Université Paris Cité

diiP Summer School

2024

Introduction
AI for Research

–
Yvonne Becherini
Université de Paris Cité
Laboratoire Astroparticule et Cosmologie
Data Intelligence Institute of Paris

Welcome! Who am I

- My name is Yvonne Becherini

- I am an Astroparticle Physicist

- I work at University of Paris Cité, Laboratoire Astroparticule et Cosmologie & Data Intelligence Institute of Paris

- My principal interests are
  - Astroparticle Physics
  - Gamma-Ray & Neutrino Astronomy
  - Extragalactic sources

- I develop Data Analysis methods with Machine Learning

- I give two courses: Applied Data Analytics and Advanced Applied Data Analytics on Zoom.

diiP Courses in data science!

Applied Data Analytics

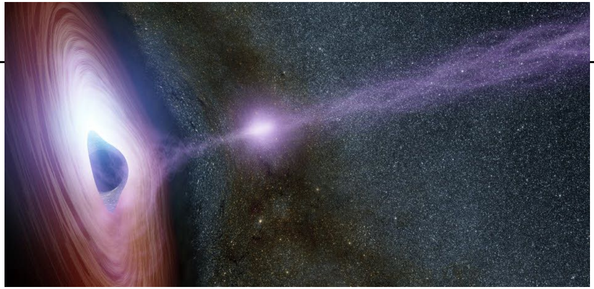
Monday 09/10/23	Tuesday 10/10/23	Wednesday 11/10/23	Thursday 12/10/23	Friday 13/10/23
10h00-11h20 **Introduction Objective of the course**	10h00-11h20 **Classic supervised Learning**	10h00-11h20 **Classic Supervised Learning Regression**	10h00-11h20 **Neural Networks**	10h00-11h20 **Unsupervised Learning & Generative Models**
11h20-11h30 Break and Poll ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕
11h30-13h00 **Data preparation**	11h30-13h00 **Classic supervised Classification**	11h30-13h00 **Notebooks**	11h30-13h00 **Deep Learning**	11h30-13h00 **Students present their data analysis challenges**

Advanced Applied Data Analytics

Monday 22/01/24	Tuesday 23/01/24	Wednesday 24/01/24	Thursday 25/01/24	Friday 26/01/24
10h00-11h20 **Introduction Objective of the course (Review of some basic topics)**	10h00-11h20 **Visualization of the latent space** **Variational Autoencoders**	10h00-11h20 **Graph Neural Networks**	10h00-11h20 **Bayesian Deep Learning**	10h00-11h20 **Self-supervised Learning**
11h20-11h30 Break and Poll ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕	*11h20-11h30 Break and Poll* ☕
11h30-13h00 **Generative Adversarial Networks**	11h30-13h00 **Transformers**	11h30-13h00 **Graph Embedding Methods**	11h30-13h00 **Bayesian Deep Learning**	11h30-13h00 **Self-supervised Learning**

... and now the Summer School! →

Monday	Tuesday	Wednesday	Thursday	Friday
	Supervised Learning	**Knowledge-guided Data Science**	**Anomaly detection for Time Series**	**Generative Adversarial Networks**
	AI in Medicine/Biology		***AI in Industry***	***AI in Particle Physics***
Large Language Models	**Representation Learning**	**High-Dimensional Vector Similarity Search**	**Graph Neural Networks**	
Social Dinner		**Bateau Mouche**		



The goal is to understand:

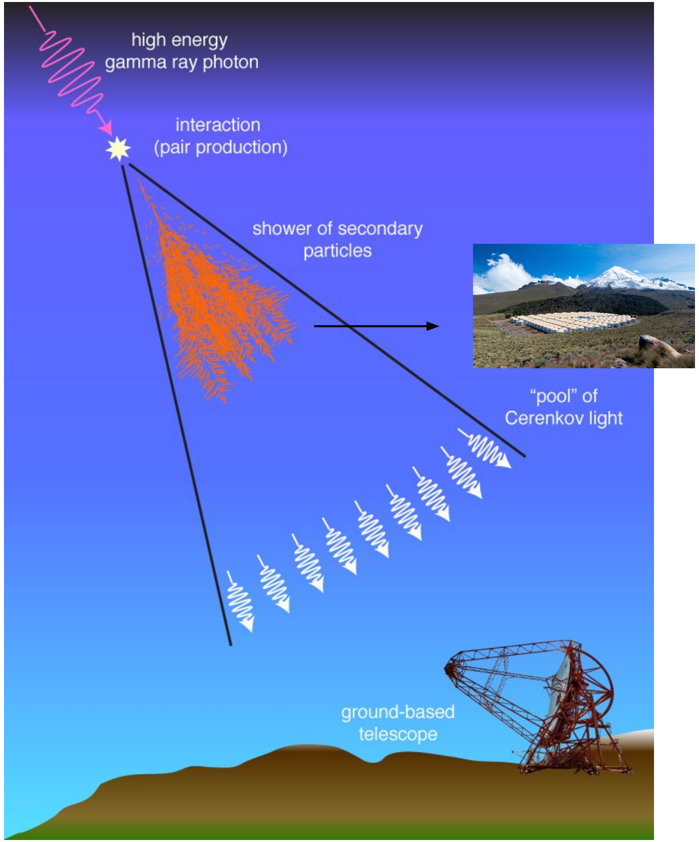
- ➔ The mechanisms of generation of energy in the Universe
- ➔ The creation and propagation of energetic particles in the Universe: gamma-rays, neutrinos, protons
- ➔ The nature of Dark Matter

Methods used:

- ➔ Observation of phenomena through ultra-precise and ultra-sensitive particle detectors
- ➔ The analysis of the data acquired is often complex for one main reason: the signal searched is tiny, compared to a huge amount of background

Very High Energy Gamma Ray Astronomy Experiments

- Observations are made:
    - With Imaging Atmospheric Cherenkov Telescopes detecting the Cherenkov light generated in the atmosphere by the passage of highly relativistic charged particles
    - Wide field of view detector arrays: Surface detectors catching the particles in the atmospheric showers
- Gamma rays from astrophysical sources are rare, and at the same time we receive a huge amount of background events from cosmic rays (very similar)
- The amount of data generated can be huge: several TB per month



Reconstruction of event parameters

After data are calibrated, we need to perform the reconstruction of the kinematics of the gamma ray: incoming direction and energy

This can be done with algorithms or using ML regression

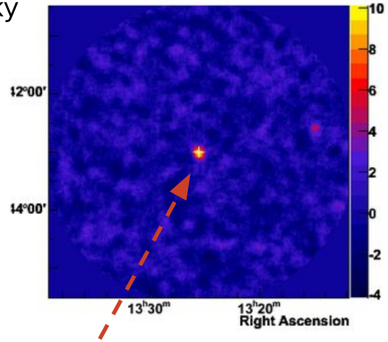
Regression is a method allowing to estimate the value of a variable associated with the signal (or the background).

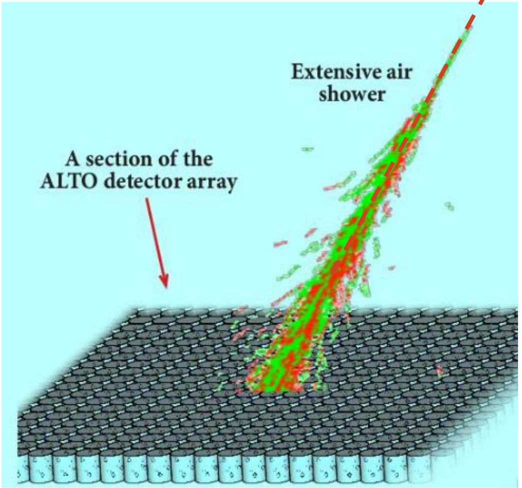
With the help of simulations and data analysis, we can infer what the values expected for a particular event are

When we run the reconstruction algorithms, we run them on all data, i.e. on all gamma ray signal and all proton background events.

Then everything is passed to the next step, the suppression of the background

Astrophysical source in the sky





What the gamma ray would look like close to the detector

Extract a “rare” signal through classification

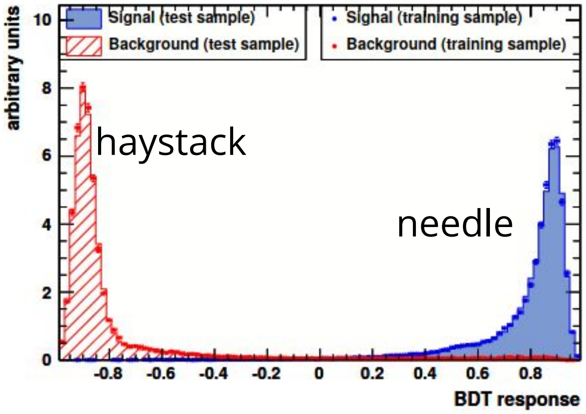
Extract a rare “signal” in the presence of a large amount of noise, which is equivalent to finding a needle in a haystack

Need to develop powerful methods to extract these rare events.

For most problems, after having cleaned properly the data, the answer is

- Event classification through user-defined features

But if you wish to achieve a better classification performance in difficult analysis regions, better to switch to Deep Learning

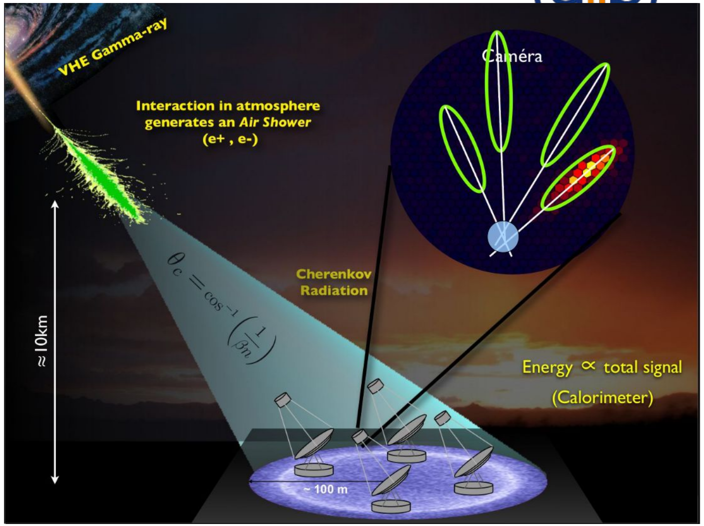


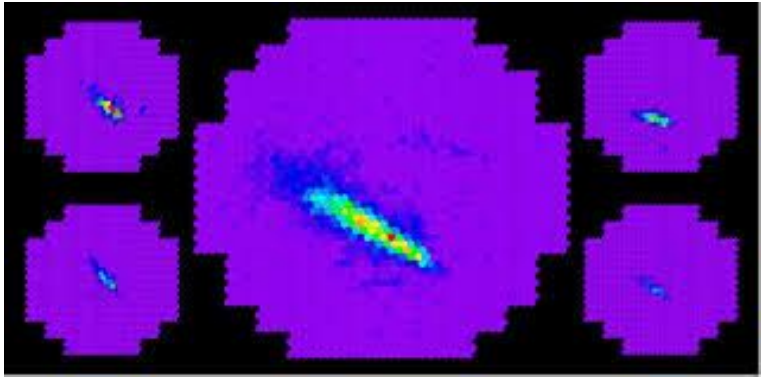
→ Y. Becherini et al., Astroparticle Physics (2011)
→ M. Senniappan, Y. Becherini et al, JINST (2021)

Gamma/Proton classification in H.E.S.S. data

The HESS telescope array is located in Namibia and is a 5-tel array of Imaging Atmospheric Cherenkov Telescopes detecting the Cherenkov light created in the atmosphere by the passage of highly relativistic charged particles.

One of the most crucial steps in the analysis of data, is the suppression of the cosmic ray background to extract the “signal” of gamma-rays.

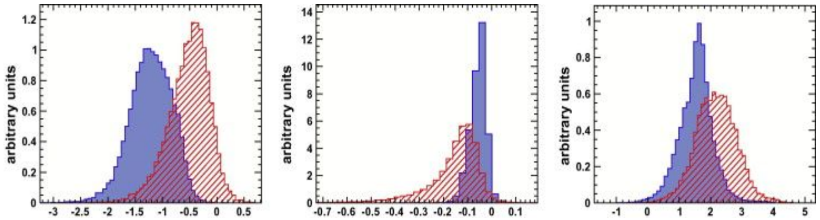


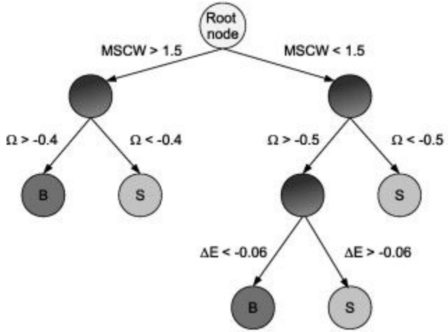


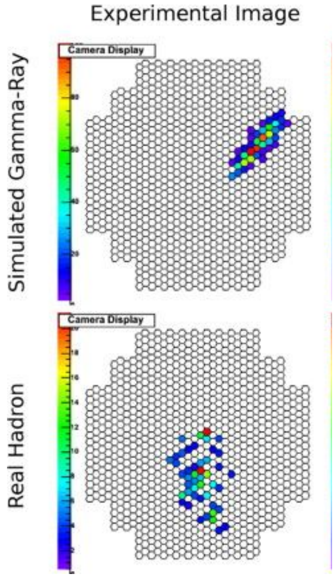
H.E.S.S.

Gamma/Proton classification in H.E.S.S. data

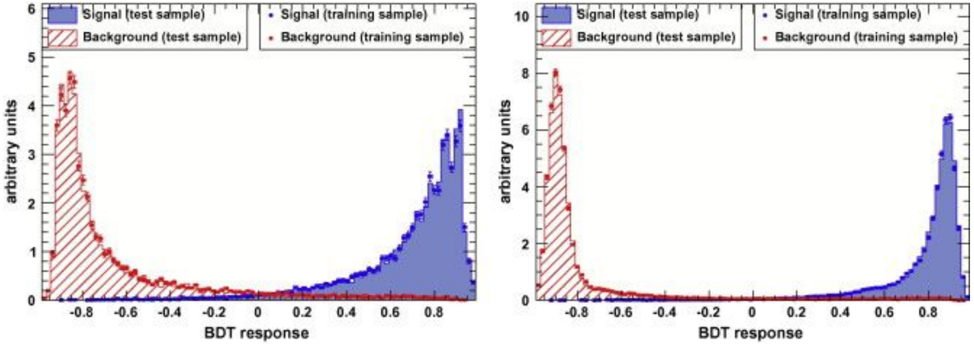
- Supervised feature-based ML
- Classification of gamma-rays and protons using a set of user-defined input variables
- The algorithm performing the separation is the Boosted Decision Trees method.



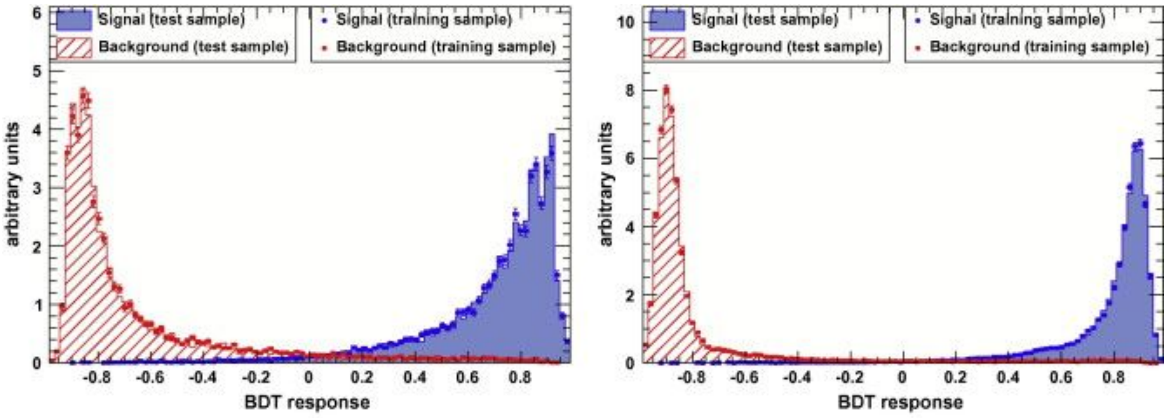




[“A new analysis strategy for detection of faint gamma-ray sources with Imaging Atmospheric Cherenkov Telescopes”, Astroparticle Physics, (2011)]



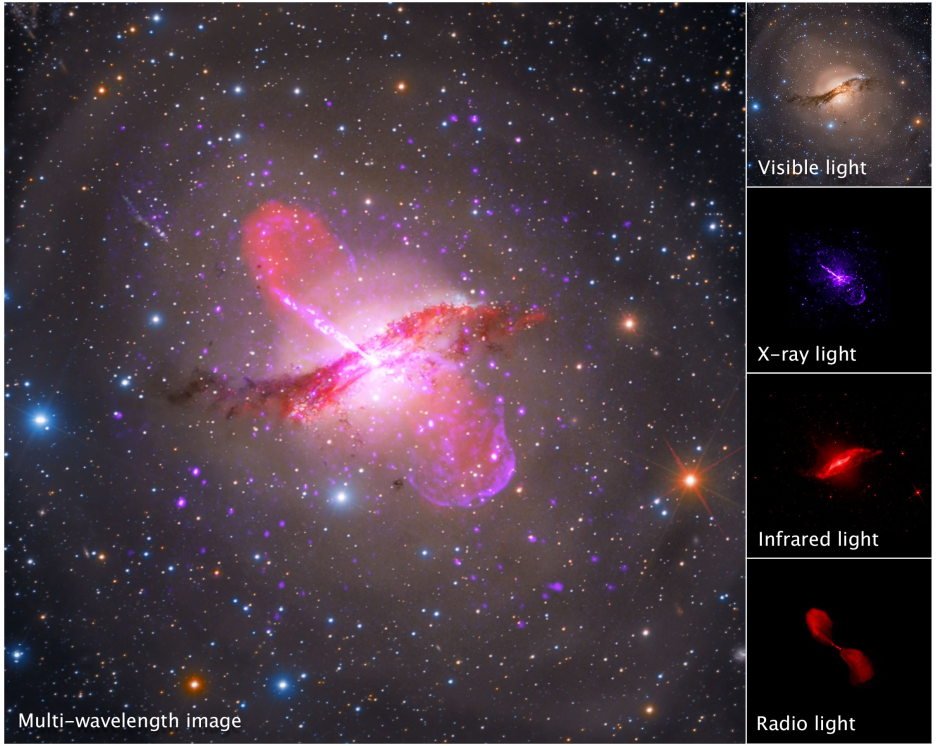
Gamma/Proton classification in H.E.S.S. data



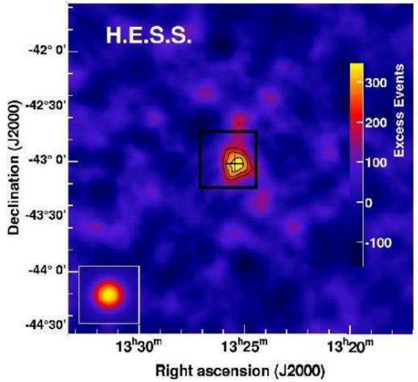
- The final response of the algorithm to an independent set of data (test data) allows defining an analysis cut before looking at the real data.
- The final analysis cut can be based on a desired gamma-ray efficiency. Example: if I say I will cut at 0.4 on the right plot, I will have 95% of gamma-rays and a contamination of less than 1% of protons.
- When the analysis cuts are frozen, you are allowed to look at the real data.

Implications of having a better event classification based on Artificial Intelligence

The Active Galactic Nucleus
Centaurus A

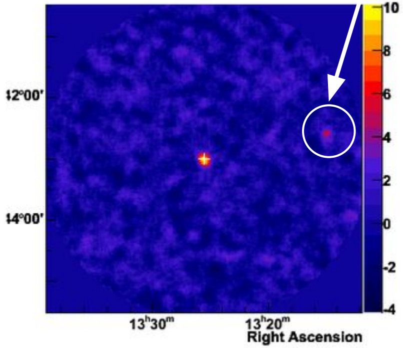


2008: discovery with a detection significance of 5σ with standard analyses (no ML)



2016: Re-analysis of data using supervised ML 9.8σ !

Appearance of a second source in the field of view!



A clear gain in the detectability of gamma-ray emitting sources

Current contributions to the field through the diiP Chair

Search for hidden information in AGN data

Enzo Oukacha, PhD student

Search for neutrino-gamma correlations in the high-energy extragalactic sky
Project **CoNIC**

CoNIC

Pranju Goswami (Postdoc) & Wilma Kiviaho (Master)

Supervised Inference for Redshift Exploration
Project **SIREX**

SIREX

Ankur Sharma (Postdoc)

Blazar Labelling and SED Tagging
Project **BLAST**

BLAST

AstrodiiP focus:

Neutrino/Gamma connection in the Extragalactic field

Max Eff, PhD student

Ankur Sharma, Pranju Goswami, Enzo Oukacha

Boost sensitivity & accelerate science

New sensitive analyses based on Graph Neural networks for Gamma-Ray and Neutrino Astronomy

Project **ADAPT**

Foula Vagena

Accelerate Discoveries (boosting) Astroparticle Physics (analysis) Techniques

Max Eff

a) Unsupervised Learning for data filtering (FiBER)
b) Supervised Learning for regression and final signal extraction

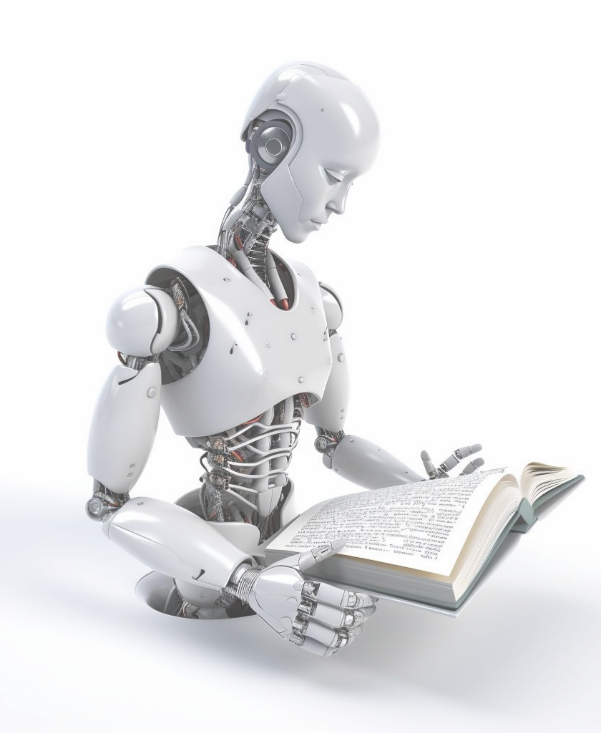
ADAPT

Two axes of research:

- Search for hidden information in data
- Boost sensitivity of current measurements

Planning

Monday	Tuesday	Wednesday	Thursday	Friday
	Supervised Learning	**Knowledge-guided Data Science**	**Anomaly detection for Time Series**	**Generative Adversarial Networks**
	AI in Medicine/Biology		***AI in Industry***	***AI in Particle Physics***
Large Language Models	**Representation Learning**	**High-Dimensional Vector Similarity Search**	**Graph Neural Networks**	
Social Dinner		**Bateau Mouche**		



Key Points

- LLMs are advanced models trained on massive text data to understand and generate human language.
- They are used in applications such as chatbots, automated content creation, and language translation.
- Examples include GPT-4 and BERT, both of which have shown impressive capabilities in language processing tasks.
- Key concepts include transformers and attention mechanisms, which allow for efficient processing of long-range dependencies in text.

Today

LLMs can automatically summarize large volumes of text, highlighting key points concisely.

This helps researchers quickly understand the core of the research, saving time and aiding in relevance assessment.

LLMs like GPT can be integrated into Electronic Health Record (EHR) systems to help doctors generate comprehensive patient notes during consultations, reducing the time spent on paperwork.

Finance sectors simplify financial reporting and customer support with LLMs.

Legal professionals use LLMs for efficient document review and legal research.

E-commerce leverages LLMs for product descriptions and customer interactions.

Education can benefit from intelligent tutoring systems and educational content generation using LLMs.

"Train on Labeled Data to Make Accurate Predictions"

1. Supervised learning utilizes **labeled data**, where each example is paired with an output label.

2. It is used for both **classification** tasks (e.g., image recognition) and **regression** tasks (e.g., predicting house prices).

3. Common algorithms include **linear regression, logistic regression, support vector machines, and neural networks**.

4. The data is typically split into **training, validation, and test sets** to evaluate model performance.

5. **Overfitting** occurs when a model learns the training data too well, capturing noise instead of the underlying pattern.

6. **Underfitting** happens when a model is too simple to capture the underlying pattern in the data.

"Discover underlying patterns and features in data for various machine learning tasks"

NLP Applications

Includes tasks like translation and sentiment analysis.

Computer Vision Applications

Includes tasks like object detection and image classification.

Autoencoders

Learn efficient codings of input data by compressing and reconstructing it.

Word Embeddings

Used in NLP to represent words in dense vector spaces, capturing semantic meanings.

Feature Discovery

Trains models to automatically discover features from raw data.

Reduced Need for Manual Engineering

Reduces the need for manual feature engineering.

"Integrate domain-specific knowledge into data science for improved accuracy and better interpretability"

Uses techniques like **rule-based systems** and hybrid models for better decision-making processes.

Employs concepts like **knowledge graphs** to represent and utilize complex relationships in data.

Applied in fields such as **healthcare** and **environmental science** to leverage specific domain expertise.

Enhances **data-driven insights with expert knowledge** for more accurate and reliable outcomes.

Integrates domain knowledge into data science to enhance **model performance and interpretability**.

Facilitates better understanding of data patterns and trends through domain-specific knowledge.

"Find Similarities Fast in Complex Data with High-Dimensional Vector Search"

- Techniques like Approximate Nearest Neighbors (ANN) help in finding similar vectors in high-dimensional spaces.
- Locality-Sensitive Hashing (LSH) is used to hash input items so that similar items map to the same buckets with high probability.
- Applications include recommendation systems where user preferences are matched with similar items, and image retrieval systems that find images with similar features.
- Key concepts include vector spaces, where data points are represented as vectors, and similarity metrics, which measure how alike two vectors are.

"Identify unexpected patterns and outliers in time series"

- Anomaly detection in time series identifies unusual patterns that deviate from expected behavior.
- Methods include statistical models (e.g., ARIMA), machine learning models (e.g., isolation forest), and deep learning models (e.g., LSTM).
- Applications include financial fraud detection, where unusual transactions are flagged, and equipment failure prediction, where abnormal sensor readings indicate potential issues.
- Key concepts involve understanding time series data, seasonality (regular patterns), and trend analysis (long-term movement).

GNNs are neural networks designed to work with graph-structured data, capturing relationships between nodes and edges.

Techniques include Graph Convolutional Networks (GCNs) which apply convolution operations on graphs, and Graph Attention Networks (GATs) that use attention mechanisms.

Applications include social network analysis, where GNNs help to understand user interactions, and drug discovery, where they model molecular structures.

Key concepts include nodes (entities), edges (relationships between entities), and graph embedding (vector representation of nodes in a graph).

GANs consist of two neural networks, a generator and a discriminator, that compete against each other.

The generator creates fake data samples, while the discriminator tries to distinguish between real and fake samples.

Applications include image generation, data augmentation, and style transfer.

Key concepts involve adversarial training, where the generator and discriminator are trained simultaneously.

Recap of Basic Topics in Deep Learning

Training Process

Involves techniques like backpropagation for error correction and gradient descent for optimizing the model. Key for model accuracy.

Feedforward Networks

Simplest type, where connections do not form cycles. Used in applications like image and speech recognition.

Neural Networks

Basic types include feedforward, convolutional, and recurrent neural networks, each suited for different tasks.

Convolutional Networks

Specialized for processing grid-like data, such as images. Utilize convolutional layers to detect features.

Recurrent Networks

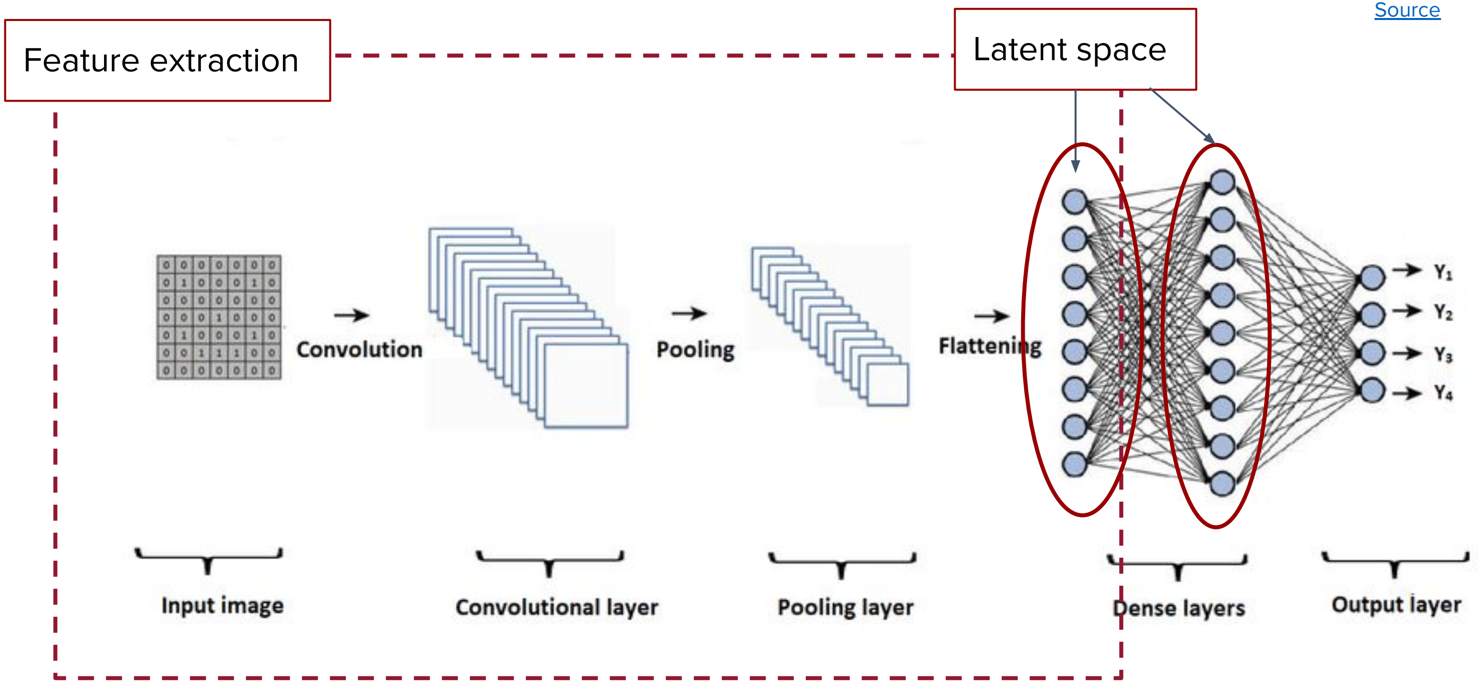
Designed for sequential data, like time series or text. They have loops to maintain memory of previous inputs.

Regularization Techniques

Methods like dropout and batch normalization help prevent overfitting and improve model generalization.

Convolutional Neural Networks

Source



The latent space is an abstract multidimensional space that encodes a meaningful internal representation of externally observed events

Samples that are similar “in the external world” are positioned close to each other in the latent space.

Latent space

In the feature extraction phase, the model has captured the important patterns of the input that are needed for the image classification task.

In the latent space, images that depict the same object have very close representations.

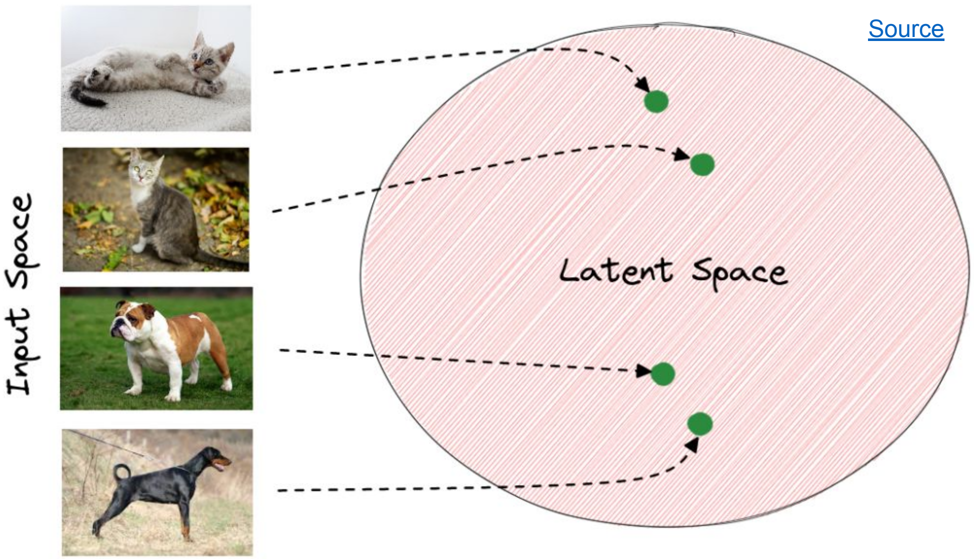
Generally, the distance of the vectors in the latent space corresponds to the semantic similarity of the raw images.

The green points correspond to the latent vector of each image extracted from the last layer of the model.

We observe that vectors of the same animals are closer to the latent space.

Therefore, it is easier for the model to classify the input images using these feature vectors instead of the raw pixel values.

Source



But ... difficult to interpret the latent space meaning

Recap of Basic Topics in Data Analytics

Normalization

Scaling data to a standard range to improve the performance of machine learning algorithms.

Cleaning

Removing or correcting errors and inconsistencies in the data to ensure quality.

Data Preprocessing

Involves cleaning, normalization, and transformation of raw data to prepare it for analysis.

Transformation

Converting data into a suitable format or structure for analysis.

Exploratory Data Analysis (EDA)

Using statistical summaries and visualization techniques to understand data patterns and distributions.

Model Evaluation

Assessing model performance using metrics like accuracy, precision, recall, and cross-validation.

Coffee Break