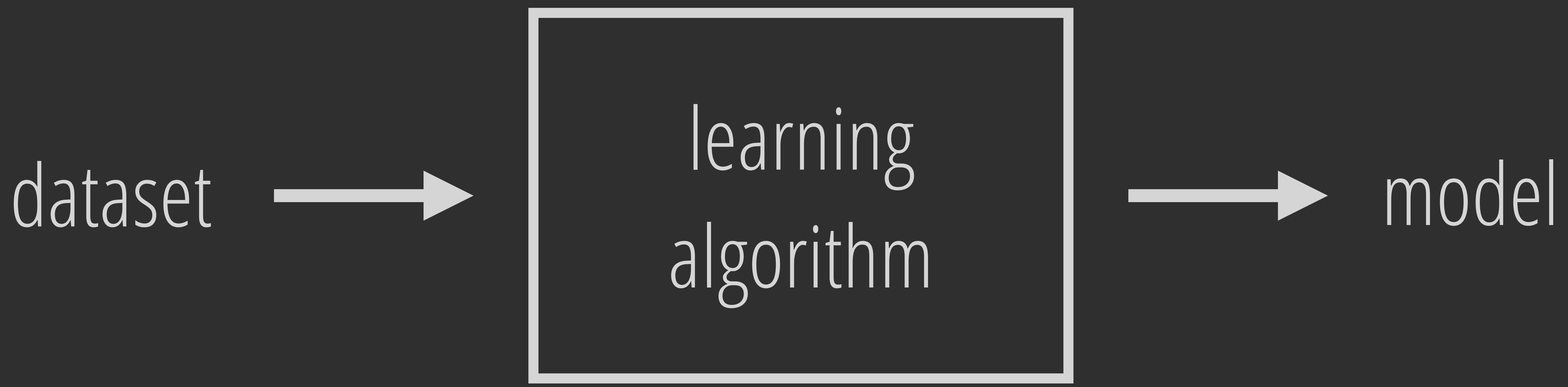
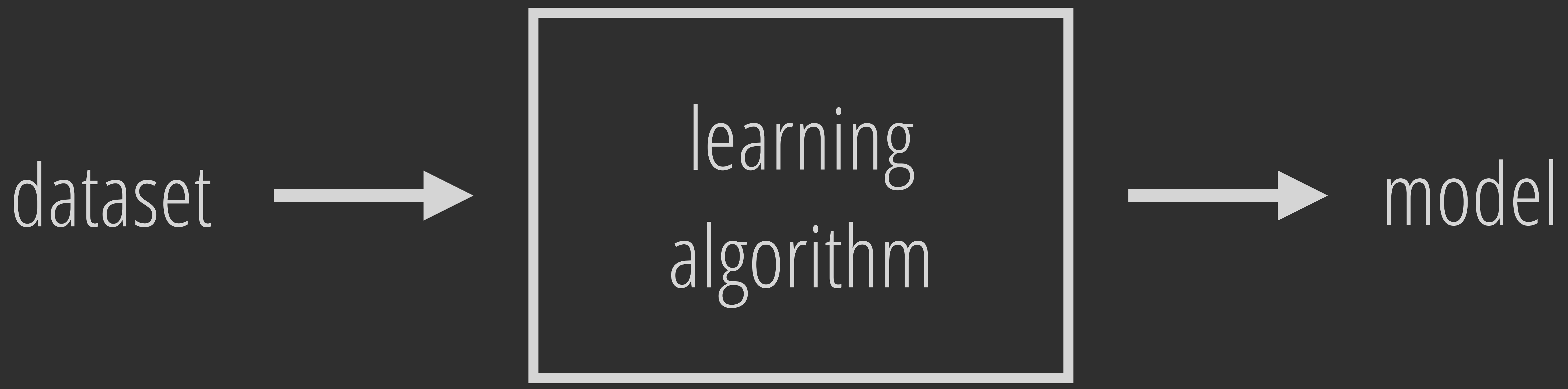
Certifying Robustness to Programmable Data Bias in Decision Tree Learning

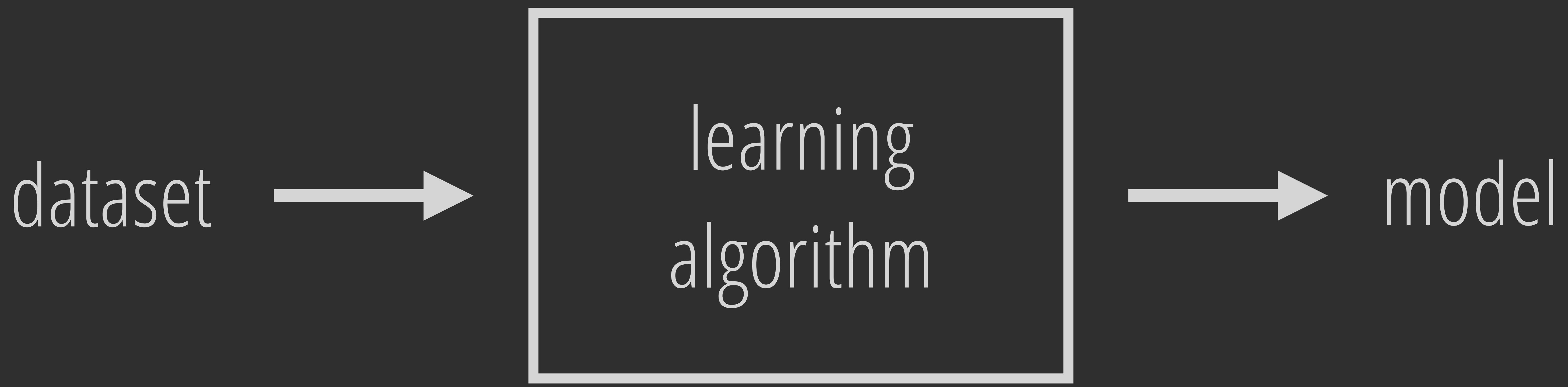
Anna P. Meyer

Joint work with Aws Albarghouthi and Loris D'Antoni


dataset
learning
algorithm
model



Is the dataset unbiased?
complete?
representative?



Is the dataset biased?
complete?
representative?

Probably not.

What is the impact on the model's predictions?

Goal: certify robustness to training-data bias

Types of dataset bias

- Label-flipping
- Missing data
- Fake data (data-poisoning)

Types of dataset bias

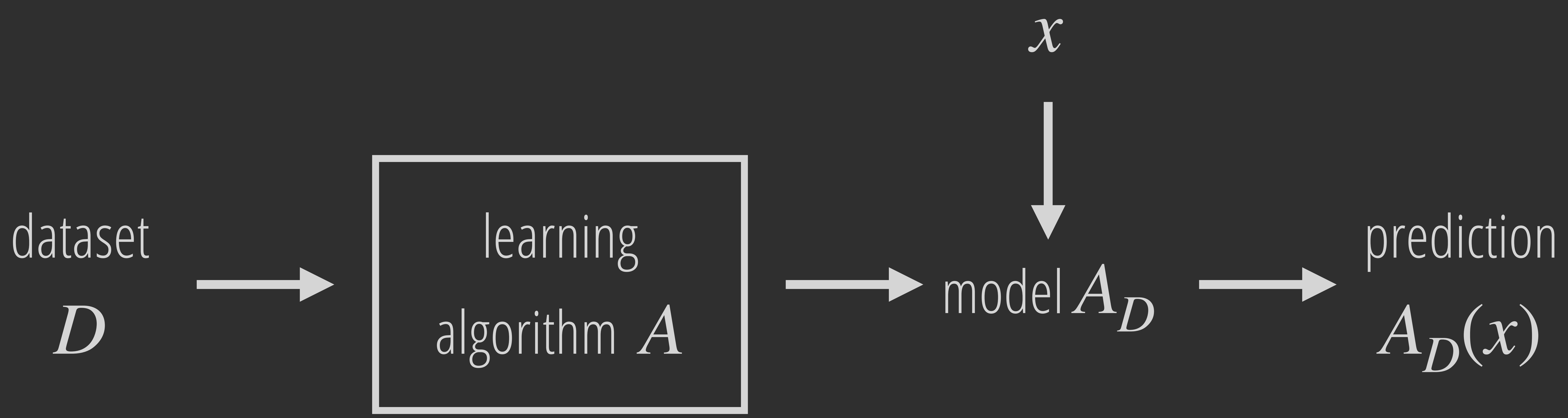
- Label-flipping
- Missing data
- Fake data (data-poisoning)

Assume fixed amount and type of data bias

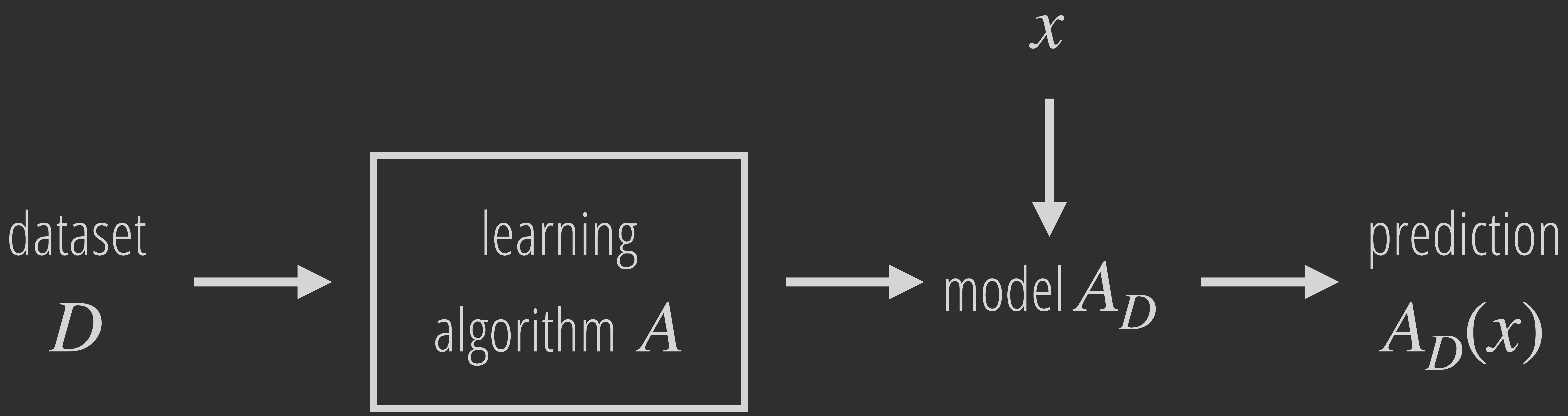
Types of dataset bias

- Label-flipping
- Missing data
- Fake data (data-poisoning)

Each type can be general or targeted



dataset
D
learning
algorithm A
model A_D
x
prediction
$A_D(x)$



bias robustness of x

for all D' that disagree with D on $\leq n$ labels

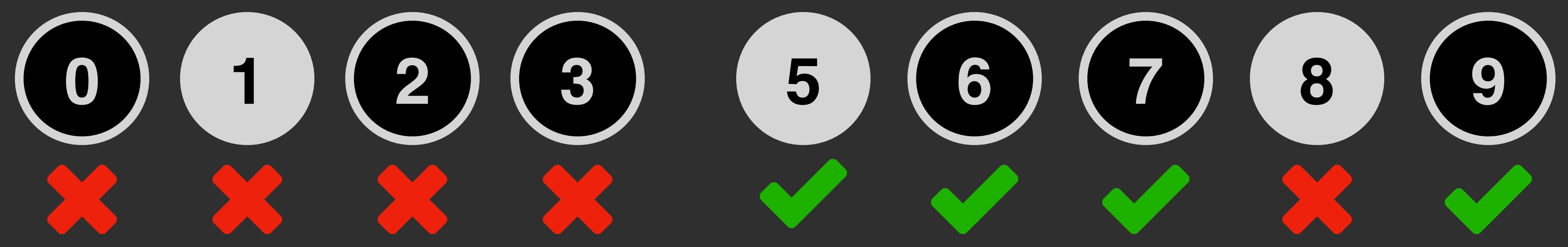
show that $A_{D'}(x) = A_D(x)$

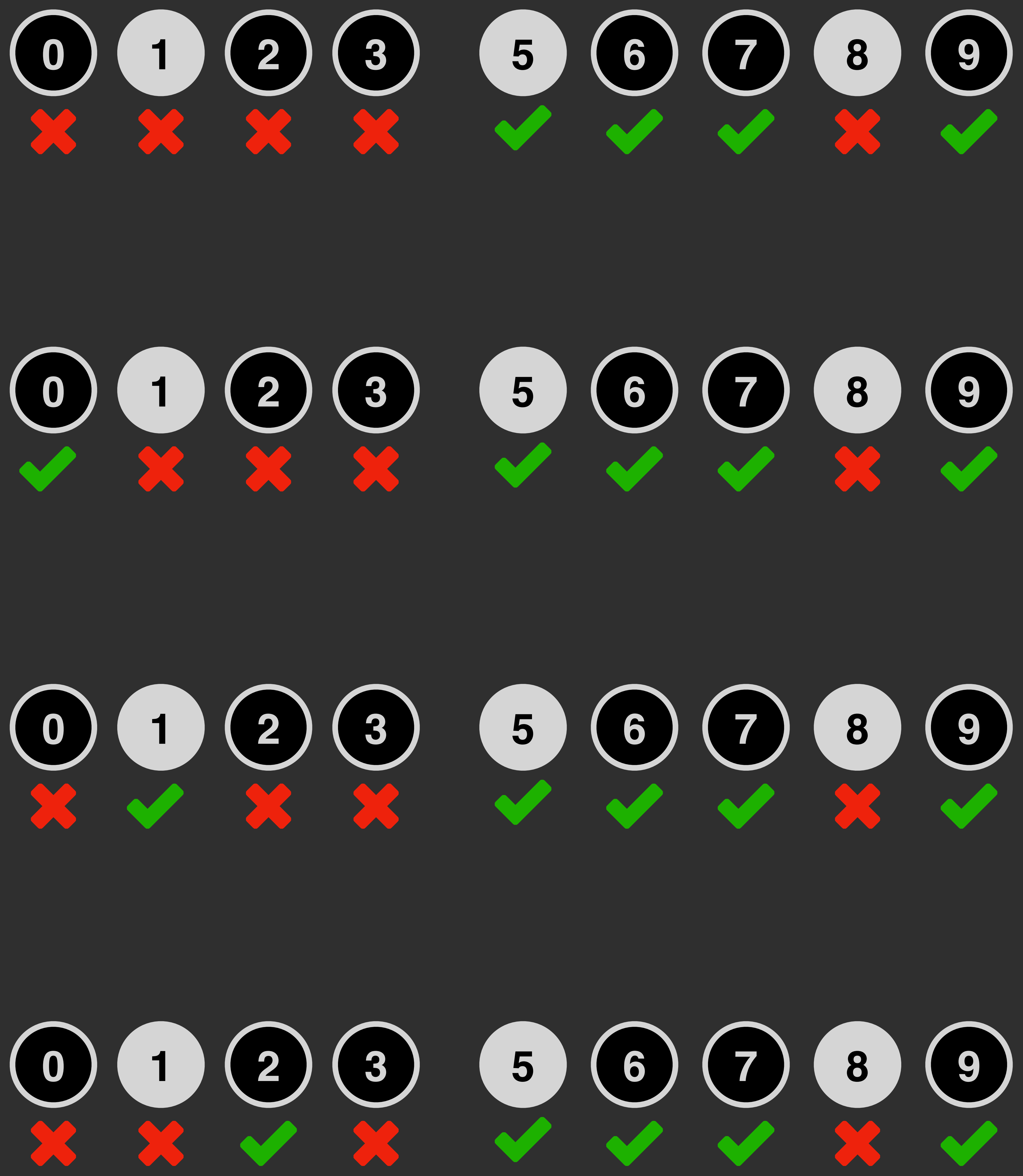
bias robustness of x

for all D' that disagree with D on $\leq n$ labels

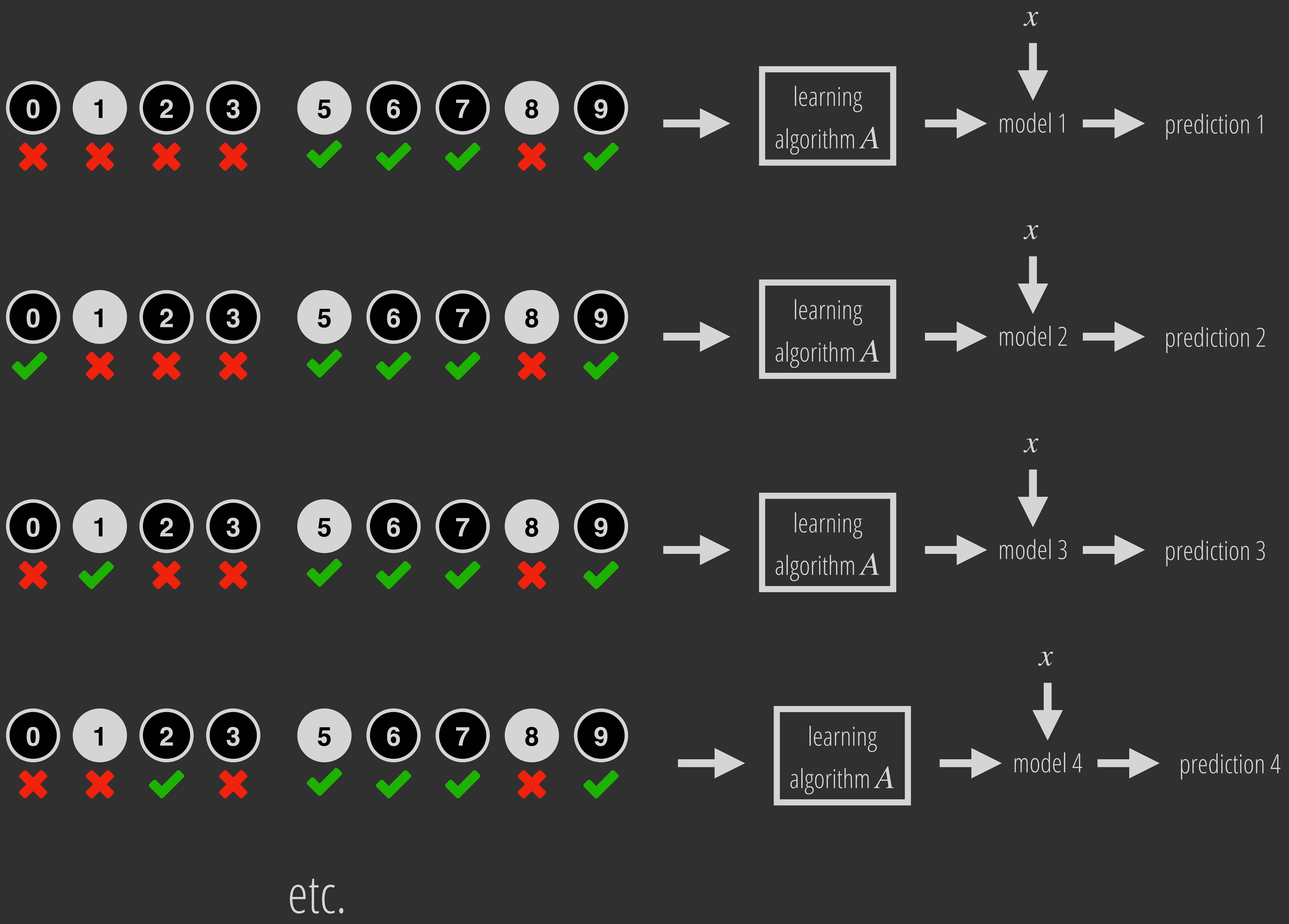
show that $A_{D'}(x) = A_D(x)$

Dataset D





etc.


0 1 2 3 5 6 7 8 9
learning algorithm A
x
model 1
prediction 1
0 1 2 3 5 6 7 8 9
learning algorithm A
x
model 2
prediction 2
0 1 2 3 5 6 7 8 9
learning algorithm A
x
model 3
prediction 3
0 1 2 3 5 6 7 8 9
learning algorithm A
x
model 4
prediction 4
etc.

$|D| = 1000$

$n = 10$

$\sim 10^{23}$ datasets!

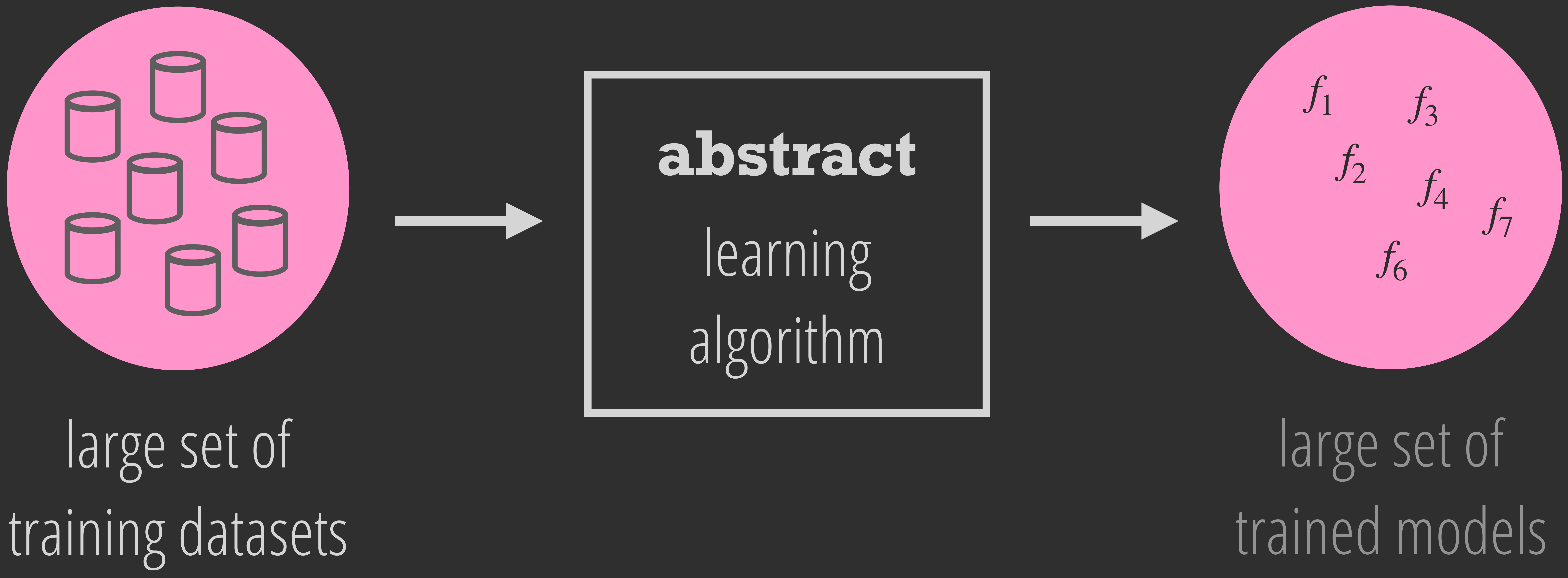
bias robustness of x

for all D' that disagree with D on $\leq n$ labels

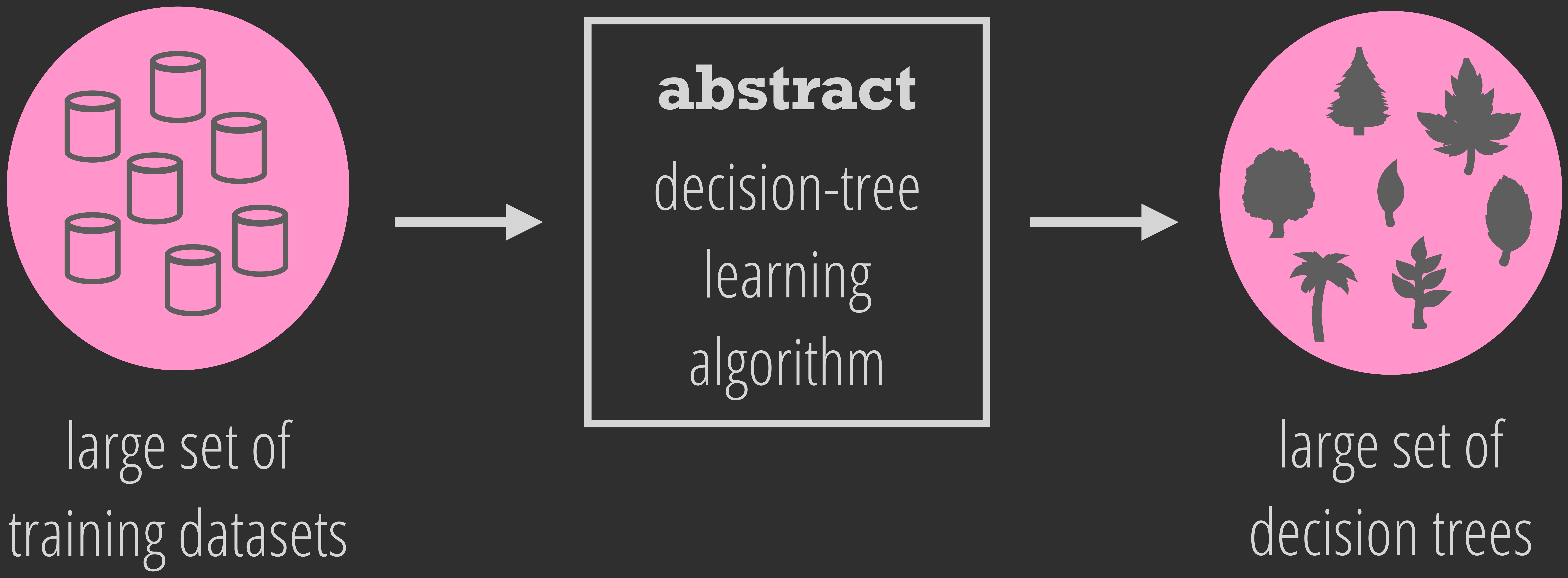
show that $A_{D'}(x) = A_D(x)$

Key challenge

Combinatorial explosion in the number of datasets



f_1
f_3
f_2
f_4
f_7
f_6
abstract
learning
algorithm
large set of
training datasets
large set of
trained models



large set of training datasets
abstract
decision-tree learning algorithm
large set of decision trees

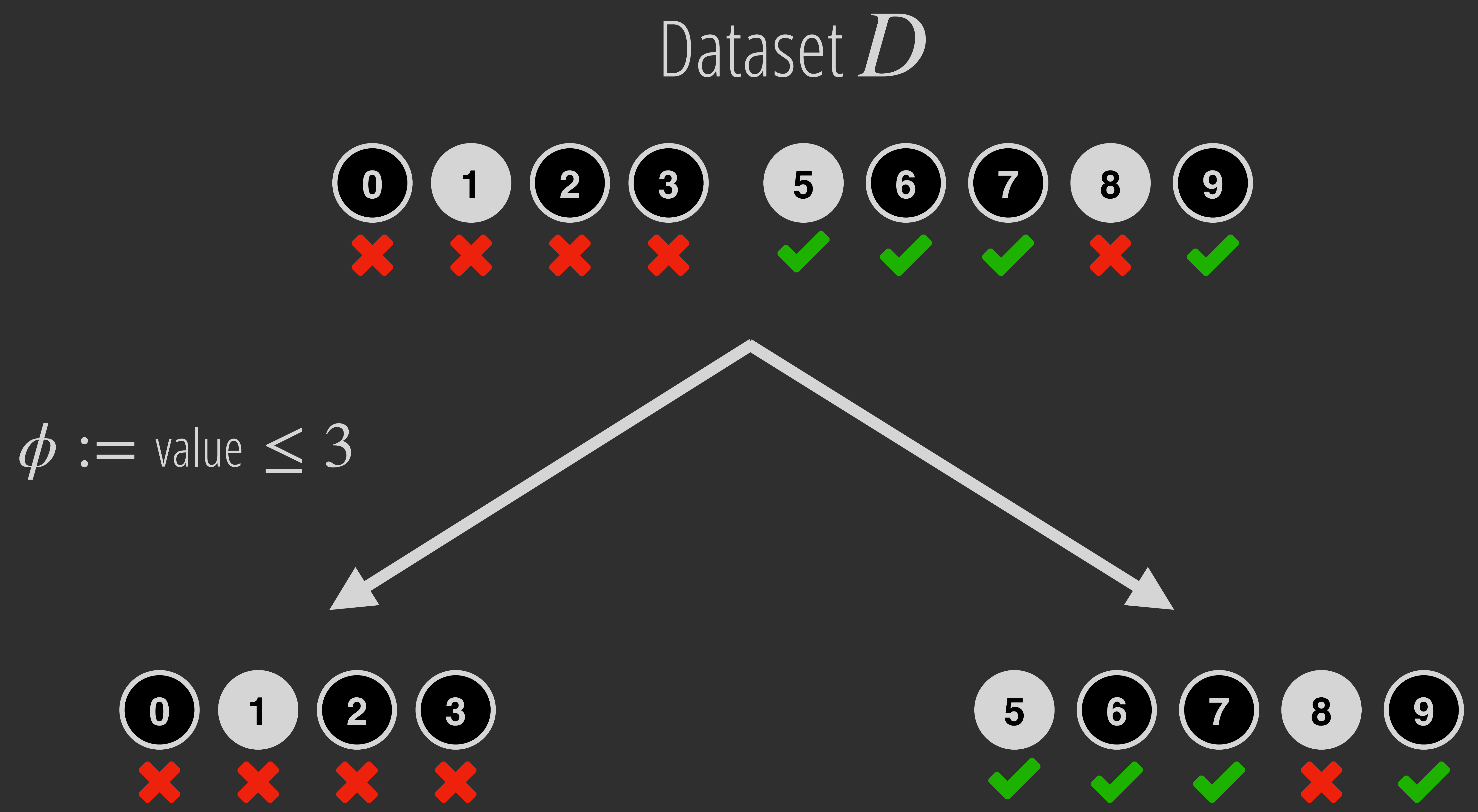
(Very) simplified decision tree algorithm

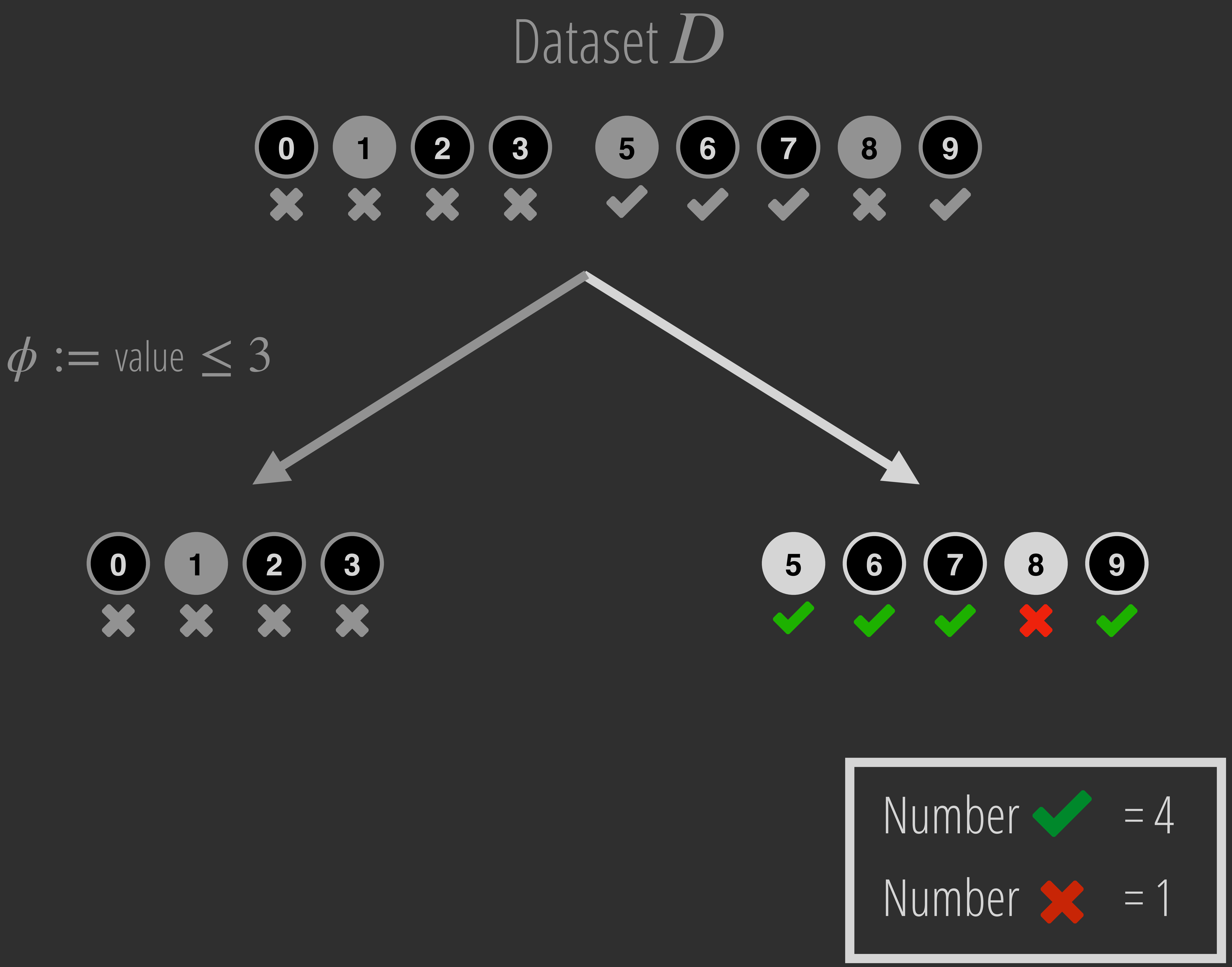
1. Choose the predicate that minimizes entropy on the data
2. Split the data according to that predicate
3. Repeat on child nodes until entropy is 0, or maximum depth is reached

(Very) simplified decision tree algorithm

1. Choose the predicate that minimizes entropy on the data
2. Split the data according to that predicate
3. Repeat on child nodes until entropy is 0, or maximum depth is reached

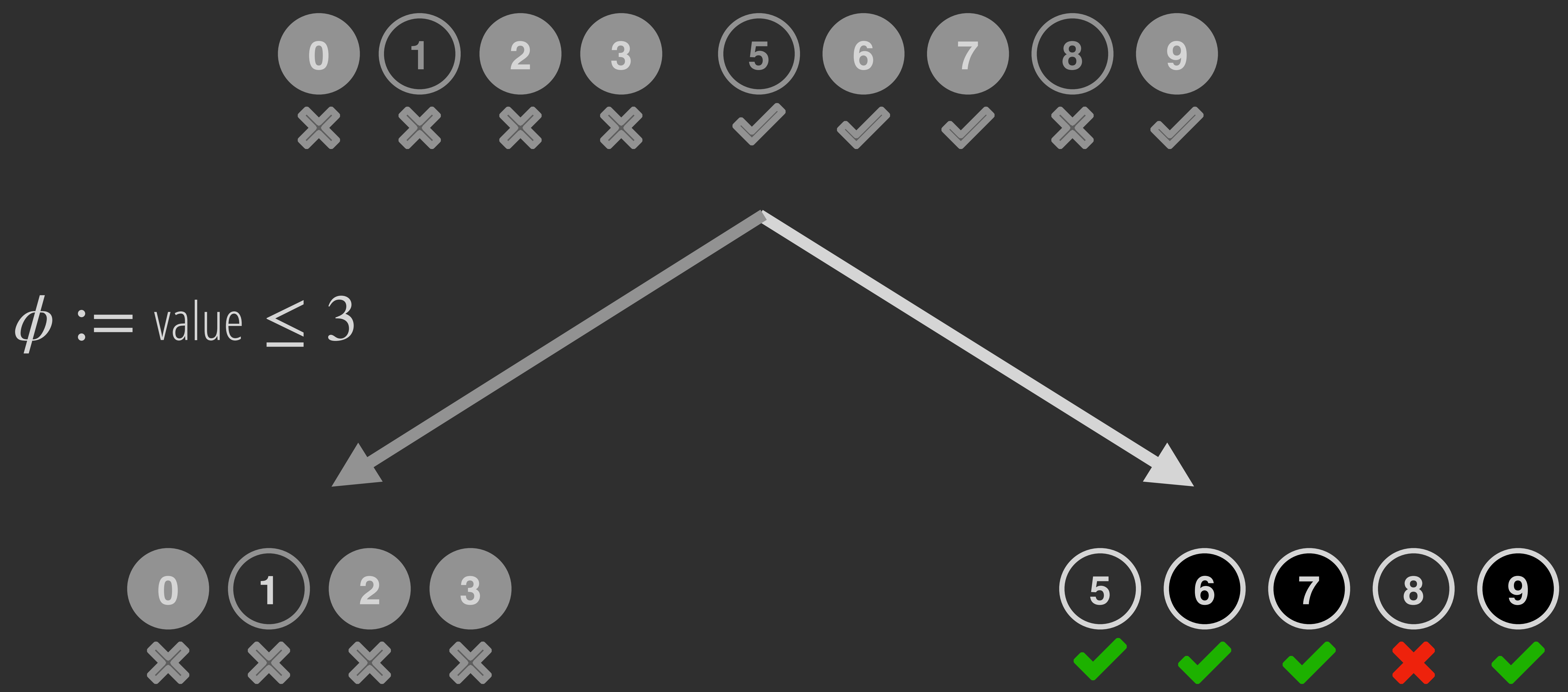
Dataset D


Dataset D
0 1 2 3 5 6 7 8 9
$\phi :=$ value ≤ 3
0 1 2 3
5 6 7 8 9



Dataset D
0 1 2 3 5 6 7 8 9
$\phi :=$ value ≤ 3
0 1 2 3
5 6 7 8 9
Number ✓ = 4
Number ✗ = 1

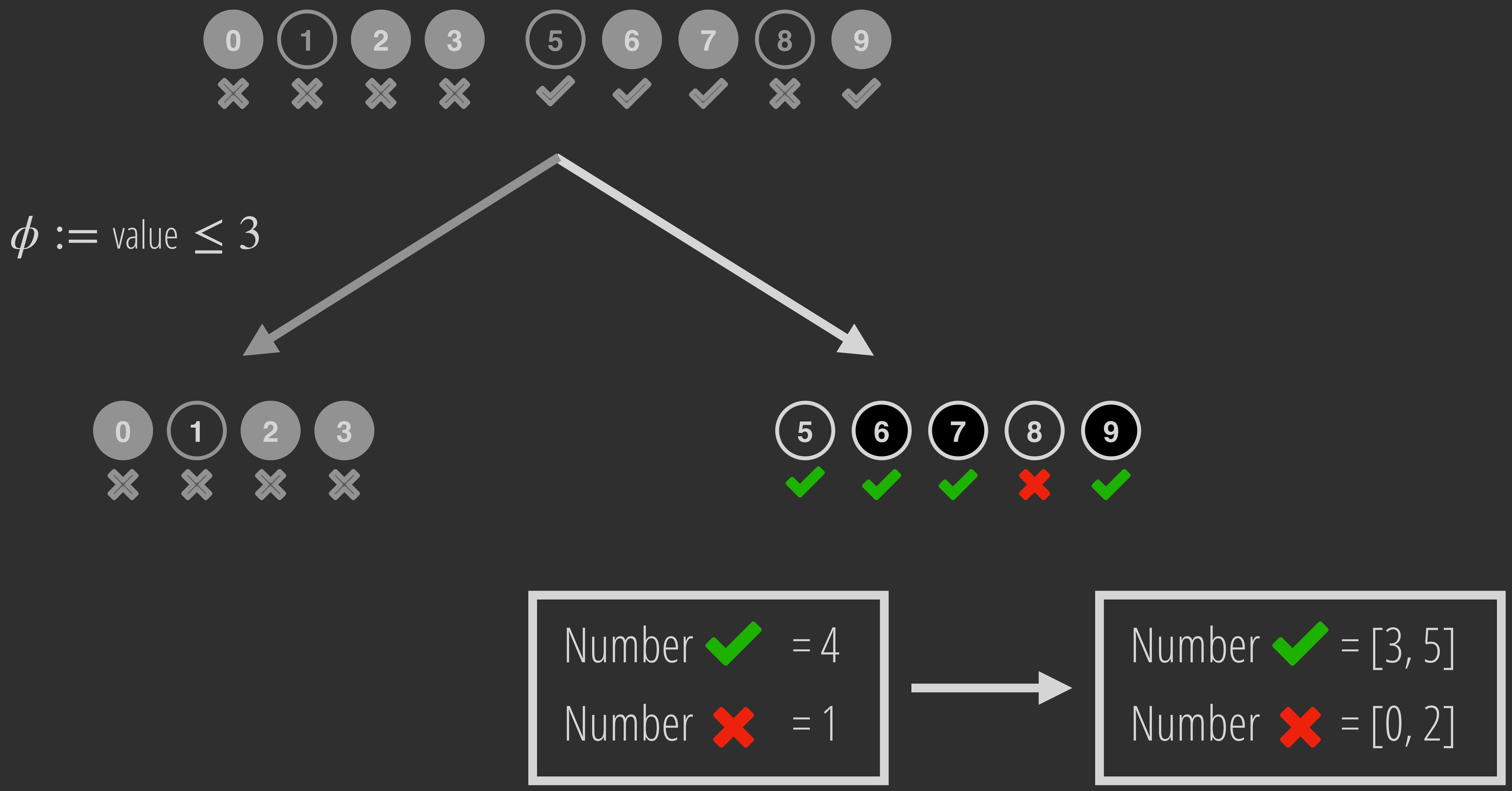
Dataset D



$$\text{Gini Impurity} = \checkmark \cdot (1 - \checkmark) + \times \cdot (1 - \times)$$

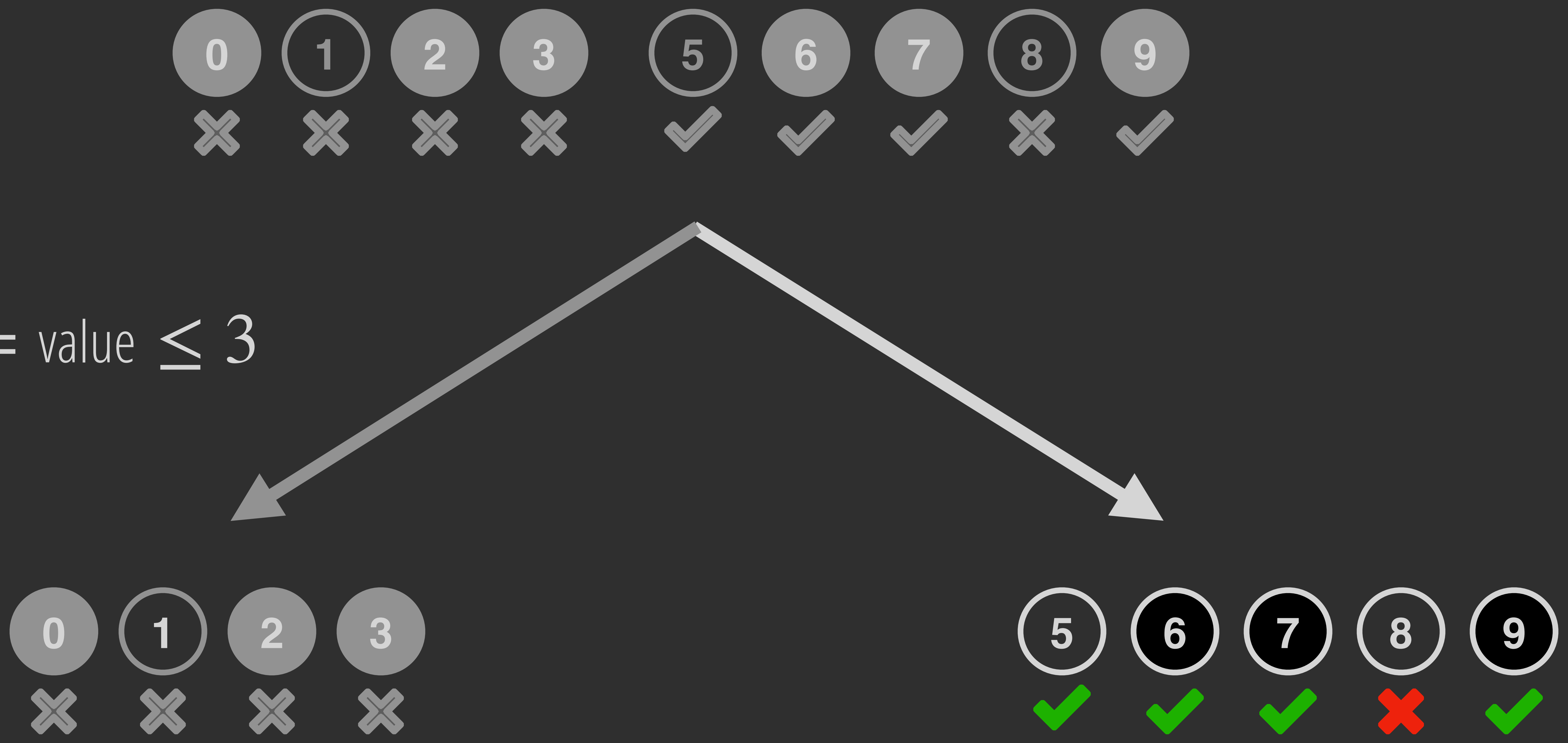
$$= \frac{4}{5}(1 - \frac{4}{5}) + \frac{1}{5}(1 - \frac{1}{5})) \quad = 0.32$$

Abstraction of Dataset D



Abstraction of Dataset D

$\phi :=$ value ≤ 3



$$\text{Gini Impurity} = \checkmark \cdot (1 - \checkmark) + \times \cdot (1 - \times)$$

$$= \frac{[3,5]}{5}\left(1 - \frac{[3,5]}{5}\right) + \frac{[0,2]}{5}\left(1 - \frac{[0,2]}{5}\right) \quad = [0,\ 0.8]$$

5 6 7 8 9

✔ ✔ ✔ ✖ ✔

$$\text{Gini Impurity} = ✔ \cdot (1 - ✔) + ✖ \cdot (1 - ✖)$$

$$= \frac{[3,5]}{5}\left(1 - \frac{[3,5]}{5}\right) + \frac{[0,2]}{5}\left(1 - \frac{[0,2]}{5}\right) \quad = [0,\ 0.8]$$

Aside: We can be more precise!

E.g., we can't simultaneously have 5 ✔ and 2 ✖ .

—–> details omitted from this presentation, but in our experiments we use the precise version

Experimental results

Certification rate

Given n% bias, what percentage of test data points are certifiably robust?

Bias type	Dataset	Bias amount as a percentage of training set					
		0.05	0.1	0.2	0.4	0.7	1.0
FLIP (label-flipping)	Drug Consumption	94.5	94.5	94.5	94.5	92.1	85.1
	COMPAS	100.0	89.0	81.5	71.5	47.8	39.7
	Adult Income	98.4	96.6	85.2	64.4	42.2	14.8
	COMPAS targeted	100.0	89.0	89.0	81.9	76.2	53.0
	AI targeted	98.8	98.6	96.6	81.6	69.0	52.8

Certification rate

Given n% bias, what percentage of test data points are certifiably robust?

Bias type	Dataset	Bias amount as a percentage of training set					
		0.05	0.1	0.2	0.4	0.7	1.0
FLIP (label-flipping)	Drug Consumption	94.5	94.5	94.5	94.5	92.1	85.1
	COMPAS	100.0	89.0	81.5	71.5	47.8	39.7
	Adult Income	98.4	96.6	85.2	64.4	42.2	14.8
	COMPAS targeted	100.0	89.0	89.0	81.9	76.2	53.0
	AI targeted	98.8	98.6	96.6	81.6	69.0	52.8

Bias-set size color scheme	$< 10^{10}$	$< 10^{50}$	$< 10^{100}$	$< 10^{500}$	$> 10^{500}$	infinite

Future work

- Extensions to other ML algorithms
- Counter-examples to robustness