Introduction to SANE

Final EPI conference, 7 March 2024

Lucas van der Meer
Ahmad Hesam
Martin Brandt
Annette Langedijk
Freek Dijkstra

Agenda

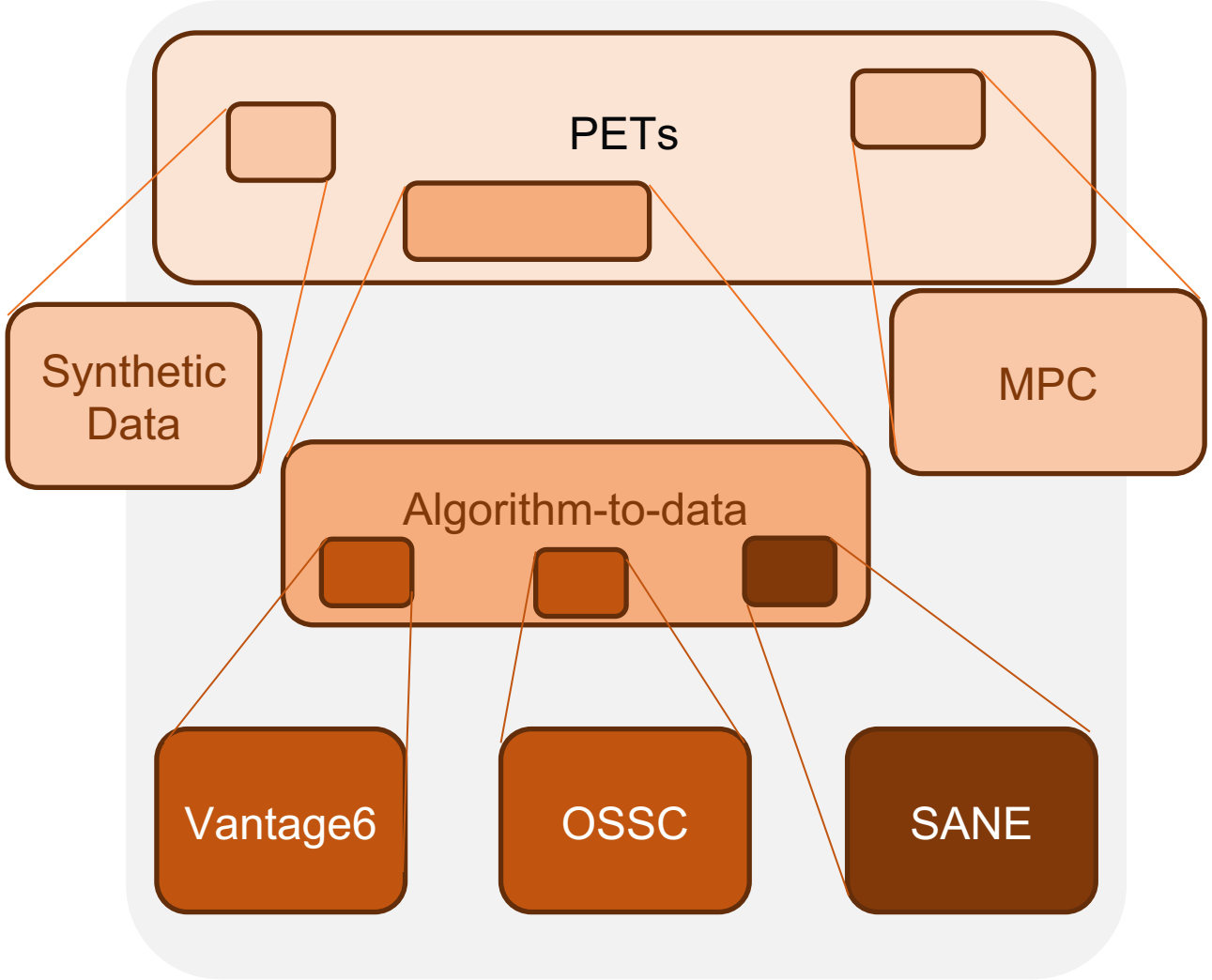
01
SSH Data sharing challenges

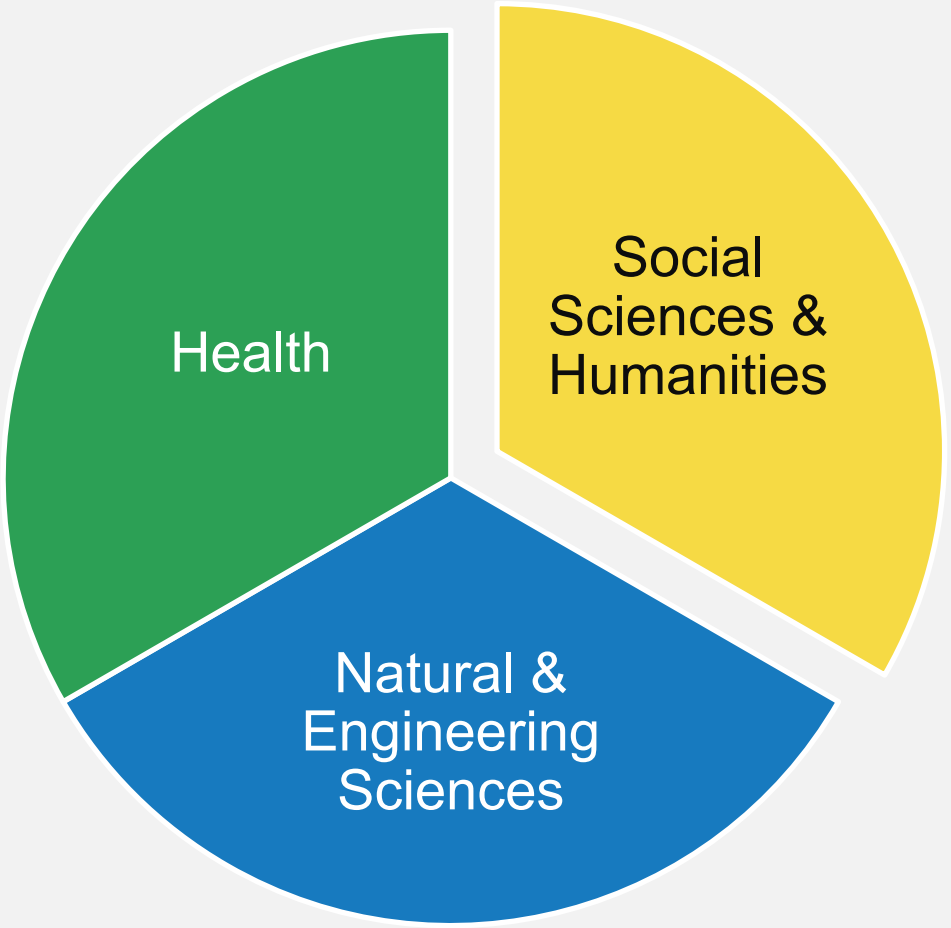
02
What is SANE?

03
Relation to other projects

04
Q&A

Scope


PETs
Synthetic Data
MPC
Algorithm-to-data
Vantage6
OSSC
SANE


Health
Social Sciences & Humanities
Natural & Engineering Sciences

data sharing challenges in social sciences and humanities

Non-research data has great scientific value, but providers often struggle to share

Privacy

Copyright

Competition

Typical data provider would need (1)

Complete control over the data:

- Data not sent to researcher
- “Non-consumptive use”
- Review any output

Pseudonymised data

Data upload not allowed

Check on researcher & purpose

Trust the research software

Typical data provider would need (2)

Additional requirements:

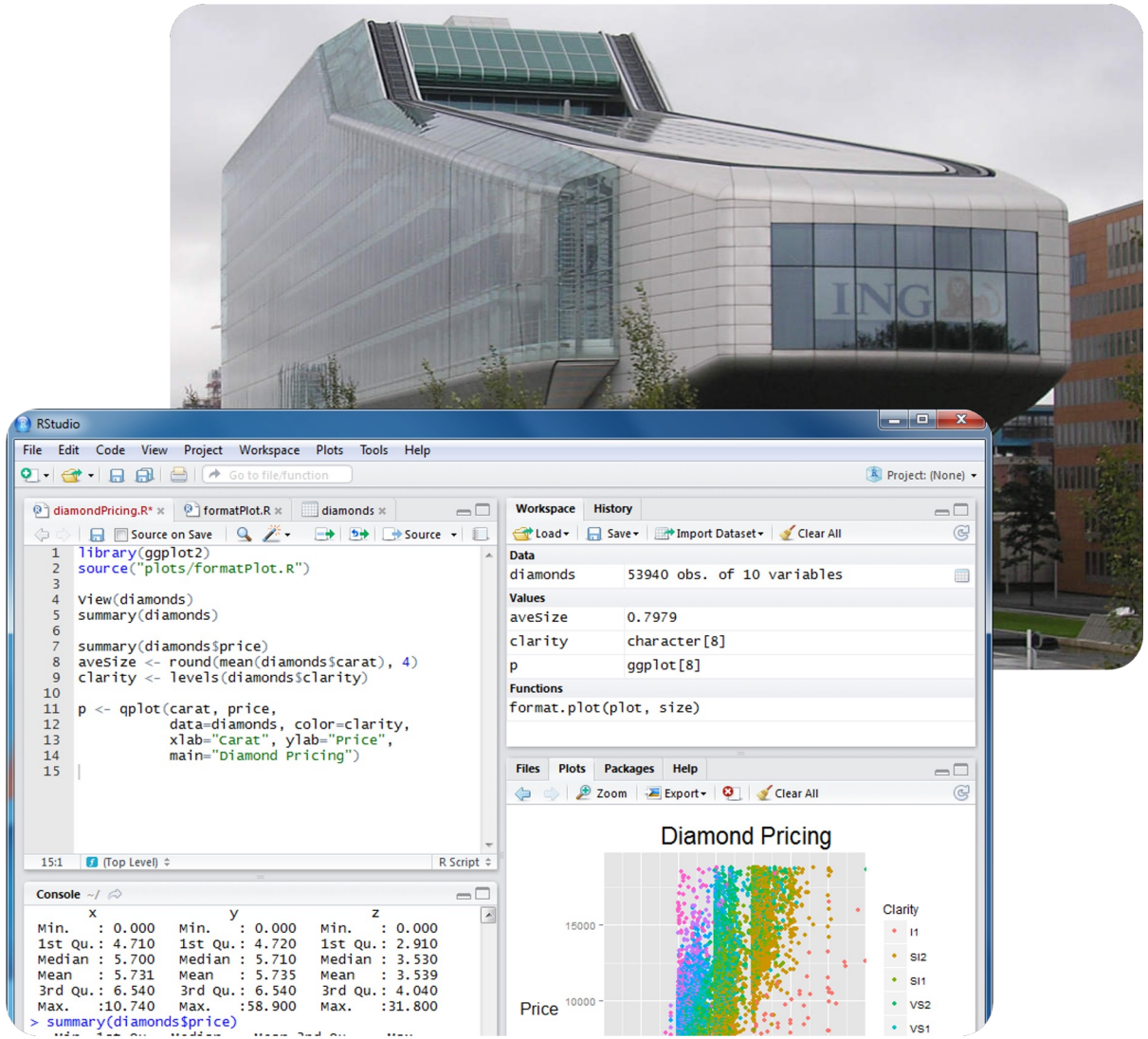
- Researcher cannot see the data
- Data & software upload allowed

Image: Wikimedia ©

Researcher would need (1)

“To what extent does PIN transaction match official inflation statistics?”

- Play (tinker) with the data
- Specific characteristics of the combined data determine consequent analytical steps
- Combine it with my own data
- Extract and download aggregated non-sensitive results
- Use R, Python



Researcher would need (2)

“How can a publisher evaluate the impact of converting any backlist ebooks titles?”

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

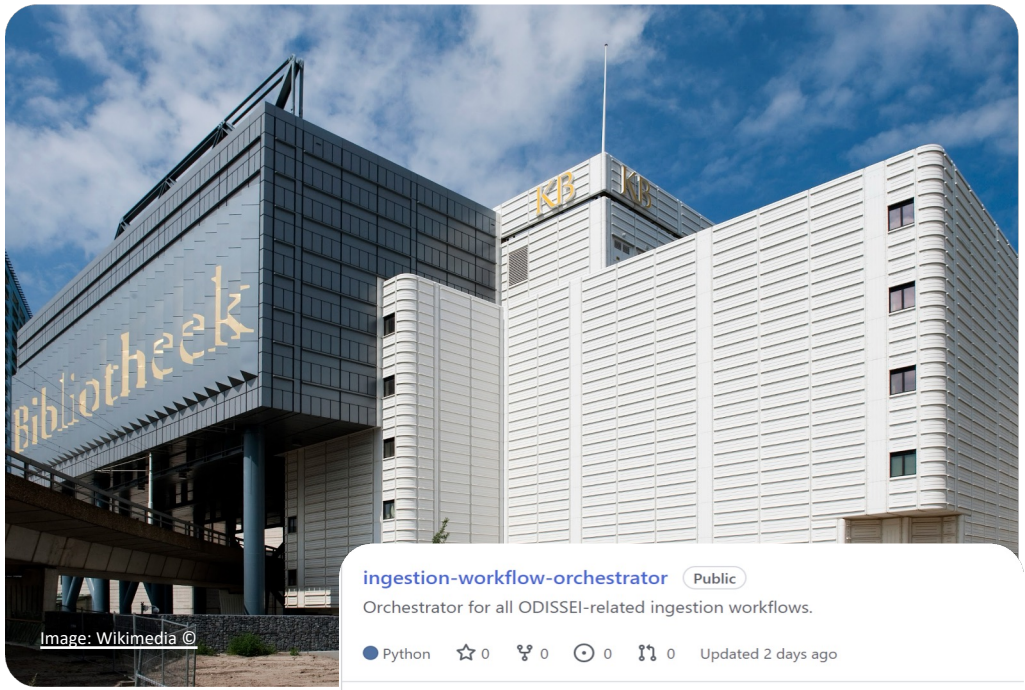
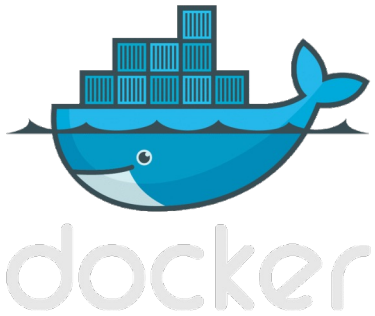


Image: Wikimedia ©



Best practice:
Five safes

Office for National Statistics

UK • DATA ARCHIVE

SANE:
Secure ANalysis Environment

What is SANE?

SANE – Secure ANalysis Environment

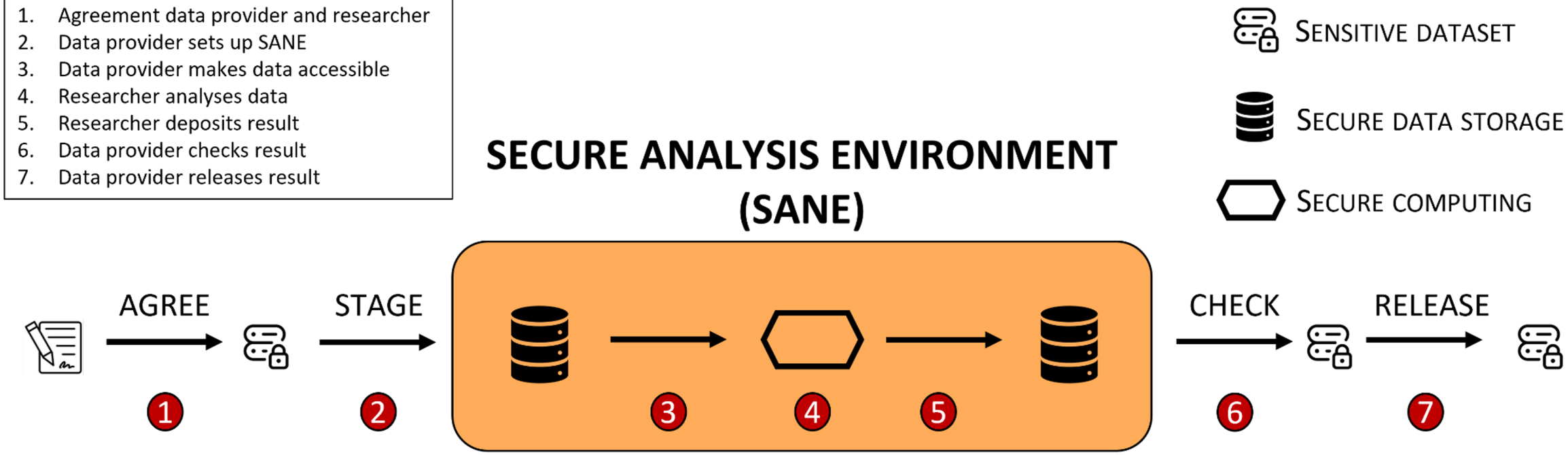
- Virtual, fully shielded computing environment
- Enable access to the sensitive data
- Pre-approved analysis software
- Data provider maintain complete control
- Penetration tested
- Runs on ISO-27001 certified SURF Research Cloud

Tinker SANE: allows the researcher to see, manipulate and play with the data.

Blind SANE: the researcher submits an algorithm without being able to see the data and the data provider approves the algorithm and output.

SANE – Workflow

STEPS
1. Agreement data provider and researcher
2. Data provider sets up SANE
3. Data provider makes data accessible
4. Researcher analyses data
5. Researcher deposits result
6. Data provider checks result
7. Data provider releases result



Information Security Management ISO 27001 Certified

Screenshots ~~Demo~~

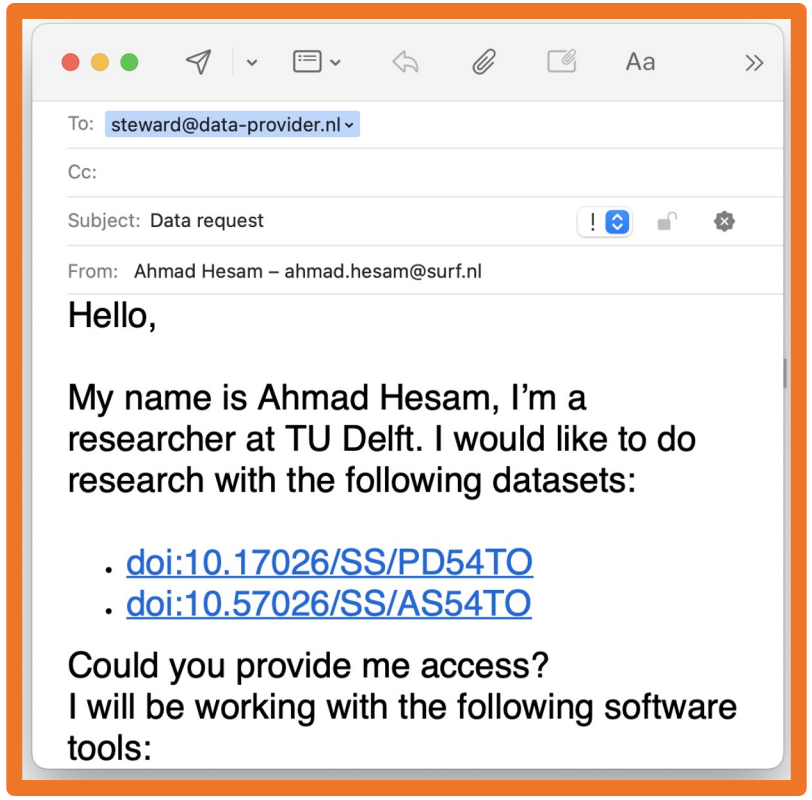
Roles

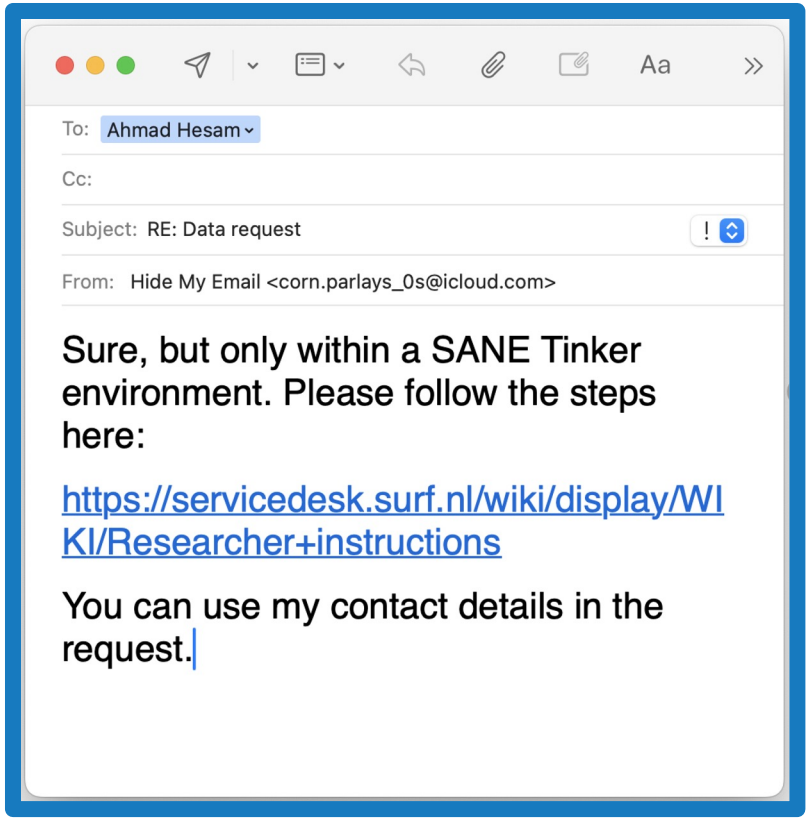
Data Provider

Researcher

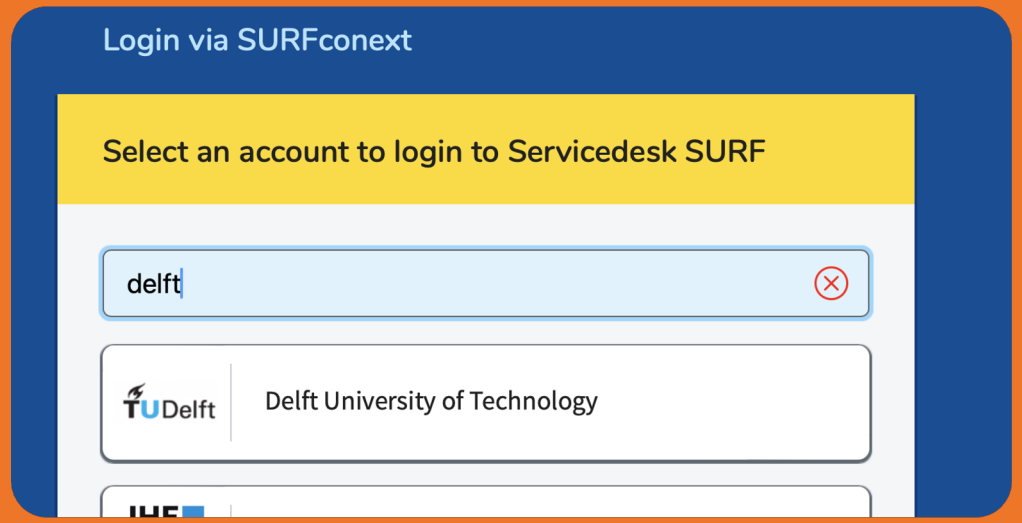
SURF Advisor

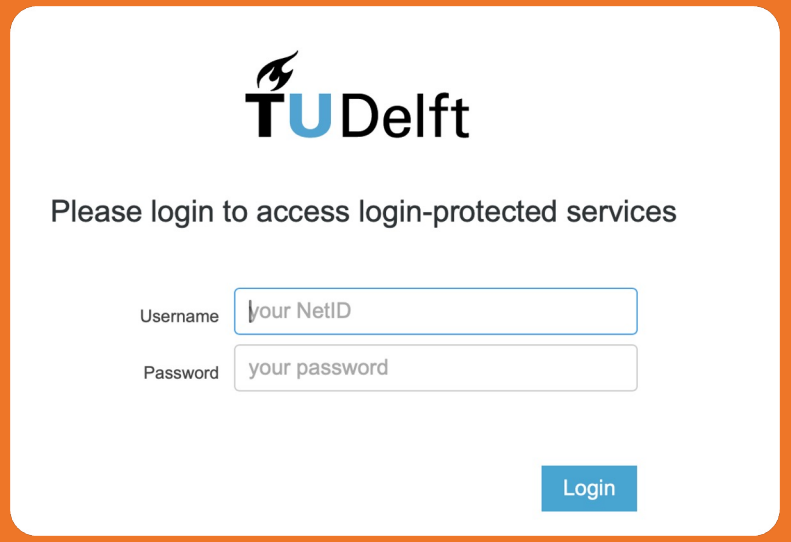
Data Agreement Step

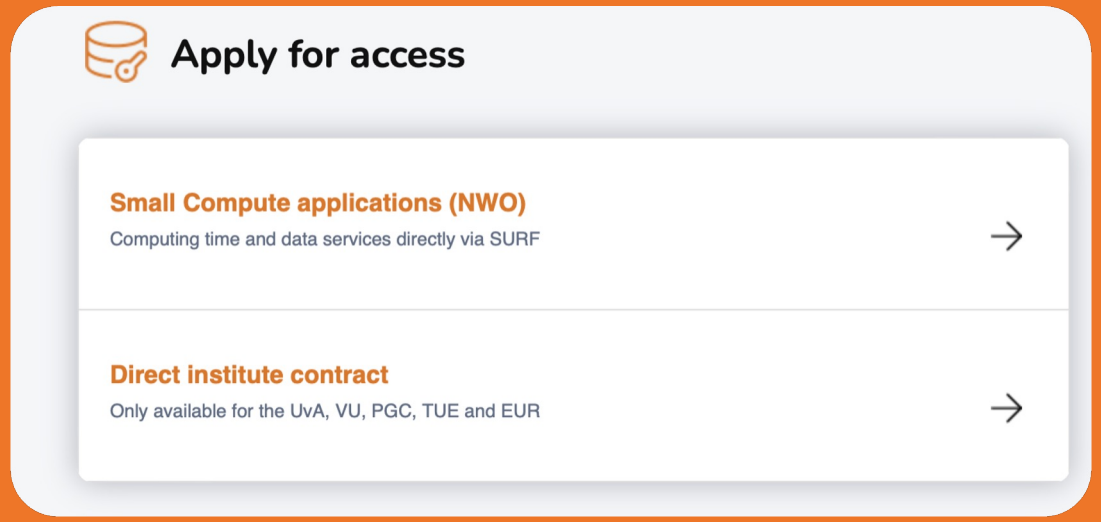




SANE Project Request







SANE Project Request

Title of the project

[SANE] Socio-economic agent-based modeling

Description

Aa

Dear Servicedesk,

In the context of SANE on SURF Research Cloud I wish to request the following collaboration:

Name: Socio-economic ABM

Description
Using the following datasets I wish to calibrate an agent-based model to discover new socio-economic phenomena....

<rest of description>

Data provider contact person: Martin Brandt, steward@data-provider.com

Kind regards,
Ahmad Hesam

Technical project requirements

I expect to need 100GB of storage for the datasets, and about 10GB for my results, so 200GB in total should be enough.

I will need at least 20 (iterations) x 500 (core-hours per iteration) = 10k core hours to perform my analysis.

Resources

- [ ] Snellius
- [x] Research Cloud - HPC Cloud
- [ ] Data processing - Grid
- [ ] Data processing - Spider

Signing Authority e-mail

john.doe@tudelft.nl

Recognizable institutional e-mail required

Signing Authority Staff Position within the organization

Professor

https://www.surf.nl/en/themes/research-infrastructure/sane-secure-environment-for-analysing-sensitive-data

SURF processes request

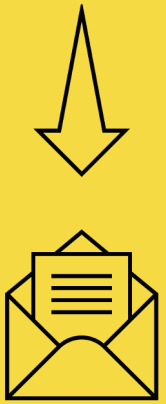
Dear data provider,

You have been invited to the SANE

CO: “Socio-economic ABM”

Please accept this invite to start the

SANE project



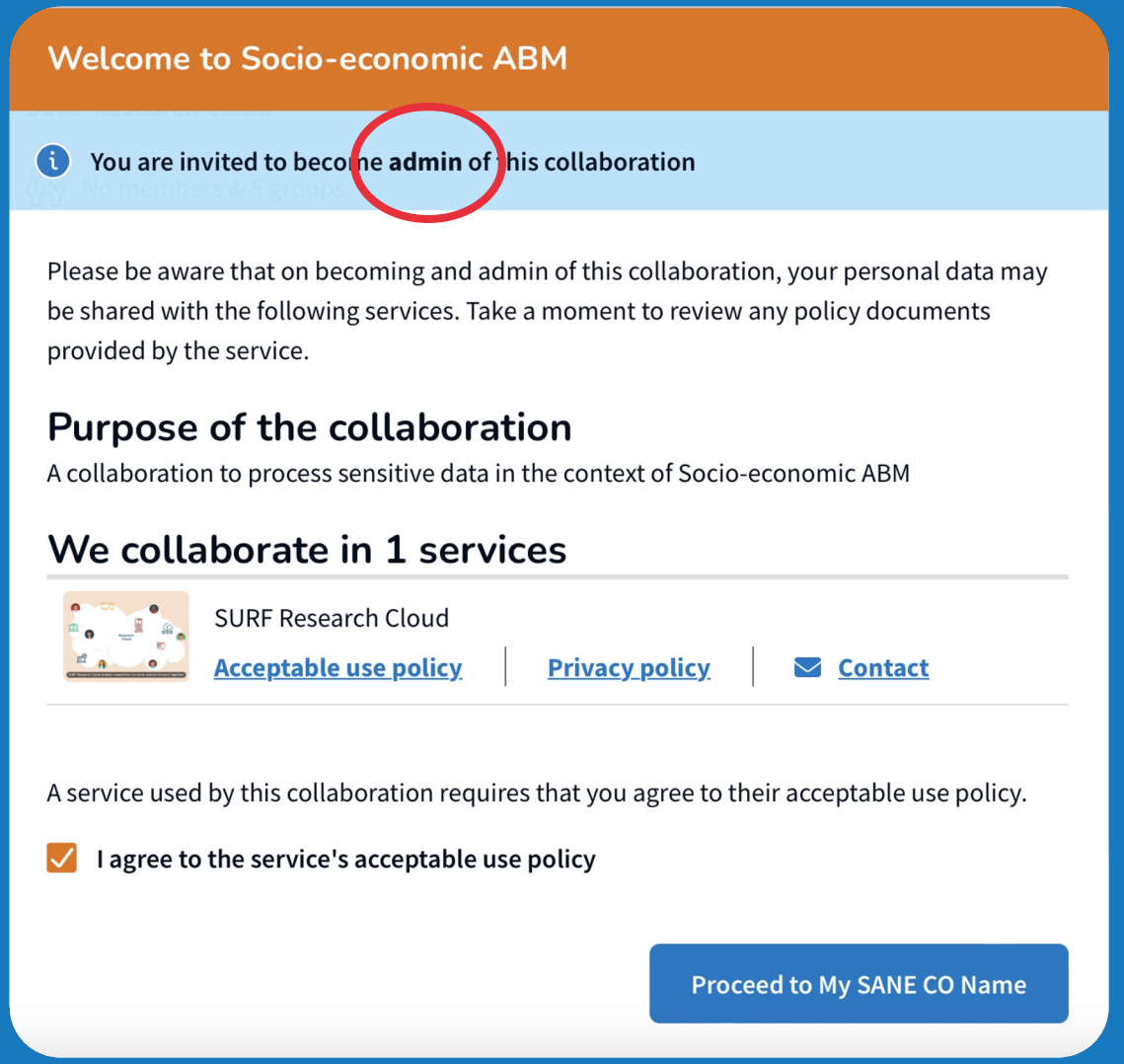
Accept invitation to CO

Dear data provider,

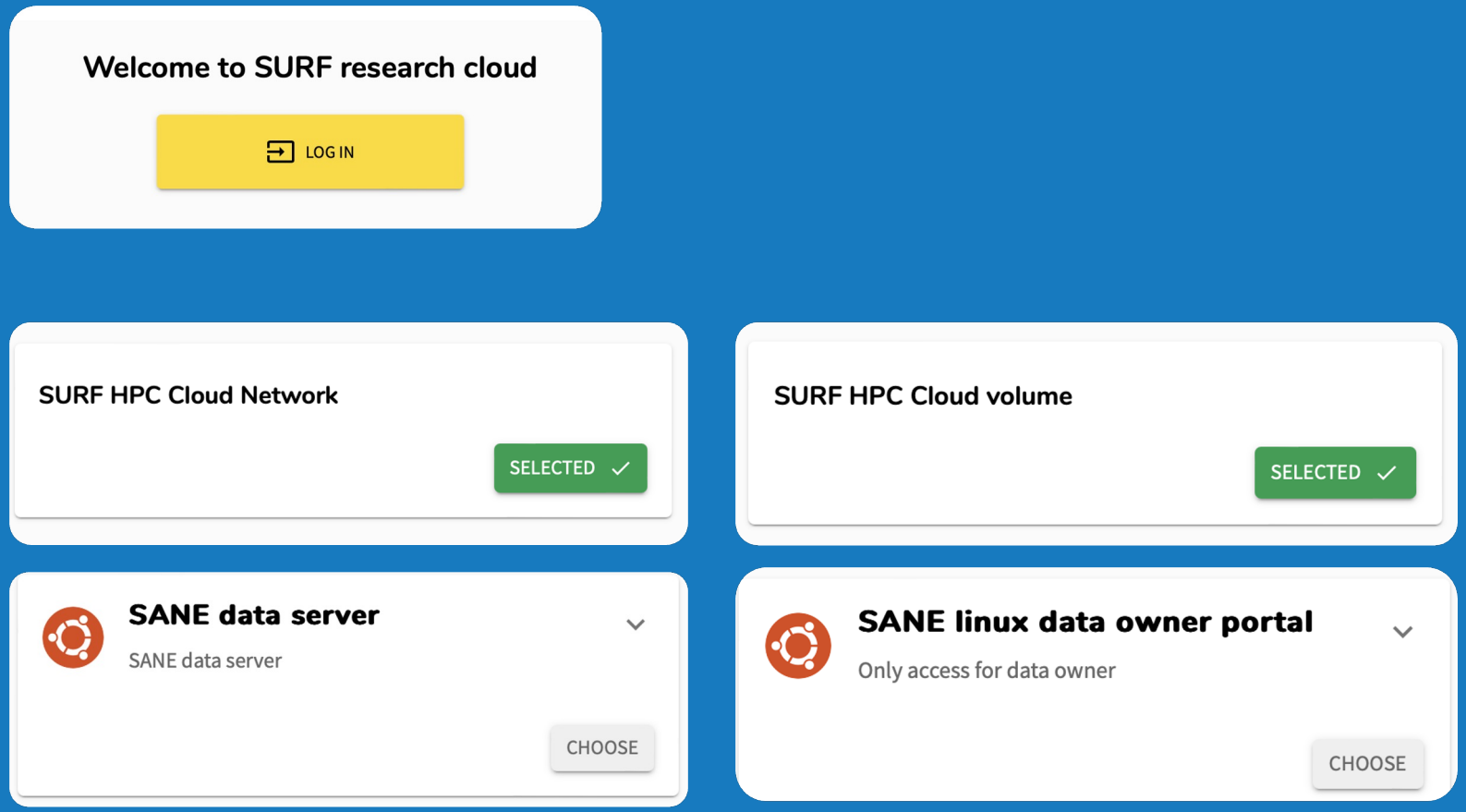
You have been invited to the SANE CO: “Socio-economic ABM”

Please accept this invite to start the SANE project

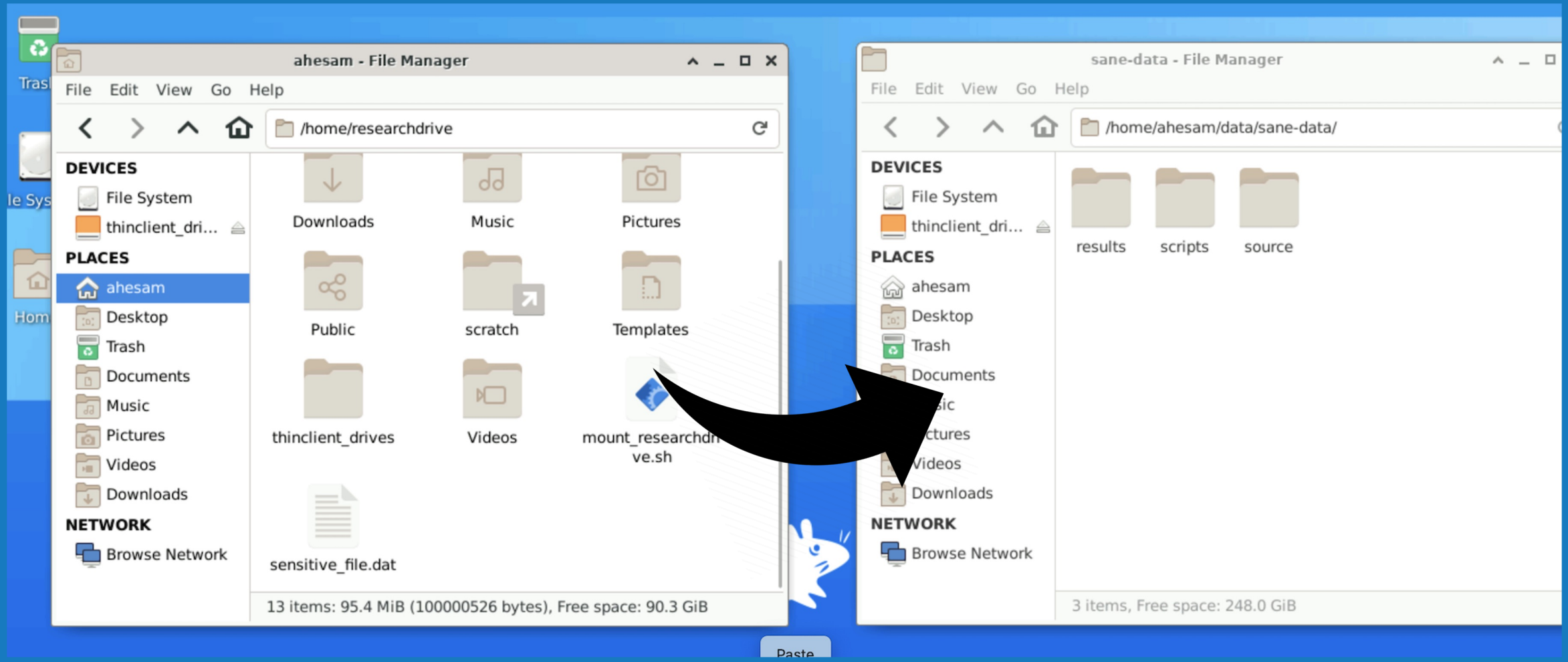
Accept



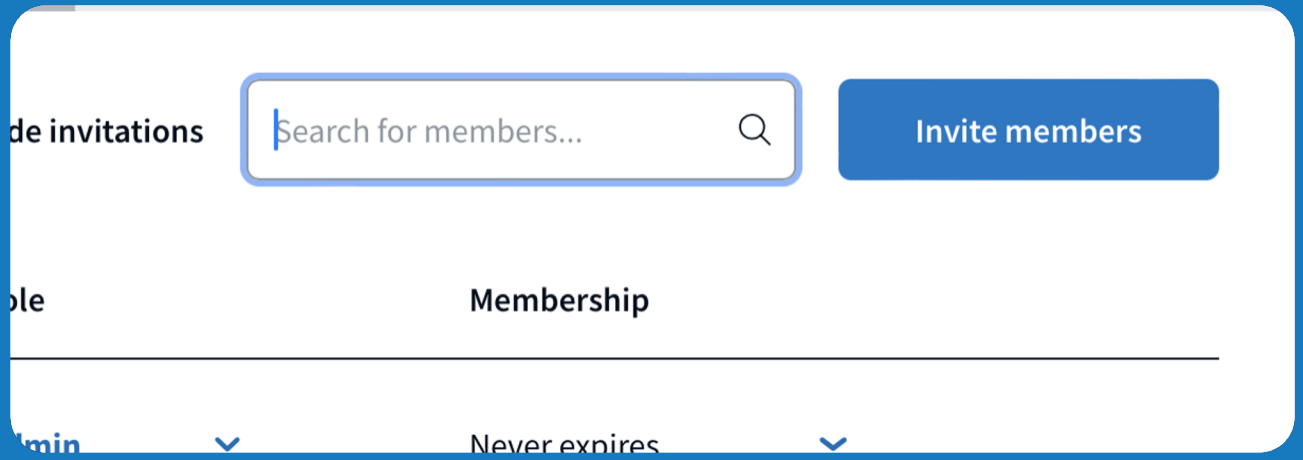
Setup SANE environment



Upload sensitive data



Invite Researcher



Invitation details

Invitees

Invitee email addresses

Role

Regular user

Group membership

Select groups

Membership valid until

Message

Personal message to the invitees

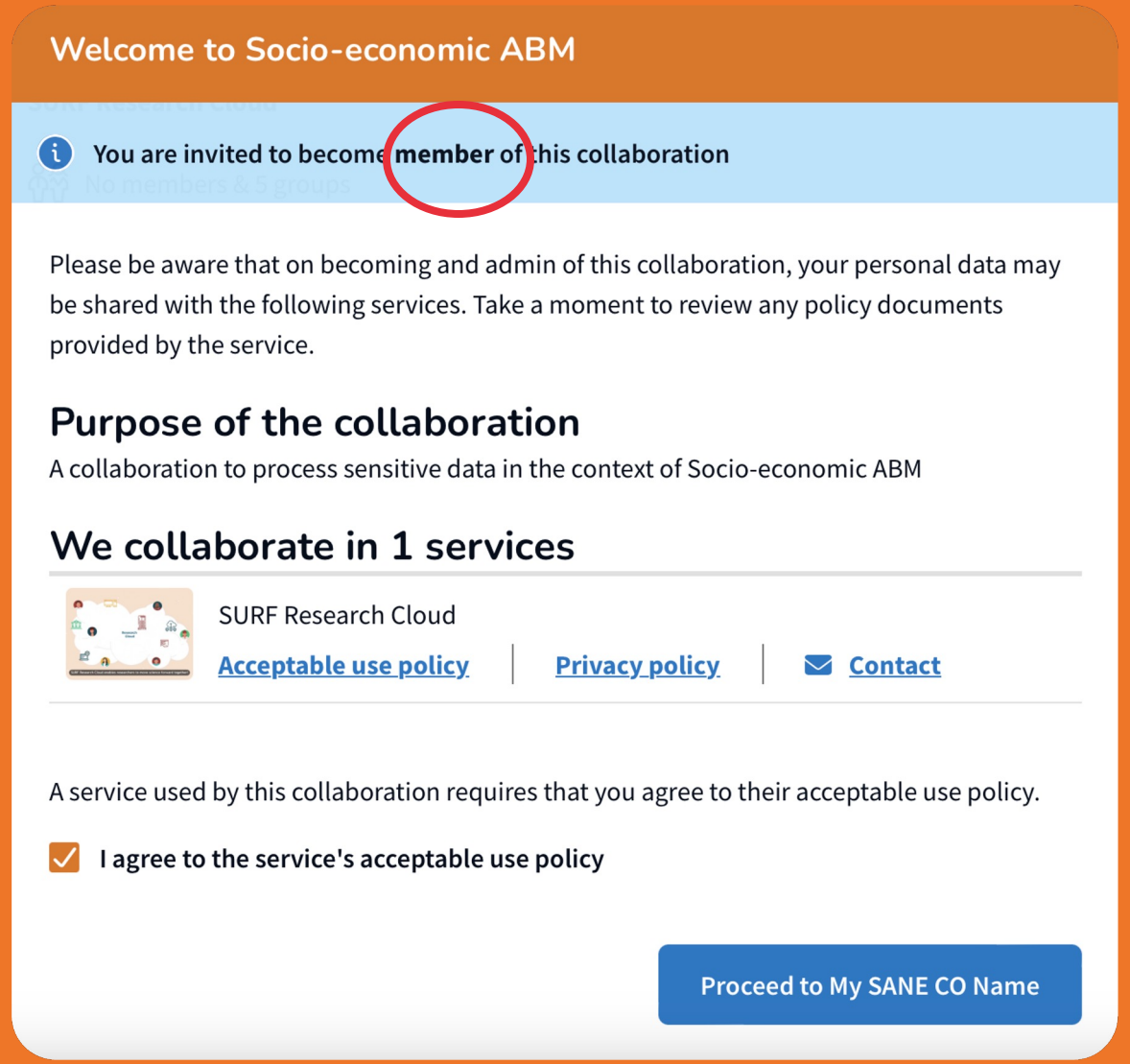
Accept invitation to CO

Dear researcher,

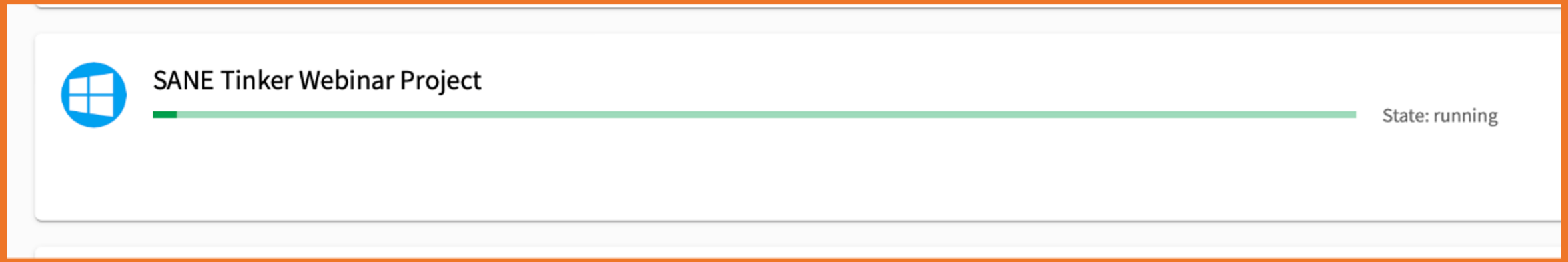
You have been invited to the SANE CO: “Socio-economic ABM”

Please accept this invite to start the SANE project

Accept

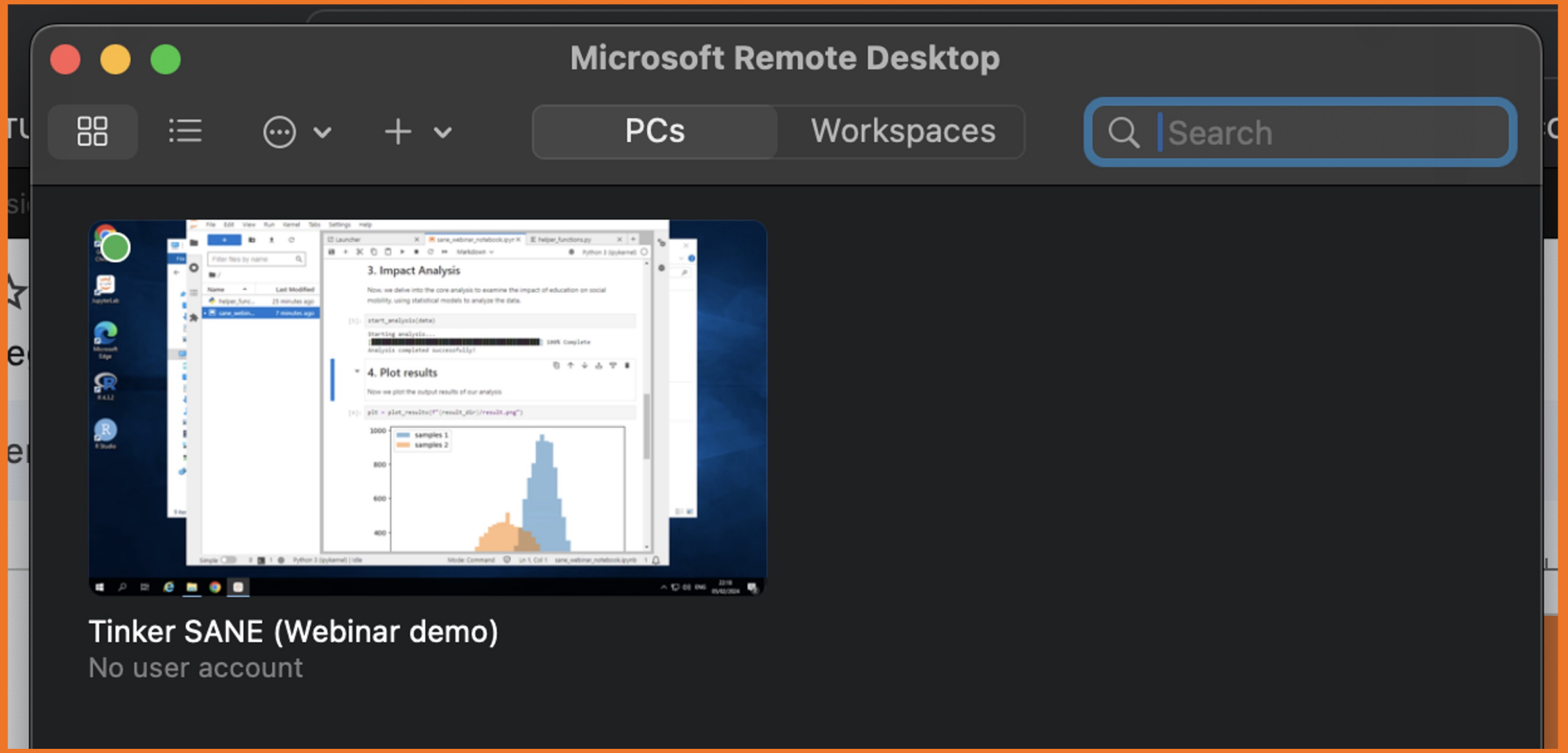


Start SANE workspace

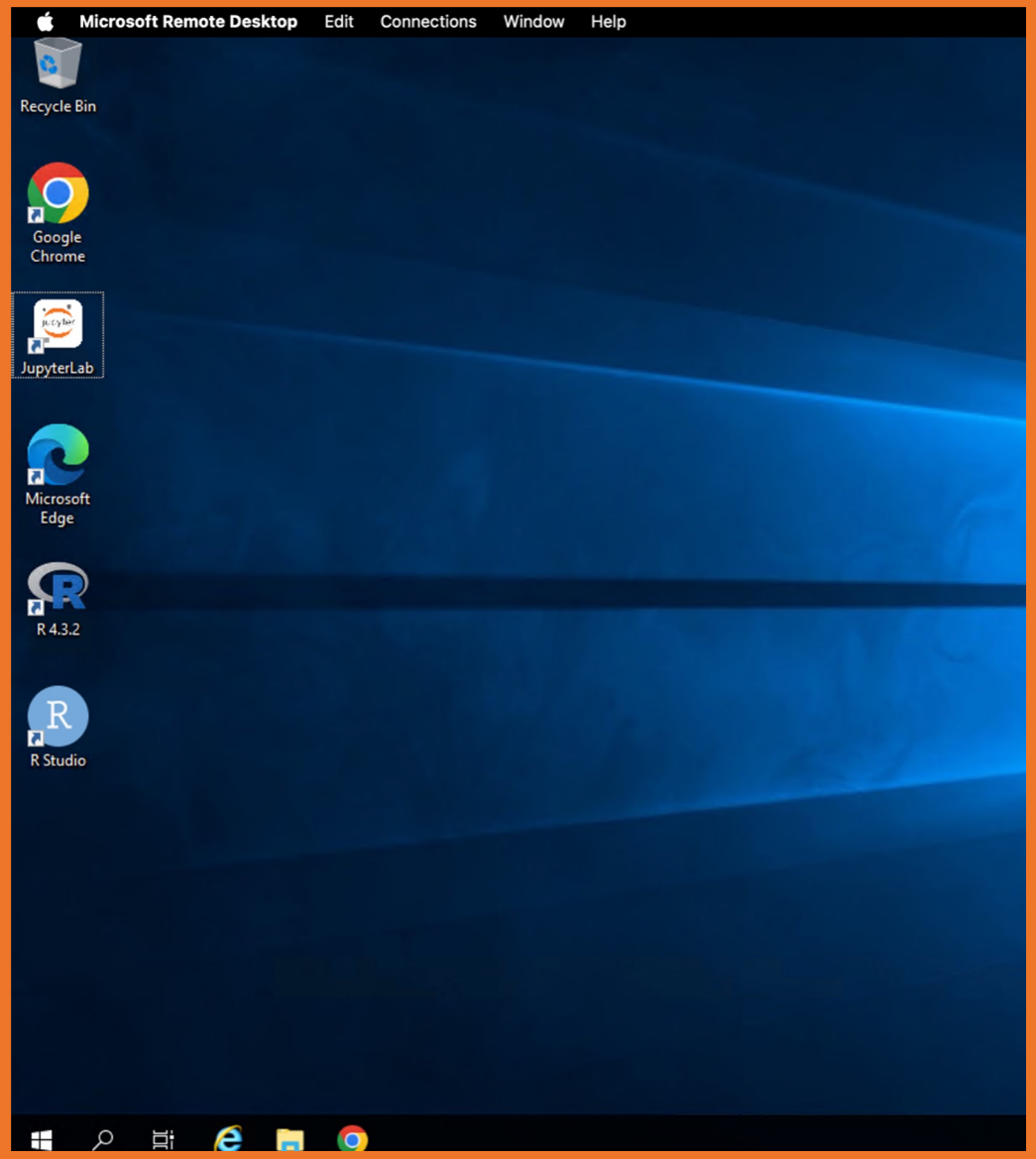


Use your favorite Remote Desktop Client to access your Tinker SANE workspace

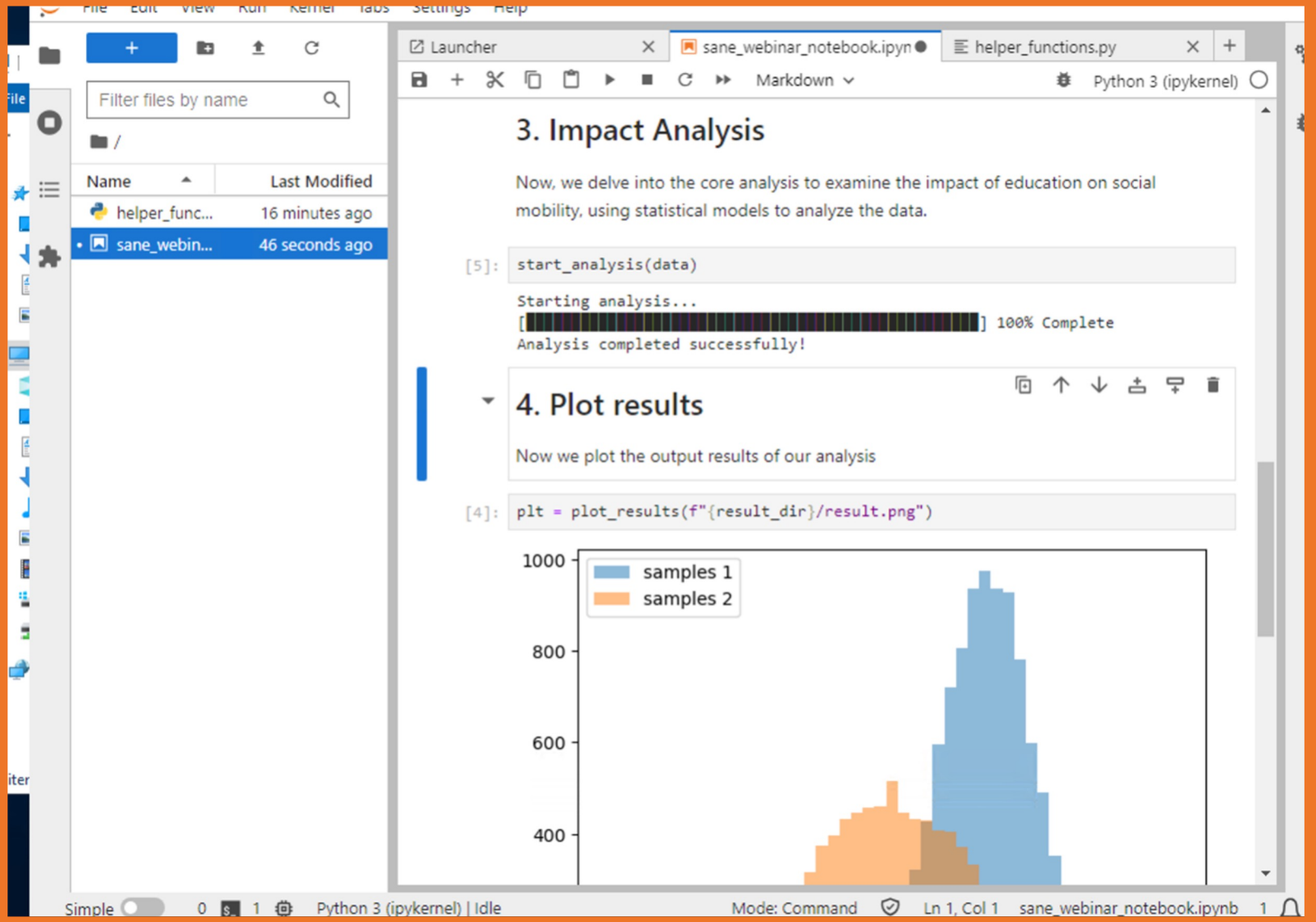
Tinker SANE workspace



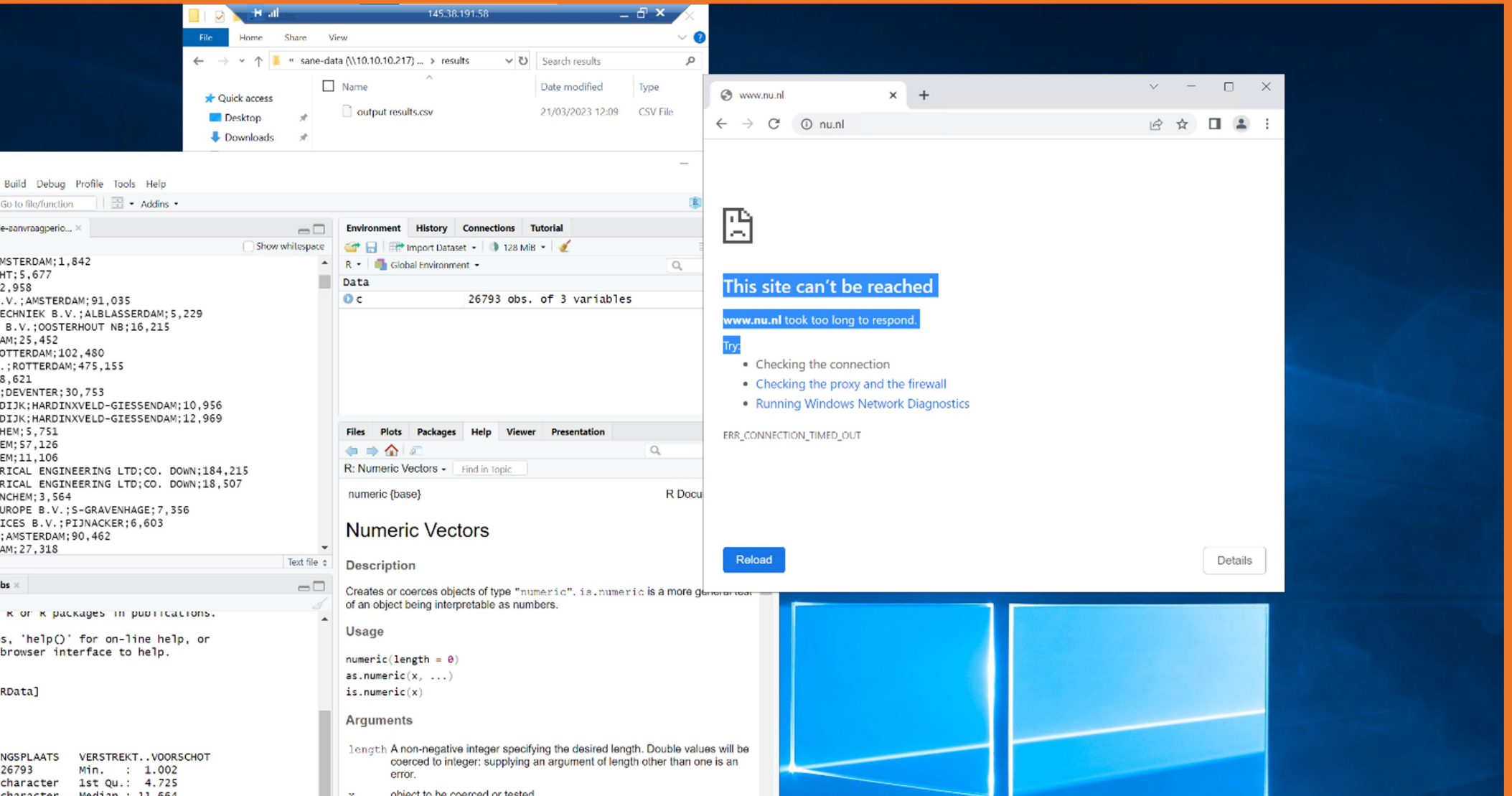
Tinker SANE workspace



Perform analysis in Tinker SANE



Closed Down Environment



Start Blind SANE workspace

1. Choose the expiration date of the machine

10-02-2024 (22:16)

2. Workspace name, domain name and description

Name

SANE Blind Webinar Project

Hostname

saneblindwebin

Description

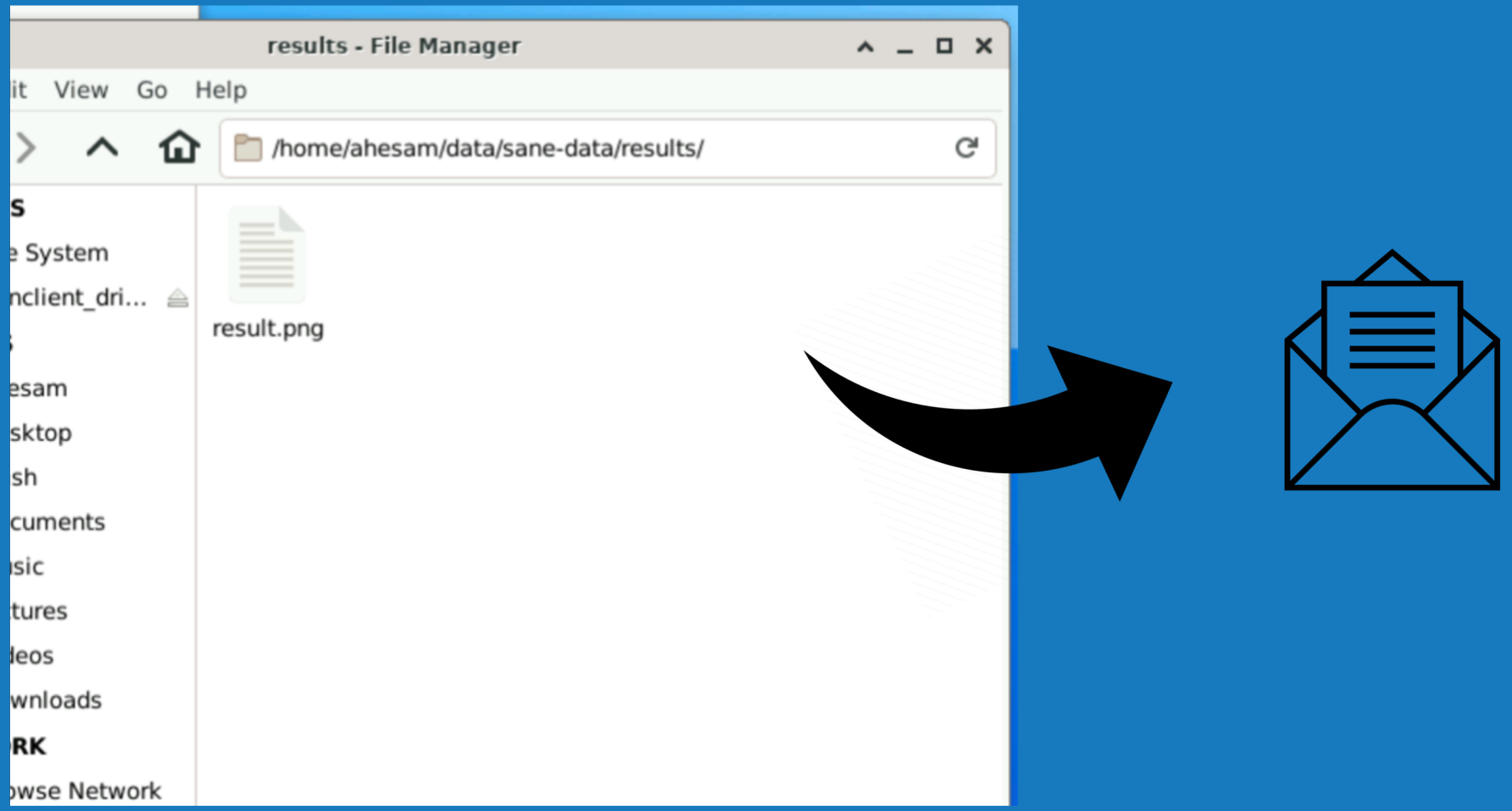
Description

3. Workspace parameters

Script location: (script subfolder or GIT repository url)

Script location

Inspect and release output results



Relation to other projects

Use Cases:

FirmBackbone

Open-science principles and restricted data

Peter Gerbrands, Wolter Hassink,
Daniel Oberski, Rutger Schilpzand,
Arjen van Witteloostuijn

Universiteit Utrecht

VU VRIJE UNIVERSITEIT AMSTERDAM

PDI SSH Platform Digitale Infrastructuur Social Sciences & Humanities

ODISSEI

Use Cases

FirmBackbone

Universiteit Utrecht

VU VRIJE UNIVERSITEIT AMSTERDAM

PDI SSH Platform Digitale Infrastructuur Social Sciences & Humanities

Lisa Het werkgelegenheidsregister van Nederland

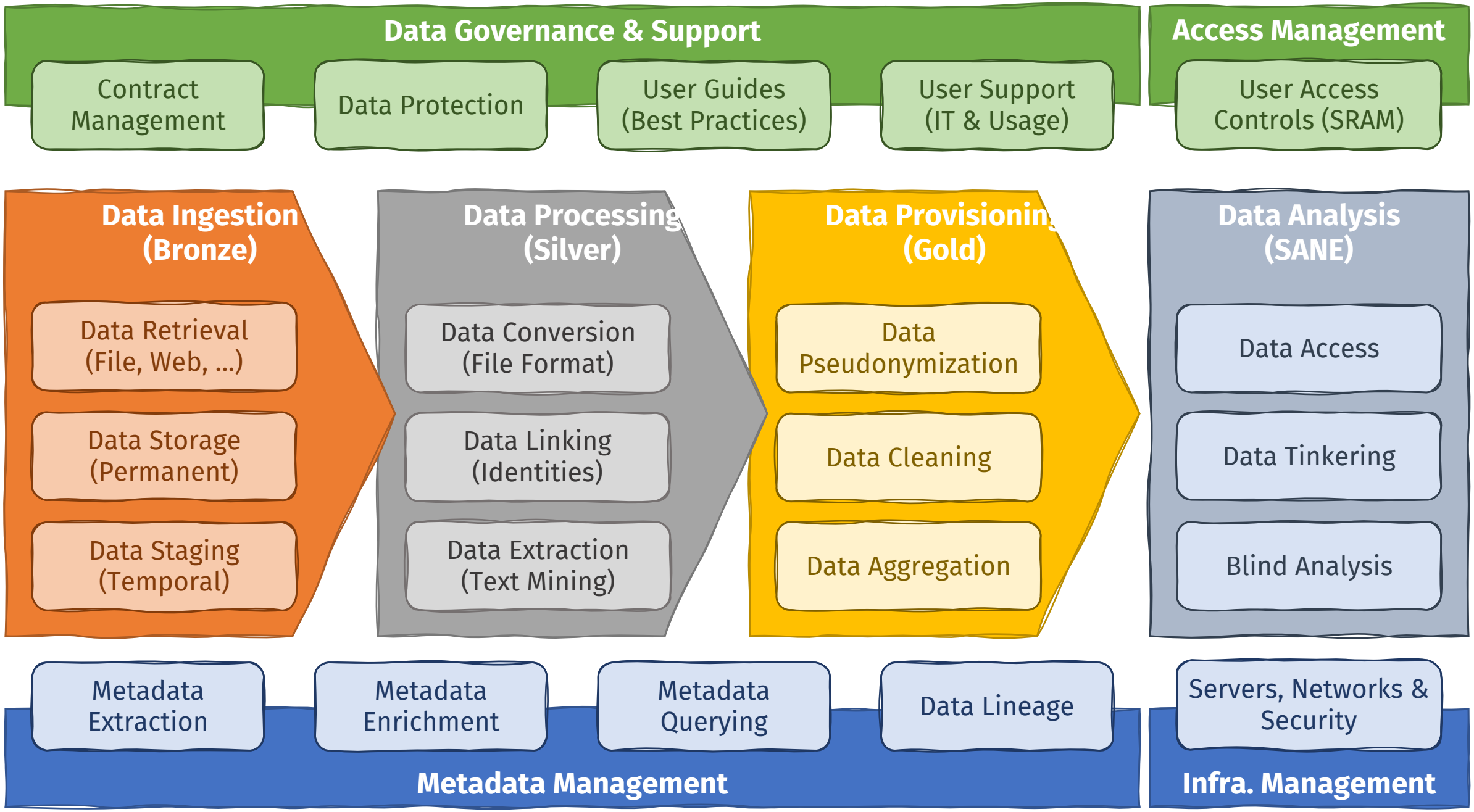
ODISSEI

KVK HOUVAST VOOR ONDERNEMERS

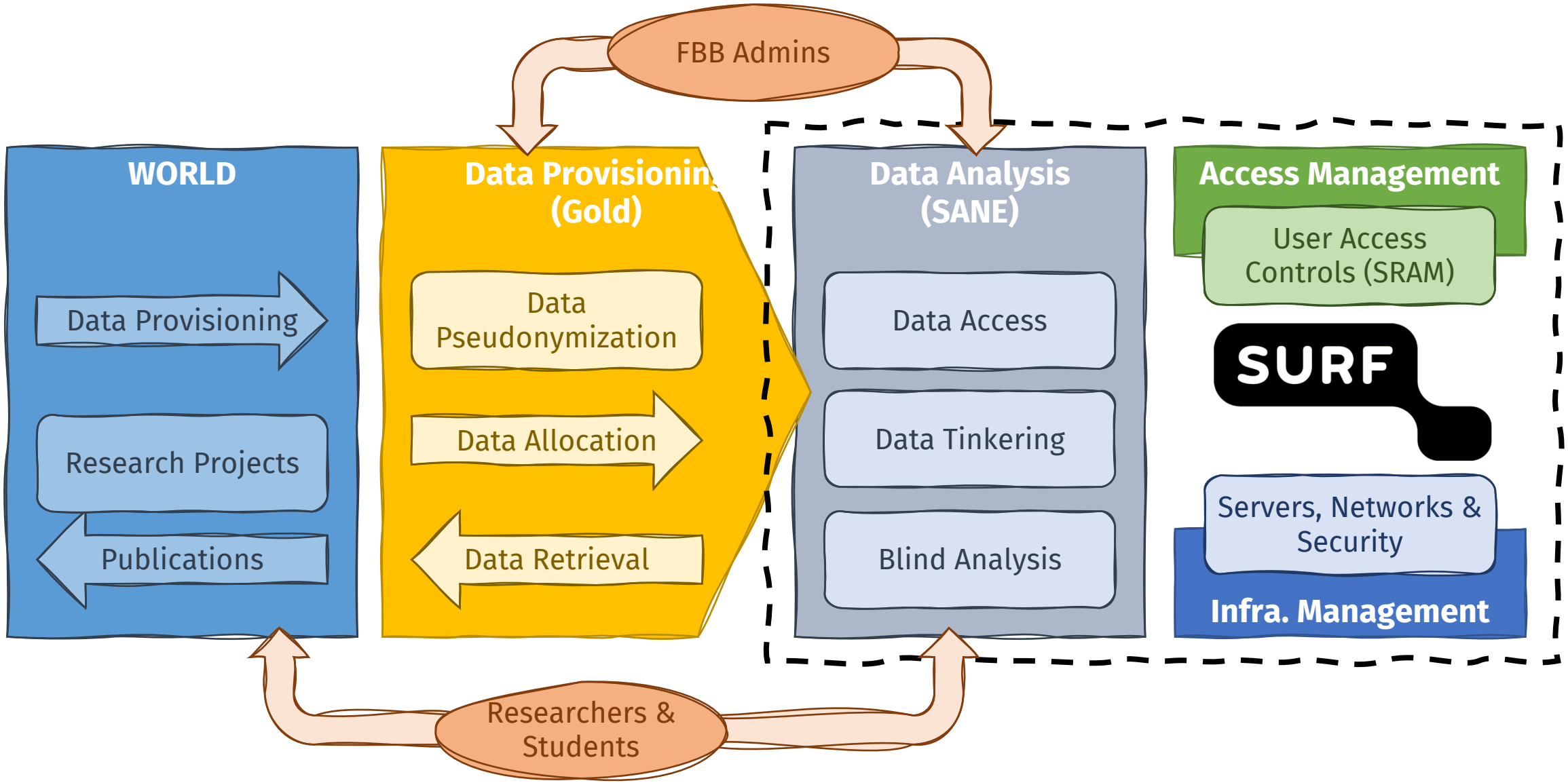
KB nationale bibliotheek

Clariah

SURF


Data Governance & Support
Contract Management
Data Protection
User Guides (Best Practices)
User Support (IT & Usage)
Access Management
User Access Controls (SRAM)
Data Ingestion (Bronze)
Data Retrieval (File, Web, ...)
Data Storage (Permanent)
Data Staging (Temporal)
Data Processing (Silver)
Data Conversion (File Format)
Data Linking (Identities)
Data Extraction (Text Mining)
Data Provisioning (Gold)
Data Pseudonymization
Data Cleaning
Data Aggregation
Data Analysis (SANE)
Data Access
Data Tinkering
Blind Analysis
Metadata Extraction
Metadata Enrichment
Metadata Querying
Data Lineage
Metadata Management
Servers, Networks & Security
Infra. Management

FirmBackbone – Architecture highlight



Q&A

Thank you for your attention!

- annette.langedijk@surf.nl
- lucas@odissei-data.nl
- ahmad.hesam@surf.nl
- freek.dijkstra@surf.nl

SURF Clariah ODISSEI