Machine Learning (ML) for Data-Driven Decision-Making (DDDM)

Module 5: Ethical and Responsible AI

Tobias Rebholz

University of Tübingen

Summer Term 2024

Explainable AI

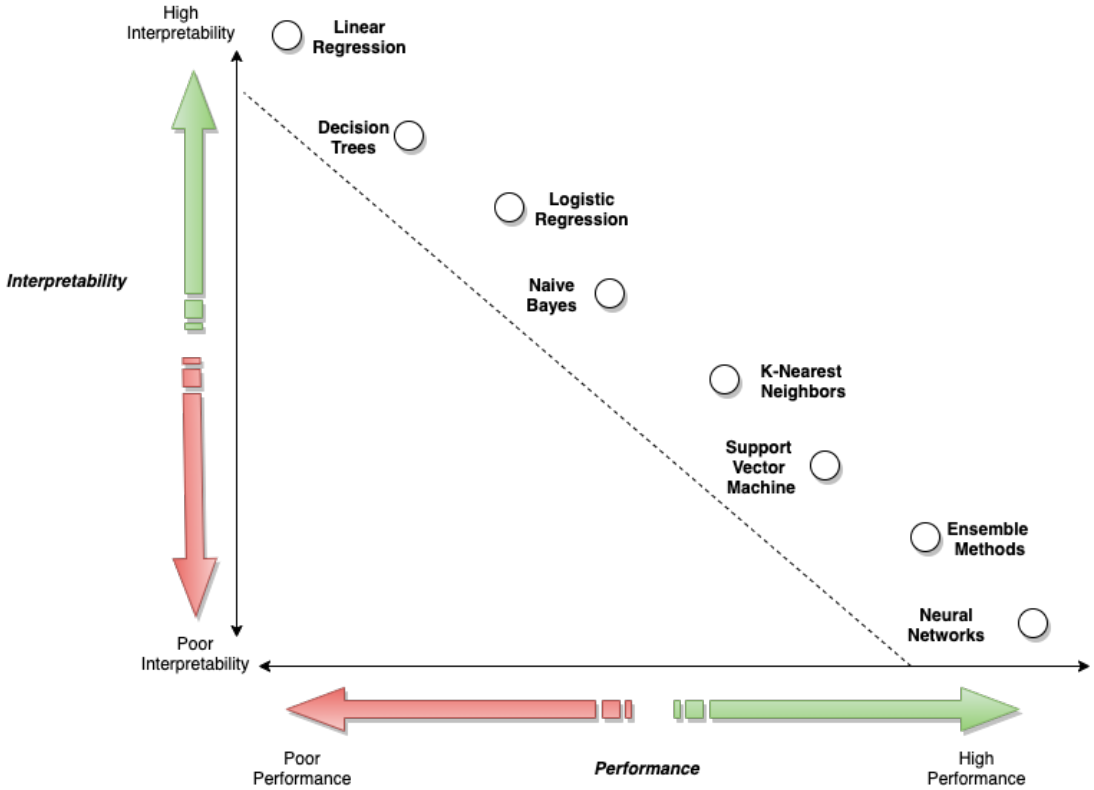
Explainable AI

In the history of science and technology, the engineering artifacts have almost always preceded the theoretical understanding

(Yann LeCun, Turing Award Winner)

Model Interpretability

- Problem: More powerful models are less understandable



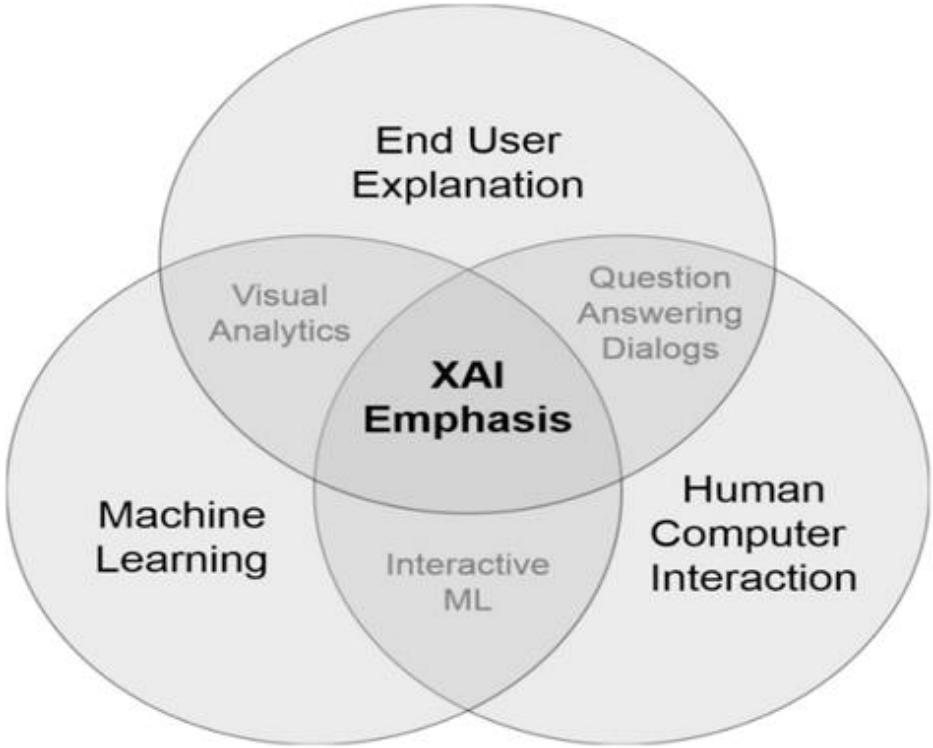
(https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html)

Model Interpretability

- Black box problem: ML algorithms, and particularly Deep Learning methods, solve problems in inscrutable ways because they can refine themselves autonomously and with an idiosyncrasy beyond the scope of human comprehension
- This is especially worrisome when the algorithm is making decisions with real-world consequences for human well-being, such as:
  - Determining whether a blip on a scan has the potential to be cancerous
  - Applying the brakes in an autonomous vehicle
  - Granting a loan to buy a house
  - ...

Explainable AI

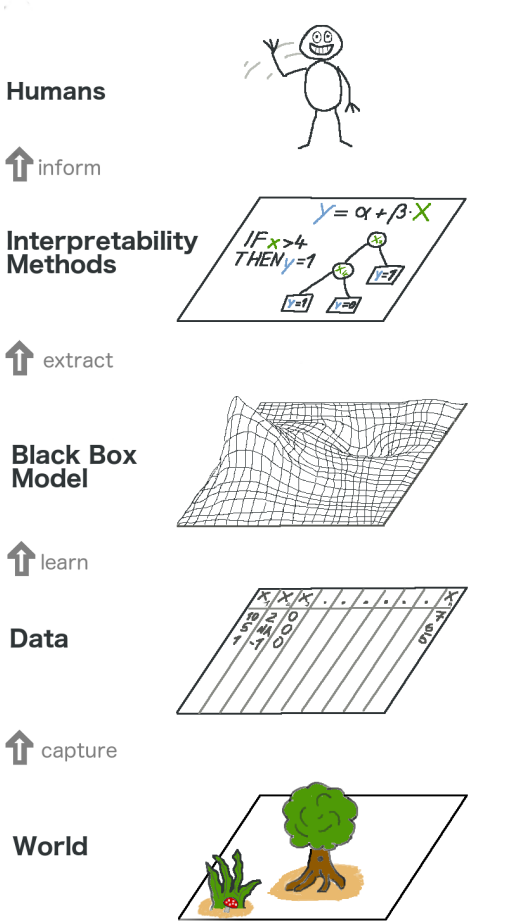
- Explainable AI (XAI): Subfield of ML that aims to address the black box problem by attempting to increase the interpretability, transparency, and ultimately also fairness of such methods



(Dağlarli, 2019, Figure 1)

Explainable AI

- The XAI process:



(Molnar, 2022, Figure 6.1)

Explainable AI vs. Interpretable ML

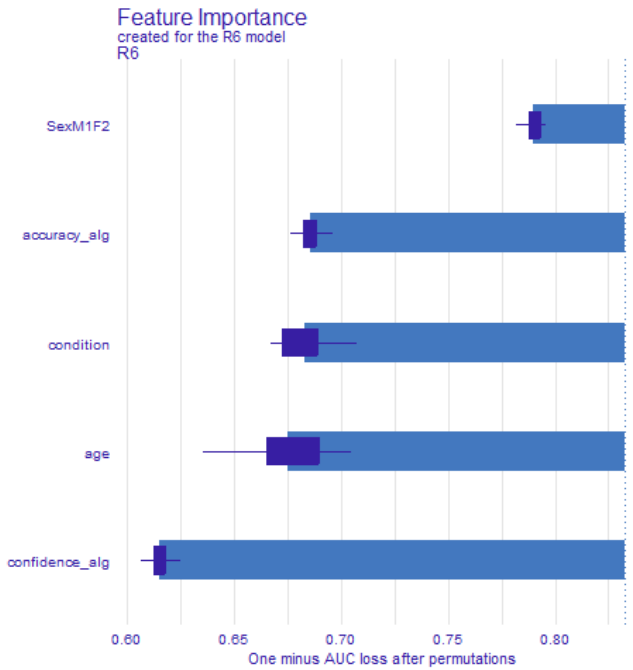
- Explainable AI (XAI): Methods and techniques used to make complex, opaque ML models more understandable to humans
  - This includes post-hoc explanations of models that are not inherently interpretable (e.g., a random forest with many decision trees).
- Interpretable ML (IML): Developing models that are inherently human understandable
  - This means that the inner workings of the model and how it makes decisions can be directly understood without additional tools or techniques (e.g., a single decision tree)
- Since we covered interpretable models at the beginning of the FS, we will focus primarily on XAI techniques in this final module
  - However, we will use both terms (XAI/IML) interchangeably, referring to their common goal of making the predictions of ML methods more transparent and understandable

Model Agnosticity

- Some interpretation methods are model-specific
  - I.e., they can only be applied to a certain model (family)
  - Vs. model agnosticity: Separating the explanations from the ML model
- Typically, not just one, but many types of ML models are compared against each other in their performance to solve a specific DDDM task (cf. PA-Projects)
  - Biggest advantage model-agnostic interpretation methods: Can be applied to any ML model $\Rightarrow$ Freedom to use any model, essentially including the most powerful ones
- For this reason, and for the sake of brevity, we will only discuss model-agnostic XAI/IML techniques in this module

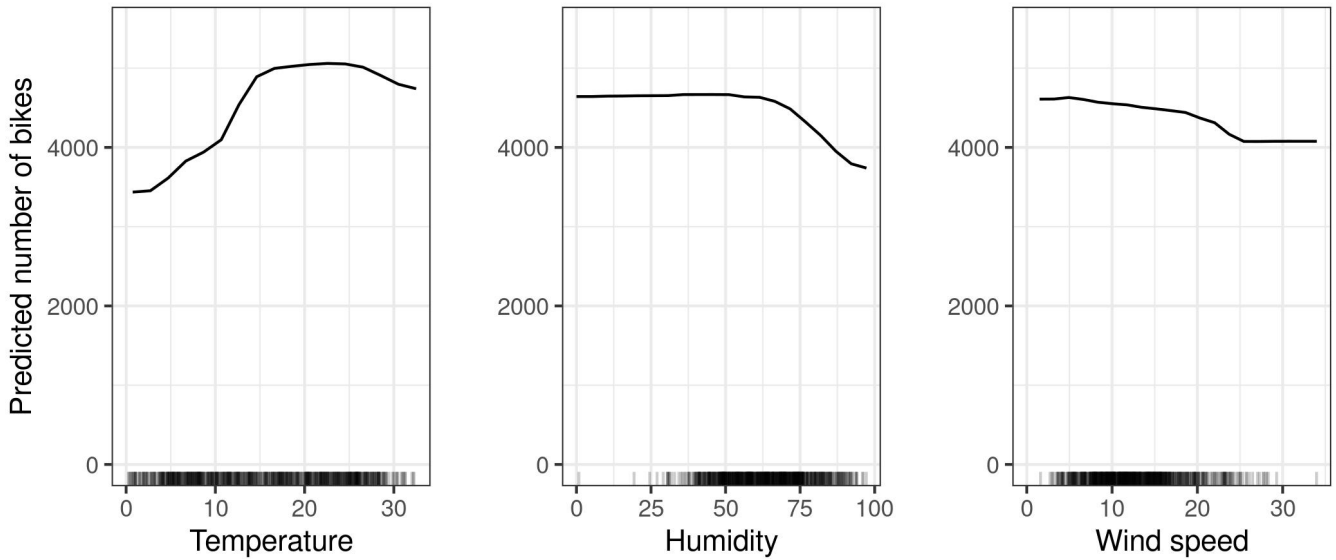
Important XAI/IML Techniques

- We already know an important model-agnostic XAI technique that was introduced for random forests in module 2
- Permutation Feature Importance: Measuring the increase in the prediction error of the model after randomly permuting the feature's values
  - Idea: Breaking the original relationship between this feature and the target



Important XAI/IML Techniques

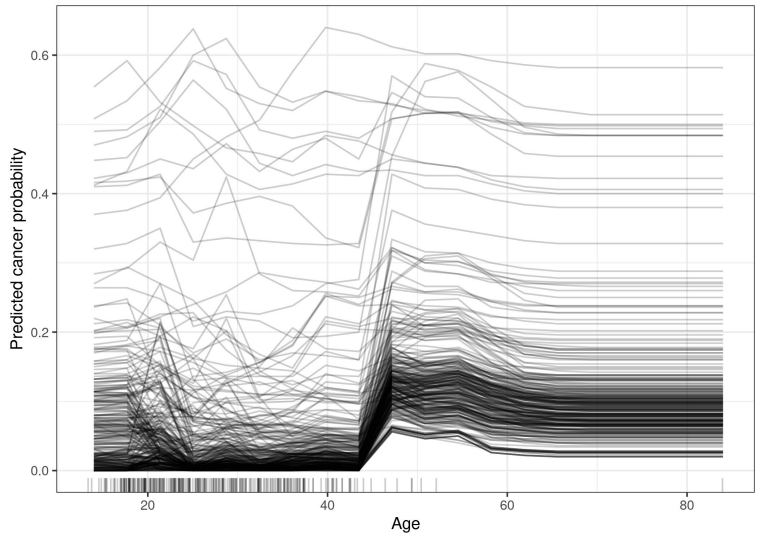
- Partial Dependence Plots (PDPs): Visualizing the relationship between a feature and the predicted outcome, averaging out the effects of all other features
  - I.e., showing the so-called "marginal effect" of a feature on a target
  - The relationship between a feature and the target can be linear, monotonic, . . . , or much more complex



(Molnar, 2022, Figure 8.1)

Important XAI/IML Techniques

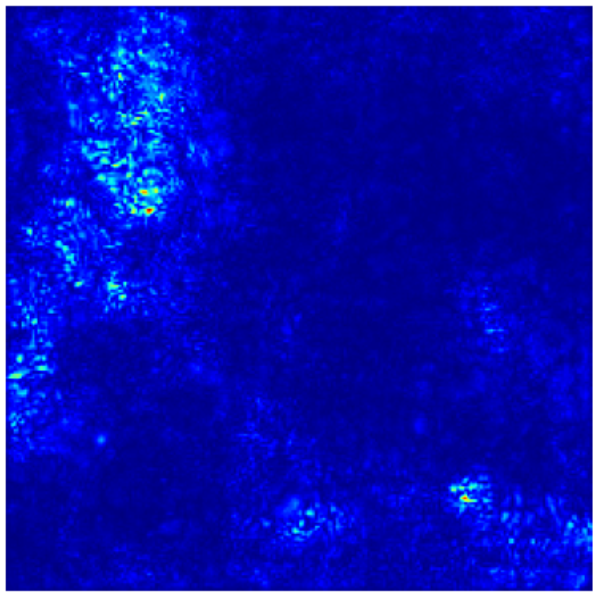
- Individual Conditional Expectation (ICE) plots: Visualizing how changes in a single feature affect the predictions of a ML model for (all) individual instances
  - Allows for an assessment of the heterogeneity of marginal effects across instances in a data set
  - In this sense, ICE plots are an extension of PDPs



(Molnar, 2022, Figure 9.1)

Important XAI/IML Techniques

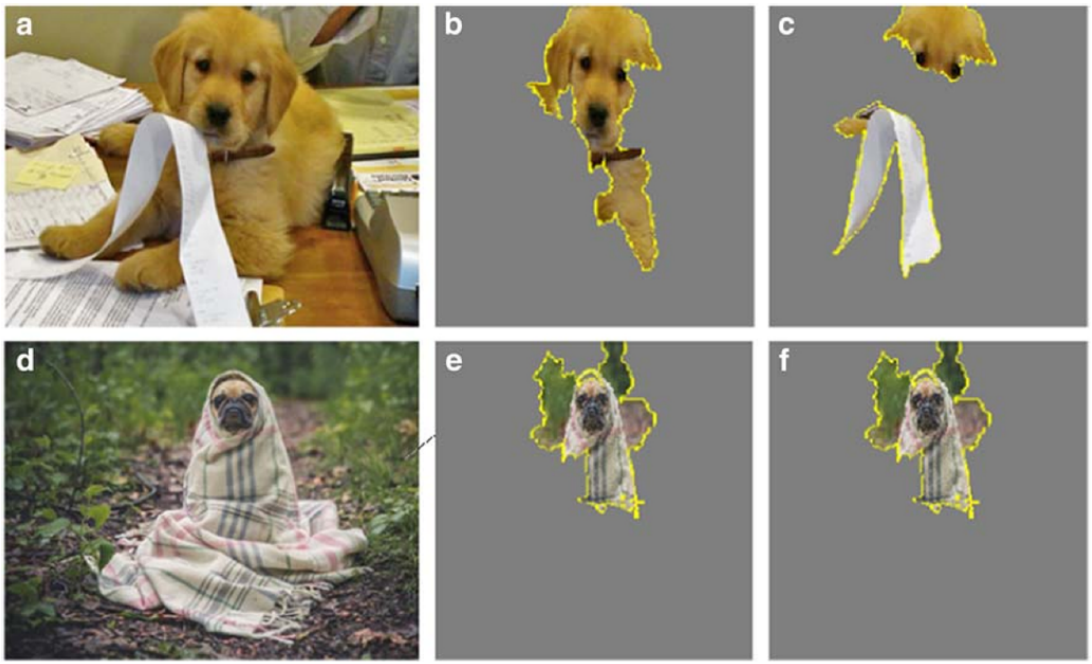
- Saliency Maps: Highlighting regions in input data (e.g., image pixels) that are most influential in the model's prediction
  - E.g., pixels are colored according to their contribution to the classification:



(Molnar, 2022, Figure 10.8)

Important XAI/IML Techniques

- Local Interpretable Model-Agnostic Explanations (LIME): Highlighting the superpixel areas that are most important for an image classification
  - LIME-explanations for Inceptionv3's classifications:
    - b: golden retriever; c: toilet tissue
    - e: bath towel; f: three-toed sloth



(Taylor & Taylor, 2021, Figure 2)

Excurse: Further Important XAI/IML Techniques

- General idea behind LIME (Ribeiro et al., 2016): Explaining individual predictions by approximating the complex model (e.g., neural network) locally with a simpler, more interpretable so-called "surrogate model" (e.g., linear regression)
- SHapley Additive exPlanations (SHAP; Lundberg and Lee, 2017): Providing a measure for the contribution of each feature to the prediction
  - Cf. permutation feature importance
  - But: Based on the game theoretically optimal "Shapley values": How to fairly distribute the "payout" (= prediction) among the "players" (= features)?
- Counterfactual explanations: Describe what minimal changes to the input features would have changed the model's output
  - E.g., achieving a desired prediction
  - Goal: Providing actionable insights about model sensitivity
- ...

XAI/IML in R

```
"Not discussed!"
```

- If you want to use any of these methods in your PA-projects, please refer to the references on the last slide!

Artificial Cognition

Artificial Cognition

- Computer scientists explain artificial intelligence primarily by tinkering under the hood of the black box, including attempts such as:
  - Generating explanations through more interpretable architectures
  - Introducing a second AI that examines another AI's decision and learns to generate explanations for it
- Problems:
  - Most XAI techniques are merely correlational in nature: Estimating which features the model cares about by displaying hypothetical predictions based on previously learned correlations
    - This is not bad per se: E.g., we can see when a model may be unfairly over-indexing on protected attributes, like gender or race (discussed in more detail later!)
  - Malleable introspection: Similar to asking a human how they made a decision, AI systems generate explanations that may not be the true explanation for their behavior

Artificial Cognition

- The human mind is also a black box!
  - Cognitive psychology: A science of behavior that works without opening its black box
- Artificial Cognition (Ritter et al., 2017; Taylor and Taylor, 2021): Translating the methods and rigor of cognitive psychology to the study of artificial black boxes to enable explainability
  - Satisfactory explanations: Causal, rather than correlative $\Rightarrow$ Require experimental attempts of falsification
  - Subfield of the so-called "Machine Behavior" movement toward XAI: All interpretation methods that do not rely on AIs trying to explain themselves

Artificial Cognition

- Procedure: Using behavioral experiments to infer the properties of invisible artificial mental processes
  - E.g., by asking questions like: What if the input was a little bit different, would the output of the model be different as well?
  - Cf. approaching the black box problem for the human mind
- Incl. identifying boundary conditions: Being able to explain when a behavior occurs implies that we should also be able to account for when the behavior stops occurring
  - In psychology, the term "boundary conditions" refers to the limits or constraints that affect the validity of a theory or model
- Due to huge differences between ML model architectures, in contrast to human brain architectures, more akin to the psychology of individual differences
  - I.e., instead of central tendencies, deviations from the mean represent true data

Ethical Considerations

Model Bias

- The parameters of a fitted ML model reflect the truth in the training data, not necessarily the real world!
- Problems:
  - Discrepancies between the training data and the real world are inevitable
    - This can create several vulnerabilities for stakeholders (i.e., users and other parties affected by a model's outcomes)
    - E.g., training datasets that are biased against a particular group (e.g., gender, race) will yield model predictions that are biased against that particular group
  - Without a causal model, it is challenging to distinguish between direct discrimination (e.g., women earning less because they are women) and indirect discrimination (e.g., women earning less because they choose lower-income professions)

Model Fairness

- Closely related to IML is the concept of model fairness: A trained ML model is deemed fair if it does not discriminate against protected subgroups
  - Fairness is context-dependent and involves normative consensus
  - Transparent documentation on model performance and intended use cases is crucial for assessing model fairness
- Fairness is particularly relevant for psychological assessments using ML, but potential remedies exist, such as:
  - Avoiding to directly includ protected attributes (e.g., gender) in the model to prevent it from using these information
    - However, flexible ML models can infer group membership on the basis of available features that are related to the protected attribute (e.g., socio-economic status from ZIP code)
  - Explicitly evaluating fairness by comparing predictions for different values of the protected attribute using XAI teachniques (e.g., feature importance or PDPs/ICE)

Fairness Challenges

- Fairness and ethics are flexibly interpreted, and XAI-based explanations will only be satisfactory if they enable stakeholders to judge whether the decision was appropriate in that particular situation
  - Cultural differences: Influence perceptions of AI ethics, especially in high-stakes scenarios (e.g., car accidents of autonomous vehicles)
  - Perspectives: Can shift based on stakeholders' roles (e.g., passenger vs. pedestrian in the context of autonomous driving)

Model Accountability

- Model Accountability: ML models should be accompanied by clear documentations, detailing their performance characteristics, limitations, and intended use cases
  - Transparency helps stakeholders understand the decision-making process and trust the model's outputs

Selected Accountability Mechanisms I

- Establishing frameworks that can be used to assess whether ML systems operate ethically
- Ensuring that shareholders (i.e., organizations that deploy ML models) can be held responsible for the outcomes
- Continuous monitoring and evaluation of deployed models to detect and mitigate biases over time
- Ensuring that data used for training ML models respects users' privacy and complies with relevant regulations (e.g., GDPR)
  - Incl. implementing data anonymization and secure data handling practices to protect sensitive information
- Addressing vulnerabilities in ML models that could be exploited maliciously
  - Developing robust defenses against adversarial attacks and ensuring the integrity and security of ML systems

Selected Accountability Mechanisms II

- Assessing the broader societal implications of deploying ML models
  - Incl. potential impacts on employment, social equity, and human rights
- Promoting ethical AI use by considering long-term consequences and aiming for beneficial outcomes for all stakeholders
- Involving diverse groups in the development and deployment process to ensure that the perspectives and needs of different communities are considered
  - Incl. encouraging open dialogue and collaboration

Trust in Technology

- Psychological factors that were found to influence people's trust in technology:
  - Providing clear, understandable information about—particularly black box—models' judgment and decision-making processes
  - Consistency and predictability of a model's behavior (e.g., in terms of performance)
  - User-centered design: E.g., (conversational) user interfaces that facilitate access and understanding
  - Users' prior experiences with similar technologies and their individual differences (e.g., tech-savviness, cognitive style)
  - ...

Trust Measurement Issues

- In general, measuring trust (e.g., using psychometric tools and surveys) poses considerable challenges, such as:
  - Trust is not static; it evolves over time based on user interactions with specific technology
    - Longitudinal studies and repeated measures are necessary to understand how trust develops and changes
    - Continuous user feedback and iterative design processes can help build and maintain trust over time
  - Trust levels can vary greatly depending on the context of use and cultural background of the stakeholders
    - Cross-cultural studies are important to identify and account for these differences in trust
- Trust can also be inferred from user behavior, such as frequency of use or adherence to recommendations
  - Self-reported measures of trust in technology should be complemented by behavioral data

FP-Projects

Google Cheat Sheet

- How to find studies/datasets for your projects?

Example	Result: webpages with
psychology animals	the words **psychology** *and* **animals**
"Vaticano**"**	the *exact word* **Vaticano**
"Dutch language**"**	the *exact phrase* **Dutch language**
"global warming" **OR** "greenhouse effect"	the phrase **global warming** *and/or* the phrase **greenhouse effect**
post * syndrome	the words **post** and **syndrome** *separated by one word*
salsa -dancing	the word **salsa**, *not* the word **dancing**

(adapted from https://de.slideshare.net/slideshow/google-cheatsheetenglish/81625186)

Google Cheat Sheet

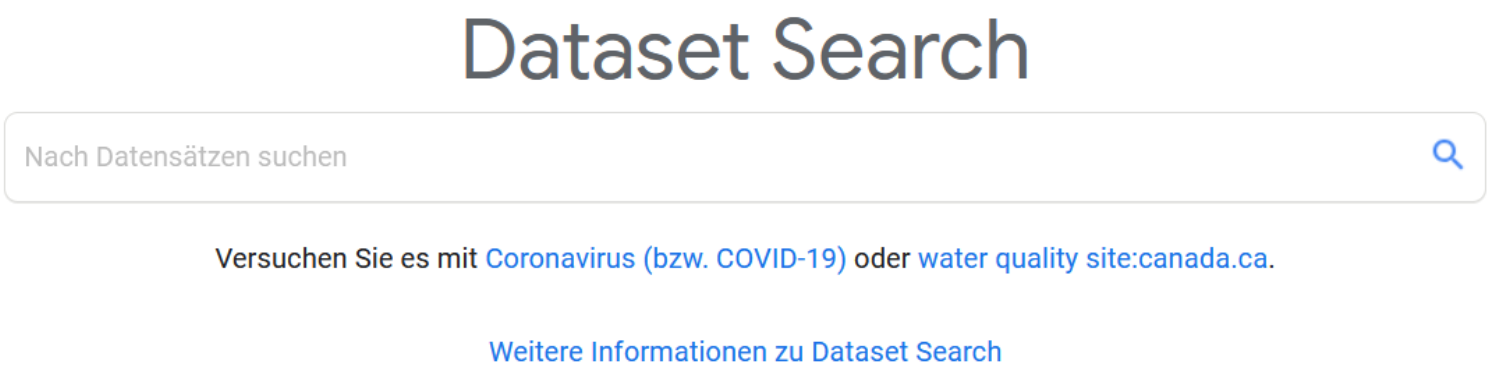
- How to find studies/datasets for your projects?

Operator	Meaning	Example
##..##	Search within number range	"Kyoto protocol" 1990..2000
site:	Search specific website or specific domain	"human rights" site:un.org
-site:	Exclude website or domain	"global warming" -site:wikipedia.org
filetype:	Search specific file format	"social media" filetype:pdf
-filetype	Exclude specific file format	psychology -filetype:xls
( )	Formulate complex queries	Siberia (site:gov OR site:edu)

(adapted from https://de.slideshare.net/slideshow/google-cheatsheetenglish/81625186)

Google Dataset Search

- https://datasetsearch.research.google.com/



Evaluation Criteria FS-Presentations

	Student:	**Name**
	Weight:	**75.00%**
Content	Meaningful and appropriate summary (incl. criticism) of the original study	
	Topic/core research question clearly named and well justified	
	Own considerations included (e.g., examples, critical comments, conclusions)	
	Appropriate focus and factual accuracy (esp. in describing the planned analytical approach)	
	Logical argumentation and comprehensibility (e.g., precise explanation and definition of technical terms)	
	Embedding in the seminar topic (e.g. references to other topics/methods/modules/applications in behavioral science)	
	Sources, citations, (own) figures/tables etc. formally correct (according to APA7)	
	Grade (rounded):	**#DIV/0!**
	Weight:	**25.00%**
Form / Style / Design	Clear structuring (i.e., using an outline, including transitions)	
	Presentation style (i.e., free, fluent, easy to understand, appropriate speed and choice of words)	
	Contact with the audience (e.g., involving the participants by actively asking questions)	
	Clarity, readability, and appropriate use of slides *(and other materials, if applicable)*	
	Good coordination between the speakers (if applicable); group effort recognizable	
	Grade (rounded):	**#DIV/0!**
	Total grade:	**#DIV/0!**
	Final grade (rounded):	**#DIV/0!**

Summary

Summary

- XAI/IML: Attempts to open the black box of ML/AI
  - Ethical considerations are important in light of the increasing implementations of black box systems in the digital ecosystem
    - E.g., fairness, transparency, accountability, ...
  - Vs. Artificial Cognition: Cognitive psychological approach (i.e., experimentation/falsification) to explain black box algorithmic behavior
- Trust in technology (incl. measurement issues):
  - Ensuring ethical use of ML/AI by addressing privacy, security, fairness etc. concerns is crucial for fostering trust
  - Transparent communication about the limitations and potential biases of ML/AI systems can help manage user expectations and trust

Homework

- FS:
  - Further readings:
    - XAI: Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000560
    - IML: Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar. https://christophm.github.io/interpretable-ml-book/
    - I.a., ethics: Van Dis, E. A. M., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7
- FP:
  - Finish the group project planning
  - Prepare a short (**20-25 minutes**) presentation of the group project proposal (see module 4 for suggested structure)