

# The MAterials Simulation Toolkit for Machine Learning (MAST-ML): Automating Development and Evaluation of Machine Learning Models for Materials Property Prediction

**Ryan Jacobs, Tam Mayeshiba, Ben Afflerbach, Dane Morgan**  
*(University of Wisconsin – Madison, WI USA)*

**Luke Miles, Max Williams, Matthew Turner, Raphael Finkel**  
*(University of Kentucky, Lexington, KY USA)*

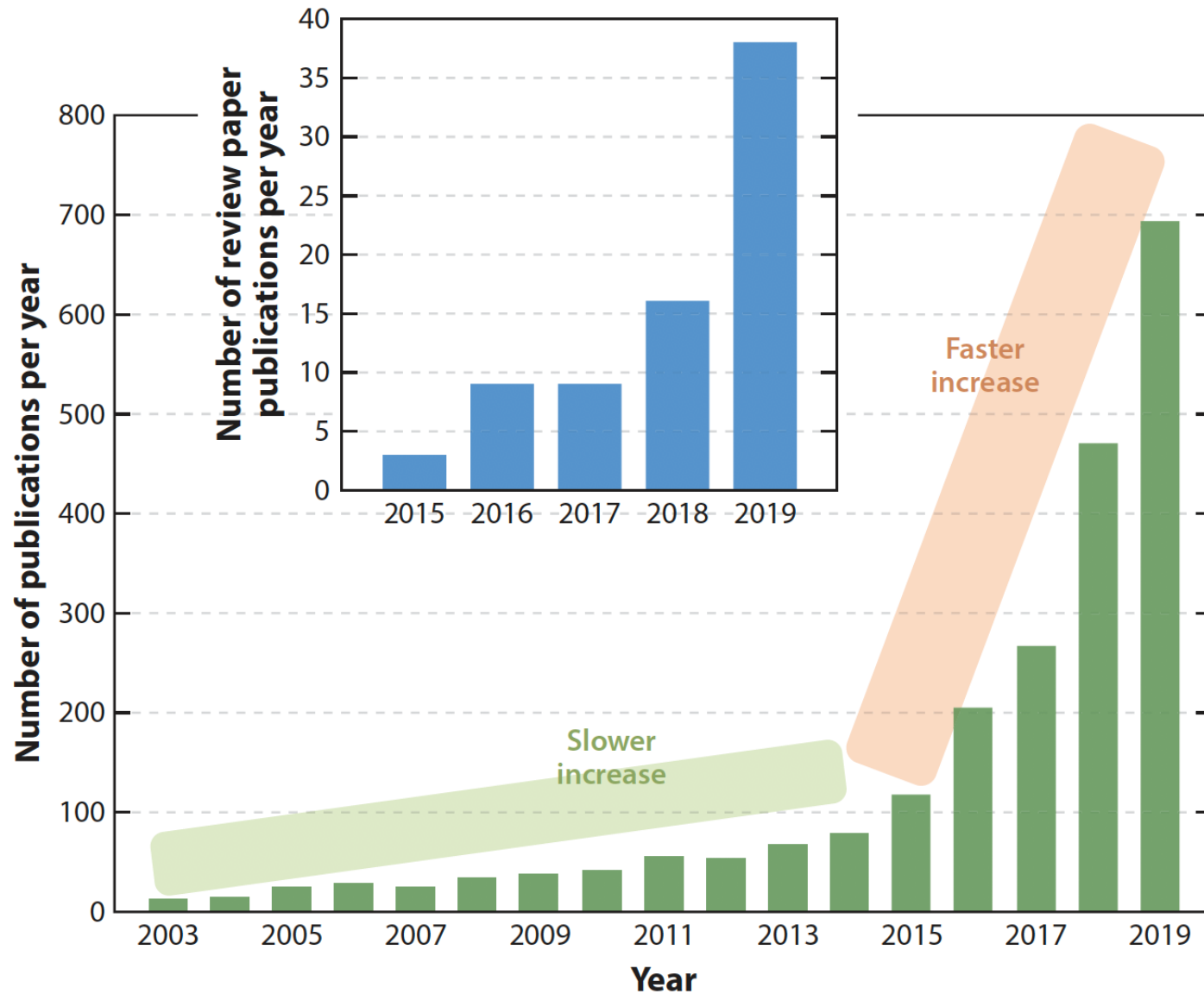
*Most Recent Skunkworks MASTML members:*  
**Avery Chan, Hock Lye Lee, Min Yi Lin**

<https://github.com/uw-cmg/MAST-ML>

**NanoHub ML Workshop**  
**5/19/2021**

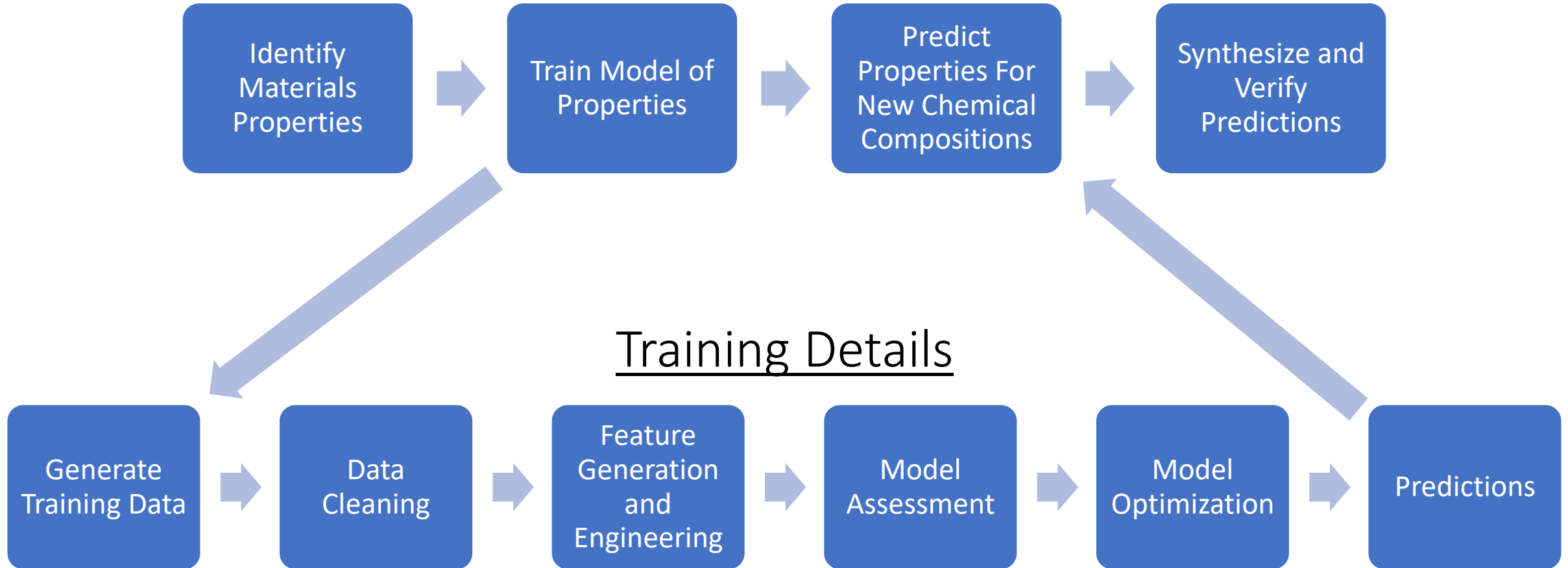


# Machine learning in Materials Science is Exploding



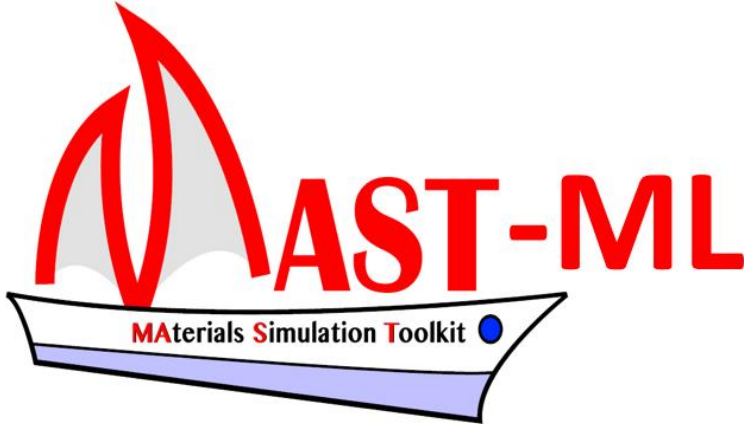
# A Basic Materials Design Workflow

---



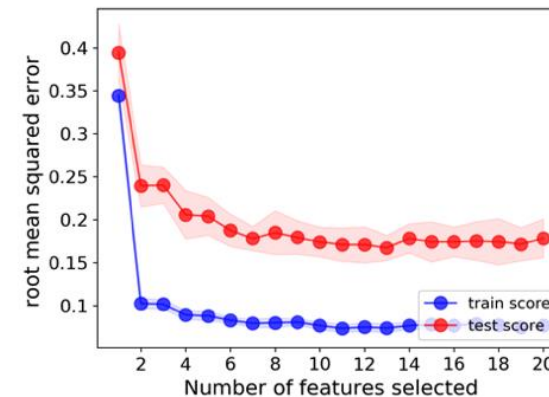
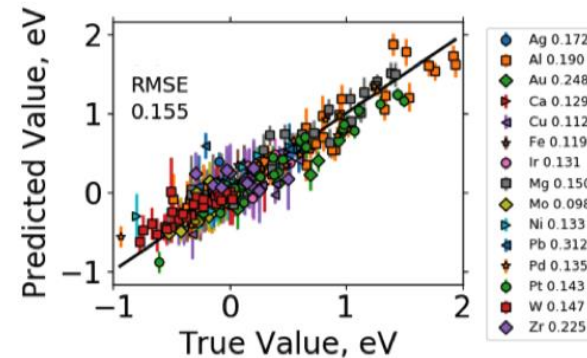
# What is MAST-ML?

MAST-ML is an open-source Python package designed to broaden and accelerate the use of machine learning in materials science research, particularly for non-experts.

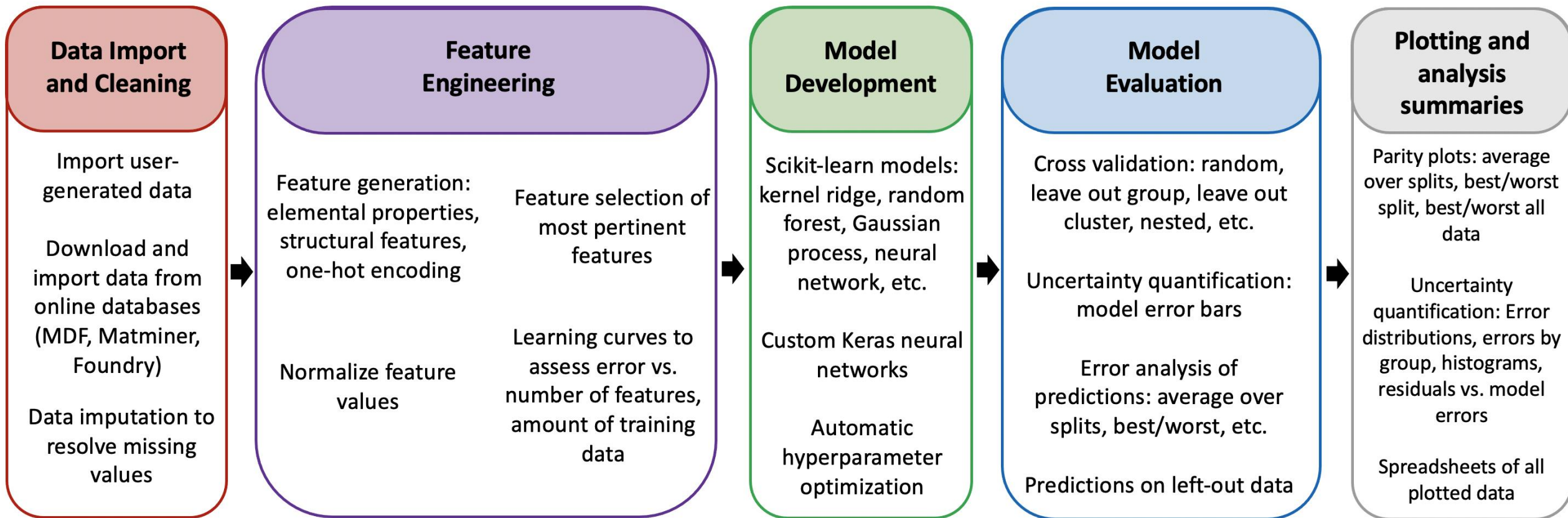


Automated machine learning tools for materials informatics research (MAST-ML)

<https://github.com/uw-cmg/MAST-ML>

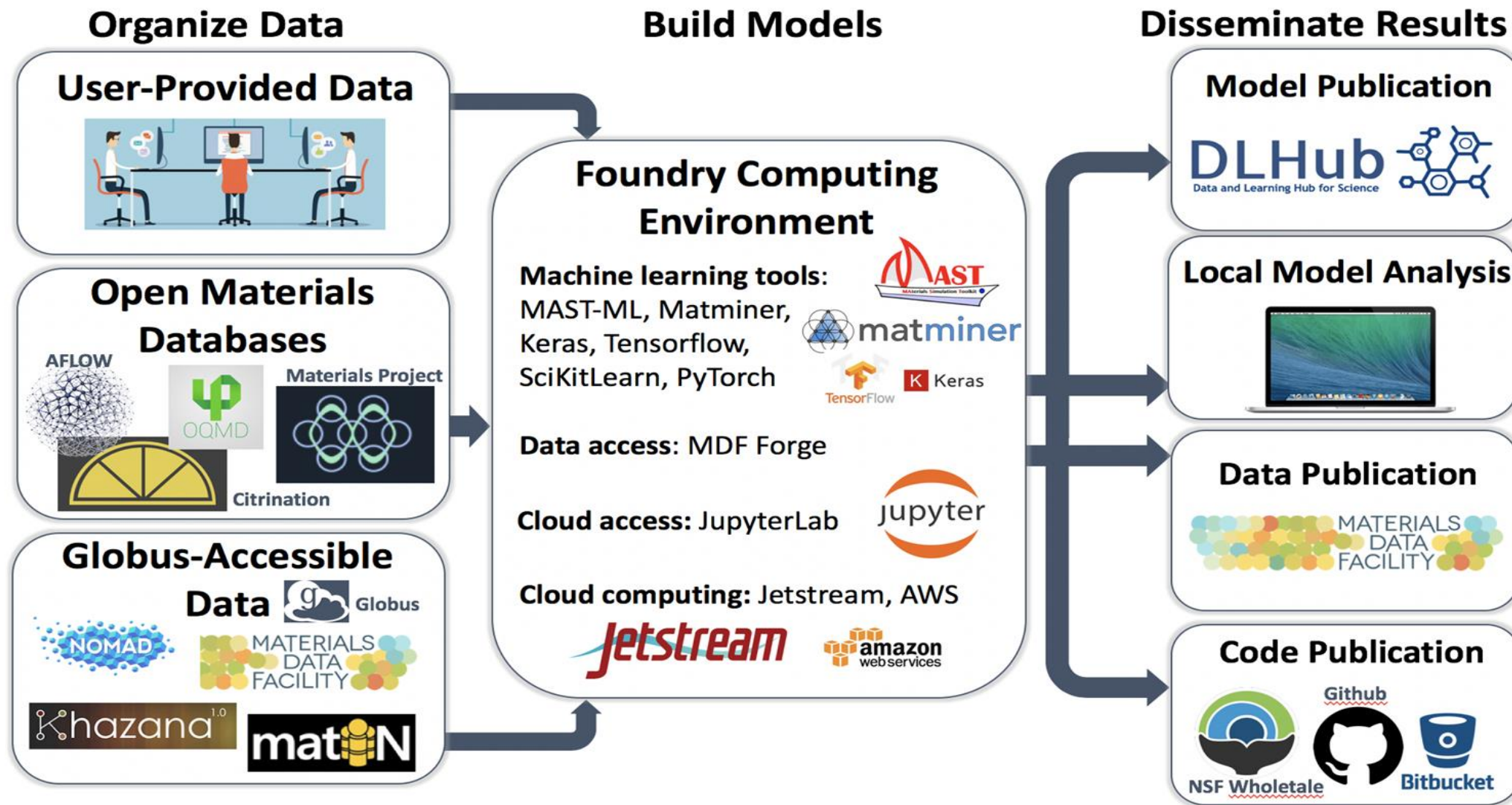


# MAST-ML automates the supervised learning workflow



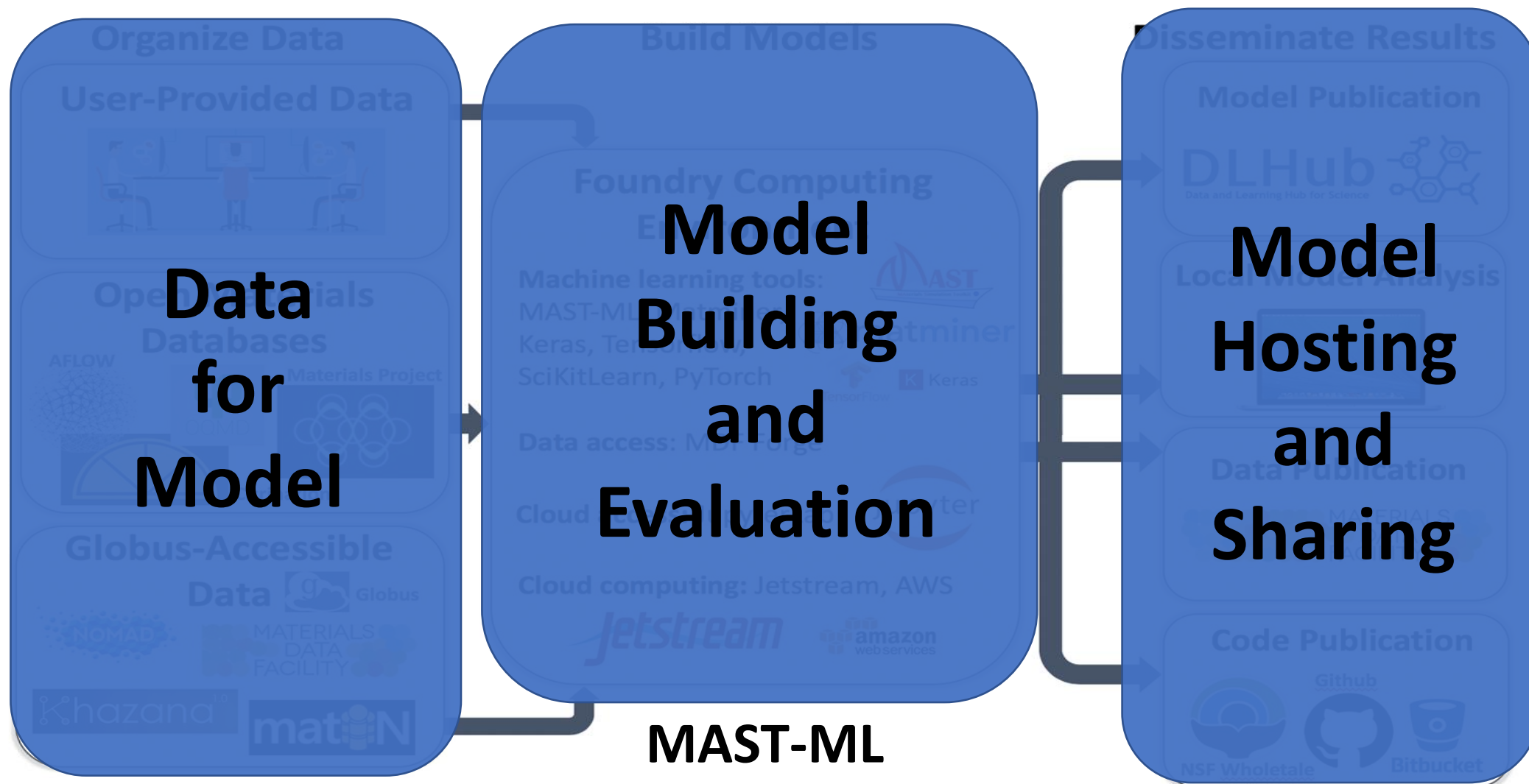
- MAST-ML supports the full library of scikit-learn modules, and can be used to construct neural networks with Keras (based on tensorflow)
- MAST-ML allows for the simultaneous execution of an arbitrary combination of data preprocessing, feature generation/selection, model types and model evaluation metrics

# (NSF CSSI) Machine Learning Materials Innovation Infrastructure



(PIs Dane Morgan, Paul Voyles, Michael Ferris, Ryan Jacobs, Ben Blaiszik)

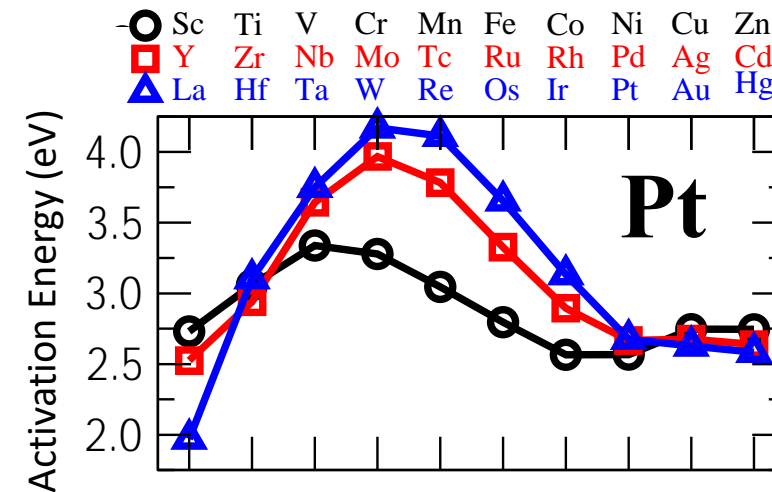
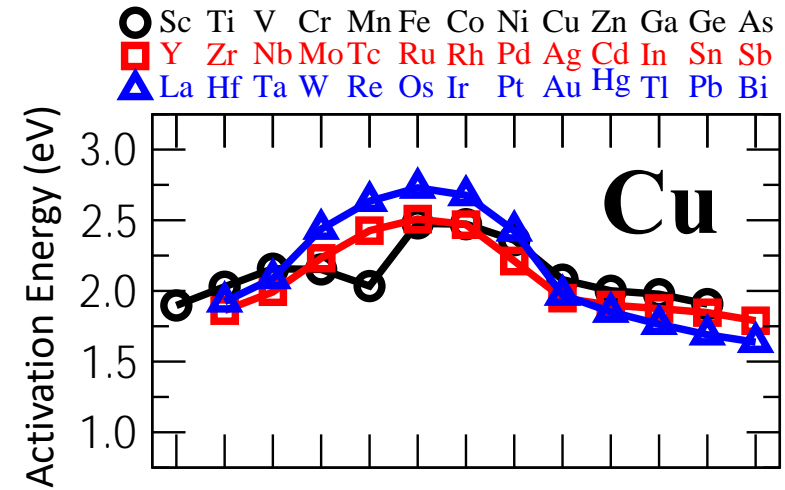
# (NSF CSSI) Machine Learning Materials Innovation Infrastructure



Model building, evaluation, and key connections  
between data and model dissemination

# Test Problem: Impurity Diffusion Database

- Diffusion of dilute impurity X in host H. We have DFT calculations of 440 values, but want ~4,000. [1, 2]
- Assume Y= Activation energies measured relative to host, X= Host descriptors, Impurity descriptors. Find  $Y=F(X)$ .
- Descriptors = elemental properties like melting temperature, bulk modulus, electronegativity, ... and their ratios, differences, etc. (MAGPIE set)[3]
- F is determined using standard machine learning regression methods (e.g., Gaussian Process Regression (Gaussian Kernel) (GPR), Random Forest (RF), neural network).
- Fit F with calculated data (15 hosts, 440 M-X pairs)

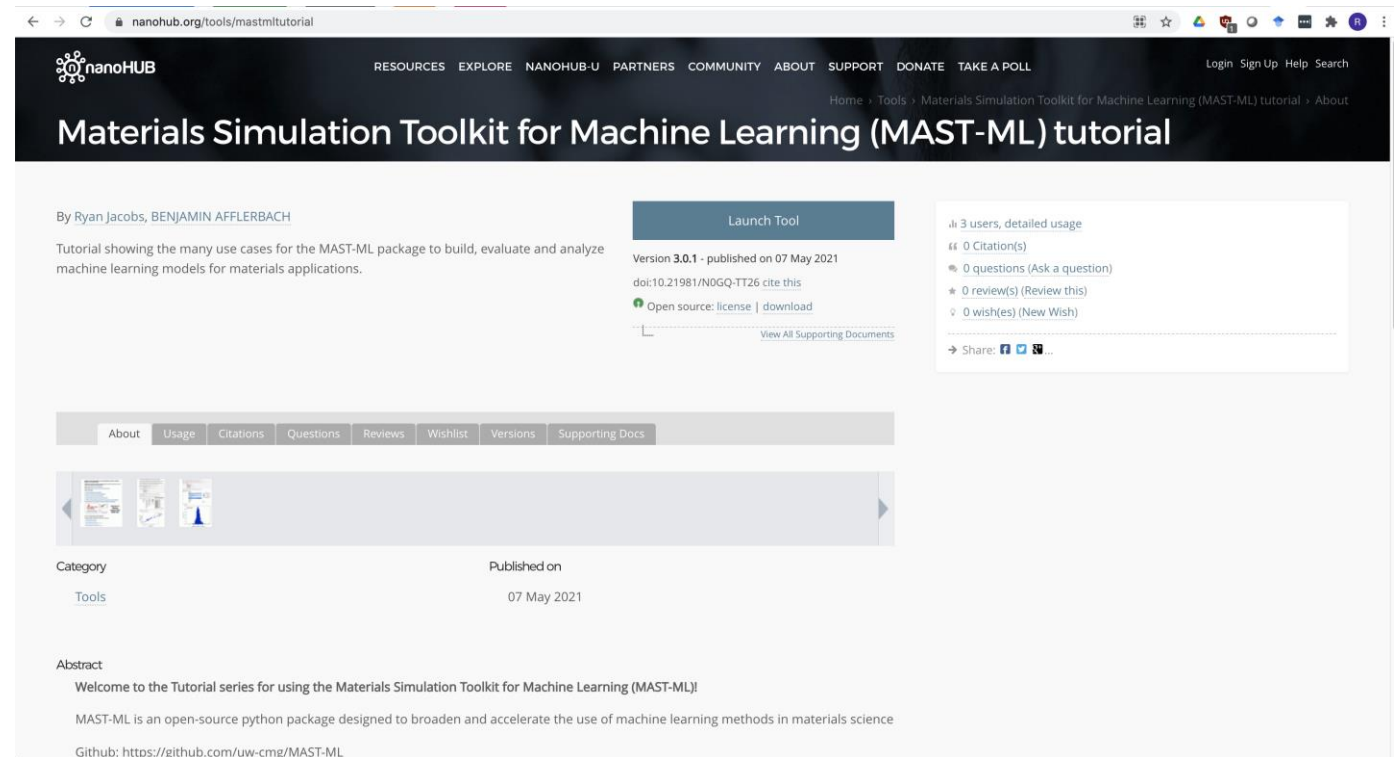


<http://diffusiondata.materialshub.org/>



# Getting Started with the MAST-ML tutorial on NanoHub

- Link to Tool:  
<https://nanohub.org/tools/mastmltutorial>
- Select “Launch Tool”
- A Jupyter notebook environment will open (may take a minute)
- Click on cell and run with Shift+return
- Data will be saved to local directory, see next slides for how to download results



The screenshot displays the NanoHub website interface for the MAST-ML tutorial. The page title is "Materials Simulation Toolkit for Machine Learning (MAST-ML) tutorial". It is authored by Ryan Jacobs and Benjamin Afflerbach. A prominent "Launch Tool" button is visible. The page includes a description of the tutorial, version information (3.0.1, published 07 May 2021), and options to open source, license, or download. A sidebar on the right shows user statistics and social sharing options. Below the main content, there are tabs for "About", "Usage", "Citations", "Questions", "Reviews", "Wishlist", "Versions", and "Supporting Docs". The "Usage" tab is currently selected, showing a category of "Tools" and a published date of "07 May 2021". An abstract section provides a brief overview of the MAST-ML package and its purpose in materials science.

# Downloading results from MAST-ML tutorial

---

**Each numbered step has a screenshot on the next few slides**

Step 1.) Can visualize and zip saved files by clicking File-> Open in Jupyter notebook. A new window will open displaying your folders

Step 2.) To download a saved folder, need to zip it first. On right-hand side of window displaying folders, go to New->Terminal. A new window will open displaying a Unix-style Terminal.

Step 3.) Zip the folder you want to download with the command:

- `tar -zcvf folder_name.tar.gz folder_name`

Step 4.) Look back at your displayed folders: there is now a new .tar.gz folder. You can click the box next to this folder and click the “Download” button.

# Downloading results from MAST-ML tutorial: Step 1

Step 1.) Can visualize and zip saved files by clicking File-> Open in Jupyter notebook. A new window will open displaying your folders

proxy.nanohub.org/weber/1836304/1t5dFNuiSP5c5R8/1/notebooks/mastmltutorial.ipynb?

nanoHUB jupyter mastmltutorial (read only)

File Edit View Insert Cell Kernel Widgets Help Snippets

New Notebook

Open...

Preferences

Share Session

Make a Copy...

Save as...

Rename...

Save and Checkpoint

Revert to Checkpoint

Print Preview

Download as

Trust Notebook

Close and Halt

come to the Tutorial series for using the Materials Simulation Toolkit for Machine Learning (MAST-ML)!

MAST-ML is an open-source python package designed to broaden and accelerate the use of machine learning methods in materials science

GitHub: <https://github.com/uw-cmg/MAST-ML>

Citation: <https://doi.org/10.1016/j.commatsci.2020.109544>

Table of Contents

- [Tutorial 1: Getting Started with MAST-ML](#)
- [Tutorial 2: Data Import and Cleaning with MAST-ML](#)
- [Tutorial 3: Feature Engineering with MAST-ML](#)
- [Tutorial 4: Models and Data Splitting Tests with MAST-ML](#)
- [Tutorial 5: Left out data, nested cross validation, and optimized models with MAST-ML](#)
- [Tutorial 6: Model error analysis and uncertainty quantification with MAST-ML](#)

**MAST-ML**  
Materials Simulation Toolkit

Automated machine learning tools for

Predicted Value, eV

True Value, eV

RMSE 0.155

an squared error

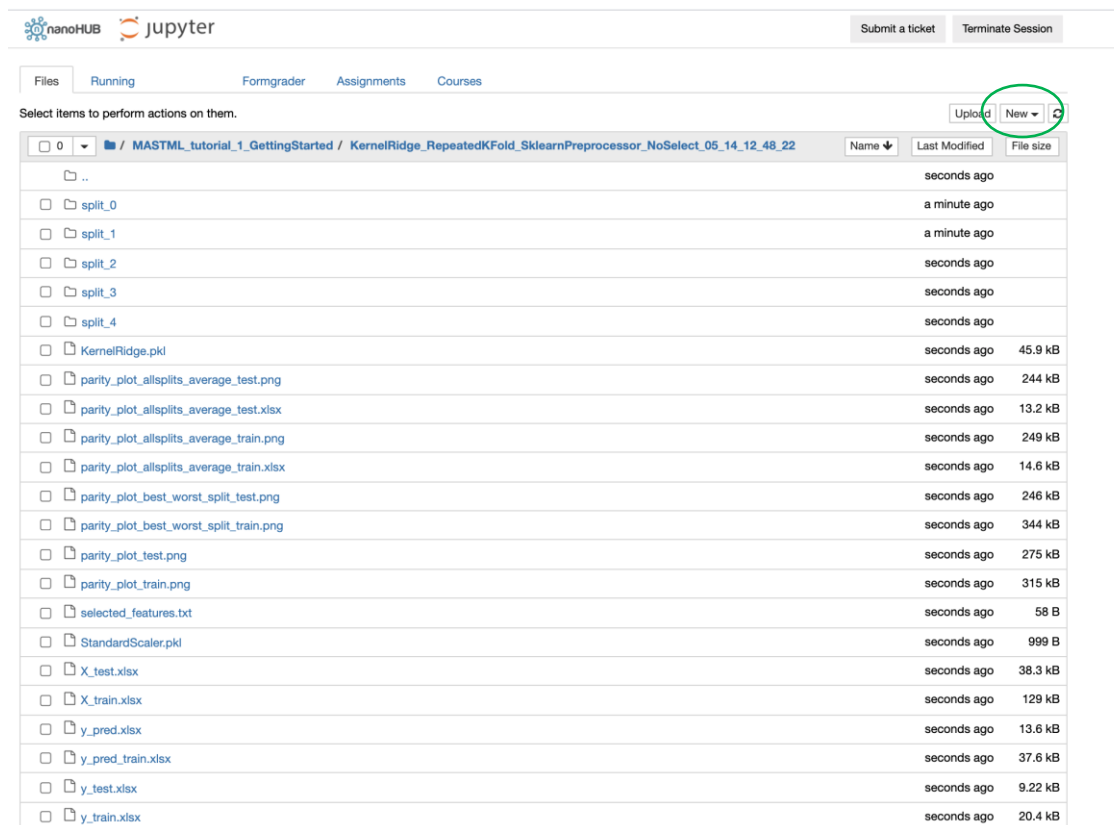
**Tutorial # 1:  
Getting  
Started with**

Ag 0.172  
Al 0.190  
Au 0.248  
Ca 0.129  
Cu 0.112  
Fe 0.139  
H 0.131  
Mg 0.150  
Mo 0.098  
Ni 0.133  
Pb 0.312  
Pd 0.135  
Pt 0.143  
W 0.147  
Zr 0.225

<https://proxy.nanohub.org/weber/1836304/1t5dFNuiSP5c5R8/1/notebooks/mastmltutorial.ipynb?#>

# Downloading results from MAST-ML tutorial: Step 2

Step 2.) To download a saved folder, need to zip it first. On right-hand side of window displaying folders, go to New->Terminal. A new window will open displaying a Unix-style Terminal.



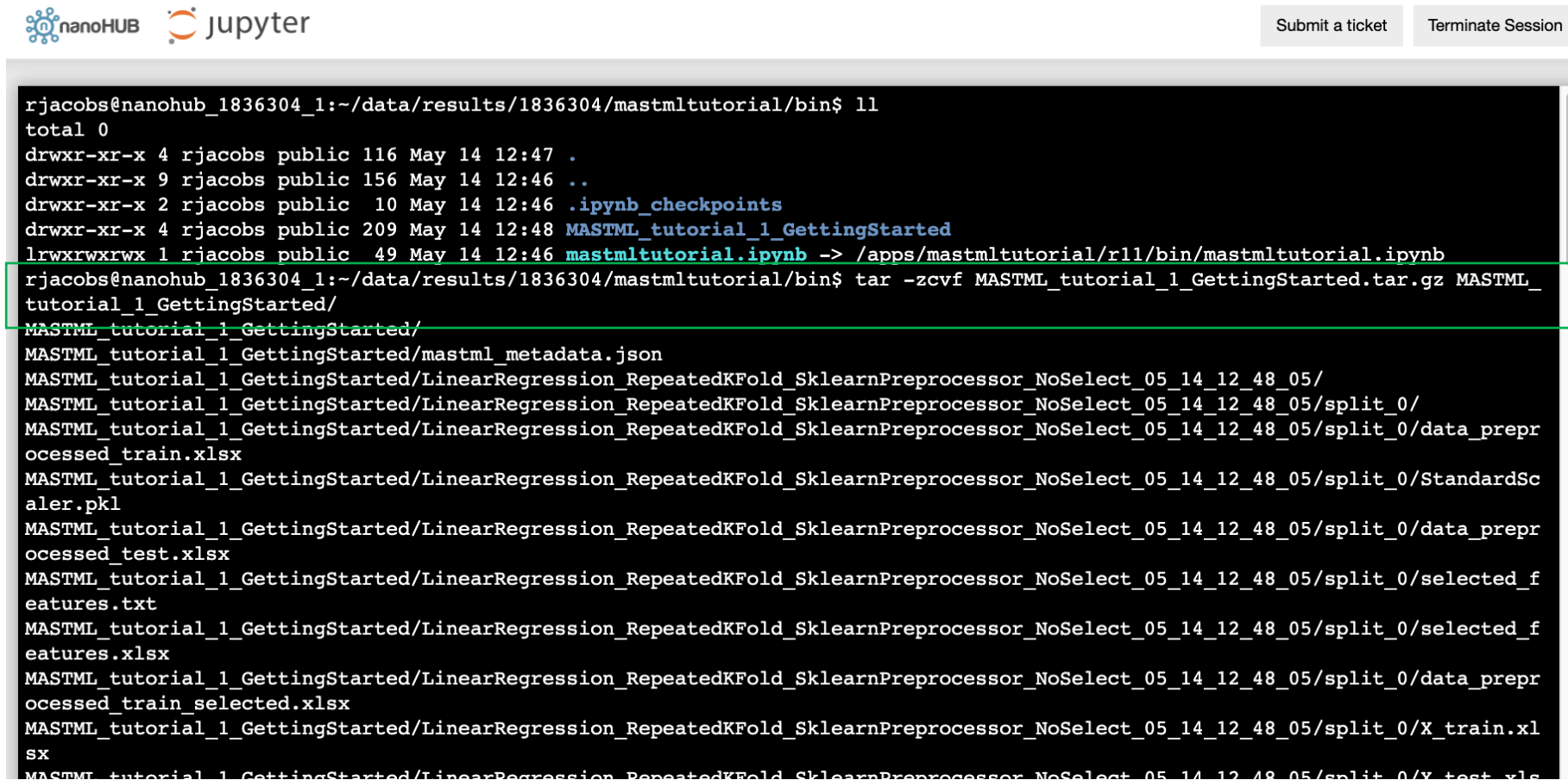
The screenshot shows the nanoHUB Jupyter interface. At the top, there are logos for nanoHUB and Jupyter, along with buttons for 'Submit a ticket' and 'Terminate Session'. Below the logos, there are tabs for 'Files', 'Running', 'Formgrader', 'Assignments', and 'Courses'. The main area displays a file browser for the path '/ MASTML\_tutorial\_1\_GettingStarted / KernelRidge\_RepeatedKFold\_SklearnPreprocessor\_NoSelect\_05\_14\_12\_48\_22'. A 'New' button is circled in green. Below the file browser, there is a table of files and folders.

Name	Last Modified	File size
..	seconds ago	
split_0	a minute ago	
split_1	a minute ago	
split_2	seconds ago	
split_3	seconds ago	
split_4	seconds ago	
KernelRidge.pkl	seconds ago	45.9 kB
parity_plot_allplits_average_test.png	seconds ago	244 kB
parity_plot_allplits_average_test.xlsx	seconds ago	13.2 kB
parity_plot_allplits_average_train.png	seconds ago	249 kB
parity_plot_allplits_average_train.xlsx	seconds ago	14.6 kB
parity_plot_best_worst_split_test.png	seconds ago	246 kB
parity_plot_best_worst_split_train.png	seconds ago	344 kB
parity_plot_test.png	seconds ago	275 kB
parity_plot_train.png	seconds ago	315 kB
selected_features.txt	seconds ago	58 B
StandardScaler.pkl	seconds ago	999 B
X_test.xlsx	seconds ago	38.3 kB
X_train.xlsx	seconds ago	129 kB
y_pred.xlsx	seconds ago	13.6 kB
y_pred_train.xlsx	seconds ago	37.6 kB
y_test.xlsx	seconds ago	9.22 kB
y_train.xlsx	seconds ago	20.4 kB

# Downloading results from MAST-ML tutorial: Step 3

Step 3.) Zip the folder you want to download with the command:

- `tar -zcvf folder_name.tar.gz folder_name`
- In this case, I ran Tutorial 1, and the folder name was “MASTML\_tutorial\_1\_GettingStarted”



The screenshot shows a terminal window from a JupyterLab session on nanoHUB. The terminal prompt is `rjacobs@nanohub_1836304_1:~/data/results/1836304/mastmltutorial/bin$`. The user has entered the command `ll`, which displays the contents of the current directory. The output shows several files and directories, including `.ipynb_checkpoints` and `MASTML_tutorial_1_GettingStarted`. The user then enters the command `tar -zcvf MASTML_tutorial_1_GettingStarted.tar.gz MASTML_tutorial_1_GettingStarted/`, which is highlighted with a green box. The terminal output shows the progress of the tar command, listing the files being archived, such as `MASTML_tutorial_1_GettingStarted/mastml_metadata.json`, `MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/`, and `MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/`.

```
rjacobs@nanohub_1836304_1:~/data/results/1836304/mastmltutorial/bin$ ll
total 0
drwxr-xr-x 4 rjacobs public 116 May 14 12:47 .
drwxr-xr-x 9 rjacobs public 156 May 14 12:46 ..
drwxr-xr-x 2 rjacobs public 10 May 14 12:46 .ipynb_checkpoints
drwxr-xr-x 4 rjacobs public 209 May 14 12:48 MASTML_tutorial_1_GettingStarted
lrwxrwxrwx 1 rjacobs public 49 May 14 12:46 mastmltutorial.ipynb -> /apps/mastmltutorial/r11/bin/mastmltutorial.ipynb
rjacobs@nanohub_1836304_1:~/data/results/1836304/mastmltutorial/bin$ tar -zcvf MASTML_tutorial_1_GettingStarted.tar.gz MASTML_tutorial_1_GettingStarted/
MASTML_tutorial_1_GettingStarted/
MASTML_tutorial_1_GettingStarted/mastml_metadata.json
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/data_preprocessed_train.xlsx
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/StandardScaler.pkl
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/data_preprocessed_test.xlsx
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/selected_features.txt
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/selected_features.xlsx
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/data_preprocessed_train_selected.xlsx
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/X_train.xlsx
MASTML_tutorial_1_GettingStarted/LinearRegression_RepeatedKFold_SklearnPreprocessor_NoSelect_05_14_12_48_05/split_0/X_test.xlsx
```

# Downloading results from MAST-ML tutorial: Step 4

Step 4.) Look back at your displayed folders: there is now a new .tar.gz folder. You can click the box next to this folder and click the “Download” button.



Submit a ticket

Terminate Session

Files **Running** Formgrader Assignments Courses

Duplicate Rename Move **Download** View Edit

Upload New

		Name	Last Modified	File size
<input type="checkbox"/>		MASTML_tutorial_1_GettingStarted	2 minutes ago	
<input type="checkbox"/>		mastmltutorial.ipynb	Running 4 minutes ago	49 B
<input checked="" type="checkbox"/>		MASTML_tutorial_1_GettingStarted.tar.gz	seconds ago	10.1 MB