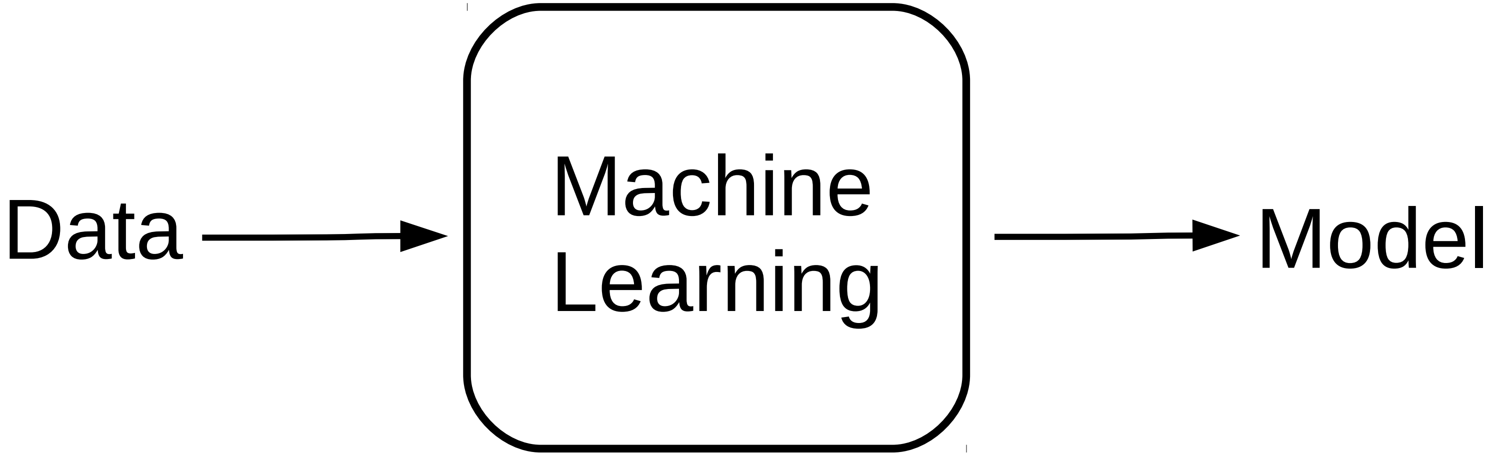
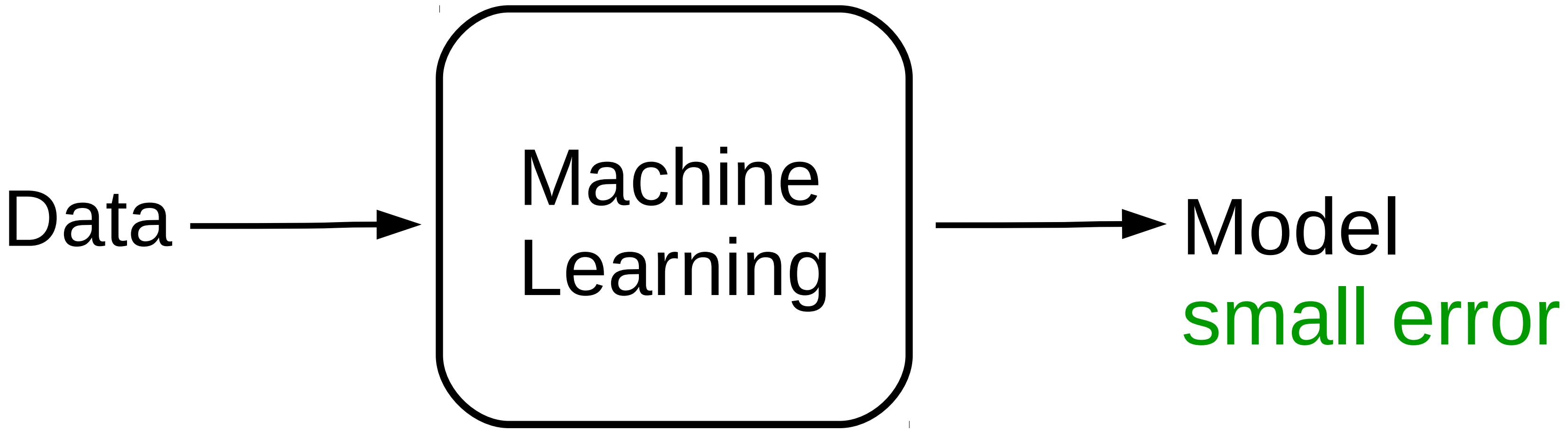
FairSquare:
Probabilistic Verification of Program Fairness

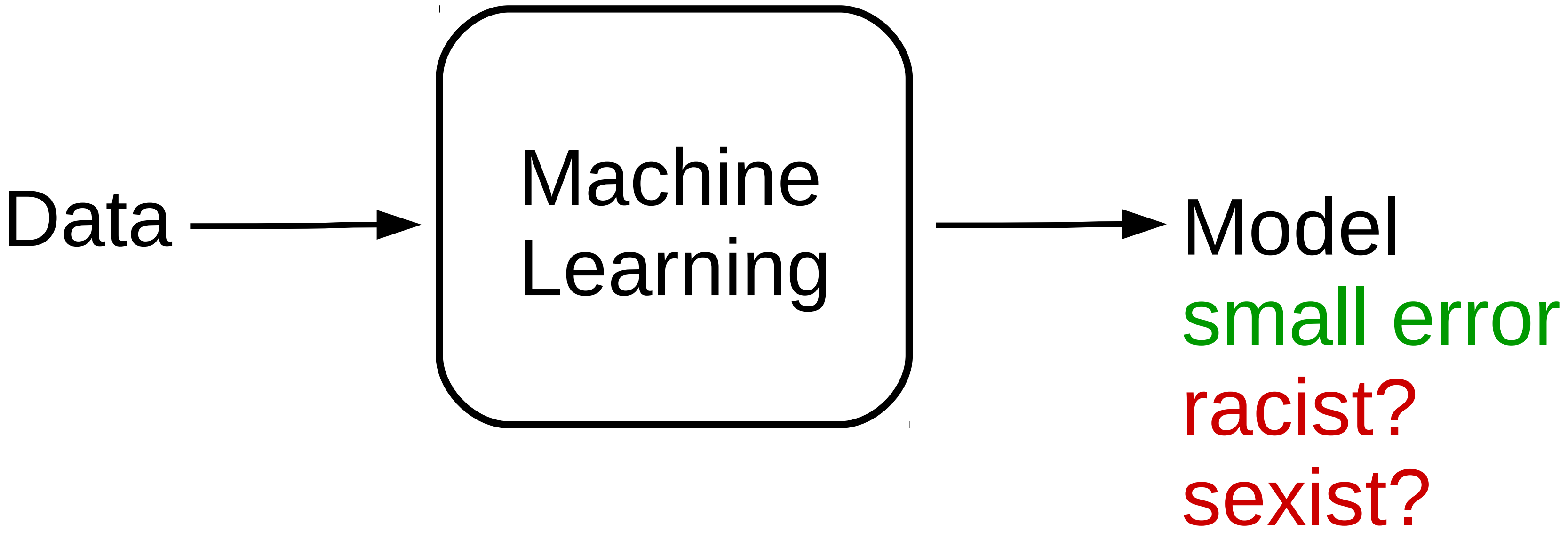
Aws Albarghouthi
Loris D'Antoni
Samuel Drews
University of Wisconsin-Madison
madPL

Aditya V. Nori
Microsoft Research





Data
Machine
Learning
Model


Data
Machine
Learning
Model
small error


Data
Machine
Learning
Model
small error
racist?
sexist?

SundayReview | OPINION

Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

Bianca Bagnarelli

TheUpshot

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Group Fairness

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

decision-
making
program

Group Fairness

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

sensitive feature (e.g. minority)

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

Group Fairness

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

sensitive feature (e.g. minority)

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

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

Group Fairness

population model

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

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

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

Individual Fairness

$$\{v_1, v_2 \sim \mathcal{M}\}$$

$$h_1 \leftarrow \mathcal{D}(v_1)$$

$$h_2 \leftarrow \mathcal{D}(v_2)$$

$$\{\Pr[h_1 \neq h_2 \mid v_1 \sim v_2] < \epsilon\}$$

similarity

Q: What are *good* definitions of fairness?

Q: What are *good* definitions of fairness?

A: Someone else's problem

Contributions

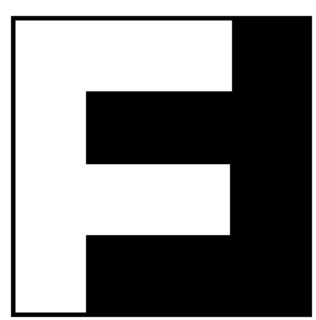
- Formalize fairness definitions as probabilistic verification problems

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

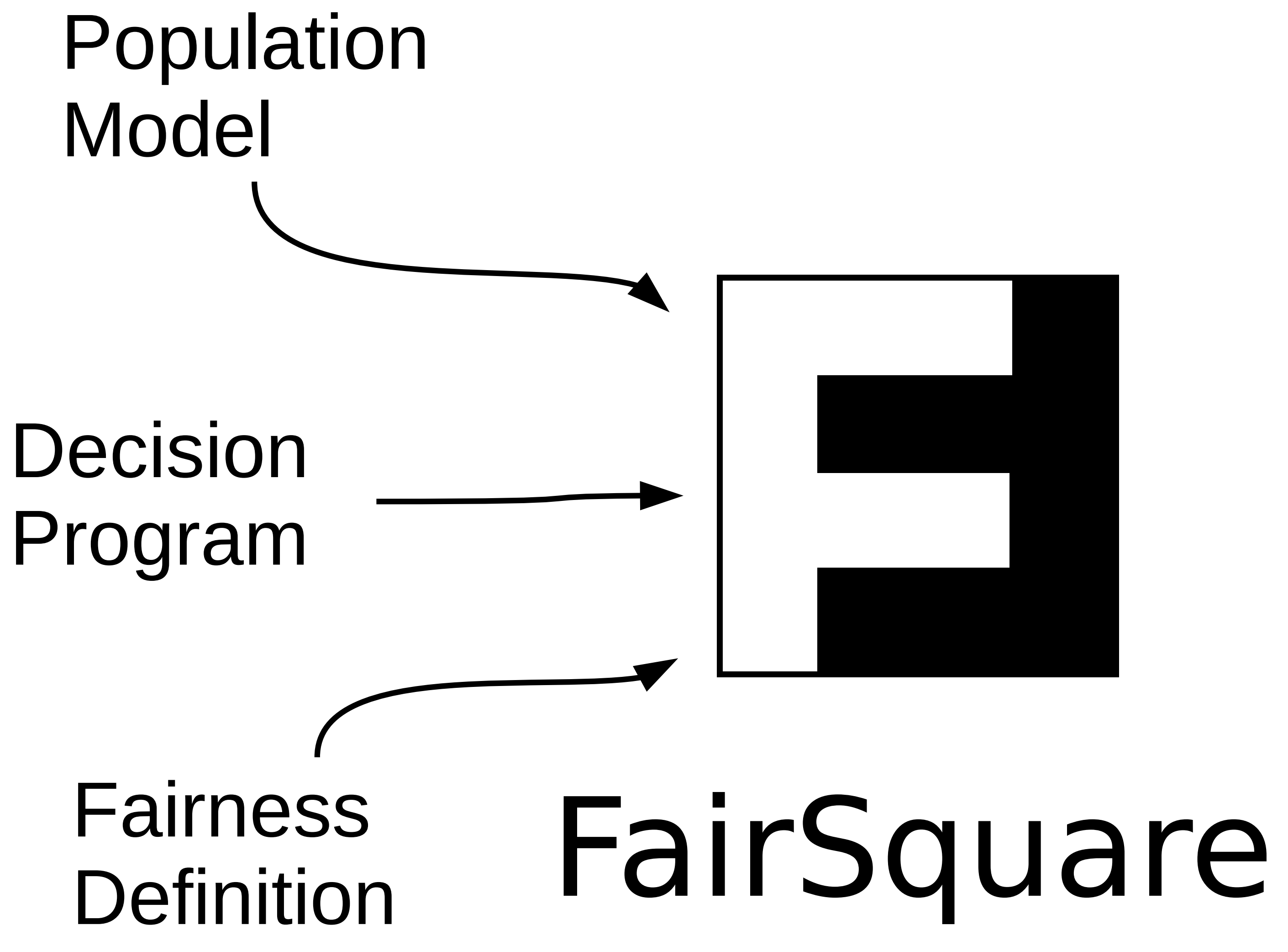
$$h \leftarrow \mathcal{D}(v)$$

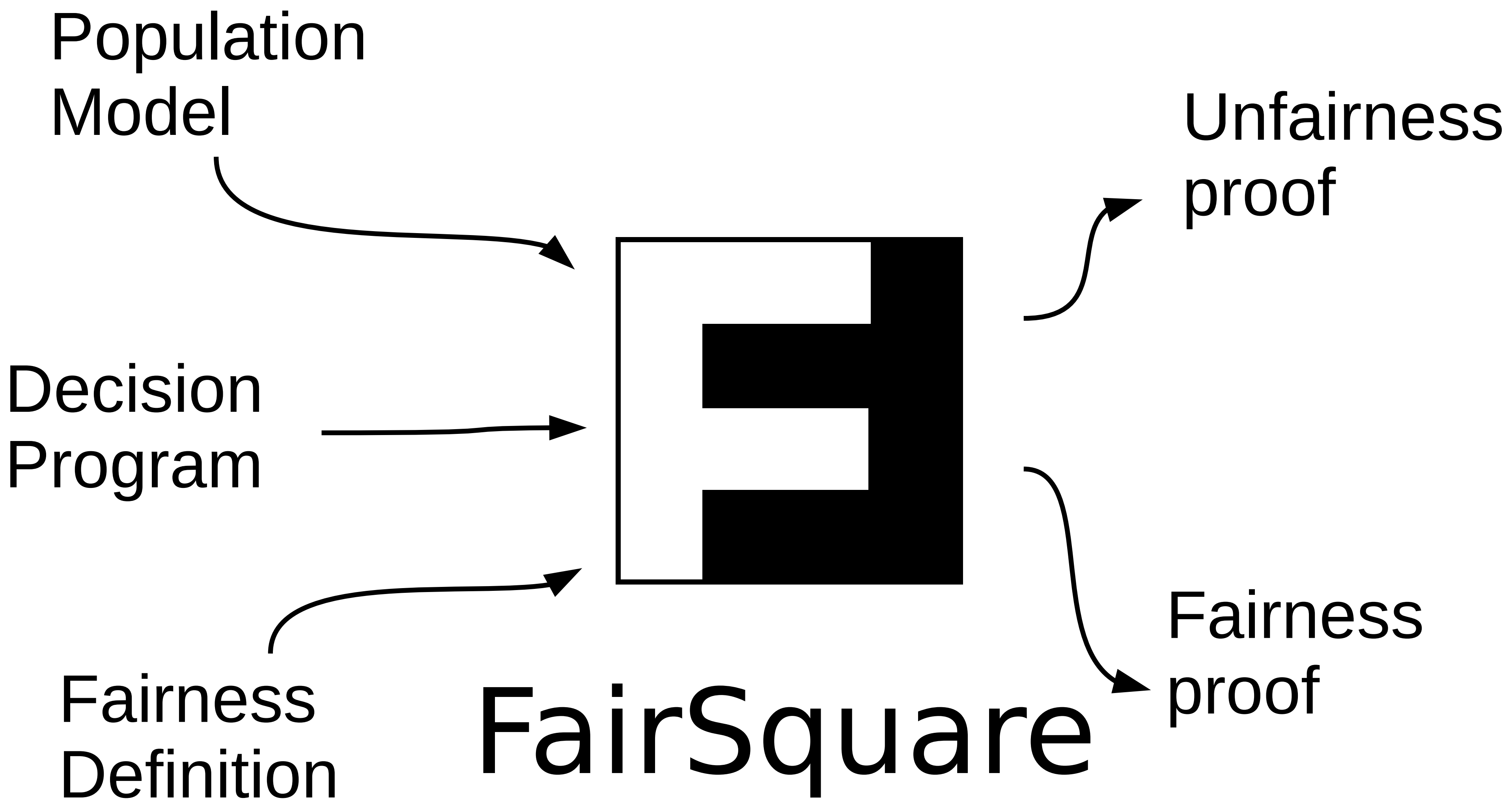
$$\left\{\frac{\Pr[h \mid v_s]}{\Pr[h \mid \neg v_s]} > 1 - \epsilon\right\}$$

- Decision procedure for evaluating probabilistic postconditions



FairSquare


Population
Model
Decision
Program
Fairness
Definition
FairSquare


Population Model
Decision Program
Fairness Definition
FairSquare
Unfairness proof
Fairness proof

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

```
define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

$$\left\{ \frac{\Pr[\texttt{hire} \mid \texttt{ethnicity > 10}]}{\Pr[\texttt{hire} \mid \texttt{ethnicity <= 10}]} > 1 - \epsilon \right\}$$

population model

```
define popModel()
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if (ethnicity > 10)
    colRank ← colRank + 5
  return colRank, yExp
```

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

decision-making program

```
define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

population model

```
define popModel()
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if (ethnicity > 10)
    colRank ← colRank + 5
  return colRank, yExp
```

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

decision-making program

```
define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

```
dec(popModel())
```

```
define popModel()
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if (ethnicity > 10)
    colRank ← colRank + 5
  return colRank, yExp
```

```
define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

$$\left\{ \frac{\Pr[\texttt{hire} \mid \texttt{ethnicity > 10}]}{\Pr[\texttt{hire} \mid \texttt{ethnicity <= 10}]} > 0.9 \right\}$$

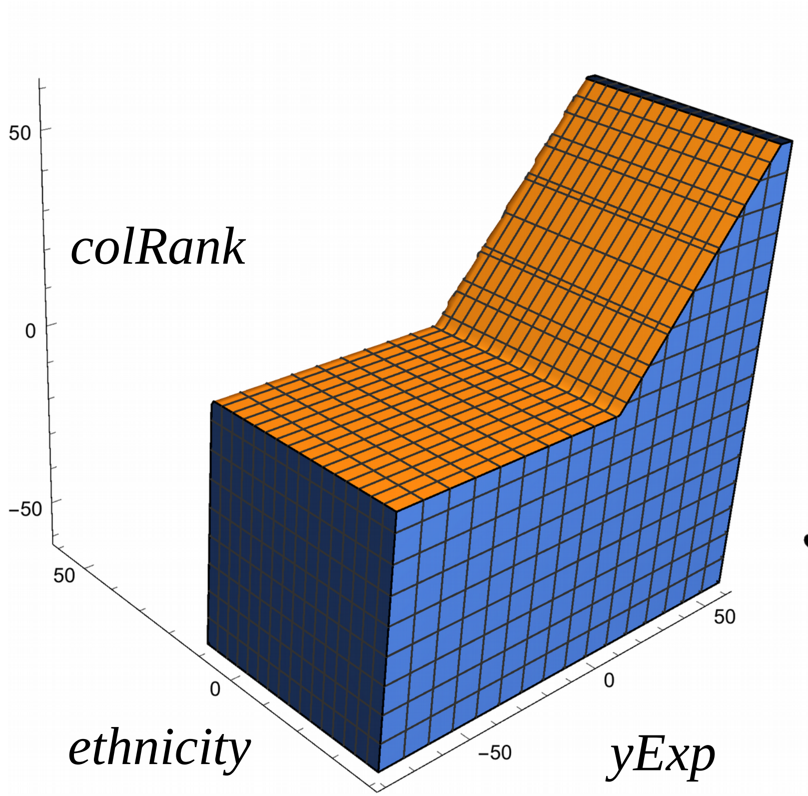
```
define popModel()
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if (ethnicity > 10)
    colRank ← colRank + 5
  return colRank, yExp

define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

$$\left\{ \frac{\Pr[\texttt{hire} \mid \texttt{ethnicity > 10}]}{\Pr[\texttt{hire} \mid \texttt{ethnicity <= 10}]} > 0.9 \right\}$$

$$\Pr[\texttt{hire} \wedge \texttt{ethnicity <= 10}]$$

represent assignments as a region $\varphi \subseteq \mathbb{R}^3$

(LRA formula)



$$\int_{\varphi} p_e(e) p_c(c) p_y(y) \, de \, dc \, dy$$

"weighted volume"

```
define popModel():
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if ethnicity > 10:
    colRank ← colRank + 5
  return colRank, yExp
```

```
define dec(colRank, yExp):
  expRank ← yExp − colRank
  if colRank <= 5:
    hire ← true
  elif expRank > -5:
    hire ← true
  else:
    hire ← false
  return hire
```

$$
\begin{array}{ll}
\texttt{popModel} & \left\{ \begin{array}{l} \varphi = \ ethnicity > 10 \rightarrow colRank' = colRank + 5 \\ \qquad \wedge\ ethnicity \leq 10 \rightarrow colRank' = colRank \end{array} \right. \\
\texttt{dec} & \left\{ \begin{array}{l} \qquad \wedge\ expRank = yExp - colRank' \\ \qquad \wedge\ (colRank' \leq 5 \vee expRank > -5) \rightarrow hire \\ \qquad \wedge\ \neg(colRank' \leq 5 \vee expRank > -5) \rightarrow \neg hire \end{array} \right. \\
\Pr & \left\{ \begin{array}{l} \qquad \wedge\ hire \wedge ethnicity \leq 10 \end{array} \right.
\end{array}
$$

```
define popModel():
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if ethnicity > 10:
    colRank ← colRank + 5
  return colRank, yExp
```

```
define dec(colRank, yExp):
  expRank ← yExp − colRank
  if colRank <= 5:
    hire ← true
  elif expRank > -5:
    hire ← true
  else:
    hire ← false
  return hire
```

$$
\begin{array}{ll}
\texttt{popModel} & \left\{ \begin{array}{l} \varphi = \boxed{ethnicity > 10 \rightarrow colRank' = colRank + 5} \\ \wedge\ ethnicity \leq 10 \rightarrow colRank' = colRank \end{array} \right. \\
\texttt{dec} & \left\{ \begin{array}{l} \wedge\ expRank = yExp - colRank' \\ \wedge\ (colRank' \leq 5 \vee expRank > -5) \rightarrow hire \\ \wedge\ \neg(colRank' \leq 5 \vee expRank > -5) \rightarrow \neg hire \end{array} \right. \\
\Pr & \left\{ \begin{array}{l} \wedge\ hire \wedge ethnicity \leq 10 \end{array} \right.
\end{array}
$$

```
define popModel():
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if ethnicity > 10:
    colRank ← colRank + 5
  return colRank, yExp
```

```
define dec(colRank, yExp):
  expRank ← yExp - colRank
  if colRank <= 5:
    hire ← true
  elif expRank > -5:
    hire ← true
  else:
    hire ← false
  return hire
```

popModel

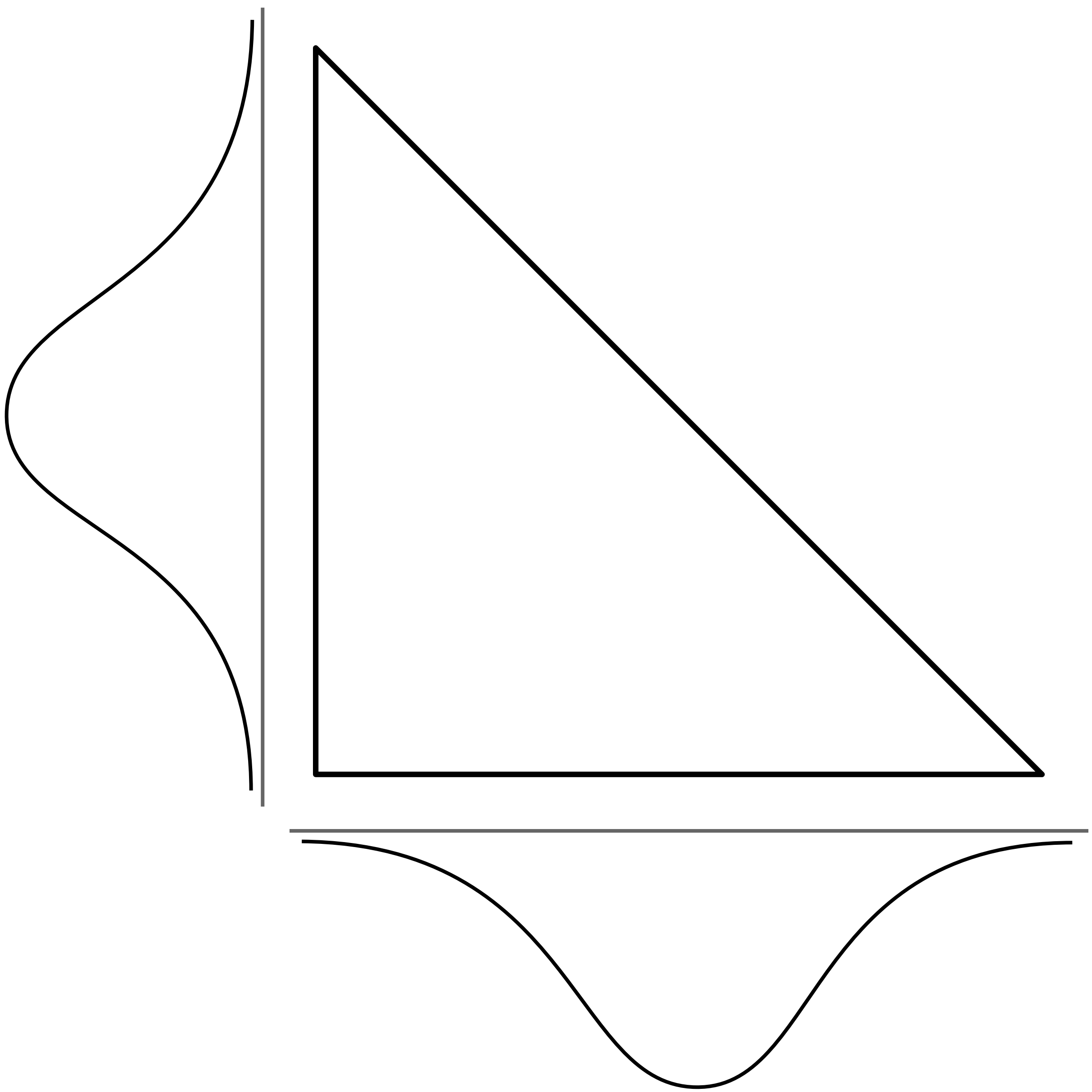
$$\varphi = \mathit{ethnicity} > 10 \rightarrow \mathit{colRank}' = \mathit{colRank} + 5$$
$$\wedge\ \mathit{ethnicity} \leq 10 \rightarrow \mathit{colRank}' = \mathit{colRank}$$

dec

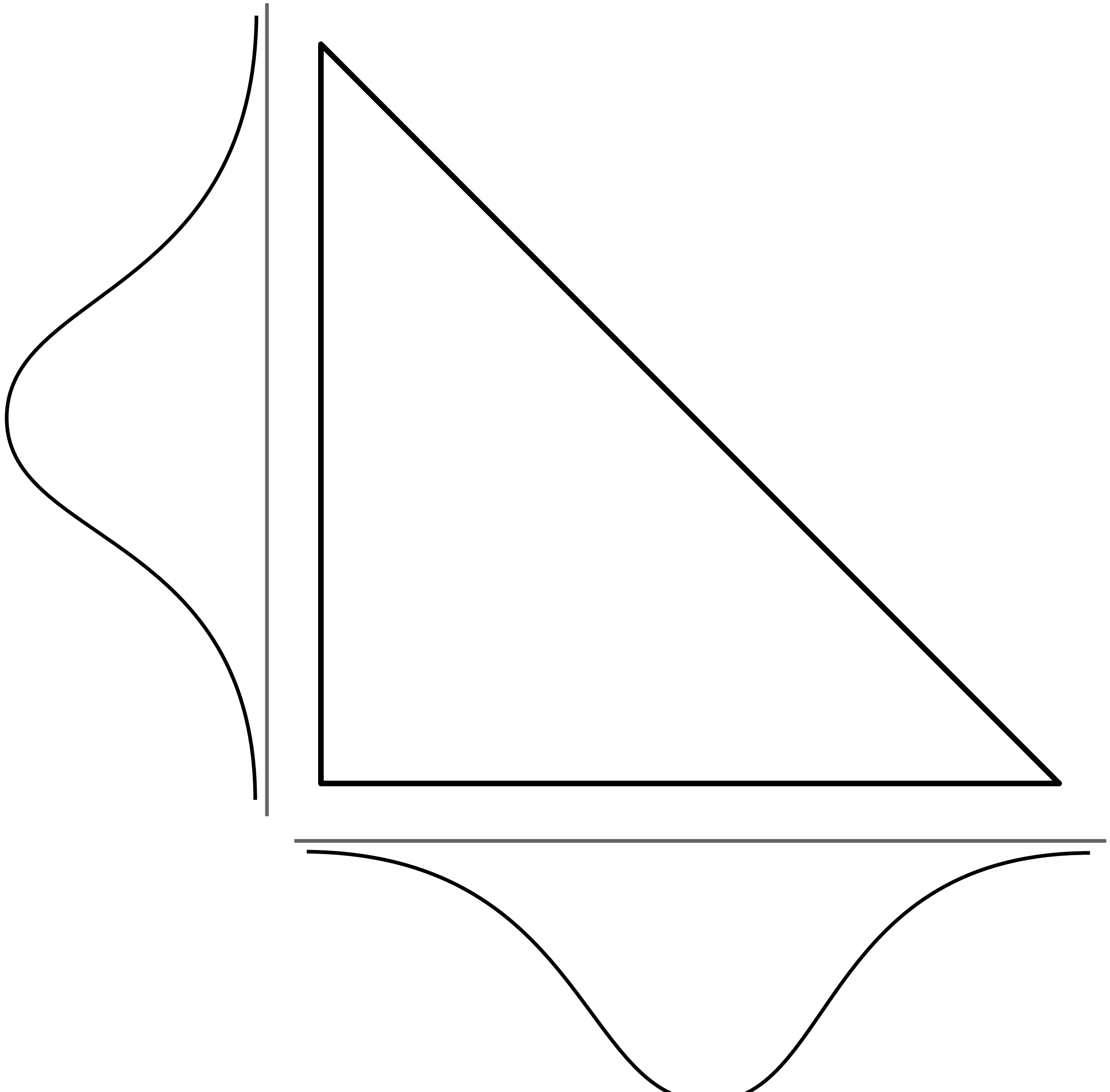
$$\wedge\ \mathit{expRank} = \mathit{yExp} - \mathit{colRank}'$$
$$\wedge\ (\mathit{colRank}' \leq 5 \vee \mathit{expRank} > -5) \rightarrow \mathit{hire}$$
$$\wedge\ \neg(\mathit{colRank}' \leq 5 \vee \mathit{expRank} > -5) \rightarrow \neg \mathit{hire}$$

Pr

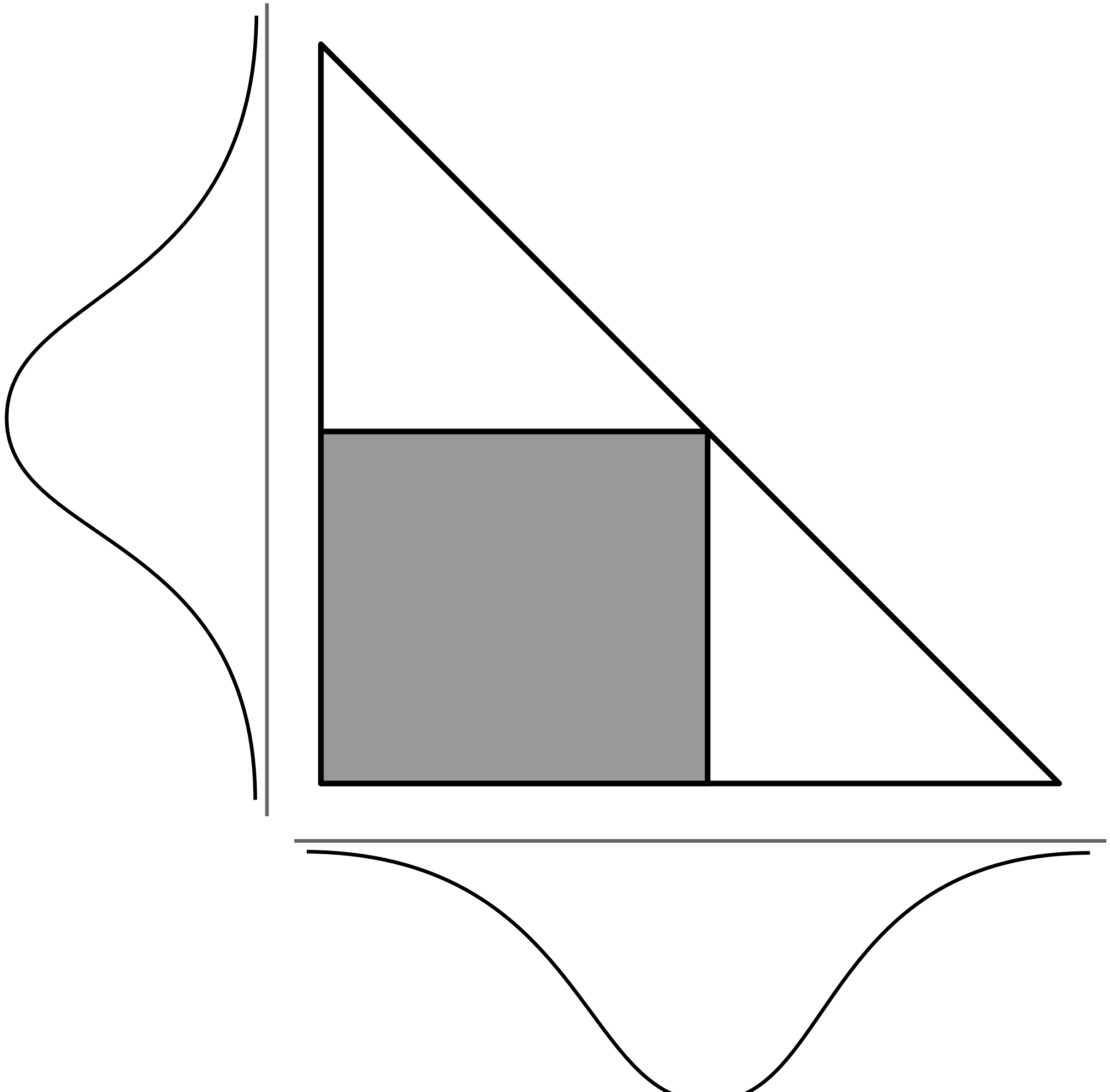
$$\boxed{\wedge\ \mathit{hire} \wedge \mathit{ethnicity} \leq 10}$$



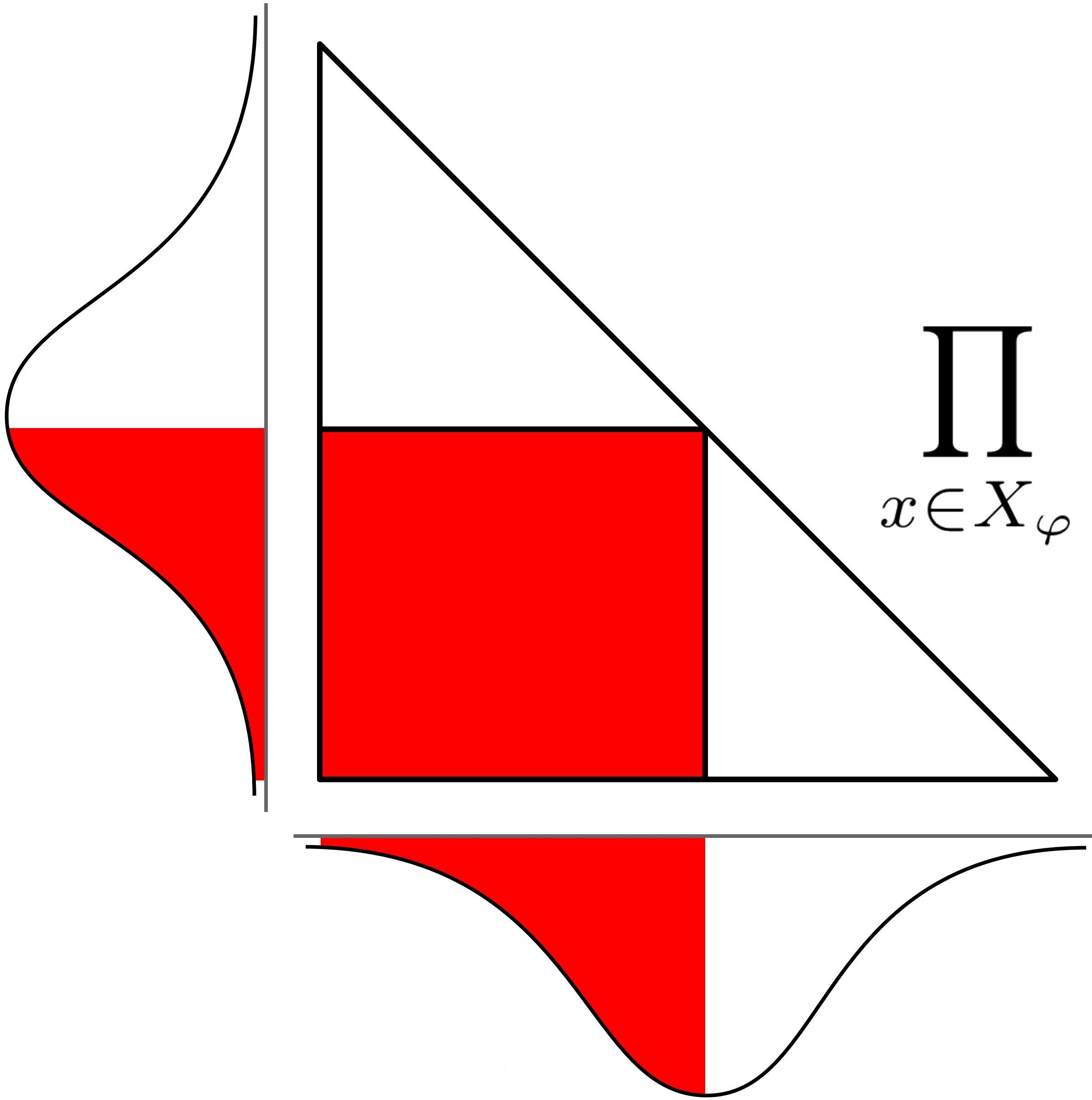
$$\square_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x \right) \wedge \forall X_{\varphi}. \left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x \right) \Rightarrow \varphi \right)$$



$$\Box_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x \right) \wedge \forall X_{\varphi}. \left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x \right) \Rightarrow \varphi \right)$$

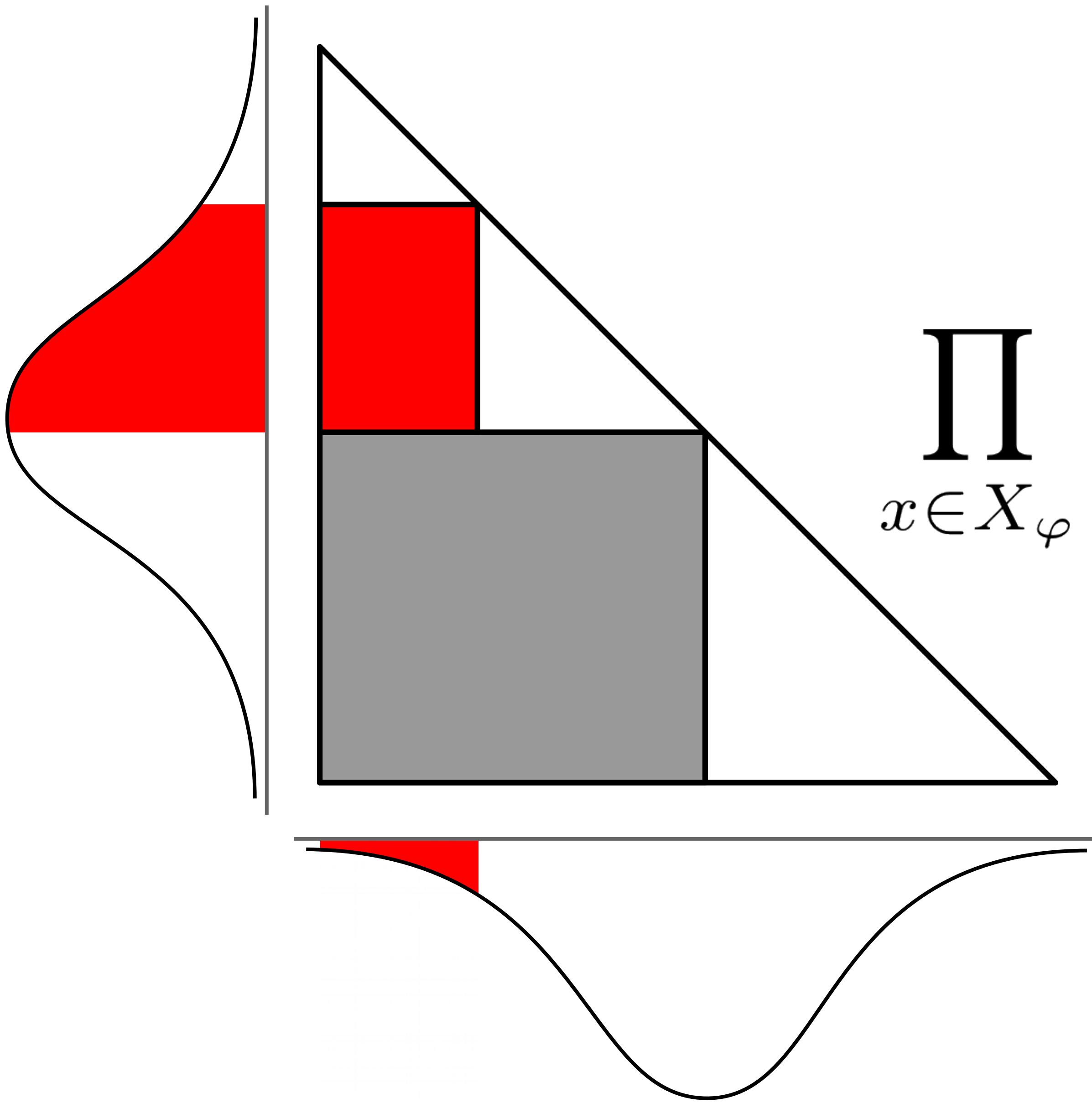


$$\square_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x \right) \wedge \forall X_{\varphi}. \left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x \right) \Rightarrow \varphi \right)$$



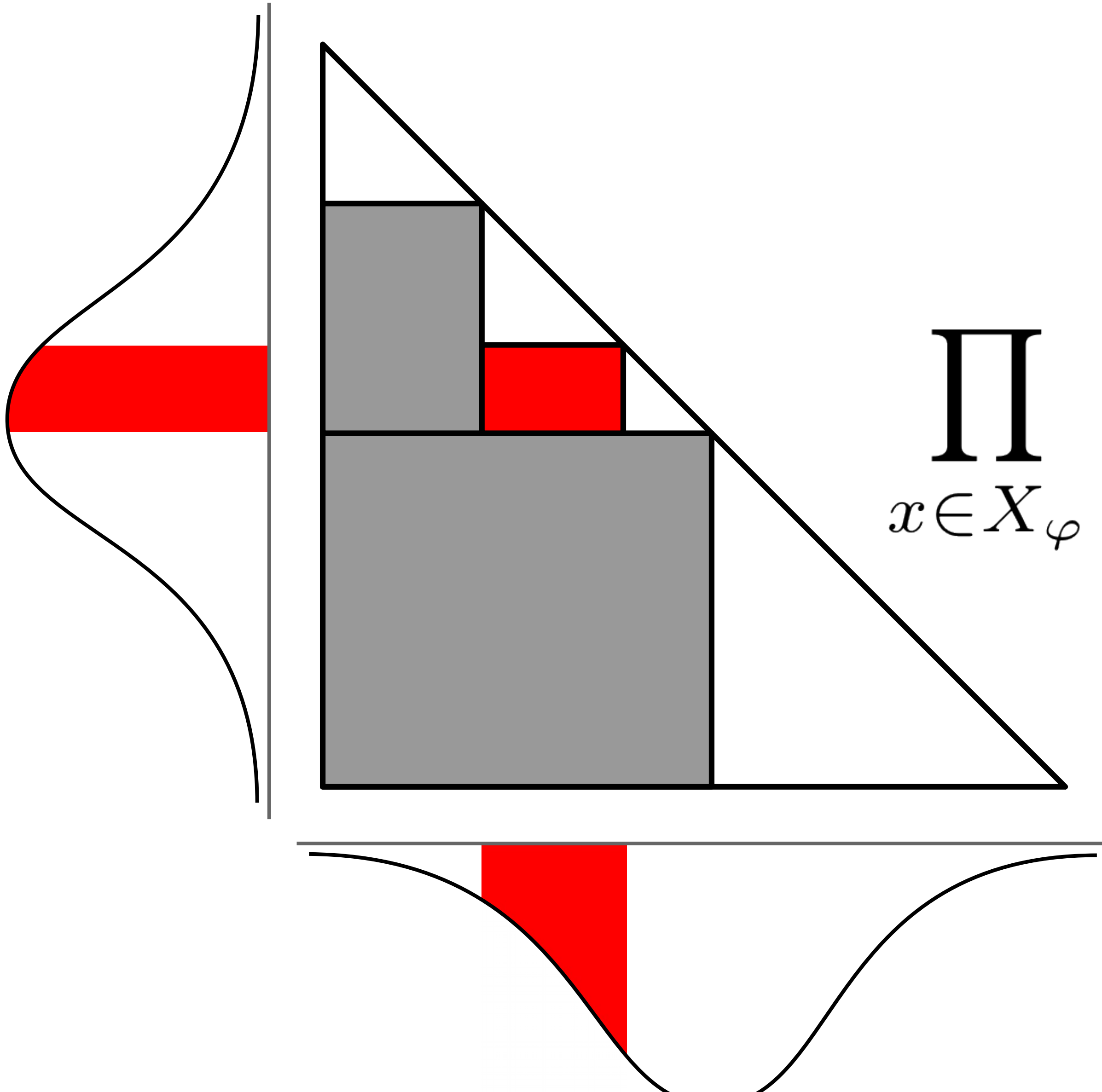
$$\prod_{x \in X_{\varphi}} \int_{l_x}^{u_x} p_x(v)\, dv$$

$$\square_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x\right) \wedge \forall X_{\varphi}.\left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x\right) \Rightarrow \varphi\right)$$



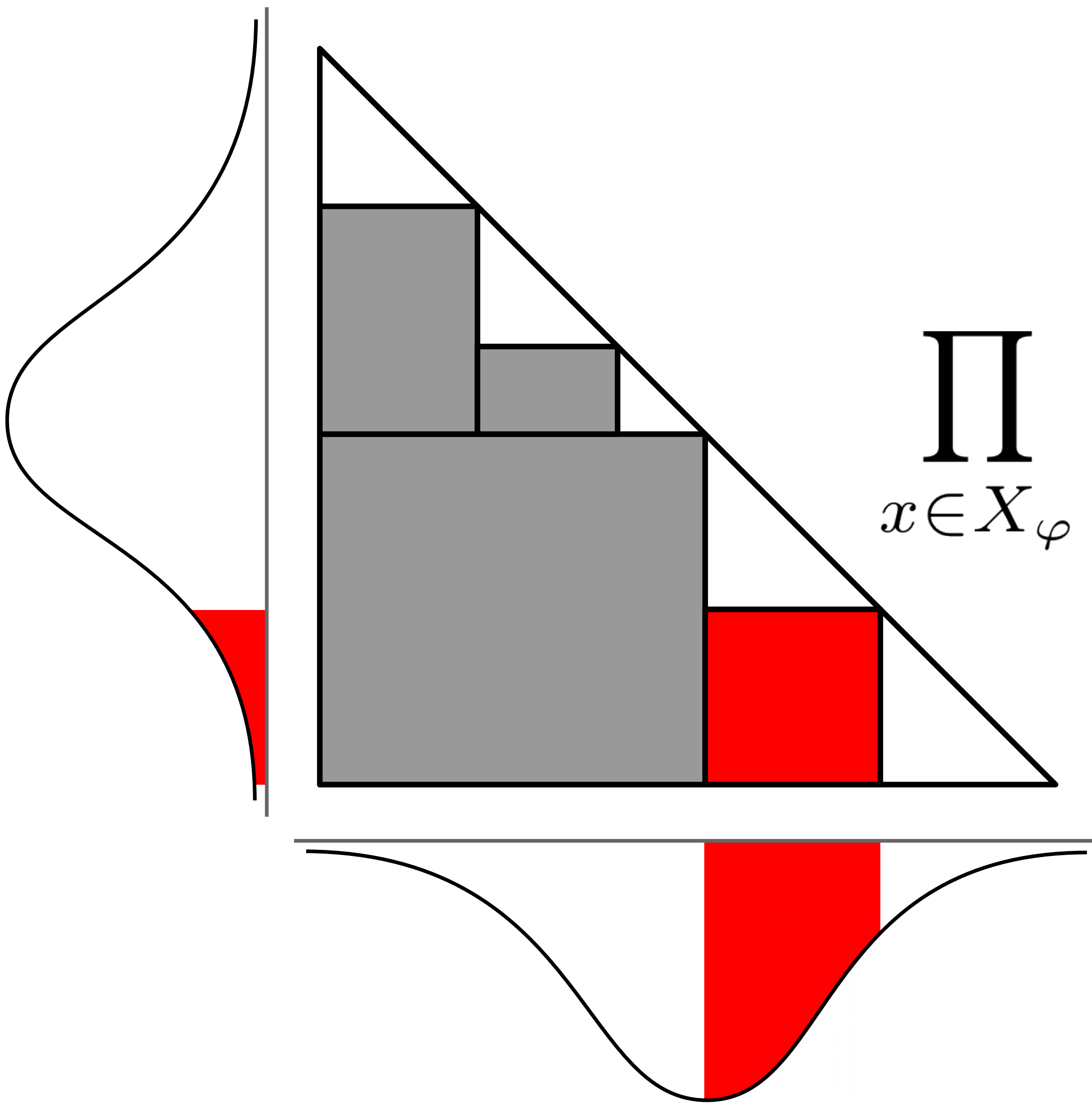
$$\prod_{x \in X_{\varphi}} \int_{l_x}^{u_x} p_x(v)\, dv$$

$$\boxed{\perp}_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x \right) \wedge \forall X_{\varphi}. \left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x \right) \Rightarrow \varphi \right)$$



$$\prod_{x \in X_{\varphi}} \int_{l_x}^{u_x} p_x(v)\, dv$$

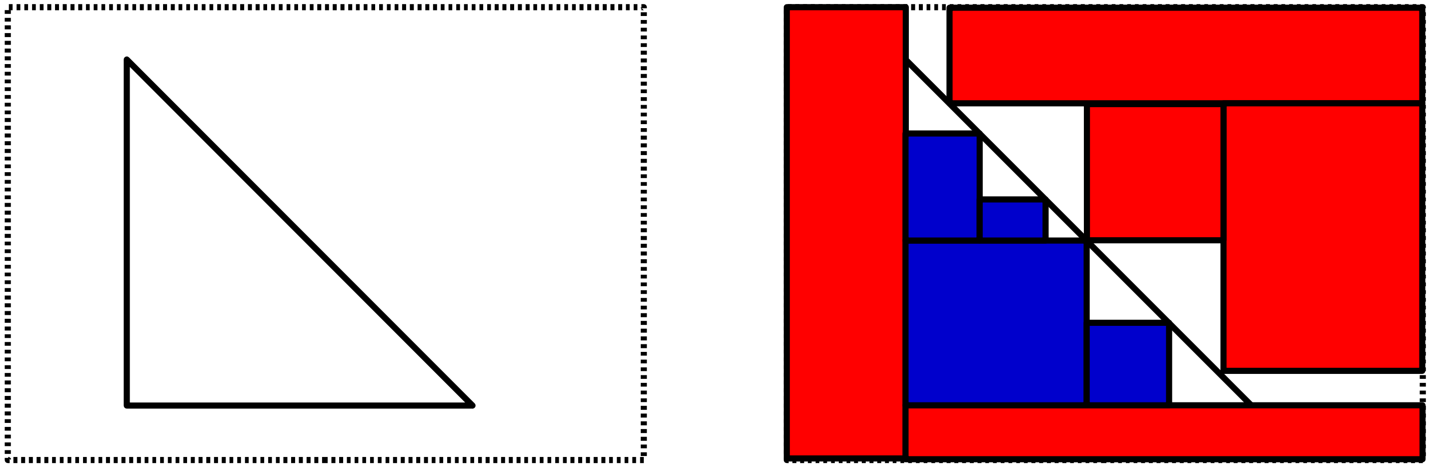
$$\boxed{\perp}_{\varphi} \equiv \left(\bigwedge_{x \in X_{\varphi}} l_x < u_x \right) \wedge \forall X_{\varphi}. \left(\left(\bigwedge_{x \in X_{\varphi}} l_x \leqslant x \leqslant u_x \right) \Rightarrow \varphi \right)$$



$$\prod_{x \in X_{\varphi}} \int_{l_x}^{u_x} p_x(v)\, dv$$

Evaluating *post*

- Obtain formula φ for each probability in *post*
- Underapproximate the weighted volume of φ
- Overapproximate by doing the same for $\neg\varphi$



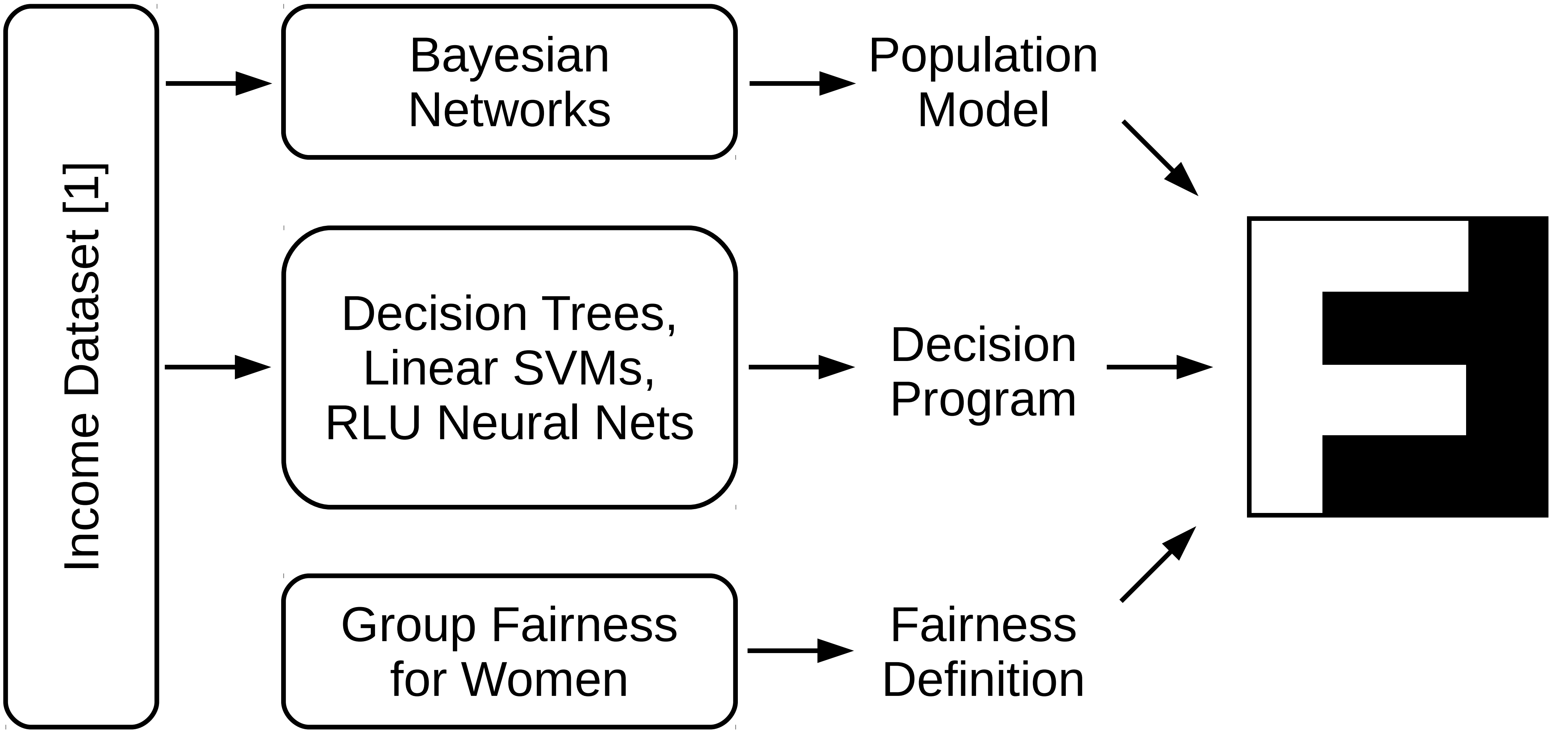
eventually, approximations are *good enough*

“Live” demo

```
define popModel()
  ethnicity ~ gauss(0,10)
  colRank ~ gauss(25,10)
  yExp ~ gauss(10,5)
  if (ethnicity > 10)
    colRank ← colRank + 5
  return colRank, yExp
define dec(colRank, yExp)
  expRank ← yExp - colRank
  if (colRank <= 5)
    hire ← true
  elif (expRank > -5)
    hire ← true
  else
    hire ← false
  return hire
```

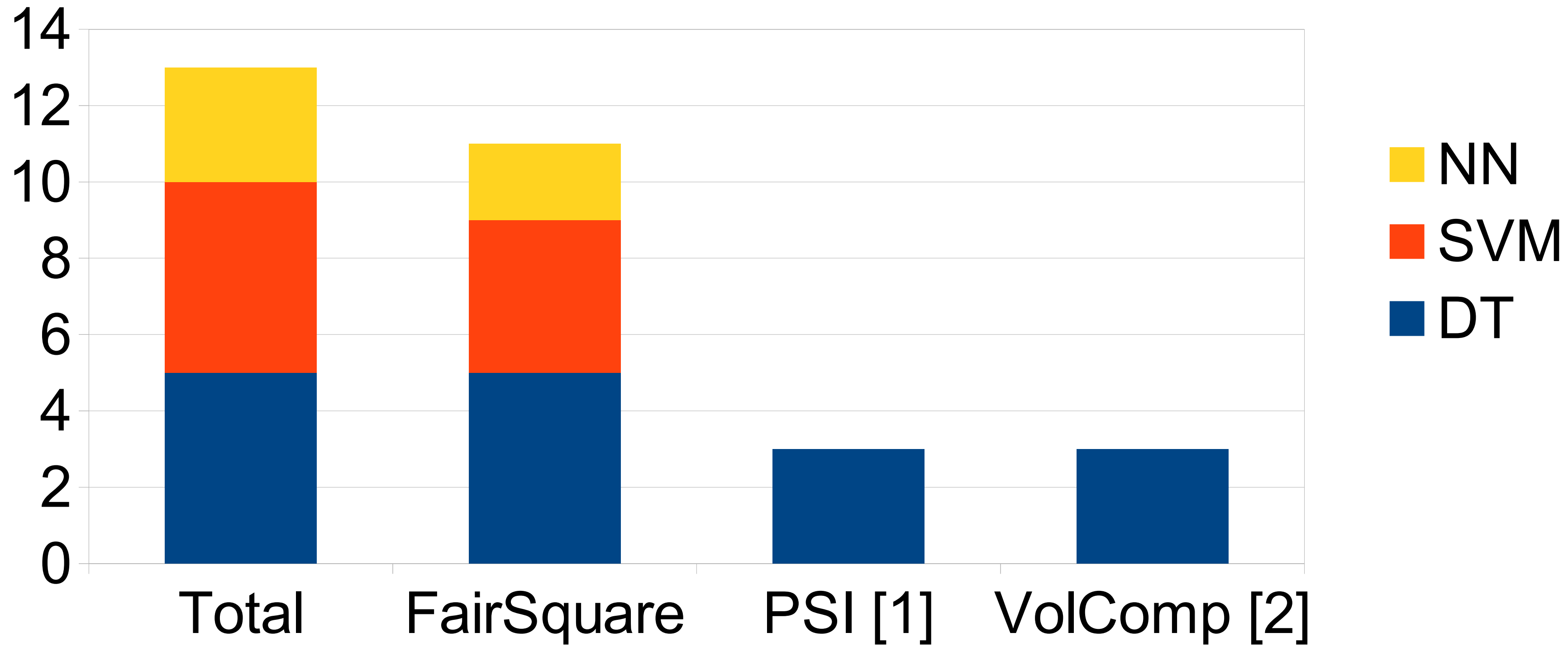
$$\left\{ \frac{\Pr[\texttt{hire} \mid \texttt{ethnicity > 10}]}{\Pr[\texttt{hire} \mid \texttt{ethnicity <= 10}]} > 0.9 \right\}$$

Case Study

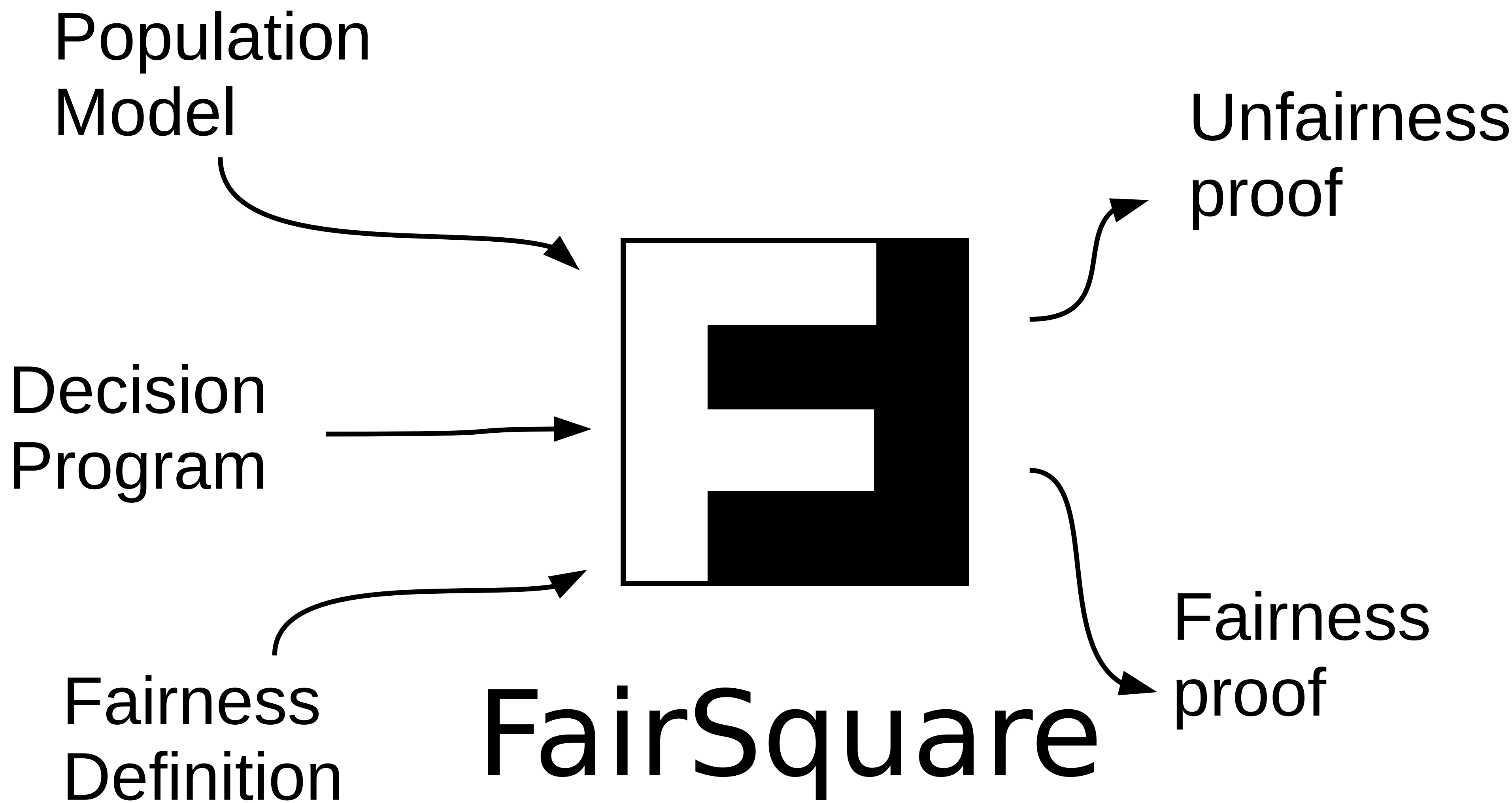


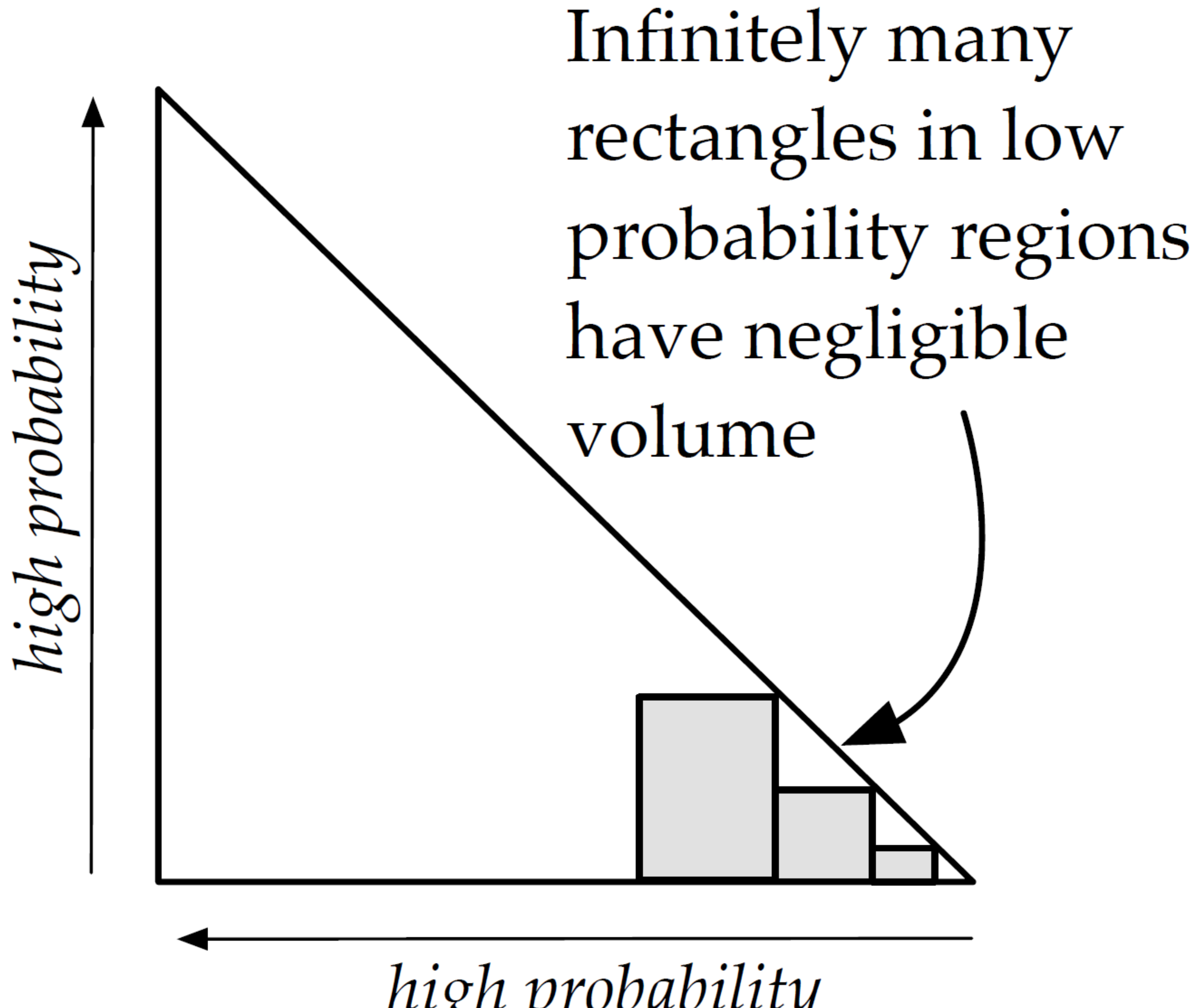
[1] https://archive.ics.uci.edu/ml/datasets/adult

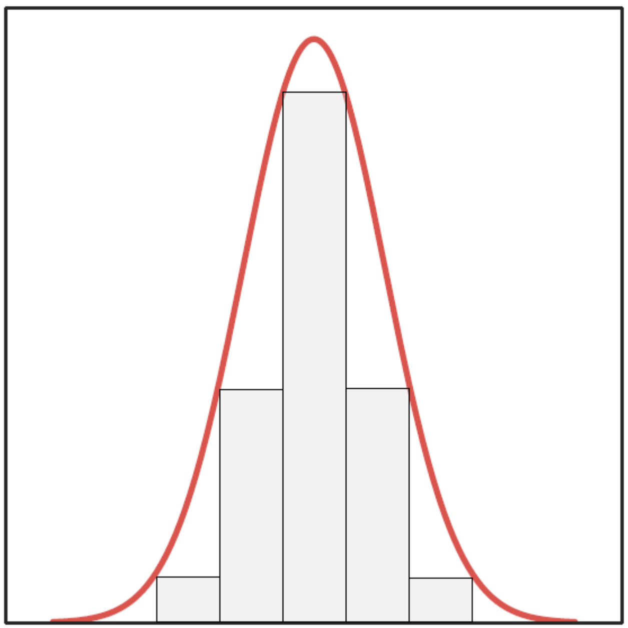
Fairness Verification Problems



[1] Gehr et al. CAV 2016 [2] Sankaranarayanan et al. PLDI 2013


Population
Model
Decision
Program
Fairness
Definition
FairSquare
Unfairness
proof
Fairness
proof


Infinitely many rectangles in low probability regions have negligible volume
high probability
high probability



$$step(x) = \begin{cases} c_i, & x \in [a_i, b_i) \text{ for } 1 \leqslant i \leqslant n \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_x = \sum_{i=1}^{n} c_i \cdot \left| [a_i, b_i) \cap [l_x, u_x] \right|$$