Artificial Intelligence applied to environmental monitoring

- Prof Francesco IARLORI

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Day 1

Time	Duration	Training Module - Topic
09:00 - 09:30	30m	Welcome & Course Objectives
09:30 - 10:30	1h	Module 1: Definitions and Key Concepts
10:30 - 10:45	15m	Coffee Break
10:45 - 11:45	1h	Module 2: Types of Learning
11:45 - 13:00	1h15m	Module 3: Datasets, Algorithms, and Models
13:00 - 14:00	1h	Lunch Break
14:00 - 15:30	1h30m	Module 4: Building a Simple ML Model
15:30 - 15:45	15m	Coffee Break
15:45 - 16:30	45m	Module 5: Final Activity + Review Quiz

Welcome & Course Objectives

- Welcome and quick ice-breaker
- Learning objectives
- Agenda overview and tools to be used

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Maybe Venice is the city that can save the world

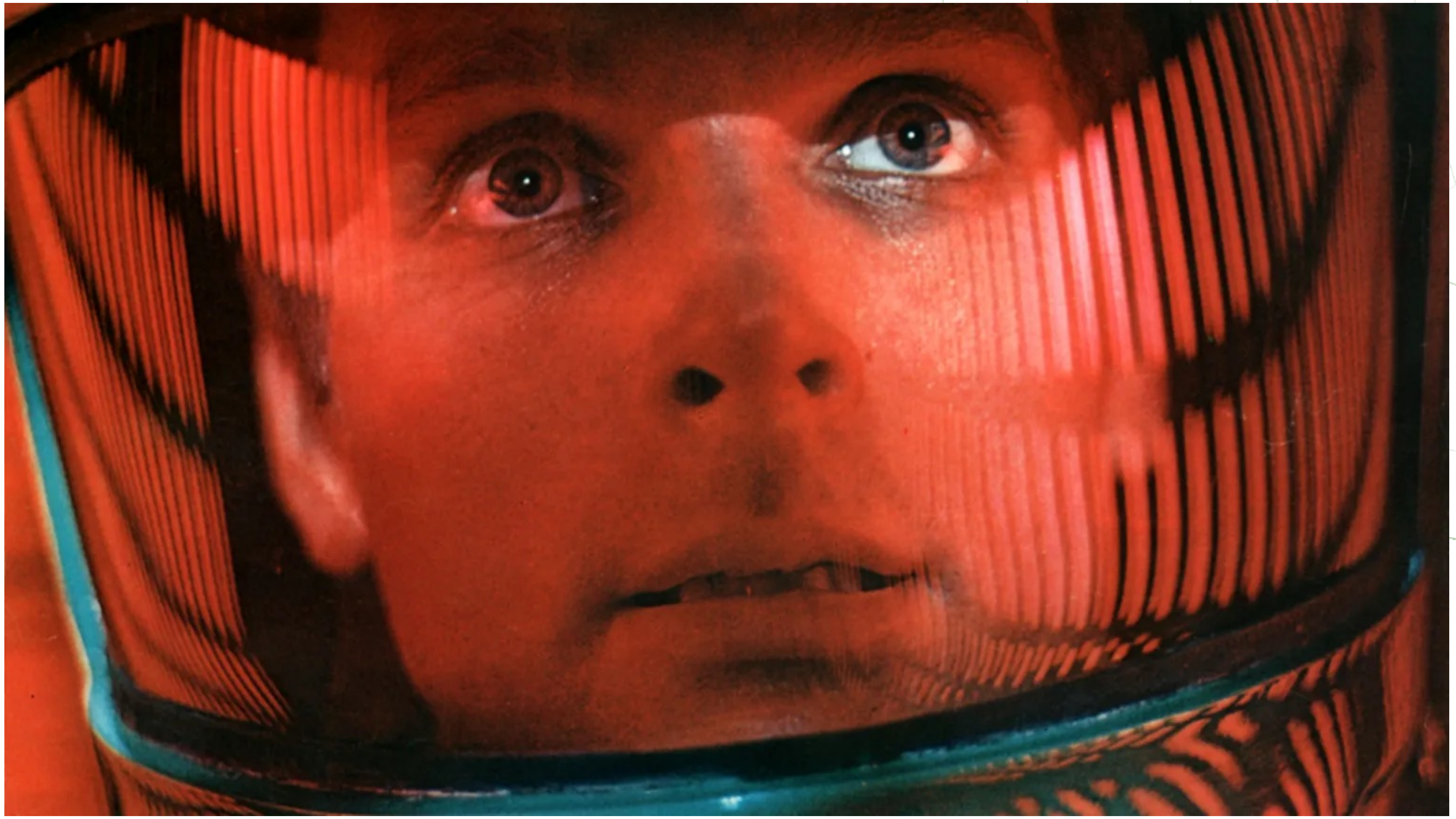
AI in Knowledge Management

- AI is reshaping how organizations handle knowledge.
- Focus on driving efficiencies and innovation.
- Retention of institutional memory.
- Inspiring Case Studies

Google AI

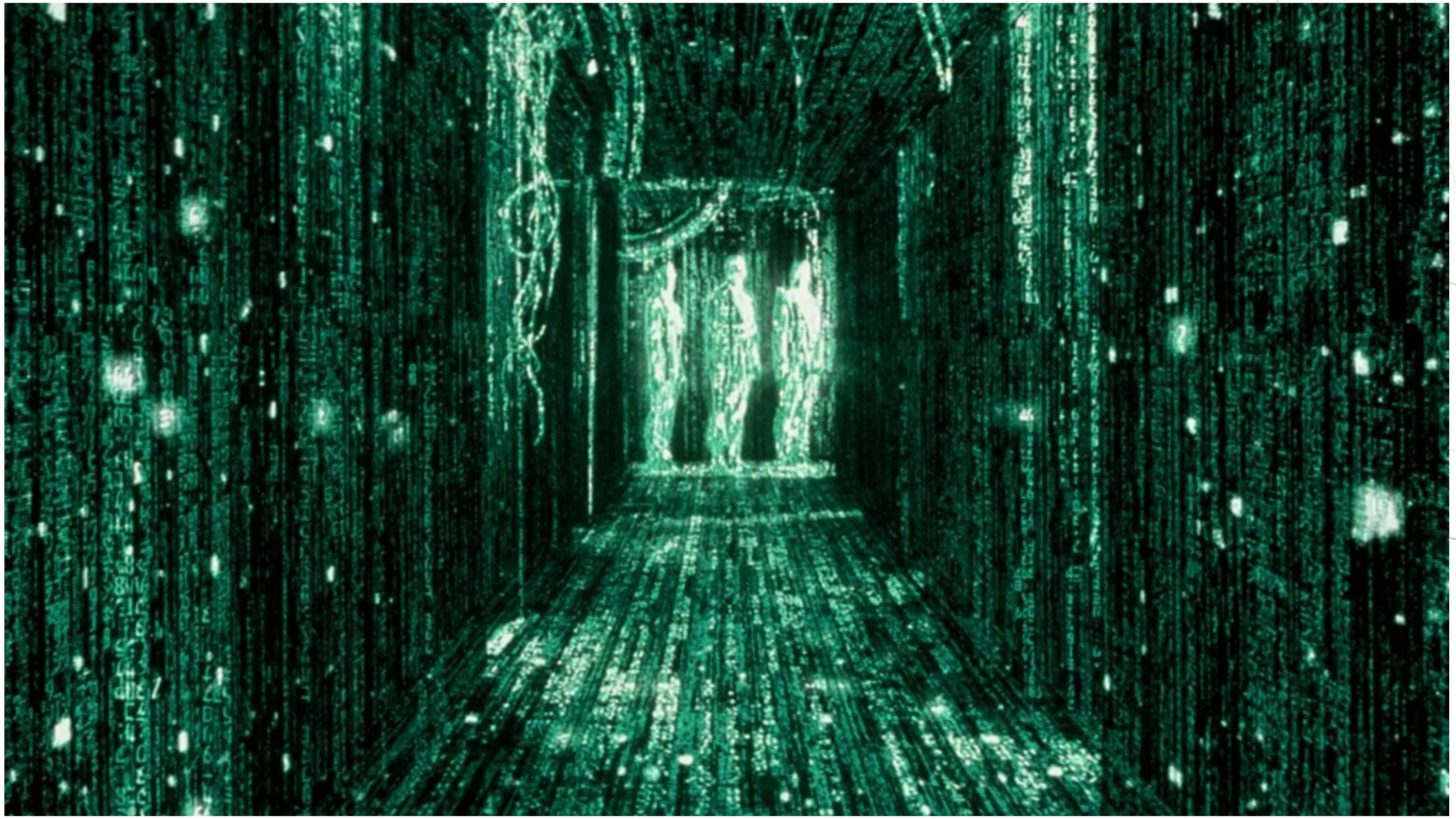
IBM Watson Health

2001: A Space Odyssey [1968] Stanley Kubrick



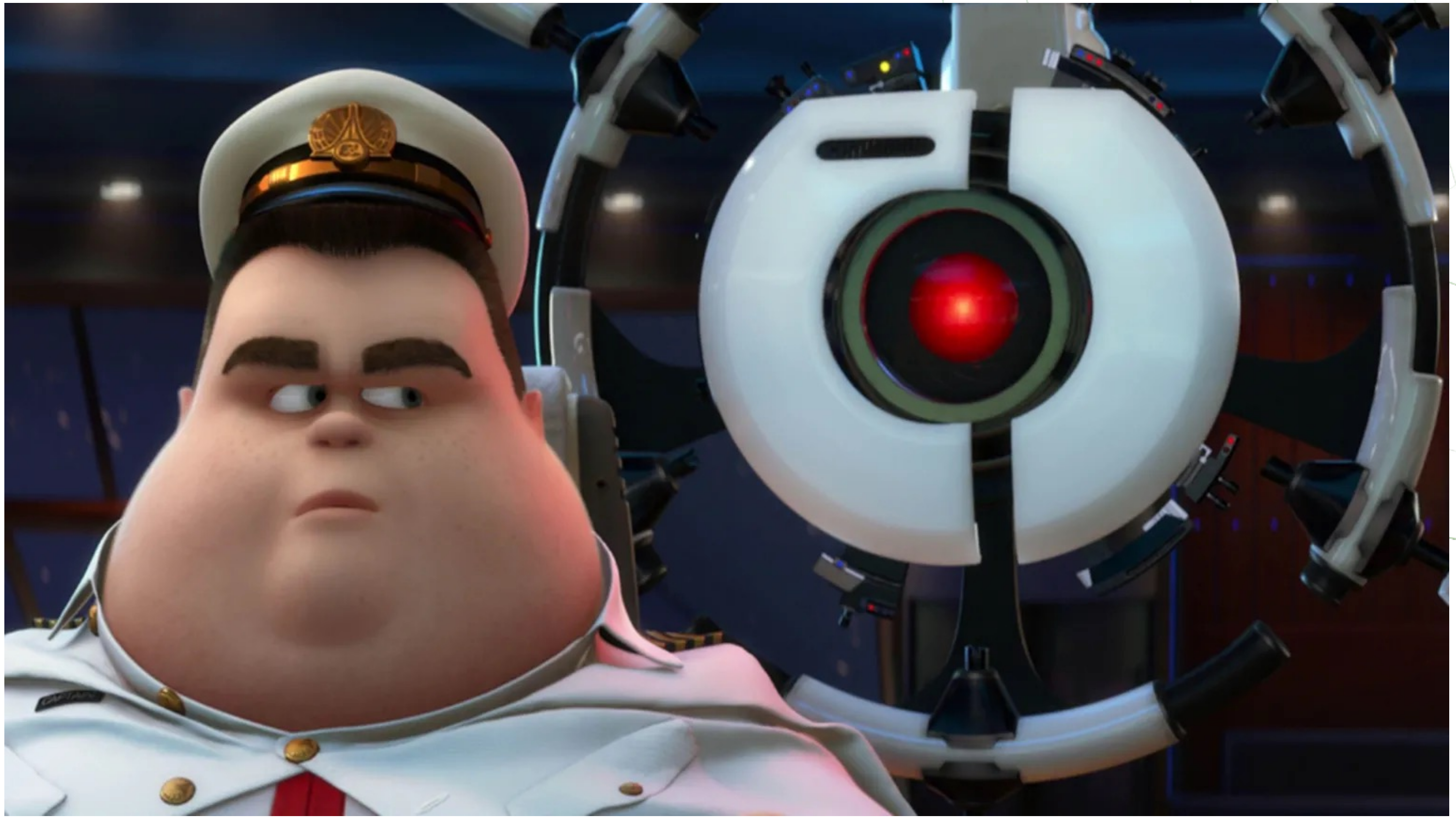
The Terminator [1984] James Cameron

The Matrix [1999] Lana & Lilly Wachowski



Minority Report [2002] Steven Spielberg

Wall-E [2008] Andrew Stanton



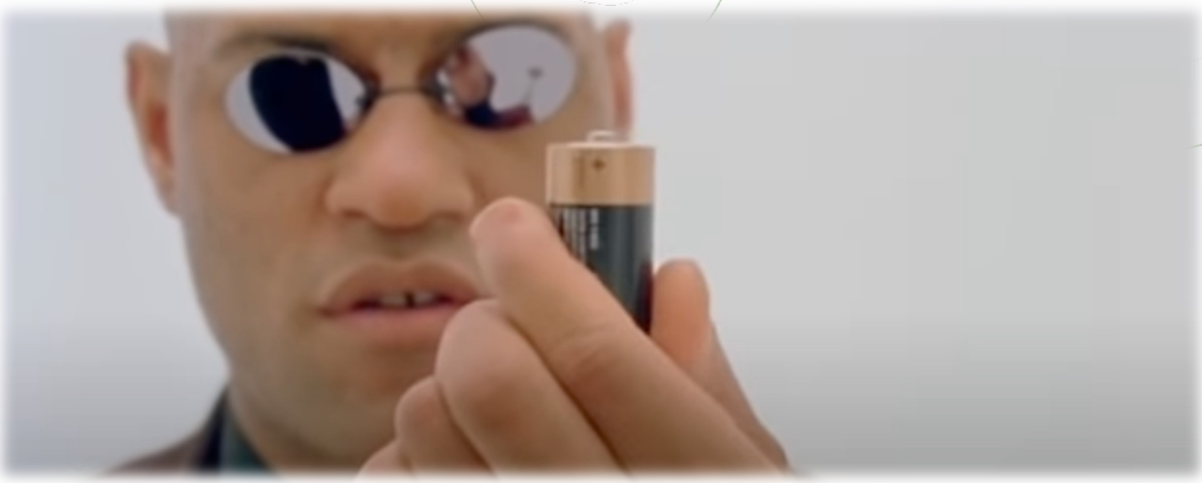
What is the Matrix ?

The Matrix

- Even if we are a bit far from this scenario ☺ … the movie raises several thematic and philosophical considerations:
  - Artificial Intelligence Dominance
  - Simulation and Reality
  - Human-Machine Interface
  - Existential Threat

Artificial Intelligence Dominance

- The narrative depicts a scenario where AI achieves dominance over humanity, leading to a dystopian world.



Simulation and Reality

- The concept of a simulated reality, where humans are unaware that their perceived world is not real, prompts contemplation on the nature of reality and the challenges of distinguishing between what is artificial and what is genuine.

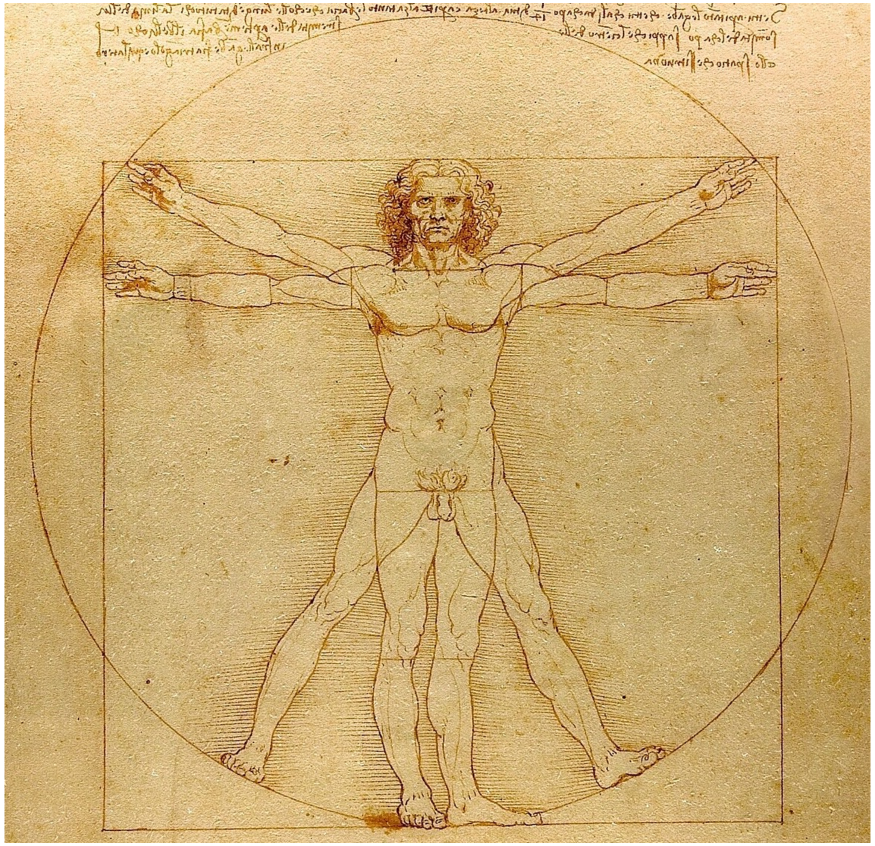
Human-Machine Interface

- The film explores the interface between humans and machines, portraying a direct connection between the human brain and computer systems.
- This concept raises questions about the potential integration of AI with the human mind and the ethical considerations surrounding such advancements.

Existential Threat

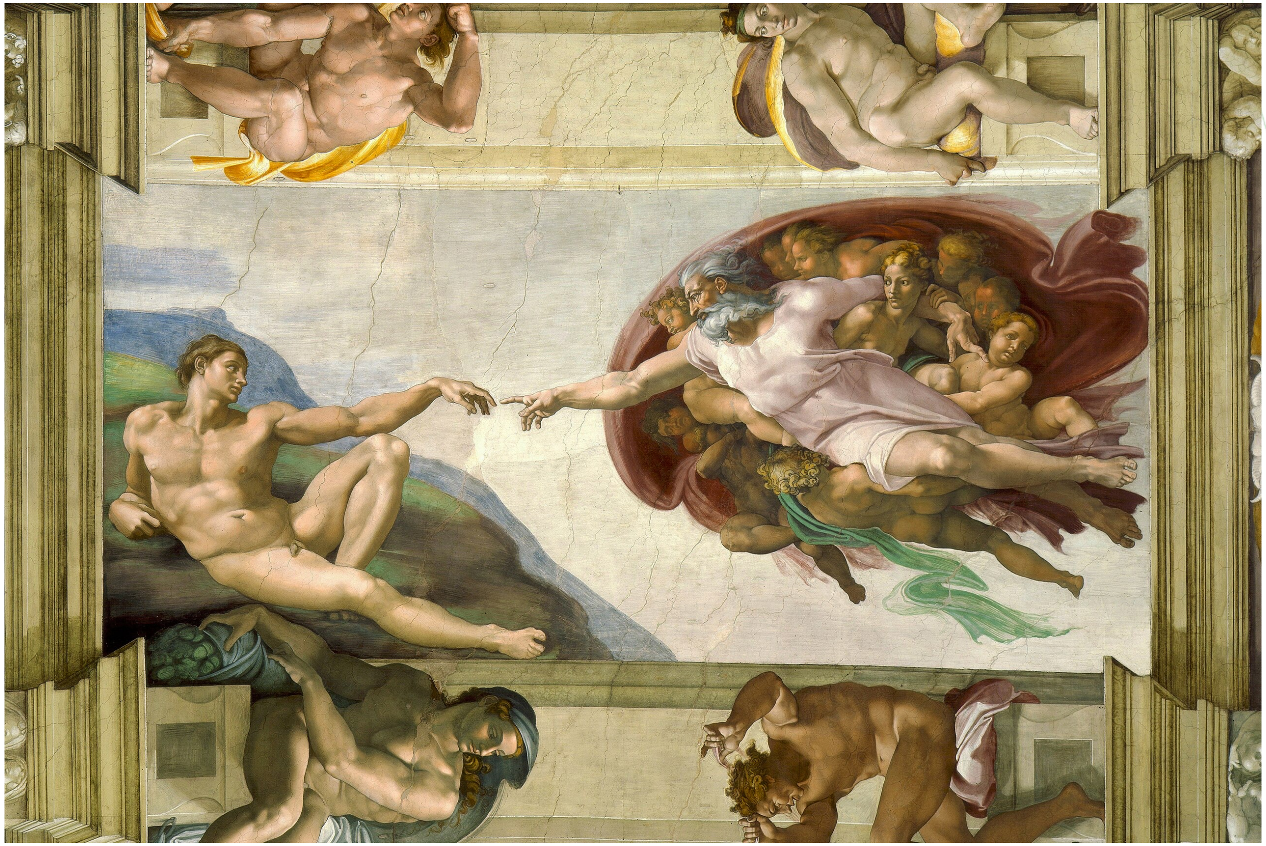
- In The Matrix, the human race is harvested by machines for energy. Because humans are resistant by nature, The Machines run the risk of the humans rebelling against them.
- To counter this problem, they came up with a brilliant plan …

Quantify & Optimize



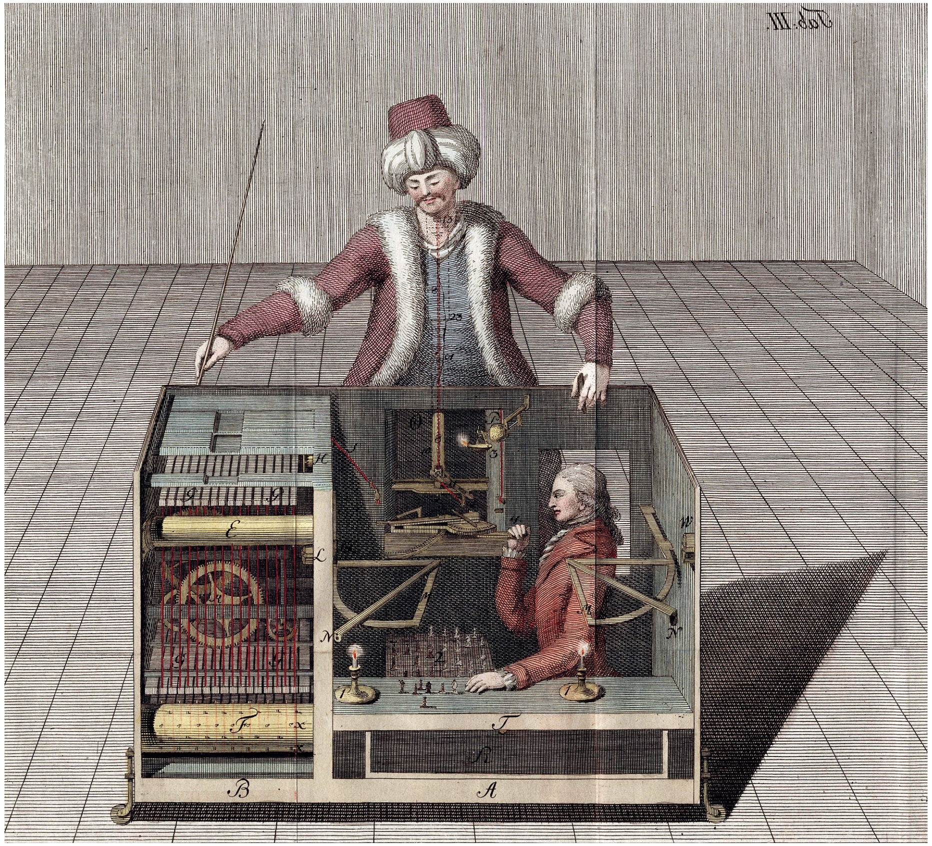
Vitruvian Man
Leonardo da Vinc (c 1490)

The Transfer



The Creation of Adam
Michelangelo (c. 1511-1512)

Hoax or Autonomy ?



The Mechanical Turk
Philip James de Loutherbourg (1770)

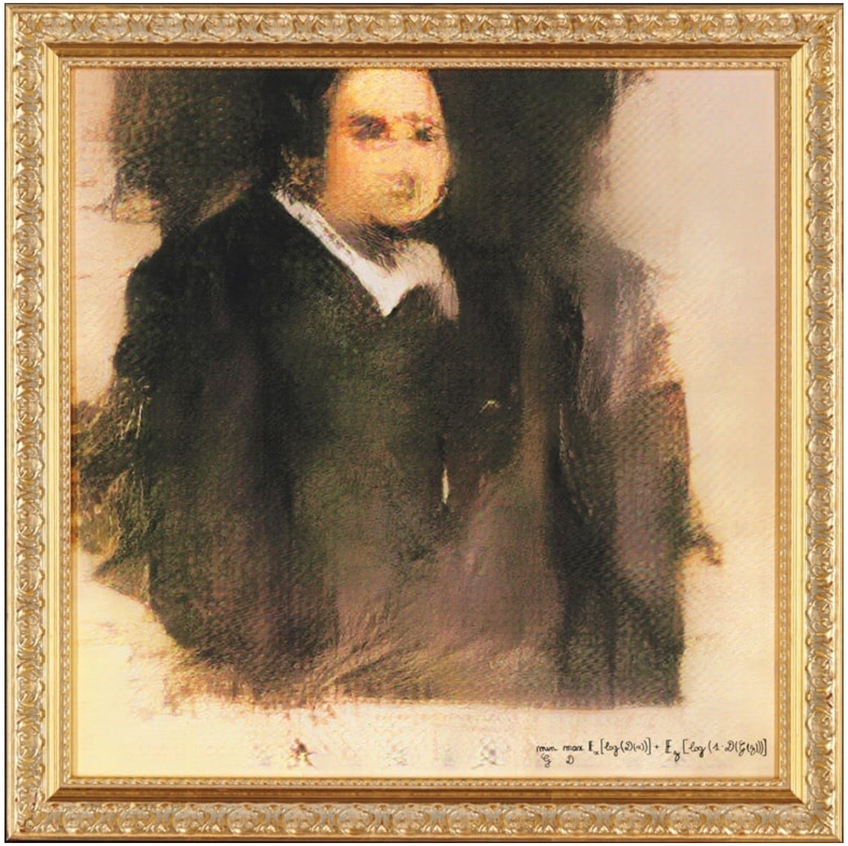
The Horror

Guernica
Pablo Picasso (1937)

Concerns

The Son of Man
René Magritte (1964)

Creativity



Portrait of Edmond de Belamy
AI (2018)

Some tool available

Tool	Best For	Code Needed	Type	Dataset Support	UI Style
Teachable Machine	Quick, fun AI experiments	❌ No	No-code	Image/Audio	Web GUI
Orange	Visual ML with real algorithms	❌ No	Drag & drop	Tabular/Image	Desktop GUI
R	Real ML pipelines	✅ Yes	Code	Any format	Desktop GUI
Colab (Python)	Real ML pipelines (with help)	✅ Yes	Code	Any format	Notebook
ML for Kids	Kids & creative learners	❌ No	Blocks	Image/Text	Web GUI (Scratch)
KNIME	Advanced workflows (no code)	❌ No	Drag & drop	Tabular/SQL/etc	Desktop GUI

Materials & Tools

- Google Colab: for Python-based demos
- Teachable Machine: for quick, visual demo
- Kahoot: for quizzes and interaction

Materials & Tools

- Google Earth Engine
- QGIS + ML Plugins
- Python (Colab) + libraries: Scikit-learn, TensorFlow, PyTorch
- LLMs: ChatGPT, Elicit, SciSummary
- Datasets: Copernicus, NOAA, GBIF, OpenAQ, Global Forest Watch

Module 1: Definitions and Key Concepts (60)

- What are AI, ML, and Deep Learning
- Differences and relationships between them
- Real-world applications of AI
- 📌 Work on real-world examples of AI or interactive quiz

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

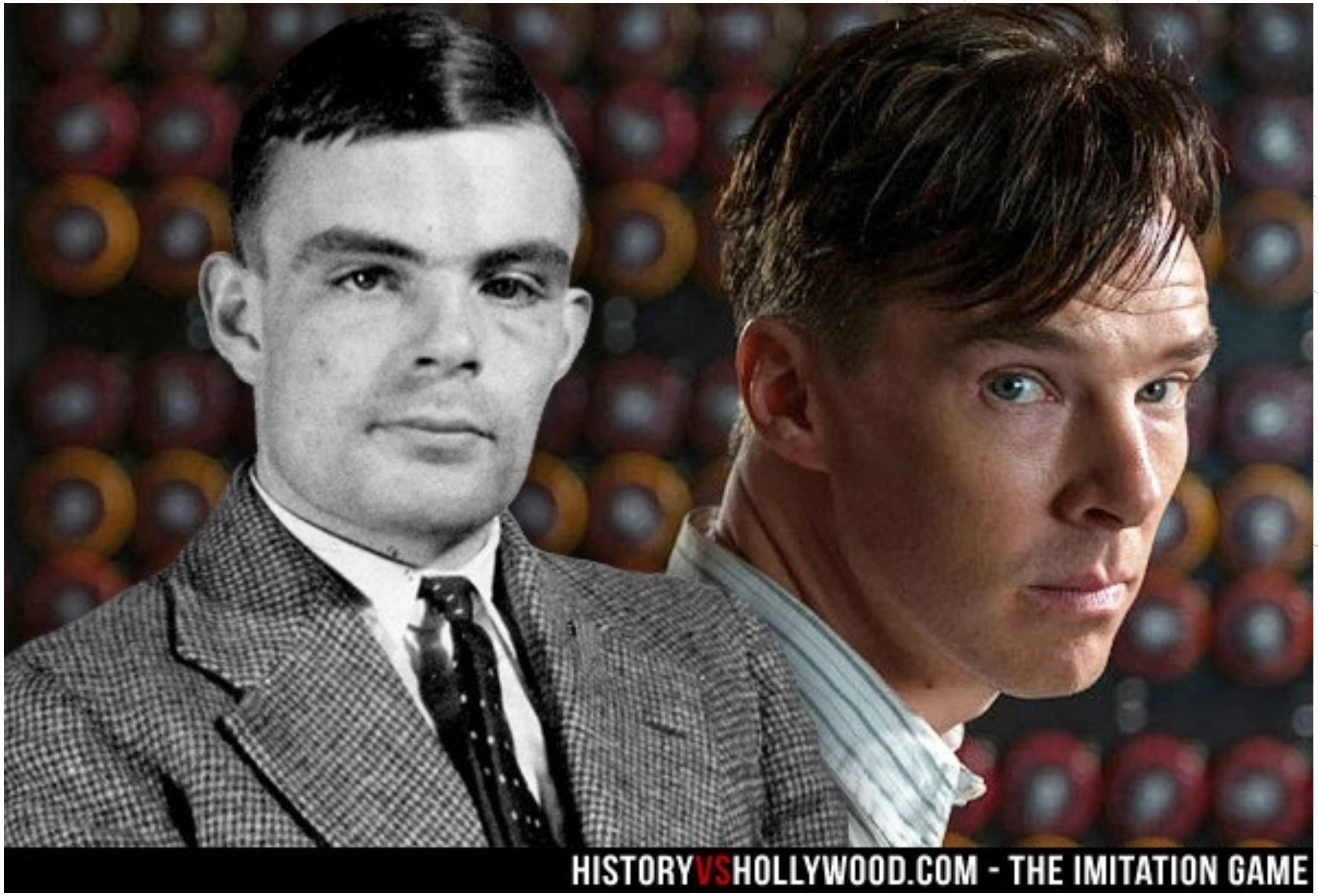
Ministero dell'Università e della Ricerca

Italiadomani

The Artificial Intelligence evolution

- In 1956 the term Artificial Intelligence was coined by John McCarthy
- *The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.*

The Turing test



The Turing test

Turing test

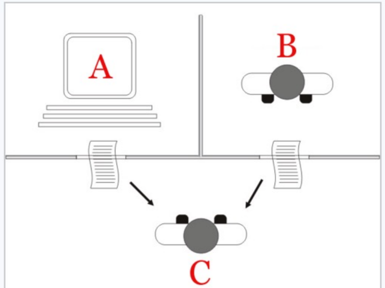
64 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

"Imitation game" redirects here. For the film, see The Imitation Game. For other uses, see Turing test (disambiguation).

The **Turing test**, originally called the **imitation game** by Alan Turing in 1949,[2] is a test of a machine's ability to exhibit intelligent behaviour equivalent to that of a human. In the test, a human evaluator judges a text transcript of a natural-language conversation between a human and a machine. The evaluator tries to identify the machine, and the machine passes if the evaluator cannot reliably tell them apart. The results would not depend on the machine's ability to answer questions correctly, only on how closely its answers resembled those of a human. Since the Turing test is a test of indistinguishability in performance capacity, the verbal version generalizes naturally to all of human performance capacity, verbal as well as nonverbal (robotic).[3]

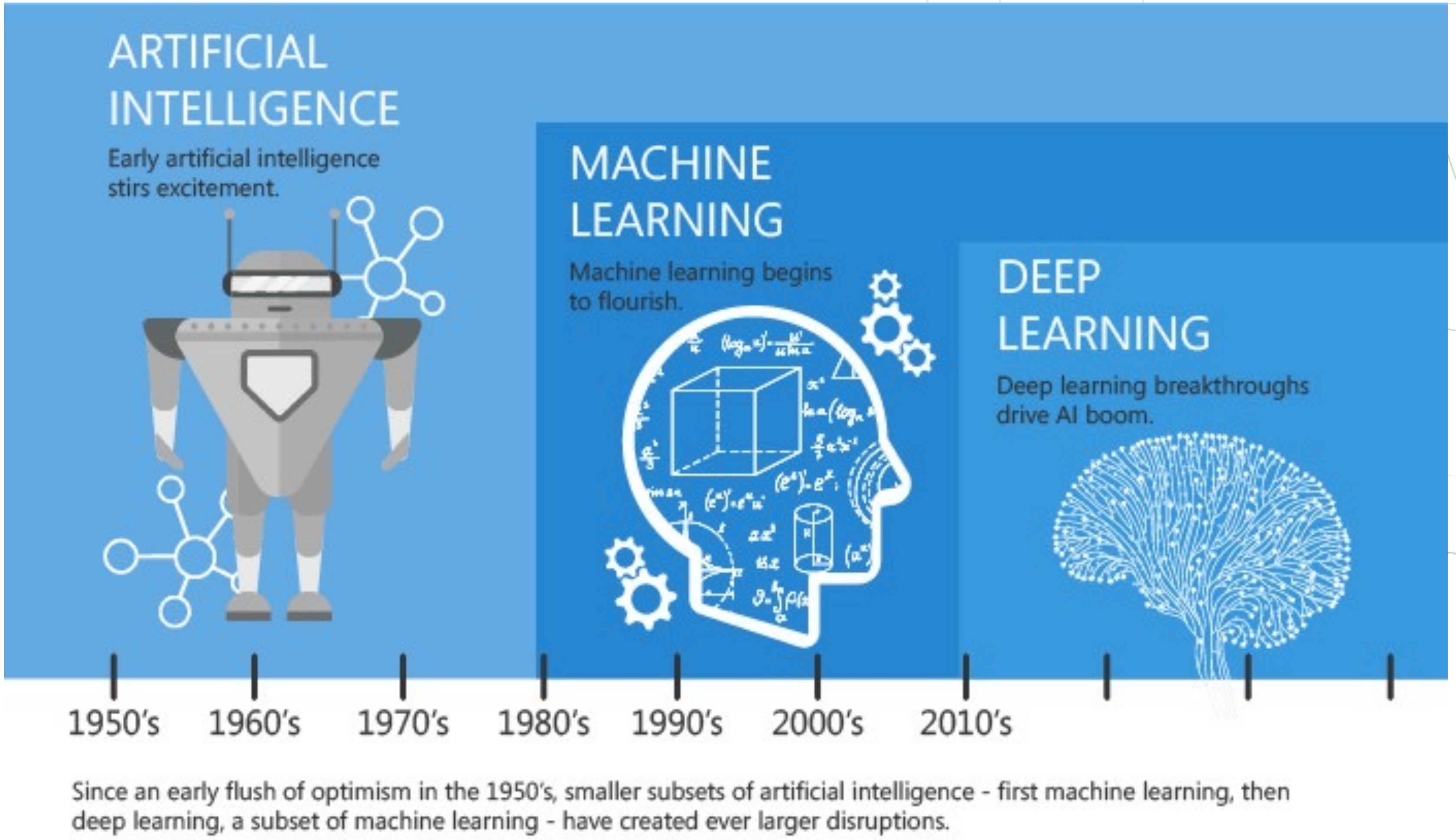


The "standard interpretation" of the Turing test, in which player C, the interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination.[1]

The test was introduced by Turing in his 1950 paper "Computing Machinery and Intelligence" while working at the University of Manchester.[4] It opens with the words: "I propose to consider the question, 'Can machines think?'" Because "thinking" is difficult to define, Turing chooses to "replace the question by another, which is closely related to it and is expressed in relatively unambiguous words".[5] Turing describes the new form of the problem in terms of a three-person party game called the "imitation game", in which an interrogator asks

Part of a series on
Artificial intelligence (AI)

The AI evolution



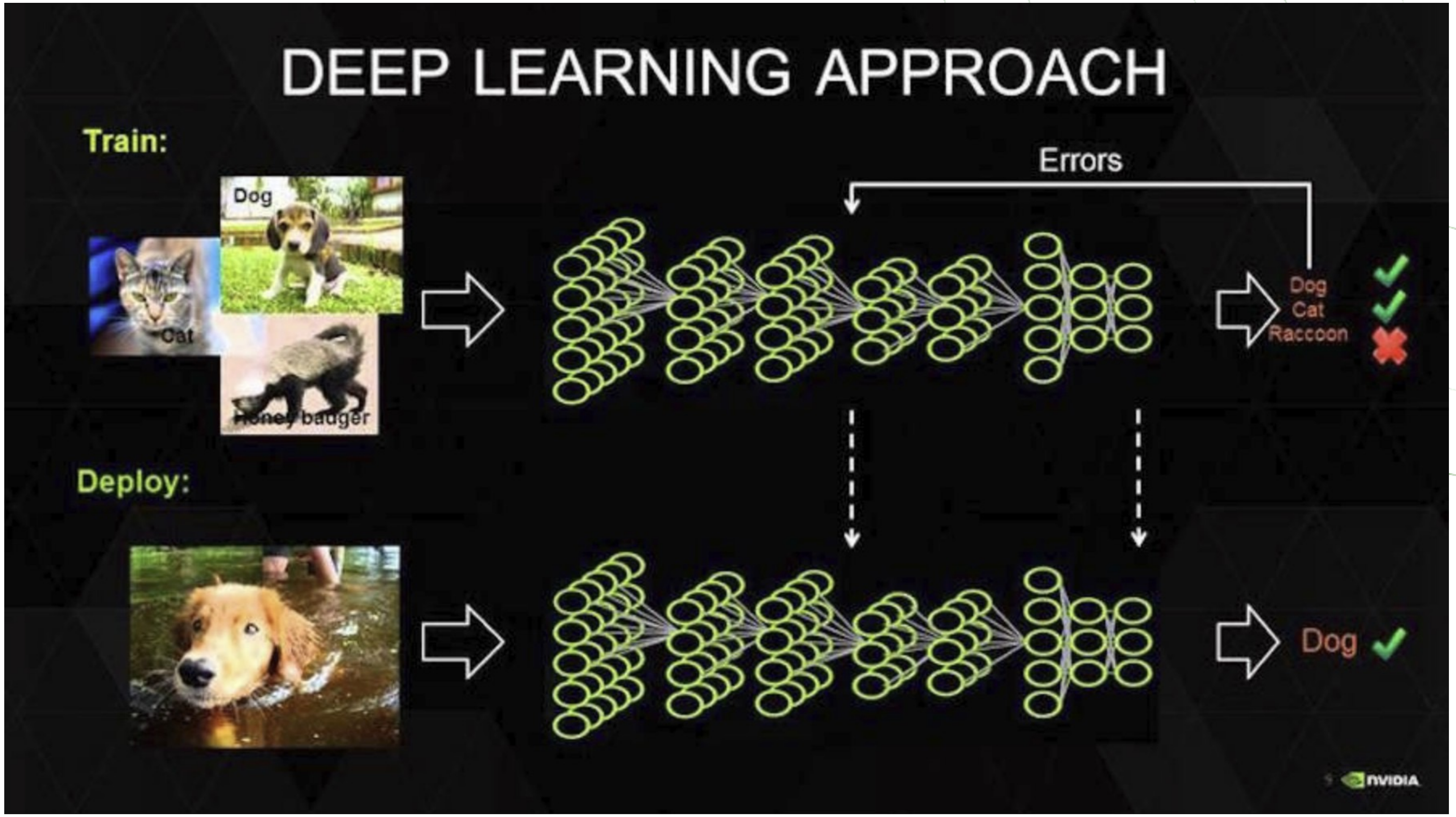
A new vision of Artificial Intelligence today

- Before: Intelligence was hardcoded into machines
- Today: Machines learn by observing Big Data
- BigData -> AI

The old vs the new school

- In the past, many attempts to make machines ”Intelligent”: Expert systems, Artificial Intelligence, etc.
- Today, Big data/Artificial Intelligence is about deriving math models (insights) from huge data bases
- Being able to observe and learn models leads to intelligent behaviour
  - IBM Watson
  - AlphaGo

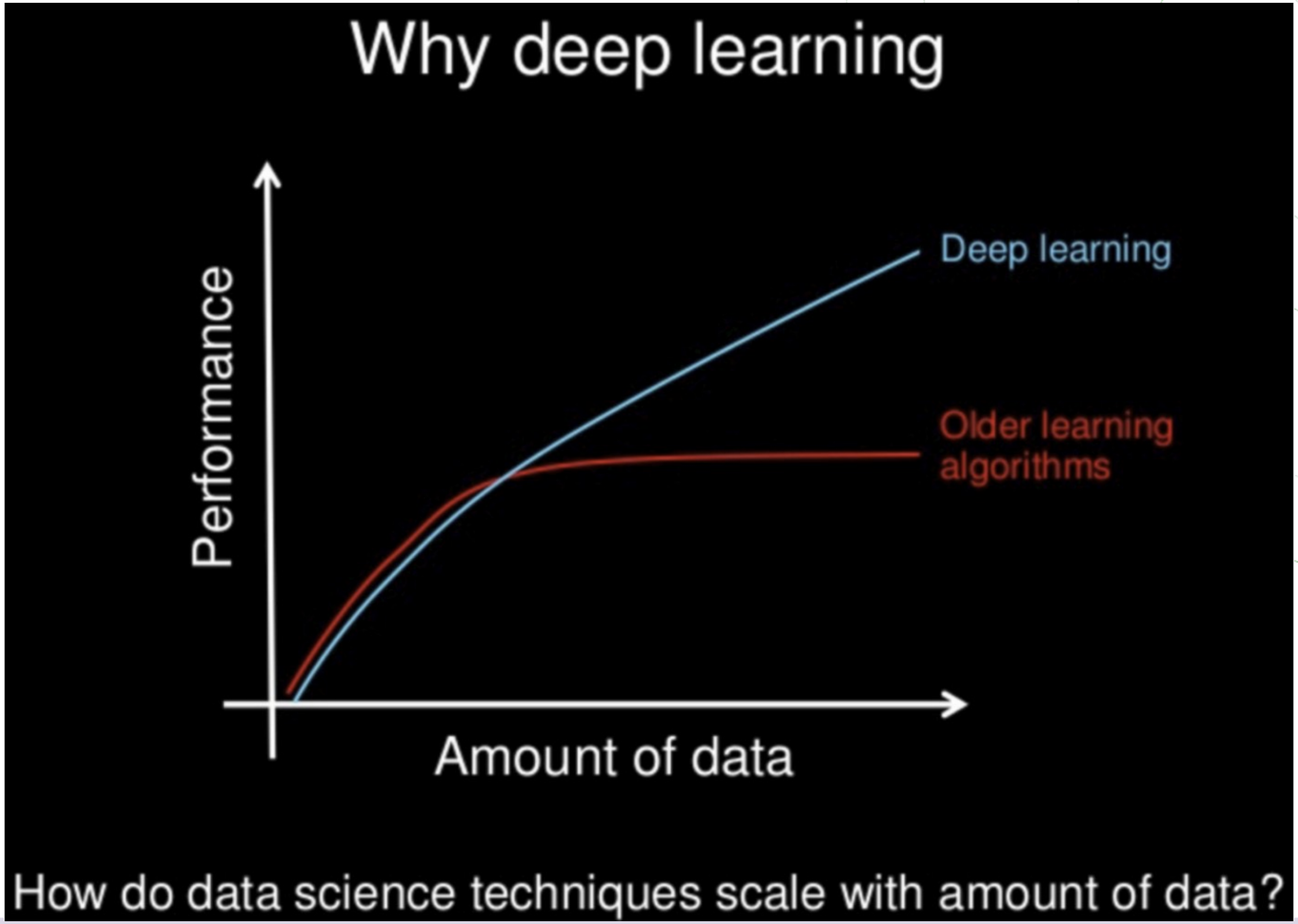
Deep Learning



Self learning example

- Learning to walk

Learning from data



In 1997

Alpha GO

- 3000 years old game
- Simple board
- Before 2016 it was considered to be impossible to model
- Many (many) more combinations compared to chess
- It was said:
  - ”the most elegant game that humans have ever invented”;
  - ”simple rules that give rise to endless complexity”;
  - ”more possible Go positions than there are atoms in the universe”
- Mostly based on intuition

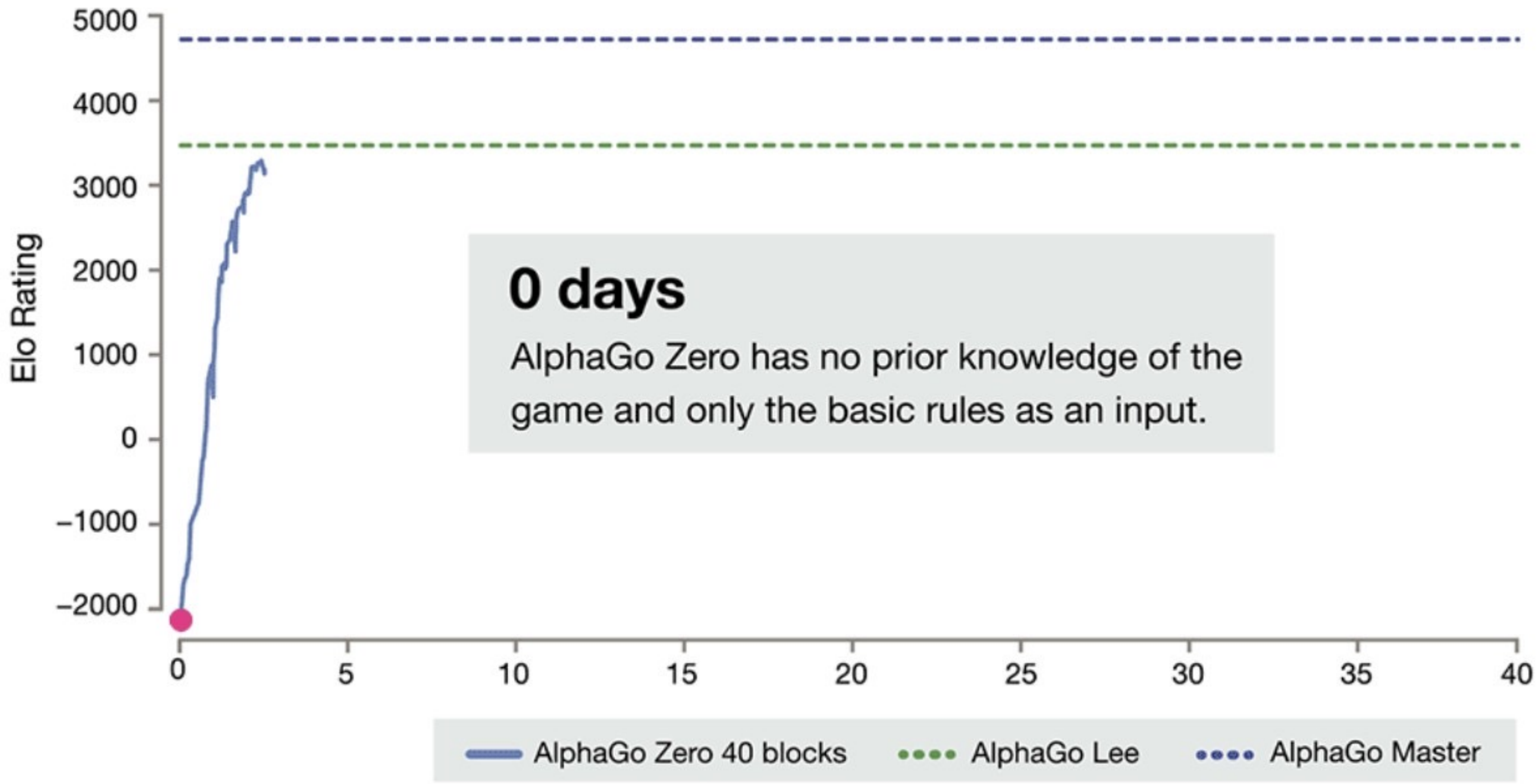
In 2016

It gets better

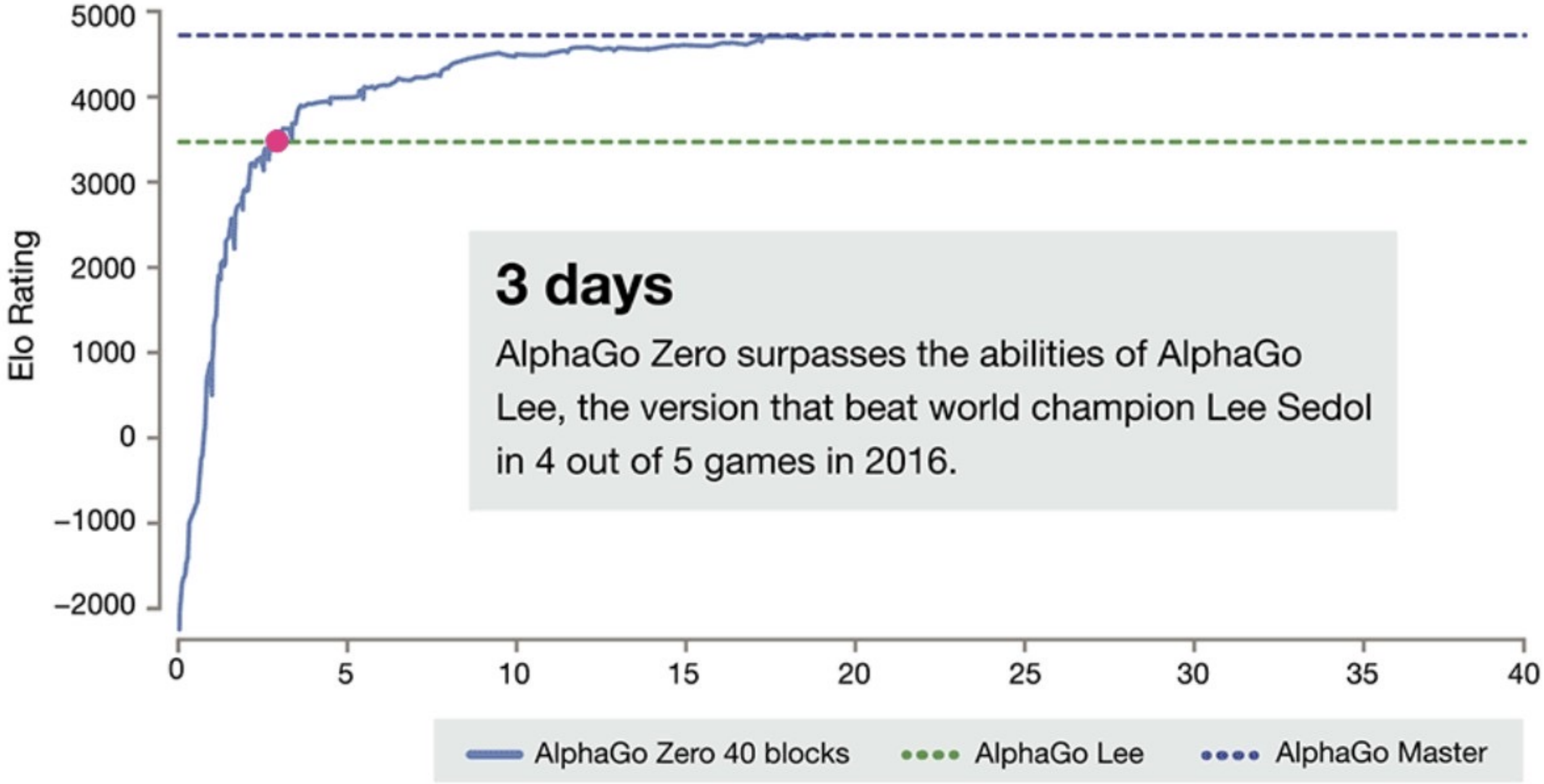
- In 2018 AlphaGo-Zero
- A new version based on Deep Learning techniques

Previous versions of AlphaGo initially trained on thousands of human amateur and professional games to learn how to play Go. AlphaGo Zero skips this step and learns to play simply by playing games against itself, starting from completely random play. In doing so, it quickly surpassed human level of play and defeated the previously published champion-defeating version of AlphaGo by 100 games to 0.

At the beginning



After 3 days



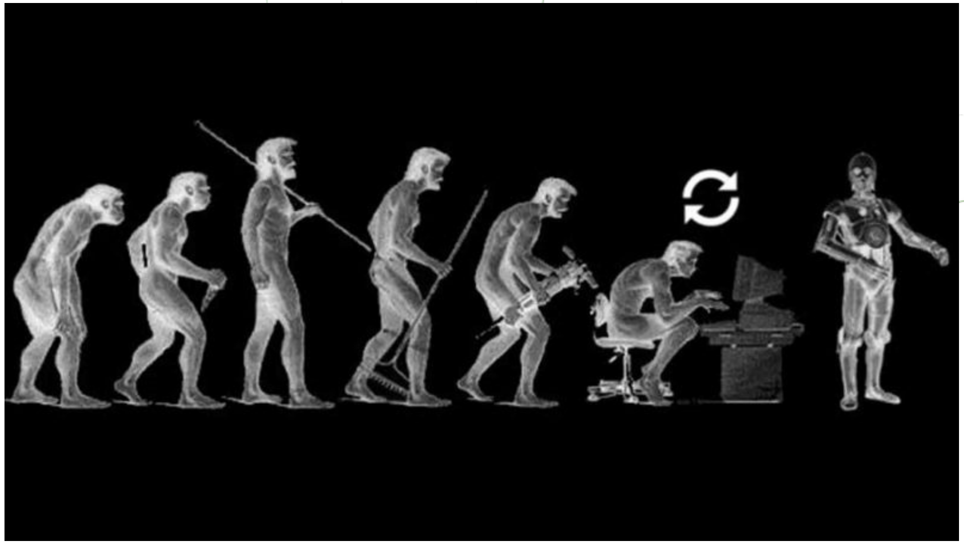
5000
4000
3000
2000
1000
0
−1000
−2000
Elo Rating
0
5
10
15
20
25
30
35
40
3 days
AlphaGo Zero surpasses the abilities of AlphaGo Lee, the version that beat world champion Lee Sedol in 4 out of 5 games in 2016.
AlphaGo Zero 40 blocks
AlphaGo Lee
AlphaGo Master

After 21 days

After 40 days

The old vs the new school

- CYC vs Watson
- Two (very) different approaches
- CYC was “embedding” knowledge
- Watson is able to “learn” from huge amount of data



CYC

ITINERIS

Cyc

From Wikipedia, the free encyclopedia

For other uses, see CYC (disambiguation).

Cyc (/ˈsaɪk/) is the world's longest-lived artificial intelligence project,[citation needed] attempting to assemble a comprehensive ontology and knowledge base that spans the basic concepts and "rules of thumb" about how the world works (think common sense knowledge but focusing more on things that rarely get written down or said, in contrast with facts one might find somewhere on the internet or retrieve via a search engine or Wikipedia), with the goal of enabling AI applications to perform human-like reasoning and be less "brittle" when confronted with novel situations that were not preconceived.

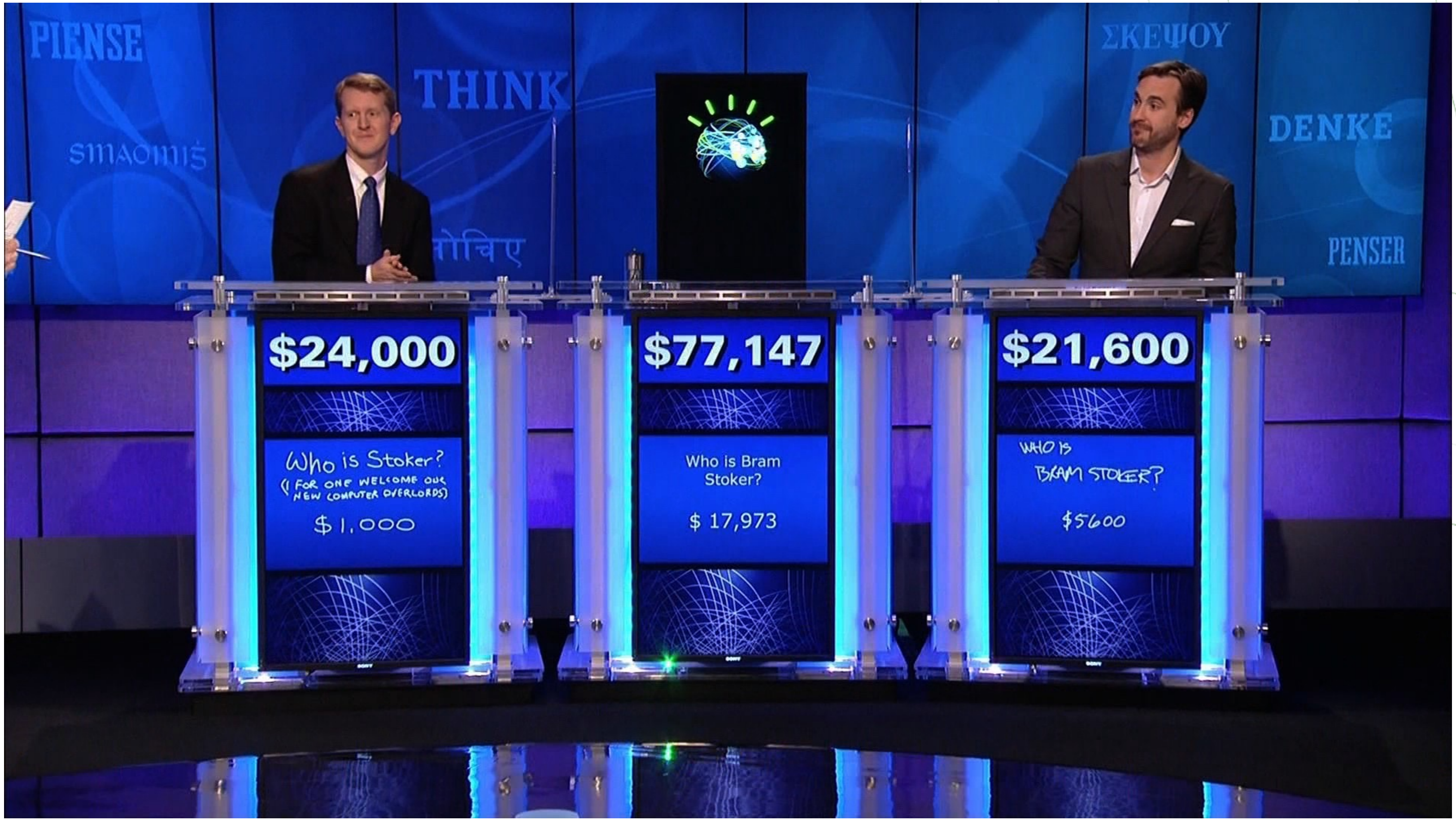
Douglas Lenat began the project in July 1984 at MCC, where he was Principal Scientist 1984–1994, and then, since January 1995, has been under active development by the Cycorp company, where he is the CEO.

Cyc

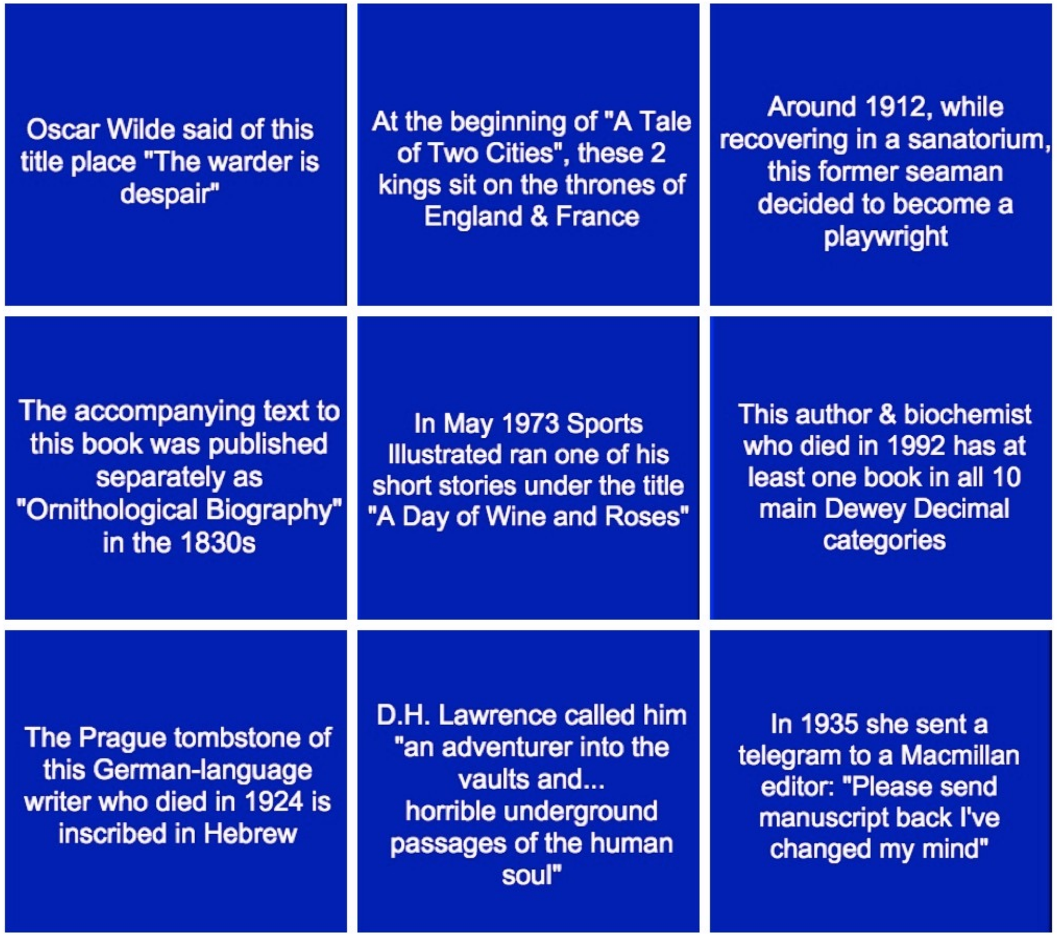
Original author(s)	Douglas Lenat
Developer(s)	**Cycorp, Inc.**
Initial release	1984; 35 years ago
Stable release	6.1 / 27 November 2017; 15 months ago
Written in	Lisp, CycL
Type	Ontology and Knowledge Base and Knowledge Representation Language and Inference engine
Website	www.cyc.com

Contents [hide]

Watson


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Jeopardy



IBM Watson vs. Ken Jennings & Brad Rutter
📅 **Date**: February 14–16, 2011
📍 **Show**: *Jeopardy!* (special three-day exhibition match)

Watson at work

WIRED

Technology | Science | Culture | Gear | Business |

Credit **IBM**

IBM's Watson -- the language-fluent computer that beat the best human champions at a game of the US TV show *Jeopardy!* -- is being turned into a tool for medical diagnosis. Its ability to absorb and analyse vast quantities of data is, IBM claims, better than that of human doctors, and its deployment through the cloud could also reduce healthcare costs.

Watson at work

Two years ago, IBM announced that Watson had "learned" the same amount of knowledge as the average second-year medical student. For the last year, IBM, Sloan-Kettering and Wellpoint have been working to teach Watson how to understand and accumulate complicated peer-reviewed medical knowledge relating to oncology. That's just lung, prostate and breast cancers to begin with, but with others to come in the next few years). Watson's ingestion of more than 600,000 pieces of medical evidence, more than two million pages from medical journals and the further ability to search through up to 1.5 million patient records for further information gives it a breadth of knowledge no human doctor can match.

Watson at work

ITINERIS

According to Sloan-Kettering, only around 20 percent of the knowledge that human doctors use when diagnosing patients and deciding on treatments relies on trial-based evidence. It would take at least 160 hours of reading a week just to keep up with new medical knowledge as it's published, let alone consider its relevance or apply it practically. Watson's ability to absorb this information faster than any human should, in theory, fix a flaw in the current healthcare model. Wellpoint's Samuel Nessbaum has claimed that, in tests, Watson's successful diagnosis rate for lung cancer is 90 percent, compared to 50 percent for human doctors.

What is Artificial Intelligence (AI)?

- Many Interpretations
- Artificial General Intelligence (aka. Strong AI or Full AI)
  - General intelligent actions
  - Discerning problems
  - Acting as humans would
  - ... up to self-consciousness
- Restricted AI (aka. Weak or Applied AI or Narrow AI)
  - Implementing intelligence in specific applications or problem solving tasks
  - No full cognitive abilities

What Is Artificial Intelligence (AI)?

- Mimics human intelligence
- Applications: **speech recognition**, **robotics**, **decision-making**
- Types: Narrow, General, Superintelligence

What Is Machine Learning (ML)?

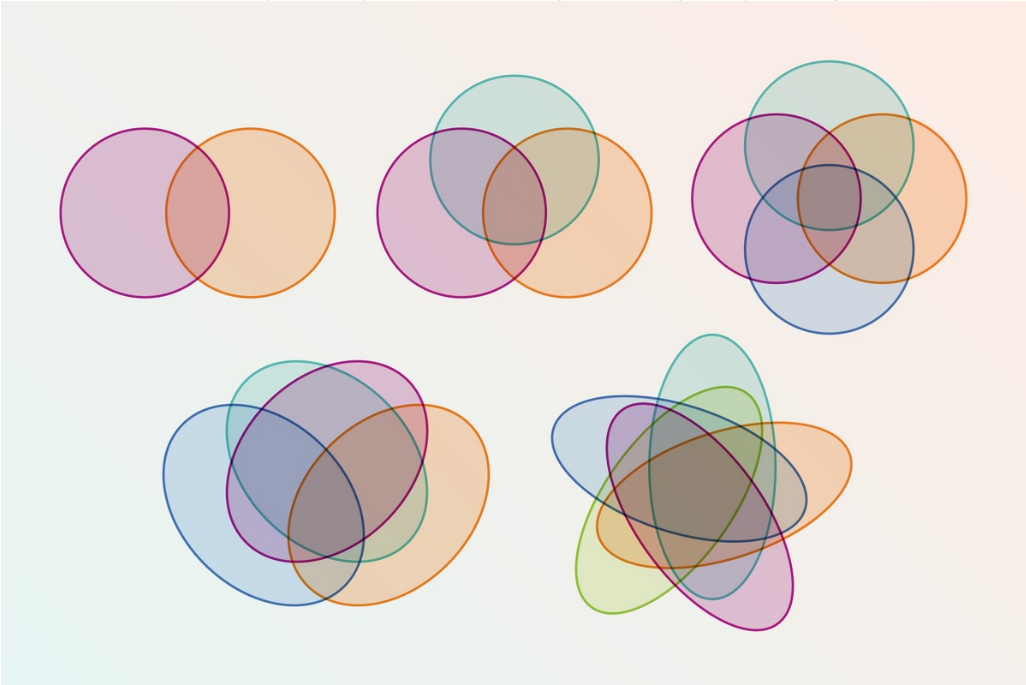
- Subset of AI
- Learns patterns from data
- Example: teaching a machine to fish

What Is Deep Learning (DL)?

- Subset of ML
- Uses neural networks with many layers
- Works best with large datasets

The Relationship: AI > ML > DL

- Hierarchical structure
- DL is part of ML, which is part of AI



Key Differences Between AI, ML, and DL

- Scope
- Data needs
- Algorithm type
- Hardware requirements

Real-World Applications of AI

- Healthcare: diagnosis, drug discovery
- Finance: fraud detection
- Retail: recommendations, chatbots
- Manufacturing: predictive maintenance

Deep Learning in Action

- Self-driving cars
- Face recognition
- Chatbots like ChatGPT
- Image generation (e.g., DALL·E)

Myths and Misconceptions

- AI is not conscious
- DL isn't always the best option
- Data is key

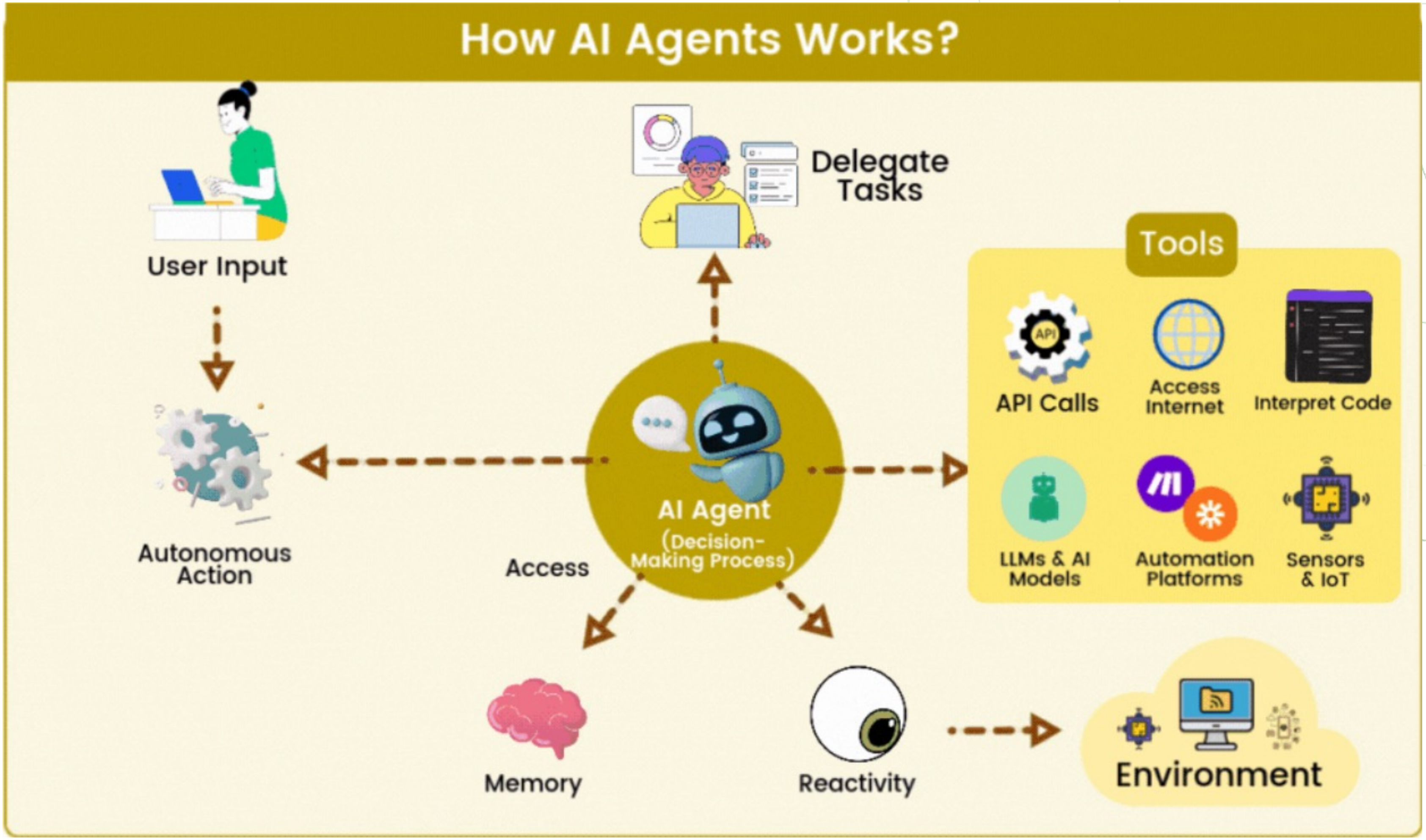
Challenges and Considerations

- Data quality and bias
- Explainability
- Ethics and responsible AI

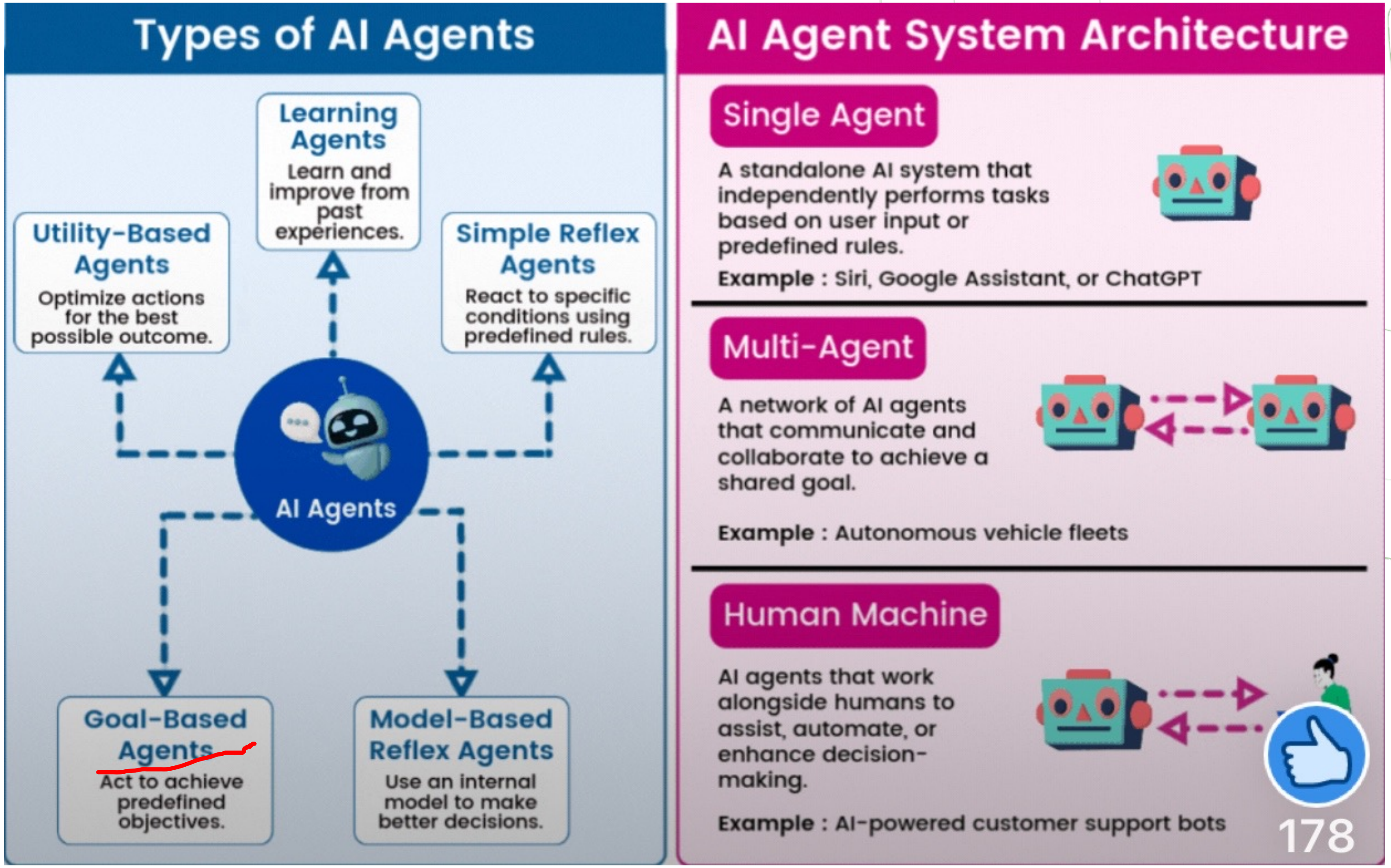
Future Trends

- Generative AI
- AI for everyone (low-code/no-code)
- Autonomous systems and AGI

What is an AI Agent ?



What is an AI Agent ?



Summary and Takeaways

- AI, ML, DL defined and compared
- How they relate
- Real-world impact

Module 2: Types of Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Use cases for each learning type
- 📌 Demo: Teachable Machine (e.g., image classification)
- 📌 Discussion: Which learning type fits which kind of problem?

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell’Università e della Ricerca
Italiadomani

Why Learning Types Matter

- They define how models learn from data
- Help select the right approach for a task
- Real-world relevance and use cases

What Is Supervised Learning?

- Learns from labelled data
- Predicts outcomes based on input-output pairs
- Examples: classification, regression

How Supervised Learning Works

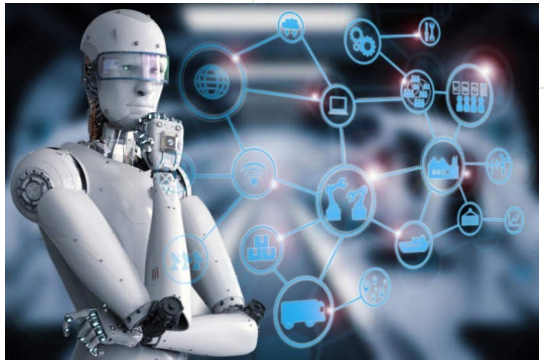
- Training with input-output examples
- Objective: minimize prediction error
- Requires a labelled dataset

Use Cases of Supervised Learning

- Email spam detection
- Fraud detection in finance
- Medical image classification
- Sales forecasting

Supervised learning

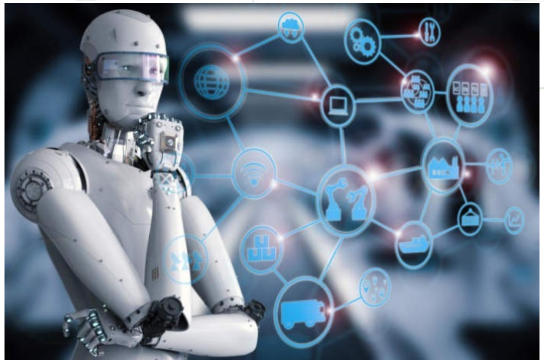
- Data are labelled
- Labels are the targets (or output, or class): what we want to learn
- So, for each observation we have:
  - Input values
  - Label



Supervised learning

- Data are labelled
- Labels are the targets (or output, or class): what we want to learn
- So, for each observation we have:
  - Input values
  - Label

The machine learning algorithm learns such associations over time



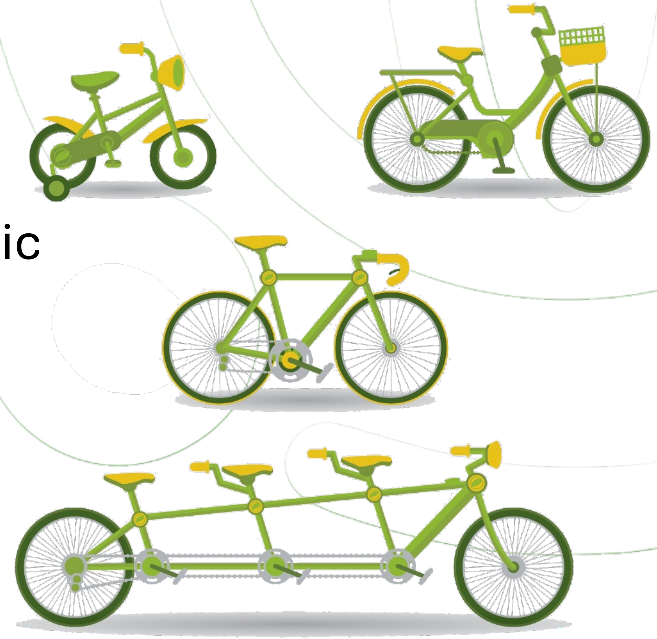
Supervised learning - example

- We want to teach a small kid how to distinguish a bike from a car
- He has not ever seen those before

Input = *a set of labelled images*

Supervised learning - example

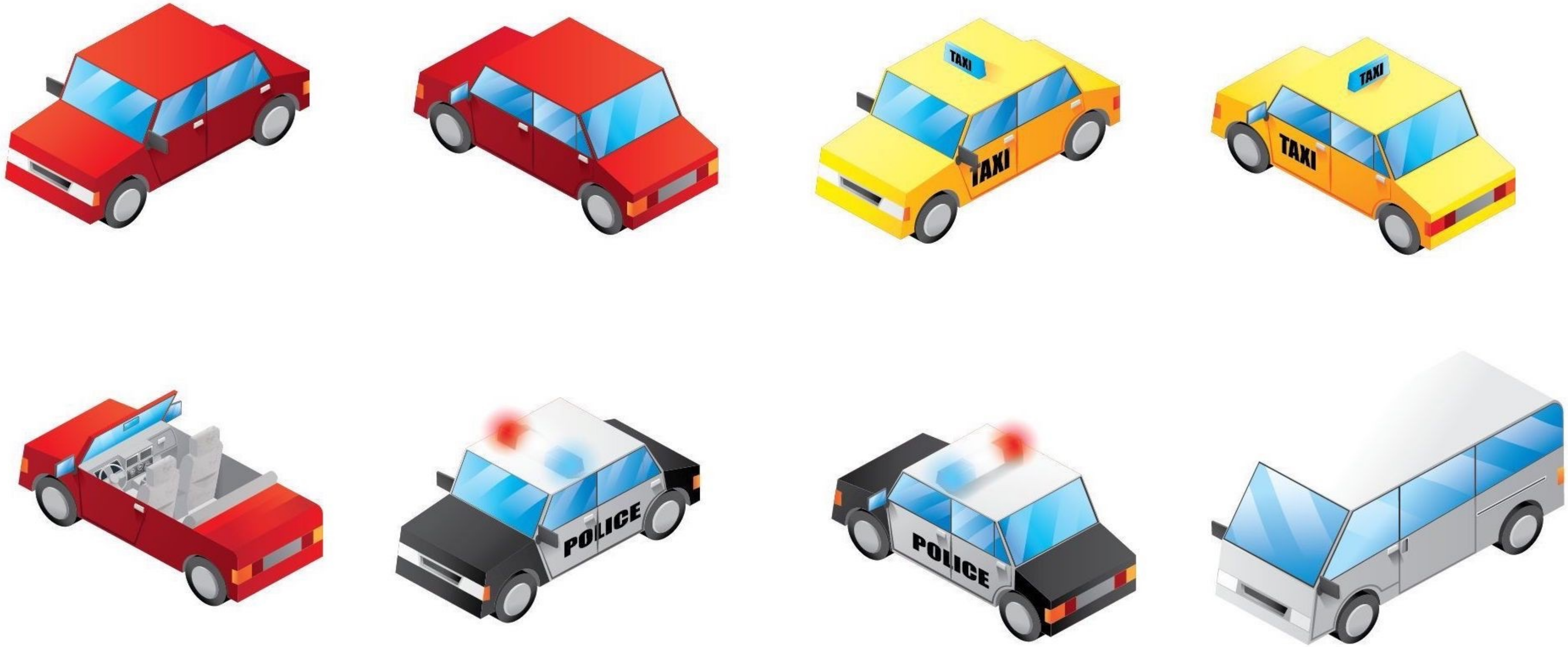
- Let’s proceed as follows:
  - Let’s show the images of the bikes
  - We tell him those are ”bikes”
  - We do not teach him about any specific characteristic



So we let the kid analyse those images to understand what makes those objects a “bike”

Supervised learning - example

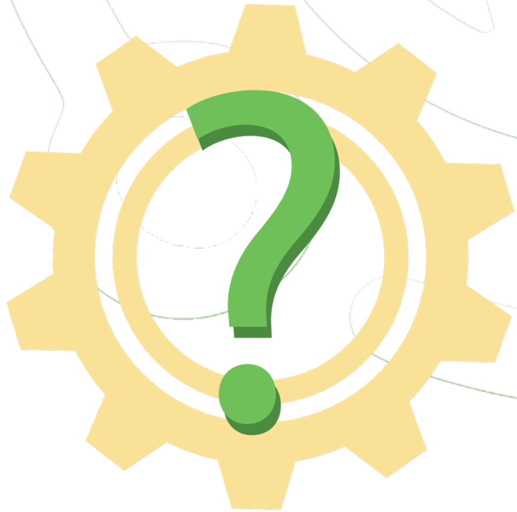
- We do the same with the cars



We let him “think and learn”

Supervised learning - example

- Eventually, we show him a picture and ask him to identify it



Notice: It’s a new picture, he has not seen it before

What Is Unsupervised Learning?

- Learns patterns from unlabelled data
- No predefined output
- Examples: clustering, dimensionality reduction

How Unsupervised Learning Works

- Finds structure in data
- Groups similar items or reduces data complexity
- Often used for exploration

Use Cases of Unsupervised Learning

- Customer segmentation
- Anomaly detection
- Market basket analysis
- Topic modelling in documents

Unsupervised learning

- Here the algorithm learns without any label
- The input to the algorithm is just a set of observations

In general, this is a more challenging class of problems

Unsupervised learning - Example

- Let’s repeat the previous example with no supervision
- This time we show the kid the images at once, bikes and cars together
- **We don’t tell him anything about the two type of objects!**

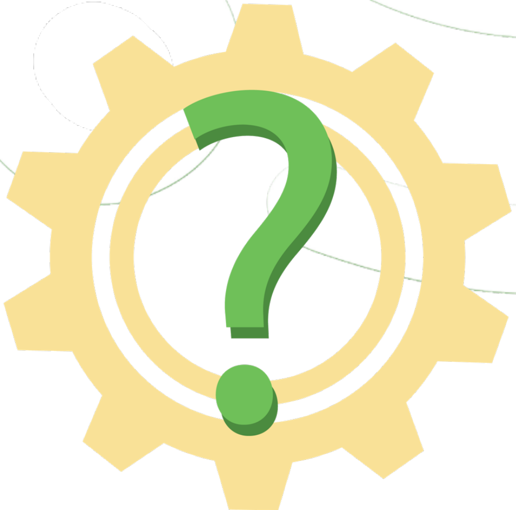
Unsupervised learning

- The kid has to learn by himself the two categories and what makes those different from each other

He will use a different logical path to cluster the input images

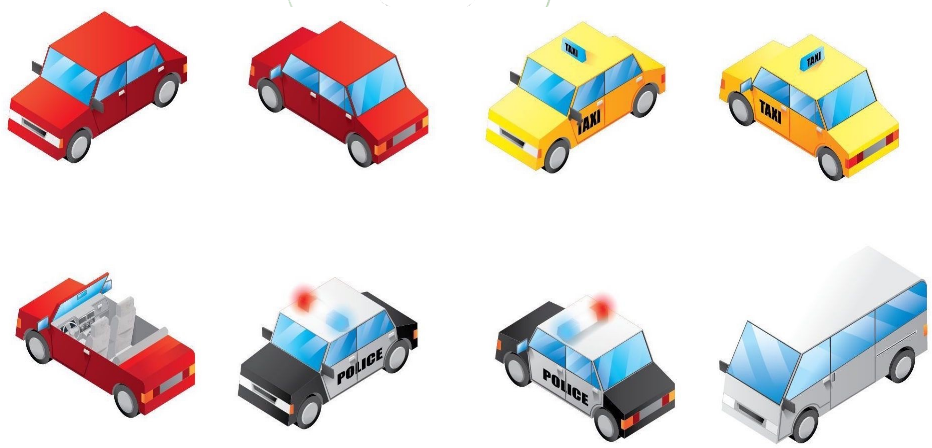
Unsupervised learning - Example

- Then, like before, we show him a new unseen image



Unsupervised learning - Considerations

- The kid in his learning may use more than two categories or a very different set of categories of what we expect
- For instance, he may decide to put together objects based on color, size, or number of wheels (that he sees!)



Unsupervised learning - Considerations

- The results is greatly dependent upon the quality of the input images
- As usual, the more data in input the more accurate the learning, at least, until a certain point

What Is Reinforcement Learning?

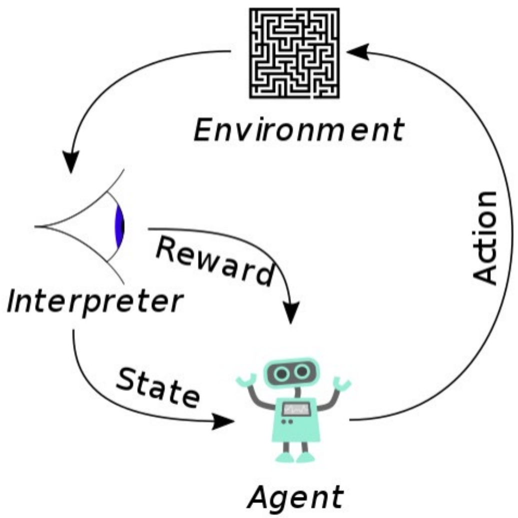
- Learns by interacting with an environment
- Receives rewards or penalties
- Examples: game AI, robotics

How Reinforcement Learning Works

- Agent takes actions to maximize cumulative reward
- Trial-and-error learning
- Uses policy, value functions

Reinforcement learning - The basic principle

- Learning is based on a gain function
- Each time the machine reaches a positive state it gains something
- The objective is to maximize gain



Use Cases of Reinforcement Learning

- Used by DeepMind to develop AlphaGo
- Game playing (e.g., AlphaGo)
- Autonomous vehicles
- Dynamic pricing
- Industrial automation

Supervised Learning

ITINERIS

- Task to be learnt: to extract a description / labelling or pattern from the data, based on the training
- Training examples labelled by a (human) supervisor
- Use it to predict the output for further examples
- Performance measured as how accurate the output is
- Example applications
  - Credit approval
  - Medical diagnosis Fraud detection
  - Text and image labelling or classification

Unsupervised Learning

- Task to be learnt: finding interesting patterns/ groups/ categories in the data based on evidence
- No pre-labeled data detection of facts from raw data
- Performance measured as how good / meaningful the groups / patterns are
- Example applications
  - Customer segmentation
  - User behavior categorization
  - Grouping of items by similarity

Reinforcement Learning

- The machine learns through trial-and-error interactions
- The goal is to maximize the amount of reward received from the environment
- Iterative process
  - Trained through interactions with the environment
  - Rewards assigned upon success
  - Performance measured as amount of rewards collected
- Example applications
  - Robot learning
  - Games

Summary and Comparison

- **Supervised**: labelled data, known outputs
- **Unsupervised**: no labels, finds hidden patterns
- **Reinforcement**: learns via rewards and penalties

Module 3: Datasets, Algorithms, and Models (75)

- What is a dataset: structure and data quality
- Concepts: features, labels, training/test sets
- Intro to popular algorithms: linear regression, decision trees, k-means
- 📌 Demo: Google Colab – Basic example using linear regression
- 📌 Activity: Guided questions + small group discussion

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Why Data Matters in ML

- Algorithms are only as good as the data they use
- Data quality drives performance
- Foundation for training accurate models

What Is a Dataset?

- Structured collection of data
- Rows = instances or samples
- Columns = features or attributes
- May include a target variable (label)

Dataset Structure and Quality

- Features: measurable properties
- Labels: known outcomes (for supervised learning)
- Importance of clean, complete, and relevant data

Training and Test Sets

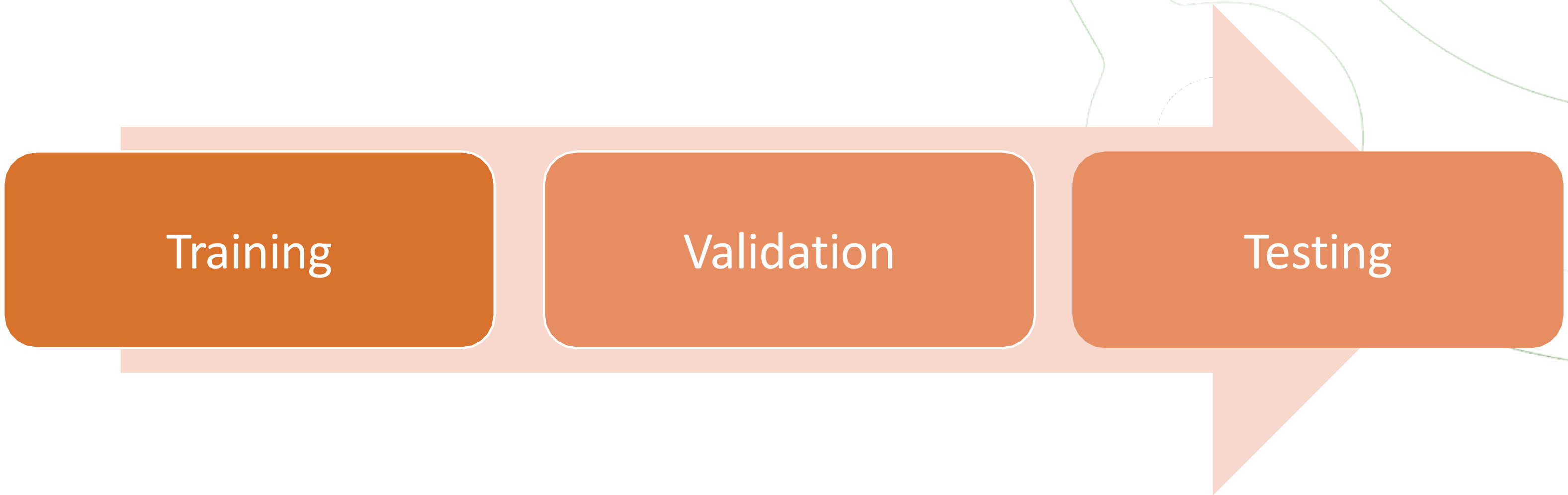
- **Training set**: used to train the model
- **Test set**: used to evaluate performance
- Often split: 70/30 or 80/20

Common Dataset Pitfalls

- Missing values
- Noisy or inconsistent data
- Data leakage
- Imbalanced classes

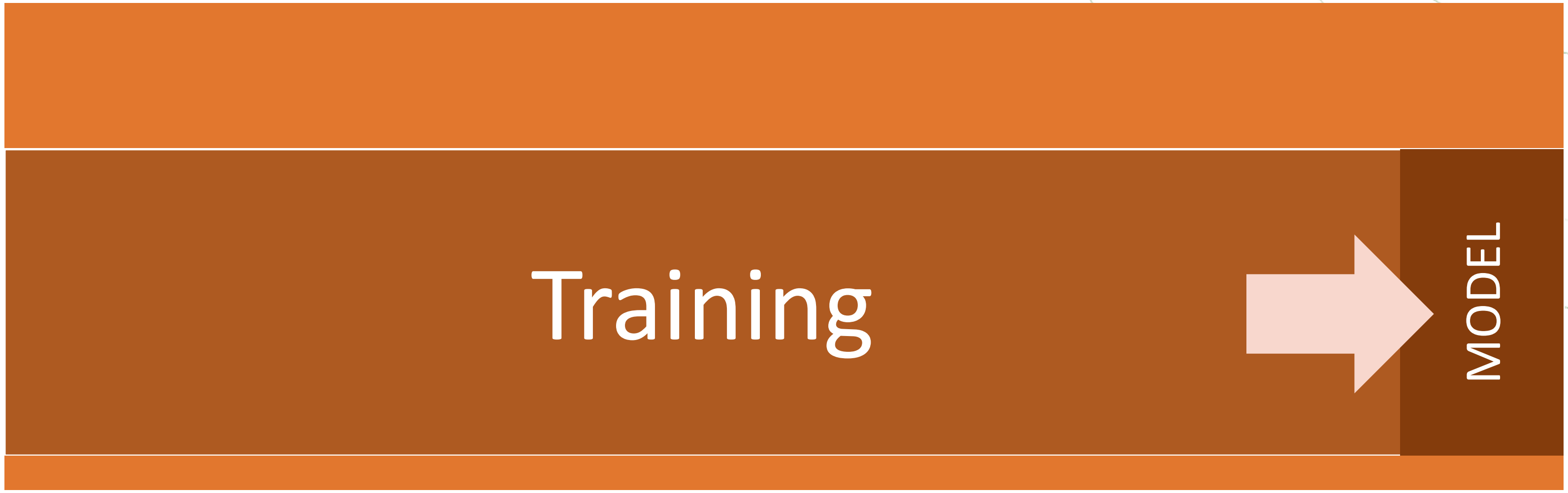
Training, Validation, and Test

- Split process into: training, validation, and test



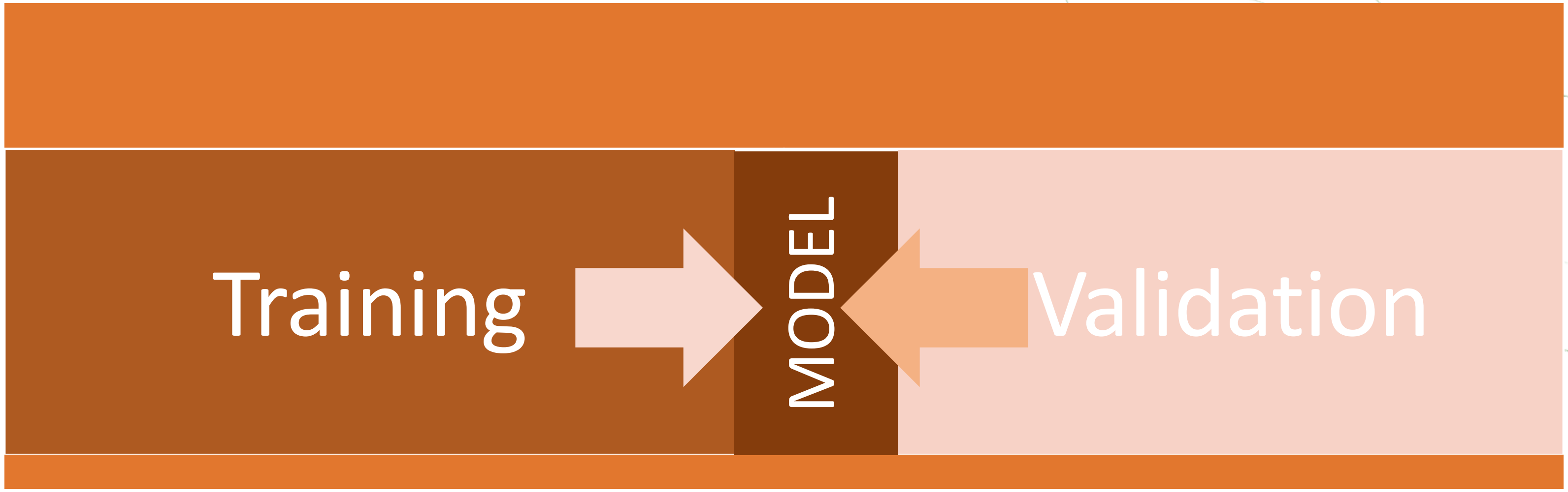
Training

- Learn the model over the training set



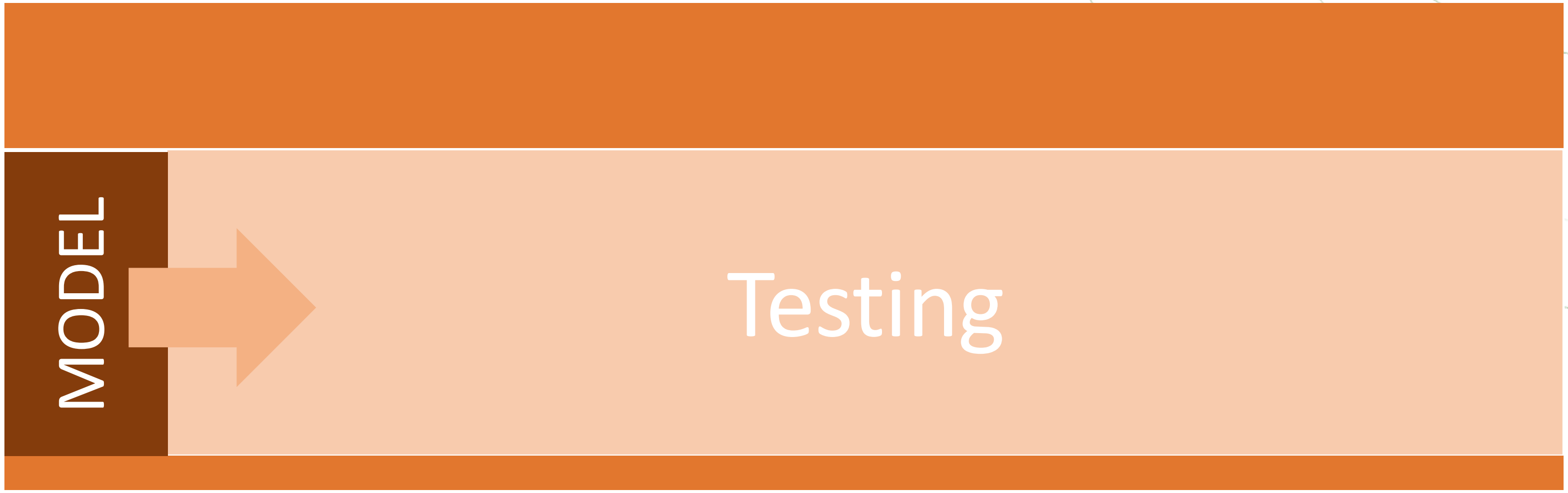
Validation

- Optimize the predictive capability of the model using the validation set



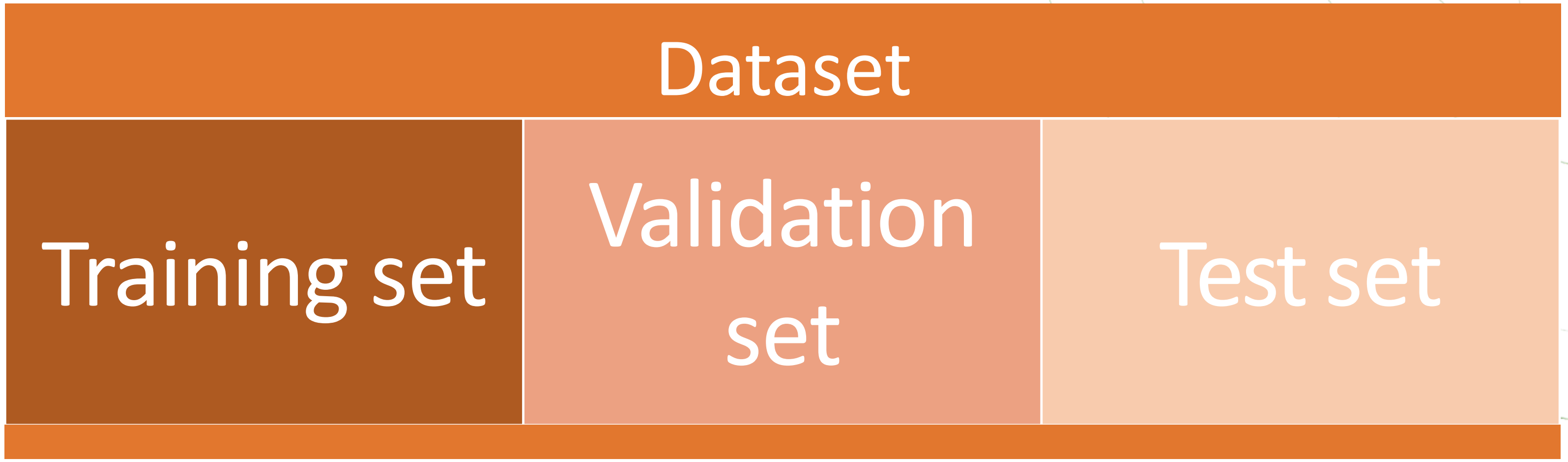
Testing

- See how well the model works on the test set (unseen before)
- The error gives an unbiased estimate of the predictive power of a model



Training, Validation, and Test

- Split data into three sets: training, validation, and test



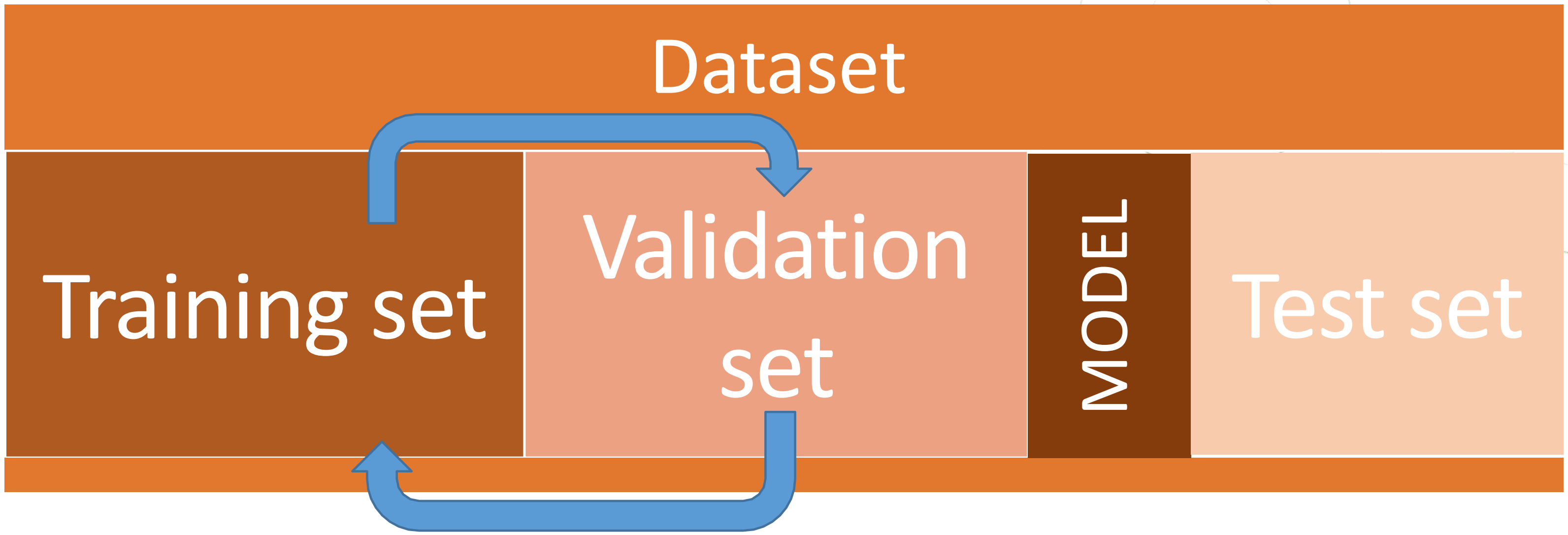
Training, Validation, and Test

- Split data into three sets: training, validation, and test

Training set	Validation set	Test set
70%	15%	15%

Cross-Validation

- Repeat the iteration on training + validation multiple times
- 10-fold cross-validation: pick 10 times random subsets as training and validation, and average the quality of the results



Intro to Algorithms: **Linear Regression**

- Predicts **continuous** values
- Models relationships between features and target
- Simple, interpretable model

Intro to Algorithms: **Decision Trees**

- Tree-like structure for decisions
- Easy to understand and visualize
- Handles classification and regression

Intro to Algorithms: **K-Means**

- Unsupervised clustering algorithm
- Groups data into k clusters
- Useful for segmentation tasks

Supervised Learning Models

Model	Description
Linear Regression	Predicts a continuous value using a linear relationship between input and output.
Logistic Regression	Used for binary classification (yes/no, 0/1).
Decision Trees	Tree-like models used for both classification and regression tasks.
Random Forest	Ensemble of decision trees for more robust predictions.
Support Vector Machines (SVM)	Finds the best boundary between classes. Effective in high-dimensional spaces.
K-Nearest Neighbors (K-NN)	Classifies a sample based on the majority class of its 'k' nearest neighbors.
Naive Bayes	Probabilistic model based on Bayes’ theorem, good for text classification.
Gradient Boosting Machines (GBM)	Powerful ensemble model that builds trees sequentially (e.g., XGBoost, LightGBM).
Neural Networks	Highly flexible models that mimic the human brain, used in deep learning.

(Used when the data has labeled outputs)

Unsupervised Learning Models

Model	Description
K-Means Clustering	Groups data into ‘k’ clusters based on feature similarity.
Hierarchical Clustering	Builds a tree of clusters (dendrogram) for hierarchical grouping.
DBSCAN	Density-based clustering for discovering clusters of varying shapes.
Principal Component Analysis (PCA)	Dimensionality reduction technique that keeps variance.
t-SNE / UMAP	Non-linear dimensionality reduction for visualization.
Autoencoders	Neural network-based model for data compression and reconstruction.

(Used when the data has no labels)

Reinforcement Learning Models

Model	Description
K-Means Clustering	Groups data into ‘k’ clusters based on feature similarity.
Hierarchical Clustering	Builds a tree of clusters (dendrogram) for hierarchical grouping.
DBSCAN	Density-based clustering for discovering clusters of varying shapes.
Principal Component Analysis (PCA)	Dimensionality reduction technique that keeps variance.
t-SNE / UMAP	Non-linear dimensionality reduction for visualization.
Autoencoders	Neural network-based model for data compression and reconstruction.

(Learning via interaction and feedback in an environment)

Other Important Models / Techniques

Model	Description
Ensemble Methods	Combines multiple models (e.g., Bagging, Boosting, Stacking).
Time Series Models	e.g., ARIMA, LSTM (for forecasting).
Transformer Models	State-of-the-art models in NLP (e.g., BERT, GPT).
GANs (Generative Adversarial Networks)	Generates new data similar to training data (used in image generation).

Choosing the Right Algorithm

- Depends on task: classification, regression, clustering
- Consider data size, type, quality
- Trade-off: accuracy vs interpretability

Summary and Q&A

- Datasets = fuel for ML
- Features, labels, train/test split
- Intro to core algorithms: **regression**, **trees**, **clustering**

Bias

Over the Town
Marc Chagall (1918)

Module 4: Building a Simple ML Model

- ML pipeline: problem definition, algorithm selection, training, evaluation
- Overfitting/underfitting, model validation basics
- 📌 Live Demo: Step-by-step model building in Colab using a simple dataset

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell’Università e della Ricerca

Italiadomani

Link

- https://drive.google.com/drive/folders/1co_PgxnccqMLkXadfZtHjpF5cBbJR6ms?usp=sharing
- Link … https://shorturl.at/5Blm2

Pattern recognition

- The human brain… the most powerful pattern recognition machine?

What’s a “chair”?

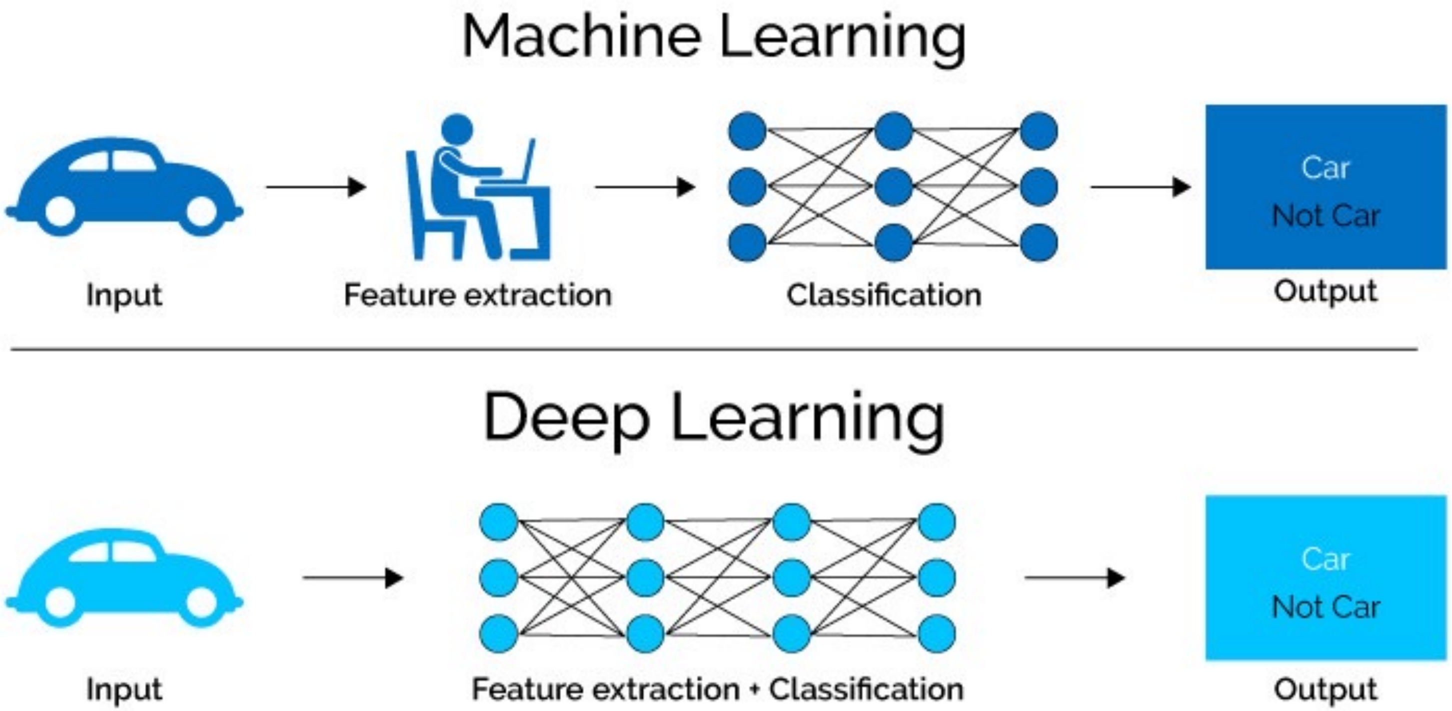
What's a "chair"?

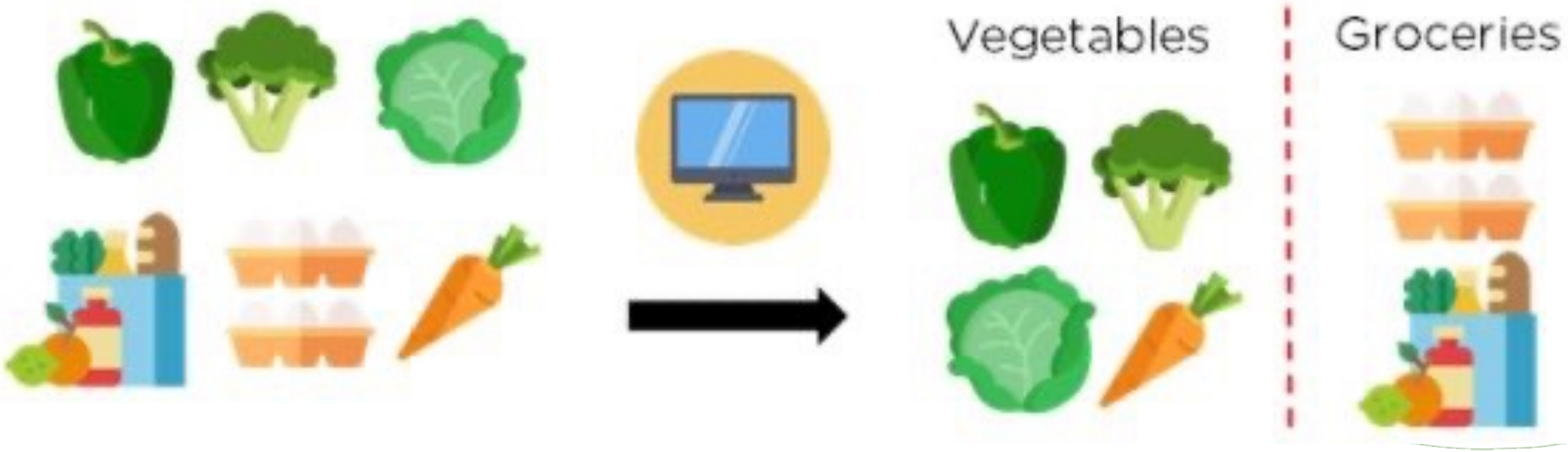
What’s a “chair”?

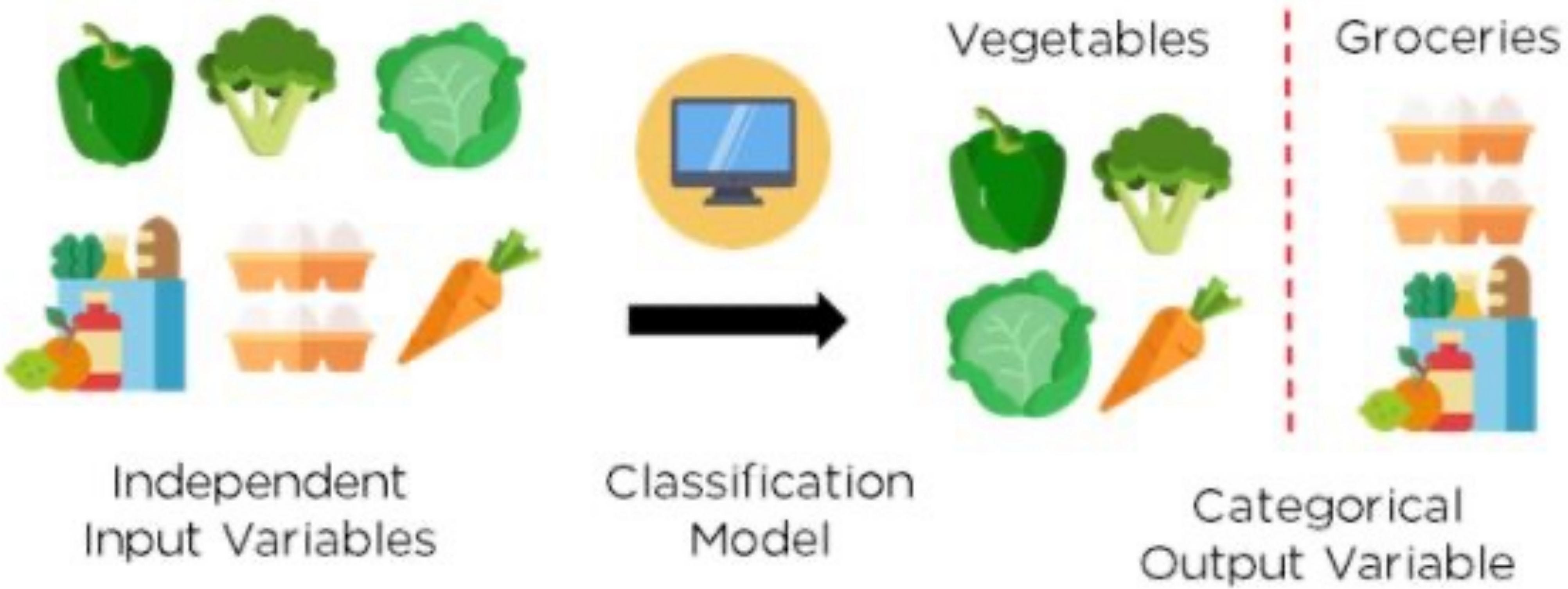
What’s a “chair”?

What’s a “chair”?

Fully automated learning




Vegetables
Groceries

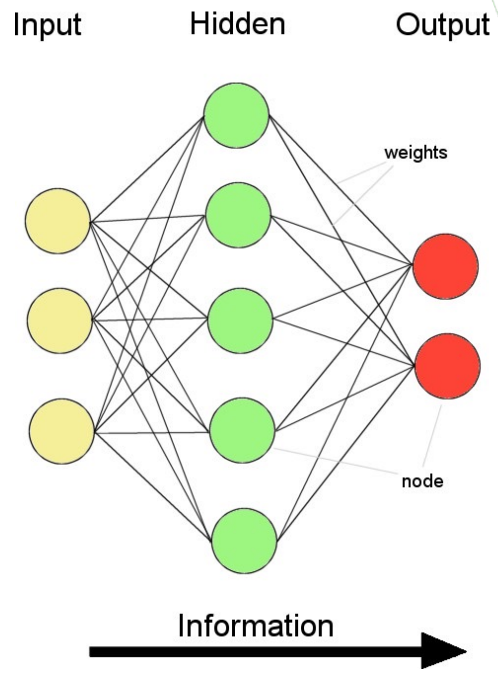

Vegetables
Groceries
Independent
Input Variables
Classification
Model
Categorical
Output Variable

Some questions

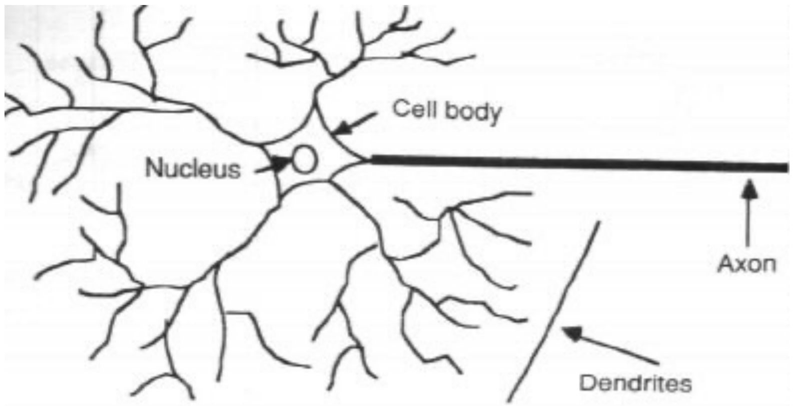
- What is a neural network?
- How does it work and why now?
- Why is it generally better than other methods on image, speech and certain other types of data?

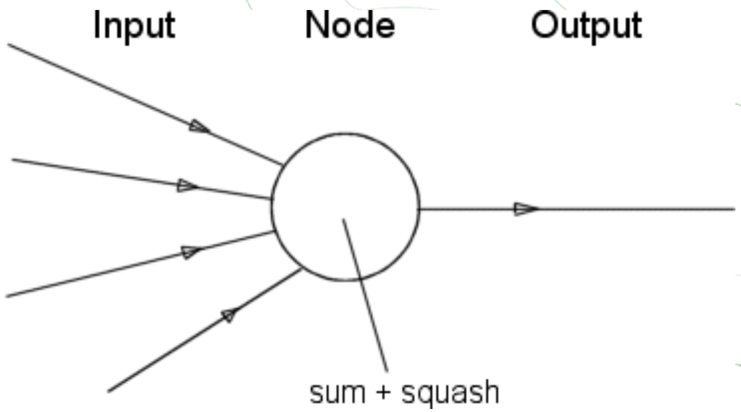
Artificial Neural Networks (ANNs)

- ANNs incorporate the two fundamental components of biological neural nets:
  - Neurones (nodes)
  - Synapses (weights)



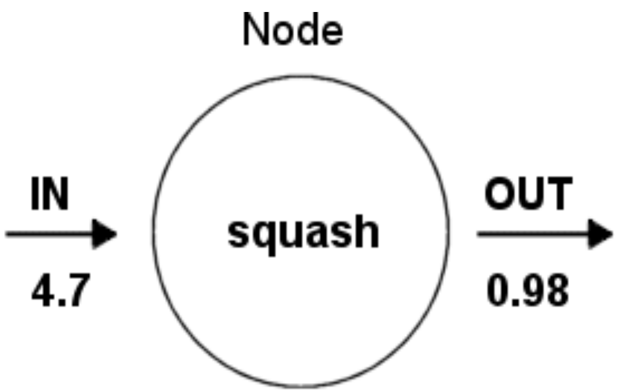
Neuron vs. Nodes



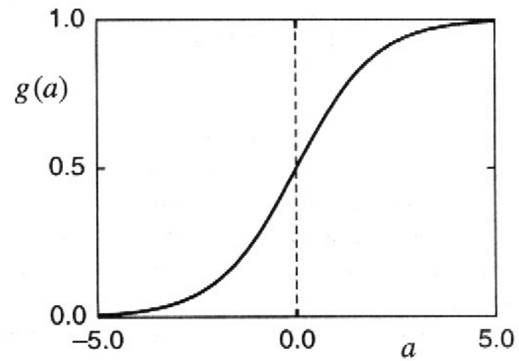


131

Structure of a node

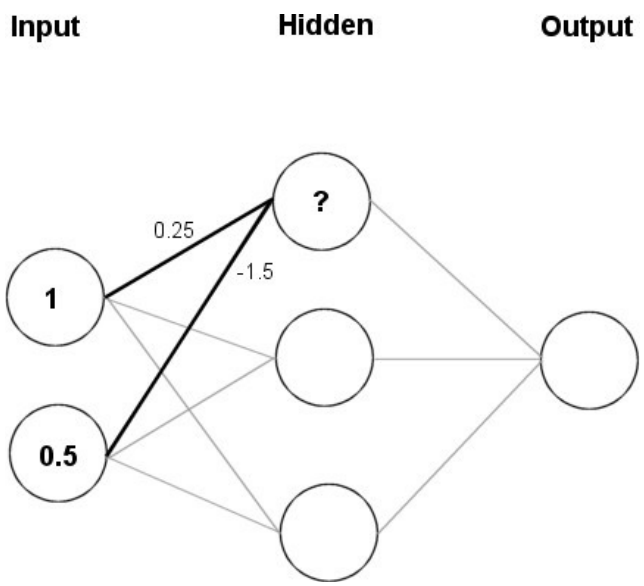


Squashing function limits node output:



132

Feeding data through the net



$(1 \times 0.25) + (0.5 \times (-1.5)) =$

$0.25 + (-0.75) \quad = \ -\mathbf{0.5}$

Squashing:

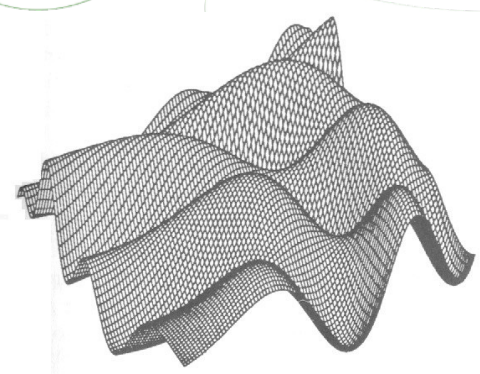
$$\frac{1}{1+e^{0.5}} = 0.3775$$

Weight settings determine the behaviour of a network

→ How can we find the right weights?

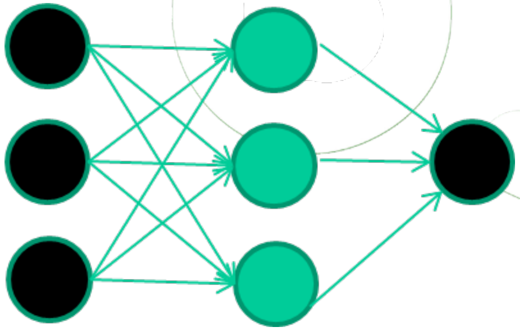
Training – Backpropagation

- The Learning Phase
  - Requires training set (input / output pairs)
  - Starts with small random weights
  - Error is used to adjust weights (supervised learning)
- -> Gradient descent on error landscape



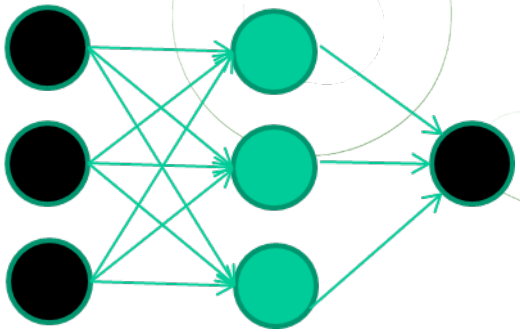
A dataset

Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



Training the neural network

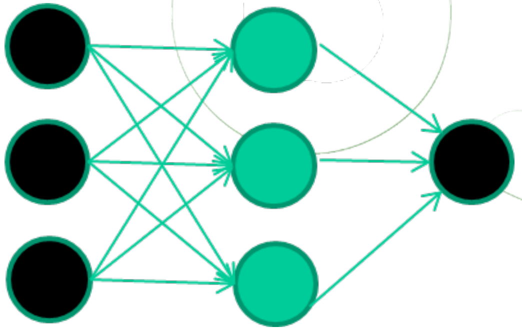
Fields			*class*
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



Training data

- Initialise with random weights

Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			

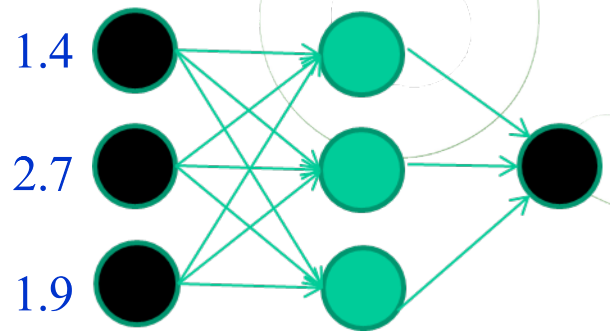


138

Training data

- Present a training pattern

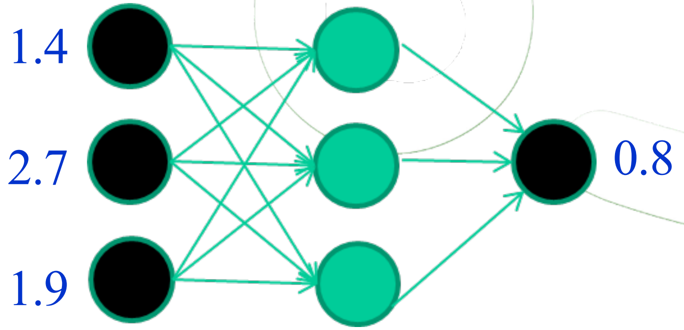
Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



Training data

- Feed it through to get output

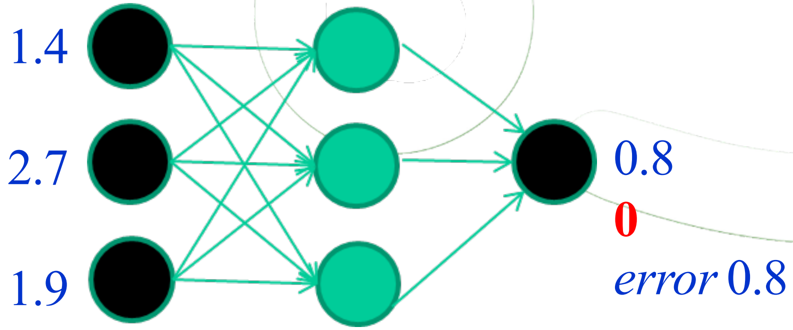
Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



Training data

- Compare with target output

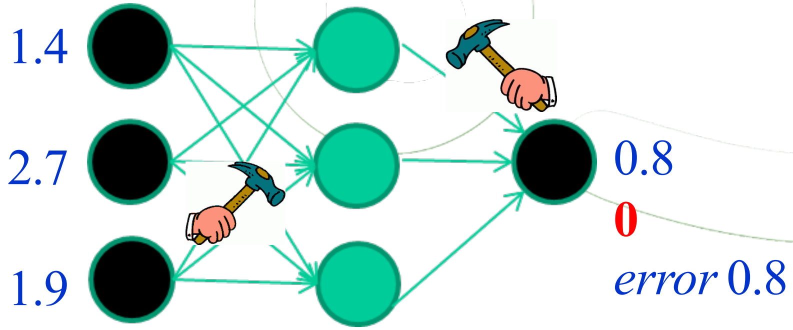
Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



Training data

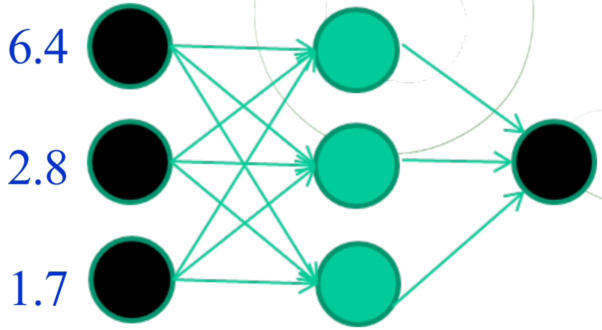
- Adjust weights based on error

Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



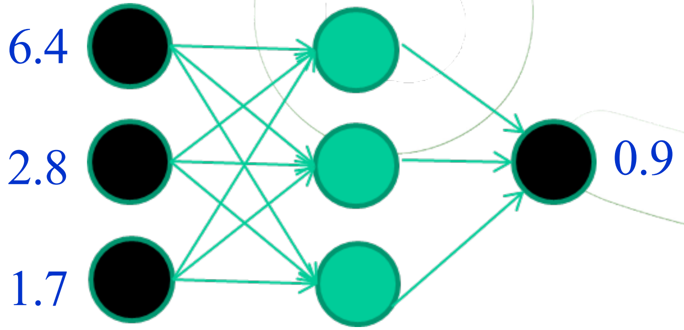
Training data

Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



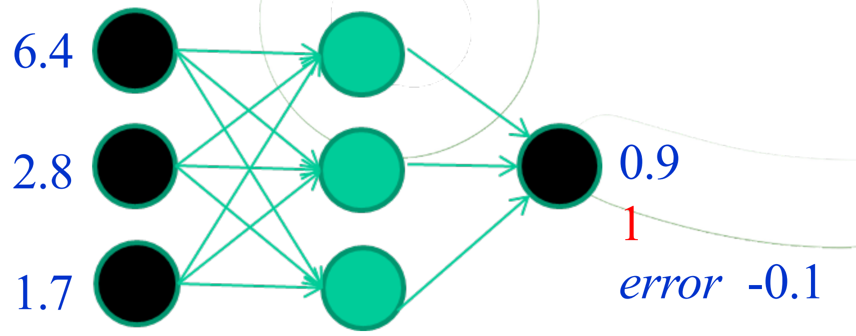
Feed it through to get output

	Fields		***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



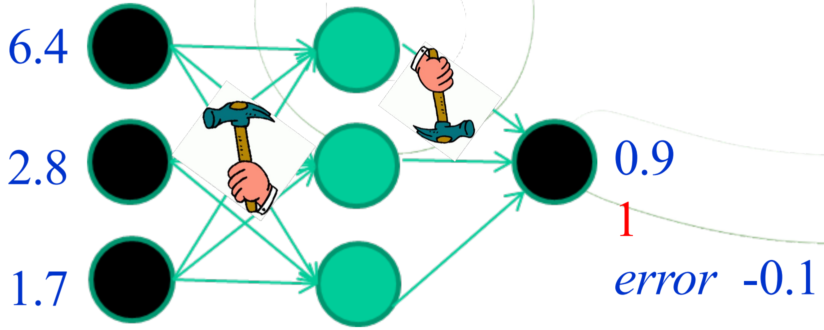
Training data

Fields			class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc ...			



Training data

	Fields		class
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			

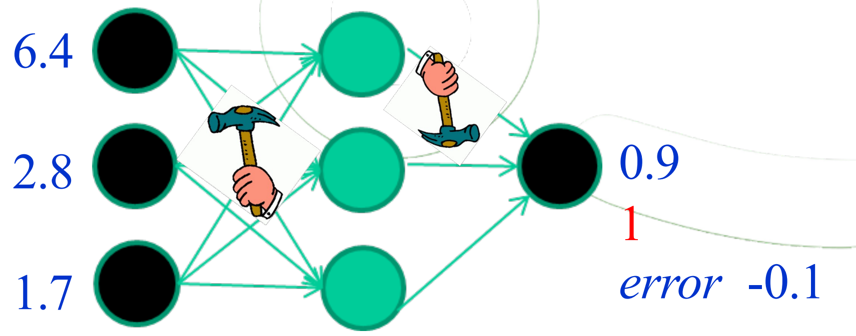


Training data

- Repeat this thousands, maybe millions of times – each time taking a random training instance, and making slight weight adjustments

Training data

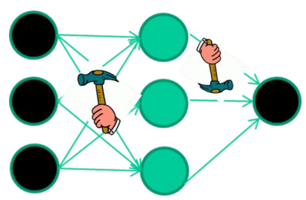
Fields			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			



And so on ….

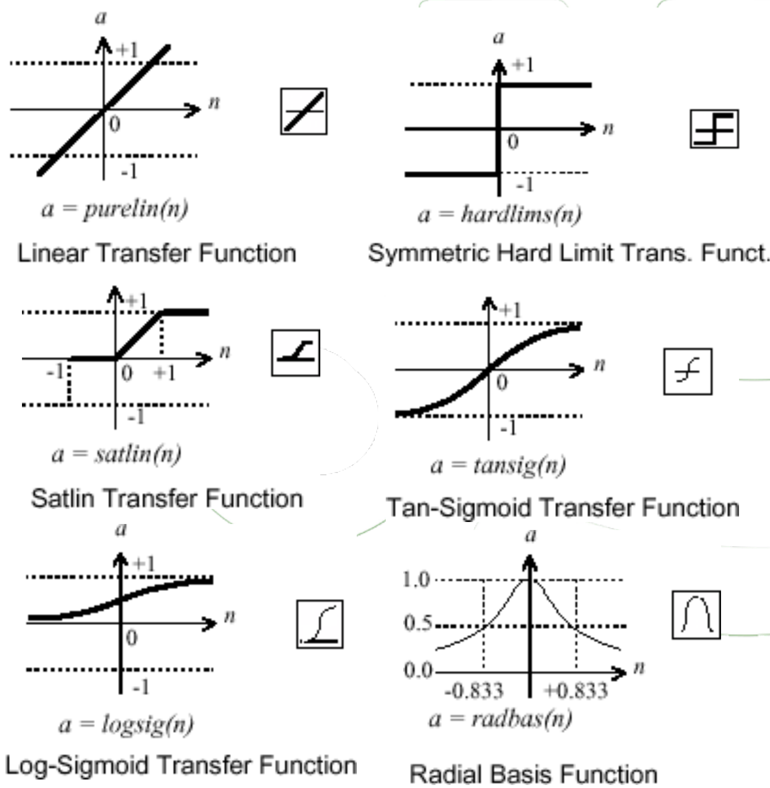
Dumbness and volume

- Weight-learning algorithms for ANNs are dumb
- They work by making thousands and thousands of tiny adjustments, each making the network do better at the most recent pattern, but perhaps a little worse on many others
- But, by dumb luck, eventually this tends to be good enough to learn effective classifiers for many real applications



Activation functions

- The activation function is generally non-linear.
- Linear functions are limited because the output is simply proportional to the input.



Some other points

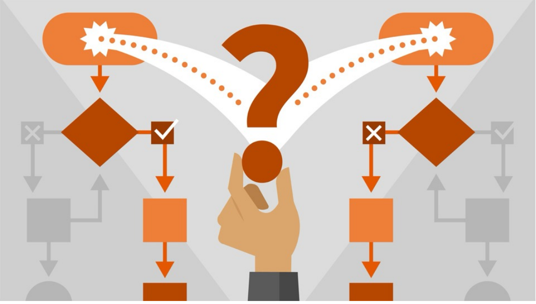
- If f(x) is non-linear, a network with 1 hidden layer can, in theory, learn perfectly any classification problem.
- A set of weights exists that can produce the targets from the inputs.
- The problem is finding them.

GOAL and MEANS

- You must choose the right method
- You must run statistical validity tests

The basic of “Modeling”

- We do it naturally in life, maybe without knowing it
  - Basic task for ”data miners”
  - Statisticians have been doing it for many years
  - It takes many different form
  - Today, all managers should have at least a basic understanding



Modeling: A simple "supervised" example

Age	Income	**Out**
30	65k	Y
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

Modeling: A simple “supervised” example

Age	Income	**Out**
30	65k	Y
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

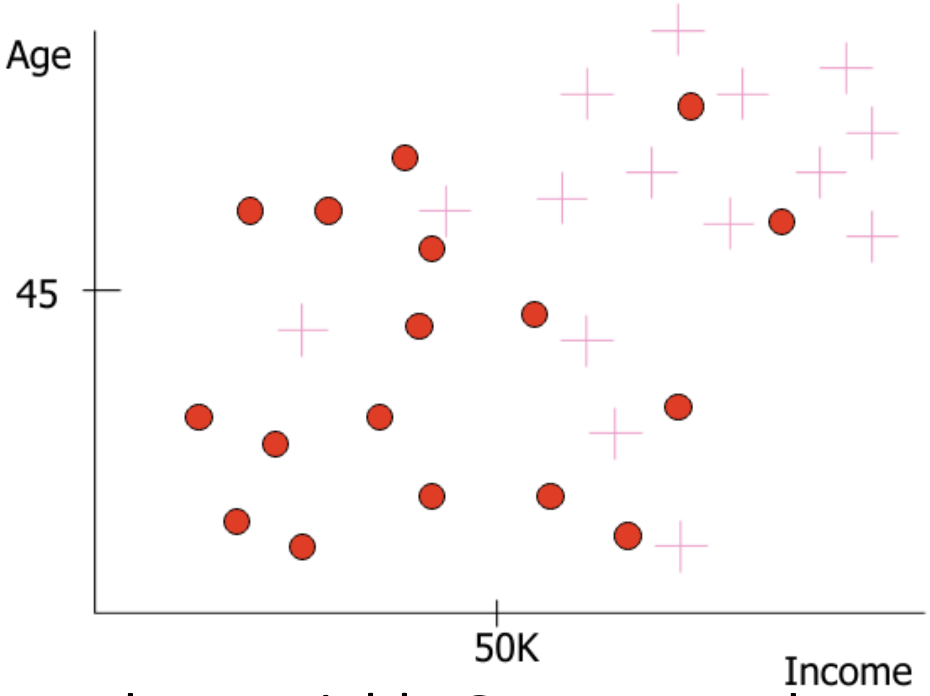
- Goal: Build a model to predict the dependent variable Outcome

Modeling: A simple “supervised” example

Age	Income	**Out**
~~30~~	~~65k~~	~~Y~~
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

- **Out** is a categorical variable
- *Age* and *Income* are the independent variables

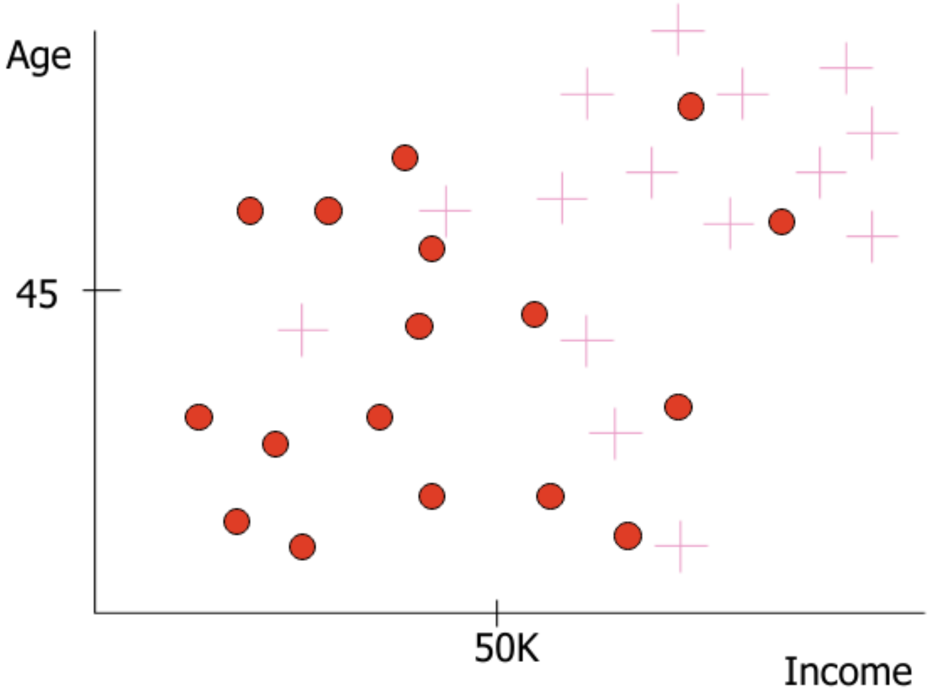
Let’s model



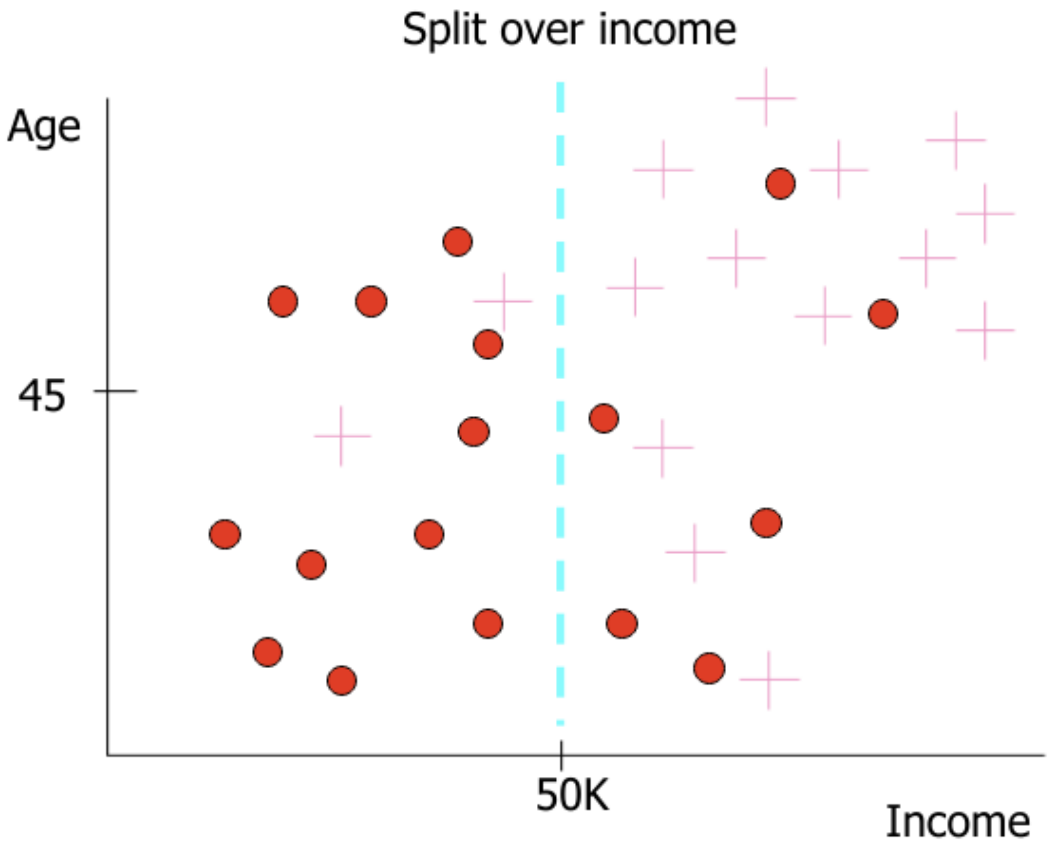
- We project the dependent variable **Out**, as + and ∘, on a two-dimensional space

Let’s model

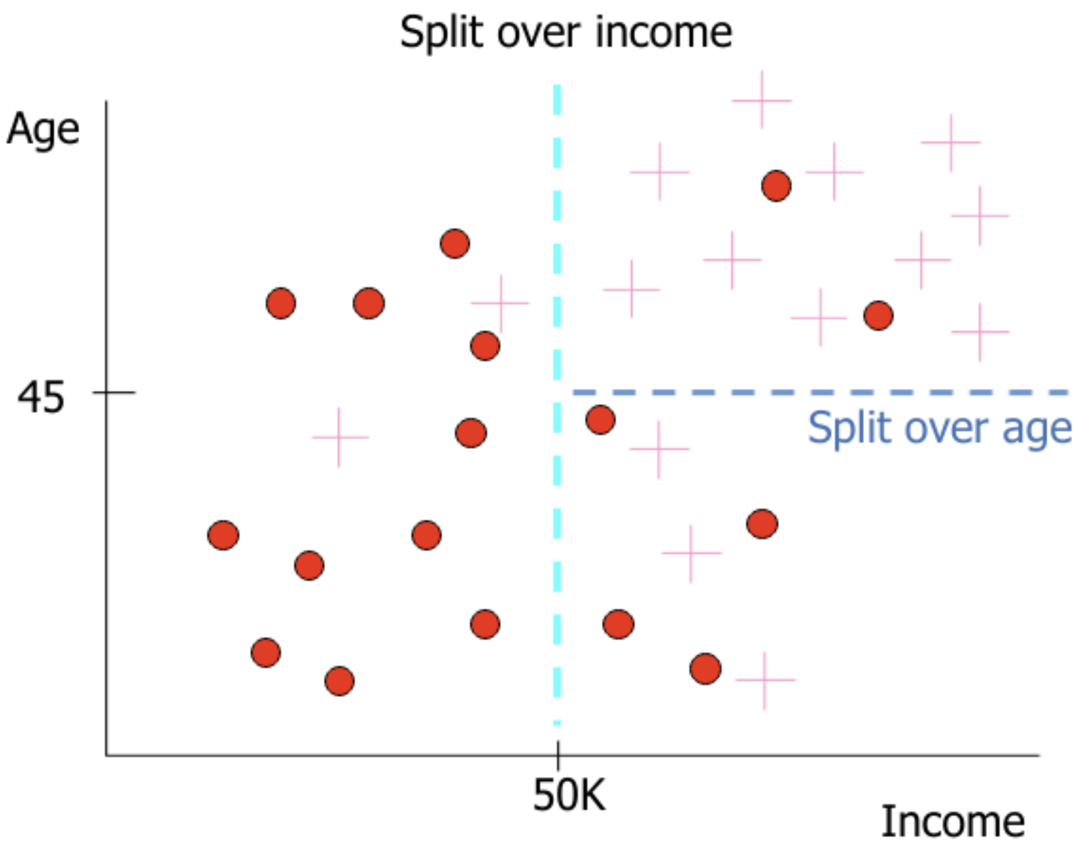
- Do we notice anything?



Let’s model

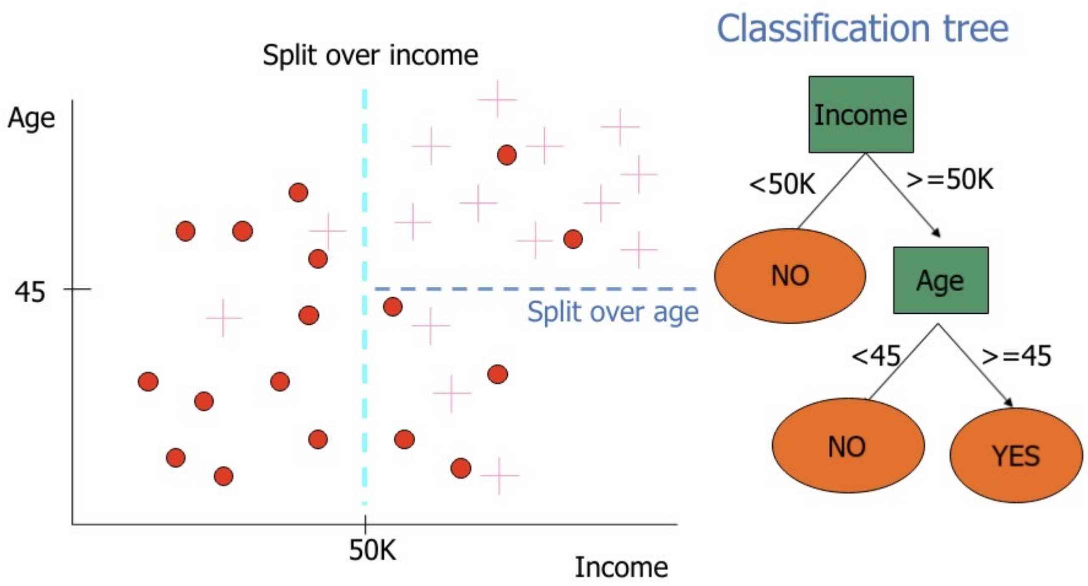


Let’s model



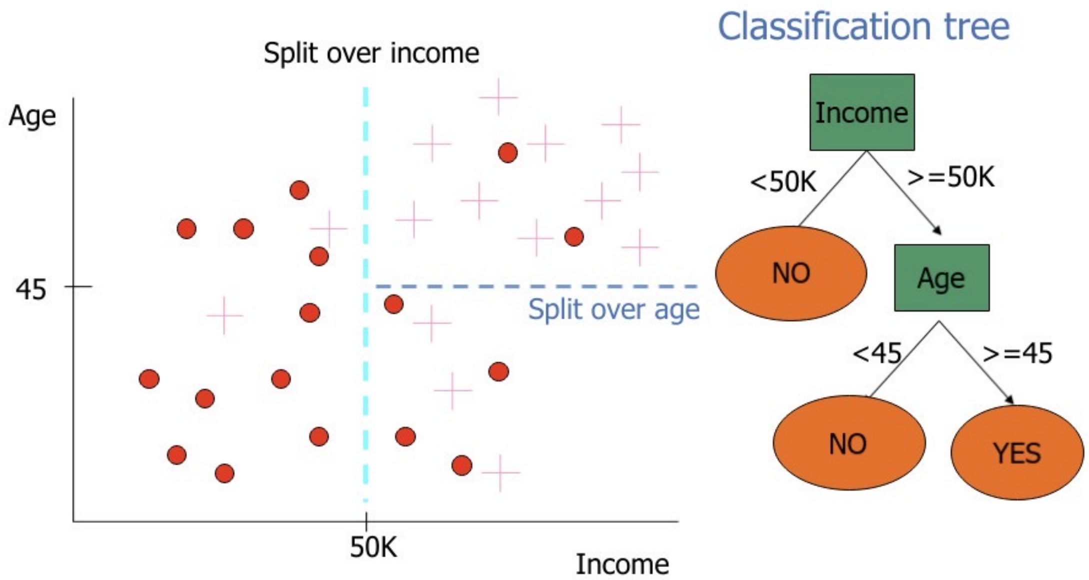
Let’s model

- We can model it through a Classification tree



Let's model

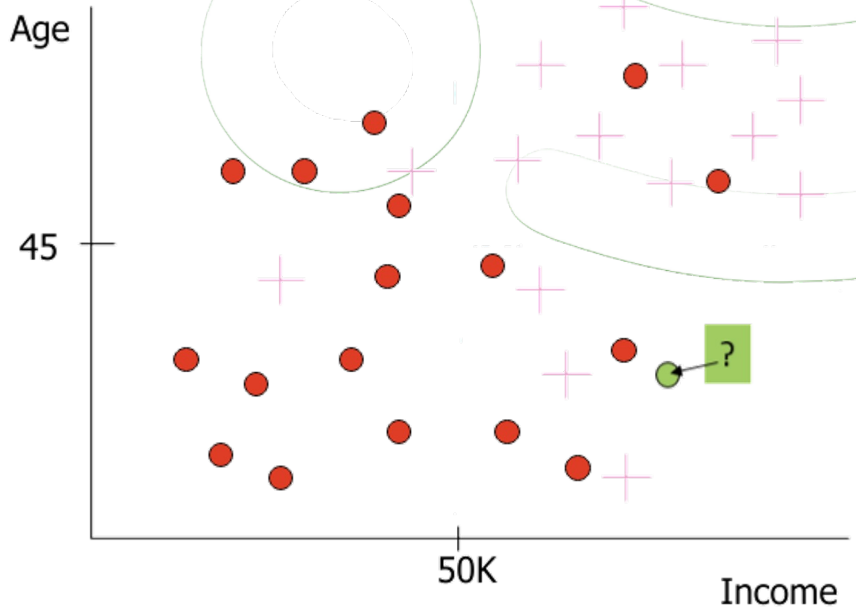
- We can model it through a Classification tree



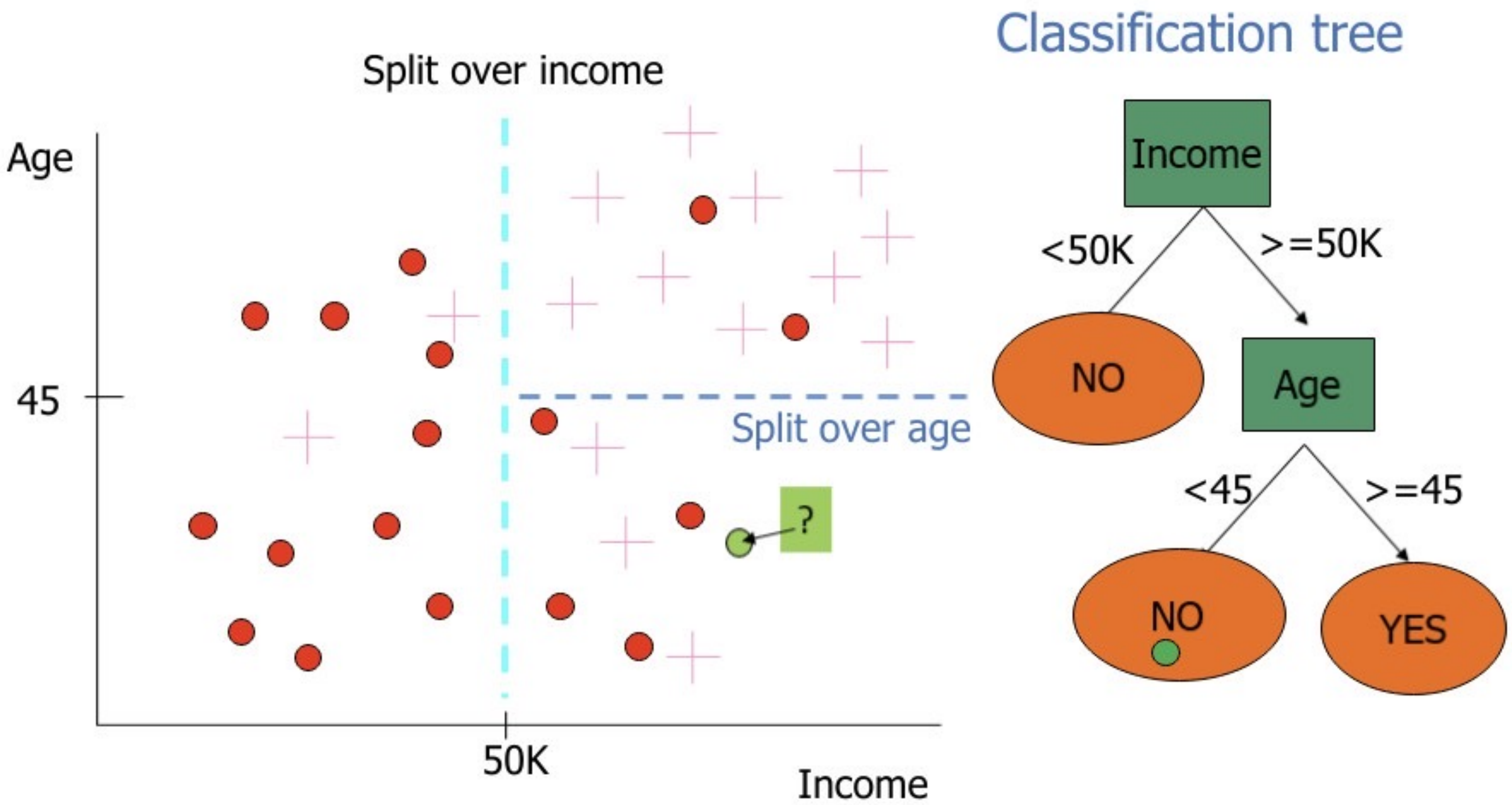
- The algorithm finds the optimal splits
- It maximizes prediction confidence

Let’s apply the model now

- Let’s now generalize the model
- Say, we receive new data and we want to predict the **Out** variable
  - For instance, a new person 25 years old and with an income of 70k

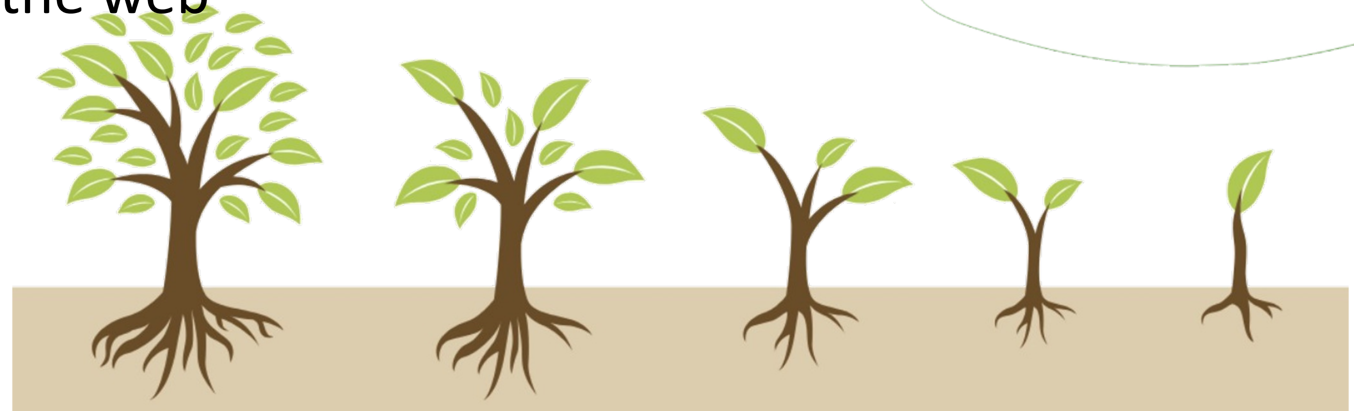


We can now predict



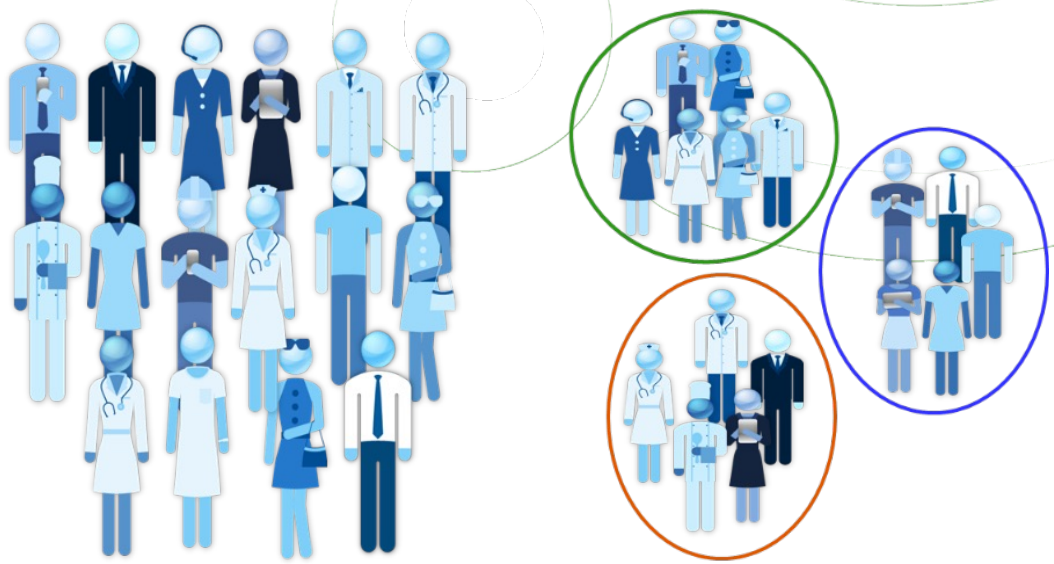
Classification tree considerations

- C4.5 was the original idea
- Old technique very well established
- Works for categorical dependent variable
- It works with both continuous and categorical independent variables
- Fast to execute
- Easy to find free code on the web



Clustering

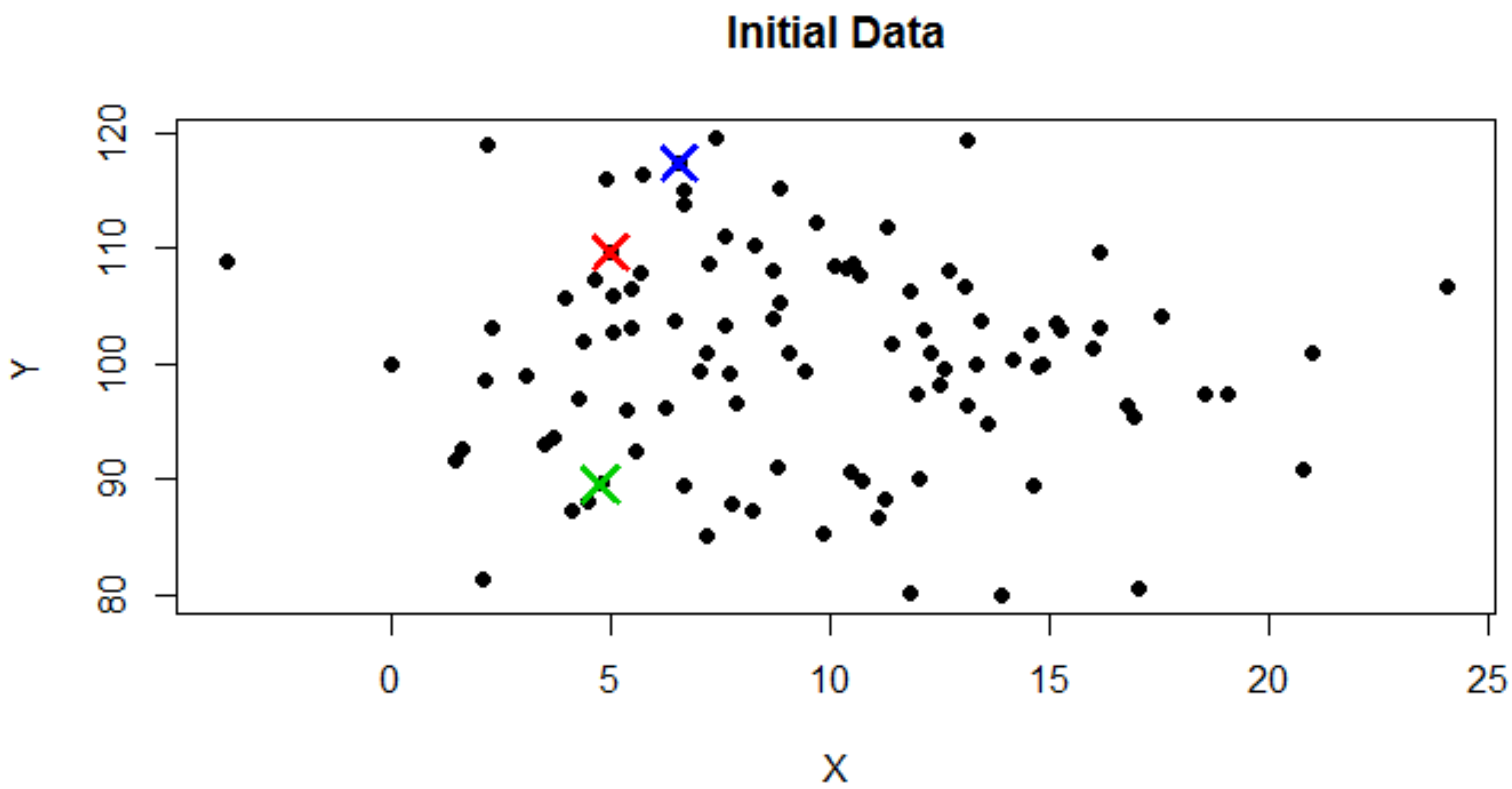
- **Unsupervised learning model**
- Widely used in business/marketing applications
- Gives a structure to unsorted data points
- In simpler words: Aggregate similar items



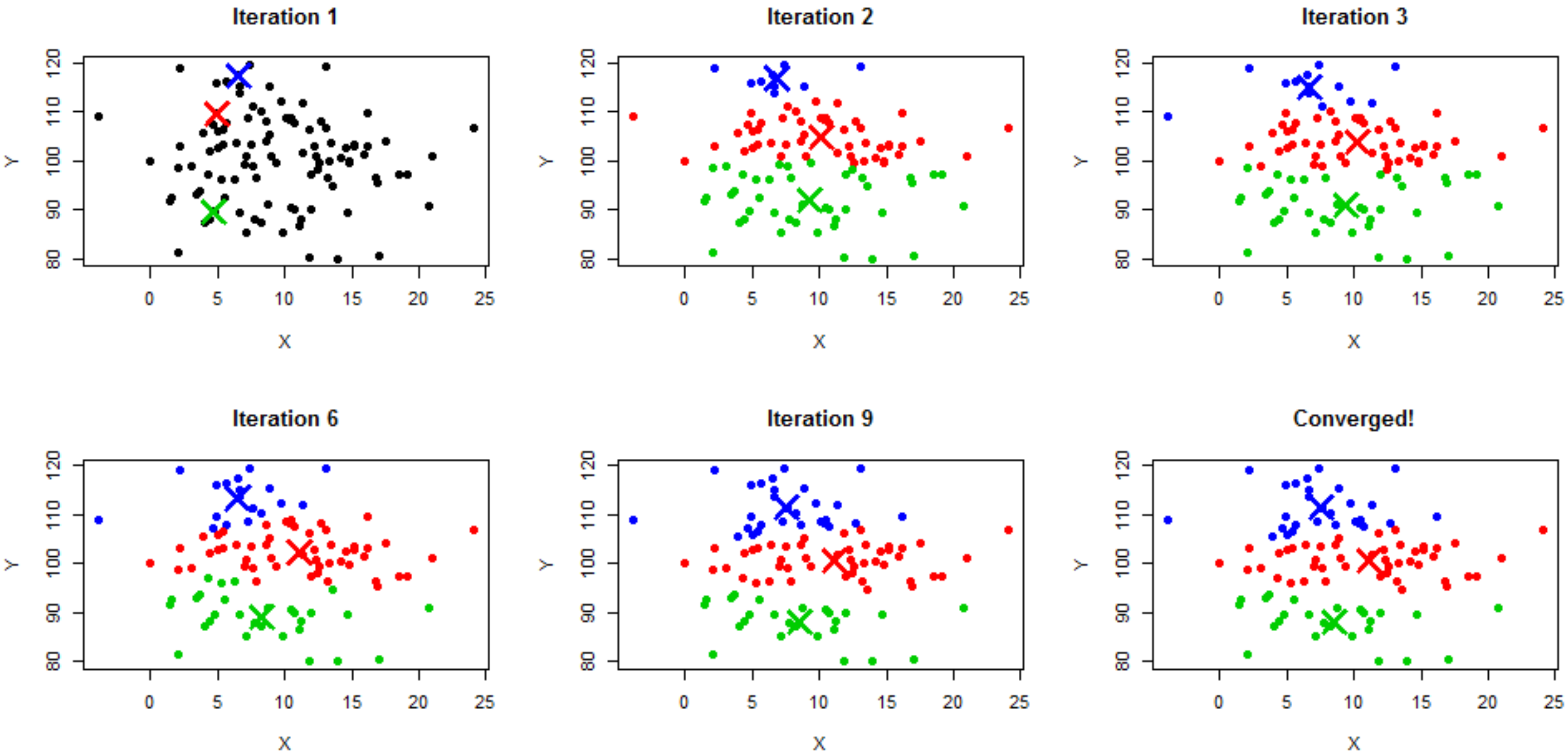
k-means Clustering algorithm

- Step 0: Initialize K random centroids (just pick randomly K data points)
- Step 1 For every data point:
  - assign it to the closest centroid (any distance metric works);
- Step 2 For every centroid
  - move the centroid to the average among all its points;
- Repeat Step 1 and Step 2 until all centroids do not change anymore;
- The algorithm converged

k-means example... Initial step



k-means example... Iterations



k-means, some considerations

- Easy to apply
- Various algorithms available
  - You can find free, ready to use, code on the web
- Not suitable for categorical variables
- Need to normalize variable for scale uniformity
  - Otherwise, distance calculation may be affected
- It scales on Big Data, that is, it can be parallelized
- How to pick initial K?

Associations

What can we infer just by observing?

Association rules

- Market-basket analysis: Understanding meaningful patterns by analysing baskets
- A basket is a generic set of items
  - Pattern: A set of items
  - Frequent pattern: A pattern that appears frequently
  - We infer rules such as: $A, B, \ldots, C \Longrightarrow E$
  - Beer and diapers on friday evening?!
  - It may depend on the context: ”Have kids?”, ”Travelling fo

What Is the ML Pipeline?

- Structured process for building ML models
- Ensures repeatability and clarity
- Common across all ML workflows
- Steps: problem → data → model → evaluation

Step 1 – Problem Definition

- Understand the business or scientific objective
- Define input/output clearly
- Set measurable success criteria
- Decide on ML type: supervised, unsupervised, or reinforcement

Step 2 - Data Collection & Preparation

- Collect relevant data from appropriate sources
- Clean the data: remove duplicates, handle missing values
- Feature engineering: extract and format meaningful attributes
- Normalize/standardize data if needed

Step 3 - Algorithm Selection

- Choose based on task: classification, regression, clustering
- Consider:
  - Data size and quality
  - Training time
  - Interpretability vs performance
- Start simple (e.g., linear models) and iterate

Step 4 – Model Training

- Use the training data to teach the model
- Fit parameters to minimize error
- Adjust model settings (hyperparameters)
- Watch for convergence or signs of overfitting

Step 5 – Model Evaluation

- Test on unseen data (test set)
- Use metrics:
  - Classification: accuracy, precision, recall, F1
  - Regression: RMSE, MAE, R^2
- Compare multiple models for best performance

Overfitting vs Underfitting

- Overfitting: memorizes training data, **fails on new data**
- Underfitting: **too simple**, misses important patterns
- Visualization: training vs test accuracy curve
- Ideal: good generalization to new, unseen data

Model Validation Basics

- Use a validation set or cross-validation
- Helps tune hyperparameters and assess performance
- K-Fold Cross-Validation = robust, avoids random split bias
- Use validation before final test

Improving Generalization

- Techniques:
  - Regularization (L1/L2)
  - Dropout (in neural nets)
  - Pruning (in trees)
  - Ensembling (bagging, boosting)
- Collect more or better data
- Perform feature selection or engineering

Summary & Q&A

- ML pipeline provides structure to model building
- Each stage impacts final performance
- Validation helps build robust, trustworthy models
- Avoid both overfitting and underfitting

Module 5: Final Activity + Review Quiz

- Look at a real-world problem and choose an ML approach
- Final quiz (via Kahoot)
- Class discussion and wrap-up

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

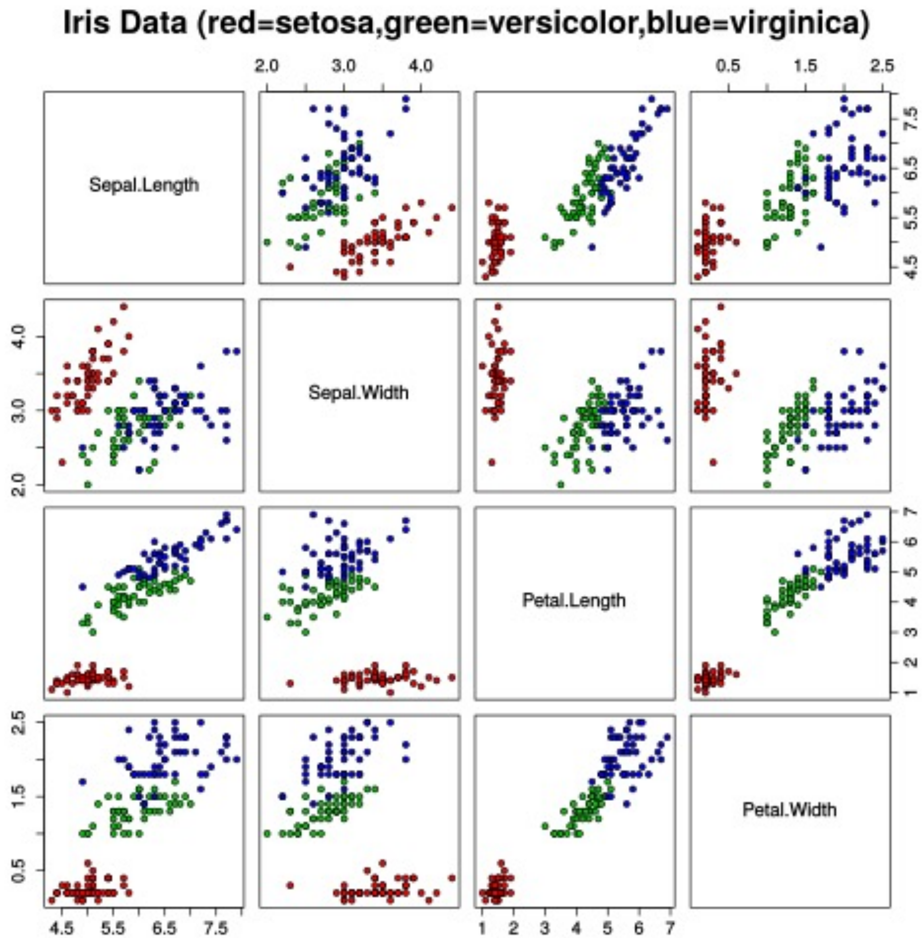
Iris flower data set (Fisher's Iris by the British statistician)

The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Iris virginica* and *Iris versicolor*).

Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish each species.

Fisher's paper was published in the *Annals of Eugenics* (today the *Annals of Human Genetics*).



ITINERIS

THANKS!

Francesco Iarlori

https://uk.linkedin.com/in/thefrankie/

@thefrankie

Francesco@Iarlori.com

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

ITINERIS

BackUp

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

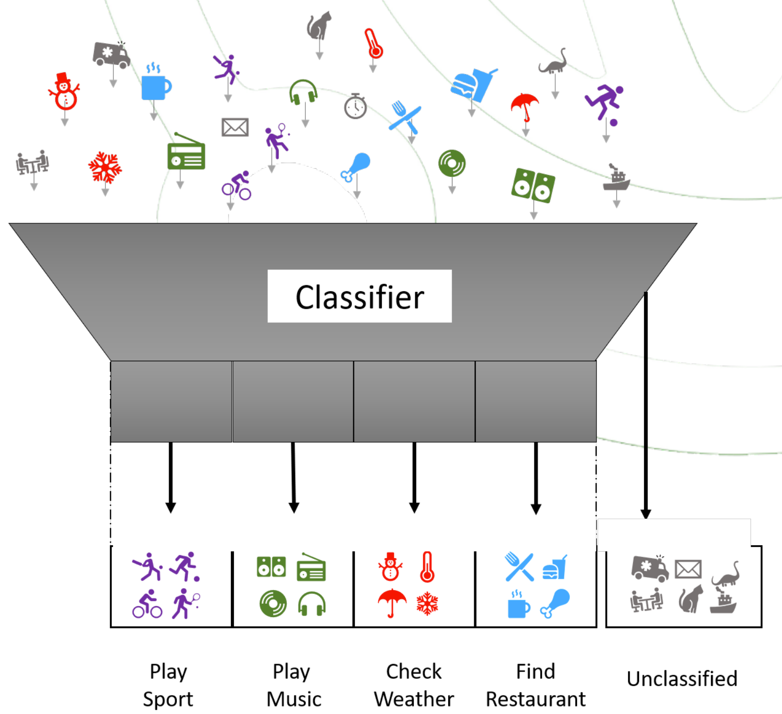
Italiadomani

Word embeddings

- Idea: learn an embedding from words into vectors
- Need to have a function W(word) that returns a vector encoding that word.

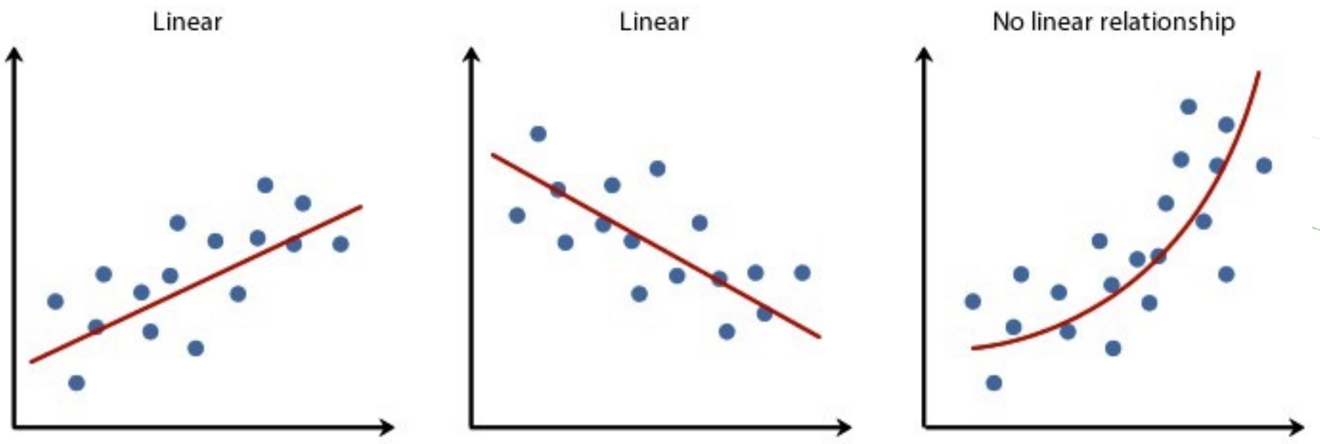
ML problems

- **Supervised**. Provide samples of data and labels.
- Classification of objects in categories: cats and dogs in pictures



ML problems (cont.)

- **Supervised**. Provide samples of data and labels.
- Regression. Estimating the relationships between a dependent variable and one or more independent variables



ML problems (cont.)

- **Unsupervised**. Provide data items and group them
- Clustering of objects in groups by similarity