ETH-RAT

REPRODUCIBILITY AND REPLICABILITY:
OPEN SCIENCE ON TRUSTED DATA

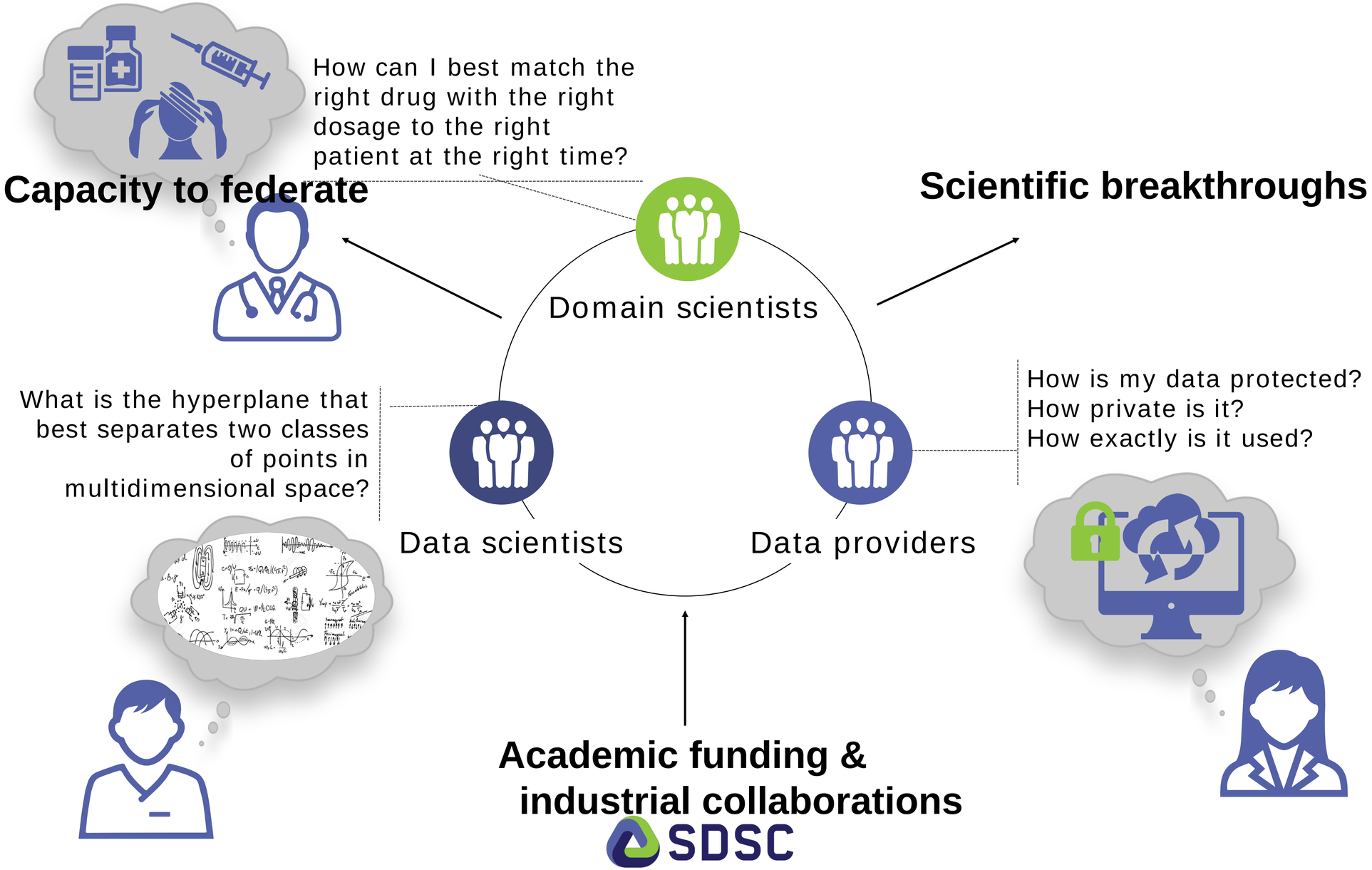
Christine Choirat | Swiss Data Science Center

SDSC

https://datascience.ch/
@SDSCdatascience

FOREWORD: SWISS DATA SCIENCE CENTER

Closing the gaps in the data science journey



A multinational team of professionals

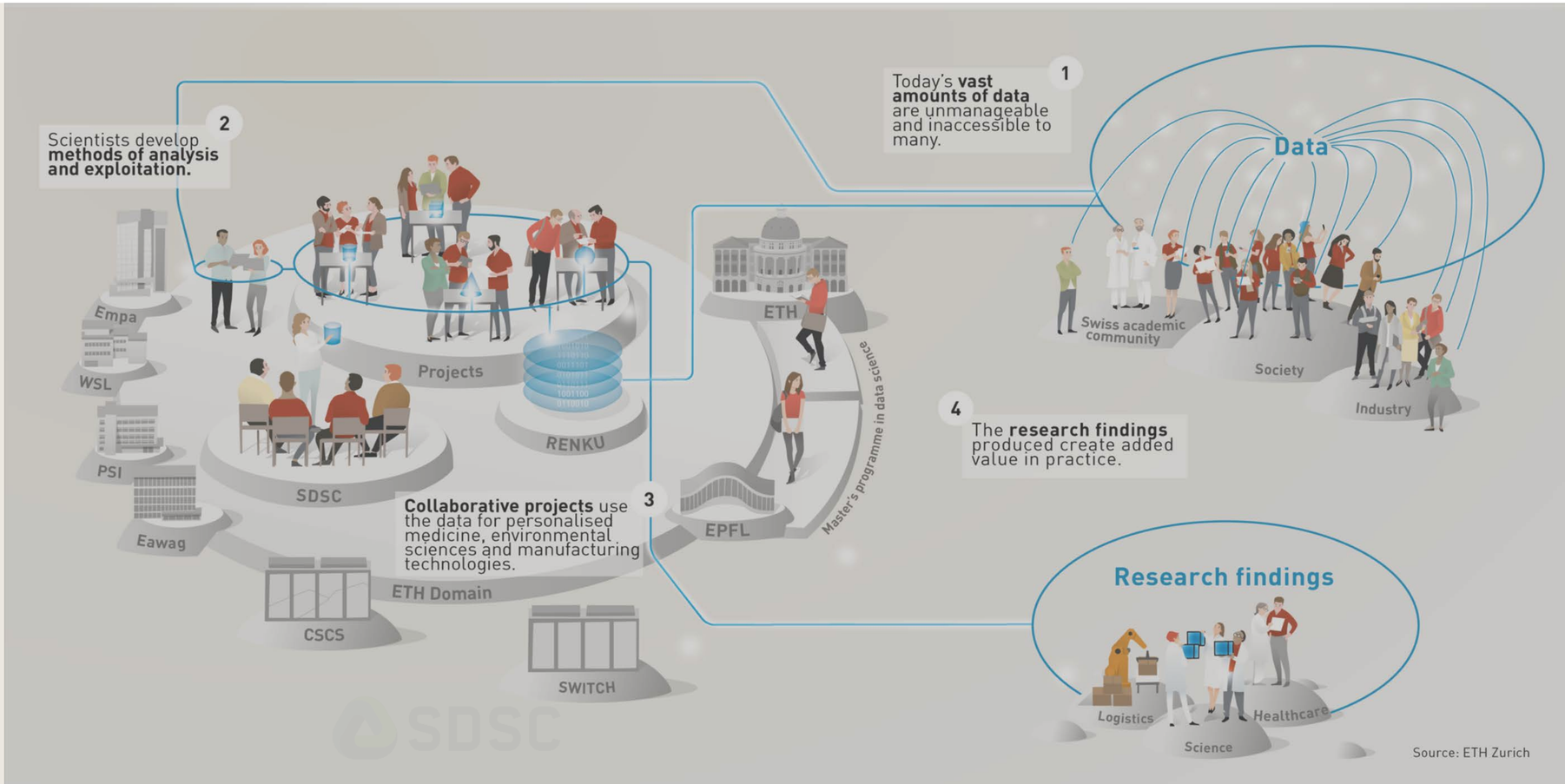
Joint venture between EPFL and ETH Zurich, with offices in Lausanne and Zurich
Fully operational since January 2017, growing to 50+ full-time professionals
Mission: accelerate the adoption of data science in academia and industry

EPFL
ETH zürich

Uniquely positioned in a broad landscape

- Very effective collaborations between two world-renowned engineering schools
- Dedicated team of scientists to bridge the various gaps in the data science journey

… enabling adoption of data science **at a scale only possible when pooling resources**



Main focuses

Embedded R&D collaboration

Academic and industry research collaborations

- 30 academic projects in personalized health and environmental science
- Several signed industry partnerships
- Initiating discussions at the international level

Insights as a Service

RENKU, the SDSC analytics platform (Open Source)

- Facilitates multi-disciplinary collaborations in data science & AI
- Promotes reusability / reproducibility of science (respects FAIR principles)
- Extremely positive response from academia and industry

Swiss Data Custodian

Data Custodian, a multisided platform to establish trust and transparency in data usage

- A data vault and secure multiparty compute ecosystems (reference architecture lead)
- Enables cooperation between mutually non-trusting parties (trusted intermediary)
- Promotes a shift from data ownership to the ownership of its use (preserve data sovereignty)

Educational Services

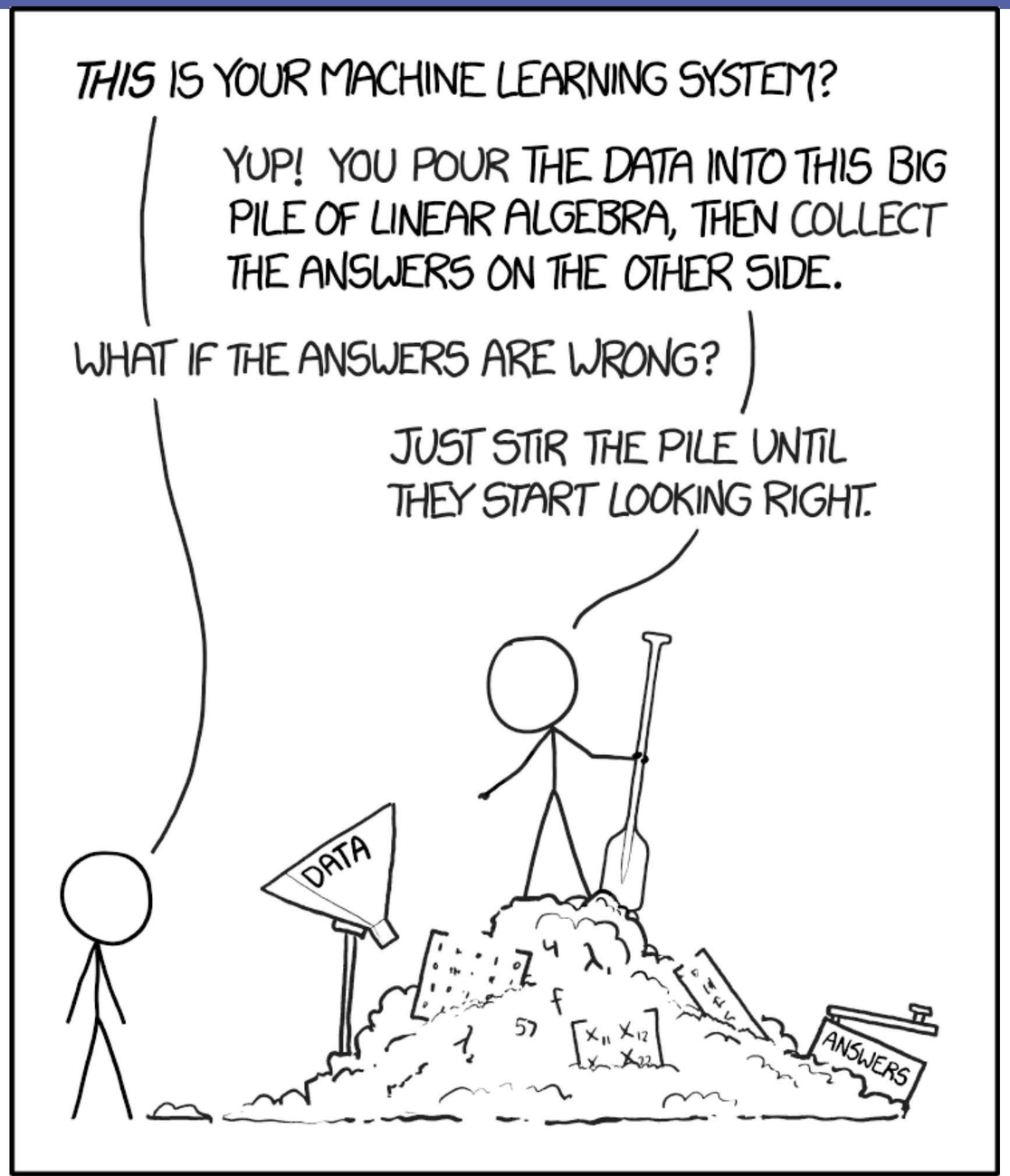
Contribution to the education in Data Science

- Involved in EPFL and ETH Zurich Master's Degrees in Data Science
- Contributes to continuous education through EPFL / HEC Lausanne CAS and ETH Zurich DAS programs
- Coaching programs for industry partners

SDSC

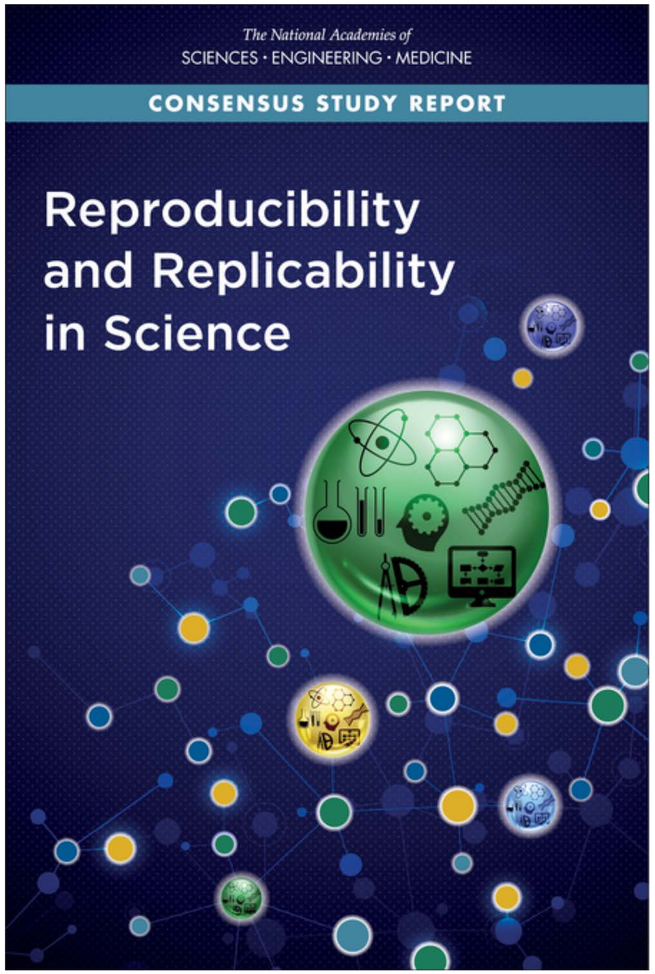
REPRODUCIBILITY “CRISIS”

PR situation to avoid…



Guidelines

- Researchers: include a clear, specific, and complete description of how the reported results were reached
- Funding agencies and organizations: should consider investing in research and development of open-source, usable tools and infrastructure that support reproducibility
- Journals: should consider ways to ensure computational reproducibility
- The NSF: should take steps to facilitate the transparent sharing and availability of digital artifacts, such as data and code



SDSC

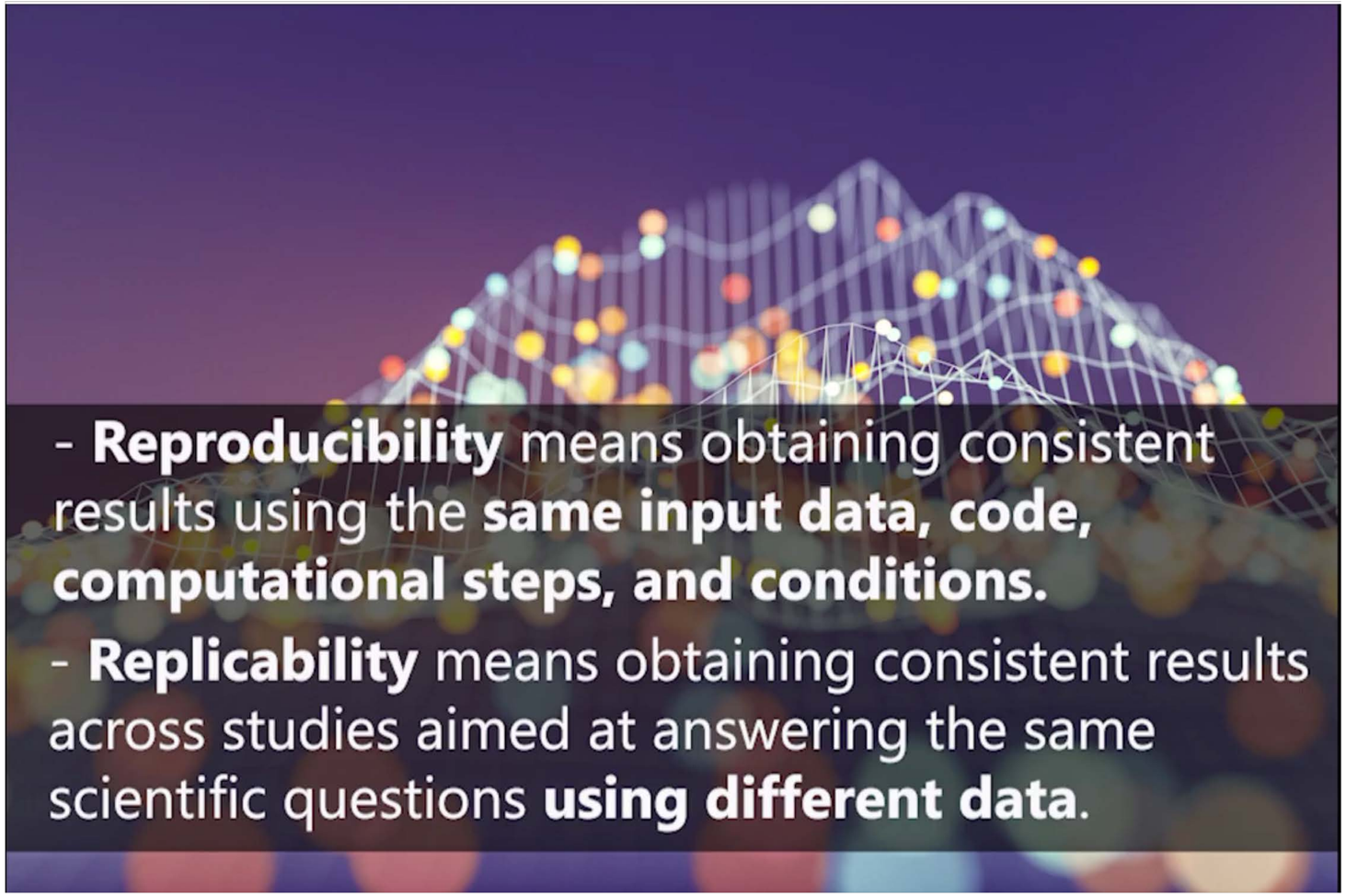
Guidelines

Source: NASEM

Reproducibility and Replicability in Data Science

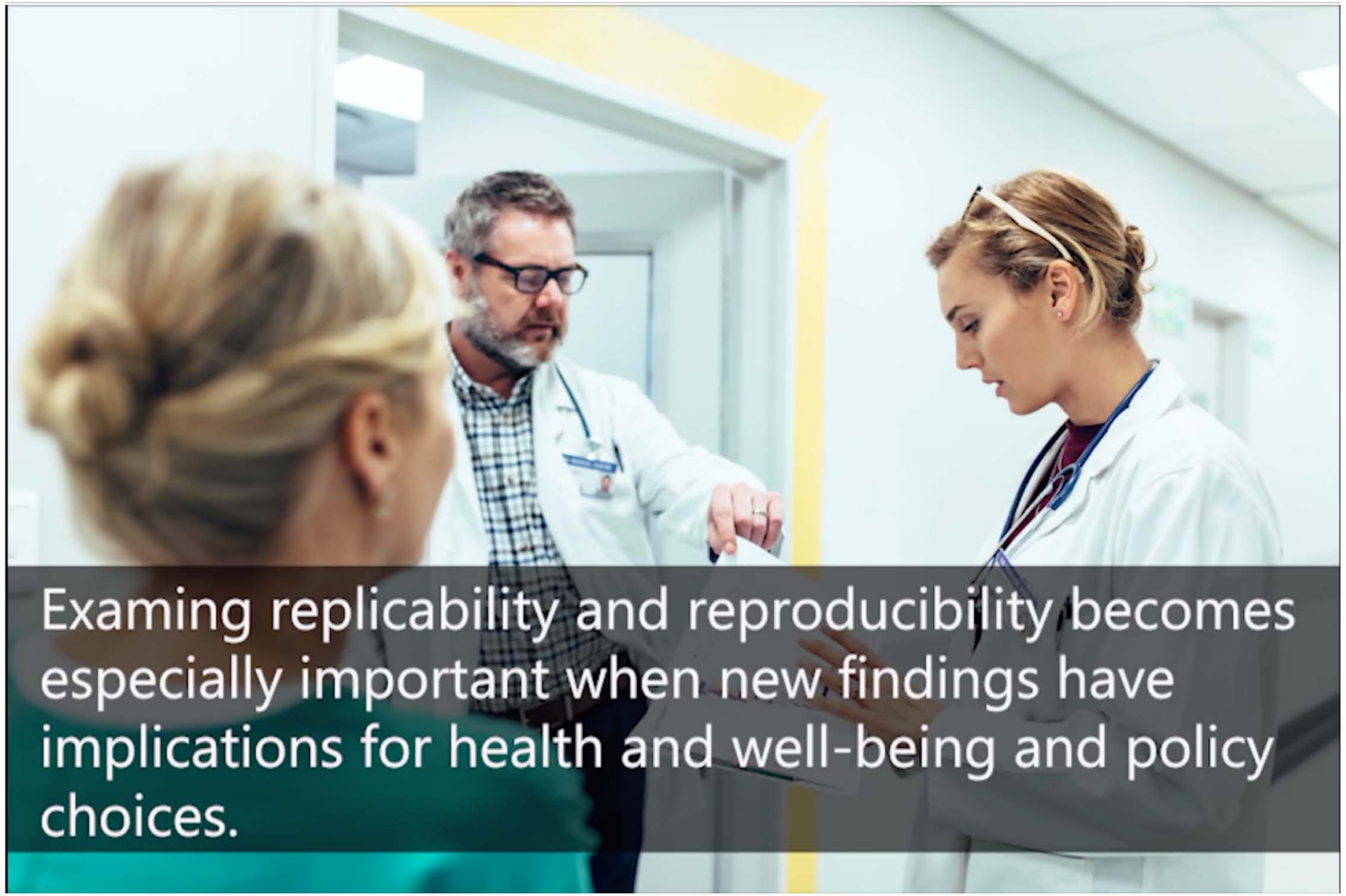
SDSC

Reproducibility and Replicability in Data Science



SDSC

Reproducibility and Replicability in Data Science



SDSC

Reproducibility and Replicability in Data Science

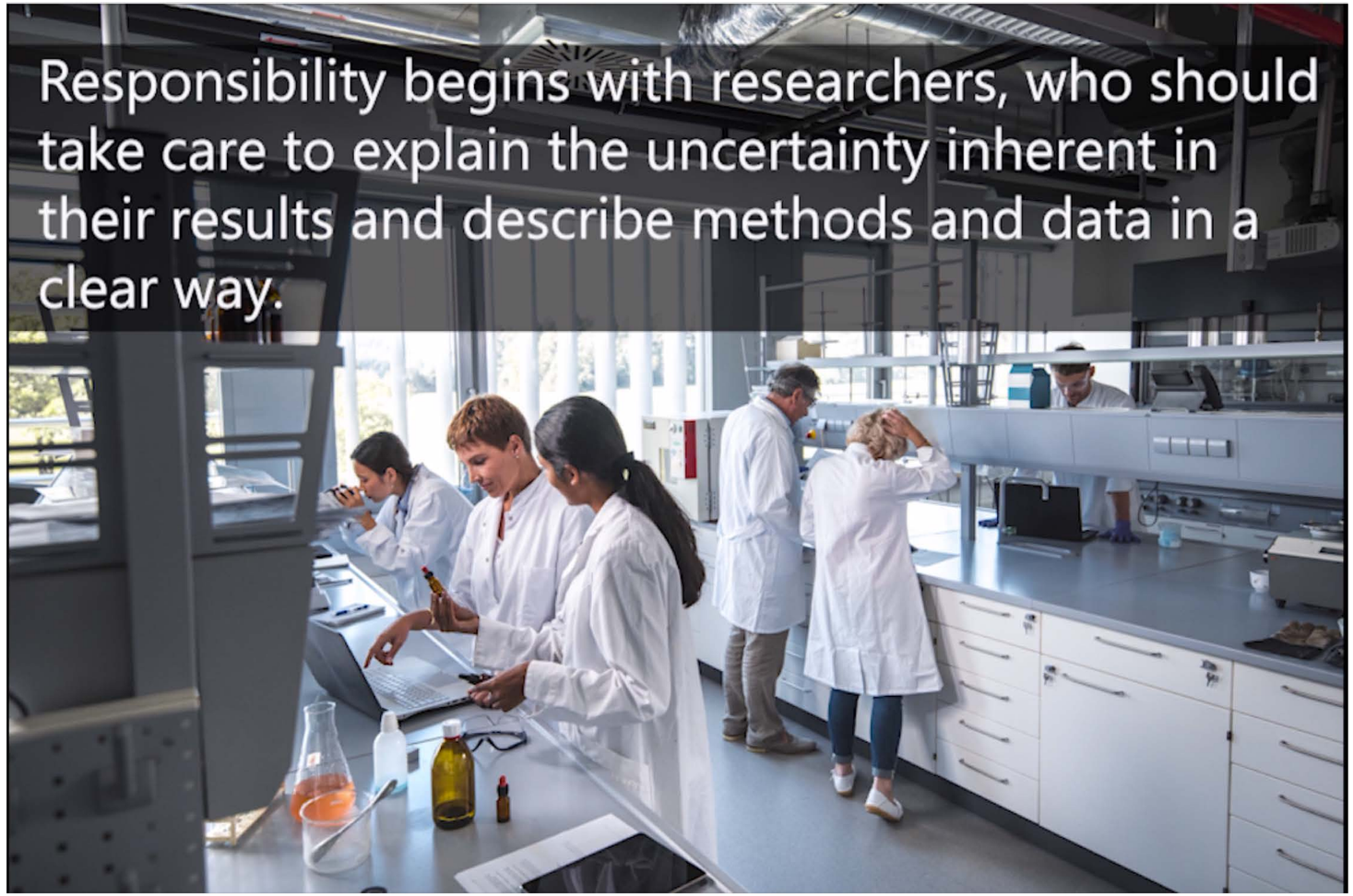
Reproducibility and Replicability in Data Science

SDSC

Reproducibility and Replicability in Data Science

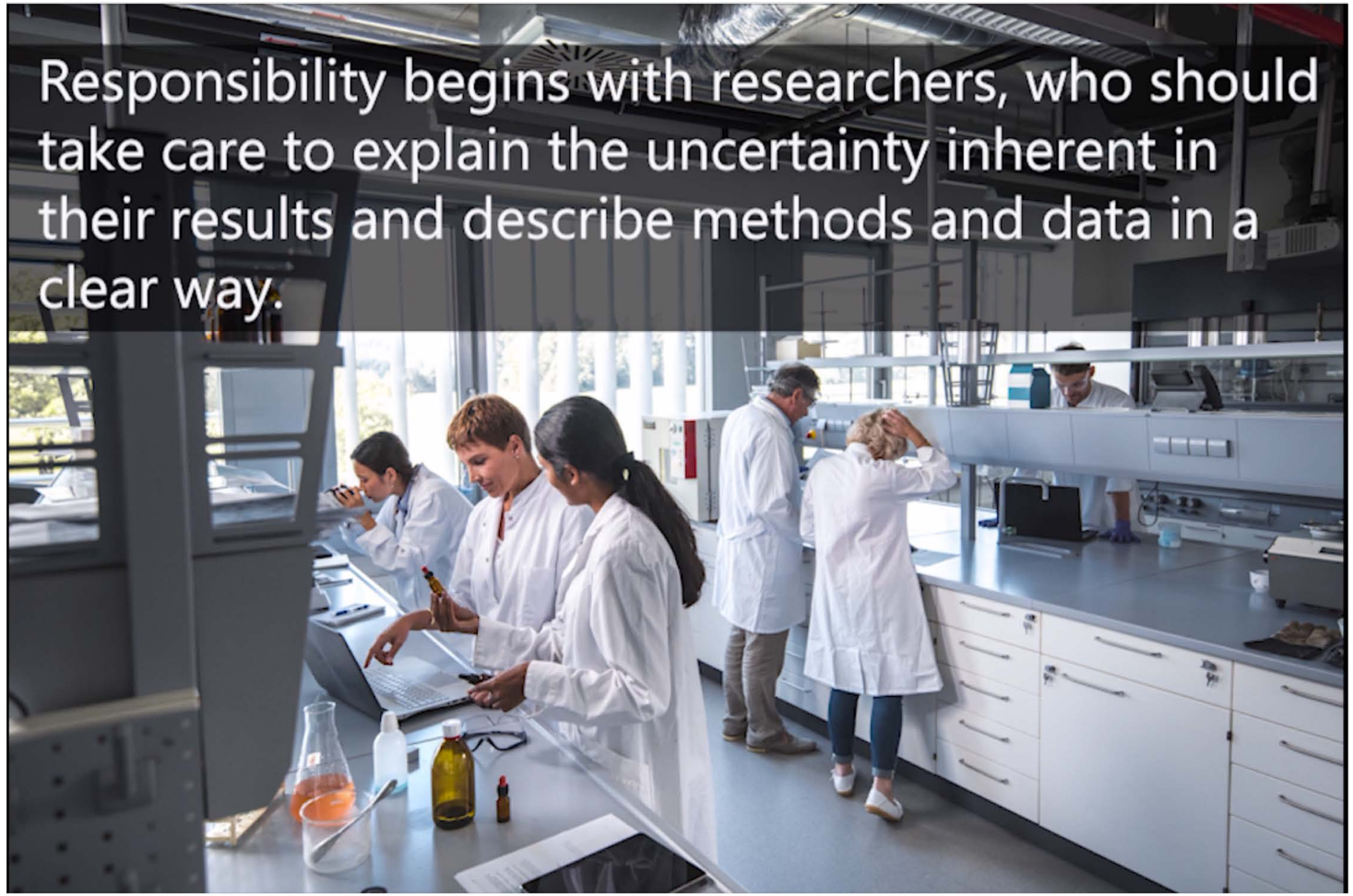
SDSC

Reproducibility and Replicability in Data Science



SDSC

Reproducibility and Replicability in Data Science



Five FAQs in Data-Driven Research

1. How did I compute this result?

2. How does new data change this result?

3. How did you compute *your* result?

Can I use your data to reproduce it?
With your code?
On your infrastructure?

4. Has anyone ever used an <XYZ-algorithm> on this data?

   Did it work?

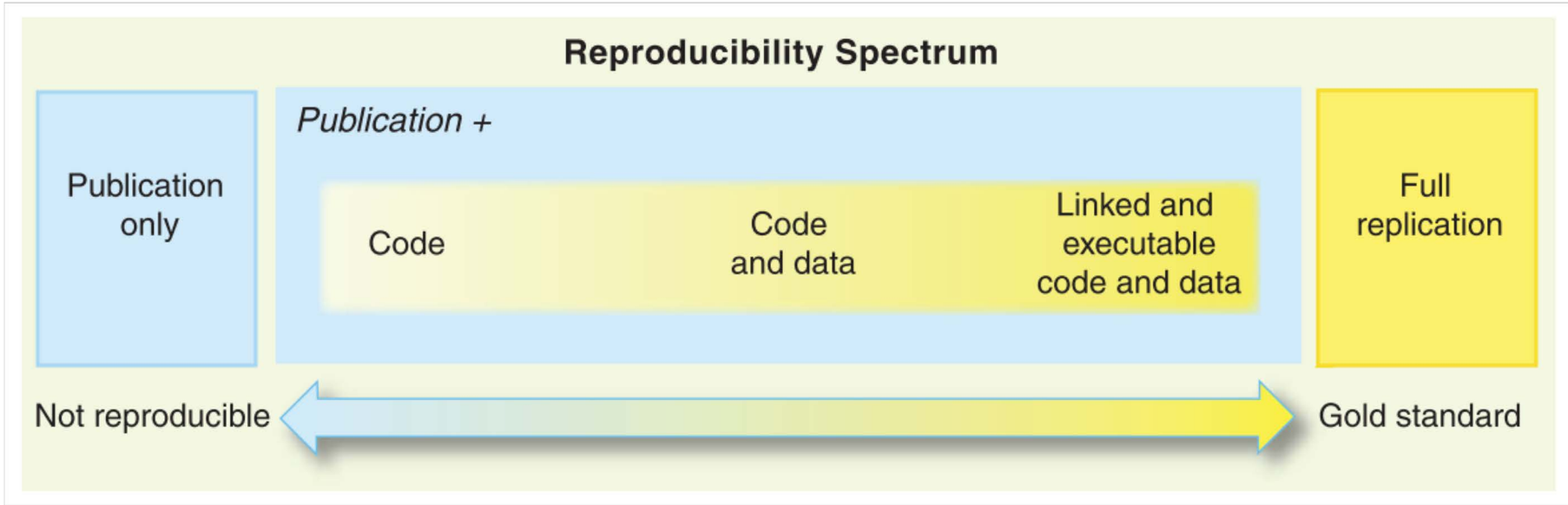
5. Who is using the data and algorithms?

In which context?

Five Questions —> Three Words

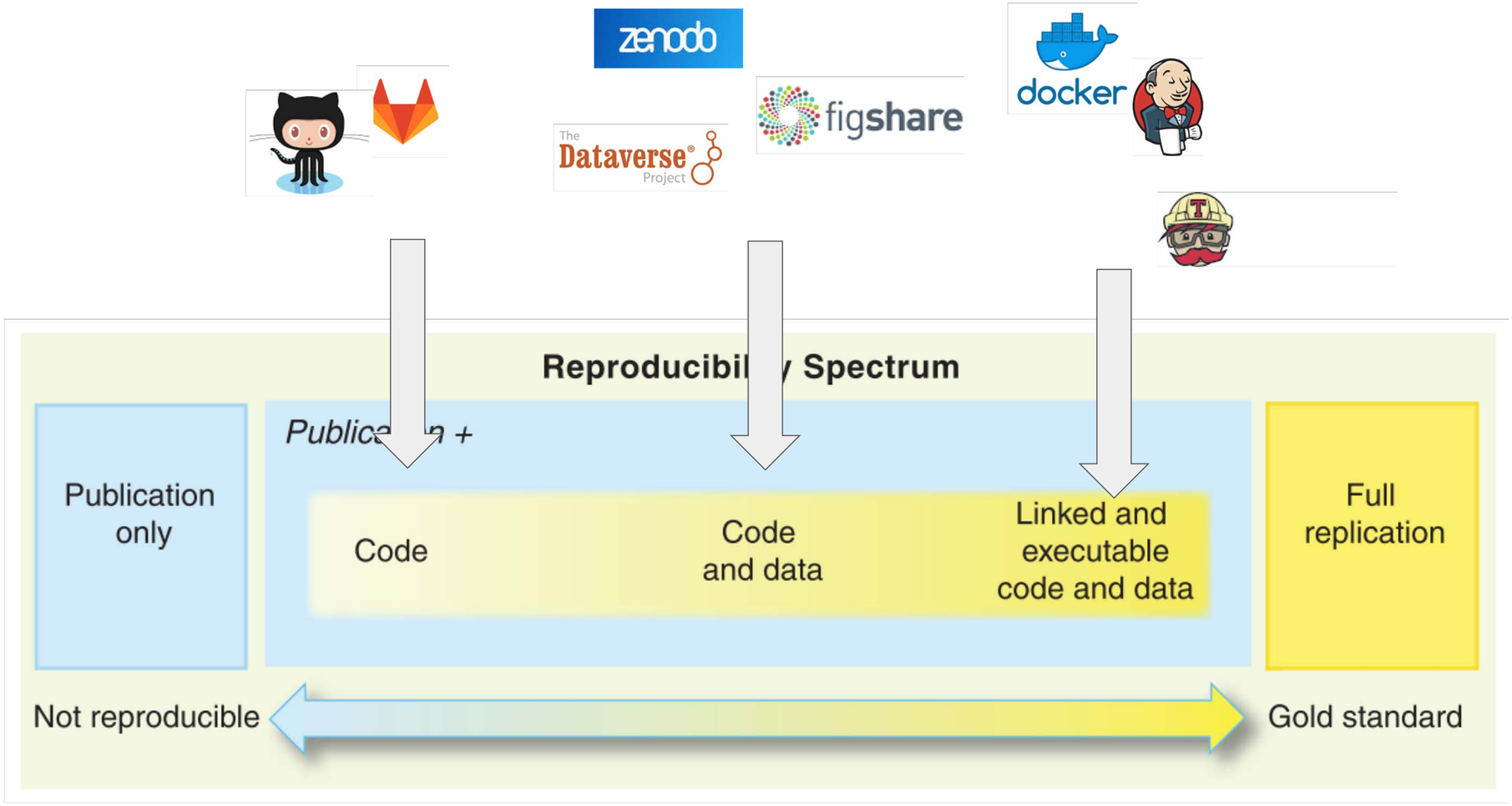
Reproducibility
Reusability
Collaboration

Reproducibility spectrum



Source: [Peng, 2011]

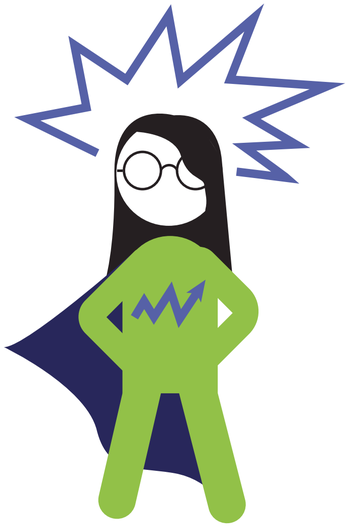
Reproducibility spectrum



Source: [Peng, 2011]

A Myriad of Tools

Hard **work** to make science reproducible, accessible, open etc…

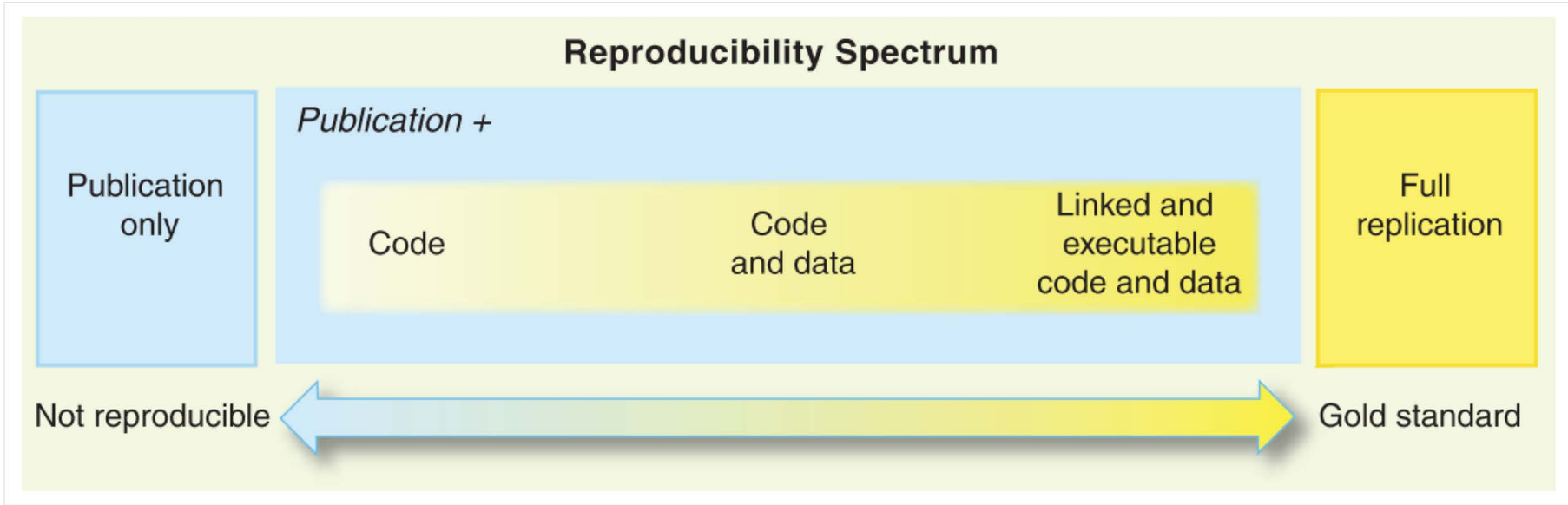


- Version control (git)
- Code sharing (GitHub, GitLab, BitBucket)
- Data sharing (Zenodo, Figshare, institutional digital archives)
- Presentation and communication (Jupyter, RStudio)
- Correctness – testing (CI, e.g. travis)
- Packaging, containerization (docker, singularity)

Difficult to stay **productive** and worry about all of the above!

Reproducibility spectrum

RENKU 連句



Source: [Peng, 2011]

SDSC

RENKU - 連句 - A platform for reproducible (data) science

local

Jupyter Notebook

RStudio

MATLAB

…

RENKU

- Reproducibility
- Reuse
- Collaboration

Gitlab | RDF

Docker | Keycloak

JupyterLab

Kubernetes

On premise,
Or on the cloud

SDSC

Terminology

- We borrow the RENKU name from the Japanese word for *linked-verse poetry*
- A "ku" is a verse in a renku poem
- We use "ku" to mean a piece of the data analysis process – includes discussion, code, and results

Renku

1.
Provide the means to create reproducible (data) science

2.
Facilitate the sharing and reuse of research artefacts

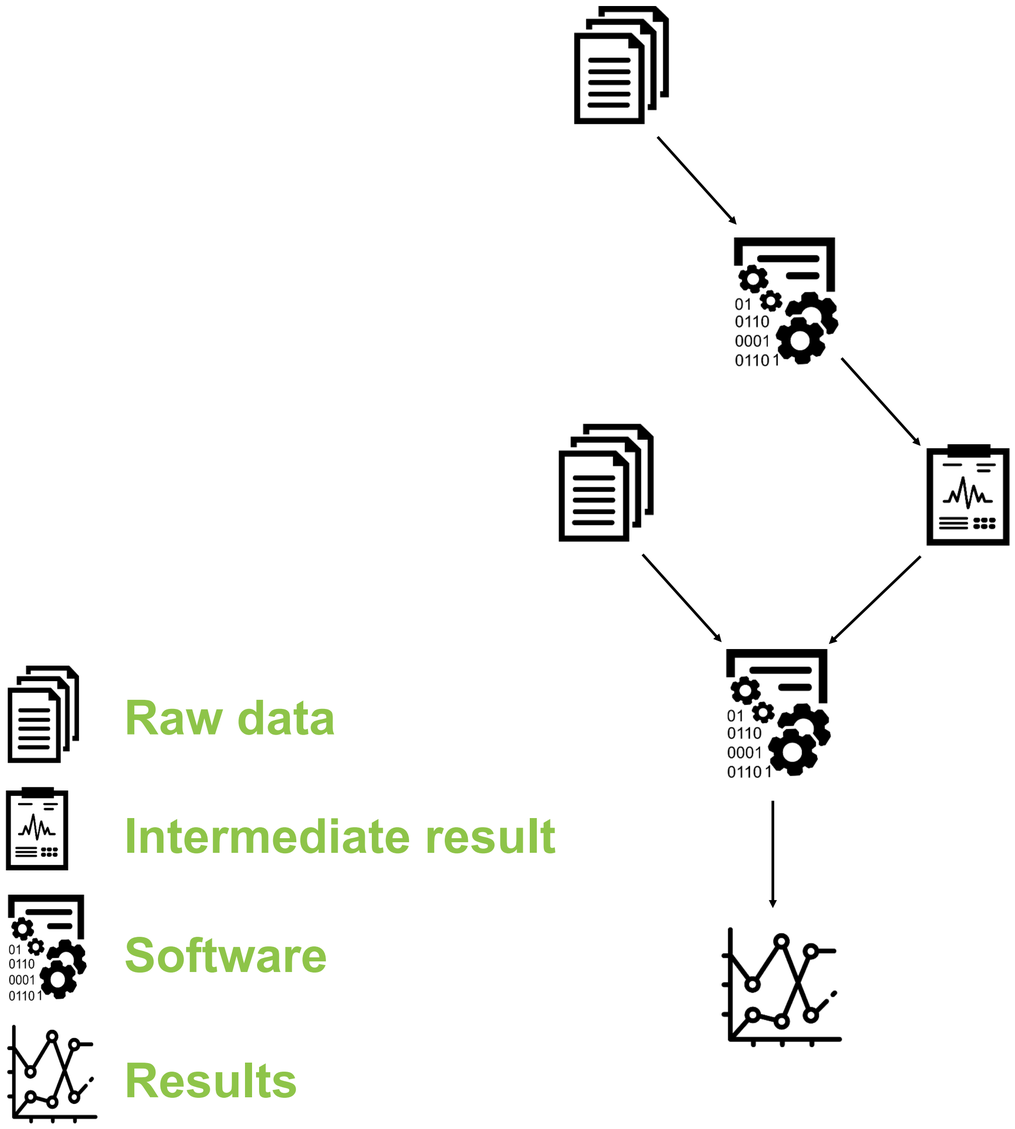
3.
Foster a collaborative environment for interactive prototyping

4.
Enable the discovery of relevant data and methods

5.
Allow federated access across institutions giving each the freedom to impose its own access controls over resources

SDSC

Capture the scientific process

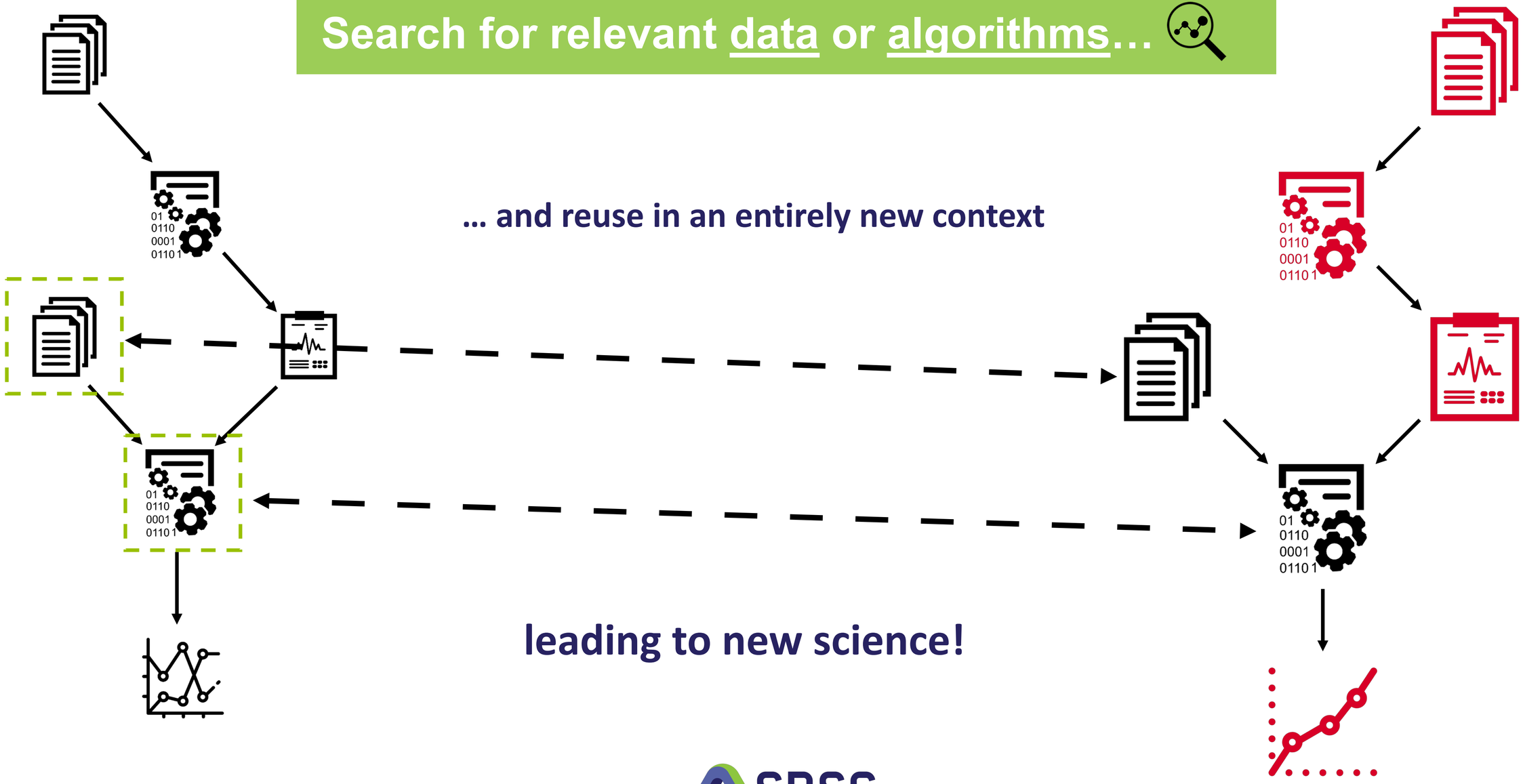


1. Inputs and outputs of analysis steps are recorded into a **knowledge graph** *while the work is being done*
2. Steps can be **repeated** or integrated into more complex **workflows**
3. **Provenance** of all data products is always accessible via simple tools
4. **Version control** is built-in for data, code, and workflows

SDSC

Reuse and repeat

Search for relevant data or algorithms…



… and reuse in an entirely new context

leading to new science!

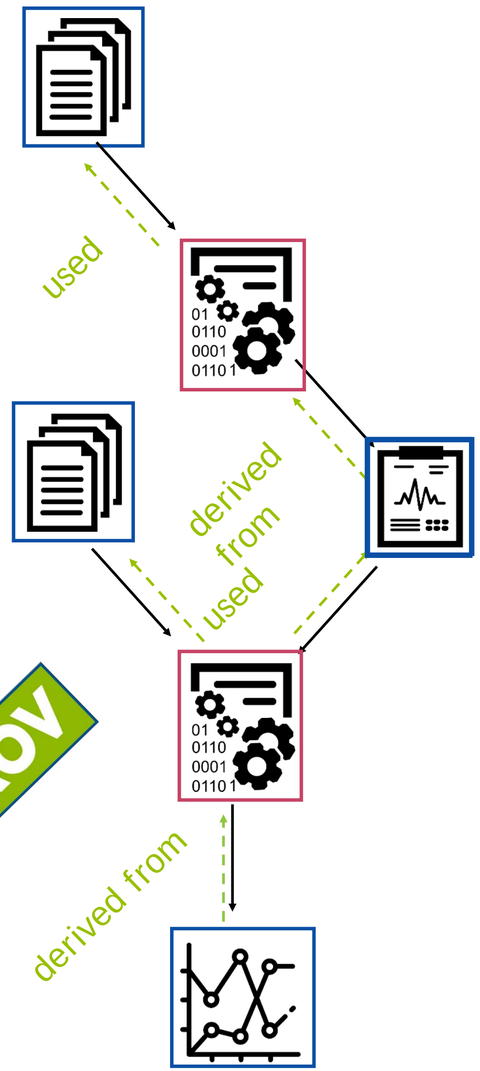
SDSC

Encapsulate with rich metadata

JSON-LD

- Metadata use Dublin Core, FOAF, and Schema.org
- Provenance graph is based on PROV-O W3C recommendation

W3C PROV



COMMON WORKFLOW LANGUAGE

- CWL for representing all computational steps
- Capture individual steps from user input
- Tools for constructing workflows from basic pieces
- Rely on container technologies to ensure reproducibility

SDSC

Reusability and collaboration

- Who is using the data and how?
- Which algorithms are used to answer which questions?
- How to regenerate results if new data becomes available? If old data is now off-limits?
- Who to credit?
- How popular is my work/the work of my lab/my unit?

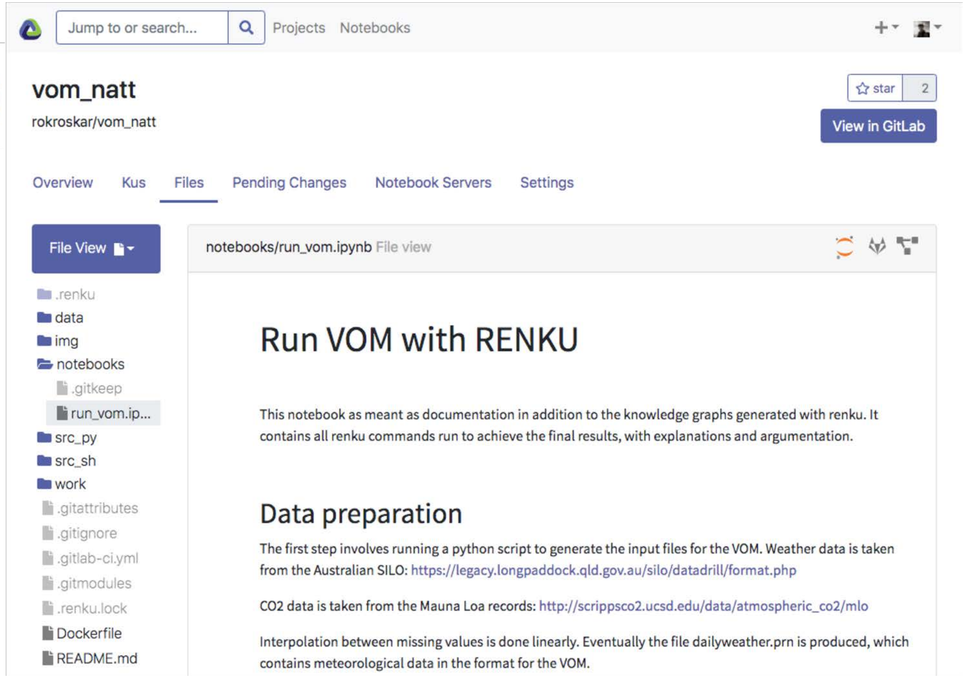
TRUST

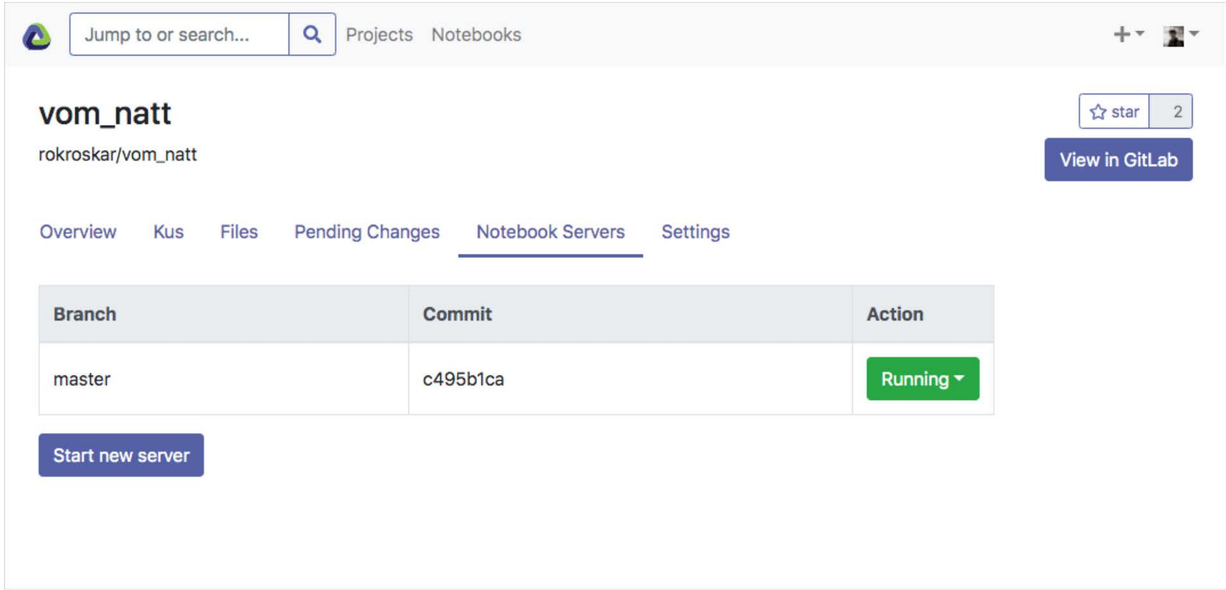
SDSC

What is Renku, really?

A platform integrating:

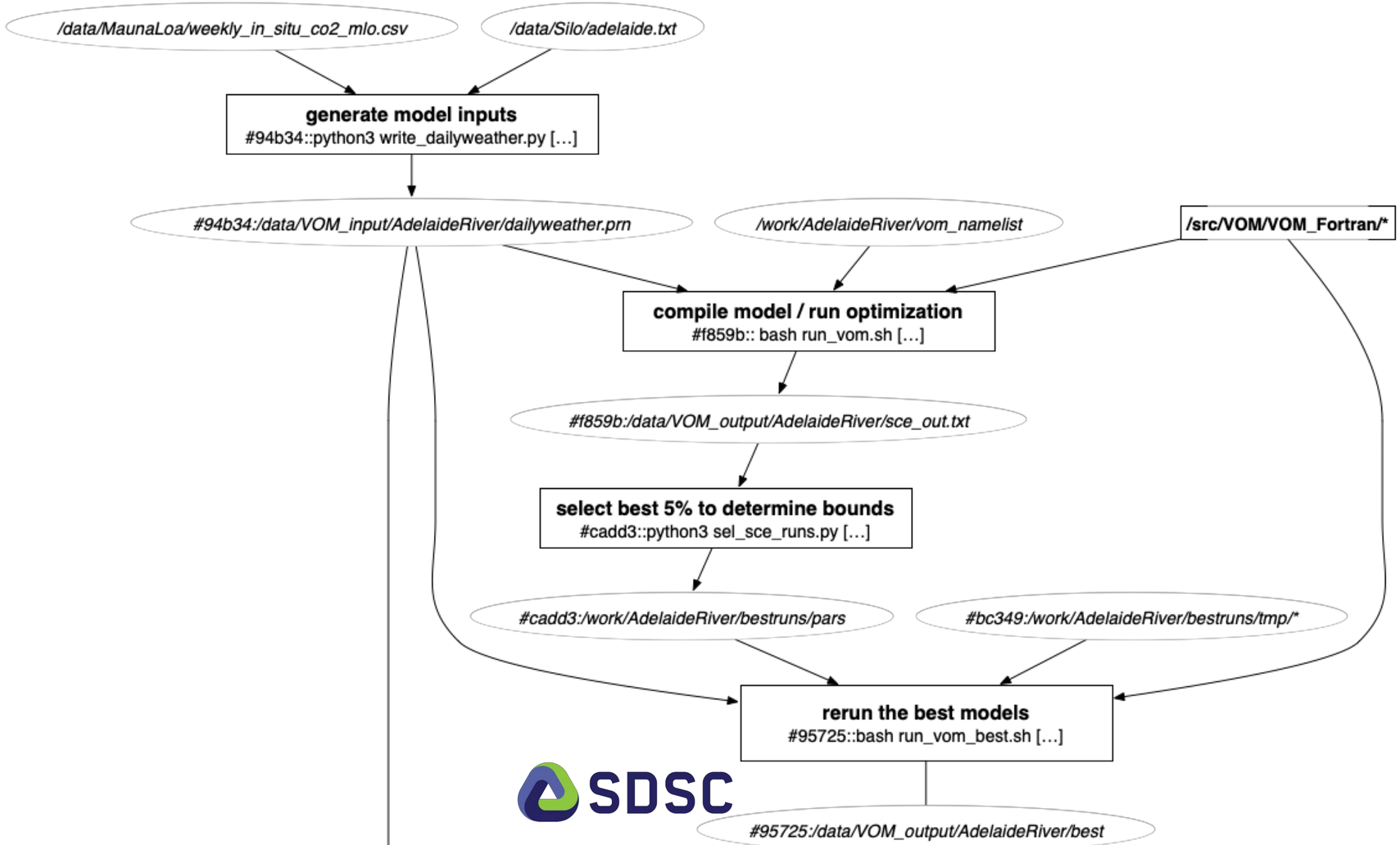
git, **Jupyter/RStudio**, **docker**, analysis **workflows** linked with a **knowledge graph**



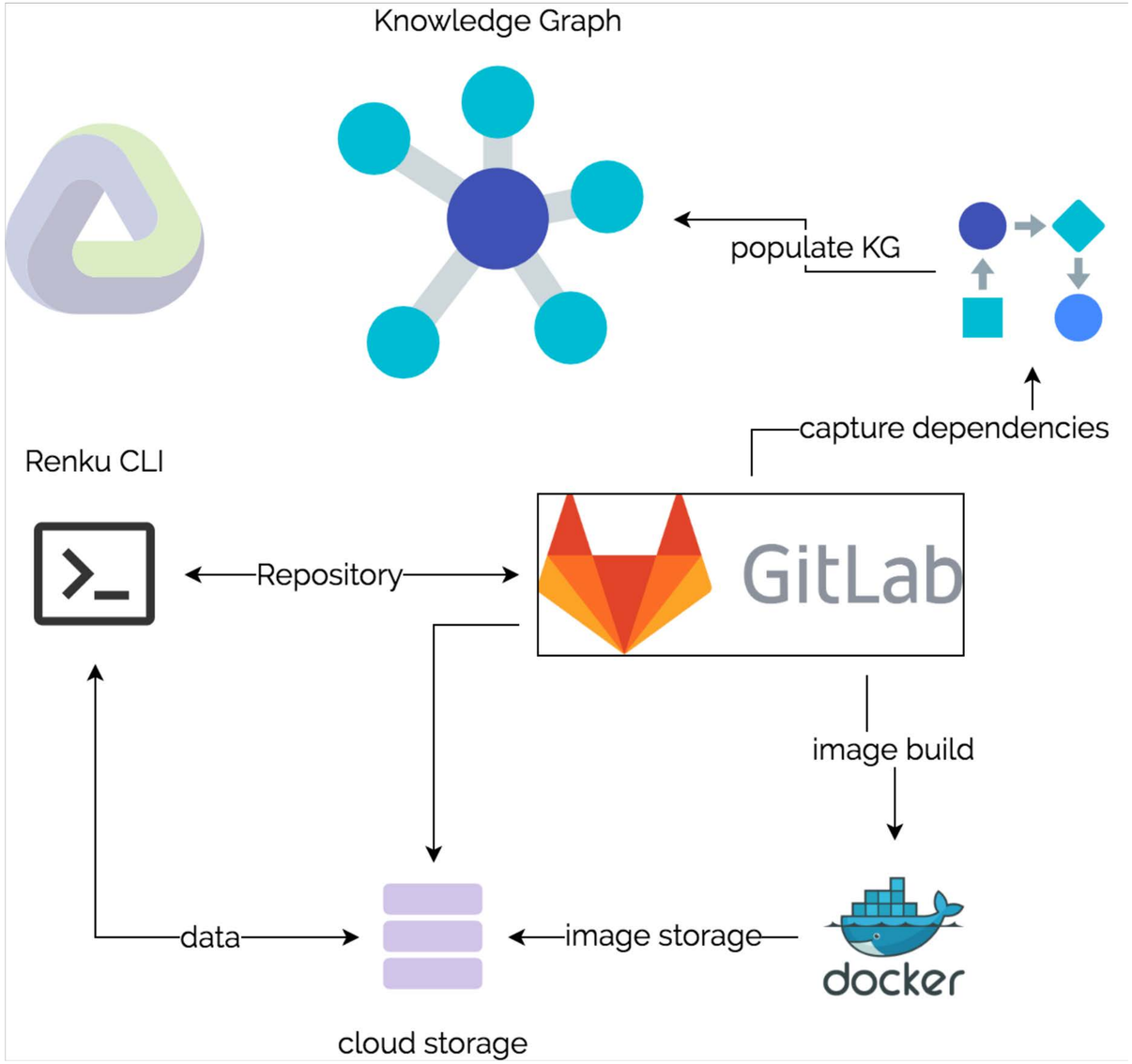


What is Renku, really?

A platform integrating:
git, **Jupyter/RStudio**, **docker**, analysis **workflows** linked with a **knowledge graph**



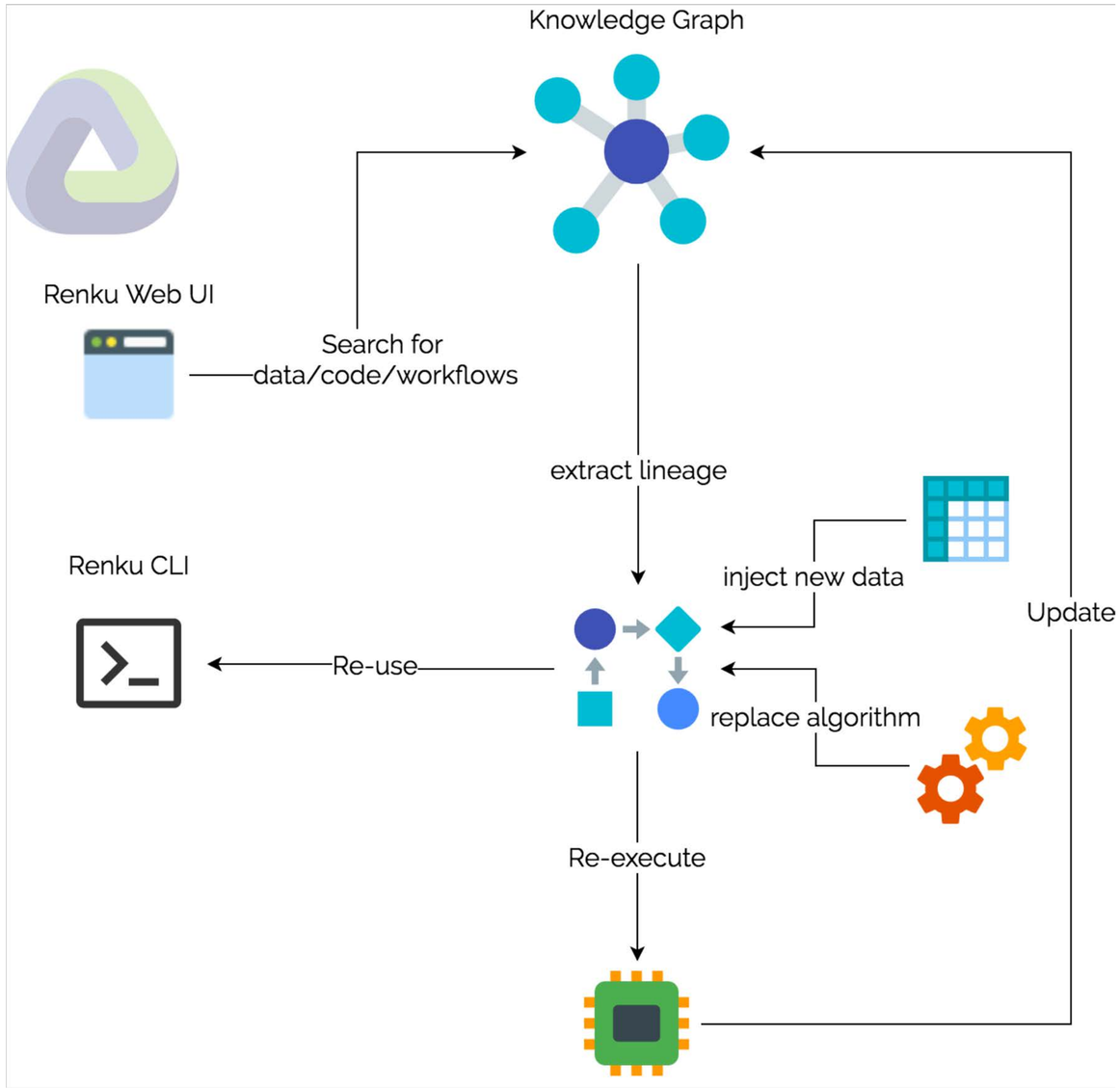
Reproducibility



- CLI developed from ground up taking advantage of Git to capture lineage
- Designed with interoperability in mind, using linked data standards to express dependencies
- Using version control for data and code – overcoming usability challenges

SDSC

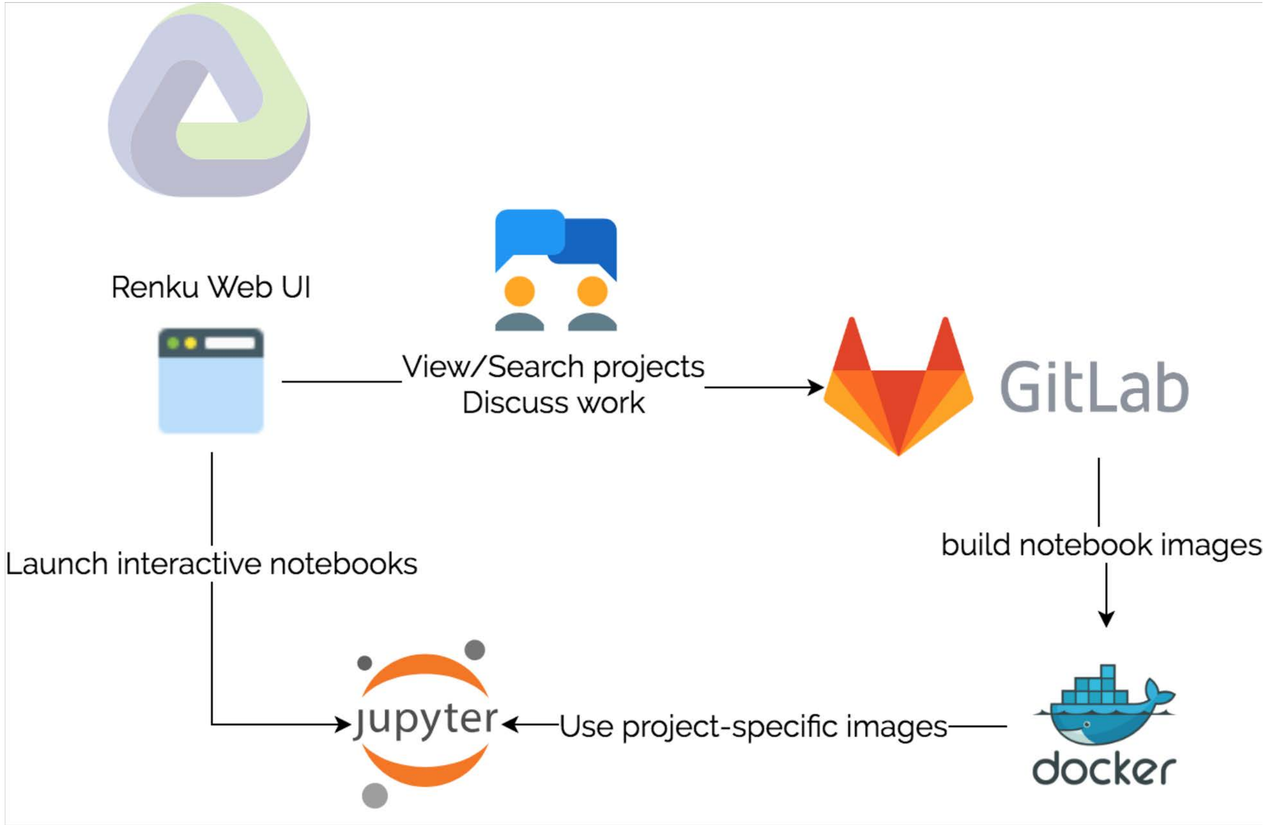
Reusability



- Understanding lineage means we can always re-execute and update results as methods or data change
- Workflow construction and re-execution possible with the local client
- Developing the means to execute workflows in the cloud and on HPC (REANA, Cromwell, Toil, …)
- Search for graph artifacts being developed

SDSC

Collaboration



- Web UI serves as primary point of contact for users
- Allows creation of projects, discussions with media embedding
- Creation of hosted interactive sessions with version-controlled environments (Docker images)
- Various supported languages, e.g. Python, MATLAB (with GUI), R, RStudio

Where is Renku used?

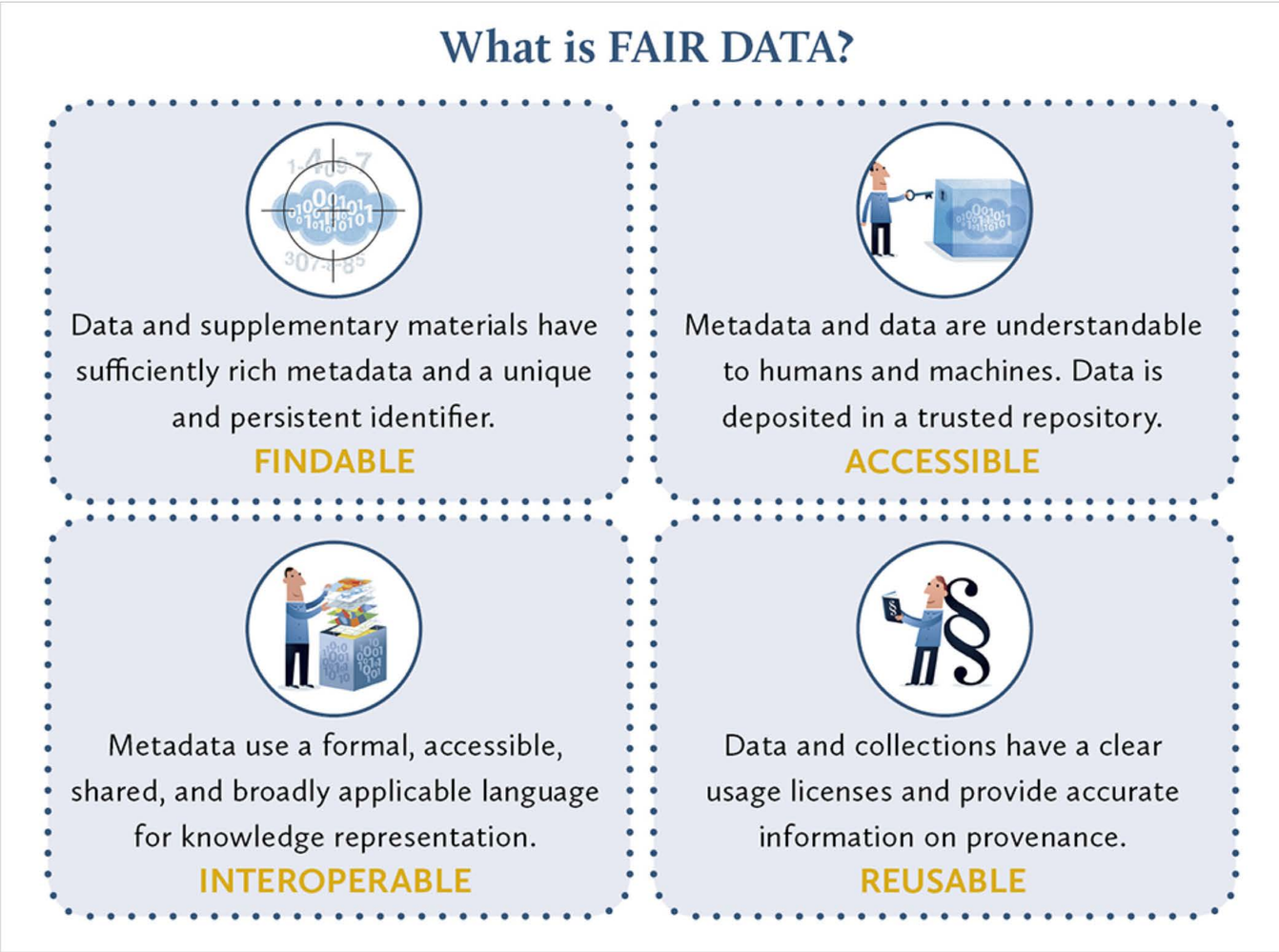
- We currently run 6 deployments, 4 are “production”
- Several SDSC academic projects are using RENKU from the start
- Large use-case from EPFL life sciences
  - reproducibility/collaboration
  - end-to-end lineage (from wet lab through computation)

Status and Roadmap Q3/Q4 2019

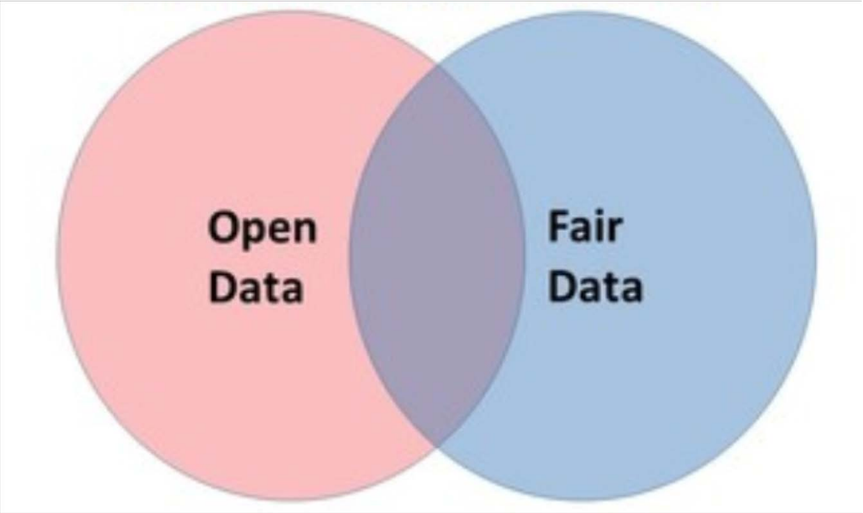
- Knowledge graph is live and first use by UI coming to renkulab.io
- Many UI/UX improvements, more coming
- Integrations for data import/export with external repositories e.g. Zenodo

- UI features for easier adoption e.g. environment configurations
- Tighter integration of UI with the knowledge graph, improving metadata
- Support for more environments, e.g. HPC
- Workflow execution in the cloud/HPC
- Solidifying our operations and deployment practices

Open Science ≠ Open Data



Source: LIBER



SDSC

Swiss Data Custodian

Swiss Data Custodian

- A multisided service to maximize economic opportunities and meet the societal challenges brought about by the digital revolution
- It is a data vault + secure multiparty compute ecosystems
  - Assign data ownership to the rightful person
  - Preserve data sovereignty
  - Establish trust and transparency in data sharing
  - Promote economic and societal incentives for sharing data
  - Enable cooperation between mutually non-trusting parties
- Governed by a trusted entity
  - Monitor and maintain compliance with DPA obligations

Thank you!

Renku is under very active development

Preview with tutorial at: https://renkulab.io

Open-source:
https://github.com/SwissDataScienceCenter

All feedback welcome!

https://datascience.ch/
@SDSCdatascience