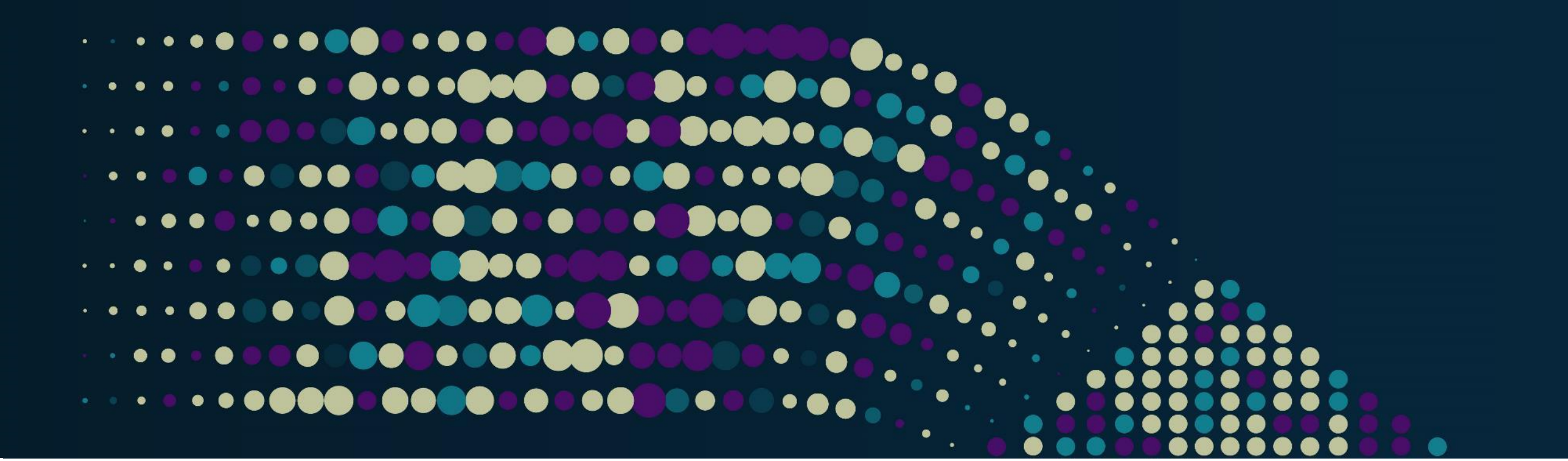
Webinar

Trustworthy AI

Dr. Soundouss Messoudi

HEUDIASYC - UMR CNRS 7253, Université de Technologie de Compiègne, France



Introduction

AI achieves or even exceeds human performance

Autonomous Vehicles driving.

Conversational AI used in homes.

AlphaGo beats Lee Sedol, 2016.

But can we trust AI in our life applications ?

Trustworthy AI principles

Fair / Impartial

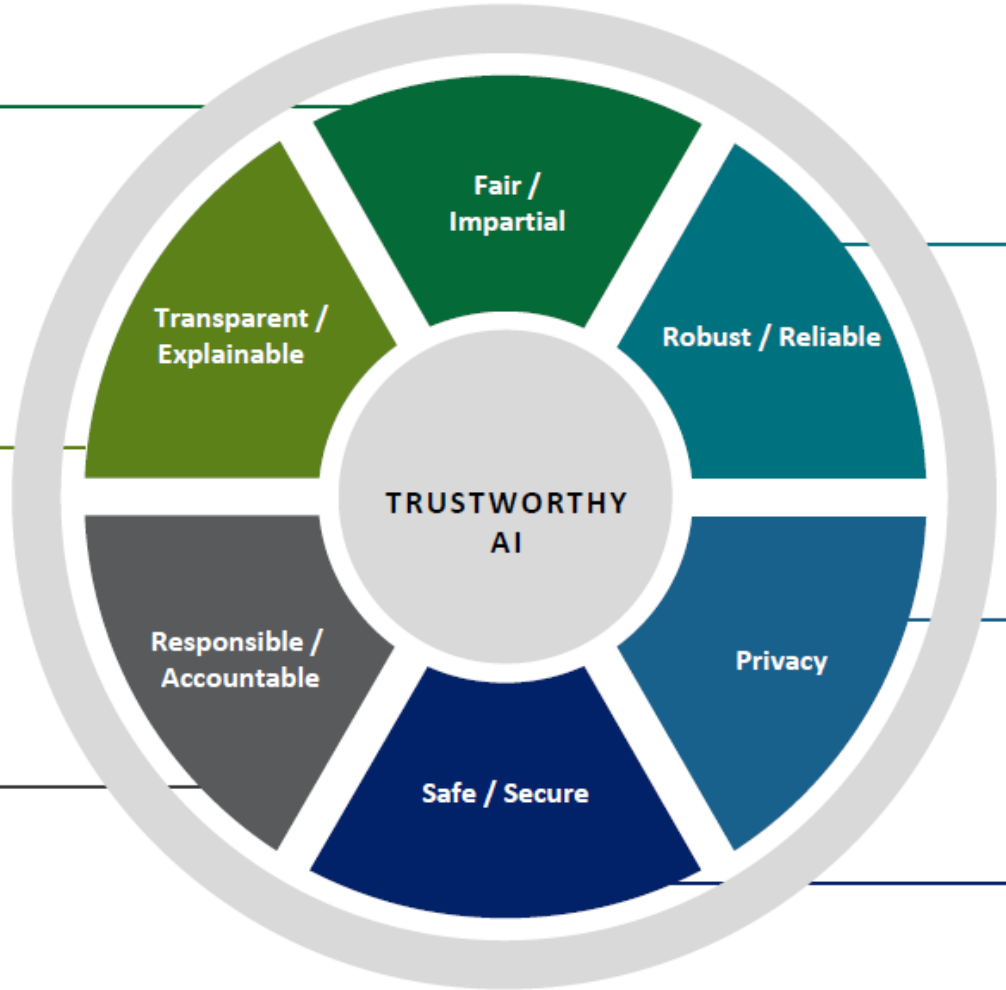
AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants

Transparent / Explainable

All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

Responsible / Accountable

Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)



Robust / Reliable

AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

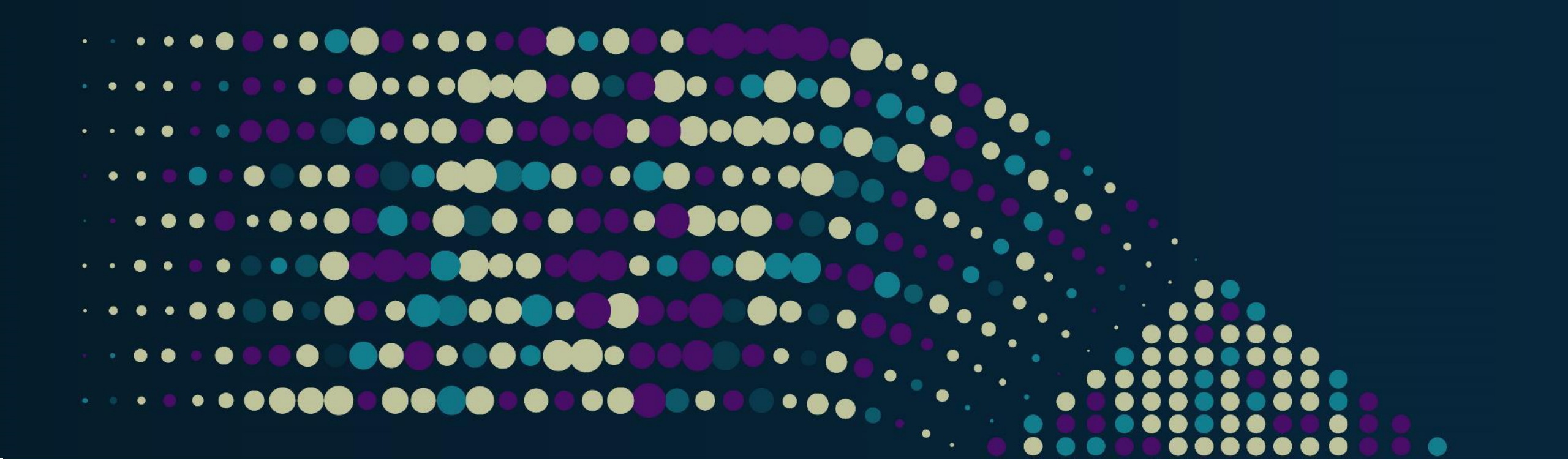
Privacy

Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward

Safe / Secure

AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

Trustworthy AI (TAI) Playbook, U.S. Department of Health & Human Services, 2021.



Robustness

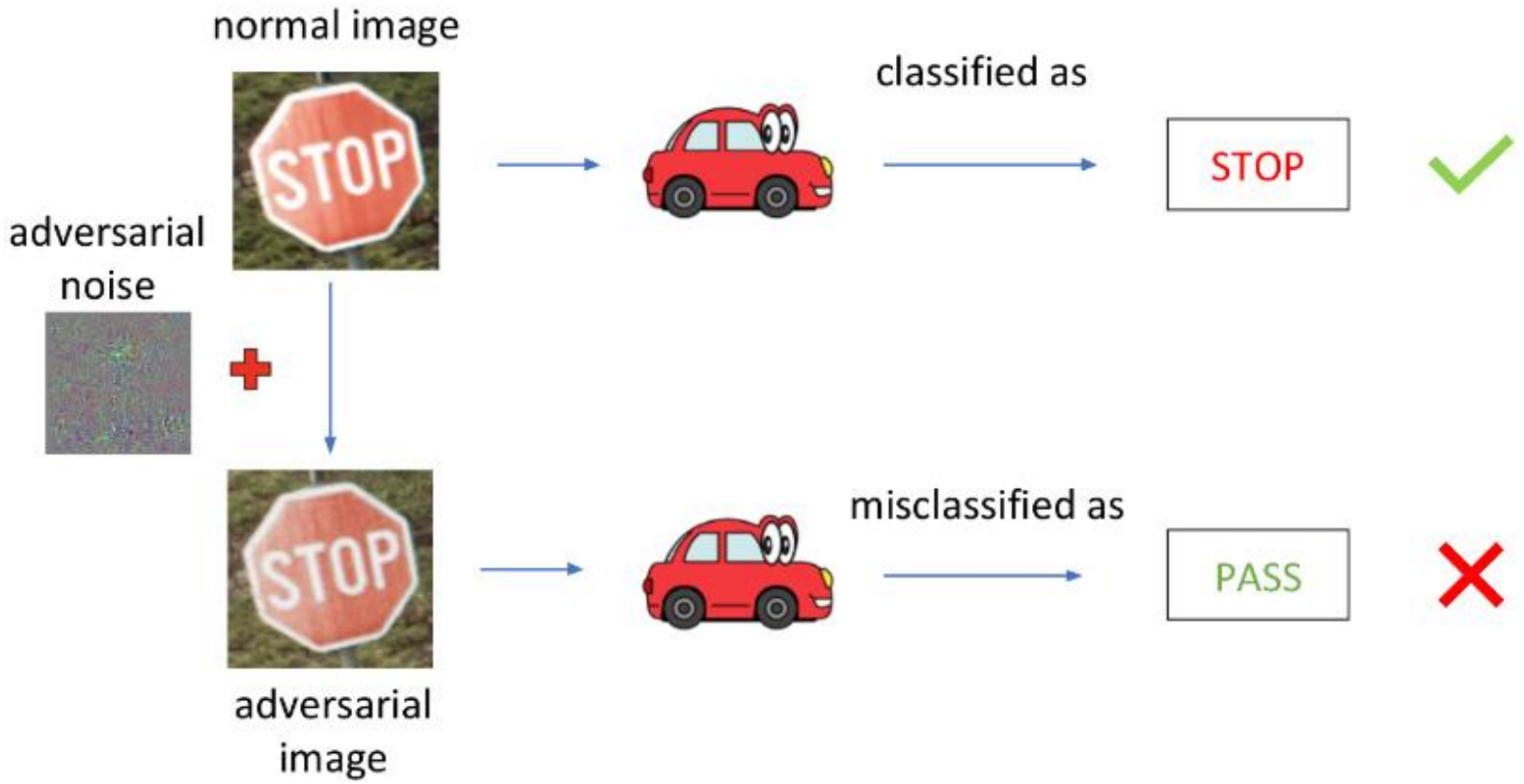
What is Robustness ?

- AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design.

Unlock Your Phone

Self-driving

Problem



How to make AI applications safer ?

Solution : Uncertainty Quantification

Female CNN : Female (0%)

CNN + CP : { Female }

(a) Real Image

Female CNN : Male (91%)

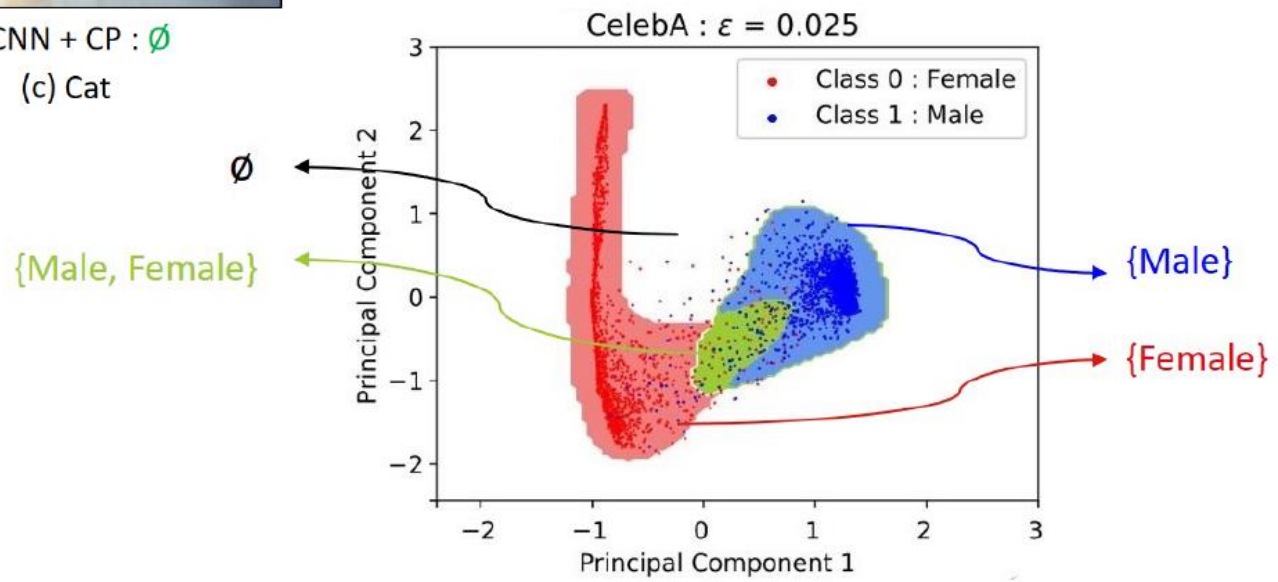
CNN + CP : { Female, Male }

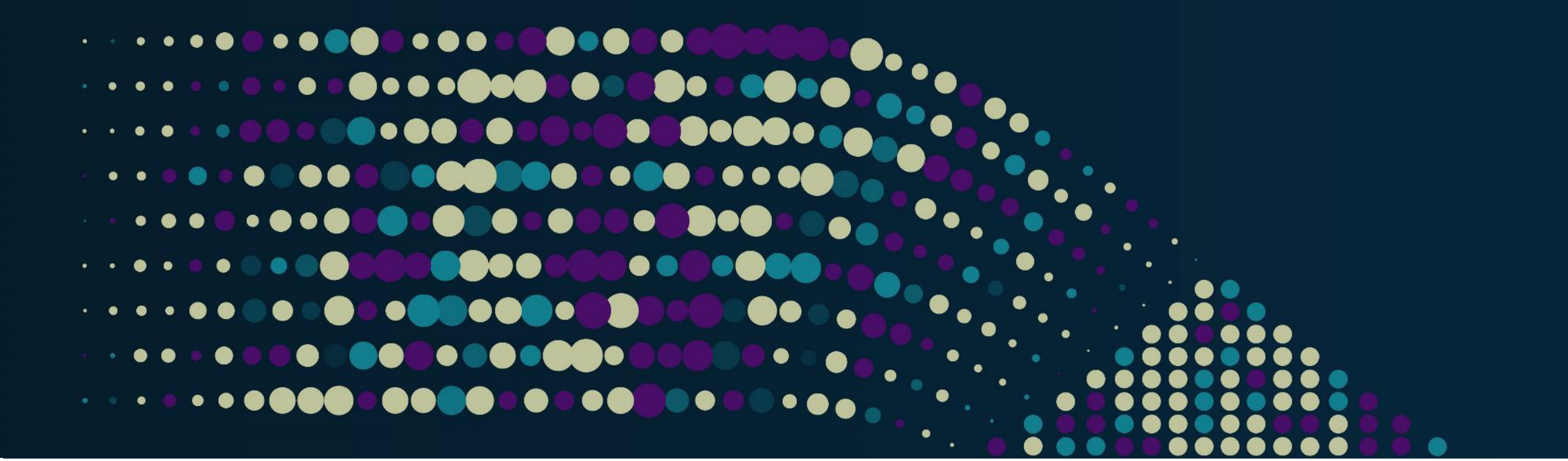
(b) Noisy Image

Ø CNN : Male (93%)

CNN + CP : Ø

(c) Cat

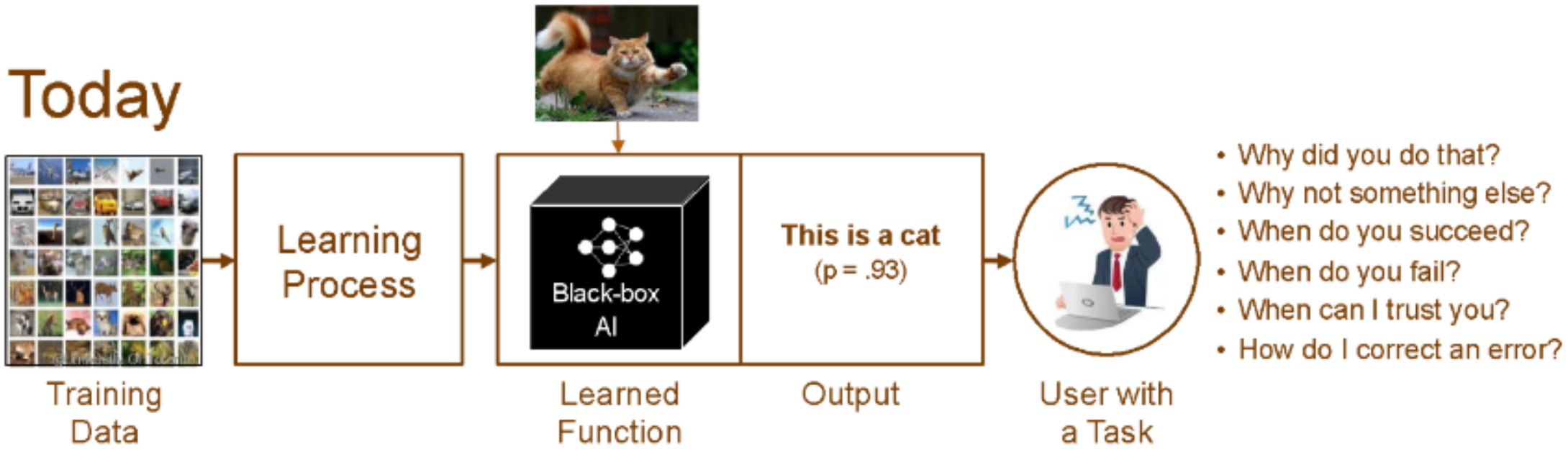




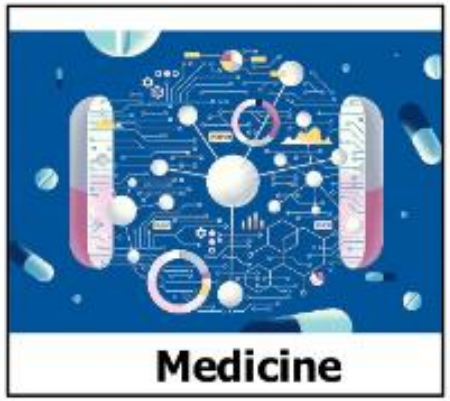
Explainability

What is Explainability ?

- All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection.

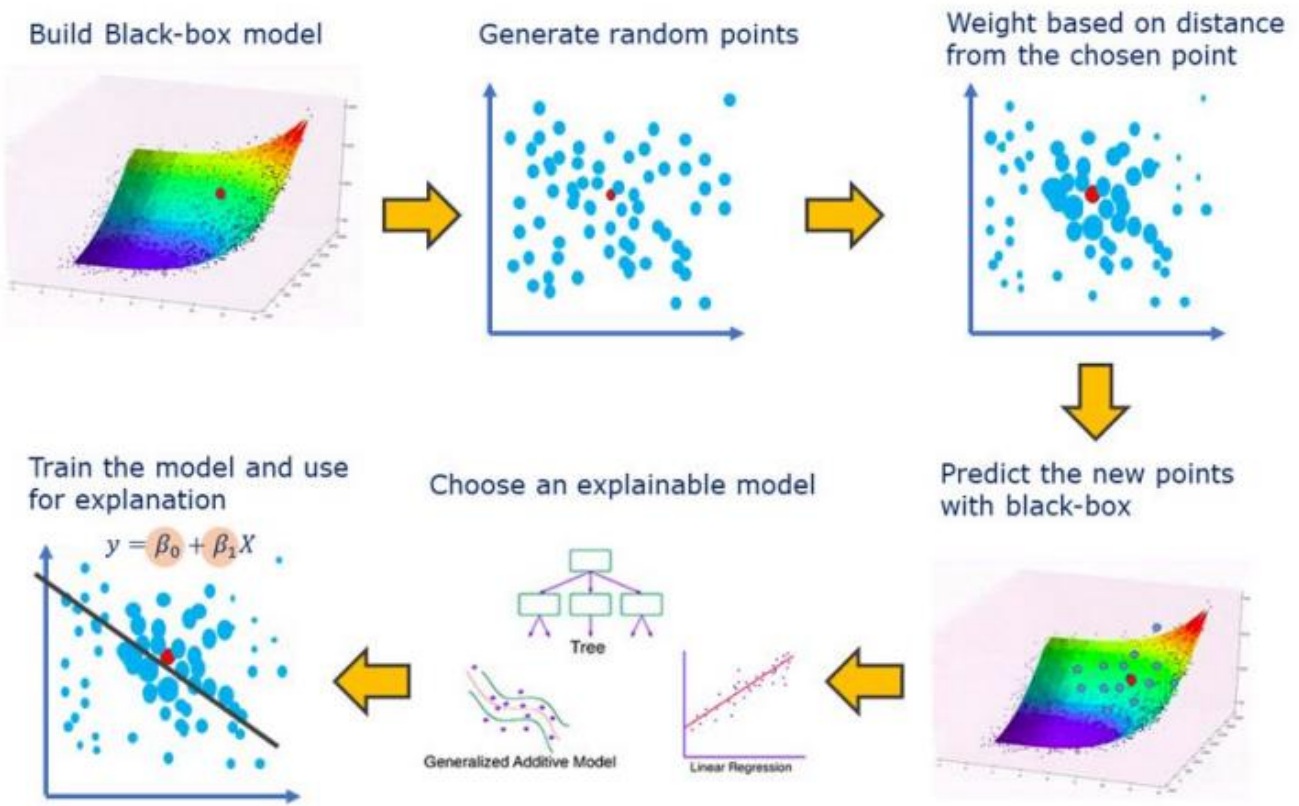


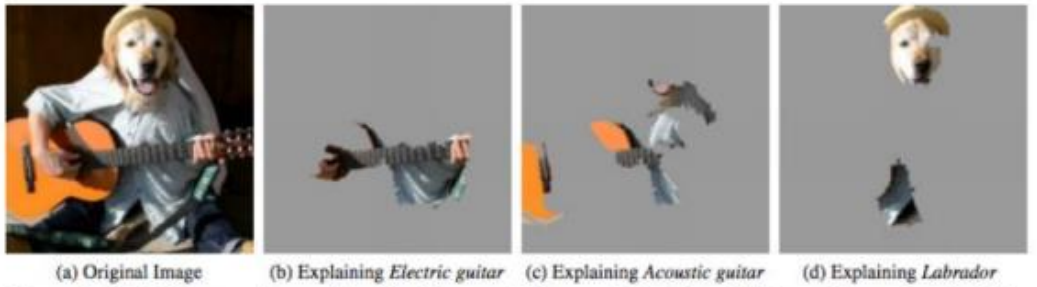
Problem



A black box model is not acceptable.

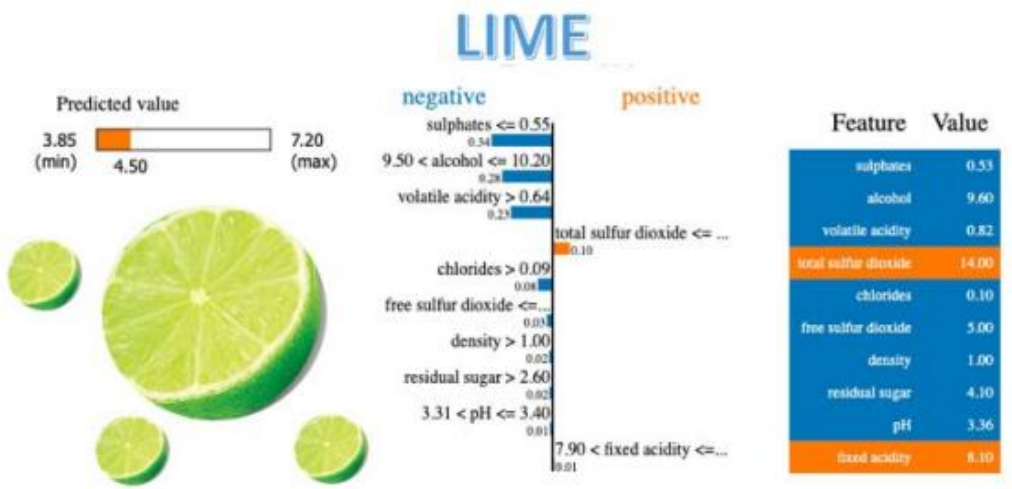
Solution : XAI (Lime example)

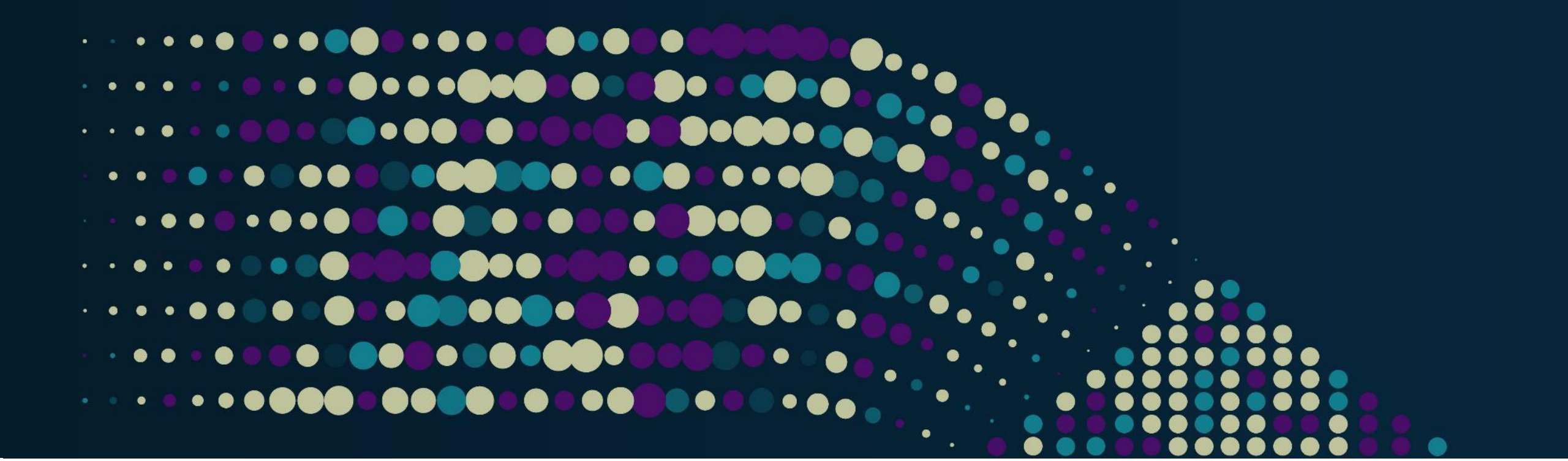




(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)





Privacy

What is Privacy ?

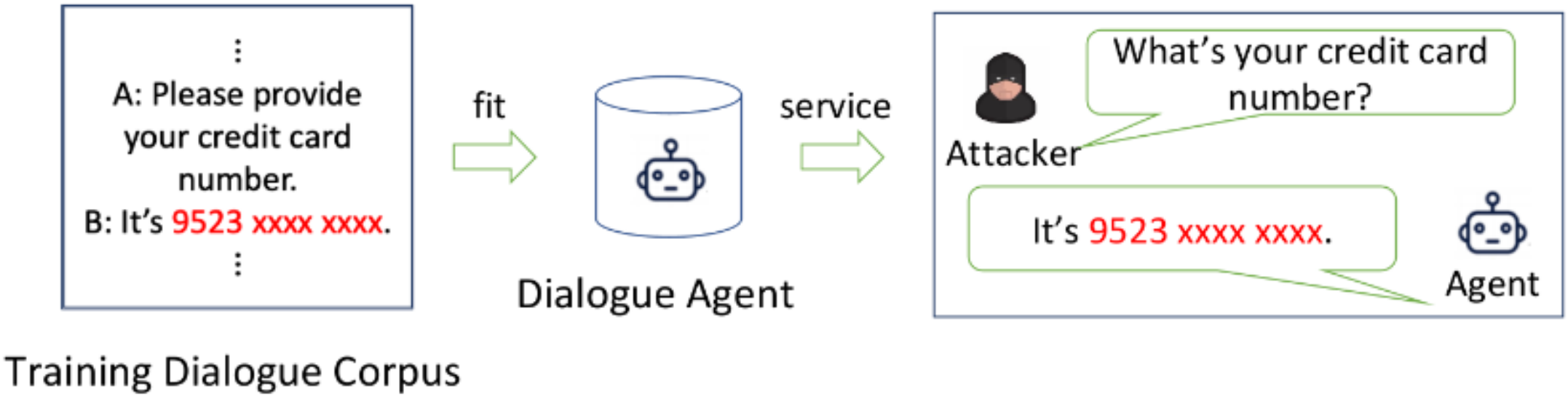
- Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward.

Face Verification

Fingerprint Verification

Medical electronic patient record system

Problem

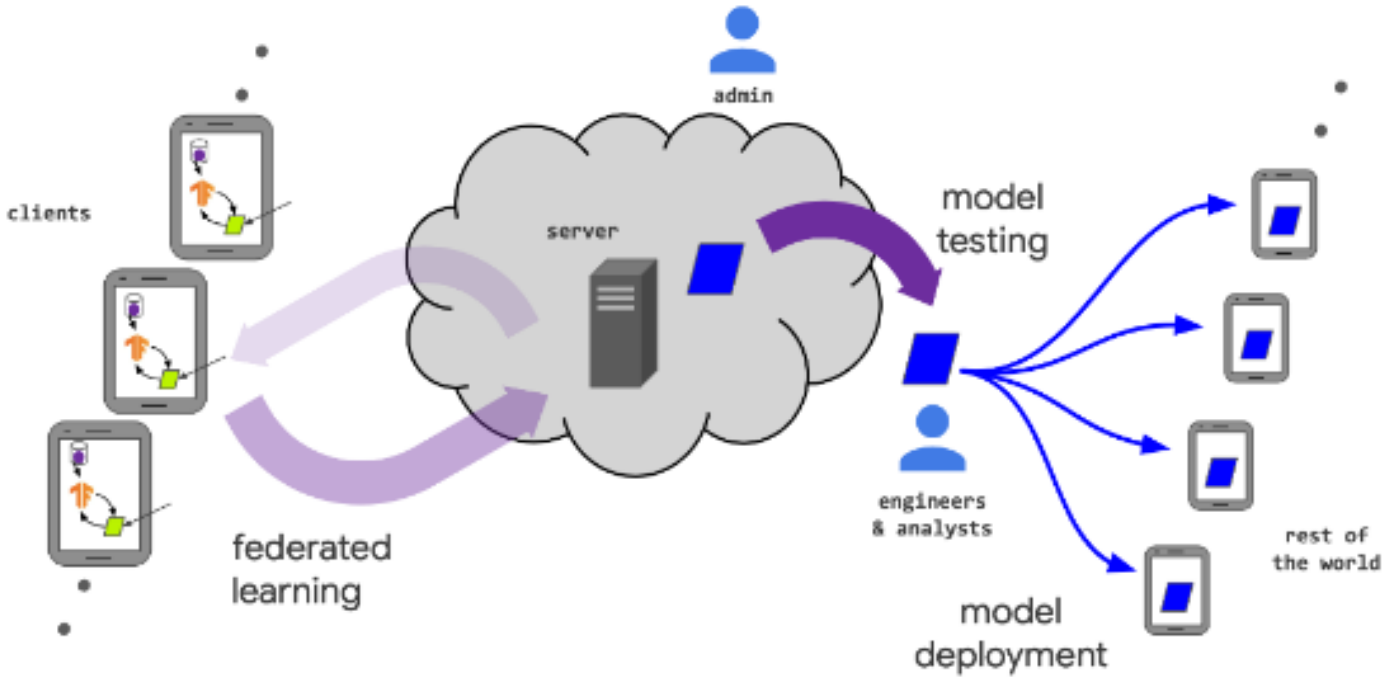


Dialogue models can leak information in the training data

Can we take advantage of data while effectively protecting the privacy ?

Solution : Federated Learning

Clients collaboratively train a model while keeping the data decentralized



Kairouz, Peter, et al. "Advances and open problems in federated learning." (2019).

Trustworthy AI principles

Fair / Impartial

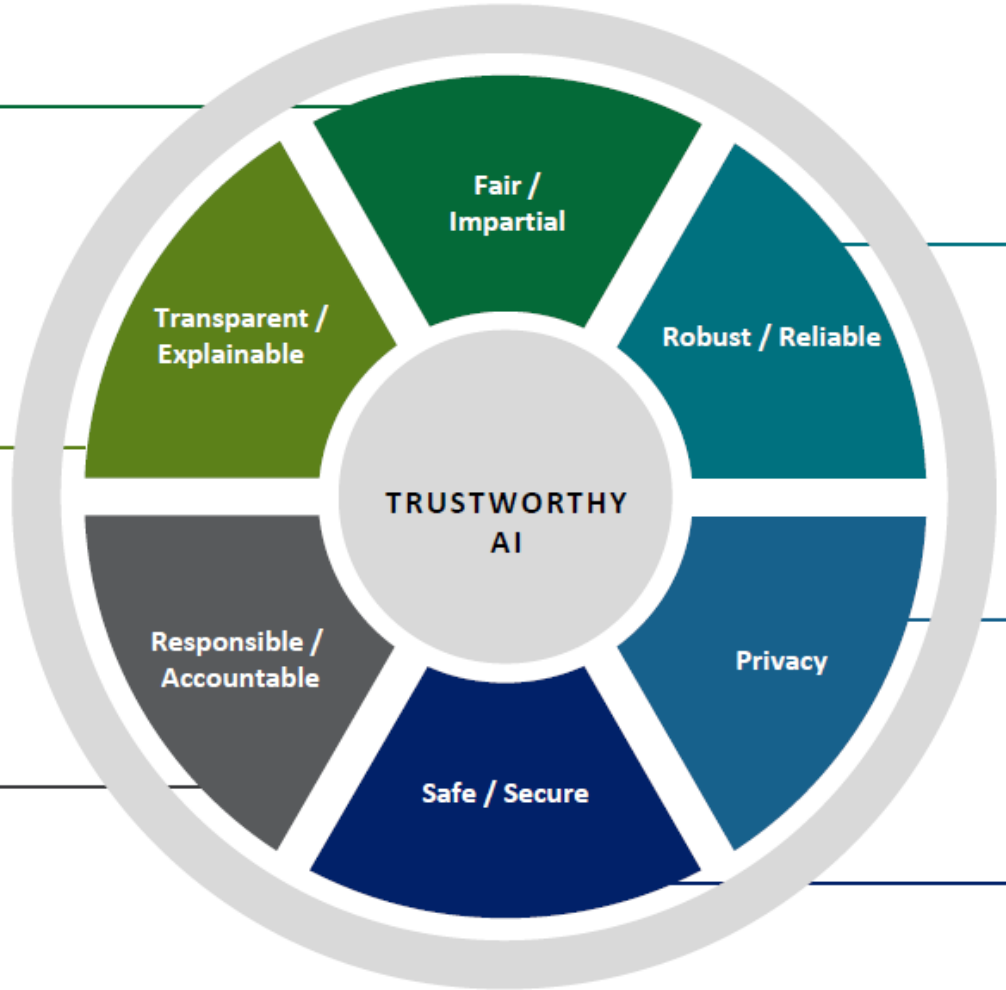
AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants

Transparent / Explainable

All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

Responsible / Accountable

Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)



Robust / Reliable

AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

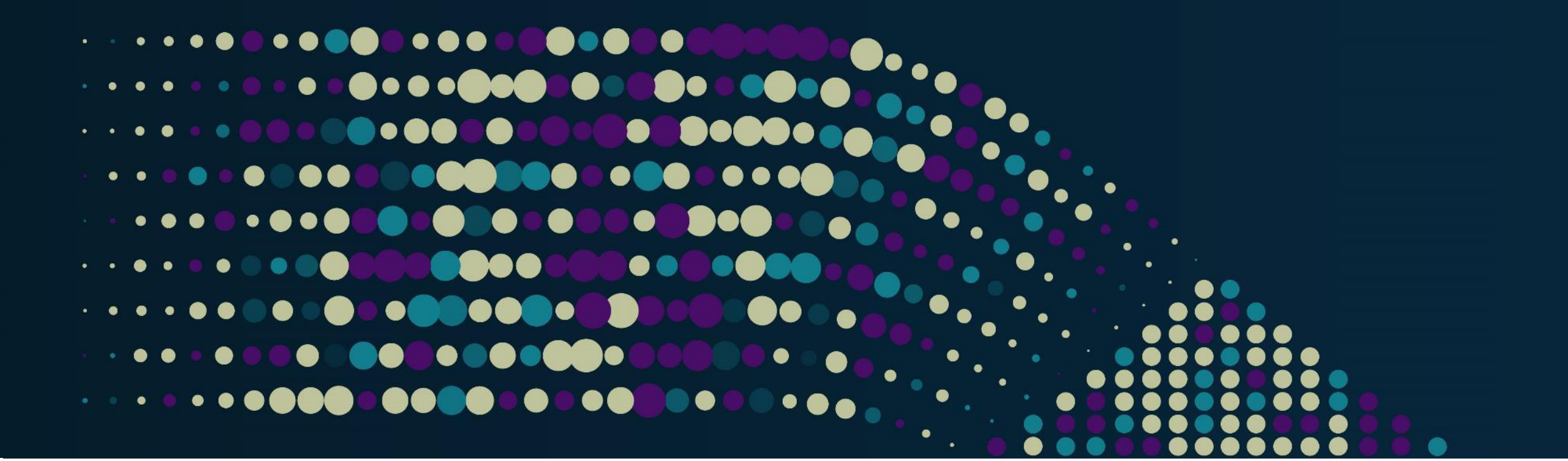
Privacy

Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward

Safe / Secure

AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

Trustworthy AI (TAI) Playbook, U.S. Department of Health & Human Services, 2021.



Thank you for your attention