ActivePapers: a platform for performing and publishing reproducible research

Konrad HINSEN

Centre de Biophysique Moléculaire, Orléans, France
and
Synchrotron SOLEIL, Saint Aubin, France

February 14, 2017

ActivePapers is. . .

- a research project on the integration of computer-aided research into the scientific record
- two software tools developed as a proof of concept
- a set of data storage conventions used by these tools

ActivePapers is **not**

- an easy-to-use tool vying for popularity
- a proposed standard looking for adopters

For more information:

- project Web site
- paper drawing first conclusions

Scientific publications

Traditional research

Communicate ideas, observations, hypotheses, deductions for

- **critical examination** to generate **trust**
- **reuse** to foster **progress**
- **credit** to build one's **reputation**

Computer-aided research

Communicate software and electronic datasets

- How do we generate **trust** in a computation?
- How do we enable **reuse** of code and data?
- How do we give **credit** to software authors?

Trust

To err is human (and sometimes they even cheat)

- Software has bugs (and sometimes hidden features).
- People make mistakes when using software.
- Users don't know/understand what their software does.

A societal issue beyond science

- Do you trust your smartphone not to spy on you?
- Do you trust your car not to cheat with emission control?
- Did you trust Google not to be evil (ending October 2015)?
- Do you trust Amazon to keep credit card info safe from hackers?

The technological Stockholm syndrome

We are so dependent on computing technology that we close our eyes to the trust issues that stem from its complexity.

Trust-generating measures

- Independent reimplementation → ReScience
- Transparency: publish all code and all data
- Quality control: version control, unit tests, . . .
- More understandable code: notebooks, literal programming
- Re-use trusted components.

All these measures act on the **human** face of software.

This is not enough!

What does this program do?

data_analysis.py

```
from datalib import Dataset

points = [(1, 1), (-1, 1), (2, 4)]

data = Dataset()
for x, y in points:
    if x > 0:
        data.add_value(y)
print(data.average())
```

Quick answer:

Libraries *and languages* matter

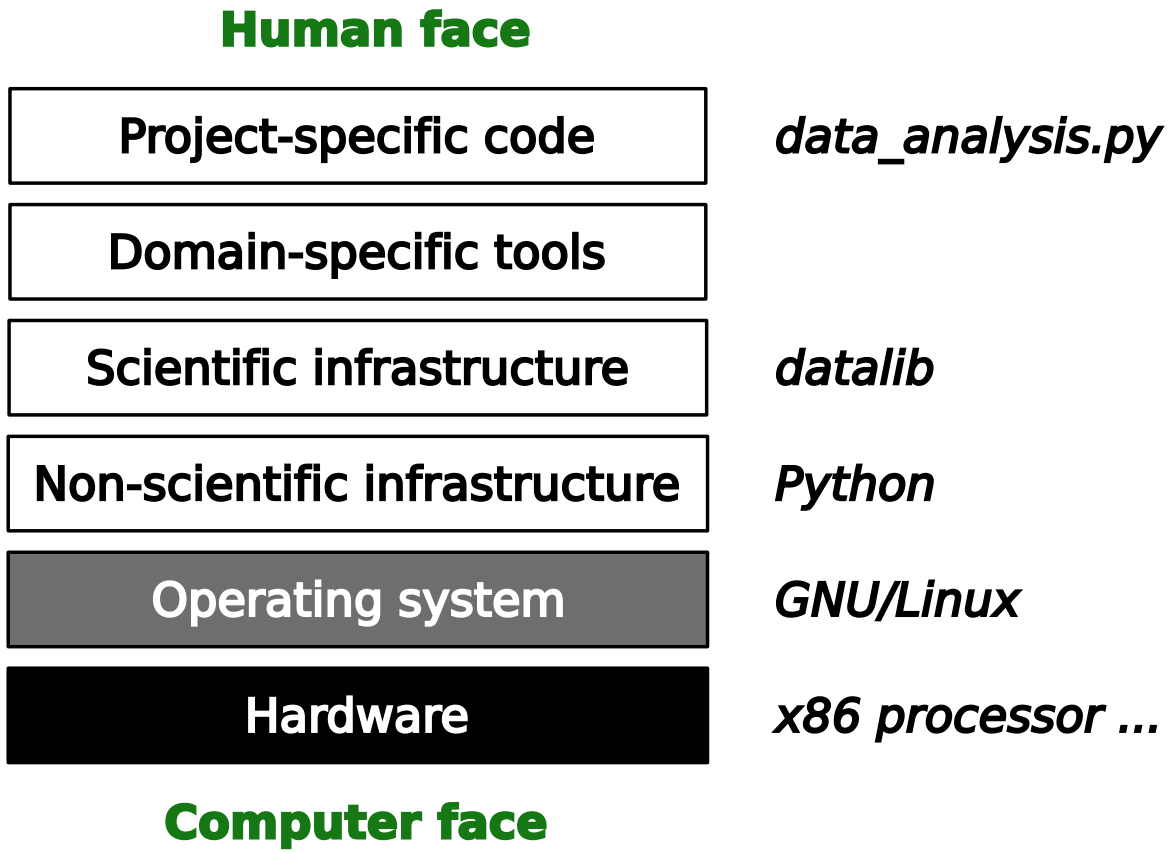
datalib.py

```
class Dataset(object):

    def __init__(self):
        self.values = []

    def add_value(self, value):
        self.values = [value]

    def average(self):
        return sum(self.values, 0)/len(self.values)
```

There's a bug! `add_value` keeps only the last value of each dataset!

So the result of `data_analysis.py` is 4. More precisely: it's 4 in Python 2 but 4.0 in Python 3.

The scientific software stack

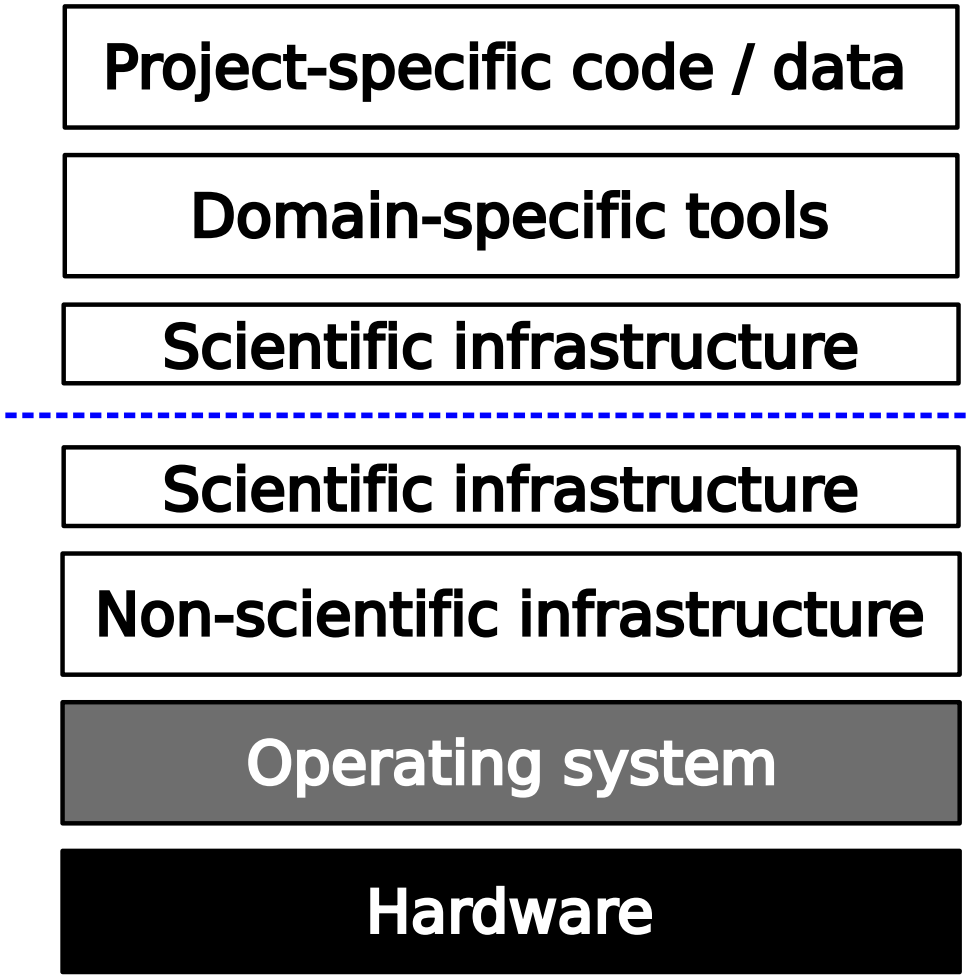


The meaning of each layer is defined by **all the layers below it**.

Full reproducibility requires publishing all of them.

Even the hardware!

The ActivePapers platform – content model



Content:
published

Trust:
inspection,
quality control

Interface:
stable

Platform:
professionally
maintained

Trust:
crowd testing,
quality control

Digital media for science

	Platform	Interface	Content
Paper	PDF reader LaTeX Word ...	PDF format	PDF files
Video	MP4 player Camera Video editor ...	MP4 format	MP4 files
Computation	Data inspector Code editor Validator ...	ActivePapers format	ActivePapers

ActivePapers history

Idealist phase (2010-2011)

- Do the best possible job with available technology.
- ... even if this makes it difficult to use.
- ActivePapers JVM edition
- Finalist in the Executable Papers Challenge at ICCS 2011

Pragmatist phase (2012-)

- Compromises to make it usable with today's software.
- Priority: biomolecular simulations (my field of research)
- ActivePapers Python edition

Full story: K. Hinsen, F1000Research 2015 **3** 289

The ActivePapers platform

Data container

HDF5

- efficient for big datasets (binary)
- attributes facilitate metadata storage

Code execution

Java Virtual Machine

- long-term stability
- secure execution
- good performance

Python+NumPy

- popular in science

References / links

DOI + HDF5 path

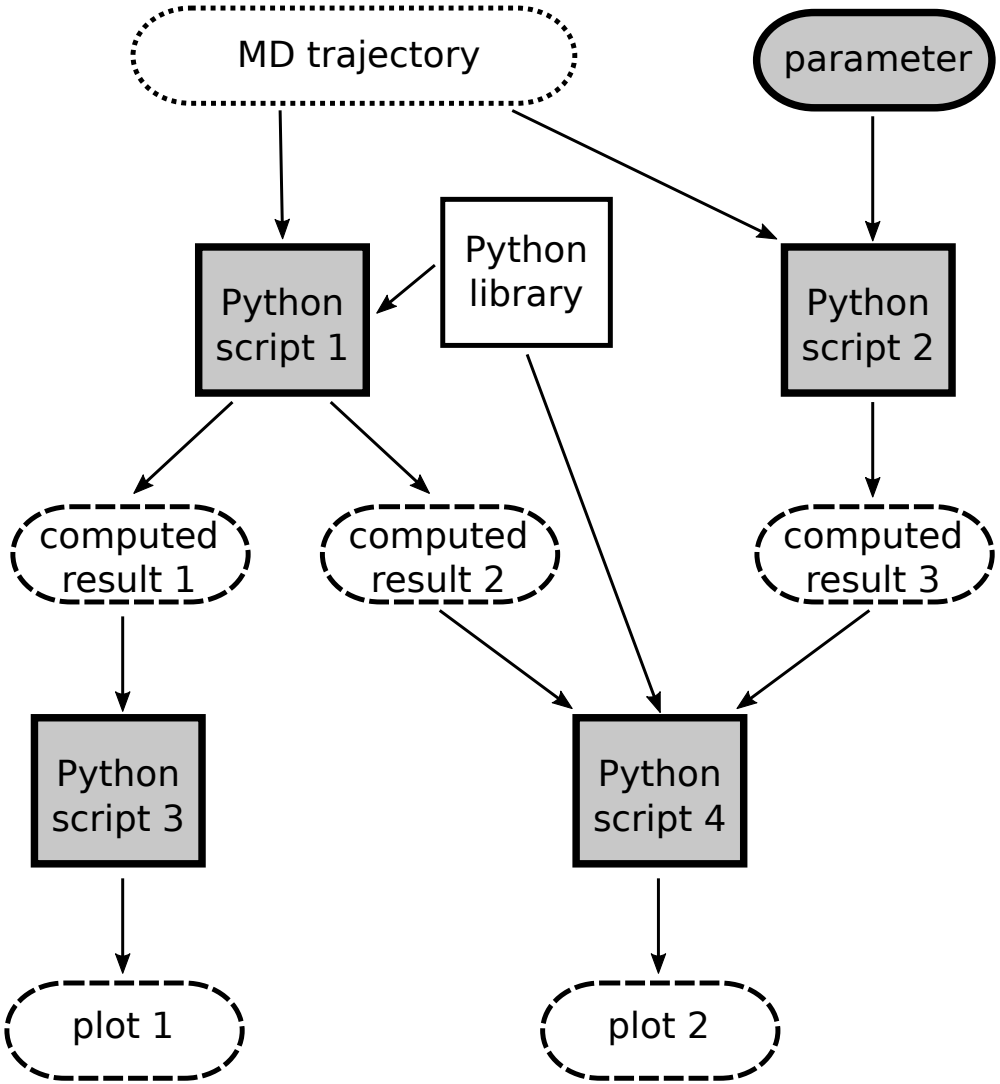
- works with existing citation mechanisms, bibliometry, ...

ActivePapers in bibliometry

Use = citation

for software and data

Inside an ActivePaper



ActivePapers in practice

- Used in five research projects
- 12 ActivePapers published on Zenodo
- 5 ActivePapers published on figshare
- Two types of published ActivePapers:
  - Software libraries
  - Data plus scripts

Demo time!