Analysis of Trustworthiness in Machine Learning and Deep Learning

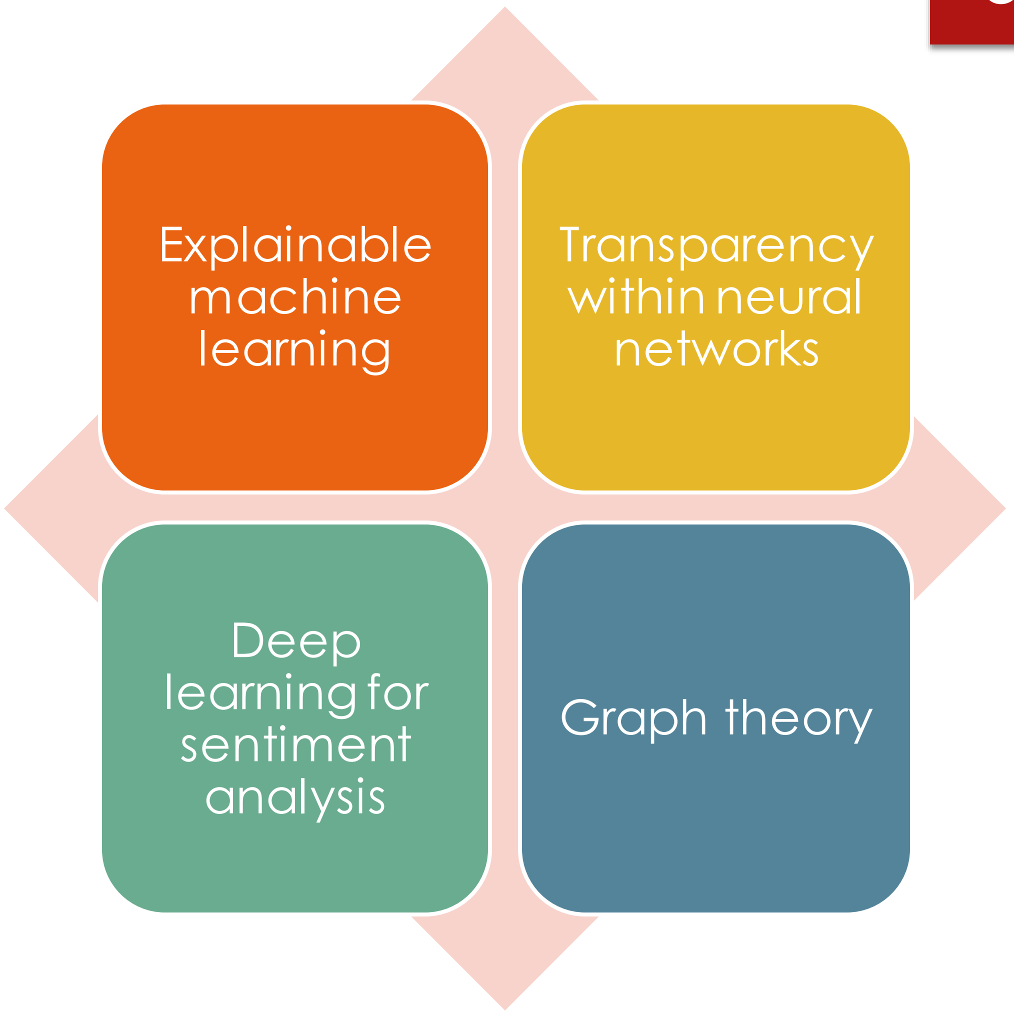
- **AUTHORS**: MOHAMED KENTOUR, JOAN LU
- **PRESENTER:** MOHAMED KENTOUR
  - University of Huddersfield, UK
  - Email: mohamed.kentour@hud.ac.uk

Presenter's short Bio

- MSc Software Engineering and information Processing, University of Boumerdes, Algeria(2018)
- MSc Computing, University of Huddersfield, UK (2020)
- Currently doing a PhD in Computer Science and Informatics, University of Huddersfield, UK

Website: [KentMoh/GitHub.uk]

Research Interest



Plan

Introduction

- Overview
- Objectives

Background

- Interpretable vs Explainable ML models
- Case of DL
- Limits

Insights

- Model decomposition
- Bring users' perceptual metrics into the learning flow

Demonstration

Legal concerns

Conclusion and future work

1. Introduction

- Data deluge and decision making [1]
  - Performance vs transparency
  - Need for transparency

- Data-science life cycle [2]
  - Trustworthiness within ML and DL life cycle

- Users' behavior changing
- Users' cognitive level

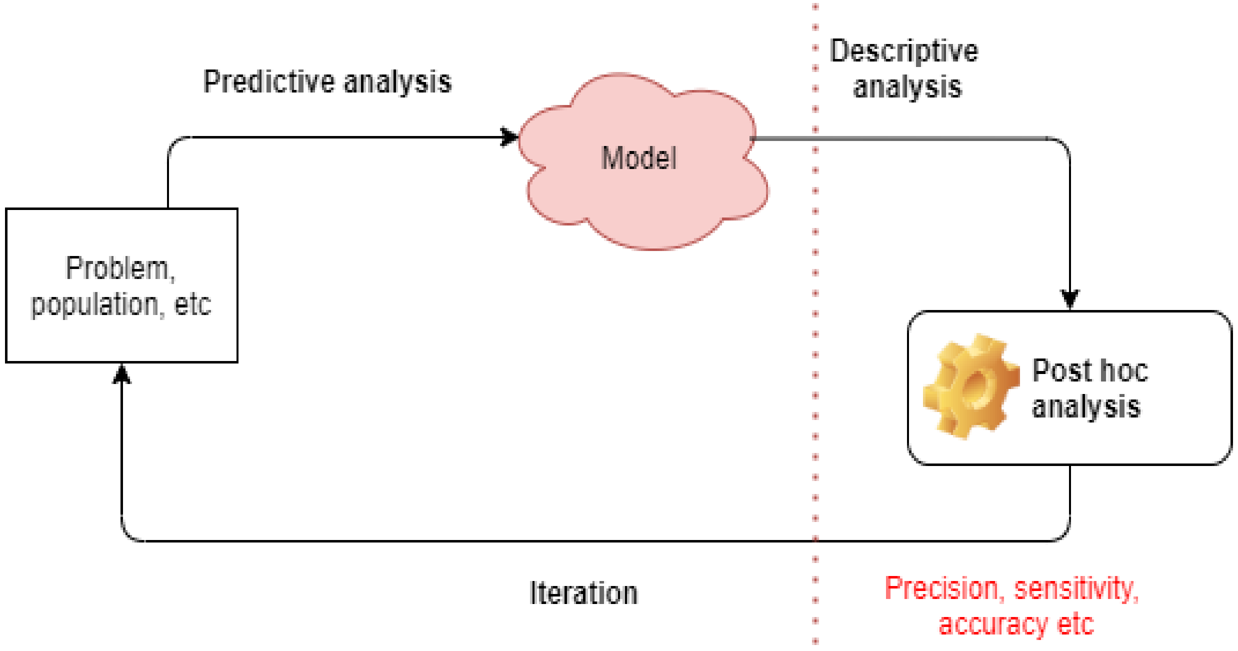


Figure 1. Data-science life cycle.

1. Introduction

1.1 Overview

- New data sampling
- New perceptual dimension
- Adjust to the users' requirements

1.2 Objectives

- Analyse literature models
- Show the impact of the perceptual metrics on models' performance
- Increase the trustworthiness of the model through a demonstration

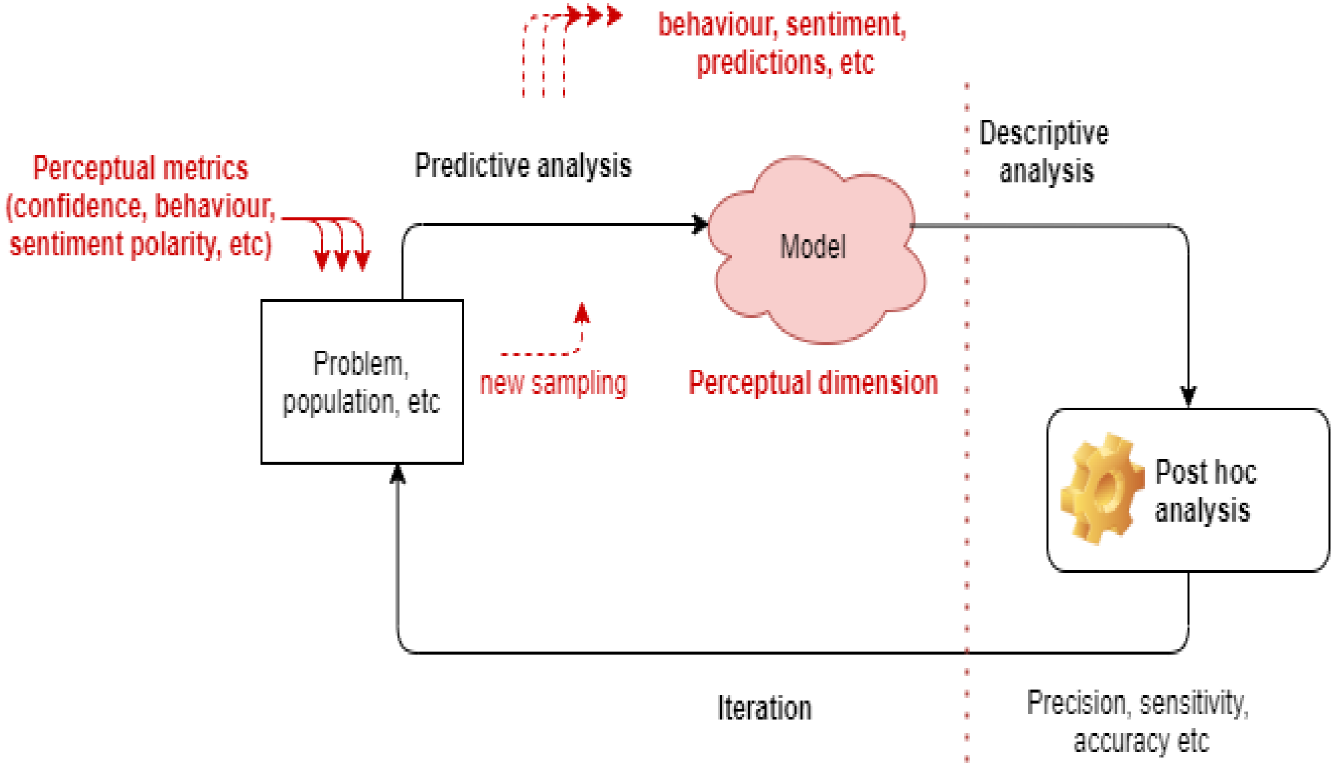


Figure 2. Perceptual metric's inclusion within ML life cycle.

2. Background

2.1 Interpretable ML

- Post deployed analysis

-> Challenge

- Bridge learning theory with quantifiable metrics

-> Solutions

- Matrix factorization [knowledge, method]
- Fuzzy System and ontology for decision trees
- Invoke explainable models (LIME, COVAR) to measure quantifiable metrics

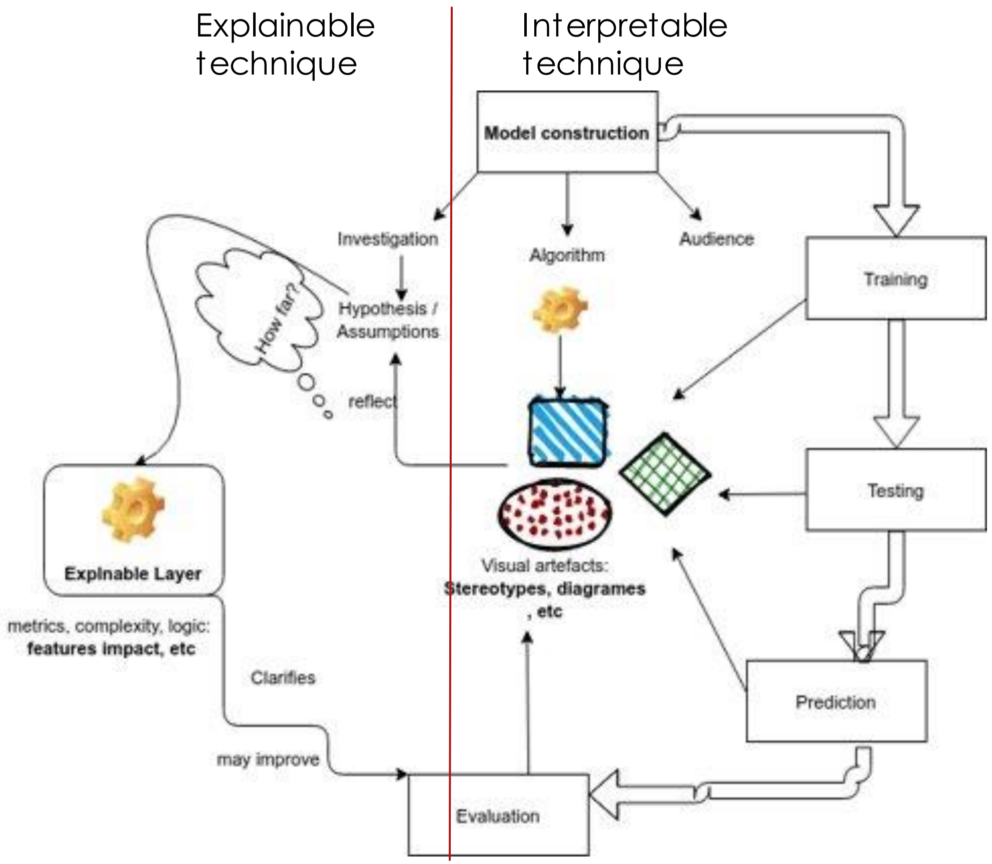


Figure 3. Interpretable vs Explainable ML.

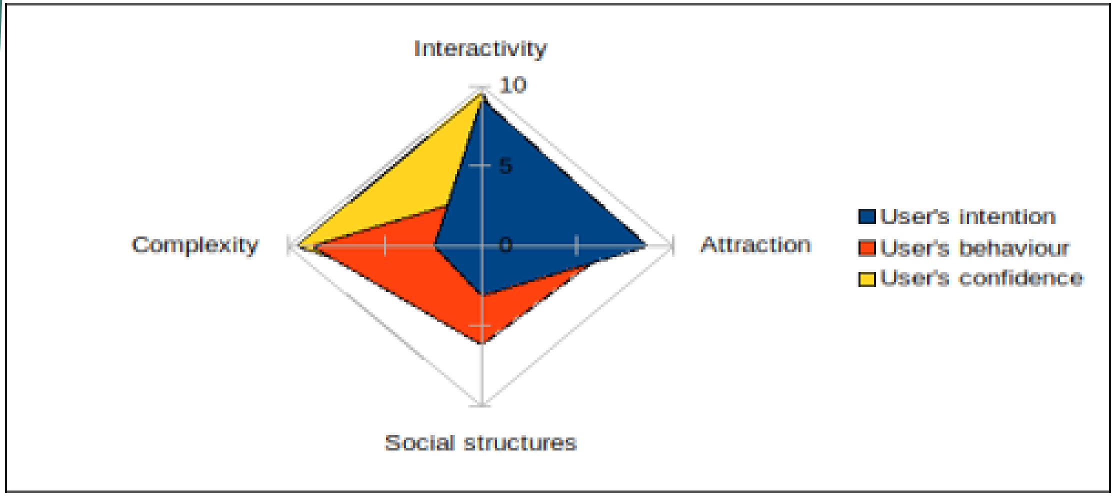


Figure 4. Interpretable trust dimensions and their impact on users' reaction.

2. Background

2.2. Explainable ML models

LIME / SP-LIME [3]

- Goes beyond a single trust of a prediction (trade-off approximation/complexi_ty)
- Sub-modular Pick LIME to study features impact on explanations

IBM 360° [4]

- More flexible: separating obvious explanation from black-boxes (local/global variables)

DARPA [5]

- Highest accuracy and lowest complexity by mapping from high-level to low-level features which is part of learning process (backpropagation)

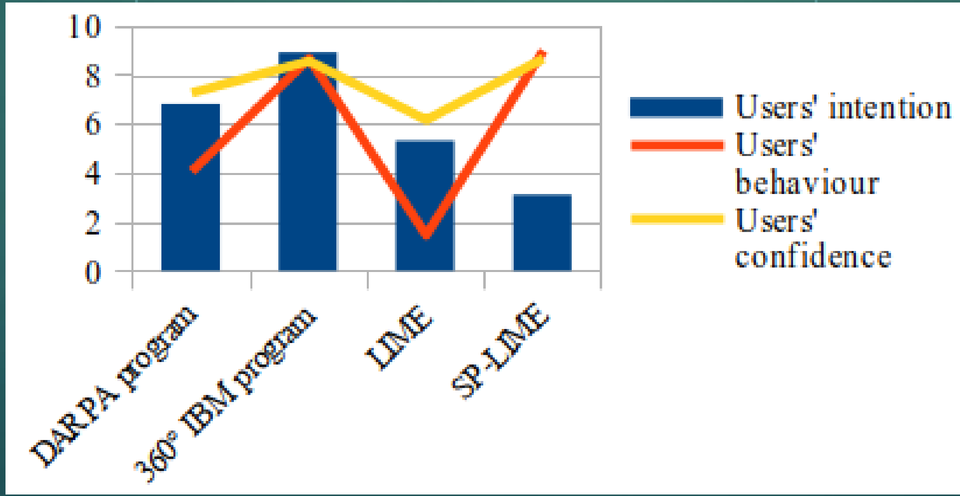


Figure 5. Users' reactions on explainable models.

2. Background
2.2. Case of DL

- (1) Generative modeling
- (2) Post-hoc techniques

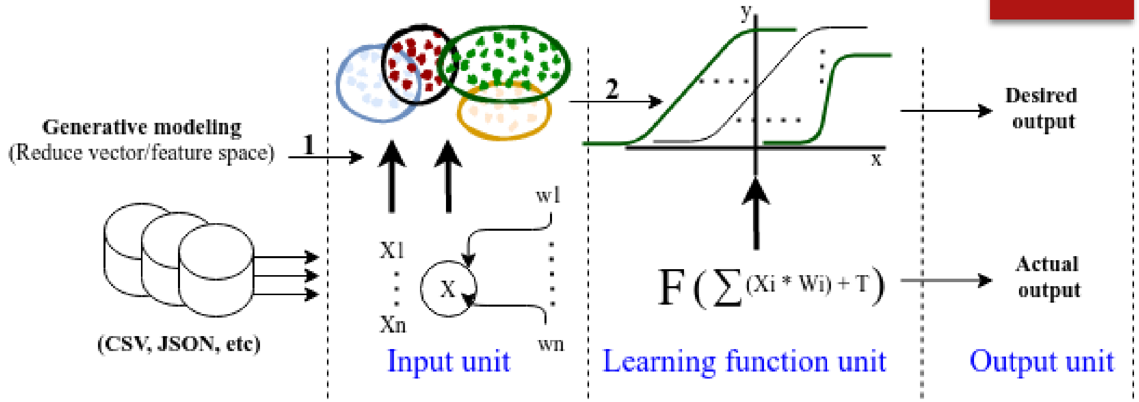


Figure 6. Explainable DL "Computational units".

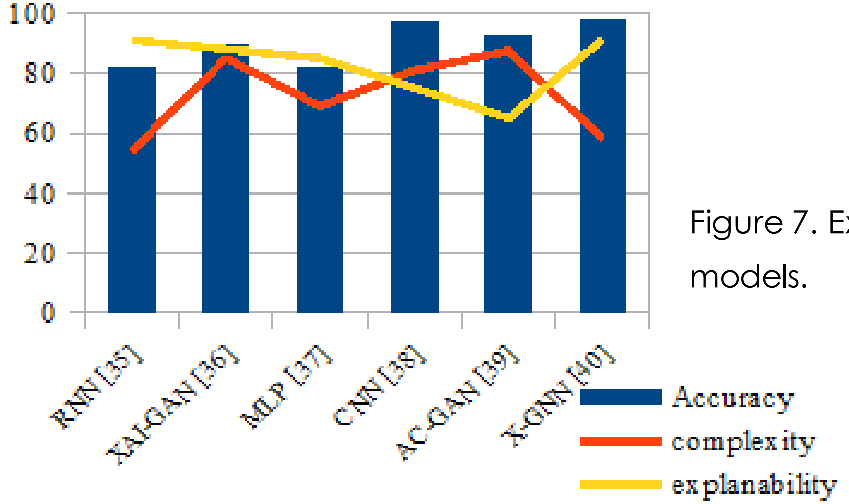


Figure 7. Explainable DL models.

2. Background

2.3. Limits

- Although explainable models show high performance, they fail to infer missing concepts.
- Data sparsity within perceptual metrics remains an issue.
- Users may express a changing behavior regarding any explainable model.

3. Insight

- Model decomposition (Figure 6) allows a link between explanability and a learning theory.
  - Credit assignment path.
  - Abductive learning, etc.
- To justify new features (trustworthy metrics) based on their impact on the whole performance.
  - I.e., active neurons in neural networks; feature selection, etc.

4. Demonstration

Importance of handling users' changing behavior in a recommender system.

-> Does the recommender explain or support a user changing behavior ?

-> Solution:

CHR (Constraint handling rules)

:- chr_constraint actor/1, actress/1.

:- chr_constraint movie/10, recommendation/3.

- movie(_,_,_,_,_,_,_,X,_,_) ==> actor(X); actress(X).
- recommendation(_, _, A) ==> movie(_,_,_,A,_,_,_,_,_,_)

Model's resilience against undeclared instances

```
?- movie(_,_,_,'drama, comedy',_,_,_,'Mr Bean',_,_).
false.
```

```
?- movie(_,_,_,'drama, comedy',_,_,_,'Mr Bean',_,_).
actor('Mr Bean'),
```

```
?- recommendation(male, sad, A), movie(_,_,_,A,_,_,_,'Mr Bean',_,_).
false.

?-
```

```
?- recommendation(male, sad, A), movie(_,_,_,A,_,_,_,'Mr Bean',_,_).
A = 'drama, comedy',
actor('Mr Bean'),
```

5. Legal concerns

- Exposure of explainable models and data privacy.
  - How far shall we explain?
  - How far shall we contextually adapt data?
- Explanability vs adversarial attacks.
  - e.g., IBM (predicted behabvior of "WayBlazer app") [6]
- Fairness and the need to introduce new regulatory metrics [7].

6. Conclusion and future research

- User-centered analysis.
- Gap: explanability/interpretability.

Reasearch perspective

- Logical reasoning for model certainty.
  - Perceptual metrics could be formalized before being trained.
  - Perceptual metrics could be typed and attributed for model exceptions.
- AI policy [8] for an easy disparate behavior deletion.

References

[1] D. F. Reding and J. Eaton, Science and Technology Trends, Exploring the S&T Edge NATO Science & Technology Organization. 2020. Retrieved from http://www.sto.nato.int/. Accessed on 21/02/2021 22:10.

[2] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asi, and B. Yu,"Definitions, methods, and applications in interpretable machine learning". PNAS, 116 (44) 22071-22080, 2019.

[3] M. T. Ribeiro, S. Singh, and C. Guestren, "Why Should I Trust You? Explaining the Predictions of Any Classifier". KDD 2016 San Francisco, CA, USA. doi: http://dx.doi.org/10.1145/2939672.2939778, 2016.

[4] A. Mojsilovic, "Introducing AI Explainability 360. Retrieved from https://www.ibm.com/blogs/research/2019/08/ai-explainablility-360/, 2019. Visited on 03/02/2019 01:12.

[5] M. Turek, "Explainable Artificial Intelligence (XAI). DEFENCE ADVANCED RESEARCH PROJECT AGENCY. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence, 2019.Visited on 06/03/2020 20:11.

[6] I. Portilla, "WayBlazer Cognitive Computing Application Powered by IBM Watson & Neo4j". GraphConnect Europe, 2016. Accessed from WayBlazer Cognitive Computing Application Powered by IBM Watson & Neo4j | LaptrinhX. Visited on 18/01/2019.

[7] UK Gov, "General Data Protection Regulation (GDPR)", 2020. Data protection - GOV.UK (www.gov.uk). Accessed on 02/05/2020 14:25.

[8] O. Dowden, "New strategy to unleash the transformational power of Artificial Intelligence". Retrieved from https://www.gov.uk/government/organisations/office-for-artificial- intelligence, 12 March 2021. Visited on 25/04/2021 22:45

Thanks for listening