# Actual causality, responsibility, explanations, and fairness – a bird's eye view

We are hiring!! Please contact me at hana@causalens.com


Hana Chockler

causaLens

and

Department of Informatics

King's College, London

# Humans trust Causal AI with complex decisions

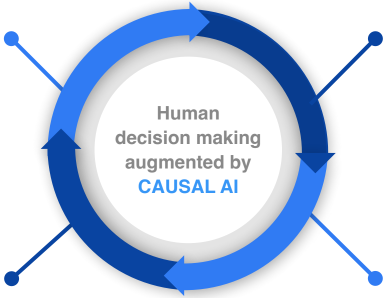
**Correlation ML systems learn to perform simple predictions**

But predictions are a very small element of decision making.

## Causal AI is the only technology that can augment human decision making

**DOMAIN EXPERTISE**
Considering the wider context when making decisions

**SCENARIOS**
Evaluating hypothetical scenarios

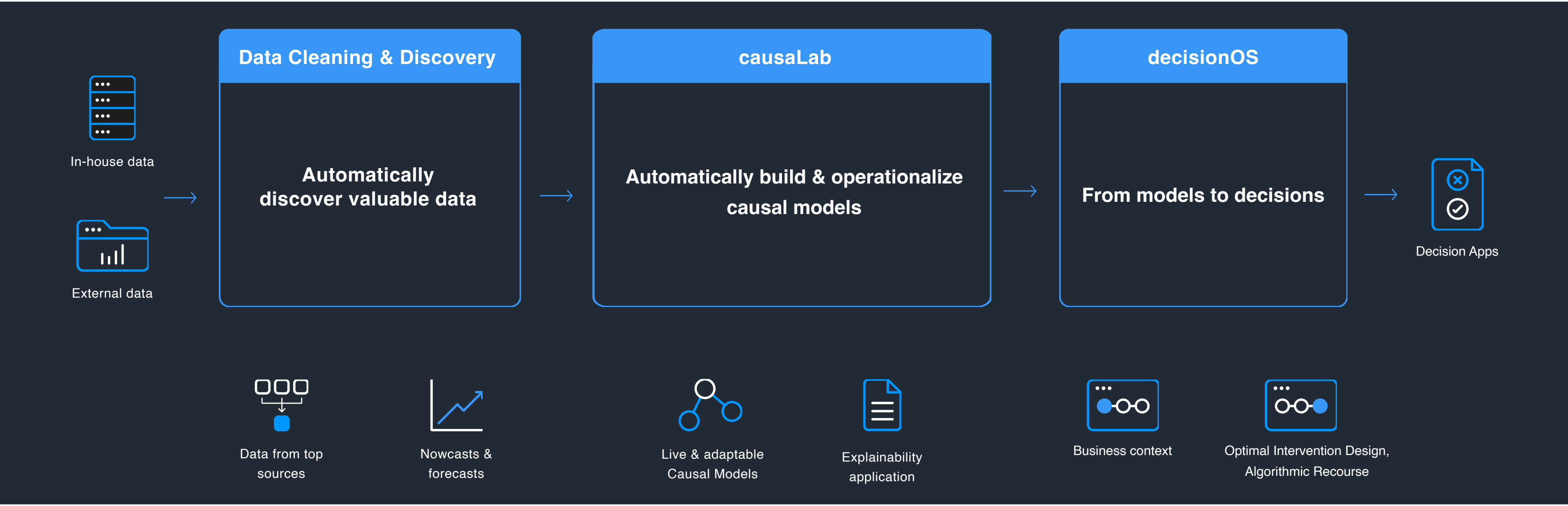



**CONSTRAINTS**
Respecting human, legal and environmental constraints

**INTERVENTIONS**
Planning and designing optimal interventions

# World’s First Full-Stack Causal AI Platform

**We launched the World’s First Causal AI Enterprise Platform, which automates everything from Raw Data to Improved Business Decisions.**



# Actual causality, responsibility, explanations, and fairness – a bird's eye view

Hana Chockler
causaLens
and
Department of Informatics
King's College, London

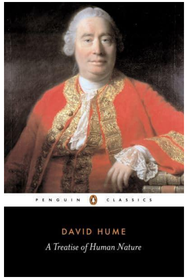
**Background:**

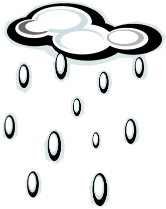
# Causality

When do we say that **A** is a cause of **B**?

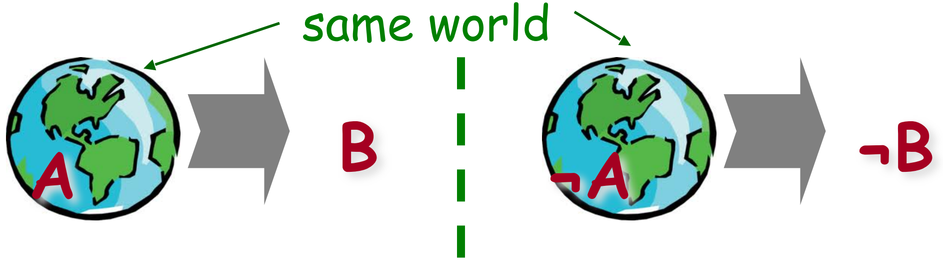
Common approach: **counterfactual causality**.

**A** is a **cause** of **B** if, had **A** not happened, then **B** would not have happened.



Rain is a cause of me being drenched with water.



# Causality

When do we say that **A** is a cause of **B**?

Common approach: **counterfactual causality**.

**We need to capture more complex causal connections!**


Rain or car is a cause of me being drenched?

Rain is a cause of me being drenched.

# Causality

When do we say that **A** is a cause of **B**?

Common approach: **counterfactual causality**.

**We need to capture more complex causal connections!**

preemption

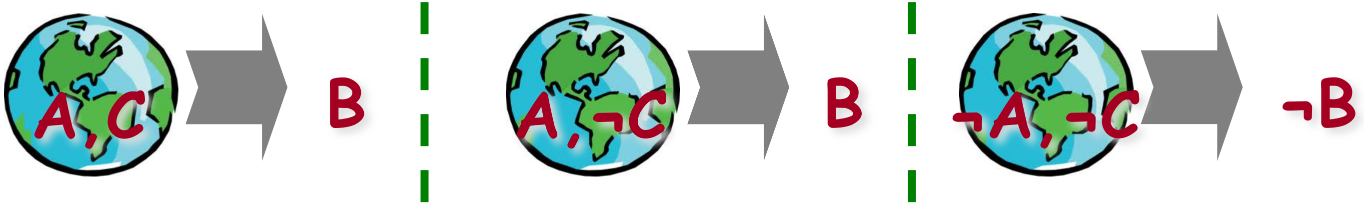


Car is a cause of me being drenched, but not the rain

# Actual causality

## Extends the counterfactual reasoning by having expressive causal models allowing redundancy, preemption, and complex causal structures

Redundancy: A is a cause of B if there exists some contingency ***C***
(change in the current world)
in which B counterfactually depends on A.



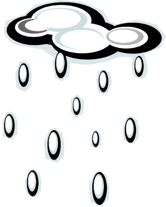

original world    contingency    counterfactual dependence

# Illustration of redundancy in actual causality

Rain is an actual cause of me being drenched.

Contingency = the car

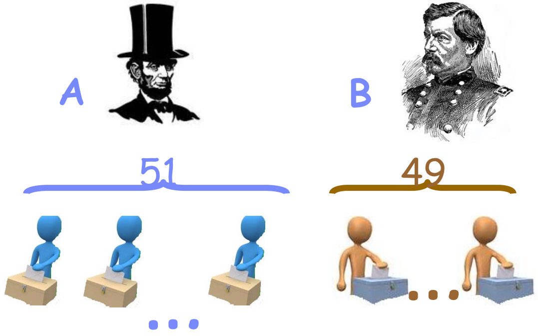
Rain is
a counterfactual cause



# Responsibility: a quantitative measure of causality

## Voting example

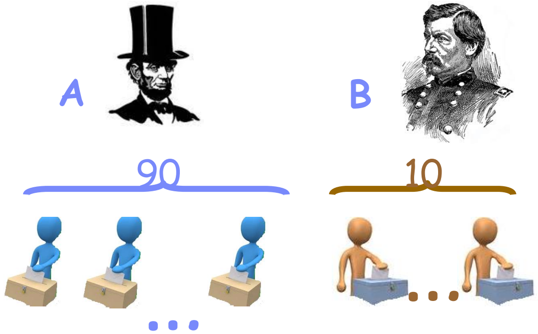
**California:**




Each **blue** voter is a cause of Lincoln's win

**We need to distinguish between the cases!**

**New York:**




Each **blue** voter is a cause of Lincoln's win

# Responsibility: a quantitative measure of causality

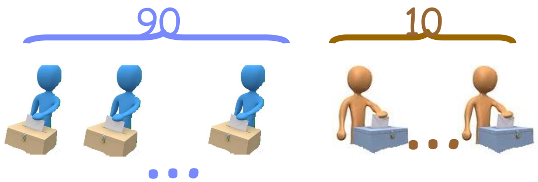
## Voting example


California:

Each **blue** voter is 1-responsible for Lincoln's win

**Responsibility allows to rank causes**


New York:




Each **blue** voter is 1/40-responsible for Lincoln's win

# Complexity of Computing Causality and Responsibility

## Causality:

- $\Sigma_2$- complete for singleton causes.
- $D_2$- complete in general case.

$D_2$ is the difference class of $\Sigma_2$ and $\Pi_2$

## Responsibility:

- $FP^{\Sigma_2}[\log(n)]$- complete.

INTRACTABLE

## Complexity of Computing Causality and Responsibility

**Causality:**

- $\Sigma_2$- complete for singleton causes.

**Responsibility:**

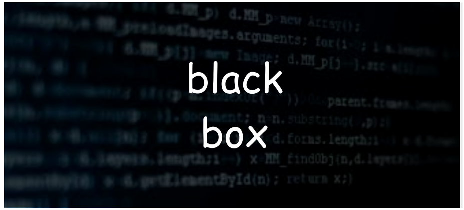
- $FP^{\Sigma_2[\log(n)]}$- complete.

INTRACTABLE

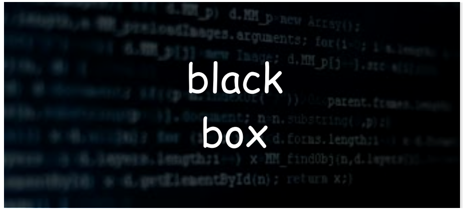
# The good news:

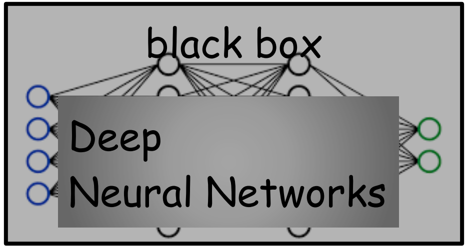
- There are linear-time approximation algorithms
  - Accurate on most problems
- We usually care only about highest-ranked causes
  - Polynomial to compute the exact set

Modern computerized systems are
**huge and difficult to understand**



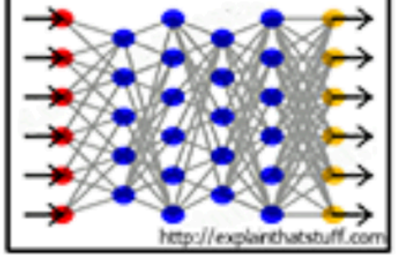
Modern computerized systems are
huge and difficult
**or even impossible to understand**



# From DARPA:

**AI System**




- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

**DoD and non-DoD Applications**

**Transportation**

**Security**

**Medicine**

**Finance**

**Legal**

**Military**

**User**

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

ico.
Information Commissioner's Office

The UK's independent authority set up to uphold information rights in the public interest, promoting openness by public bodies and data privacy for individuals.

Home | Your data matters | For organisations | Make a complaint | Action we've taken

For organisations / Guide to Data Protection / Key DP themes /
Explaining decisions made with Artificial Intelligence

# Explaining decisions made with AI

ico. The Alan Turing Institute

GDPR right to explanation

EUROPEAN COMMISSION

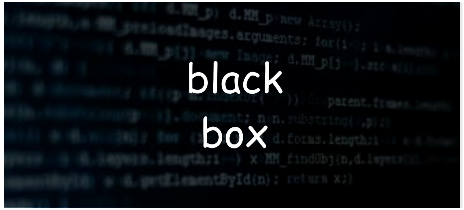
Brussels, 19.2.2020
COM(2020) 65 final

**WHITE PAPER**

**On Artificial Intelligence - A European approach to excellence and trust**

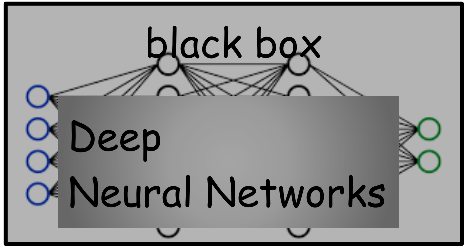

Modern computerized systems are
huge and difficult
or even impossible
to understand

Can we understand and fix errors?




Can we trust the system?

Can we explain the system's decisions?



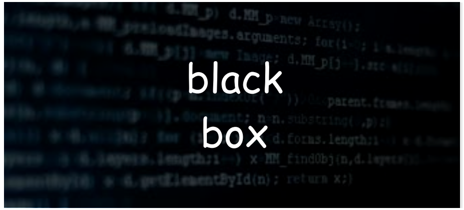

Is the system fair?

Modern computerized systems are huge and difficult

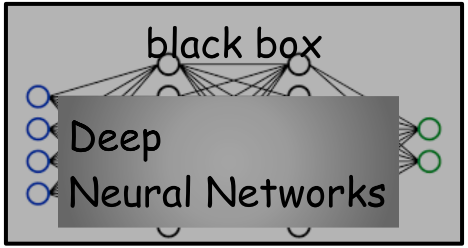
or even impossible to understand


Can we understand and fix errors?

Can we trust the system?



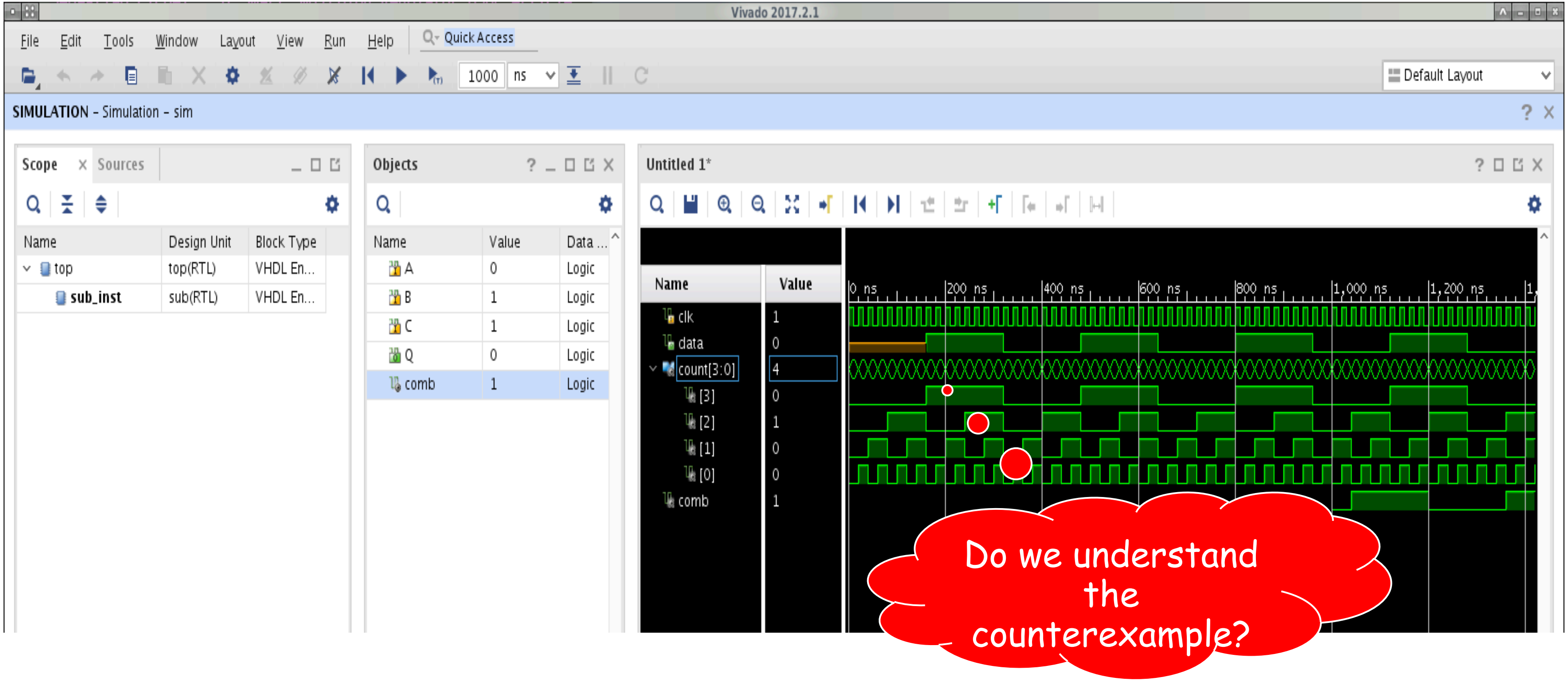

Can we explain the system's decisions?



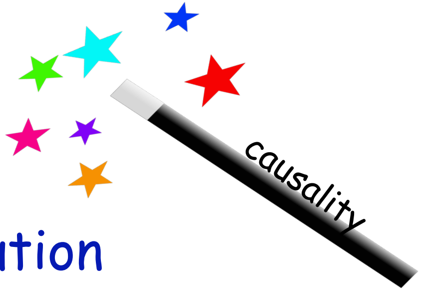

Is the system fair?

# Counterexamples in hardware

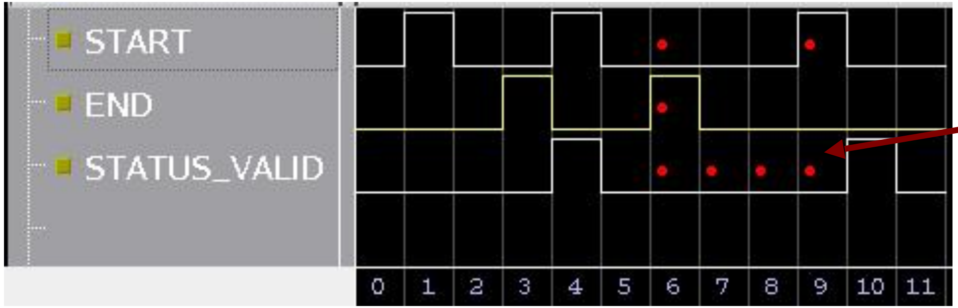
A huge timing diagram that is very difficult to understand



Explaining counterexamples using causality
(Red Dots)
part of IBM tool




# A timing diagram of a buggy hardware execution

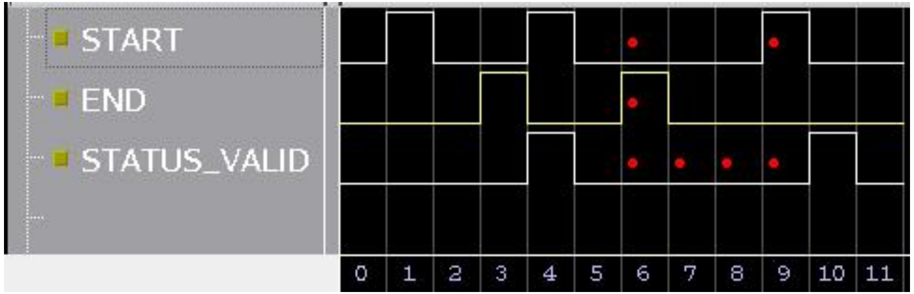


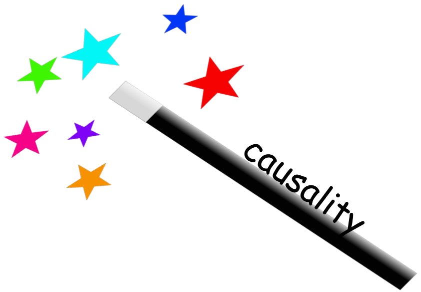

causes marked as red dots

$\varphi$ = always ((!START and !STATUS_VALID and END( ->
next(!START Until (STATUS_VALID and READY))

works and is really useful!

Explaining counterexamples using causality
(Red Dots)
part of IBM tool




# Following this work…

Many applications of causality and responsibility to software engineering

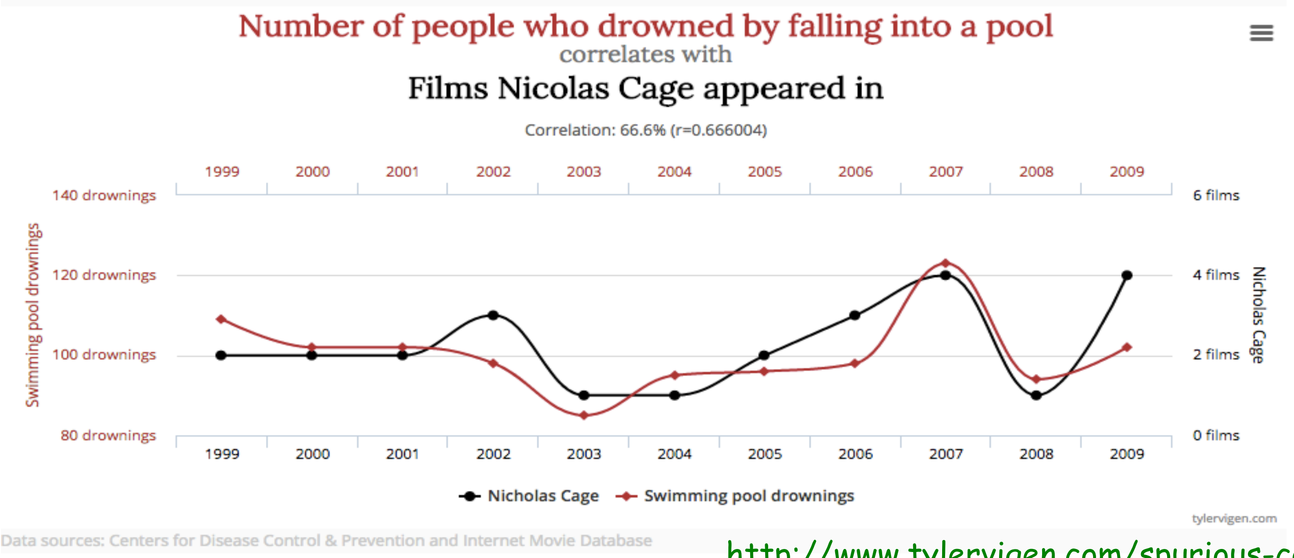
**CREST workshop**

EUROPEAN JOINT CONFERENCES ON THEORY & PRACTICE OF SOFTWARE

causal debugging for software

# Explanation of faults in software testing - SOA

- **Statistical Analysis for Fault Localisation**
  - Looks for **correlation** – elements that appear more in failing traces than in passing ones are suspicious
  - Elements are ordered by their degree of suspiciousness

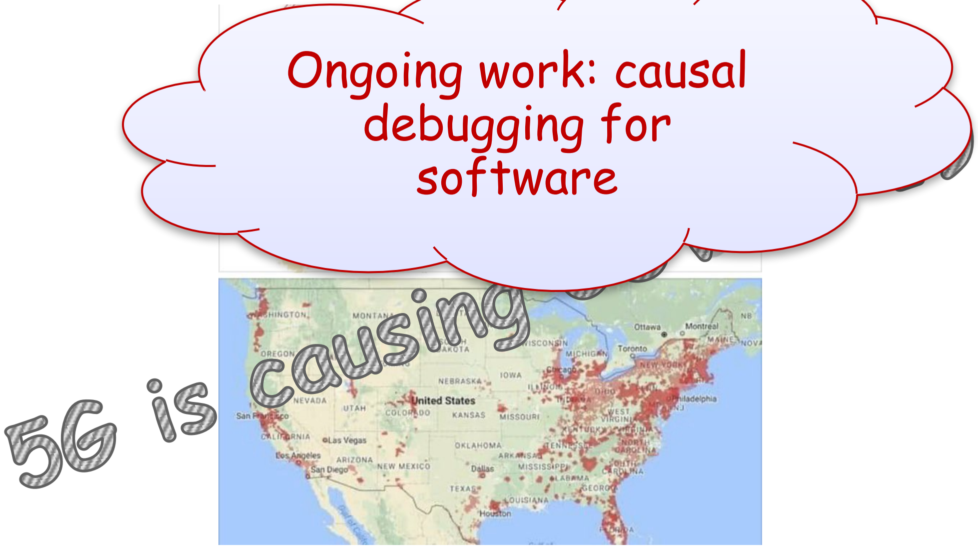



http://www.tylervigen.com/spurious-correlations

# Explanation of faults in software testing - SOA

- **Statistical Analysis for Fault Localisation**
  - Looks for **correlation** – elements that appear more in failing traces than in passing ones are suspicious
  - Elements are ordered by their degree of suspiciousness

Ongoing work: causal debugging for software



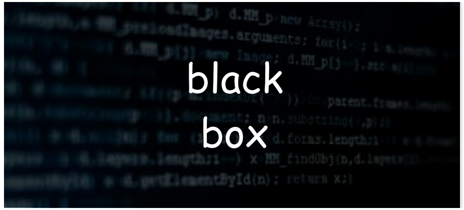

http://www.medium.com

Modern computerized systems are
huge and difficult

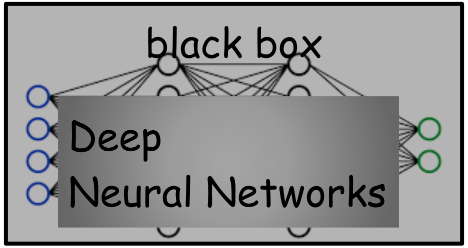
# or even impossible to understand

Can we understand and fix errors?

Can we trust the system?




Can we explain the system's decisions?



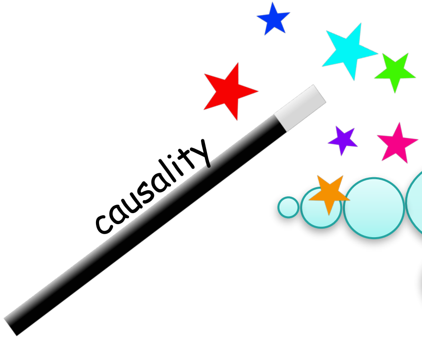

Is the system fair?

# Explanations for Deep Neural Network's decisions

DNN for classifying animals

red panda

Because of this part:




**Explanation:** minimal, sufficient, non-trivial subset of the pixels of the image

## How to detect misclassification?

### Example: wolves vs huskies

Training phase: Pictures of wolves and huskies → DNN

Classification phase:

(husky)

→ DNN → ~~wolf~~

© Ribeiro, Singh, Guestrin.
"Why Should I Trust You?", KDD'16

# Subtle misclassification - uncovered by explanations

DNN for classifying images

seems ok

cowboy hat

Explanation uncovered misclassification!

Because of this part: 

# Explanations for DNN's decisions - based on ranking of causes

DNN for classifying

red panda

Because  of this part:

How to compute a minimal sufficient subset efficiently?


Use causal analysis for ranking causes & choose from the top

# Photobombing (Partially occluded images)

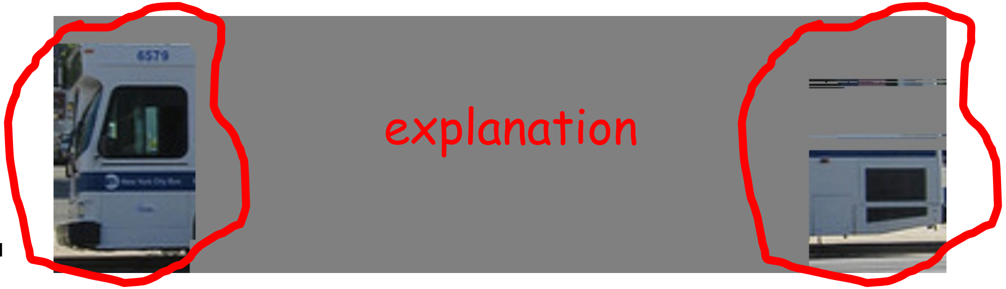
~~dog~~ people

Partially occluded images
have non-contiguous explanations

# Explanations for DNN's decisions - Photobombing

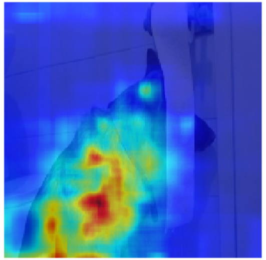

DNN

bus



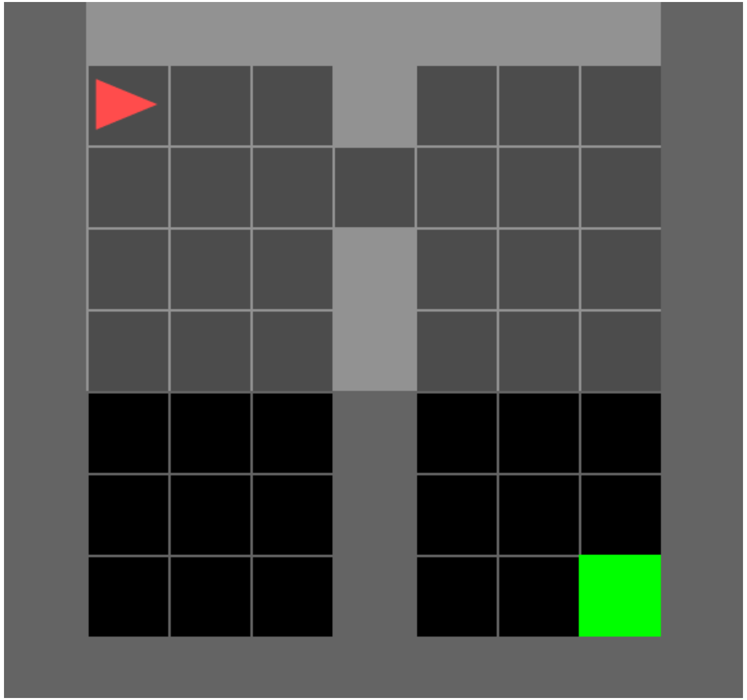

DNN

dog

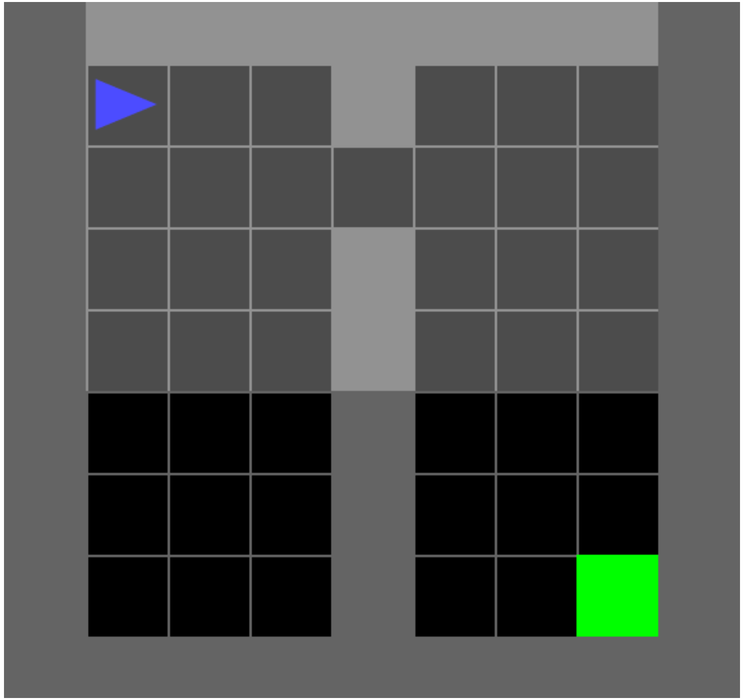


ranking

# Reinforcement learning – causal simplification of policies



Original policy

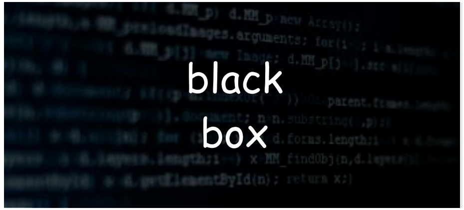


Simplified policy

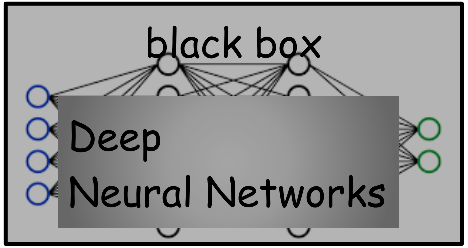
Modern computerized systems are huge and difficult or even impossible to understand

Can we understand and fix errors?

Can we trust the system?



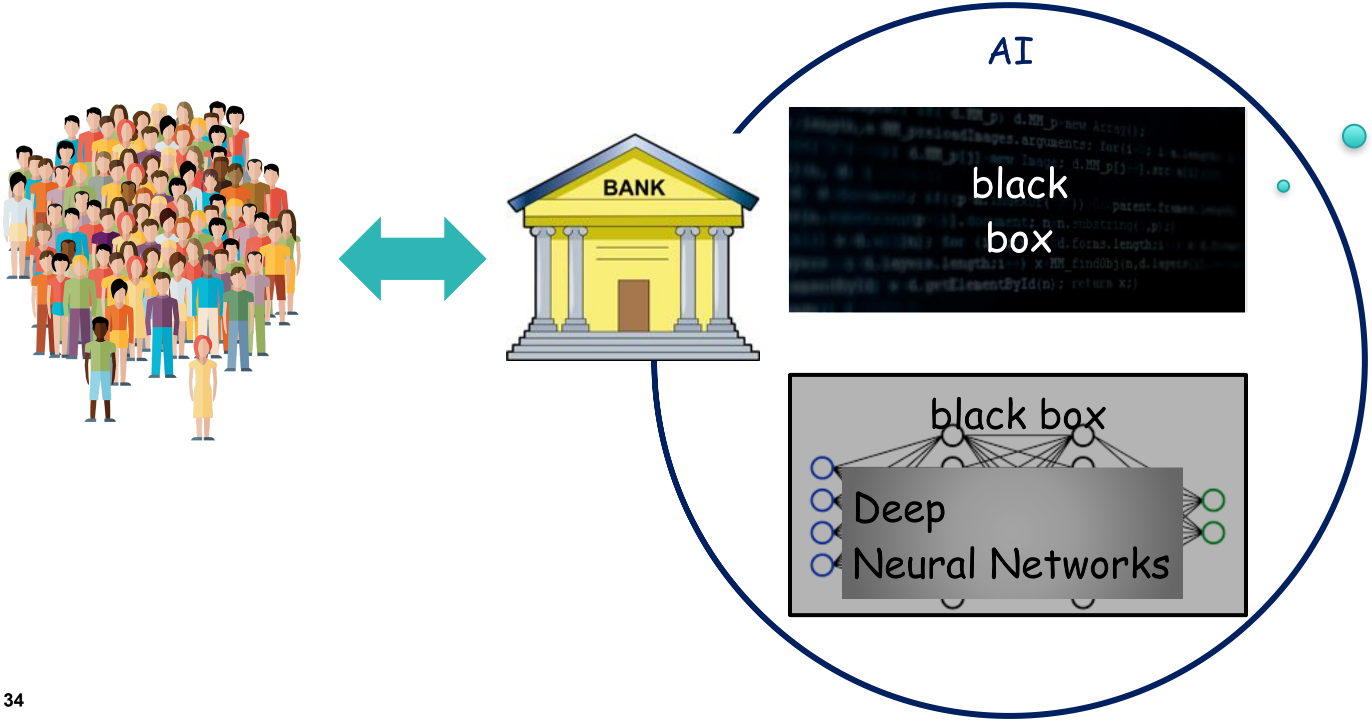

Can we explain the system's decisions?




Is the system fair?

**AI black-box systems are widely used**

**Their decisions affect people**




# Fairness - Motivation

## What is fair? How do we detect unfairness?
## What should the regulations require?

**DHH** @dhh · Nov 9, 2019
Replying to @dhh

To be fair, this is an even more egregious version of the same take. THE ALGORITHM is always assumed to be just and correct. It's verdict is thus predestined to be a reflection of your failings and your sins.

> **Isles47** @isles47
> Replying to @dhh and @AppleCard
>
> Haha this is absurd. Litterally none of the things you list here have any effect on credit approval. Whats her existing line of credit, what's her credit score, what outstanding debt does she have? How old is her original line of credit?

**Steve Wozniak**
@stevewoz

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

6:58 AM · Nov 10, 2019

181 Reply Copy link to Tweet

Read 33 replies

GOOGLE WEB ENTERTAINMENT

# Google's algorithms advertise higher paying jobs to more men than women

*Study suggests how 'impartial' data can encode real-life prejudices*

By James Vincent | Jul 7, 2015, 5:40am EDT

Via MIT Technology Review

# Women less likely to be shown ads for high-paid jobs on Google, study shows

**Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs**

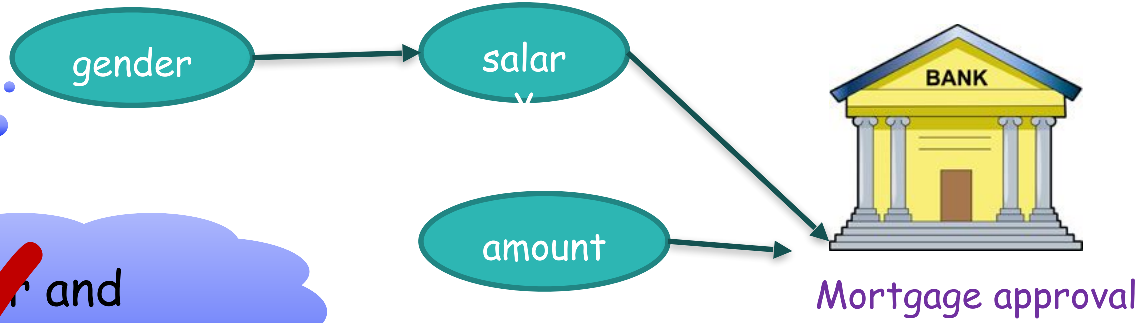
Ensure that gender is not a part of the model?

# Fairness
## 1st attempt

Ensure that gender is not a part of the model?

## The State of the Gender Pay Gap in 2021

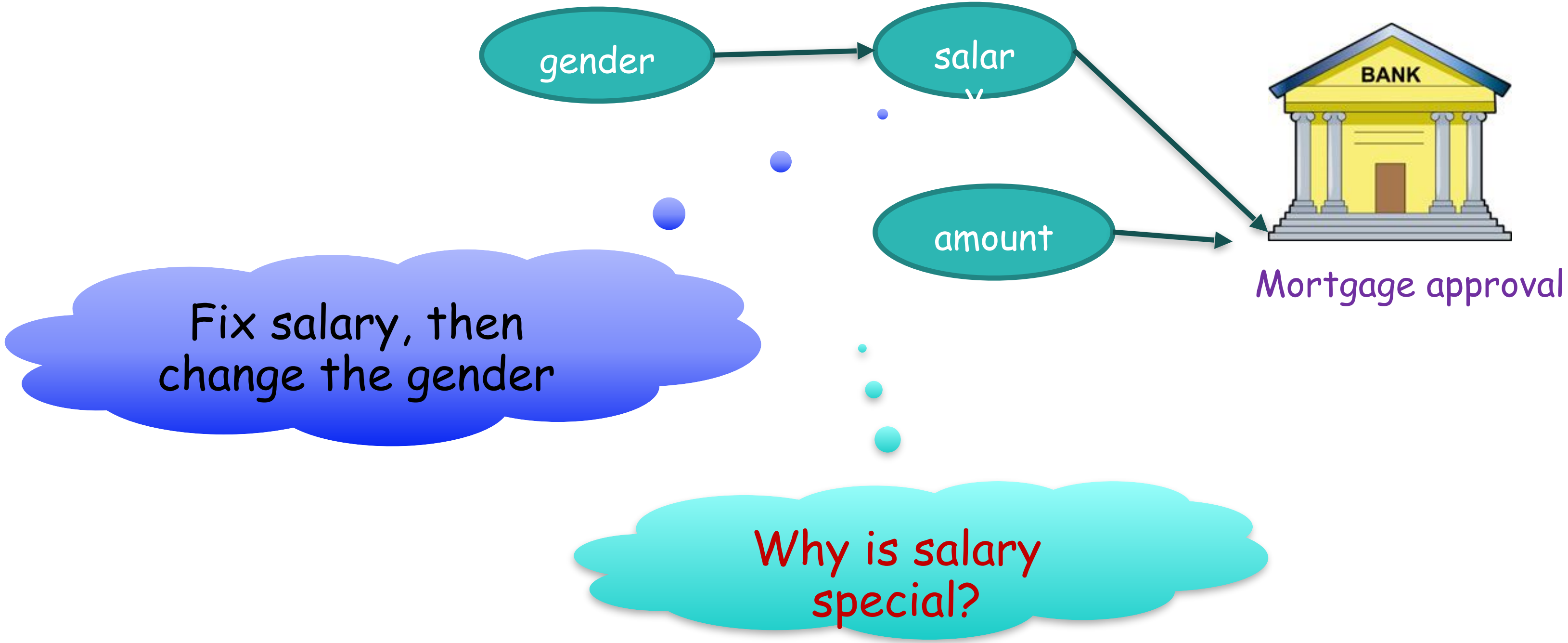
**In 2021, women earn 82 cents for every dollar earned by men.**




~~Flip the gender and check the decision?~~

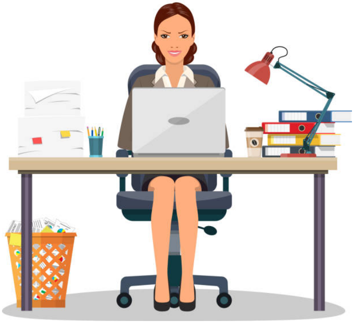
Won't work because salary will change

# Fairness
## 2nd attempt

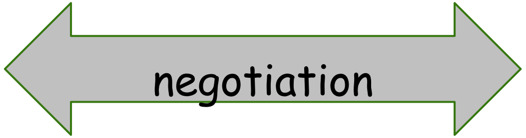


# Fairness

- The rough idea: define the set of **sensitive** variables and the set of **allowed** variables



regulator




black


<u>Sensitive variables:</u>
- Gender
- Ethnicity
- Age
- …

✅

<u>Proposed allowed variables:</u>
- Salary
- Savings
- House price
- …

# Which variables should be allowed?

Business necessity ⟷ Fairness

- Salary – for mortgage applications
- University rank – for job applications
    - Is it fair?

All variables are allowed — No allowed variables

Unfair decisions — regulation — Risky decisions

# Fairness of a system – for certification

A model M is **fair** wrt the set X of sensitive variables and the set Y of allowed variables if for any setting, changing the values of sensitive variables has no effect on the outcome of M if the allowed variables are fixed.

How can we check this?

Certification process:

co-NP complete

interventions

black Box AI

# Fairness for a single applicant (verifying a lawsuit)

A model M is **fair** wrt the set X of sensitive variables and the set Y of allowed variables for a given applicant Alice if for the values describing Alice, changing the values of sensitive variables has no effect on the outcome of M for Alice if the allowed variables are fixed.

How can we check this?

Polynomial in the number of settings of sensitive variables

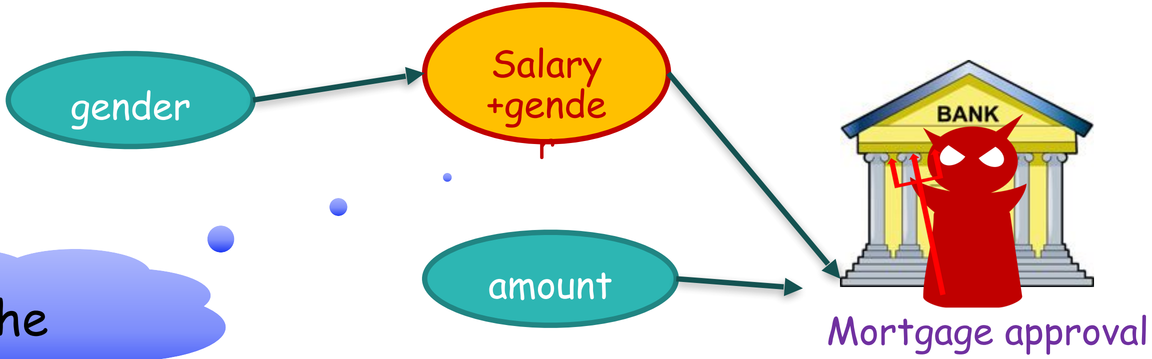
Check for a single applicant

interventions

black Box AI

# Proxy variables

What if the bank hides the sensitive variables in allowed ones?




We will miss the unfairness

# Proxy variables

Not always clear whether some variables are proxy

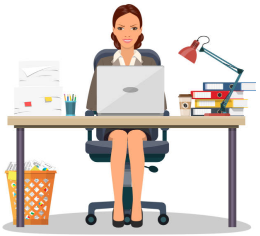

black
Box AI

Public information

Social networks

Religious holidays celebrated

A proxy for religious affiliation and/or ethnicity!



# Summary:
# Proposed regulation



Mortgage approval

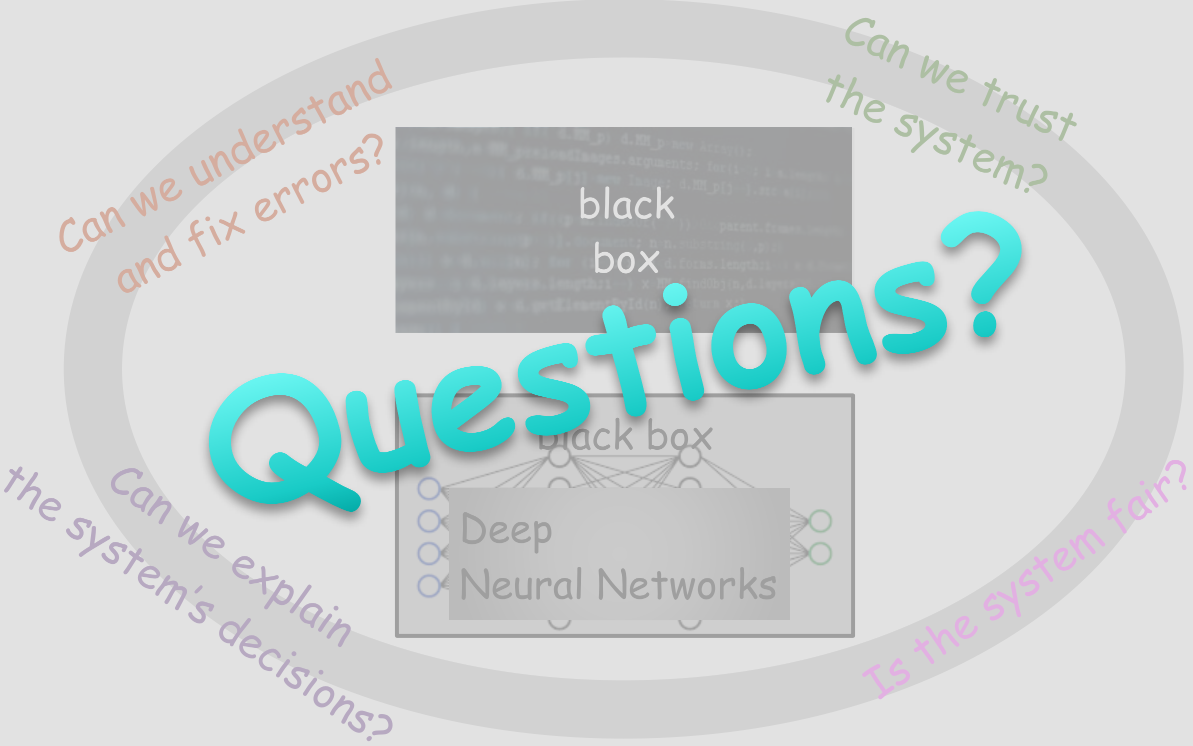
Fairness + no proxy variables

Precise algorithm or approximation

Statistical independence

# Bibliography


- Chockler & Halpern. "Responsibility and Blame: A Structural-Model Approach". J. Artif. Intell. Res. 22: 93-115 (2004)
- Beer, Ben-David, Chockler, Orni, Trefler. "Explaining Counterexamples Using Causality". CAV'09: 94-108.
- Aleksandrowicz, Chockler, Halpern, Ivrii. "The Computational Complexity of Structure-Based Causality". AAAI'14: 974-980.
- Alrajeh, Chockler, Halpern. "Combining Experts' Causal Judgments". AAAI'18: 6311-6318.
- Sun, Chockler, Huang, Daniel Kroening. "Explaining Image Classifiers Using Statistical Fault Localization". ECCV'20: 391-406.
- Chockler, Kroening, Sun. "Explanations for Occluded Images". ICCV'21: 1234-1243.
- Pouget, Chockler, Sun, Kroening. "Ranking Policy Decisions". NeurIPS'21.
- Chockler, Halpern. "On Testing for Discrimination Using Causal Models". AAAI'22.


Questions?
Can we understand
and fix errors?
Can we trust
the system?
black
box
black box
Deep
Neural Networks
Can we explain
the system's decisions?
Is the system fair?