Essentials of Machine Learning

Carlos Cotrini

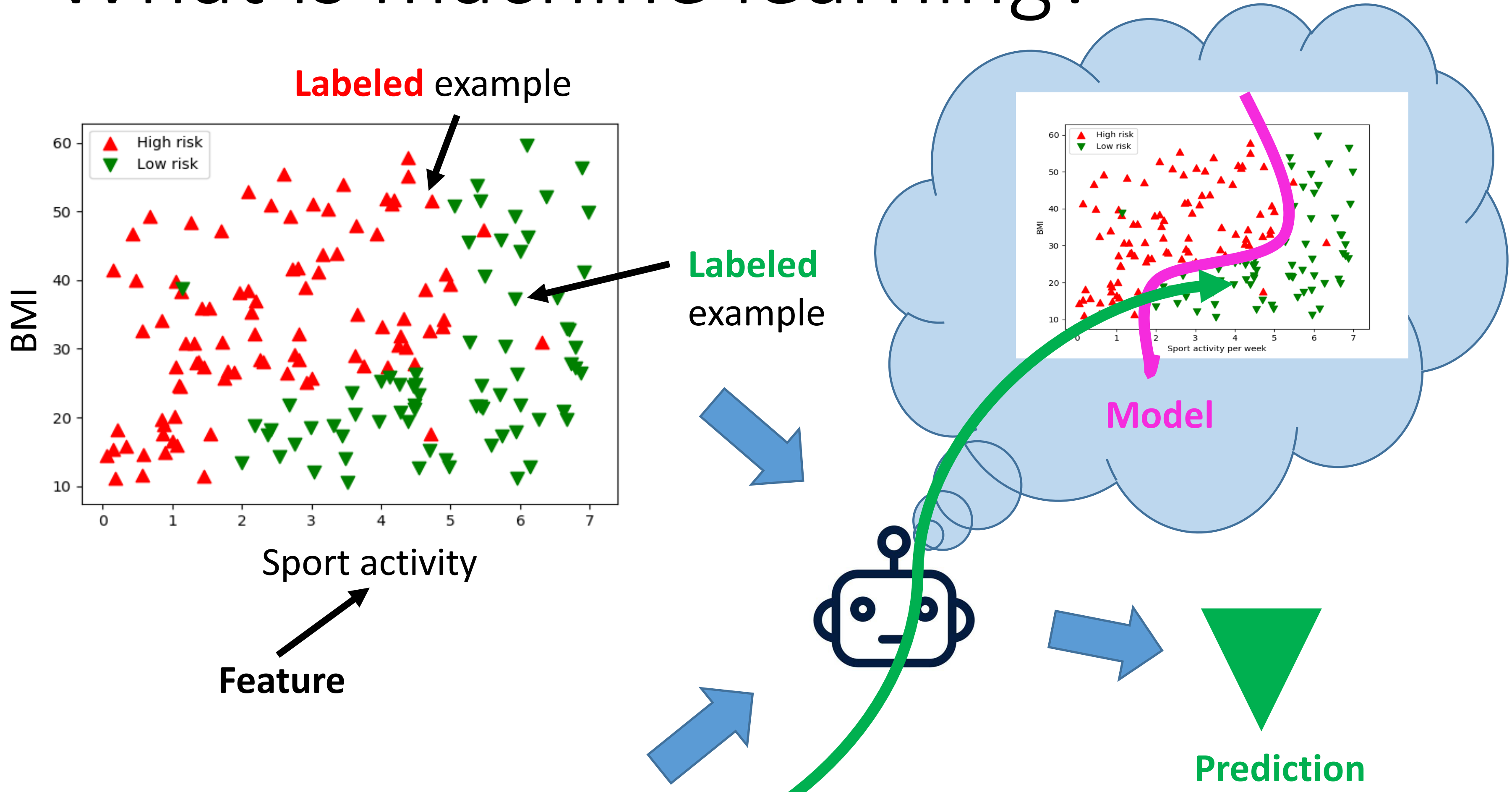
2019

Agenda

- What is machine learning?
- How models work
    - Classification trees
- Parameter selection via cross-validation

What is machine learning?

What is machine learning?



Labeled example

Labeled example

Feature

Model

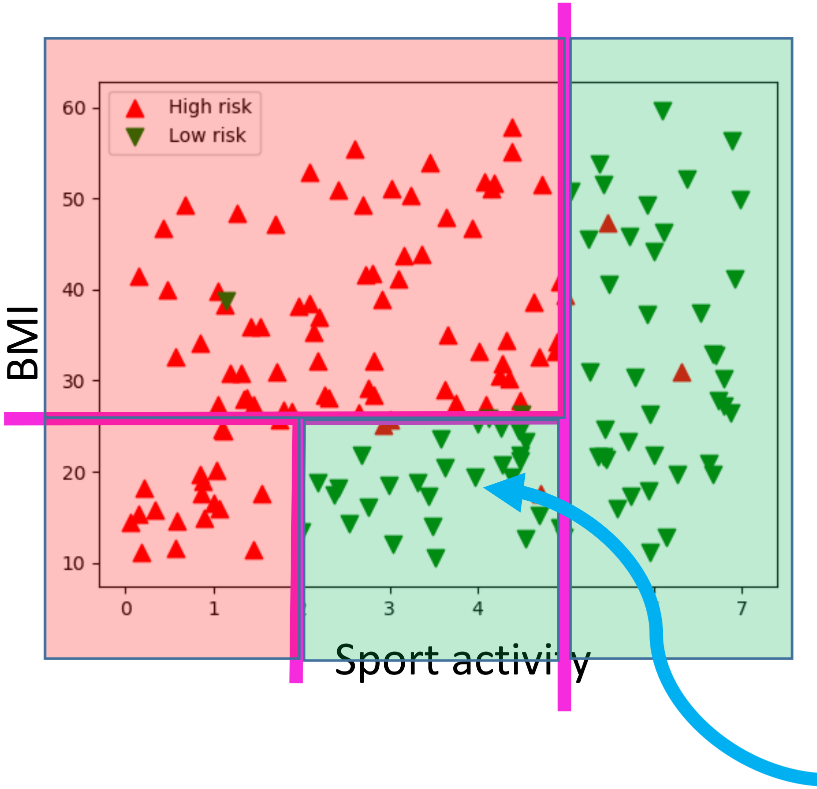
Prediction

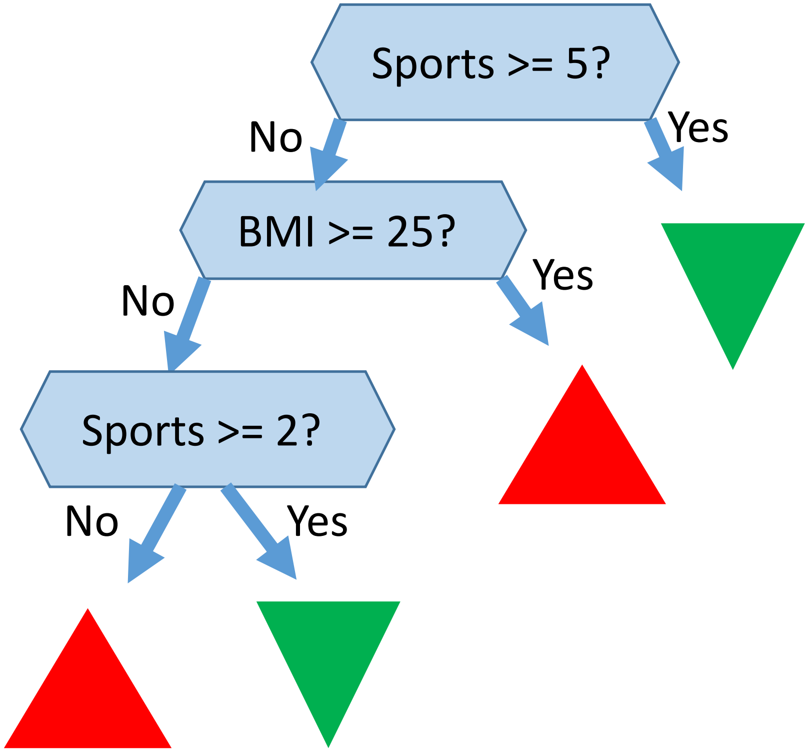
Name	BMI	Sport act.
Daniel	20	4

New (unlabeled) example

How models work

Classification trees




Sports >= 5?

No

Yes

BMI >= 25?

No

Yes

Sports >= 2?

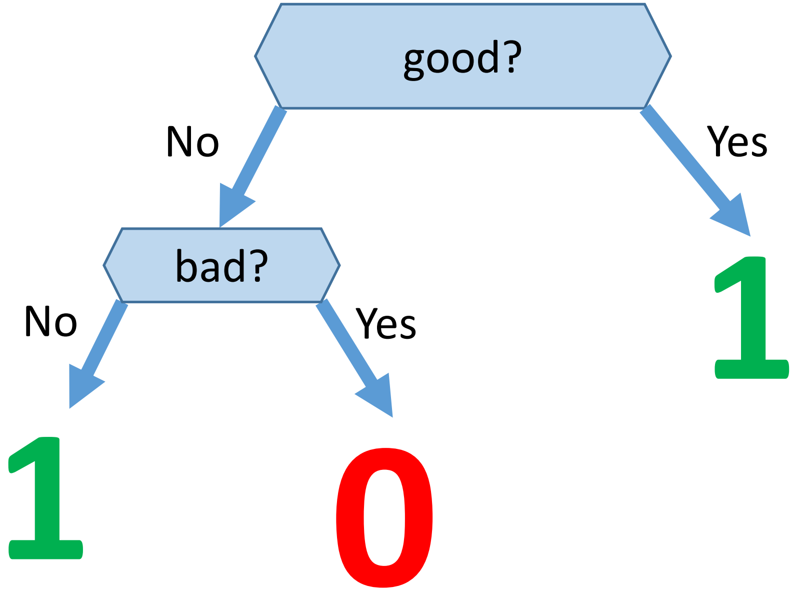
No

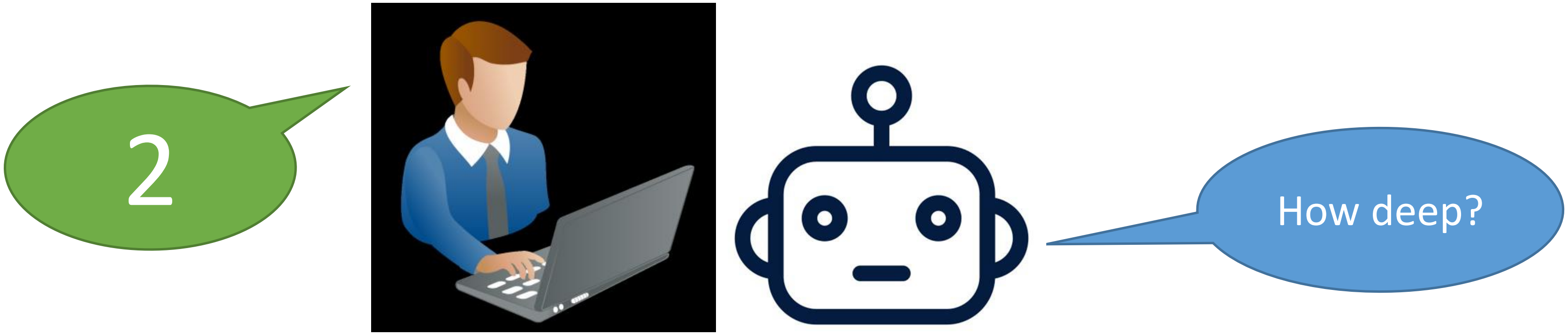
Yes

Name	BMI	Sport act.
Daniel	20	4

When training trees, you must specify their depth (and other parameters)

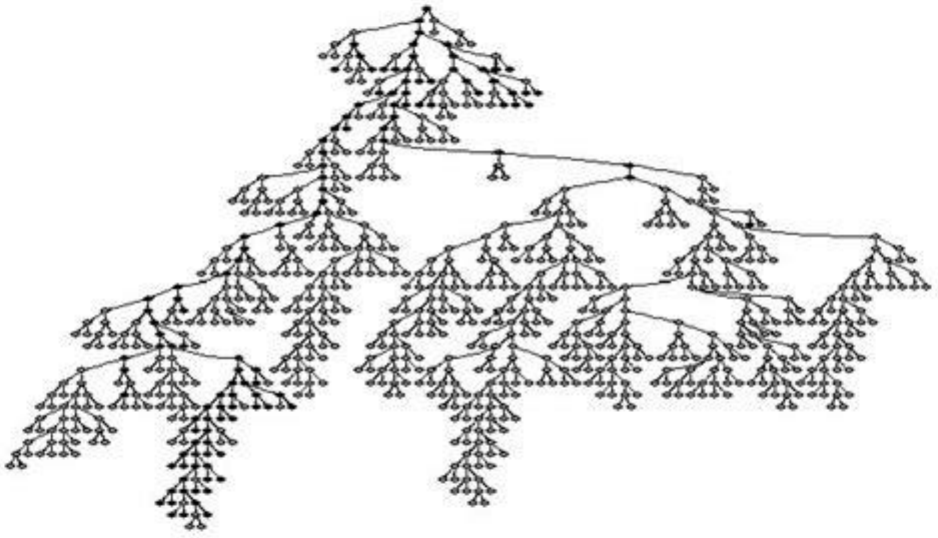
Review	Positive?
This is a good movie	1
What a good film!	1
Bad film	0
It was a bad movie	0

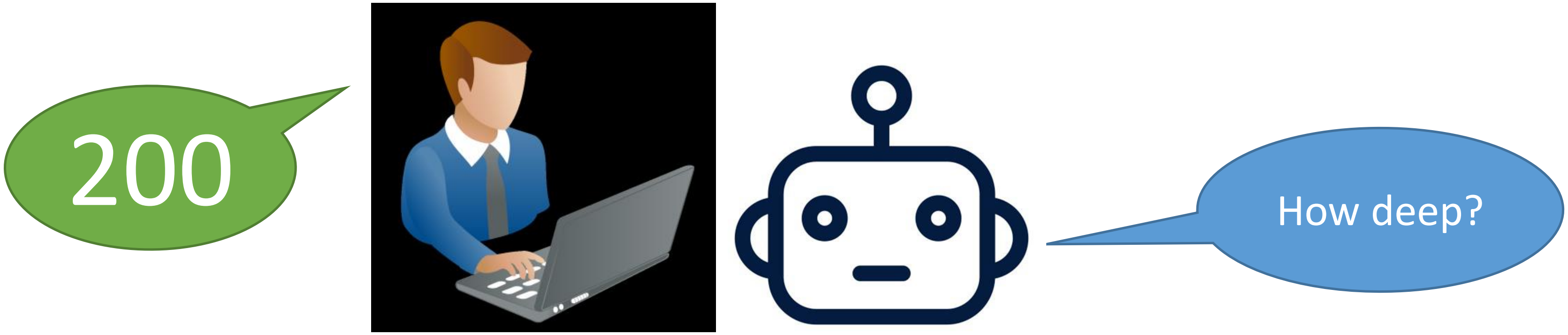




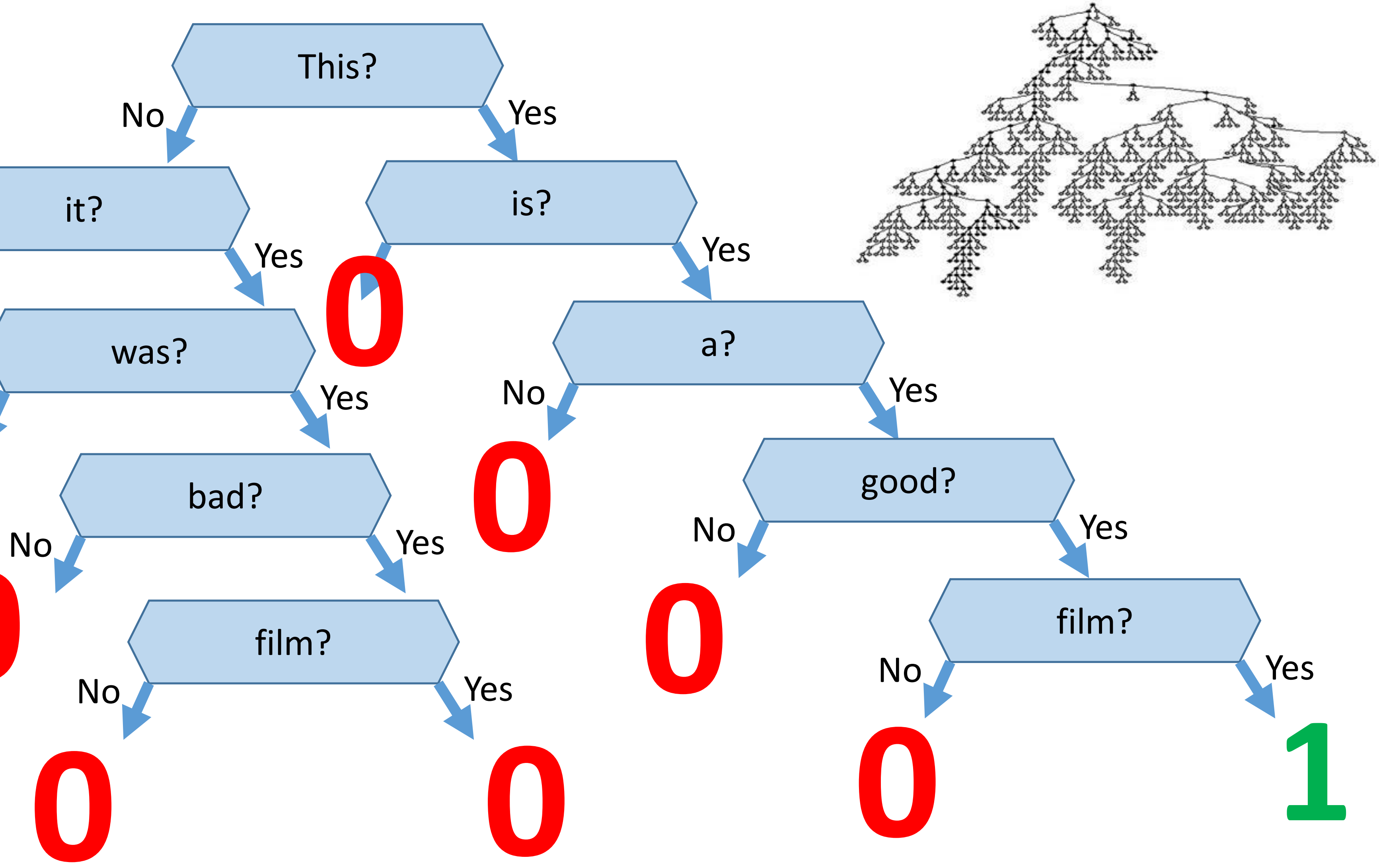
When training trees, you must specify their depth (and other parameters)

Review	Positive?
This is a good movie	1
What a good film!	1
Bad film	0
It was a bad movie	0





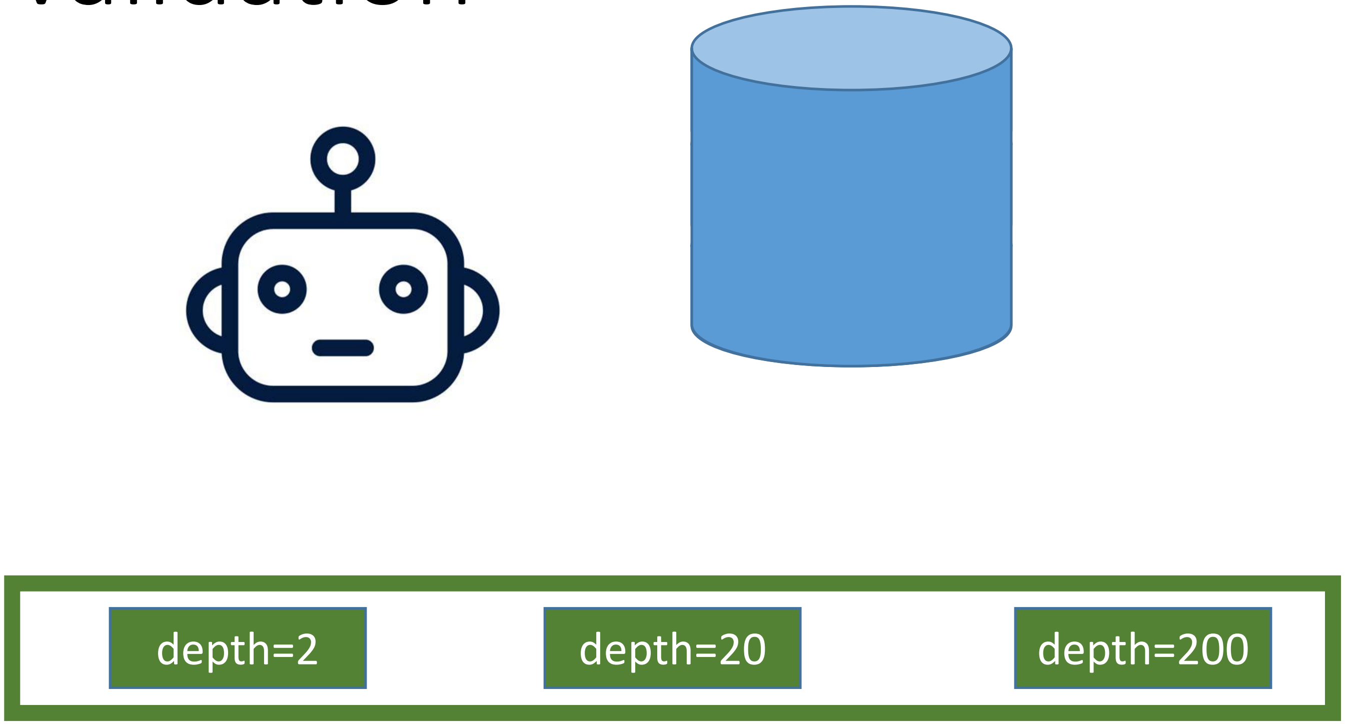
Overfitting: When complex models "memorize" the data



This?
No
Yes
it?
is?
Yes
Yes
0
was?
a?
Yes
No
Yes
bad?
good?
No
Yes
0
No
Yes
0
film?
film?
No
Yes
No
Yes
0
0
0
1

Grid search: Parameter selection by cross-validation

Parameter selection by cross-validation

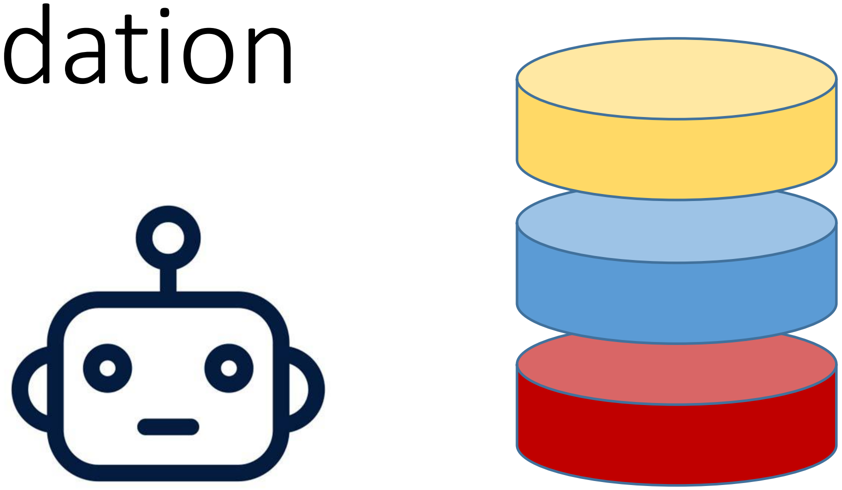


depth=2

depth=20

depth=200

Parameter selection by cross-validation

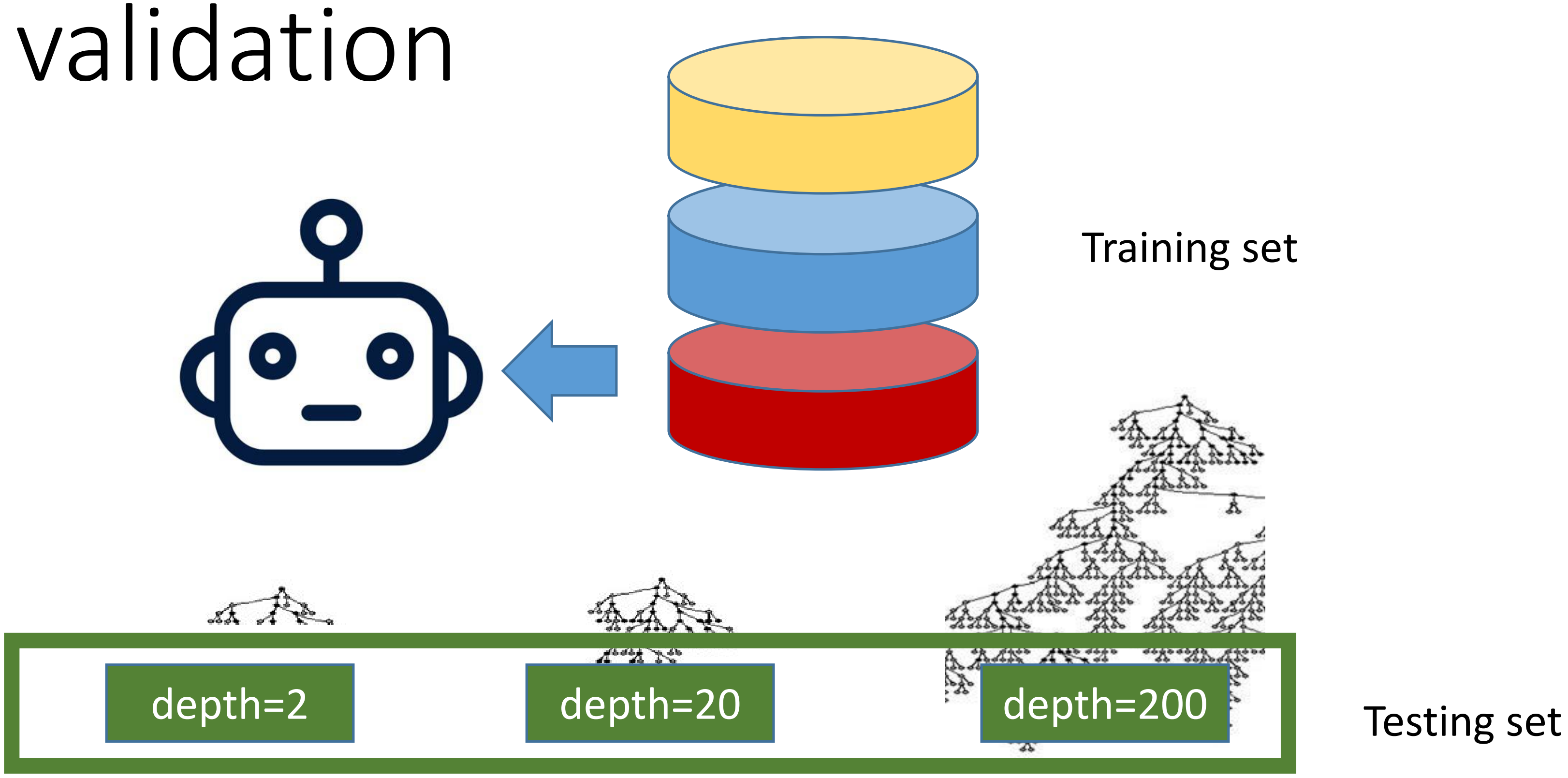


depth=2

depth=20

depth=200

Parameter selection by cross-validation



Training set

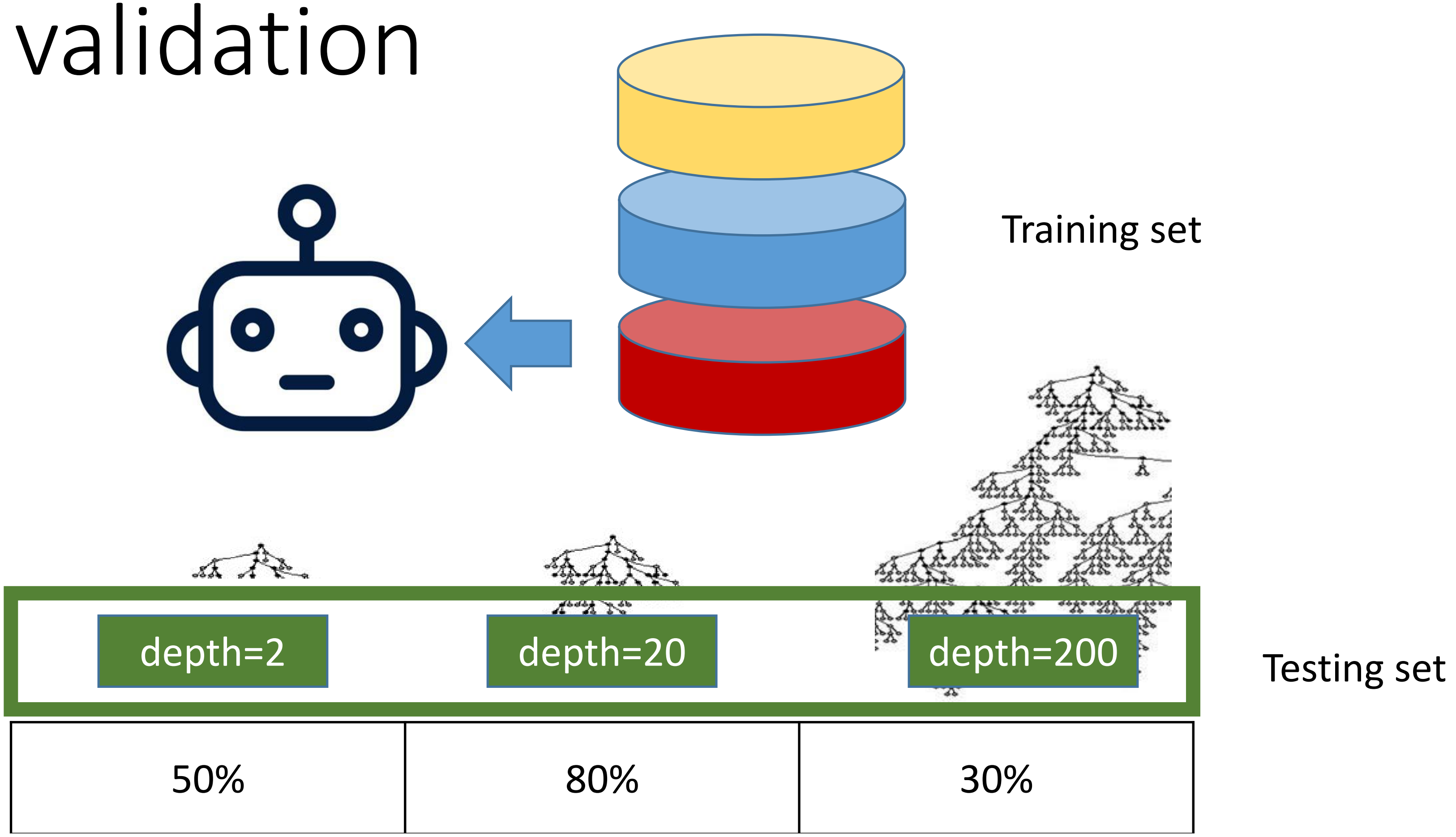
depth=2

depth=20

depth=200

Testing set

Parameter selection by cross-validation

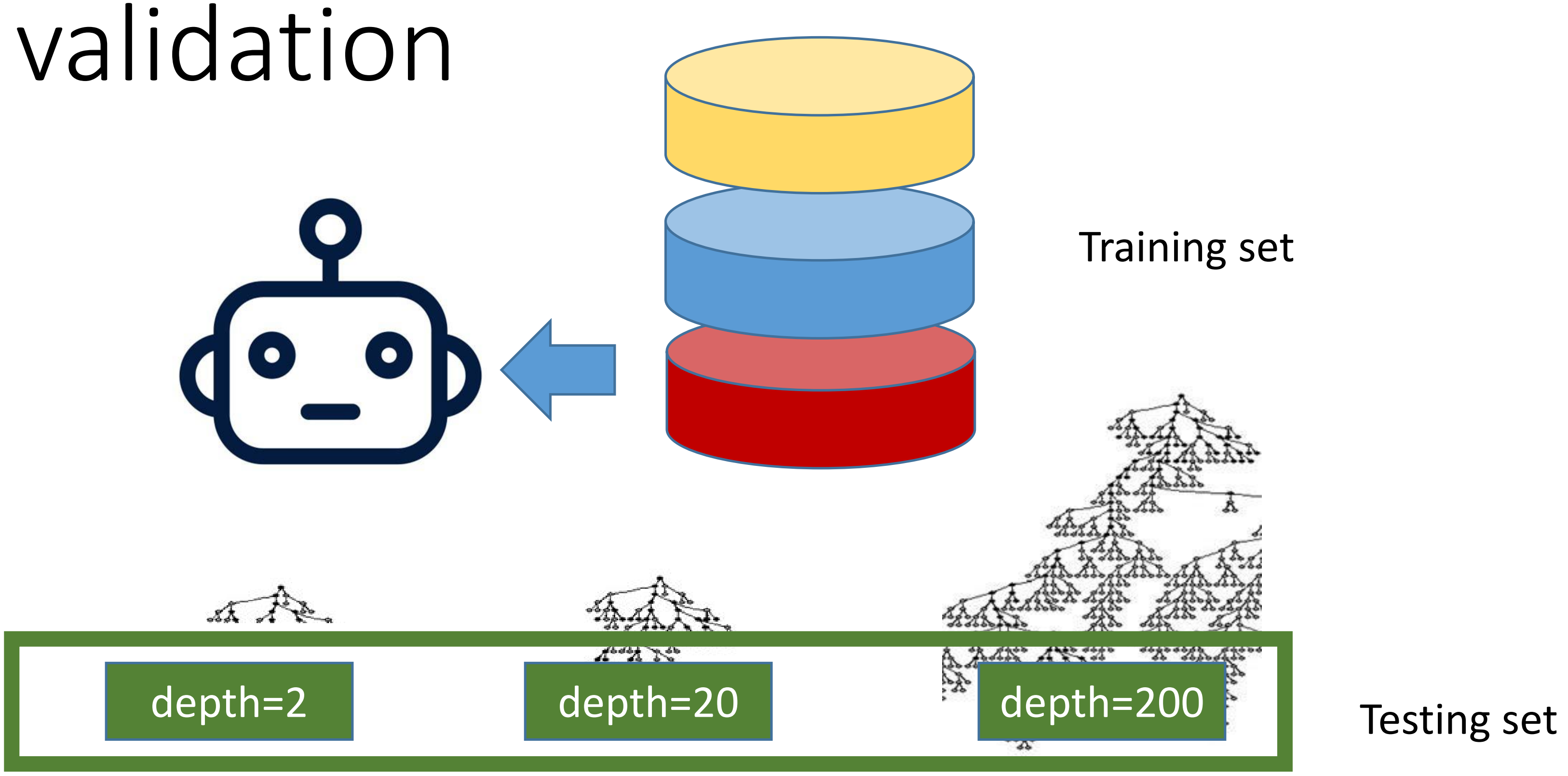


Training set

Testing set

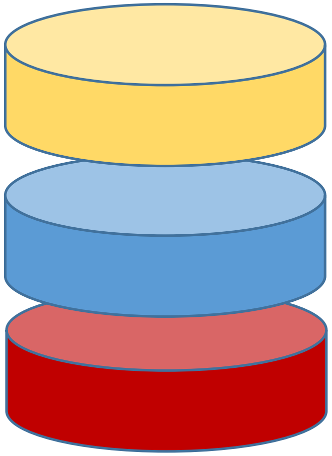
depth=2	depth=20	depth=200
50%	80%	30%

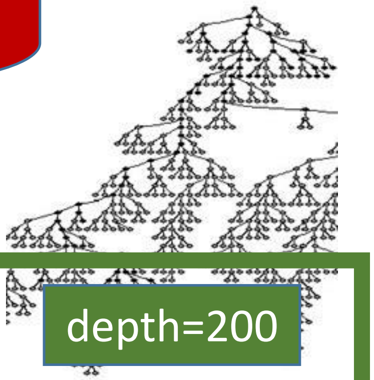
Parameter selection by cross-validation



50%	80%	30%
60%	90%	20%

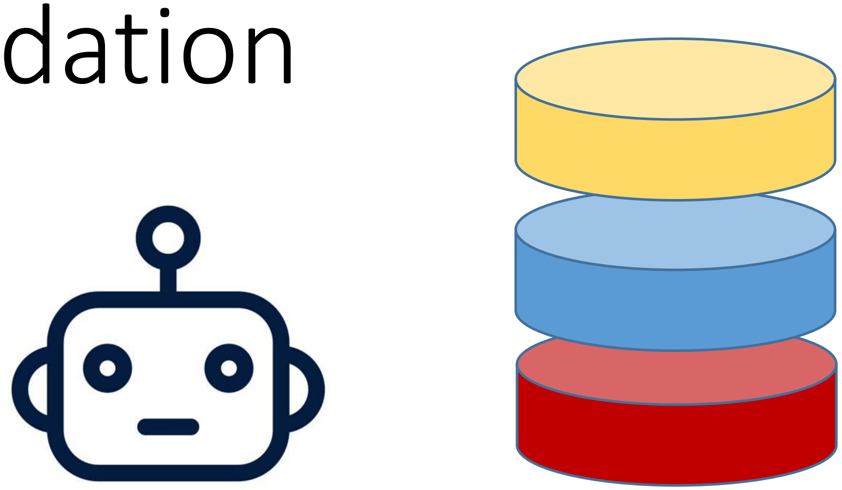
Parameter selection by cross-validation





depth=2	depth=20	depth=200
50%	80%	30%
60%	90%	20%
40%	90%	40%

Parameter selection by cross-validation



depth=2	depth=20	depth=200
50%	80%	30%
60%	90%	20%
40%	90%	40%

Flower classification

Iris setosa

Iris tectorum

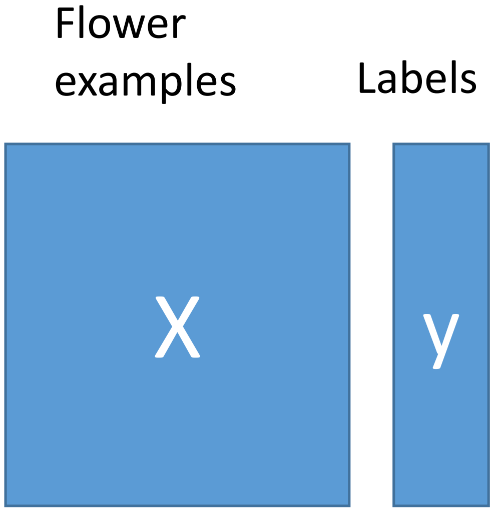
Iris latifolia

Data representation

Sepal length	Sepal width	Petal length	Petal width	Is setosa?
5.1	3.5	1.4	0.2	1
2.1	1.2	3.3	3.2	0
3.1	1.6	2.2	4.1	1
2.2	4.1	1.3	1.4	1

Data representation

- X[i, j]: Value of column j for flower i. (4 columns)
- y[i]: 1 if flower i is an iris setosa and 0 otherwise.



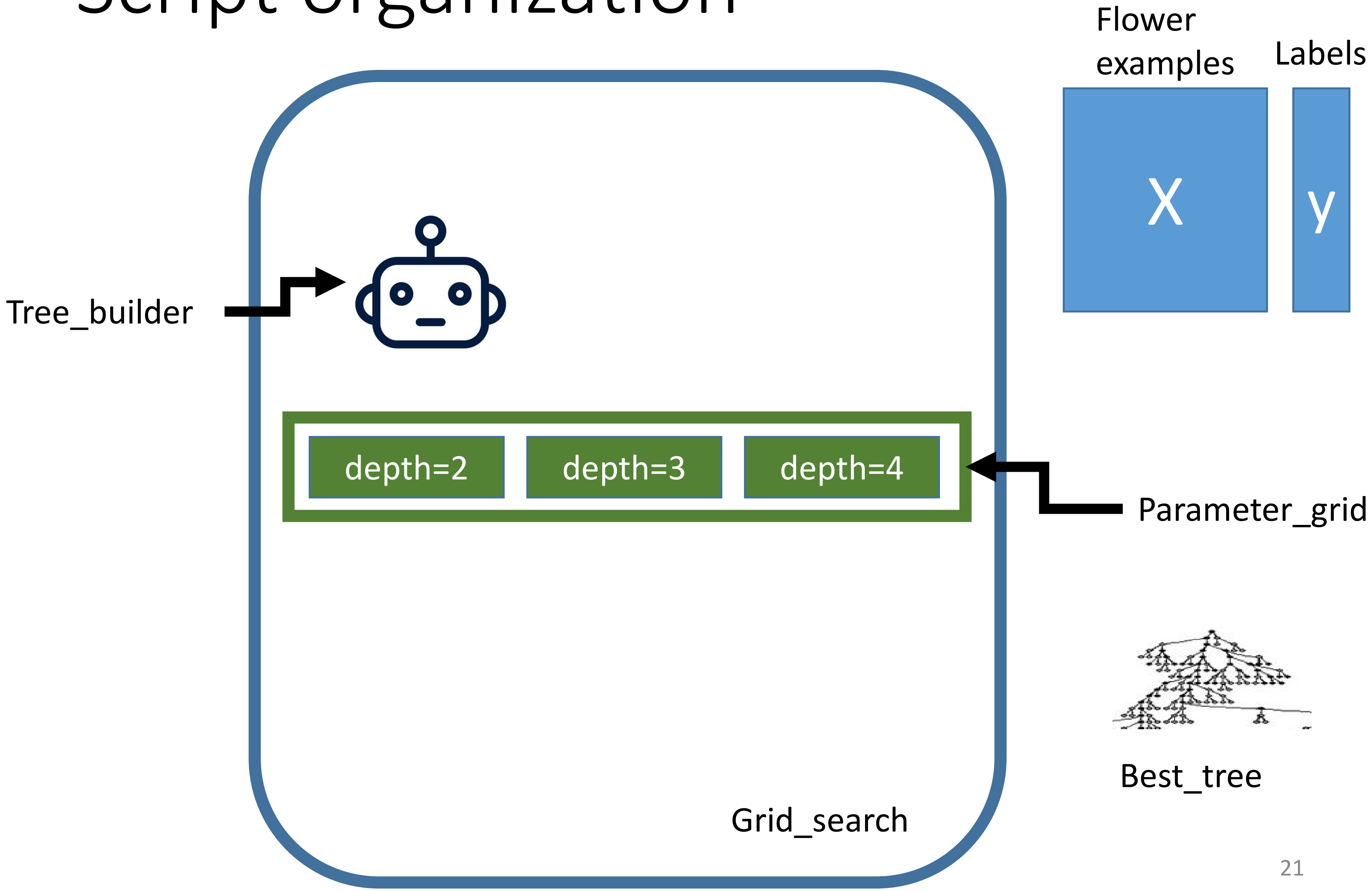
Flower examples

Labels

X

y

Script organization



Flower examples

Labels

X

y

Tree_builder

depth=2

depth=3

depth=4

Parameter_grid

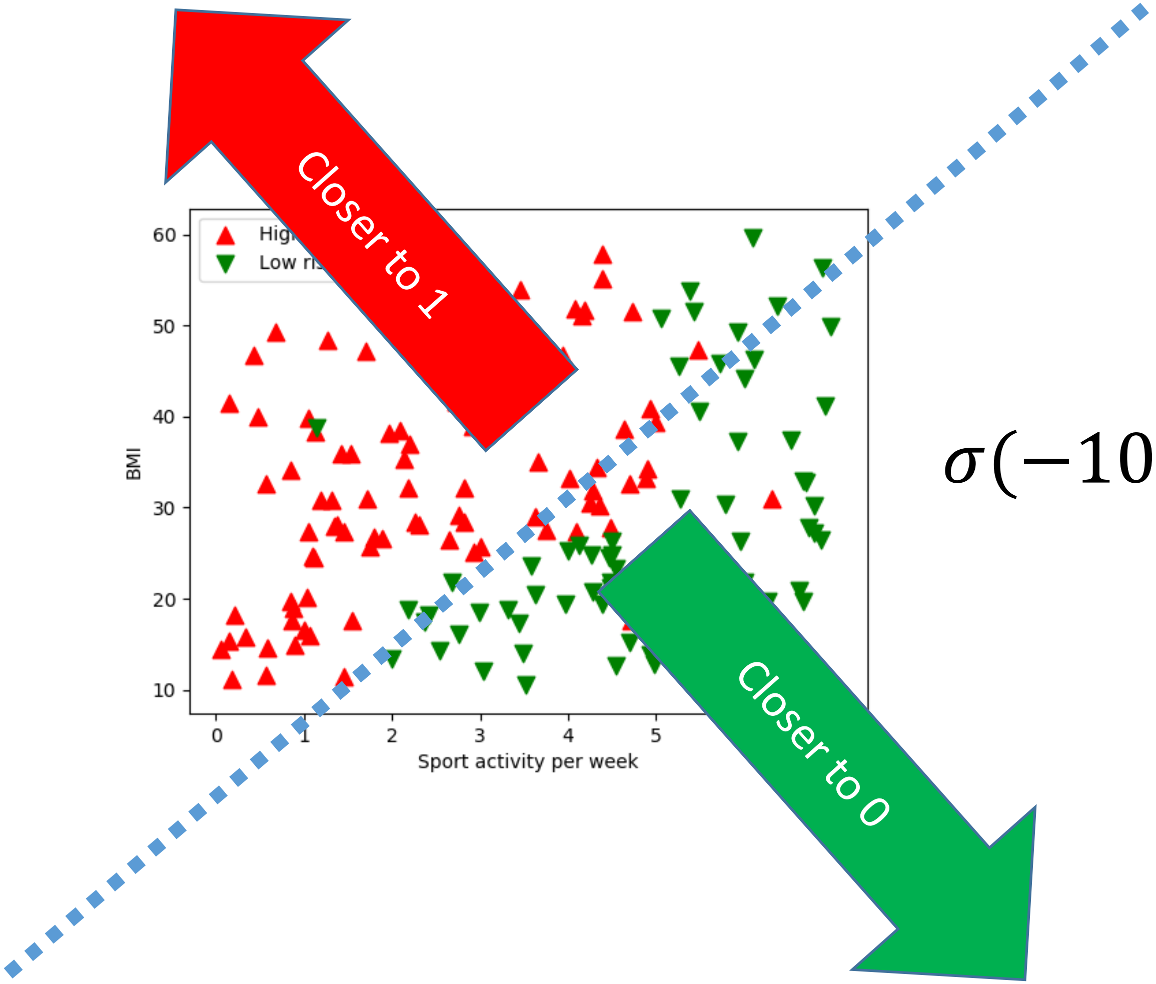
Best_tree

Grid_search

Agenda

- Other types of models:
    - Logistic models
    - Support-vector machines
    - Many others…
- How models are computed.
- How to deal with non-numeric data.

Logistic model

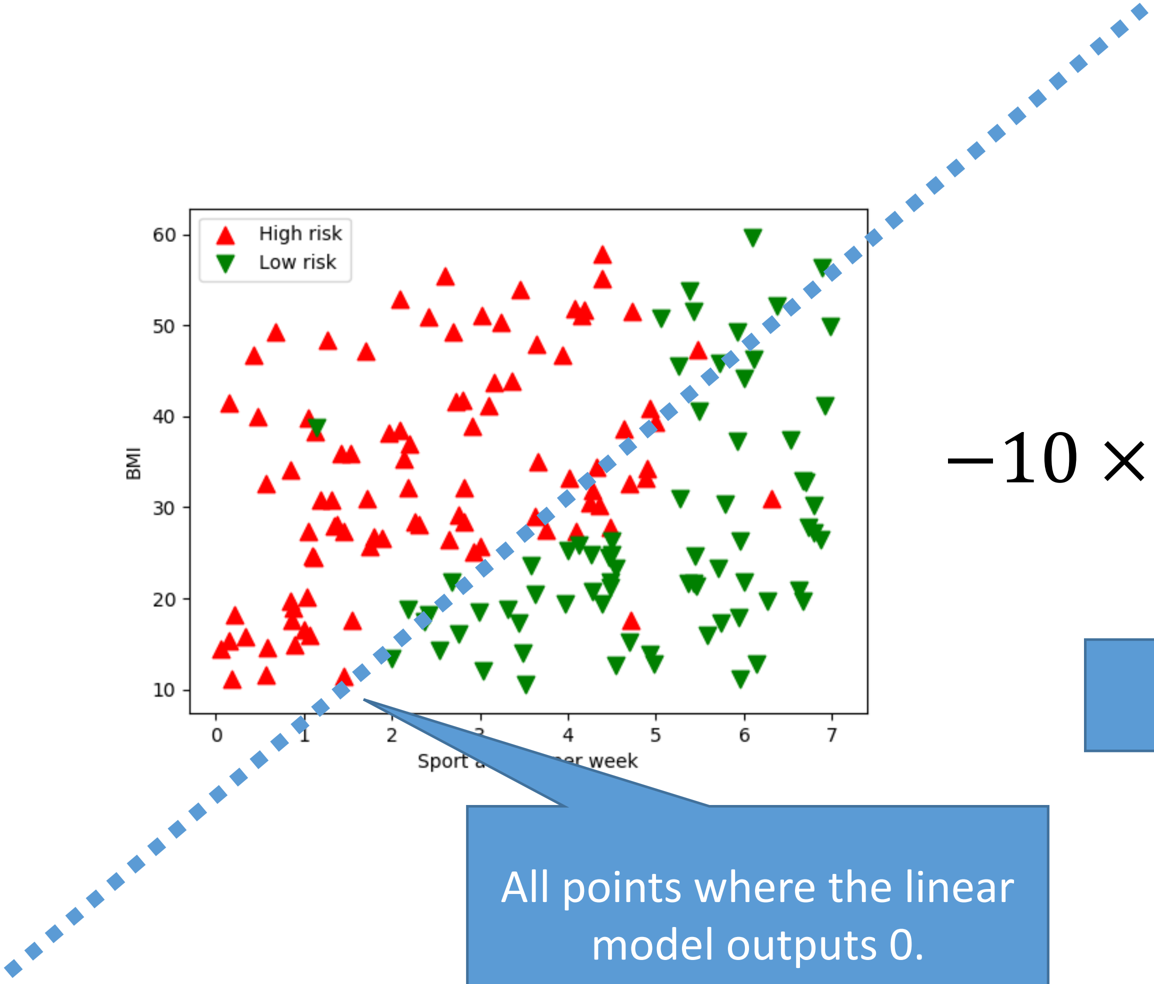


Closer to 1

Closer to 0

$$\sigma(-10 \times \boldsymbol{SW} + \boldsymbol{BMI} + 10)$$

Logistic model

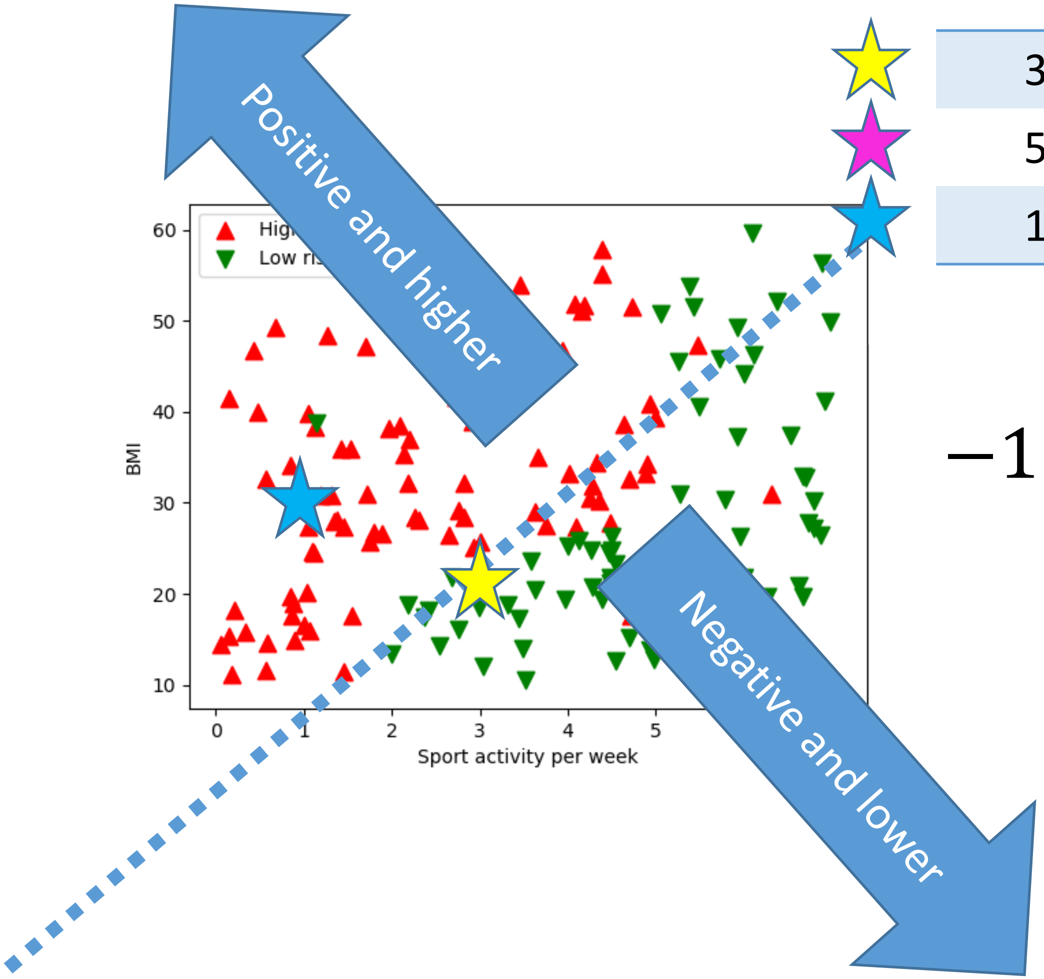


All points where the linear model outputs 0.

$$-10 \times \boldsymbol{SW} + \boldsymbol{BMI} + 10$$

Linear model

Logistic model



Positive and higher

Negative and lower

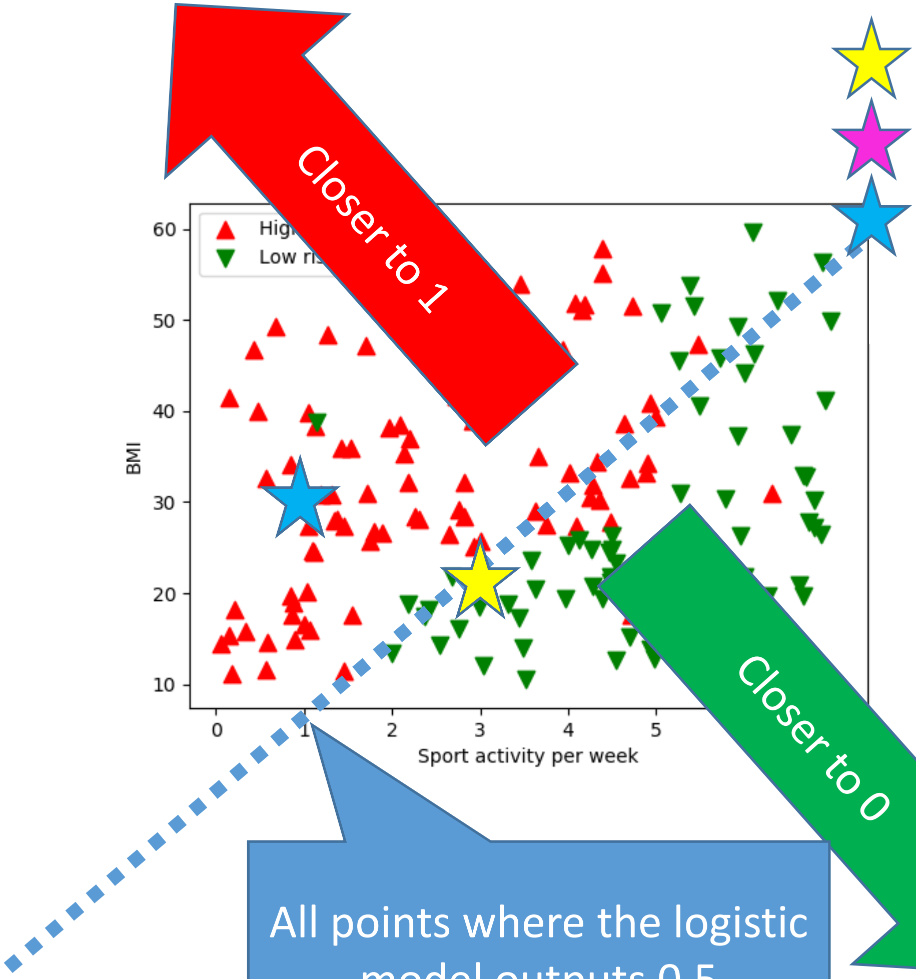
	SW	BMI
★	3	20
★	5	20
★	1	30

$$-10 \times \boldsymbol{SW} + \boldsymbol{BMI} + 10$$

Logistic model

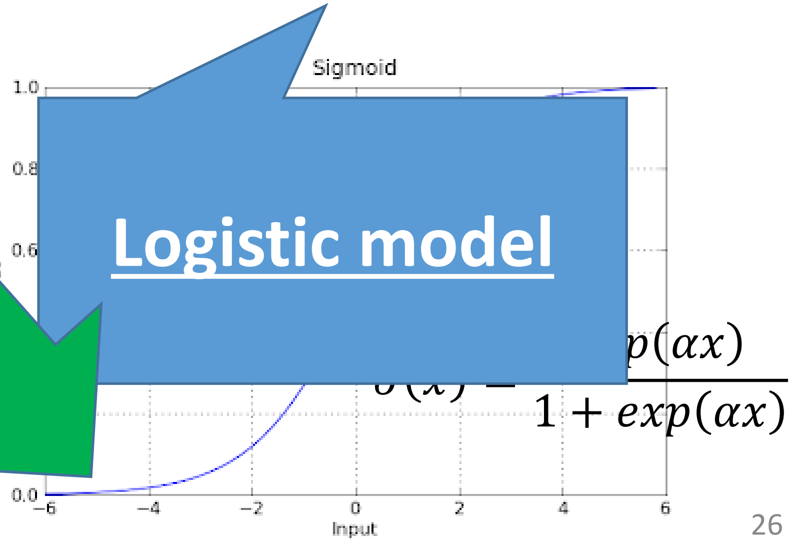
SW	BMI	Linear model
3	20	0
5	20	-20
1	30	30

$$\sigma(-10 \times SW + BMI + 10)$$



Closer to 1

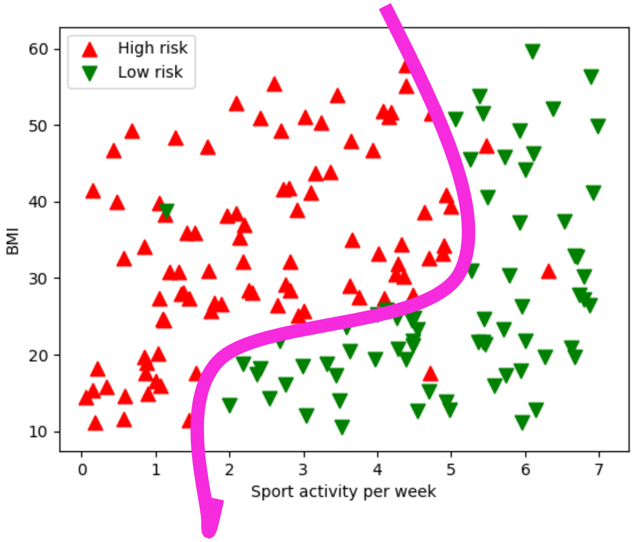
All points where the logistic model outputs 0.5



Closer to 0

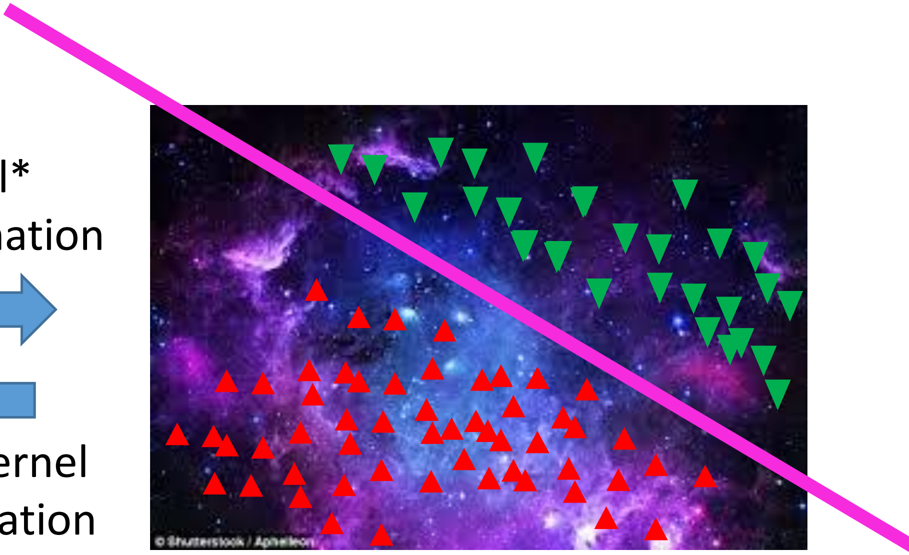
Logistic model

Support-vector machines



Kernel* transformation

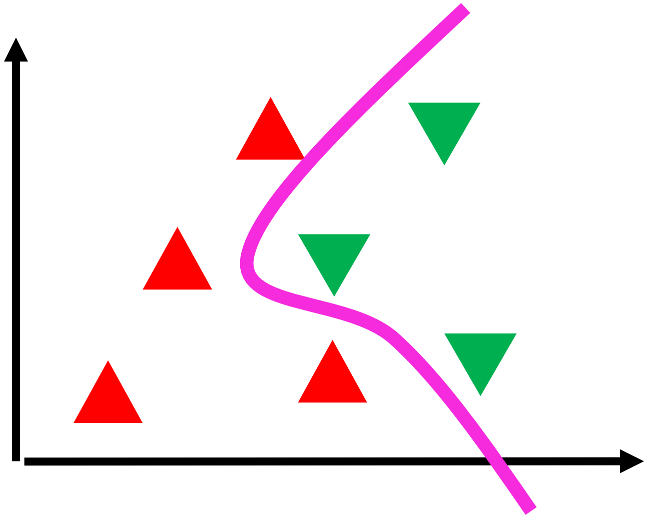
Inverse Kernel transformation

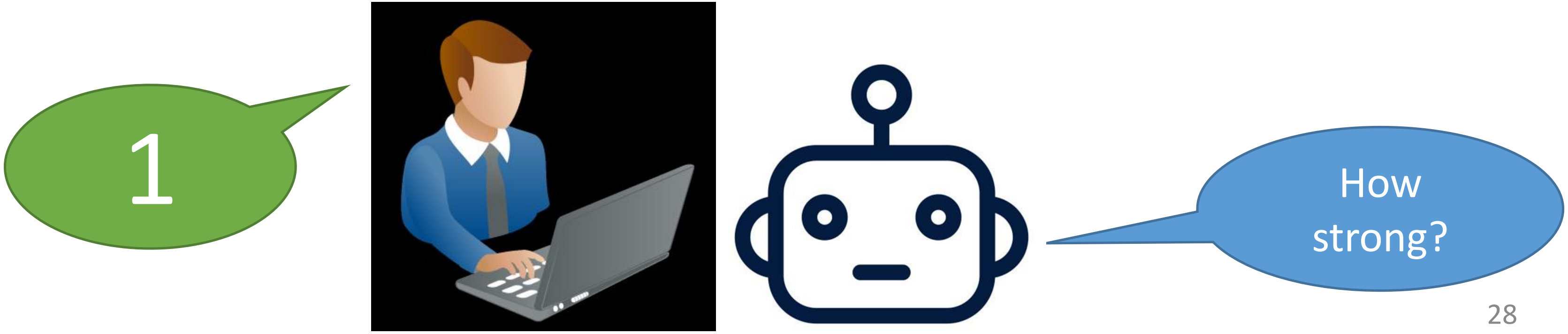


$$\sigma(-2 \times SW^3 \times BMI^2 + 4 \times BMI^3)$$

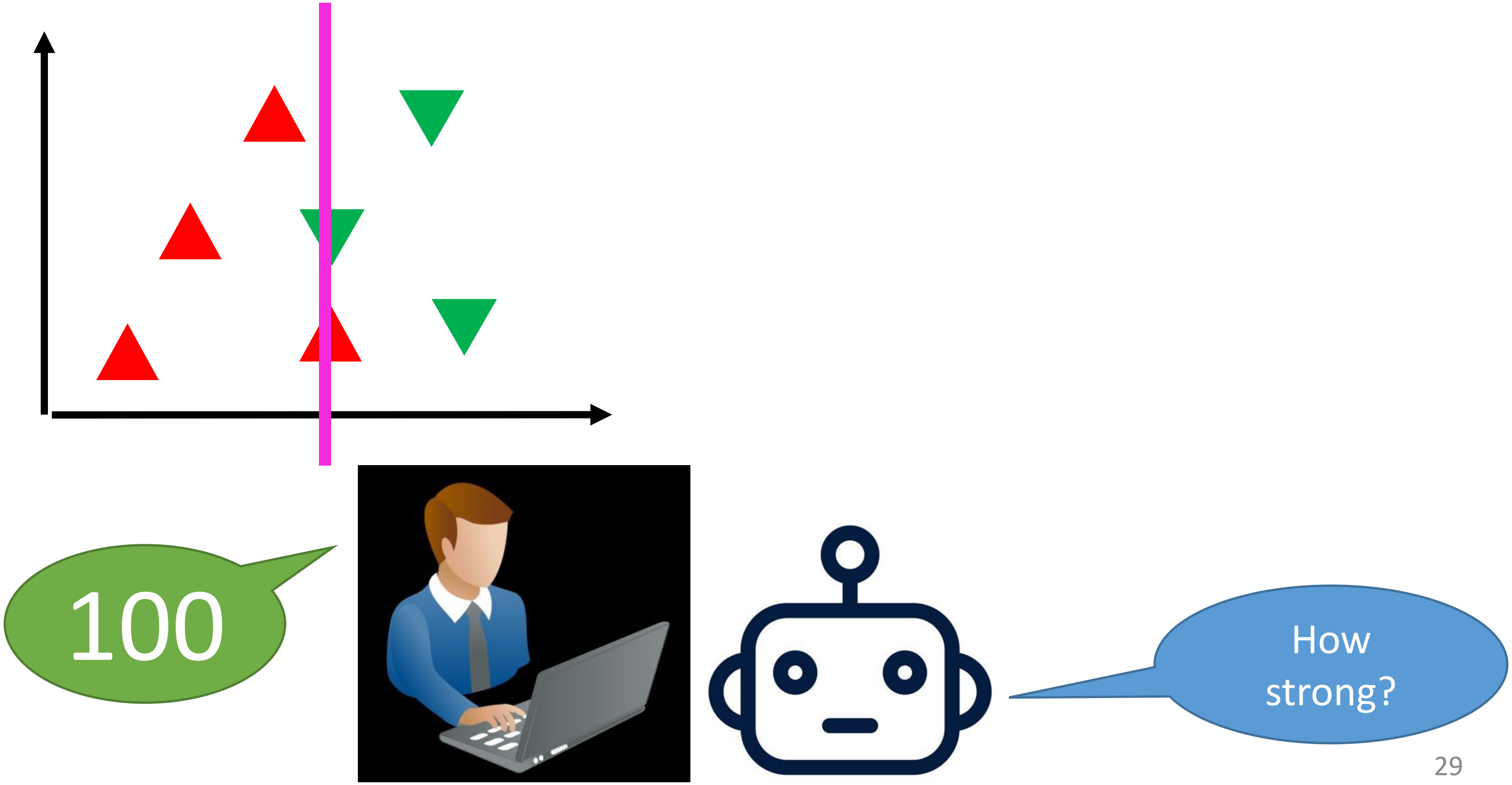
* Radial basis function kernel

When you train support-vector machines, you must specify the **regularization strengths** and other parameters.

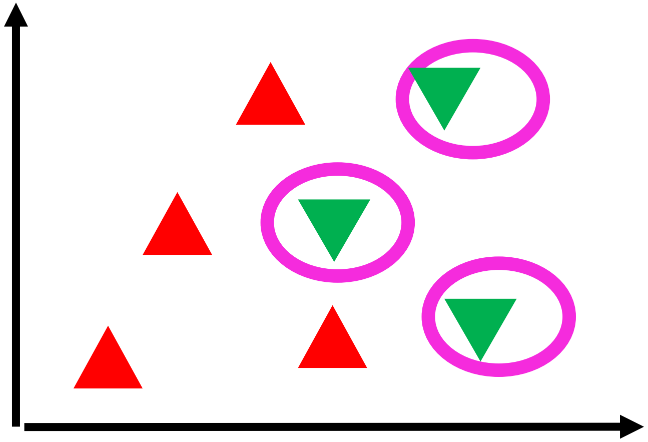




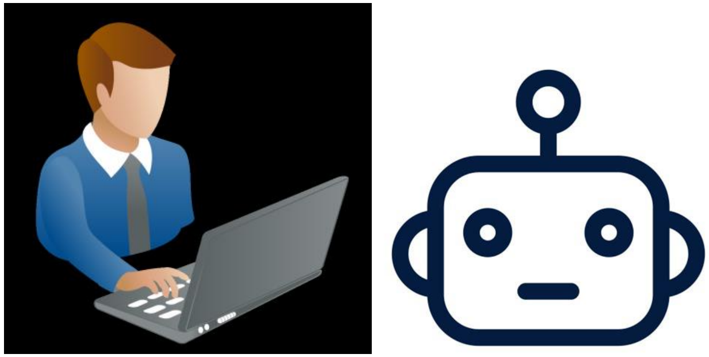
When you train support-vector machines, you must specify the **regularization strengths** and other parameters.



When you train support-vector machines, you must specify the **regularization strengths** and other parameters.



1e-5



How strong?

How to deal with non-numeric data and texts?

What to do if the data is not numeric?

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

One-hot encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

One-hot encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survive d?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survive d?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

Processing text data

Review	Positive review?
“Nice film”	1
“OK film”	1
“Bad movie”	0
“Terrible!”	0

Bag-of-words vectorization

Review	Positive review?
“Nice film”	1
“OK film”	1
“Bad movie”	0
“Terrible!”	0

bad	film	movie	nice	ok	terrible	Positive?
0	1	0	1	0	0	1
0	1	0	0	1	0	1
1	0	1	0	0	0	0
0	0	0	0	0	1	0

Conclusion

- What is machine learning?
- How models work and how to compute them
  - Classification trees
  - Logistic models
  - Support-vector machines
- How models are built
  - You don’t need to know how to build them in order to use them!
- How to deal with non-numeric data
- Parameter selection via cross-validation