Using Jupyter to Take Your Data Science Workflow to the Next Level

Hunter Glanz

California Polytechnic State University
San Luis Obispo, California, USA

July 24, 2019

Aims/Themes of This Talk

Reproducibility

Aims/Themes of This Talk

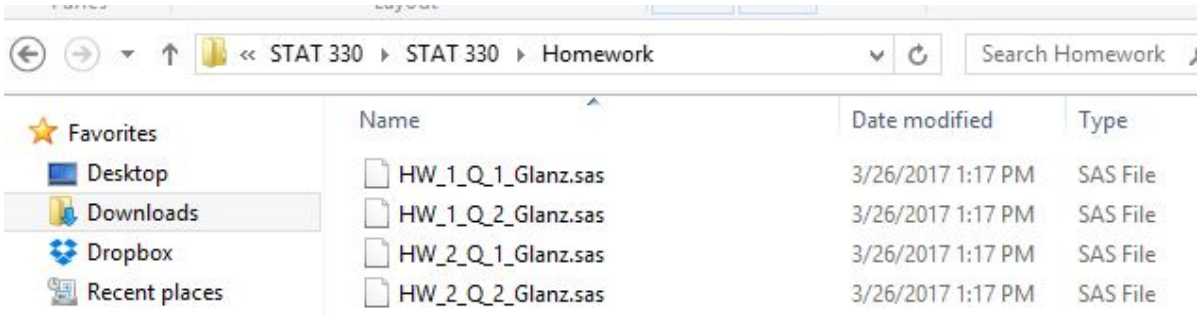
Reproducibility
and
Integration of SAS with open source tools like Jupyter

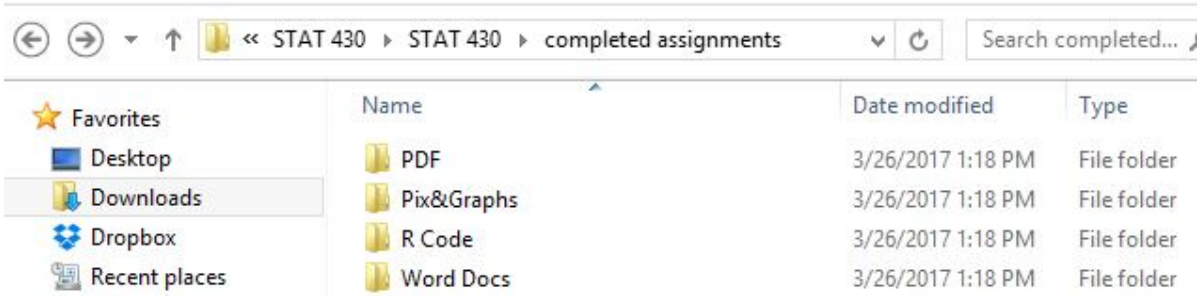
Historical Collaboration on Projects

- Organized, but fragmented...

Historical Collaboration on Projects

- Organized, but fragmented...

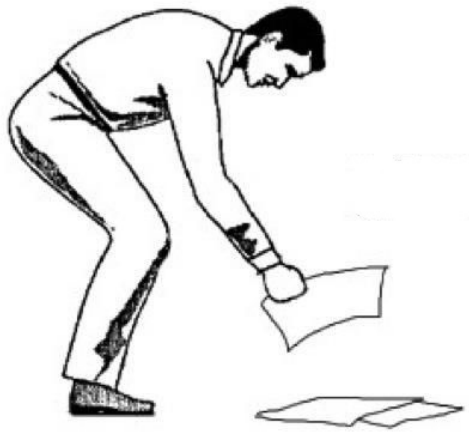




Story is Similar between Academia and Industry

- Clean, wrangle, manage data
- Summarize and visualize data
- Analyze and model
- Synthesize and report

"Some" assembly required



Historical Deficiencies

- Fragmented collection of files:
  - Code, text, images, data...
- Communication, readability, reproducibility suffer
- Unnecessarily large distance between data and story

Does Anything Out There Work?

- Make computing easier:
  - IDEs and editors (Emacs, Notebook++, Vim, SAS Studio, RStudio, etc.)
- Make report-writing easier:
  - SAS*, RStudio, LaTeX

```
5  /* Question 2*/
6
7  data increment;
8
9
10 do i = 0 to 15000 by 1000;
11 /* saving the dates for ecah increment*/
12 caldate = i;
13 /* finding the day of the week*/
14 dayofweek = weekday(i);
15 output;
16 end;
17
18
19 run;
20
21 proc print data = increment;
22 title "Dates";
23 /*formatting the SAS date to a readable date*/
24 format caldate mmddyy10. ;
25 run;
```

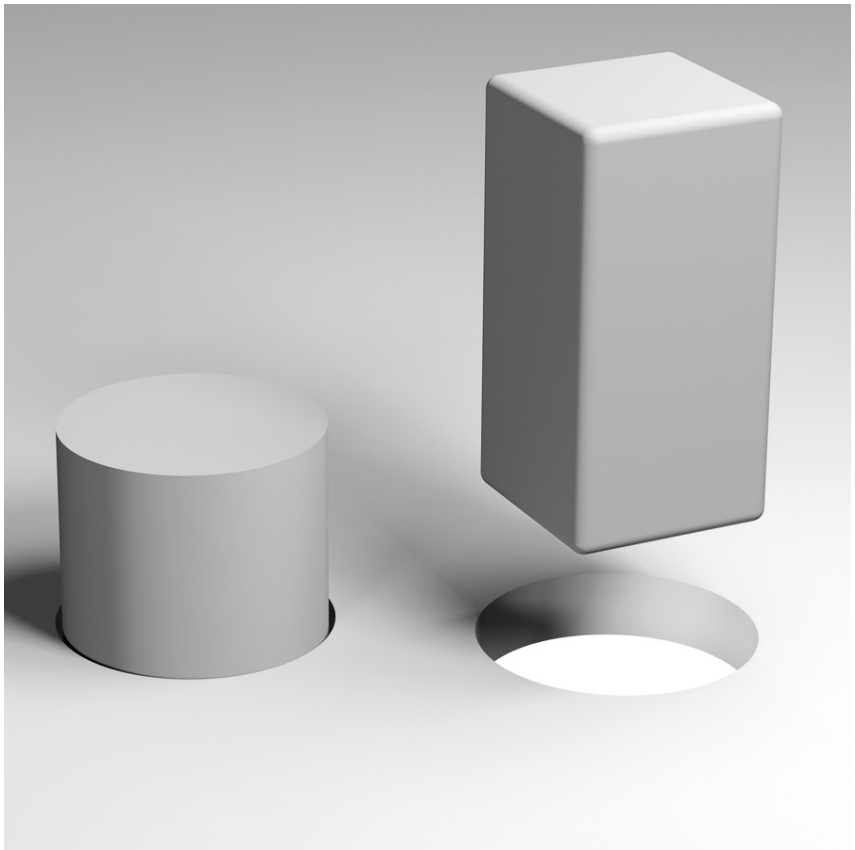
Style=Pearl and Highlight Color =Chartreuse

Country	# Medalists	Total Medals			Age (yrs)	Weight (kg)	Gender	
		Sum	Ratio	Max per Athlete	Average	Average	Female	Male
Australia	25	32	1.28	4	24.4	73.2	76.0%	24.0%
Azerbaijan	1	1	1.00	1	18.0	56.0	---	100.0%
Belarus	4	4	1.00	1	29.5	68.3	75.0%	25.0%
Belgium	2	2	1.00	1	26.0	61.0	50.0%	50.0%
Brazil	5	5	1.00	1	23.4	81.6	40.0%	60.0%
Canada	25	25	1.00	1	28.2	77.8	56.0%	44.0%

Nothing Addresses Our Needs

- IDEs/Editors
  - SAS Studio and RStudio
    - Built-in documentation
  - Color coding, formatting, etc.
- Documenting
  - RMarkdown & Report-Writing Interface
  - Highly customizable LaTeX

None of these integrate coding and documentation!



Capabilities of Ideal Tool

- Color coding, formatting, organization of existing tools
- Documentation and report writing organically coexist with coding
- Streamline the process of:
  - Exploratory data analysis
  - Data science
  - Data journalism
  - Research
  - Analytics, etc.

Cue the Jupyter Project

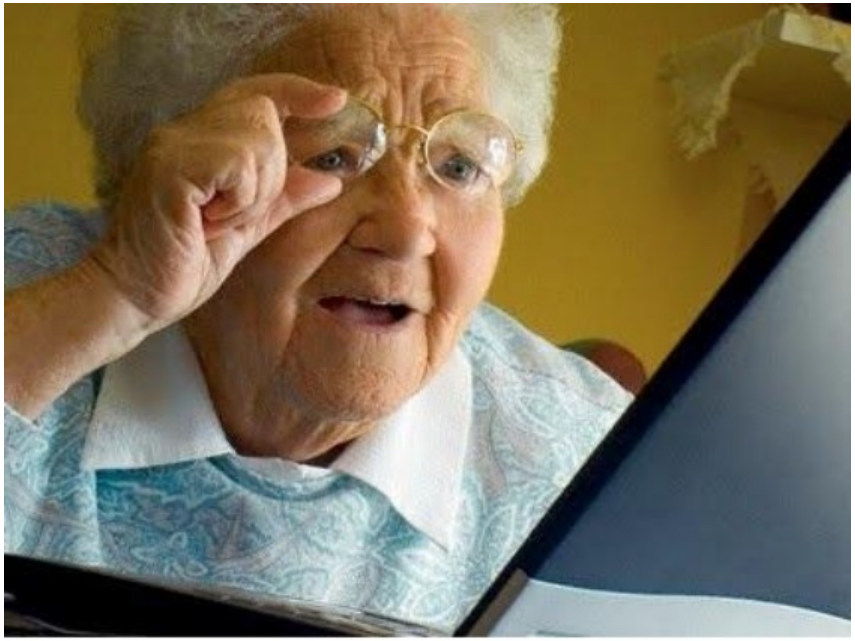
- Supports over 40 software languages!

You May Be....

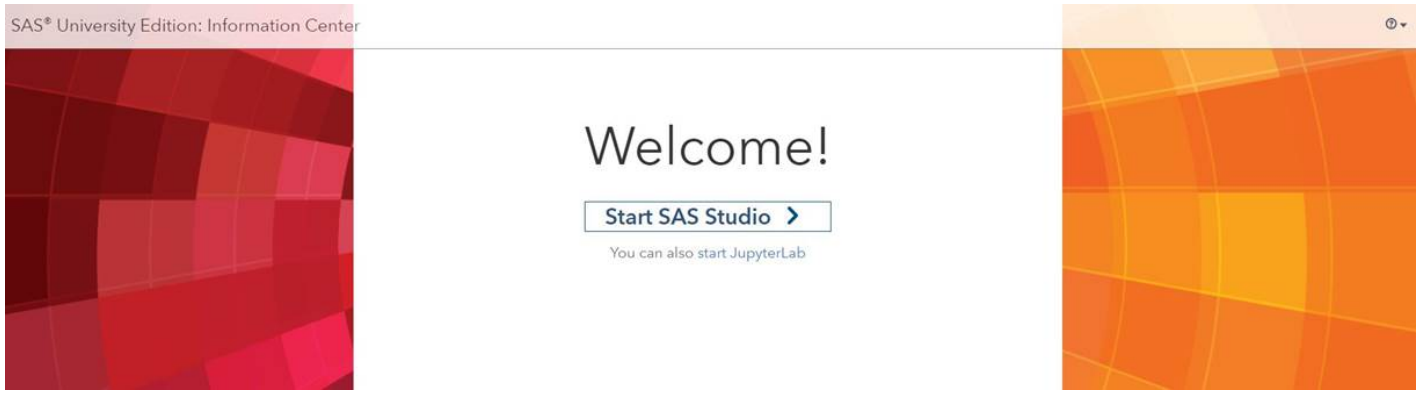

CALIFORNIA · POLYTECHNIC · STATE · UNIVERSITY · 1901 ·
DISCERE FACIENDO

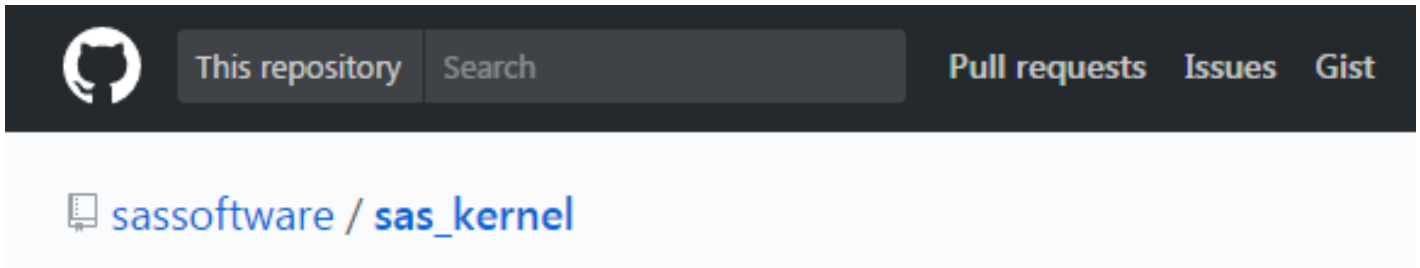
You May Be....

You May Be....



Getting Started with Jupyter Notebooks





Live Demo of Jupyter via SAS University Edition!

Live Demo!

Project Jupyter Summary Part I

- Web application enabling creation of documents that contain live code, equations, visualizations and explanatory text.
- Over 40 languages are supported, including SAS, Python, and R!
- Code within the notebook can produce rich output such as images, videos, LaTeX and JavaScript.
- Interactive widgets can be used to manipulate and visualize data in real time.

Project Jupyter Summary Part II

- Nbviewer
- GitHub
- Binder
- JupyterHub (for organizations)

Conclusions: Ease of Use

- From personal use to JupyterHub for organizations, Jupyter Notebooks make statistical computing easier to do AND share than every before!

  "The need to minimize thought-to-execution friction is probably the single biggest productivity requirement in corporate America."

- Jupyter Notebooks let us take a GIANT leap toward achieving this when it comes to data science and statistical computing.

Conclusions: The Tool We Need and Deserve

- Jupyter Notebooks
  - A single vehicle for live code, text and images.
  - Dynamic documents that enhance statistical computing and statistical communication.
  - Streamline the analytics process
    - The color coding, formatting and organization of your favorite editor
    - Report writing capabilities
    - The integration and synthesis of these things into a single environment

Thank You! Questions?

- Slides and Live Demo at
  https://github.com/hglanz/PugSUG2019-GlanzTalk