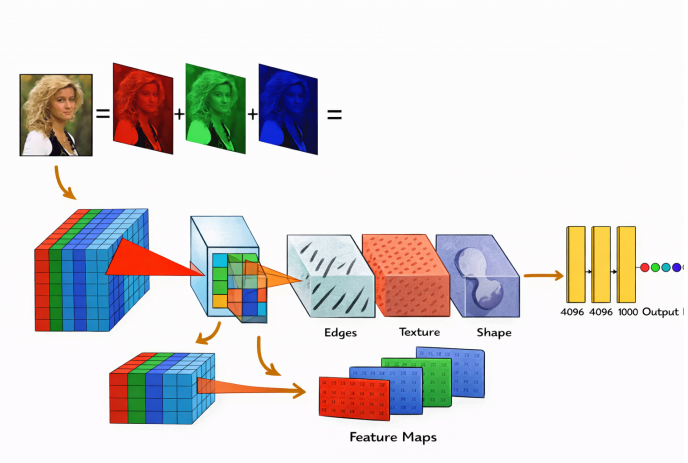
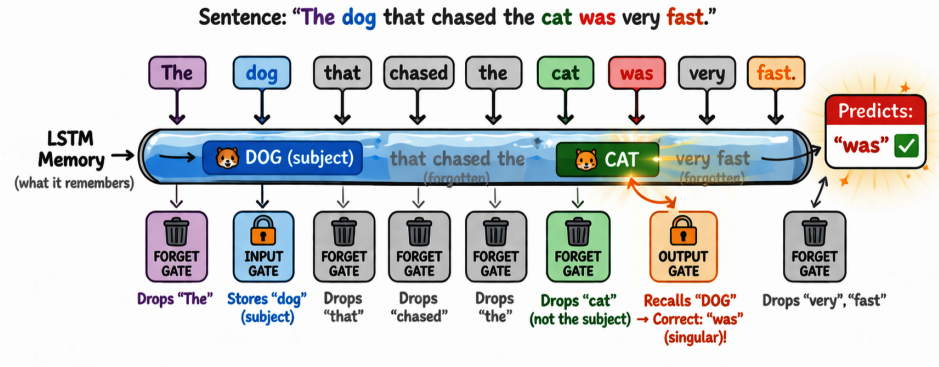
The Tutor-Pupil Augmentation: Enhancing Learning and Interpretability via Input Corrections

Darya Biparva, Maarten Schoukens, and Donatello Materassi
bipar001@umn.edu

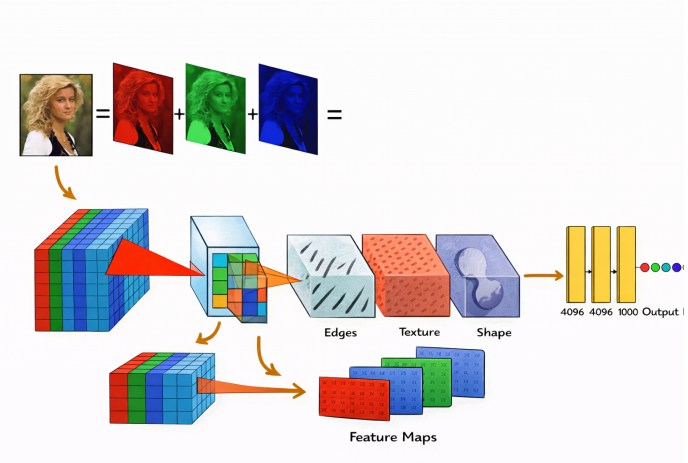
The Fourteenth International Conference on Learning Representations (ICLR2026)

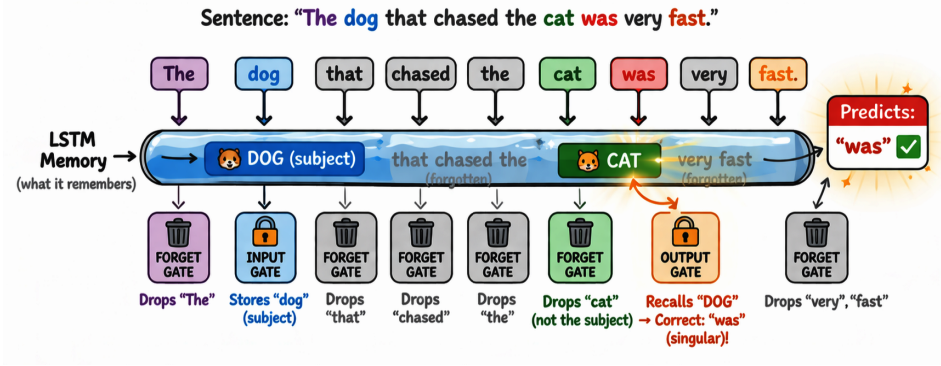
Incorporating prior knowledge into our models





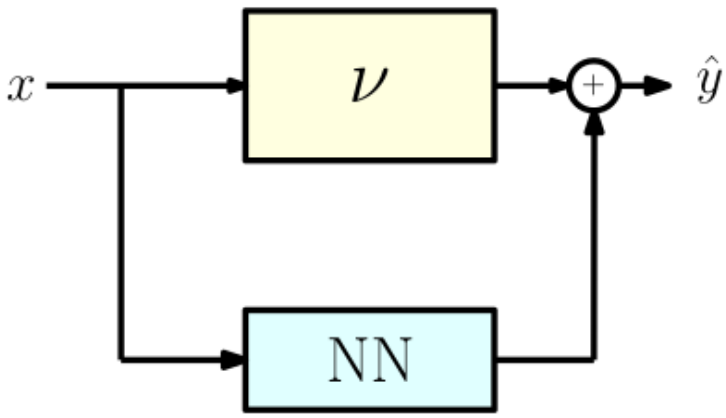
Incorporating prior knowledge into our models



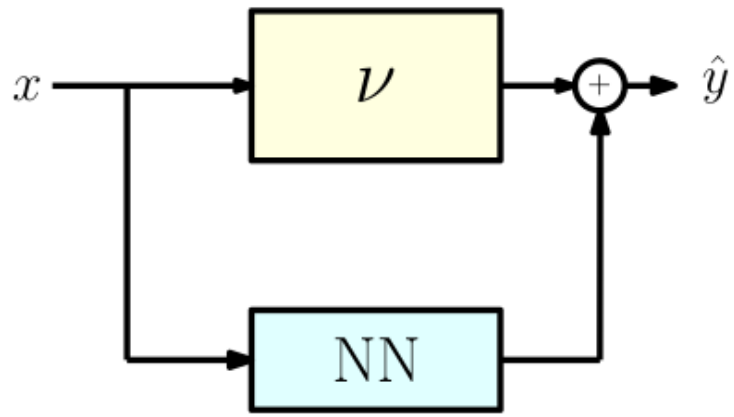


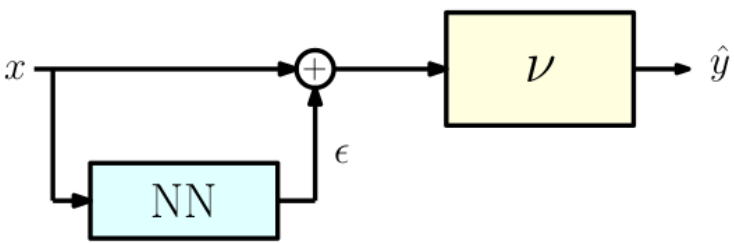
- What if our prior knowledge is a model?
  - First principles: $P = \frac{nRT}{V}$ } Physics-informed ML / Physics-guided ML
  - Previously tested simple model
  - Trusted model

Model Augmentation

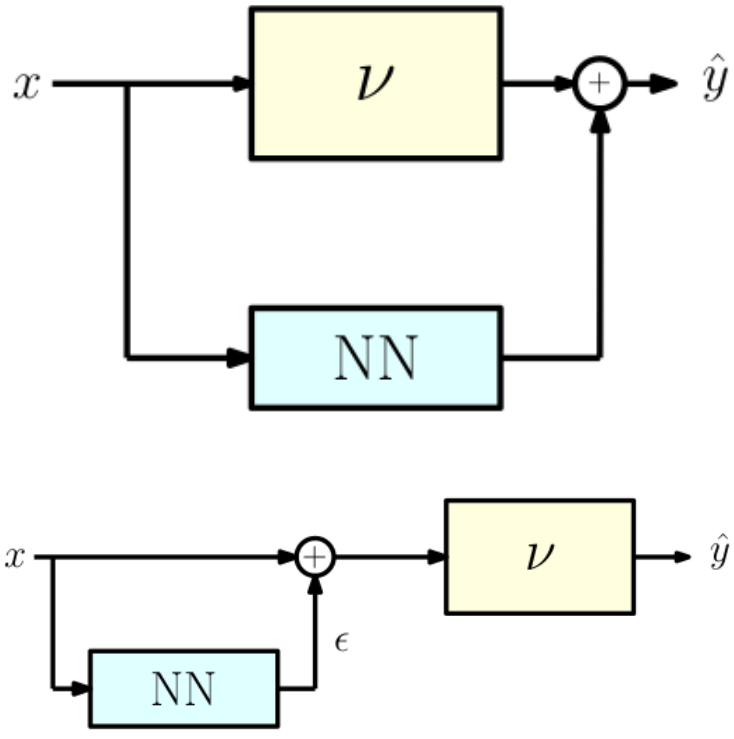


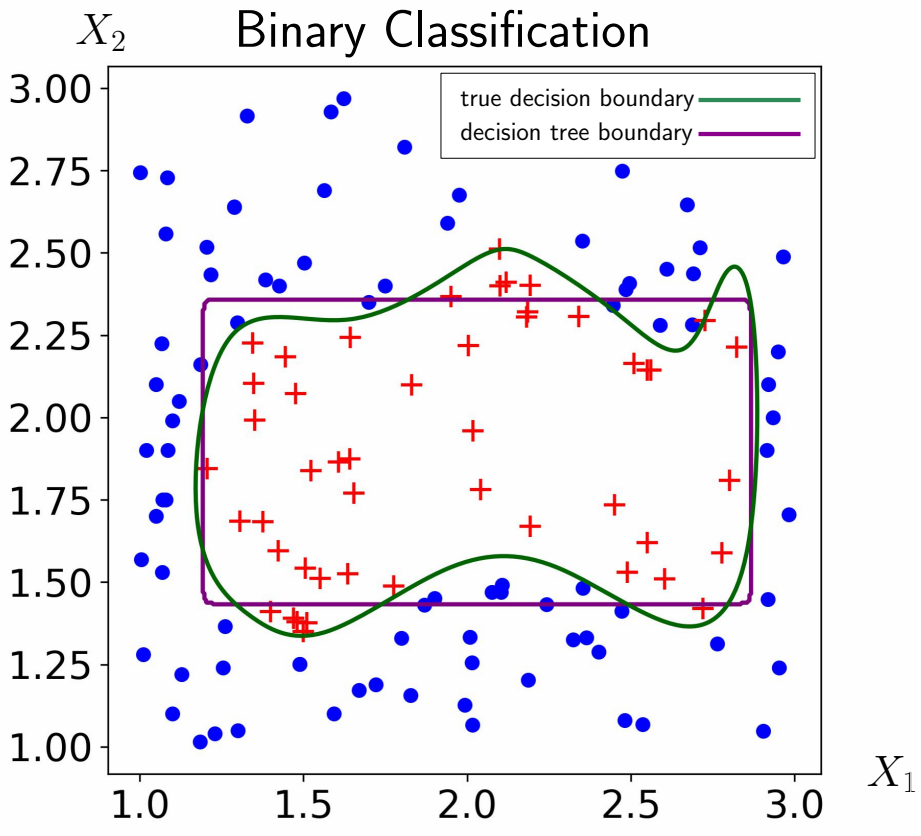
Model Augmentation





Model Augmentation

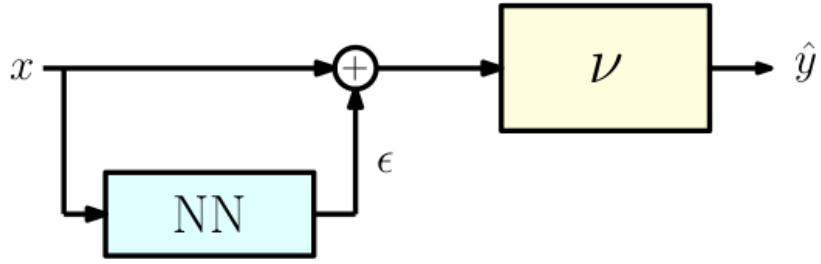


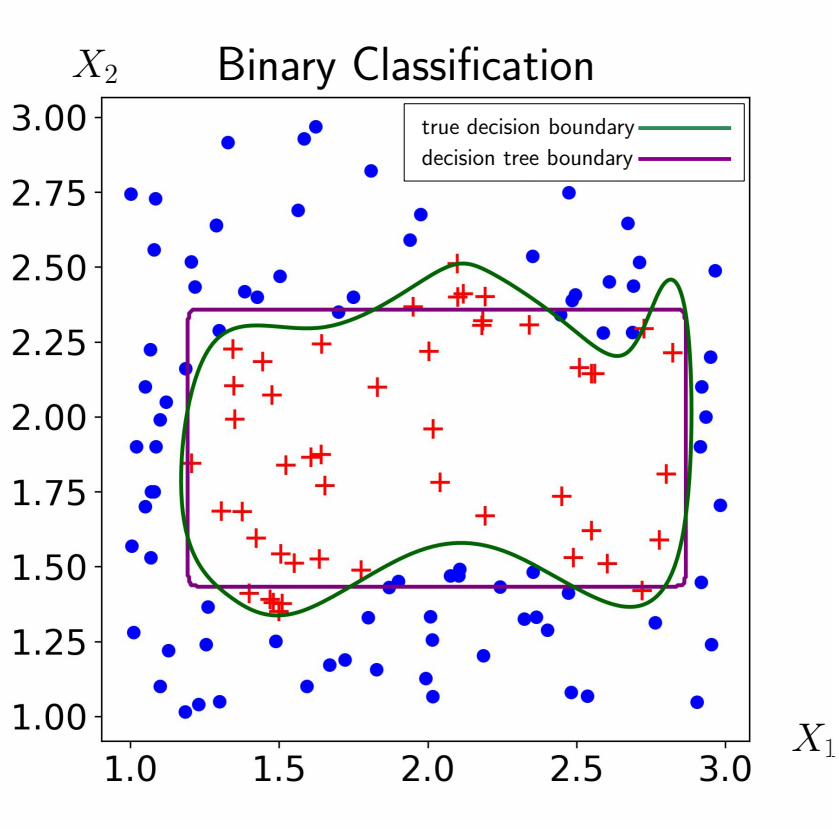


The polygon represents the interpretable function ν

How to incorporate the known model into a neural network?

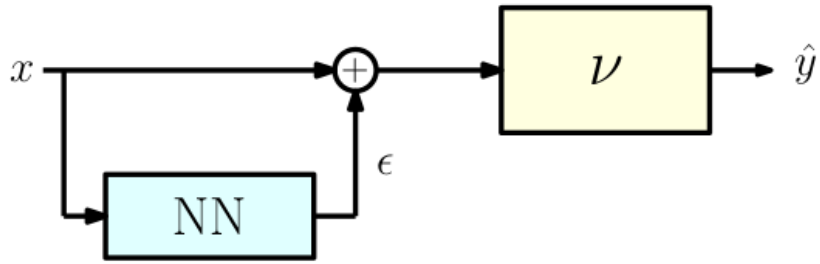
Tutor-Pupil Augmentation

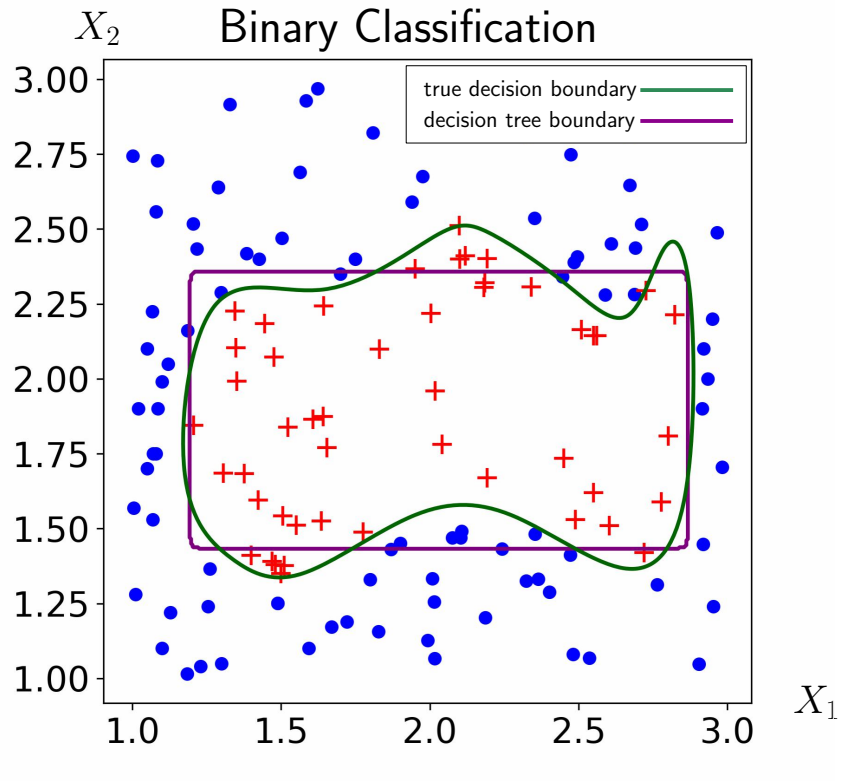




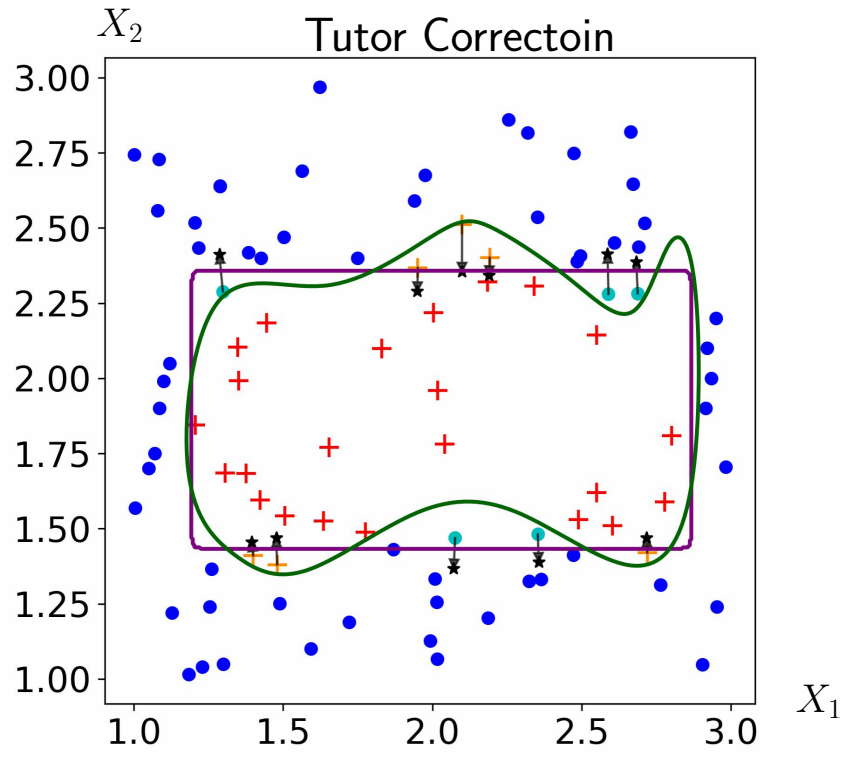
The polygon represents the interpretable function ν

Tutor-Pupil Augmentation





The polygon represents the interpretable function ν



Data represented in orange and light blue are data that the augmented model classified them differently than ν
The minimal change ϵ identified by the neural network is depicted with a black arrow, and the modified $\epsilon + x$ is depicted with a black star.

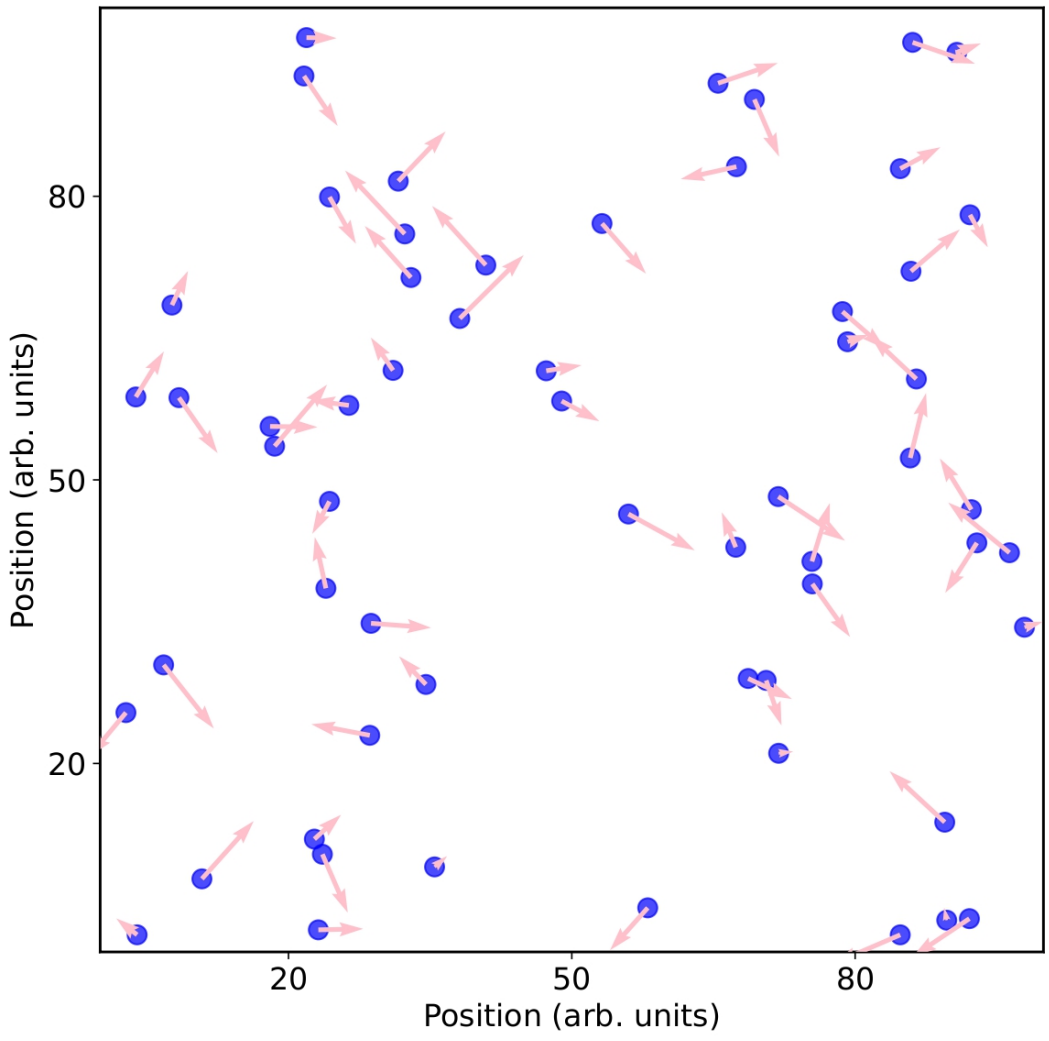
Physics based example

- Theoretically driven
- Assumptions:

  infinitesimal radius of particles
  elastic collisions

$$P = \frac{nRT}{V}$$

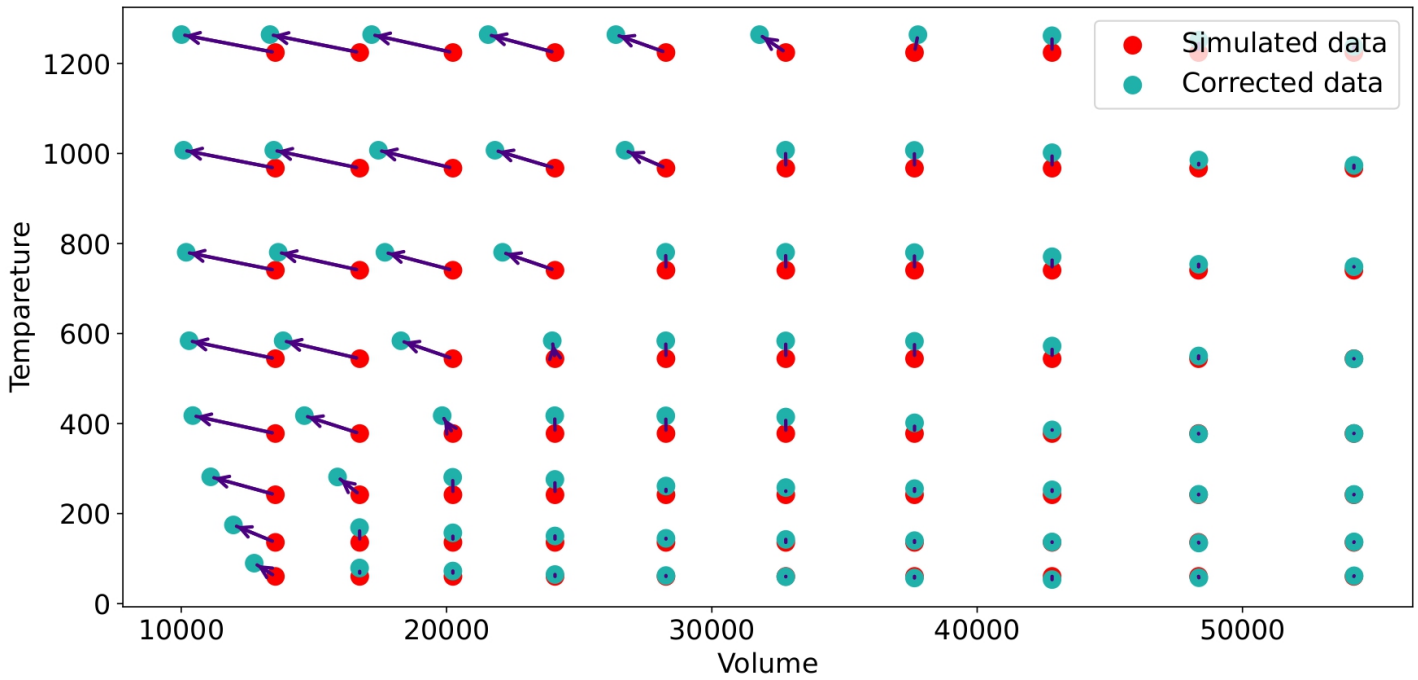
Physics based example



$$P = \frac{nRT}{V}$$

Physics based example

$$\hat{P} = \frac{nR(T+\epsilon_T)}{V+\epsilon_V}$$



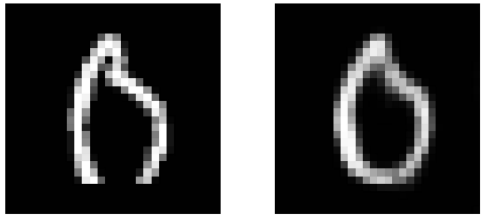
Van der Waals equation: $$P = \frac{nRT}{(V-b)} - \frac{a}{V^2}$$

Image classification

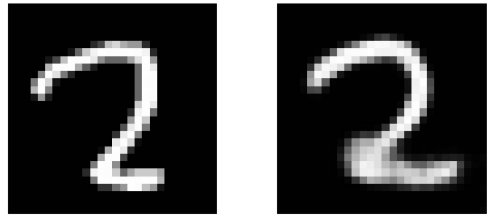
- Image classification task
- Train a logistic regression

Image classification

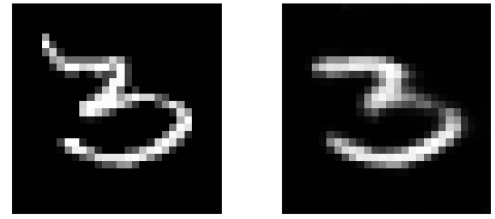
- 98.5% accuracy



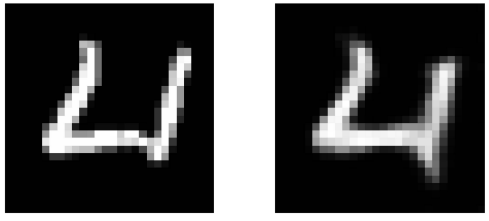
Pupil: 9, Tutor-Pupil: 0



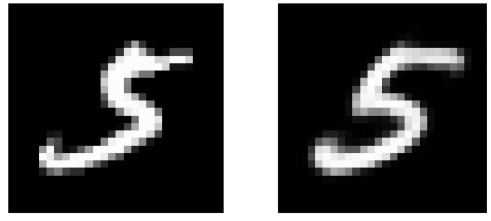
Pupil: 7, Tutor-Pupil: 2



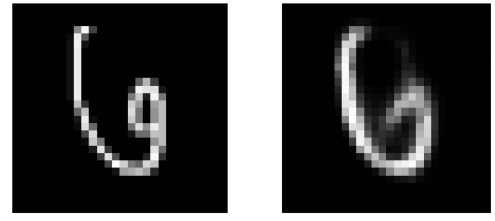
Pupil: 5, Tutor-Pupil: 3



Pupil: 6, Tutor-Pupil: 4

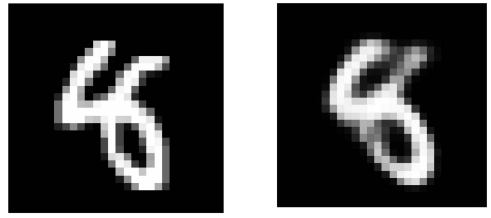


Pupil: 3, Tutor-Pupil: 5

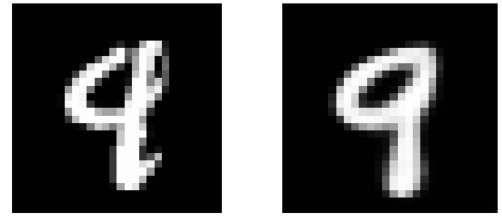


Pupil: 4, Tutor-Pupil: 6

Pupil: 9, Tutor-Pupil: 7



Pupil: 4, Tutor-Pupil: 8



Pupil: 4, Tutor-Pupil: 9