

Repeatable and reproducible workflows using the RENKU open science platform

L. Krieger¹, R. Nijzink¹, G. Thakur¹, C. Ramakrishnan², R. Roskar², S. Schymanski¹

*¹ Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, supported by FNR
ATTRACT (A16/SR/11254288)*

² Swiss Data Science Center, Zurich, Switzerland

Contact: louis.krieger@list.lu

MOTIVATION: REPRODUCIBILITY CRISIS

NATURE VIDEO · 25 MAY 2016

Is there a reproducibility crisis in science?

Charlotte Stoddart

Archive > Volume 533 > Issue 7604 > News Feature > Article

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

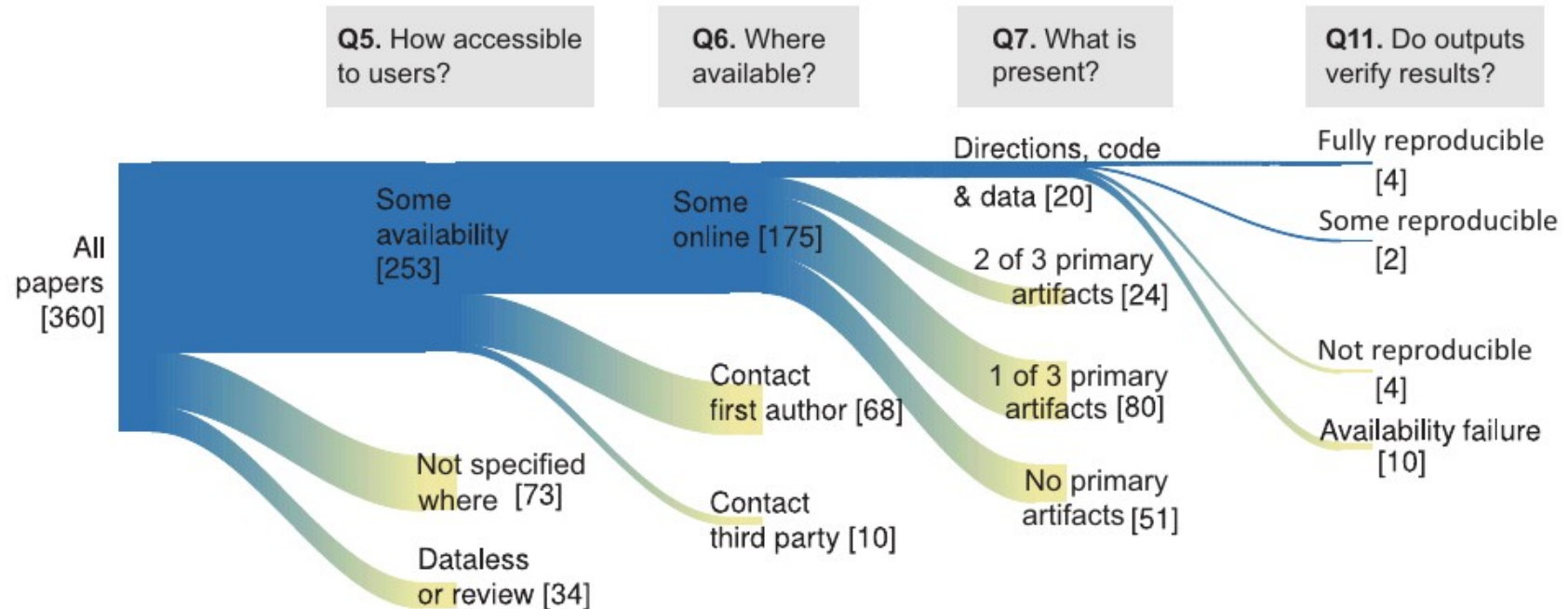
25 May 2016 | Corrected: 28 July 2016

"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."

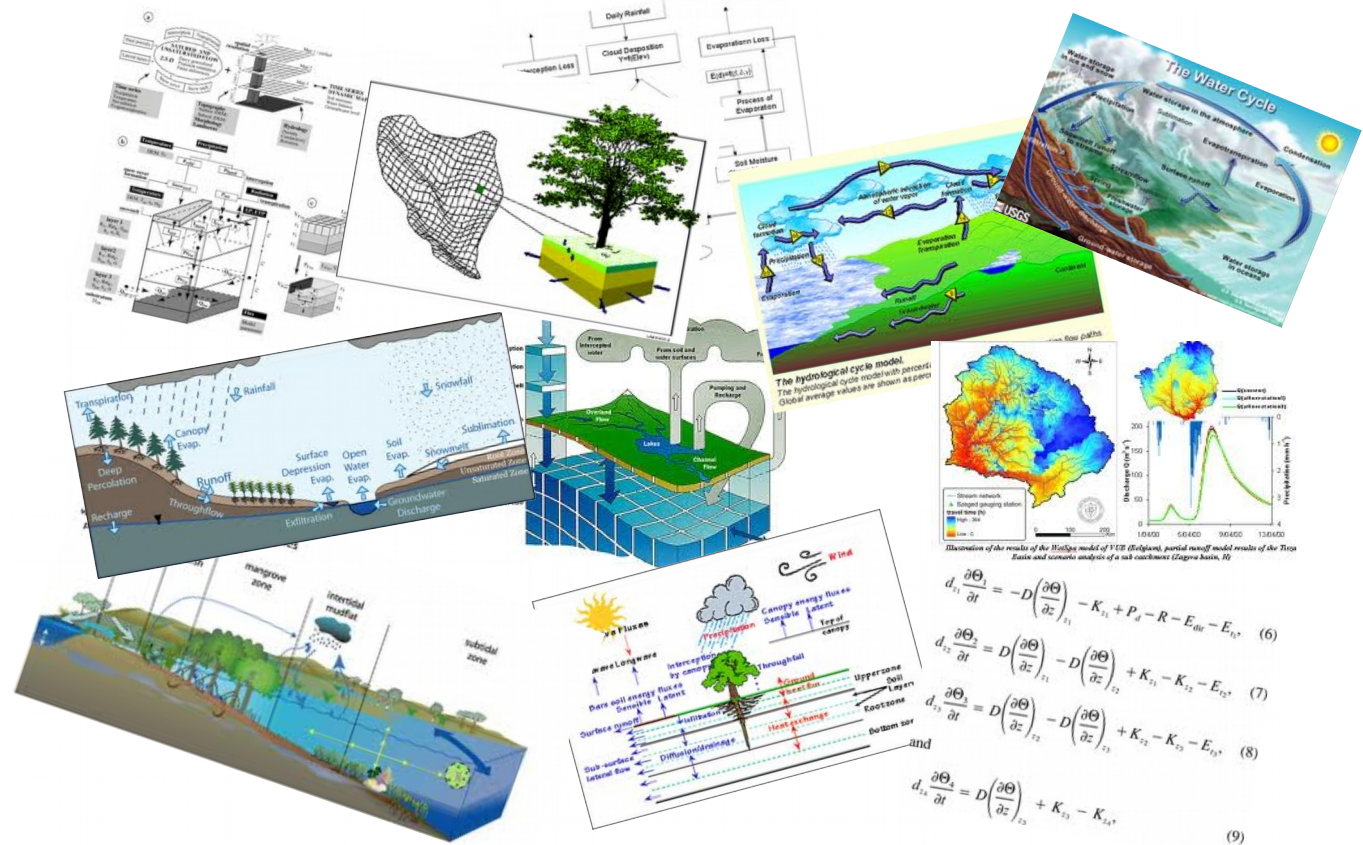
SAD REALITY IN HYDROLOGY: 1% OF RESULTS REPRODUCIBLE

Stagge et al. (2019): **Assessing data availability and research reproducibility in hydrology and water resources.**

SCIENTIFIC DATA, 6:190030, <https://doi.org/10.1038/sdata.2019.30>



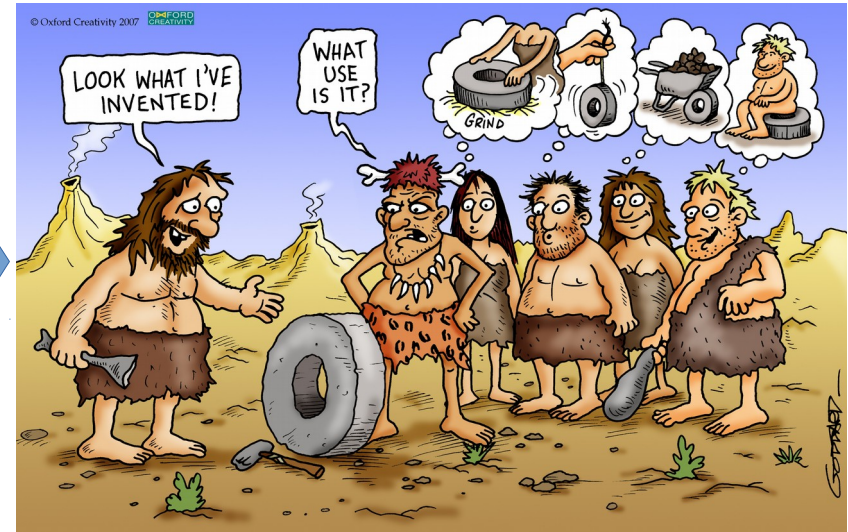
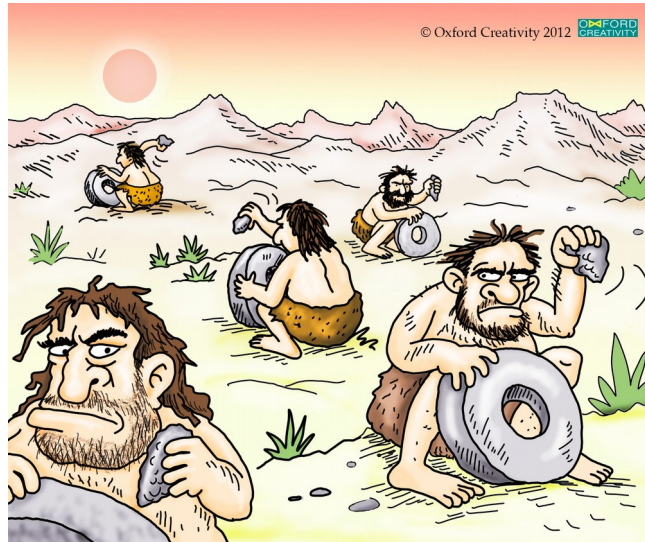
SAD REALITY IN HYDROLOGY: A MODEL FOR EACH DATA SET



OPEN SCIENCE: SHARING → COMMUNITY → PROGRESS

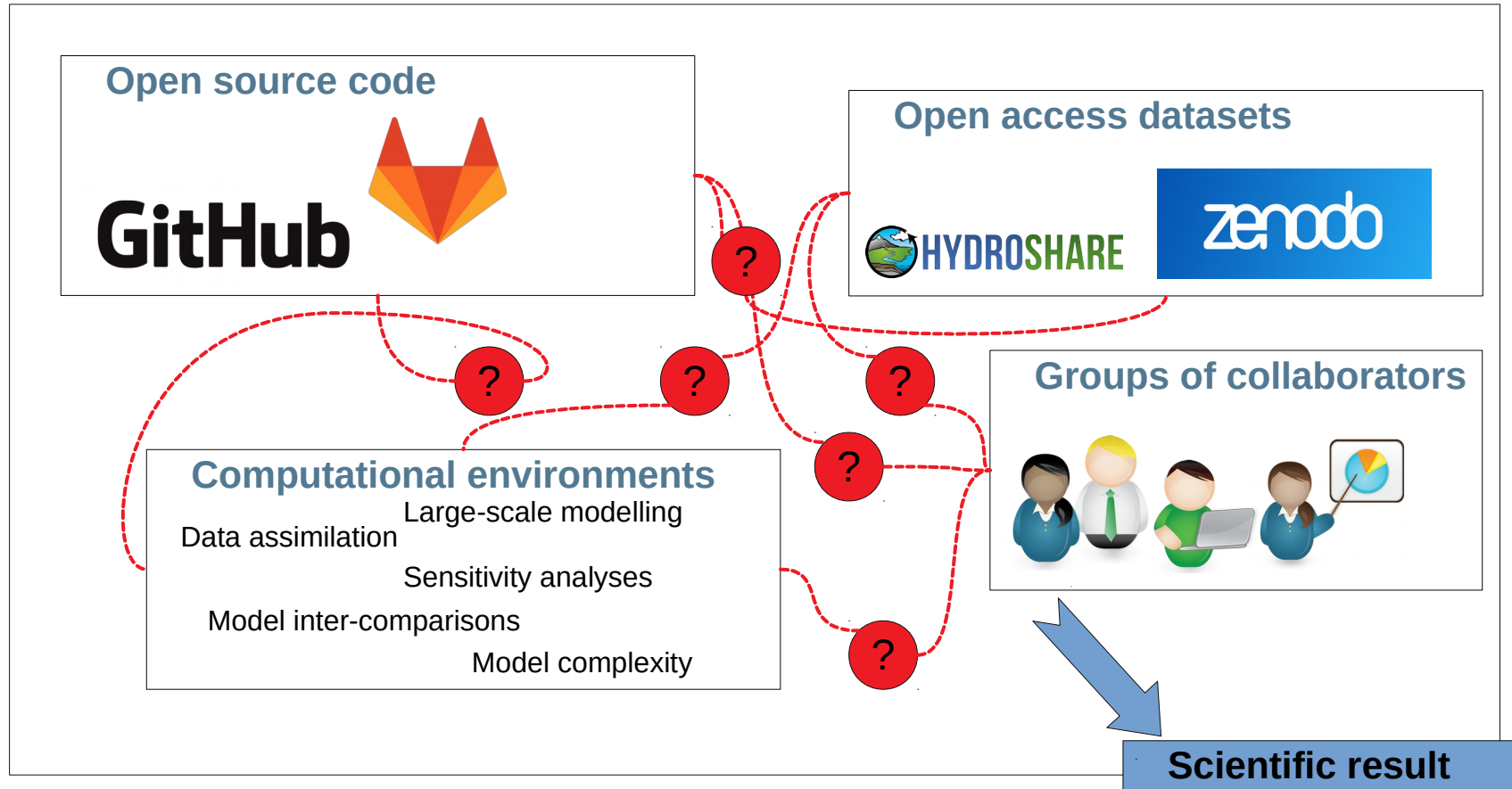
Lack of openness forces us to
keep re-inventing the wheel

How much more progress could
we make if we shared all our
methods, tools and inventions?



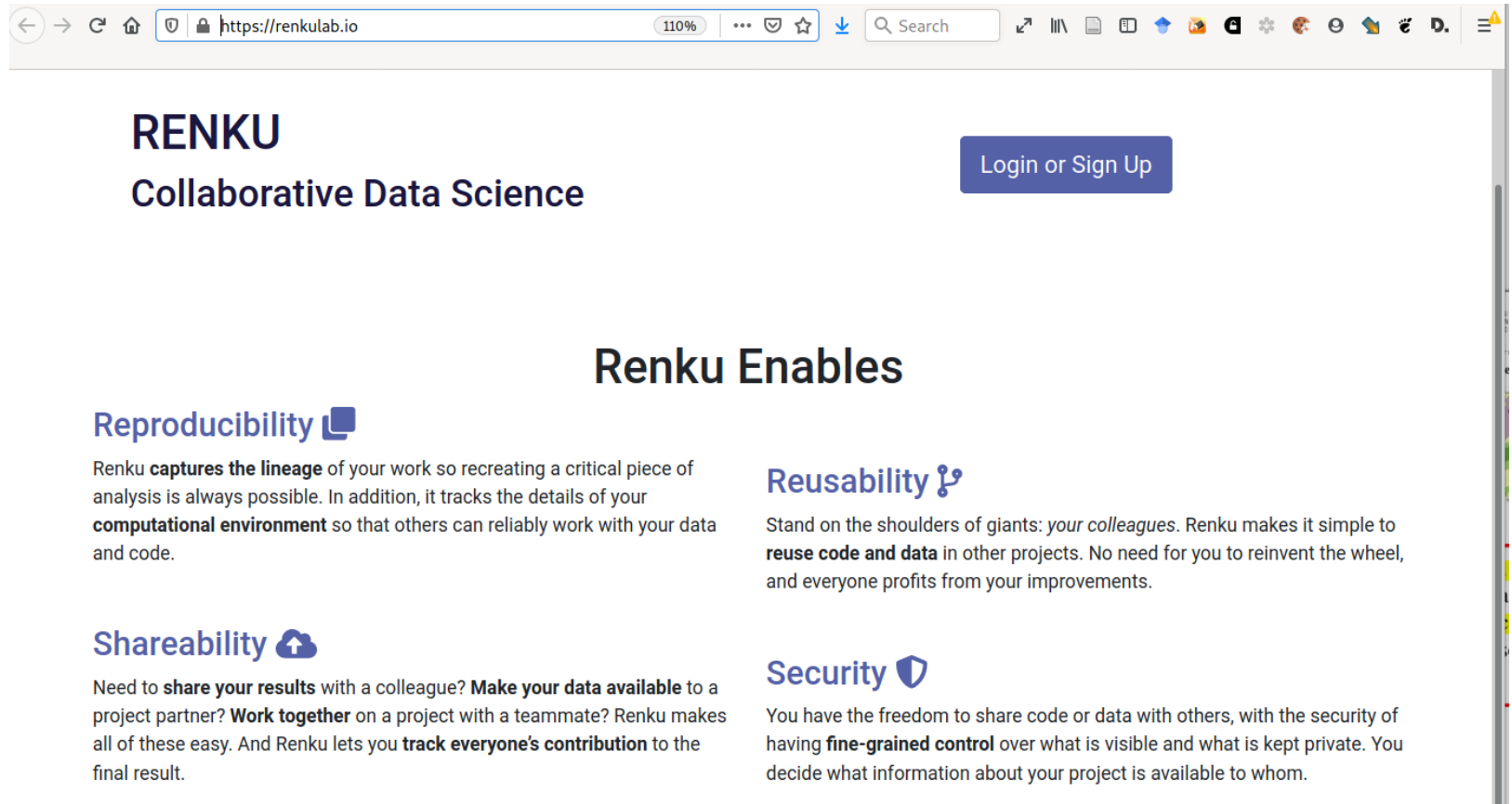
IMPORTANCE OF LINEAGE

Even if we have all elements, connections often unknown: Who used which code and which data in which environment to produce the results?



FULLY REPRODUCIBLE SCIENCE: RENKULAB.IO

LUXEMBOURG
INSTITUTE OF SCIENCE
AND TECHNOLOGY



The screenshot shows a web browser window with the address bar displaying 'https://renkulab.io'. The page content includes the 'RENKU Collaborative Data Science' header, a 'Login or Sign Up' button, and a section titled 'Renku Enables' with four sub-points: Reproducibility, Reusability, Shareability, and Security. Each sub-point has a brief description of its feature.

RENKU

Collaborative Data Science

[Login or Sign Up](#)

Renku Enables

Reproducibility

Renku **captures the lineage** of your work so recreating a critical piece of analysis is always possible. In addition, it tracks the details of your **computational environment** so that others can reliably work with your data and code.

Reusability

Stand on the shoulders of giants: *your colleagues*. Renku makes it simple to **reuse code and data** in other projects. No need for you to reinvent the wheel, and everyone profits from your improvements.

Shareability

Need to **share your results** with a colleague? **Make your data available** to a project partner? **Work together** on a project with a teammate? Renku makes all of these easy. And Renku lets you **track everyone's contribution** to the final result.

Security

You have the freedom to share code or data with others, with the security of having **fine-grained control** over what is visible and what is kept private. You decide what information about your project is available to whom.



RENKU 連句

Environment for collaborative, reproducible data science

Concepts

- Tracking of scientific steps to create data lineage, i.e. a knowledge graph
- Updating of out-dated results
- Tool to re-use or re-run analyses
- Sharing of analyses

Usage

- Command Line Interface
- Web platform

Features

RENKU is based on :

- Gitlab
- JupyterHub
- Kubernetes
- Keycloak
- Common Workflow Language

More info 

Lineage of code and data

renku run

- Tracks the used command
 - Lineage between inputs and outputs preserved
- Detects arguments
 - Settings, options and flags saved

renku dataset

- Stores meta-data:
 - link to original data
 - author, description, ...
- Convenient update function
- Git large-file storage (lfs)
 - No issues with big data

Reproducible and repeatable

renku log

- Shows lineage from inputs to outputs
- Creates **knowledge graph**

renku update

- Updates all outputs
 - Results based on old date automatically updated

Shared computational environments

<https://renkulab.io>

- Interactive environment platform
- Docker-based :
 - Same environment can be reproduced elsewhere
- Integrated gitlab server

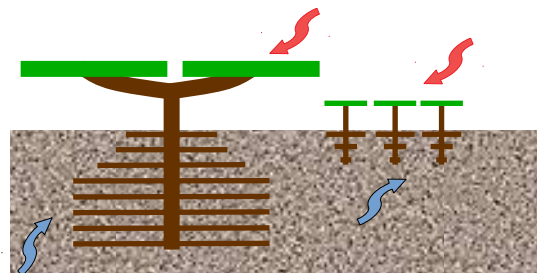
2 examples for the use of renku in research:

- **CASE 1 : LARGE NUMERICAL EXPERIMENTS**
- **CASE 2 : REPRODUCING LAB EXPERIMENTS**

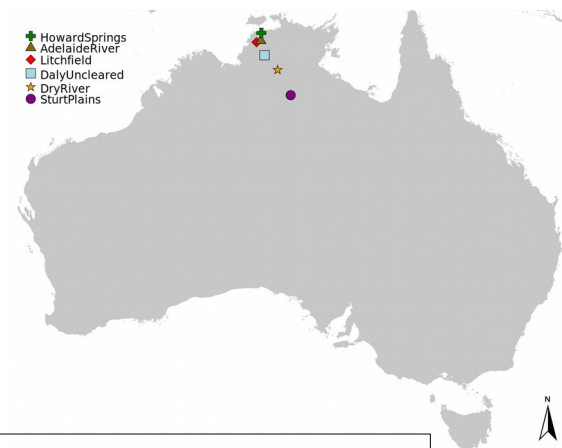
CASE 1 : LARGE NUMERICAL EXPERIMENTS

Vegetation Optimality Model

Optimizes vegetation properties to maximize the **Net Carbon Profit**, i.e. the difference of carbon uptake by photosynthesis and carbon costs of the system.



LUXEMBOURG
INSTITUTE OF SCIENCE
AND TECHNOLOGY



Challenges

- Parameter optimization:
 - Large computations
- Multiple study sites:
 - Complex lineage from inputs to outputs
- Multiple analyses:
 - Complex computations with many options, flags, settings etc.
- Different versions of data, model-code and pre- and post processing scripts

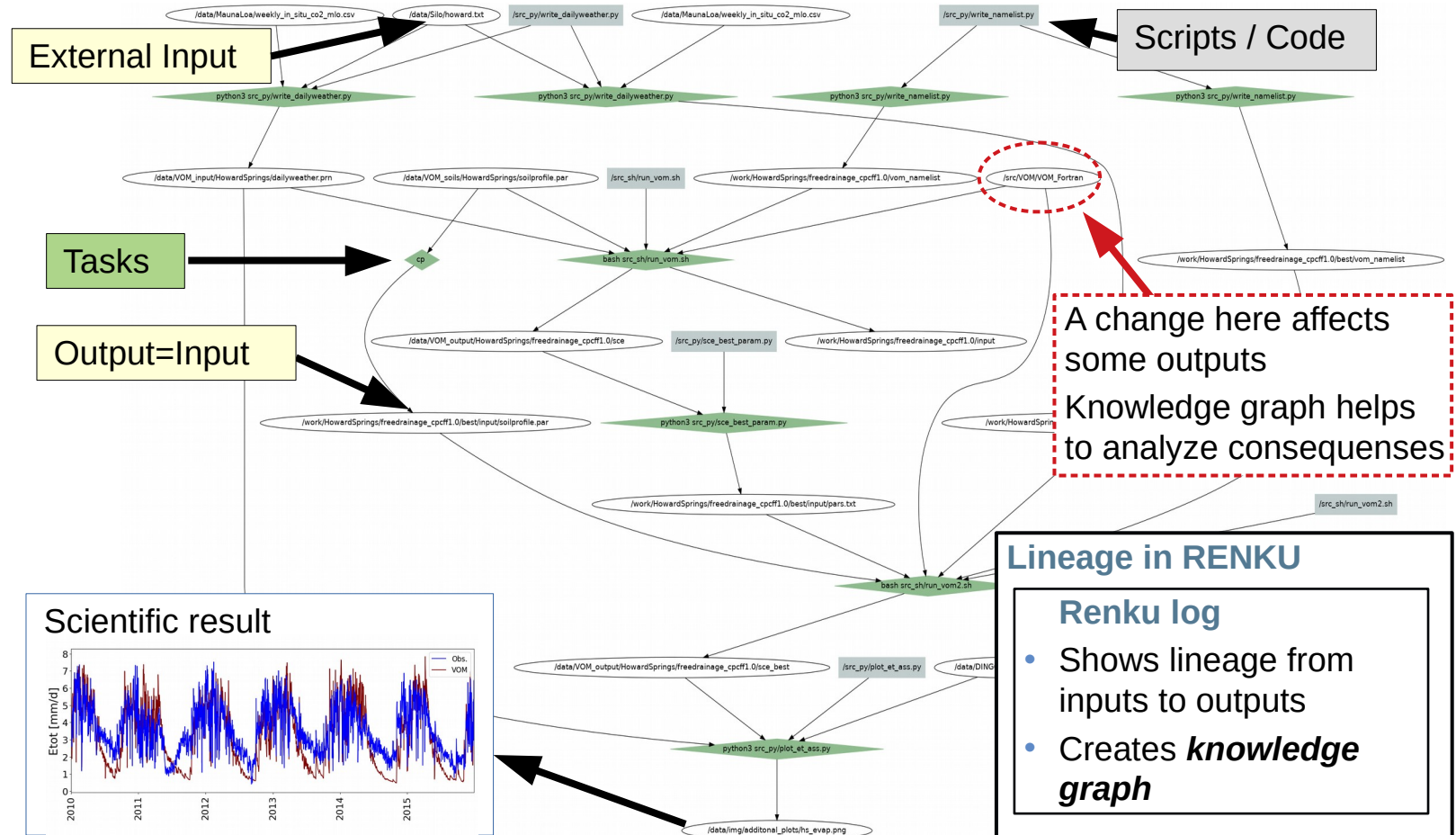


Why use renku in this scenario?

- Complex lineage visualized by a **knowledge graph**
 - Shows how input, outputs and scripts are connected
- Easy to repeat and reproduce analyses

Go to repository

CASE 1 : LARGE NUMERICAL EXPERIMENTS

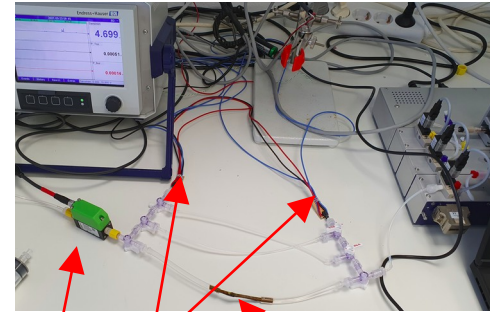


*This is a simplified knowledge graph. See original [repository](#).

CASE 2 : REPRODUCING EXPERIMENTS

Plant hydraulic experiments

Mathematical derivations are combined with experimental measurements to produce graphs of hydraulic permeability.



LUXEMBOURG
INSTITUTE OF SCIENCE
AND TECHNOLOGY



syringe pump

flow
sensor

pressure
sensors

plant twig

Challenges

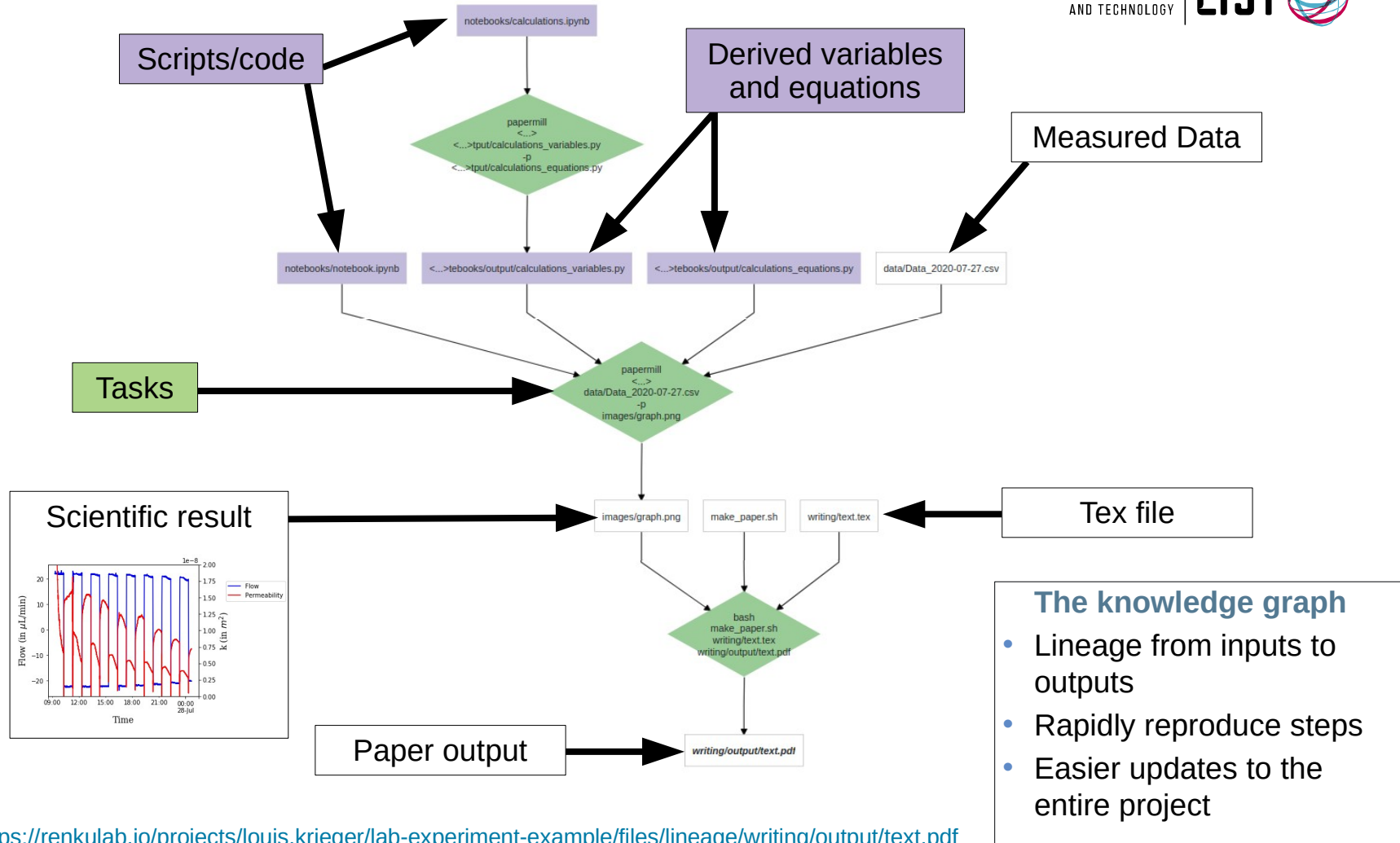
- Various sources of inputs :
 - Import of equations and derivations into code
 - Data and code used in creation of graphs
- Connect final data and plots to original data and docu, calibration functions, processing scripts
- Long chain of steps to follow if any changes are made, e.g. due to sensor re-calibration or updated processing scripts

Why use renku in this scenario ?

- Lineage tracking and knowledge graph generation
- Easy reproduction of workflows
- Easy update of affected results if calibration data or a processing script changes

Go to repository

CASE 2 : REPRODUCING EXPERIMENTS



Challenges

- Reproducibility and verifiability of research results not only requires data and code, but also:
 - Documentation
 - Versioning
 - Lineage between files
 - Computational environment

Solution

- Renku and renkulab.io provide:
 - Versioning (git-based)
 - Automatic lineage recording and knowledge graph generation
 - Reproducible and re-usable computational environments (docker-based)
 - Online collaboration platform, allowing to copy each others' projects and work with them without the need for local software installation!

