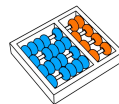


INTRODUÇÃO À PANDAS

Capacitação Profissional em Tecnologias de Inteligência Artificial

Allan M. de Souza

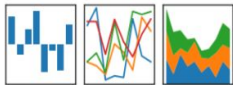
Instituto de Computação - Universidade Estadual Campinas



O que é pandas?

pandas

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

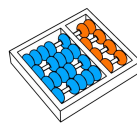


Data



PANDAS

— — —

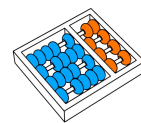


<https://pandas.pydata.org/>

pandas

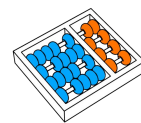
pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

Install pandas now!



PANDAS

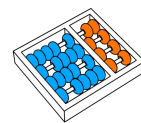
- É uma biblioteca open source (BSD-licensed), a qual fornece estruturas de **alta-performance** e fácil de usar
- Além disso, pandas também fornece ferramentas para análise e processamento de dados
- Fornece uma API eficiente para análise de dados em python
- É uma ferramenta poderosa e flexível para análise de dados



PANDAS FEATURES

— — —

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

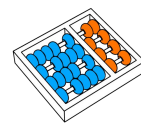


ESTRUTURAS DE DADOS DO PANDAS

— — —

Existem duas estruturas de dados principais no pandas

- Series
- DataFrames



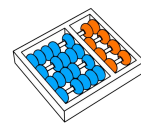
PANDAS SERIES

- **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 rolados
- Ndarrays só permite a indexação por números inteiros
- Exemplo, caso desejamos acessar o elemento 2.2 da sequência abaixo:
 - Usando um **ndarray** podemos acessar usando com Seq[1]
 - Usando uma **serie** podemos acessar usando com Seq['b']

Seq =

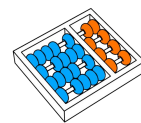
0	1	2	3
1.1	2.2	3.3	3.4

ndarray

Seq =

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

series



PANDAS DATAFRAMES

Dataframes é uma estrutura de dados no formato de tabela com rótulos nas linhas e nas colunas

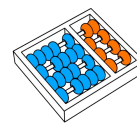
Dataframes podem ser indexados tanto por linha quanto por coluna

Column Label/ Header		0	1	2	3	4
Index Label		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

Diagram illustrating a Pandas DataFrame structure with labels for indices and values.

Labels and Annotations:

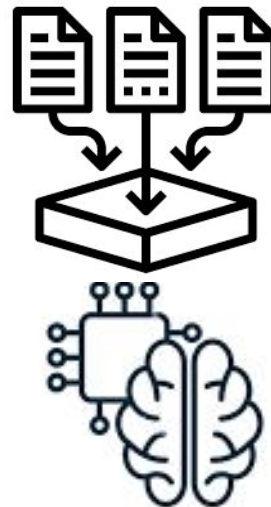
- Column Index:** Points to the header row (0-4).
- Row Index:** Points to the first column (S1-S5).
- Column:** Points to the 'Marks' column (index 2).
- Row:** Points to the 'S4' row (index 3).
- Element/ Value/ Entry:** Points to the value '88.78' at the intersection of Row 3 and Column 2.



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



INTRODUÇÃO À PANDAS

Aula prática

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

pandas

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

