NHL STENDEN
hogeschool

Looking to the Future of

Machine Learning

Guest lecture by Saul Johnson

Who am I?

I'm Saul, a final-year information security Ph.D. candidate at Teesside University in the north-east of England.

I'm also Head of Software Engineering at BreachLock B.V. an Amsterdam-based offensive security firm.

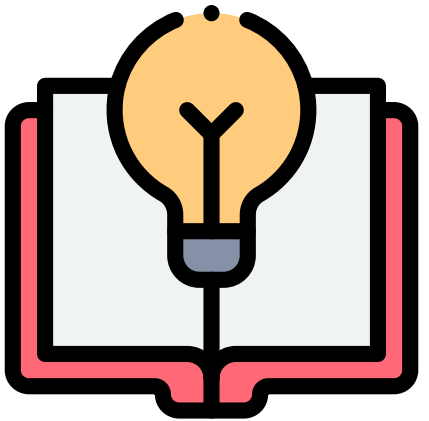
GitHub: @lambdacasserole
Twitter: @lambdacasserole
Website: https://sauljohnson.com/
Linkedin: https://www.linkedin.com/in/sauljohnson/

NHL STENDEN hogeschool

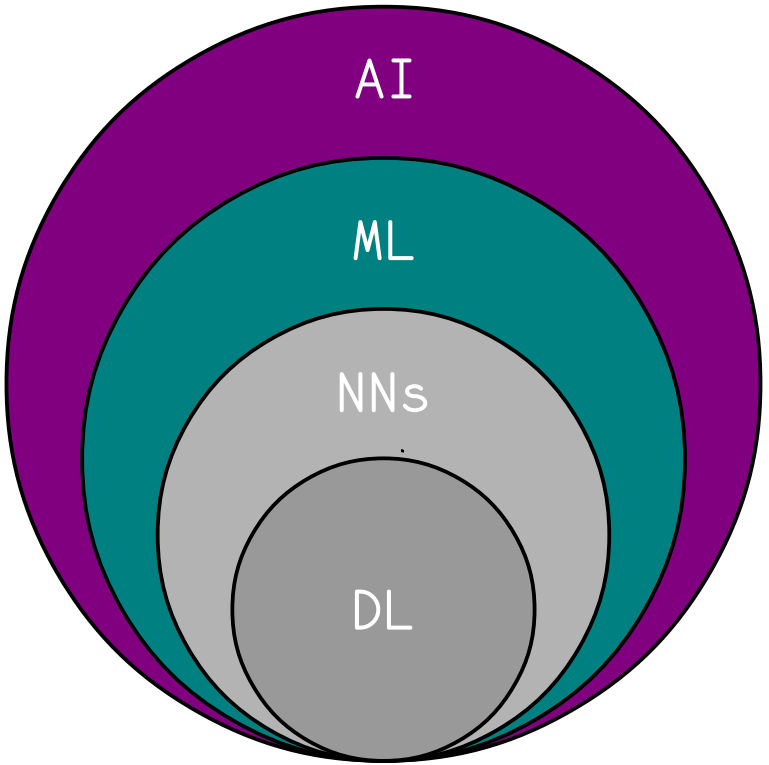
What are we going to cover?

By the end of this session, we'll aim to:

- **Understand** the difference between AI, ML and DL.
- **Know** the threat posed by adversarial inputs to systems that rely on ML models.
- **Be aware** of the concept of Machine Learning as a Service (MLaaS) and how we can leverage it
- **Be familiar in-depth** with the Classr MLaaS platform.
- **Be aware** of some current hot topics and research directions in AI/ML/DL
- **Begin to appreciate** ethical/legal/accessibility concerns around mainstream AI/ML/DL adoption



First, some terminology...

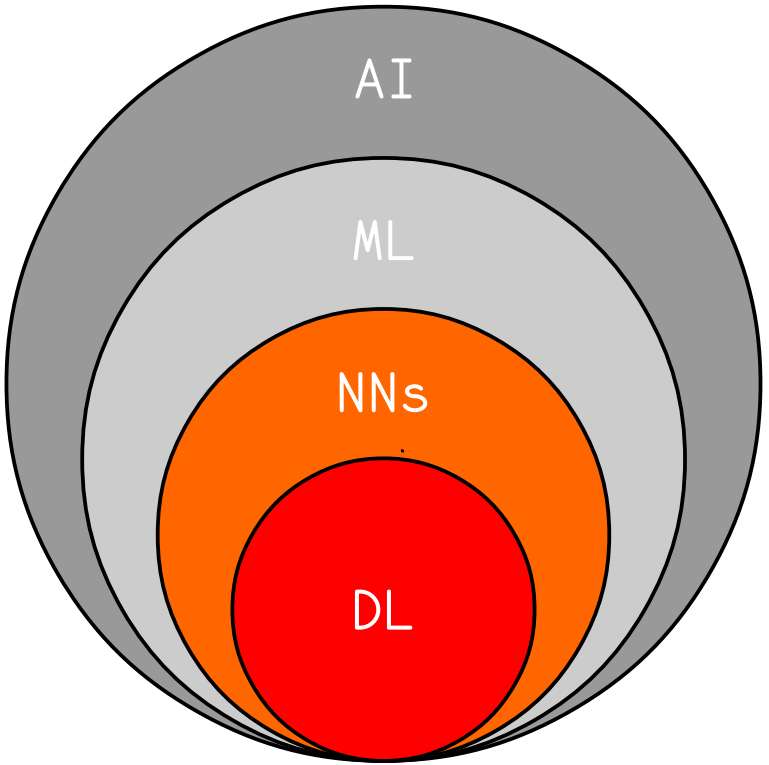


You've covered this before, but let's recap:

- **AI** stands for **artificial intelligence**, which encompasses any technology that allows machines to act in a way that mimics human intelligence. This includes pathfinding, planning etc. as well as ML.

- **ML** stands for **machine learning**, which is a subset of AI in which a model (basically a data structure) is fed inputs to *train* it to do something. We looked at naive bayes classifiers, SVMs and random forests in our last lecture. These are types of ML model.

First, some terminology... (cont.)



Now, for what we didn't cover:

• **NNs** stands for **neutral networks**. These are diverse and powerful ML models that are constructed from layers of neurons. Input is provided by *activating* neurons in the input layer, which produces neuron activations in the output layer, which is our result.

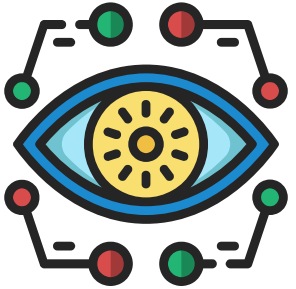
• **DL** stands for **deep learning** which refers to ML using NNs that have multiple layers between the input and output layers (called *hidden layers*) which allows for the creation of extremely powerful models.

ML/DL is very hot right now across...

Psychology

Medical

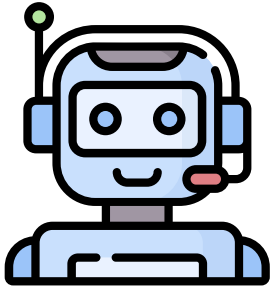
Speech



Vision

Language

Multimedia



Robotics

Learning

NHL STENDEN
hogeschool

Adversarial Inputs

Hacking ML models for fun, profit and chaos.

Adversarial Examples

"panda"
57.7% confidence

noise

"gibbon"
99.3% confidence

[1]

[1] Goodfellow, Ian J. et al. “Explaining and Harnessing Adversarial Examples.” CoRR abs/1412.6572 (2014): n. pag.

Adversarial Examples (cont.)

→

→

[2]

"Stop"
Correct!

"Yield"
A bit like stop..?

"Speed limit"
Oh no.

[2] Tang, R. et al. "An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks." KDD '20 (2020): p. 218–228

Is there any hope?

Yes! We can defend against adversarial inputs using two main strategies: adversarial defence and adversarial detection...

Adversarial defence

• Involves **preventing** successful adversarial attacks by neutralizing them.

• In the case of the panda example from earlier, we might **denoise** the image to defend against the attack.

Adversarial detection

• Involves **detecting** and refusing adversarial inputs.

• In the case of the stop sign example from earlier, we might **use another ML model** to detect the malicious image region and reject the input.

But... sometimes it's just a muffin?

[1]

Sometimes, it's not even an adversarial example that confuses an ML model, but just things that look **weirdly similar**.

For example, it takes more concentration than you might expect to tell which images on the left are of **chihuahuas**, and which are of **blueberry muffins**!

Try keeping your eyes on this text, and classifying the images yourself without looking directly, it's hard!

[1] Karen Zack – Twitter | https://twitter.com/teenybiscuit/status/707727863571582978

NHL STENDEN hogeschool

...or a mop? Or fried chicken?

[1]

[2]

[1] Karen Zack – Twitter | https://twitter.com/teenybiscuit/status/707670947830968320
[2] Karen Zack – Twitter | https://twitter.com/teenybiscuit/status/705232709220769792

NHL STENDEN hogeschool

NHL STENDEN
hogeschool

Taking the pain out of ML

A quick look at Machine Learning as a Service (MLaaS)

MLaaS? What?

Machine learning as a service (MLaaS) describes software platforms that are designed to make it **easier** to create, train and make use of ML models by offering this functionality **as a service**.

By signing up to one of these platforms, businesses are able to build products for classification, image/object recognition, sentiment analysis, natural language processing etc. without **all the added overhead** of building the ML stuff from scratch.

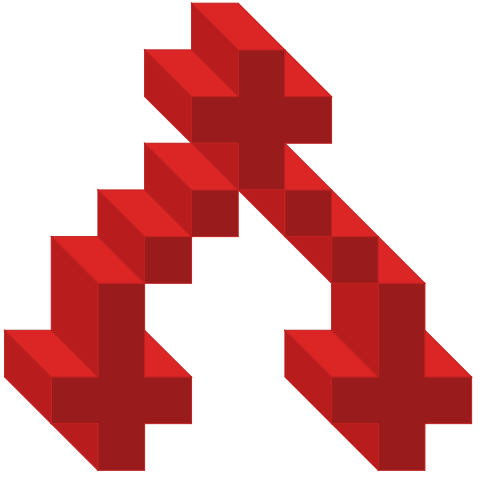
Classr: A (Very Biased) Overview

These platforms can have a bit of a learning curve, so let's start with the absolute basics.

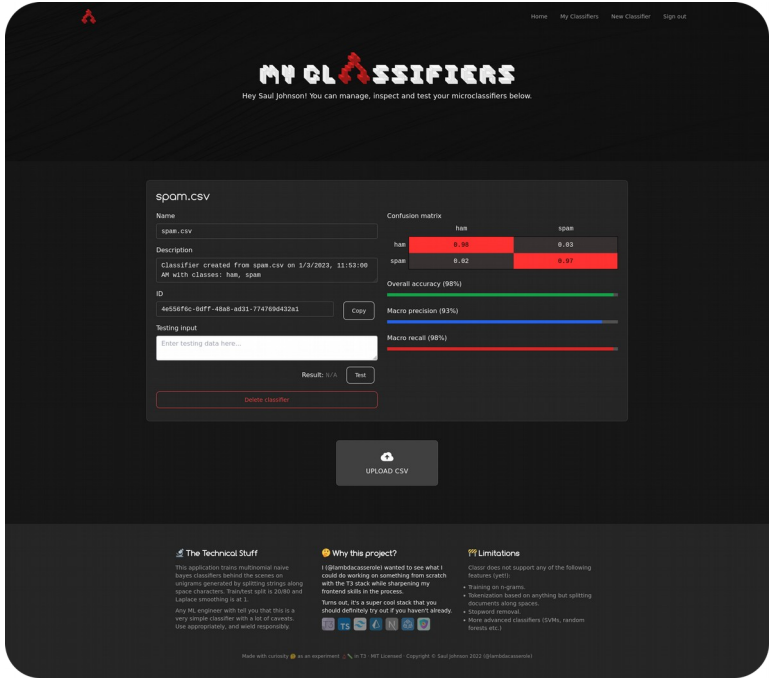
Classr is a very simple MLaaS web application I built after our last lecture together.

It's pretty much the simplest MLaaS you could imagine. It trains **naive bayes classifiers** on up to 4MB of training data from a CSV file you can upload via a simple web interface.

It also offers confusion matrices, precision/recall data, and a simple way to test and interact with you models. **Let's dive in!**



Taking our first steps into MLaaS!



You can access ***Classr*** at https://classr.dev and sign in using your **GitHub account**.

It's completely **free of charge** to use. Feel free to use it when you have simple text classification problems to tackle in future!

You **can** use it for your assignments if you think it's appropriate for your chosen classification problem, but please **do not** use it otherwise. Naive bayes on unigrams may not be powerful enough for your needs!

NHL STENDEN hogeschool

Let's build a spam classifier!

	A	B	C
1	label	document	
2	positive	Fantastic experience, v	
3	negative	Not great to be honest	
4	negative	Weird vibe as soon as	
5	negative	Alright, but disappointi	
6	positive	Best meal I've ever had	
7	positive	Delicious, will definitely	
8	negative	Food was cold and nas	
9	positive	Brilliant food and servic	
10	negative	Asked for medium rare	

Classr takes **2-column CSV files**, with one column titled 'label' and one column titled 'document'. This format is mandatory.

As it uses a **multinomial** model behind the scenes, we can specify as many classes as we like! Let's grab a spam dataset and put it to the test!

NHL STENDEN hogeschool

A simple little code example...

```python
from classr import Classr

# Classify input from the user using our cloud classifier!
cloud_classifier = Classr('f0a42c21-57d3-4b7e-addf-627a586ade66')
document = input('Enter your input: ')
print(cloud_classifier.classify(document))
```

By installing the **Classr SDK** (software development kit) for Python, we can make use of our cloud classifier from our own programs!

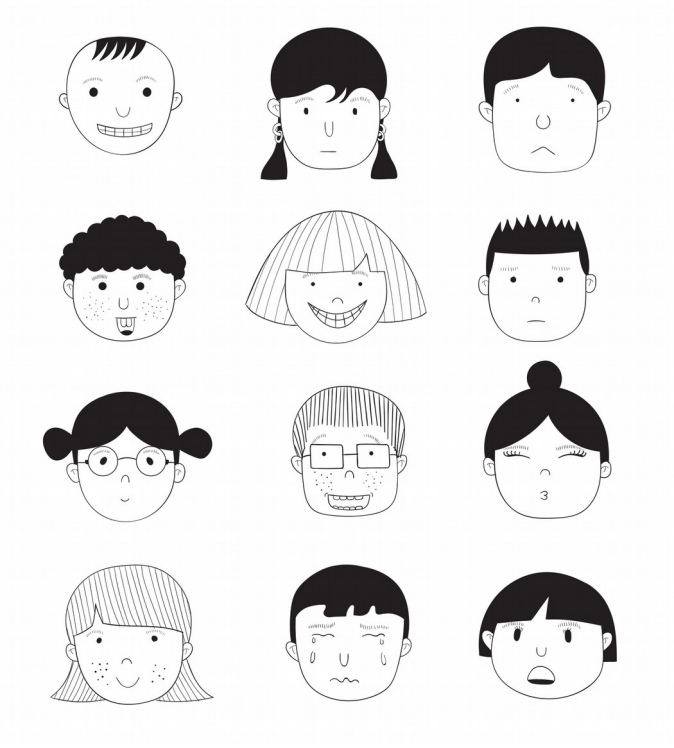
We can train the model via the web app without getting bogged down in code, and build systems to take advantage of it. **This is MLaaS in a nutshell.**

NHL STENDEN hogeschool

Artificial empathy

Let's take a brief look at emotion perception in machines!

Facial Emotion Recognition



Human language and communication obviously goes **a long way beyond** the linguistic content of speech.

Large language models like **ChatGPT (we'll get to that later)** equipped with the ability to recognize and act according to human emotional states may give a very convincing impression of a human-like intelligence.

Facial expressions are a possible starting point here!

NHL STENDEN hogeschool

What can we code up right now?

Of course, in true Python fashion, the **massively complex task** of extracting and analyzing facial expressions from images can be achieved in **just a few lines** of code!

This code sample uses the Python **DeepFace** and **OpenCV** libraries to capture images from the webcam, extract faces and analyze their emotional state.

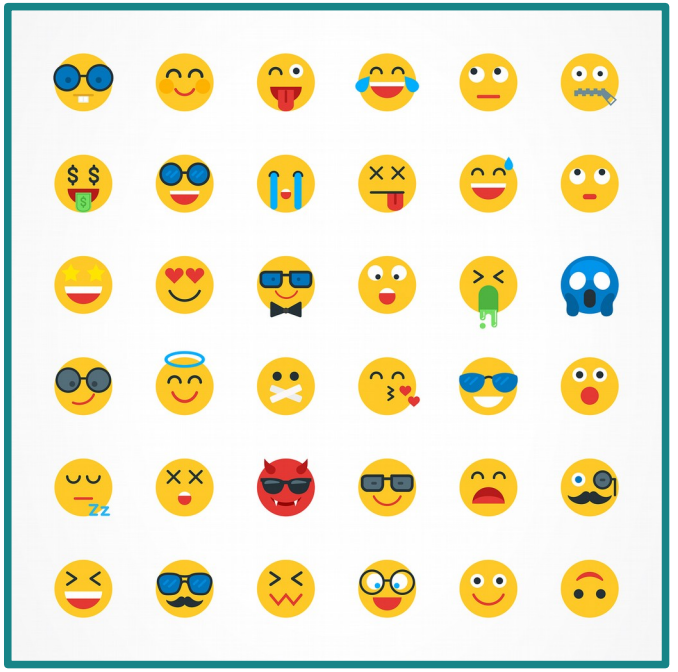
```python
from deepface import DeepFace
import cv2

# Read image from camera.
camera = cv2.VideoCapture(0)
result, image = camera.read()

# Analyze facial expression.
analysis = DeepFace.analyze(image, actions=['emotion'])
emotions = analysis['emotion']

# Print results.
print(emotions)
```

Let's deploy Classr again!

Why stop here? Let's jump back into Classr again and deploy a simple **sentiment analysis** model.

What if we were to combine **facial emotion recognition** with **SER (speech emotion recognition)** and analysis of the **linguistic content** of speech?

The possibilities are really exciting to consider. Do you think we're approaching a point where we'll be able to interact with machines as if they're people?



NHL STENDEN hogeschool

And it doesn't stop there...

```python
import speech_recognition as sr

# Initialize speech recognition engine.
recognizer = sr.Recognizer()

# Use the microphone to get audio...
with sr.Microphone() as mic:

    # Filter out background noise.
    recognizer.adjust_for_ambient_noise(mic, duration=0.2)

    # Capture audio, turn it into text.
    audio = recognizer.listen(mic)
    text = recognizer.recognize_google(audio)

    # Print results.
    print('It sounded like you said: ', text)
```

Capturing audio from the computer's microphone and turning it into text for sentiment analysis is **almost as easy** in the Python ecosystem as facial emotion recognition!

Here we use the **speech_recognition** library to do just that!

A word of caution, however...

Before we get carried away, it's worth thinking about **who** this technology will benefit and **how.**

Human beings are **incredibly diverse**. If we build a system that depends on a user's facial expressions, how will it respond to a **stroke survivor** with partial facial paralysis?

This is just the **tip of the iceberg**. AI/ML/DL could **supercharge** our ability to build accessible systems… or it could **seriously marginalise** all sorts of people. Don't let shiny new tech distract from that!

NHL STENDEN
hogeschool

ChatGPT: The Elephant in the Room

I know you know what I'm talking about!

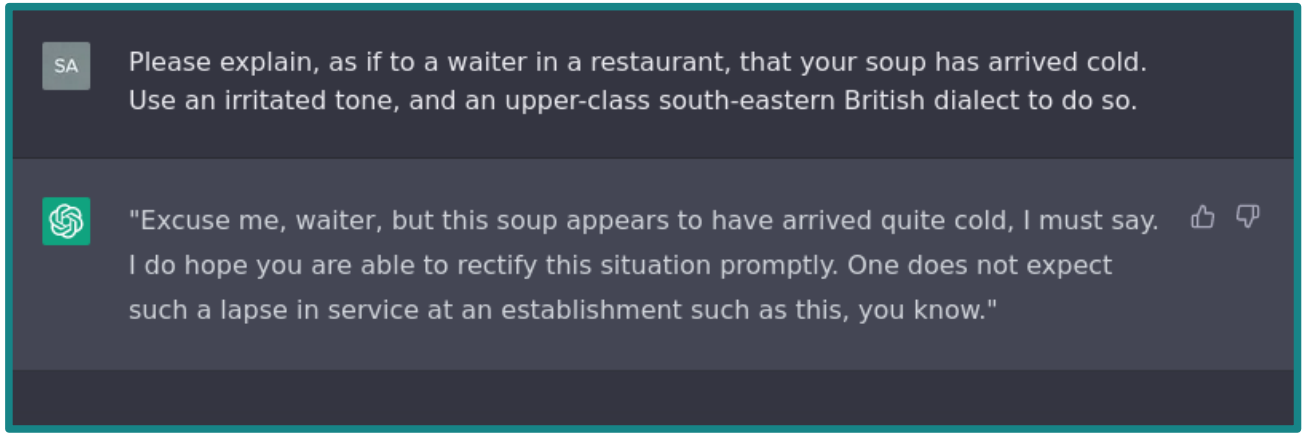
What is ChatGPT actually?

ChatGPT is a **large language model (LLM)** built using **GPT 3.5 (Generative Pre-trained Transformer v3.5)** with a chatbot-like interface.

It's a **transformer network** (a type of deep learning model) that has been trained using petabytes of data and that has over **175 billion parameters**.

The trained model requires **800GB** (almost a terabyte) of storage space, and costs **millions of dollars per month** to run. One can only imagine how much it cost to train!

Why the fuss?



Further reading:
Transfer learning. What is it? Why is it useful?

The hype around ChatGPT is **understandable.** Never before has the general public had access to an LLM **this powerful**, and it's poised to potentially **change** the way we interact with computers **forever.**

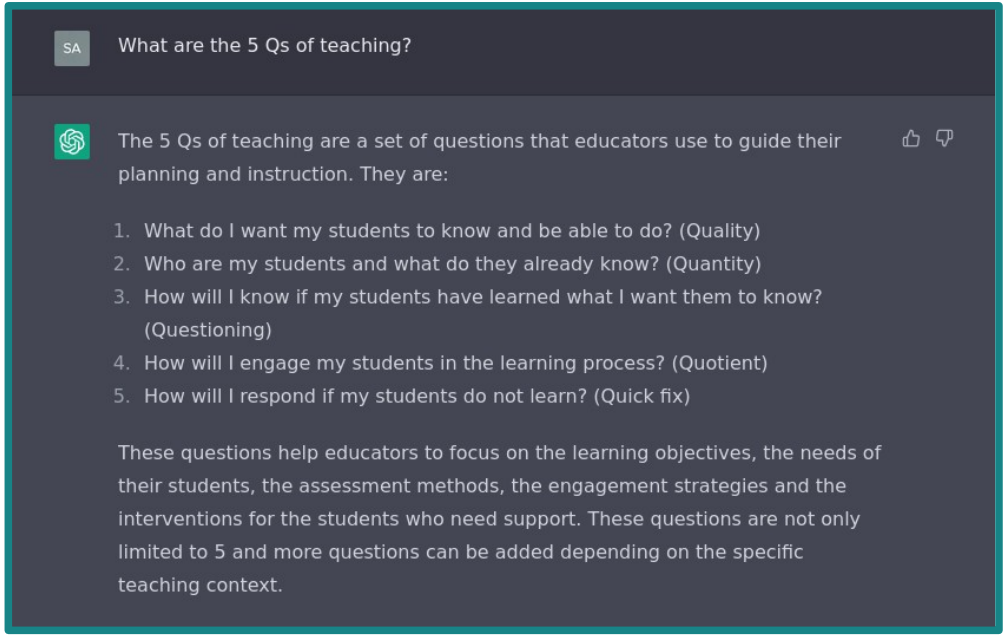
NHL STENDEN hogeschool

ChatGPT hallucinates!

Under the **slick surface** and **clever engineering**, however, lies a relatively **straightforward** premise: anticipating the most **statistically probable** text in response to a particular prompt given.

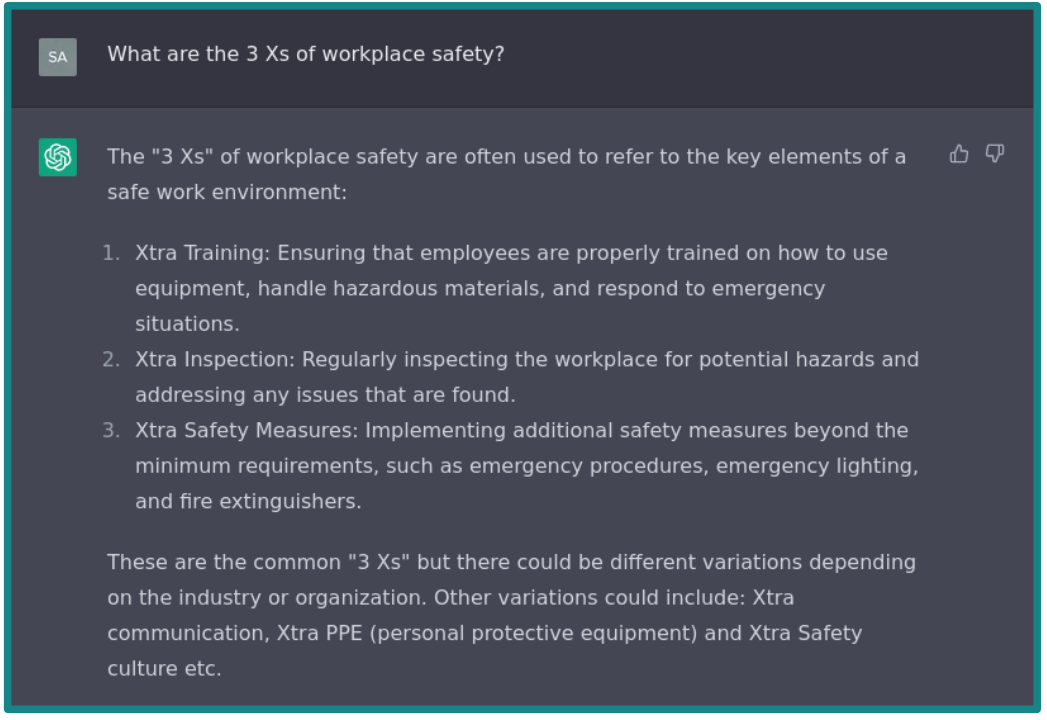
This is most easily exposed when you ask it a **seemingly sensible** (but **nonsensical**) question, which causes it to **"hallucinate"** and confidently express an "answer".

See right. The 5 Qs of teaching do not exist. They are not a thing.



NHL STENDEN hogeschool

ChatGPT hallucinates! (cont.)



The 3 Xs of workplace safety **do not exist** and are not a thing either. "Xtra training", "Xtra inspection" what?

Though it is trained on **a lot** of data, ChatGPT has nevertheless been trained on **human-generated** data. This data is not always entirely correct and is sometimes **downright wrong**.

Further reading:
Reinforcement learning.
What is it? Why is it useful?

NHL STENDEN hogeschool

NHL STENDEN hogeschool

And much, much more...

A brief tour of the latest, coolest, and most controversial in ML!

AI "Art" and Prompt Engineering

"a seagull eating a slice of pizza in the style of vincent van gogh"
- Saul Johnson 2023 (?)

I have **zero artistic ability**, but on the left you'll find "my" masterpiece, and the prompt I fed to a generative ML model called **DALL-E** (also from OpenAI) to create it.

Is this my creation? DALL-E was trained on millions of images, and the authors of those images were never consulted as to whether or not they're okay with that.

A program has taken my natural language prompt, and essentially created this image from what is in its training set. **Is this really ethical?**

NHL STENDEN hogeschool

Neural Radiance Fields (NeRFs)

Neural Radiance Fields (NeRFs) refer to a class of fully-connected (dense) **neural networks** that can construct and render complex 3D scenes from relatively few still 2D image inputs.

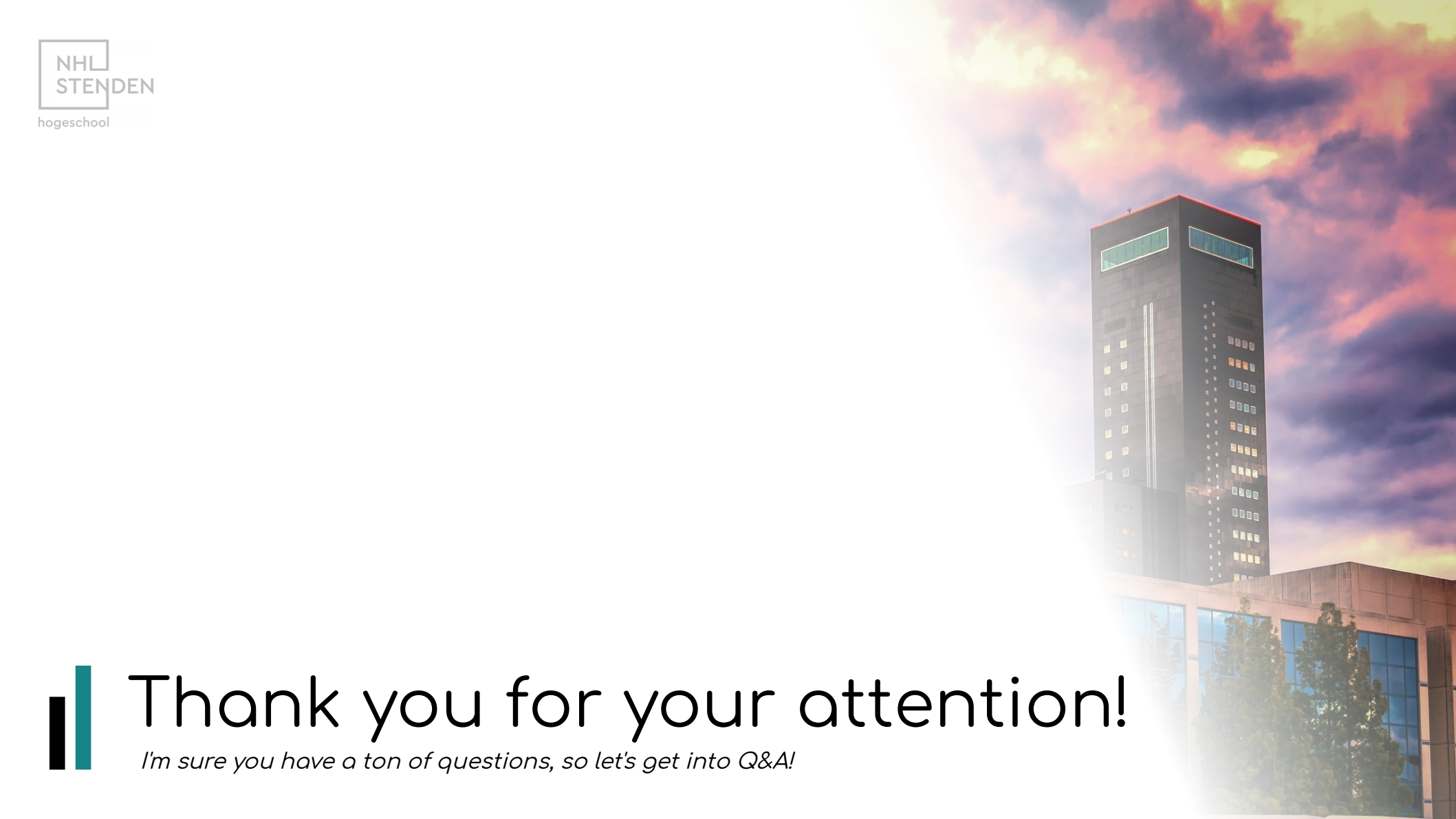
NeRFs represent an exciting research direction and are likely to play a significant role in the **future of imaging!**

Let's take a look at NeRFs in action...

Martin-Brualla et al. Google Research - https://www.youtube.com/watch?v=mRAKVQj5LRA

NHL STENDEN hogeschool

NHL STENDEN hogeschool



Thank you for your attention!

I'm sure you have a ton of questions, so let's get into Q&A!