

Productionalizing a Data Science Team

Nicole Carlson (she/her/hers)

nicole@parsingscience.com

@parsing_science

What do I mean by
productionalizing a team?

Testimonial

“I’m able to experiment a lot more at work following our guidelines because it just makes it easier to organize it all in my head and get started on the actual functionality.”



Morgan, Data Scientist

About Me

- Former physicist
- First Data Scientist hired at my last company; team was 4 people when I left
- Second Data Scientist hired at my current company; team is now 10 people



Overview

- Pre-requisites
- Project Management
- Code Organization
- Coding Standards
- Code Reviews
- Model Deployment
- Documentation

Overview

- Pre-requisites
 - Project Management
 - Code Organization
 - Coding Standards
 - Code Reviews
 - Model Deployment
 - Documentation

Hire people who want to work on a team



Hire people with strong opinions



- Ravenclaw
- Line length: 100
- Loves trailing commas
- Uses *Black* autoformatting
- Cares a lot about coding standards



- Slytherin
- Line length: 110
- Hates trailing commas
- Formats everything by hand
- Cares a lot about coding standards

Overview

- Pre-requisites
- **Project Management**
- Code Organization
- Coding Standards
- Code Reviews
- Model Deployment
- Documentation

Agile Twitter Poll



MichelangeloDAgostin

@MichelangeloDA

Question for my data scientist friends out there. How do you feel about "agile"?



39 votes · Final results

2:47 PM · Oct 2, 2019 · [Twitter Web App](#)

Come up with a process that works for your team

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Standup	Standup	Standup	Standup	Standup	Standup	Standup	Standup	Standup	Demos
Retro for prior sprint								Sprint Planning	

How my team does Agile Sprint Planning

- Each person comes up with their own tasks
- The group discusses each task and estimates its time
- R&D or Literature Review tasks are timeboxed
- Long-term project planning happens outside of the meeting
- Code reviews are assigned after the meeting

Remove unneeded parts, e.g. user stories

Description

As a Hanna Torrence and Nicole Carlson

I want to understand what the other person is doing in Ollivander

So that we are both on the same page

~~Also each person will set aside X points to make subtasks~~

Description

As a Data Scientist,

I want to figure out if we can get non-member pageview cosine similarity to run so that we can provide more recommendations

Track your work

- What is each person working on?
- What you will work on next?
- What is the overall outline of your project?
- What are potential future projects?
- What requests have you gotten from other teams?

Tools: Project Management



Testimonial

“Having a project tracking system like JIRA has been a lifesaver. I will regularly put our current sprint side-by-side with our quarterly goals to check in that we're on the right track.”



Ali, Director of Data Science








Overview








- Pre-requisites
- Project Management
- **Code Organization**
- Coding Standards
- Code Reviews
- Model Deployment
- Documentation

What's next after Jupyter notebooks?



Organization: choose a standard repo layout

 <code>.github</code>
 <code>hermiones_handbag</code>
 <code>tests</code>
 <code>.gitignore</code>
 <code>CHANGELOG.md</code>
 <code>Dockerfile</code>
 <code>JenkinsPipeline</code>

 <code>LICENSE.txt</code>
 <code>Makefile</code>
 <code>README.md</code>
 <code>docker-compose.yaml</code>
 <code>requirements.txt</code>
 <code>setup.cfg</code>
 <code>setup.py</code>

Tools: Code generators



Extra organization: Common utilities repo



Sonic Screwdrivers

What's in a common utilities repo?

- Common SQL
- Database connections
- Data transformations
- Logging helpers
- Unittest helpers
- Visualization helpers

Overview

- Pre-requisites
- Project Management
- Code Organization
- **Coding Standards**
- Code Reviews
- Model Deployment
- Documentation

Examples of coding standards

- How long will your lines be?
- How will you continue long lines?
- How will you write docstrings?
- Will you use single quotes or double quotes?
- Tabs vs spaces

Tips for developing coding standards

- Start from PEP8
- Consult with software engineers at your company
- Make a rule when all (or most) people agree
- If there is no rule, then be consistent when you write your own code
- If multiple people are working on the same repo, they should agree on a set of standards for that repo

Linters: automatically finds code violations



- A linter looks for programmatic and stylistic errors, e.g.
 - Programmatic: The code tries to use an undefined variable
 - Stylistic: Your lines are more than X characters long

Tools: Automated Code Formatters



YAPF

Testimonial

“My personal projects now follow our team coding standards because it’s painful to look at things that aren’t formatted. Double quotes burn my eyes now.”



Michael, Senior Data Scientist

Overview

- Pre-requisites
- Project Management
- Code Organization
- Coding Standards
- **Code Reviews**
- Model Deployment
- Documentation

But first ... a cautionary tale!

Why you should never push to master...

 Showing **1 changed file** with **1 addition** and **0 deletions**.

1 ■■■■ AUTHORS.txt

... .. @@ -1,2 +1,3 @@

1 1 Michelangelo D'Agostino <mdagostino@shoprunner.com>, <mdagost@gmail.com>

2 2 Nicole Carlson <ncarlson@shoprunner.com>, <nicole@parsingscience.com>

3 +Hanna Torrence <htorrenc@shoprunner.com>, <hanna.torrence.com>

Come up with a Code Review Process

- How the author should prepare a Pull Request (PR)
- How the reviewer should review the code
- What happens when a PR is approved

Teach new people how your team reviews code

- Think about how your process might be different from other places

Tools: PR Checklist

Pull Request Checklist

- ✓ All tests in the `tests` folder pass with a local build
- ✓ Pull request title or description contains the JIRA ID (if applicable)
- ✓ Pull request includes a description of why we are doing this
- ✓ Init files import new capabilities to appropriate level of package (if applicable)
- ✓ CHANGELOG has been updated
- ✓ Version in `_version.py` has been updated
- ✓ README has been updated (if applicable)

Tools: Continuous Integration



Testimonial

“I set up Continuous Integration for my personal project. I set up a linter and unittests. Our PR process has ruined me for the better.”

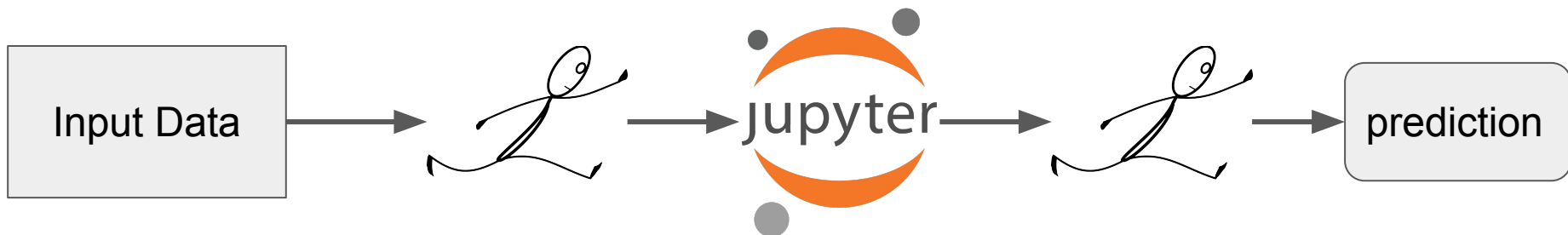


Nate, Data Scientist

Overview

- Pre-requisites
- Project Management
- Code Organization
- Coding Standards
- Code Reviews
- **Model Deployment**
- Documentation

Model predictions: the hard way



Model predictions: the easy way



Deployment suggestions

- Choose one web framework for building APIs
- Use the same one for all your projects
- Consult with software engineers

Tools: Web Frameworks



Flask

Collectively support your APIs

- Load testing
- Monitoring
- Logging

Testimonial

“Our ability to deploy APIs serving model results makes it much easier to get our models tested and in front of customers. We have many examples and extensive documentation to make it easy to get started with deploying a new model.”



Hanna, Senior Data Scientist

Overview

- Pre-requisites
- Project Management
- Code Organization
- Coding Standards
- Code Reviews
- Model Deployment
- **Documentation**

Document everything (in exactly one place)

- Workflow process
- Code review process
- Deployment process
- Instructions on how to get access to things

Example: Idiosyncrasies of your data

Column name	Description
product_id	unique product id
description	product description. This field is optional. It can be concat to more info below.
more_info	this can be treated as actual product description, crawled from vendor's website

Example: Team standards

Data Science Code Style



Most of our repos are in Python. We follow PEP8 with a few other conventions ([ShopRunner Python Coding Standards](#)). The most important thing is to be consistent within the codebase. If you're not sure about something, please ask.

We have a [yeoman generator](#) that creates a standard Python repo structure. Please follow this structure except to the extent that you have strong reasons for departing from it.

Guidelines

- We use numpy docstrings (http://www.sphinx-doc.org/en/stable/ext/example_numpy.html). The only difference is that we start the text of the docstring on the line after the triple quotes, e.g.

Example: Onboarding information

<New hire name>'s Onboarding Document



<Insert whimsical photo here> ★

Introduction

Welcome to ShopRunner Data Science! We're all super, super excited that you're here. This document is meant to do two things. First, it's meant to help you get introduced to our data, our tools, and our work. Second, it's meant to communicate your goals over the next 30 days, 90 days, and 6 months.

Testimonial

“This is the first company I've worked for where the wiki pages were actually useful. You'd go to the instructions on how to set up AWS access, follow the steps, and it would actually work!”



Peter, Senior Data Scientist

Tools: Documentation



GitBook



Recap

- Yes, you need some process
- Whatever you choose, be consistent
- Document everything
- Revisit your choices periodically



Thanks!

- ShopRunner Data Science team
- Tudor Radoaca

My info:
Nicole Carlson
@parsing_science