

# INFO-UB 23: Introduction to Programming and Data Science

---

Katherine Hoffmann Pham

July 2, 2018

NYU Stern, Department of Information Systems

# Welcome!

1. Please let me know if you are not on NYU Classes:  
<https://newclasses.nyu.edu>
2. Please log in to the Jupyterhub server:  
[jupyterhub.ipeirotis.org](https://jupyterhub.ipeirotis.org)  
If necessary, for today you can use Google Colab:  
<https://colab.research.google.com/>
3. Please ensure that you have access to our Slack account:  
<https://info-ub23-summer2018.slack.com/>
4. Please fill out the course survey:  
<https://goo.gl/forms/YDNQH7QCu0FkPW3o1>

# Overview

---

## INFO-UB23: Introduction to Programming and Data Science aka “dealing with data”

1. Getting and cleaning data (Python)
2. Database design and management (SQL)
3. Visualization and analysis (Pandas, Matplotlib)

# Prerequisites and Focus

- Prerequisites:
  - Your computer and charger
- Focus: Data ...
  - Collection
  - Storage
  - Organization
  - Management
  - Analysis
- Not the focus:
  - Machine learning
  - Data mining
  - Big data

# Why INFO-UB 23?

“Starting in 2012, my colleagues and I began taking a closer look at the hands-on experience of data scientists ... What we found was that the bulk of their time was spent manipulating data – a mix of data discovery, data structuring, and creating context. In other words, most of their time was spent turning data into a usable form rather than looking for insights.”

- HBR, “The Sexiest Job of the 21st Century is Tedious, and That Needs to Change”

# Course Philosophy

1. If it is repetitive, automate it.

# Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.



# Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.

matplotlib / matplotlib

Watch

504

★ Star

7,500

Fork

3,467

Code

Issues 1,101

Pull requests 268

Projects 5

Wiki

Insights

## Join GitHub today

GitHub is home to over 28 million developers working together to host and review code, manage projects, and build software together.

Sign up

Dismiss

matplotlib: plotting with Python <http://matplotlib.org/>

26,076 commits

11 branches

70 releases

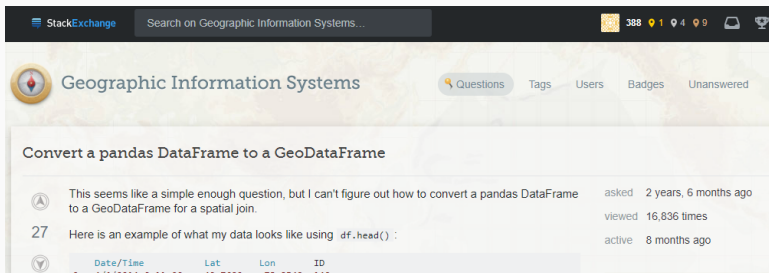
733 contributors

# Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.
4. Expertise is relative.

# Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.
4. Expertise is relative.
5. Everyone has something to contribute.



The screenshot shows a StackExchange page for the 'Geographic Information Systems' tag. The question is titled 'Convert a pandas DataFrame to a GeoDataFrame'. The user asks for help converting a pandas DataFrame to a GeoDataFrame for a spatial join. They provide an example of their data using `df.head()`. The question has 27 answers, with the most recent one from 8 months ago. The page also shows the user's profile, the number of questions (388), and various badges.

StackExchange Search on Geographic Information Systems...

Geographic Information Systems

Questions Tags Users Badges Unanswered

**Convert a pandas DataFrame to a GeoDataFrame**

This seems like a simple enough question, but I can't figure out how to convert a pandas DataFrame to a GeoDataFrame for a spatial join.

27 Here is an example of what my data looks like using `df.head()` :

Date/Time	Lat	Lon	ID
4/1/2014 0:11:00	40.7600	-73.9540	140

# Course Philosophy

Implications:

1. Stop me to ask questions at any time
2. Turn to each other for help

# Where to next?

- [INFO-UB 24](#) Projects in Programming and Data Science
- [INFO-GB 3106](#) Data Visualization
- [INFO-UB 57](#) Data Mining for Business Analytics
- [NYU Center for Data Science](#)
- Real life ...

# Introductions . . .

Let's give the short version of the survey:

- Name
- What you study
- Why this class?
- Experience working with data?
- Hobbies/interests

# Course Logistics

---

Announcements, grades, materials via NYU classes:

<https://newclasses.nyu.edu/>

Otherwise, see the course webpage:

[http://people.stern.nyu.edu/khoffman/intro\\_programming\\_datasci/](http://people.stern.nyu.edu/khoffman/intro_programming_datasci/)



The course notes are available on Github.

- Our repository is:

`https://github.com/khof312/Summer2018\_ProfHoffmannPham`

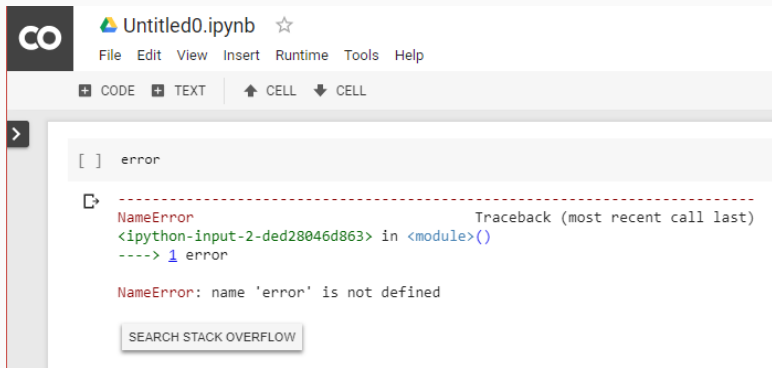
- This is a fork of a master repository:

`https://github.com/ipeirotis/dealing\_with\_data`

# Getting Help

For troubleshooting and help, I recommend:

1. Google



# Getting Help

For troubleshooting and help, I recommend:

1. Google
2. [StackOverflow](https://stackoverflow.com/): `https://stackoverflow.com/`

# Getting Help

For troubleshooting and help, I recommend:

1. Google
2. [StackOverflow](https://stackoverflow.com/): `https://stackoverflow.com/`
3. [Slack](https://info-ub23-summer2018.slack.com/): `https://info-ub23-summer2018.slack.com/`

# Getting Started

---

# Jupyterhub overview

- Jupyterhub is our online data science environment
- I encourage you to use it for several reasons:
  1. It provides a [standardized environment](#)
  2. I can [automatically update](#) the available files
  3. [Homeworks](#) can easily be tested and submitted

# Jupyterhub overview

The screenshot shows the Jupyterhub web interface with several handwritten annotations in red:

- Navigation Bar:**
  - Files** (labeled 1)
  - Running** (labeled 2)
  - Clusters**
  - Assignments** (labeled 3) - An arrow points to this tab with the text "fetch / complete / validate / submit".
  - Nbextensions**
- Top Right:**
  - Logout** (labeled 7)
  - Control Panel** (labeled 6) - An arrow points to this button with the text "worst case => shut down".
- Actions:**
  - Upload** (labeled 4) - An arrow points to this button with the text "add new files".
  - New** (labeled 8) - An arrow points to this button with the text "create files".
- File List:**
  - A list of files and folders is shown with columns for "Name" and "Last Modified".
  - The file **Summer2018\_ProfHoffmannPham** is highlighted with a red box and labeled (2).
  - A bracket on the left side of the file list is labeled "select to rename, delete".
- Bottom Left:**
  - Handwritten text: "syncs w/ Github" and "automatic conflict resolution".

# Jupyter notebook overview

The image shows a Jupyter notebook interface with several handwritten annotations in red ink:

- File menu:** An arrow points to the "File" menu with the note "save download".
- Edit menu:** An arrow points to the "Edit" menu with the note "cut/copy/split".
- Insert menu:** An arrow points to the "Insert" menu with the note "add/delete".
- Cell menu:** An arrow points to the "Cell" menu with the note "run, start, stop".
- Kernel menu:** An arrow points to the "Kernel" menu with the note "restart and clear".
- Help menu:** An arrow points to the "Help" menu with the note "rename".
- Toolbar:** An arrow points to the "Run" button with the note "change cell type". An arrow points to the "Validate" button with the note "keyboard shortcuts".
- Markdown cell:** An arrow points to a markdown cell with the note "markdown cell".
- Code cell:** An arrow points to a code cell with the note "code cell".
- Run order:** An arrow points to the "In [ ]" prompt in a code cell with the note "indicates run order, or running [\*]".
- Status bar:** An arrow points to the "Not Trusted" status with the note "status". An arrow points to the "Python 3" version indicator with the note "Python version".

The notebook content includes a title "A-Introduction to Python Notebooks (unsaved changes)", a section "Python Primer: The Basics", and a code cell containing the following text:

```
In [ ] print("Hello World")
```