

# Building scalable and robust Data Science solutions

A Developer's perspective  
A Developer's RANT

# About The Author

- Bachelor in Comp Sc
- Worked as a Web developer for several years
- Master in Information management systems (With DS as major)
- Currently working as Sr Manager of Analytics in ABInBev

## • Work Ex



## • Education



# Before we start, let's look at reddit -

*"the DS's poor excuse for a pipeline The DS pipeline: run jupyter notebook 1, copy the result and paste it into notebook 2 as a variable."*

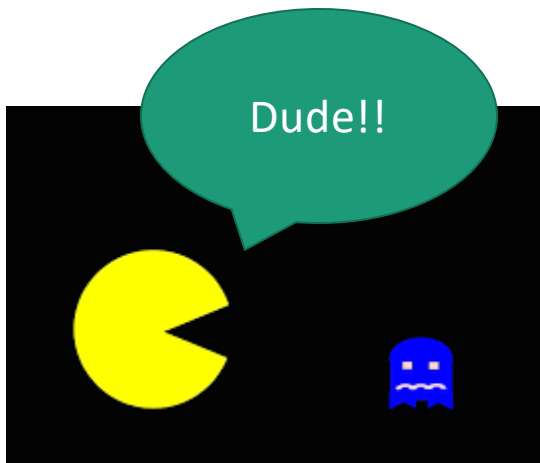
*"My org is very similar but the ML and Eng teams are also separate. Also the ML people are called "data scientists" and the DS people are called "decision scientists" so it's not at all confusing.."*

*"Wtf man. I swear that whenever I ask them to push the code to git and I'll take a look. They literally push a notebook."*

*"Fortunately our Scientist bands are higher than Engineering bands, but we hire mostly PhDs who can all code to some degree, and code as well as software engineers sometimes."*

# No Ankit, that happens very rarely...

*"We are working on a Next Gen AI that would make Skynet look like Pacman"*



```
1 {
2   "cells": [
3     {
4       "cell_type": "code",
5       "execution_count": null,
6       "metadata": {},
7       "outputs": [],
8       "source": [
9         "## Import the dependencies\n",
10        "import tabulate\n",
11        "import pandas as pd\n",
12        "import numpy as np\n",
13        "import itertools\n",
14        "import re\n",
15        "import datetime\n",
16        "import matplotlib.pyplot as plt\n",
17        "import math\n",
18        "from sklearn import datasets, linear_model, metrics\n",
19        "import seaborn as sns\n",
20        "from sklearn.ensemble import RandomForestRegressor\n",
21        "from sklearn.linear_model import LinearRegression\n",
22        "import statsmodels.api as sm\n",
23        "from scipy import stats\n",
24        "from azure.storage.blob import *\n",
25        "from io import StringIO"
26      ]
27    },
28    {
29      "cell_type": "code",
30      "execution_count": 4,
31      "metadata": {},
```

# Ankit, stop being a counterproductive bunch...

## Problem -

*"We want to use an OS python library but customize around 5 lines of the code"*

*-- New joiners in the team*

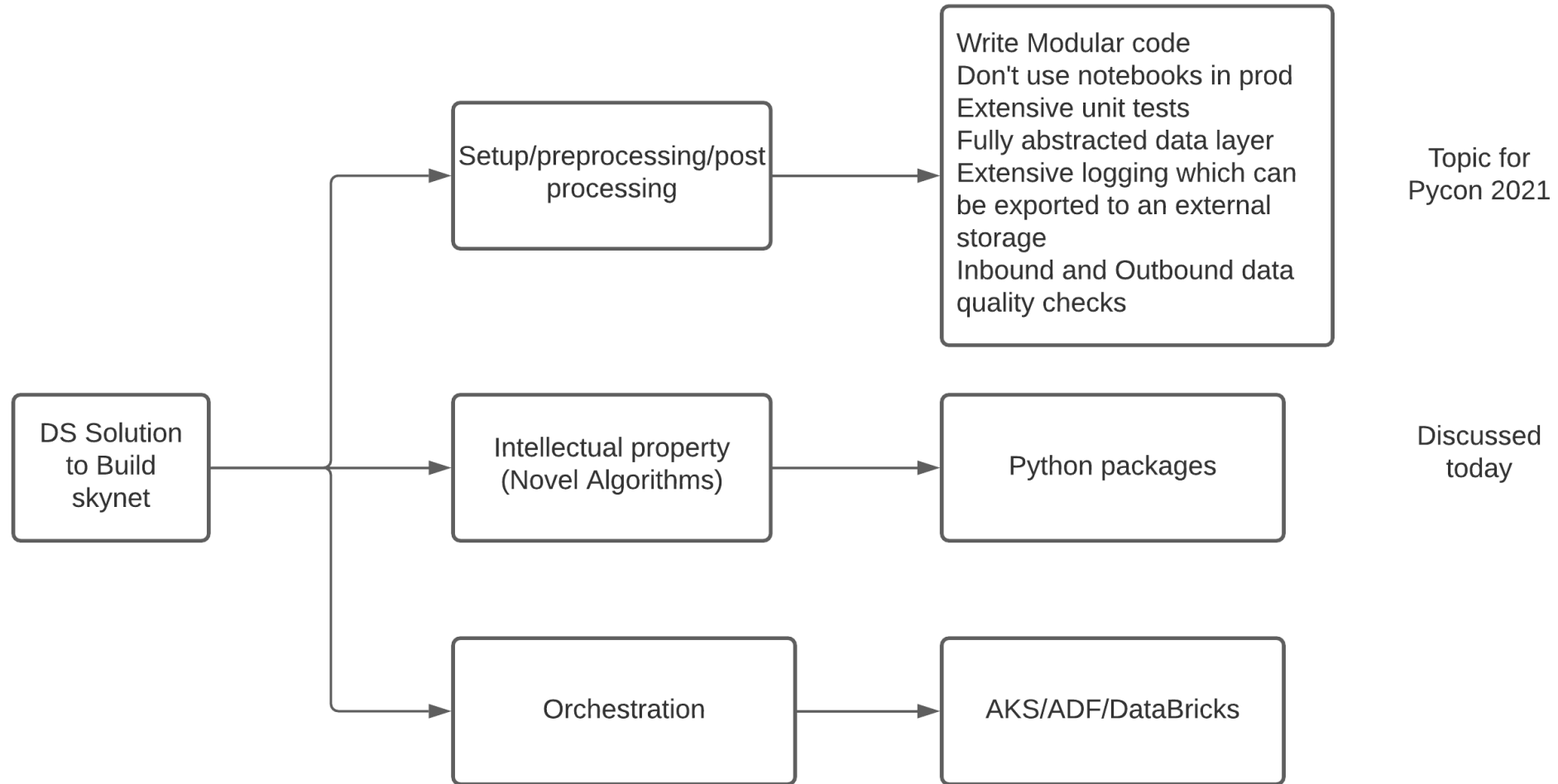
## Solution -

*"Take 500 lines of code out from the upstream package relevant to us, copy and paste those into a notebook, change the five lines that you need"*

**You ask why?**

*Cause "results" and "ease" of getting there are important*

# Ankit, Why don't you tell us what to do?



# Building packages? Isn't that simple?

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
X, y = make_classification(n_samples=1000, n_features=4,
                          n_informative=2, n_redundant=0,
                          random_state=0, shuffle=False)
clf = RandomForestClassifier(max_depth=2, random_state=0)
clf.fit(X, y)
print(clf.predict([[0, 0, 0, 0]]))
```

# Let's break it down..

## **User View -**

- Clean APIs for the user
- Excellent examples and documentations
- Reliable output
- Readability
- Stable APIs (Backwards compatibility)

## **Design View -**

- Architecture that is easy to develop and future compatible
- Coding standards – pep8/484
- Development that treats backward compatibility highly
- Extensive Unit tests
- Tests that go beyond Unit tests - How do we reliably test a process that has randomness ?

## **Process View -**

- Release cadence
- Licenses
- Bug reporting/contributing
- Announcements

## **Infra View -**

- Platforms supported
- Private/public package
- Masking code
- Sharing mechanism



# Building the package

```
setup(  
    ext_modules=generate_extensions(),  
    name="baby_skynet",  
    version="1.0",  
    author="Arnold",  
    author_email="arnold@has.to.come",  
    description="A package for reaching Transcendence",  
    long_description=long_description,  
    long_description_content_type="text/markdown",  
    url="https://github.com/ankitagarwal/pycon_2020",  
    license="GPL3",  
    packages=[],  
    classifiers=[  
        "Programming Language :: Python :: 3.7",  
        "Intended Audience :: Developers",  
        "License :: Other/Proprietary License",  
        "Operating System :: Microsoft :: Windows",  
        "Operating System :: LINUX",  
        "Operating System :: Apple :: MacOS",  
        "Topic :: Scientific/Engineering",  
        "Topic :: Scientific/Engineering :: Mathematics",  
        "Topic :: Scientific/Engineering :: Artificial Intelligence",  
        "Topic :: Software Development :: Libraries :: Python Modules",  
    ],  
    platforms=["Windows", "Linux", "MacOS"],  
    test_suite="pytest",  
    python_requires='>=3.7',  
    data_files=[  
        ('baby-skynet', ['README.md']),  
        ('baby-skynet', ['requirements.txt']),  
    ],  
)
```

Update *.pypirc* file in home Directory with the following

```
[distutils]
index-servers = local
[local]
repository: <URL>
username: <USERNAME>
password: <PASSWORD>
```



Handling  
Authentication  
against Private  
Artifactory Server

As a wheel -

***python setup.py bdist\_wheel upload -r local***

With source -

***python setup.py sdist upload -r local***

Deploying a  
version of the  
package



## Option 1

```
pip install baby_skynet --extra-index-  
url=https://user:token@<server>
```

## Option 2

Create a .netrc file with following content

**Machine <server>**

**Login <username>**

**Password <token>**

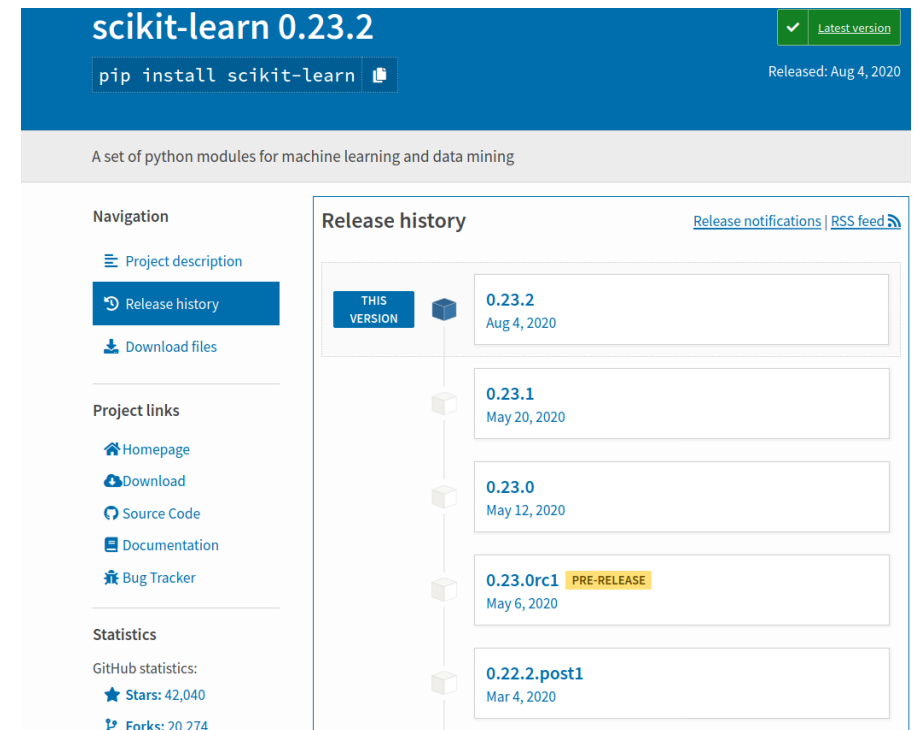
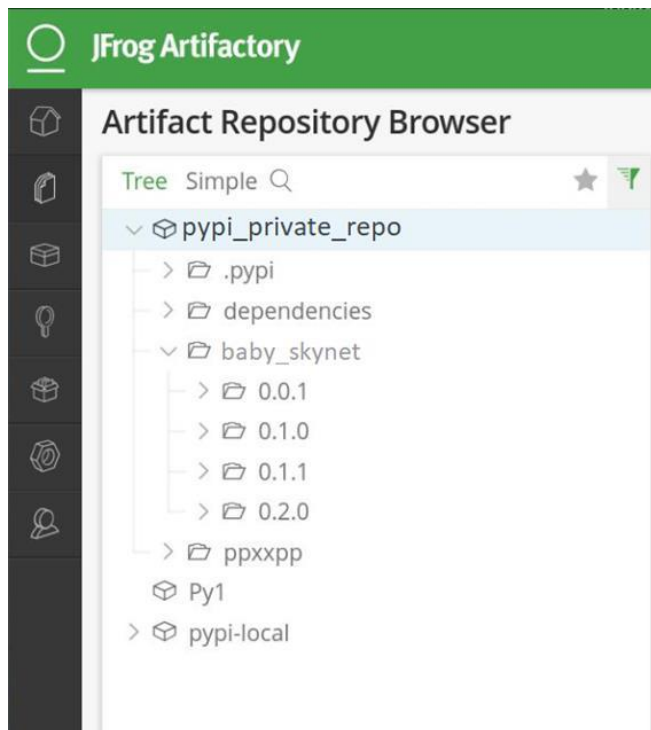
```
pip install baby_skynet --extra-index-url=<server>
```



Installing the  
package

Finally we are there..

```
Python 3.7.6 (default, Jan 8 2020, 19:59:22)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> from baby_skynet import Skynet
>>> Skynet()
Skynet is still a baby and knows nothing, please build it more.
<baby_skynet.Skynet.Skynet object at 0x7f7fb50ec810>
>>>
```



# Link to the materials used in this presentation

[Github](#)

Slides will be added by  
EOD today.

