Workshop on Reproducible Science: Practical Exercises

Stefanie Scherzinger and Stefan Klessinger,
assisted by Johannes Pietrzyk

University of Passau, Germany

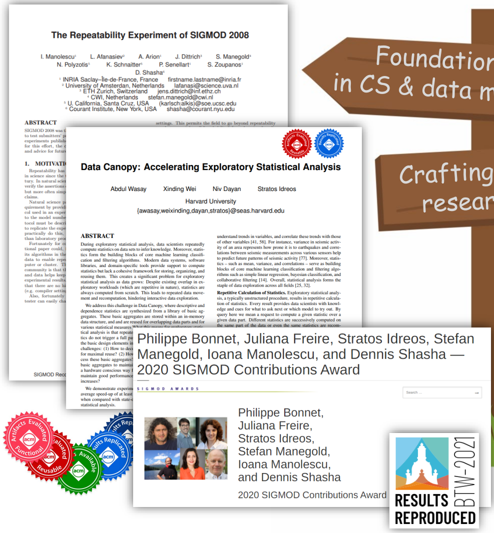
June 2022

References & Credits

Joint work with Wolfgang Mauerer and Stefan Klessinger:

- ICDE 2021 tutorial “Nullius in Verba: Reproducibility for Database Systems Research, Revisited.” (Mauerer, Scherzinger)
- SEENG@ICSE 2022 paper “Beyond the Badge: Reproducibility Engineering as a Lifetime Skill.” (Mauerer, Klessinger, Scherzinger)
- Q-SANER@SANER 2022 paper “1-2-3 Reproducibility for Quantum Software Experiments.” (Mauerer, Scherzinger)
- Flipped classroom course on “Reproducibility Engineering”, jointly taught in winter term 21/22 at OTH Regensburg (Mauerer) and Uni Passau (Scherzinger, Klessinger) (videos: `https://tinyurl.com/repeng`)

What to Expect



The Repeatability Experiment of SIGMOD 2008

I. Manolescu[1] L. Afanasiev[2] A. Arion[1] J. Dittrich[3] S. Manegold[4]
N. Polyzotis[5] K. Schnaitter[5] P. Senellart[1] S. Zoupanos[1]
D. Shasha[6]

[1] INRIA Saclay–Île-de-France, France firstname.lastname@inria.fr
[2] University of Amsterdam, Netherlands lafanasi@science.uva.nl
[3] ETH Zurich, Switzerland jens.dittrich@inf.ethz.ch
[4] CWI, Netherlands stefan.manegold@cwi.nl
[5] U. California, Santa Cruz, USA {karlsch|alkis}@soe.ucsc.edu
[6] Courant Institute, New York, USA shasha@courant.nyu.edu

SIGMOD Record

Data Canopy: Accelerating Exploratory Statistical Analysis

Abdul Wasay Xinding Wei Niv Dayan Stratos Idreos

Harvard University
{awasay,weixinding,dayan,stratos}@seas.harvard.edu

ABSTRACT

During exploratory statistical analysis, data scientists repeatedly compute statistics on data sets to infer knowledge. Moreover, statistics form the building blocks of core machine learning classification and filtering algorithms. Modern data systems, software libraries, and domain-specific tools provide support to compute statistics but lack a cohesive framework for storing, organizing, and reusing them. This creates a significant problem for exploratory statistical analysis as data grows: Despite existing overlap in exploratory workloads (which are repetitive in nature), statistics are always computed from scratch. This leads to repeated data movement and recomputation, hindering interactive data exploration.

We address this challenge in Data Canopy, where descriptive and dependence statistics are synthesized from a library of basic aggregates. These basic aggregates are stored within an in-memory data structure, and and are reused for overlapping data parts and for various statistical measures

understand trends in variables, and correlate these trends with those of other variables [41, 58]. For instance, variance in seismic activity of an area represents how prone it is to earthquakes and correlations between seismic measurements across various sensors help to predict future patterns of seismic activity [77]. Moreover, statistics – such as mean, variance, and correlations – serve as building blocks of core machine learning classification and filtering algorithms such as simple linear regression, bayesian classification, and collaborative filtering [14]. Overall, statistical analysis forms the staple of data exploration across all fields [25, 32].

Repetitive Calculation of Statistics. Exploratory statistical analysis, a typically unstructured procedure, results in repetitive calculation of statistics. Every result provides data scientists with knowledge and cues for what to ask next or which model to try out. By query here we mean a request to compute a given statistic over a given data part. Different statistics are successively computed on

Philippe Bonnet, Juliana Freire, Stratos Idreos, Stefan Manegold, Ioana Manolescu, and Dennis Shasha — 2020 SIGMOD Contributions Award

SIGMOD AWARDS

Philippe Bonnet,
Juliana Freire,
Stratos Idreos,
Stefan Manegold,
Ioana Manolescu,
and Dennis Shasha

2020 SIGMOD Contributions Award

Artifacts Evaluated Functional
Artifacts Evaluated Reusable
Artifacts Available
Results Replicated
acm
RESULTS REPRODUCED
BTW-2021

Foundational work in CS & data management
Crafting gold-standard research artefacts

git

Reproducible Builds

docker

doi

zenodo

Long-Term Maintenance in Industry

- Boeing 747 aircraft
  - Development started in 1966
  - Last machines produced in 2022 will be in service until about 2050
- Bitcoin
- Tor Browser
- Civil Infrastructure Platform Initiative: Linux Kernels

Piggyback Strategy for Research

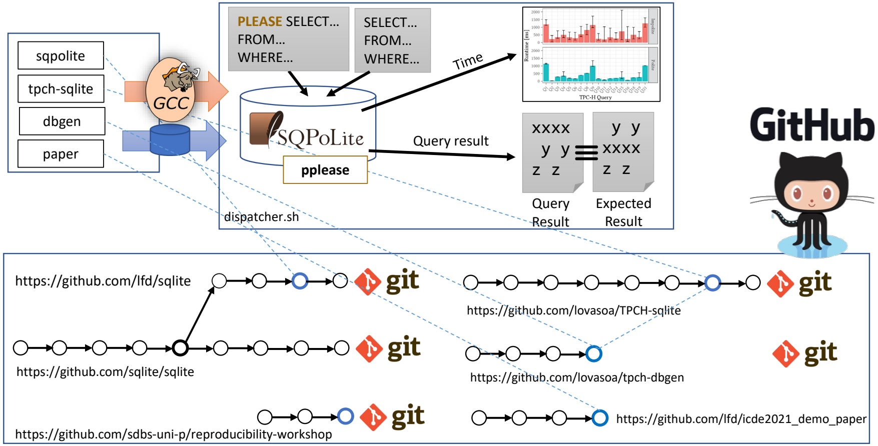
Put your money on open source tools that are massively employed:
as industry has a strong incentive (and the resources) to maintain them.

Agenda

Exercises coming up

- Comparing research results — different notions of “identity”
- Reproducible builds — common problems and how ReproTest can help spot them
- Working with a reproduction package — modify a package to get to know your way around

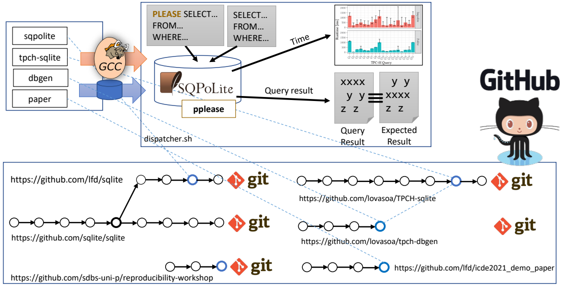
Example Scenario


sqpolite
tpch-sqlite
dbgen
paper
GCC
PLEASE SELECT...
FROM...
WHERE...
SELECT...
FROM...
WHERE...
SQPoLite
pplease
Time
Query result
xxxx
y y
z z
≡
y y
xxxx
z z
Query Result
Expected Result
TPC-H Query
Runtime [ms]
dispatcher.sh
GitHub
https://github.com/lfd/sqlite
git
https://github.com/sqlite/sqlite
https://github.com/sdbs-uni-p/reproducibility-workshop
https://github.com/lovasoa/TPCH-sqlite
https://github.com/lovasoa/tpch-dbgen
https://github.com/lfd/icde2021_demo_paper

Play-along docker recipe: `https://github.com/sdbs-uni-p/reproducibility-workshop`.

Managing Components and Processes

- Artefacts (data, code, SW tools, scripts)
- Build process
- Need to compare both the artefacts *and* the results.



sqpolite
tpch-sqlite
dbgen
paper
GCC
PLEASE SELECT... FROM... WHERE...
SELECT... FROM... WHERE...
SQPoLite
pplease
Time
Query result
xxxx y y z z
Query Result
Expected Result
dispatcher.sh
GitHub
https://github.com/lfd/sqlite
https://github.com/sqlite/sqlite
https://github.com/sdbs-uni-p/reproducibility-workshop
https://github.com/lovasoa/TPCH-sqlite
https://github.com/lovasoa/tpch-dbgen
https://github.com/lfd/icde2021_demo_paper
git

Oceans and Islands

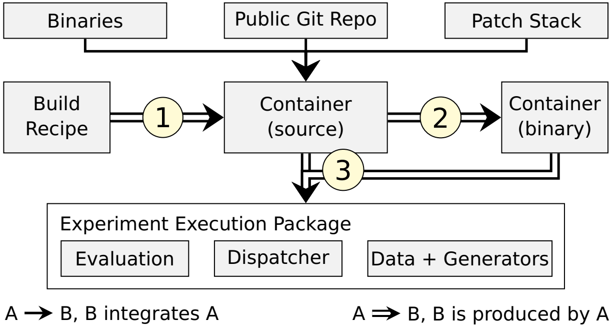
Be independent of updates to external components, because. . .

- Base system details change (package versions, etc.)
- System runtime configuration (beyond distro and kernel versions) changes
- Github repositories disappear
- Projects move between hosts (Sourceforge, GitHub, . . .)
- External software is no longer maintained, download links disappear
- . . .

Goal: Build self-contained, complete environments

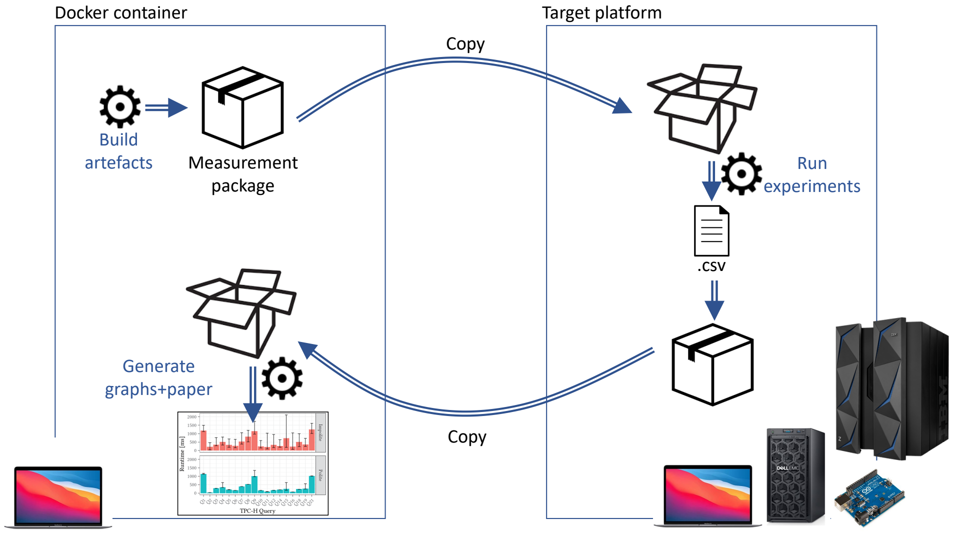
Be ready to build your stuff even when you trapped on an island
without internet access, or 20 years after all the repositories have gone.

Deliverables

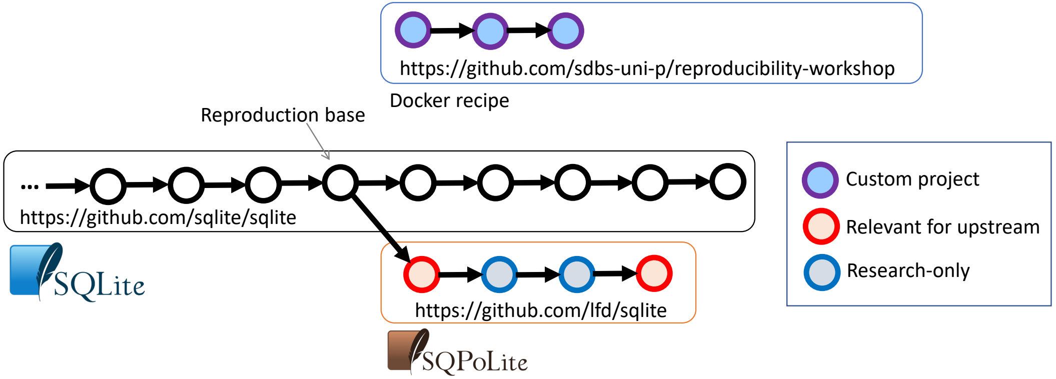


SEENG@ICSE 2022 paper “Beyond the Badge: Reproducibility Engineering as a Lifetime Skill.” (Mauerer, Klessinger, Scherzinger)

Dealing with different Target Platforms



Artefacts extended by Research

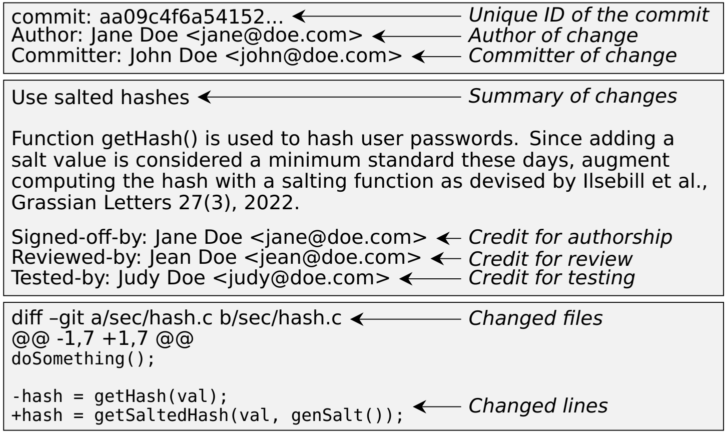

https://github.com/sdbs-uni-p/reproducibility-workshop
Docker recipe
Reproduction base
...
https://github.com/sqlite/sqlite
SQLite
https://github.com/lfd/sqlite
SQPoLite
Custom project
Relevant for upstream
Research-only

Creating Proper Histories I

Best Practices

- Presenting thoughts versus presenting results
- Rebasing, squashing, rewriting and all of that
- Self-documenting patches
- Trail of responsibility/lineage and provenance (developer certificate of origin)
- Upstream, integrate, externalise?

Creating Proper Histories II



```
commit: aa09c4f6a54152...
Author: Jane Doe <jane@doe.com>
Committer: John Doe <john@doe.com>

Use salted hashes

Function getHash() is used to hash user passwords. Since adding a
salt value is considered a minimum standard these days, augment
computing the hash with a salting function as devised by Ilsebill et al.,
Grassian Letters 27(3), 2022.

Signed-off-by: Jane Doe <jane@doe.com>
Reviewed-by: Jean Doe <jean@doe.com>
Tested-by: Judy Doe <judy@doe.com>

diff –git a/sec/hash.c b/sec/hash.c
@@ -1,7 +1,7 @@
doSomething();

-hash = getHash(val);
+hash = getSaltedHash(val, genSalt());
```

Licensing Consideration I

Steps

- Choosing a license
- Specifying a license
- SPDX identifiers

Licensing Consideration II

Proprietary, closed-source components

~~Just don't!!~~ Try not to use them.

Summary

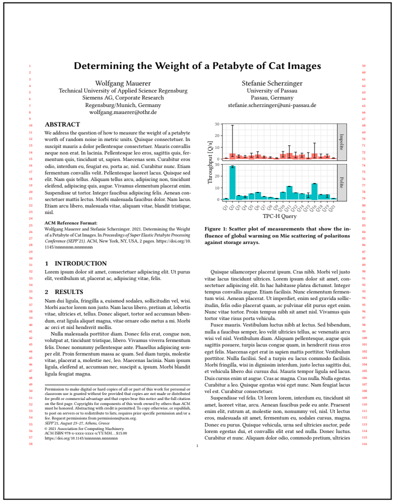
We have solved two problems:

- Our build is reproducible.
- Our results are reproducible.

Two more to go:

- We need to write a paper.
- We need to make our artifacts available.

End-to-end reproduction I



Determining the Weight of a Petabyte of Cat Images

Wolfgang Mauerer
Technical University of Applied Science Regensburg
Siemens AG, Corporate Research
Regensburg/Munich, Germany
wolfgang.mauerer@othr.de

Stefanie Scherzinger
University of Passau
Passau, Germany
stefanie.scherzinger@uni-passau.de

ABSTRACT

We address the question of how to measure the weight of a petabyte worth of random noise in metric units. Quisque consectetuer. In suscipit mauris a dolor pellentesque consectetuer. Mauris convallis neque non erat. In lacinia. Pellentesque leo eros, sagittis quis, fermentum quis, tincidunt ut, sapien. Maecenas sem. Curabitur eros odio, interdum eu, feugiat eu, porta ac, nisl. Curabitur nunc. Etiam fermentum convallis velit. Pellentesque laoreet lacus. Quisque sed elit. Nam quis tellus. Aliquam tellus arcu, adipiscing non, tincidunt eleifend, adipiscing quis, augue. Vivamus elementum placerat enim. Suspendisse ut tortor. Integer faucibus adipiscing felis. Aenean consectetuer mattis lectus. Morbi malesuada faucibus dolor. Nam lacus. Etiam arcu libero, malesuada vitae, aliquam vitae, blandit tristique, nisl.

ACM Reference Format:
Wolfgang Mauerer and Stefanie Scherzinger. 2021. Determining the Weight of a Petabyte of Cat Images. In *Proceedings of Super Elastic Petabyte Processing Conference (SEPP'21)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis.

2 RESULTS

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna.

Throughput [Q/s]
Impolite
Polite
Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q14 Q15 Q16 Q18 Q19 Q21
TPC-H Query

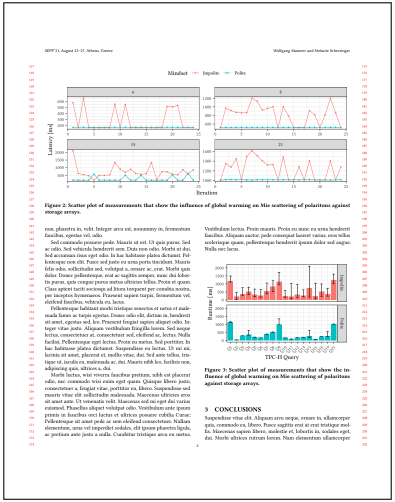
Figure 1: Scatter plot of measurements that show the influence of global warming on Mie scattering of polaritons against storage arrays.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies

1



SEPP'21, August 23–27, Athens, Greece — Wolfgang Mauerer and Stefanie Scherzinger

Figure 2: Scatter plot of measurements that show the influence of global warming on Mie scattering of polaritons against storage arrays.

non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus.

Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Runtime [ms]
Impolite
Polite
Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q14 Q15 Q16 Q18 Q19 Q21
TPC-H Query

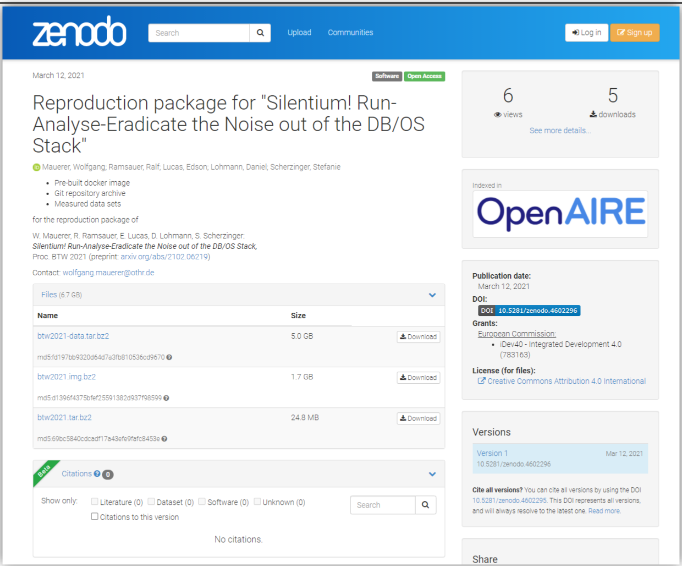
Figure 3: Scatter plot of measurements that show the influence of global warming on Mie scattering of polaritons against storage arrays.

3 CONCLUSIONS

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper

2

End-to-end Reproduction II



zenodo | Search | Upload | Communities | Log in | Sign up

March 12, 2021 — Software | Open Access

Reproduction package for "Silentium! Run-Analyse-Eradicate the Noise out of the DB/OS Stack"

Mauerer, Wolfgang; Ramsauer, Ralf; Lucas, Edson; Lohmann, Daniel; Scherzinger, Stefanie

- Pre-built docker image
- Git repository archive
- Measured data sets

for the reproduction package of

W. Mauerer, R. Ramsauer, E. Lucas, D. Lohmann, S. Scherzinger:
Silentium! Run-Analyse-Eradicate the Noise out of the DB/OS Stack,
Proc. BTW 2021 (preprint: arxiv.org/abs/2102.06219)

Contact: wolfgang.mauerer@othr.de

Files (6.7 GB)

Name	Size	
btw2021-data.tar.bz2 md5:fd197bb9320d64d7a3fb810536cd9670	5.0 GB	Download
btw2021.img.bz2 md5:d1396f4375bfef25591382d937f98599	1.7 GB	Download
btw2021.tar.bz2 md5:69bc5840cdcadf17a43efe9fafc8453e	24.8 MB	Download

Beta — Citations 0

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

No citations.

6 views — 5 downloads
See more details...

Indexed in OpenAIRE

Publication date:
March 12, 2021
DOI:
DOI 10.5281/zenodo.4602296
Grants:
European Commission:
- iDev40 - Integrated Development 4.0 (783163)

License (for files):
Creative Commons Attribution 4.0 International

Versions

Version 1 10.5281/zenodo.4602296	Mar 12, 2021

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.4602295. This DOI represents all versions, and will always resolve to the latest one. Read more.

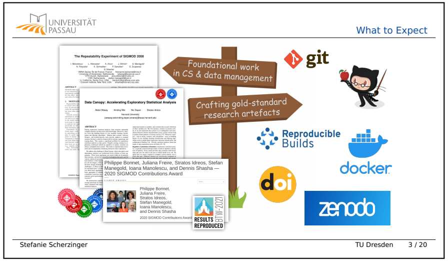
Share

What’s in it for you?

Benefits vs. Costs

- We won’t lie to you: Reproducibility engineering is a lot of work.
- Enablers to reproducibility: Testability & automation
- However, there are clear benefits for your team, just imagine:
  - Long-term, automation will save you time.
  - You can build your stuff even after the responsible PhD student has graduated.
  - You can switch to a new notebook without risking your progress in research.
  - You will no longer have to negotiate within the team why s.th. works for you, but only you.

Summary

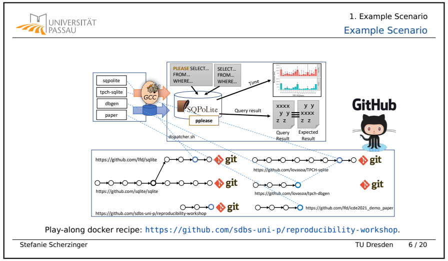


UNIVERSITÄT PASSAU

What to Expect

The Repeatability Experiment of SIGMOD 2008
Data Canopy: Accelerating Exploratory Statistical Analysis
Philippe Bonnet, Juliana Freire, Stratos Idreos, Stefan Manegold, Ioana Manolescu, and Dennis Shasha — 2020 SIGMOD Contributions Award
Philippe Bonnet, Juliana Freire, Stratos Idreos, Stefan Manegold, Ioana Manolescu, and Dennis Shasha
2020 SIGMOD Contributions Award
RESULTS REPRODUCED
BTW-2021
Foundational work in CS & data management
Crafting gold-standard research artefacts
git
Reproducible Builds
docker
doi
zenodo

Stefanie Scherzinger TU Dresden 3 / 20



UNIVERSITÄT PASSAU

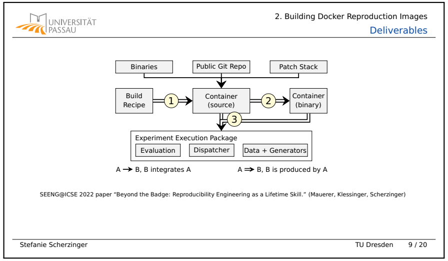
1. Example Scenario

Example Scenario

sqpolite
tpch-sqlite
dbgen
paper
GCC
PLEASE SELECT... FROM... WHERE...
SELECT... FROM... WHERE...
SQPoLite
pplease
Time
Query result
xxxx y y z z
y y xxxx z z
Query Result
Expected Result
dispatcher.sh
GitHub
https://github.com/lfd/sqlite
git
https://github.com/sqlite/sqlite
https://github.com/sdbs-uni-p/reproducibility-workshop
https://github.com/lovasoa/TPCH-sqlite
https://github.com/lovasoa/tpch-dbgen
https://github.com/lfd/icde2021_demo_paper

Play-along docker recipe: `https://github.com/sdbs-uni-p/reproducibility-workshop`.

Stefanie Scherzinger TU Dresden 6 / 20



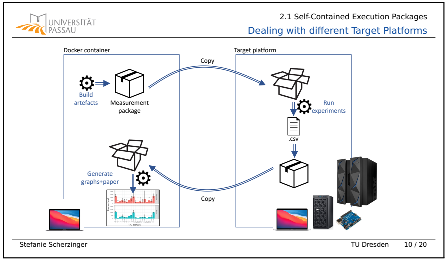
UNIVERSITÄT PASSAU

2. Building Docker Reproduction Images

Deliverables

SEENG@ICSE 2022 paper “Beyond the Badge: Reproducibility Engineering as a Lifetime Skill.” (Mauerer, Klessinger, Scherzinger)

Stefanie Scherzinger TU Dresden 9 / 20



UNIVERSITÄT PASSAU

2.1 Self-Contained Execution Packages

Dealing with different Target Platforms

Stefanie Scherzinger TU Dresden 10 / 20