Journals, languages and their reproducibility problems

Bartłomiej Eljasiak, Konrad Komisarczyk, Mariusz Słapek

Warsaw University of Technology

Agenda

1. How do we define reproducibility?
2. Our topic
3. Why this topic?
4. What do we want to test?
5. Differences in examples
6. How do we score articles?
7. Which journals and why?
8. How do we chose articles?
9. What is to be done?

Reproducibility definition

“Reproducibility of a method/test can be defined as the closeness of the agreement between independent results obtained with the same method on the identical subject(s) (or object, or test material) but under different conditions (different observers, laboratories etc.)”

- Slezak and Waczulikova, 2011

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Goodman, Steven N. and Fanelli, Daniele and Ioannidis, John P. A. What does research reproducibility mean? 2016. Science Translational Medicine 8, 341. URL https://stm.sciencemag.org/content/8/341/341ps12

Our topic

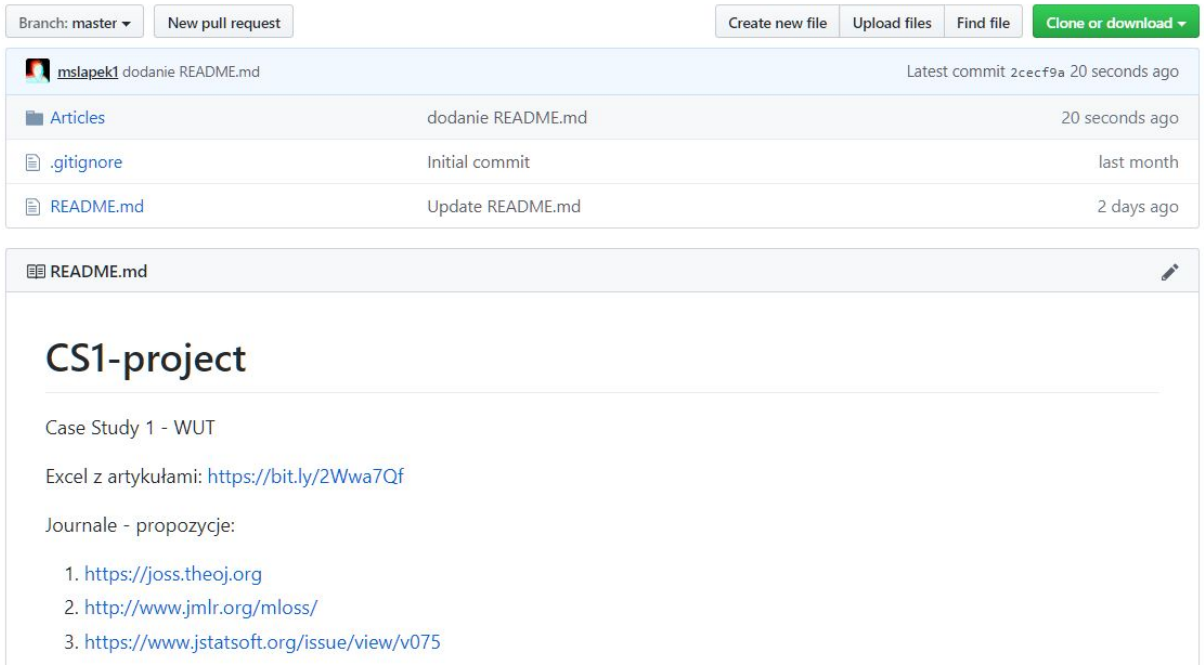
Reproducibility differences of articles published in various journals and using R or Python language

Why this topic?

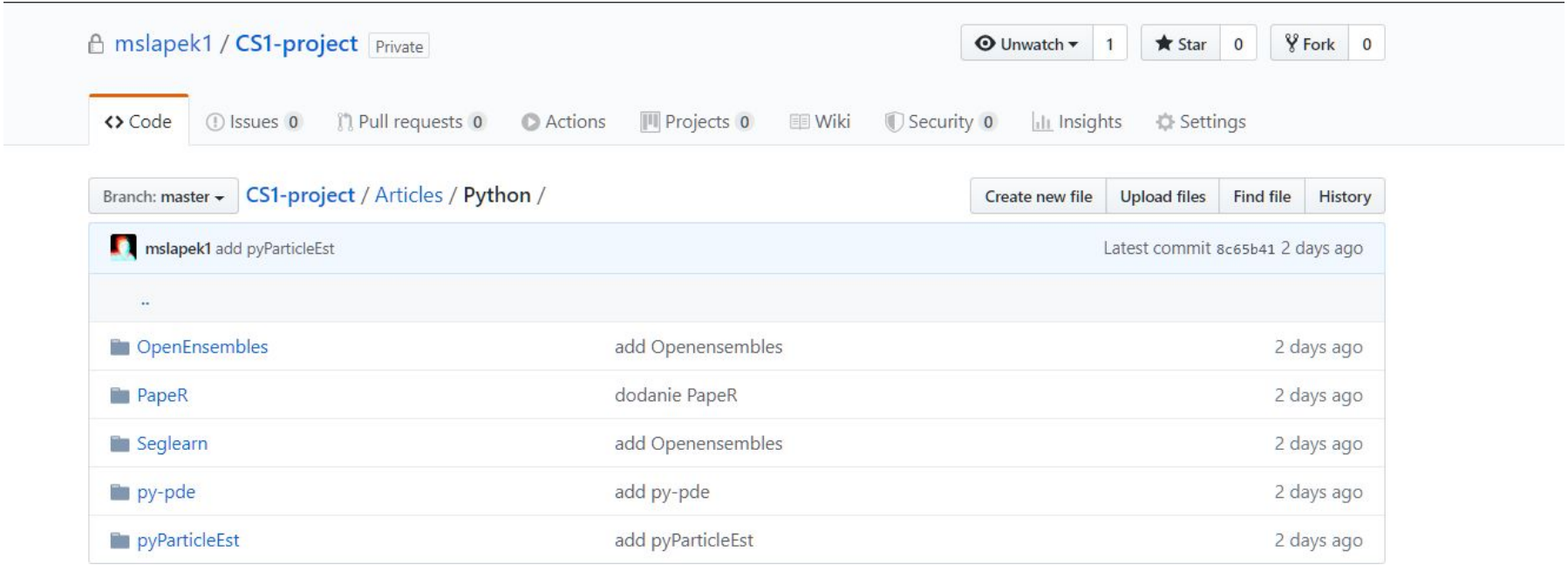
1. Compare two of the most popular programming languages in data science publications.
2. No one has done this before.
3. Different journals also compete between each other.

Journals notice the importance of this subject (McNutt 2014). Also according to scientists journals should take some responsibility for this subject (Eisner 2018).

Our github


Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
mslapek1 dodanie README.md
Latest commit 2cecf9a 20 seconds ago
Articles
dodanie README.md
20 seconds ago
.gitignore
Initial commit
last month
README.md
Update README.md
2 days ago
README.md
CS1-project
Case Study 1 - WUT
Excel z artykułami: https://bit.ly/2Wwa7Qf
Journale - propozycje:
1. https://joss.theoj.org
2. http://www.jmlr.org/mloss/
3. https://www.jstatsoft.org/issue/view/v075

Project structure



Repo example

OpenEnsembles

What do we want to test?

1. Articles
2. Usage examples
3. Github tests

Differences in examples

1. Journals focus on different things
2. microtests vs one test to rule them all
3. If example fails, do we care why?

Our scoring method must...

1. Not discriminate any journal
2. Be prepared for any number of test/examples
3. Allow us to compare every two articles
4. Be as simple as possible

How do we score articles?

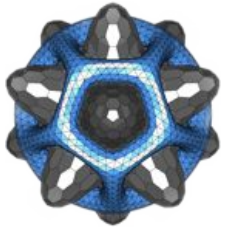
The more test we do the more we know
how to score them

Well… we don’t score them just yet

but after we gather enough data we will!

Which journals and why?

- The Journal of Open Source Software
  joss.theoj.org



- Journal of Machine Learning Research
  jmlr.org

JMLR

- Journal of Statistical Software
  jstatsoft.org

How do we chose articles?

1. N Python and N R articles from each journal.
2. N ~= 7 - about 40 articles in total.
3. Randomly chosen from time range 2016-present.

Insights that we can give you

based on what we know now

Python has more complex test/examples

Style of journal might define its reproducibility problems

What is to be done?

1. Further reproducibility testing.
2. Reproducibility metrics.
3. Presentation of results and conclusions.

Thank you for your time

Questions?