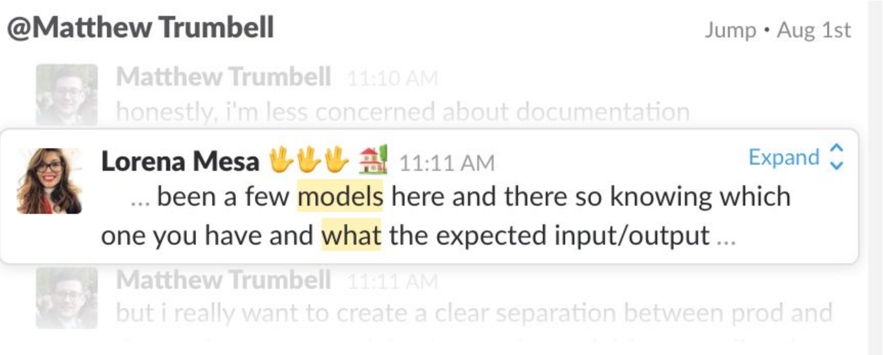
Using Django, Docker and Scikit-learn to bootstrap your Machine Learning Project

Lorena Mesa @loooorenanicole

sproutsocial

http://bit.ly/2wMv2RC

Something that happened recently …

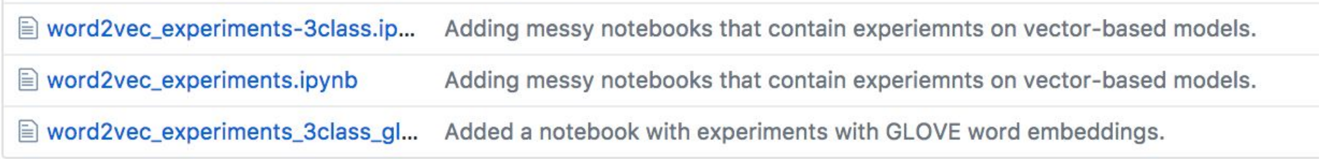


@Matthew Trumbell — Jump • Aug 1st

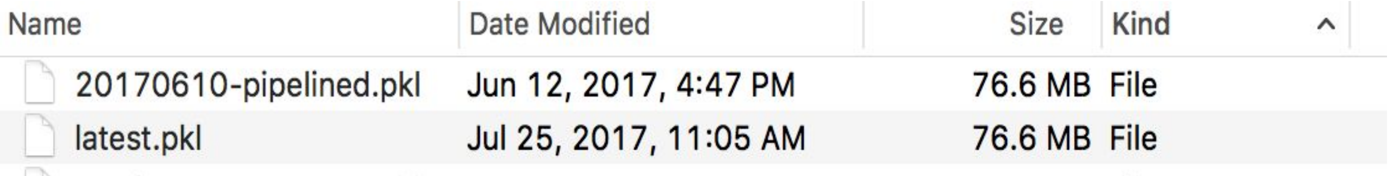
Matthew Trumbell 11:10 AM
honestly, i'm less concerned about documentation

Lorena Mesa 🖖🖖🖖 🏡 11:11 AM
… been a few models here and there so knowing which one you have and what the expected input/output …

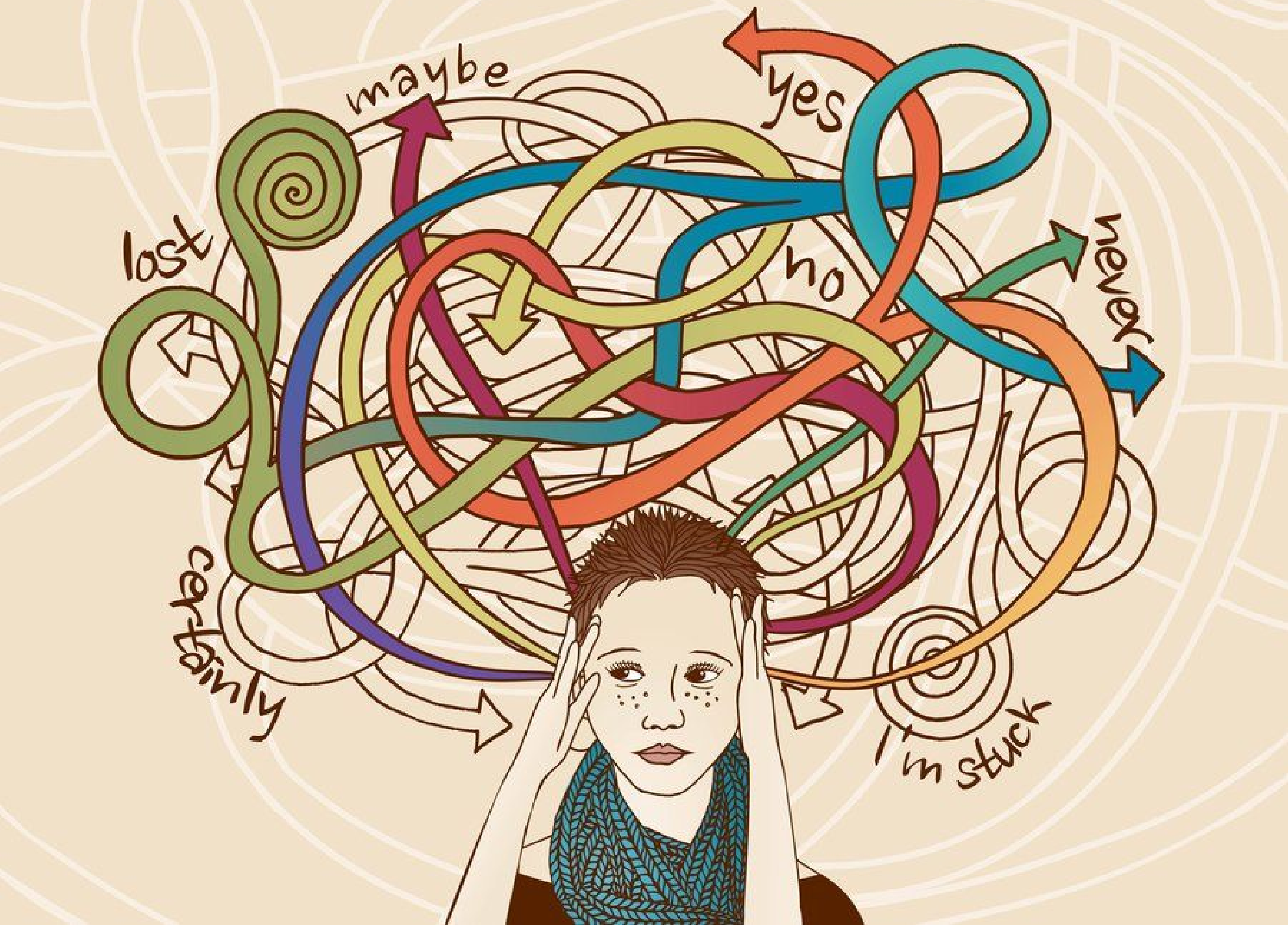
Matthew Trumbell 11:11 AM
but i really want to create a clear separation between prod and



word2vec_experiments-3class.ip...	Adding messy notebooks that contain experiemnts on vector-based models.
word2vec_experiments.ipynb	Adding messy notebooks that contain experiemnts on vector-based models.
word2vec_experiments_3class_gl...	Added a notebook with experiments with GLOVE word embeddings.

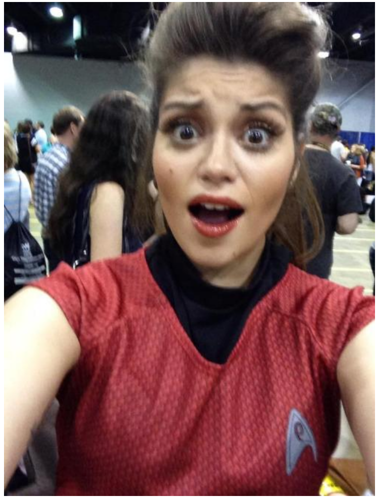


Name	Date Modified	Size	Kind
20170610-pipelined.pkl	Jun 12, 2017, 4:47 PM	76.6 MB	File
latest.pkl	Jul 25, 2017, 11:05 AM	76.6 MB	File


maybe
yes
lost
no
never
certainly
i'm stuck

Hello, I’m Lorena Mesa.

TECH LADIES



In reply to Đave Hoover

Lorena Mesa @loooorenanicole · 7h

@davehoover and we are super thankful!!!! Look at these Squirrels 2014!! #dbc @devbootcamp

SYSTERS

AN ANITA BORG INSTITUTE COMMUNITY

<write/speak/code>

python SOFTWARE FOUNDATION

pyladies CHICAGO

sproutsocial

http://bit.ly/2wMv2RC

How I'll approach today's chat

Review of machine learning

●

Anatomy of a data science team

●

Engineering a machine learning problem

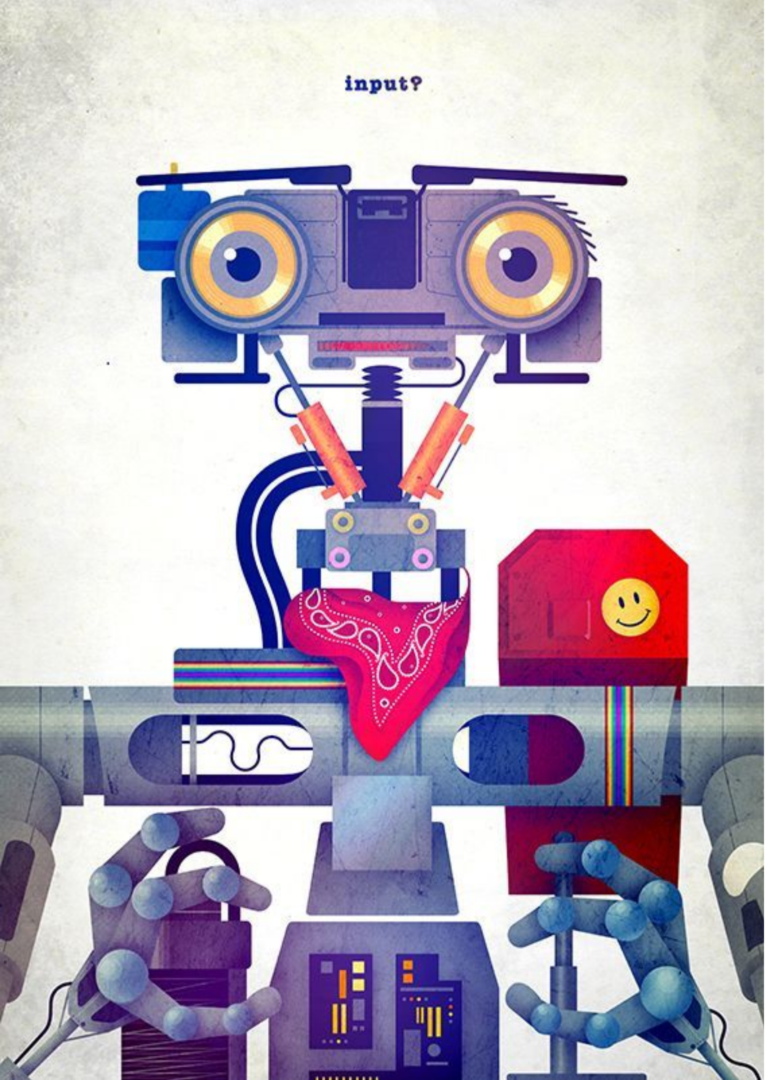
●

Iterating on machine learning engineering with Docker, Django, and scikit-learn (sklearn)

●

What next?

What is machine learning?



Machine Learning

is a subfield of computer science [that] stud[ies] pattern recognition and computational learning [in] artificial intelligence. [It] explores the construction and study of ***algorithms*** that can learn from and ***make predictions on data***.

sproutsocial

Machine Learning, another definition

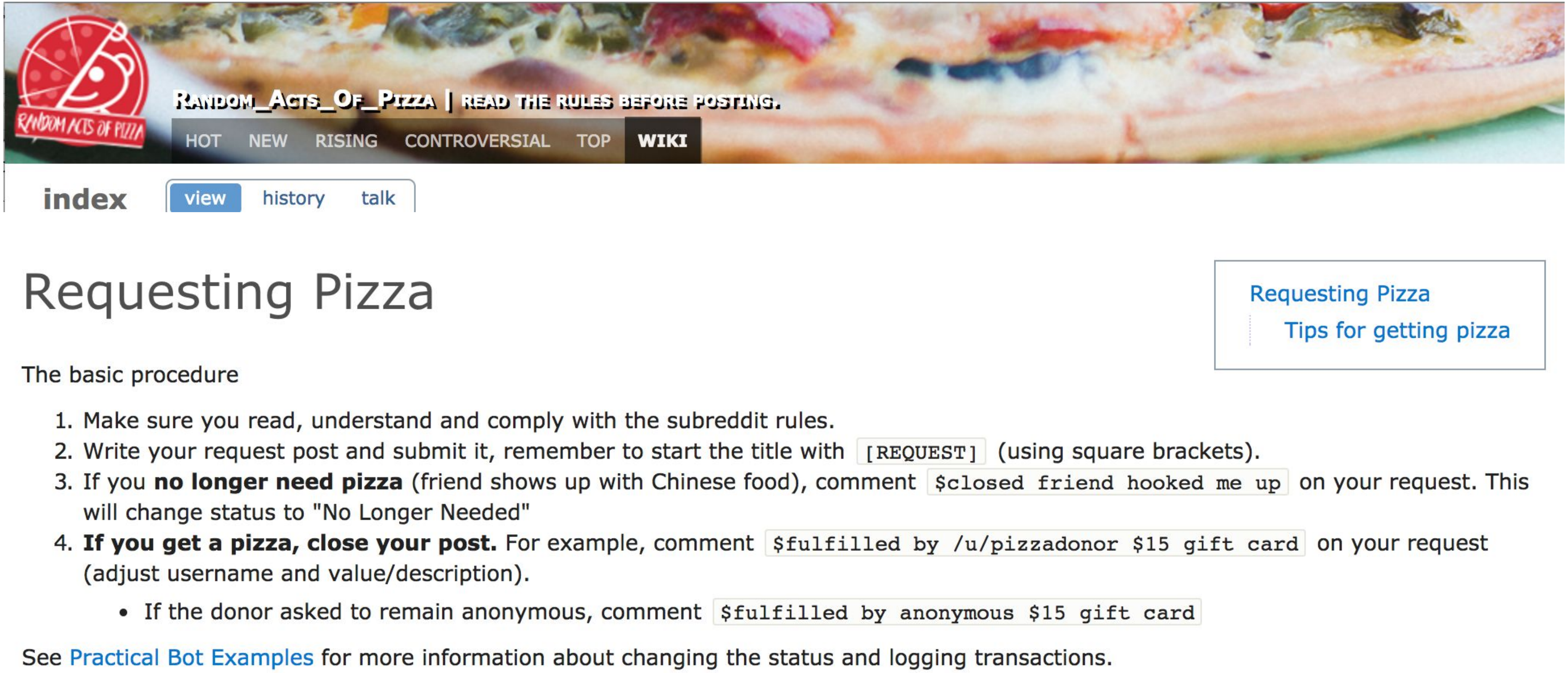
A computer program is **said to learn** from **experience** (E) with respect to some **task** (T) and some performance **measure** (P), if its performance on T, as measured by P, improves with experience E.

(Ch. 1 - Machine Learning Tom Mitchell)

http://bit.ly/2wMv2RC

Example Project:
Predicting Altruism with a Naive Bayes Classifier

FREE



Requesting Pizza

The basic procedure

1. Make sure you read, understand and comply with the subreddit rules.
2. Write your request post and submit it, remember to start the title with `[REQUEST]` (using square brackets).
3. If you **no longer need pizza** (friend shows up with Chinese food), comment `$closed friend hooked me up` on your request. This will change status to "No Longer Needed"
4. **If you get a pizza, close your post.** For example, comment `$fulfilled by /u/pizzadonor $15 gift card` on your request (adjust username and value/description).
    - If the donor asked to remain anonymous, comment `$fulfilled by anonymous $15 gift card`

See Practical Bot Examples for more information about changing the status and logging transactions.

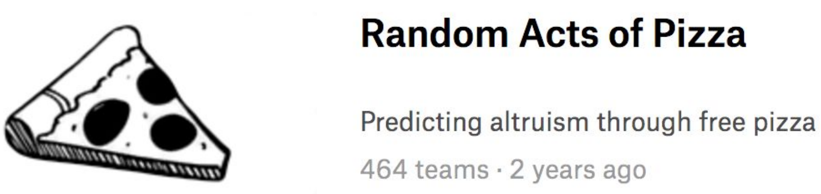
Free acts of pizza, a Reddit subreddit

http://bit.ly/2wMv2RC

I actually have money for the pizza but all I have is a 50 dollar bill, and the delivery boys don't accept anything larger than a 20 dollar bill.

I've got a guitar in one hand and the other is a pizza pie?

Free acts of pizza



Random Acts of Pizza

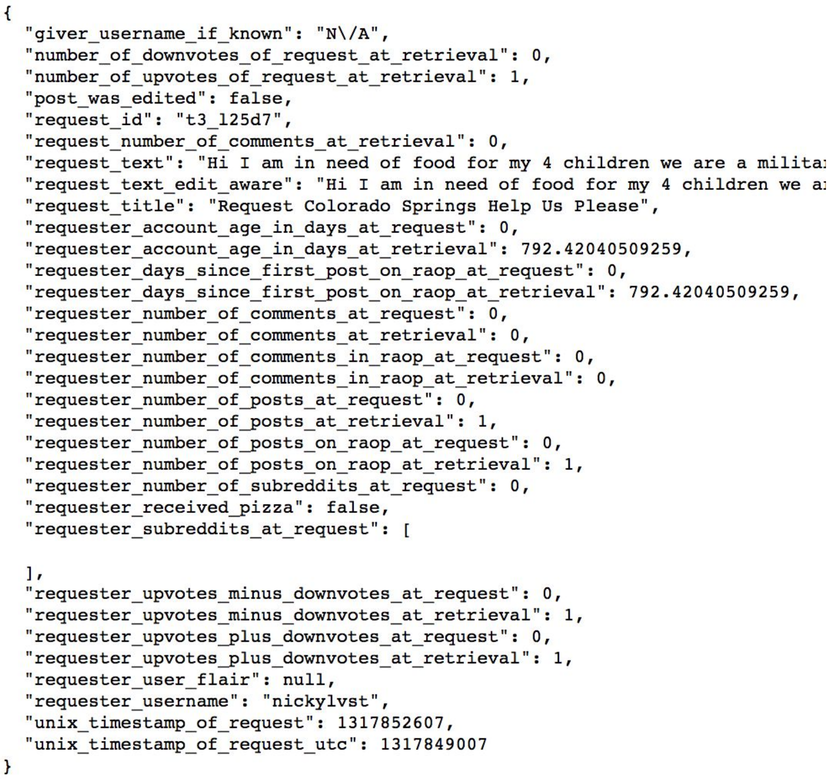
Predicting altruism through free pizza

464 teams · 2 years ago

Training data contains:

- 5671 requests
- Successful (994) labelled as True
- Unsuccessful (3046) labelled as False.

Unlabeled data has 1631 requests.



```
{
  "giver_username_if_known": "N\/A",
  "number_of_downvotes_of_request_at_retrieval": 0,
  "number_of_upvotes_of_request_at_retrieval": 1,
  "post_was_edited": false,
  "request_id": "t3_l25d7",
  "request_number_of_comments_at_retrieval": 0,
  "request_text": "Hi I am in need of food for my 4 children we are a milita
  "request_text_edit_aware": "Hi I am in need of food for my 4 children we a
  "request_title": "Request Colorado Springs Help Us Please",
  "requester_account_age_in_days_at_request": 0,
  "requester_account_age_in_days_at_retrieval": 792.42040509259,
  "requester_days_since_first_post_on_raop_at_request": 0,
  "requester_days_since_first_post_on_raop_at_retrieval": 792.42040509259,
  "requester_number_of_comments_at_request": 0,
  "requester_number_of_comments_at_retrieval": 0,
  "requester_number_of_comments_in_raop_at_request": 0,
  "requester_number_of_comments_in_raop_at_retrieval": 0,
  "requester_number_of_posts_at_request": 0,
  "requester_number_of_posts_at_retrieval": 1,
  "requester_number_of_posts_on_raop_at_request": 0,
  "requester_number_of_posts_on_raop_at_retrieval": 1,
  "requester_number_of_subreddits_at_request": 0,
  "requester_received_pizza": false,
  "requester_subreddits_at_request": [

  ],
  "requester_upvotes_minus_downvotes_at_request": 0,
  "requester_upvotes_minus_downvotes_at_retrieval": 1,
  "requester_upvotes_plus_downvotes_at_request": 0,
  "requester_upvotes_plus_downvotes_at_retrieval": 1,
  "requester_user_flair": null,
  "requester_username": "nickylvst",
  "unix_timestamp_of_request": 1317852607,
  "unix_timestamp_of_request_utc": 1317849007
}
```

http://bit.ly/2wMv2RC

Task:
Classify a piece of data

Is a pizza request successful?
Is it altruistic or not?

Experience:
Labeled training data

Request_id | No
Request_id | Yes

Performance Measurement:
Is the label correct?

Verify if the request is successful or not

Anatomy of a Data Science Team

sproutsocial

IBM UX Personas Applied to Engineering

" What I seek as a developer is a chance to think about problems and to stretch the bounds of my understanding and skills. "
" What motivates me is what I can do with data to impact a population and make a change. "
" The look on people's faces, when you deliver data sets that they need, is a great feeling. "

App Developer

- **Kai, the Application Developer,** builds applications that interact with data and implements data models.

Data Scientist

- **Chris, the Data Scientist** gets deep into the data to draw hidden insights and influence business decisions.

Data Engineer

- **Harley, the Data Engineer,** builds and maintains a scalable data infrastructure to make relevant data available to teams.

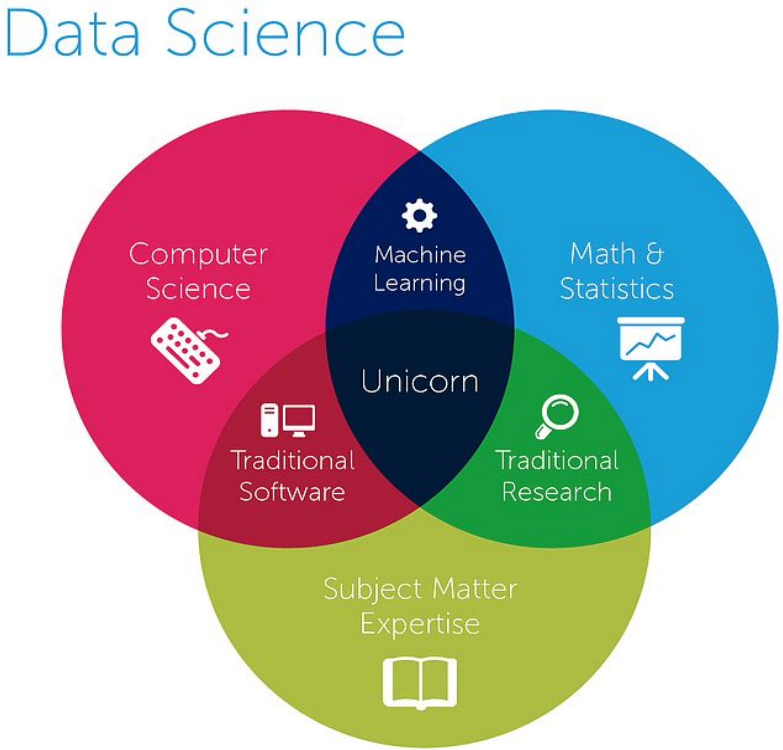
http://bit.ly/2wMv2RC

Data Science Teams

Complementary skill sets, for example consider my team:

- (4) Data scientists: PhD Natural Language Processing, Predictive Analytics, Economics
- (1) Software engineer: historically platform engineering and data analyst
- Designated Infrastructure support

http://bit.ly/2wMv2RC



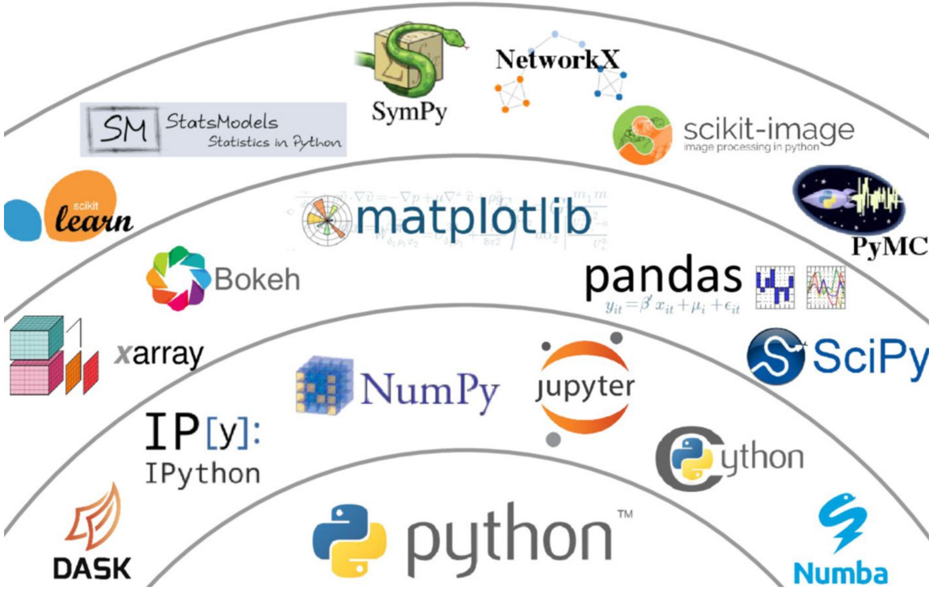
Engineering a Machine Learning Project

Python Tools Used by Data Scientists

- Executable code + analysis environments: Jupyter notebooks
- Machine learning: sklearn
- Database: DataGrip, or another database IDE
- Data analysis: Pandas
- Plotting: matplotlib, bokeh
- Data visualization: seaborn

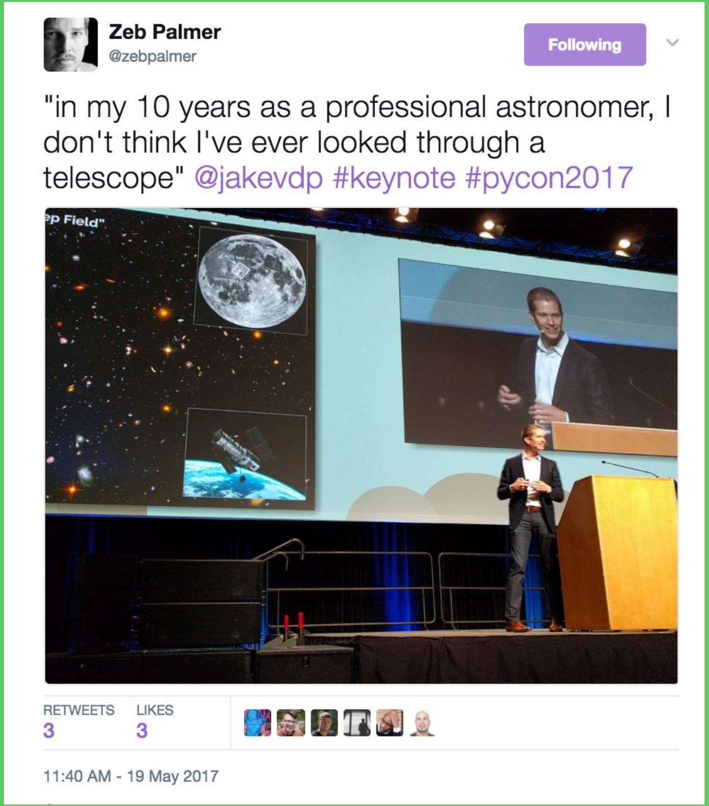
http://bit.ly/2wMv2RC

Python's Scientific Stack


SymPy
NetworkX
SM StatsModels Statistics in Python
scikit-image image processing in python
scikit learn
matplotlib
PyMC
Bokeh
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
xarray
NumPy
jupyter
SciPy
IP[y]: IPython
Cython
DASK
python™
Numba

Jake VanderPlas, PyCon 2017 keynote

sproutsocial



Zeb Palmer
@zebpalmer

"in my 10 years as a professional astronomer, I don't think I've ever looked through a telescope" @jakevdp #keynote #pycon2017

RETWEETS 3 LIKES 3

11:40 AM - 19 May 2017

Python adoption in science community

- Python is glue (plays well with other languages)
- “Batteries included” + 3rd party modules
- Simple + dynamic
- Open ethos is well suited to science

> “Before ~~software~~ machine learning can be usable, it must first be reusable.

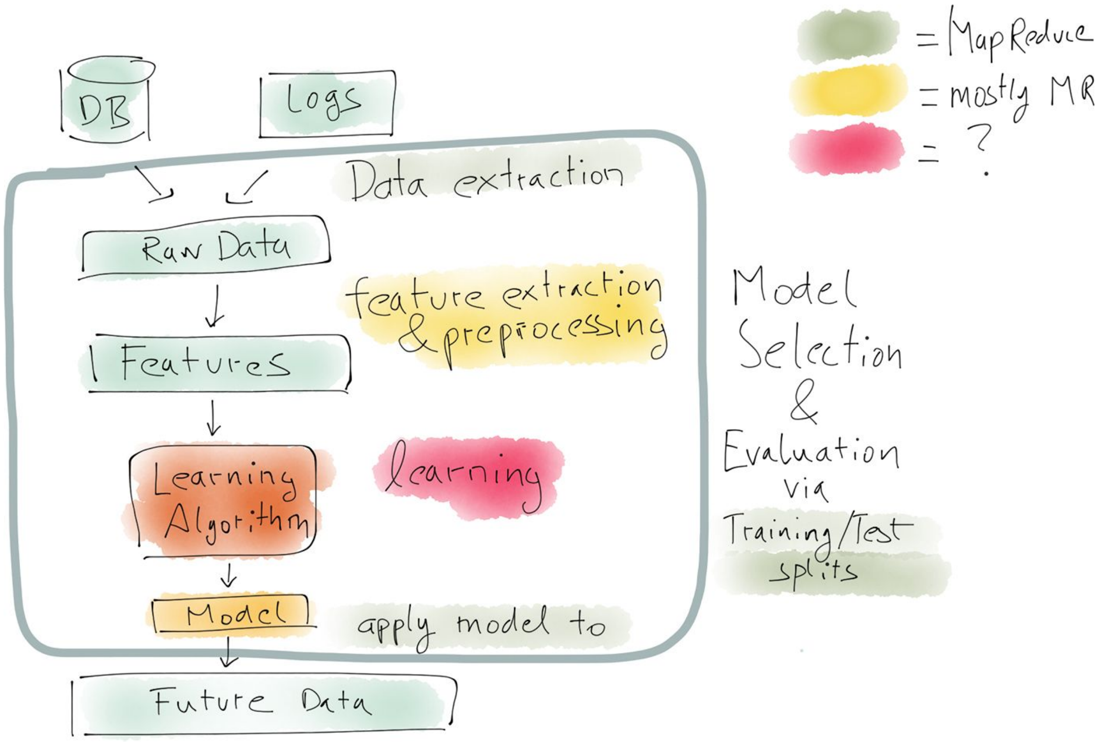
(modified) **Ralph Johnson**

UIUC Computer Science

Design Patterns: Elements of Reusable Object-Oriented Software

sproutsocial

Simple Machine Learning Pipeline

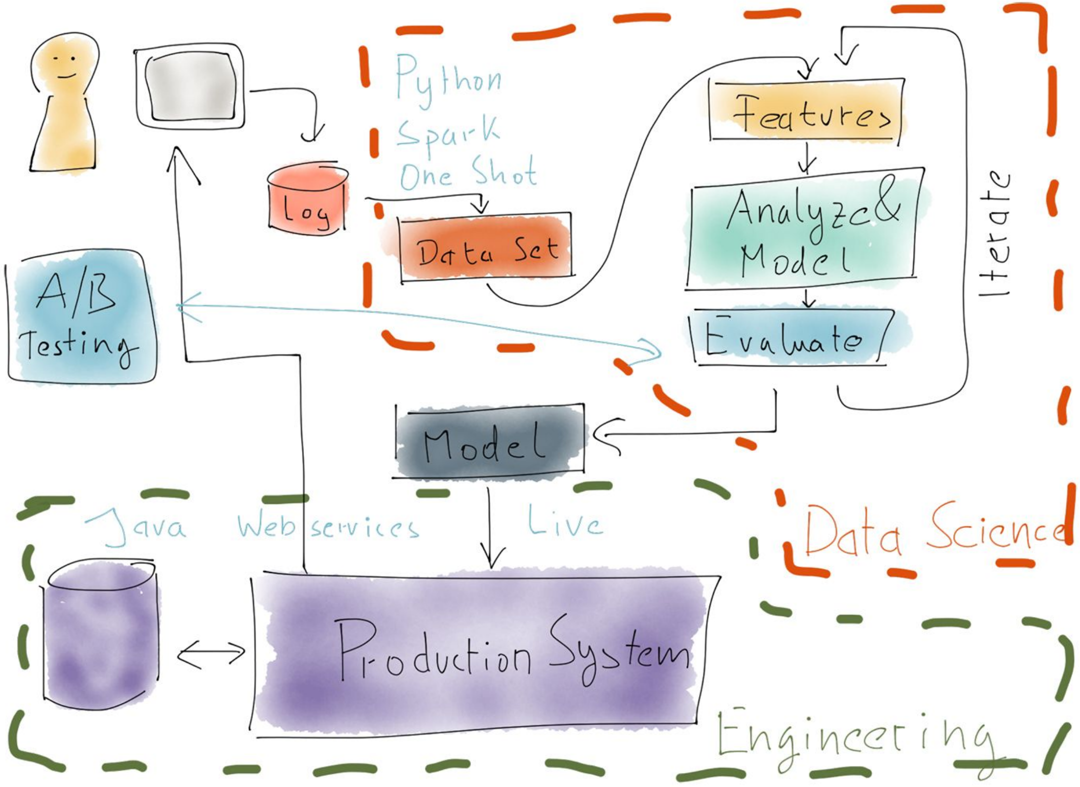


http://bit.ly/2wMv2RC

Feature engineering is expensive, it takes time to:

- Shape the data
- Select which features to use
- Collect data!

Simple Machine Learning Pipeline



http://bit.ly/2wMv2RC

Data science is fundamentally embedded within a different system from production

What is the handoff between data science and production?

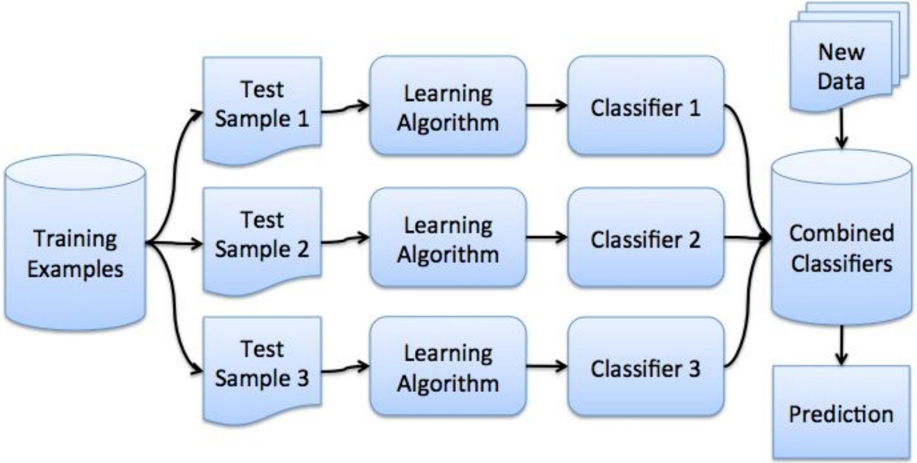
sproutsocial

Simplified Machine Learning Project

1. Get and shape the data
2. Train the model on the data
3. Pickle the model, save with joblib
4. User it! Predict on the data

```
from sklearn.naive_bayes import MultinomialNB

X, y = get_xy()
X_train, X_test, y_train, y_test =
train_test_split(X, y, random_state=1111)

model = MultinomialNB().fit(X_train, y_train)
filename = 'pizza_classifier_latest.pkl'
pickle.dump(model, open(filename, 'wb'))
```

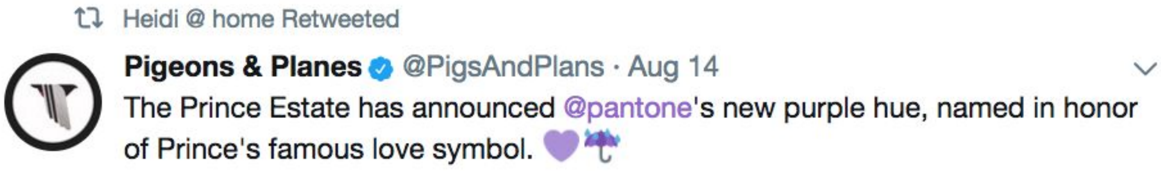
You can use sklearn pipelines to apply transformations with scoring indicators as well



http://bit.ly/2wMv2RC

sproutsocial

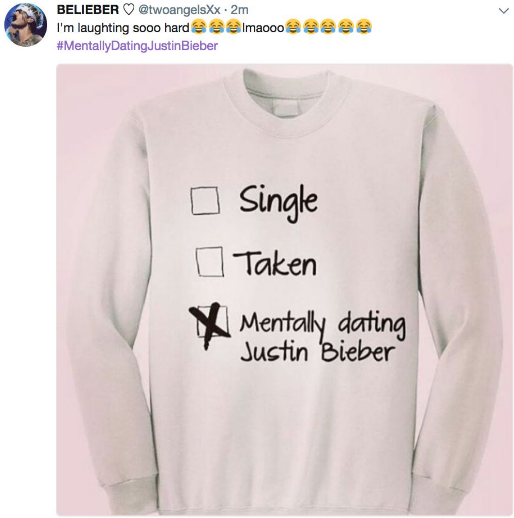
Reproducibility matters.
How do we engineer for that?



Heidi @ home Retweeted

Pigeons & Planes @PigsAndPlans · Aug 14

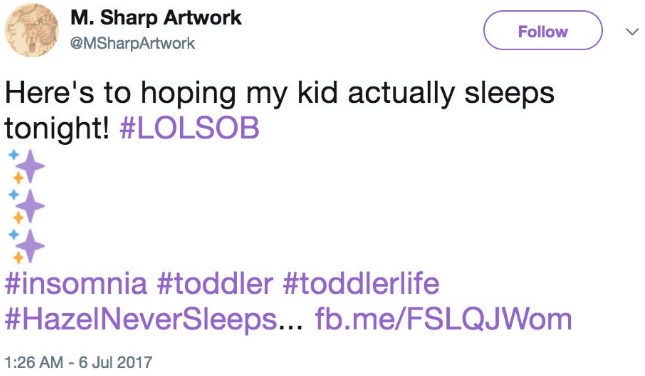
The Prince Estate has announced @pantone's new purple hue, named in honor of Prince's famous love symbol. 💜☔



BELIEBER ♡ @twoangelsXx · 2m

I'm laughing sooo hard😂😂😂lmaooo😂😂😂😂😂 #MentallyDatingJustinBieber

#LOLSOB



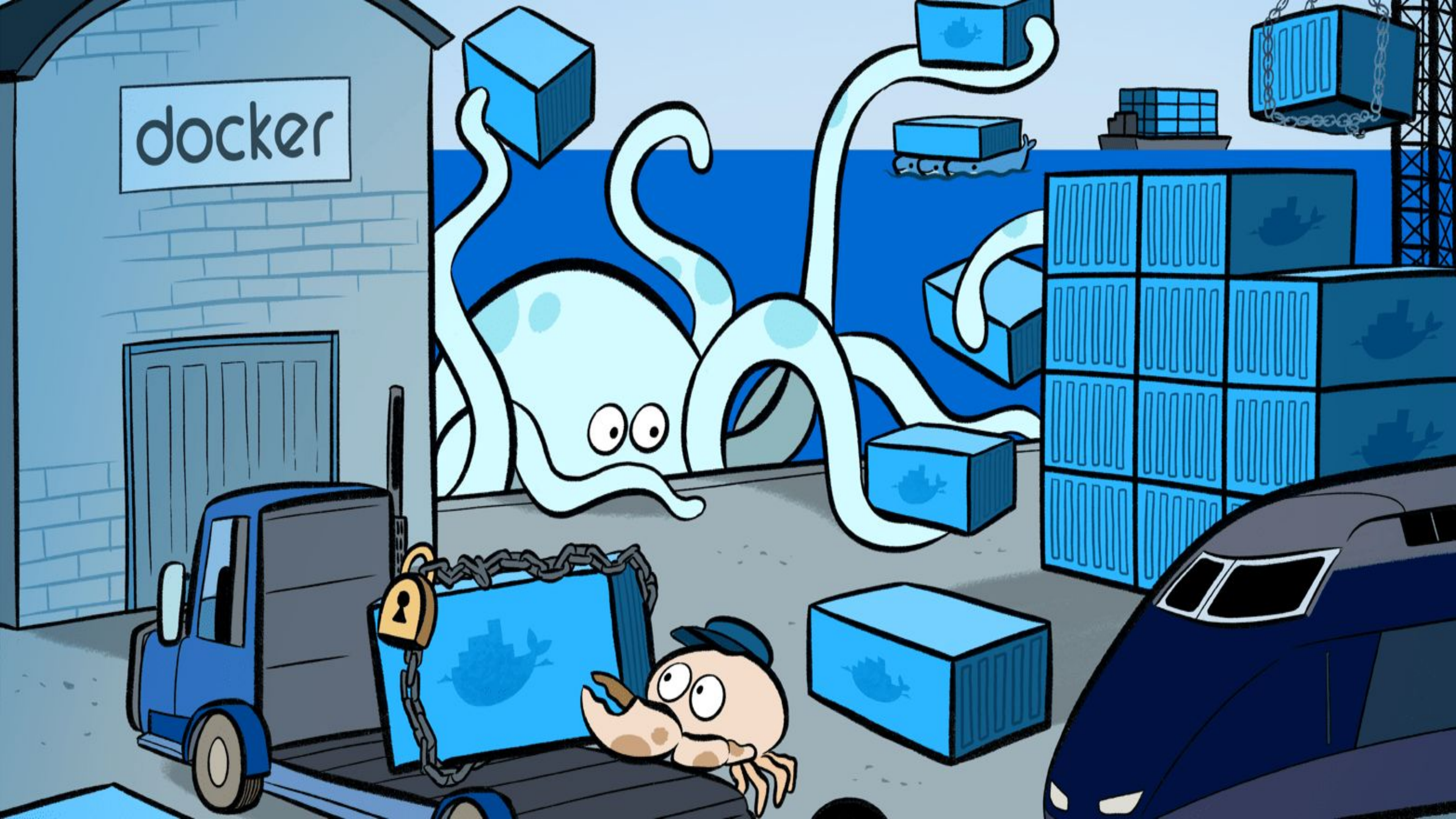
M. Sharp Artwork
@MSharpArtwork

Here's to hoping my kid actually sleeps tonight! #LOLSOB
✨
✨
✨
#insomnia #toddler #toddlerlife #HazelNeverSleeps... fb.me/FSLQJWom

1:26 AM - 6 Jul 2017

http://bit.ly/2wMv2RC

http://bit.ly/2wMv2RC


docker

Docker

Docker containers are a big executable tarball (with explicit format) that includes everything needed to run it: code, system tools, libraries, settings!

Also, according to Kelsey Hightower, ***"the first rule of Python is you don't use the system installed version of Python"***

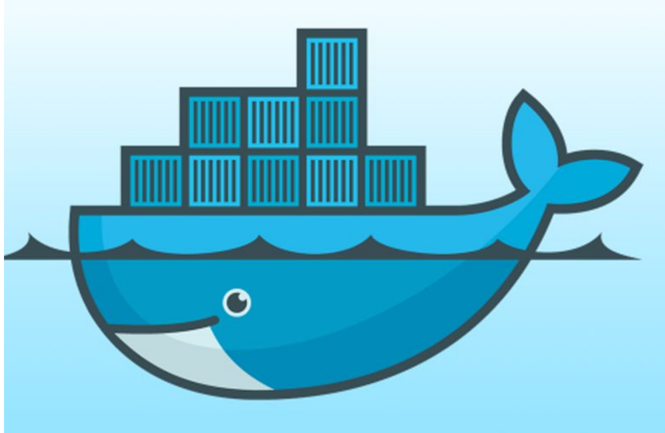
Step 1: Write a Dockerfile (cached layers)

Step 2: Build the Docker image

```
    docker build -t 'predicting-altruism:latest' .
```

Step 3: Run the Docker image in a container

```
    docker run -d -ti -p 8888:8888 -v
    ~/local_path/to/notebooks:/home/jupyter/notebooks
    predicting-altruism .
```



http://bit.ly/2wMv2RC

Example Dockerfile

```
FROM python:3
RUN pip install virtualenv

RUN useradd jupyter
RUN adduser jupyter sudo
RUN mkdir /home/jupyter/
ADD entrypoint.sh /home/jupyter/
RUN chmod +x /home/jupyter/entrypoint.sh
ADD requirements.txt /home/jupyter/
ADD notebooks/ /home/jupyter/notebooks

RUN chown jupyter:jupyter /home/jupyter/
VOLUME ["/home/jupyter/notebooks"]

WORKDIR /home/jupyter

RUN virtualenv myenv && pip install -r /home/jupyter/requirements.txt
ENV SHELL=/bin/bash
ENV USER=jupyter
EXPOSE 8888:8888

ENTRYPOINT ["/bin/bash", "/home/jupyter/entrypoint.sh"]
```

http://bit.ly/2wMv2RC

Model versioning

Docker + Volumes!

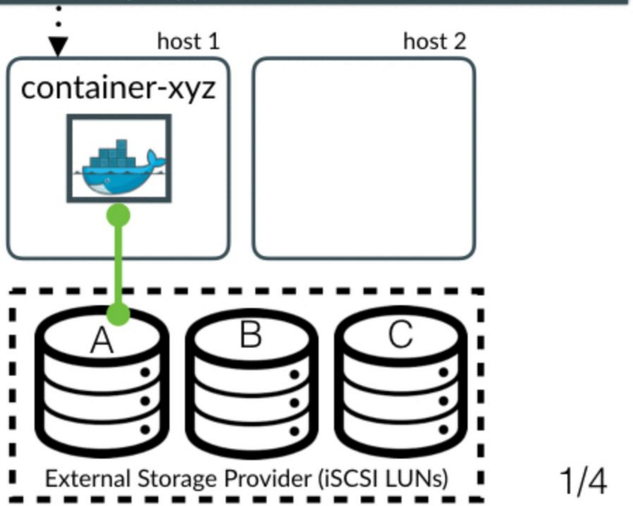
Docker volumes allow a mountable data directory

```
ADD notebooks/ /home/jupyter/notebooks

...

VOLUME ["/home/jupyter/notebooks"]
```

Whenever the data scientist and/or other team member is ready to save the code, data, & model any work done in the Docker image updates the mountable data volume

Create and Run Container on Host 1

```
$ docker run -d -v /host/data:/data --volume-driver=flocker
—name=container-xyz app
```



http://bit.ly/2wMv2RC

sproutsocial

Django-izing Docker + sklearn model

Process for updating a model

- Create a Dockerfile with a mountable data volume
- Add the Dockerfile to a Django API
- Add Jupyter notebook into the mountable data volume in the Django API
- Call the build endpoint to build the new Docker image
- Spin up Docker container
- Data science magic
- Save the model
- Update the model to wherever it needs to live for productionalizing it

Wrap docker-py into Django endpoint

```
from docker import APIClient
from io import BytesIO

def create_image(request, model):
  if not request.POST.get('path'):
      path = BASE_DIR
  else:
      path = request.POST[‘path’]

try:
      with open(path, 'r') as d:
          dockerfile = [x.strip() for x in d.readlines()]
          dockerfile = '\n'.join(dockerfile)

      f = BytesIO(dockerfile.encode('utf-8'))

      # Point to the Docker instance
      cli = APIClient(base_url='unix://var/run/docker.sock')
      response = [line for line in cli.build(
          fileobj=f, rm=True, tag=model + ‘:latest‘, stream=True
      )]

      return JsonResponse({'image': response})
  except:
      return HttpResponseServerError()
```

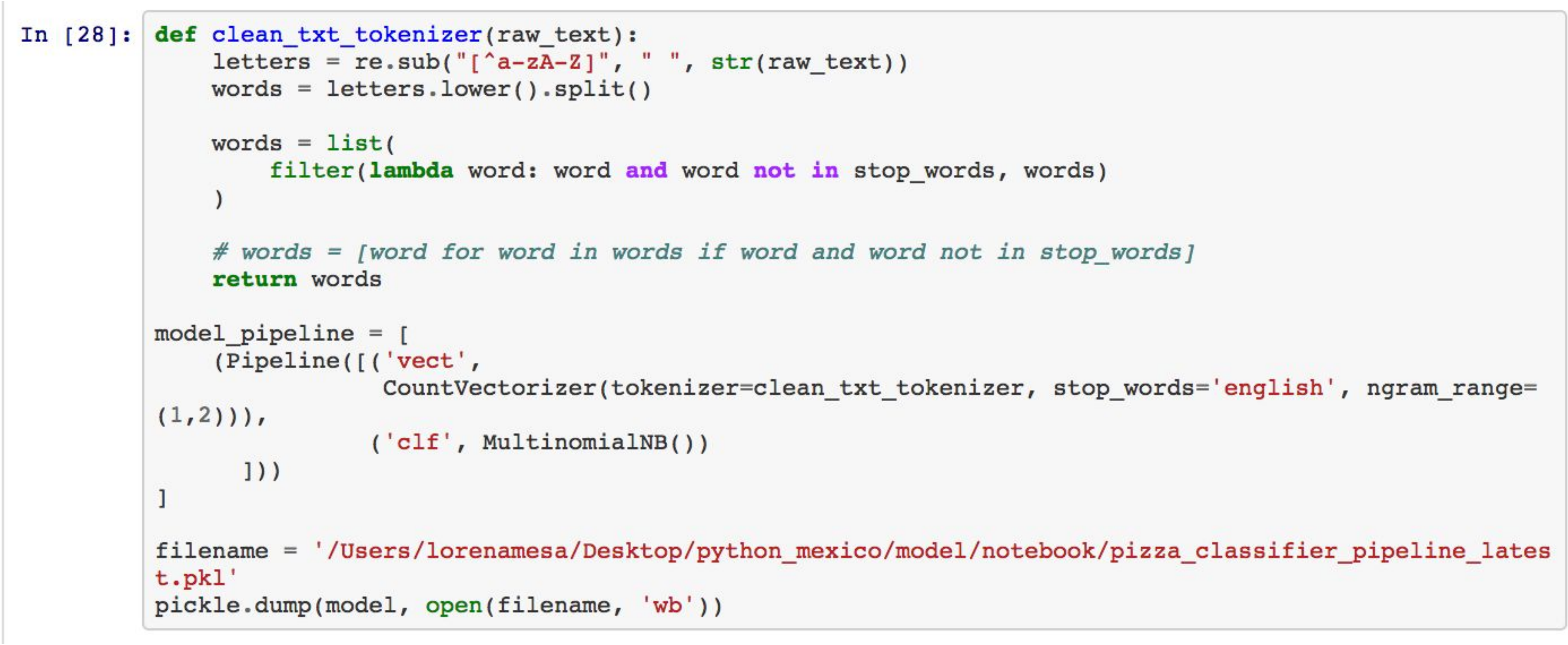
```
urlpatterns = [
   url(r'^create/image/(?P<model>\w{0,50})',
create_image, name='create_image'),
]
```

For more information on the Docker Python SDK reference the docs on the low level API here.

Entire demo code at here.

sproutsocial

Jupyter Notebook Pickles the Model



In [28]:

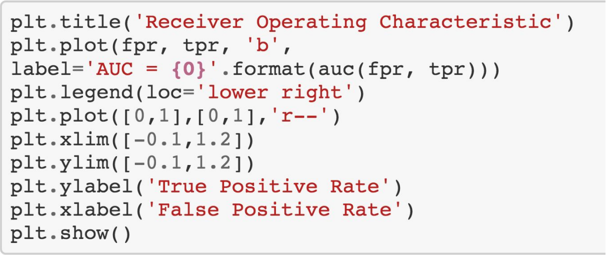
```
def clean_txt_tokenizer(raw_text):
    letters = re.sub("[^a-zA-Z]", " ", str(raw_text))
    words = letters.lower().split()

    words = list(
        filter(lambda word: word and word not in stop_words, words)
    )

    # words = [word for word in words if word and word not in stop_words]
    return words

model_pipeline = [
    (Pipeline([('vect',
                  CountVectorizer(tokenizer=clean_txt_tokenizer, stop_words='english', ngram_range=
(1,2))),
                ('clf', MultinomialNB())
        ]))
]

filename = '/Users/lorenamesa/Desktop/python_mexico/model/notebook/pizza_classifier_pipeline_lates
t.pkl'
pickle.dump(model, open(filename, 'wb'))
```

```
docker build -t ‘predict-altruism’ .
```

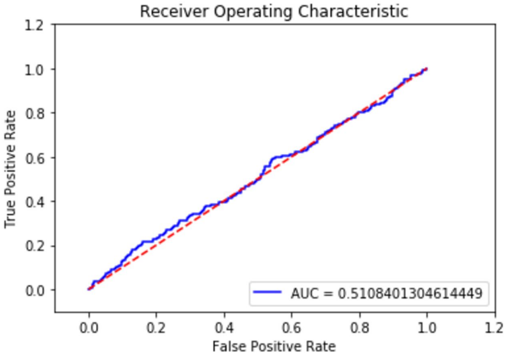
(Also known as … a POST request to your Django API. Yes, it can be that simple.)

What next?

- Store analytics on different versions performance on a fixed training set
- Surface analytics in user friendly dashboard via admin panel
- Integrate Django, Docker, scikit-learn project into Kubernetes (4:10pm - End to end Django on Kubernetes)
- Add CI to automate deployment of models
- Evaluate if Django is the best tool? (5:00pm - Django vs Flask)



```
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = {0}'.format(auc(fpr, tpr)))
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



Continued Learning

Talks:

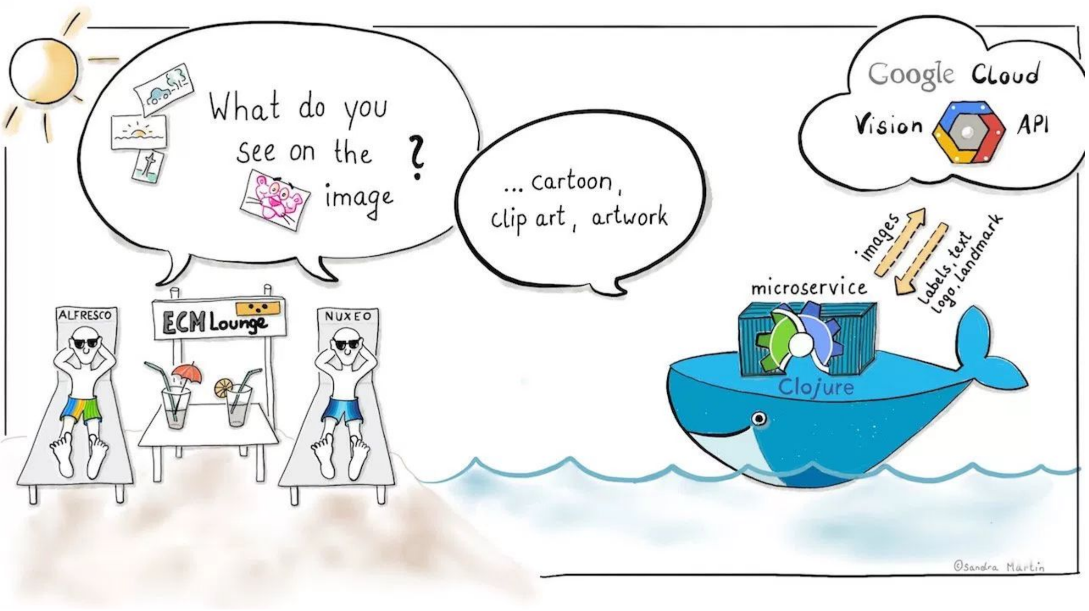
- Kelsey Highwater - PyCon 2017 closing keynote on Docker + Kubernetes
- Jake VanderPlas - The Python Visualization Landscape
- Kevin Goetsch - Deploying Machine Learning using sklearn pipelines
- Lorena Mesa - Predicting free Pizza with Python
- Rob Story - Bridging from Python to JVM

Books:

- Introduction to Machine Learning with Python, Sarah Guido, O'Reilly's

GitHub:

- Docker with Jupyter Notebook mountable volume
- Demo code from this talk!
- Docker-py (Read the Docs)


What do you see on the image ?
... cartoon, clip art, artwork
Google Cloud Vision API
images
labels, text logo, landmark
microservice
Clojure
ALFRESCO
ECM Lounge
NUXEO
©Sandra Martin

Thank you!

http://bit.ly/2s5R01V | @loooorenanicole