IJCAI-ECAI 2026 TUTORIAL 7

Data-Centric AI: Addressing Missing Data Imputation in Image and Tabular Contexts

Joana Cristo Santos[1], Ricardo Cardoso Pereira[2], Pedro Henriques Abreu[2]

[1] University of Lisbon, Portugal (LASIGE) [2] University of Coimbra, Portugal (CISUC)

jcsantos@ciencias.ulisboa.pt, rdpereira@dei.uc.pt, pha@dei.uc.pt

Funded by the PRR Project NEURASPACE - AI Fights Space Debris, 02/C05-i01.01/2022.PC644877546-00000020.

Overview

Theory

Missingness mechanisms (MCAR, MAR, MNAR), evaluation strategies, and the trade-offs of removal vs. imputation.

Methods

From statistical baselines through autoencoders to GANs, both for tabular records and image reconstruction.

Practice

Two hands-on Python sessions on real-world tabular and healthcare imaging data, comparing methods end-to-end.

tinyurl.com/ijcai26-t7

“

The world's most valuable resource is no longer oil, but data.

— The Economist, Cover story (May 6, 2017)

Free Dataset Repositories

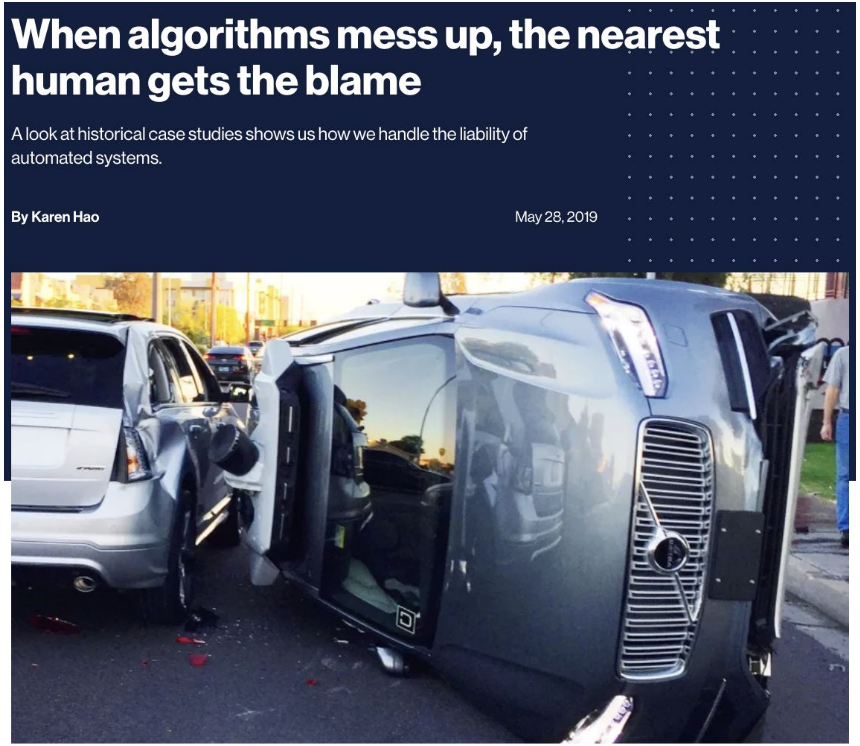
A small sample of the many public sources now available to practitioners

1. Kaggle
2. UCI ML Repository
3. NASA Open Data
4. Reddit Datasets
5. Amazon AWS Open Data
6. Google Cloud Platform
7. data.world
8. Quandl
9. Weather Underground
10. Quantopian

At this point, we might be led to think we are living in a perfect world of abundant, clean data. The reality is not so perfect.

Quality ≠ Quantity

When Algorithms Are Trained With Erroneous Data



Data Quality Is a Growing Concern

- Recent attention has turned to the role of low-quality data in the gap between **AI proof-of-concept** and **production**
- Andrew Ng and others have identified this as a key obstacle to machine learning models succeeding when deployed in the real world

> Recently, however, more attention has been paid to the role of low-quality data is playing in what Ng has identified as the proof-of-concept to production gap, or the inability of AI projects and machine learning models to succeed when they are deployed in the real world.

The Data-Centric AI Movement

dc ai DATA-CENTRIC AI

Model-Centric AI vs. Data-Centric AI

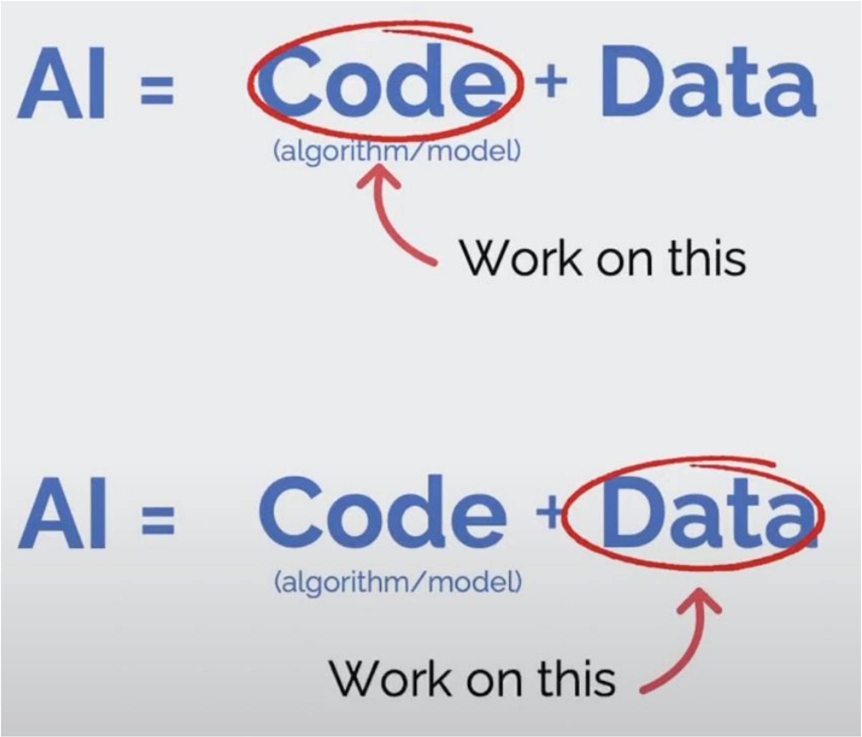
Model-Centric AI

- Given a fixed dataset, try to produce the best possible model
- Improve performance on a task by changing the model itself - the loss function, hyper-parameters, architecture, and so on

Data-Centric AI

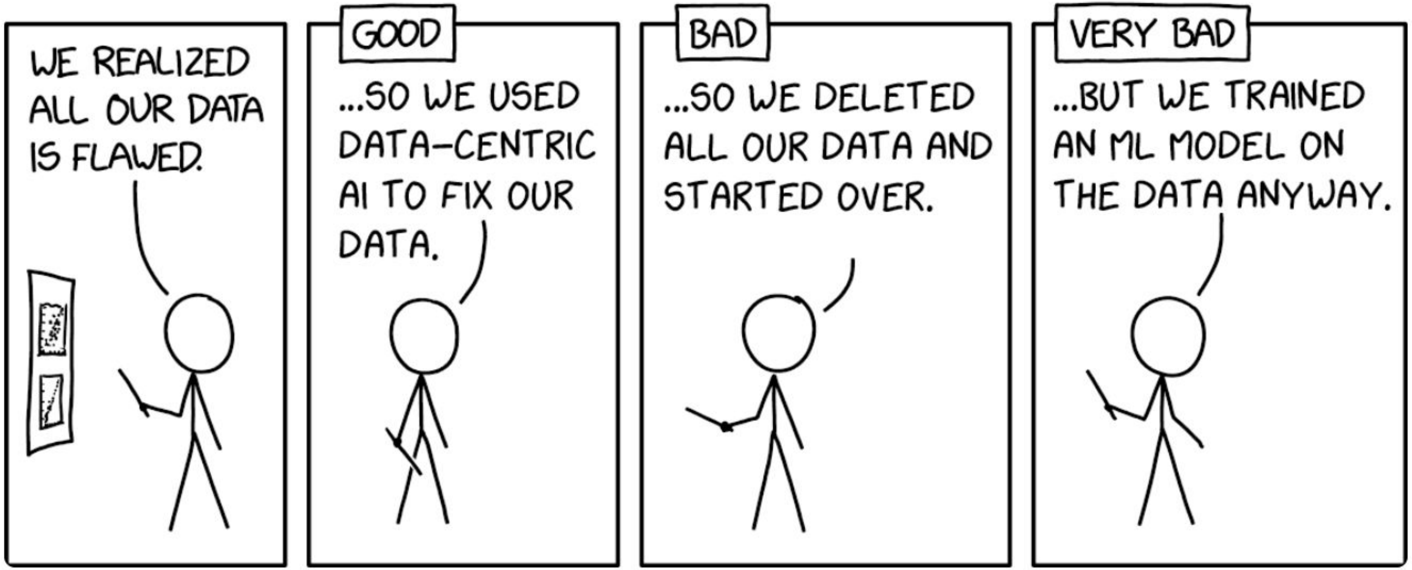
- Given a fixed model, try to improve the training dataset
- Systematically and algorithmically change the dataset to improve performance on a task

Shifting From Model-Centric to Data-Centric AI



Andrew Ng, Data + AI Summit, 2022, AI = Code + Data

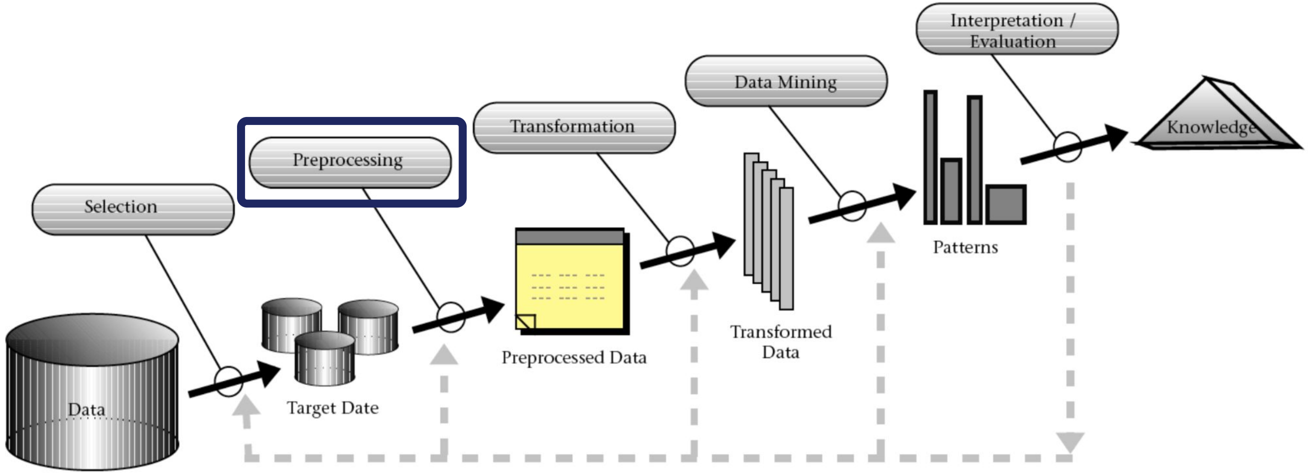
A Familiar Story



Inspired by XKCD 2494 “Flawed Data”

Data Challenges

- Imbalanced Data - when classes are not equally represented
- **Missing Data - when observations are incomplete (this tutorial)**
- Among others …



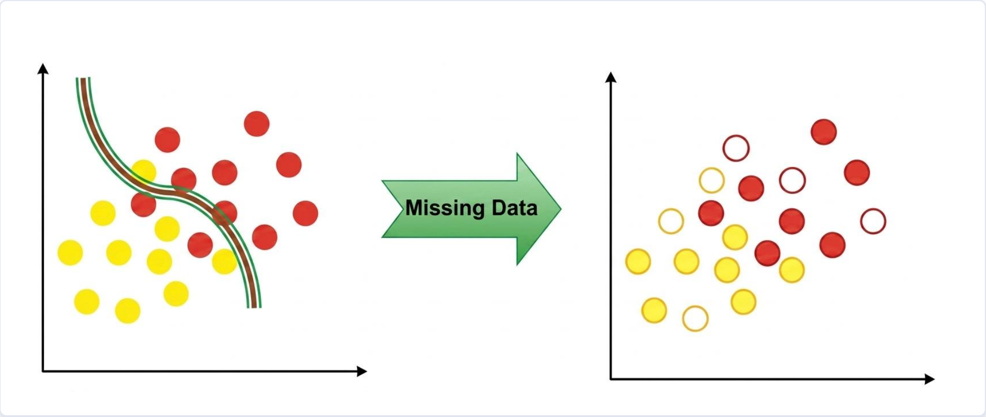
KDD Process, Fayyad et al., 1996

Missing Data

Mechanisms and Imputation Pipeline

What Is Missing Data?

- There are many possible definitions in the literature
- In general: missing data, or data incompleteness, is characterised by absent observations within a dataset, across one or multiple features
- Left unaddressed, missing values bias downstream analysis and degrade classifier performance



Examples of Possible Situations

- A physician did not record a piece of information about a patient by mistake
- A faulty scale failed to send some measurements to the central system
- A survey respondent did not answer a particular question - typically one related to demographic data
- **These three situations are completely different!**

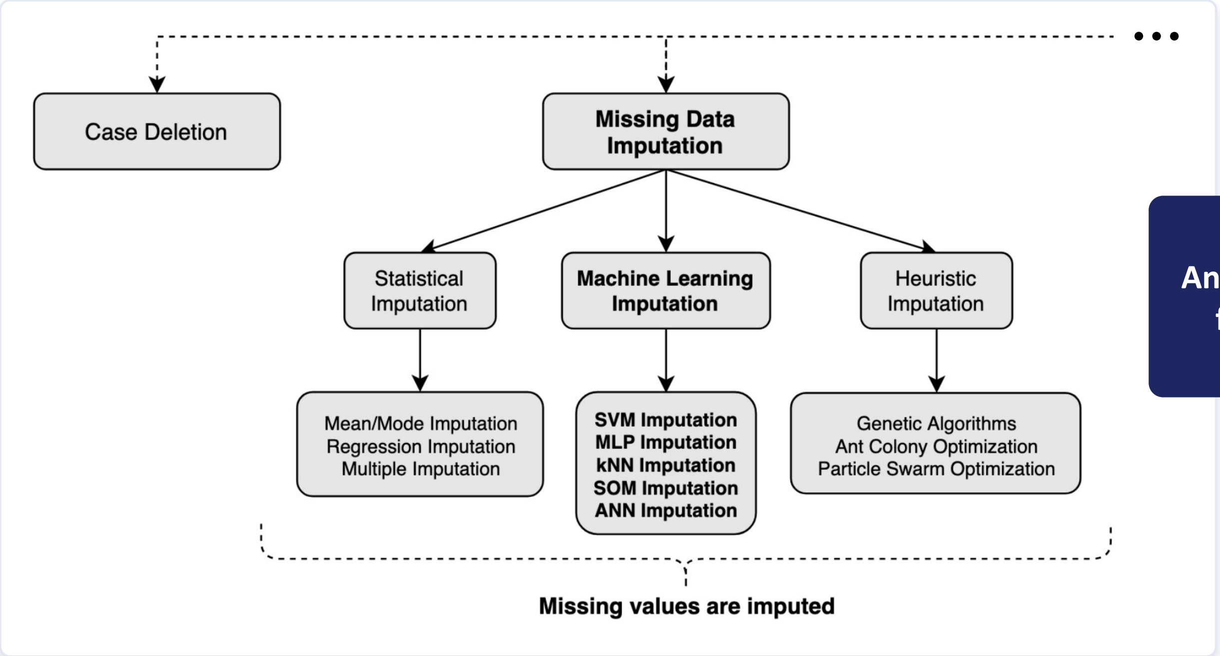
Missing Data Mechanisms

- **Missing Completely At Random (MCAR):** the probability of absence is constant - there is no relationship between the causes of missingness and the observed data
- **Missing At Random (MAR):** the probability of missing data is equal only within specific observed groups
- **Missing Not At Random (MNAR):** the reasons for missingness are unknown, and are linked to unobserved factors

Example: Job Performance Rates

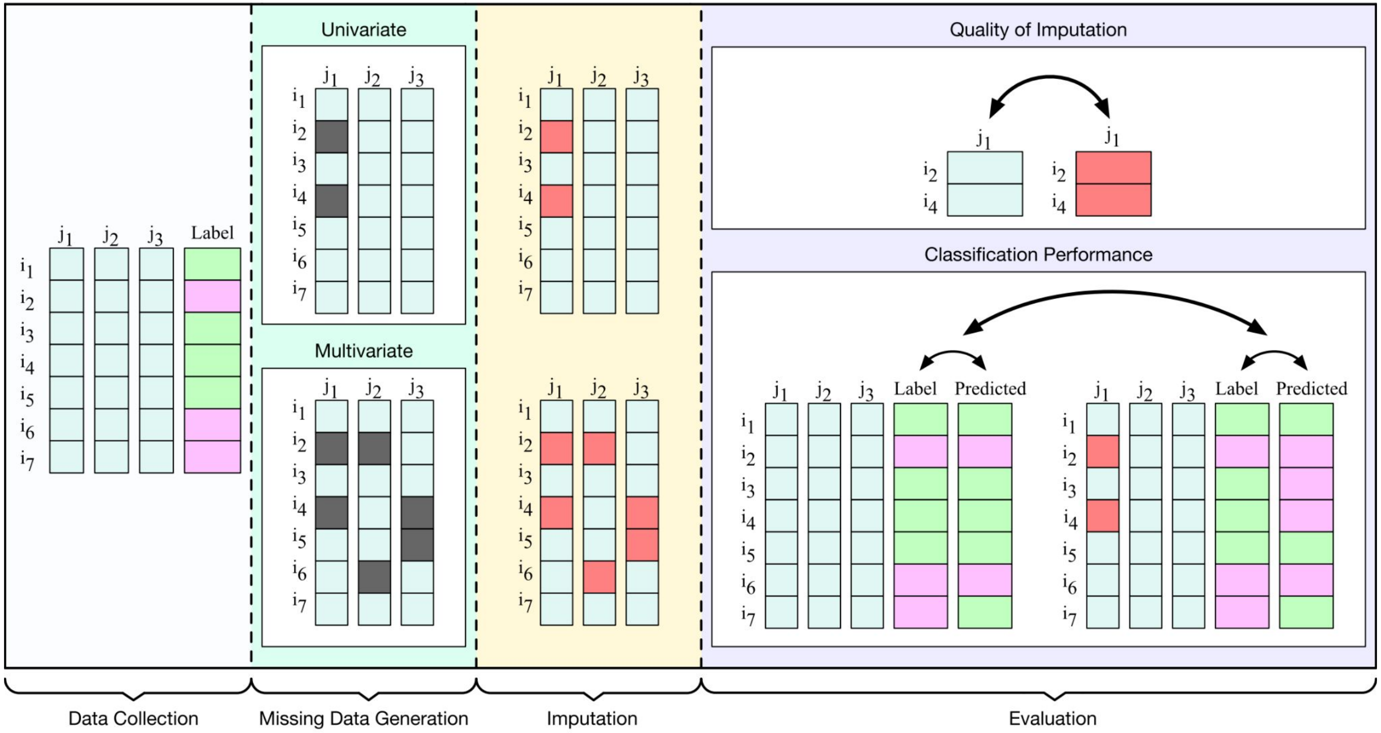
IQ	Job Performance Rates					Age*
	Complete	MCAR	MAR	MNAR1	MNAR2	
78	9	?	?	9	?	66
84	13	13	?	13	13	45
84	10	?	?	10	10	28
85	8	8	?	?	8	22
87	7	7	?	?	7	42
91	7	7	7	?	7	40
92	9	9	9	9	?	65
94	9	9	9	9	9	55
94	11	11	11	11	11	33
96	7	?	7	?	?	60
99	7	7	7	?	7	48
105	10	10	10	10	10	28
105	11	11	11	11	11	33
106	15	15	15	15	15	23
108	10	10	10	10	?	64
112	10	?	10	10	10	25
118	16	16	16	16	16	45
134	12	?	12	12	12	38

Handling Missing Data



Any ML Python package can be used for imputation, e.g., scikit-learn

Missing Data Imputation Pipeline



Missing Data Pipeline Steps

- **Dataset collection:** the dataset used cannot itself contain missing values - its values serve as ground truth
- **Missing data generation (amputation):** missing values are synthetically introduced into one or more features, at varying missing rates, following one or more mechanisms of missingness
    - **How can we do this?**
- **Imputation strategy:** missing data are replaced with plausible values, and the imputation error can then be measured against the original values
- **Evaluation:** split into the imputation error itself, and the resulting classifier performance

FEATURED WORK 1

Generating Synthetic Data: A Review by Missing Mechanism

Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Costa, Jastin Soares, João Santos, Pedro Henriques Abreu

IEEE Access, 7, 11651–11667

2019

For MNAR, generation methods are still relatively limited, and are mainly applicable to numeric features (since missingness here relates a value to itself)

FEATURED WORK 2

Imputation of Data Missing Not at Random: Artificial Generation and Benchmark Analysis

Ricardo Cardoso Pereira, Pedro Henriques Abreu, Pedro Pereira Rodrigues, Mário A. T. Figueiredo

Expert Systems with Applications, 249

2024

Three New Generation Strategies for MNAR

- **Missingness Based on Own Values (MBOV):** missingness probability of a value depends on itself, typically anchored on the lowest or highest value. Two new variants are proposed, one a where parameter *p* controls the percentage of random missingness, and another where values near the centre of the distribution are removed
- **Missingness Based on Unobserved Values (MBUV):** a new feature is created from a normal distribution, and its highest and lowest values are then set as missing in the target observed feature
- **Missingness Based on Own and Unobserved Values (MBOUV):** an overall missingness rate is set and randomly distributed across features, with MBUV applied to nominal features and half of the numeric ones, and MBOV to the rest

FEATURED WORK 3

mdatagen: A Python Library for the Artificial Generation of Missing Data

Arthur Dantas Mangussi, Miriam Seoane Santos, Filipe Loyola Lopes, Ricardo Cardoso Pereira, Ana Carolina Lorena, Pedro Henriques Abreu

Neurocomputing, 625

2025

pypi.org/project/mdatagen

A Python package which implements these new MNAR (and other mechanism) generation strategies

Thank You

Ricardo Cardoso Pereira
rdpereira@dei.uc.pt

Hands-On Session: Missing Data in Tabular Datasets

tinyurl.com/ijcai26-t7-l2