DELIVERING ML MODELS THE SAFE AND SANE WAY

David Tan

ThoughtWorks®

THE PLAN TODAY

Why *do we need to improve the ML workflow?*

What *are some better practices?*

How *can we practice these practices?*

TEMPERATURE CHECK

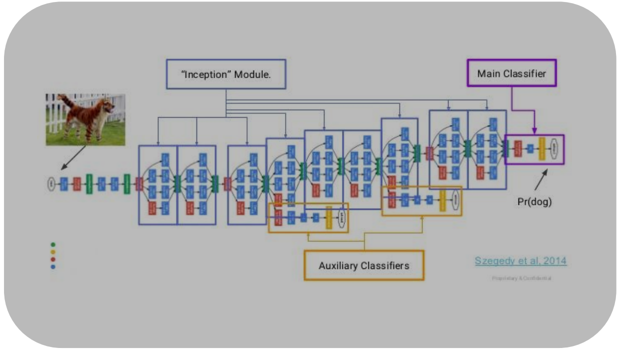
Who has…

- Trained a ML model before?
- Deployed a ML model for fun?
- Deployed a ML model at work?
- Deployed a model using an automated CI pipeline?

WHAT’S THE PROBLEM?

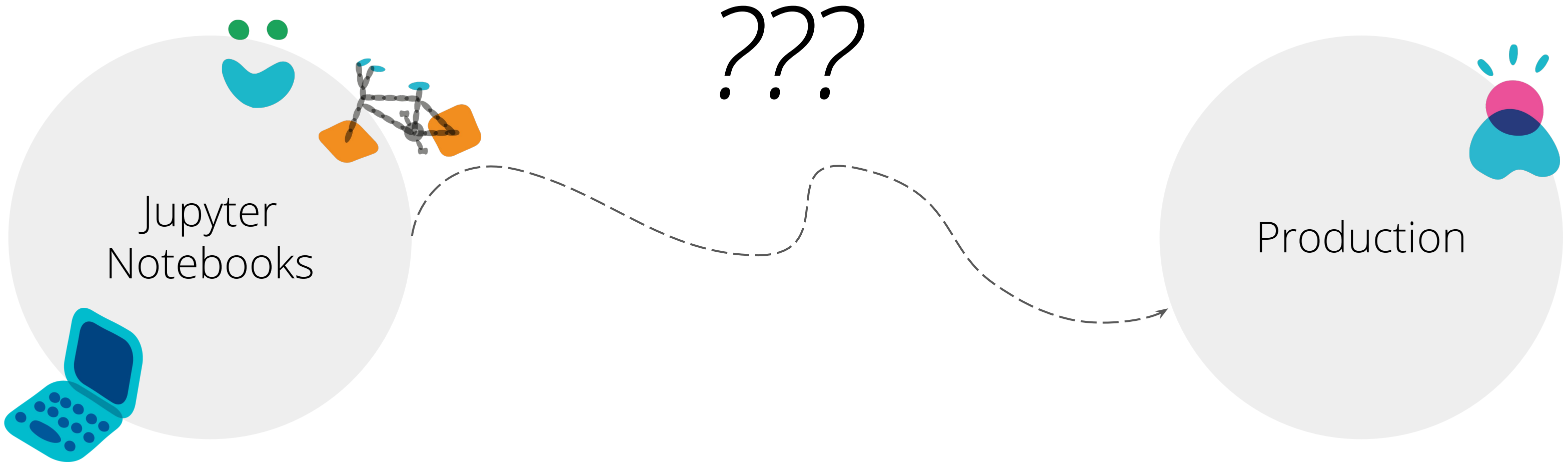
OBSERVATION

One of these is not like the others



Continuous delivery practices can help us change this.

JOURNEY ON AN ML PROJECT



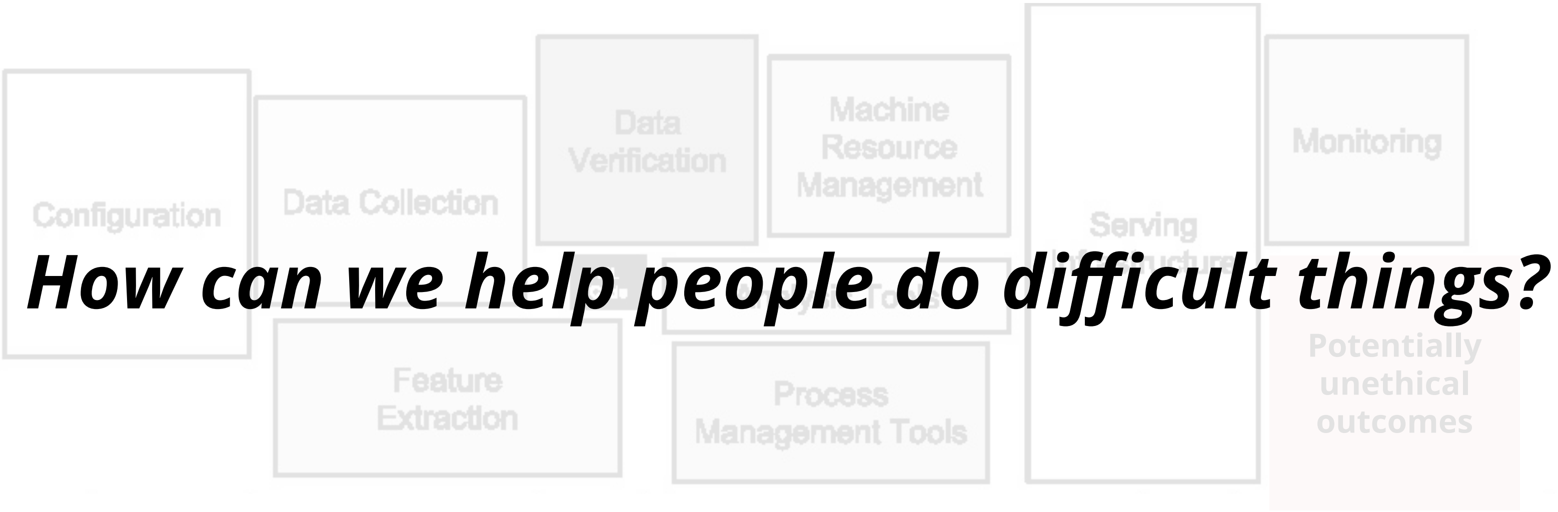
Jupyter Notebooks

???

Production

IT'S NEVER JUST ML

We got 99 problems and machine learning ain't one



How can we help people do difficult things?

Source: Machine Learning: The High Interest Credit Card of Technical Debt (Google, 2015)

HELPING PEOPLE DO DIFFICULT THINGS

Sensible defaults

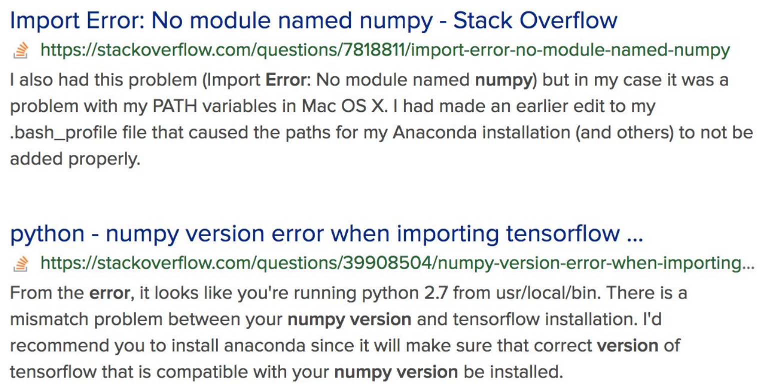
- Two reference repos:
  - `github.com/ThoughtWorksInc/ml-cd-starter-kit`
  - `github.com/ThoughtWorksInc/ml-app-template`
- 6 better habits
  - Common problems and suggested solutions

LET’S GO

6 common problems and suggested solutions

1

WORKS ON MY MACHINE



Import Error: No module named numpy - Stack Overflow

https://stackoverflow.com/questions/7818811/import-error-no-module-named-numpy

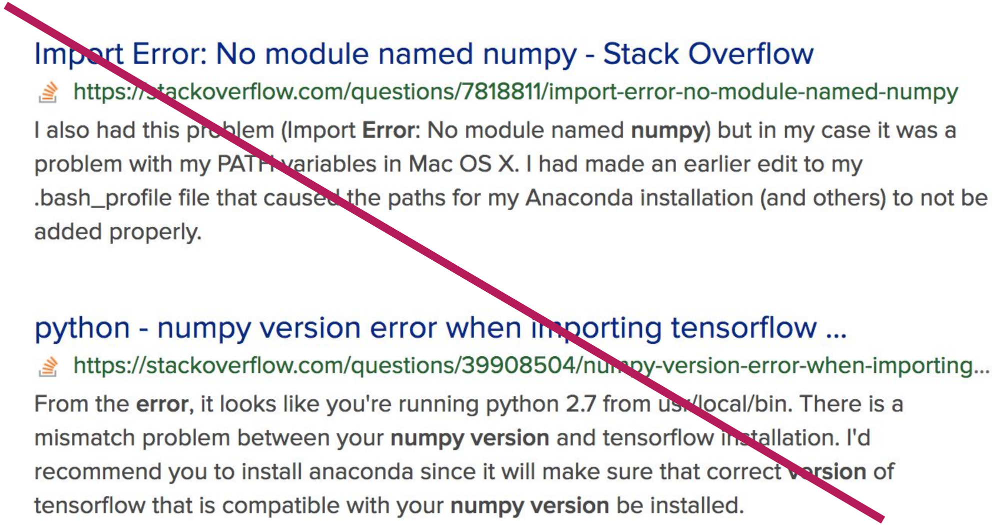
I also had this problem (Import **Error**: No module named **numpy**) but in my case it was a problem with my PATH variables in Mac OS X. I had made an earlier edit to my .bash_profile file that caused the paths for my Anaconda installation (and others) to not be added properly.

python - numpy version error when importing tensorflow ...

https://stackoverflow.com/questions/39908504/numpy-version-error-when-importing...

From the **error**, it looks like you're running python 2.7 from usr/local/bin. There is a mismatch problem between your **numpy version** and tensorflow installation. I'd recommend you to install anaconda since it will make sure that correct **version** of tensorflow that is compatible with your **numpy version** be installed.

1. WORKS ON MY MACHINE



Import Error: No module named numpy - Stack Overflow

https://stackoverflow.com/questions/7818811/import-error-no-module-named-numpy

I also had this problem (Import **Error**: No module named **numpy**) but in my case it was a problem with my PATH variables in Mac OS X. I had made an earlier edit to my .bash_profile file that caused the paths for my Anaconda installation (and others) to not be added properly.

python - numpy version error when importing tensorflow ...

https://stackoverflow.com/questions/39908504/numpy-version-error-when-importing...

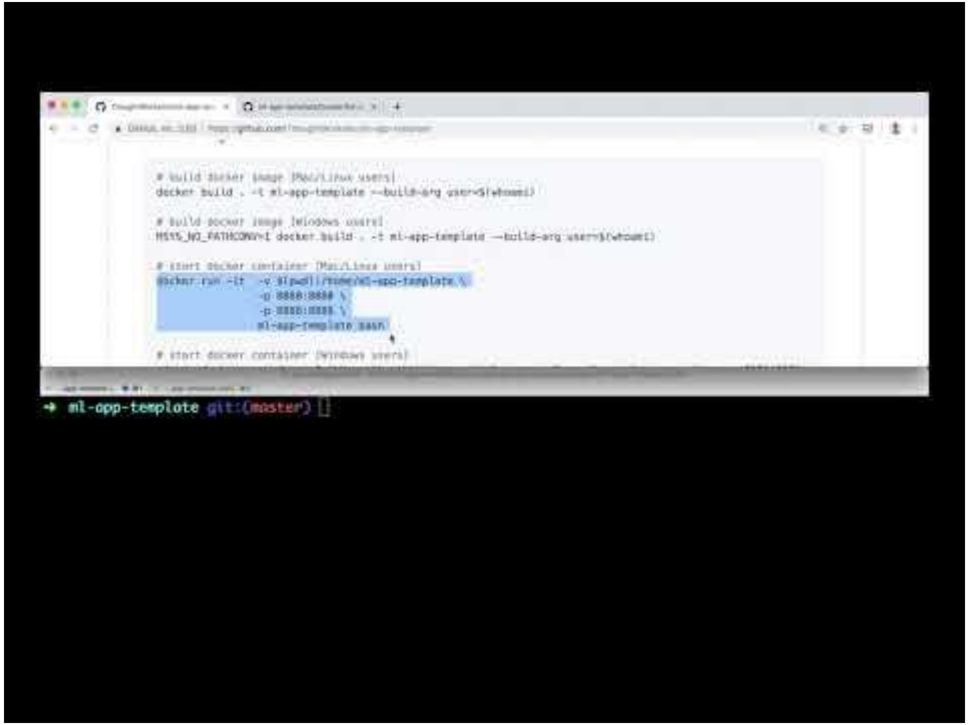
From the **error**, it looks like you're running python 2.7 from usr/local/bin. There is a mismatch problem between your **numpy version** and tensorflow installation. I'd recommend you to install anaconda since it will make sure that correct **version** of tensorflow that is compatible with your **numpy version** be installed.

To start, simply:

- **docker build …**
- **docker run   …**

1. WORKS ON MY MACHINE

Demo



```
# build docker image [Mac/Linux users]
docker build . -t ml-app-template --build-arg user=$(whoami)

# build docker image [Windows users]
MSYS_NO_PATHCONV=1 docker build . -t ml-app-template --build-arg user=$(whoami)

# start docker container [Mac/Linux users]
docker run -it  -v $(pwd):/home/ml-app-template \
                -p 8888:8888 \
                -p 8888:8888 \
                ml-app-template bash

# start docker container [Windows users]
```

```
➜ ml-app-template git:(master) 
```

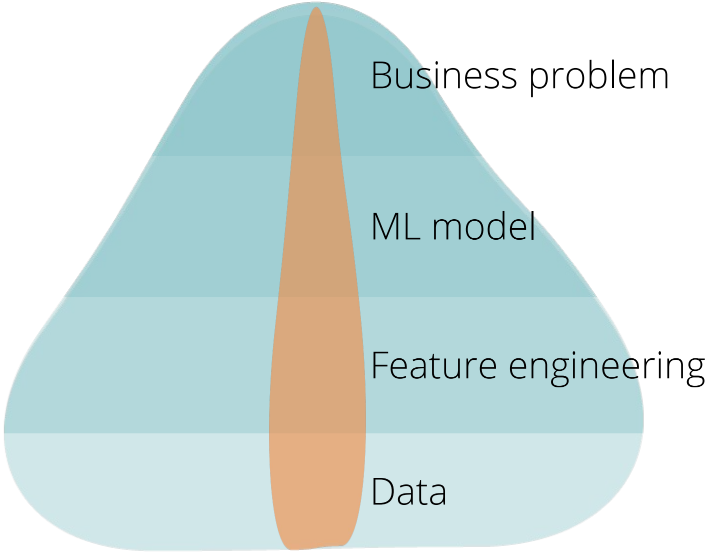
2

NO DATA / DATA SUCKS

2. NO DATA / DATA SUCKS

Mitigation measures

- Think about data access before starting project
- “Wizard of Oz” / Fake it till we make it
    - Provide interface but without ML implementation
- ry release
  om now)



3

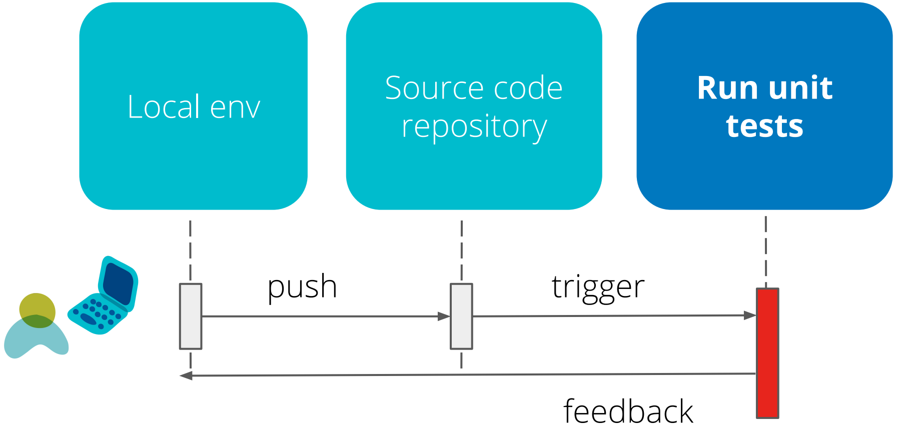
DEPLOYMENTS ARE COMPLICATED

3. DEPLOYMENTS ARE COMPLICATED

Mitigation measures

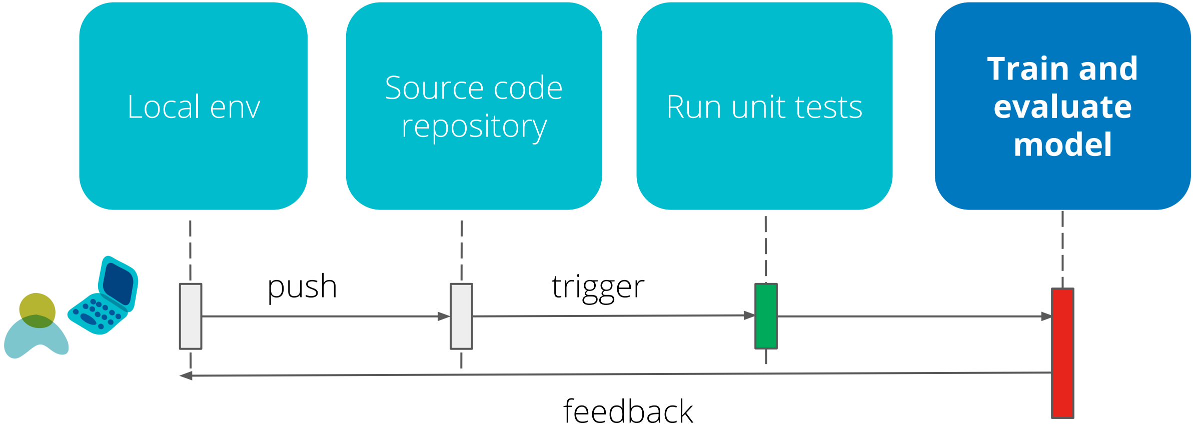
- CI pipeline
- Deploy early and often
- Tracer bullet
- Bring the pain forward

3. DEPLOY EARLY AND OFTEN



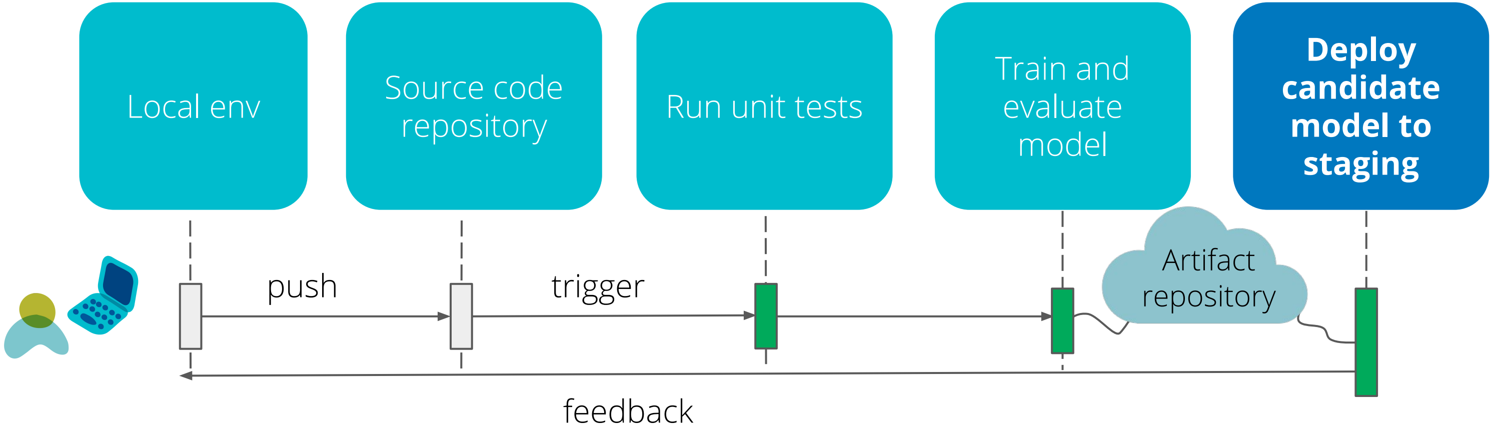
Source: Continuous Delivery (Jez Humble, Dave Farley)

3. DEPLOY EARLY AND OFTEN



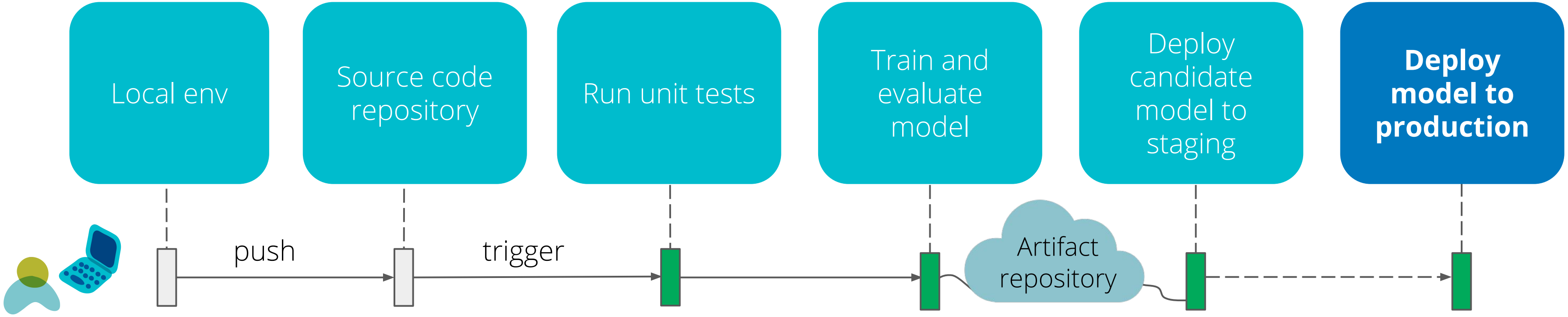
Source: Continuous Delivery (Jez Humble, Dave Farley)

3. DEPLOY EARLY AND OFTEN



Source: Continuous Delivery (Jez Humble, Dave Farley)

3. DEPLOY EARLY AND OFTEN



Local env

Source code repository

Run unit tests

Train and evaluate model

Deploy candidate model to staging

Deploy model to production

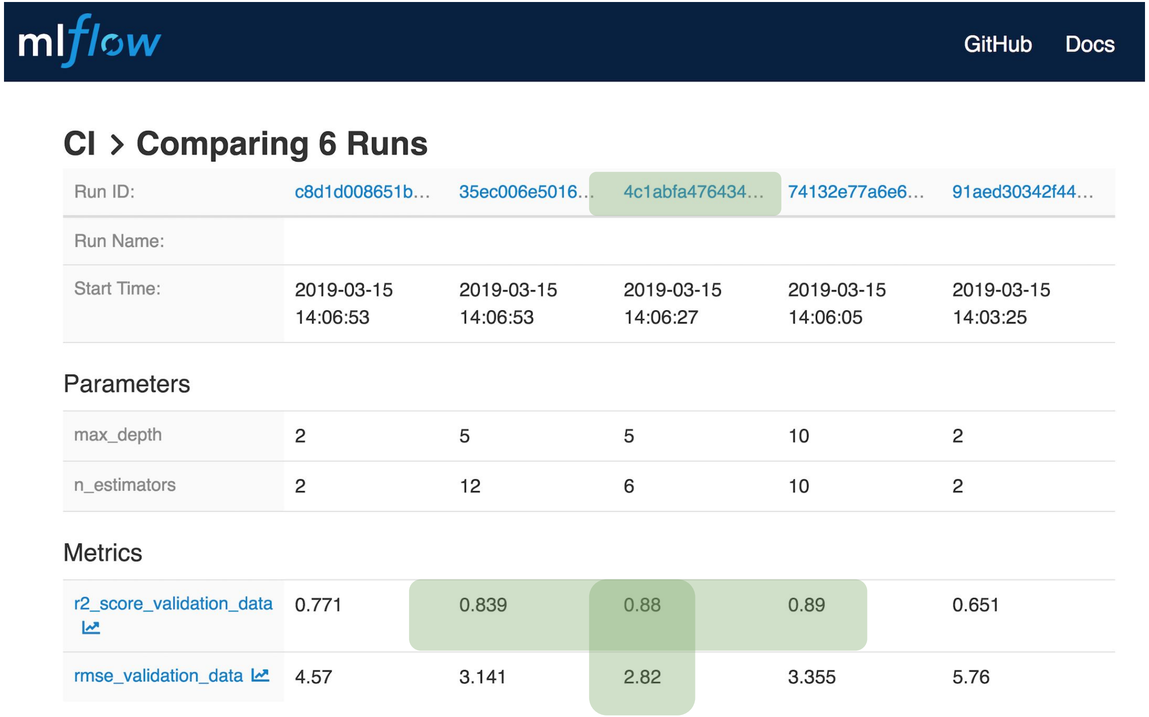
Source: Continuous Delivery (Jez Humble, Dave Farley)

4

WHICH CANDIDATE MODEL TO DEPLOY TO PROD?

4. CHOOSING BUILDS

Include model evaluation metrics in CI pipeline



CI > Comparing 6 Runs

Run ID:	c8d1d008651b...	35ec006e5016...	4c1abfa476434...	74132e77a6e6...	91aed30342f44...
Run Name:					
Start Time:	2019-03-15 14:06:53	2019-03-15 14:06:53	2019-03-15 14:06:27	2019-03-15 14:06:05	2019-03-15 14:03:25

Parameters

max_depth	2	5	5	10	2
n_estimators	2	12	6	10	2

Metrics

r2_score_validation_data	0.771	0.839	0.88	0.89	0.651
rmse_validation_data	4.57	3.141	2.82	3.355	5.76

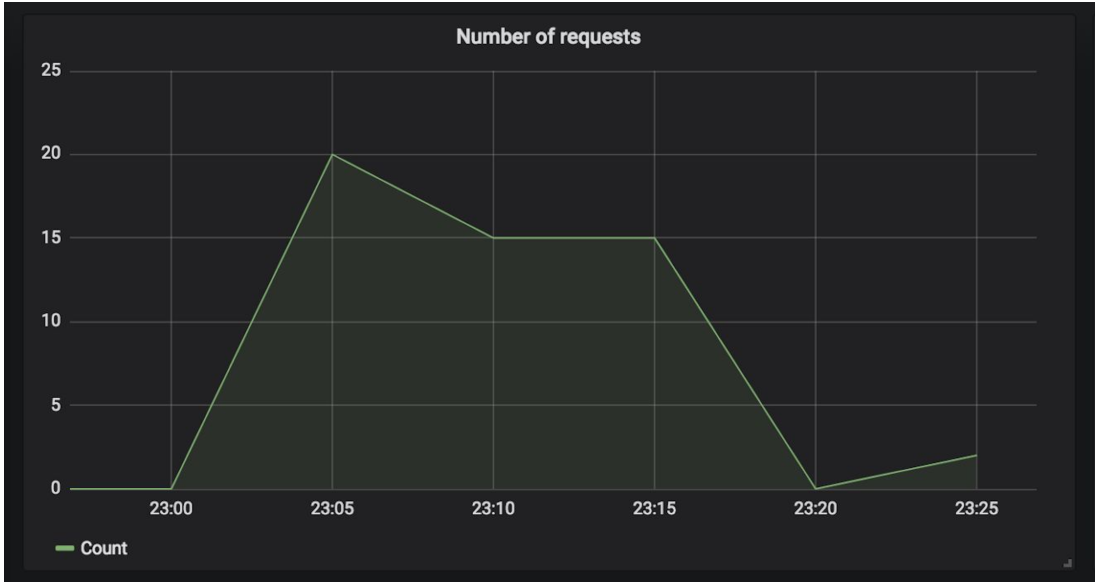
5

HOW'S THE MODEL DOING IN THE WILD?

5. OBSERVE!

Benefit #1: Feedback on production model

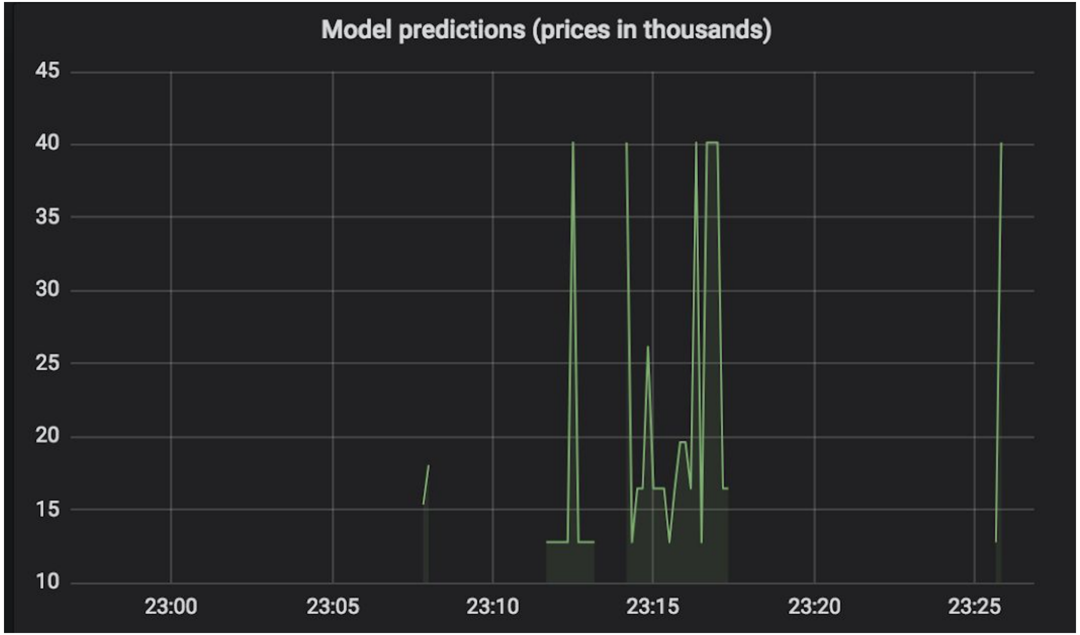
Monitoring **service usage**



5. OBSERVE!

Benefit #1: Feedback on production model

Monitoring **model output**

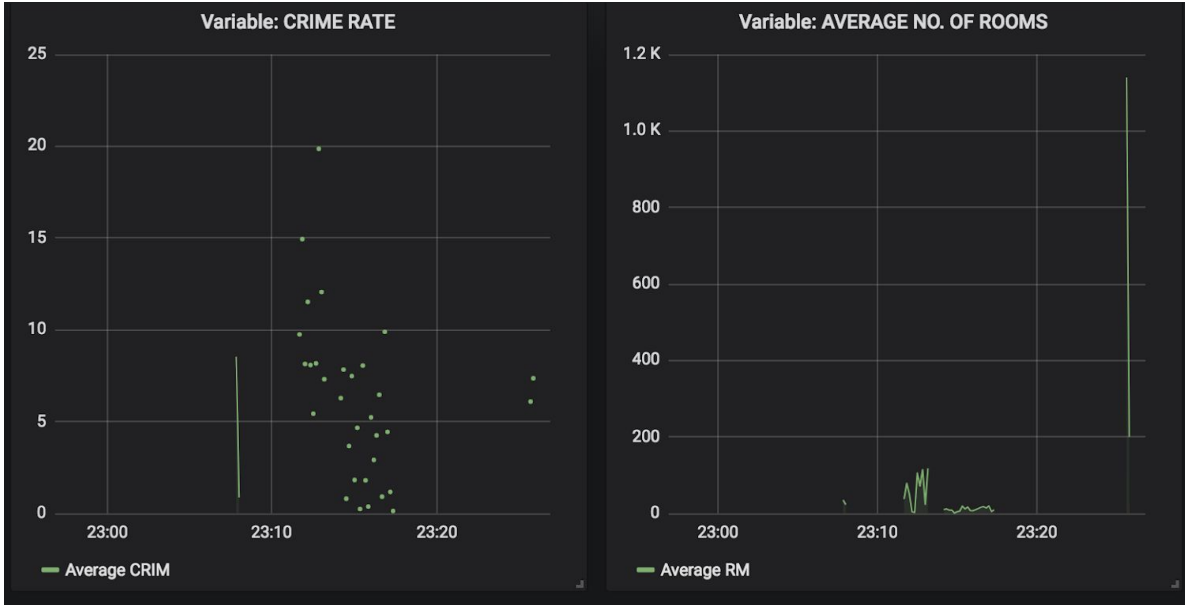


5. OBSERVE!

Benefit #1: Feedback on production model

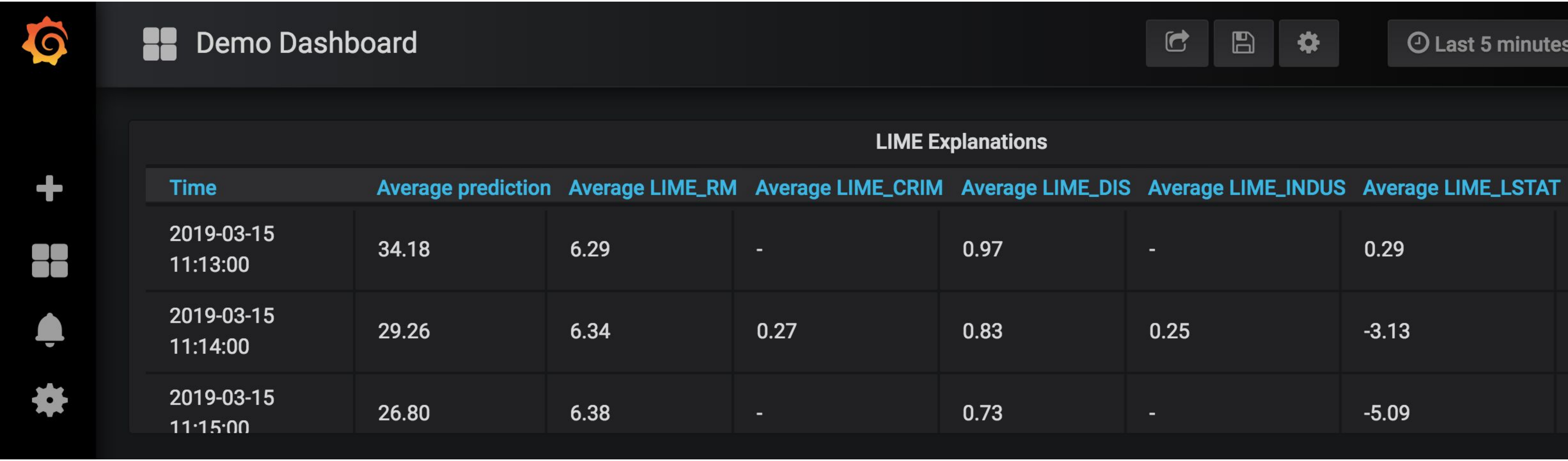
Monitoring **model inputs**

- Could help identify training-serving skew



5. OBSERVE!

Benefit #2: Interpretability of predictions

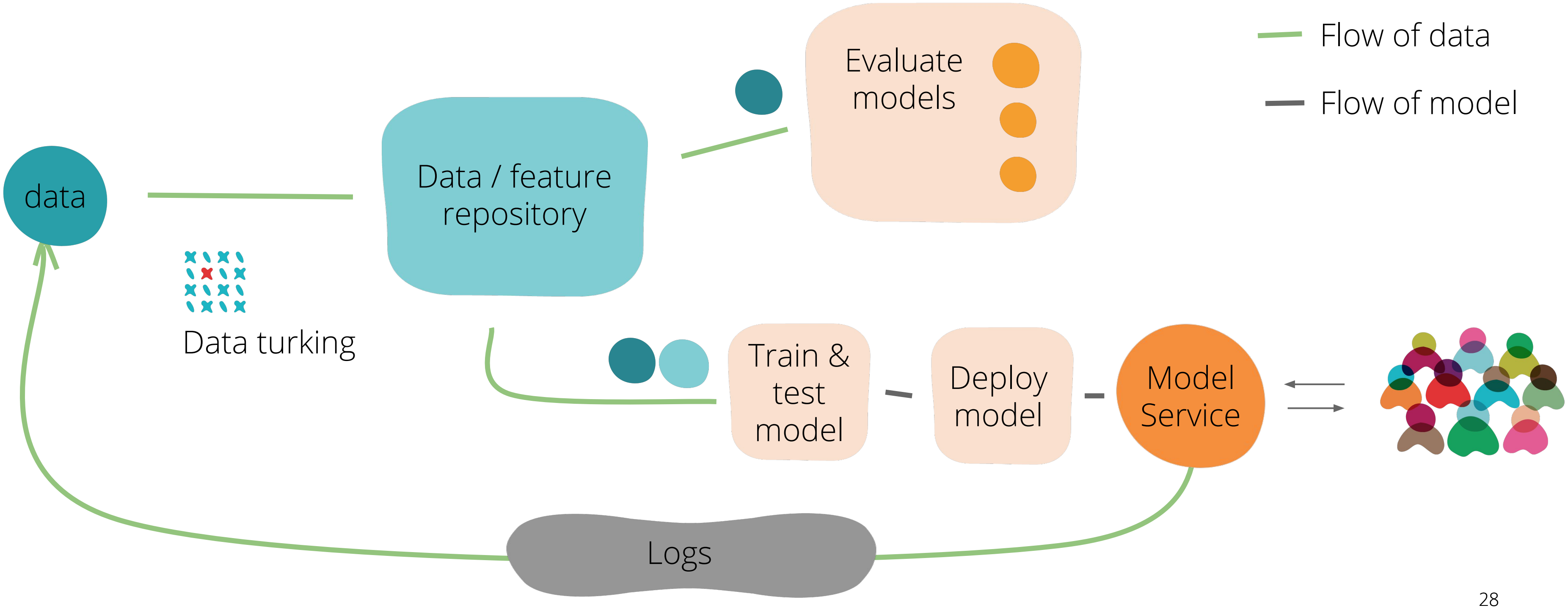


LIME Explanations

Time	Average prediction	Average LIME_RM	Average LIME_CRIM	Average LIME_DIS	Average LIME_INDUS	Average LIME_LSTAT
2019-03-15 11:13:00	34.18	6.29	-	0.97	-	0.29
2019-03-15 11:14:00	29.26	6.34	0.27	0.83	0.25	-3.13
2019-03-15 11:15:00	26.80	6.38	-	0.73	-	-5.09

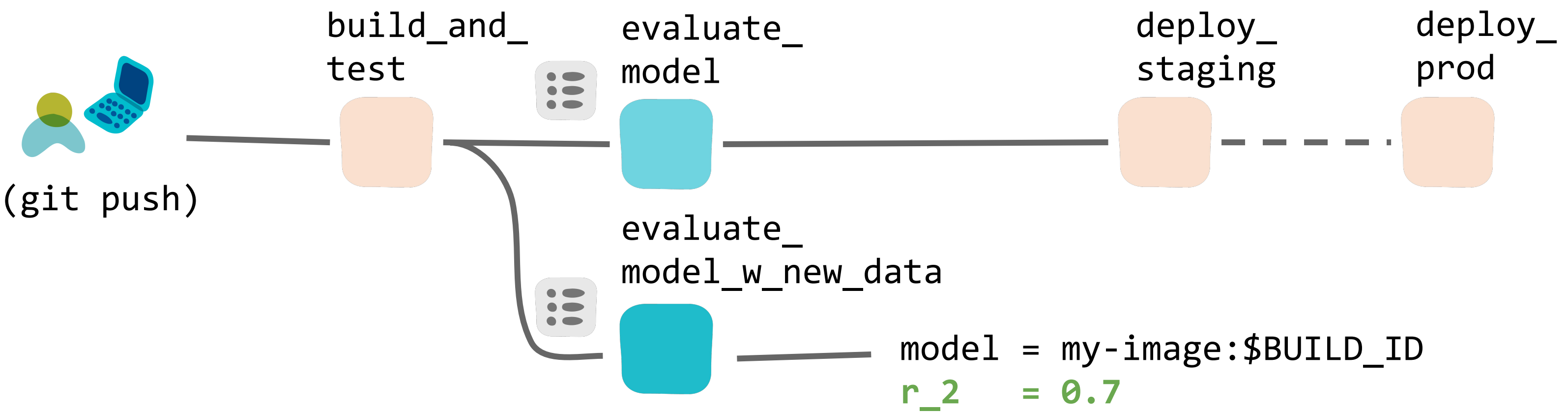
5. OBSERVE!

Benefit #3: Closing the data collection loop



5. OBSERVE!

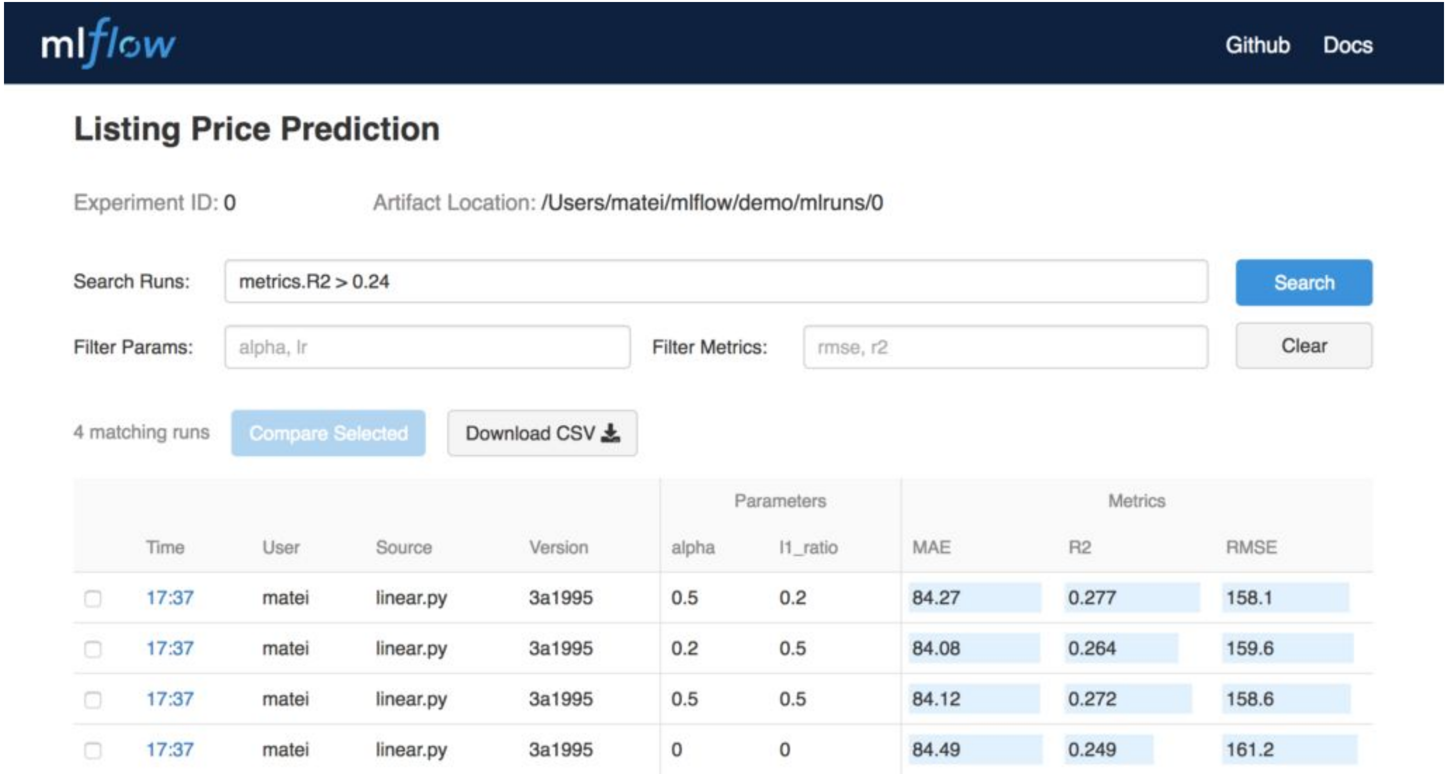
Benefit #4: Ability to measure goodness of any model



```
model = my-image:$BUILD_ID
r_2   = 0.7
rmse  = 42
```

5. OBSERVE!

Benefit #4: Ability to measure goodness of any model



Listing Price Prediction

Experiment ID: 0 Artifact Location: /Users/matei/mlflow/demo/mlruns/0

Search Runs: metrics.R2 > 0.24

Filter Params: alpha, lr

Filter Metrics: rmse, r2

4 matching runs

					Parameters		Metrics		
	Time	User	Source	Version	alpha	l1_ratio	MAE	R2	RMSE
□	17:37	matei	linear.py	3a1995	0.5	0.2	84.27	0.277	158.1
□	17:37	matei	linear.py	3a1995	0.2	0.5	84.08	0.264	159.6
□	17:37	matei	linear.py	3a1995	0.5	0.5	84.12	0.272	158.6
□	17:37	matei	linear.py	3a1995	0	0	84.49	0.249	161.2

5. HOW'S THE MODEL IN THE WILD?

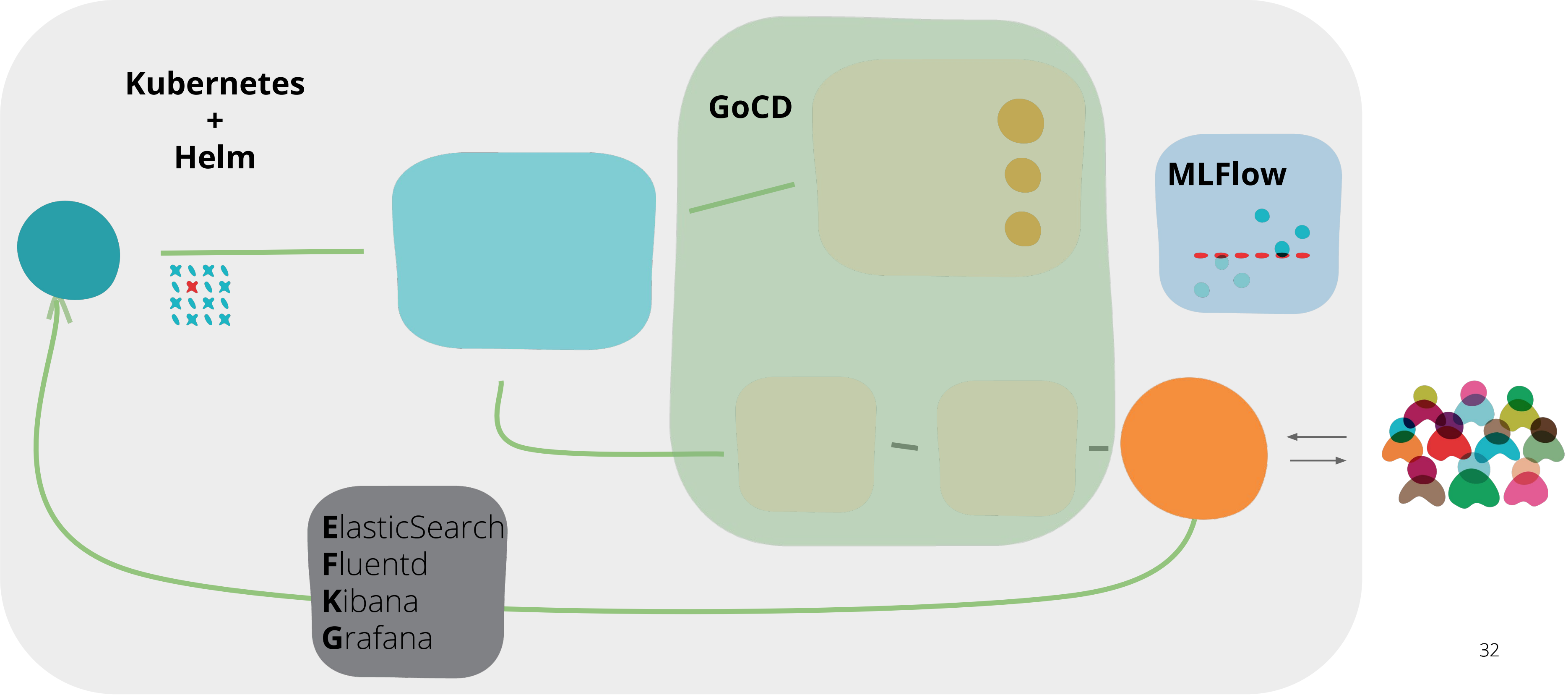
OBSERVE!

Summing up

- Mitigation measures
  - Logging + Monitoring
- Benefits
  - Feedback on production models
  - Interpretability (how did the model decide on this particular prediction?)
  - Better data for training
  - Better (unseen) data for evaluating candidate/champion models

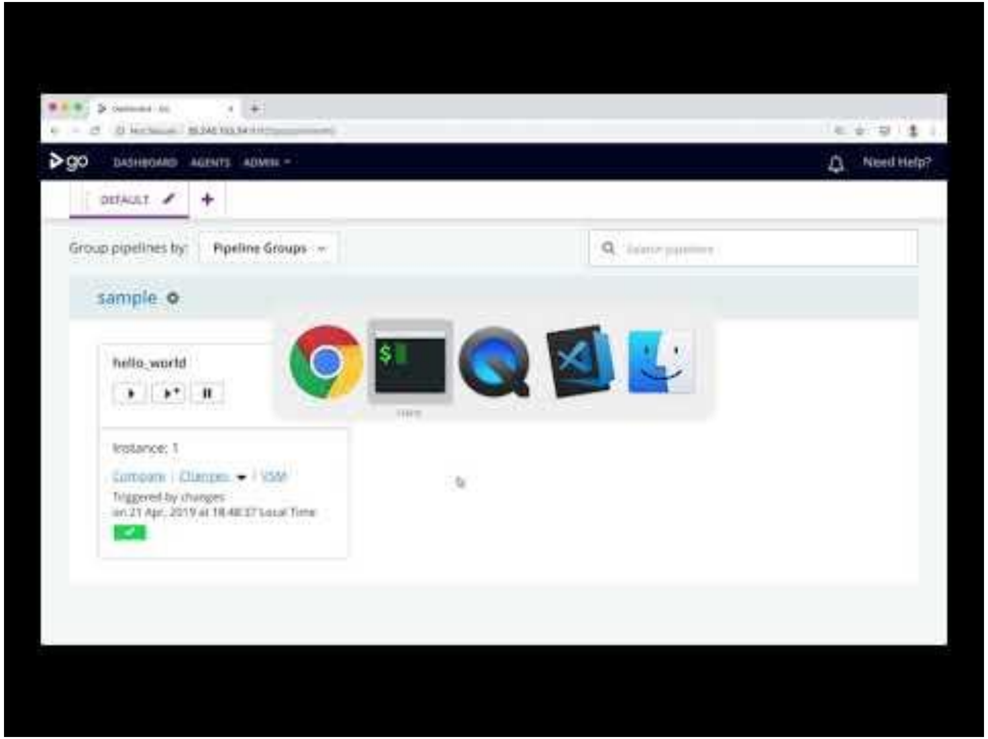
5. HOW'S THE MODEL IN THE WILD?

Demo



5. HOW'S THE MODEL IN THE WILD?

Demo



hello_world

Instance: 1

Compare | Changes | VSM

Triggered by changes
on 21 Apr, 2019 at 18:48:37 Local Time

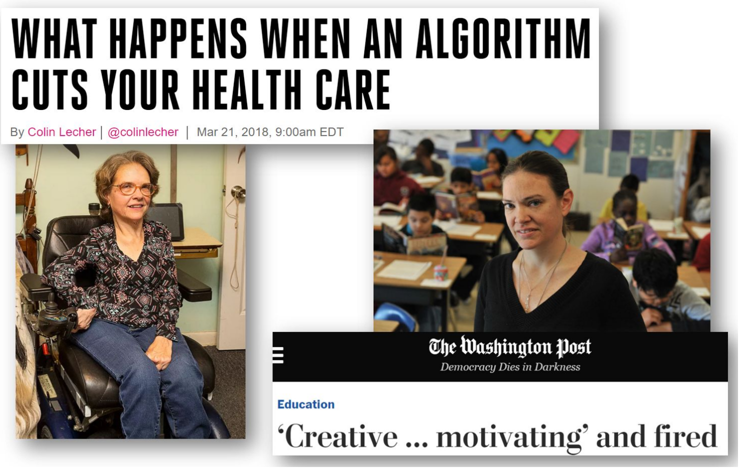
6

HARMFUL MODELS IN PRODUCTION

6. HARMFUL MODELS IN PRODUCTION

Actual news headlines

- PredPol algorithm reinforces racial biases in policing data
- Recruiting tool shows bias against women



WHAT HAPPENS WHEN AN ALGORITHM CUTS YOUR HEALTH CARE

By Colin Lecher | @colinlecher | Mar 21, 2018, 9:00am EDT

The Washington Post

Democracy Dies in Darkness

Education

'Creative ... motivating' and fired

Image source: I'm an AI researcher, and here's what scares me about AI (Rachel Thomas)

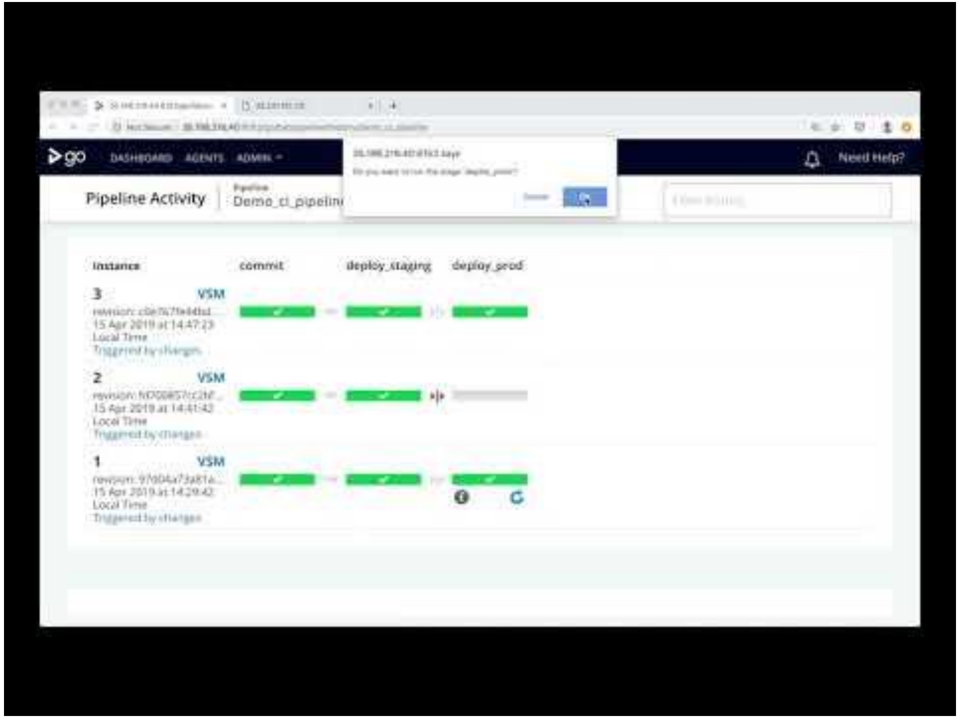
6. HARMFUL MODELS IN PRODUCTION

Mitigation measures

- Discuss and define what “bad” looks like in our context
- “Black mirror” retros
- Measure unfairness
    - Make fairness a measurable fitness function
- Data ethics checklist (link)
- Human-in-the-loop / appeal processes
- Ability to recover from harmful models

6. HARMFUL MODELS IN PRODUCTION

Demo: rollback to last good build



Do you want to run the stage 'deploy_prod'?

instance	commit	deploy_staging	deploy_prod
3 VSM revision: c0e7e7fe44bd... 15 Apr 2019 at 14:47:23 Local Time Triggered by changes	✓	✓	✓
2 VSM revision: 5f00857cc2bf... 15 Apr 2019 at 14:41:42 Local Time Triggered by changes	✓	✓	
1 VSM revision: 97004a73a81a... 15 Apr 2019 at 14:29:42 Local Time Triggered by changes	✓	✓	✓

SUMMING UP

How can we make easier to do the right thing?

MAKE IT EASIER TO DO THE RIGHT THING

- Better habits
  - Environment management
  - Closing the data collection loop
  - Deploy early and often
  - Automated tracking of hyperparameters and metrics
  - Logging and monitoring
  - Do no harm
- Two reference repos:
  - `github.com/ThoughtWorksInc/ml-cd-starter-kit`
  - `github.com/ThoughtWorksInc/ml-app-template`

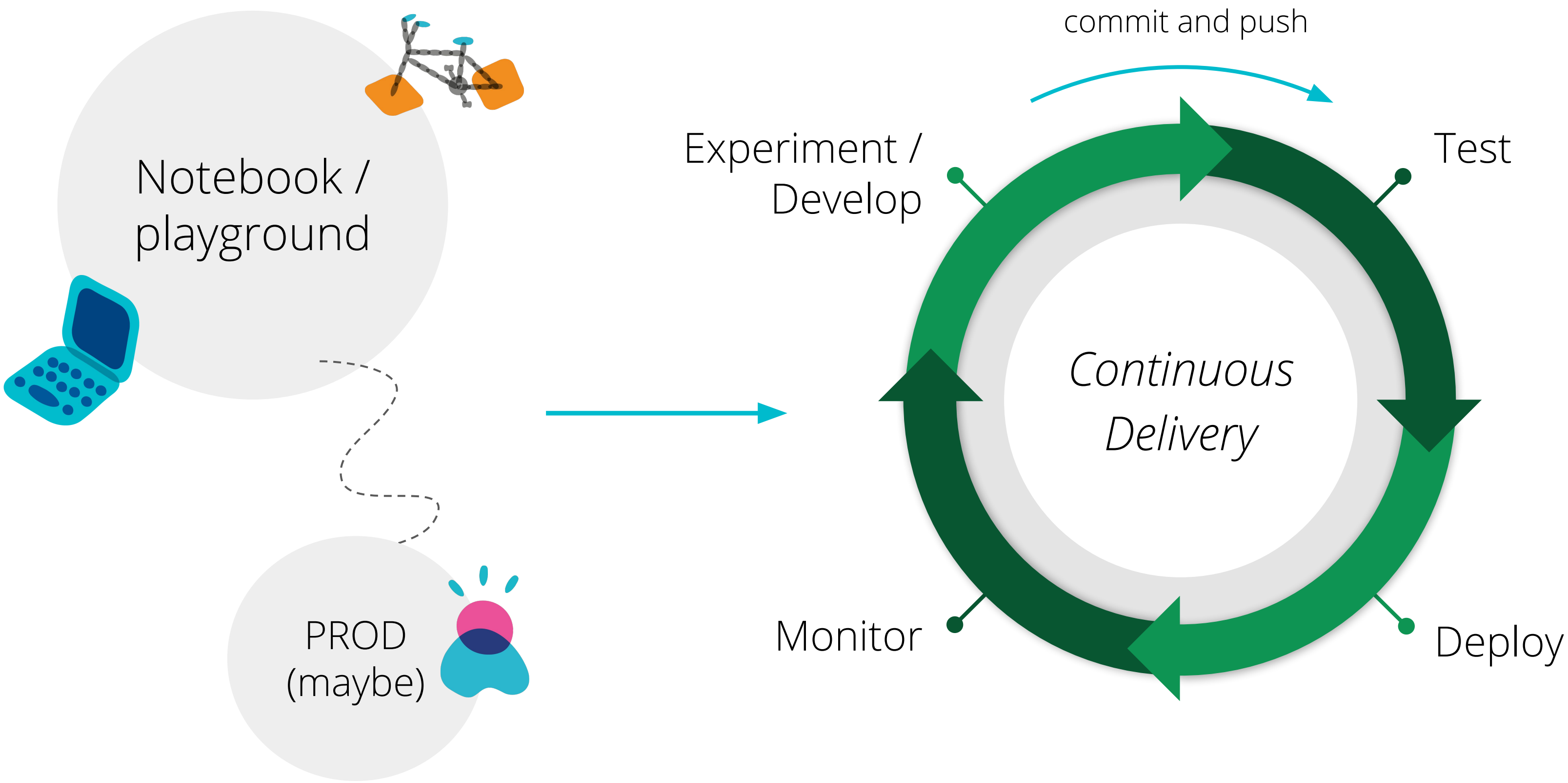
`github.com/ThoughtWorksInc/ml-cd-starter-kit`

- ☑ Provision and configure cross-cutting services
  - ☑ GoCD
  - ☑ EFKG
  - ☑ MLFlow

`github.com/ThoughtWorksInc/ml-app-template`

- ☑ Project boilerplate template
  - ☑ Unit tests
  - ☑ Train model
  - ☑ Test model metrics
  - ☑ Dockerised setup
- ☑ Store CI pipeline as code
- ☑ Track hyperparameters and metrics of each training run on CI
- ☑ Logging (predictions, inputs, explanatory variables)

SUMMING UP



Notebook / playground

PROD (maybe)

commit and push

Experiment / Develop

Test

Deploy

Monitor

Continuous Delivery

FURTHER READING

- Slides: tinyurl.com/ml-xconf
- Search: ThoughtWorks CD4ML

THANK YOU.

David Tan
davidtan@thoughtworks.com

tinyurl.com/ml-xconf