COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Master's Research and Projects: FNLCR and Columbia University Partnership

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Master's Research and Projects: FNLCR and Columbia University Partnership

2:00 PM: Introduction Prof. Michael Robbins

2:10 PM: Opening remarks Dr. Ethan Dmitrovsky

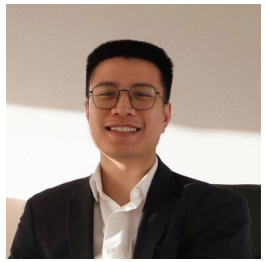
2:20 PM: Student Presentations

- Cloud Deployment, Optimization Strategies for Teaching, Training and Collaborative Reproducible Research
- Survey to Identify Emerging Infectious Disease Datasets for Machine Learning
- Survey to Identify Cancer Datasets for Machine Learning
- Q & A

3:20 PM: Closing remarks Dr. Eric Stahlberg

Project Team

Mahitha Kotipalli

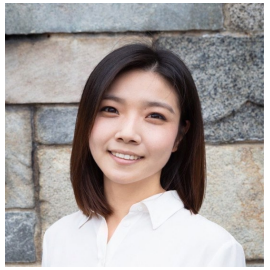


Jim Hu

Niranjana Moleyar

Malin Ortenblad

Kerry Hu



Jie Chen

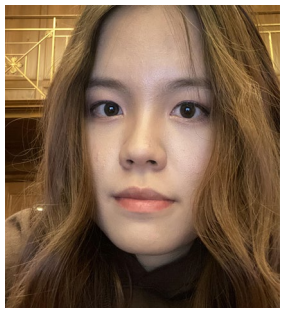
Mengyao He

Om Vaghasia

Panagiotis Misirlis

Jiaxi Zhou

Xinyao Wang



Qinwei Zhang

Yue Hu

Zihui Zhou

Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Soder
TA, Project Manager
Columbia University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

Cloud Deployment, Optimization Strategies for Teaching, Training and Collaborative Reproducible Research

Team Members: Jiaxi Zhou, Xinyao Wang, Zihui Zhou, Yue Hu, Qinwei Zhang
MSOR from Columbia University
Sep. 8, 2020

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Team Introduction

Jiaxi Zhou
Columbia University
M.S. in Operations Research
Software: Binder
Email: jz3150@columbia.edu

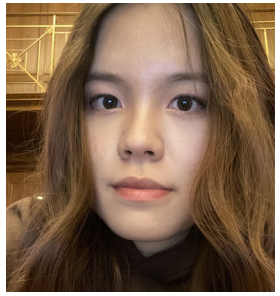
Jiaxi joined the Food and Agriculture Organization of the United Nations as a gender Intern to improve women's access to agricultural resources with a B.S. in Logistic Management.

Yue Hu
Columbia University
M.S. in Operations Research
Software: Azure Notebooks
Email: yh3218@columbia.edu

With a B.S. in Insurance from Nanjing University, Yue has experience in the CICC, the top investment bank in China, where she leveraged quantitative and computational methods to asset management work.

Zihui Zhou
Columbia University
M.S. in Operations Research
Software: CoCalc
Email: zz2694@columbia.edu

Zihui completed her double majors in Mathematics and Studio Art at Boston College. She is a data enthusiast, who applied data science knowledge while interning at BrainCo, a biotechnology company, and Intellipro Group, targeting companies including Waymo, ByteDance, Instagram, etc.



Xinyao Wang
Columbia University, M.S. in Operations Research
Software: Colab
Email: xw2675@columbia.edu

With a BS in Mathematics from WIlliam & Mary, Xinyao has intern experience in equity investment department at ICBC, a $340B bank, and a data analyst the Port Authority, which controls flight operations including JFK, EWR and LGA airports.

Qinwei Zhang
Columbia University, M.S. in Operations Research
Software: Kaggle Kernel
Email: qz2391@columbia.edu

Qinwei holds a Bachelor degree in Engineering Mechanics from University of Illinois. She works on research projects with Prof. Yuri Faenza while applying programming skills for Terra, a digital media startup.

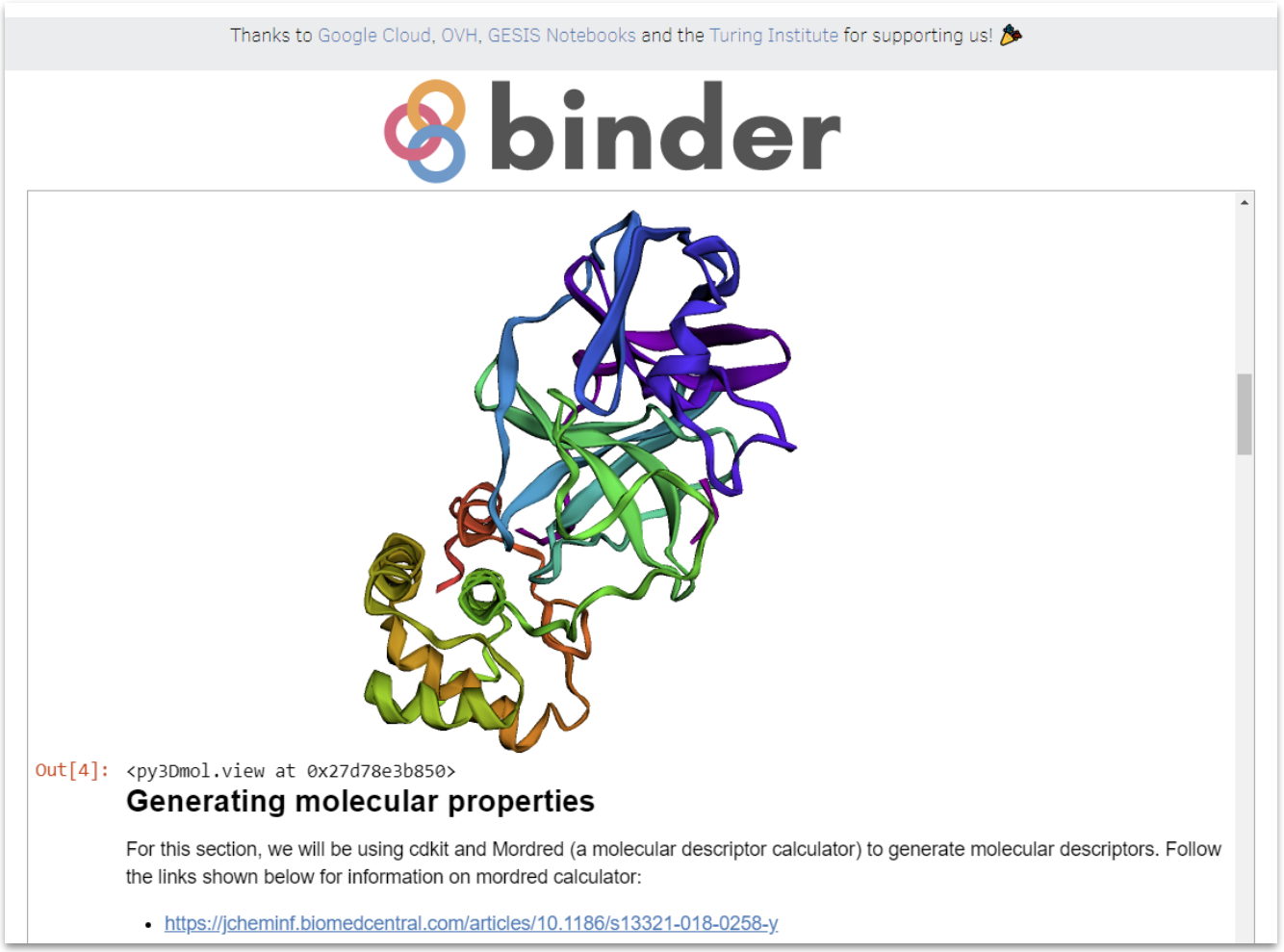
Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Soder
TA, Project Manager
Columbia University

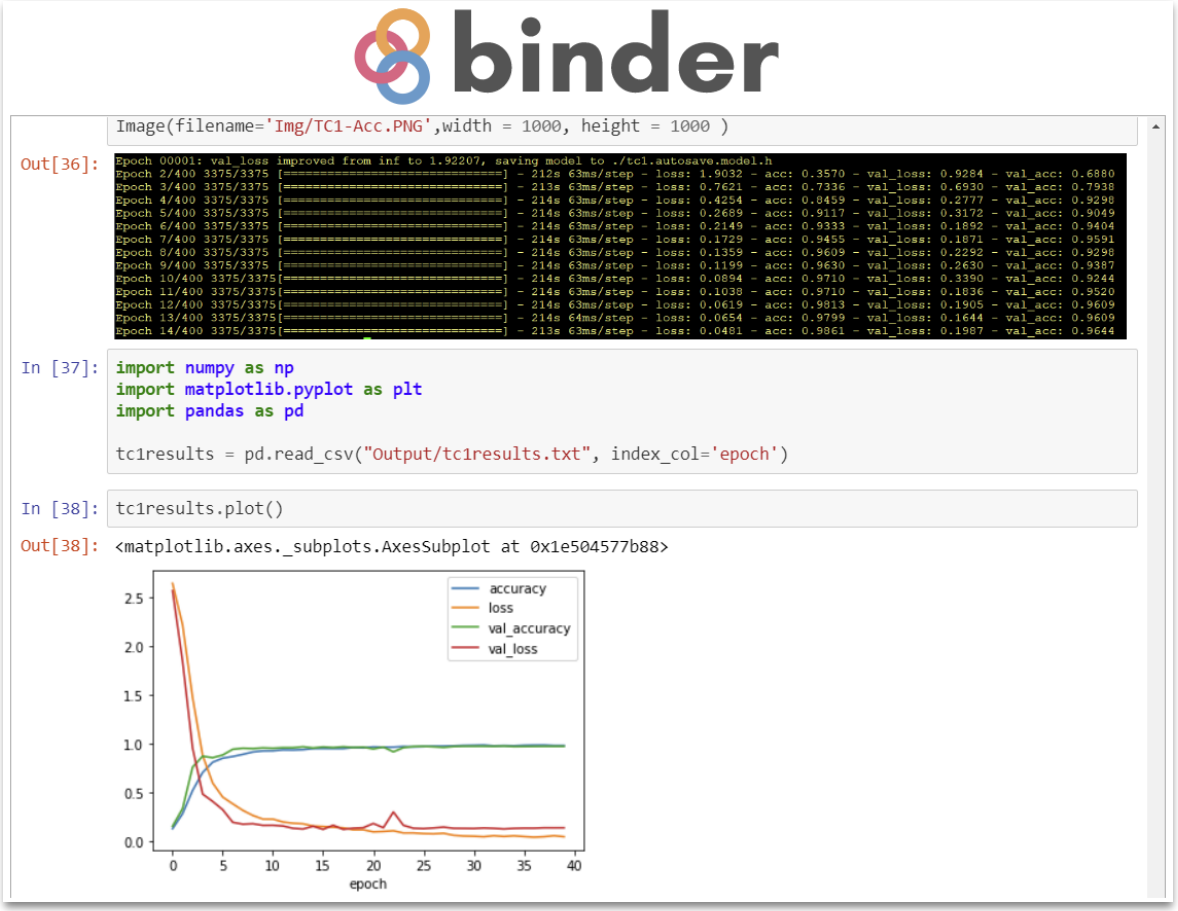
Project Goals



- **Compare between different software and recommend the best software**
- **Use to promote reproducible work**
- **Create live demonstration and share interactive live notebook for workshop/training**

How we tested the cloud computing software?

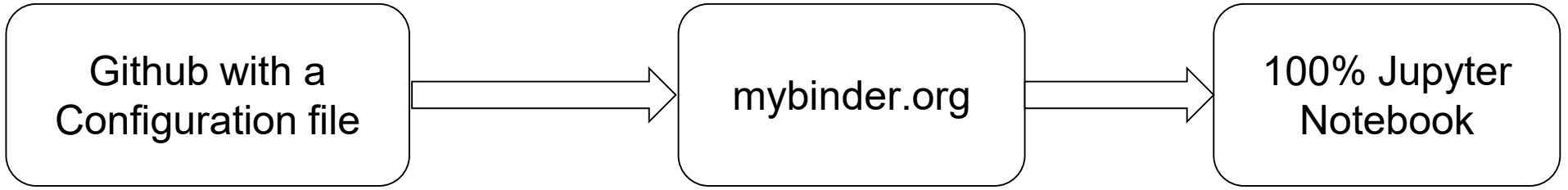
- **Test Codes**
  - Generating Molecular Features for Drug Function Classification: https://github.com/ravichas/ML-predict-drugclass
  - Cancer Site/Type classification using Convolutional Neural Network (**NCI-DOE collaborative project**) TC1: https://github.com/ravichas/ML-TC1
- **Criteria**
  - User friendliness
  - Configurations (CPU, RAM, disk space, etc.)
  - Customizability
  - GitHub compatibility
  - Supporting languages and file formats



Binder (https://mybinder.org/)

Jiaxi Zhou

Introduction



Supported languages: Python (2 and 3), R, Julia

Charge: No

Shareable: Yes

Account : No

Traffic

- 100 (claimed)
- Around 130 (self-test)

Number of visits sending	Launch Successfully	Still Loading after 10 min	Too Many Users ERROR
200	43	125	32
150	16	95	39
150	52	86	11
140	50	89	0
140	0	87	53
135	54	78	3
130	53	77	0
125	28	97	0
100	41	59	0

Too Many Users ERROR

Error loading ravichas/ML-predict-drugclass/master!
See logs below for details.

The Binder team has a site reliability guide that talks about what it is like to run a BinderHub.

Build logs hide

```
Found built image, launching...
Too many users running https://github.com/ravichas/ML-predict-drugclass! Try again soon.
```

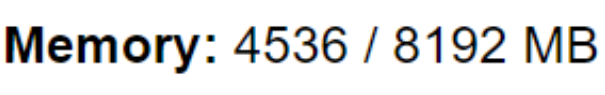
Memory

- 1-2 GB (claimed)
- 2 GB (self-test)

The warning will be given above 1800MB

The kernel will collapse when ram is above 2 GB immediately

Solution: Create Binder repository by GESIS Notebooks beta



Optimized Configuration File

https://en.wikipedia.org/wiki/YAML

For file Pedict_Drug_Class:

Reduced **16** unnecessary packages

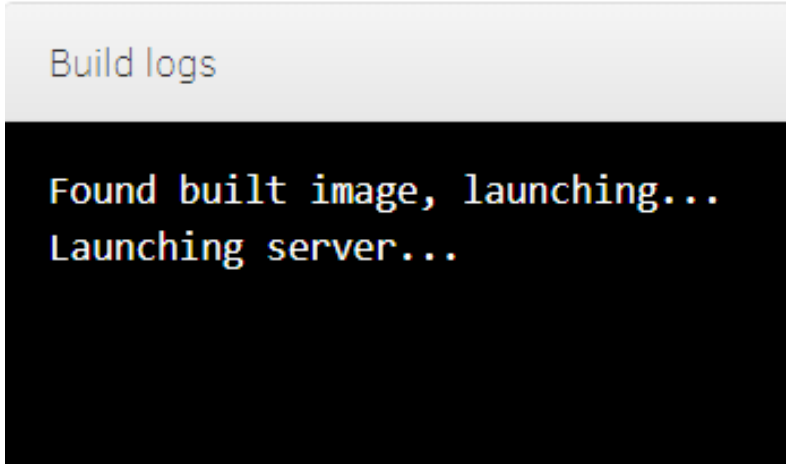
https://mybinder.org/v2/gh/Jiaxi-Zhou/test3/master

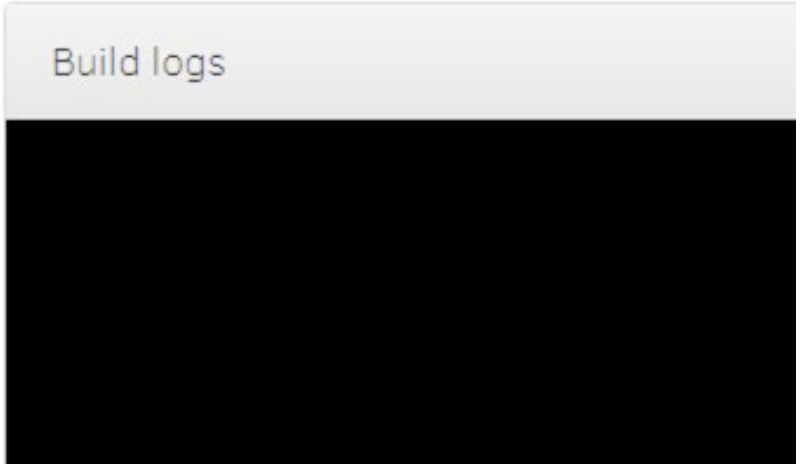

```
name: tutorial
channels:
 - anaconda
dependencies:
 - python
 - numpy
 - pip
 - matplotlib
 - ipython=7.10.0
 - mordred
 - numpy
 - pandas
 - rdkit=2019.09.2
 - jupyterlab
 - py3Dmol
 - pip:
   - scikit-learn==0.22
   - ipymol
```

Launching Time

Startup: 20min

Link Launching Time	Original YML	Optimized YML
ML_drug_class	25s-180s	10s-150s
ML_TC1	30s-240s	/





Conclusion & Suggestion

- **Pros**
  - Free
  - Easy to connect with github
  - Convenient to use
- **Cons**
  - Unstable
  - Small RAM

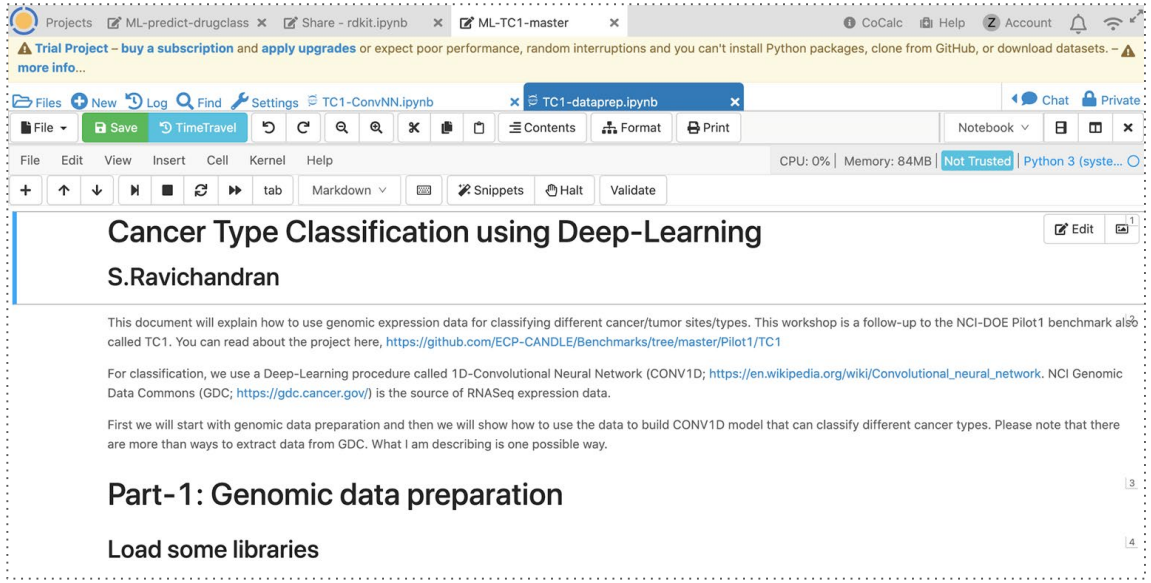
- **Good for**
  - Workshop
  - Training
  - Coding lessons
- **Less appropriate for**
  - Formal business meeting
  - Formal Presentation

CoCalc (cocalc.com)

Zihui Zhou

CoCalc (free plan) Introduction

- **User friendliness**
  - Easy to deploy (quick setup)
  - Excellent version control
  - Real-time collaboration: Yes
  - Shareable: Yes (add collaborators)

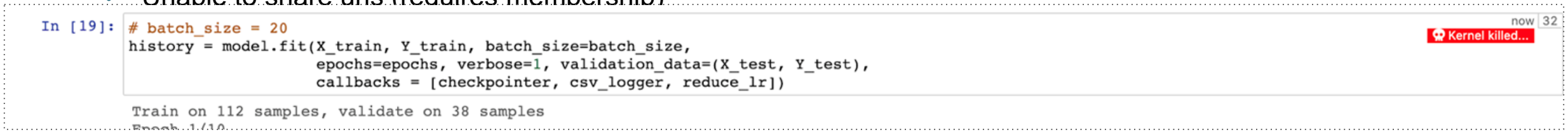


CoCalc (free plan) Introduction

- **User friendliness (continue)**
  - Configurations
    - 1-core shared CPU
    - 1 GB of shared memory
    - 3 GB of disk space
  - Incompatible with GitHub (upload files manually)
  - Internet access: No
  - Packages: require membership and installation requests
- **Supporting languages:** over 10 languages, including Python, C++, R
- **Customizability (& ease) of the configuration file:** ipynb, txt, html, md, rst, tex, etc.

CoCalc (free plan) Summary

- **Pros**
  - 95% similarity to Jupyter Notebook (easy to use)
  - provides version control, "time travel", with excellent functionality
  - easy to deploy (quick to set up)
- **Cons**
  - incompatibility with GitHub
  - installing additional packages (rdkit, py3Dmol, mordred) requires membership and request submissions
  - limited shared 1GB memory and 3GB disk space
  - Unable to share urls (requires membership)



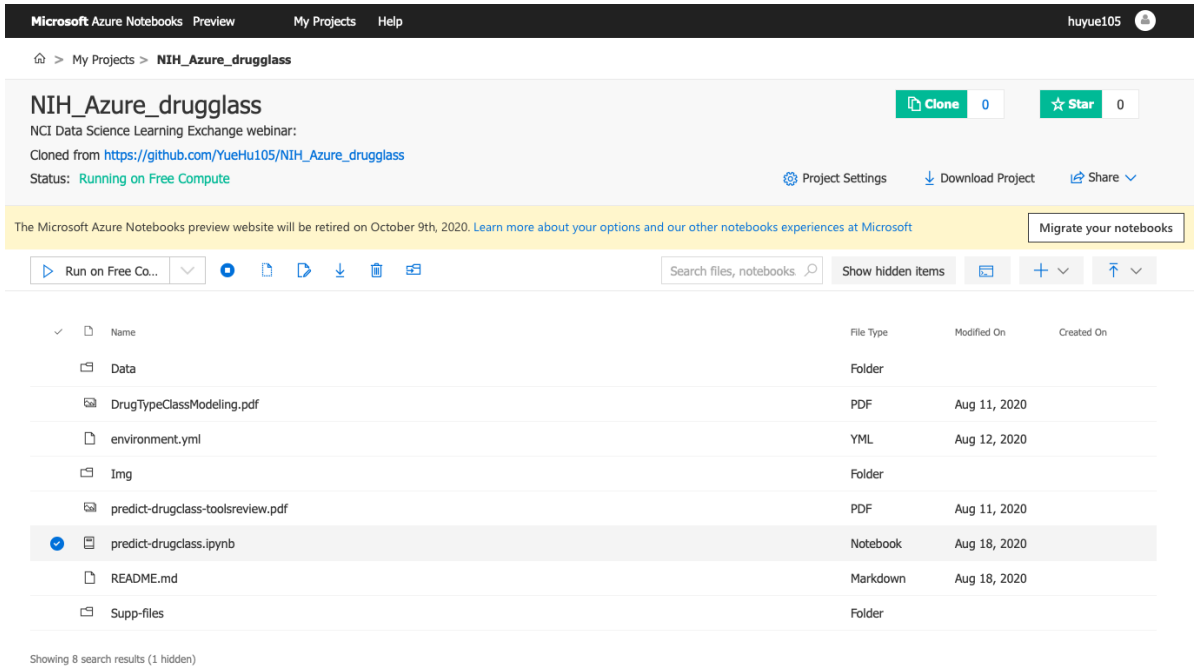
Microsoft Azure Notebooks

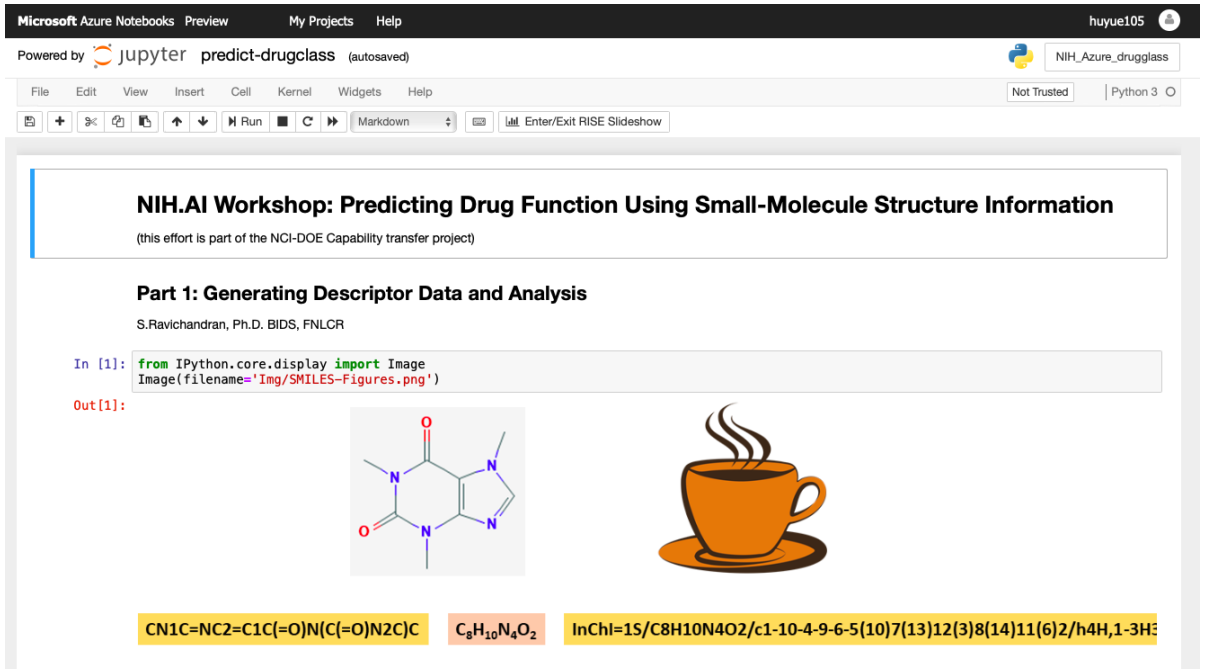
Azure Notebook (https://azure.microsoft.com/)

Yue Hu

GitHub + Jupyter Notebook Interface

Microsoft Azure Notebooks





Brief Introduction

Microsoft Azure Notebooks

- **Free for use**
- **Available languages:**

  Python (2.7, 3.5, 3.6), R, F
- **GitHub support:**

  clone the entire GitHub repository
- **Pre-configured Jupyter extensions/pre-installed packages**
- **executive slide-like codes**
- **Future development:**

  On October 9, 2020 the Azure Notebooks public preview site will be retired and replaced with integrated services from Visual Studio, Azure, and GitHub.

Cons

Microsoft Azure Notebooks

- No version control system
- Extremely Time-consuming for installing extra packages

```
!conda install rdkit -y
```

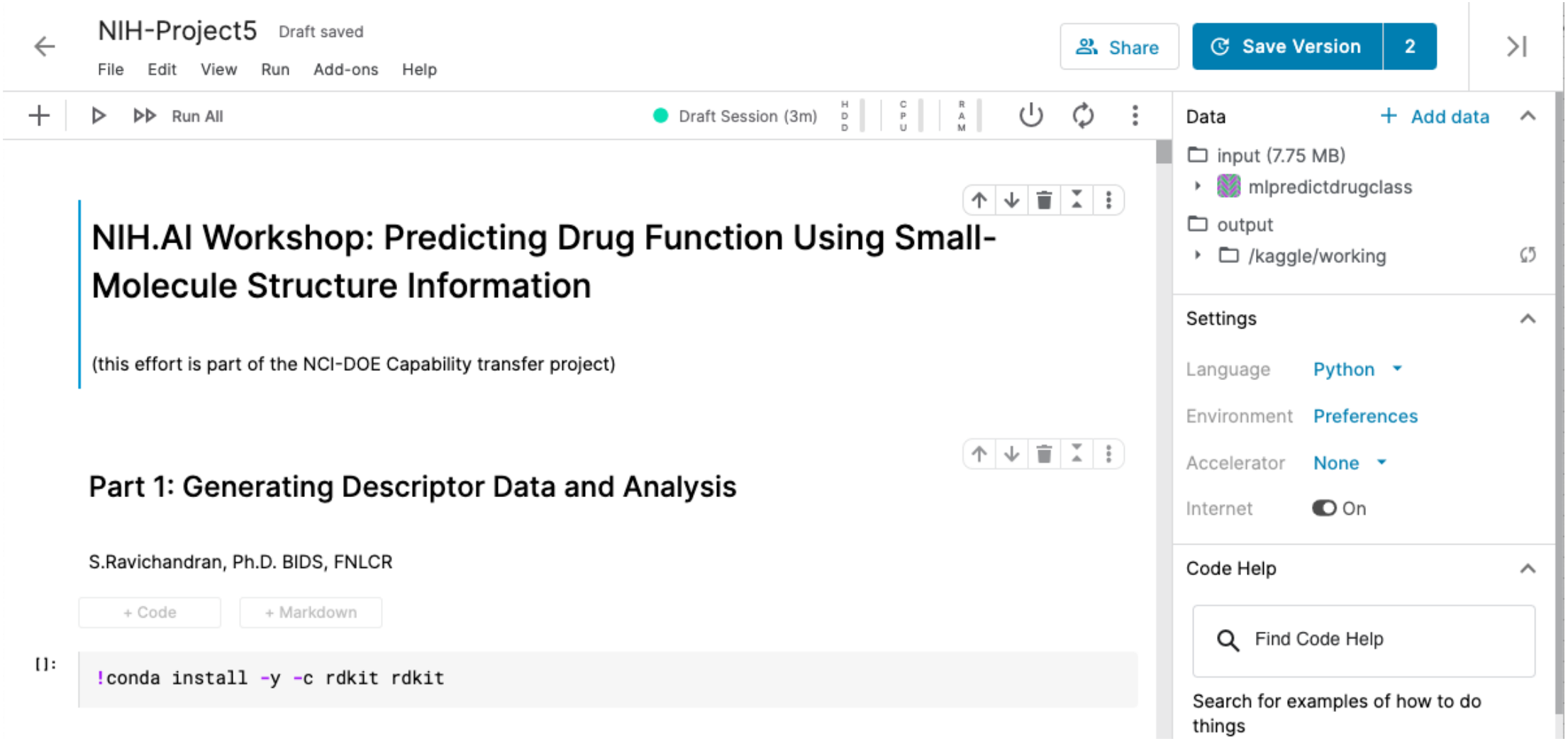
```
Fetching package metadata .............
Solving package specifications: .

Package plan for installation in environment /home/nbuser/anaconda3_420:
```

- Unclear future development

Kaggle Notebook (https://www.kaggle.com/notebooks)

Qinwei Zhang

User Interface



Brief Introduction

- Available languages:
  - Python 3 & R
- Configurations:
  - 4 CPU cores and 16GB of RAM. 9 hours execution time. 5GB of auto-saved disk space
- GitHub Support:
  - Whole repository can be imported as a Dataset. Kernel can access the Dataset. But for jupyter notebook, it cannot be opened directly and needs to create a new notebook first and import through URL or local drive.
- Custom Packages:
  - Hundreds of packages pre-installed. Additional packages can be added through pip, conda or by specifying the GitHub repository of a package, i.e. rdkit, py3Dmol, mordred in ML-predict-drugclass notebook. But they need to be reinstalled at the start of every session.

Pros and Cons

Pros:

- No installation.
- Instantly deploy and shareable. Can invite other Kaggle users to collaborate.
- Free access to computational resources, including sufficient GPU and disk space, with no credit card required. Can link to Google Cloud account for more compute power.

Cons:

- Need to sign up a Kaggle account ahead of time to edit the notebook. No real-time collaboration.
- Not able to set up your own environment. Custom packages need to be re-installed every session.
- Need to revise the file path.

Colab (https://colab.research.google.com/)

Xinyao Wang

Google Colaboratory Introduction

Google Colab is a jupyter notebook environment. It is free source provided by google wherein we can write and execute code. We can use Google Colab with ease just as we use local jupyter.

1. Packages: Many packages are already installed for users. You can check using !pip freeze command. You can install any packages you want using !pip install command.(! is needed before pip)
2. Configuration/traffic:
    i. Default 13GB of RAM with maximum extension of 25GB
    ii. Disk Space: free 100 GB
    iii. 2 vCPU @2.2GHz
    iv. Idle cut-off 90 minutes
    v. Maximum run time of 12 hours
3. Real-time Collaboration: Yes.
4. Shareable: Yes.
5. Keep as a private / local file: Yes
6. Internet access: Yes
7. Supporting Languages: over 40 programming languages including Python, R, Julia, Scala, etc.

Google Colaboratory Introduction

Google Colab is a jupyter notebook environment. It is free source provided by google wherein we can write and execute code. We can use Google Colab with ease just as we use local jupyter.

1. Packages: Many packages are already installed for users. You can check using !pip freeze command. You can install any packages you want using !pip install command.(! is needed before pip)
2. Configuration/traffic:
    i. Default 13GB of RAM with maximum extension of 25GB
    ii. Disk Space: free 100 GB
    iii. 2 vCPU @2.2GHz
    iv. Idle cut-off 90 minutes
    v. Maximum run time of 12 hours
3. Real-time Collaboration: Yes.
4. Shareable: Yes.
5. Keep as a private / local file: Yes
6. Internet access: Yes
7. Supporting Languages: over 40 programming languages including Python, R, Julia, Scala, etc.

Google Colaboratory Pros & Cons

Pros:

- Directly show the output for Github notebook
- Easy to link with Github
- High speed for installing packages and running codes
- Large RAM and disk space
- Reader could run and edit the code directly
- Free
- Many resources and tutorials online

Cons:

- Need to use personal Gmail rather than Edu Gmail
- Need to refresh and save file often when multiple people working on same Colab file

Google Colaboratory Pros & Cons

Pros:

- Directly show the output for Github notebook
- Easy to link with Github
- High speed for installing packages and running codes
- Large RAM and disk space
- Reader could run and edit the code directly
- Free
- Many resources and tutorials online

Cons:

- Need to use personal Gmail rather than Edu Gmail
- Need to refresh and save file often when multiple people working on same Colab file

Cloud Computing Software Ranking by Popularity

1. Google Colab
2. Kaggle Kernel
3. Binder
4. Azure Notebooks
5. CoCalc

Conclusion: Recommend Choosing Colab

	Download packages speed	Memory	Disk Space	Linked with Github	Capacity
Colab	1 min	12GB~25GB	100 GB	Easy	Unlimited
Kaggle	2-3 mins	16 GB	5 GB	Repository import as a Dataset, Notebook through URL	Uneditable public link: Unlimited
Azure Notebooks	20 mins per package	4 GB (with 1GB stored dataset)	1 GB	Easy (clone the whole repo)	N/A
Binder	20 mins	2 GB	> 2 GB	Easy（clone the whole repo）	130-135
CoCalc	Unable to download packages	1 GB	3 GB	incompatible	Urls not shareable

THANK YOU!

Thank you for this opportunity, especially to Ravi and Naomi. We learned about FNL and about collaborating to solve real machine learning problems in medical science. We've learned a great deal and hope to be able to work with FNL again.

We have identified ***Colab*** as the dominant technology for critical presentations that combine machine learning with medical data.

And, we encourage FNL to begin plans for deploying presentations using a hospital-wide standard for ***efficiency*** and to leverage ***cooperation***.

Any questions?