Applied Deep Learning

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

www.bayes.city.ac.uk

Learning objectives of today

Goals:

- Recognize the vast potential of using pre-trained neural networks for our (computer vision) tasks, instead of spending months developing architectures and training complex networks
- Debate the risks of bias inherent to machine learning in general and deep learning applications in particular

How will we do this?

- We start with a quick overview of transfer learning, and we will see how it can be applied to a previous problem we faced. We will move to actually applying transfer learning in the tutorial
- We will then see a concrete example of bias in facial detection tools
- We discuss how we can overcome bias and get to know a recent autoencoder-based development

Transfer learning

Transfer learning

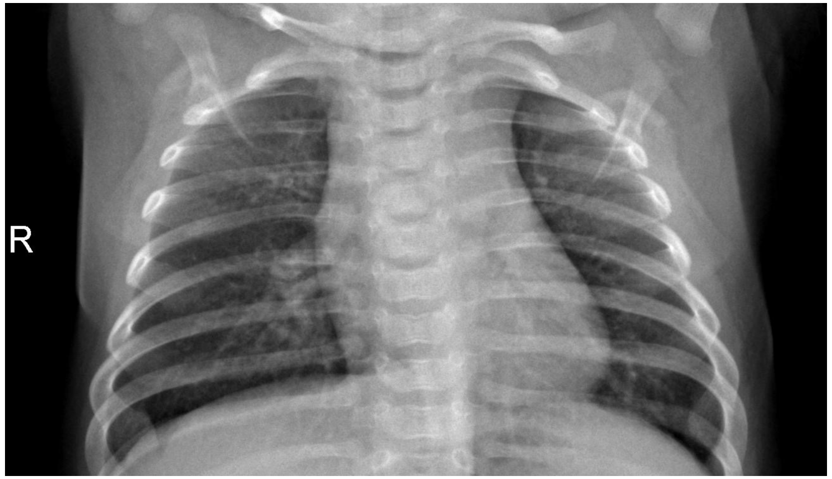
- Say you want to create a neural network to classify patients as having pneumonia or not based on a lung X-Ray

- Instead of developing a new neural network, maybe we could use an existing network. Say one that won the ImageNet classification challenge. After all, they can differentiate between lots of different objects

- Going through the list of objects in the challenge, you realize there are none capturing healthy lungs or lungs with pneumonia

- Do you think the pre-trained neural network you identified could still be useful for the task at hand?

Pneumonia or normal?

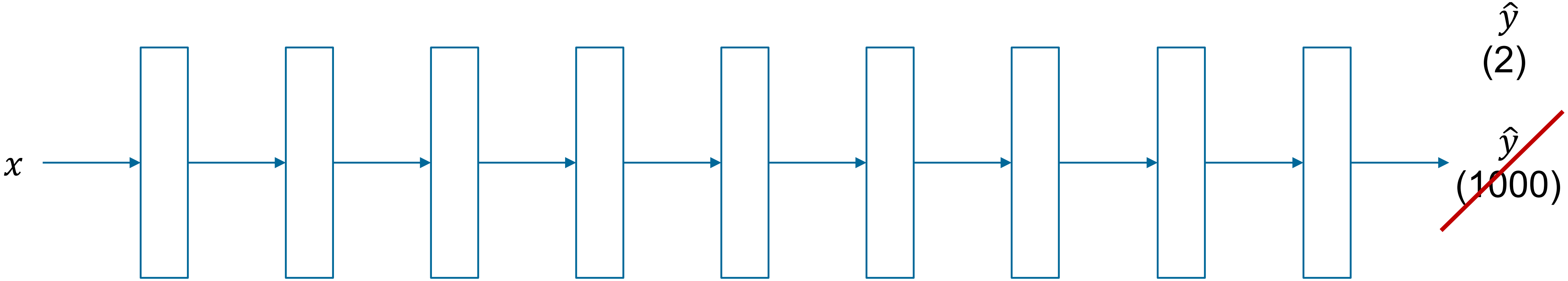


Repurposing a neural network

- Naïve approach: take the existing (trained) neural network

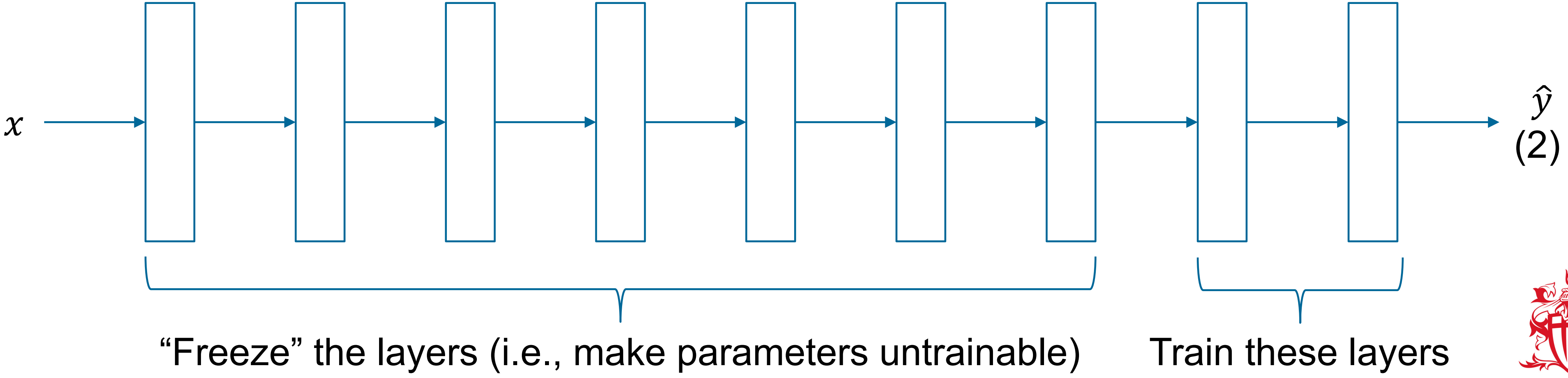
- Adjust the output layer

- Train some more with your data set



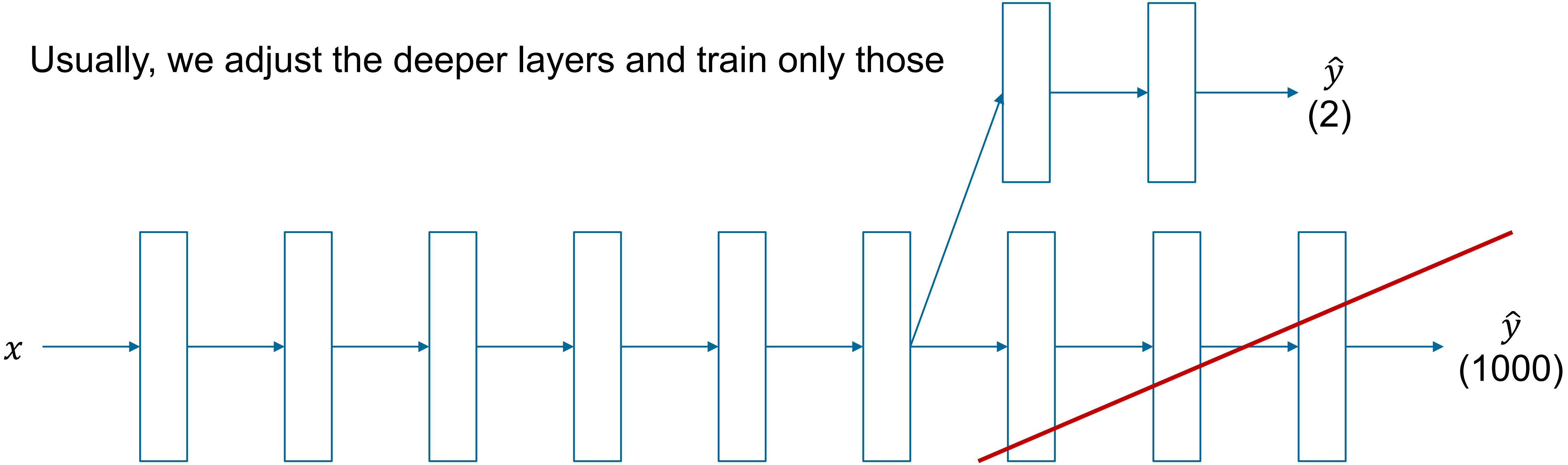
Difference between low-level and high-level features

- Problem with the previous approach: training may be very slow
- But: early layers capture low-level features that are unlikely to be different
- Deeper layers capture high-level features that are likely to be different



Difference between low-level and high-level features

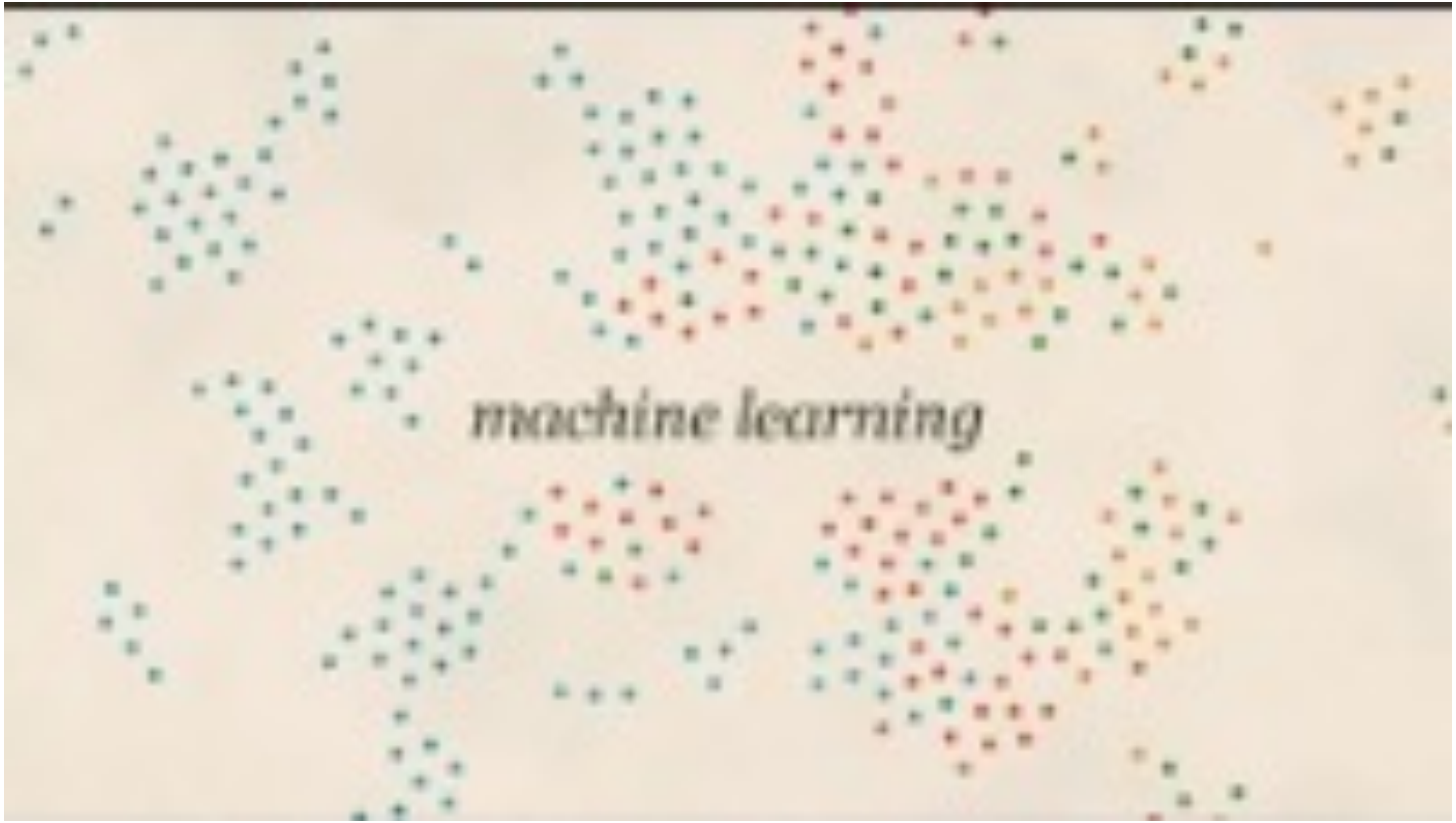
- We can go further, by adjusting some of the layers to fit better with our context
- Usually, we adjust the deeper layers and train only those



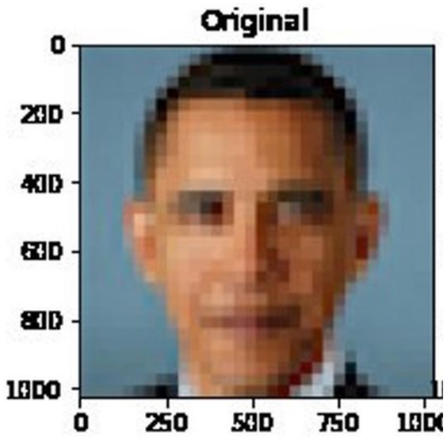
Let’s see how transfer learning can help us with our lung classification problem

Human bias in machine learning

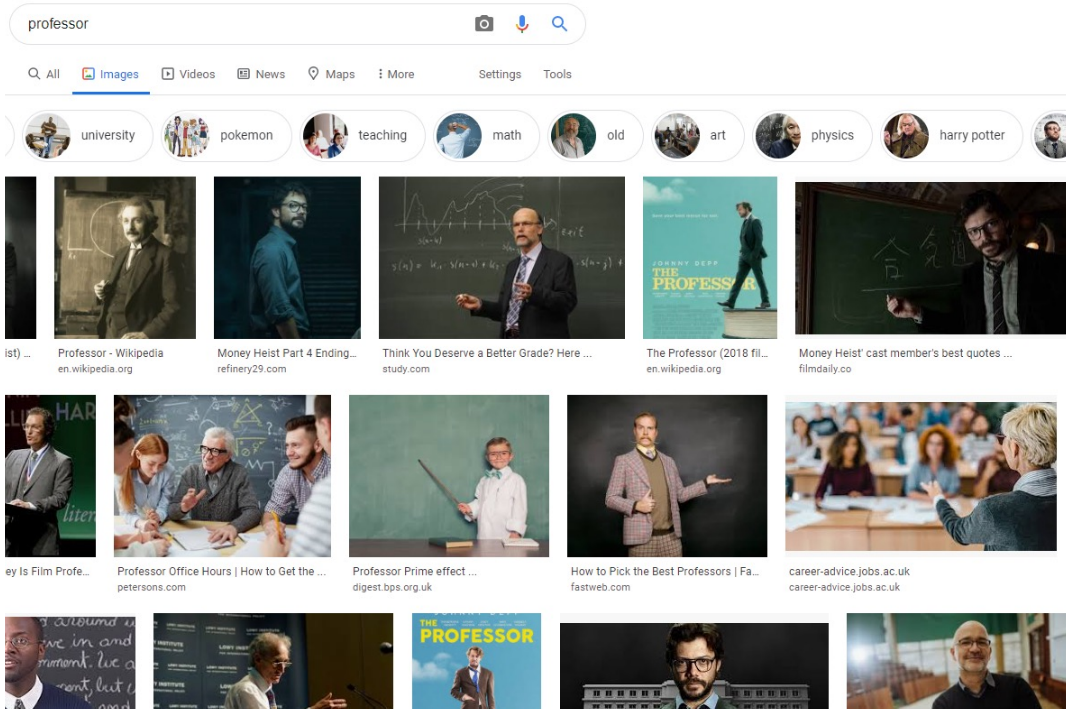
Human bias in machine learning



Other examples of human bias in machine learning



Source: theverge



COMPAS Recidivism Racial Bias

Racial Bias in inmate COMPAS reoffense risk scores for Florida (ProPublica)

Algorithm to score probability of recidivism used by judges and parole officers. Predicted blacks would reoffend more often than they actually did.

Issues:

- **Sample bias:** data used for training contains preponderance of one category of person
- **Prejudice bias:** the prejudices encoded into society are reflected in the data

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Your turn

- In the notebook “*ADL_Week 8_Facial detection and bias.ipynb*” we build a classifier to detect whether an image contains the face of a person or not.

- To train the model, we use roughly 50k images from celebrity faces and 50k images from other objects

- We then test our model on a new set of faces and see how well it does based on skin tone and gender

- Your task:
  - Go through parts 1 and 2 of the notebook (note that model training may take 3-5 minutes)
  - With your classmates, discuss the questions at the end of part 2

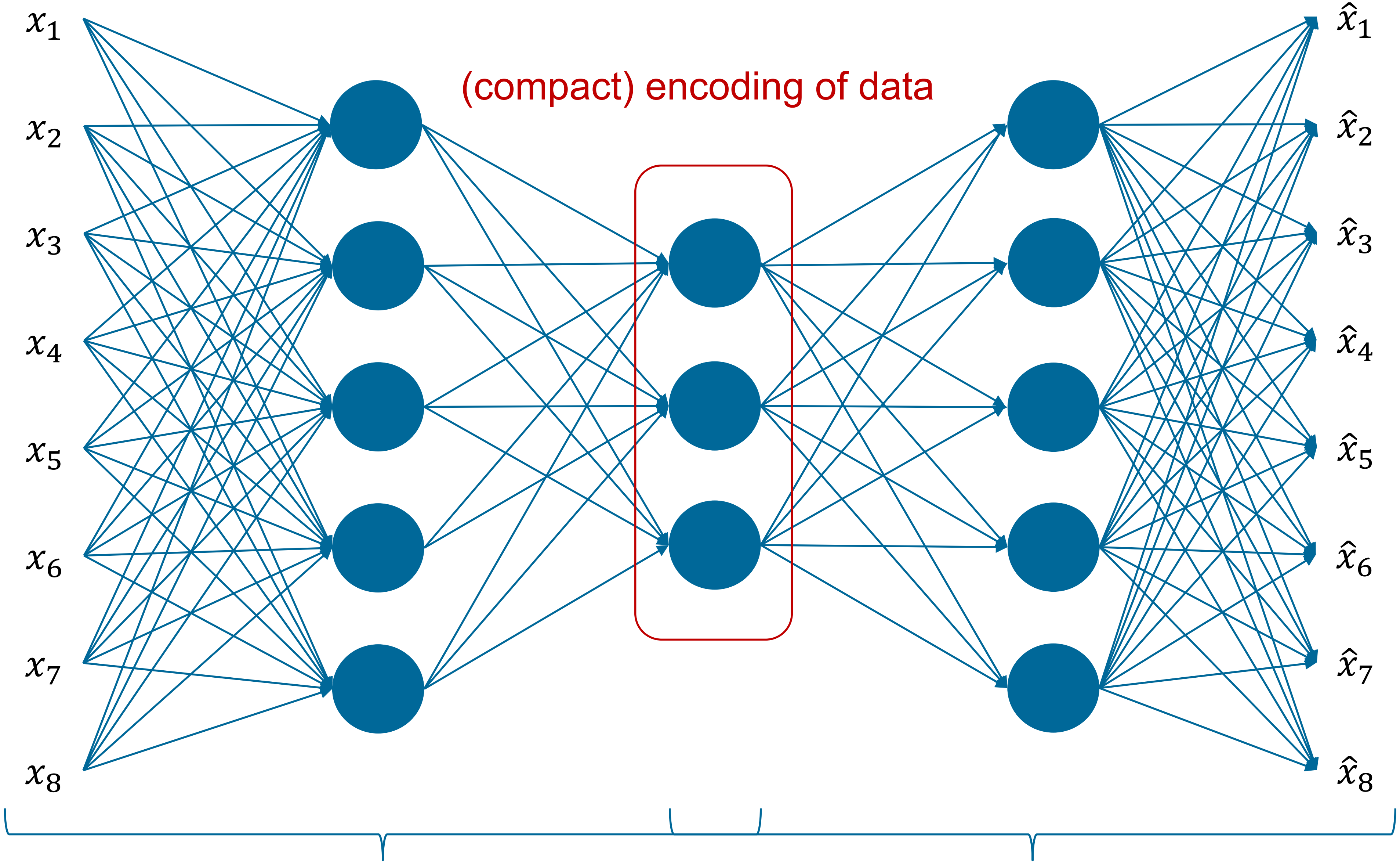
BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Let’s discuss your observations

Go to pollev.com/adl2022

Debiasing with a variational autoencoder

Recall an autoencoder

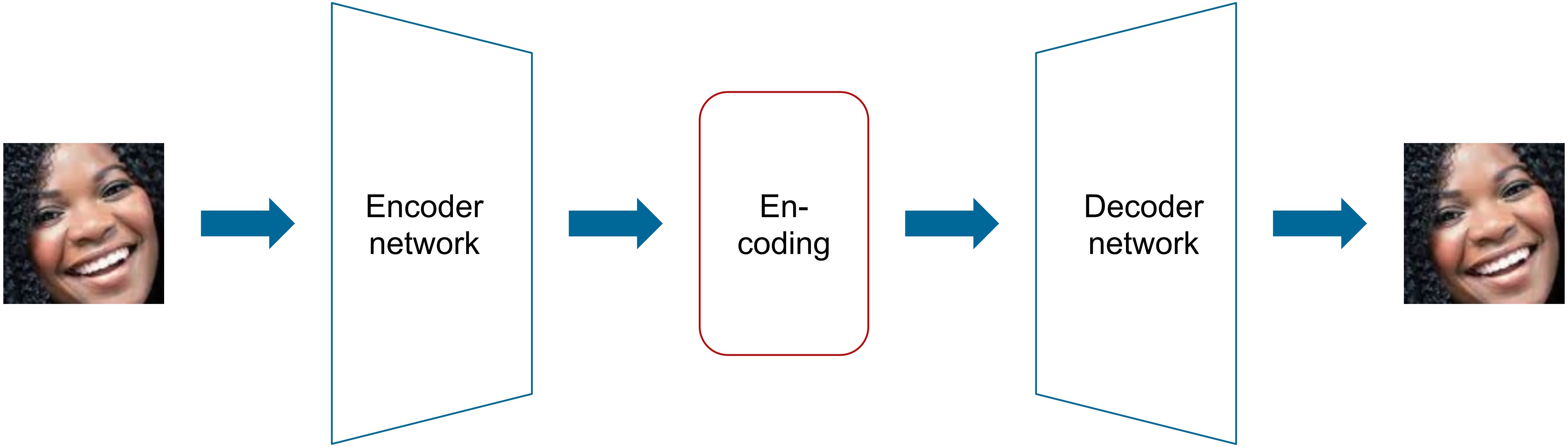


"Encoder": find a representation $f(x)$

"Decoder": unpack $x = g\big(f(x)\big)$

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

More abstractly

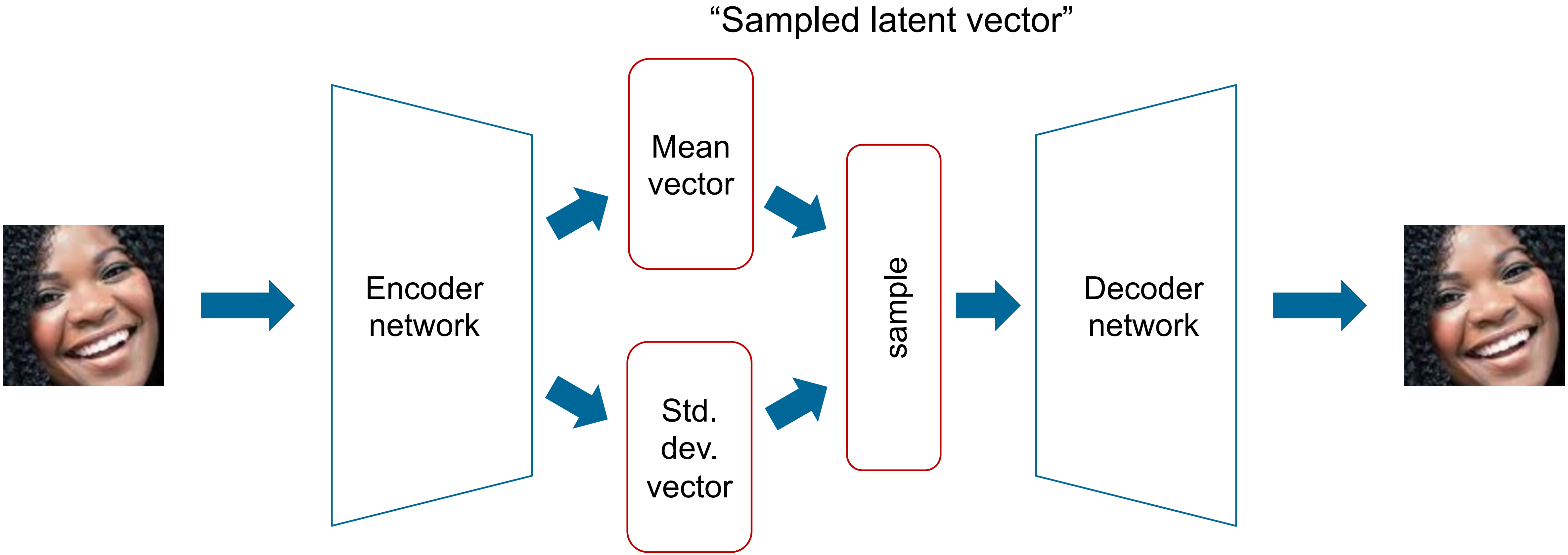


Encoder network

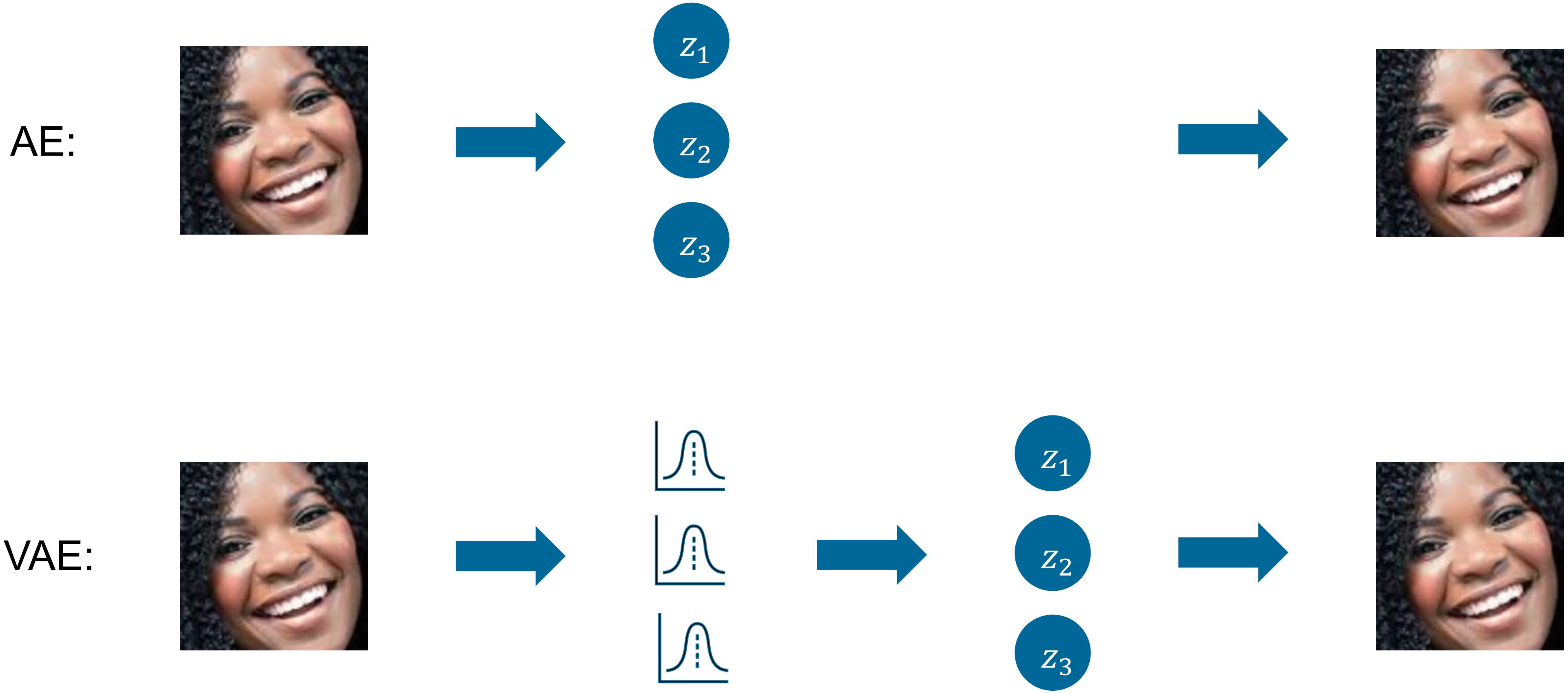
Encoding

Decoder network

A variational autoencoder



The difference between an autoencoder (AE) and a variational autoencoder (VAE)

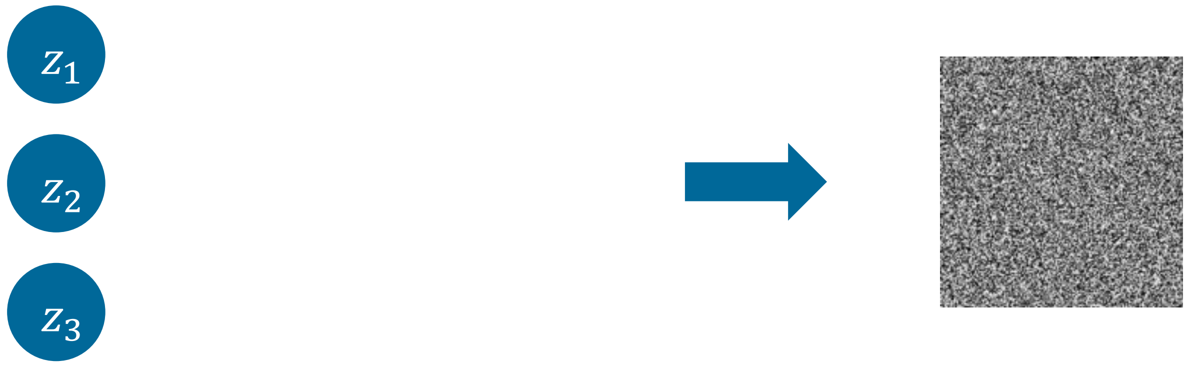


AE:

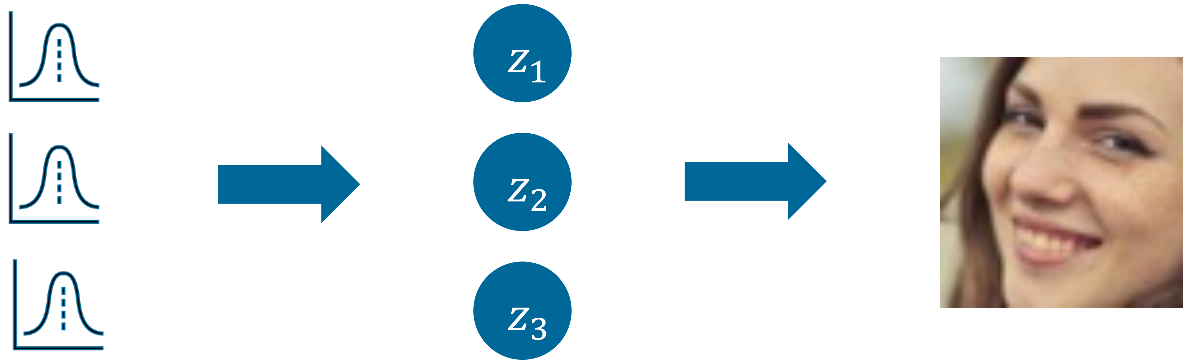
VAE:

Creating new content from a VAE

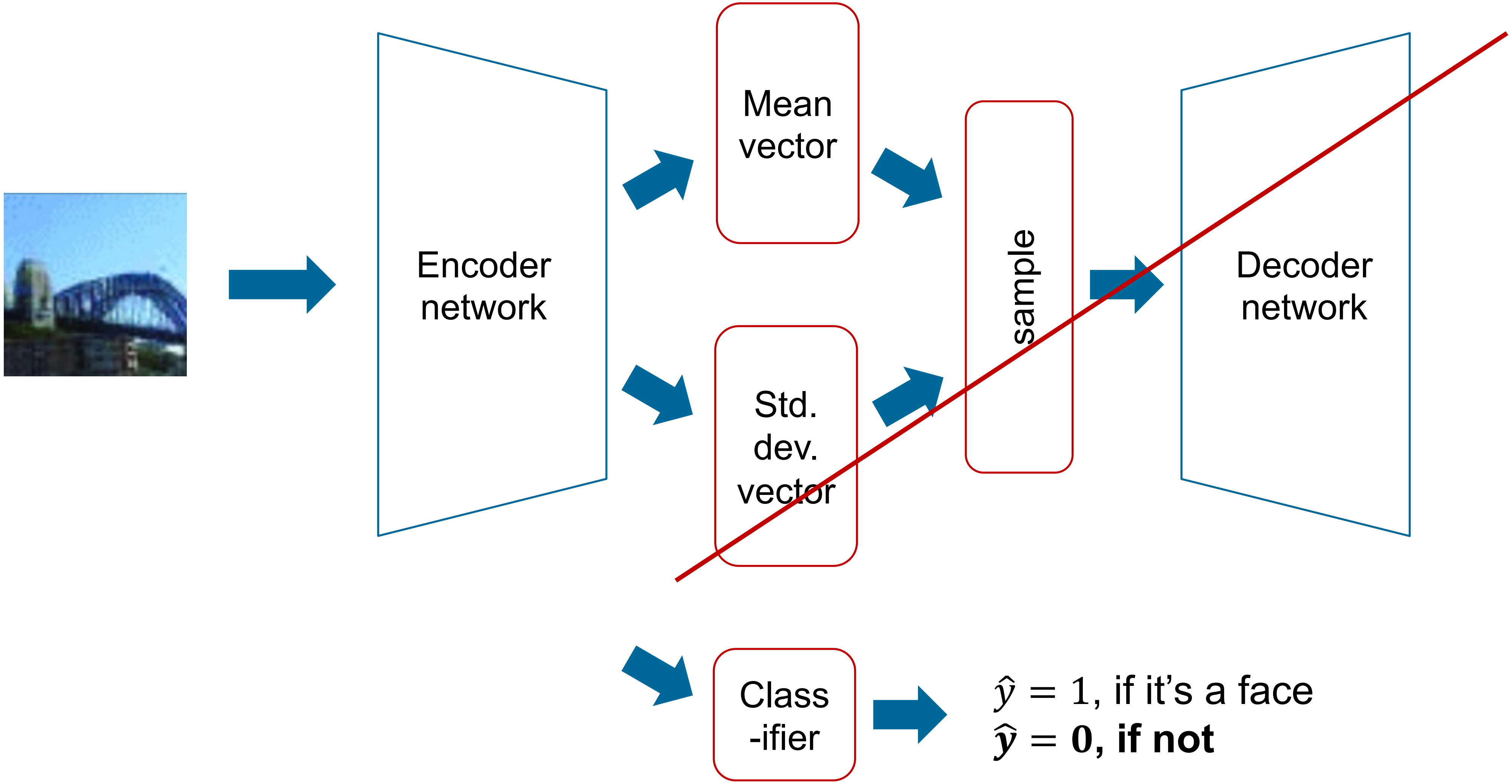
AE:



VAE:



Debiasing variational autoencoder (DB-VAE) by Amini et al.



Encoder network

Mean vector

Std. dev. vector

sample

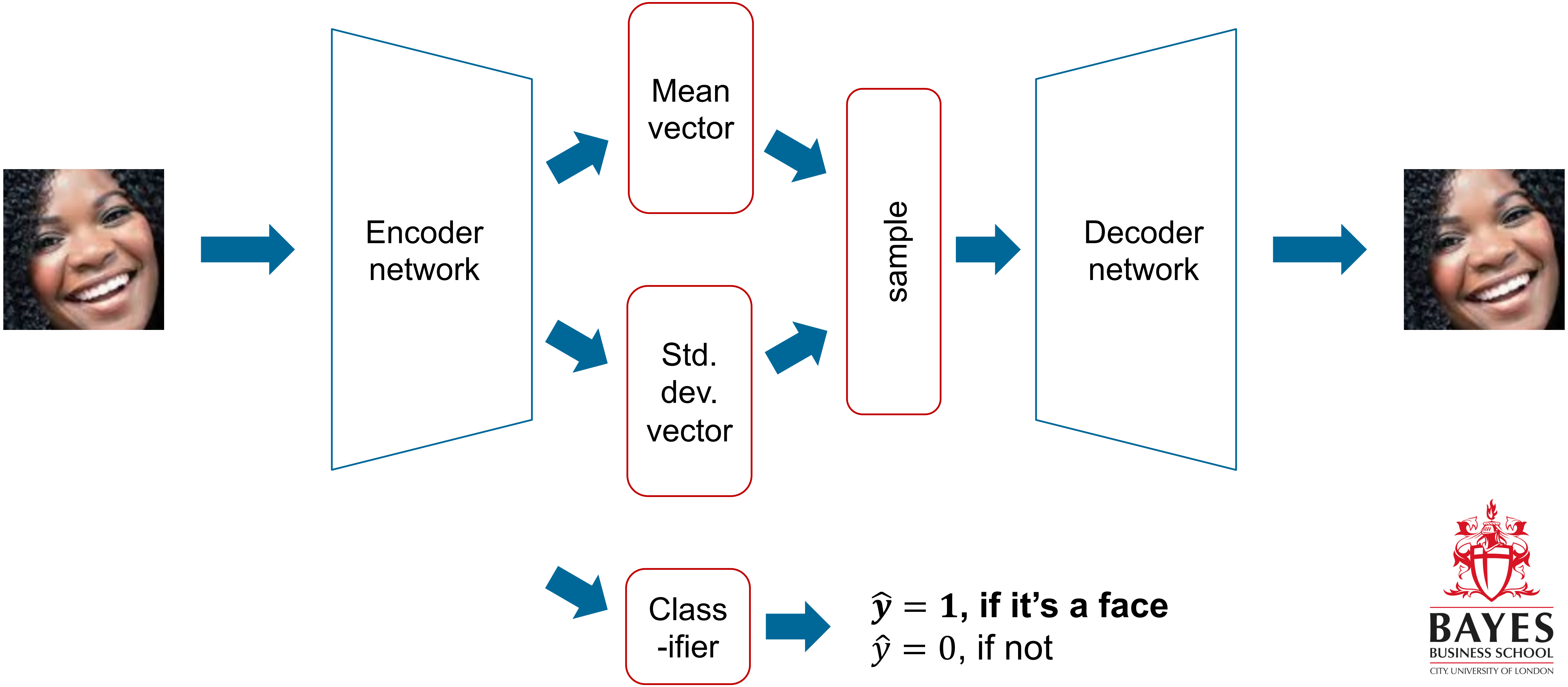
Decoder network

Class -ifier

$\hat{y} = 1$, if it’s a face

$\hat{\boldsymbol{y}} = \mathbf{0}$, **if not**

Debiasing variational autoencoder (DB-VAE) by Amini et al.



Encoder network

Mean vector

Std. dev. vector

sample

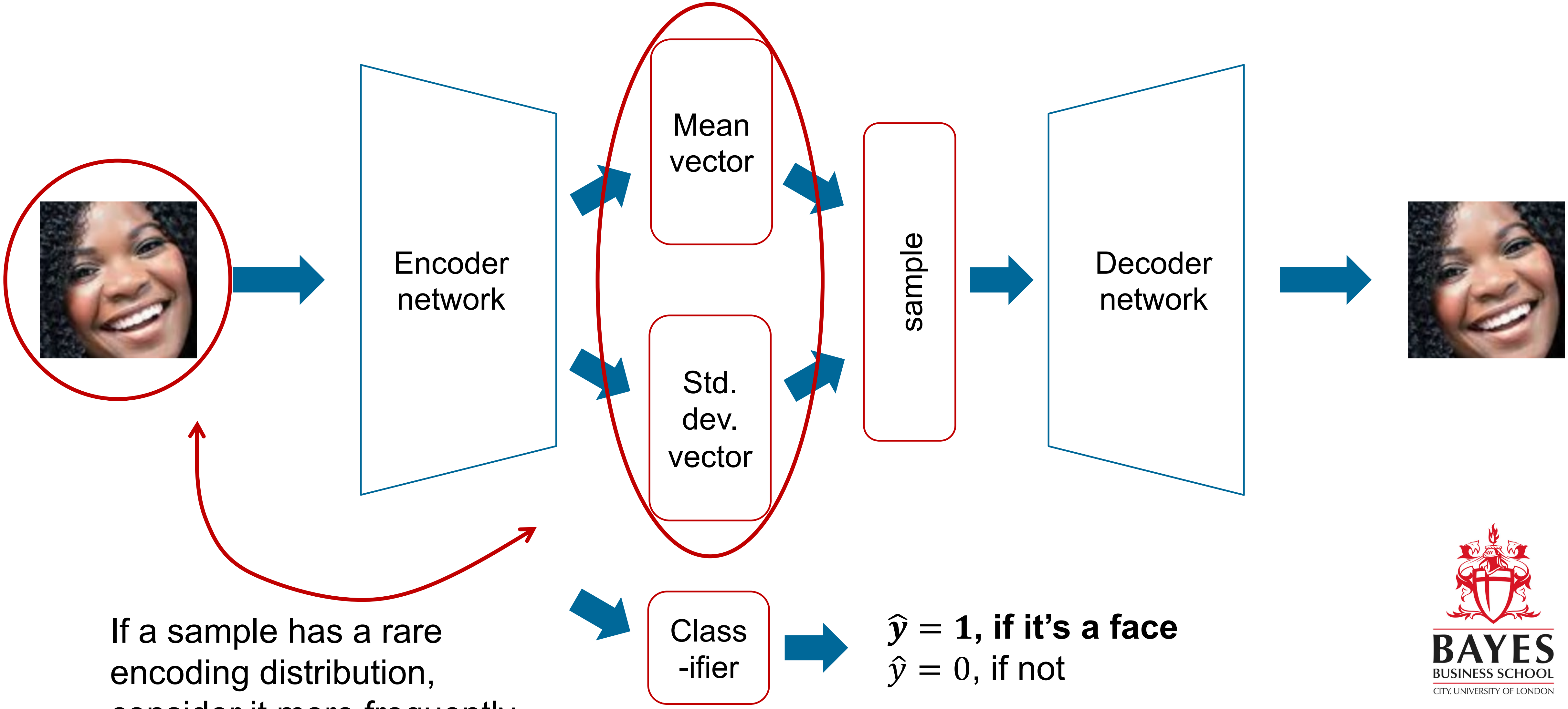
Decoder network

Class -ifier

$\hat{y}$ **= 1, if it's a face**

$\hat{y} = 0$, if not

Debiasing variational autoencoder (DB-VAE) by Amini et al.



$\hat{y} =$ **1, if it's a face**
$\hat{y} = 0$, if not

If a sample has a rare encoding distribution, consider it more frequently

Your turn

- Take another look at “*ADL_Week 8_Facial detection and bias.ipynb*”

- We build a variational autoencoder for debiasing. The details are a bit more complex, so don’t worry about it (the appendix has a lot of extra information)

- We then test the model again on the new set of faces, to see how we are doing this time regarding bias

- Your task:
  - Go through part 3 (training takes a bit longer, so the key outputs are already there for you to look at). Your focus should be on the final evaluation
  - With your classmates, discuss the questions at the end of part 3

Debrief

- How do accuracy and bias of the approach compare to the (non-debiased) baseline? Is the result surprising?

- In which applications (facial detection or not) is debiasing important? Are there some where you don’t want to debias?

Debrief

- Do you have other ideas for ways to address issues of human bias in models?

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Debrief

- Should it be necessary that companies demonstrate that models are not biased?

See you next week!

Sources

- Amini et al., 2019, Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure: http://introtodeeplearning.com/AAAI_MitigatingAlgorithmicBias.pdf
- Géron, 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow
- Goodfellow, Bengio, Courville, 2016, The Deep Learning Book: http://www.deeplearningbook.org
- Google, 2017, Machine Learning and Human Bias: https://www.youtube.com/watch?v=59bMh59JQDo
- Manyika et al., 2019, What Do We Do About the Biases in AI? https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai