SMARTHEP

REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Why do Machines Learn?

Introduction to ML Theory & Common Misconceptions in ML-dev

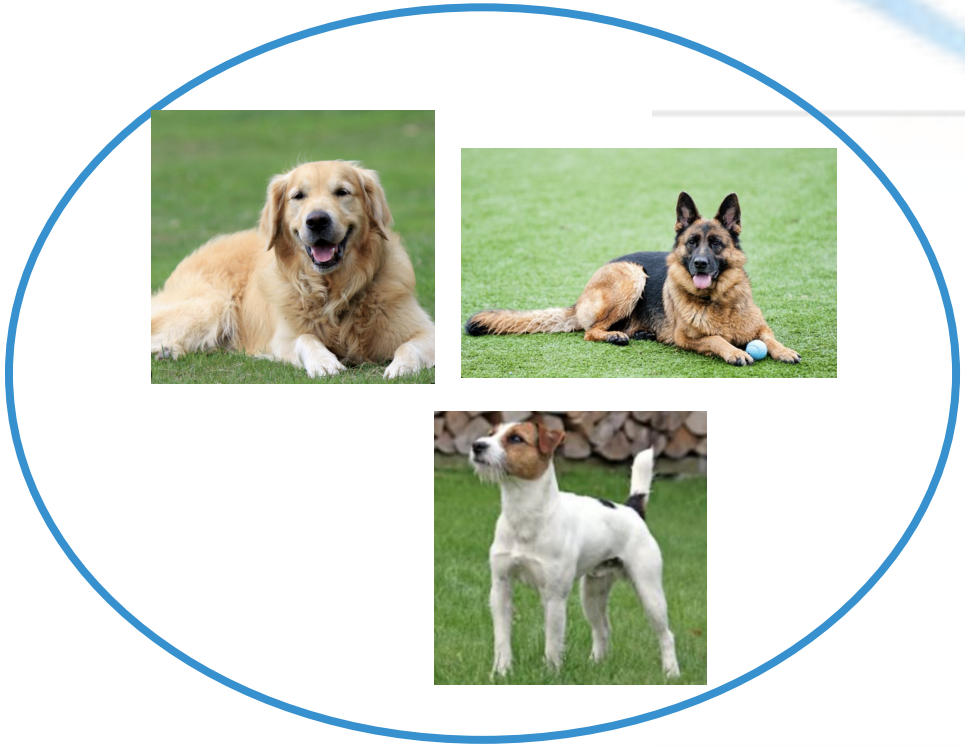
SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Pratik Jawahar - iCSC '24

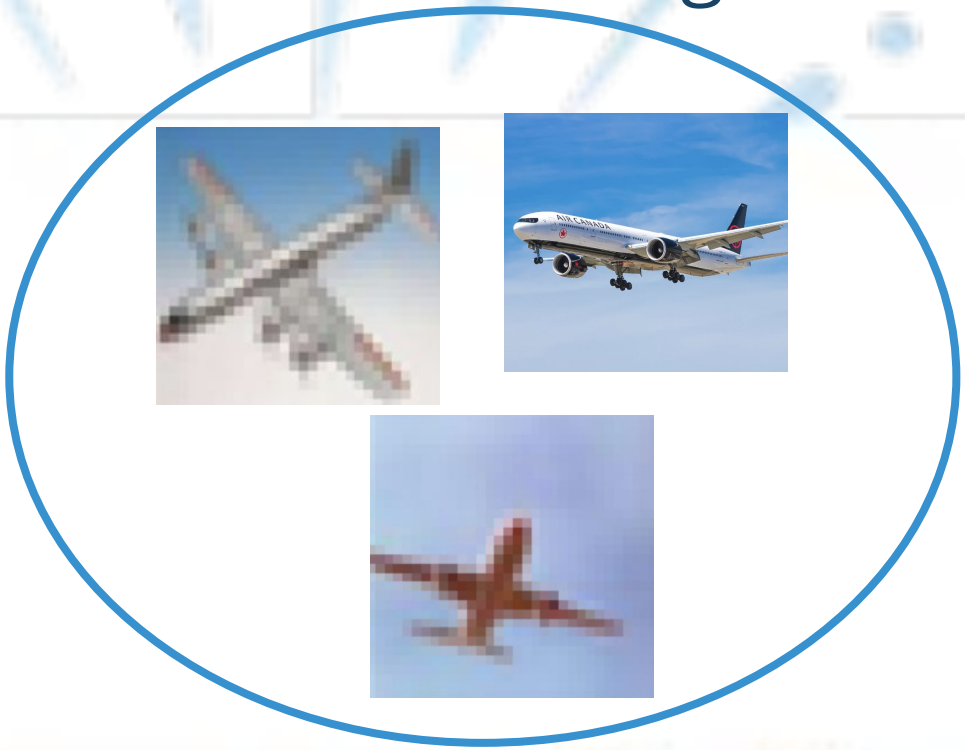
MANCHESTER 1824
The University of Manchester

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Consider a classifier trained on these 6 labeled images



Class A



Class B

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

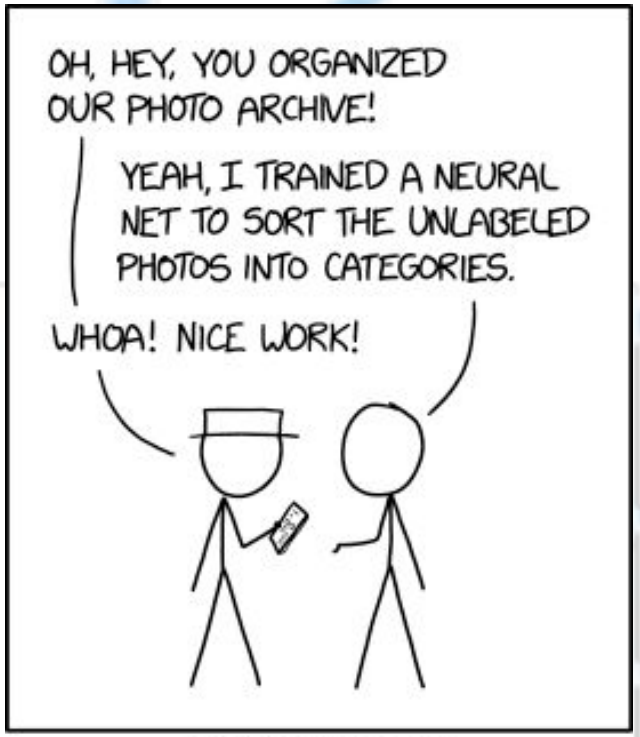
What class will the trained classifier predict here?

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

What if I told you the classifier model was a BNN!

xkcd's stick figure scientists are now upset and they are ready to cancel you!



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

What information do you need to be able to answer this question?

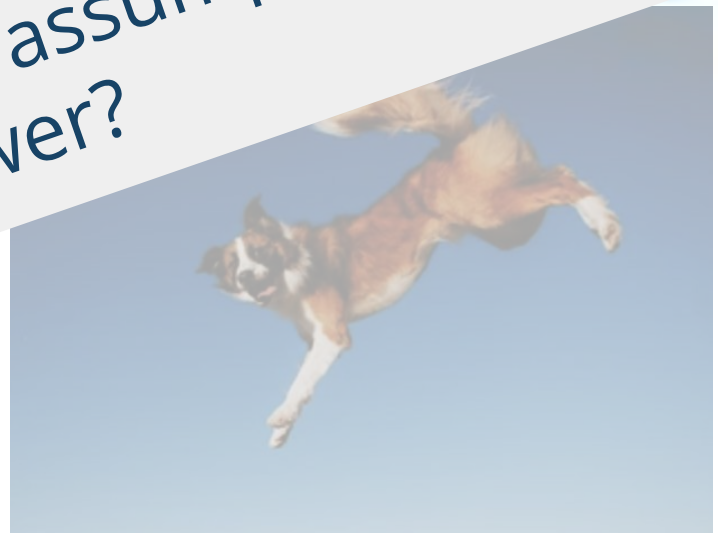
- What class will the model predict here?

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

- What d[illegible]

Based on the given information (this is an ML talk, pictures you saw on the previous slide etc.) what assumptions did you make before deciding on your answer?



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

The Example Bias

- Examples provided in documentation are almost never inclusive of all capabilities
  - But they are easy to {cmd+c; cmd+v}
- The problem:
  - Its easy to copy examples as is from research papers
  - Researchers building on top of such a paper, propagate the example to the point where the example becomes convention

```
import pandas as pd
pd.DataFrame({'A': [1, 2, 3]})
```

```
>>> import numpy as np
>>> a = np.arange(15).reshape(3, 5)
```

```
import h5py
f = h5py.File('mytestfile.hdf5', 'r')
```

```
import torch.nn as nn
import torch.nn.functional as F

class Model(nn.Module):
    def __init__(self):
        super().__init__()
        self.conv1 = nn.Conv2d(1, 20, 5)
        self.conv2 = nn.Conv2d(20, 20, 5)

    def forward(self, x):
        x = F.relu(self.conv1(x))
        return F.relu(self.conv2(x))
```

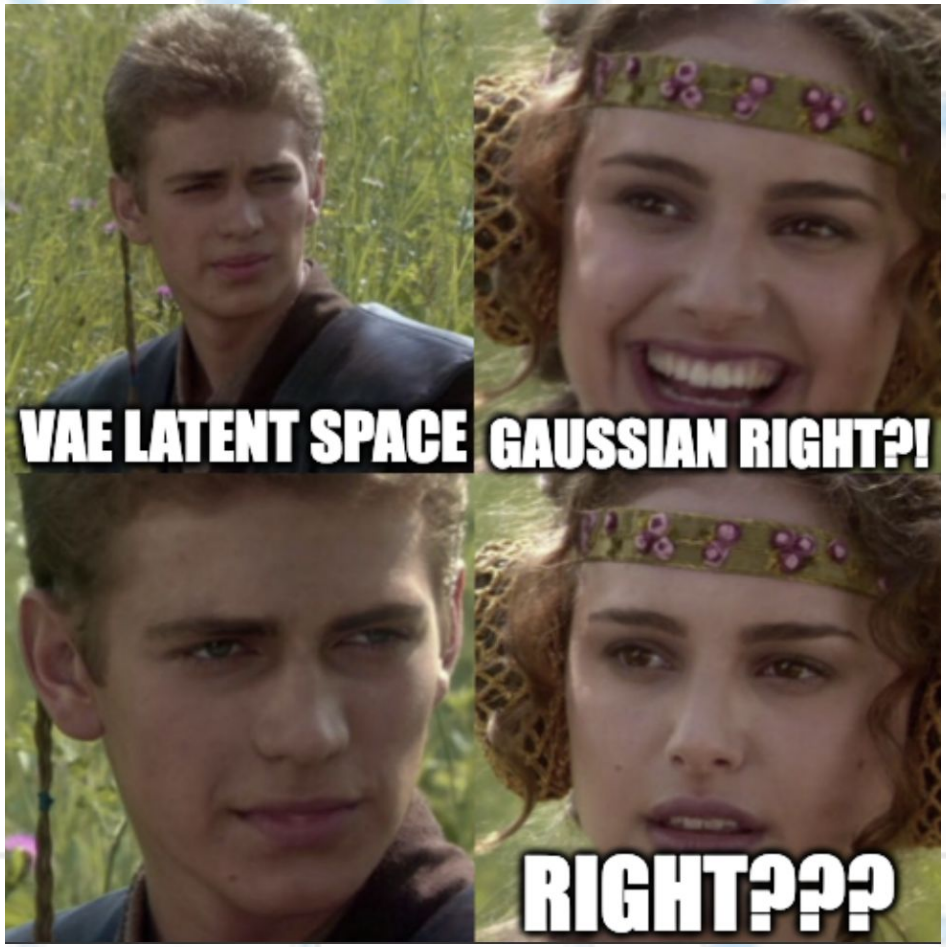
SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

The Example Bias

- The VAE: Extract from Tutorial on Medium [link]
- Posterior is approximated as a multi-variate normal distribution as defined in the original VAE paper

```
def reparameterization(self, mean, var):
    epsilon = torch.randn_like(var).to(device)
    z = mean + var*epsilon
    return z
```

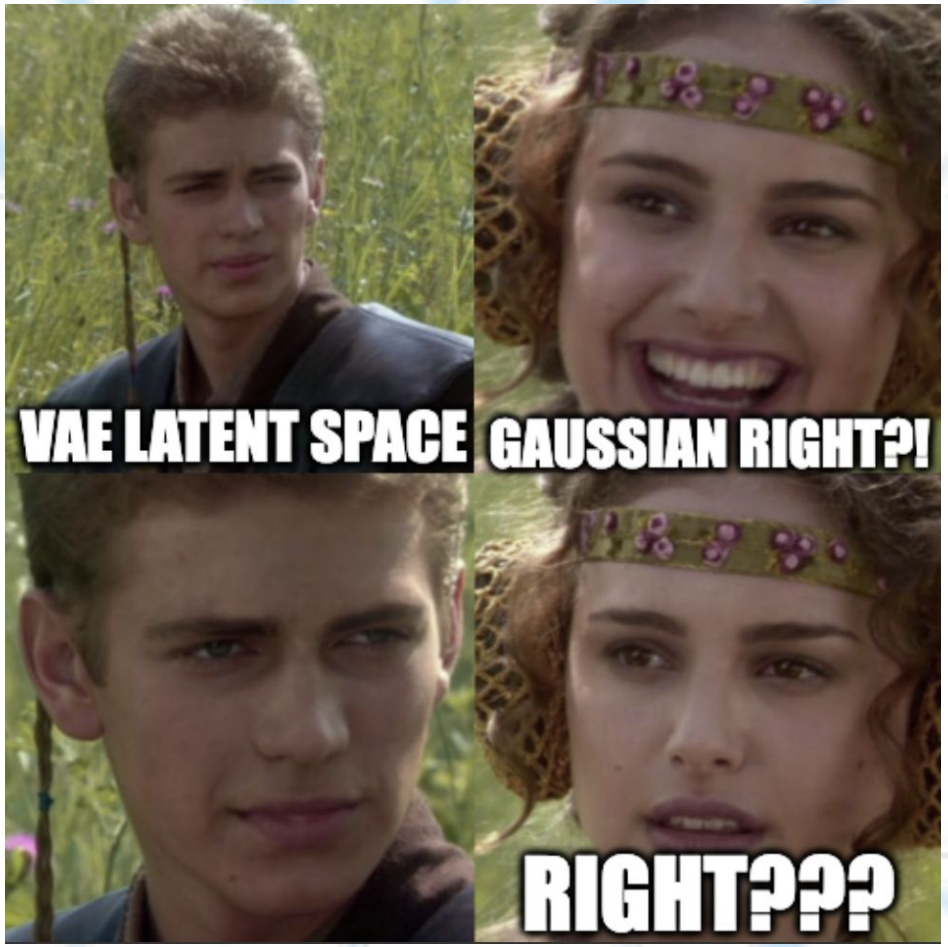
```
def loss_function(x, x_hat, mean, log_var):
    reproduction_loss = nn.functional.binary_cross_entropy(x_hat, x,
    KLD = - 0.5 * torch.sum(1+ log_var - mean.pow(2) - log_var.exp())
```



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

The Example Bias

- The VAE:
  - The original VAE paper lists more possible posterior distributions for which the reparametrization trick works just as easily: https://arxiv.org/abs/1312.6114
  - Hyperspherical VAE: https://arxiv.org/abs/1804.00891
  - Complex Latent Spaces using Normalizing Flows: https://arxiv.org/abs/1505.05770



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

The Example Bias

- Will let you uncover this misconception yourself

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

The Solution?

- IDK!

A Solution?

- Develop theory-informed intuitions for conventional choices
- This talk is meant to be a preliminary synopsis of resources
  - theory, on most moving parts of an ML workflow, to be considered before diving into ML-Dev!

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

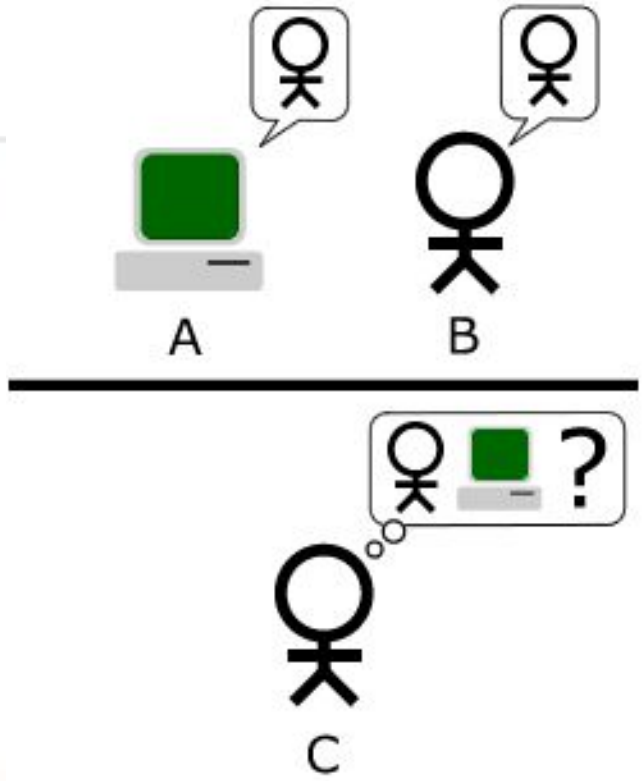
What is ML?

- First recorded mentions of ML come from Alan Turing in the 1940s via meetings of the Ratio Club (a dining club for researchers discussing cybernetics)
- Turing uses the term *"Machine Intelligence"* in 1947 and publishes the foundational paper *"Computing Machinery and Intelligence"* in 1956 which is seen as the formal inception of AI
  - Turing Test is used to define Computer Intelligence
- ML algorithms and their bases go way back, rooted in(non-exhaustive):
  - Statistics {late 1800s}
  - Psychology (psychometrics - latent variable models {1900s})
  - Cybernetics (feedback models {1940s})
  - Neurobiology (McCulloch-Pitts neurons {1943})
  - Mathematics (Backprop, gradient descent are interpretations of the chain rule in calculus {1670s})
  - Early Deep Neural Network designs (ELM, Hopfield Networks, Helmhotlz machines, Boltzmann machines etc. {1950s on})

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

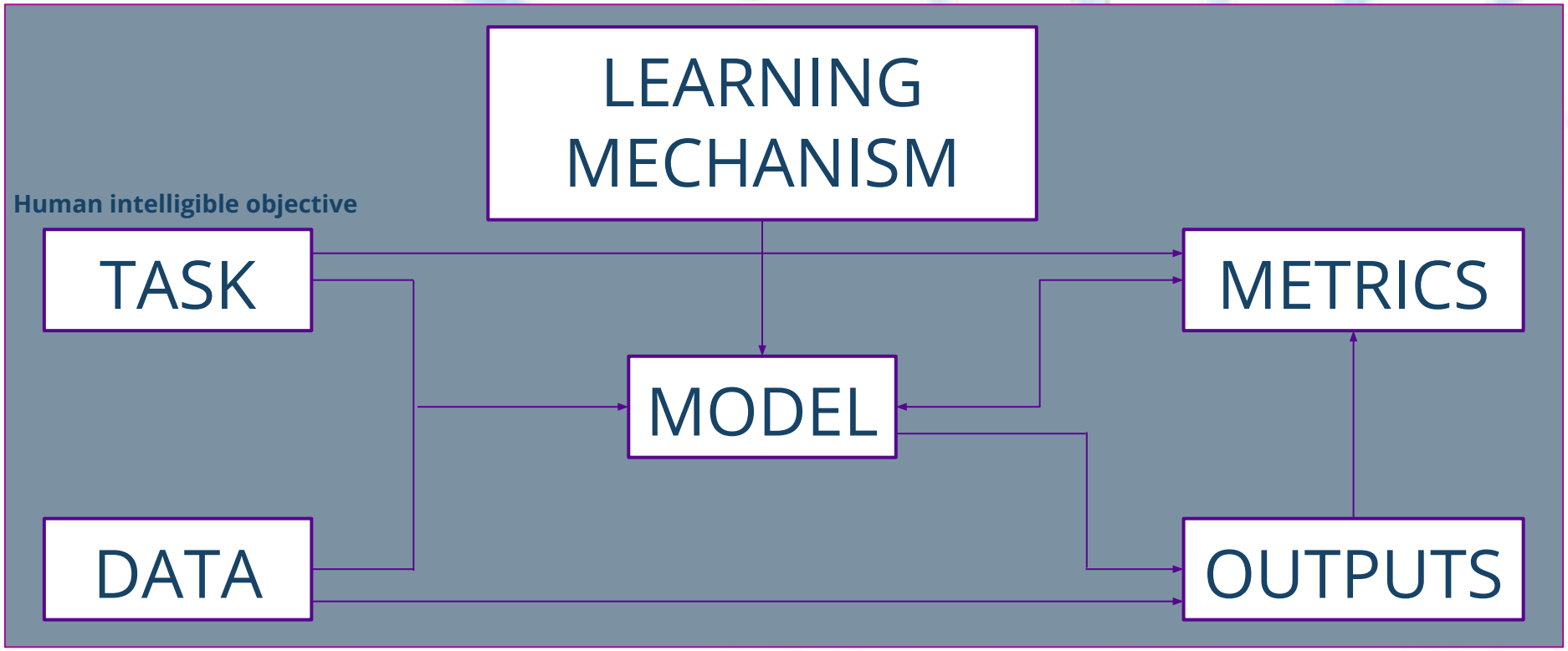
Turing Test

- A is a man, B is a woman, C is a moderator
  - C can only ask questions via written notes
  - A, B respond via notes from separate hidden rooms
  - C has to identify the man and the woman correctly
  - A tries to trick C into making an incorrect decision while B tries to assist C
- Now replace A with a computer
  - Can A trick C into thinking it is the human as opposed to B?
- A computer that can consistently trick C is considered to be an *"Intelligent Machine"*



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

The General Pipeline

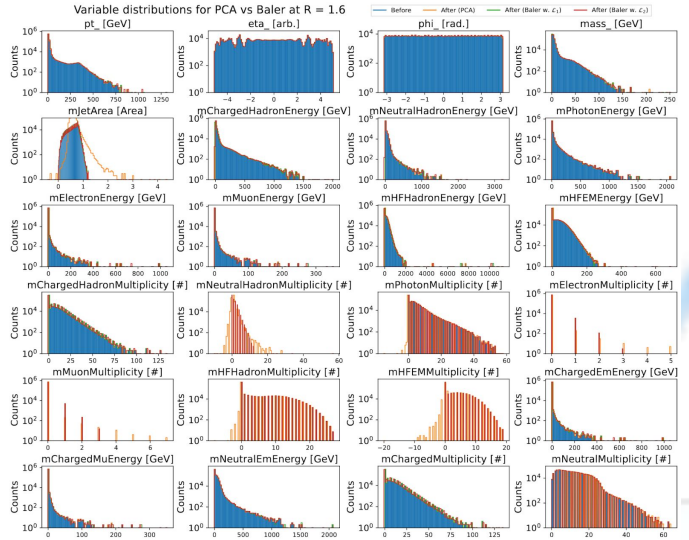


SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

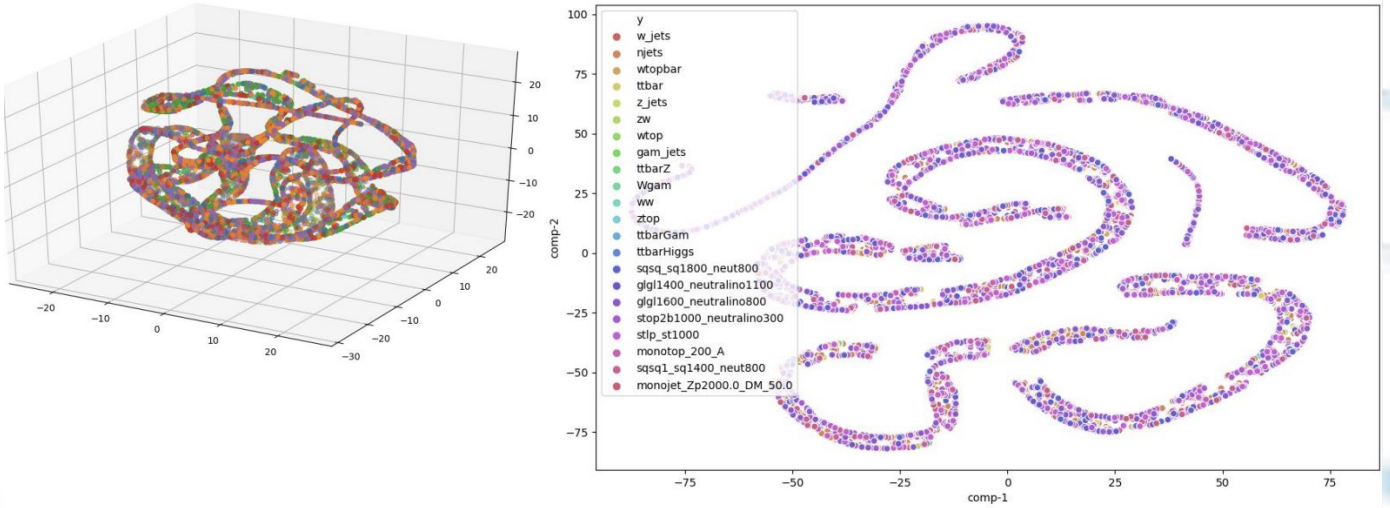
SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Data

- Collection source
  - Awareness on data collection process
  - Ensuring data collected is per expectation
    - Perform EDA checks (EDA is more of an art!)
      - Visualizing the dataset to understand its characteristics
        - PCA, TSNE, TriMap etc.
      - Books: [Philosophy of EDA]; [Practical Guide to EDA]



Histogram of all features in a jet-dataset in CMS Open Data

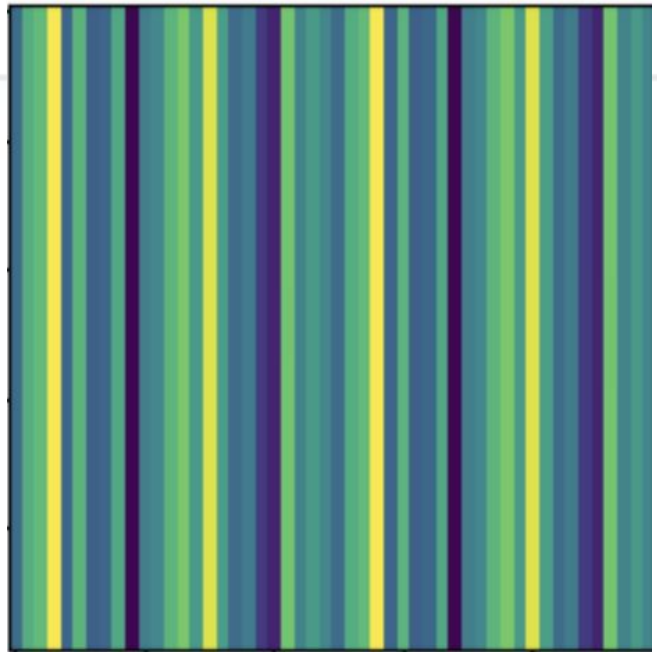


TSNE viz of the Darkmachines Anomaly Challenge Dataset

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Data

- Modalities
  - Representing human-level data in computer-readable formats
  - Choosing the right {data-modality, model} pair is essential
    - A non-sequence preserving model architecture to process video data gives up {vital} information along the time-dimension
    - Representing the 4-momentum as a .PNG and applying a CNN adds unnecessary spatial correlations between features that don't actually exist in the data
  - Understranding optimal pre-processing techniques for the chosen modality - **NO** one **method-fits-all** solution
    - [Preproc methods]; [Book on Preproc]

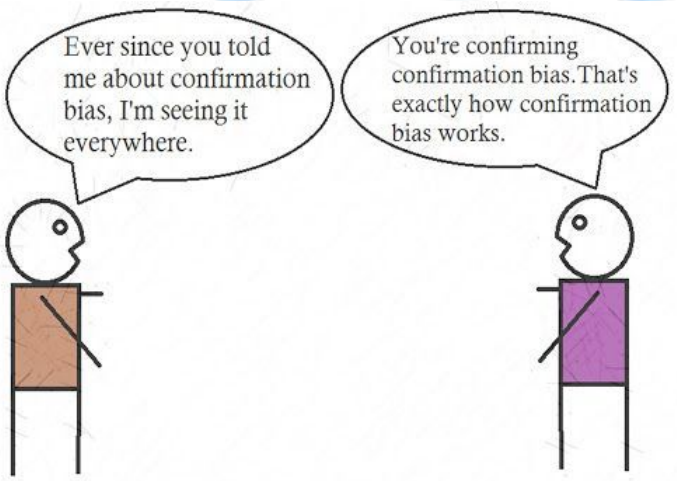


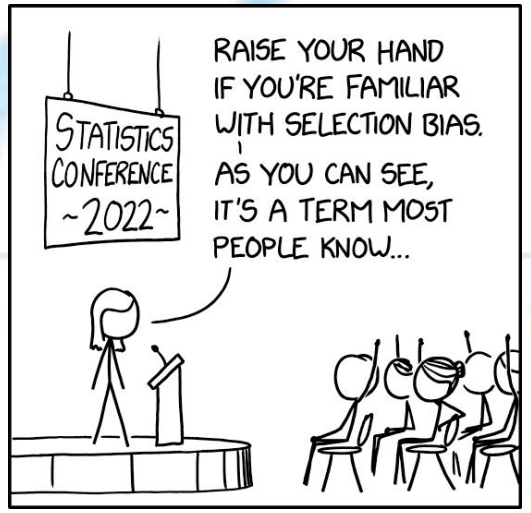
SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Data

- Bias:
  - Systemic
  - Automation
  - Selection
  - Reporting
  - Overgeneralization
  - Implicit
  - Group Attribution
- [Google developers blog]



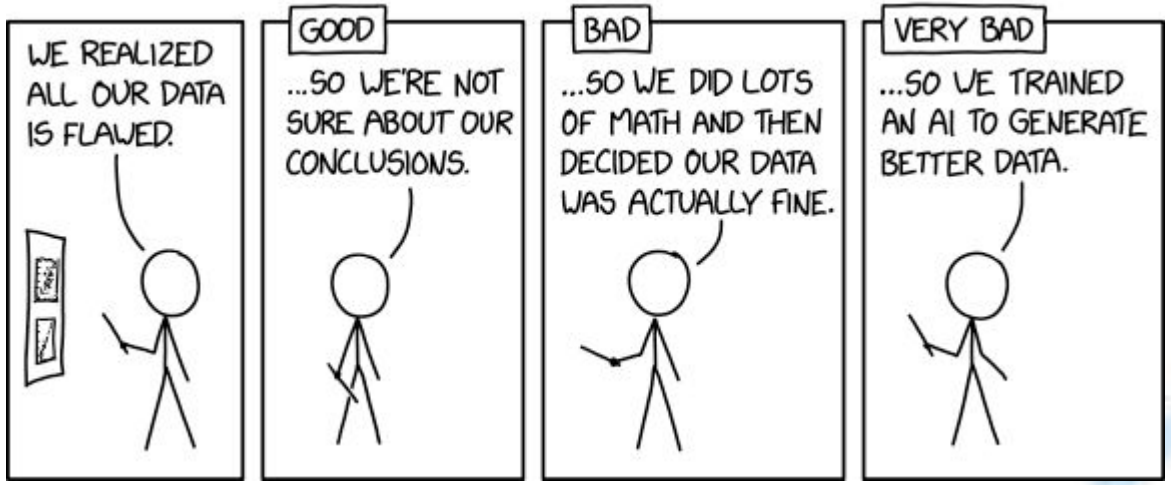


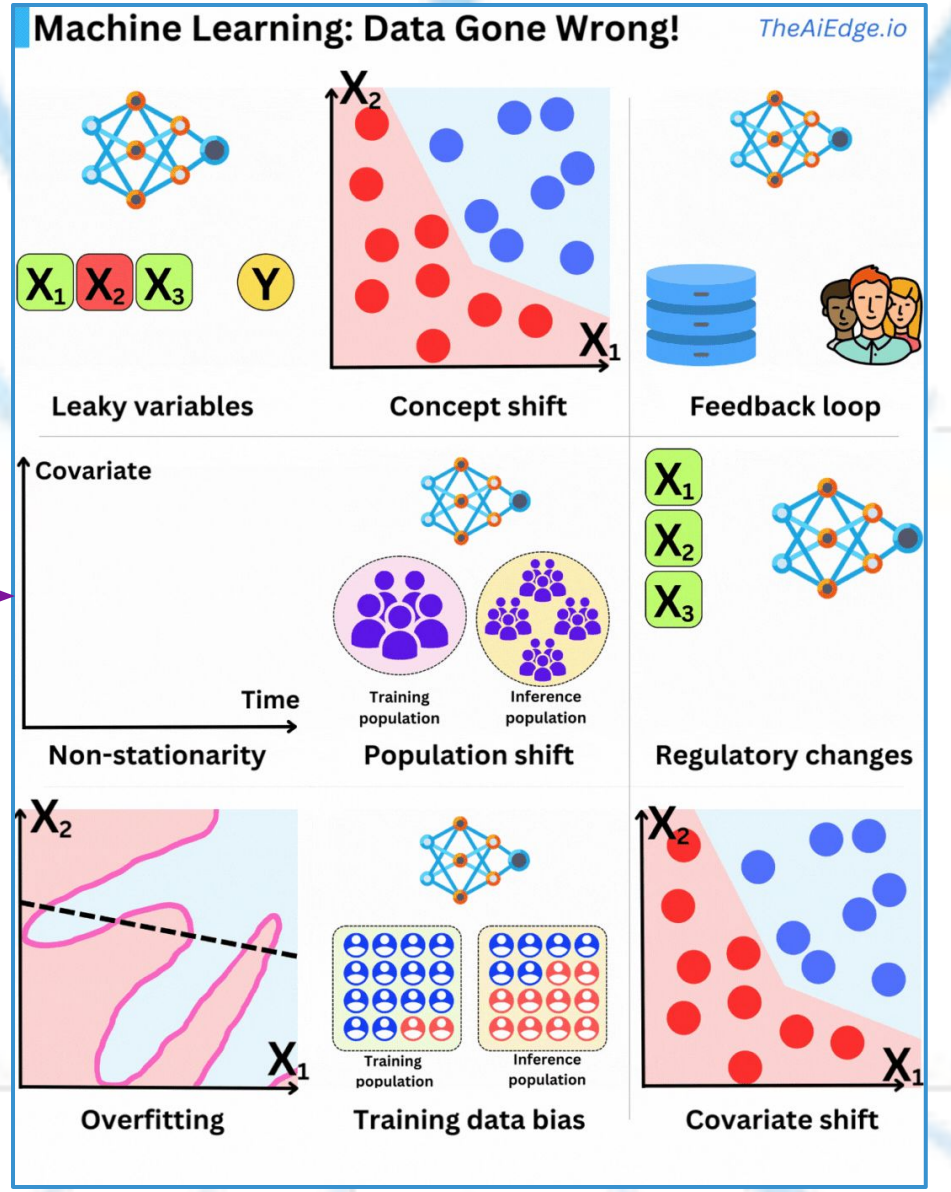
SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Data

Other common issues with data to be considered during dev: Blog



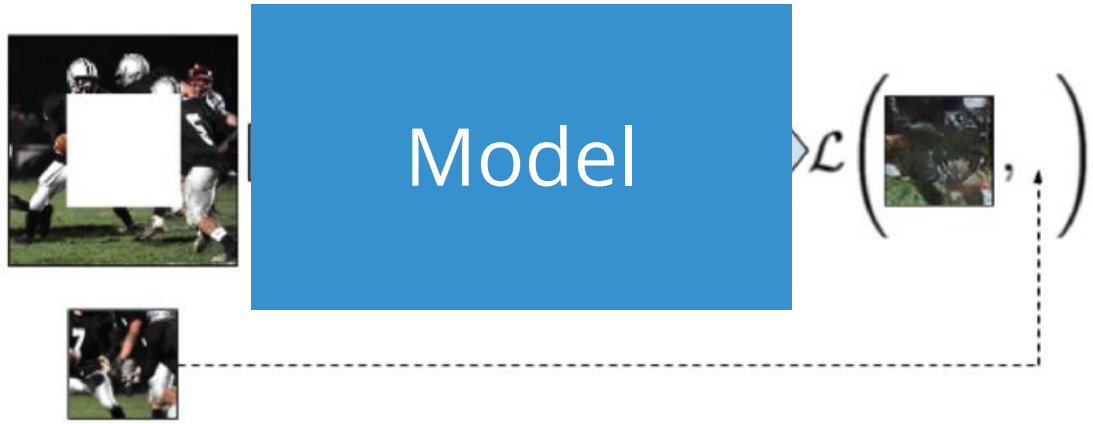


SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

The University of Manchester

Learning Mechanism

- Data:Target based classification:
  - Supervised
    - Data comes with labels
    - Model outputs are defined
  - Self-Supervised
    - No explicit labels
    - Train a model to define "its own" labels
  - Semi-Supervised
    - Somewhere on the spectrum - part labeled data, rest unlabeled
  - Unsupervised
    - No labels; No definition for outputs until a model is chosen
    - Model learns to map input data to points in an abstract feature space
      - For eg. in clustering, the abstract space could be the space containing the centers, boundaries of the clusters
  - Reinforcement
    - Optimal control paradigm that rewards/punishes "agents" trying to achieve a defined goal



Dont let the example bias get you!

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Model

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Model

- An ML model is an algorithm, not a black box
  - For eg. *standard* training of a single node perceptron is a convex optimization problem (i.e. all achievable local minima are the global minimum)
  - Many ML algorithms (eg. SVMs, logistic regression etc.) have polynomial time guarantees
- The macroscopic effects of a *complex* ML (read: DL) model can be blackbox-like (read: NP-hard optimization problem) [S. Judd's thesis is foundational work]
  - A perceptron as small as 2 layers with 3 nodes each is intrinsically hard to optimize [Blum, Rivest 1993]
- So are DL models blackboxes? Or the optimization algorithms used to train them?
  - These optimization algorithms are part of the Learning Mechanism, but first lets focus on models

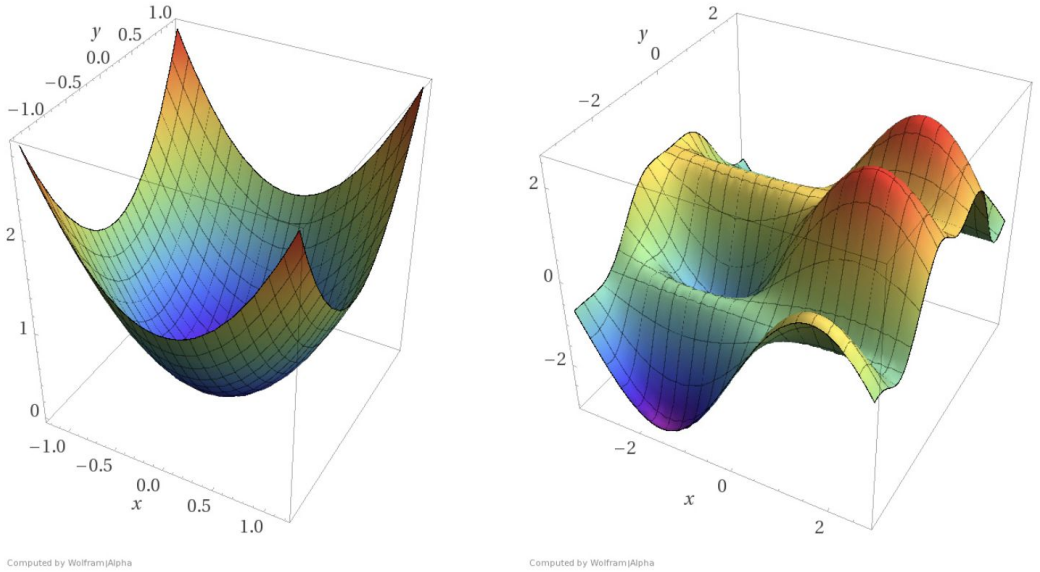


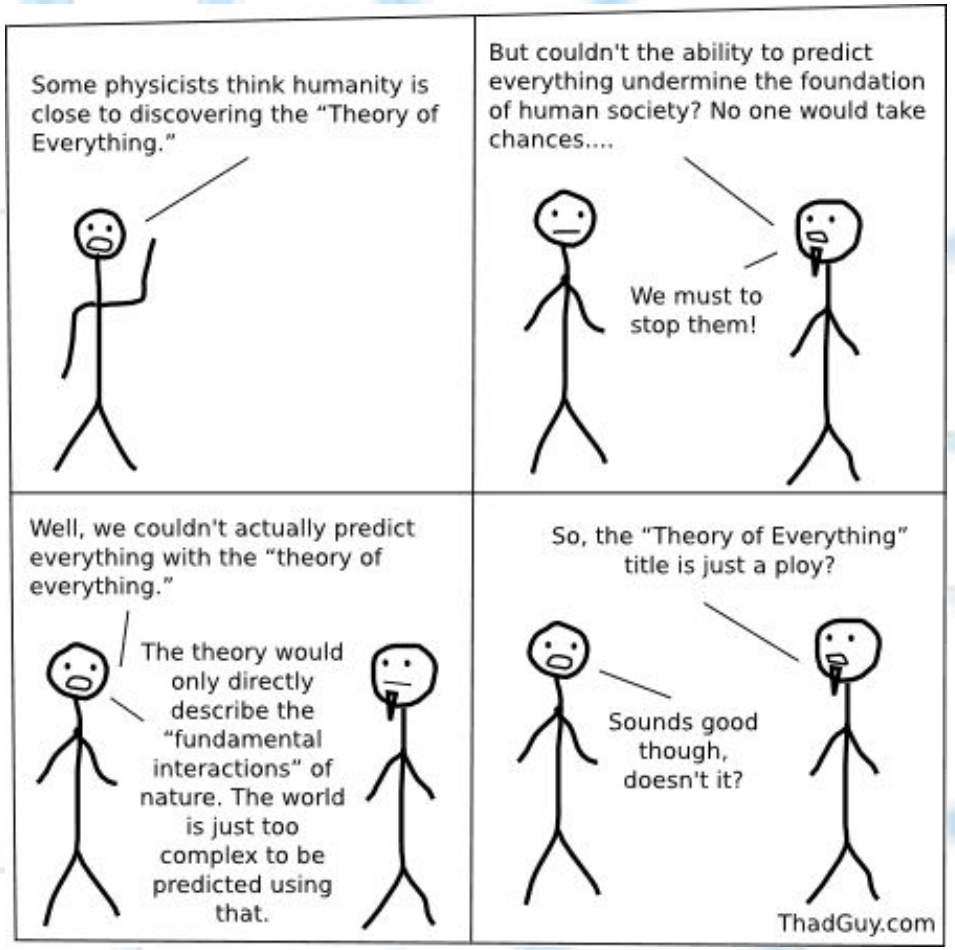
Figure 1. Left: a convex function. Right: a non-convex function. It is much easier to find the bottom of the surface in the convex function than the non-convex surface. (Source: Reza Zadeh)

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Model

- How do we define every possible ML model under a common mathematical framework?
  - Yes, Theoretical ML is also on the open quest for a "Theory of Everything" for ML models
  - An *elegant framework* from which all models can be derived
- Efforts so far:
  - Top-Down: Define task based constraints that drive the model design
    - eg. LSTMs can be defined as *sequence preserving* models, CNNs as models sensitive to *spatial-correlations* etc.
  - Bottom-Up: Define models based on the tensor level operations/transformations they perform
    - eg. LSTMs defined by their operational gate diagram, CNNs with the convolution operator applied under constraints of parameters such as stride, edge-handling etc.
- How do you bridge these two approaches? Open Question!

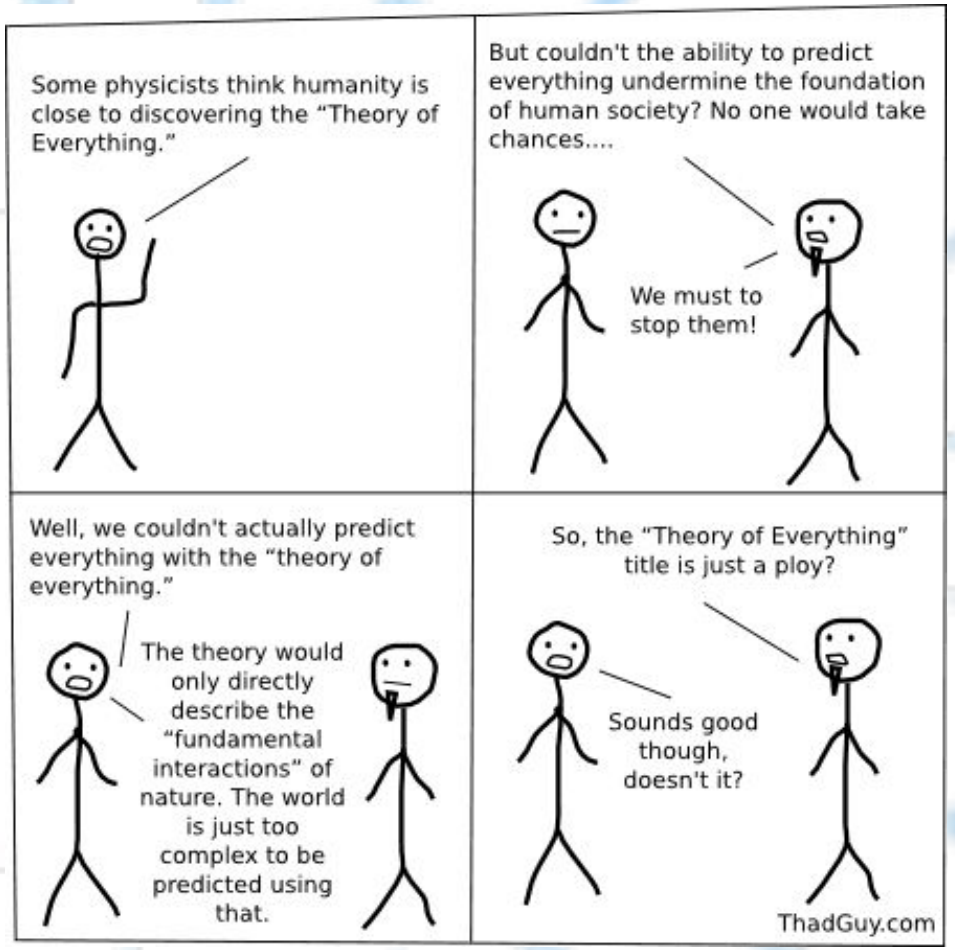


SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Model

- How do we define every possible ML model under a common mathematical framework?
  - Yes, Theoretical ML is also on the open quest for a "Theory of Everything" for ML models
  - An *elegant framework* from which all models can be derived
- But why do we care about a "Theory of Everything" in ML?
  - A neat framework to define all existing algorithms (Helps zoom out and see the bigger picture; given how much of a depth-first search ML research has become)
  - Helps in structuring the discovery of future architectures/algorithms



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Model

- So how do we unify the top-down and bottom-up approaches?
- Progress so far:
  - Kernel methods
    - Kernels are formal definitions of dot products
    - Any algorithm that interacts with data only via dot products - it is a kernel method
      - Perceptrons, SVMs, linear regression, k-means clustering etc.
    - Bottom-up definition because you define the class of operations first and build up
    - If you can get an idea of the distribution of mappings of input datapoints in the kernel-space, based on the margins you can say for eg. if a Perceptron will be able to converge to a minimum efficiently or how many samples an SVM will require to be able to generalize

The following are a few examples of important kernels.

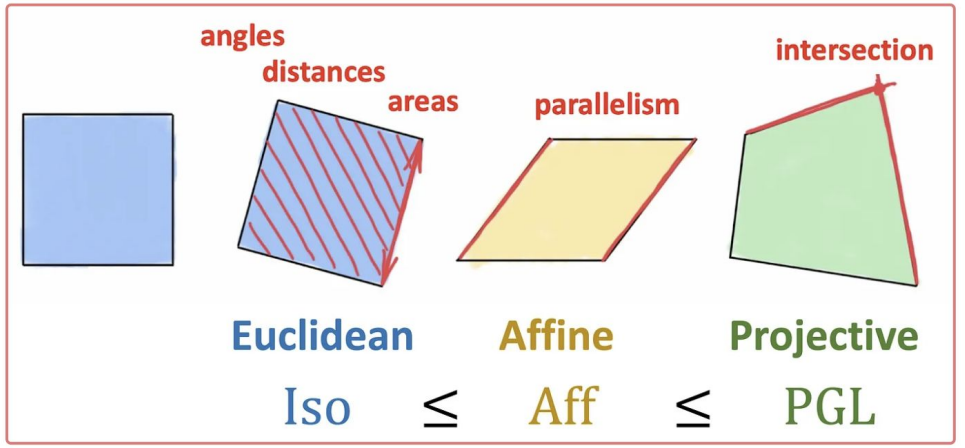
- **Linear:** $K(x, z) = x \cdot z$.
- **Polynomial:** $K(x, z) = (x \cdot z)^d$ or $K(x, z) = (1 + x \cdot z)^d$.
- **Gaussian:** $K(x, z) = e^{-||x-z||^2/(2\sigma^2)}$.
- **Laplace:** $K(x, z) = e^{-||x-z||/(2\sigma^2)}$.

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

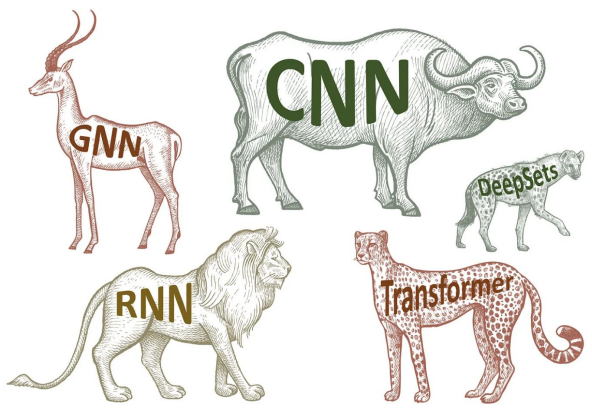
SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Model

- So how do we unify the top-down and bottom-up approaches?
- Progress so far:
  - Geometric DL (2021)
    - Follows *Erlangen Program* philosophies of looking at geometry as a study of invariants
      - Find transformations under which the properties you care about a specific geometry are invariant and use this basis of transformations as your geometric definition
    - Successfully describes most commonly used mechanisms eg. in describing:
      - Conv layers as an exact solution of linear translation equivariance in grids
      - Message-passing and self-attention as instances of permutation equivariant learning over graphs
      - It also extends naturally to exotic spaces such as spheres, meshes etc.
    - But not all ML transformations we'd like to study are invertible - and thus can't be studied as equivariance relations!



Klein's Erlangen Programme approached geometry as the study of properties remaining invariant under certain types of transformations. 2D Euclidean geometry is defined by rigid transformations (modeled as the isometry group) that preserve areas, distances, and angles, and thus also parallelism. Affine transformations preserve parallelism, but neither distances nor areas. Finally, projective transformations have the weakest invariance, with only intersections and cross-ratios preserved, and correspond to the largest group among the three. Klein thus argued that projective geometry is the most general one.



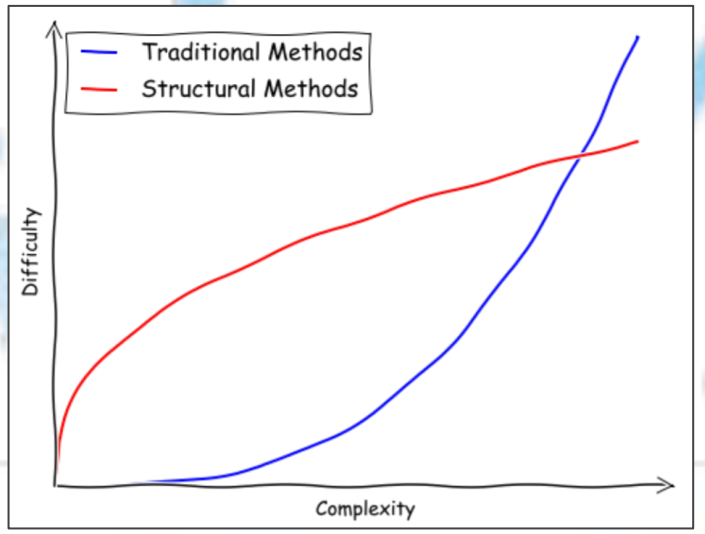
Deep learning today: a zoo of architectures, few unifying principles. Animal images: ShutterStock.

Source

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Model

- So how do we unify the top-down and bottom-up approaches?
- Progress so far:
  - Categorical DL (2024)
    - Most recent attempt at ToE based on *compositionality*
    - Categories are a collection of **Objects** and **Morphisms** between any two objects in the category
      - eg. A Set-Category has sets as objects and functions as the morphisms between sets
    - Homomorphisms are then used to generalize equivariance relations described in GDL
    - They go further by using the homomorphisms to also define constraints that describe the control flow of NNs, thereby beginning to address the top-down approach
    - Limitations: Currently only works for individual layers, not weight sharing between layers, which is essential to describe the non-linear maps in most Deep Networks



Plot by a researcher who loves structural methods

Source

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

So what's my point?!

- DL operation is not a blackbox mechanism (more grey tbh, like really dark grey)!
- It is getting less opaque and will continue to!!
- Knowing how to represent all models under a common theory also gives us a mathematical framework to:
  - Choose the best model for a set of task-level constraints (Top-down problem)
  - Choose the best model for the available compute resources (Bottom-up problem)

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

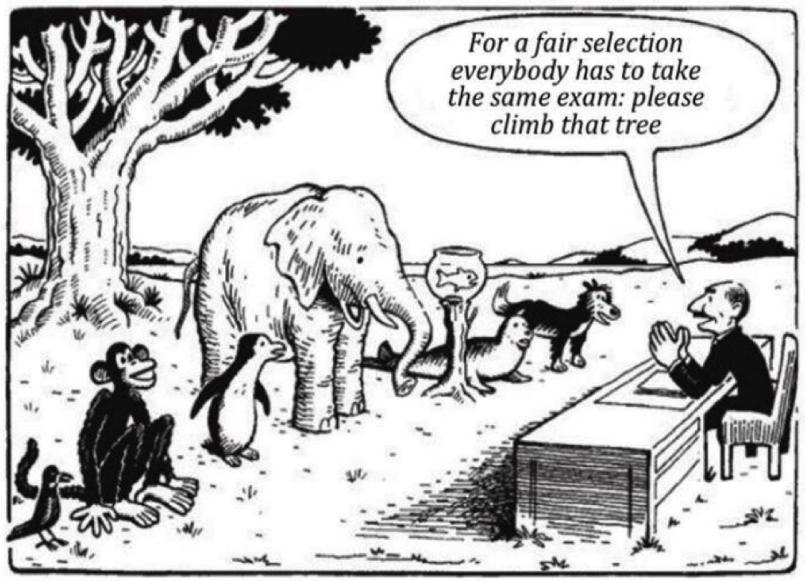
Metrics

- We now have
  - Shiny data
  - A loose Learning Mechanism (LM)
  - A model
- How do we make the model actually learn from the data based on the constraints laid by the LM?
- We need to add more features to the LM to enable this
  - What are computers better than humans at?
    - {Solving P problems; Verifying NP problems} faster than humans
- So let our LM convert the "learning problem" into an "optimization problem"
  - What do we optimize?
    - METRICS!! (Not really, we optimize objective functions against a metric, but go with the emotion not the words)

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Metrics

- Under this loose definition,
  - There is some cost function (loss) that most likely contains the inputs and the outputs of the model - we reduce the loss via optimizers
  - But how long do we keep going? Are we guaranteed to converge to the global minimum? If it doesn't do I have to pay electricity bills for a computer running for infinite time??
- This is where model performance metrics come in
  - Metrics are our way of defining when the training is good enough for us to stop
  - So the metric you pick is your way of judging if the model is good enough
    - Pick meaningful metrics based on your task requirements - do not set your model to fail
      - No it isn't enough to just read motivational quotes like the one on the right - you have to put it in practice



"Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid."

- Albert Einstein

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

Learning Mechanism (LM)

- Currently most LMs rely heavily on:
  - Loss function choice
    - The example bias plays in heavily here to make people assume there's a finite (read: countable on one hand) number of loss functions
    - Any function can be a loss function if it:
      - Can be used as an optimization objective (requires definition of optimization algorithm)
      - Is differentiable - optimization algorithms are nosy and usually want to know gradients
        - Gradients are easy/fast to compute
      - Incorporates the objectives of the task
      - Lays desired constraints on model updates
  - Optimization algorithms
  - Backpropagation to update model weights

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) \equiv \mathbb{E}((\hat{\theta}-\theta)^2) &= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}(\hat{\theta})+\mathbb{E}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}(\hat{\theta})\right)^2+2\left((\hat{\theta}-\mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta})-\theta)\right)+\left(\mathbb{E}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}(\hat{\theta})\right)^2\right]+2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta})-\theta)\right]+\mathbb{E}\left[\left(\mathbb{E}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}(\hat{\theta})\right)^2\right]+2(\mathbb{E}(\hat{\theta})-\theta)\overbrace{\mathbb{E}(\hat{\theta}-\mathbb{E}(\hat{\theta}))}^{=\mathbb{E}(\theta)-\mathbb{E}(\theta)=0}+\mathbb{E}\left[\left(\mathbb{E}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}(\hat{\theta})\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathrm{Var}(\hat{\theta})+\mathrm{Bias}(\hat{\theta},\theta)^2
\end{aligned}
$$

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Conclusions

- There are no one-size-fits-all answers to any ML question you may have while diving into ML-dev
- Steer clear of the example bias; Know your tools not just their usage
    - Form your own implicit bias instead (atleast that is conscious)
- A few hours on ML theory can take you a long way in ML-dev
- ML theory could use some physics theory at this point - consider joining the workforce
- Don't @ me if you spent hours designing the best loss function for your {data-task-model-LM} set and ended up reinventing MSE
    - **Sometimes**, things are popular for a reason

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

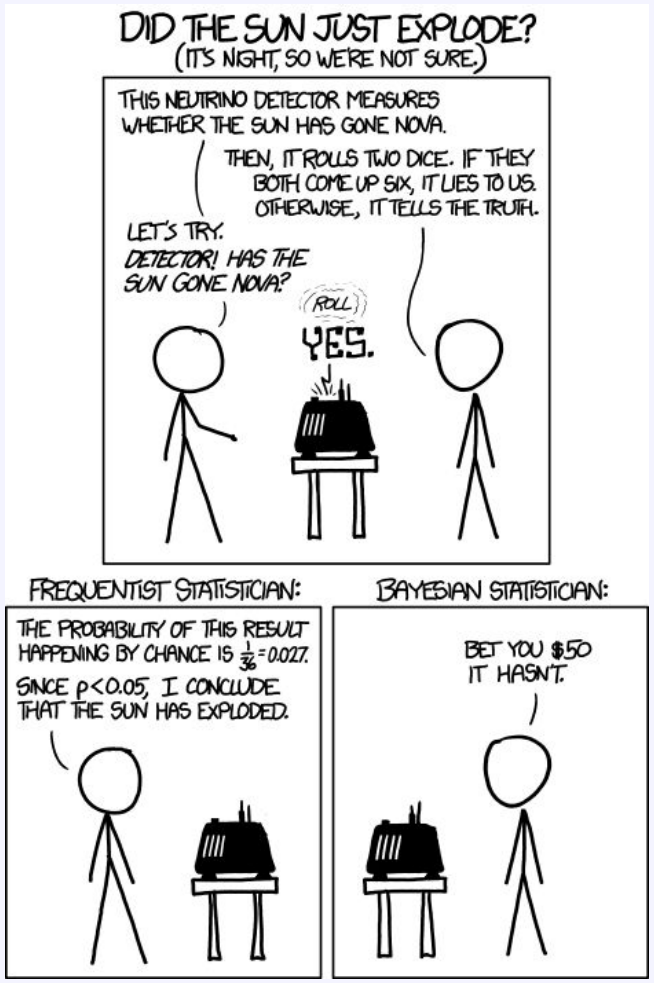
Observed Empirical Effects

- Fun stuff for you to google!
  - Bias Variance Trade-off
    - Latest opinion on this topic: [LeCun 2023]
  - Grokking
  - Loss landscapes
    - For that matter any paper that has "___ is all you need" or "AGI" in the title is best enjoyed with a tub of popcorn
  - Hyperparameter tuning
  - xAI
  - x-bit LLMs where 'x' reduces faster than *insert Elon joke here*

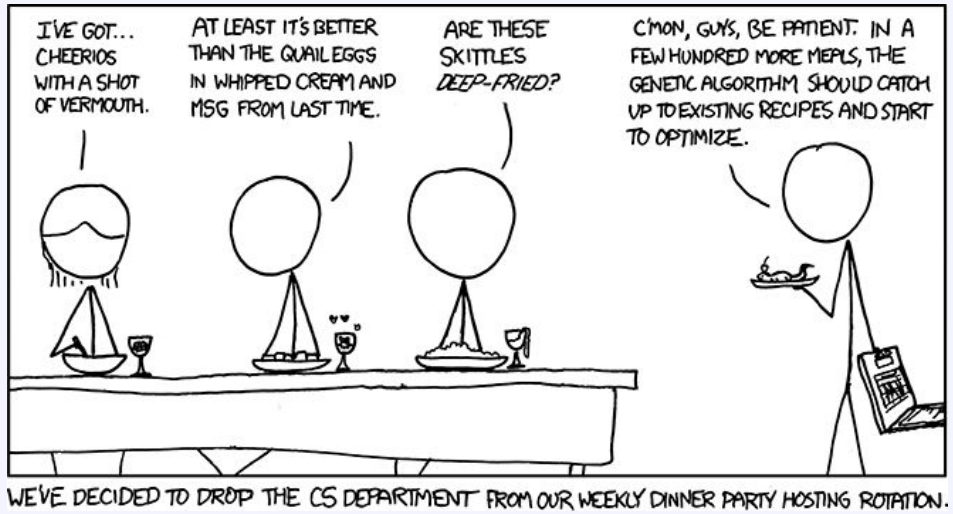
SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

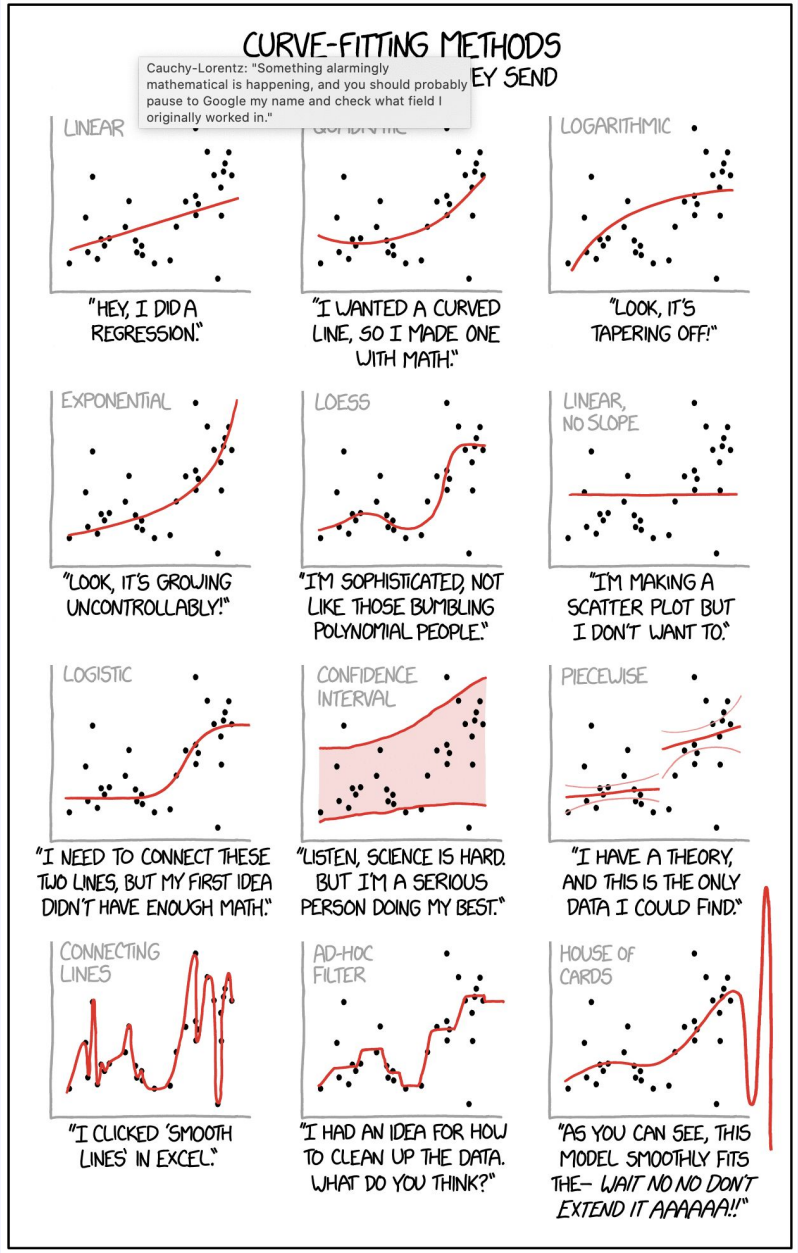
SMARTHEP

REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY



END





SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

MANCHESTER 1824
The University of Manchester

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY

Useful Resoruces:

- Michael Bronstein's Medium Blogs
- Geoffrey Hinton's Lectures
- This course - because at CERN we are obsessed with making ML models faster (courtesy of DHCP)
- Twitter (block Elon first for mental sanity) Academia
  - Get access to papers hot off the press
  - Watch top level academics beef with each other
    - Most times just googling terms from twitter disses from academics hurled at other academics teaches me more than well designed courses

SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086