

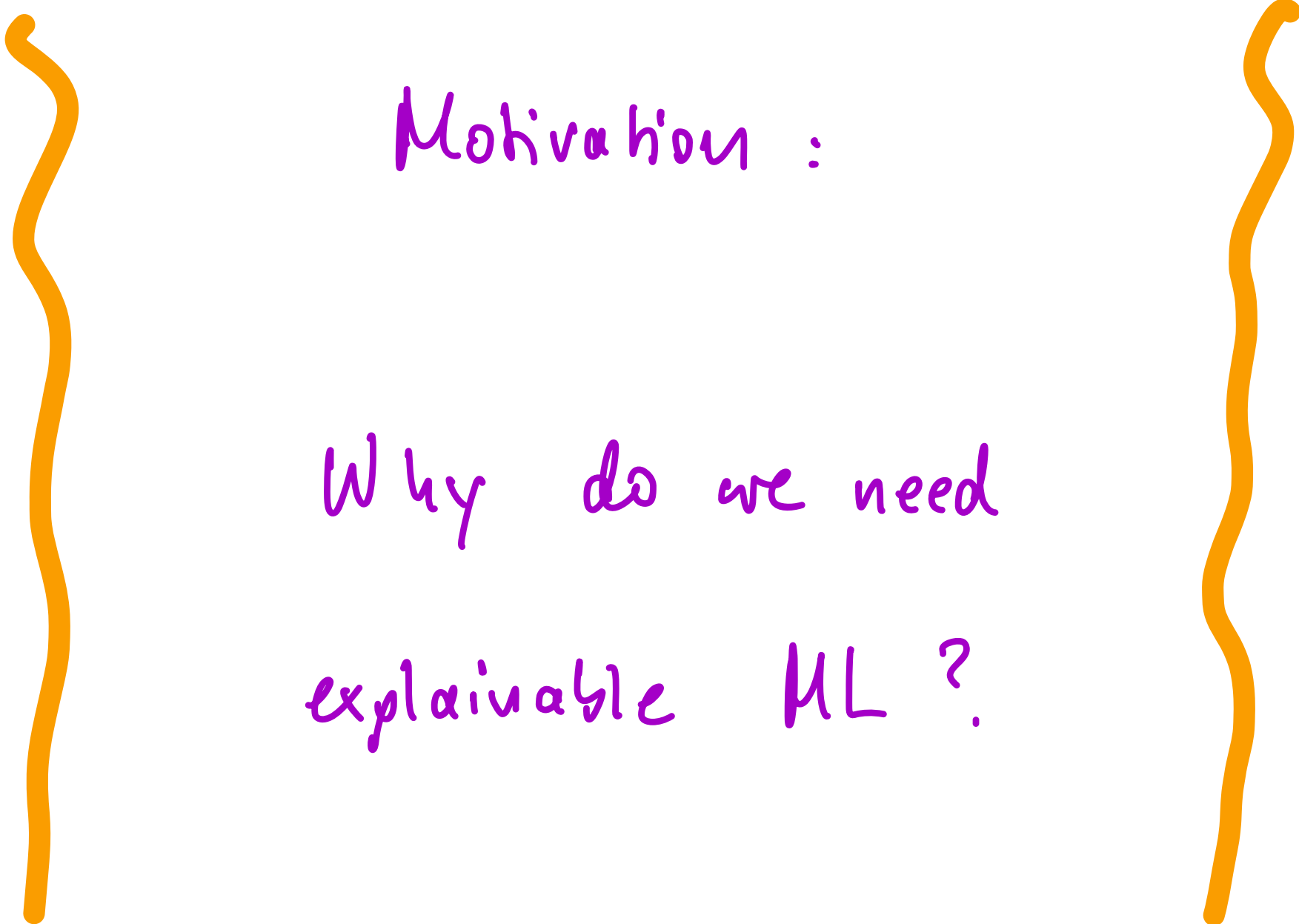
Seminar

Explainable machine learning

Max Thierse, Gunnar König,

Ulrike von Luxburg

Summer term 2026



Motivation :

Why do we need
explainable ML ?

Why do we need explainable machine learning?

Black box models with millions of parameters

Hard to debug, hard to trust

Might want to "understand" what the model does:

- in science
- for critical applications (medicine)
- in society (loan, university admission, ..)

AI regulation requires transparency!

Purposes of explainable AI

Would like to be able to query for "explanations" why the ML model acts / decides / predicts in a specific way.

Might serve to

- debug your code
- establish trust (society applications; medicine)
- balance power / control between model provider and "customer"
- give hints for concrete actions (eg recourse)
- reveal scientific insights (truth?)

AI Regulation

EU AI Act:

- AI applications are assigned to a risk category
- High-risk applications need to satisfy specific transparency criteria.
- Explainability is out of them!



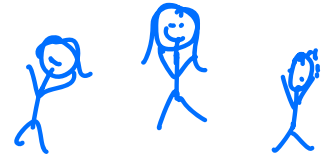
Very brief introduction
to explainable AI



Setup for now

Machine learning on feature-based data (not text, not images)

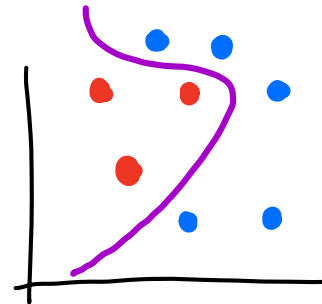
- Patients described by numeric values
- Loan applicants "



Data is high-dimensional (many features)

$$\begin{pmatrix} 1 \\ -2.3 \\ 1.5 \\ 17 \\ 32 \end{pmatrix} \text{ features}$$

Task: classification or regression

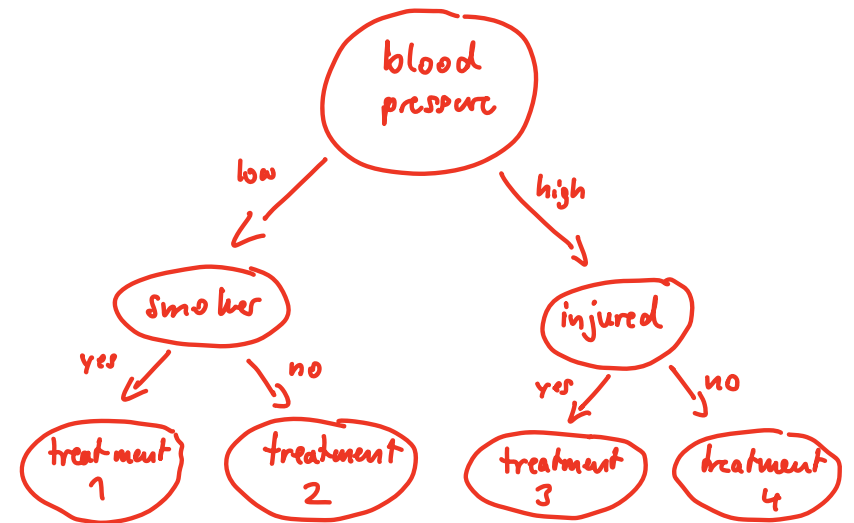


Interpretable vs. explainable ML

The gold standard: interpretable ML

Use models that are so simple that we can understand/judge their underlying decision mechanism:

- decision trees:



- linear models:

$$f(x) = \sum a_i x_i$$

- generalized additive models (GAMs)

$$f(x) = \sum f_i(x_i)$$

But ...

Interpretable ML might not always be powerful enough
to solve complex prediction problems.

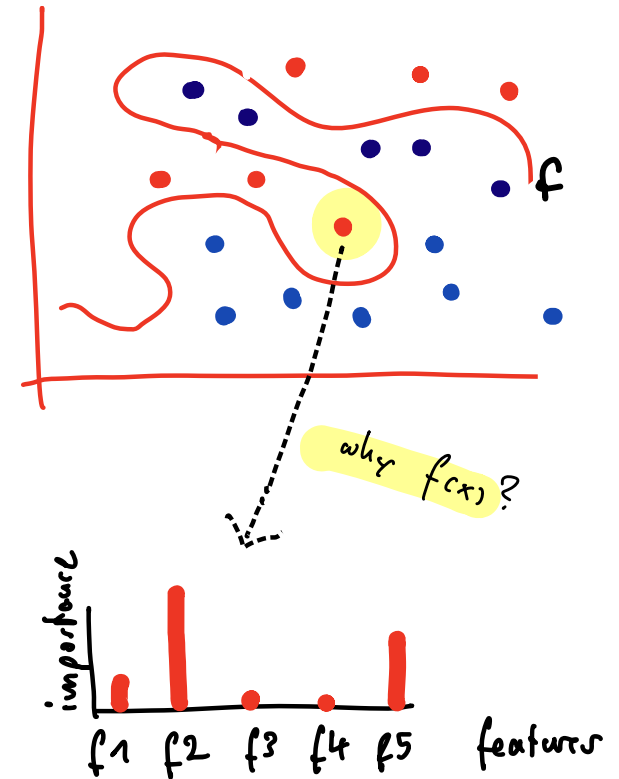
Local post-hoc explanations

- Consider a complex decision model f (cannot explain it as a whole)
- Request local post-hoc explanation for individual decisions:

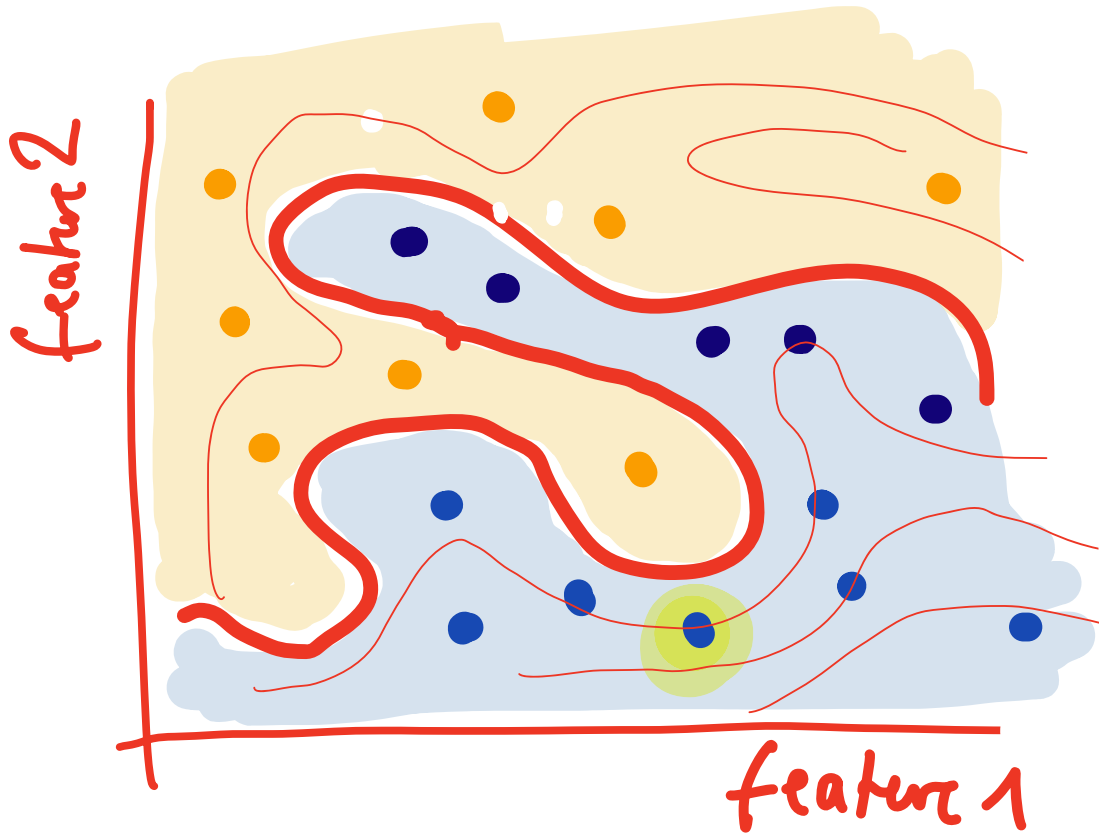
why $f(x)$?

- "Explanation" often comes in form of a feature attribution:

which of the input features was important for the particular decision?



Rationale for creating such explanations

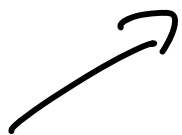


If small changes in feature i of x change the decision about x , then feature i is considered important.

Often boils down to approximating f locally at x by a simple model, for example a linear function.

Explanations for images

original
image



creating
"features"



(b) explaining 'terrapiin'



(c) explaining 'strawberry'

Data attribution vs feature attribution



Which of the other points in the training set were particularly important for the prediction $f(x)$?

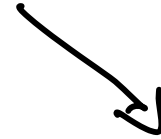


Which of the features of x were particularly important for the prediction $f(x)$?

Global vs local explanations



Want to understand/explain
the decision mechanism/rationale
globally



Want to explain only
decisions for individual
local data points

Model-specific vs model-agnostic explanations



Explanations depend on the model architecture, say the specific architecture of a neural network, eg attention scores



Shap, LIME, Anchors, ...

Explanation does not make (strong) assumptions on the form of f or how it has been generated

Cooperative

vs

adversarial scenario



explanation provider and
explanation recipient share the
same goal

Examples:

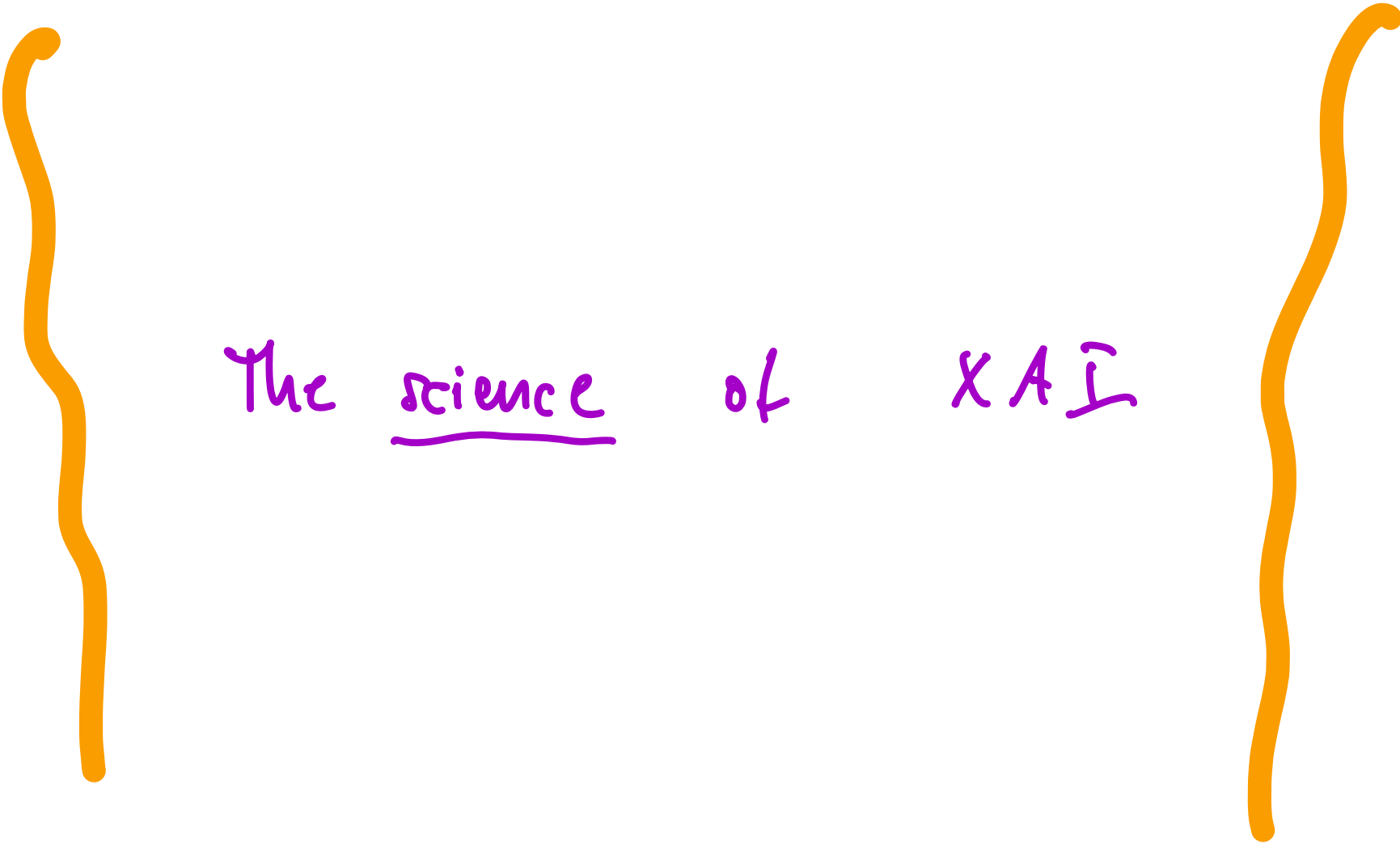
- Model / code debugging
- Scientific discovery



... have different goals

Example: loan application

No incentive to provide
"true explanation"



The science of XAI

Engineering vs. Science of XAI

- One can **engineer** many different heuristics, criteria, algorithms that spit out explanations.
- Many of them often produce explanations that "make sense"

But what kind of **scientific guarantee** can we give?

- When are explanations "true"?
- By which standards should we evaluate explanations?
- What kind of guarantee can we give, under which assumptions?

When is the scientific approach important?

AI act: (eg, credit scoring)

Medicine

Critical applications

...

When is an engineering approach justified?

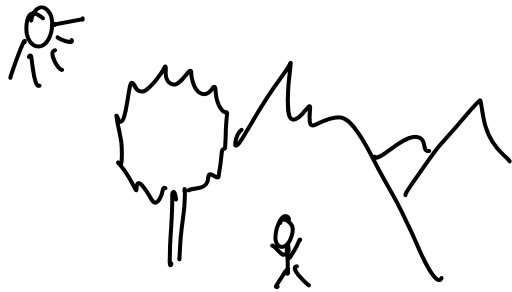
Exploratory data analysis, say in science

Code debugging

Whenever a useful explanation can help, but even many wrong explanations do not really hurt

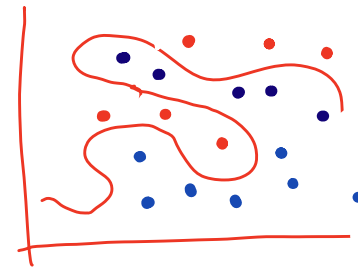
Can we assert that an explanation is true?

True world:



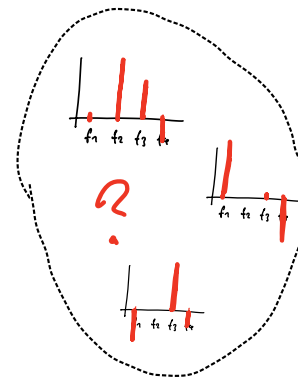
? ! ?
incomplete
knowledge

The algorithm's world



Geometry:
distances, features, labels

? ! ?
ambiguity



... rather not ...

Then how should we evaluate explanations?

- Benchmarks?
- Statistical tests?
- Ad hoc properties?

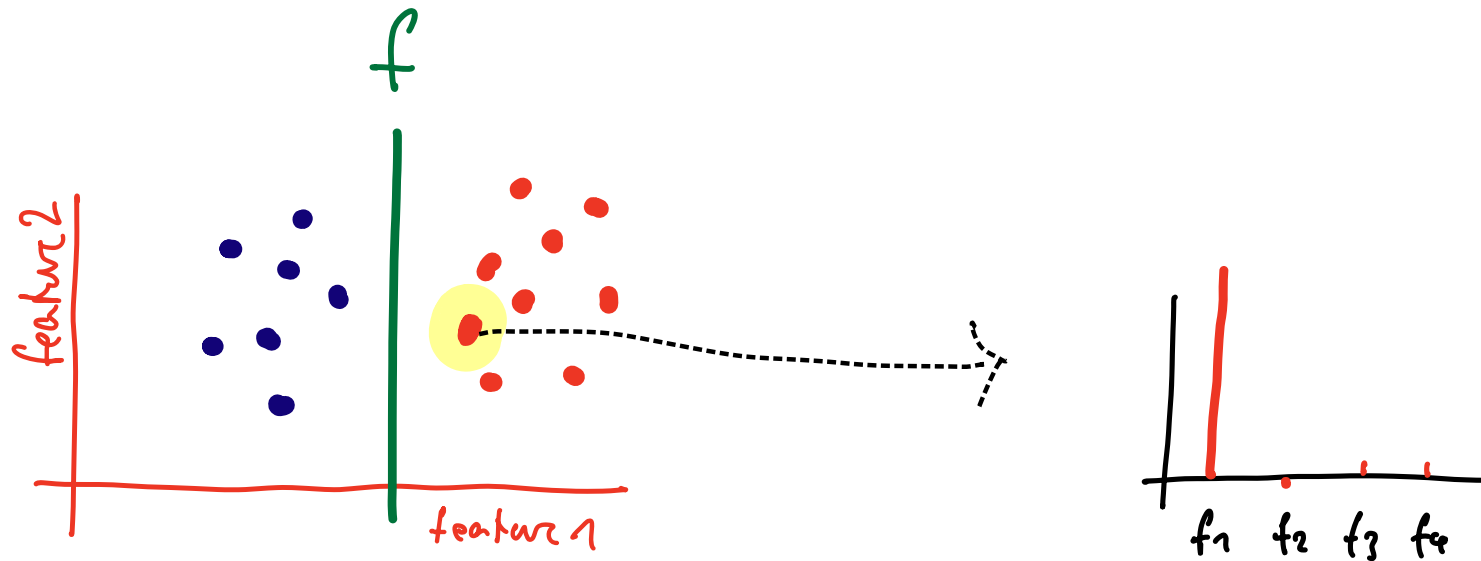
Explanations can be manipulated / cheated

If you control both f and the explanation algorithm, you can change f in a subtle way to produce a desired explanation.

... see later ...

Theoretical guarantees ?

So far, guarantees only exist if the underlying function f is simple, say for a linear function :



Theoretical guarantees ?

If the underlying function is complex,

then theoretical guarantees cannot exist

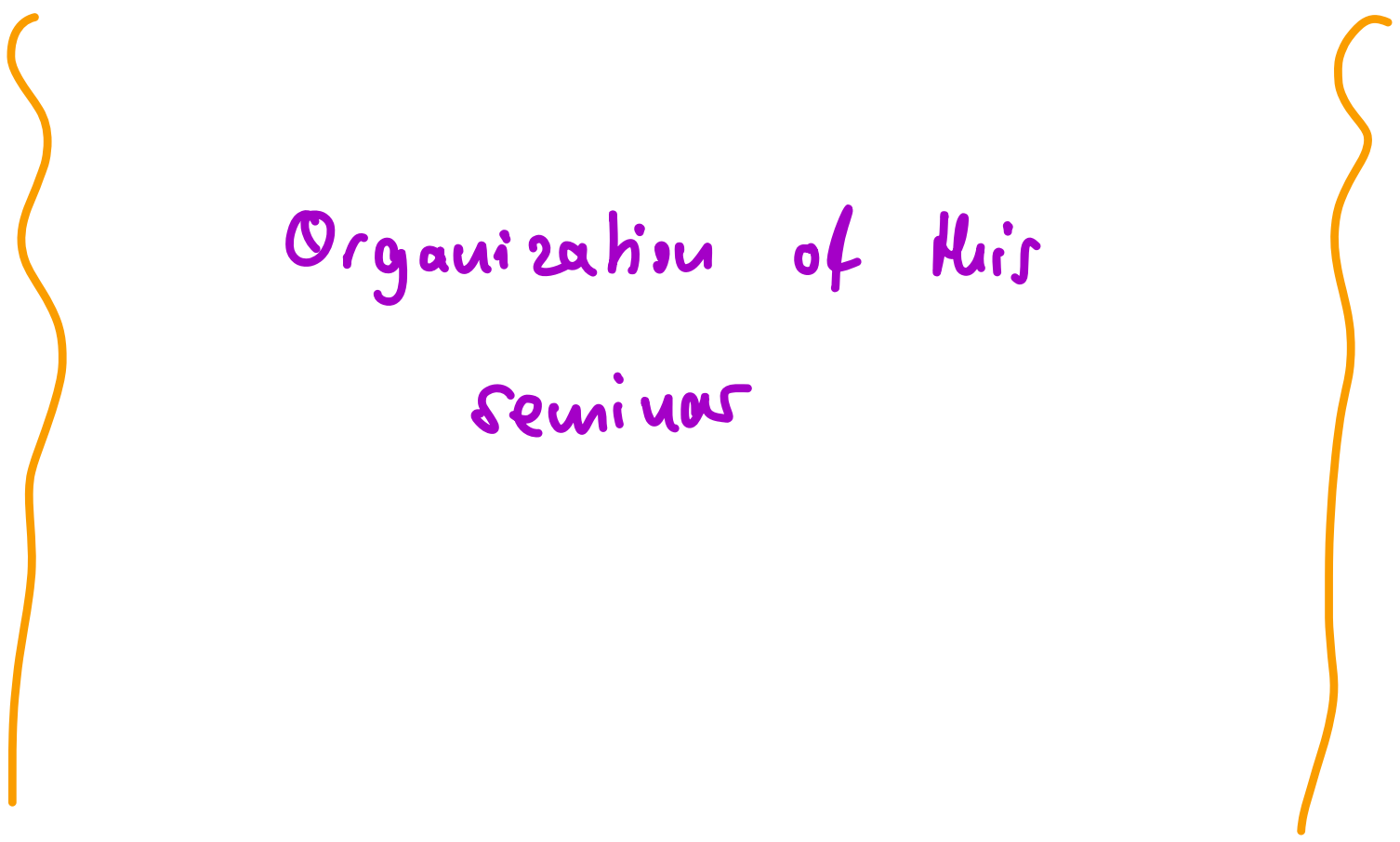
(Informative explanations only exist for simple functions)

Different goals of explanations

To make progress, it might make sense to differentiate between different goals for which one requests an explanation:

- detect discrimination
- make suggestions for recourse
- causal understanding

Or it might make sense to look into specific ML models together with specific explanation algorithms



Organization of this seminar

~ Webpage, lias !

Time line

Phase 1 : • two lectures to cover the basics and establish terminology (next two weeks)

- Students select paper to present and get assigned

Phase 2 : Students prepare their presentations
(meeting with their supervisor before christmas)

Phase 3: Presentations by students on two full days
(tentative : Jan 21 + 22)

Paper bidding & assignment

- Paper bidding by next wednesday, Oct 21
- If you bid for a paper, also register in Iliad, so we have your data
- assignment in the following days, to be confirmed by wed Oct 28
- you get assigned a supervisor
- Once you confirmed, you cannot step back and not participate.

Preparing your Presentation

- By end oct you get assigned a paper and a supervisor
- By Dec 1st you need to have contacted your supervisor and agreed on a first meeting time.
- By dec 11 you need to submit a first version of the slides on lias and discuss it with your supervisor before christmas.
- final presentation Jan 21/22

Preparing your Presentation

- Your presentation will be 20 minutes plus 10 min of discussion, in english

How to prepare the presentation

- Read paper thoroughly, understand all of it
- 10 - 20 slides
 - what is the problem addressed in the paper?
 - what is the suggested solution? Main message?
 - How do the authors substantiate their claims?
Experiments? Theory? Present it
 - Be critical: does the evidence support the author's claim?
 - Is the question / answer interesting?
- last slide: why did you choose the paper, and did you now like it?

Prachice !

- Practice you talk. often! Do it!

Requirements and grading

To pass the seminar, you need to

- Be able to explain the contents of the paper to your supervisor before christmas
- Discuss a first draft of the slides with your supervisor before christmas
- give a presentation about your paper
- Be able to answer questions about the paper.
- be present at both compact days

Chat GPT / Claude

- You need to fully understand the contents of your paper.
- You need to prepare the slides yourself, without any use of LLMs

Selection of participants

If we have too many participants...

- restriction to MSc students
- restriction to students with ML background
(at least one of the MSc lectures)

Among the remaining students:

- everybody can vote for a paper to present
- we will then select randomly

If you did not get assigned a paper, you can still stay and listen, but you won't get any credits for it.

Questions ?