

# The value of research data for scientific communication practices

Stephanie van de Sandt

University supervisor:  
Prof. Vivien Petras, PhD

CERN supervisors:  
Lars Holm Nielsen  
Artemis Lavasa  
Dr. Kamran Naim  
Dr. Sünje Dallmeier-Tiessen



28/10/2020

18th Gentner Day

Or:

Research data scary tales  
from a PhD in Library and Information Science  
on Data Parasites

# Science is important... and expensive!

“After a marathon five-day summit of the European Council, on which EU heads of state sit, leaders agreed to give €81 billion to the upcoming flagship research programme, Horizon Europe, which starts in January 2021. [...]

nature View all Nature Research Journals Search Login

Explore our content Journal information Subscribe

nature > news > article


NEWS · 22 JULY 2020

## Science money slashed in EU's €1.8-trillion budget deal

European Union's flagship research programme allocated €81 billion in latest negotiations – substantially less than previously proposed.

Quirin Schiermeier

[Twitter](#) [Facebook](#) [Email](#)



European Union leaders assembled in Brussels for their first in-person summit since the coronavirus pandemic began to escalate in March. Credit: Stephanie Lecocq/EPA-EFE/Shutterstock

### Sign up to Nature Briefing

An essential round-up of science news, opinion and analysis, delivered to your inbox every weekday.

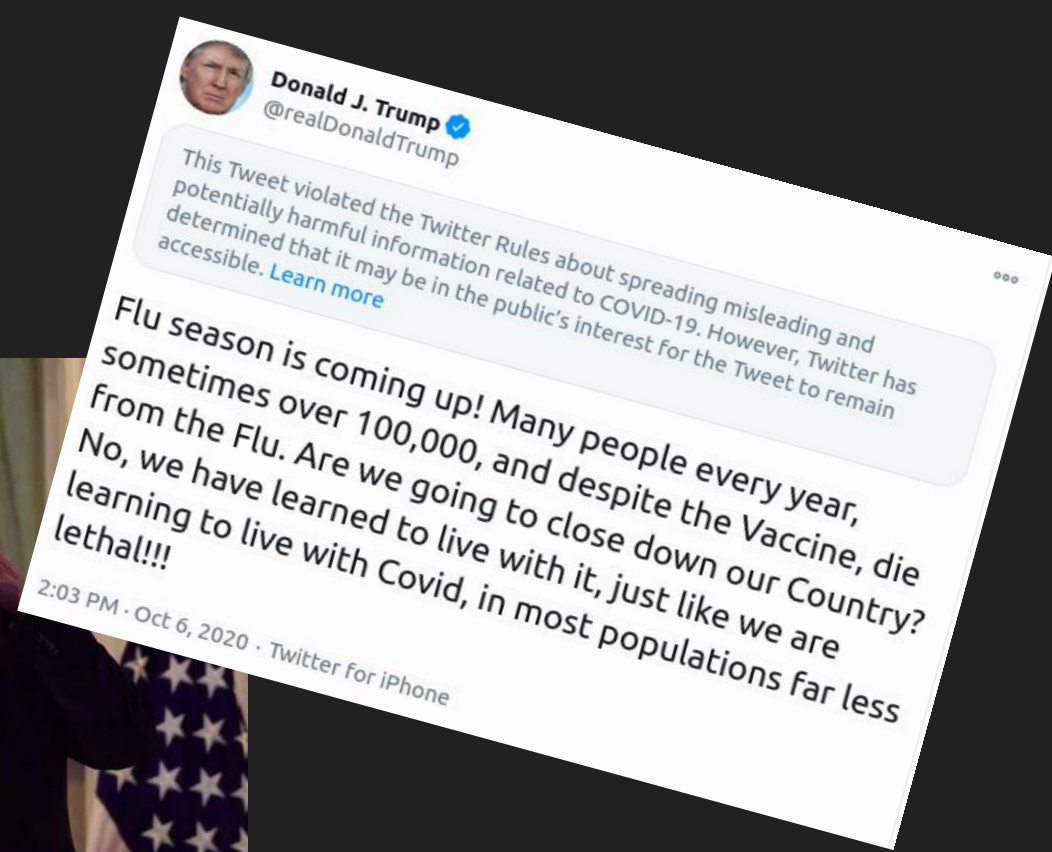
**Email address**

Yes! Sign me up to receive the daily Nature Briefing email. I agree my information will be processed in accordance with the Nature and Springer Nature Limited Privacy Policy.

**Sign up**

Worth it?

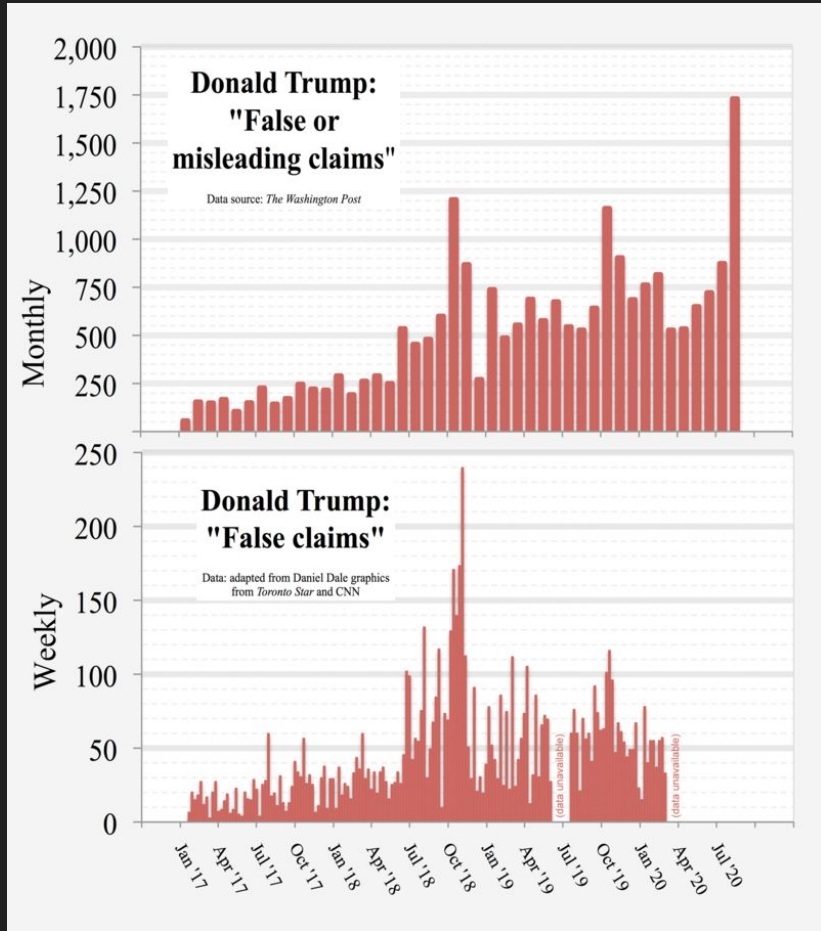
# Trick or Treat?



# Trick! .. or is it?

<https://www.thestar.com/news/world/analysis/2019/06/05/donald-trump-has-now-said-more-than-5000-false-claims-as-president.html>

<https://www.washingtonpost.com/graphics/politics/trump-claims-database/>



# Trick or Treat?

“Universal mask use could save an additional 129,574 (85,284–170,867) lives from September 22, 2020 through the end of February 2021, or an additional 95,814 (60,731–133,077) lives assuming a lesser adoption of mask wearing (85%), when compared to the reference scenario.



The screenshot shows the top portion of a web page for a Nature Medicine article. At the top left is the 'nature medicine' logo. Below it are two navigation links: 'Explore our content' and 'Journal information', both with downward-pointing chevrons. A red horizontal line separates this header from the main content area. Below the line is a breadcrumb trail: 'nature > nature medicine > articles > article'. The main title of the article is 'Modeling COVID-19 scenarios for the United States' in a large, bold, black font. Above the title, there are three links: 'Article', 'Open Access', and 'Published: 23 October 2020'. Below the title is the author information: 'IHME COVID-19 Forecasting Team'. Further down are two more links: 'Nature Medicine (2020)' and 'Cite this article'. At the bottom of the visible section are three metrics: '84k Accesses', '3774 Altmetric', and 'Metrics'.

<https://doi.org/10.1038/s41591-020-1132-9>

# Treat! ... Trick??!

“A review of this source indicates, however, that public mask use for the United States sat at a significantly higher rate of 68% as of 21 September, the stated date. This higher number is also consistent with more recent survey data, suggesting U.S. mask usage in public spaces has consistently hovered between 75 and 80% since mid-July 2020 – a figure much closer to the IHME’s own targeted mask compliance rates.”

## Widely cited COVID-19-masks paper under scrutiny for inaccurate stat



Retraction  
Watch



Data is needed  
to verify research statements

# Research Data Horror Week



# Research is in a (trust) crisis

The image shows a screenshot of the Nature journal website. At the top, the 'nature' logo is displayed in white on a dark red background, with the tagline 'International weekly journal of science' below it. A search bar with a 'Go' button is in the top right. A navigation menu includes 'Home', 'News & Comment', 'Research', 'Careers & Jobs', 'Current Issue', 'Archive', 'Audio & Video', and 'For Authors'. Below this, a breadcrumb trail shows 'Archive > Volume 533 > Issue 7604 > Editorial > Article'. The article title is 'Reality check on reproducibility' under the 'NATURE | EDITORIAL' section. The main text begins with 'A survey of Nature readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.' The date is '25 May 2016'. There are buttons for 'PDF' and 'Rights & Permissions'. A social media section shows 'Like' and 'Share' buttons, with a notification that 'Eamonn Maguire and 258,314 others like this.' A 'Crisis talks' sidebar features an illustration of two flasks, one with red liquid and one with yellow liquid and a red 'X', with the headline '1,500 scientists lift the lid on reproducibility' and a sub-headline 'Survey sheds light on the 'crisis' rocking research.' Below this is a 'Related stories' section with two links: 'The pressure to publish pushes down quality' and 'Research data: Silver lining to irreproducibility'. At the bottom right, there is a blue banner for 'Sign up for FREE today' with a green circular logo. A navigation bar at the very bottom has 'Recent', 'Read', and 'Commented' tabs.

**nature** International weekly journal of science

Search   [Advanced search](#)

[Home](#) | [News & Comment](#) | [Research](#) | [Careers & Jobs](#) | [Current Issue](#) | [Archive](#) | [Audio & Video](#) | [For Authors](#)

[Archive](#) > [Volume 533](#) > [Issue 7604](#) > [Editorial](#) > [Article](#)

NATURE | EDITORIAL

## Reality check on reproducibility

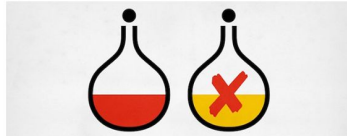
A survey of *Nature* readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.

25 May 2016

Is there a reproducibility crisis in science? Yes, according to the readers of *Nature*. [Two-thirds of researchers who responded to a survey by this journal](#) said that current levels of reproducibility are a major problem.

The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite. But, the survey suggests, there is a bigger issue — and something that needs to be fixed. [One-third of the survey respondents said that they think](#)

**Crisis talks**




**1,500 scientists lift the lid on reproducibility**  
Survey sheds light on the 'crisis' rocking research.

Eamonn Maguire and 258,314 others like this.

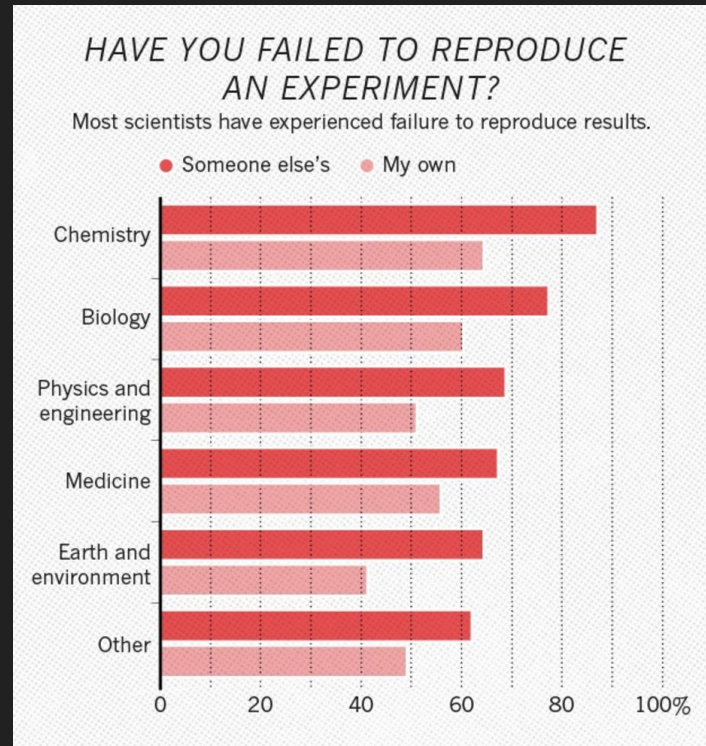
**Related stories**

- [The pressure to publish pushes down quality](#)
- [Research data: Silver lining to irreproducibility](#)

Sign up for **FREE** today 

Recent **Read** Commented

# (Most) Research results are not reproducible



<https://dx.doi.org/10.1038/533452a>

# When data is too good to be true

The screenshot shows the homepage of 'Dutch Daily News', which is described as 'DAILY DUTCH NEWS IN ENGLISH'. The navigation menu includes categories: NEWS, BUSINESS, TECHNOLOGY, ENTERTAINMENT, HEALTH, TRAVEL, and NETHERLANDS. The breadcrumb trail indicates the current page is 'Home » food » Meat eaters are selfish and less social'. The main article title is 'Meat eaters are selfish and less social', posted in the 'food, News' category with 29 comments. The article text reads: "Meat brings out the worst in people". This is what psychologists of the Radboud University Nijmegen and Tilburg University concluded from varrious studies on the psychological significance of meat. To the right, under 'RECENT NEWS', there are two featured articles: 'Why The Netherlands dominates the World Happiness Report rankings' and 'More British people moving to the Netherlands', both posted by Dutch Daily News.

<https://dutchdailynews.com/meat-eaters-selfish-less-social/>

Meat eaters are “selfish bastards”

# Misconduct and data fraud

“I failed as a scientist. I adapted research data and fabricated research. Not once, but several times, not for a short period, but over a longer period of time.”

International

## Levelt: fraud detected in 55 publications

Redactie • 28 november 2012

The fraud of Diederik Stapel could go unnoticed because of the failure of criticism in a culture of poor science. This statement is one of the conclusions of the final report of joint committees Levelt, Noort and Drenth (from Tilburg, Groningen and Amsterdam respectively) investigating the Stapel case. The results are being presented today in Amsterdam.

... if data is available at all



# And CERN?

# LHCb External Data Access Policy

## 2018 CMS data preservation, re-use and open access policy

CMS collaboration

Cite as: CMS collaboration (2018). 2018 CMS data preservation, re-use and open access policy. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.7347.JDWH

### ATLAS Data Access Policy

#### Introduction

ATLAS has fully supported the principle of open access. This document outlines the policy of ATLAS as regards open access to data described in the DPHEP [1] model. The main objective is to make data available in a usable way to people external to the ATLAS collaboration.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some elements with open access. To support open access to data additional resources are needed to develop and support the tools to make the data available.

#### Policies for Different Data Levels

Open access to ATLAS data by people outside the collaboration can be provided at different levels of increasing complexity, listed below, with associated conditions.

Documentation

## ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community. These data embed unique scientific information for the in depth understanding of the profound nature and origin of matter. Because of their uniqueness, long term preservation must be an essential objective of the data processing framework and will lay the foundations of the ALICE Collaboration legacy to the scientific community as well as to the general public. These considerations call for a detailed assessment of the ALICE data preservation strategy and policy. Documentation, long term preservation at various levels of abstraction, data access and analysis policy and software availability constitute the key elements of such a data preservation strategy allowing future collaborators, the wider scientific community and the general public to analyze data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the redaction addressed by the ALICE data preservation policy.



# CERN Open Data Portal

opendata  
CERN

Explore more than **two petabytes**  
of open data from particle physics!

Start typing...  Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

**Explore**

- [datasets](#)
- [software](#)
- [environments](#)
- [documentation](#)

**Focus on**

- [ATLAS](#)
- [ALICE](#)
- [CMS](#)
- [LHCb](#)
- [OPERA](#)
- [Data Science](#)


Get started

# What is the benefit of Open Data?


- The availability of the underlying data makes research results **transparent & verifiable**
- Research results are **reproducible**, also in the long-term
- Research data can be **reused** by others
- Publicly available data and software can be **properly cited**
- Research data citations provide **credit** to the data producer(s)
- Metrics based on research data citations may demonstrate the **impact** of data and software projects
- Data metrics have the potential to prove the **impact** of a research project to research funders

# Data citations as incentives for Open Science

Tibor Simko hat retweetet





 **Kati Lassila-Perini** @KatiLassila · 3. Juni

Feeling motivated: see refs [33] & [34] in @CMSExperiment paper [journals.aps.org/prl/pdf/10.1103...](https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.119.132003) !! That's how it goes: 1. Release #OpenData - cool! 2. External people do great work and publish papers - awesome! 3. The results are relevant to @CMSExperiment - mind blowing!

 Tweet übersetzen

[33] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, and W. Xue, Exposing the QCD Splitting Function with CMS Open Data, *Phys. Rev. Lett.* **119**, 132003 (2017).

[34] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler, Jet substructure studies with CMS open data, *Phys. Rev. D* **96**, 074003 (2017).

 1  5  11 



**Kyle Cranmer**

@KyleCranmer

I love watching the citation count grow on @ATLASexperiment Higgs data @inspirehep @HepData @datacite #DOI <http://t.co/5i2pjDINdy>

07 Dec 2014

# The issue with data / software citations

The screenshot shows a Zenodo dataset page for 'Citations to software and data in Zenodo via open sources'. The page includes a search bar, user profile, and navigation tabs for 'Dataset' and 'Open Access'. The main content area features a title, author list, and a descriptive paragraph. Below this is a 'Files' section with a table of two CSV files. A 'Citations' section is highlighted with a red box, showing a 'Beta' status and zero citations. The right sidebar contains 'Pending approvals', view/download statistics (120 views, 2,375 downloads), and 'Indexed in OpenAIRE' information.

zenodo Search Upload Communities stephanie.van.de.sandt@cern.ch

October 11, 2019 Dataset Open Access Edit

## Citations to software and data in Zenodo via open sources

van de Sandt, Stephanie; Ioannidis, Alex; Nielsen, Lars Holm

In January 2019, the Asclepias Broker harvested citation links to Zenodo objects from three discovery systems: the NASA Astrophysics Datasystem (ADS), Crossref Event Data and Europe PMC. Each row of our dataset represents one unique link between a citing publication and a Zenodo DOI. Both endpoints are described by basic metadata. The second dataset contains usage metrics for every cited Zenodo DOI of our data sample.

Preview

Files (1.9 MB)

Name	Size	Preview	Download
asclepias_broker_citations_201901_processed.csv md5:778e3273db371bf2cef87644f24c2796	1.3 MB	Preview	Download
asclepias_broker_citations_201901_usagemetrics.csv md5:7b40886bf3ce75147e9b64e633a00ace	672.5 kB	Preview	Download

**Beta** Citations 0

Show only:  Literature (0)  Dataset (0)  Software (0)  Unknown (0)  
 Citations to this version

No citations.

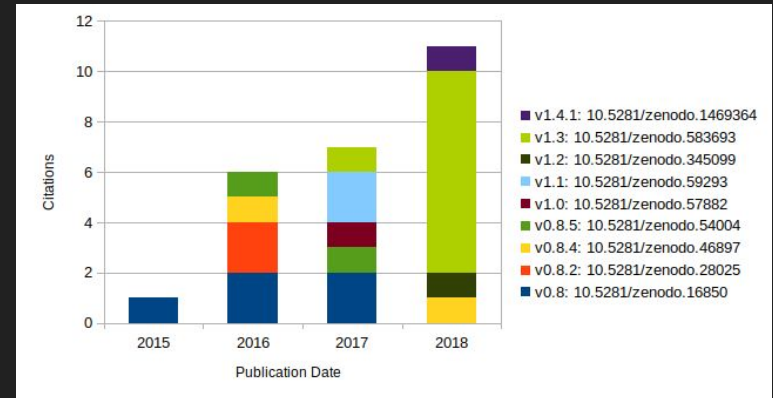
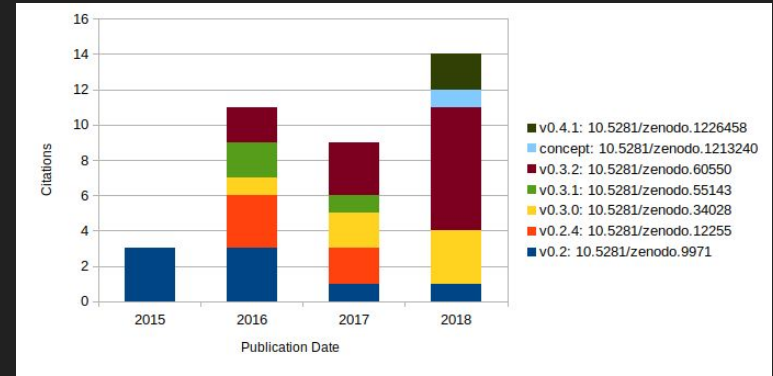
120 views 2,375 downloads  
See more details...

Indexed in OpenAIRE

Publication date: October 11, 2019  
DOI: 10.5281/zenodo.3482927

# Data citations are hard to capture

- Only 0.33% of all DOIs registered by the research data repository Zenodo are traceably cited at least once
- 98.5% of all analyzed citations to datasets on Zenodo proved to be self-citations
- The impact of software is hard to track as citations may point to multiple versions



# Data citations are idiosyncratic

## B Some issues in cleaning the 32 s record

In figure 1 of the main text, we showed the 32 s raw data taken from the LIGO archive (<https://losc.ligo.org/events/GW150914/>). The LIGO team used a 4096 s record for cleaning the data sets, which reveals the same peak-like structure in the power spectra for Hanford and Livingston detectors. In figure 11 we show the power spectra for the 4096 s raw data

[76] "Jet primary dataset in AOD format from RunB of 2010 (/Jet/Run2010B-Apr21ReReco-v1/AOD)," CMS Collaboration, CERN Open Data Portal (), 10.7483/OPENDATA.CMS.3S7F.2E9W.

HEP grid systems. Functionality, to allow users to perform various individual HEP grid systems and grid sites. Flexibility, to make H open for the various distributed grid computing environments. Differer to connect to the Ganga job monitoring system and to check the perfo among the grid sites, have been implemented. The new HappyFace syst integrated and now it displays the information and the status of both t

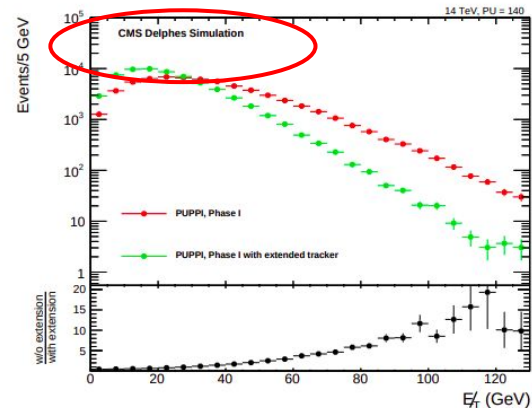
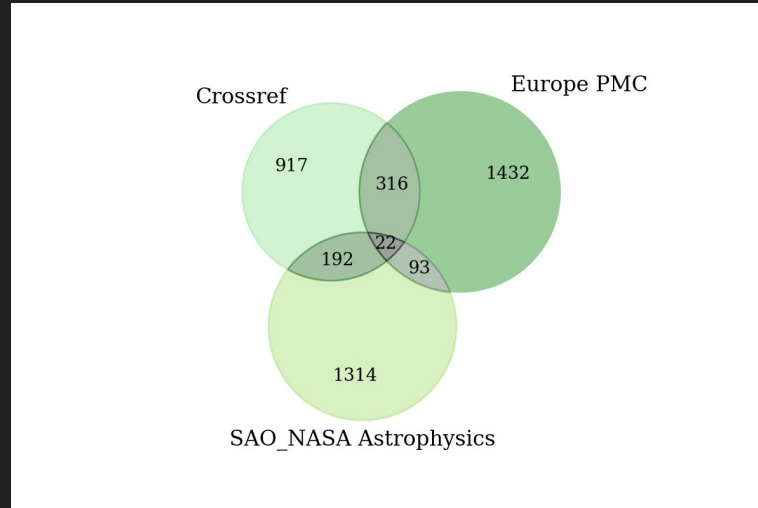


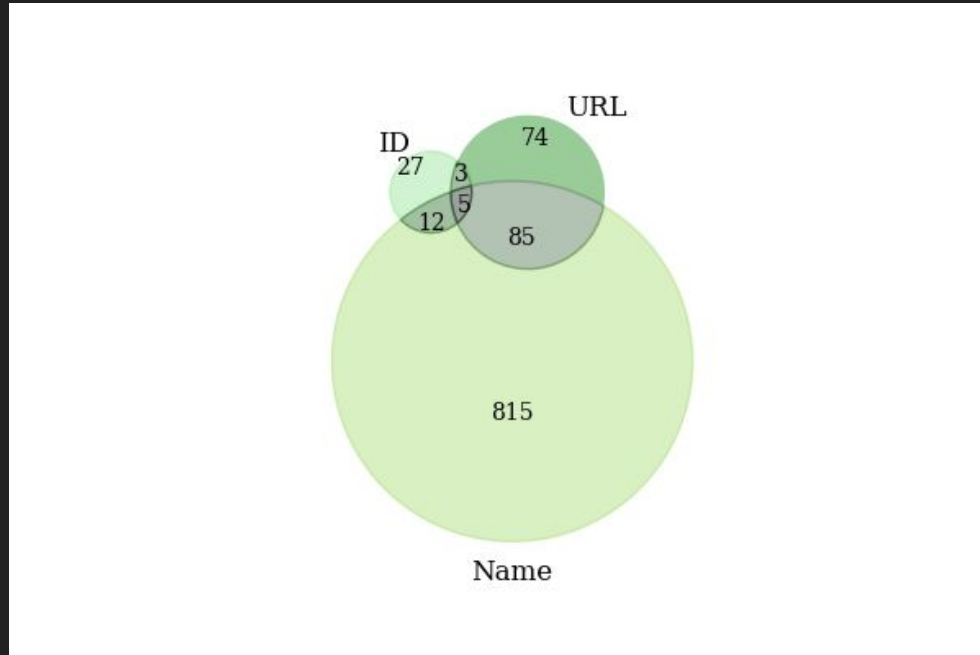
Figure 1: Comparison of the  $E_{miss}$  resolution with the present CMS tracker and with an hypothetical extended tracker up to  $|\eta| = 4$ . [1, 6]

# Data citation discovery tools are insufficient



Overlap of tracked citations by multiple citation discovery services

...and rely on formal data citations



Overlap of search term efficiency



# The tracking of research data citations

- We can only observe a fraction of real data and software usage.
- Tracking research data reuse in publications is challenging due to technical and social issues.
- Citations to research data and software are not standardized, and not yet adopted in common publication practices.
- Data creators and software developers do not receive credit for their work.

Research data and software are valuable and should be available as open as possible and as soon as possible.

It is everybody's responsibility to improve scholarly communication by giving credit where credit is due.

# When it works...

## corner.py

build passing coverage 27% license BSD DOI 10.5281/zenodo.53155 JOSS 10.21105/joss.00024

Read [the documentation](#).

If you make use of this code, please cite [the JOSS paper](#).

```
@article{corner,  
  doi = {10.21105/joss.00024},  
  url = {https://doi.org/10.21105/joss.00024},  
  year = {2016},  
  month = {jun},  
  publisher = {The Open Journal},  
  volume = {1},  
  number = {2},  
  pages = {24},  
  author = {Daniel Foreman-Mackey},  
  title = {corner.py: Scatterplot matrices in Python},  
  journal = {The Journal of Open Source Software}
```

README.rst

## triangle.py

Make some beautiful corner

**Corner plot / körnär plät/**  
An illustrative representation of a corner plot promise.

Built by [Dan Foreman-Mackey](#)  
Licensed under the 2-clause

## Installation

Just run

build passing coverage 87% license BSD DOI 10.5281/zenodo.53155

## Documentation

- Installation
  - Dependencies
  - Using pip
  - From source
  - Tests
- Getting started
- A note about sigmas
- Custom plotting
- Detailed API documentation

## Attribution

If you make use of this code, please cite [the JOSS paper](#):

```
@article{corner,  
  Author = {Daniel Foreman-Mackey},  
  Doi = {10.21105/joss.00024},  
  Title = {corner.py: Scatterplot matrices in Python},  
  Journal = {The Journal of Open Source Software},  
  Year = 2016,  
  Volume = 24,  
  Url = {http://dx.doi.org/10.5281/zenodo.45906}  
}
```

July 9, 2020

## dfm/corner.py: corner.py v2.1.0.rc1

Dan Foreman-Mackey, Adrian Price-Whelan, Will Youdens, Geoffrey Ryan, Matt Pitkin, Victor Zabeltz, jshejy, Arfon Smith, Gregory Ashton, Michael Smith, Emily Rice, Brendon J. Brewer, Brigitta Sipőcz, David W. Hogg, Eric Gonthier, Hanno Rein, Hennadii Madar, Ian Czekajka, James Tocknell, Kyle Barbary, Remy Frechelt, Stephan Hoyer, Thomas A Caswell, Wolfgang Kerzendorf, Kelle Cruz

Release candidate: A maintenance release to keep infrastructure up to date.

Preview

dfm/corner.py: corner.py v2.1.0.rc1.zip	
dfm-corner.py-ef62538	61 Bytes
coverage	2.4 KB
github	143 Bytes
workflows	56 Bytes
tests.yml	1.5 KB
gitignore	138 Bytes
.rst-environment.yml	1.4 KB
LICENSE	688.7 KB
MANIFEST.in	1.2 KB
README.rst	36 Bytes
corner.png	8.3 KB
demo.py	
docs	
gitignore	
Makefile	
static	

Beta

Citations 189

Show only:  Literature (184)  Unknown (4)  Software (1)  Dataset (0)

Citations to this version

	<b>Credit Lost: Two Decades of Software Citation in Astronomy</b> Bouquin, Daina R. et al. (DOI: 10.3847/1538-4365/ab7be6)	2020	<a href="#">ADS</a> <a href="#">DOI</a>
	<b>Digging for Relics of the Past: The Ancient and Obscured Bu...</b> Cadelano, M. et al. (DOI: 10.3847/1538-4357/ab88b3)	2020	<a href="#">ADS</a> <a href="#">ARXIV</a>

Confusing citation recommendation in a messy identifier universe - solved! (for now)

Thank you for listening!