# Establishing a Productive Machine Learning Workflow: Syllabus

January 13, 2020 / 6:00 PM - 7:30PM EST

## Important Links

### Workshop Hackpack

Pre-workshop checklist, and resources to explore during and after the workshop.

### Hack the North 2020++ Event Schedule

Check this out to stay up-to-date on activities, workshops, and other key happenings this weekend.

## Motivator

Conventional software engineering workflows are suboptimal for a Data Scientist. But, there are elements of a SWE workflow we can tweak to boost productivity.

Through this workshop, you'll learn about best practices for experiment configuration, productivity tools, continuous integration for data science, static analysis, and more.

## Prerequisite Knowledge

To get the most out of this workshop, you should have completed at least one project in Python, ideally a data science project. Most of the content will not depend on much more than basic Python syntax, but you will have a better sense of why some of the suggestions are useful with some background.

## Learning Outcomes

This is what you will walk away from the workshop able to do:

- Create reproducible data science code using python environments
- Write tests suitable for machine learning research.
- Use static analysis tools appropriate for data science.
- Set up basic pre-commit and continuous integration with GitHub Actions
- Configure projects using Hydra

## Timeline (1 hour)

| Time | Module | Description |
|---|---|---|
| 5 min. | Introduction | |
| 10 min. | Python Environments | We'll talk about why you should use PIP Environments, how they compare to Conda Envs/Docker/etc, and how to use pip-tools to smooth out your workflow. |
| 10 min. | Testing | We'll motivate testing for data science with pytest (it's not just for engineers!), talk about how to test for an exploratory project, and some alternative frameworks. |
| 10 min. | Static Analysis | We'll cover type checking and other static analysis in Python, specifically what types are worth the time investment for data science (focusing on Pyflakes). |
| 15 min. | Version Control Tips | We'll set up pre-commit, basic continuous integration with GitHub Actions, and talk about what specifically is useful for research projects. |
| 5 min. | Break | |
| 15 min. | Configuring Experiments | We'll focus on configuring projects in a clean, sustainable way using Hydra. |
| 5 min. | Conclusion and Q&A | |