SML 312 – Research Projects in Data Science

Course Overview & Project Workflow

Three Mini-Projects + One Final Project

- Mini-Project #1 – Linear Regression
- Mini-Project #2 – Classification
- Mini-Project #3 – NLP / Image Processing
- Final Project – up to you!

Class Scheduling:

- M/Tu 7:30 – 9pm Course Meeting (w/10 min break)
- Precepts – Fridays – 1:30pm (w/Daniel Melese)
- Office Hours – After Class or by Appt

SML 312 – Fall 2022 Calendar

	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
September	9/5	9/6 7:30-9pm Class 1a					
	9/12 7:30-9pm Class 1b	9/13 7:30-9pm Class 2a					
	9/17 7:30-9pm Class 2b	9/18 7:30-9pm Class 3a					
October	9/24 7:30-9pm Class 3b	9/25 7:30-9pm Class 4a					
	10/3 7:30-9pm Class 4b	10/4 7:30-9pm Class 5a					
	10/10 7:30-9pm Class 5b	10/11 7:30-9pm Class 6a					
	10/17 Fall Break	10/18 Fall Break					
November	10/24 (Mid Grades due) 7:30-9pm Class 6b	10/25 (Mid Grades due) 7:30-9pm Class 7a					
	10/31 7:30-9pm Class 7b	11/1 7:30-9pm Class 8a					
	11/7 7:30-9pm Class 8b	11/8 7:30-9pm Class 9a					
	11/14 7:30-9pm Class 9b	11/15 7:30-9pm Class 10a					
	11/21 7:30-9pm Class 10b	11/22 Thanksgiving Break					
December	11/28 7:30-9pm Class 11a	11/29 7:30-9pm Class 11b					
	12/5 7:30-9pm Class 12a	12/6 7:30-9pm Class 12b			12/9 Reading Period Final Papers Due	12/10 Day 1 of Final Presentations	12/11 Day 2 of Final Presentations
					12/16 Dean's Date All Final Projects Due		

SML 312 – Fall 2022 Calendar

	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
September	9/5	9/6 7:30-9pm Class 1a					
	9/12 7:30-9pm Class 1b	9/13 7:30-9pm Class 2a					
	9/17 7:30-9pm Class 2b	9/18 7:30-9pm Class 3a	Project #1 – Linear Regression				
October	9/24 7:30-9pm Class 3b	9/25 7:30-9pm Class 4a					
	10/3 7:30-9pm Class 4b	10/4 7:30-9pm Class 5a					
	10/10 7:30-9pm Class 5b	10/11 7:30-9pm Class 6a	Project #2 – Classification				
	10/17 Fall Break	10/18 Fall Break					
November	(Mid Grades due) 10/24 7:30-9pm Class 6b	(Mid Grades due) 10/25 7:30-9pm Class 7a					
	10/31 7:30-9pm Class 7b	11/1 7:30-9pm Class 8a	Project #3 – NLP / Neural Networks				
	11/7 7:30-9pm Class 8b	11/8 7:30-9pm Class 9a					
	11/14 7:30-9pm Class 9b	11/15 7:30-9pm Class 10a					
December	11/21 7:30-9pm Class 10b	11/22 Thanksgiving Break					
	11/28 7:30-9pm Class 11a	11/29 7:30-9pm Class 11b					
	12/5 7:30-9pm Class 12a	12/6 7:30-9pm Class 12b			Reading Period 12/9 Final Papers Due	Day 1 of 12/10 Final Presentations	Day 2 of 12/11 Final Presentations
					Dean's Date 12/16 All Final Projects Due		

Partial Cloud of Possible Lecture Topics

- Under/ Overfitting
- Exploratory Data Analysis
- Error Analysis
- Project Lifecycle
- Regularization
- Curse of Dimensionality
- Dimensionality Reduction
- Measures of Goodness -- ROC & Confusion Matrix
- Sample Data Science Papers
- Learning Rate and Stability

- Test/Train Split
- Cross-validation
- Out of sample
- Dummy Variables
- RANSAC
- Linear Regression
- Naïve Bayes
- Decision Trees & Random Forests
- NLP
- Vector Embeddings
- Neural Networks

- Clustering
- Nearest Neighbors
- K-means
- SVMs
- Feature Extraction

- Python / Jupyter
- Github
- Slack
- Google Colab
- Cloud
- Data Ingestion
- Data Cleaning
- Feature Engineering
- Reinforcement Learning?
- Web Scraping / Packages
- Ensemble Methods – Bagging and Boosting

Other ideas for Possible Topics?

- Your suggestions here!
- Python Modules – Seaborn & Matplotlib.
- Computing on the Cloud… (AWS/GCP/Azure)
- Computer Vision…

Data Science Project Lifecycle

Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:

Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:

1. Ask a Question

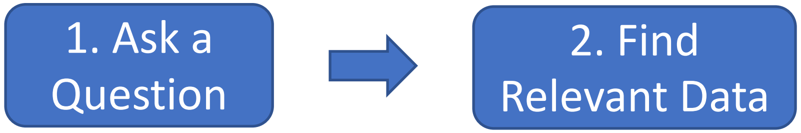
Data Science Project Lifecycle

Think about the data science projects that you’ve worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



1. Ask a Question

2. Find Relevant Data

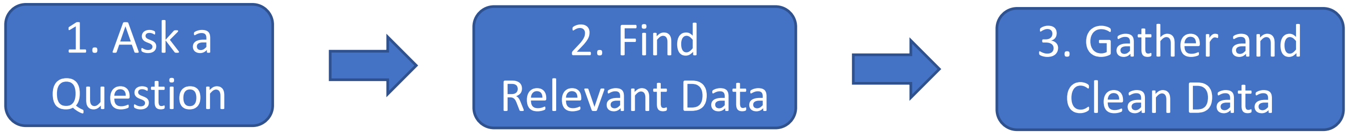
Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



1. Ask a Question

2. Find Relevant Data

3. Gather and Clean Data

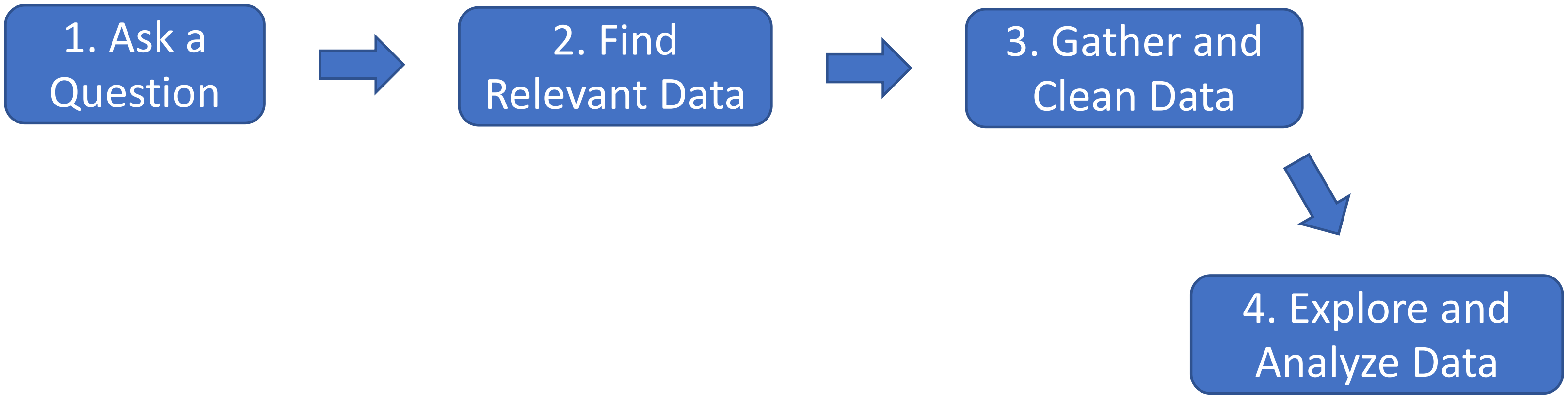
Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



(c) 2022 Jonathan Hanke for Princeton Fall 2022 Course SML 312

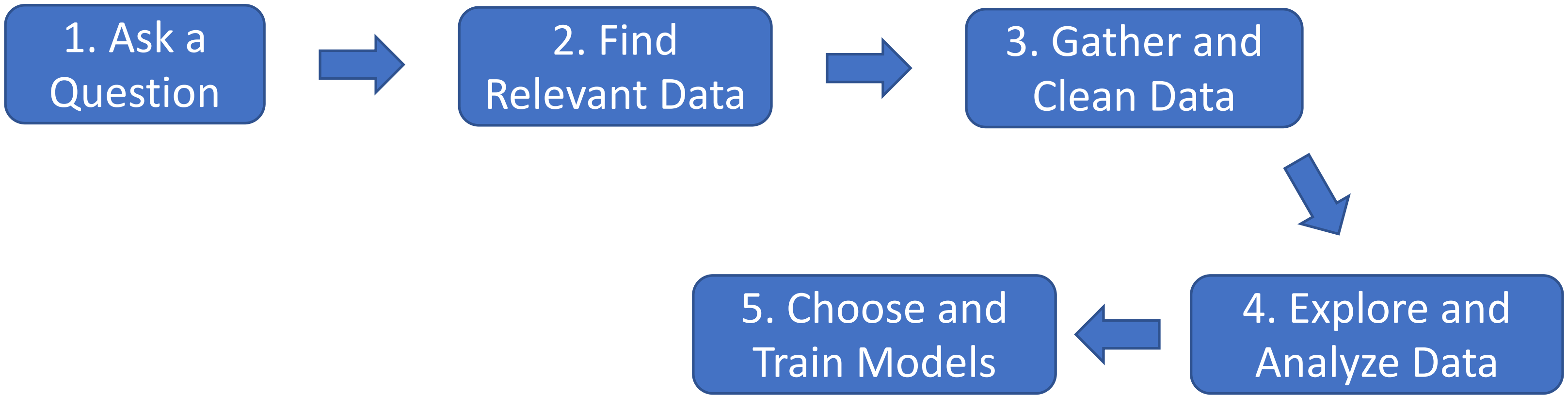
Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



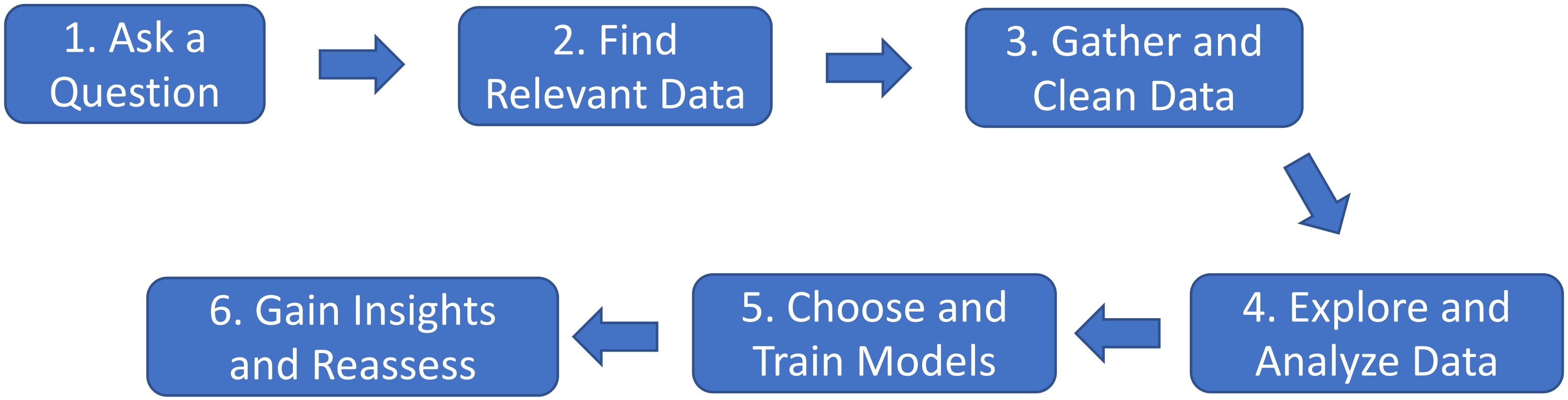
Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



1. Ask a Question

2. Find Relevant Data

3. Gather and Clean Data

4. Explore and Analyze Data

5. Choose and Train Models

6. Gain Insights and Reassess

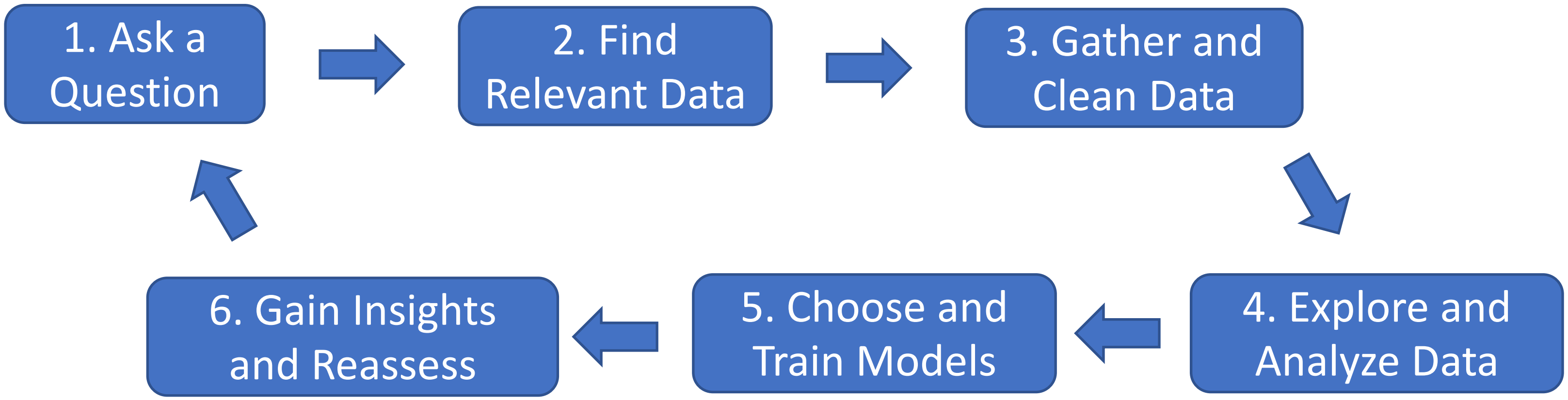
Data Science Project Lifecycle

Think about the data science projects that you've worked on -- what steps did you go through as you did you do them?

Did you repeat any of the steps?

Which steps took the most thought?

One Possible Data Science Project Lifecycle:



1. Ask a Question

2. Find Relevant Data

3. Gather and Clean Data

4. Explore and Analyze Data

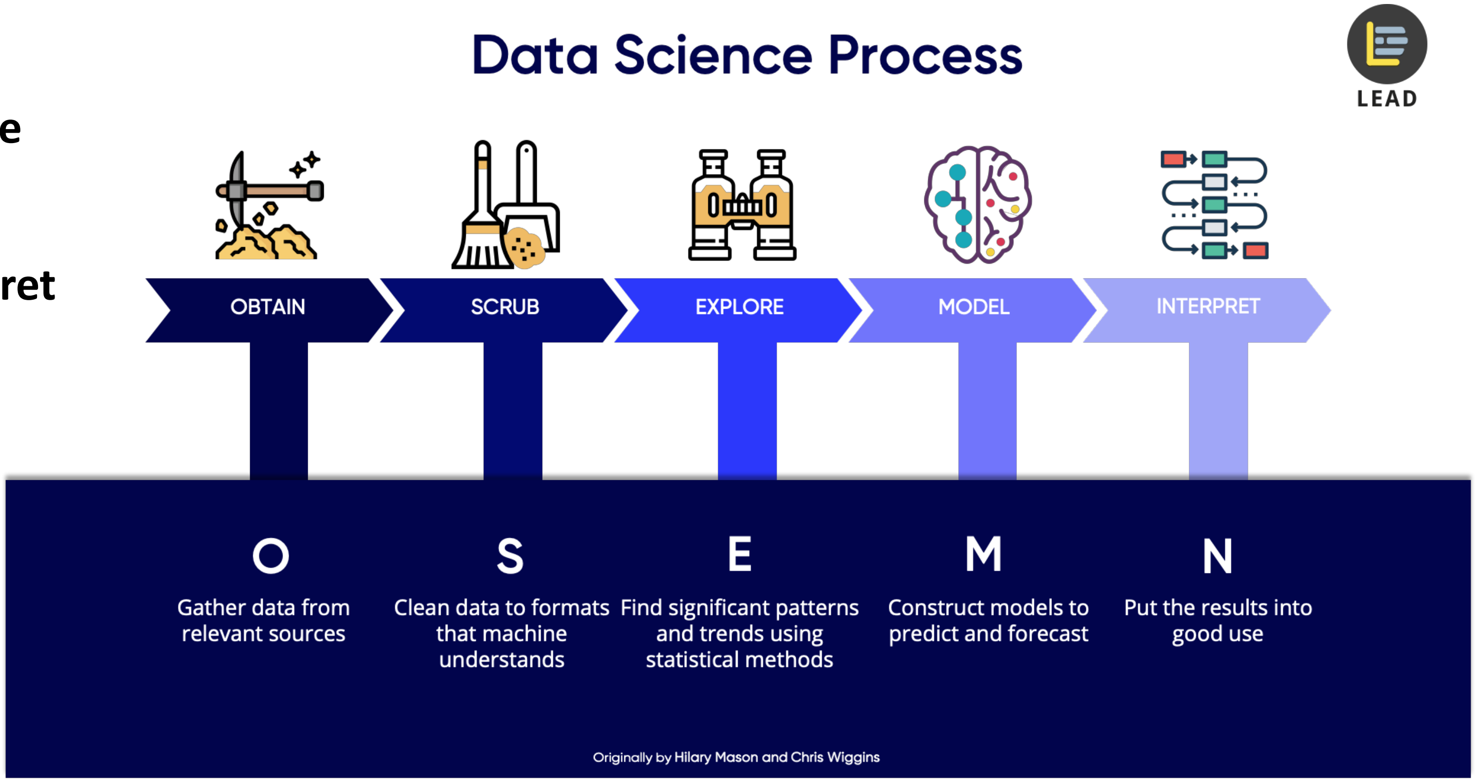
5. Choose and Train Models

6. Gain Insights and Reassess

Data Science Project Lifecycle

Another Version – OSEMN “Awesome” (originated in 2010 by Hillary Mason):

1. **Obtain**
2. **Scrub**
3. **Explore**
4. **Model**
5. **iNterpret**



Data Science Process

OBTAIN — SCRUB — EXPLORE — MODEL — INTERPRET

O Gather data from relevant sources

S Clean data to formats that machine understands

E Find significant patterns and trends using statistical methods

M Construct models to predict and forecast

N Put the results into good use

Originally by Hilary Mason and Chris Wiggins

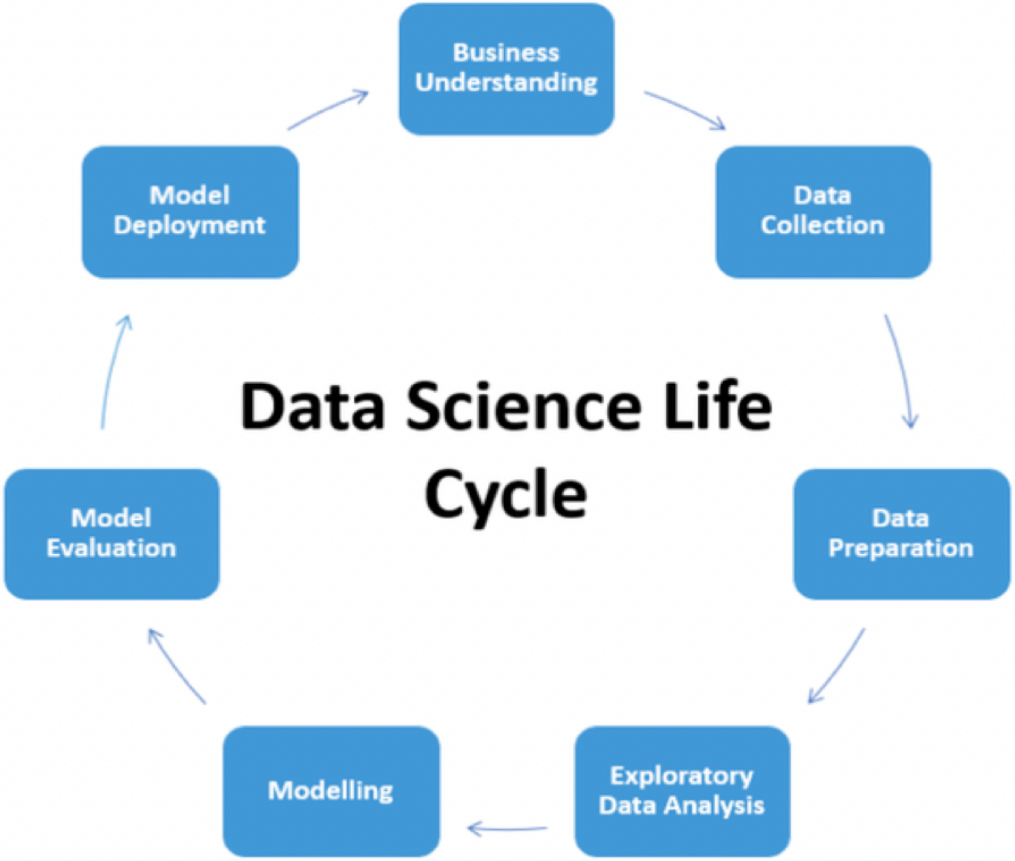
Reference: 1/3/2019 Article “5 Steps of a Data Science Project Lifecycle”

https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Data Science Project Lifecycle

Still Other Versions:

1. **Business Understanding**
2. **Data Collection**
3. **Data Preparation**
4. **Exploratory Data Analys**
5. **Modelling**
6. **Model Evaluation**
7. **Model Deployment**



Reference: 4/15/2021 Article: ”A Complete Tour of the Data Science Lifecycle”

https://analyticsindiamag.com/a-complete-tour-of-data-science-project-life-cycle/

Your Data Science Final Project Lifecycle

Using a process like this (you decide*), please start to work on your final project (through some initial EDA) this week so you have some data to consider as we start talking about modelling next week. Taking time to get to know your data will help you decide how to think about what it can tell you!

It's ok if your process doesn't feel perfect the first time... it is perfectly normal to iterate the process several times with your project by the end of the course.

Starting early with your question and data exploration will give you time and options for next steps if your initial data isn't useful for answering the question you're interested in, or if you need to refine your question based on what you learn from initial explorations!

***For Reference, more than you ever wanted to know about Data Science Workflows is here:**

https://resources.github.com/downloads/development-workflows-data-scientists.pdf