Repairing Decision-Making Programs Under Uncertainty

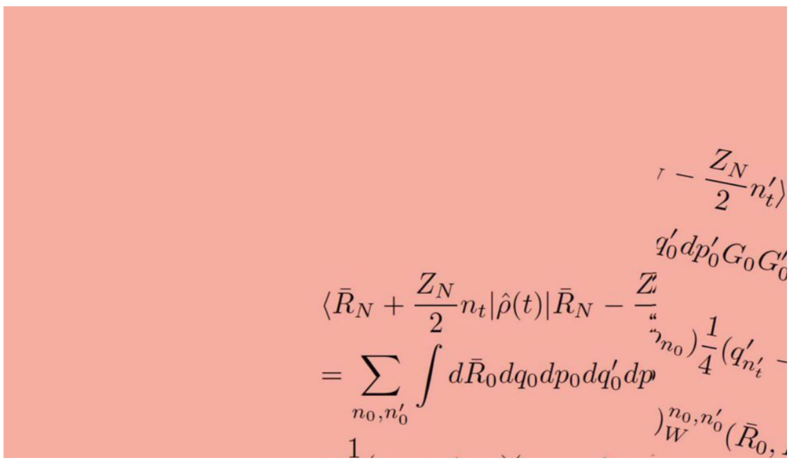


Aws Albarghouthi   Loris D'Antoni   **Samuel Drews**

University of Wisconsin-Madison   madPL

Internet Of Things

Who do you blame when an algorithm gets you fired?



TheUpshot

HIDDEN BIAS

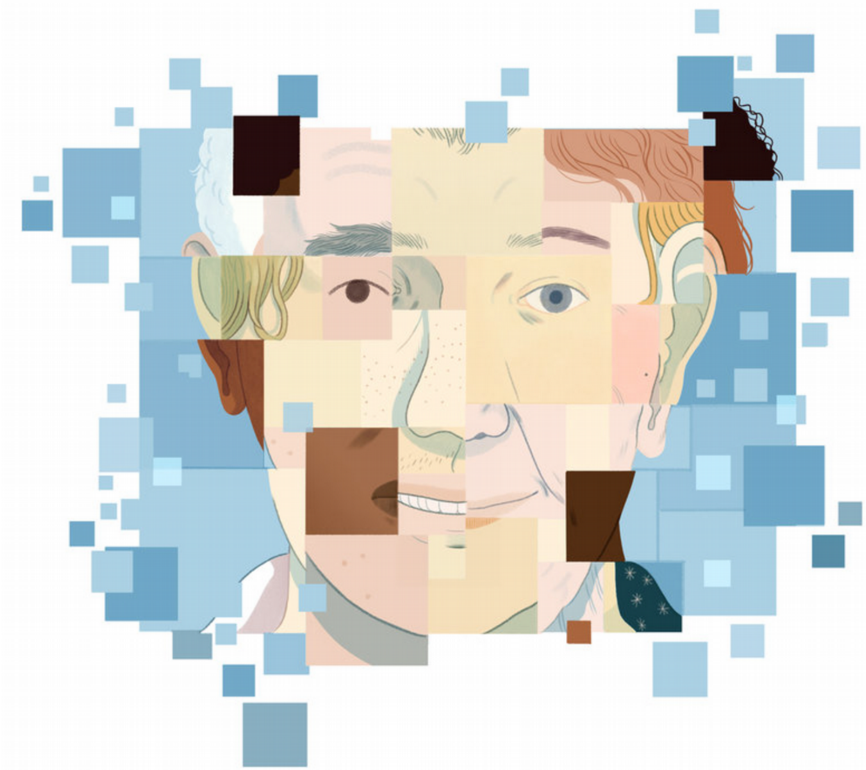
When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

SundayReview | OPINION

Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016



Bianca Bagnarelli

Example Fairness Condition

$$h \leftarrow \mathcal{D}(\boldsymbol{v})$$

decision-making program

Example Fairness Condition

$$\{\boldsymbol{v} = (v_1, \ldots, v_s, \ldots)\}$$

sensitive feature (e.g. minority)

$$h \leftarrow \mathcal{D}(\boldsymbol{v})$$

$$\left\{ \frac{\Pr[h \mid v_s]}{\Pr[h \mid \neg v_s]} > 1 - \epsilon \right\}$$

Example Fairness Condition

probabilistic
precondition

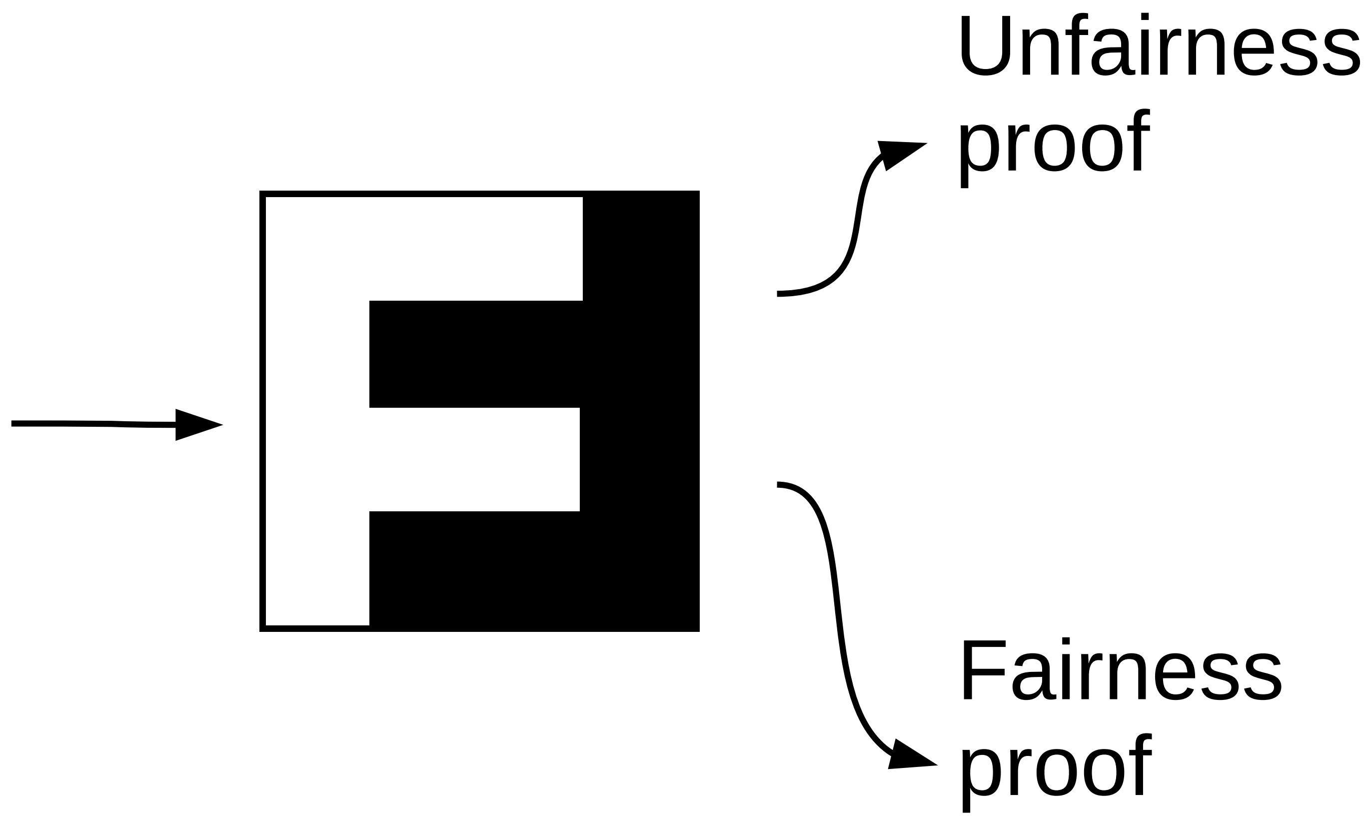
$$\{v \sim \mathcal{M}\}$$

$$h \leftarrow \mathcal{D}(v)$$

$$\left\{ \frac{\Pr[h \mid v_s]}{\Pr[h \mid \neg v_s]} > 1 - \epsilon \right\}$$

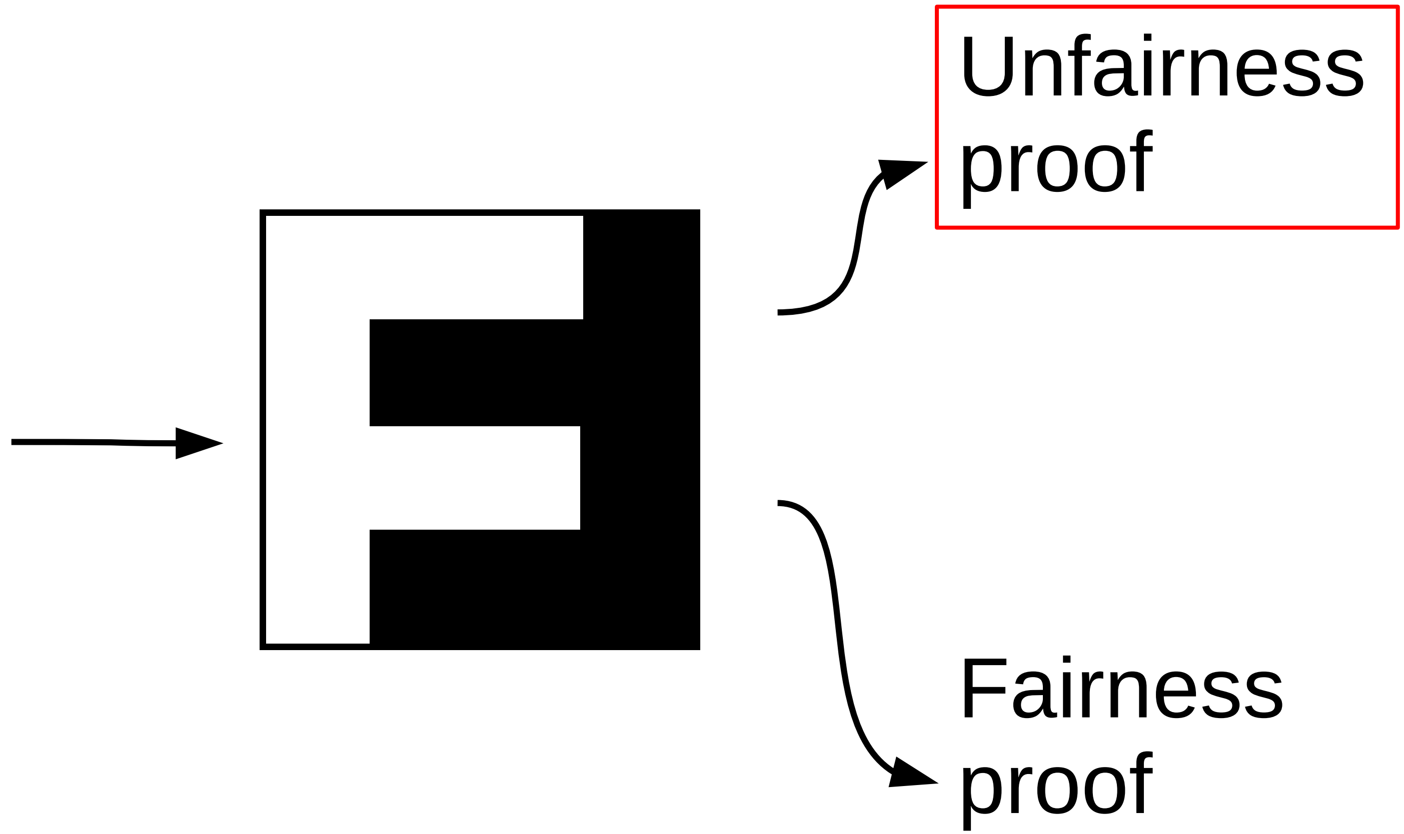
FairSquare [OOPSLA 17]

$$\{v \sim \mathcal{M}\}$$
$$h \leftarrow \mathcal{D}(v)$$
$$\{post\}$$



FairSquare [OOPSLA 17]

$$\{v \sim \mathcal{M}\}$$
$$h \leftarrow \mathcal{D}(v)$$
$$\{post\}$$



Probabilistic program repair: definition

Input

$$\{v \sim \mathcal{M}\}$$

$$h \leftarrow \mathcal{D}(v)$$

$$\{post\}$$

The postcondition
does not hold

Probabilistic program repair: definition

Input	Output

Input

$$\{v \sim \mathcal{M}\}$$
$$h \leftarrow \mathcal{D}(v)$$
$$\{post\}$$

The postcondition **does not hold**

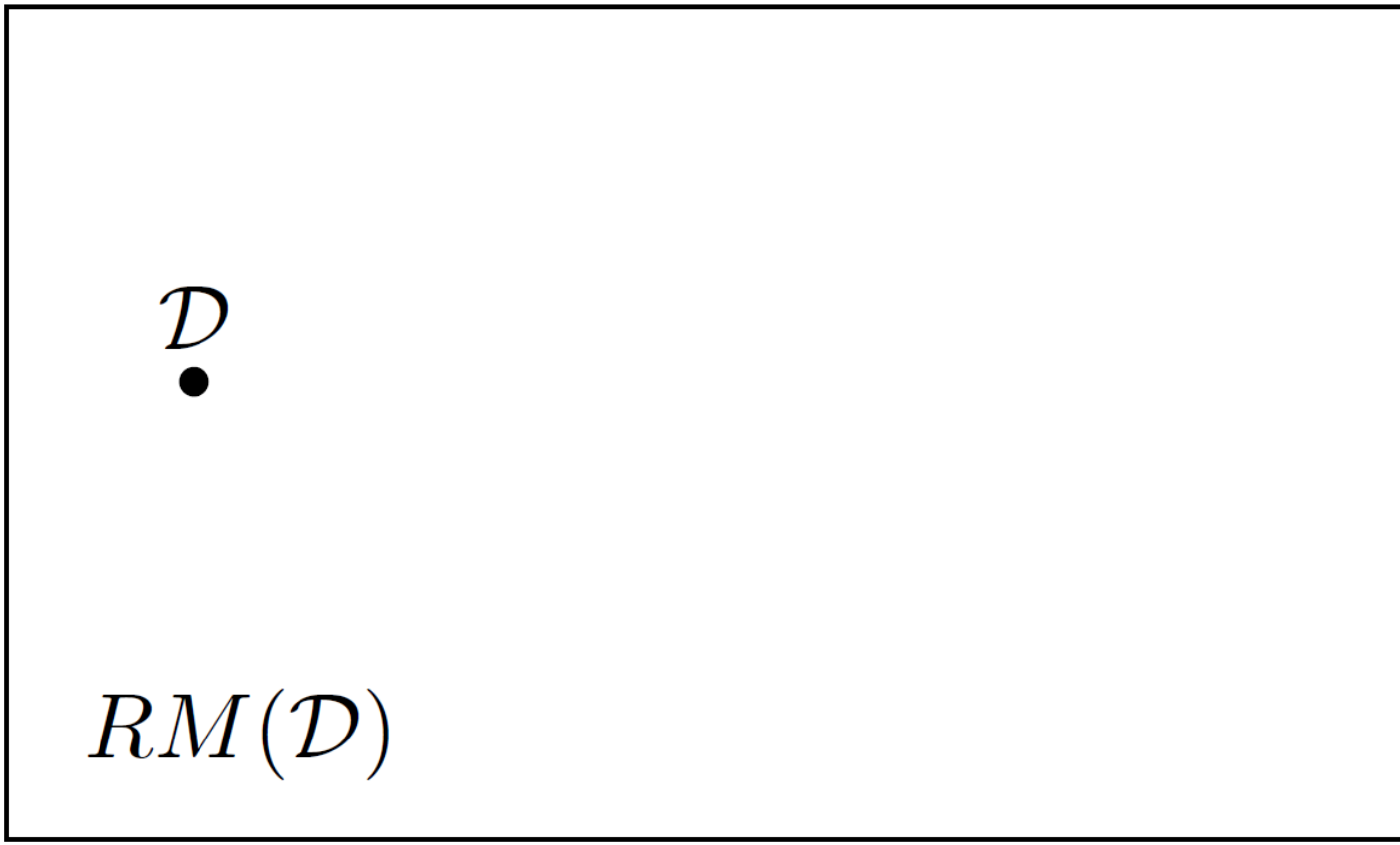
Output

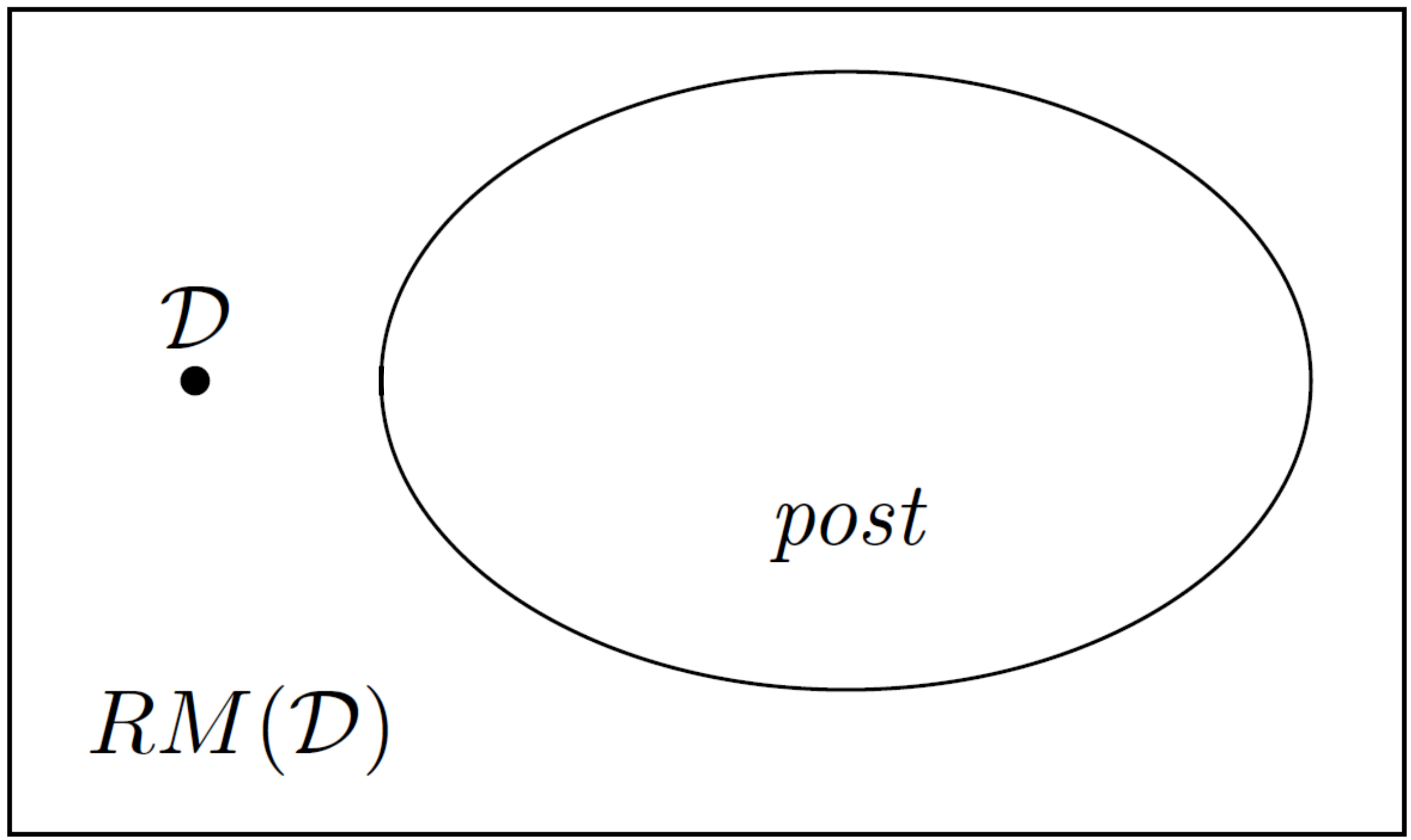
Program $\mathcal{D}'$ in repair model $RM(\mathcal{D})$

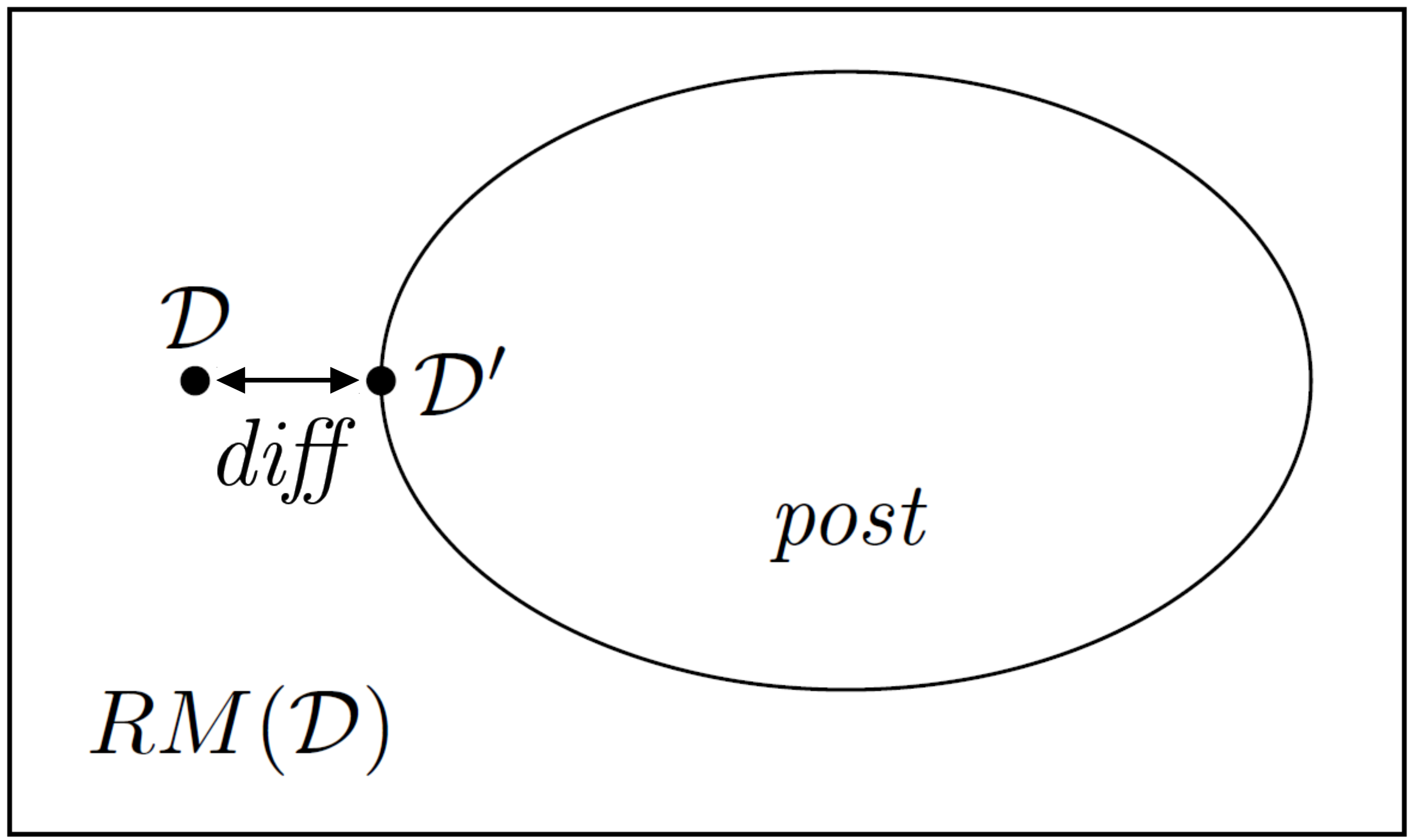
$$\{v \sim \mathcal{M}\}$$
$$h \leftarrow \mathcal{D}'(v)$$
$$\{post\}$$

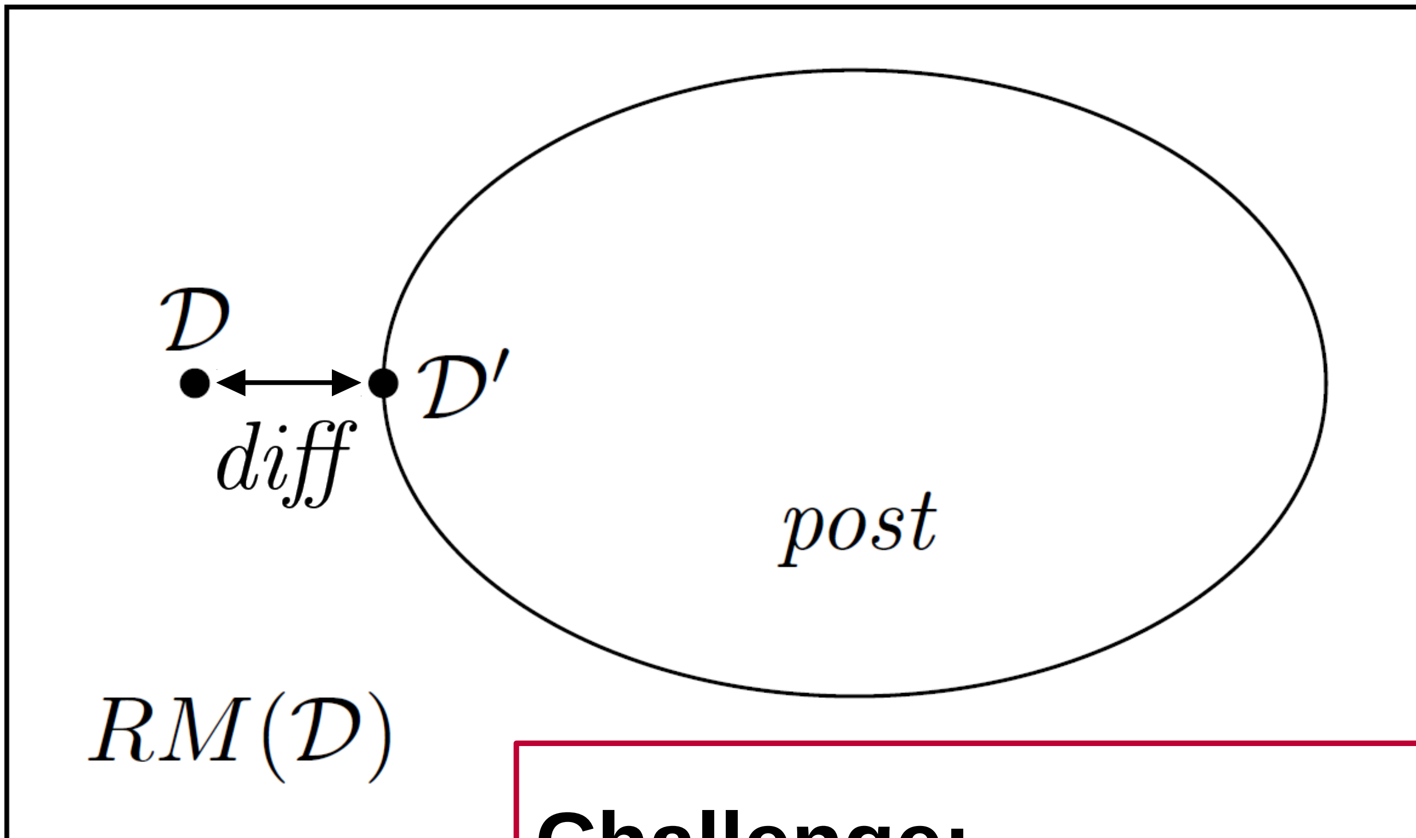
The postcondition **holds**, Difference between $\mathcal{D}$ and $\mathcal{D}'$ is minimal

$$\mathit{diff}(\mathcal{D}, \mathcal{D}') = \Pr[\mathcal{D}(v) \neq \mathcal{D}'(v) \mid v \sim \mathcal{M}]$$


$\mathcal{D}$
$RM(\mathcal{D})$


$\mathcal{D}$
$post$
$RM(\mathcal{D})$


$\mathcal{D}$
$\mathcal{D}'$
$diff$
$post$
$RM(\mathcal{D})$



Challenge:
What should the output of $\mathcal{D}'$ be on each input?

$\mathcal{D}$

```
fun hire(min,urank)
    dec = 1 <= urank <= 10
    return dec
```

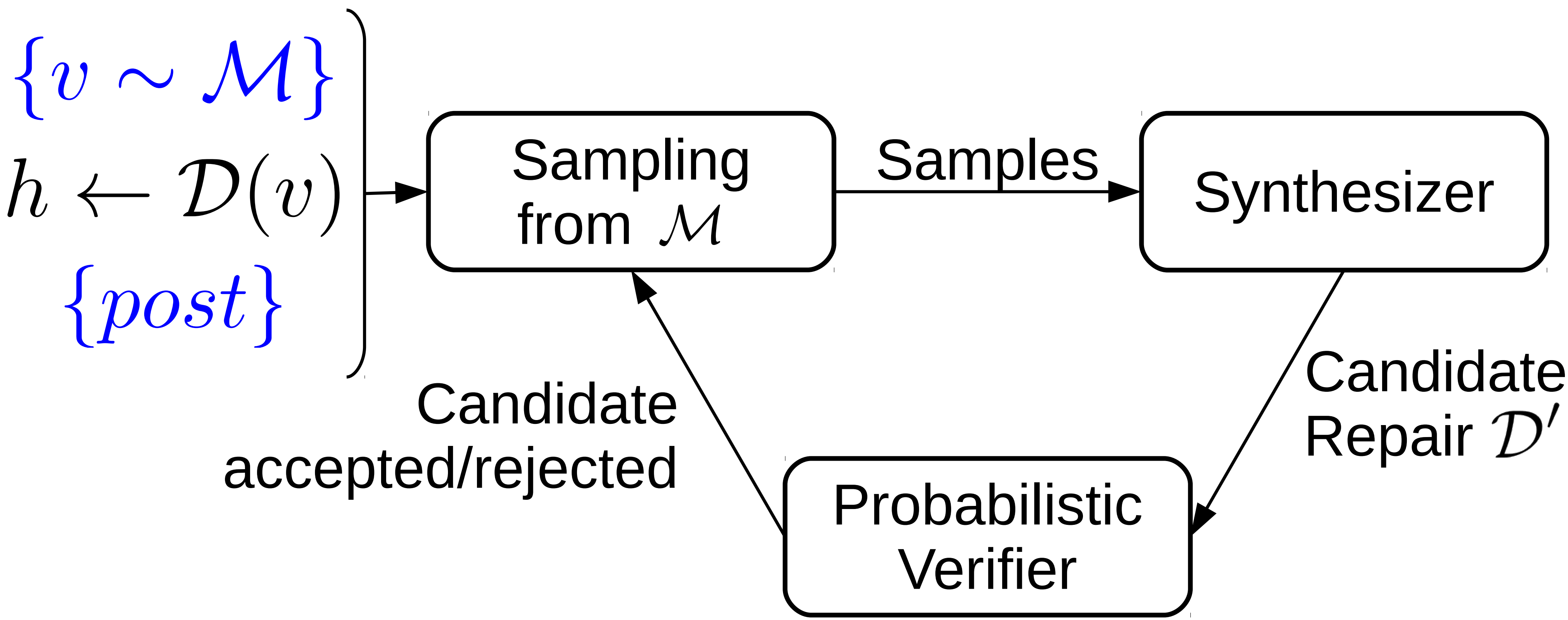
$\mathcal{D}$

```
fun hire(min,urank)
  dec = 1 <= urank <= 10
  return dec
```

$RM(\mathcal{D})$

```
fun hireRep(min,urank)
  dec = ●₁ <= urank <= ●₂
  return dec
```

DIGITS:
DIstribution-Guided InducTive Synthesis



DIGITS:
DIstribution-Guided InducTive Synthesis

Sample **n** inputs
from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n

DIGITS:
DIstribution-Guided InducTive Synthesis

Sample **n** inputs
from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n
T	T	T	…	T
T	T	T	…	F
T	T	F	…	T
T	T	F	…	F
T	F	T	...	T
..	..	..	…	…

DIGITS:
DIstribution-Guided InducTive Synthesis

Sample **n** inputs from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n
T	T	T	…	T
T	T	T	…	F
T	T	F	…	T
T	T	F	…	F
T	F	T	...	T
..	..	..	…	…

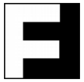
For each labeling, synthesize one program consistent with it and check postcondition

$$\{v \sim \mathcal{M}\}\ \mathcal{D}_1\ \{post\}$$

DIGITS:
DIstribution-Guided InducTive Synthesis

Sample **n** inputs from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n
T	T	T	…	T
T	T	T	…	F
T	T	F	…	T
T	T	F	…	F
T	F	T	...	T
..	..	..	…	…

For each labeling, synthesize one program consistent with it and check postcondition 

$$\{v \sim \mathcal{M}\}\ \mathcal{D}_1\ \{post\}$$
$$\{v \sim \mathcal{M}\}\ \mathcal{D}_2\ \{post\}$$
$$\{v \sim \mathcal{M}\}\ \mathcal{D}_3\ \{post\}$$
$$\{v \sim \mathcal{M}\}\ \mathcal{D}_4\ \{post\}$$
$$\{v \sim \mathcal{M}\}\ \mathcal{D}_5\ \{post\}$$

...

DIGITS: DIstribution-Guided InducTive Synthesis

Sample **n** inputs from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n
T	T	T	…	T
T	T	T	…	F
T	T	F	…	T
T	T	F	…	F
T	F	T	...	T
..	..	..	…	…

For each labeling, synthesize one program consistent with it and check postcondition

$\{v \sim \mathcal{M}\}\ \mathcal{D}_1\ \{post\}$

~~$\{v \sim \mathcal{M}\}\ \mathcal{D}_2\ \{post\}$~~

$\{v \sim \mathcal{M}\}\ \mathcal{D}_3\ \{post\}$

$\{v \sim \mathcal{M}\}\ \mathcal{D}_4\ \{post\}$

~~$\{v \sim \mathcal{M}\}\ \mathcal{D}_5\ \{post\}$~~

...

DIGITS:
DIstribution-Guided InducTive Synthesis

Sample **n** inputs from precondition $\mathcal{M}$

inp_1	inp_2	...	…	inp_n
T	T	T	…	T
T	T	T	…	F
T	T	F	…	T
T	T	F	…	F
T	F	T	...	T
..	..	..	…	…

For each labeling, synthesize one program consistent with it and check postcondition $\models$

$\{v \sim \mathcal{M}\}\ \mathcal{D}_1\ \{post\}$

~~$\{v \sim \mathcal{M}\}\ \mathcal{D}_2\ \{post\}$~~

$\{v \sim \mathcal{M}\}\ \mathcal{D}_3\ \{post\}$

$\{v \sim \mathcal{M}\}\ \mathcal{D}_4\ \{post\}$

~~$\{v \sim \mathcal{M}\}\ \mathcal{D}_5\ \{post\}$~~

...

Output $\mathcal{D}_i$ that has minimal $diff(\mathcal{D}, \mathcal{D}_i)$

```
fun hireRep(min,urank)
  dec = ●1 <= urank <= ●2
  return dec
```

$$\varphi = ((H_1 \leq urank \leq H_2) == dec)$$

$$((H_1 \leq 12 \leq H_2) == true)$$

$$((H_1 \leq 17 \leq H_2) == false)$$

$$((H_1 \leq 12 \leq H_2) == true)$$

$$((H_1 \leq 17 \leq H_2) == false)$$

$$H_1 = 10, H_2 = 15$$

```
fun hireRep1(min,urank)
  dec = 10 <= urank <= 15
  return dec
```

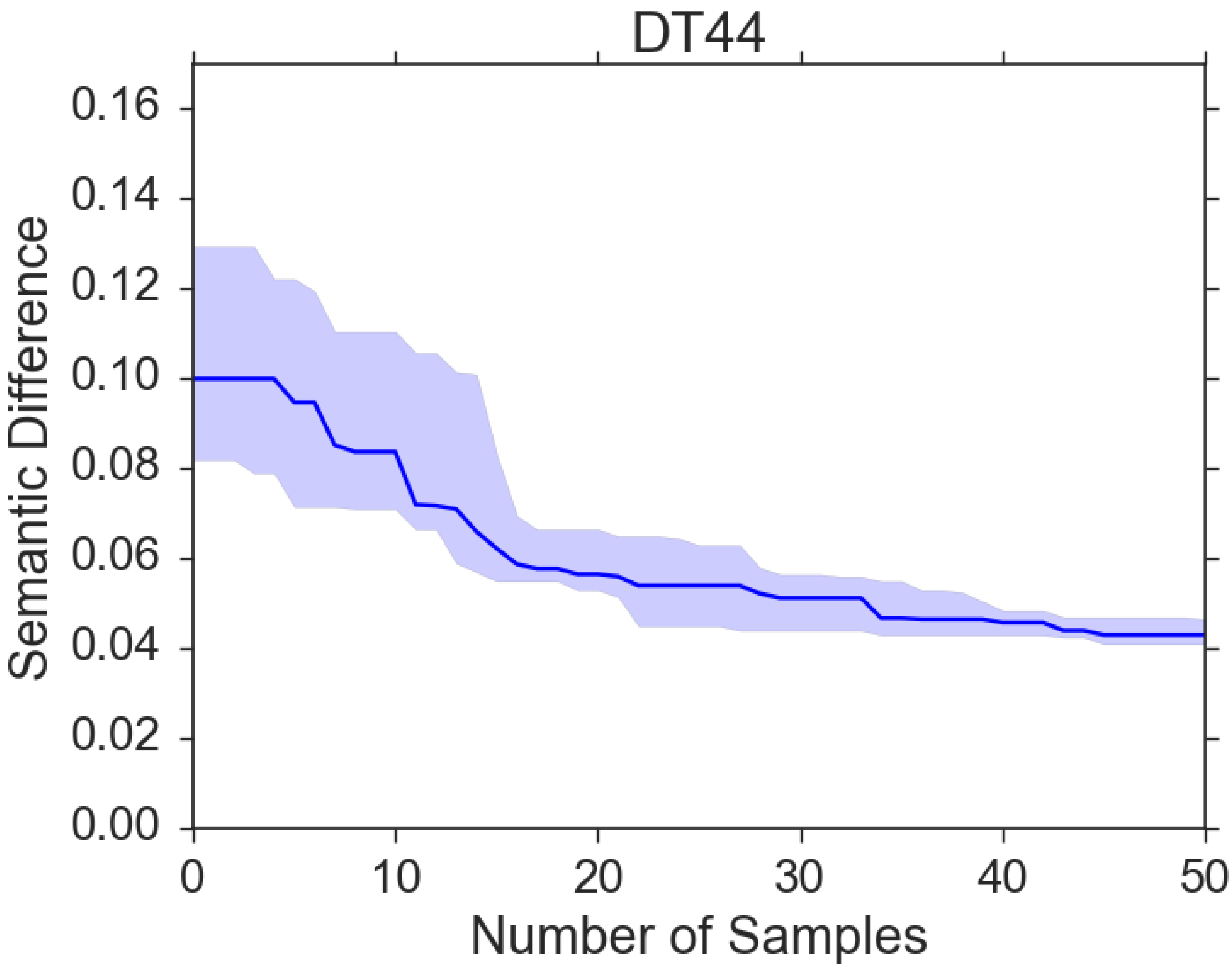
Does DIGITS work?

18 repair problems:
decision trees, support vector machines

10 minute best-effort period

All repaired!

Does DIGITS work?



Does DIGITS converge?

Yes*

* terms and conditions apply:

repair model $RM(\mathcal{D})$ has finite VC-dimension d

postcondition has some extra continuity property (see paper)

VC dimension of a set of programs

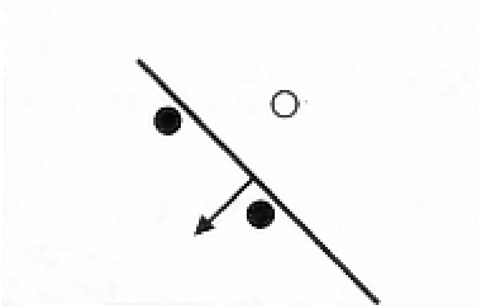


image: V. Kecman, 2001

VC dimension of a set of programs

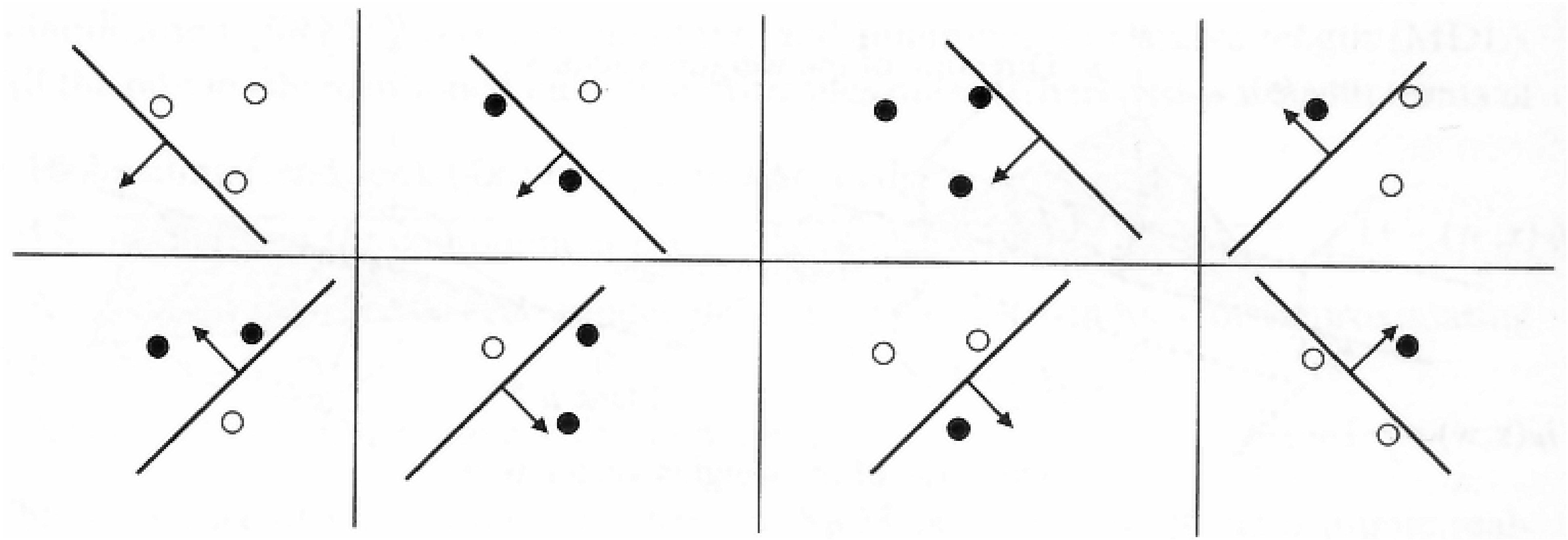


image: V. Kecman, 2001

Does DIGITS work?

Assume:

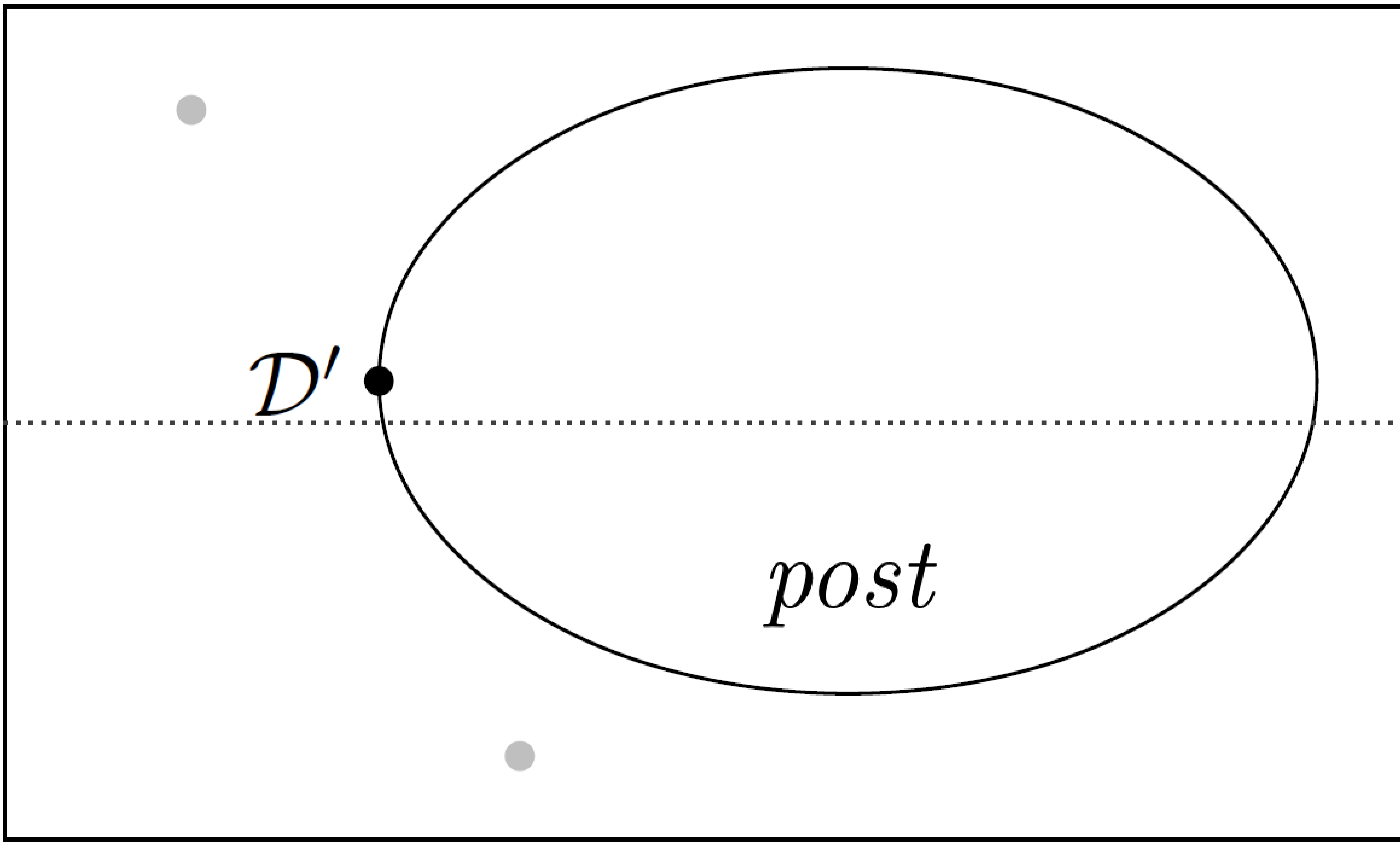
- there exists an optimal solution $\mathcal{D}^*$
- repair model $RM(\mathcal{D})$ has finite VC-dimension d

Does DIGITS work?

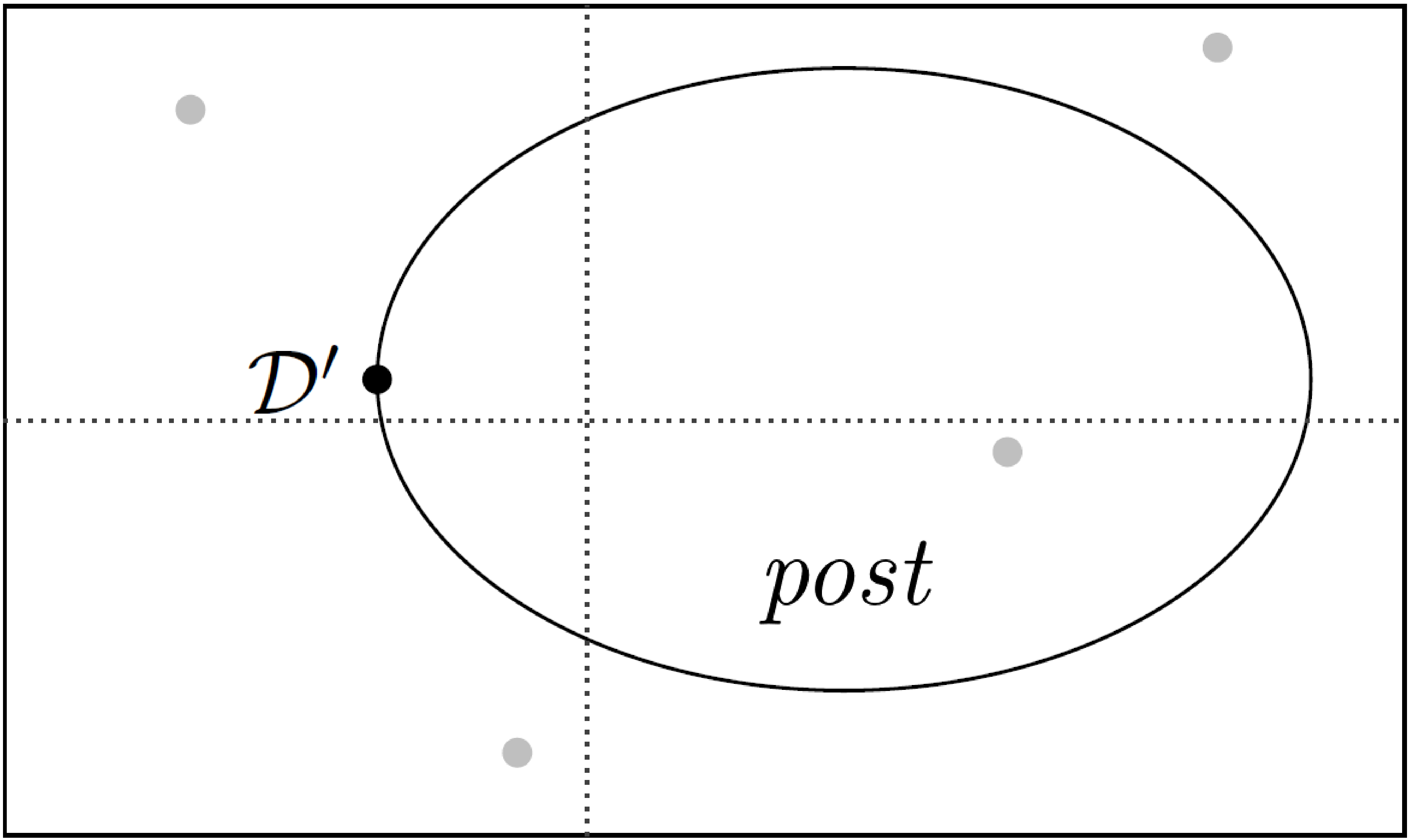
Assume:

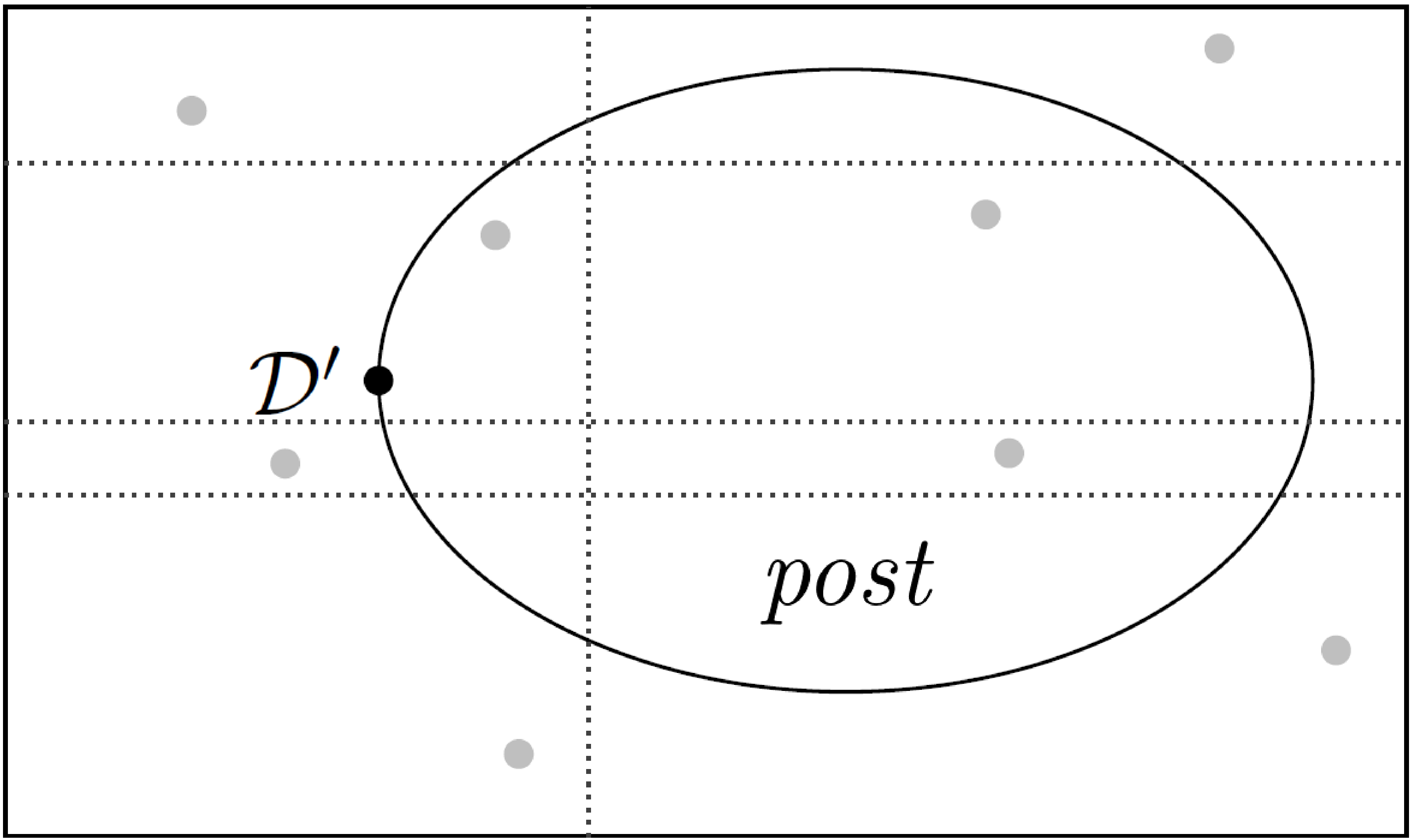
- there exists an optimal solution $\mathcal{D}^*$
- repair model $RM(\mathcal{D})$ has finite VC-dimension d

For every $\varepsilon > 0, \delta > 0$,
if we run DIGITS on $n = poly(d, \varepsilon, \delta)$ samples,
then with probability $> 1 - \delta$ we find a solution $\mathcal{D}'$
with $diff(\mathcal{D}^*, \mathcal{D}') < \varepsilon$.

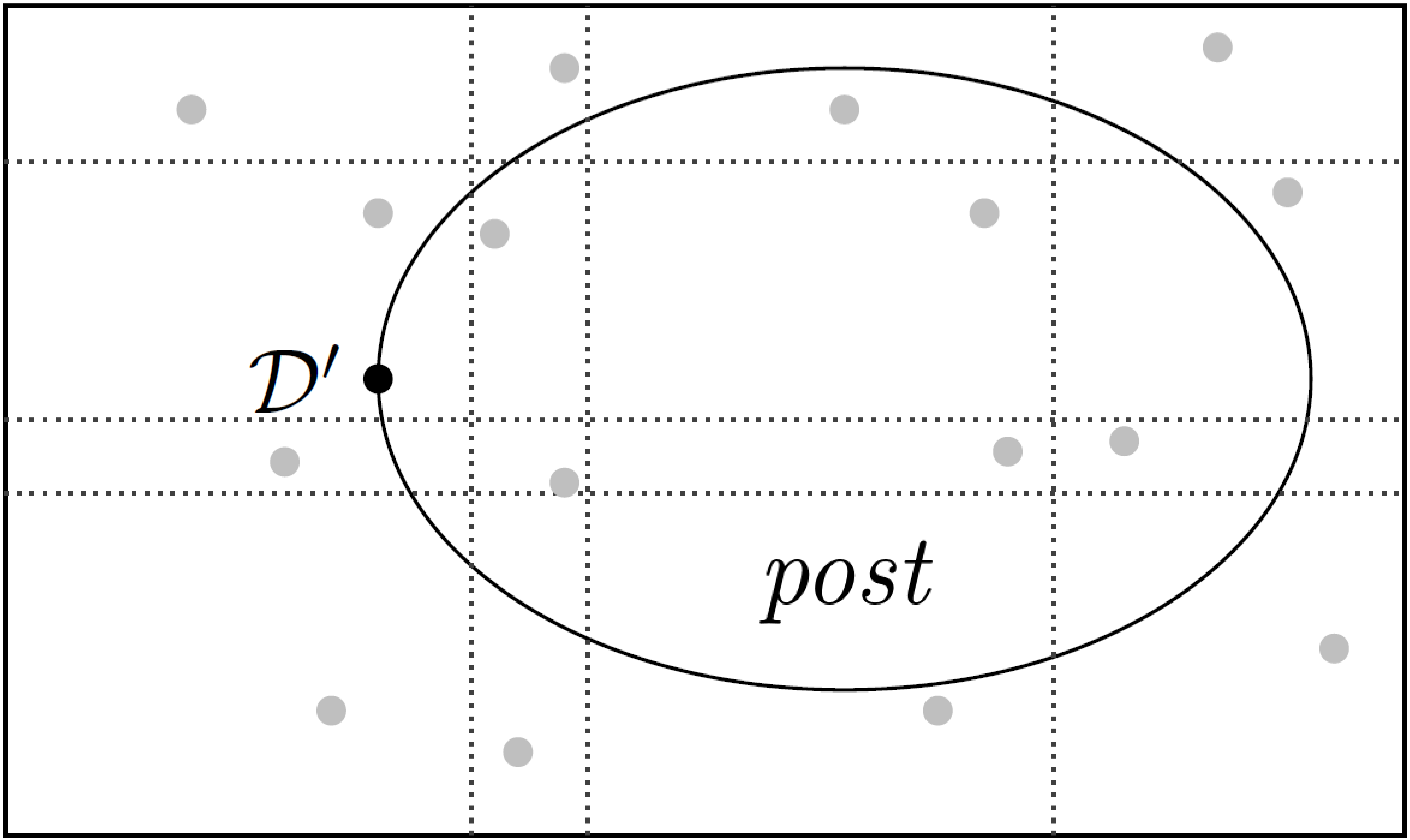


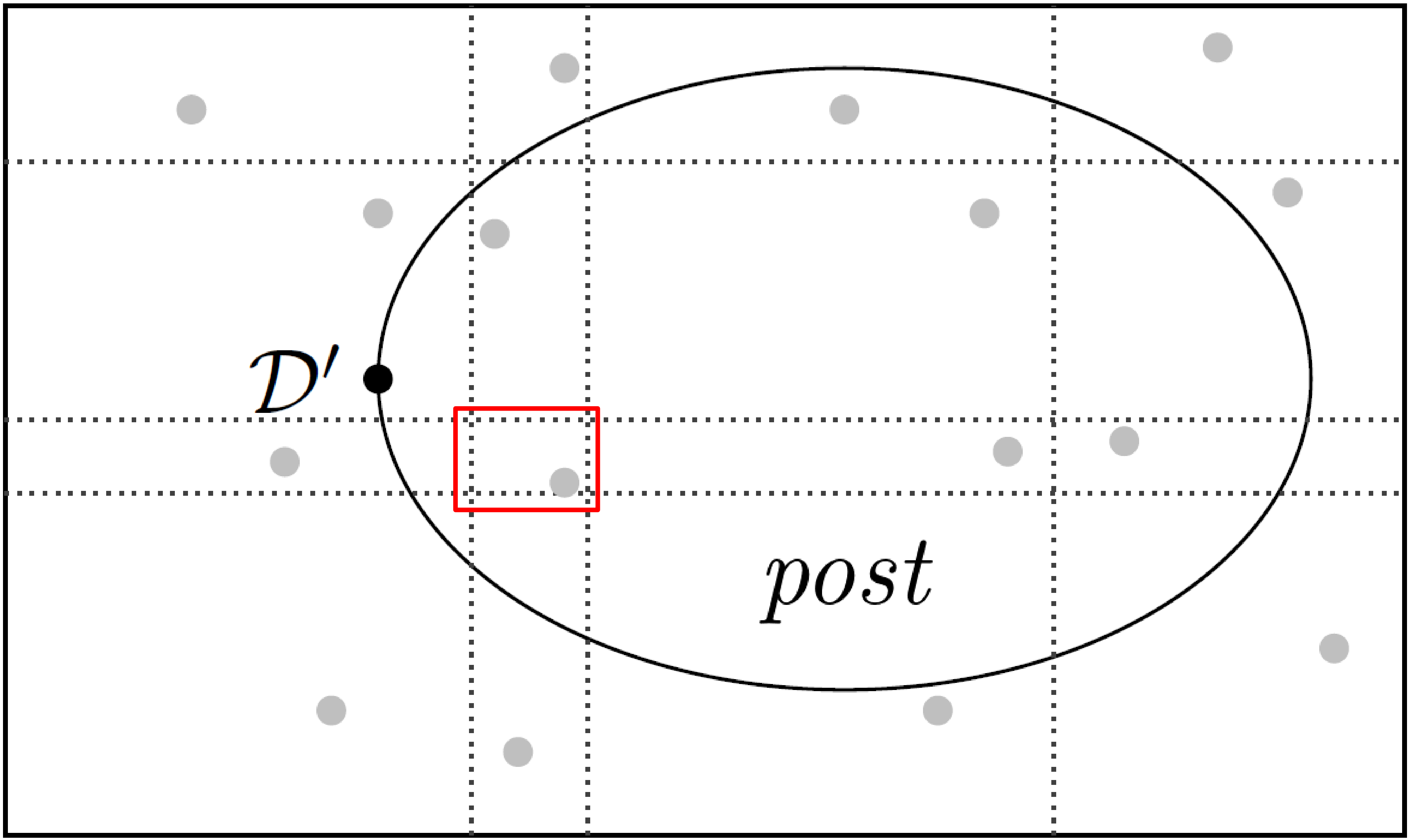
$\mathcal{D}'$
$post$


$\mathcal{D}'$
post



$\mathcal{D}'$
post


$\mathcal{D}'$
$post$



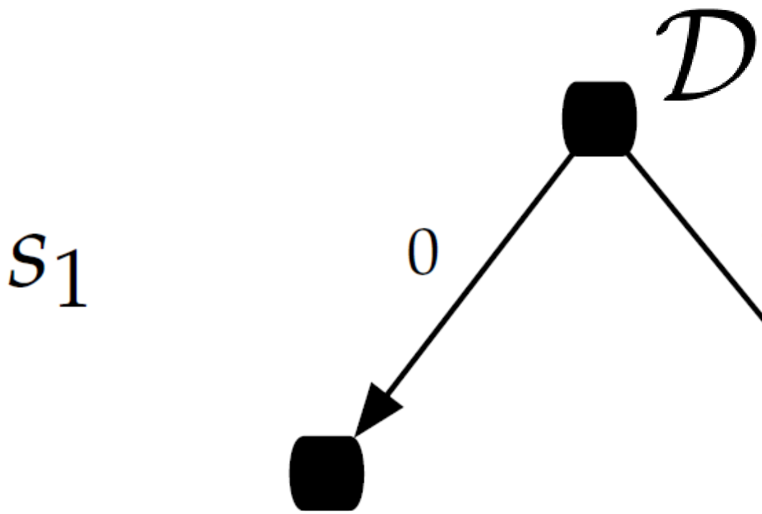
$\mathcal{D}'$
$post$

Trie Structure

Sample one point at a time

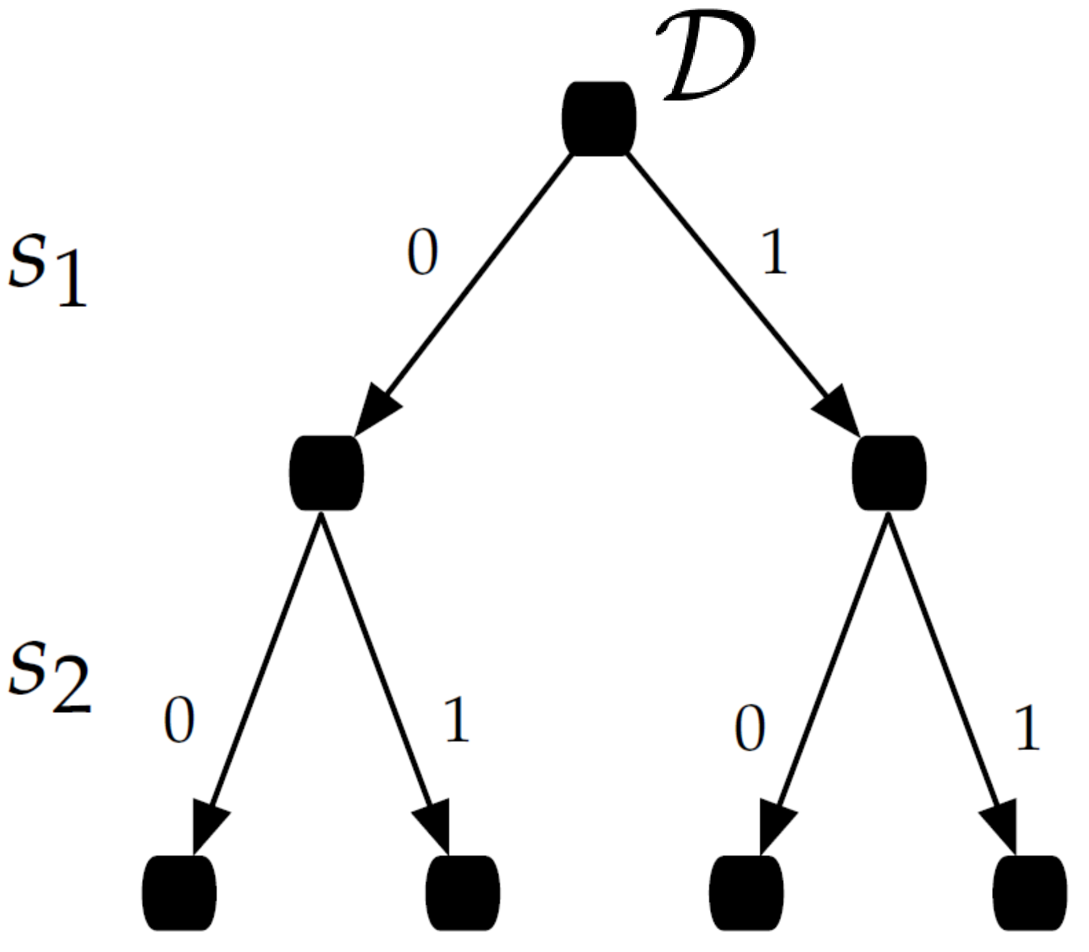
Trie Structure

Sample one point at a time



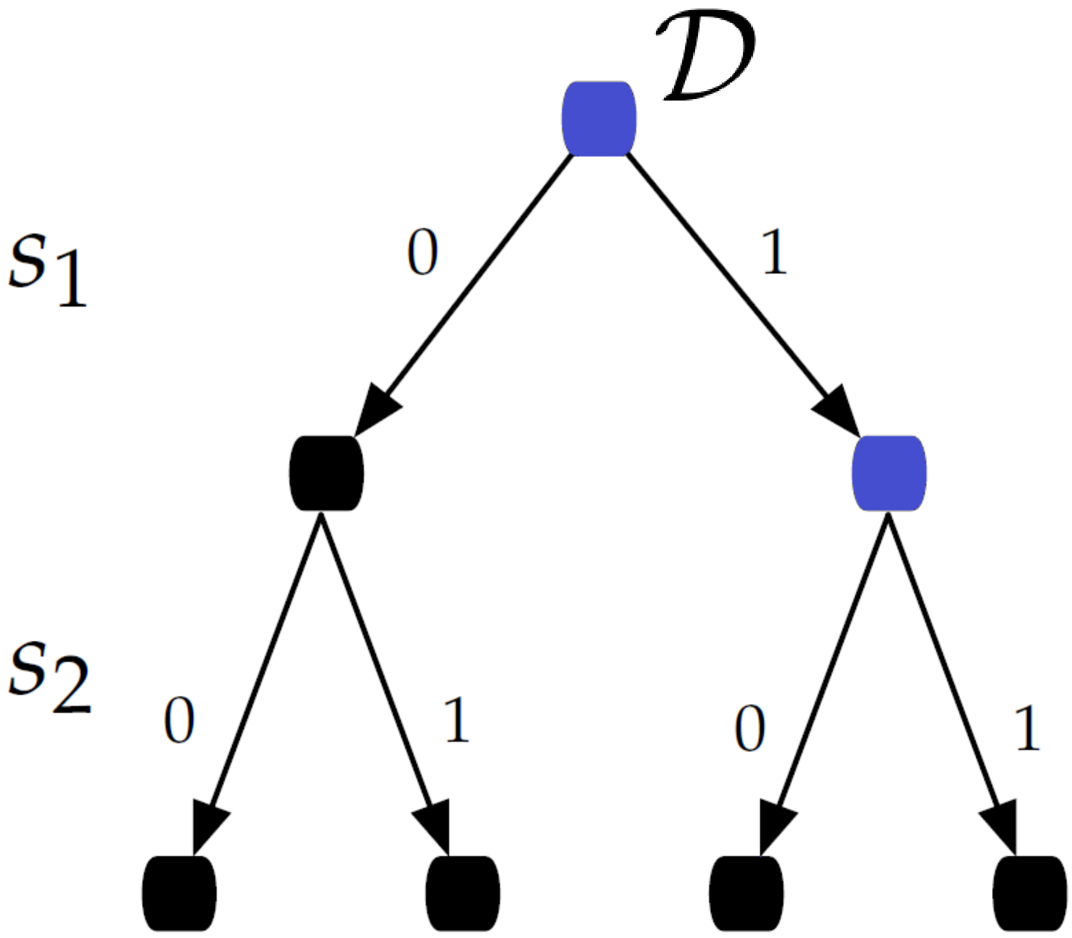
Trie Structure

Sample one point at a time



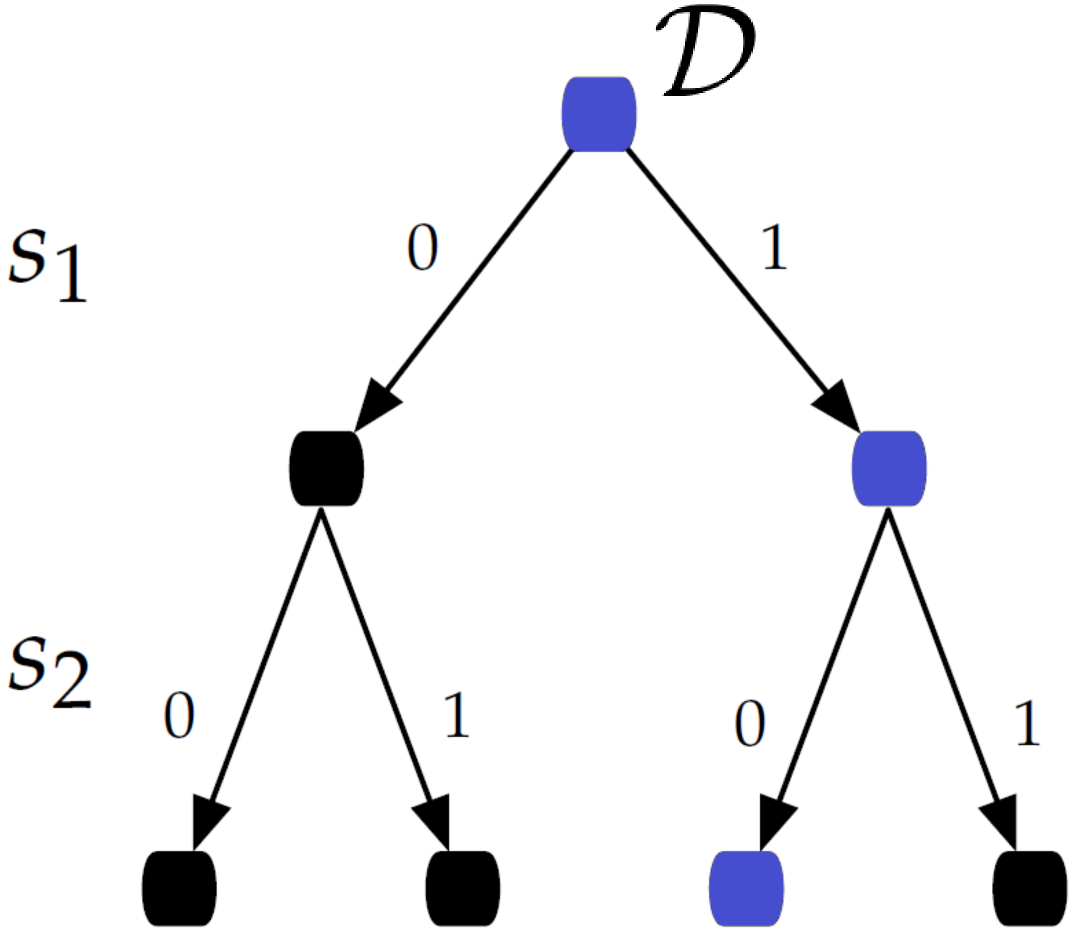
Trie Structure

Sample one point at a time



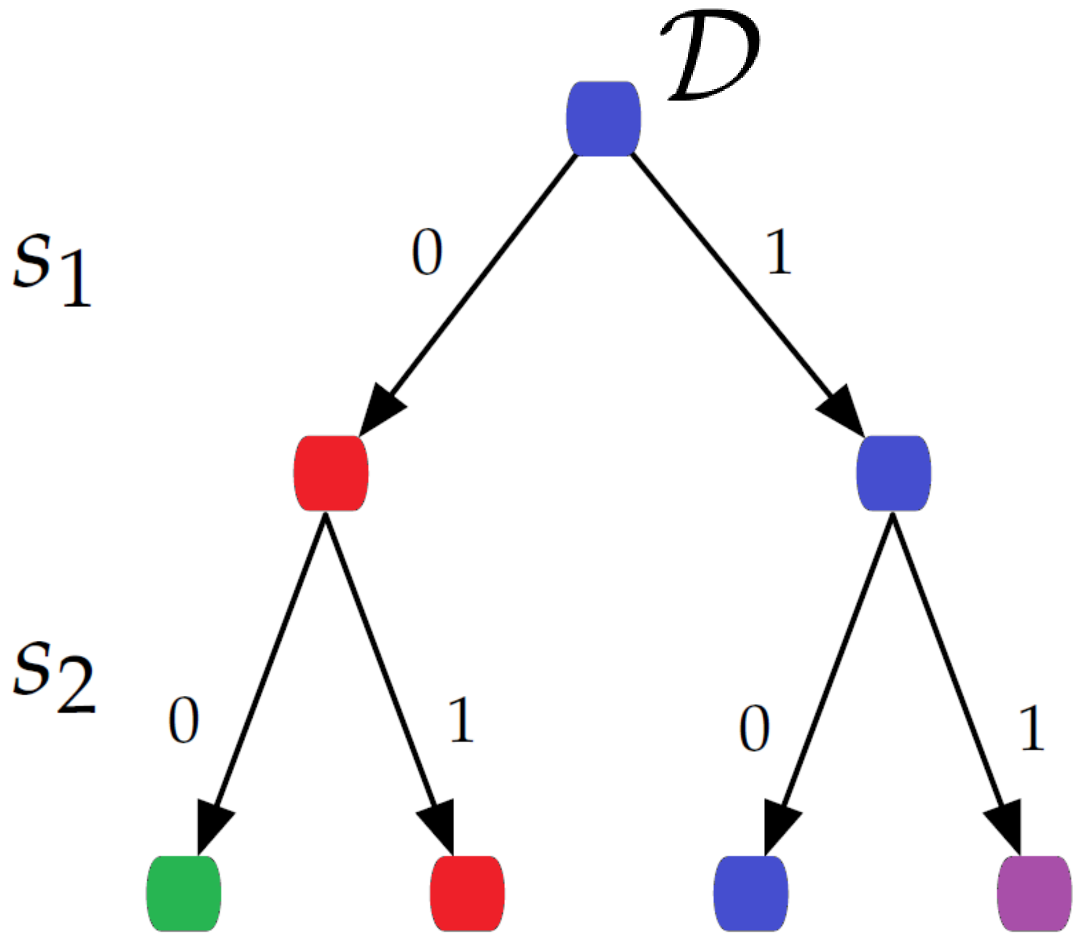
Trie Structure

Sample one point at a time



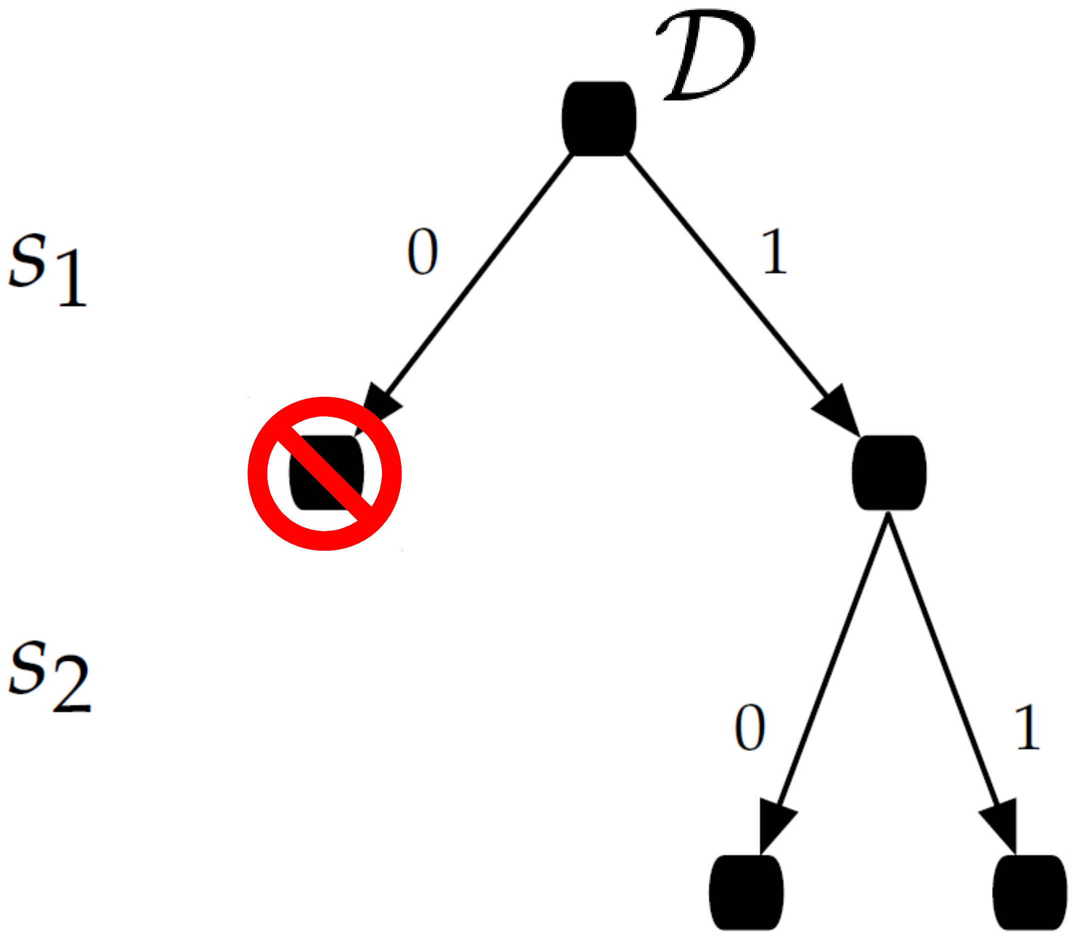
Trie Structure

Sample one point at a time



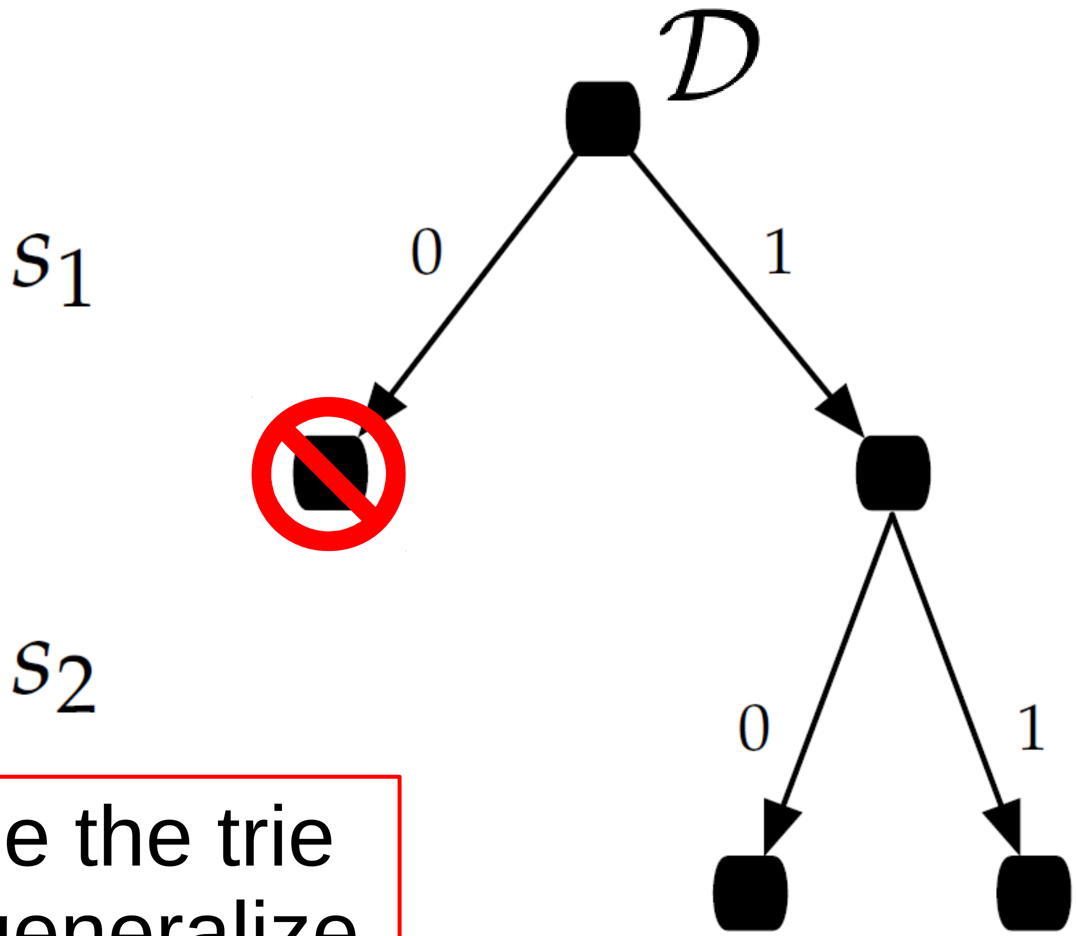
Trie Structure

Sample one point at a time



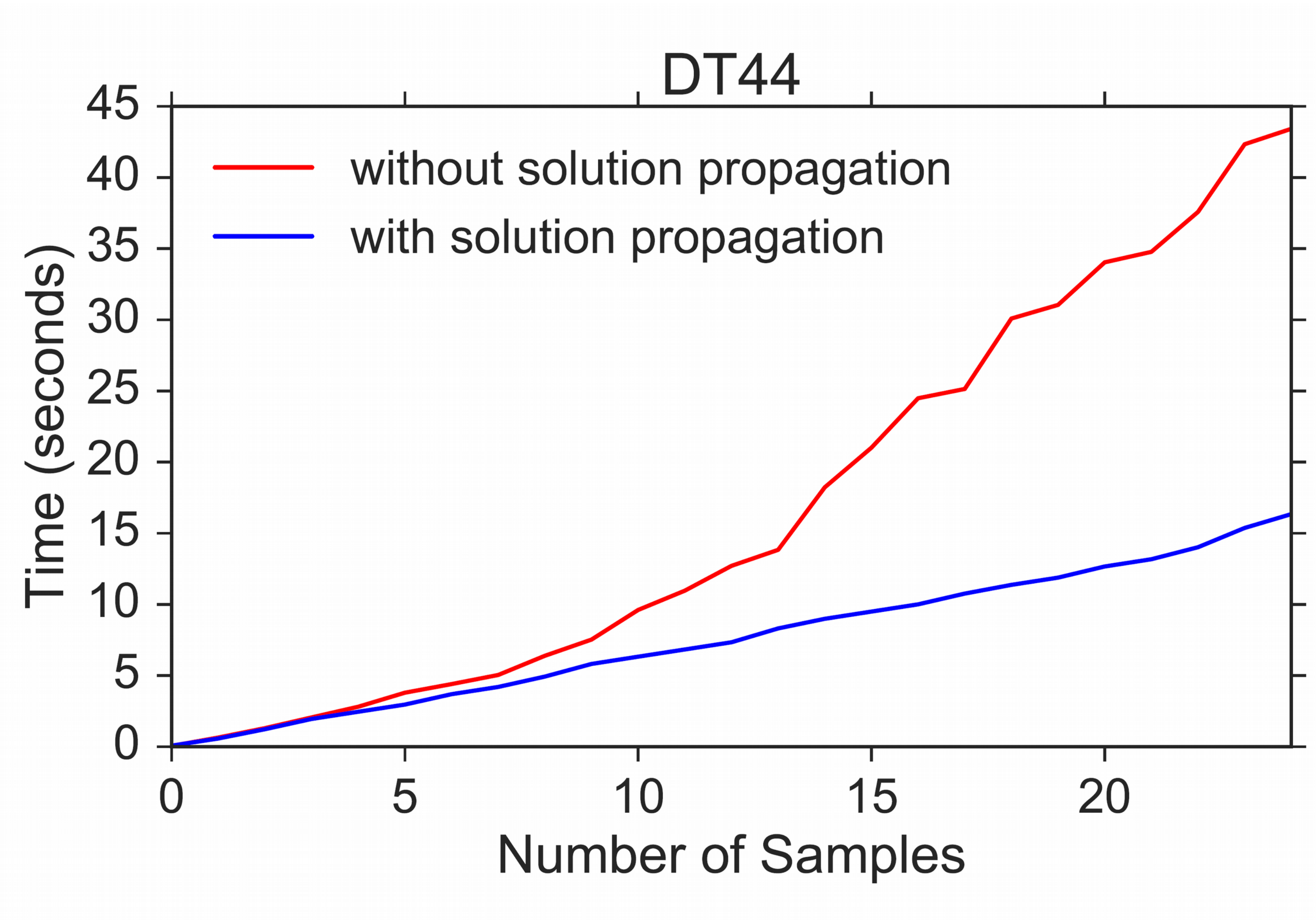
Trie Structure

Sample one point at a time

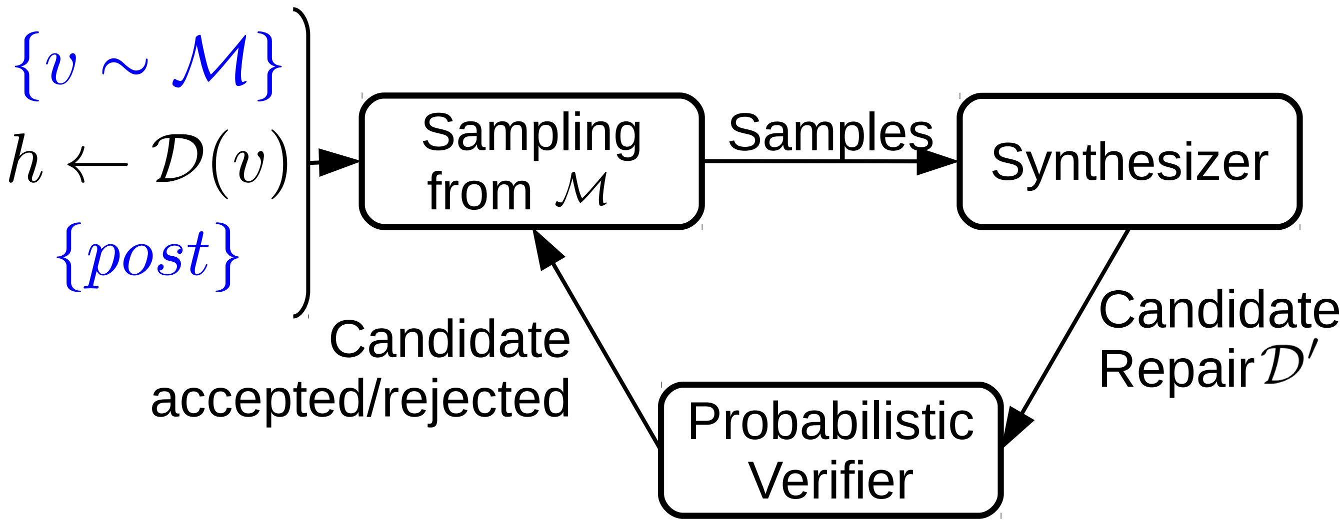


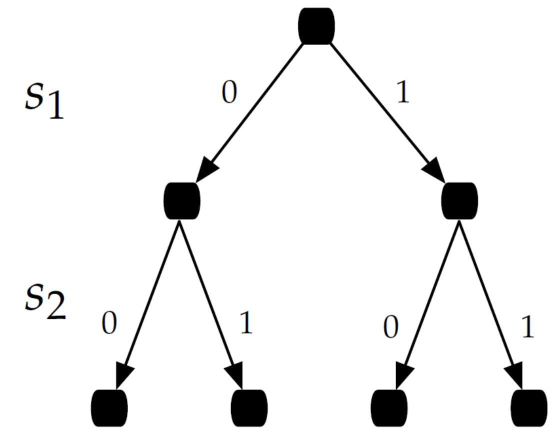
Prune the trie
and generalize

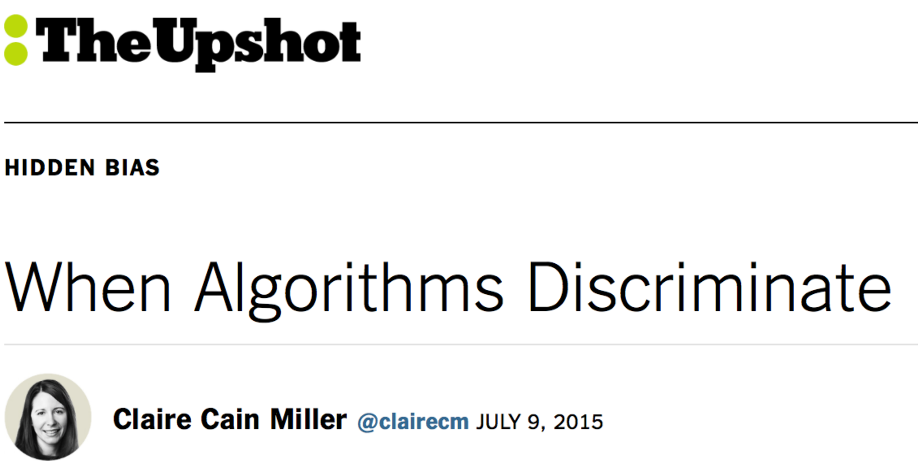
Trie Savings



DIGITS







TheUpshot

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015