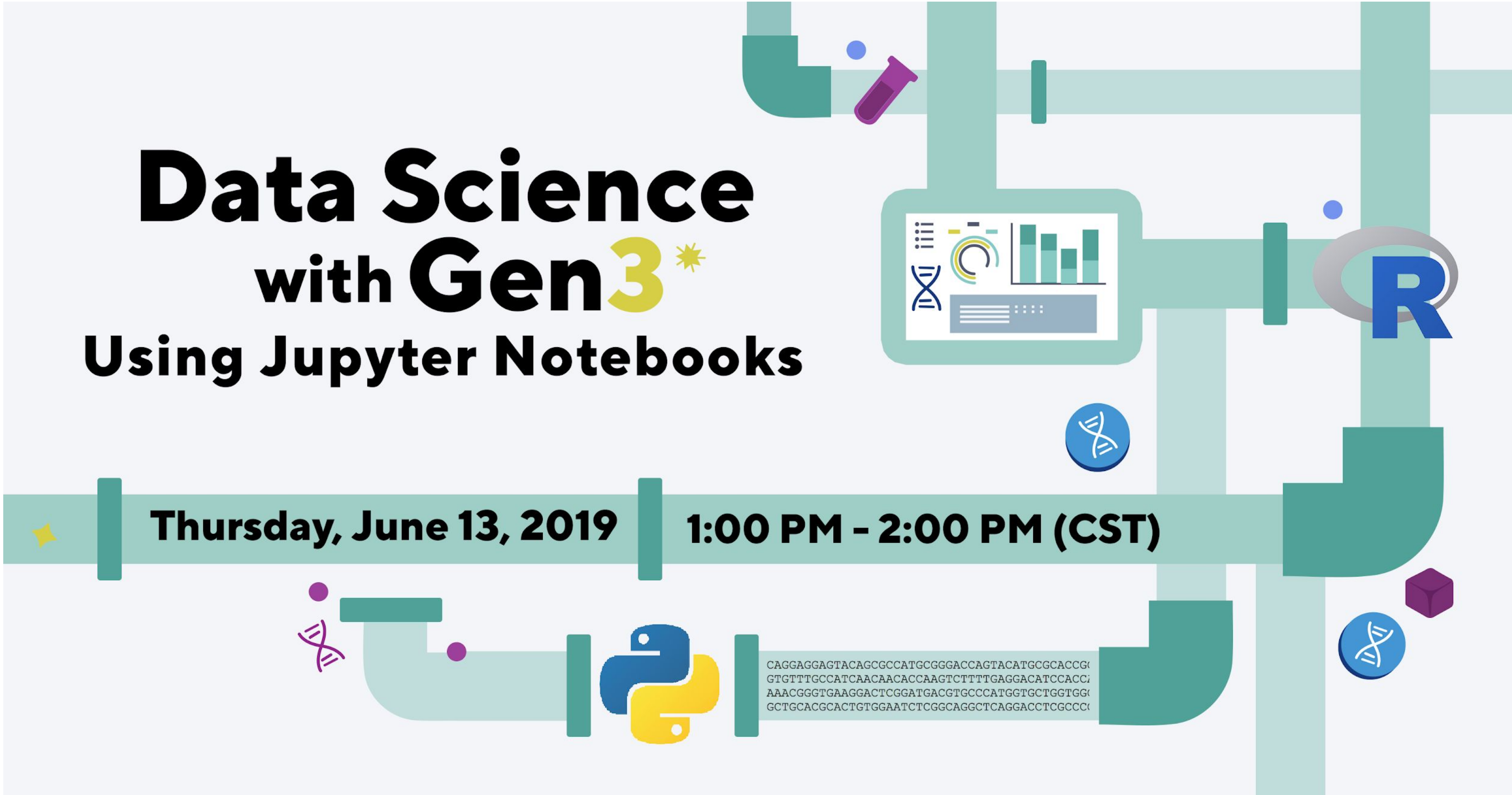
Data Science with Gen3 Using Jupyter Notebooks


R
CAGGAGGAGTACAGCGCCATGCGGGACCAGTACATGCGCACCG
GTGTTTGCCATCAACAACACCAAGTCTTTTGAGGACATCCACC
AAACGGGTGAAGGACTCGGATGACGTGCCCATGGTGCTGGTGG
GCTGCACGCACTGTGGAATCTCGGCAGGCTCAGGACCTCGCCC

Thursday, June 13, 2019

1:00 PM - 2:00 PM (CST)

Data Science with Gen3

Using Jupyter Notebooks

Yilin Xu, Ph.D.

Center for Translational Data Science,
University of Chicago

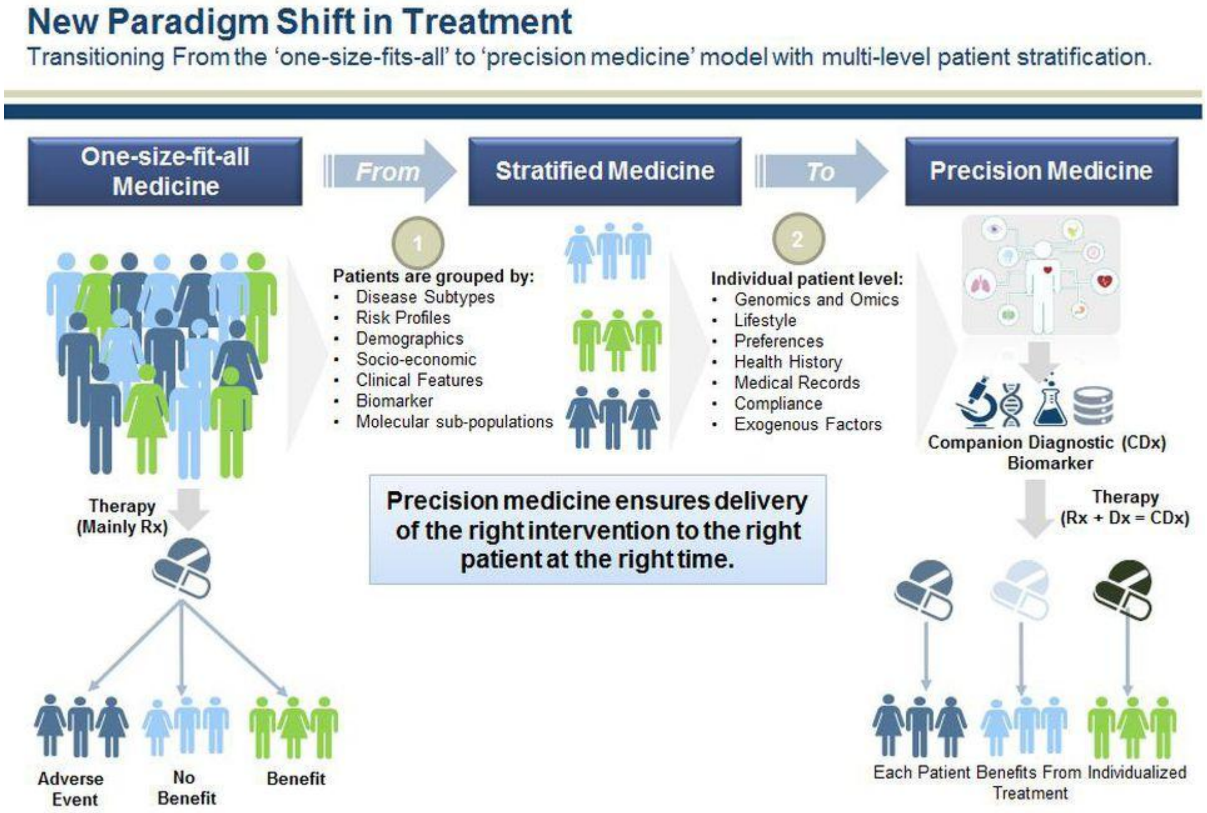
June 13, 2019

Gen3 in data science

The Gen3 data model is flexible and able to host data to be analyzed in different scientific fields.

In this webinar, we will speak about using Gen3 for data analysis in general and show an example of Gen3 used for precision medicine.

Precision Medicine and Data Science



New Paradigm Shift in Treatment

Transitioning From the 'one-size-fits-all' to 'precision medicine' model with multi-level patient stratification.

Patients are grouped by:

- Disease Subtypes
- Risk Profiles
- Demographics
- Socio-economic
- Clinical Features
- Biomarker
- Molecular sub-populations

Individual patient level:

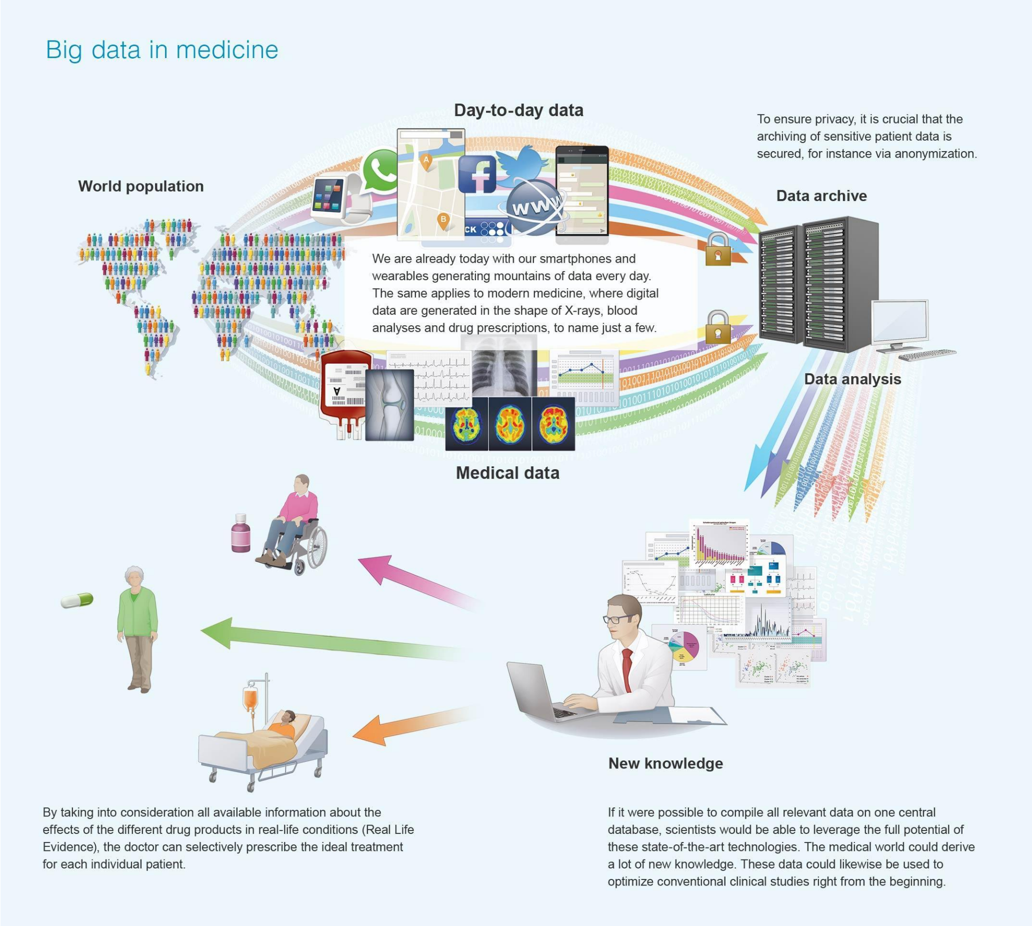
- Genomics and Omics
- Lifestyle
- Preferences
- Health History
- Medical Records
- Compliance
- Exogenous Factors

Precision medicine ensures delivery of the right intervention to the right patient at the right time.

Therapy (Mainly Rx)
Adverse Event
No Benefit
Benefit
Companion Diagnostic (CDx) Biomarker
Therapy (Rx + Dx = CDx)
Each Patient Benefits From Individualized Treatment

Source: Frost & Sullivan -Figure 1: New Paradigm Shift in Treatment, as referenced in this forbes article

Big Data in Medicine



Big data in medicine

Day-to-day data

To ensure privacy, it is crucial that the archiving of sensitive patient data is secured, for instance via anonymization.

World population

Data archive

We are already today with our smartphones and wearables generating mountains of data every day. The same applies to modern medicine, where digital data are generated in the shape of X-rays, blood analyses and drug prescriptions, to name just a few.

Data analysis

Medical data

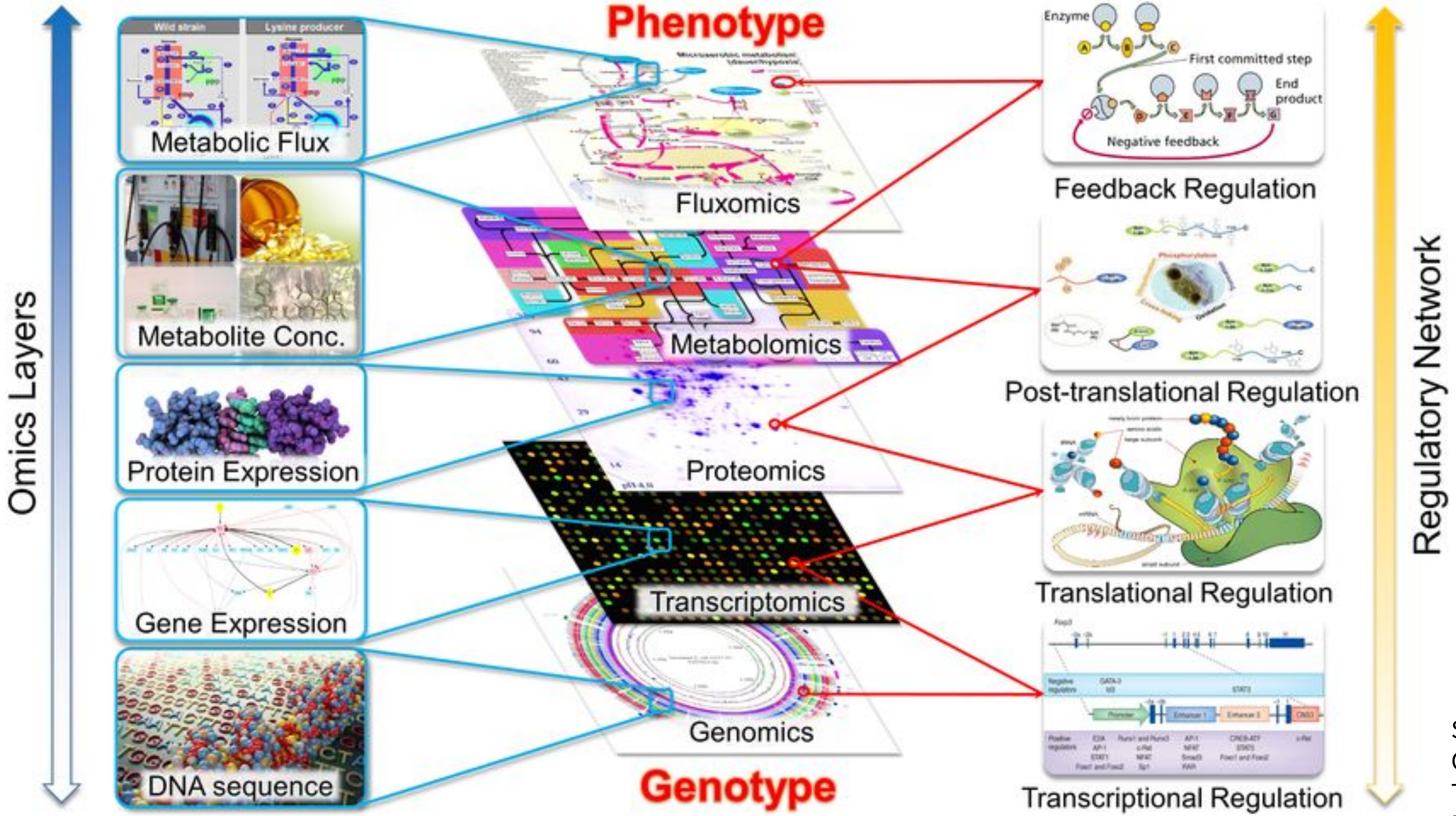
New knowledge

By taking into consideration all available information about the effects of the different drug products in real-life conditions (Real Life Evidence), the doctor can selectively prescribe the ideal treatment for each individual patient.

If it were possible to compile all relevant data on one central database, scientists would be able to leverage the full potential of these state-of-the-art technologies. The medical world could derive a lot of new knowledge. These data could likewise be used to optimize conventional clinical studies right from the beginning.

Source: Bayer Research, 30 November 2016 Big data in medicine

Big Data in Medicine



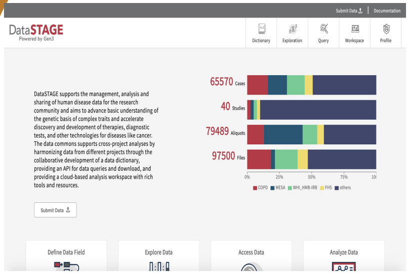
Source: Guo W, Feng X (2016) OM-FBA: Integrate Transcriptomics Data with Flux Balance Analysis to Decipher the Cell Metabolism.

GEN3 Data Commons

*Data commons co-locate data, storage and computing infrastructure with commonly used software services, **tools & apps** for analyzing and sharing data to create a resource for the research community.*

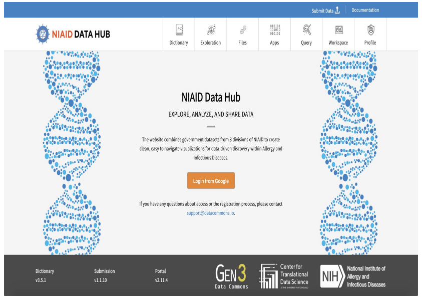
Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.

The Gen3 Ecosystem



DataSTAGE supports the management, analysis and sharing of human disease data for the research community and aims to advance basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cancer. The data commons supports cross-project analyses by harmonizing data from different projects through the collaborative development of a data dictionary, providing an API for data queries and download, and providing a cloud-based analysis workspace with rich tools and resources.

NHLBI Data Stage



NIAID Data Hub

EXPLORE, ANALYZE, AND SHARE DATA

The website combines government datasets from 3 divisions of NIAID to create clean, easy to navigate visualizations for data-driven discovery within Allergy and Infectious Diseases.

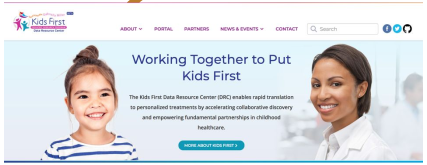
If you have any questions about access or the registration process, please contact support@datacommons.io.

Dictionary v3.5.1

Submission v1.1.10

Portal v2.11.4

NIAID Data Hub

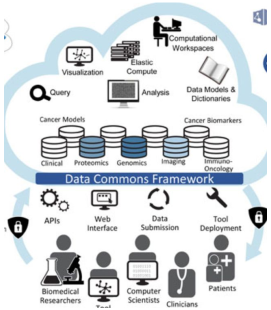


Working Together to Put Kids First

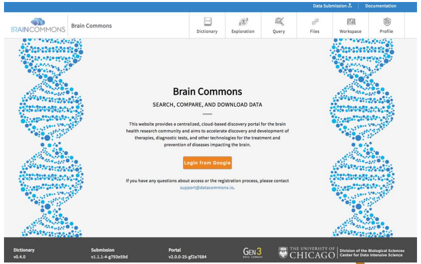
The Kids First Data Resource Center (DRC) enables rapid translation to personalized treatments by accelerating collaborative discovery and empowering fundamental partnerships in childhood healthcare.

Kids First Data Resource

NHGRI AnVIL



NCI CRDC



Brain Commons

SEARCH, COMPARE, AND DOWNLOAD DATA

This website provides a centralized, cloud-based discovery portal for the brain health research community and aims to accelerate discovery and development of therapies, diagnostic tests, and other technologies for the treatment and prevention of diseases impacting the brain.

If you have any questions about access or the registration process, please contact support@datacommons.io.

Dictionary v0.4.0

Submission v1.1.1-4-g793e59d

Portal v2.0.0-25-gf2a7684

Data commons from other foundations

Narrow Middle Architecture for Data Ecosystem

GEN3 Data Commons

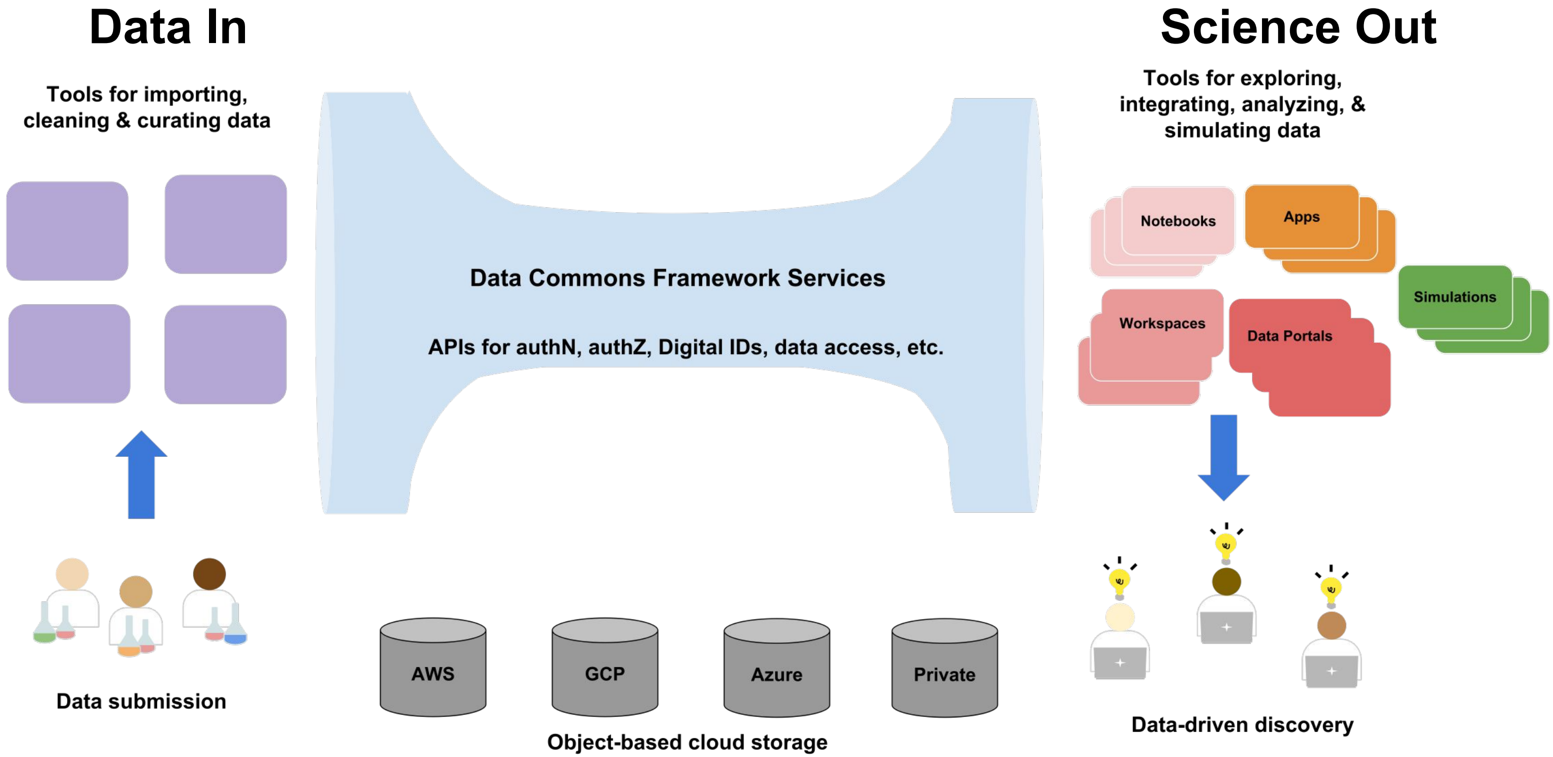

Data In
Tools for importing, cleaning & curating data
Data submission
Data Commons Framework Services
APIs for authN, authZ, Digital IDs, data access, etc.
AWS
GCP
Azure
Private
Object-based cloud storage
Science Out
Tools for exploring, integrating, analyzing, & simulating data
Notebooks
Apps
Simulations
Workspaces
Data Portals
Data-driven discovery

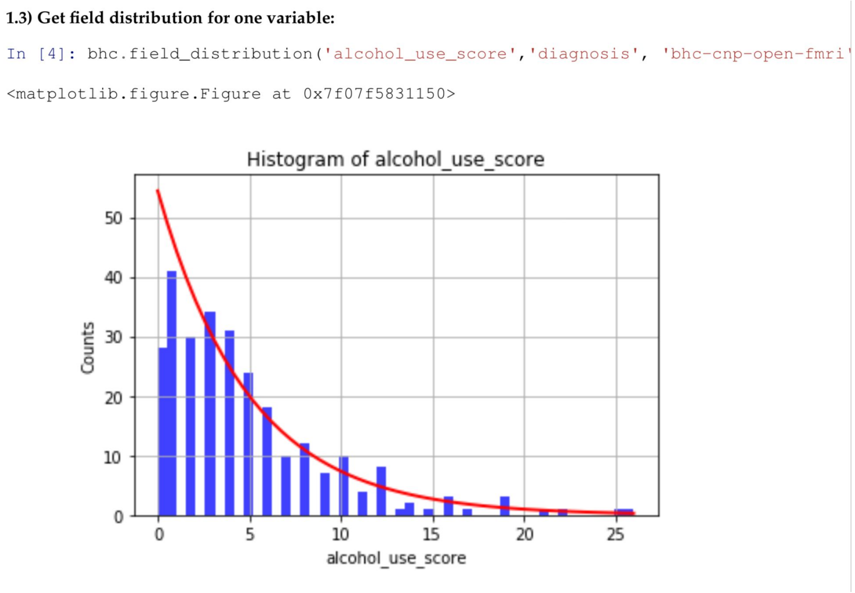
Diagram: Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles & Practice of Oncology, 2018, Volume 24, Number 3, May/June 2018.

Outline

- Build Notebook in Gen3
- Select virtual cohort in data portal
- Notebook example
- Coming feature for analysis

Notebooks in the Workspace

- Notebooks combine annotation, code, and output visualization



1.3) Get field distribution for one variable:

```
In [4]: bhc.field_distribution('alcohol_use_score','diagnosis', 'bhc-cnp-open-fmri'
```

```
<matplotlib.figure.Figure at 0x7f07f5831150>
```

- Gen3 currently supports Jupyter notebooks for a “lightweight workspace”

Notebooks in Workspace

- An authorized user’s workspace in a given commons includes a persistent drive in which analysis notebooks, scripts, data files, etc., are saved and persist even after logout

- Gen3 Jupyter notebooks support both R and Python language

- User Documentation available at https://gen3.org/resources/user/analyze-data/

jupyterhub

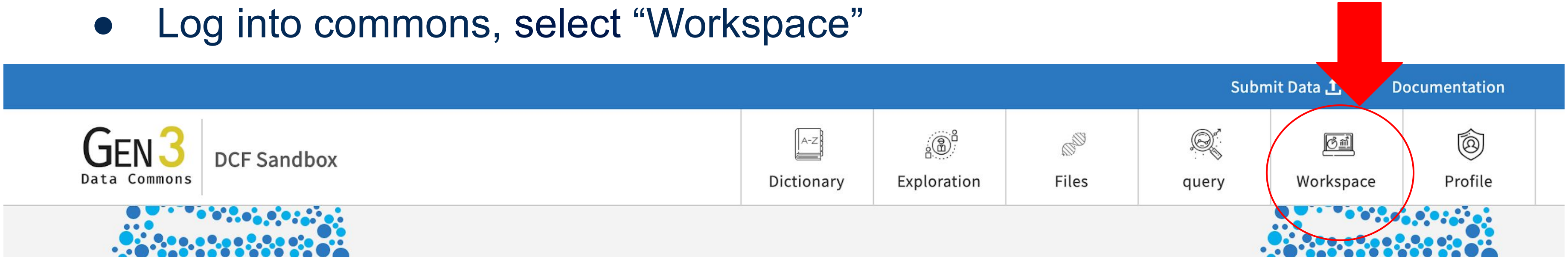
Gen3 Python SDK

GEN3 Data Commons

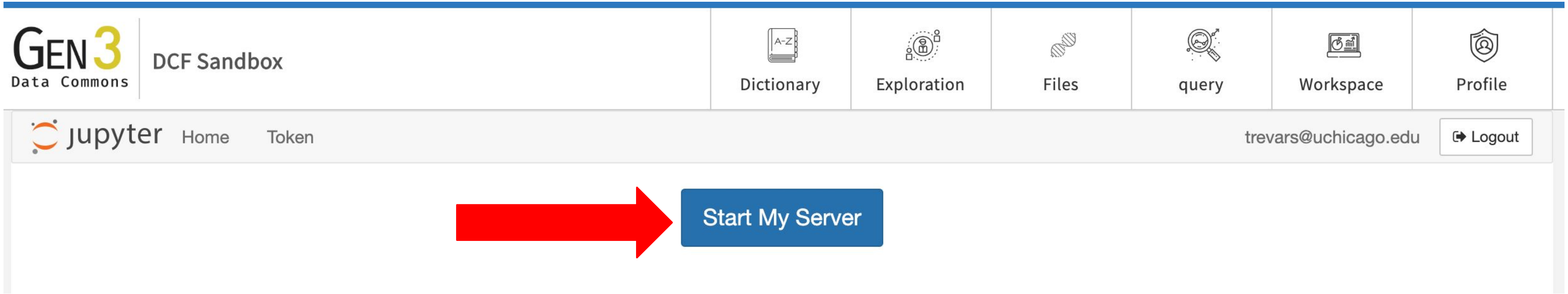
- The Gen3 SDK facilitates data analysis in notebooks by providing a Python library that makes calls to Gen3 APIs easier.
- Gen3 Python SDK has three classes:
  - Gen3 Auth Helper: Support json web token authentication
  - Gen3 Submission: Submit/Export/Query data from Gen3 submission system
  - Gen3 File Class: Interact with Gen3 file management features
- The Gen3 SDK is available on the python package index (PyPI) at https://pypi.org/project/gen3
- For detailed information on how to use the Gen3 SDK, see the Gen3 SDK documentation at http://gen3sdk-python.rtfd.io.

Access workspace

- Log into commons, select "Workspace"

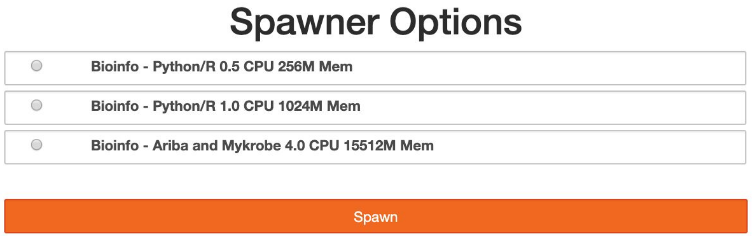


- Click "Start My Server" to start the Jupyter server in your Workspace



Set up workspace

- Choose a virtual machine flavor with the appropriate memory and compute space required for your analysis
- As a Gen3 Data Commons operator, you can configure the different flavors based on the resources available to you, your user community’s needs, and what prices you’re willing to pay
- The notebook runs a container image that is deployed by kubernetes. The tools and packages in the container are available to anyone selecting the flavor.



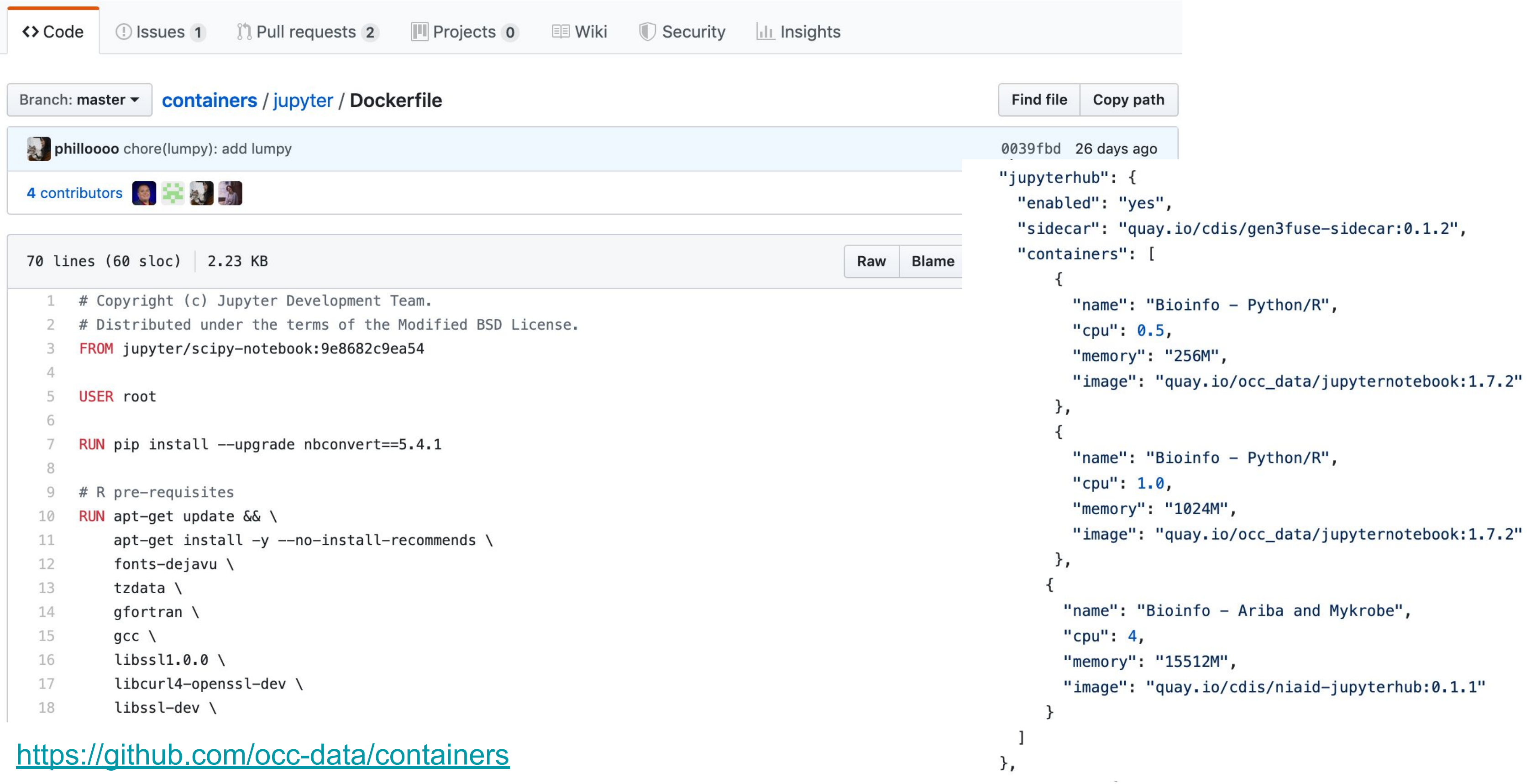
Spawner Options

○ Bioinfo - Python/R 0.5 CPU 256M Mem

○ Bioinfo - Python/R 1.0 CPU 1024M Mem

○ Bioinfo - Ariba and Mykrobe 4.0 CPU 15512M Mem

Docker for notebook



containers / jupyter / Dockerfile

philloooo chore(lumpy): add lumpy 0039fbd 26 days ago

4 contributors

70 lines (60 sloc) | 2.23 KB

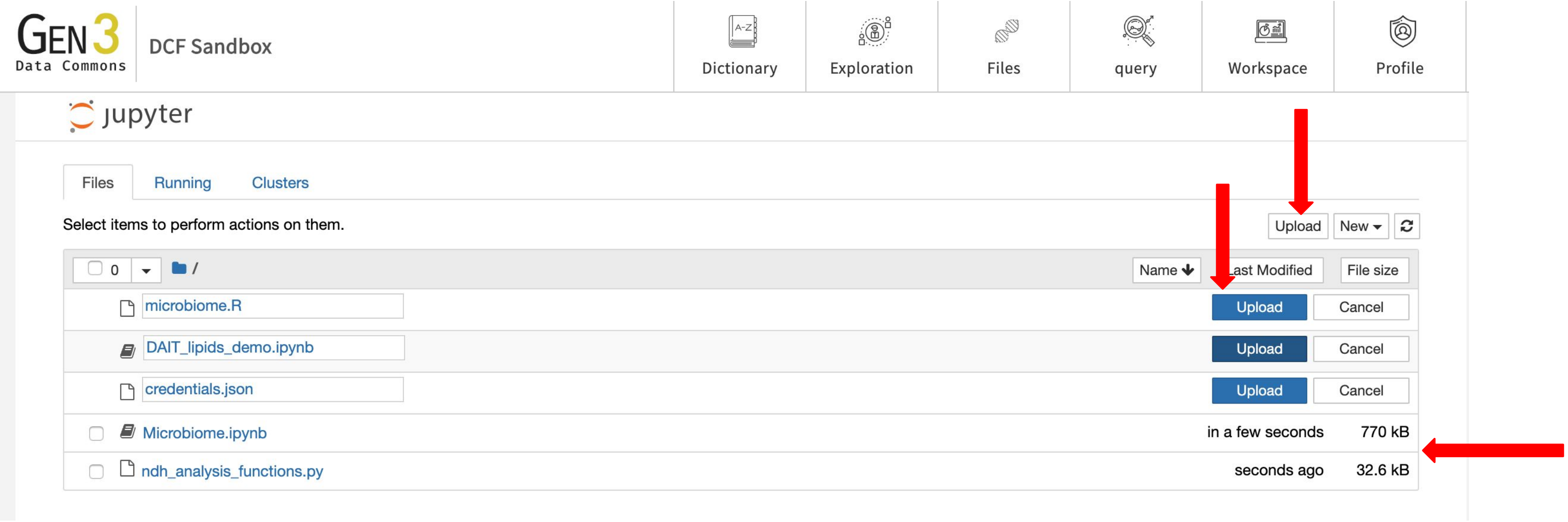
```
# Copyright (c) Jupyter Development Team.
# Distributed under the terms of the Modified BSD License.
FROM jupyter/scipy-notebook:9e8682c9ea54

USER root

RUN pip install --upgrade nbconvert==5.4.1

# R pre-requisites
RUN apt-get update && \
    apt-get install -y --no-install-recommends \
    fonts-dejavu \
    tzdata \
    gfortran \
    gcc \
    libssl1.0.0 \
    libcurl4-openssl-dev \
    libssl-dev \
```

```
"jupyterhub": {
  "enabled": "yes",
  "sidecar": "quay.io/cdis/gen3fuse-sidecar:0.1.2",
  "containers": [
      {
        "name": "Bioinfo - Python/R",
        "cpu": 0.5,
        "memory": "256M",
        "image": "quay.io/occ_data/jupyternotebook:1.7.2"
      },
      {
        "name": "Bioinfo - Python/R",
        "cpu": 1.0,
        "memory": "1024M",
        "image": "quay.io/occ_data/jupyternotebook:1.7.2"
      },
     {
       "name": "Bioinfo - Ariba and Mykrobe",
       "cpu": 4,
       "memory": "15512M",
       "image": "quay.io/cdis/niaid-jupyterhub:0.1.1"
     }
  ]
},
```

https://github.com/occ-data/containers

Run notebook in Gen3 Workspace

- If using an existing notebook and library:
  - Upload any necessary reference files needed for the analysis to your workspace
  - Upload existing Python or R libraries to your workspace
  - You will access clinical data and object files from the data commons within the notebook



Select items to perform actions on them.

microbiome.R

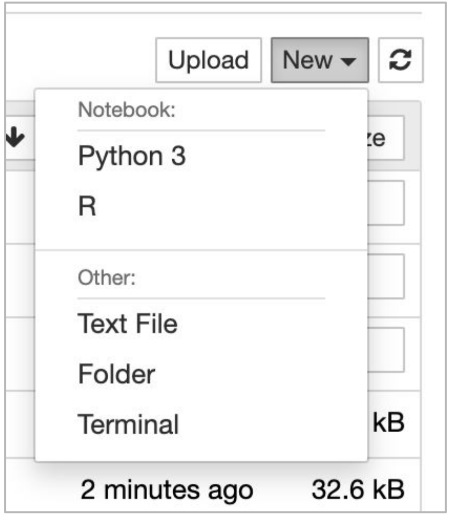
DAIT_lipids_demo.ipynb

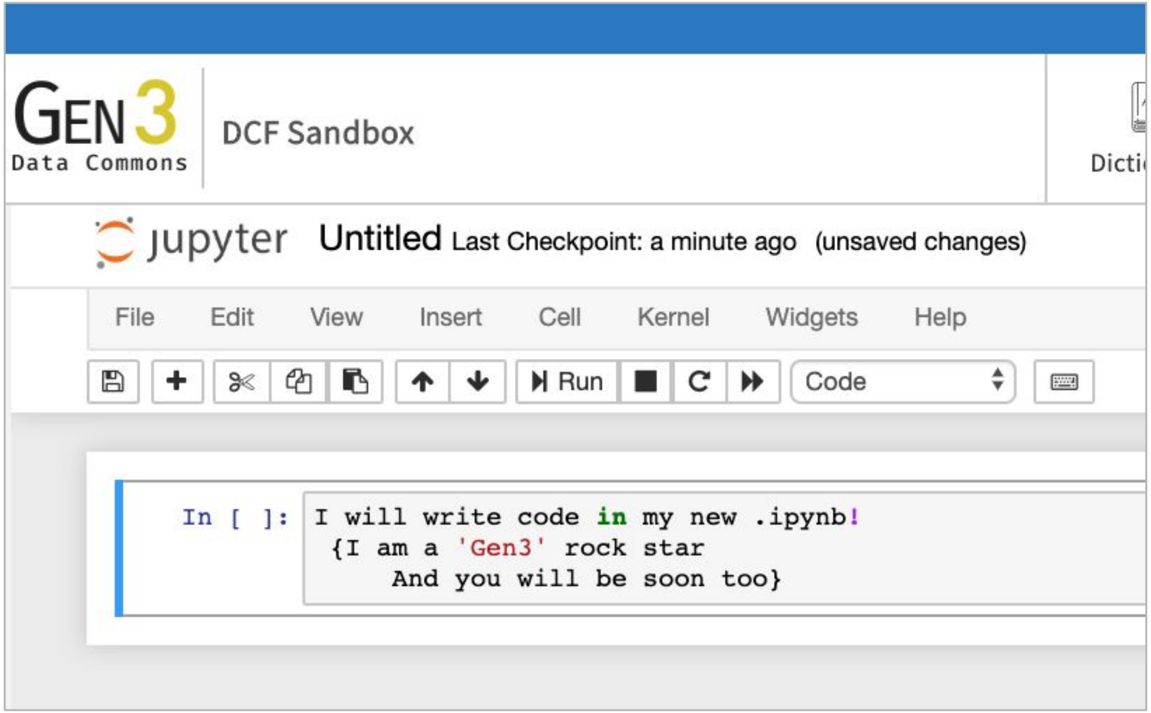
credentials.json

Microbiome.ipynb in a few seconds 770 kB

ndh_analysis_functions.py seconds ago 32.6 kB

Creating notebook and libraries from scratch in the Workspace

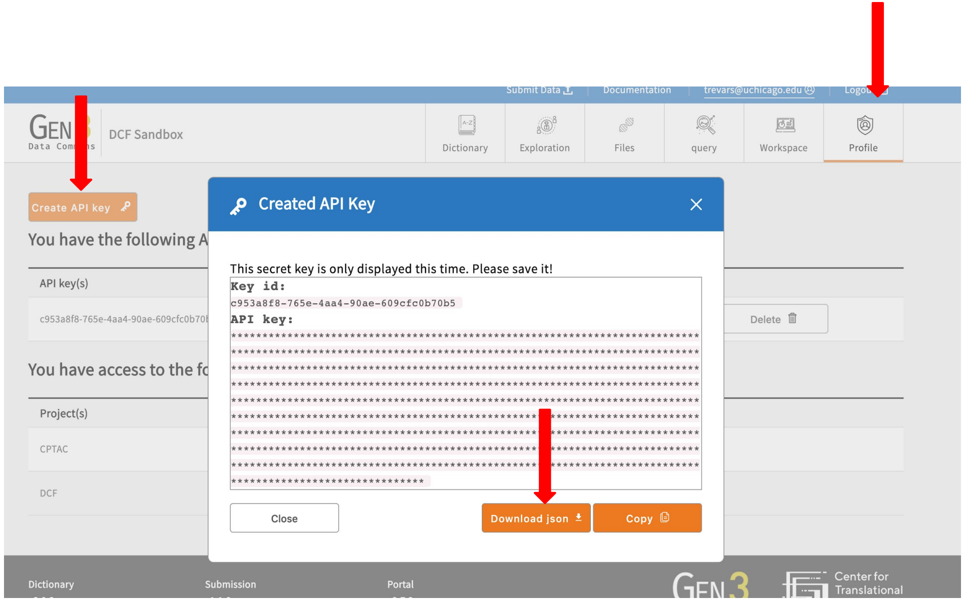




```
I will write code in my new .ipynb!
 {I am a 'Gen3' rock star
     And you will be soon too}
```

Prepare your API key for data accessing

- Create or manage your API keys
  - API keys are valid for a month
  - Used to receive temporary access token that is valid for only 30 minutes
  - Access token must be sent to Gen3 API to access data in the commons
- Upload credentials.json to the workspace to allow you to access data within your commons
  - Be sure your API credentials JSON matches the name of the JSON as you call it in your notebook



You have the following A

API key(s)

c953a8f8-765e-4aa4-90ae-609cfc0b70

You have access to the fo

Project(s)

CPTAC

DCF

Created API Key

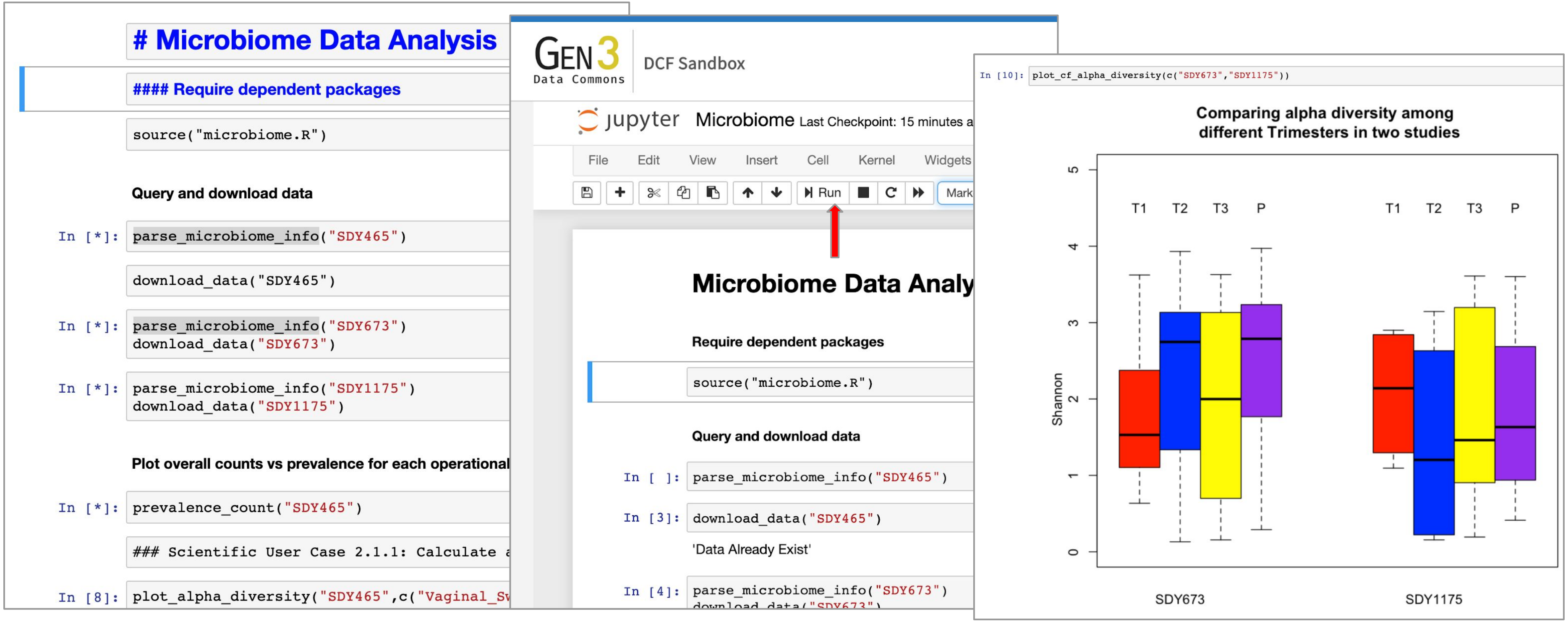
This secret key is only displayed this time. Please save it!

```
Key id:
c953a8f8-765e-4aa4-90ae-609cfc0b70b5
API key:
************************************************************************
************************************************************************
************************************************************************
************************************************************************
************************************************************************
************************************************************************
************************************************************************
************************************************************************
************************************************************************
*****************************
```

Writing and running Jupyter notebook

Start Writing!

Start Analyzing!



```
# Microbiome Data Analysis

#### Require dependent packages

source("microbiome.R")
```

Query and download data

```
In [*]: parse_microbiome_info("SDY465")

        download_data("SDY465")

In [*]: parse_microbiome_info("SDY673")
        download_data("SDY673")

In [*]: parse_microbiome_info("SDY1175")
        download_data("SDY1175")
```

Plot overall counts vs prevalence for each operational

```
In [*]: prevalence_count("SDY465")

        ### Scientific User Case 2.1.1: Calculate a

In [8]: plot_alpha_diversity("SDY465",c("Vaginal_Sw
```

DCF Sandbox

jupyter Microbiome Last Checkpoint: 15 minutes a

Microbiome Data Analy

Require dependent packages

```
source("microbiome.R")
```

Query and download data

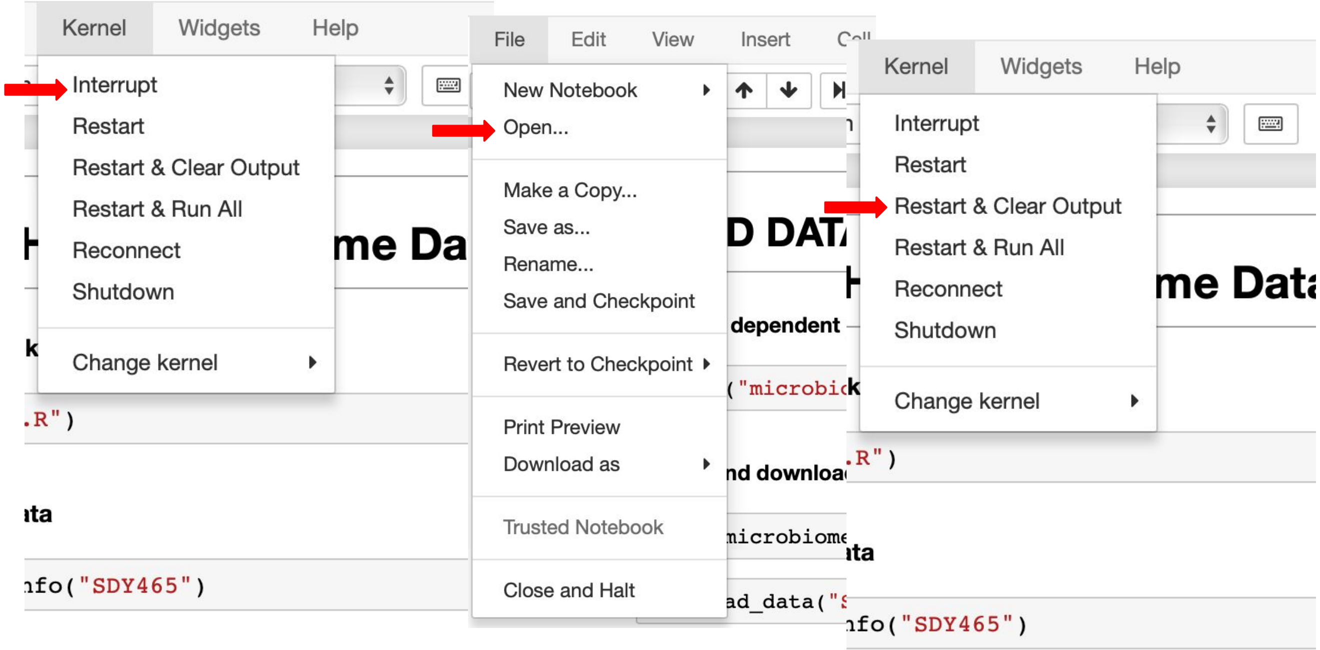
```
In [ ]: parse_microbiome_info("SDY465")

In [3]: download_data("SDY465")

        'Data Already Exist'

In [4]: parse_microbiome_info("SDY673")
        download_data("SDY673")
```

```
In [10]: plot_cf_alpha_diversity(c("SDY673","SDY1175"))
```

Comparing alpha diversity among different Trimesters in two studies

Tune your source code

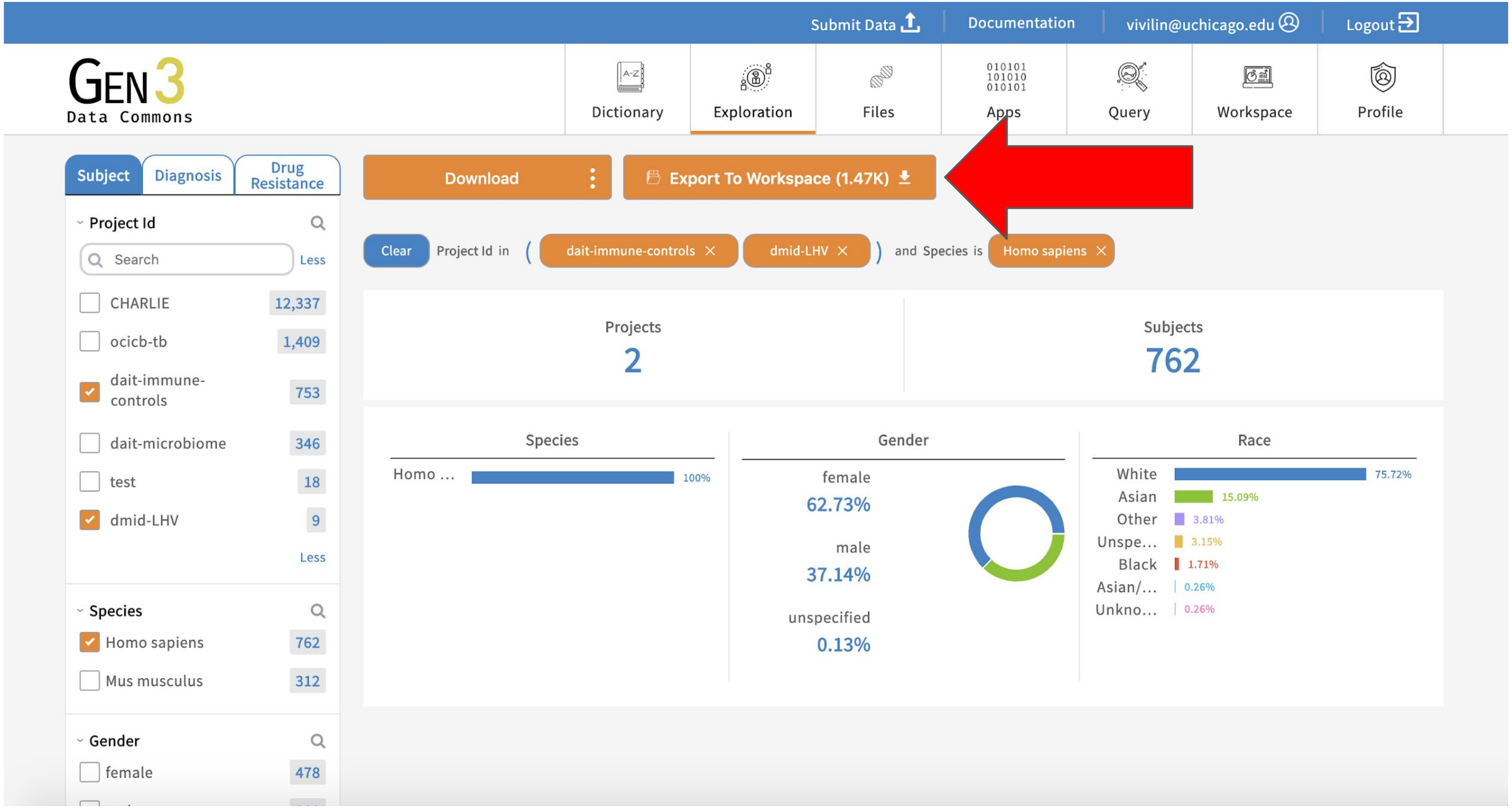
- You can stop your notebook to manage your resources responsibly
- If you update your source code or library, you can restart to use the updated code



Sharing notebooks

- Reasons to share a notebook:
    - Review and feedback on methods
    - Other scientists use your analysis on different data based on their access
    - Accompany publication
- Suggestions for sharing notebooks
    - Remove results before sharing
    - Consider GitHub repositories for community accessible notebooks with associated files and libraries
- How to use a shared notebook
    - Upload all libraries and necessary files to your workspace, including notebook
    - Ensure your credentials.json is current and in your workspace

Virtual cohort selection in data portal



Project Id in (dait-immune-controls ✕ dmid-LHV ✕) and Species is Homo sapiens ✕

Projects: 2

Subjects: 762

Project Id

Search · Less

☐	CHARLIE	12,337
☐	ocicb-tb	1,409
☑	dait-immune-controls	753
☐	dait-microbiome	346
☐	test	18
☑	dmid-LHV	9

Less

Species

☑	Homo sapiens	762
☐	Mus musculus	312

Gender

☐	female	478

Species

Homo …	100%

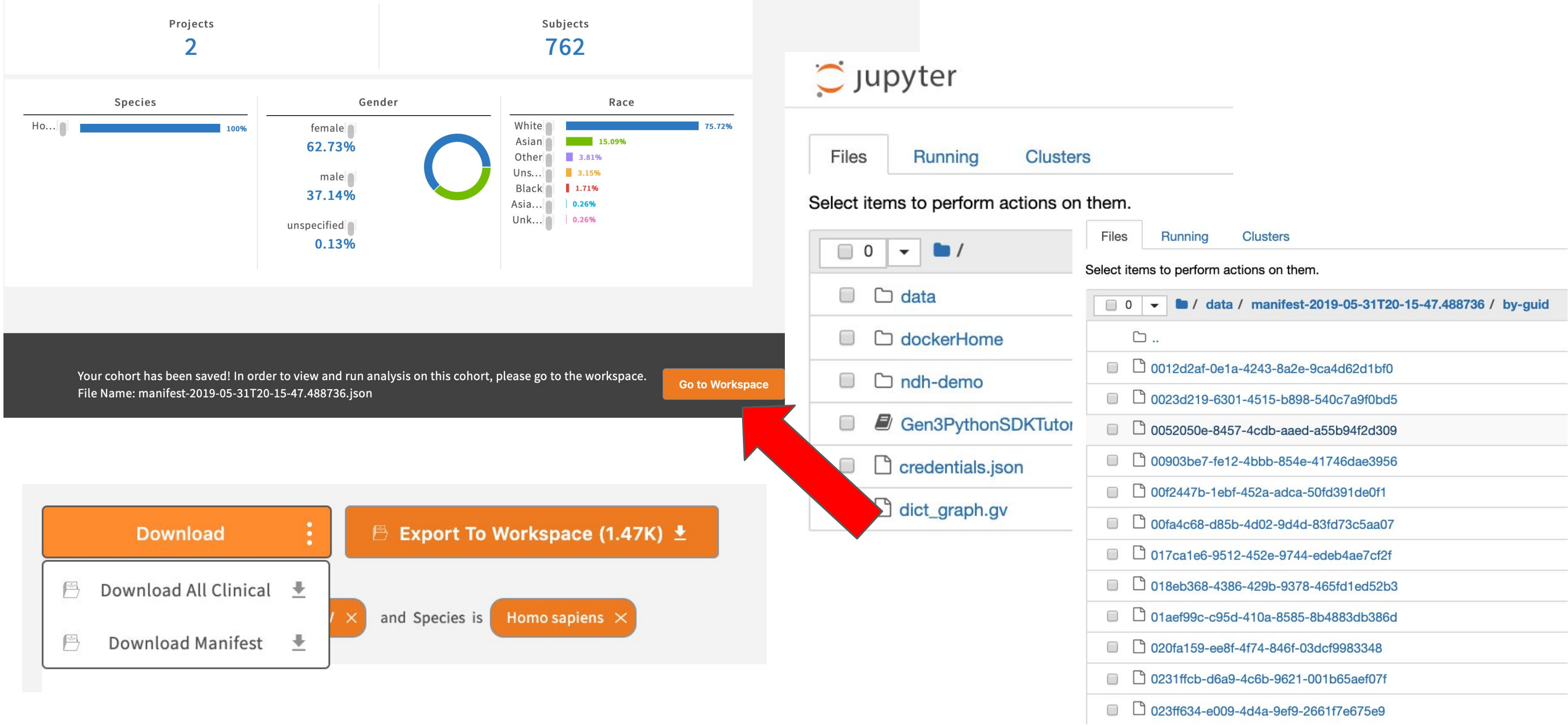
Gender

female	62.73%
male	37.14%
unspecified	0.13%

Race

White	75.72%
Asian	15.09%
Other	3.81%
Unspe…	3.15%
Black	1.71%
Asian/…	0.26%
Unkno…	0.26%

Virtual cohort selection in data portal



Projects 2

Subjects 762

Your cohort has been saved! In order to view and run analysis on this cohort, please go to the workspace.
File Name: manifest-2019-05-31T20-15-47.488736.json

Select items to perform actions on them.

- 0 /
- data
- dockerHome
- ndh-demo
- Gen3PythonSDKTutoı
- credentials.json
- dict_graph.gv

Select items to perform actions on them.

0 / data / manifest-2019-05-31T20-15-47.488736 / by-guid

- ..
- 0012d2af-0e1a-4243-8a2e-9ca4d62d1bf0
- 0023d219-6301-4515-b898-540c7a9f0bd5
- 0052050e-8457-4cdb-aaed-a55b94f2d309
- 00903be7-fe12-4bbb-854e-41746dae3956
- 00f2447b-1ebf-452a-adca-50fd391de0f1
- 00fa4c68-d85b-4d02-9d4d-83fd73c5aa07
- 017ca1e6-9512-452e-9744-edeb4ae7cf2f
- 018eb368-4386-429b-9378-465fd1ed52b3
- 01aef99c-c95d-410a-8585-8b4883db386d
- 020fa159-ee8f-4f74-846f-03dcf9983348
- 0231ffcb-d6a9-4c6b-9621-001b65aef07f
- 023ff634-e009-4d4a-9ef9-2661f7e675e9

Summary of using notebook in Gen3

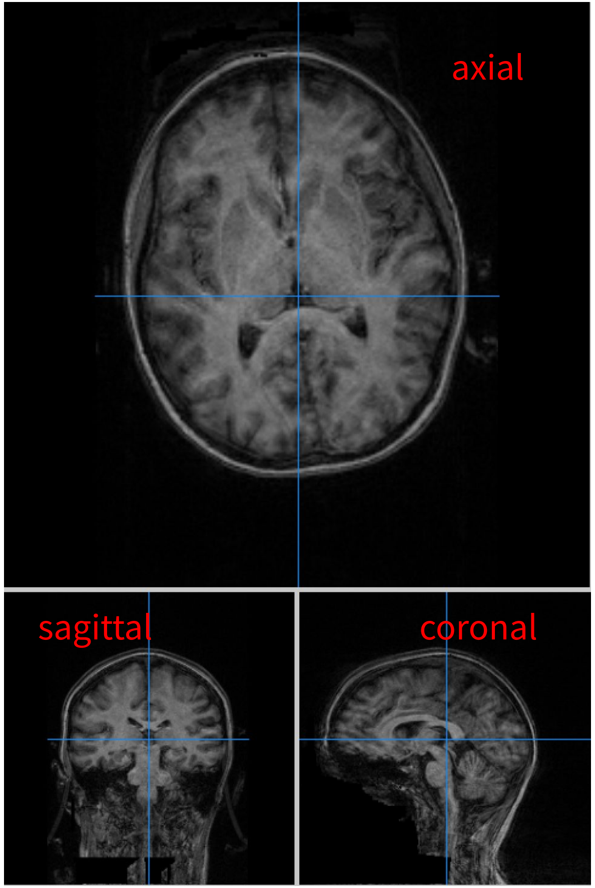
Gen3 Data Commons

- Select virtual machine flavor with the appropriate memory and compute space required for your analysis
- Import API key and upload or write reference files, library and Jupyter notebook
- Select virtual cohort from data portal and import clinical and object data in virtual machine to run the notebook and tune the library
- Share notebook with community

Notebook example

Outline of the notebook:

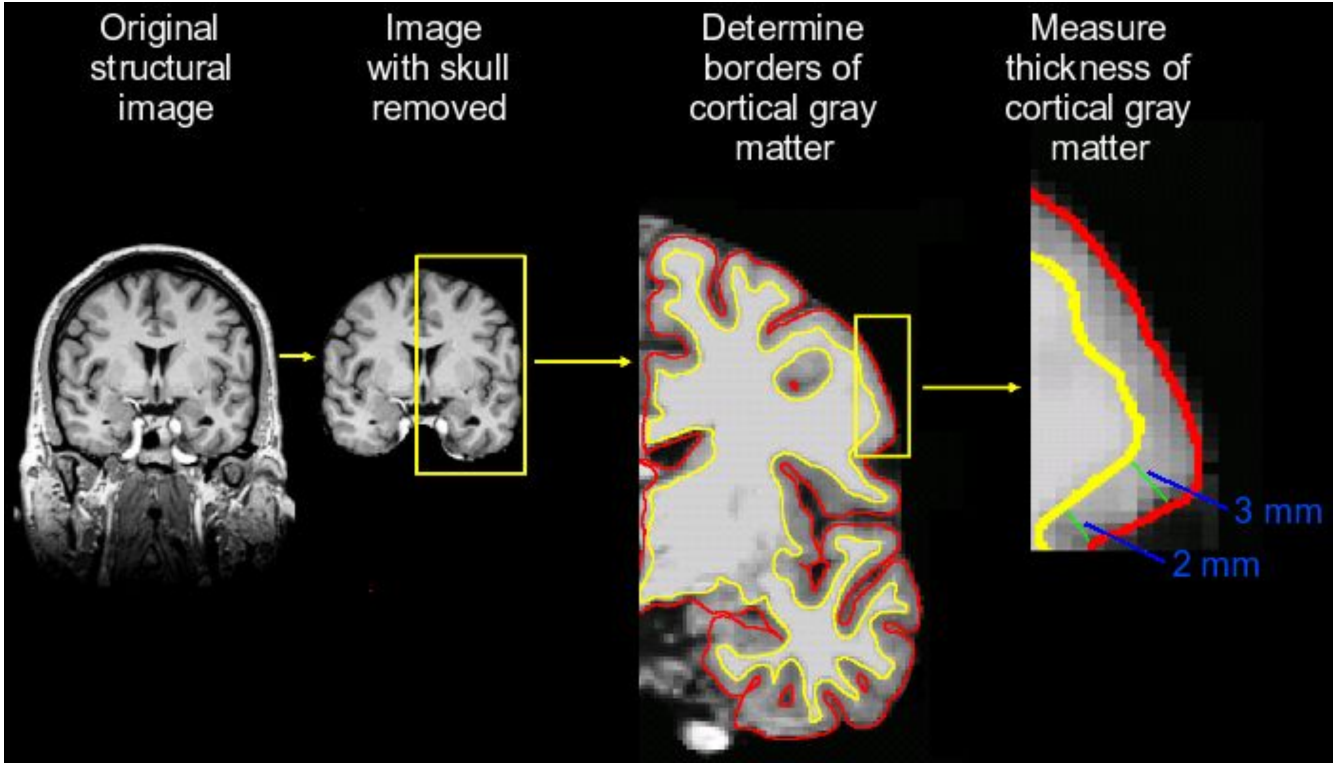
- Analyze MRI images getting average cortical thickness measurement in different regions
- Visualize brain surface segmented into different regions
- Compare cortical thickness across groups of patients with different brain disease



Source: https://www.openfmri.org/dataset/ds000030/

Notebook example

Cortical thickness measurement


Original structural image
Image with skull removed
Determine borders of cortical gray matter
Measure thickness of cortical gray matter
3 mm
2 mm

Source: http://www.martinos.org/neurorecovery/technology.htm

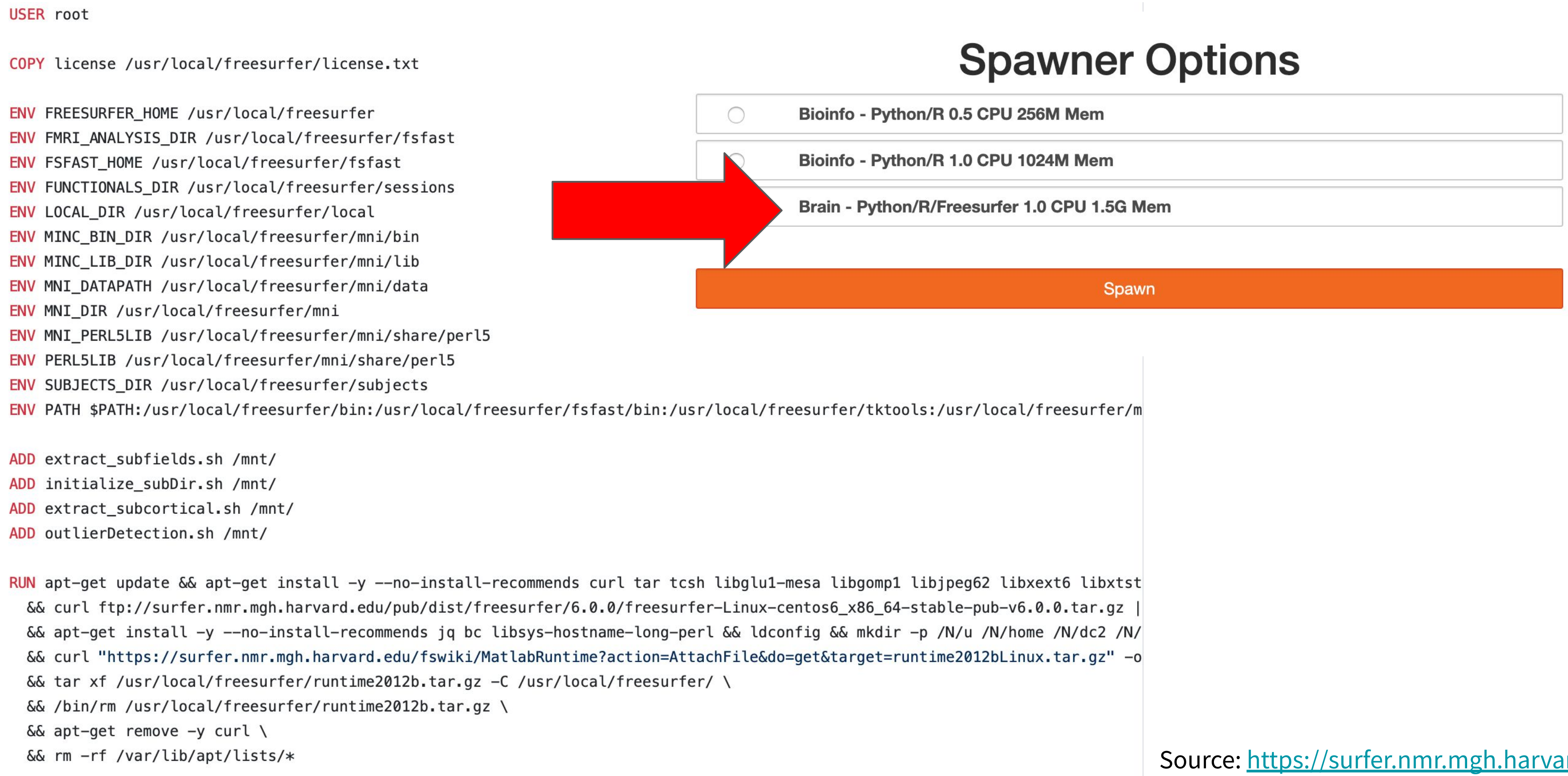
Notebook example

Freesurfer Enigma pipeline: recon-all

- Normalize brain signal intensity, skull-stripping, white matter and gray matter segmentation, and delineation of the gray-white interface
- Divide surface into separate cortical regions
- Surface area and mean cortical thickness was extracted for each of the 68 regions (34 per hemisphere)

Source: https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all

Notebook example



```
USER root

COPY license /usr/local/freesurfer/license.txt

ENV FREESURFER_HOME /usr/local/freesurfer
ENV FMRI_ANALYSIS_DIR /usr/local/freesurfer/fsfast
ENV FSFAST_HOME /usr/local/freesurfer/fsfast
ENV FUNCTIONALS_DIR /usr/local/freesurfer/sessions
ENV LOCAL_DIR /usr/local/freesurfer/local
ENV MINC_BIN_DIR /usr/local/freesurfer/mni/bin
ENV MINC_LIB_DIR /usr/local/freesurfer/mni/lib
ENV MNI_DATAPATH /usr/local/freesurfer/mni/data
ENV MNI_DIR /usr/local/freesurfer/mni
ENV MNI_PERL5LIB /usr/local/freesurfer/mni/share/perl5
ENV PERL5LIB /usr/local/freesurfer/mni/share/perl5
ENV SUBJECTS_DIR /usr/local/freesurfer/subjects
ENV PATH $PATH:/usr/local/freesurfer/bin:/usr/local/freesurfer/fsfast/bin:/usr/local/freesurfer/tktools:/usr/local/freesurfer/m

ADD extract_subfields.sh /mnt/
ADD initialize_subDir.sh /mnt/
ADD extract_subcortical.sh /mnt/
ADD outlierDetection.sh /mnt/

RUN apt-get update && apt-get install -y --no-install-recommends curl tar tcsh libglu1-mesa libgomp1 libjpeg62 libxext6 libxtst
  && curl ftp://surfer.nmr.mgh.harvard.edu/pub/dist/freesurfer/6.0.0/freesurfer-Linux-centos6_x86_64-stable-pub-v6.0.0.tar.gz |
  && apt-get install -y --no-install-recommends jq bc libsys-hostname-long-perl && ldconfig && mkdir -p /N/u /N/home /N/dc2 /N/
  && curl "https://surfer.nmr.mgh.harvard.edu/fswiki/MatlabRuntime?action=AttachFile&do=get&target=runtime2012bLinux.tar.gz" -o
  && tar xf /usr/local/freesurfer/runtime2012b.tar.gz -C /usr/local/freesurfer/ \
  && /bin/rm /usr/local/freesurfer/runtime2012b.tar.gz \
  && apt-get remove -y curl \
  && rm -rf /var/lib/apt/lists/*
```

Spawner Options

○ Bioinfo - Python/R 0.5 CPU 256M Mem

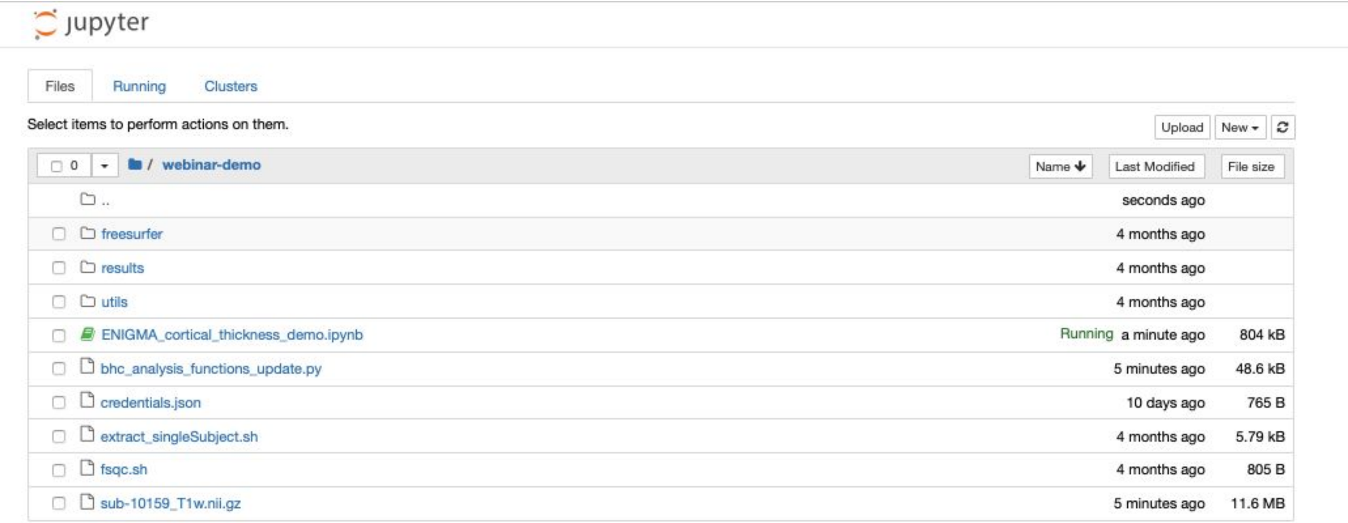
○ Bioinfo - Python/R 1.0 CPU 1024M Mem

Brain - Python/R/Freesurfer 1.0 CPU 1.5G Mem

Source: https://surfer.nmr.mgh.harvard.edu

Workspace jupyter notebook demonstration

- Now, we will take a look at the Jupyter notebook



Select items to perform actions on them.

/ webinar-demo

	Name ↓		Last Modified	File size
	..		seconds ago	
□	freesurfer		4 months ago	
□	results		4 months ago	
□	utils		4 months ago	
□	ENIGMA_cortical_thickness_demo.ipynb	Running	a minute ago	804 kB
□	bhc_analysis_functions_update.py		5 minutes ago	48.6 kB
□	credentials.json		10 days ago	765 B
□	extract_singleSubject.sh		4 months ago	5.79 kB
□	fsqc.sh		4 months ago	805 B
□	sub-10159_T1w.nii.gz		5 minutes ago	11.6 MB

Coming soon for analysis

- Additional tools for the workspace is in development, including R Studio notebooks, Galaxy, and more

Coming soon!

- Clinical data export to workspace

- Gen3 workflow execution service. The Gen3 workflow execution service will use its own cwl engine, developed in-house, to execute workflows. User passes the cwl workflow ("packed") as a JSON, as well as a JSON specifying workflow inputs, to the workflow execution service API.

Learn More

- github.com/uc-cdis
- gen3.org
- Gen3 Community on Slack
- support@datacommons.io
- ctds.uchicago.edu

Selected Data Commons Using Gen3

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

NIH National Heart, Lung, and Blood Institute
DataSTAGE

NIAID DATA HUB PILOT

AnVIL

BRAINCOMMONS

NIH National Human Genome Research Institute

Next Webinar

Gen3 Data Commons

Gen3 Webinar

Thursday, August 8, 2019
1:00 PM – 2:00 PM (CST)

Questions?

GEN3 Data Commons