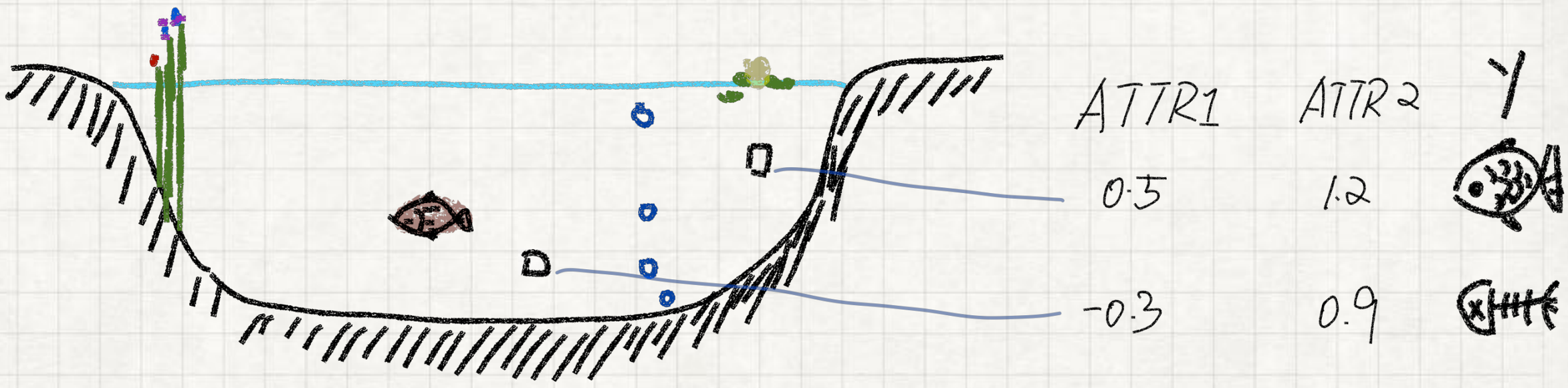
$$P(|E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$



MOD2
LEARNING AS A STOCHASTIC PROCESS

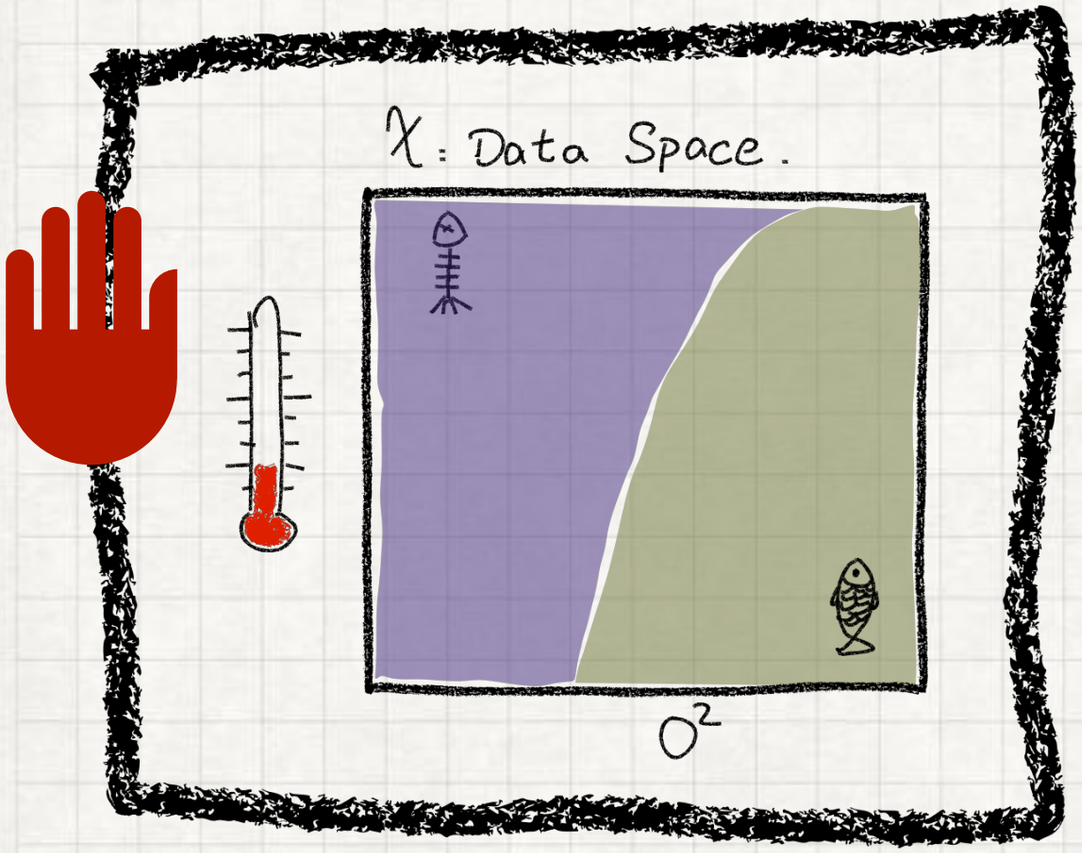
- What are the “things” in the statement of Hoeffding’s inequality: bad things happen rarely?
- And how they arise, so we can make sense of “rarely” / “often”?

A PICTURESQUE EXAMPLE



- Consider the problem of predicting whether the water condition is healthy for fish using two easy measurements: temperature and oxygen density.

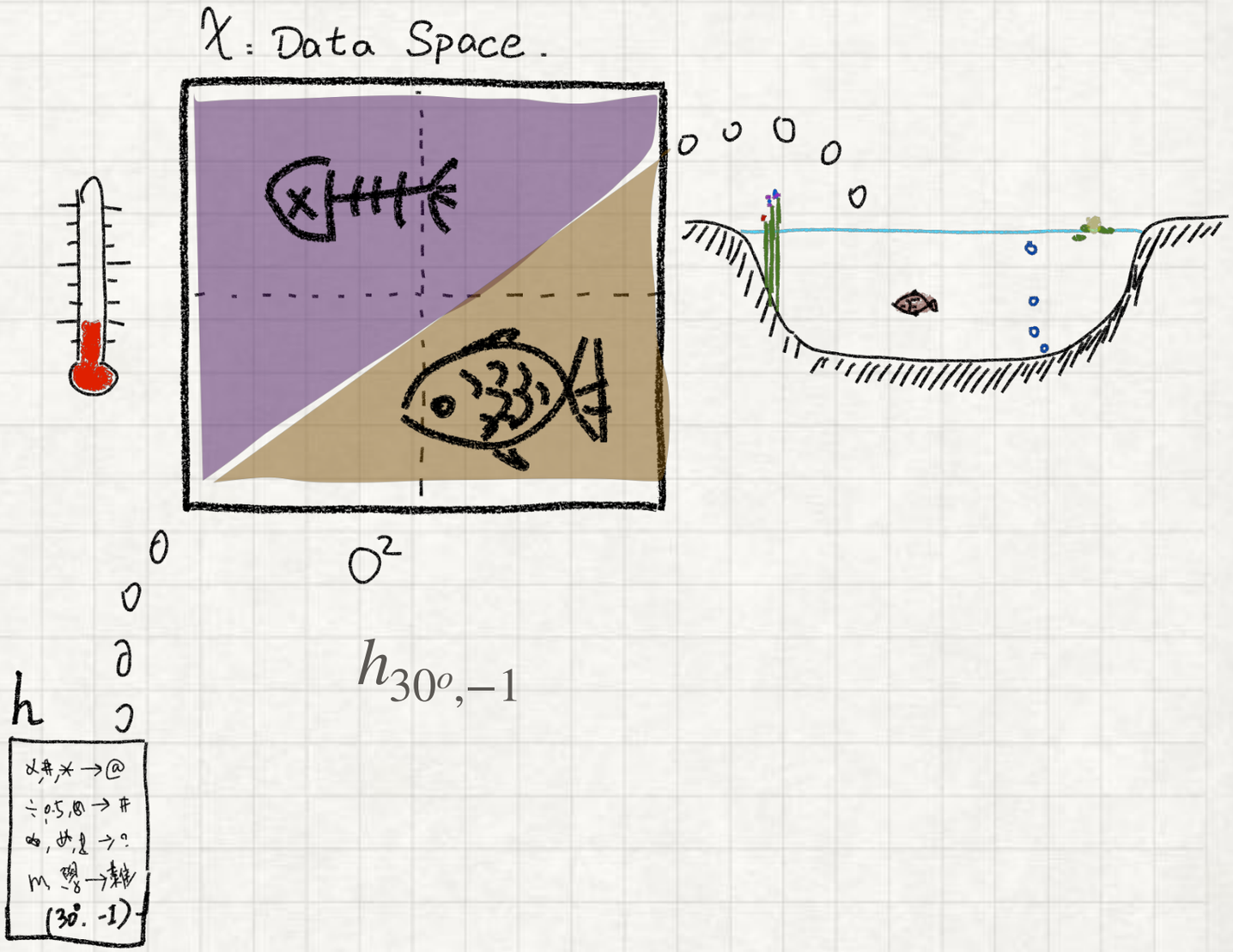
TRUE RELATIONSHIP IS UNAVAILABLE



- And you don't want to test by killing tons of fish!

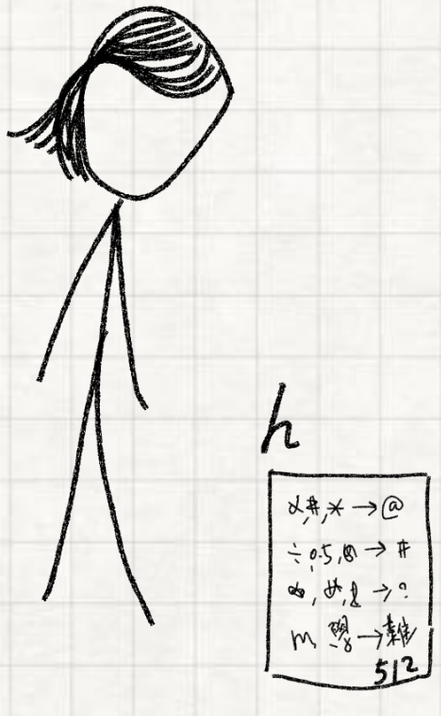
HYPOTHESES: MODELLING THE OUTCOME IN DATA SPACE

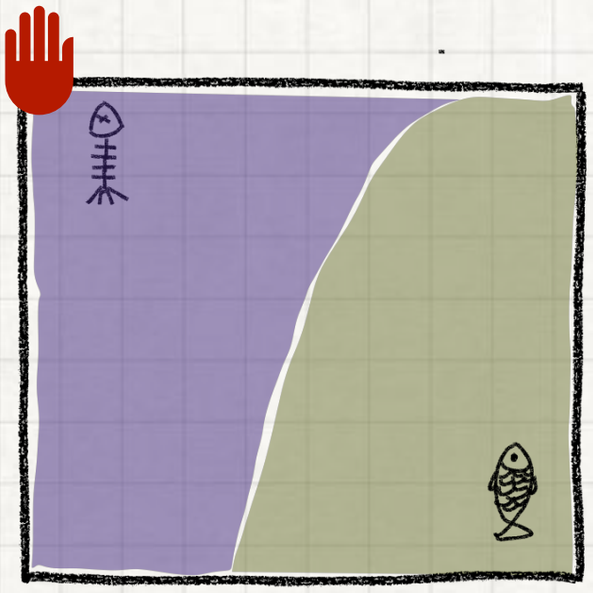
- E.g. a linear model uses a pair of (θ, b) — angle and y-intercept — to describe the the class boundary.

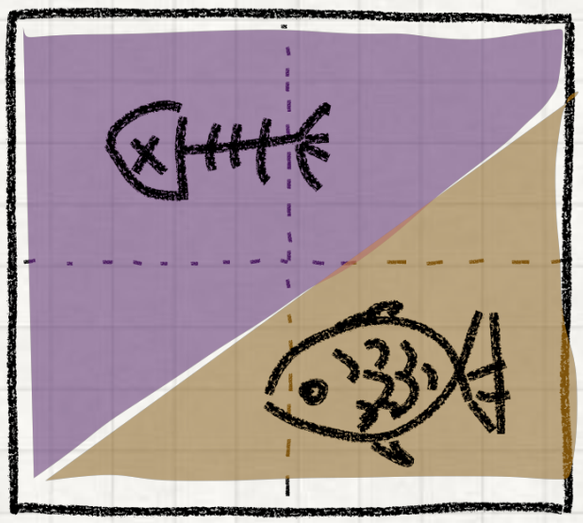


$h_{30^o,-1}$

GOLDEN STANDARD TO EVALUATE

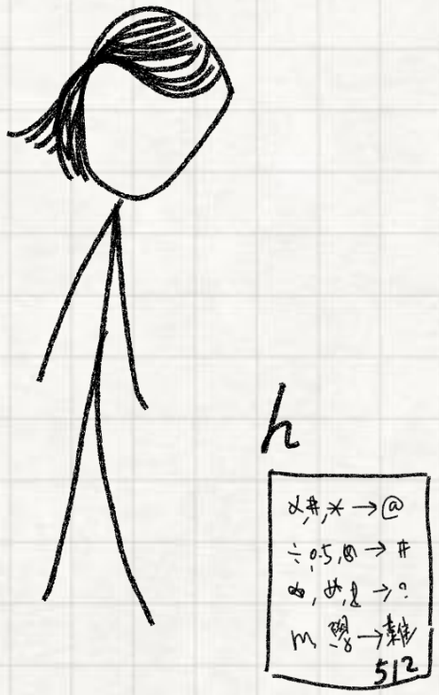


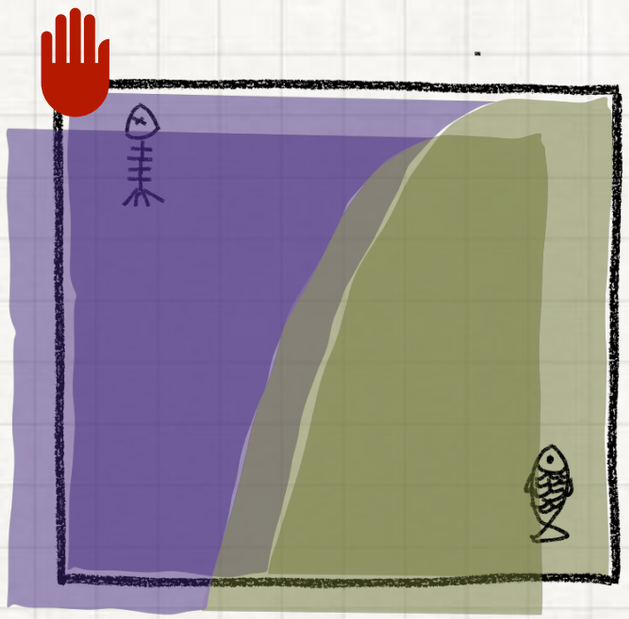


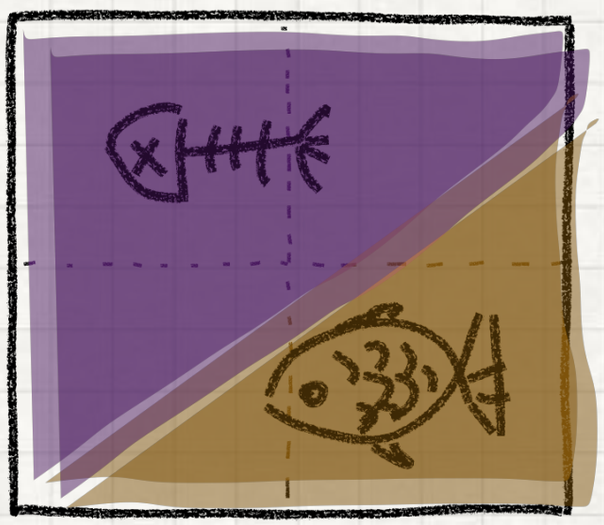


We want to know the generalisation performance, the potential error of a hypothesis h over the entire $\mathcal{X}$-space — $E_{out}[h]$

GOLDEN STANDARD TO EVALUATE

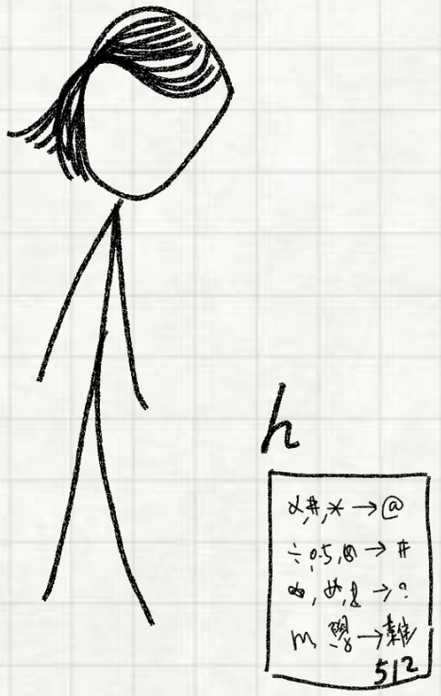


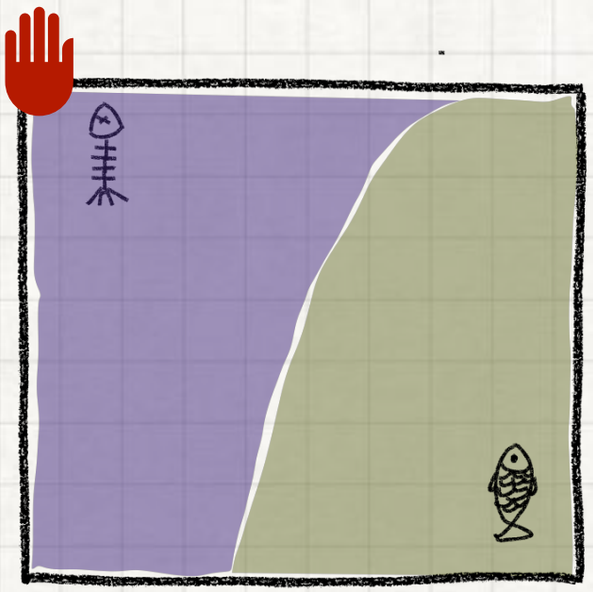


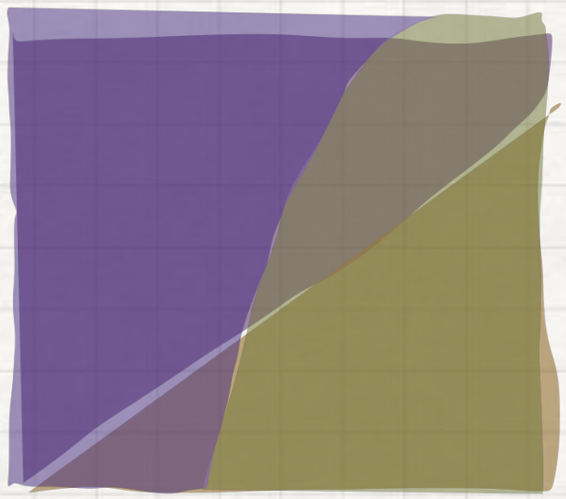


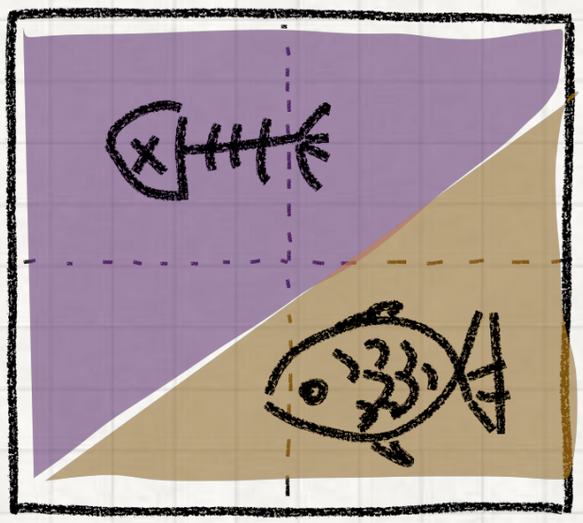
We want to know the generalisation performance, the potential error of a hypothesis h over the entire $\mathcal{X}$-space — $E_{out}[h]$

GOLDEN STANDARD TO EVALUATE



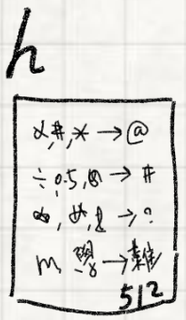


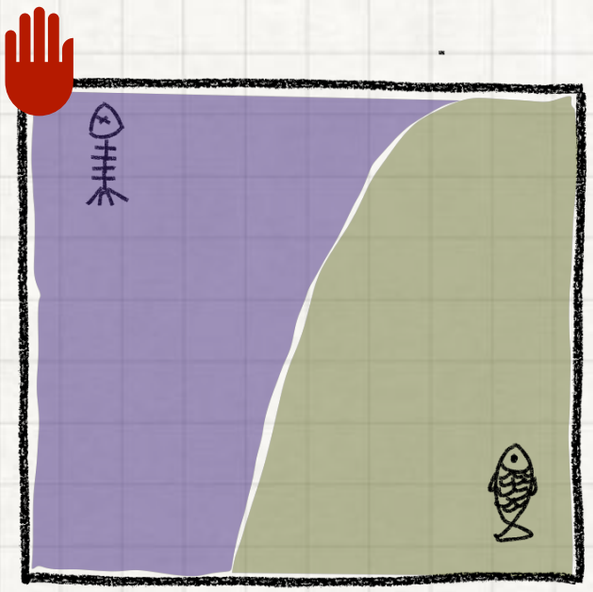


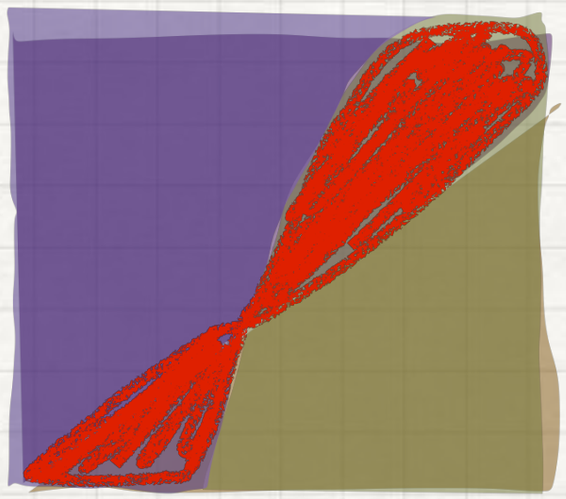


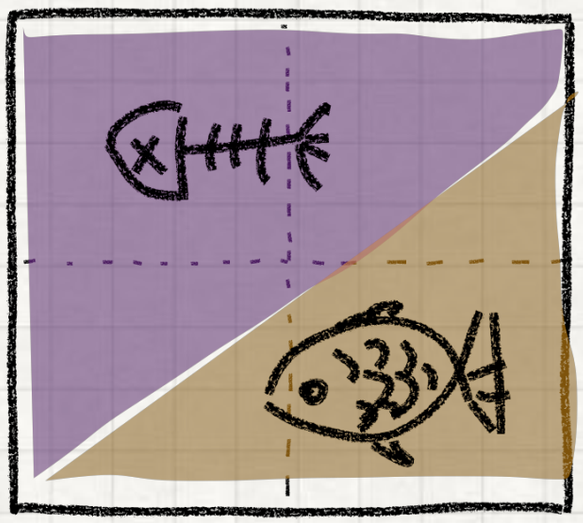
We want to know the generalisation performance, the potential error of a hypothesis h over the entire $\mathcal{X}$-space — $E_{out}[h]$

GOLDEN STANDARD TO EVALUATE



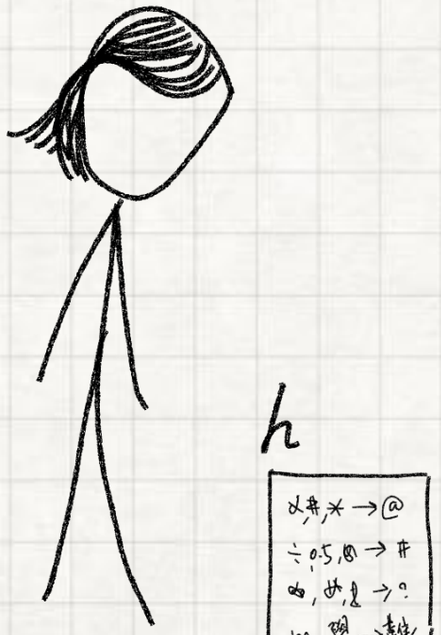


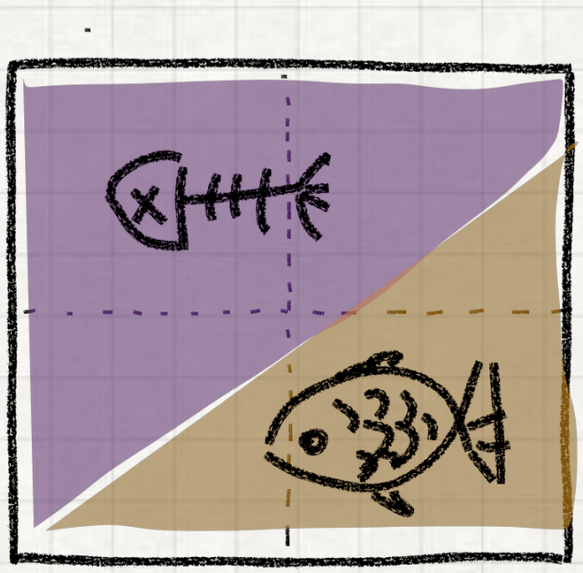


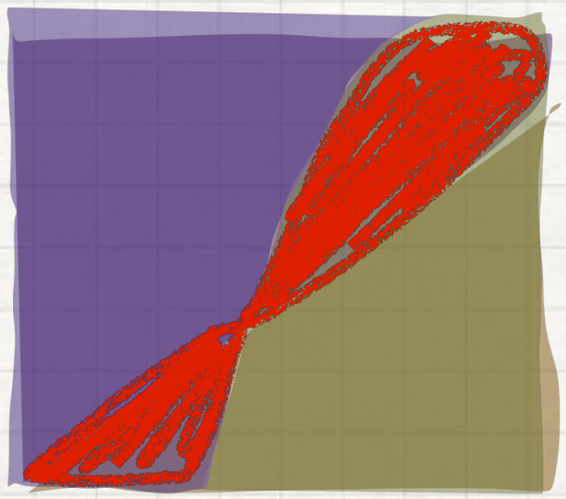


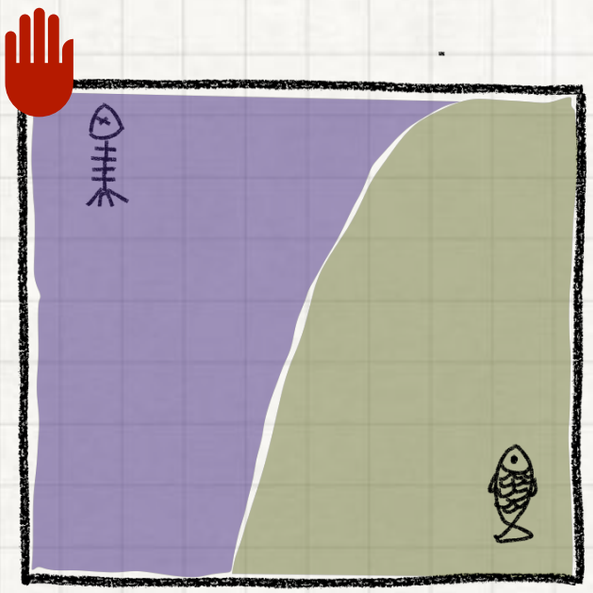
We want to know the generalisation performance, the potential error of a hypothesis h over the entire $\mathcal{X}$-space — $E_{out}[h]$

GOLDEN STANDARD TO EVALUATE





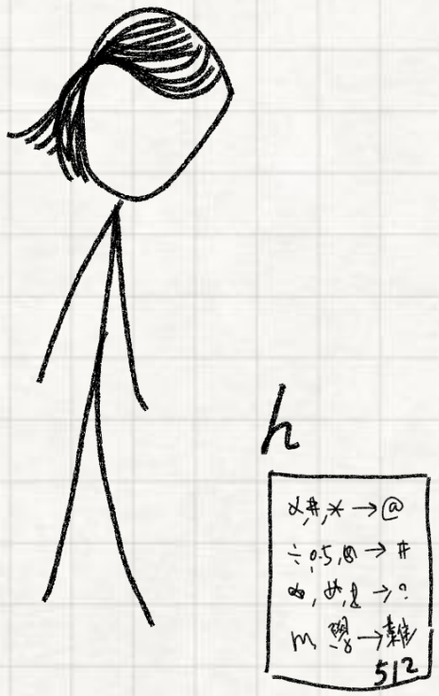


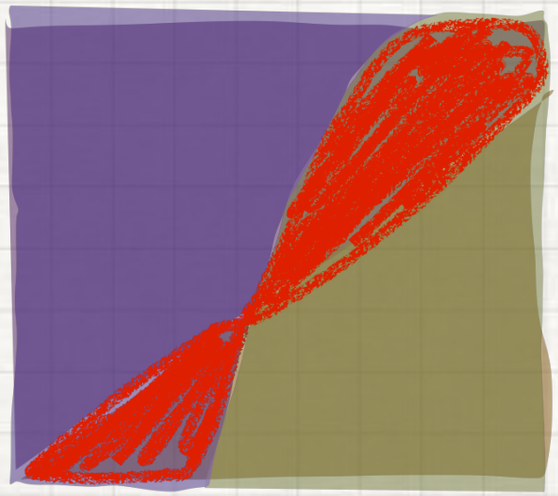


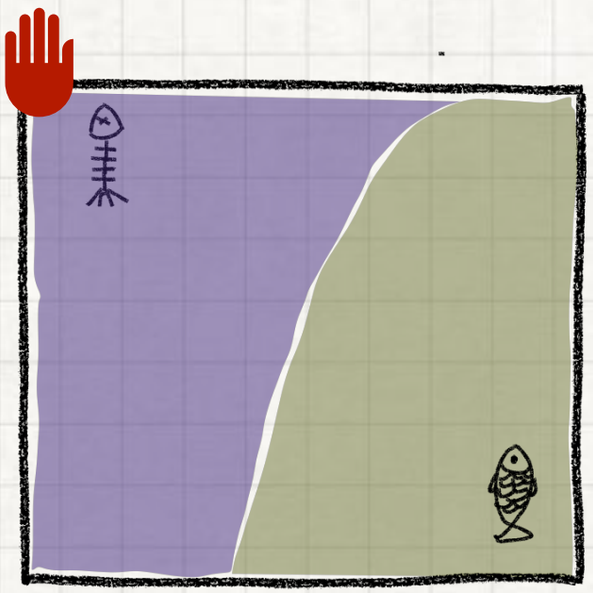
We want to know the generalisation performance, the potential error of a hypothesis h over the entire $\mathcal{X}$-space — $E_{out}[h]$

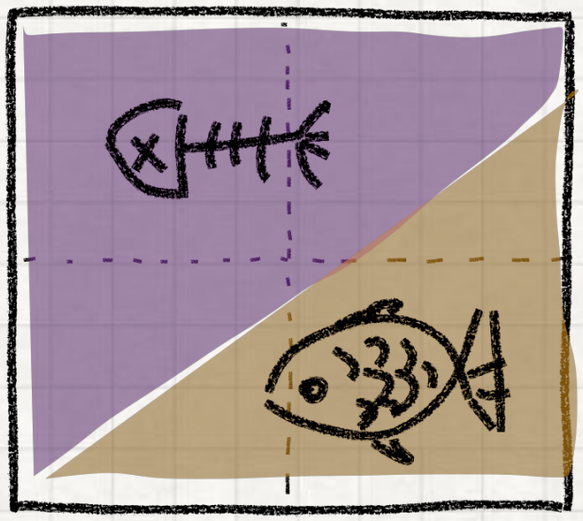
$$E_{out}[h] = \int_{x \in \mathcal{X}} [h(x) \neq y(x)] p(x) dx$$

GOLDEN STANDARD IS NOT DIRECTLY ACCESSIBLE



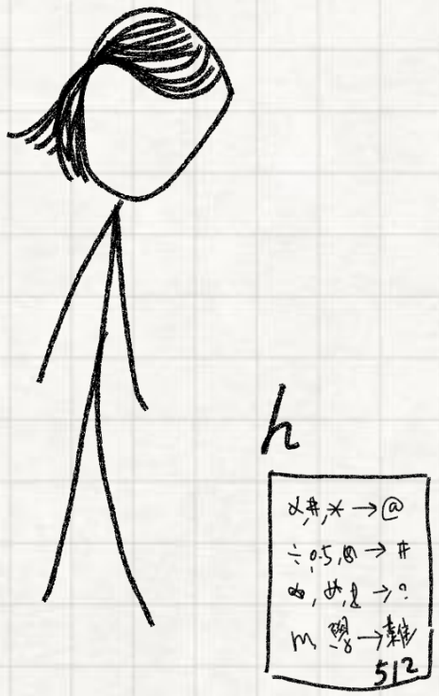


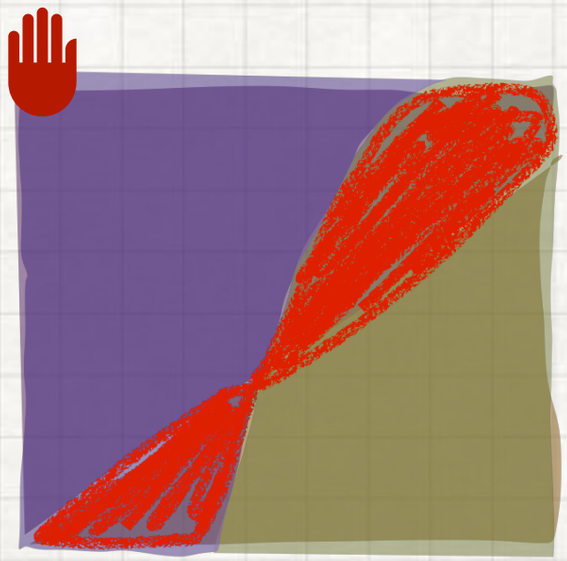


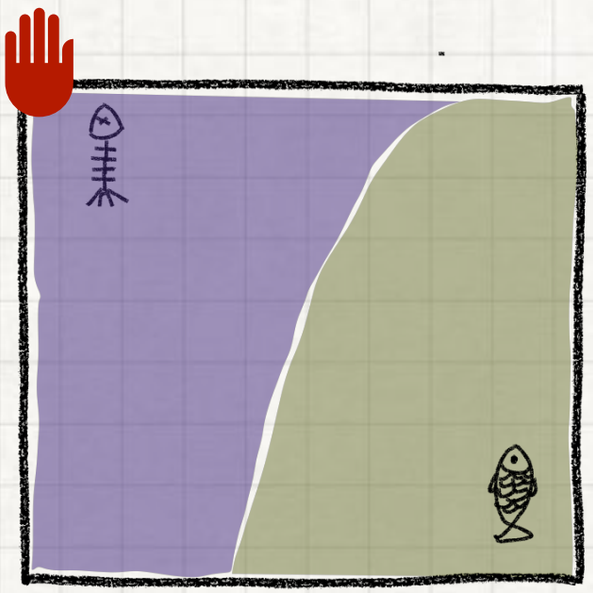


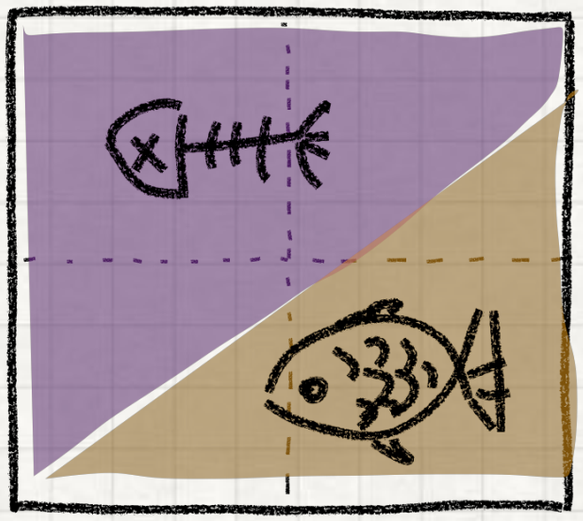
$$E_{out}[h] = \int_{x \in \mathcal{X}} [h(x) \neq y(x)] p(x) dx$$

GOLDEN STANDARD IS NOT DIRECTLY ACCESSIBLE



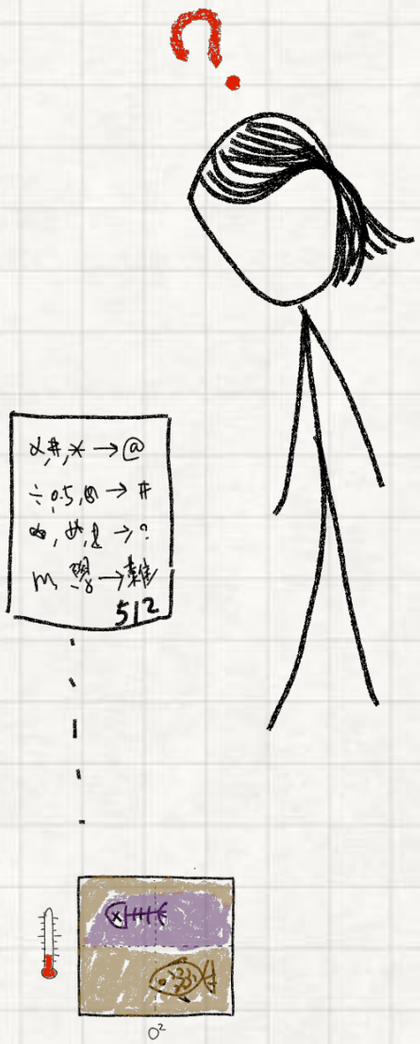






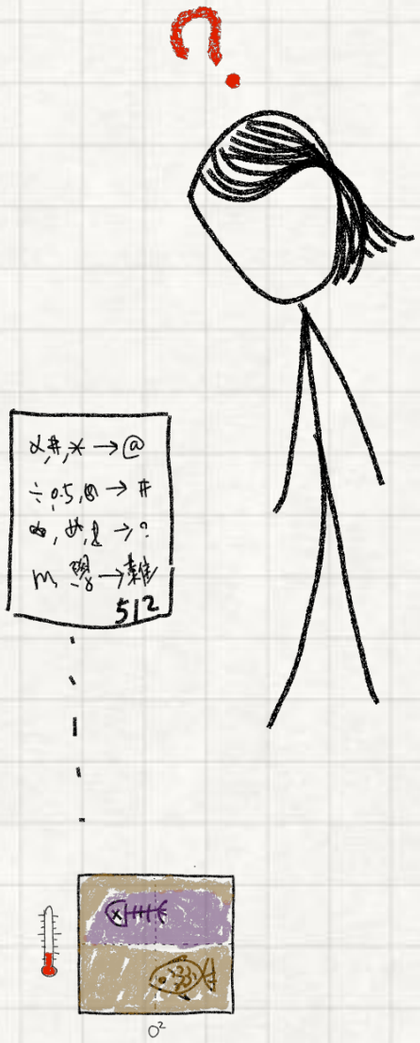
$$E_{out}[h] = \int_{x \in \mathcal{X}} [h(x) \neq y(x)]p(x)dx$$

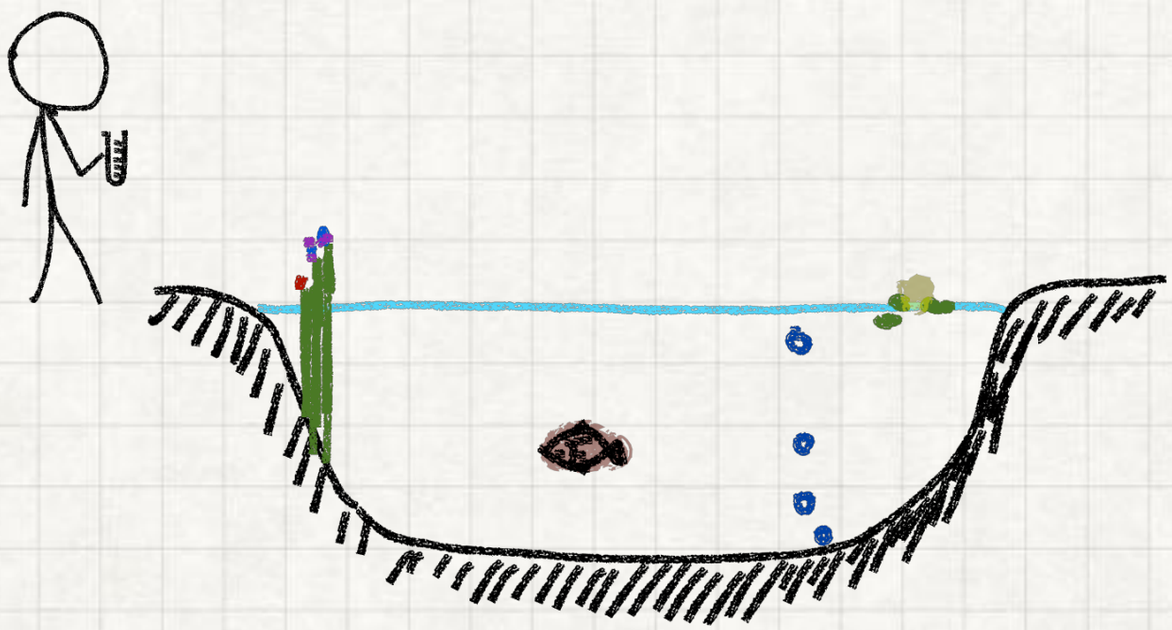
TEST A HYPOTHESIS



- Now given a hypothesis, we want to determine if it is an accurate predictor for all possible water conditions …

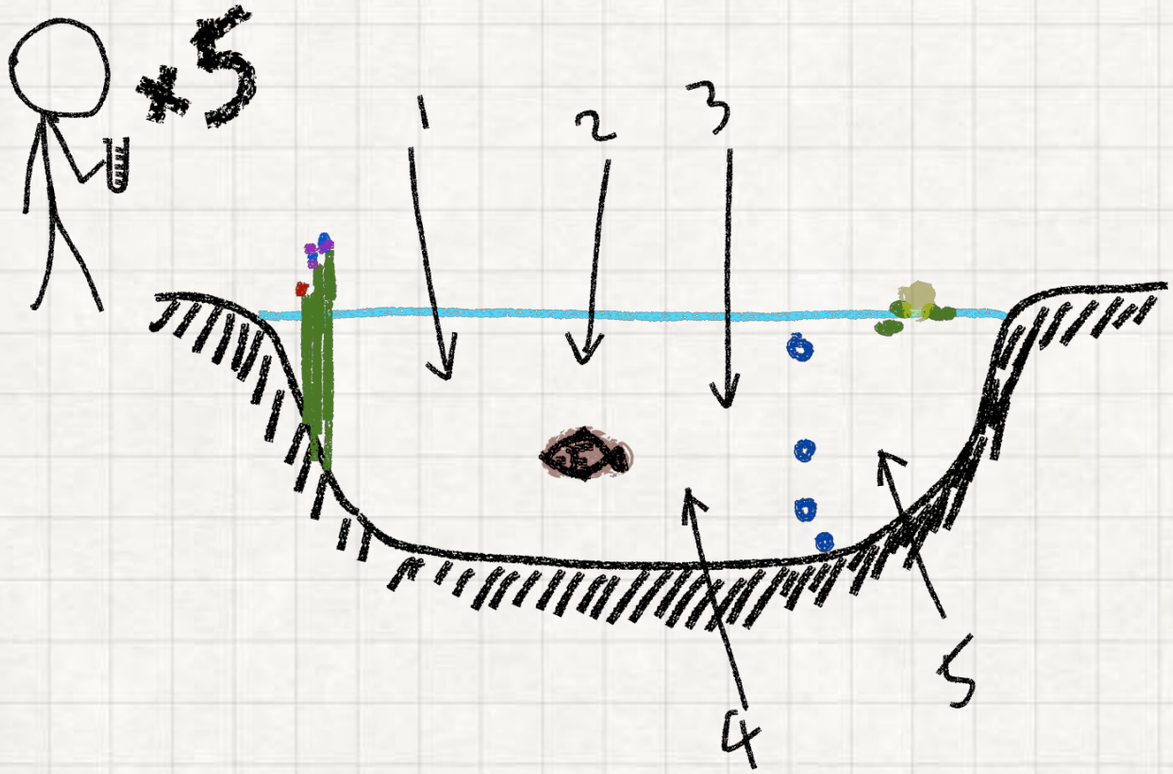
TEST A HYPOTHESIS

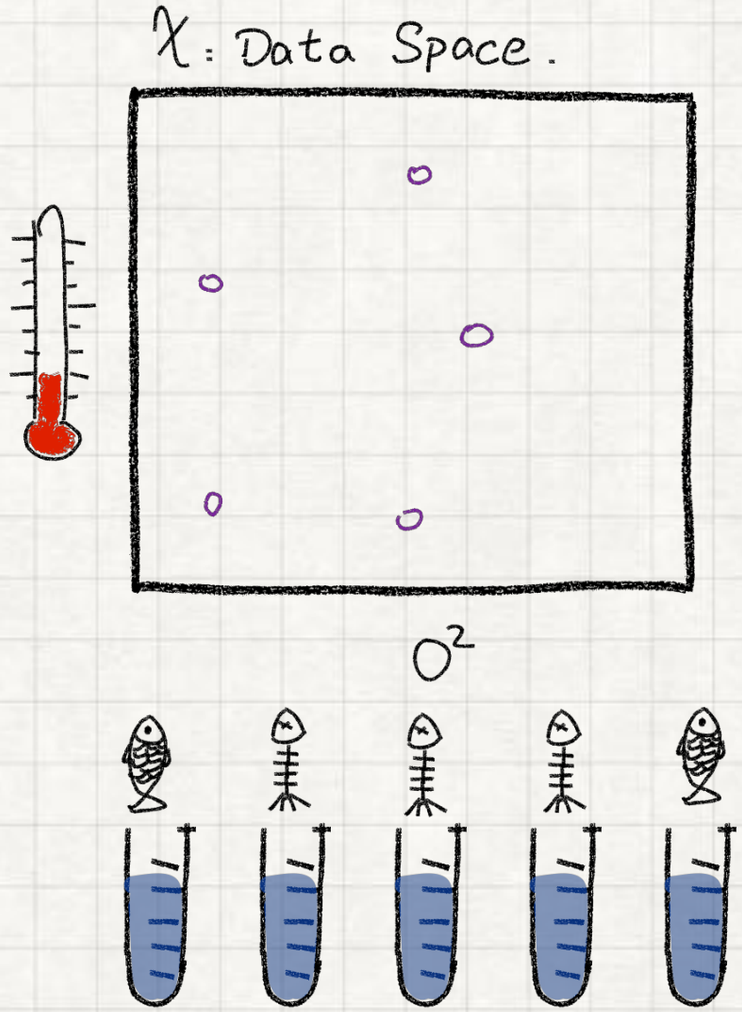




- Now given a hypothesis, we want to determine if it is an accurate predictor ... by sending people to the pond.

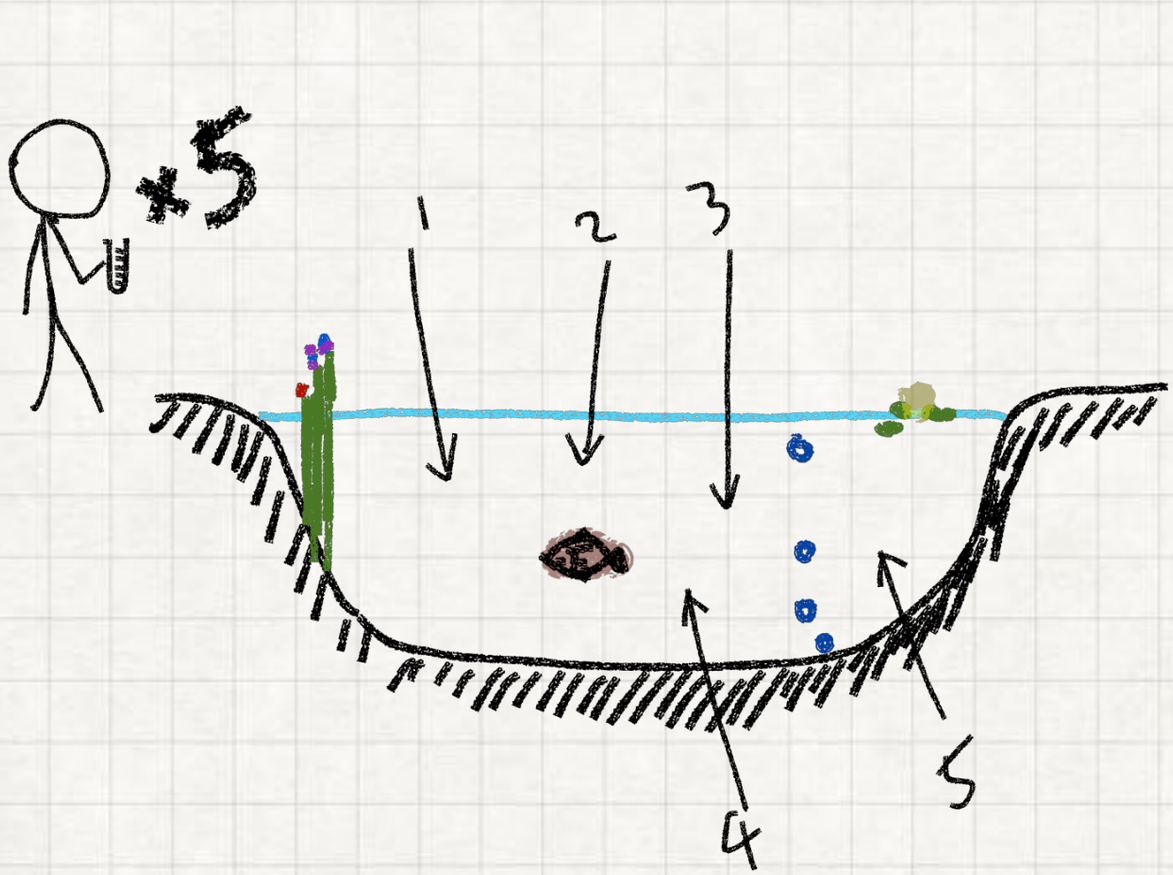
N-SAMPLING IN THE DATA SPACE

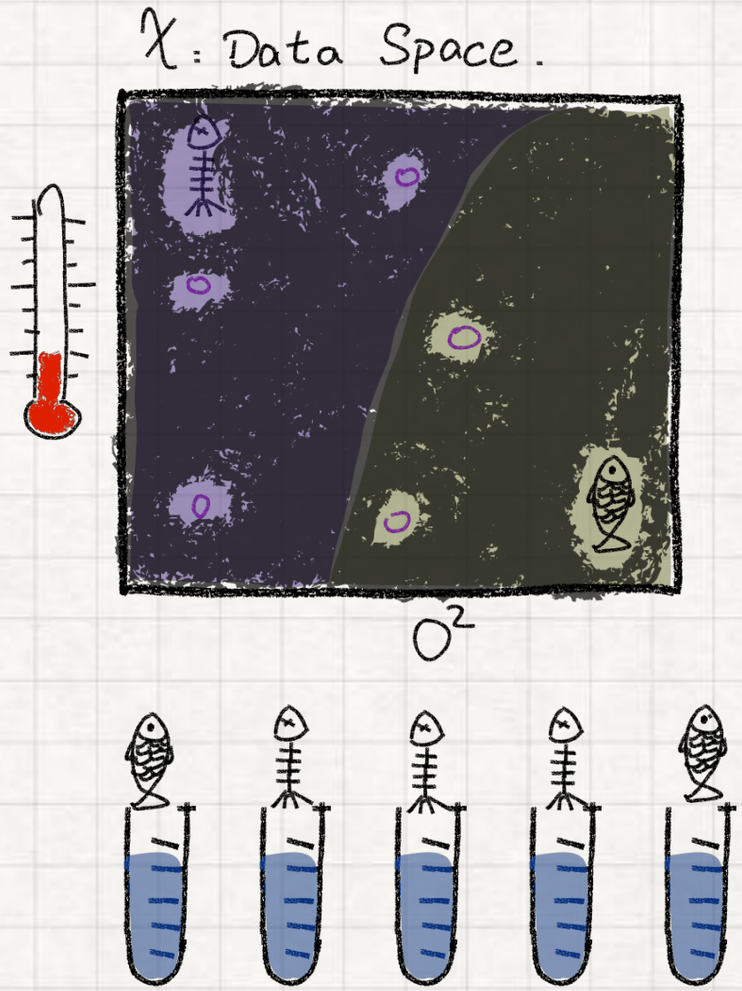




Now we have observations of the real-pond relationship between the measurements of temperature and oxygen and the condition for fish (R.I.P. for some).

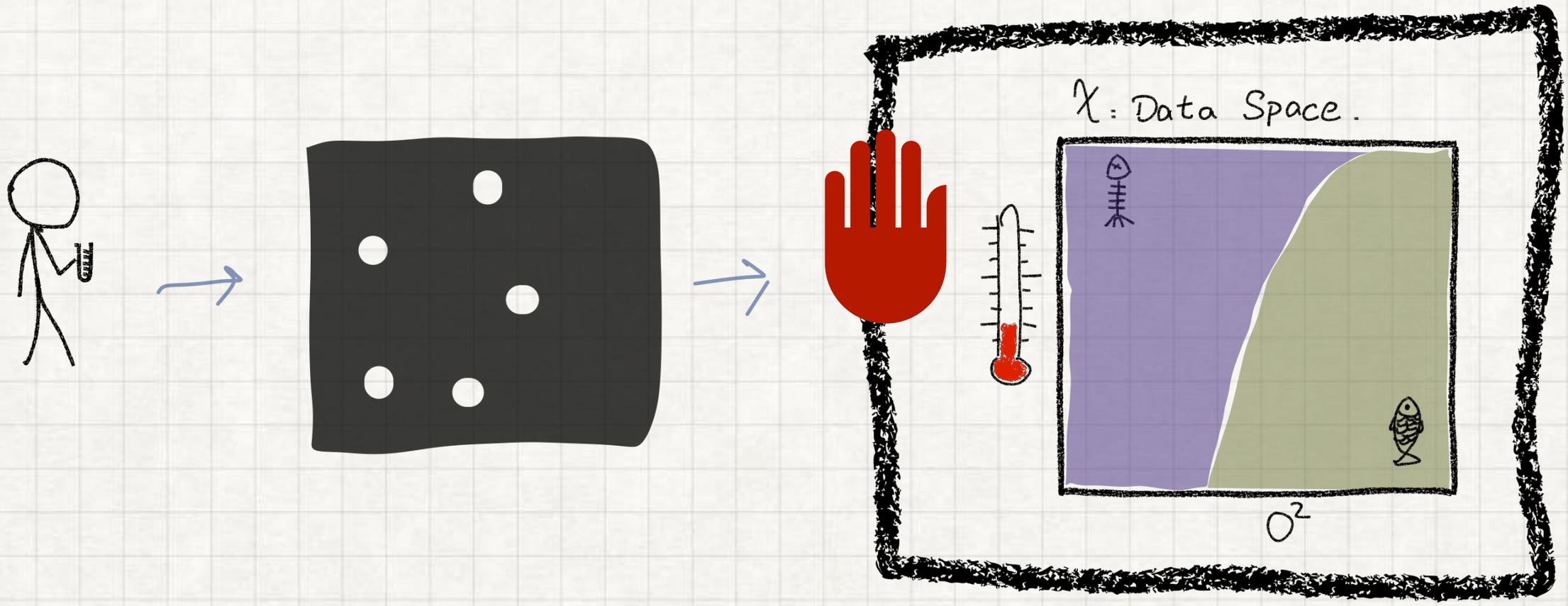
INFORMATION GAINED BY N-SAMPLES



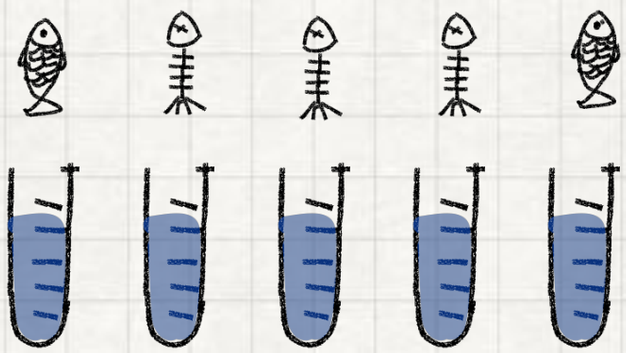


We are peeking through the observations into ...

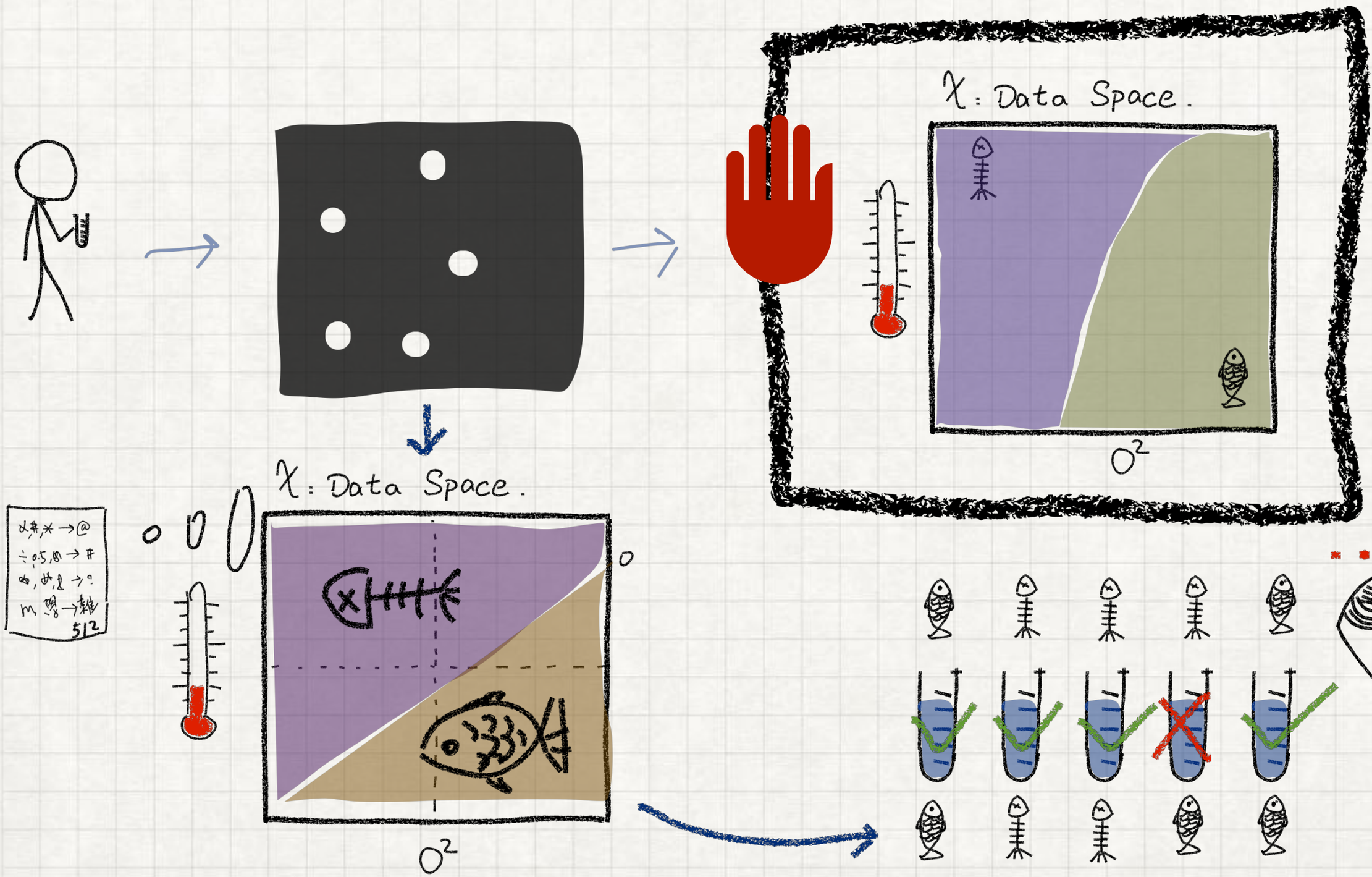
EVALUATION BY USING N SAMPLES



We are peeking through the observations into the black box of the real target concept / relationship between X and y of interest.

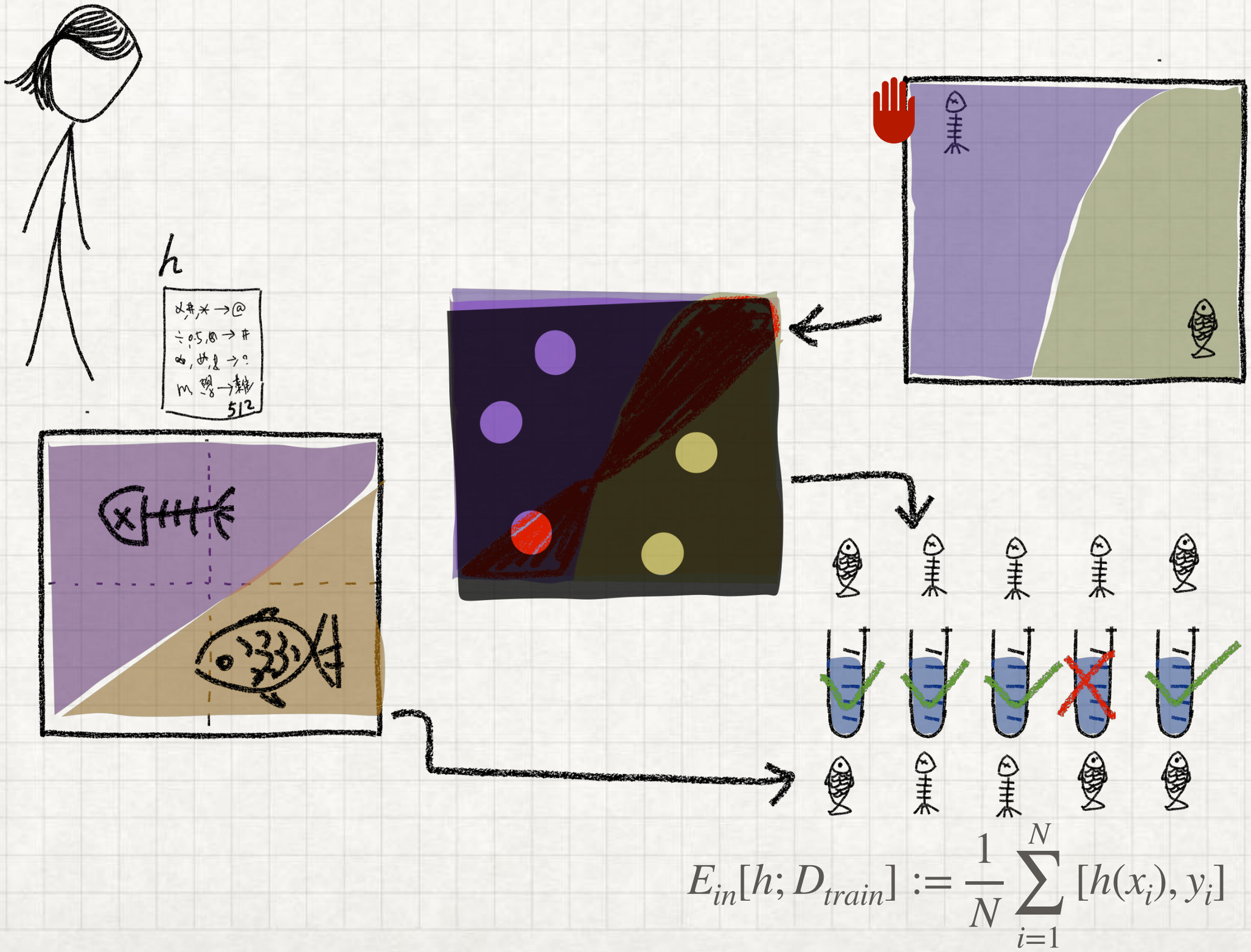


EVALUATION BY USING N SAMPLES



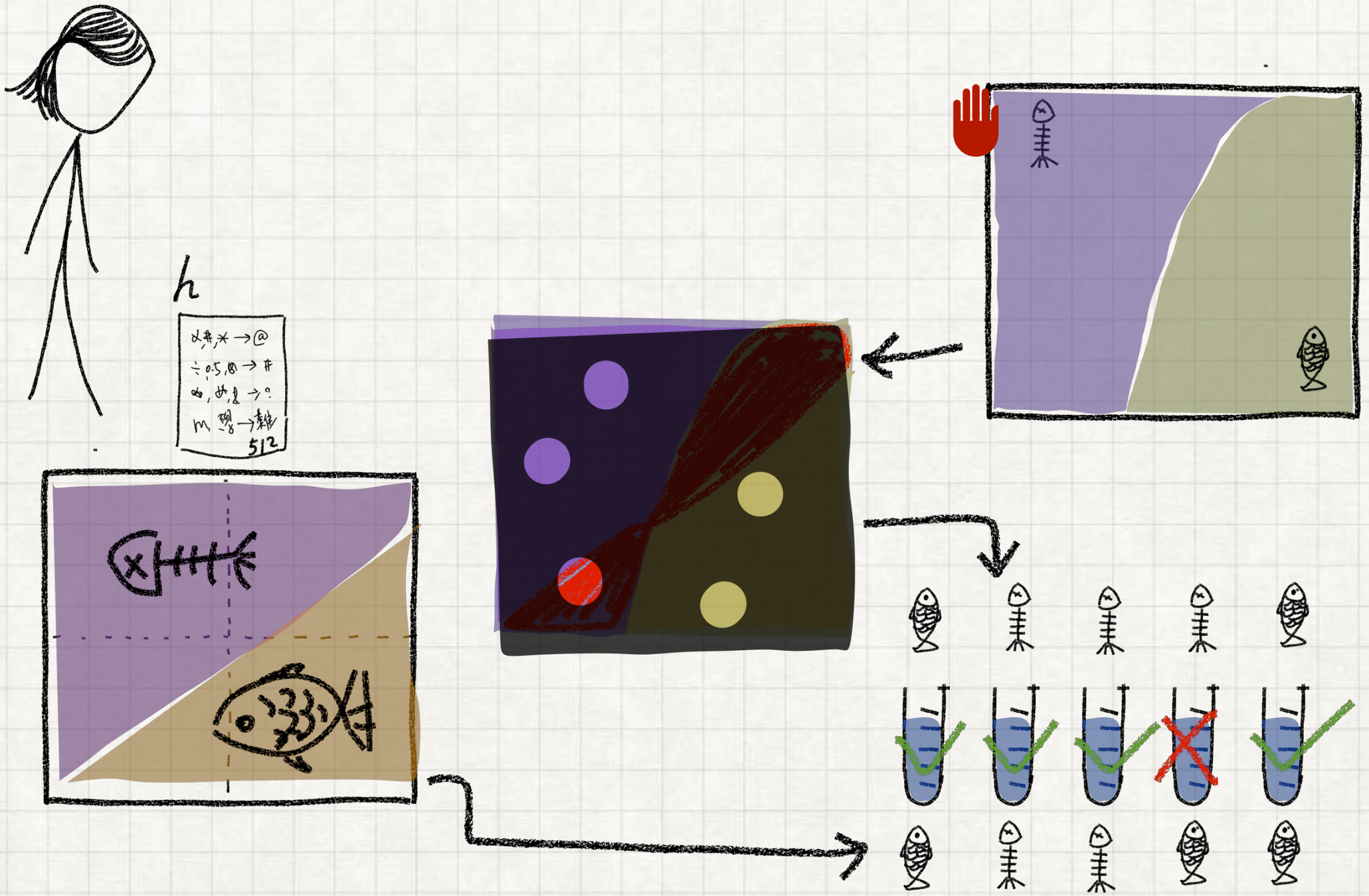
We compare the observations to our hypothesis.

USE N SAMPLES INSTEAD OF GOLDEN STANDARD



$$E_{in}[h; D_{train}] := \frac{1}{N}\sum_{i=1}^{N}[h(x_i), y_i]$$

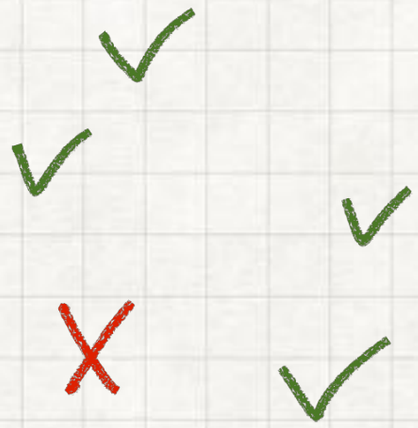
USE N SAMPLES INSTEAD OF GOLDEN STANDARD

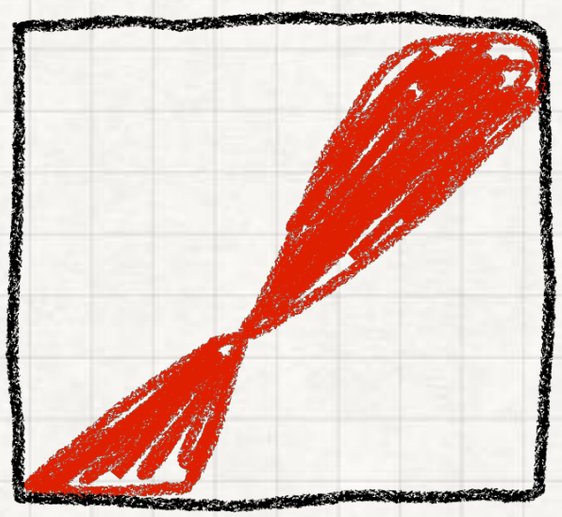


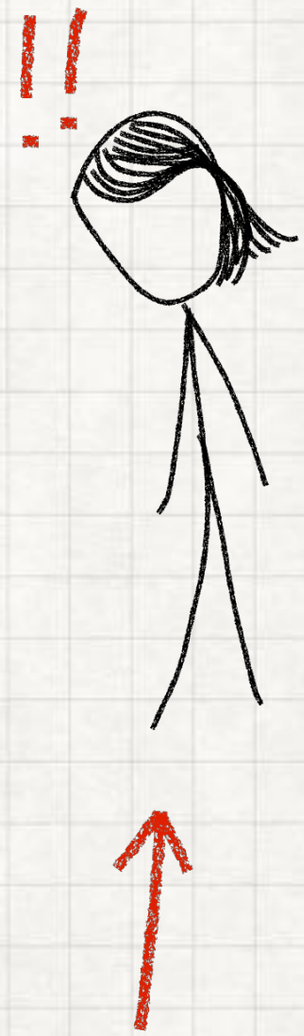
$$E_{in}[h; D_{train}] := \frac{1}{N}\sum_{i=1}^{N}[h(x_i) \neq y_i]$$

$y(x)@\{x_i\}$

RISK OF EMPIRICAL EVALUATION







$$\left| E_{in}[h; D_{train}] - E_{out}[h] \right|$$

$$\left| 0.2 - 0.22 \right| = 0.02$$

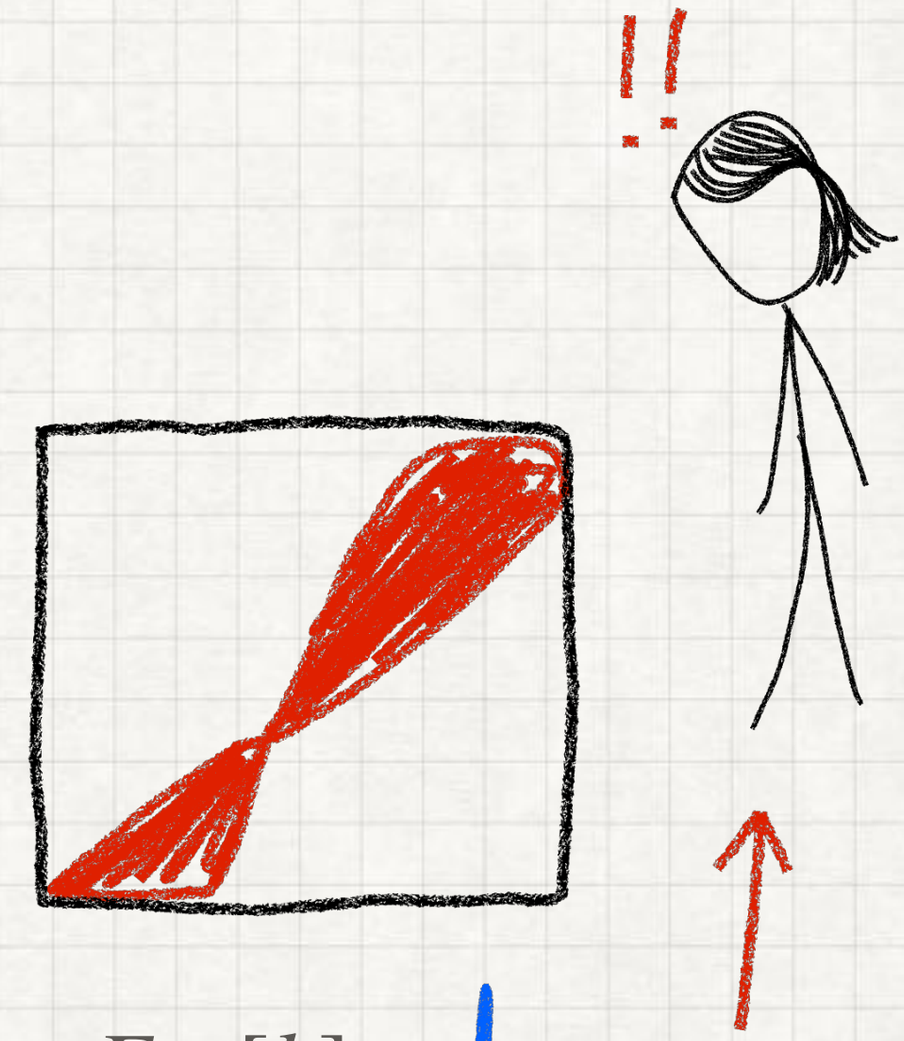
RISK OF EMPIRICAL EVALUATION

Q: If you have a magic wand, which will make data more likely to arise from where h makes errors, how would that affect the expected error?

A. No influence.

B. E_{in} tends to be an overestimate of E_{out}, i.e. it is more likely $E_{out} < E_{in}$.

C. E_{in} tends to be an underestimate of E_{out}.



$$|E_{in}[h; D_{train}] - E_{out}[h]|$$

$$|0.2 - 0.22| = 0.02$$

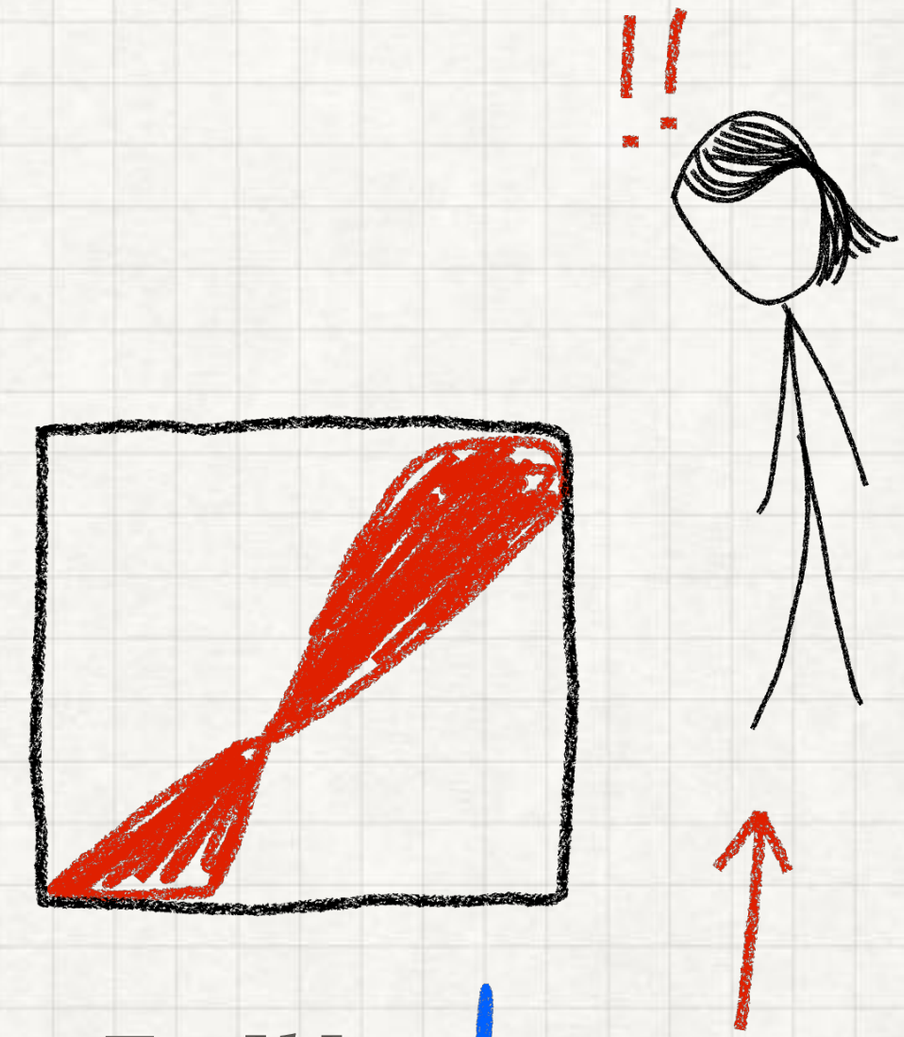
RISK OF EMPIRICAL EVALUATION

Q: If you have a magic wand, which will make data more likely to arise from where h makes errors, how would that affect the expected error?

A. No influence.

B. E_{in} tends to be an overestimate of E_{out}, i.e. it is more likely $E_{out} < E_{in}$.

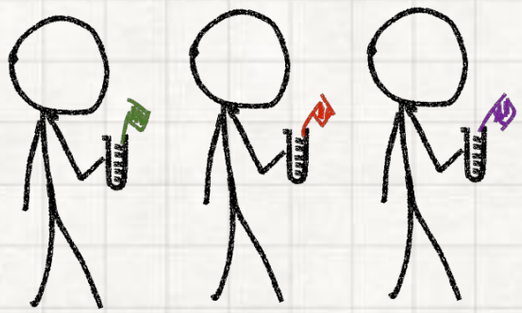
C. E_{in} tends to be an underestimate of E_{out}.

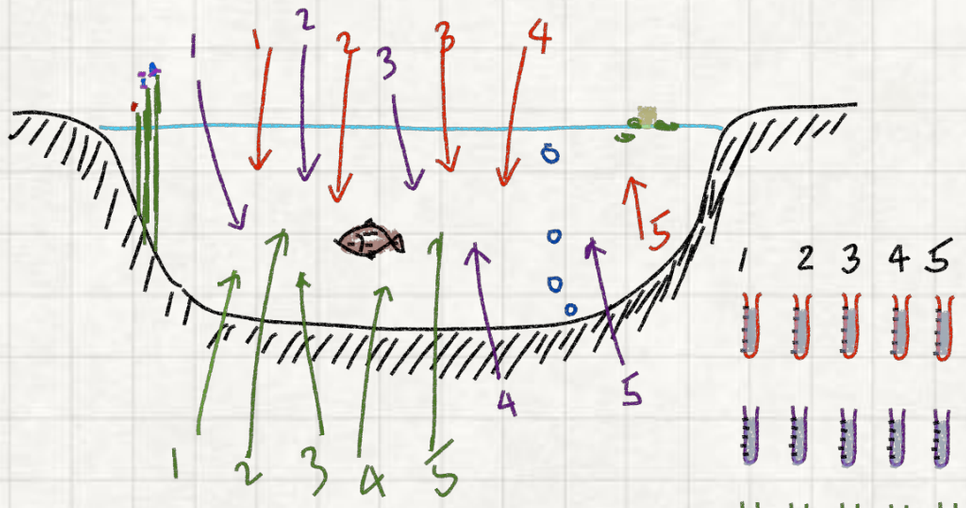


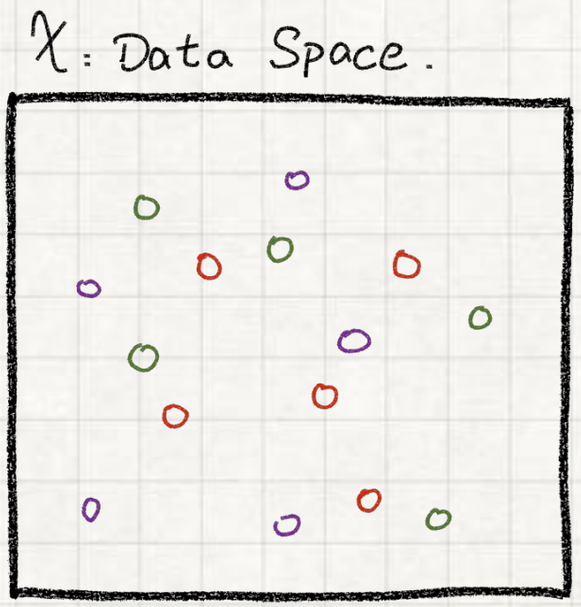
$$|E_{in}[h; D_{train}] - E_{out}[h]|$$

$$|0.2 - 0.22| = 0.02$$

RANDOM RISK FROM RANDOM DATA

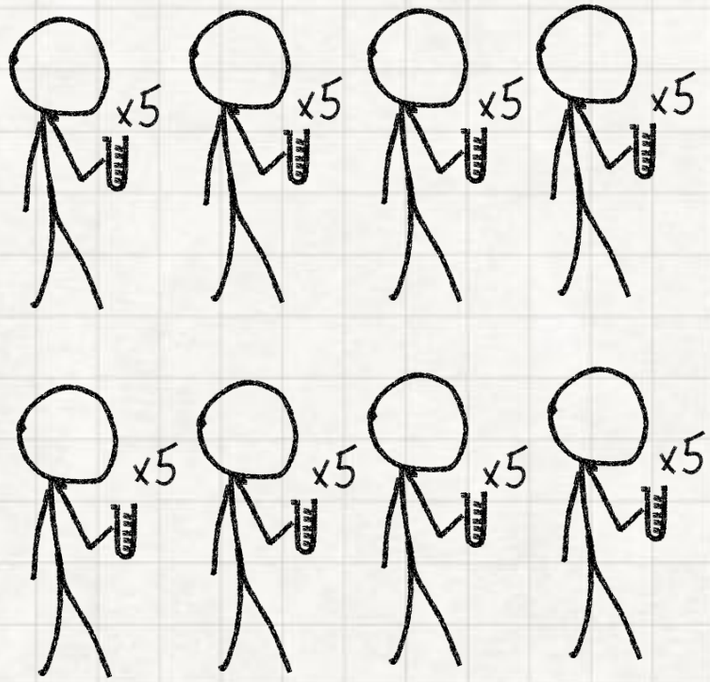


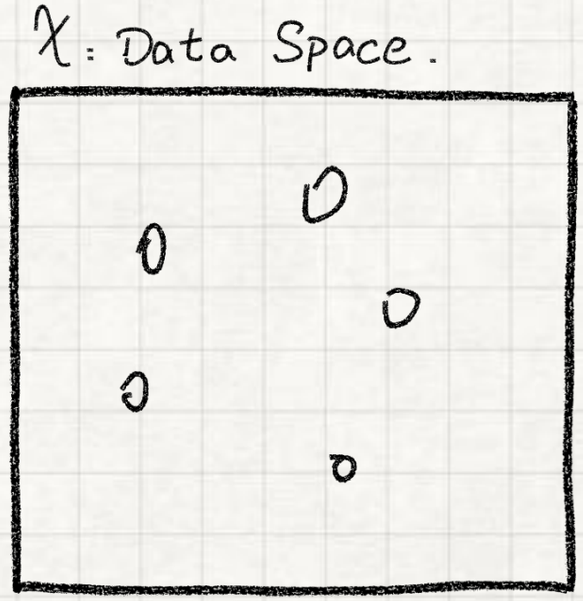


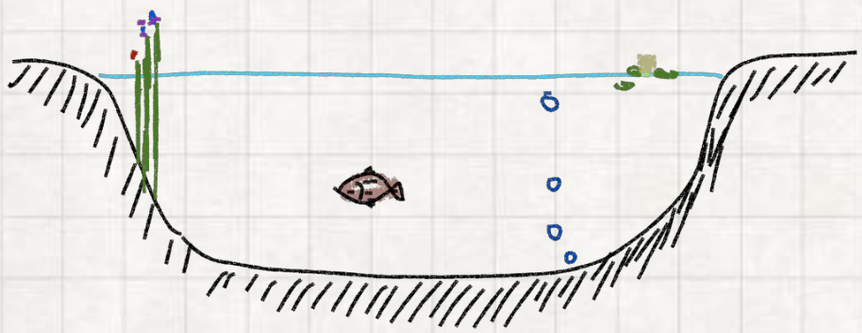


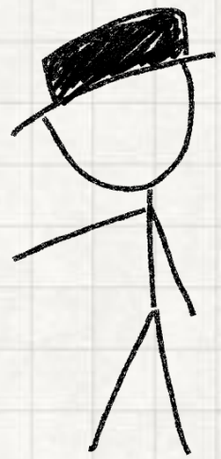
If you sent different testers to the pond and let each of them take 5 samples, they would have come back with different 5-sample sets, corresponding to different sampling in the data space.

SPACE OF N-SAMPLE DATASETS (NOT DATA SPACE $\mathcal{X}$)



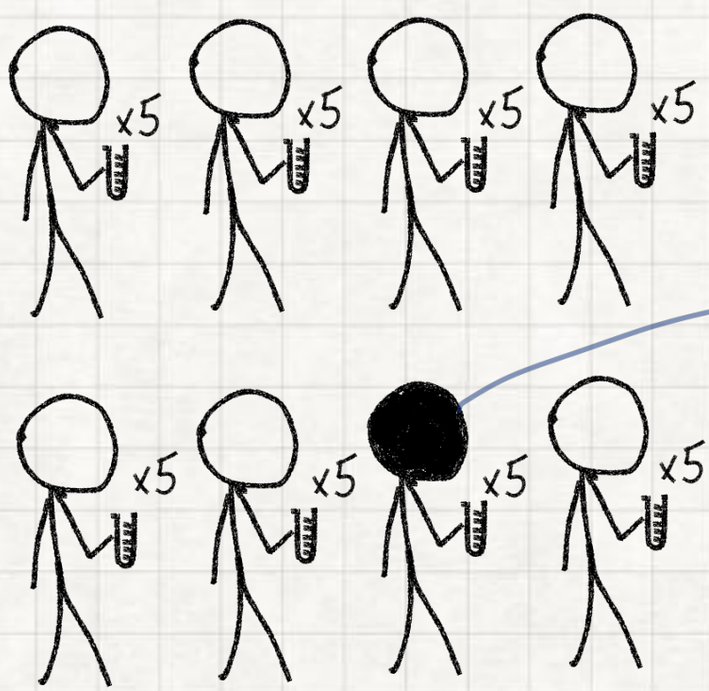


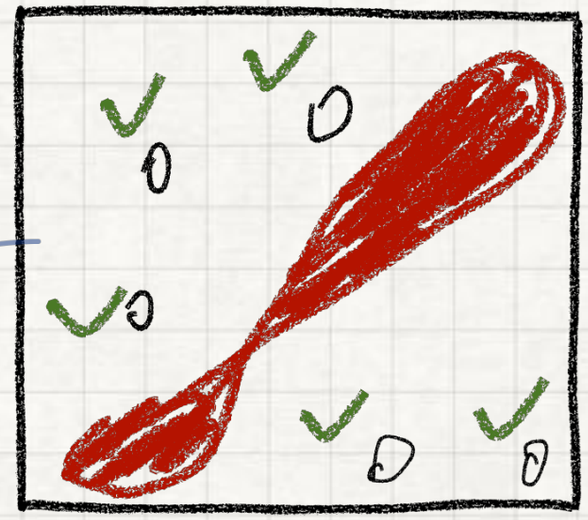




Now consider OUTSOURCING the pond-visiting task to a sampling business. One of their testers will return you 5 samples. You don't have control on or information of whom they would dispatch to sample from the pond.

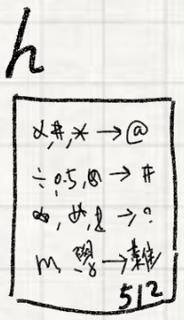
RISK IN THE SPACE OF N-SAMPLE DATASETS

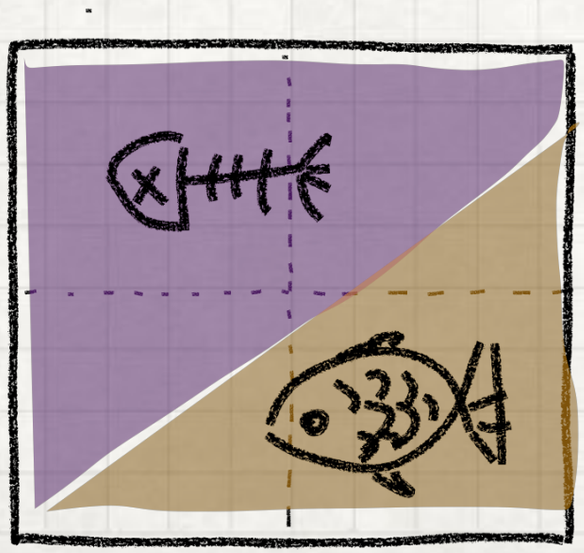




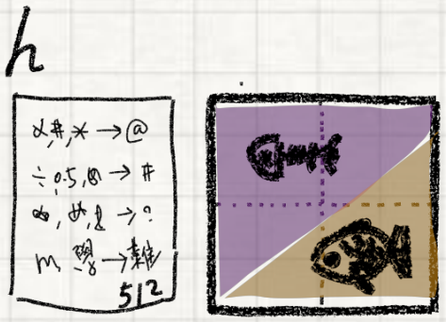
100%!

Some testers, though unintentionally, will return datasets, which lead to a highly inaccurate evaluation of your hypothesis. We mark them in black for our h.

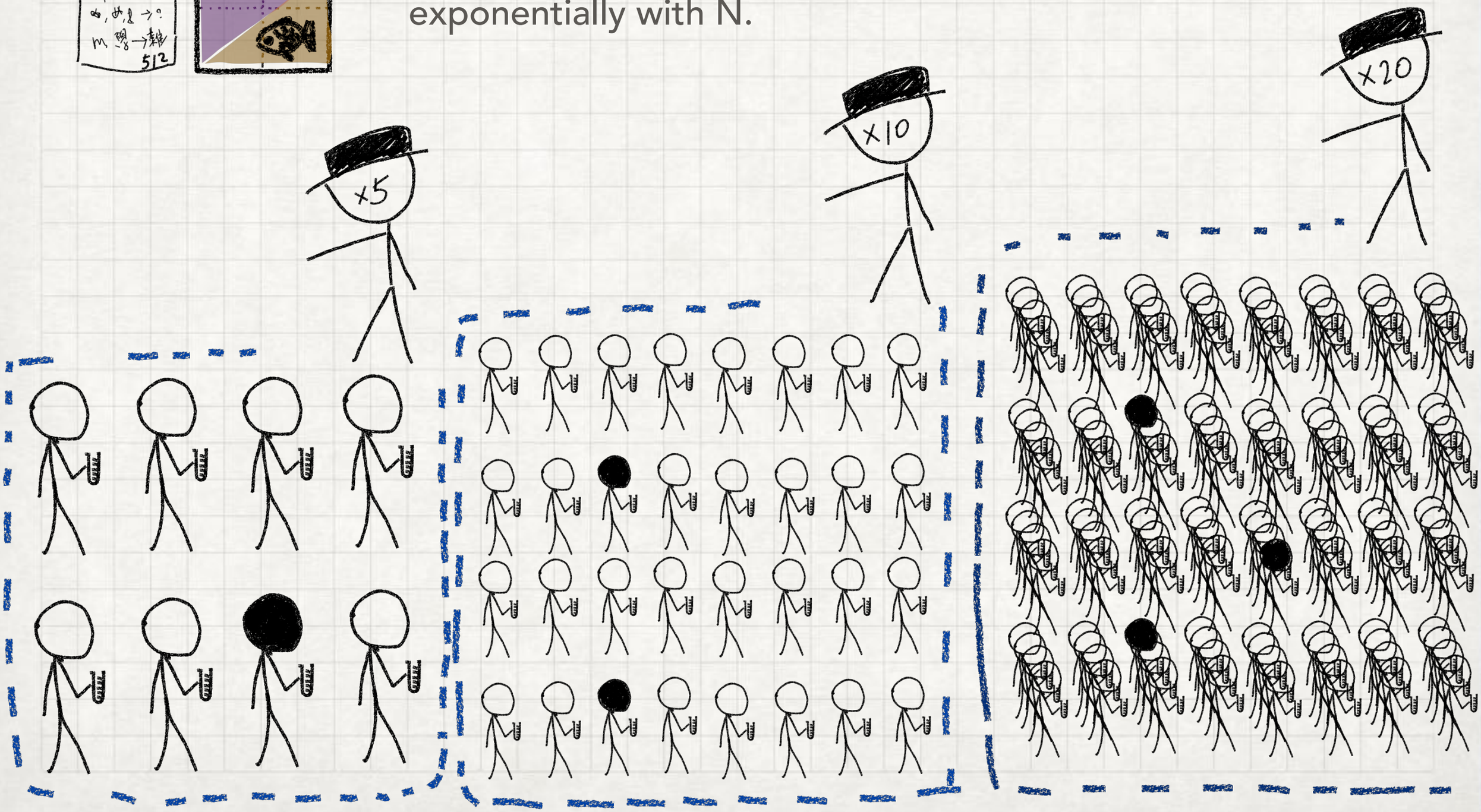




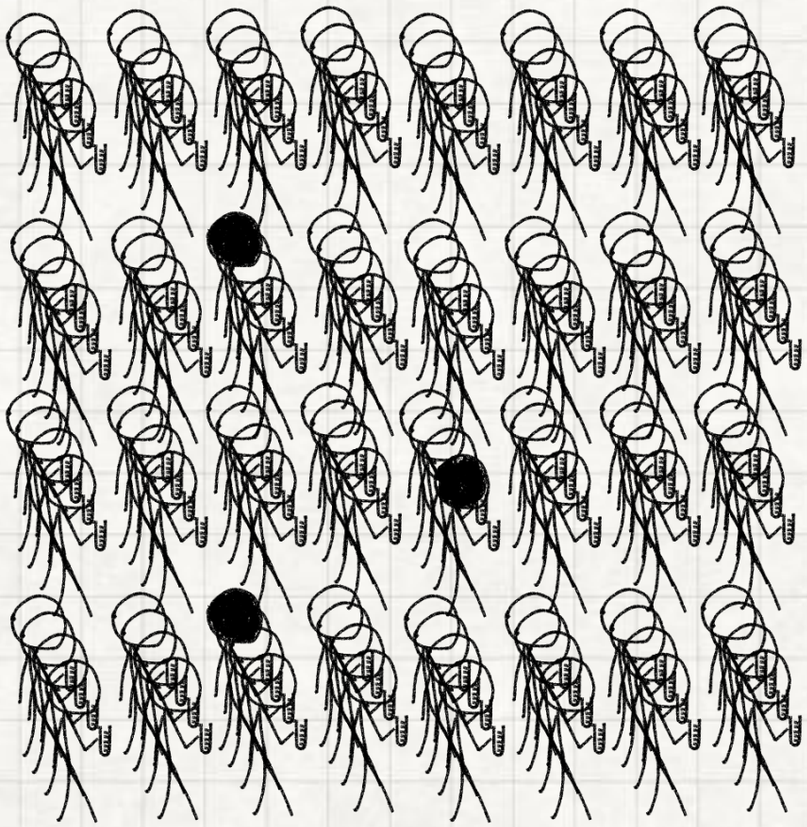
RISK REDUCES WHEN N INCREASES



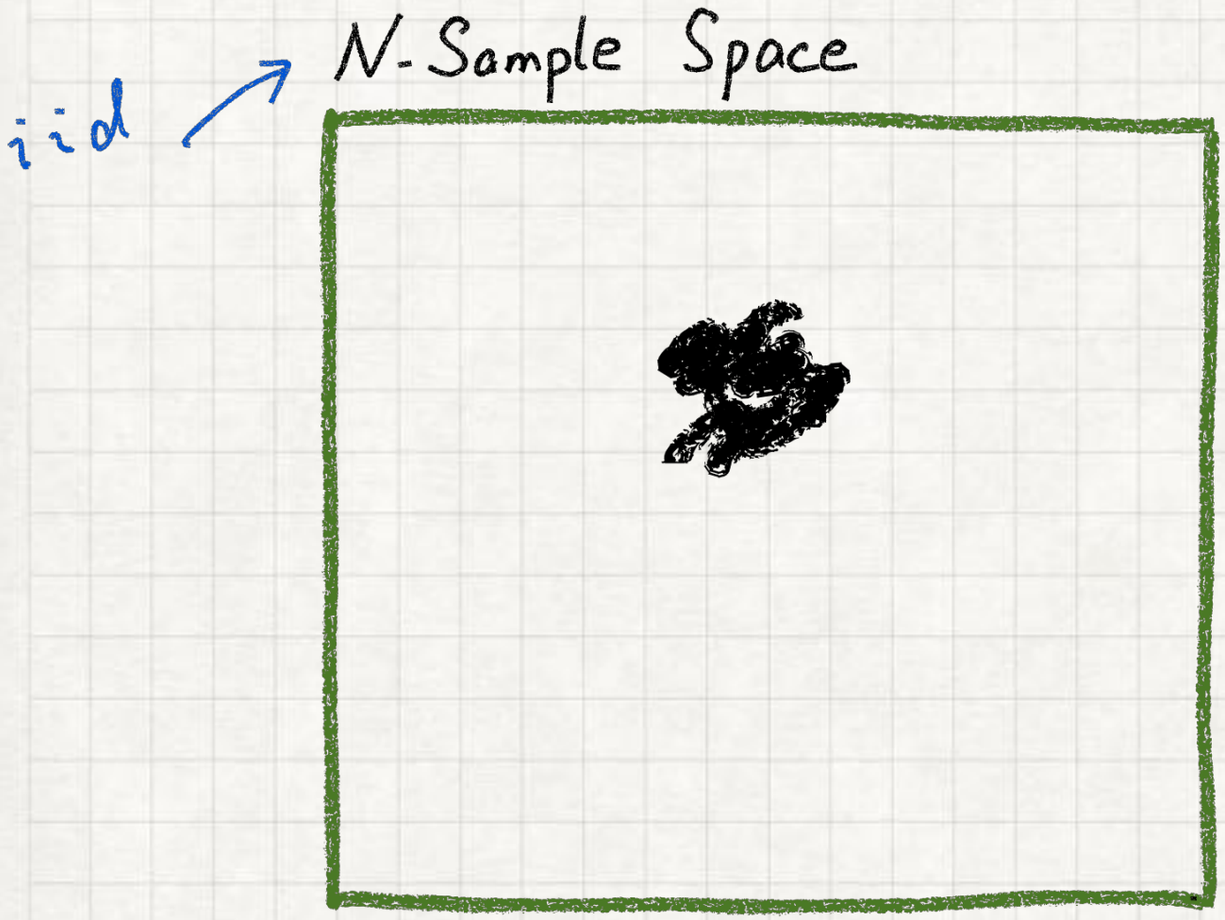
Hoeffding's inequality says when N increases, the portion of such "black-tester" for a hypothesis drops exponentially with N.



BAD AREA FOR h IN N-SAMPLE SPACE



BAD AREA FOR h IN N-SAMPLE SPACE



BAD AREA FOR h IN N-SAMPLE SPACE

$$Area^{Bad} \leq B$$

$$B \propto e^{-2\epsilon^2 N}$$

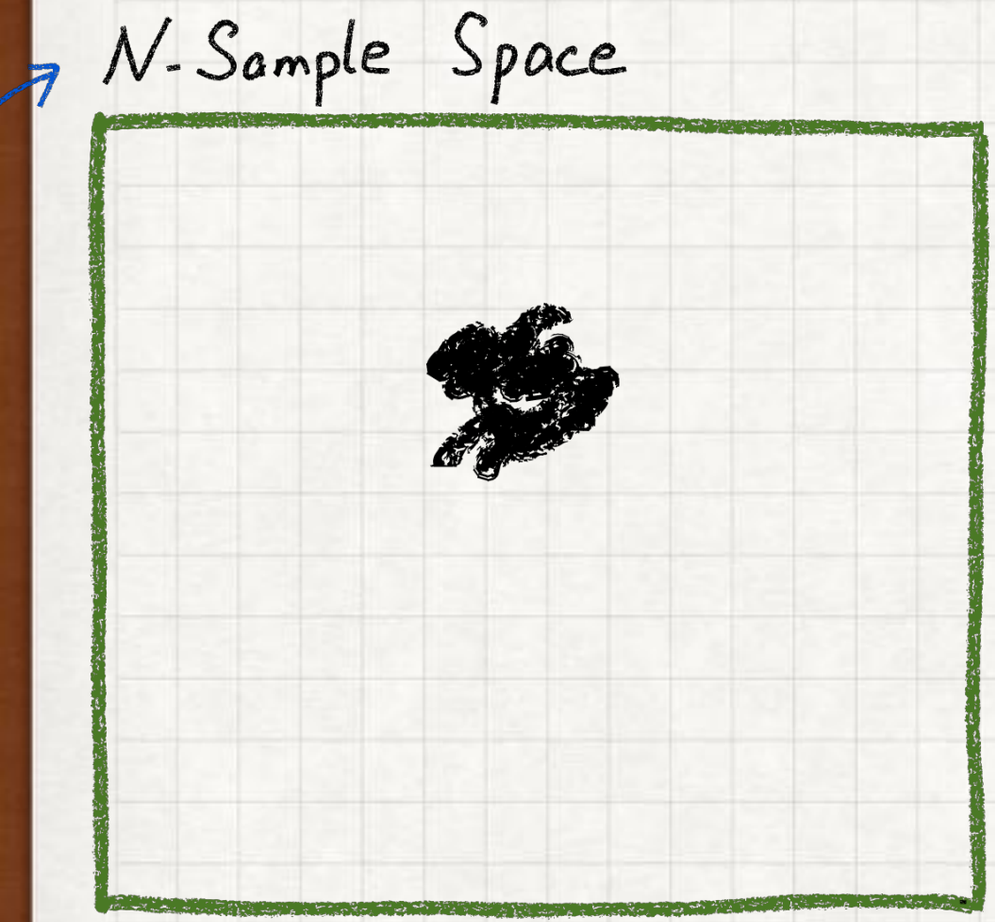
Q: What is ϵ?

A. ϵ is directly determined by the training error E_{in}.

B. ϵ is directly determined by the generalisation error E_{out}.

C. Black-marker: ϵ is the predetermined tolerance on the difference between E_{in} and E_{out}. E_{in} is evaluated on a dataset D, say, $E_{in}[h; D]$. If $|E_{out} - E_{in}| > \epsilon$, the dataset D will be marked black in the N-Sample space.

BAD AREA FOR h IN N-SAMPLE SPACE



$$Area^{Bad} \leq B$$

$$B \propto e^{-2\epsilon^2 N}$$

Q: What is ϵ?

A. ϵ is directly determined by the training error E_{in}.

B. ϵ is directly determined by the generalisation error E_{out}.

C. Black-marker: ϵ is the predetermined tolerance on the difference between E_{in} and E_{out}. E_{in} is evaluated on a dataset D, say, $E_{in}[h; D]$. If $|E_{out} - E_{in}| > \epsilon$, the dataset D will be marked black in the N-Sample space.

BAD AREA FOR h IN N-SAMPLE SPACE

$$Area^{Bad} \leq B$$

$$B \propto e^{-2\epsilon^2 N}$$

Q: What is ϵ?

A. ϵ is directly determined by the training error E_{in}.

B. ϵ is directly determined by the generalisation error E_{out}.

C. Black-marker: ϵ is the predetermined tolerance on the difference between E_{in} and E_{out}. E_{in} is evaluated on a dataset D, say, $E_{in}[h; D]$. If $|E_{out} - E_{in}| > \epsilon$, the dataset D will be marked black in the N-Sample space.

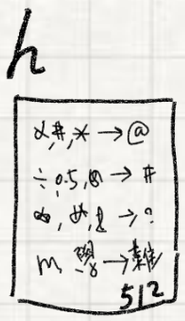
MOD3
BOUND FOR TRAINED MODEL GENERALISATION ERROR

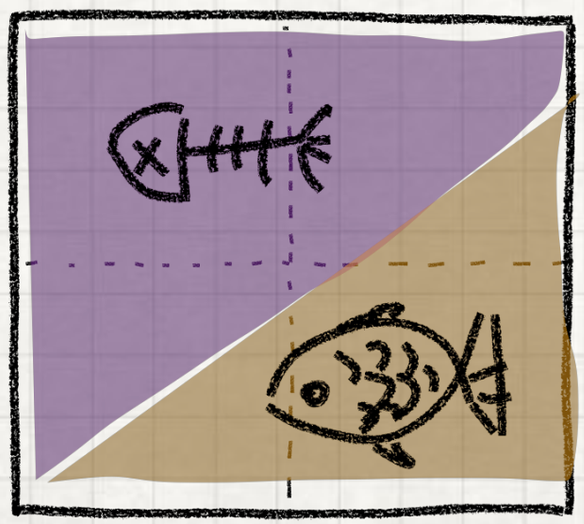
BAD AREA FOR FIXED h

Q: Machine Learning is simple as we need only to increase the number of samples so the estimation of error is accurate enough.

A. Yes.

B. No.



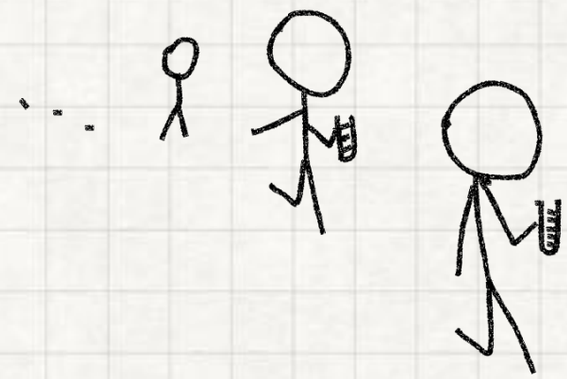


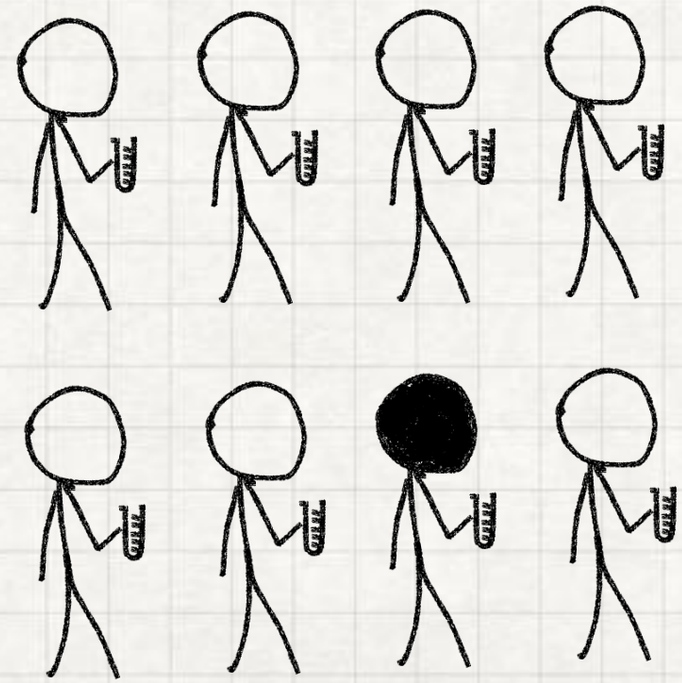
BAD AREA FOR FIXED h

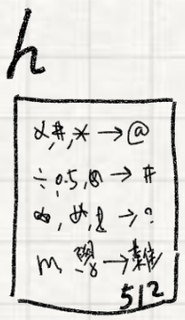
Q: Machine Learning is simple as we need only to increase the number of samples so the estimation of error is accurate enough.

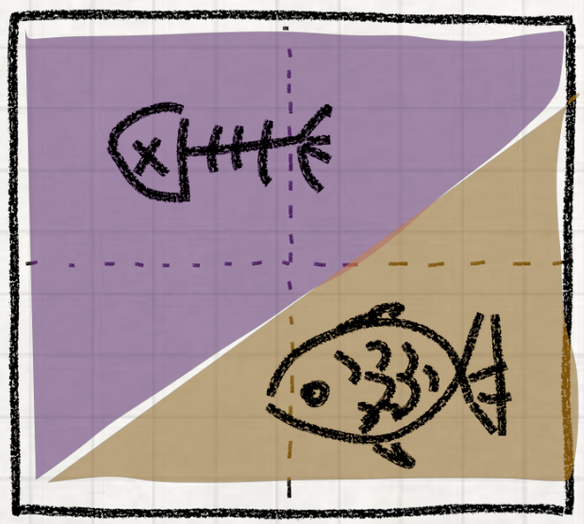
A. Yes.

B. No.



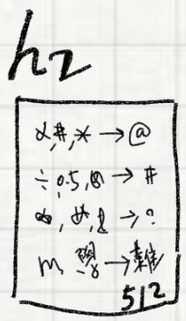


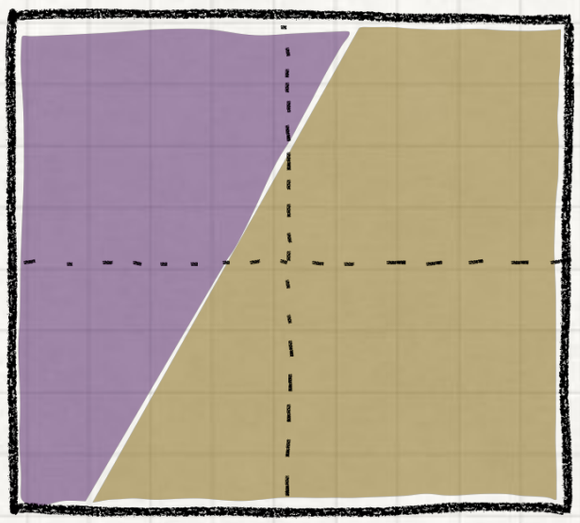


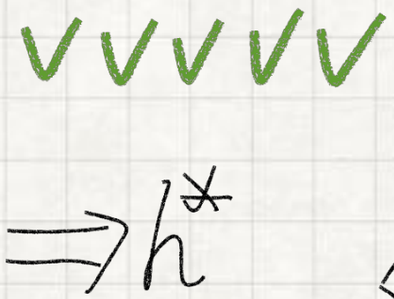


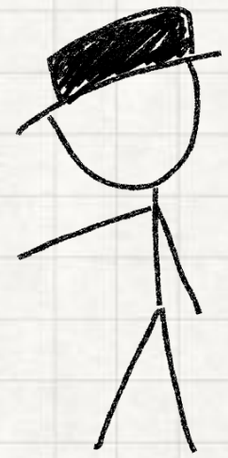
CONSIDER LEARNED h

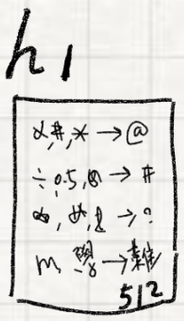
We consider multiple hypotheses and select the one that minimises $E_{in}[h^*; D_{train}]$.

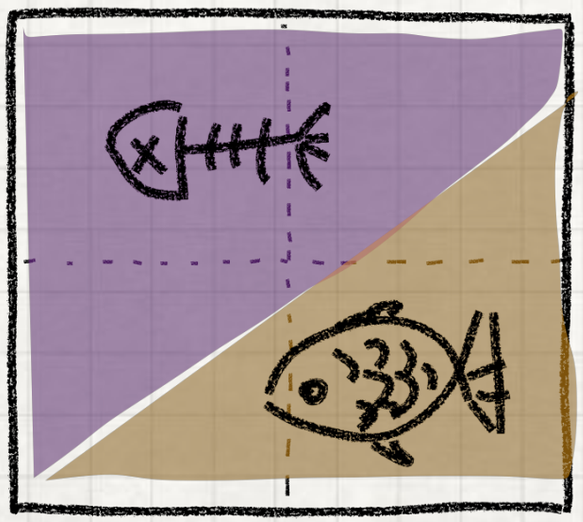




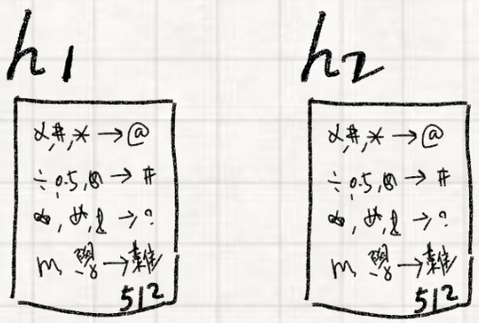


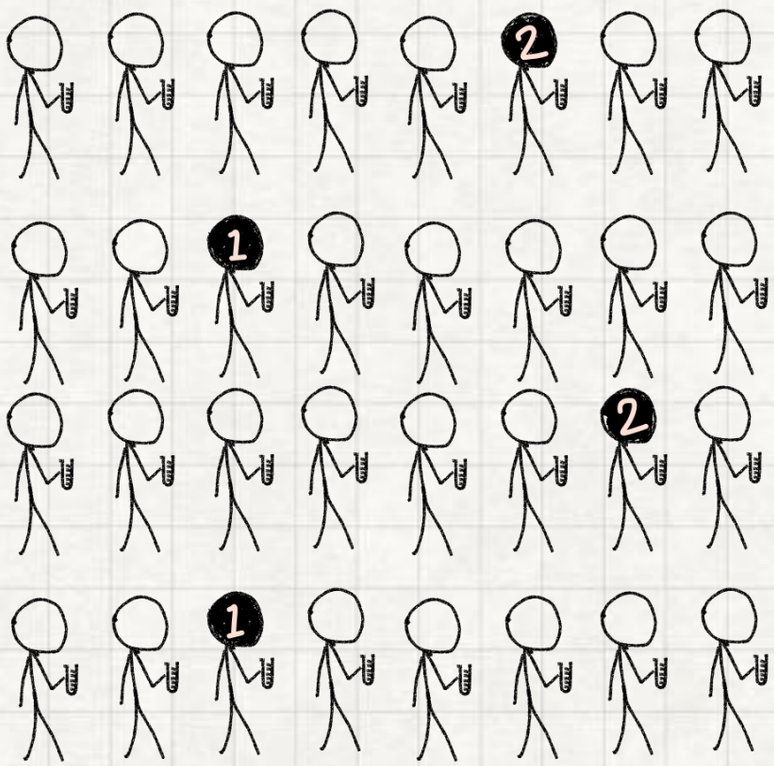






MULTIPLIED RISK FOR LEARNED h





Q: Consider the **risk**

$$|E_{in}[h^*; D_{train}] - E_{out}[h^*]| > \epsilon,$$

i.e. our D_{train} is black. How this risk changes in the following two cases:

i.) h^* is selected from two hypotheses $\{h_1, h_2\}$ by comparing $E_{in}[h_1; D_{train}]$ and $E_{in}[h_2; D_{train}]$ and taking the one with small E_{in}

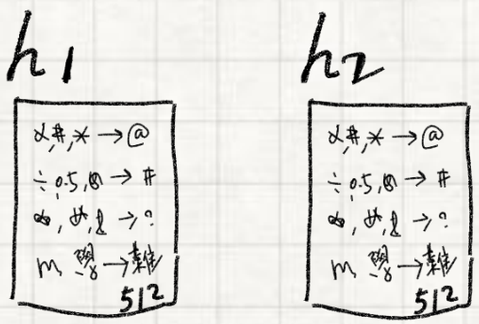
ii) h^* is fixed?

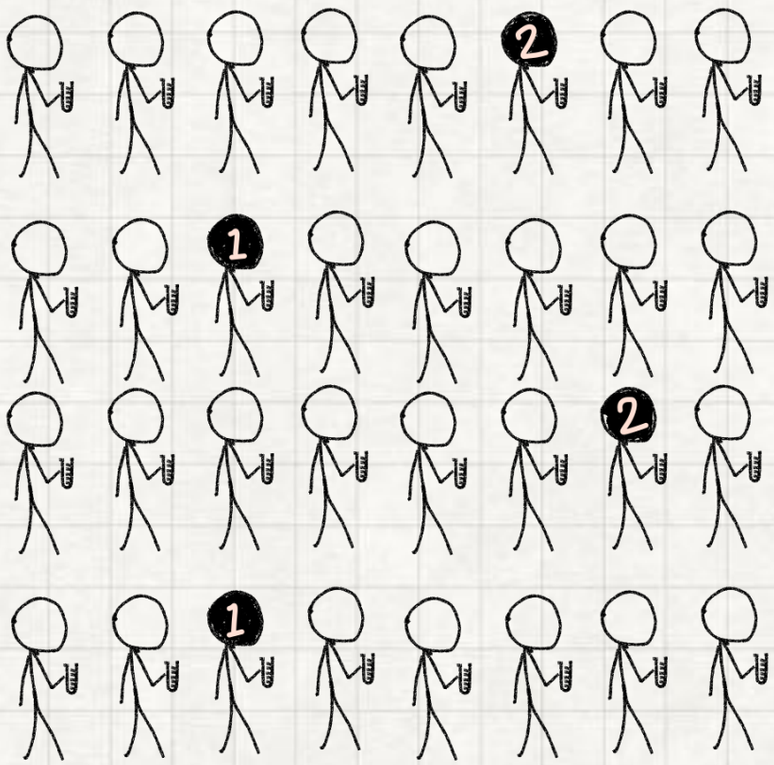
A. risk of i) is smaller, as we may reduce training error E_{in}

B. risk of i) is greater, as it is more probable that D_{train} happens to be a bad dataset for either hypotheses (than it would have been for only one of them).

C. risk of i) and ii) is the same, as we only select one hypothesis, the risk is only for D_{train} being bad for the selected hypothesis.

MULTIPLIED RISK FOR LEARNED h FROM A LARGE $\mathscr{H}$

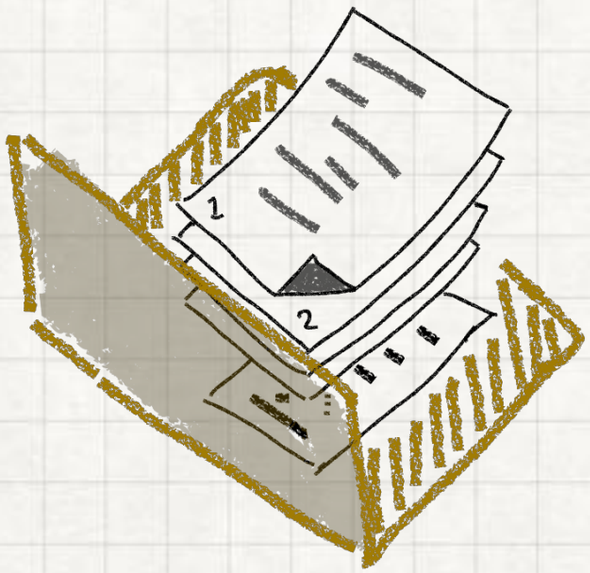


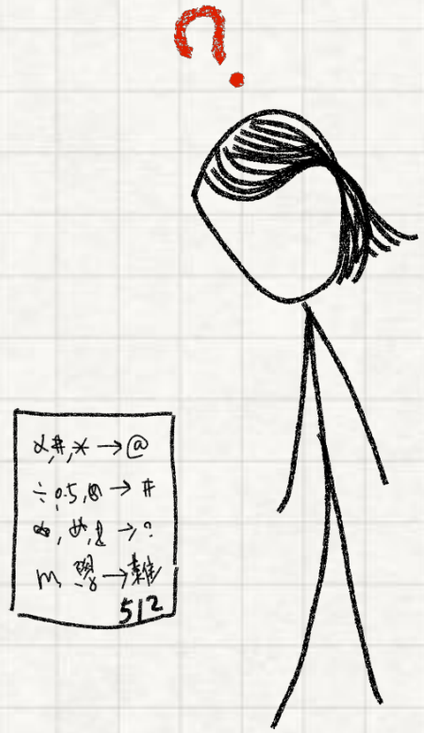


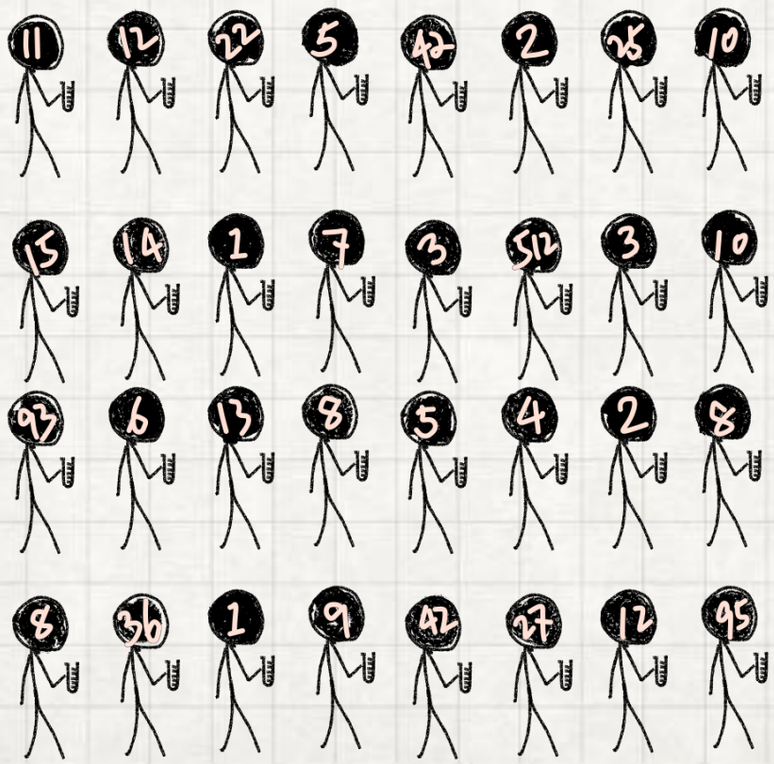
The risk increases for the learned hypothesis. Any guarantee about the result of learning, must take into account the risk of the training dataset is "black" for all hypotheses candidates.

The problem is that any sensibly useful family of candidate hypotheses contains TOO MANY members ...

RISK EXPLOSION FOR LEARNED h FROM A LARGE $\mathscr{H}$

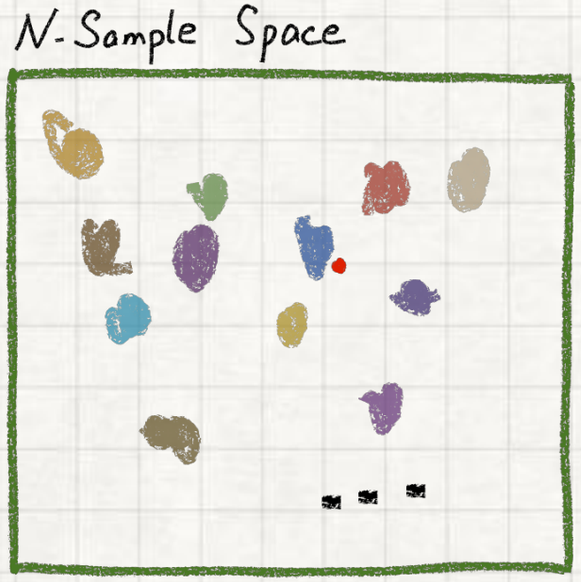


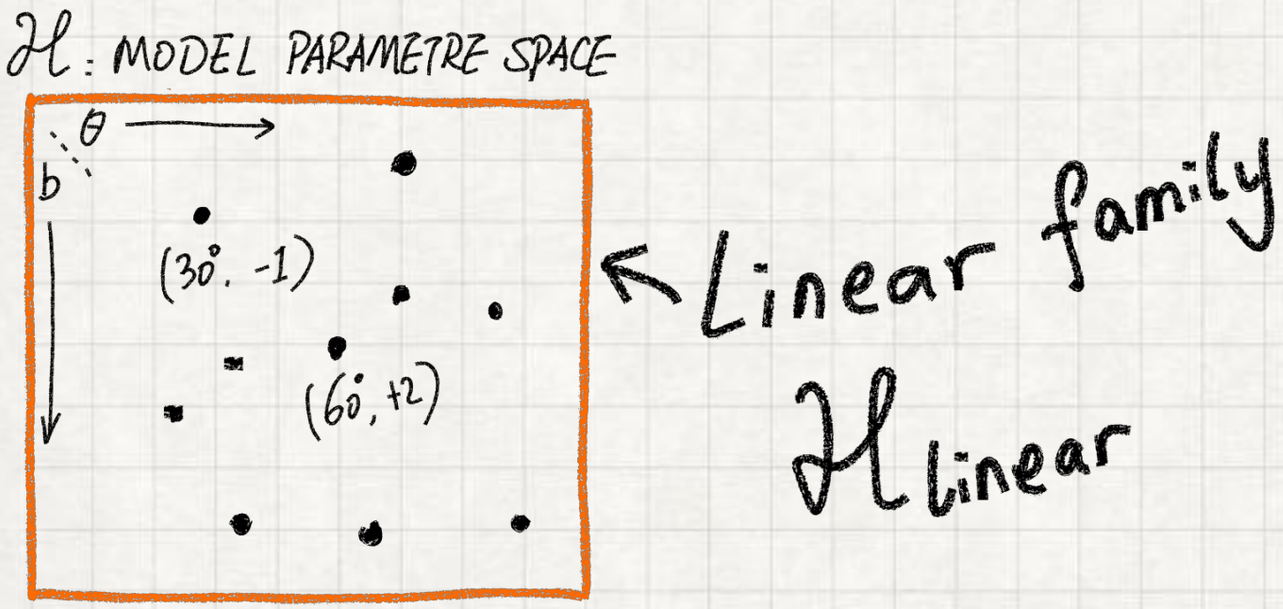




The problem is that any sensibly useful family of candidate hypotheses contains TOO MANY members ...

GREATER N CANNOT HELP



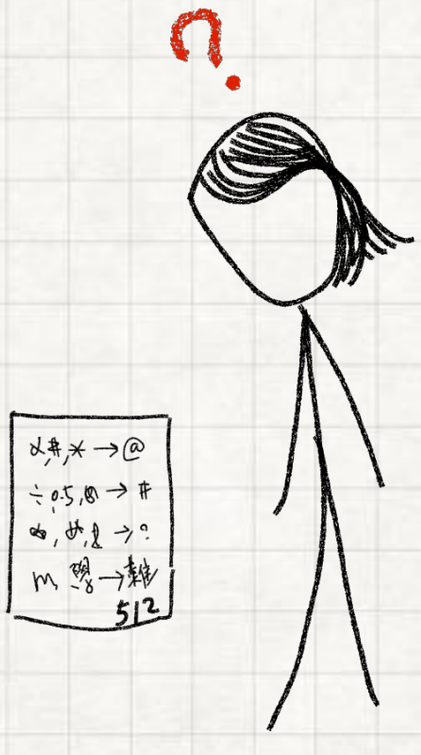


- Let N increase. Hoeffding's inequality tells us the areas will shrink $\propto e^{-2\epsilon^2 N}$.

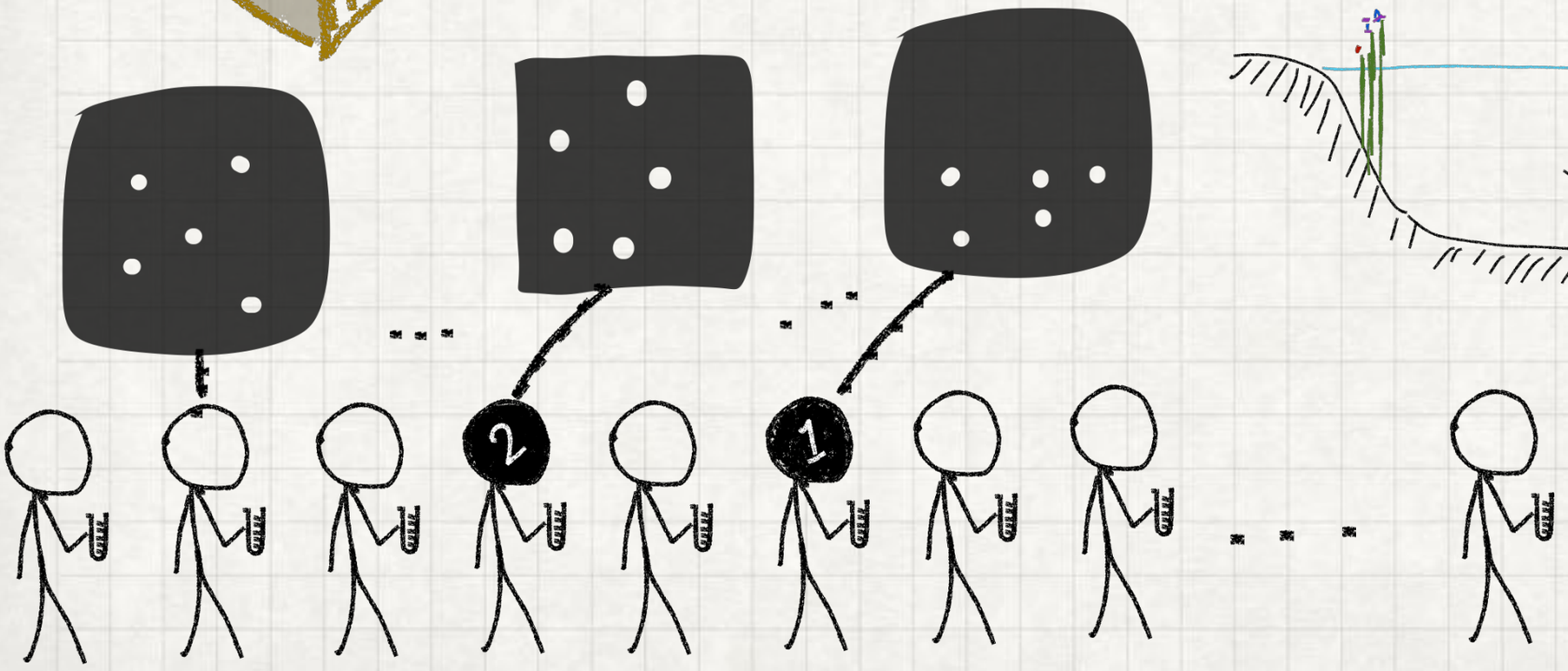
- But we are considering infinitely many hypotheses, the gross total bad areas will cover everywhere ($|\mathcal{H}| = \infty$ times *anything* has no bound) regardless how small an individual bad area is. And we have no confidence on what has been learned.

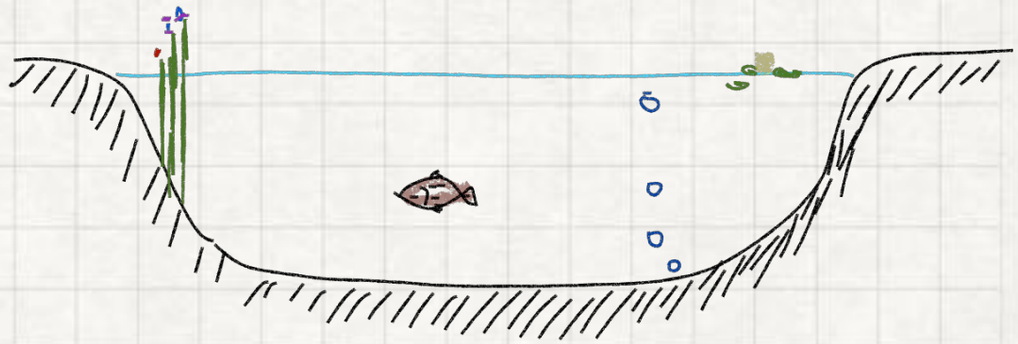
MOD4
CLEVER COUNTING: VC-DIMENSION , MODEL COMPLEXITY AND ERROR BOUND

ALL HOPE IS NOT LOST

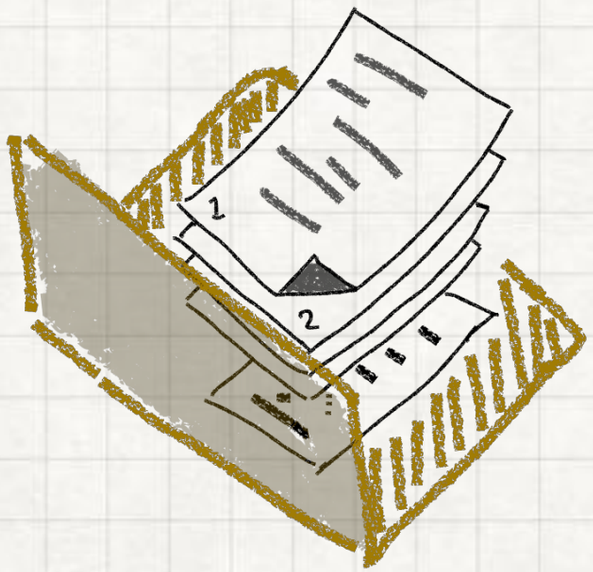


Recall that a field worker provides N "holes" (a dataset $D^{(N)}$ of N samples) through which we can evaluate hypothesis $E_{in}[h; D^{(N)}]$.

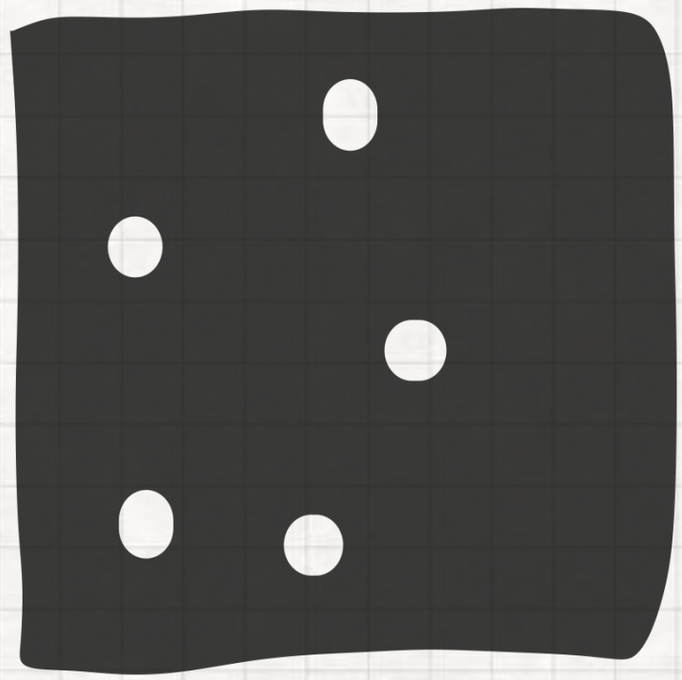


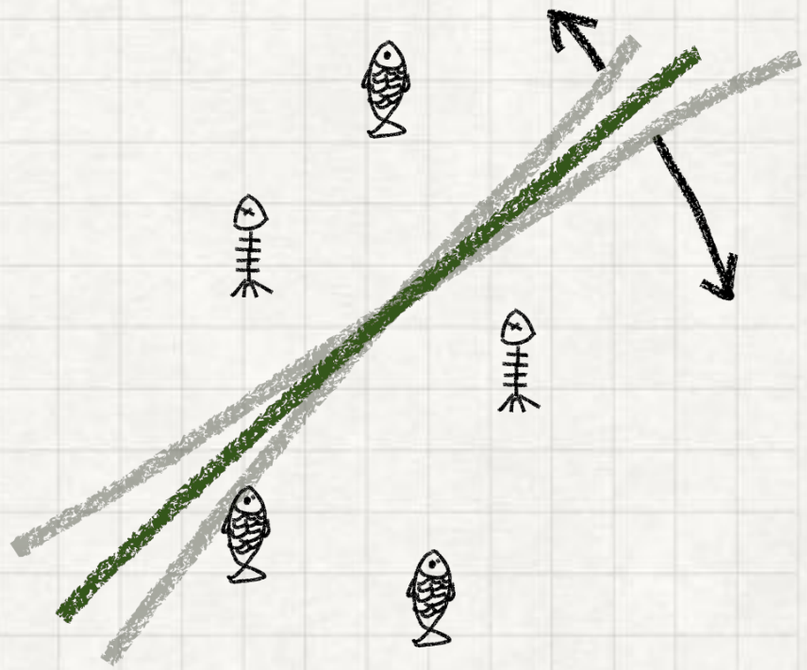


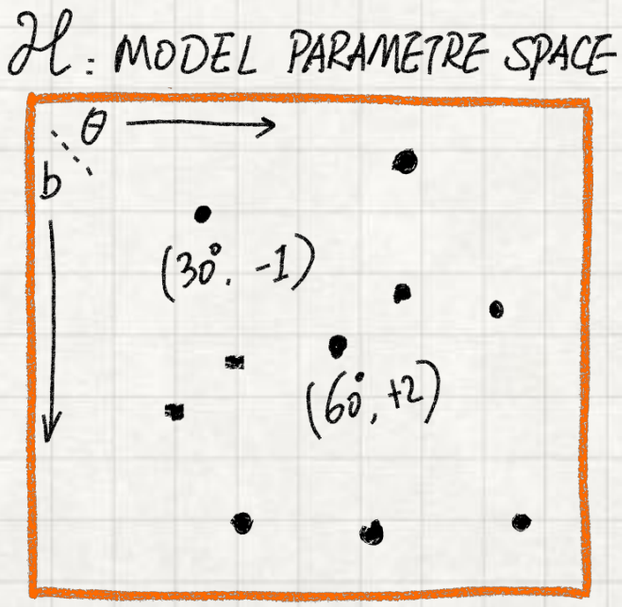
"IMPLEMENTABLE DICHOTOMIES" ON N SAMPLES



- Risk is caused by the dataset happen to generate a bad E_{in}
- Since any training error $E_{in}[h; D^{(N)}]$ is the result of classifying N training samples.
- If a family $\mathscr{H}$ is "weak", it can only produce a subset of all possible classification results of N samples.

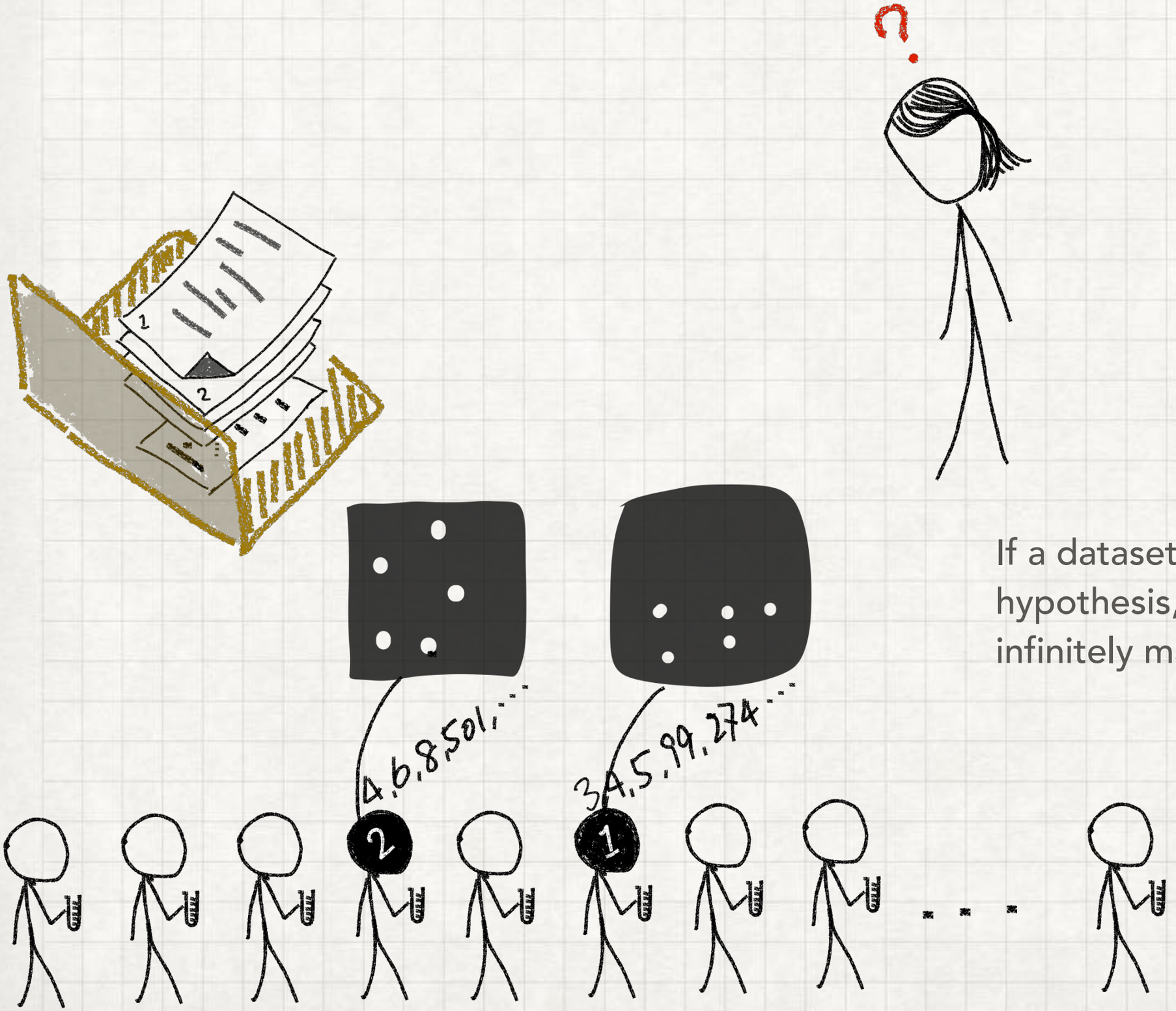






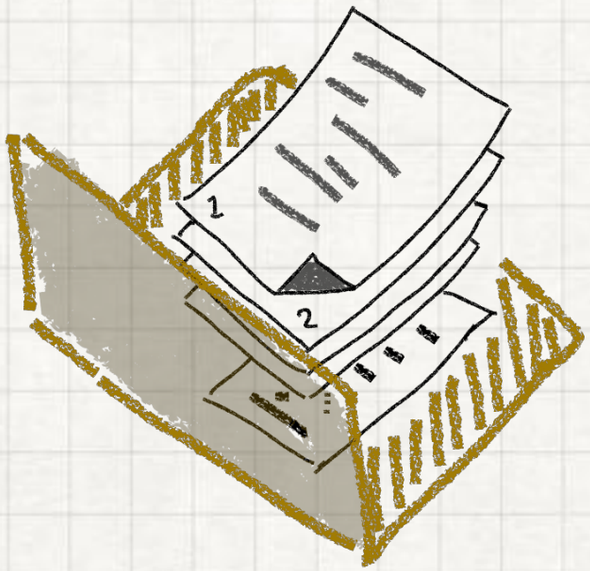
- Consider linear family — it is impossible to generate all +/- combinations.

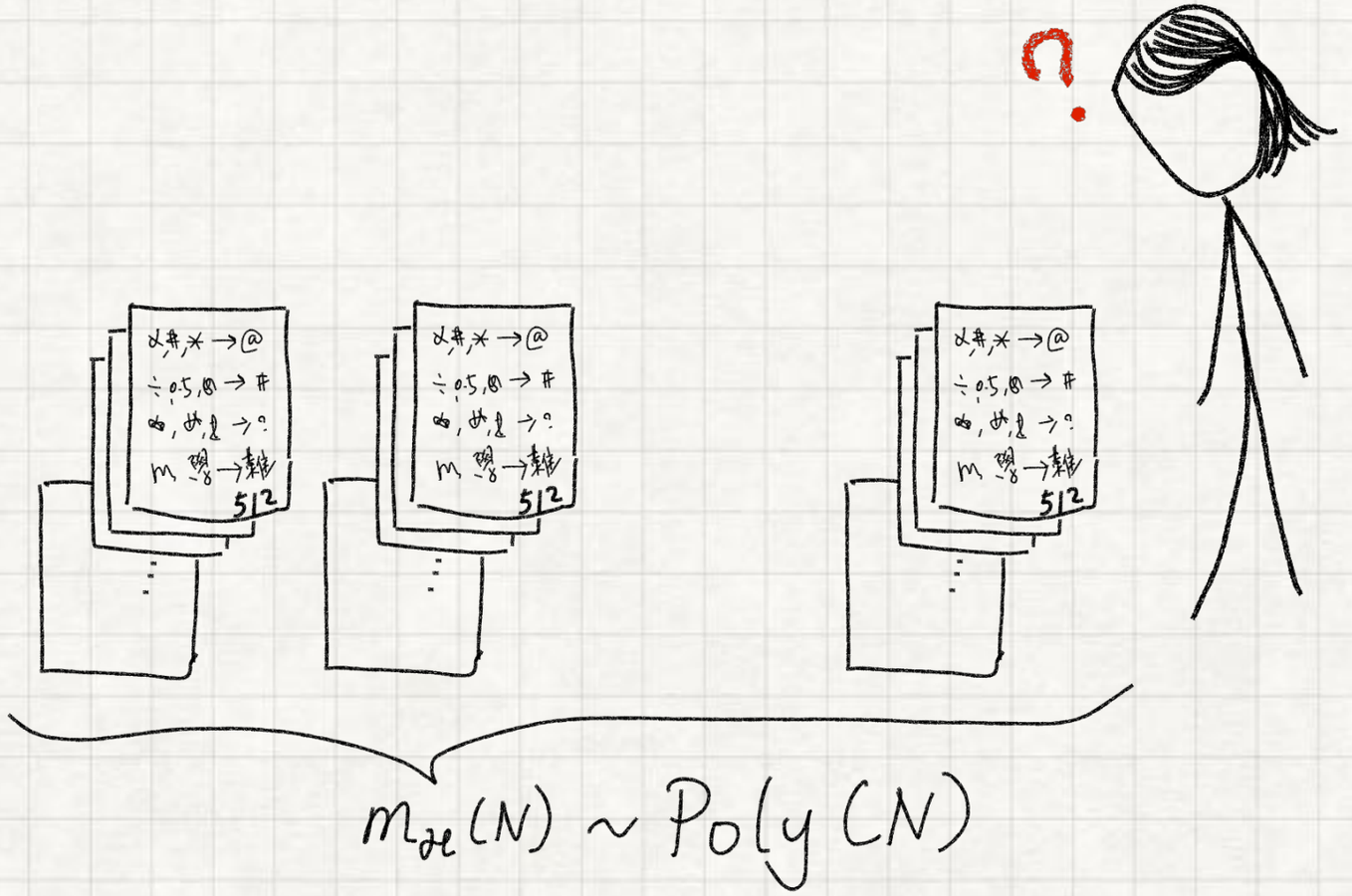
STOP OVER COUNTING BAD DATASETS



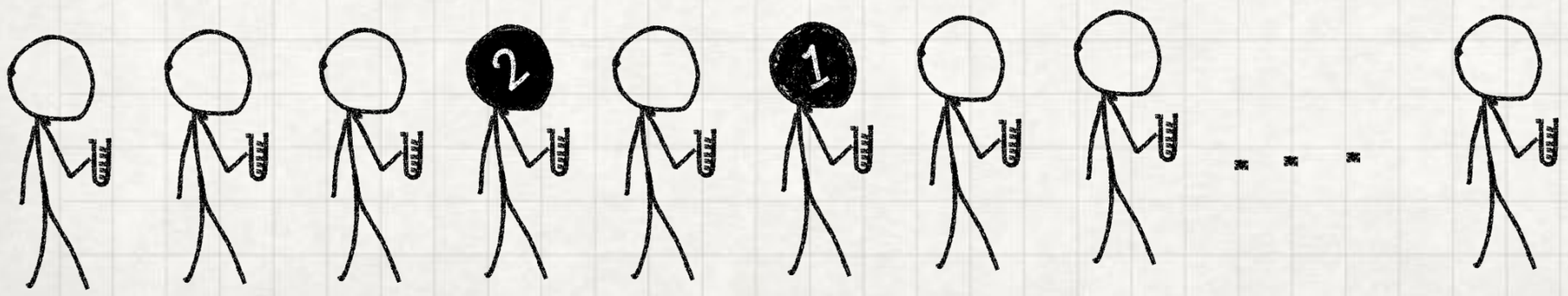
If a dataset is bad for one hypothesis, it is bad for many, infinitely many.

A LIMITED MULTIPLICATION FACTOR

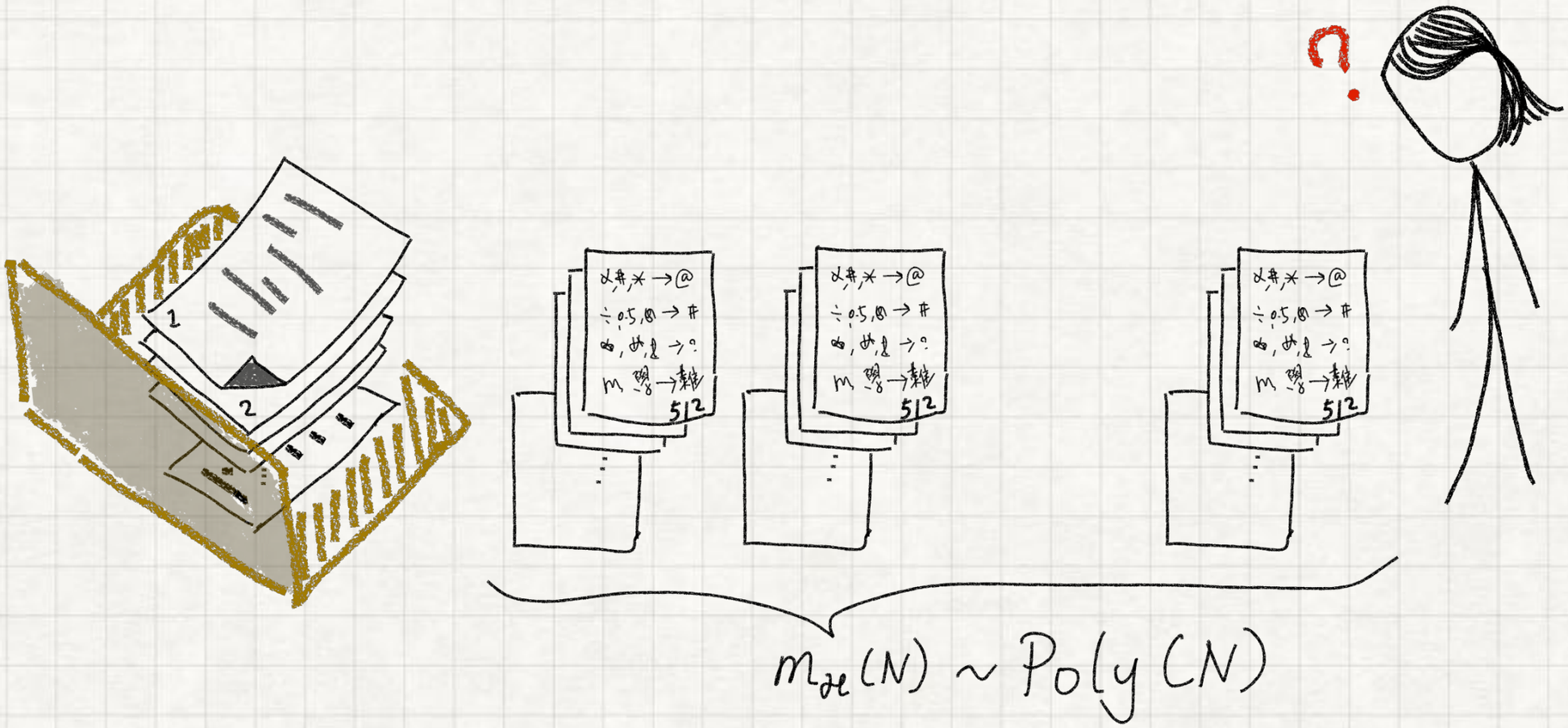




The multiplication factor is much less than $|\mathscr{H}|$, and less than e^N, it is polynomial of N. ($O(N^5)$, $O(N^2)$, etc.)



PUT TOGETHER



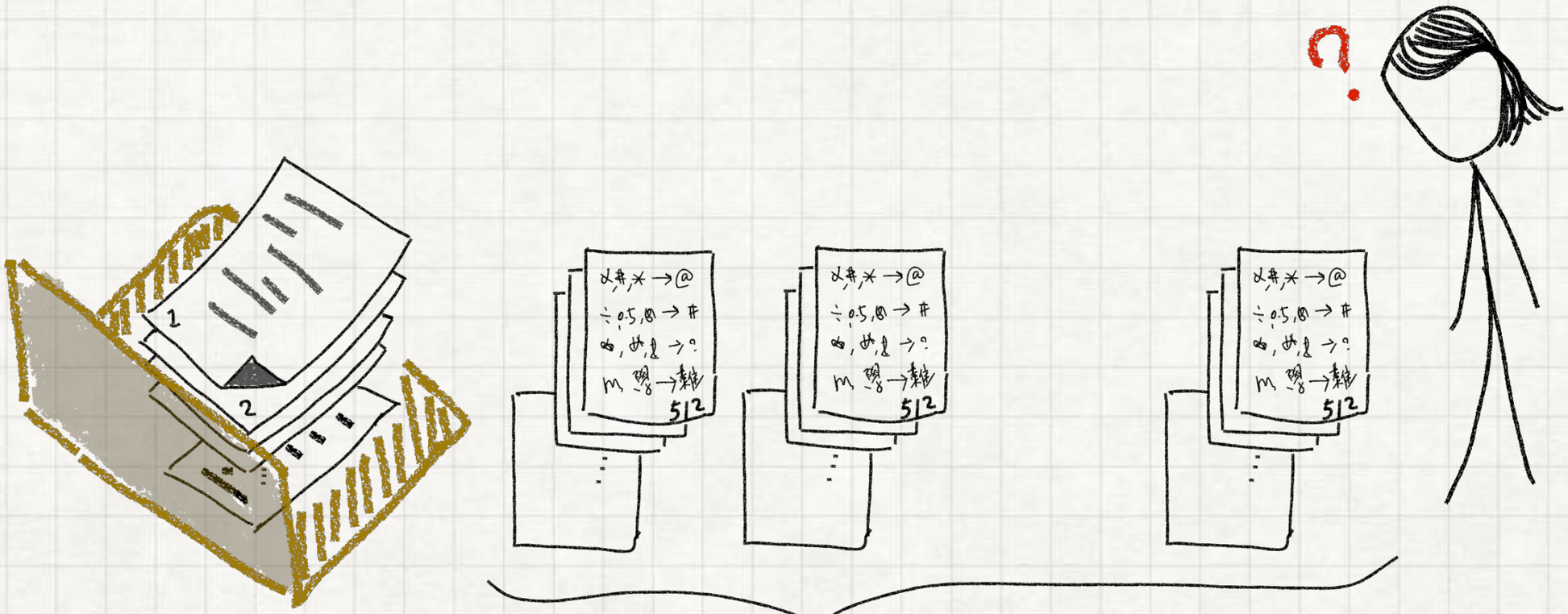
Multiplication factor $m_{\mathcal{H}} \ll e^N \ll |\mathcal{H}|$

$$P\{h \text{ learned from } \mathcal{H} : |E_{in}[h] - E_{out}[h]| > \epsilon)\}$$

$$\leq m_{\mathcal{H}}(N) \times \text{Hoeffding's-Ineq.-said-about-}N$$

$$= m_{\mathcal{H}}(N) 2 \cdot e^{-2\epsilon^2 N}$$

PUT TOGETHER



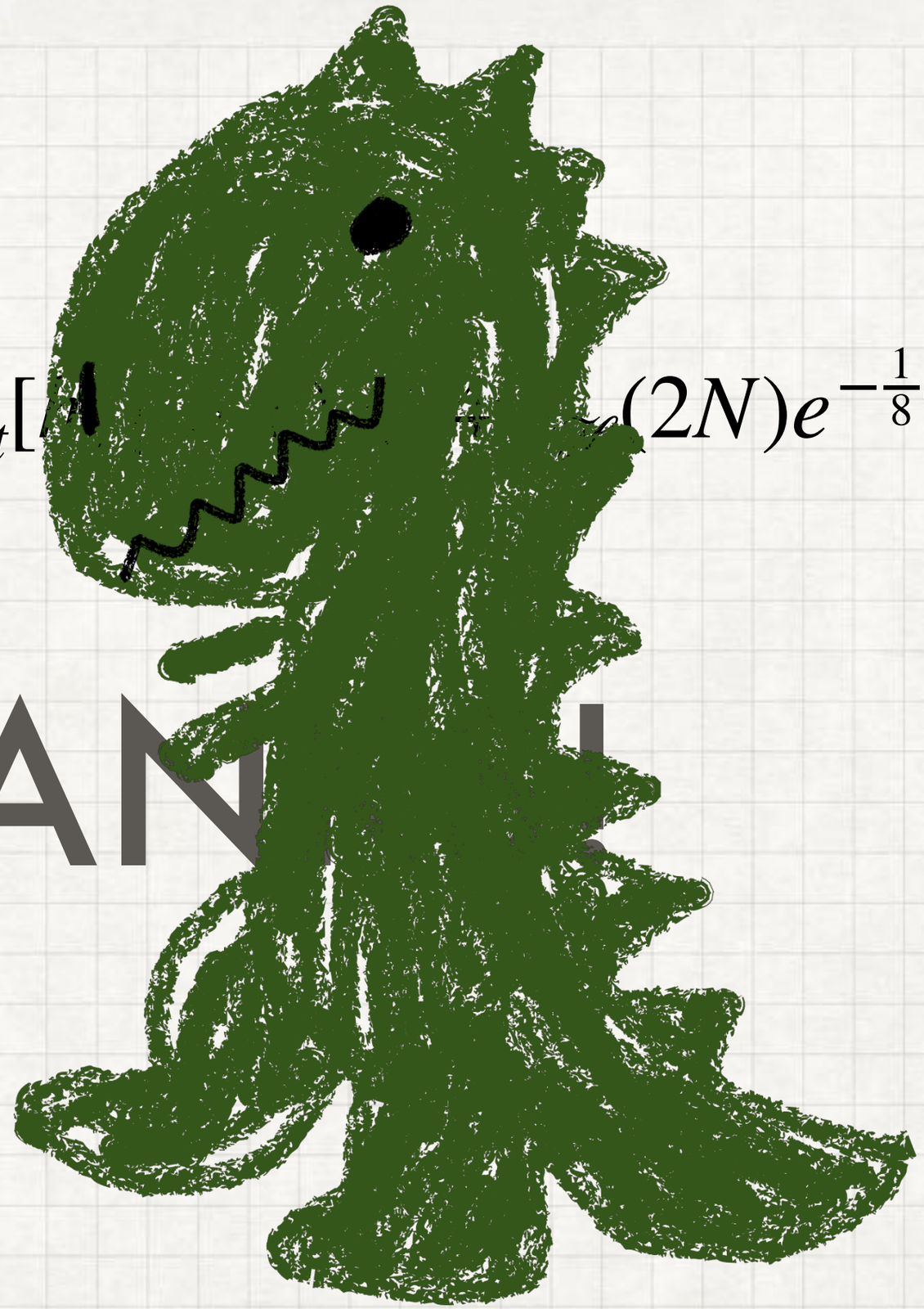
Multiplication factor $m_{\mathcal{H}} \ll e^N \ll |\mathcal{H}|$

$$P\{h \text{ learned from } \mathcal{H} : |E_{in}[h] - E_{out}[h]| > \epsilon)\}$$

$$\leq m_{\mathcal{H}}(N) \times \text{Hoeffding's-Ineq.-said-about-}N$$

$$= m_{\mathcal{H}}(N) 2 \cdot e^{-2\epsilon^2 N}$$

$$P(\sup_{h \in \mathcal{H}} |E_{in}[h] - E_{out}[h]| > \epsilon) \leq \mathcal{H}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$



THANK

$$P(\sup_{h \in \mathcal{H}} |E_{in}[h] - E_{out}[h]| > \epsilon) \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}$$

THANKS!

LAST BUT NOT LEAST …

TEST DATASET

- $m_{\mathcal{H}}$ still tends to be very high for many useful families, say deep networks.
- So after selecting h^* from $\mathcal{H}$ ...

- We test h^* on a separate dataset D_{test}
- D_{test} never affected h^*
- The fixed-h Hoeffding's inequality applies — with high probability

$$|E_{in:test} - E_{out}| < \epsilon$$

Never corrupt your test set!

THANKS!

THANKS!