Outline

Nauta et al. 2023

Paper Selection

- Literature from 2014-2020
- 12 conferences
- Query: explain*|explanat*|interpret*
- Search on 04.05.2021: 606 Results
- Without workshop papers and tutorials: 494
- After inclusion criterion 361:

  Original work introducing, applying, and/or evaluating one or more methods for explaining a machine learning model.

- only papers that introduce a new xai technique: 312
  - the reduced 49 papers were still concidered for evaluation metrics

Categorization of the papers

There were 6 dimensions for paper categorization:

- Type of data (time series, graph, image...)
- Type of predictive model (NN, SVM, Tree Ensemble...)
- Type of method used for to explain (built-in, post-hoc...)
- Type of explanation (Heatmap, Feature Plot...)
- Type of problem (Model explanation, outcome explanation...)
- Type of task (classification, regression...)

XAI Explanation Quality Properties

The authors defined 12 quality properties to be examined:

- Correcteness
- Completeness
- Consistency
- Continuity
- Contrastivity
- Covariate complexity
- Compacteness
- Composition
- Confidence
- Context
- Coherence
- Controllability

General

- Extensive approach focusing on finding trends and maybe blindspots in research
- Due to the high volume, statistic evaluation is possible
- Points to automated, quantitative evaluation methods (could be interesting for us -> TimeXAI)
- Maybe cherry-pick from their quality properties?

Mohseni et al 2021

Paper Selection

- choose from multiple disciplines: ML, HCI, Visualization, Psychology
- iterative approach chosing 40 papers as a start and then doing upwards/downwards literature research with some refinement, resulting in 226 papers
- keywords: interpretability, explainability, intelligibility, transparency, algorithmic decision-making, fairness, trust, mental model, and debugging in machine learning and intelligent systems

Summary

- derived a general framework from a more "distanced" view for a whole design process of an XAI system used by novices and experts alike
- split design goals between novice users, data experts and AI experts
- Introduce 5 evaluation measures for XAI systems
- In general more HCI view

General

- HCI view might be interesting for TimeXAI, do we want to incorporate this?
- iterative approach maybe interesting for us?
- cherry-pick goals and evaluation measures?

Das & Rad 2020

Paper selection:

- focuses on milestone papers from the last 15 years
- good overview over most interesting papers

Main categorization:

- Scope: Local/global explanations
- Methodology: Perturbation/Backpropagation
- Usage: Model-intrinsic/post-hoc

Summary:

- compares XAI methods directly based on methodology
- most research focuses on model-agnostic post-hoc explainability due to easy integration and wide reach

Schwalbe & Finzel 2023

- Reviewed 50 surveys on XAI in meta survey
- there is no definite taxonomy for XAI
- they tried to introduce one (pretty recent 01/2023), maybe we can adapt to this?

Paper collection:

Large scale approach with database

Pros	Cons
* captures most papers and will result in an extensive amount of data	* filtering and reading will be a lot of work
* if done with a database approach might be a start for a knowledge database for the future	* worst case: many papers are not insignificant (=can't be discarded) but also not helpful
* enables statistic analysis	
* best chances to get impulses for TimeXAI	

Iterative approach

+ hopefully less "bad" papers to read and include