GakuNin RDMデータ活用セミナー ： これからの研究データ管理を探る

下山, 武司
国立情報学研究所オープンサイエンス基盤研究センター

藤原, 一毅
国立情報学研究所オープンサイエンス基盤研究センター

清水, 敏之
九州大学データ駆動イノベーション推進本部研究データ管理支援部門

https://doi.org/10.15017/7159632

出版情報：2023-11-16. National Institute of Informatics
バージョン：

NII RCOS
Research Center for Open Science and Data Platform

GakuNin RDM
データ解析機能の使い方

藤原一毅
国立情報学研究所オープンサイエンス基盤研究センター

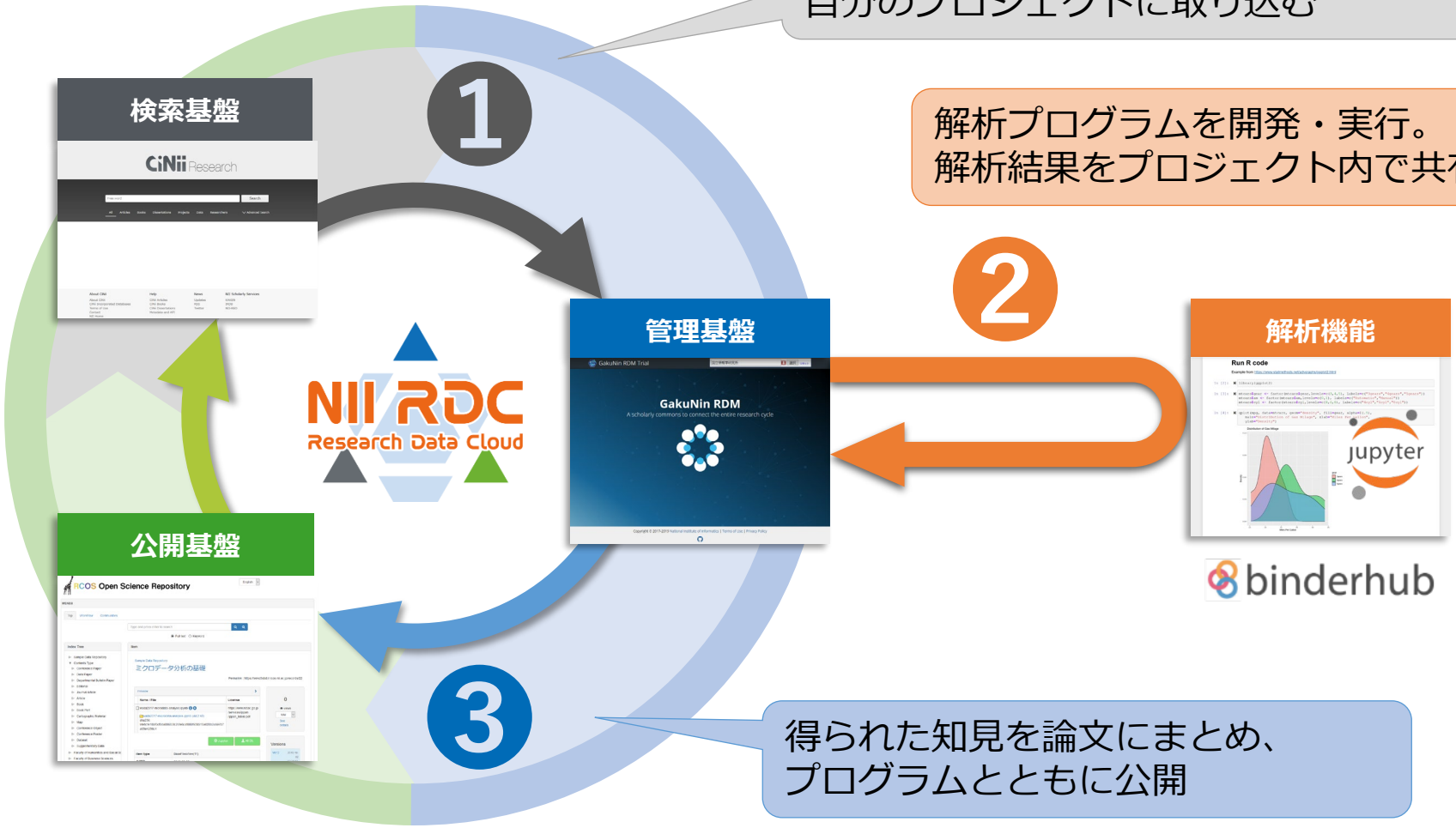
2023/11/16
九州大学 GakuNin RDM データ活用セミナー

データとコードが循環する世界

先行研究のデータを発見。GakuNin RDM の
自分のプロジェクトに取り込む

解析プログラムを開発・実行。
解析結果をプロジェクト内で共有

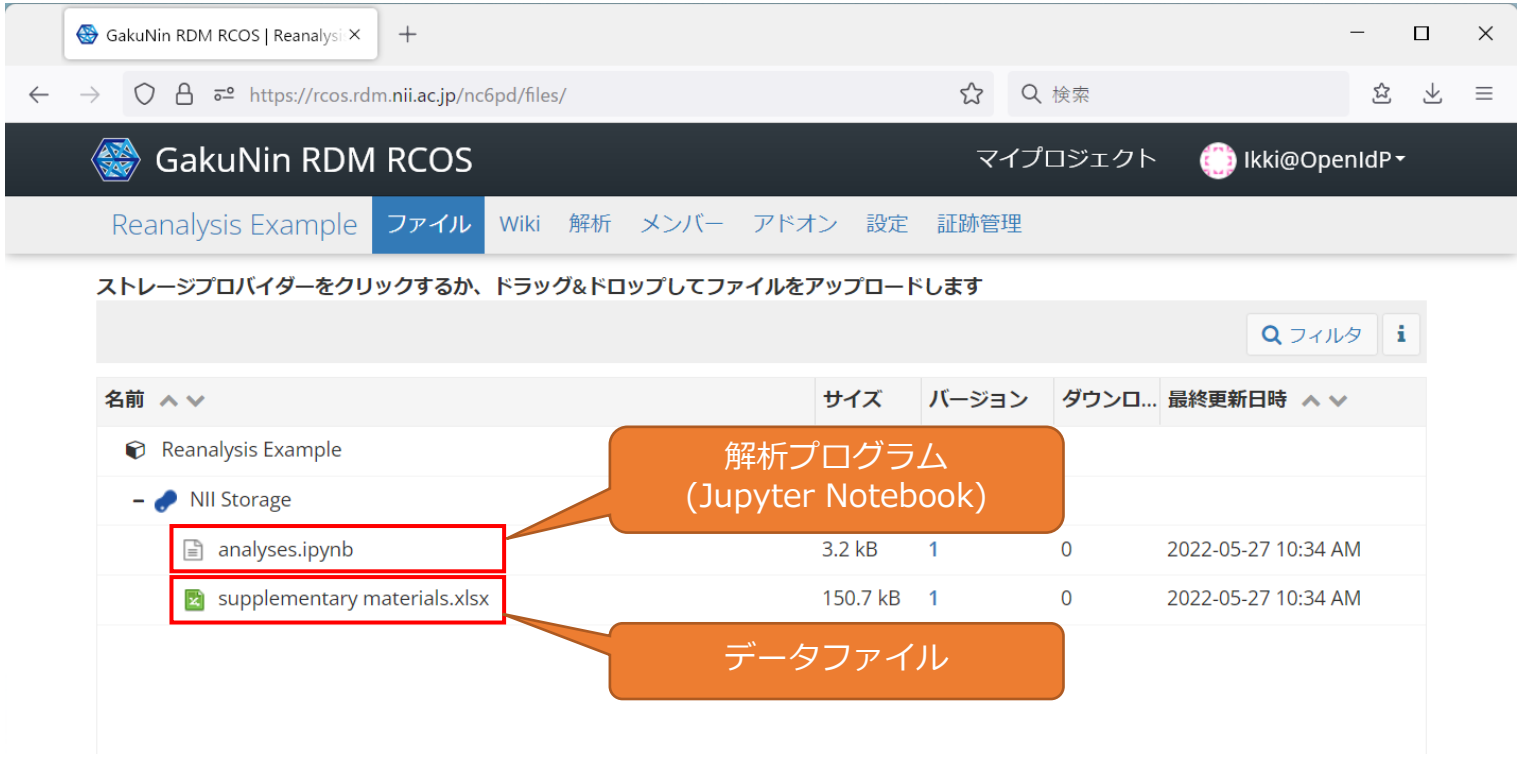
得られた知見を論文にまとめ、
プログラムとともに公開



GakuNin RDM データ解析機能

- JupyterHub がインストールされた計算機と連携し、データ解析環境をGakuNin RDMから1クリックで構築
- NII所有の計算機のほか、クラウド上のVMなど外部計算機とも連携可能

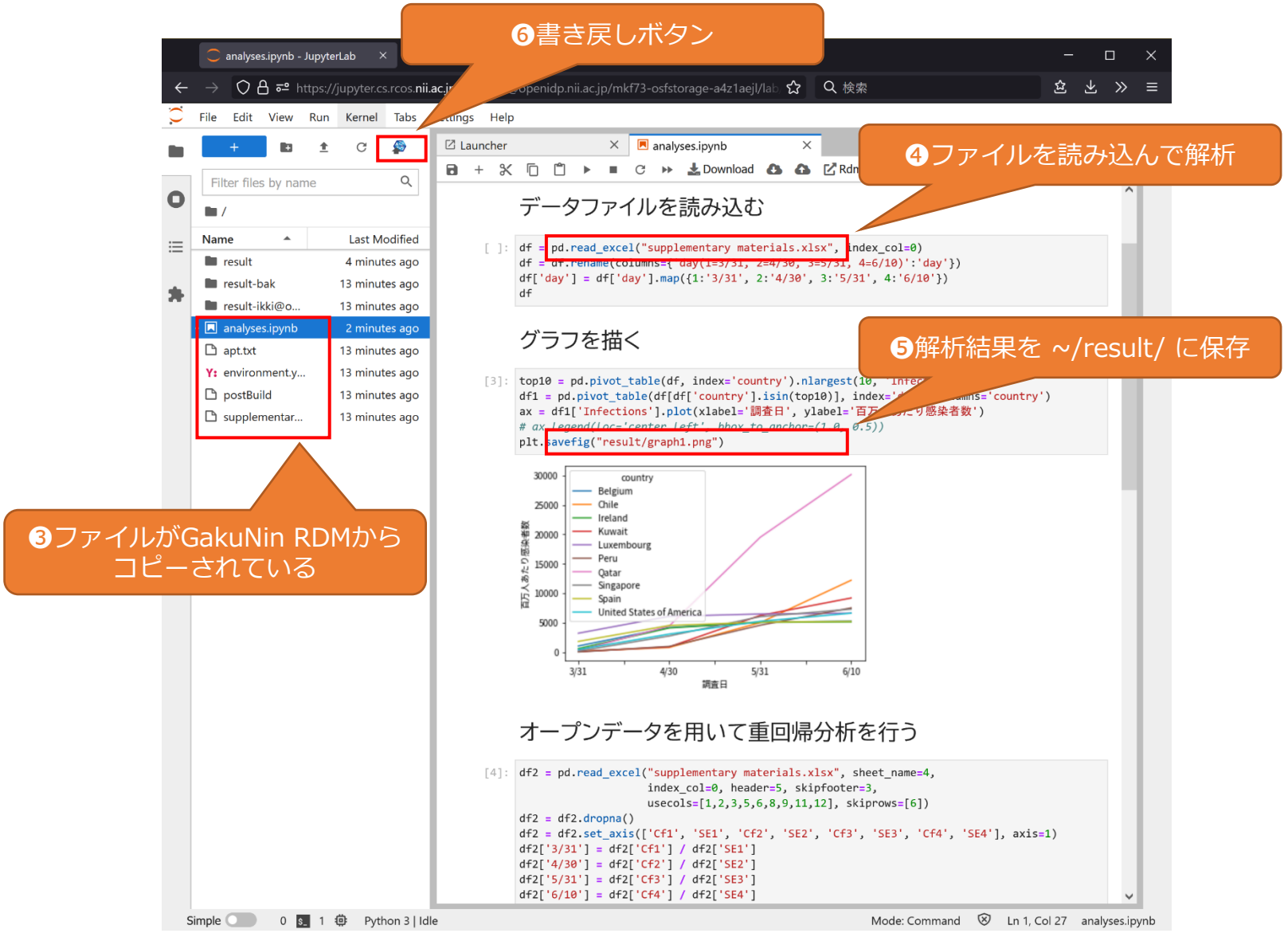
デモ (1/4)


GakuNin RDM RCOS | Reanalysi
https://rcos.rdm.nii.ac.jp/nc6pd/files/
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki@OpenIdP
Reanalysis Example
ファイル
Wiki
解析
メンバー
アドオン
設定
証跡管理
ストレージプロバイダーをクリックするか、ドラッグ&ドロップしてファイルをアップロードします
フィルタ
名前
サイズ
バージョン
ダウンロ...
最終更新日時
Reanalysis Example
NII Storage
analyses.ipynb
3.2 kB
1
0
2022-05-27 10:34 AM
supplementary materials.xlsx
150.7 kB
1
0
2022-05-27 10:34 AM
解析プログラム
(Jupyter Notebook)
データファイル

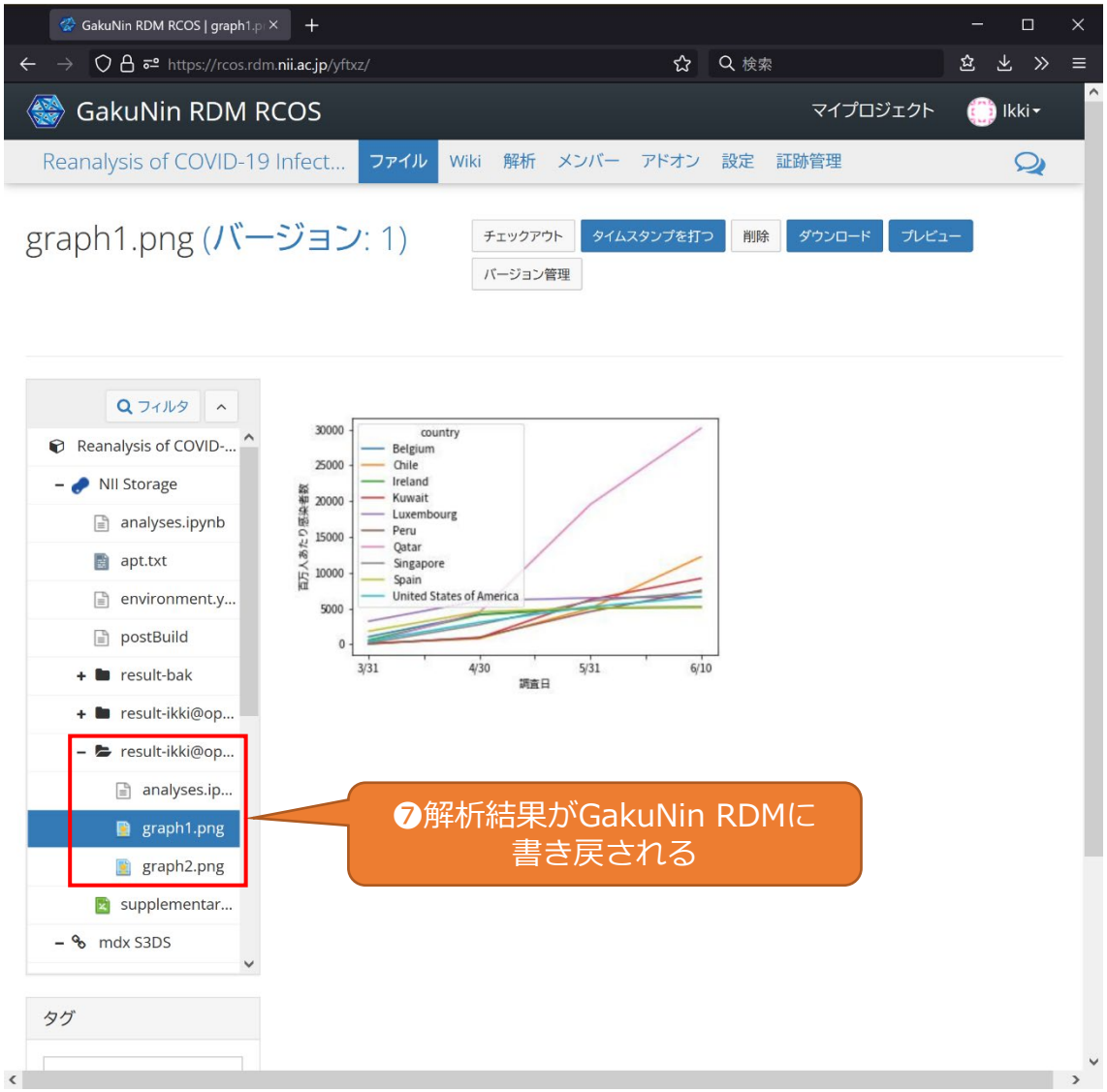
デモ (2/4)


Reanalysis Example BidnerHub
https://rcos.rdm.nii.ac.jp/nc6pd/binderhub
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki@OpenIdP
Reanalysis Example
ファイル
Wiki
解析
メンバー
アドオン
設定
新しい解析環境
❶解析環境を構成
基本イメージ
Python 3.9 + R 4.1.3
Jupyter Notebook, JupyterLab, RStudio, Shinyが使えます。
変更
追加パッケージ
apt-get
fonts-noto-cjk:
+追加
conda
seaborn:
openpyxl:
+追加
pip
+追加
R (MRAN)
+追加
自動実行スクリプト
#!/bin/bash
set -x
...
保存
❷計算機を選択して起動！
環境作成
このプロジェクトのデフォルトストレージの内容がコピーされます。
新しい解析環境を作成: https://binder.cs.rcos.nii.ac.jp

デモ (3/4)


❻書き戻しボタン
❹ファイルを読み込んで解析
❺解析結果を ~/result/ に保存
❸ファイルがGakuNin RDMからコピーされている
analyses.ipynb - JupyterLab
File Edit View Run Kernel Tabs Settings Help
Launcher
analyses.ipynb
Download
Filter files by name
Name
Last Modified
result 4 minutes ago
result-bak 13 minutes ago
result-ikki@o... 13 minutes ago
analyses.ipynb 2 minutes ago
apt.txt 13 minutes ago
environment.y... 13 minutes ago
postBuild 13 minutes ago
supplementar... 13 minutes ago
データファイルを読み込む
df = pd.read_excel("supplementary materials.xlsx", index_col=0)
df['day'] = df['day'].map({1:'3/31', 2:'4/30', 3:'5/31', 4:'6/10'})
df
グラフを描く
plt.savefig("result/graph1.png")
country
Belgium
Chile
Ireland
Kuwait
Luxembourg
Peru
Qatar
Singapore
Spain
United States of America
3/31
4/30
5/31
6/10
オープンデータを用いて重回帰分析を行う
df2 = pd.read_excel("supplementary materials.xlsx", sheet_name=4,
index_col=0, header=5, skipfooter=3,
usecols=[1,2,3,5,6,8,9,11,12], skiprows=[6])
df2 = df2.dropna()
df2 = df2.set_axis(['Cf1', 'SE1', 'Cf2', 'SE2', 'Cf3', 'SE3', 'Cf4', 'SE4'], axis=1)
df2['3/31'] = df2['Cf1'] / df2['SE1']
df2['4/30'] = df2['Cf2'] / df2['SE2']
df2['5/31'] = df2['Cf3'] / df2['SE3']
df2['6/10'] = df2['Cf4'] / df2['SE4']
Simple
Python 3 | Idle
Mode: Command
Ln 1, Col 27
analyses.ipynb

デモ (4/4)


GakuNin RDM RCOS | graph1.p
https://rcos.rdm.nii.ac.jp/yftxz/
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki
Reanalysis of COVID-19 Infect...
ファイル
Wiki
解析
メンバー
アドオン
設定
証跡管理
graph1.png (バージョン: 1)
チェックアウト
タイムスタンプを打つ
削除
ダウンロード
プレビュー
バージョン管理
フィルタ
Reanalysis of COVID-...
NII Storage
analyses.ipynb
apt.txt
environment.y...
postBuild
result-bak
result-ikki@op...
result-ikki@op...
analyses.ip...
graph1.png
graph2.png
supplementar...
mdx S3DS
タグ
country
Belgium
Chile
Ireland
Kuwait
Luxembourg
Peru
Qatar
Singapore
Spain
United States of America
30000
25000
20000
15000
10000
5000
0
3/31
4/30
5/31
6/10
調査日
❼解析結果がGakuNin RDMに
書き戻される

こんな用途に使えます

研究

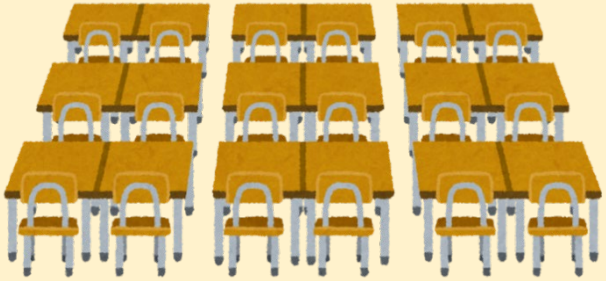
- ご自身の研究のためのデータ分析

公開・共有

- 他の研究者の二次分析に資するデータとプログラムの公開

教育・学習

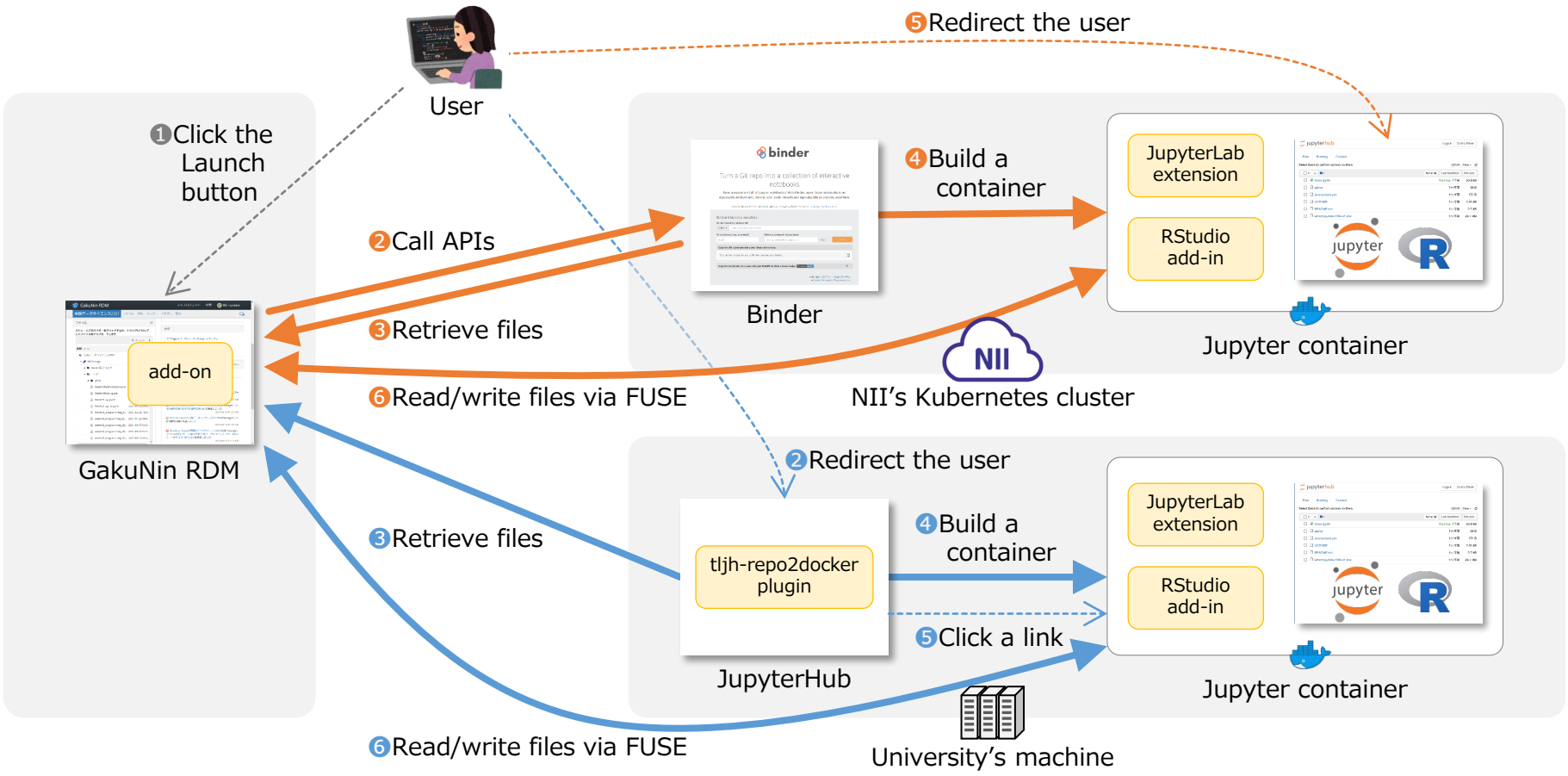
- 学生たちにデータ分析をさせるゼミ・講義・演習など



引き継ぎ

- 先輩の研究環境を後輩が再現し、研究を継続する

システム構成


User
❺Redirect the user
❶Click the Launch button
❷Call APIs
❸Retrieve files
❻Read/write files via FUSE
add-on
GakuNin RDM
binder
Binder
❹Build a container
JupyterLab extension
RStudio add-in
jupyter
Jupyter container
NII
NII's Kubernetes cluster
❷Redirect the user
❸Retrieve files
tljh-repo2docker plugin
JupyterHub
❹Build a container
❺Click a link
❻Read/write files via FUSE
University's machine

類似サービスとの比較

	GakuNin RDM データ解析機能	GESIS Notebooks	mybinder.org	Google Colab	Microsoft Codalab
対象分野	汎用	社会科学	汎用	主に深層学習	深層学習
提供元	NII （日・学術機関）	GESIS （独・学術機関）	Project Jupyter （任意団体）	Google （米・民間企業）	Microsoft （米・民間企業）
アカウント	学認	GESIS	不要	Google	Codalab
対応言語	Python, R, Shiny, MATLAB（準備中）	Python, R, Julia	Python, R, Julia	Python, R, Julia, Swift	Python
対応リポジトリ	GakuNin RDM, JDCat (WEKO3), GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	Google Drive, GitHub	?
メモリ CPU ストレージ	3GB 36コア共有 10GB	32GB 2コア 10GB	2GB 1コア ?	13GB 2コア 40GB	?
タイムアウト	半永続 30日不使用で消去	40分	10分	90分 / 12時間	?
インフラ	オンプレ	オンプレ	Google, OVH, Turing Institute	Google	Microsoft

詳しい情報

誰が使えますか？

- 九州大学のユーザーは全員、GakuNin RDM データ解析機能をご利用いただけます。

どこを見れば分かりますか？

- マニュアル https://support.rdm.nii.ac.jp/
  → ユーザーマニュアル → データ解析機能
- 解説動画 https://youtu.be/_FzOpDTQrBQ

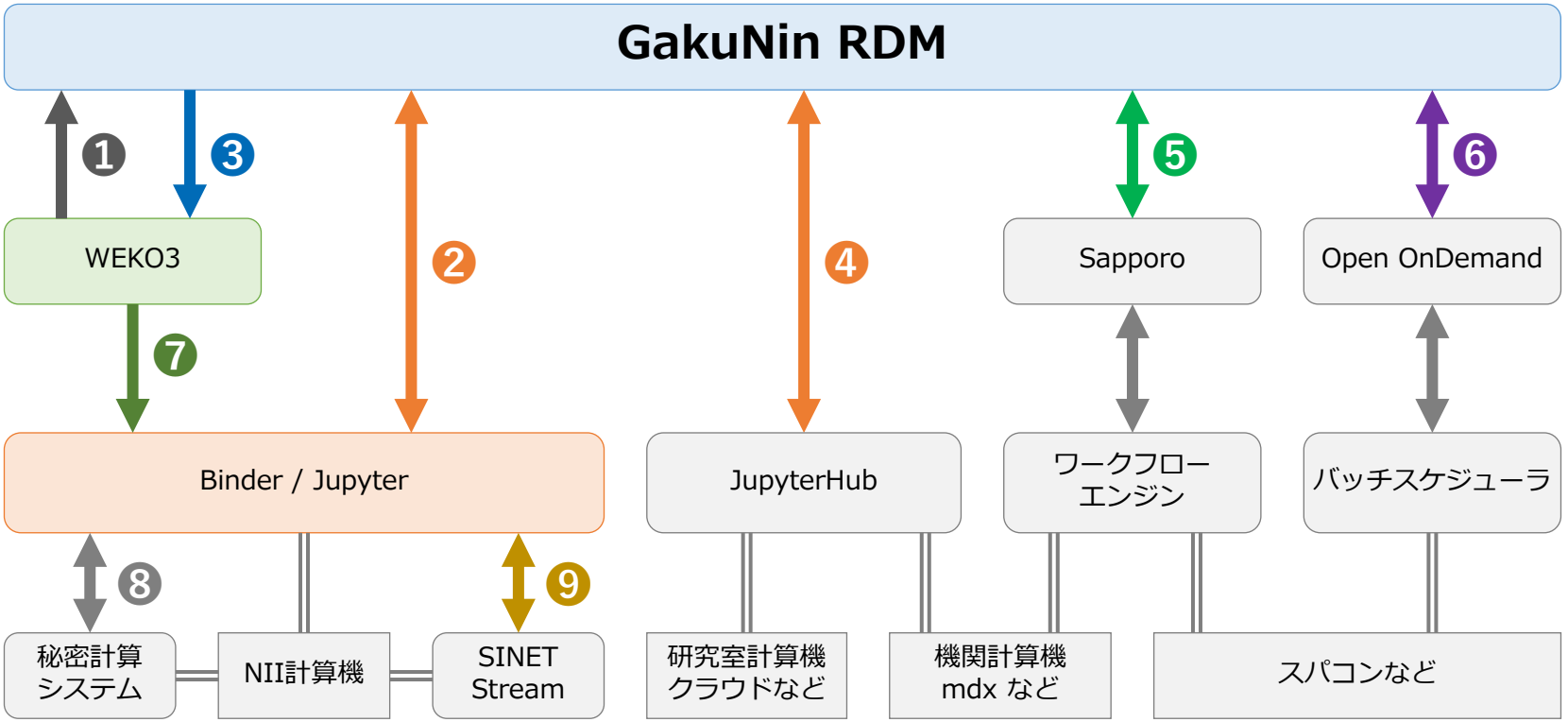
分からないときは誰に聞けばいいですか？

- 九州大学 DX推進本部 研究データ管理支援部門
  rds_help@dx.kyushu-u.ac.jp

外部計算機連携のバリエーション

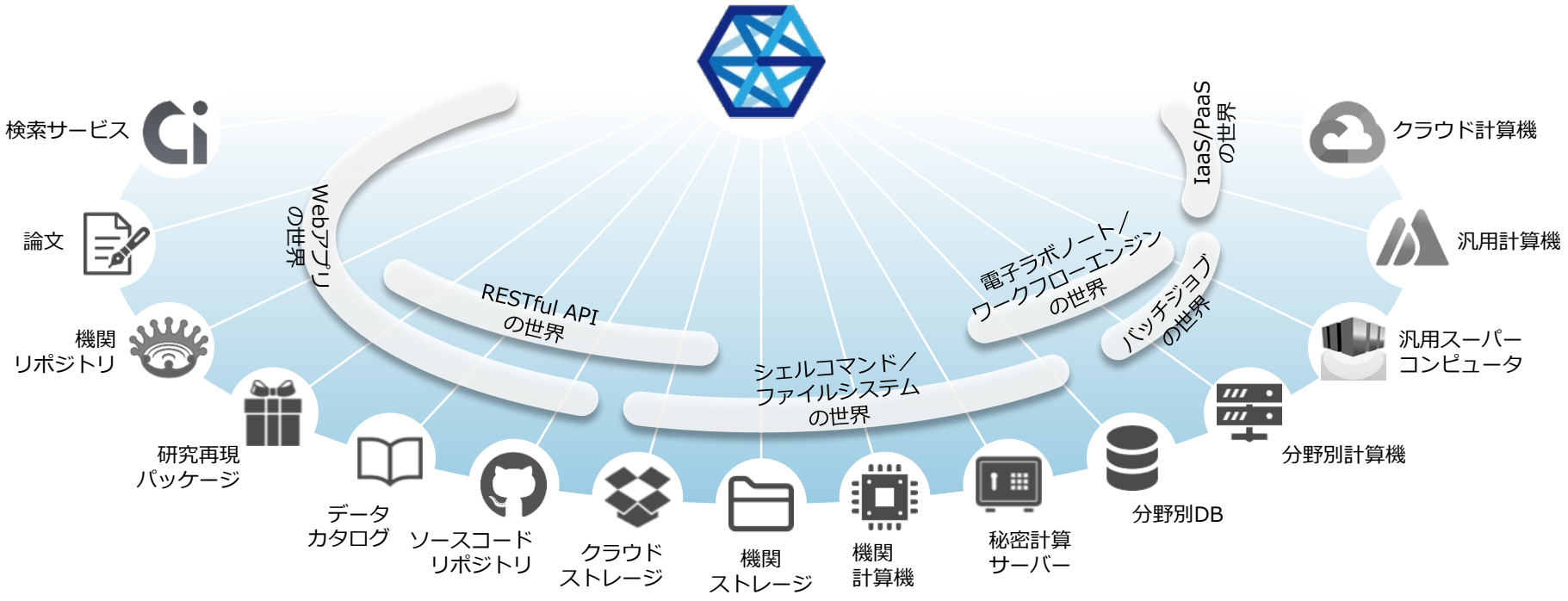
	システム	運用主体	認証方法	同時起動数	ドメイン名＋サーバ証明書	バックエンド
A	Binder	NII	学認	10個／ユーザ	○	Kubernetes
B	Binder	機関	学認, OAuth, LDAP, etc.	任意設定	必要	Kubernetes
C	JupyterHub	研究室等	OAuth, LDAP, ローカル	1個／ユーザ	不要	Linux VM

コード付帯機能群



❶❸ 計算再現パッケージ機能	GRDMプロジェクトをWEKOで公開、他者がGRDMに取り込み再利用	開発中
❷❹ GakuNin RDMデータ解析機能	Jupyterによるデータ解析環境をGRDMから構築	運用中
❺ 外部ワークフローエンジン連携機能	Sapporo 経由でワークフローを実行、結果をGRDMに回収	設計中
❻ 外部HPC連携機能	Open OnDemand からGRDMのファイルを読み書き	提供中
❼ WEKOオンライン分析機能	Jupyterによるデータ解析環境をWEKOから構築	運用中
❽ 秘密計算システム統合機能	秘密分散によるセキュアな解析環境をJupyterから利用	設計中
❾ SINETStream連携検討	SINETStreamによるリアルタイムデータ収集環境を構築	開発中

目指したい将来像



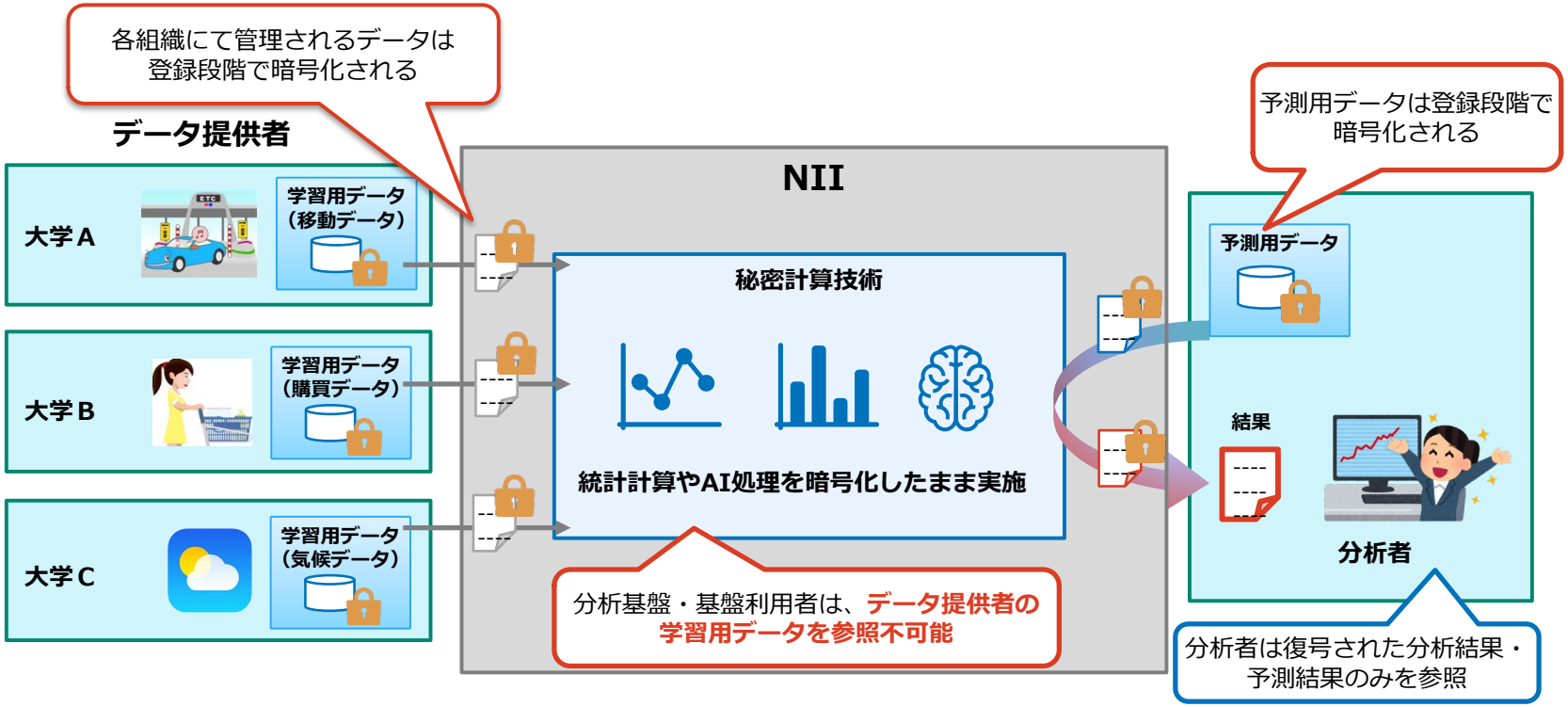
GakuNin RDM を核として
データの世界と計算機の世界を結ぶ

RCOS

rcos@nii.ac.jp

秘密計算システムの大学向けトライアル

- NIIの計算機上にNTTの秘密計算システムを構築し、大学等の研究者に無償提供。現在、実験パートナーを募集中
- 研究者は、システムに手持ちのデータを登録し、秘密計算ライブラリを用いたデータ分析プログラムを書いて実行
- 秘密分散に基づくマルチパーティ計算の実用性を評価していただく → 有用な秘匿解析機能の開発へ



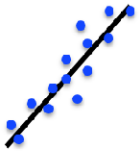
秘密計算システム「算師」

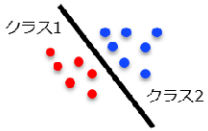
NII RCOS
Research Center for Open Science and Data Platform

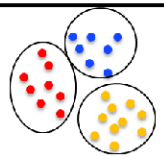
秘密計算AIで実現したAI 4大カテゴリの主要なアルゴリズム

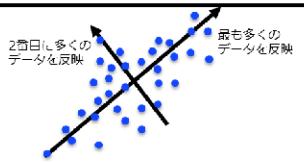
NTT

①回帰（連続値の推定）	②クラス分類（種類毎の分類）	③クラスタリング（類似値による分類）	④データ次元圧縮（主成分の抽出）
教師あり学習		教師なし学習	









ニューラルネットワーク（FFNN、CNN、RNN）
- データ量：大
- 特徴パラメータ量：極めて大
- 予想外の発見的な結果を得やすい（ビッグデータの共有による解析精度向上）

街の店舗横断の購買予測

複数の病院横断の血液分析による疾患予測

決定木、GBDT
- データ量が小～中
- 特徴パラメータ量 小～大
- データが少量でも精度を出しやすい（データ数が揃わない時の分析精度向上）

複数の病院横断の希少疾患予測

クレジットカード会社横断の不正取引検知

Lasso回帰

ロジスティック回帰

階層型クラスタリング
k-means
- データ間の類似度にもとづいてデータをグループ分けする手法

複数の病院横断の疾病患者グループ分析

街の店舗横断の顧客グループ分析

主成分分析
- 多量の変数を少量の変数に置換、要約
- 総合力に影響している項目が把握できることで分析を効率化する

街の店舗横断の購買傾向分析

製造メーカ横断の製品評価分析

スライド提供: NTT社会情報研究所

詳しい情報

お申し込み方法

- 以下の内容を記載したメールをお送りください。
  - 研究者の氏名、所属、連絡先メールアドレス
  - 研究内容の概略、秘密計算を利用するメリット
- 個別にヒアリングを行い、詳細を決定します。
  - 原則として1チームにつき1セットの秘密計算環境を提供します。

資料

- トライアル概要
  https://rcos.nii.ac.jp/news/2023/01/20230123-0/
- ニュースリリース
  https://www.nii.ac.jp/news/release/2023/0123.html

お問い合わせ

- sc-trial2023@nii.ac.jp