ETH-RAT

SDSC

RENKU - 連句

Reproducible Data Science

Sandra Savchenko-de Jong for the SDSC Renku team

JupyterCon NY | 24.8.2018

Issues in modern Data Science

- Where did the data for this plot come from?
- What does this new data mean for last year's Nature paper?
- How did my predecessor create these results?
- Can I use your (confidential) data? With my code? In your environment? Online?
- Has anyone ever trained a GAN on this data?
- Who is using my data and code? Why are they not citing me?!

Issues in modern Data Science

Many solutions exist to address part of these questions

- Version control & collaboration: Gitlab/Github
- Collaborate on papers: Overleaf, Google drive
- Re-usable environment/code: Docker Containers
- Re-runnable pipelines: Luigi, CWL

Issues in modern Data Science

Many solutions exist to address part of these questions

- Version control & collaboration: Gitlab/Github
- Collaborate on papers: Overleaf, Google drive
- Re-usable environment/code: Docker Containers
- Re-runnable pipelines: Luigi, CWL

Renku combines existing & new technologies to provide a one-stop shop for data science

Overview

- ***SDSC : a Swiss national initiative***

- Renku : a platform for multi-disciplinary collaboration
  - Big picture
  - System aspects
  - Interacting with the platform
  - What’s next

- Conclusion

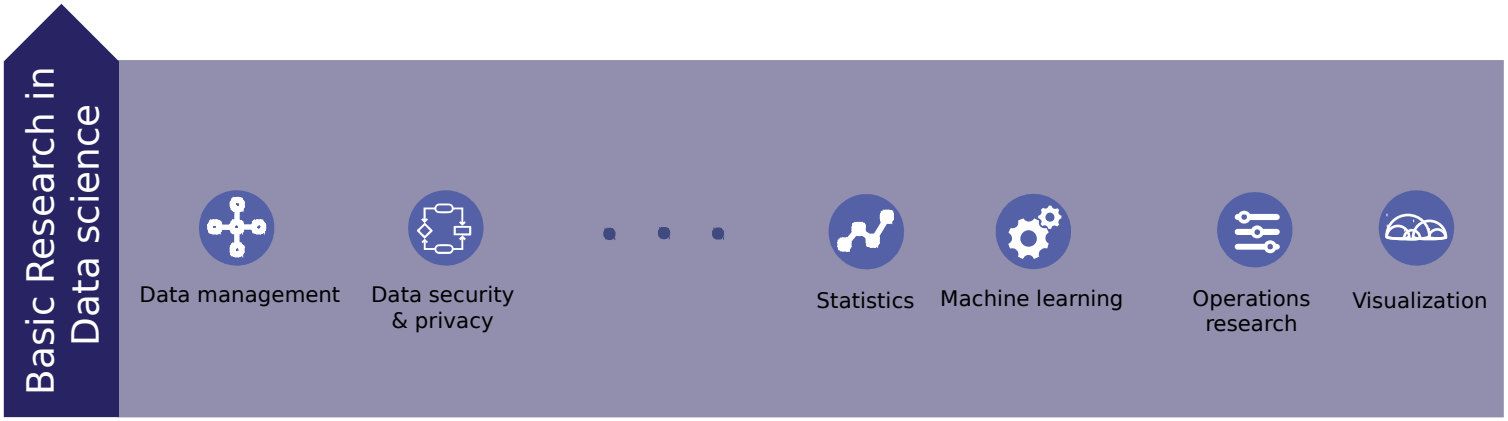
What is SDSC ?

- SDSC - Swiss Data Science Center
- National project
- Not profit-driven
- Joint venture between EPFL and ETH Zurich
- Started 01-2017

- Currently: ~25 people (goal 2018: 40 people ... hiring!)
  (8 software engineers, 10 data scientists, 7 management & admin)

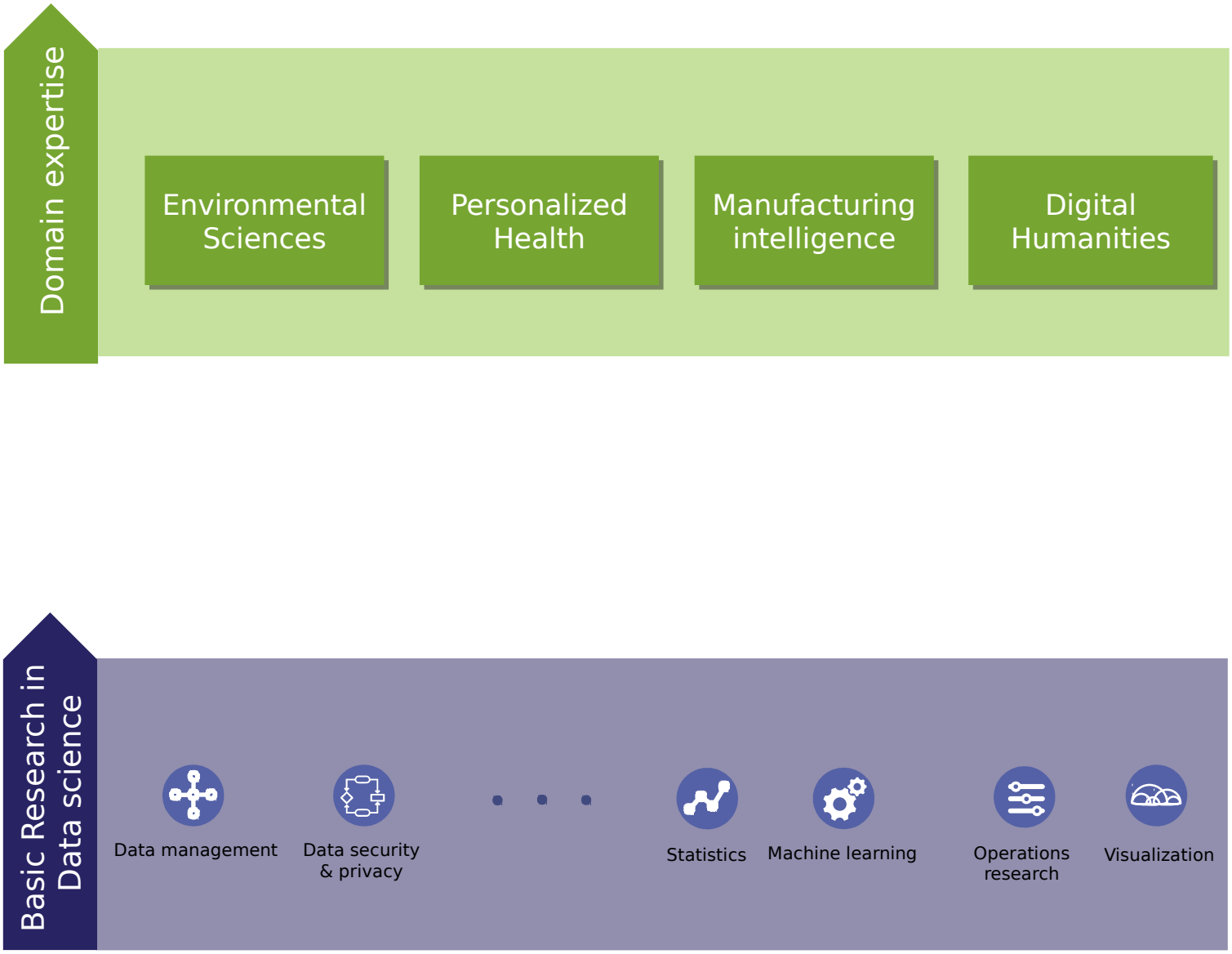
- Open data science
- Involvement in industry & academic projects
- Renku Platform

Interfacing domain and research



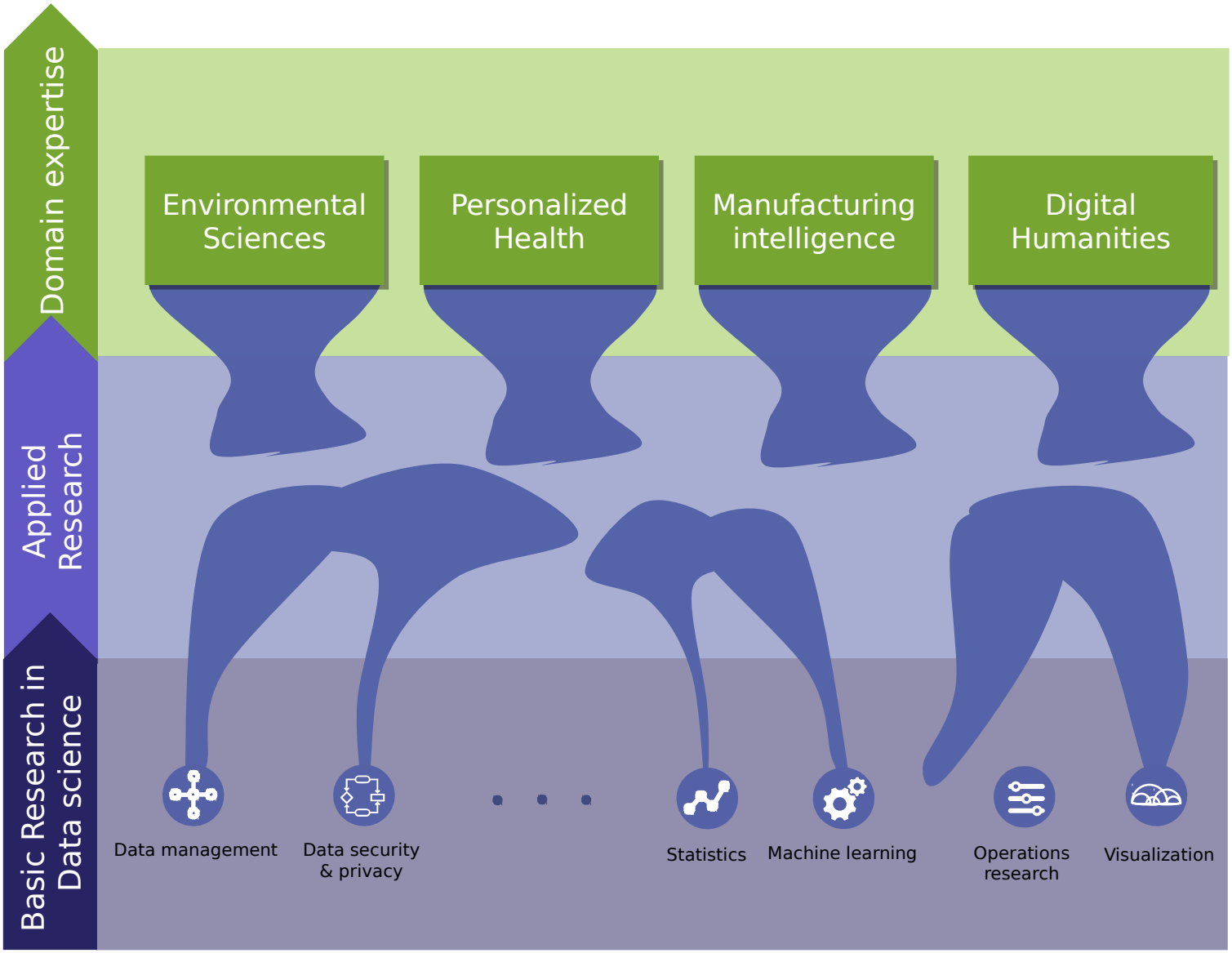
SDSC

Interfacing domain and research



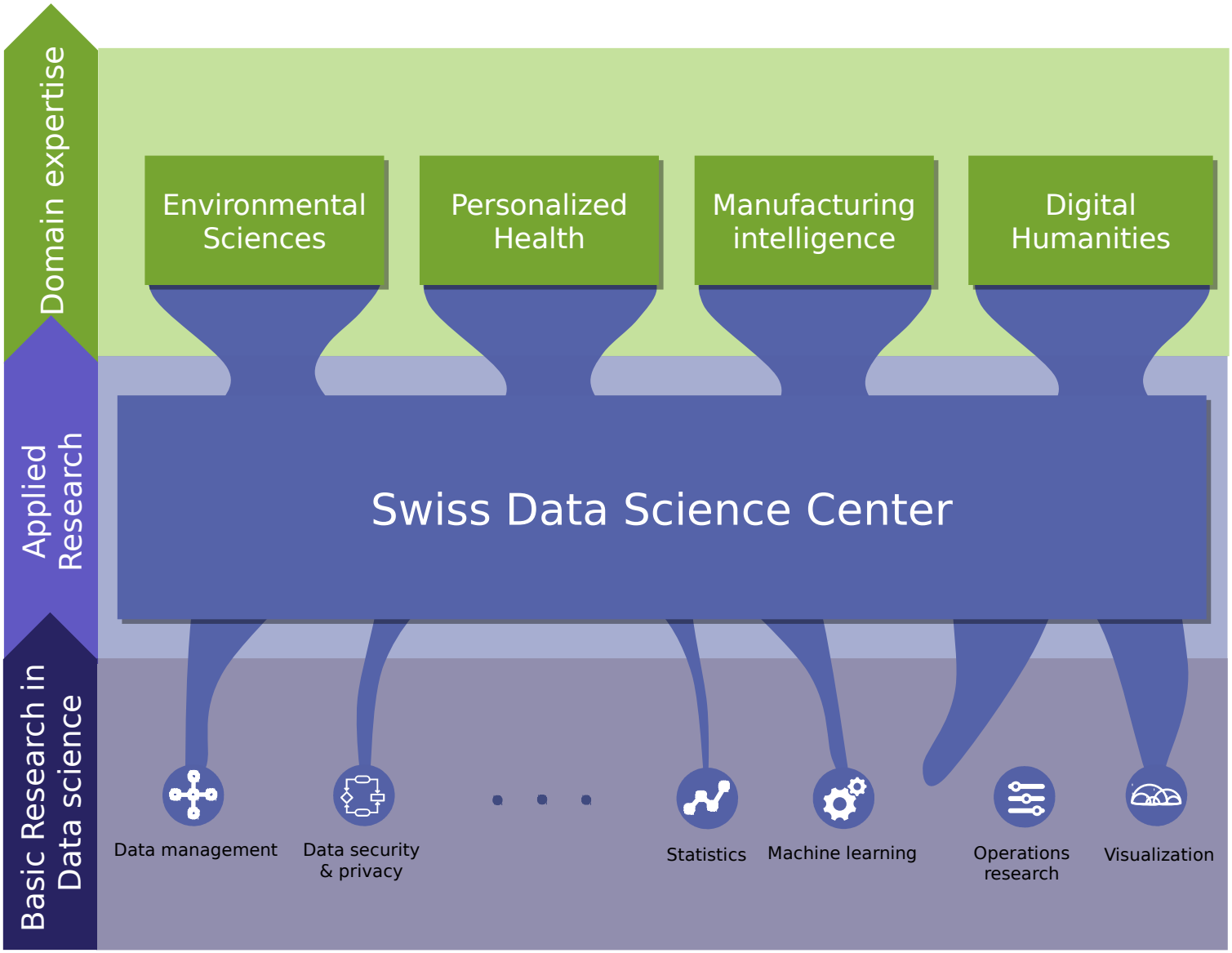
SDSC

Interfacing domain and research



SDSC

Interfacing domain and research



SDSC

Mission of the SDSC

Accelerate the adoption of data science and machine learning techniques within the academic community and the industrial sector

Overview

- SDSC : a Swiss national initiative

- ***Renku : a platform for multi-disciplinary collaboration***

  - Big picture

  - System aspects

  - Interacting with the platform

  - What’s next

- Conclusion

Renku

Renku（連句 "linked verses")

(n)

1. a Japanese form of popular collaborative linked verse poetry

2. SDSC platform for reproducible science

Renku: overview

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

Core of Renku

Capturing, recording and utilizing the lineage of results is the core of Renku

FAIR principles

- **Findable**
  - *“Data and meta-data should be easy to find by both humans and computers”*
- **Accessible**
  - *“Data and meta-data should be stored for the long-term, such that they can be easily accessed and downloaded using standard communication protocols.”*
- **Interoperable**
  - *“Data is ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets”*
- **Reusable**
  - *“Data and metadata are well-described and can be reused in future research. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans”*

https://www.force11.org/group/fairgroup/fairprinciples
EOSC Declaration 10.2017 – Data Culture and FAIR Data

FAIR principles are enabled by Renku

- **Findable**

  → All entities and meta-data are properly labelled, lineage is tracked, easy search functionality

- **Accessible**

  → One-stop shop with secure REST APIs to access data and code

- **Interoperable** :

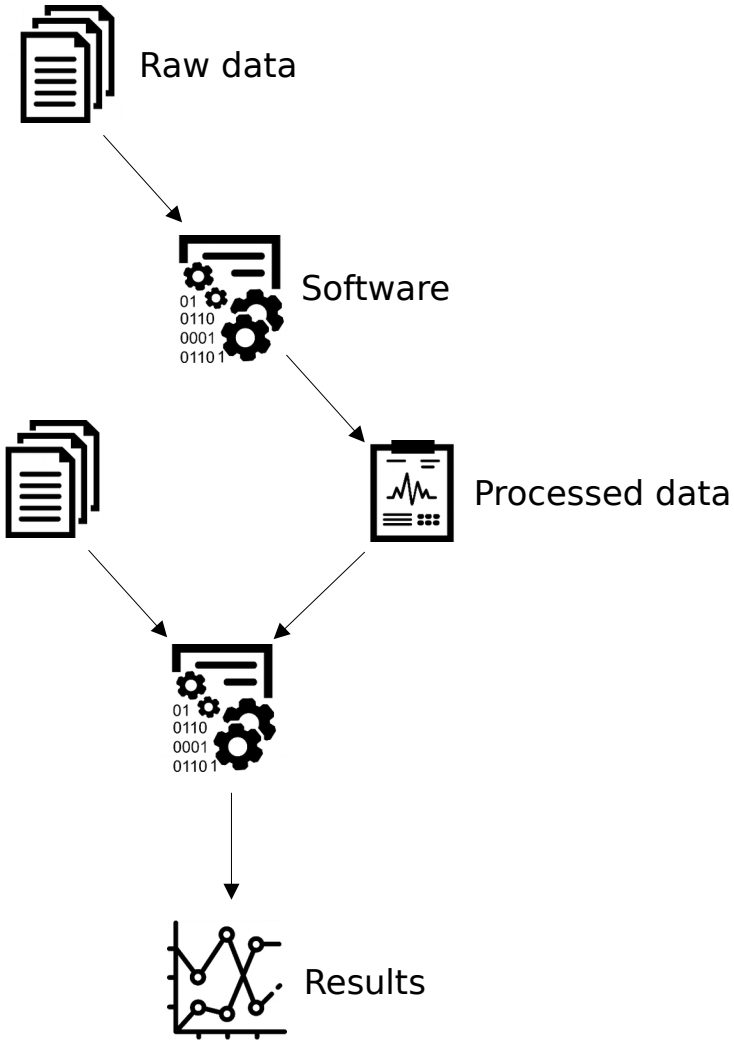
  → Data and metadata are in a standard form ready for use by humans and machines

- **Re-usable**

  → Enabled by tracing and storing of lineage

.... And more!

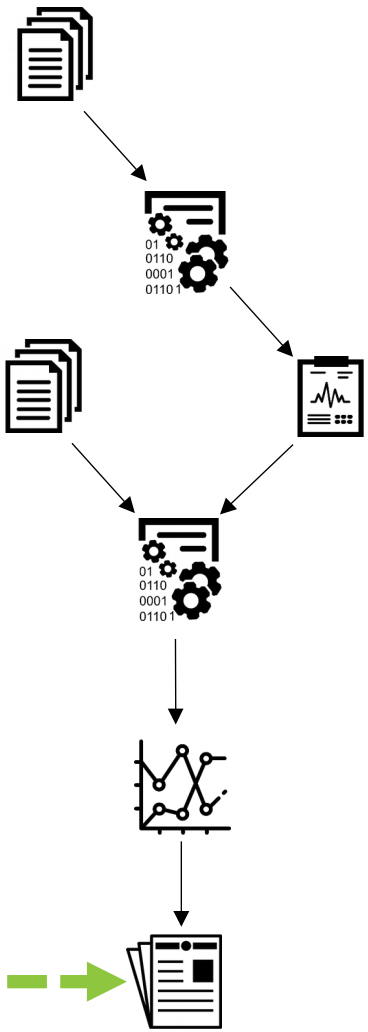
Capture your scientific process



1. Lineage is recorded into a knowledge graph
2. Steps can be repeated and reused
3. Version control is built-in for data, code, and workflows
4. Lineage accessible via simple tools

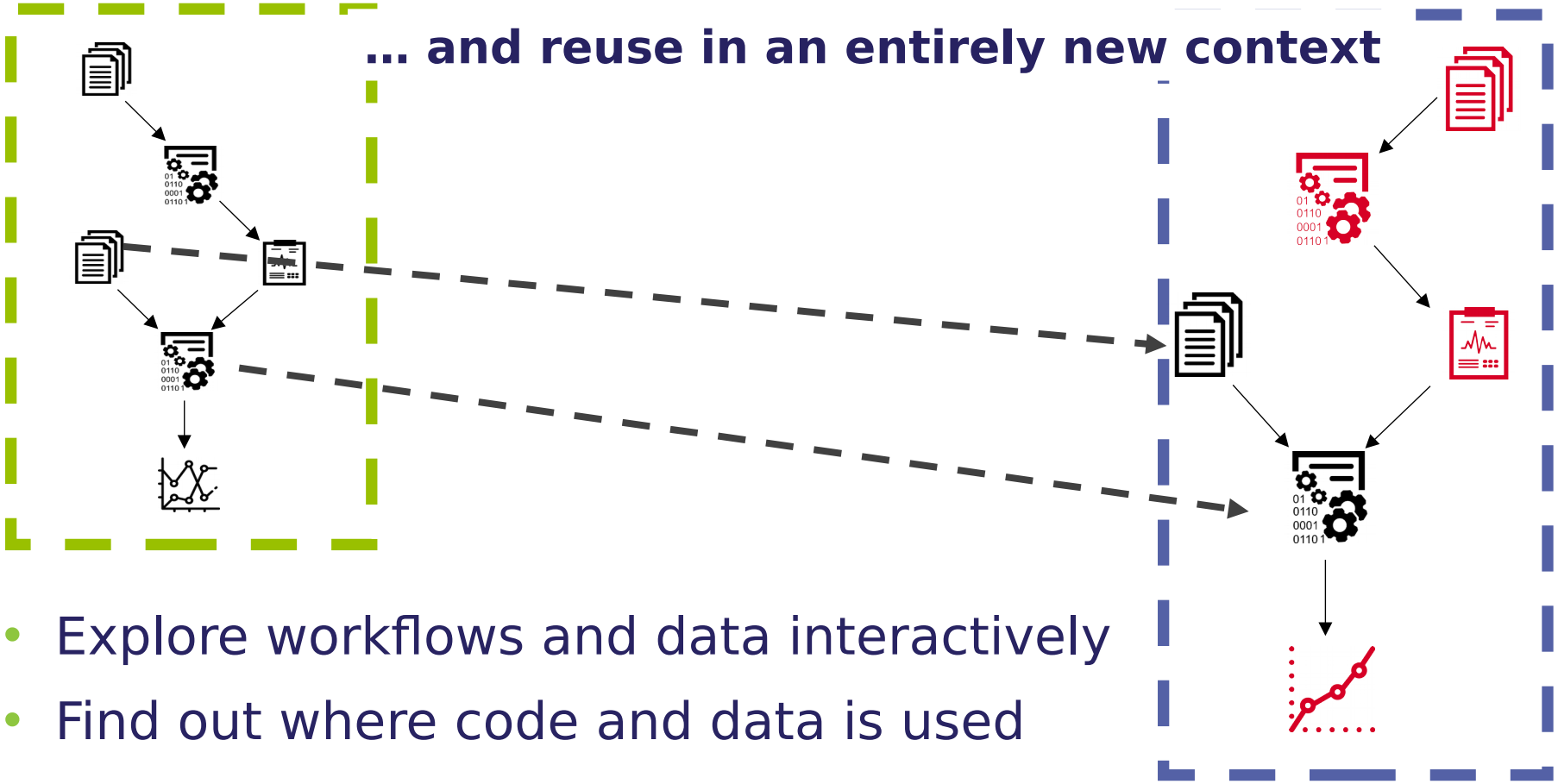
Discover and understand the work of others

Search ...



Ex: search for a publication, obtain a full view of how the results were obtained, search for data sets, algorithms, relationships patterns ...

Reuse and repeat



- Explore workflows and data interactively
- Find out where code and data is used
- Easily reuse work from others, preserving lineage
- Identify popular datasets and algorithms across the platform

SDSC

Overview

- SDSC : a Swiss national initiative

- ***Renku : a platform for multi-disciplinary collaboration***

  - Big picture
  - ***System aspects***
  - Interacting with the platform
  - What’s next

- Conclusion

Renku platform

- Modular architecture:
  - easily extendable
  - re-usable components
- Open source
- Events-based
- Proven standards & technologies
- Written in JS, Scala, and Python

Technologies used

GitLab

jupyter

docker

Kubernetes

HELM

jupyterhub

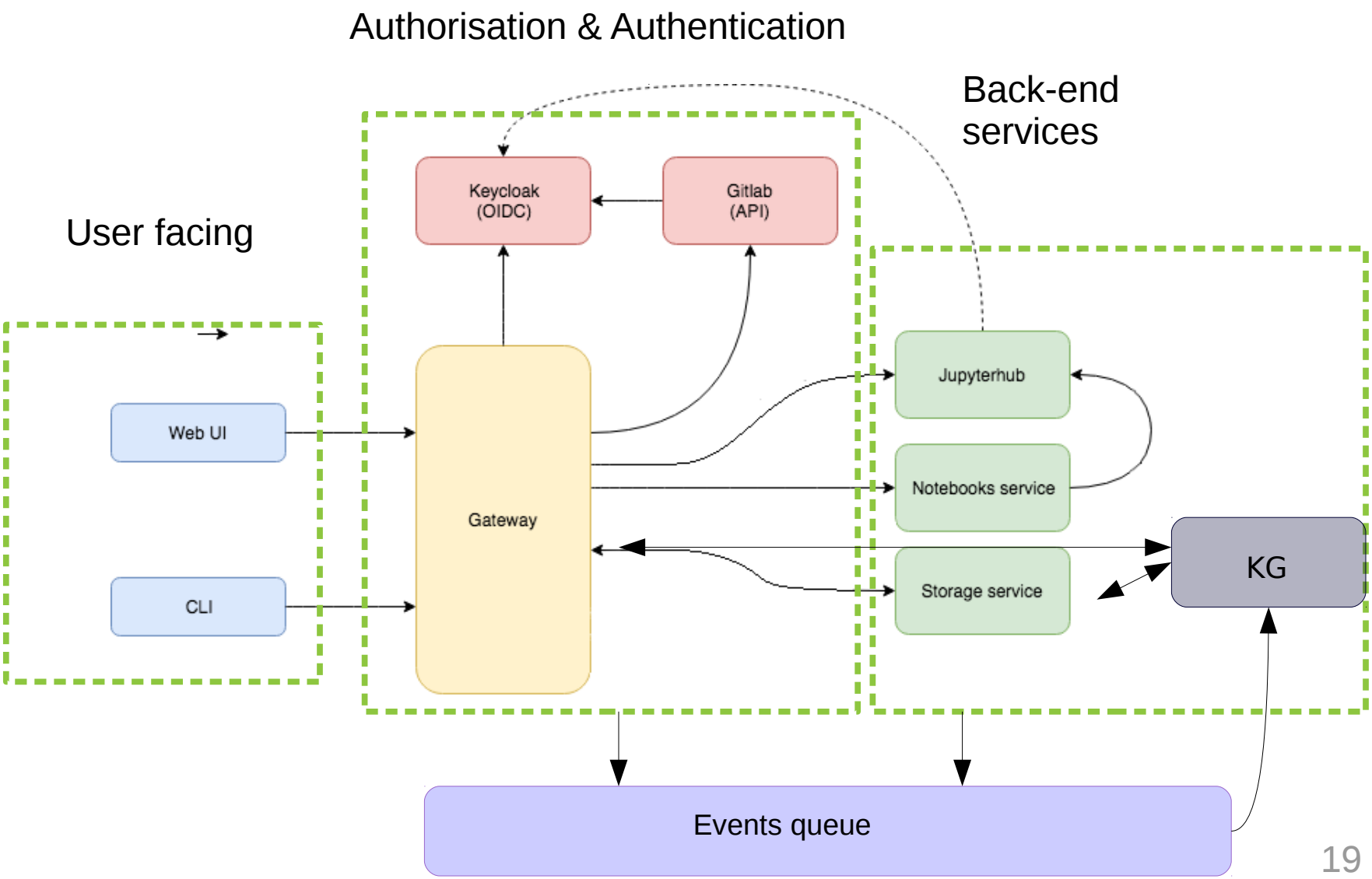
KEYCLOAK

COMMON
WORKFLOW
LANGUAGE

...and more

SDSC

Renku Architecture



SDSC

Events and Knowledge Graph

- Each component generates events which are piped to the event queue
- Events feed into the Knowledge Graph
- Knowledge graph is immutable
- Graph contains information on the data and meta-data
- Graph can be queried by other services to get the state

Overview

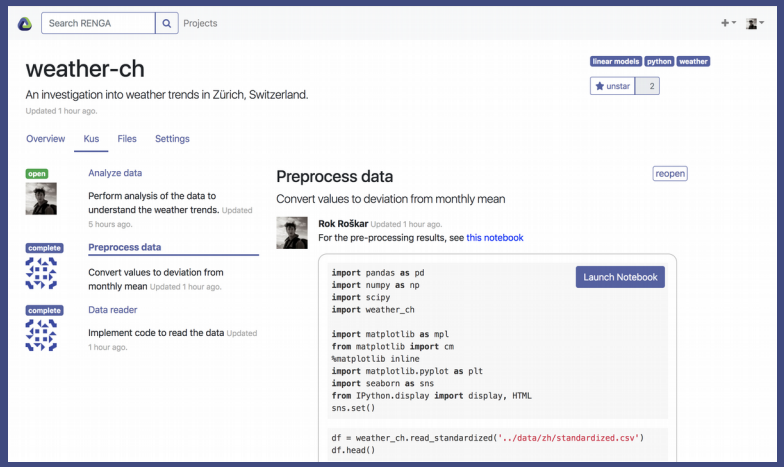
- SDSC : a Swiss national initiative

- ***Renku : a platform for multi-disciplinary collaboration***

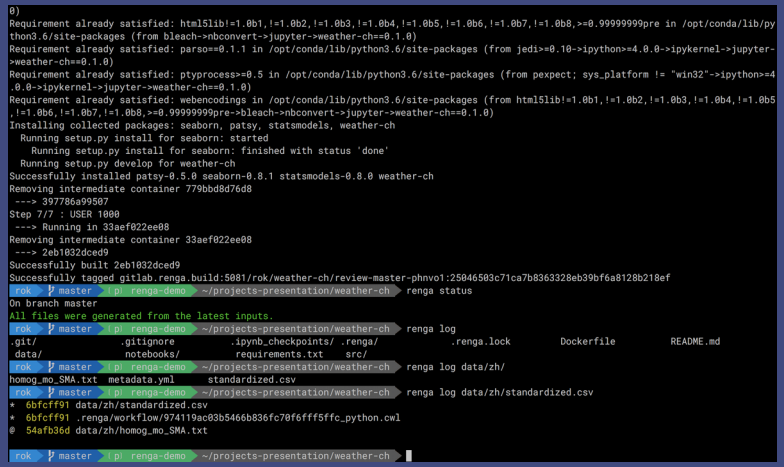
  - Big picture
  - System aspects
  - ***Interacting with the platform***
  - What's next

- Conclusion

Interacting with the platform



Web-based front-end



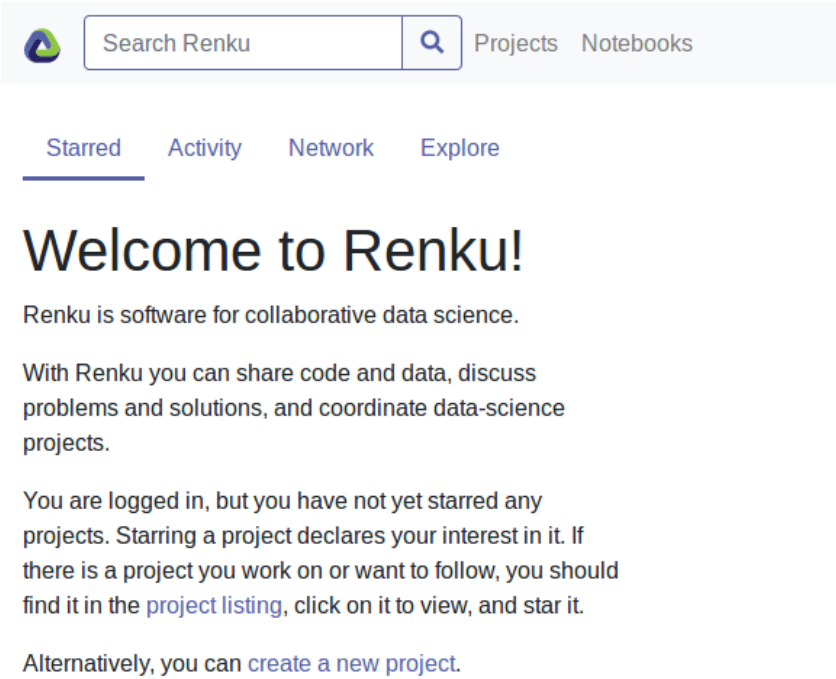
Command-line interface

SDSC

Interacting with the platform: UI

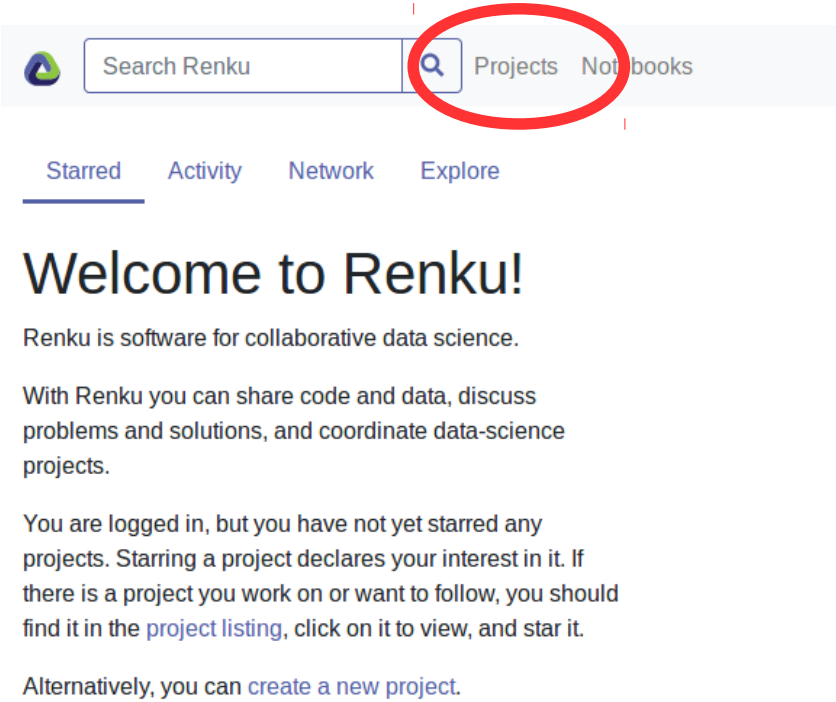
- Online component
- Share and collaborate
- Run notebooks online with JupyterHub
- Easily compare changes made to notebooks before commiting

Interacting with the platform: UI 1

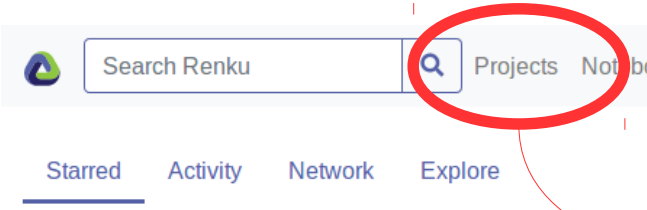


SDSC

Interacting with the platform: UI 1



Interacting with the platform: UI 1



Starred Activity Network Explore

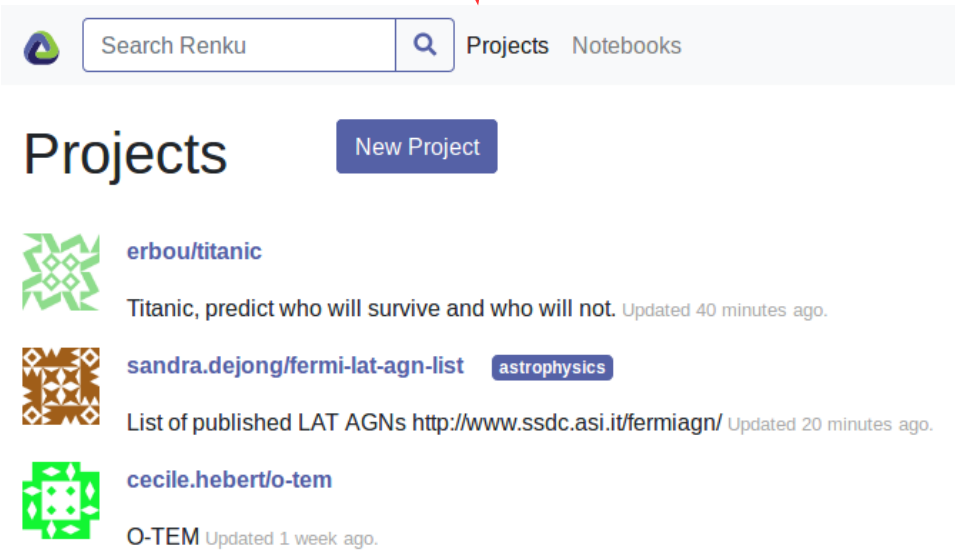
Welcome to Renku!

Renku is software for collaborative data science.

With Renku you can share code and data, discuss problems and solutions, and coordinate data-science projects.

You are logged in, but you have not yet starred any projects. Starring a project declares your interest in it. If there is a project you work on or want to follow, you should find it in the project listing, click on it to view, and star it.

Alternatively, you can create a new project.



SDSC

Interacting with the platform: UI 2

New Project

Title

My new cool project

Id: my-new-cool-project

Description

Let's data science!

A description of the project helps users understand it and is highly recommended.

Visibility

Public

Create

Interacting with the platform: UI 2

New Project

Title

My new cool project

Id: my-new-cool-project

Description

Let's data science!

A description of the project helps users understand it and is highly recommended.

Visibility

Public

Create

My new cool project

Let's data science!

Updated 53 seconds ago.

star 0

Notebooks

Overview Kus Files Pending Changes Settings

All
Data
Notebooks
Workflows

.gitignore
.gitlab-ci.yml
.renku/metadata.yml
Dockerfile
requirements.txt

SDSC

Interacting with the platform: UI 2

New Project

Title

My new cool project

Id: my-new-cool-project

Description

Let's data science!

A description of the project helps users understand it and is highly recommended.

Visibility

Public

Create

Custom notebook image is being built based on Dockerfile + requirements.txt

My new cool project

Let's data science!

Updated 53 seconds ago.

star 0

Notebooks

Overview Kus Files Pending Changes Settings

All

Data

Notebooks

Workflows

.gitignore

.gitlab-ci.yml

.renku/metadata.yml

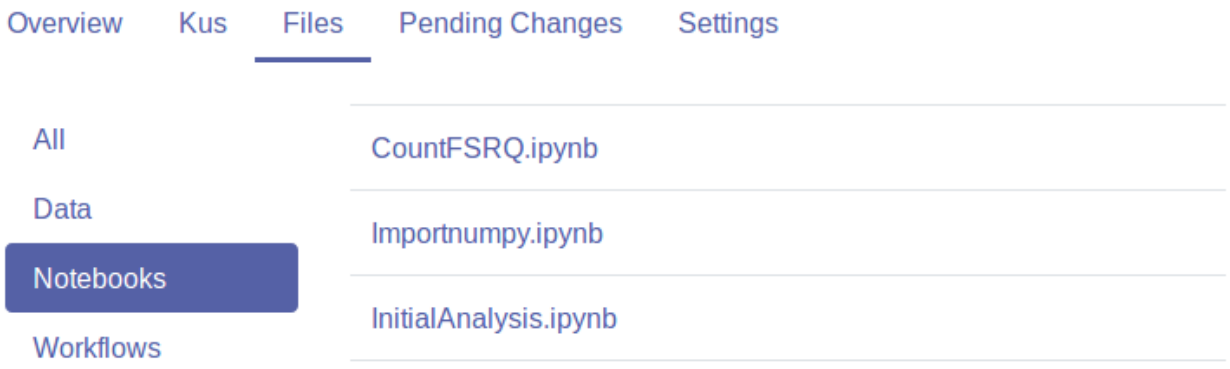
Dockerfile

requirements.txt

Initial files upon creation

SDSC

Interacting with the platform: UI 3



SDSC

Interacting with the platform: UI 3

Overview Kus Files Pending Changes Settings

All
Data
Notebooks
Workflows

CountFSRQ.ipynb
Importnumpy.ipynb
InitialAnalysis.ipynb

Preview of notebook & quick launch button

Overview Kus Files Pending Changes Settings

All
Data
Notebooks
Workflows

Launch Notebook

Initial analysis: check content of data set and count occurances of sources

```
[ ]  import numpy as np
     import pandas as pd
```

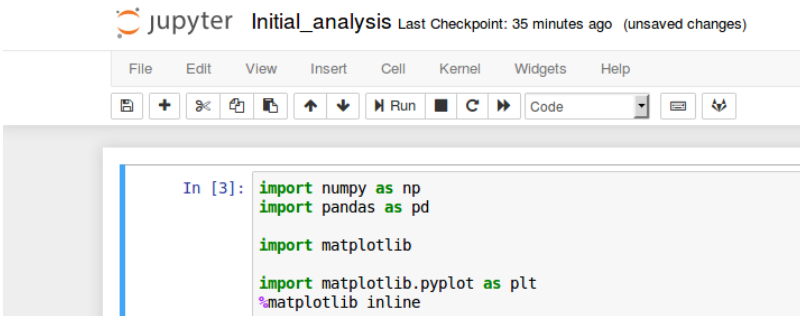
```
[ ]  data = pd.read_csv('data/dataset/fermi3lac.csv')
```

data.count()

SDSC

Interacting with the platform: UI 4

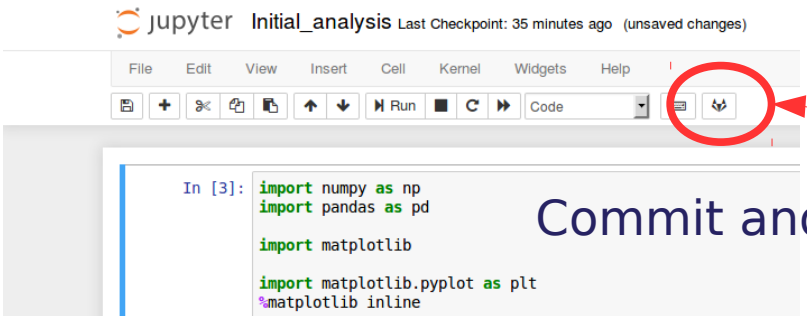
Jupyter notebook:



SDSC

Interacting with the platform: UI 4

Jupyter notebook:

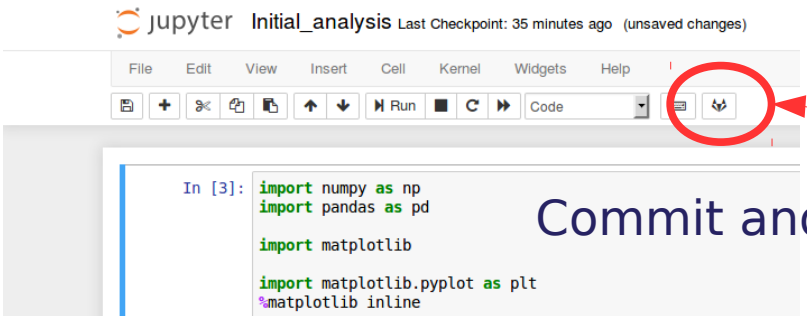


Commit and push changes

SDSC

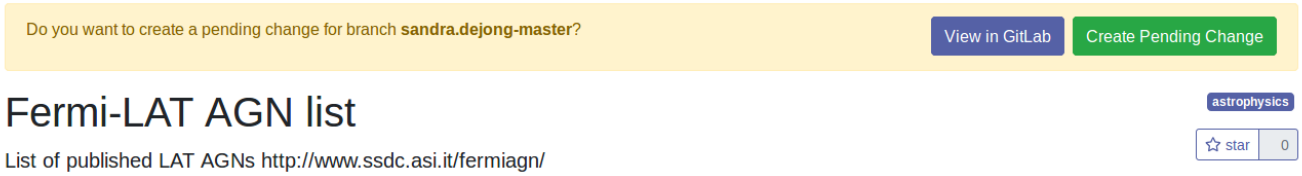
Interacting with the platform: UI 4

Jupyter notebook:



Commit and push changes

In the Renku UI:



Fermi-LAT AGN list

astrophysics

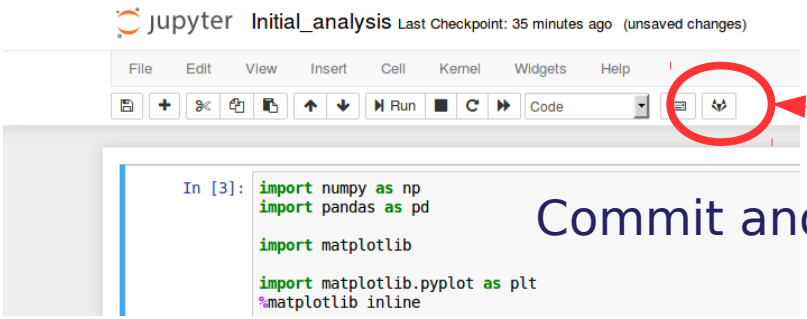
star 0

List of published LAT AGNs http://www.ssdc.asi.it/fermiagn/

SDSC

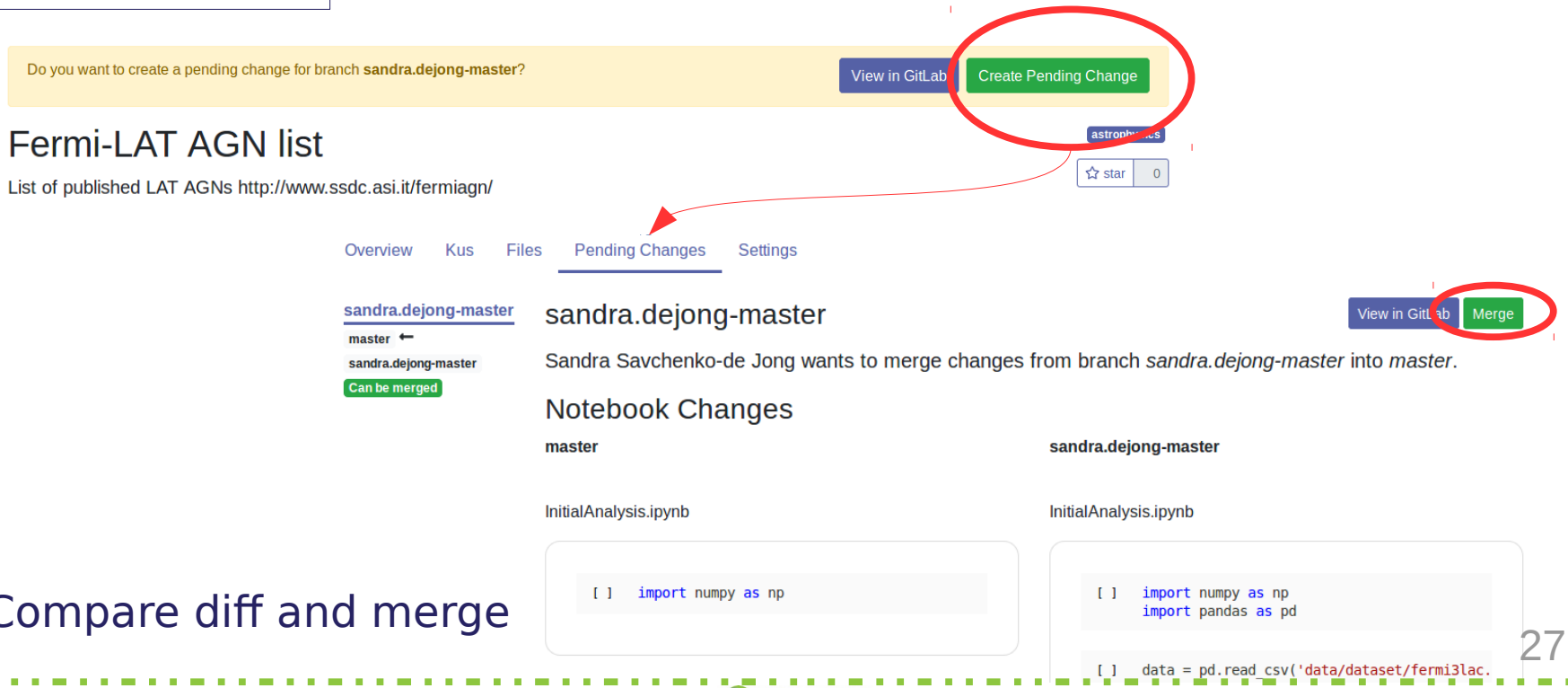
Interacting with the platform: UI 4

Jupyter notebook:



Commit and push changes

In the Renku UI:

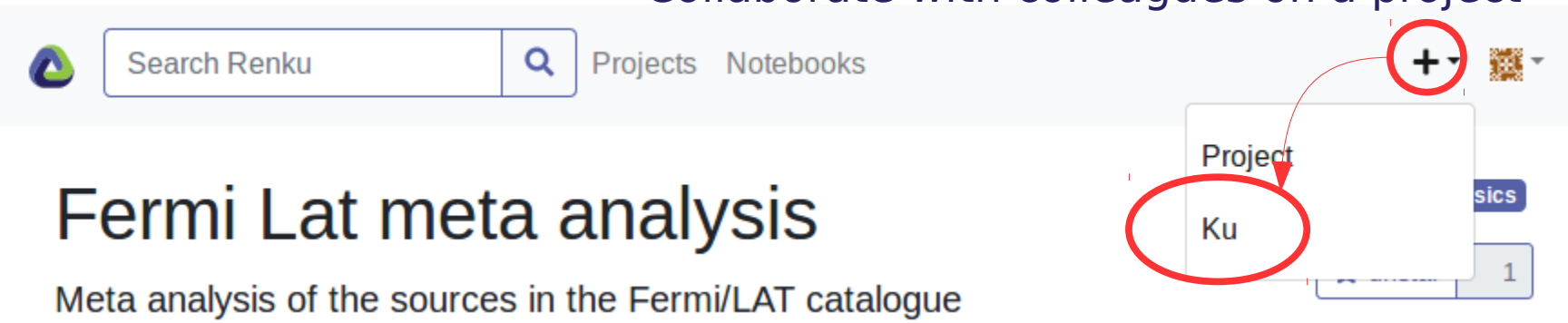


Compare diff and merge

SDSC

Collaborating using Ku

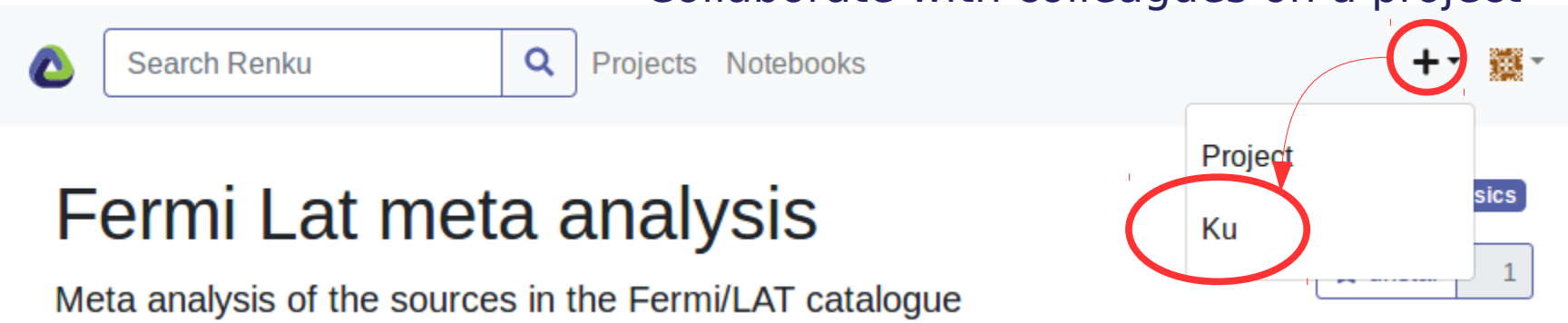
Collaborate with colleagues on a project



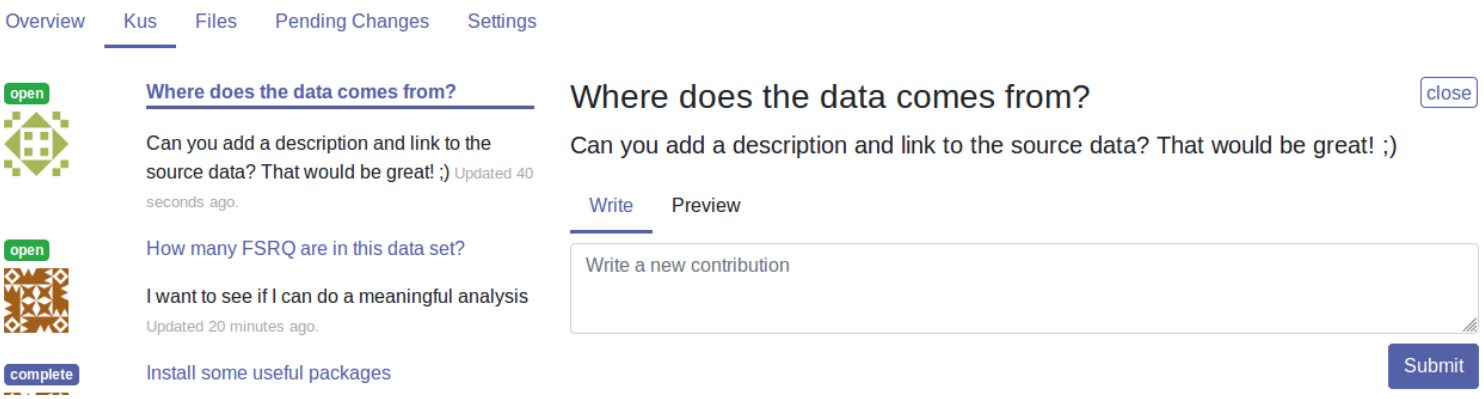
Extention of Gitlab ‘Issues’

Collaborating using Ku

Collaborate with colleagues on a project



Extention of Gitlab ‘Issues’



SDSC

Collaborating using Ku 2

Refer to and open notebooks from a Ku

How many FSRQ are in this data set?

close

I want to see if I can do a meaningful analysis

Sandra Savchenko-de Jong Updated 2 days ago.

You can use this

Launch Notebook

```
[1]  import numpy as np
     import pandas as pd
```

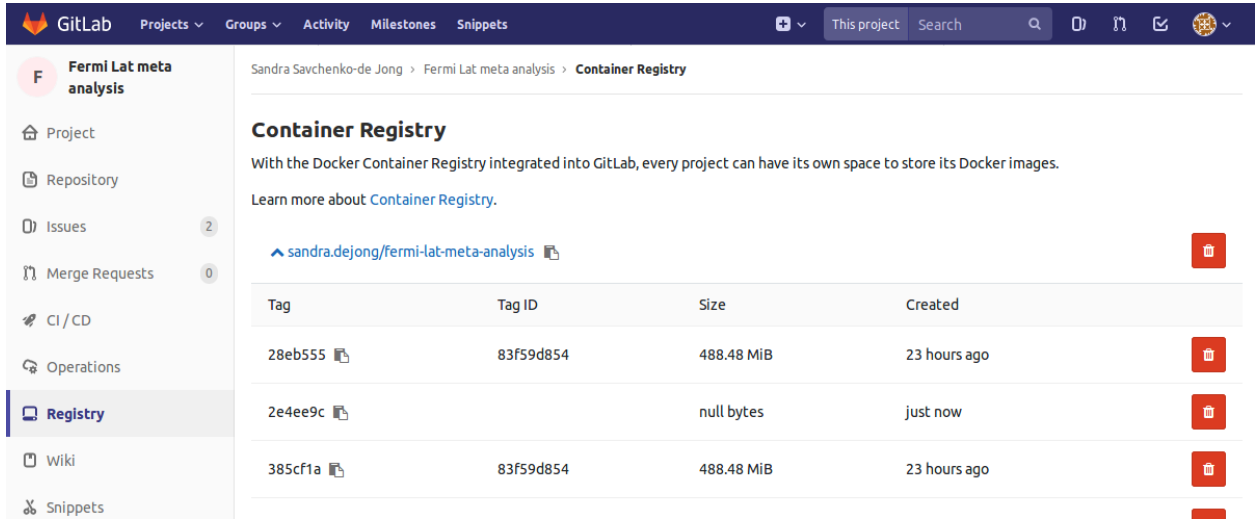
```
[2]  fermisources=pd.read_csv('data/dataset/fermi3lac.csv'
```


SDSC

Jupyter notebook service

Notebook service links GitLab and JupyterHub:

- At each push an image is built according to Dockerfile + requirements in project
- URL is provided to launch a Jupyter server based on project & commit hash



→ Retrieve state of project at each commit !

SDSC

Interacting with the platform: CLI

- Online & Offline component
- Can run without the full platform
- Run reproducible workflows

Interacting with the platform: CLI

- Create data sets and add data

```
$ renku dataset create mydata
$ renku dataset add mydata file
```

Interacting with the platform: CLI

- Create data sets and add data

```
$ renku dataset create mydata
$ renku dataset add mydata file
```

- Run analysis on data

```
$ renku run analysis mydata > output
```

Interacting with the platform: CLI

- Create data sets and add data

```
$ renku dataset create mydata
$ renku dataset add mydata file
```

- Run analysis on data

```
$ renku run analysis mydata > output
```

- View lineage of data

```
$ renku log output
```

Interacting with the platform: CLI 2

```
sandra@Charon ~fgl $ renku run grep -i fsrq data/dataset/fermilac3.csv > fsrq
sandra@Charon ~fgl $ renku log fsrq
*   e4d16b65 fsrq
*   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

```
sandra@Charon ~fgl $ renku run grep -i blazar data/dataset/fermi3lac.csv > blazar
sandra@Charon ~fgl $ renku log blazar
*   640802b9 blazar
*   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Interacting with the platform: CLI 2

```
sandra@Charon ~fgl $ renku run grep -i fsrq data/dataset/fermilac3.csv > fsrq
sandra@Charon ~fgl $ renku log fsrq
*   e4d16b65 fsrq
*   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

```
sandra@Charon ~fgl $ renku run grep -i blazar data/dataset/fermi3lac.csv > blazar
sandra@Charon ~fgl $ renku log blazar
*   640802b9 blazar
*   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Interacting with the platform: CLI 2

```
sandra@Charon ~fgl $ renku run grep -i fsrq data/dataset/fermilac3.csv > fsrq
sandra@Charon ~fgl $ renku log fsrq
*   e4d16b65 fsrq
*   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

```
sandra@Charon ~fgl $ renku run grep -i blazar data/dataset/fermi3lac.csv > blazar
sandra@Charon ~fgl $ renku log blazar
*   640802b9 blazar
*   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Use previous output as input for new pipelines

```
sandra@Charon ~/fgl $ renku run wc fsrq blazar > wc_all.out
sandra@Charon ~/fgl $ renku log wc_all.out
*   8a572a25 wc_all.out
*     8a572a25 .renku/workflow/53f3edbc917c4d7b8a07334984089b11_wc.cwl
|\
* |   e4d16b65 fsrq
* |   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
| *   640802b9 blazar
| *   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
|/
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Interacting with the platform: CLI 2

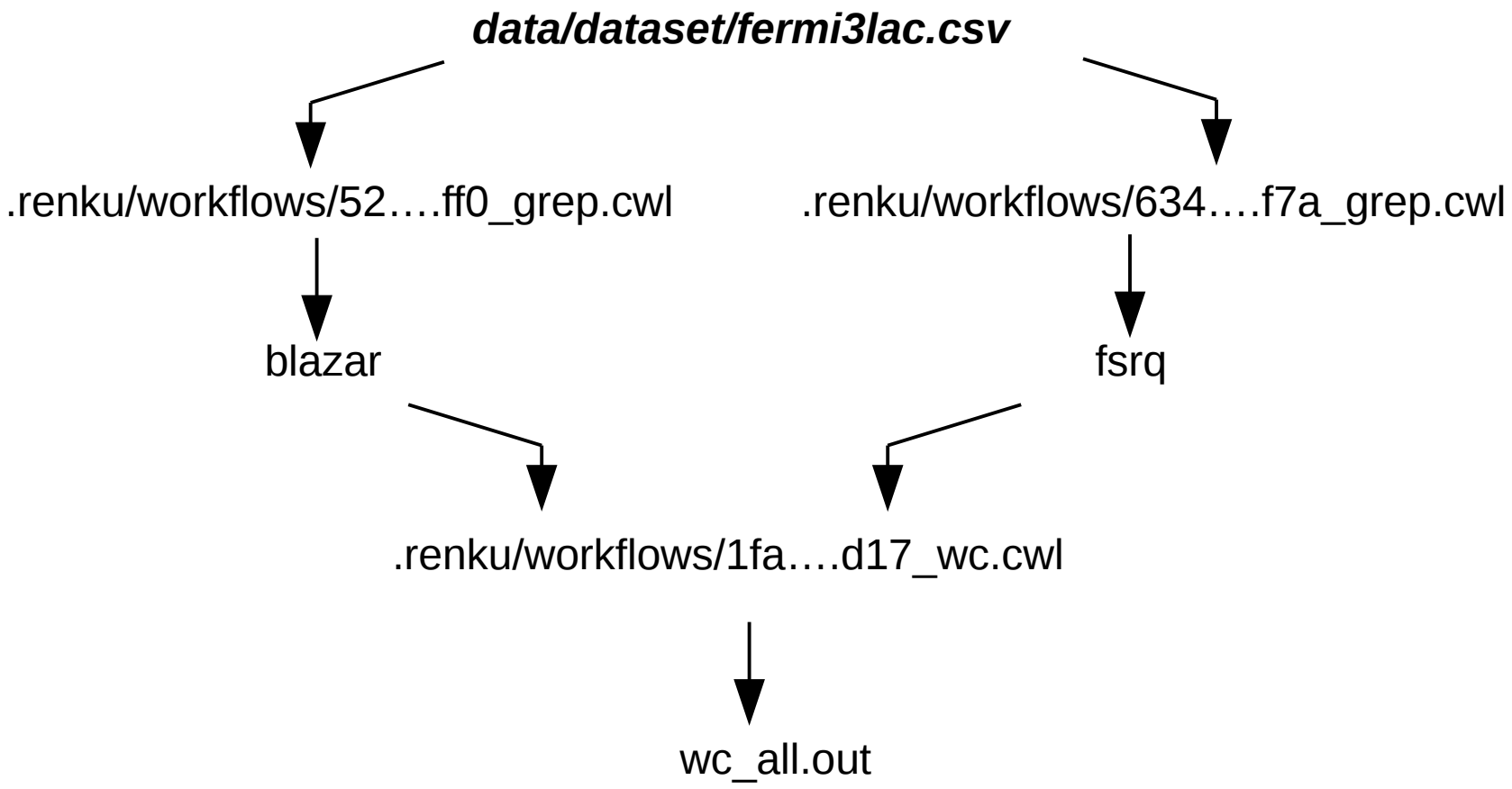
```
sandra@Charon ~fgl $ renku run grep -i fsrq data/dataset/fermilac3.csv > fsrq
sandra@Charon ~fgl $ renku log fsrq
*   e4d16b65 fsrq
*   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

```
sandra@Charon ~fgl $ renku run grep -i blazar data/dataset/fermi3lac.csv > blazar
sandra@Charon ~fgl $ renku log blazar
*   640802b9 blazar
*   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Use previous output as input for new pipelines

```
sandra@Charon ~/fgl $ renku run wc fsrq blazar > wc_all.out
sandra@Charon ~/fgl $ renku log wc_all.out
*   8a572a25 wc_all.out
*     8a572a25 .renku/workflow/53f3edbc917c4d7b8a07334984089b11_wc.cwl
|\
* |   e4d16b65 fsrq
* |   e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
| *   640802b9 blazar
| *   640802b9 .renku/workflow/56594c43e4b74934b649d1127650663e_grep.cwl
|/
@   a5c2fcfc data/dataset/fermi3lac.csv
```

Interacting with the platform: UI Lineage



SDSC

Overview

- SDSC : a Swiss national initiative

- ***Renku : a platform for multi-disciplinary collaboration***

  - Big picture
  - System aspects
  - Interacting with the platform
  - ***What's next***

- Conclusion

What to expect in 12 months

- Access control : ABAC
- Federated mode
- Support for workflow execution in the cloud
- Use ontology and metadata standards for better interoperability e.g. PROV-O/JSON-LD
- Graph-search functionality

Plugins:

- Data and code discovery
- Recommender systems based on the KG
- And more !

Overview

- SDSC : a Swiss national initiative

- Renku : a platform for multi-disciplinary collaboration

  - Big picture
  - System aspects
  - Interacting with the platform
  - What’s next

- ***Conclusion***

Issues in modern Data Science

- Where did the data for this plot come from?
- What does this new data mean for last year's Nature paper?
- How did my predecessor create these results?
- Can I use your (confidential) data? With my code? In your environment? Online?
- Has anyone ever trained a GAN on this data?
- Who is using my data and code? Why are they not citing me?!

...addressed by Renku

- I know where the data for this plot came from (reproducibility)
- I can rerun my analysis with this new data and compare with last year's Nature paper (repetition)
- I know how my predecessor created these results (reproducibility)
- I can use your (confidential) data, with my code, on your cluster or online if I have the right permissions (collaboration, federation)
- I can search if someone ever trained a GAN on this data (discovery)
- I know who is using my data and code.... And I am automatically cited through the lineage (reproducibility)

For more information

@SDSCdatascience

- https://renku.readthedocs.io/
- https://github.com/SwissDataScienceCenter/renku
- https://datascience.ch/renku-platform/
- https://renku.readthedocs.io/en/latest/user/firststeps.html

For more information

@SDSCdatascience

- https://renku.readthedocs.io/
- https://github.com/SwissDataScienceCenter/renku
- https://datascience.ch/renku-platform/
- https://renku.readthedocs.io/en/latest/user/firststeps.html

Try Renku on renkulab.io

SDSC