ITINERIS

Data Mining and Machine Learning

Fundamentals of AI

Elena Grimaccia

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

Training objectives

- Understand Fundamental Concepts: Define and distinguish between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning, including various learning paradigms such as supervised, unsupervised, and reinforcement learning.
- Knowledge of key machine learning models: Supervised learning approaches (generalised lineal models such as linear and not linear regression, logistic regression, predictions, qualitative predictors), Unsupervised learning techniques (Principal Component Analysis, Cluster Analysis) relevant to environmental data, Training- versus Test-Set Performance.
- Apply AI in Environmental Contexts: Utilize AI tools and techniques for environmental applications, including biodiversity monitoring, climate modelling, remote sensing, and big data analysis related to air, water, and soil quality.
- Evaluate Ethical Implications: Assess the ethical considerations of deploying AI in environmental contexts, focusing on privacy, data sovereignty, automation impacts, and equitable decision-making.

Training objectives

- Understand Fundamental Co nts: Def nd d guish between Artificial Intelligence (AI), Machine Learning (ML), a n L ous learning paradigms such as supervised, unsupervised g.
- Knowledge of key machi proaches (generalised lineal models such as linea gression, predictions, qualitative predictors), Unsupe ipal Component Analysis, Cluster Analysis) relevant to en st-Set Performance.
- Apply AI in Environmental exts niques for environmental applications, including biodiversity ling, remote sensing, and big data analysis related to air, wat and qu
- Evaluate Ethical Implications: Assess the ical co derations of deploying AI in environmental contexts, focusing on priva , data so ereignty, automation impacts, and equitable decision-making.

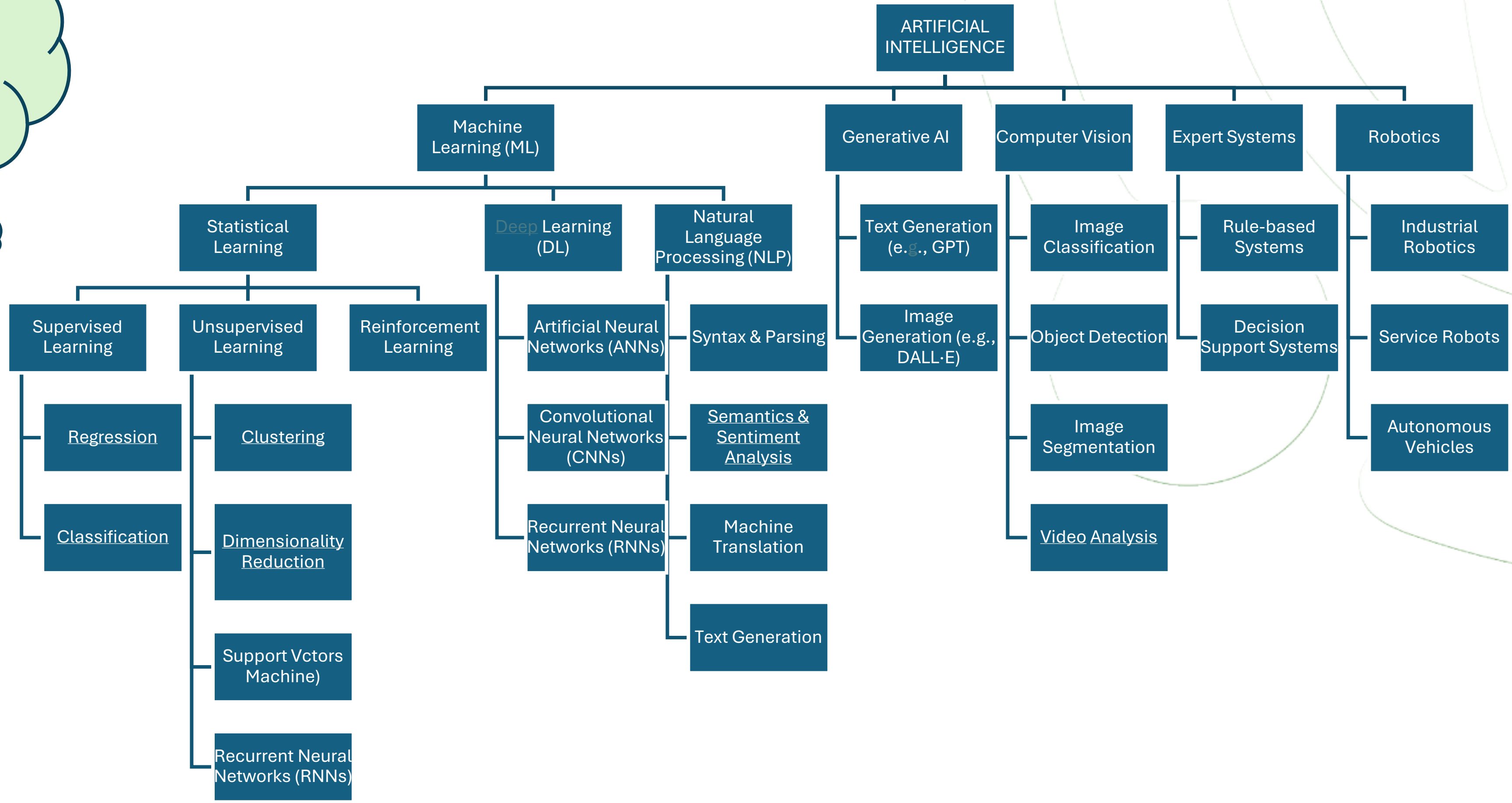
It’s a lot!!

Aim of this lesson

- Understand the key concepts and categories of AI.
- Understand Fundamental Concepts: Define and distinguish between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning.

Our first aim: definition of AI

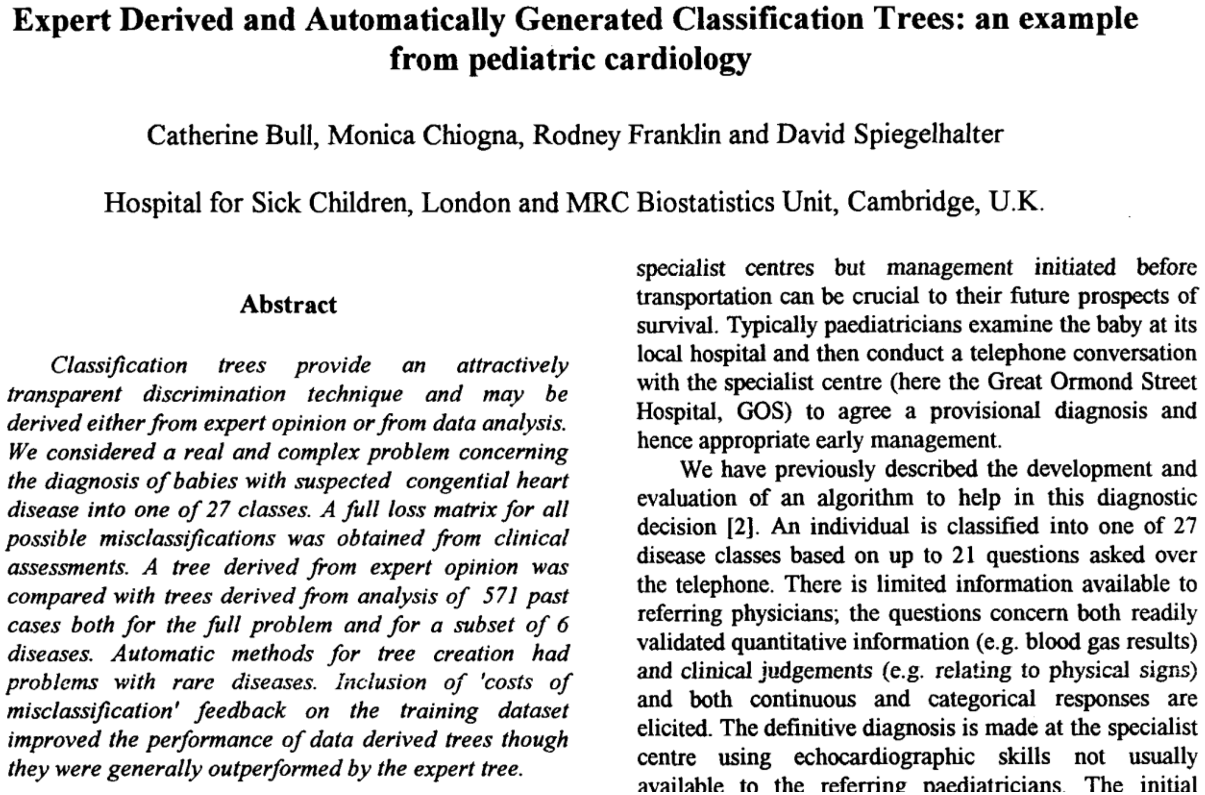
What is AI?



- ARTIFICIAL INTELLIGENCE
  - Machine Learning (ML)
    - Statistical Learning
      - Supervised Learning
        - Regression
        - Classification
      - Unsupervised Learning
        - Clustering
        - Dimensionality Reduction
        - Support Vctors Machine)
        - Recurrent Neural Networks (RNNs)
      - Reinforcement Learning
    - Deep Learning (DL)
      - Artificial Neural Networks (ANNs)
      - Convolutional Neural Networks (CNNs)
      - Recurrent Neural Networks (RNNs)
    - Natural Language Processing (NLP)
      - Syntax & Parsing
      - Semantics & Sentiment Analysis
      - Machine Translation
      - Text Generation
  - Generative AI
    - Text Generation (e.g., GPT)
    - Image Generation (e.g., DALL·E)
  - Computer Vision
    - Image Classification
    - Object Detection
    - Image Segmentation
    - Video Analysis
  - Expert Systems
    - Rule-based Systems
    - Decision Support Systems
  - Robotics
    - Industrial Robotics
    - Service Robots
    - Autonomous Vehicles

Something for everyone in AI

Classification trees



Expert Derived and Automatically Generated Classification Trees: an example from pediatric cardiology

Catherine Bull, Monica Chiogna, Rodney Franklin and David Spiegelhalter

Hospital for Sick Children, London and MRC Biostatistics Unit, Cambridge, U.K.

Abstract

Classification trees provide an attractively transparent discrimination technique and may be derived either from expert opinion or from data analysis. We considered a real and complex problem concerning the diagnosis of babies with suspected congential heart disease into one of 27 classes. A full loss matrix for all possible misclassifications was obtained from clinical assessments. A tree derived from expert opinion was compared with trees derived from analysis of 571 past cases both for the full problem and for a subset of 6 diseases. Automatic methods for tree creation had problems with rare diseases. Inclusion of 'costs of misclassification' feedback on the training dataset improved the performance of data derived trees though they were generally outperformed by the expert tree.

specialist centres but management initiated before transportation can be crucial to their future prospects of survival. Typically paediatricians examine the baby at its local hospital and then conduct a telephone conversation with the specialist centre (here the Great Ormond Street Hospital, GOS) to agree a provisional diagnosis and hence appropriate early management.

We have previously described the development and evaluation of an algorithm to help in this diagnostic decision [2]. An individual is classified into one of 27 disease classes based on up to 21 questions asked over the telephone. There is limited information available to referring physicians; the questions concern both readily validated quantitative information (e.g. blood gas results) and clinical judgements (e.g. relating to physical signs) and both continuous and categorical responses are elicited. The definitive diagnosis is made at the specialist centre using echocardiographic skills not usually available to the referring paediatricians. The initial

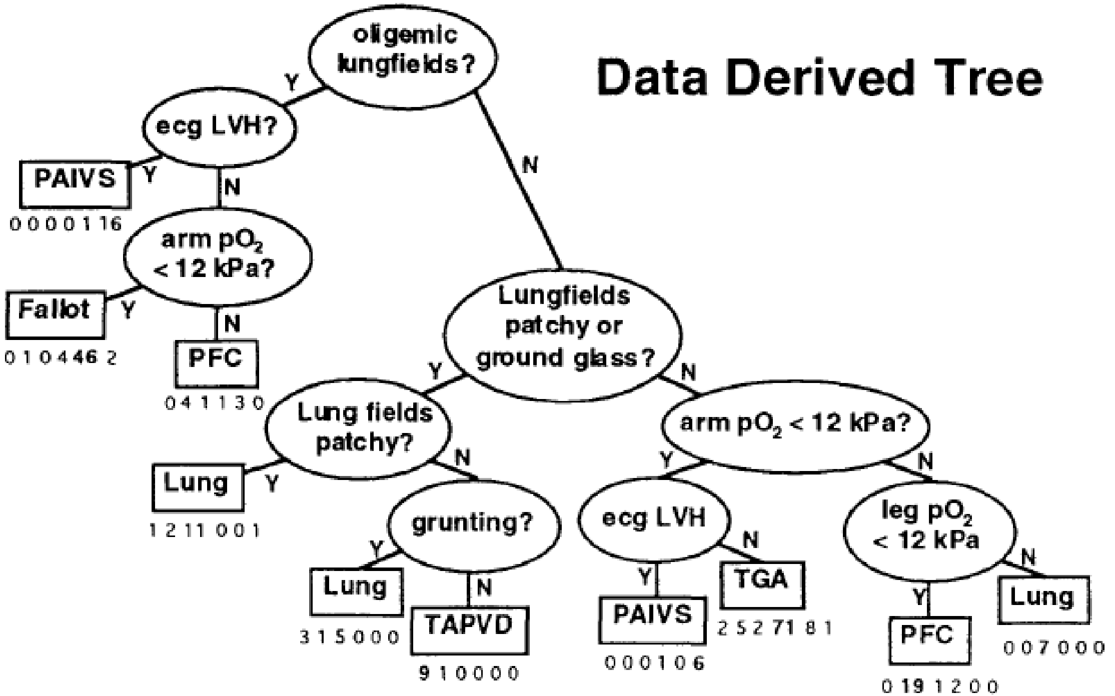
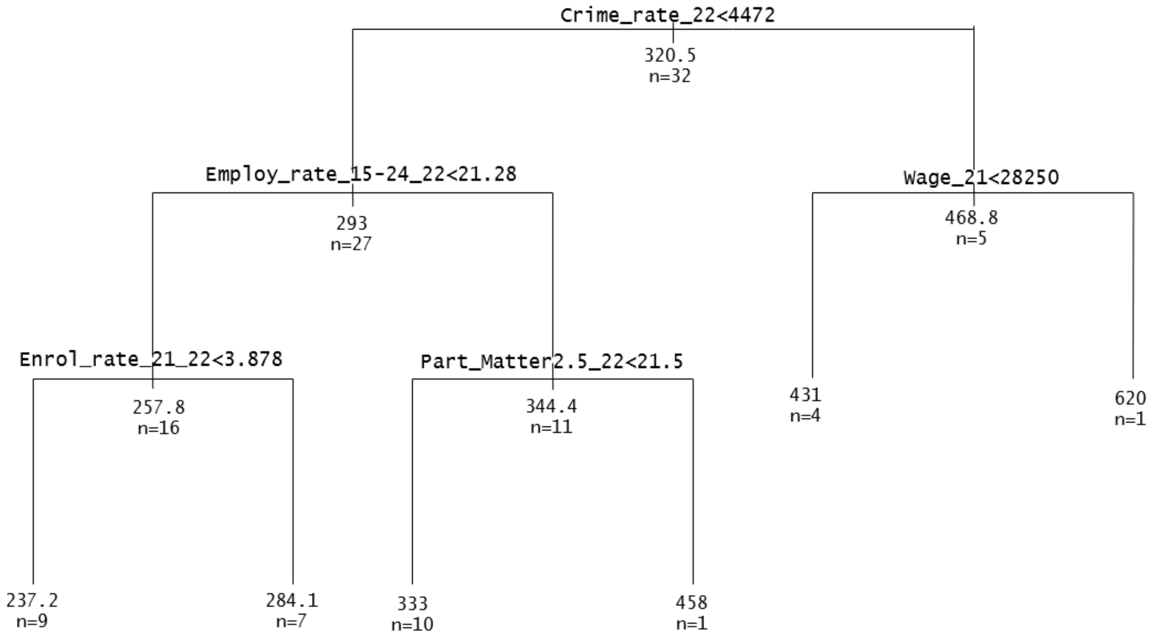


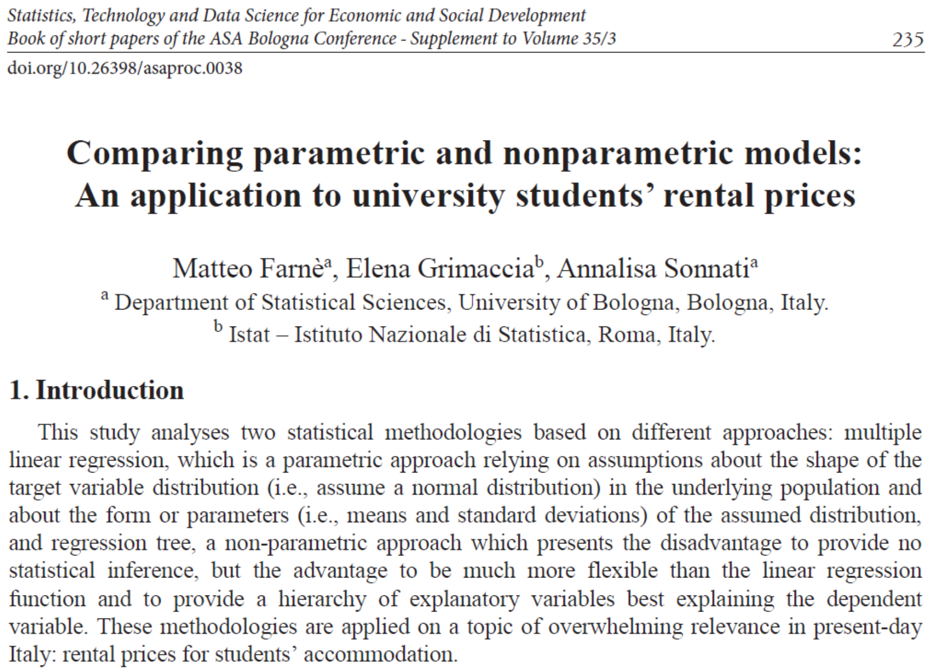
Figure 2. Data derived tree trained from the full dataset of 238 patients with 6 cyanotic diseases. Using the same dataset as testset, the numbers of times each diagnosis was reached is shown under each leaf

C. Bull, M. Chiogna, R. Franklin and D. Spiegelhalter, "Expert derived automatically generated classification trees: an example from pediatric cardiology," Proceedings of Computers in Cardiology Conference, London, UK, 1993, pp. 217-220, doi: 10.1109/CIC.1993.378465.

Regression trees

University students' rental prices=f (cities characteristics)




Statistics, Technology and Data Science for Economic and Social Development
Book of short papers of the ASA Bologna Conference - Supplement to Volume 35/3 235
doi.org/10.26398/asaproc.0038

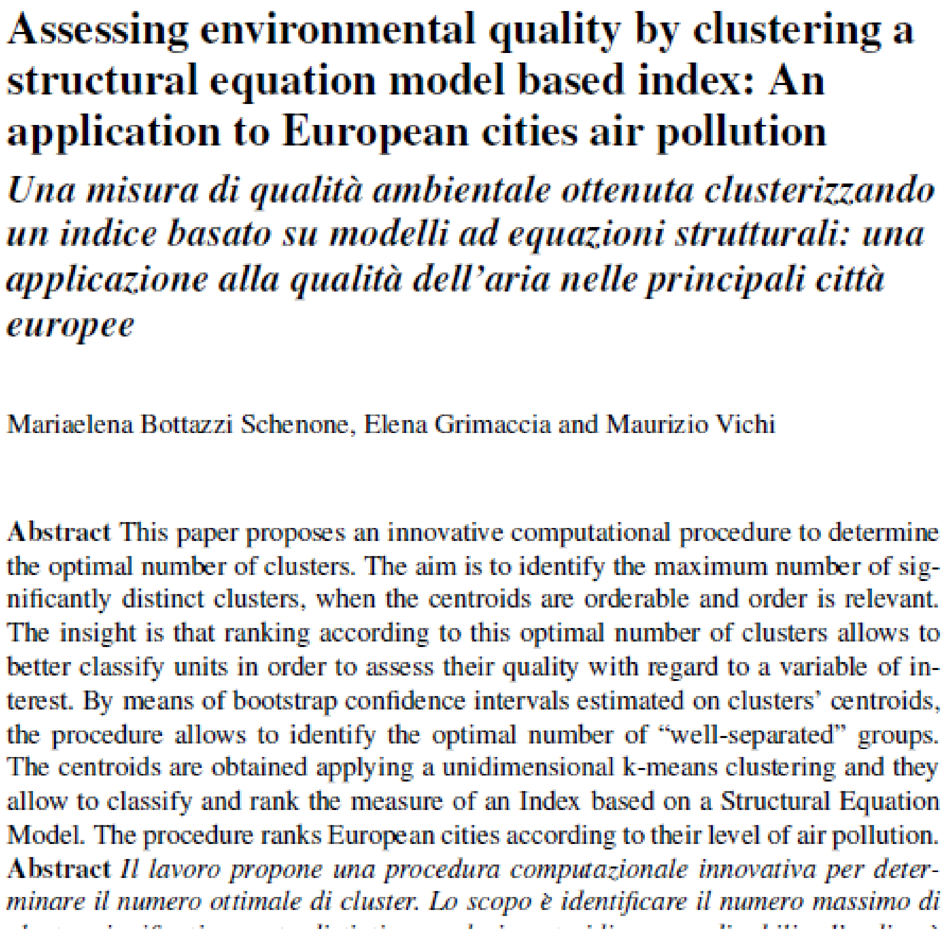
Comparing parametric and nonparametric models: An application to university students' rental prices

Matteo Farnè[a], Elena Grimaccia[b], Annalisa Sonnati[a]
[a] Department of Statistical Sciences, University of Bologna, Bologna, Italy.
[b] Istat – Istituto Nazionale di Statistica, Roma, Italy.

1. Introduction

This study analyses two statistical methodologies based on different approaches: multiple linear regression, which is a parametric approach relying on assumptions about the shape of the target variable distribution (i.e., assume a normal distribution) in the underlying population and about the form or parameters (i.e., means and standard deviations) of the assumed distribution, and regression tree, a non-parametric approach which presents the disadvantage to provide no statistical inference, but the advantage to be much more flexible than the linear regression function and to provide a hierarchy of explanatory variables best explaining the dependent variable. These methodologies are applied on a topic of overwhelming relevance in present-day Italy: rental prices for students' accommodation.

Cluster analysis: an application on air pollution



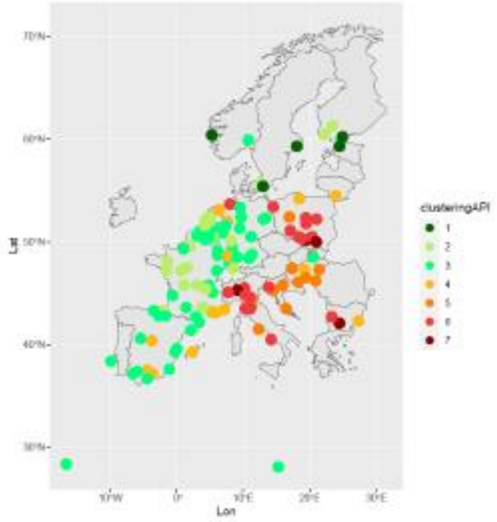
Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution

Una misura di qualità ambientale ottenuta clusterizzando un indice basato su modelli ad equazioni strutturali: una applicazione alla qualità dell'aria nelle principali città europee

Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Abstract This paper proposes an innovative computational procedure to determine the optimal number of clusters. The aim is to identify the maximum number of significantly distinct clusters, when the centroids are orderable and order is relevant. The insight is that ranking according to this optimal number of clusters allows to better classify units in order to assess their quality with regard to a variable of interest. By means of bootstrap confidence intervals estimated on clusters' centroids, the procedure allows to identify the optimal number of "well-separated" groups. The centroids are obtained applying a unidimensional k-means clustering and they allow to classify and rank the measure of an Index based on a Structural Equation Model. The procedure ranks European cities according to their level of air pollution.

Abstract *Il lavoro propone una procedura computazionale innovativa per determinare il numero ottimale di cluster. Lo scopo è identificare il numero massimo di*



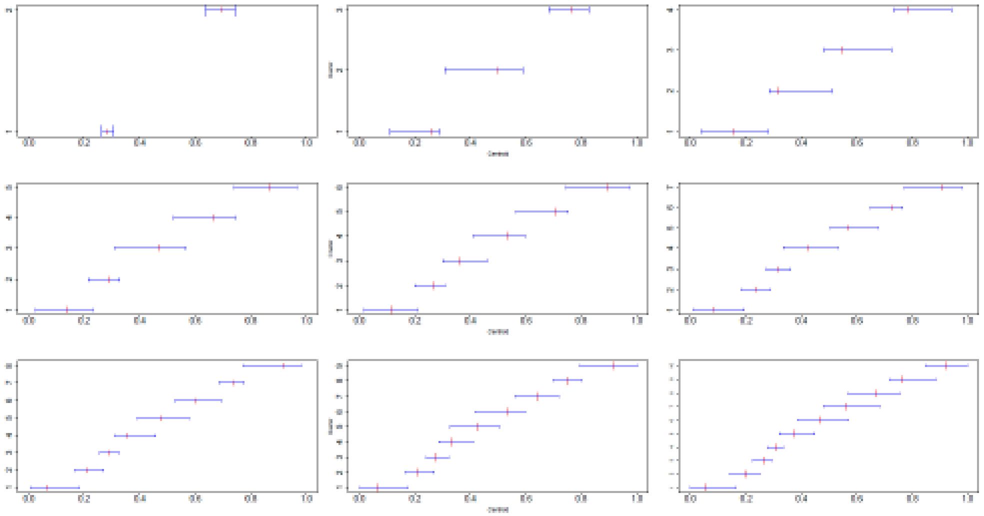
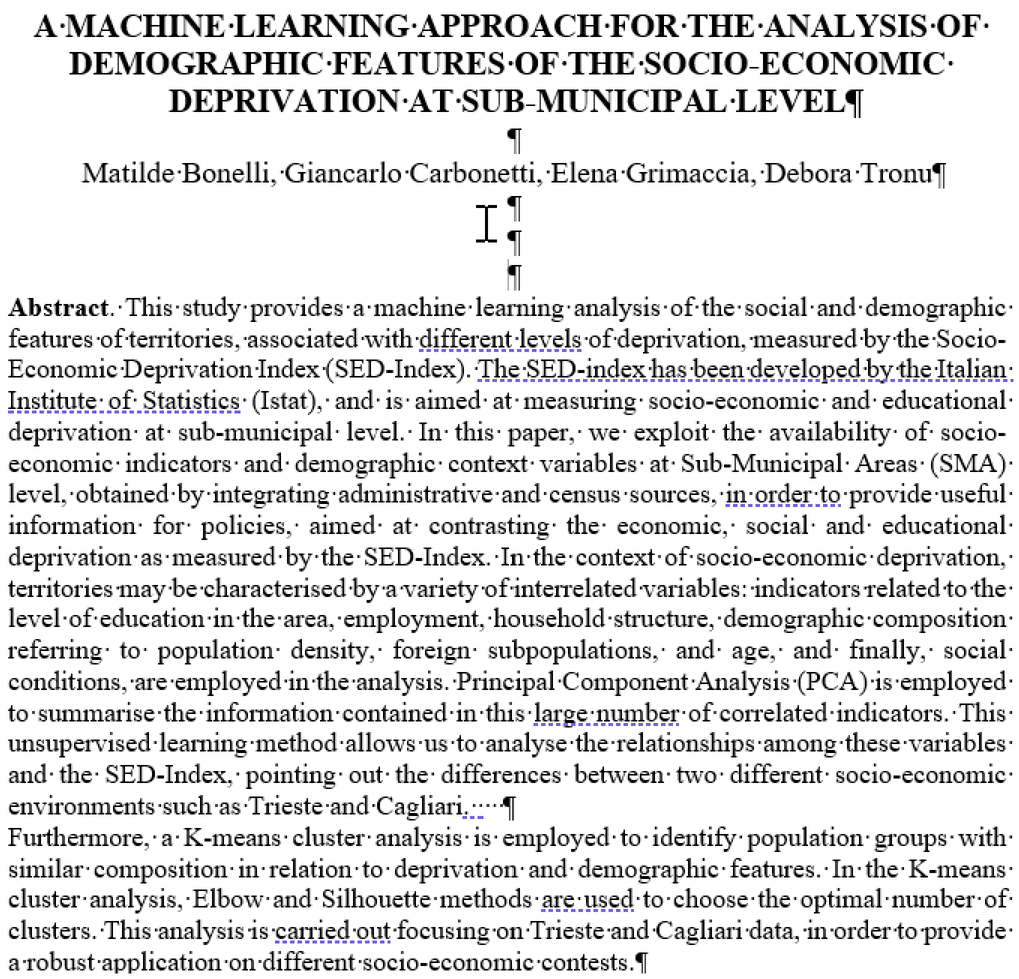


Fig. 1 95% bootstrap confidence intervals for k-means centroids. k ranges in 2-10.

Dimensionality reduction: Principal Components Analysis

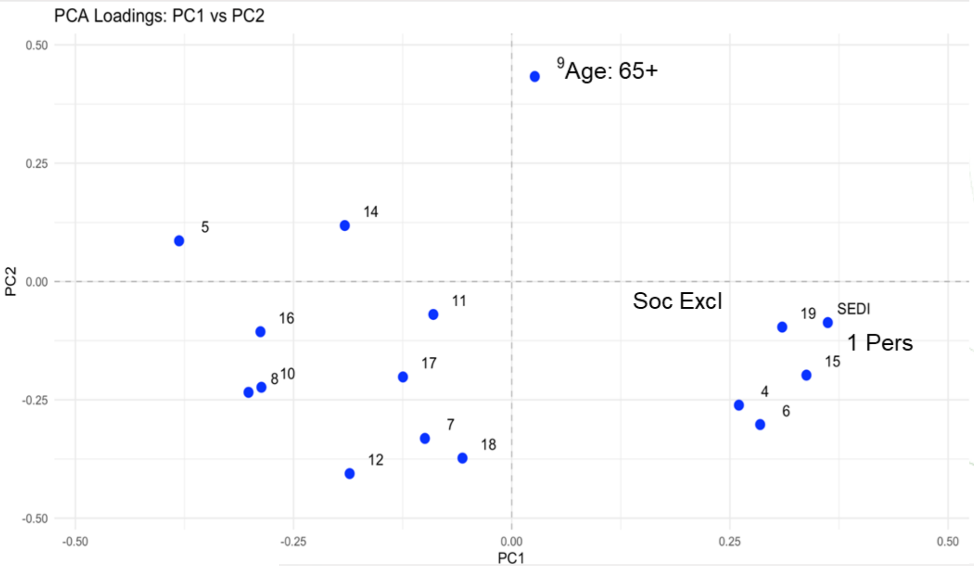


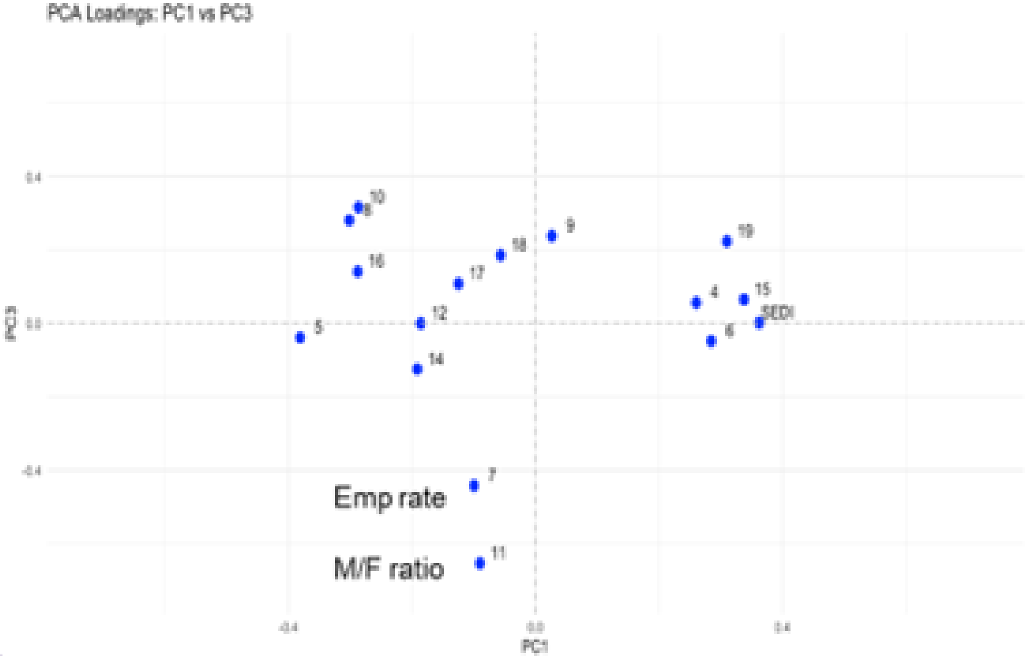
A MACHINE LEARNING APPROACH FOR THE ANALYSIS OF DEMOGRAPHIC FEATURES OF THE SOCIO-ECONOMIC DEPRIVATION AT SUB-MUNICIPAL LEVEL

Matilde Bonelli, Giancarlo Carbonetti, Elena Grimaccia, Debora Tronu

Abstract. This study provides a machine learning analysis of the social and demographic features of territories, associated with different levels of deprivation, measured by the Socio-Economic Deprivation Index (SED-Index). The SED-index has been developed by the Italian Institute of Statistics (Istat), and is aimed at measuring socio-economic and educational deprivation at sub-municipal level. In this paper, we exploit the availability of socio-economic indicators and demographic context variables at Sub-Municipal Areas (SMA) level, obtained by integrating administrative and census sources, in order to provide useful information for policies, aimed at contrasting the economic, social and educational deprivation as measured by the SED-Index. In the context of socio-economic deprivation, territories may be characterised by a variety of interrelated variables: indicators related to the level of education in the area, employment, household structure, demographic composition referring to population density, foreign subpopulations, and age, and finally, social conditions, are employed in the analysis. Principal Component Analysis (PCA) is employed to summarise the information contained in this large number of correlated indicators. This unsupervised learning method allows us to analyse the relationships among these variables and the SED-Index, pointing out the differences between two different socio-economic environments such as Trieste and Cagliari.

Furthermore, a K-means cluster analysis is employed to identify population groups with similar composition in relation to deprivation and demographic features. In the K-means cluster analysis, Elbow and Silhouette methods are used to choose the optimal number of clusters. This analysis is carried out focusing on Trieste and Cagliari data, in order to provide a robust application on different socio-economic contests.



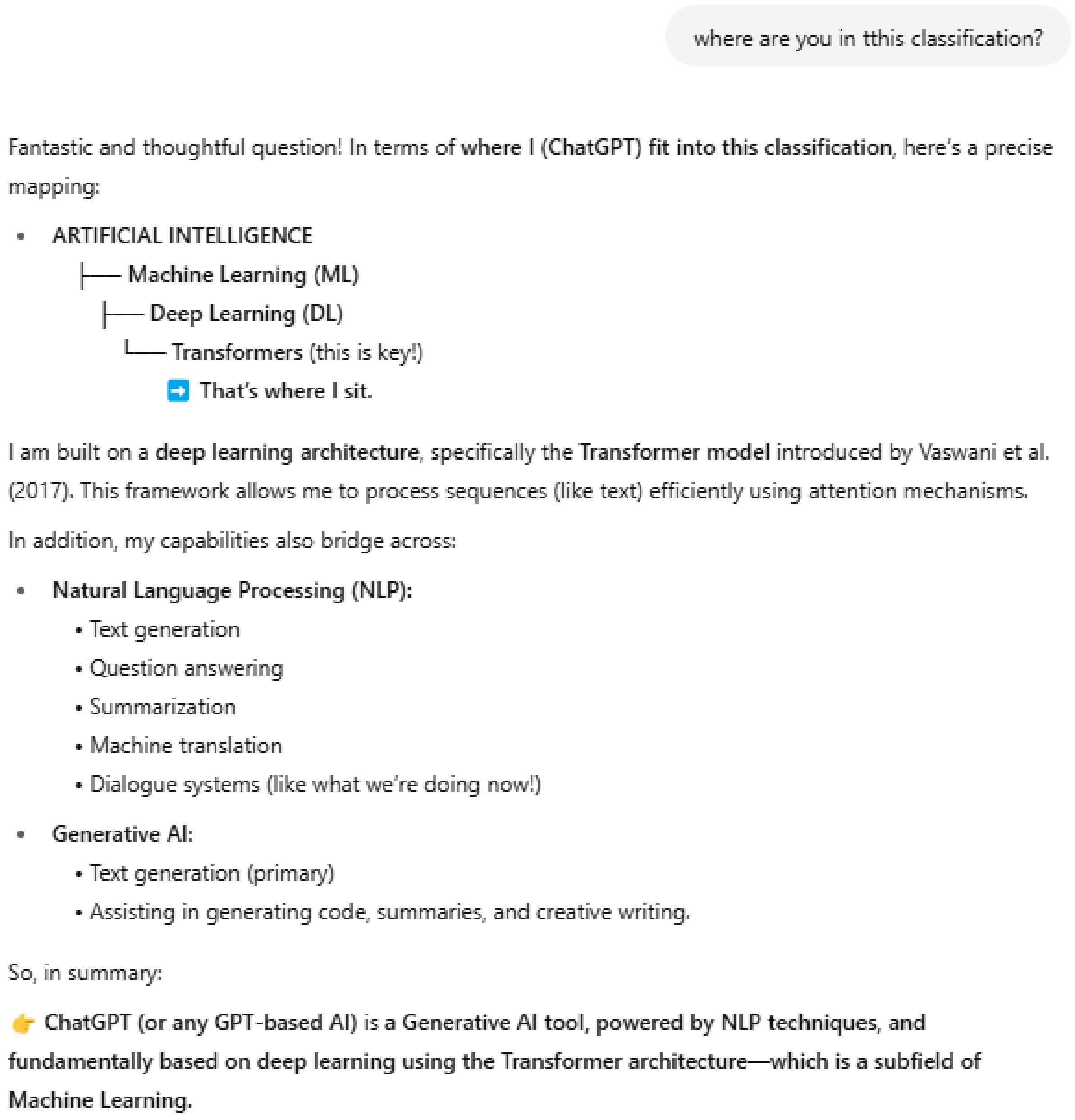


DeepLearning: Chat GPT

Chatboxes also employ:

🌐Generative AI

🌐Natural Language Processing



where are you in tthis classification?

Fantastic and thoughtful question! In terms of **where I (ChatGPT) fit into this classification**, here's a precise mapping:

- ARTIFICIAL INTELLIGENCE
  ├── **Machine Learning (ML)**
  ├── **Deep Learning (DL)**
  └── **Transformers** (this is key!)
  ➡️ **That's where I sit.**

I am built on a **deep learning architecture**, specifically the **Transformer model** introduced by Vaswani et al. (2017). This framework allows me to process sequences (like text) efficiently using attention mechanisms.

In addition, my capabilities also bridge across:

- **Natural Language Processing (NLP):**
  - Text generation
  - Question answering
  - Summarization
  - Machine translation
  - Dialogue systems (like what we're doing now!)
- **Generative AI:**
  - Text generation (primary)
  - Assisting in generating code, summaries, and creative writing.

So, in summary:

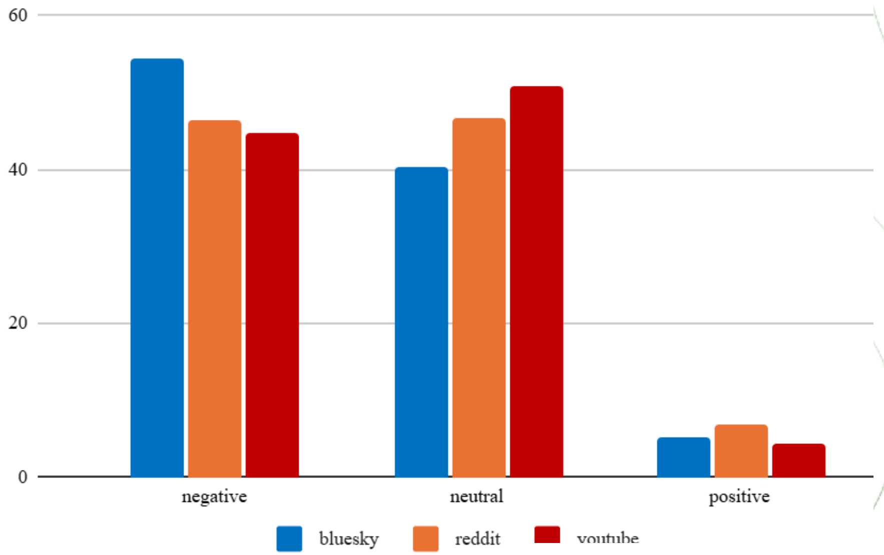
👉 **ChatGPT (or any GPT-based AI) is a Generative AI tool, powered by NLP techniques, and fundamentally based on deep learning using the Transformer architecture—which is a subfield of Machine Learning.**

Public Sentiment on Social platforms

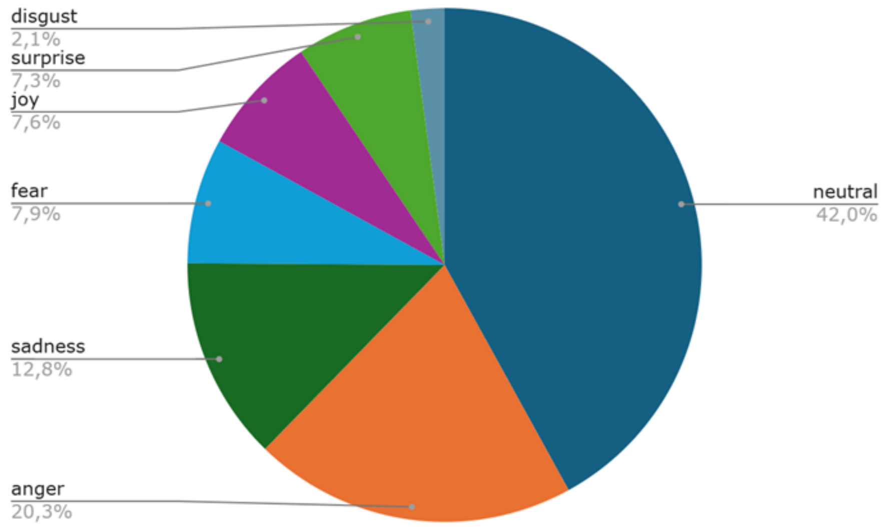
- Social Media and Public Opinion on Healthcare: insights from Google Trends, Bluesky, Reddit and YouTube
- Objective: Analyze public sentiment surrounding the UnitedHealthcare CEO killing using data from the chosen platforms.
- Data collection:
- Keywords: "Brian Thompson," "United CEO," "health insurance," "healthcare."
- Time Frame: December 4–19, 2024.

Seniment analysis and Emotion detection results

- Different online platforms reflect and shape differently public sentiment on the UnitedHealthcare Affair and healthcare in general.



- The prevailing emotions emerging from the comments on social platforms where of anger



Generative AI: The Call

Exhibition

- Holly Herndon & Mat Dryhurst
- Serpentine North Gallery, London
- 4 October 2024 - 2 February 2025
- The Call proposes new cultural, legal, and technical rituals for art in the age of AI

https://www.serpentinegalleries.org/whats-on/holly-herndon-mat-dryhurst-the-call

So…

What is AI?

In your opinion, how can we define AI?
ps: this is not a test ;)

Possible definitions of AI

There has been considerable confusion regarding a precise definition of Artificial Intelligence. The definitions have also changed over the years.

- AI is the science and engineering of making intelligent machines, especially intelligent computer programs. (Turing, 1950).
- Artificial Intelligence is the science of making machines do things that would require intelligence if done by man. (Raphael, 1976).
- Artificial intelligence (A.I.) is the study of how to make computers do things that people are better at or would be better at if they could extend what they do to a worldwide web-sized amount of data and not make mistakes. (Rich, 1985).
- The capacity to “imitating intelligent human behavior” (Kok et al., 2009).

Possible definitions of AI

Artificial Intelligence (AI) means different things to different people.

- “Some believe that AI is synonymous with any form of intelligence achieved by nonliving systems; they maintain that it is not important if this intelligent behavior is not arrived at via the same mechanisms on which humans rely” (Lucci & Kopec, 2015).
- “Artificial intelligence is a computerized system that exhibits behavior that is commonly thought of as requiring intelligence” (National Science and Technology Council - NSTC, 2016).

In AI, a computer system can perform tasks that normally require human intelligence: visual perception, speech recognition, decision making, translation between languages…..

Operational definition of AI

Key features of Artificial Intelligence

- Capable of imitating human behavior in an intelligent way
- Interacts with other machines to control and communicate with them, can also communicate with humans
- Enables automation of routine processes
- Applies various methods such as Machine Learning, Deep Learning, natural language programming, etc. to train machines to conduct tasks in human-like ways

Artificial intelligence describes developing computer systems that can carry out tasks that require human intelligence.

Hidden figures (il diritto di contare)
These ladies were working at Nasa as "computers"

https://www.youtube.com/watch?v=nWeCJTyho8Y

AI a bit of background

- Artificial intelligence describes developing computer systems that can carry out tasks that require human intelligence.
- After initial programming, nonhuman technologies with AI can learn and operate independently. Based on learned data inputs, these devices execute adaptive tasks and function with increasing degrees of autonomy from direct human direction. ML is an essential component of AI that improves machine performance
- by exposing it to cumulative data inputs.
- Machine learning and deep learning algorithms were first developed in the 1950s, initially sparking enthusiasm but remaining inactive for many decades.
- The development of AI can be divided into two stages: symbolic AI (i.e., good old-fashioned AI) and modern AI (i.e., machine learning AI).
- The 1950s to 1980s were known as the era of symbolic AI, which attempted to imitate human-level intelligence by manually building massive sets of explicit rules to deal with knowledge. Symbolic AI solves clearly defined, logical problems such as rule-based inquiries and response systems. For example, A symbolic AI-based medical diagnostic system leverages a knowledge base of explicitly defined rules to infer and deduce plausible diagnoses from the symptoms reported by a patient. For instance, the system may employ a rule that states: IF the patient exhibits a fever, a cough, and difficulty breathing, THEN the patient may be suffering from pneumonia.

Let’s discuss this topic: definition of AI

- 1. “Universal Intelligence: A Definition of Machine Intelligence” Shane Legg & Marcus Hutter (2007)This paper proposes a formal, mathematical definition of intelligence based on an agent’s ability to maximize its performance across a wide range of environments.
- 2. “A Definition of Artificial Intelligence” Dimiter Dobrev (2012)This work provides a precise, formal definition of AI without referring to human capabilities. It focuses on measurable, task-oriented behavior to define intelligence levels.
- 3. “Defining AI in Policy versus Practice” Krafft, Young, Katell et al. (2019)This paper examines the gap between how AI is defined in public policy and how it is applied in technical practice. It stresses the need for clear operational definitions in regulatory and real-world settings.
- 4. “Towards a Sociotechnical Definition of Artificial Intelligence” Johannes Dahlke (2024)Argues that definitions of AI should include not only technical features but also human purposes, institutional goals, and dynamic public expectations. A broader framework for what constitutes AI today.
- 5. “Defining Artificial Intelligence 1.0 / 2.0” – AI Watch Reports (EU JRC) European Commission Joint Research Centre (2020 / 2022)These reports offer a regulatory and operational definition of AI used in EU monitoring frameworks. They include practical taxonomies and keyword-based identification approaches for detecting AI systems in real-world settings.

Summary of Definitions

Paper	Type of Definition	Key Feature
Legg & Hutter (2007)	Formal-mathematical	Intelligence as universal performance
Dobrev (2012)	Formal-pragmatic	Task-based AI levels, non-anthropocentric
Krafft et al. (2019)	Policy vs. practice	Highlights definitional gaps
Dahlke (2024)	Sociotechnical	Includes goals, expectations, institutions
AI Watch (2020/2022)	Operational/regulatory	Taxonomy and detection criteria

Please **summarise**:

- The problem the paper addresses.
- The methodology and key contributions.
- One limitation you can identify.
- An ethical or societal consideration related to the paper.

Then,

- Present a critical review.
- Identify strengths, weaknesses, and reproducibility concerns.
- Suggest extensions or alternative approaches.

1. Universal Intelligence (Legg & Hutter, 2007)

- • Proposes a mathematical definition of AI based on performance across a wide range of environments.
- • Intelligence is defined as an agent's ability to achieve goals in different settings.
- • Highly cited in theoretical AI research.

- What is intelligence? Humans have a very questionable and uncleae idea of intelligence.
- The goal is to define “intelligence”: universal intelligence
- Eg: Kolmogorov
- Limitation: non computability!
- How can we develop a machine “IQ”?

2. A Formal Definition of AI (Dobrev, 2005)

- • Defines AI without reference to human intelligence.
- • Emphasizes task performance as a basis for intelligence.
- • Focuses on clarity, measurability, and comparability.
- • Link: arxiv.org/abs/1209.4838

- At the time the topic was not so developed
- The paper contains some personal ideas of authors, including the concept of “world”: good or bad, as well as life
- If you fail a game, “STOP”:
  - If games are not interrelated, then it is a good thing, and “life” could be continued
  - Otherwise, your game is finished

3 Policy vs. Practice (Krafft et al., 2019)

- • Analyzes how AI is defined differently in policy documents and technical practice.
- • Stresses the importance of developing clear, operational definitions.
- • Relevance for governance, regulation, and ethics in AI systems.

- Definition of AI for researchers and policy makers
- People are different: for researchers AI is a tool to find meaningful results, while policy makers can use it as a source of information
- The paper contains the results of a survey on the meaning of AI are very different between policy makers and researchers:
  - The first are worried about the reliability of the tool and possible inequalities
  - The latter see AI as a tool to achieve better or quicker results.

4. Sociotechnical Perspective (Dahlke, 2024)

- • Proposes a definition of AI that includes technical capabilities and human-centered purposes.
- • Considers public expectations, institutional settings, and dynamic goals.
- • Advocates for a broader, context-aware understanding.
- • Link: arxiv.org/abs/2410.13452

5 Operational AI Definitions (AI Watch, EU)

Provides taxonomies and keyword systems for AI detection.

- • Used in the EU’s regulatory and monitoring activities.
- • Practical approach to identify and classify AI systems.
- • Used in 'AI Watch' and AI Act documentation.

- Guidance on AI
- State of the art definition of AI
- Logical approach to categorize those definition: according to different tools employed:
  - 1. perception of the environment
  - 2. processing of decision making
  - 3. decision making
  - 4 achieving of definitive goal

5 Operational AI Definitions (AI Watch, EU)

Provides taxonomies and keyword systems for AI detection.

- They are looking for a practical way to define AI for law issues: copyrights, etc.
- At the end of the day, they do not find a precise definition but they find categories that can be useful to the definition
- AI is acquiring more independence from its developers
- AI should be able to recognize Paths!
- Taxonomies and groupings can help the identification of definition
- BUT: logical and theoretical approach was a bit too general

References

General literature

- Michie, D., Spiegelhalter, D.J., Taylor, C.C. and Campbell, J. eds., 1995. *Machine learning, neural and statistical classification*. Ellis Horwood.
- Russell & Norvig (2020). *Artificial Intelligence: A Modern Approach*.
- Goodfellow, Bengio, & Courville (2016). *Deep Learning*.
- Jurafsky & Martin (2023). *Speech and Language Processing*.
- Chollet (2017). *Deep Learning with Python*.
- LeCun, Bengio & Hinton (2015). “Deep Learning.” *Nature*.
- Floridi & Cowls (2022). *The Ethics of Artificial Intelligence*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All You Need*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008. https://arxiv.org/abs/1706.03762