Pharo DataFrame:
Past, Present, and Future

Larisa Safina, Oleksandr Zaitsev, Cyril Ferlicot-Delbecque,
Papa Ibrahima Sow

International Workshop on Smalltalk Technologies 2023, Lyon, France

Agenda

What is Data Frames and what they are good for

Evolution of Data Frames in Pharo

Data Frames outside of Pharo

Data Frames

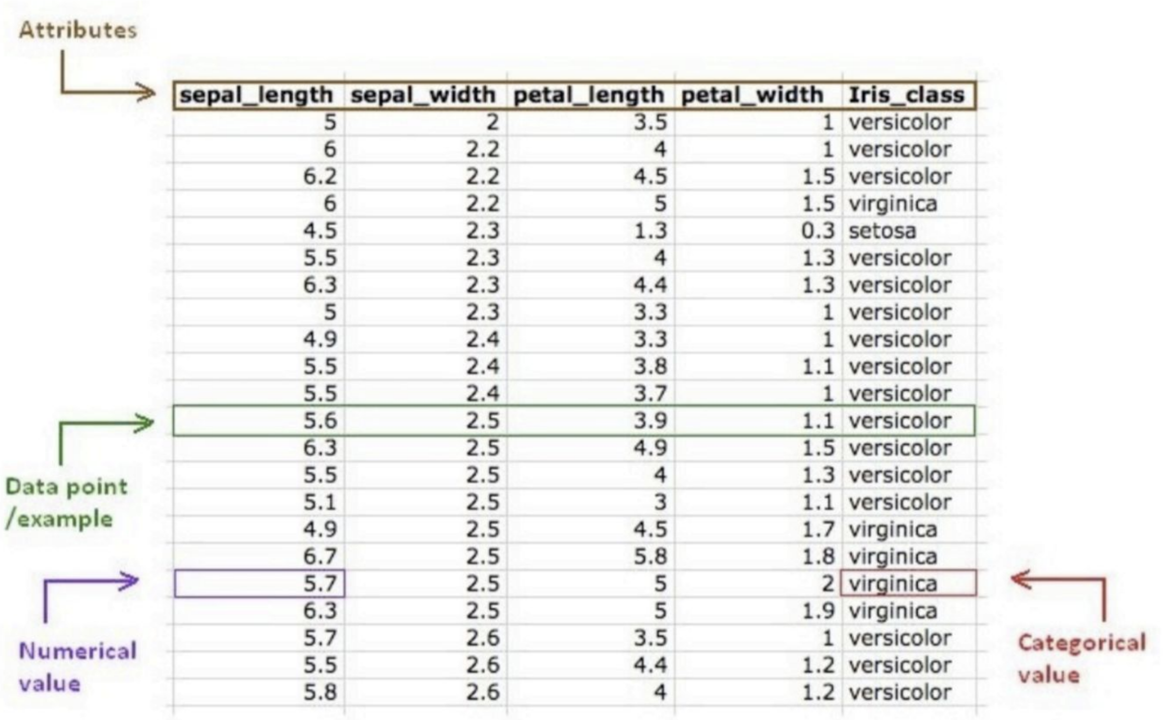
Tabular structure

Columns with named headers

Columns/Rows index

Mixed data types in rows

Primitive data types


Attributes
Data point /example
Numerical value
Categorical value

sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

```
irisDataFrame := DataFrame readFromCsv: 'iris.csv'.
irisDataFrame inspect
```

*The table is courtesy of Zaitsev et al (ESUG22)

Data Frames API

Data Import/Export

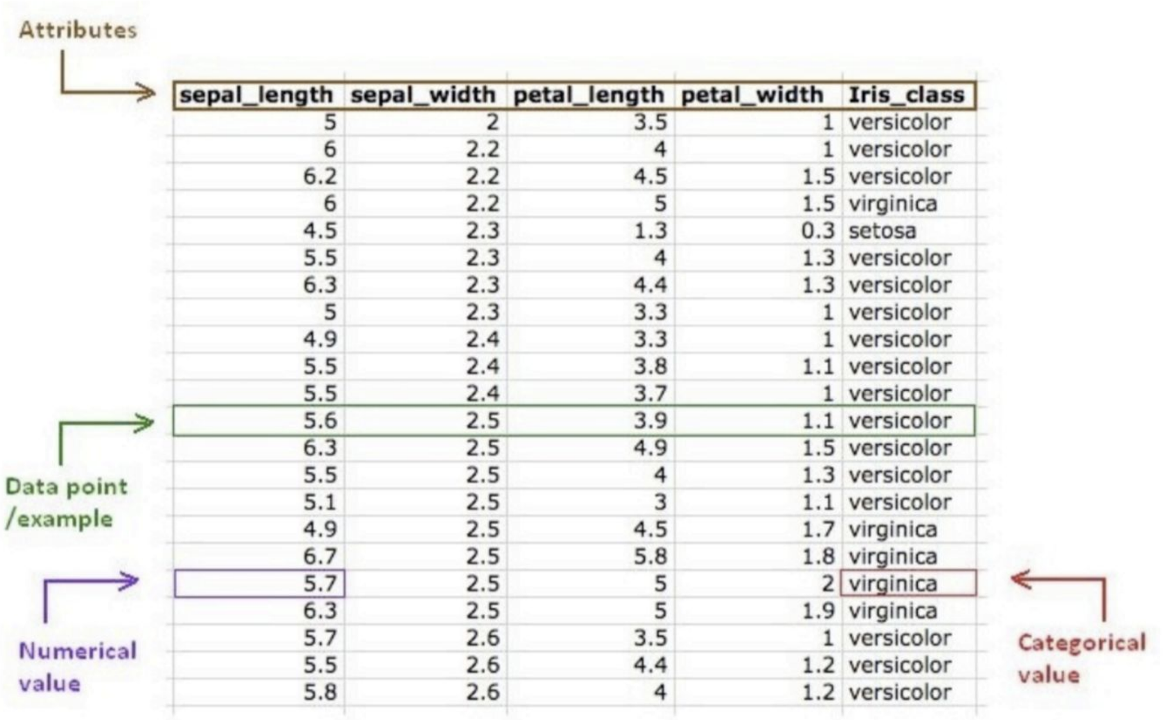
Grouping and Aggregation

Missing Data Handling

Statistical Operations

Time Series Analysis

Visualization


Attributes
Data point /example
Numerical value
Categorical value

sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

```
irisDataFrame := DataFrame readFromCsv: 'iris.csv'.
irisDataFrame inspect
```

*The table is courtesy of Zaitsev et al (ESUG22)

Pharo DataFrame. Past

Started as a GSoC 2017 project (by Oleks)

Google Summer of Code

Focus of the two followed-up GSoC

Contributors 12

Contributions from external developers

Pharo DataFrame. **Past**

Problems:

- No stable maintenance
- Lack of functionality
- Low performance
- Incomplete coherence with Pharo collections
- Lack of detailed documentation

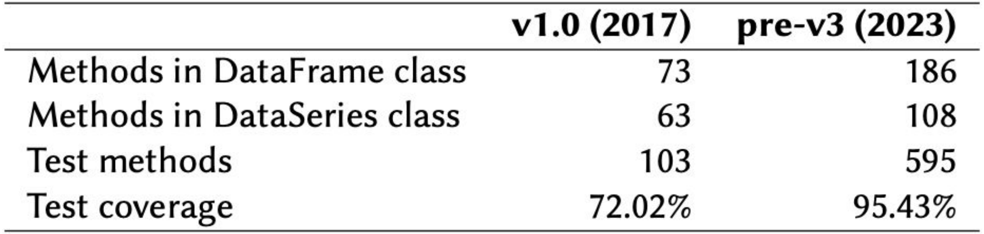
Pharo DataFrame. Present

Stable project with a permanent developer (Cyril)

Code optimization

- code quality
- speed and volume improvements

Adding new functionality

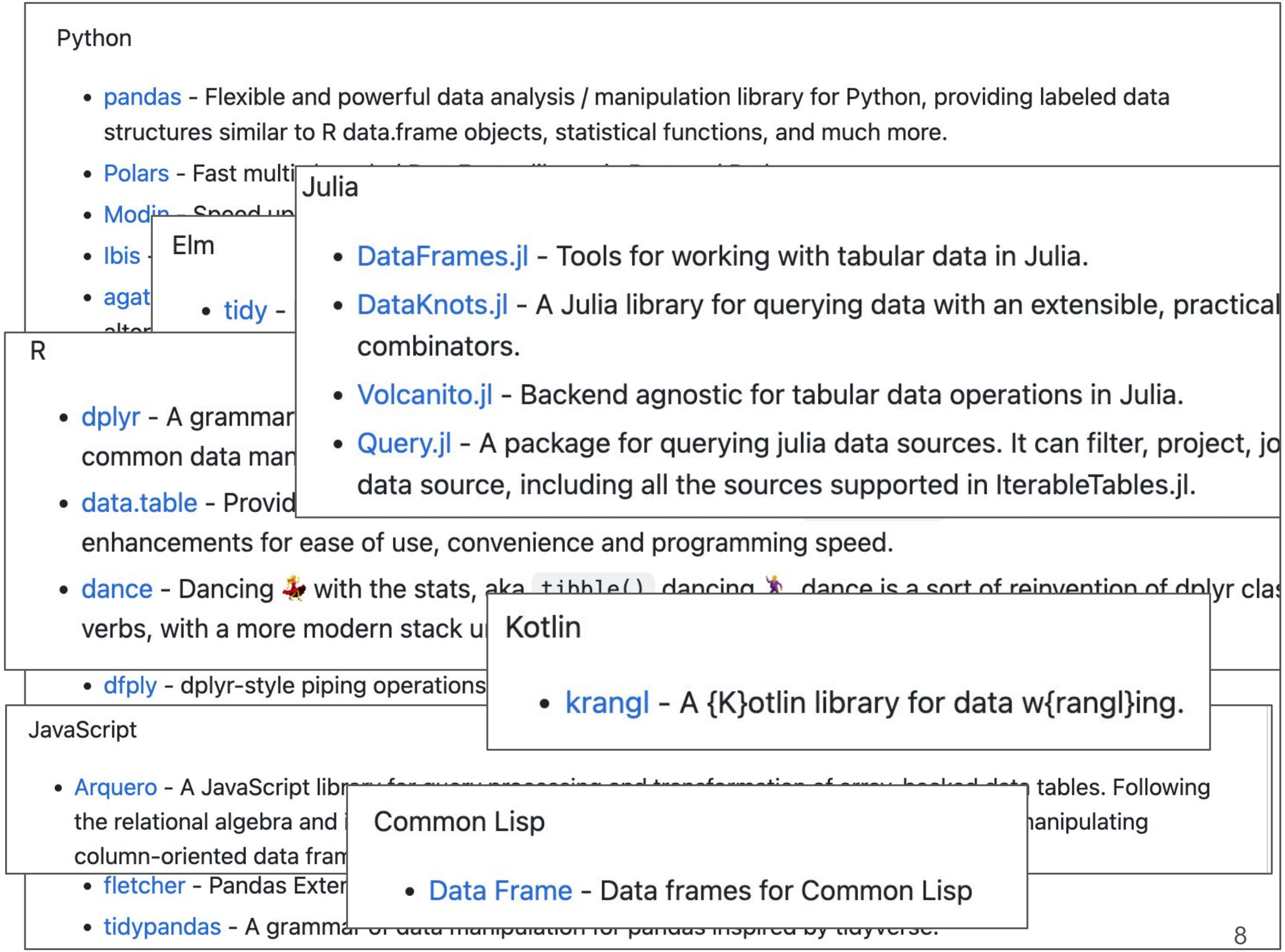


	v1.0 (2017)	pre-v3 (2023)
Methods in DataFrame class	73	186
Methods in DataSeries class	63	108
Test methods	103	595
Test coverage	72.02%	95.43%

Awesome Data Frames

Kyle Mitchell (jcmkk3) and
Uwe L. Korn (xhochy)

https://github.com/jcmkk3/awesome-dataframes



Python

- pandas - Flexible and powerful data analysis / manipulation library for Python, providing labeled data structures similar to R data.frame objects, statistical functions, and much more.
- Polars - Fast multi
- Modin - Speed up
- Ibis
- agat
  alter
- dfply - dplyr-style piping operations
- fletcher - Pandas Exter
- tidypandas - A gramma r of data manipulation for pandas inspired by tidyverse.

Elm

- tidy -

R

- dplyr - A grammar
  common data man
- data.table - Provid
  enhancements for ease of use, convenience and programming speed.
- dance - Dancing 💃 with the stats, aka `tibble()` dancing 🕺. dance is a sort of reinvention of dplyr clas
  verbs, with a more modern stack u

Julia

- DataFrames.jl - Tools for working with tabular data in Julia.
- DataKnots.jl - A Julia library for querying data with an extensible, practical combinators.
- Volcanito.jl - Backend agnostic for tabular data operations in Julia.
- Query.jl - A package for querying julia data sources. It can filter, project, jo data source, including all the sources supported in IterableTables.jl.

Kotlin

- krangl - A {K}otlin library for data w{rangl}ing.

JavaScript

- Arquero - A JavaScript libr … tables. Following the relational algebra and … anipulating column-oriented data fram

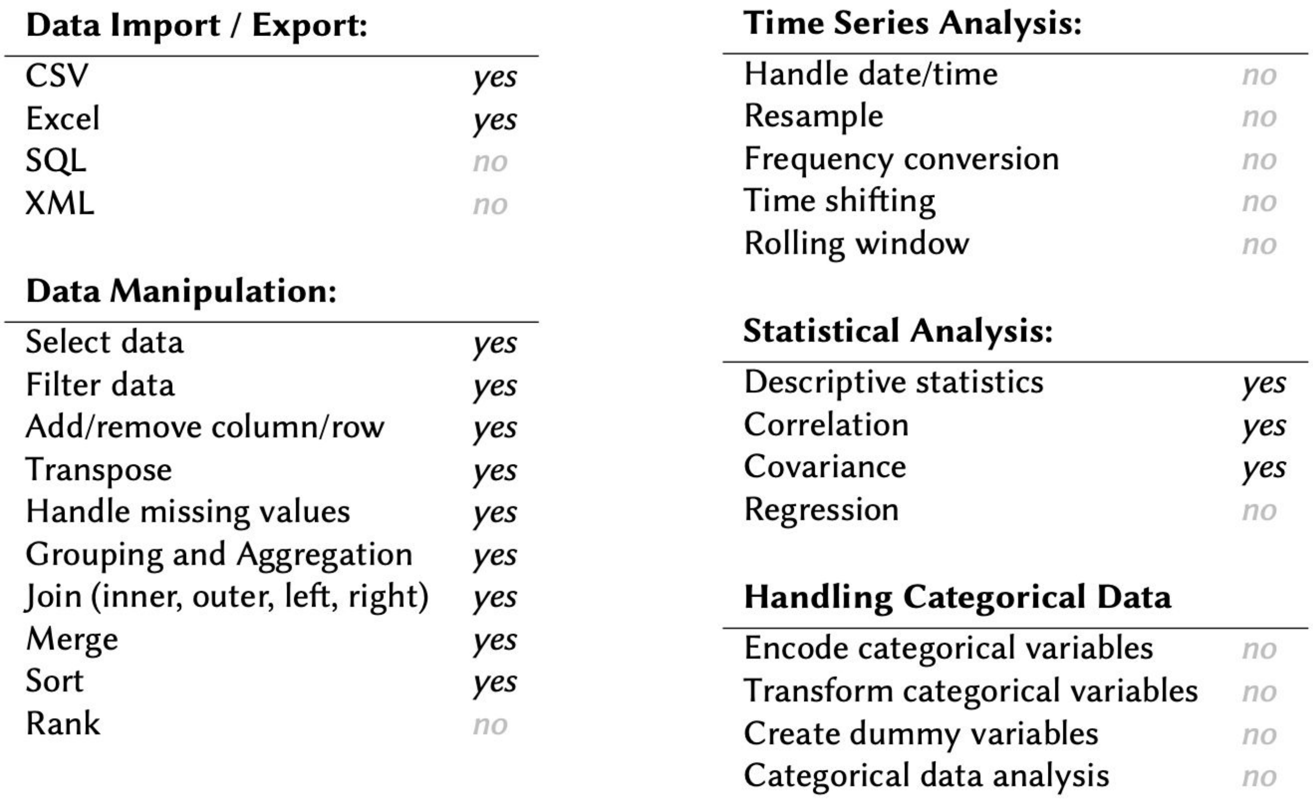
Common Lisp

- Data Frame - Data frames for Common Lisp

Pharo DataFrame

vs

Pandas



Data Import / Export:	
CSV	*yes*
Excel	*yes*
SQL	*no*
XML	*no*

Data Manipulation:	
Select data	*yes*
Filter data	*yes*
Add/remove column/row	*yes*
Transpose	*yes*
Handle missing values	*yes*
Grouping and Aggregation	*yes*
Join (inner, outer, left, right)	*yes*
Merge	*yes*
Sort	*yes*
Rank	*no*

Time Series Analysis:	
Handle date/time	*no*
Resample	*no*
Frequency conversion	*no*
Time shifting	*no*
Rolling window	*no*

Statistical Analysis:	
Descriptive statistics	*yes*
Correlation	*yes*
Covariance	*yes*
Regression	*no*

Handling Categorical Data	
Encode categorical variables	*no*
Transform categorical variables	*no*
Create dummy variables	*no*
Categorical data analysis	*no*

Pharo DataFrame. Future

✨ Better performance ✨

Functionality Enhancements

Better synchronisation with PolyMath and pharo-ai

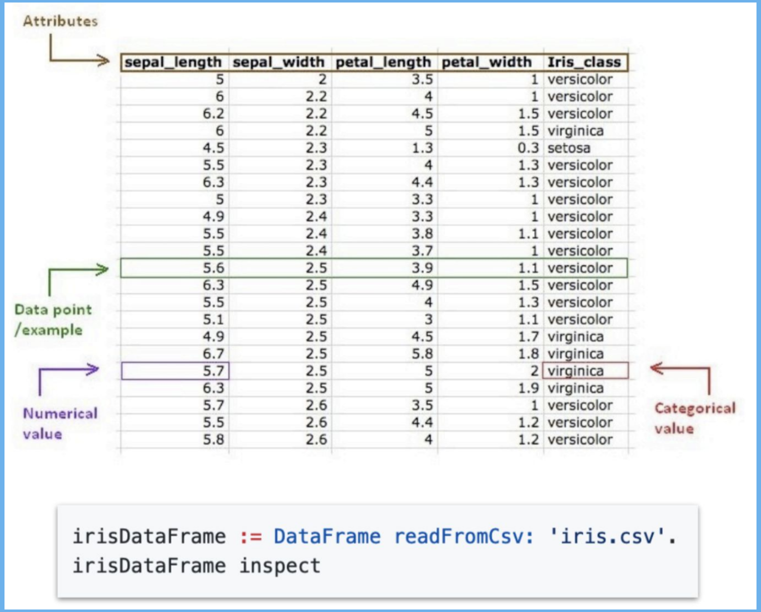
Big Data Support

Evaluation ❤️

Toy Story 1995

Do you want to contribute?

https://github.com/PolyMathOrg/DataFrame



Attributes
Data point /example
Numerical value
Categorical value

sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

```
irisDataFrame := DataFrame readFromCsv: 'iris.csv'.
irisDataFrame inspect
```

Better performance

Functionality Enhancements

Evaluation

Student project → Mature project with engineers