MDMC
Master in Data Management and Curation

AREA SCIENCE PARK

SISSA

Consiglio Nazionale delle Ricerche

Introduction to Statistical Data Analysis and Machine Learning

Matteo Biagetti & Tommaso Rodani

Trieste, 22/10/2024

Outline

- Why a data analysis course at MDMC?
- Objectives of the course
- Topics of the course
- Your data

Why data analysis at MDMC?

Why data analysis at MDMC?

Research Lab

Generated with AI

Why data analysis at MDMC?

FAIRification: data and metadata acquisition procedure compliant with FAIR principles

Research Lab

Generated with AI

Why data analysis at MDMC?

Automatisation: Unsupervised data and metadata handling and pre-processing

Research Lab

Generated with AI

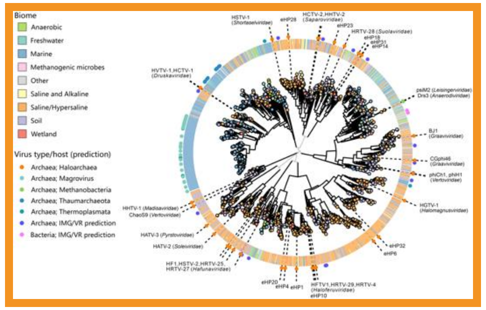
FAIRification: data and metadata acquisition procedure compliant with FAIR principles

Why data analysis at MDMC?

MDMC
Master in Data Management and Curation

Automatisation: Unsupervised data and metadata handling and pre-processing

Data-driven approach: Innovation through interdisciplinary research



WikiMedia Commons

Research Lab

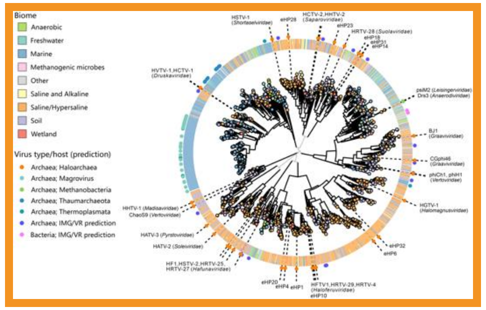
Generated with AI

FAIRification: data and metadata acquisition procedure compliant with FAIR principles

Why data analysis at MDMC?

Automatisation: Unsupervised data and metadata handling and pre-processing

Data-driven approach: Innovation through interdisciplinary research



WikiMedia Commons

FAIR Research Lab

FAIRification: data and metadata acquisition procedure compliant with FAIR principles

Generated with AI

Infrastructure-Augmentation: Data and metadata analysis suggests improvements to the lab itself

Objectives

Suppose you have spent some effort to build a data workflow in your lab so that data and meta data are managed in a proper (FAIR) way.

Now what?

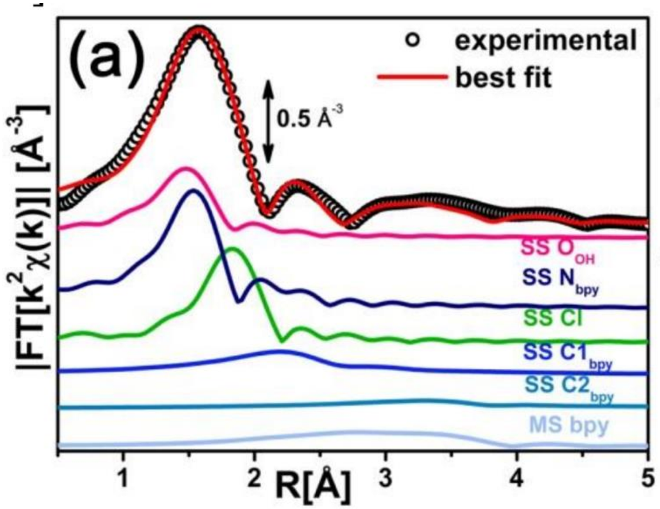
• What type of information do these data and metadata contain?	**Problem Identification**
• Can they be preprocessed to simplify information extraction?	**Problem Investigation**
• What tools can I use to extract information?	**Problem Solution**

Topics

1. REGRESSION

Given an input-output pair, I want to learn the relation between the two

- To predict a new output given a new input
- To learn the physical, biological, chemical, etc, law between input and output
- To generate a new pair
- …



Braglia et al. 2017

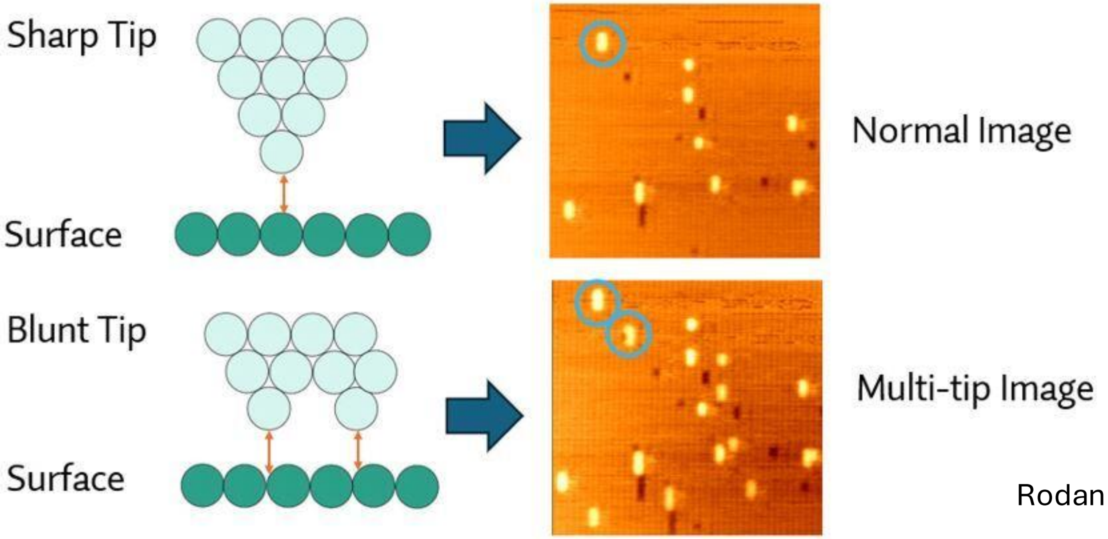
AlphaFold (DeepMind)
WikiMedia Commons

Topics

2. CLASSIFICATION

Given set of data with given properties, I want to classify them based on these properties

- To single out specific properties
- To filter data for purification
- …



Rodani et al 2024

Topics

3. UNSUPERVISED LEARNING

Given set of data, I want to discover its properties

- To discover relations among quantities (new physical, biological, chemical laws)
- Reduce dimension of dataset
- …

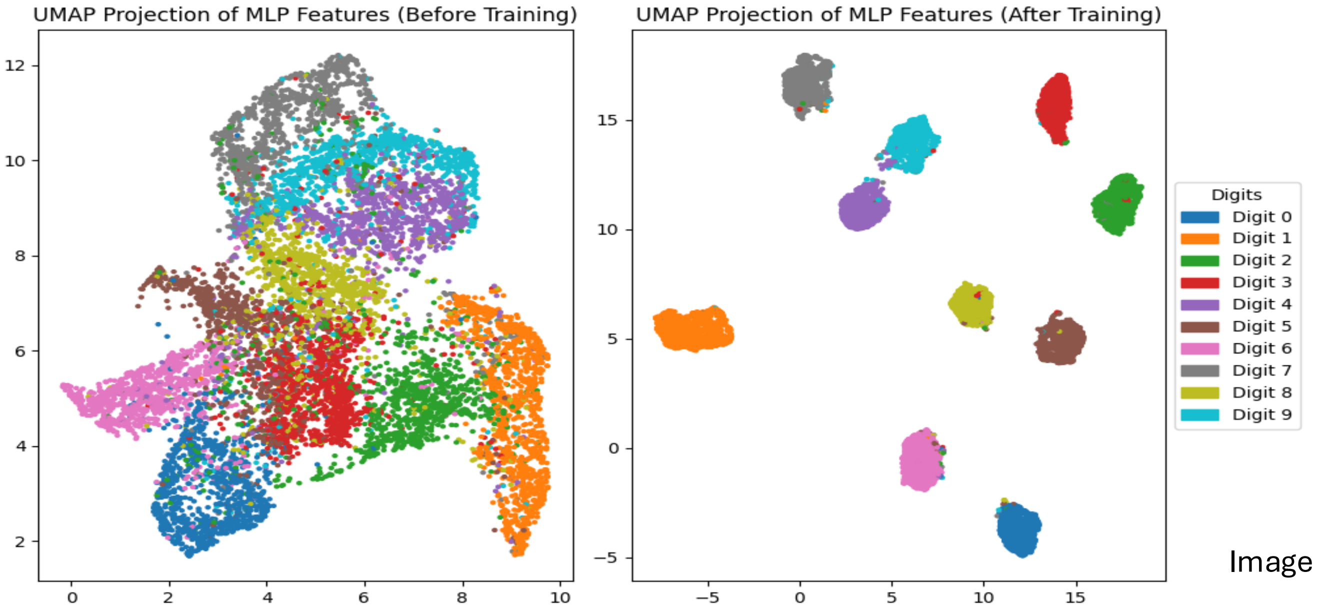


Image from T.Rodani

Your Data!

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

nffa-di

Nano Foundries Fine Analysis
Digital Infrastructure

prp

Pathogen Readiness Platform
for CERIC-ERIC Upgrade

This Pilot training activity has been funded by the European Union – NextGenerationEU within the project PNRR "PRP@CERIC" IR0000028 and «NFFA-DI" IR0000015 - Missione 4, “Istruzione e Ricerca” – Componente 2, “Dalla ricerca all'impresa” – Linea di investimento 3.1, “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione” – Azione 3.1.1, “Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe
e costituzione di reti”.