Statistical Monitoring

with focus on Algorithmic Fairness

Konstantin Kueffner

Institute of Science and Technology Austria

Online Monitoring.

More information, but less time.

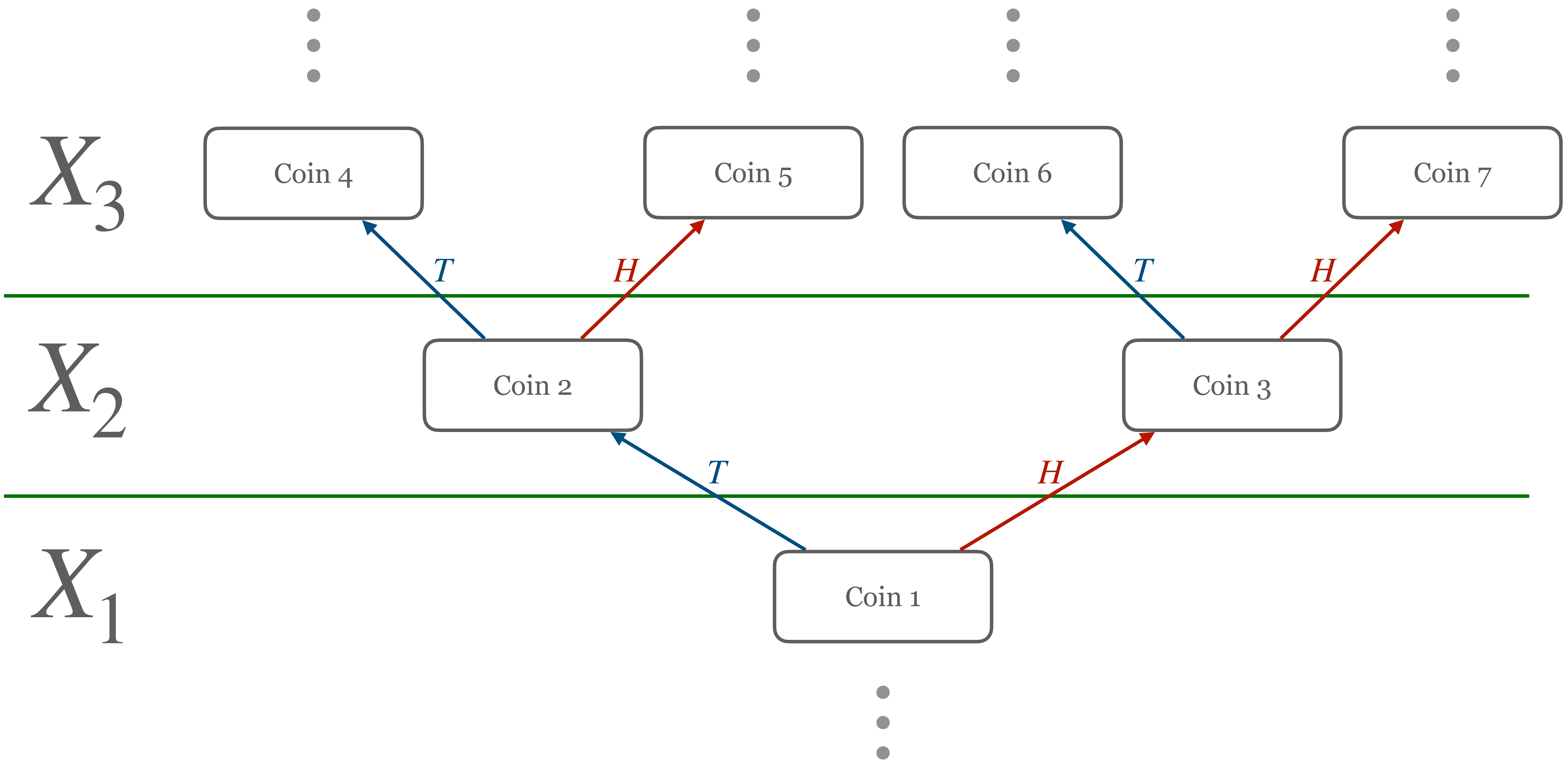
Example.

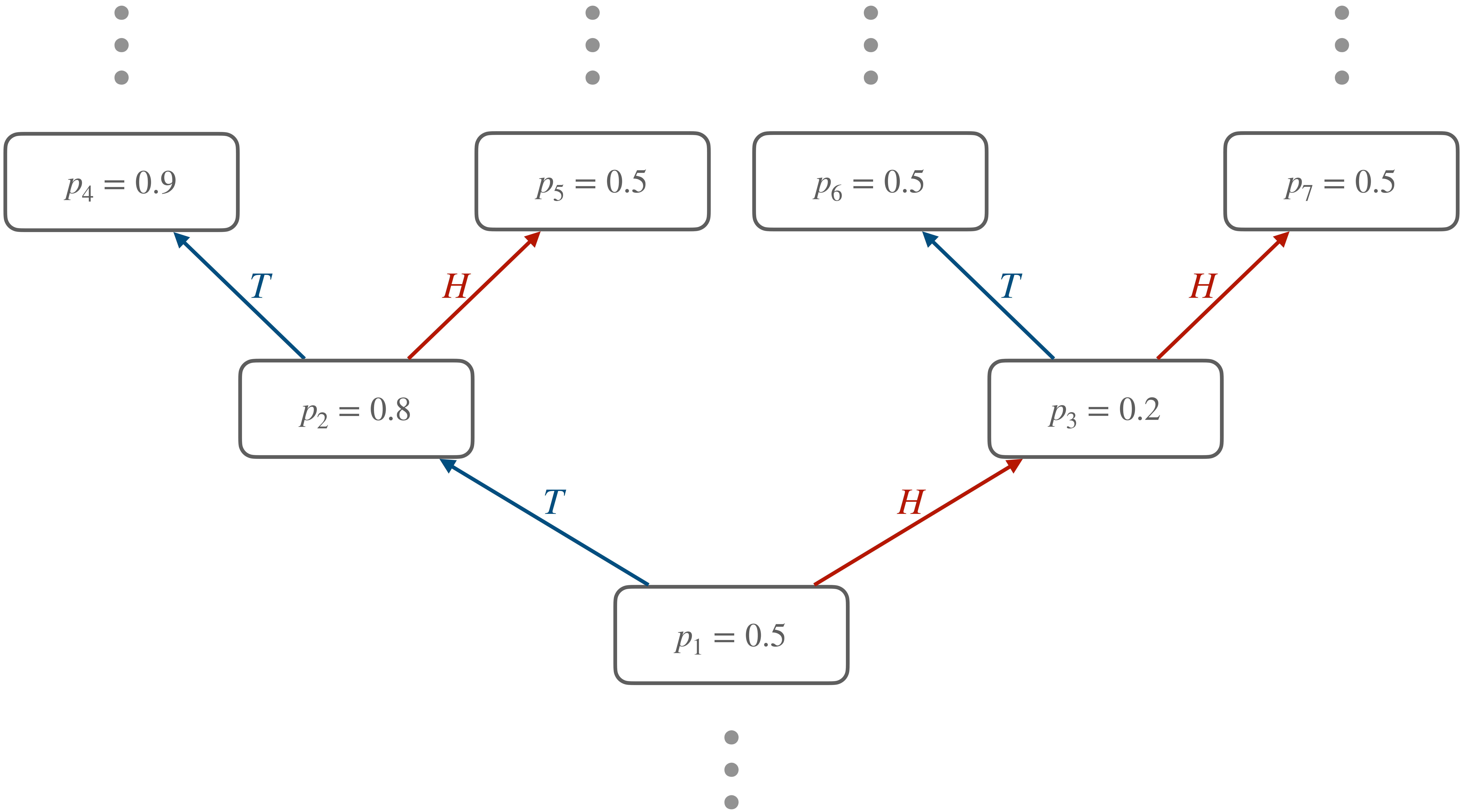
Too many coins.

X_3

X_2

X_1


X_3
X_2
X_1
Coin 4
Coin 5
Coin 6
Coin 7
Coin 2
Coin 3
Coin 1
T
H
T
H
T
H


$p_4 = 0.9$
$p_5 = 0.5$
$p_6 = 0.5$
$p_7 = 0.5$
T
H
T
H
$p_2 = 0.8$
$p_3 = 0.2$
T
H
$p_1 = 0.5$

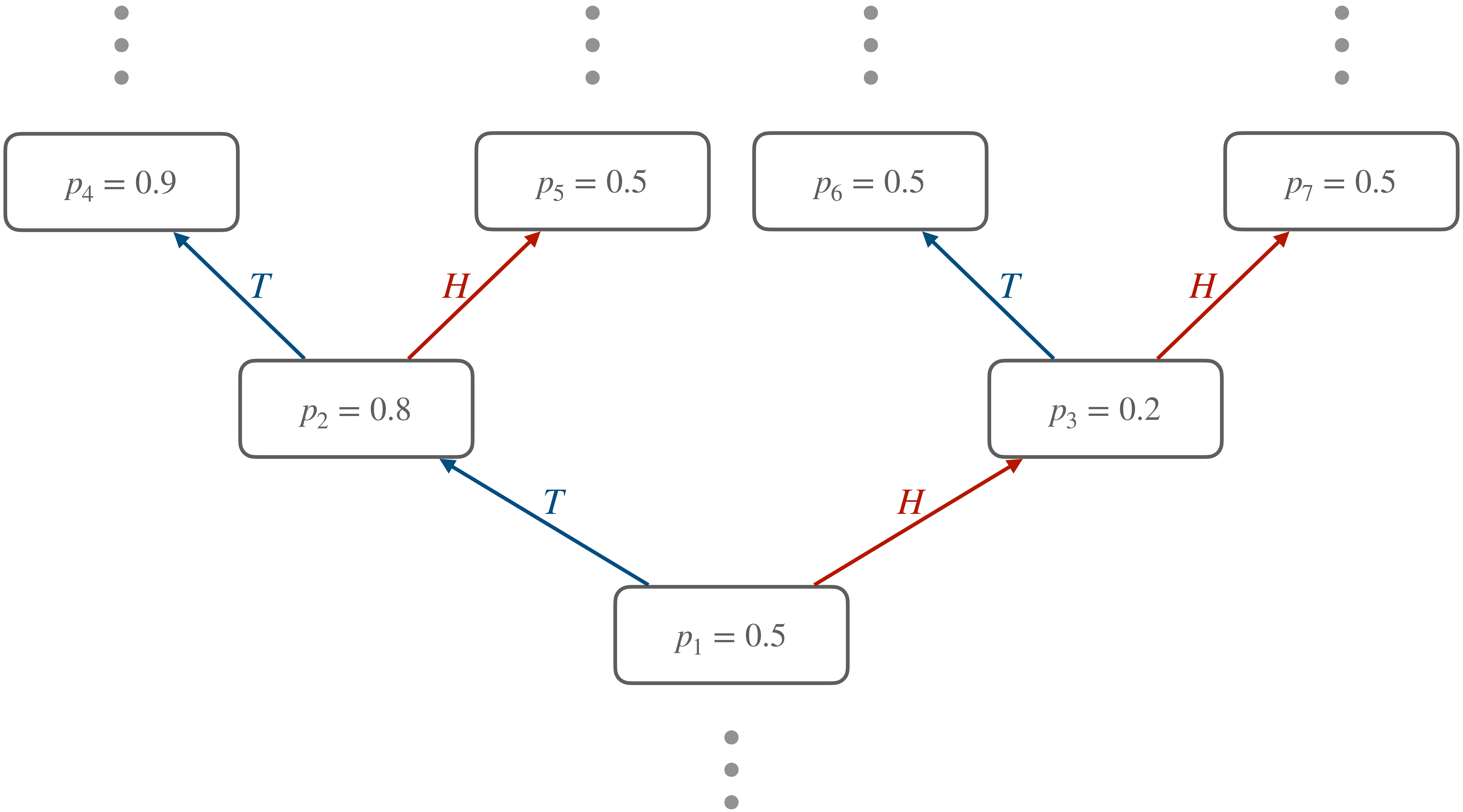
How “fair” is this process?

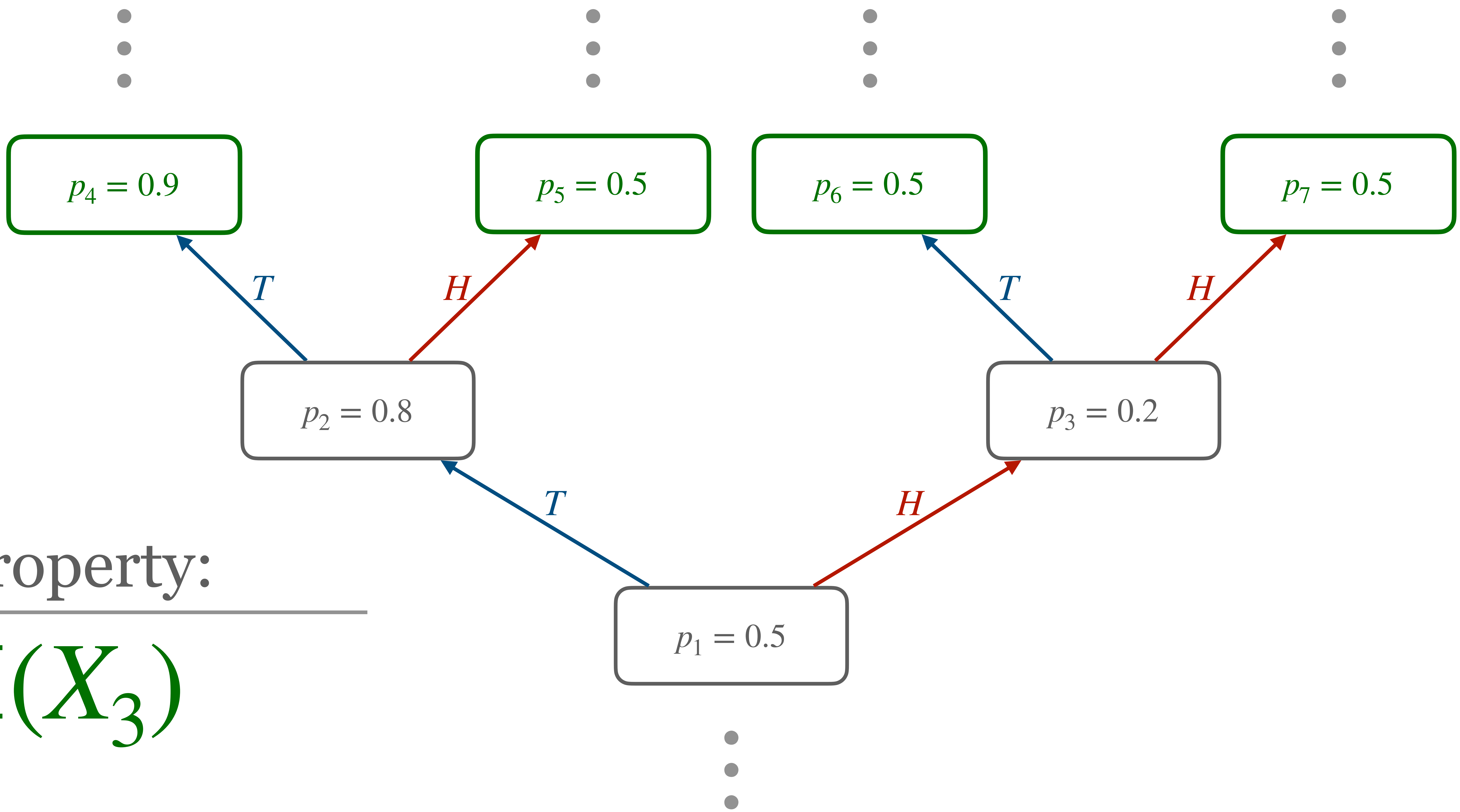
Multiple interpretations.

$$\mathbb{P}(\mathbf{H}) - \mathbb{P}(\mathbf{T})$$

Static Fairness

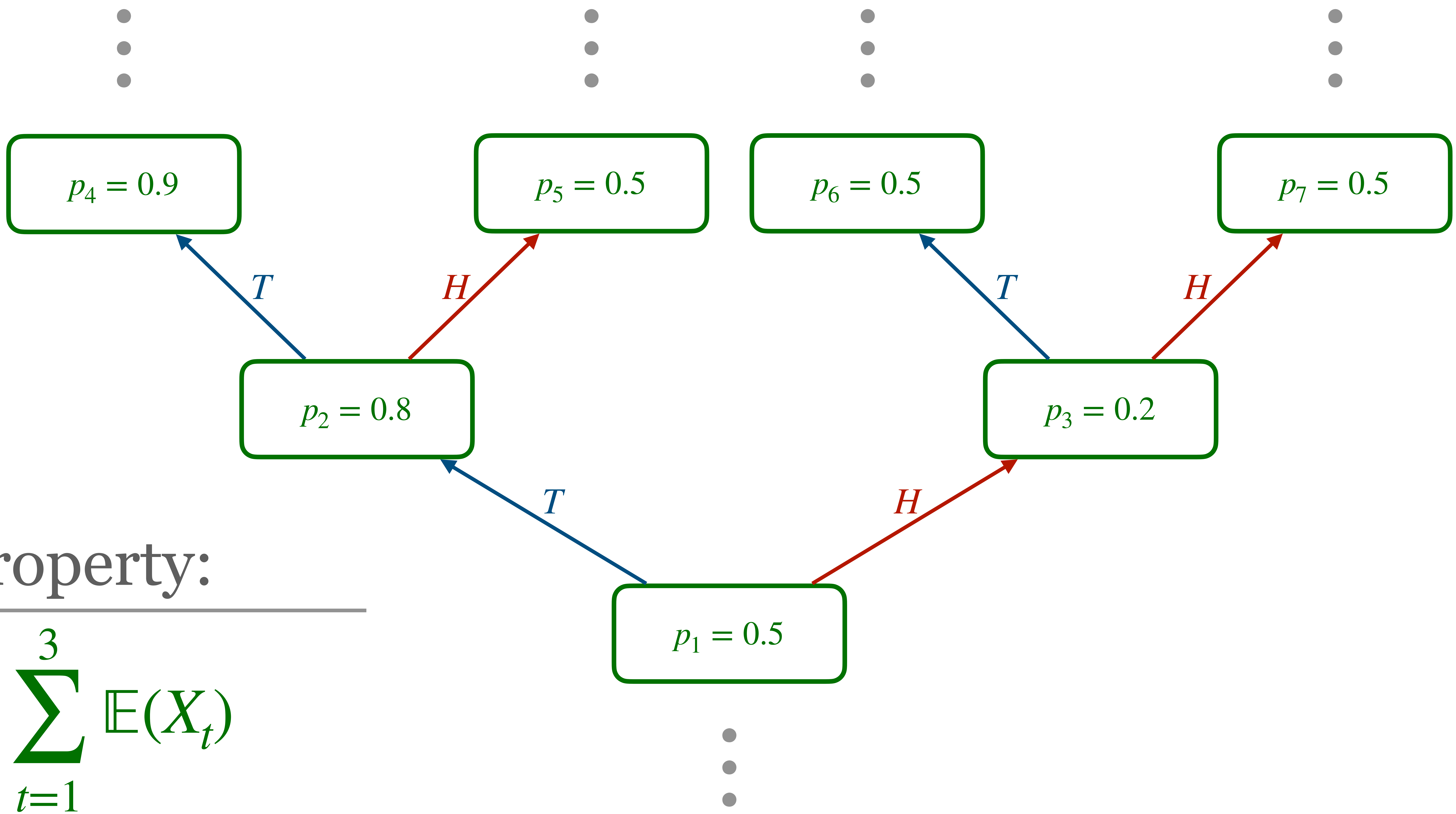
The offline perspective.


$p_4 = 0.9$
$p_5 = 0.5$
$p_6 = 0.5$
$p_7 = 0.5$
T
H
T
H
$p_2 = 0.8$
$p_3 = 0.2$
T
H
$p_1 = 0.5$



Property:

$$\mathbb{E}(X_3)$$



Property:

$$\frac{1}{3}\sum_{t=1}^{3}\mathbb{E}(X_t)$$

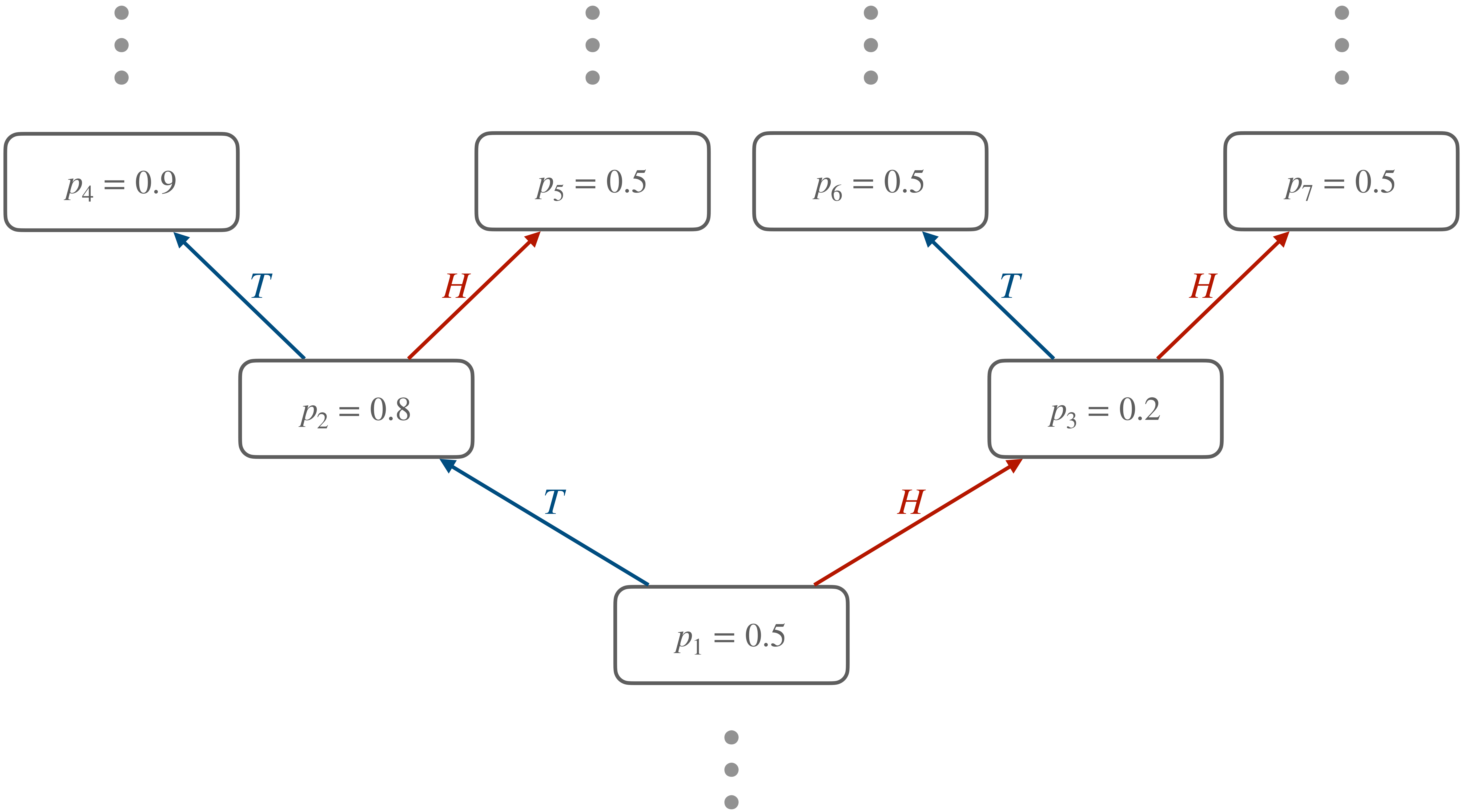
Dynamic Fairness

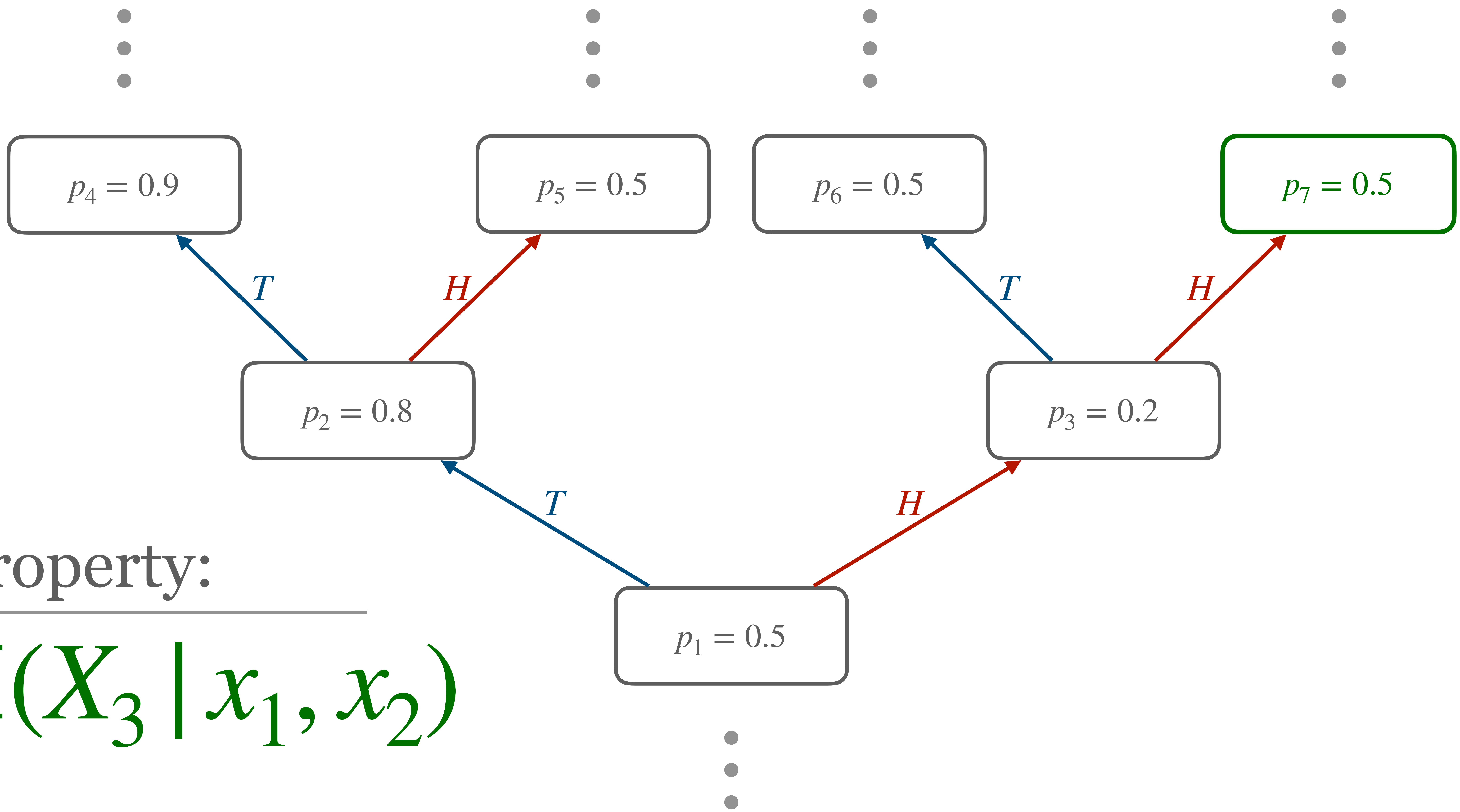
The runtime perspective.

$$x_3 = T$$

$$x_2 = H$$

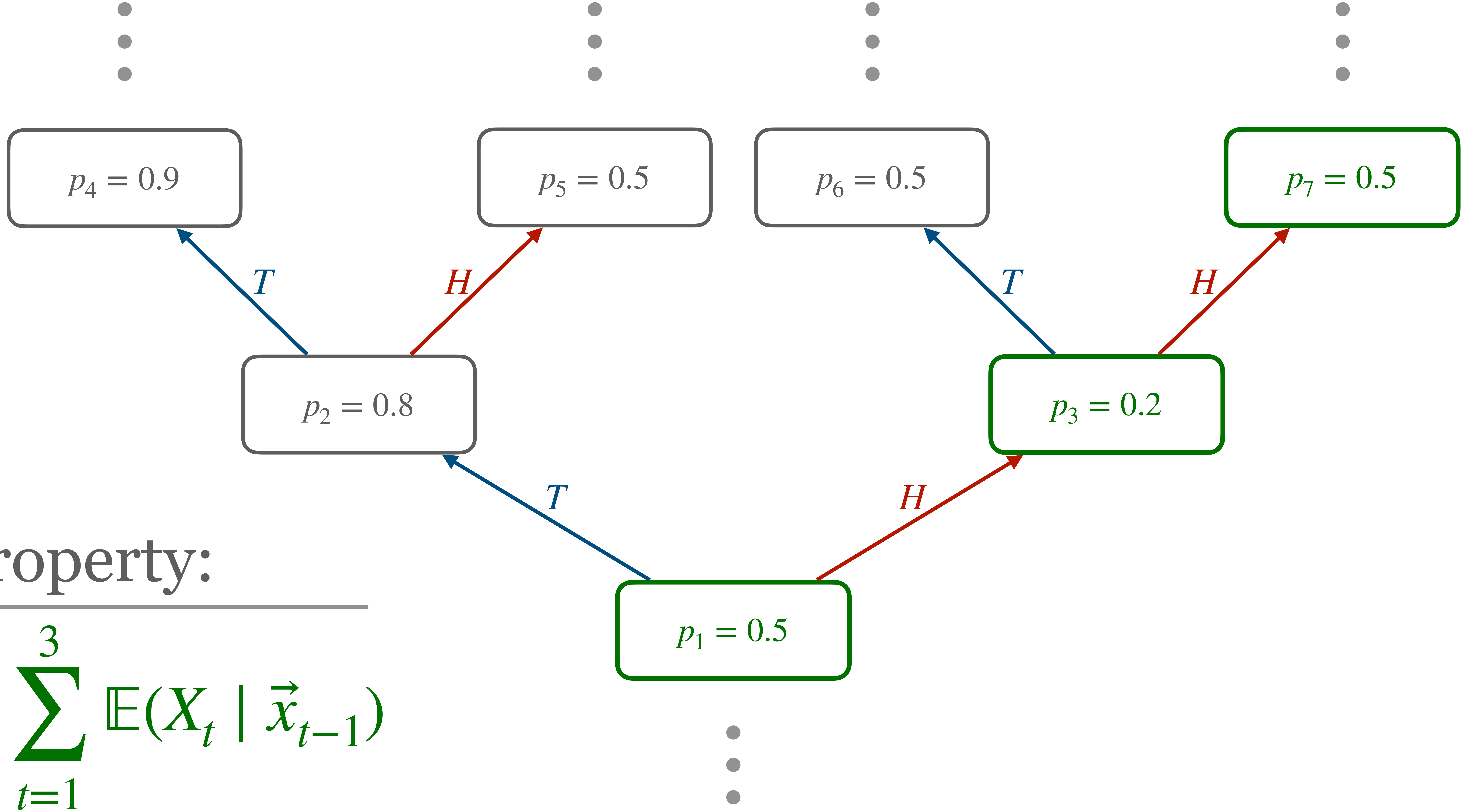
$$x_1 = H$$


$p_4 = 0.9$
$p_5 = 0.5$
$p_6 = 0.5$
$p_7 = 0.5$
T
H
T
H
$p_2 = 0.8$
$p_3 = 0.2$
T
H
$p_1 = 0.5$



Property:

$$\mathbb{E}(X_3 \mid x_1, x_2)$$



Property:

$$\frac{1}{3}\sum_{t=1}^{3} \mathbb{E}(X_t \mid \vec{x}_{t-1})$$

Inherently Runtime.

Specification is w.r.t. the observed trace.

Generalisation.

What are we getting at?

$$f : \Sigma^n \to \mathbb{R}$$

$$X_1, X_2, X_3, X_4, X_5, X_6, \ldots$$

$$X_1, X_2, X_3, \underline{X_4, X_5, X_6}, \dots$$

$$\mathbb{E}(f(X_4, X_5, X_6))$$

$$\underbrace{X_1, X_2, X_3}, \underbrace{X_4, X_5, X_6}, \dots$$

$$\mathbb{E}(f(X_4, X_5, X_6) \mid x_1, x_2, x_3)$$

$$\underline{X_1, X_2, X_3}, \underline{X_4, X_5, X_6}, \ldots$$

$$\mathbb{E}(f(X_4, X_5, X_6) \mid x_2, x_3)$$

$$\mathbb{E}(f(\overrightarrow{X}_{t:t+n}) \mid B(\vec{x}_t))$$

Neighbourhood around $\vec{x}_t \in \Sigma^t$

Difficult to compute…

…with the model.
But what if the only thing you have is…

…a Black Box?

…a Black Box?

and only a finite realisation of the stochastic process.

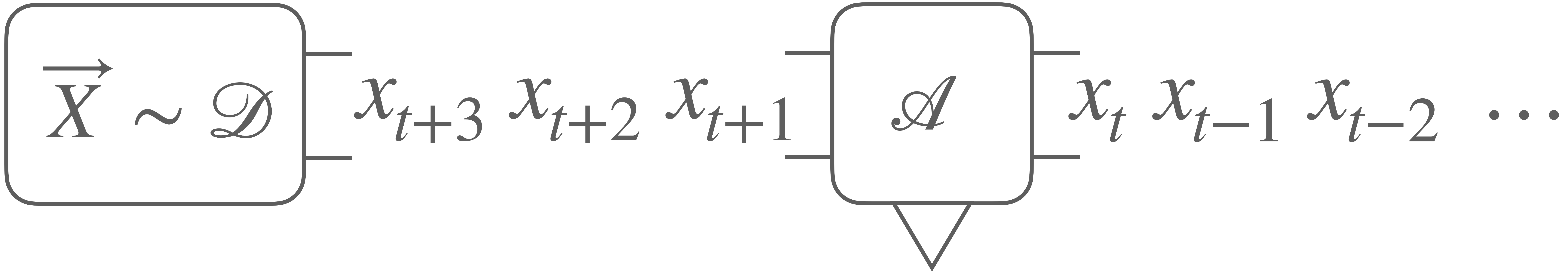
…a Black Box?

and only a finite realisation of the stochastic process.
(…and some assumptions)

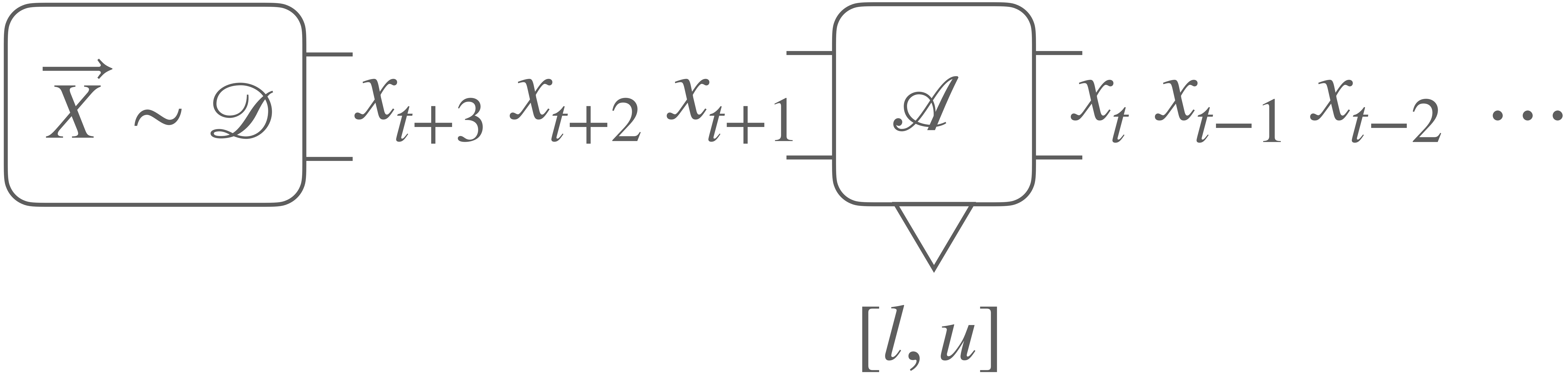
Well…you estimate.

At least you try to.

$$\boxed{\overrightarrow{X} \sim \mathscr{D}} \; x_{t+3}\; x_{t+2}\; x_{t+1}\; x_t\; x_{t-1}\; x_{t-2}\; \cdots$$


$\overrightarrow{X} \sim \mathscr{D}$
x_{t+3} x_{t+2} x_{t+1}
$\mathscr{A}$
x_t x_{t-1} x_{t-2} $\cdots$

$$\mathbb{E}(f(\overrightarrow{X}_{t:t+n}) \mid B(\vec{x}_t)) \in \mathscr{A}(\overrightarrow{x_t}) \text{ with probability } 1 - \delta$$



Projects:

Monitoring Algorithmic Fairness (CAV23)
Runtime Monitoring of Dynamic Fairness Properties (FAccT23)
Monitoring Algorithmic Fairness under Partial Observations (RV23)

Projects:

Monitoring Algorithmic Fairness (CAV23)

Runtime Monitoring of Dynamic Fairness Properties (FAccT23)

Monitoring Algorithmic Fairness under Partial Observations (RV23)

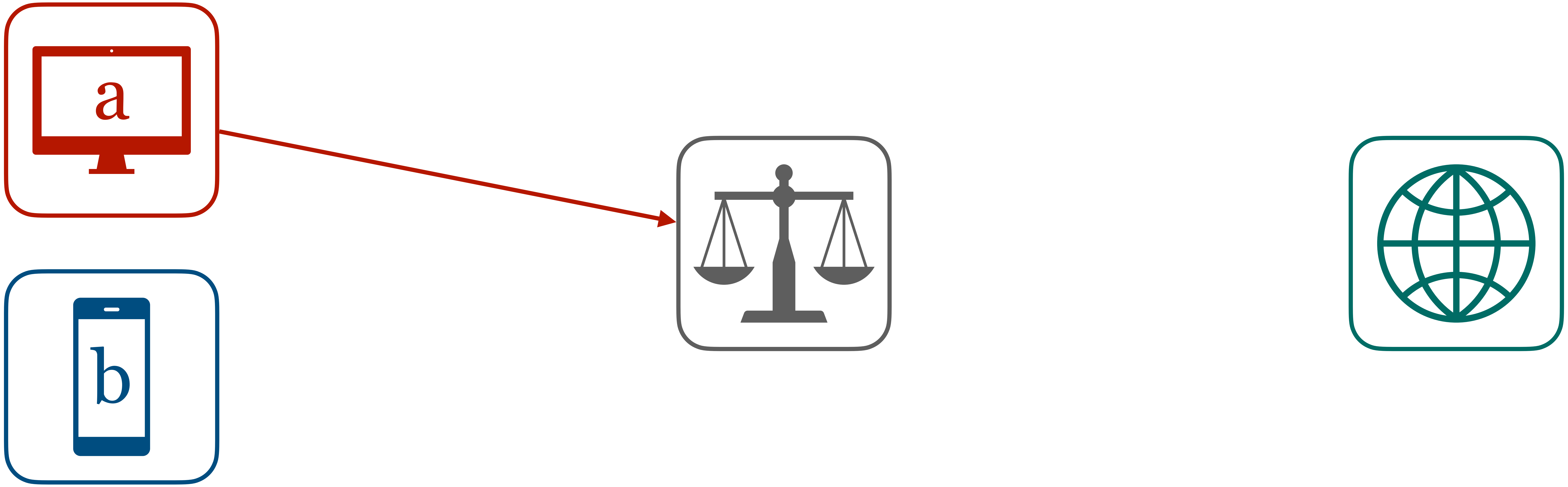
Monitoring Algorithmic Fairness

under Partial Observations (RV23)

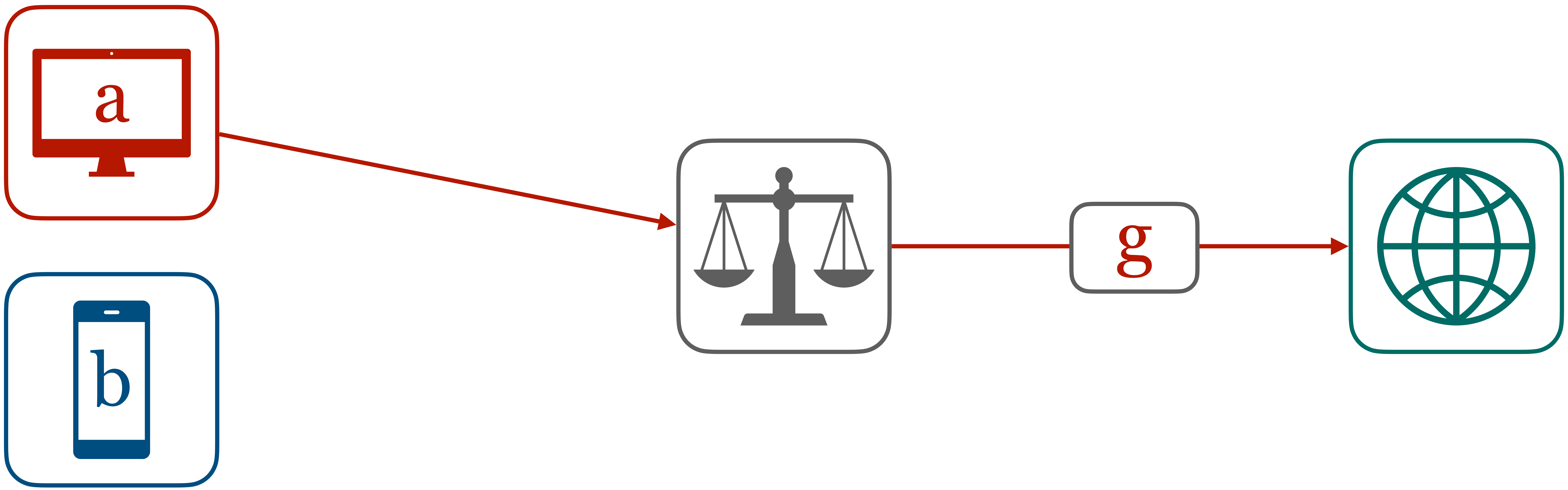
Example.

A simple resource allocation problem.

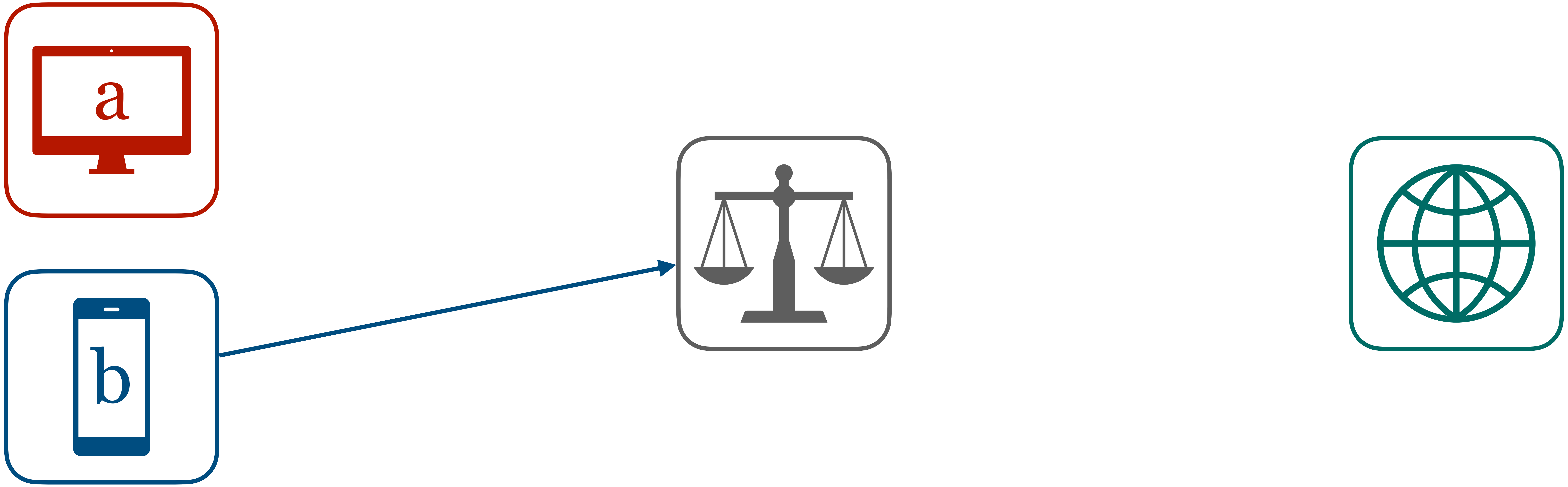

a
b



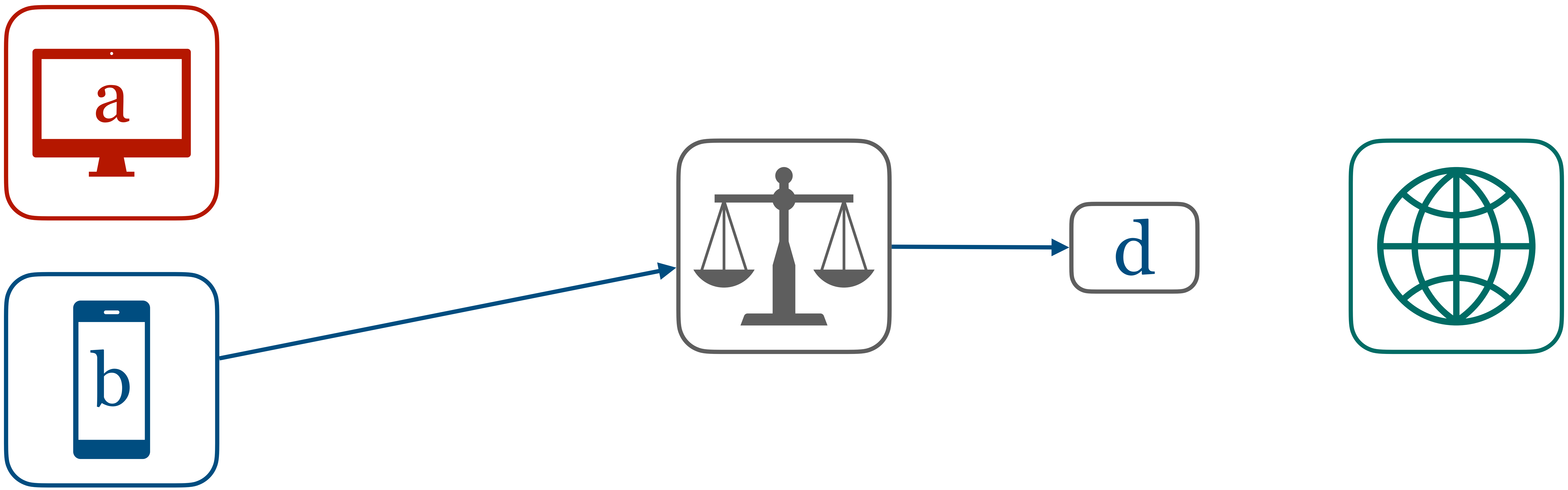
a



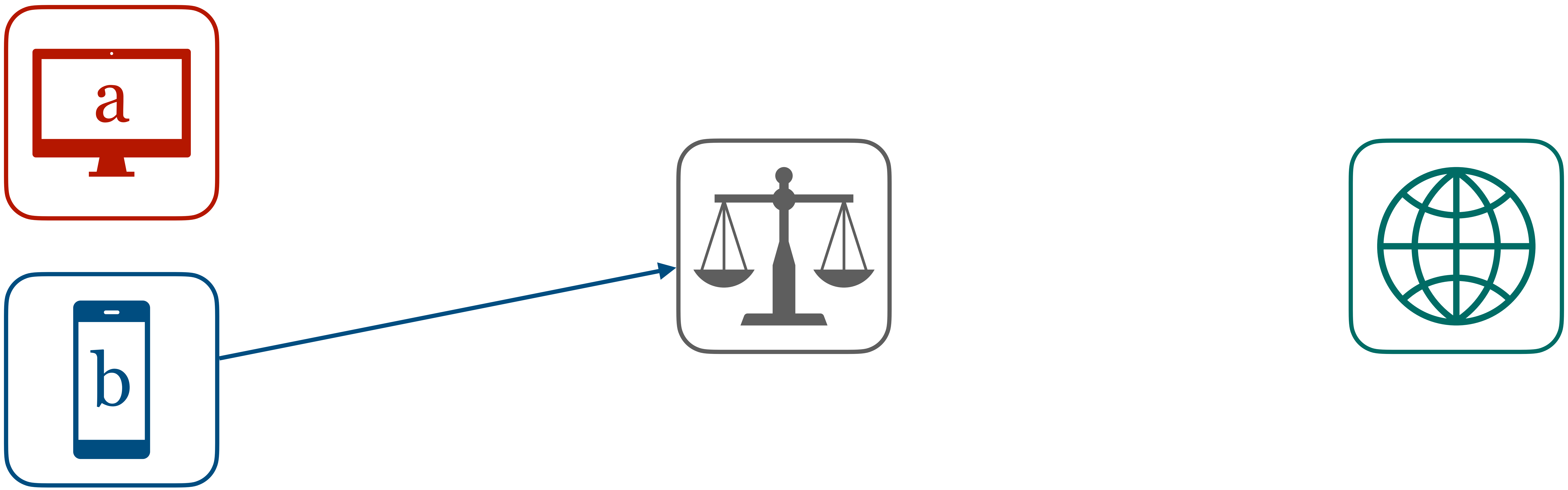
a g



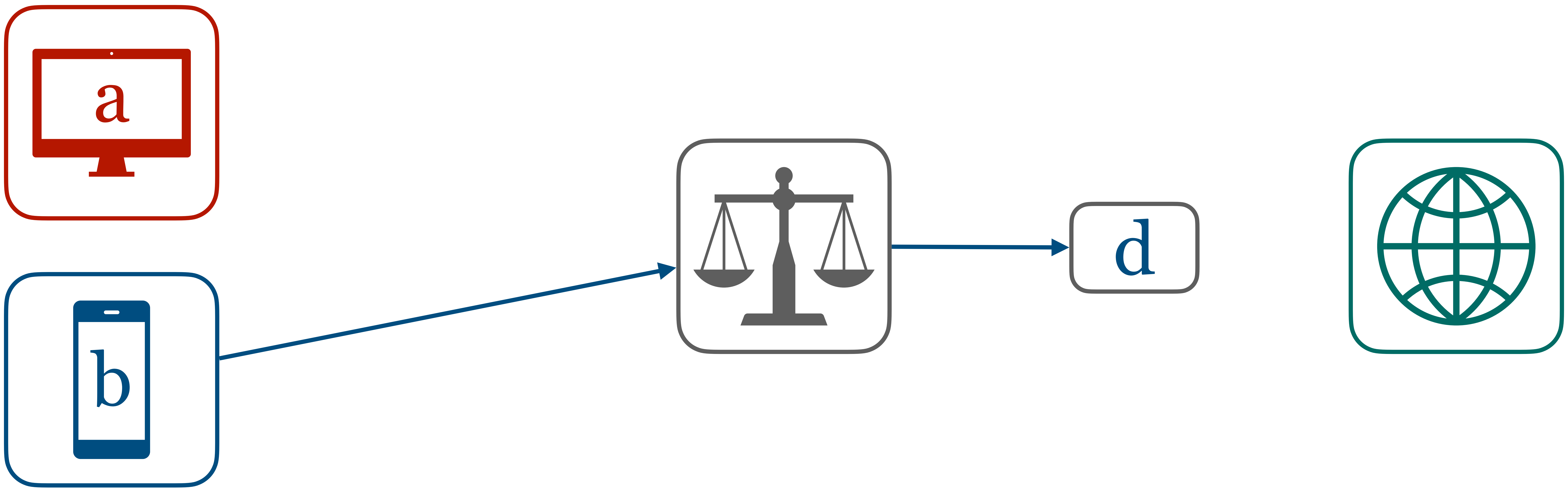
a g b



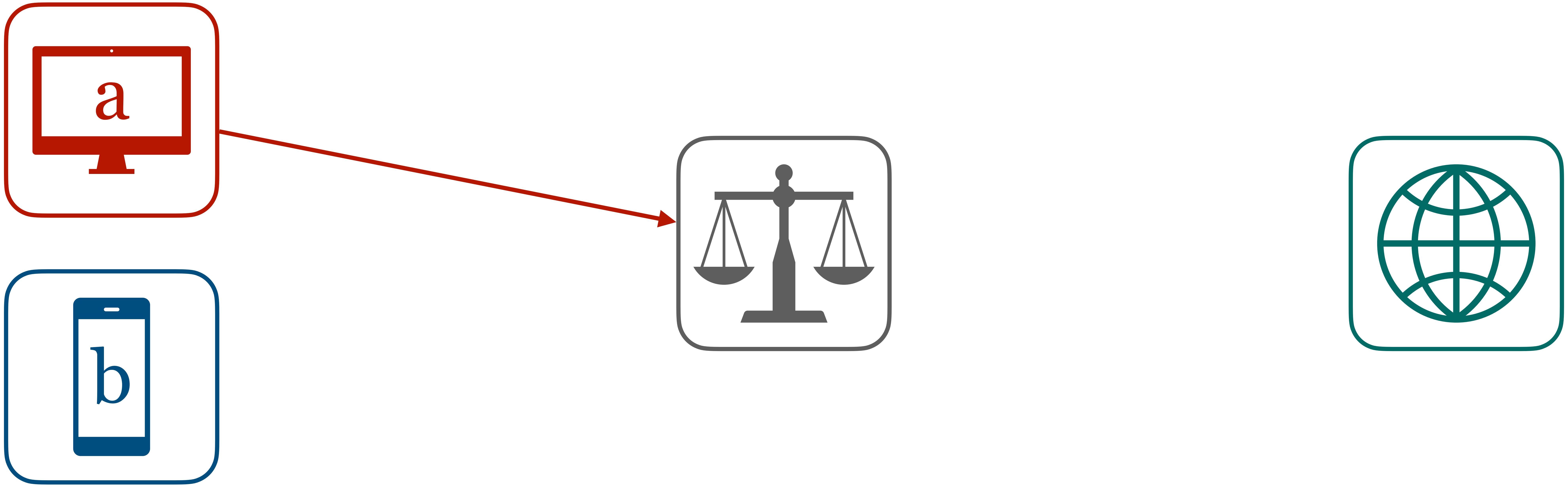
a g b d



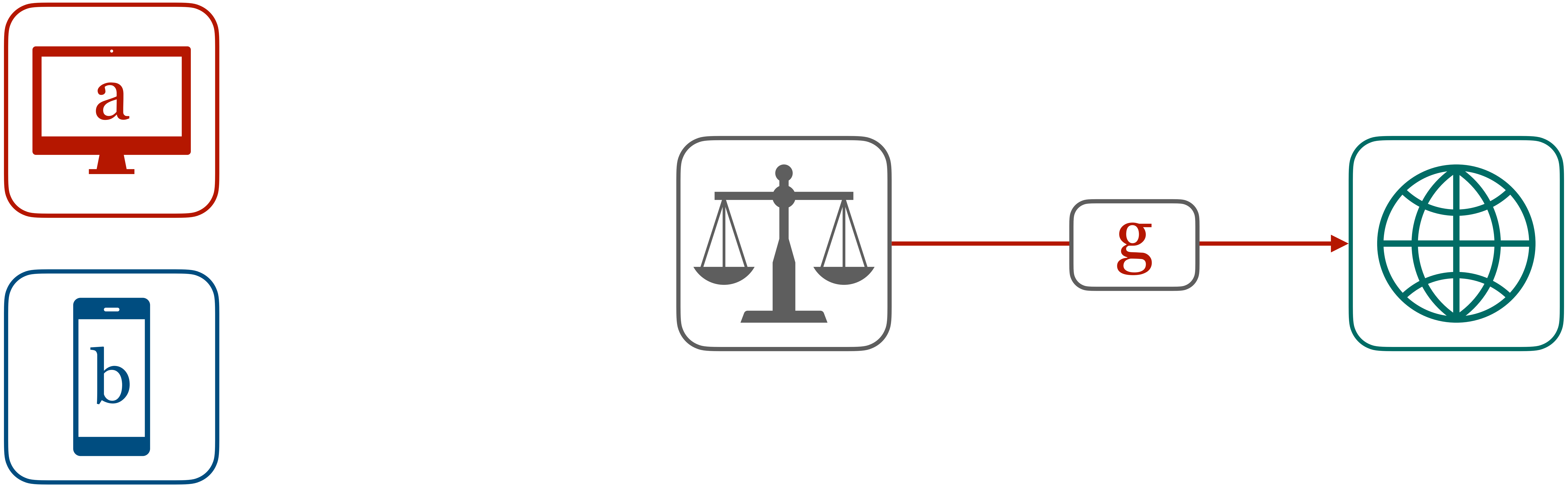
a g b d b



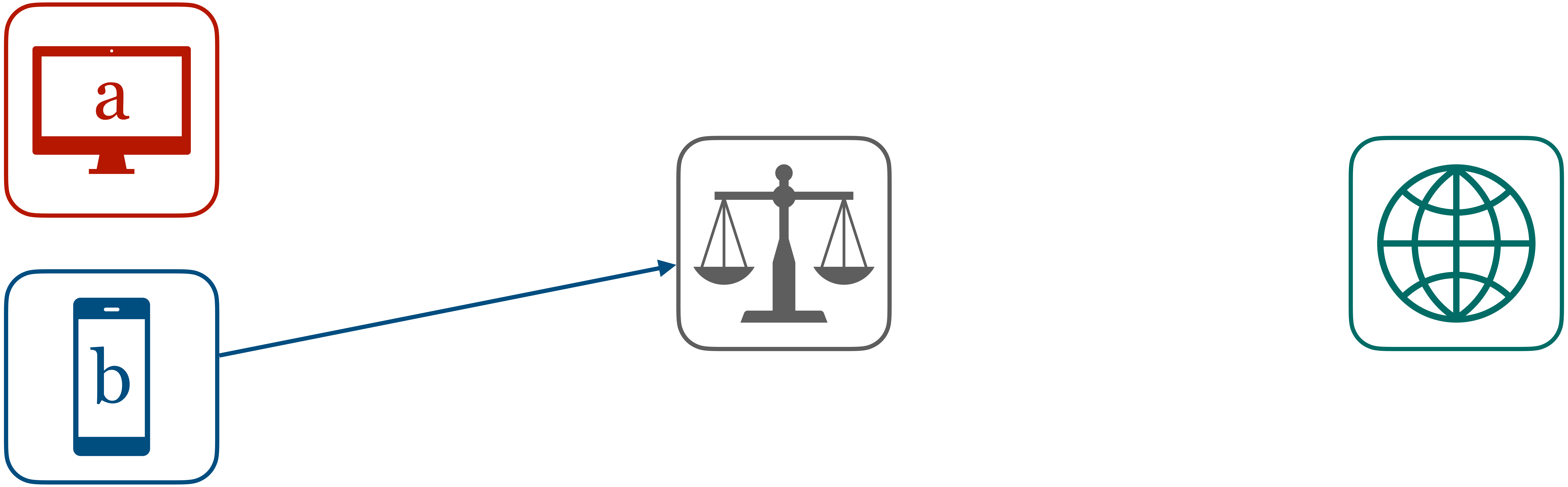
a g b d b d



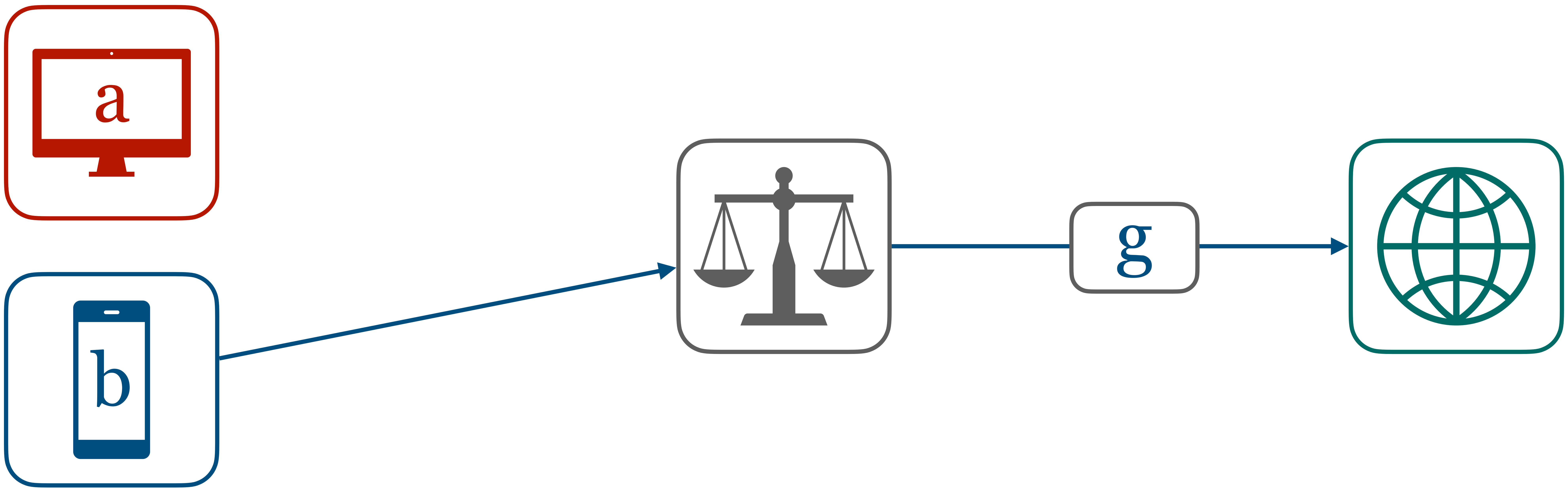
a g b d b d a



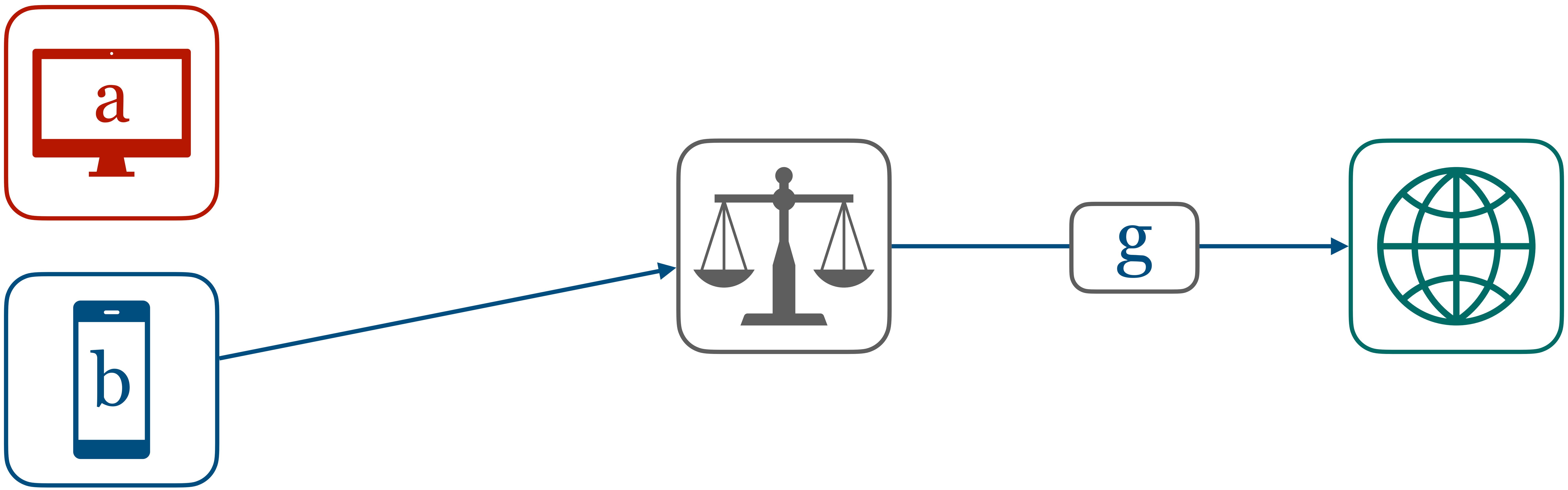
a g b d b d a g



a g b d b d a g b



a g b d b d a g b g



a g b d b d a g b g . . . . . .

a g b d b d a g b g . . . . . .

$$\mathbb{P}(\mathbf{g}) - \mathbb{P}(\mathbf{g})$$

$$\mathbb{P}(\mathbf{g} \mid a) - \mathbb{P}(\mathbf{g} \mid b)$$

Problem Statement.

What are we trying to do?

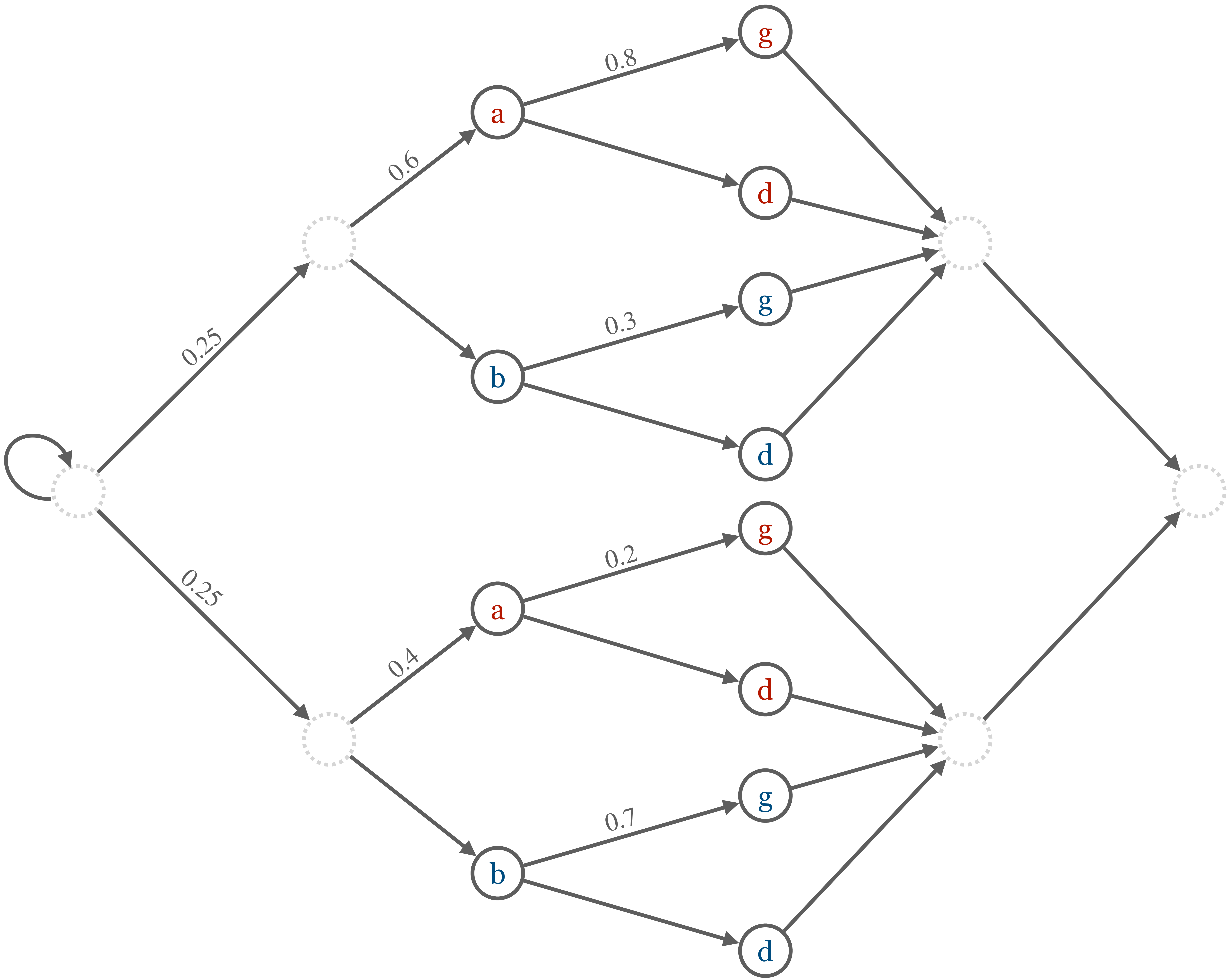
Properties.

Arithmetic expressions over
$\mathbb{E}(f(\overrightarrow{X}_{t:t+n}))$ *for* $f : \Sigma^n \to \mathbb{R}$
and any $t > 0$.

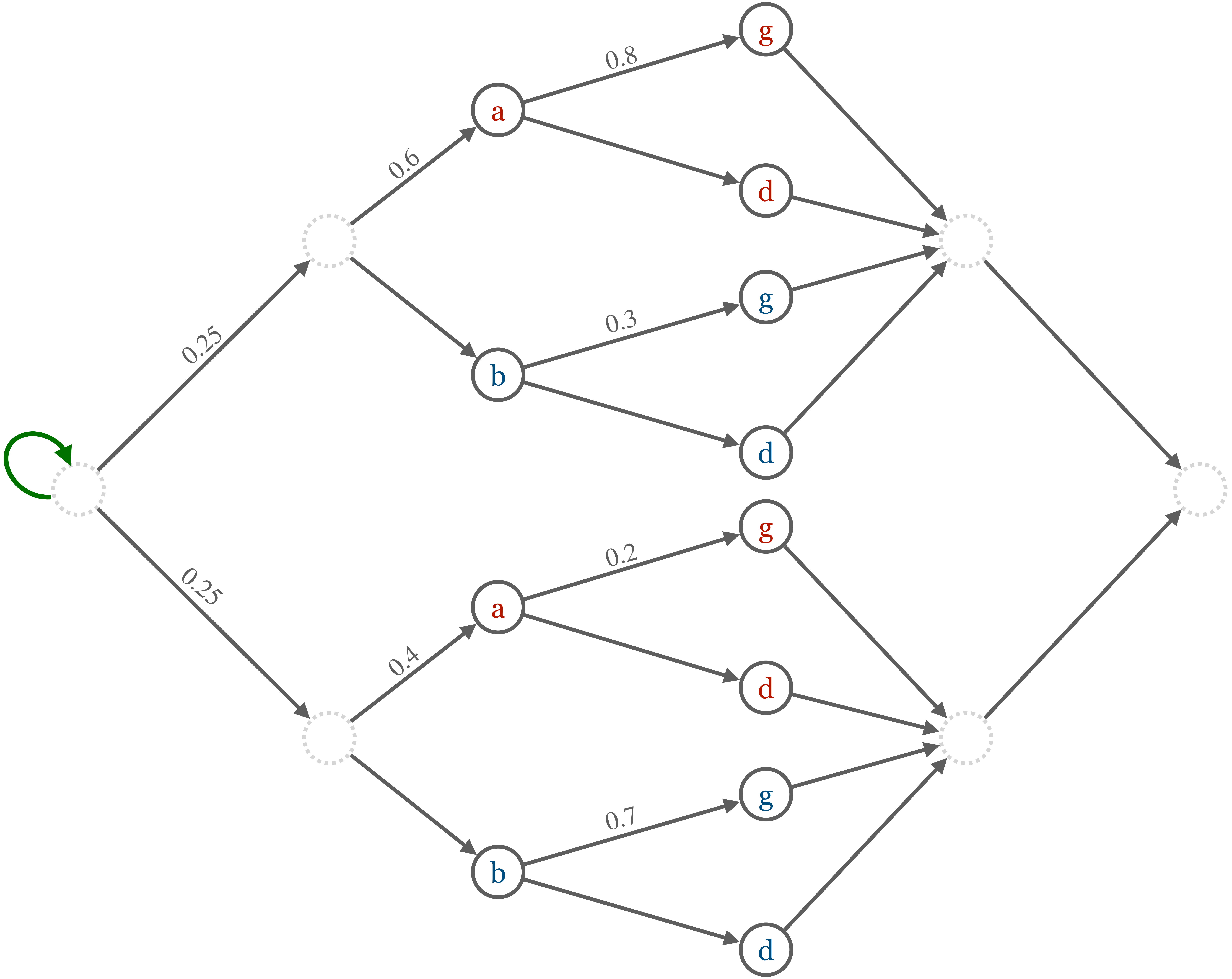
$$\mathbb{P}\left(\mathbb{E}(f(\overrightarrow{X}_{t:t+n})) \in \mathscr{A}(\overrightarrow{X}_t)\right) \geq 1 - \delta$$

Assumptions.

The system is a
stationary, aperiodic, labelled Markov chain
with known mixing time τ_{mix}.



0.25
0.25
0.6
0.4
0.8
0.3
0.2
0.7
a
b
a
b
g
d
g
d
g
d
g
d


0.25
0.25
0.6
0.4
a
b
a
b
0.8
0.3
0.2
0.7
g
d
g
d
g
d
g
d

Stationarity.

…the distribution over states does not change.

$$\pi = \pi \cdot P$$

Mixing Time.

…first time the total variation distance from stationarity distribution drops below ε.

$$\tau_{mix}(\varepsilon) = \min_{t} \left\{ \sup_{\mu} \|\mu \cdot P^t - \pi\|_{TV} \leq \varepsilon \right\}$$

Algorithm.

A sketch.

$$\mathbb{E}\left(f(\overrightarrow{X}_{t:t+n})\right) = \mathbb{E}\left(f(\overrightarrow{X}_{t+k:t+k+n})\right)$$

From stationarity

$$\hat{f}(\vec{x}_t) := \frac{1}{t-n+1} \sum_{i=1}^{t-n+1} f(\vec{x}_{i:i+n+1})$$

Estimator

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$

$$f(x_1, x_2, x_3).$$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$

$$\mathbb{E}(\hat{f}(\overrightarrow{X}_t)) = \mathbb{E}(f(\overrightarrow{X}_{t:t+n}))$$

Unbiased

$\vec{x}_t$ and $\vec{x}'_t$ differ only in position i

$$|\hat{f}(\vec{x}_t) - \hat{f}(\vec{x}'_t)| \leq c_i(t)$$

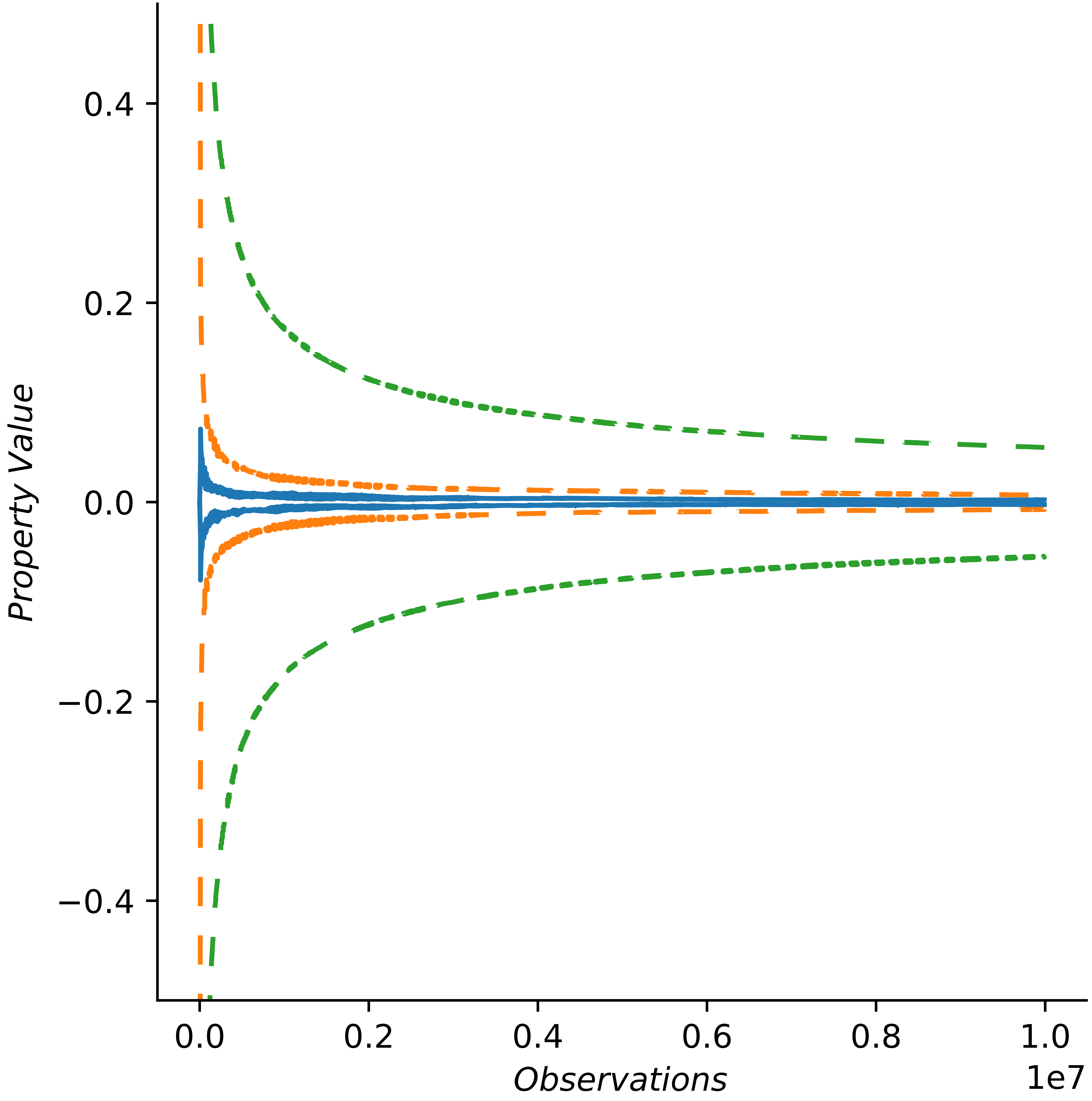
Lipschitz continuous

$$\mathbb{P}\left(| \mathbb{E}(f(\overrightarrow{X}_{t:t+n})) - \hat{f}(\overrightarrow{X}_{t}) | \geq \varepsilon \right) \leq \gamma(\varepsilon, \tau_{mix}, \{c_i(t)\}_i)$$

McDiarmid's inequality for MCs

Experiments.

3D-Hypercube (i.e. a cube).



Property Value
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
1e7
Observations

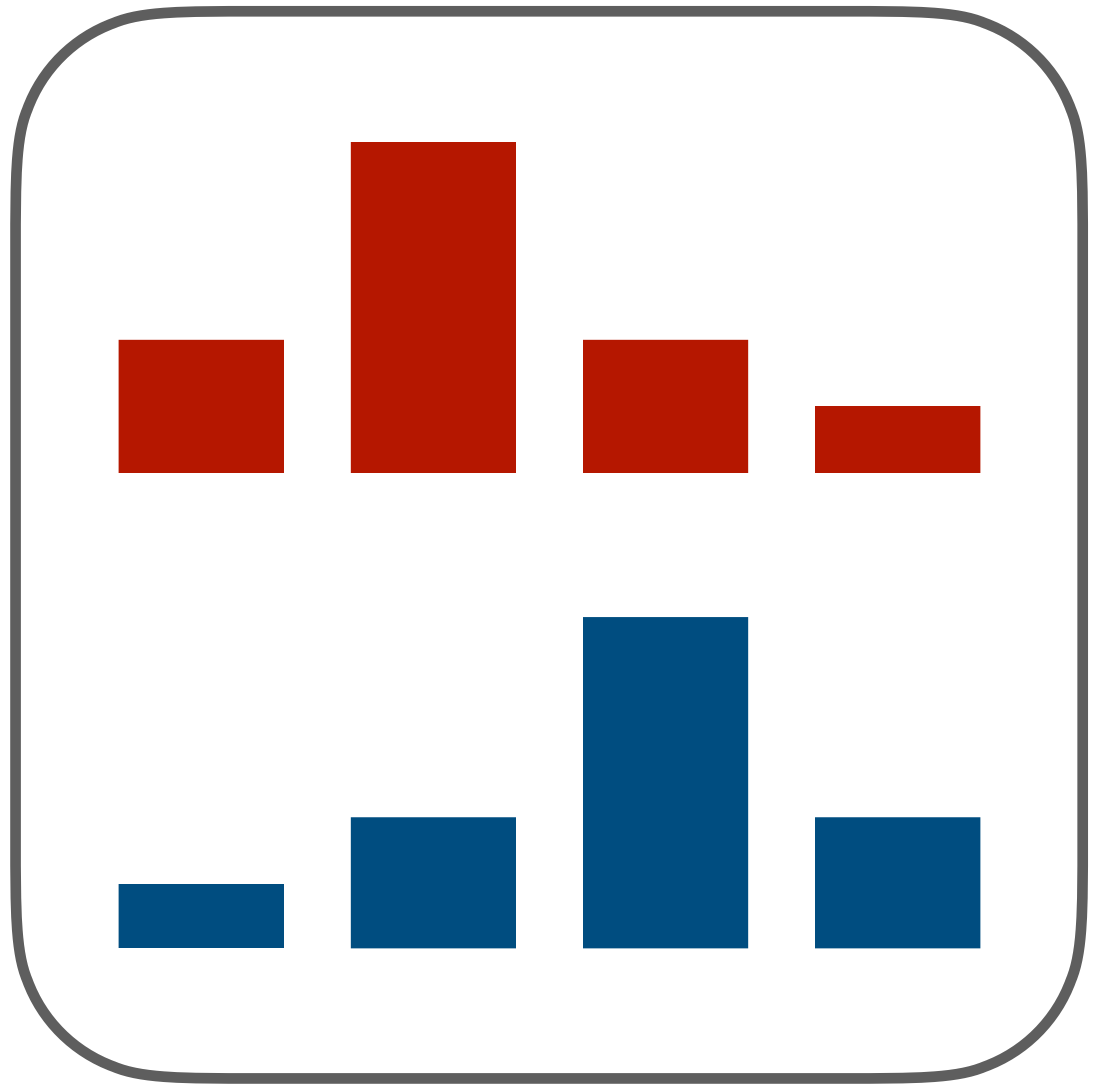
Runtime Monitoring

of Dynamic Fairness Properties (FAccT23)

Example.

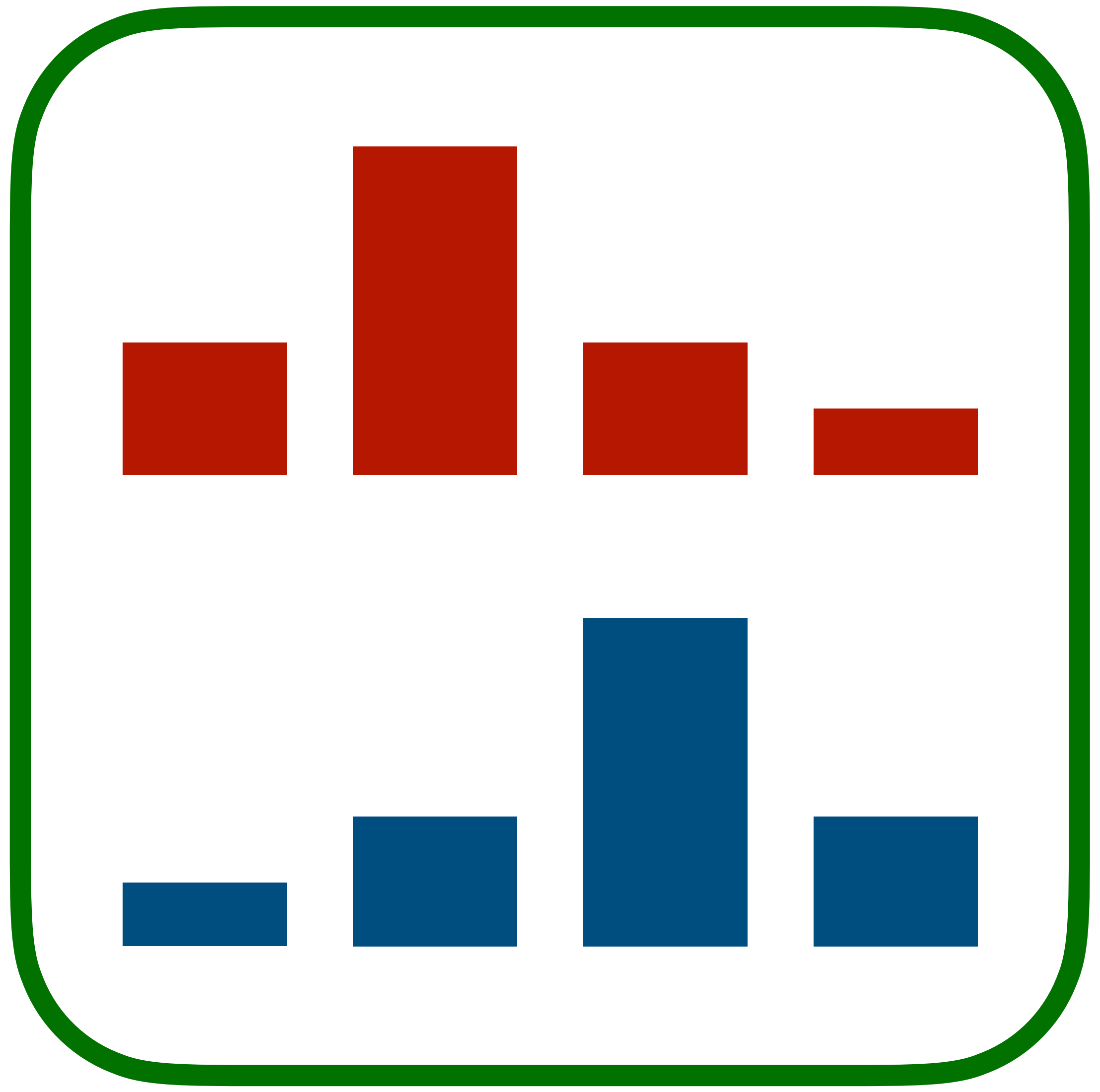
Dynamic Lending Problem (D'Amour 2020).

D'Amour et al. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies



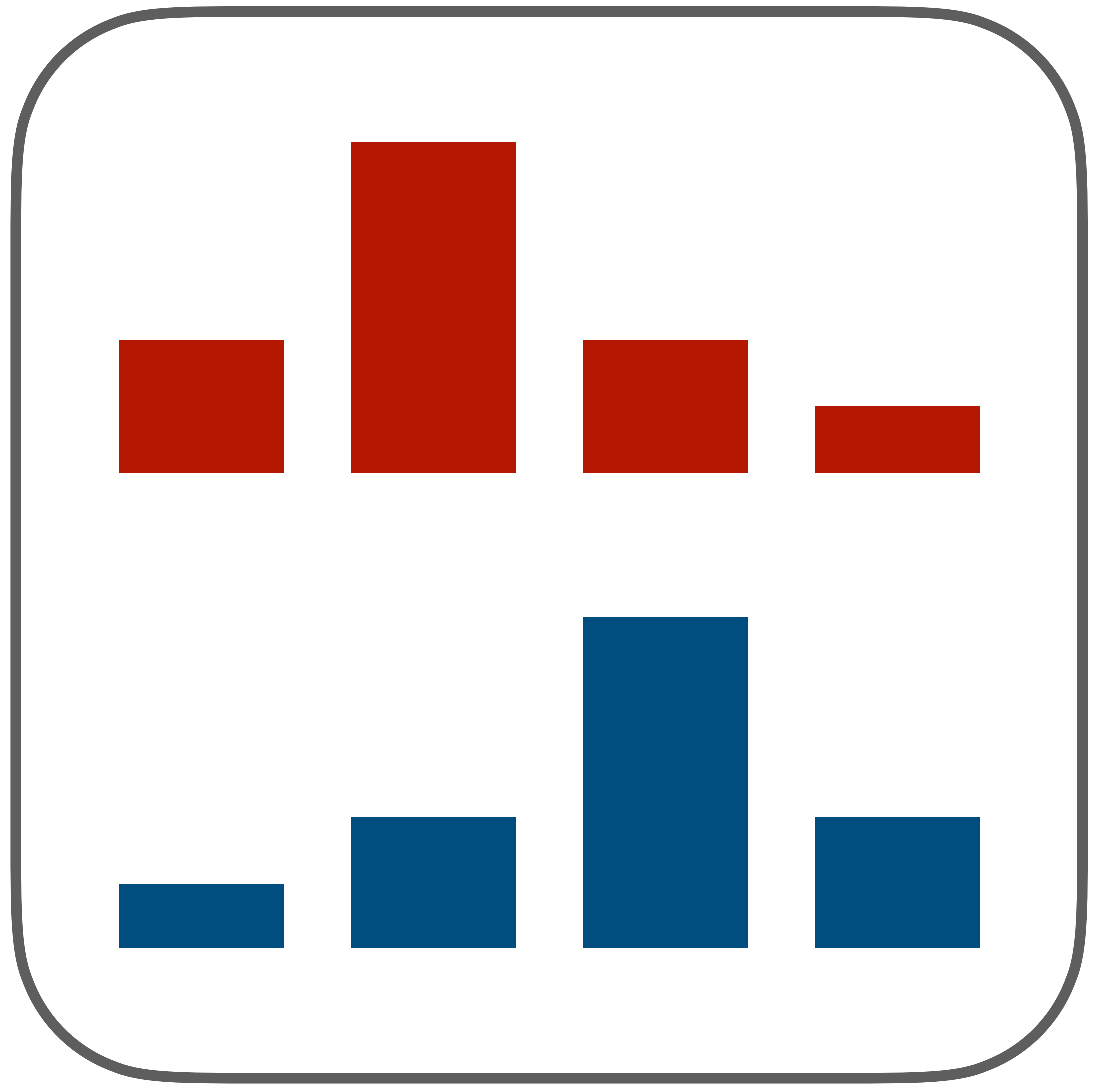
Bank:
***g**rant*
or
***d**eny*

Customer:
***r**epay*
or
***d**efault*



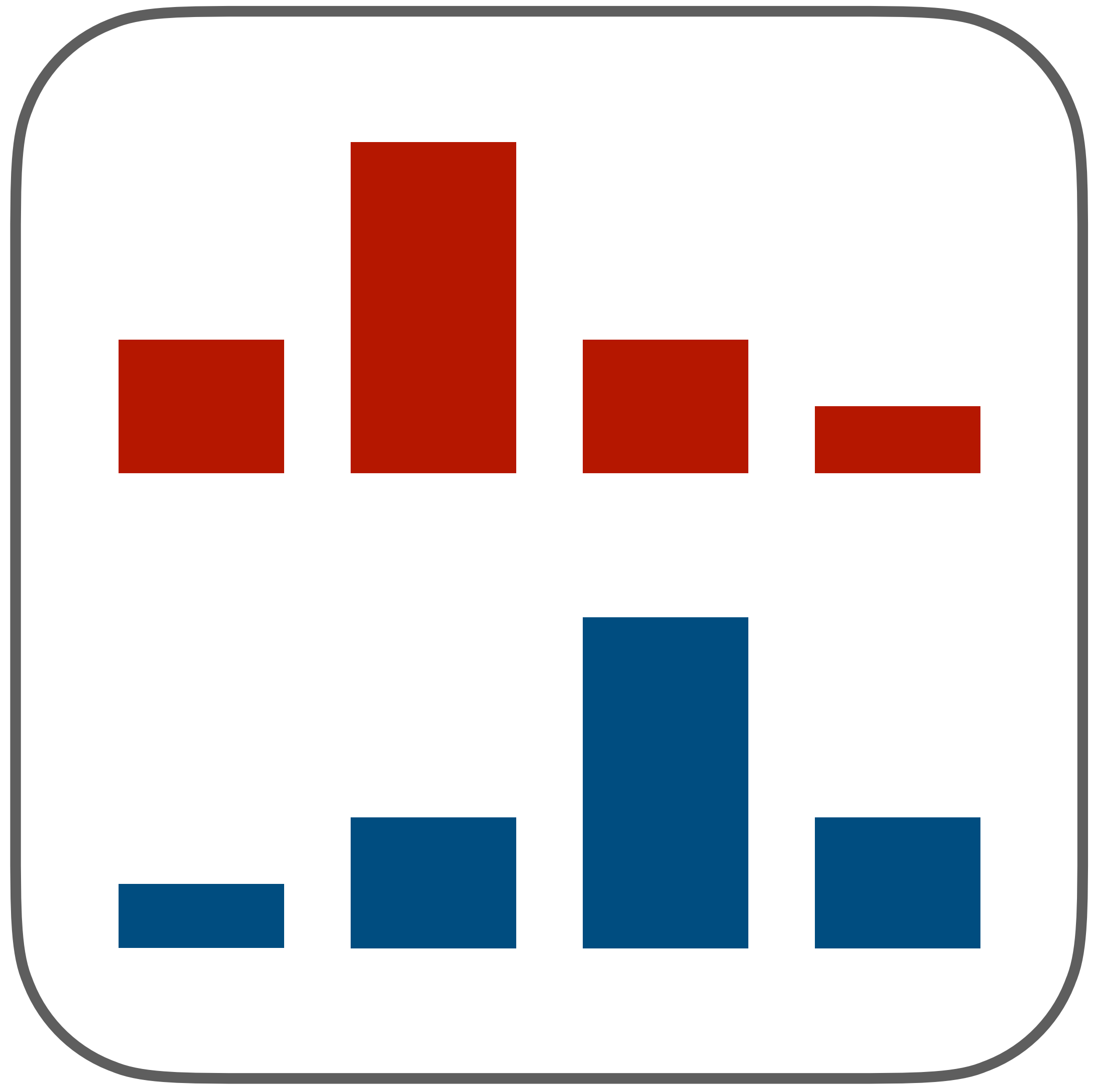
Bank:
grant
or
deny

Customer:
repay
or
default



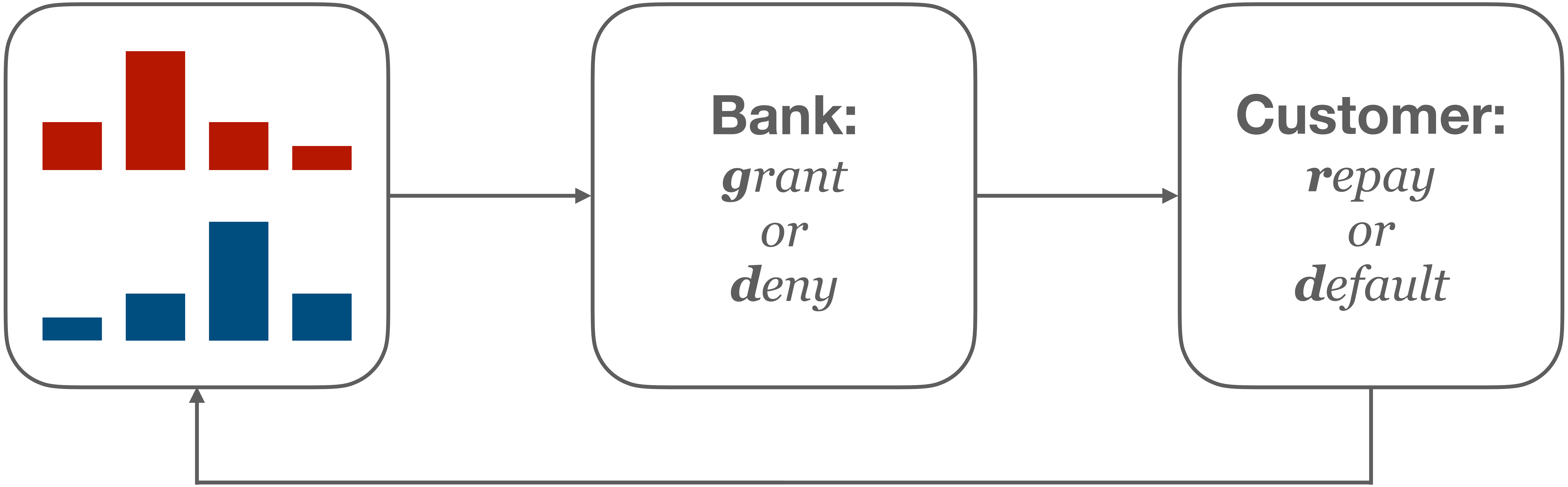
Bank:
grant
or
deny

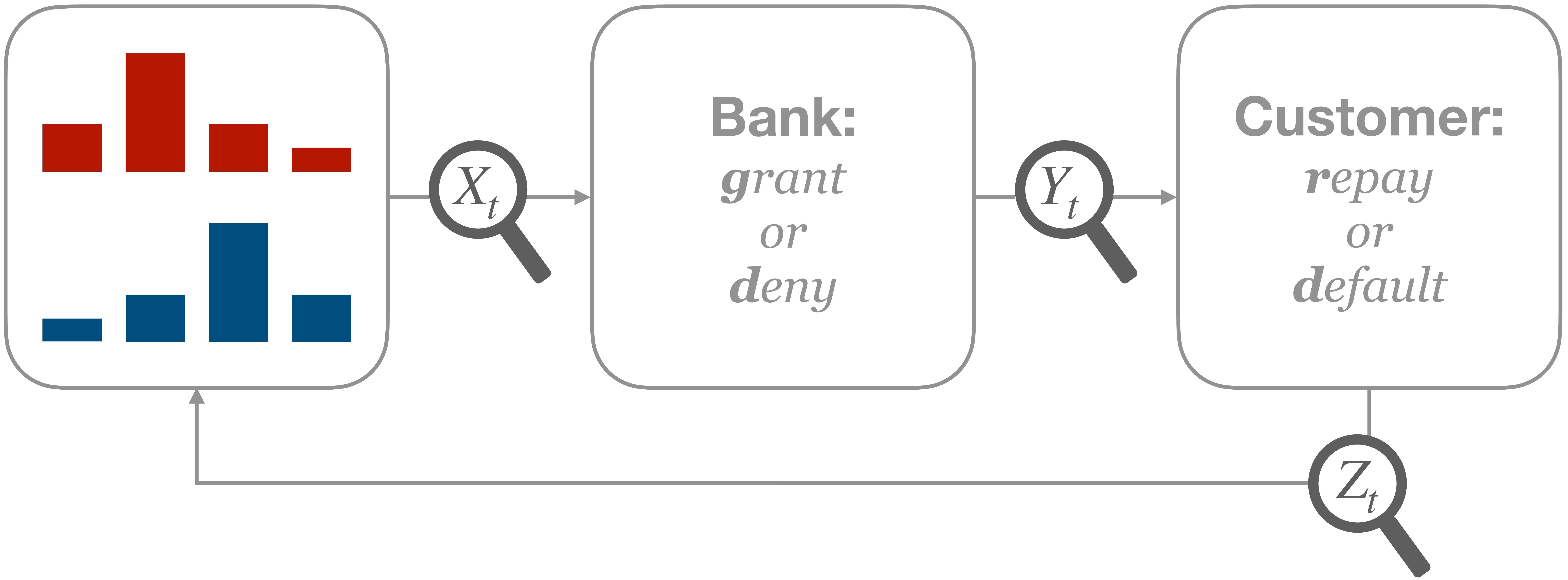
Customer:
repay
or
default



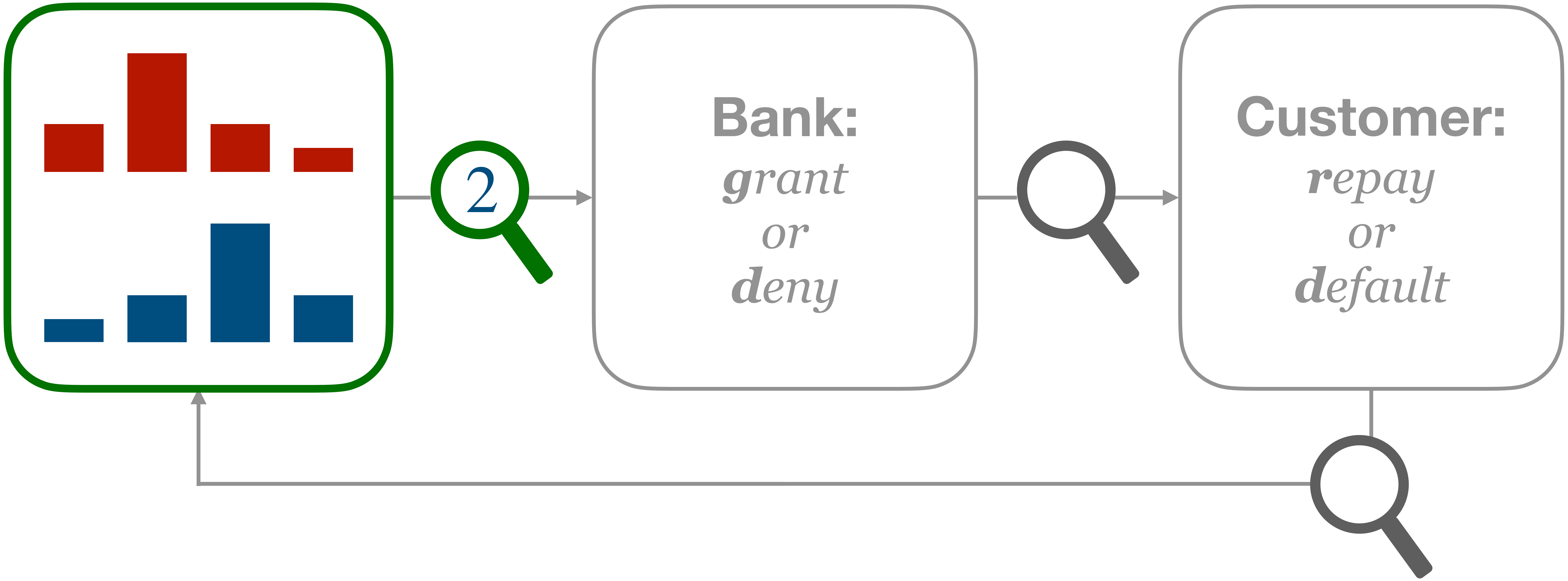
Bank:
grant
or
deny

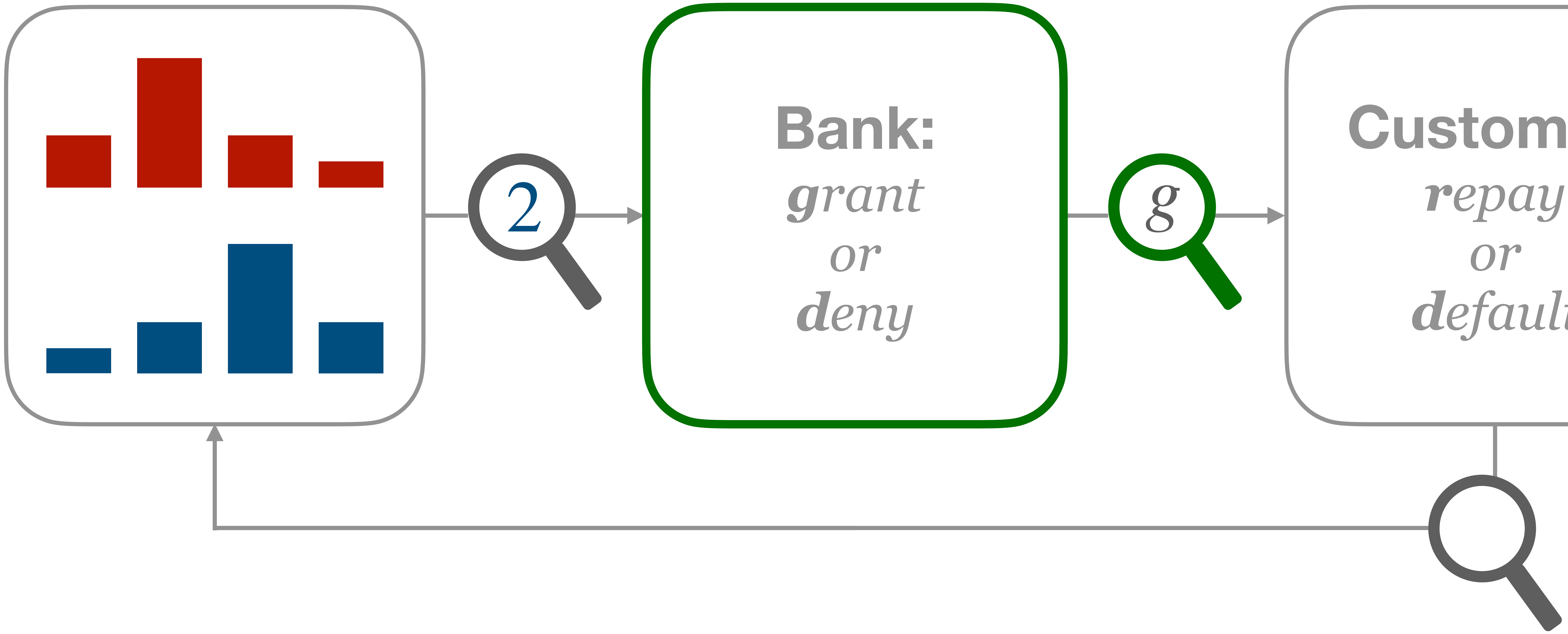
Customer:
repay
or
default

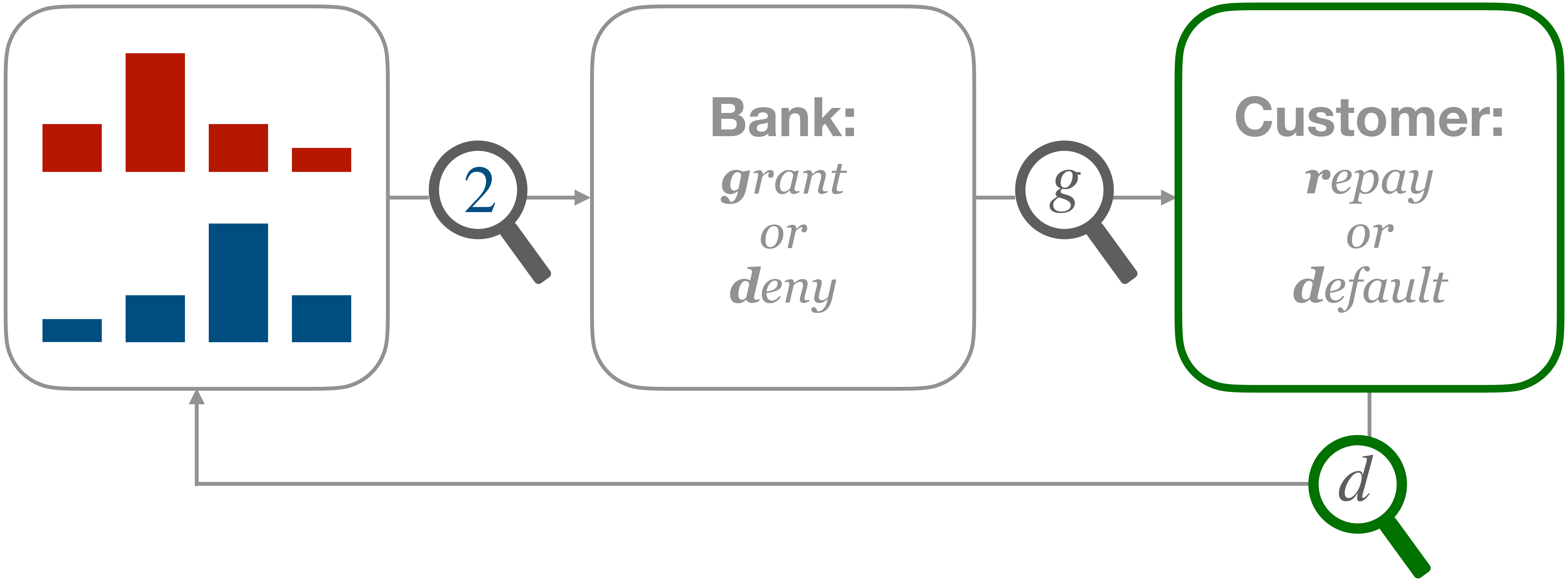

Bank:
grant
or
deny
Customer:
repay
or
default

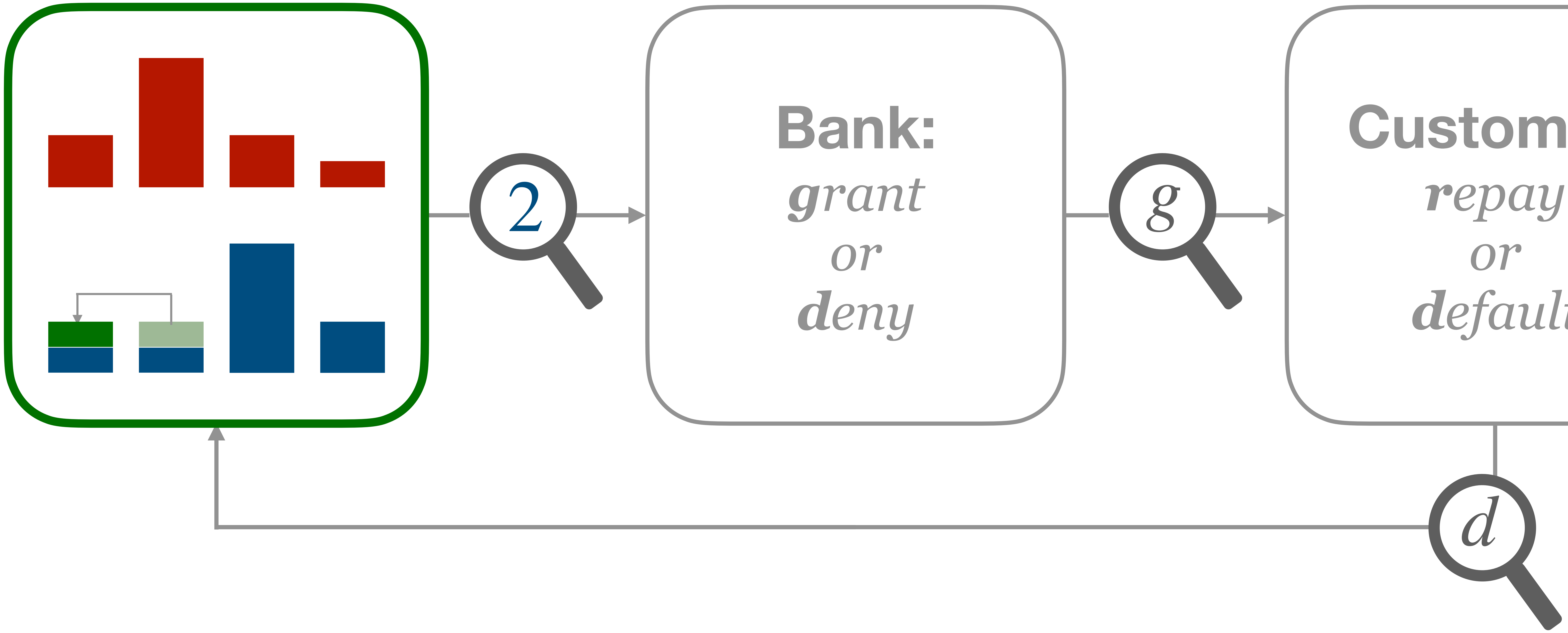


X_t
Bank:
grant
or
deny
Y_t
Customer:
repay
or
default
Z_t


2
Bank:
grant
or
deny
Customer:
repay
or
default

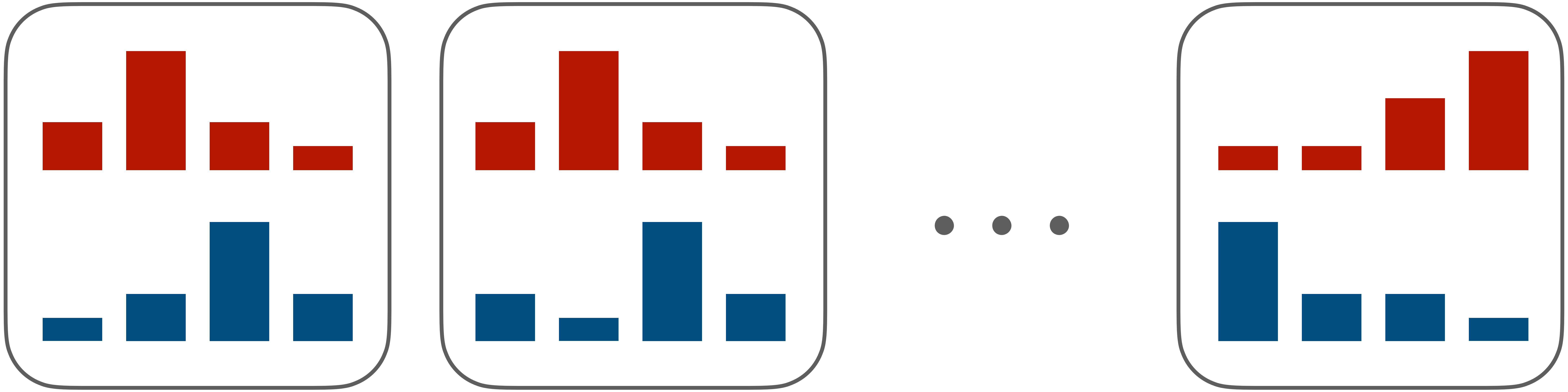

2
Bank:
grant
or
deny
g
Customer:
repay
or
default


2
Bank:
grant
or
deny
g
Customer:
repay
or
default
d



Group ***Red:*** $2.2 \xrightarrow[0]{} 2.2$

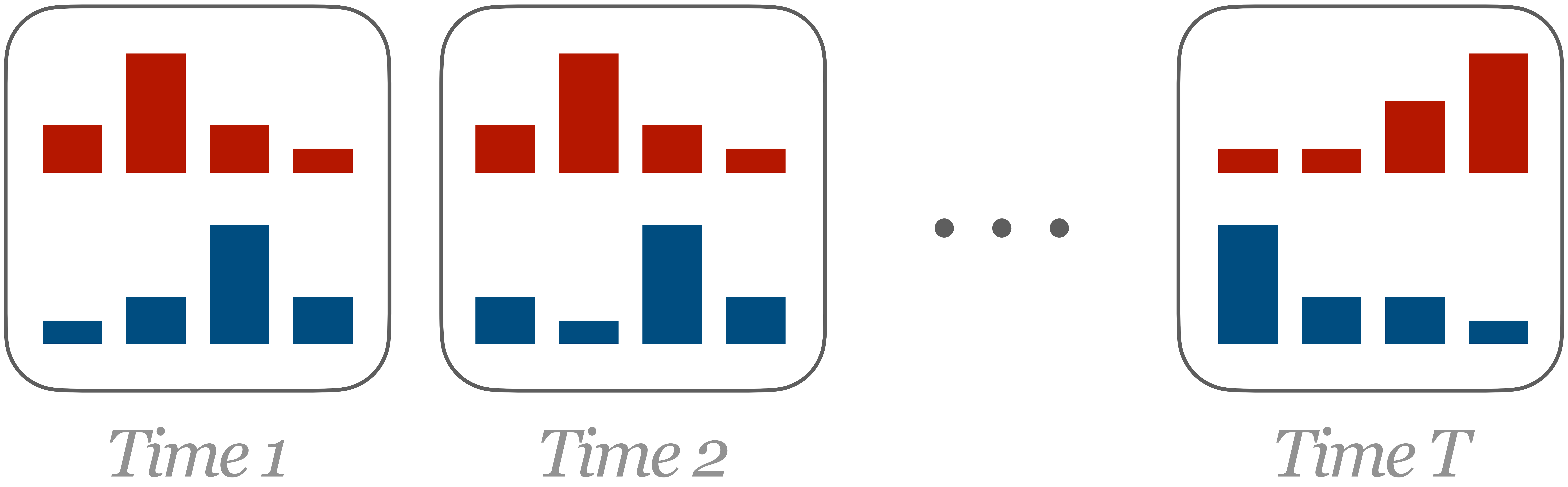
Group ***Blue:*** $2.8 \xrightarrow[\frac{1}{10}]{} 2.7$



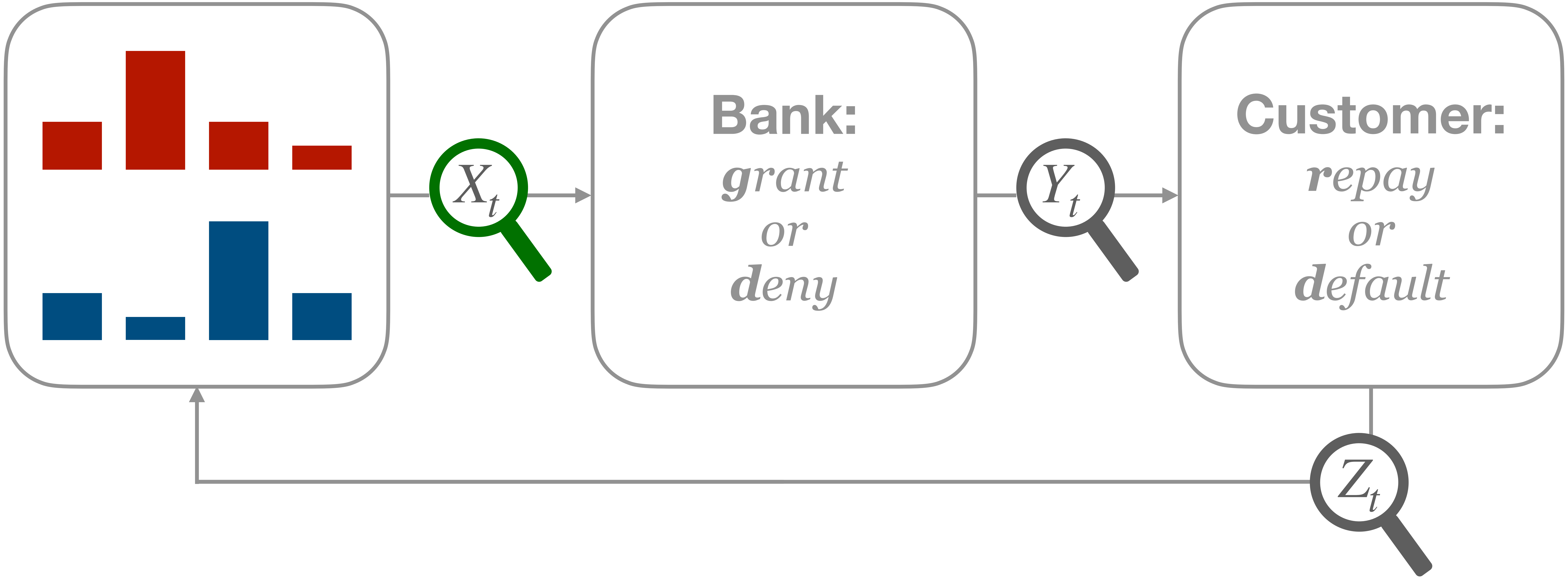
Time 1 *Time 2* ... *Time T*

2.2 2.2 ... 3.2

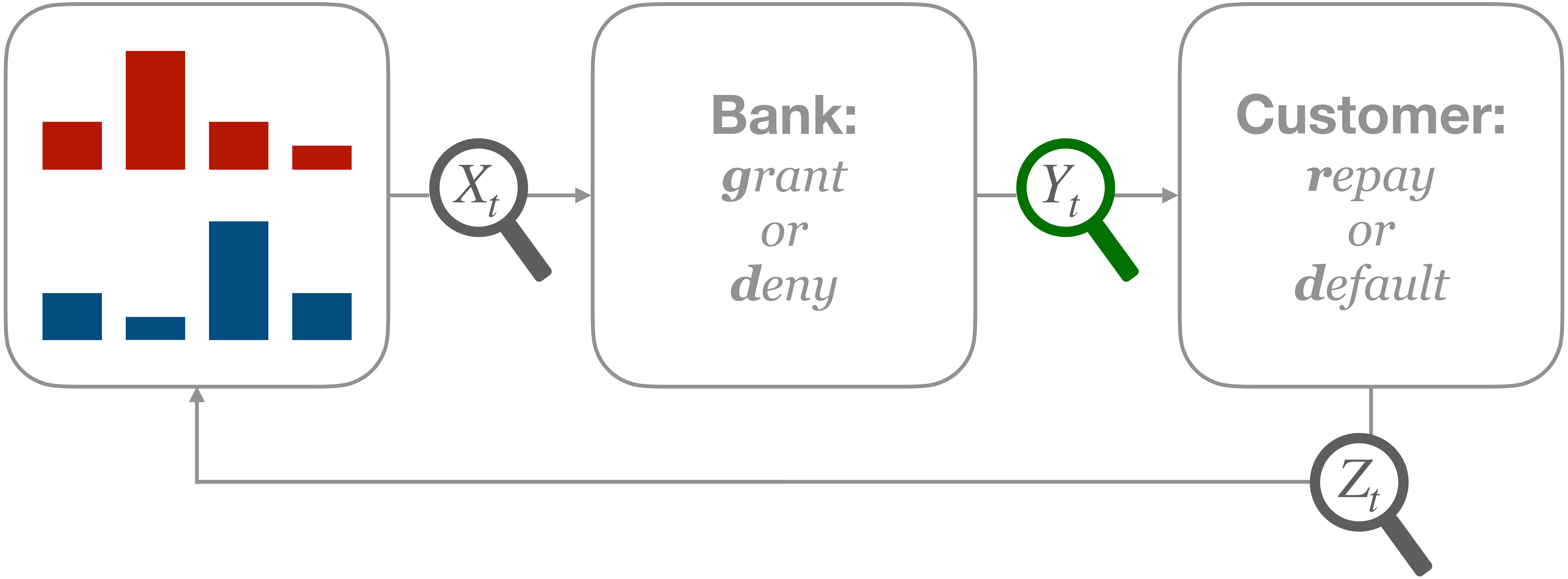
2.8 2.7 ... 1.9



Estimate the current disparity in average credit scores between Group Red and Group Blue

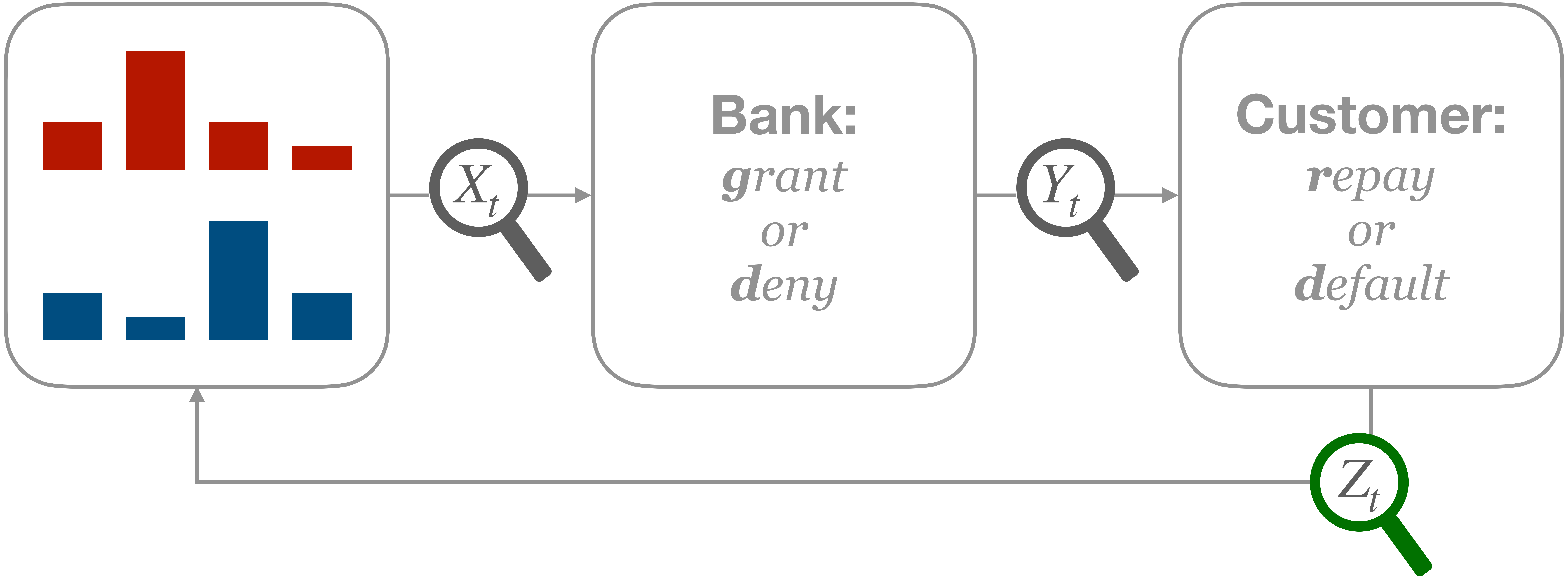


$$\overrightarrow{O}_t := O_1, \ldots, O_t = \underbrace{(X_1, Y_1, Z_1)}_{\text{Sample}}, \ldots, (X_t, Y_t, Z_t)$$



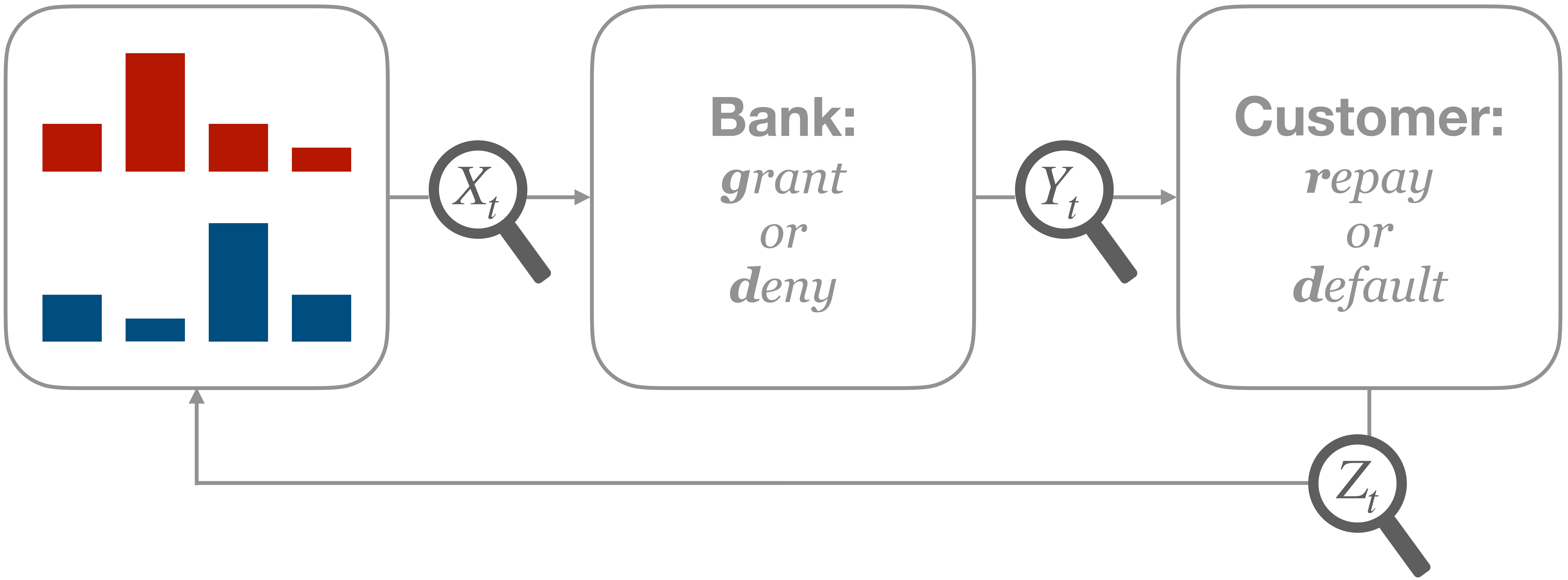
$$\overrightarrow{O}_t := O_1, \ldots, O_t = (X_1, Y_1, Z_1), \ldots, (X_t, Y_t, Z_t)$$

Decision



$$\overrightarrow{O}_t := O_1, \ldots, O_t = (X_1, Y_1, Z_1), \ldots, (X_t, Y_t, Z_t)$$

Outcome



$$\varphi(\vec{o}_t) = \mathbb{E}_R(X_t \mid \vec{o}_{t-1}) - \mathbb{E}_B(X_t \mid \vec{o}_{t-1})$$

$\perp$

Past

Problem Statement.

What are we trying to do?

Properties.

Arithmetic expressions over

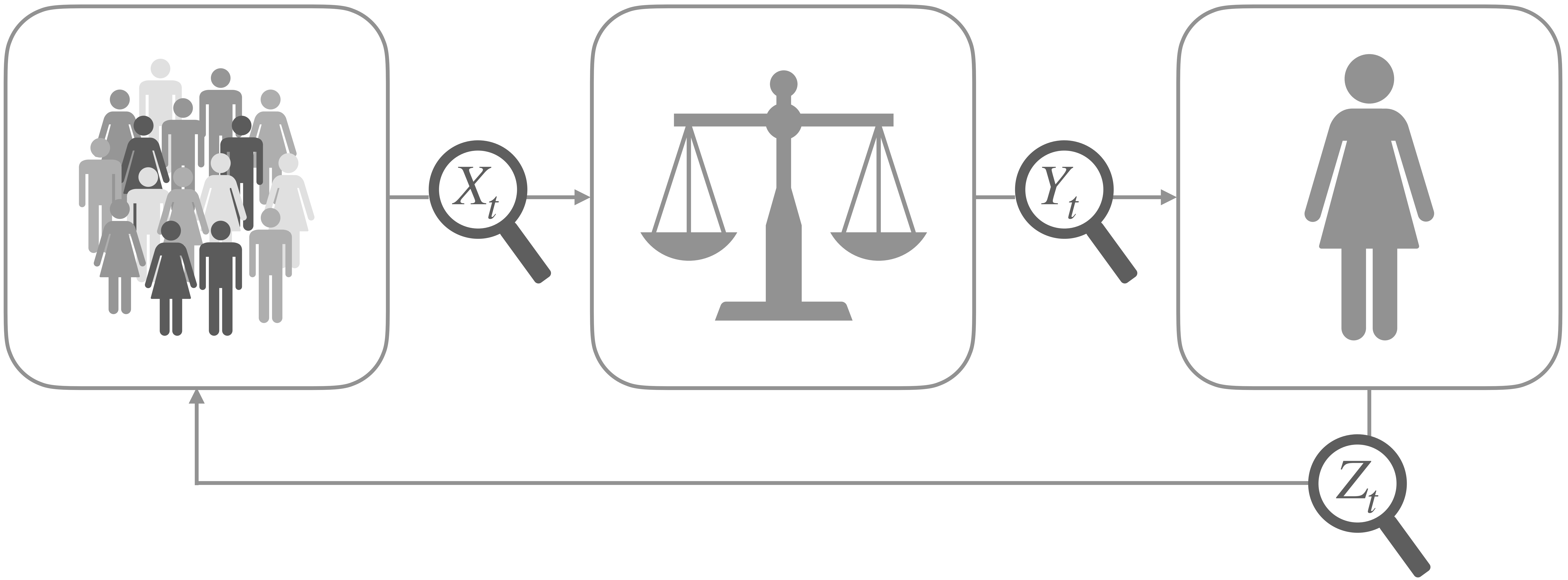
$\mathbb{E}(f(X_t) \mid \overrightarrow{O}_{t-1})$ *for* <u>*some*</u> $f : \Sigma \to \mathbb{R}$

and any $t > 0$.

$$\mathbb{P}\left(\mathbb{E}(f(X_t) \mid \overrightarrow{O}_{t-1}) \in \mathscr{A}(\overrightarrow{O}_t)\right) \geq 1 - \delta$$

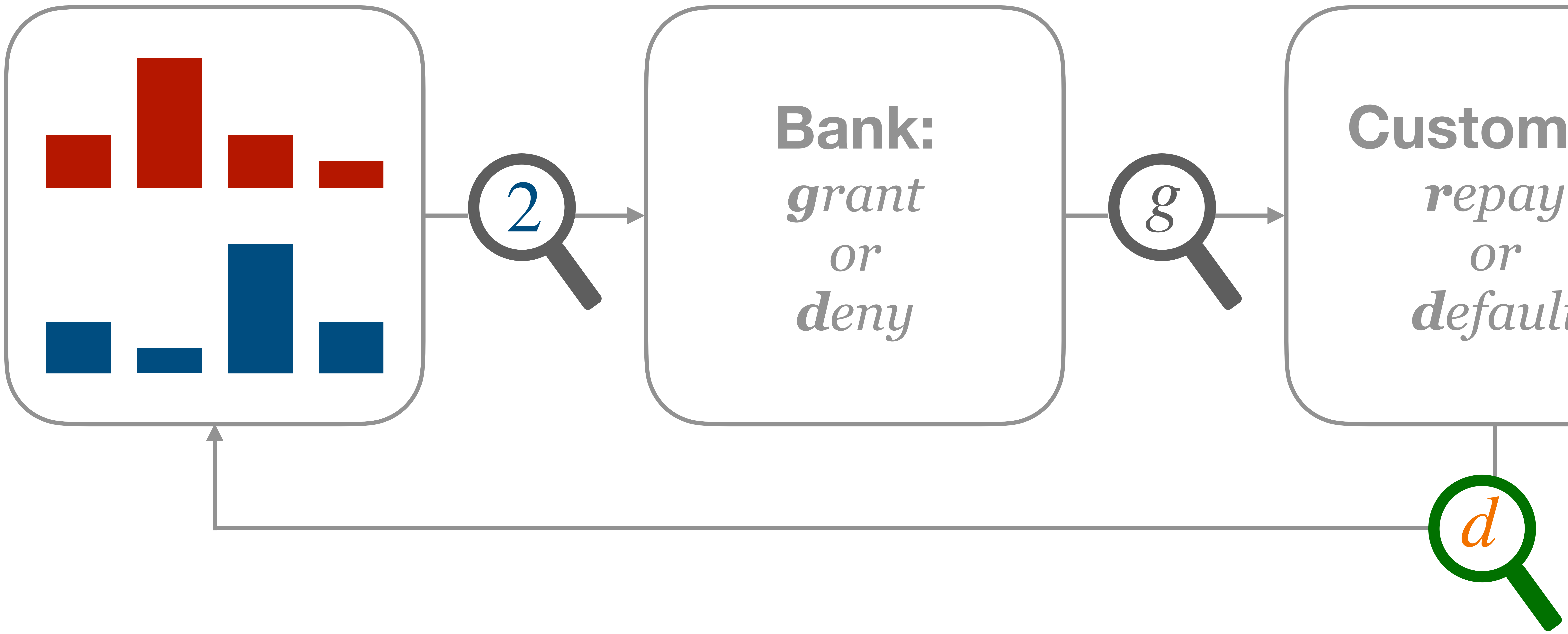
Assumptions.

Knowledge about how the expected value changes (and that X_t is sub-exponential).



$$\mathbb{E}_G(X_{t+1} \mid \vec{o}_t) = \mathbb{E}_G(X_t \mid \vec{o}_{t-1}) + \Delta(o_t)$$

Change Function



$$\mathbb{E}_B(X_{t+1} \mid \vec{o}_t) = \mathbb{E}_B(X_t \mid \vec{o}_{t-1}) - \frac{1}{n_B}$$

Algorithm.

A sketch.

Estimate $\mathbb{E}_G(X_t \mid y_t, z_t, \vec{o}_{t-1})$ for each group G.

Estimate $\mathbb{E}_G(X_t \mid y_t, z_t, \vec{o}_{t-1})$ for each group G.

Compute confidence interval of estimates.

Estimate $\mathbb{E}_G(X_t \mid y_t, z_t, \vec{o}_{t-1})$ for each group G.

Compute confidence interval of estimates.

Push confidence intervals through $f(\cdot)$.

Estimate $\mathbb{E}_G(X_t \mid y_t, z_t, \vec{o}_{t-1})$ for each group G.

Compute confidence interval of estimates.

Push confidence intervals through $f(\cdot)$.

Apply union bound (and interval arithmetic) to compute confidence interval of the property.

Confidence Interval.

Doob-Martingales and Azuma's Inequality

$$\hat{E}_1(\vec{o}_t) = \frac{1}{t}\sum_{i=1}^{t}\left(X_i - \sum_{j=1}^{i-1}\Delta(\vec{o}_j)\right)$$

Estimator accounts for the shift

$$\mathbb{E}(X_{t+1} \mid \vec{o}_t) - \mathbb{E}(X_t \mid \vec{o}_{t-1}) = \Delta(\vec{o}_t)$$

$$\mathbb{E}(\hat{E}_1(\overrightarrow{O}_t)) = \mathbb{E}(X_1)$$

Unbiased

$$\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\right), \mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_1\right), \ldots, \mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_t\right)$$

Doob-Martingale

$$\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_{k+1}\right) - \mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_k\right)$$

Bound Difference

$$\mathbb{P}\left(\left|\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\right) - \mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_t\right)\right| \geq \varepsilon\right) \leq \delta$$

Azuma's inequality

$$\mathbb{P}\left(\left|\overbrace{\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\right)}^{\mathbb{E}(X_1)} - \mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_t\right)\right| \geq \varepsilon\right) \leq \delta$$

Azuma's inequality

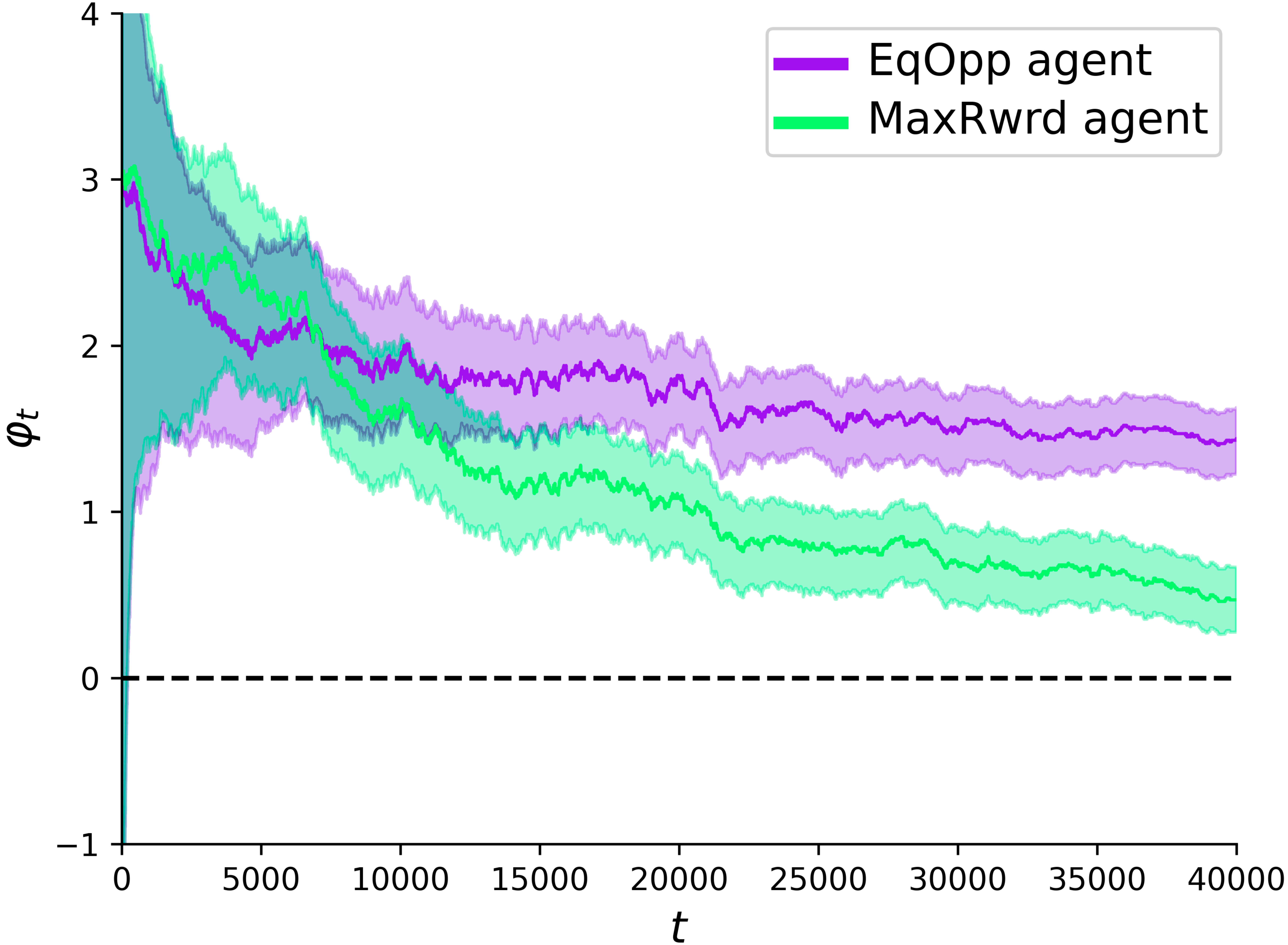
$$\mathbb{P}\left(\left|\overbrace{\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\right)}^{\mathbb{E}(X_1)} - \overbrace{\mathbb{E}\left(\hat{E}_1(\overrightarrow{O}_t)\middle|\overrightarrow{O}_t\right)}^{\hat{E}_1(\overrightarrow{O}_t)}\right| \geq \varepsilon\right) \leq \delta$$

Azuma's inequality

Experiments.

Lending and Attention (D'Amour 2020).

D'Amour et al. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies


EqOpp agent
MaxRwrd agent
φ_t
t
4
3
2
1
0
−1
0
5000
10000
15000
20000
25000
30000
35000
40000

Related Work.

What has been done so far?

Static verification of algorithmic fairness

Albarghouthi, et al. "Fairsquare: probabilistic verification of program fairness." OOPSLA 2017.

Bastani et al. "Probabilistic verification of fairness properties via concentration." OOPSLA 2019.

Ghosh et al. "Justicia: A stochastic sat approach to formally verify fairness." AAAI 2021.

Sun, et al. "Probabilistic verification of neural networks against group fairness." FM 2021.

Ghosh, et. al. "Algorithmic fairness verification with graphical models." AAAI 2022.

Monitoring algorithmic fairness

Albarghouthi and Vinitsky. "Fairness-aware programming." FAccT 2019.

Henzinger et al. "Monitoring Algorithmic Fairness." CAV 2023.

Henzinger et al. "Runtime Monitoring of Dynamic Fairness Properties." FAccT 2023.

Henzinger et al. "Monitoring Algorithmic Fairness under Partial Observations." RV 2023.

Summary.

Main points.

Interested in monitoring “distributional” properties, e.g. conditional expectation, of stochastic processes.

Leverage tools from non-asymptotic statistics to provide valid guarantees for each time step.

We focused on monitoring Algorithmic Fairness, but those techniques have wide applicability.

Institute of
Science and
Technology
Austria