O'REILLY®

LIVE ONLINE TRAINING

Data Engineering for Data Scientists

Build resilient pipelines to support stronger models

February 6, 2020

1:00pm – 3:00pm EST

This course is full.

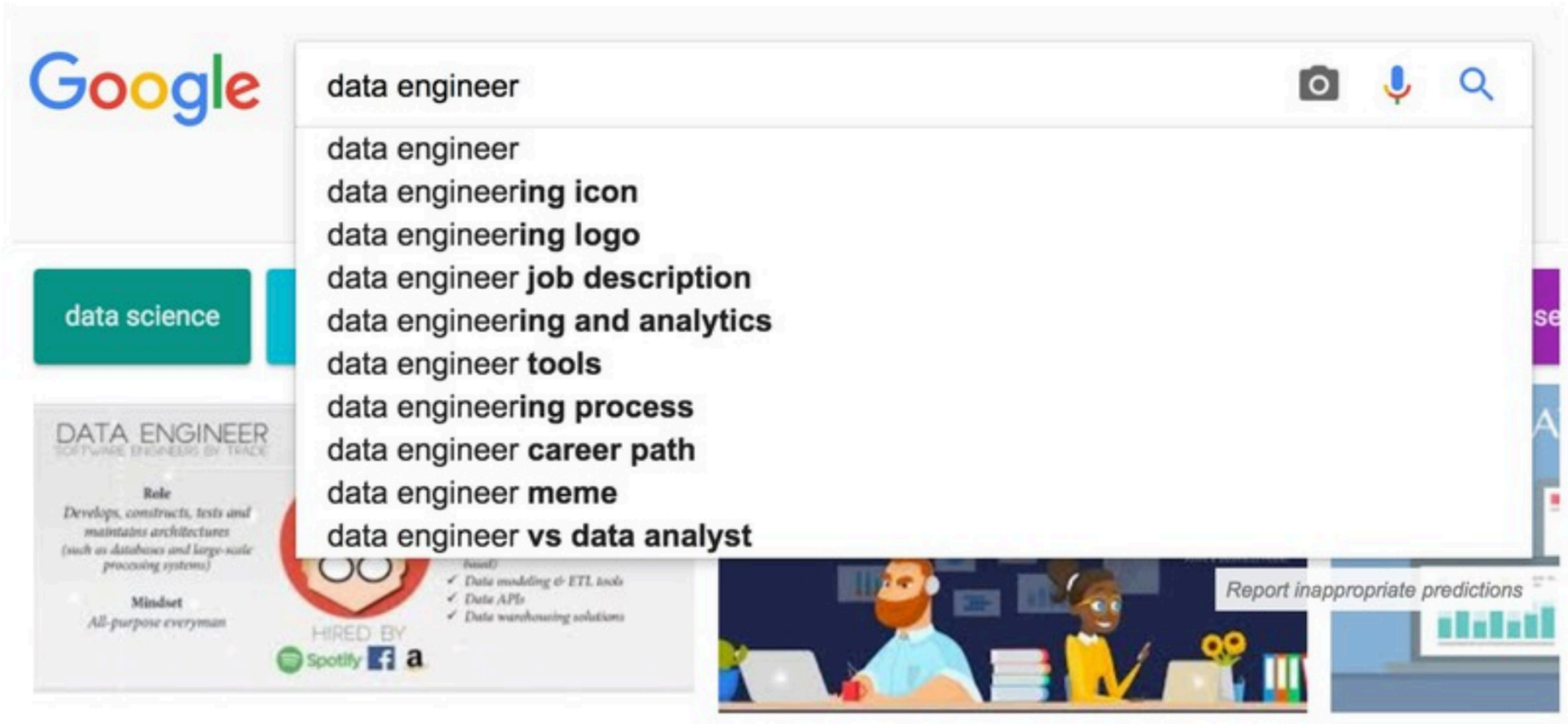
Agenda

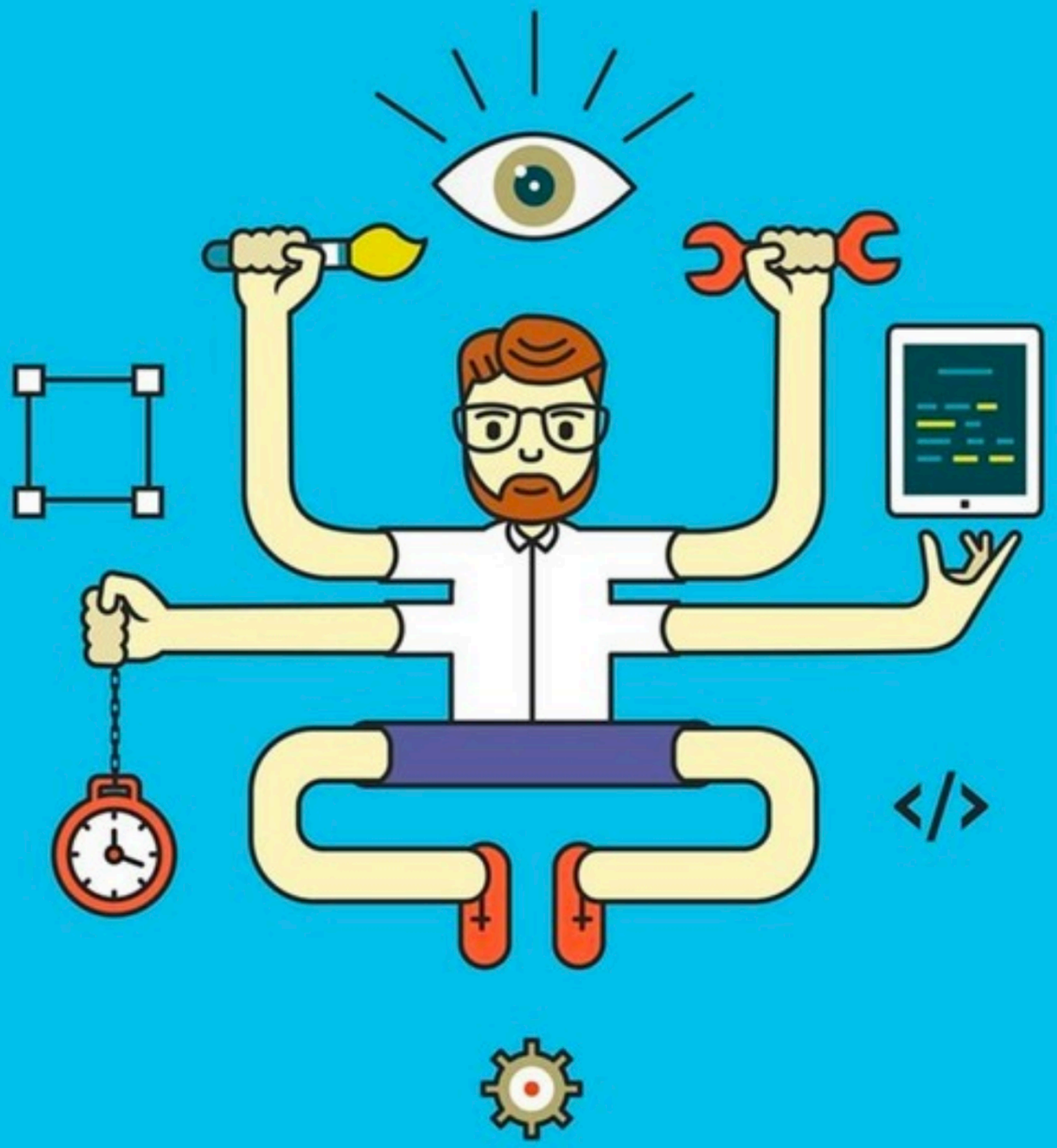
1 Extend

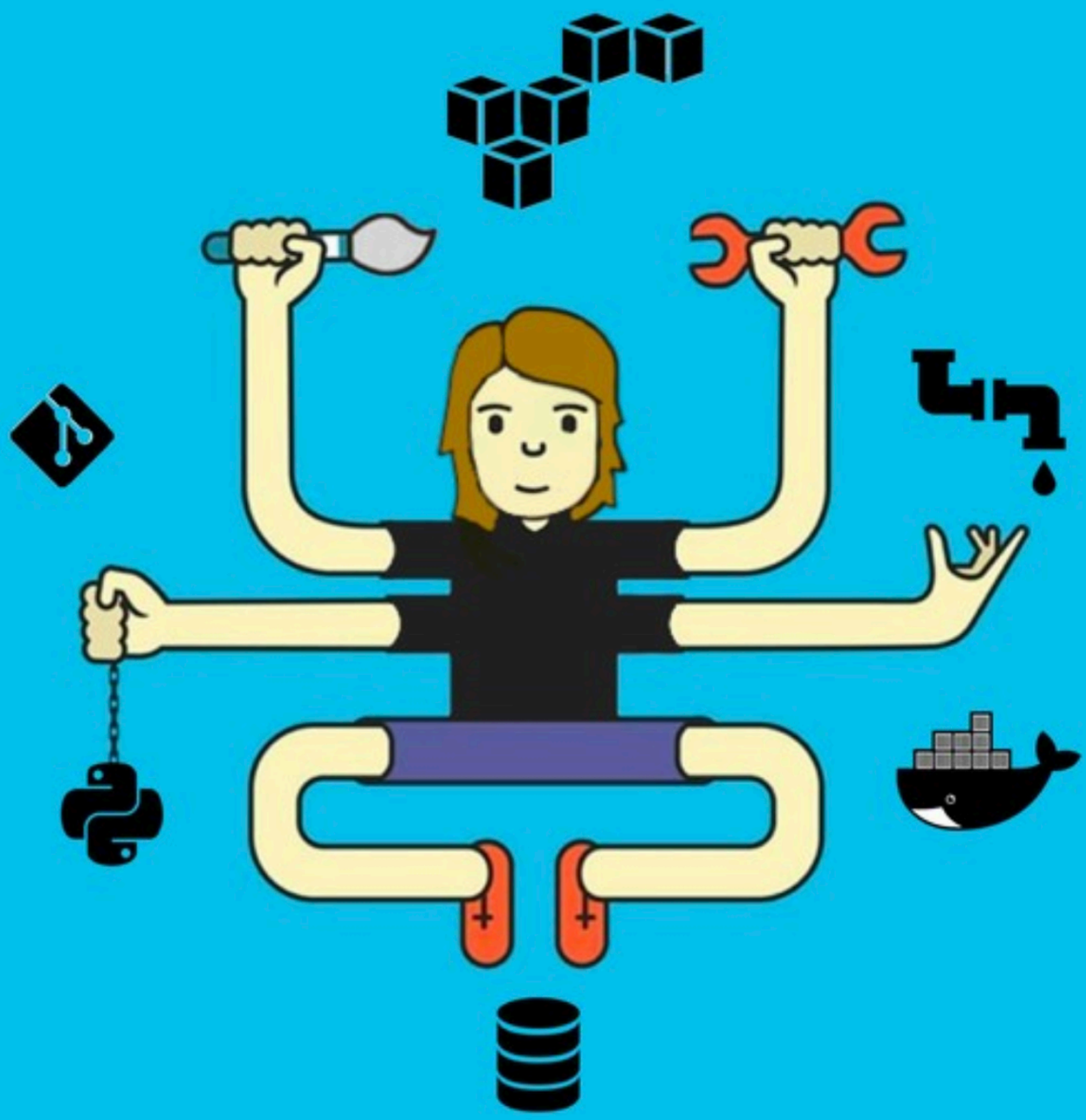
2 Save

3 Schedule

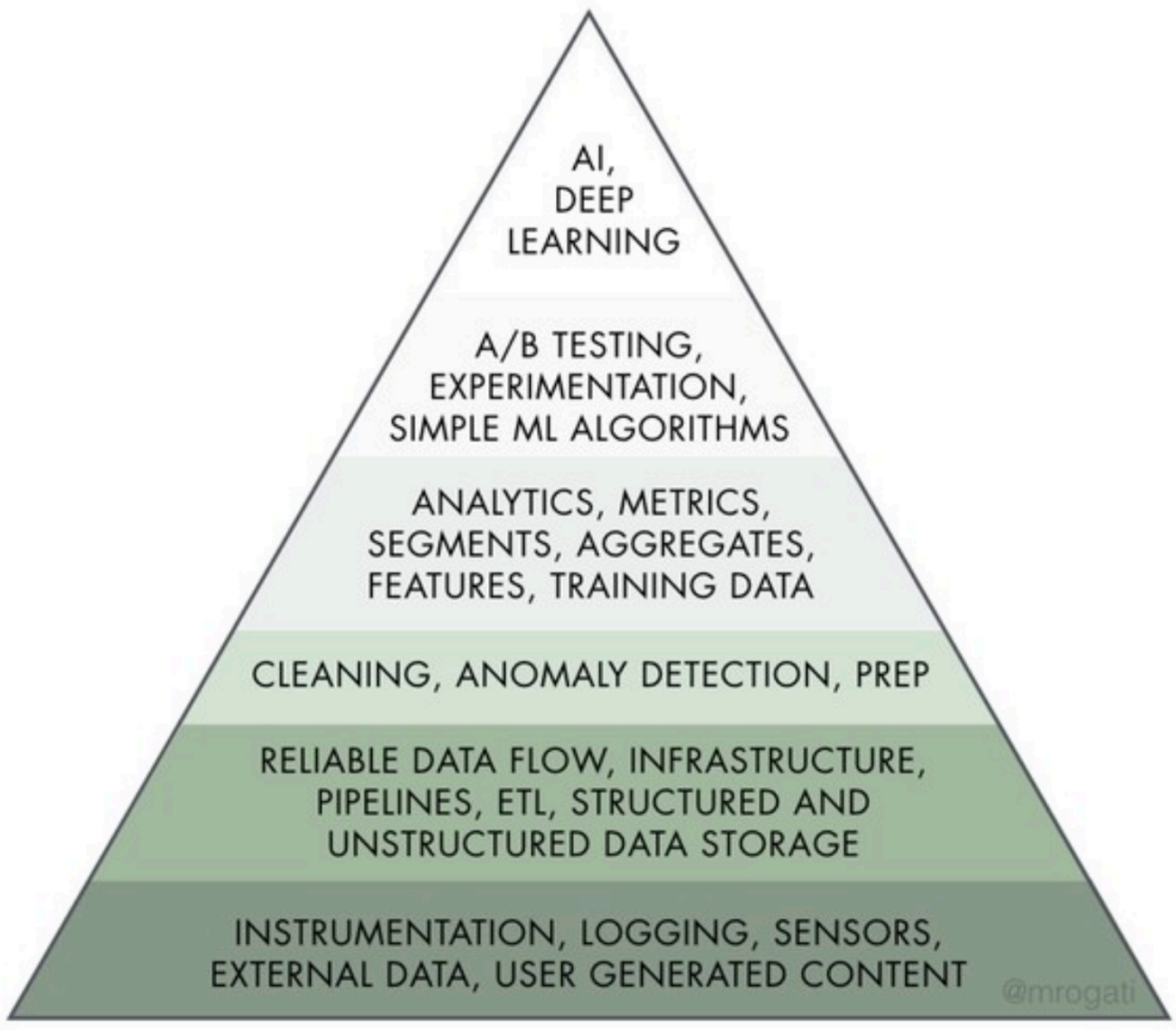
Data Engineering



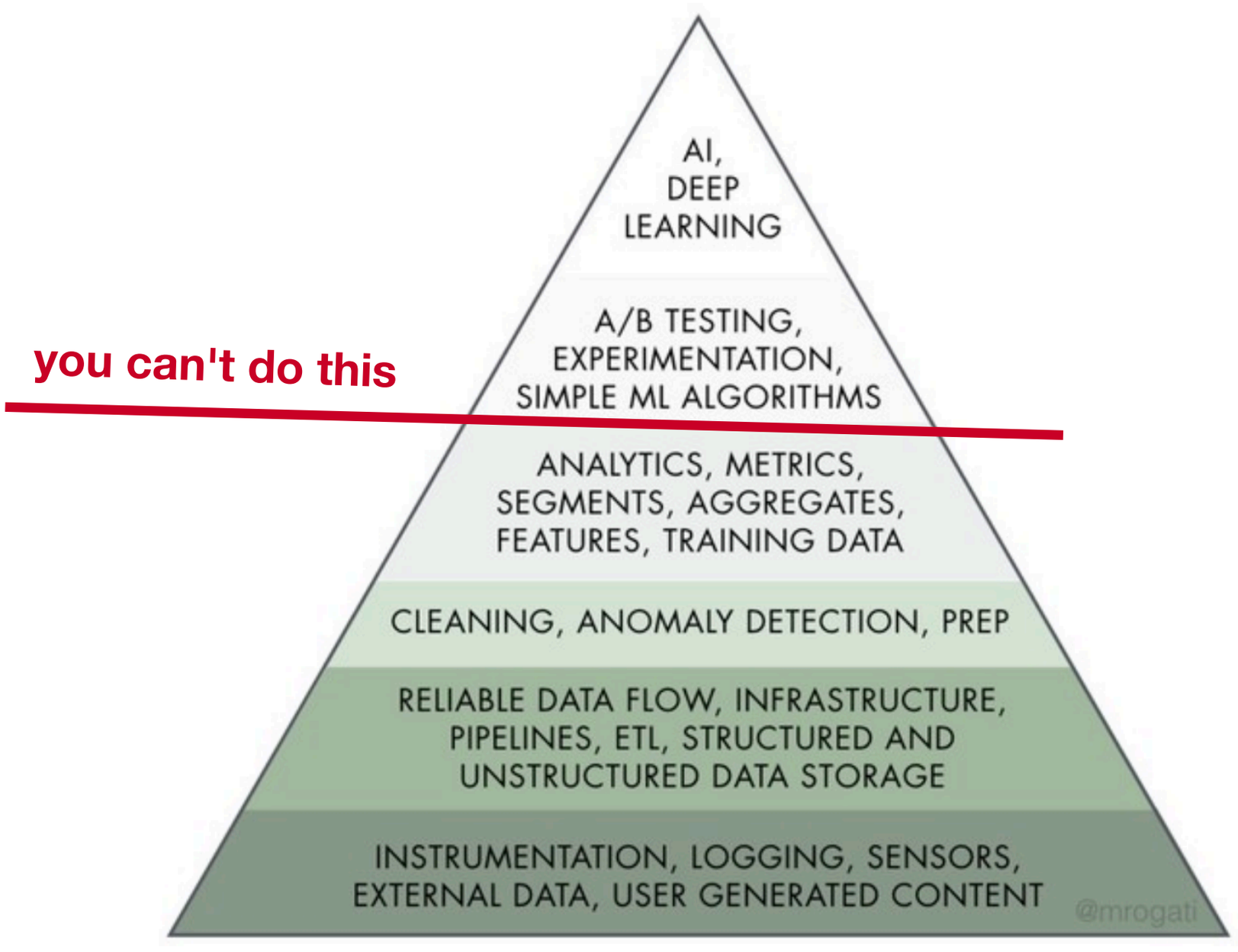




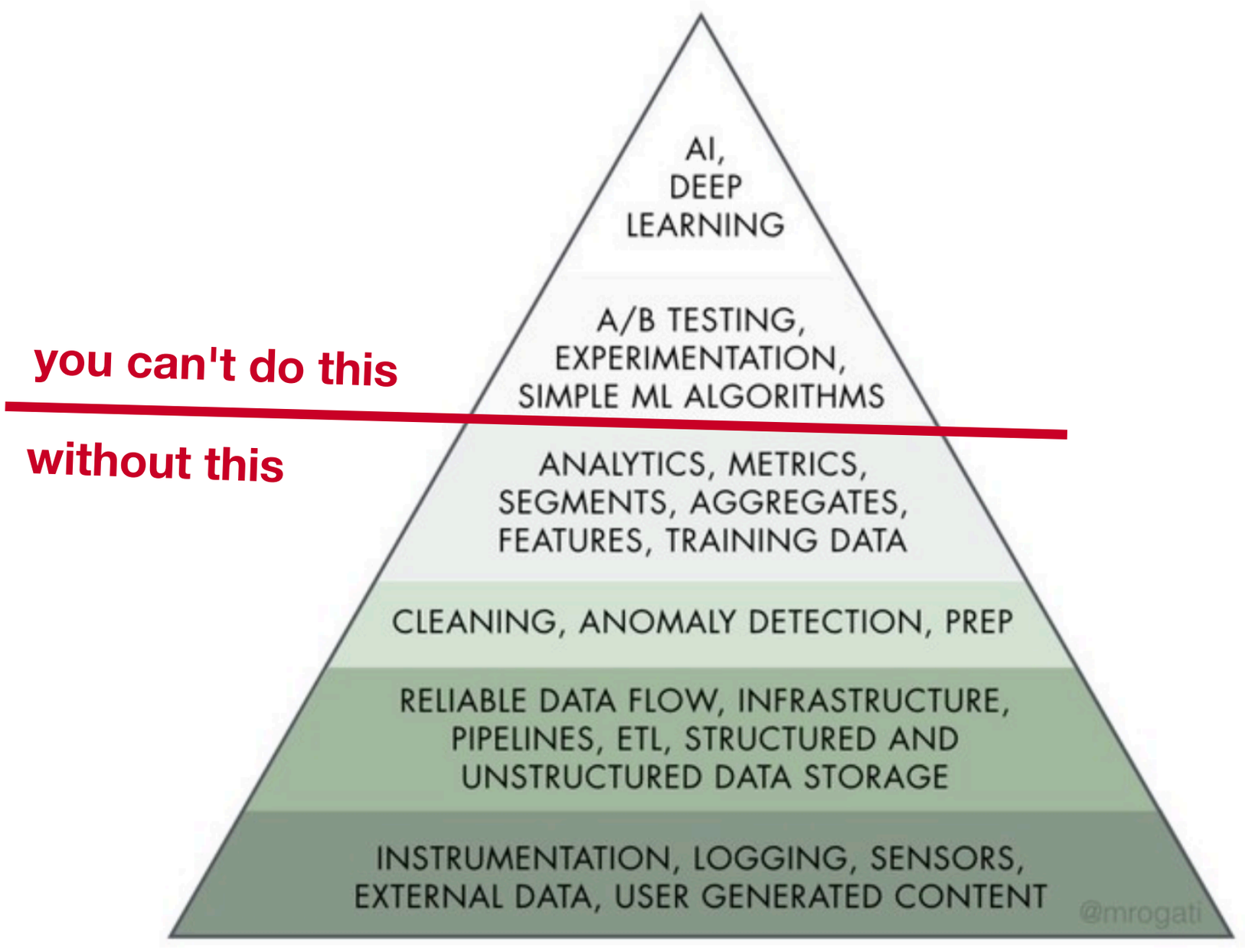
Data Hierarchy of Needs



Data Hierarchy of Needs



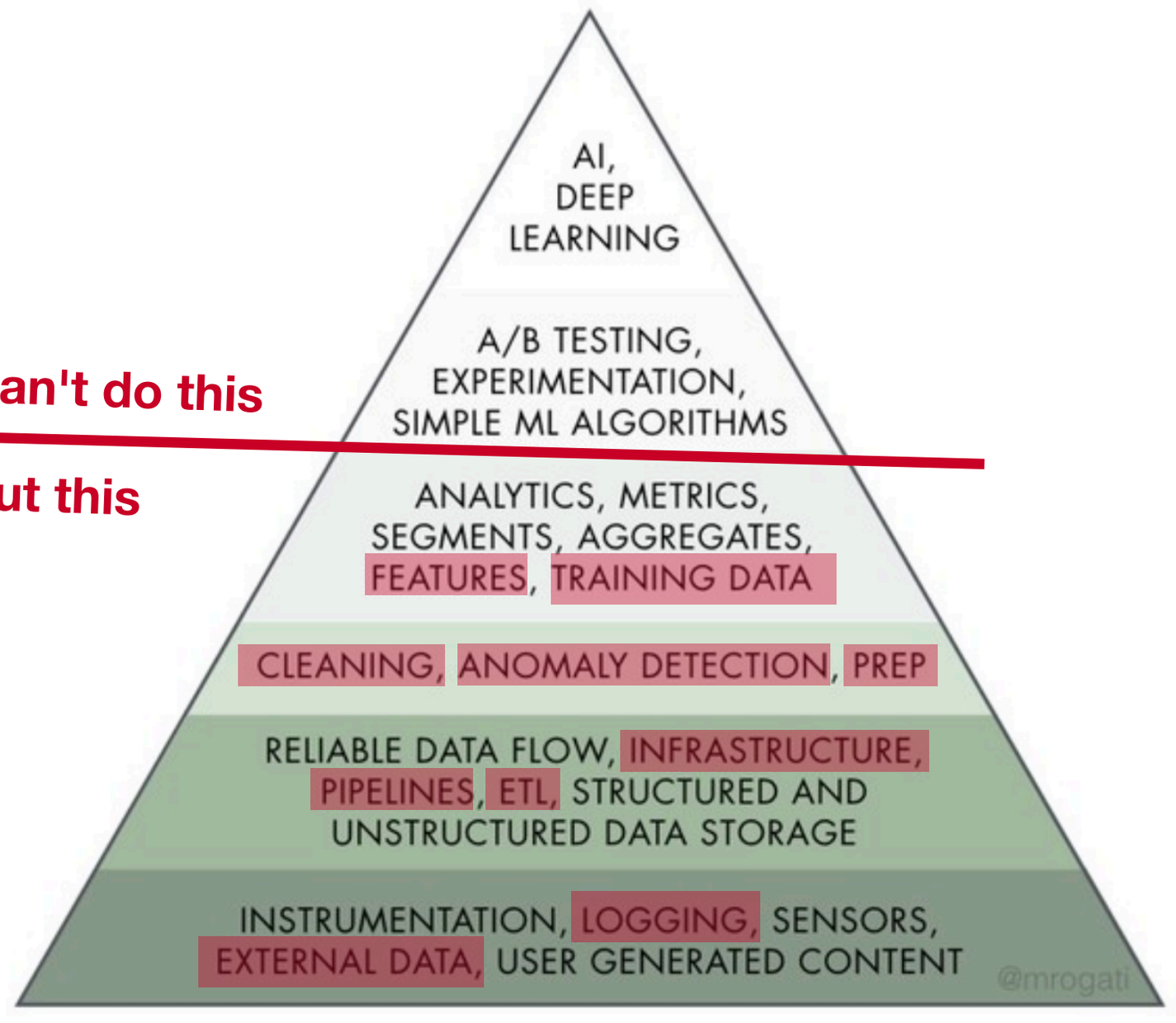
Data Hierarchy of Needs



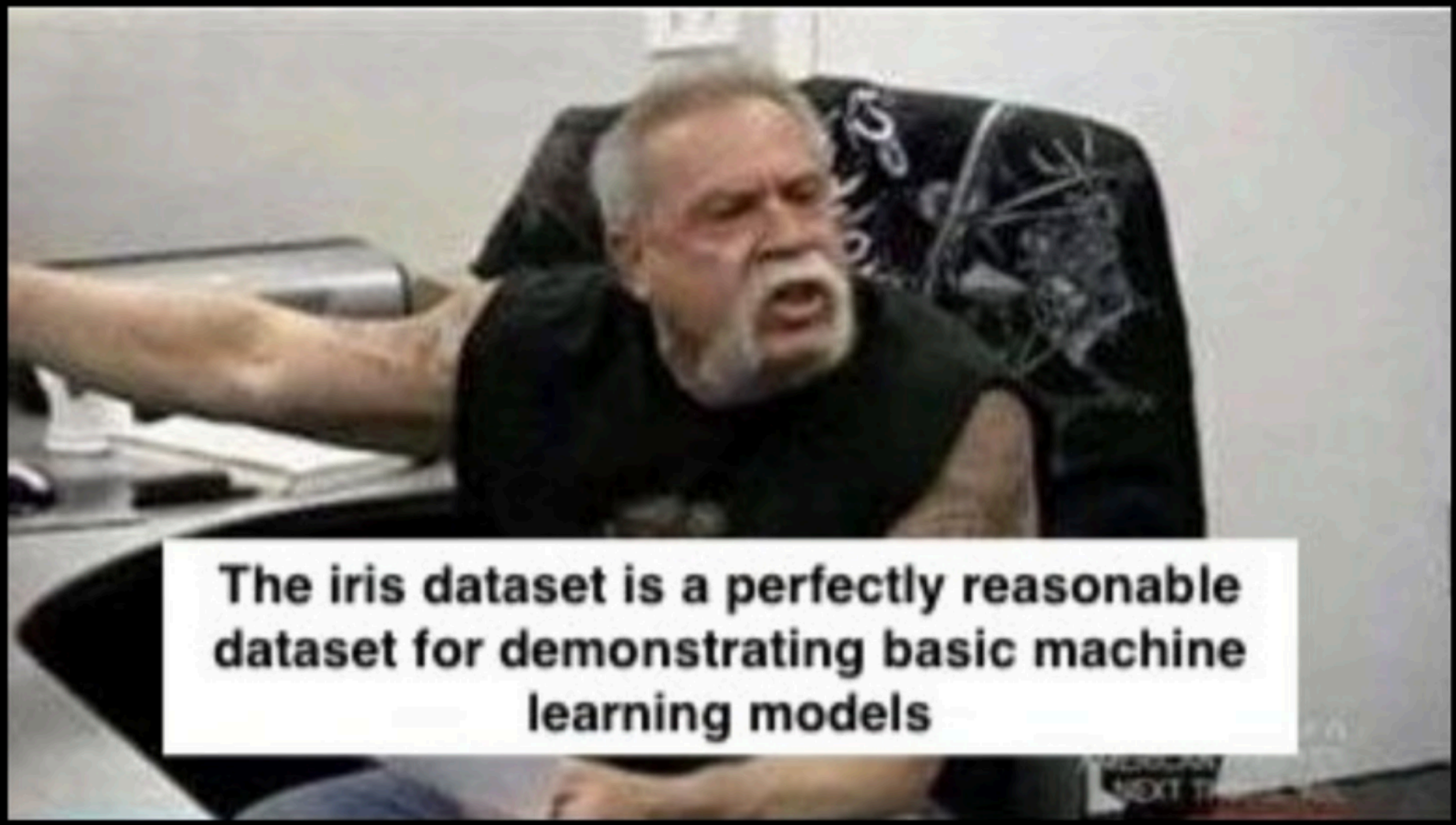
Data Hierarchy of Needs

you can't do this

without this

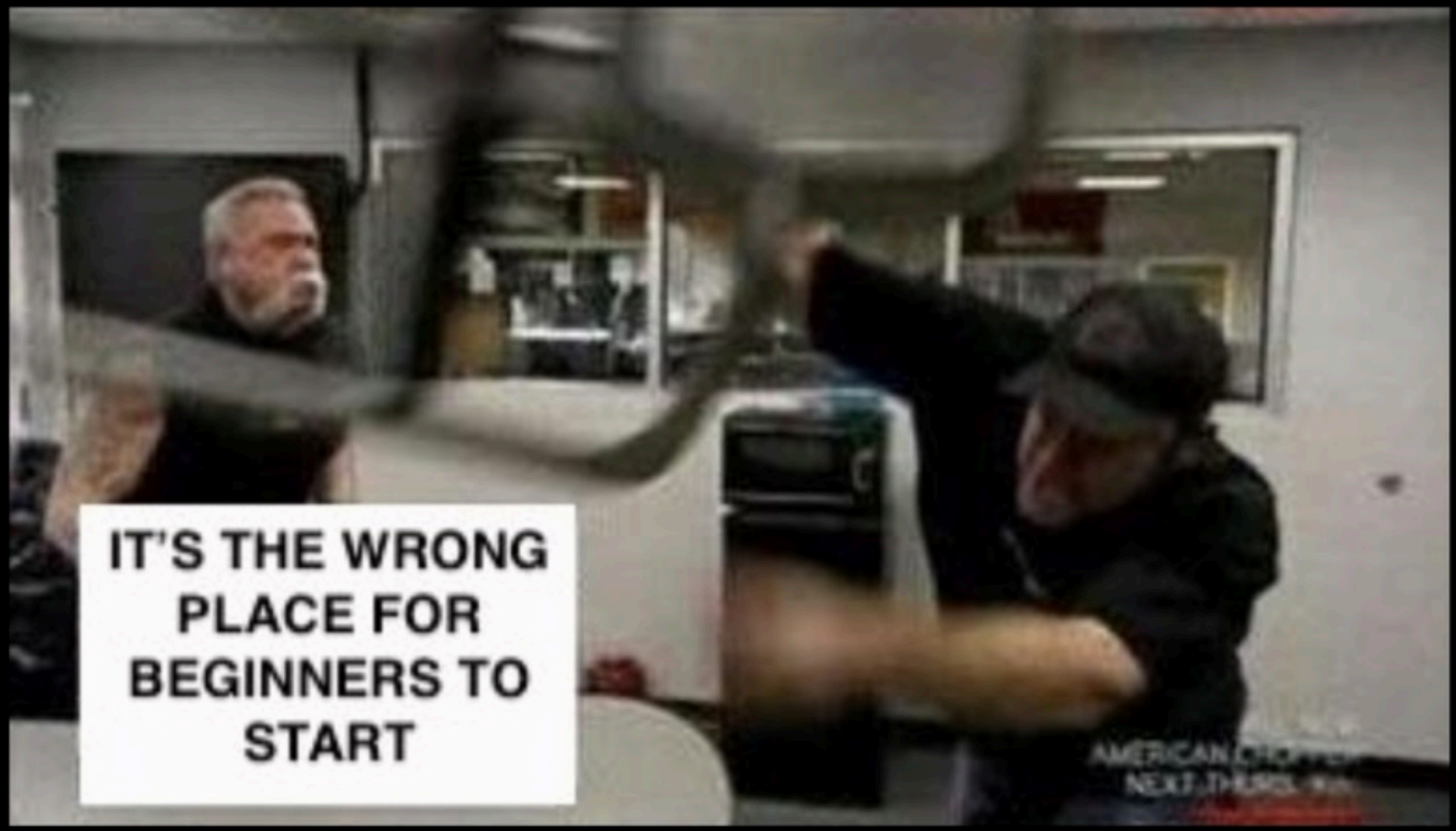



ELMER'S
Washable, No Run
School Glue
#1 Teacher Brand
Safe, Non-Toxic
AP
7.625 FL OZ (225 mL)


The iris dataset is a perfectly reasonable dataset for demonstrating basic machine learning models


It lulls beginners into falsely thinking working with data is easy


EVERYONE NEEDS TO START
SOMEWHERE


IT'S THE WRONG PLACE FOR BEGINNERS TO START


YOU DON'T
REMEMBER
WHAT IT'S LIKE
TO BE A
BEGINNER


ZAITSEV
22
KAPANEN
28
31
BROWN
12
52
65
14


FANTASY
HOCKEY

Max
102 - 48 - 15 | 1st

8.00 vs. 2.00

ibringdog debrincat
Kyle
76 - 69 - 20 | 4th

Team	G	A	+/-	PPP	SOG	HIT	BLK	W	GA	SV	SHO	Score
	18	**24**	**19**	**8**	**91**	**64**	**41**	1	11	**103**	0	**8**
ibringdog debrincat	4	14	4	6	52	22	20	**2**	**8**	89	0	2

Today's Stats Matchup Totals

Standings

Current Standings

Standings | Schedule | Playoffs | Ratings & Levels New

Rank	Team	W-L-T	Pct	Pts	Last Week	Waiver	Moves
1	[redacted]	102-48-15	.664	219	9-1-1	1	38
2	Sharva's a Rat	83-63-19	.561	185	1-9-1	6	10
3	Adam's Cool Team	79-65-21	.542	179	5-5-1	8	12
4	ibringdog debrincat	76-69-20	.521	172	3-6-2	5	19
5	Isaac's Team	74-72-19	.506	167	6-3-2	10	15
6	👉👌	71-70-24	.503	166	7-2-2	9	15
7	My team sucks	72-75-18	.491	162	5-5-1	3	40
8	Dex's Dumb and Dumba	66-87-12	.436	144	6-5-0	4	11
9	Alex's Amazing Team	55-88-22	.400	132	5-6-0	2	12
10	Taylor's Team	54-95-16	.376	124	2-7-2	7	4

Last standings update: Sat Feb 01 02:01am EST

1

Extend

model.ipynb

ACTIVITY

activities/mapper.ipynb

Q&A

Jupyter to Hydrogen

Jupyter to Hydrogen

.ipynb

- ✅ exploratory analysis
- ✅ visualizing ideas
- ✅ prototyping

- ❌ messy
- ❌ bad at versioning
- ❌ not ideal for production

Jupyter to Hydrogen

.ipynb

.py

```
$ jupyter nbconvert --to script model.ipynb
```

Part I: Extend

Need data to talk about data, and a model to talk about models...

The Data

In [1]:
```
import pandas as pd

df = pd.read_csv('data/hockey.csv')
```

In [2]:
```
df.shape
```

Out[2]:
```
(28917, 12)
```

In [3]:
```
df[df['name'] == 'Auston Matthews'].head(3)
```

Out[3]:

	player_id	name	position	date	team	venue	opponent	outcome	goals	assists	sho
10	matthau01	Auston Matthews	C	2019-10-02	TOR	Home	OTT	W	2	0	

```
$ jupyter nbconvert --to script model.ipynb
```

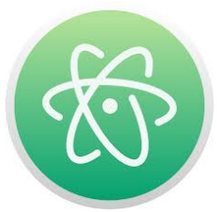
model.py

```
#!/usr/bin/env python
# coding: utf-8

# # Part I: Extend

# Need data to talk about data, and a model to talk about models...

# ### The Data

# In[1]:


import pandas as pd

df = pd.read_csv('data/hockey.csv')


# In[2]:
```

Hydrogen

Hydrogen is an interactive coding environment that supports Python, R, JavaScript and other Jupyter kernels.

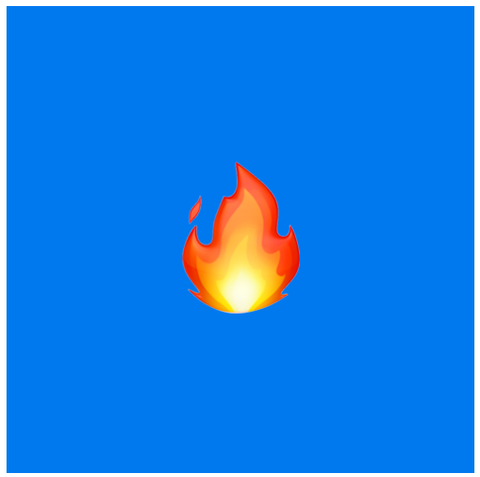
Checkout our Documentation and Medium blog post to see what you can do with Hydrogen.



https://atom.io/packages/hydrogen

model_slim.py

Q&A




input
output

Python Fire 

Python Fire is a library for automatically generating command line interfaces (CLIs) from absolutely any Python object.

- Python Fire is a simple way to create a CLI in Python. [1]
- Python Fire is a helpful tool for developing and debugging Python code. [2]
- Python Fire helps with exploring existing code or turning other people's code into a CLI. [3]
- Python Fire makes transitioning between Bash and Python easier. [4]
- Python Fire makes using a Python REPL easier by setting up the REPL with the modules and variables you'll need already imported and created. [5]

Installation

To install Python Fire with pip, run: `pip install fire`

To install Python Fire with conda, run: `conda install fire -c conda-forge`

To install Python Fire from source, first clone the repository and then run: `python setup.py install`

predict_scaffold.py

predict.py

```
🔥 python predict.py ovechal01
```

ACTIVITY

💻 activities/fire_starter.py

Q&A


2
Save

predict_to_database.py

```
🔥 python predict_to_database.py ovechal01
```

Monitor


Rollbar


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

Waiting for data from your Python app...

This page will update once we receive an error from your app.

Don't want to start with Python? Choose a different SDK.

Installation

To send errors to Rollbar from your Python application you should use the pyrollbar SDK. Install pyrollbar with pip:

```
pip install rollbar
```

COPY

Send a Message

You'll need your project's server-side access token to initialize the pyrollbar SDK. Sending a message to the Rollbar server is as simple as:

```
import rollbar

rollbar.init('49e86c1178e346899c2f29e3744420a9')
rollbar.report_message('Rollbar is configured correctly')
```

COPY

A few seconds after you execute this code, the message should appear on your project's "Items" page.

remove password

Search

- Repositories 244
- Code 50M+
- **Commits 508K**
- Issues 280K
- Packages 1
- Marketplace 0
- Topics 0
- Wikis 34K
- Users 0

Advanced search Cheat sheet

Showing 445,215 available commit results

Sort: **Recently committed**

removed password ···

Verified c64833c

reharr4 committed to reharr4/blamazon 2 hours ago

Onmibus updates (#7) ···

Verified b1b80c9

stevector committed to stevector/internet-button-playground 4 hours ago ✔

Merge pull request #534 in EDD/edd-django from ~WCMORRELL/edd:feature... ···

8aa8978

wmorrell committed to JBEI/edd 7 hours ago

File restructuring ···

8735106

caitlin authored and **PilotBob** committed to CassandraWerewolf/CassandraWeb 7 hours ago

removed password

removed password from bamazonCustomer.js

master

reharr4 committed 2 hours ago Verified 1 parent 7e38084 commit c64833c915cc0f604cb0eeab

Showing 1 changed file with **2 additions** and **2 deletions**.

4 bamazonCustomer.js

```
@@ -9,7 +9,7 @@ var connection = mysql.createConnection({
9  9          // username
10 10         user: "root",
11 11         // password
12    -       password: "L!sb0n13",
   12 +       password: "",
13 13         database: "bamazon_db"
14 14     });
15 15
@@ -94,4 +94,4 @@ connection.connect(function (err) {
94 94     //         })
95 95     //     })
96 96     // }
97    - connection.end();
   97 + connection.end();
```


PSSST...!


Secrets

python-dotenv |

Reads the key,value pair from `.env` file and adds them to environment variable. It is great for managing app settings during development and in production using 12-factor principles.

> Do one thing, do it well!

Installation

```
pip install -U python-dotenv
```

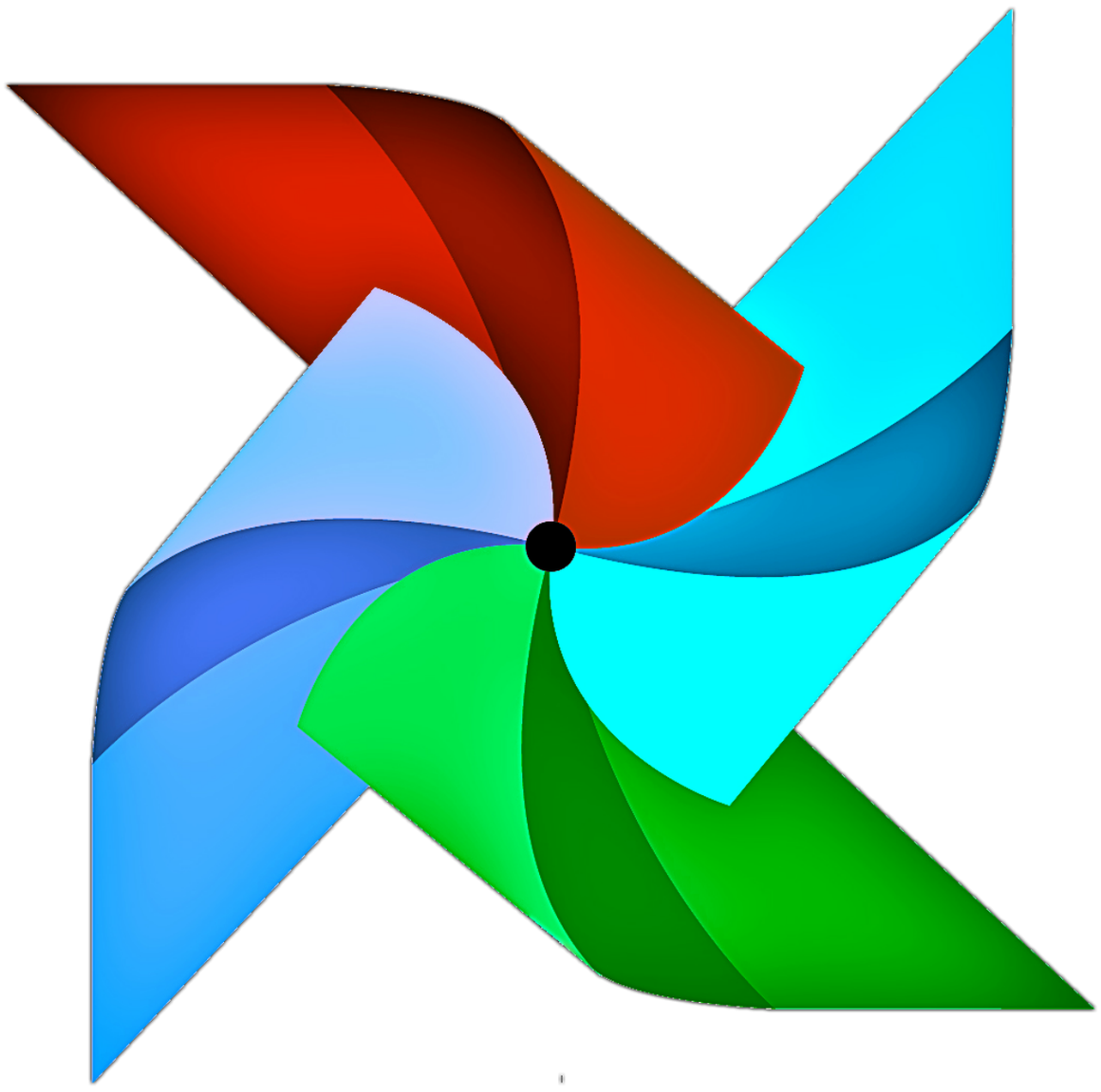
model_rollbar.py

ACTIVITY

💻 activities/secret_sandbox.py

3

Schedule



Airflow

Apache Airflow

Apache Airflow (or simply Airflow) is a platform to programmatically author, schedule, and monitor workflows.

When workflows are defined as code, they become more maintainable, versionable, testable, and collaborative.

Use Airflow to author workflows as directed acyclic graphs (DAGs) of tasks. The Airflow scheduler executes your tasks on an array of workers while following the specified dependencies. Rich command line utilities make performing complex surgeries on DAGs a snap. The rich user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed.

Getting started

Please visit the Airflow Platform documentation (latest **stable** release) for help with installing Airflow, getting a quick start, or a more complete tutorial.

Documentation of GitHub master (latest development branch): ReadTheDocs Documentation

For further information, please visit the Airflow Wiki.

Tutorial

This tutorial walks you through some of the fundamental Airflow concepts, objects, and their usage while writing your first pipeline.

Example Pipeline definition

Here is an example of a basic pipeline definition. Do not worry if this looks complicated, a line by line explanation follows below.

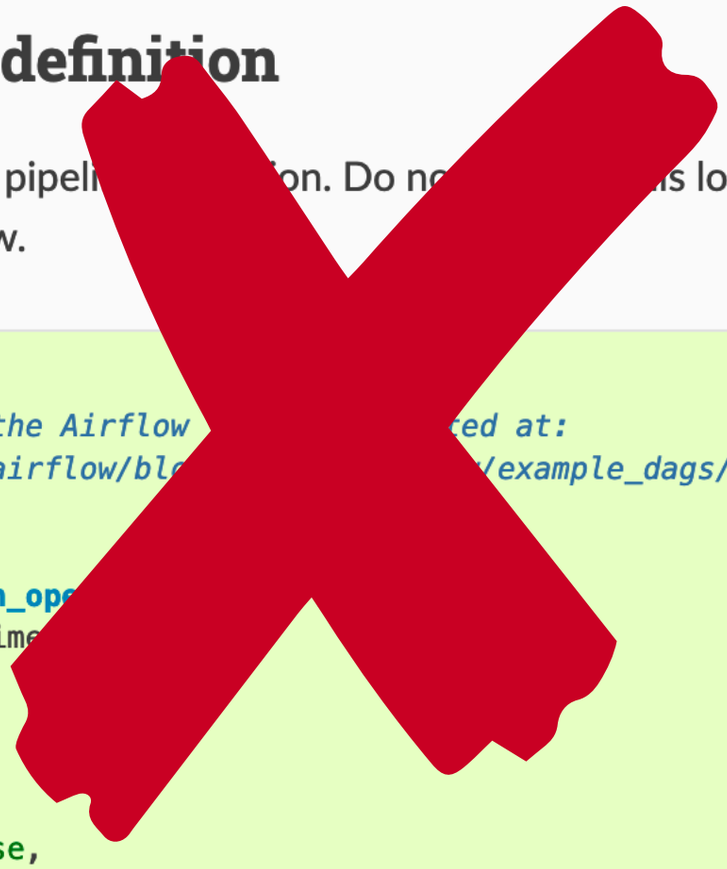
```
"""
Code that goes along with the Airflow tutorial located at:
https://github.com/apache/airflow/blob/master/airflow/example_dags/tutorial.py
"""
from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from datetime import datetime, timedelta


default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime(2015, 6, 1),
    'email': ['airflow@example.com'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=5),
    # 'queue': 'bash_queue',
    # 'pool': 'backfill',
    # 'priority_weight': 10,
    # 'end_date': datetime(2016, 1, 1),
}

dag = DAG('tutorial', default_args=default_args, schedule_interval=timedelta(days=1))
```

Tutorial

This tutorial walks you through some of the fundamental Airflow concepts, objects, and their usage while writing your first pipeline.

Example Pipeline definition

Here is an example of a basic pipeli on. Do n s looks complicated, a line by line explanation follows below.



```
"""
Code that goes along with the Airflow                ted at:
https://github.com/apache/airflow/bl                /example_dags/tutorial.py
"""
from airflow import DAG
from airflow.operators.bash_ope
from datetime import datetime


default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime(2015, 6, 1),
    'email': ['airflow@example.com'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=5),
    # 'queue': 'bash_queue',
    # 'pool': 'backfill',
    # 'priority_weight': 10,
    # 'end_date': datetime(2016, 1, 1),
}

dag = DAG('tutorial', default_args=default_args, schedule_interval=timedelta(days=1))
```

live coding

```
cd DE4DS

# airflow needs a home
export AIRFLOW_HOME=`pwd`/airflow

pip install -U apache-airflow

# in airflow.cfg set:
# load_examples = False

# initialize the database
airflow initdb
```

airflow/dags/hockey.py

💨 `airflow test hockey fetch <date>`

```
airflow test hockey fetch <date>


# start the web server
airflow webserver -p 8080

# start the scheduler
airflow scheduler
```

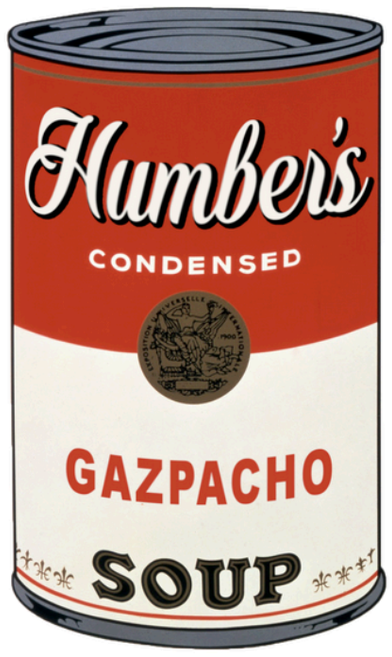
Homework

💻 add t3 to save predictions to the db

Q&A

That's all Folks!

packages



dependencies 0 build passing downloads 6k

build passing pypi v1.0.2 downloads 1k

March 9, 2020

Minimum viable machine learning

Presented by **Max Humber**

158 SPOTS REMAINING

LEARN MORE