CS60050 : Machine Learning (Spring 2024)

Auxiliary Materials for Class Teaching

Dr. Aritra Hazra

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

West Bengal, India - 721302

Examples (Concept Learning)

Training Examples

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport?
1	Sunny	Warm	Normal	Strong	Warm	Same	*Yes*
2	Sunny	Warm	High	Strong	Warm	Same	*Yes*
3	Rainy	Cold	High	Strong	Warm	Change	*No*
4	Sunny	Warm	High	Strong	Cool	Change	*Yes*

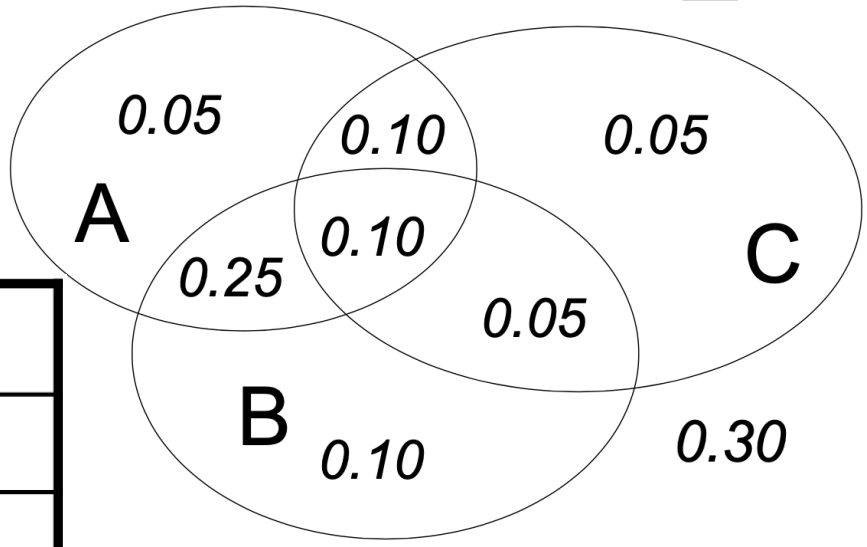
Testing Examples

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport?
1	Sunny	Warm	Normal	Strong	Cool	Change	*?*
2	Rainy	Cold	Normal	Light	Warm	Same	*?*
3	Sunny	Warm	Normal	Light	Warm	Same	*?*
4	Sunny	Cold	Normal	Strong	Warm	Same	*?*

Training Examples (Decision Tree Learning)

Day	Outlook	Temp	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Joint Probability Distribution Table



A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Joint Probability Distribution Example

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

Correlation is NOT Causation

Bayesian Learning Example

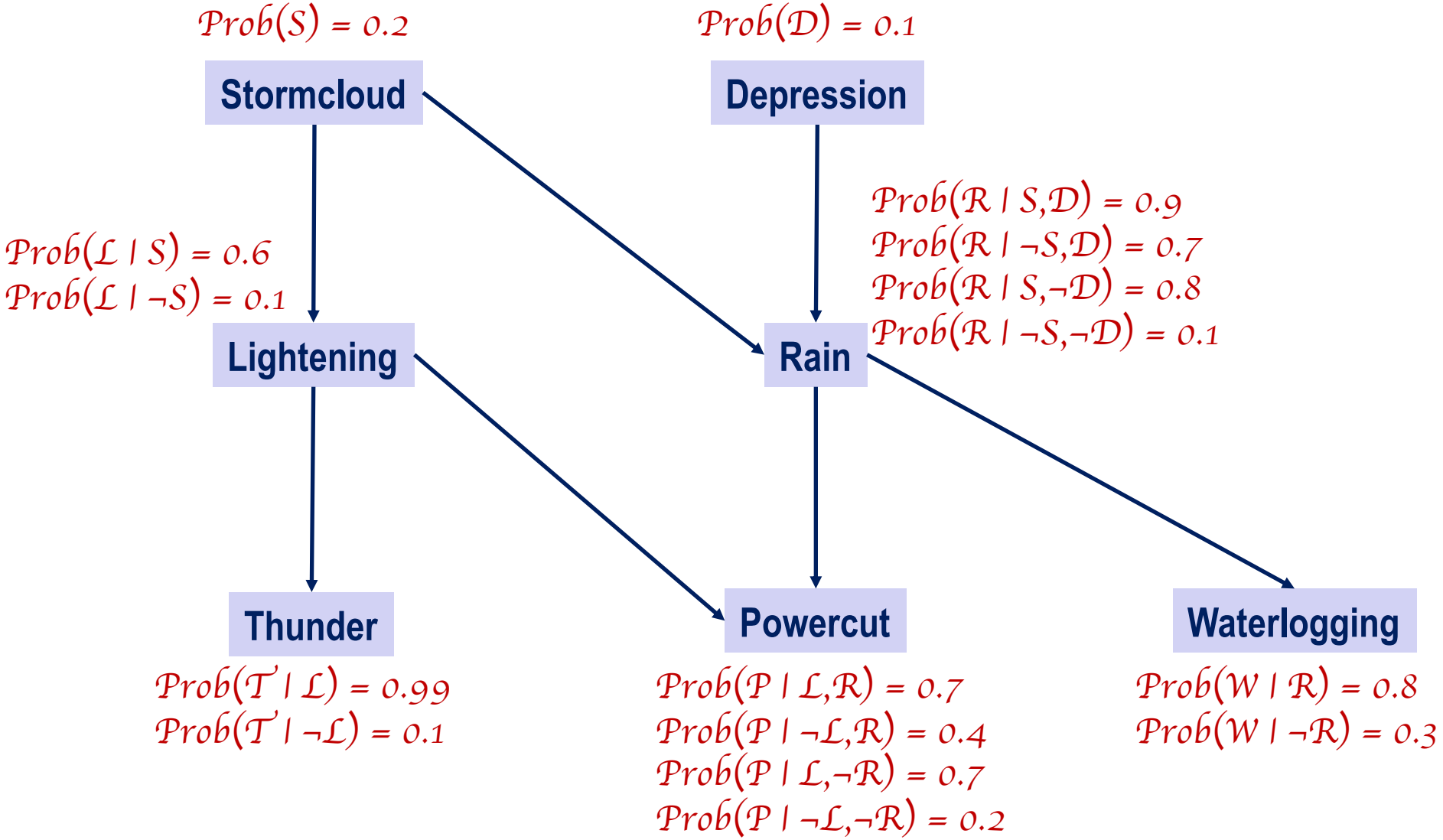
Training Example

Attendance	Book Choice	Practice	Result
High	Text	Yes	Pass
High	Ref	No	Fail
Low	Text	Yes	Pass
Low	Ref	Yes	Fail
High	Ref	No	Pass

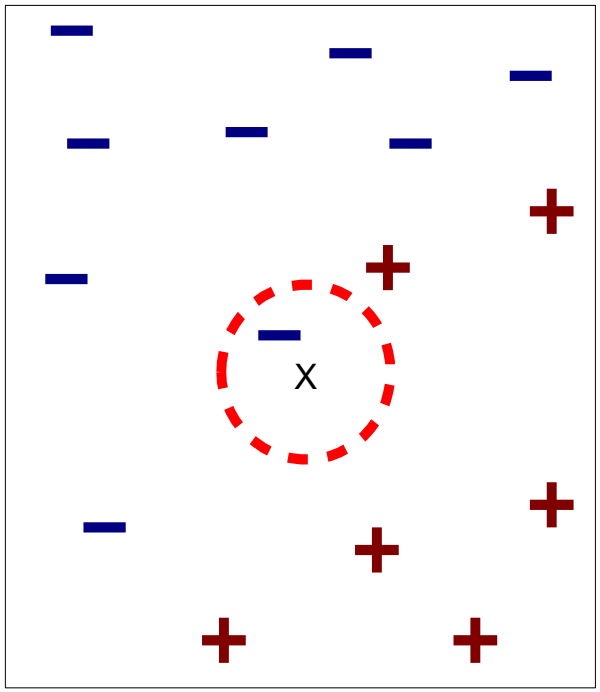
Test Example

Attendance	Book Choice	Practice	Result
High	Ref	Yes	?
Low	Ref	No	?

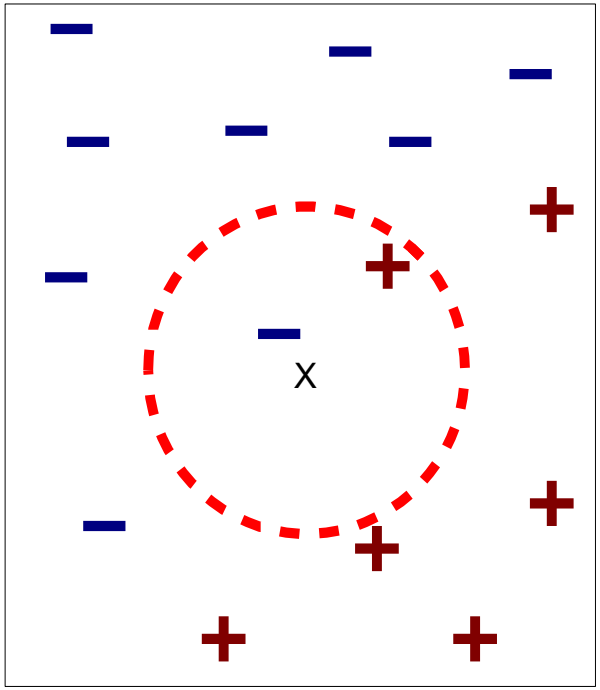
Bayesian Network Example



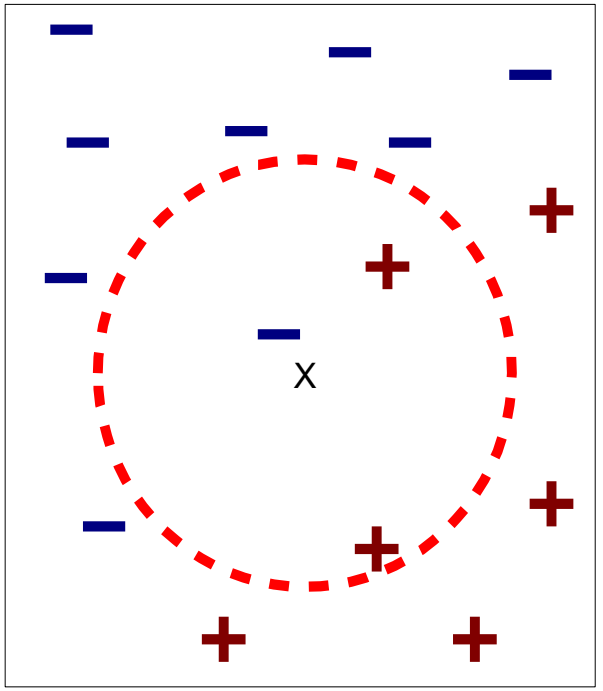
k-Nearest Neighbor



(a) 1-nearest neighbor

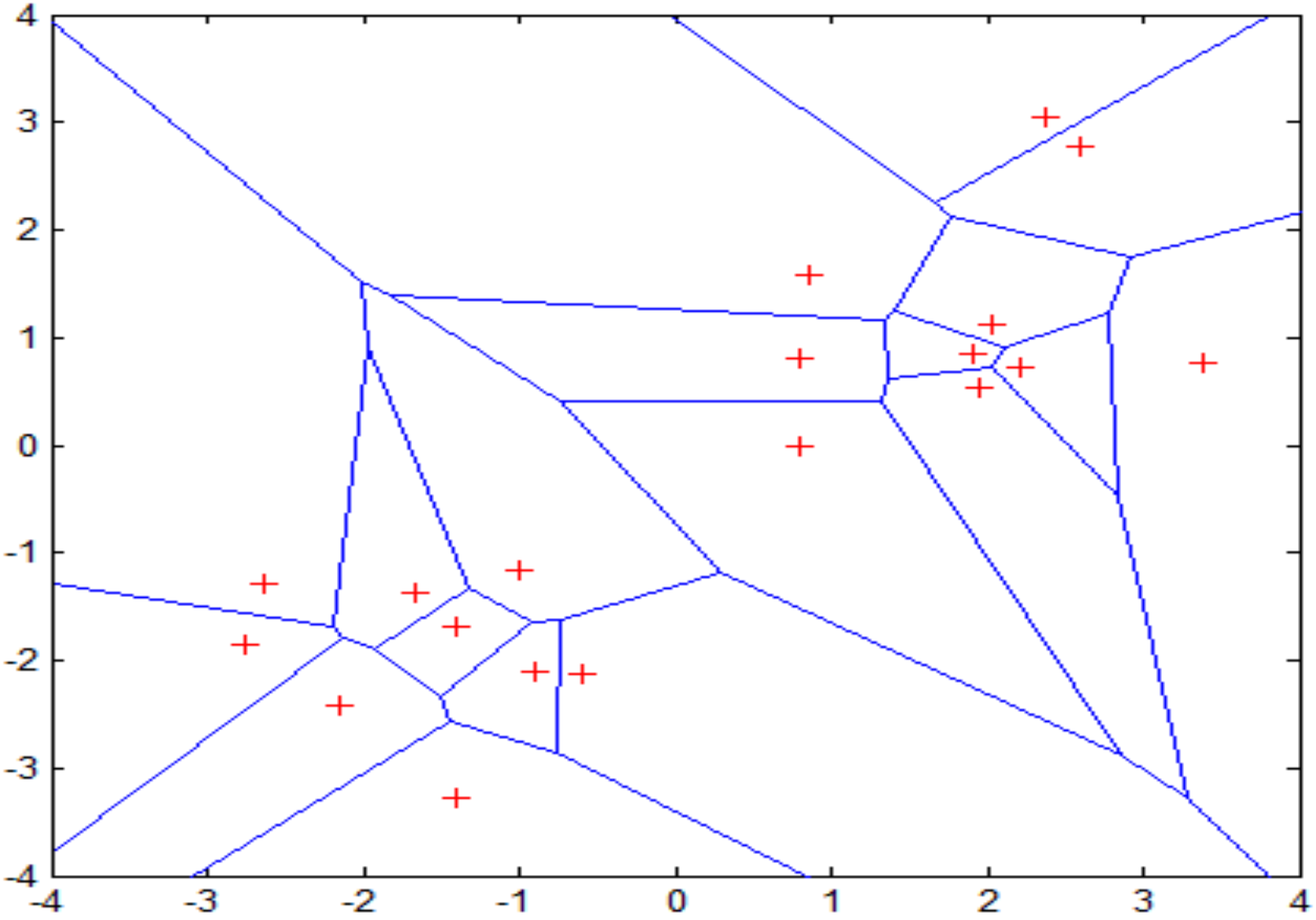


(b) 2-nearest neighbor



(c) 3-nearest neighbor

Voronoi Diagram (Tessellations)



Distance Metrics

Minkowsky:

$$D(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r}$$

Euclidean:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan / city-block:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

Camberra:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Chebychev:

$$D(\boldsymbol{x},\boldsymbol{y}) = \max_{i=1}^{m} |x_i - y_i|$$

Quadratic:

$$D(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T Q(\boldsymbol{x} - \boldsymbol{y}) = \sum_{j=1}^{m} \left(\sum_{i=1}^{m} (x_i - y_i) q_{ji} \right) (x_j - y_j)$$

Q is a problem-specific positive definite $m \times m$ weight matrix

Mahalanobis:

$$D(\boldsymbol{x},\boldsymbol{y}) = [\det V]^{1/m} (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} V^{-1} (\boldsymbol{x} - \boldsymbol{y})$$

V is the covariance matrix of $A_1..A_m$, and A_j is the vector of values for attribute j occuring in the training set instances $1..n$.

Correlation:

$$D(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{m} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x_i})^2 \sum_{i=1}^{m} (y_i - \overline{y_i})^2}}$$

$\overline{x_i} = \overline{y_i}$ and is the average value for attribute i occuring in the training set.

Chi-square:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{1}{sum_i} \left(\frac{x_i}{size_x} - \frac{y_i}{size_y} \right)^2$$

sum_i is the sum of all values for attribute i occuring in the training set, and $size_x$ is the sum of all values in the vector $\boldsymbol{x}$.

Kendall's Rank Correlation:

$$D(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{i-1} \mathrm{sign}(x_i - x_j)\,\mathrm{sign}(y_i - y_j)$$

sign(x)=-1, 0 or 1 if $x < 0$, $x = 0$, or $x > 0$, respectively.

Figure 1. Equations of selected distance functions. ($\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of m attribute values).

kNN Example

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

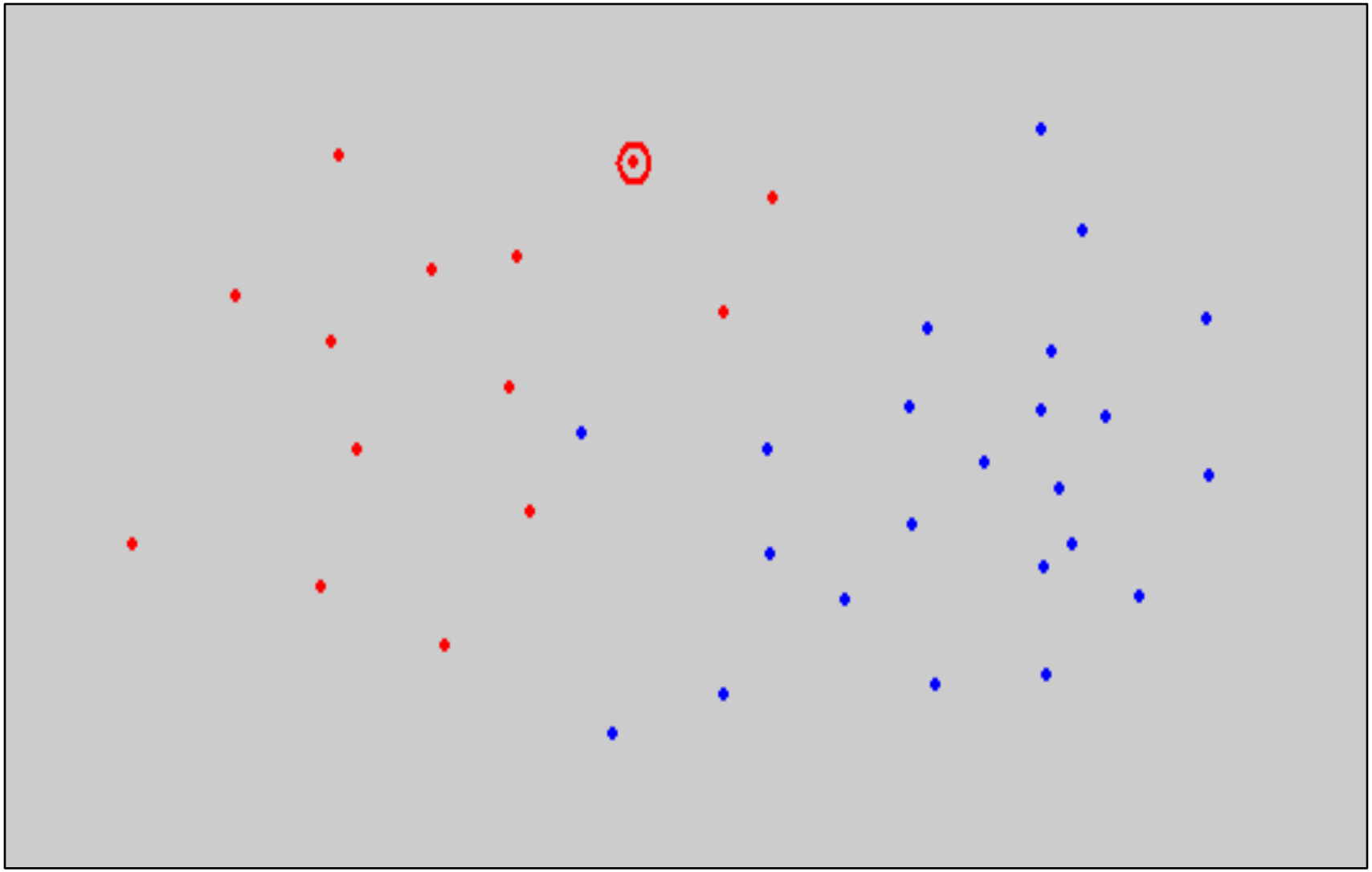
Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Mammals ?

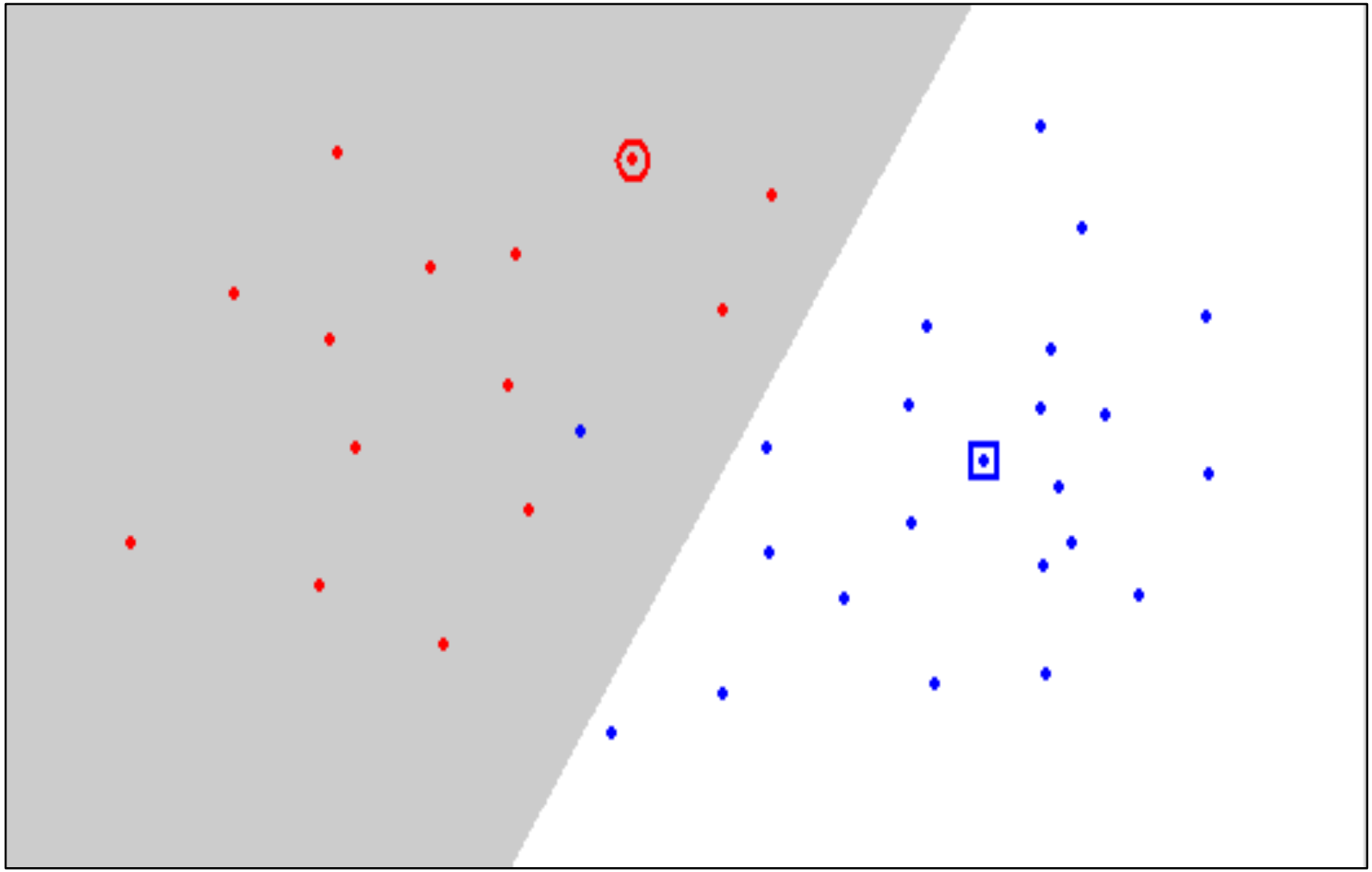
or

Non-Mammals ?

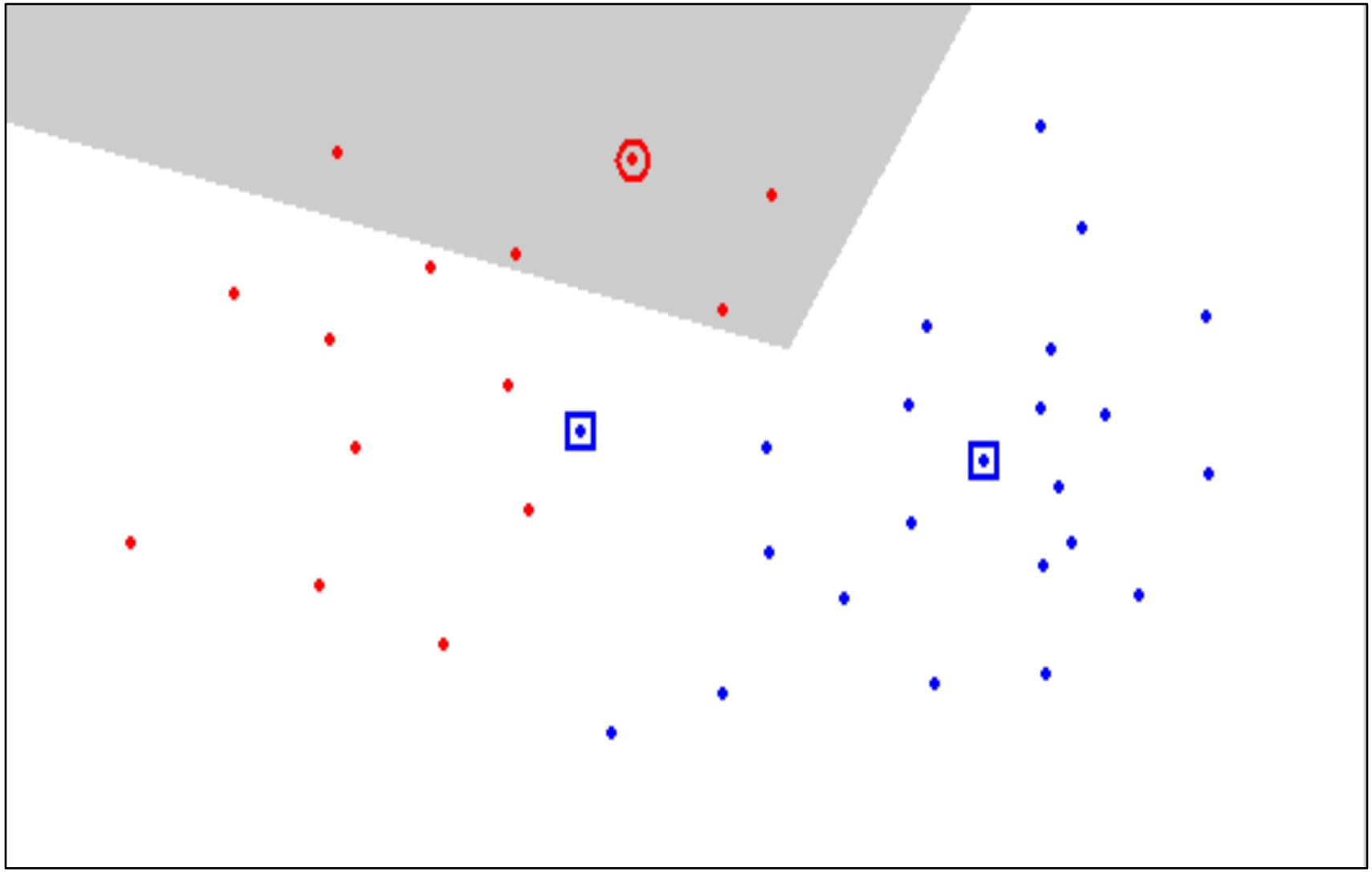
Condensing



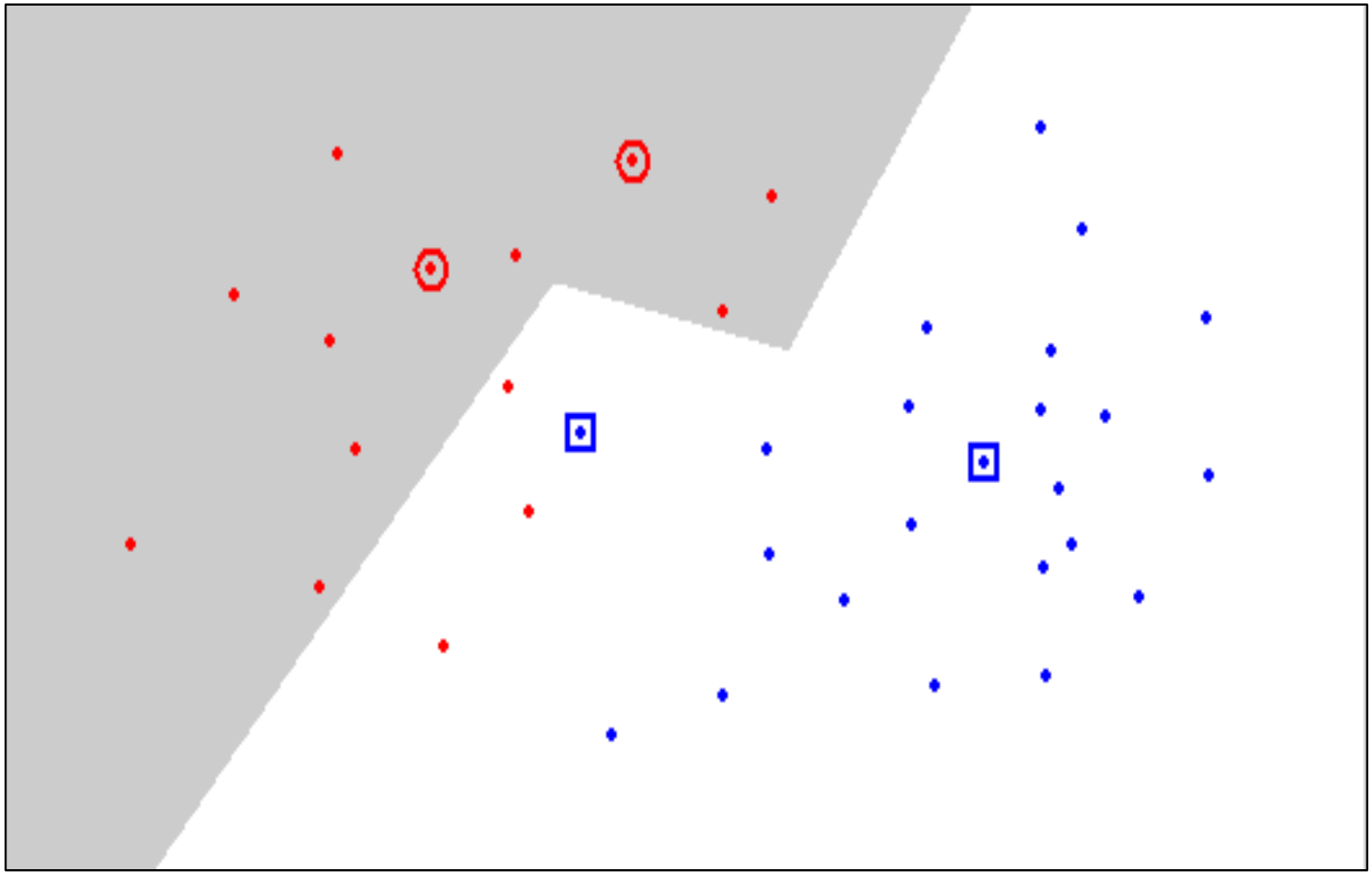
Condensing



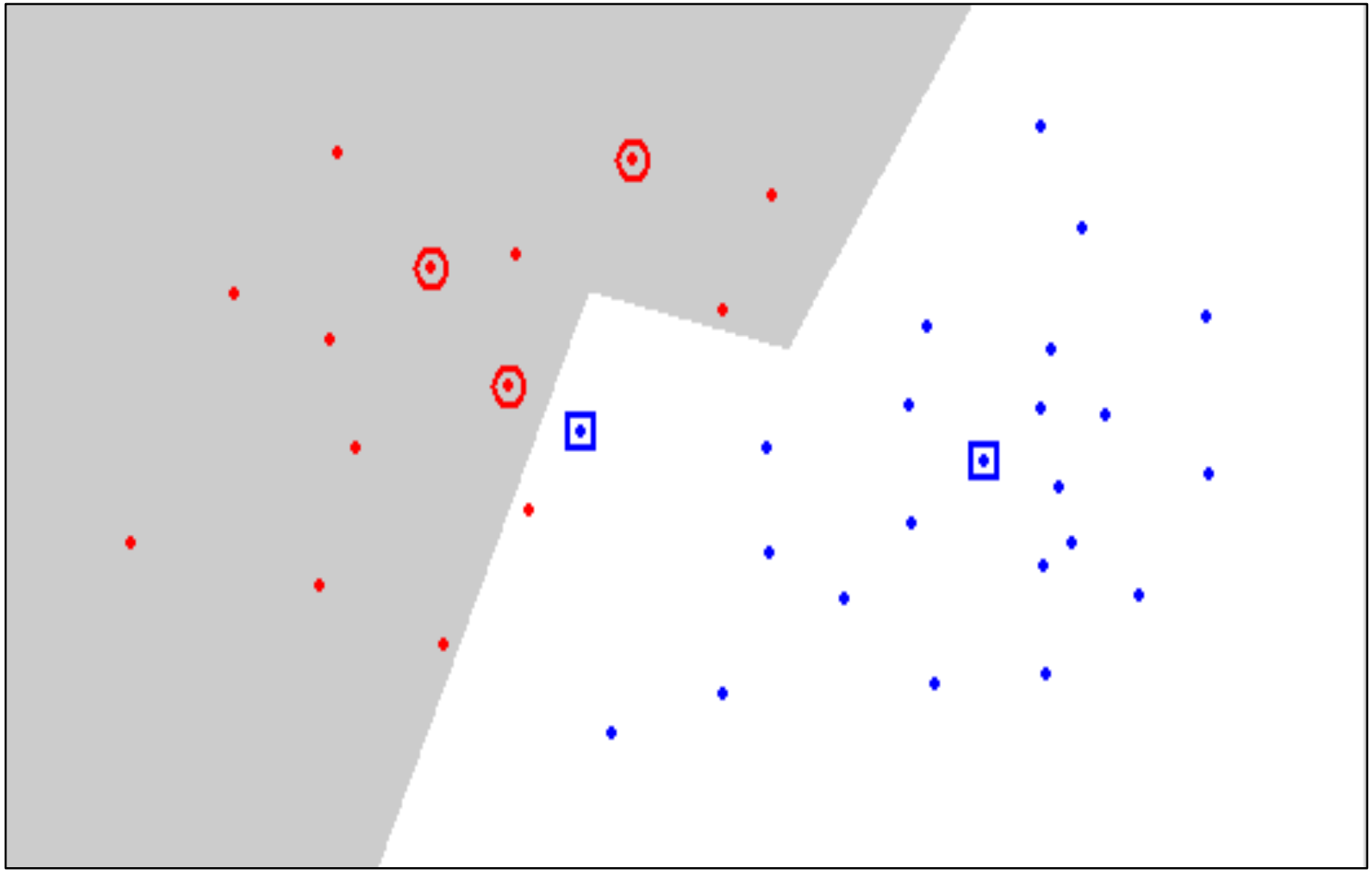
Condensing



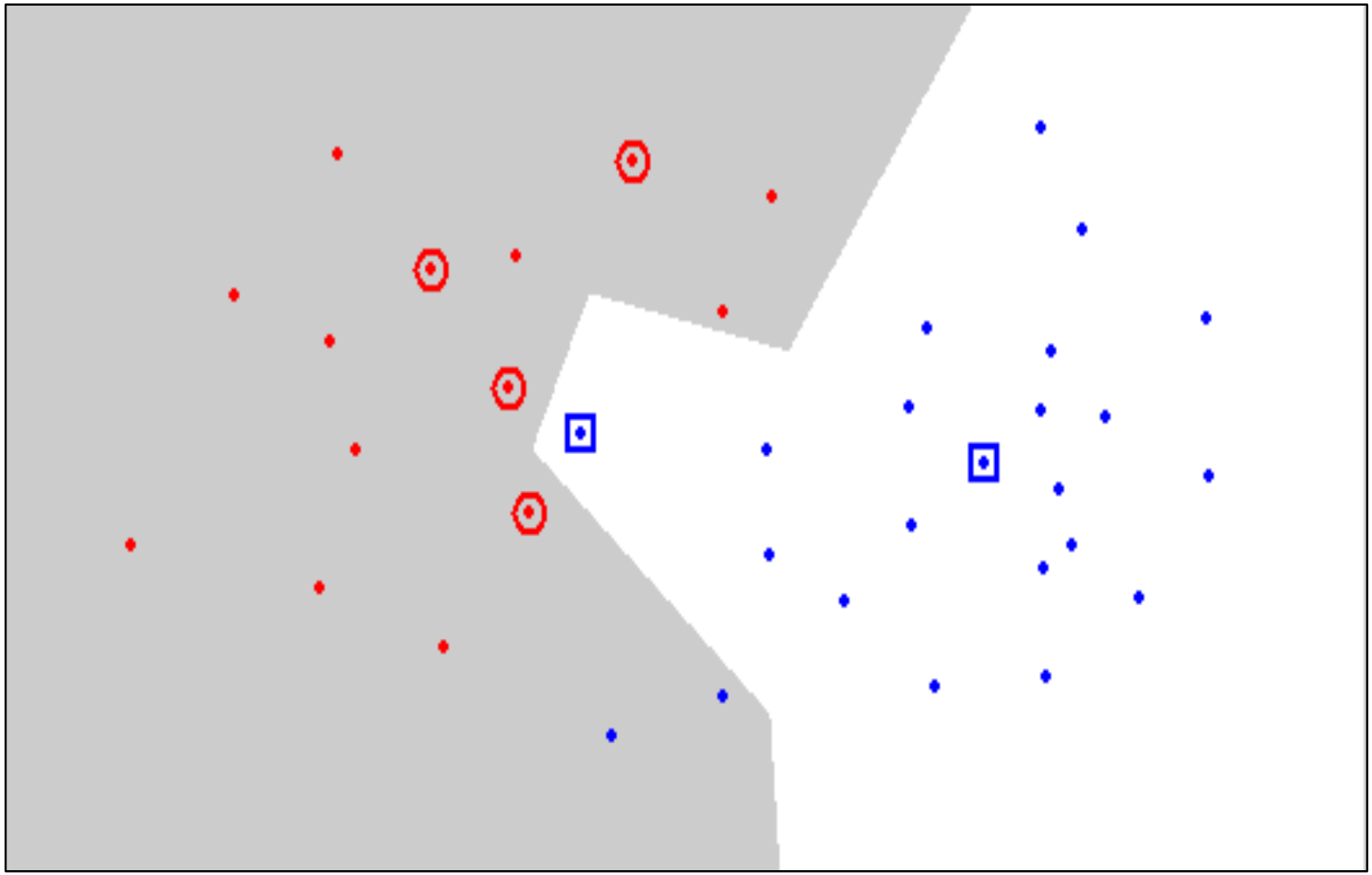
Condensing



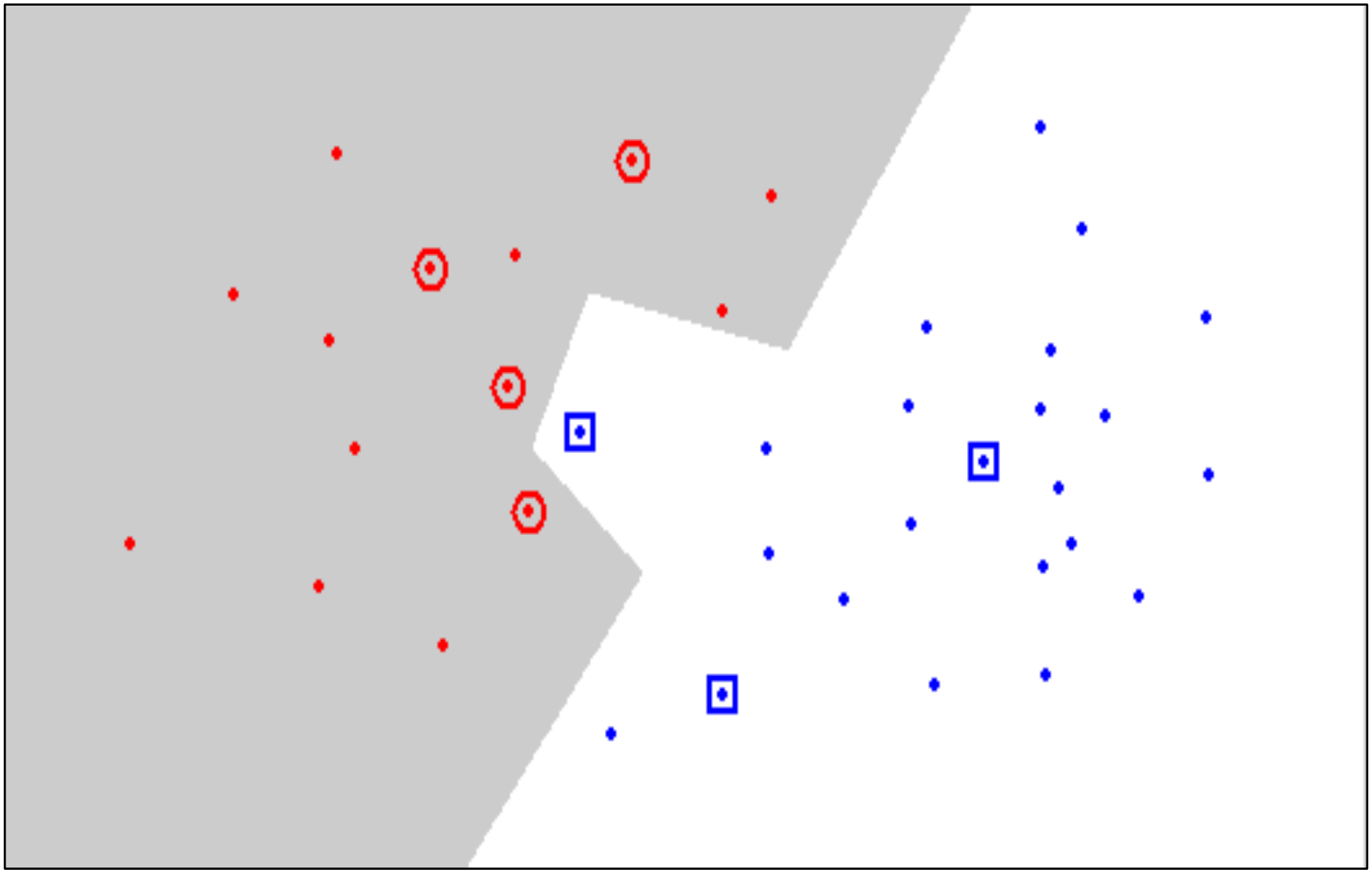
Condensing



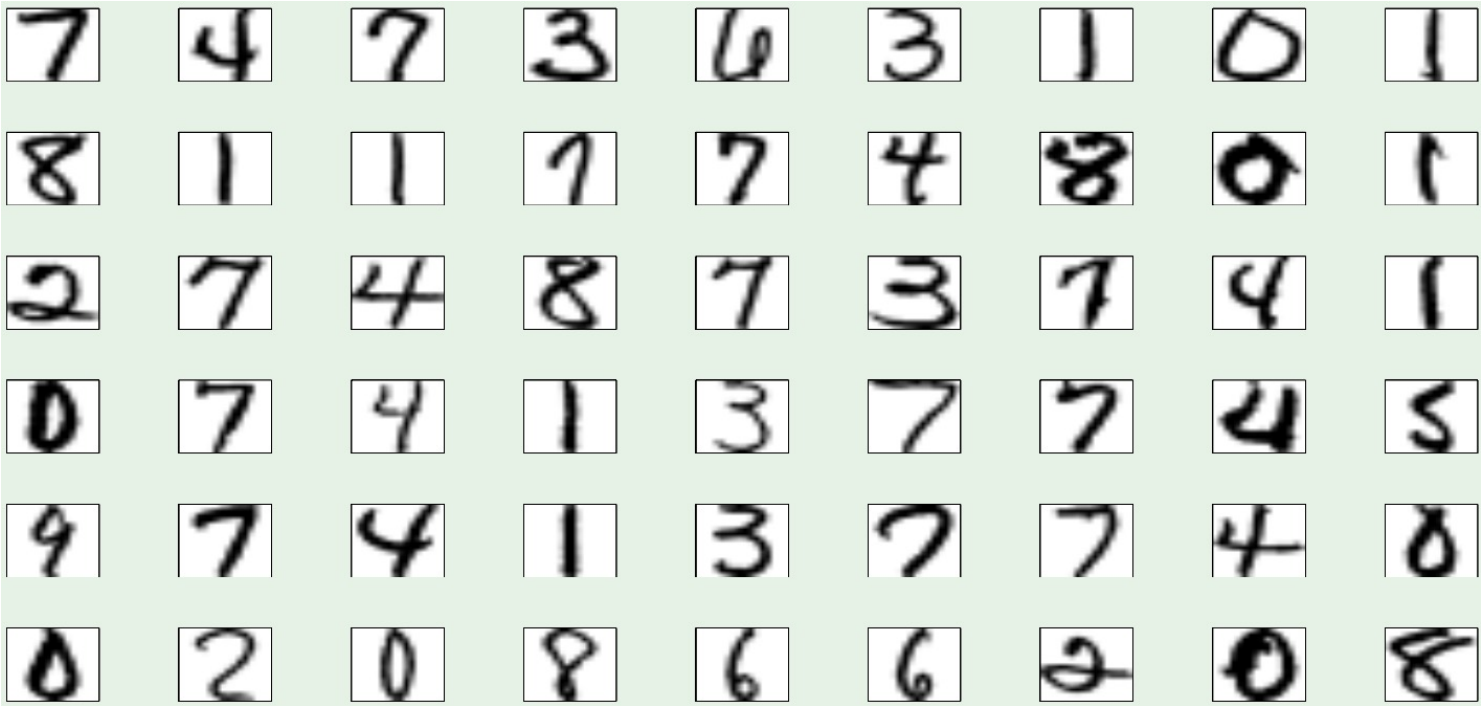
Condensing

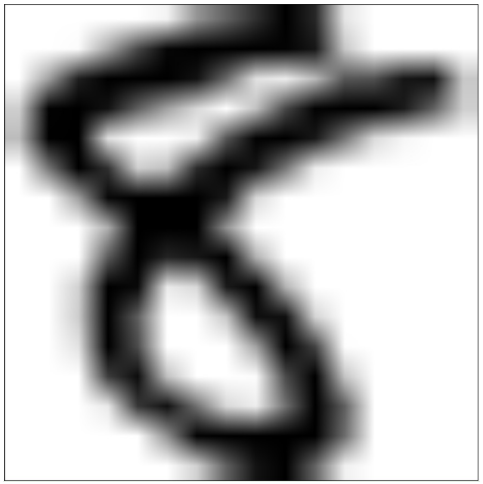


Condensing

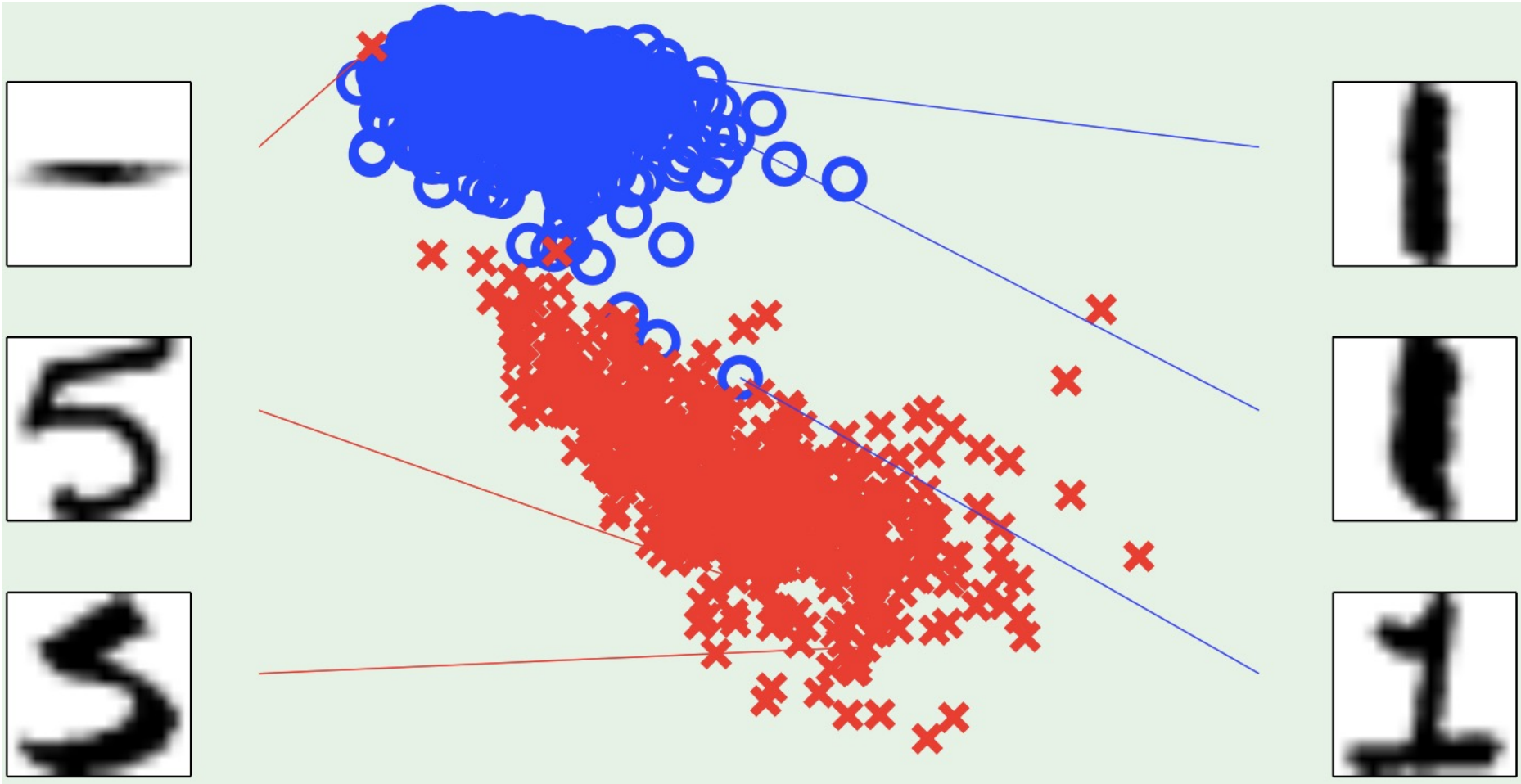


Handwritten Digit Recognition Dataset

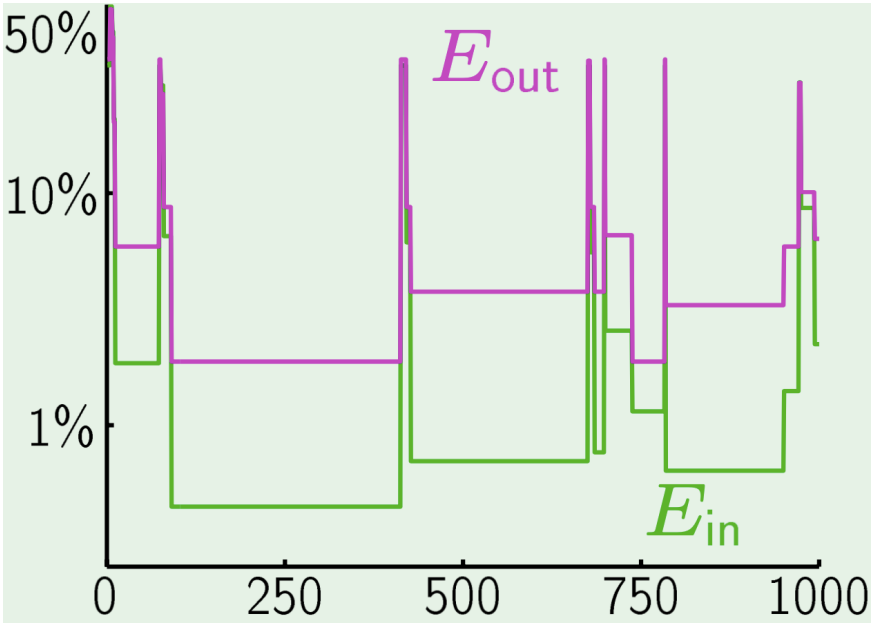


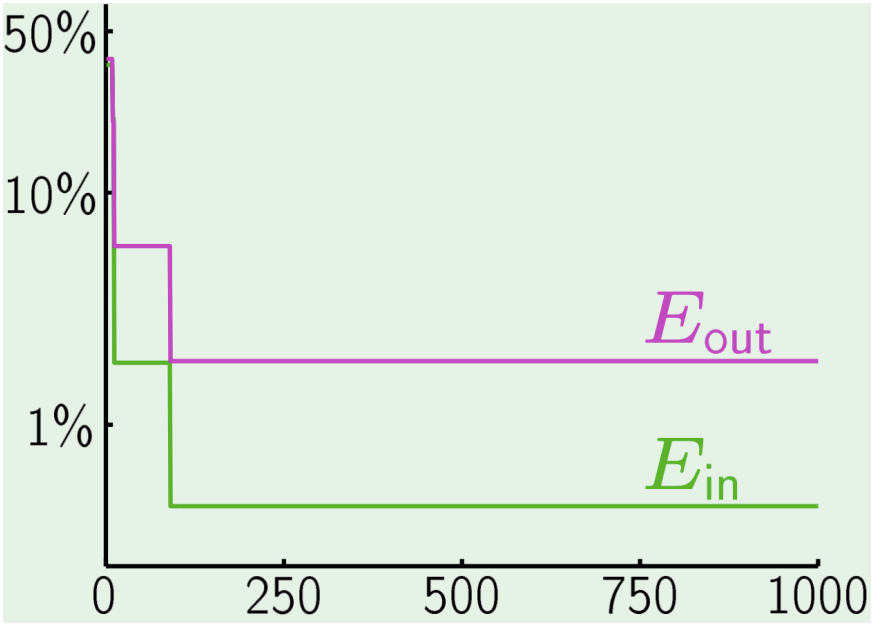


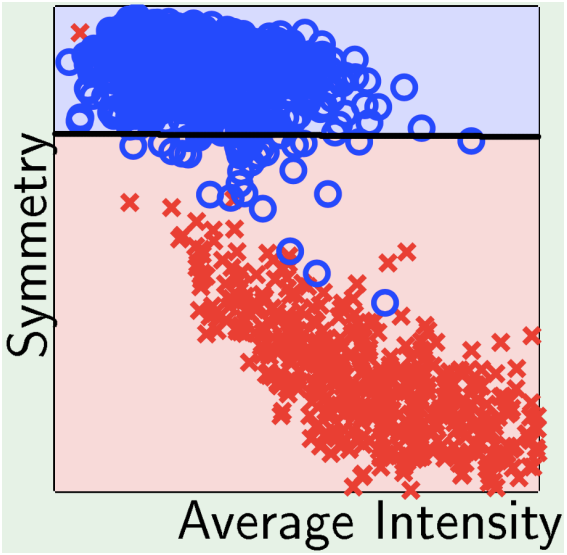
Feature Illustration

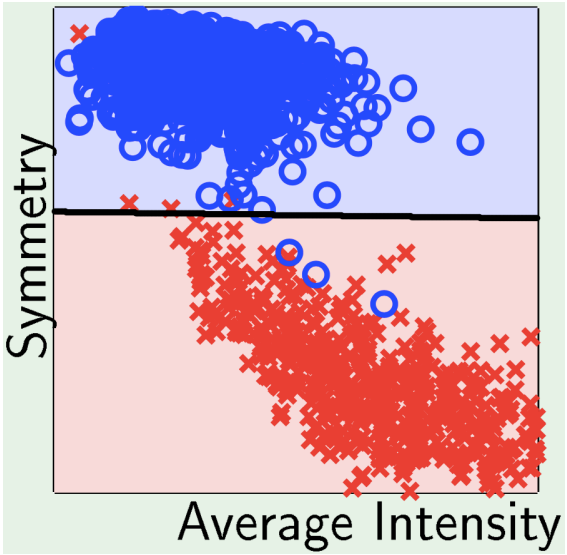


Perceptron Learning vs. Pocket Learning

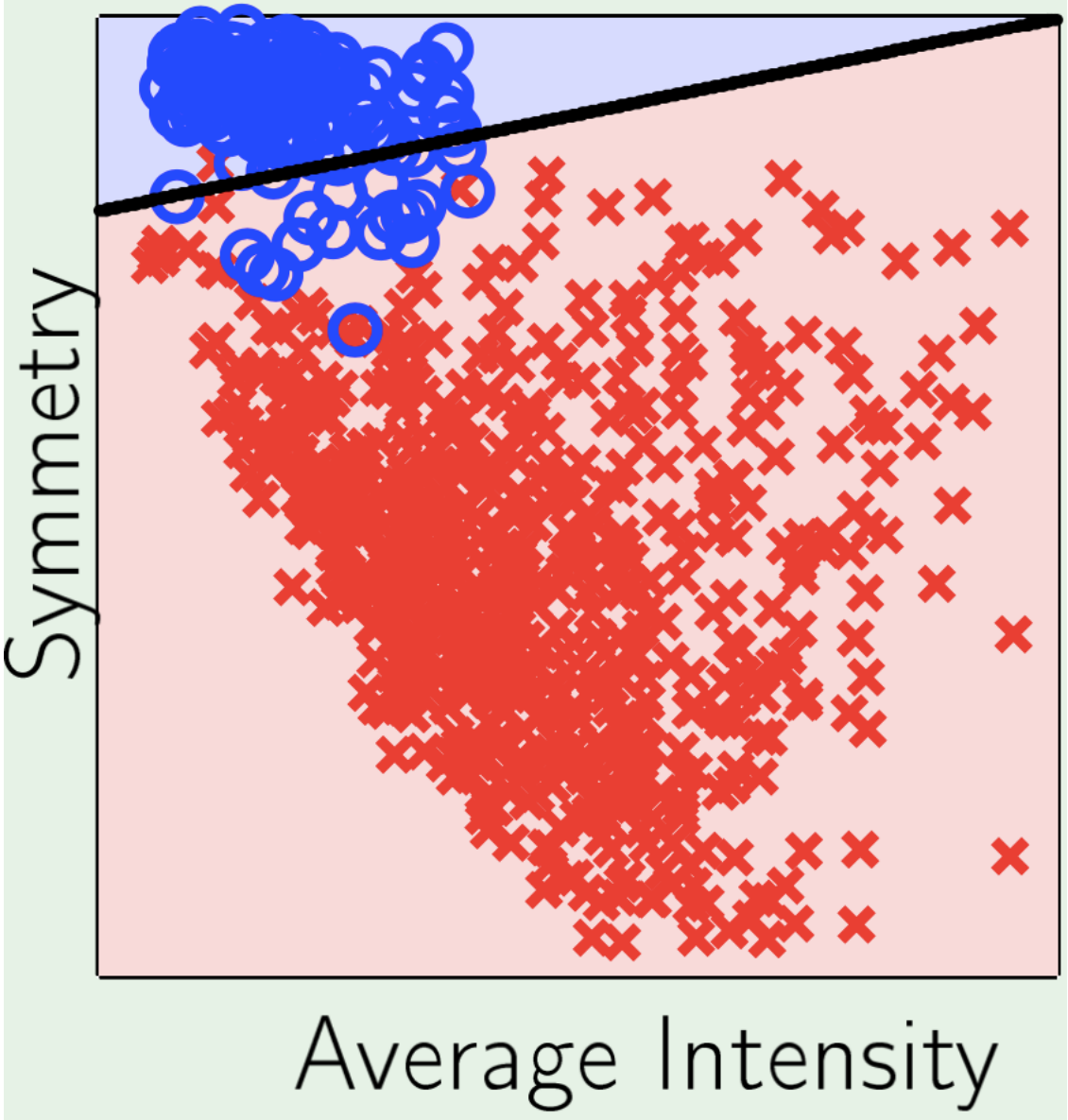




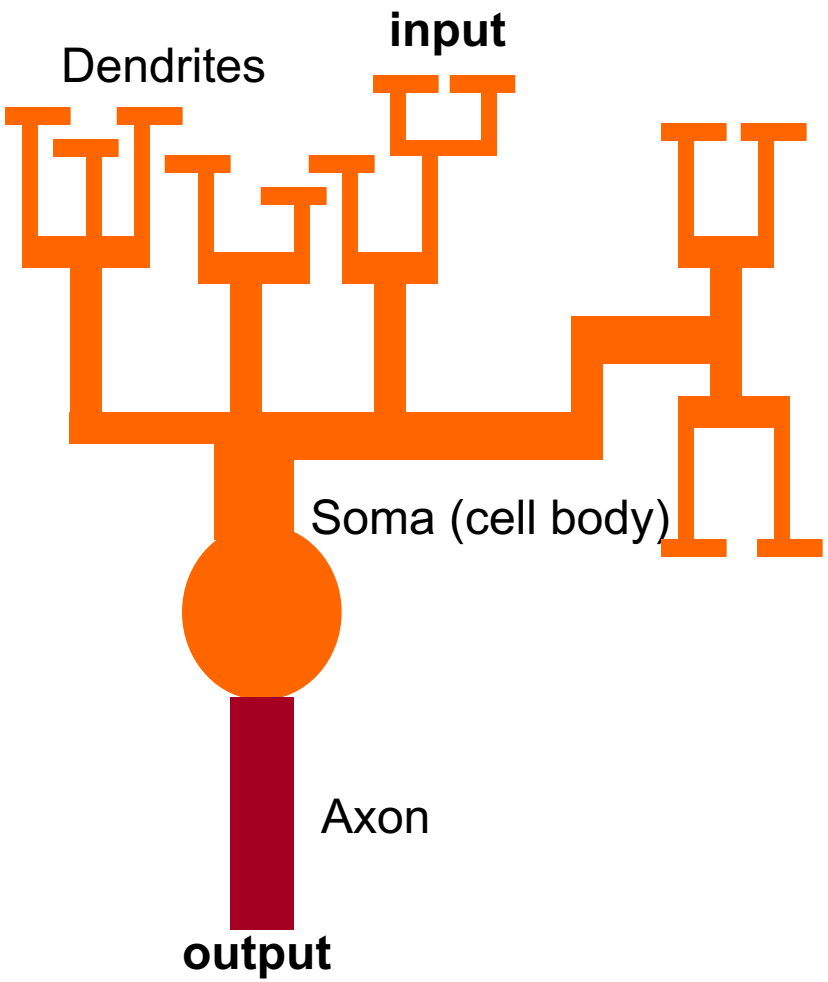


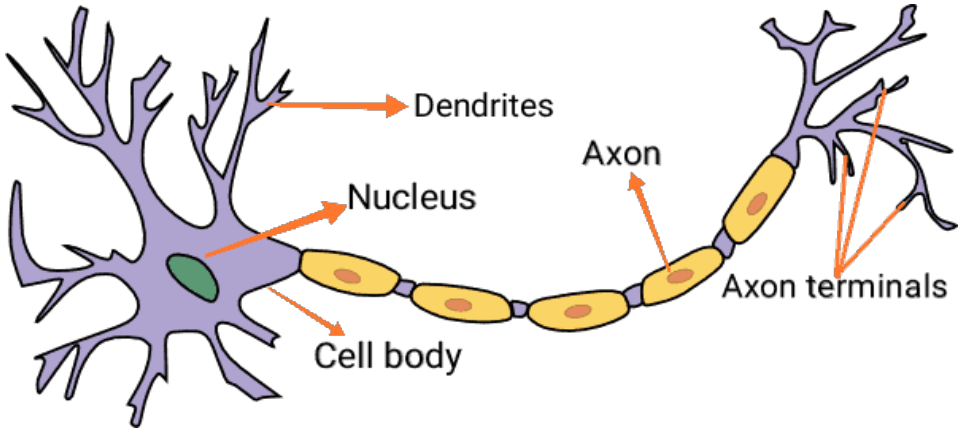


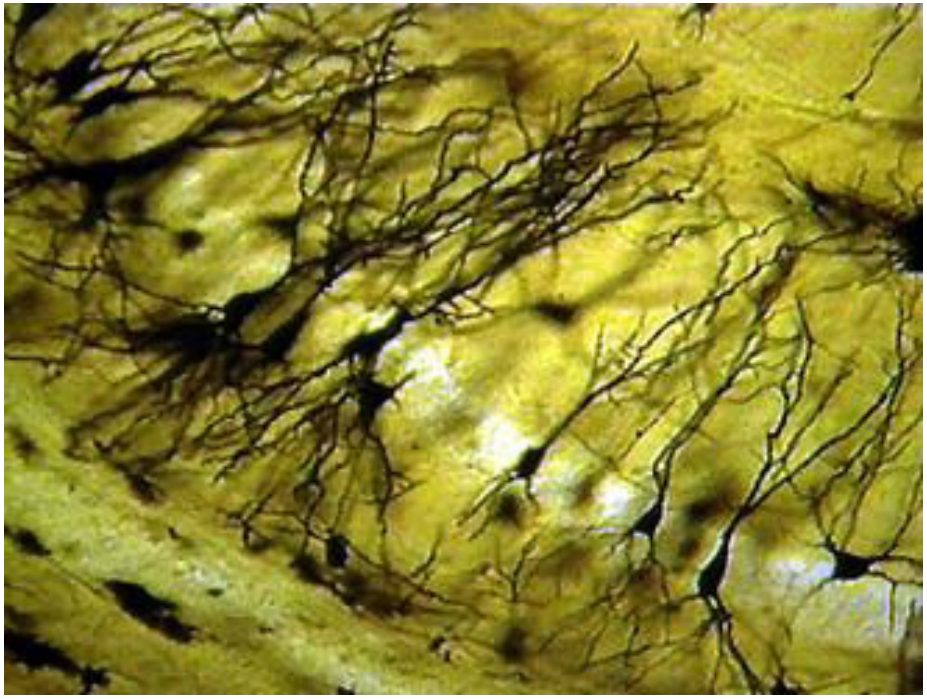
Linear Regression Boundary



Biological Inspiration





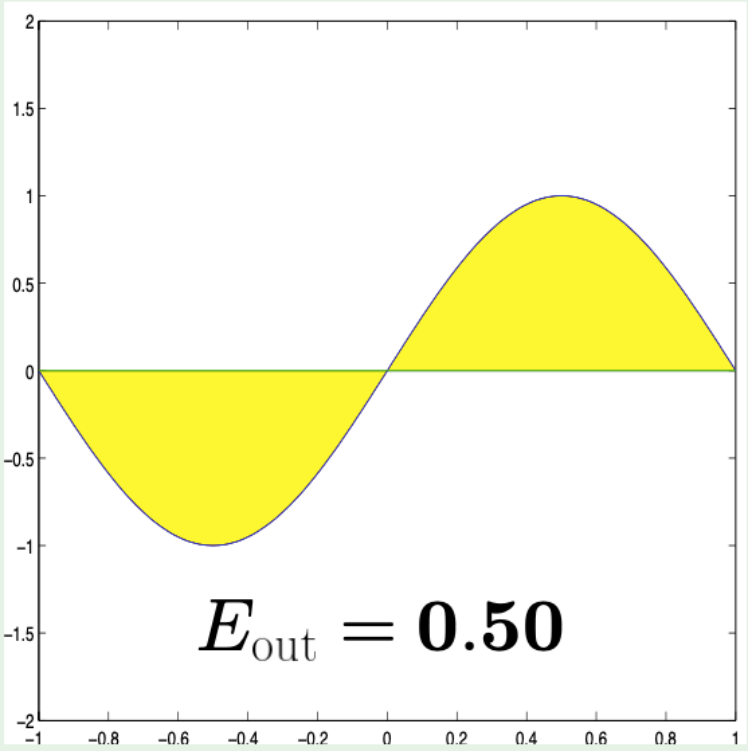


Recursive Bound on Growth Function

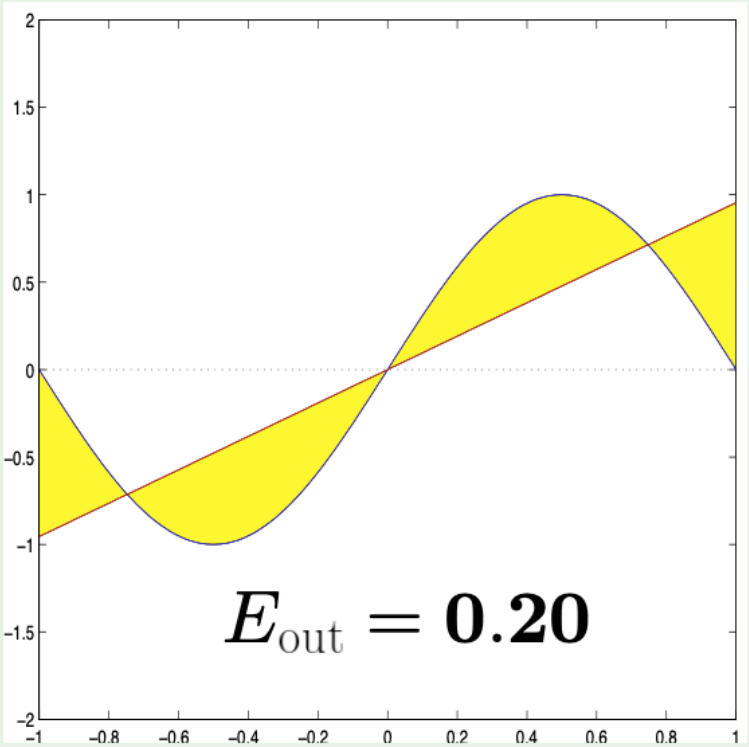
	# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1	α	$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
S_2						
S_2^+	β	$+1$	-1	...	$+1$	$+1$
		-1	-1	...	$+1$	$+1$
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
S_2^-	β	$+1$	-1	...	$+1$	-1
		-1	-1	...	$+1$	-1
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

Approximation – $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

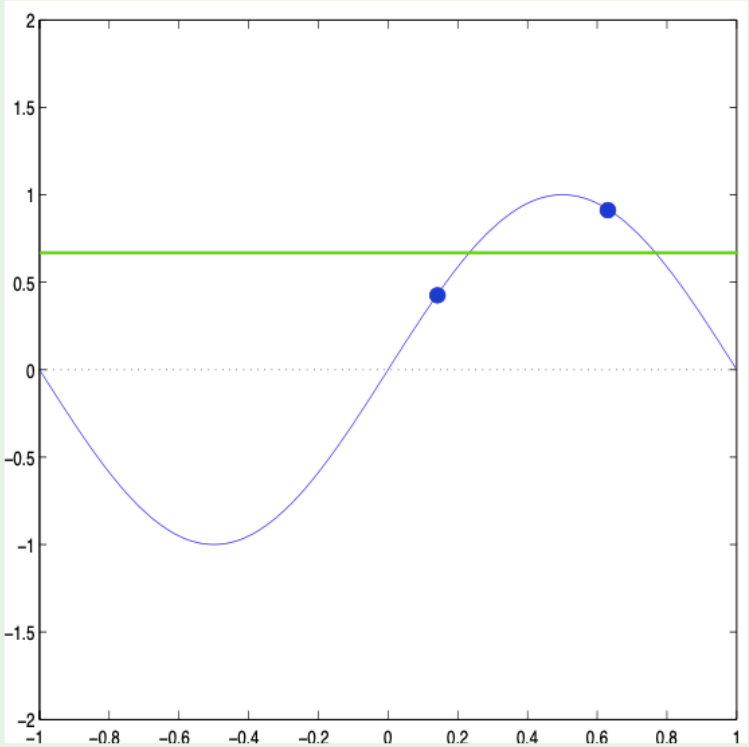


$\mathcal{H}_1$

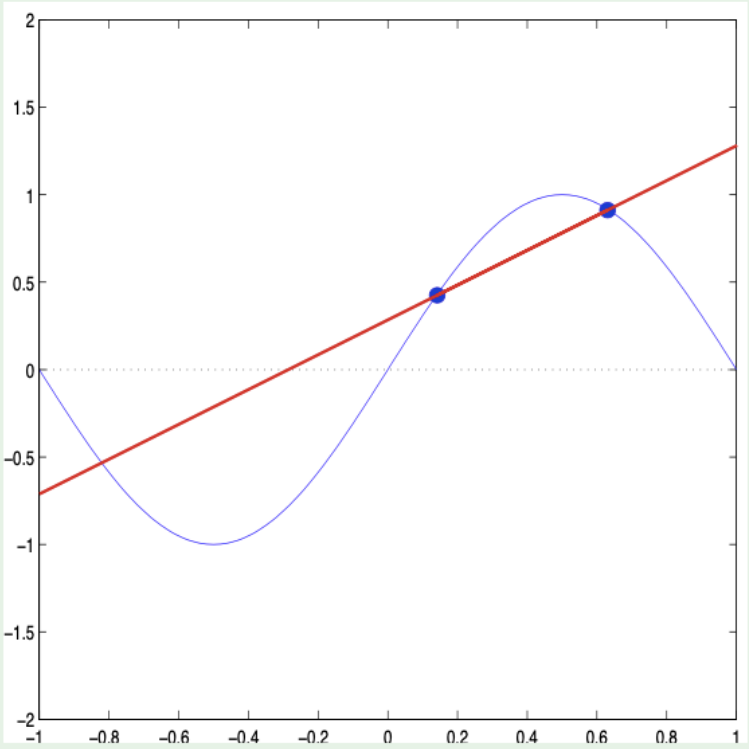


Learning – $\mathcal{H}_0$ versus $\mathcal{H}_1$

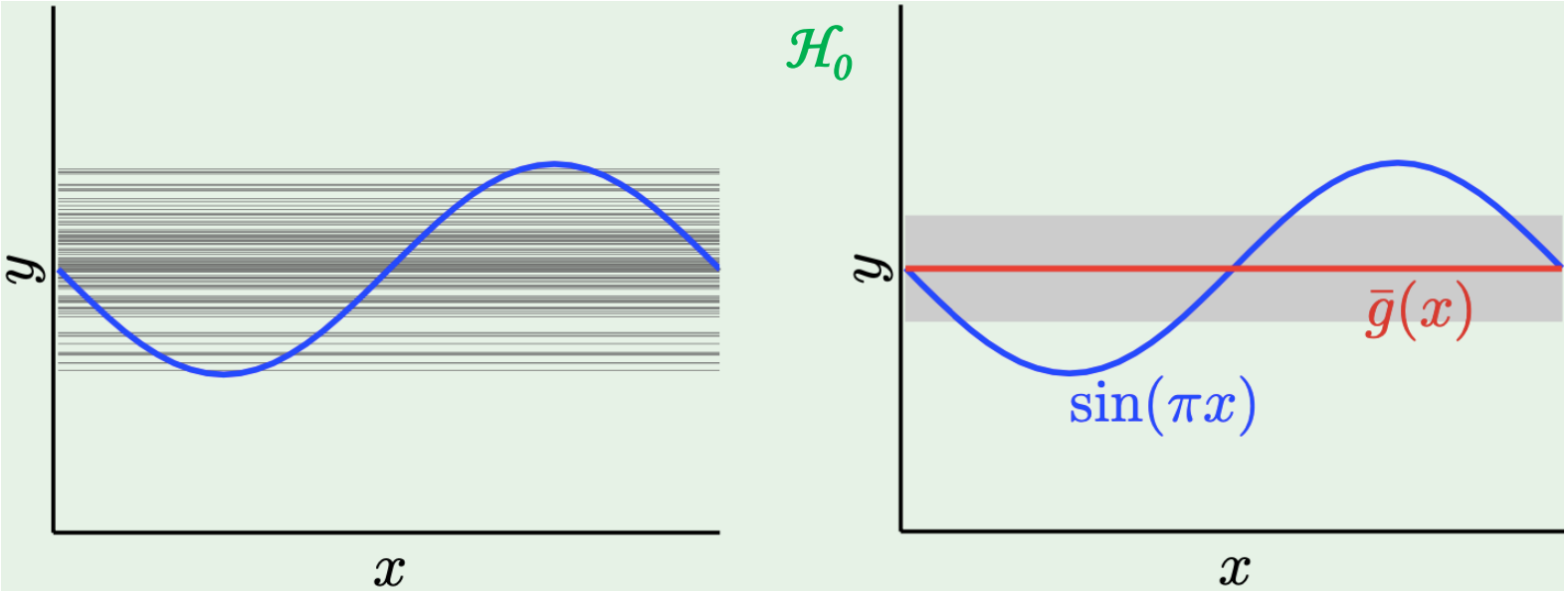
$\mathcal{H}_0$



$\mathcal{H}_1$

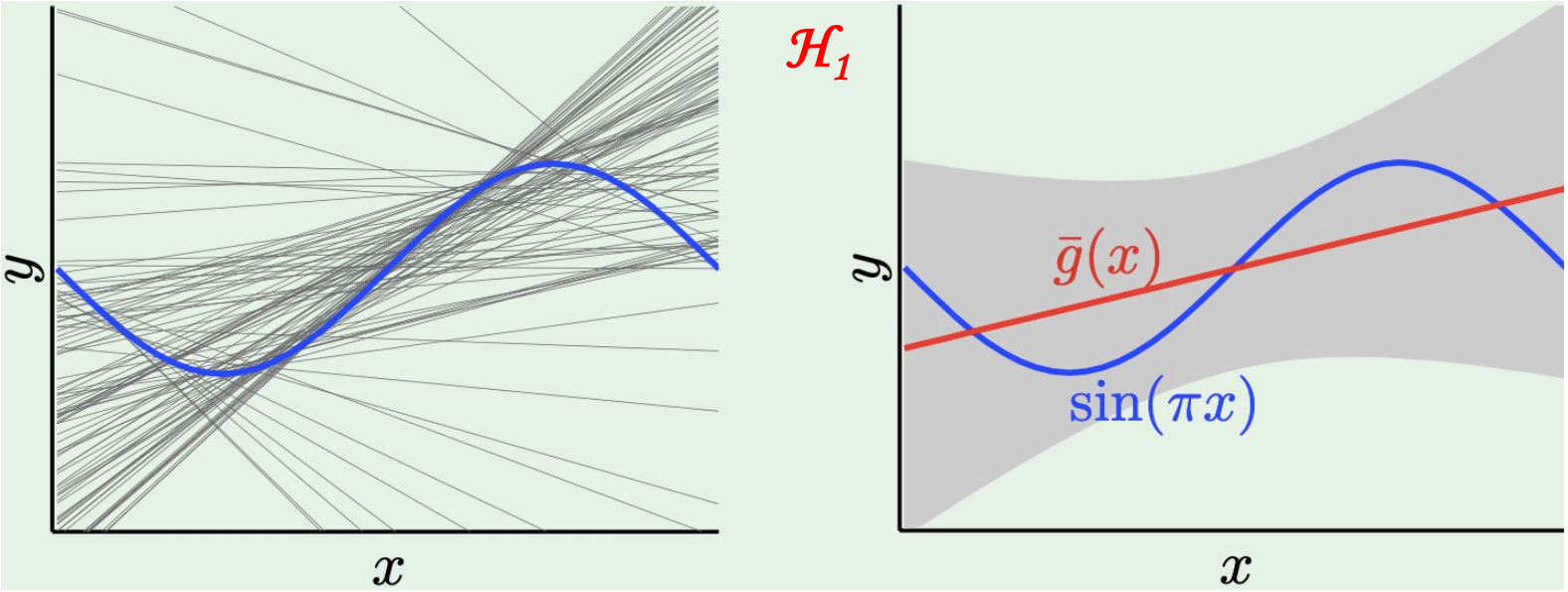


Bias and Variance – $\mathcal{H}_0$ versus $\mathcal{H}_1$



bias = 0.5

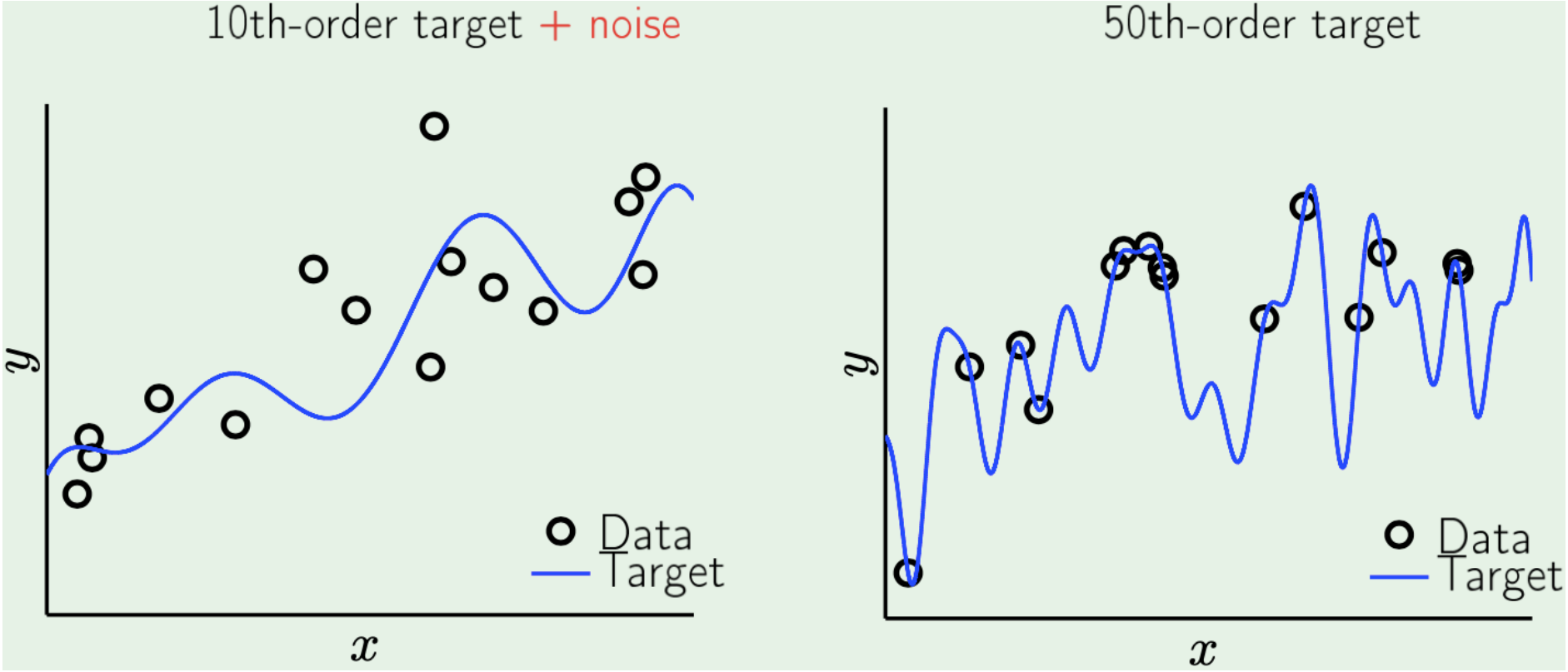
var. = 0.25



bias = 0.21

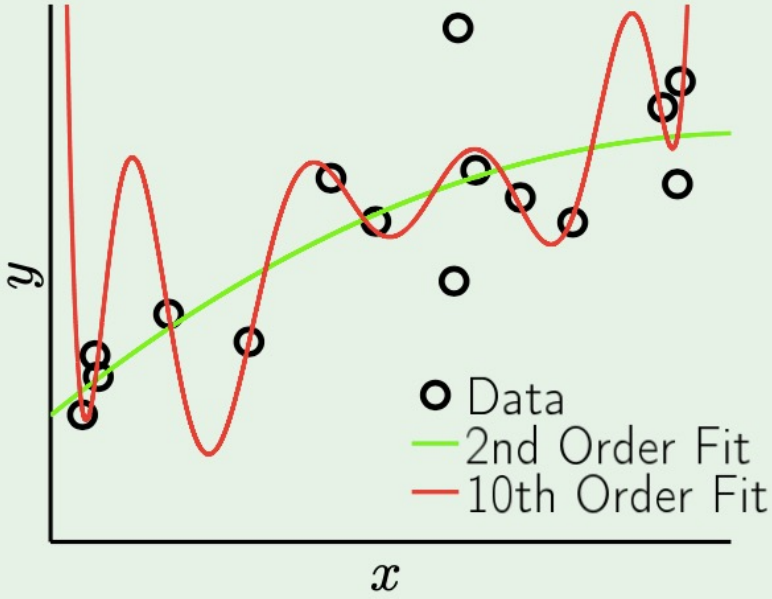
var. = 1.69

Overfitting – Case Study

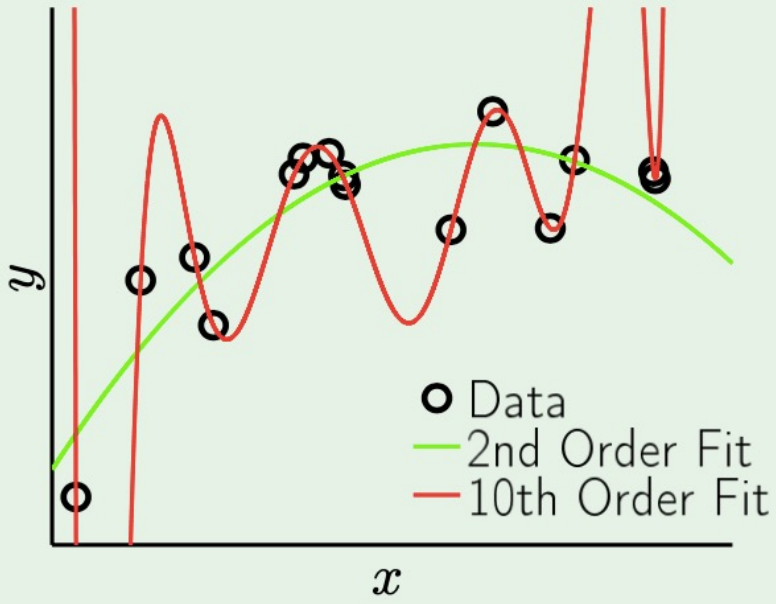


2nd Order Fit versus 10th Order Fit

10th-order target + noise



50th-order target



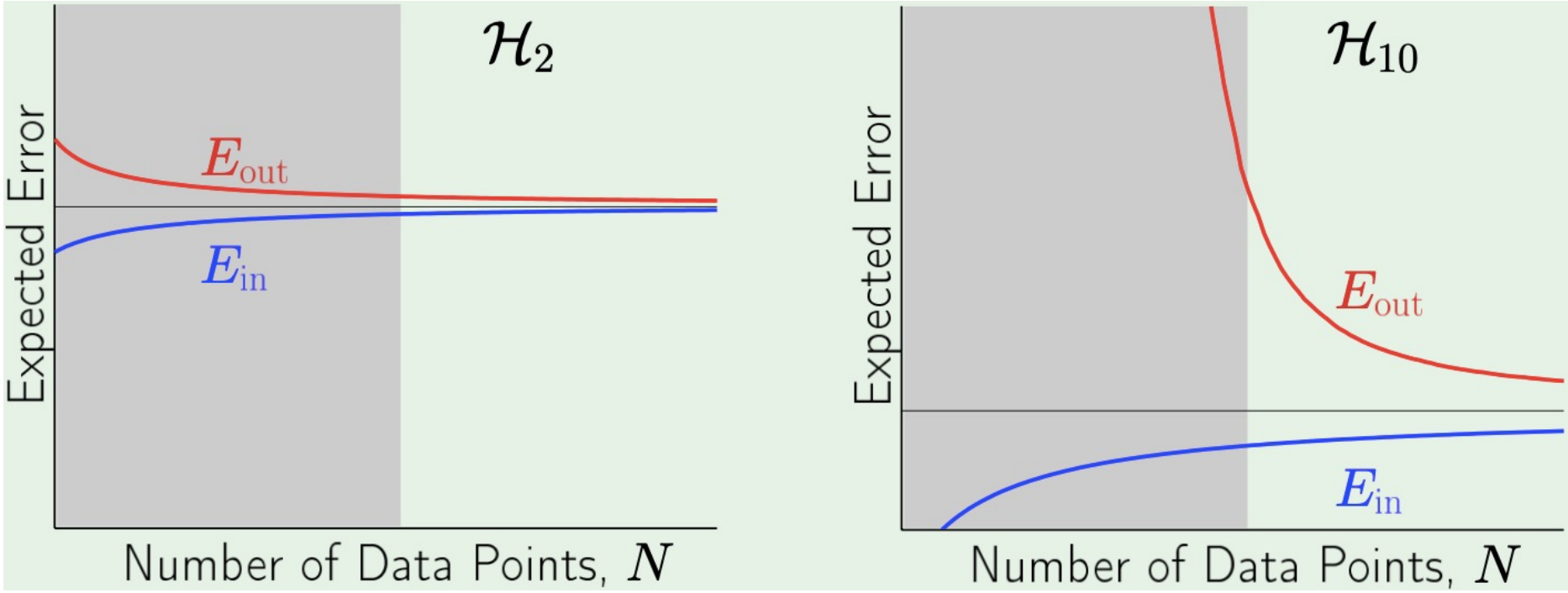
Noisy low-order target

	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	9.00

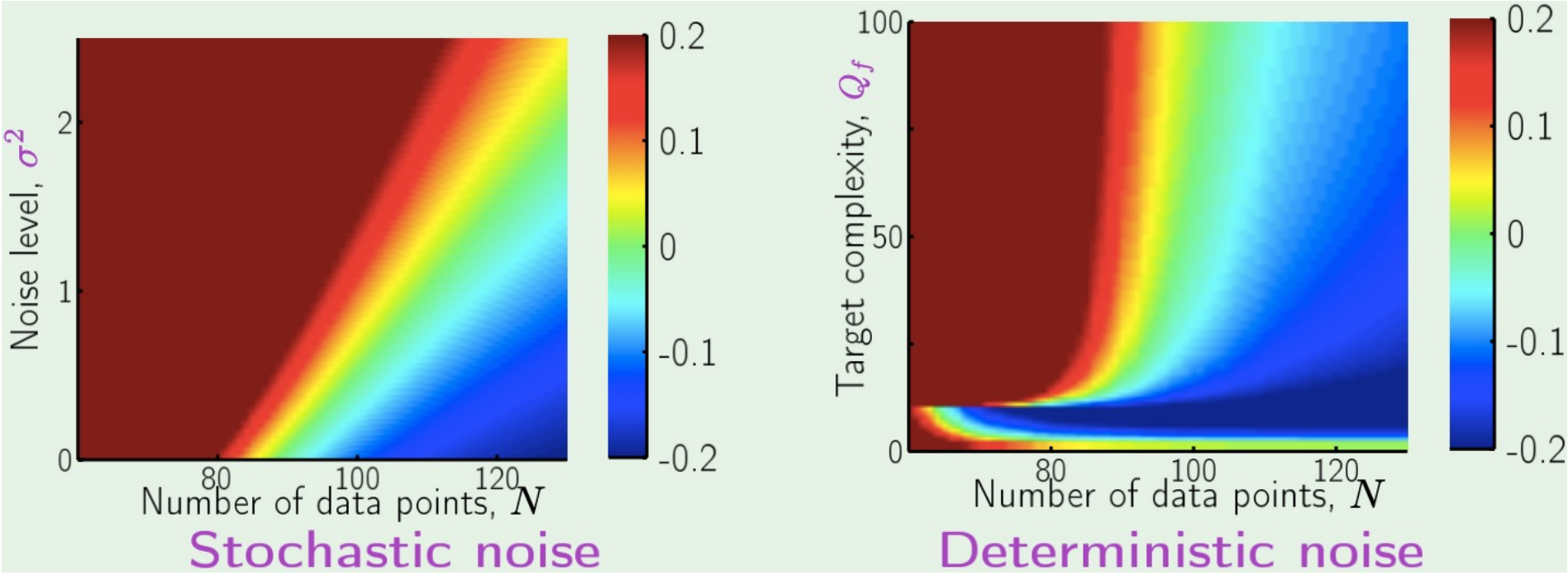
Noiseless high-order target

	2nd Order	10th Order
E_{in}	0.029	10^{-5}
E_{out}	0.120	7680

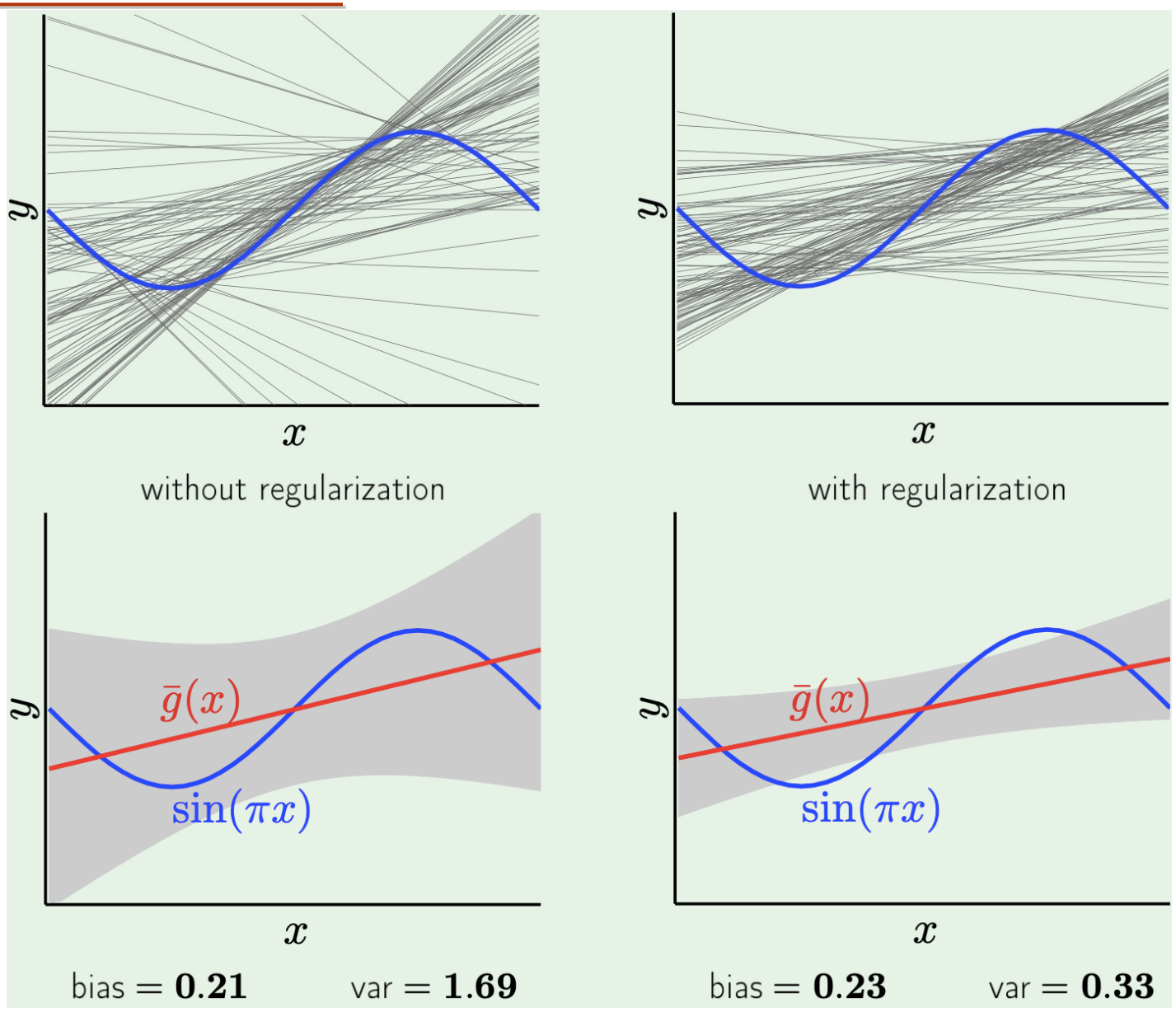
Overfitting Zone



Impact of Noise to Overfitting



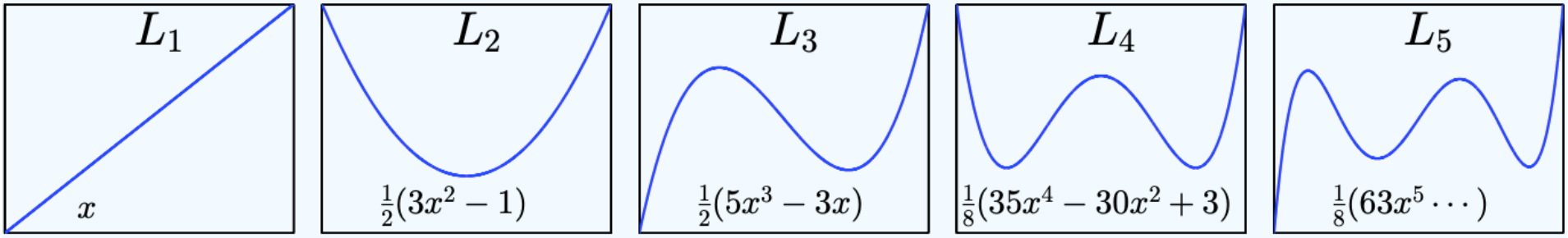
Benefit of Regularization – Example



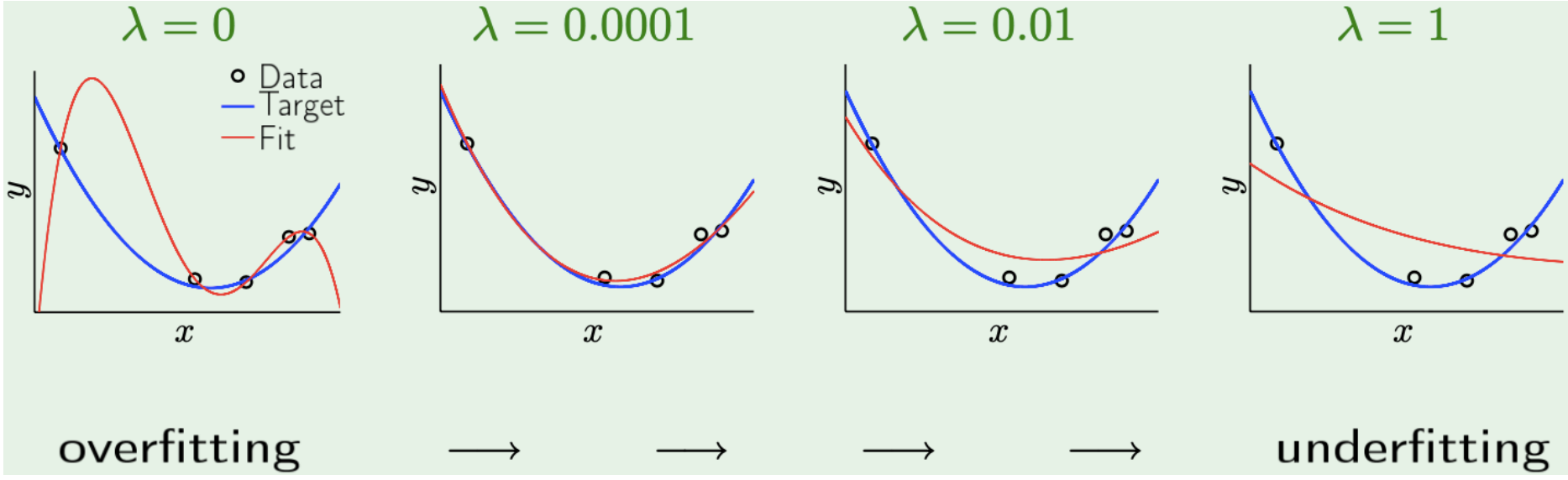
Polynomial Hypothesis Model

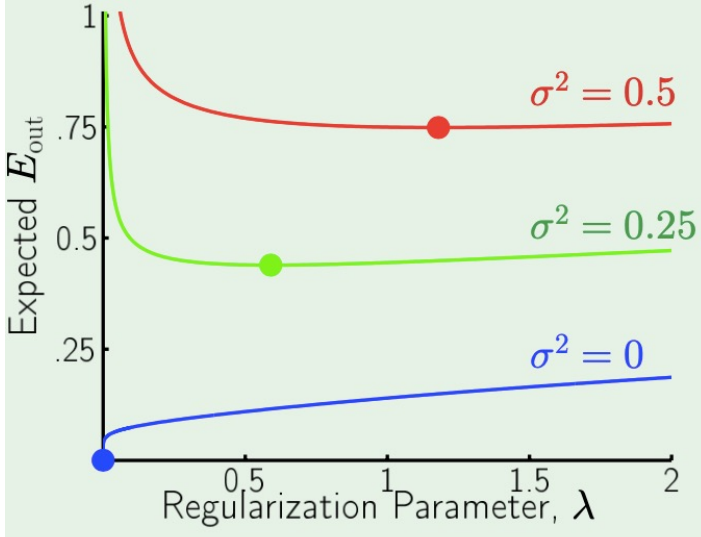
$\mathcal{H}_{\mathrm{Q}}$: polynomials of order Q — linear regression in $\mathcal{Z}$ space

$$\mathbf{z} = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_{\mathrm{Q}}(x) \end{bmatrix} \qquad \mathcal{H}_{\mathrm{Q}} = \left\{ \sum_{q=0}^{Q} w_q \, L_q(x) \right\}$$

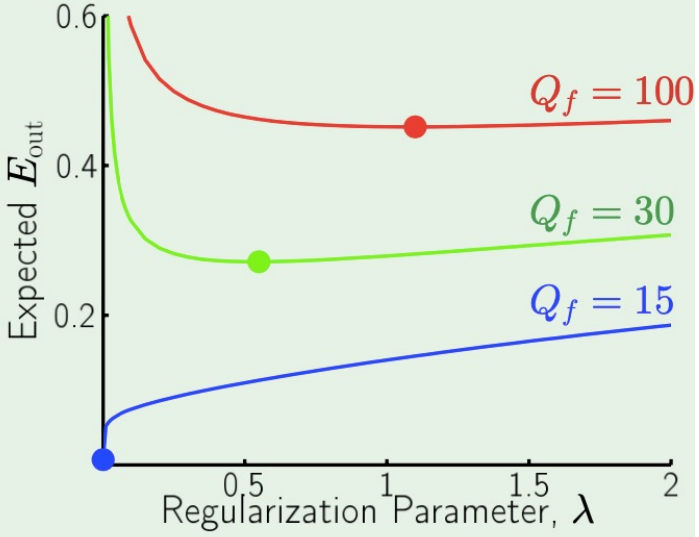


Impact of Regularization Parameter (λ)





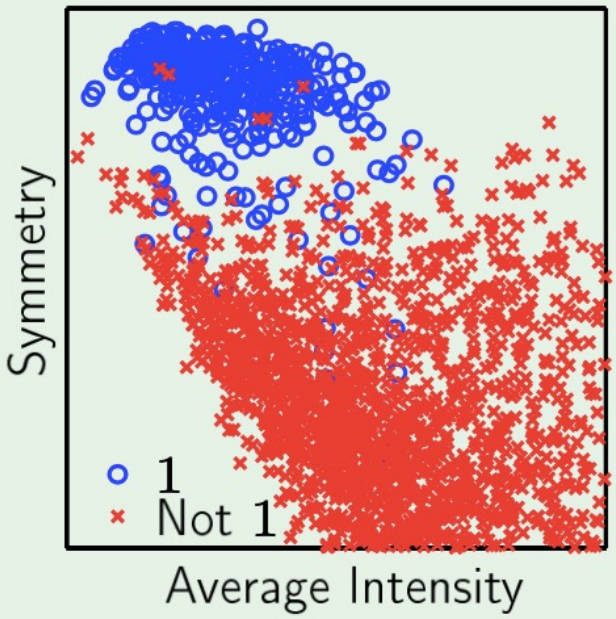
Stochastic noise



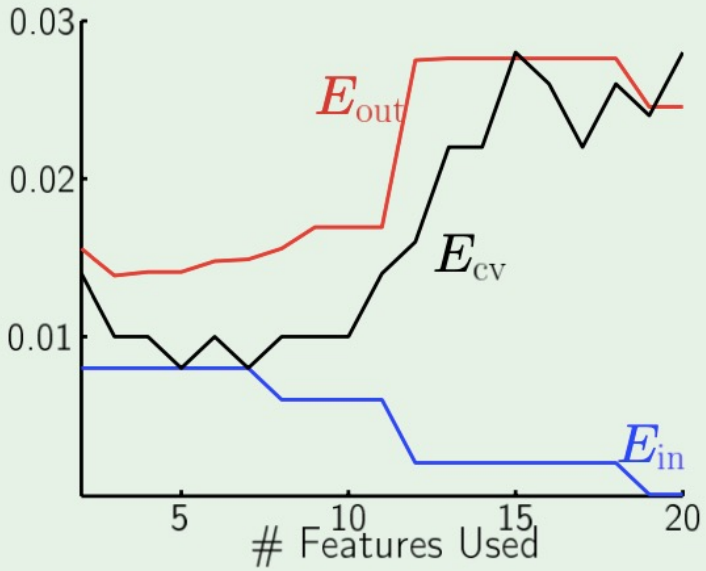
Deterministic noise

Cross Validation – Case Study

Digits classification task



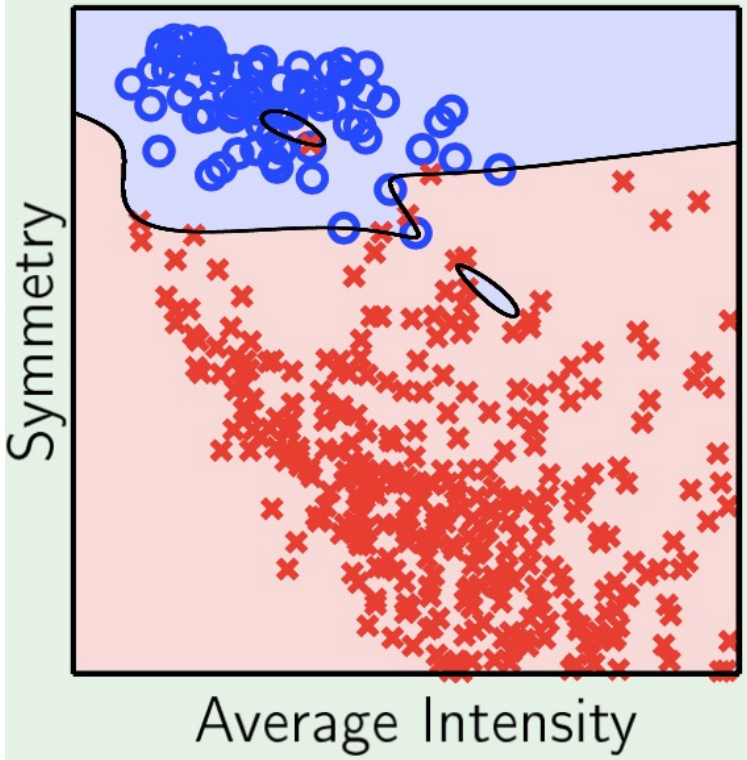
Different errors



$$(1, x_1, x_2) \rightarrow (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, \ldots, x_1^5, x_1^4x_2, x_1^3x_2^2, x_1^2x_2^3, x_1x_2^4, x_2^5)$$

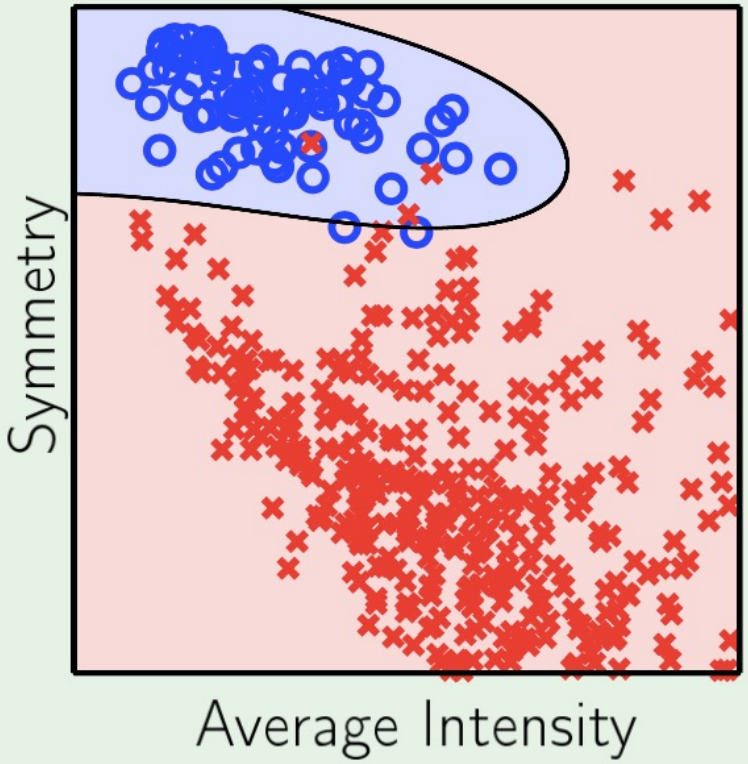
Cross Validation – Case Study

without validation



$E_{\text{in}} = 0\%$ $E_{\text{out}} = 2.5\%$

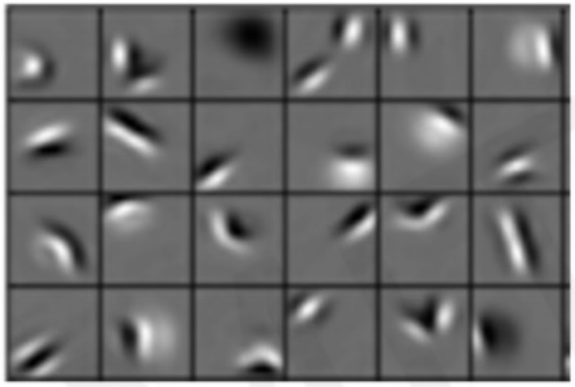
with validation



$E_{\text{in}} = 0.8\%$ $E_{\text{out}} = 1.5\%$

ConvNet: Feature Representation Hierarchy

Low level features



Edges, dark spots

Mid level features

Eyes, ears, nose

High level features

Facial structure

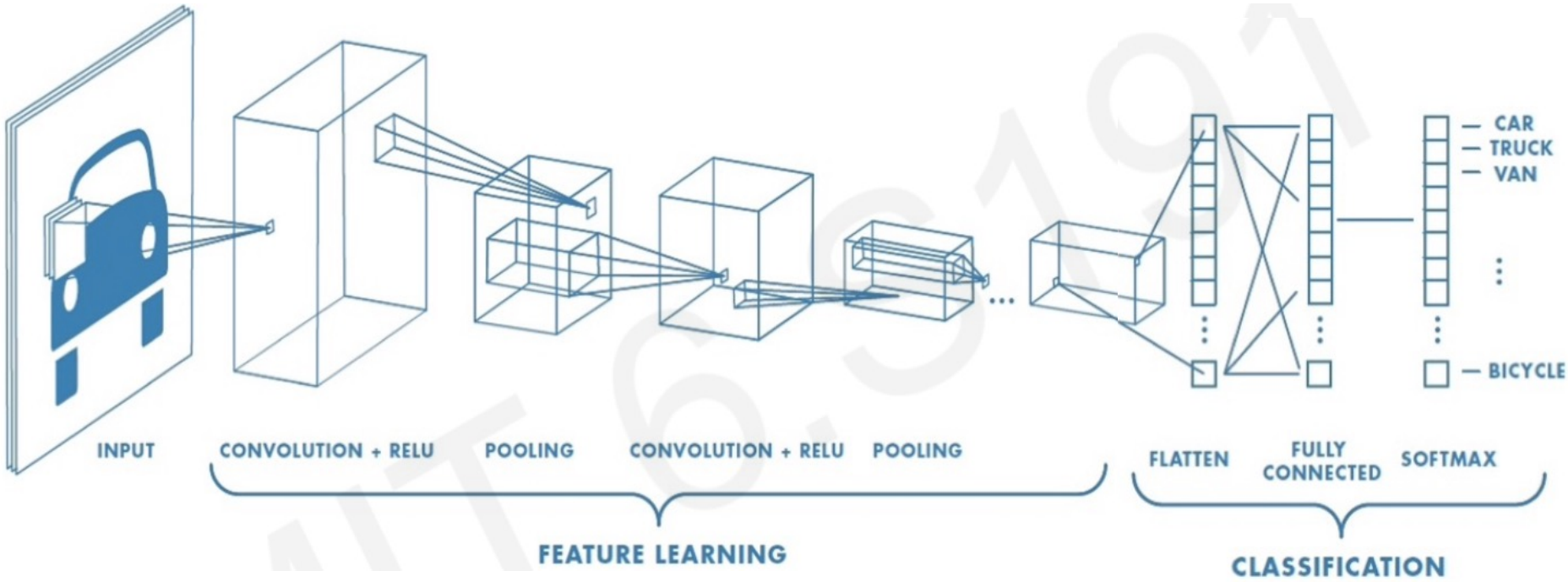
Original

Sharpen

Edge Detect

"Strong" Edge Detect

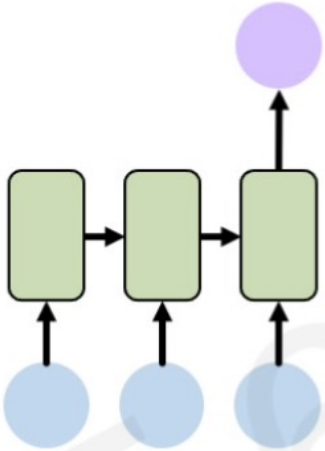
CNN Architecture



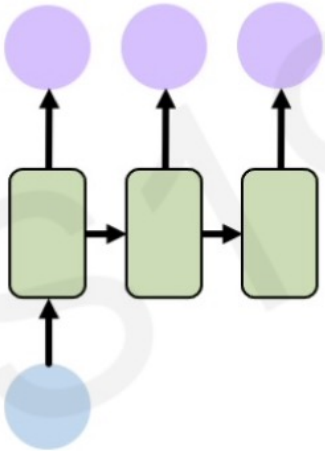
Sequence Modeling Applications

One to One
Binary Classification

"Will I pass this class?"
Student → Pass?

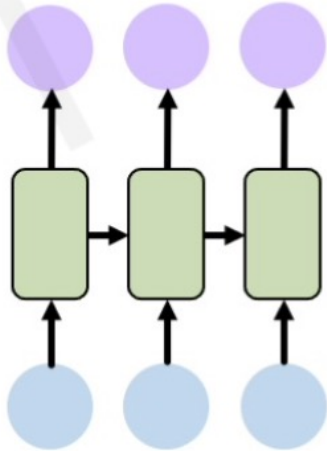


Many to One
Sentiment Classification



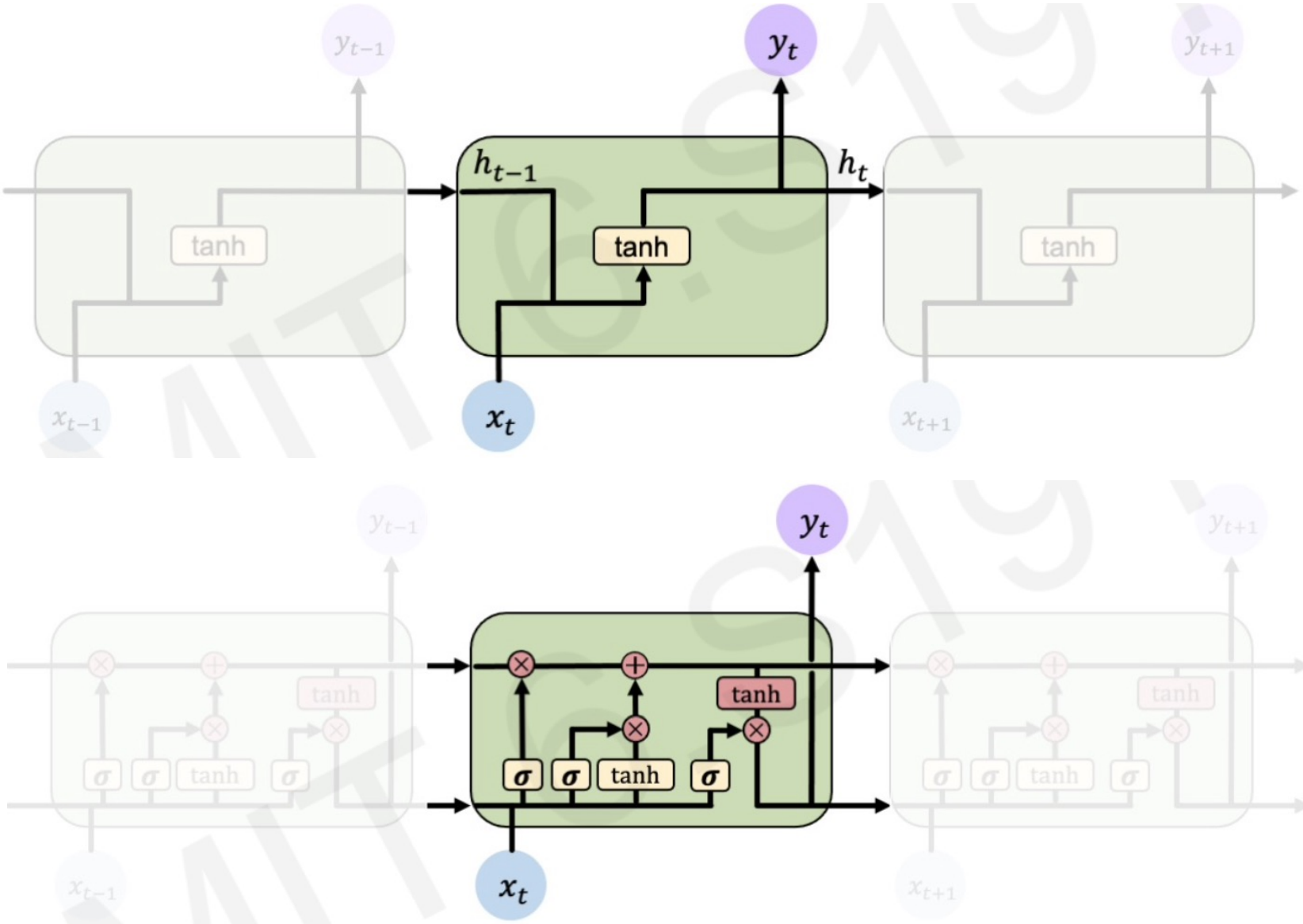
One to Many
Image Captioning

"A baseball player throws a ball."

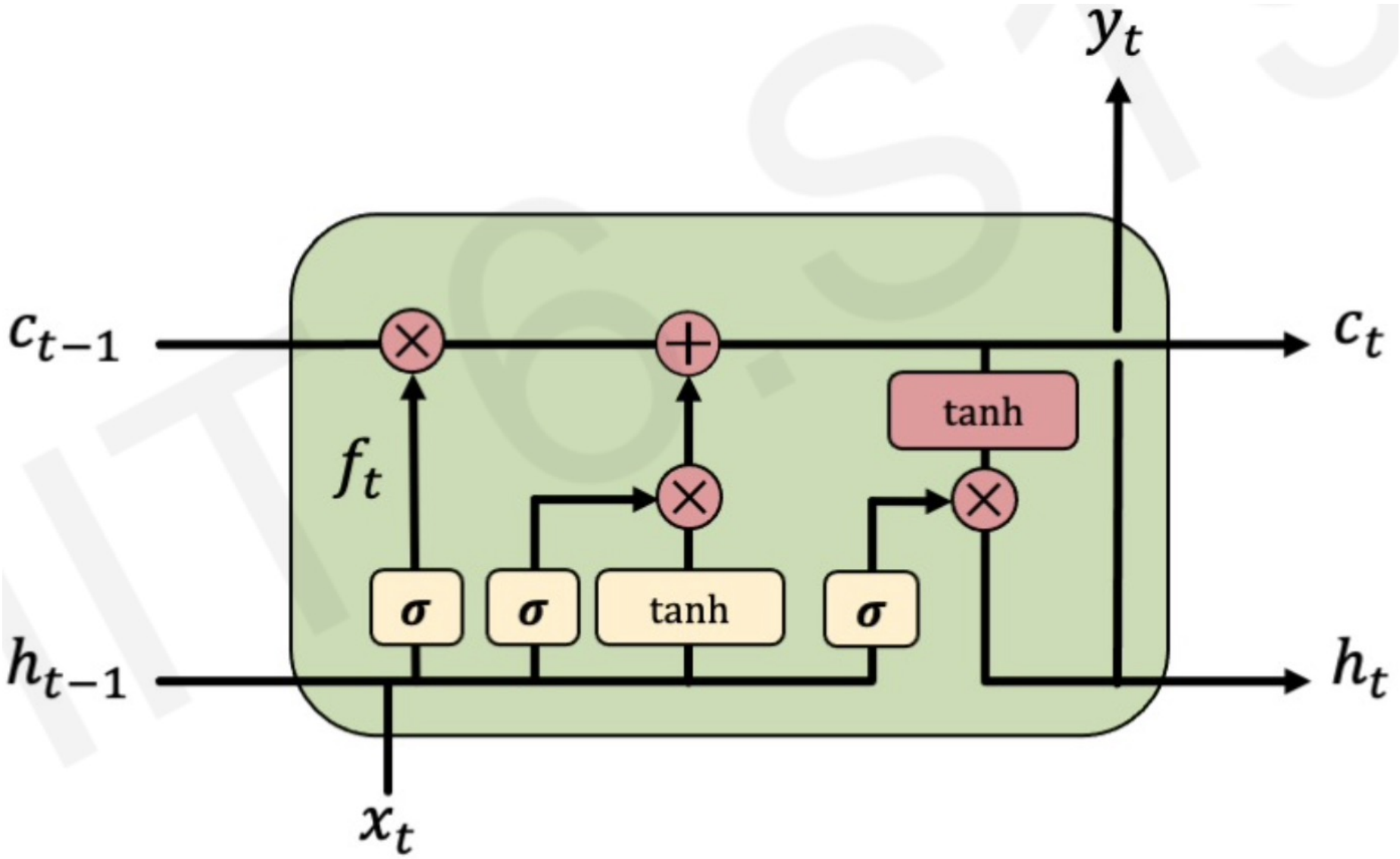


Many to Many
Machine Translation

RNN vs. LSTM


y_{t-1}
y_t
y_{t+1}
h_{t-1}
h_t
tanh
tanh
tanh
x_{t-1}
x_t
x_{t+1}
y_{t-1}
y_t
y_{t+1}
tanh
tanh
σ
σ
tanh
σ
σ
σ
tanh
σ
tanh
x_{t-1}
x_t
x_{t+1}

Long Short Term Memory (LSTM)



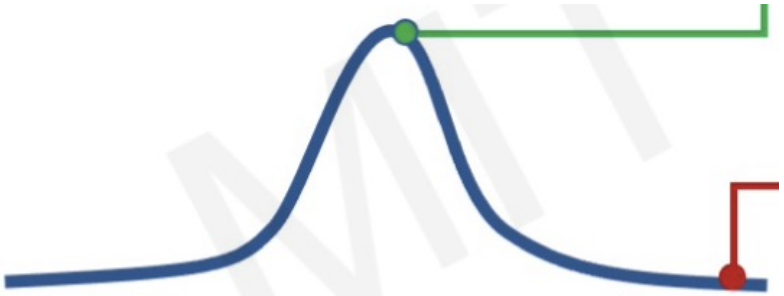
Generative Modeling Example

Homogeneous skin color, pose

VS

Diverse skin color, pose, illumination

Why Generative Models?



95% of Driving Data:
(1) sunny, (2) highway, (3) straight road

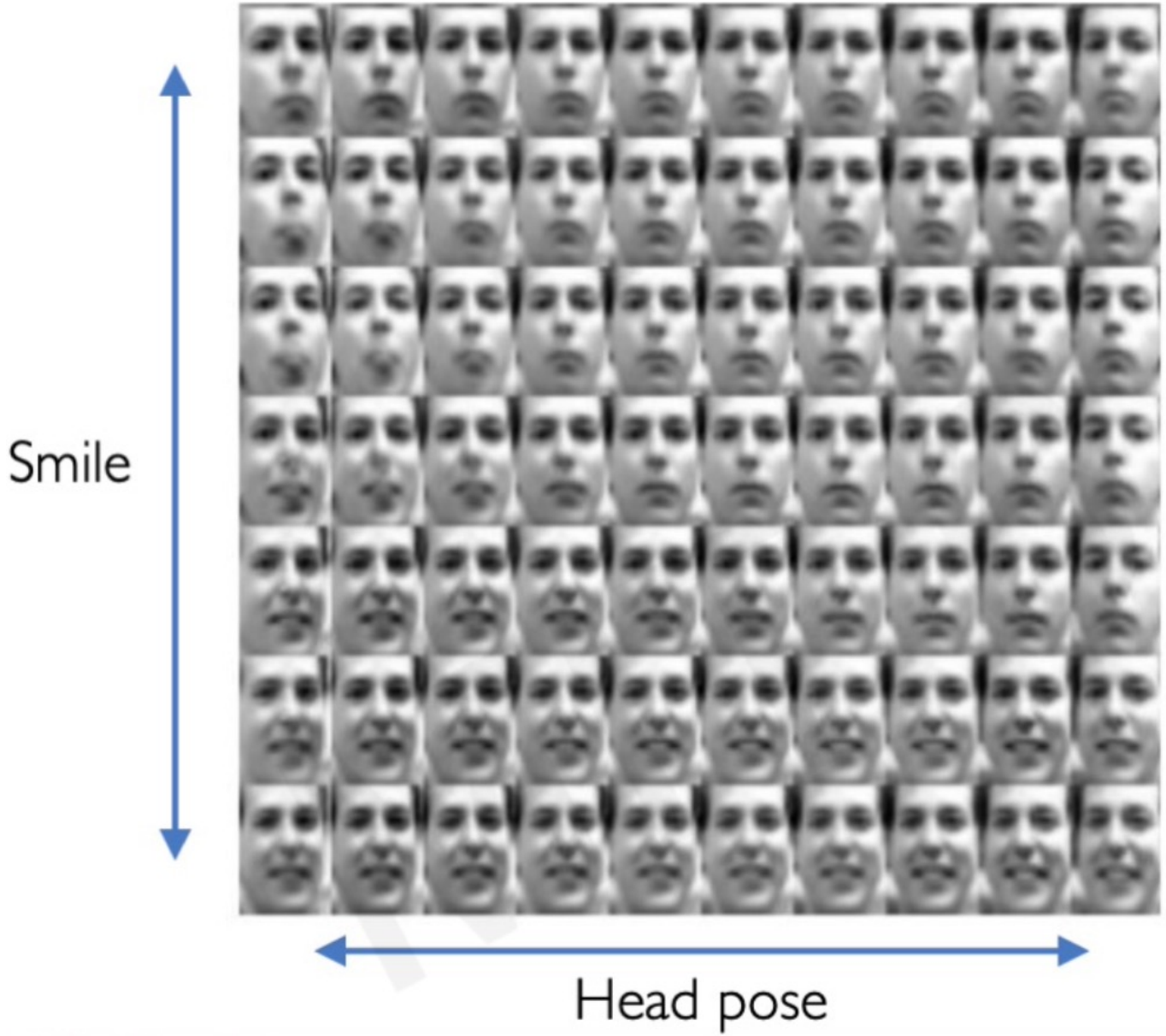
Detect outliers to avoid unpredictable behavior when training

Edge Cases

Harsh Weather

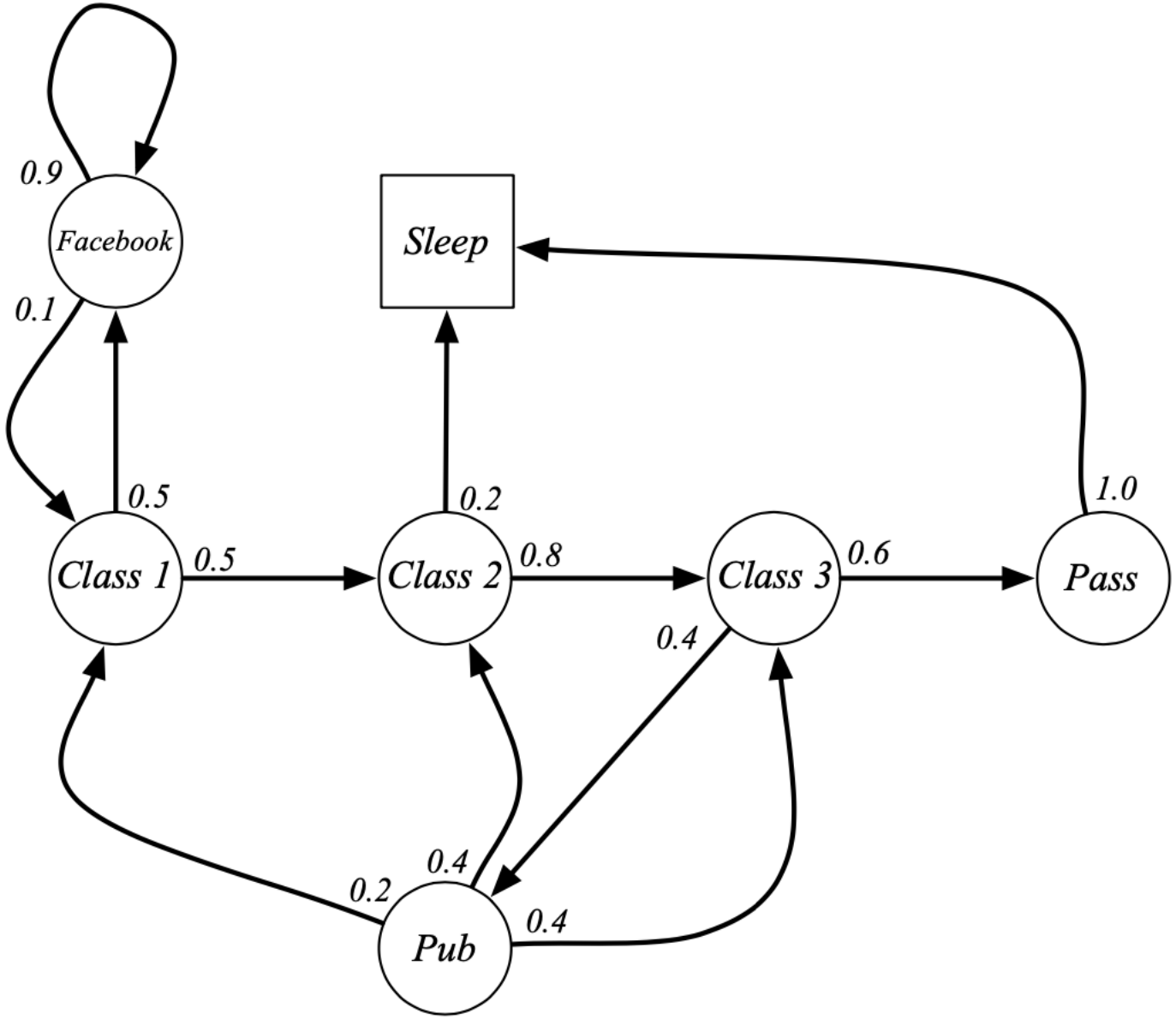
Pedestrians

Variational Auto-Encoders (VAE)

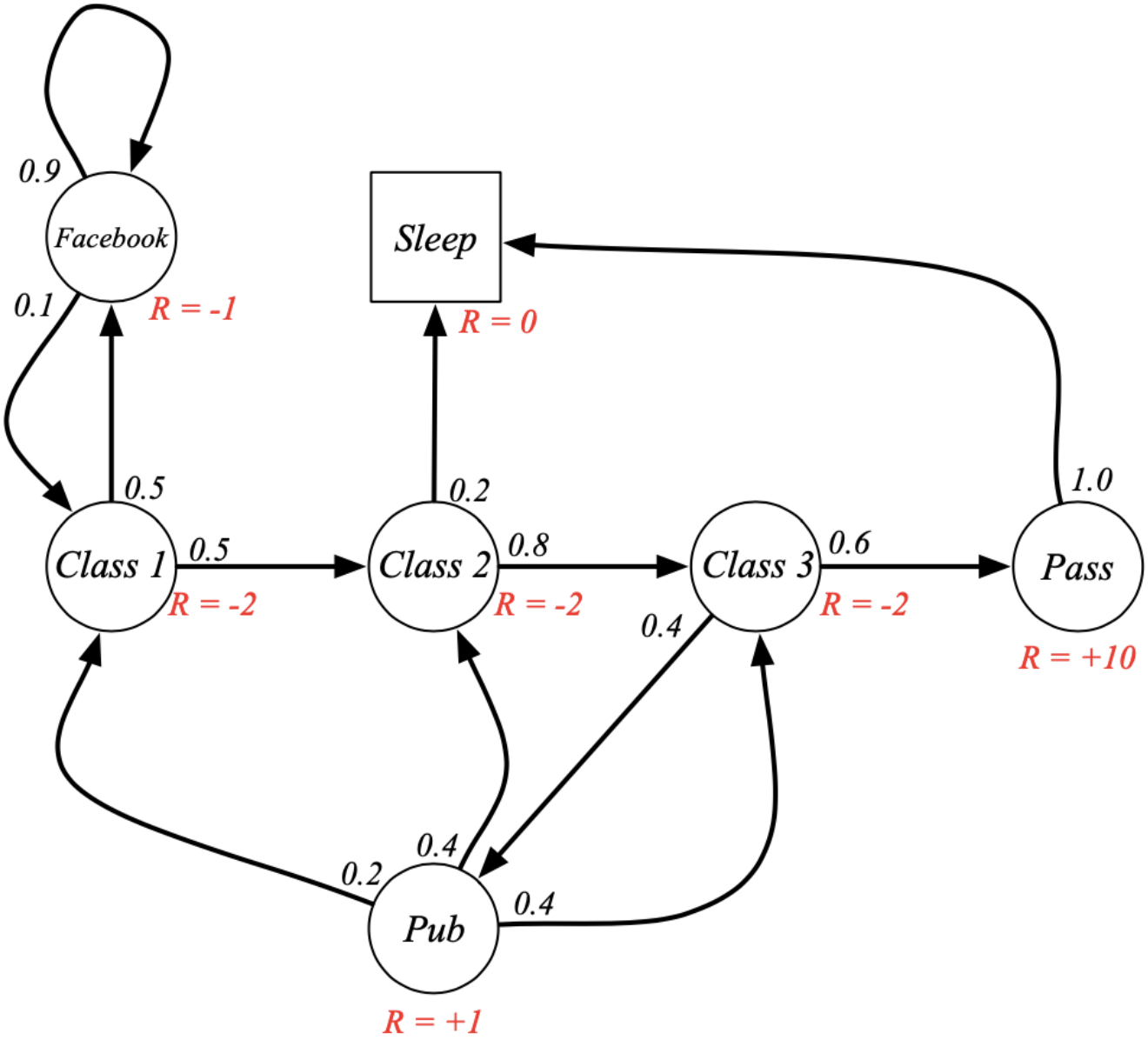


Generative Adversarial Networks (GAN)

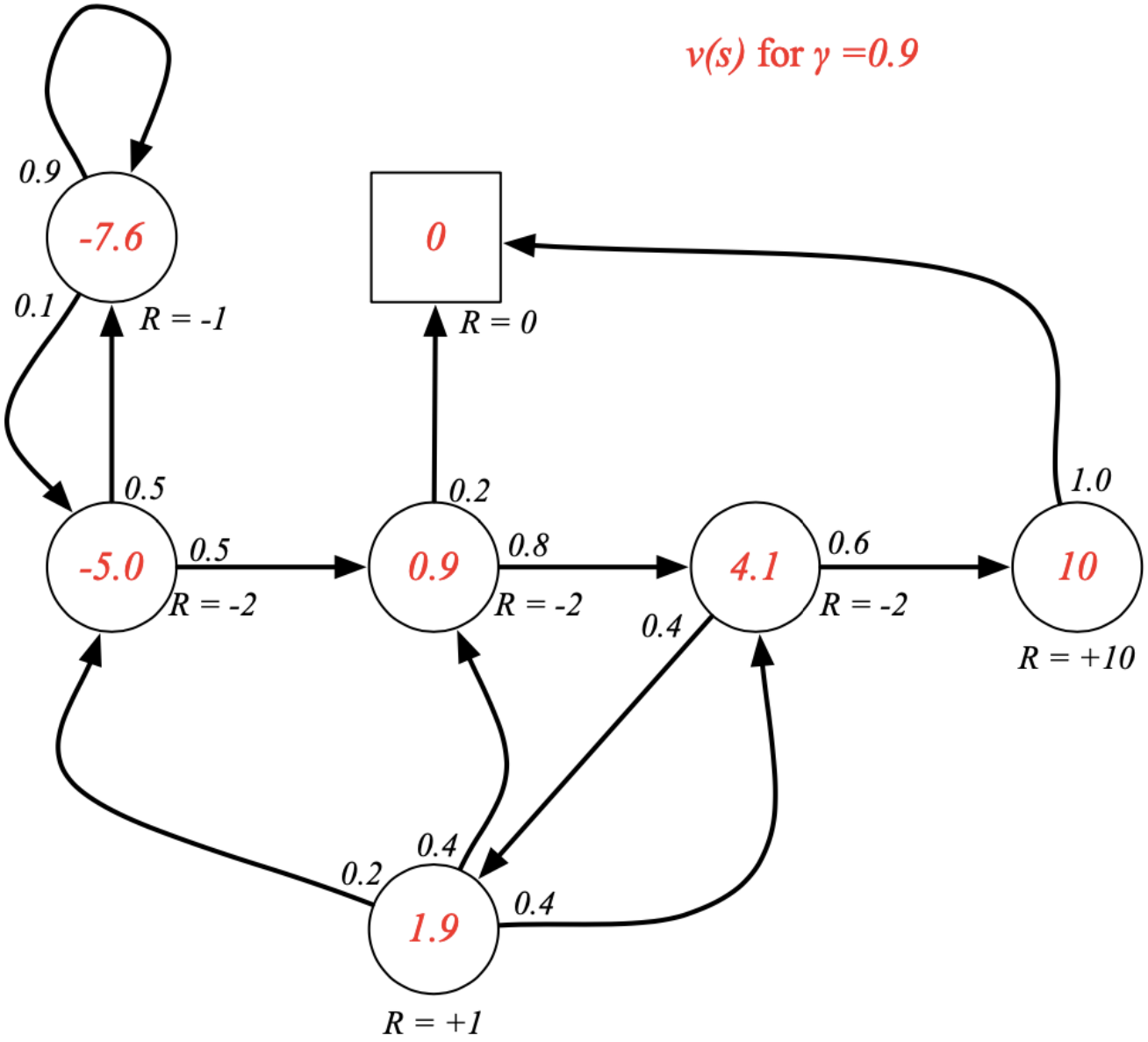
Markov Process/Chain – Example



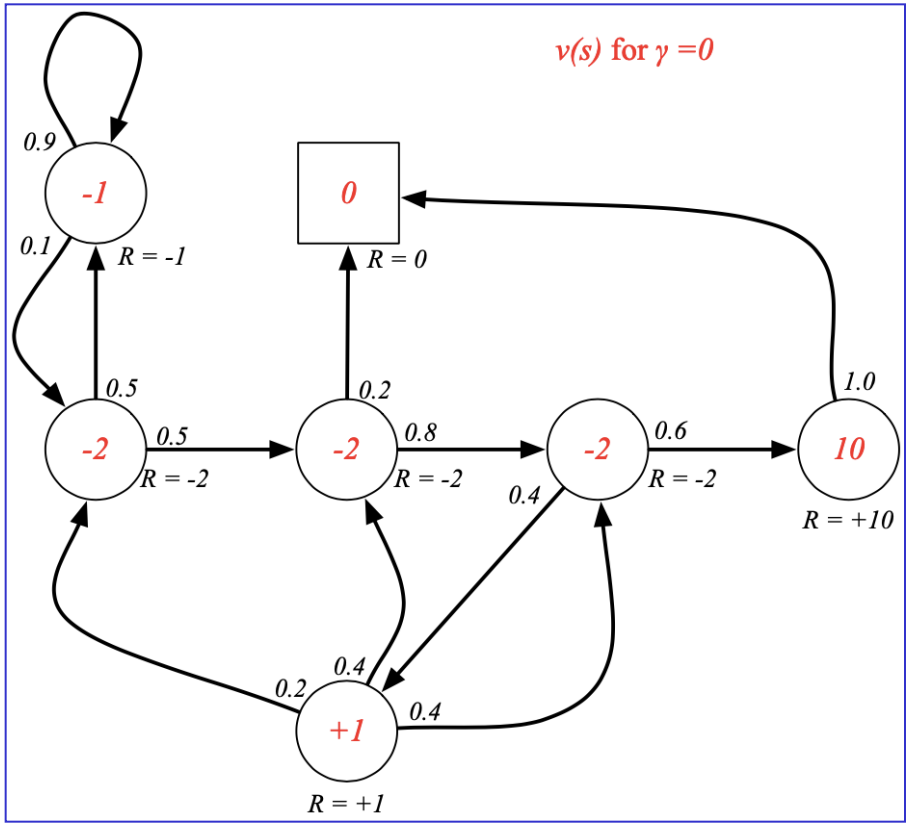
Markov Reward Process – Example

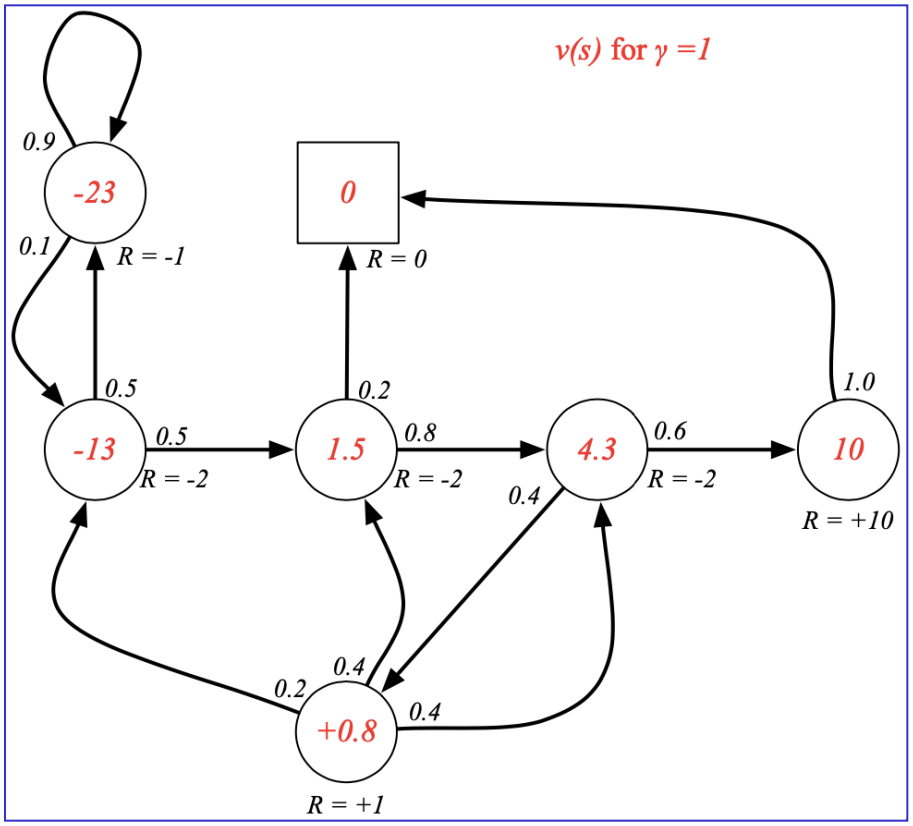


Markov Reward Process – Value Function

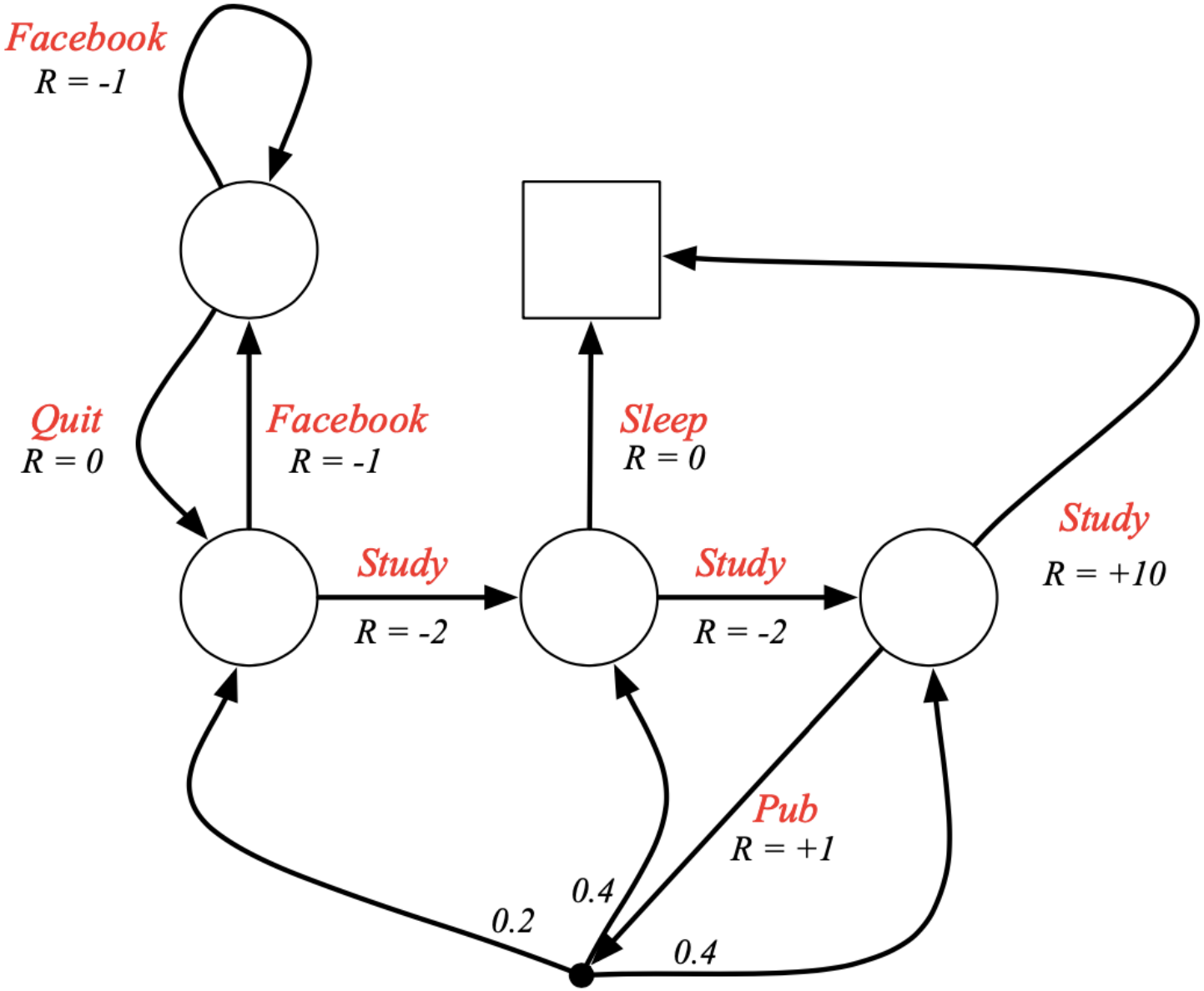


Markov Reward Process – Value Function

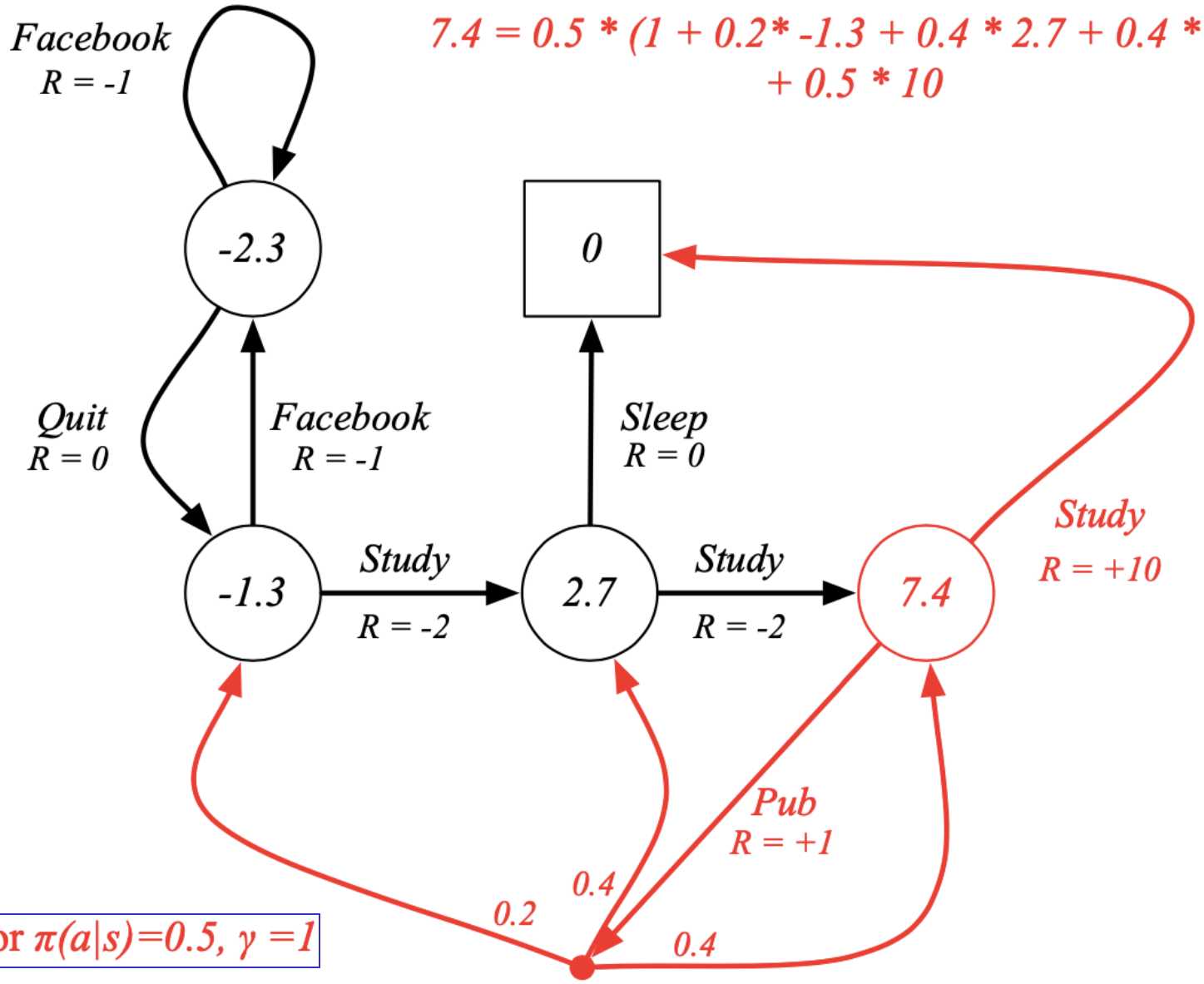




Markov Decision Process – Example

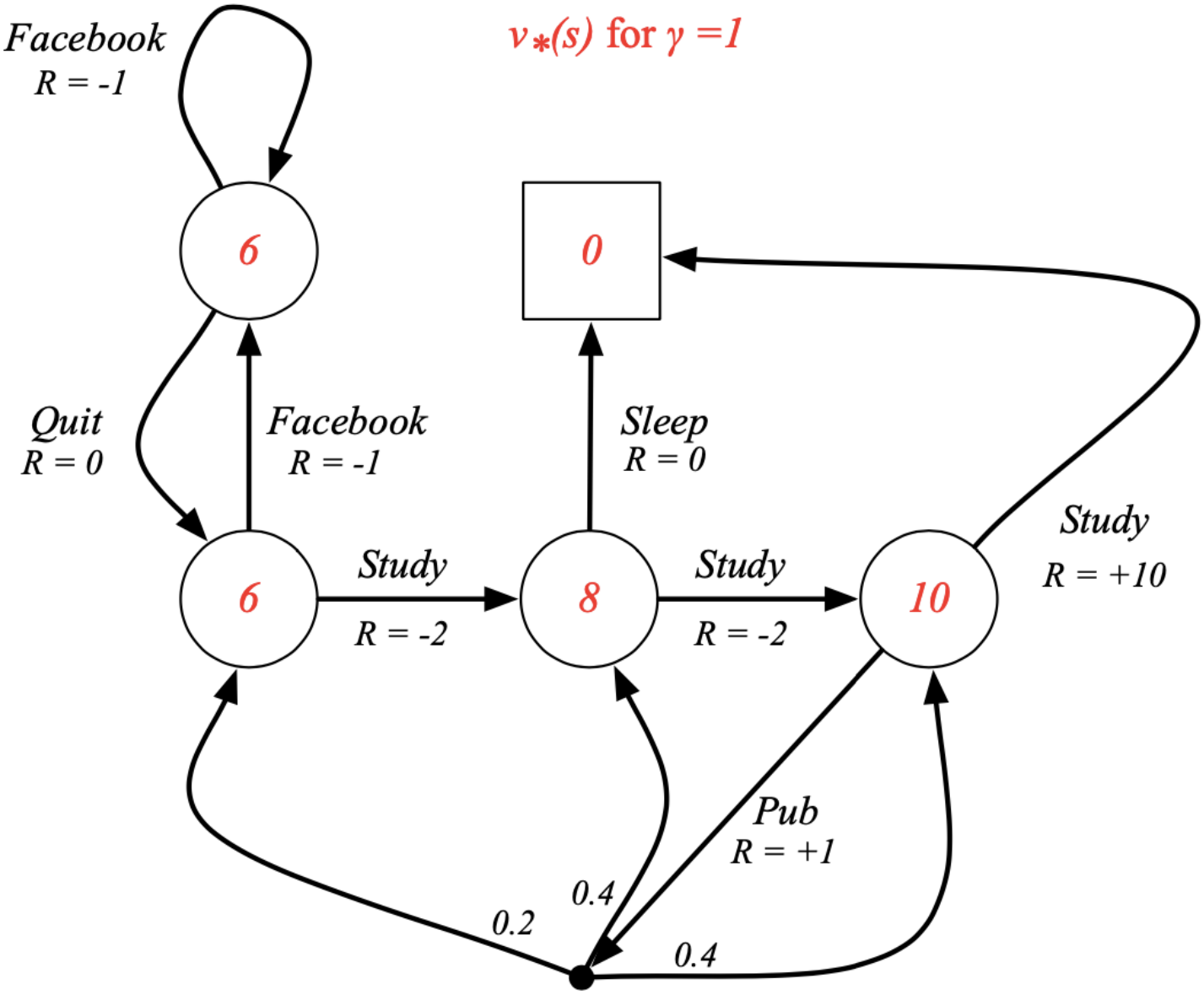


Markov Decision Process – Value Function

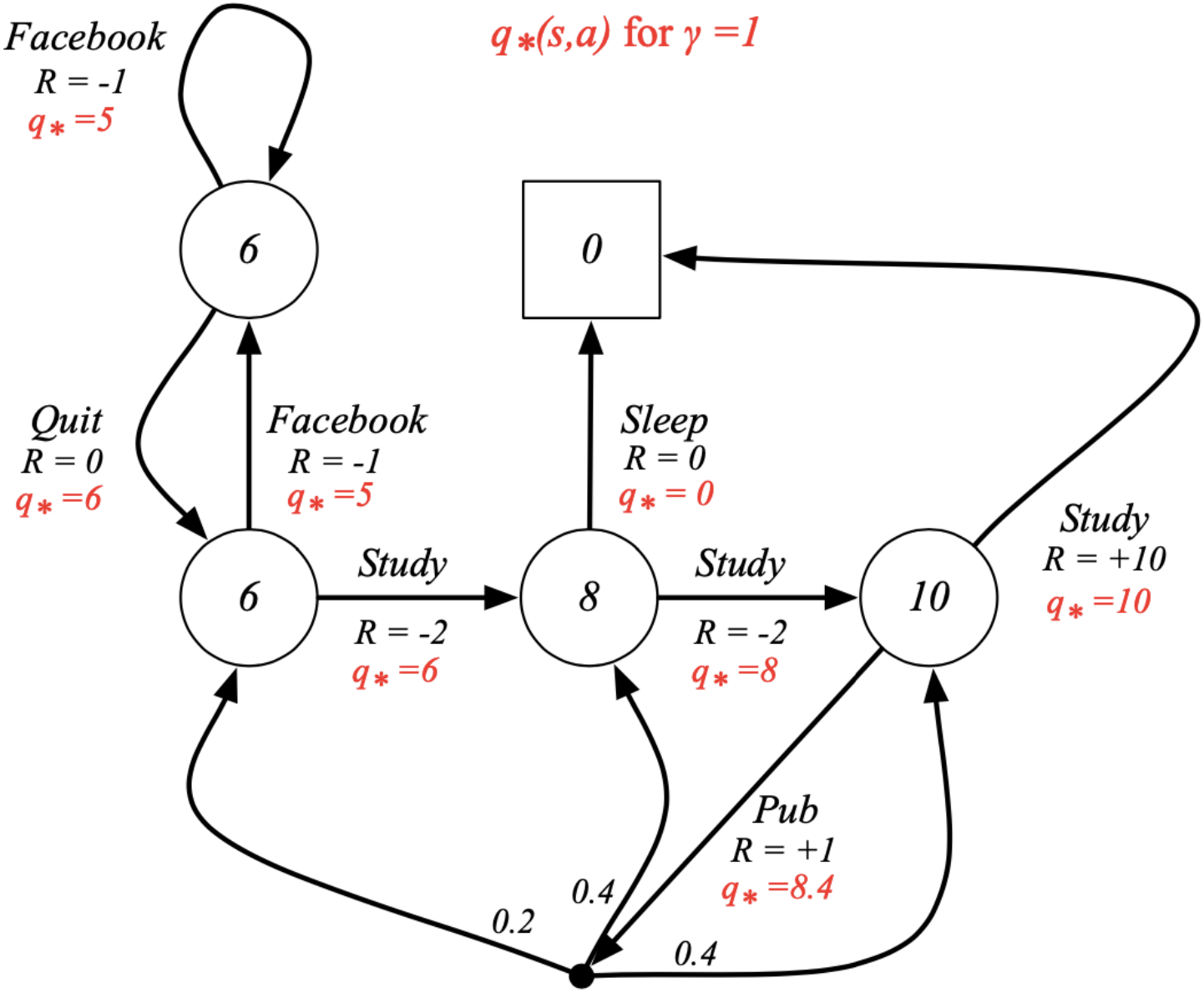


$v\pi(s)$ for $\pi(a|s)=0.5$, $\gamma = 1$

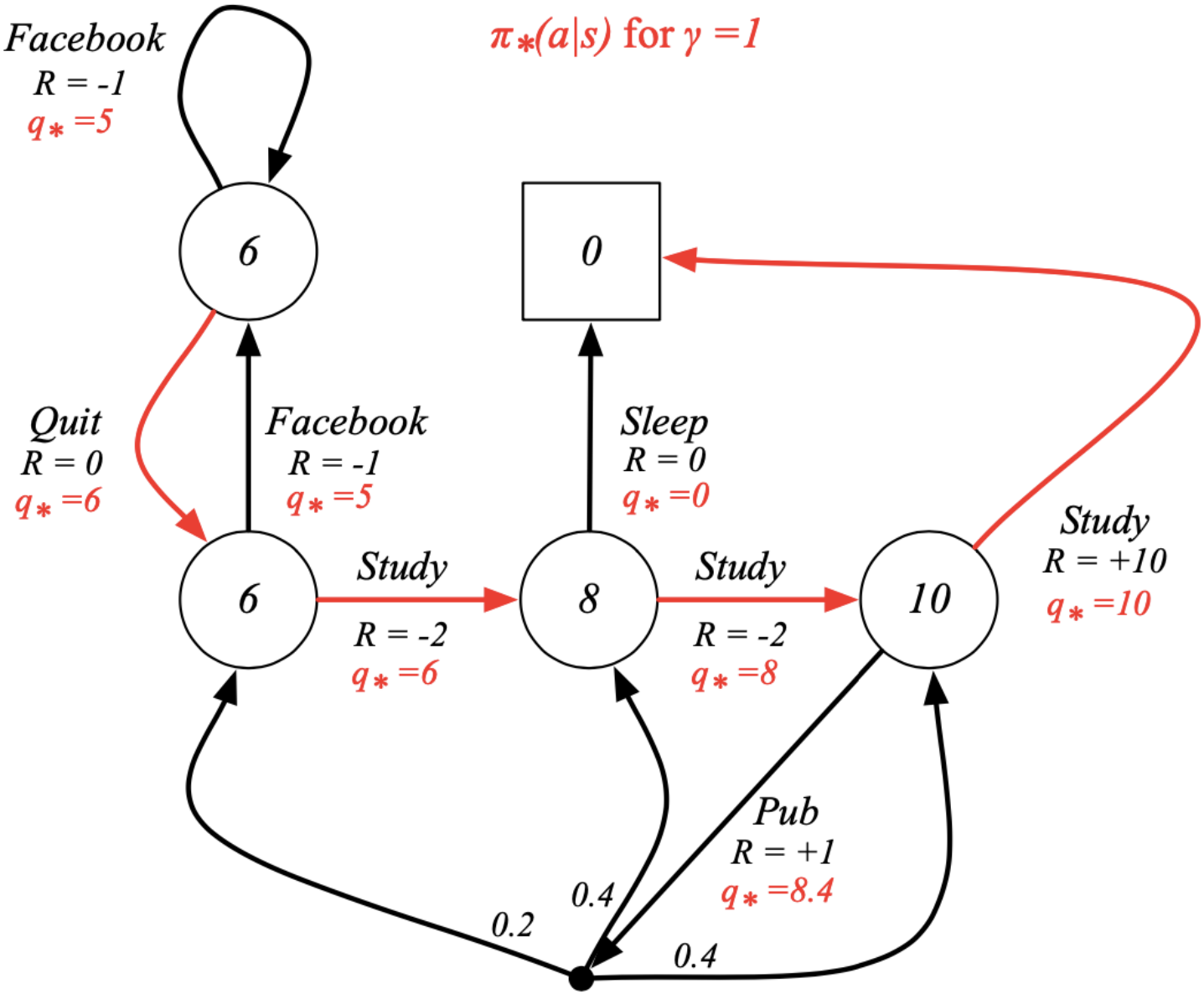
Markov Decision Process – Optimal Value Fn



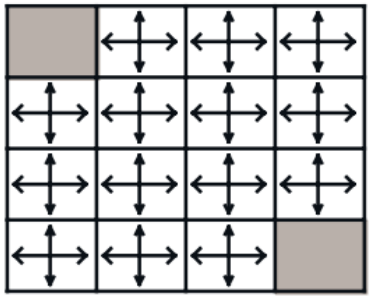
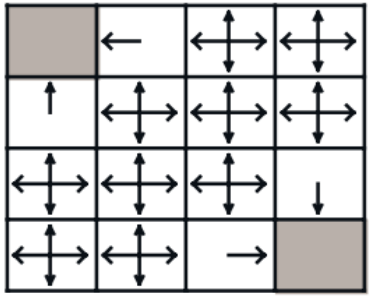
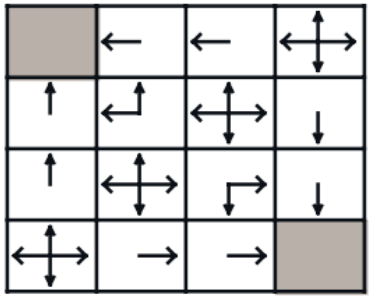
Markov Decision Process – Optimal Value Fn

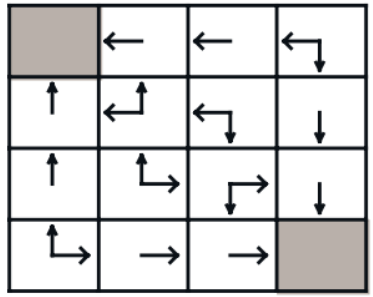
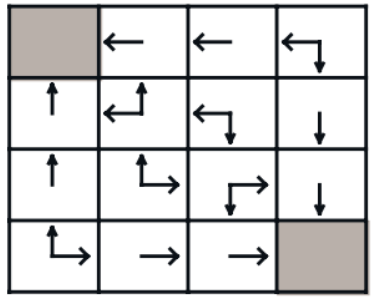
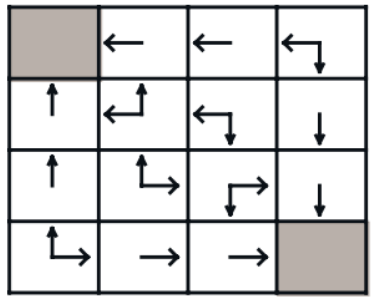


Markov Decision Process – Optimal Policy



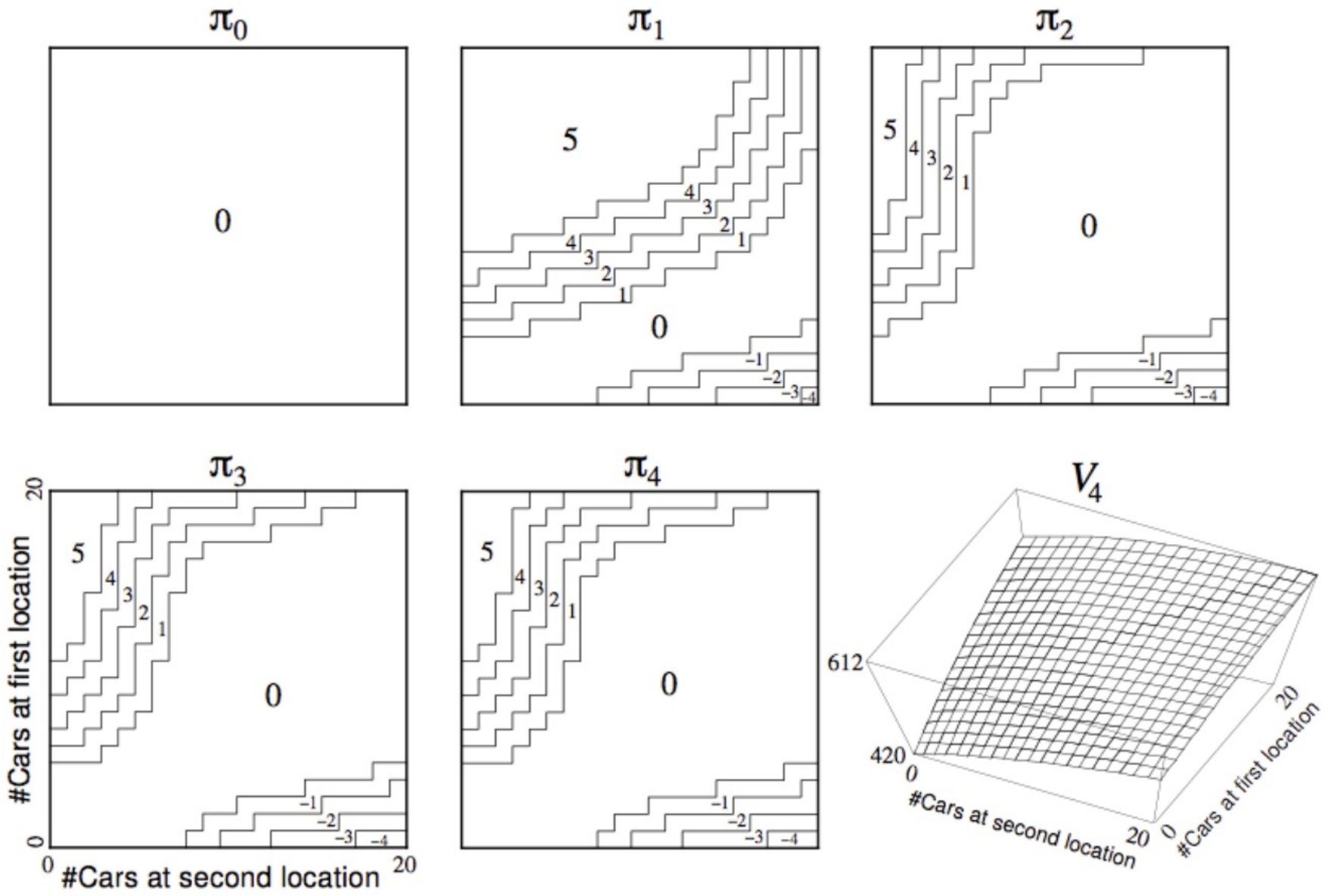
Policy Evaluation – Example

v_k for the Random Policy					Greedy Policy w.r.t. v_k
	0.0	0.0	0.0	0.0	
	0.0	0.0	0.0	0.0	
	0.0	0.0	0.0	0.0	
	0.0	0.0	0.0	0.0	
k = 0					
	0.0	-1.0	-1.0	-1.0	
	-1.0	-1.0	-1.0	-1.0	
	-1.0	-1.0	-1.0	-1.0	
	-1.0	-1.0	-1.0	0.0	
k = 1					
	0.0	-1.7	-2.0	-2.0	
	-1.7	-2.0	-2.0	-2.0	
	-2.0	-2.0	-2.0	-1.7	
	-2.0	-2.0	-1.7	0.0	
k = 2					

0.0	-2.4	-2.9	-3.0	
-2.4	-2.9	-3.0	-2.9	
-2.9	-3.0	-2.9	-2.4	
-3.0	-2.9	-2.4	0.0	
k = 3				
0.0	-6.1	-8.4	-9.0	
-6.1	-7.7	-8.4	-8.4	
-8.4	-8.4	-7.7	-6.1	
-9.0	-8.4	-6.1	0.0	
k = 10				
0.0	-14.	-20.	-22.	
-14.	-18.	-20.	-20.	
-20.	-20.	-18.	-14.	
-22.	-20.	-14.	0.0	
k = ∞				

optimal policy

Policy Iteration – Example (Jack's Car Rental)



Value Iteration – Example

g			

Problem

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

V_1

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1

V_2

0	-1	-2	-2
-1	-2	-2	-2
-2	-2	-2	-2
-2	-2	-2	-2

V_3

0	-1	-2	-3
-1	-2	-3	-3
-2	-3	-3	-3
-3	-3	-3	-3

V_4

0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-4
-3	-4	-4	-4

V_5

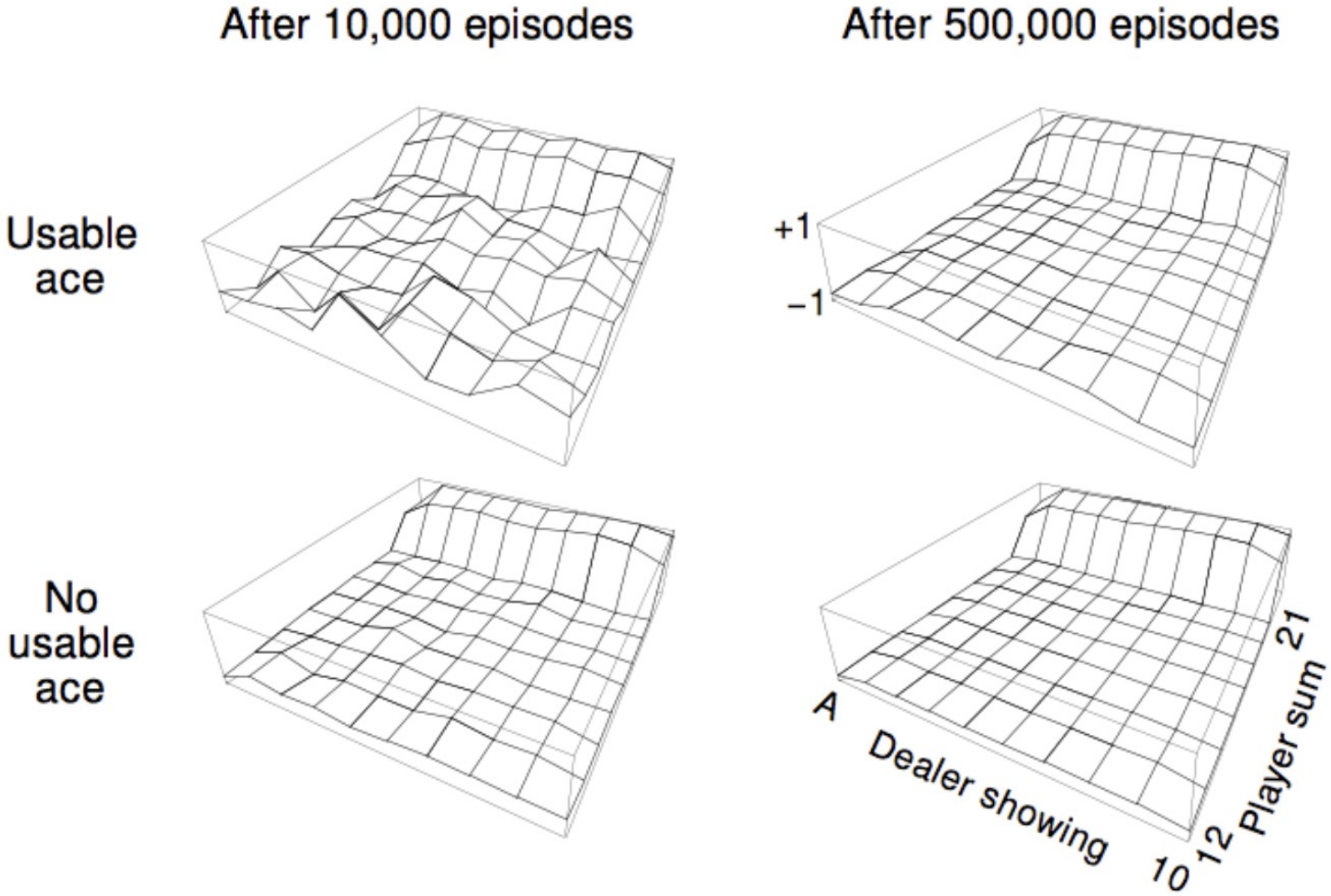
0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-5
-3	-4	-5	-5

V_6

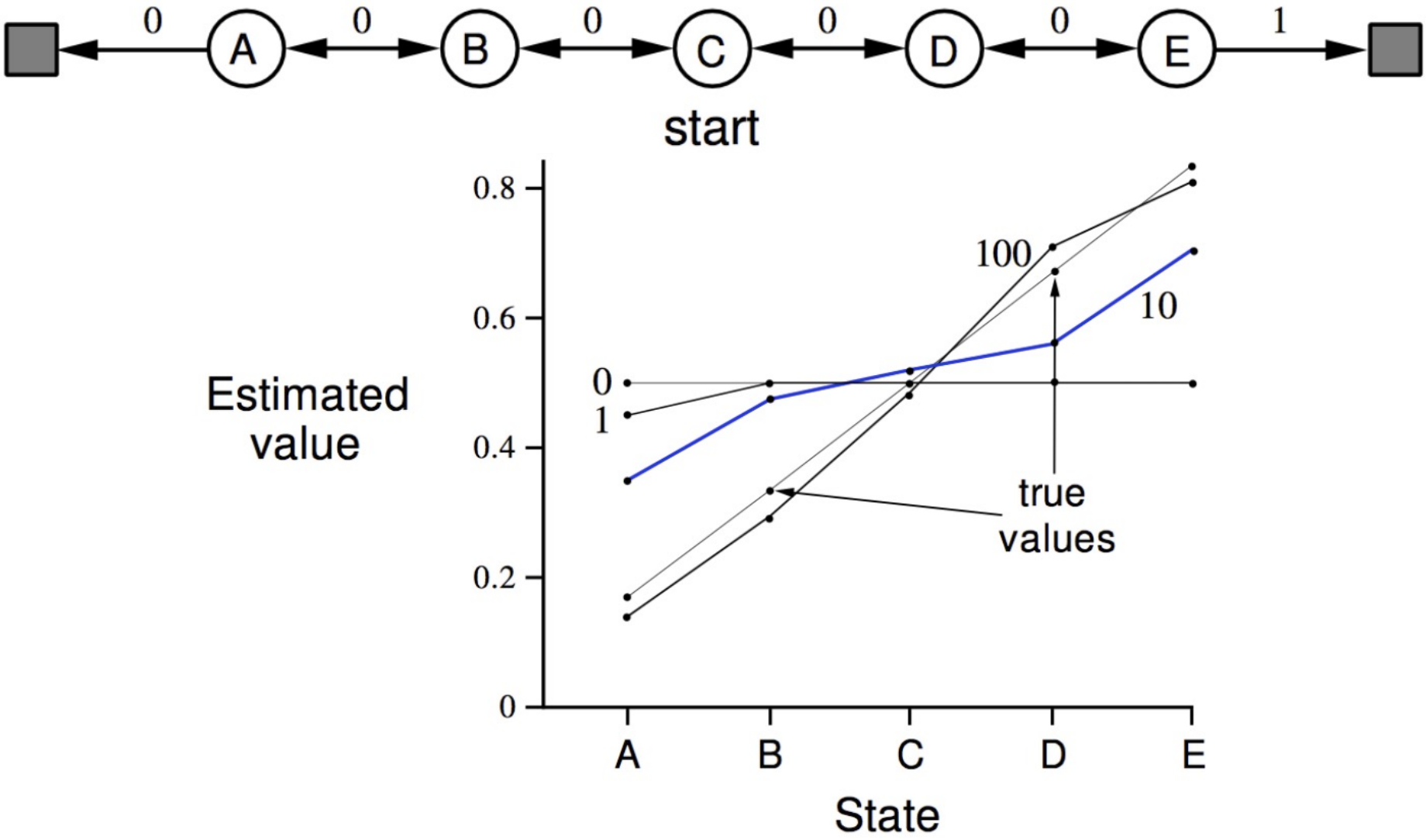
0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-5
-3	-4	-5	-6

V_7

Model Free Prediction – MC Example



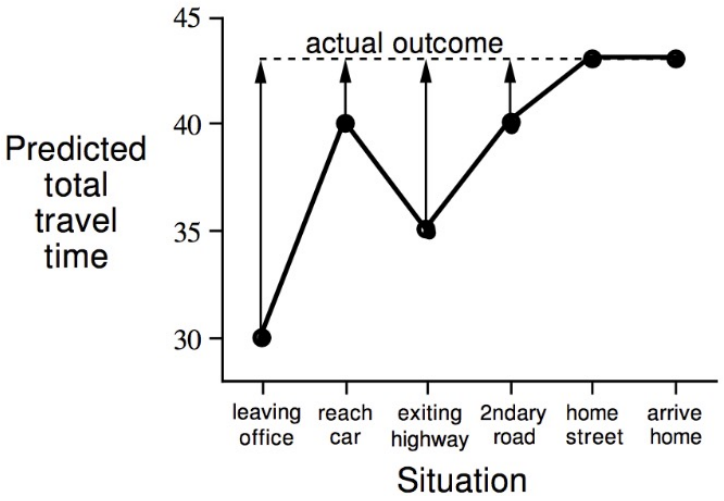
Model Free Prediction – TD Example



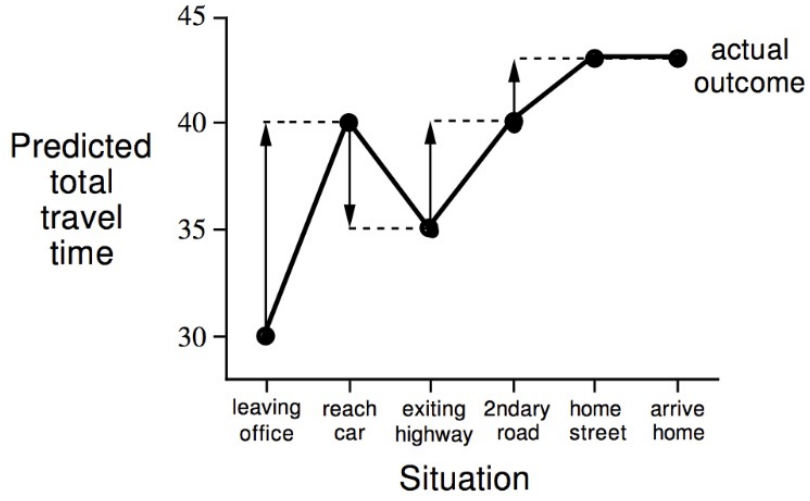
MC vs. TD – Driving Example

State	Elapsed Time (minutes)	Predicted Time to Go	Predicted Total Time
leaving office	0	30	30
reach car, raining	5	35	40
exit highway	20	15	35
behind truck	30	10	40
home street	40	3	43
arrive home	43	0	43

Changes recommended by Monte Carlo methods (α=1)



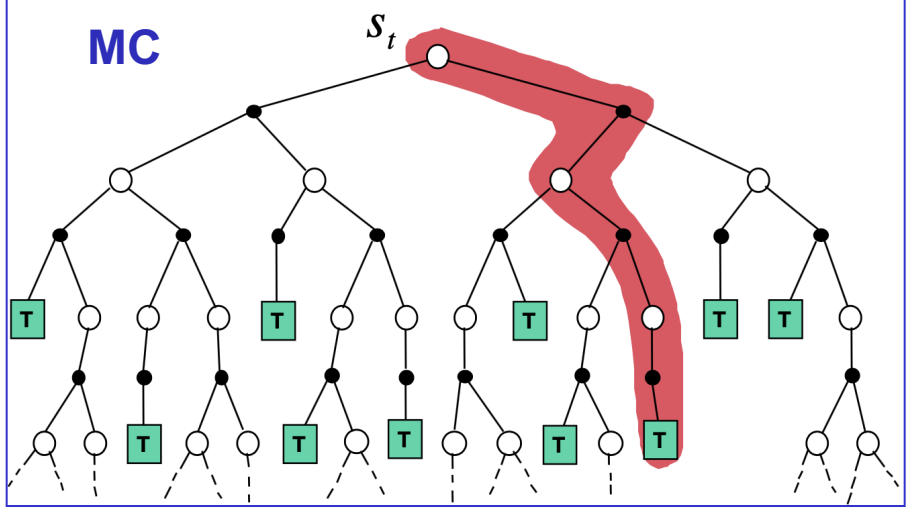
Changes recommended by TD methods (α=1)



Backup Principles

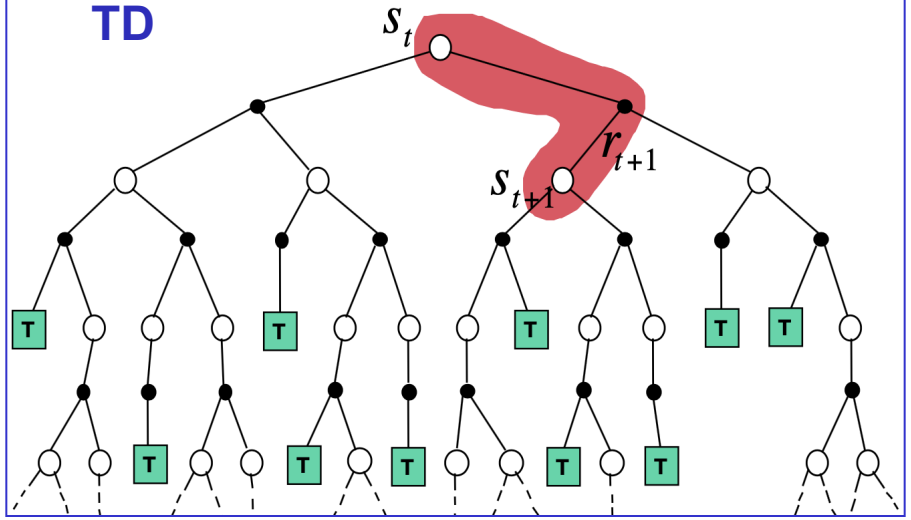
$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right)$$

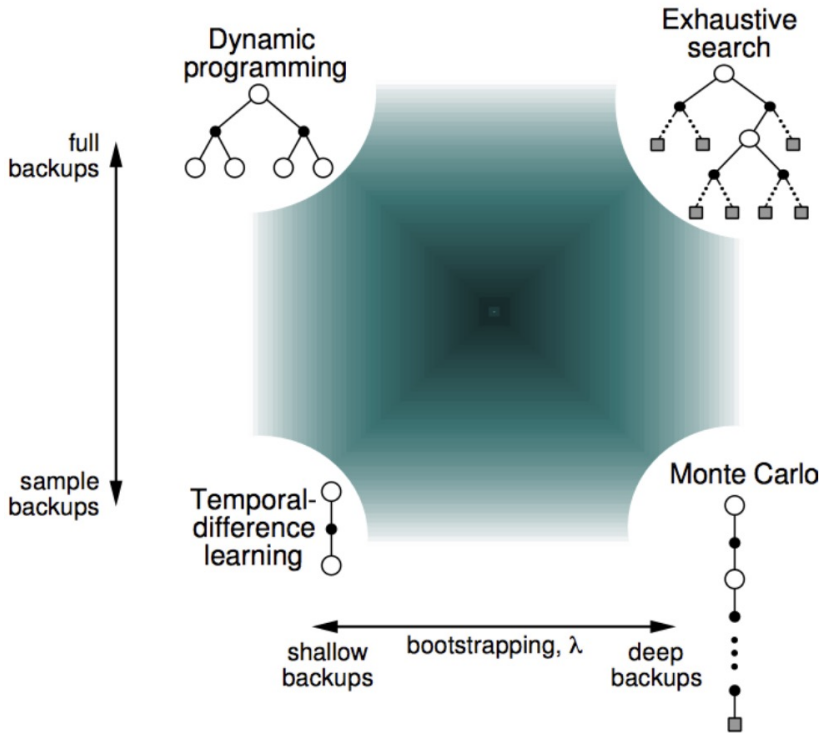
MC



$$V(S_t) \leftarrow V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$$

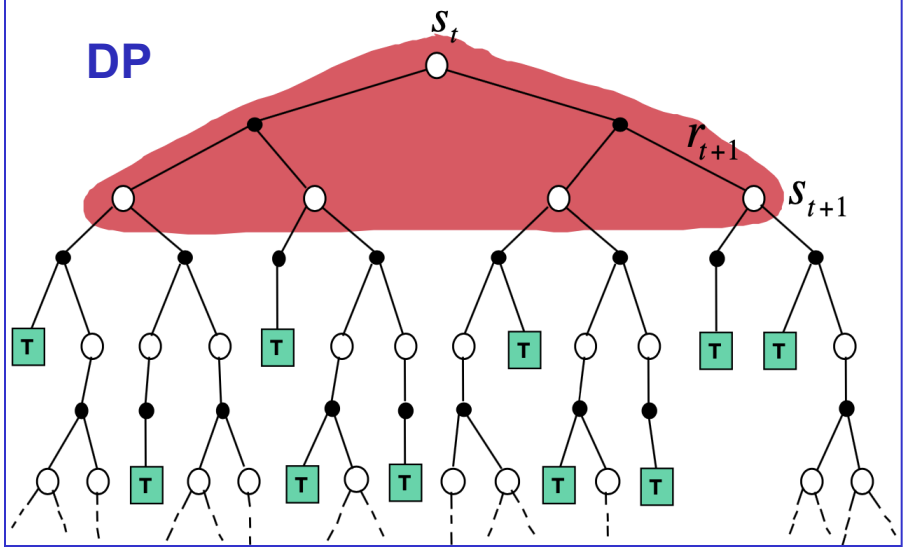
TD



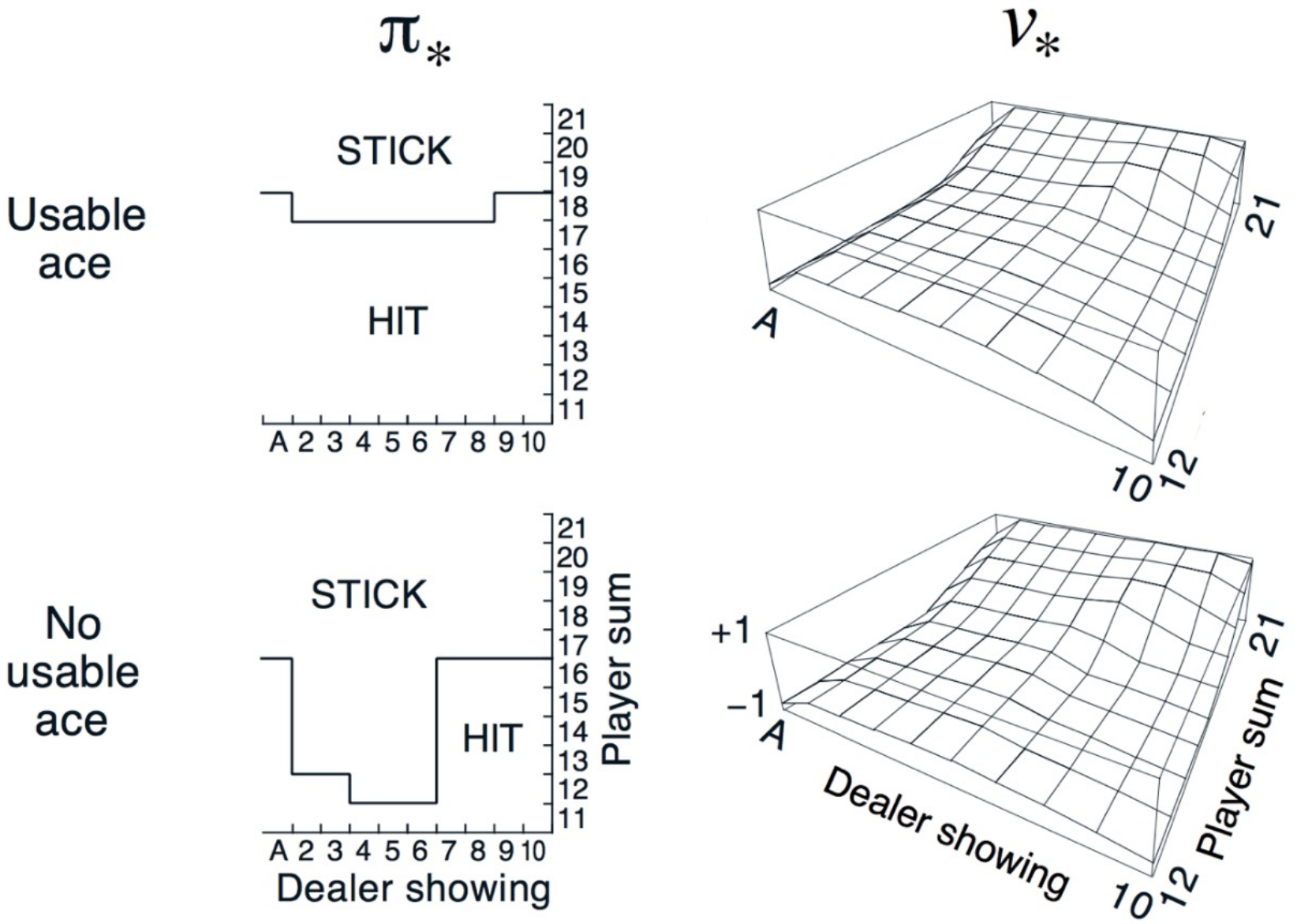


$$V(S_t) \leftarrow \mathbb{E}_\pi\left[R_{t+1} + \gamma V(S_{t+1})\right]$$

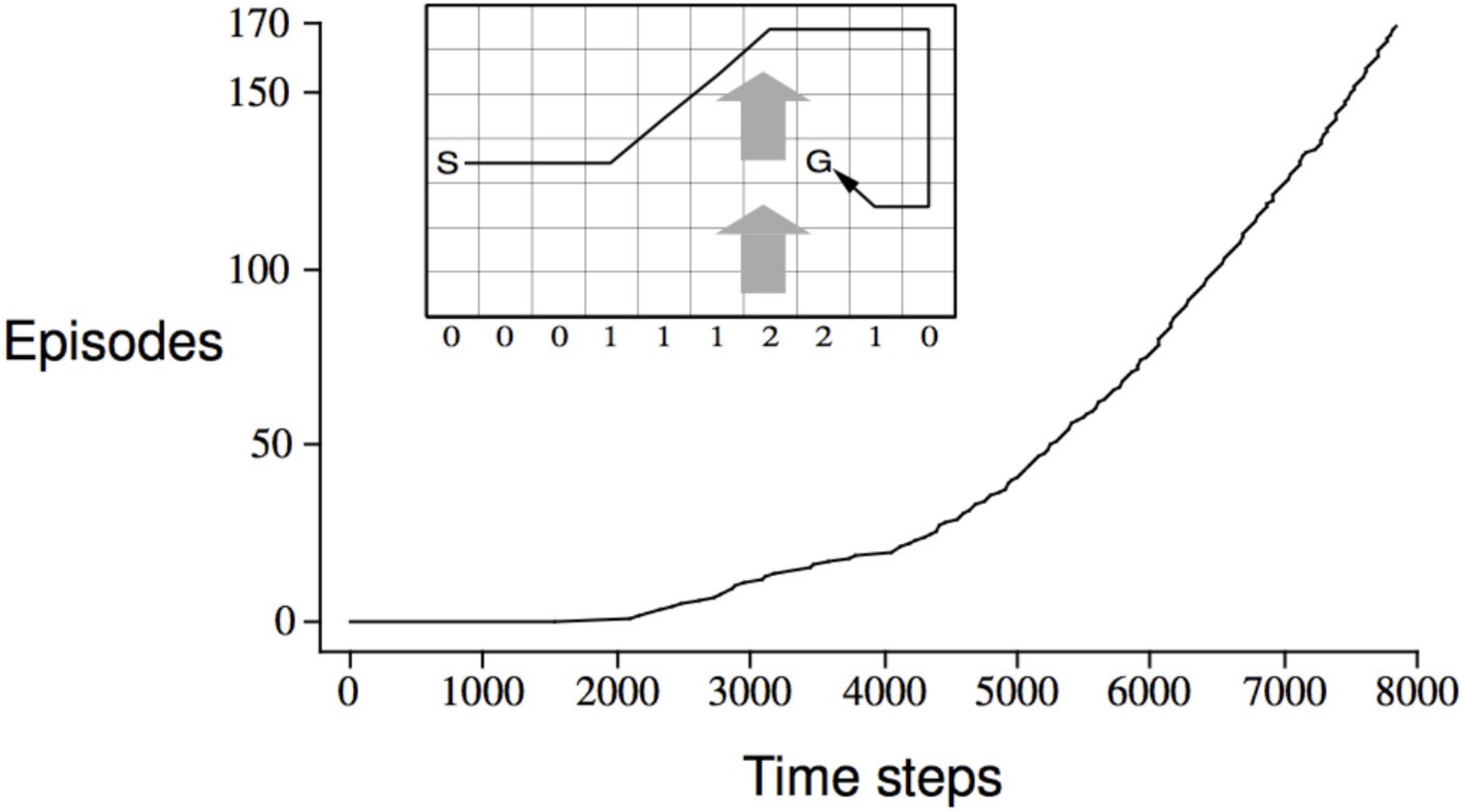
DP



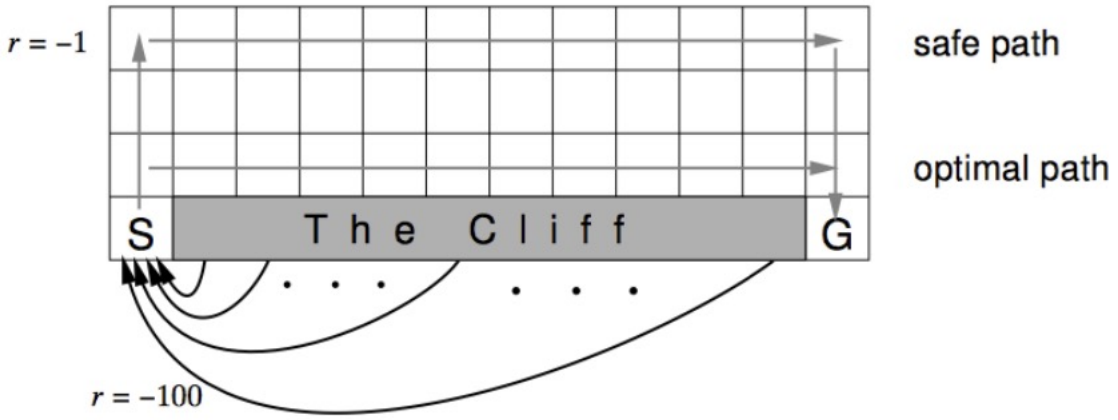
Model Free Control – MC Example

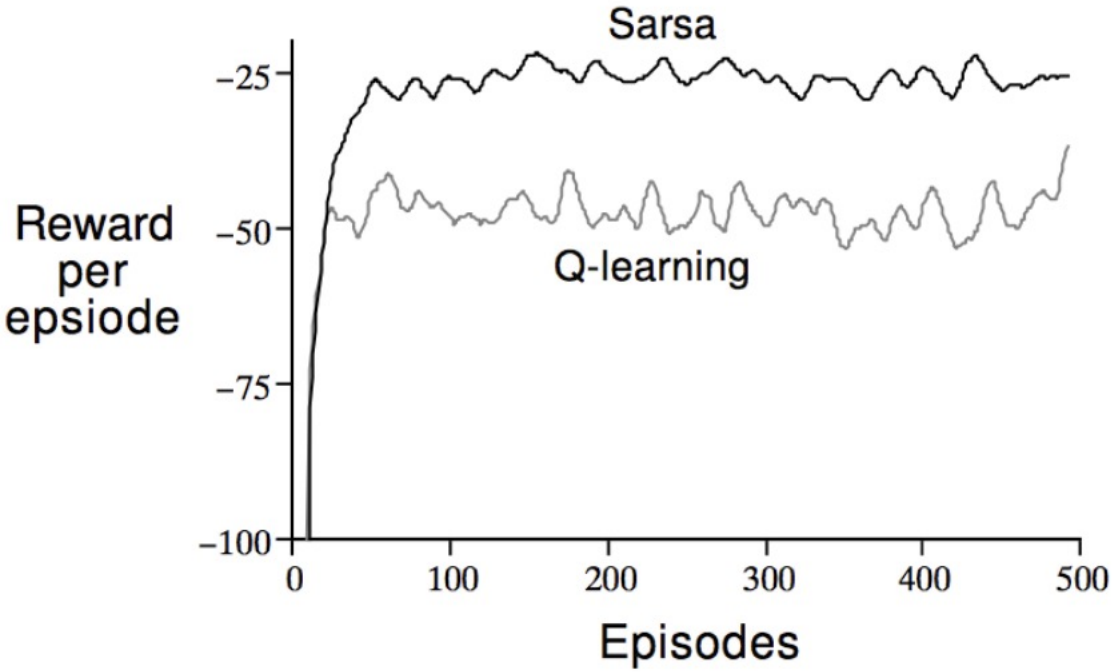


Model Free Control – SARSA Example



Model Free Control – SARSA vs. Q-Learning

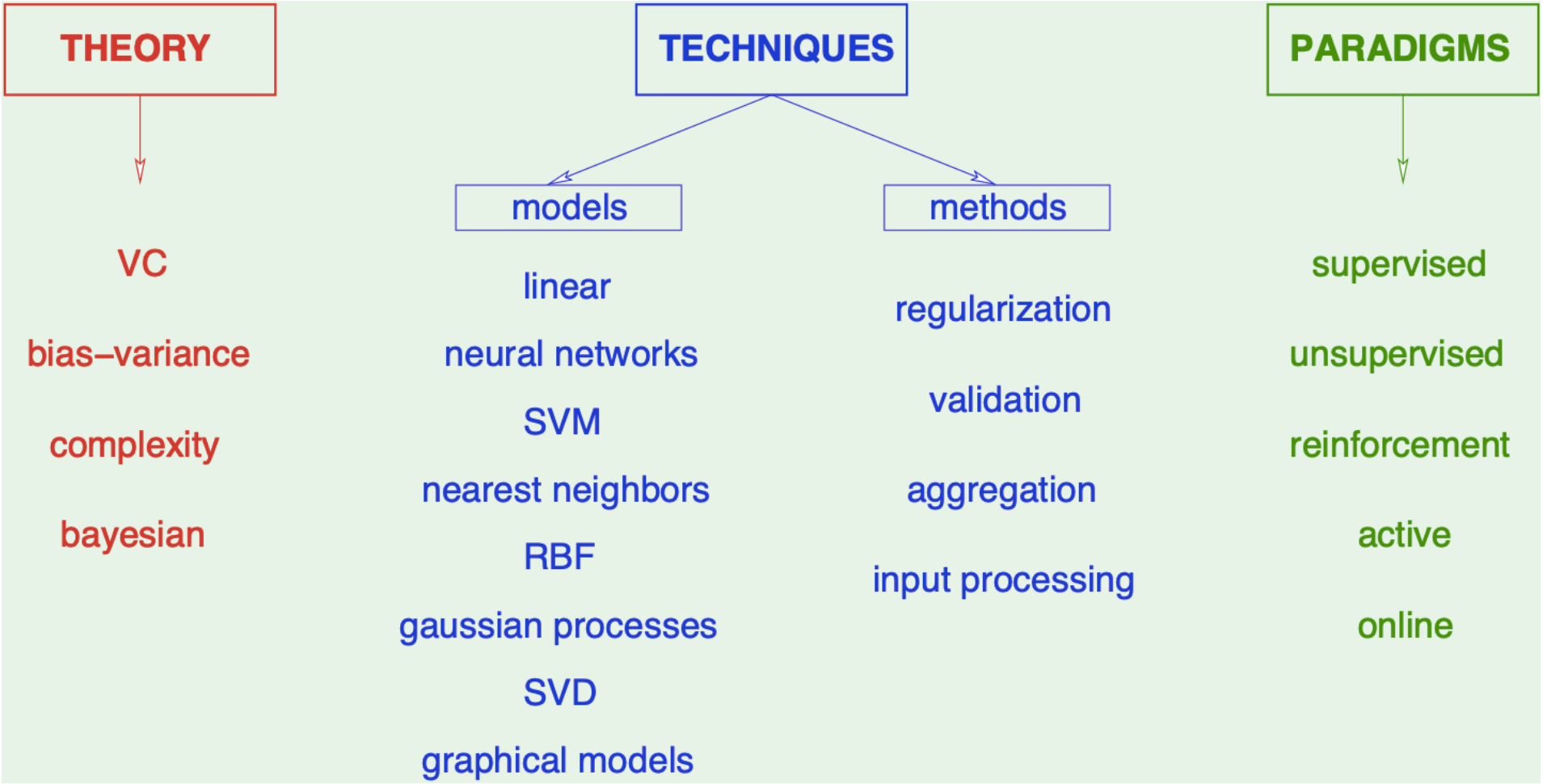




ML Jungle

semi–supervised learning
overfitting
stochastic gradient descent
SVM
Q learning
Gaussian processes
deterministic noise
data snooping
distribution–free
linear regression
VC dimension
learning curves
collaborative filtering
mixture of expe
decision trees
nonlinear transformation
sampling bias
neural networks
no free
RBF
training versus testing
noisy targets
Bayesian prior
active learning
linear models
bias–variance tradeoff
weak learners
ordinal regression
cross validation
logistic regression
data contamination
ensemble learning
types of learning
perceptrons
hidden Markov mo
error measures
kernel methods
graphical models
ploration versus exploitation
soft–order constraint
is learning feasible?
clustering
weight decay
Boltzmann mach
regularization
Occam's razor

ML (Travelled) Road Map



Thank You!