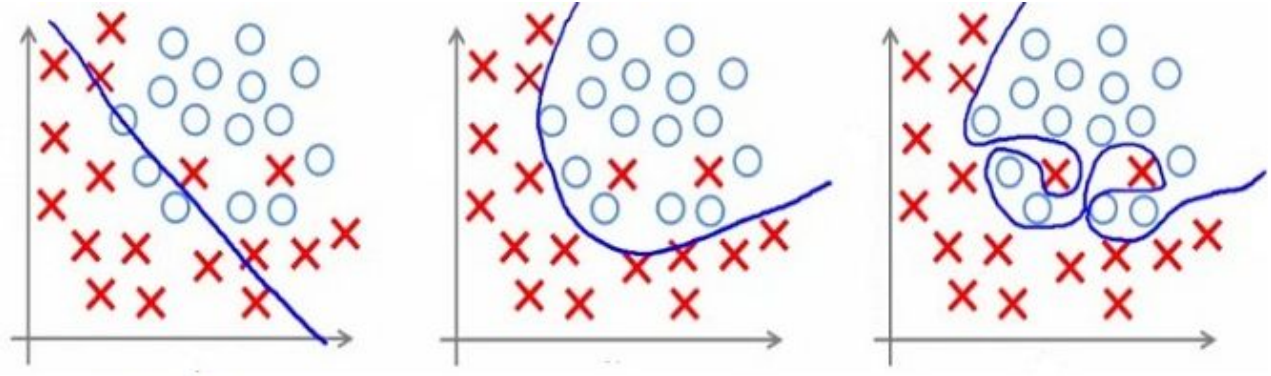
# Generalization Guarantees

Nupur Chatterji, Kenny Marino, Colin White

# Generalization

- The ability to *generalize* beyond the training set is the *essence* of machine learning
- Assume the training and test data both come from the same fixed distribution

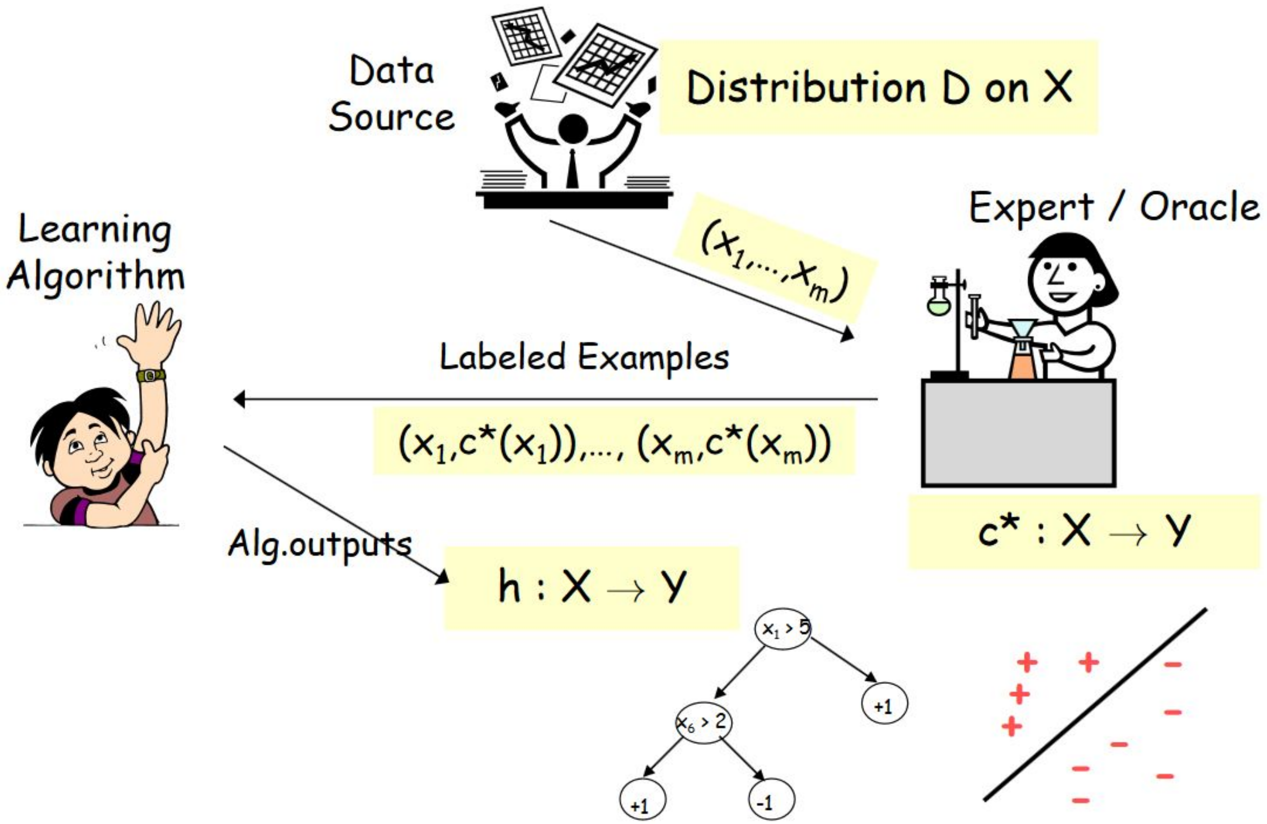


Under-fitting

Appropriate-fitting

Over-fitting

# PAC/SLT models for Supervised Learning



Data Source

Distribution D on X

$(x_1,\dots,x_m)$

Expert / Oracle

Learning Algorithm

Labeled Examples

$(x_1,c^*(x_1)),\dots,(x_m,c^*(x_m))$

$c^* : X \to Y$

Alg.outputs

$h : X \to Y$

# PAC/SLT models for Supervised Learning

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, $x_i$ i.i.d. from D
- Does optimization over S, find hypothesis $h \in H$.
- Goal: h has small error over D.

True error: $\mathrm{err}_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$

How often $h(x) \neq c^*(x)$ over future instances drawn at random from D

- But, can only measure:

Training error: $\mathrm{err}_S(h) = \frac{1}{m}\sum_i I\big(h(x_i) \neq c^*(x_i)\big)$

How often $h(x) \neq c^*(x)$ over training instances

**Sample complexity: bound $err_D(h)$ in terms of $err_S(h)$**

# Sample Complexity: Realizable Case

- First, assume there exists *h* in *H* consistent with the sample

**Theorem:** $m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$

labeled examples are sufficient s.t. with prob >1-δ, all $h \in H$

with $err_D(h) \geq \varepsilon$ have $err_S(h) \geq 0$

- What does this tell us?

# Sample Complexity: Agnostic Case

- What if there is no *h* in *H* consistent with the sample?

**Theorem:** $$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

labeled examples are sufficient s.t. with prob >1-δ, all *h*∈*H*

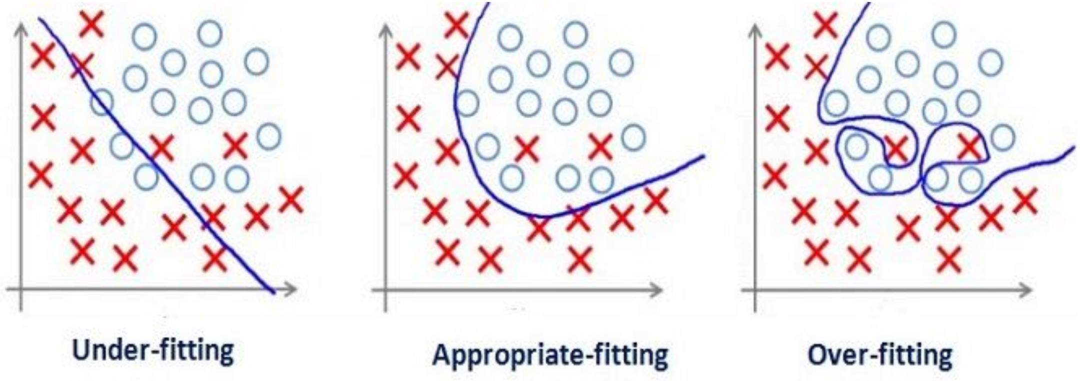
satisfy $|err_D(h) - err_S(h)| < \varepsilon$

- What does this tell us?

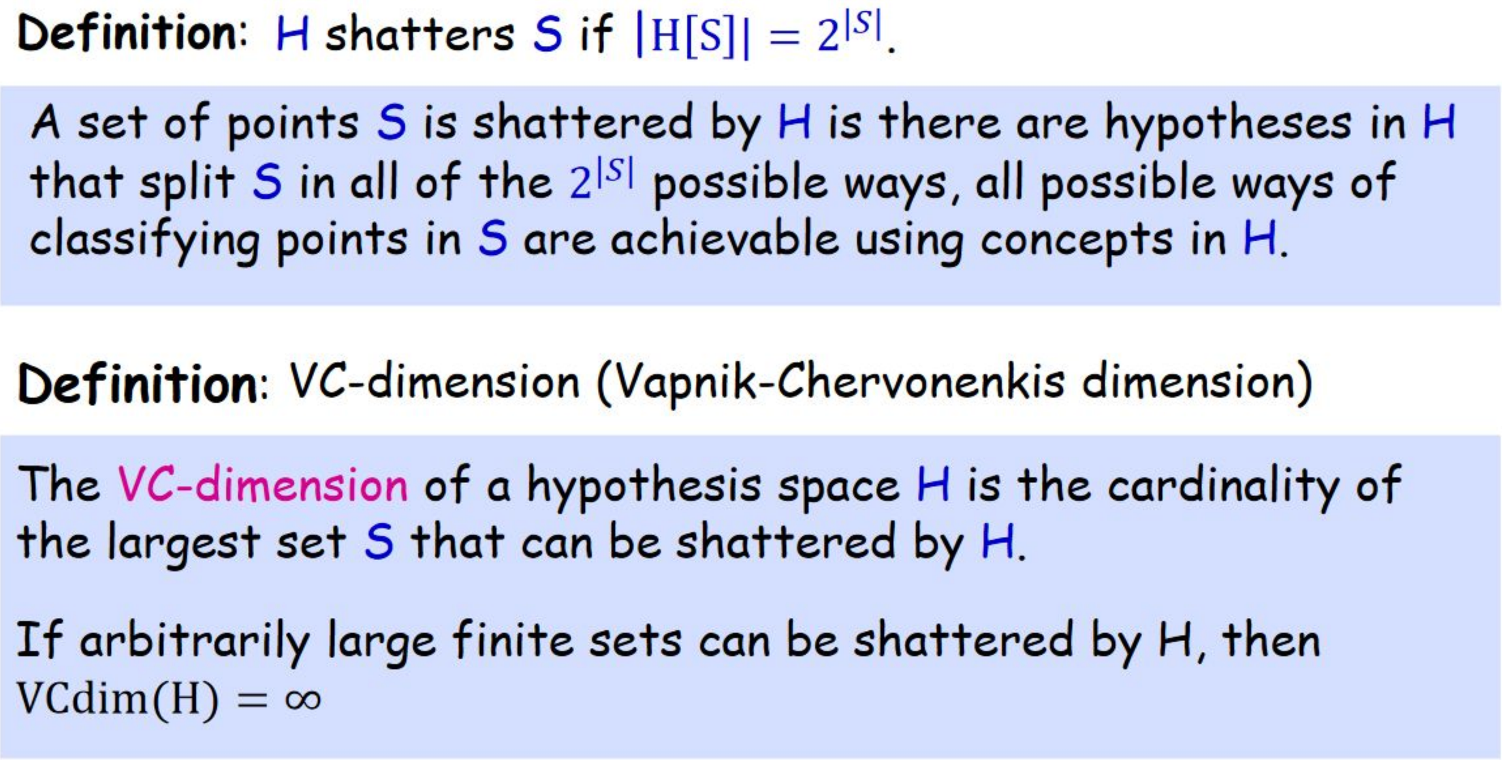
# But what if our hypothesis class is infinite??

- Many hypothesis classes we’ve seen are infinite.
- Linear separators
- Thresholds on a real line

VC Dimension! Measures the *complexity* of the hypothesis class



# VC Dimension



**Definition:** H shatters S if $|\mathrm{H}[\mathrm{S}]| = 2^{|S|}$.

A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways, all possible ways of classifying points in S are achievable using concepts in H.

**Definition:** VC-dimension (Vapnik-Chervonenkis dimension)

The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(\mathrm{H}) = \infty$

# VC Dimension Bounds

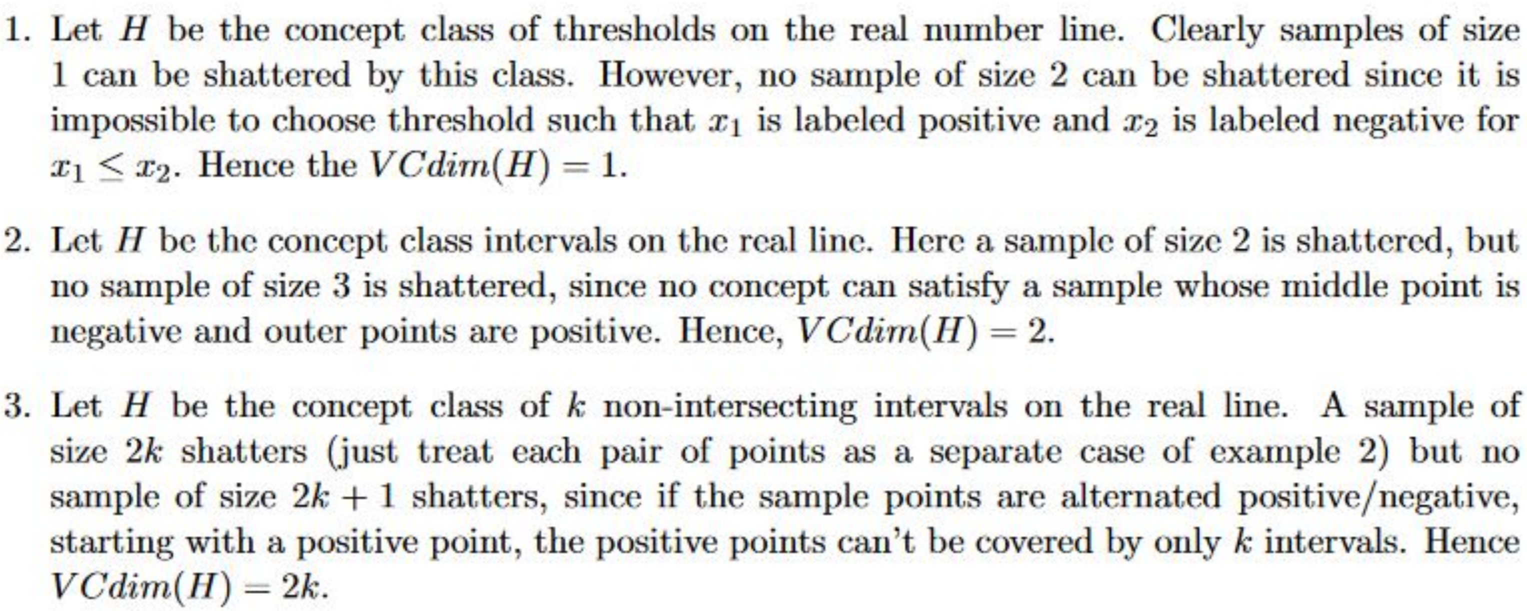
**Theorem:** $$m = \mathrm{O}\left(\frac{1}{\varepsilon}\left[VCdim(H)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient s.t. with prob >1-δ, all *$h \in H$*

with *$err_D(h) \geq \varepsilon$* have *$err_S(h) \geq 0$*

- Examples??

# Examples of VC dimension

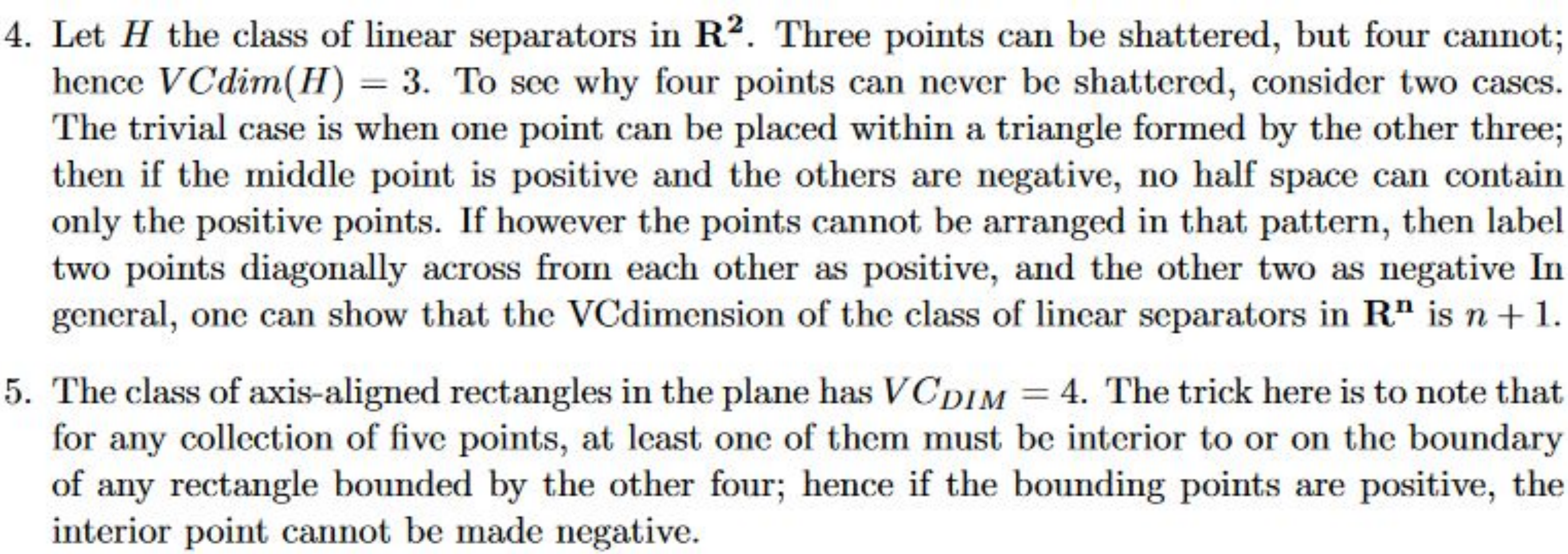


1. Let $H$ be the concept class of thresholds on the real number line. Clearly samples of size 1 can be shattered by this class. However, no sample of size 2 can be shattered since it is impossible to choose threshold such that $x_1$ is labeled positive and $x_2$ is labeled negative for $x_1 \leq x_2$. Hence the $VCdim(H) = 1$.

2. Let $H$ be the concept class intervals on the real line. Here a sample of size 2 is shattered, but no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and outer points are positive. Hence, $VCdim(H) = 2$.

3. Let $H$ be the concept class of $k$ non-intersecting intervals on the real line. A sample of size $2k$ shatters (just treat each pair of points as a separate case of example 2) but no sample of size $2k + 1$ shatters, since if the sample points are alternated positive/negative, starting with a positive point, the positive points can't be covered by only $k$ intervals. Hence $VCdim(H) = 2k$.

# Examples of VC dimension



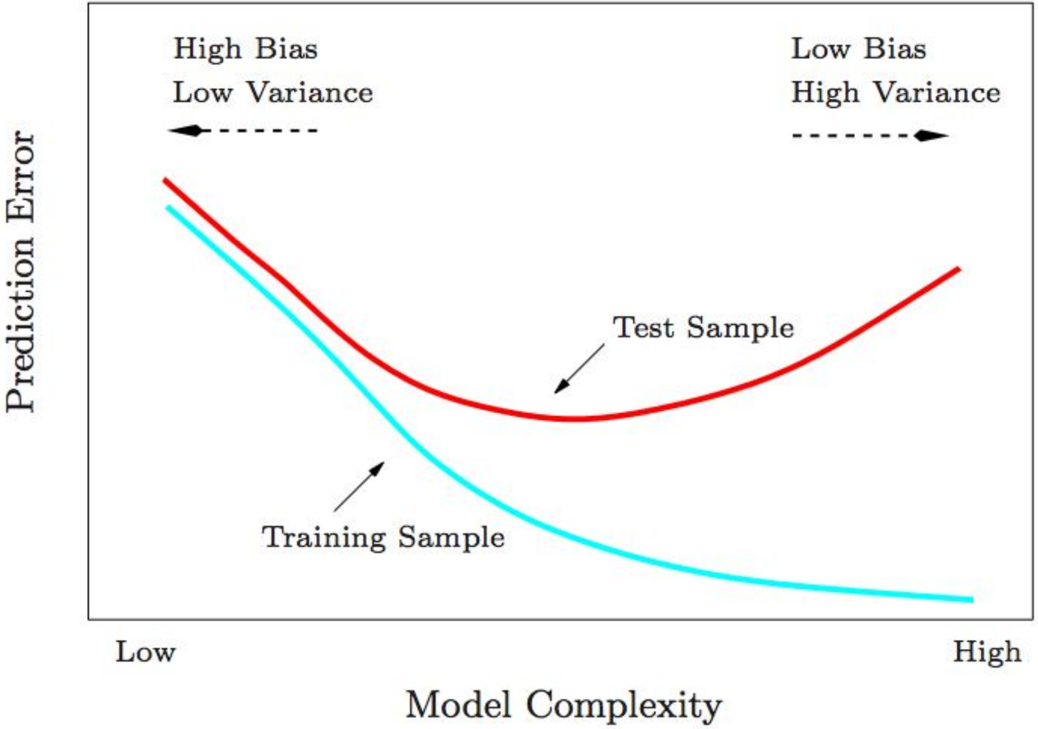
4. Let $H$ the class of linear separators in $\mathbf{R^2}$. Three points can be shattered, but four cannot; hence $VCdim(H) = 3$. To see why four points can never be shattered, consider two cases. The trivial case is when one point can be placed within a triangle formed by the other three; then if the middle point is positive and the others are negative, no half space can contain only the positive points. If however the points cannot be arranged in that pattern, then label two points diagonally across from each other as positive, and the other two as negative In general, one can show that the VCdimension of the class of linear separators in $\mathbf{R^n}$ is $n+1$.

5. The class of axis-aligned rectangles in the plane has $VC_{DIM} = 4$. The trick here is to note that for any collection of five points, at least one of them must be interior to or on the boundary of any rectangle bounded by the other four; hence if the bounding points are positive, the interior point cannot be made negative.

# Examples of VC Dimension

What is the VC dimension of $f(x) = \sin(\alpha x)$, for all $\alpha$ ?

# Training Error vs Test Error


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity