Secure ANalysis Environment (SANE)

23 June 2022

Steven Claeyssens (KB/CLARIAH)
Lucas van der Meer (ODISSEI)
Martin Brandt (SURF)

Sharing research data

- Increasing desire to use data not made for research purposes

Sharing research data

- Increasing desire to use data not made for research purposes

Sharing research data

- Increasing desire to use data not made for research purposes

Foto: de Sleutels ©

Foto: Wikimedia ©

Sharing research data

- Increasing desire to use data not made for research purposes

Foto: de Sleutels ©

Foto: Wikimedia ©

Foto: Wikimedia ©

Sharing research data

- Increasing desire to use data not made for research purposes

Foto: de Sleutels ©

Foto: Wikimedia ©

Foto: Wikimedia ©

"I'm willing to share my data, but I can't" 🤷 → SANE

- McKinsey: ~1% of the world's data used for analytics purposes

Sharing research data

- Increasing desire to use data not made for research purposes

Foto: de Sleutels ©

Foto: Wikimedia ©

copyright
all rights reserved

Foto: Wikimedia ©

"I'm willing to share my data, but I can't” 🤷 → SANE

- McKinsey: ~1% of the world’s data used for analytics purposes

2 Mentimeter

3 Use cases

Data provider perspective: KvK

- Pseudonymised data
- Complete control over the data
- Data cannot leave environment
- Trust the research software
- Research purpose
- Approve data upload

- Tinker SANE

Foto: Wikimedia ©

Data provider perspective: KB

- Collections as data
- In copyright
- Non-consumptive use
- Research purpose
- Data upload allowed
- Software upload allowed
- Review any output

- Blind SANE

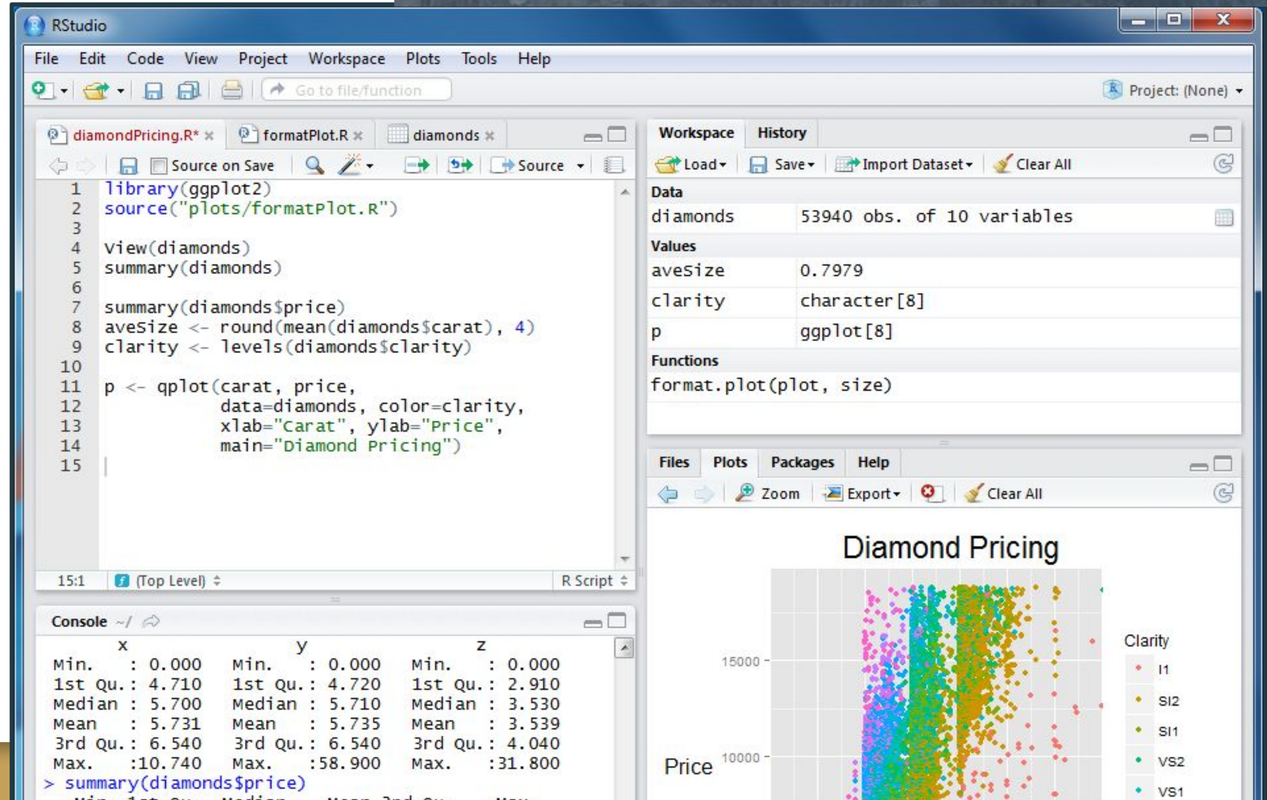
Foto: Wikimedia ©

Researcher perspective: KvK

“To what extent does the proportion of part-time employees affect firm closure?”

- Combine it with my own data
- Play (Tinker with the data)
- Specific characteristics of the combined data determine consequent analytical steps.
- Use R, Python

Foto: Wikimedia ©



Researcher perspective: KB

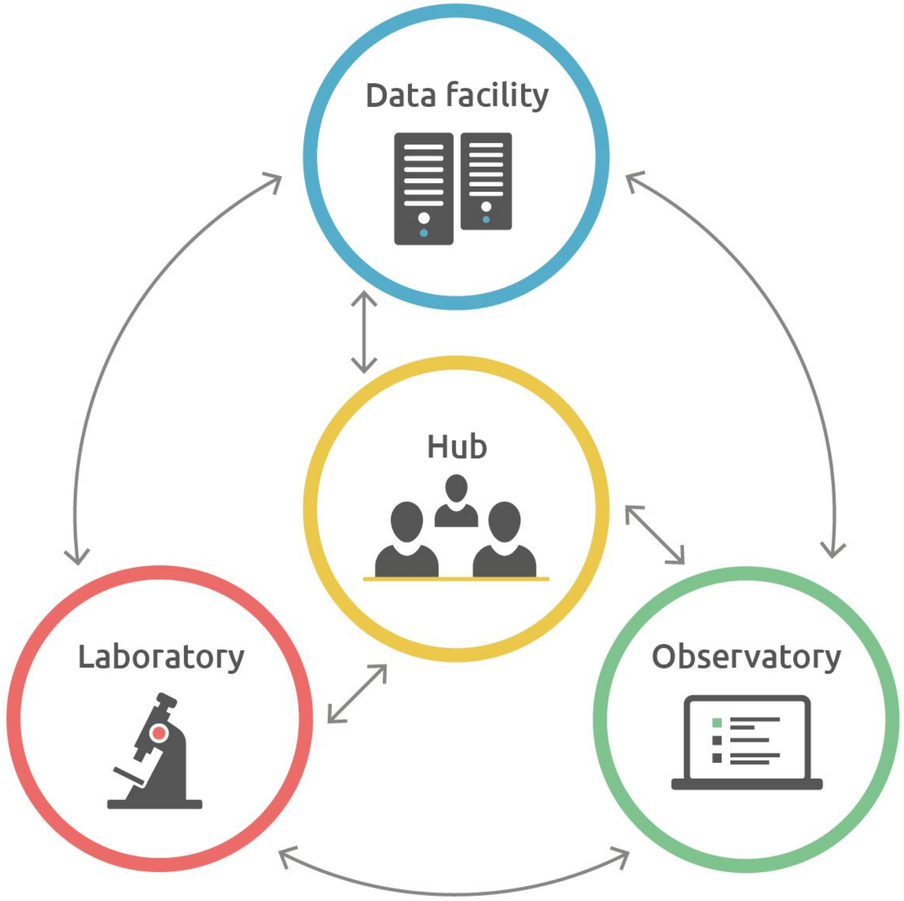
- Mine the KB collections
- Use existing software and algorithms, e.g. NLP pipelines
- Use my own software or algorithm
- Upload a container
- Extract and download non-copyrighted data

Foto: Wikimedia ©

A national infrastructure for Social Science

ODISSEI creates a federated data infrastructure for the social and economic sciences in the Netherlands, on behalf of more than 40 member organisations.

The ODISSEI research infrastructure



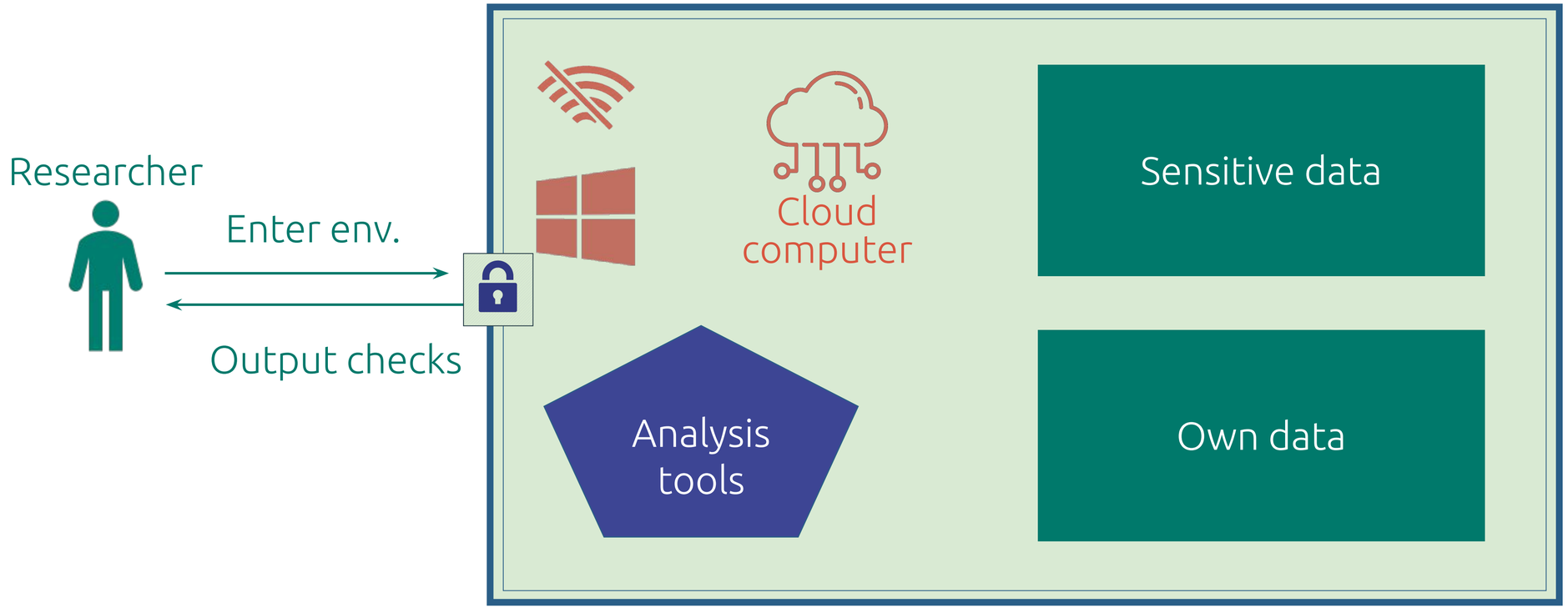
CLARIAH

A distributed research infrastructure, part of the National Roadmap for Large-Scale Research Facilities (NWO)

CLARIAH, as a national counterpart to CLARIN and DARIAH, is directly and deeply embedded in the Europe-wide ESFRI enterprise. CLARIN and DARIAH are the only two humanities infrastructures under development on the ESFRI Roadmap, both projects were on the 2008 national roadmap, and they joined forces in CLARIAH, which was put on the 2012 national roadmap.

4 SANE

Secure ANalysis Environment (SANE)



Five Safes

Afbeelding: UKDS ©

Customisations

- Blind
- API connection to Data owner
- Temporary environment
- …

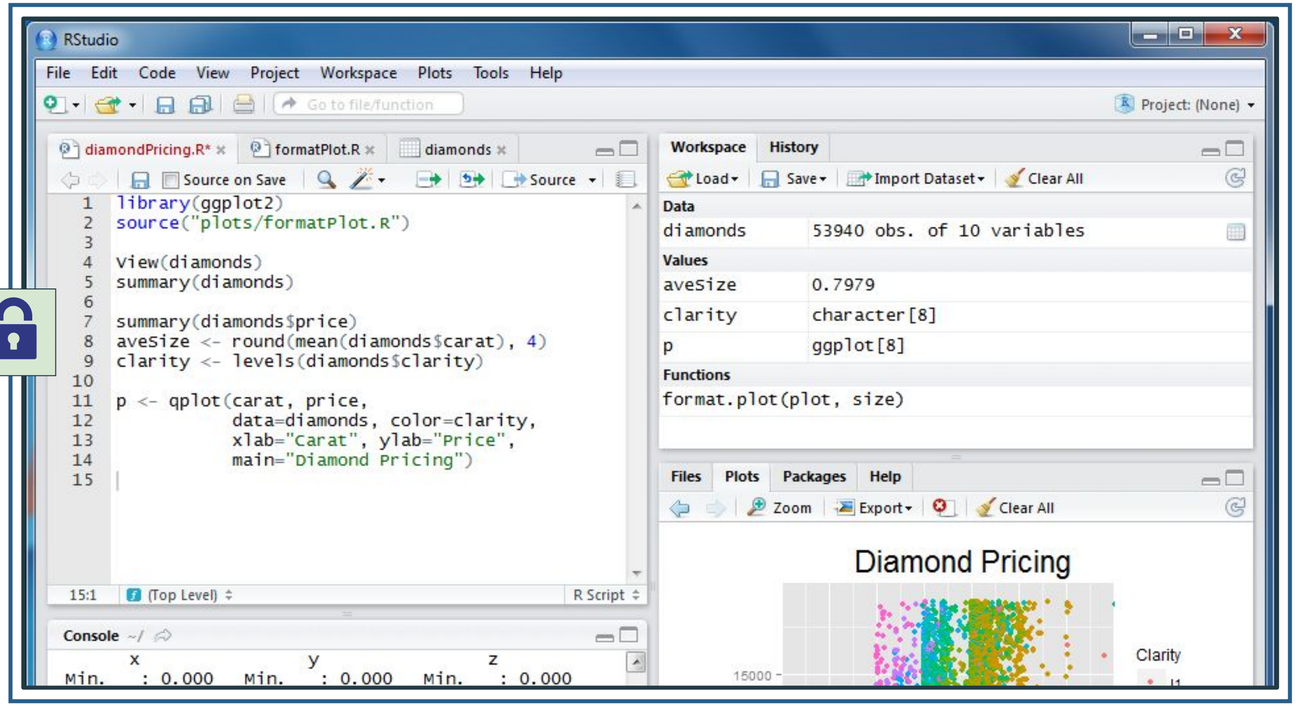
5 Similar initiatives

CBS Remote Access Environment

Researcher

Input

Output checks



Hathi Trust

Data Capsules

Secure virtual environments for non-consumptive text analysis, where researchers can implement their own data analysis and visualization tools.

Switch to Secure mode

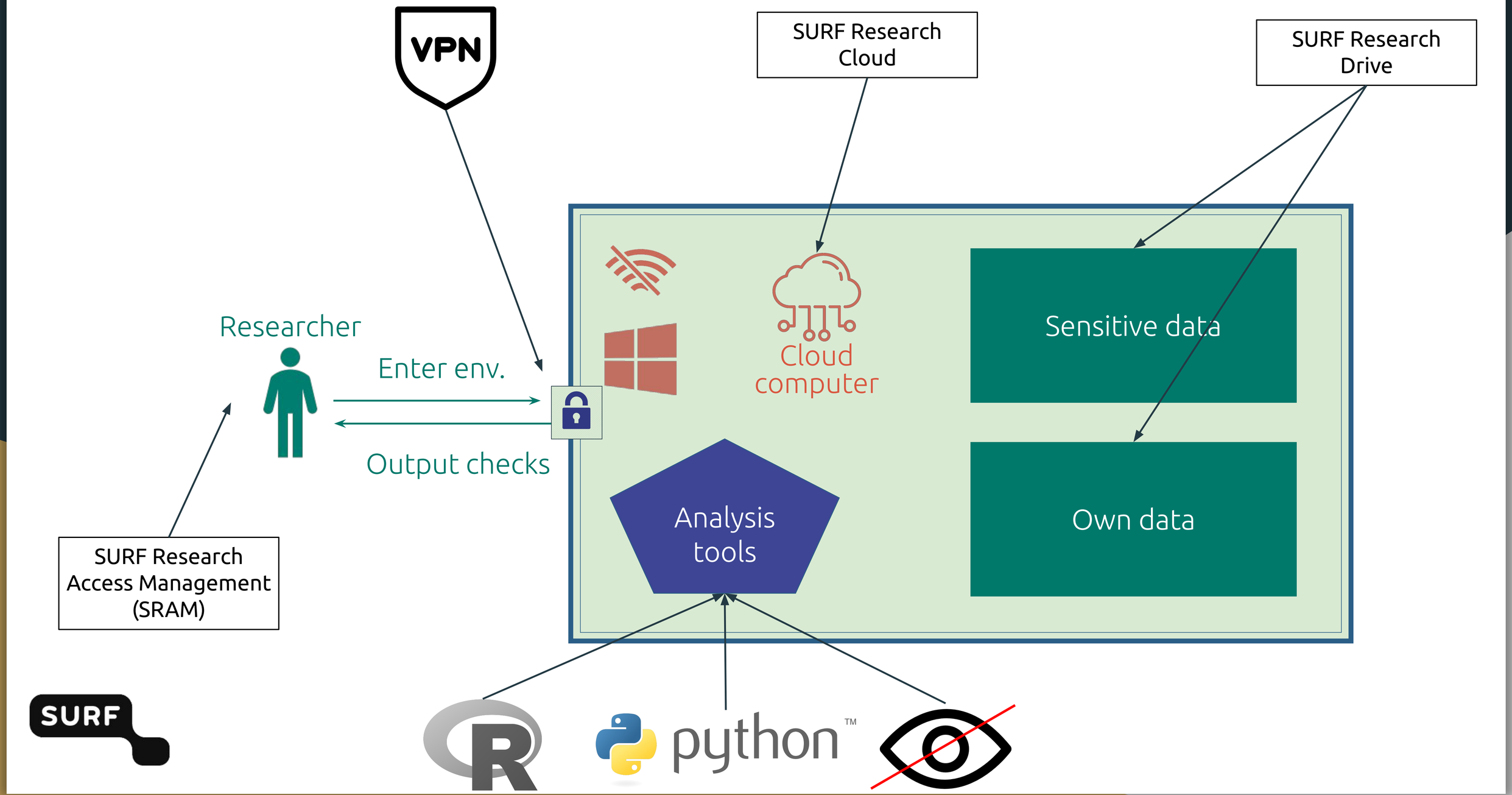
Network access in Secure mode is limited. You will be able to access the HTRC Data API only. While in Secure Mode, save your data or results to the Secure Volume. Data saved elsewhere in your Capsule **will be lost** when you switch your Capsule back to the Maintenance mode.

Your research practices are expected to comply with the HTRC's Policy of Non-consumptive Research, and your capsule is not to be used for close reading. HathiTrust text data, page images, or metadata records from HTRC may not be copied from a capsule, and only approved data exports will be released from a capsule.

Review terms of use.

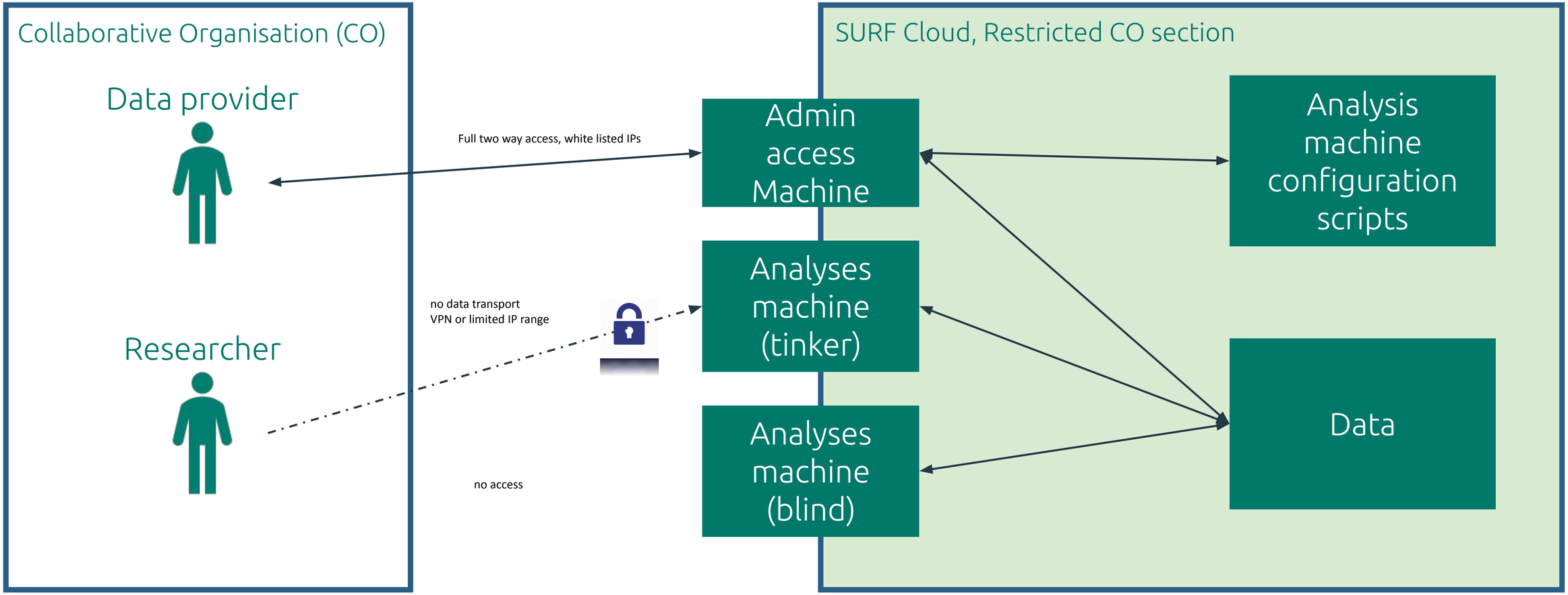
Agree & Switch Mode | Cancel

6 Available building blocks


VPN
SURF Research Cloud
SURF Research Drive
Researcher
Enter env.
Output checks
Cloud computer
Sensitive data
Own data
Analysis tools
SURF Research Access Management (SRAM)
SURF
python™

7 Technical blueprint

Blueprint using SURF Research Cloud (SRC) platform



Researcher flow for starting an analysis environment

1. Researcher is invited into the CO of the project based on Federated Identity
2. Researcher enters SRC portal, selects catalog item, and starts it with his budget
3. SRC platform starts VM and applies configuration based on script
4. After applying configuration SRC closes environment from internet
5. SRC platform links to CO managed data in cloud
6. Analysis phase
    a. (tinker) Researcher gets access over secure connection that prevent data transport
    b. (blind) Algorithm of researcher is started on the data
7. Results are saved to CO managed storage, either by user or by algorithm
8. Researcher requests Data provider to give him the results

Analysis machine features

- Network access control on cloud level
  - Security groups setting port and IP restrictions
  - Provisioned by SRC platform
  - Based on configuration scripts managed by Data provider
- Researcher has control for starting, stopping
- Researcher can request extra packages from Data provider
  - These packages are added to the configuration scripts by Data provider
  - Machine doesn’t need access to internet after configuration phase