Postgraduate course
Universitat de Valencia 2020

Introduction to Machine Learning for physicists

Veronica Sanz (UV/IFIC)

LECTURE 8
XAI

Explainable AI?

During this course
we have learned a good array of techniques in ML
accuracy in benchmark datasets has been ~90%
& often running a NN one feels a good job means that kind of accuracy

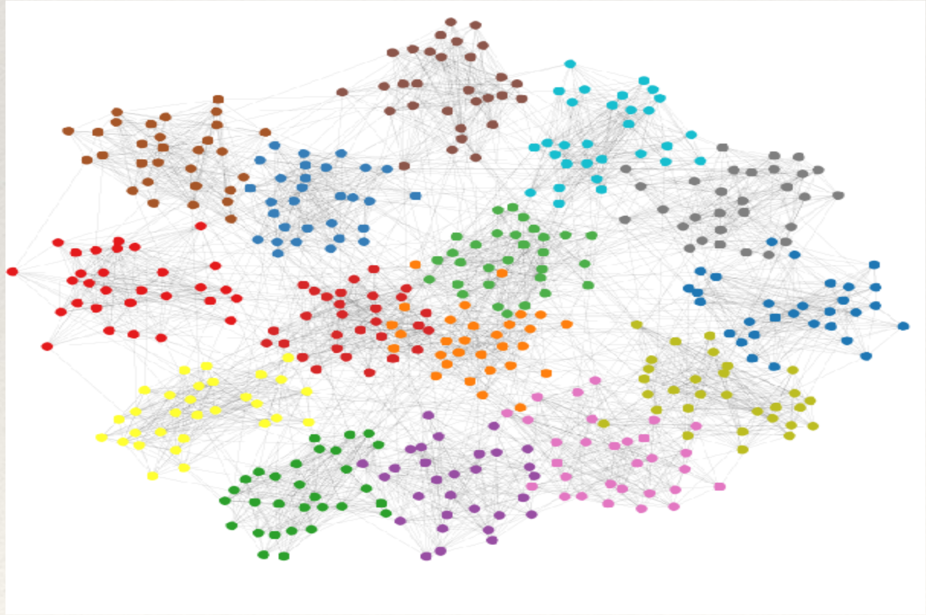
Also in this course we learned that ML can
beat humans in their highest-level strategic tasks,
help speed up difficult computations billions of times,
handle symbolic expressions,
solve *impossible* inverse problems,
find hidden symmetries….

Explainable AI?

We are talking about powerful stuff
with direct societal impact

With a simple hardware setup we can track and ID hundreds
of people in real time

we can scout online posts to gauge sentiment, cluster
individuals based on electricity use, predict sexual/political
orientation from a few clicks…



Explainable AI?

So, yes, we cannot just hold AI’s hand,
close our eyes and jump with it

From an ethical perspective: we need to make sure decisions based on AI comply with human policies

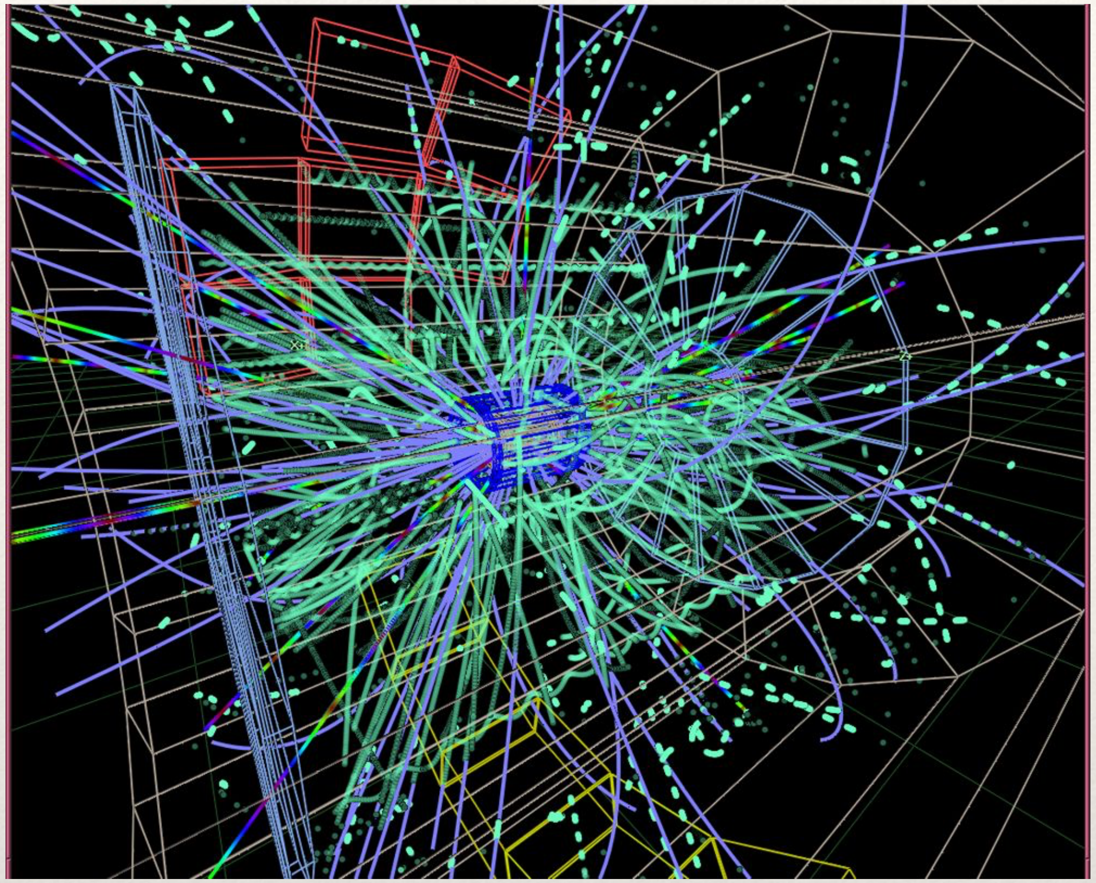
AI is a tool, not an aim

From a practical perspective: breakthroughs come from poking around big solid castles like AI

Finding what AI *does* can help us discover new techniques

To trust AI’s decisions and help on improving them
we need AI to become more ‘human-readable’

The use of XAI: Particle Physics example



Let's say you train a DNN to learn from
X=huge dataset of raw collision images
to identify
y=New Physics/Known Physics
and your algorithm gets super-good at it

Super-good, but not good enough
because you expect too few events to ever
discover this, even with 10 ab^-1

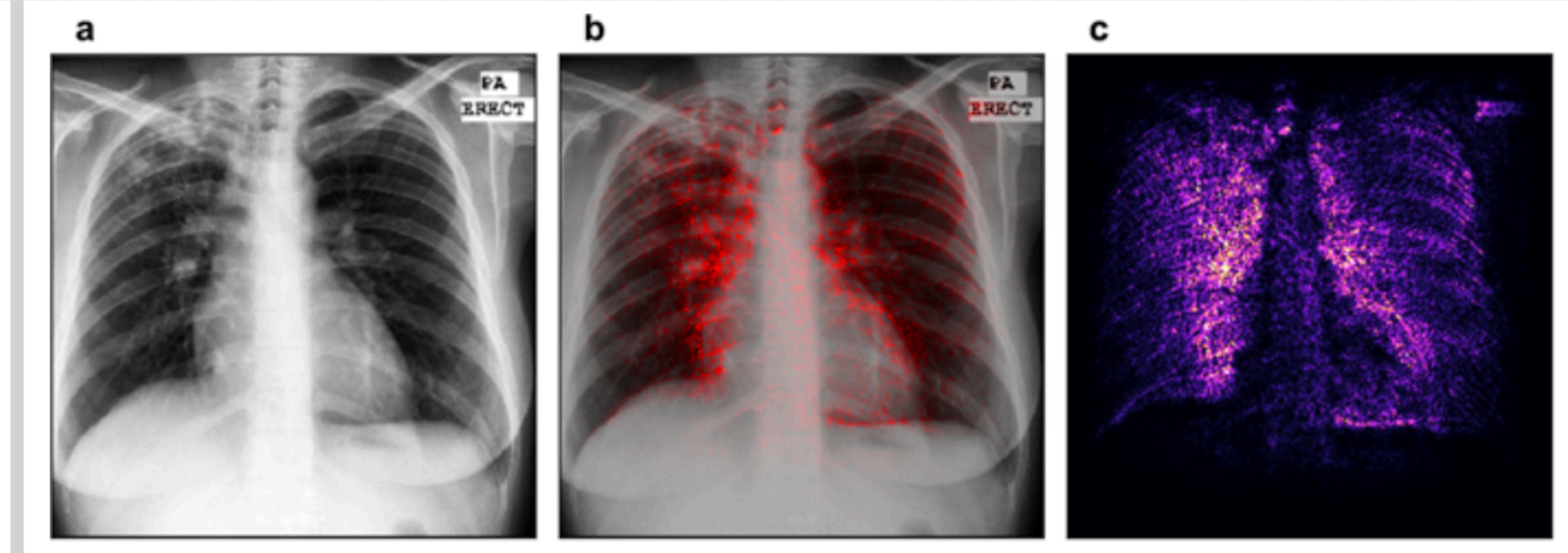
What else could your algorithm do for you? It became very good at finding new phenomena, so it must be that it *saw* something in the data

How?

The use of XAI: Particle Physics example

For example, if you have used some level of CNNs
some inputs layers are images and the training will give us regions of
these images which activate more strongly the neutrons
[*saliency map*]
there are some techniques to use saliency maps to
visualise AI's inner workings, see e.g. this paper for detector monitoring
using these techniques

For example, one widely used in ML&medical is Grad-Cam



Example from Nature paper

The use of XAI: Particle Physics example

Back to New Physics
by doing an XAI analysis on the results of your algorithm
you may realise there are some typical salient features in the images it identifies as new phenomena

Then you may realise that this particular set of configurations could be enhanced by changing the selection trigger at the analysis level
or by proposing a modification on the trigger menu

Re-running with an improved trigger, you may go from S/B <<1 to ~1

The use of XAI: Ethical example

Let’s say you are working in a company in the insurance sector

Your task is to assess the level of risk of customers to fix a premium

You have a huge dataset you can mine

X=customer descriptors

and

y=history of claims

X=name, DOB, address, level of studies, gender, level of studies, medical history details, facebook friends…

You run a ML algorithm and become extremely good at predicting **y** hence you can compute the most *fair* premium for your company/customer

The use of XAI: Ethical example

Let’s say you are working in a company in the insurance sector

Your task is to assess the level of risk of customers to fix a premium

You have a huge dataset you can mine

X=customer descriptors

and

y=history of claims

X=name, DOB, address, level of studies, gender, level of studies, medical history details, facebook friends…

You run a ML algorithm and become extremely good at predicting **y** hence you can compute the most *fair* premium for your company/customer

Then you start getting itchy thinking that you allowed a blackbox take decisions which affect people’s life e.g. some customers won’t be able to get insured or better, you get an audit and have to explain why you took these decisions

The use of XAI: Ethical example

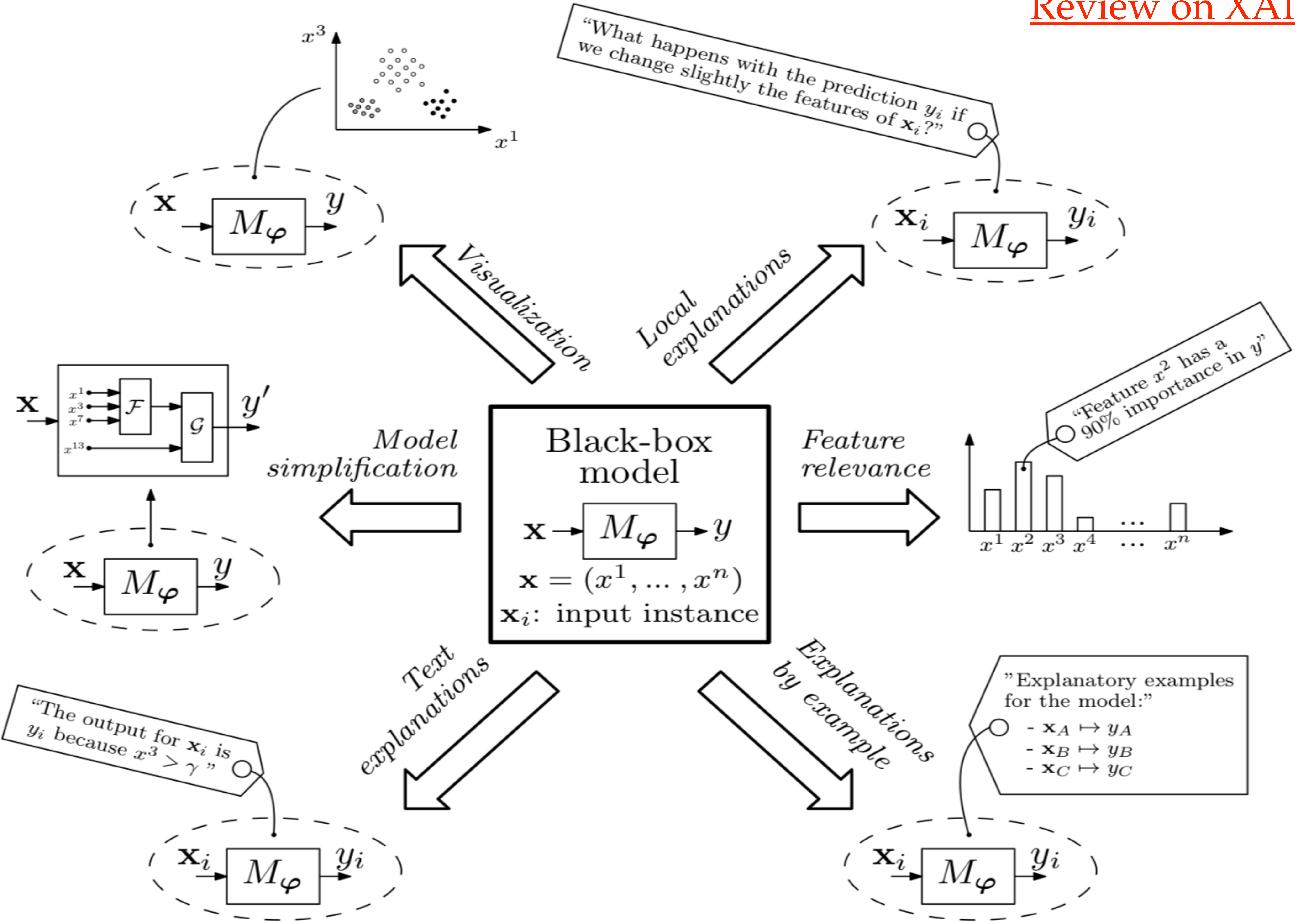
To post-hoc understand the ML
you may want to run the same dataset over a Boosted Decision Tree
you get less accuracy but can do feature importance
or you may drop features and realise these were important as accuracy drops
or you may want to run PCA and clustering to understand features in the data

Let's say that after all this digging you realise the main predictor for
y = accepted/refused application
is some combination of
(Address/Name) which seems correlated with religion, or
(level of studies/age/gender) which seems correlated with political orientation
and that predictions for minorities were substantially worse than the rest

So you have to conclude that your company is making discriminatory
decisions based on legally protected characteristics of individuals
hence is breaking the law

See for example this project DiCE

Review on XAI



Black-box model

$\mathbf{x} \rightarrow M_{\varphi} \rightarrow y$

$\mathbf{x} = (x^1, \dots, x^n)$

$\mathbf{x}_i$: input instance

Visualization

Local explanations

"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"

Model simplification

Feature relevance

"Feature x^2 has a 90% importance in y"

Text explanations

"The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$"

Explanations by example

"Explanatory examples for the model:"

- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$

Today

notebook on

X=characteristics of citizens US census

y=salary above or below $50K

binary logistic regression

I run a shallow NN and get 85% accuracy

(best state-of-the-art)

We will have **no** 12:30 catch-up

From now on, you will keep working on your choice of assignment and sending questions via Slack