Silicon Brains:

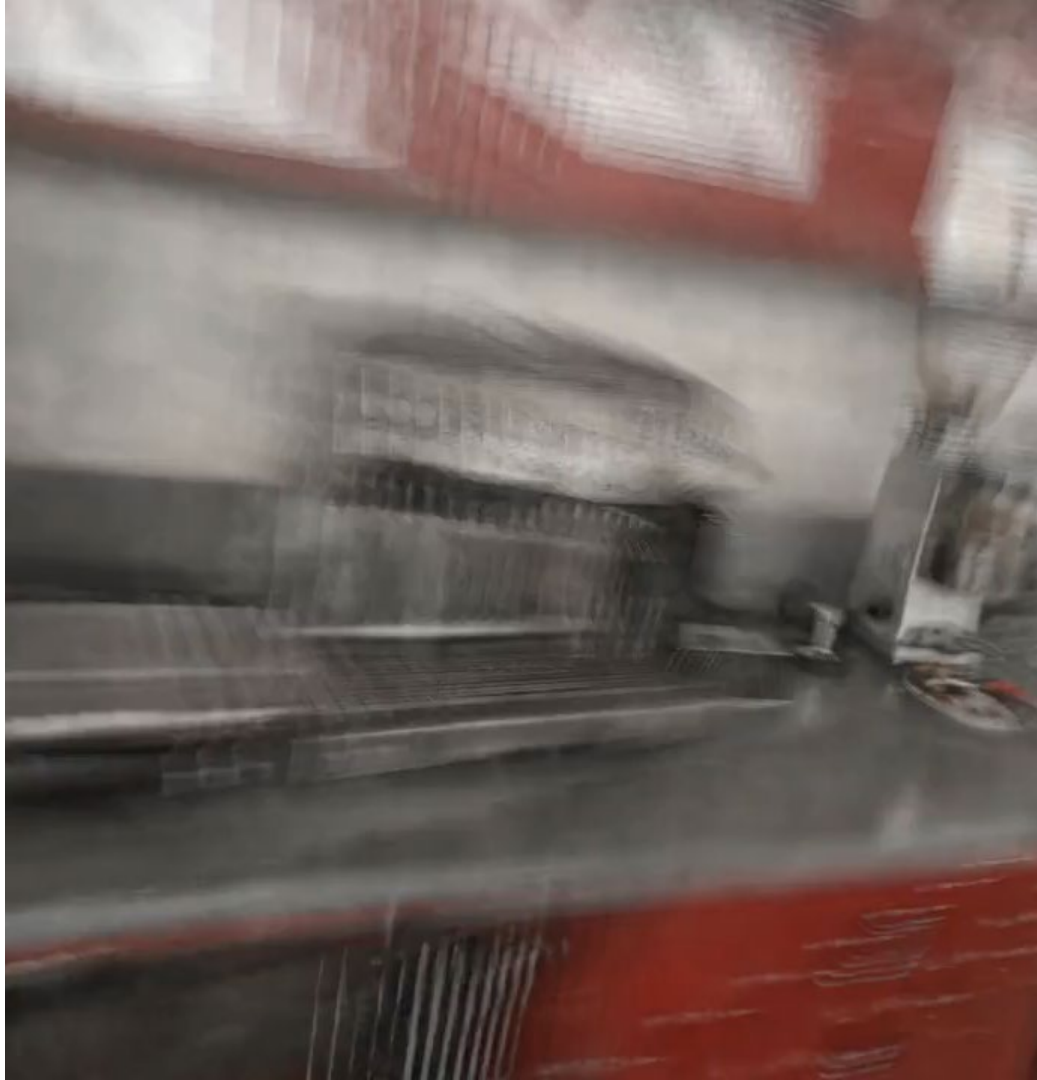
A Field Guide to the Dangers and Drawbacks of Machine Learning

Jack Naylor
PhD Candidate
Australian Centre for Field Robotics,
Faculty of Engineering

Sydney Innovation Program - 7/2/23,
The University of Sydney Law School

THE UNIVERSITY OF SYDNEY

ACFR
AUSTRALIAN CENTRE FOR FIELD ROBOTICS

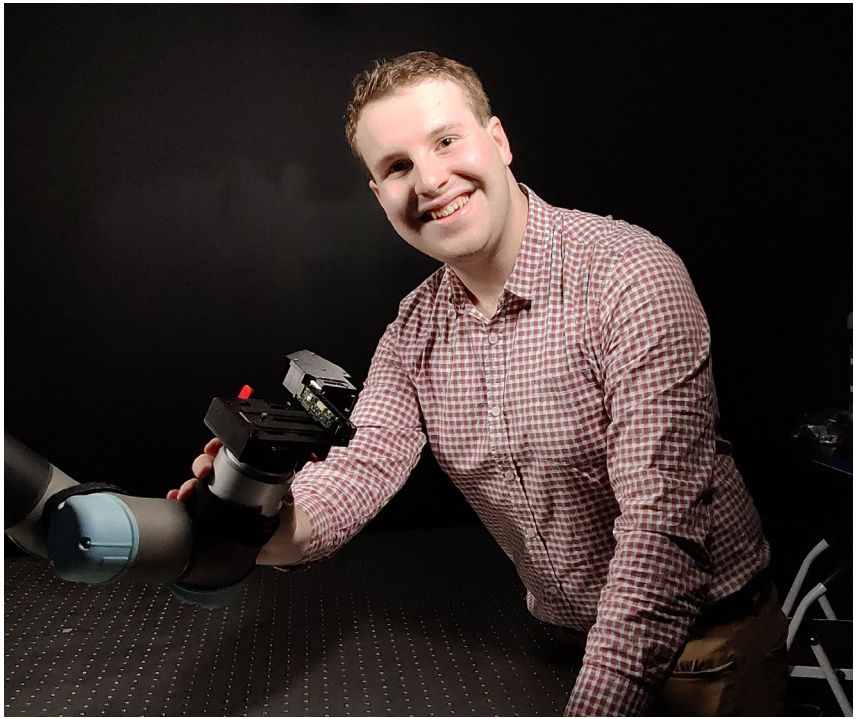


A Crash Course in 2 Hours

- What is Machine Learning, how good is it really? (15 mins)
- Classification (20 mins)
- Unsupervised Machine Learning (20 mins)
- 5 min break
- Deep Learning w/ Images (20 mins)
- Large Language Models (15 mins)
- Ethical AI and Safeguarding Users (20 mins)
- Q&A (5 mins)

About Me

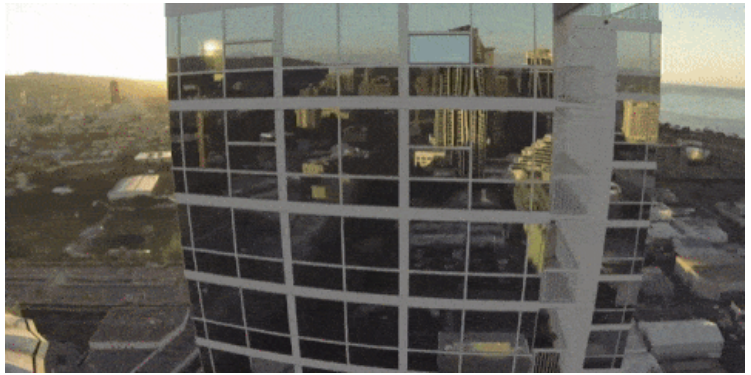
- BEng(Mech)(Hons. 1)/BSc(Adv)
- Majors in Space Engineering/Physics
- Honours @ Nearmap on small object reconstruction
- Teach Experimental Robotics, System Dynamics & Control + Mechatronic Systems Design
- Leading collaboration w/ Physics & Geosciences on AI-enabled edge data curation for remote sensing




BOLD
GORGE

Appearances Are Deceiving to Robots

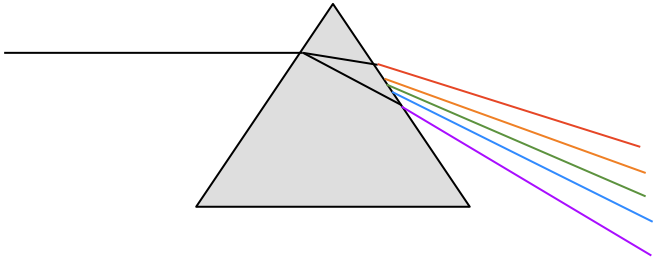
- **Features are hard to track with refraction**
- **Reflection at non-normal angles mis-represents distance**
- **Result:**
  - Drones crash
  - Spot walks onto black ice!



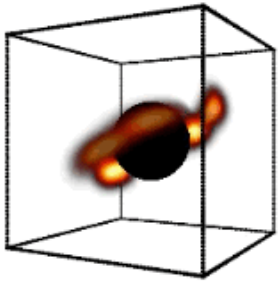
Neural Fields

- **Fields represent the underlying physical quantities which we measure**
- **Gravity, acoustics, fluids, light, electromagnetism are the language of nature**
- **To reconstruct the field is exceedingly challenging – so we use a neural network! A.k.a Neural Fields!**

Acoustics



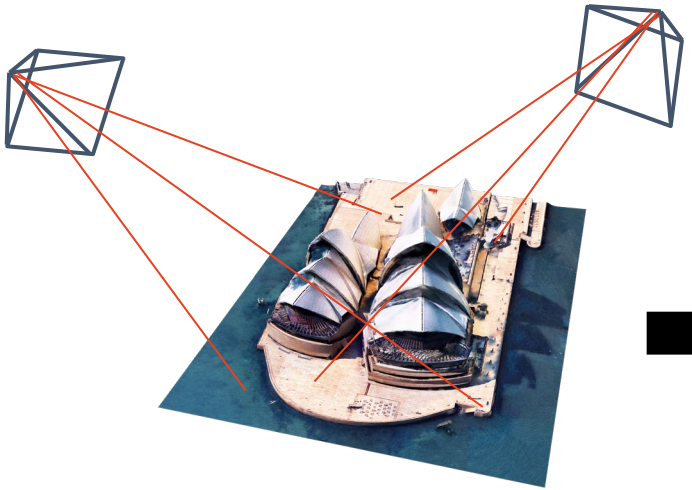
Light



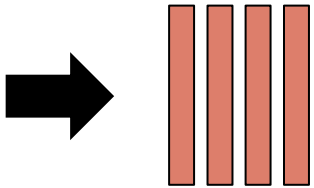
… even Black Hole Emissions!

Levis et al. 2021

Reconstructing Light



Discrete Sampling of Continuous Field (Image)



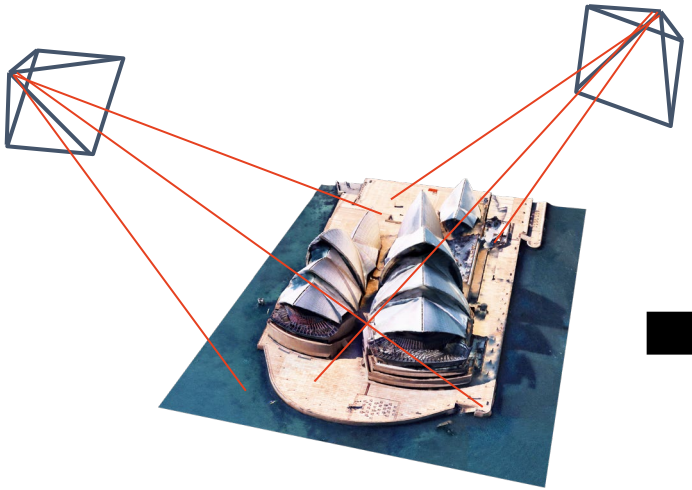
F_Θ

Network Approximating Plenoptic Function

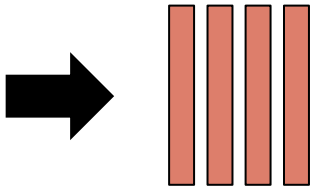
?

Continuous Scene Representation

Reconstructing Light



Discrete Sampling of Continuous Field (Image)



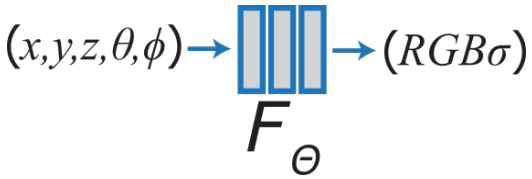
F_Θ

Network Approximating Plenoptic Function

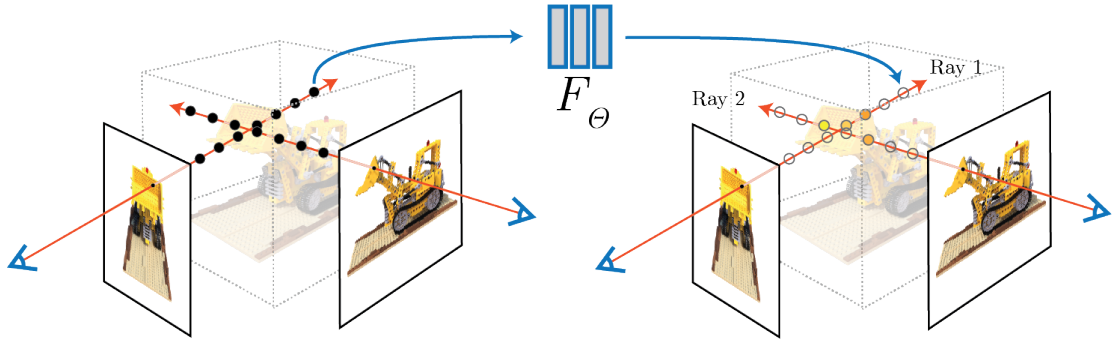
NeRF

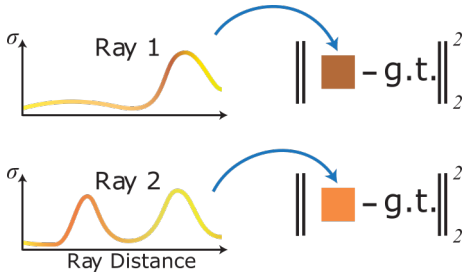
Continuous Scene Representation (Neural Radiance Field)

“It’s NeRF or nothing.”



- Learn a 5D representation of light in a NN.
- Spatially varying density, spatially/view varying colour.
- Novel view synthesis.

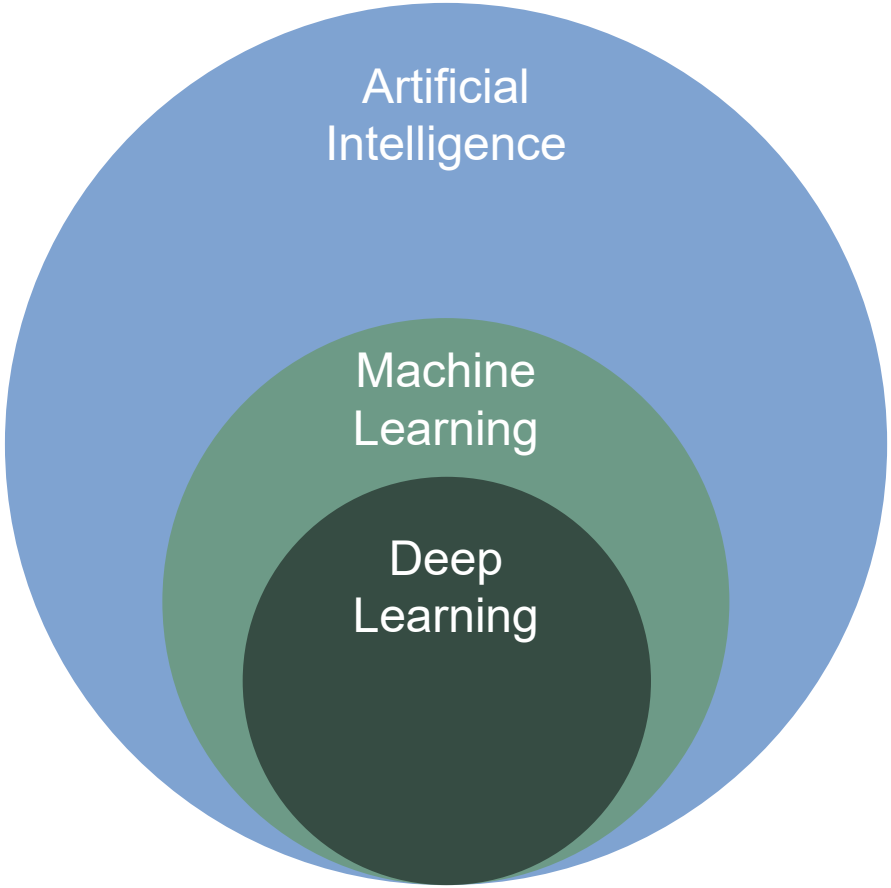




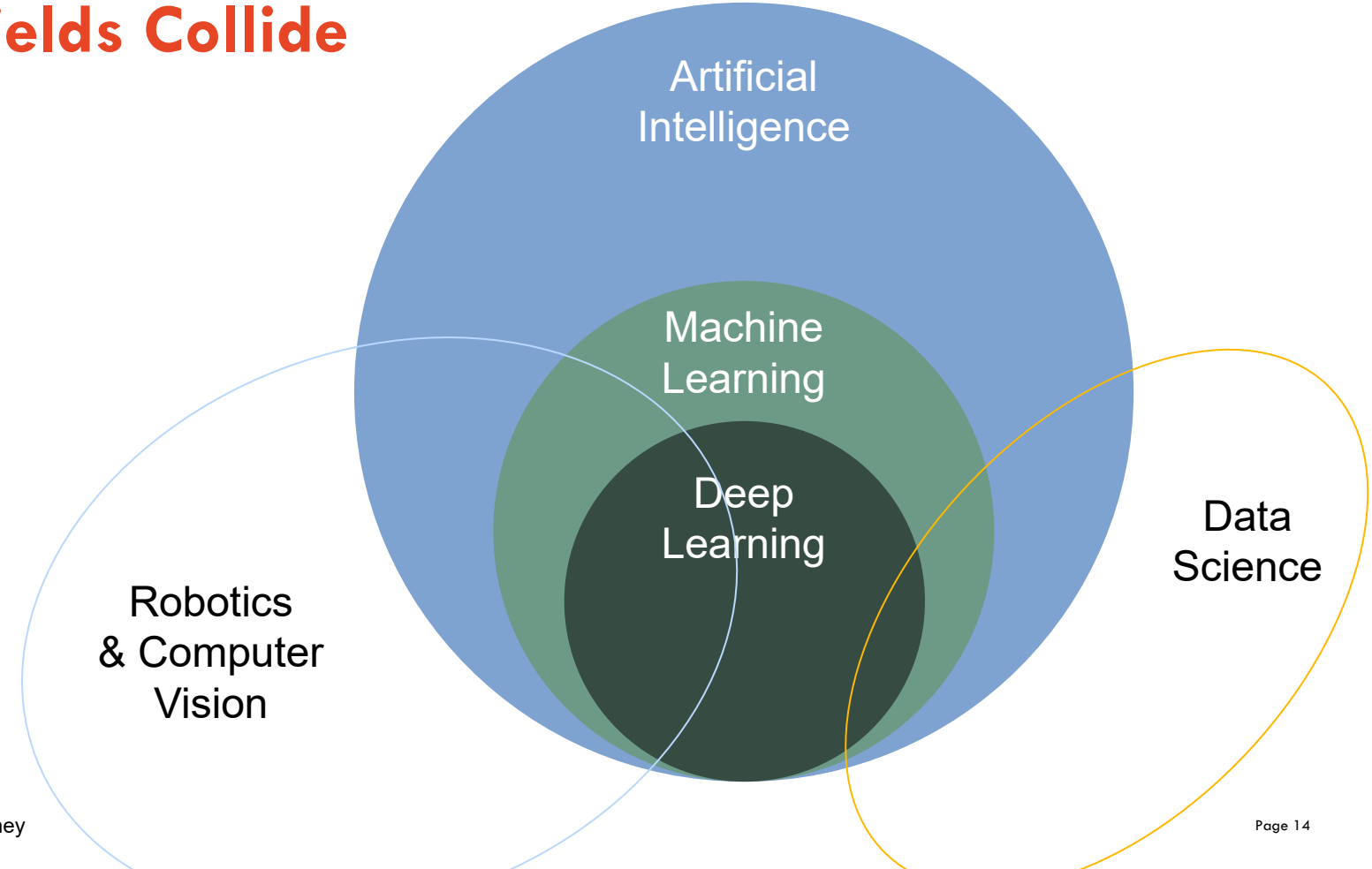
Mildenhall et al. 2020

What is Machine Learning, how good is it really?

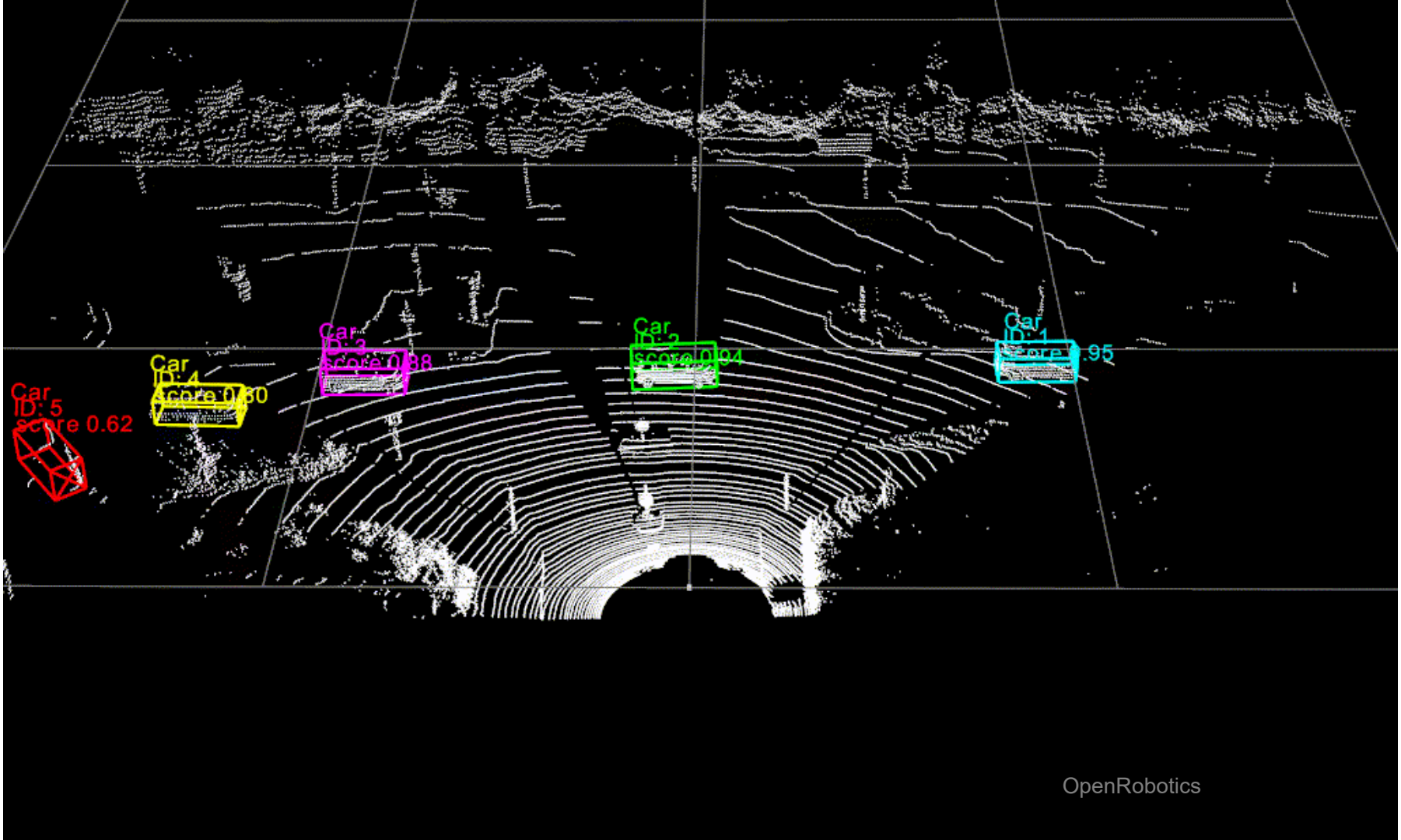
When Fields Collide



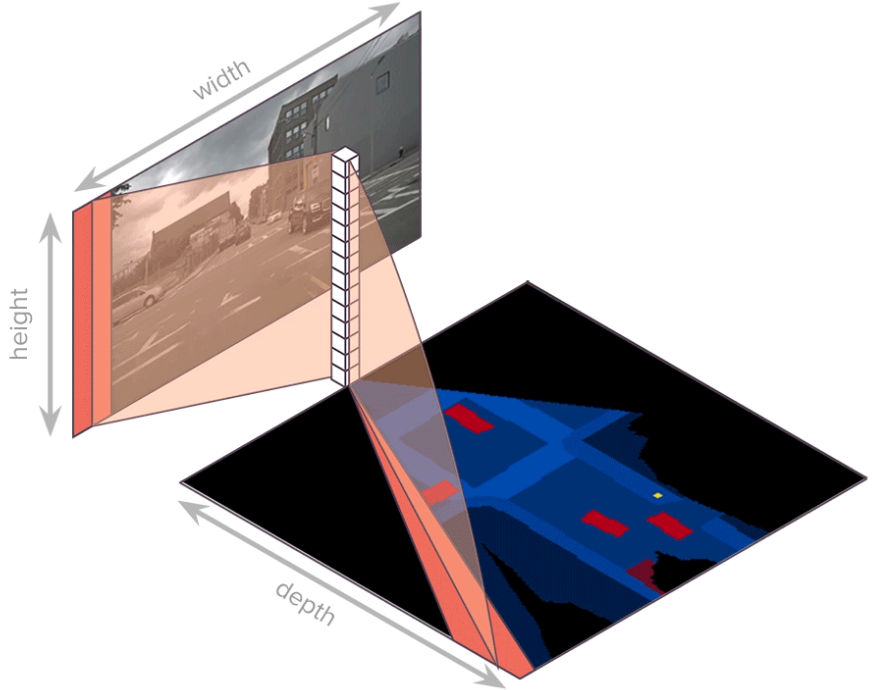
When Fields Collide




100M
WORLD
RECORD


Car
ID: 5
score 0.62
Car
ID: 4
score 0.80
Car
ID: 3
score 0.88
Car
ID: 2
score 0.94
Car
ID: 1
score 0.95
OpenRobotics

Waymo



Amazon Science

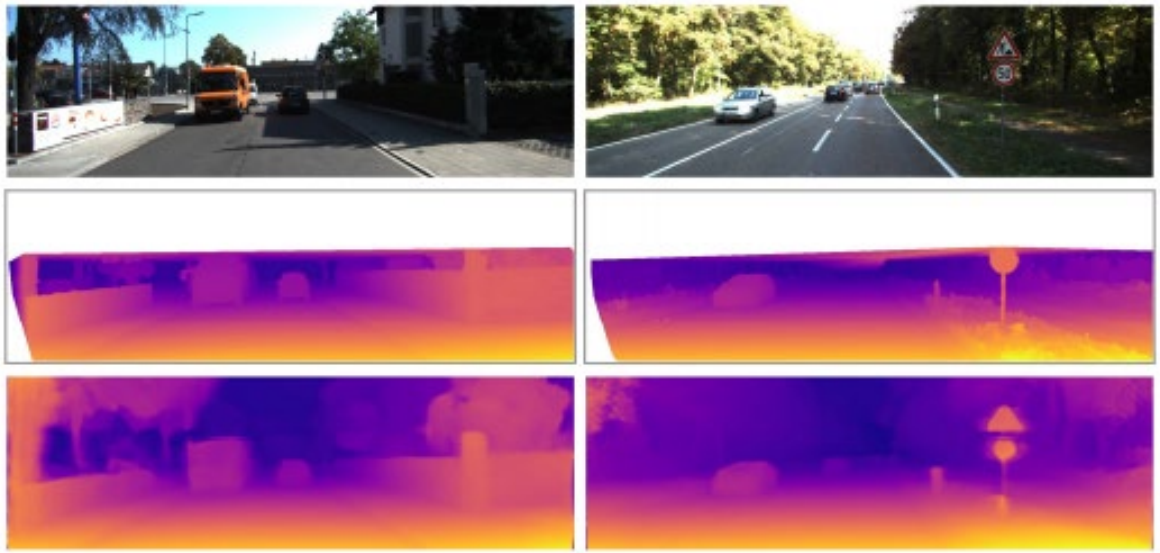
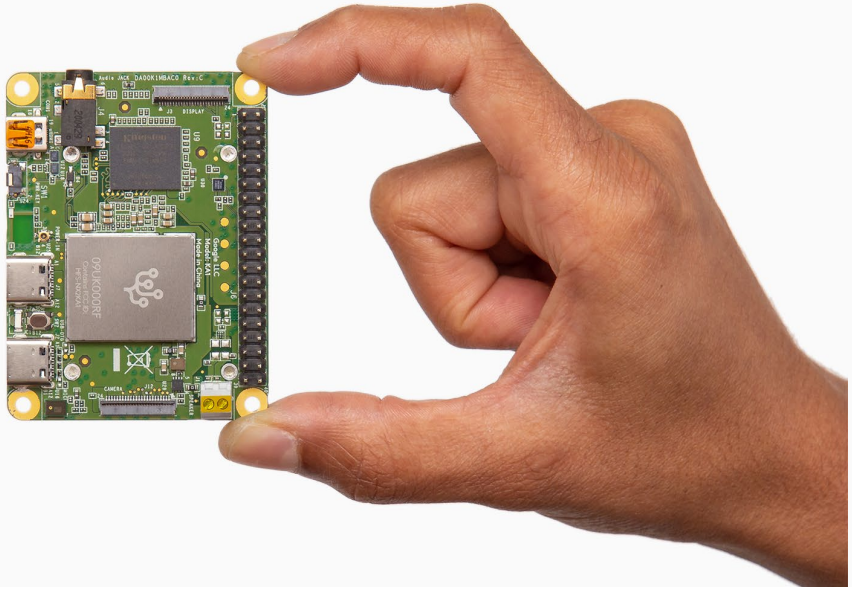


Figure 1. Our depth prediction results on KITTI 2015. Top to bottom: input image, ground truth disparities, and our result. Our method is able to estimate depth for thin structures such as street signs and poles.

Goddard et al.


StyleGAN2
StyleGAN3 (Ours)

No longer need big GPUs! (Well, kinda)



Classification

How to automate mundane, monotonous tasks

Supervised Machine Learning

- **Label data and learn a function to map input to output**
- **2 types:**
  - Classification: break data into categories
  - Regression: give a numerical result
- **Most useful to predict outputs from unseen data**

Classification

- Those rewards programs, aren't just for you

- The same algorithms we might want to work out whether fruits are citrus, berries etc. from their sugar content are what power retail stores, dating apps

The New York Times Magazine

How Companies Learn Your Secrets

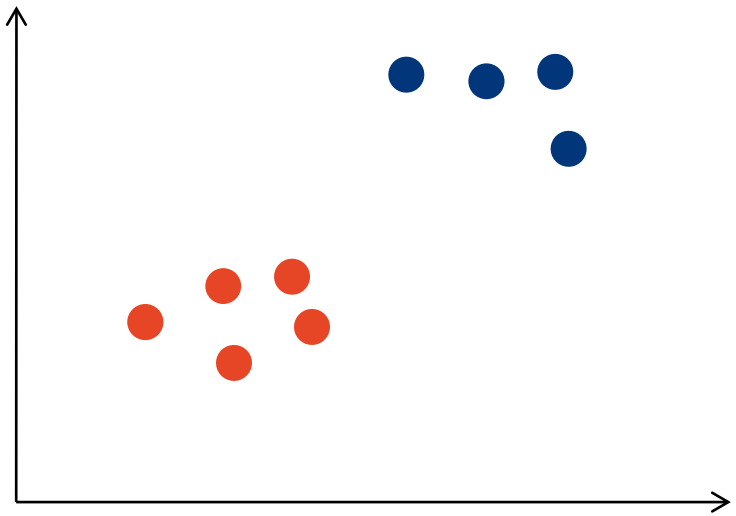
Give this article 570

Antonio Bolfo/Reportage for The New York Times

By Charles Duhigg
Feb. 16, 2012

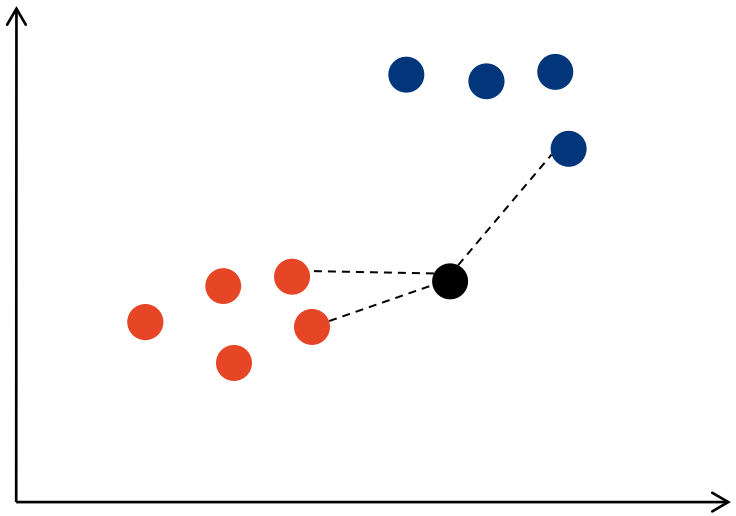
Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

How Might We Do This?



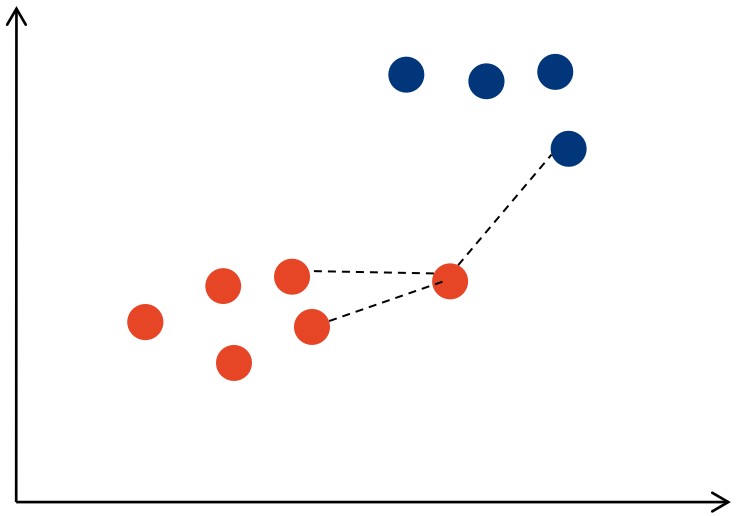
- Might look at how data is clustered?
- K-Nearest Neighbours

How Might We Do This?



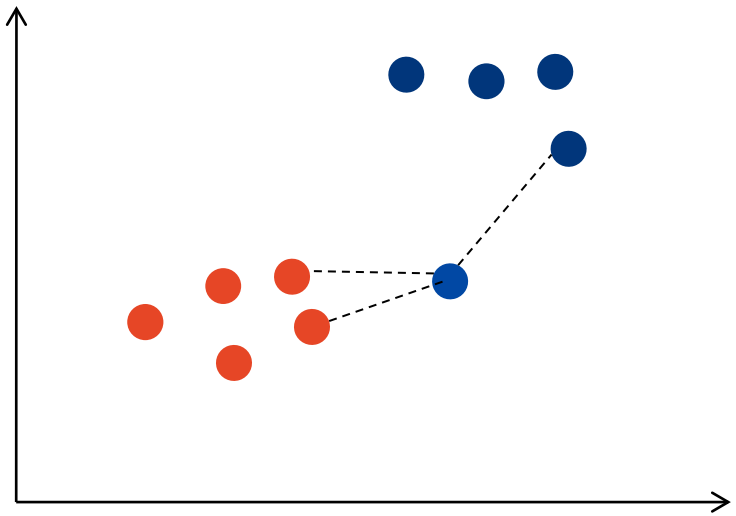
- Might look at how data is clustered?
- K-Nearest Neighbours

How Might We Do This?



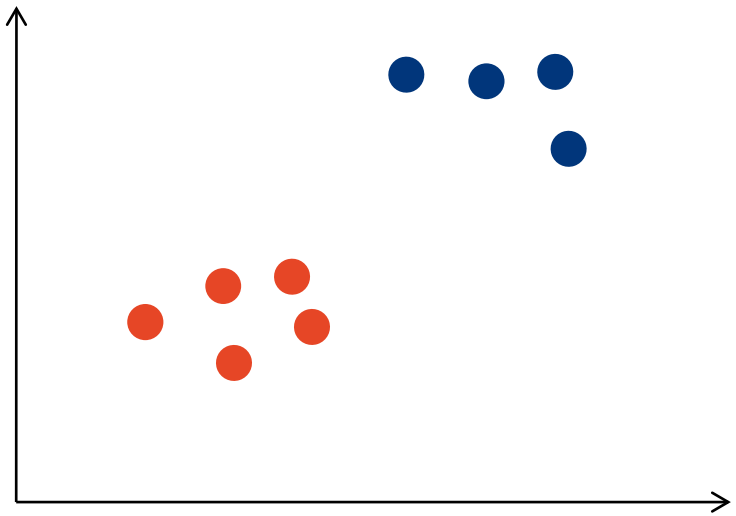
- Might look at how data is clustered?
- K-Nearest Neighbours

How Might We Do This?



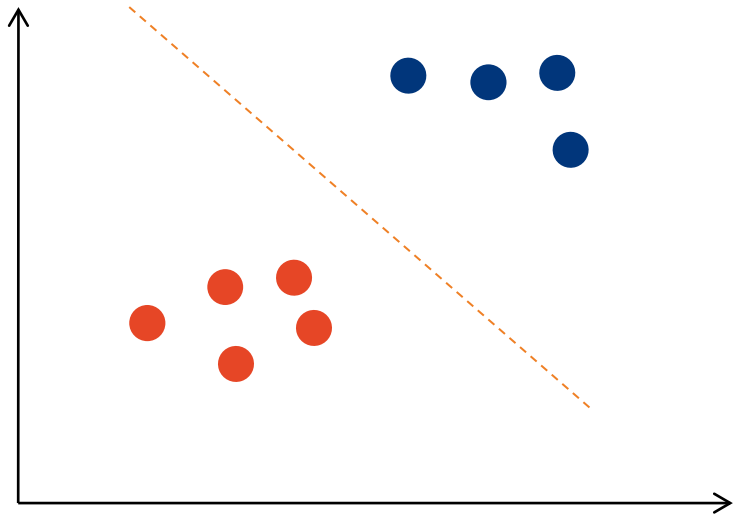
- Might look at attributes and make a decision statistically
- Naïve Bayes
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

How Might We Do This?



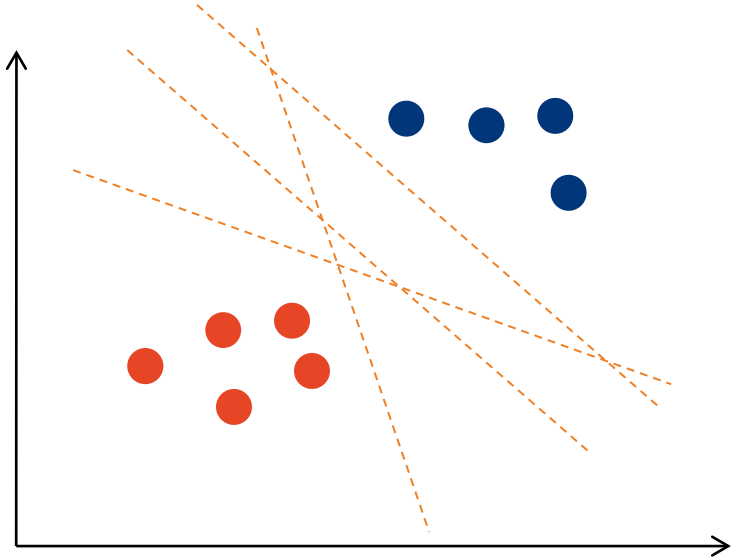
- Let us look for a boundary between the data
- Enter: the Support Vector Machine

How Might We Do This?



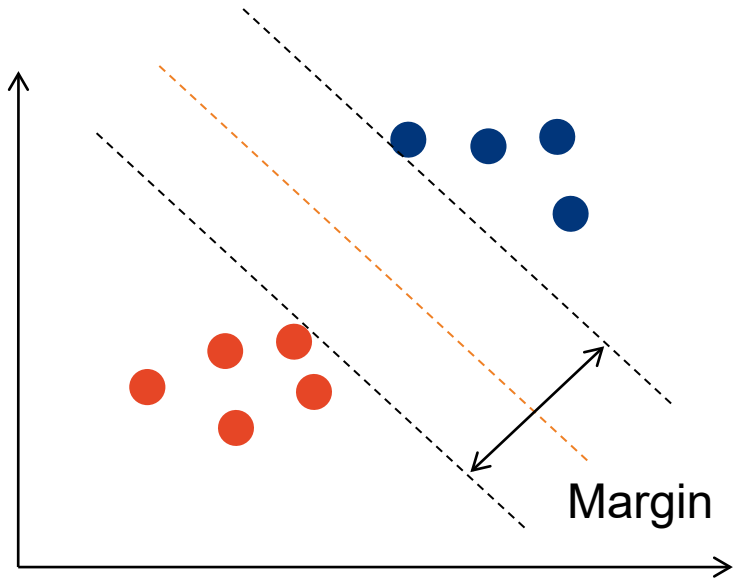
- Let us look for a boundary between the data
- Enter: the Support Vector Machine

How Might We Do This?



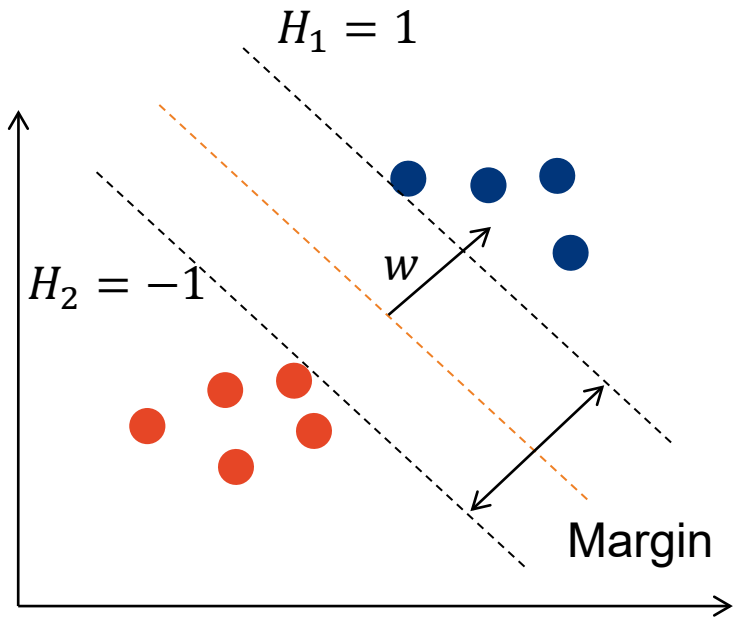
- Let us look for a boundary between the data
- Enter: the Support Vector Machine

How Might We Do This?



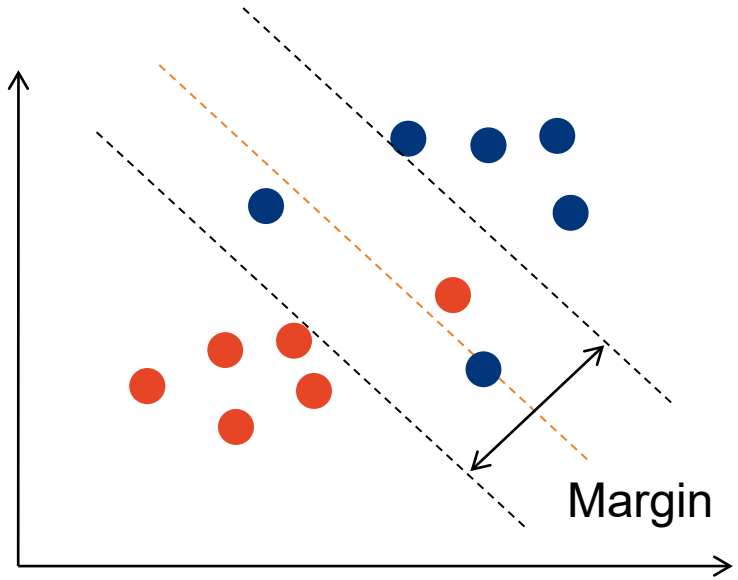
- Let us look for datapoints known as support vectors
- Use these to place a decision boundary w/ max margin
- Larger margin = more robust

How Might We Do This?



- Decision Boundary = Hyperplane
- $H_{1/2}: w \cdot x + b = \pm 1$
- w is vector which defines hyperplane placement
- Margin = $\frac{2}{||w||}$, so we want to minimise w

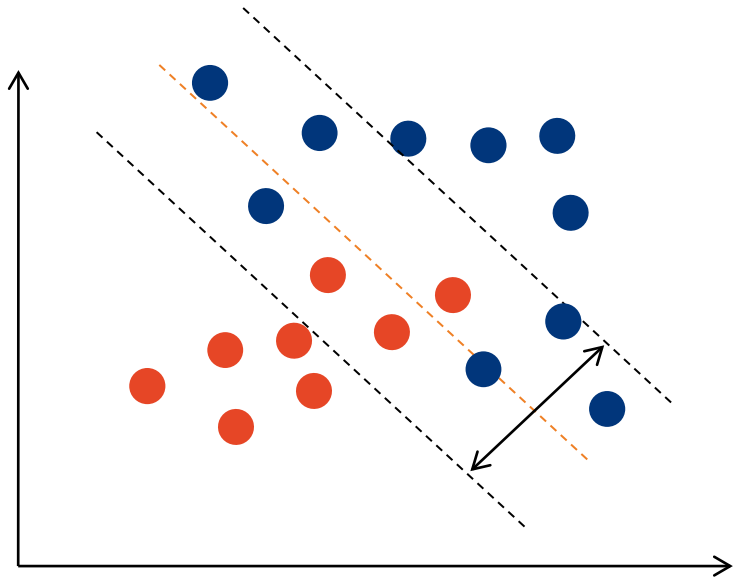
How Might We Do This?



- Outliers? We make optimise such that margin is best, so some classifications might be incorrect.

- What's more troubling is data which is not well separated by lines…

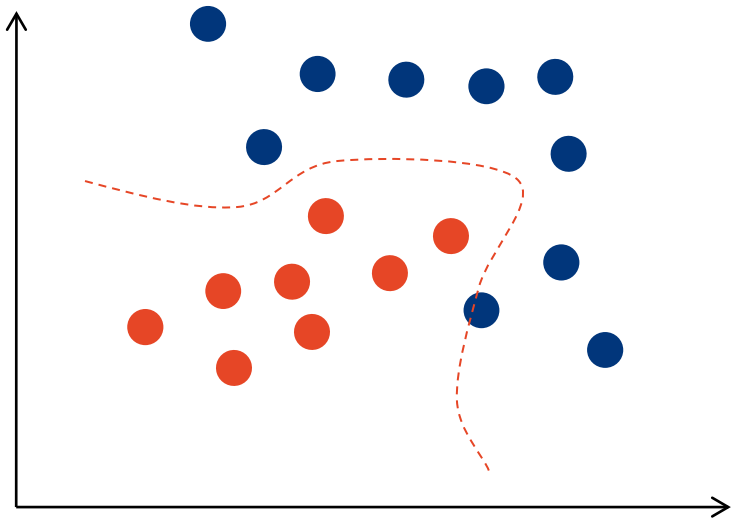
How Might We Do This?



- Outliers? We make optimise such that margin is best, so some classifications might be incorrect.

- What's more troubling is data which is not well separated by lines…

How Might We Do This?



- I won't go into the maths here, but…
- Transform data to higher dimensions, we can draw straight lines to cut the distribution.

An example with Ping Pong balls

Let's try out an SVM to try and detect breast cancer.

What worked best?

Which non-linear kernels?

Why is it not 100% accurate?

Would you trust this instead of a Doctor?

Unsupervised Machine Learning

Letting the data make decisions

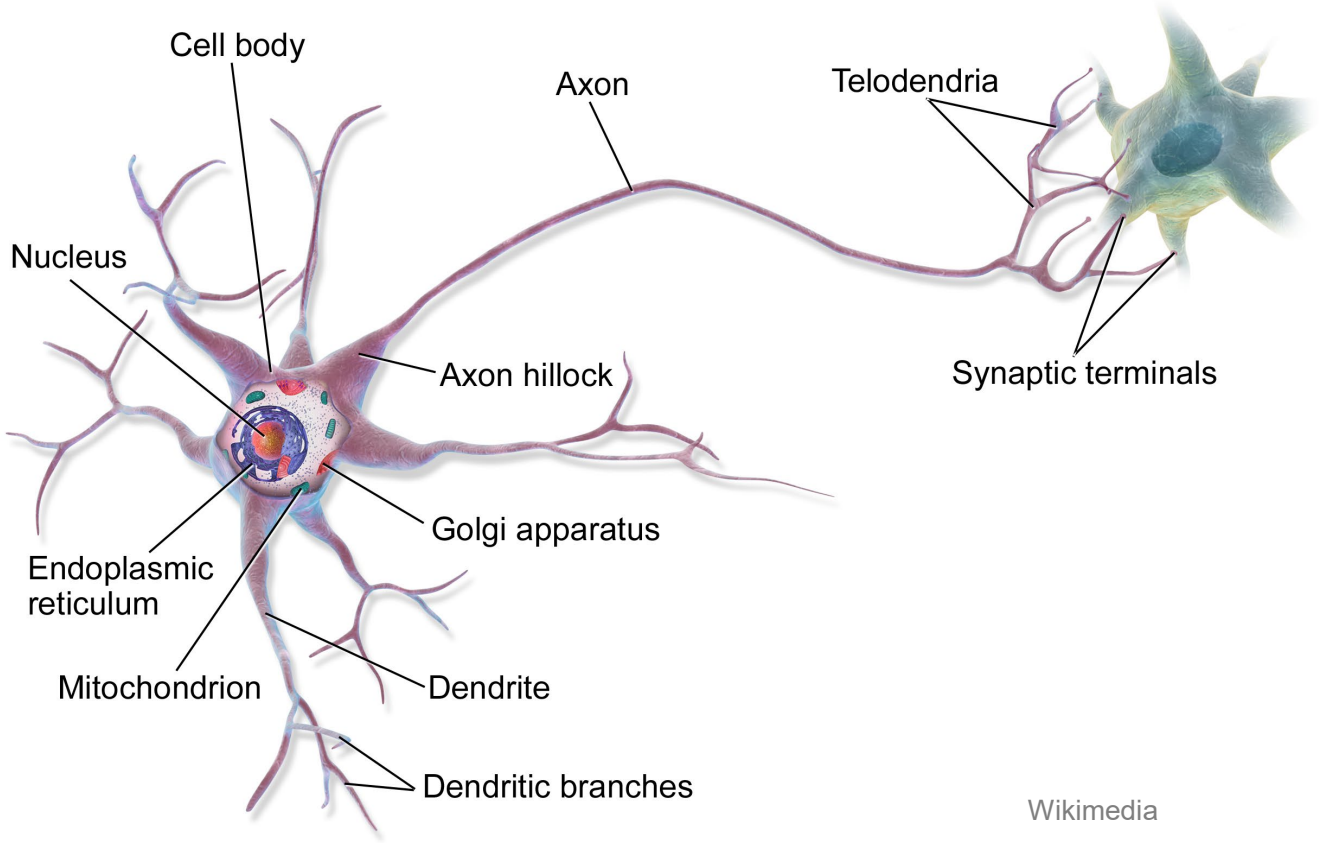
Why would we not want to supervise data?

Supervision = a form of prior, something we know

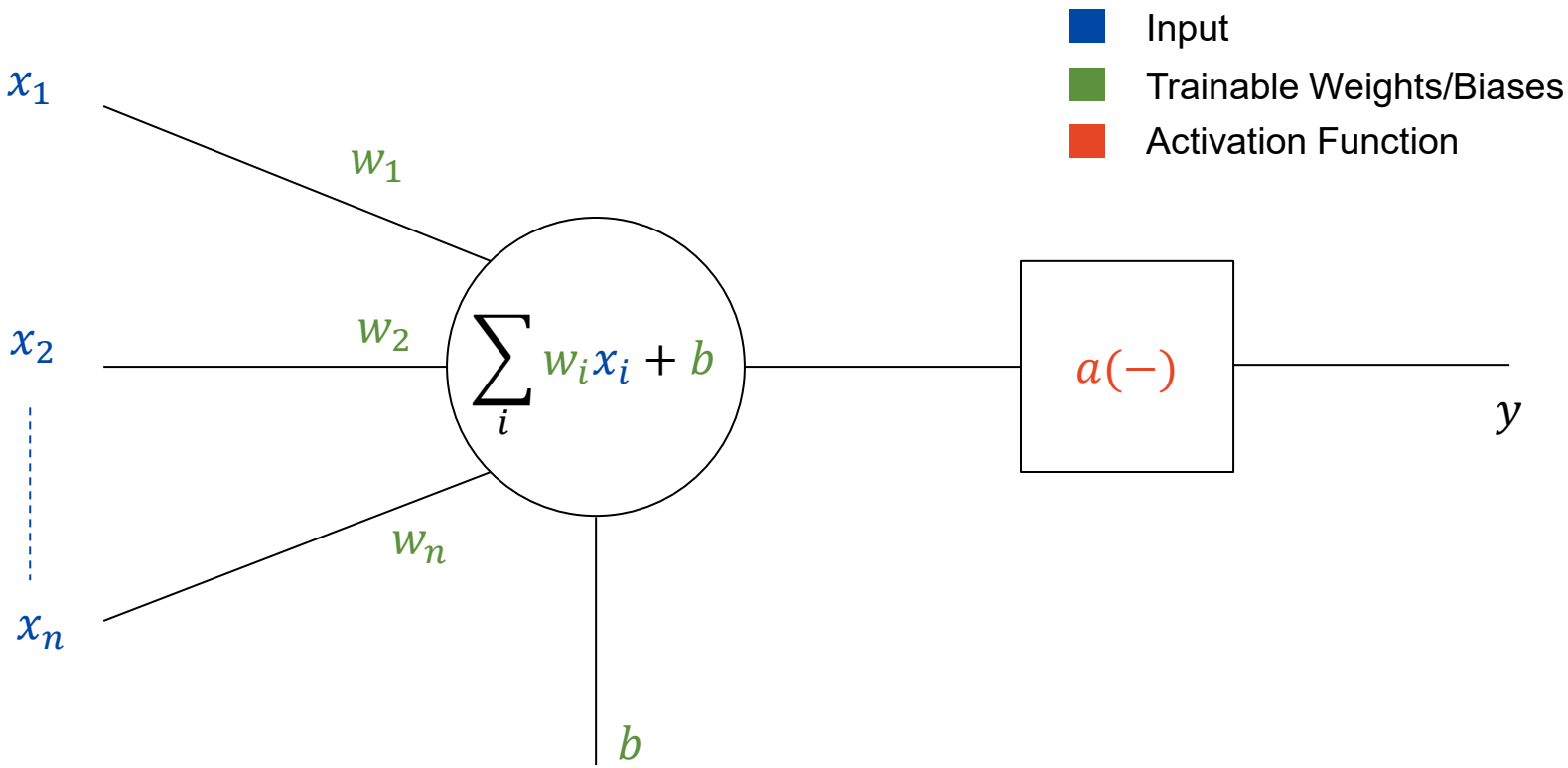
E.g. horses do not have stripes, and therefore are not zebras

We can let models decided which parts of the data are useful to accomplish a task

Neural Networks

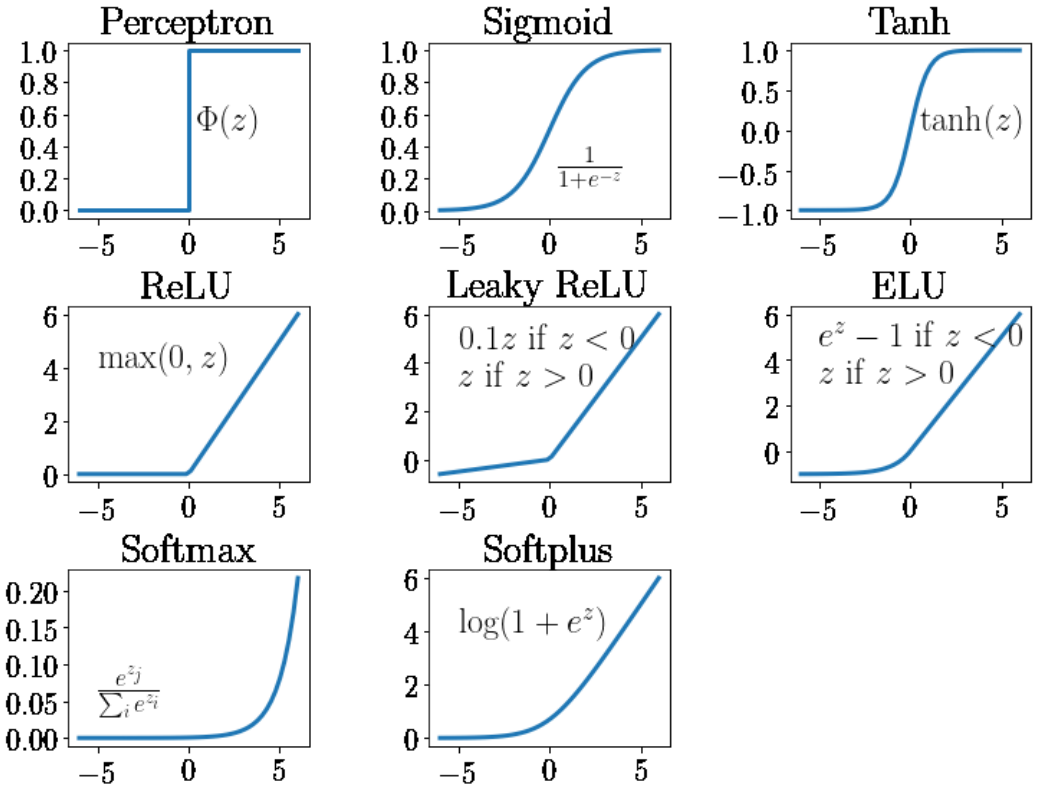


Wikimedia


Input
Trainable Weights/Biases
Activation Function
x_1
x_2
x_n
w_1
w_2
w_n
$\sum_i w_i x_i + b$
b
$a(-)$
y

Activation Functions

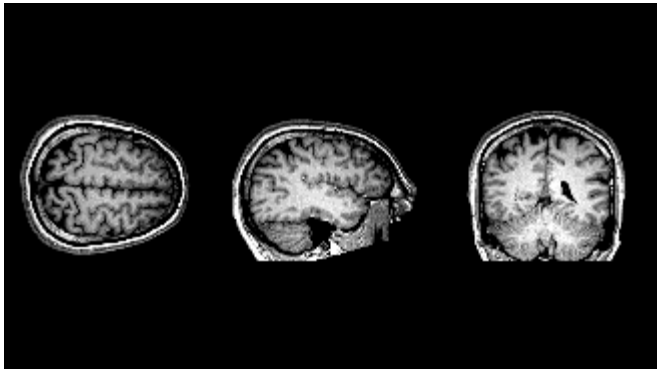
- **Remember the classifier? Not all decisions are linear.**
- **Non-linear activations give a non-linear response**

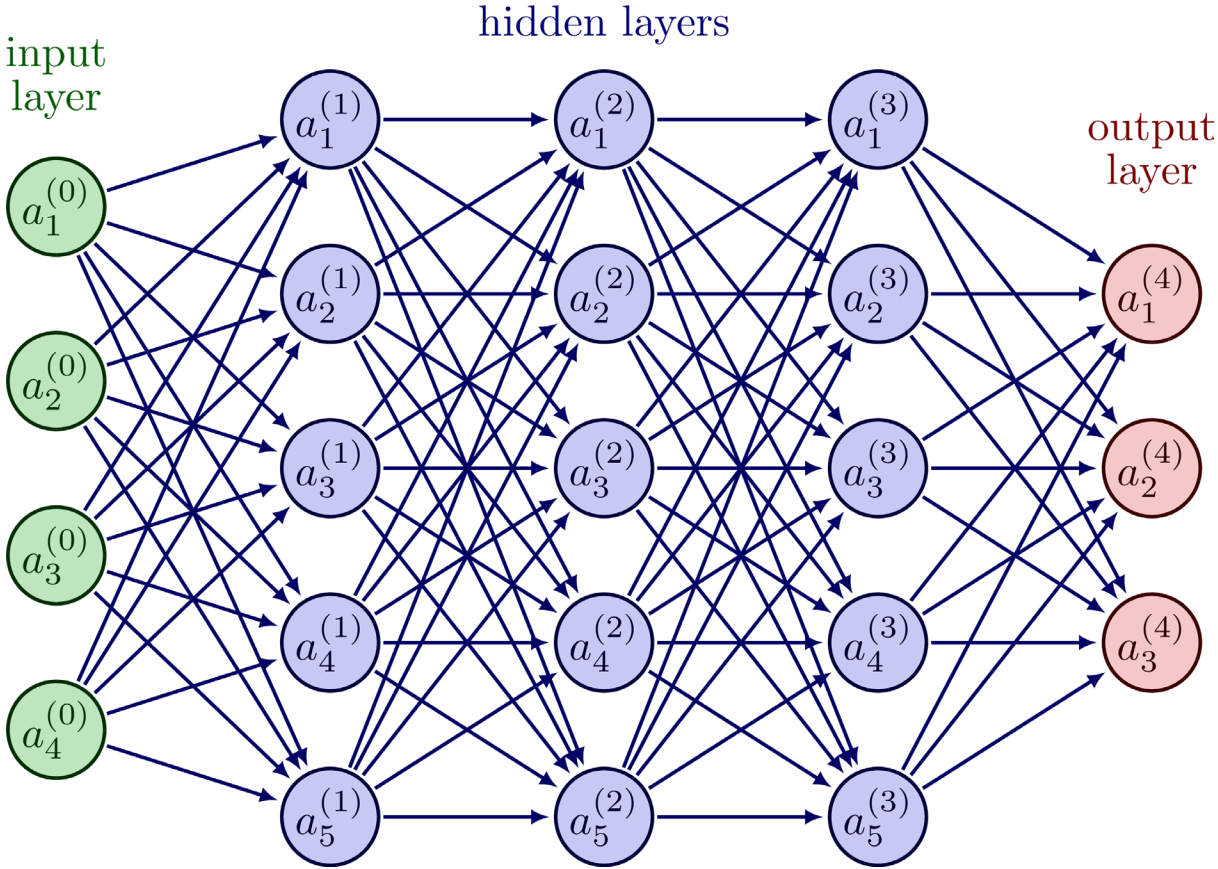


Johnson et al. 2020

More Neurons

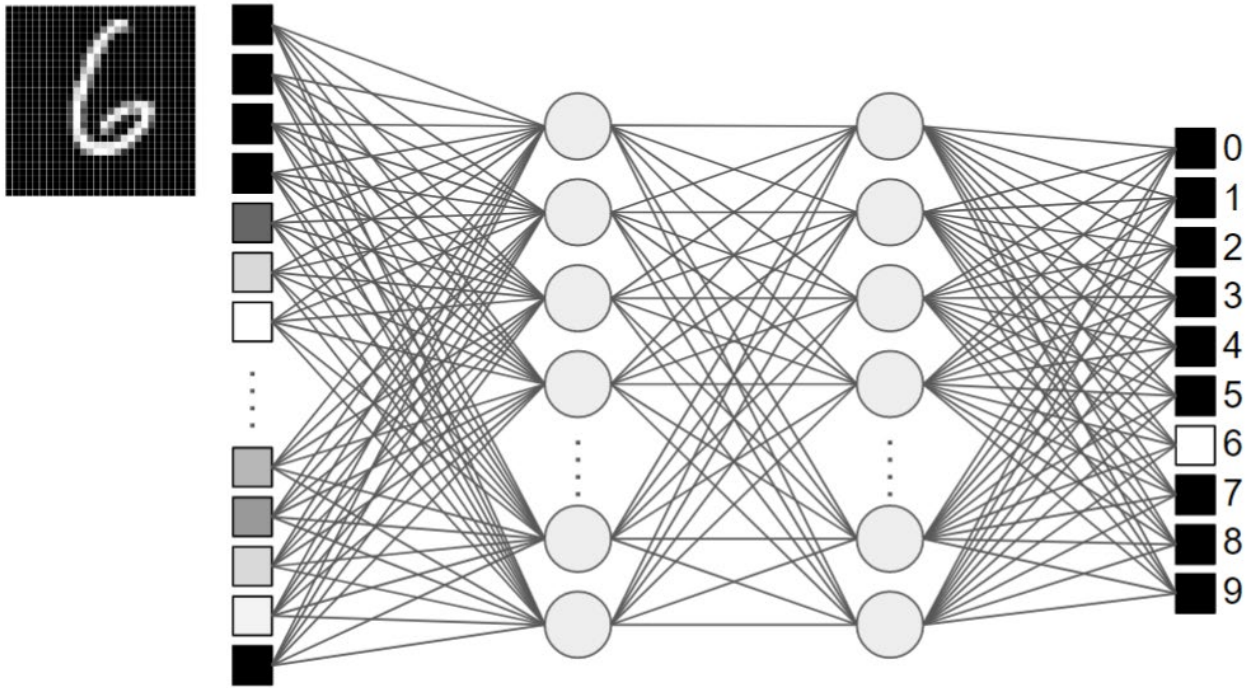
- Complex decisions require something that can approximate the function which creates them
- Single neurons suited to linear, binary problems on low-dimensional data
- We link many together to form a network, consisting of layers



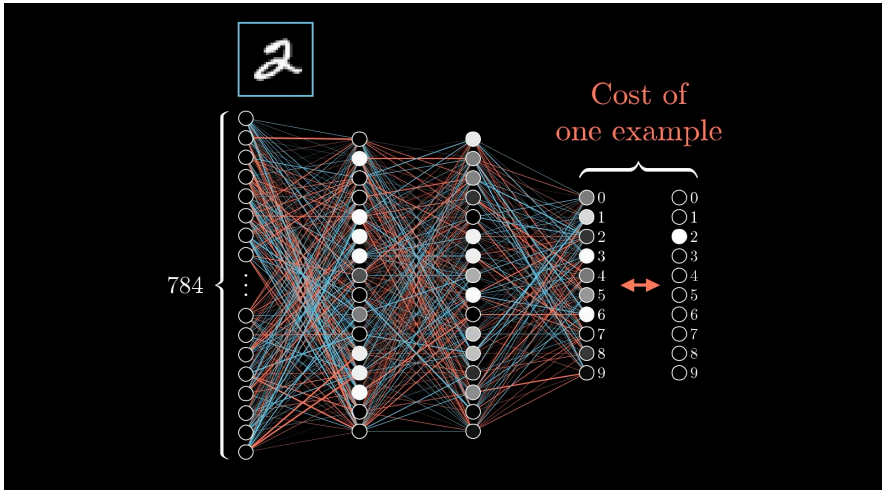

hidden layers
input layer
output layer
$a_1^{(0)}$
$a_2^{(0)}$
$a_3^{(0)}$
$a_4^{(0)}$
$a_1^{(1)}$
$a_2^{(1)}$
$a_3^{(1)}$
$a_4^{(1)}$
$a_5^{(1)}$
$a_1^{(2)}$
$a_2^{(2)}$
$a_3^{(2)}$
$a_4^{(2)}$
$a_5^{(2)}$
$a_1^{(3)}$
$a_2^{(3)}$
$a_3^{(3)}$
$a_4^{(3)}$
$a_5^{(3)}$
$a_1^{(4)}$
$a_2^{(4)}$
$a_3^{(4)}$

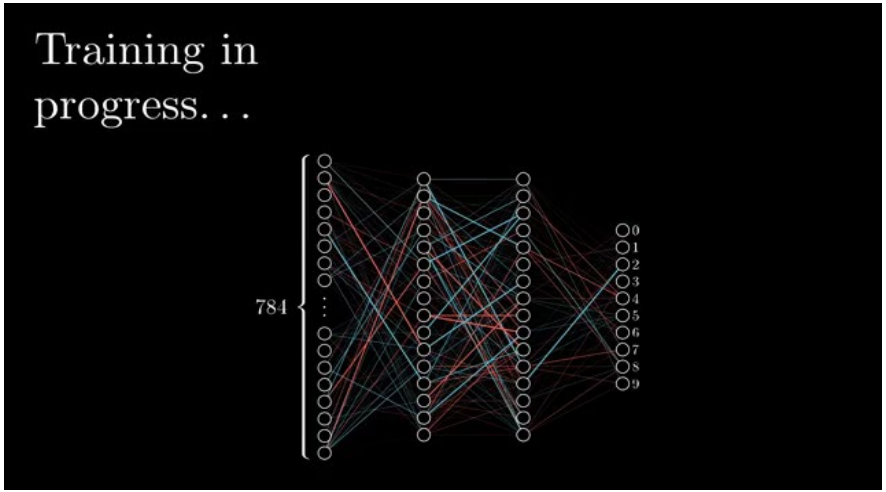
Tikz.net

Recognising Handwritten Digits (A Classifier!)



Training (3blue1brown visualisation)

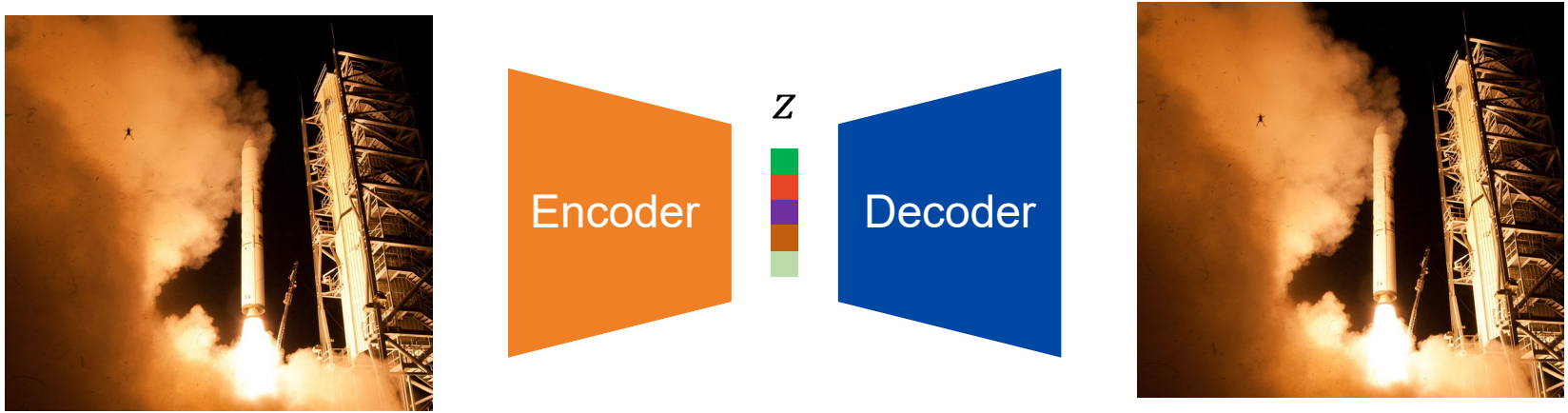




3blue1brown

Isn’t this meant to be a section on unsupervised learning?

An Unsupervised Task - Reconstruction



Auto-Encoders

No human labels - all data is its own label

Network decides the intermediary features which are important

Able to compress high dimensional data into a lower dimension

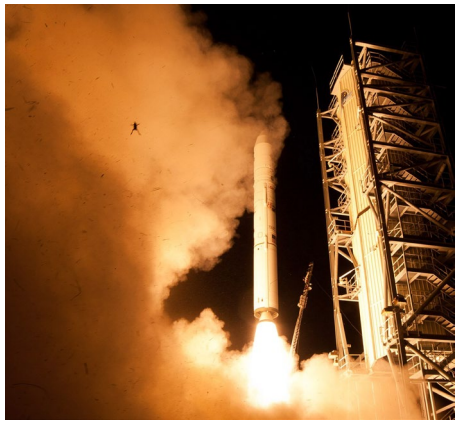
Auto-Encoding MNIST

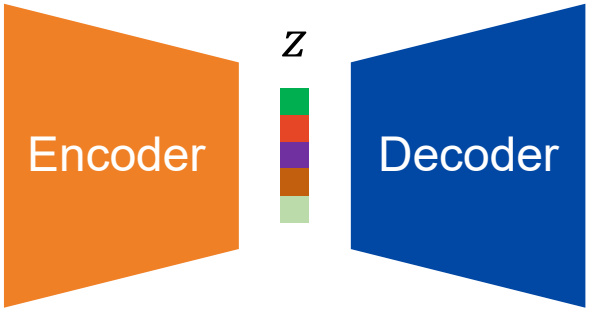
5 mins Break

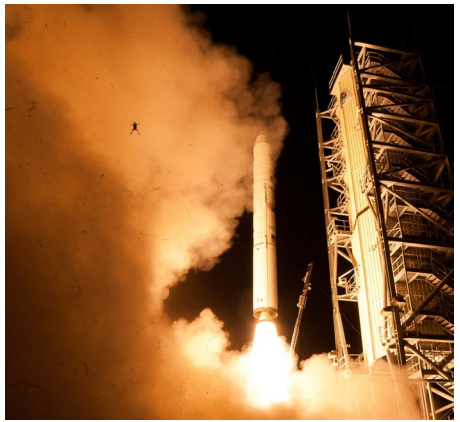
Deep Learning with Images

Turning cameras into eyes

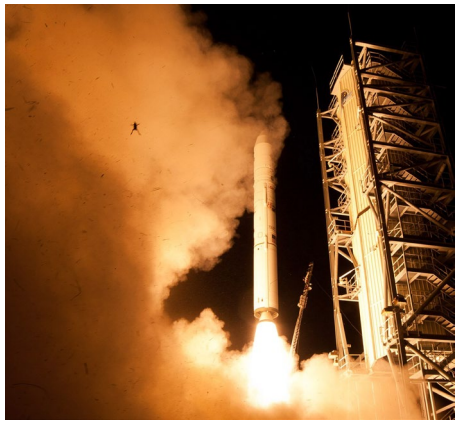
Remember those Auto-Encoders?

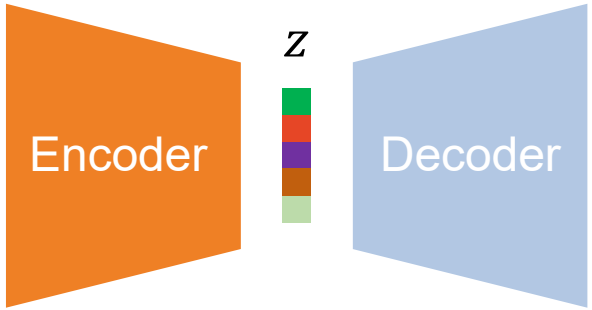


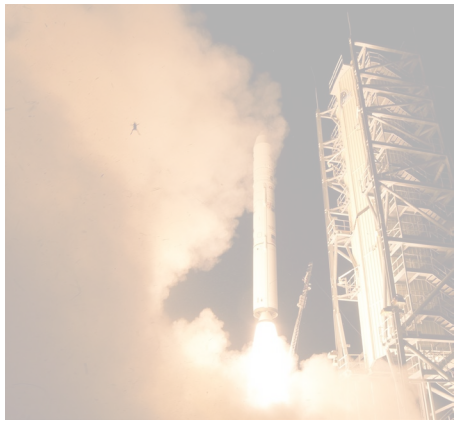




Remember those Auto-Encoders?





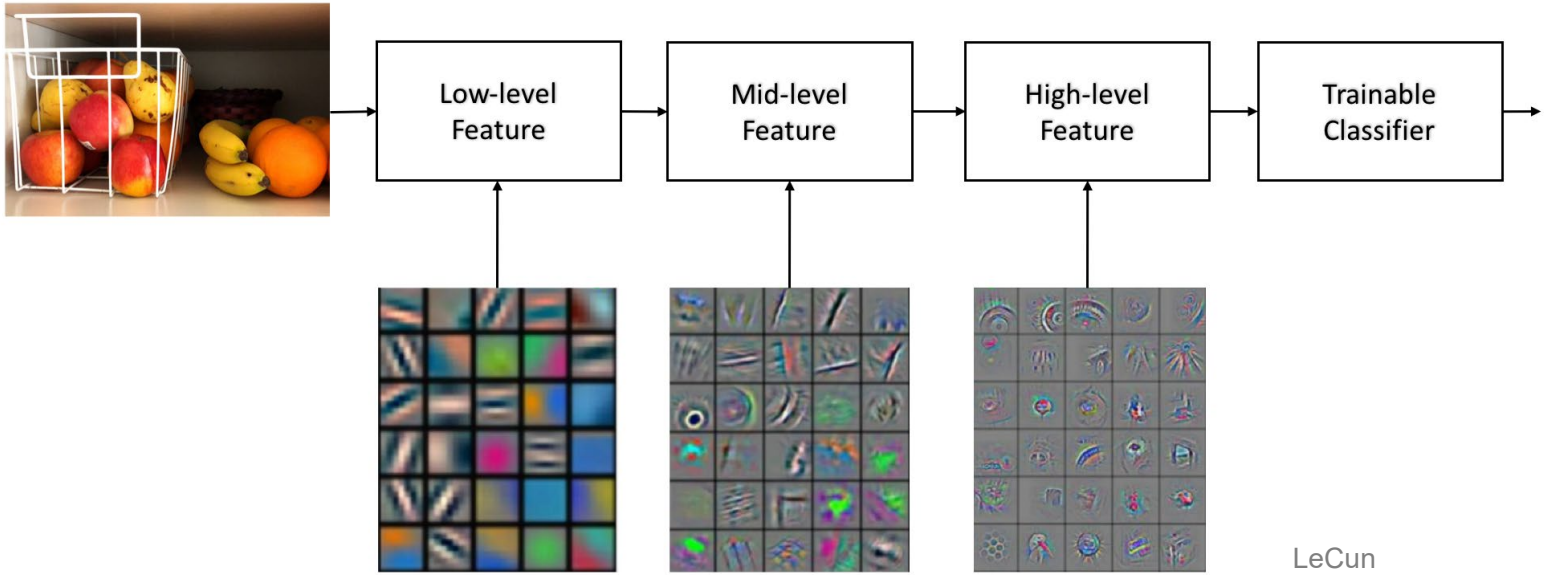


Efficiently reduces dimensions
by learning features!

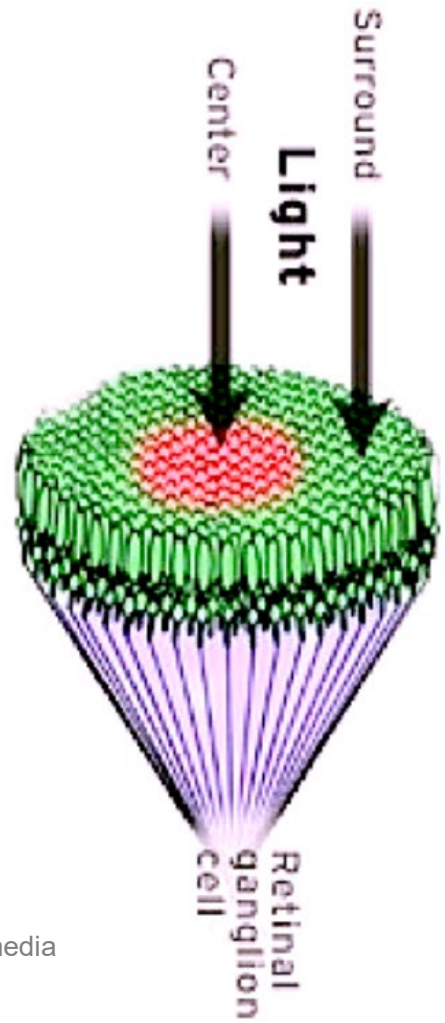
Deep Learning

- **Uses something called representation learning**

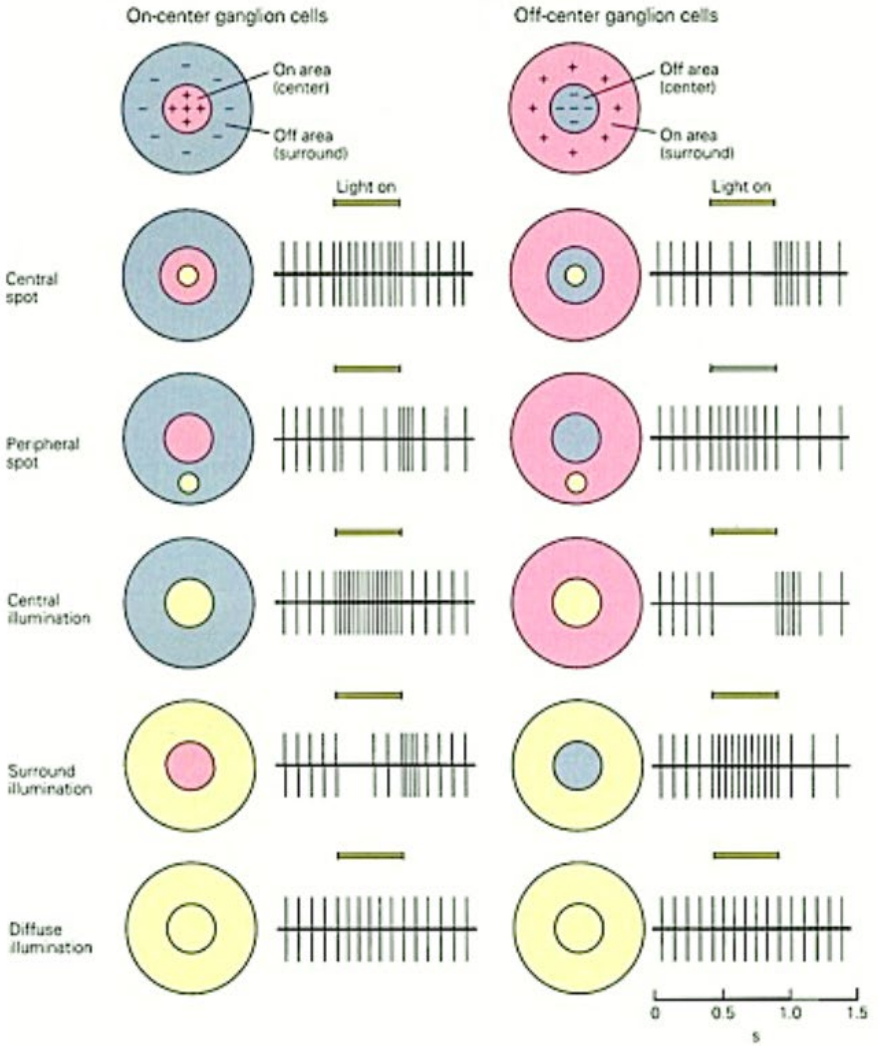
- **Learns increasingly abstract representations of data known as features**



LeCun

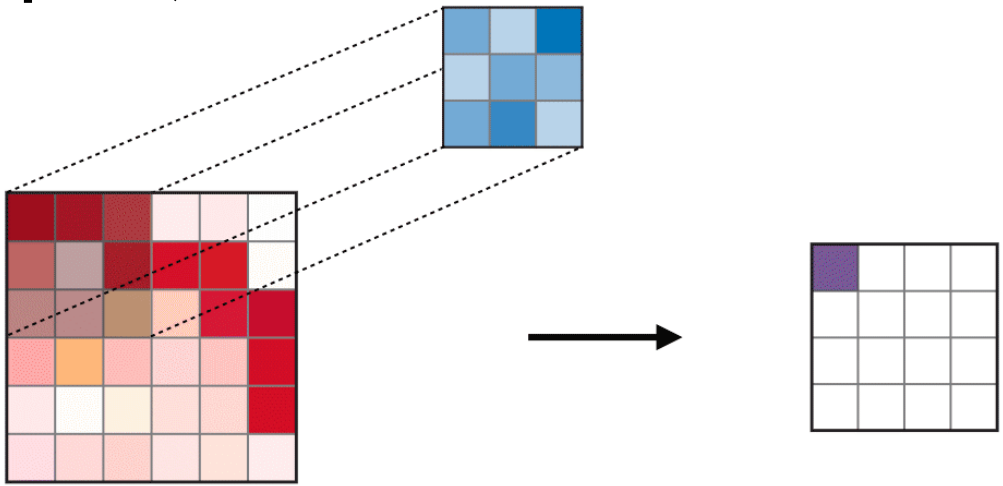

Surround
Light
Center
Retinal ganglion cell

Wikimedia


On-center ganglion cells
Off-center ganglion cells
On area (center)
Off area (surround)
Off area (center)
On area (surround)
Light on
Light on
Central spot
Peripheral spot
Central illumination
Surround illumination
Diffuse illumination
0
0.5
1.0
1.5
s

Convolutions

- The basis for most (though transformers have become very popular) modern deep learning networks
- Exploit locality of features (important stuff will not be a single point!)



Transfer Learning

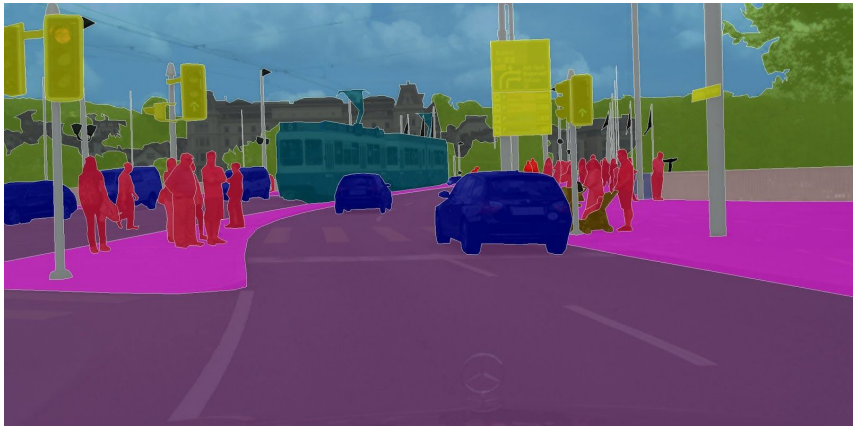
- Lots of features are the same
- Edges, textures, eyes, hands, wheels, tables etc.
- Leverage networks with lots of experience obtaining features

Koala helps a network with plane?

Semantic Segmentation

- Deep learning where each pixel is given a classification

- Allows a network to tell us which groups of pixels are together, and where they are in an image.



KITTI

Semantic Segmentation of Cats & Dogs

Large Language Models

Or: How I Learned to Stop Worrying and Love ChatGPT

Not good until recently!

- Landmark paper introduced the Transformer architecture

- Attention is fundamentally letting the network learn what inputs to base an output on

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

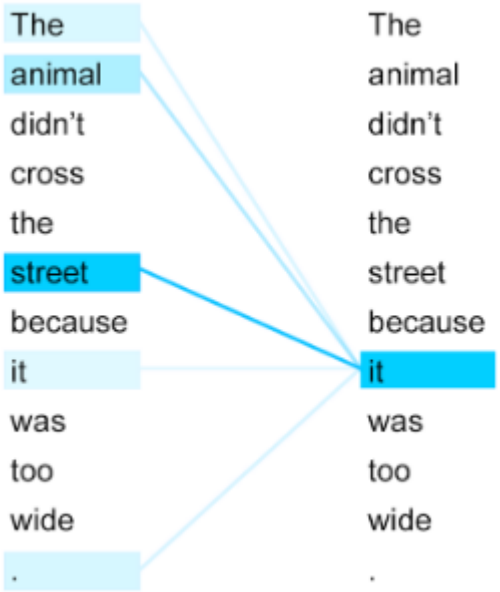
Llion Jones* Google Research llion@google.com

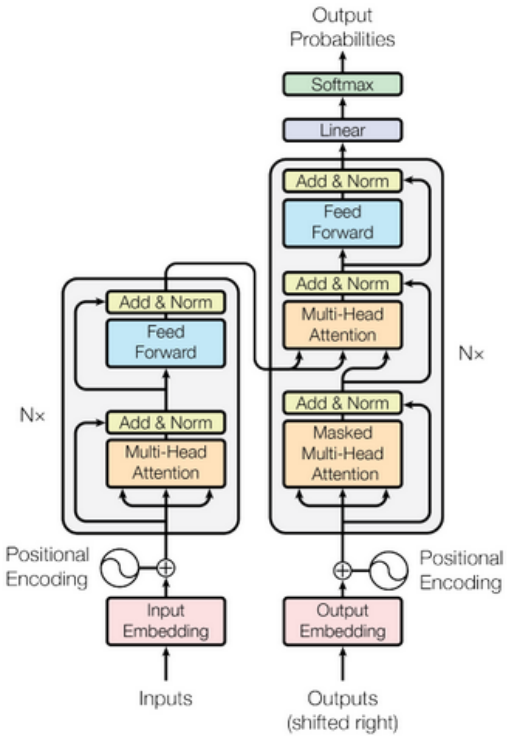
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

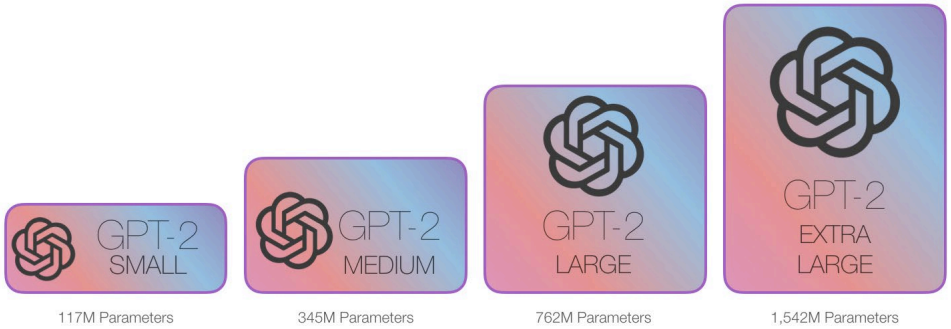
The Transformer Architecture

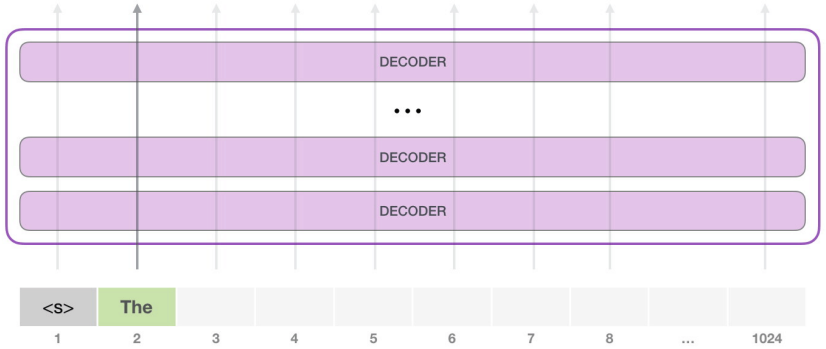




Viswani et al.

GPT2


GPT-2
SMALL
117M Parameters
GPT-2
MEDIUM
345M Parameters
GPT-2
LARGE
762M Parameters
GPT-2
EXTRA
LARGE
1,542M Parameters


DECODER
...
DECODER
DECODER
<s>
The
1
2
3
4
5
6
7
8
...
1024

Fine-Tuning

- Similar in a way to transfer learning
- Start with something that has been trained well on everything, then show it specific data at the end
- The breadth of training should prevent overfitting
- Good to LLM to mimic a certain style of writing

Smithsonian

Let's generate some Honi Soit extracts

Here's one I prepared earlier…

The public relations disaster that has befallen academia is not just a setback for the profession. It is also a serious cause for concern for the future of the humanities as a whole. Primarily because of the pervasive nature of the pandemic, many university degrees have been rendered pointless for many students. "The university should be a place where students make decisions for themselves," said USyd's Simon Rice. "It should be a place where people make decisions for themselves," an Australian higher education activist I spoke with echoed. "It shouldn't be a place where students are put through the rigours of a higher education process, where they are given a choice between a miserable future and an appalling one."

Ethical AI and Safeguarding Users

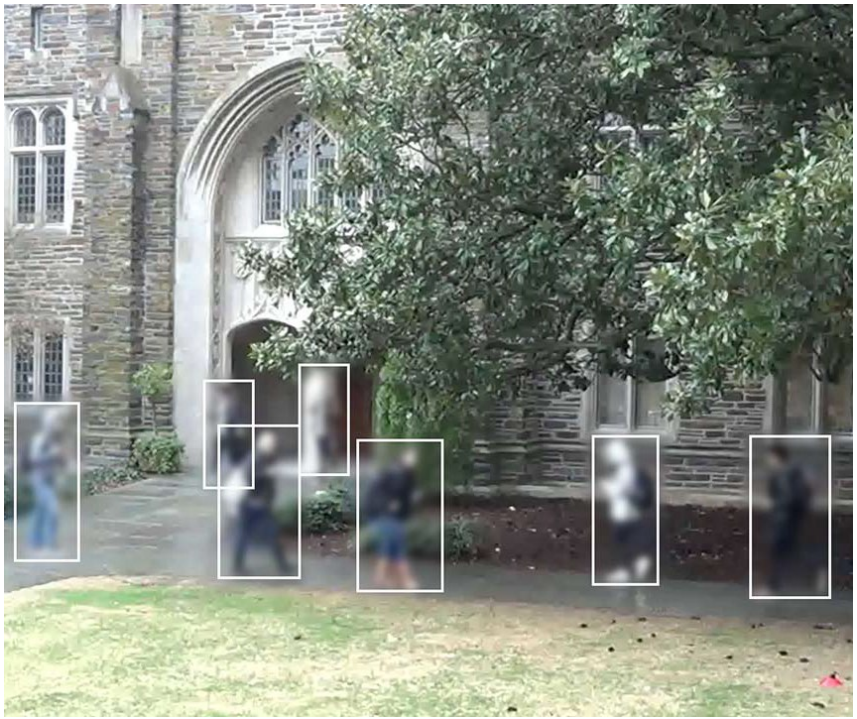
Why you don’t have to worry quite yet…

AI Researchers Now Have Ethical Principles

- **These are taken from the ICLR Code of Ethics**
  - Contribute to Society and to Human Well-being
  - Uphold High Standards of Scientific Excellence
  - Avoid Harm
  - Be Honest, Trustworthy and Transparent
  - Be Fair and Take Action not to Discriminate
  - Respect the Work Required to Produce New Ideas and Artefacts
  - Respect Privacy
  - Honour Confidentiality

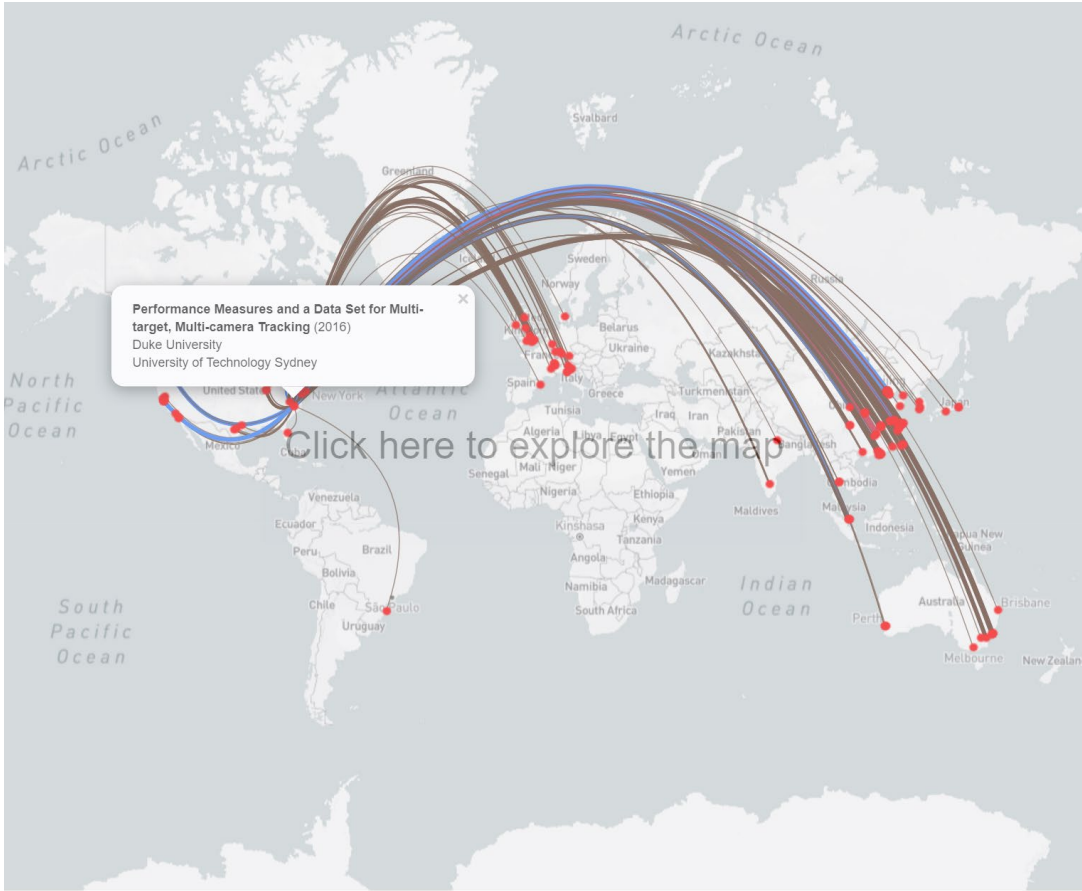
Where this hasn't been done…

- Duke MTMC (Multi-Target, Multi-Camera) was a dataset of surveillance footage
- 14hours @ 1080p, 60fps of 2000 students on 8 cameras
- Designed for individual recognition, tracking and re-identification
- Now no longer publicly available.



Where was this used?

– **These are just where there's citations!**



Academic Commercial Military / Government

Citation data is collected using SemanticScholar.org then dataset usage verified and geolocated. Citations are used to provide an estimated overview of how and where images were used based on institutional affiliations. Thicker lines represent more citations. Please zoom in to see all institutions, as cities may have multiple points very close together.

Representative Data

- Models are only as good as their data
- If the data is not representative, the models are not representative

Facial recognition

How white engineers built racist code – and why it's dangerous for black people

As facial recognition tools play a bigger role in fighting crime, inbuilt racial biases raise troubling questions about the systems that create them

Ali Breland

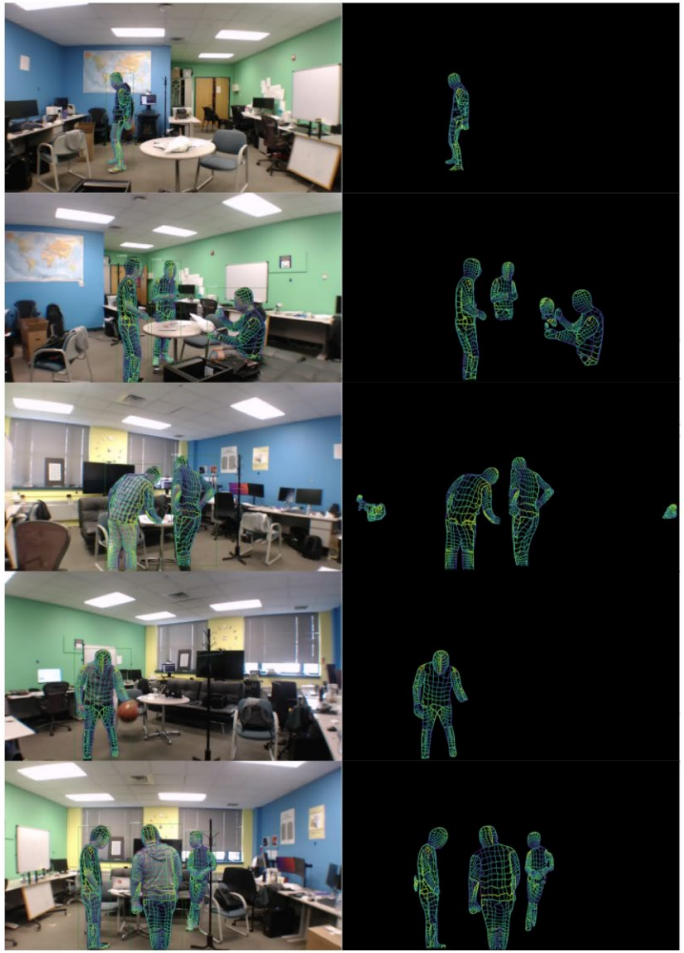
Mon 4 Dec 2017 20.00 AEDT

A protest over police violence against black communities. Photograph: Alamy Stock Photo

Generative Models + Deepfakes

DALL-E

MIT



Qualitative comparison using synchronized images and WiFi signals. (Left Column) image-based DensePose (Right our WiFi-based DensePose.

DensePose From WiFi

Jiaqi Geng
jiaqigen@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Dong Huang
donghuang@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Fernando De la Torre
ftorre@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Advances in computer vision and machine learning techniques have led to significant development in 2D and 3D human pose estimation from RGB cameras, LiDAR, and radars. However, human pose estimation from images is adversely affected by occlusion and lighting, which are common in many scenarios of interest. Radar and LiDAR technologies, on the other hand, need specialized hardware that is expensive and power-intensive. Furthermore, placing these sensors in non-public areas raises significant privacy concerns.

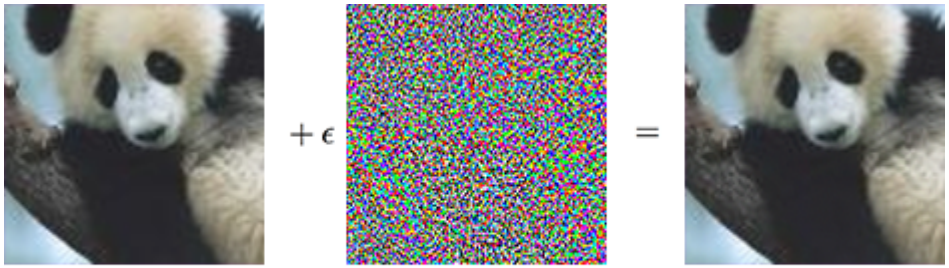
To address these limitations, recent research has explored the use of WiFi antennas (1D sensors) for body segmentation and key-point body detection. This paper further expands on the use of the WiFi signal in combination with deep learning architectures, commonly used in computer vision, to estimate dense human pose correspondence. We developed a deep neural network that maps the phase and amplitude of WiFi signals to UV coordinates within 24 human regions. The results of the study reveal that our model can estimate the dense pose of multiple subjects, with comparable performance to image-based approaches, by utilizing WiFi signals as the only input. This paves the way for low-cost, broadly accessible, and privacy-preserving algorithms for human sensing.

Adversarial Attacks

- An emergent field: protecting users of machine learning from bad actors

- We can prevent this by designing our networks carefully (Lipschitz bounded NNs for example, see papers by my colleague Patricia Pauli @ Uni. Of Stuttgart)

OpenAI



Eykholt et al. 2018

The Answer Might Lie in the Data

Learning Privacy-preserving Optics for Human Pose Estimation

Carlos Hinojosa[1], Juan Carlos Niebles[2], Henry Arguello[1]
[1]Universidad Industrial de Santander [2]Stanford University
https://carloshinojosa.me/project/privacy-hpe/

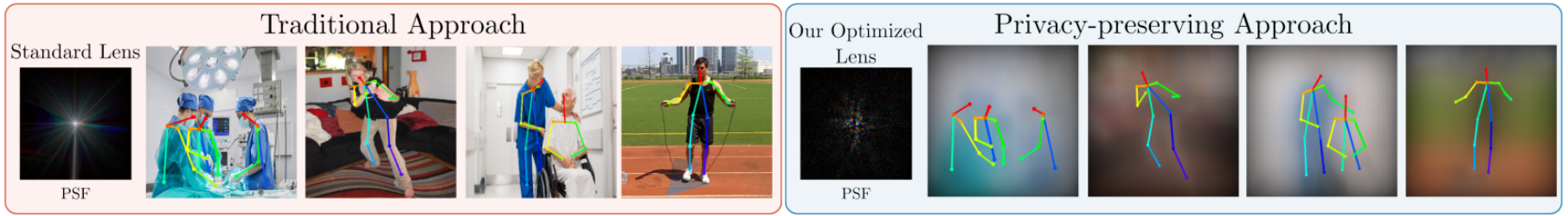
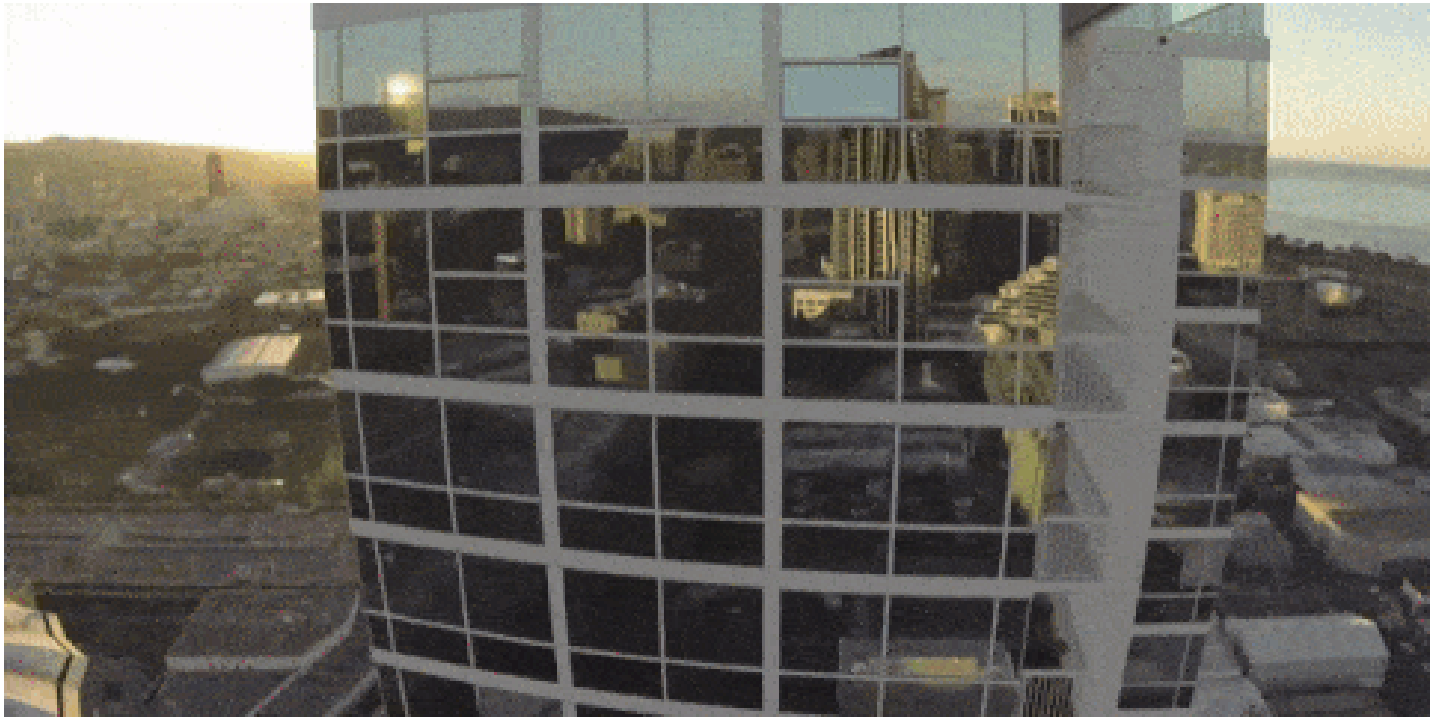


Figure 1: Standard cameras acquire visual details from the scene that could lead to privacy issues. In this work, we propose to learn privacy-preserving optics to perform human pose estimation (HPE). Our optimized lens incorporates several optical aberrations that degrade the image to hide private visual details while it still captures enough visual information to perform human pose estimation.

Killer Robots? Your Job?



Jack Naylor

Email: jack.naylor@sydney.edu.au
Website: nackjaylor.github.io
Twitter: @nackjaylor

THE UNIVERSITY OF SYDNEY

ACFR
AUSTRALIAN CENTRE FOR FIELD ROBOTICS