

RENKU - 連句

Reproducible Data Science

Sandra Savchenko-de Jong for the SDSC Renku team

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

- ***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

Basic Research in Data science



Data management



Data security
& privacy



Statistics



Machine learning



Operations
research



Visualization

Interfacing domain and research

Domain expertise

Environmental
Sciences

Personalized
Health

Manufacturing
intelligence

Digital
Humanities

Basic Research in
Data science



Data management



Data security
& privacy



Statistics



Machine learning

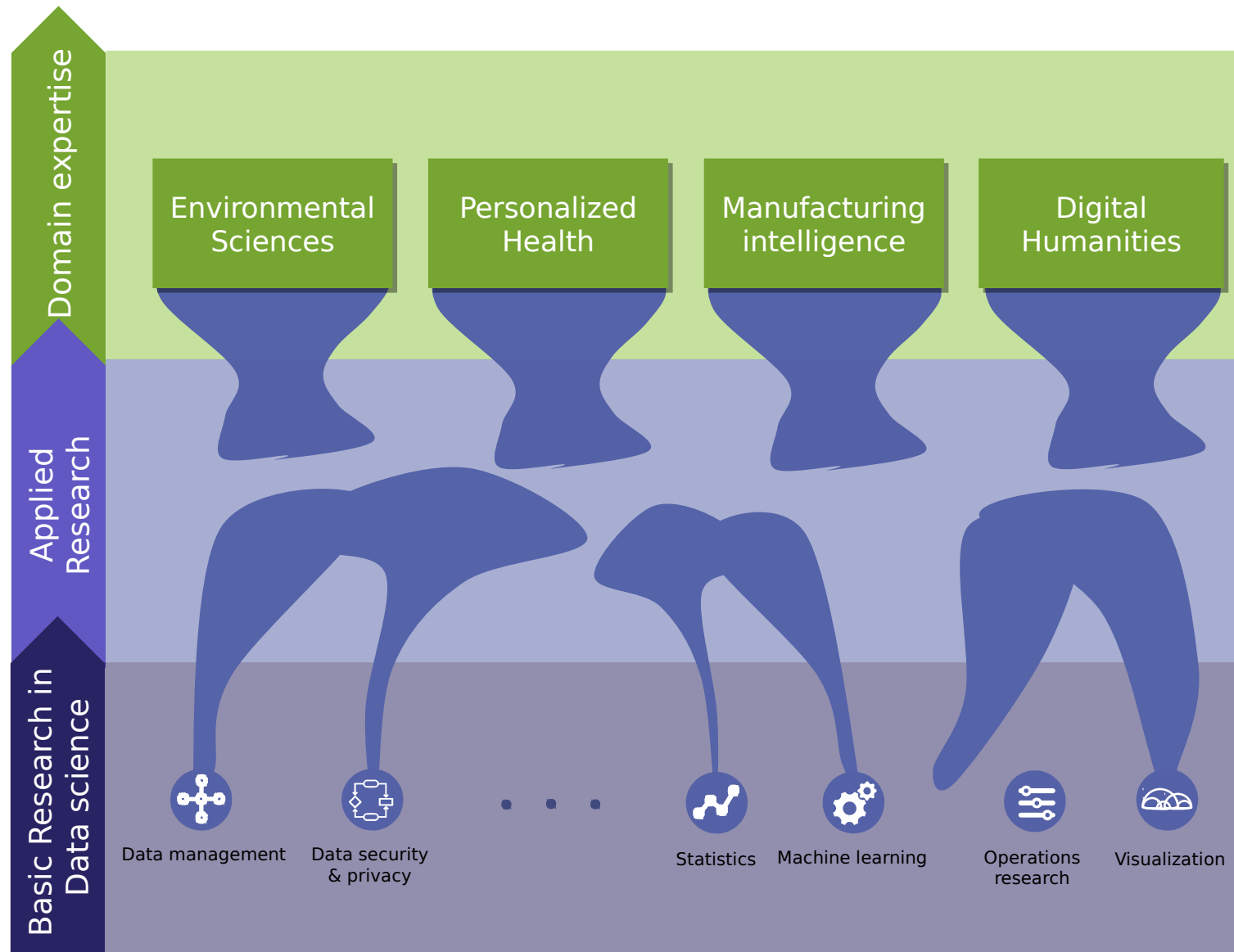


Operations
research

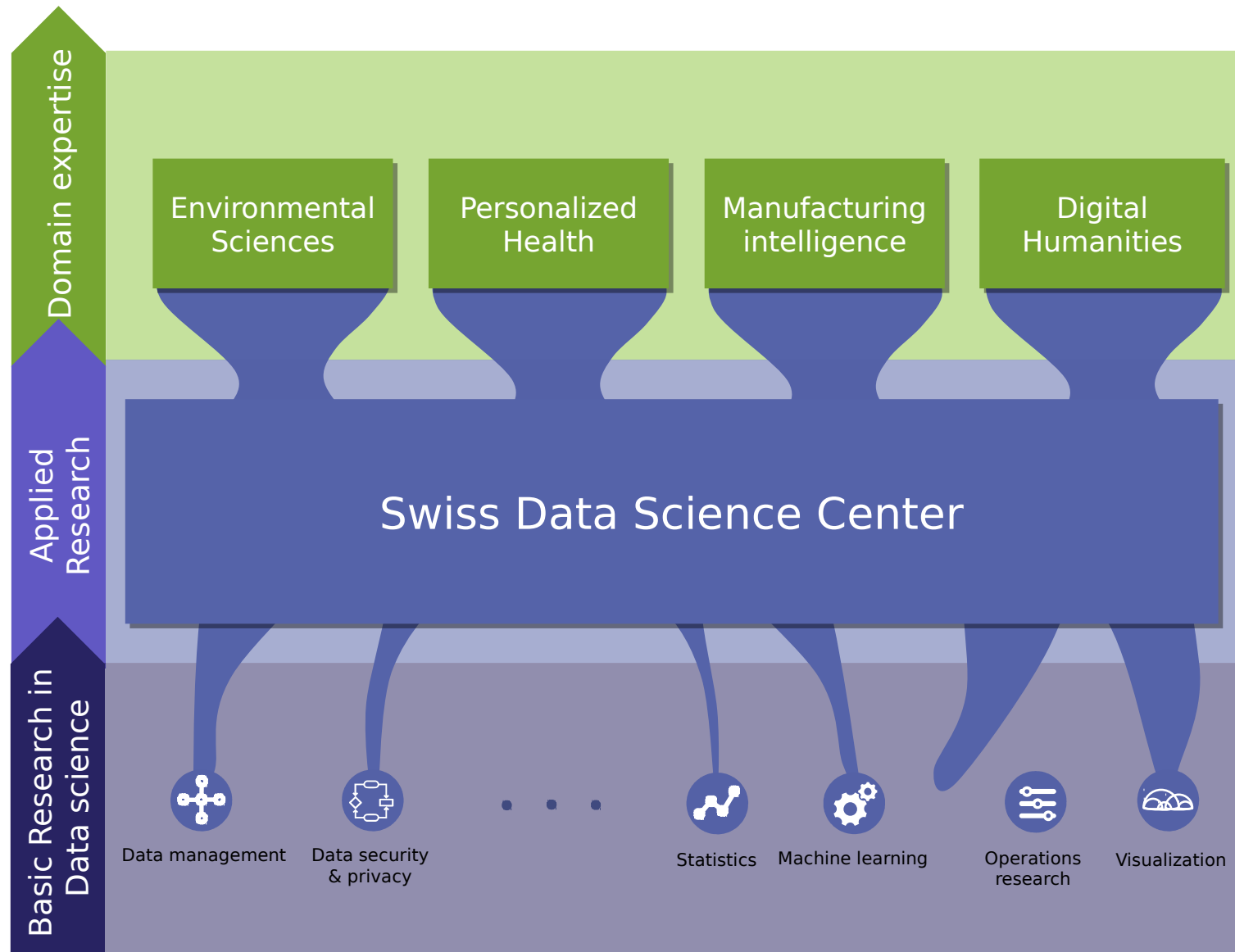


Visualization

Interfacing domain and research



Interfacing domain and research



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 (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 :**

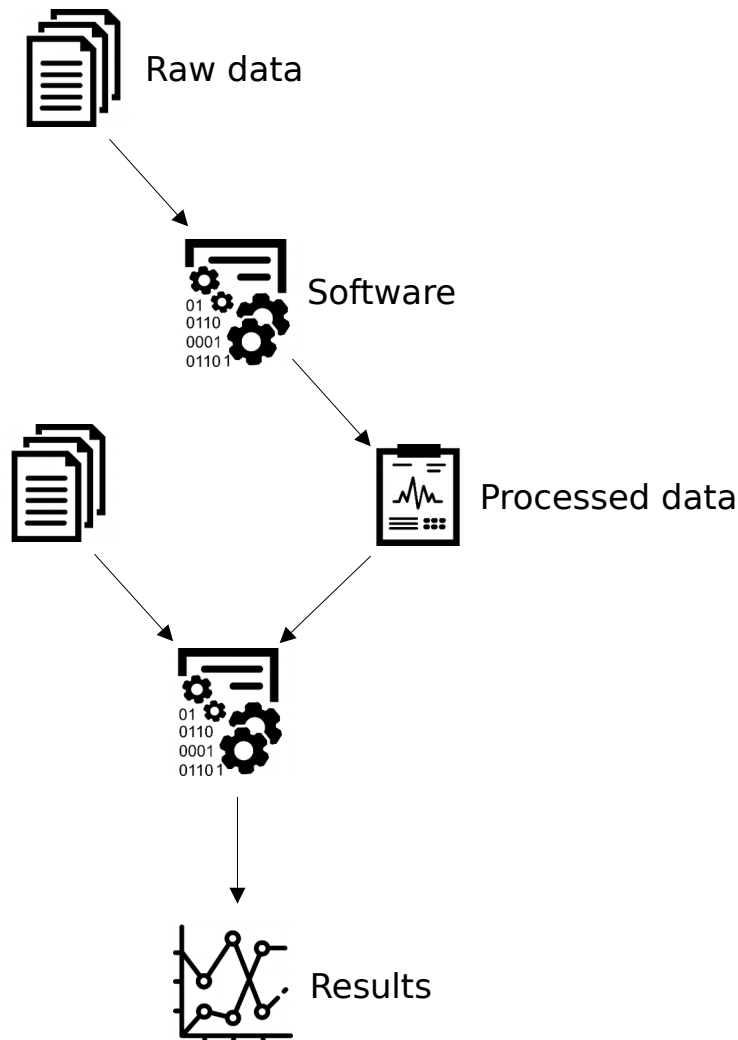
→ Data and metadata are in a standard form ready for use by humans and machines

- **Re-usable**

→ Enabled by tracing and storing of lineage

.... And more!

Capture your scientific process



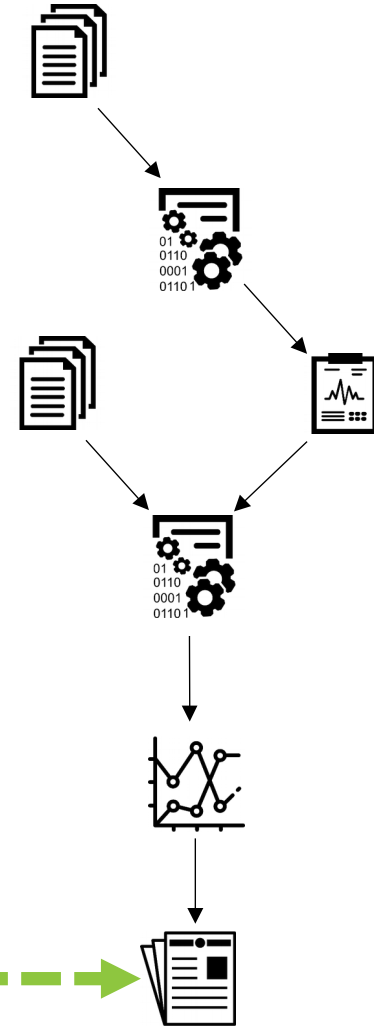
1. Lineage is recorded into a knowledge graph
2. Steps can be repeated and reused
3. Version control is built-in for data, code, and workflows
4. Lineage accessible via simple tools

Discover and understand the work of others

Search ...

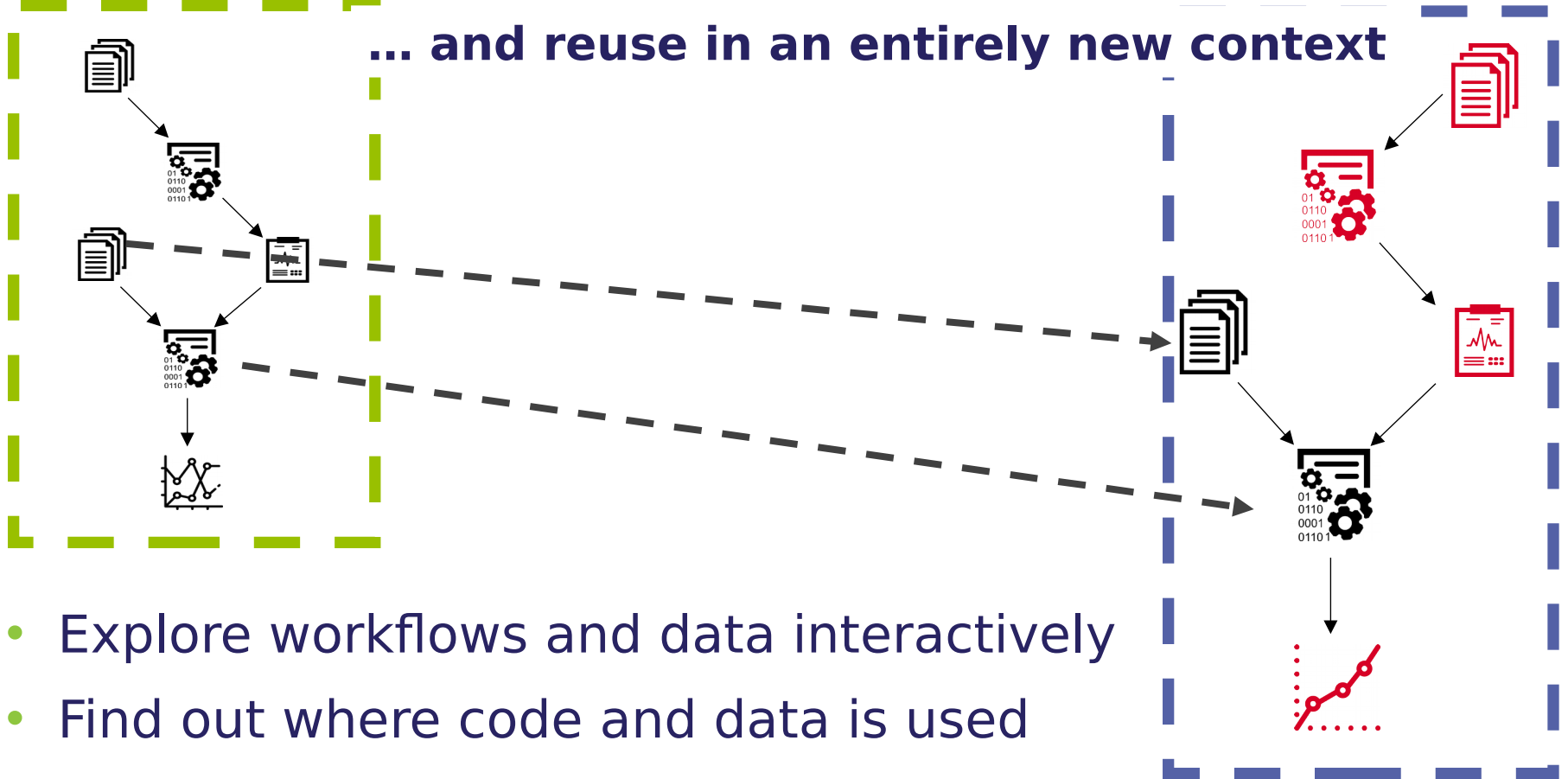


Ex: search for a publication,
obtain a full view of how the
results were obtained, search
for data sets, algorithms,
relationships patterns ...



Reuse and repeat

... and reuse in an entirely new context



- 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

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



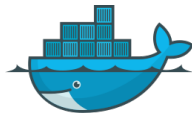
- Modular architecture:
 - easily extendable
 - re-usable components
- Open source
- Events-based
- Proven standards & technologies
- Written in JS, Scala, and Python



Technologies used



GitLab



docker



Kubernetes



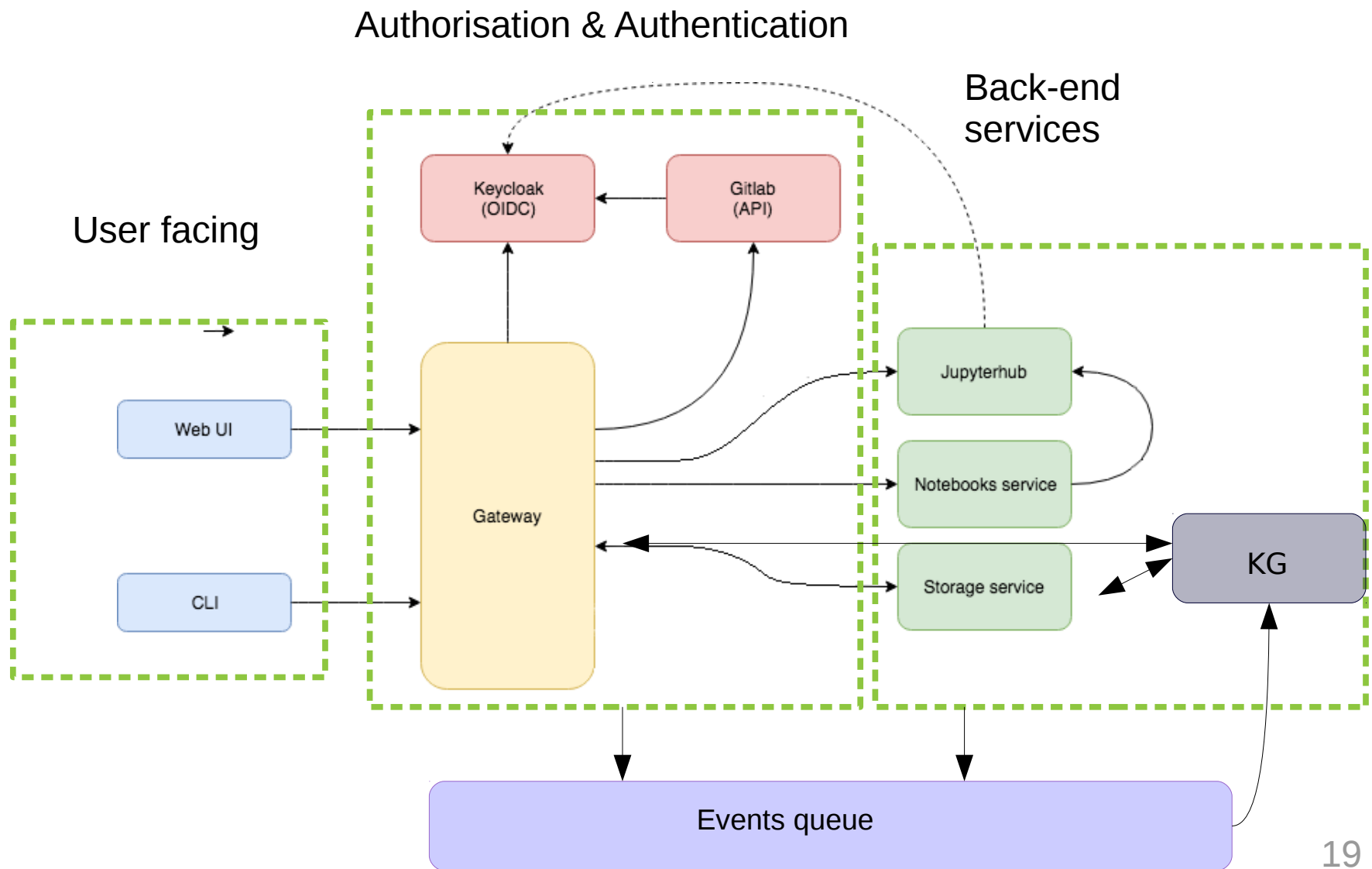
KEYCLOAK



COMMON
WORKFLOW
LANGUAGE

...and more

Renku Architecture



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



Search RENG A Projects

weather-ch

An investigation into weather trends in Zürich, Switzerland.
Updated 1 hour ago.

Overview Kus Files Settings

Analyze data
Perform analysis of the data to understand the weather trends. Updated 5 hours ago.

Preprocess data
Convert values to deviation from monthly mean. Updated 1 hour ago.

Data reader
Implement code to read the data. Updated 1 hour ago.

Preprocess data
Convert values to deviation from monthly mean
For the pre-processing results, see this notebook

Launch Notebook

```
import pandas as pd
import numpy as np
import scipy
import weather_ch

import matplotlib as mpl
from matplotlib import cm
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display, HTML
sns.set()

df = weather_ch.read_standardized('.../data/zh/standardized.csv')
df.head()
```

Web-based front-end



```
0)
Requirement already satisfied: html5lib=1.0b1,!=1.0b2,!=1.0b3,!=1.0b4,!=1.0b5,!=1.0b6,!=1.0b7,!=1.0b8,!=0.99999999pre in /opt/conda/lib/python3.6/site-packages (from bleach->nbconvert->jupyter->weather-ch==0.1.0)
Requirement already satisfied: parso==0.1.1 in /opt/conda/lib/python3.6/site-packages (from jedi==0.10.0->ipython==4.0.0->ipykernel->jupyter->weather-ch==0.1.0)
Requirement already satisfied: ptyprocess==0.5 in /opt/conda/lib/python3.6/site-packages (from pexpect; sys_platform != "win32"-->ipython==4.0.0->ipykernel->jupyter->weather-ch==0.1.0)
Requirement already satisfied: webencodings in /opt/conda/lib/python3.6/site-packages (from html5lib=1.0b1,!=1.0b2,!=1.0b3,!=1.0b4,!=1.0b5,!=1.0b6,!=1.0b7,!=1.0b8,!=0.99999999pre->bleach->nbconvert->jupyter->weather-ch==0.1.0)
Installing collected packages: seaborn, patsy, statsmodels, weather-ch
Running setup.py install for seaborn: started
Running setup.py install for seaborn: finished with status 'done'
Running setup.py develop for weather-ch
Successfully installed patsy-0.5.0 seaborn-0.8.1 statsmodels-0.8.0 weather-ch
Removing intermediate container 779bdd8d76d8
--> 397786a99597
Step 7/7 : USER 1000
--> Running in 33aef022ee08
Removing intermediate container 33aef022ee08
--> 2eb1832dced9
Successfully built 2eb1832dced9
Successfully tagged gitlab.renga build:5881/rok/weather-ch/review-master-phnvol:26946593c71ca7b836332eb39b6a8128b218ef
rok @ master ~$ ip renga-demo ~/projects/presentation/weather-ch renga status
On branch master
All files generated from the latest inputs.
rok @ master ~$ ip renga-demo ~/projects/presentation/weather-ch renga log
.git/ .gitignore .ipynb_checkpoints/ .renga/ .renga.lock Dockerfile README.md
data/ notebooks/ requirements.txt src/
rok @ master ~$ ip renga-demo ~/projects/presentation/weather-ch renga log data/zh/
homog_mo_SMA.txt metadata.yml standardized.csv
rok @ master ~$ ip renga-demo ~/projects/presentation/weather-ch renga log data/zh/standardized.csv
* 6bfcff91 data/zh/standardized.csv
* 6bfcff91 .renga/workflow/974119ac83b5466b836fc78f6ffff5ffc.python.cwl
@ 54afb36d data/zh/homog_mo_SMA.txt
rok @ master ~$ ip renga-demo ~/projects/presentation/weather-ch
```

Command-line interface

Interacting with the platform: UI

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

Interacting with the platform: UI 1

  [Projects](#) [Notebooks](#)

[Starred](#) [Activity](#) [Network](#) [Explore](#)

Welcome to Renku!

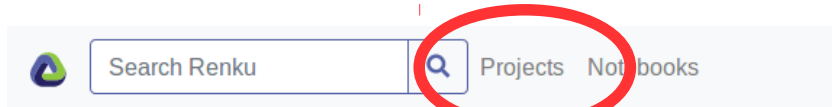
Renku is software for collaborative data science.

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](#).

Interacting with the platform: UI 1



Starred Activity Network Explore

Welcome to Renku!

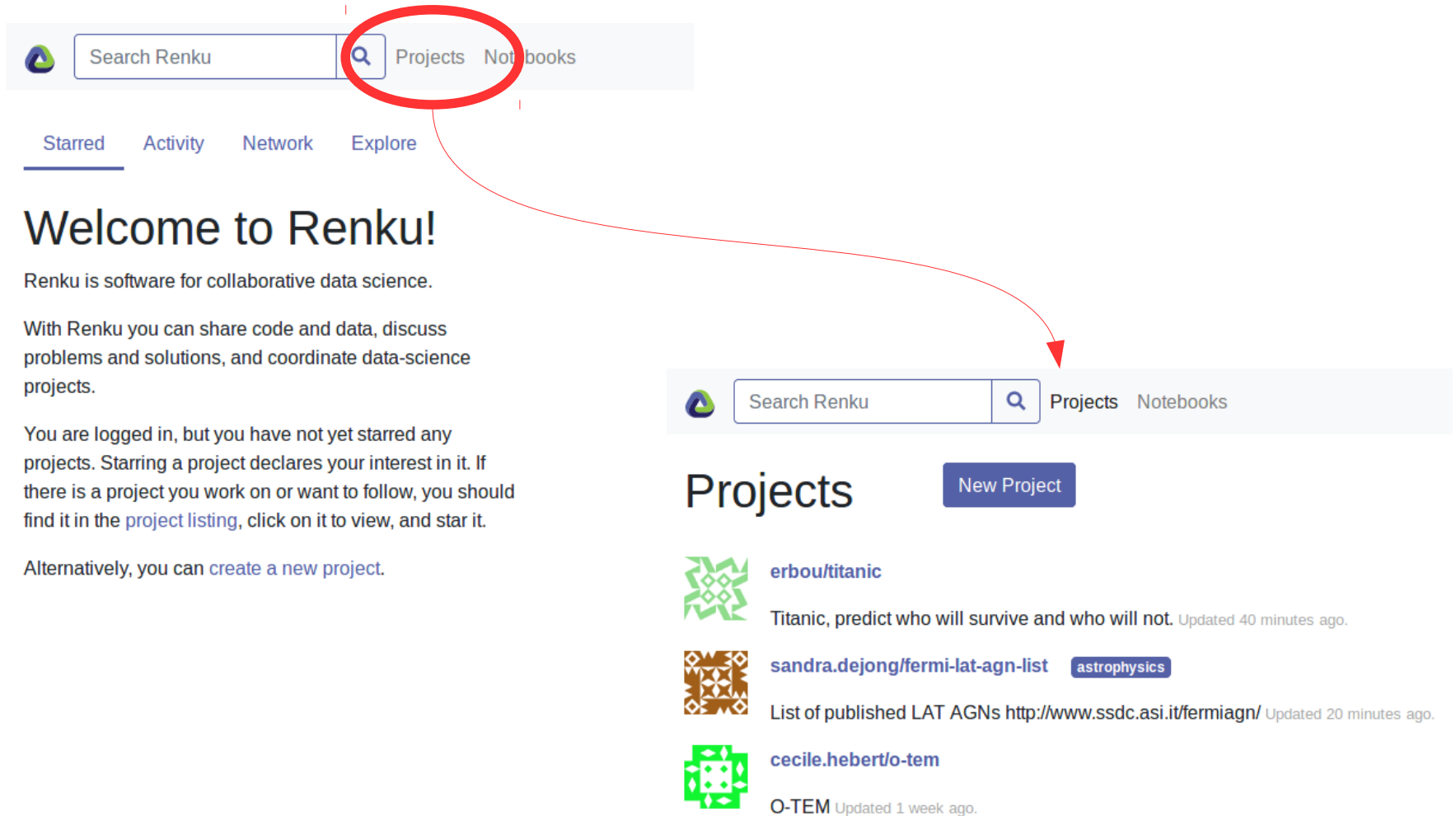
Renku is software for collaborative data science.

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](#).

Interacting with the platform: UI 1



Search Renku **Projects** Notebooks

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](#).

Projects

[New Project](#)

-  **erbou/titanic**
Titanic, predict who will survive and who will not. Updated 40 minutes ago.
-  **sandra.dejong/fermi-lat-agn-list** astrophysics
List of published LAT AGNs <http://www.ssdsc.asi.it/fermiagn/> Updated 20 minutes ago.
-  **cecile.hebert/o-tem**
O-TEM Updated 1 week ago.

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

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.

Overview Kus Files Pending Changes Settings

All

Data

Notebooks

Workflows

.gitignore

.gitlab-ci.yml

.renku/metadata.yml

Dockerfile

requirements.txt

☆ star 0

Notebooks

Initial files upon creation

Interacting with the platform: UI 3

Overview

Kus

Files

Pending Changes

Settings

All

Data

Notebooks

Workflows

CountFSRQ.ipynb

Importnumpy.ipynb

InitialAnalysis.ipynb

Interacting with the platform: UI 3

Overview Kus Files Pending Changes Settings

All

CountFSRQ.ipynb

Data

Importnumpy.ipynb

Notebooks

InitialAnalysis.ipynb

Workflows

Preview of notebook &
quick launch button

Overview Kus Files Pending Changes Settings

All

Launch Notebook

Data

Notebooks

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

Workflows

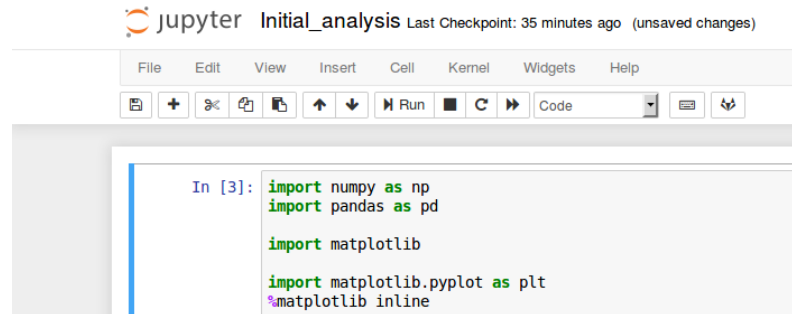
```
[ ] import numpy as np
import pandas as pd
```

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

```
data.count()
```

Interacting with the platform: UI 4

Jupyter notebook:



The screenshot displays the Jupyter Notebook interface. At the top, the title bar reads "jupyter Initial_analysis Last Checkpoint: 35 minutes ago (unsaved changes)". Below this is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. A toolbar contains icons for file operations (new, open, save, reload), navigation (up, down), execution (run, stop, step), and other functions. The main area shows a code cell with the following Python code:

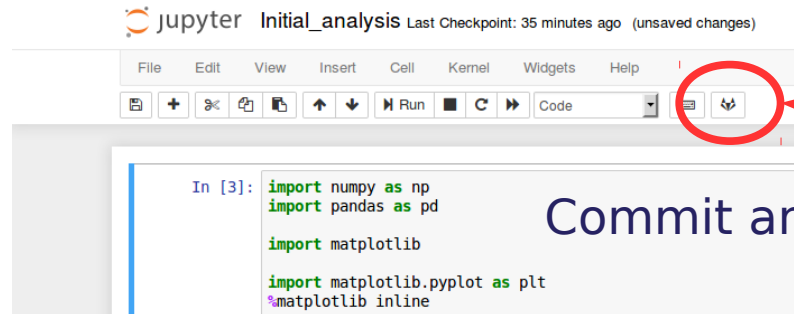
```
In [3]: import numpy as np
import pandas as pd

import matplotlib

import matplotlib.pyplot as plt
%matplotlib inline
```

Interacting with the platform: UI 4

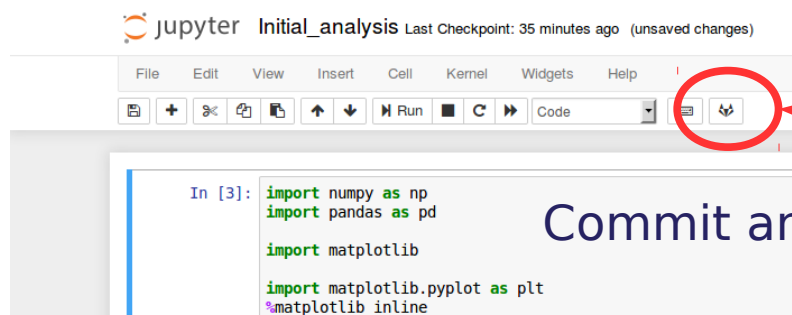
Jupyter notebook:



Commit and push changes

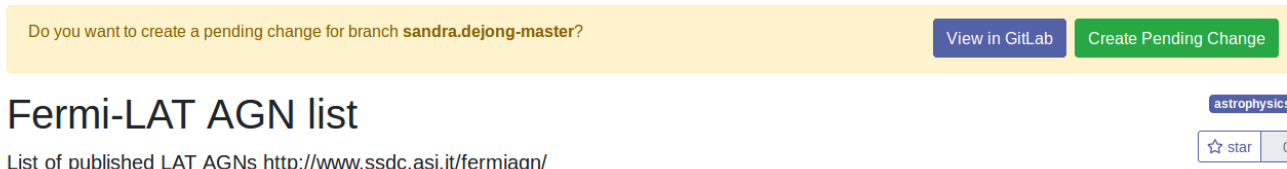
Interacting with the platform: UI 4

Jupyter notebook:



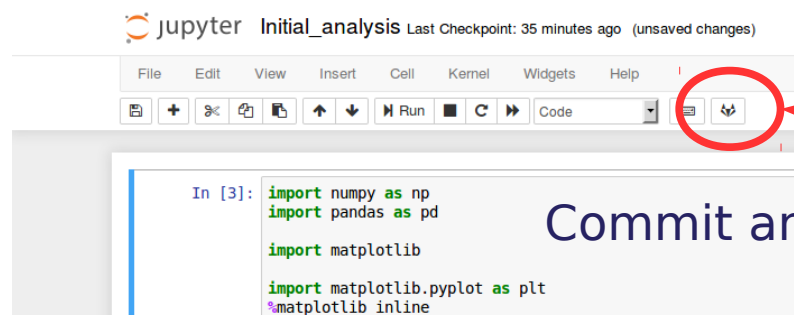
Commit and push changes

In the Renku UI:



Interacting with the platform: UI 4

Jupyter notebook:



Commit and push changes

In the Renku UI:

Do you want to create a pending change for branch **sandra.dejong-master**?

View in GitLab

Create Pending Change

Fermi-LAT AGN list

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

Overview Kus Files **Pending Changes** Settings

sandra.dejong-master

master ←

sandra.dejong-master

Can be merged

sandra.dejong-master

Sandra Savchenko-de Jong wants to merge changes from branch *sandra.dejong-master* into *master*.

Notebook Changes

master

InitialAnalysis.ipynb

```
[ ] import numpy as np
```

sandra.dejong-master

InitialAnalysis.ipynb

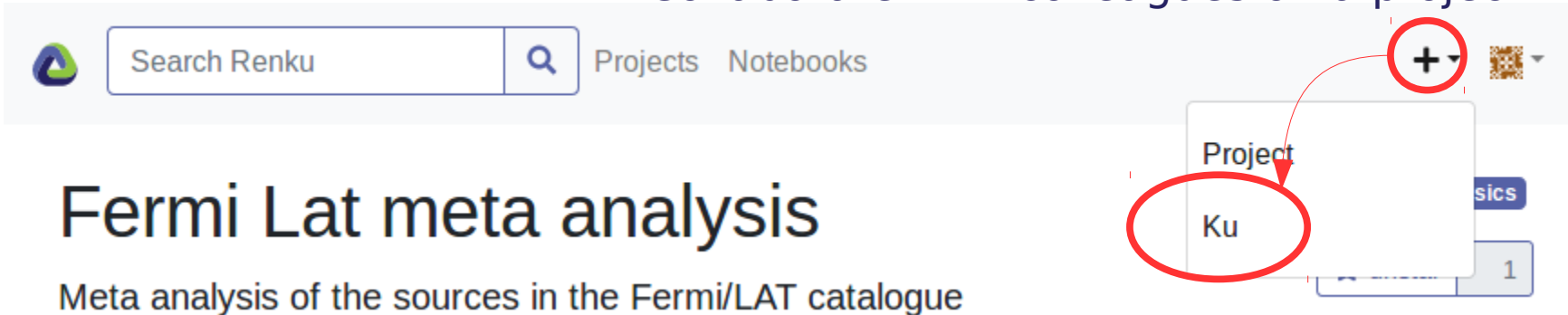
```
[ ] import numpy as np
import pandas as pd
```

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

Compare diff and merge

Collaborating using Ku

Collaborate with colleagues on a project

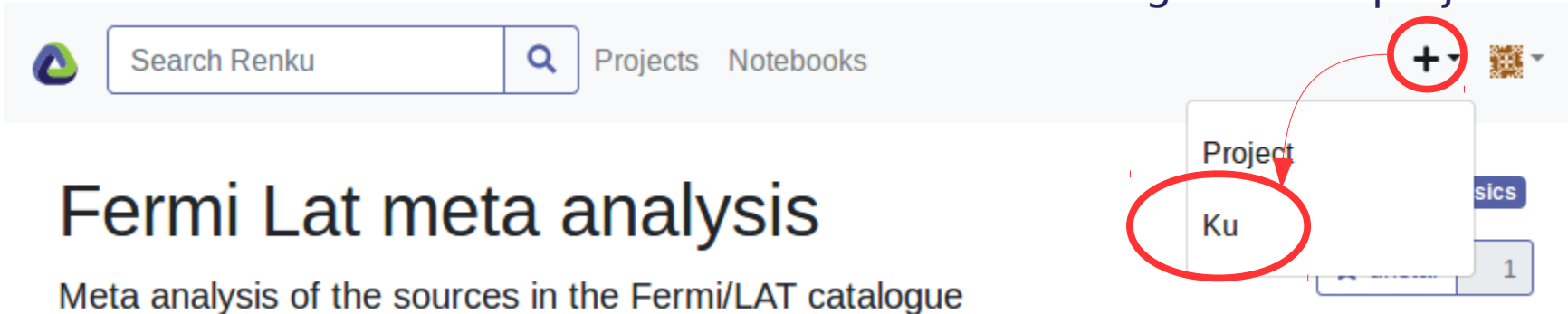


The screenshot shows the Renku web interface. At the top, there is a search bar labeled 'Search Renku' and a magnifying glass icon. To the right of the search bar are the words 'Projects' and 'Notebooks'. Below the search bar, the main heading is 'Fermi Lat meta analysis', followed by the subtitle 'Meta analysis of the sources in the Fermi/LAT catalogue'. On the right side of the interface, there is a sidebar with a 'Project' section. In this section, a card labeled 'Ku' is highlighted with a red circle. Above the 'Project' section, there is a red circle containing a plus sign and a dropdown arrow, with a red arrow pointing from it to the 'Ku' card. To the right of the 'Project' section, there is a 'Issues' section with a card labeled '1'.

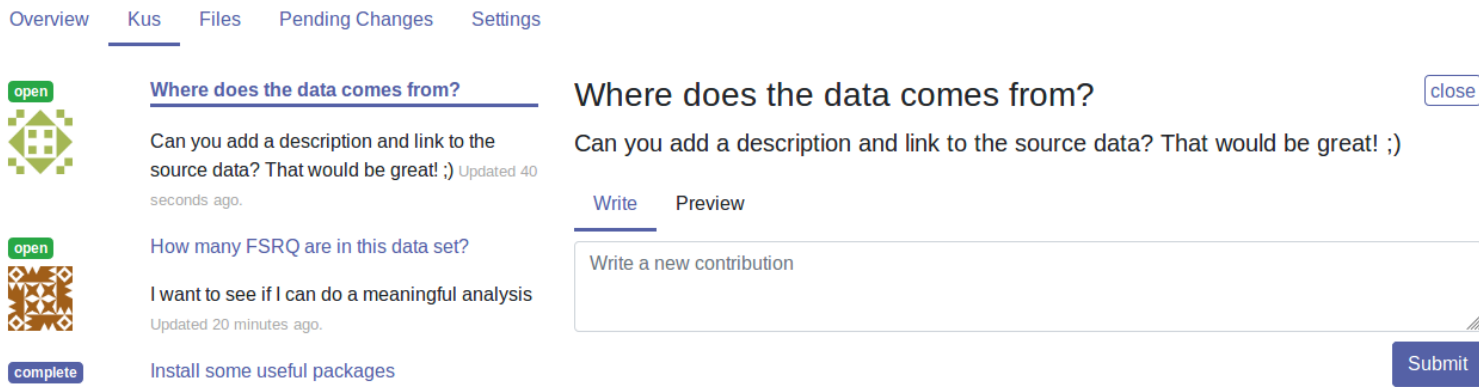
Extention of Gitlab 'Issues'

Collaborating using Ku

Collaborate with colleagues on a project



Extention of Gitlab 'Issues'



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

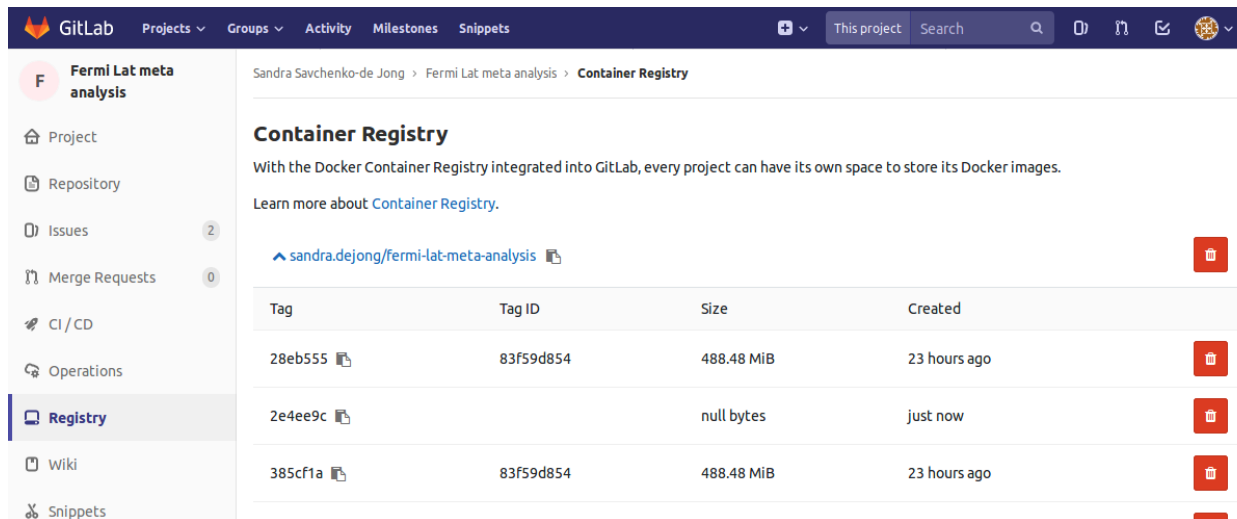
```
[1] import numpy as np
import pandas as pd
```

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

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



The screenshot shows the GitLab web interface for a project named 'Fermi Lat meta analysis'. The left sidebar contains navigation links: Project, Repository, Issues (2), Merge Requests (0), CI / CD, Operations, Registry (selected), Wiki, and Snippets. The main content area is titled 'Container Registry' and includes a description: 'With the Docker Container Registry integrated into GitLab, every project can have its own space to store its Docker images.' Below this, there is a link to 'Learn more about Container Registry.' and a link to the specific registry path 'sandra.dejong/fermi-lat-meta-analysis'. A table lists the Docker images stored in the registry:

Tag	Tag ID	Size	Created
28eb555	83f59d854	488.48 MiB	23 hours ago
2e4ee9c		null bytes	just now
385cf1a	83f59d854	488.48 MiB	23 hours ago

→ Retrieve state of project at each commit !

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

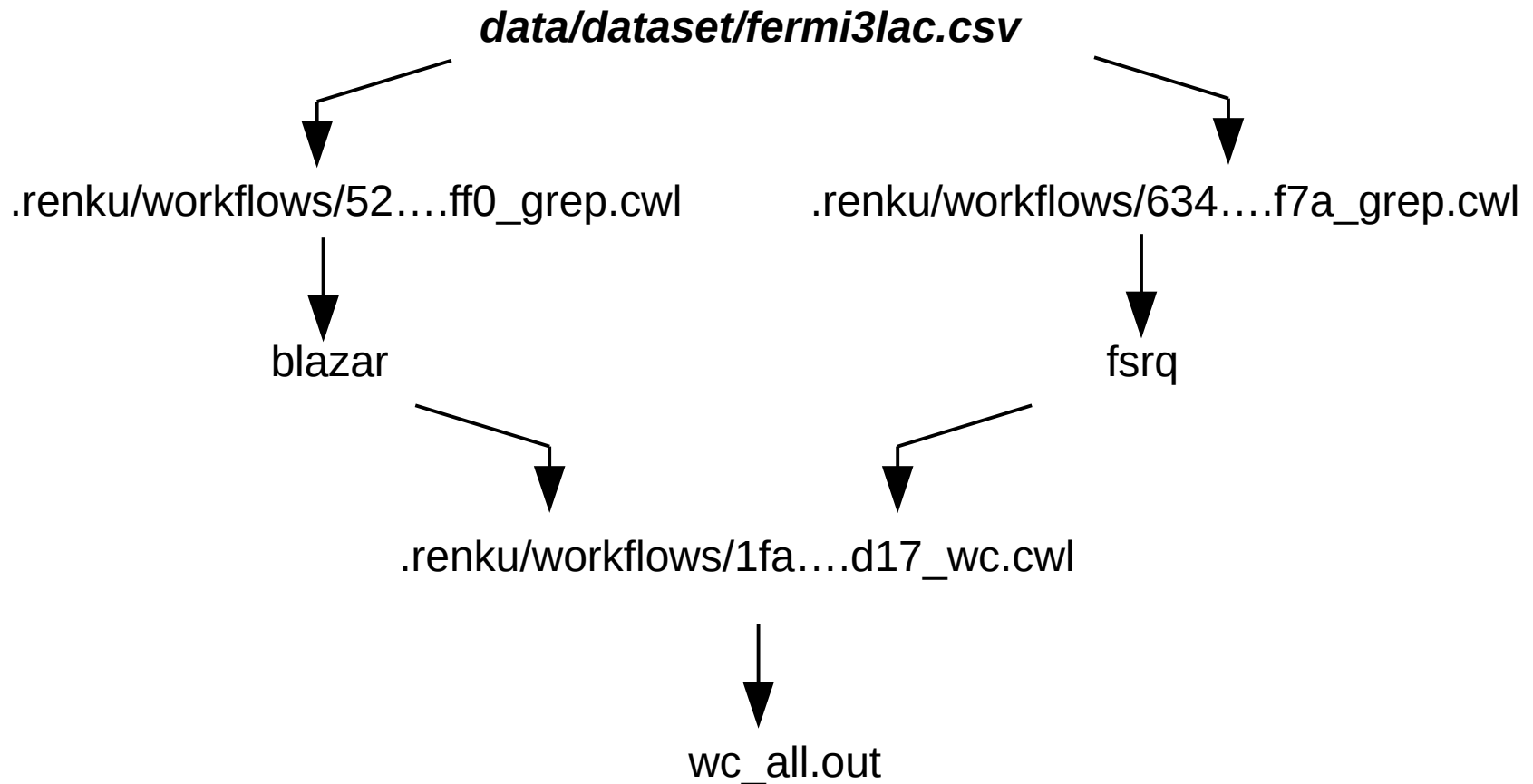
```
sandra@Charon ~fgl $ renku run grep -i fsrq data/dataset/fermilac3.csv > fsrq
sandra@Charon ~fgl $ renku log fsrq
* e4d16b65 fsrq
* e4d16b65 .renku/workflow/3e43b888a4c848528d80dda6be7c1c10_grep.cwl
@ a5c2fcfc data/dataset/fermi3lac.csv
```

```
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



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