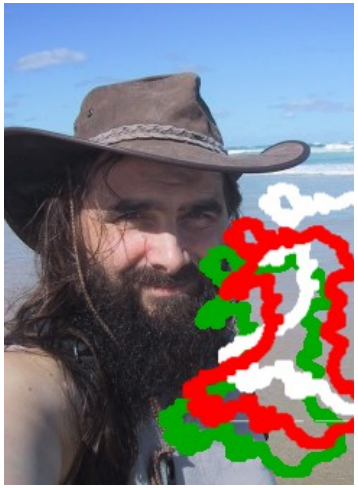
sources of bias and explanation

Alan Dix

Computational Foundry

Swansea



http://alandix.com/academic/talks/PIT-2019-bias-and-explanation/

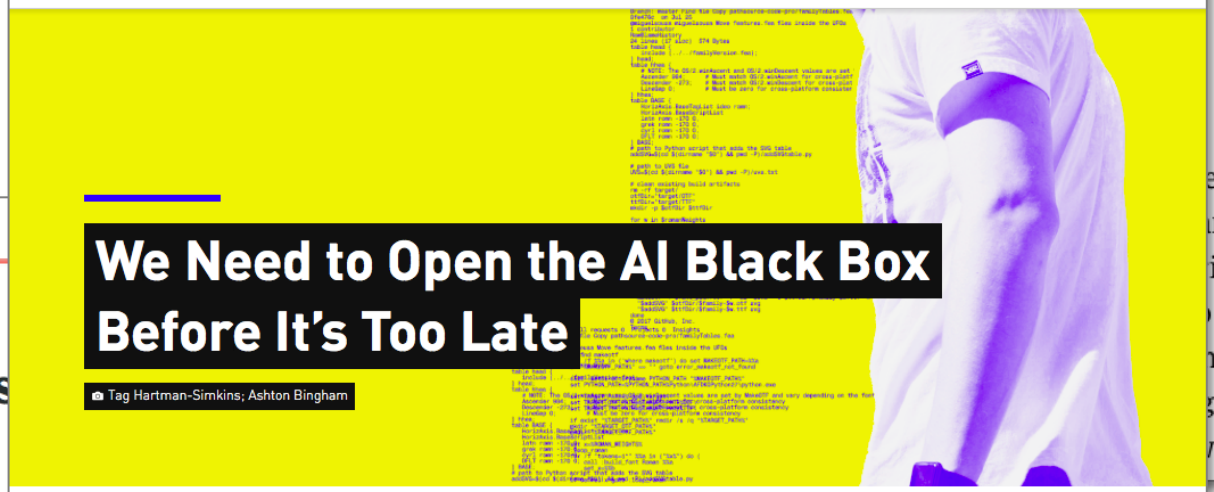
WIRED

BRIAN BARRETT GEAR 05.23.15 07:00 AM

GOOGLE MAPS IS RACIST BECAUSE THE INTERNET IS

es of
rching
ising
an
his to
gn. But
vas

Futurism



We Need to Open the AI Black Box Before It's Too Late

Tag Hartman-Simkins; Ashton Bingham

by Bahar Gholipour | January 18, 2018 | Future Society 16251

TIME

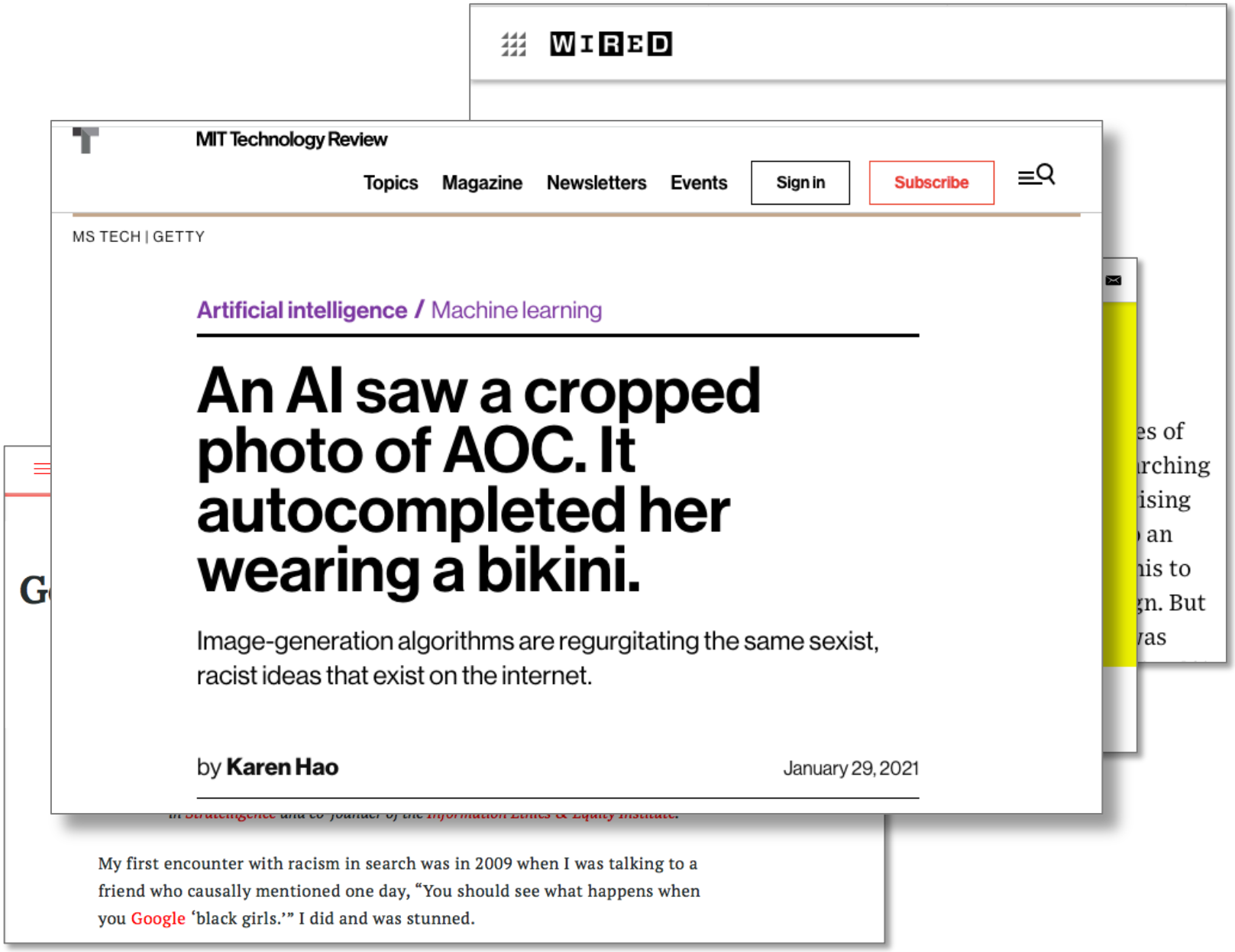
Google Has

By SAFIYA NOBLE M

IDEAS

Dr. Safiya U. Noble is the author of Algorithms of Oppression: How Search Engines Reinforce Racism and is an assistant professor of communication at the University of Southern California, Annenberg School of Communication & Journalism. She is a partner in Stratelligence and co-founder of the Information Ethics & Equity Institute.

My first encounter with racism in search was in 2009 when I was talking to a friend who causally mentioned one day, "You should see what happens when you Google 'black girls.'" I did and was stunned.


WIRED
MIT Technology Review
Topics
Magazine
Newsletters
Events
Sign in
Subscribe
MS TECH | GETTY
Artificial intelligence / Machine learning
An AI saw a cropped photo of AOC. It autocompleted her wearing a bikini.
Image-generation algorithms are regurgitating the same sexist, racist ideas that exist on the internet.
by Karen Hao
January 29, 2021
My first encounter with racism in search was in 2009 when I was talking to a friend who causally mentioned one day, "You should see what happens when you Google 'black girls.'" I did and was stunned.

types of algorithms …

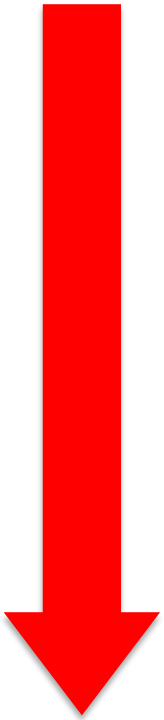
rules and regulations

ordinary code

classic AI

machine learning and neural nets

increasing opacity



when things go wrong – deliberate

misuse

- hacking

bad use

- cyberwarfare – Stuxnet, etc.
- autonomous weapons

when things go wrong – well meaning

accidents

autonomous car crashes

need for transparency

unintended consequences

bias (gender, ethnicity)

disproportionate social effects

https://www.bbc.co.uk/sounds/play/m00017s4 (report @ 1:41:00 in)

25 years back …

Human Issues in the use of Pattern Recognition Techniques

Alan Dix *

October 1991

1 Introduction

The purpose of this chapter is to emphasise that when including neural nets or similar techniques in systems with a human component, the technological issues are far easier to address than the attendant human ones. It highlights the need for a thorough theoretical understanding of the behaviour of the computer-based techniques in order to be able to assess the human consequences of their use.

The chapter focuses on two applications of pattern recognition. One is an innovative example based method of query construction and the other is the more established use of neural nets for routine decision making such as credit vetting.

In the latter example the 'user' of the system is seen as not just the operative who directly uses the computer, but also the client who is the target of the process. This wide view of human-computer interaction means we have to deal not 'just' with the usability of systems but also the entailing ethical and legal responsibilities.

Range of systems covered

This chapter concerns the use of example based or taught pattern recognition techniques. This includes most neural net or connectionist approaches and also inductive learning. These techniques all operate by being given a set of examples and from them generalising to unseen data. They are essentially

*work funded by SERC Advanced Fellowship B/89/ITA/220

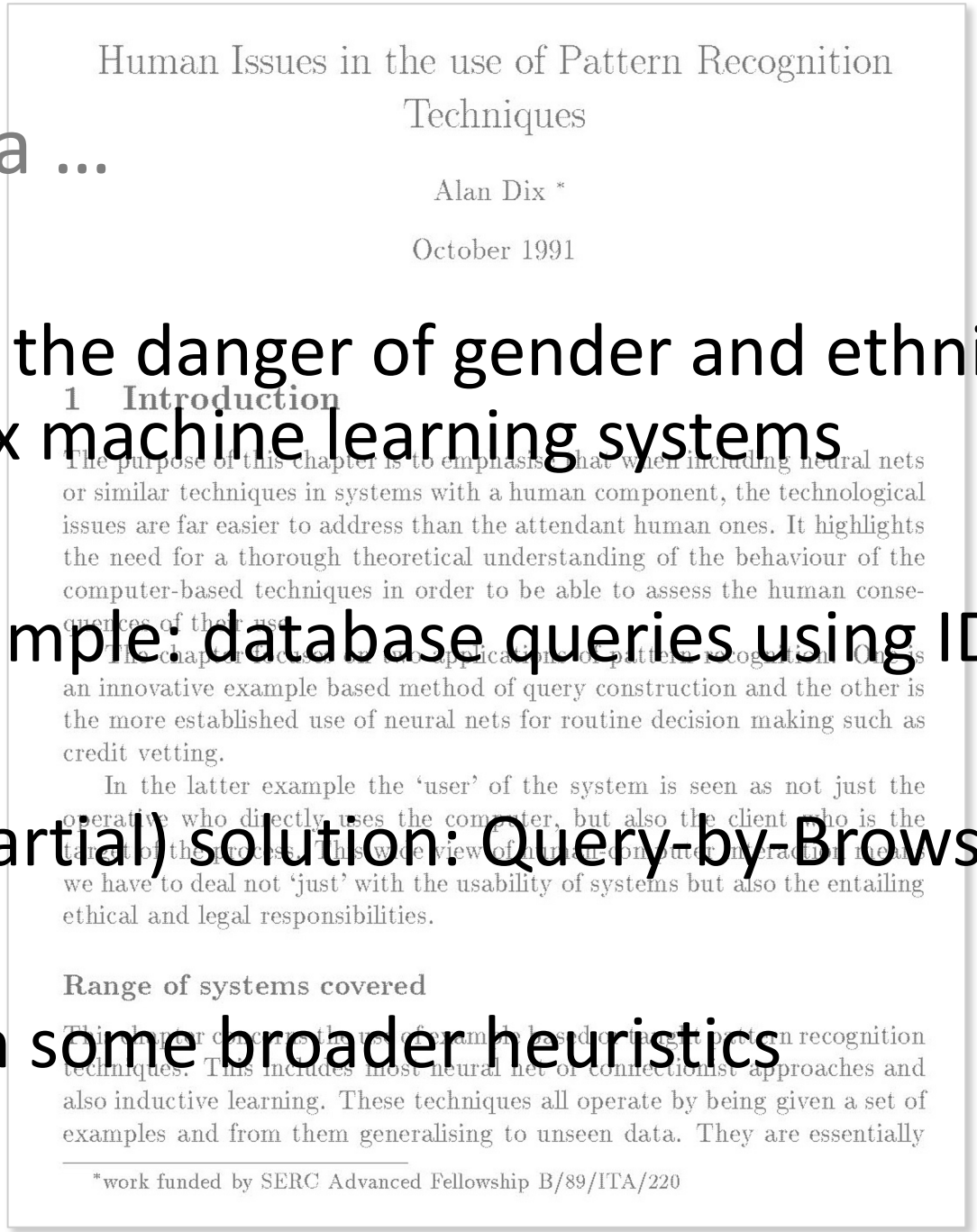
inter alia …

warns of the danger of gender and ethnic bias in black-box machine learning systems

gives example: database queries using ID3

offers (partial) solution: Query-by-Browsing

and even some broader heuristics



Human Issues in the use of Pattern Recognition Techniques

Alan Dix *

October 1991

1 Introduction

The purpose of this chapter is to emphasise that when including neural nets or similar techniques in systems with a human component, the technological issues are far easier to address than the attendant human ones. It highlights the need for a thorough theoretical understanding of the behaviour of the computer-based techniques in order to be able to assess the human consequences of their use.

The chapter focuses on two applications of pattern recognition. One is an innovative example based method of query construction and the other is the more established use of neural nets for routine decision making such as credit vetting.

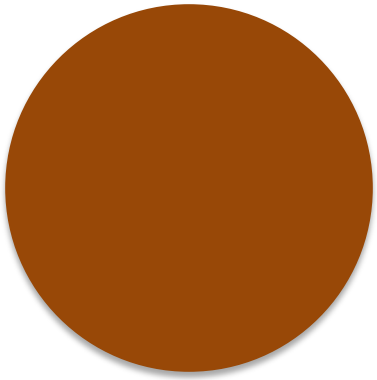
In the latter example the 'user' of the system is seen as not just the operative who directly uses the computer, but also the client who is the target of the process. This wide view of human-computer interaction means we have to deal not 'just' with the usability of systems but also the entailing ethical and legal responsibilities.

Range of systems covered

This chapter concerns the use of example based pattern recognition techniques. This includes most neural net or connectionist approaches and also inductive learning. These techniques all operate by being given a set of examples and from them generalising to unseen data. They are essentially

*work funded by SERC Advanced Fellowship B/89/ITA/220

yes, 25 years ago!



Query-by-Browsing

creating scructable
internal representations

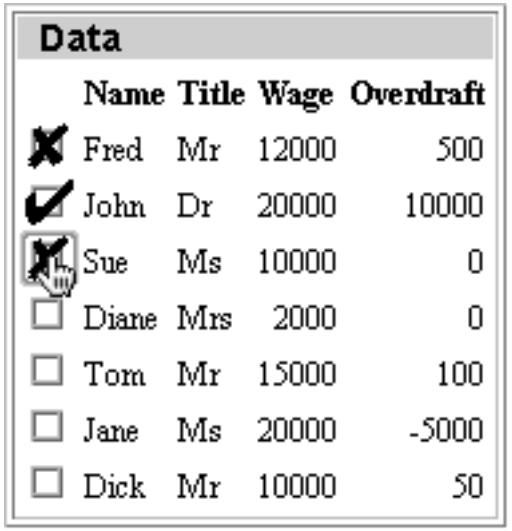
Query by Browsing

user chooses records of interest

- ✓ tick for those wanted
- ✗ cross for those not wanted

system infers query

- web version uses rule induction
- variant of Quinlan's ID3



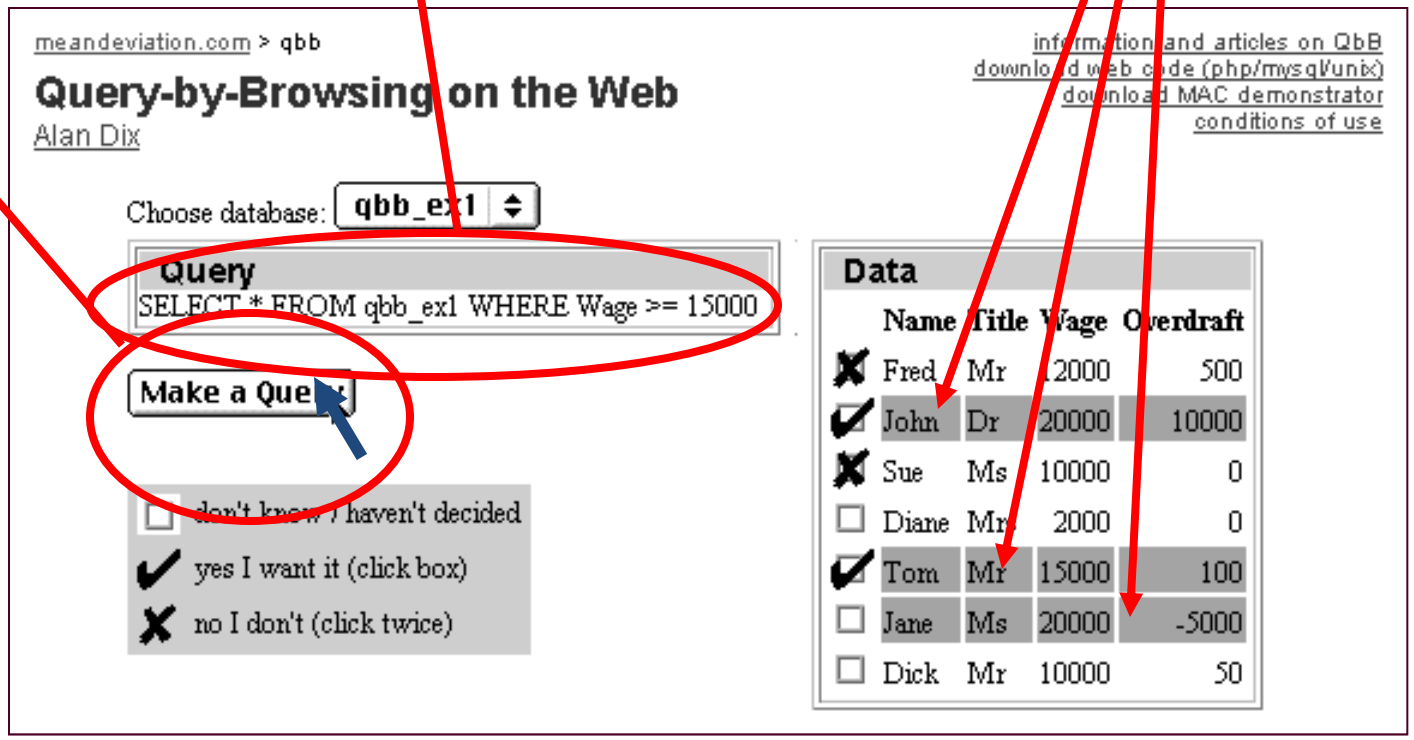
www.meandeviation.com/qbb

Query by Browsing
what it looks like

user asks system to make a query

system infers SQL query

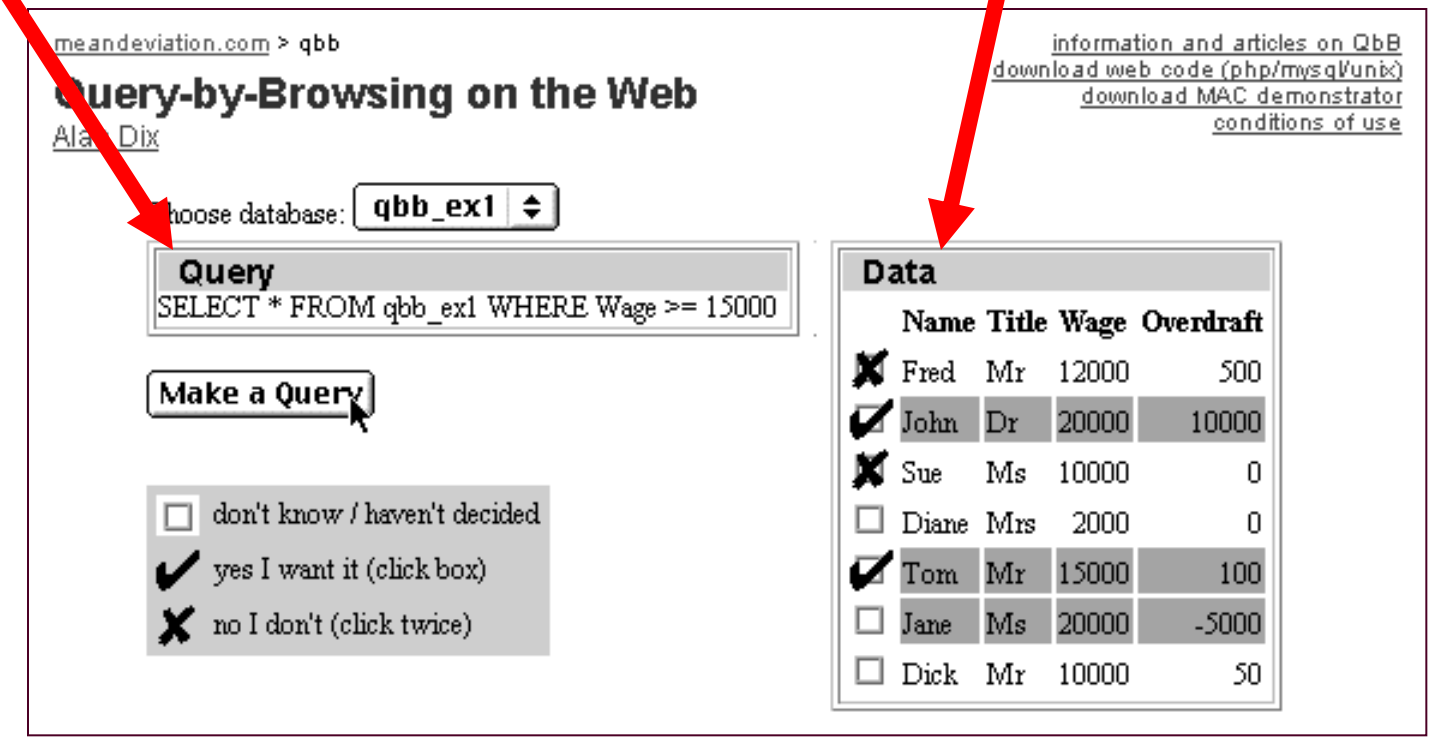
query results highlighted



Query by Browsing
dual representation

query (intensional)
for precision

listing (extensional)
for understanding



Query by Browsing – how it works

examples

Data

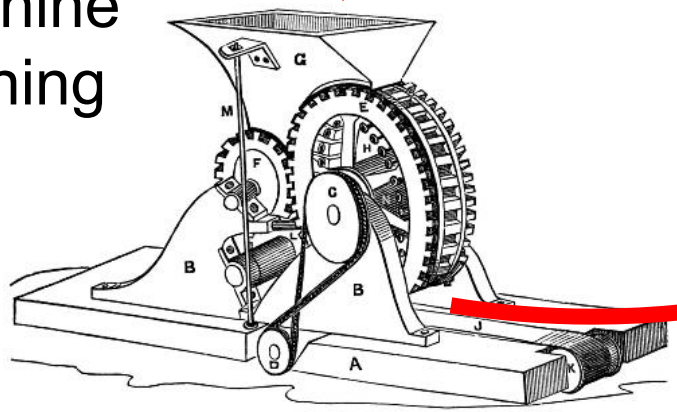
	Name	Title	Wage	Overdraft
✗	Fred	Mr	12000	500
✓	John	Dr	20000	10000
✗	Sue	Ms	10000	0
☐	Diane	Mrs	2000	0
✓	Tom	Mr	15000	100
☐	Jane	Ms	20000	-5000
☐	Dick	Mr	10000	50

SQL query

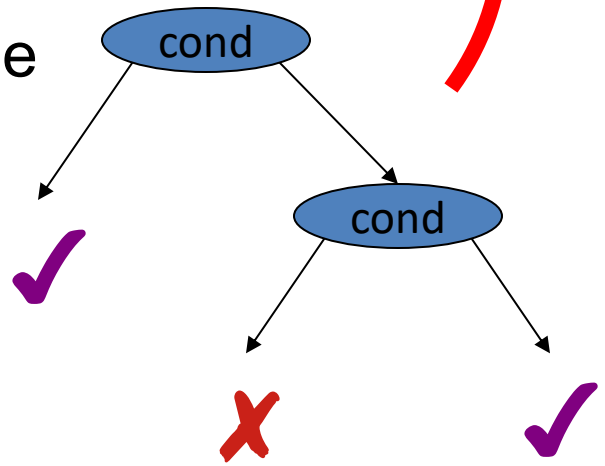
Query

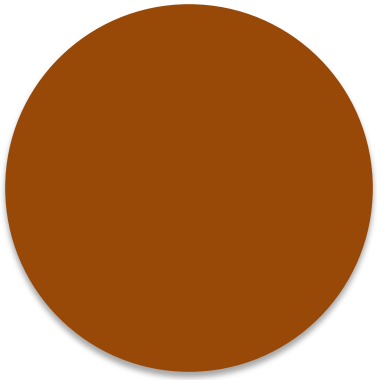
```
SELECT * FROM qbb_ex1 WHERE Wage >= 15000
```

machine learning



decision tree

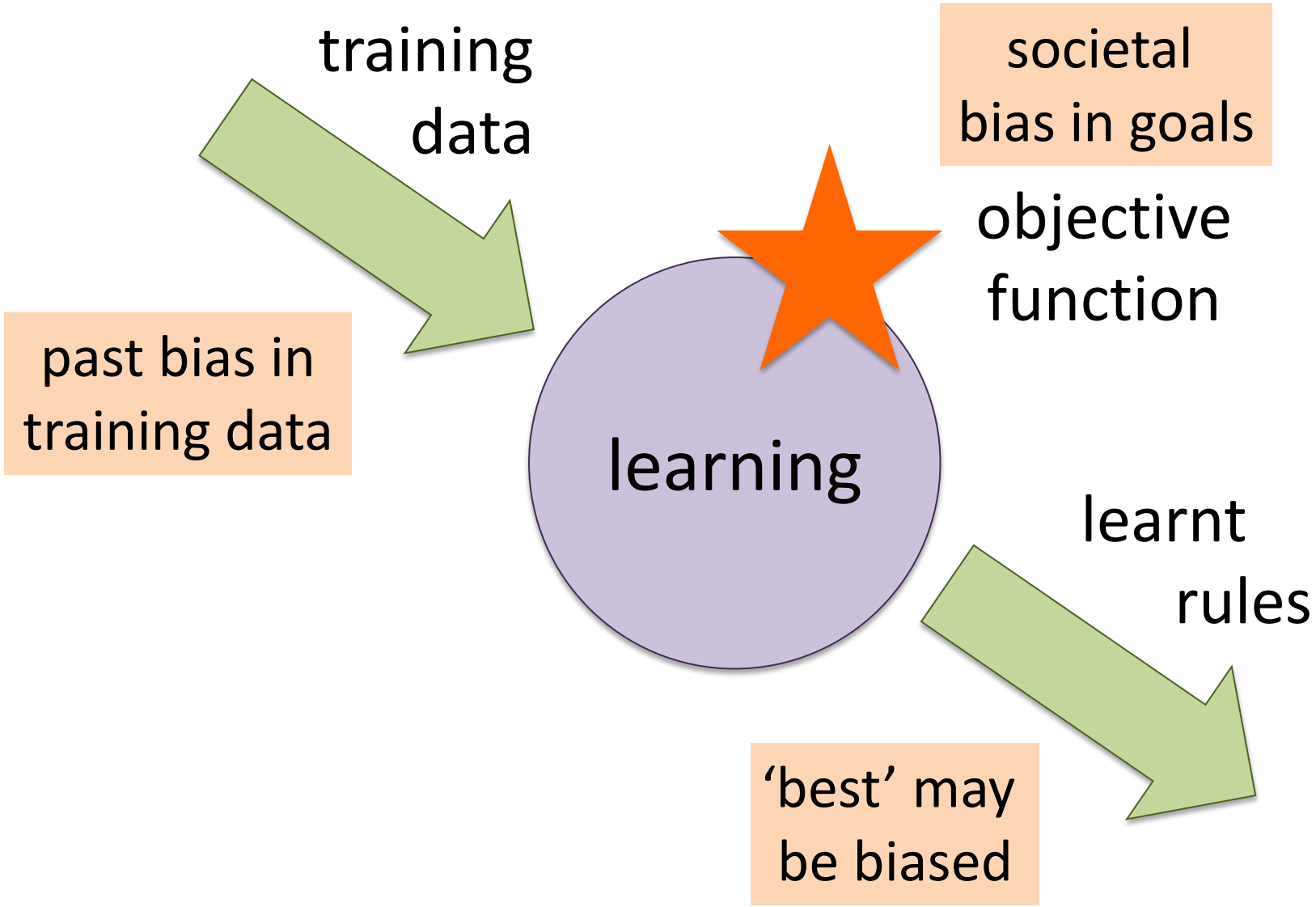


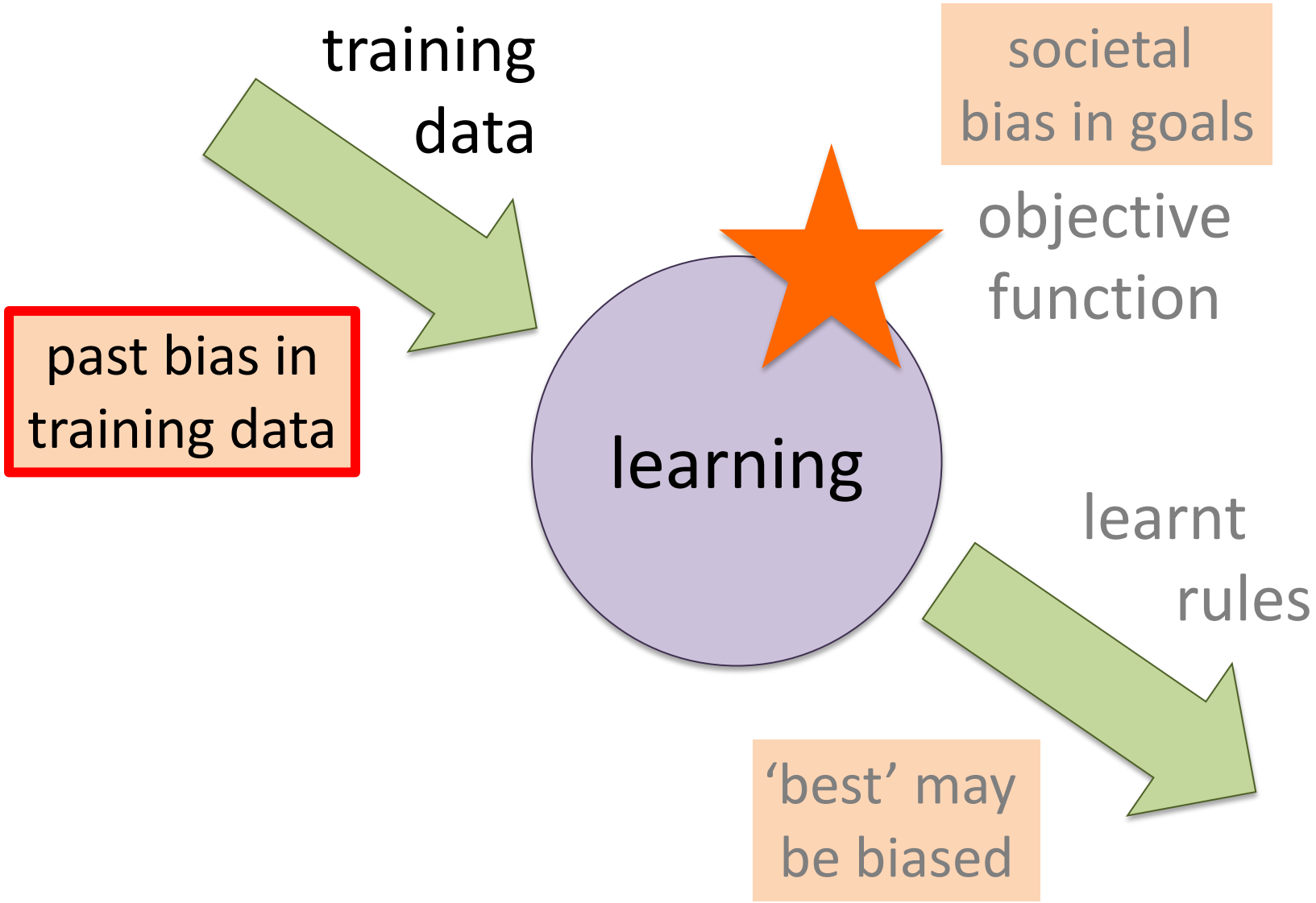


it is not just about being accurate

not just right

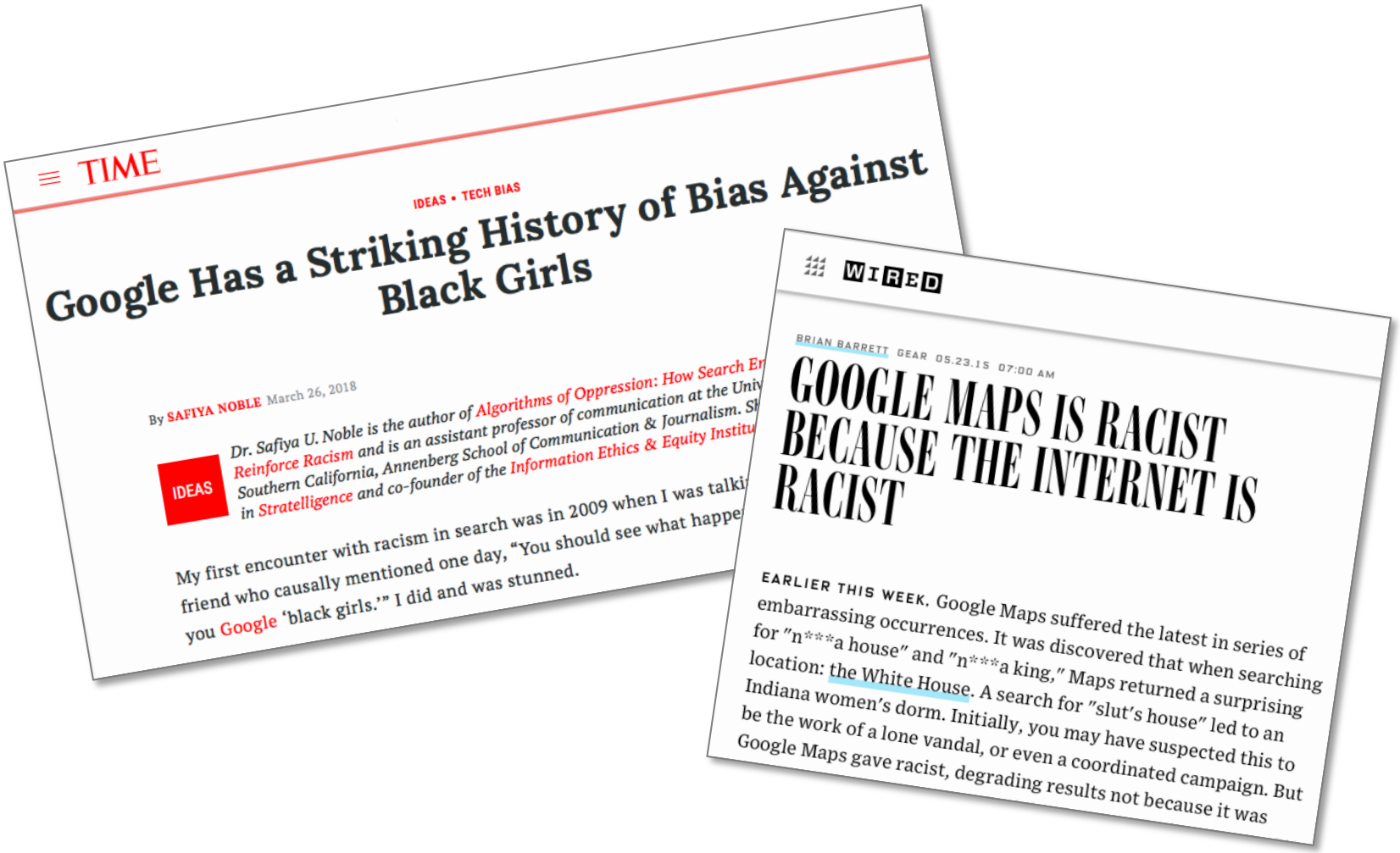
but also upright


training data
societal bias in goals
objective function
past bias in training data
learning
learnt rules
'best' may be biased


training data
societal bias in goals
objective function
past bias in training data
learning
learnt rules
'best' may be biased

algorithms reflect society

mimicking human behaviour and choices



TIME

IDEAS • TECH BIAS

Google Has a Striking History of Bias Against Black Girls

By SAFIYA NOBLE March 26, 2018

IDEAS

Dr. Safiya U. Noble is the author of Algorithms of Oppression: How Search Engines Reinforce Racism and is an assistant professor of communication at the University of Southern California, Annenberg School of Communication & Journalism. She is in Stratelligence and co-founder of the Information Ethics & Equity Institute

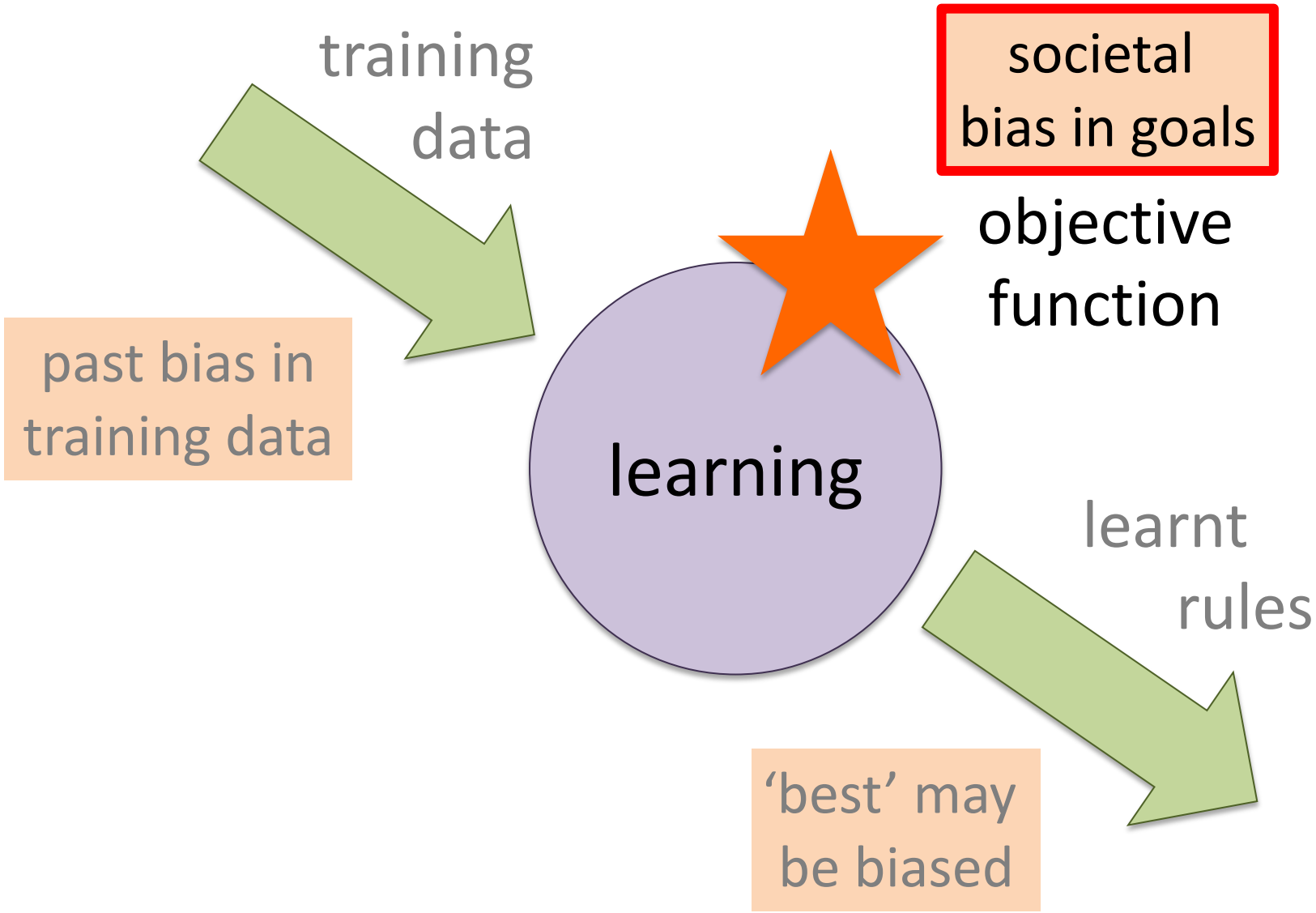
My first encounter with racism in search was in 2009 when I was talking to a friend who causally mentioned one day, "You should see what happens when you Google 'black girls.'" I did and was stunned.

WIRED

BRIAN BARRETT GEAR 05.23.15 07:00 AM

GOOGLE MAPS IS RACIST BECAUSE THE INTERNET IS RACIST

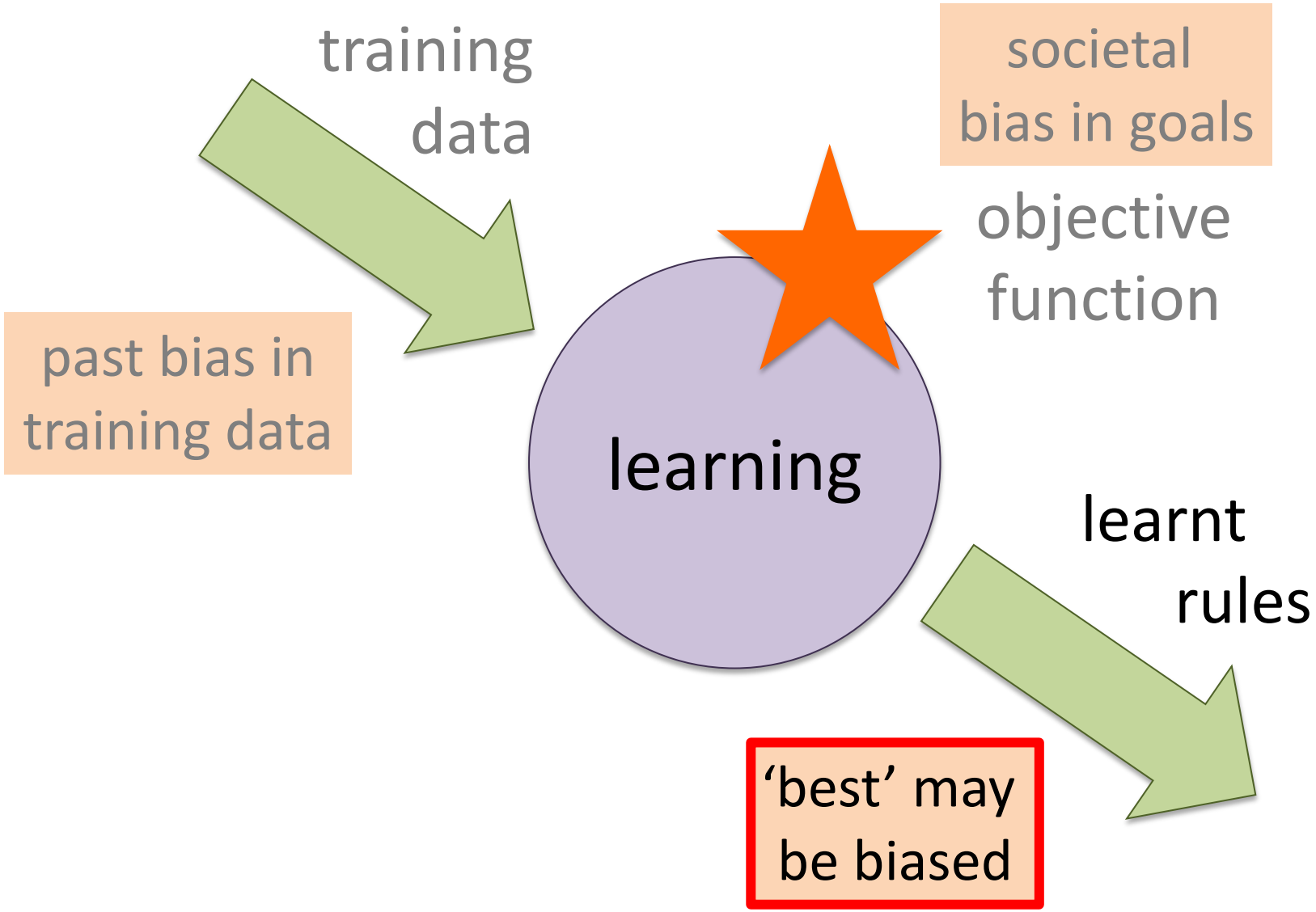
EARLIER THIS WEEK. Google Maps suffered the latest in series of embarrassing occurrences. It was discovered that when searching for "n***a house" and "n***a king," Maps returned a surprising location: the White House. A search for "slut's house" led to an Indiana women's dorm. Initially, you may have suspected this to be the work of a lone vandal, or even a coordinated campaign. But Google Maps gave racist, degrading results not because it was


training data
societal bias in goals
objective function
past bias in training data
learning
learnt rules
'best' may be biased

pandering to human bias (effective outcomes?)

- dating sites using ethnicity (CHI 2018!)
- young pretty waitresses sell more drinks
- Trump (reportedly) hiding black employees at casino when certain rich customers arrived
- BBC (& others) paying male presenters more because they are more popular

‘good’ business
but is it good?


training data
societal bias in goals
objective function
past bias in training data
learning
learnt rules
'best' may be biased

accurate does not mean right

THE TIMES
TUESDAY OCTOBER 15 2019
Log in
Subscribe

It's hysteria, not a heart attack, GP app Babylon tells women

Shanti Das

October 13 2019, 12:01am, The Sunday Times

The weekly magazine for higher education

PhD pipeline: who imports the most foreign talent? 6

No. 2,427
26 September-2 October 2019

THE

State of anxiety
Arrest keeps Hong Kong institutions on edge 9

Strategic questions
International educators reveal priorities 22

Far away, not distant
Remote event attendees are fully engaged 26

Grave business
How Dublin's anatomists

News | Opinion | Sport | Culture | Lifestyle

Automating poverty

Digital dystopia: how algorithms punish the poor

▲ Illustration: Francisco Navas/Guardian Design

In an exclusive global series, the Guardian lays bare the tech revolution transforming the welfare system worldwide – while penalising the most vulnerable

by Ed Pilkington in New York

NEWS

AI in admissions is a 'big concern'

Algorithms may simply lead to 'self-fulfilling prophecies', says ethicist. David Matthews writes

Using artificial intelligence to decide which students to admit or researchers to hire risks creating "self-fulfilling prophecies" that could simply reinforce which kind of people win opportunities, an expert in digital ethics has warned universities.

AI in admissions has recently emerged on universities' agenda: the president of Imperial College London, Alice Gast, said last year that she expected AI would "augment" the process, while several Hong Kong universities have said that they are using the technology to find the student characteristics that predicted future success.

But speaking at a major conference on AI hosted by the University of Oxford on 18 September, Carissa Véliz, a research fellow at Oxford's Uehiro Centre for Practical Ethics, said that she had several worries about the technology being deployed in education.

It is a "big concern" that AI was being used to "assess people, whether it's professors, teachers, students, and to filter candidates", she told *Times Higher Education*.

For example, an AI system might analyse data on researcher career trajectories and find that people who did a PhD at certain universities had more success in the future.

On this basis, the AI might conclude that it made sense to award grants to applicants from those universities over others.

But this risked simply reinforcing existing patterns, Dr Véliz warned.

"We will never know whether the postdoc who did not receive a grant might have become a successful academic," she argued. "As long as predictions are used to allocate resources and opportunities, the risk of self-fulfilling prophecies seems inevitable."

Another concern is that algorithms use proxy data – which could be as arbitrary as where people live, or their Facebook friends – to predict how well someone will fare in the future, she said.

Advocates for using AI in admissions argue that machines are less prone to the cognitive biases that could unfairly sway the decision of a human admissions officer.

But some types of AI, such as those based on so-called neural networks that mimic the human brain, are seen as "black boxes" because it can be unclear why they have made a particular decision.

"When the algorithm recommends someone, or brands someone as risky, we may not know why that is," Dr Véliz said.

Humans have to justify their decisions with reasons, she pointed out, but with algorithms, "reasons are missing".

Just as with new drugs, there should be randomised controlled trials to see what impact AI systems have on the distribution of opportunities, she argued, before they are "let loose on the world".

Dr Véliz also took aim at universities introducing what in some cases might be "tech for the sake of it".

She asked: "When we introduce tech into universities, are we doing it for the benefit of students, and are the benefits really worth the risk? And what are the alternatives?"

"Sometimes, low tech is surprisingly robust, and cheaper, and safer. If you think about books as a technology, they are incredibly robust, and much more so than any kind of digital tech that is glitchy, and has security issues and so on," Dr Véliz said.

For example, the filming and recording of lectures is a form of "surveillance" that "diminishes creativity and independent thinking", she added. "When I lecture in university classrooms where there are cameras and microphones, there is typically less debate on sensitive issues, for instance. No one likes to be on record exploring tentative ideas."

david.matthews@timeshighereducation.com

reinforcing societal/cultural norms

at school

boys more likely to study STEM subjects

girls more likely to study humanities

so, on average, with no other information

gender is an (albeit poor) predictor

of communication skills

and engineering knowledge

as a society we choose to use other (and better) predictors

innate (but largely irrelevant) differences

men are (on average) larger and stronger

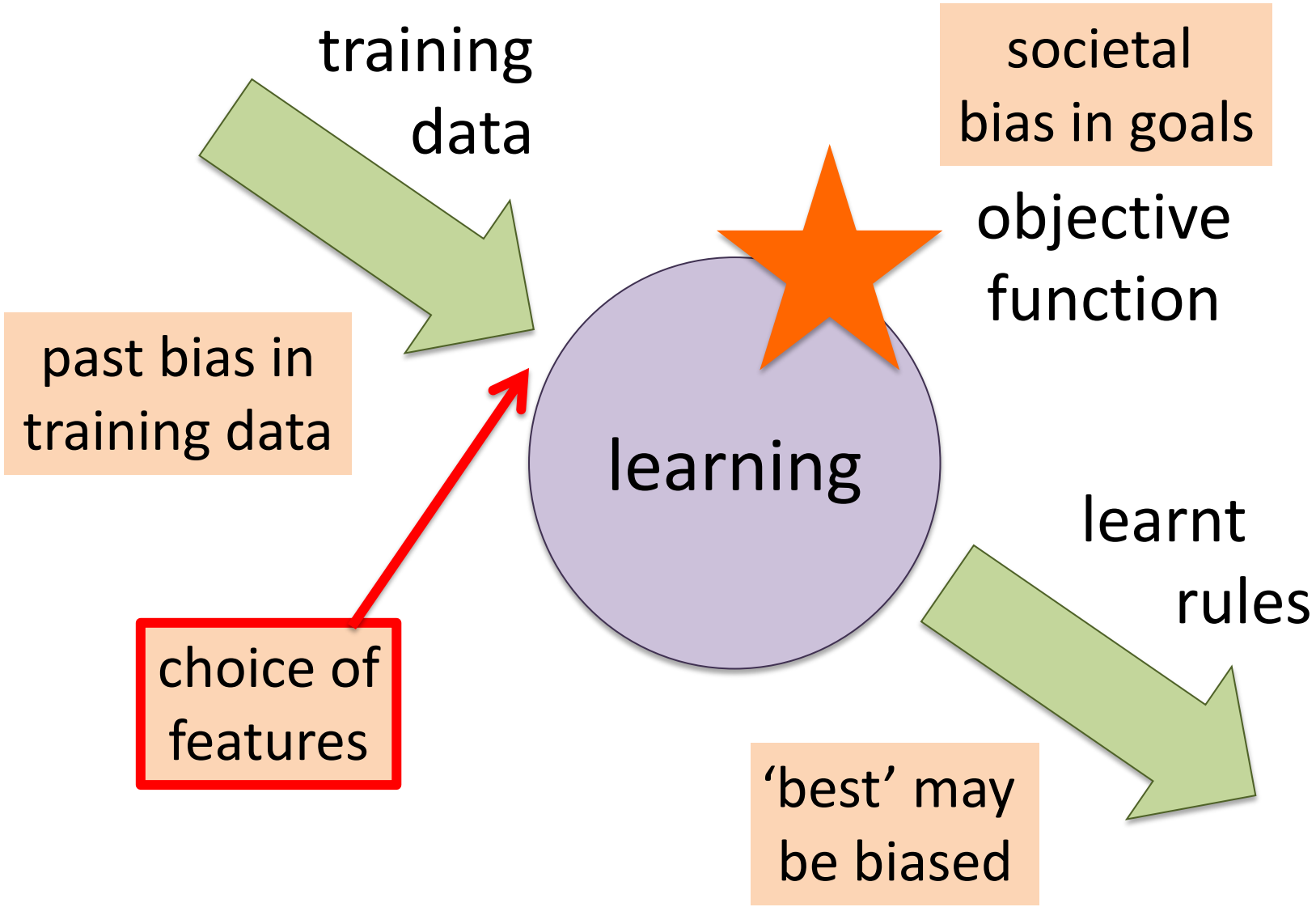
so gender is a Bayesian predictor of strength

this may *explain* gender differences in some jobs

but …

it does NOT justify employment discrimination

bias is not about
algorithmic correctness
it is about social choice


training data
societal bias in goals
objective function
past bias in training data
learning
learnt rules
choice of features
'best' may be biased

the choice of input features often critical in creating or controlling bias

more data not always better!

Note:

human reasoning is
poor at ignoring low quality cues
even when we have better ones

algorithms may be better?

however …

not sufficient to remove explicit indicators:

gender/ethnicity/disability/religion

potential correlating factors e.g. clothing

algorithms need to *actively avoid* discrimination

and how do we know our algorithms are OK?

Not just bias

safety – e.g. autonomous cars

democracy – e.g. social media, fake news

health and well being – e.g. soft-drink adverts

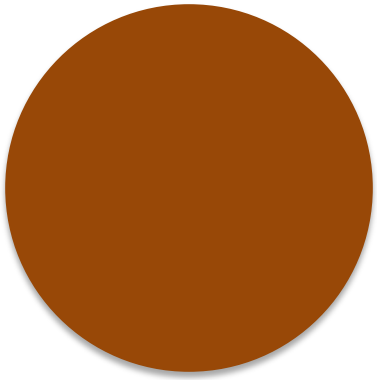
social issues – e.g. credit ratings

we need to ask

Why?

algorithmic transparency

c.f. court judgment



an AIX Kitbag

AI explainability

how to make sense of
black-box machine-learning algorithms

crucial insight …

human–human explanations

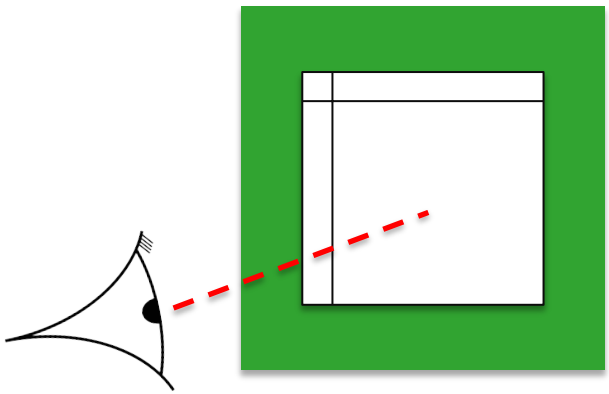
rarely utterly precise or reproducible

but are

sufficient to inspire confidence and trust

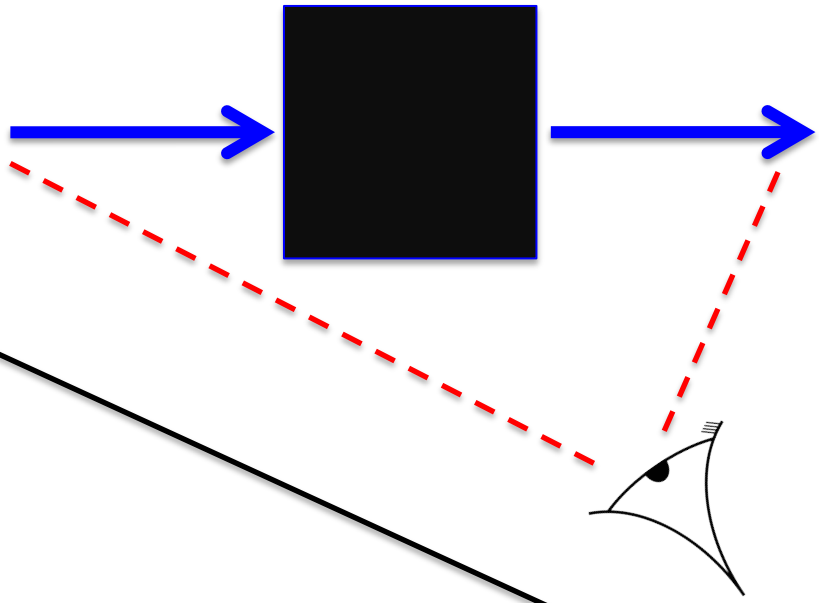
white-box

creating scructable
internal representations



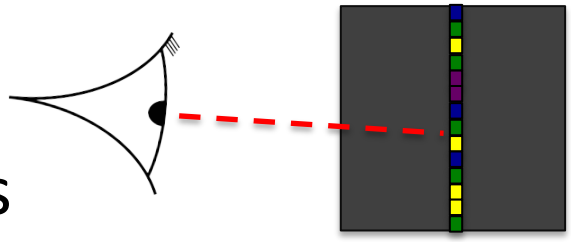
black-box

analysing and
understanding
from the outside



grey-box

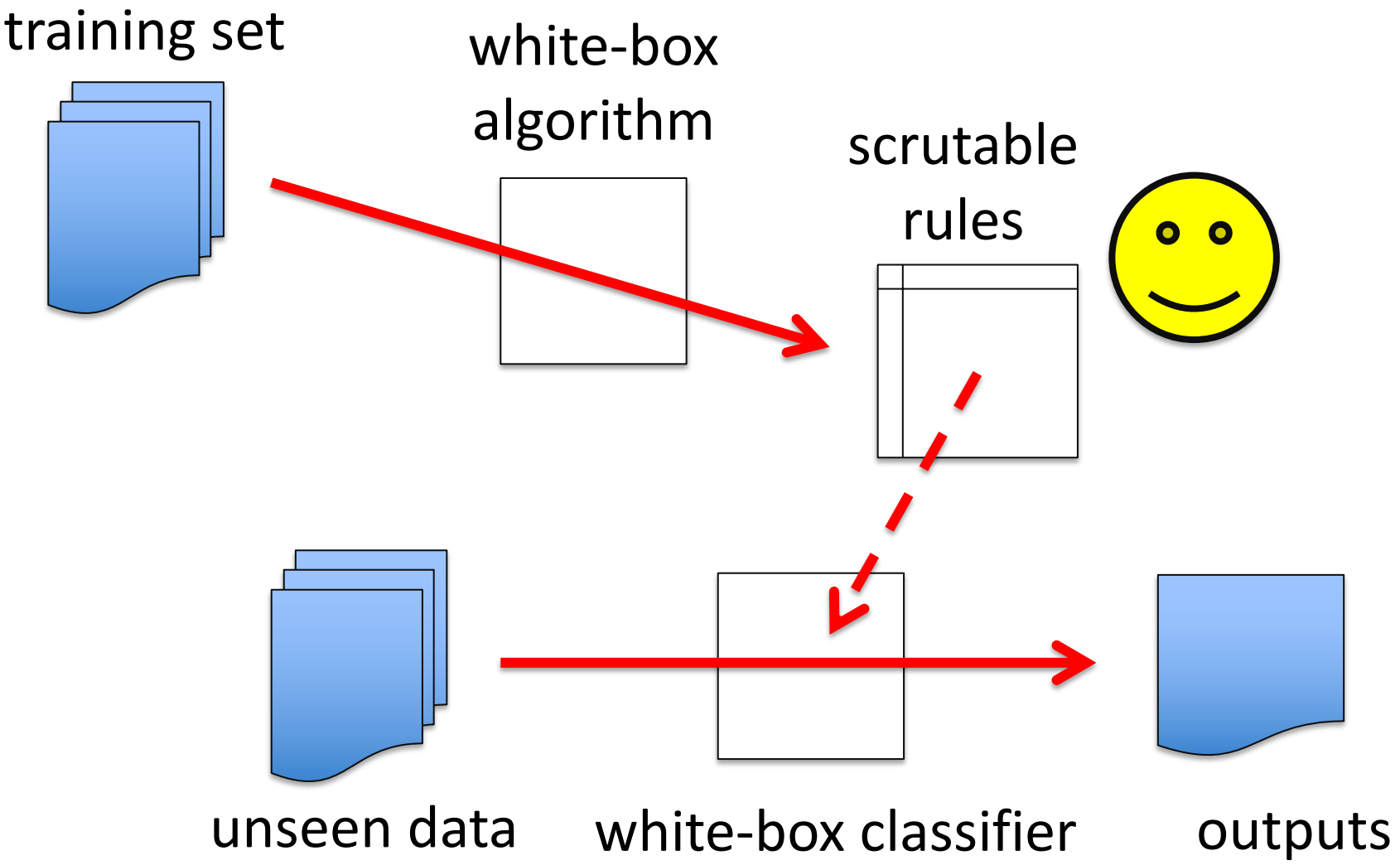
peeking within
understanding
internal representations



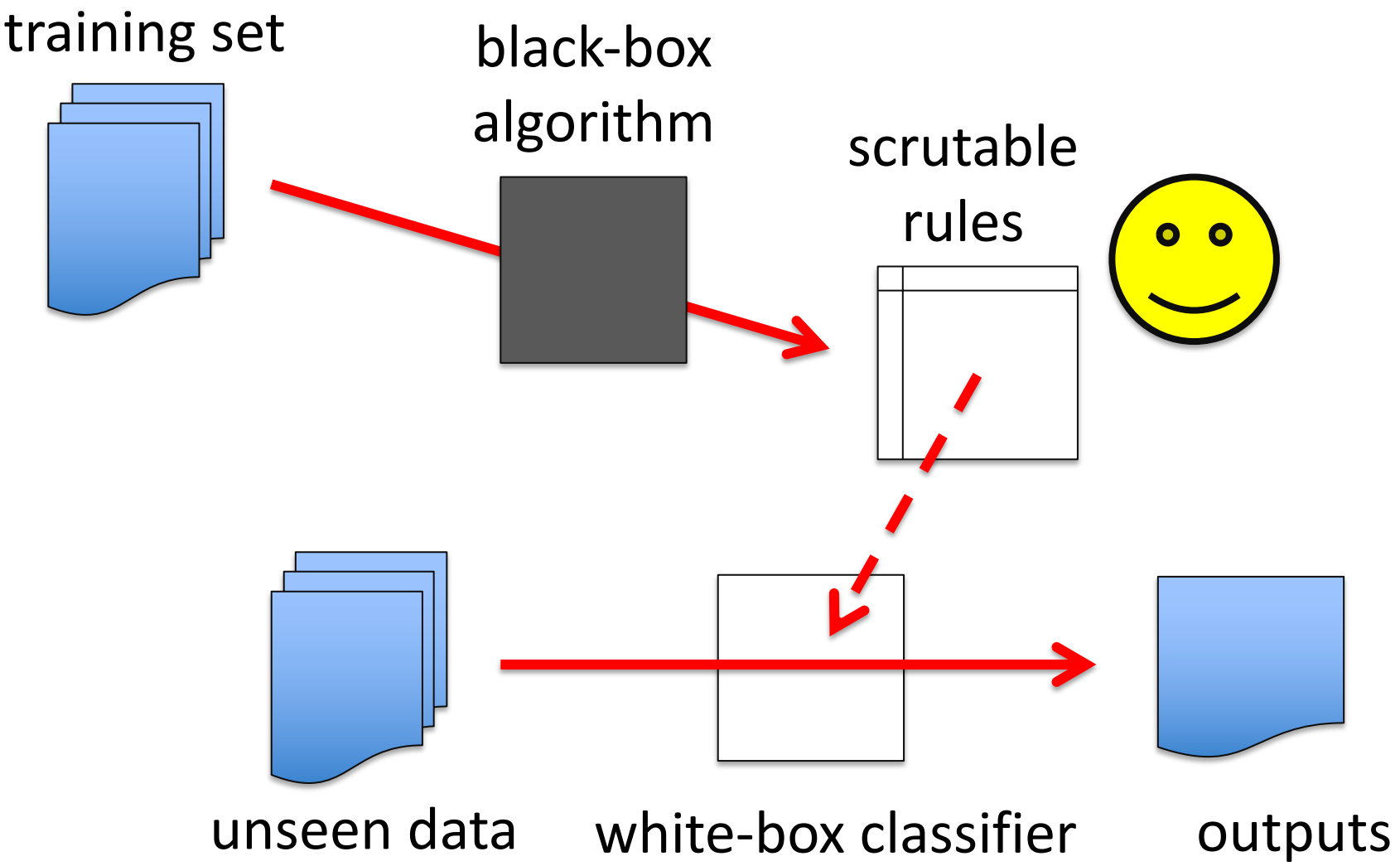
white-box methods

creating scructable
internal representations

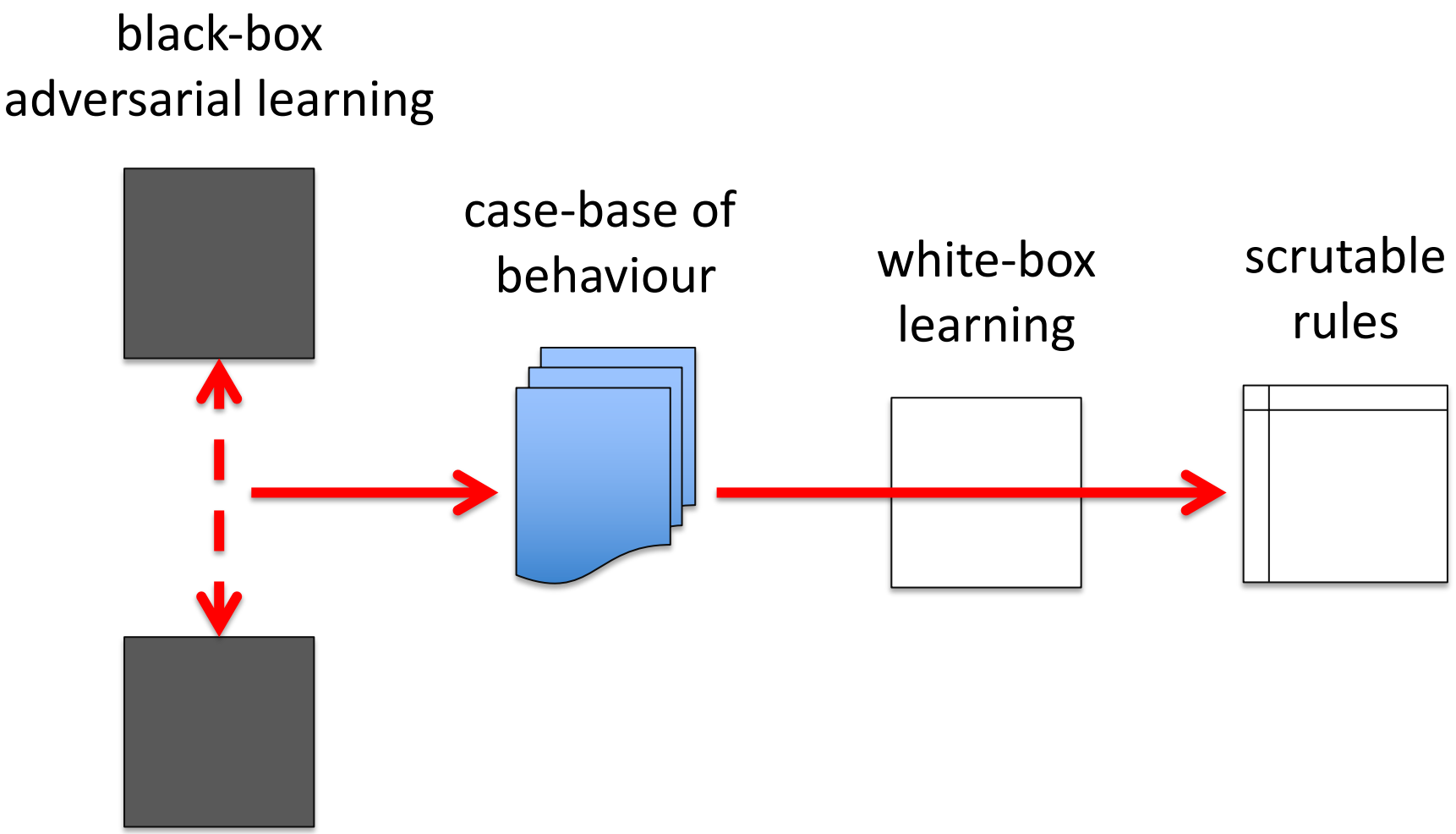
WB0. choose a white box classifier!



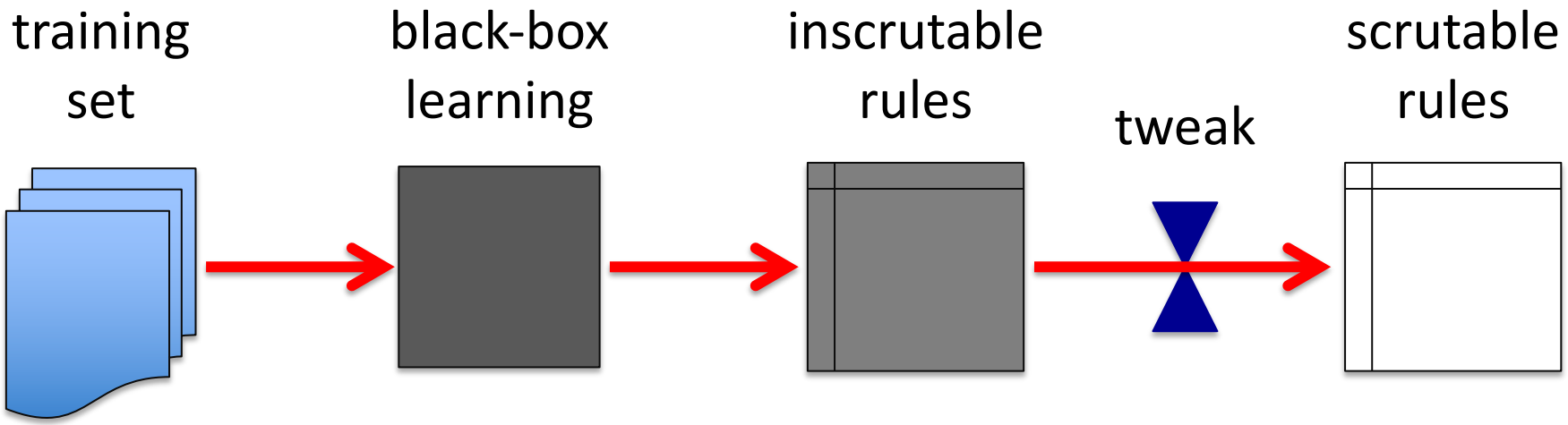
WB1. black-box generation of white box classifier



WB2. Adversarial examples for white-box learning



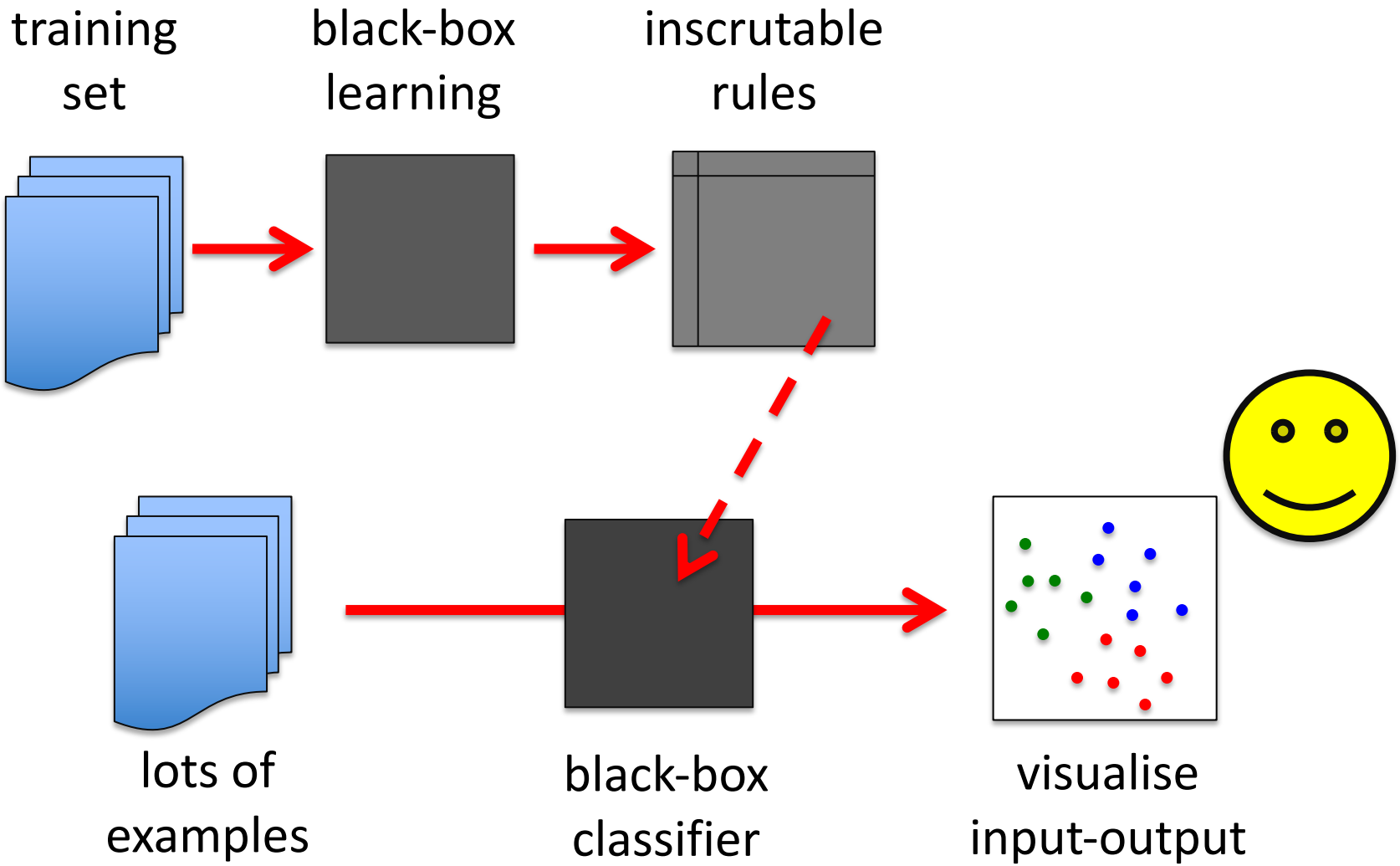
WB3. Simplification of rule set



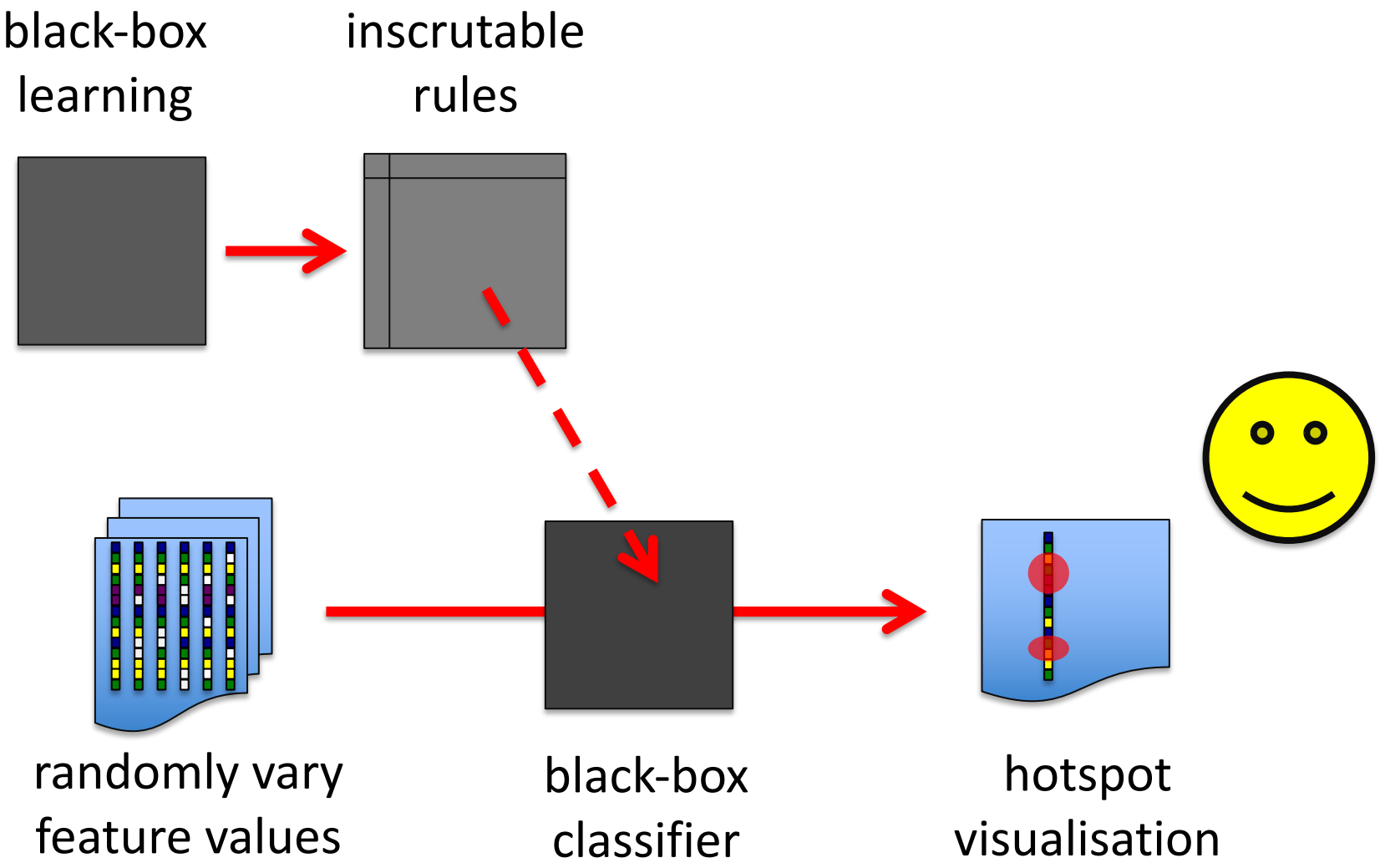
black-box methods

analysing and understanding
from the outside

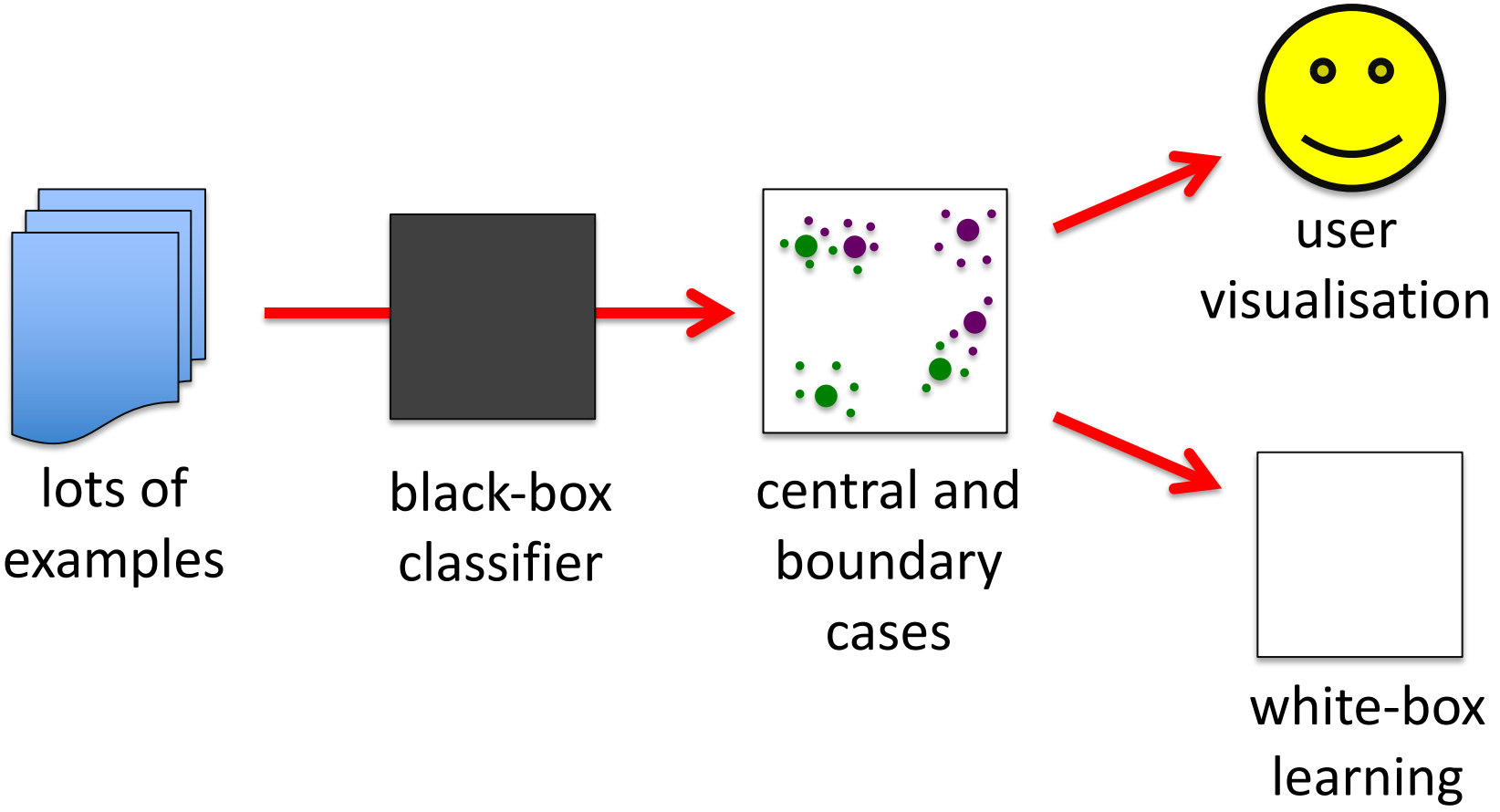
BB1. exploration analysis for human visualisation



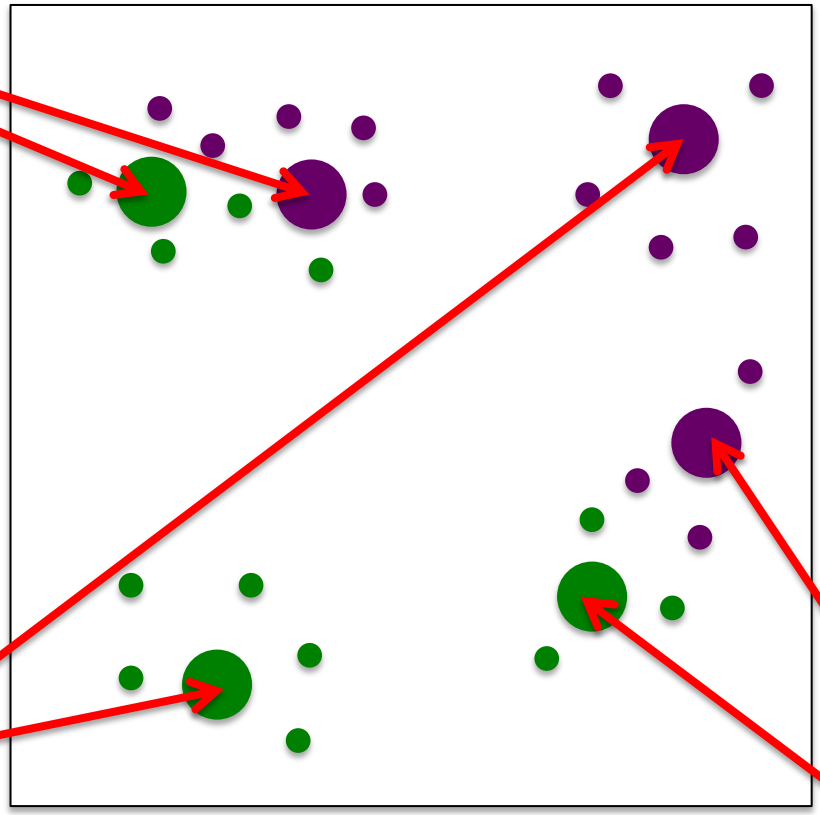
BB2. perturbation/exploration analysis for key feature detection



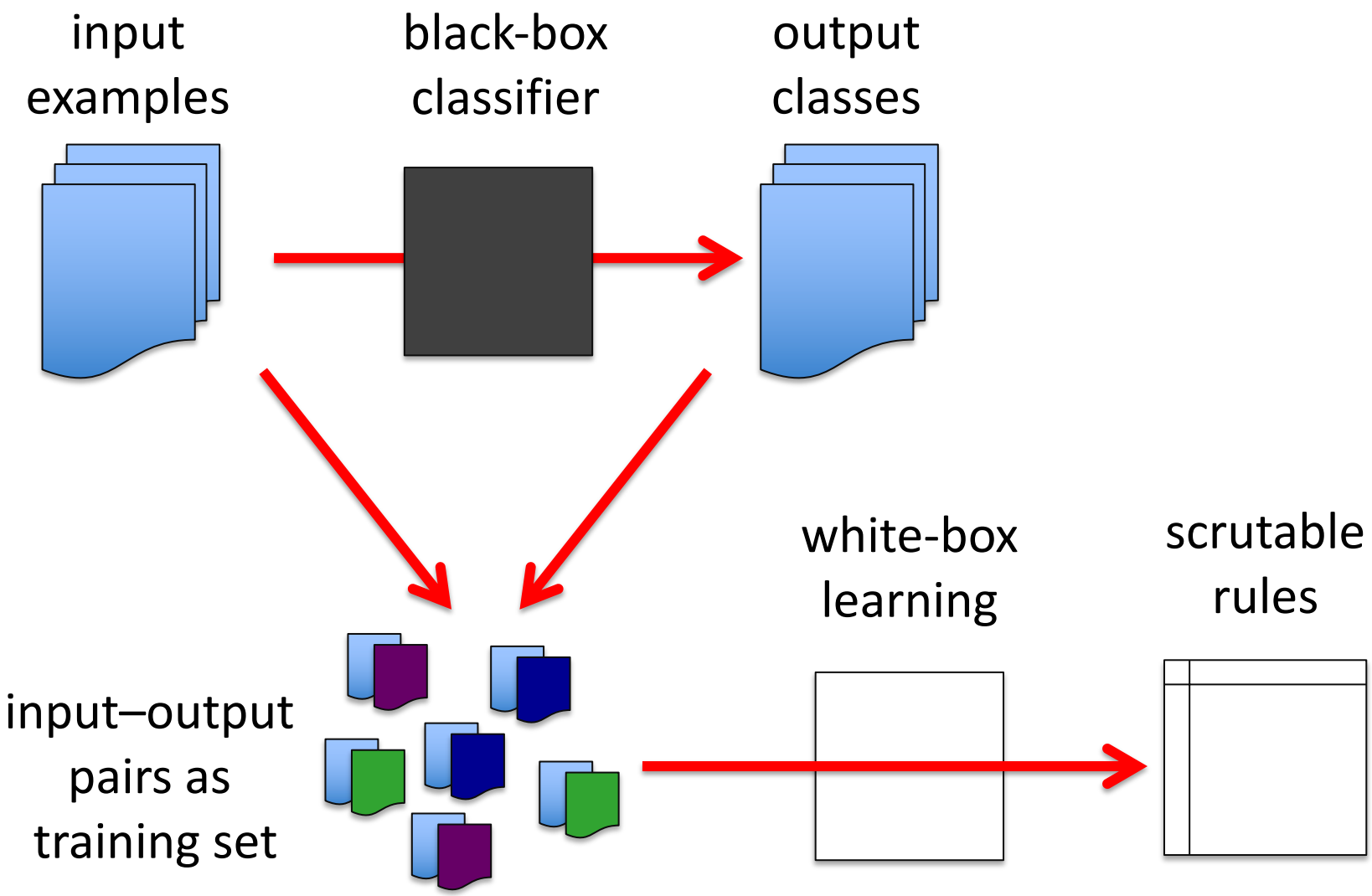
BB3. perturbation analysis for central and boundary cases



BB3. close up


boundary cases
small perturbations
change class
central cases
perturbations
do not change class
penumbra
larger perturbations
change class

BB4. black-box oracle – white-box learning

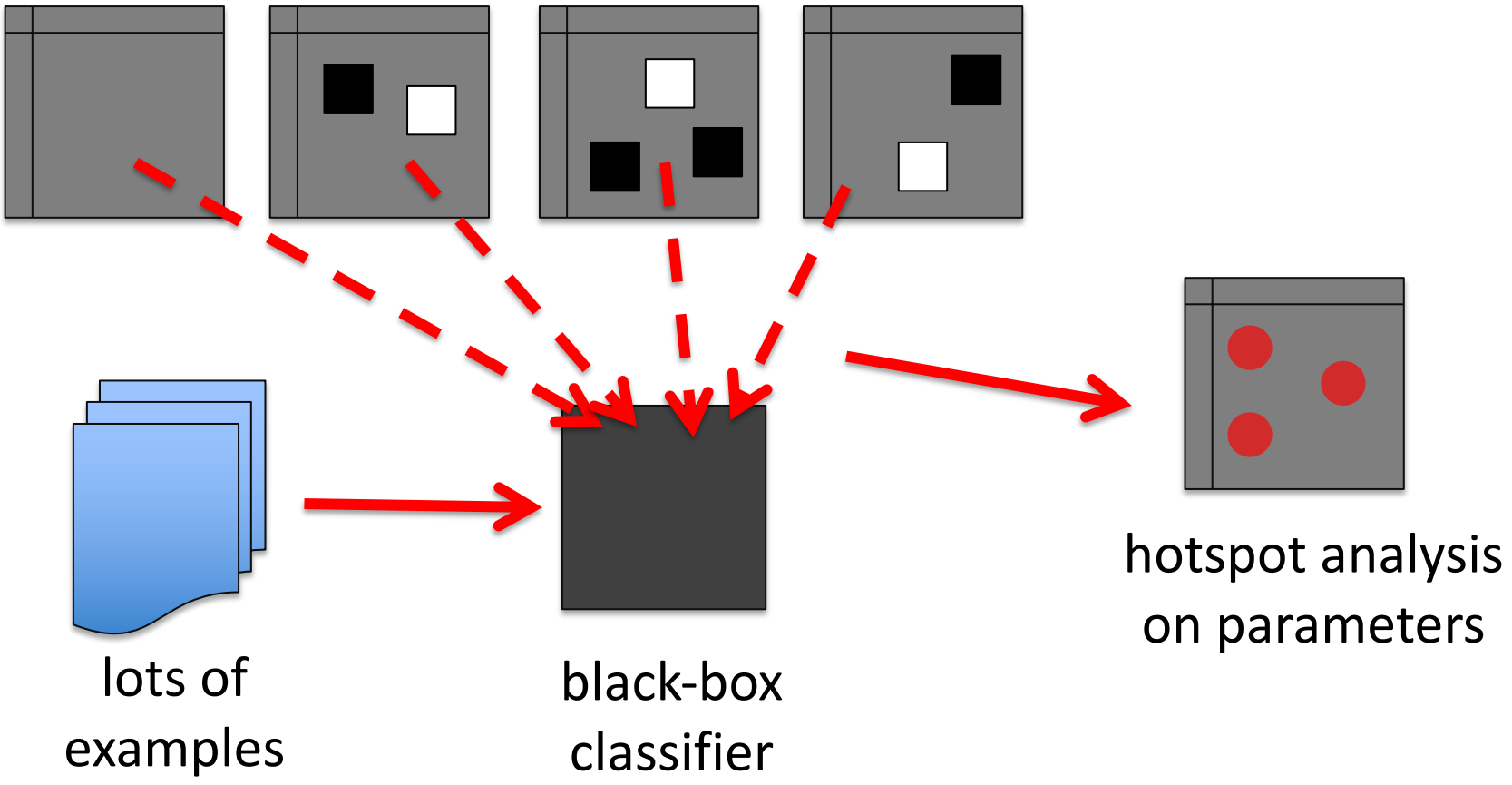


grey-box methods

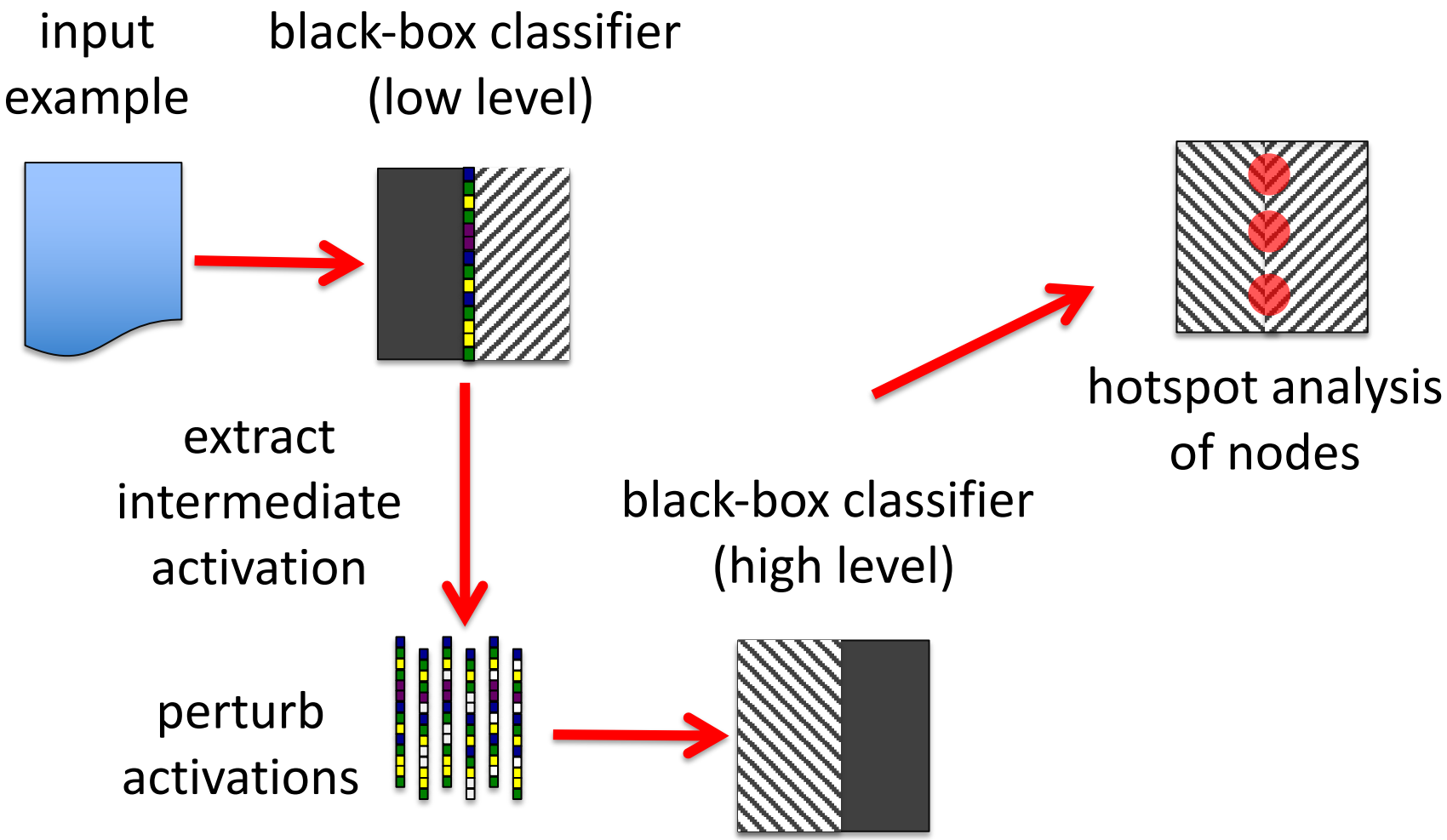
peeking within

GB0a. sensitivity analysis – weights

perturb parameters in the inscrutable rules



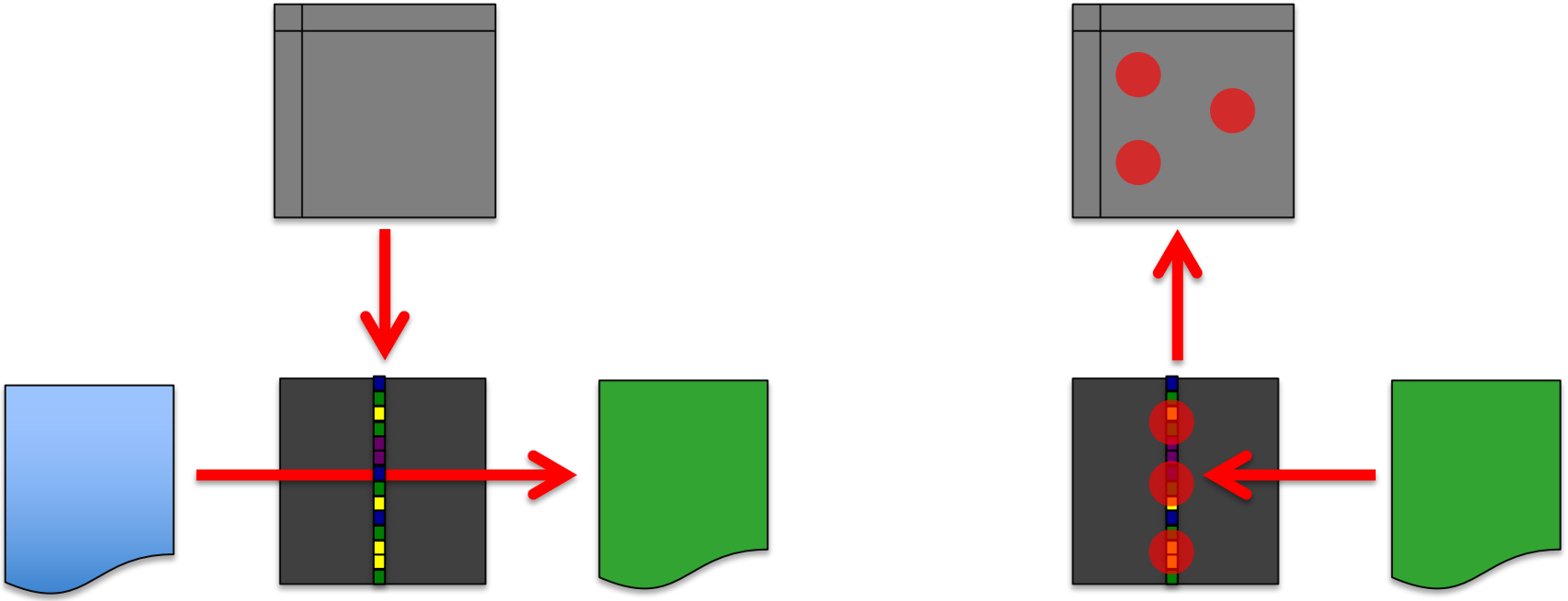
GB0b. sensitivity analysis – activation



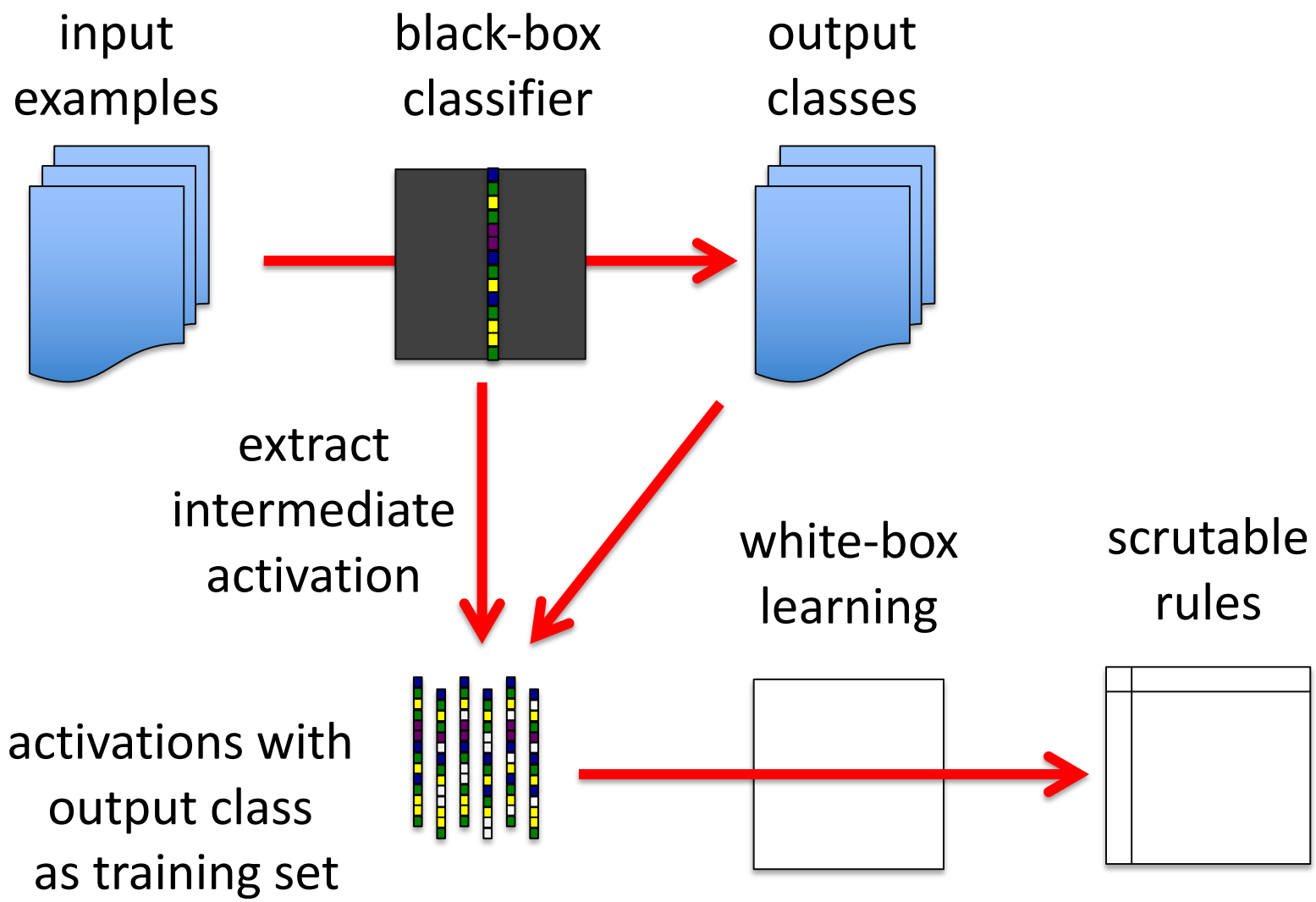
GB0c. sensitivity analysis – algorithmic

apply black-box algorithm

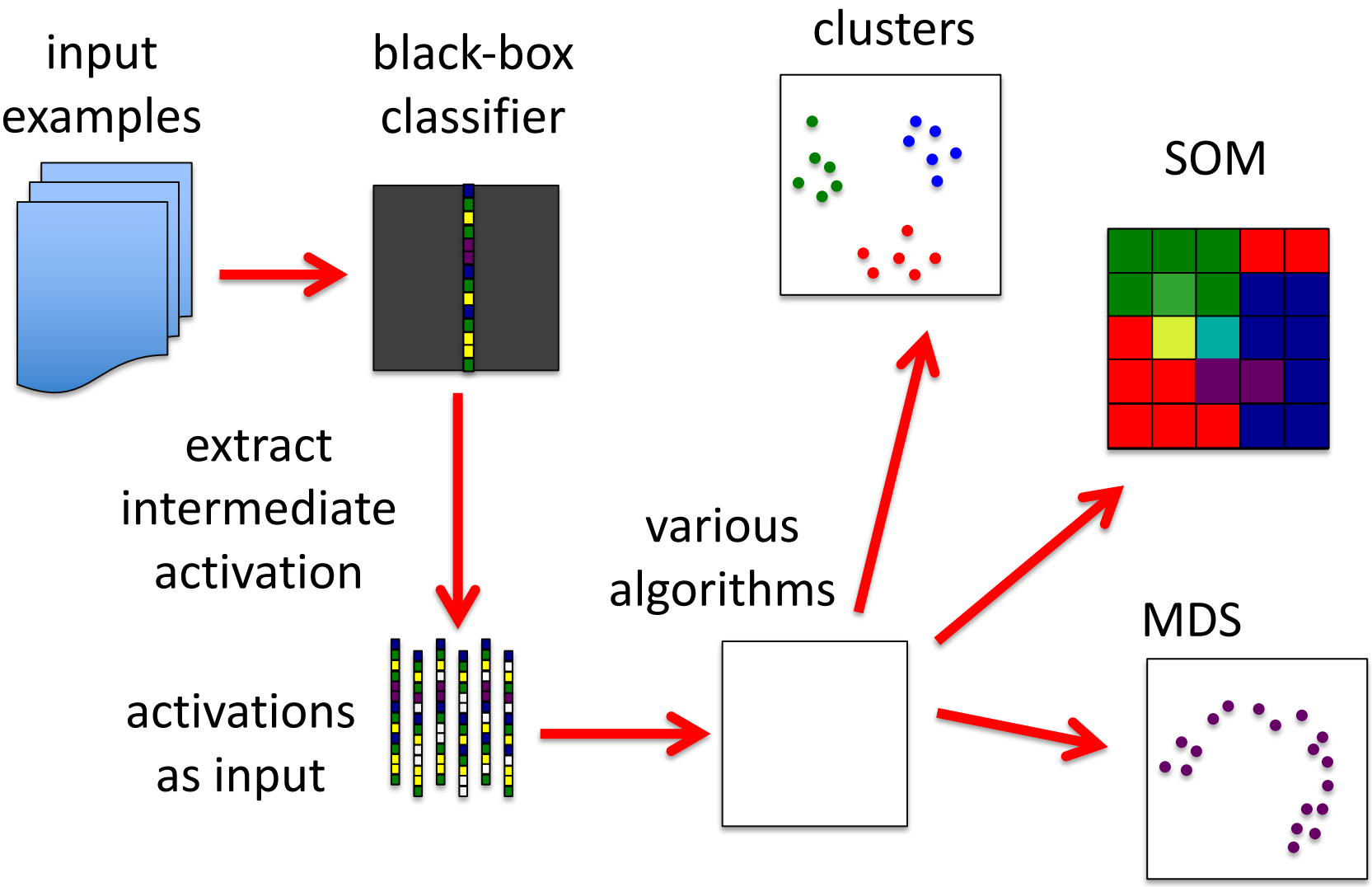
inverse algorithm



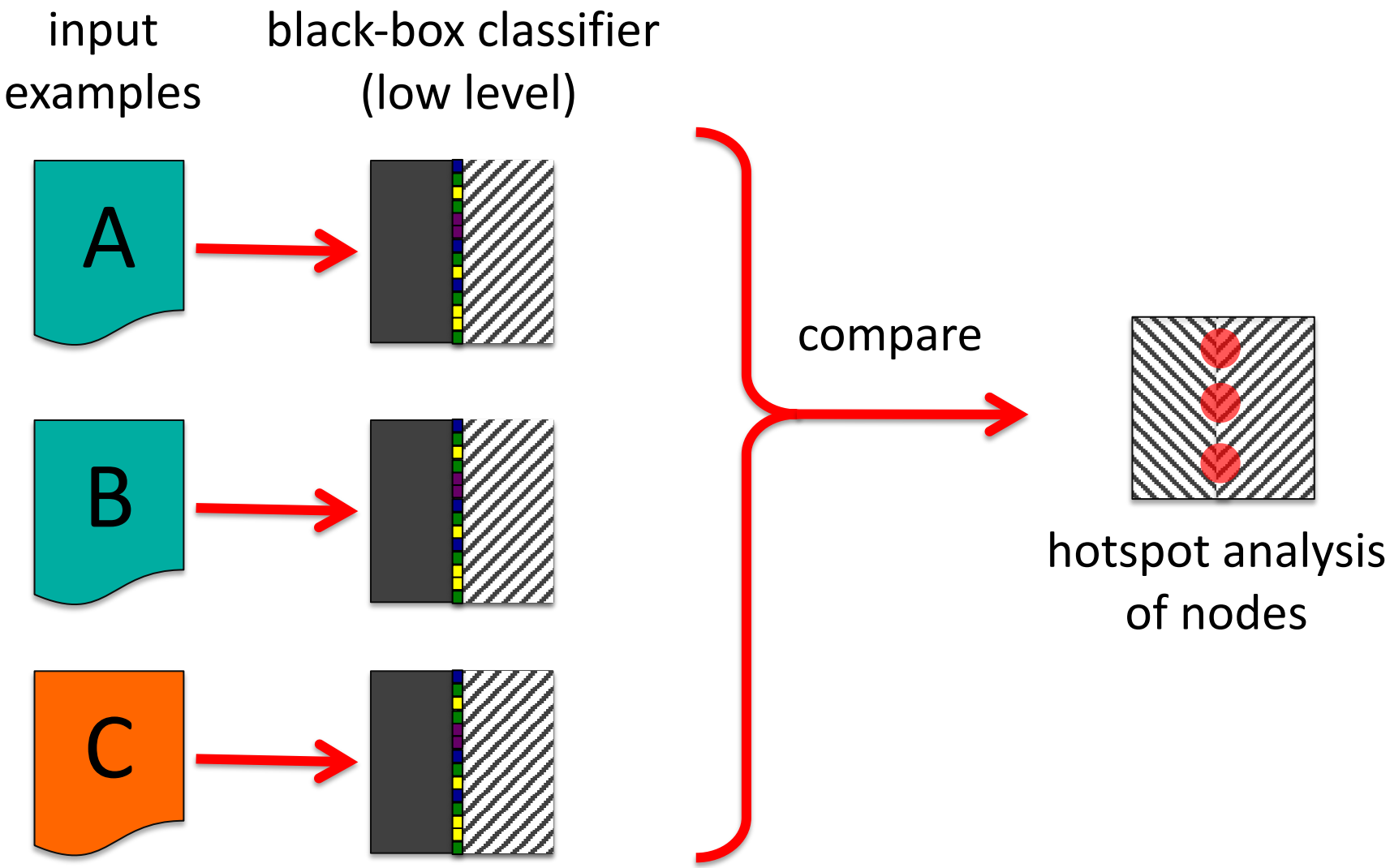
GB1. high level model generation



GB2. Clustering and comprehension of low level

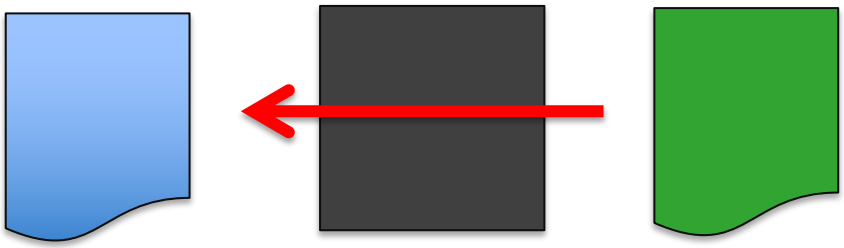


GB3. triad distinctions

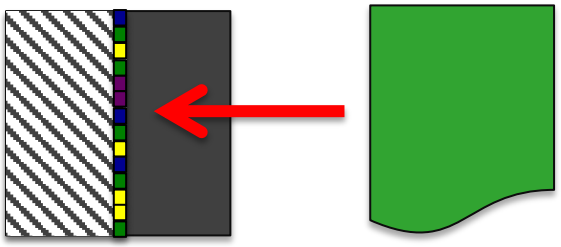


GB4. apply generatively

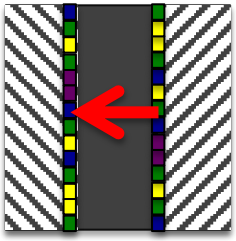
output to input



output to activation



between layers



activation to input