Analysis Reproducibility in Data (and Software) Dependent Research

What is reproducibility?

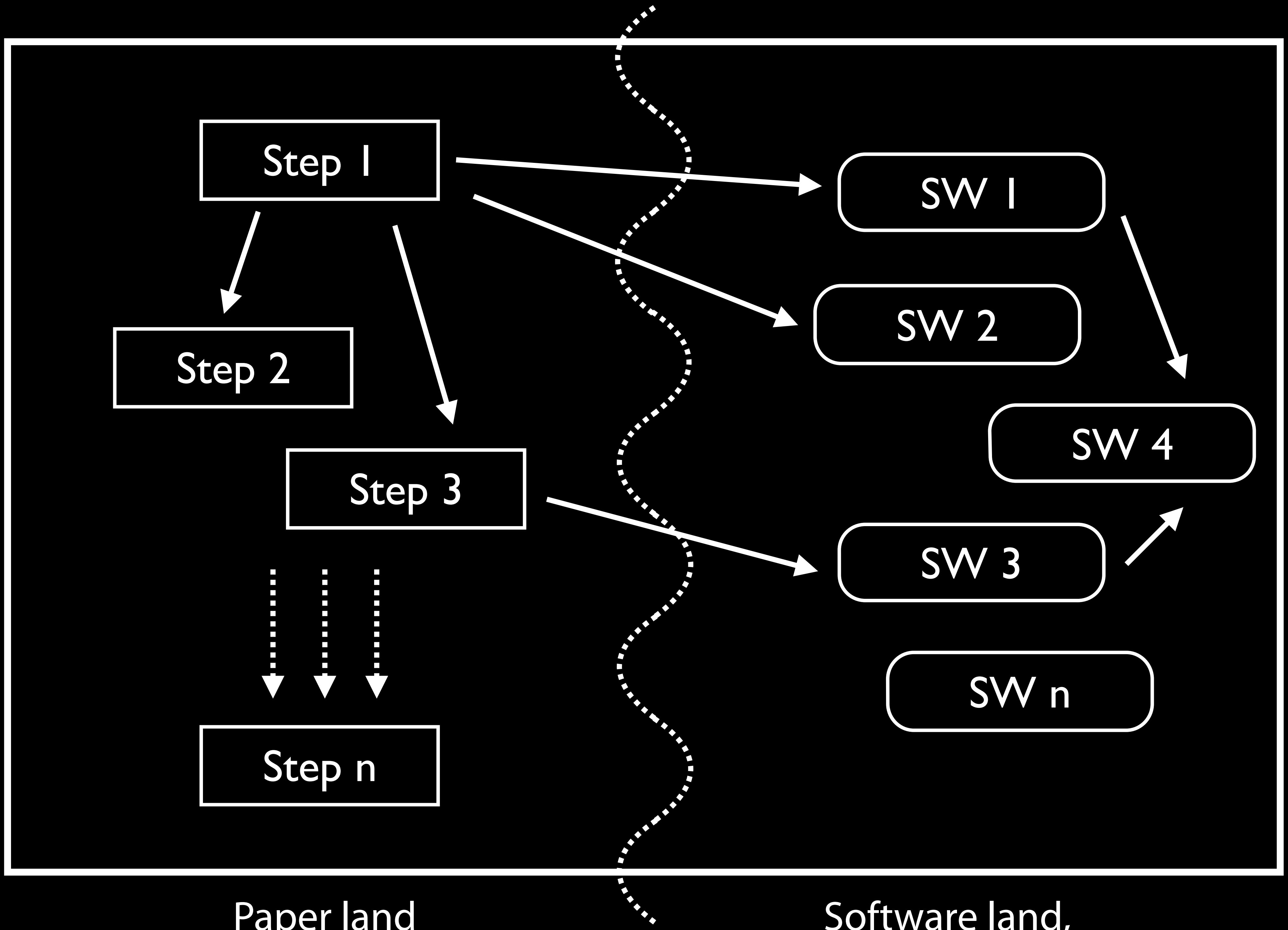
Provenance *is not* reproducibility

Reproducibility *is not* reusability

Reproducibility *is certainly not* correctness

Reproducibility means that an analysis is described in sufficient detail that it *can* be precisely reproduced

(by another person, in another environment)



Paper land
one shot analyses

Software land,
reusable components

Core reproducibility tasks

1. Capture the precise description of the experiment (either as it is being carried out, or after the fact)

2. Assemble all of the necessary data and software dependencies needed by the described experiment

3. Combine the above to verify the analysis

Most published analysis are not reproducible.

Missing software, versions, parameters (even data)

Recommendations

1. Accept that computation is an integral component of biomedical research

2. Always provide access to raw primary data

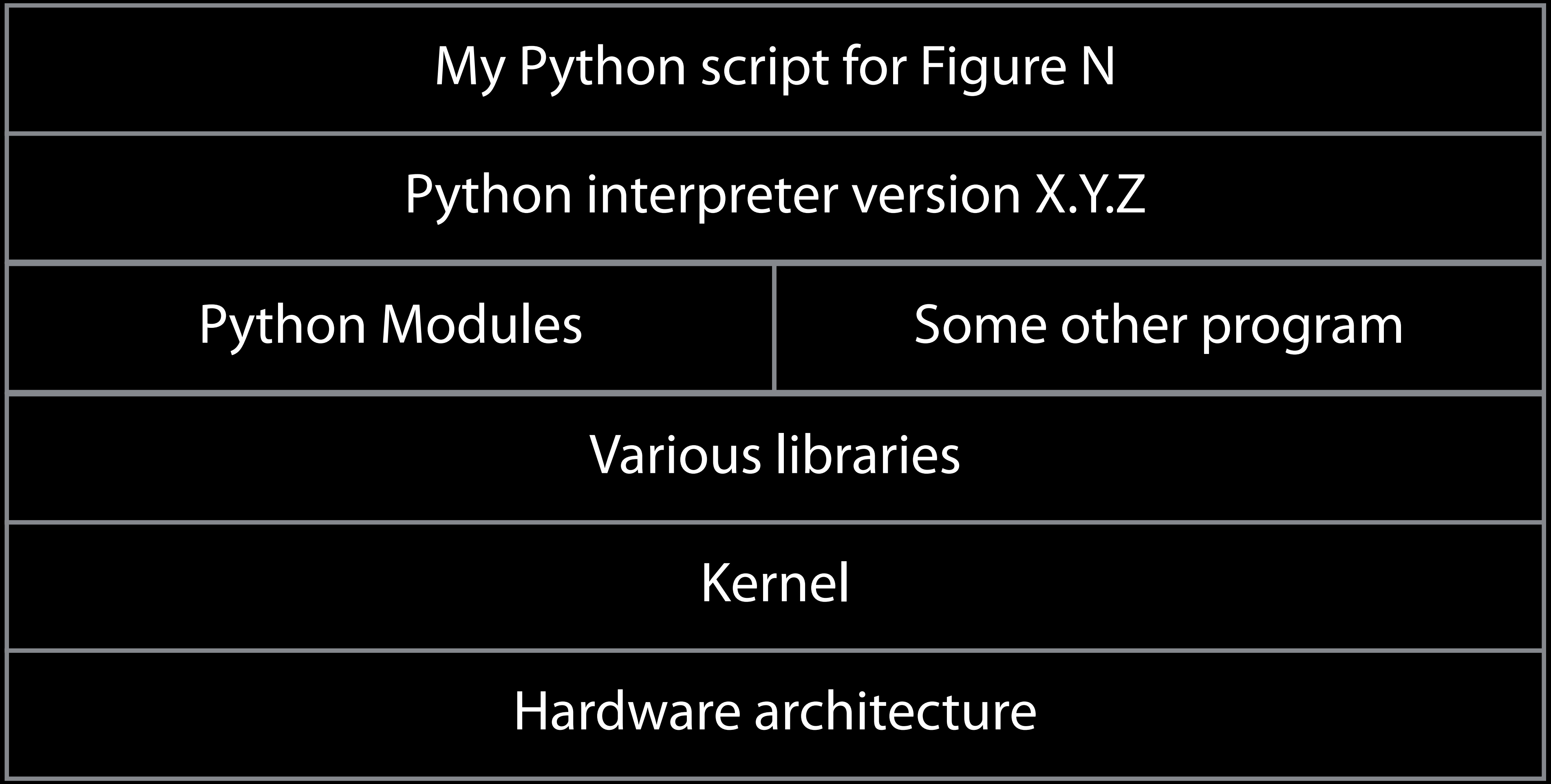
3. Record versions of all auxiliary datasets, or archive

4. Store the exact versions of *all* software used. Ideally archive the software

5. Record *all* parameters, even if default values are used.

(Abridged from Nekrutenko and Taylor, *Nature Reviews Genetics,* 2012)

How far down the stack is it realistic to go?



My Python script for Figure N

Python interpreter version X.Y.Z

Python Modules

Some other program

Various libraries

Kernel

Hardware architecture

Is reproducibility really a technical problem?

A spectrum of solutions

Analysis environments (Galaxy, GenePattern, Mobyle, …)

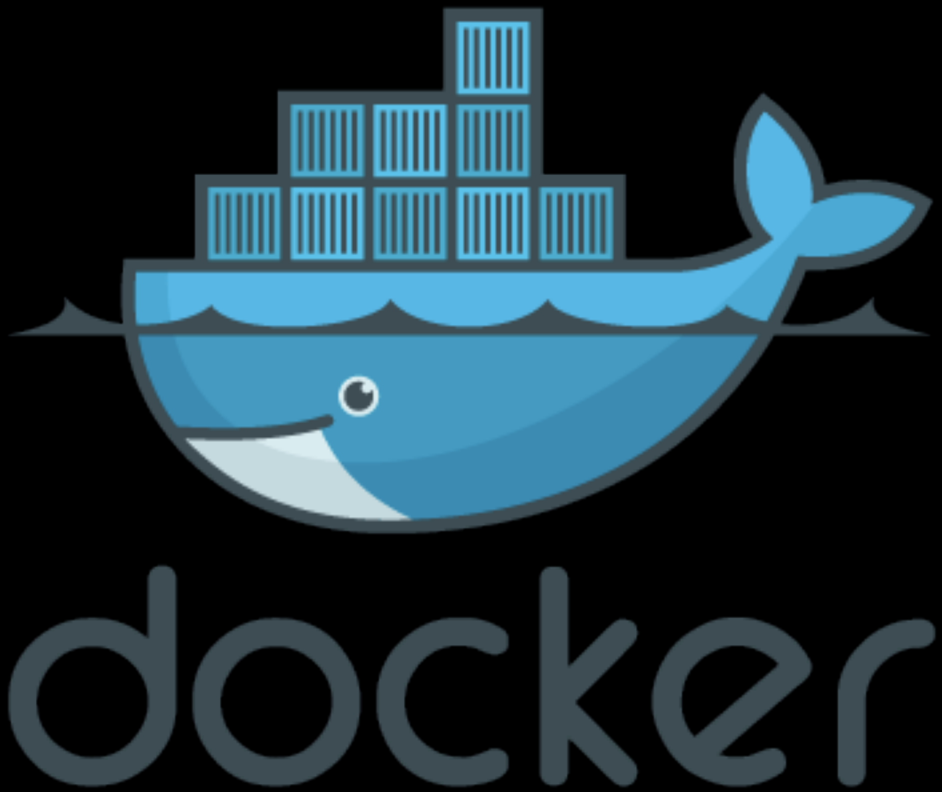
Workflow systems (Taverna, Pegasus, VisTrails, …)

Notebook style (iPython notebook, …)

Literate programming style (Sweave/knitR, …)

System level provenance capture (ReproZip, …)

Complete environment capture (VMs, containers, …)


docker

We have the technology!

Complete precise reproducibility IS POSSIBLE

Why are we not seeing widespread adoption?

Many approaches to reproducibility appropriate for different types of analysis and domains

However all these solutions have some barriers, either through constraining the user to ensure reproducibility or requiring complex packaging procedures after the fact

Ideally we should make reproducibility the norm for all analysis

Capture the description of the experiment transparently during analysis, rather than assembling after the fact

Easier for the analyst, and allows capturing the true workflow rather than just an idealized version

Can this be done without adding substantial barriers or constraints?

Final thoughts

Reproducibility requires archives, version capture and discovery are insufficient for long-term reliability

Licenses that do not allow archiving are… a problem

Reproducibility alone is not enough, it needs to be easy — if we are going to create an expectation of reproducibility it *must* be easy to validate at the peer review stage