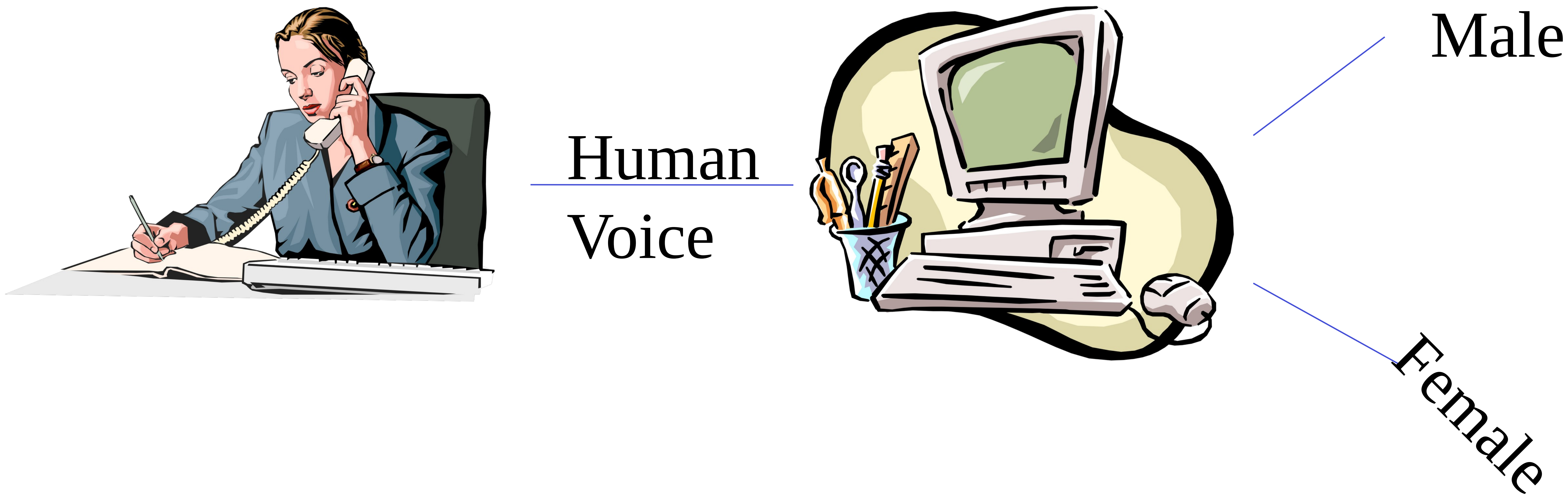
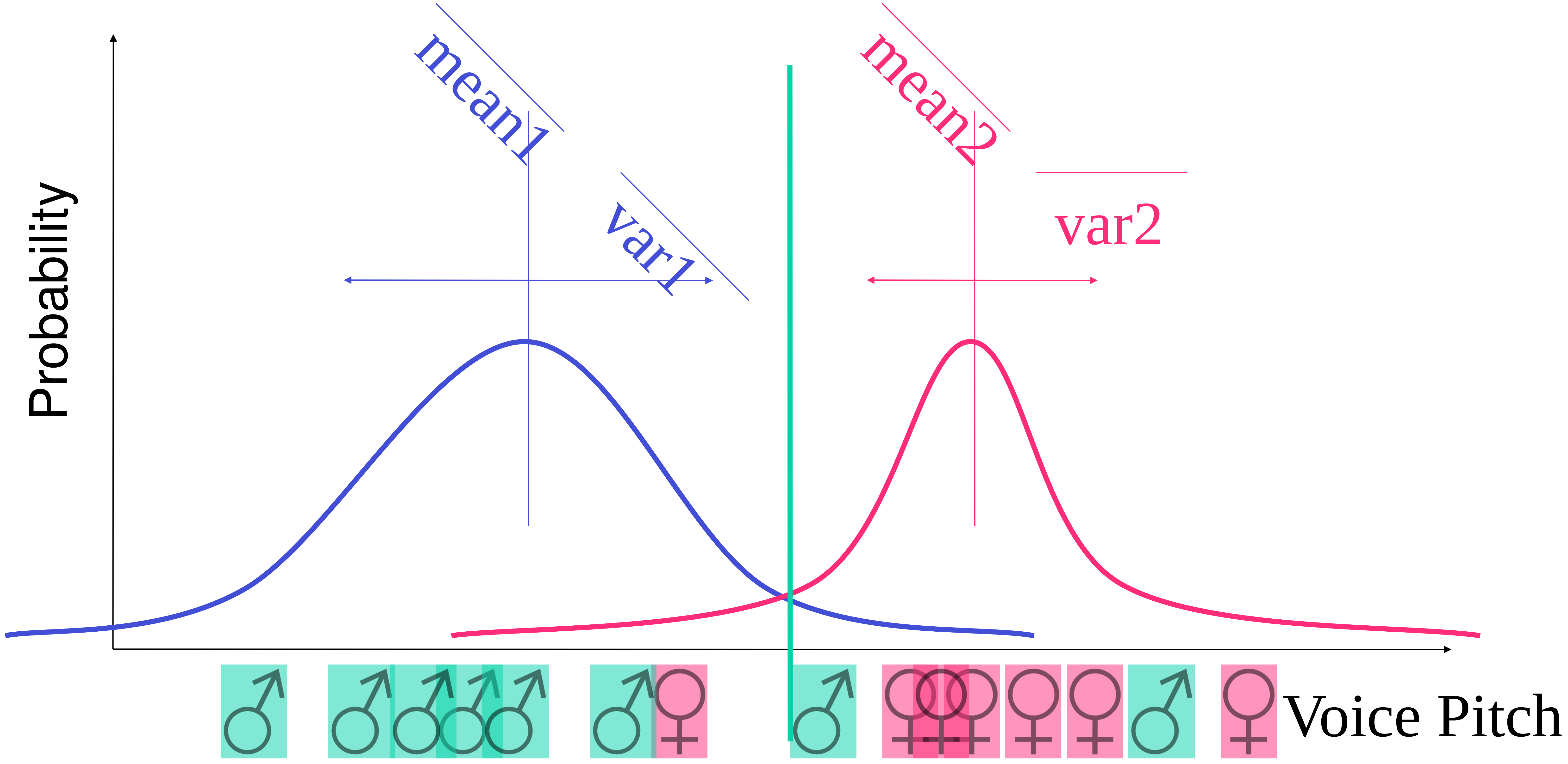
Generative vs. Discriminative Optimization vs. Elimination

Toy Example

- Computer receives telephone call
- Measures Pitch of voice
- Decides gender of caller

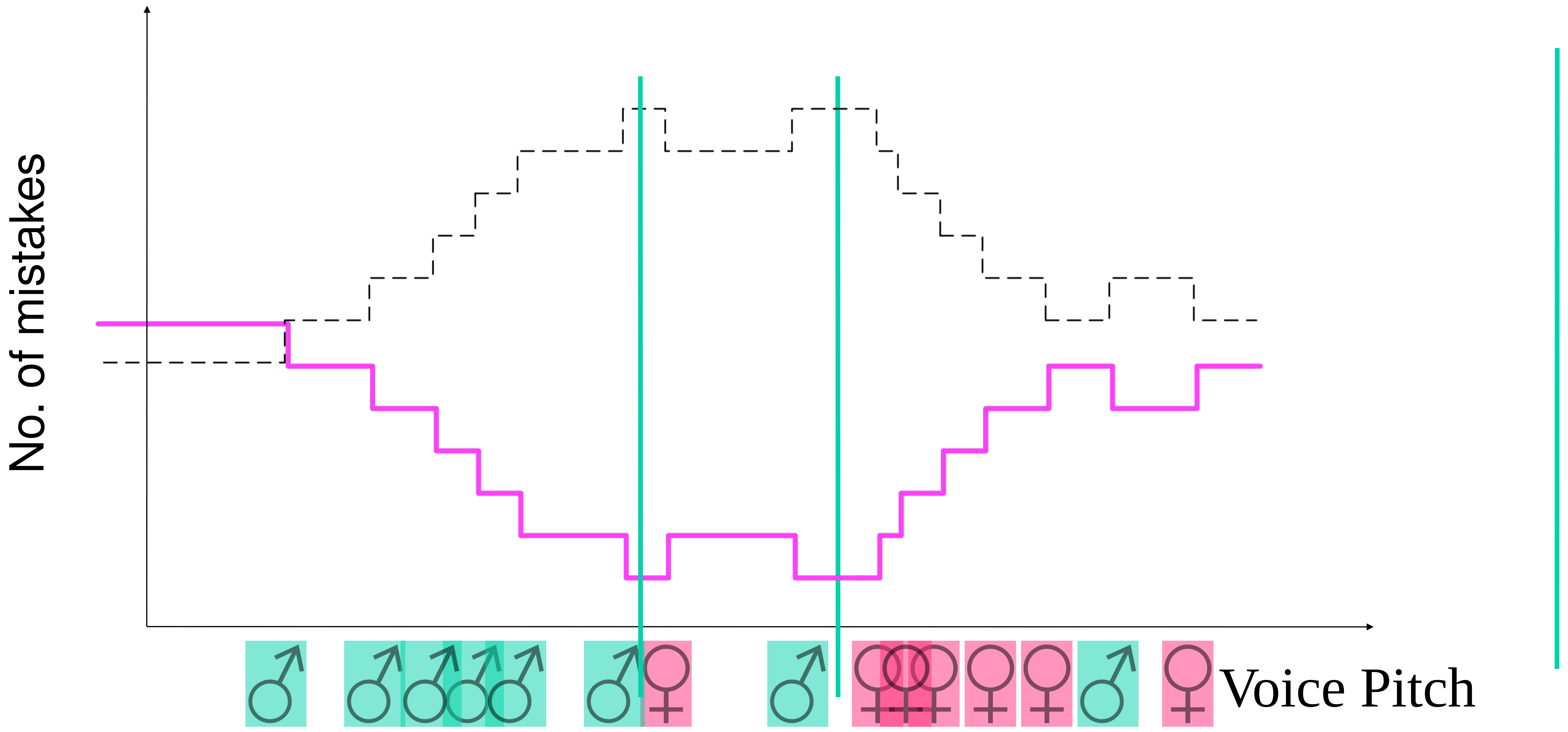


Generative modeling

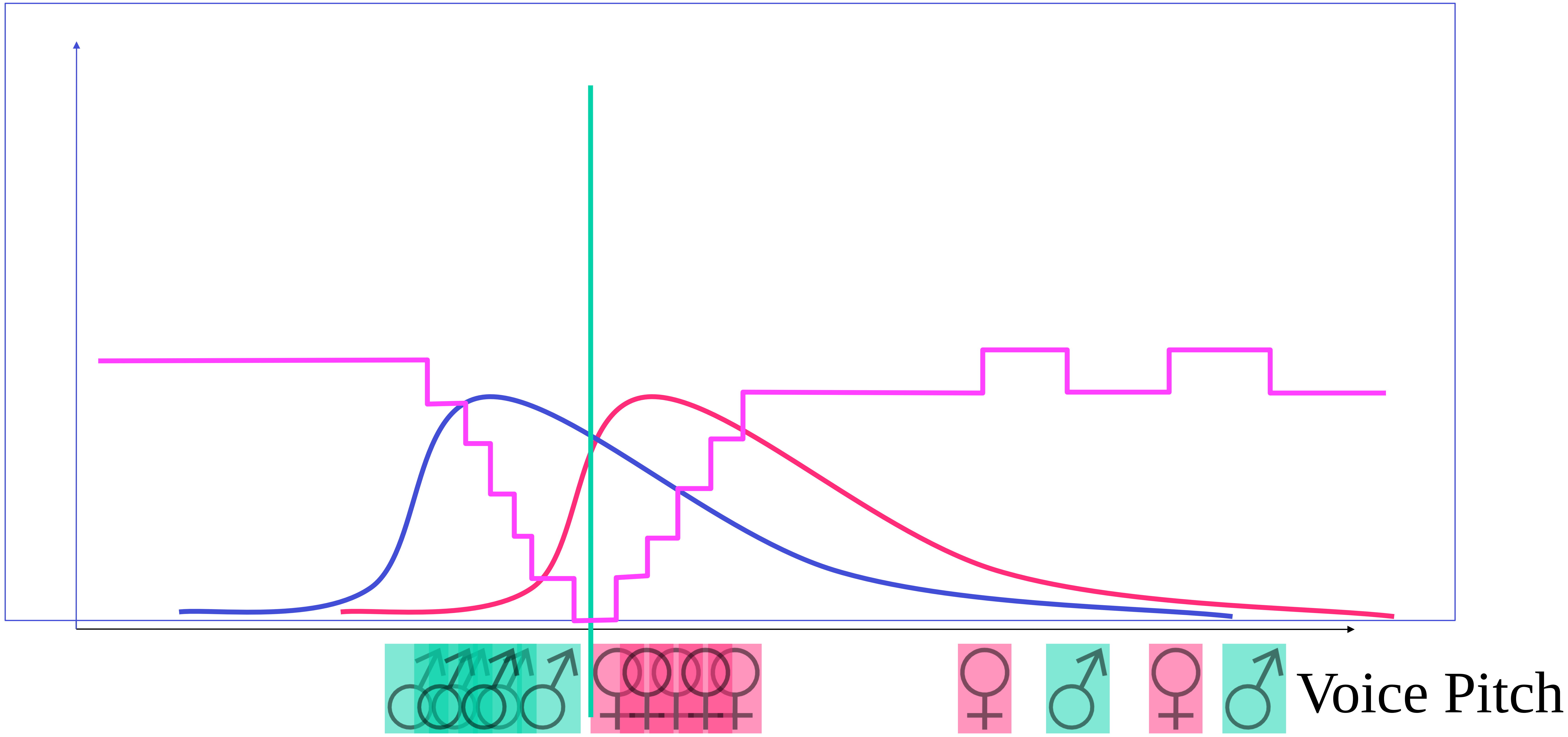


Discriminative approach

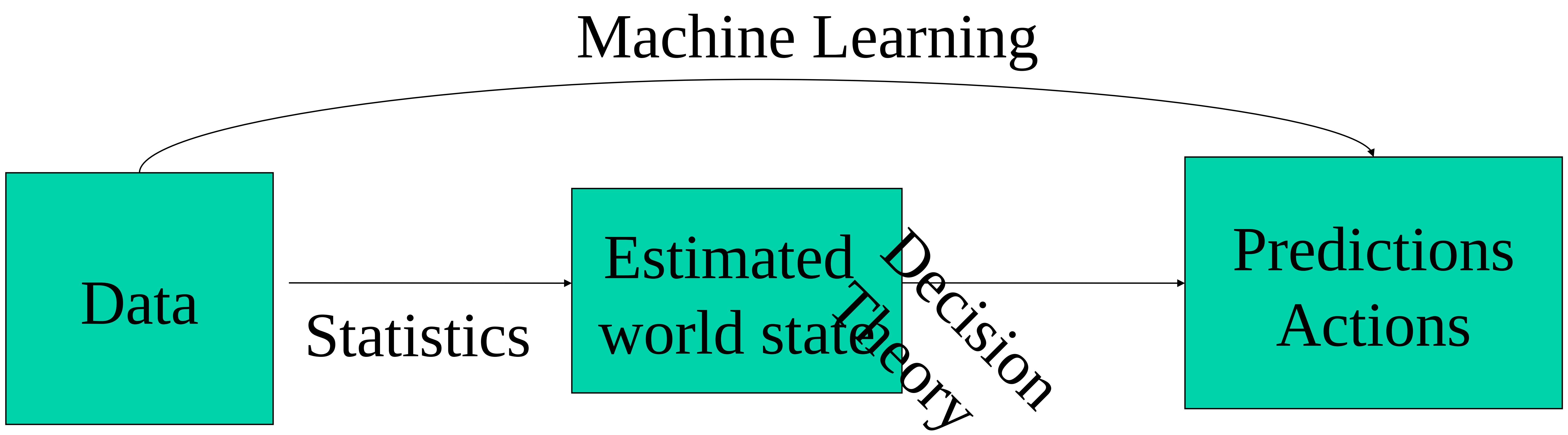
[Vapnik 85]



Ill-behaved data



Traditional Statistics vs. Machine Learning



Comparison of methodologies

Model	Generative	Discriminative
Goal	Probability estimates	Classification rule
Performance measure	Likelihood	Misclassification rate
Mismatch problems	Outliers	Misclassifications

Summary

- Generative models: goal is to explain how data is generated.
- Discriminative models: goal is to predict a property of the data (such as label)
- Generative models are more accurate when they are correct.
- Discriminative models are more robust against poor modeling or outliers.

Confidence vs. Certainty

- **Certainty** is a statement about the true distribution.
  $$P(y = +1|x) > 0.99 \; or \; P(y = +1|x) < 0.01$$
- **Confidence** is a statement about my knowledge.
  $$P(y = +1|x, training\ set) > 0.95\ (< 0.05)$$
- **Confidence** depends on the training set, **certainty** does not.

Generative vs. Discriminative vs. Robust discriminative

- Generative: Data is generated by model
- Discriminative: there is a model whose error rate is low.
- Robust discriminative:
  - There are many models whose error rate is small. (the good set)
- Easy Examples: most of the good set predicts the same way.

confidence $\neq$ certainty

- **High Certainty and high confidence:**
  - The sun will rise tomorrow.
  - This spring quarter will be over in June
  - There will be new common variant of covid in the coming year.
- **Low certainty but high Confidence:**
  - You go fishing in a new location. You know nothing about the probability of catching a a fish: low confidence.
  - You go fishing for 100 days. You have high confidence that your probability of catching a fish is larger than 5%

More examples of high confidence, low certainty

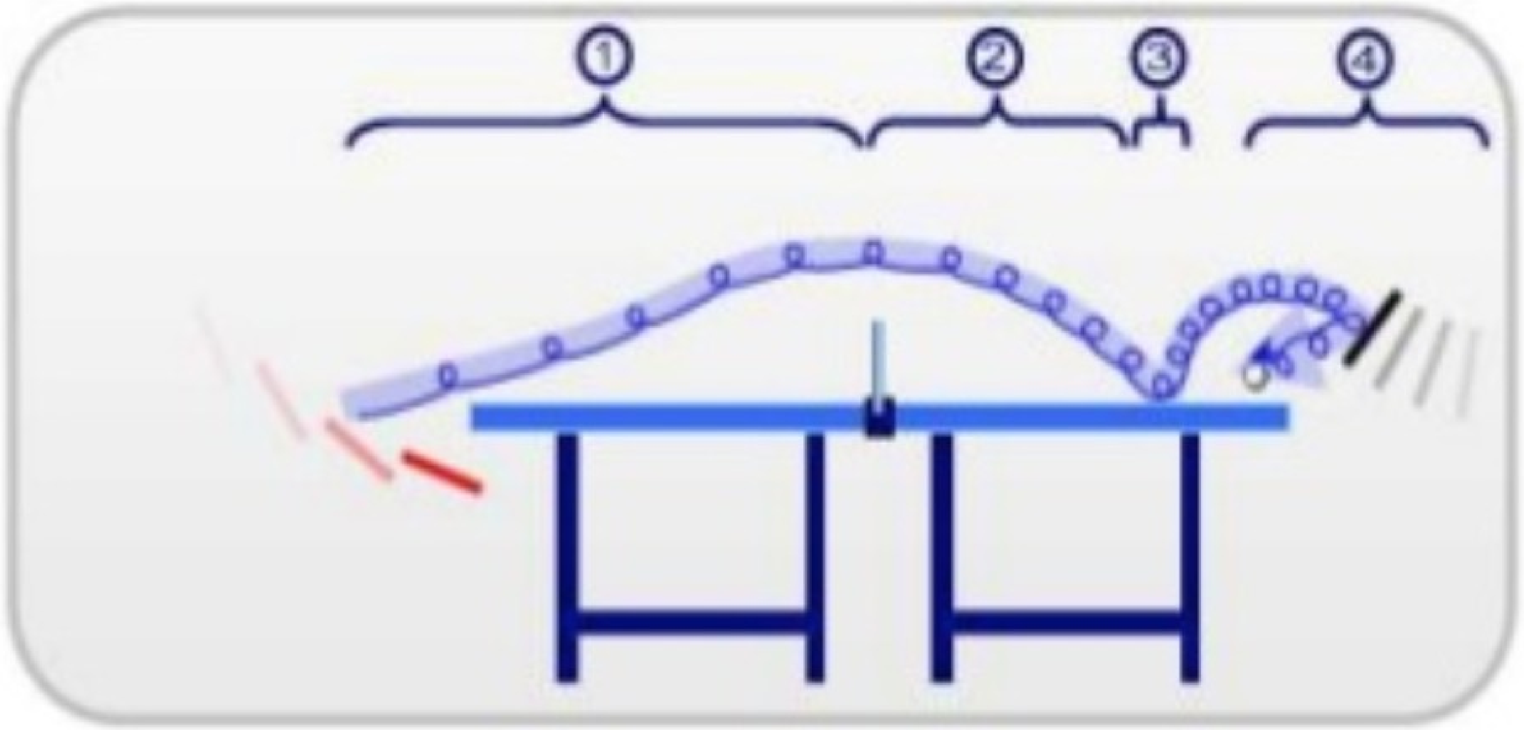
- **Poker**: deciding whether to raise or fold: need to decide correctly more than your opponent.
- **Medical Diagnosis:** "In my experience, the majority of patients who present symptoms X suffer from condition Y"
- **Sport**: read the body language of your opponent, don't let your opponent read your body language.

Prediction in sports

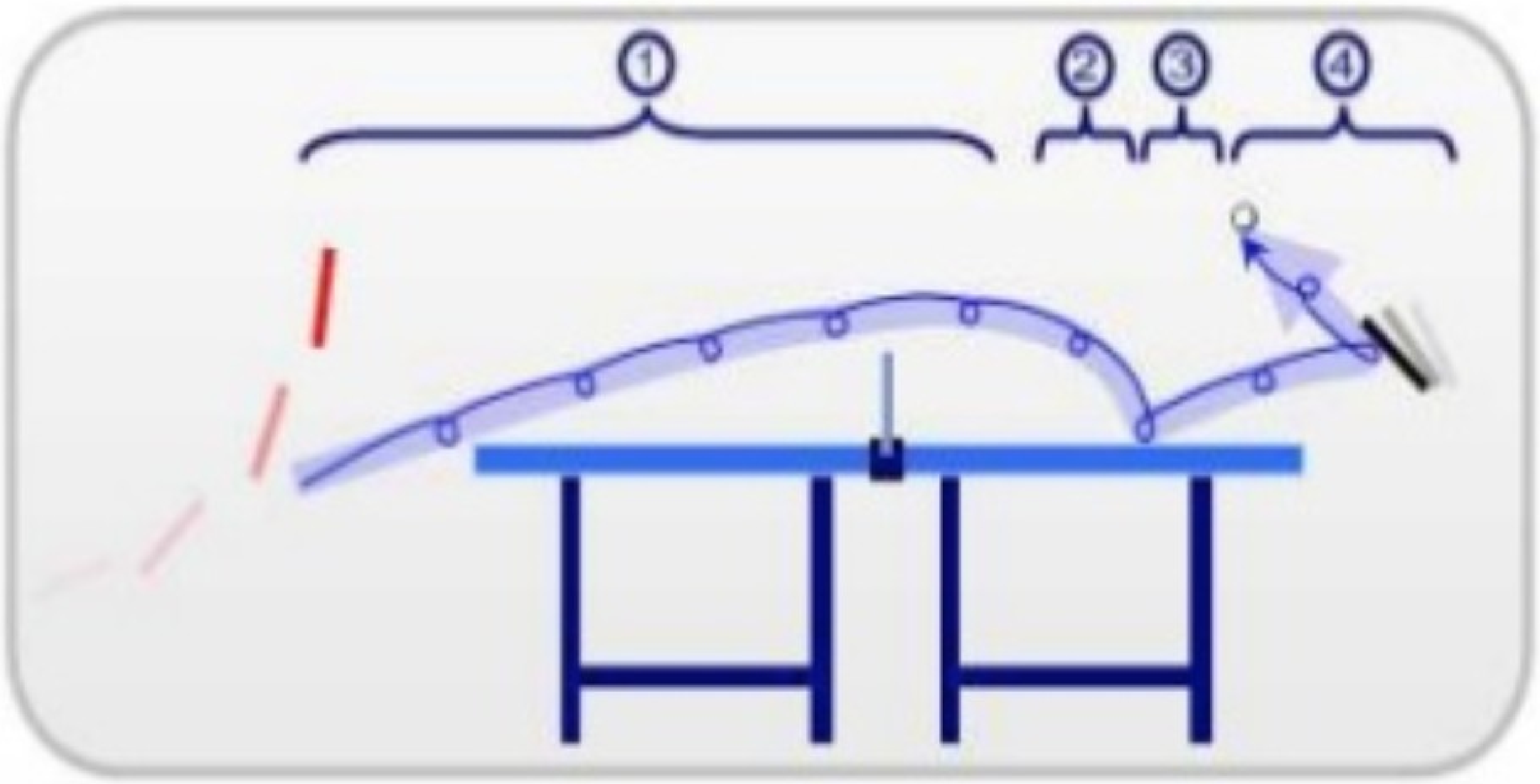
Prediction of spin in table tennis

- To respond correctly, we need to predict whether the ball has backspin or topspin.
- By the time the ball is hit, it is too late.
- We predict from body posture, from experience with individual player,…
- Definite action according to prediction.
- Prediction has to be correct more often than incorrect

Backspin



Topspin



Optimization
vs
Elimination

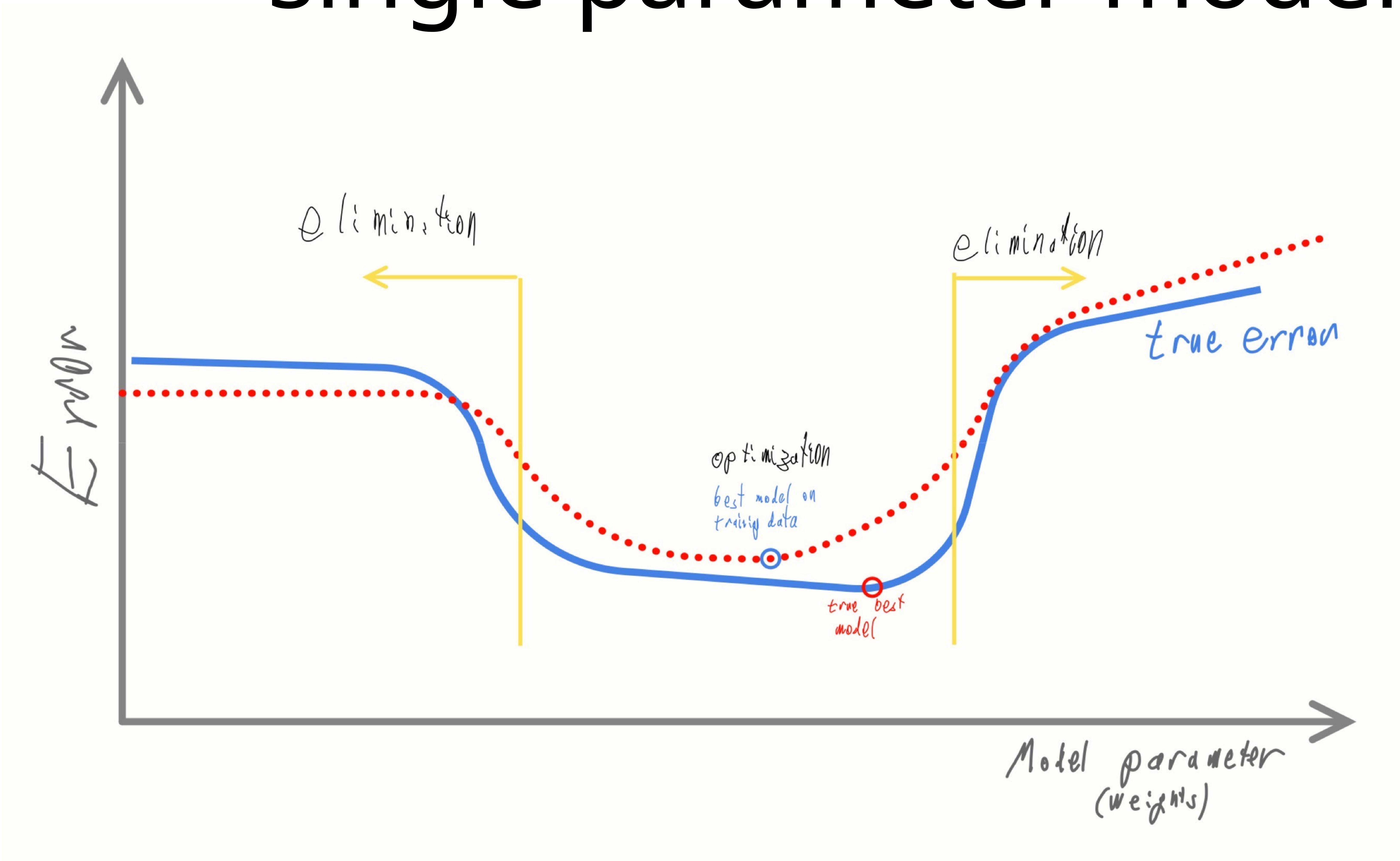
Two approaches to learning

- **Optimization:** find the model with the smallest loss on the training data.
- **Elimination:** use the training data to eliminate the models whose error

When you have eliminated all
which is impossible,
then whatever remains,
however improbable, must be
the truth.

—Arthur Conan Doyle

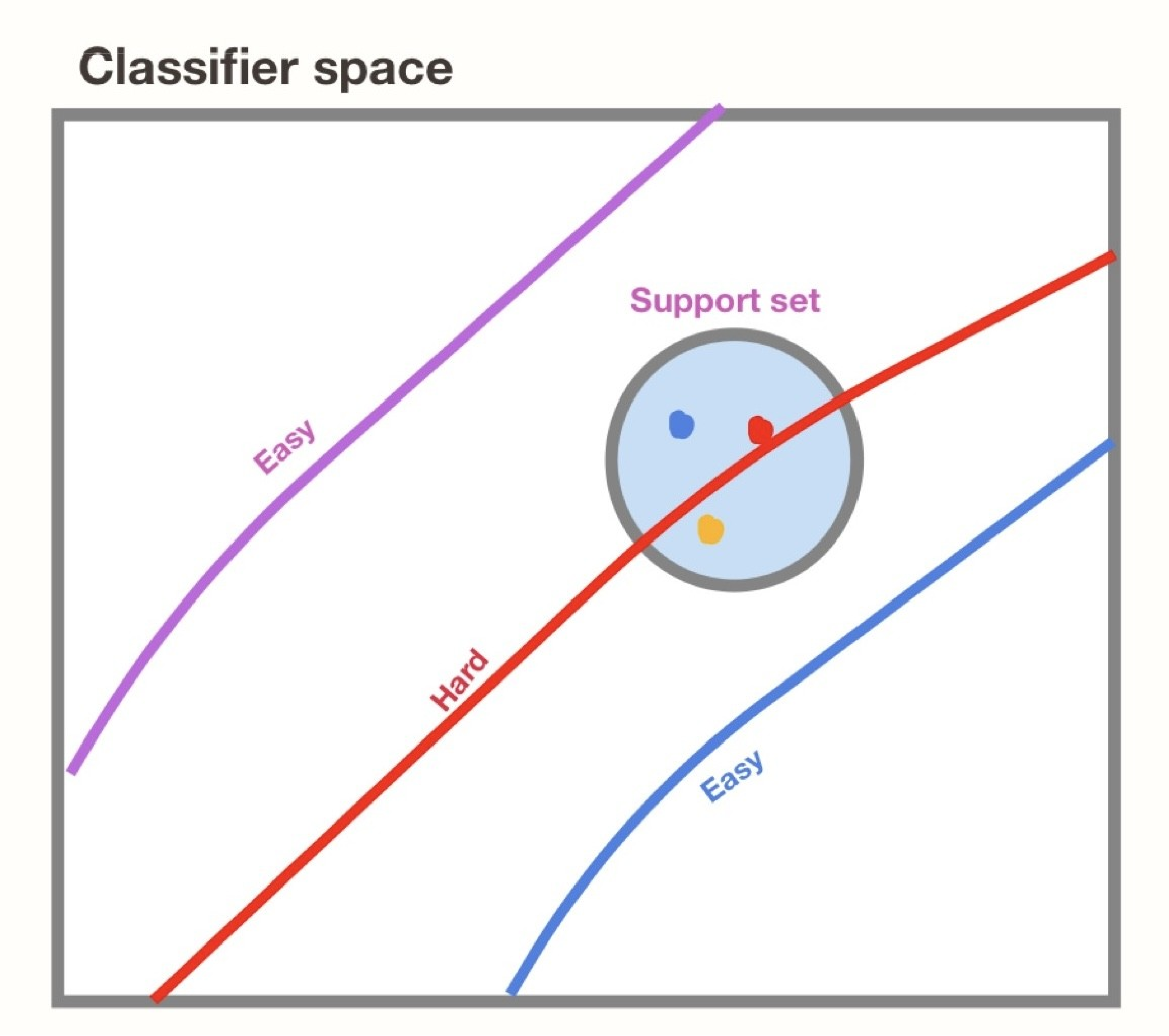
Elimination vs. Optimization
single parameter model

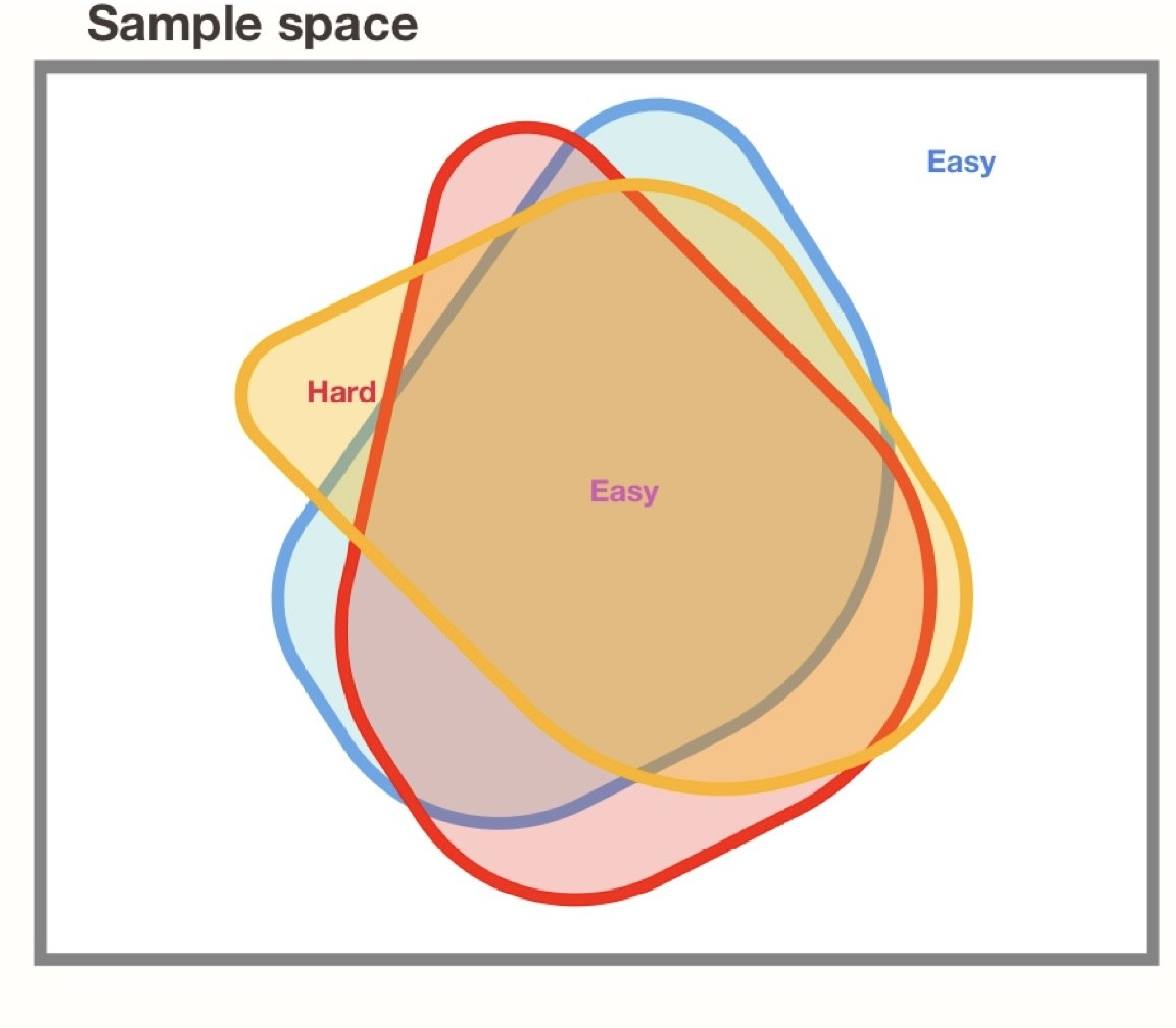


Properties of the elimination methods

- Instead of finding a single best model. You find a **Support SET** of acceptable models.
- When asked to predict, you check the predictions of all models in the support set.
- If there is a clear consensus – you predict **with confidence**
- If there is no consensus – you output “I don’t know” or “low confidence”

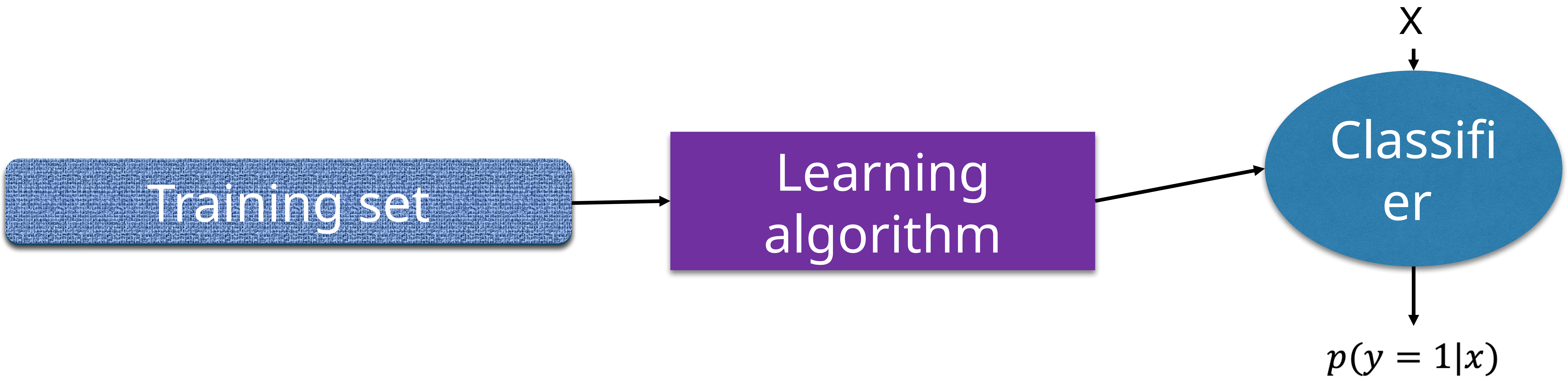
Elimination: Classifier space and example space



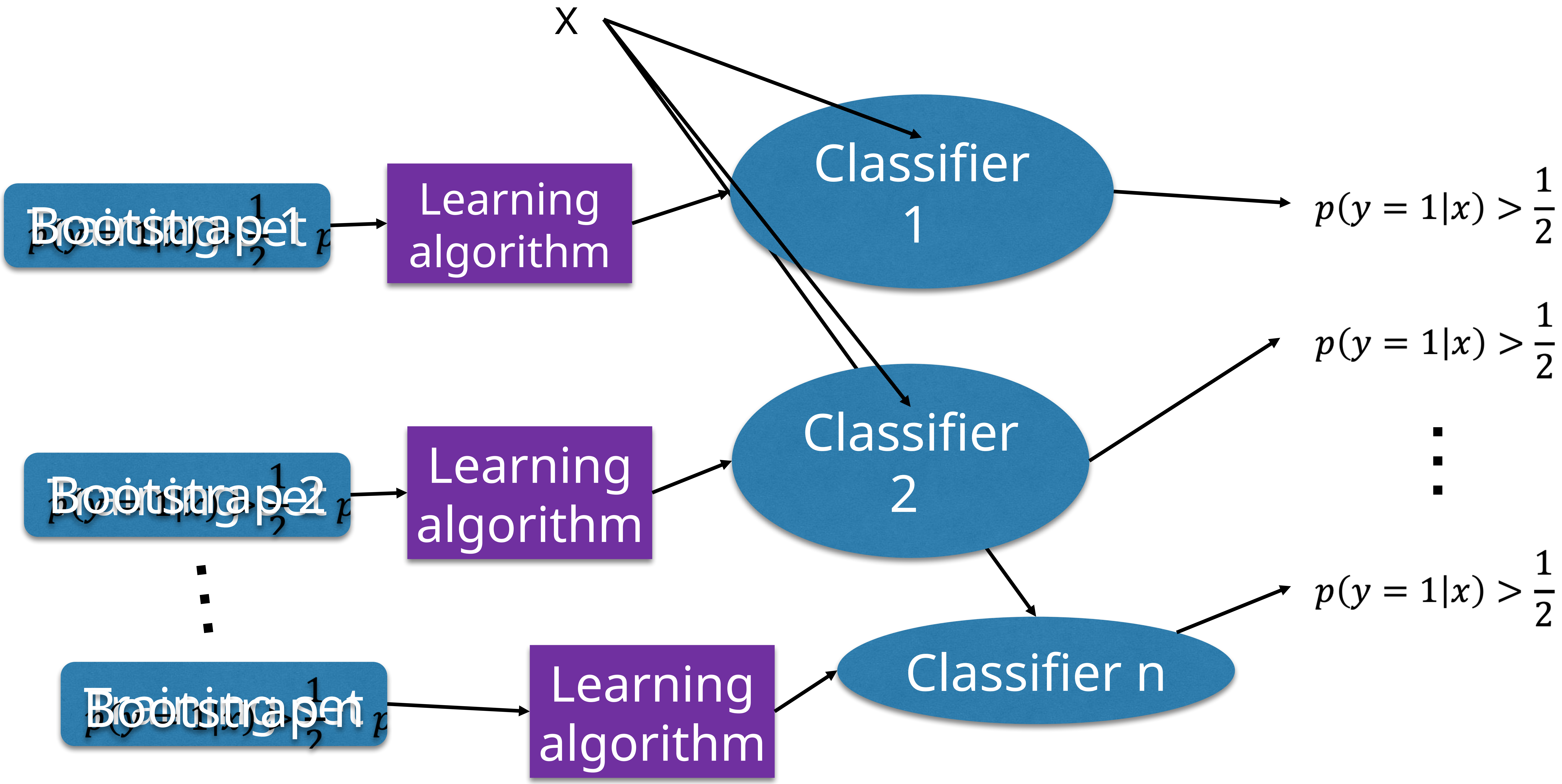


Confidence in classification

- High Certainty: $P(y|x) < \epsilon \; or \; p(y|x) > 1 - \epsilon$
- High Confidence: Switching from $p(y|x) > 1/2$ to $p(y|x) < 1/2$ (or vice versa) would require large (and unlikely) changes to the training set.



Bootstrap Ensemble



Degrees Of Confidence

Votes

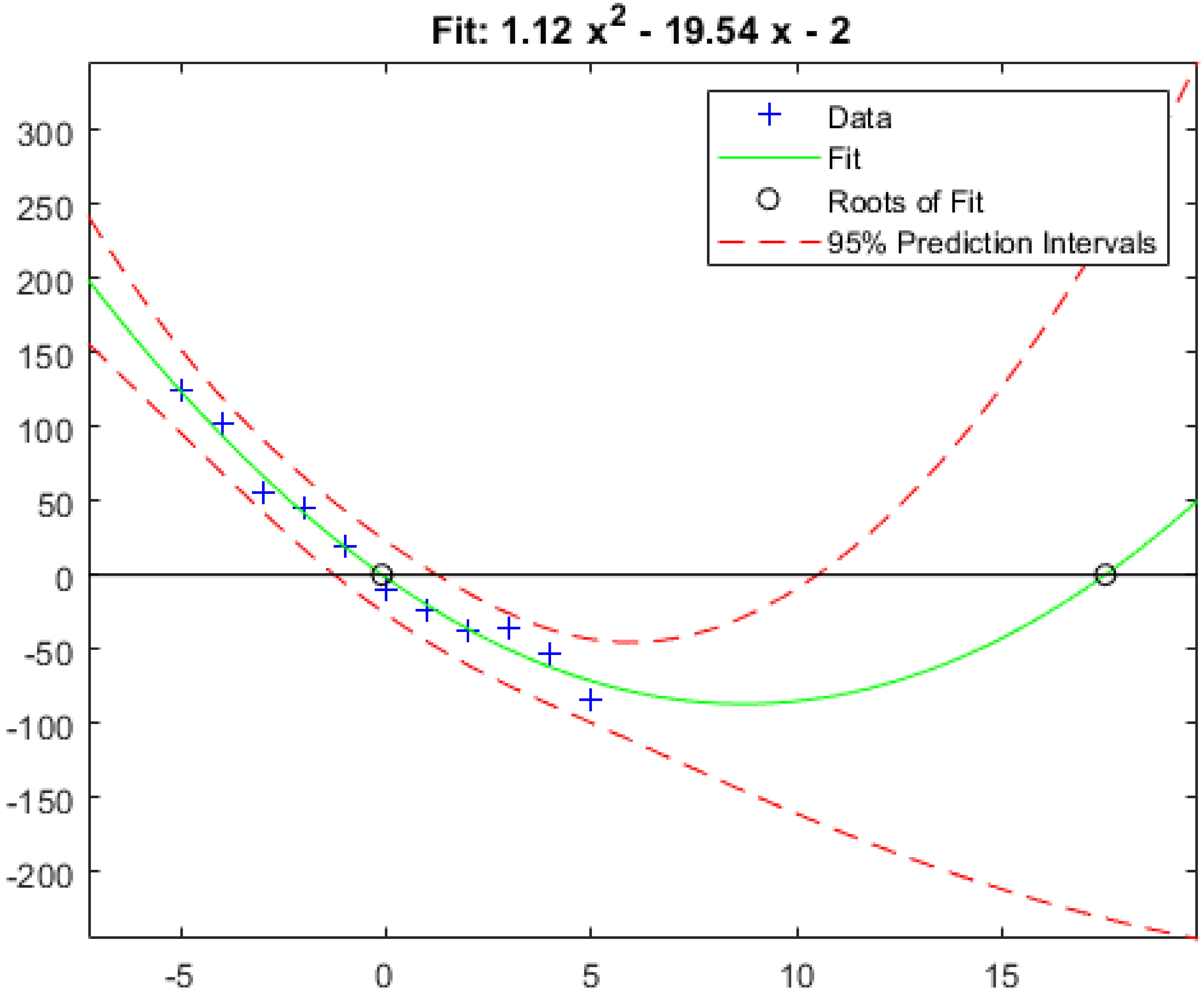
No of +1	No of -1
n	0
…	…
…	….
…	….
n/2	n/2
…	….
…	….
…	….
0	n

Confident positive | Easy positive

Not Confident $\frac{n}{2} \pm \sqrt{n}$ | **Hard / I don't know**

Confident negative | Easy negative

Confidence Intervals



experiments

Multi-label classification

- Example CIFAR100: 100 classes
- Most common measure of performance: fraction of time top score is correct
- Common Variant: fraction of time correct is among the top 5 scores.
- Alternative: the algorithm outputs a set of labels A we consider two measures of performance:
  - Fraction of time the correct label is in the set.
  - Distribution of the size of the set.
- A refinement of the “I don’t know” in the binary case

Easy and hard examples in the multilabel case

- The prediction on easiest examples is a set of size one.
- Larger prediction sets indicates harder examples.
- Example: distinguishing between dogs and cats is easier than distinguishing between subspecies.
- A prediction set that contains all labels == IDK

Does IDK matter?

- Not if
  - correct prediction = gain of $1,
  - Incorrect prediction = loss of $1
- Yes if
  - IDK = no gain or loss.
  - Correct prediction = gain of $1
  - Incorrect prediction = loss of $10