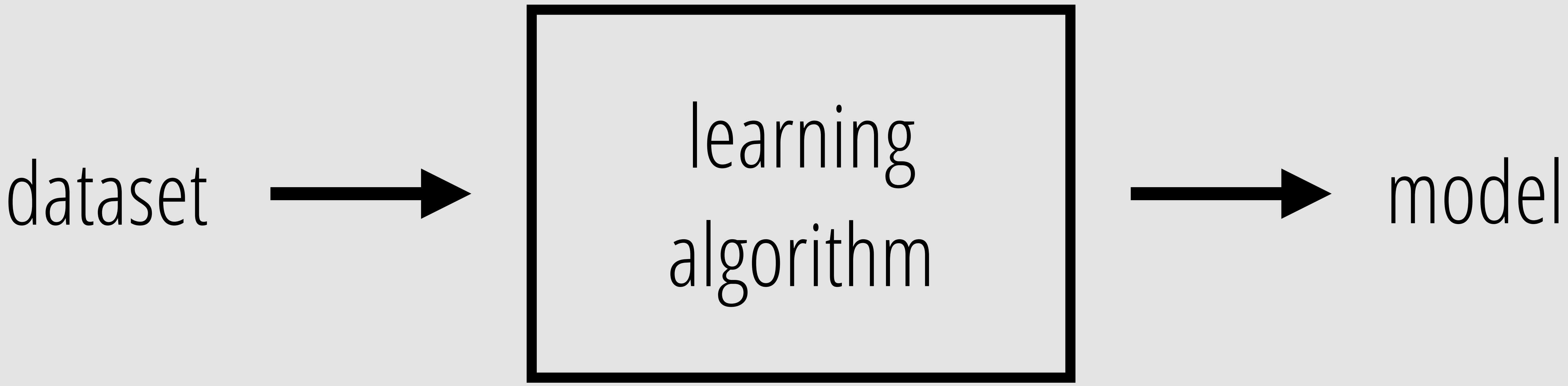
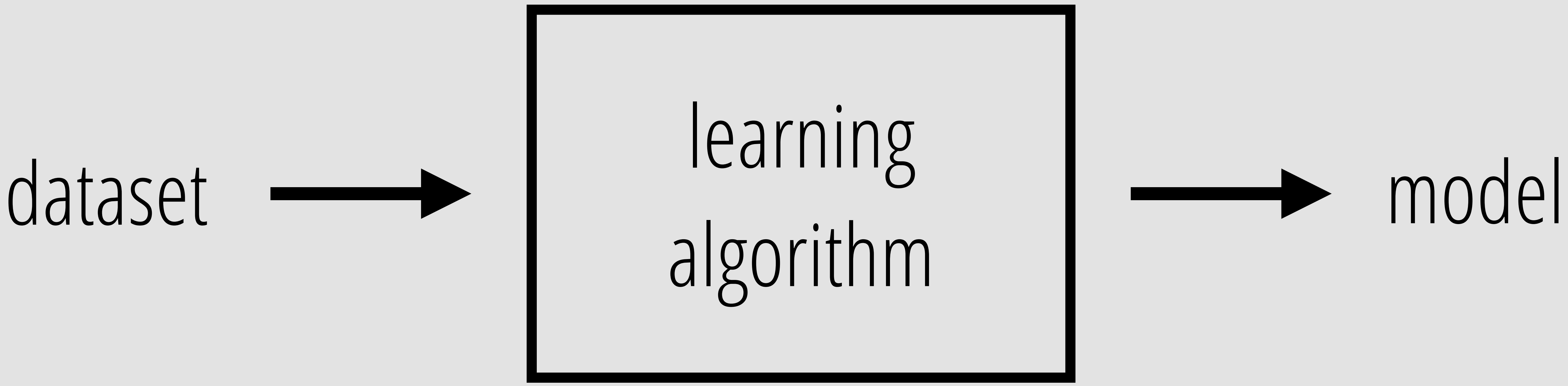
# Certifying Robustness to Programmable Data Bias in Decision Trees

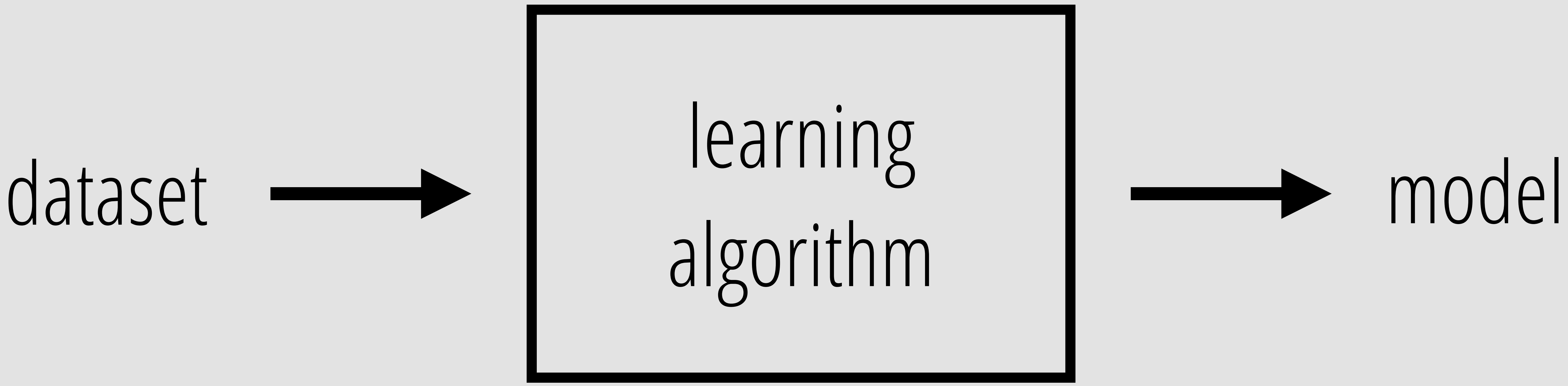
Anna P. Meyer, Aws Albarghouthi, and Loris D'Antoni

NEURAL INFORMATION PROCESSING SYSTEMS

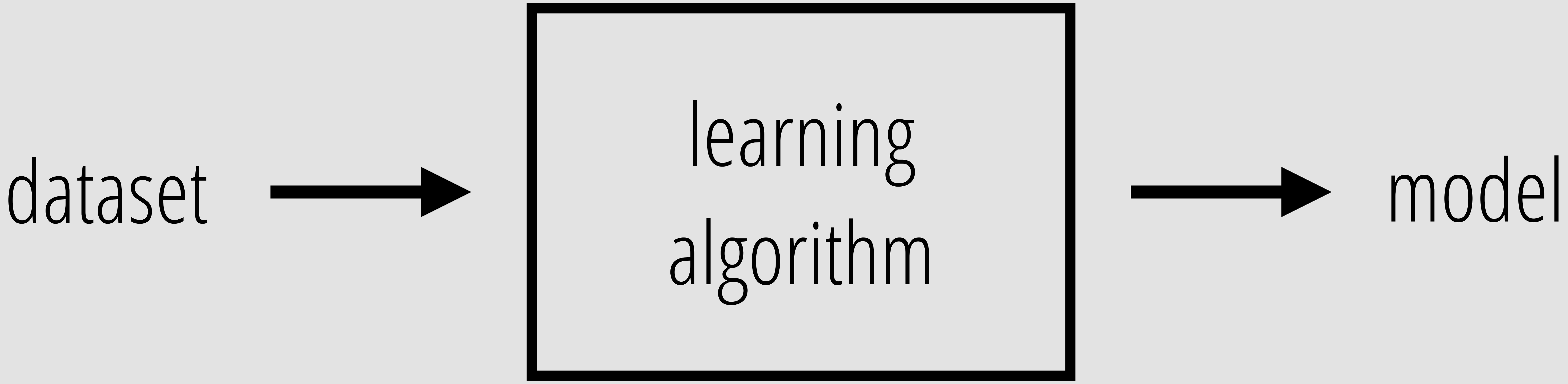

dataset
learning algorithm
model




Is the model fair?
accurate?
trustworthy?




Is the dataset biased?
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 data bias

Incorrect labels

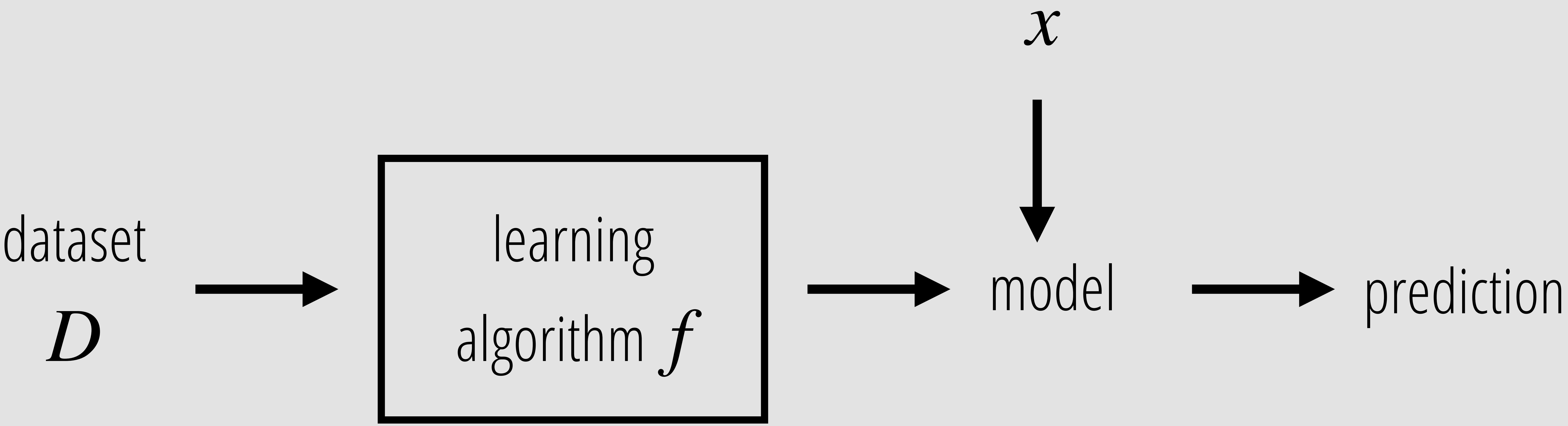
e.g., historical biases like women marked as “not hired” for a job even though they were qualified

Missing data

e.g., neglected to collect data from a minority neighborhood

Fake data

e.g., fake answers submitted through crowdsourcing




bias robustness of $x$

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

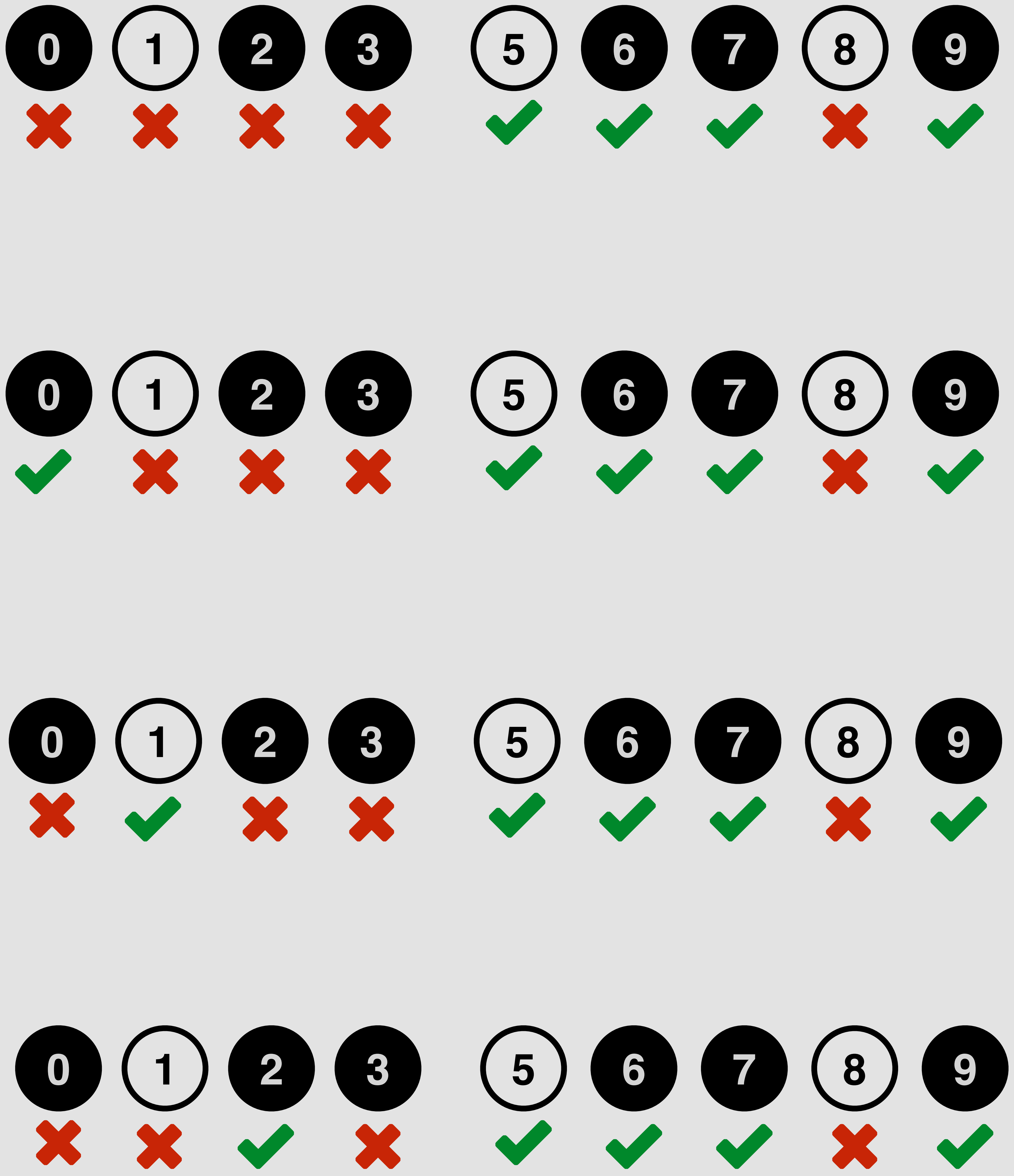
show that $f_{D'}(x) = f_D(x)$

bias robustness of $x$

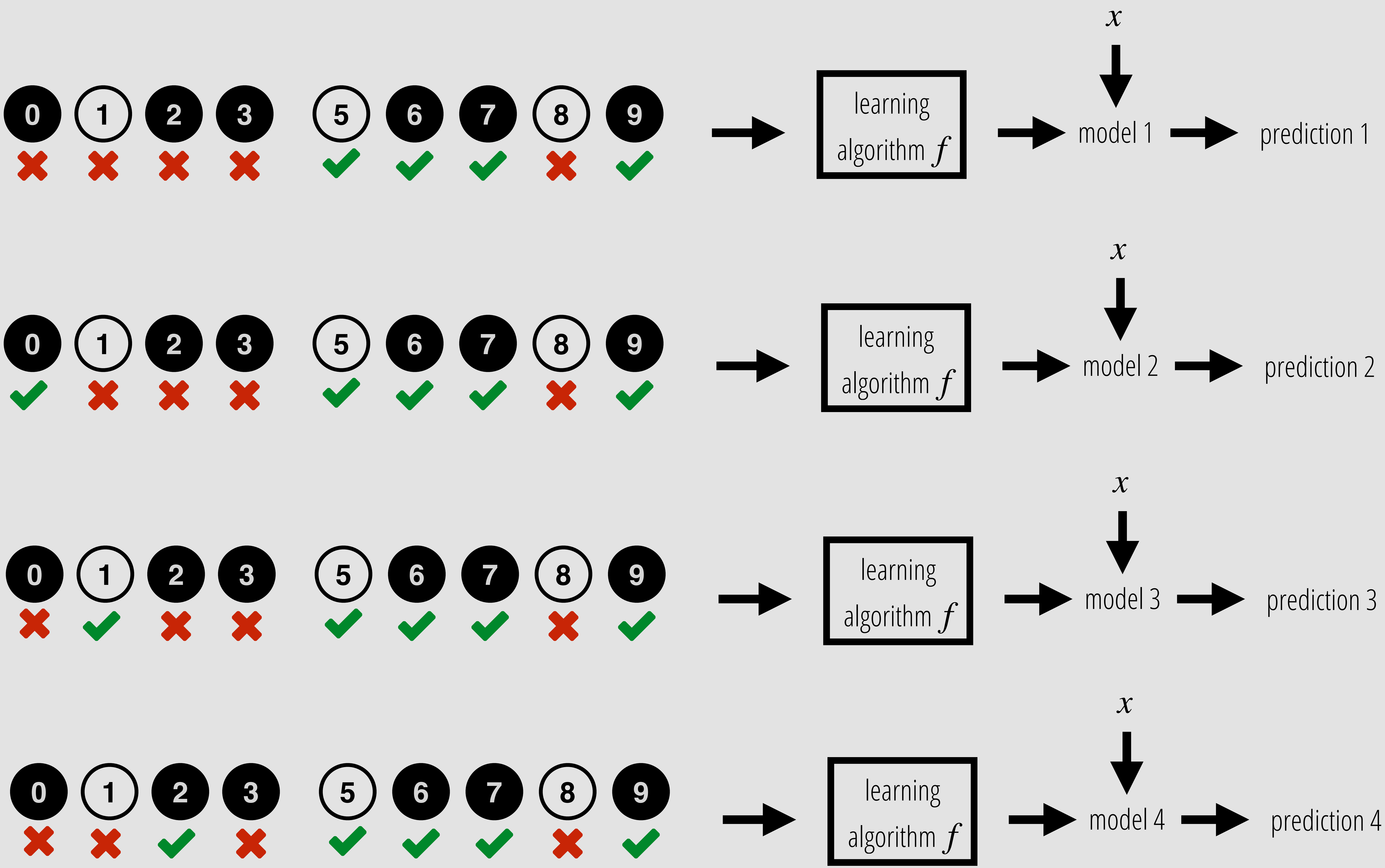
for all $D'$ that disagree with $D$ on $\leq n$ labels

show that $f_{D'}(x) = f_D(x)$

Dataset $D$




etc.


0 1 2 3 5 6 7 8 9
learning algorithm $f$
$x$
model 1
prediction 1
0 1 2 3 5 6 7 8 9
learning algorithm $f$
$x$
model 2
prediction 2
0 1 2 3 5 6 7 8 9
learning algorithm $f$
$x$
model 3
prediction 3
0 1 2 3 5 6 7 8 9
learning algorithm $f$
$x$
model 4
prediction 4

etc.

$|D| = 1000$

$n = 10$

$\sim 10^{23}$ datasets!

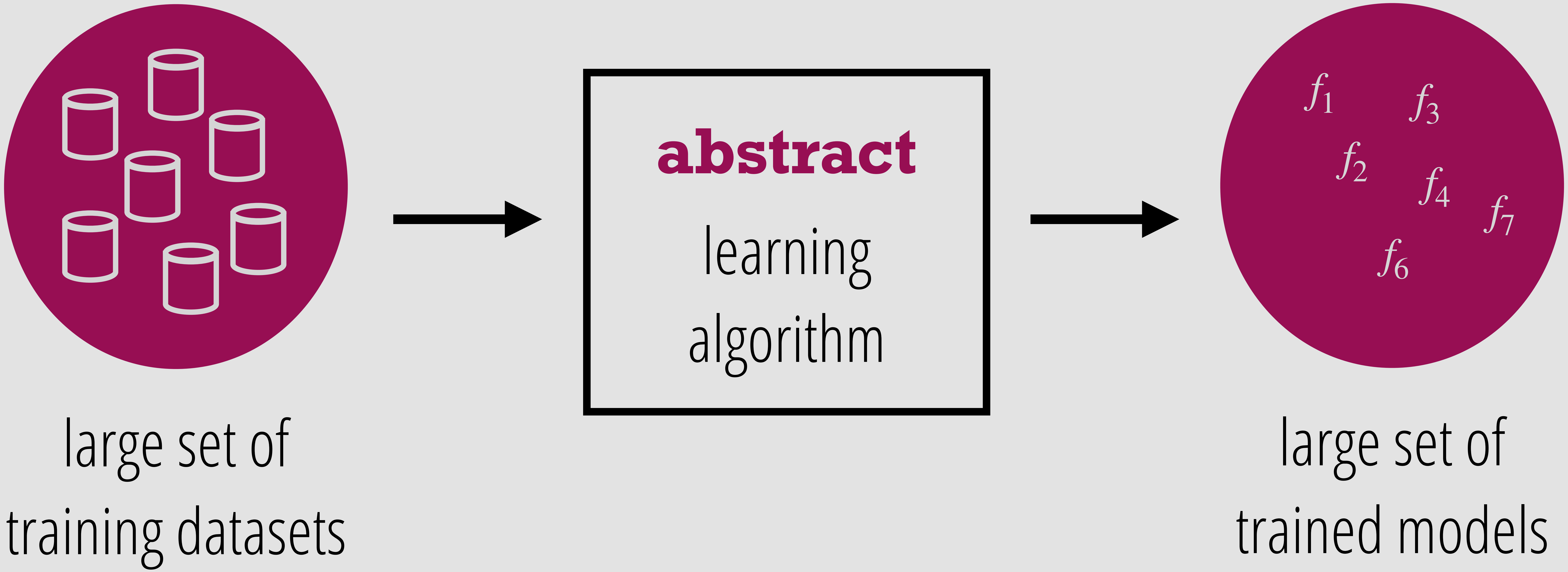
bias robustness of $x$

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

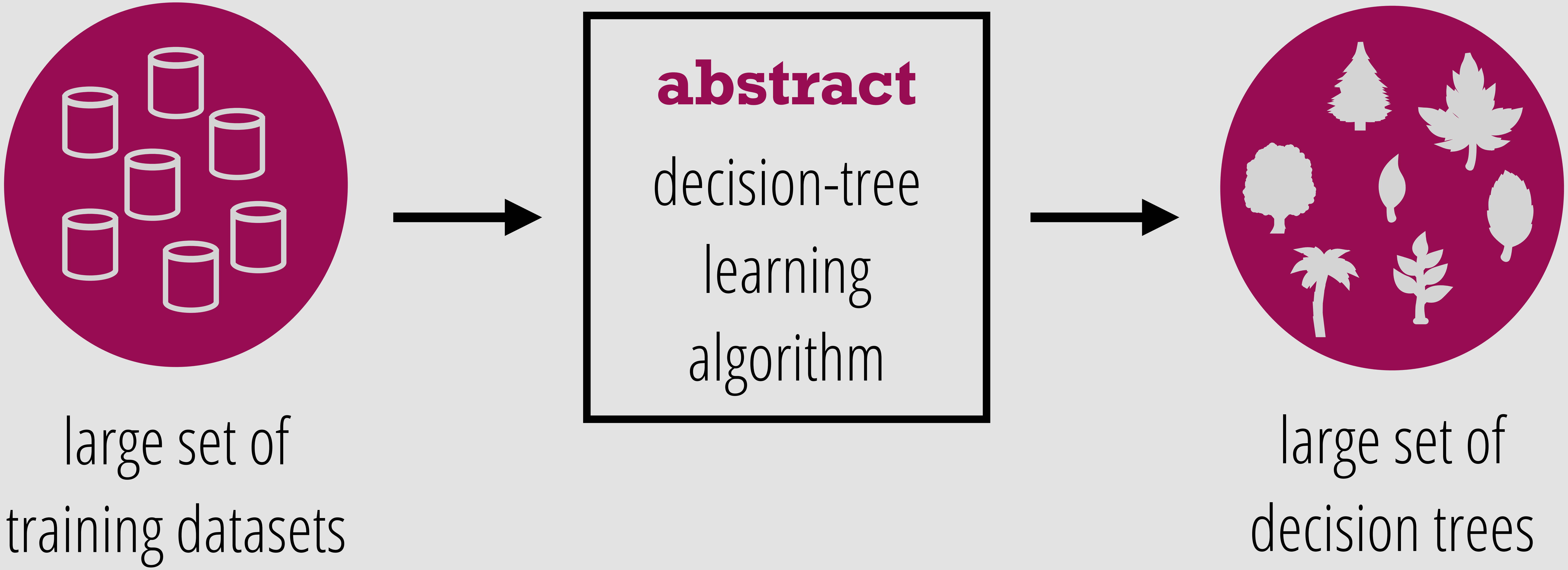
show that $f_{D'}(x) = f_D(x)$

Key challenge

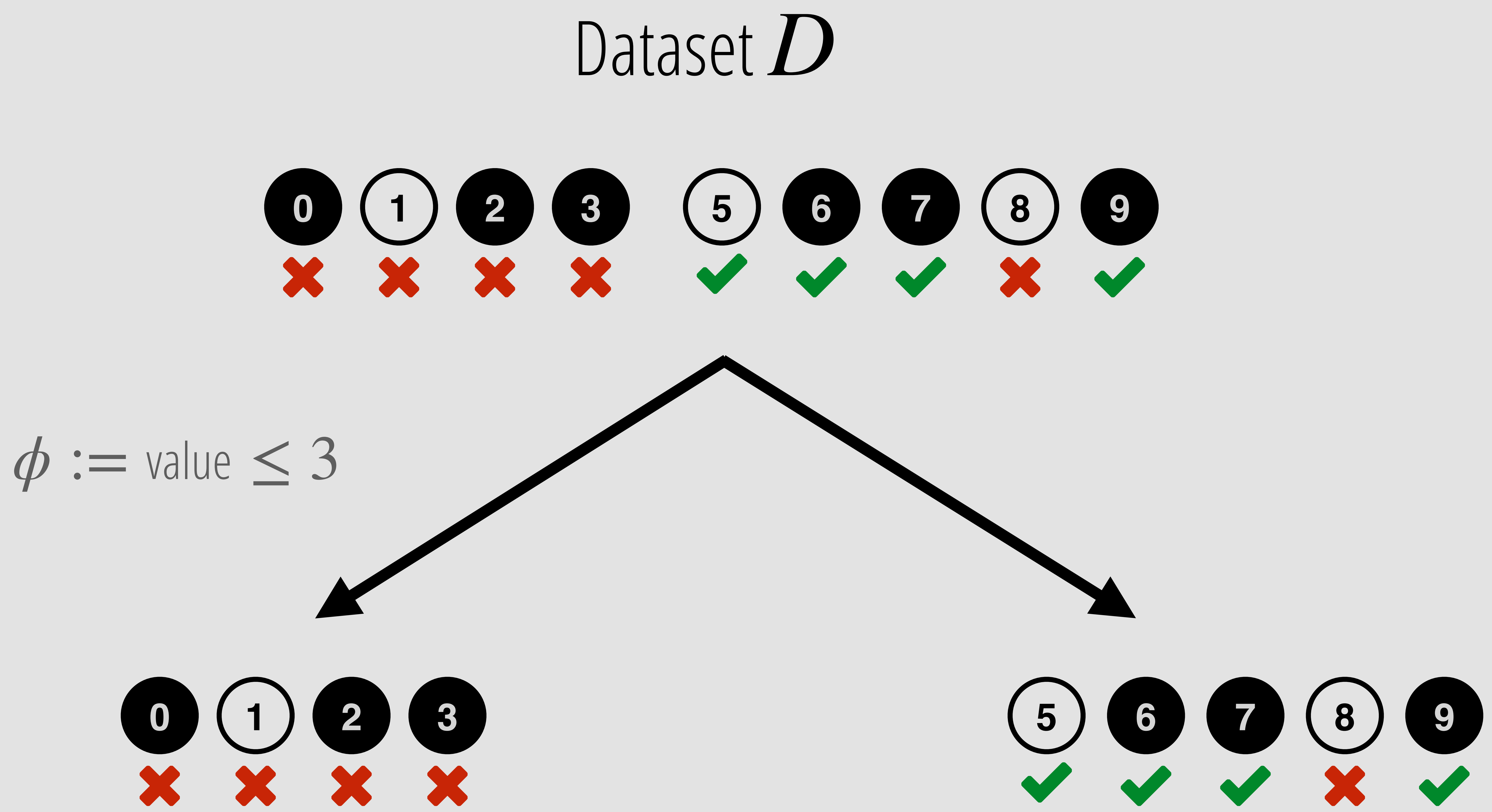
Combinatorial explosion in the number of datasets

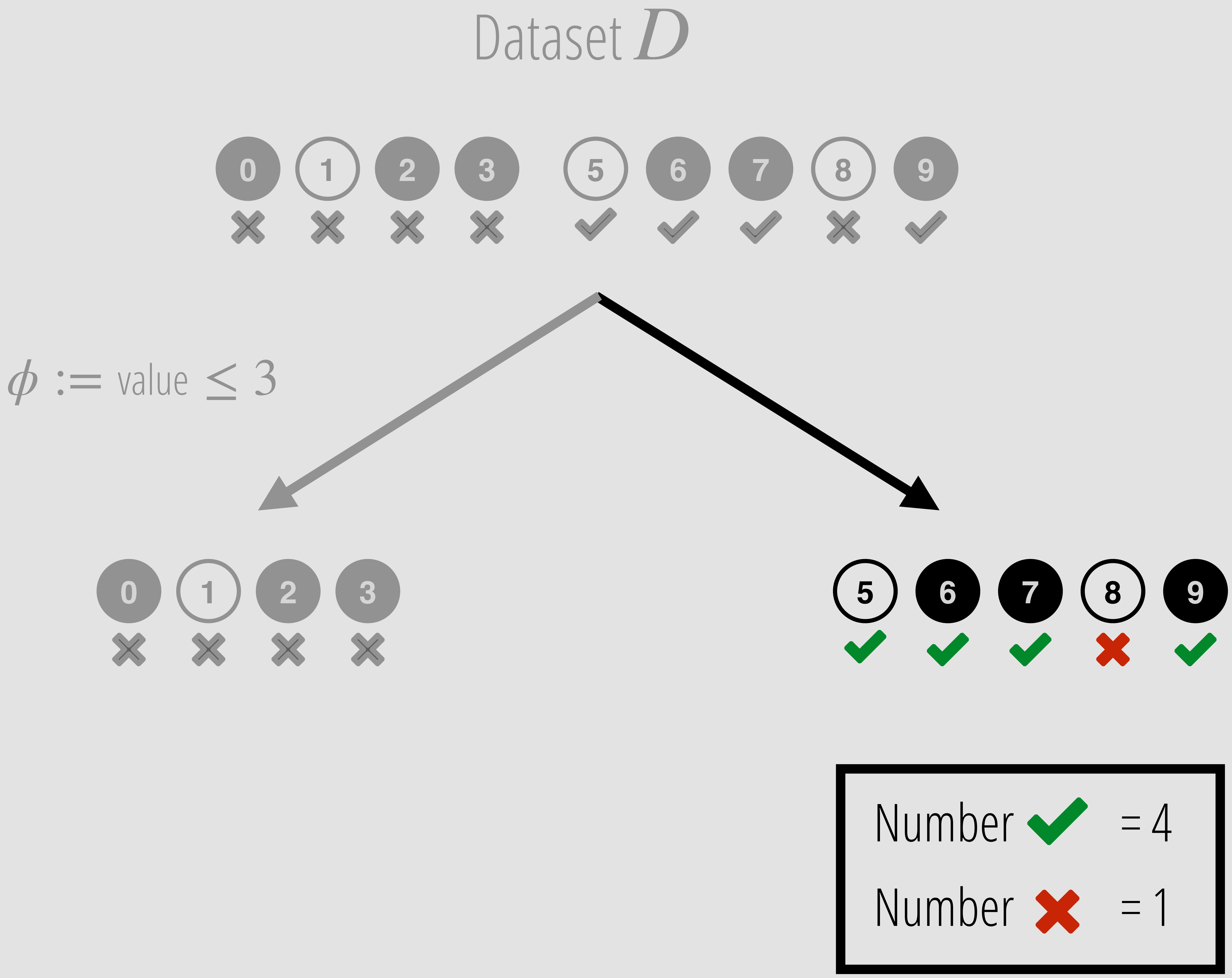


large set of training datasets
abstract
learning algorithm
$f_1$
$f_3$
$f_2$
$f_4$
$f_7$
$f_6$
large set of trained models

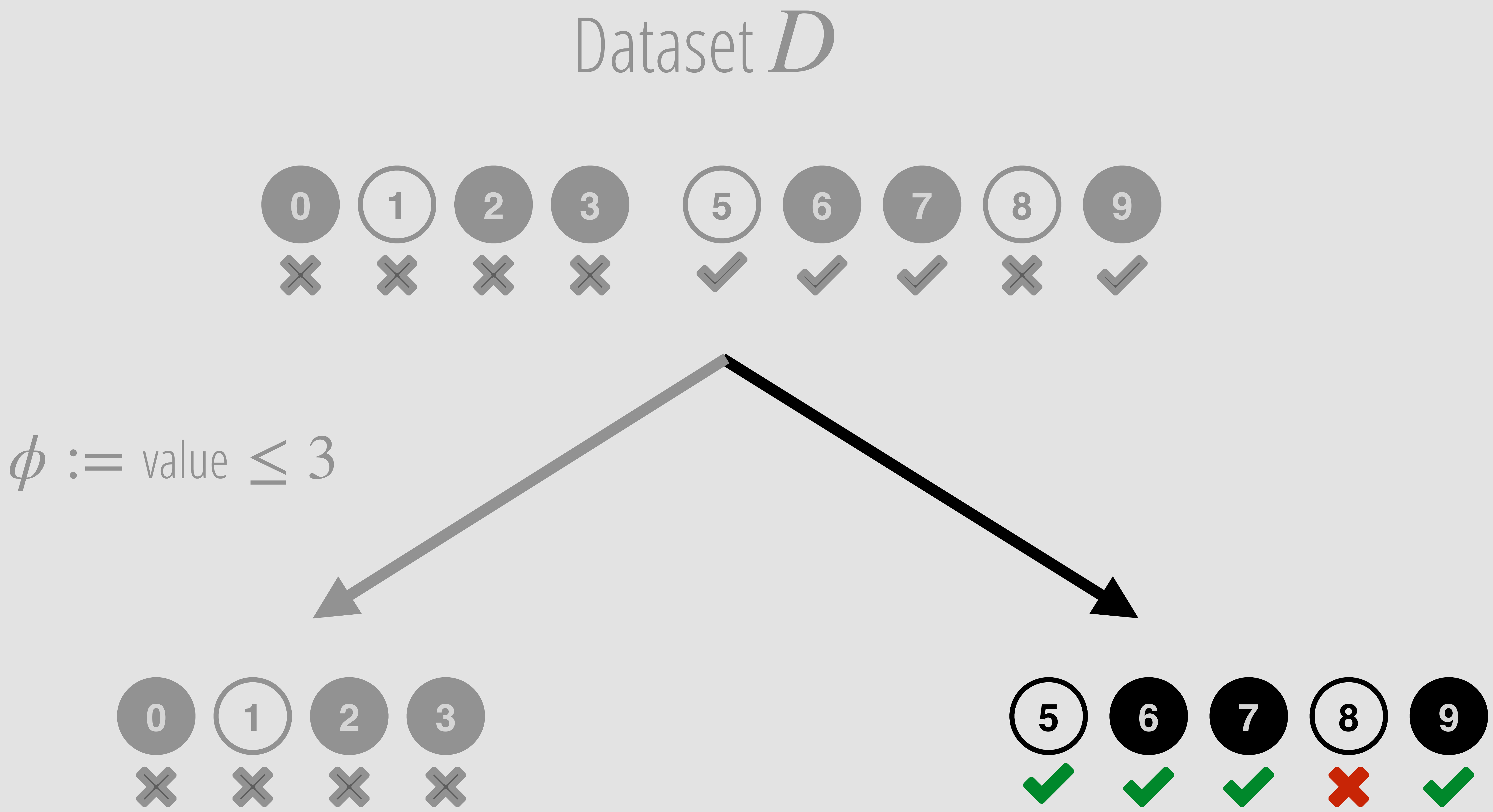

large set of
training datasets
abstract
decision-tree
learning
algorithm
large set of
decision trees

Dataset $D$


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

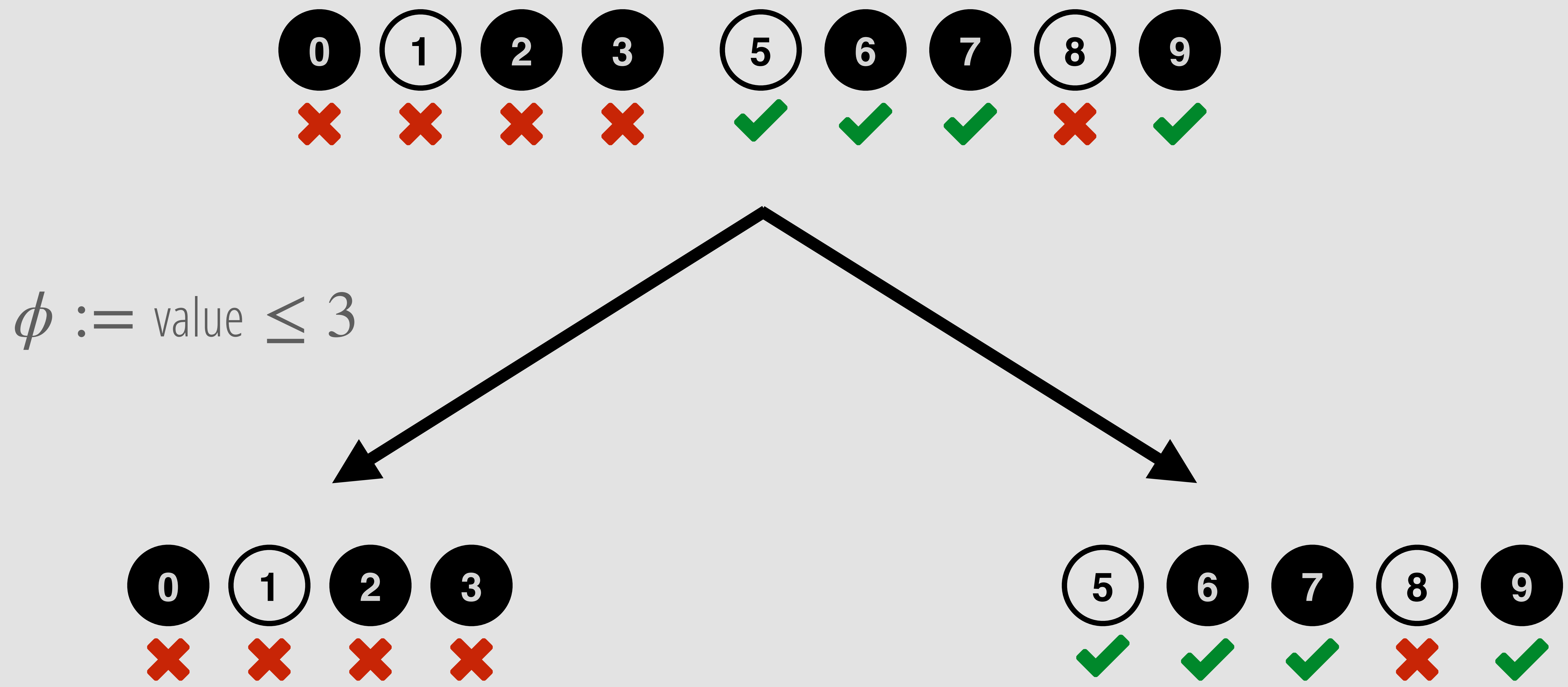


Dataset $D$
0 1 2 3 5 6 7 8 9
$\phi :=$ value $\leq 3$
0 1 2 3
5 6 7 8 9
Number ✓ = 4
Number ✗ = 1

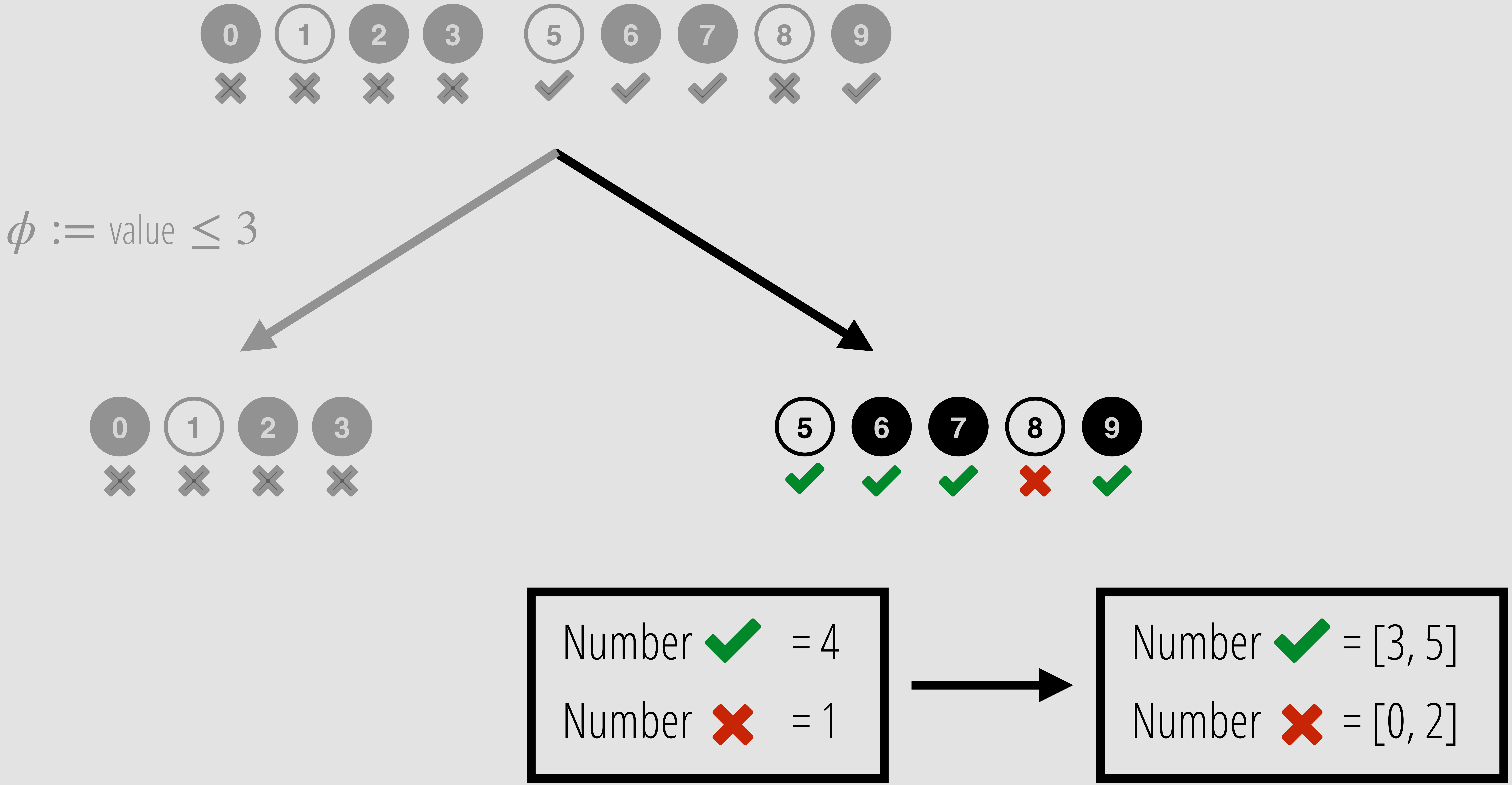



$$
\begin{aligned}
\text{Gini Impurity} &= \checkmark \cdot (1-\checkmark) + \times \cdot (1-\times) \\
&= (4/5)(1-(4/5)) + (1/5)(1-(1/5)) = 0.32
\end{aligned}
$$

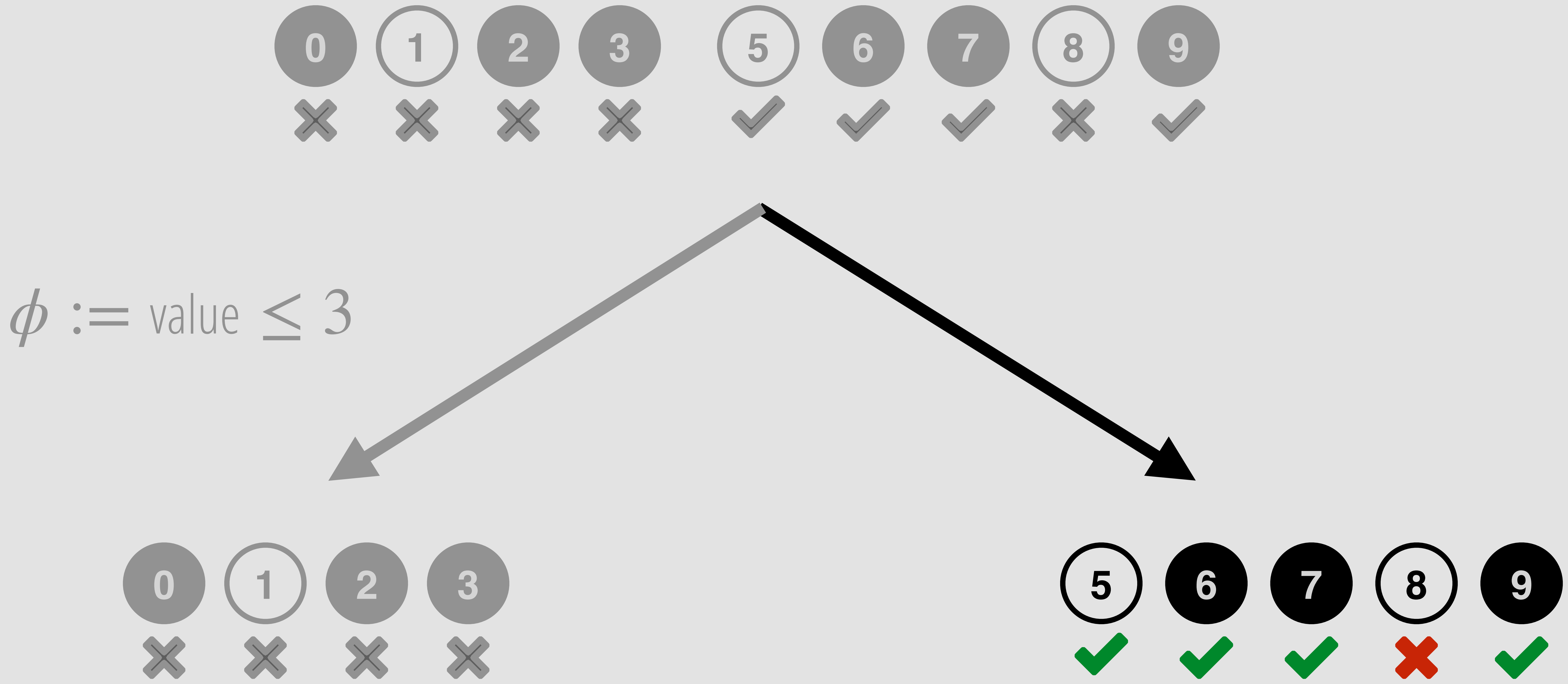
# Abstraction of Dataset $D$



# Abstraction of Dataset $D$



# Abstraction of Dataset $D$




$$\text{Gini Impurity} = ✓ \cdot (1 - ✓) + ✗ \cdot (1 - ✗)$$
$$= ([3,5]/5)(1 - ([3,5]/5)) + ([0,2]/5)(1 - ([0,2]/5))$$
$$= [0,\ 0.8]$$

# Abstract decision-tree-learner pipeline

1. Build an abstract decision tree

2. Find the prediction of $\boldsymbol{x}$ under each of the trees constructed with the best predicates

3. See whether all predictions agree

   If so, $\boldsymbol{x}$ is certifiably robust!

   If not, inconclusive.

# 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 |
| MISS (missing data) | Drug Consumption | 94.5 | 94.5 | 94.5 | 94.5 | 85.1 | 85.1 |
| | COMPAS | 89.0 | 81.9 | 52.9 | 45.3 | 9.3 | 9.2 |
| | Adult Income (AI) | 96.0 | 86.9 | 72.8 | 60.9 | | |
| | COMPAS targeted | 89.0 | 89.0 | 81.9 | 52.9 | 47.8 | 42.3 |
| | AI targeted | 98.8 | 97.2 | 86.6 | 73.0 | 62.0 | 31.6 |

# Certification rate

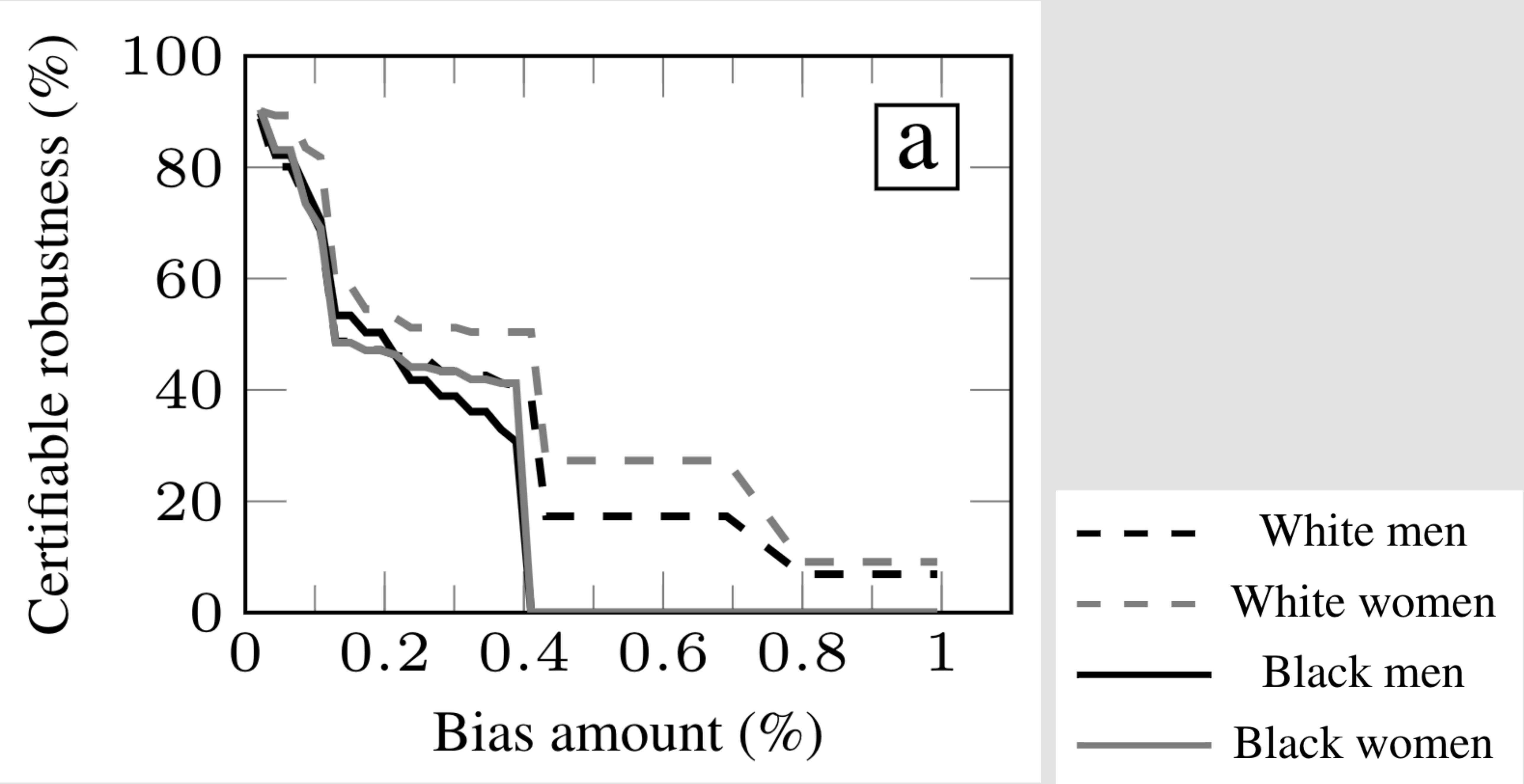
Given n% bias, what percentage of test data points are certifiably robust?

| Bias type | Dataset | Bias amount as a percentage of training set | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.05 | 0.1 | 0.2 | 0.4 | 0.7 | 1.0 |
| MISS (missing data) | Drug Consumption | 94.5 | 94.5 | 94.5 | 94.5 | 85.1 | 85.1 |
| | COMPAS | 89.0 | 81.9 | 52.9 | 45.3 | 9.3 | 9.2 |
| | Adult Income (AI) | 96.0 | 86.9 | 72.8 | 60.9 | | |
| | COMPAS targeted | 89.0 | 89.0 | 81.9 | 52.9 | 47.8 | 42.3 |
| | AI targeted | 98.8 | 97.2 | 86.6 | 73.0 | 62.0 | 31.6 |

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

# Certification discrepancy between demographic groups

COMPAS dataset (but discrepancies exist for Adult Income, too)



# Future work

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