

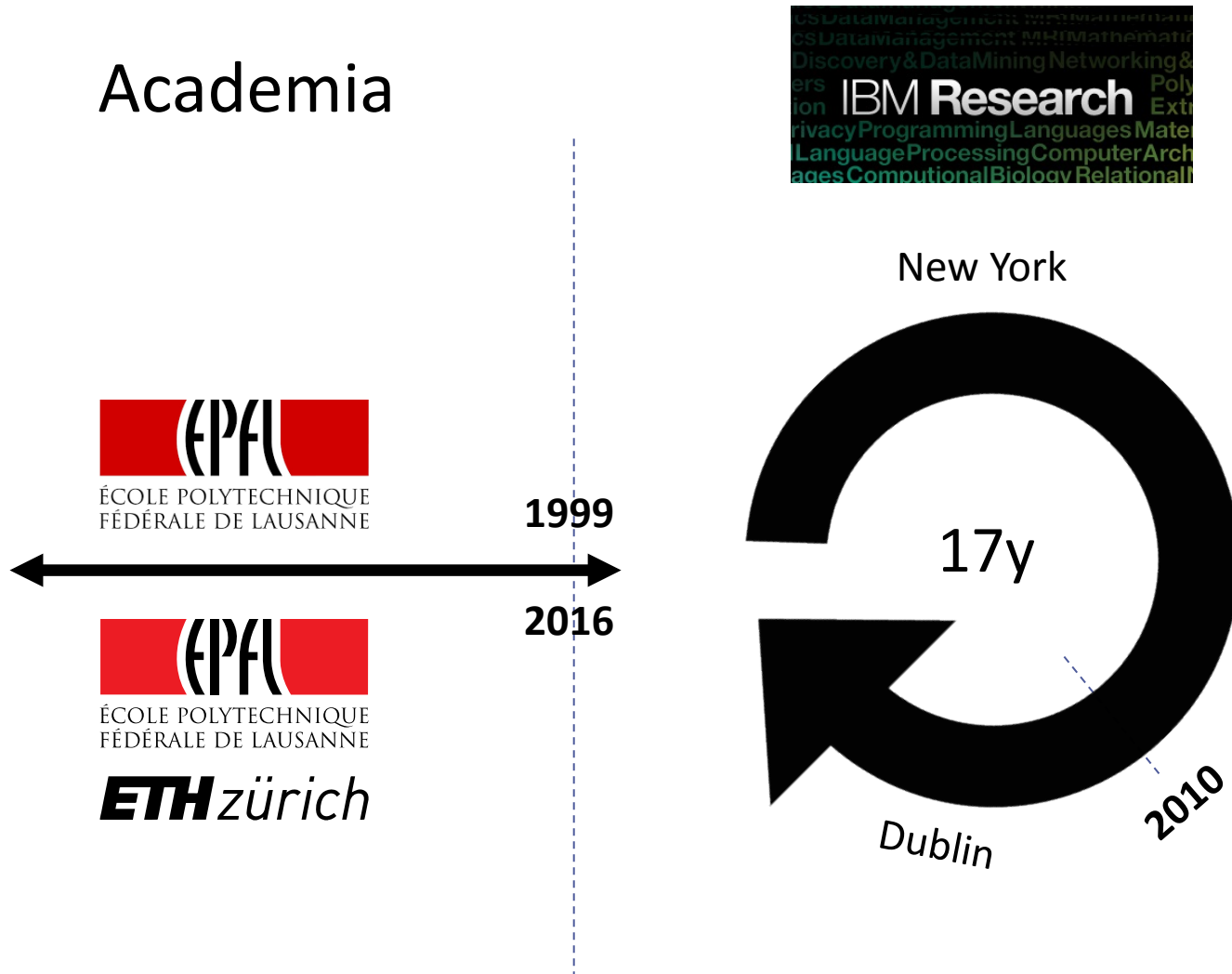
# SWISS DATA SCIENCE CENTER

A joint center between EPFL and ETH Zürich

*Olivier Verscheure*

# About me

## Academia



# What Do You See?



O'Connell Bridge / D'Olier St. Dublin City CCTV  
8 Apr 2013 18:31:50 GMT Daylight Time

# Dublin City Data Hub



# Big Data, Bad Data



"YEP... GOT MY CELLPHONE, MY PAGER, MY INTERNET LINK, MY WIRELESS FAX, AND THANKS TO THIS NIFTY SATELLITE NAVIGATION SYSTEM, I KNOW PRECISELY WHERE I AM AT ALL TIMES!"

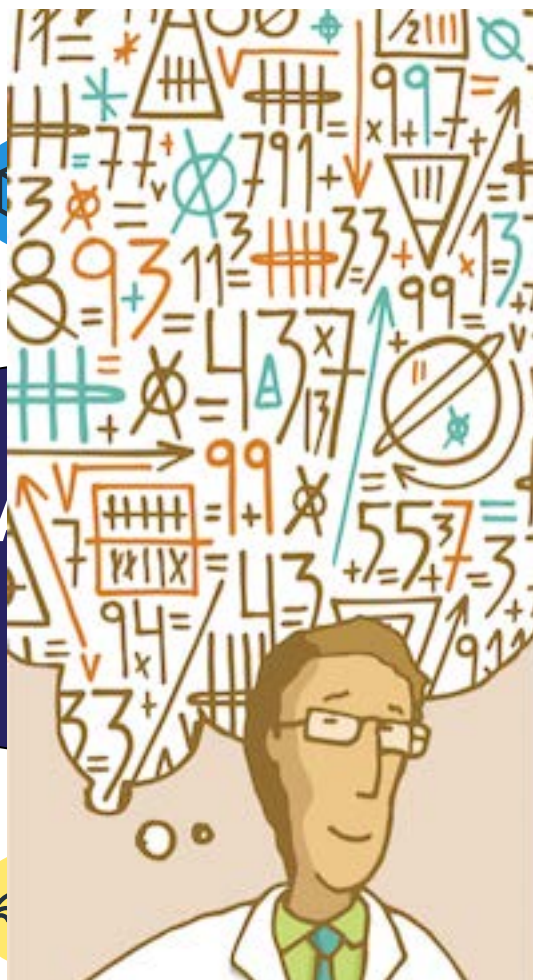
BY LOWE FOR THE SUN-SENTINEL, FLOA



# A fragmented ecosystem



DATA  
SCIENCE



Visual What is the hyperplane that best separates two classes of points in multidimensional space?

↔  
**GAP**

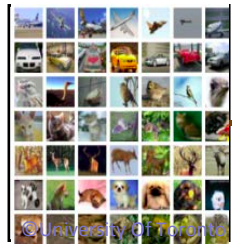


How can I best match the right drug with the right dosage to the right patient at the right time?

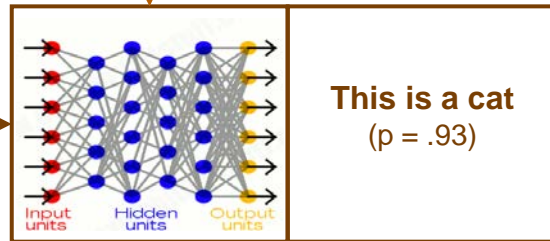


# Explainable AI – What Are We Trying To Do?

Today



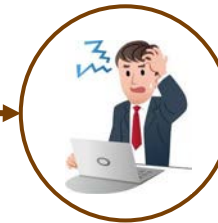
Training Data



Learned Function

This is a cat  
( $p = .93$ )

Output



User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

# Fooling deep neural net classifiers

**Title:** Universal adversarial perturbations

**Authors:** [Moosavi-Dezfooli](#), [Seyed-Mohsen](#); [Fawzi, Alhussein](#); [Fawzi, Omar](#); [Frossard, Pascal](#)

**Publication:** eprint arXiv:1610.08401

**Publication Date:** 10/2016



This is not a woolen sock

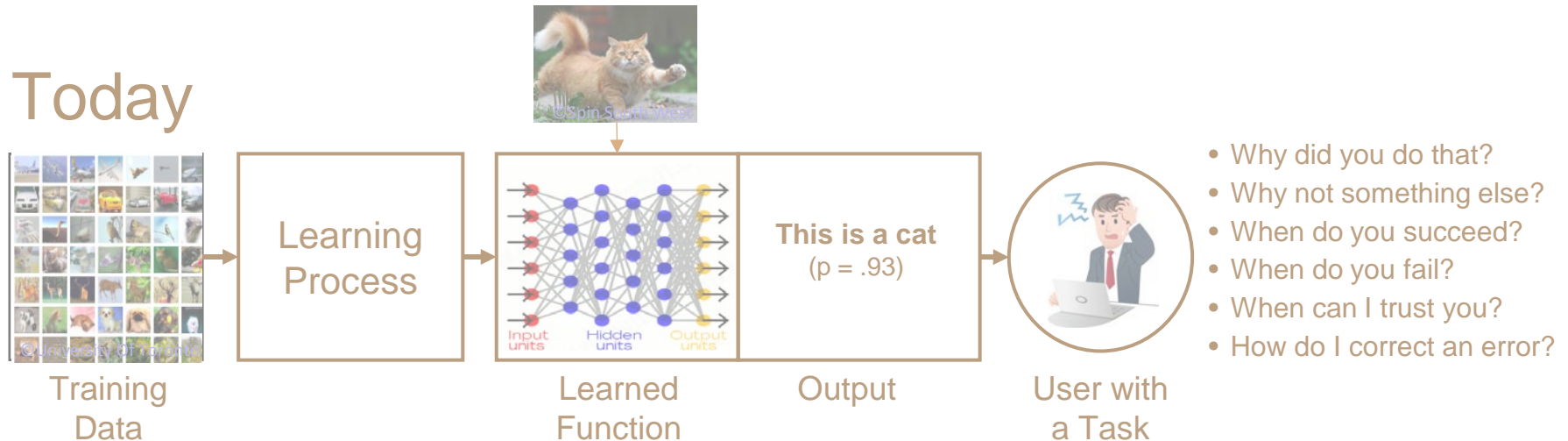
- It's an Indian elephant!
- At least after adding a universal noise to the image
- Deep learning models do not mimic brain activity



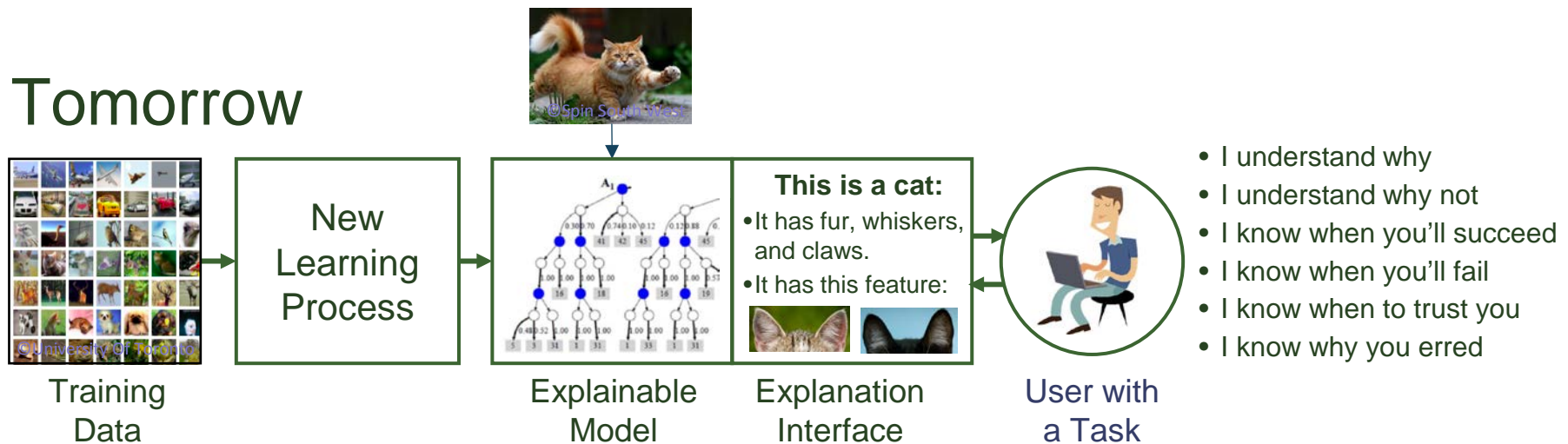


# Explainable AI – What Are We Trying To Do?

## Today



## Tomorrow



# Swiss Data Science Center (SDSC)

Foster adoption of data science by both academia and industry scientists, and domain experts



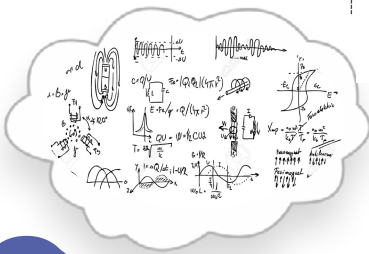
How can I best match the right drug with the right dosage to the right patient at the right time?

ETH zürich

+



What is the hyperplane that best separates two classes of points in multidimensional space?



Domain experts

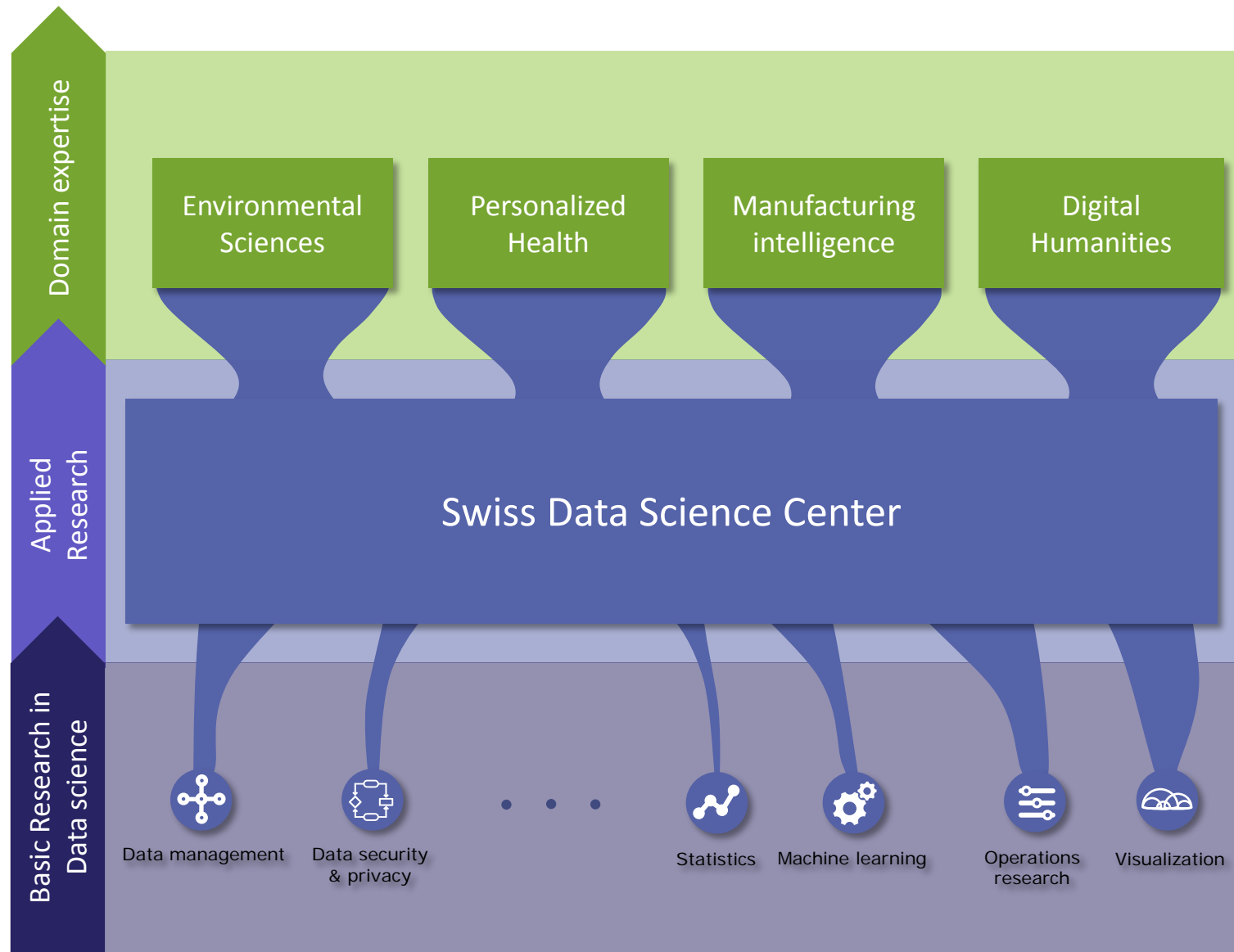
Data scientists

Data providers

How is my data protected?  
How private is it?  
How exactly is it used?

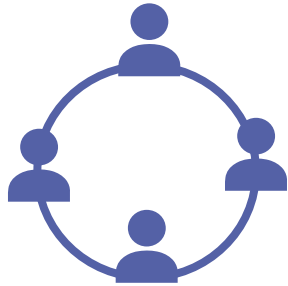


# Where does SDSC fit?



# What will the SDSC offer?

Excellence in academic research backed by strong industrial experience



## Embedded R&D collaboration

We engage in academic and industrial collaborations requiring large-scale distributed data processing (Big & Fast Data) and/or advanced analytics (machine learning & statistics) combined with an in-depth knowledge in select domains



## Domain-specific Insights as a Service

We provide secure access to our cloud-hosted analytics platform - RENGa, a highly scalable open software platform offering a one-stop-shop for hosting and exploring curated, calibrated and possibly anonymized data at scale, at-rest or in-motion.



## Open (Data) Science

RENGa offers user-friendly tooling and services to help with the adoption of Open Science, fostering research productivity and excellence.

# Answering Researchers Challenges

- A data lake, not a data swamp!
  - Where can I upload my data, and make it available?
  - What other data is available? And where is it?
  - How was this data created? Who created it?
  - How trustful is it? Can I build my research on it?

... impedes **collaboration** between scientists, and **reusability** and **reproducibility** of research

- Data science made simple & trustable
  - Combining human expertise and machine intelligence
  - Making learning methods robust against uncertainties
  - Designing methods for interpretable machine learning

# RENGA - 連歌, collaborative poetry

**Renga** (連歌), plural **renga**, a genre of Japanese linked-verse poetry in which two or more poets supply alternating sections of a poem linked by verbal and thematic associations.

—Encyclopædia Britannica

- **Renga** is a highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaboration across mutually untrusted academic and industrial institutions.
- The platform allow practitioners to:
  - Securely manage, share and process large-scale data across untrusted parties operating in a federated environment.
  - Capture complete lineage automatically up to original raw data for detailed traceability (auditability & reproducibility)
  - Adhere to FAIR principles and enable reproducible research



# RENGA's Software Stack

Users, data scientists

Domain specific notebook and analytics

SDK

Environmental Science

SDK

Transport

SDK

Personalized Health

SDK

Social science

SDK

SDSC and partners

Micro services

- Security
- Privacy
- Federation
- Reproducible research
- Data protection

play

akka

lagom

Common API

Data Processing Applications



Open Big Data Platform Stack

DC/OS

APACHE Spark

kafka

APACHE HBASE

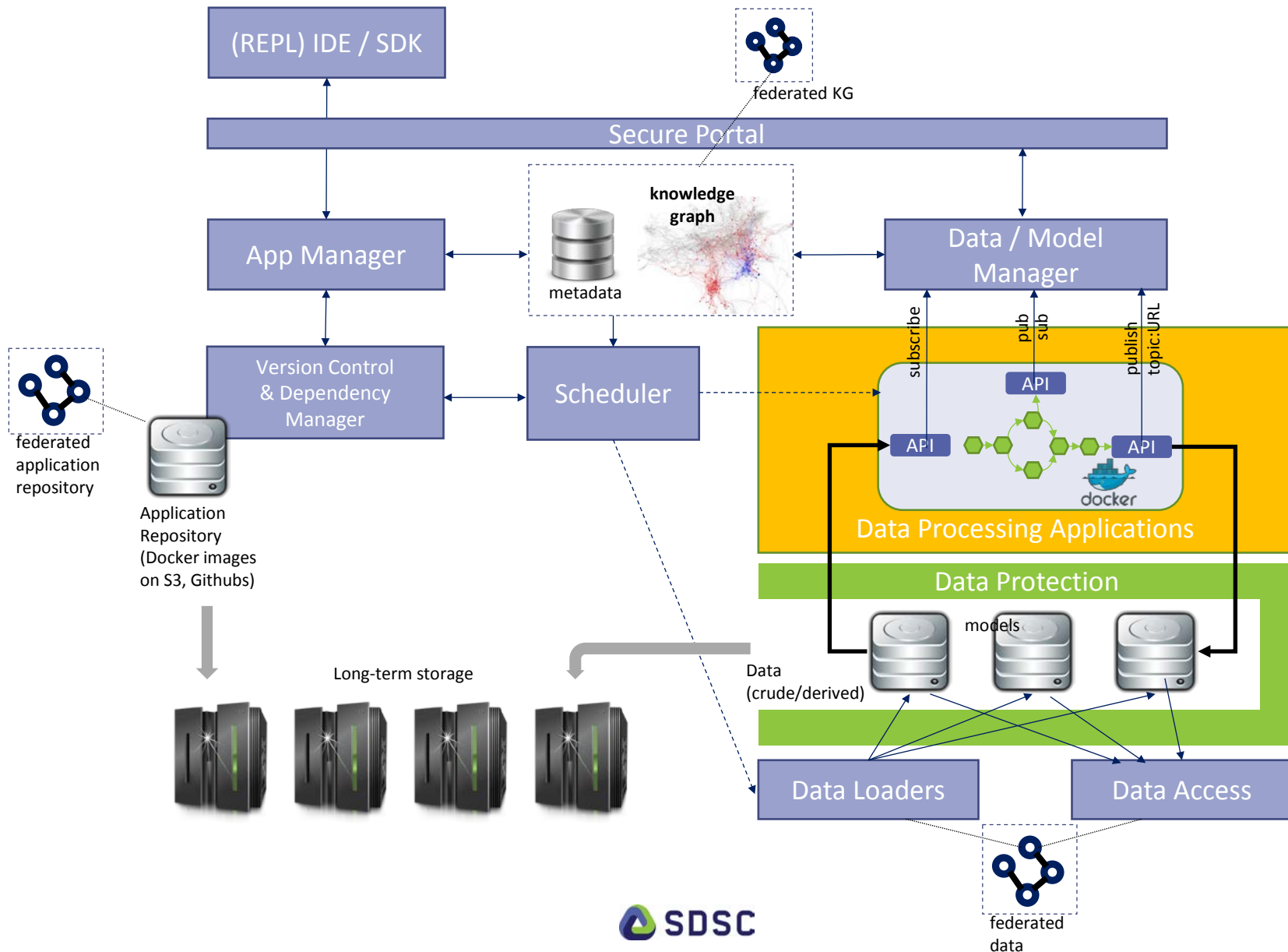
Flink

IaaS

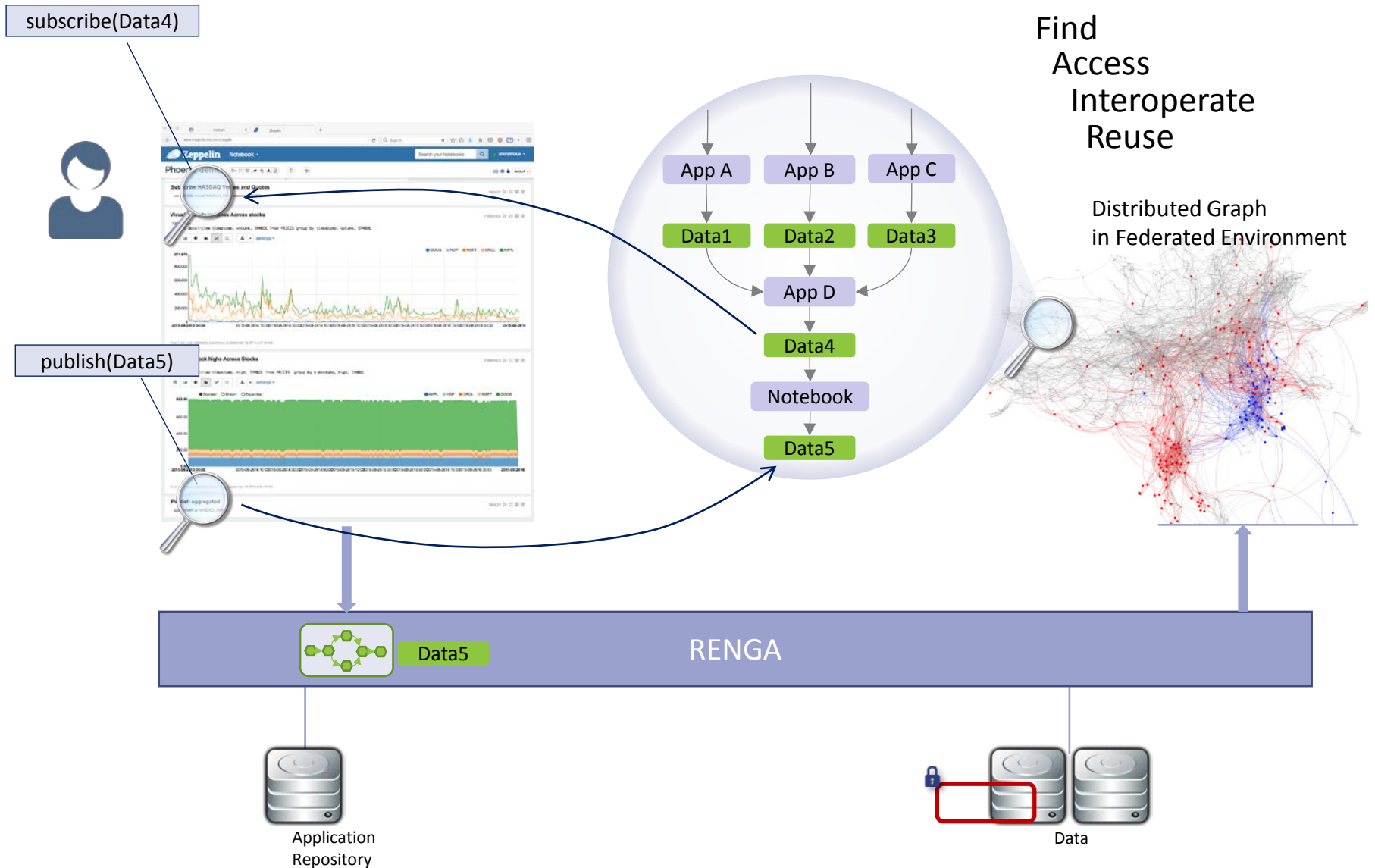
Geographically distributed cloud and on premise infrastructure(s) + long term storage solution providers for archiving



# Building the Knowledge Graph

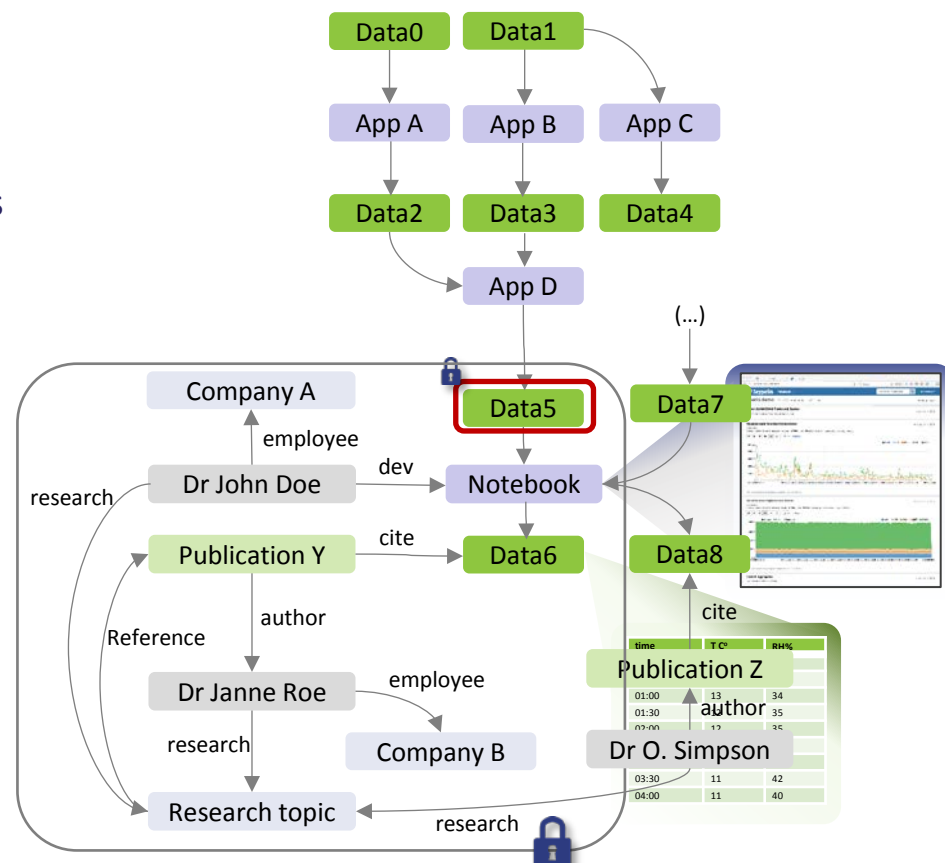


# Provenance of Data Science



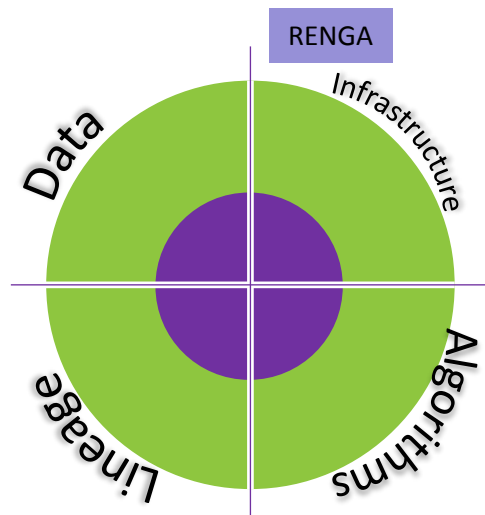
# Automated Open Science

- Reproducible Research
  - See the (versioned) algorithms
  - See the data
  - Replay a workflow
  - Compare workflows, validate robustness
- Reusability
  - Reuse data on new workflows
  - Clone and modify workflows
- Knowledge Graph
  - Data popularity, H-index
  - Who is using the data?
  - For what?
- IP Protection
  - Decide who sees the data,
  - The algorithms,
  - The data I use,
  - And how I use it

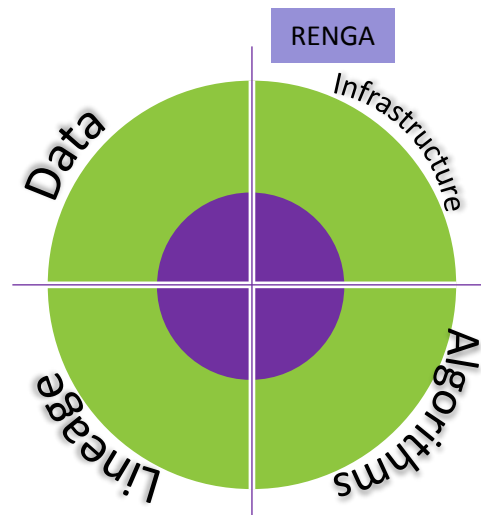


# Abstract Use Case

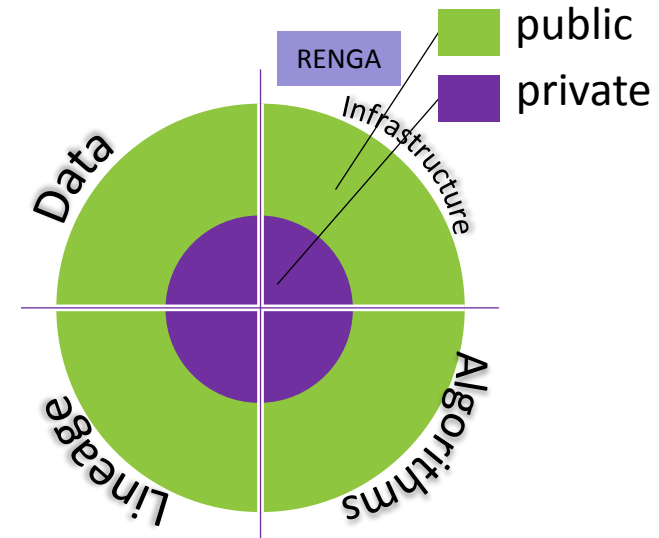
Federated Parties



Genome Center



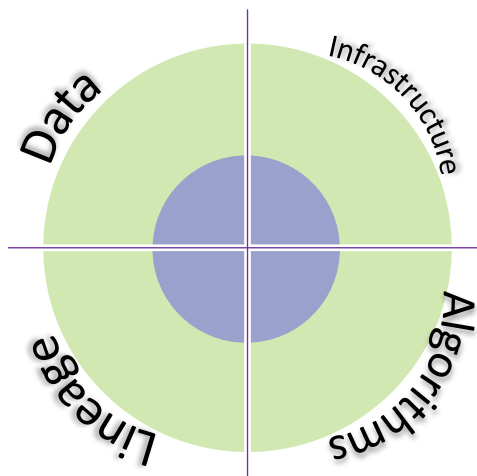
Proteome Center



USZ

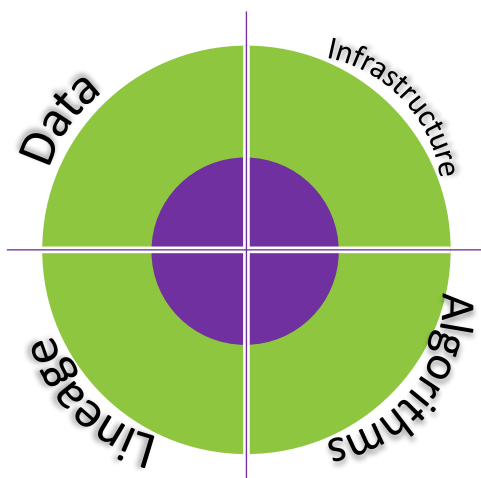
# Abstract Use Case

One stop shop

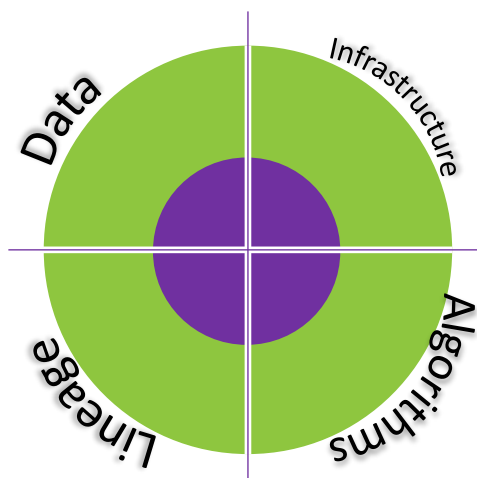


RENGA

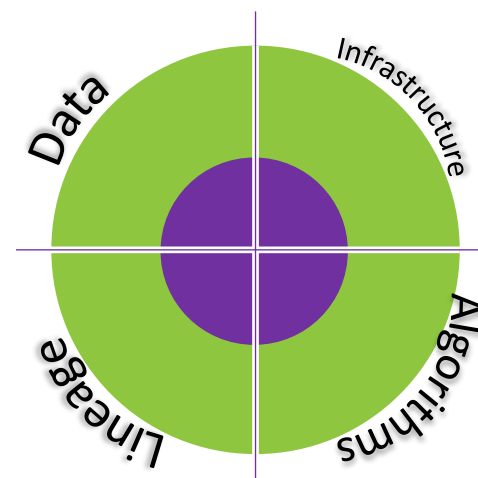
Federated Parties



Genome Center



Proteome Center

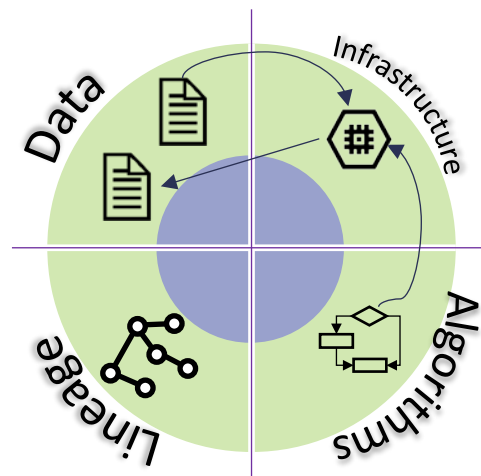


USZ



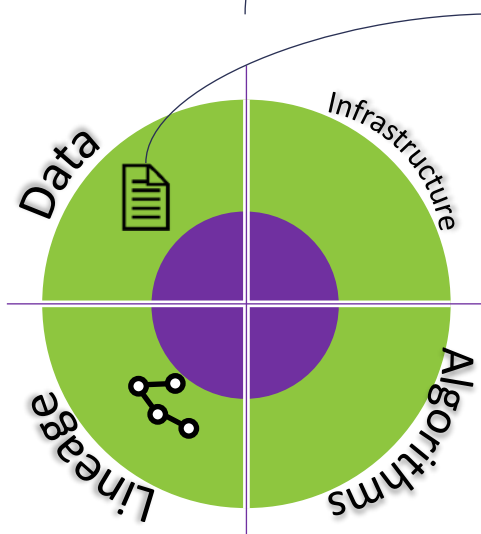
# Abstract Use Case

One stop shop

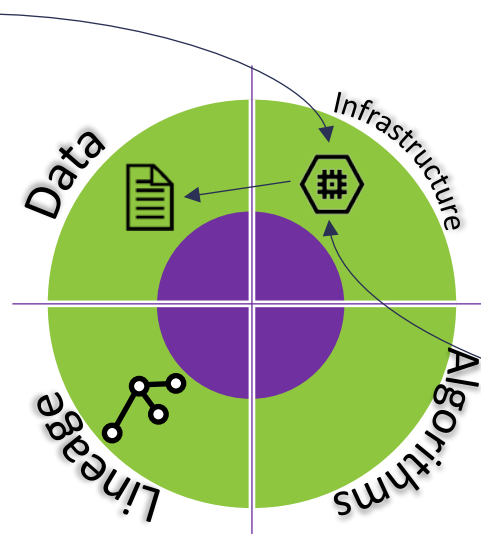


RENGA

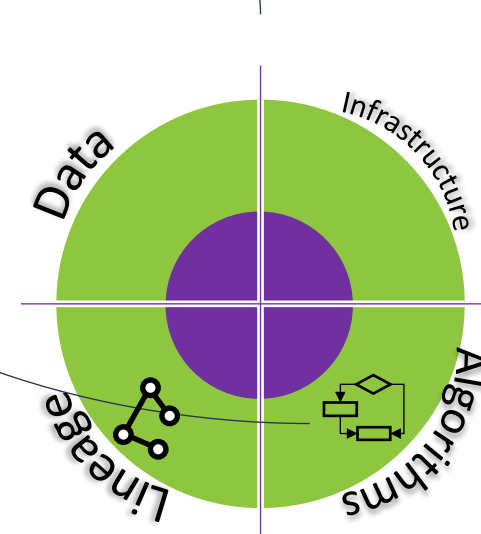
Federated Parties



Genome Center



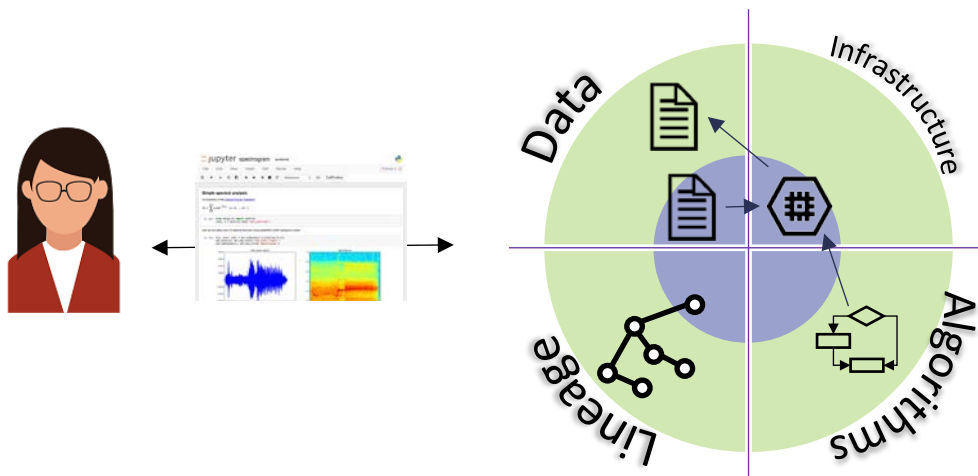
Proteome Center



USZ

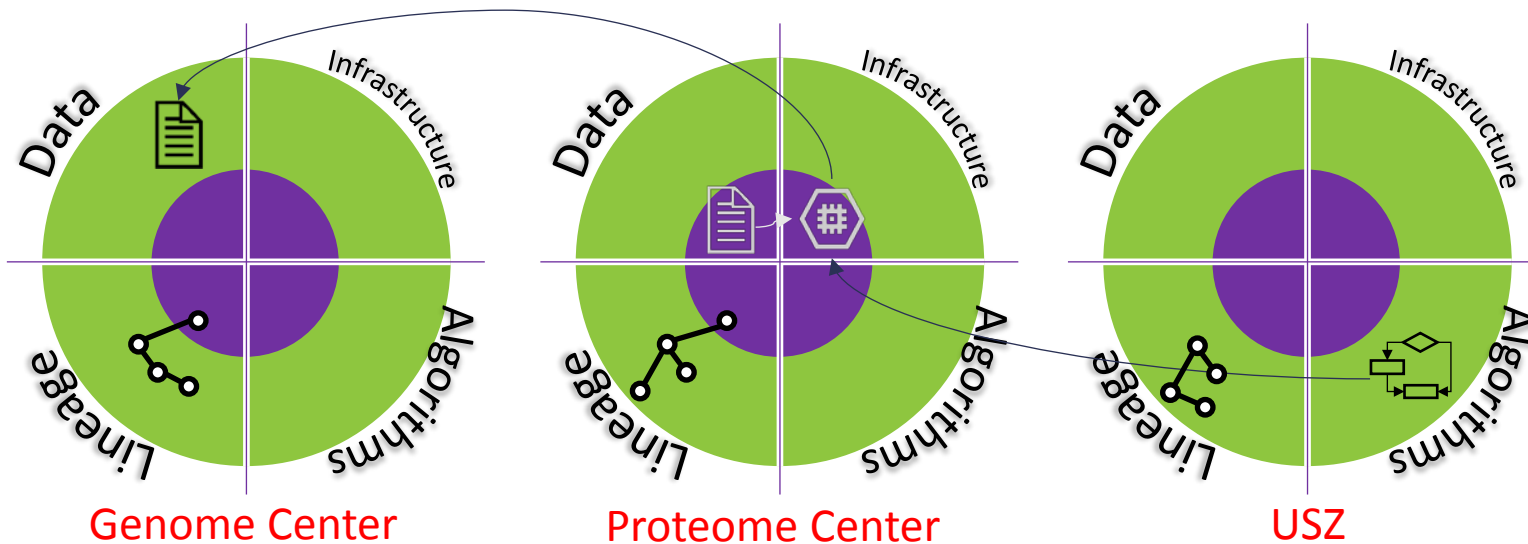
# Abstract Use Case

One stop shop



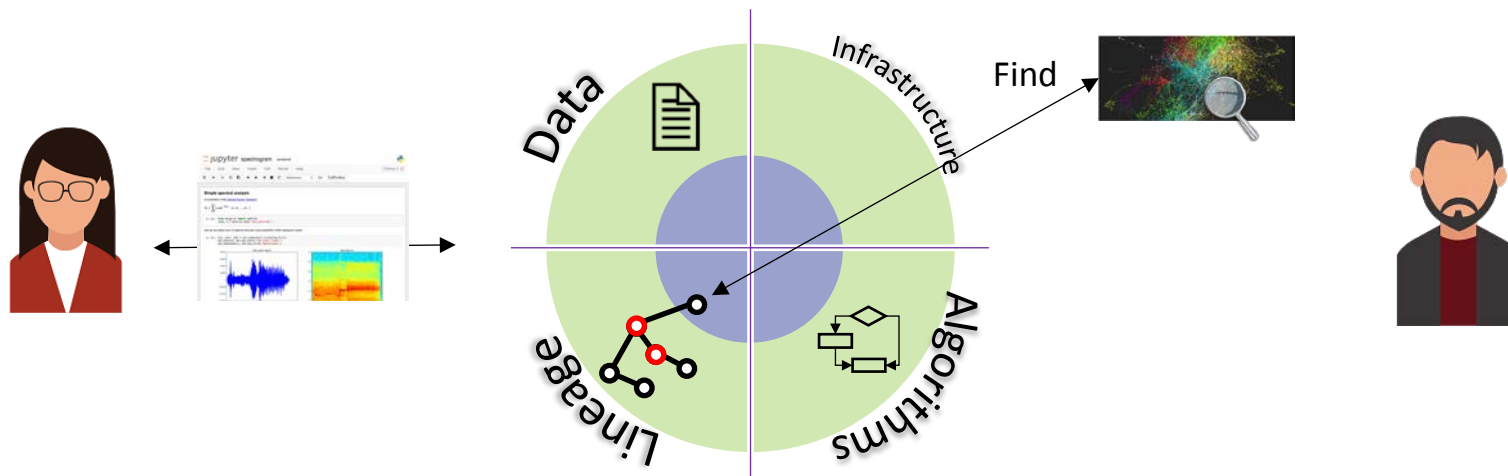
RENGA

Federated Parties



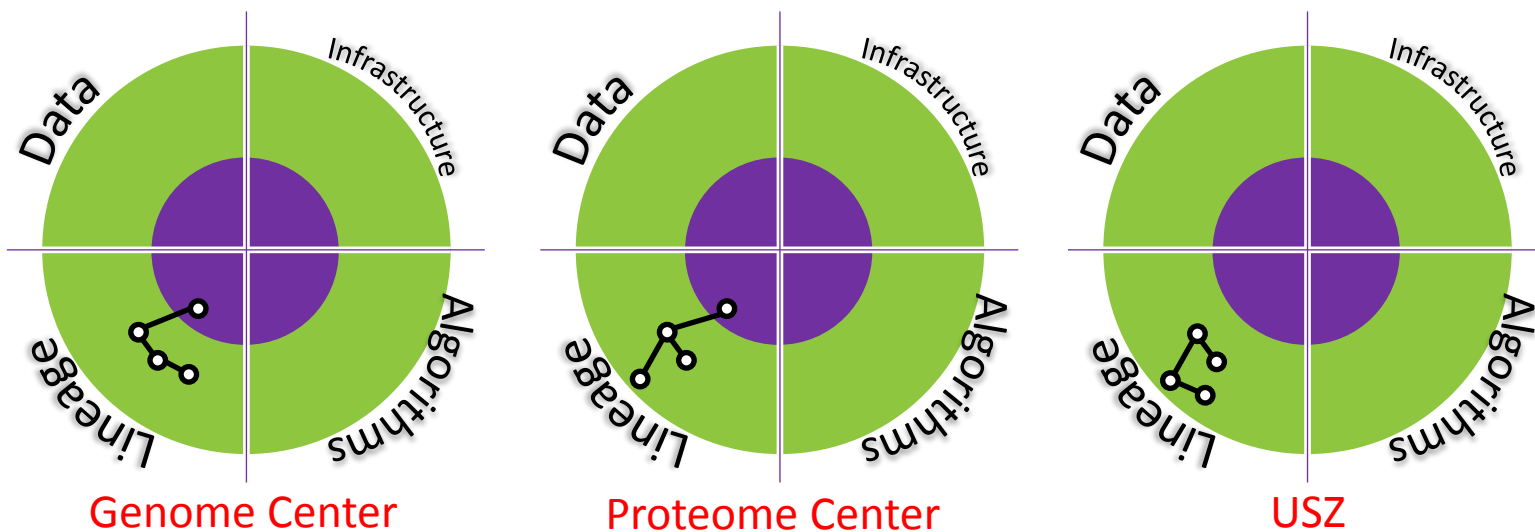
# Abstract Use Case

One stop shop



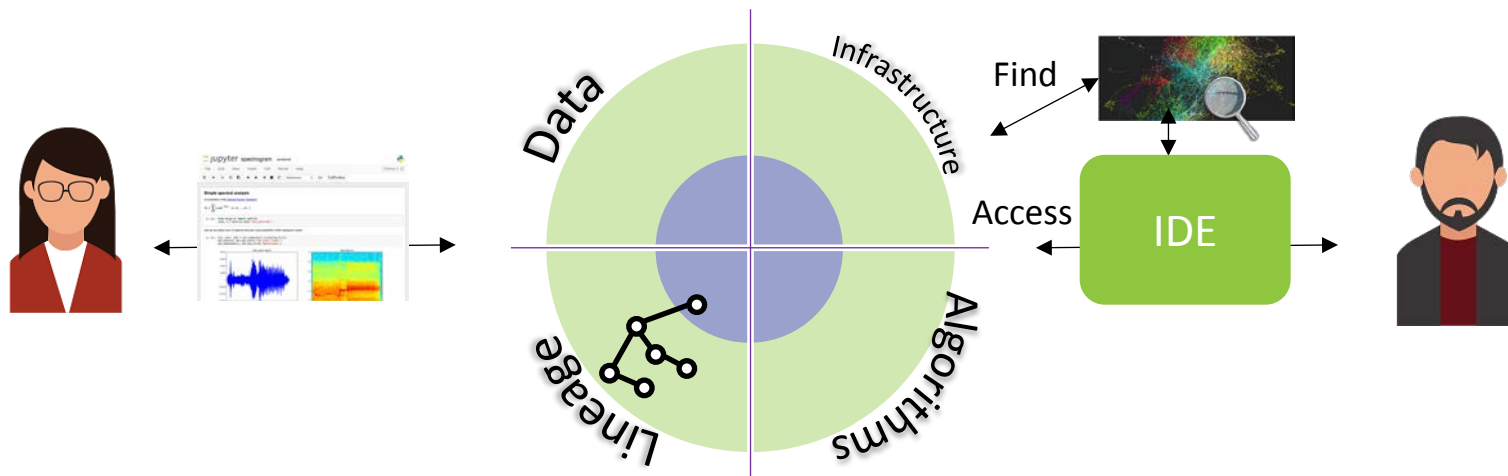
RENGA

Federated Parties

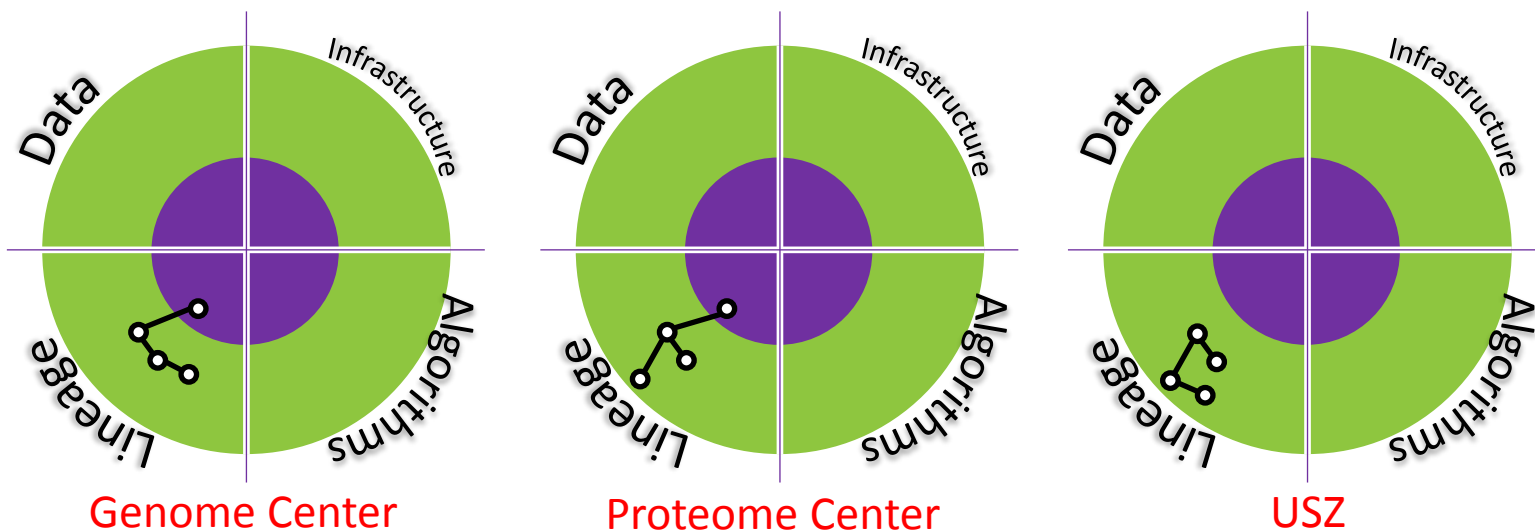


# Abstract Use Case

One stop shop

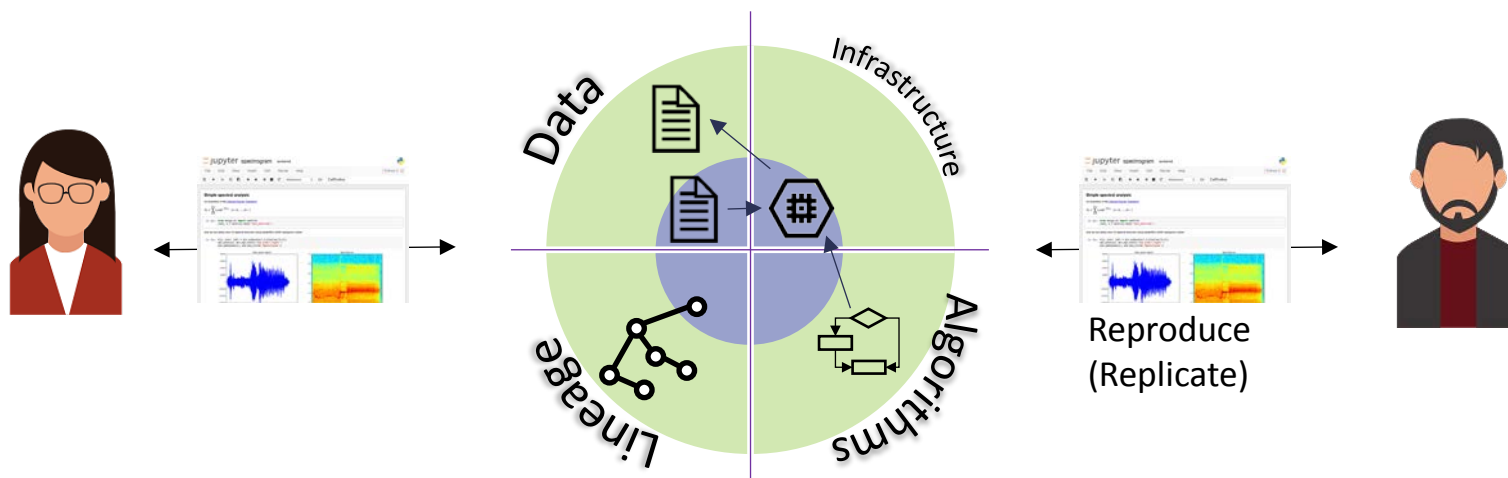


Federated Parties



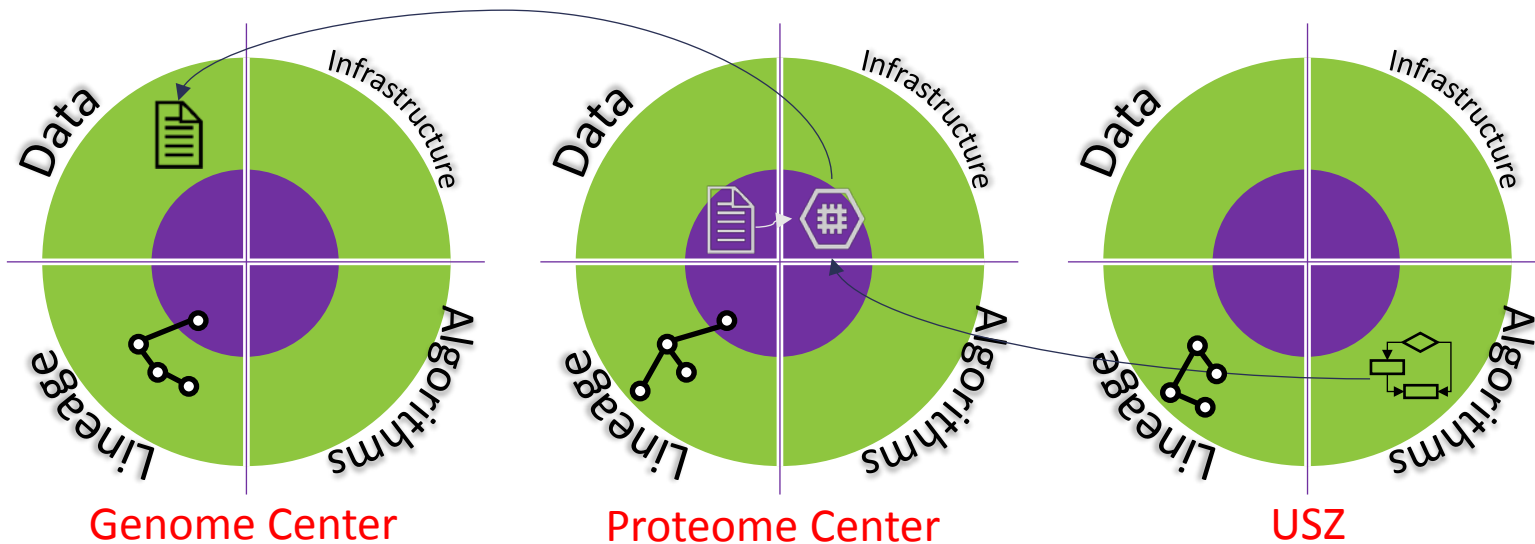
# Abstract Use Case

One stop shop



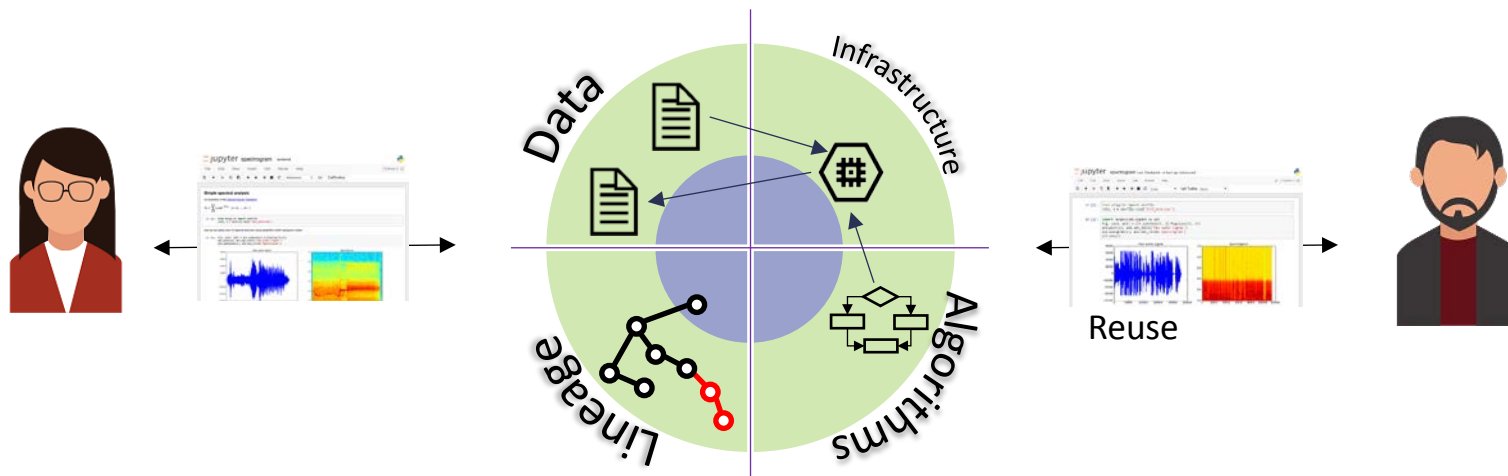
RENGA

Federated Parties

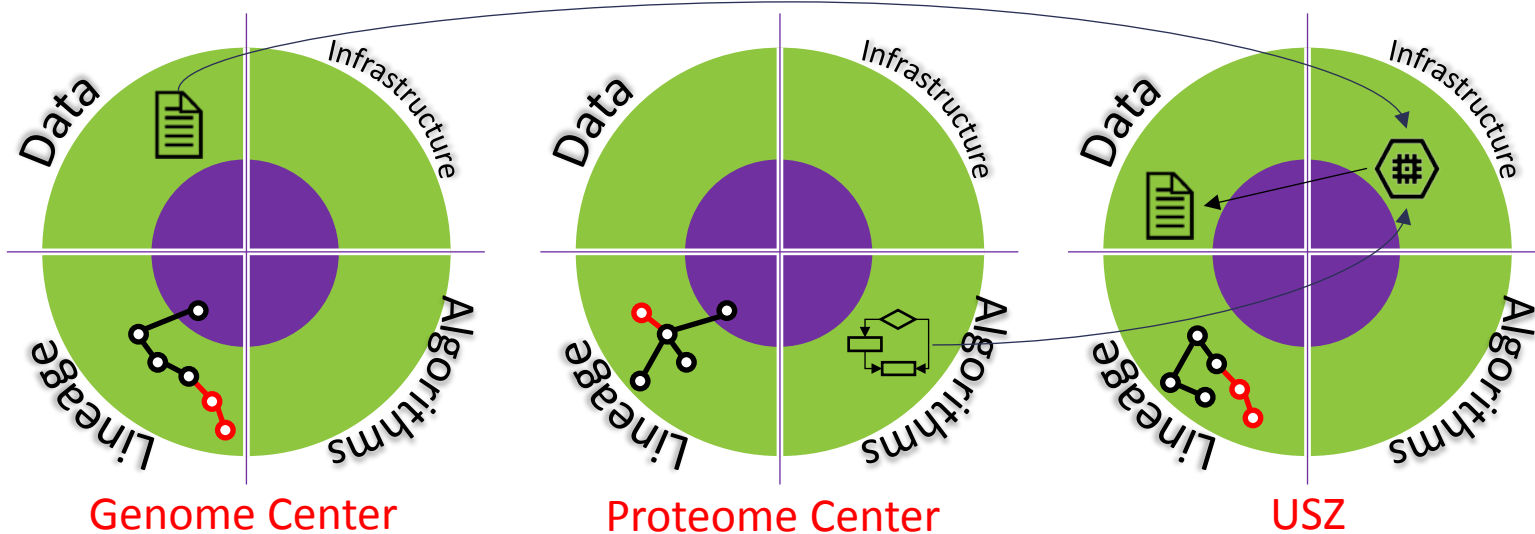


# Abstract Use Case

One stop shop



Federated Parties





# SDSC Academic Projects

1. Attribution of regional changes in the water cycle
2. ACE: Delivering Added-value To Antarctica (ACE-DATA)
3. CrowdAI – an open data science challenge platform
4. Reducing data needs and improving robustness of deep learning methods for segmentation in microscopic images (DeepMICROIA)
5. Interpretable learning methods for immunotherapy (iLearn)
6. Four-dimensional mapping of carbon dioxide using a low power sensor network
7. Improving species biodiversity analyses and citizen science feedback through mining data (SPEEDMIND)
8. Deep learning for observational cosmology

# Get Renga (get-renga.io)



MULTIDISCIPLINARY DATA SCIENCE COLLABORATIONS MADE EASY

GETTING STARTED

DOWNLOAD 0.1.0

Preview beta 0.1.0 released on September 15, 2017

Renga (連歌), plural renga, a genre of Japanese linked-verse poetry in which two or more poets supply alternating sections of a poem linked by verbal and thematic associations.

—Encyclopædia Britannica

RENGA is a highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaborations.

## SECURE

Securely manage, share and process large-scale data across untrusted parties operating in a federated environment.

## TRACEABLE

Capture complete lineage automatically up to original raw data for detailed traceability (auditability & reproducibility).

## FAIR

Work in a Findable, Accessible, Interoperable and Re-usable principles environment

## RENGA 連歌

Highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaboration across mutually untrusted academic and industrial institutions.

## Navigation

Table of Contents:

[Introduction](#)

[First steps](#)

- Our first Renga project
- Creating a project deployment
- Creating and populating a storage bucket

[User documentation](#)

[Developer](#)

[documentation](#)

[License](#)

[renga@GitHub](#)

## Quick search

## First Steps

To try out Renga, you first need a platform to connect to: see [Running the platform](#) for instructions on how to get one running on your own machine in a few minutes.

Interaction with the platform takes place via the Python-based command-line interface (CLI) and the Python API. You can get both via pip:

```
$ pip install renga
```

### Note:

We recommend using [virtualenv](#) when installing the Renga package.

## Our first Renga project

First, create a project directory:

```
$ mkdir -p ~/renga-projects/test-project  
$ cd ~/renga-projects/test-project
```

Set up your platform credentials (using the demo `docker-compose` configuration, enter `demo/demo` for username/password:

```
$ renga login http://localhost  
Username:  
Password:  
Access token has been stored in: ~/Library/Application Support/Renga/
```



# Current Status & Next Steps

The center is fully operational as of January 2017

## Center set-up

## SDSC Industry Day

## Building the community beyond the ETH Domain

In progress

October 2017

2018 & beyond

- Hiring R&D staff
- Developing hosted platform
- **Collaborating across the Swiss academic community**
  - Personalized health
  - Environmental science
- **Engaging with industry**
  - Bühler Group, HESS, P&G, Medtronic and others on preventive maintenance

By invitation only

### Objectives

- Showcase R&D activities of the center
- Offer a platform for industry to engage with SDSC

### National

- SNSF NRP 75 'Big Data'
- Strategic partner

### International

- Organizing IEEE workshop on Data Science sponsored by the Big Data SIG of the IEEE Signal Processing Society on June 2018
- 300 participants
- Open Challenges



THANK YOU!

<http://www.datascience.ch>

**Twitter:** @SDSCdatascience