

INTRODUÇÃO À PANDAS

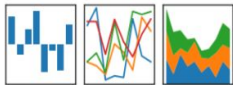
Prof. Dr. Julio Cesar dos Reis

Instituto de Computação - Universidade Estadual Campinas

O que é pandas?

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

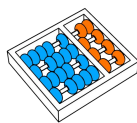


Data



PANDAS

— — —

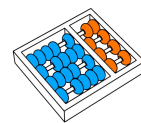


<https://pandas.pydata.org/>

pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

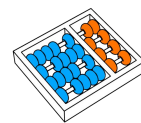
Install pandas now!



PANDAS

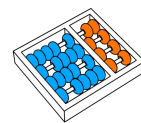
— — —

- Biblioteca open source (BSD-licensed), a qual fornece estruturas de **alta-performance** e fácil de usar
- Fornece ferramentas para análise e processamento de dados
- Fornece uma API eficiente para análise de dados em python
- Ferramenta poderosa e flexível para análise de dados



PANDAS FEATURES

- Acessível para todos
- Free para uso
- Permite modificações (BSD-licensed)
- Ponderosa
- Fácil de usar
- Eficiente (rápida)

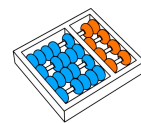


ESTRUTURAS DE DADOS DO PANDAS

— — —

Existem duas estruturas de dados principais no pandas

- Series
- DataFrames



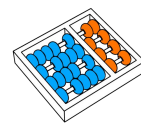
PANDAS SERIES

- **Series** é um array de uma dimensão **rotulado** capaz de armazenar qualquer tipo de dados (integer, string, float, objetos, etc)
- Os labels podem ser indexados
- Podemos criar **series** baseado em:
 - Listas
 - Dicionários
 - ndarrays

		Data
Index	p	1.0
	q	2.0
	r	2.0
	s	NaN

dtvpe: float64

Series



PANDAS SERIES vs NDARRAYs

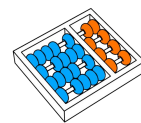
- A principal diferença entre ndarrays e series é que nas séries os índices podem ser **rotulados**
- *Ndarrays* só permite a indexação por números inteiros
- Caso desejamos acessar o elemento 2.2 da sequência abaixo:
 - Usando um **ndarray** podemos acessar usando com Seq[1]
 - Usando uma **series** podemos acessar usando com Seq['b']

	0	1	2	3
Seq =	1.1	2.2	3.3	3.4

ndarray

	'a'	'b'	'c'	'd'
Seq =	1.1	2.2	3.3	3.4

series



PANDAS DATAFRAMES

- Estrutura de dados no formato de tabela com rótulos nas linhas e nas colunas
- Podem ser indexados tanto por linha quanto por coluna

Diagram illustrating the structure of a Pandas DataFrame, showing a table with 5 columns and 5 rows. The columns are labeled 0 to 4, and the rows are labeled 0 to 4.

Column Label/ Header: 0 1 2 3 4

Index Label: 0 1 2 3 4

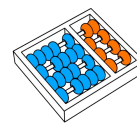
Column Index: Name Age Marks Grade Hobby

Row Index: S1 S2 S3 S4 S5

Column: Marks

Element/ Value/ Entry: 88.78

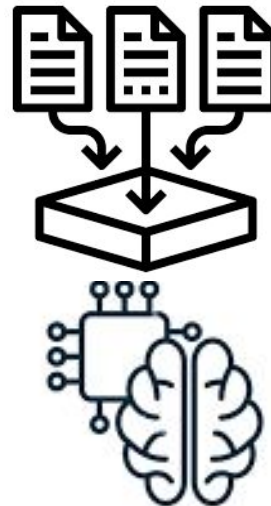
	0	1	2	3	4
	Name	Age	Marks	Grade	Hobby
0	S1	Joe	20	85.10	A Swimming
1	S2	Nat	21	77.80	B Reading
2	S3	Harry	19	91.54	A Music
3	S4	Sam	20	88.78	A Painting
4	S5	Monica	22	60.55	B Dancing



POR QUE USAR PANDAS?

— — —

- Simples de usar
- Integrado com diversas outras ferramentas de ciência de dados
- Ajuda a preparar os dados para usar no aprendizado de máquina

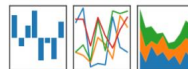


Prática

- Funções úteis
- Tipos de dados
- Importar e exportar dados
- Análise inicial de dados
- Visualizar e selecionar dados
- Manipular dados

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Data

