Building Reproducibile AI

”The Earth is Flat!! My ambigously defined experiment with says so”

J. Setpal

October 6, 2022

Table of Contents

Table of Contents

1. Why it's Worth Your Time
2. Package Management
3. Tracking Code
4. Tracking Data
5. Project Managment

ML as a Science

- Foundationally, Machine Learning is a beautiful clusterf*ck.

ML as a Science

- Foundationally, Machine Learning is a beautiful clusterf*ck.
- Proudly.

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

ML as a Science

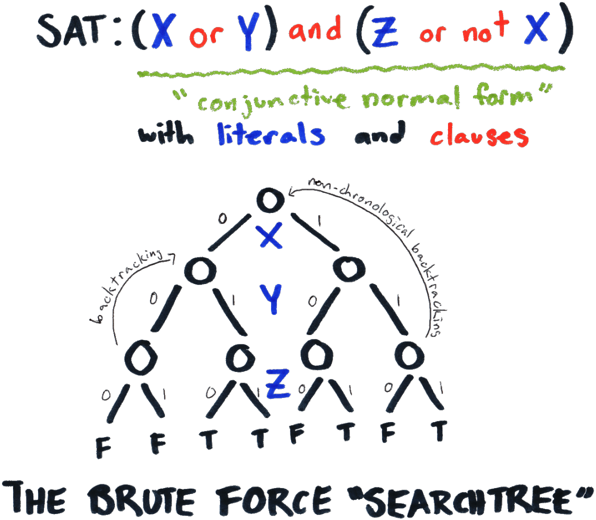
Imagine having to go debug that.

Table of Contents

1. Why it’s Worth Your Time
2. Package Management
3. Tracking Code
4. Tracking Data
5. Project Managment

Pip

Package managers are SAT solvers:


SAT: (X or Y) and (Z or not X)
" conjunctive normal form "
with literals and clauses
non-chronological backtracking
backtracking
X
Y
Z
0
1
F F T T F T F T
THE BRUTE FORCE "SEARCHTREE"

Pip, chill!

Running `pip freeze` returns every dependency within the environment.

Pip, chill!

Running `pip freeze` returns every dependency within the environment.

This is not required to generate the exact package set – `pip` can resolve it autonomously!

```
(ml) [jinen@workstation ~]$ pip freeze | wc -l
263
(ml) [jinen@workstation ~]$ pip-chill | wc -l
29
```

Pip, chill!

Running `pip freeze` returns every dependency within the environment.

This is not required to generate the exact package set – `pip` can resolve it autonomously!

```
(ml) [jinen@workstation ∼]$ pip freeze | wc -l
263
(ml) [jinen@workstation ∼]$ pip-chill | wc -l
29
```

Both achieve the same result; `pip-chill` is just more readable and less cluttered.

Table of Contents

1. Why it's Worth Your Time
2. Package Management
3. Tracking Code
4. Tracking Data
5. Project Managment

Git

Git is a version control system for text-based files.

Git

Git is a version control system for text-based files.

It has a lot of additional functionality; file merging, branching and the utilities to observe, modify and update any commit from the repository's git history.

GitHub

Git $\neq$ GitHub!

GitHub

Git $\neq$ GitHub!

GitHub is merely a service that hosts git servers. Above git, it adds CI/CD scripting, code scanning, as well as release hosting.

Authentication

SSH and **GPG** are two critical security mechanisms used within development.

Authentication

SSH and **GPG** are two critical security mechanisms used within development.

SSH provides a secure interface to communicate with GitHub over public networks.

Authentication

SSH and **GPG** are two critical security mechanisms used within development.

SSH provides a secure interface to communicate with GitHub over public networks.

GPG validates the authenticity of the commit itself.

Authentication

SSH and **GPG** are two critical security mechanisms used within development.

SSH provides a secure interface to communicate with GitHub over public networks.

GPG validates the authenticity of the commit itself.

Per GitHub's recommended security policy, GitHub highly recommends commits to be signed to merge code from a feature branch into the main branch.

Authentication

SSH and **GPG** are two critical security mechanisms used within development.

SSH provides a secure interface to communicate with GitHub over public networks.

GPG validates the authenticity of the commit itself.

Per GitHub's recommended security policy, GitHub highly recommends commits to be signed to merge code from a feature branch into the main branch.

Linus Torvalds didn't sign commits; as a result: [link]

Branching Strategy

Feature branches on projects with a with a lot of contributors can get cluttered.

Using a `<contributor>/<feature>` naming strategy allows developing branches that are easy to recognize and classify.

Table of Contents

1. Why it’s Worth Your Time
2. Package Management
3. Tracking Code
4. Tracking Data
5. Project Managment

DVC

Naive solution for data versioning.

DVC

Naive solution for data versioning. It works by bucketing data and storing it into a cache.

But it works!

MLFlow

MLFlow is a framework for experiments logging.

MLFlow

MLFlow is a framework for experiments logging.

It allows us to make observations between two runs without an active involvement within the experiments.

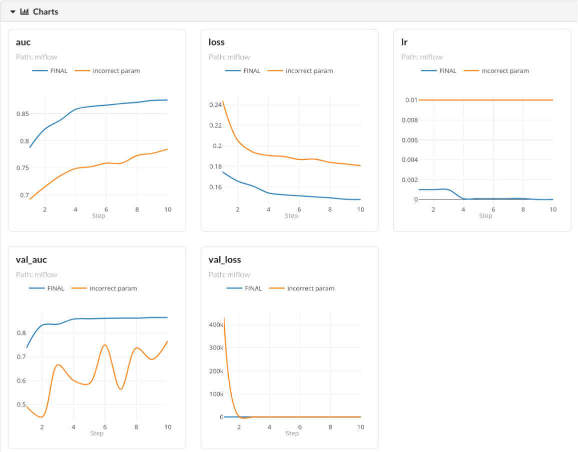


Table of Contents

1. Why it’s Worth Your Time
2. Package Management
3. Tracking Code
4. Tracking Data
5. Project Managment

CookieCutter

This forms a skeletal for our repository.

CookieCutter

This forms a skeletal for our repository.

Developing a codebase using python modules, is made much easier, due to the structure offered by CookieCutter. Directory Structure: [link]

Death to Jupyter Notebooks

Jupyter Notebooks are *fantastic* for experimentation, but unusable in a production context.

Death to Jupyter Notebooks

Jupyter Notebooks are *fantastic* for experimentation, but unusable in a production context.

We can use **Module-Based Development** to ensure to ensure path conditions are maintained without updating the environment.

Module-Based Development

Functionally, it follows the structure define within CookieCutter. As for execution:

```
$ python -m folder.subfolder.subfolder.pythonscript
```

Module-Based Development

Functionally, it follows the structure define within CookieCutter. As for execution:

```
$ python -m folder.subfolder.subfolder.pythonscript
```

Benefits:

- Structured, debuggable code.
- `$PYTHONPATH` is automatically resolved.
- Relative imports work by default!

Let's Code!

```
$ ssh -L 8888:localhost:<n> \
workshop@schema.acm.cs.purdue.edu
```

Password: `workshop`; replace `<n>` with 2000 `< n <` 65000.

```
$ start-exercise
```

Workshop Series Review

Any questions for me?

Thank you!

Have an awesome rest of your day!

Slides: `https://cs.purdue.edu/homes/jsetpal/mlops.pdf`
Exercise: `https://cs.purdue.edu/homes/jsetpal/code.tar.gz`