# Artificial Intelligence applied to enviromental monitoring

## Technical challenges and limitations of environmental AI

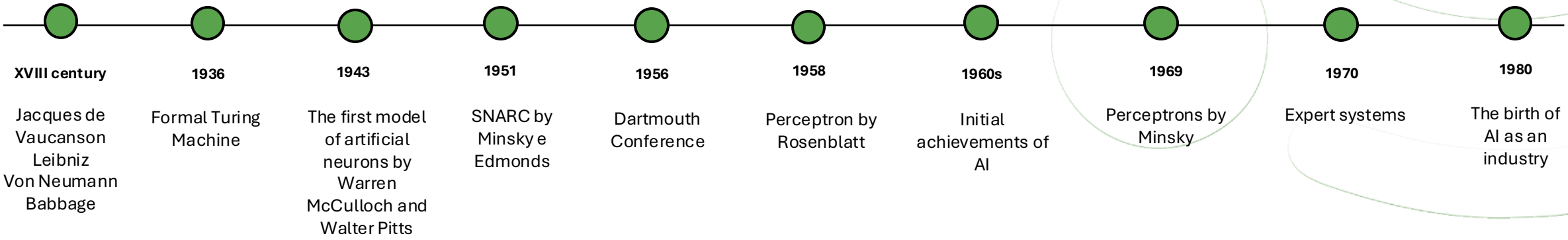
Vittoria Mascellaro

**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”

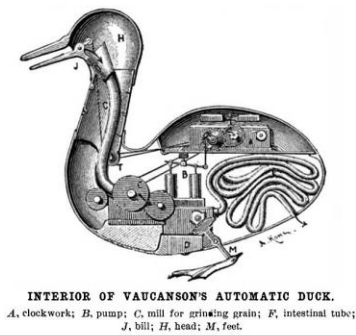
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

# Module 1: AI and Data

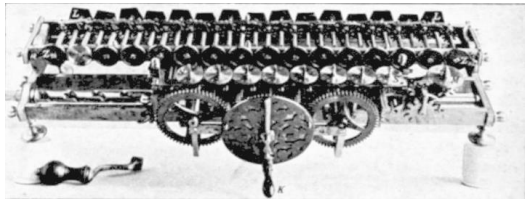
# Historical perspective



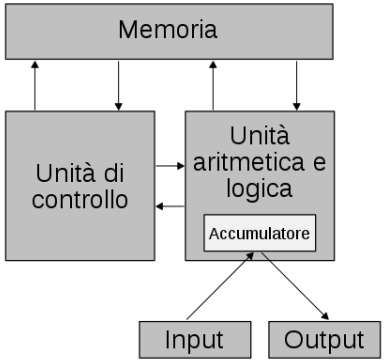
# AI Ethics


INTERIOR OF VAUCANSON'S AUTOMATIC DUCK.
*A*, clockwork; *B*, pump; *C*, mill for grinding grain; *F*, intestinal tube; *J*, bill; *H*, head; *M*, feet.

The *Digesting Duck* by Jacques de Vaucanson (1738)

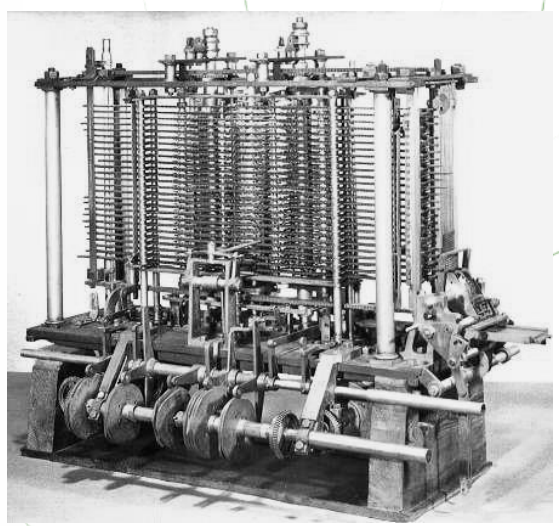


Calculus ratiocinator By Leibniz




Von Neumann architecture

Charles Babbage's Difference Engine



Charles Babbage's Analytical Engine

- These figures allow us to refer to the tradition of formalist research.
- This tradition helps us understand how artificial performance is part of human practice.
- In particular, through the projects of mathematician Charles Babbage, we see the human tendency to self-imitate using machines.
- Mathematician Ada Lovelace, in 1840, recognized the potential of Babbage's Analytical Engine.
- Lovelace was interested in the machine's ability to process symbols that could represent all objects.
  She was the first to foresee the advent of a form of Artificial Intelligence.
- Artificial Intelligence was possible, but it was not yet clear how to achieve it.

**ARTIFICIAL INTELLIGENCE**

**=**

**THE SCIENCE THAT ADDRESSES THE PROBLEM OF HOW TO REPRESENT AND BUILD KNOWLEDGE**

# Data as the foundation

- Artificial Intelligence relies on large volumes of data.
- Data fuels machine learning models: **more data → better accuracy**.
- **Raw data → Information → Knowledge → Automated decisions**
- AI doesn't just analyze data — it transforms it into **intelligent actions** (e.g., predictions, recommendations, classifications).

# Definition of data

Data are **original representations** — that is, not yet interpreted — **of a phenomenon, event, or fact**, conveyed through symbols, combinations of symbols, or any other expressive form associated with a medium

# Definition of data

Data are **original representations** — that is, not yet interpreted — **of a phenomenon, event, or fact**, conveyed through symbols, combinations of symbols, or any other expressive form associated with a medium

Data are representations of events or facts:

– **Not interpreted (original)**
– **Expressed through symbols (or combinations of symbols)**
– **Stored or conveyed on some medium (expressive form)**

# Structured data vs Unstructered data

*Structured data* refers to data that follows a predefined and expected format → as a table in a database, with columns for name, date, temperature — each entry follows a set structure

**VS**

*Unstructured data* lacks a predefined format (e.g.Podcast, video files…)

# Quality of data

**HOW GOOD IS THIS DATA?**

# What makes data "High quality"?

- **Accuracy**
- **Consistency**
- **Timeliness**
- **Completeness**
- **Spatial and temporal resolution**
- **Metadata and documentation**

# Data Quality and validation according to ISTAT

According to ISTAT, the final output of a statistical survey can be broken down into **three levels of information**:

- **Microdata** = individual data points
- **Macrodata** = statistical summaries
- **Metadata** = documentation about the data

Together represent the **statistical information** produced by a survey.
That's why ISTAT refers not just to **data quality**, but more broadly to the **quality of information >** we must define what "quality" means at **each of the three levels** — individual data, aggregated results, and metadata.

# Data Quality and validation according to ISTAT

ISTAT adopts a definition of quality originally proposed by **O. Arkhipoff** in 1986:

**"The quality of a product is its ability to meet the guarantees provided by the producer."**

These guarantess includes both the **design characteristics and tolerance**

# Data Quality and validation according to ISTAT

**Design guarantess**

1. Timeliness
2. Theoretical relevance
3. Effective relevance
4. Transparency
5. Tolerence

**Tolerance guarantess**

1. Sampling precision
2. Non-sampling precision

# Dimensions of data quality

| Dimension | Definition | Defined by |
|---|---|---|
| **1. Relevance** | The extent to which statistics meet the real needs of users. | Eurostat |
| **2. Accuracy** | The closeness between statistical estimates and the true values. | Eurostat |
| **3. Timeliness** | The delay between the reference period and the availability of data. | Eurostat |
| **4. Punctuality** | The degree to which data is released according to the planned schedule. | Eurostat |
| **5. Accessibility** | The ease with which users can access the data. | Eurostat |
| **6. Clarity (Transparency)** | The clarity of presentation and documentation, enabling users to understand and interpret data. | Eurostat |
| **7. Comparability** | The possibility of comparing data across time, regions, or countries. | Eurostat |
| **8. Coherence** | The internal consistency of data and its compatibility with other datasets. | Eurostat |
| **9. Completeness** | The extent to which required data are available without gaps. | Eurostat |
| **10. Confidentiality Protection** | Ensuring the privacy of respondents and secure handling of individual data. | **ISTAT** (added) |

# Data validation

Data validation involves examining all the characteristics that define the **dimensions of data quality**, and it has two main objectives:

a) To assess whether the **quality of the data is sufficient** for public dissemination.

b) To identify the **most significant sources of error**, and to introduce changes in the production process in order to reduce errors in future surveys.

**Four key validation measures:**

**Facilitating user assessments**



**TRANSPARENCY**

**Calculating process quality indicators**

**Estimating the main components of the error profile**

**Conducting consistency studies**

# Data validation

According to a definition provided by **Marescotti (1985)**, **environmental information** has three fundamental characteristics:

1. **Complexity**
2. **Uncertainty**
3. **Conflict**

# Data abundance vs data scarcity

- **Data Abundance:**
  When there is a large volume of data available, often from multiple sources, sometimes even overwhelming in size.
  *Example:* Social media data, satellite imagery, sensor networks producing continuous streams of information.
- **Data Scarcity:**
  When data is limited, either in quantity, quality, or both. This can happen due to cost, accessibility, or rarity of events.
  *Example:* Rare disease cases, remote environmental measurements, early-stage research data.

# Challenges of Big Data:

STORAGE

PROCESSING

NOISE

# Challenges of Small Data:

OVERFITTING

LACK OF REPRESENTATIVENESS

# Group activity: Exploring data quantity challenges

## Group 1

**Small Sample Size**

- **Scenario:** A city wants to model traffic flow but only has traffic count data from 3 days in a year.
- **Challenge Questions:**
  - What problems might arise using such a small sample?
  - How might this affect the model's reliability and predictions?
  - What strategies could improve data quantity or address this issue?

## Group 2

**Missing Data**

- **Scenario:** A weather dataset has temperature readings for every day, but 20% of the data is missing randomly.
- **Challenge Questions:**
  - How could missing data impact analysis?
  - What are possible risks when building models with this dataset?
  - What are common techniques to handle missing data?

## Group 3

**Uneven Sampling**

- **Scenario:** Environmental sensors are deployed in a forest, but some sensors record data hourly while others record daily.
- **Challenge Questions:**
  - What challenges could uneven sampling frequencies cause?
  - How might this bias the results or the model?
  - How could you standardize or correct this inconsistency?

## Group 4

**Excessive Data / Overfitting Risk**

- **Scenario:** A model uses a very large dataset with thousands of features but limited observations (high dimensionality).
- **Challenge Questions:**
  - What issues can arise from having too many features relative to data points?
  - How can this affect the model's performance?
  - What approaches can reduce this risk?

# Bias

The occurrence of distorted outcomes due to human prejudices that alter the original training data or the AI algorithm itself, leading to skewed and potentially harmful outputs

# Types of AI bias

- Algorithmic bias
- Cognitive bias
- Confirmation bias
- Exclusion bias
- Measurement bias
- Out-group homogeneity bias
- Prejudice bias
- Recall bias
- Sampling/Selection bias
- Stereotype bias

# Algorithms that amplify biases in present data

- Algorithms can not only **absorb** those biases, but actually **amplify** them

**How does this happen?**

1. **Learning from biased data**
2. **Reinforcing existing trends**
3. **Feedback loops**

**Why does this matter?**

We risk making inequalities worse

**Biases that were once hidden in society can become embedded in technology.**

# **Group activity:** Invisible pollution

# The problem

After a year of implementation, environmental NGOs and citizen groups noticed something strange. The model consistently reported **lower pollution levels** in certain low-income neighborhoods — despite clear evidence of heavy traffic, industrial activity, and frequent respiratory issues reported by local clinics.

# Investigation findandings

- These disadvantaged areas had **fewer air quality sensors**, due to underinvestment.
- Training data was heavily weighted toward **central, wealthier zones**.
- The model assumed that areas with green spaces nearby had low pollution — but failed to account for illegal waste burning or aging heating systems common in poorer districts.
- Complaints from residents were not included in the model's input, as they were seen as "anecdotal" data.

# Discussion question

- **What types of bias are present in this case?**
- **How could these biases affect policy and public health?**
- **What changes would you suggest to make the model more fair and accurate?**
- **Can “less data” about an area be considered a form of bias in itself? Why or why not?**
- **How can local communities be involved in improving these models?**

# Divide in small teams

- City government
- Data scientists
- Local community representatives
- Environmental health experts

**Each group must:**

1. Identify their key concern
2. Propose one concrete change to the AI model or the data collection process

# How to mitigate bias

**At the data level (pre-processing)**

- **Data augmentation**: generate new examples for underrepresented groups
- **Resampling**: apply over/undersampling to balance classes/groups
- **Reweighing**: assign different weights to samples based on their group membership

**At the model level (in-processing)**

- **Fairness-aware algorithms**: models designed to incorporate fairness constraints
- **Fairness regularization**: penalize bias during training

**At the output level (post-processing)**

- **Equalized odds post-processing**: adjust predictions to reduce bias
- **Threshold optimization** for different groups
- **Reject option classification**: alter decisions in uncertain cases to promote fairness

# Case study: 4 dimensional enviromental-monotoring at ECMWF

A standard method to estimate observational bias in satellite observations is to monitor first-guess departures for a certain period of time

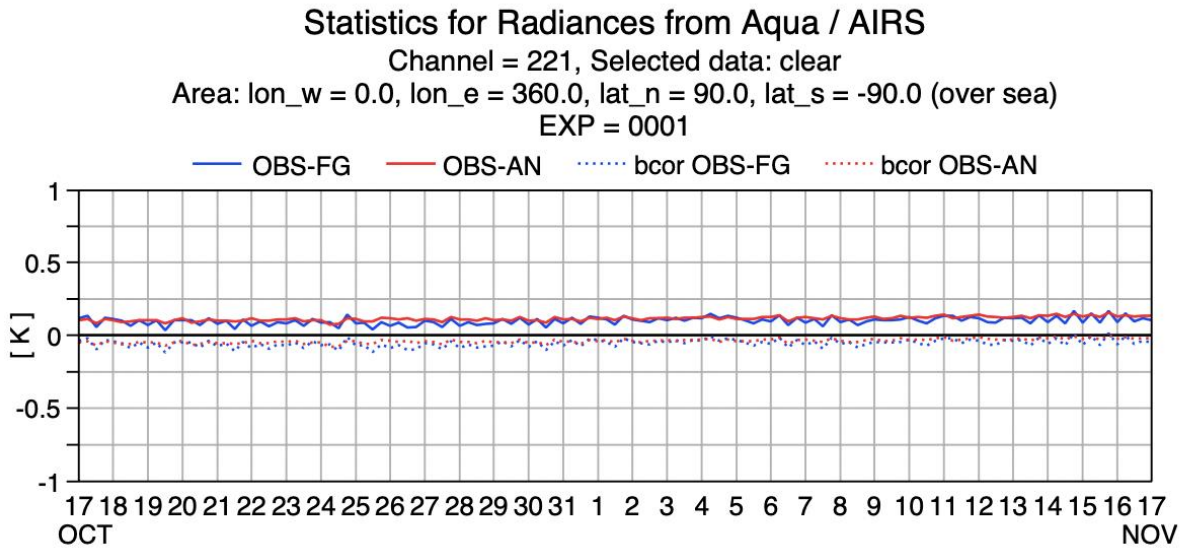



*Figure 1: AIRS monitoring.*

# Case study: 4 dimensional enviromental-monotoring at ECMWF

- Use clear-sky data to isolate systematic differences
- Differences caused by:
  - Observation errors (e.g. radiative transfer)
  - Model errors (minimized near radiosonde data)

**AIRS $CO_2$ Channel**

- Bias: small and stable → easy to correct
- $CO_2$ signal ≈ bias magnitude → risk of removing real signal or misinterpreting bias

# Case study: 4 dimensional enviromental-monotoring at ECMWF

**Bias in $CO_2$ estimates due to cloud detection issues**

- Cloud detection for AIRS radiances generally works well
- However, assumption of no systematic errors fails in tropical convective regions
- Thin cirrus clouds (allowing atmosphere/surface visibility) are hard to detect
- Large systematic errors in background water vapor profiles affect lowest peaking longwave channels (water vapor sensitive)

# BREAK