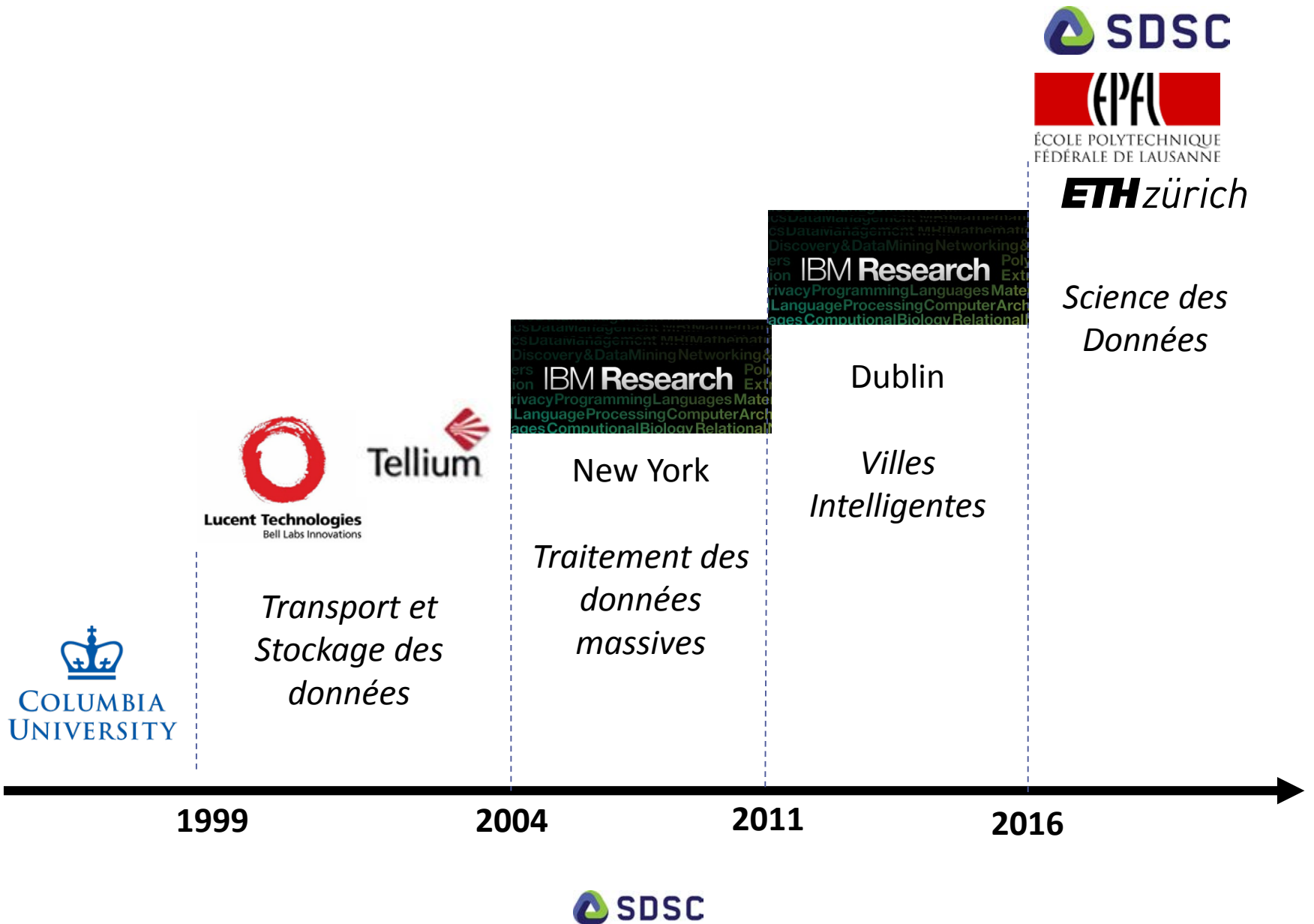




SWISS DATA SCIENCE CENTER

Une collaboration entre EPFL et ETH Zürich

Eric Bouillet



Que voyez vous?



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

Dublin City Data Hub



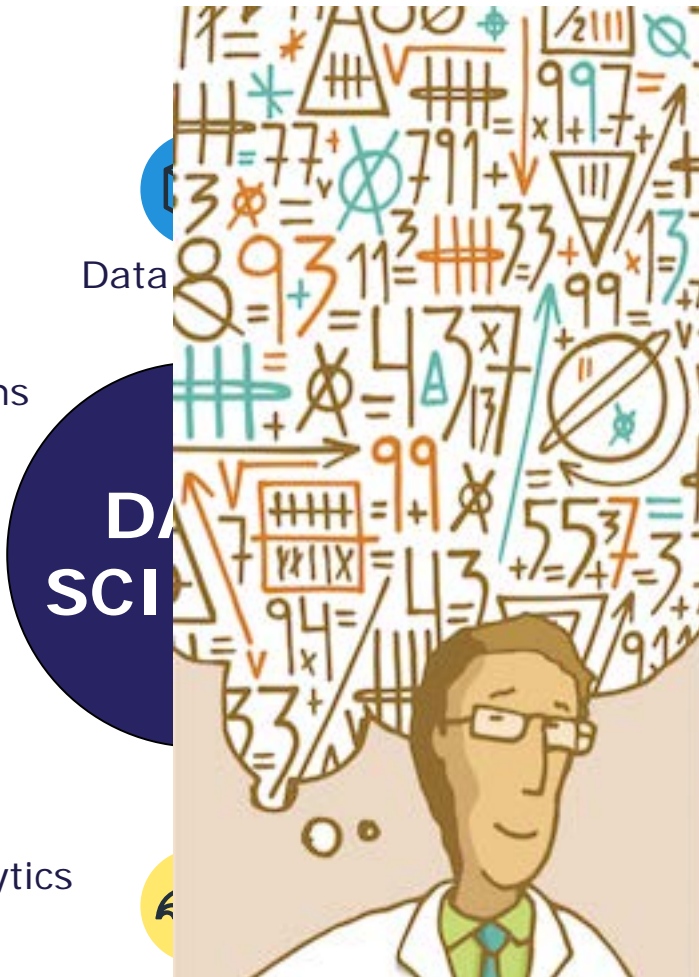
Données massives, données brutes



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

BY LOWE FOR THE SUN-SENTINEL, FLO

Un écosystème fragmenté



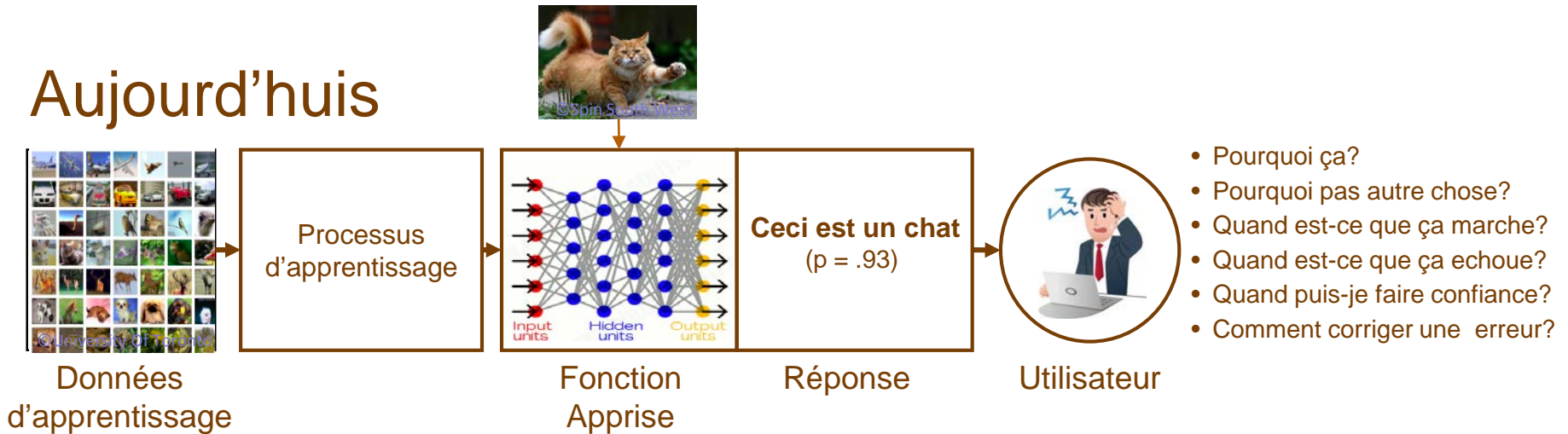
Quel est l'hyperplan qui sépare le mieux deux classes de points dans l'espace multidimensionnel?

↔
Fossé Digital



Comment puis-je adapter au mieux le bon médicament au bon dosage au bon patient au bon moment?

Aujourd'hui



Explainable AI – Dans quel but?

Title: Universal adversarial perturbations
Authors: [Moosavi-Dezfooli](#), [Seyed-Mohsen](#); [Fawzi, Alhussein](#); [Fawzi, Omar](#); [Frossard, Pascal](#)
Publication: eprint arXiv:1610.08401
Publication Date: 10/2016



Ceci n'est pas une chaussette

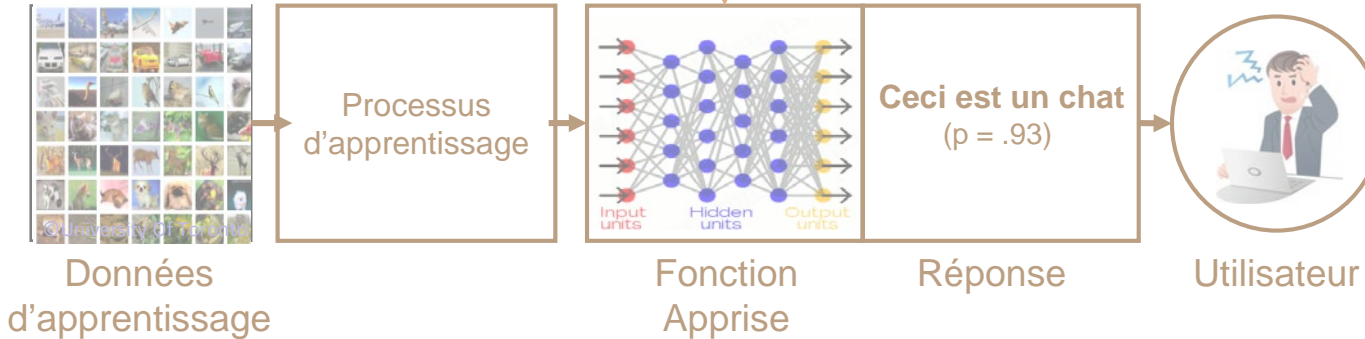


45mph sign?!

(<http://arxiv.org/abs/1707.08945>)

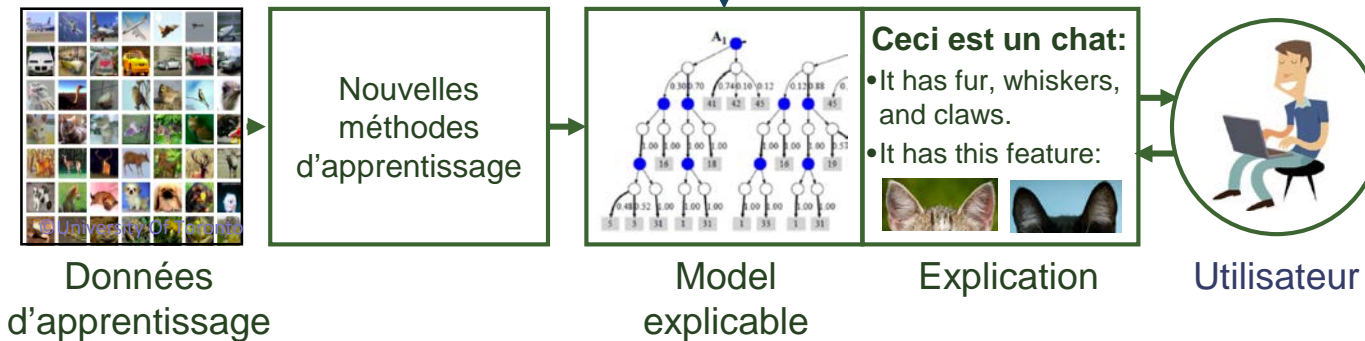
- C'est un éléphant d'Asie!
- En tout cas après avoir ajouté un bruit universel à l'image
- les méthodes d'apprentissage profonds ne mimiquent pas l'activité cérébrale humaine

Aujourd'hui



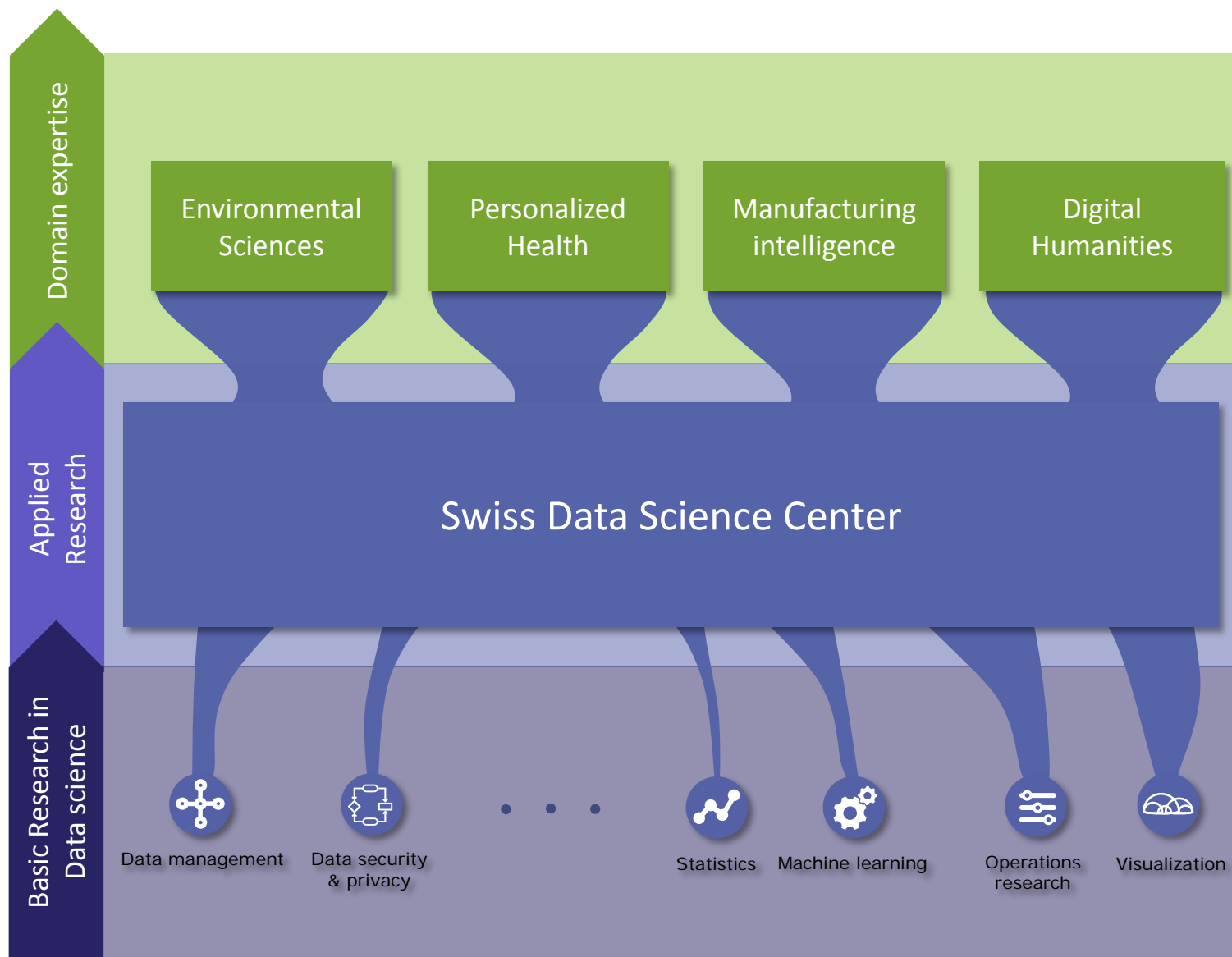
- Pourquoi ça?
- Pourquoi pas autre chose?
- Quand est-ce que ça marche?
- Quand est-ce que ça échoue?
- Quand puis-je faire confiance?
- Comment corriger une erreur?

Demain



- Je sais pourquoi
- Je sais pourquoi pas
- Je connais la robustesse
- Je connais les faiblesses
- Je fais confiance
- Je sais améliorer le model

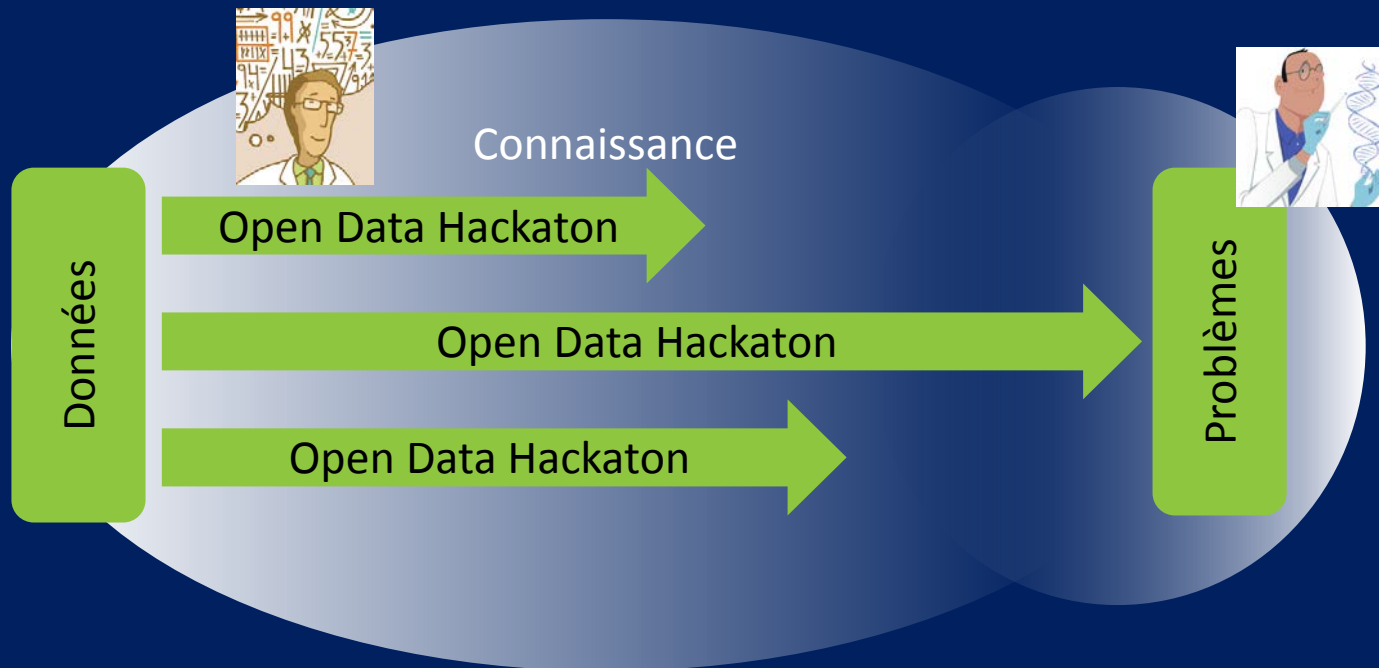
La mission du SDSC



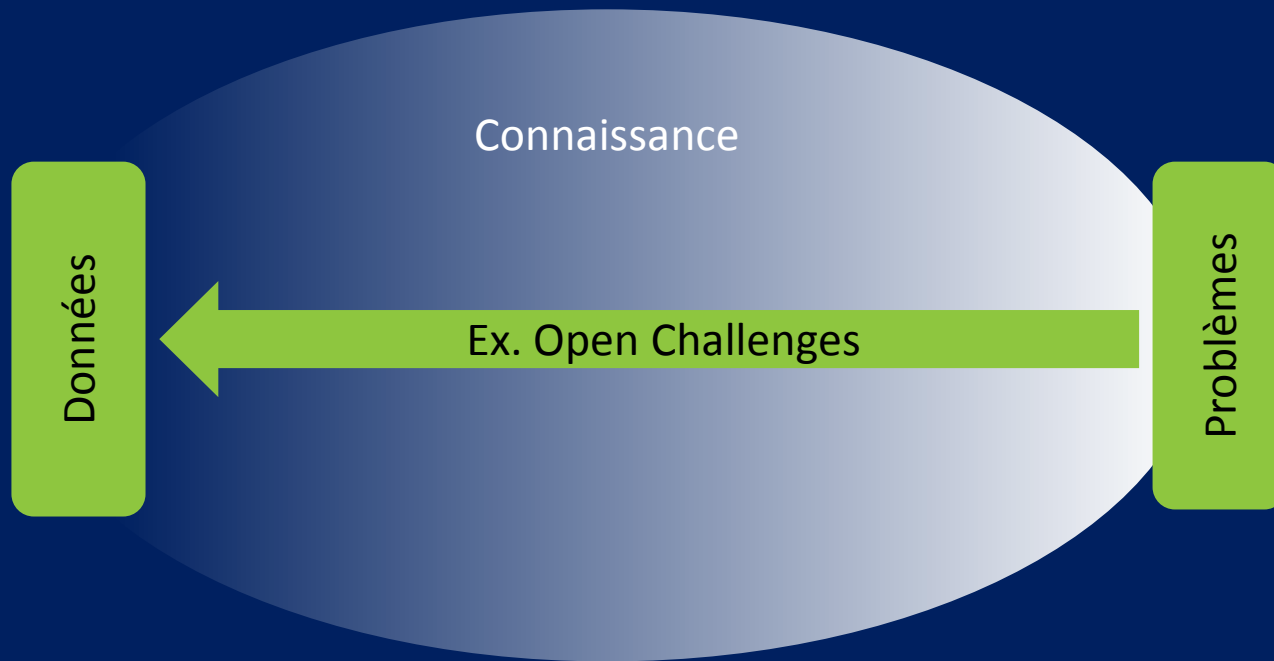
Les défis de la science des données



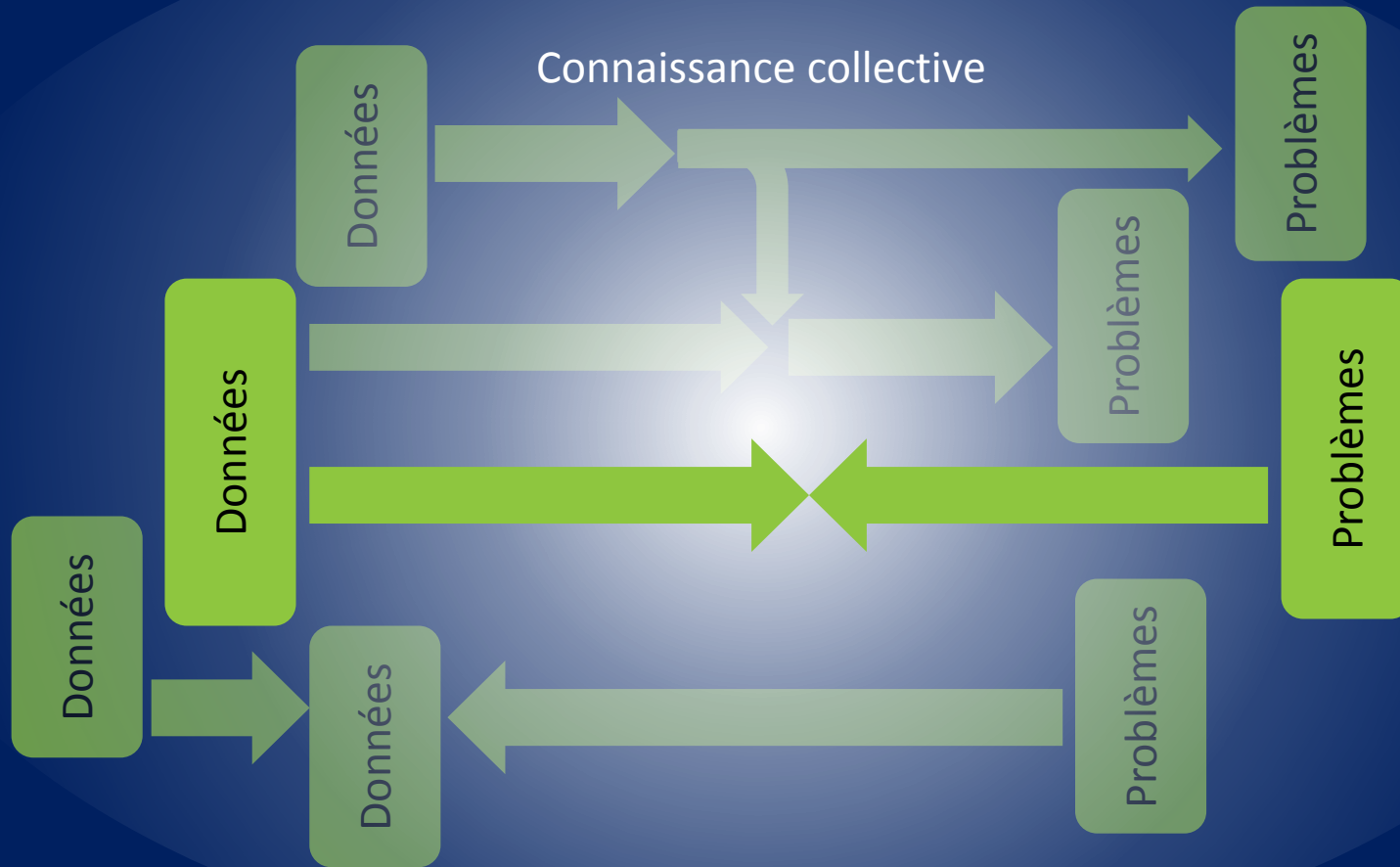
Les défis de la science des données



Les défis de la science des données



Les défis de la science des données



Les défis de la science des données

- Un reservoir de données, pas un marécage
 - Comment rendre ma donnée disponible et facilement trouvable?
 - Ma donnée est-elle protégée
 - Qui utilise ma donnée pour quoi faire
 - Quelles autres données sont disponibles? Ou?
 - Comment la donnée a-t-elle été créée? par qui?
 - Puis-je faire confiance à cette donnée?
 - Puis-je baser ma recherche sur cette donnée?
- La science des données rendue simple & de confiance
 - Combiner l'expertise humaine avec des méthodes d'apprentissage machine intelligentes
 - Une recherche transparente
 - Une recherche reproductible

... obstacles à la **collaboration** entre scientifiques, freine la **reproductibilité** et **réutilisation** des résultats de la recherche



RENGA - 連歌, POÉSIE EN COLLABORATION

RENGA (連歌)

Renga (連歌), est un genre de poésie japonaise collaborative dans laquelle deux ou plusieurs poètes doivent compléter des vers en alternance à un poème lié par des associations verbales et thématiques.

— Traduit de l'Encyclopædia Britannica

- **Renga** est une plateforme logicielle ouverte hautement évolutive et sécurisée, conçue pour favoriser la collaboration multidisciplinaire (scientifique) entre des institutions académiques et industrielles qui ne se font pas mutuellement confiance.
- La plateforme permet aux pratiquants de:
 - Gérez, partagez et traiter en toute sécurité des données à grande échelle à travers des institutions opérant dans un environnement fédéré.
 - Capture automatiquement la lignée complète jusqu'aux données brutes d'origine pour une traçabilité détaillée (auditabilité et reproductibilité)
 - Adhère aux principes "FAIR" et permet une recherche reproductible.

Principes FAIR

- Findable
 - données et meta-données doivent être facilement trouvable par les humains et les systèmes
- Accessible
 - données et meta-données doivent être préservés sur du long-terme, et doivent être facilement accessibles et téléchargeables avec des protocoles de communication standards, (en respectant les droits d'accès).
- Interoperable
 - données sont prêtes à être échangées, interprétées et combinées de façon (semi)automating avec d'autres données (et méthodes de traitement de données)
- Reusable
 - Les données et les métadonnées sont bien décrites et peuvent être réutilisées dans de futures recherches. Une citation correcte doit être facilitée, et les conditions dans lesquelles les données peuvent être utilisées doivent être claires pour les machines et les humains

Imposé par European Open Science Cloud (ESOC) pour l'hébergement et le traitement de la donnée liée à la recherche:

<https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>

Explications des principes FAIR:

http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

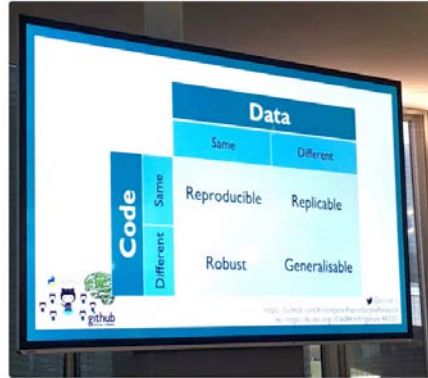
Drivers and barriers of open science:

<https://ec.europa.eu/research/openscience/index.cfm?pg=drivers§ion=monitor>

Reproductible, répliquable, robuste, généralisable

- https://twitter.com/lore_salvatori/status/912723417832595456

Reproductible / Replicable / Robust /
Generalisable @kirstie_j #OSET2017

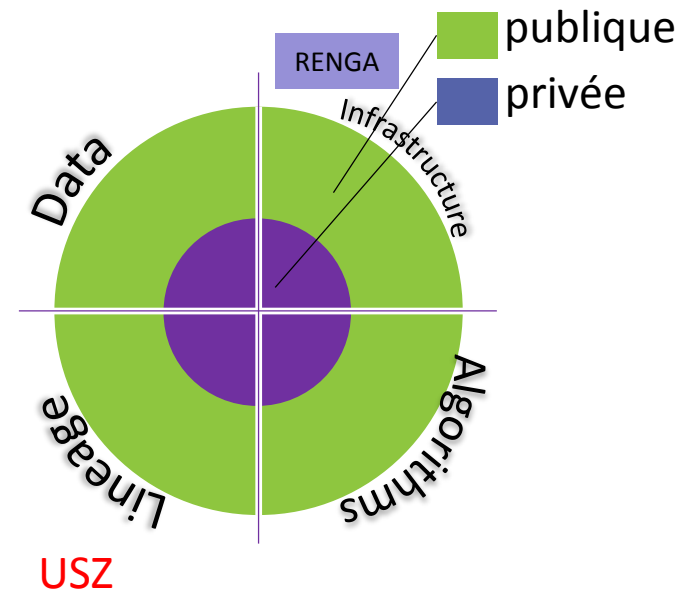
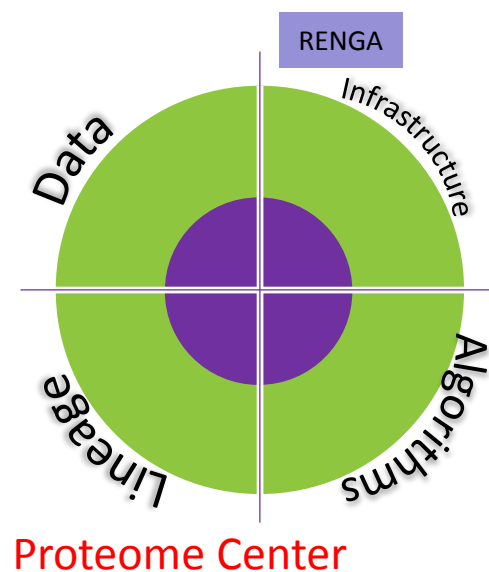
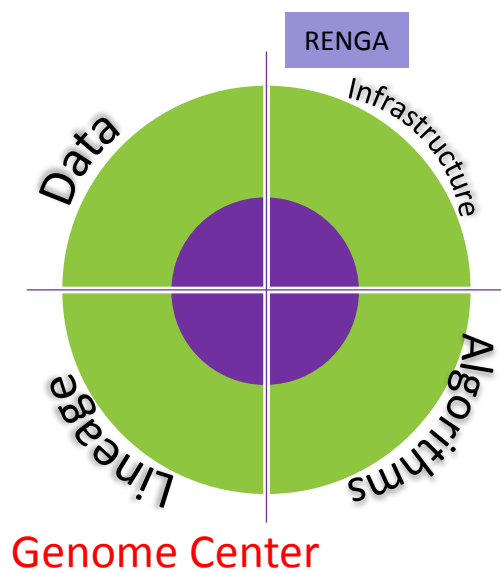


9:59 AM - 26 Sep 2017

		Donnée	
		Identique	Différente
Méthode	Identique	Reproductible	Répliquable
	Différent	Robuste	Généralisable

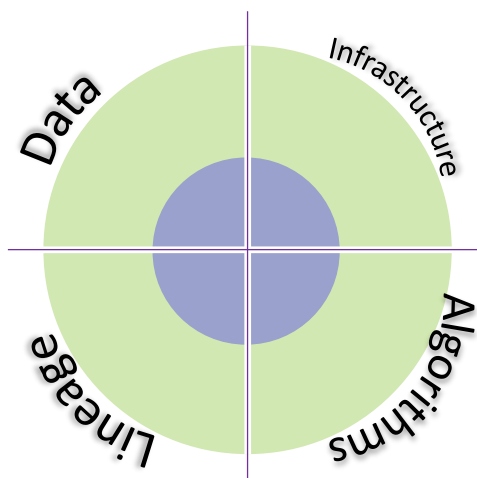
RENGA Cas d'usage

Fédératation



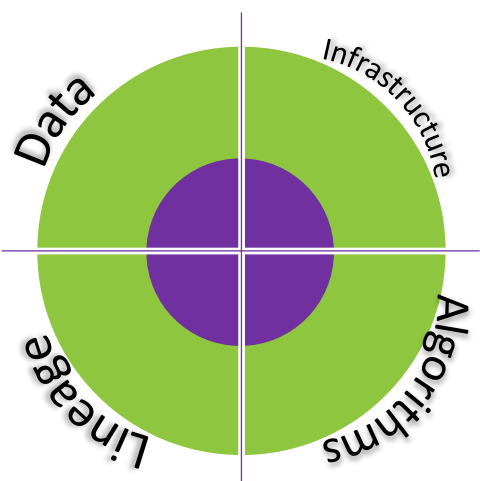
RENGA Cas d'usage

Vue unifiée

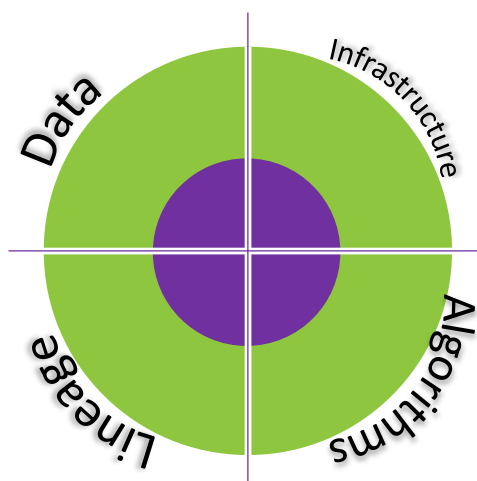


RENGA

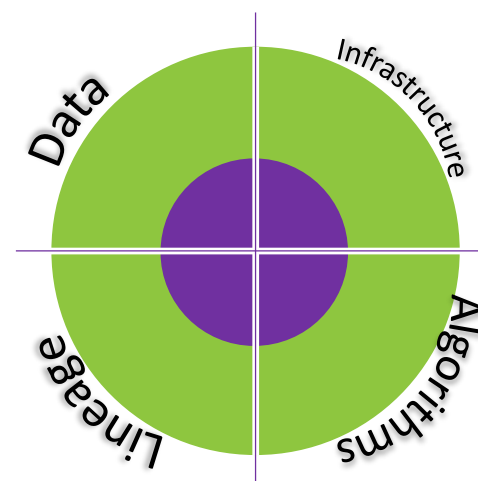
Fédération



Genome Center



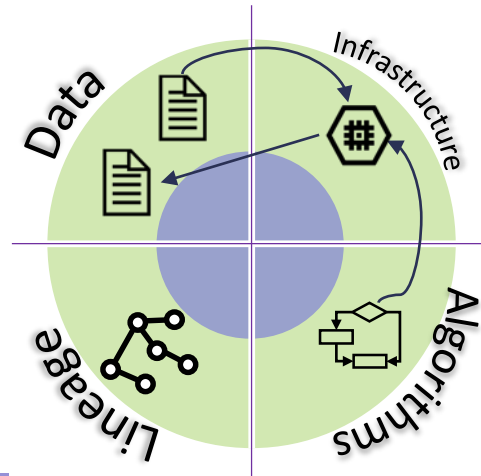
Proteome Center



USZ

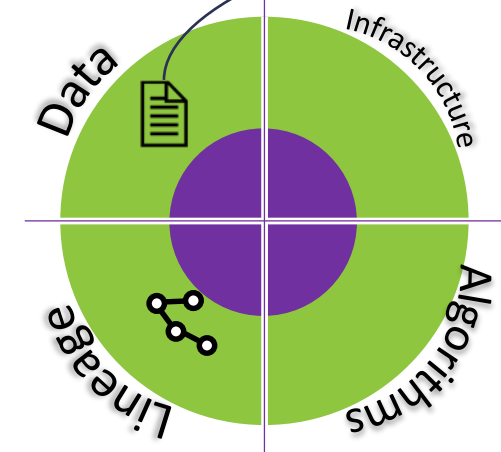
RENGA Cas d'usage

Vue unifiée

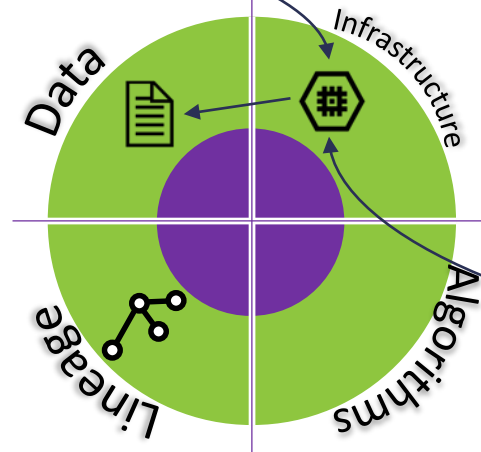


RENGA

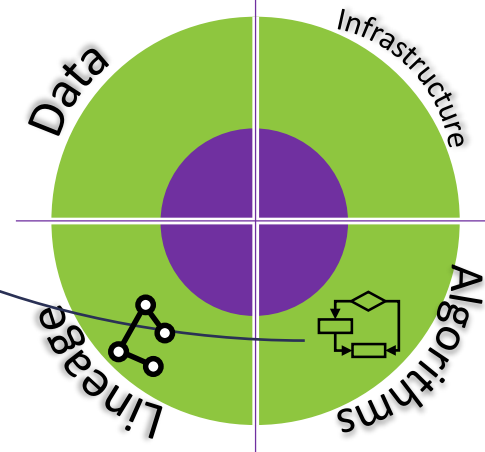
Fédération



Genome Center



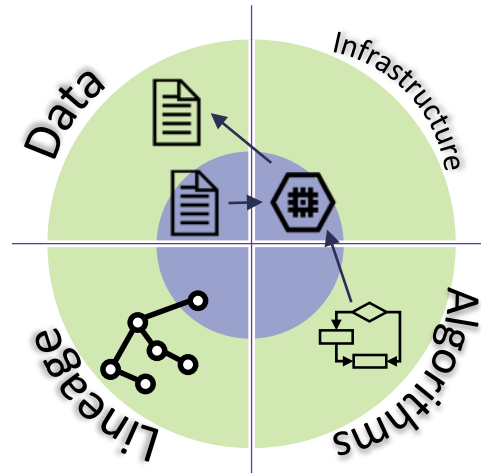
Proteome Center



USZ

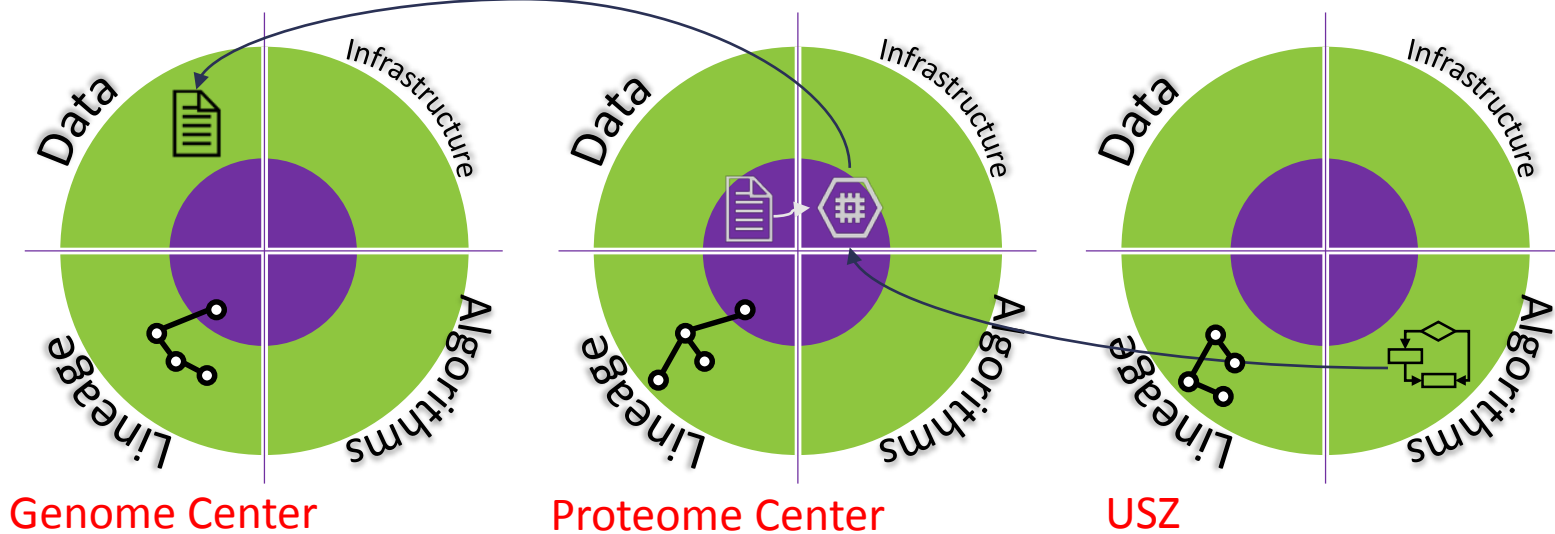
RENGA Cas d'usage

Vue unifiée



RENGA

Fédération



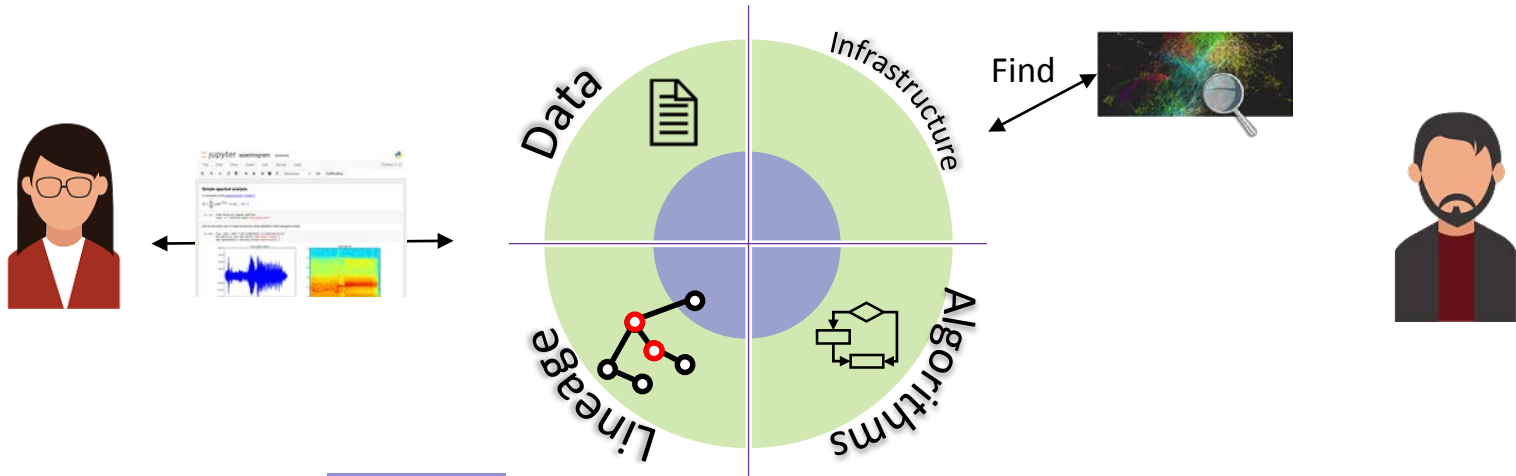
Genome Center

Proteome Center

USZ

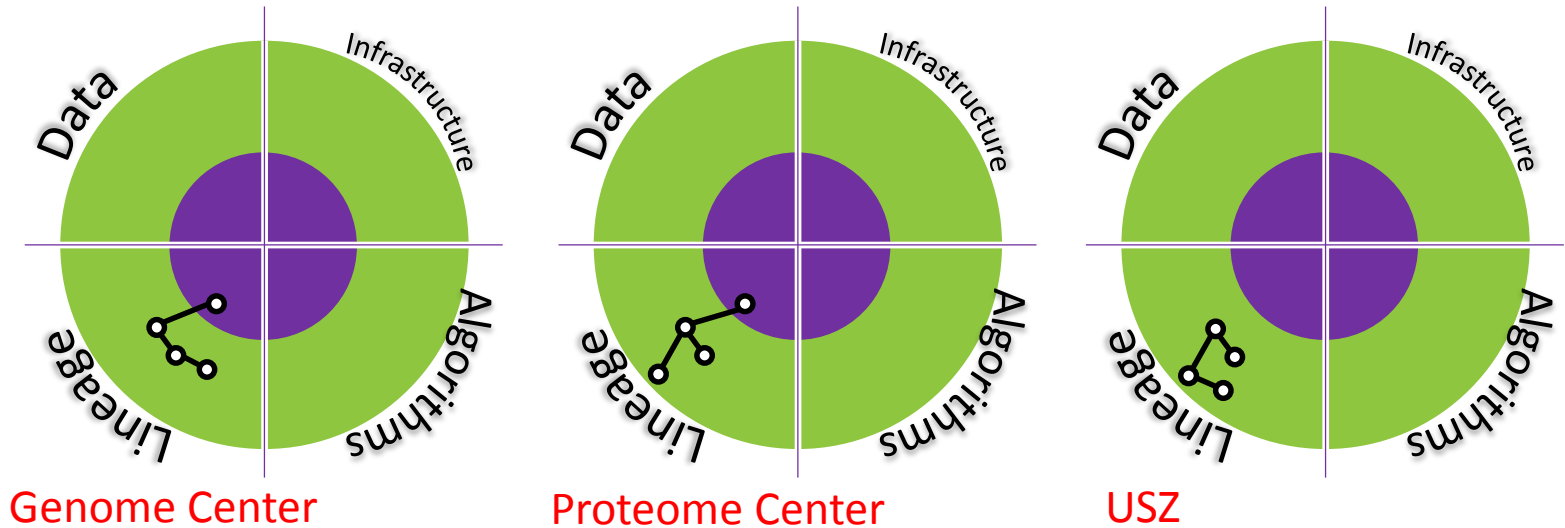
RENGA Cas d'usage

Vue unifiée



RENGA

Fédération



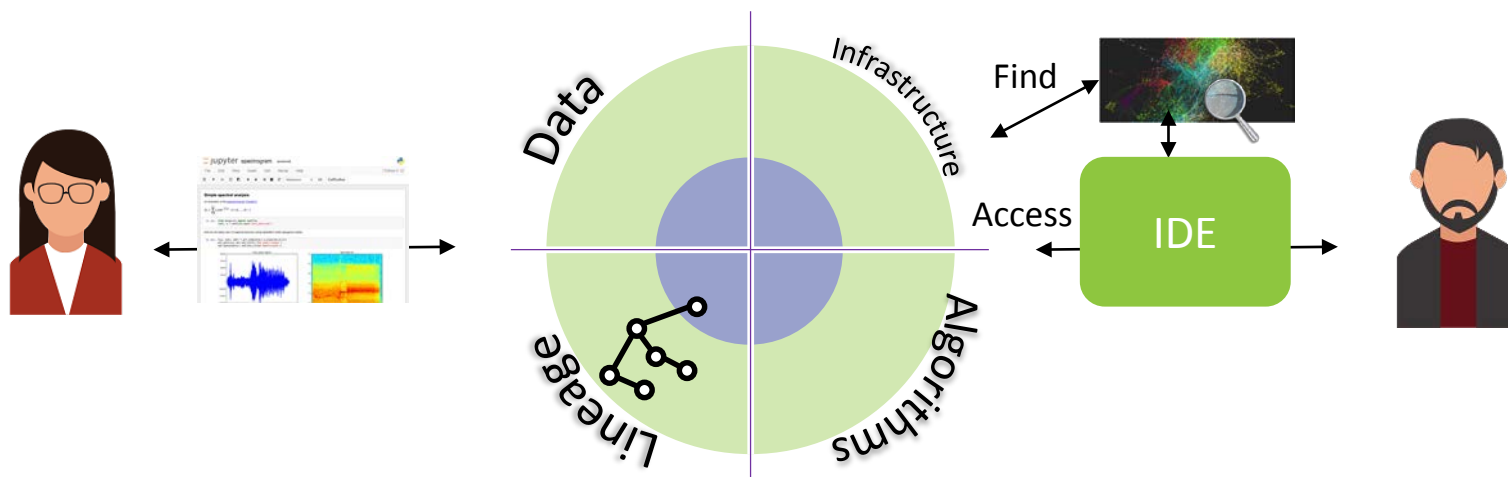
Genome Center

Proteome Center

USZ

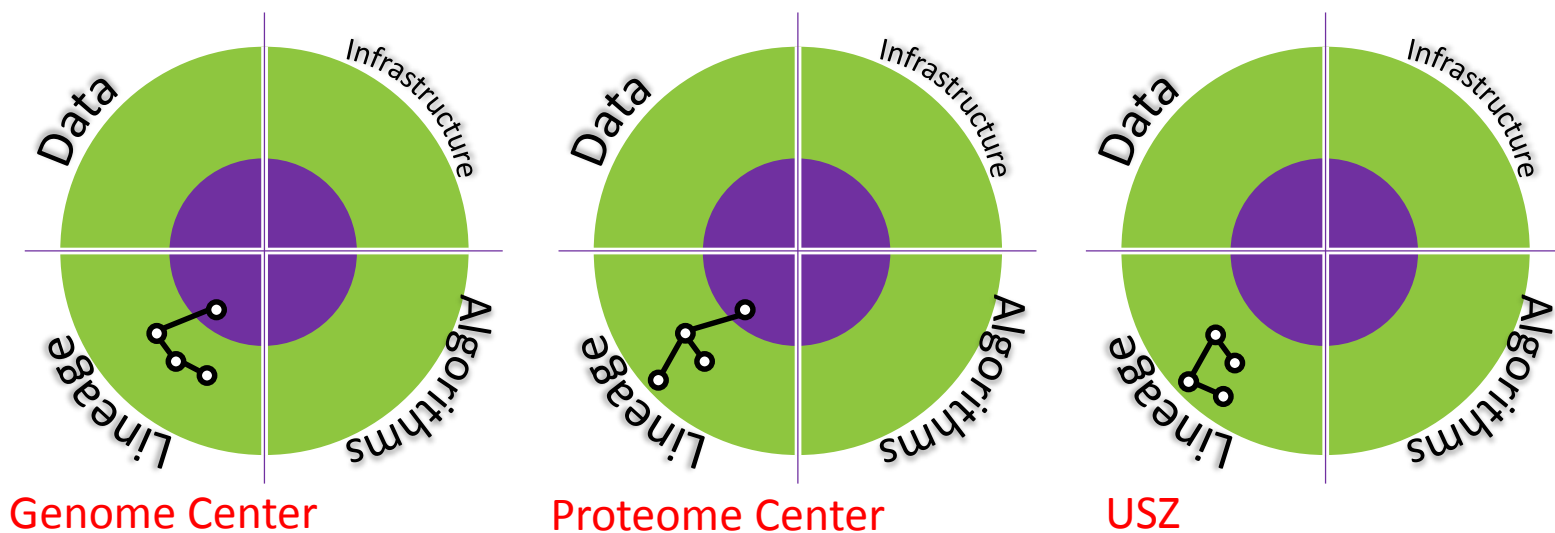
RENGA Cas d'usage

Vue unifiée



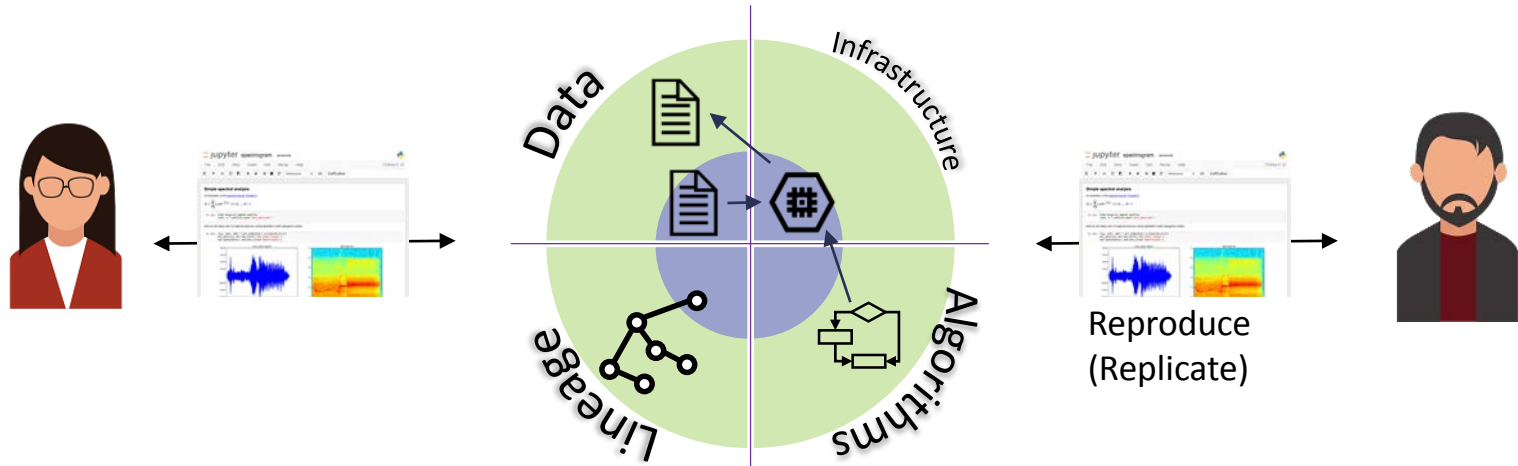
RENGA

Fédération

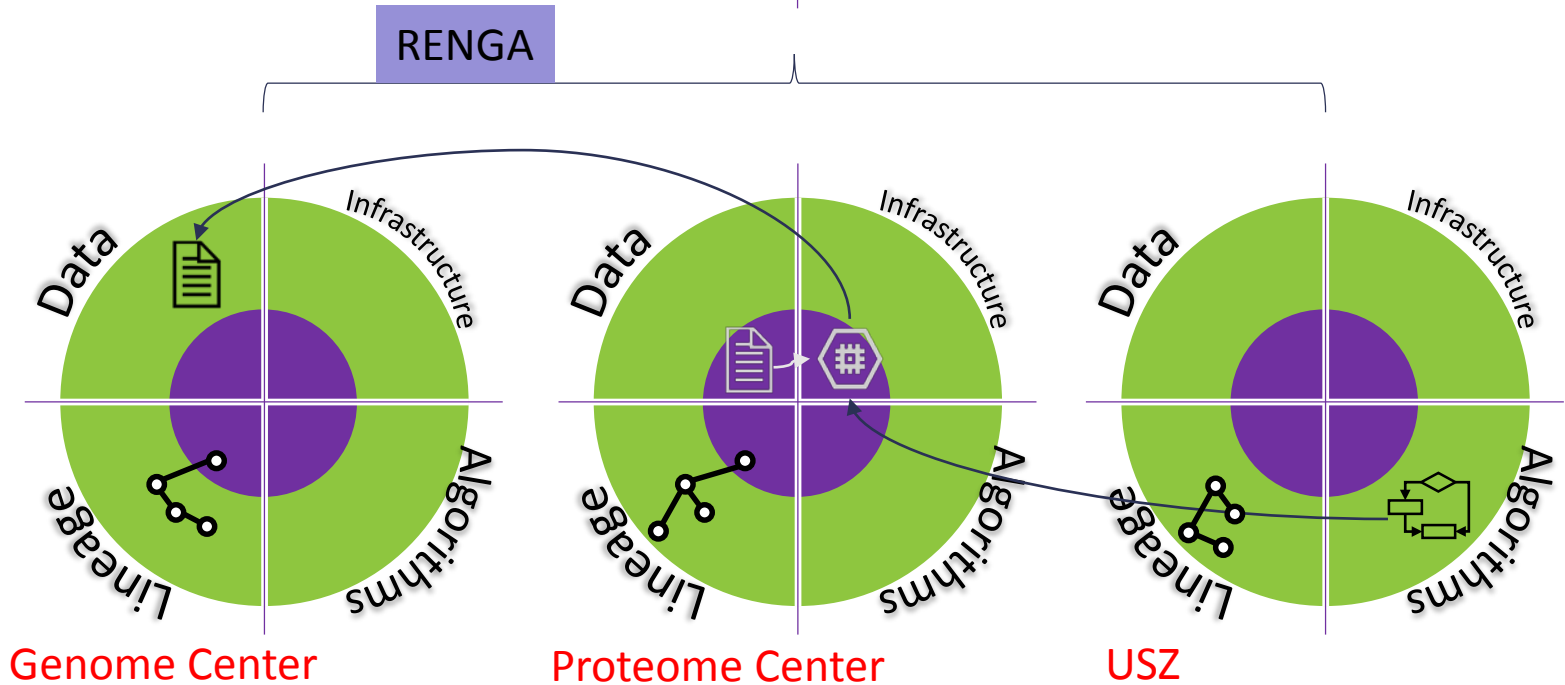


RENGA Cas d'usage

Vue unifiée

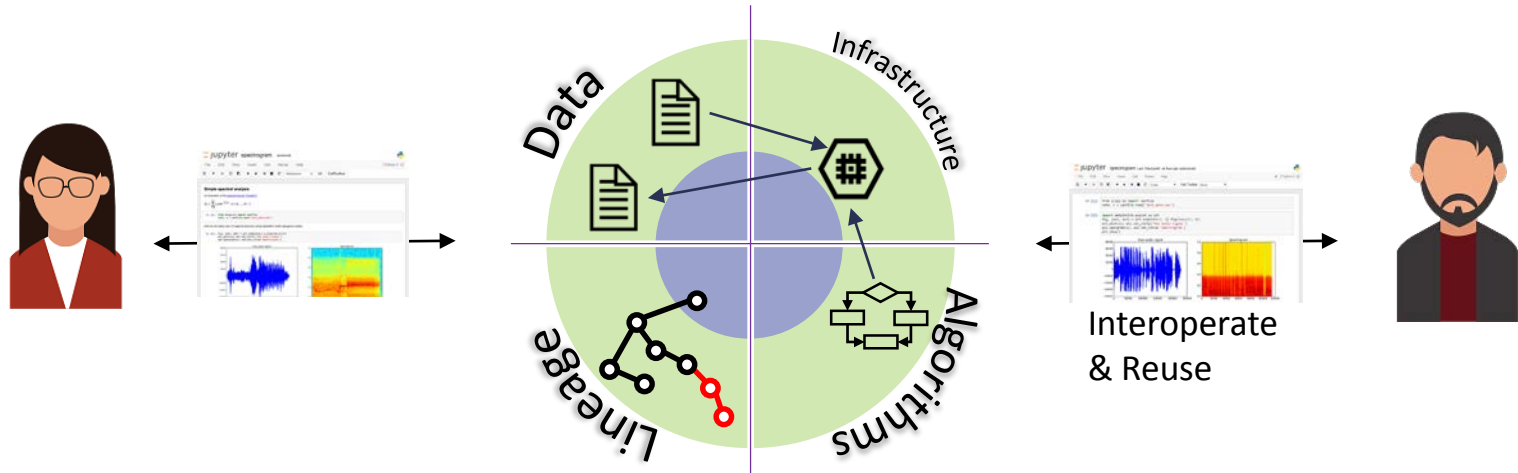


Fédération

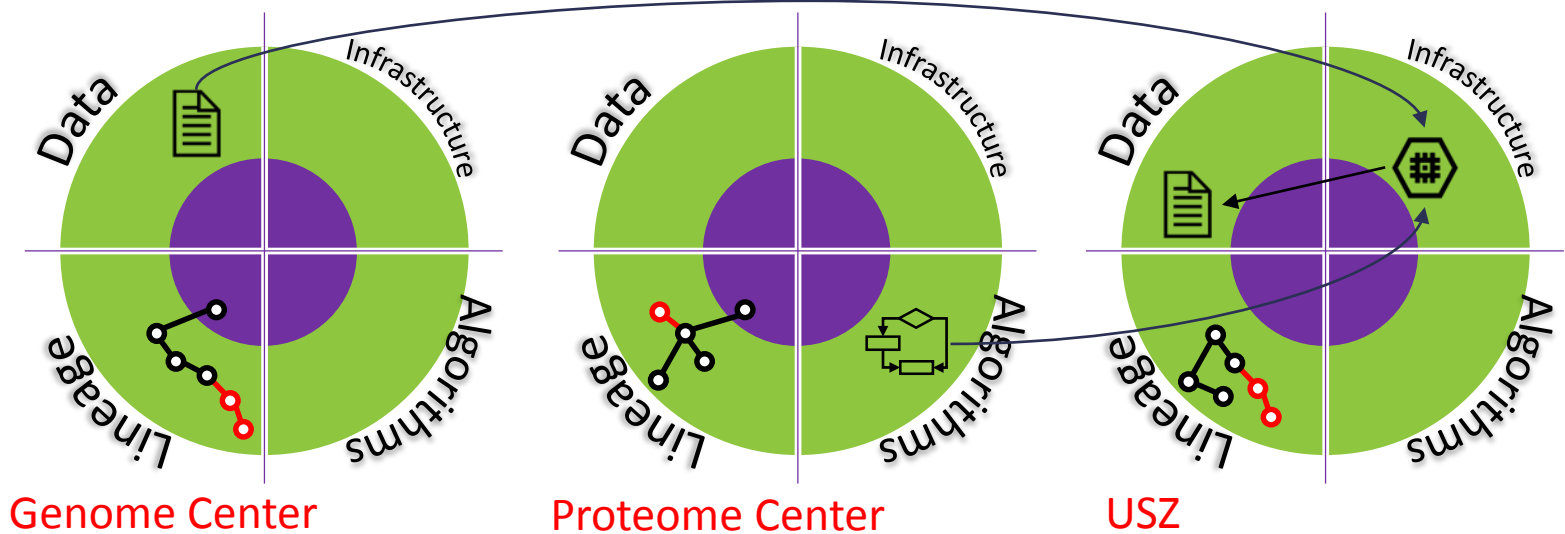


RENGA Cas d'usage

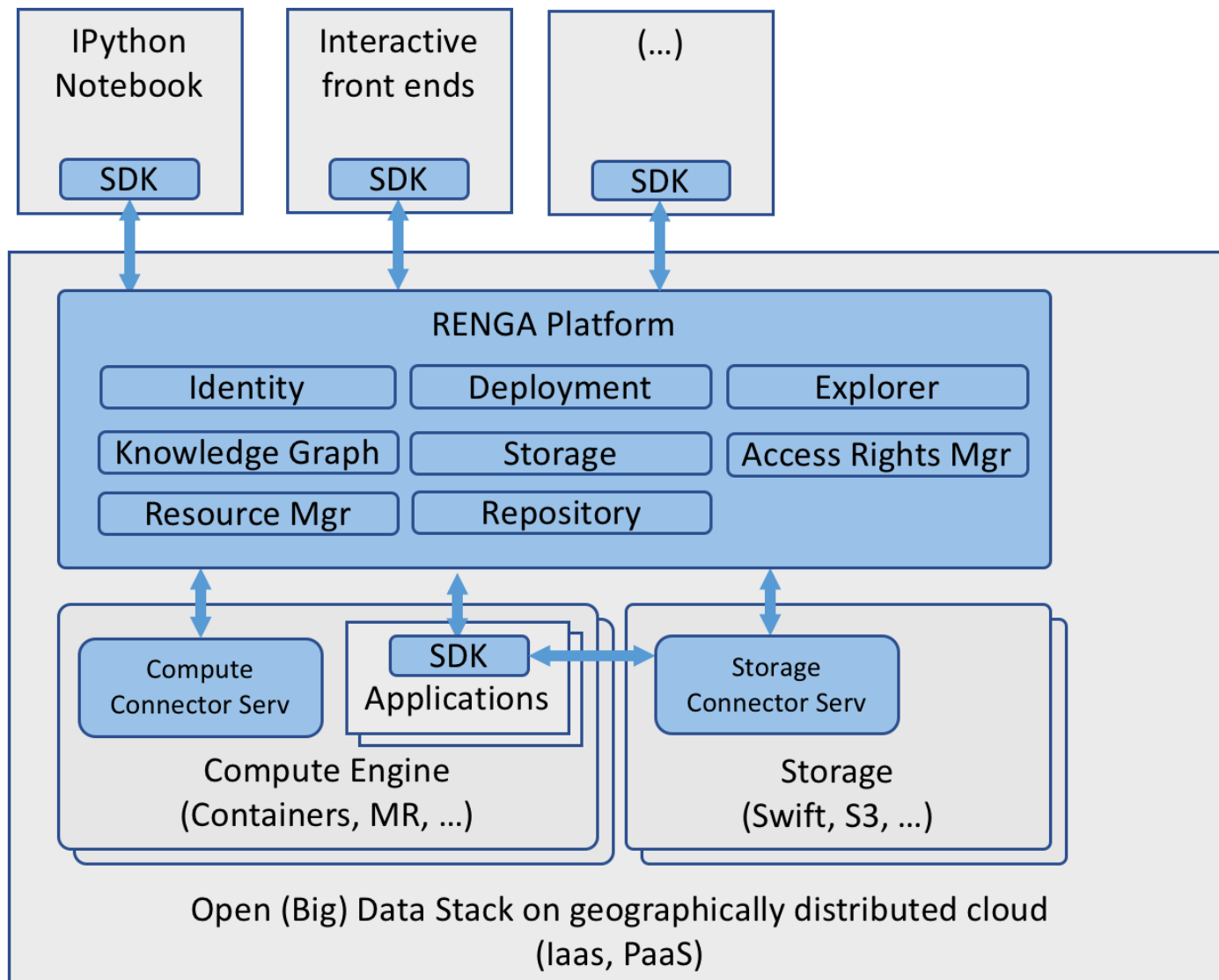
Vue unifiée



Fédération



RENGA – architecture



RENGA – architecture

Users, data scientists
SDSC and partners
IaaS

Domain specific notebook and analytics SDK

Environmental Science SDK

Transport SDK

Personalized Health SDK

Social science SDK

Micro services

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

play akka

JanusGraph

KEYCLOAK

RENGA Common API

Data Processing Applications

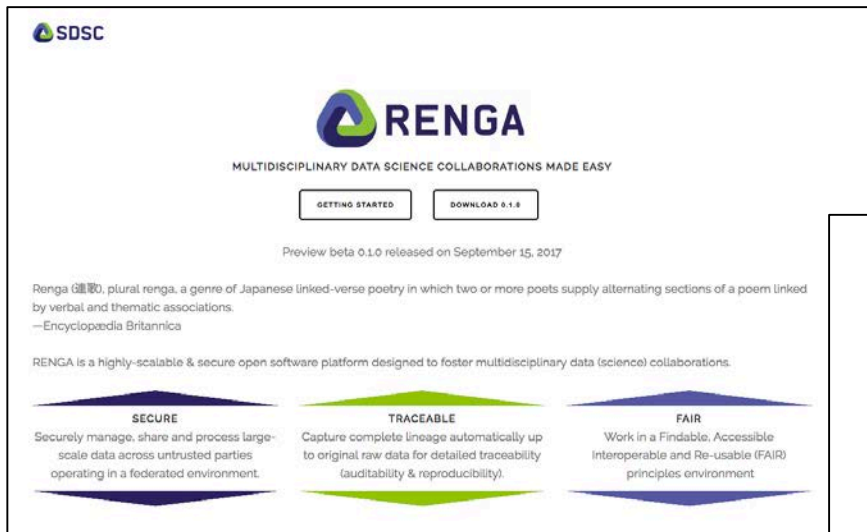
Backend (Open Big Data Platform Stack, Containers)

DC/OS APACHE Spark kafka APACHE HBASE Flink docker

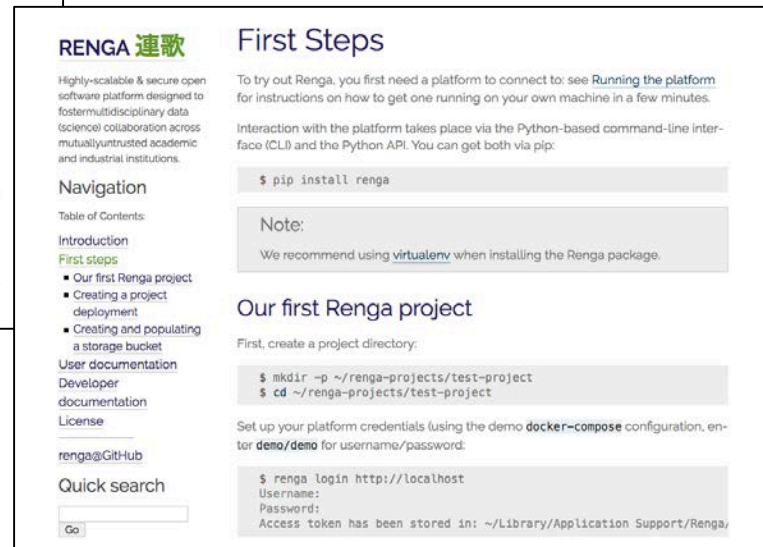
Geographically distributed cloud and on premise infrastructure(s) + long term storage solution providers for archiving

Object Store

GET RENGA (get-renga.io)



The screenshot shows the Renga website homepage. At the top left is the SDSC logo. In the center is the Renga logo, a stylized 'R' made of two overlapping triangles, followed by the word 'RENGA' in a bold, sans-serif font. Below the logo is the tagline 'MULTIDISCIPLINARY DATA SCIENCE COLLABORATIONS MADE EASY'. There are two buttons: 'GETTING STARTED' and 'DOWNLOAD 0.1.0'. Below these is a note: 'Preview beta 0.1.0 released on September 15, 2017'. A definition of Renga is provided: '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'. Below this is a description: 'RENGA is a highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaborations.' Three key features are highlighted in diamond-shaped boxes: '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).), and 'FAIR' (Work in a Findable, Accessible Interoperable and Re-usable (FAIR) principles environment).



The screenshot shows the 'First Steps' page of the Renga documentation. The title is 'RENGA 連歌 First Steps'. The text describes Renga as a 'Highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaboration across mutually untrusted academic and industrial institutions.' It explains that interaction with the platform is via a Python-based command-line interface (CLI) and the Python API. A code block shows the installation command: `$ pip install renga`. A 'Note' section states: 'We recommend using [virtualenv](#) when installing the Renga package.' The 'Our first Renga project' section instructs to create a project directory with the following commands: `$ mkdir -p ~/renga-projects/test-project` and `$ cd ~/renga-projects/test-project`. It then says to set up platform credentials using a demo `docker-compose` configuration, entering `demo/demo` for username/password. A final code block shows the login process: `$ renga login http://localhost`, followed by prompts for Username and Password, and a confirmation message: 'Access token has been stored in: ~/Library/Application Support/Renga/'. The page also includes a 'Navigation' section with links for 'Introduction', 'First steps', 'Our first Renga project', 'Creating a project deployment', and 'Creating and populating a storage bucket'. There are also links for 'User documentation', 'Developer documentation', 'License', and 'renga@GitHub'. A 'Quick search' section features a search input field and a 'Go' button.



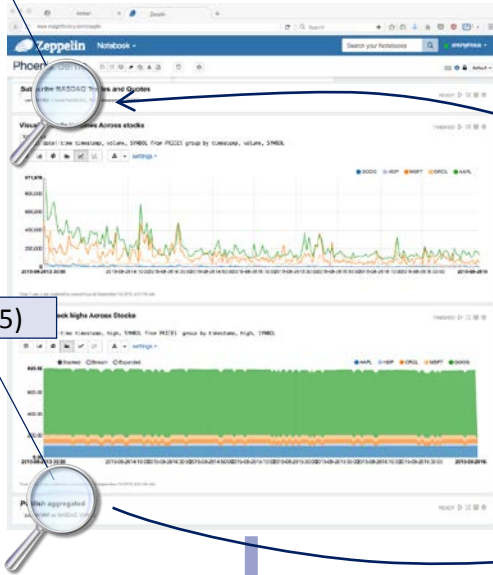
THANK YOU!

<http://www.datascience.ch>

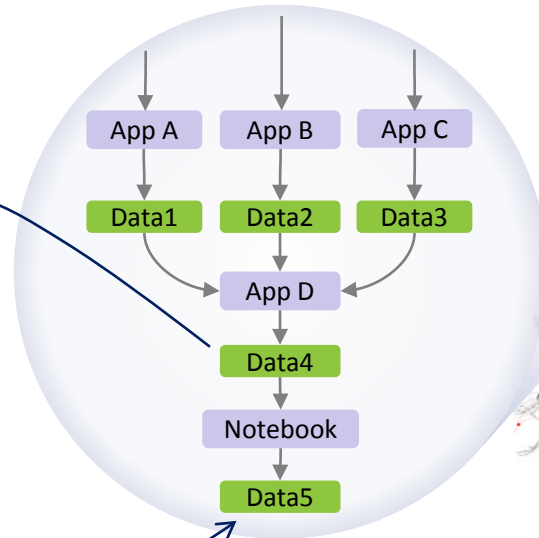
Twitter: @SDSCdatascience

Data Science Governance

subscribe(Data4)

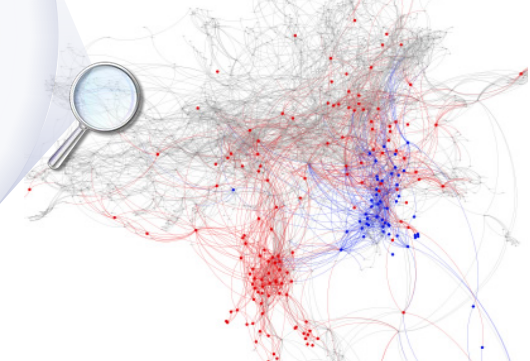


publish(Data5)



Find
Access
Interoperate
Reuse

Distributed Graph
in Federated Environment



Application
Repository



Data

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

