Explainable AI

Saja Tawalbeh

UAntwerp imec-IDLab

IDLab, imec research group at Ghent University and University of Antwerp

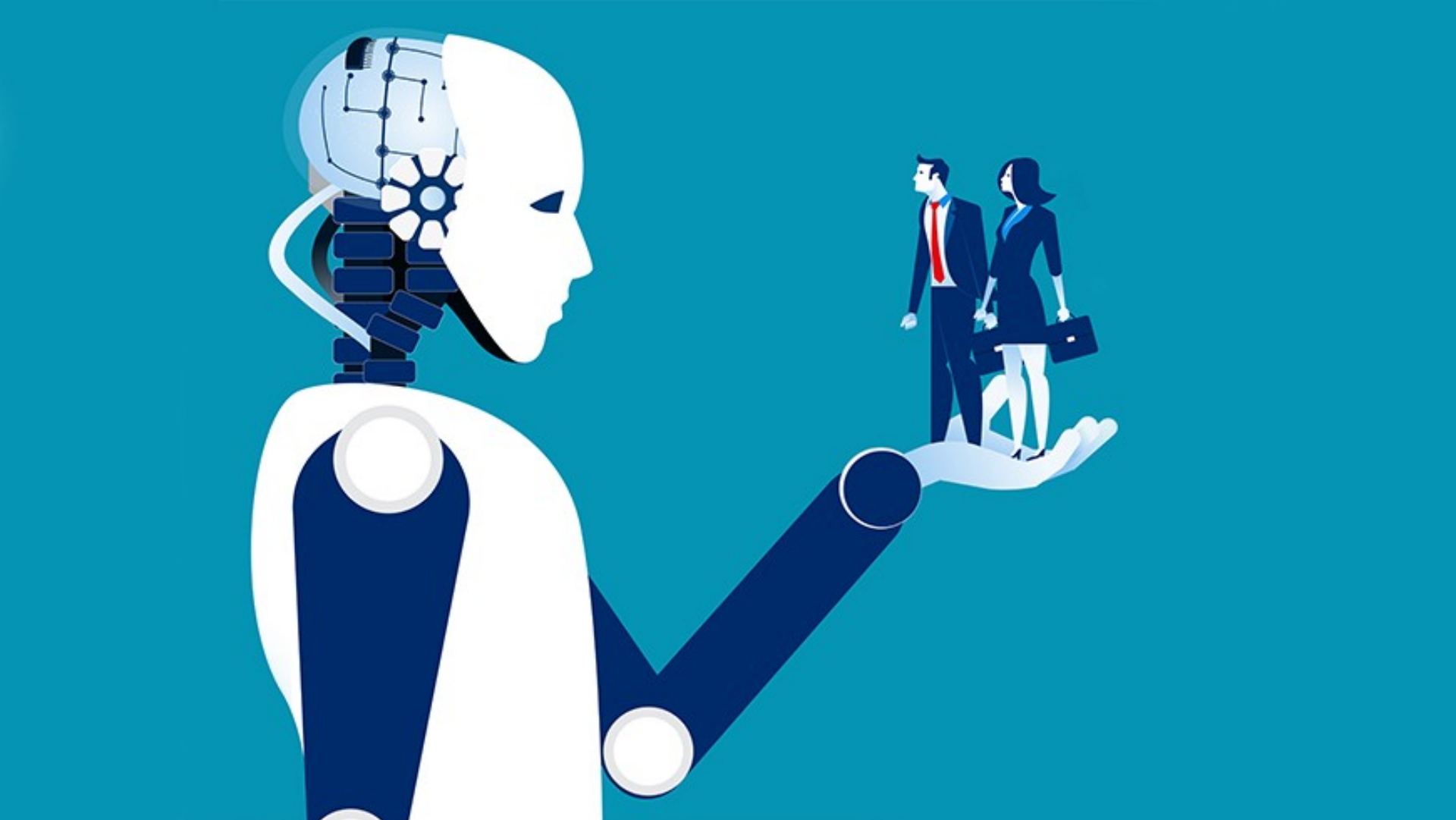
Personal Context

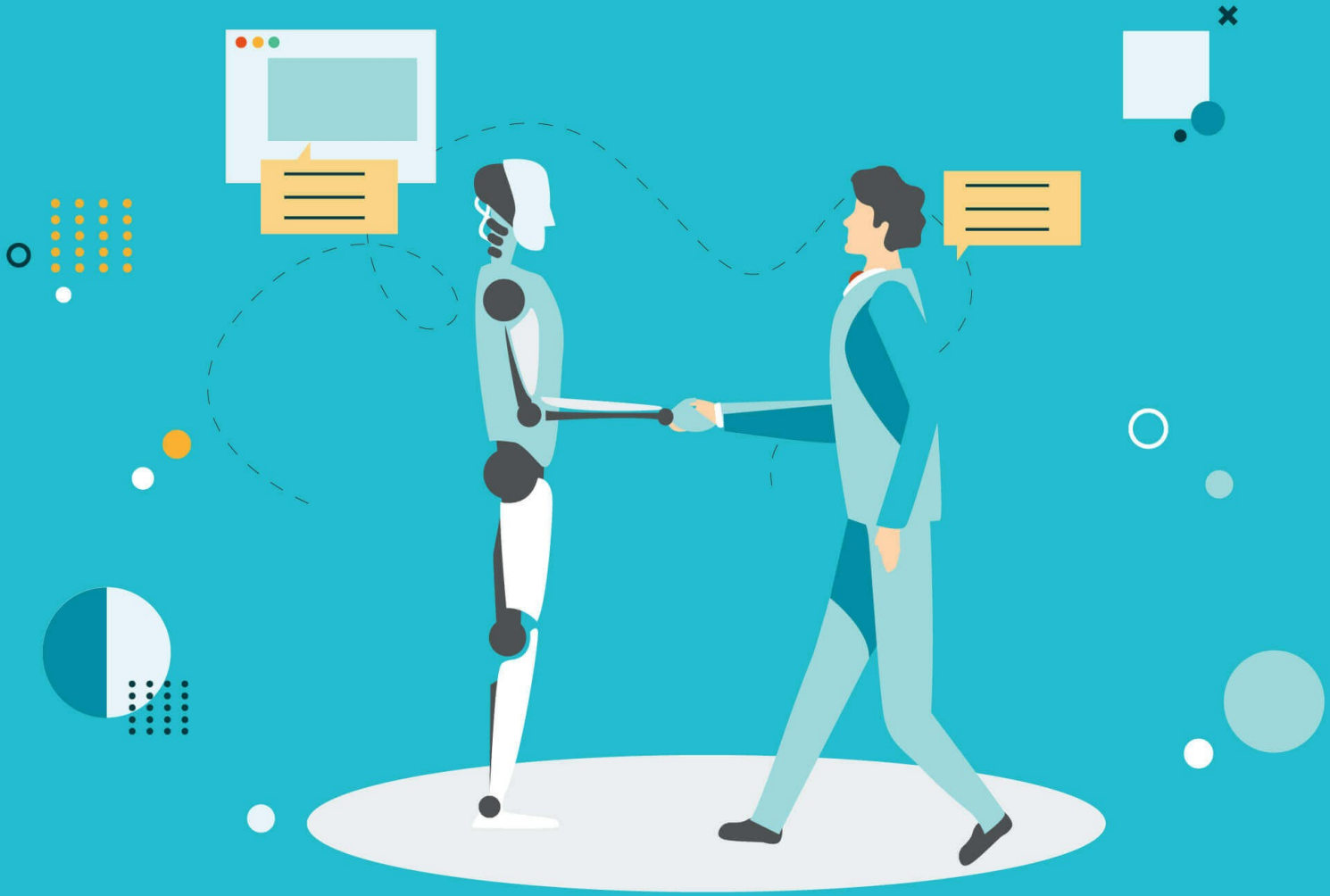
Research Lab

A unique research infrastructure used in numerous national and international collaborations

Research Team

The only way to do great work..
Is to love what YOU do..





Artificial Intelligence

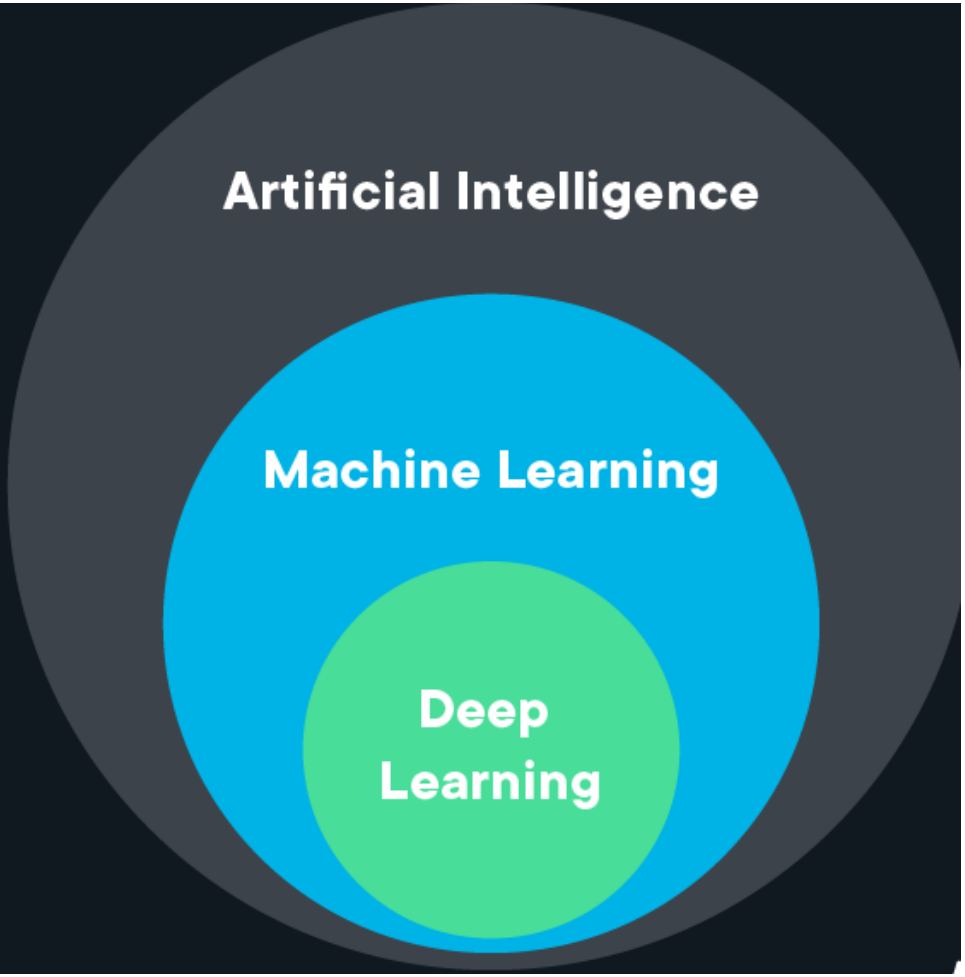
A science devoted to making machines think and act like humans.

Machine Learning

Focuses on enabling computers to perform tasks without explicit programming.

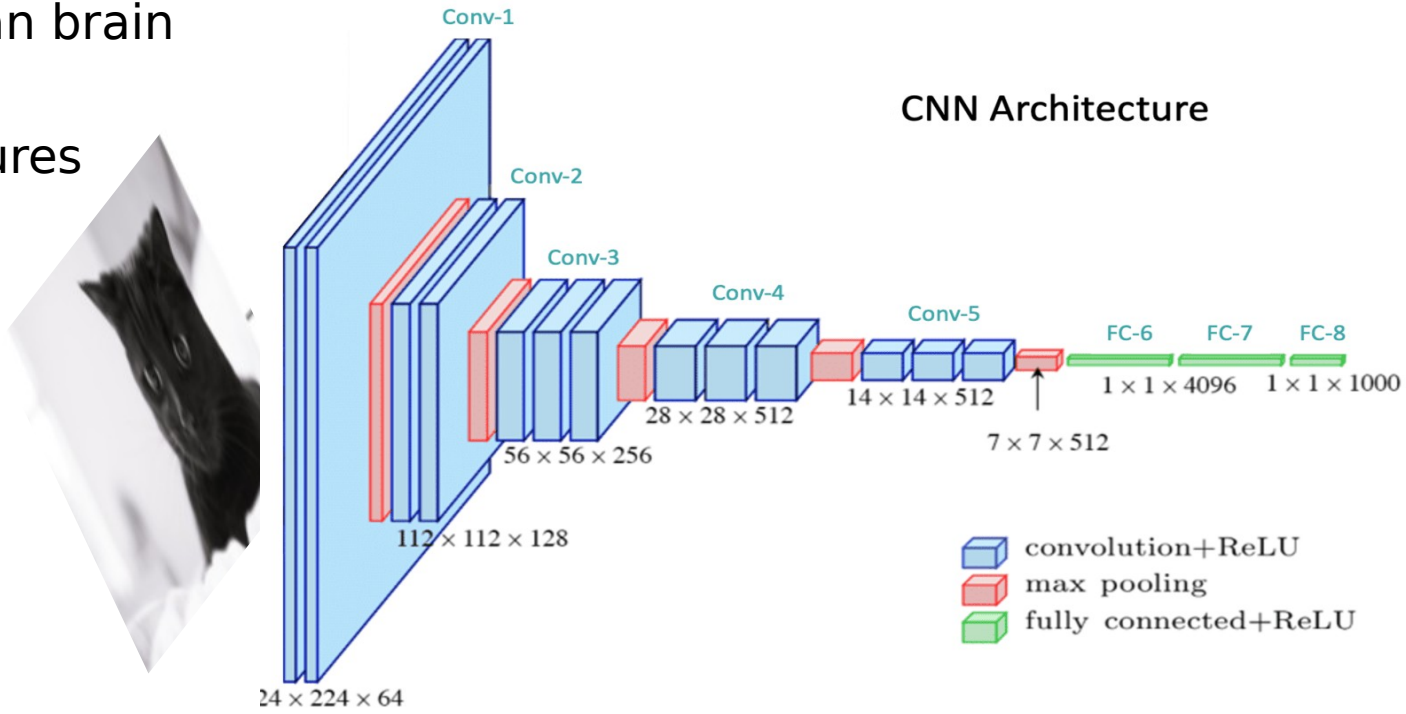
Deep Learning

A subset of machine learning based on artificial neural networks.



Convolutional Neural Networks (CNNs)

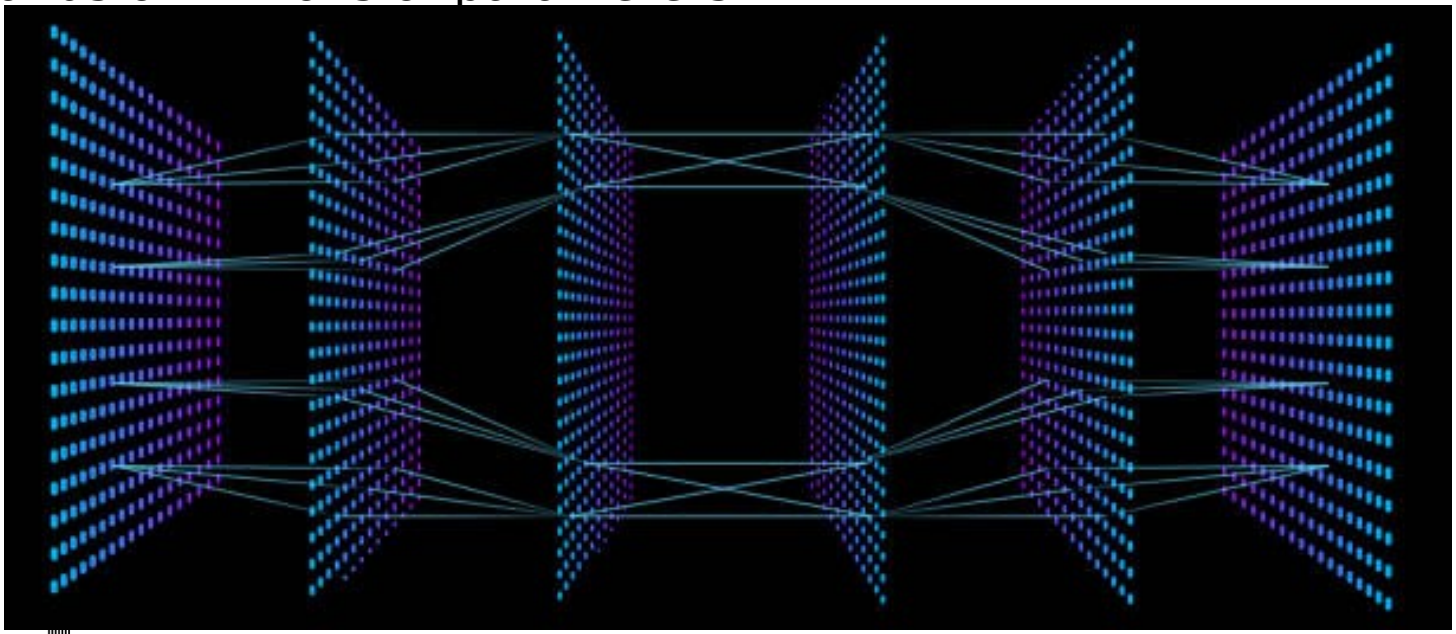
- Similar to human brain
- Deep Architectures
  - Layers



Convolutional Neural Networks (CNNs)

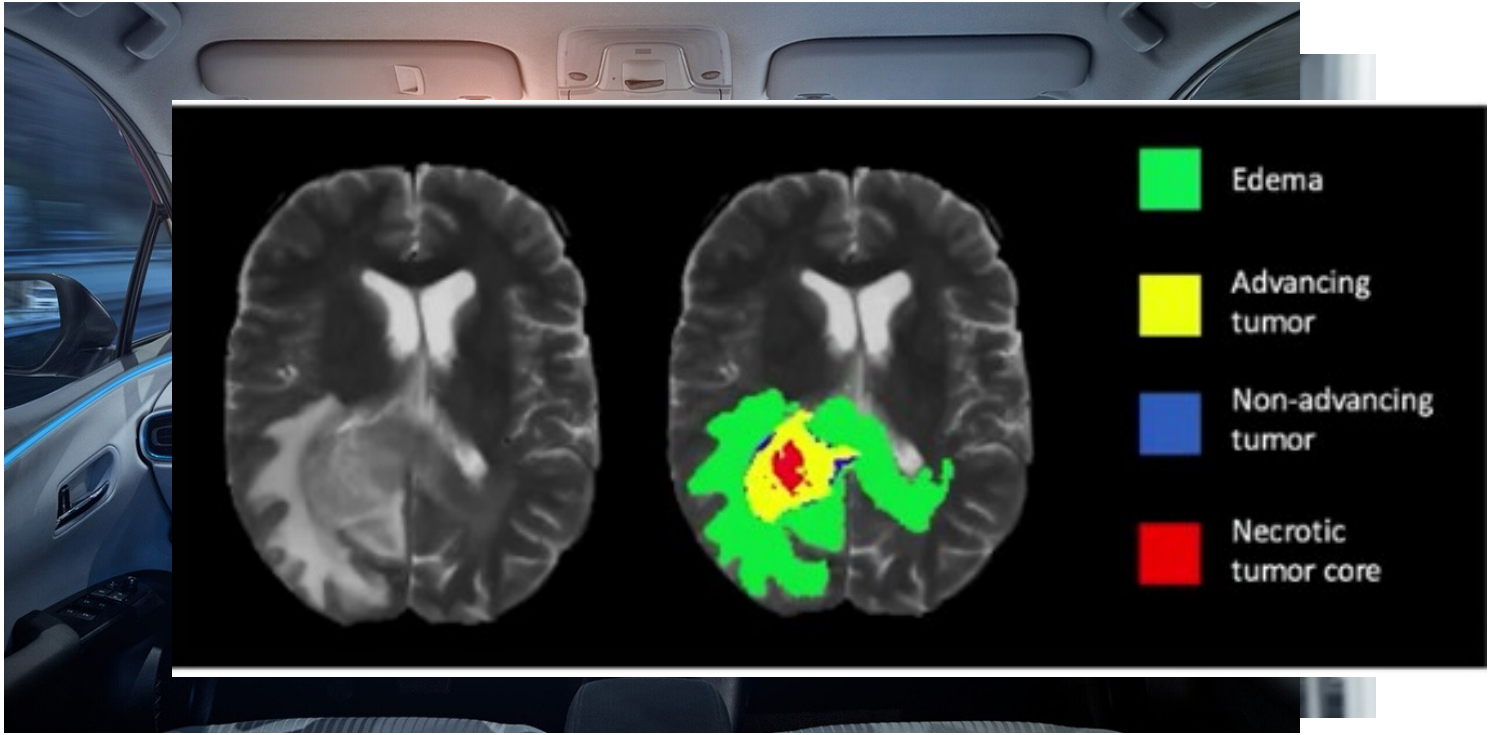
- Deep Architectures

- Thousands or millions of parameters

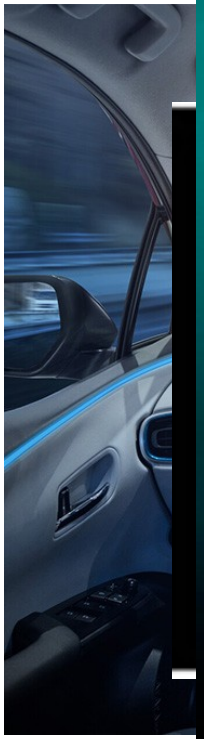


Artificial Intelligence Examples

Artificial Intelligence Examples



Artificial Intelligence Examples



Input

Input Image

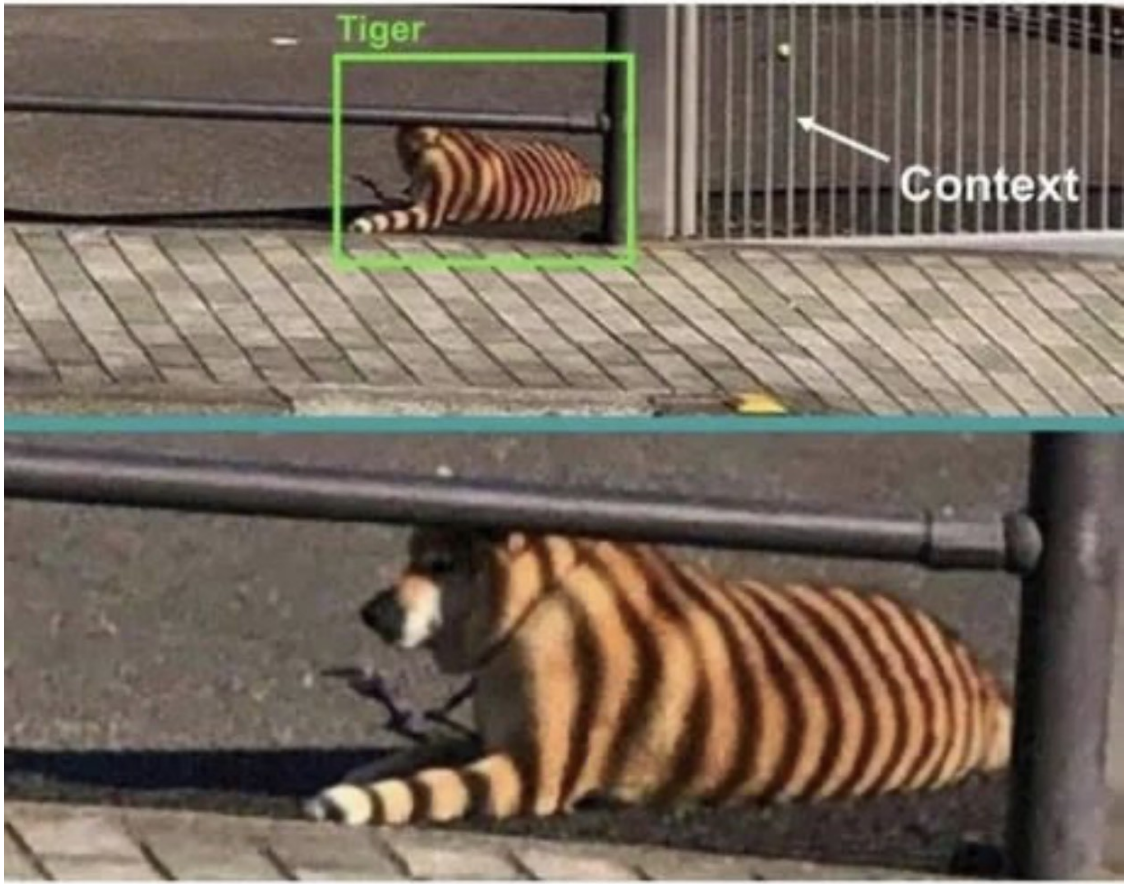
CNNs

Output

Tiger – 97 %

Dog – 2%

AI Will takeover the world



But if **CNNs** have high performance...

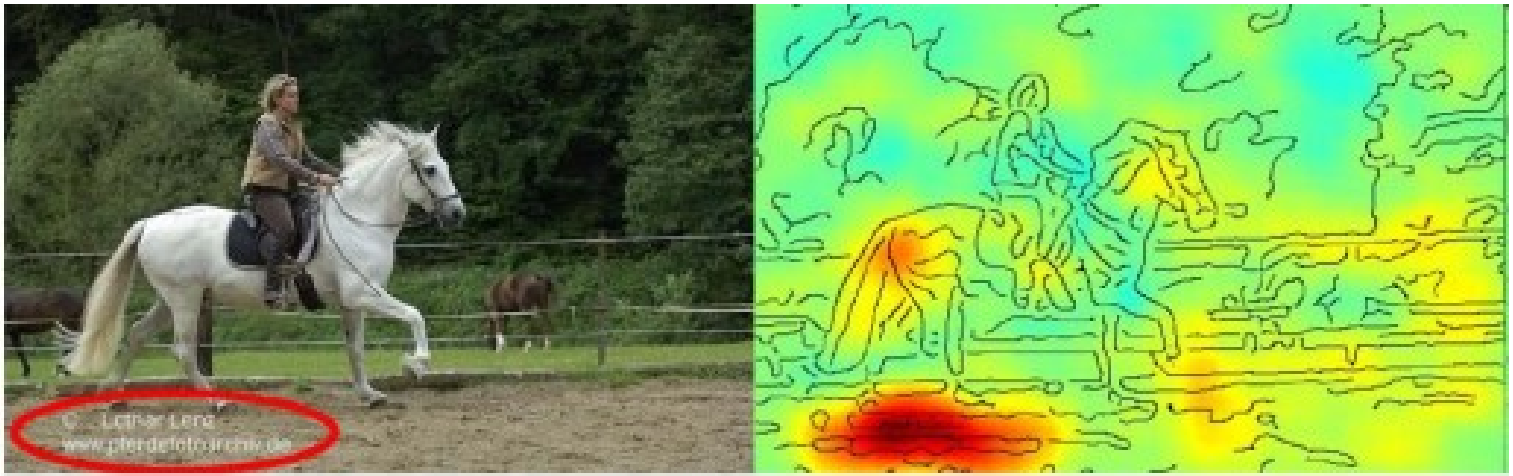
Why is this desirable?

Explaining CNNs Architectures

- Motivation
  - Detection of undesirable properties in the model

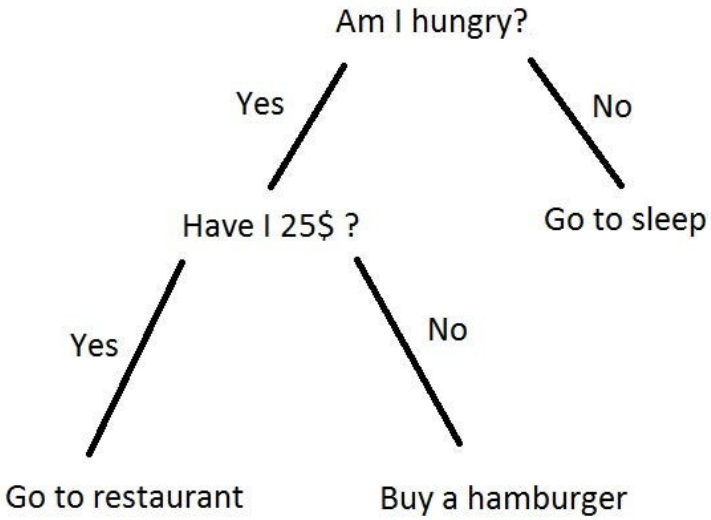
Explaining CNNs Architectures

- Motivation
  - Detection of undesirable properties in the model
  - Horse (80%)



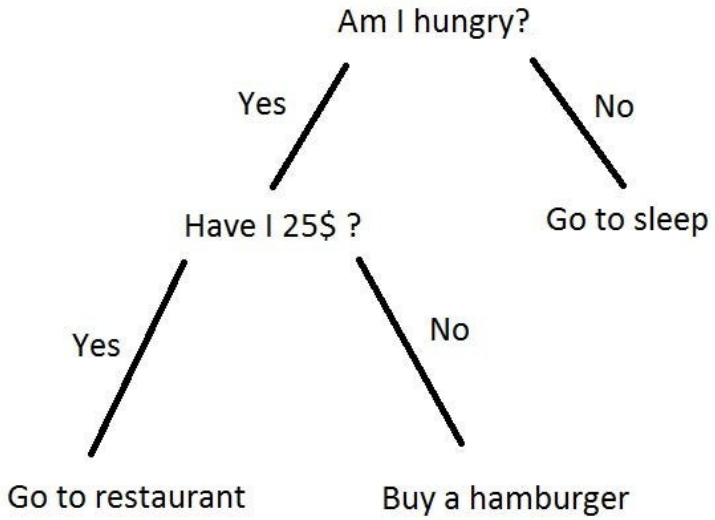
Explainable Artificial Intelligence (XAI) Categories

White Box

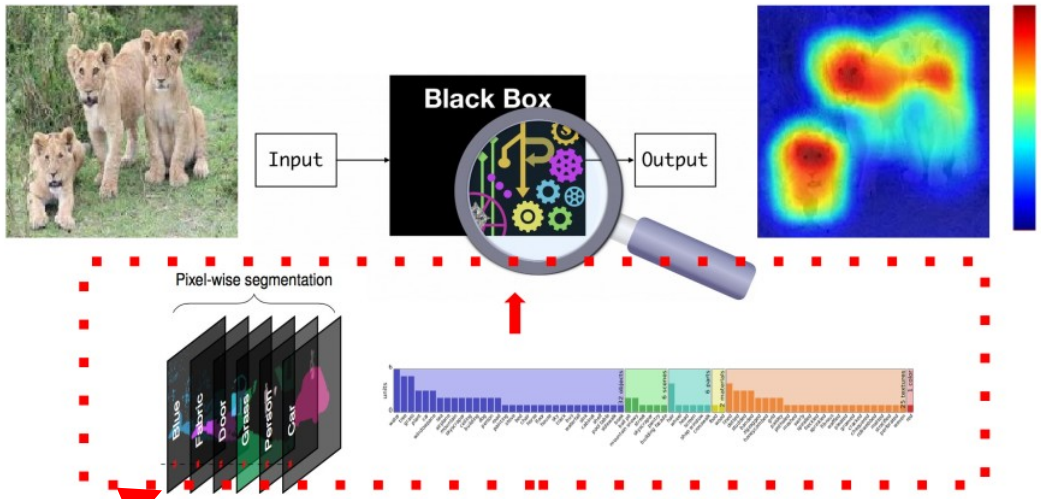


Explainable Artificial Intelligence (XAI) Categories

White Box



Black Box (POST-HOC)



Explainable Artificial Intelligence (XAI) Categories

White Box

- Transparent models
- Human understandable representations
  - If-else conditions
- Machine learning algorithms
  - Decision trees

Explainable Artificial Intelligence (XAI) Categories

White Box

- Transparent models
- Human understandable representations
  - If-else conditions
- Machine learning algorithms
  - Decision trees

Black Box (POST-HOC)

- Complex internal structure
- Lack of transparency
- Convolutional Neural Networks
  - Class Activation Mapping family methods (CAM)

Research Questions

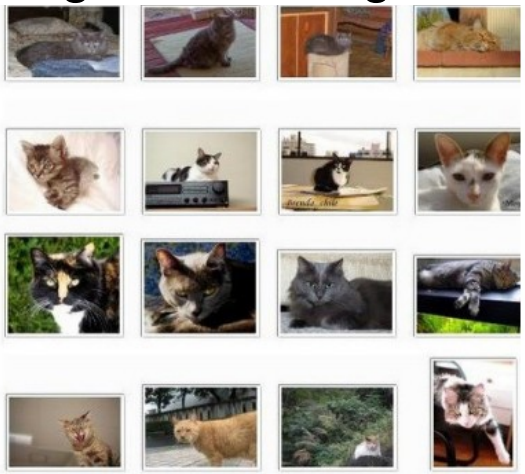
- Q1: What the model has actually learned? → Interpretation

Oramas, et. al. "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks." (2019).

Research Questions

- Q1: What the model has actually learned? → Interpretation

Original Training Data



Relevant Features

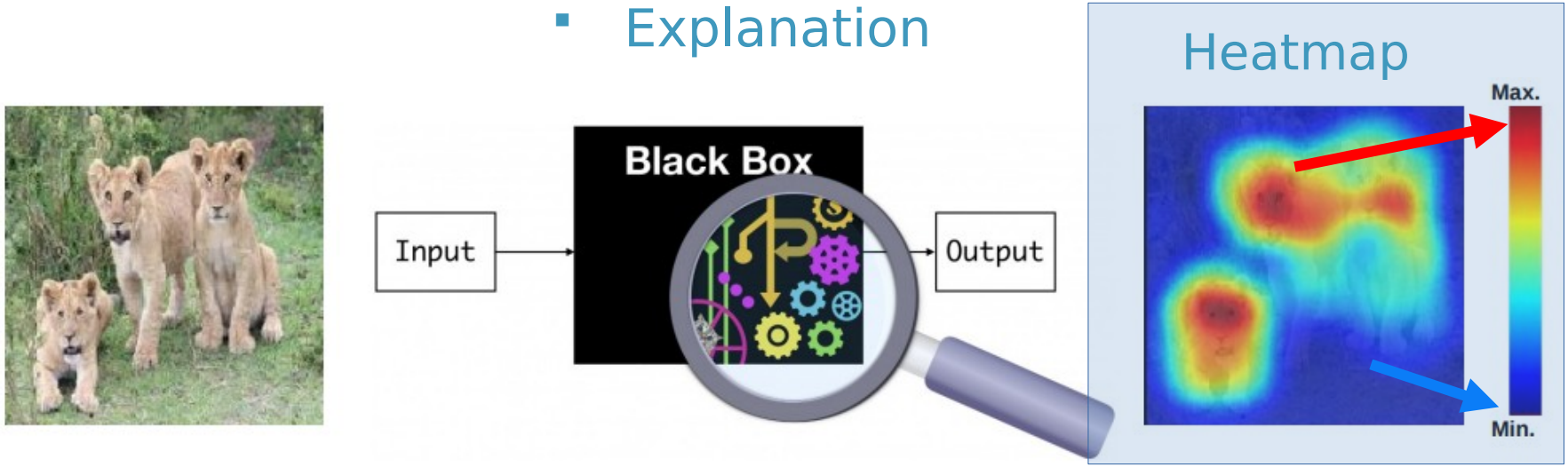
Oramas, et. al. "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks." (2019).

Research Questions

- Q2: What information from the input the model is using to make predictions? →
  - Explanation

Research Questions

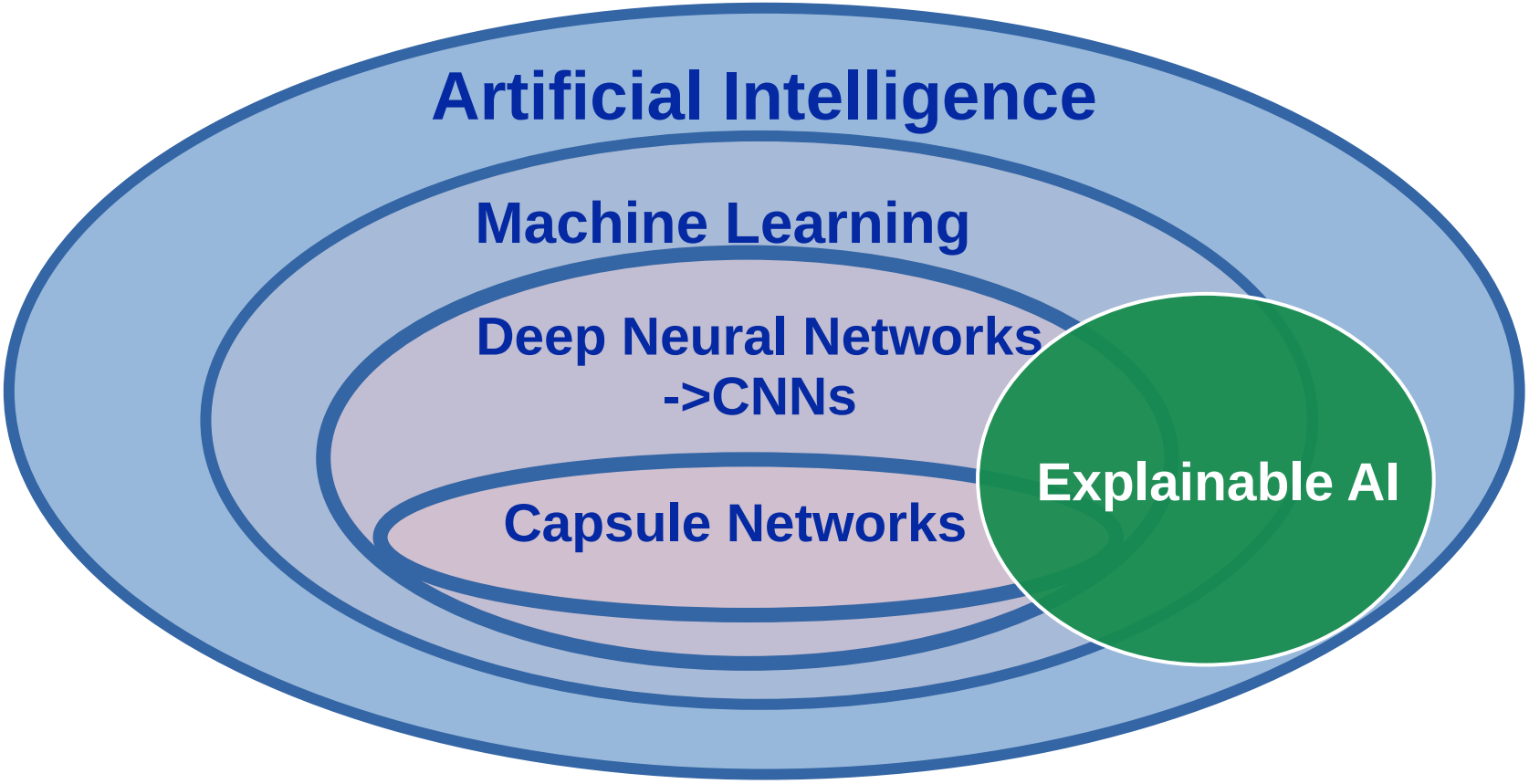
- Q2: What information from the input the model is using to make predictions? →
  - Explanation



- Higher heatmap values indicate higher influence in the prediction

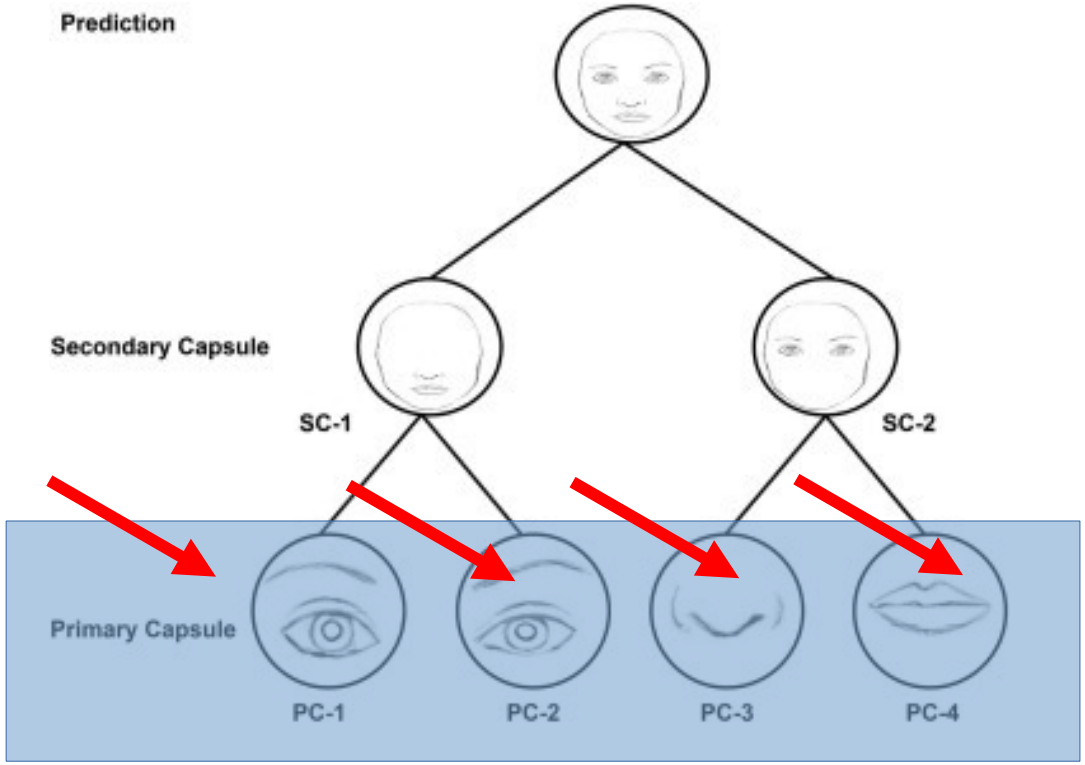
Face or NOT a Face?

Interpretable by Design Networks



Designed with interpretable mind

What is Capsule Networks (CapsNets)?

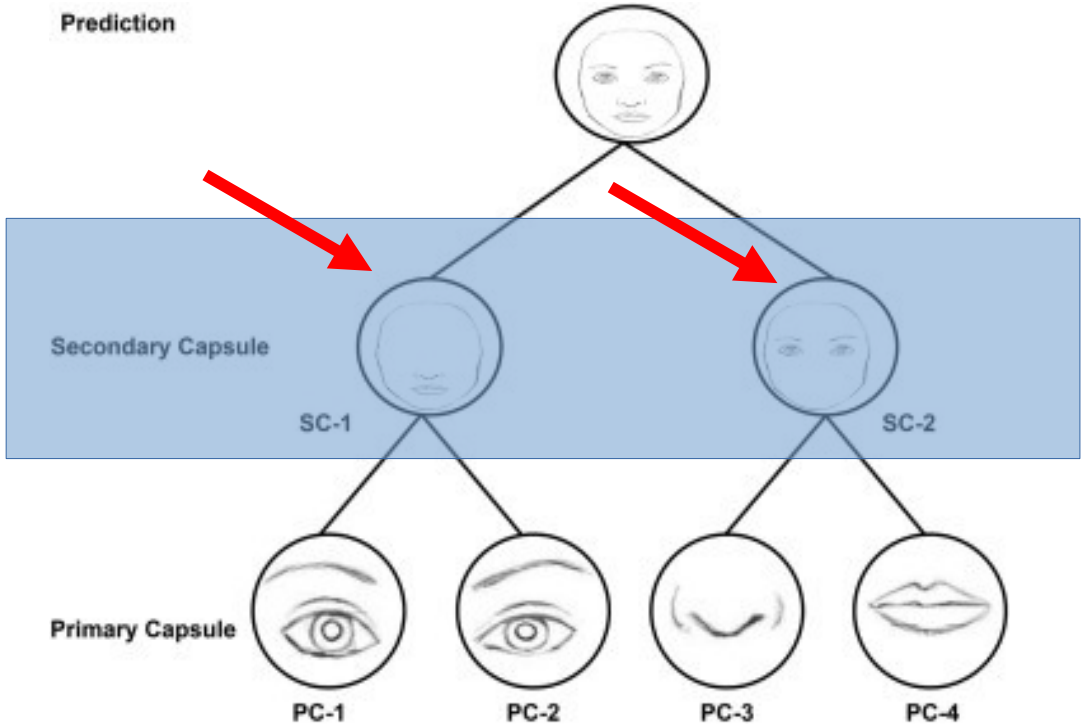


Parts (low-level) Capsules

Pawan,et. at., "Capsule networks for image classification: A review" (2022)

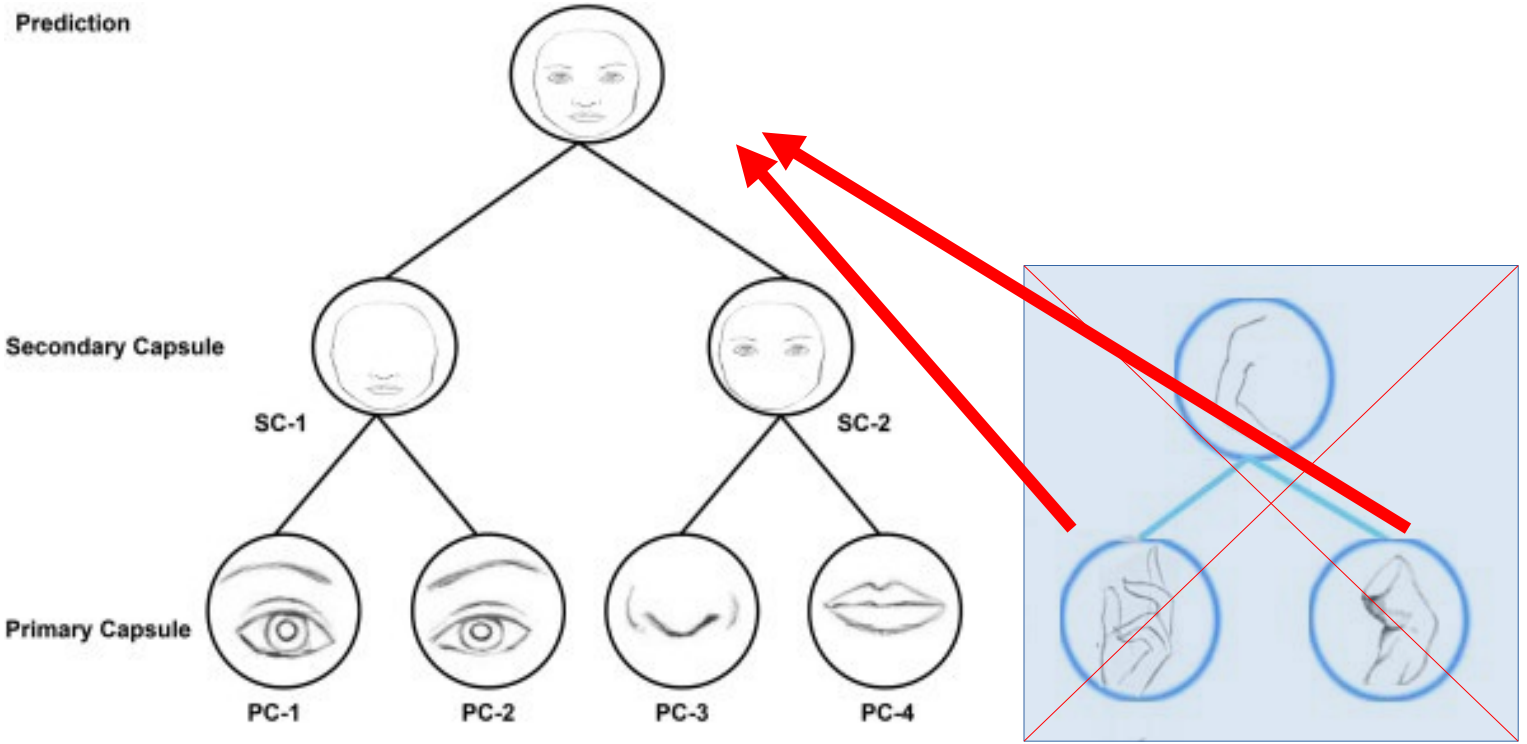
What is Capsule Networks (CapsNets)?

Whole (high-level) Capsule



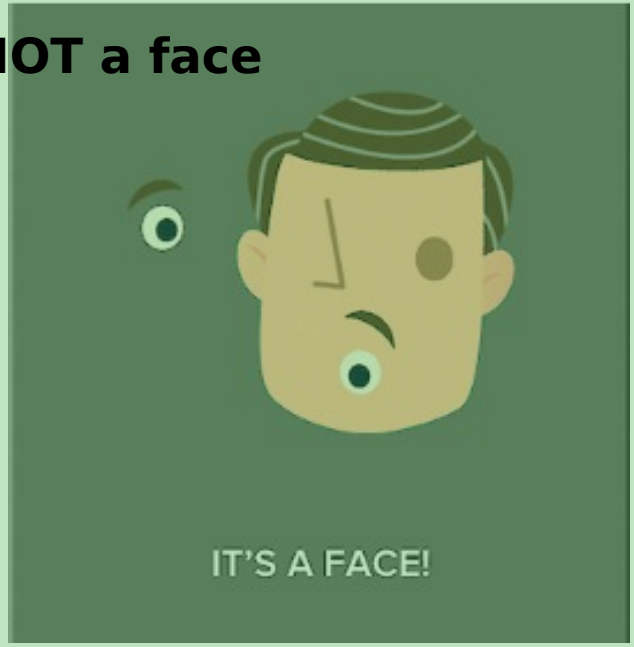
Pawan,et. at., "Capsule networks for image classification: A review" (2022)

What is Capsule Networks (CapsNets)?



Face or NOT a Face?

Capsule Networks



CNNs

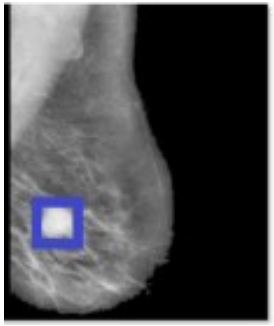
imec IDLab INTERNET & DATA LAB UNIVERSITEIT GENT Universiteit Antwerpen

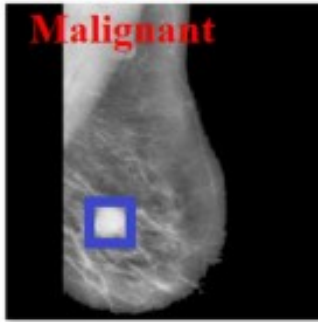
Why Do We Care about Capsule Networks!!

Diabetic detection

Breast Cancer Diagnosis

Automatic Target Recognition





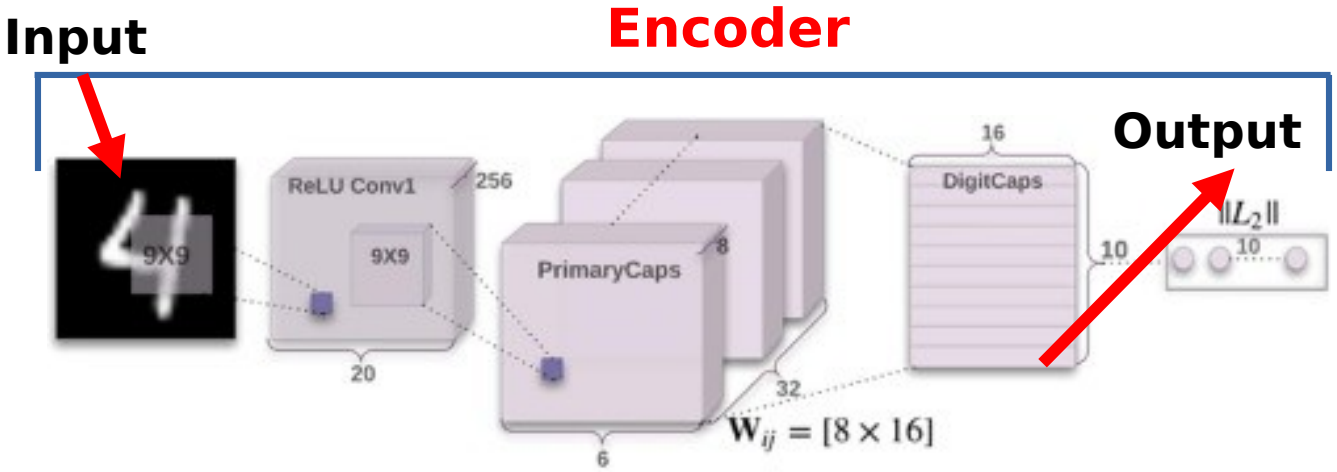
Kalyani, et al. "Diabetic retinopathy detection and classification using capsule networks." Complex & Intelligent Systems (2021)

Anupama, et al. "Breast cancer classification using capsule network with preprocessed histology images." 2019 International conference on communication and signal processing (2019)

Shah, et al. "Automatic target recognition from SAR images using capsule networks." Pattern Recognition and Machine Intelligence (2019)

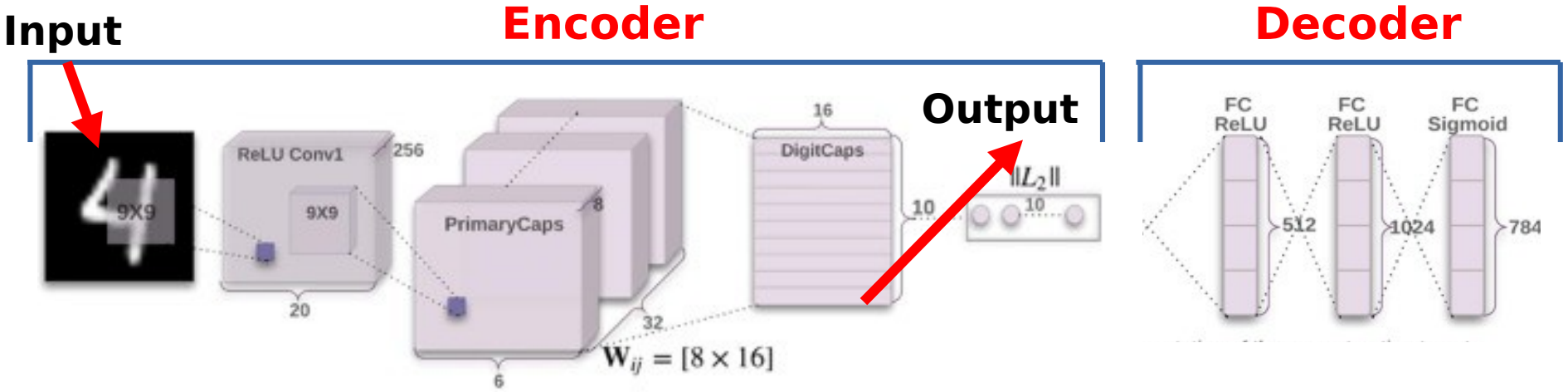
Understanding the inner working and behavior of CapsNet

- Define the path in CapsNet
  - Tracking...



Understanding the inner working and behavior of CapsNet

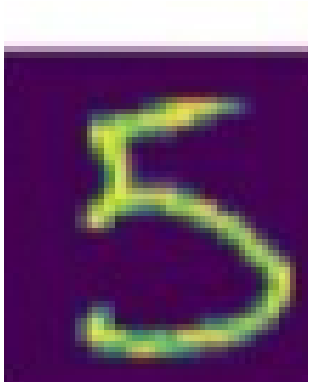
- Define the path in CapsNet
  - Tracking...



Understanding the inner working and behavior of CapsNet

- Define the path in CapsNet
  - Tracking...

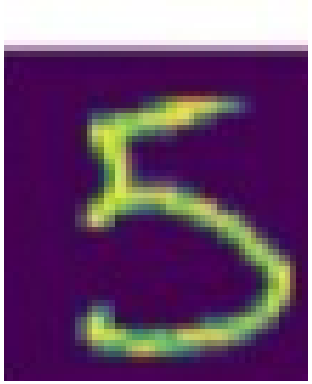
Input



Understanding the inner working and behavior of CapsNet

- Define the path in CapsNet
  - Tracking...

Input



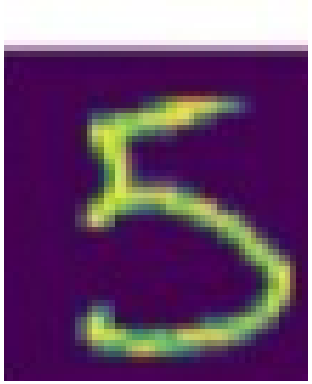
Encoder

```
-2.05661259e+01 -1.18764219e+01 -1.86646614e+01
-1.29753962e+01 -2.87226486e+01 -2.33451080e+01
-2.44269047e+01 -1.37865725e+01 -2.18092003e+01
-1.53510189e+01 -1.45434084e+01 -1.38706608e+01]
-1.33989143e+00  5.98526537e-01  2.65634656e-02
-1.09416890e+00 -4.28804219e-01  3.08536917e-01
-1.02626455e+00 -6.84777558e-01  1.63087773e+00
 5.69257021e-01  3.01515818e+00  5.26641071e-01
-9.91878688e-01  1.09790301e+00  2.77628839e-01
 8.92293453e-03  2.71537900e-01 -5.34587085e-01
 1.47699445e-01  3.78361434e-01  1.39822870e-01
 1.51214051e+00  1.61666870e+00 -7.11558908e-02]
 9.10256803e-02  1.80516332e-01  3.44895065e-01
 2.77635545e-01  1.34544238e-01  1.24773756e-01
 1.79092154e-01  1.98206961e-01  1.85325250e-01
 1.63378552e-01  2.85243630e-01  2.31984437e-01
 2.33559057e-01  1.41272530e-01  2.46041313e-01
 1.53355926e-01  2.90040970e-01  2.36210048e-01
 1.64508194e-01  2.02810913e-01  3.44647467e-01
 1.91309452e-01  2.35761121e-01  1.49378166e-01]
```

Understanding the inner working and behavior of CapsNet

- Define the path in CapsNet
  - Tracking...

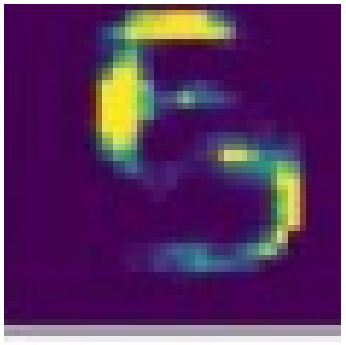
Input



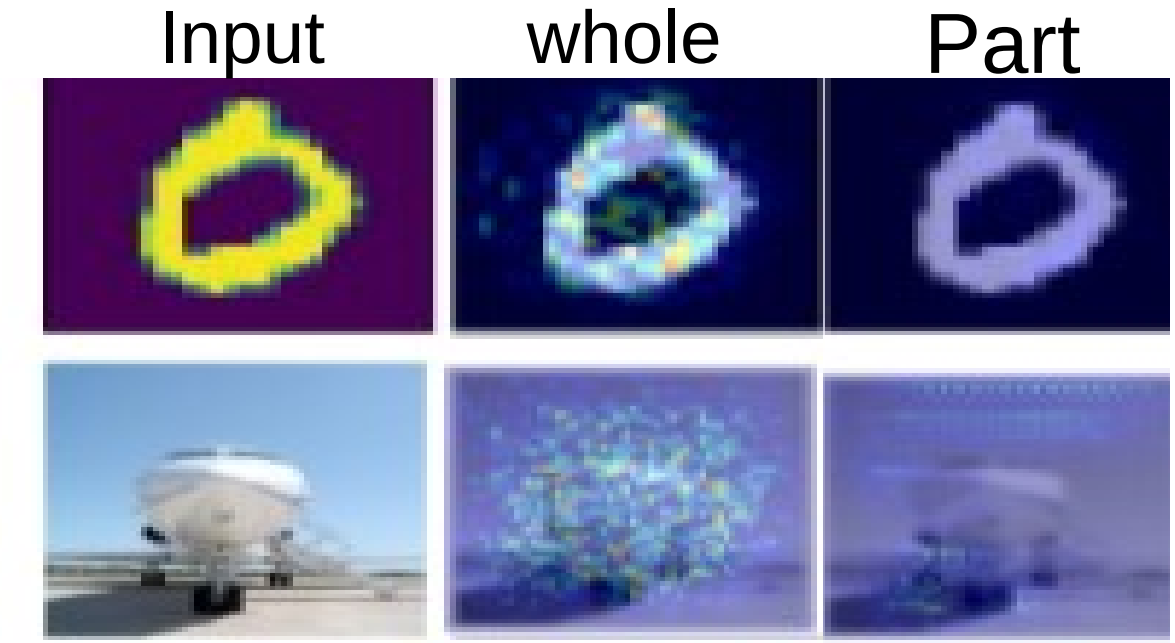
Encoder

```
-2.05661259e+01 -1.18764219e+01 -1.86646614e+01
-1.29753962e+01 -2.87226486e+01 -2.33451080e+01
-2.44269047e+01 -1.37865725e+01 -2.18092003e+01
-1.53510189e+01 -1.45434084e+01 -1.38706608e+01]
-1.33989143e+00  5.98526537e-01  2.65634656e-02
-1.09416890e+00 -4.28804219e-01  3.08536917e-01
-1.02626455e+00 -6.84777558e-01  1.63087773e+00
 5.69257021e-01  3.01515818e+00  5.26641071e-01
-9.91878688e-01  1.09790301e+00  2.77628839e-01
 8.92293453e-03  2.71537900e-01 -5.34587085e-01
 1.47699445e-01  3.78361434e-01  1.39822870e-01
 1.51214051e+00  1.61666870e+00 -7.11558908e-02]
 9.10256803e-02  1.80516332e-01  3.44895065e-01
 2.77635545e-01  1.34544238e-01  1.24773756e-01
 1.79092154e-01  1.98206961e-01  1.85325250e-01
 1.63378552e-01  2.85243630e-01  2.31984437e-01
 2.33559057e-01  1.41272530e-01  2.46041313e-01
 1.53355926e-01  2.90040970e-01  2.36210048e-01
 1.64508194e-01  2.02810913e-01  3.44647467e-01
 1.91309452e-01  2.35761121e-01  1.49378166e-01]
```

Decoder



Measuring Part-Whole Relationship (Hierarchical Relationship)



AL-Tawalbeh, et. at., "Towards the Characterization of Representations Learned via Capsule-based Network Architectures" (2023)

Conclusion

- AI also make mistakes

Towards the Characterization of Representations Learned via Capsule-based Network Architectures

Saja AL-Tawalbeh and José Oramas

University of Antwerp, imec-IDLab

Conclusion

- AI also make mistakes
- The visualizations are understandable by humans
- Capsule network may have a weak hierarchical relationship
- Interesting research ideas

Towards the Characterization of Representations Learned via Capsule-based Network Architectures

Saja AL-Tawalbeh and José Oramas

University of Antwerp, imec-IDLab

Future Challenges

- Capsule networks (2017)
  - Fundamental
  - Applying these methods to real world challenges

- Explainable artificial intelligence
  - Text

- Understand the behavior / Chat GPT
  - Common sense as international collaboration

Contact Information

- Saja Tawalbeh
  - Saja.Tawalbeh@uantwerpen.be

- Where can you find us?
  - The Beacon (Sint-Pietersvliet 7, 2000 Antwerp)

ORCID

Linkedin

embracing a better life

imec IDLab INTERNET & DATA LAB UNIVERSITEIT GENT Universiteit Antwerpen