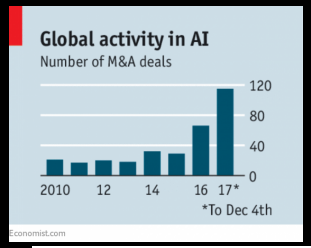
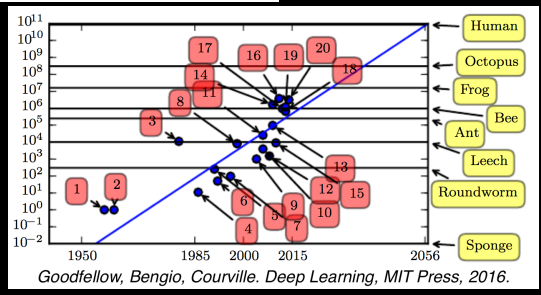
Interpreting Artificial Intelligences

Fabio G. Cozman
Universidade de São Paulo

Hot summer in AI

- Pragmatic victory:
  banking, commerce, medicine,
  agriculture, industry;
  artistic, social and political debate.





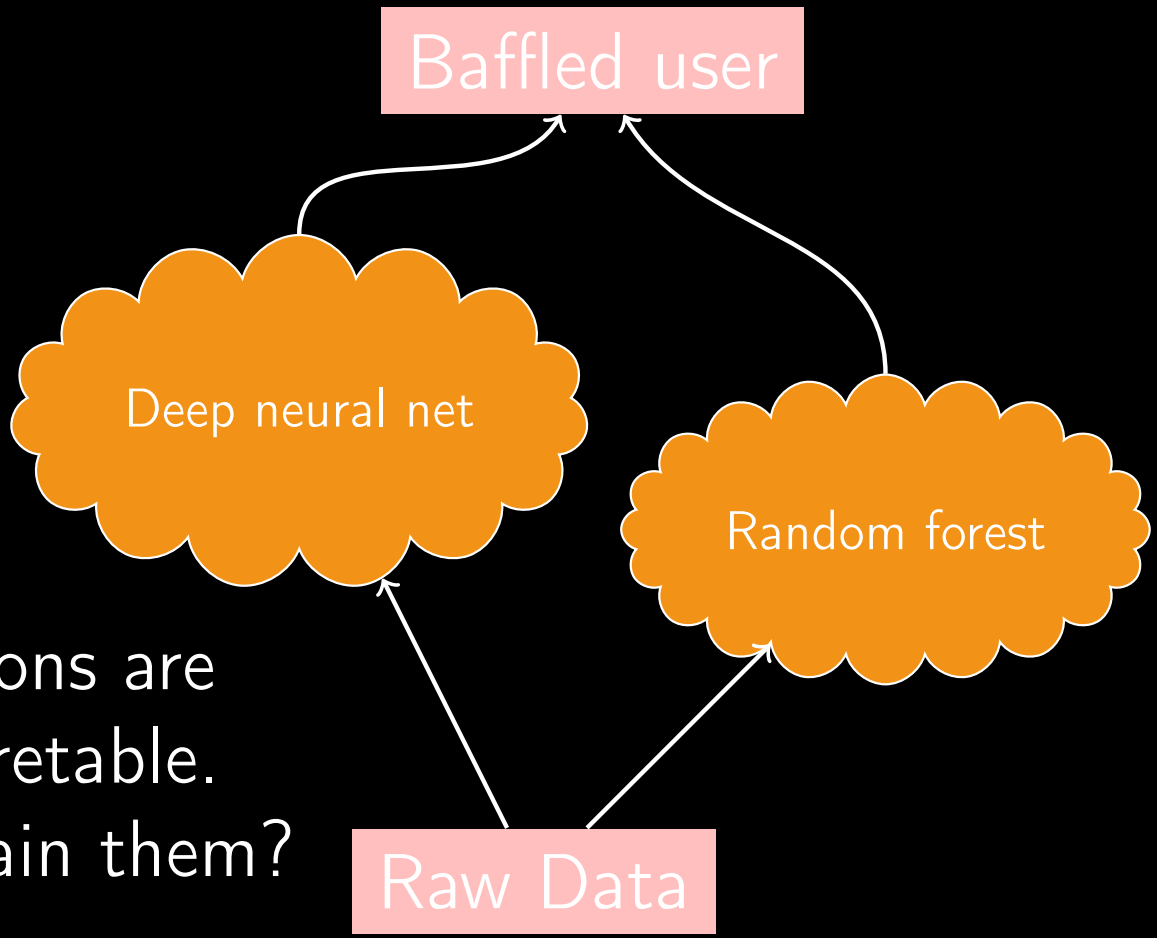
- Why? More power, more data, many insights...

And back to old questions

- Old one: is AI possible?
- Old one: is AI dangerous?
  ...new twists:
  - Robust.
  - Easy to use.
  - Fair.
  - Ethical.
  - Transparent.
  - **Interpretable.**
  - **Explainable.**
  - Human-friendly and job-neutral.

Interpreting/Explaining ML

- Often, machine learning resorts to very complex models.



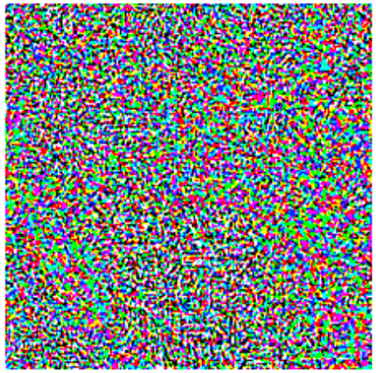
But decisions are not interpretable.
How to explain them?

Weird mistakes

“panda”

57.7% confidence

$+ .007 \times$



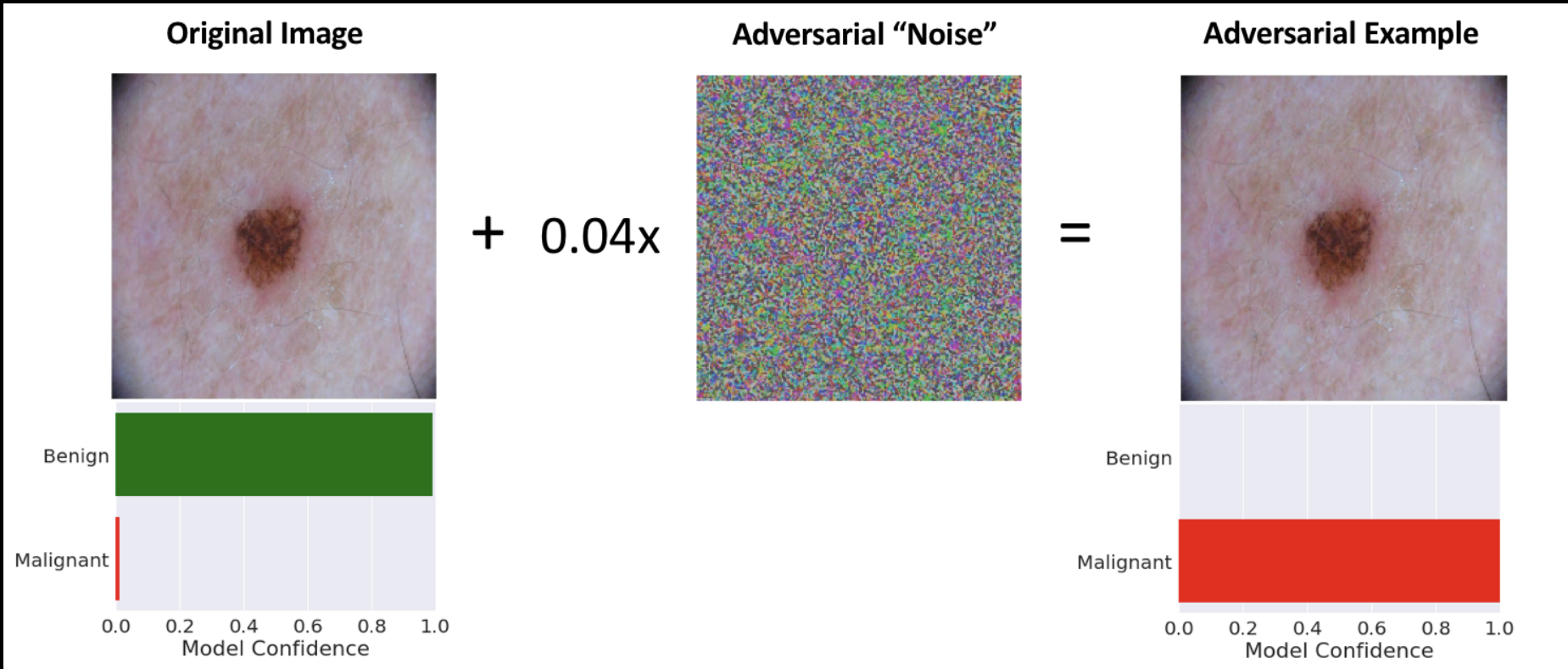
noise

$=$

“gibbon”

99.3% confidence

Weird mistakes



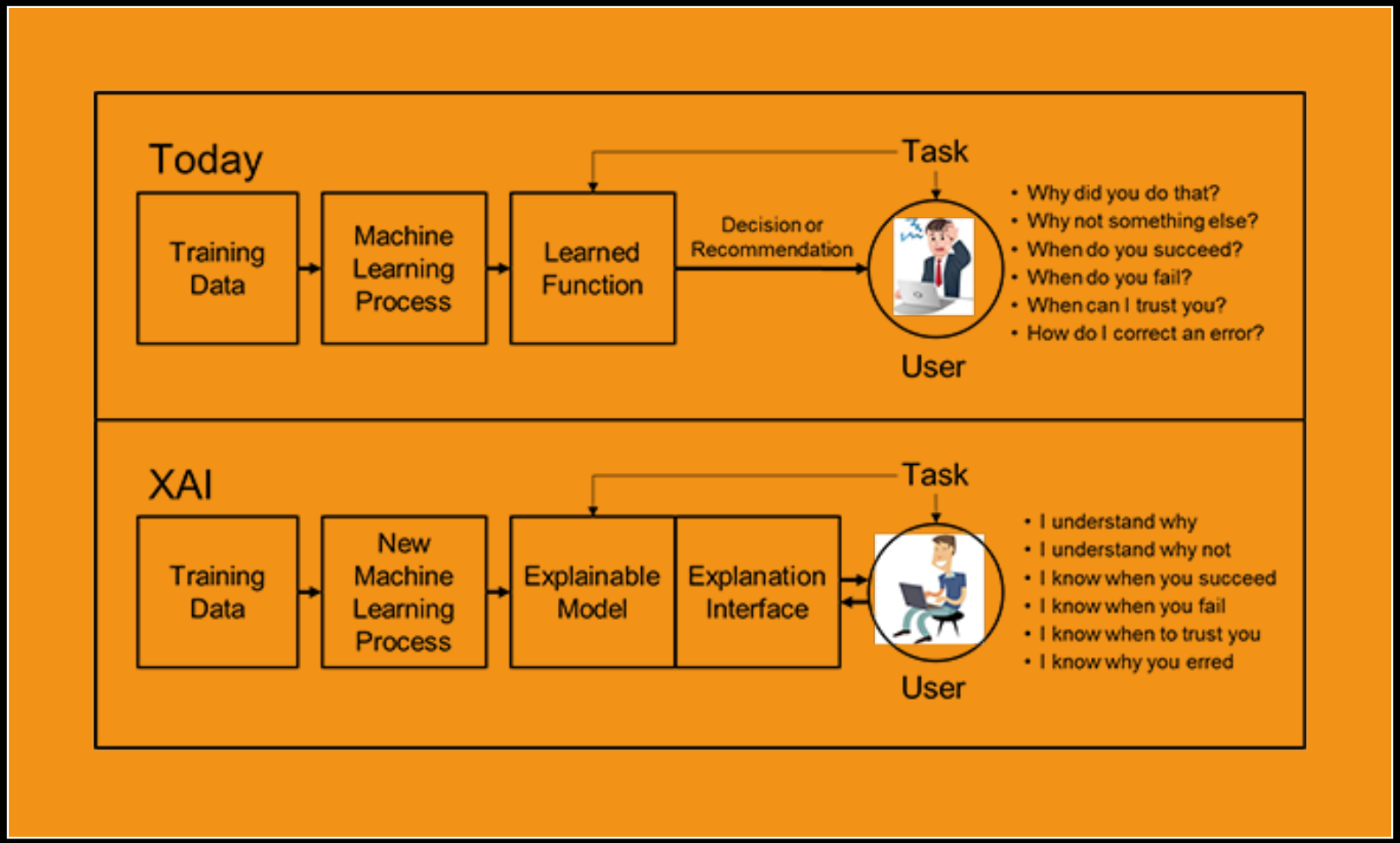
Interpretability

- A possible definition (no real consensus so far):

  Interpretability is the degree to which a human can understand the cause of a decision.

- Interpretability depends on user, on common knowledge, on system interface, on the kind of application...
  - High-stakes scenarios may require explanations.
  - Recommendation systems may truly benefit from engaging interactions.

DARPA's Explainable AI



Legal demands

- Since 25 May 2018, the General Data Protection Regulation (GDPR - EU) establishes right to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making” takes place.

Main idea about Explainable AI

- Goal is to explain the decision made by classifier, not to *justify* whether decisions themselves are good or bad.
- We wan to to expose bad decisions; to detect them so as to fix them.
- We want users to understand the decisions so that they can trust the (reasonable) ones.

Obvious idea: Use interpretable language

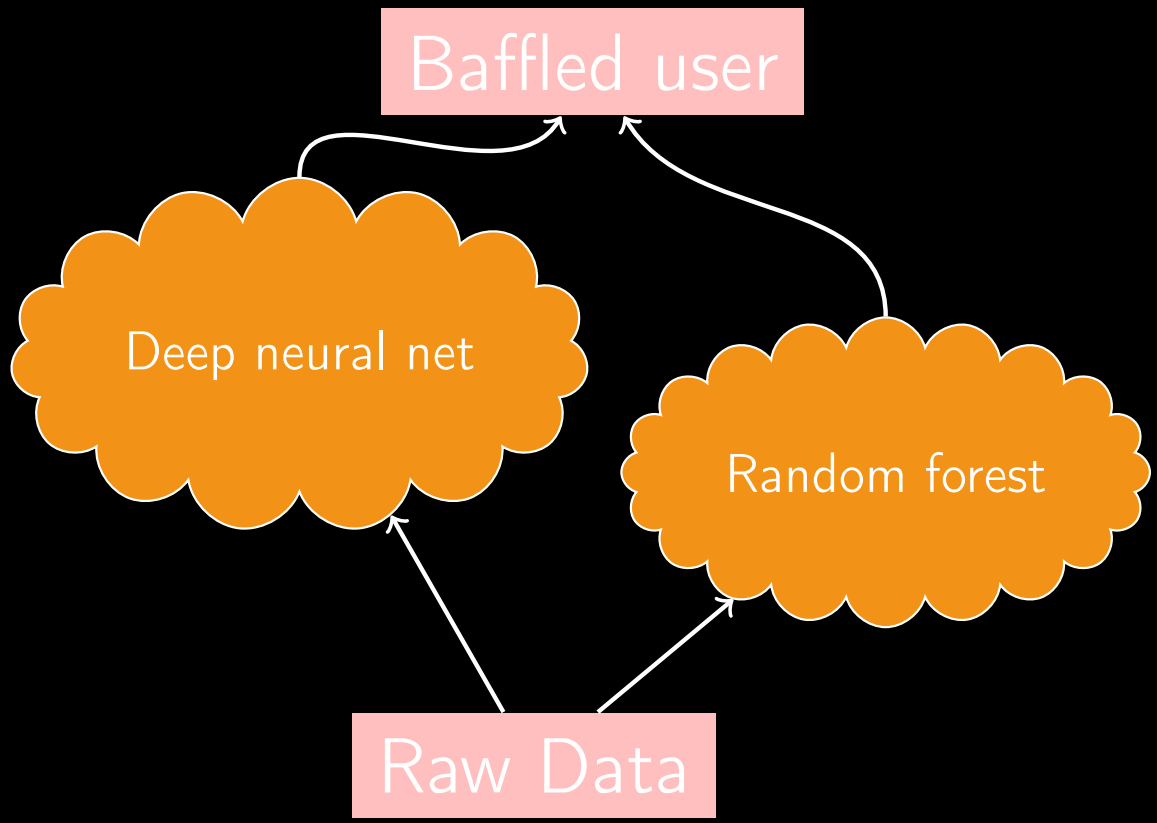
- Simple technique such as decision tree, nearest-neighbor.
  - Problem: tension between accuracy and interpretability.
- Complex model with interpretable semantics.
  - Perhaps a model from which explanations can be generated.

That is,

- Simple models are (often) interpretable.
- Complex models are (sometimes) explainable.

- For all other cases, we must generate explanations.

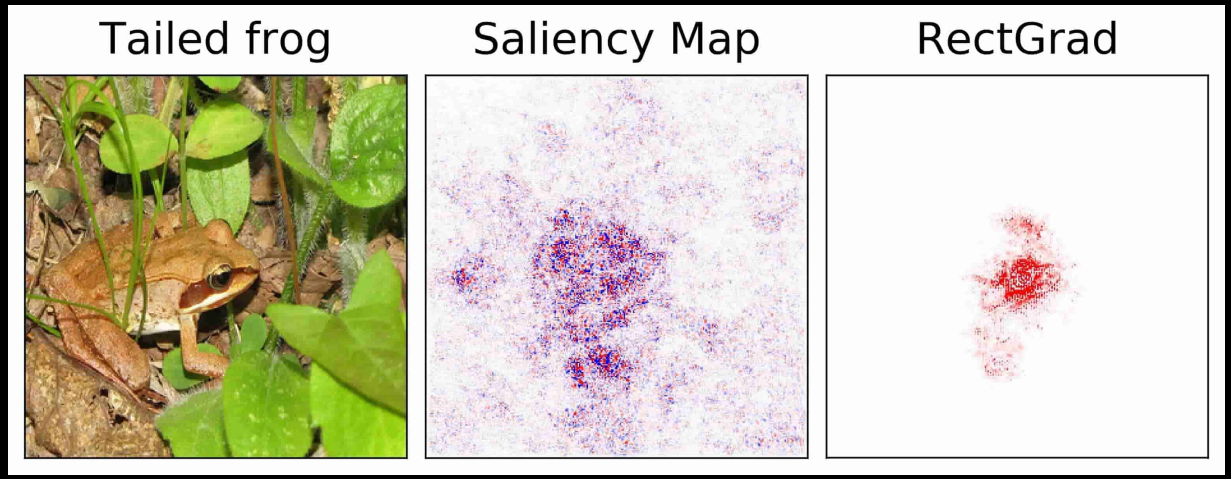
Explainable AI: some strategies

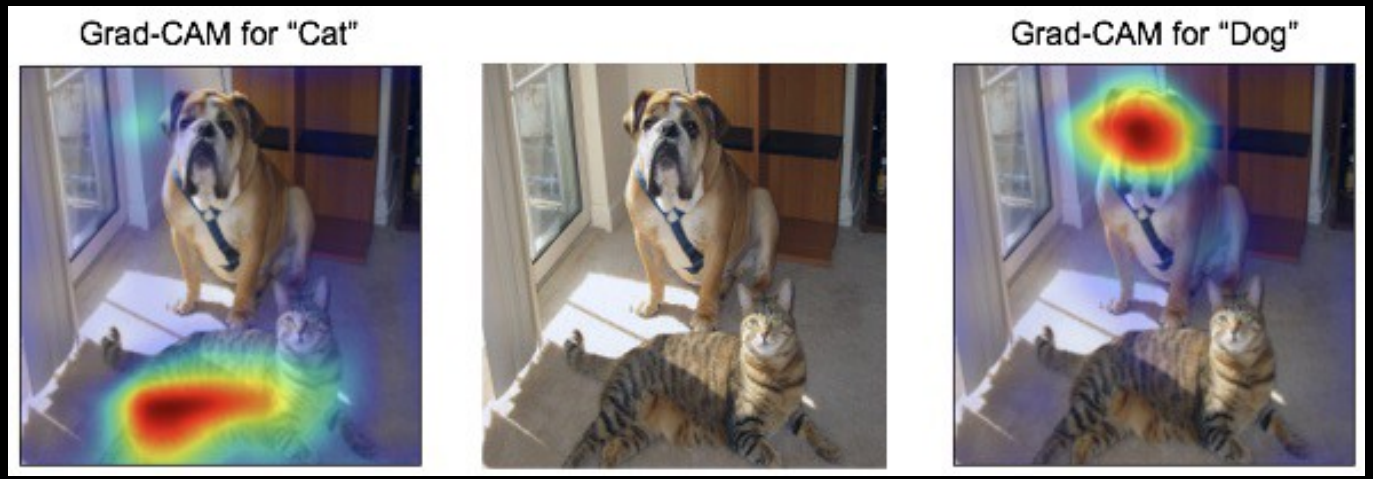


- After the model is built: sensitivity analysis (determine most relevant feature / datapoint)
- Another strategy: decompositional schemes.
- Yet another strategy: model-agnostic schemes.

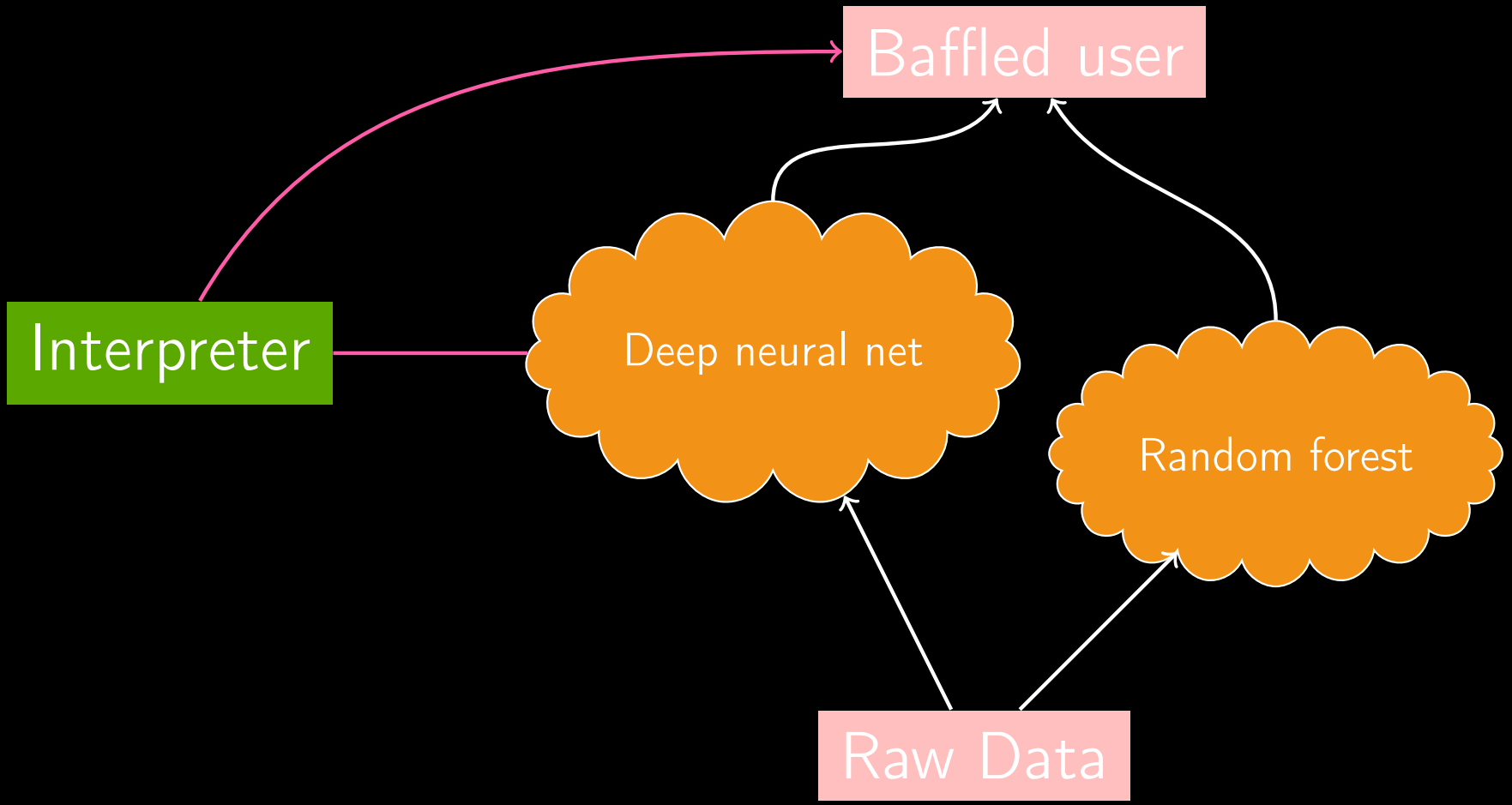
Understanding neural networks

- Saliency maps, Grad-CAM.

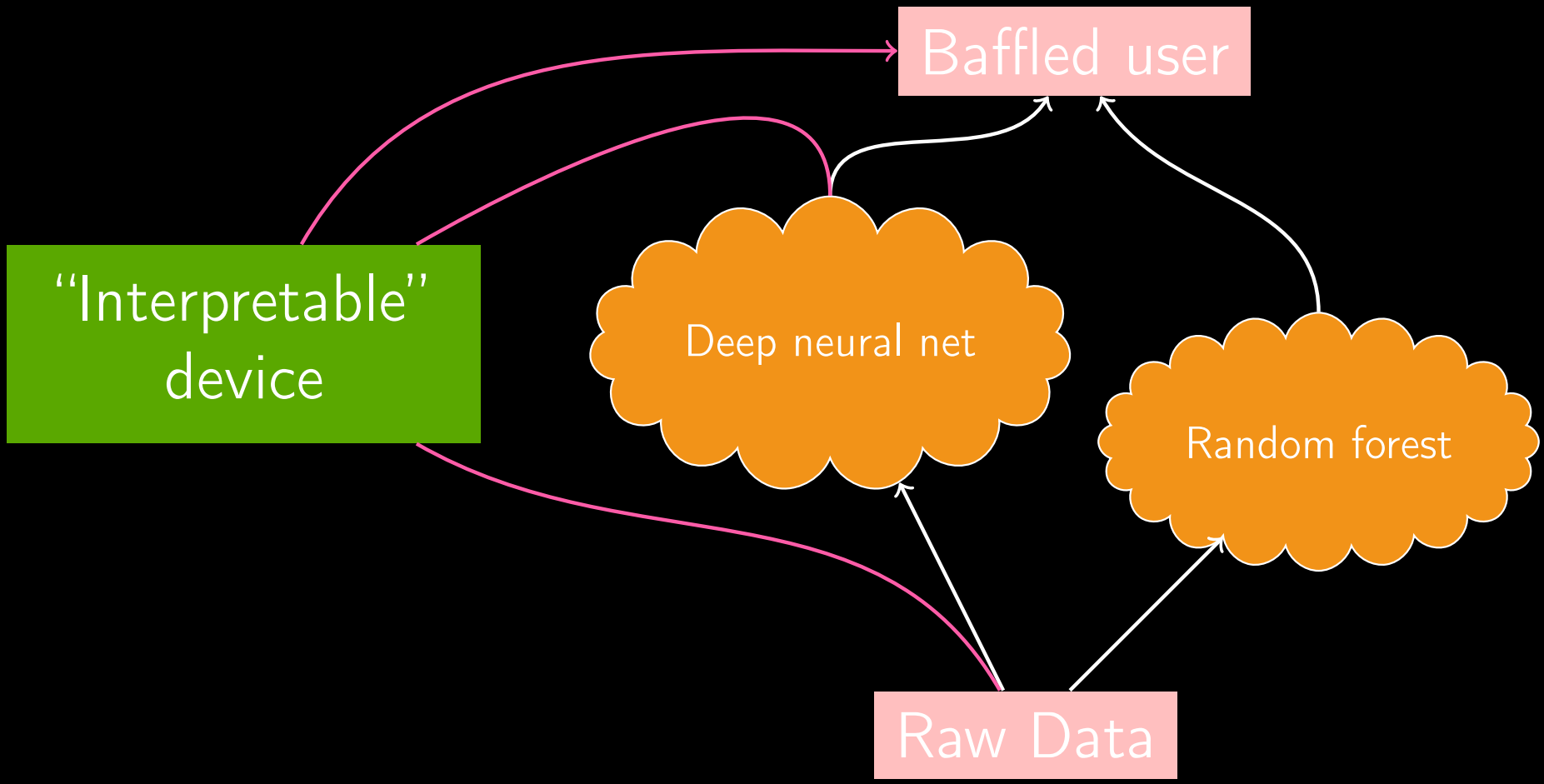




One strategy: Decompositional



Another strategy: Model agnostic

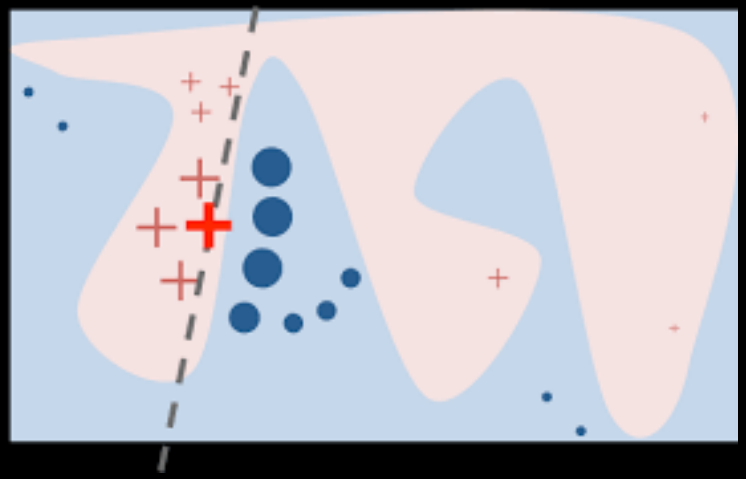


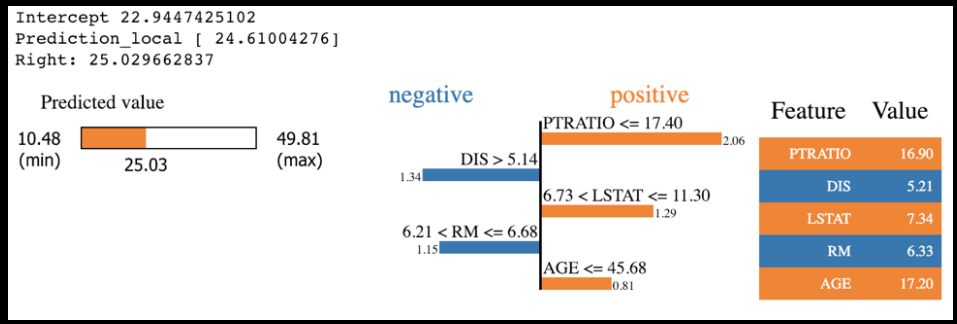
Local / global

- Local strategies explain a particular decision.
- Global strategies explain the behavior of the whole model.

Local Interpretable Model-agnostic Explanations

- Points generated around decision to be explained.
- Interpretable classifier adjusted locally.
- LIME package automates it.





Conclusion

- Current success of AI brings new challenges; among them, interpretability / explainability.
- There are many different concepts and strategies.
- Several challenges remain open!

Thanks.