Machine Learning

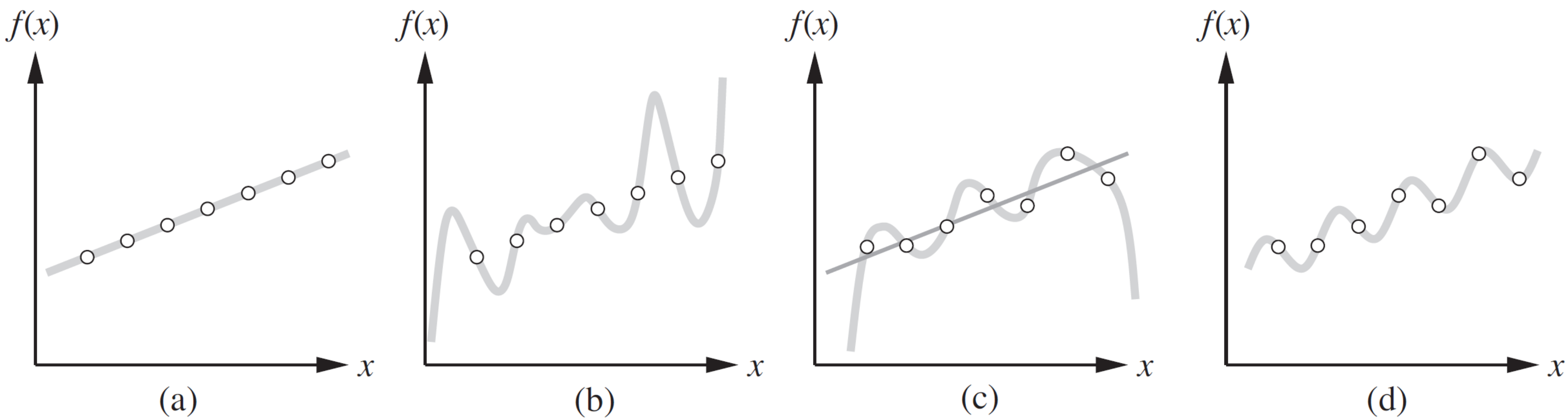
CS161

Prof. Guy Van den Broeck

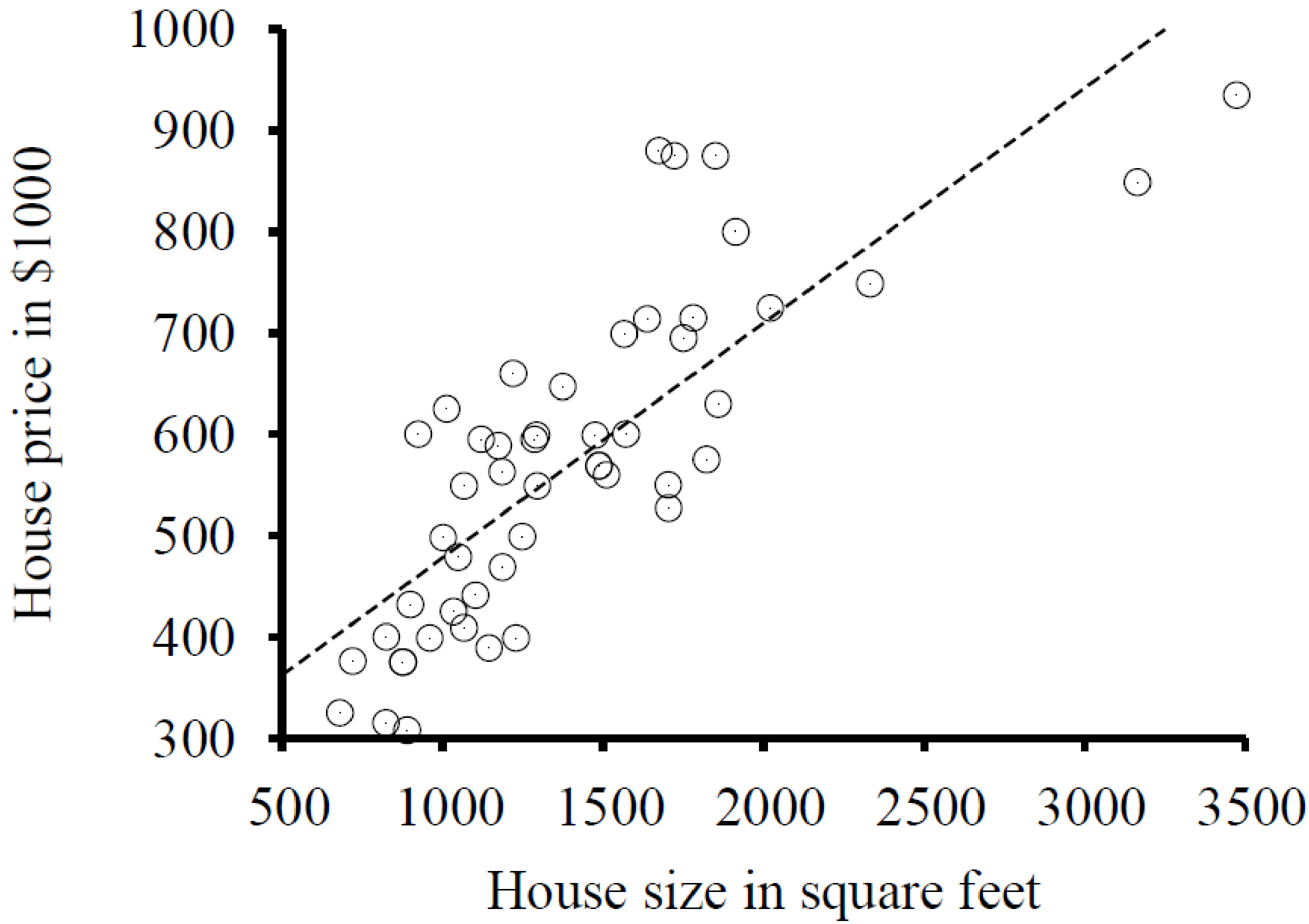
Data comes from Nature

Learning Settings

Fitting Data



Regression



1000
900
800
700
600
500
400
300
House price in $1000
500
1000
1500
2000
2500
3000
3500
House size in square feet

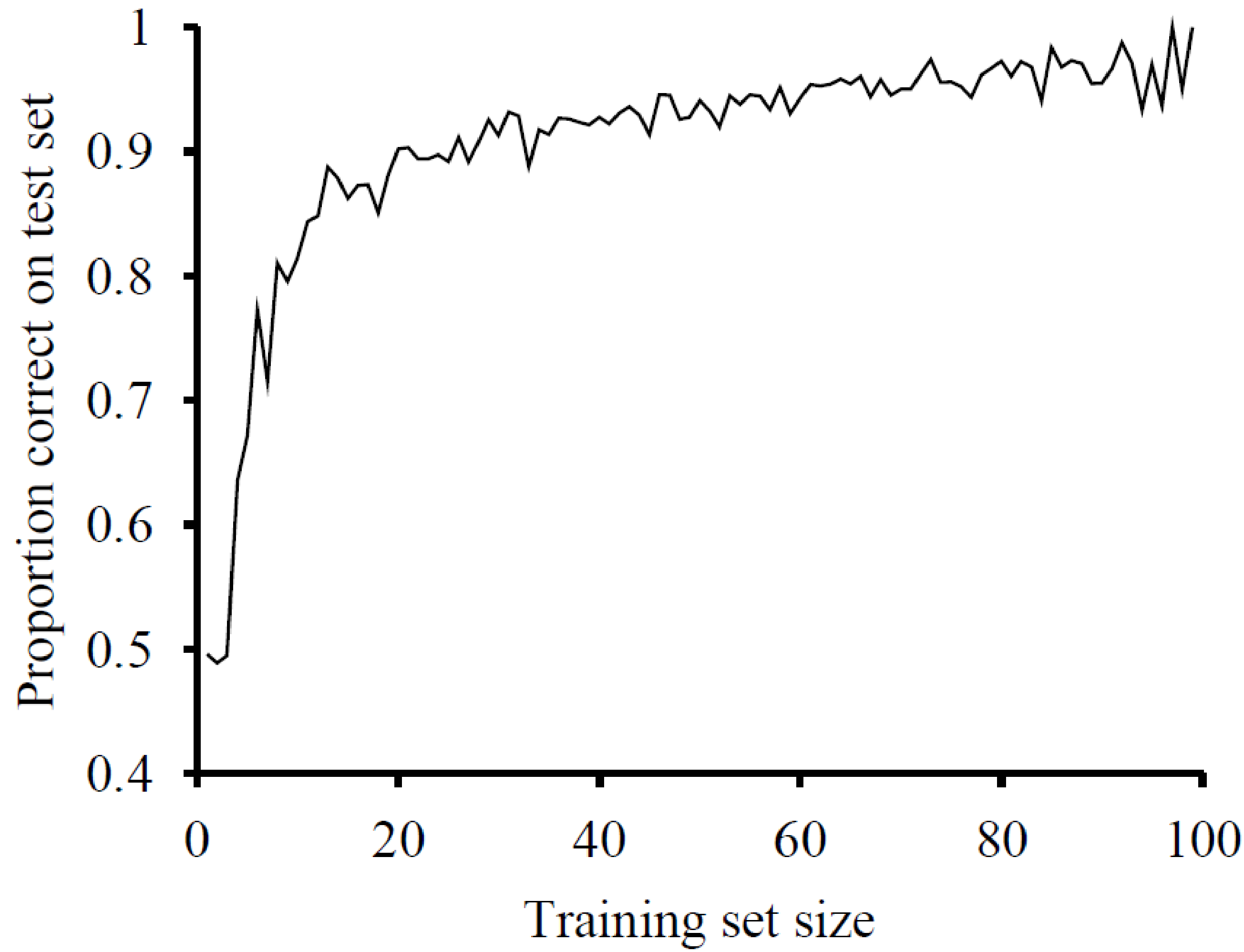
Classification Data

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

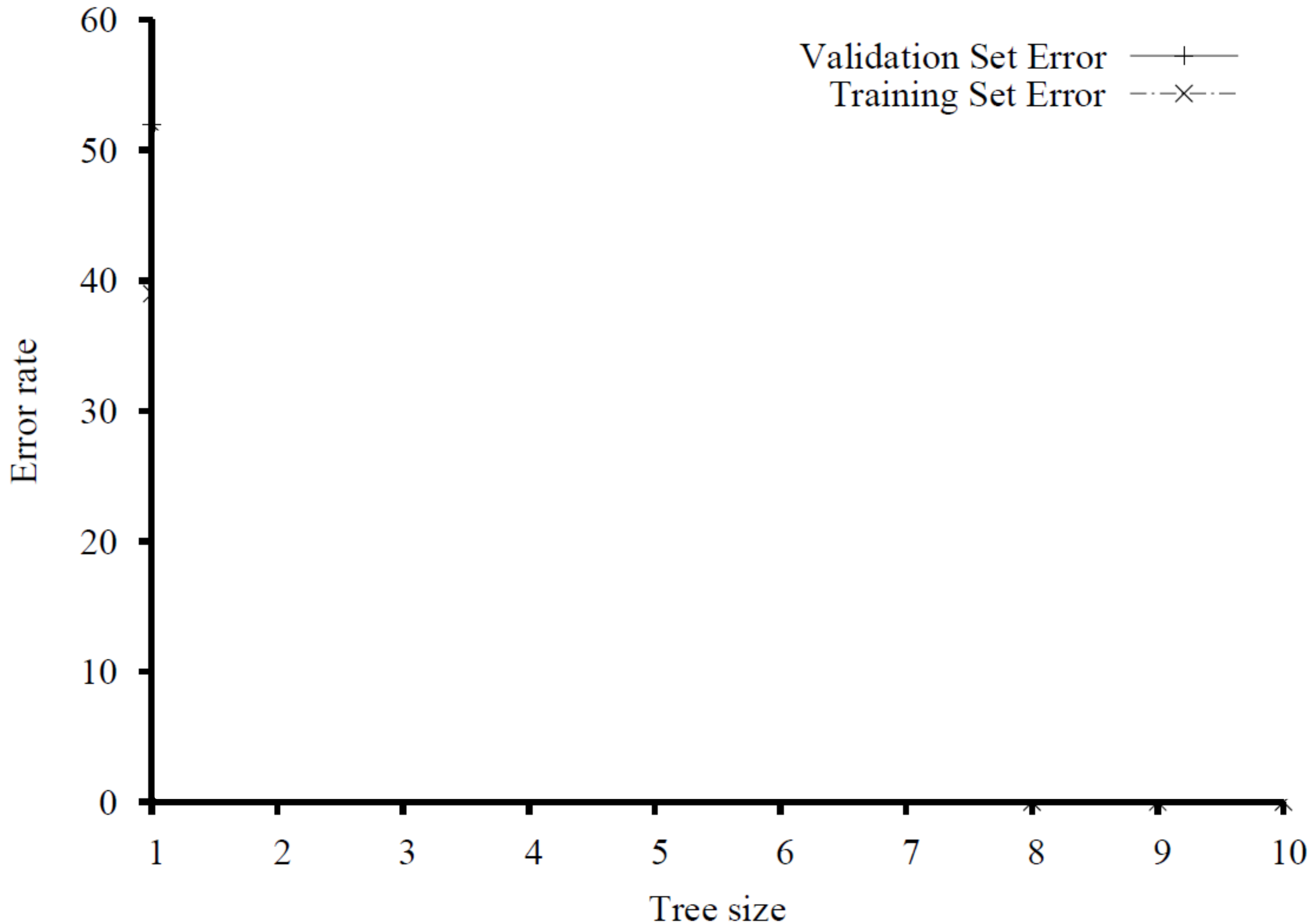
How to evaluate?

- Unsupervised learning of a Pr:
  - Likelihood: Pr(data)
- Supervised learning of binary classification:
  - Some combination of True Positive, True Negative, False Positive, False Negative
  - E.g., accuracy
- Many more possibilities

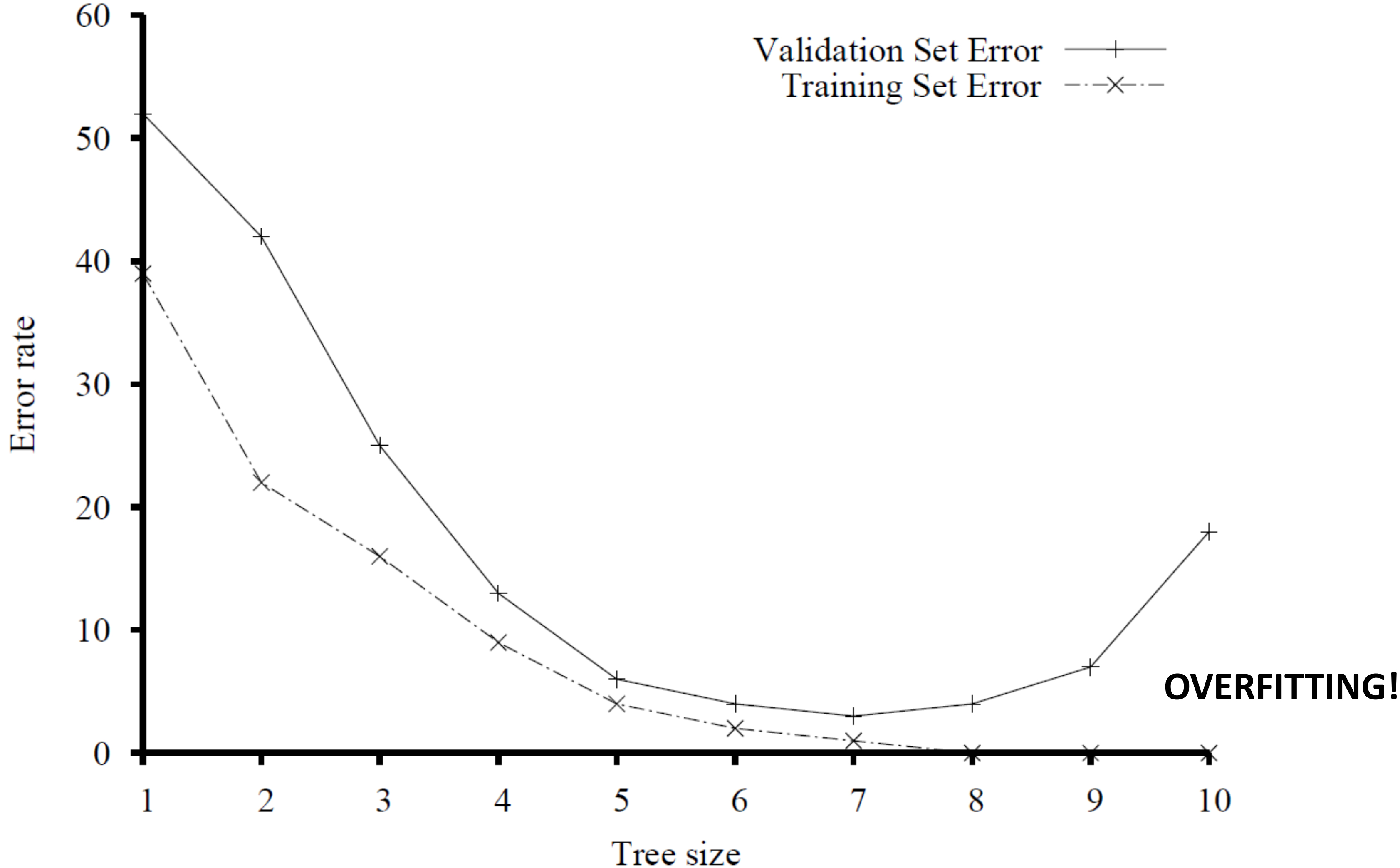
More data is better!



More model complexity is better?

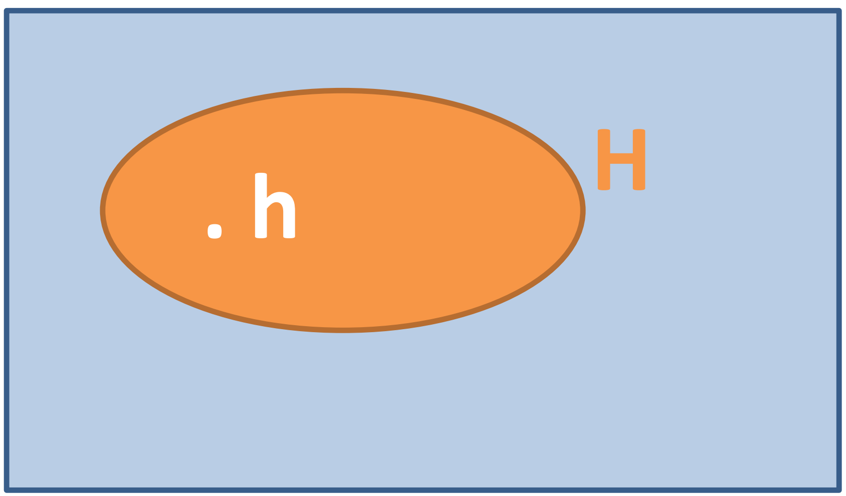


Model complexity is better?



Hypothesis Space H

All functions

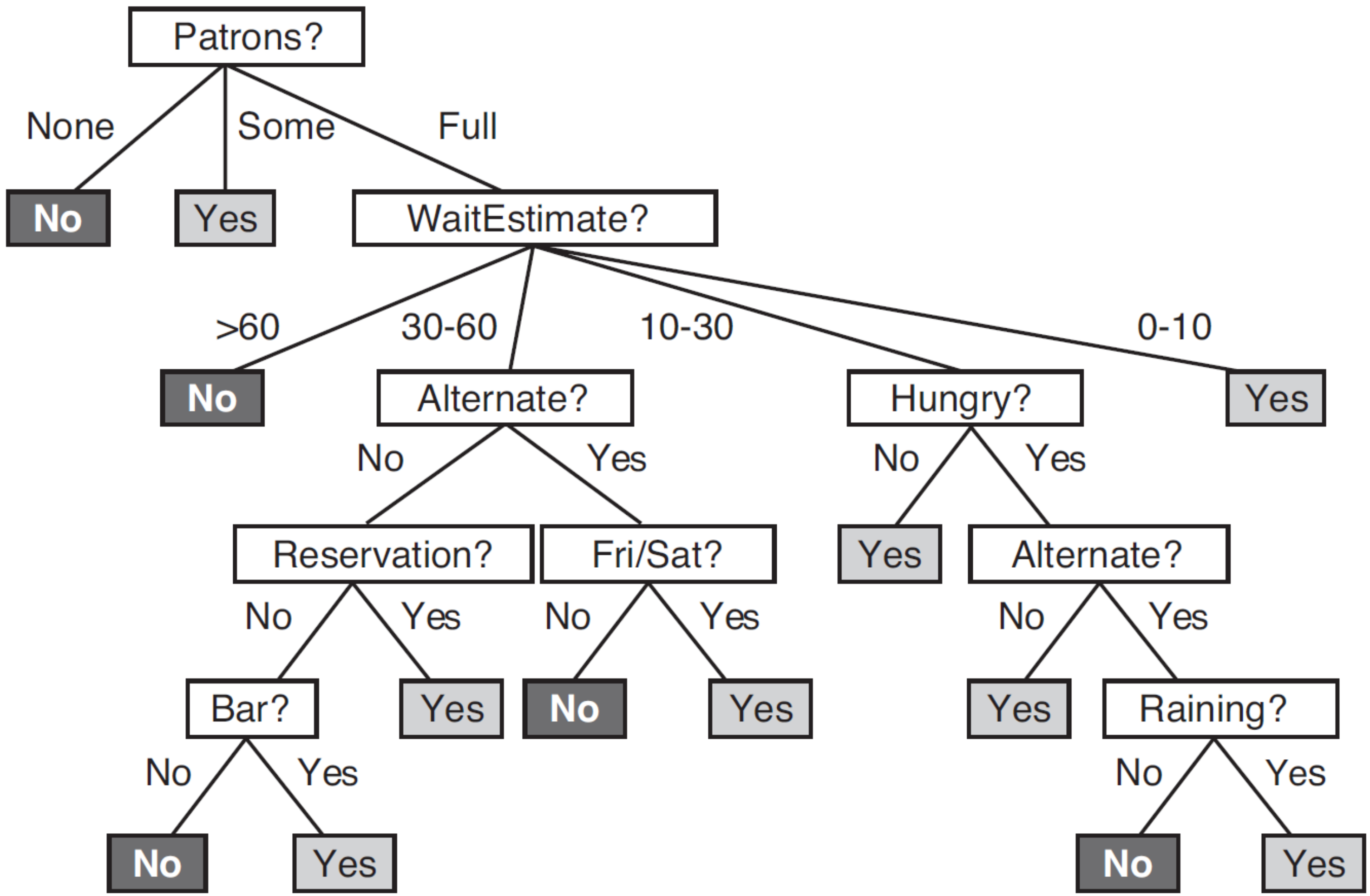


Given data about $f(x)$
Find $h(x) \approx f(x)$
Where $h \in$ **H**

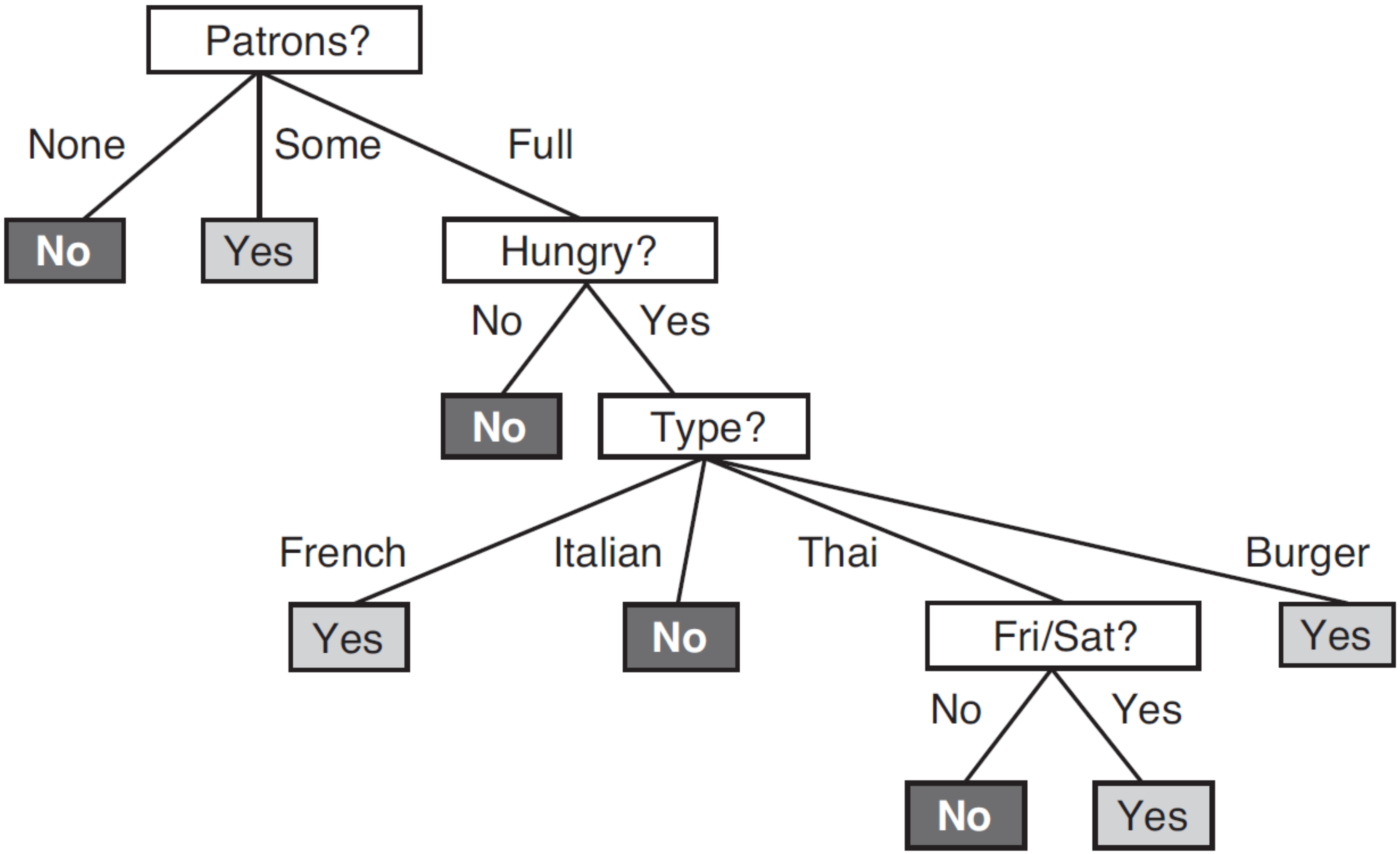
"Bias-variance tradeoff":

- Large |H|: difficult to find h, need a lot of data
- Small |H|: difficult to match true f, not enough options

The true function as a decision tree



Induced decision tree from data



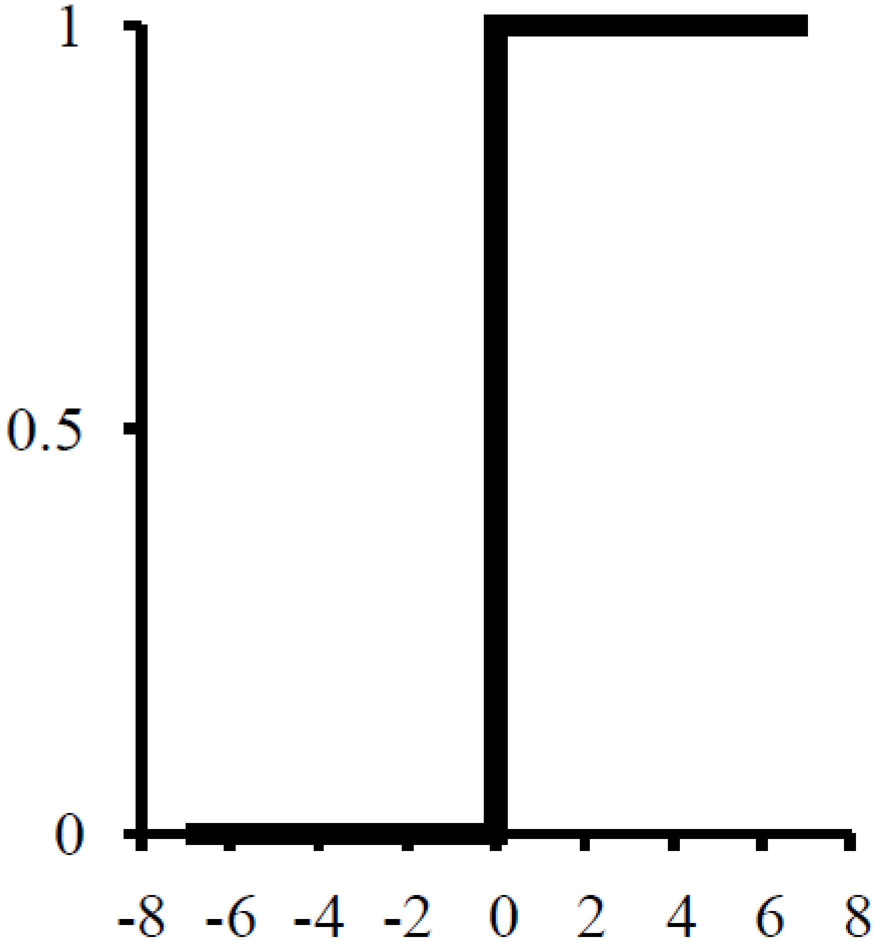
How to learn Bayesian networks?

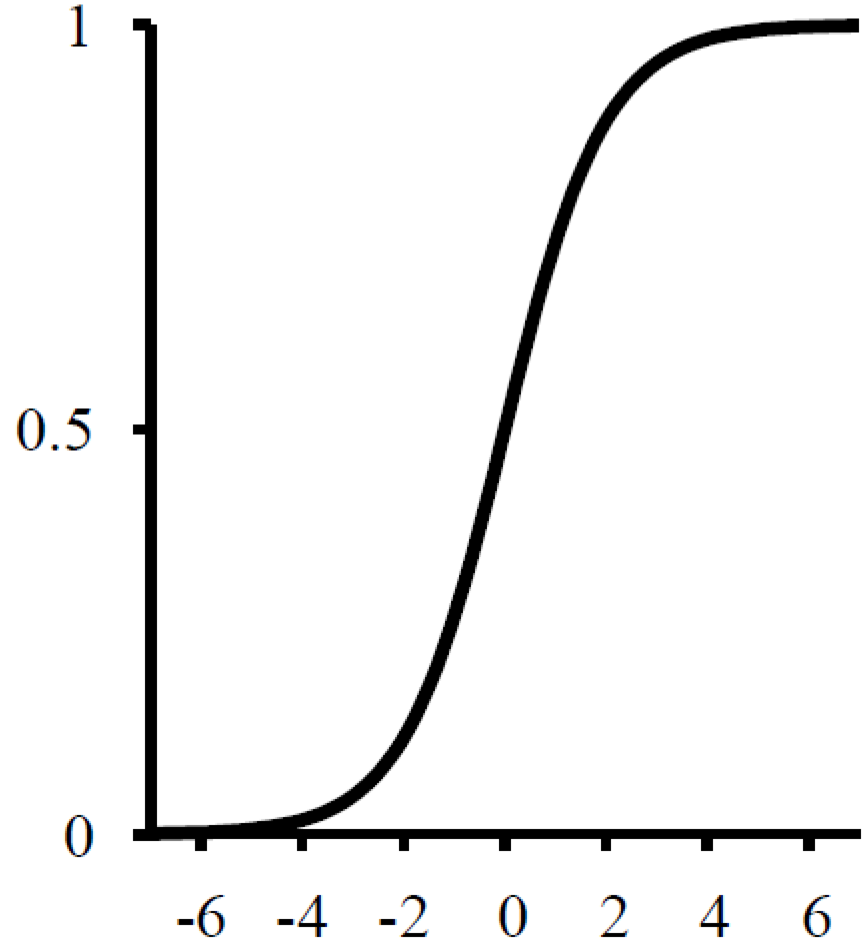
- For example: Naïve Bayes
- Parameters are conditional probability $P(x|y)$
- Estimate this probability:
  - Count how often y is true in the data
  - Count how often $x \wedge y$ is true in the data
  - Take the ratio as your estimate
- Overfitting is still a problem
  - Make parameter estimates "more conservative"

Linear Regression

- Consider a linear function
$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$
- Given data $\{(x_i, y_i)\}$, find w-vector
- Minimize loss function, for example
$$L(w) = \sum_i (h_w(x_i) - y_i)^2$$
- Overfitting is still a problem:
  - make weights prefer to be "close to 0."
  - A "regularizer"

From numbers to probabilities





Logistic Regression

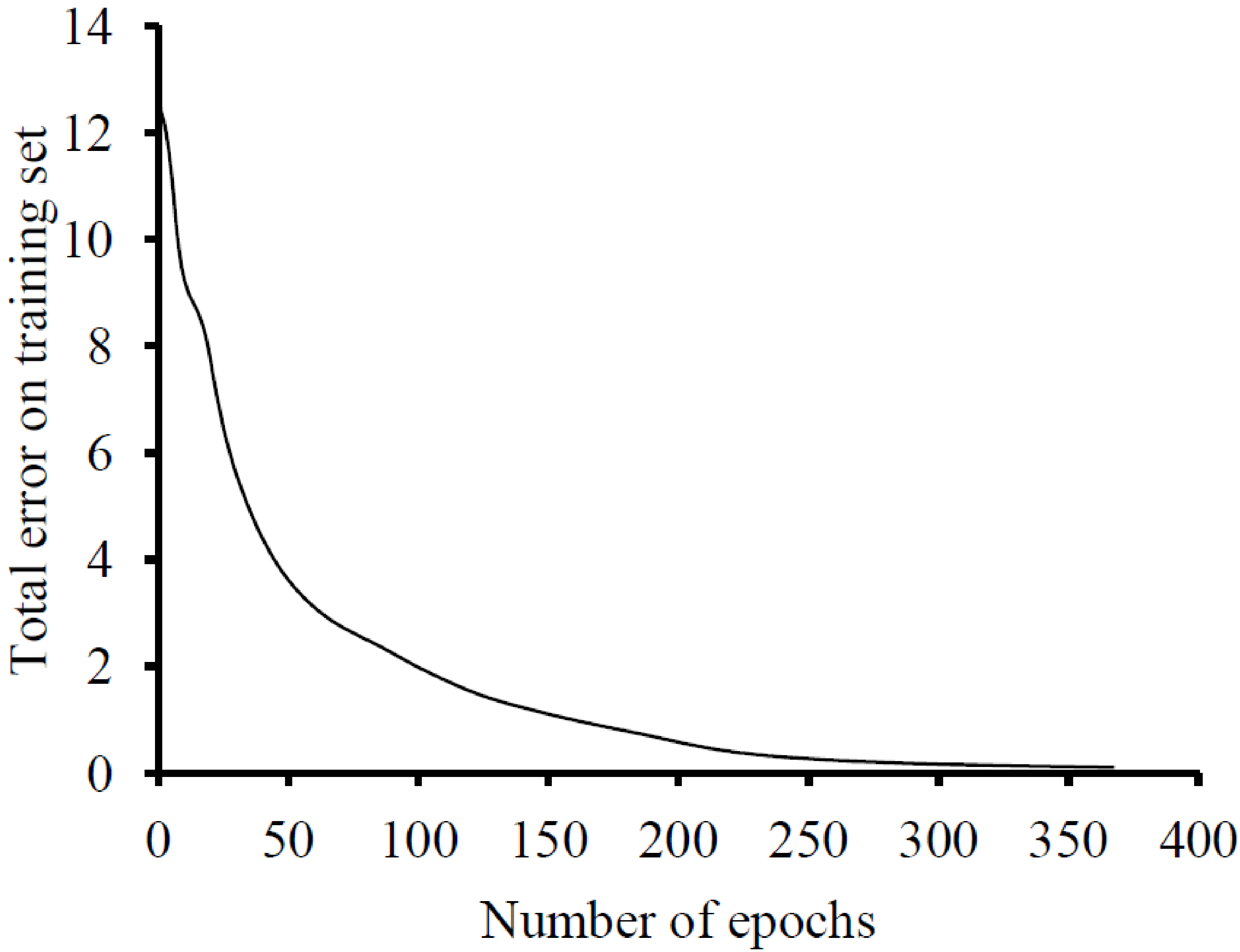
- Push linear prediction through sigmoid activation function:

$$g_w(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

$$h_w(x) = 1/(1 + \exp(\, g_w(x)))$$

- Real numbers become probabilities
- Now we have a classifier!
- Overfitting: make weights close to 0
- Training: by gradient descent on a loss function

Logistic Regression Training

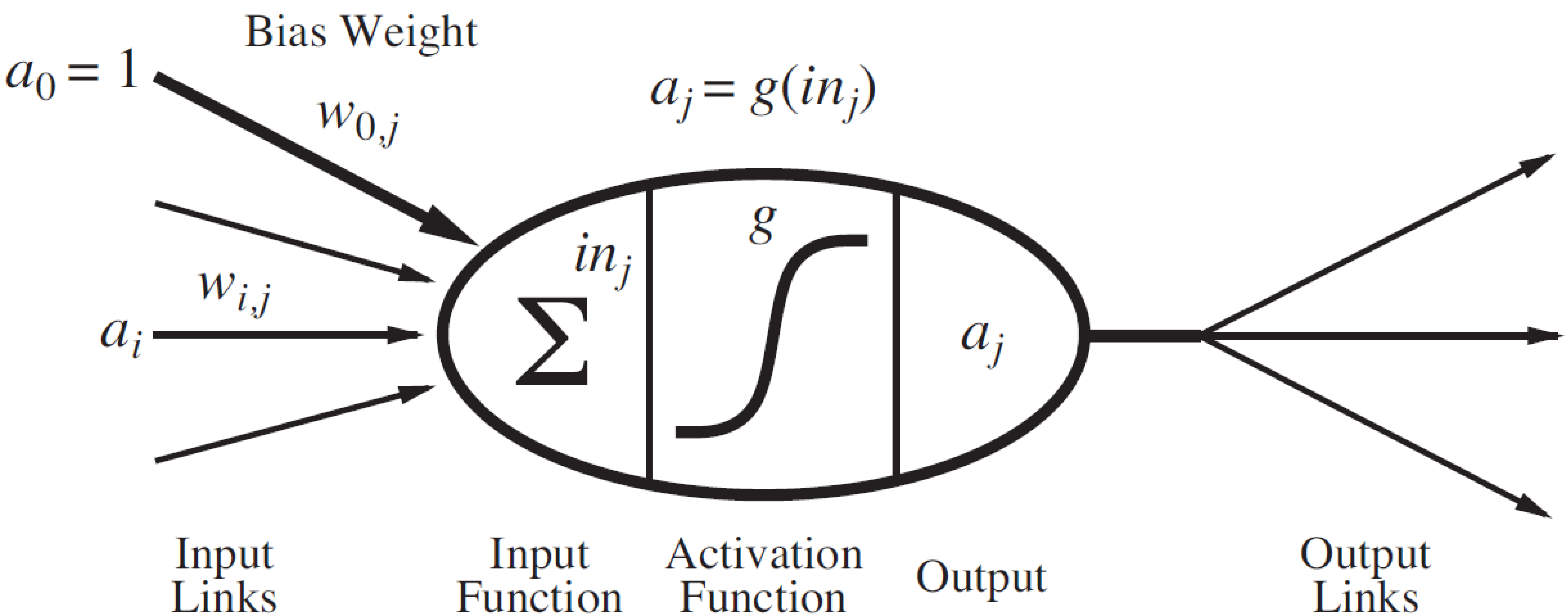


Logistic Regression vs Naïve Bayes

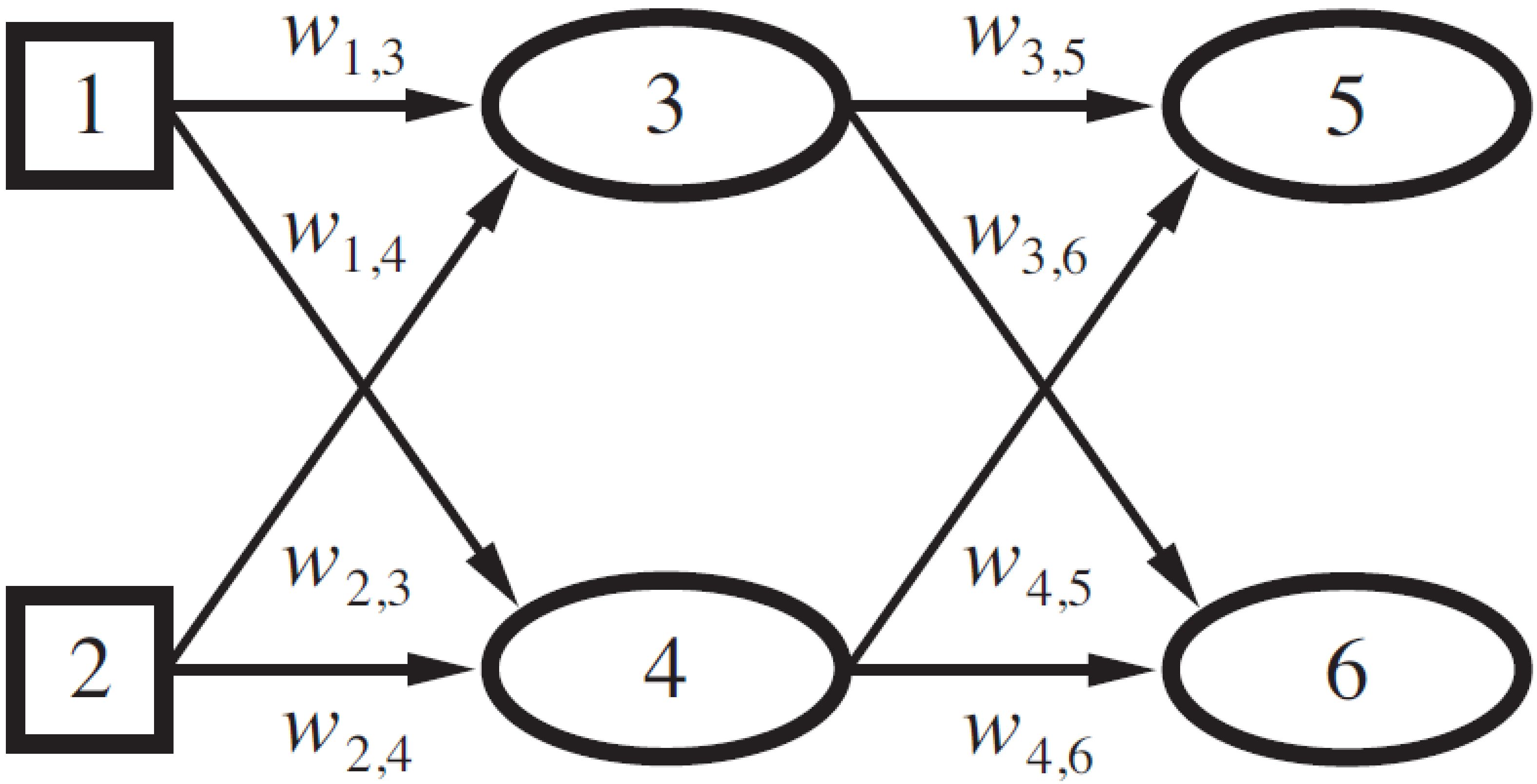
Example: MNIST Digit Classification

93% accuracy with logistic regression
99% accuracy with nested logistic regression: neural networks

Deep Neural Networks

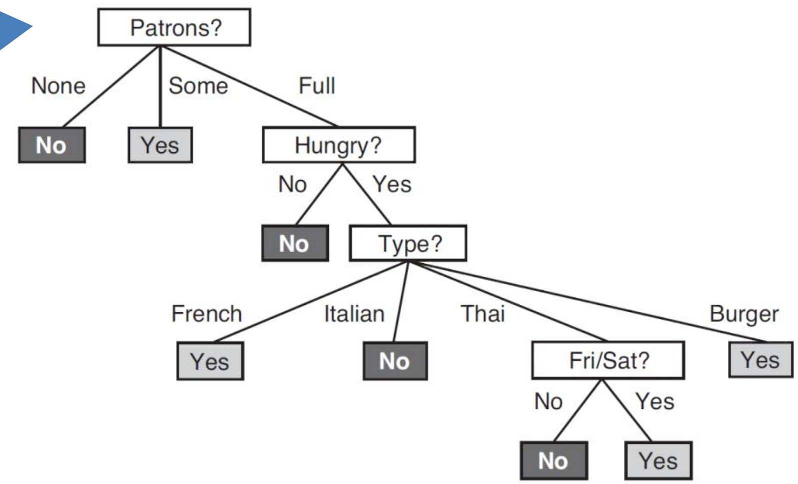


Deep Neural Networks



How to learn Decision Trees?

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$



What is the size of the hypothesis space?

- How many trees over n Boolean features?
- How many conjunctions?

We’ll do greedy search!

Which Splits?