The AI black box explanation problem

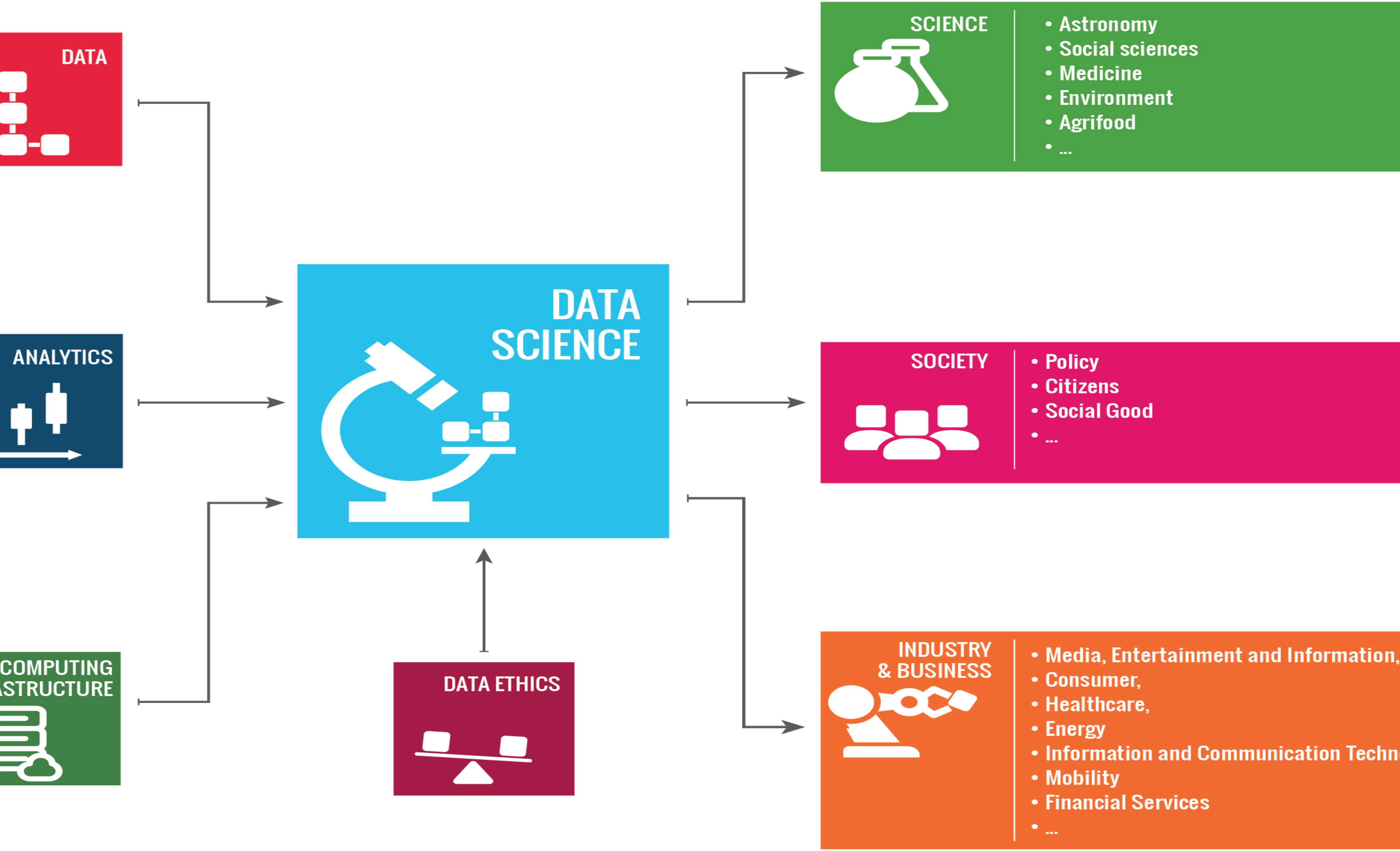
Dino Pedreschi

KDD Lab Dip. Informatica, Univ. Pisa & ISTI-CNR

www.sobigdata.eu

UNIVERSITÀ DI PISA

ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"



DATA

ANALYTICS

COMPUTING
ASTRUCTURE

DATA ETHICS

DATA
SCIENCE

SCIENCE

- Astronomy
- Social sciences
- Medicine
- Environment
- Agrifood
- ...

SOCIETY

- Policy
- Citizens
- Social Good
- ...

INDUSTRY
& BUSINESS

- Media, Entertainment and Information,
- Consumer,
- Healthcare,
- Energy
- Information and Communication Techn
- Mobility
- Financial Services
- ...

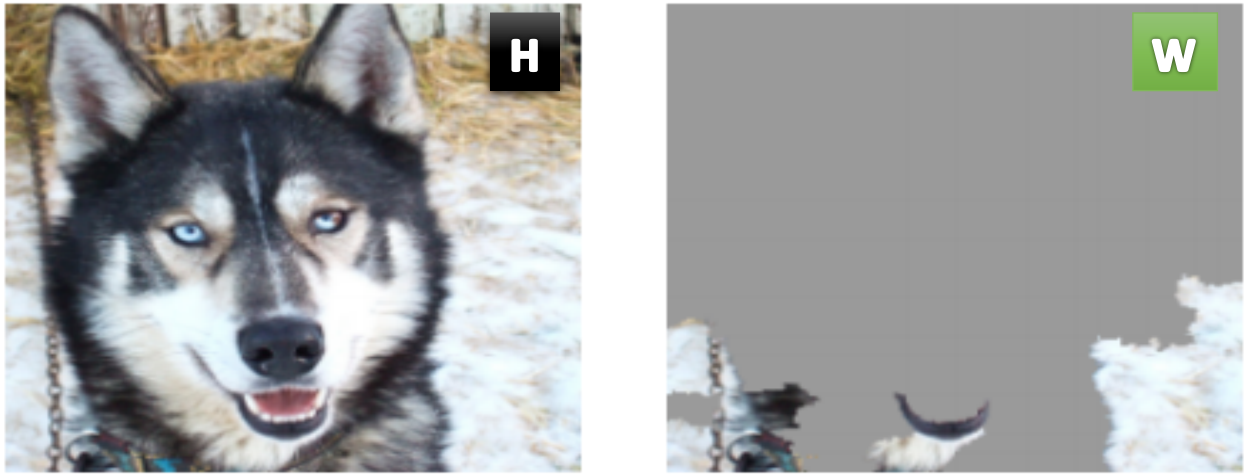
n, Giannotti, Pedreschi, G7 2017 Position paper on Data Science

H

H

W

Deep learning classifies a wolf in a picture based on … **the snow in the background!**



(a) Husky classified as wolf

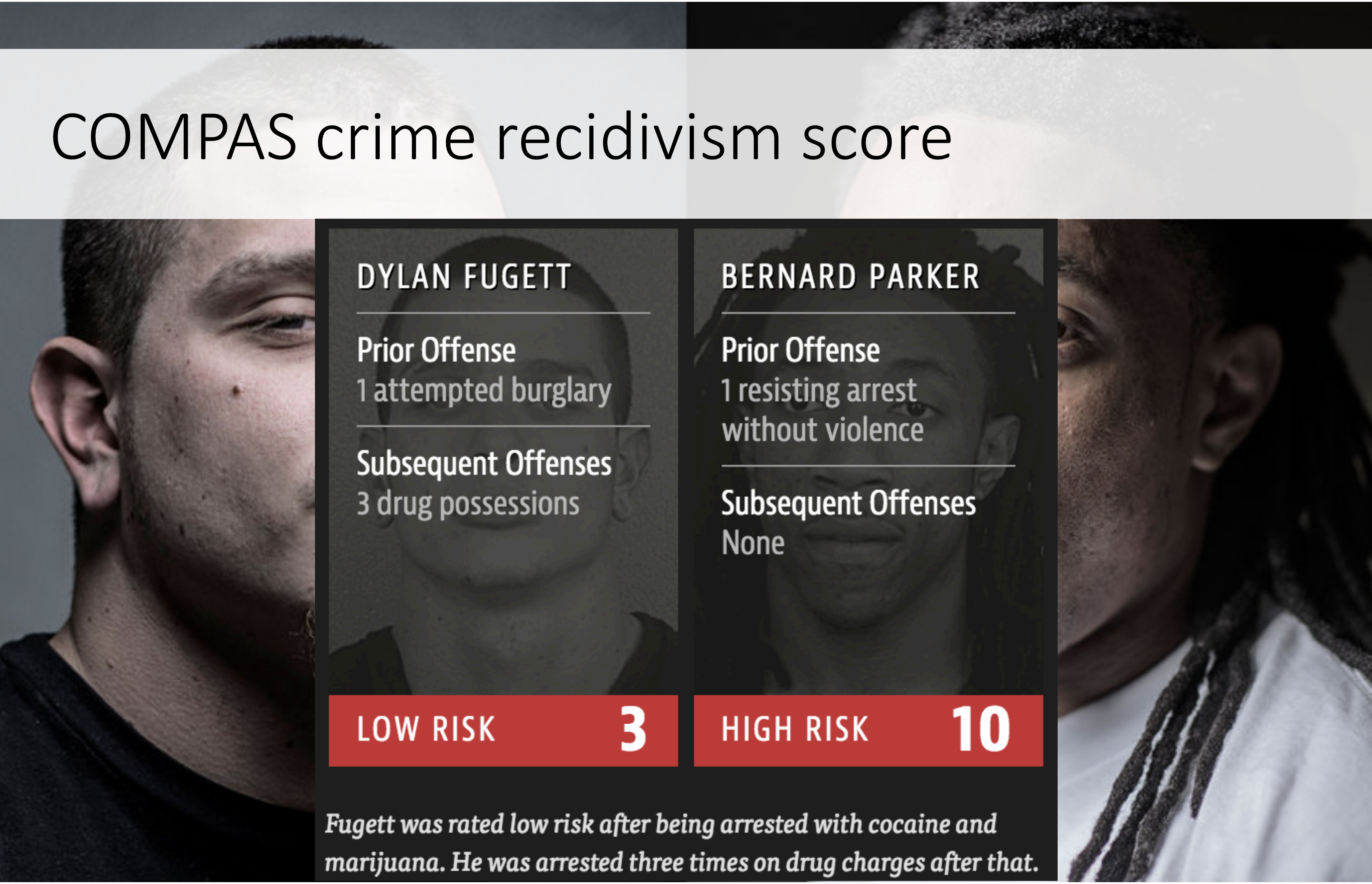
(b) Explanation

Ribeiro et al., KDD 2016

ilitary tank classification depends on the backgrou

ure **558**, 357-360 (2018)

COMPAS crime recidivism score



DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3

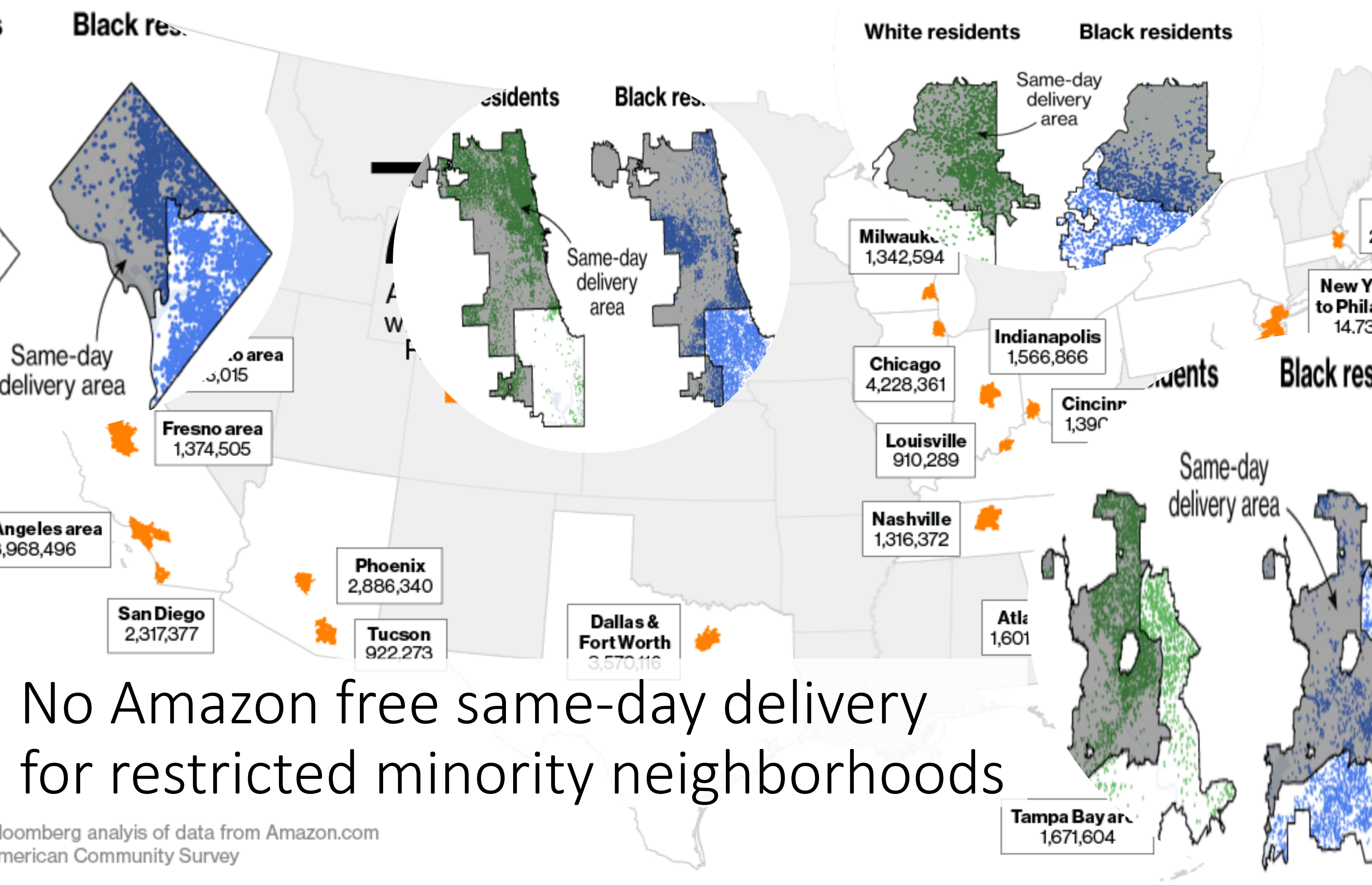
BERNARD PARKER

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

HIGH RISK 10

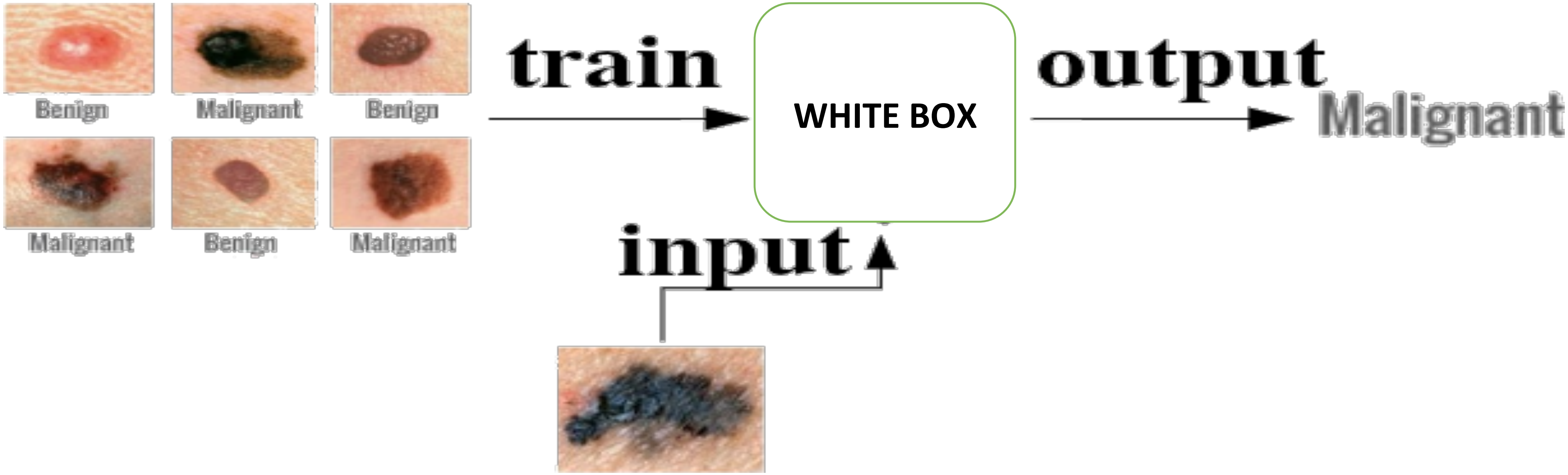
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

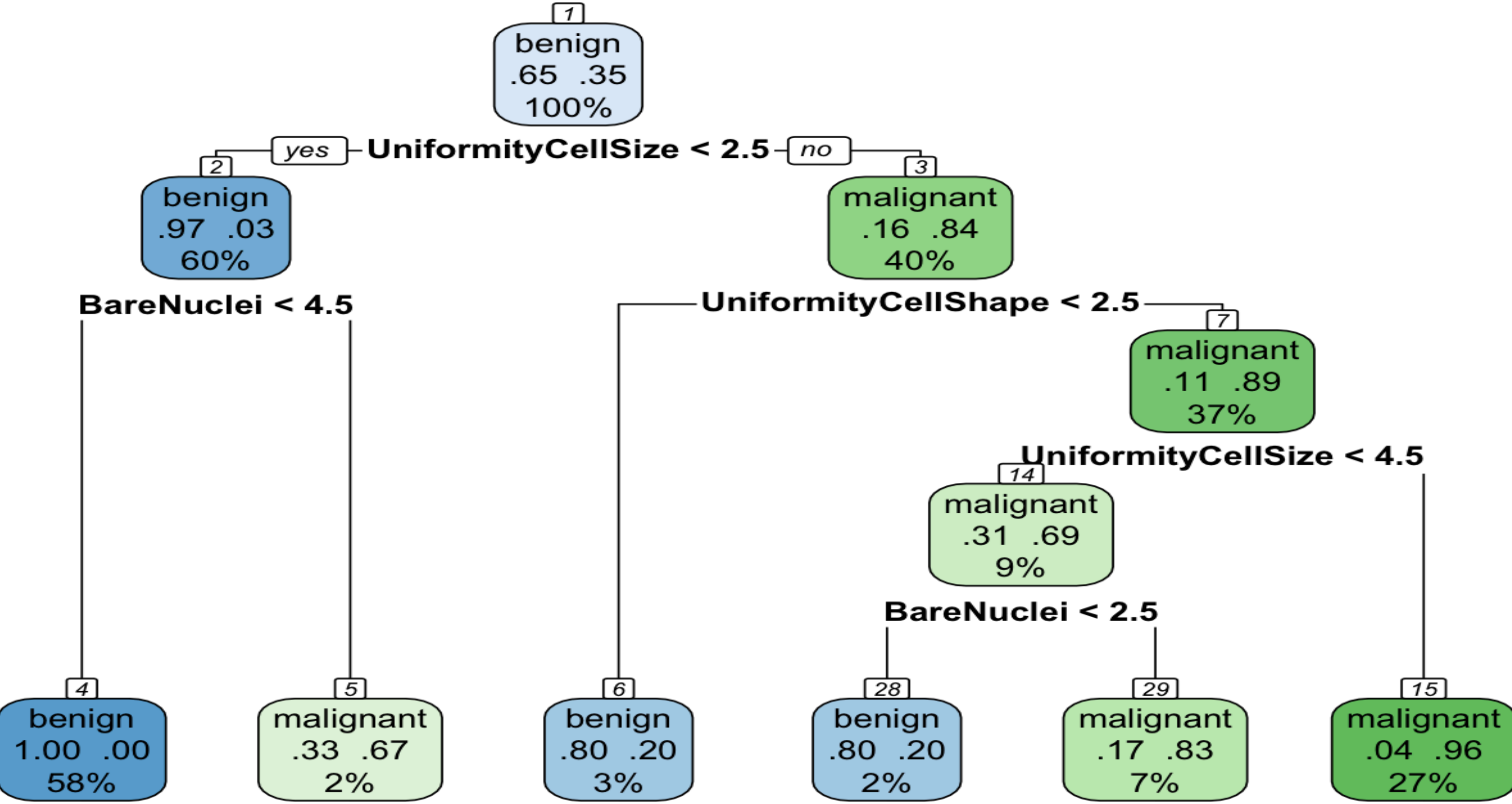


No Amazon free same-day delivery for restricted minority neighborhoods

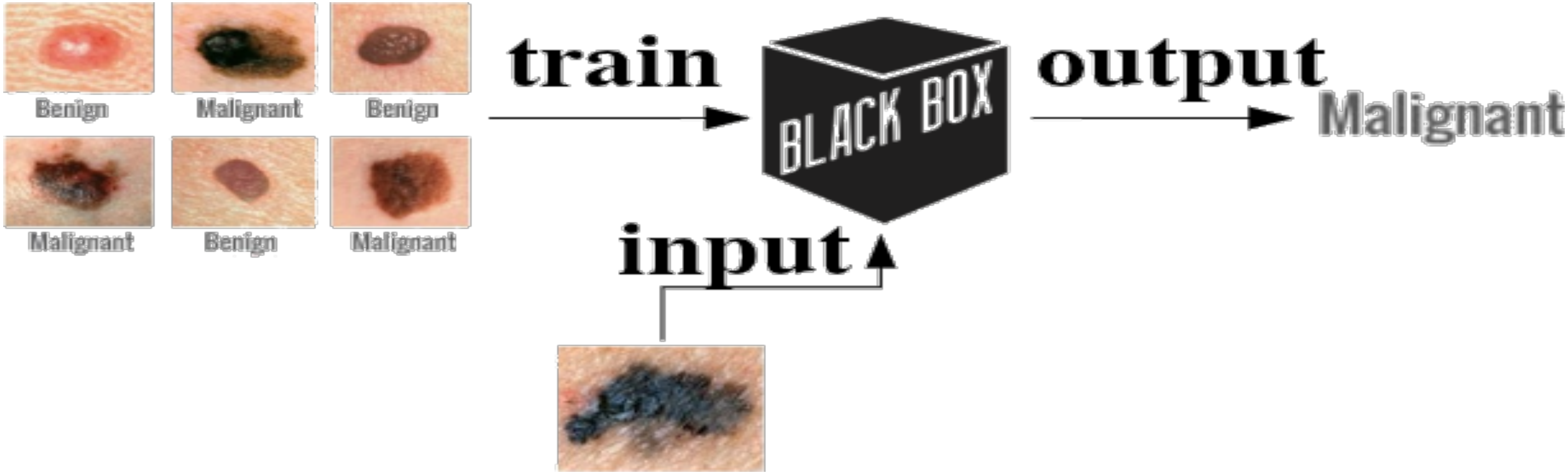
loomberg analyis of data from Amazon.com
merican Community Survey

AI = Machine Learning + Big Data

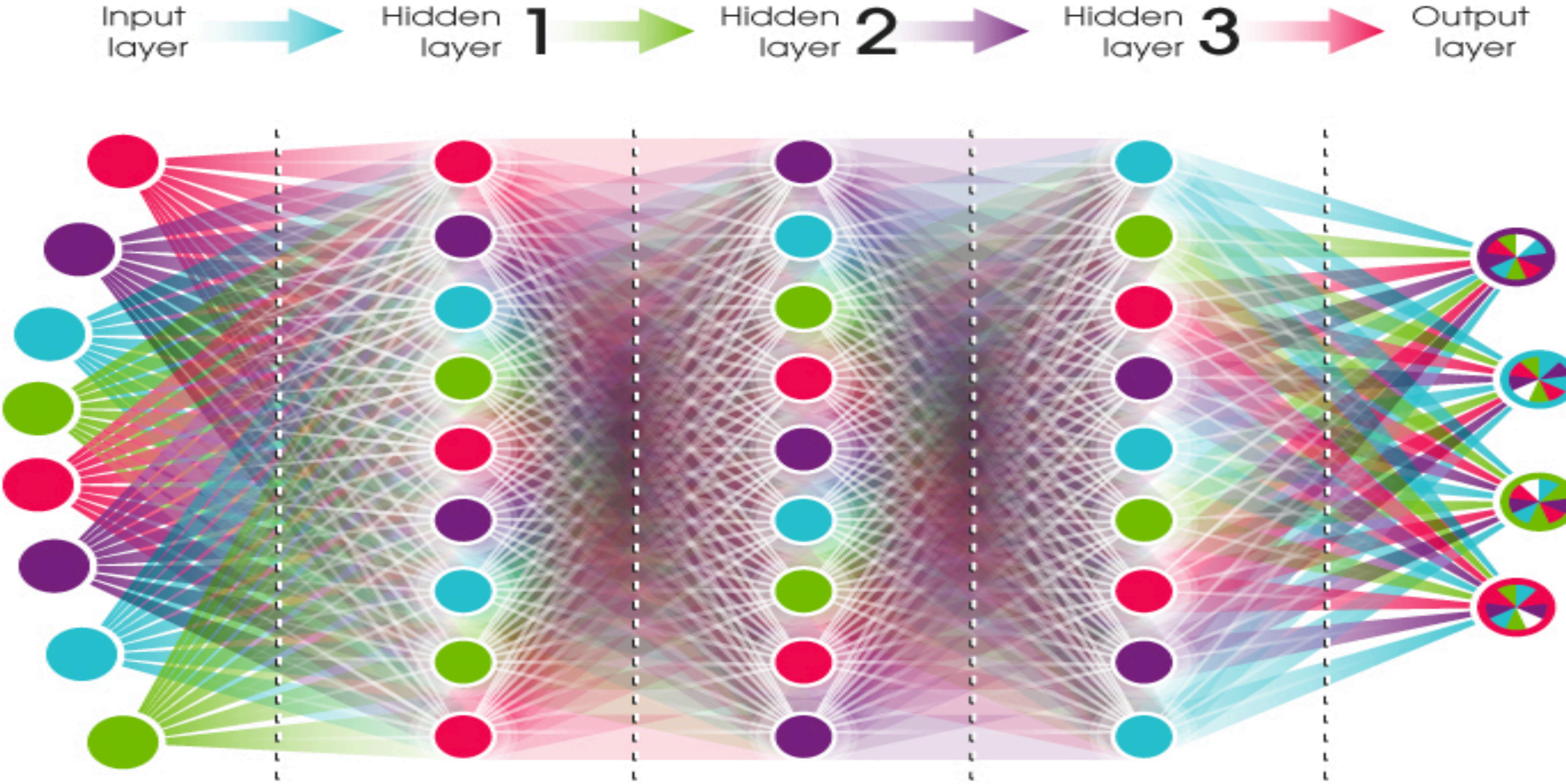



1
benign
.65 .35
100%
yes
UniformityCellSize < 2.5
no
2
benign
.97 .03
60%
BareNuclei < 4.5
3
malignant
.16 .84
40%
UniformityCellShape < 2.5
7
malignant
.11 .89
37%
UniformityCellSize < 4.5
14
malignant
.31 .69
9%
BareNuclei < 2.5
4
benign
1.00 .00
58%
5
malignant
.33 .67
2%
6
benign
.80 .20
3%
28
benign
.80 .20
2%
29
malignant
.17 .83
7%
15
malignant
.04 .96
27%

AI = Machine Learning + Big Data

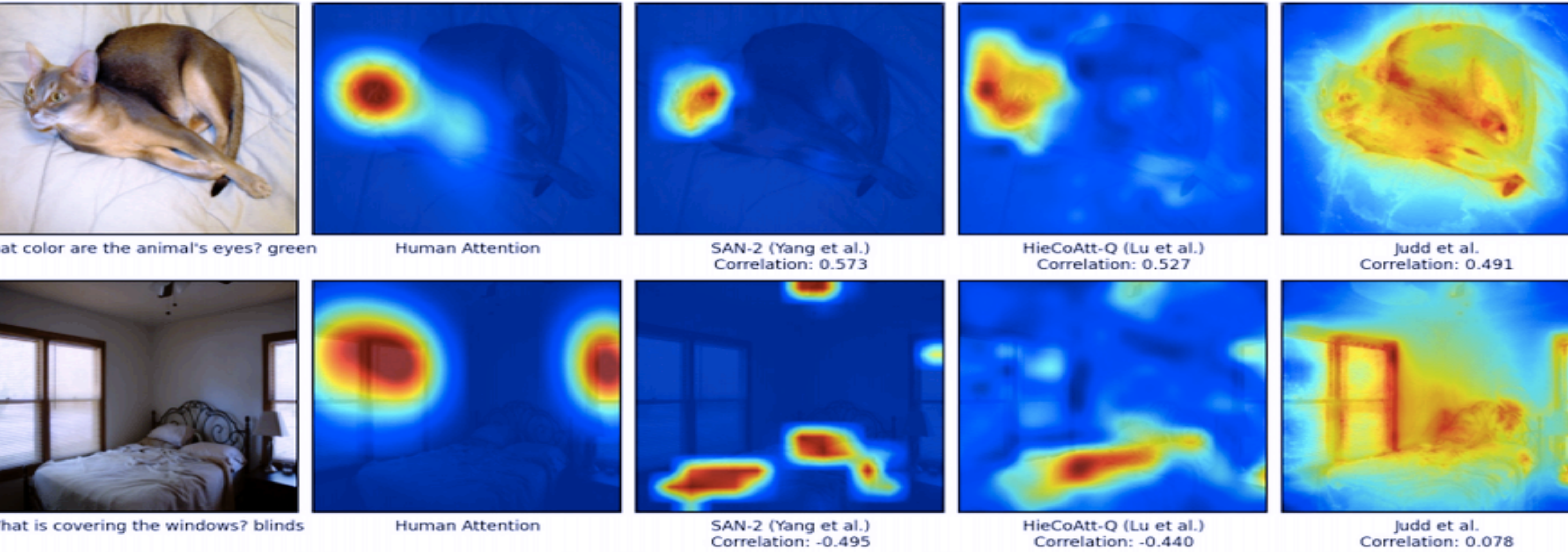


DEEP NEURAL NETWORK



alnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

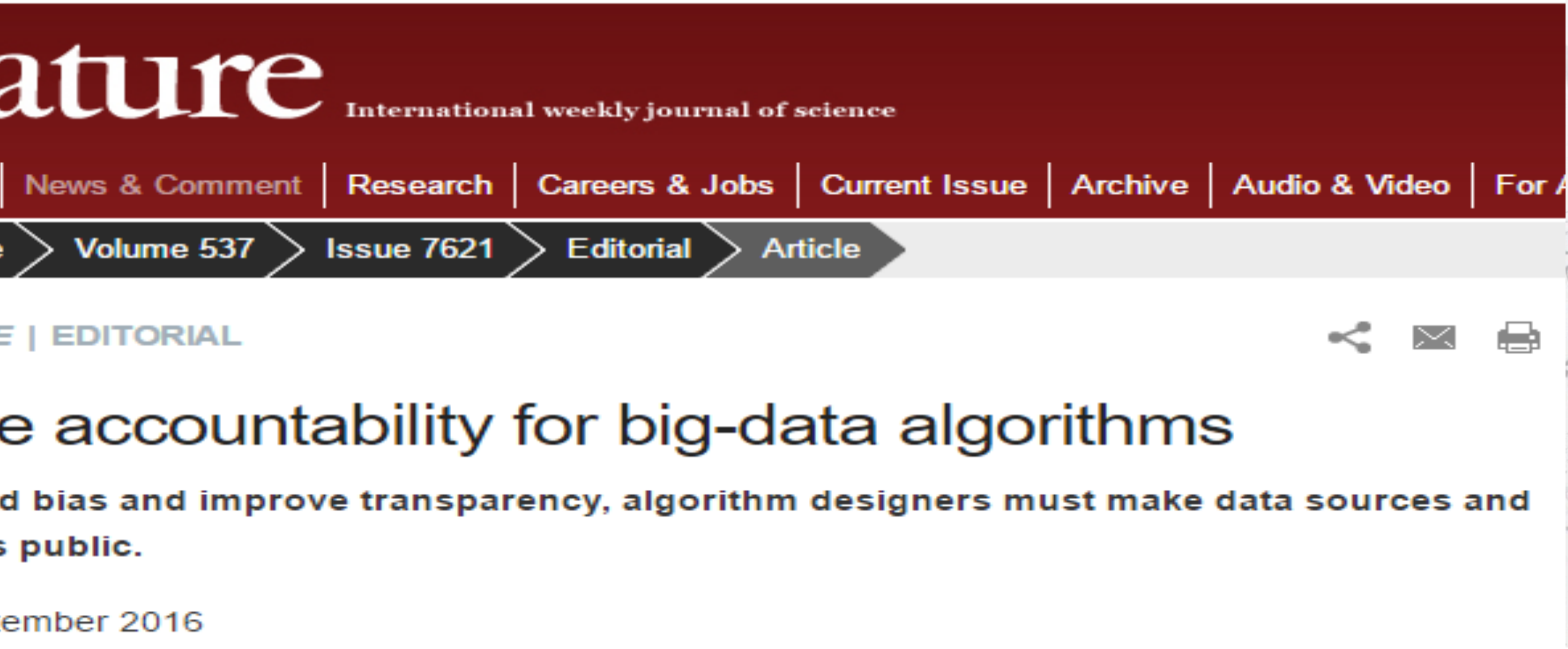
We risk to create algorithms we don’t fully understand



stems that recommend humans making a decision ould explain why

requires **transparency** and **responsibility**

E.g., in healthcare, justice, surveillance, predictive policing, autonomous vehicles, risk scoring, ...



ature International weekly journal of science

News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

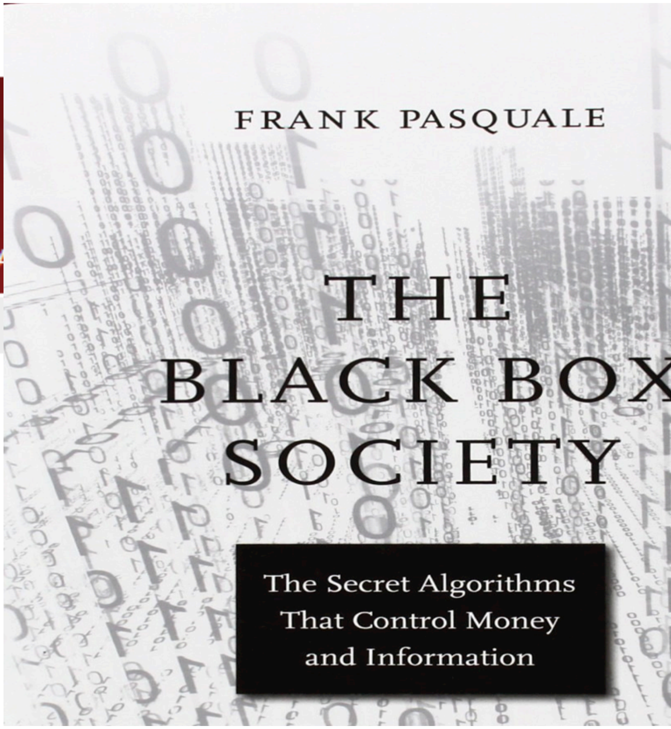
e > Volume 537 > Issue 7621 > Editorial > Article

E | EDITORIAL

e accountability for big-data algorithms

d bias and improve transparency, algorithm designers must make data sources and s public.

ember 2016



FRANK PASQUALE

THE BLACK BOX SOCIETY

The Secret Algorithms That Control Money and Information

ght of explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain “**meaningful explanations of the logic involved**” when “**automated (algorithmic) individual decision making**”, including profiling, takes place.