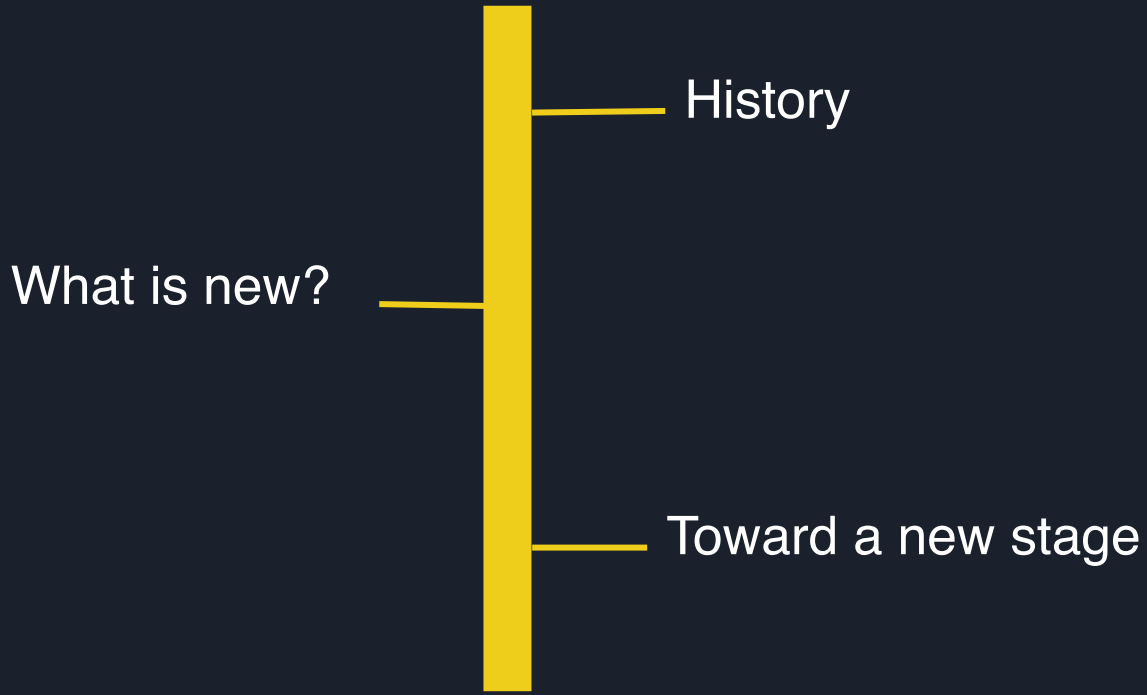
Science and data manipulation in Pharo

Polymath, Pharo-AI and DataFrame

Ferlicot-Delbecque Cyril | ESUG 2023

cyril@ferlicot.fr

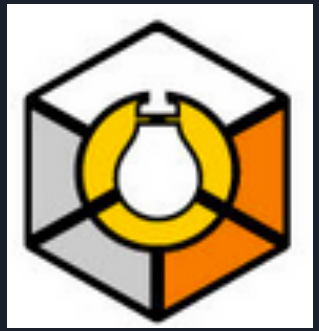
Summary



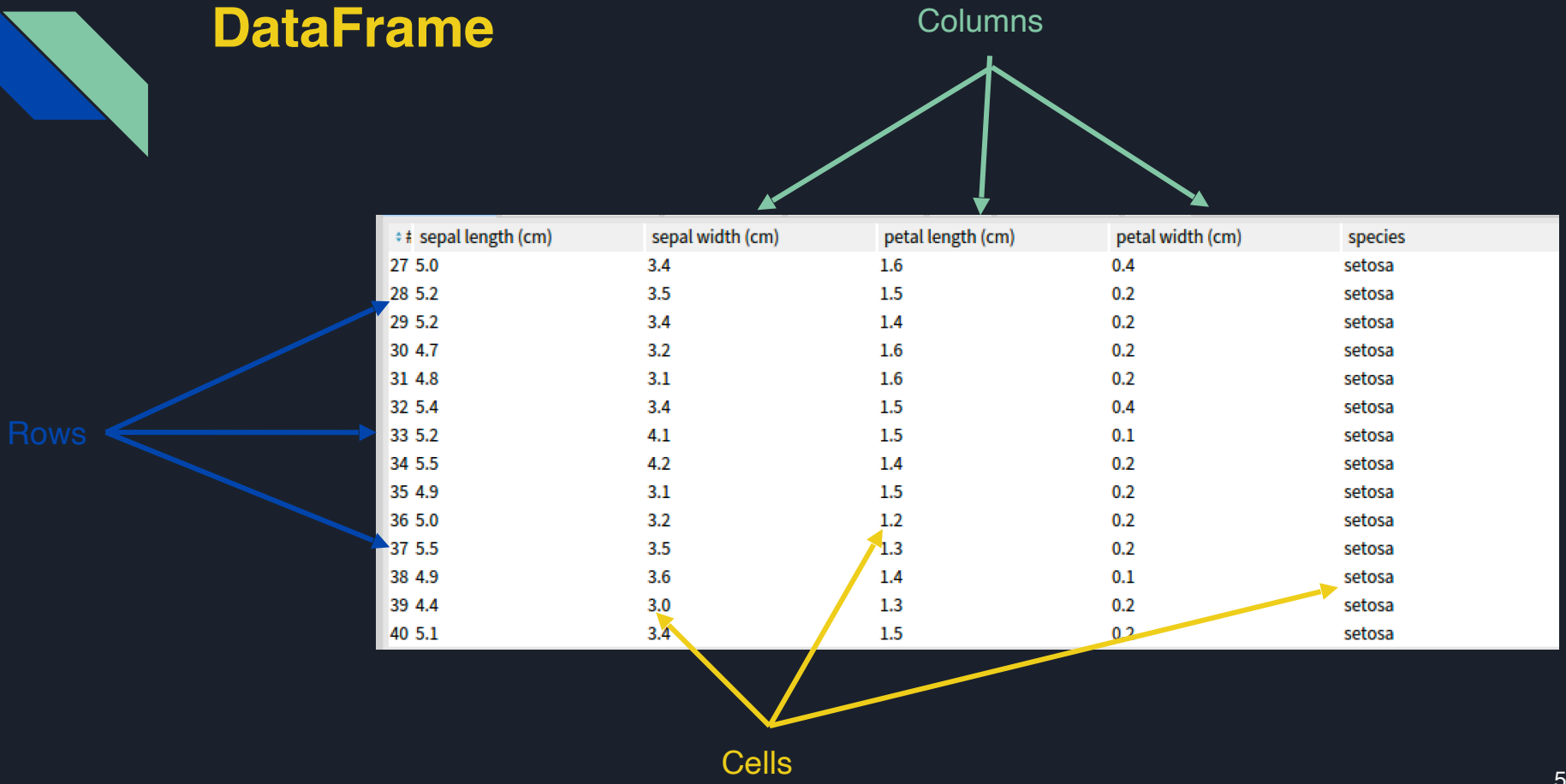
History

PolyMath

- Computation library for Pharo
- Similar to NumPy and SciPy in Python or SciRuby in Ruby
- Originally SciSmalltalk in Squeak
- Present in Pharo since a long time



DataFrame



#	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa

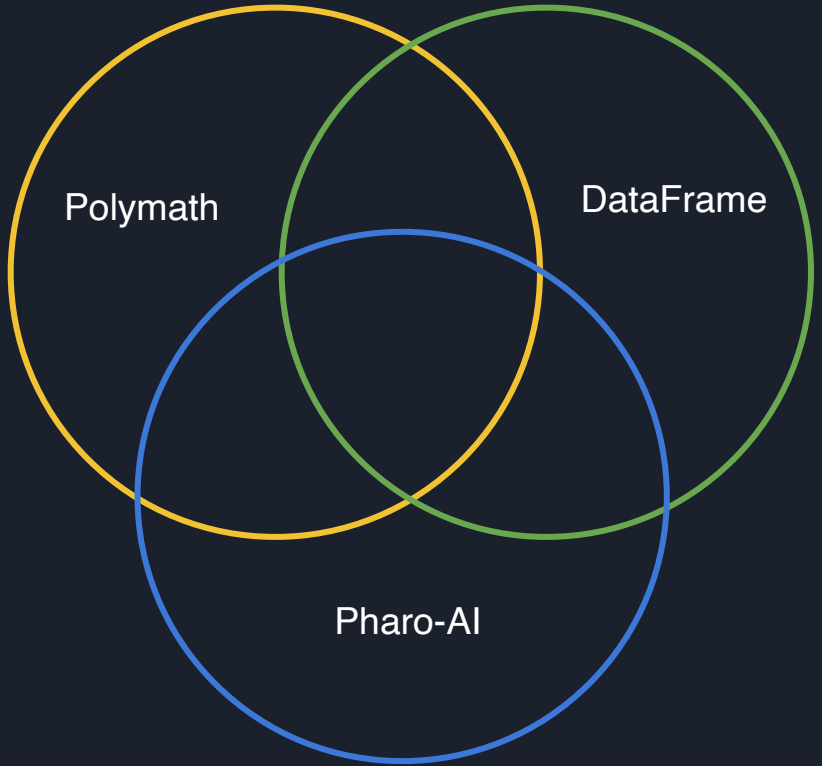
DataFrame

- Table data structure
- Similar to DataFrame in Pandas, Julia, …
- Heavily used in data science
- Created in 2017 during GSoC

Pharo-AI

- Created in 2020
- Implements classical machine learning algorithms (not deep learning)
  - K-Means, Linear Regression, N-Gram Model, …

Harmony of communities



What is new?

Polymath: Modularization

- Rearchitecture
  - Extraction of data structures and random generators
  - Extraction of distributions in progress
- Cleaning of internal dependencies

Polymath

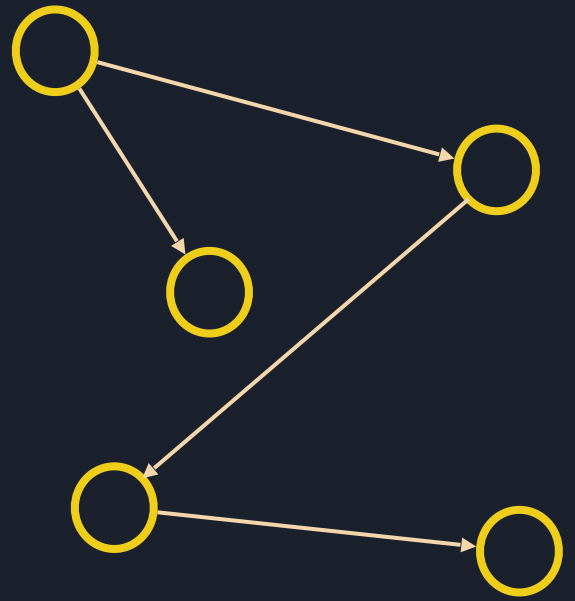
- Improvement of the CI robustness
- Align some conventions with Pharo-AI
- Divers cleanings and bug fixes
- Pharo 11 compatibility

Pharo-AI : Data manipulation

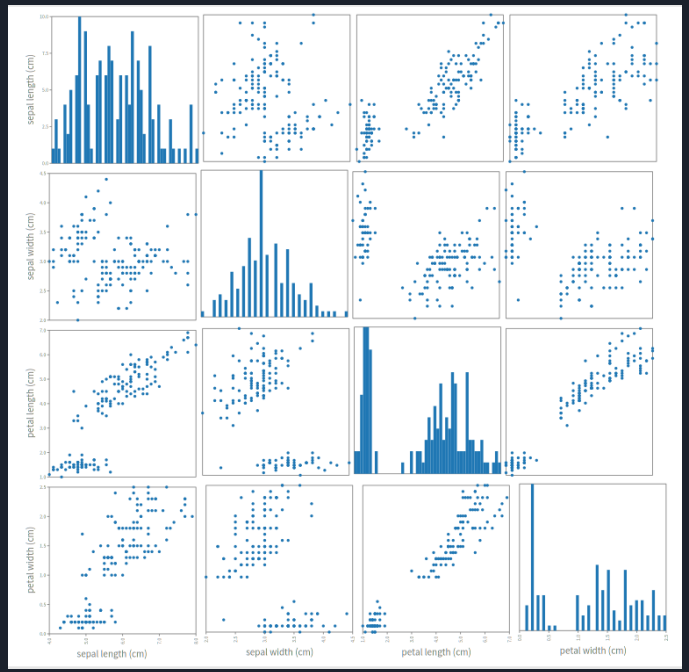
- Data partitioners : create tests sets
- Imputers : fill missing values
- Encoders : Standardize your datas
- Normalizer : Use common scales in your project

Pharo-AI

- Uniformization of projects
- Documentation
- Graph algorithms updates
- Divers speed up
- Cleaning and bug fixes in algos



DataFrame



- Speed up
- Better integration with other collections
- New visualizations based on DataFrames
- Integration with pharo-AI data preprocessing

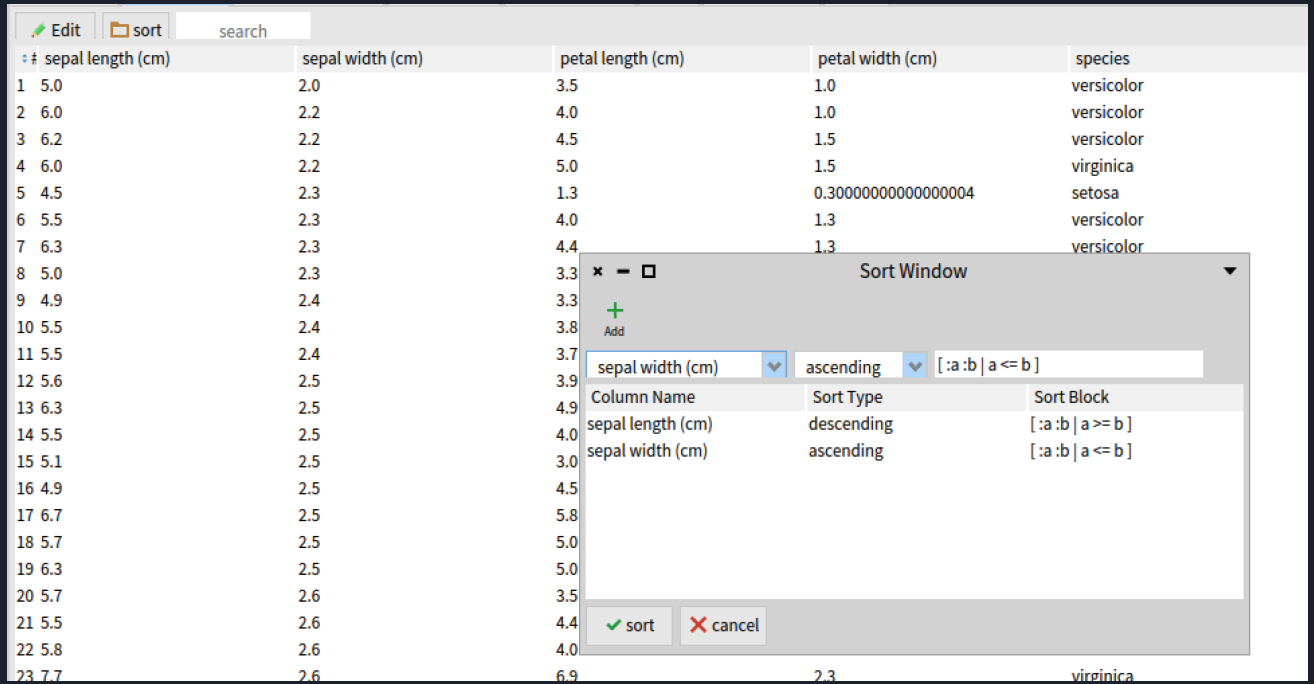
DataFrame : GSoC 2023

- GSoC of Joshua Jose Dias Barreto
- Implementation of missing features
  - Better sorting
  - Data manipulation
  - Missing values management
  - …

Google
Summer of Code

DataFrame : GSoC 2023

Further improvements of DataFrame inspector



#	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
1	5.0	2.0	3.5	1.0	versicolor
2	6.0	2.2	4.0	1.0	versicolor
3	6.2	2.2	4.5	1.5	versicolor
4	6.0	2.2	5.0	1.5	virginica
5	4.5	2.3	1.3	0.30000000000000004	setosa
6	5.5	2.3	4.0	1.3	versicolor
7	6.3	2.3	4.4	1.3	versicolor
8	5.0	2.3	3.3		
9	4.9	2.4	3.3		
10	5.5	2.4	3.8		
11	5.5	2.4	3.7		
12	5.6	2.5	3.9		
13	6.3	2.5	4.9		
14	5.5	2.5	4.0		
15	5.1	2.5	3.0		
16	4.9	2.5	4.5		
17	6.7	2.5	5.8		
18	5.7	2.5	5.0		
19	6.3	2.5	5.0		
20	5.7	2.6	3.5		
21	5.5	2.6	4.4		
22	5.8	2.6	4.0		
23	7.7	2.6	6.9	2.3	virginica

Toward a new stage

A push from the students

- DataFrame was started as a GSoC
- First AI algorithm were students projects

=> Data science interest more and more people


Agile Artificial Intelligence
Alexandre Bergel
RelationalAI, Switzerland
https://bergel.eu

We are answering to this call

- Engineers are now pushing those projects
- Projects are maintained
- Speed is becoming correct

Are you using scientific computing or data science?

Is the speed enough for you?

Are you encountering any problem?

Are you missing features for you projects?

Let us know ;)