

Data Manipulation with PYTHON

marco milanesio
MScDSAI 2021-2022

Who am I?

- PhD [2010]
 - Distributed systems
 - Network measurements and performances
 - Distributed storage
- @Inria
 - Optimization
 - Image processing
 - Spark
- @MSI - MDLab
 - Dev-ops / Databases
 - Implementation - optimization
 - Virtualization
 - CNN + image analysis

Inria
inventeurs du monde numérique

Epione
e-patient / e-medicine

UNIVERSITÉ CÔTE D'AZUR 
CENTER OF MODELING
SIMULATION AND
INTERACTIONS

UNIVERSITÉ CÔTE D'AZUR 
MEDICAL DATA
LABORATORY

Who are you?

- Curious, open minded
 - You know how to use a PC 🧐
 - Wanting to learn some cool stuff
 - Not feared of tackling problems
 - Not feared by errors
 - (optional) Some coding experience
 - (bonus) Some Python experience
 - (bonus) “LMGTFY” skills 🧐
-
- Ideal profile:
 - 50% data scientist: exploit the data
 - 50% software developer: you need to code

Motivation

A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance

Ga Y



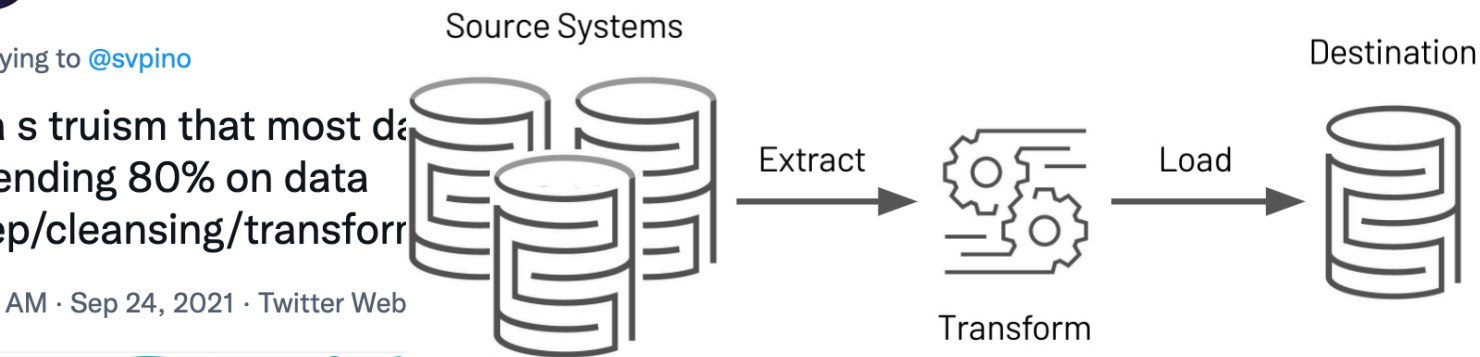
Wilson Chua
@wilsonchua

Replying to @svpino

It's a truism that most data scientists are spending 80% on data prep/cleansing/transform

3:19 AM · Sep 24, 2021 · Twitter Web

ETL Process



Data Validation



Data Modelling

trends

- M — Modeling our data will give us our predictive power as a wizard
- N — Interpreting our data

Data Exploration & Analysis



find patterns and

Course Overview

- Introduction
- Recap on Python3 (3.7 to 3.9.7)
- Data Manipulation
 - cleaning
 - preprocessing
- Basic data analysis
 - builtins
- Advanced data analysis
 - numpy
 - pandas
 - scikit-learn

Course Overview

- Syntax
- Data structures, types
- Software development
 - Testing
- Data preprocessing
 - management
 - cleaning
- Principles of functional programming
- A glimpse on ML

Course Overview

- REPL + scripts + notebooks
- 10 lessons
 - ~ 30% lectures
 - ~ 70% lab sessions
- Evaluation (to be defined)

Contacts

marcomilanesio.github.io/teaching.html

marco.milanesio@univ-cotedazur.fr