The→
Incredible
→Machine

Sprint 6 — Ideas for 3rd party scrutiny of algorithms

context

Governments increasingly rely on algorithms to support decision making...

... which has led to several cases of discrimination.

NOS Nieuws • Maandag 16 januari, 20:28

Extra toezicht moet einde maken aan 'levensgevaarlijke algoritmes'

De Autoriteit Persc
Die kunnen bijvoor
een sollicitatiegesp
controle op fraude

Het extra toezicht
kabinet er een milj
naar structureel 3,
bedrijfsleven krijge
Wolfsen is het har

NIEUWS

'Belastingdienst gebruikte algoritme dat lage inkomens selecteerde voor extra fraudecontroles'

Voor de controle op fraude heeft de Belastingdienst
flerend algoritme dat onder
. Huishoudens met een laag
hogere risicoscore dan die

Alleenstaande vrouwen en anderstaligen eerst: Rotterdam gebruikte discriminerend algoritme om bijstandsfraude op te sporen

· Nieuws · 06-03-2023 · leestijd 2 minuten · 1767 keer bekeken · bewaren

context

Dutch gov’t promised to improve oversight on algorithms and data.

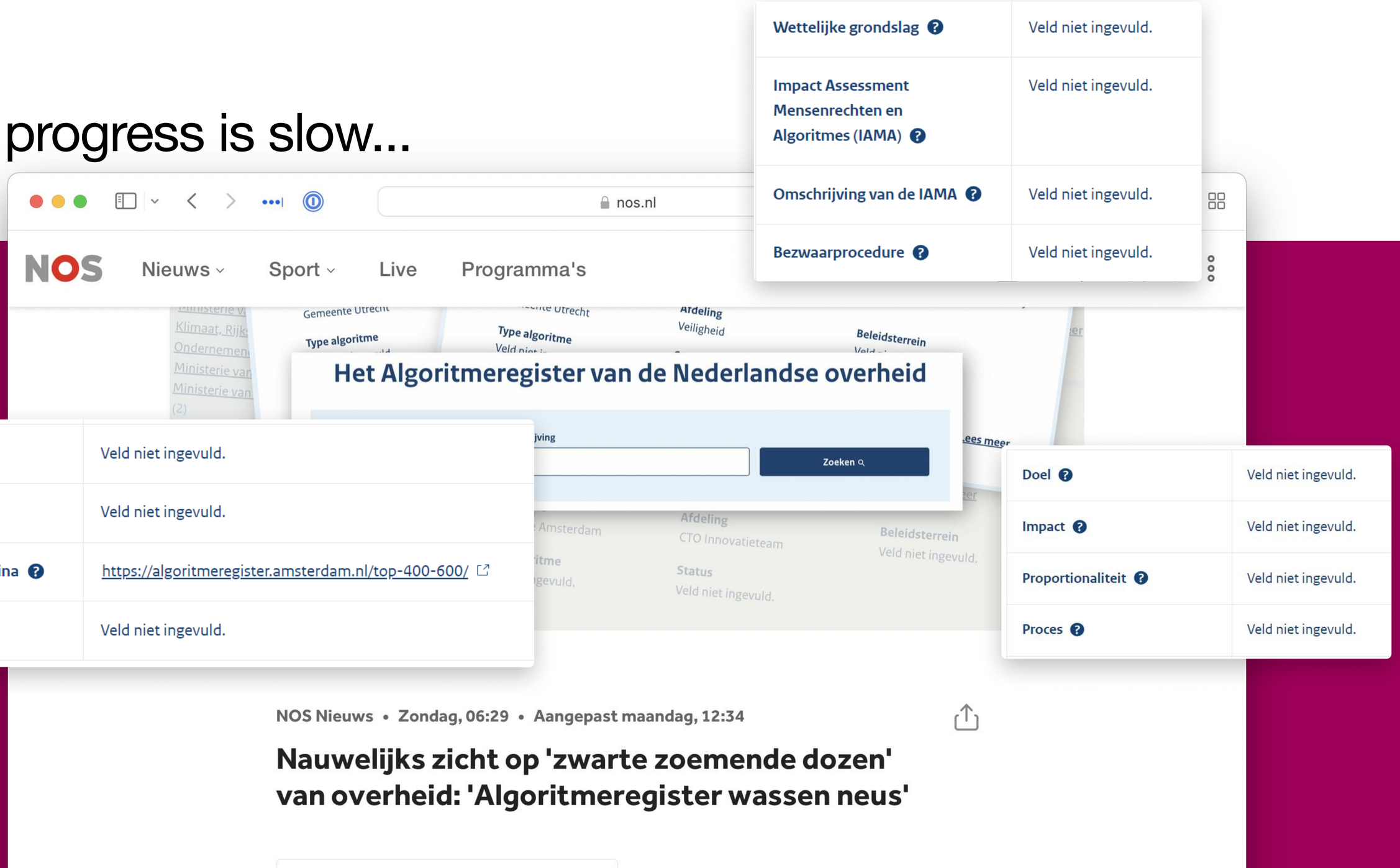
From the 2022 government coalition agreement

> “De overheid verzamelt en deelt (onderling) niet meer data dan nodig is en ontwikkelt regels voor data ethiek in de publieke sector.”
>
> “The government does not collect and share (with each other) more data than necessary and develops rules for data ethics in the public sector.”

> “Algoritmes worden wettelijk gecontroleerd op transparantie, discriminatie en willekeur [...] een algoritmetoezichthouder bewaakt dit.”
>
> “Algorithms are legally checked for transparency, discrimination and arbitrariness [...] an algorithm regulator monitors this.”

context

But progress is slow...



challenge

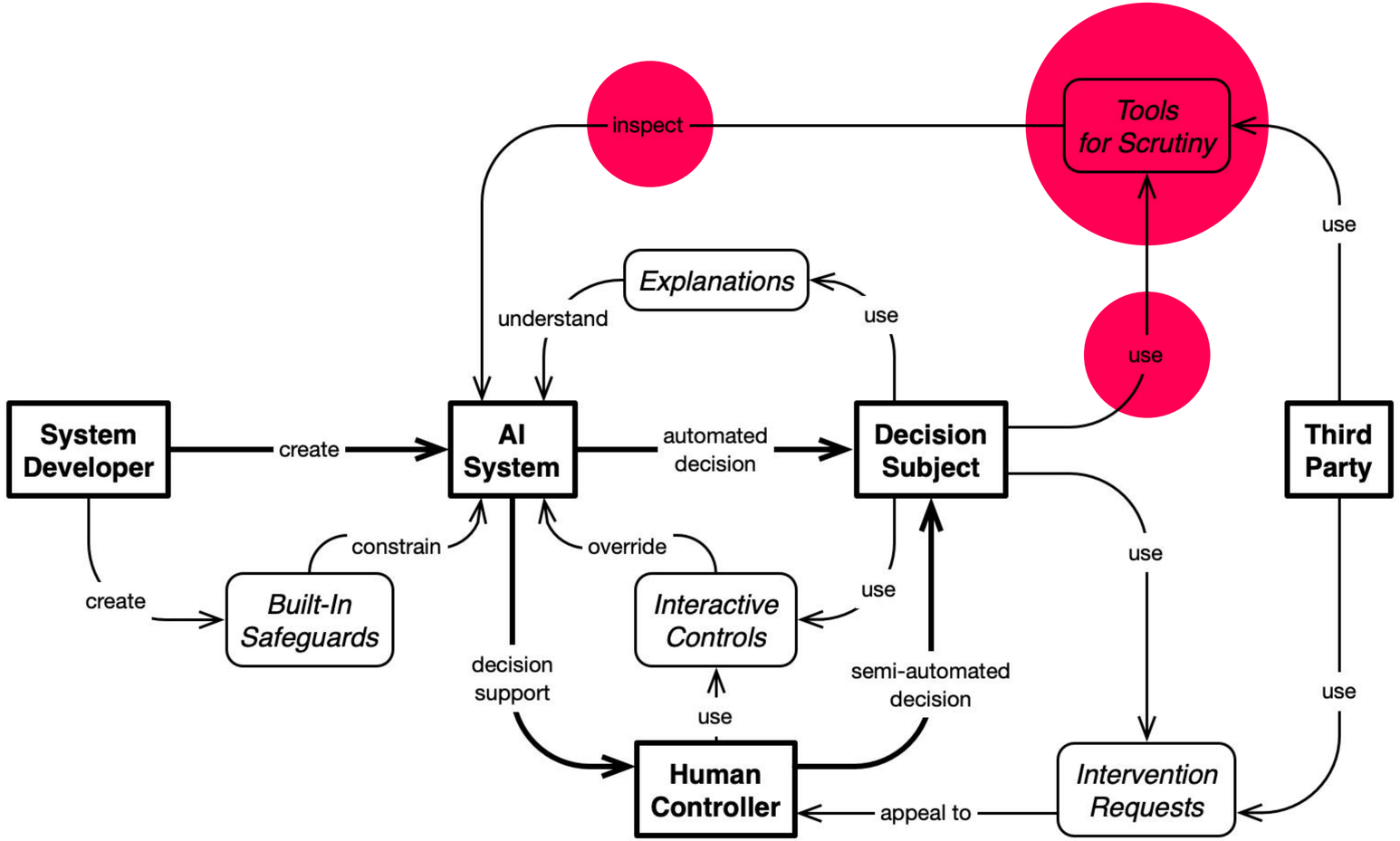
How can we — as society — scrutinize algorithmic outcomes to monitor effectiveness, discrimination, or arbitrariness...

...without exposing privacy sensitive or confidential information?

3rd Parties can be human rights organizations, advocacy groups, journalists, researchers, etc.

focus

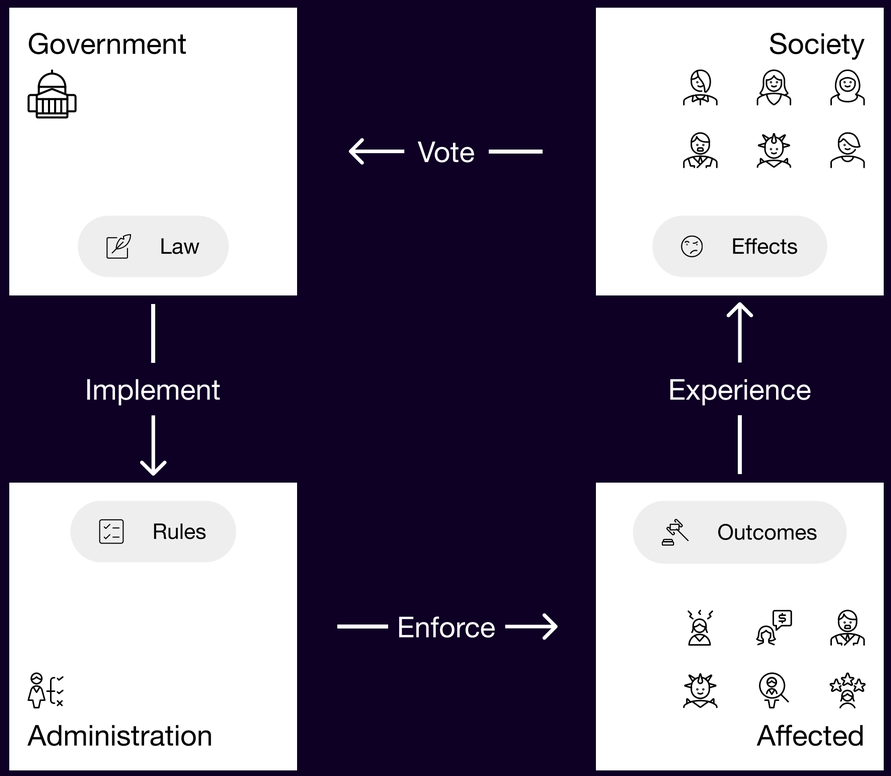
Leverage point: Tools for Scrutiny



(Alfrink et al, 2022)

context

Classic enforcement



The administration employees decide how implement the law enforce the rules.

The experienced effects of this enforcement can be affected by political action that inform new laws and how they should be enforced.

context

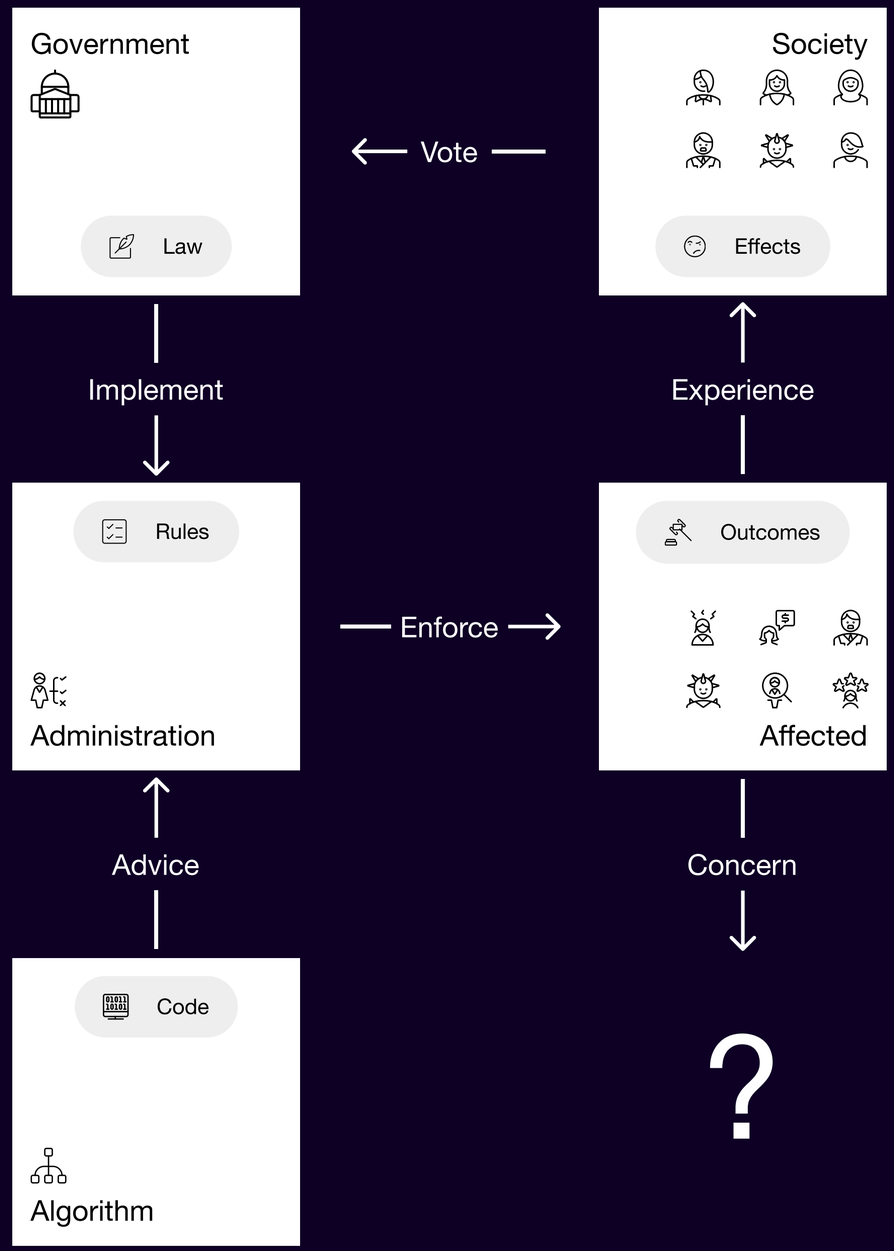
Algorithmic advice

The administration employees implement the rules in a algorithm that helps with enforcement (who to target).

The experienced effects of this enforcement are harder to affect, since the enforcement by the algorithm is a black box.

How can concerned citizens understand how the algorithm implements the rules?

Where can they go with their concerns?

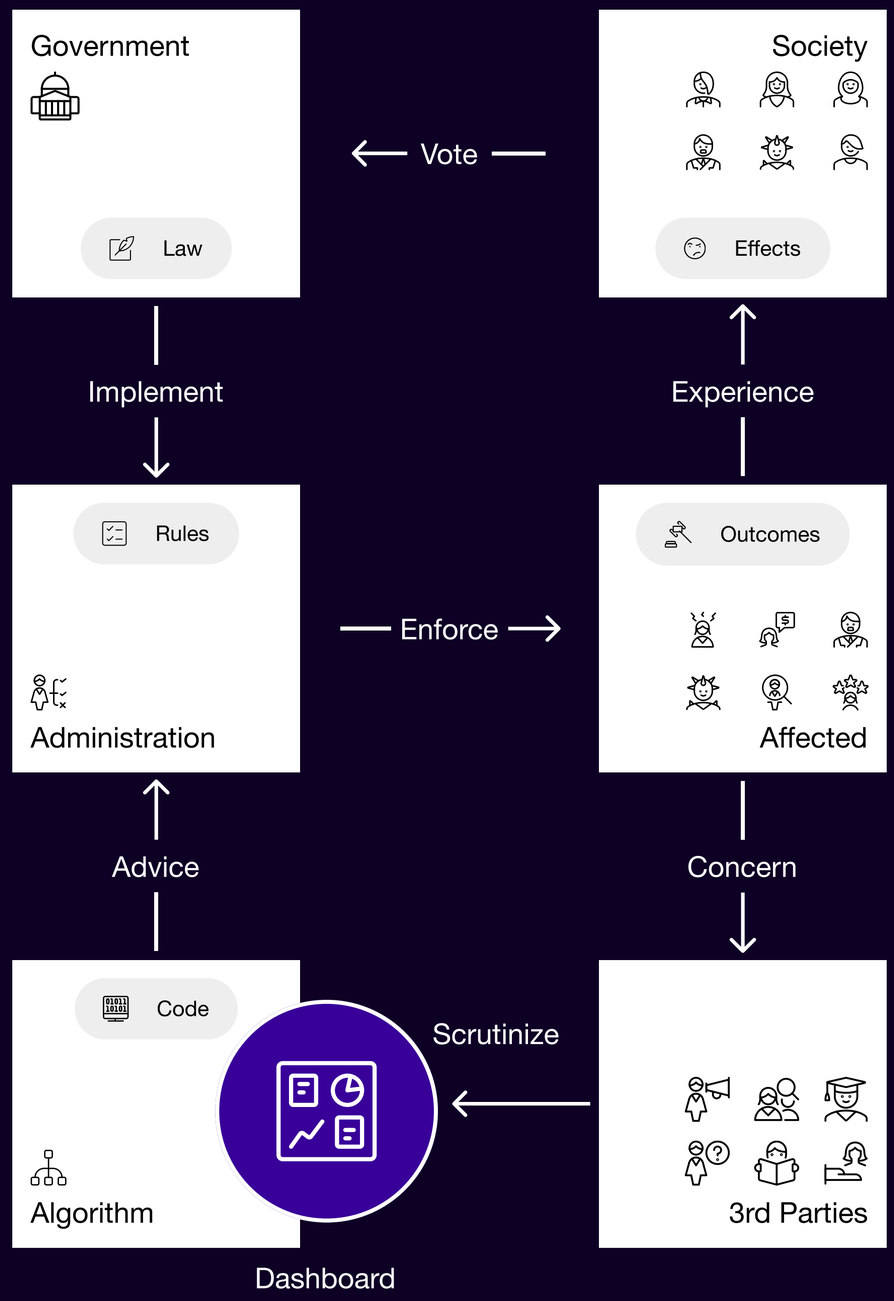


context

Scrutiny with a dashboard

A dashboard can help 3rd parties to scrutinise the working of an algorithm.

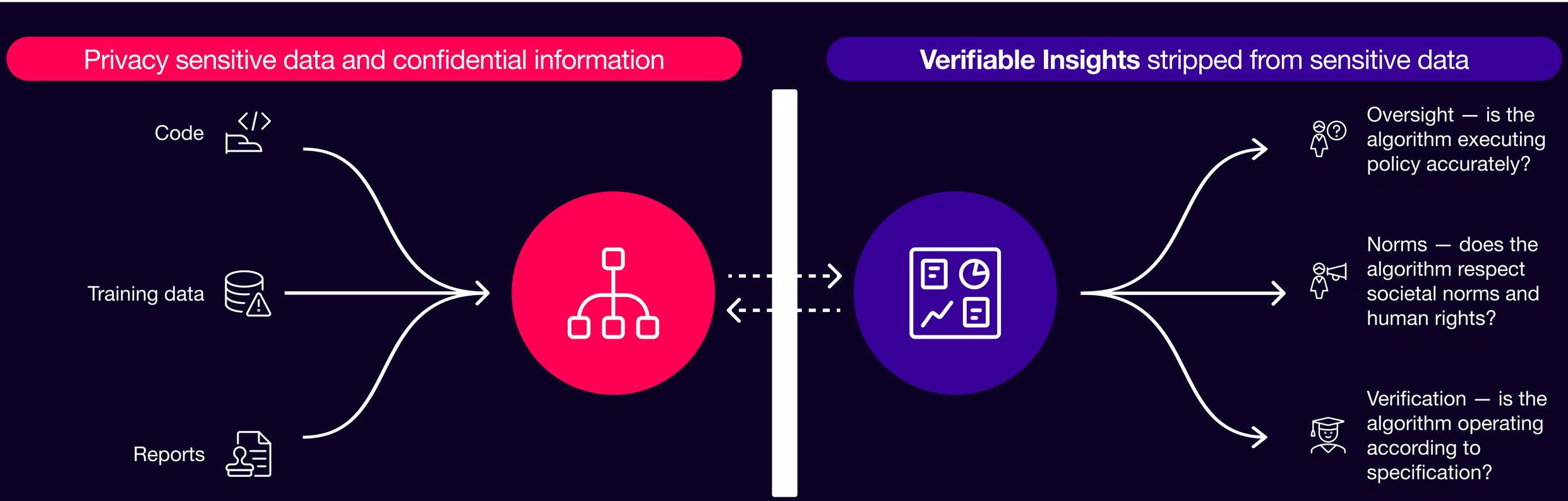
Concerned citizens can turn to them to understand how the rules are implemented and are informed enough to try and affect the implementation.



solution

The Algorithm Dashboard

The algorithm dashboard allows 3rd parties to monitor the algorithm without exposing sensitive data.



solution

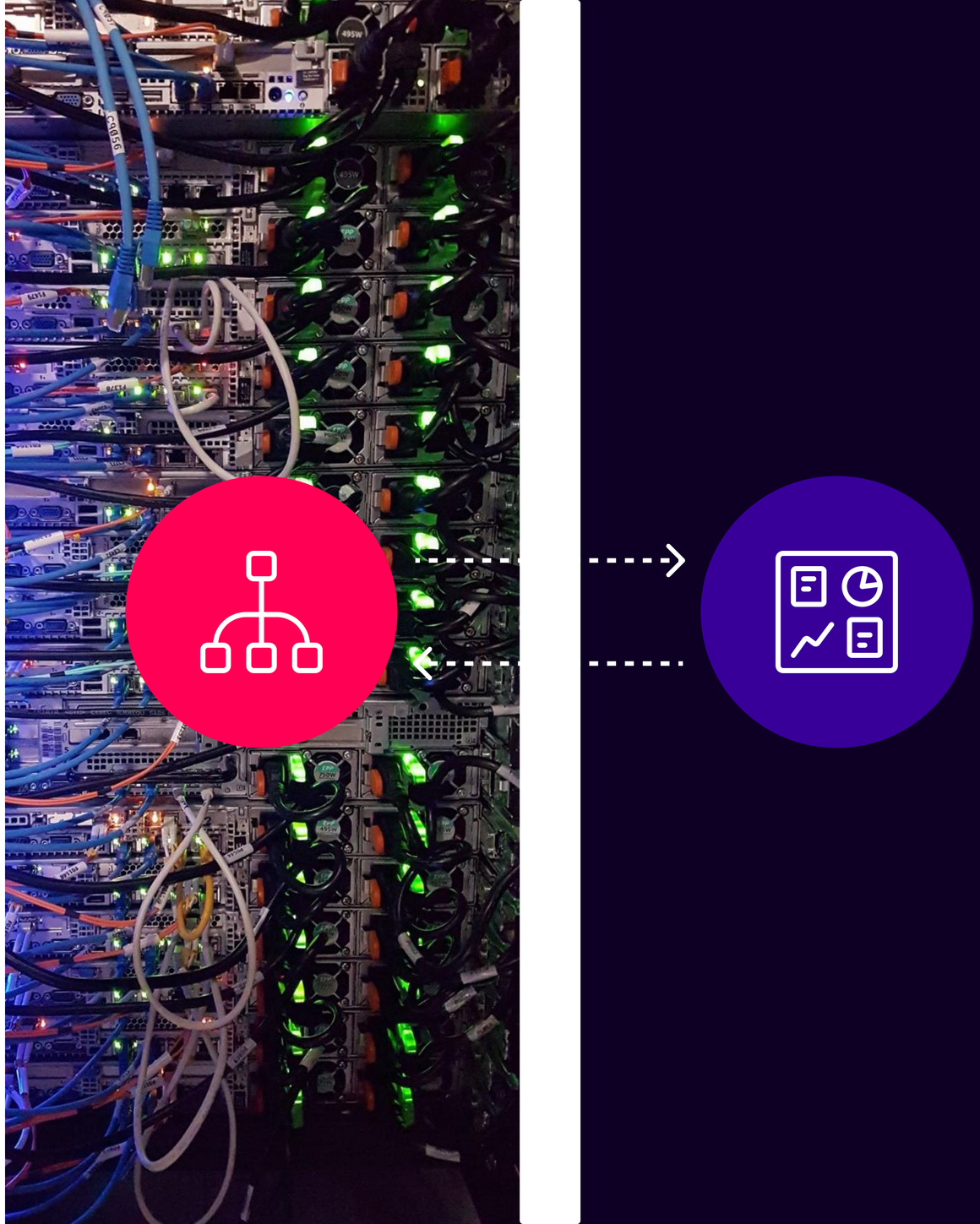
Scrutiny strategies

How can you monitor an algorithm if you are not allowed to see the data that is processes?

Strategy 1: Audit

Strategy 2: Twin

Strategy 3: Watchdog



strategy 1:

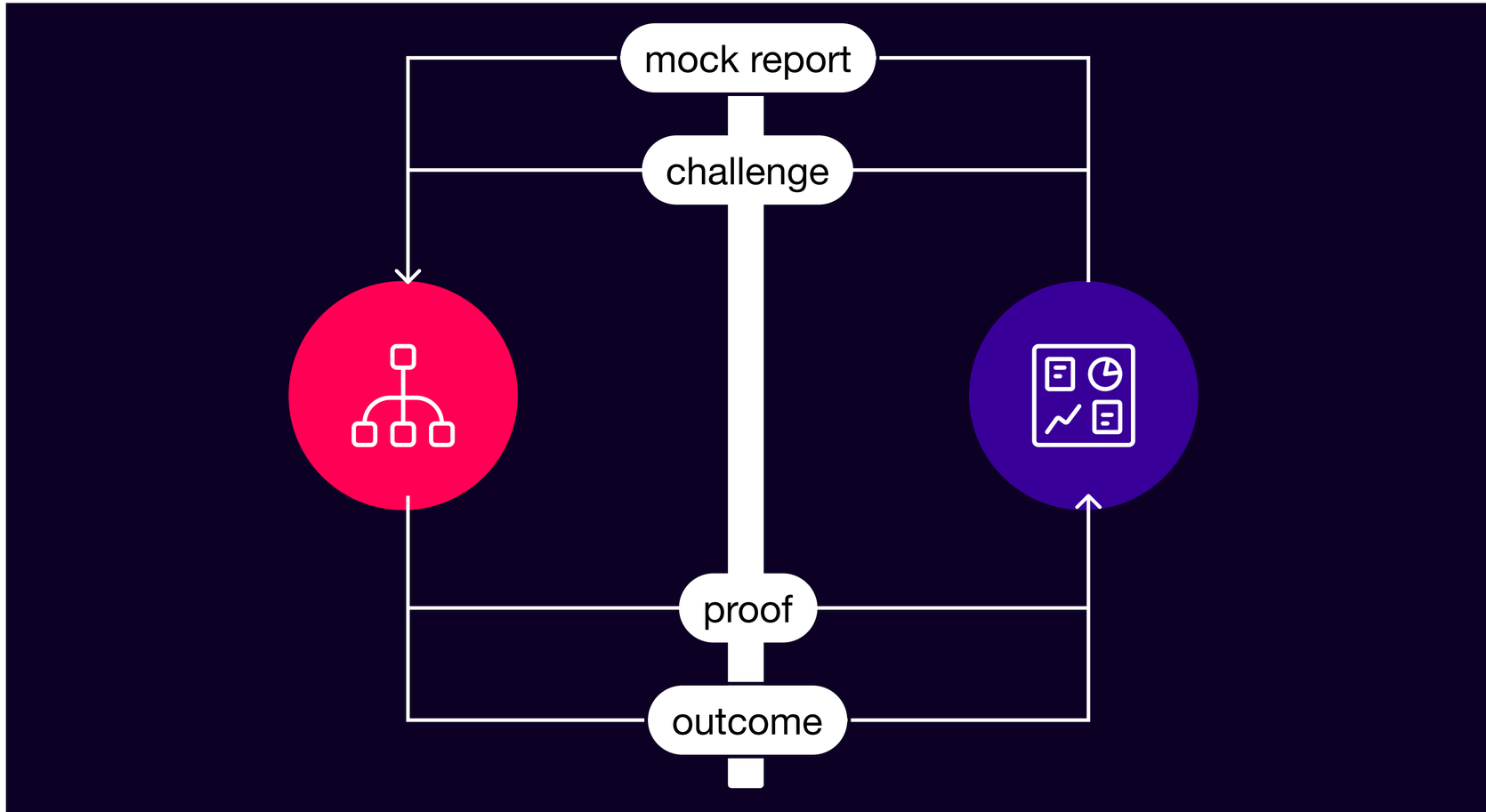
Audit

- Test implemented algorithm
- No view on data
- Occasional testing

The Audit strategy can be used to occasionally send a mock report to the algorithm and will return the output.

It allows for testing the implemented algorithm and verify whether no adjustments have been made recently.

The input data contains a challenge to create a cryptographic proof that the algorithm processed the report as claimed.



strategy 2:

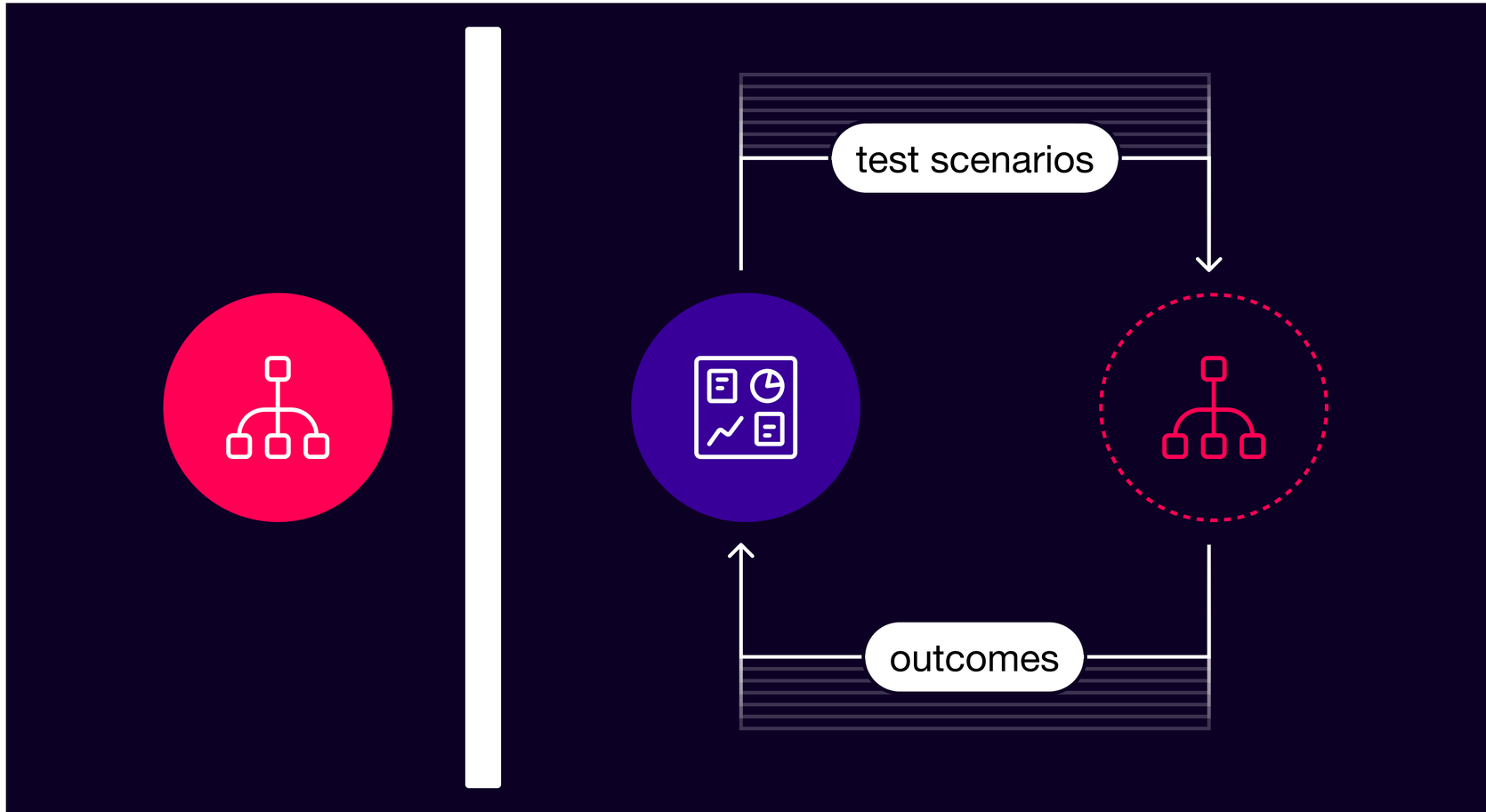
Twin

- Run unlimited test scenarios
- Inspect code
- Implemented could be different

The Twin strategy uses instance of the algorithm source code installed on a different server, managed by a 3rd party.

This allows the 3rd party to generate insights from large scale tests.

The outcomes can be used to verify whether outcomes from audits match.



strategy 3:

Watchdog

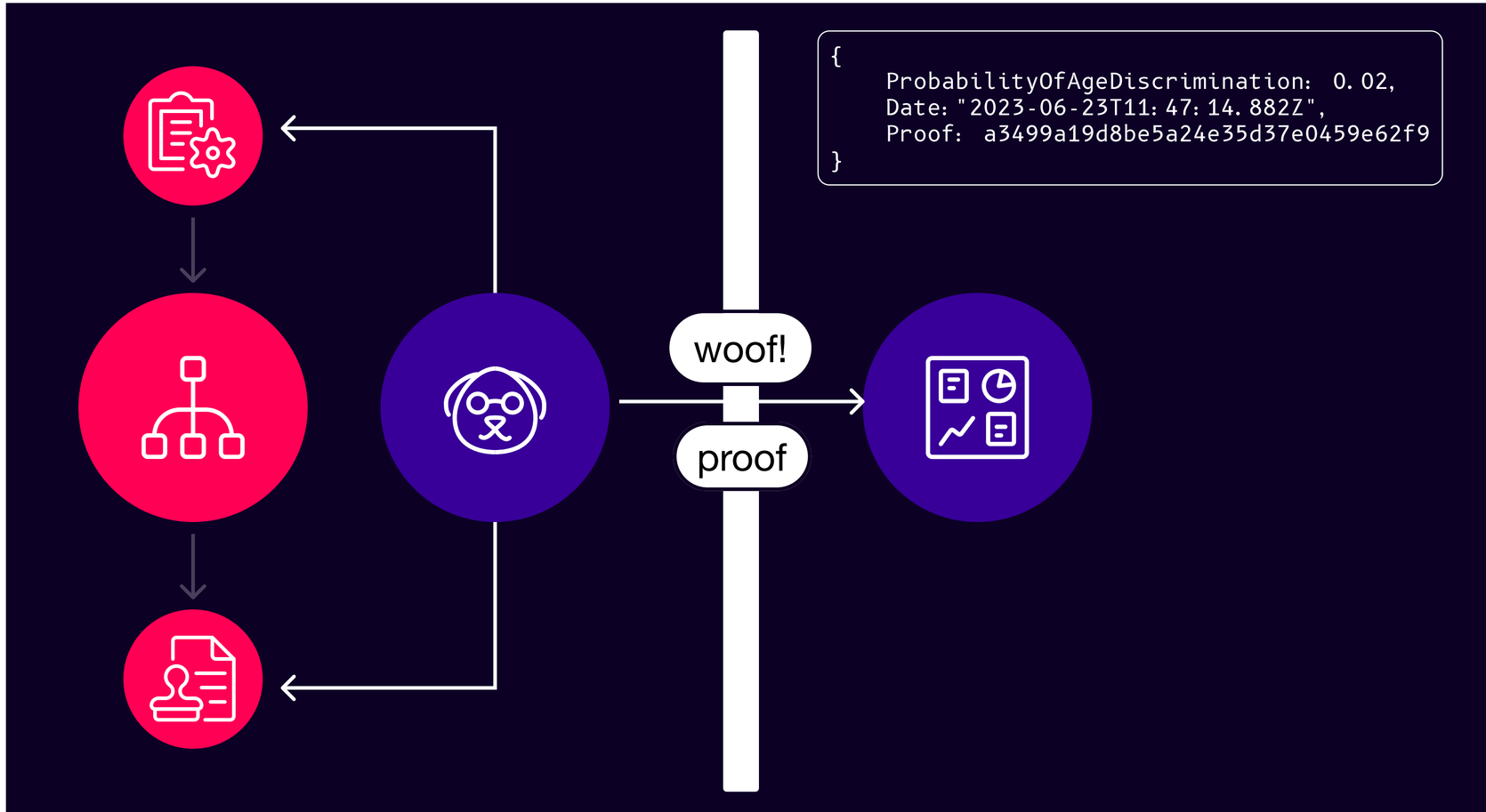
- Can inspect real in- and output data
- Can analyse against reference data
- Can only send occasional indicator

The Watchdog strategy allows running a piece of code behind the privacy wall.

It can monitor in and out coming data, as wel as compare it with data sets it knows.

Every day, the watchdog can report 1 indicator to the dashboard, hence protecting sensitive data from leaking.

Once installed the watchdog can no longer be accessed by 3rd parties.



summary

Complete perspective

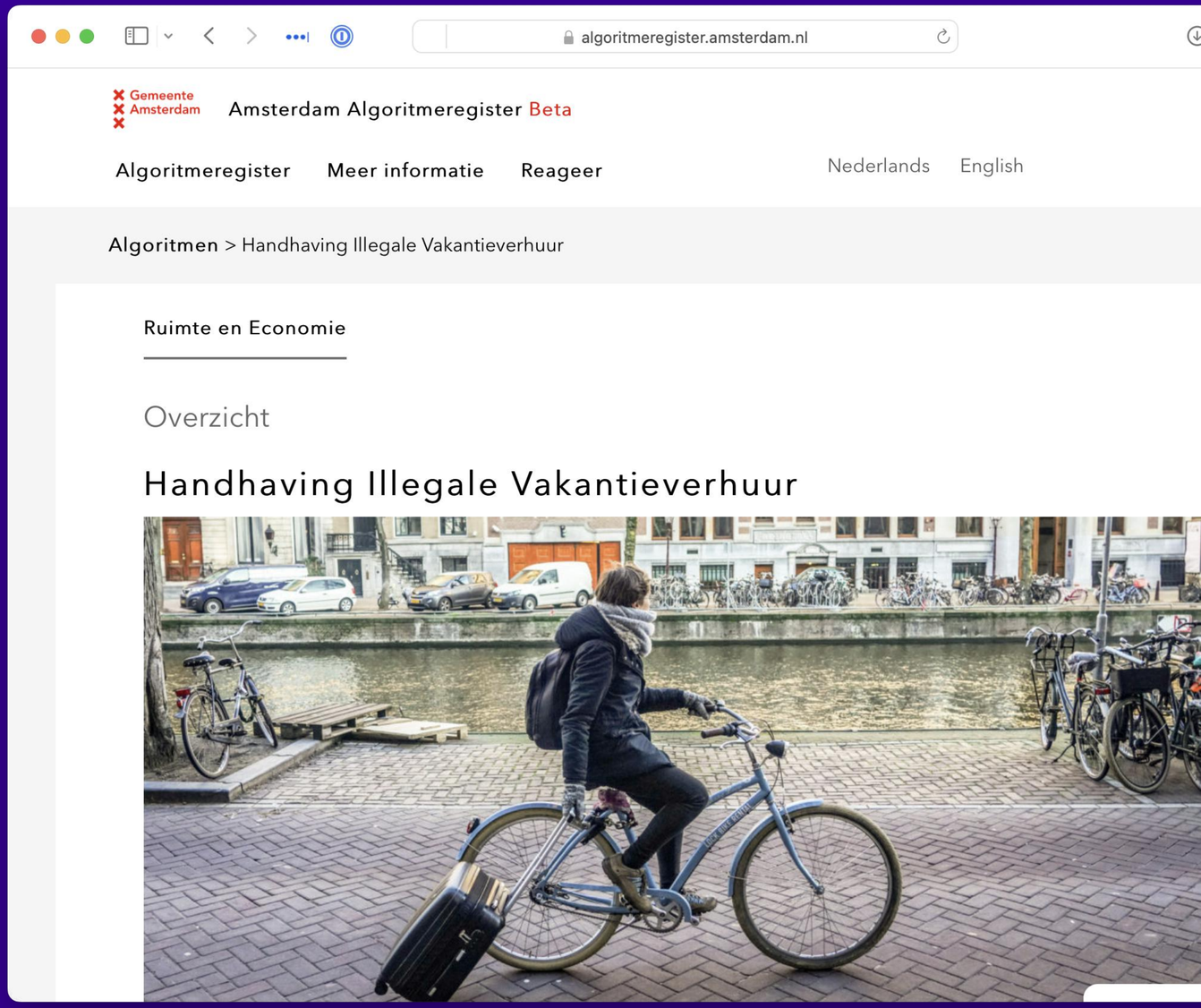
	implemented algorithm	actual data	deep inspection
Twin	✗	✗	✓
Audit	✓	✗	✗
Watchdog	✓	✓	✗

context

Illegal holiday rental housing algorithm

An algorithm that supports the employees of the department of Surveillance & Enforcement in their investigation of the reports made concerning possible illegal holiday rentals.

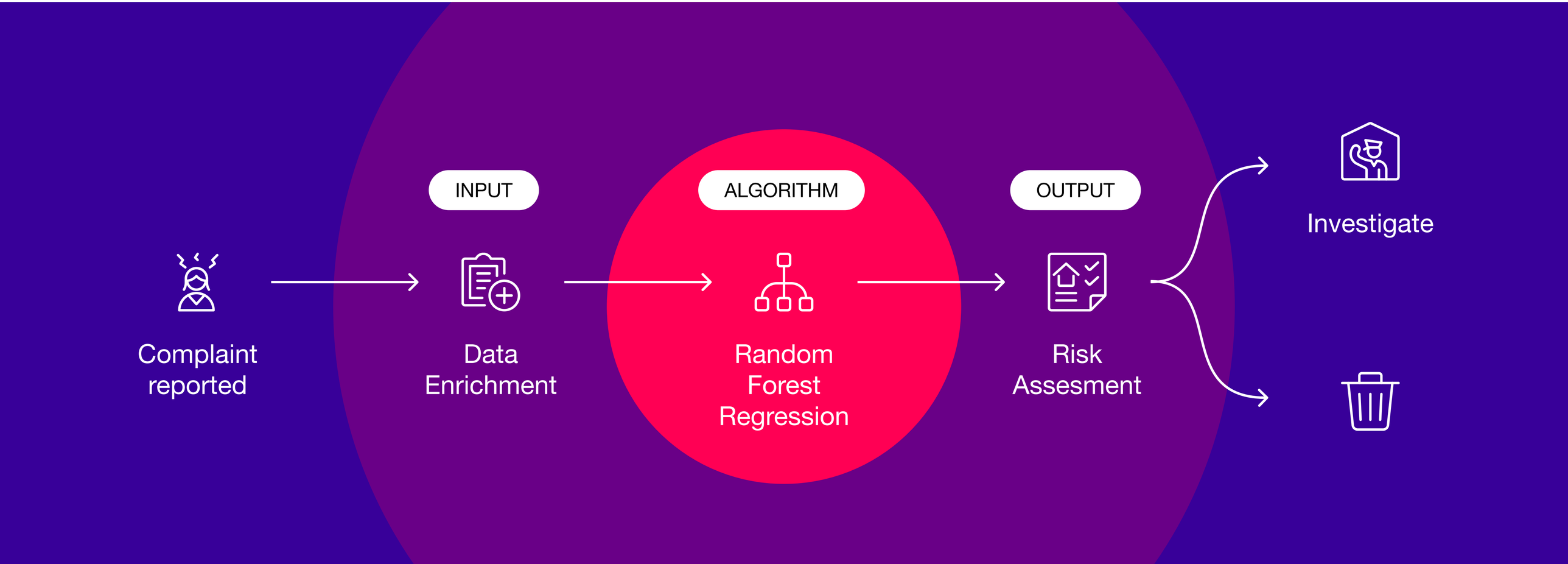
Note: This algorithm will not be actually used by the city of Amsterdam



context

How does it work? (Simplified)

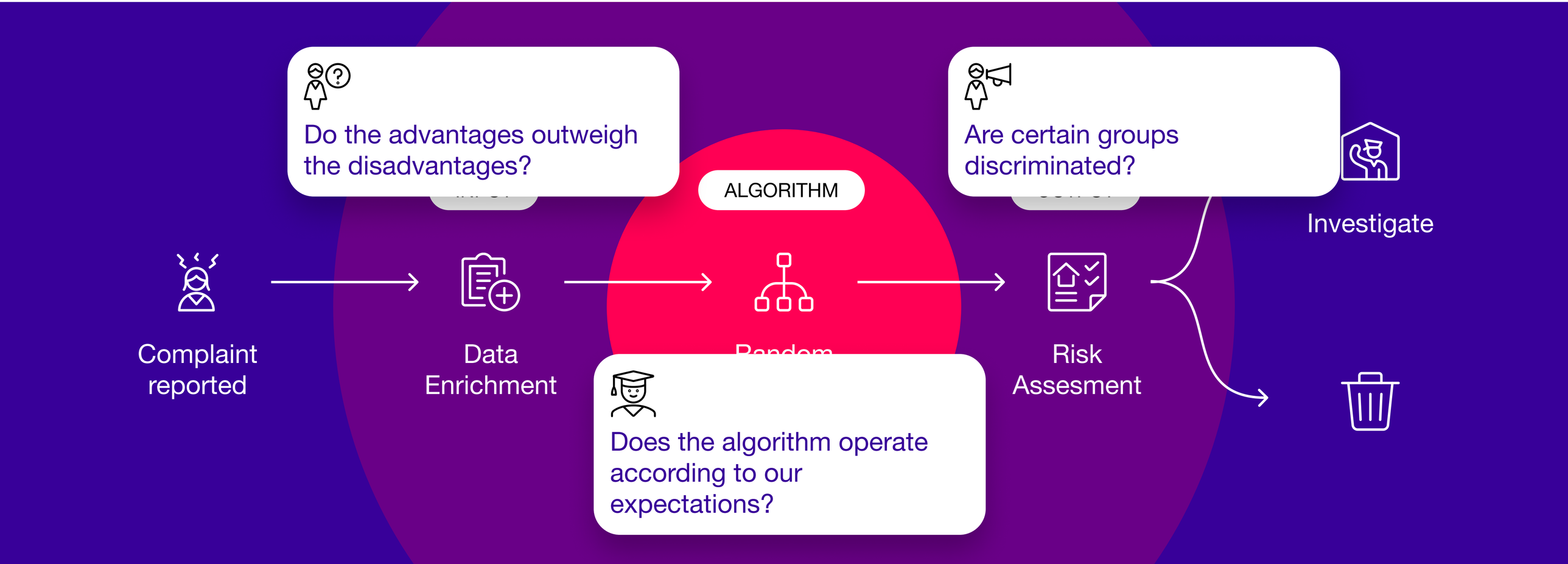
When a citizen files a complaint about illegal holiday rental activities, their report is enriched with data (e.g. building data), and then processed by the algorithm. The algorithm makes a risk assessment for the case which is then selected for further investigation or not.

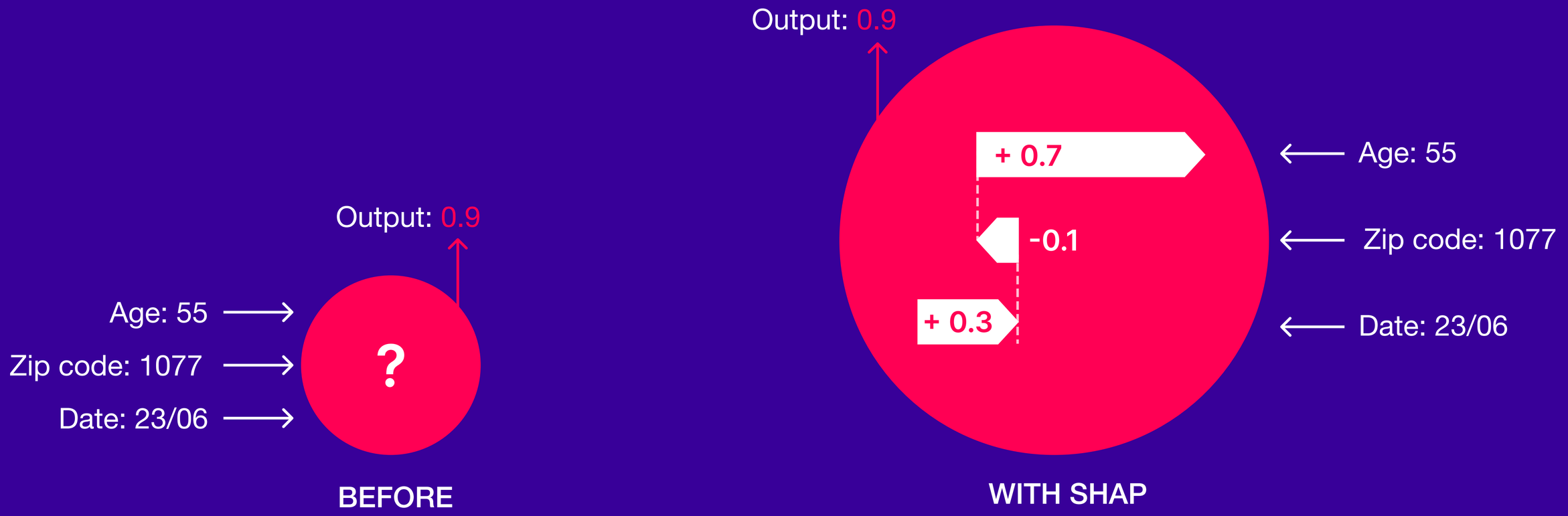


context

How does it work?
(Simplified)

When a citizen files a complaint about illegal holiday rental activities, their report is enriched with data (e.g. building data), and then processed by the algorithm. The algorithm makes a risk assessment for the case which is then selected for further investigation or not.





context

Risk assessment with SHAP

Currently, overseers are provided with an explanation of the algorithm's output. This explanation reveals the contribution of each feature towards the final model output, or ‘risk indication score’, illustrating the extent to which different characteristics of the dataset has influenced the results.

Google

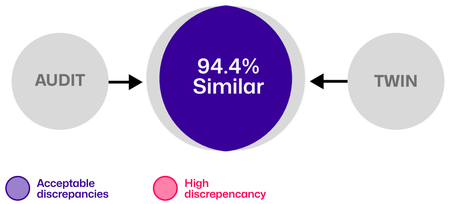
google.com

Dashboard Illegal Holiday Rental Algorithm

Audit | Twin | Verify Operation

Is the algorithm implemented as claimed?

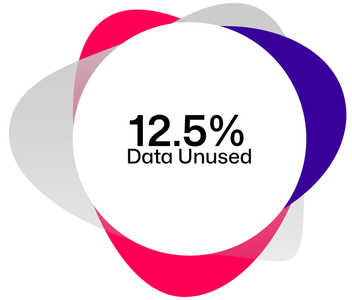
Compare test server data with audits on implemented algorithm.



Twin | Verify Operation

Has data been minimized sufficiently?

Run cases through the test server to determine to what extent the data that is claimed to be necessary for conducting a risk assessment is actually utilized.



Author: Bits of Freedom
Inspect: www.bitsoffreedom.nl
Last updated: 18 hours ago

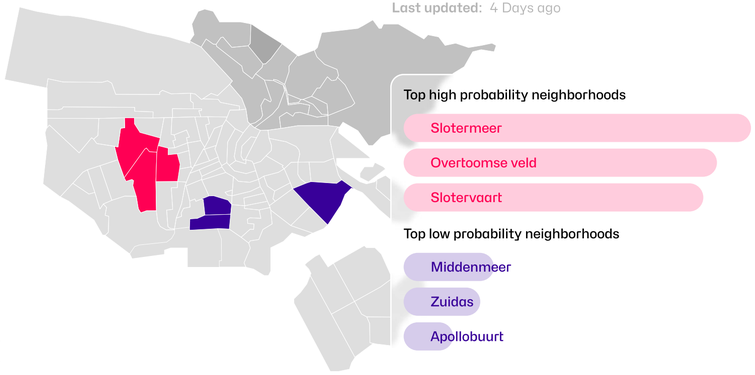
Personal data | Building data | Case description

Watchdog | Discover bias

Areas categorized as high / low probability

Measure the ratio of live high probability cases in each neighbourhood to discover possible patterns.

Author: Nationale ombudsman
Inspect: www.nationaleombudsman.nl
Last updated: 4 Days ago



Top high probability neighborhoods

- Slotermeer
- Overtoomse veld
- Slotervaart

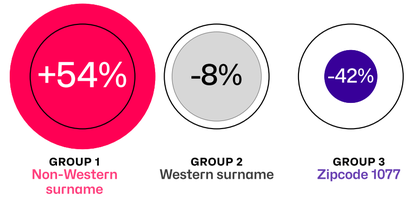
Top low probability neighborhoods

- Middenmeer
- Zuidas
- Apollobuurt

Watchdog | Monitor bias

Is there disproportional representation of certain groups among cases with high or low risk profiles?

Compare the proportion of selected groups in the output of the algorithm to the expected proportions from scientific and statistical studies.



Author: Rekenkamer
Inspect: www.amsterdam.nl/rekenkamer/
Last updated: 2 hours ago

Underrepresented | Overrepresented

widget

Bias monitor

Watchdog

Test for bias in the output of the algorithm. Any group that the 3rd party suspects are subject to bias can be tested. The output of the algorithm is never exposed to the widget.

A 3rd party submits a watchdog

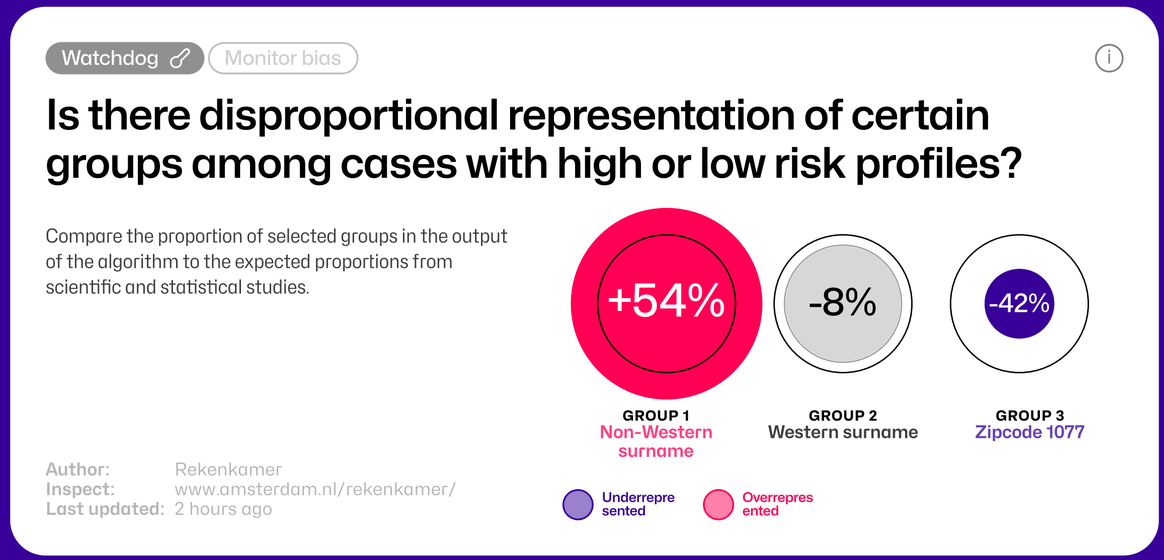
The watchdog runs alongside the algorithm

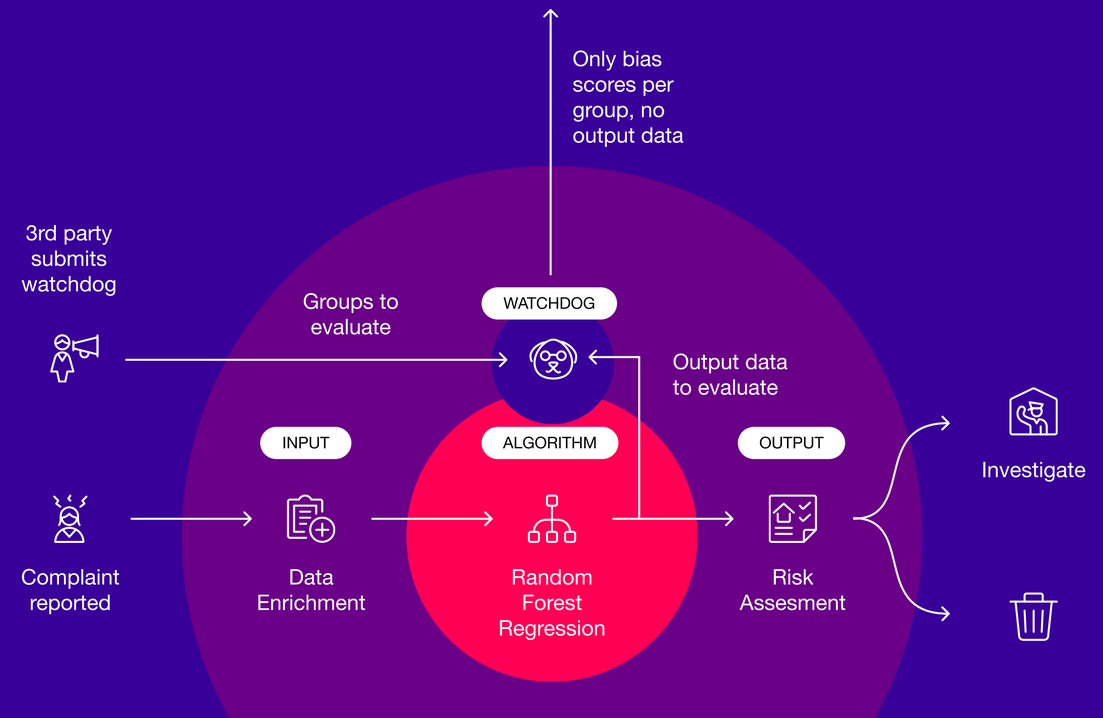
Definitions of groups are created by the 3rd party

The watchdog monitors the actual output

The watchdog is creates bias scores per group

Only the bias scores are available in the widget





widget

Verify operation

Audit Twin

See if the algorithm runs without alterations by comparing the output of a test server with the results of an audit. If the results are similar, the algorithm is likely doing what it is supposed to do.

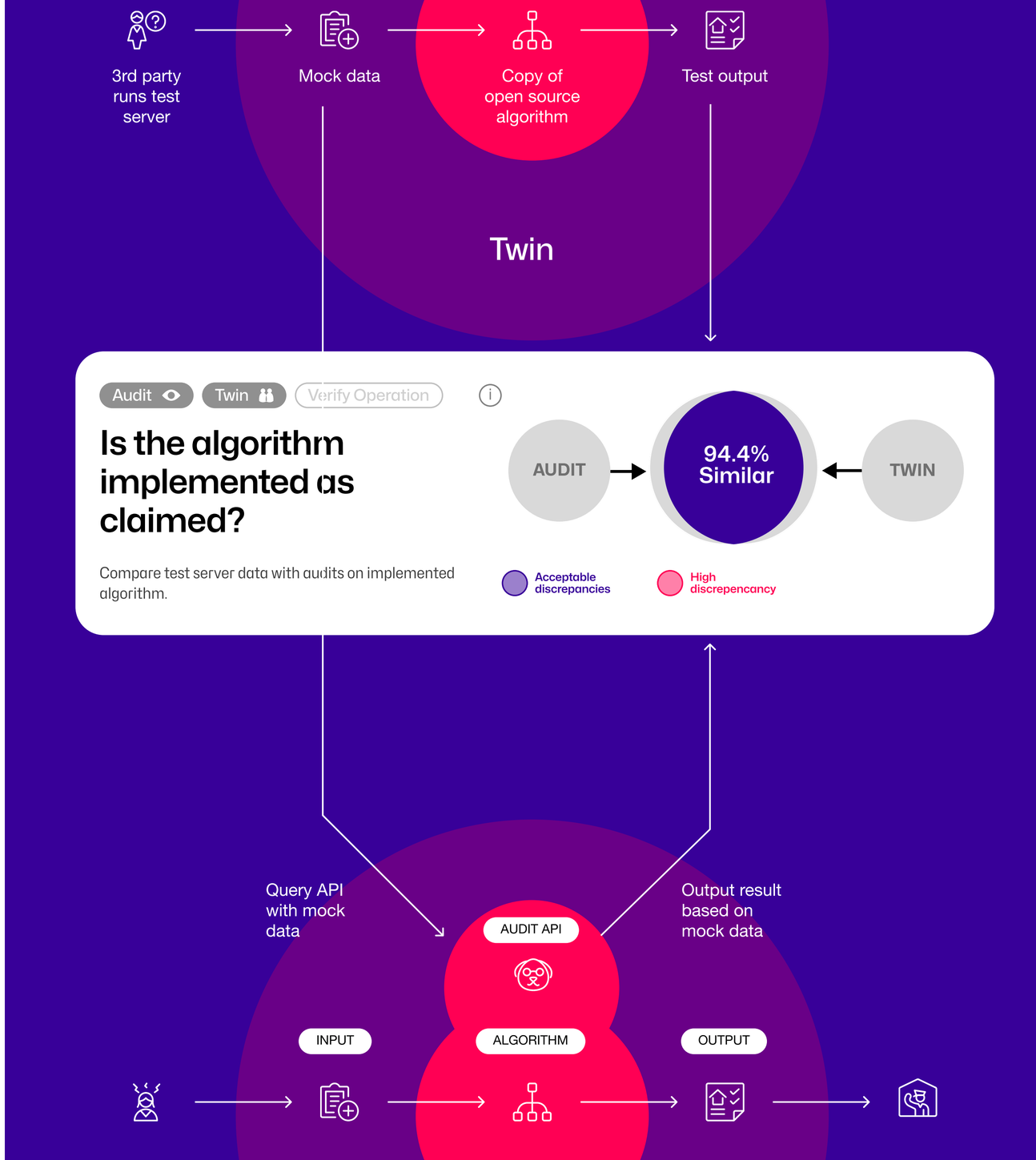
The algorithm is made open source

A 3rd party runs a test server

Mock data is sent to the audit API and test server

The audit API returns results based on mock data

The widget compares that with the test server output



widget

Data minimisation optimization

Watchdog

Identify data that is not used by the algorithm that can be removed from the data input. This would help with further data minimisation.

A 3rd party submits a watchdog

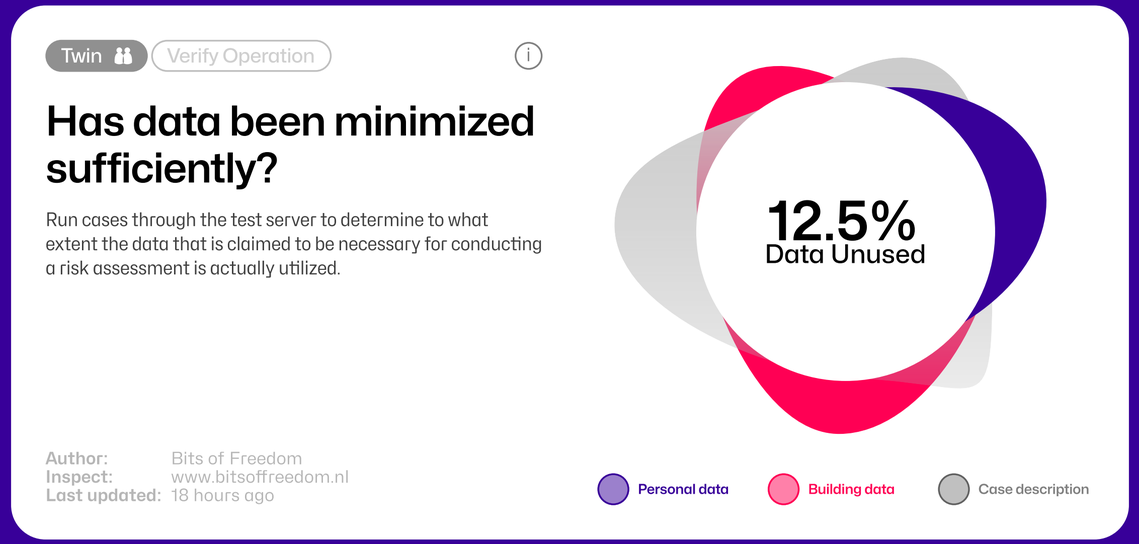
The watchdog runs alongside the algorithm

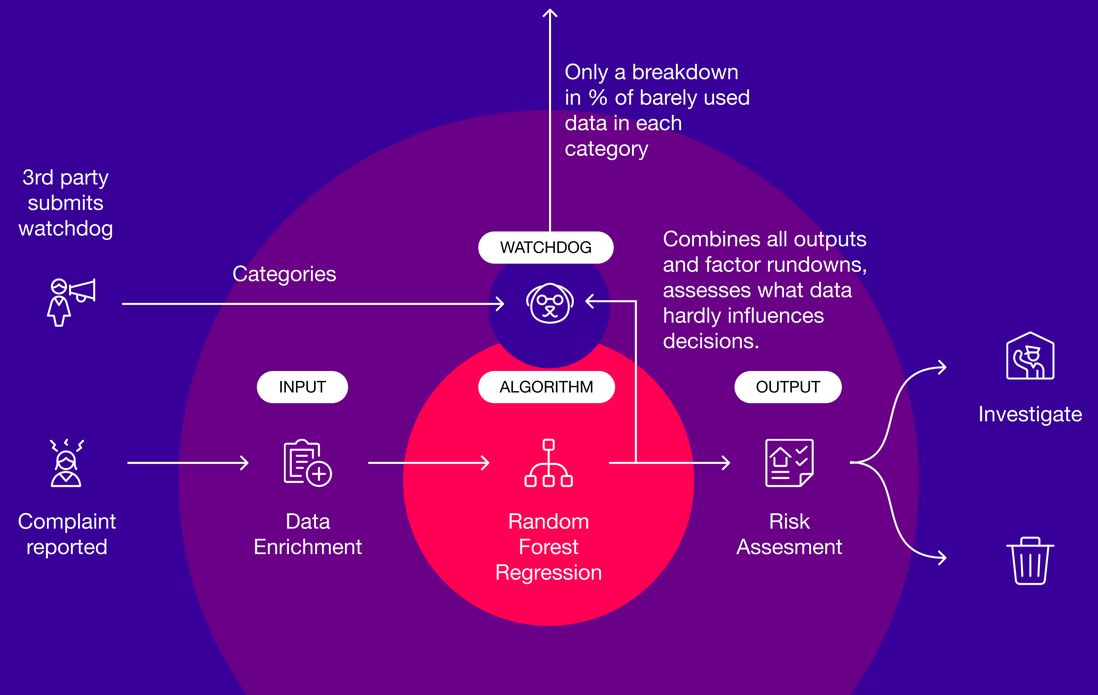
with SHAP, it sees the impact of all factors involved

The watchdog combines all the SHAP scores

See which datatypes hardly ever influence the decision

Only a breakdown of unused data is shown in widget





Algoritmeregister

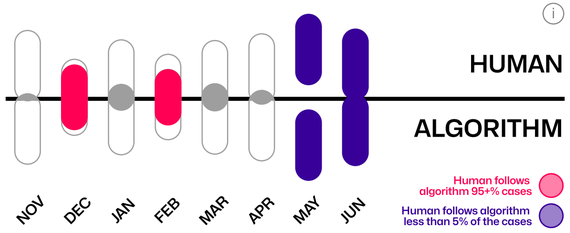
http://algoritmeregister.amsterdam.nl

Dashboard Illegal Holiday Rental Algorithm

Watchdog

Algorithmic influence on human decisions

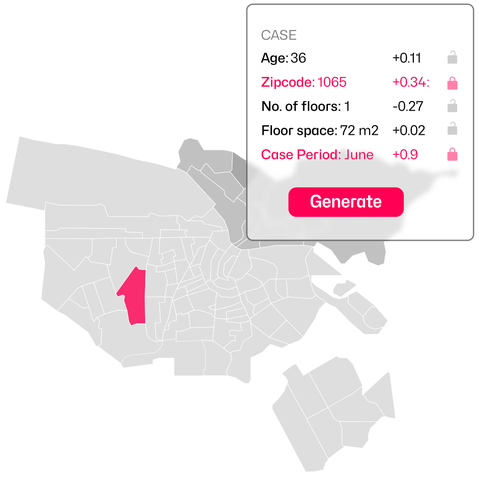
Gain Insight into Algorithmic Influence on Human Decision Making by Comparing the Algorithm's priority list and the human made list that was used for further investigations,



Twin | Discover bias

Generate a high risk profile

Discover hidden biases in the algorithm by generating random high-risk profiles.

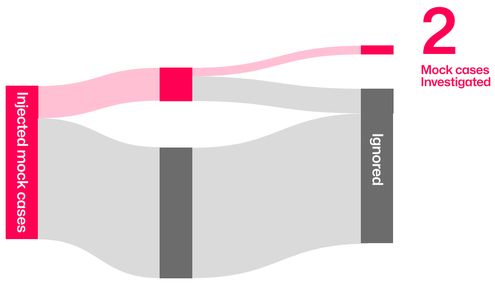


Author: Rekenkamer
Inspect: www.amsterdam.nl/rekenkamer/
Last updated: 2 hours ago

Audit

Is the algorithms trusted blindly?

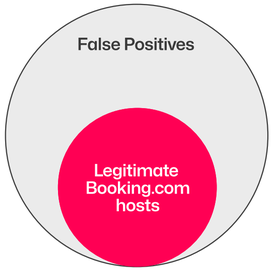
Every month, 20 obvious mock cases are reported to the live algorithm. This widget reports if any of them have made it through the selection process to be investigated.



Author: Rekenkamer
Inspect: www.amsterdam.nl/rekenkamer/
Last updated: 2 hours ago

Legitimate Hosts Under Scrutiny

Shows the share of legitimate hosts that have been investigated, out of all false positive assessments.



Author: Booking.com
Inspect: www.amsterdam.nl/rekenkamer/
Last updated: 2 hours ago

Algoritmeregister

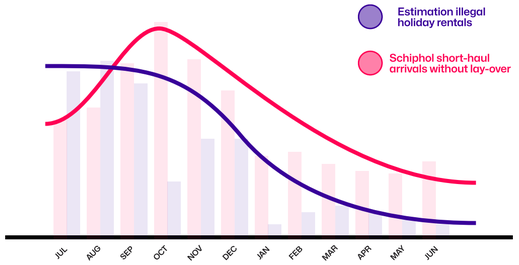
http://algoritmeregister.amsterdam.nl

Dashboard Illegal Holiday Rental Algorithm

Verify Operation

Illegal Holiday Rental Regulation Impact on Schiphol Traffic

Assessing the impact of policies aimed at reducing illegal holiday rentals on short-stay traffic in tourism by



Author: Urgenda
Inspect: www.urgenda.nl
Last updated: 11 Days ago

Chat with the Algorithm

I have a Question!

What do you think about people living in Geuzenveld?

Sorry, I cannot answer that question. **I don't have any opinion about any particular people.**

Describe an average high probability neighborhood

Middle rise housing complexes built in the 90s in the west.

That sounds like Geuzenveld_

Find out what Amsterdammers think about the algorithm

Through these 100 sims, an unfiltered perspective of Amsterdam's population is given, which reflects their concerns regarding the algorithm.

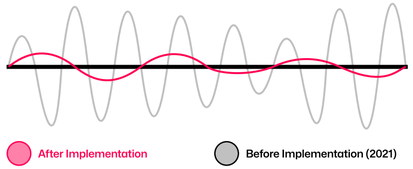
Author: SP
Inspect: www.amsterdam.sp.nl
Last updated: 11 days ago

Algorithm's effect on noise pollution

Evaluate the algorithm's effect on public space nuisance by comparing noise pollution levels in areas before and after implementation.

Author: Voor de Stad
Inspect: www.voordestad.amsterdam
Last updated: 1 day ago

Area: Middenmeer

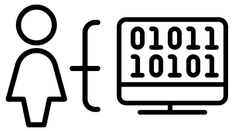


After Implementation

Before Implementation (2021)

conclusion

It is possible to monitor Algorithms while safeguarding sensitive data!



It requires thorough involvement and collaboration between government parties and algorithm programmers

An extended commitment to privacy is necessary

Standards are needed to realise models for scrutiny

... let’s discuss!