Medical Imaging Research Laboratory
www.creatis.insa-lyon.fr

Lessons learnt with ReproVIP

Sorina Camarasu-Pop | CREATIS, CNRS (UMR 5220), INSERM (U1294), INSA Lyon, Université de Lyon, France

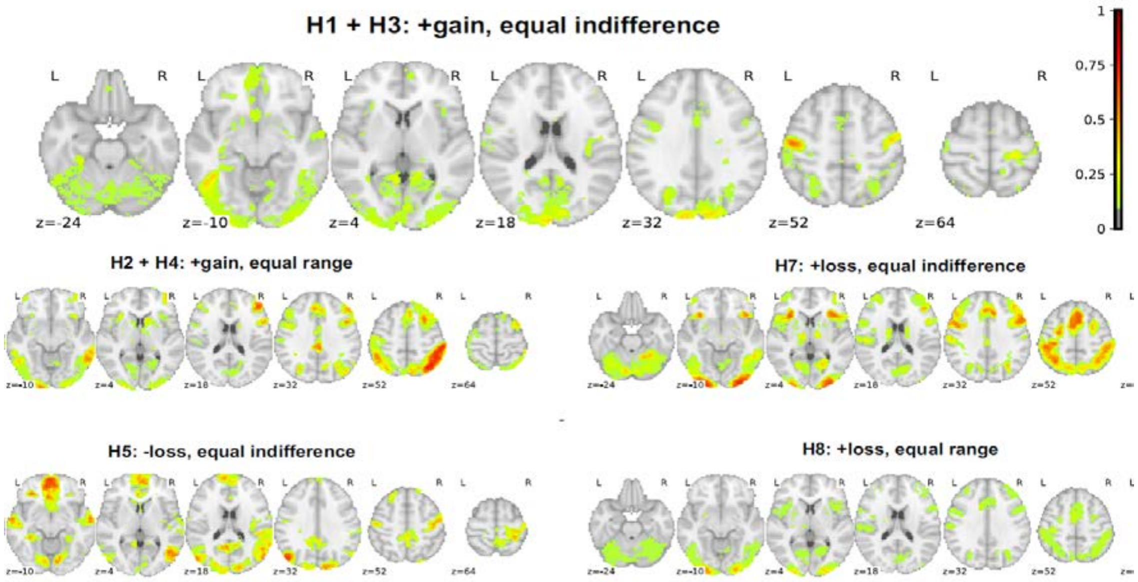
Reproducibility Issues in Neuroimaging

Setup

- **1** dataset
- **70** teams
- **9** hypotheses

Findings

- Analytical flexibility
- Variability of results
- Optimism bias



H1 + H3: +gain, equal indifference

H2 + H4: +gain, equal range

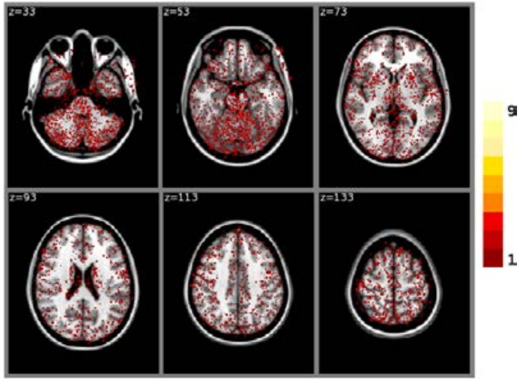
H7: +loss, equal indifference

H5: -loss, equal indifference

H8: +loss, equal range

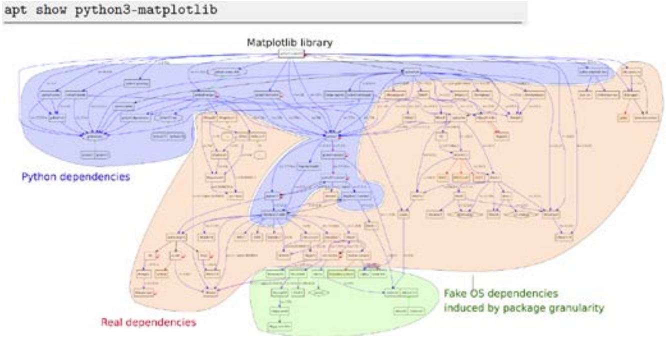
R. Botvinik-Nezer *et al*, “Variability in the analysis of a single neuroimaging dataset by many teams” *Nature* 2020

Computational reproducibility issues

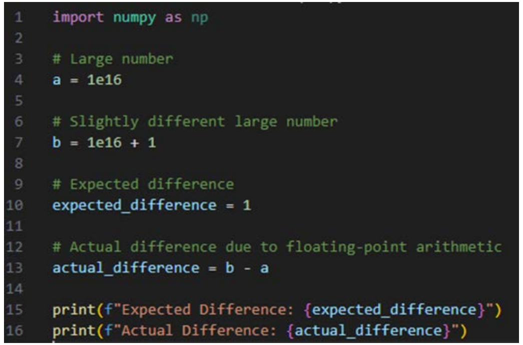


expf(1.540518522262573242187500000000)
=4.6670093536376953125000 (glibc 2.5)
expf(1.540518522262573242187500000000)
=4.6670098304748535156250 (glibc 2.18)

Same FSL version (5.0.6) and different versions of GNU/Linux
Sum of binarized differences between cortical tissue classifications obtained on cluster A (CentOS) and cluster B (Fedora) (FSL FAST, build 1, *n* = 150 subjects). Credits: Tristan Glatard, https://www.frontiersin.org/articles/10.3389/fninf.2015.00012/full



Complex dependencies. Credits: Arnaud Legrand

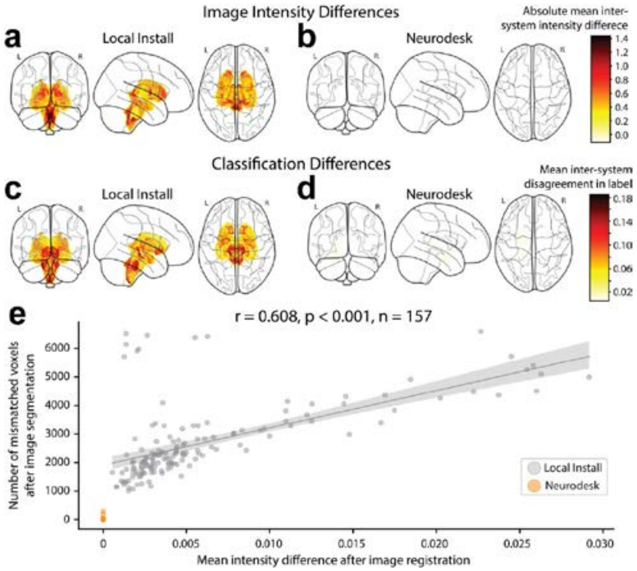


```
import numpy as np

# Large number
a = 1e16

# Slightly different large number
b = 1e16 + 1

# Expected difference
expected_difference = 1

# Actual difference due to floating-point arithmetic
actual_difference = b - a

print(f"Expected Difference: {expected_difference}")
print(f"Actual Difference: {actual_difference}")
```

Floating point arithmetic: rounding errors

Computational reproducibility

- Main causes
  - Software dependencies and their evolution
  - Numerical instability due to floating point arithmetic
- Containerization
  - Package and run an application and its dependencies
- Guix
  - Functional package manager
  - Reproducible computational environments



A.I. Renton *et al*, *Neurodesk Nature 2024*

ReproVIP

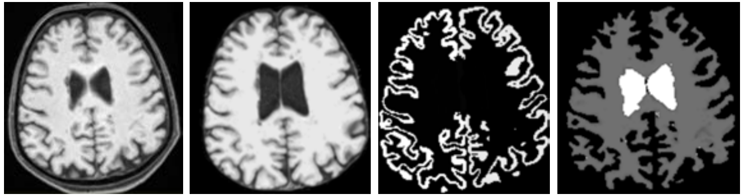
- French ANR JCJC project
  - Partners: CREATIS, IPHC, Concordia University
  - https://reprovipgroup.pages.in2p3.fr/documentation

- Main objectives
  - Evaluate and improve the **reproducibility** of scientific results: **same result when the code is executed with the same set of inputs**
  - Provide an **integrated, end to end solution**, allowing to launch reproducible executions in a transparent manner
  - Evaluate the proposed methods and tools on medical imaging studies

The Virtual Imaging Platform

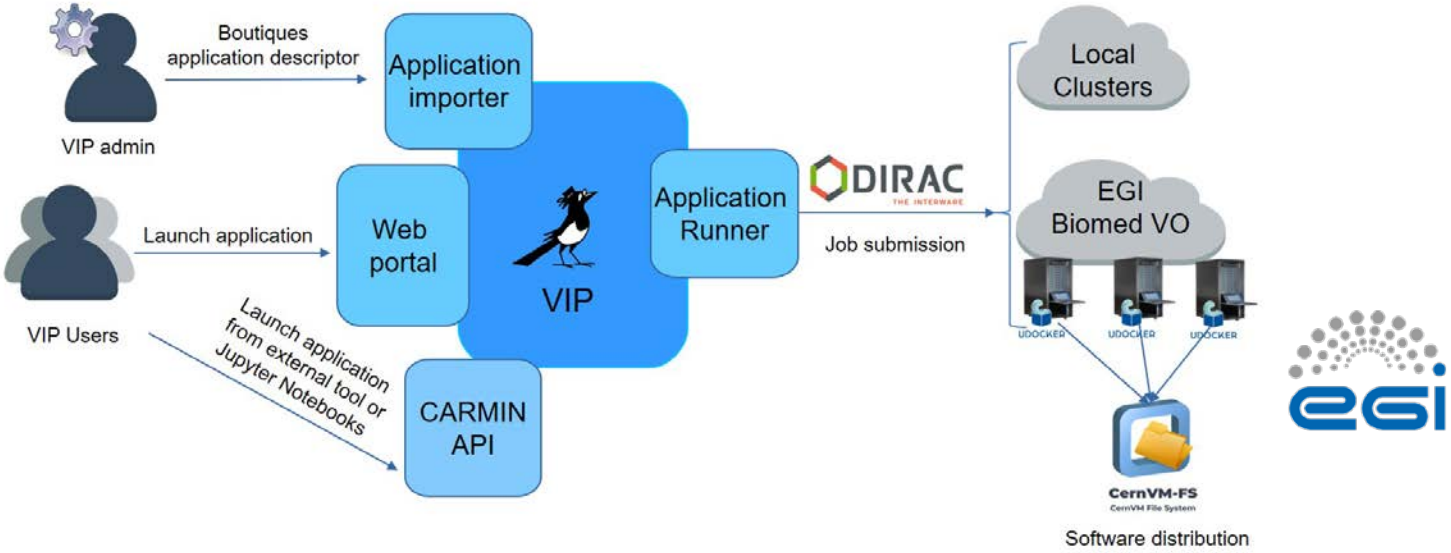
- Scientific applications as a Service
  - More than 25 applications publicly available
  - https://vip.creatis.insa-lyon.fr/home.html
- Transparent access to computing resources
  - 110+ CPU years (EGI biomed VO) used in the last 12 months
  - 77 publications with results obtained on VIP
- Large community
  - More than 1500 registered users
- Open and reproducible science
  - Zenodo, DOIs, Containers, Boutiques

https://vip.creatis.insa-lyon.fr



Example of white/grey matter brain segmentation with Freesurfer on VIP
Credits : Berardino Barile and Dominique Sappey-Marinier, Creatis

VIP eco-system

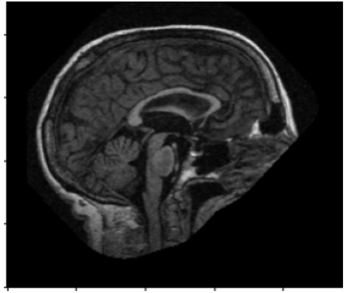


The Impact Of Hardware Variability

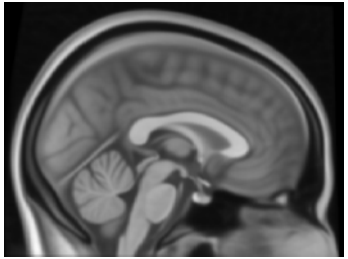
- The Impact Of Hardware Variability On Applications Packaged With Docker And Guix: A Case Study In Neuroimaging
  - ACM REP'24 Best Paper award ☺
  - https://hal.science/hal-04480308v2

- Objectives
  - Evaluate the impact of hardware variability
  - Compare and correlate hardware variability to
    - Software variability encountered in different software packages
    - Numerical variability resulting from Monte Carlo Arithmetic (MCA) Random Rounding (RR)

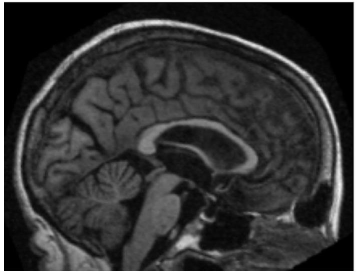
FSL FLIRT



original image

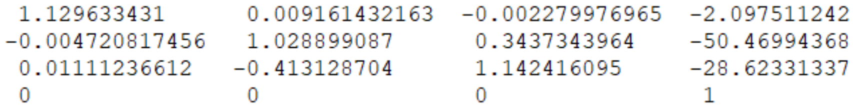


reference



registered image

- FMRIB Software Library ([FSL](FSL))
  - Library of analysis tools for FMRI, MRI and diffusion brain imaging data
- FLIRT: FMRIB's Linear Image Registration Tool
  - Affine brain registration: align a brain scan with another one through rotation, translation, scaling and shearing
- FLIRT outputs
  - Registered brain image in NIfTI format (.nii.gz)
  - Transformation matrix in text format (.mat)

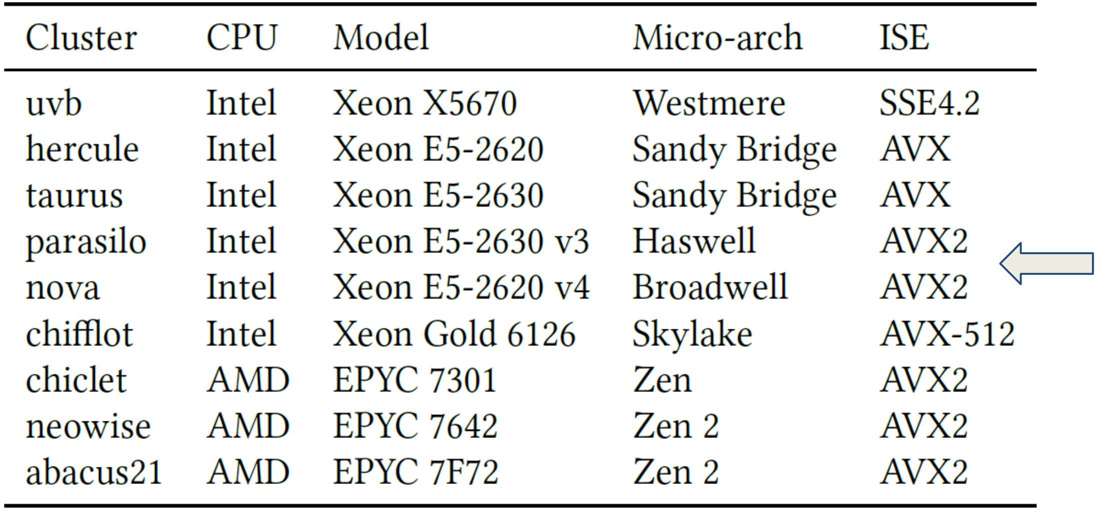


```
 1.129633431       0.009161432163  -0.002279976965  -2.097511242
-0.004720817456    1.028899087      0.3437343964    -50.46994368
 0.01111236612    -0.413128704      1.142416095     -28.62331337
 0                 0                0                 1
```

Example of transformation matrix (.mat file)

Grid'5000 research infrastructure

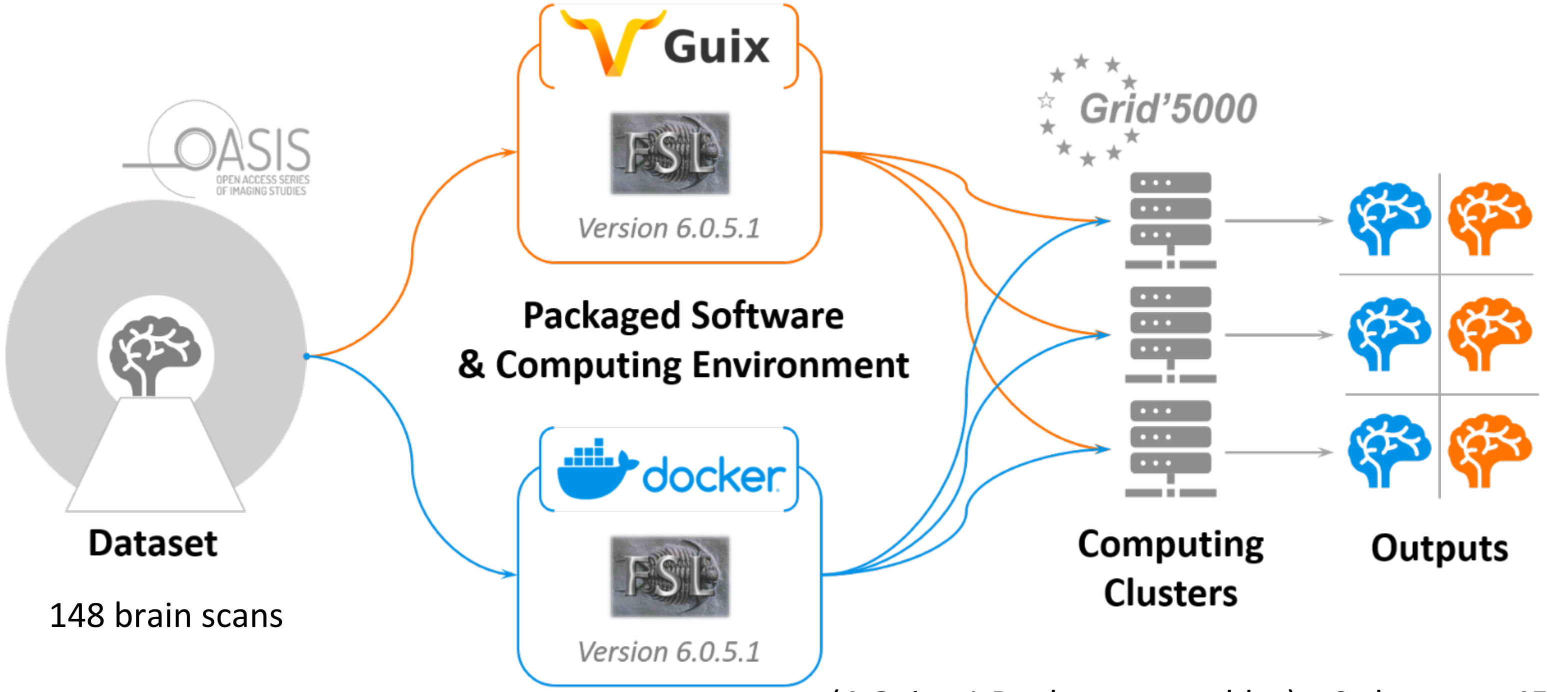
- Large-scale testbed for experiment-driven research in computer science
- Access to a wide spectrum of hardware



Cluster	CPU	Model	Micro-arch	ISE
uvb	Intel	Xeon X5670	Westmere	SSE4.2
hercule	Intel	Xeon E5-2620	Sandy Bridge	AVX
taurus	Intel	Xeon E5-2630	Sandy Bridge	AVX
parasilo	Intel	Xeon E5-2630 v3	Haswell	AVX2
nova	Intel	Xeon E5-2620 v4	Broadwell	AVX2
chifflot	Intel	Xeon Gold 6126	Skylake	AVX-512
chiclet	AMD	EPYC 7301	Zen	AVX2
neowise	AMD	EPYC 7642	Zen 2	AVX2
abacus21	AMD	EPYC 7F72	Zen 2	AVX2

Fused Multiply-Add (FMA)

Overview of experiments on Grid’5000



Dataset

148 brain scans

Packaged Software
& Computing Environment

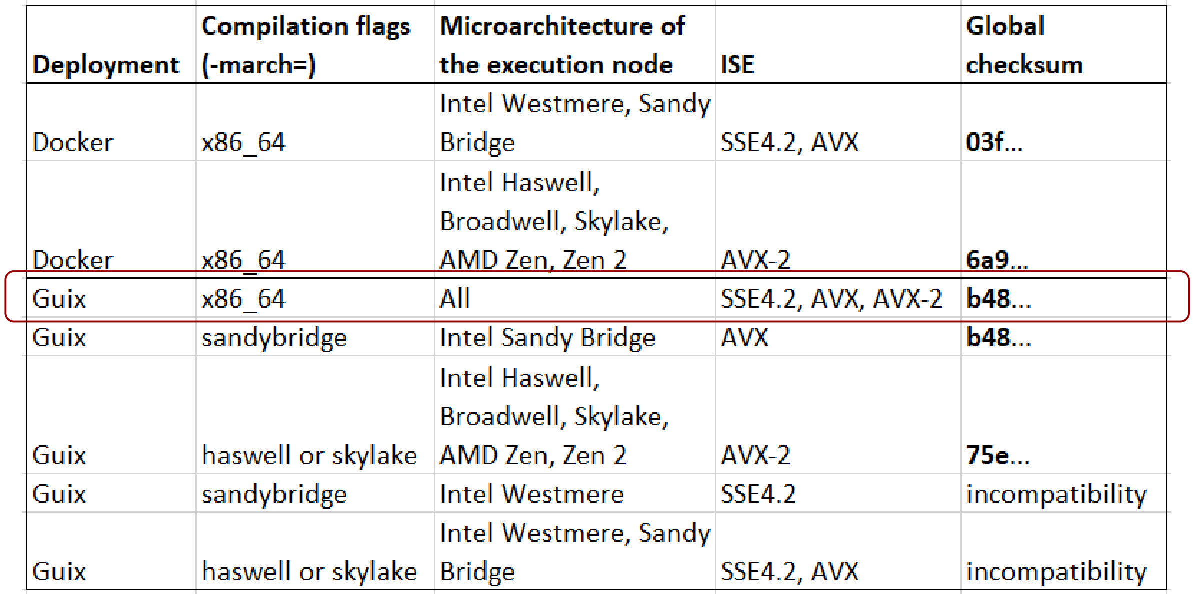
Computing Clusters

Outputs

(4 Guix +1 Docker executables) x 9 clusters = 45 experiments

Hardware variability

- Comparison of global checksums
  - tarball of the 148 results for each one of the 45 experiments



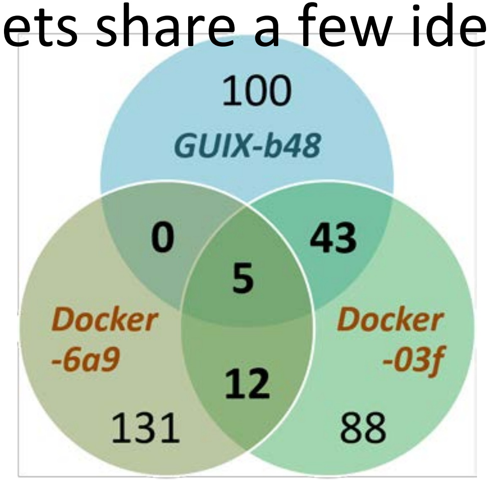
Deployment	Compilation flags (-march=)	Microarchitecture of the execution node	ISE	Global checksum
Docker	x86_64	Intel Westmere, Sandy Bridge	SSE4.2, AVX	**03f…**
Docker	x86_64	Intel Haswell, Broadwell, Skylake, AMD Zen, Zen 2	AVX-2	**6a9…**
Guix	x86_64	All	SSE4.2, AVX, AVX-2	**b48…**
Guix	sandybridge	Intel Sandy Bridge	AVX	**b48…**
Guix	haswell or skylake	Intel Haswell, Broadwell, Skylake, AMD Zen, Zen 2	AVX-2	**75e…**
Guix	sandybridge	Intel Westmere	SSE4.2	incompatibility
Guix	haswell or skylake	Intel Westmere, Sandy Bridge	SSE4.2, AVX	incompatibility

Four different global checksums

Two micro-architecture subsets: with and without AVX-2

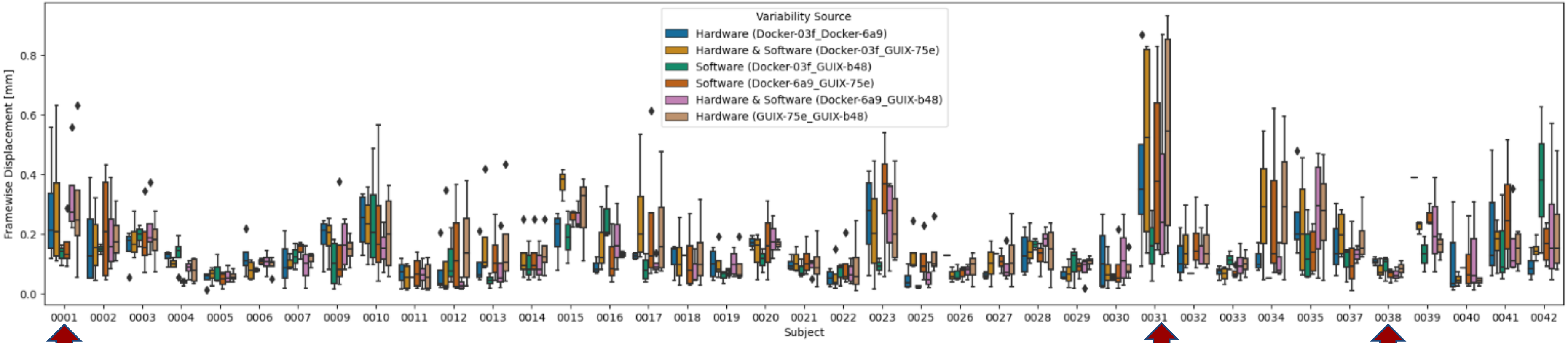
Variability depends on input data

- Comparison of the 148 individual results among the four sets of results
- Three of the four sets share a few identical results



Intersections between result sets (individual matrix files) for three of the four experiments.

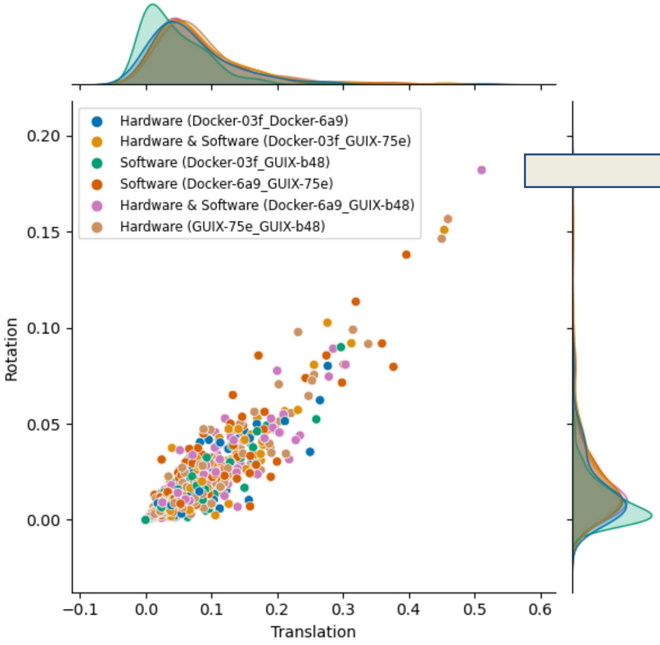
Variability across subjects



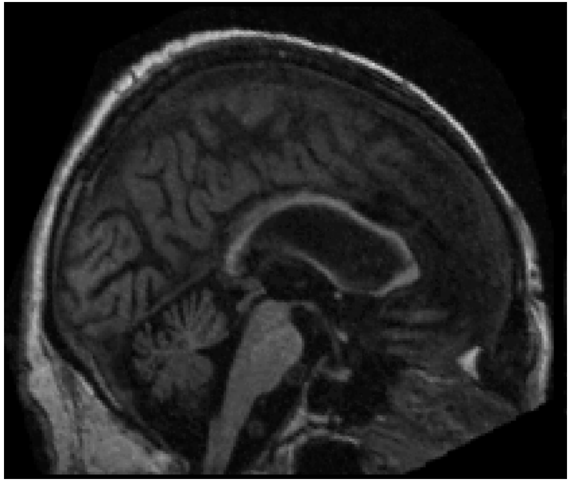
Framewise displacement across subjects

=> importance of using large image databases

Effects on the registration



Distributions of rotation and translation differences in the transformation matrix results ('.mat' result files)

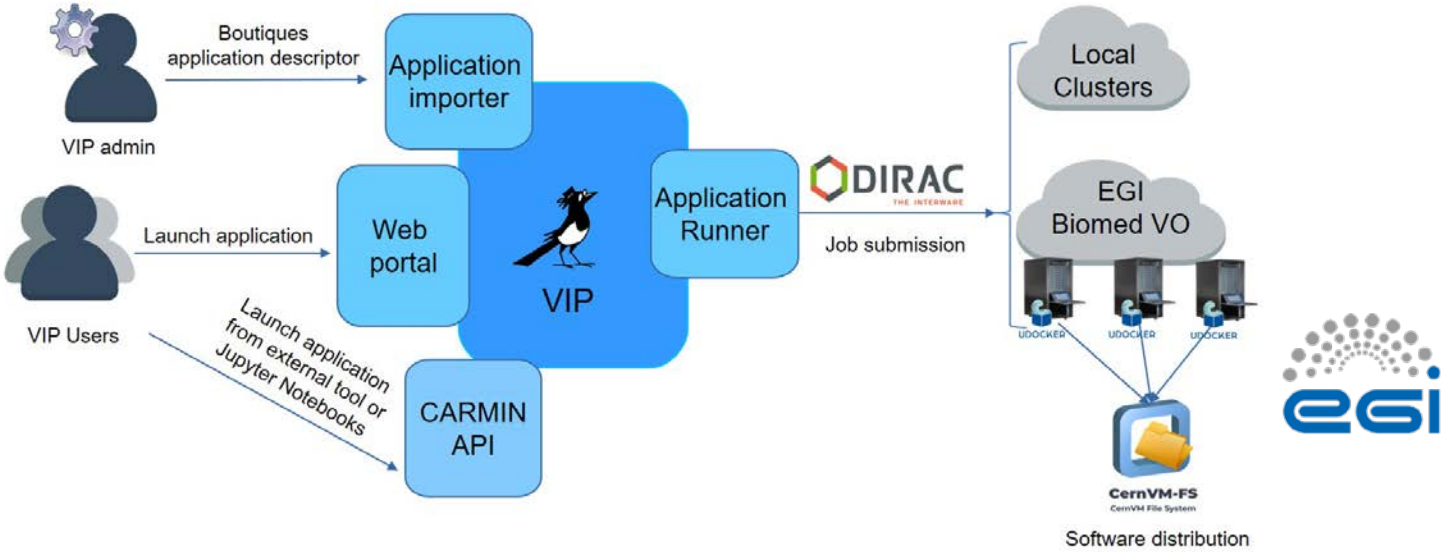


Differences between outputs (belonging to groups Docker-6a9 and Guix-b48) with the largest difference in translation and rotation (subject 31, scan 2)

Study conclusions

- Hardware, software and numerical variability lead to variations
  - of similar magnitudes but
  - uncorrelated with each other
- Variations remained moderate but might impact downstream analyses
- In our case hardware variability was due to AVX-2 support
  - Further work is needed for a finer analysis of the differences observed
- Both packaging solutions (Docker and Guix) produced
  - Each one bit-wise reproducible results when using the same packaged FLIRT executable on equivalent micro-architectures
  - Different outputs from one another due to the software variability

Back to production

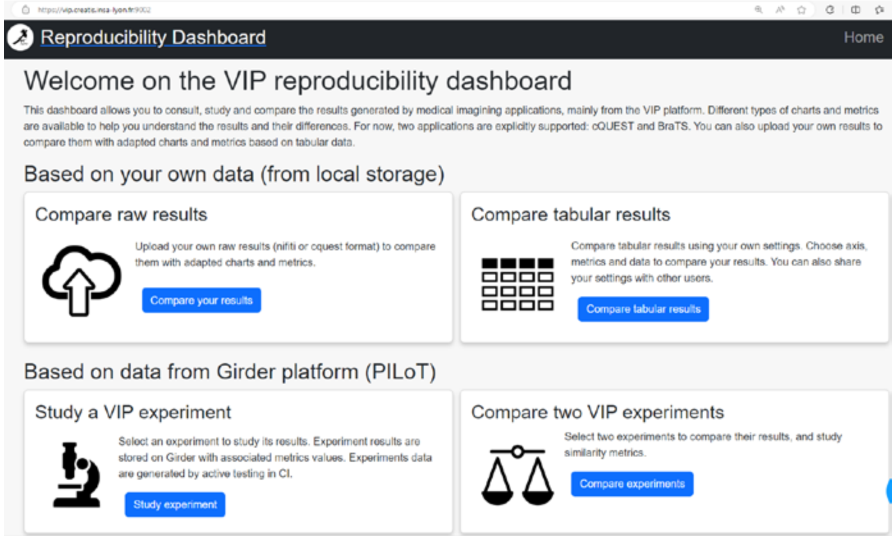


Back to production

- Use of CVMFS for software distribution
  - CVMFS uploader maintained by RAL
- Tests with FSL Guix modules (relocatable packages options)
  - Guix build sever on a VM on the SCIGNE infrastructure
  - CVMFS test server in a VM on the SCIGNE infrastructure
    - Large number of files
- Current Guix-CVMFS based solution available in VIP in test mode only

Integrated end to end solution

- VIP portal
  - Applications as a service
  - Execution sharing (Zenodo)
- Automation
  - Jupyter Notebooks (templates)
  - Python client, REST API
- Reproducibility Dashboard
  - https://vip.creatis.insa-lyon.fr:9002
- Continuous Integration (CI)
- Integration with storage platforms
  - Girder, Shanoir



Welcome on the VIP reproducibility dashboard

This dashboard allows you to consult, study and compare the results generated by medical imagining applications, mainly from the VIP platform. Different types of charts and metrics are available to help you understand the results and their differences. For now, two applications are explicitly supported: cQUEST and BraTS. You can also upload your own results to compare them with adapted charts and metrics based on tabular data.

Based on your own data (from local storage)

Compare raw results

Upload your own raw results (nifiti or cquest format) to compare them with adapted charts and metrics.

Compare tabular results

Compare tabular results using your own settings. Choose axis, metrics and data to compare your results. You can also share your settings with other users.

Based on data from Girder platform (PILoT)

Study a VIP experiment

Select an experiment to study its results. Experiment results are stored on Girder with associated metrics values. Experiments data are generated by active testing in CI.

Compare two VIP experiments

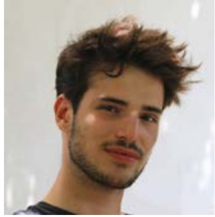
Select two experiments to compare their results, and study similarity metrics.

ReproVIP reproducibility dashboard

Conclusions

- Computational reproducibility
  - Challenging and often over-looked
  - Various, possibly complex solutions
- VIP provides an integrated, end-to-end solution for reproducible executions of scientific applications available in VIP
  - Playground for reproducible experiments
- Reproducible and generalisable software solutions
  - Computational reproducibility is only a small aspect of a larger issue

Acknowledgements

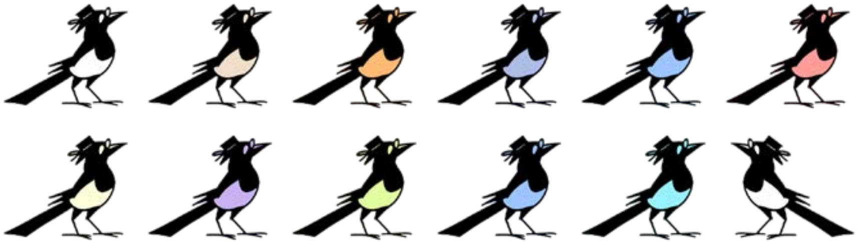


CREATIS

UNIVERSITÉ Concordia UNIVERSITY

iPHC Institut Pluridisciplinaire Hubert CURIEN STRASBOURG

ReproVIP anr



Thank you for your attention!
Questions?

ReproVIP