Introduction to Machine Learning and Deep Learning

Vietnam Academy of Science and Technology
March 17th-19th, 2025

Kevin Webster
Imperial College London

Introduction to Machine Learning and Deep Learning

Monday

- Overview of the field
- Core machine learning (ML) concepts and techniques
- Linear regression
- Logistic regression

Tuesday

- Multilayer perceptrons
- Gradient based training
- Optimisation, regularisation, validation

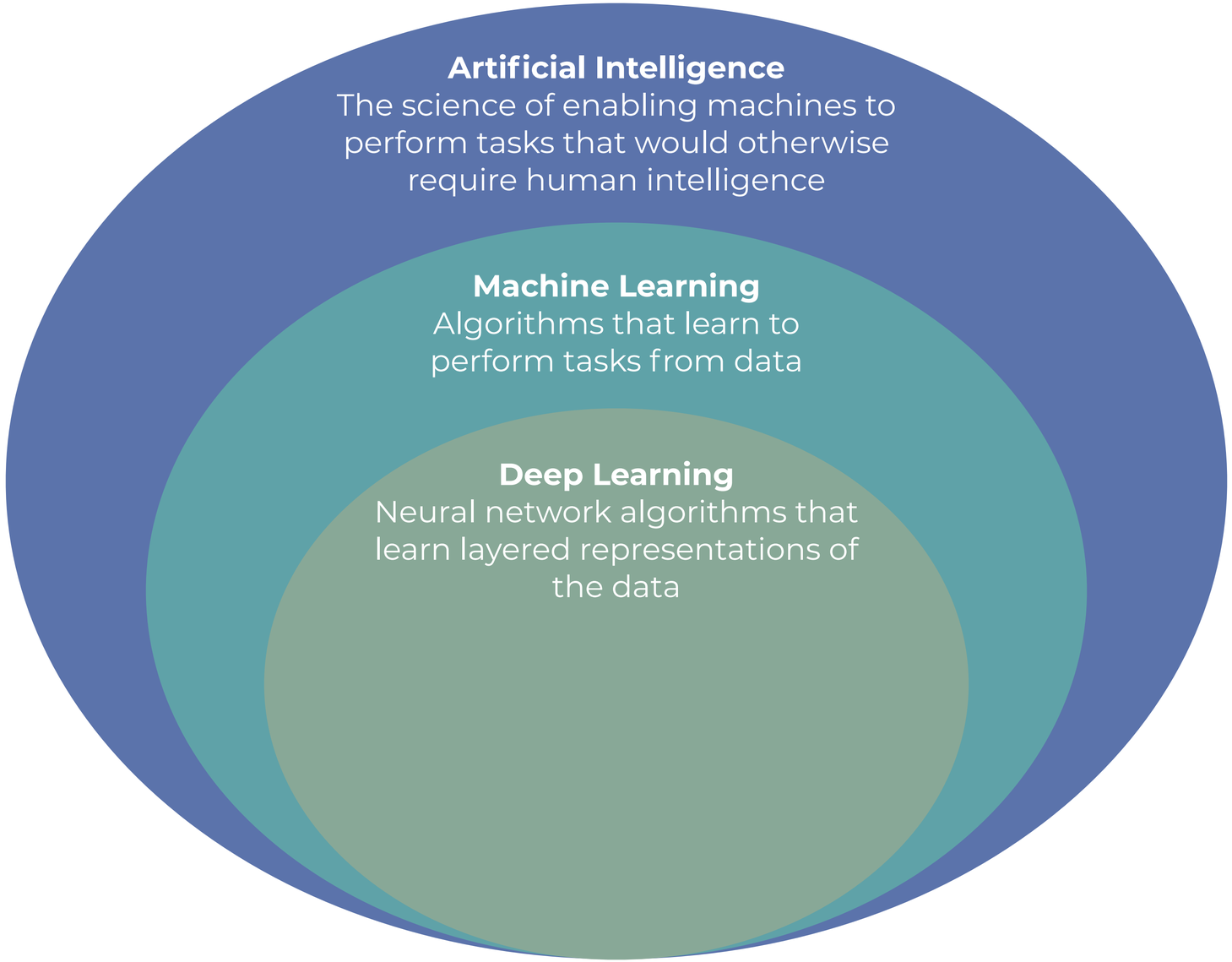
Wednesday

- Learning from time series
- Recurrent neural networks (RNNs)
- Long Short Term Memory (LSTM)

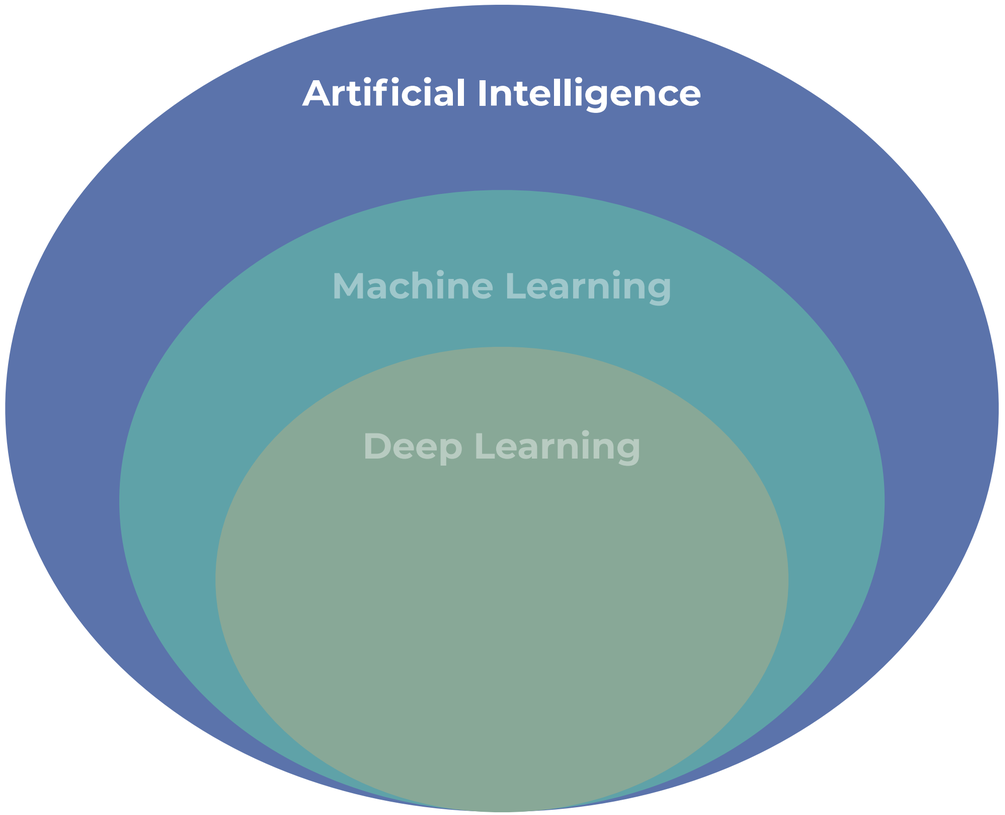
Practical sessions:

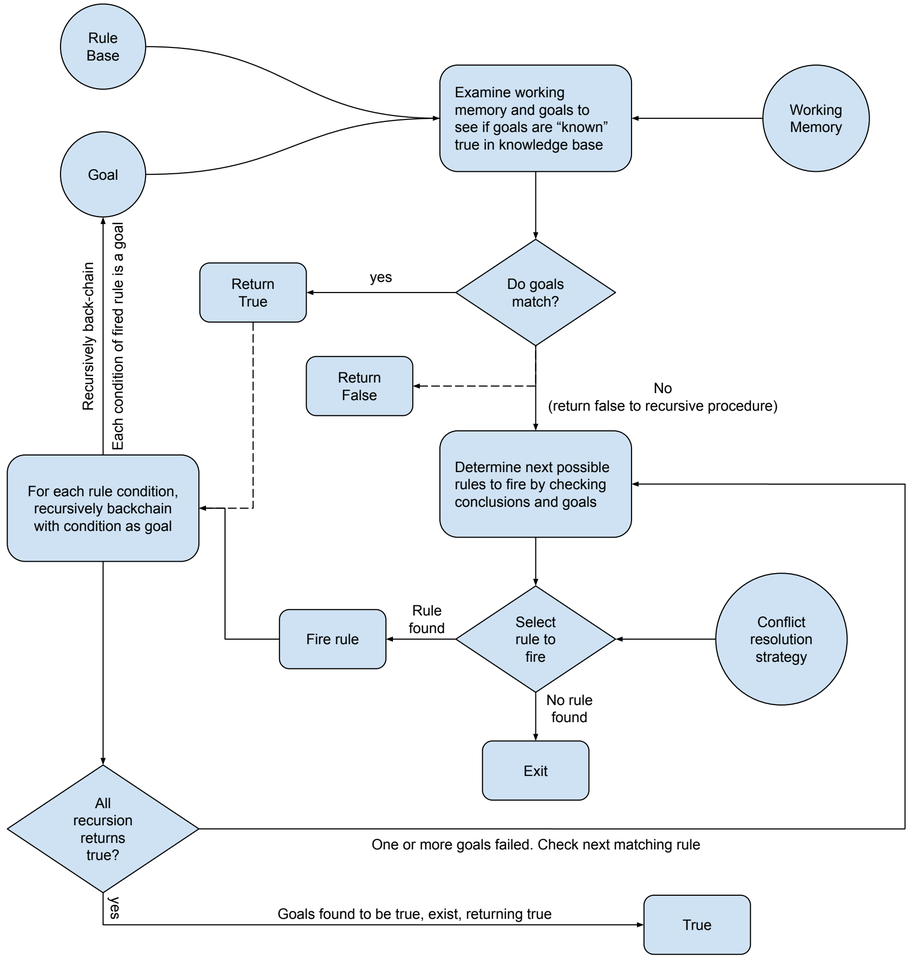
- sklearn and Keras

Artificial Intelligence, Machine Learning and Deep Learning



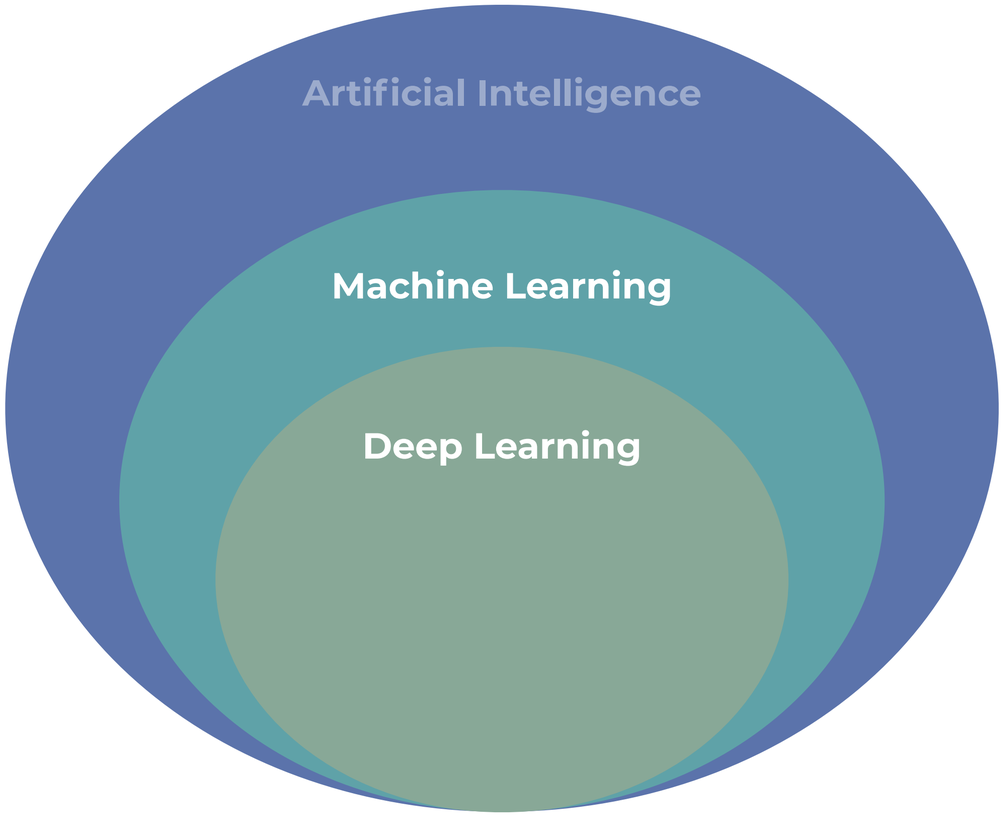
Artificial Intelligence, Machine Learning and Deep Learning

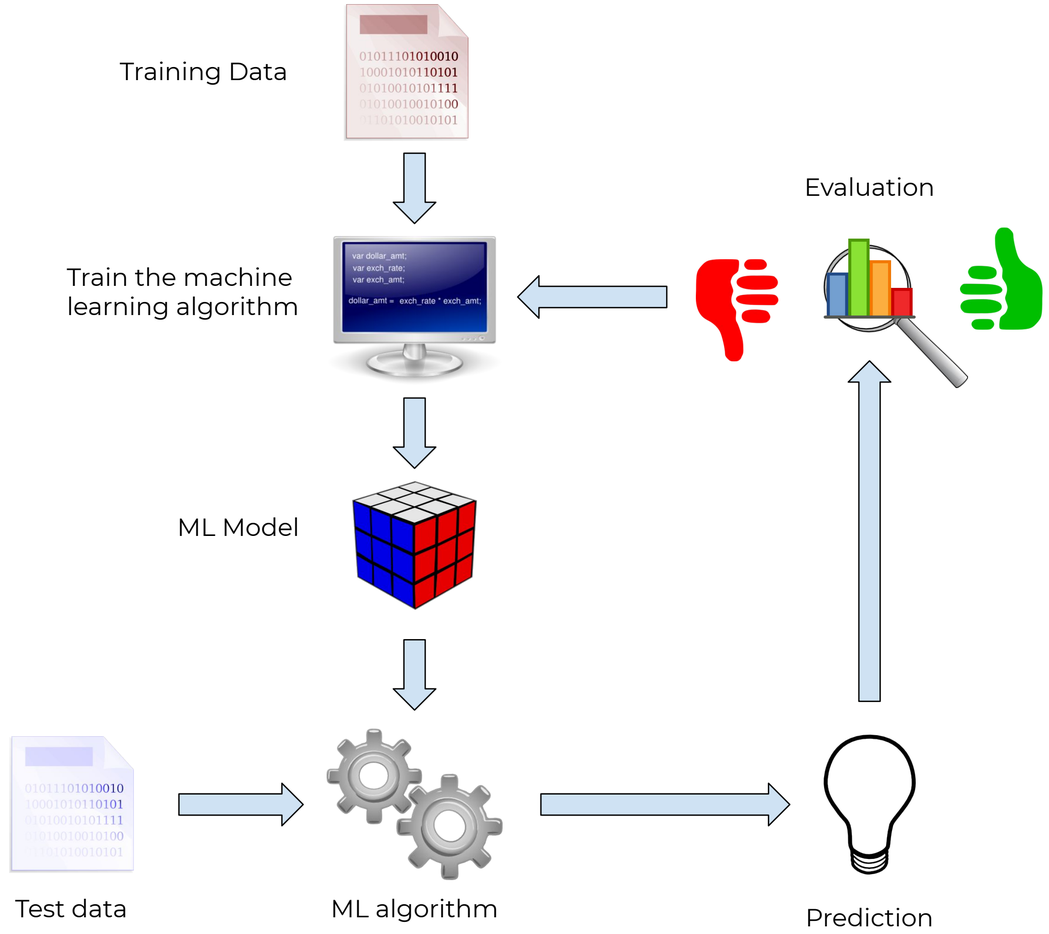




Rule-based system

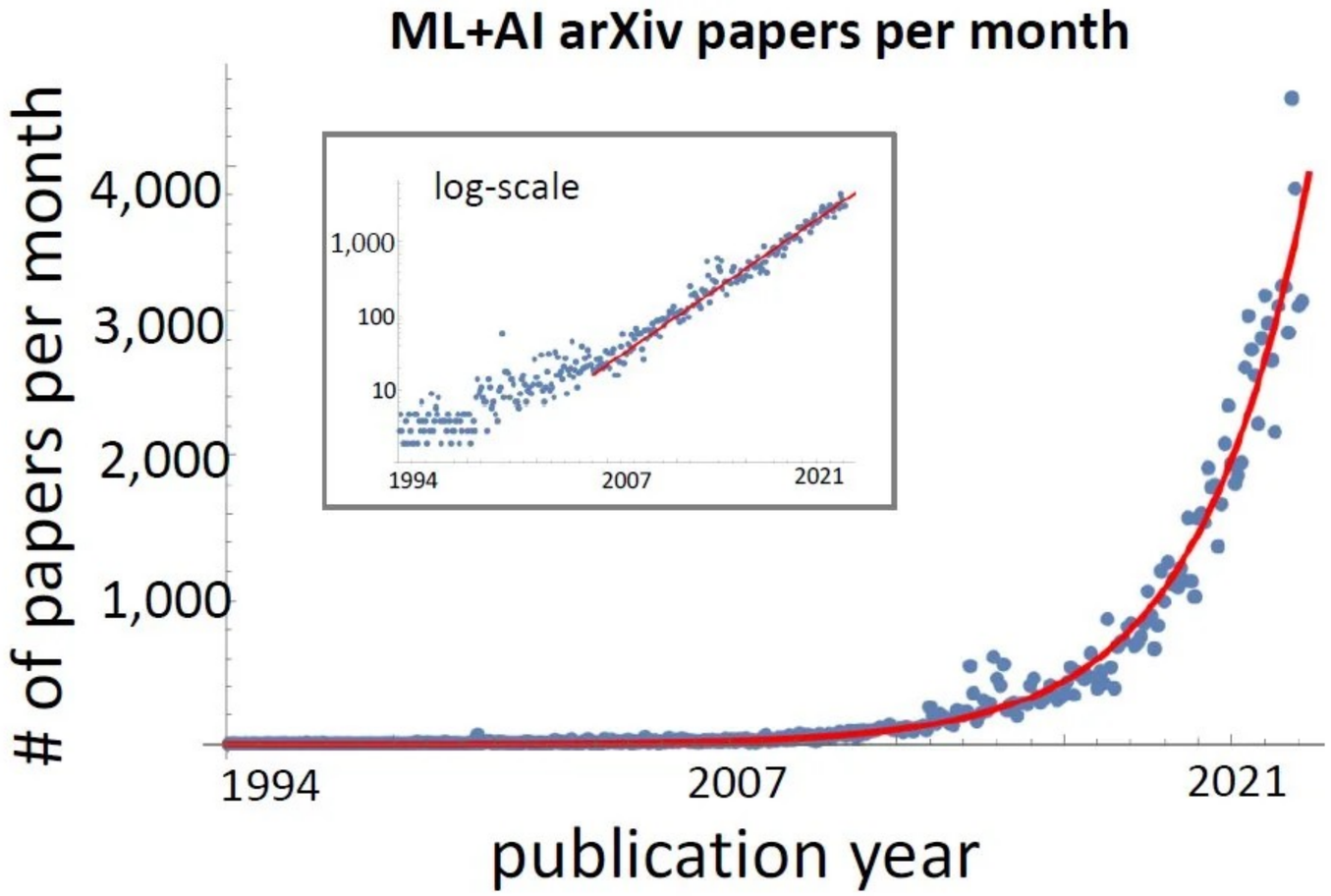
Artificial Intelligence, Machine Learning and Deep Learning





Machine learning system

Research Trends in Machine Learning



Machine learning arXiv papers per year

source: https://www.reddit.com/r/singularity/comments/xwdzr5/the_number_of_ai_papers_on_arxiv_per_month_grows/?rdt=55514

A Machine Learning Definition

> “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E”

Mitchell, T. (1997), "Machine Learning", McGraw-Hill, New York.

Tasks **T**

- Regression
- Classification (binary/multiclass)
- Clustering
- Anomaly detection
- Density estimation
- Recommender systems

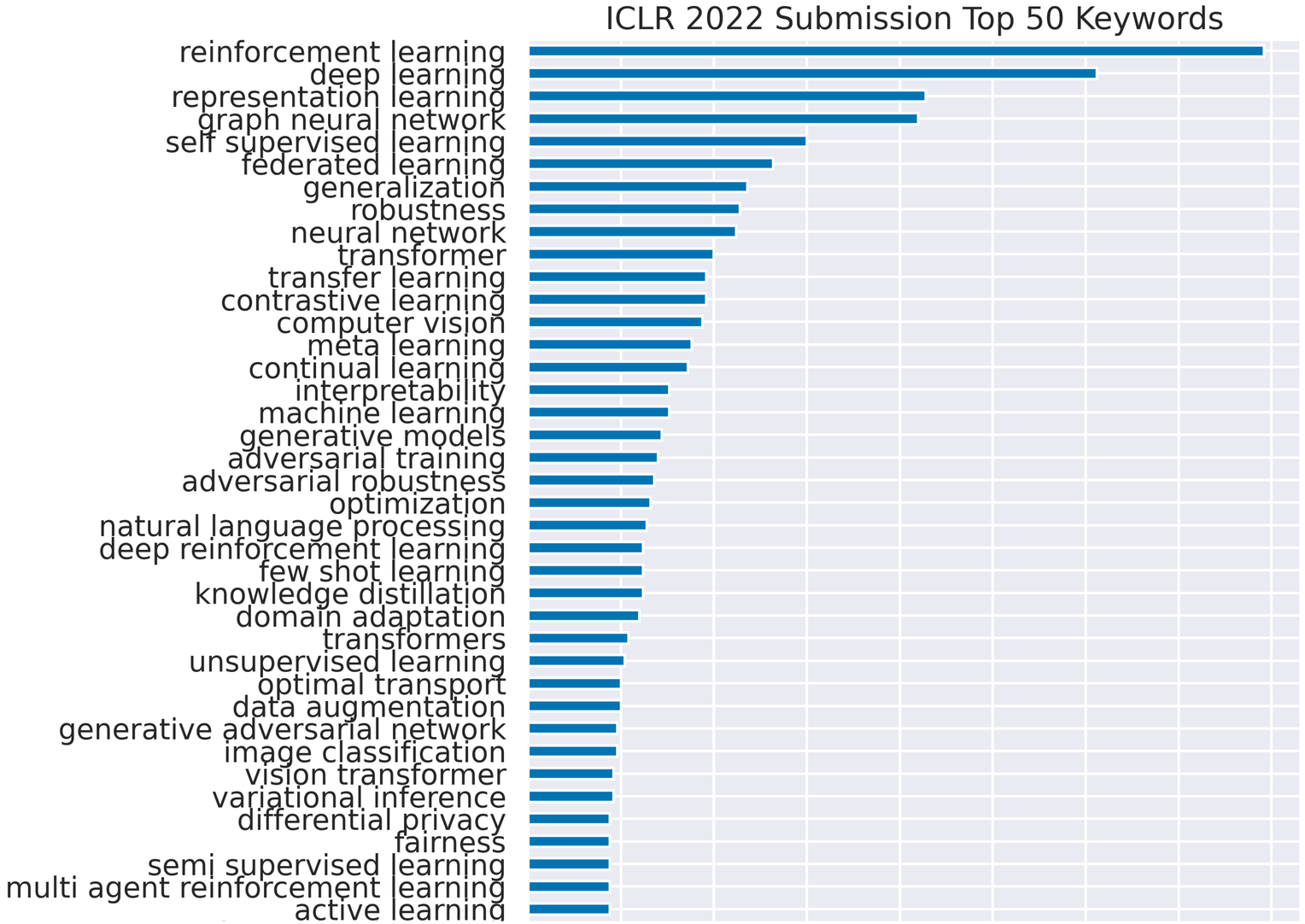
Experience **E**

- Labelled examples
- Unlabelled examples
- Streaming data (interactive)

Measure **D**

- Mean squared error (MSE)
- Cross entropy
- Mean absolute error (MAE)
- ROC-AUC
- Accuracy
- Precision/recall
- F1-score

Research Trends in Machine Learning



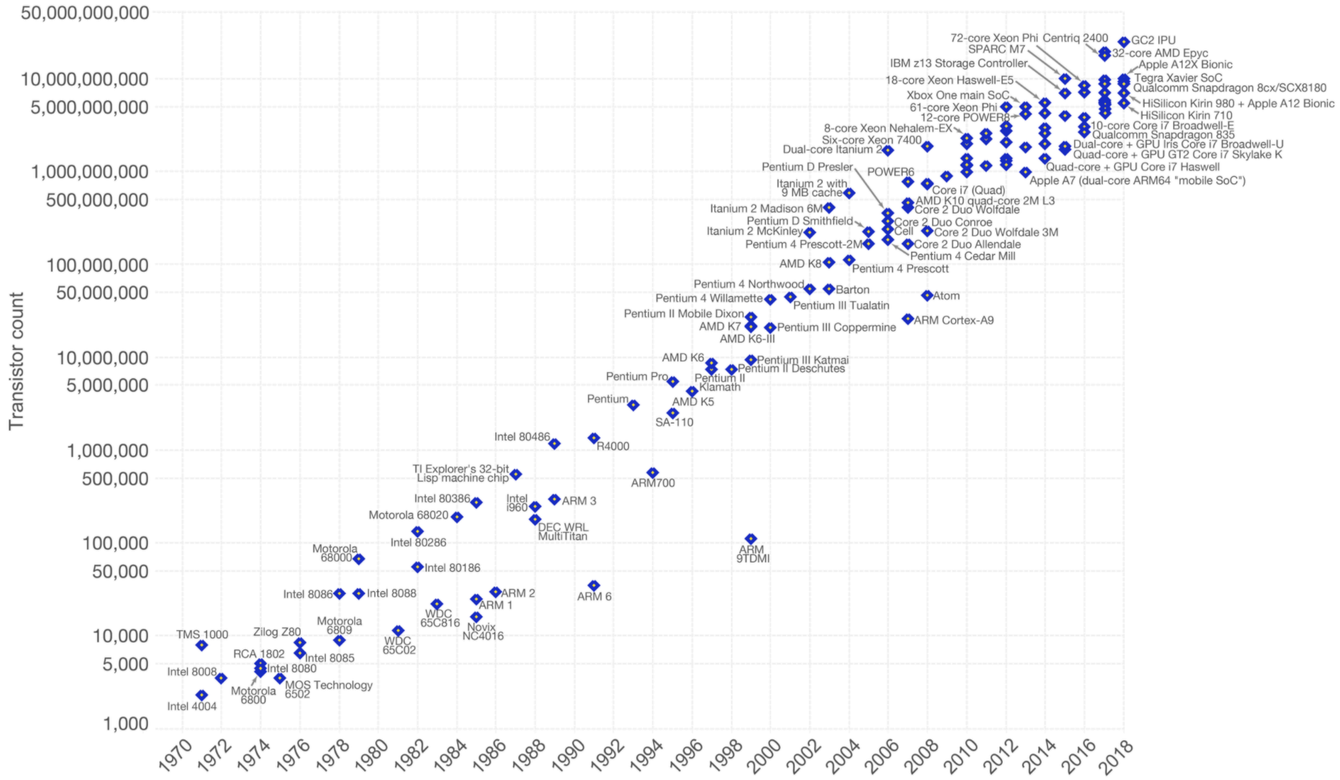
source: https://github.com/fedebotu/ICLR2022-OpenReviewData?tab=readme-ov-file

Why Deep Learning?

1. Abundance of data

- Digitisation of existing technologies
- Online shopping
- Smartphones
- Laptops
- Wearable devices
- Social media

2. Increase in computing power



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Linear regression and logistic regression

We will start by looking at two of the simplest ML models:

- Linear regression
- Logistic regression

A linear regression model is used to fit a dataset with **continuous** target values (regression), whilst logistic regression is used to fit a dataset with **discrete** target values (classification)

Linear regression

A linear regression model can be used to fit a dataset of inputs and targets $\mathcal{D} := (\mathbf{x}_i, y_i)_{i=1}^N$

$$f_\theta(\mathbf{x}) = \theta^T \phi(\mathbf{x})$$

Here the $\phi(\mathbf{x})$ is a M-dimensional vector of (nonlinear) basis functions

The parameters θ are chosen to minimise the mean squared error loss

$$L_{MSE}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (f_\theta(\mathbf{x}_i) - y_i)^2$$

Linear regression

$$f_\theta(\mathbf{x}) = \theta^T \phi(\mathbf{x})$$

$$L_{MSE}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (f_\theta(\mathbf{x}_i) - y_i)^2$$

It can be shown that the solution to this minimisation problem can be found explicitly, using the **normal equation**:

$$\hat{\theta} = (\Phi_{\mathbf{x}}^T \Phi_{\mathbf{x}})^{-1} \Phi_{\mathbf{x}}^T \mathbf{y}$$

Where $\Phi_{\mathbf{x}}$ is the NxM **design matrix**, formed by stacking the feature vectors for each example in the rows of the matrix

Exercise: prove the normal equation! Under what conditions is $\Phi_{\mathbf{x}}^T \Phi_{\mathbf{x}}$ invertible?

Logistic regression

We can also consider classification problems, where we have a dataset of inputs and targets $\mathcal{D} := (\mathbf{x}_i, y_i)_{i=1}^N$ with $y_i \in \mathcal{C}$, where $\mathcal{C}$ is some finite set of classes.

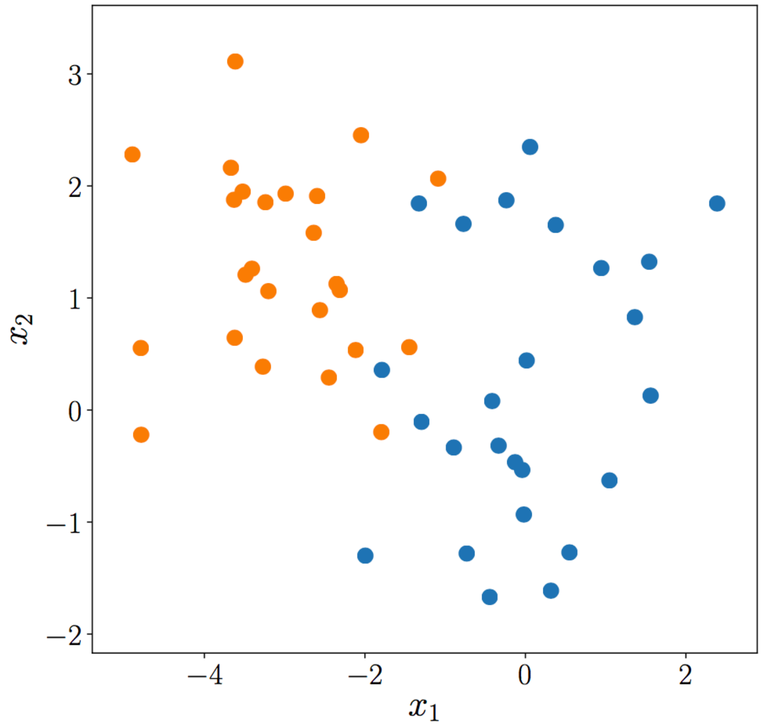
We will consider binary classification, where $\mathcal{C} = \{0, 1\}$

A logistic regression classifier is defined as a model

$$f_\theta(\mathbf{x}) = \sigma(\theta^T \phi(\mathbf{x}))$$

where σ is the logistic sigmoid function, given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Logistic regression

$$f_\theta(\mathbf{x}) = \sigma(\theta^T \phi(\mathbf{x}))$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Now, the output of our model is interpreted as $p(y = 1 \mid x)$

The loss function that we wish to minimise is the **binary cross entropy**:

$$L_{BCE}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \{y_i \log f_\theta(\mathbf{x}_i) + (1 - y_i) \log(1 - f_\theta(\mathbf{x}_i))\}$$

Logistic regression

$$L_{BCE}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\{y_i \log f_\theta(\mathbf{x}_i) + (1-y_i)\log(1-f_\theta(\mathbf{x}_i))\}$$

Unlike linear regression, there is no closed-form solution of the logistic regression problem

However, it can be shown that the loss function is convex

To optimise the parameters of the logistic regression model, we need to resort to **gradient-based optimisation**

Key machine learning concepts

We will use a linear regression model on a toy dataset as an example to demonstrate the following key ML concepts:

- Model selection
- Input and target features
- Loss function
- Basis functions
- Model capacity
- Training/validation/test splits
- Overfitting & underfitting
- Regularisation