Lecture 2:

Overview

Fall 2022

Kai-Wei Chang
CS @ UCLA
kw+cm146@kwchang.net

The instructor gratefully acknowledges Eric Eaton (UPenn), who assembled the original slides, Jessica Wu (Harvey Mudd), David Kauchak (Pomona), Dan Roth (Upenn), Sriram Sankararaman (UCLA), whose slides are also heavily used, and the many others who made their course materials freely available online.

Announcement

- There is a discussion session on Friday
  - See session/time/loc at myUCLA
- Math Review Quiz is on BruinLearn

This Lecture

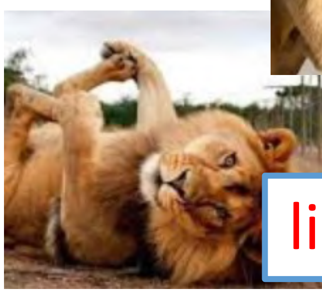
- Learning Protocols
  - Supervised Learning
  - Unsupervised Learning
- Challenges in ML
- Framing Learning Problems

Type of learning protocols

Supervised Learning

Training phase:





lion



Not lion

Test phase:

??

Supervised Learning

Input

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

An item **x** drawn from an input space $\mathcal{X}$

System

y = f(**x**)

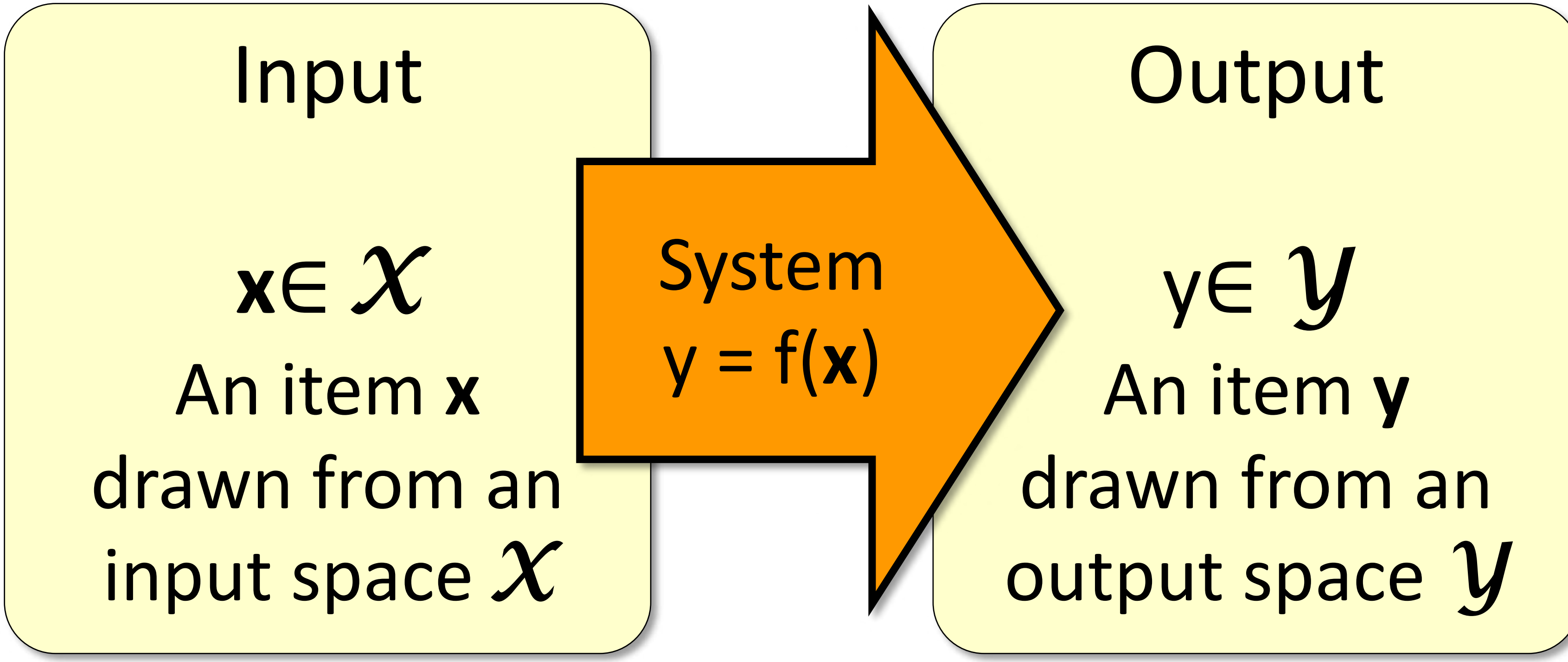
Output

$y \in \mathcal{Y}$

An item **y** drawn from an output space $\mathcal{Y}$

- We consider systems that apply a function f() to input items **x** and return an output **y** = f(**x**).

Supervised Learning



- In (supervised) machine learning, we deal with systems whose f(**x**) is learned from examples.

Supervised Learning

Input

x∈ $\mathcal{X}$

An item **x** drawn from an instance space $\mathcal{X}$

Target function

y = f(**x**)

Learned Model

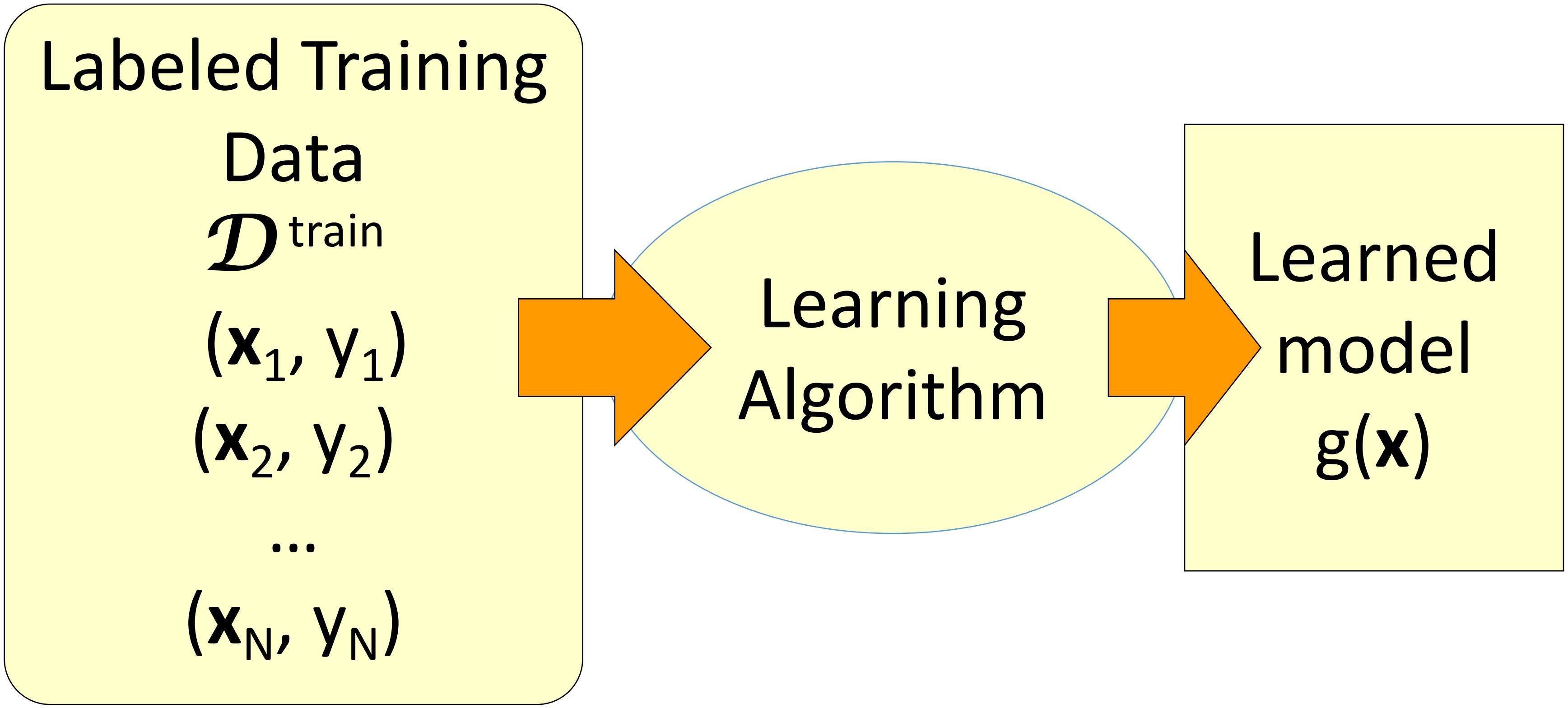
y = g(**x**)

Output

y∈ $\mathcal{Y}$

An item **y** drawn from a label space $\mathcal{Y}$

Supervised Learning: Training



- Give the learner examples in $\mathcal{D}^{\text{train}}$
- The learner returns a model g($\mathbf{x}$)

Supervised Learning: Testing

Labeled

Test Data

$\mathcal{D}^{\text{test}}$

$(\mathbf{x}'_1, y'_1)$

$(\mathbf{x}'_2, y'_2)$

...

$(\mathbf{x}'_M, y'_M)$

- Reserve some labeled data for testing

Supervised Learning: Testing

Labeled Test Data $\mathcal{D}^{\text{test}}$

$(\mathbf{x}'_1, y'_1)$

$(\mathbf{x}'_2, y'_2)$

…

$(\mathbf{x}'_M, y'_M)$

Raw Test Data $\mathcal{X}^{\text{test}}$

$\mathbf{x}'_1$

$\mathbf{x}'_2$

….

$\mathbf{x}'_M$

Test Labels $\mathcal{Y}^{\text{test}}$

y'_1

y'_2

…

y'_M

Supervised Learning: Testing

- Apply the model to the raw test data
- Evaluate by comparing predicted labels against the test labels

Raw Test Data $\mathcal{X}^{\text{test}}$

$\mathbf{x}'_1$

$\mathbf{x}'_2$

....

$\mathbf{x}'_M$

Learned model $g(\mathbf{x})$

Predicted Labels $g(\mathcal{X}^{\text{test}})$

$g(\mathbf{x}'_1)$

$g(\mathbf{x}'_2)$

....

$g(\mathbf{x}'_M)$

Test Labels $\mathcal{Y}^{\text{test}}$

y'_1

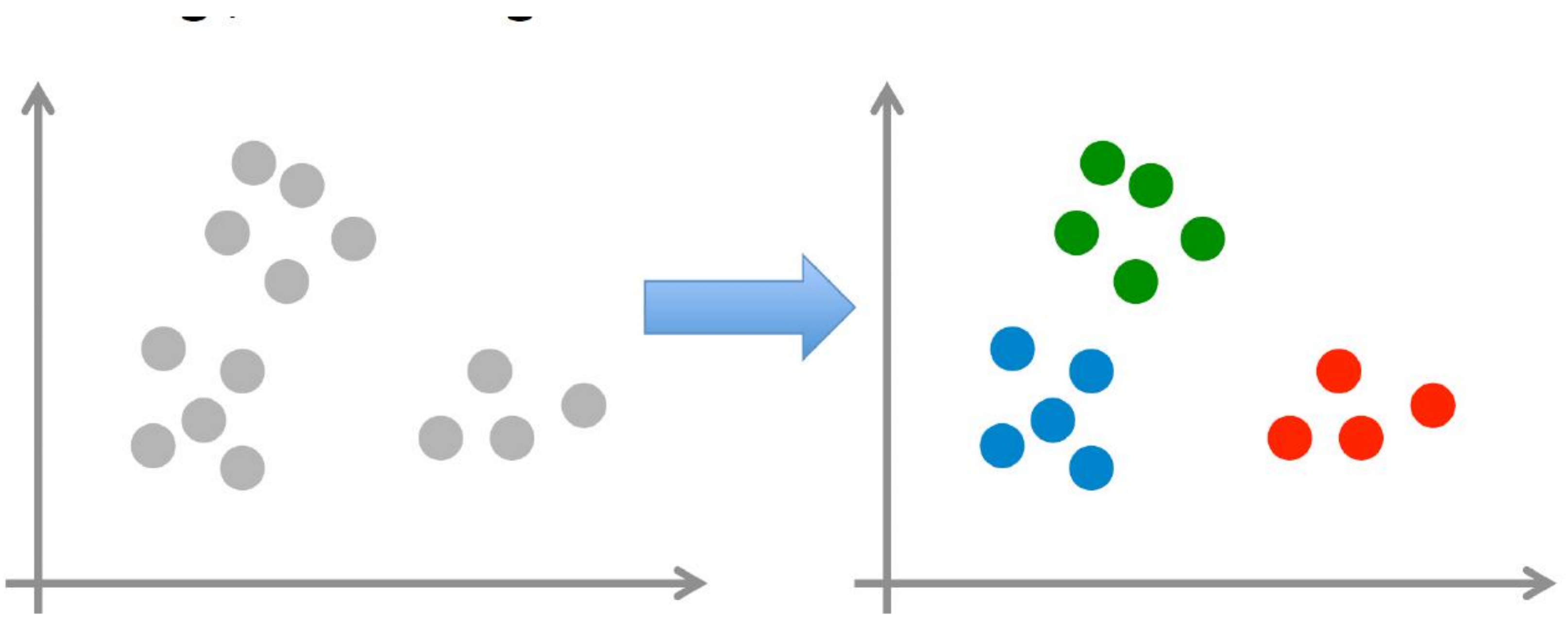
y'_2

...

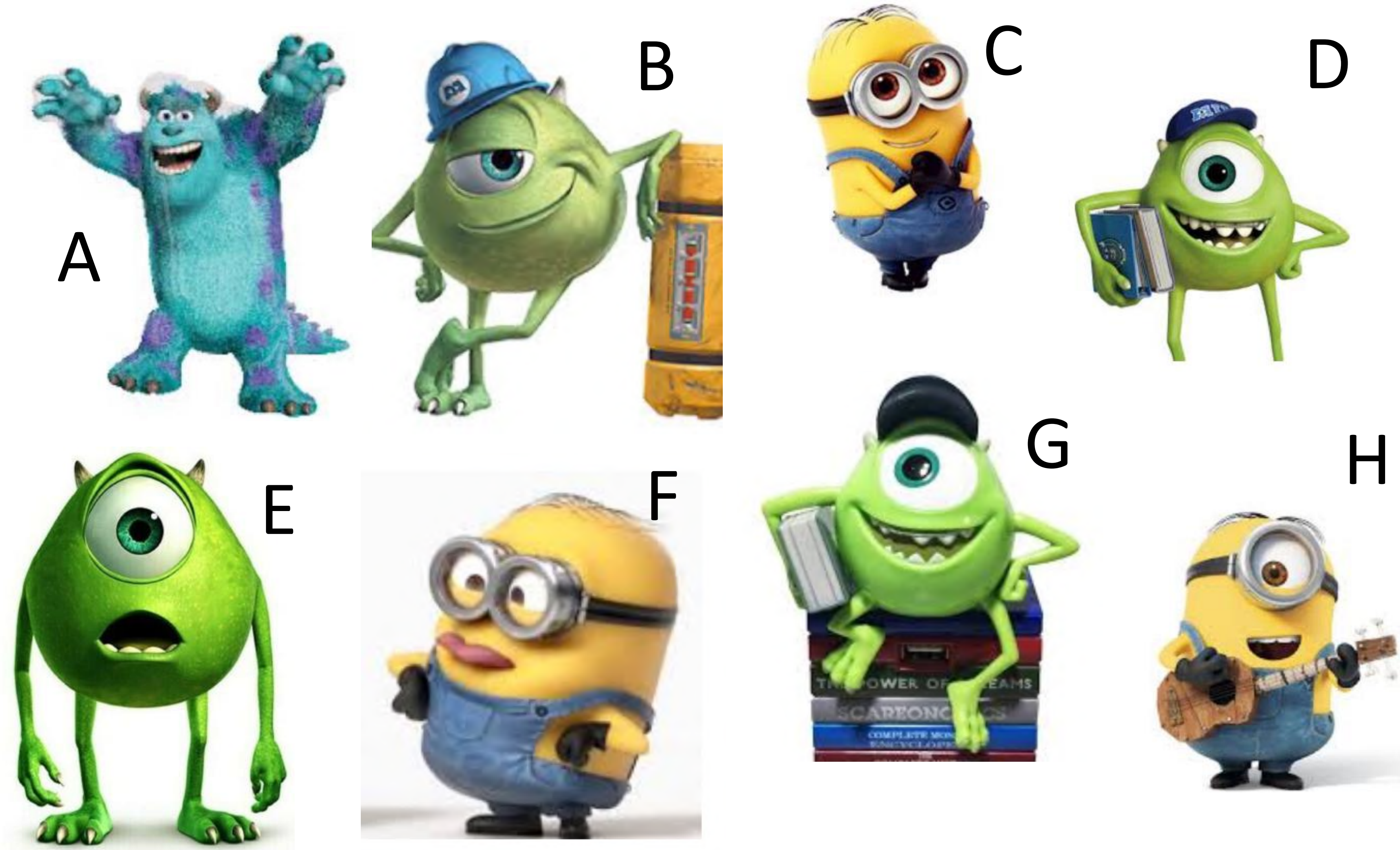
y'_M

Unsupervised learning

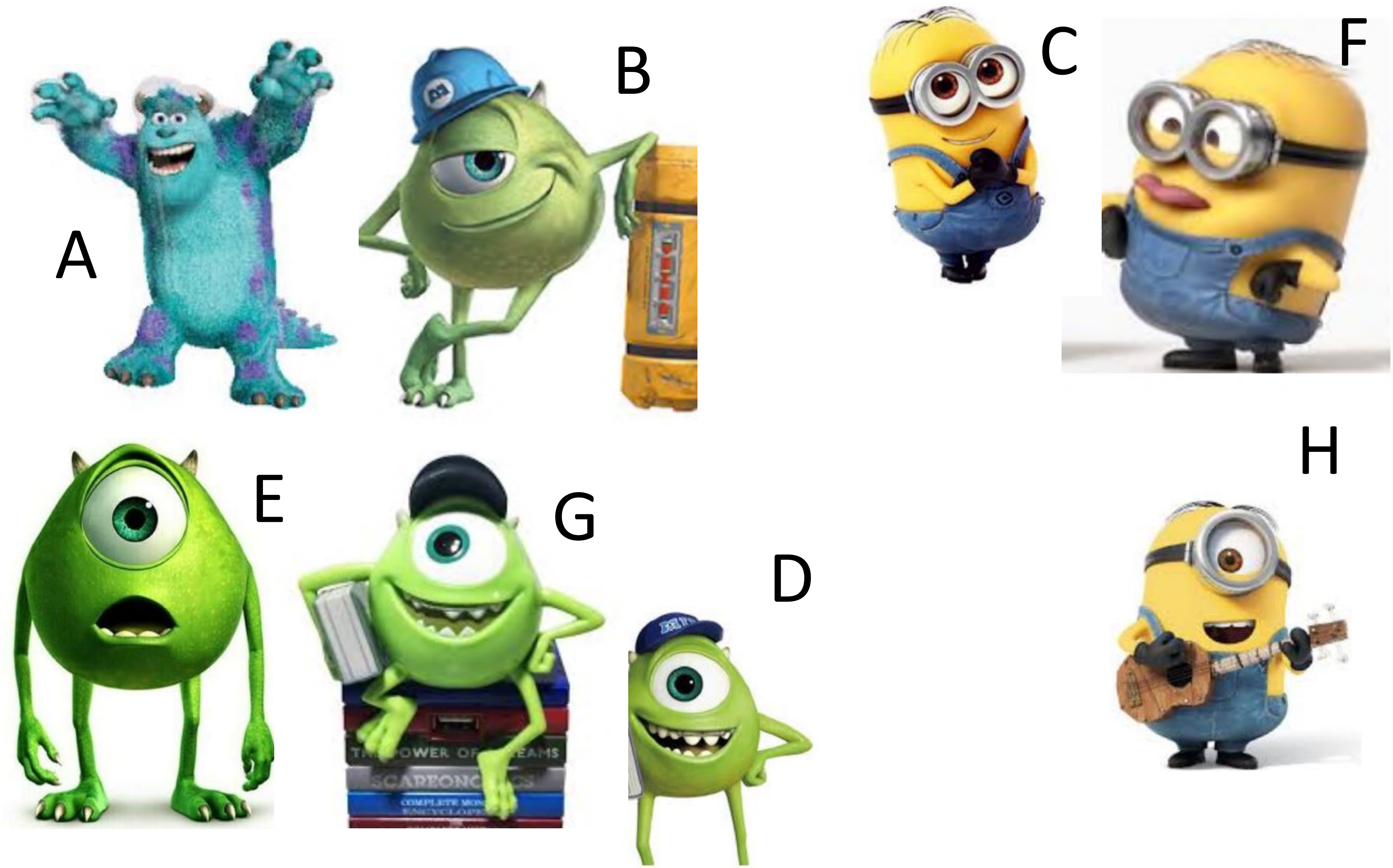
- Given: unlabeled inputs
- Goal: learn some intrinsic structure in inputs



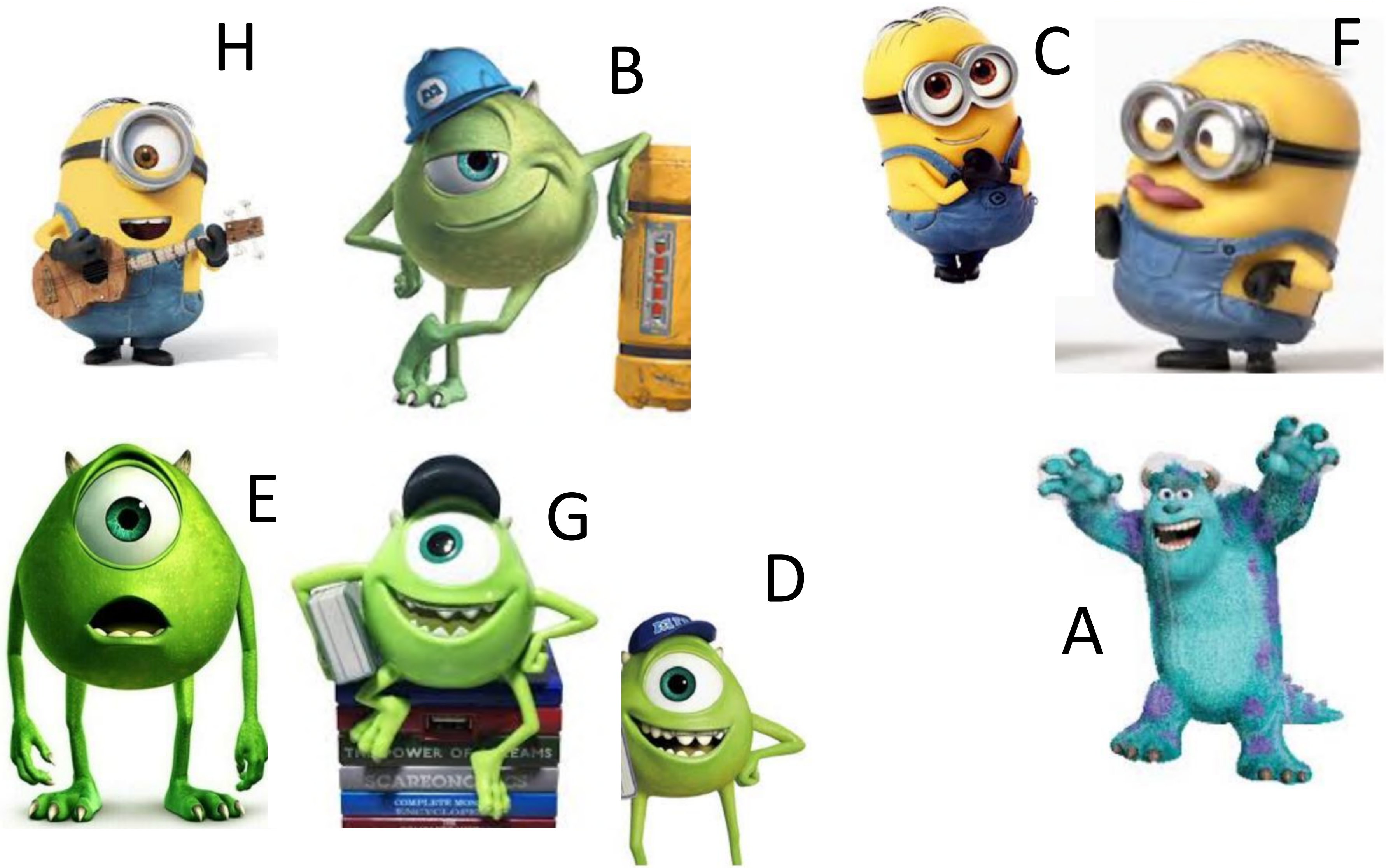
How many “kinds of monsters” are there?



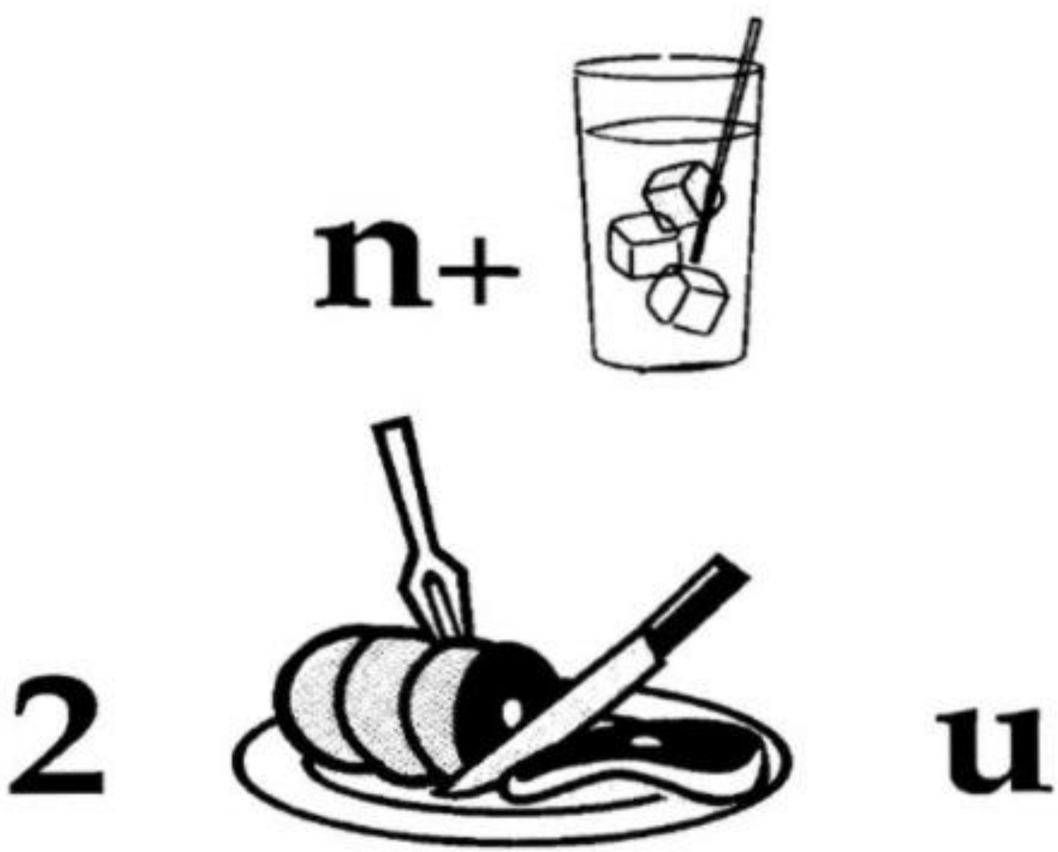
How many "kinds of monsters" are there?



How many “kinds of monsters” are there?



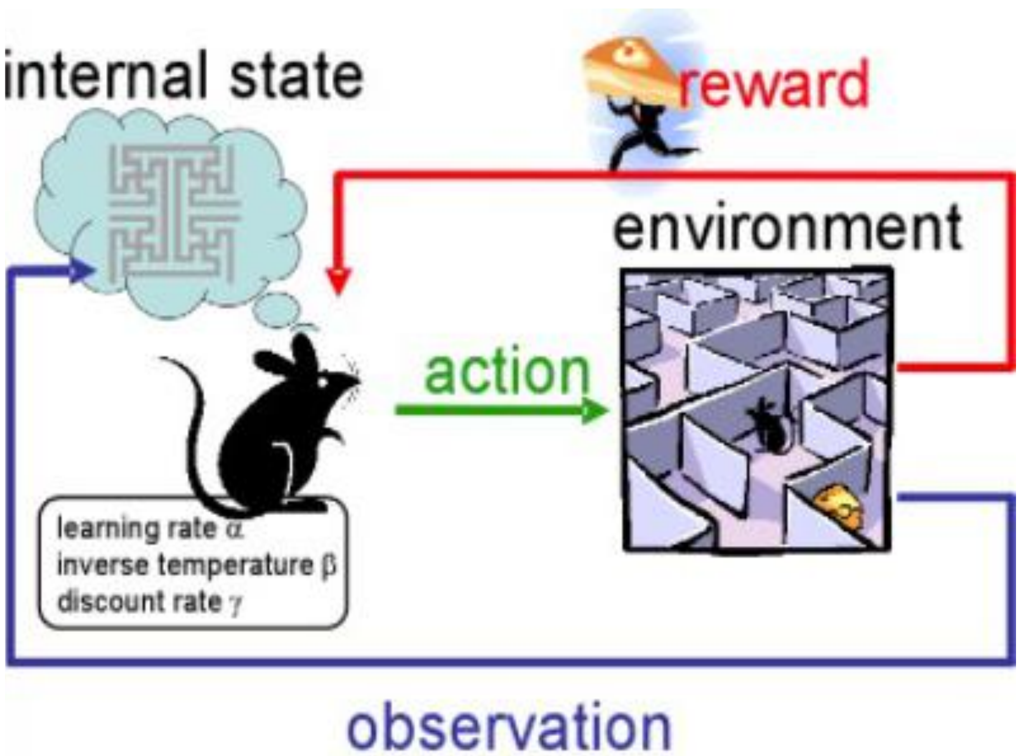
Decipher



Credit: Dan Roth

Reinforcement Learning

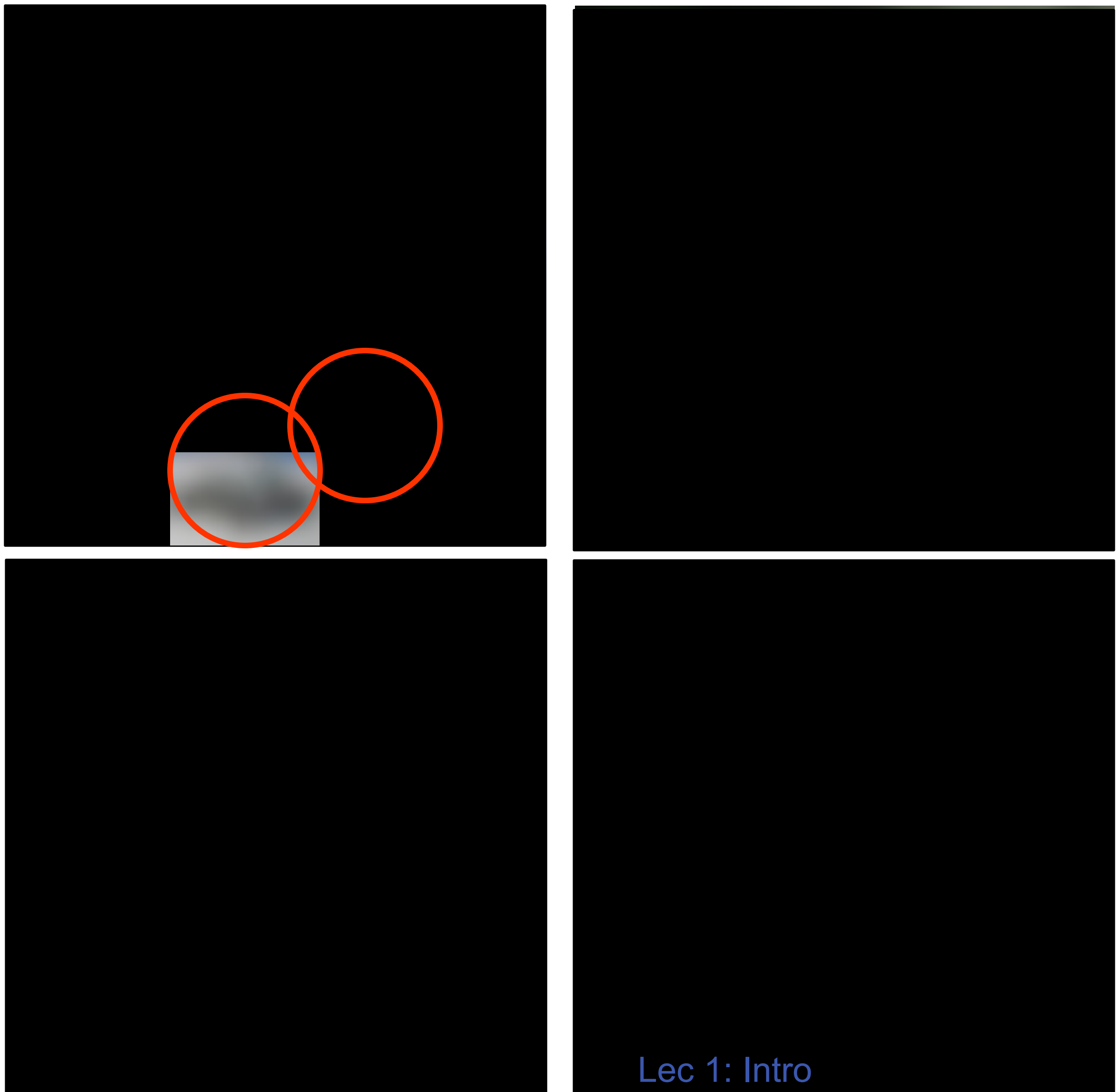
- Given sequences of states and actions with rewards
- Learn policy that maximizes agent's reward

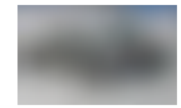


(image taken from Cyber Rodent Project)

Challenges in ML

Structured Inference

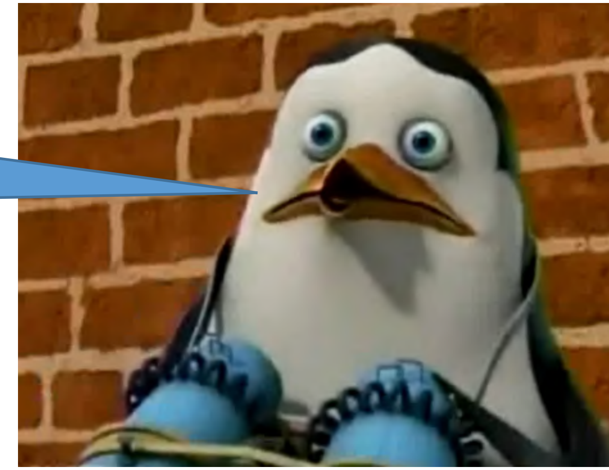




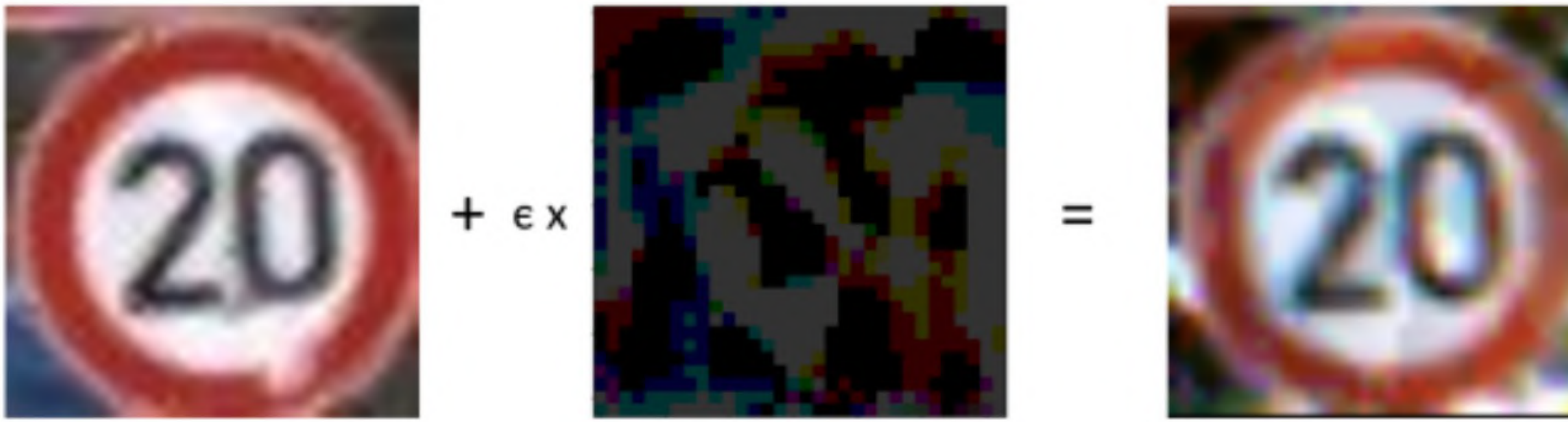
Credit: Dhruv Batra

Robustness

Car or shoe?



Adversarial Attack



93%, 20 Km/h Sign $\quad$ $sign(\nabla * J(\theta, x, y))$ $\quad$ 90%, 80 Km/h Sign

https://arxiv.org/abs/1712.09327v1

Commonsense

- Winograd Schema (1972)

The city councilmen refused the demonstrators a permit because they feared violence.

The city councilmen refused the demonstrators a permit because they advocated violence.

- Visual Commonsense

Is it raining outside?

a) Yes, it is snowing.
b) Yes, [person8] and [person10] are outside.
c) No, it looks to be fall.
d) Yes, it is raining heavily.

An example from the VCR dataset

Fairness/Inclusion in ML

Select photo

✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to

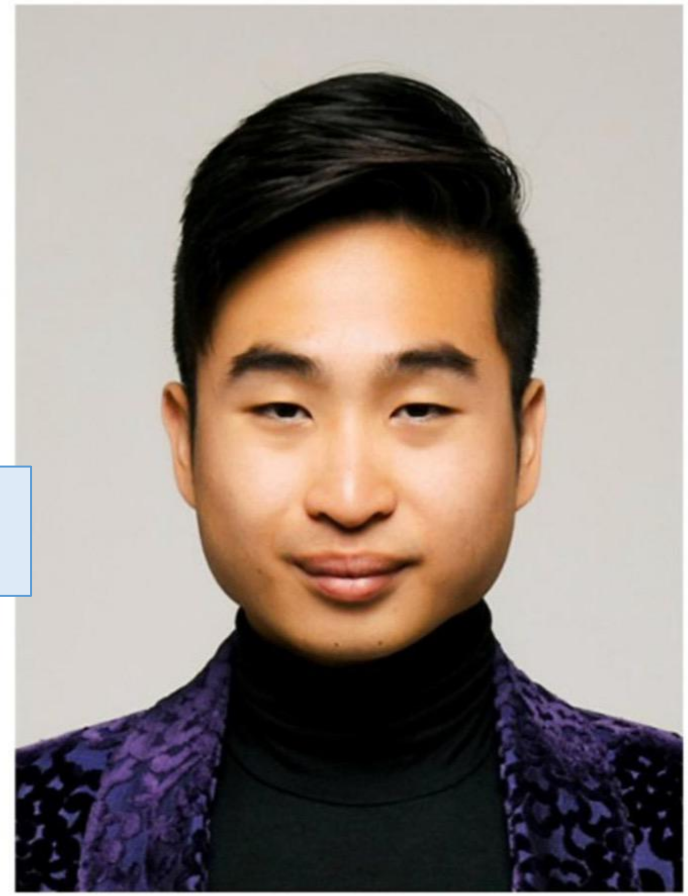
Subject eyes are closed

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.

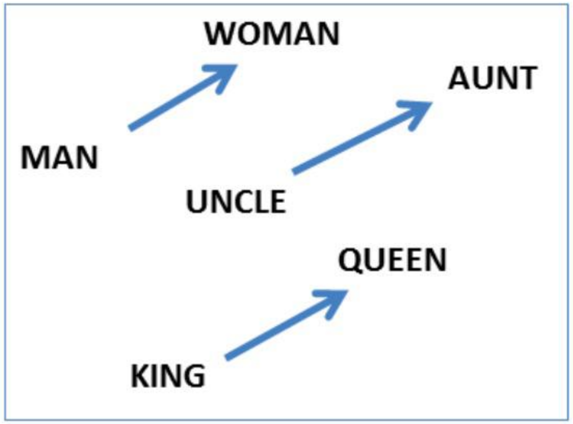


A screenshot of New Zealand man Richard Lee's passport photo rejection notice, supplied to Reuters December 7, 2016. Richard Lee/Handout via REUTERS

https://www.reuters.com/article/us-newzealand-passport-error/new-zealand-passport-robot-tells-applicant-of-asian-descent-to-open-eyes-idUSKBN13W0RL

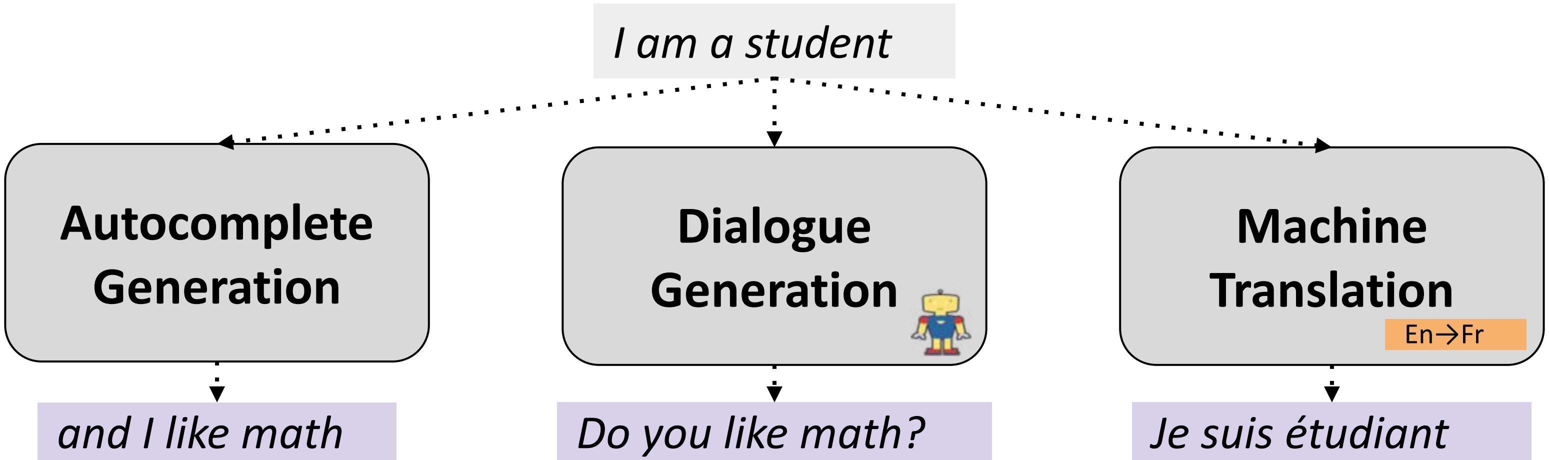
Fairness in ML-- Word embedding bias

- $v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$

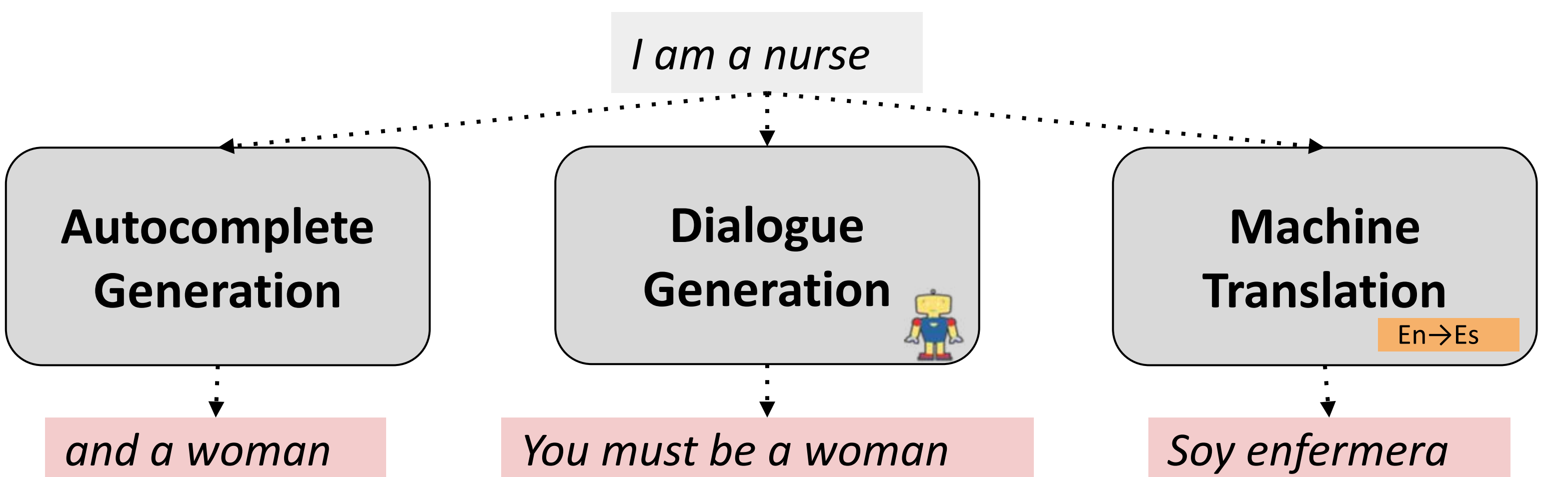


he: ___	she: ___
uncle	aunt
lion	
surgeon	
architect	
beer	
professor	

We use Google w2v embedding trained from the news



Language generations can be gendered!



[Societal Biases in Language Generation: Progress and Challenges](#)

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng, in ACL, 2021.

Misgendering in NLG

Alex went to the hospital for their appointment. [MASK] felt sick.

Prediction	Score
Alex went to the hospital for their appointment . **She** felt sick .	45.3%
Alex went to the hospital for their appointment . **He** felt sick .	36%
Alex went to the hospital for their appointment . **Alex** felt sick .	5.8%
Alex went to the hospital for their appointment . **I** felt sick .	2.3%
Alex went to the hospital for their appointment . **They** felt sick .	0.4%

https://demo.allennlp.org/masked-lm

Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang, in EMNLP, 2021.

Framing a Learning Problem

How we set up a learning problem

- The Badges Game……
  - This is an example of the key learning protocol: supervised learning

The Badges game

+ Naoki Abe	- Eric Baum

- Conference attendees to the 1994 Machine Learning conference were given name badges labeled with + or −.

- What function was used to assign these labels?

Training data

+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin
- Chidanand Apte
+ Minoru Asada
+ Lars Asker
+ Javed Aslam
+ Jose L. Balcazar
- Cristina Baroglio

+ Peter Bartlett
- Eric Baum
+ Welton Becket
- Shai Ben-David
+ George Berg
+ Neil Berkman
+ Malini Bhandaru
+ Bir Bhanu
+ Reinhard Blasig
- Avrim Blum
- Anselm Blumer
+ Justin Boyan

+ Carla E. Brodley
+ Nader Bshouty
- Wray Buntine
- Andrey Burago
+ Tom Bylander
+ Bill Byrne
- Claire Cardie
+ John Case
+ Jason Catlett
- Philip Chan
- Zhixiang Chen
- Chris Darken

Raw test data

Gerald F. DeJong
Chris Drummond
Yolanda Gil
Attilio Giordana
Jiarong Hong
J. R. Quinlan
Priscilla Rasmussen
Dan Roth
Yoram Singer
Lyle H. Ungar

Labeled test data

+ Gerald F. DeJong
- Chris Drummond
+ Yolanda Gil
- Attilio Giordana
+ Jiarong Hong
- J. R. Quinlan
- Priscilla Rasmussen
+ Dan Roth
+ Yoram Singer
- Lyle H. Ungar

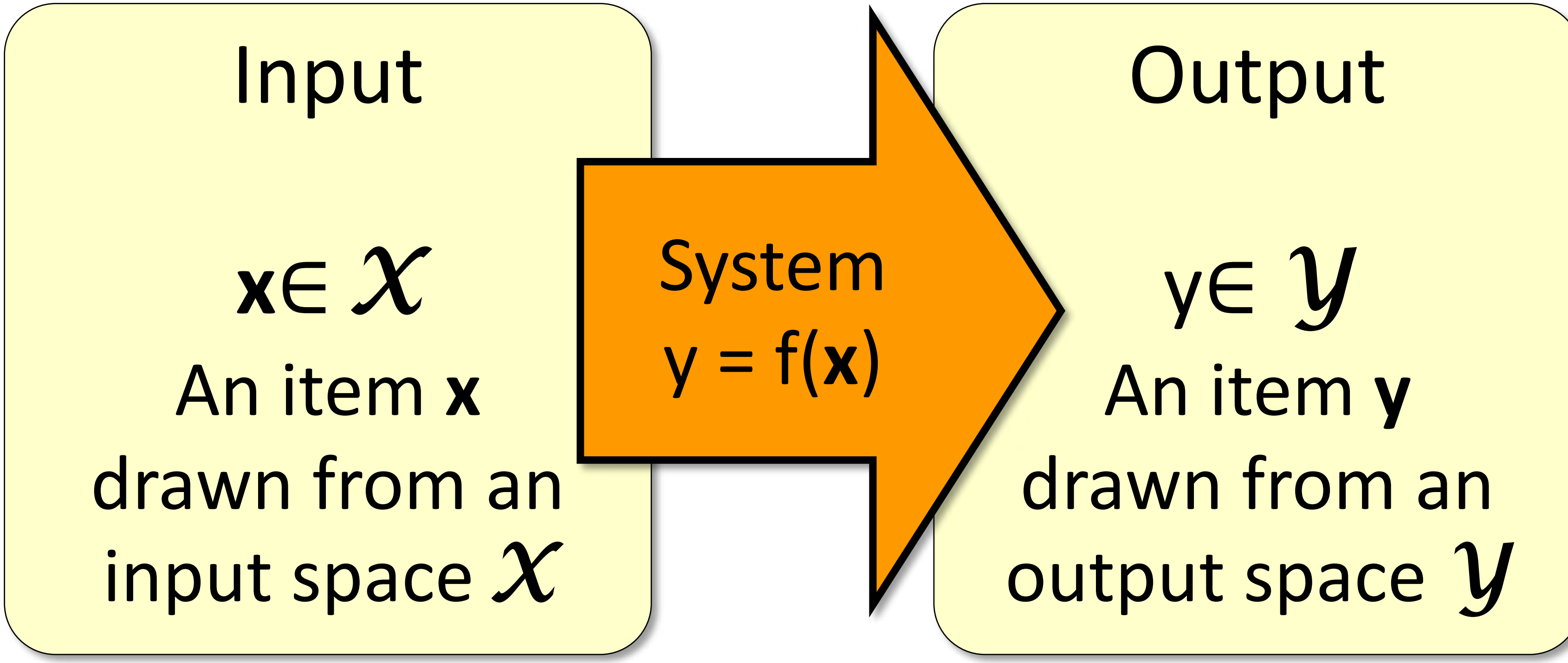
Exercise: What is the rule?

+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin
- Chidanand Apte
+ Minoru Asada
+ Lars Asker
+ Javed Aslam
+ Jose L. Balcazar
- Cristina Baroglio

+ Peter Bartlett
- Eric Baum
+ Welton Becket
- Shai Ben-David
+ George Berg
+ Neil Berkman
+ Malini Bhandaru
+ Bir Bhanu
+ Reinhard Blasig
- Avrim Blum
- Anselm Blumer
+ Justin Boyan

+ Carla E. Brodley
+ Nader Bshouty
- Wray Buntine
- Andrey Burago
+ Tom Bylander
+ Bill Byrne
- Claire Cardie
+ John Case
+ Jason Catlett
- Philip Chan
- Zhixiang Chen
- Chris Darken

Supervised Learning



- We consider systems that apply a function f() to input items **x** and return an output **y** = f(**x**).

Using supervised learning

- What is our instance space?
  - Gloss: What kind of features are we using?
- What is our label space?
  - Gloss: What kind of learning task are we dealing with?
- What is our hypothesis space?
  - Gloss: What kind of functions (models) are we learning?
- What learning algorithm do we use?
  - Gloss: How do we learn the model from the labeled data?
- What is our loss function/evaluation metric?
  - Gloss: How do we measure success? What drives learning?

1. Input: The instance space $\mathcal{X}$

Input

x$\in \mathcal{X}$

An item **x** drawn from an instance space $\mathcal{X}$

x is represented in a feature space

- Typically $x \in \{0,1\}^n$ or R^N
- Usually represented as a vector
- We call it input vector

Example:

Boolean features:

Does this email contain the word 'money'?

Numerical features:

How often does 'money' occur in this email

What is the width/height of this bounding box?

What is the length of the first name?

What's $\mathcal{X}$ for the Badges game?

- Possible features:
  - Length of their first or last name?
  - Does the name contain letter 'x'?
  - How many vowels does their name contain?
  - Is the n-th letter a vowel?

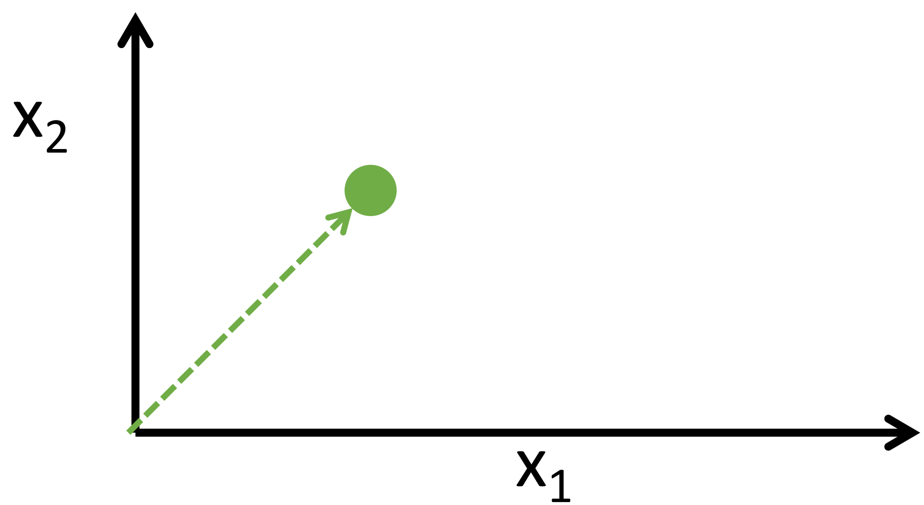
+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin

+ Peter Bartlett
- Eric Baum
+ Welton Becket
- Shai Ben-David
+ George Berg
+ Neil Berkman

+ Carla E. Brodley
+ Nader Bshouty
- Wray Buntine
- Andrey Burago
+ Tom Bylander
+ Bill Byrne

$\mathcal{X}$ as a vector space

- $\mathcal{X}$ is an N-dimensional vector space (e.g. R^N)
  - Each dimension = one feature.
- Each **x** is a feature vector (hence the boldface **x**).
- Think of $\mathbf{x} = [x_1 \ldots x_N]$ as a point in $\mathcal{X}$:



Example: the badge game

+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin

+ Peter Bartlett
- Eric Baum
+ Welton Becket
- Shai Ben-David
+ George Berg
+ Neil Berkman

+ Carla E. Brodley
+ Nader Bshouty
- Wray Buntine
- Andrey Burago
+ Tom Bylander
+ Bill Byrne

[first-char is vowel, first-char is A, first-char is N, second-char is vowel …]

+ Naoki Abe

[0 , 0 , 1 , 1 …]

- Avrim Blum

[1 , 1 , 0 , 0 …]

Good features are essential

- The choice of features is crucial for how well a task can be learned.
  - In many application areas (language, vision, etc.), a lot of work goes into designing suitable features.
  - This requires domain expertise.
- CM146 can't teach you what specific features to use for your task.
  - But we will touch on some general principles

2. Output space

y is represented in output space (label space)
Different kinds of output:

- Binary classification:
  $y \in \{-1,1\}$
- Multiclass classification:
  $y \in \{1,2,3, \ldots K\}$
- Regression:
  $y \in R$
- Structured output
  $y \in \{1,2,3, \ldots K\}^N$

Output

y$\in \mathcal{Y}$

An item **y** drawn from a label space $\mathcal{Y}$

Supervised Learning : Examples

Animal recognition

- x: Bitmap picture of the animal
- y :

Lion? Yes/No

Lion/Cat/Dog

Lion/Mammal/Dog/Fish

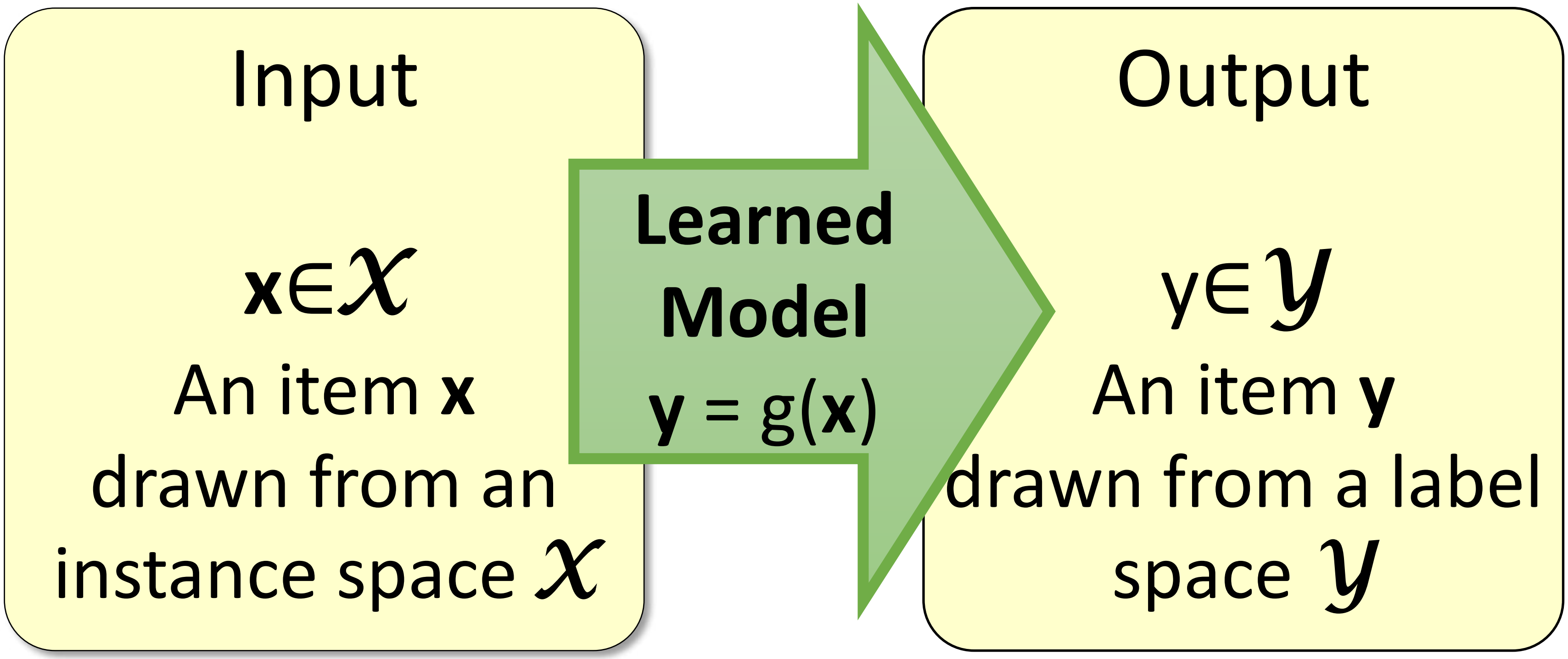
Binary output

Multiclass output

Multilabel output

Output of the applications may different from the output of ML models.

3. The model g(**x**)



- We need to choose what *kind* of model we want to learn

3. The model g(**x**)

Input

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

An item **x** drawn from an instance space $\mathcal{X}$

Learned Model

y = g(**x**)

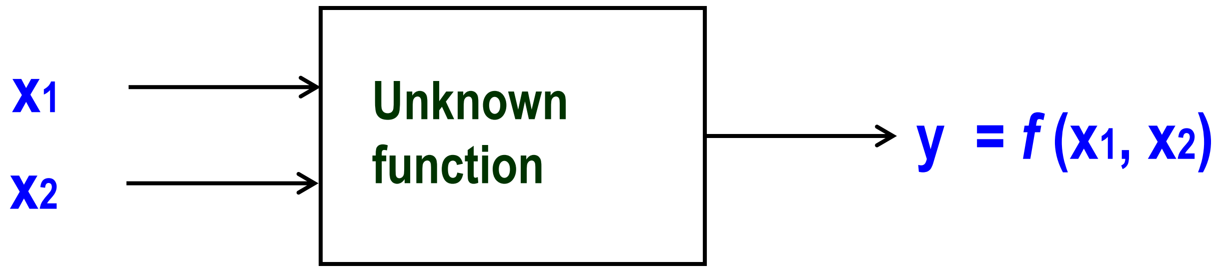
Output

$y \in \mathcal{Y}$

An item **y** drawn from a label space $\mathcal{Y}$

- We need to choose what *kind* of model we want to learn

Boolean Function



x_1	x_2	y
0	0	0
0	1	0
1	0	1
1	1	1

Function 1

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	1

Function 2

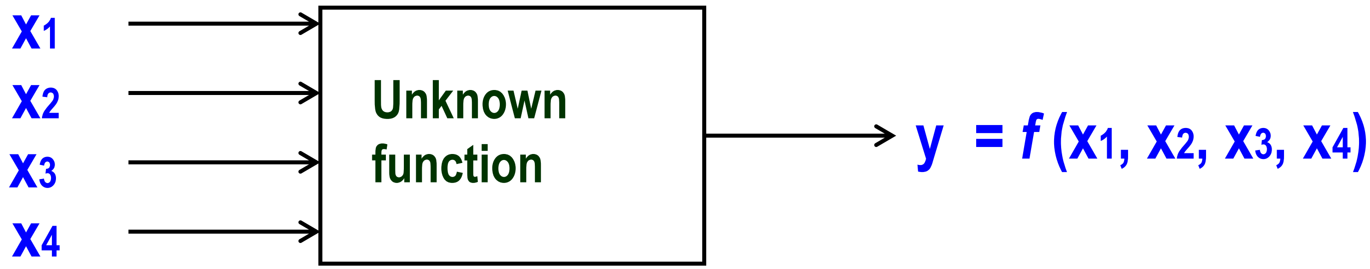
x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	0

Function 3

…

Hypothesis Space

A Learning Problem



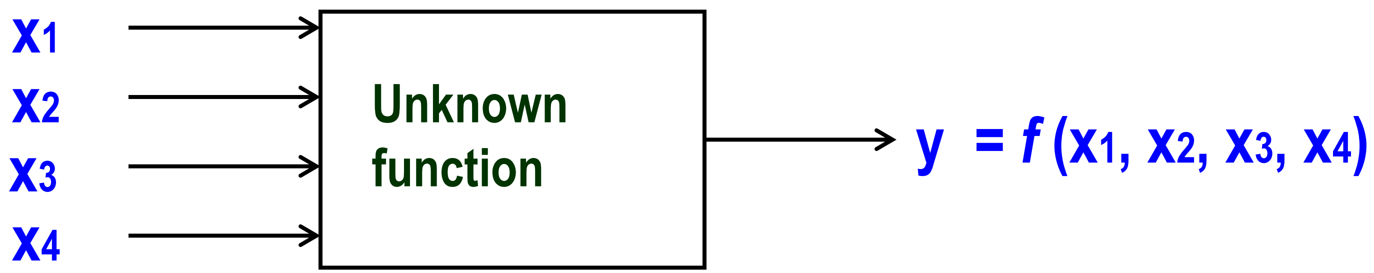
Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Can you learn this function?

What is it?

A function g is consistent to a dataset $\mathrm{D} = \{(x_i, y_i)\}$ if $g(x_i) = y_i, \forall\, i$

Discussion: A Learning Problem



Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Can you learn this function?

What is it?

A function g is consistent to a dataset $\mathrm{D} = \{(x_i, y_i)\}$ if $g(x_i) = y_i, \forall\, i$

How many possible functions over four features?

How many function is consistent to D on the left

Hypothesis Space

How many possible functions over four features?

Complete Ignorance:

There are $2^{16} = 65536$ possible functions over four input features.

Example	x_1	x_2	x_3	x_4	y
	0	0	0	0	?
	0	0	0	1	?
	0	0	1	0	?
	0	0	1	1	?
	0	1	0	0	?
	0	1	0	1	?
	0	1	1	0	?
	0	1	1	1	?
	1	0	0	0	?
	1	0	0	1	?
	1	0	1	0	?
	1	0	1	1	?
	1	1	0	0	?
	1	1	0	1	?
	1	1	1	0	?
	1	1	1	1	?

Hypothesis Space

Complete Ignorance:
There are $2^{16} = 65536$ possible functions over four input features.

We can't figure out which one is correct until we've seen every possible input-output pair.

After observing seven examples we still have 2^9 possibilities for f

Is Learning Possible?
which one is the most likely one?

Example	x_1	x_2	x_3	x_4	y
	0	0	0	0	?
	0	0	0	1	?
	0	0	1	0	0
	0	0	1	1	1
	0	1	0	0	0
	0	1	0	1	0
	0	1	1	0	0
	0	1	1	1	?
	1	0	0	0	?
	1	0	0	1	1
	1	0	1	0	?
	1	0	1	1	?
	1	1	0	0	0
	1	1	0	1	?
	1	1	1	0	?
	1	1	1	1	?

Hypothesis Space

Complete Ignorance:

There are $2^{16} = 65536$ possible functions over four input features.

We c

corre

poss

After

have 2 possibilities for f

Is Learning Possible?

Example	x_1	x_2	x_3	x_4	y
	0	0	0	0	?
	0	0	0	1	?
	0	0	1	0	0
					1
					0
					0
					0
					?
					?
					1
					?
					?
	1	1	0	0	0
	1	1	0	1	?
	1	1	1	0	?
	1	1	1	1	?

- There are $|Y|^{|X|}$ possible functions f(**x**) from the instance space **X** to the label space **Y.**
- Learners typically consider only a *subset* of the functions from **X** to **Y**, called the hypothesis space **H** . $\mathbf{H} \subseteq |Y|^{|X|}$

Hypothesis Space (2)

Simple Rules: **conjunctive rules**

of the form $y=x_i \Lambda x_j \Lambda \ldots \Lambda x_k$

e.g., $\mathbf{y = x_2 \; \Lambda \; x_3}$

$\mathbf{y = x_1 \; \Lambda \; x_2 \; \Lambda \; X_4}$

How large is the hypothesis space?

Hypothesis Space (2)

1	0	0	1	0	**0**
2	0	1	0	0	**0**
3	0	0	1	1	**1**
4	1	0	0	1	**1**
5	0	1	1	0	**0**
6	1	1	0	0	**0**
7	0	1	0	1	**0**

Simple Rules: There are only 16 simple **conjunctive rules**

of the form $y=x_i \Lambda x_j \Lambda x_k$

Rule	Counterexample	Rule	Counterexample
$y=1$		$x_2 \Lambda x_3$	
x_1		$x_2 \Lambda x_4$	
x_2		$x_3 \Lambda x_4$	
x_3		$x_1 \Lambda x_2 \Lambda x_3$	
x_4		$x_1 \Lambda x_2 \Lambda x_4$	
$x_1 \Lambda x_2$		$x_1 \Lambda x_3 \Lambda x_4$	
$x_1 \Lambda x_3$		$x_2 \Lambda x_3 \Lambda x_4$	
$x_1 \Lambda x_4$		$x_1 \Lambda x_2 \Lambda x_3 \Lambda x_4$	

Hypothesis Space (2)

1	0	0	1	0	**0**
2	0	1	0	0	**0**
3	0	0	1	1	**1**
4	1	0	0	1	**1**
5	0	1	1	0	**0**
6	1	1	0	0	**0**
7	0	1	0	1	**0**

Simple Rules: There are only 16 simple **conjunctive rules**

of the form $y=x_i \wedge x_j \wedge x_k$

Rule	Counterexample	Rule	Counterexample
$y=c$		$x_2 \wedge x_3$	0011 1
x_1	1100 0	$x_2 \wedge x_4$	0011 1
x_2	0100 0	$x_3 \wedge x_4$	1001 1
x_3	0110 0	$x_1 \wedge x_2 \wedge x_3$	0011 1
x_4	0101 1	$x_1 \wedge x_2 \wedge x_4$	0011 1
$x_1 \wedge x_2$	1100 0	$x_1 \wedge x_3 \wedge x_4$	0011 1
$x_1 \wedge x_3$	0011 1	$x_2 \wedge x_3 \wedge x_4$	0011 1
$x_1 \wedge x_4$	0011 1	$x_1 \wedge x_2 \wedge x_3 \wedge x_4$	0011 1

No simple rule explains the data. The same is true for **simple clauses**.

Hypothesis Space (3)

1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

m-of-n rules: There are 32 possible rules of the form "y = 1 if and only if at least m of the following n variables are 1"

Notation: 2 variables from the set on the left. **Value**: Index of the counterexample.

variables	1-of	2-of	3-of	4-of
{X1}				
{X2}				
{X3}				
{X4}				
{X1,X2}				
{X1, X3}				
{X1, X4}				
{X2,X3}				

variables	1-of	2-of	3-of	4-of
{X2, X4}				
{X3, X4}				
{X1,X2, X3}				
{X1,X2, X4}				
{X1,X3,X4}				
{X2, X3,X4}				
{X1, X2, X3,X4}				

Hypothesis Space (3)

1	0	0	1	0	**0**
2	0	1	0	0	**0**
3	0	0	1	1	**1**
4	1	0	0	1	**1**
5	0	1	1	0	**0**
6	1	1	0	0	**0**
7	0	1	0	1	**0**

m-of-n rules: There are 32 possible rules of the form "y = 1 if and only if at least m of the following n variables are 1"

Notation: 2 variables from the set on the left. **Value**: Index of the counterexample.

variables	1-of	2-of	3-of	4-of
$\{x_1\}$	3	-	-	-
$\{x_2\}$	2	-	-	-
$\{x_3\}$	1	-	-	-
$\{x_4\}$	7	-	-	-
$\{x_1, x_2\}$	2	3	-	-
$\{x_1, x_3\}$	1	3	-	-
$\{x_1, x_4\}$	6	3	-	-
$\{x_2, x_3\}$	2	3	-	-

variables	1-of	2-of	3-of	4-of
$\{x_2, x_4\}$	2	3	-	-
$\{x_3, x_4\}$	4	4	-	-
$\{x_1, x_2, x_3\}$	1	3	3	-
$\{x_1, x_2, x_4\}$	2	3	3	-
$\{x_1, x_3, x_4\}$	1	***	3	-
$\{x_2, x_3, x_4\}$	1	5	3	-
$\{x_1, x_2, x_3, x_4\}$	1	5	3	3

Found a consistent hypothesis.

Views of Learning

- Learning is the removal of our <u>remaining</u> uncertainty:
- Learning requires guessing a good hypothesis class
  - Start with a small class and enlarge it until it contains an hypothesis that fits the data.
- We could be wrong !
  - Our guess of the hypothesis space could be wrong
    - $y = x4 \ \wedge$ one-of $(x1, x3)$ is also consistent

General strategies for Machine Learning

- Develop flexible hypothesis spaces:
  - Decision trees, neural networks, nested collections.
- Develop representation languages for restricted classes of functions:
  - E.g., Functional representation (n-of-m); Grammars; linear functions; stochastic models

General strategies for Machine Learning

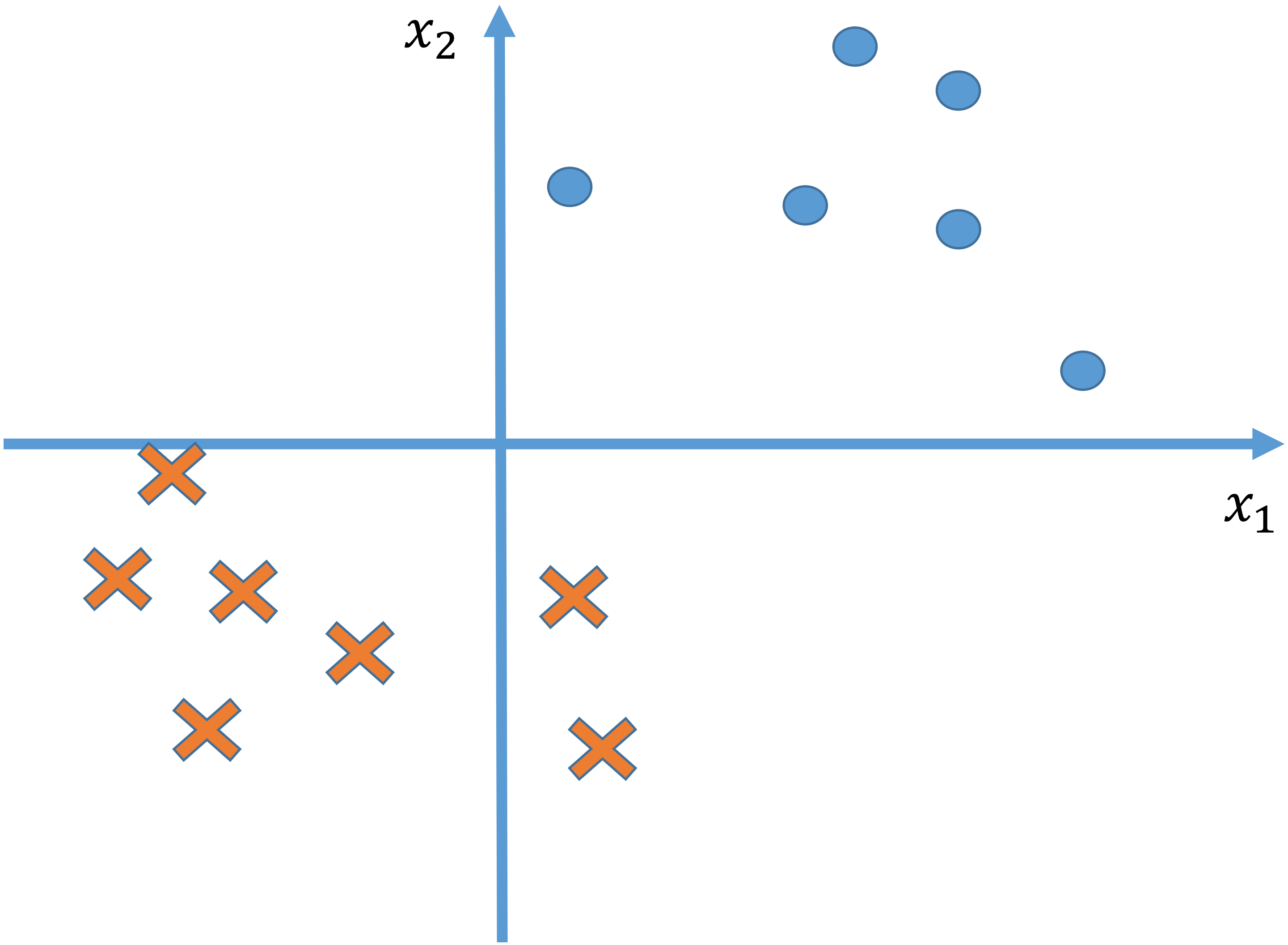
- Develop flexible hypothesis spaces:
  - Decision trees, neural networks, nested collections.
- Develop representation languages for restricted classes of functions:

In either case:

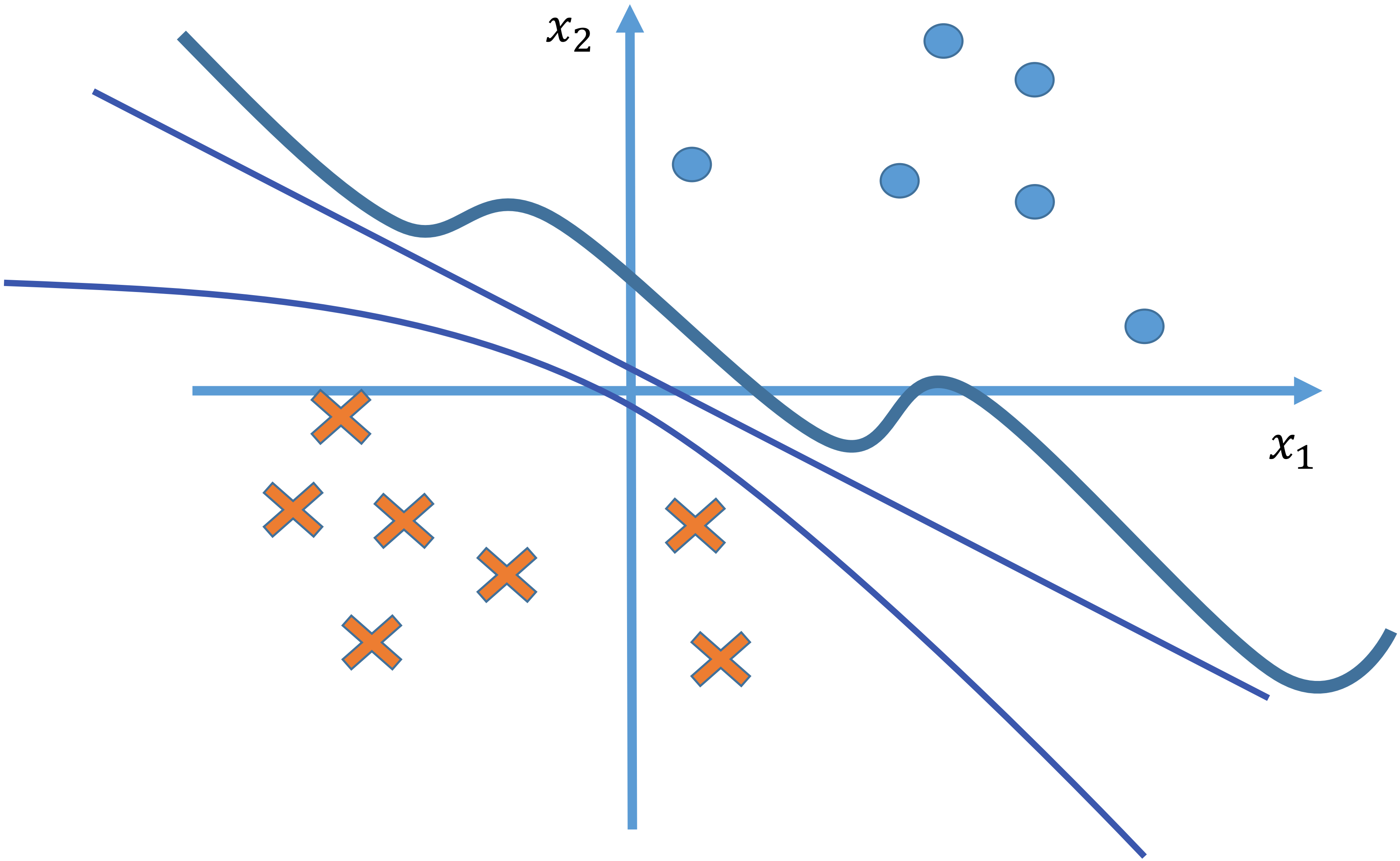
- Develop algorithms for finding a hypothesis in our hypothesis space, that fits the data
- And hope that they will generalize well

Hypothesis Space -- Real-Value Features

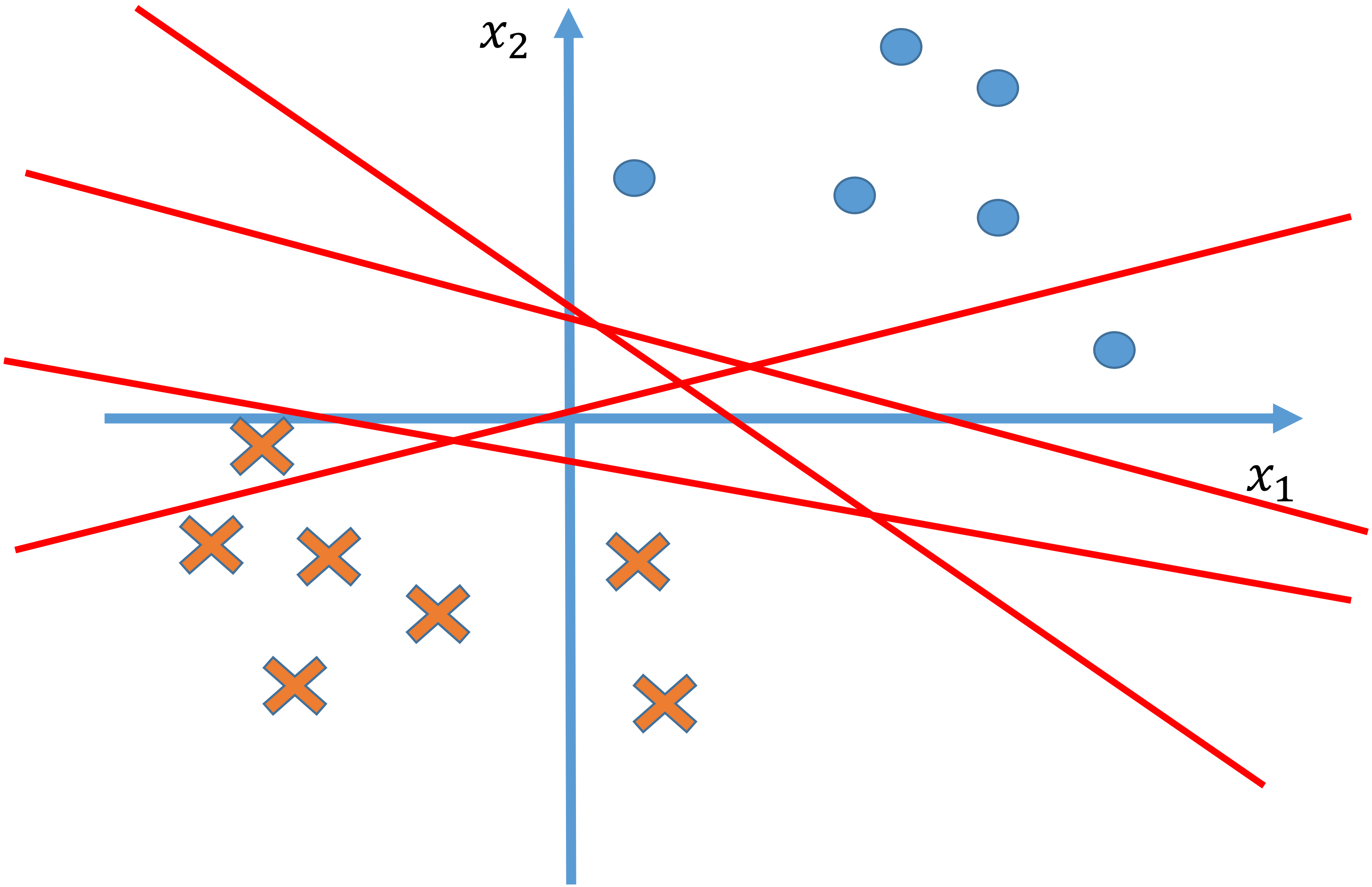
Example problem



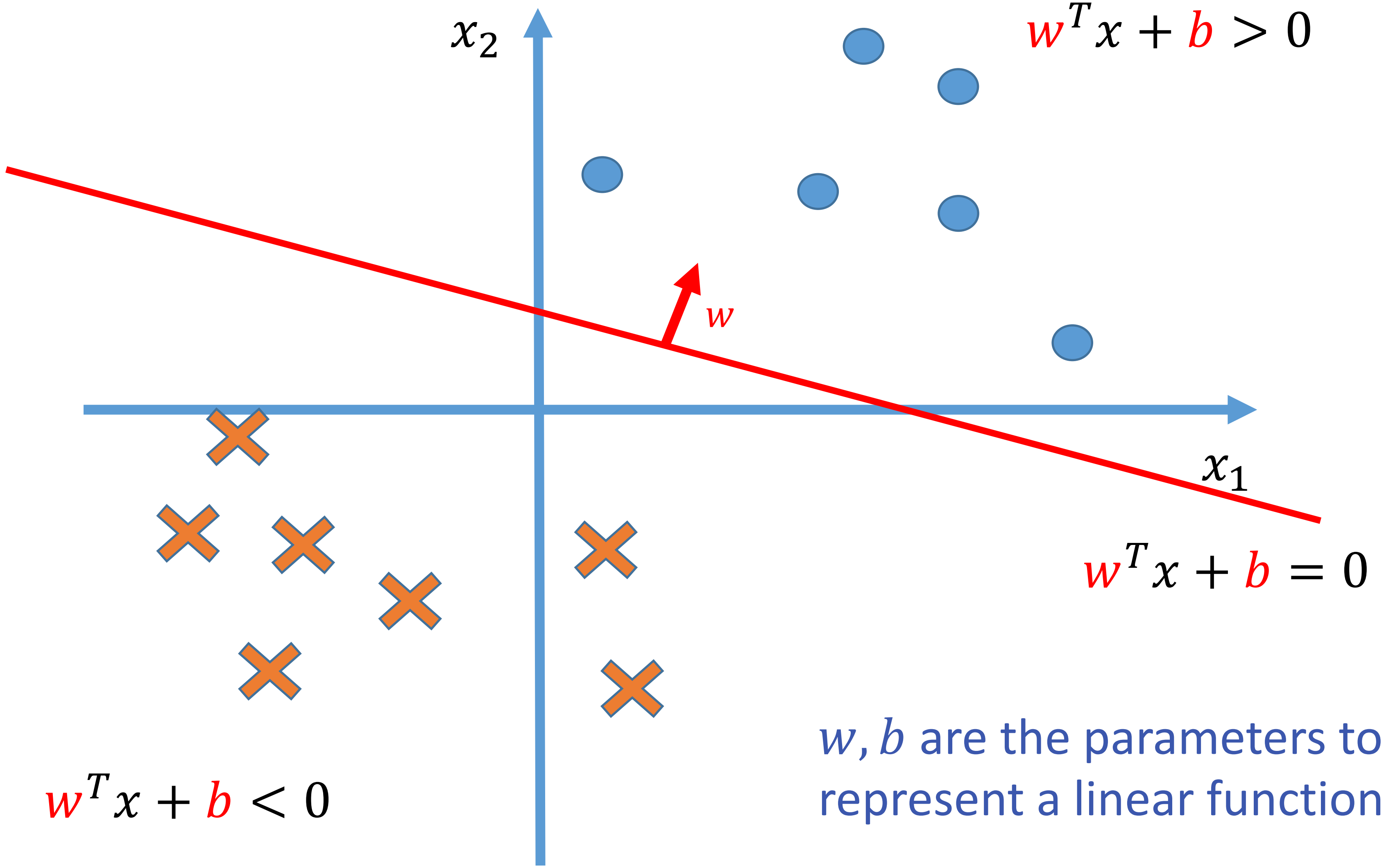
Hypothesis space:



Hypothesis space: linear model



Hypothesis space: linear model



w, b are the parameters to represent a linear function