Hands-on transparent QMSKI: Open-access data, reproducible workflows, and interactive publications

Serena Bonaretti
https://sbonaretti.github.io/

22nd International Workshop on Quantitative Musculoskeletal Imaging

February 25, 2019

CC BY NC SA

10.5281/zenodo.2577617

Disclosures

- My first workshop on transparent research: Feel free to add and discuss information
- I will show *some* of the tools for transparent research, chosen among the most commonly used
- I do not have conflict of interest

What are openness and reproducibility?

- **Open science** refers to the free availability of data, software, and methods developed by researchers with the aim to share knowledge and tools to professionals and citizens (Woelfle 2011)

- **Reproducibility** is the ability of researchers to duplicate the results of a previous study using the same data, software, and methods used by the original authors (Bollen 2015)
  - **Replicability** is the recreation of same results using new data but the same experimental design (Gorgolewski)

Note: In some fields the two definitions are inverted

Replication crisis

- Discovery of personalized cancer treatment at Duke


The NEW ENGLAND JOURNAL of MEDICINE

This article has been retracted.

A correction has been published. ↑

ORIGINAL ARTICLE

A Genomic Strategy to Refine Prognosis in Early-Stage Non–Small-Cell Lung Cancer

Anil Potti, M.D., Sayan Mukherjee, Ph.D., Rebecca Petersen, M.D., Holly K. Dressman, Ph.D., Andrea Bild, Ph.D., Jason Koontz, M.D., Robert Kratzke, M.D., Mark A. Watson, M.D., Ph.D., Michael Kelley, M.D., Geoffrey S. Ginsburg, M.D., Ph.D., Mike West, Ph.D., David H. Harpole, Jr., M.D., et al.

(Potti 2006)


THE LANCET
Oncology

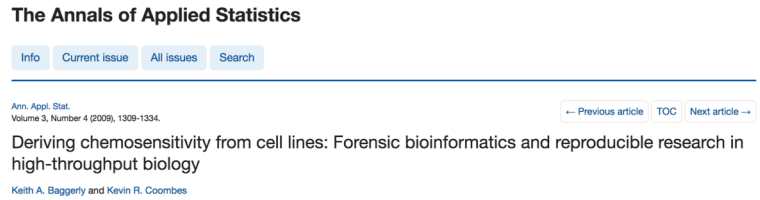
FAST TRACK — ARTICLES | VOLUME 8, ISSUE 12, P1071-1078, DECEMBER 01, 2007

RETRACTED: Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

Prof Hervé Bonnefoi, MD • Anil Potti, MD • Mauro Delorenzi, PhD • Louis Mauriac, MD • Mario Campone, MD • Michèle Tubiana-Hulin, MD • et al. Show all authors

Published: November 14, 2007 • DOI: https://doi.org/10.1016/S1470-2045(07)70345-5

(Bonnefoi 2007)


The Annals of Applied Statistics

Info | Current issue | All issues | Search

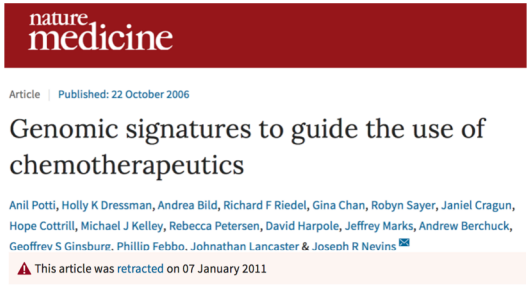
Ann. Appl. Stat.
Volume 3, Number 4 (2009), 1309-1334.

← Previous article | TOC | Next article →

Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology

Keith A. Baggerly and Kevin R. Coombes

(Baggerly 2009)


nature medicine

Article | Published: 22 October 2006

Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins

This article was retracted on 07 January 2011

(Potti 2006)


Journal of Clinical Oncology®
An American Society of Clinical Oncology Journal

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, Anil Potti

From the Division of Medical Oncology, Department of Medicine; Institute for Genome Sciences and Policy; Department of Surgery, Duke University, Durham, NC; and the Division of Gynecologic Oncology, H. Lee Moffitt Cancer Center, Tampa, FL

Show Less

A retraction has been published.

(Hsu 2007)

V. Stodden[1]: "None of that was picked up in peer-review. Nobody looks under the covers that deeply in peer-review. They were able to read the paper. This kind of thing is in the code and in the data itself"

2006-2010

[1]https://www.youtube.com/watch?v=dF1-nkqwmjI from min 13:30, and https://www.youtube.com/watch?v=eV9dcAGaVU8

Replication crisis

- Replication of studies in oncology

Comment | Published: 28 March 2012

Drug development

Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Nature **483**, 531–533 (29 March 2012) | Download Citation

(Begley 2012)

"Fifty-three papers were deemed 'landmark' studies…[S]cientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result."

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

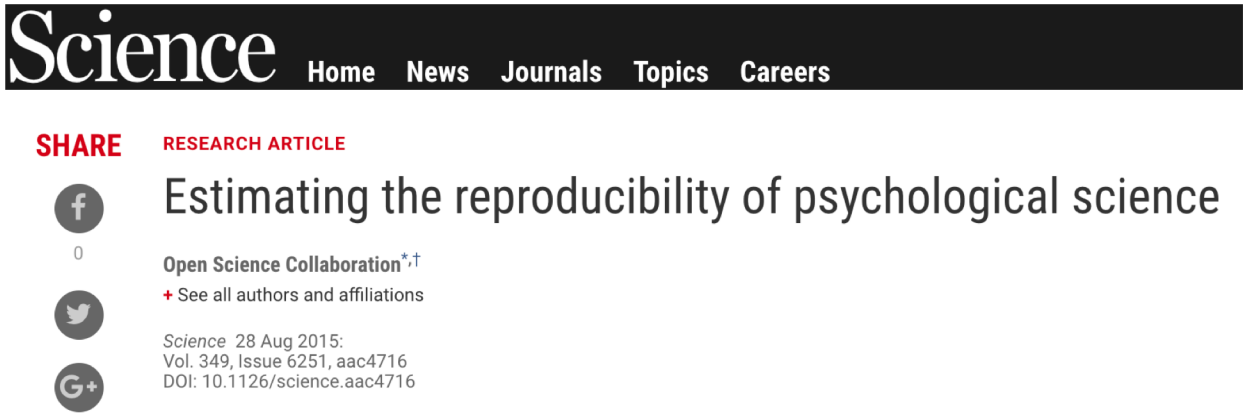
Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

https://blogs.plos.org/absolutely-maybe/2016/12/05/reproducibility-crisis-timeline-milestones-in-tackling-research-reliability/

Replication crisis

- Other fields: psychology


Science Home News Journals Topics Careers

SHARE

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

(Open Science Collaboration 2015)

"We conducted replications of 100 experimental and correlational studies […]. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval"

https://blogs.plos.org/absolutely-maybe/2016/12/05/reproducibility-crisis-timeline-milestones-in-tackling-research-reliability/

Replication crisis

- Daily update on paper retraction

https://retractionwatch.com/

Article	Year of retraction	Citing Articles before retraction	Citing Articles after retraction	Total cites (journals indexed by Web of Science)
Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N Engl J Med April 4, 2013 Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella, D, Aros F, Gomez-Gracia E, Ruiz-Gutiérrez V, Fiol M, Lapetra J, Lamuela-Raventos RM, Serra-Majem L, Pinto X, Basora J, Munoz MA, Sorli JV, Martinez JA, Martinez-Gonzalez MA, et al., for the PREDIMED Study Investigators	2018	1792	79	1917

https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/

Computational reproducibility crisis

HOW TO ENCOURAGE AND PUBLISH REPRODUCIBLE RESEARCH

Jelena Kovačević

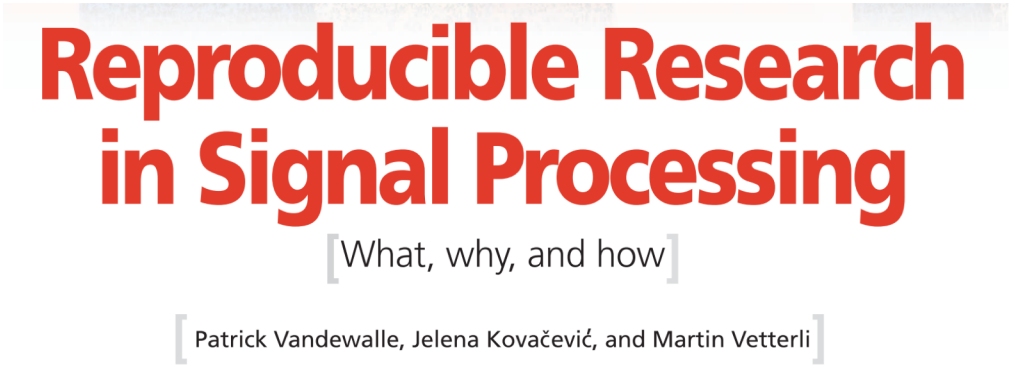
Depts. of Biomedical Engineering & Electrical and Computer Engineering
Carnegie Mellon University
Email: jelenak@cmu.edu

(Kovacevic 2007)

15 papers from IEEE Transactions on image processing:

- 33% data available
- 0% code available
- 60% pseudo code available

Computational reproducibility crisis



(Vandewalle 2009)

134 papers from IEEE Transactions on image processing in 2004:

- 33% data available
- 9% code available
- 33% pseudo code available

Computational reproducibility crisis

MIT Sloan School of Management

MIT Sloan School Working Paper 4773-10

The Scientific Method in Practice: Reproducibility in the Computational Sciences

Victoria Stodden

(Stodden 2010)

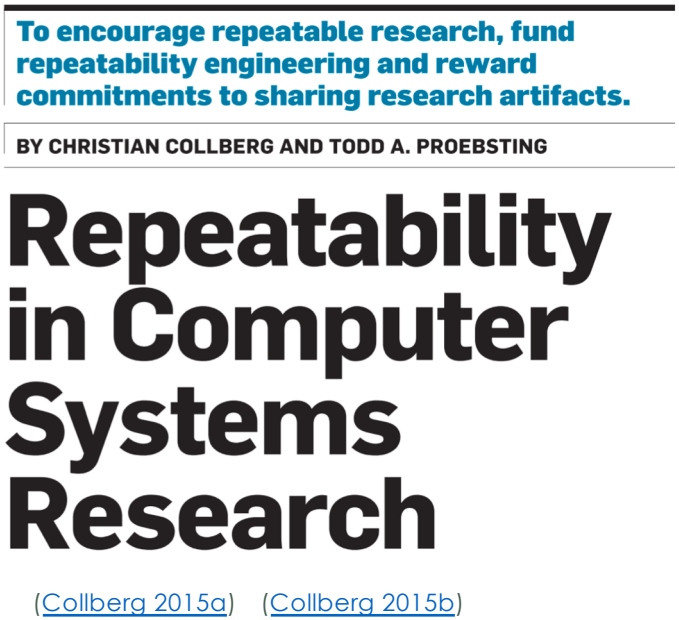
Survey at NIPS 2008
134 researchers

Before conference
- 74% willing to share code
- 67% willing to share data

After conference
- 30% shared some code
- 20% shared some data

Some reasons not to share:
- The time it takes to clean up and document for release
- The possibility that your data / code may be used without citation
- Competitors may get an advantage
- Dealing with questions from users about the code

Computational reproducibility crisis



(Collberg 2015a) (Collberg 2015b)

- 508 papers from 8 conferences and 5 journals (2012)
- Team of undergraduate students, graduate students, and postdocs
- They could reproduce algorithms of 226 (44%)

Why is it so difficult to reproduce studies?

- Papers must be concise (Vandewalle 2012)
  - Limited amount of words, figures, and tables
  - Authors have to select methods and parameters to present
- Data and software are in the supplementary material
  - Not in the paper body (Vandewalle 2012)
- Until a few years ago there were no permanent, large, and free repositories for data and software with DOI
  - Personal repositories often get deleted (Gil 2016)
- "Publish or perish" might favor quantity over quality and scientific bias (results have to be good) (Fanelli 2010)

Current way of publishing

- Publications are the tip of the iceberg
- "It's impossible to verify most of the results that computational scientists present at conferences and in papers" Donoho 2009
- "Scientific and mathematical journals are filled with pretty pictures of computational experiments that the reader has no hope of repeating" LeVeque 2009

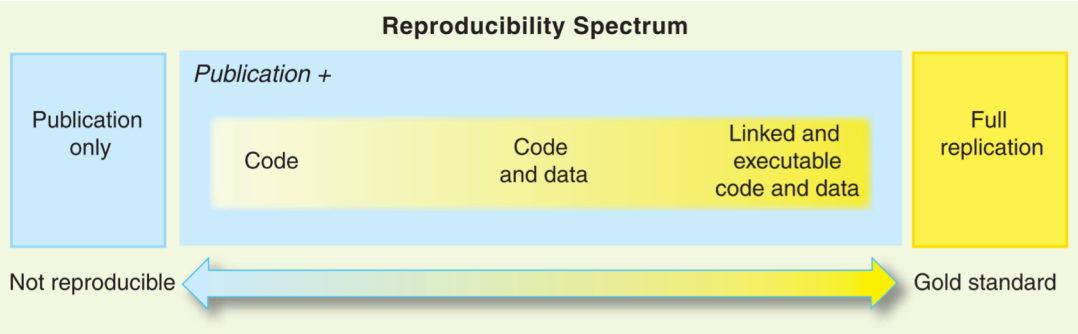
Slide modified from: http://mescal.imag.fr/membres/arnaud.legrand/teaching/2016/mosig_smpe_2_reproducible_research.pdf

Publications for transparent research

- "An author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a concrete definition of reproducibility in computationally oriented research" Claerbout 1992

- "An article about computational science in a scientific publication is not the scholarship itself, it's merely scholarship advertisement. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures" Donoho 2009

- Peng 2011:

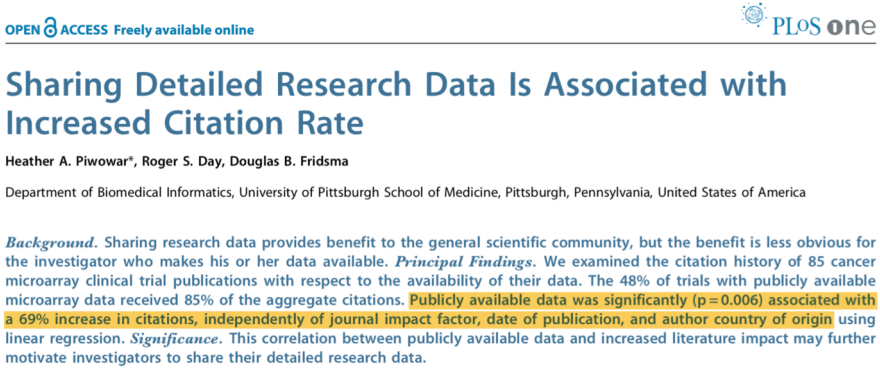


What are the benefits of transparent research?

- Openness and reproducibility are essential to researchers to:
  - Assess the value of scientific claims (Sandve 2013)
  - Compare new methods to existing ones (Freire 2018)
  - Build on the work of other scientists with confidence and efficiency, i.e. without "reinventing the wheel" (Rule 2018)
  - Collaborate to improve and expand robust scientific workflows to accelerate scientific discoveries (Donoho 2009, Munafo 2017)

What are the benefits of transparent research?

- Increased citation rate



OPEN ACCESS Freely available online

PLoS one

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Background. Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. ***Principal Findings.*** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. ***Significance.*** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

(Piwowar 2007)

REPRODUCIBLE RESEARCH FOR SCIENTIFIC COMPUTING

PATRICK VANDEWALLE

Code Sharing Is Associated with Research Impact in Image Processing

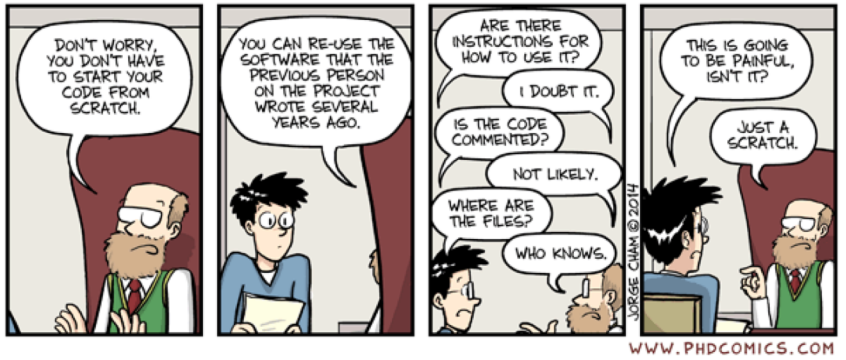
In computational sciences such as image processing, publishing usually isn't enough to allow other researchers to verify results. Often, supplementary materials such as source code and measurement data are required. Yet most researchers choose not to make their code available because of the extra time required to prepare it. Are such efforts actually worthwhile, though?

(Vandewalle 2012)

"The median number of citations […] increases with a factor of 3 when code is available online"

- Personal and group self-discipline (Donoho 2009)

- Reproducibility helps defeat self-deception (Nuzzo 2015)



Funding agencies support transparent research

- Europe: EOSC, Horizon 2020, OpenAire
- US: NIH, Gates Foundation, Chan-Zuckerberg Initiative
- Canada: Open data
- Australia, New Zealand, Asia, …

How can we conduct transparent research?

- Historically, research data, tools, and processes were rarely openly available because of limited storage and computational power (Munafo 2017)

- Nowadays there are several tools to conduct transparent research
  - Open access data
  - Reproducible workflows
  - Interactive publications

A few questions about our research practice

- How many of us have uploaded data and/or code to a public repository?
- How many of us use Jupyter notebook or R markdown for reproducible workflows?
- How many of us have written an interactive publication?

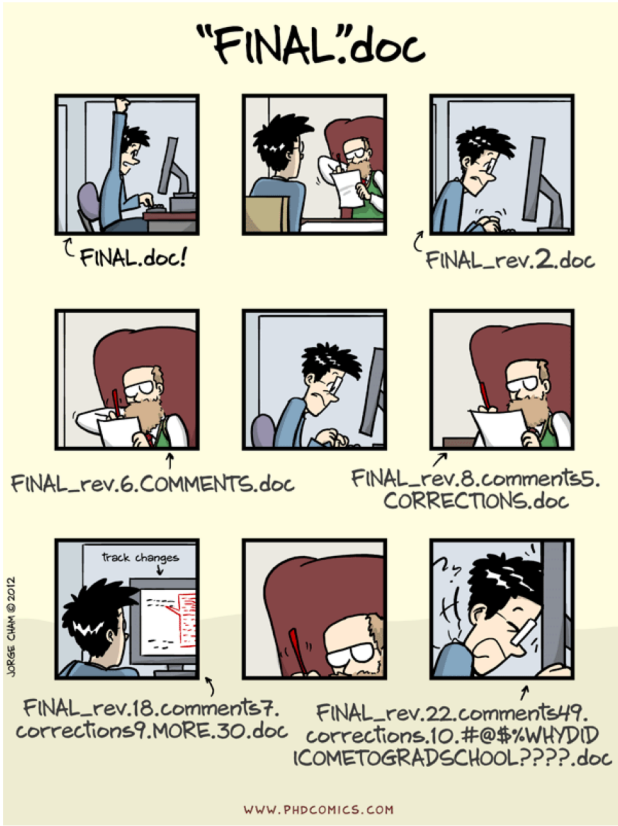
Hands-on transparent QMSKI:
Open-access data,
reproducible workflows,
and interactive publications

About open data

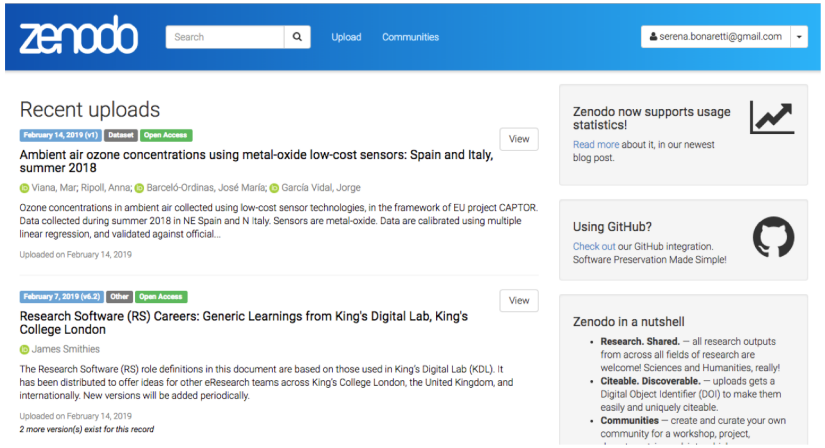
- Data = data and software

- Why open repositories?
  - Personal repositories often get deleted (Gil 2016)
  - Provide a DOI → Data and software are citable
  - Version control

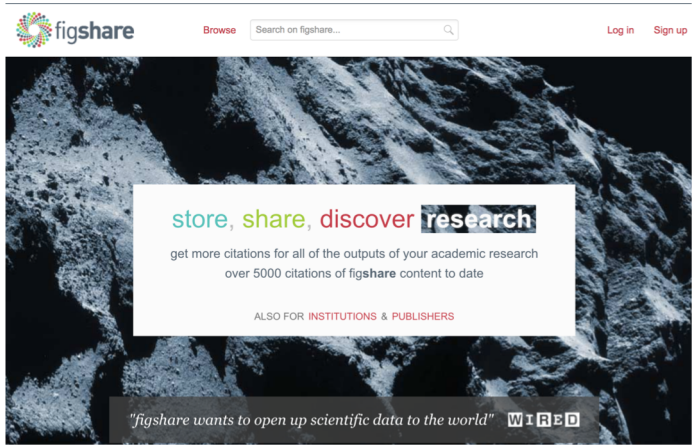
- Metadata and documentation
  - Data provenance and usage



Data repositories

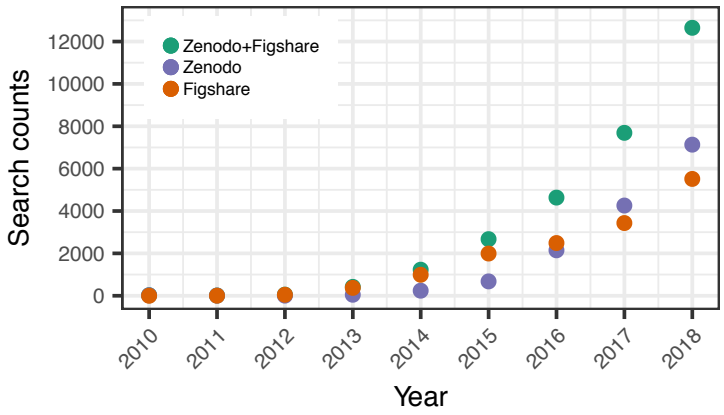


https://zenodo.org/



https://figshare.com/

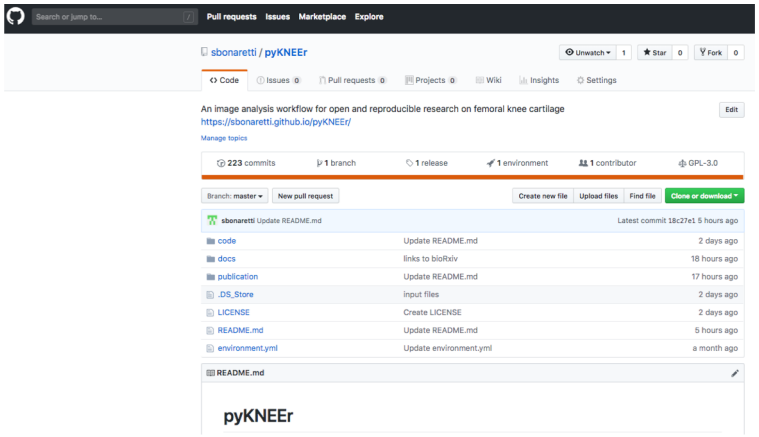
Zenodo and Figshare



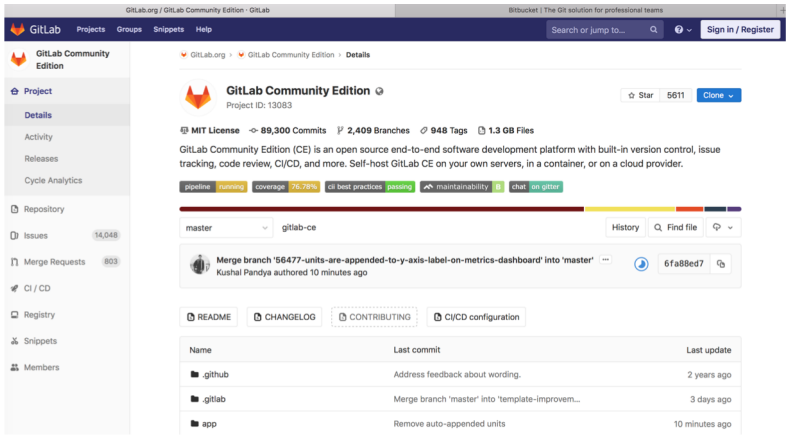
GitHub

- More repositories: a list on Nature

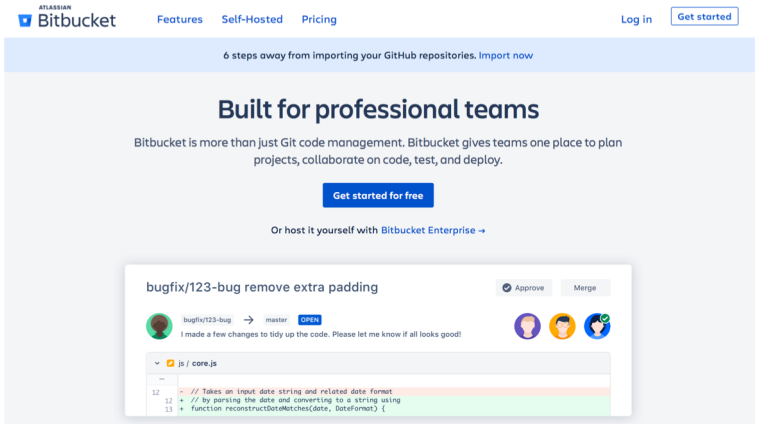
Software repositories



https://github.com/



https://about.gitlab.com



https://bitbucket.org

- Version control for reproducibility and collaborations

Metafiles and documentation

- Provenance of raw data
- Computational provenance of derived data
- Software documentation for code reuse

- Fields in repositories (e.g. Zenodo)
- Text files
- README.md in GitHub
- API
- Website
- …

License

"Free software does not mean public-domain software. Free software is copyrighted, but it comes with a public license that allows you to use it, copy it, and redistribute it provided you follow certain guidelines" Claerbout 1992

- Online material is automatically protected by copyright

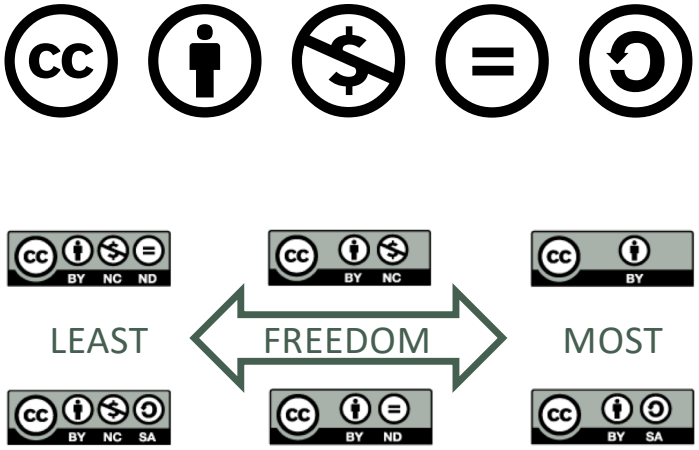
- If you want your material to be used:

Choose a license, "any" license

Data, publications, and code licenses

Data and publications

- Creative commons (https://creativecommons.org/choose)



Reproduced from: https://www.youtube.com/watch?v=8YkbeycRa2A

Code

(https://choosealicense.com/licenses/)

	GNU AGPLv3	GNU GPLv3	GNU LGPLv3	Mozilla Public License 2.0	Apache License 2.0	MIT License	The Unlicense
Commercial use	●	●	●	●	●	●	●
Distribution	●	●	●	●	●	●	●
Modification	●	●	●	●	●	●	●
Patent use	●	●	●	●	●		
Private use	●	●	●	●	●	●	●
Disclose source	●	●	●	●			
License and copyright notice	●	●	●	●	●	●	
Network use is distribution	●						
Same license	●	●	●	●	●		
State changes	●	●	●				
Liability	●	●	●	●	●	●	●
Warranty	●	●	●	●	●	●	●
Trademark use				●	●		

Hands-on transparent QMSKI: Open-access data, reproducible workflows, and interactive publications

How do we create reproducible workflows?

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, **2** Centre for Cancer Biomedicine, University of Oslo, Blindern, Oslo, Norway, **3** Department of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University Park, Pennsylvania, United States of America, **4** Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America, **5** Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway, **6** Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

(Sandve 2013)

PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Taking Advantage of Git and GitHub

Yasset Perez-Riverol[1]*, Laurent Gatto[2], Rui Wang[1], Timo Sachsenberg[3], Julian Uszkoreit[4], Felipe da Veiga Leprevost[5], Christian Fufezan[6], Tobias Ternent[1], Stephen J. Eglen[7], Daniel S. Katz[8], Tom J. Pollard[9], Alexander Konovalov[10], Robert M. Flight[11], Kai Blin[12], Juan Antonio Vizcaíno[1]*

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **2** Computational Proteomics Unit, Cambridge Systems Biology Centre, University of Cambridge, Cambridge, United Kingdom, **3** Applied Bioinformatics and Department of Computer Science, University of Tübingen, Tübingen, Germany, **4** Medizinisches Proteom-Center, Ruhr-Universität Bochum, Bochum, Germany, **5** Department of Pathology, University of Michigan, Ann Arbor, Michigan, United States of America, **6** Institute of Plant Biology and Biotechnology, University of Münster, Münster, Germany, **7** Centre for Mathematical Sciences, University of Cambridge, Cambridge, United Kingdom, **8** National Center for Supercomputing Applications and Graduate School of Library and Information Science, University of Illinois, Urbana, Illinois, United States of America, **9** MIT Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **10** Centre for Interdisciplinary Research in Computational Algebra, University of St Andrews, St Andrews, United Kingdom, **11** Department of Molecular Biology and Biochemistry, Markey Cancer Center, Resource Center for Stable Isotope-Resolved Metabolomics, University of Kentucky, Lexington, Kentucky, United States of America, **12** The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

(Perez-Riverol 2016)

PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

1 Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology, Duke University, Durham, North Carolina, United States of America, **4** Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States of America

(Wilson 2016)

arXiv.org > cs > arXiv:1810.08055

Computer Science > Other Computer Science

Ten Simple Rules for Reproducible Research in Jupyter Notebooks

Adam Rule, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, Mai H. Nguyen, Sara Brin Rosenthal, Fernando Pérez, Peter W. Rose

(Submitted on 13 Oct 2018)

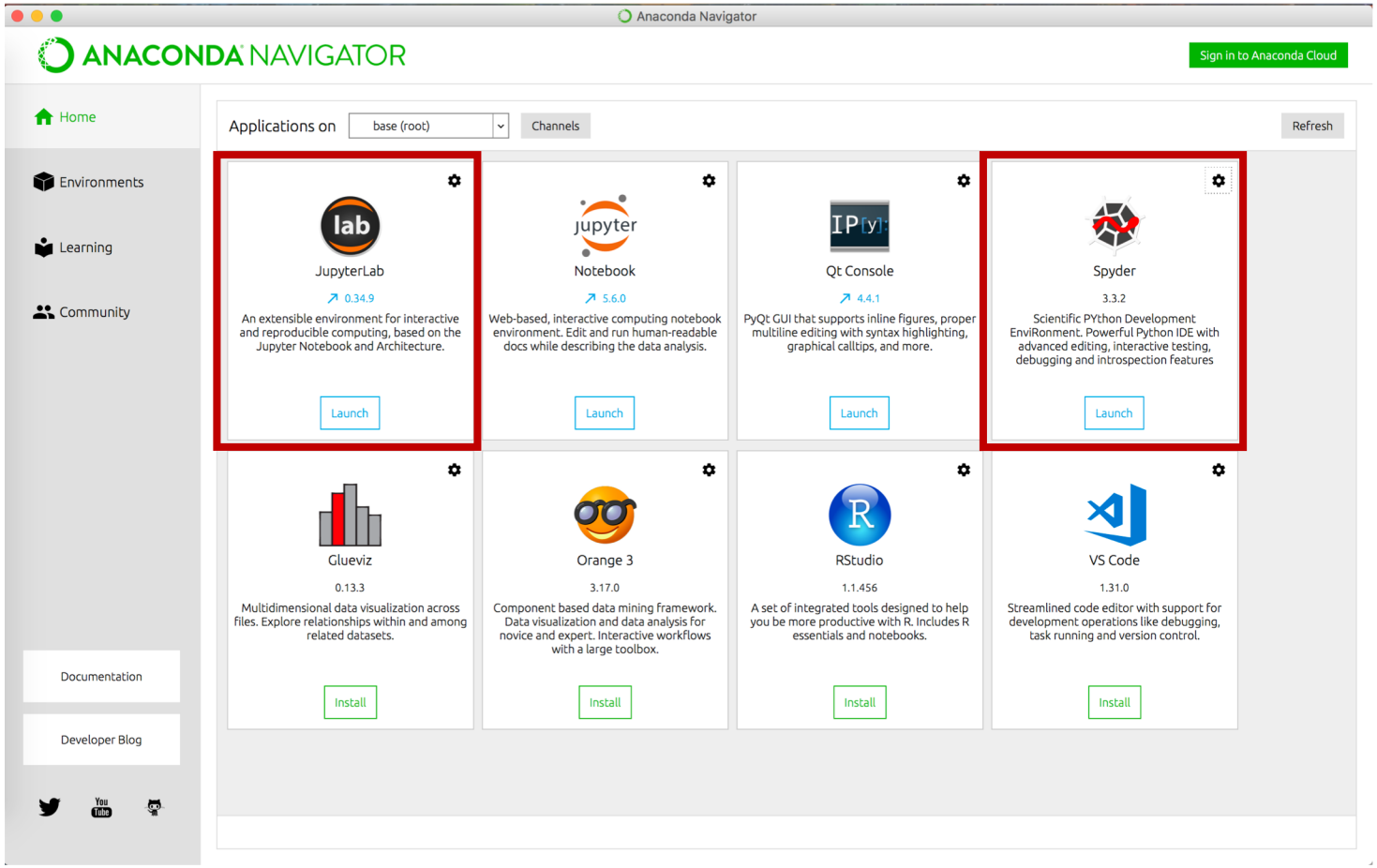
Reproducibility of computational studies is a hallmark of scientific methodology. It enables researchers to build with confidence on the methods and findings of others, reuse and extend computational pipelines, and thereby drive scientific progress. Since many experimental studies rely on computational analyses, biologists need guidance on how to set up and document reproducible data analyses or simulations.
In this paper, we address several questions about reproducibility. For example, what are the technical and non-technical barriers to reproducible computational studies? What opportunities and challenges do computational notebooks offer to overcome some of these barriers? What tools are available and how can they be used effectively?
We have developed a set of rules to serve as a guide to scientists with a specific focus on computational notebook systems, such as Jupyter Notebooks, which have become a tool of choice for many applications. Notebooks combine detailed workflows with narrative text and visualization of results. Combined with software repositories and open source licensing, notebooks are powerful tools for transparent, collaborative, reproducible, and reusable data analyses.

(Rule 2018)

Summary of rules

- Use open source programming language and file formats
  - e.g. python, R, .txt, .csv, …
- Automate data analysis
  - Avoid manual data manipulation
  - Tidy data tables with scripts
  - Always store raw data behind plots
- For every result keep track of how it was produced
  - Share and explain your data and code
  - Document the process, not just the results
  - Use version control, record dependencies
- Be your own user
  - Code well, be transparent, be simple
- Contribute to reproducible and open research
  - Don't be a perfectionist: "release early, release often"

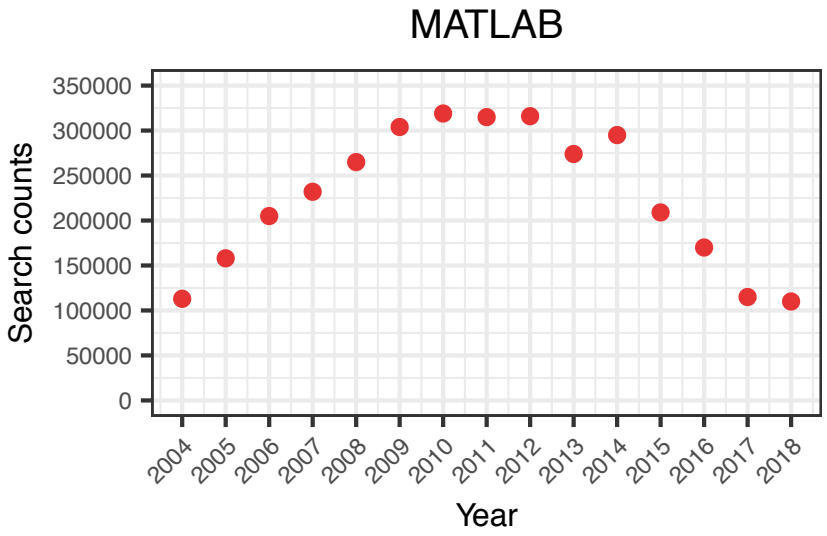
Open access language: python environment

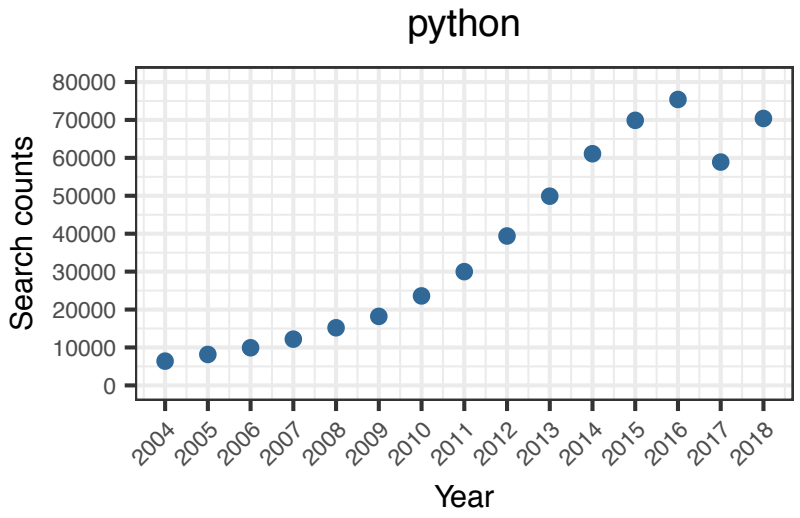


[Download Anaconda]

Open access language

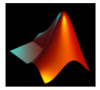
- Compatible with open and reproducible research
- A large amount of packages and shared code
- No license purchase





GitHub - Modified from Donoho 2015

From MATLAB to python: Differences



```
% initialize and print an array
% array name = [start:step:stop];


C = [2:2:8];
```

```
# initialize and print an array
# array name = arange(start,stop,step)

import numpy as np
C = np.arange(2,10,2)
```

```
% if structure

if a == 1 && b ~= 3
    fprintf('a=1 and b not 3 \n');
    fprintf('OK? \n');
end
```

```
# if structure

if a == 1 and b != 3:
    print('a=1 and b not 3');
    print('OK?')
```

```
% plotting



x = linspace(0,2*pi,100);
y = sin(x);
plot(x,y)
ylabel('sin(x)')
xlabel('x')
```

```
# plotting

import numpy as np
import matplotlib.pyplot as plt
x = np.linspace(0,2*np.pi,100)
y = np.sin(x)
plt.plot(x,y)
plt.ylabel('sin(x)')
plt.xlabel('x')
plt.show()
```

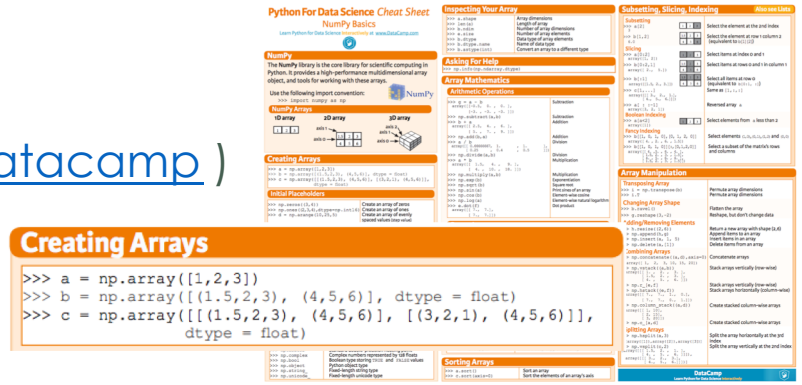
http://reactorlab.net/resources-folder/matlab/P_to_M.html

From MATLAB to python: Practical tips

- To start, translate part of your current code to python
  - Open MATLAB and Spyder, and translate lines one-to-one
  - Focus on syntax, not algorithm

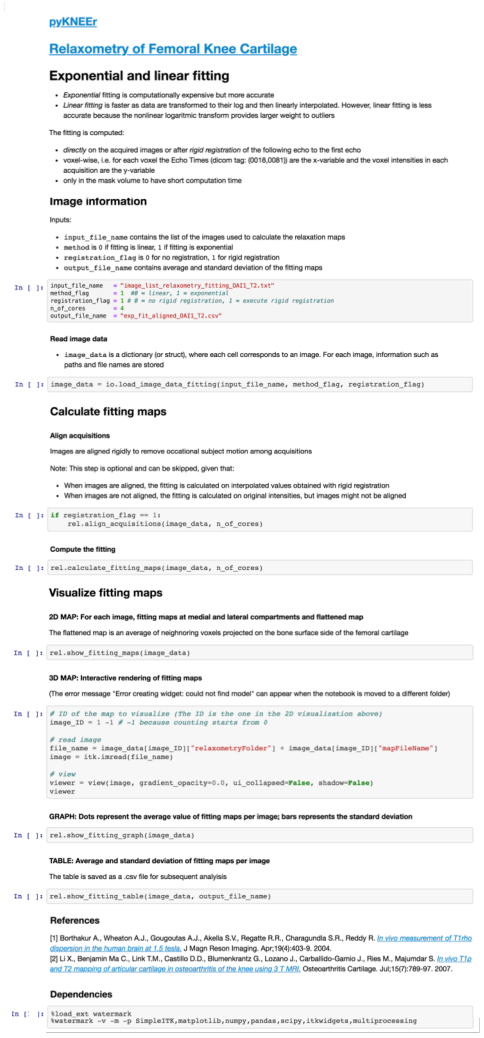
- Every time you are stuck, search for solutions online
  - There are plenty of question-and-answer sites (e.g. Stack Overflow) and blogs

- Take advantage of online material
  - Structured knowledge: Free online courses (e.g. datacamp)
  - Free style: Blogs, cheat sheet, and YouTube



- Take advantage of shared code
  - Look for already developed code, don't reinvent the wheel

Jupyter notebooks



pyKNEEr

Relaxometry of Femoral Knee Cartilage

Exponential and linear fitting

- *Exponential* fitting is computationally expensive but more accurate
- *Linear fitting* is faster as data are transformed to their log and then linearly interpolated. However, linear fitting is less accurate because the nonlinear logaritmic transform provides larger weight to outliers

The fitting is computed:

- *directly* on the acquired images or after *rigid registration* of the following echo to the first echo
- voxel-wise, i.e. for each voxel the Echo Times (dicom tag: (0018,0081)) are the x-variable and the voxel intensities in each acquisition are the y-variable
- only in the mask volume to have short computation time

Image information

Inputs:

- `input_file_name` contains the list of the images used to calculate the relaxation maps
- `method` is 0 if fitting is linear, 1 if fitting is exponential
- `registration_flag` is 0 for no registration, 1 for rigid registration
- `output_file_name` contains average and standard deviation of the fitting maps

```
input_file_name   = "image_list_relaxometry_fitting_OAI1_T2.txt"
method_flag       = 1  ## = linear, 1 = exponential
registration_flag = 1 # 0 = no rigid registration, 1 = execute rigid registration
n_of_cores        = 4
output_file_name  = "exp_fit_aligned_OAI1_T2.csv"
```

Read image data

- `image_data` is a dictionary (or struct), where each cell corresponds to an image. For each image, information such as paths and file names are stored

```
image_data = io.load_image_data_fitting(input_file_name, method_flag, registration_flag)
```

Calculate fitting maps

Align acquisitions

Images are aligned rigidly to remove occational subject motion among acquisitions

Note: This step is optional and can be skipped, given that:

- When images are aligned, the fitting is calculated on interpolated values obtained with rigid registration
- When images are not aligned, the fitting is calculated on original intensities, but images might not be aligned

```
if registration_flag == 1:
    rel.align_acquisitions(image_data, n_of_cores)
```

Compute the fitting

```
rel.calculate_fitting_maps(image_data, n_of_cores)
```

Visualize fitting maps

2D MAP: For each image, fitting maps at medial and lateral compartments and flattened map

The flattened map is an average of neighnoring voxels projected on the bone surface side of the femoral cartilage

```
rel.show_fitting_maps(image_data)
```

3D MAP: Interactive rendering of fitting maps

(The error message "Error creating widget: could not find model" can appear when the notebook is moved to a different folder)

```
# ID of the map to visualize (The ID is the one in the 2D visualization above)
image_ID = 1 -1 # -1 because counting starts from 0

# read image
file_name = image_data[image_ID]["relaxometryFolder"] + image_data[image_ID]["mapFileName"]
image = itk.imread(file_name)

# view
viewer = view(image, gradient_opacity=0.0, ui_collapsed=False, shadow=False)
viewer
```

GRAPH: Dots represent the average value of fitting maps per image; bars represents the standard deviation

```
rel.show_fitting_graph(image_data)
```

TABLE: Average and standard deviation of fitting maps per image

The table is saved as a .csv file for subsequent analysis

```
rel.show_fitting_table(image_data, output_file_name)
```

References

[1] Borthakur A., Wheaton A.J., Gougoutas A.J., Akella S.V., Regatte R.R., Charagundla S.R., Reddy R. *In vivo measurement of T1rho dispersion in the human brain at 1.5 tesla.* J Magn Reson Imaging. Apr;19(4):403-9. 2004.

[2] Li X., Benjamin Ma C., Link T.M., Castillo D.D., Blumenkrantz G., Lozano J., Carballido-Gamio J., Ries M., Majumdar S. *In vivo T1ρ and T2 mapping of articular cartilage in osteoarthritis of the knee using 3 T MRI.* Osteoarthritis Cartilage. Jul;15(7):789-97. 2007.

Dependencies

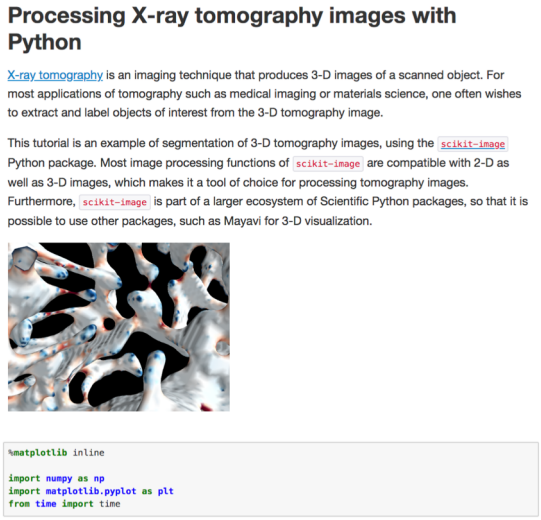
```
%load_ext watermark
%watermark -v -m -p SimpleITK,matplotlib,numpy,pandas,scipy,itkwidgets,multiprocessing
```

- Open-source web application integrating
  - Live code
  - Narrative text with equations
  - Visualizations
- Versatile
- Easy to share among researchers

launch binder

Example of notebooks in medical imaging

- Bone microCT
- SimpleITK notebooks
- SPIE 2019 workshop
- Deep Learning Toolkit
- VTK
- pyKNEEr

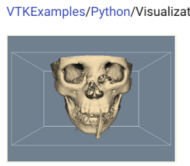


Processing X-ray tomography images with Python

X-ray tomography is an imaging technique that produces 3-D images of a scanned object. For most applications of tomography such as medical imaging or materials science, one often wishes to extract and label objects of interest from the 3-D tomography image.

This tutorial is an example of segmentation of 3-D tomography images, using the `scikit-image` Python package. Most image processing functions of `scikit-image` are compatible with 2-D as well as 3-D images, which makes it a tool of choice for processing tomography images. Furthermore, `scikit-image` is part of a larger ecosystem of Scientific Python packages, so that it is possible to use other packages, such as Mayavi for 3-D visualization.


```
%matplotlib inline

import numpy as np
import matplotlib.pyplot as plt
from time import time
```



VTKExamples/Python/VisualizationAlgorithms/HeadBone

Other Languages
See (Cxx)

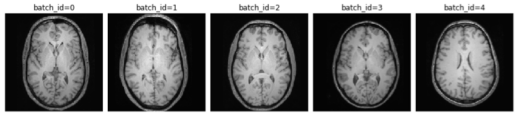
Code

HeadBone.py

```
#!/usr/bin/env python

import vtk


def main():
    fileName = get_program_parameters()

    colors = vtk.vtkNamedColors()
```

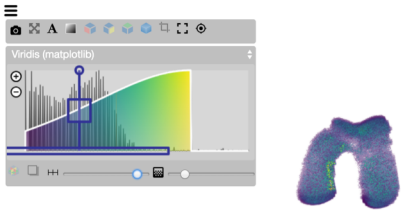
```
# Visualise the `dict_batch_feat` using matplotlib.
input_tensor_shape = dict_batch_feat.shape
center_slices = [s//2 for s in input_tensor_shape]

# Visualise the `gen_batch_feat` using matplotlib.
f, axarr = plt.subplots(1, input_tensor_shape[0], figsize=(15,5));
f.suptitle('Visualisation of the `dict_batch_feat` input tensor with shape={}'.format(input_tensor_shape))

for batch_id in range(input_tensor_shape[0]):
    # Extract a center slice image
    img_slice_ = np.squeeze(dict_batch_feat[batch_id, center_slices[1], :, :, :])
    img_slice_ = np.flip(img_slice_, axis=0)

    # Plot
    axarr[batch_id].imshow(img_slice_, cmap='gray');
    axarr[batch_id].axis('off')
    axarr[batch_id].set_title('batch_id={}'.format(batch_id))

f.subplots_adjust(wspace=0.05, hspace=0, top=0.8)
plt.show();
```

Visualisation of the `dict_batch_feat` input tensor with shape=(5, 128, 224, 224, 1)



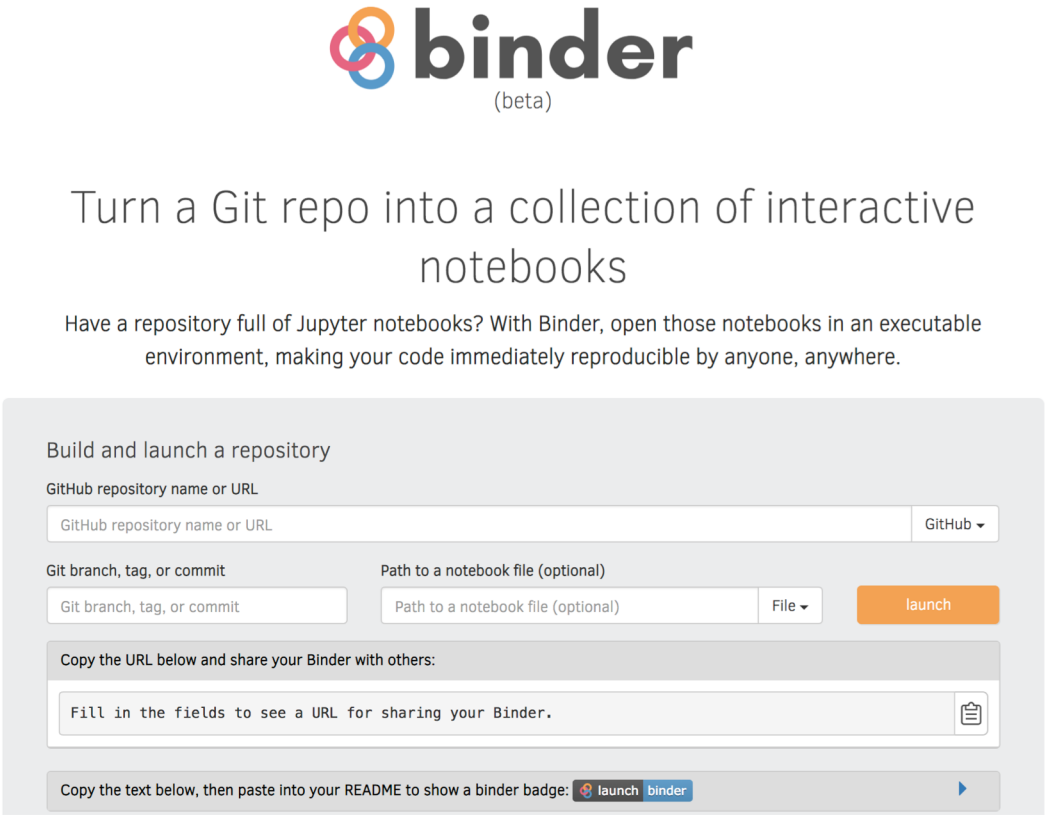
3D MAP: Interactive rendering of T_2 maps

```
# ID of the map to visualize (The ID is the one in the 2D visualization above)
image_ID = 1 -1 # -1 because counting starts from 0

# read image
file_name = image_data[image_ID]["relaxometryFolder"] + image_data[image_ID]["t2mapMaskFileName"]
image = itk.imread(file_name)

# view
viewer = view(image, gradient_opacity=0.0, ui_collapsed=False, shadow=False)
viewer
```



binder

- Online interactive computational environment



Computational environment

- Dependencies for reproducibility of computational environment
  - Package changes and future versions can be not compatible

Dependencies

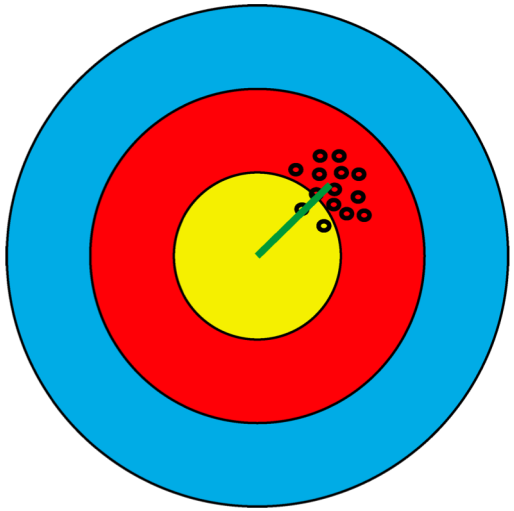
```
%load_ext watermark
%watermark -v -m -p matplotlib,numpy,pandas,scipy
```

```
CPython 3.7.1
IPython 7.2.0

matplotlib 2.2.3
numpy 1.16.1
pandas 0.24.1
scipy 1.2.1

compiler   : Clang 4.0.1 (tags/RELEASE_401/final)
system     : Darwin
release    : 17.7.0
machine    : x86_64
processor  : i386
CPU cores  : 4
interpreter: 64bit
```

Last note on reproducibility

- Reproducibility does not imply correctness
- Lack of reproducibility does not imply incorrectness

- Incorrectness can be found with reproducibility

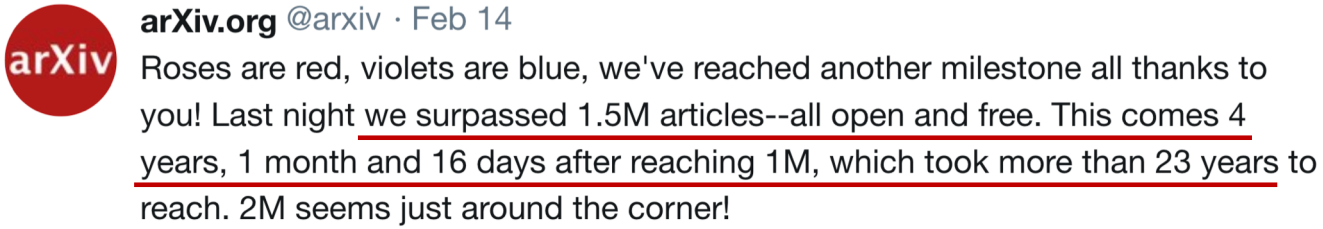


Holmes 2018

Hands-on transparent QMSKI: Open-access data, reproducible workflows, and interactive publications

Preprints and open access

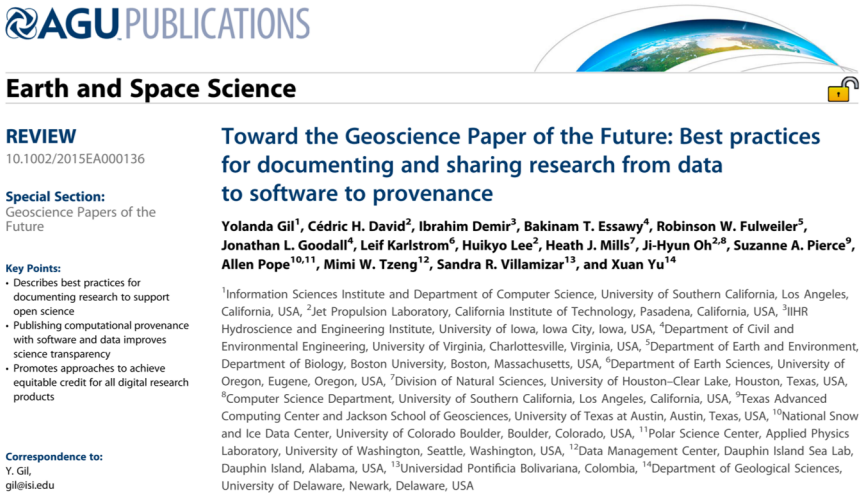
- Paper repositories: arXiv and bioRxiv



  - Major publishers accept preprints (list)

- Scientific journals:
  - Only open access: PLOS, Frontiers, F1000, …
  - Allow open access publications: Wiley, Elsevier,…
  - Allocation for open access publications in grants

How do we make interactive publications?



AGU PUBLICATIONS

Earth and Space Science

REVIEW

10.1002/2015EA000136

Special Section:
Geoscience Papers of the Future

Key Points:
- Describes best practices for documenting research to support open science
- Publishing computational provenance with software and data improves science transparency
- Promotes approaches to achieve equitable credit for all digital research products

Correspondence to:
Y. Gil,
gil@isi.edu

Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance

Yolanda Gil[1], Cédric H. David[2], Ibrahim Demir[3], Bakinam T. Essawy[4], Robinson W. Fulweiler[5], Jonathan L. Goodall[4], Leif Karlstrom[6], Huikyo Lee[2], Heath J. Mills[7], Ji-Hyun Oh[2,8], Suzanne A. Pierce[9], Allen Pope[10,11], Mimi W. Tzeng[12], Sandra R. Villamizar[13], and Xuan Yu[14]

[1]Information Sciences Institute and Department of Computer Science, University of Southern California, Los Angeles, California, USA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, [3]IIHR Hydroscience and Engineering Institute, University of Iowa, Iowa City, Iowa, USA, [4]Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, Virginia, USA, [5]Department of Earth and Environment, Department of Biology, Boston University, Boston, Massachusetts, USA, [6]Department of Earth Sciences, University of Oregon, Eugene, Oregon, USA, [7]Division of Natural Sciences, University of Houston–Clear Lake, Houston, Texas, USA, [8]Computer Science Department, University of Southern California, Los Angeles, California, USA, [9]Texas Advanced Computing Center and Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA, [10]National Snow and Ice Data Center, University of Colorado Boulder, Boulder, Colorado, USA, [11]Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, Washington, USA, [12]Data Management Center, Dauphin Island Sea Lab, Dauphin Island, Alabama, USA, [13]Universidad Pontificia Bolivariana, Colombia, [14]Department of Geological Sciences, University of Delaware, Newark, Delaware, USA

(Gil 2016)

On Reproducible AI

Towards reproducible research, open science, and digital scholarship in AI publications

Odd Erik Gundersen, Yolanda Gil and David W. Aha

Abstract

Background: Artificial intelligence, like any science, must rely on reproducible experiments to validate results. **Objective:** To give practical and pragmatic recommendations for how to document AI research so that results are reproducible. **Method:** Our analysis of the literature shows that AI publications currently fall short of providing enough documentation to facilitate reproducibility. Our suggested best practices are based on a framework for reproducibility and recommendations for best practices given by scientific organizations, scholars, and publishers. **Results:** We have made a reproducibility checklist based on our investigation and described how every item in the checklist can be documented by authors and examined by reviewers. **Conclusion:** We encourage authors and reviewers to use the suggested best practices and author checklist when considering submissions for AAAI publications and conferences.

(Gundersen 2018)

- Buttons and links to data and code
- Documentation of the computational provenance of results

Examples of interactive papers

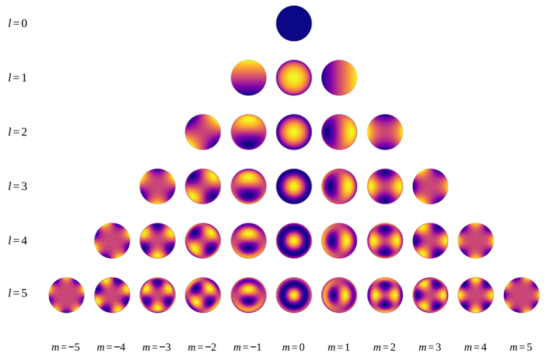


Figure 1. The real spherical harmonics up to degree $l = 5$ computed from Equation (1). In these plots, the x-axis points to the right, the y-axis points up, and the z-axis points out of the page.

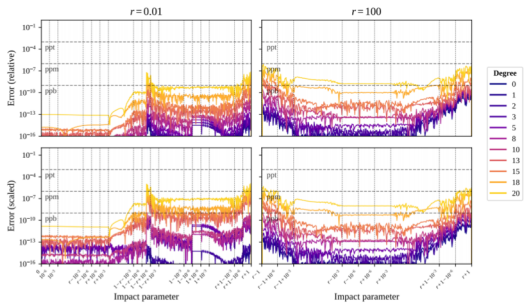


Figure 12. Similar to Figure 11, but showing instead the error on the *derivative* of the flux with respect to the impact parameter computed analytically with autodifferentiation. The error is computed relative to a numerical derivative computed at 128 bit precision.

(Luger 2018)

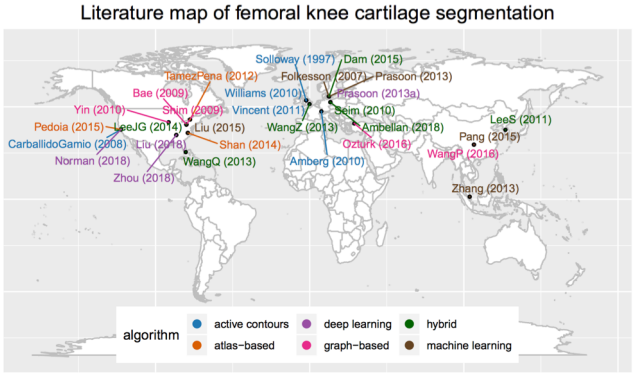


Figure 1: The visualization shows name of first author, year of publication, affiliation of last author, and segmentation method for 29 relevant publications on femoral knee cartilage segmentation from 1997 to 2018. Publications by segmentation method and in alphabetical order are: Active contours: Amberg(2010)[2], Carballido-Gamio(2008)[6], Solloway(1997)[62], Vincent(2011)[67], Williams(2010)[74]; Atlas-based: Pedoia(2015)[47], Shan(2014)[59]; Tamez-Pena(2012)[64]; Deep-learning: Liu(2018)[33], Norman(2018)[43], Prasoon(2013a)[52], Zhou(2018)[81]; Graph-based: Bae(2009)[4], Ozturk(2016)[44], Shim(2009)[60], WangP(2016)[68], Yin(2010)[78]; Hybrid: Ambellan(2018)[1], Dam(2015)[11], LeeJG(2014)[30], LeeS(2011)[31], Seim(2010)[57], WangQ(2013)[69], WangZ(2013)[70]; Machine learning: Folkesson(2007)[18], Liu(2015)[34], Pang(2015)[46], Prasoon(2013)[51], Zhang(2013)[80]. This graph and graphs in Fig. 4 and Fig. 5 were made in Jupyter notebook using `ggplot2` [71], an R package based on the grammar of graphics [72].

	Repository	Metadata / Documentation	Software / Language	License	DOI	Citation
Software Used						
Preprocessing	Bitbucket	Wiki	C++, ITK	Apache	https://doi.org/10.1016/j.media.2014.05.008*	[59]*
elastix 4.8	GitHub	Github Wiki	C++, ITK	Apache	https://doi.org/10.1109/TMI.2009.2035616*	[28]*
Developed pyKNEEr	GitHub	Website	python, Jupyter notebook	GNU GPLv3	https://doi.org/10.5281/zenodo.2574172	Bonaretti S. et al. "pyKNEEr" (v0.0.1). Zenodo. 2019. 10.5281/zenodo.2574172
Data						
Original	OAI	Website	-	Data user agreement	https://doi.org/10.1016/j.joca.2008.06.016*	[49]*
Derived (results)	Zenodo	Jupyter notebook	-	CC-BY-NC-SA	https://doi.org/10.5281/zenodo.2530609	Bonaretti S. et al. Dataset used in (Bonaretti et al. 2019). Zenodo. 2019. 10.5281/zenodo.2530609

Dataset	OAI1-DESS	OAI1-T_2	OAI2-BL	OAI2-FU	inHouse-DESS	inHouse-CQ
Number of subjects	19	19	88	88	4	4
I. Acquisition parameters						
Acquisition protocol	DESS	T_2-w	DESS		DESS	CubeQuant
Acquisition plane	sagittal	sagittal	sagittal		sagittal	sagittal
Number of images in series	2 (1 available)°	7	2 (1 available)°		2	4
In-plane spacing [mm]	0.3646 x 0.3646 (0.4270 x 0.4270)*	0.3125 x 0.3125 (0.4296 x 0.4296)*	0.3646 x 0.3646		0.3125 x 0.3125	0.3125 x 0.3125
Slice thickness [mm]	0.7 (0.75)*	3 (3.5)*	0.7		1.5	3
Echo time (TE) [ms]	4.7	10, 20, 30, 40, 50, 60, 70	4.7		42.52	-
Spin-lock time (TSL) [ms]	-	-	-		-	1, 10, 30, 60
Repetition time (TR) [ms]	16.32	2700 (2900)*	16.32		25	1302
Flip angle [°]	25	180	25		30	90
II. Ground truth segmentation						
Method	atlas-based		active models		- -	- -
Anatomy	femur, femoral cartilage		femoral cartilage		- -	- -
Type	mask		contour		- -	- -
III. Experimental results						
Image number in series	1	1 2-7	1	1	1 2	1 2-4
Preprocessing						
Spatial standardization	•	• •	•	•	• •	• •
Intensity standardization	•	• -	•	-	• -	• -
Segmentation						
Find reference	4, 8, 10, 13, 16	- -	-	-	- -	- -
Intersubject	•	- -	•	-	• -	- -
Longitudinal	-	- •	-	•	- -	- -
Multimodal	-	• -	-	-	- -	• -
Segmentation quality						
Dice coefficient	•	• -	•	•	- -	- -
Analysis						
Morphology	•○	•○ -	•○	•○	• -	• -
Relaxation	-	•○	-	-	•	•

(Bonaretti 2019)

Journal requirements

PLOS ONE

https://blog.frontiersin.org/2018/02/21/open-data-figshare-partnership/

Acceptable Data-Sharing Methods
Unacceptable Data Access Restrictions
Explanatory Notes and Guidance
Recommended Repositories
Repository Inclusion Criteria
FAQs for Data Policy
PLOS Data Advisory Board

Unacceptable Data Access Restrictions

PLOS journals will not consider manuscripts for which the following factors influence ability to share data:

- Authors will not share data because of personal interests, such as patents or potential future publications.
- The conclusions depend solely on the analysis of proprietary data, whether these data are owned by the authors, by their funders or institutions, or by other parties. We consider proprietary data to be data owned by commercial interests, or copyrighted data that the data owners will not share, e.g., data from a pharmaceutical company that will share the data only with regulatory agencies for purposes of drug approval, but not with researchers. If proprietary data are used and cannot be accessed by others (in the same manner by which the authors obtained them), the manuscript must include an analysis of public data that validates the conclusions so that others can reproduce the analysis and build on the findings.

See acceptable data access restrictions here.

https://journals.plos.org/plosone/s/data-availability#loc-unacceptable-data-access-restrictions

ELSEVIER

Home › Authors › Author services › Research Data › Open Data

https://www.elsevier.com/authors/author-services/research-data/open-data

Open Data

What is open data?

Elsevier supports the principle that "Raw research data should be made freely available to all researchers" and authors should be free to publically post their raw research data (see Author Rights for more details).

We have developed a simple way for authors to do this, by making their research data available on Mendeley Data and linking it to their article on ScienceDirect.

frontiers Science News

HOME NEWS OPEN SCIENCE POLICY MEDIA RELATIONS AWARDS

Home » Frontiers Announcements » Frontiers partners with Figshare to promote open data

Frontiers partners with Figshare to promote open data

Posted on February 21, 2018 in Frontiers Announcements, Top News

Improved visualization, citation and discoverability of supplementary research data outputs in Frontiers journals

figshare credit for all your research

London, UK, and Lausanne, Switzerland — Ahead of Open Data Day, Frontiers today announces its integration with Figshare's online digital repository. This broadens the types of supplementary data that can be included with Frontiers articles, and enhances the visualization, discoverability, citation and sharing of research data outputs. The new service — which is provided at no extra charge to authors — also helps Frontiers authors to satisfy institutional and funder requirements for open and FAIR (Findable, Accessible, Interoperable, and Reusable) data.

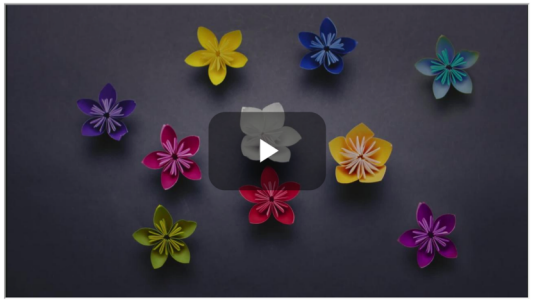
WILEY

Open Research > Open Data

https://authorservices.wiley.com/open-research/open-data/index.html

- Author Resources
- Reviewers
- Editors
- Ethics Guidelines
- Help
- Open Research
  - Open Access
  - Open Data
  - Open Practices
  - Open Collaboration
  - Open Recognition and Reward

Open Data



- Protect the long-term integrity of your research by making your data, methodologies and reporting standards openly available
- Comply with funder requests to share data
- Authors of articles published in Wiley journals are encouraged to share their research data including, but not limited to: raw data, processed data, software, algorithms, protocols, methods, materials. Visit our Author Compliance Tool for the policy of your chosen journal

More material on transparent research

- Journals
  - PLOS Computational Biology, Nature, Science, …
- Community websites
  - Data and software carpentry, …
- Platforms
  - Open Science Framework, European Science Cloud, …
- Blogs, YouTube (lectures), Twitter (follow the experts)

Take home message