Applied Deep Learning

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

www.bayes.city.ac.uk

Learning objectives of today

Goals:

- See how convolutional layers provide natural ways for understanding neural network learning
- Debate the risks of bias inherent to machine learning applications and understand how to detect and avoid bias using deep learning
- 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

How will we do this?

- We start by discussing what CNNs actually learn – and how we can use their activation to make algorithms more transparent
- We will then discuss biases introduced through our data and see how we can identify (and possibly overcome) these biases
- We end with a look at transfer learning, which enables complex applications with only limited training

The importance of transparency and interpretability

Trust

Fairness

Regulation

Informativeness

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

CNNs: not actually a black box

Visual interpretation of the learning of a CNN

- Different from neural networks more generally, CNNs are highly interpretable. Why?
  - We are representing visual concepts → this is easier to visualize by design

- Multiple approaches
  - Visualizing intermediate outputs
  - Visualizing convolutional filters
  - Visualizing heatmaps of class activation

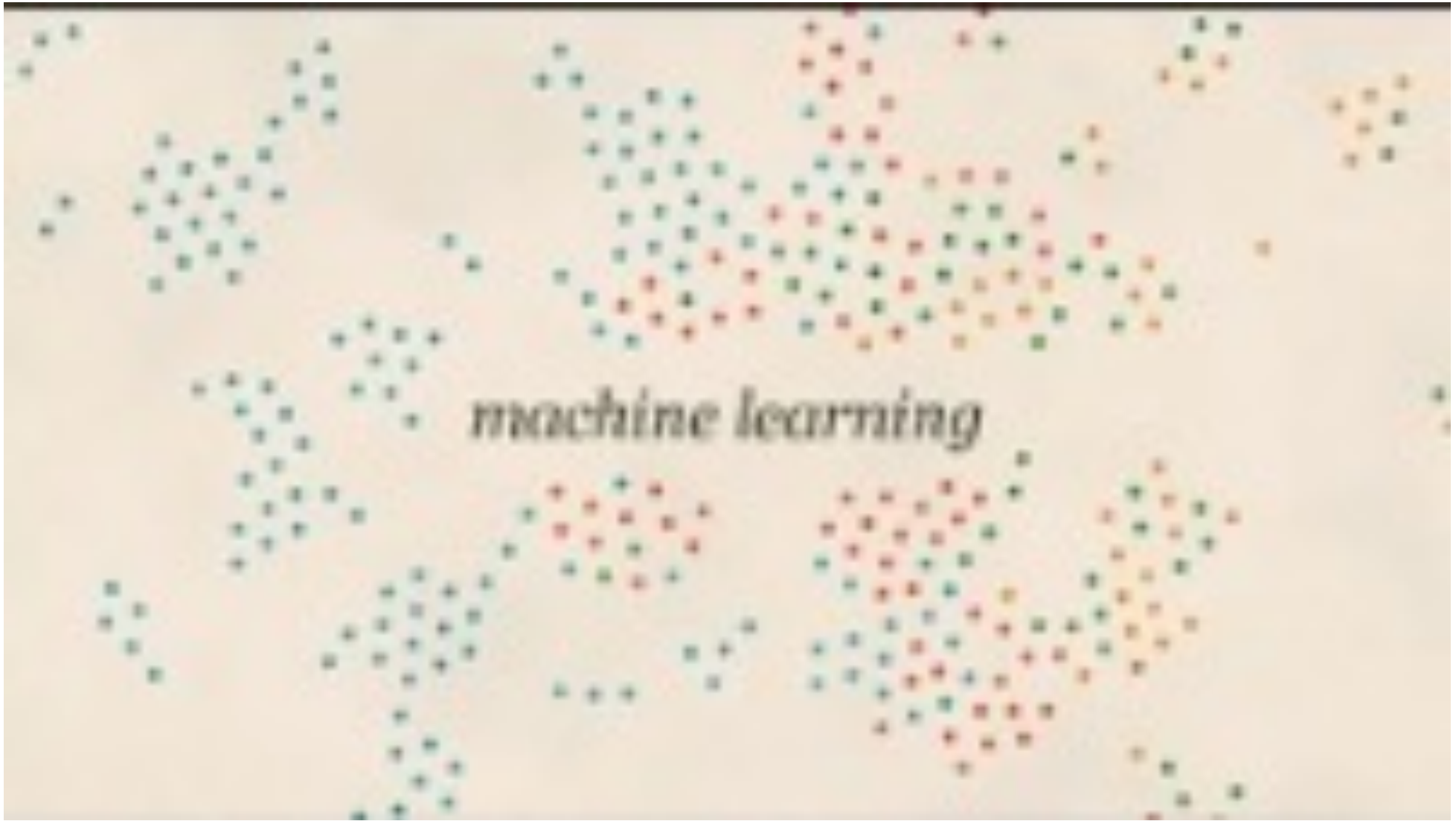
BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Different approaches and their use

- Visualizing intermediate outputs
  - Understand how layers transform their input
  - Understand meaning of individual filters
- Visualizing convolutional filters
  - Understand visual patterns and concepts filters are receptive to
- Visualizing heatmaps of class activation
  - Understand which parts of an image are used to classify
  - Direct application: Localization
  - Direct application: Checking for biases

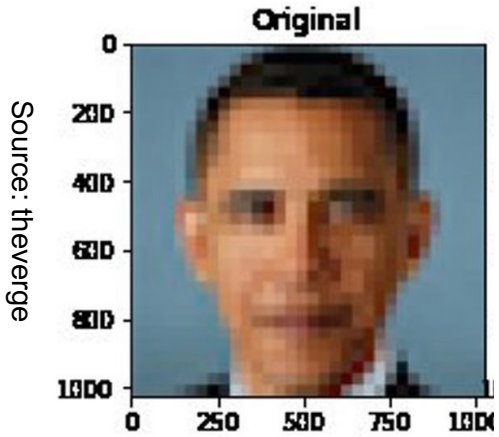
Human bias in machine learning

Human bias in machine learning

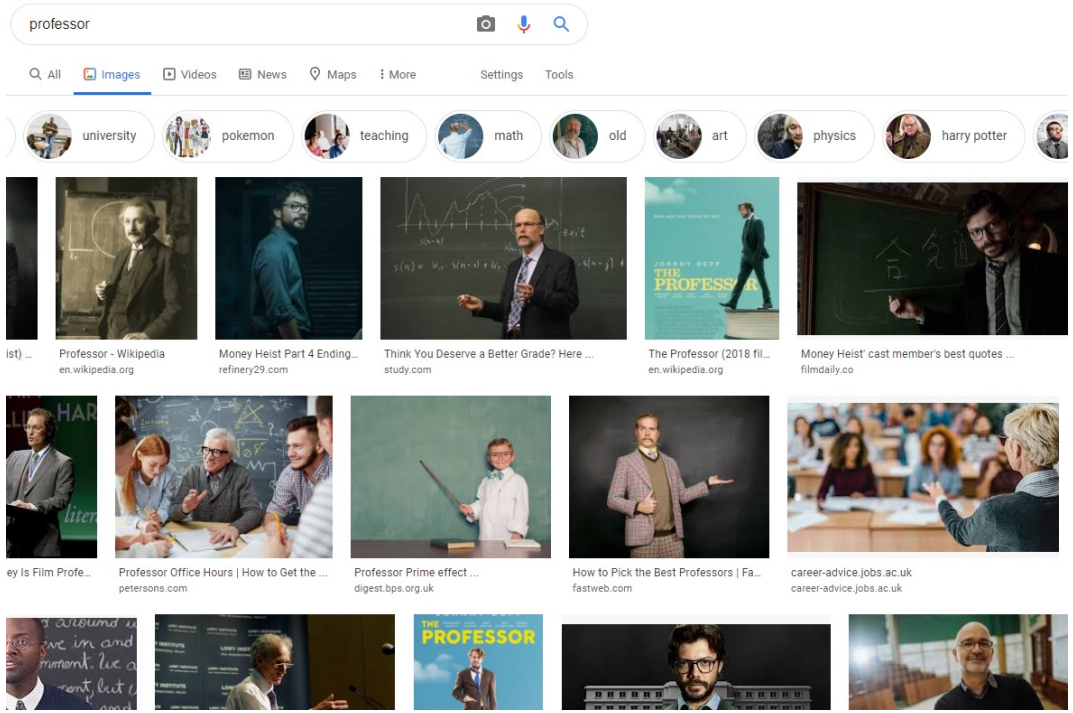


BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

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 African-Americans 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

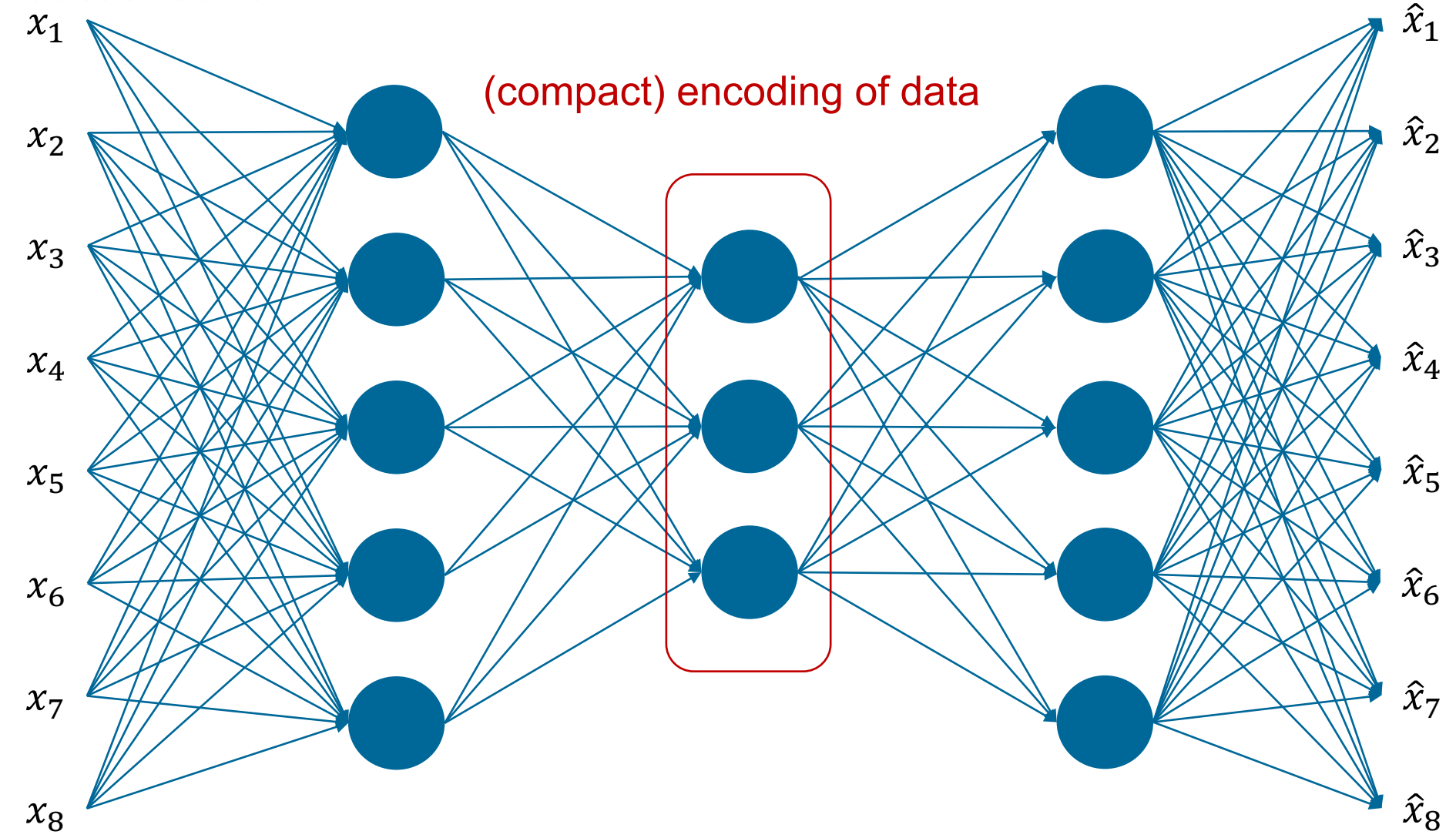
Why should we care about bias in our algorithms (as a business)?

Bias in ML – an example

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Detecting biases with a variational autoencoder

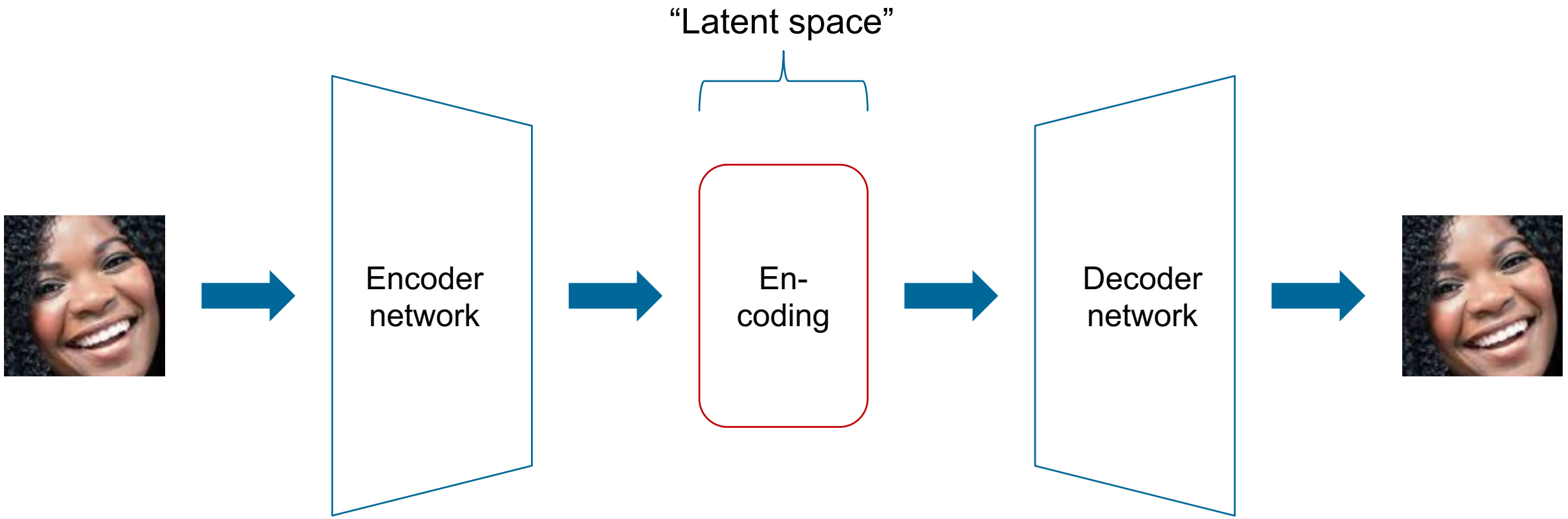
Recall an autoencoder



"Encoder": find a representation $f(x)$ "Decoder": unpack $x = g\big(f(x)\big)$

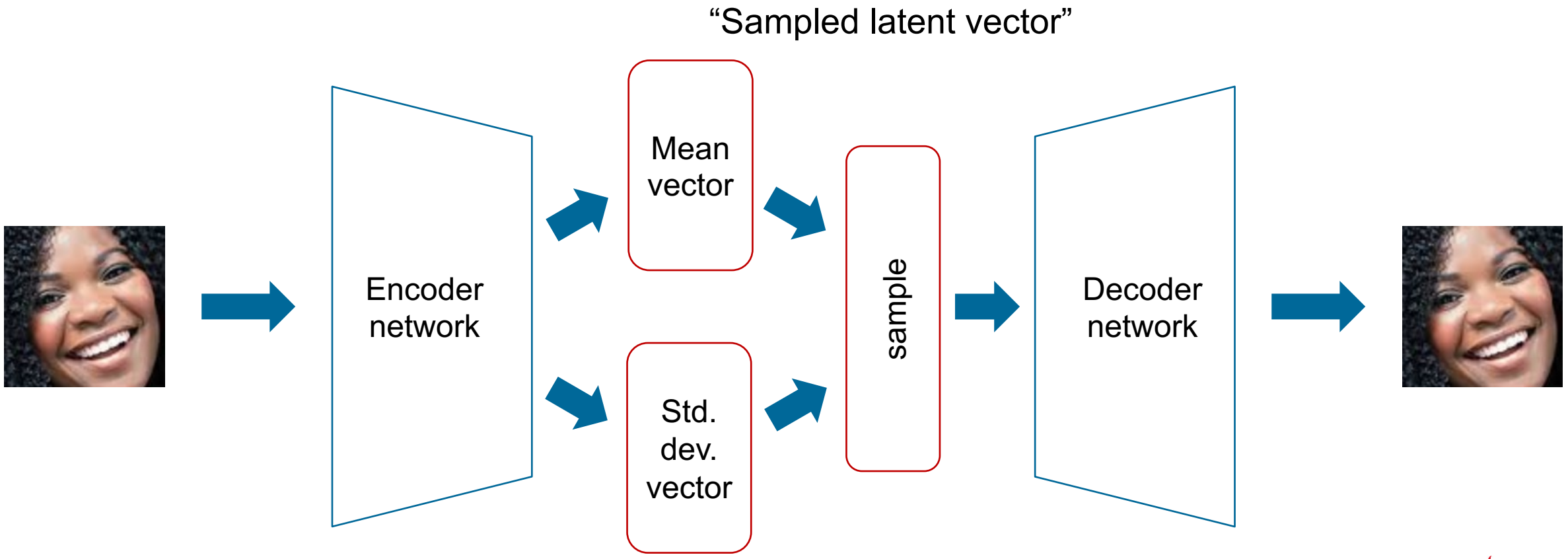
BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

More abstractly



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

A variational autoencoder

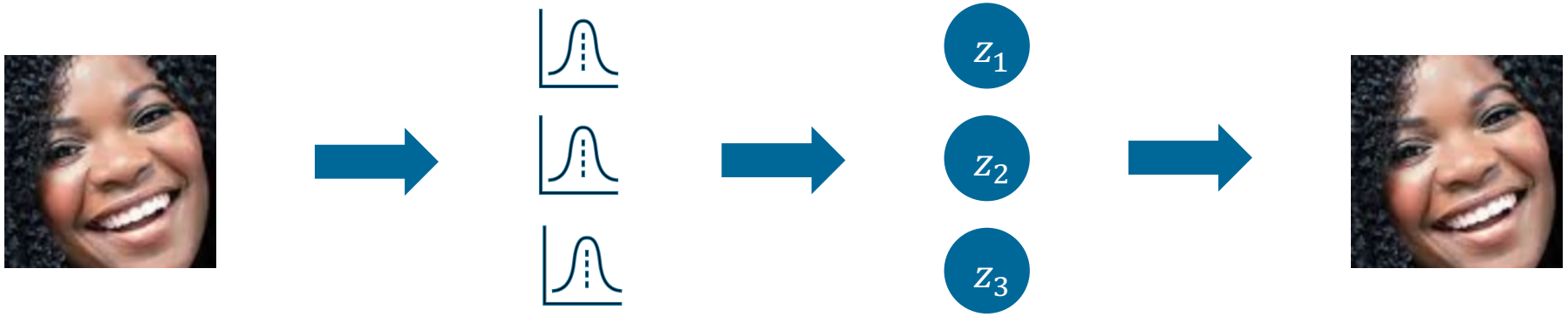


BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

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

AE:

VAE:

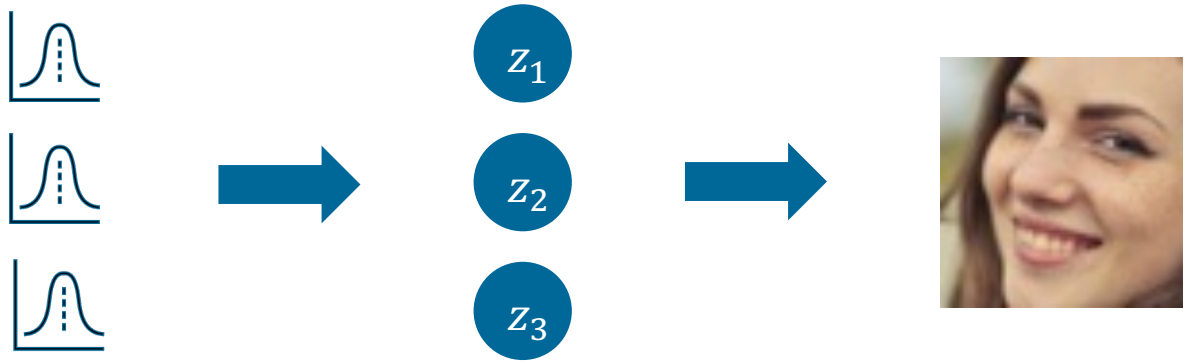


BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Creating new content from a VAE

AE:

VAE:



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Semi-supervised variational autoencoder (SS-VAE)

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

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Semi-supervised variational autoencoder (SS-VAE)

Encoder network

Mean vector

Std. dev. vector

sample

Decoder network

Class -ifier

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

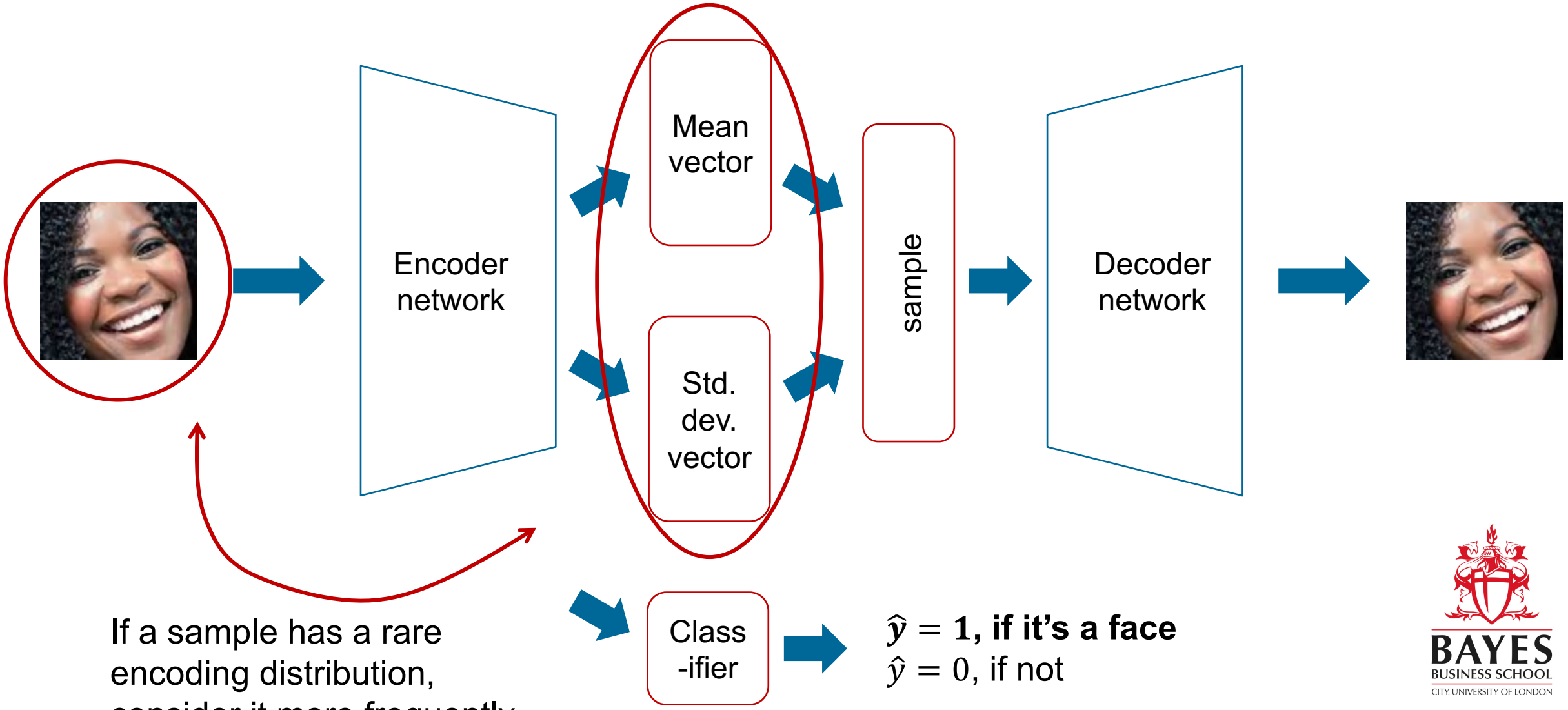
$\hat{\boldsymbol{y}} = \mathbf{0}$**, if not**

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Try it out!

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Going further: debiasing variational autoencoder (DB-VAE)



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Let’s see facial recognition – and bias – at scale

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

How can we train a powerful CNN so quickly?

- Transfer learning: we use a pre-trained model!
- If a CNN is trained on a large number of images, the spatial feature hierarchy can act as a generic model of the visual world
- Another example: take a model trained on the ImageNet dataset (1.4 million labeled images of 1,000 different classes, such as animals and everyday objects)
  - What type of classification tasks might this be useful for?

A brief recap

A brief recap

- Transparency and interpretability of ML algorithms is important for a number of reasons – luckily, CNNs can be interpreted in fairly natural ways

- One of the reasons that interpretability is important is that we want to avoid our algorithms perpetuating negative biases
  - A lot of approaches to tackle biases rely on explicitly defining discriminatory categories
  - We saw an approach to find (and possibly correct) biases in an unsupervised manner for the case of (facial) image classification

- We also saw that (facial) image classification can be done (possibly in a biased manner) with only a few training images. The “secret sauce” here is transfer learning, that is, building on pre-trained CNNs

Please fill out the module evaluation

https://city.surveys.evasysplus.co.uk/

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON


See you tomorrow!

Sources

- Amini et al., 2019, Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure: http://introtodeeplearning.com/AAAI_MitigatingAlgorithmicBias.pdf
- Chollet, 2021, Deep Learning with Python (2nd edition)
- Géron, 2022, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd edition)
- 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