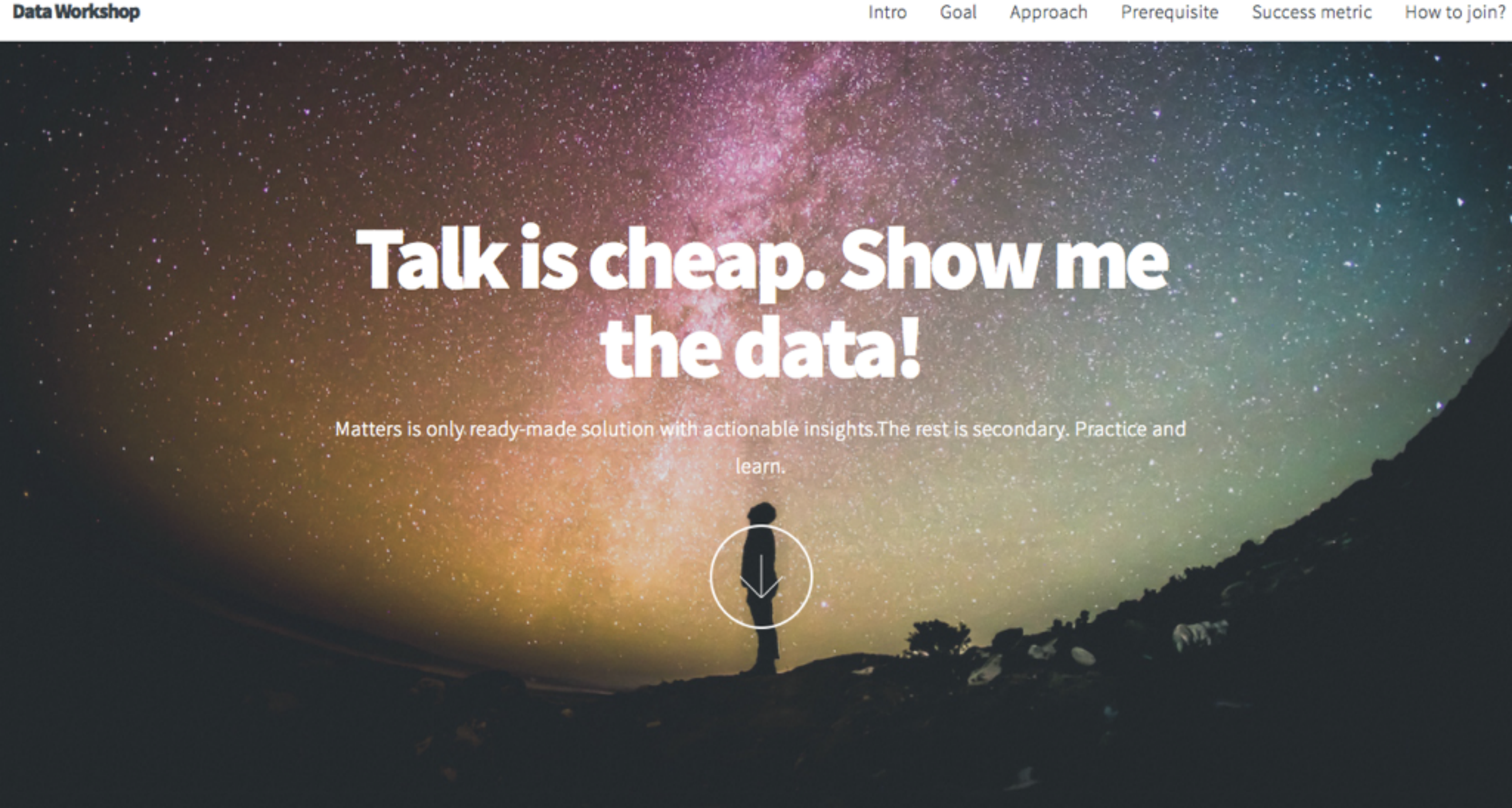
Data Workshop #3

Evaluating ML Models

dataworkshop.eu

Alekseichenko Vladimir

DataWorkshop.eu



About me

Vladimir Alekseichenko

Love analyze data

Architect

slon1024

slon1024

vova@vova.me

Disclaimer

Data Workshop *[all time]* focuses on the **intuition** and **practical** tips.

For a formal treatment, see something else.*

* papers or classical machine learning books

Environment

github.com/dataworkshop/prerequisite

github.com/dataworkshop/model_evaluation

Packages

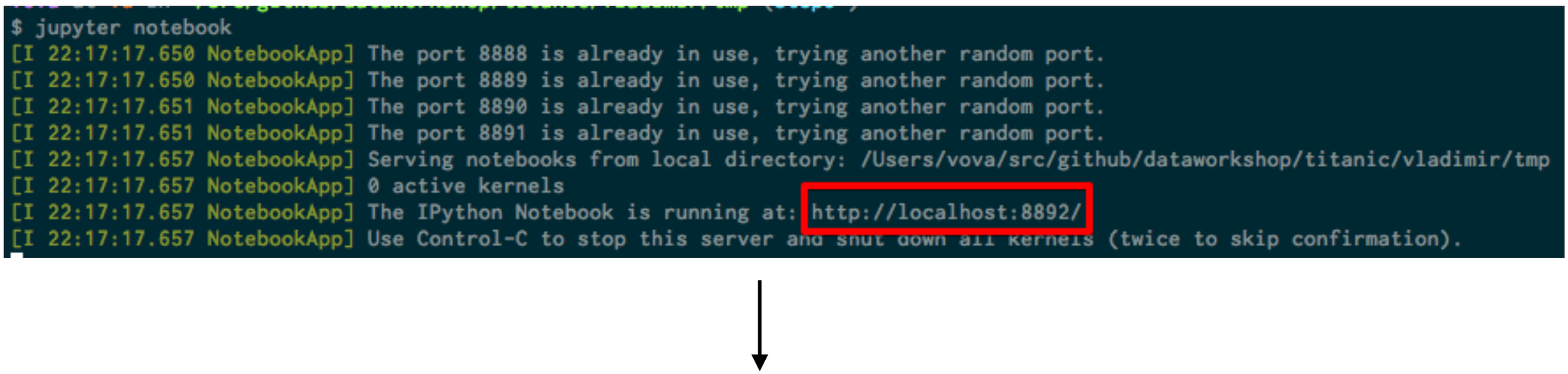
github.com/**dataworkshop/prerequisite**

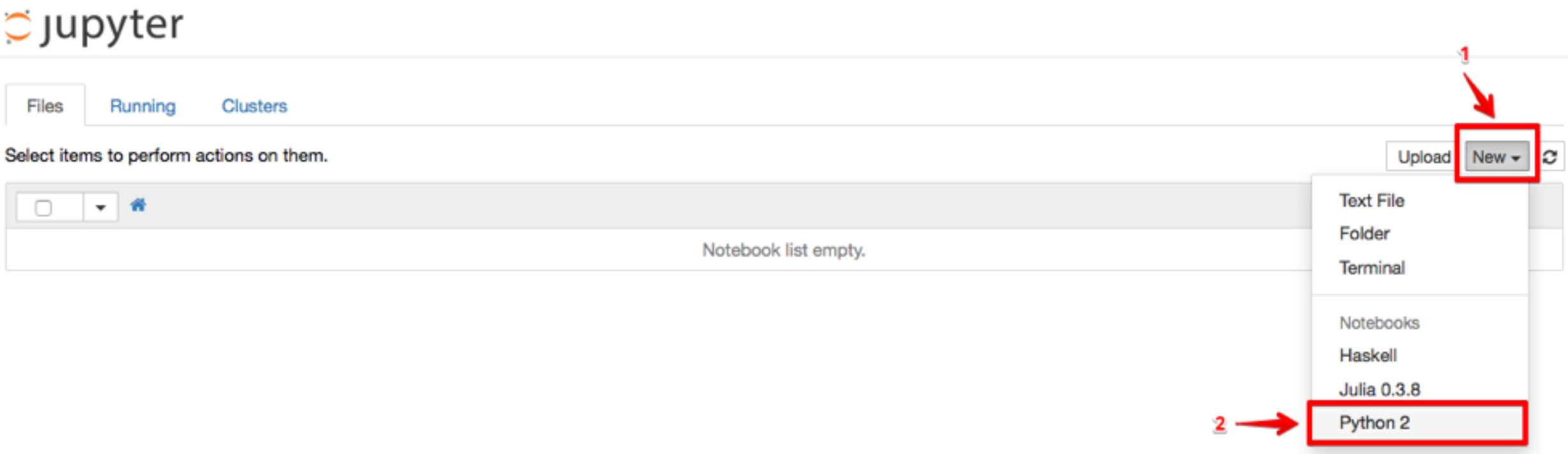
```
$ python run.py
seaborn-0.7.0 - OK
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
All right, you are ready to go on Data Workshop!
```

```
$ python run.py
seaborn-0.6 should be upgraded to seaborn-0.7
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
RECOMENDATION (without upgrade some needed features could be missing)
pip install --upgrade seaborn
```

```
$ python run.py
seaborn-0.7.0 - OK
xgboost - missing
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
REQUIRED
Please install those packages before Data Workshop: xgboost
pip install xgboost
More info how to install xgboost: http://xgboost.readthedocs.org/en/latest/build.html
```

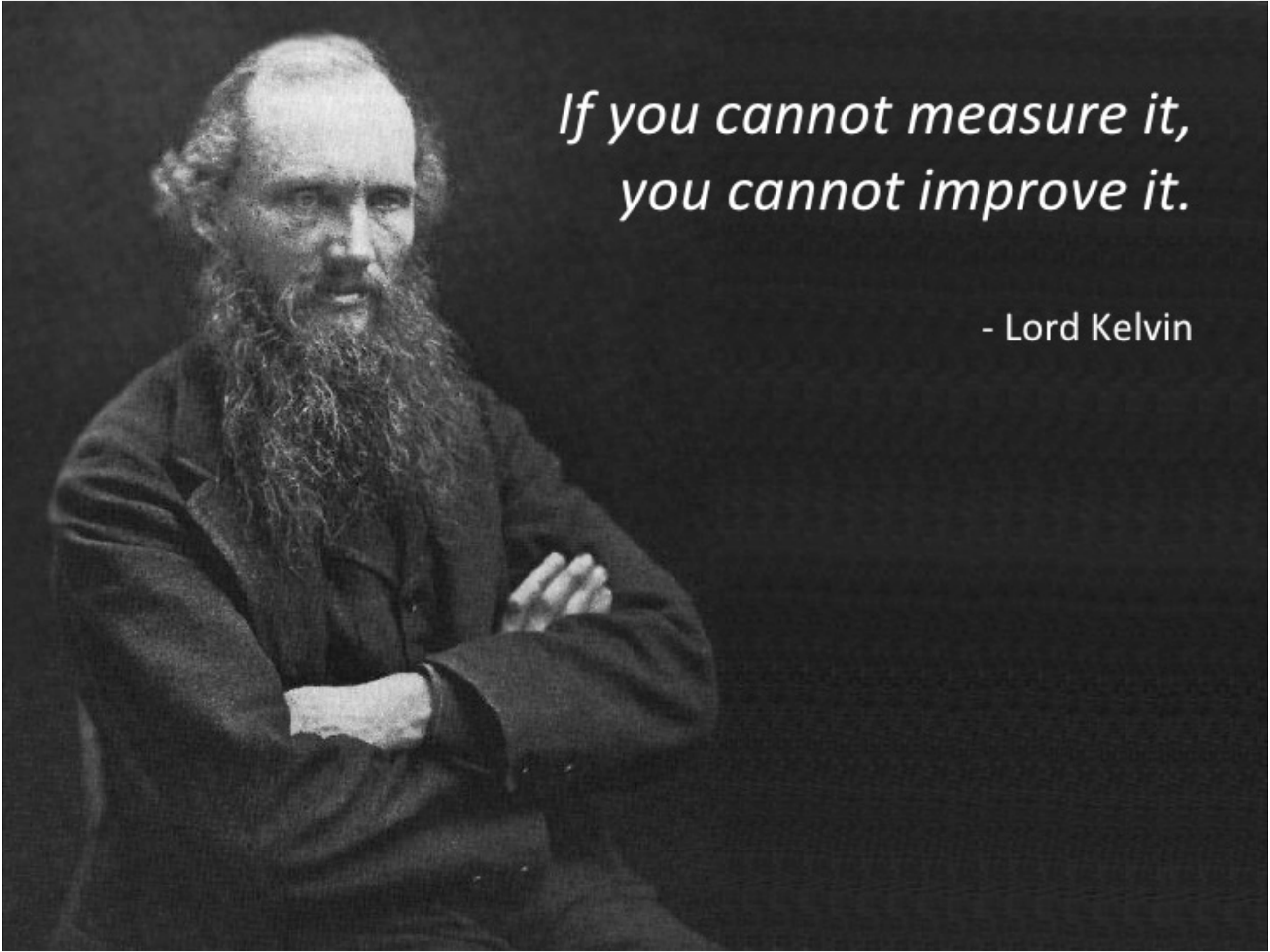
jupyter notebook



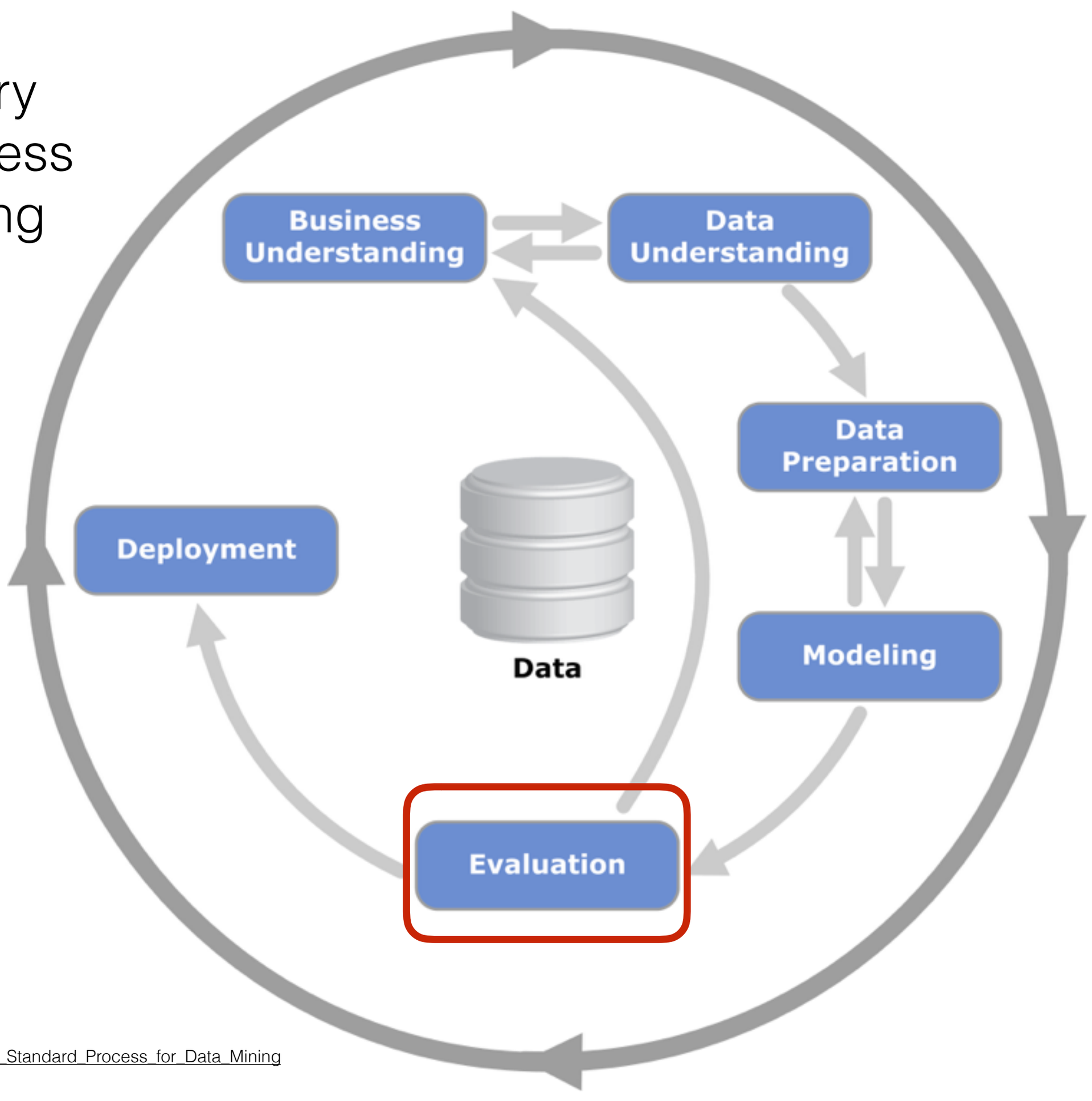


Motivation

metrics & validation


If you cannot measure it,
you cannot improve it.
- Lord Kelvin

Cross Industry Standard Process for Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Understand Business & Data

Read and explore data

Chanel_ID	Client_ID	Product_ID	Demand
1	3	2	10
1	3	5	15
3	3	6	12

Feature Engineering

Create a new ones based on already exists

Chanel_ID	Client_ID	Product_ID	…	Demand	DemandLog
1	3	2	…	10	2.303
1	3	5	…	15	2.708
3	3	6	…	12	2.485

Evaluation is needed

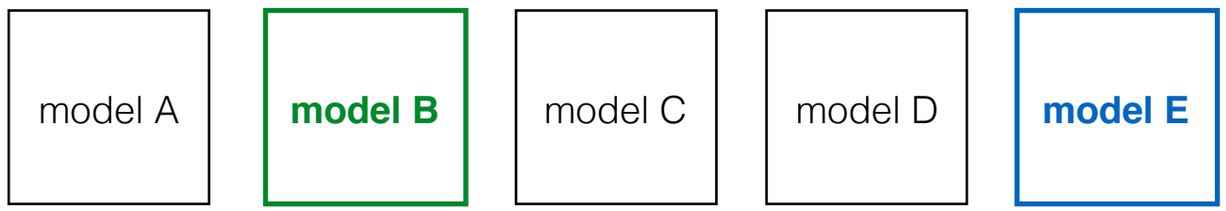
Feature Selection

Select only useful features

~~Chanel_ID~~	Client_ID	Product_ID	~~…~~	~~Demand~~	DemandLog
	3	2	…		2.303
	3	5	…		2.708
	3	6	…		2.485

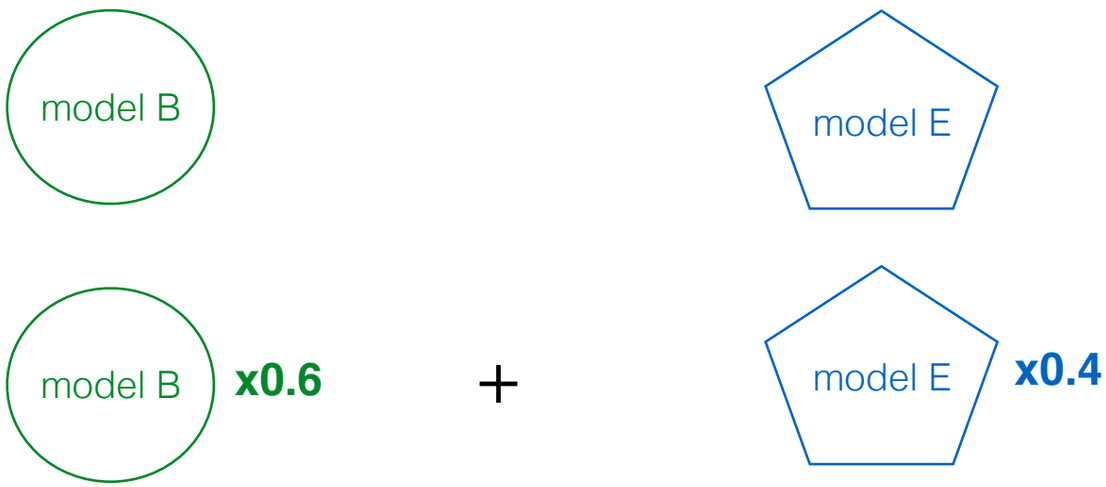
Model Selection

Find the best model(s)



Tuning Hyperparameters

Find the best hyperparameters for given model



Ensemble Modeling

Combine few models into one more better

Simple questions

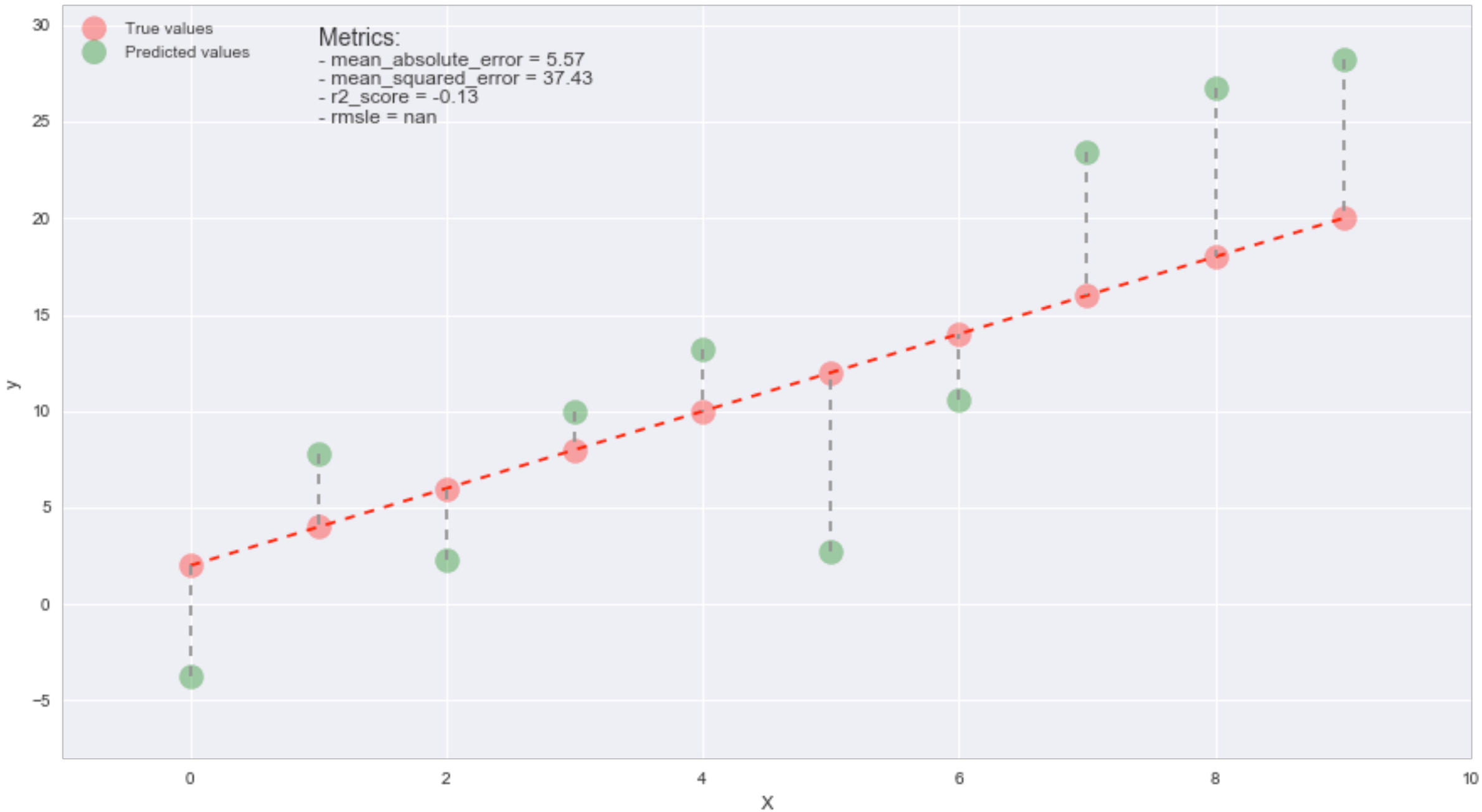
- How can i measure success for this project?
- How would i know when i’ve succeeded?
- Is it possible succeed for this project?

Regression

mse, rmse, rmsle, r2

`sklearn.metrics.*`

Regression


True values
Predicted values
Metrics:
- mean_absolute_error = 5.57
- mean_squared_error = 37.43
- r2_score = -0.13
- rmsle = nan
30
25
20
15
10
5
0
−5
y
0
2
4
6
8
10
X

Simple points about metrics

- Range
- Outliers
- Negative & Positive

Classification

accuracy, confusion-matrix, f1, log-loss, auc

```
sklearn.metrics.*
```

Classification

Predicting class labels given input data

- Binary - only two labels (e.g spam or ham)

- Multi-class - more than two possibles classes (e.g. good, neutral, bad)

Classification - binary

Ferrari

or

Lamborghini

Classification - multi-class

Apple

or

Pear

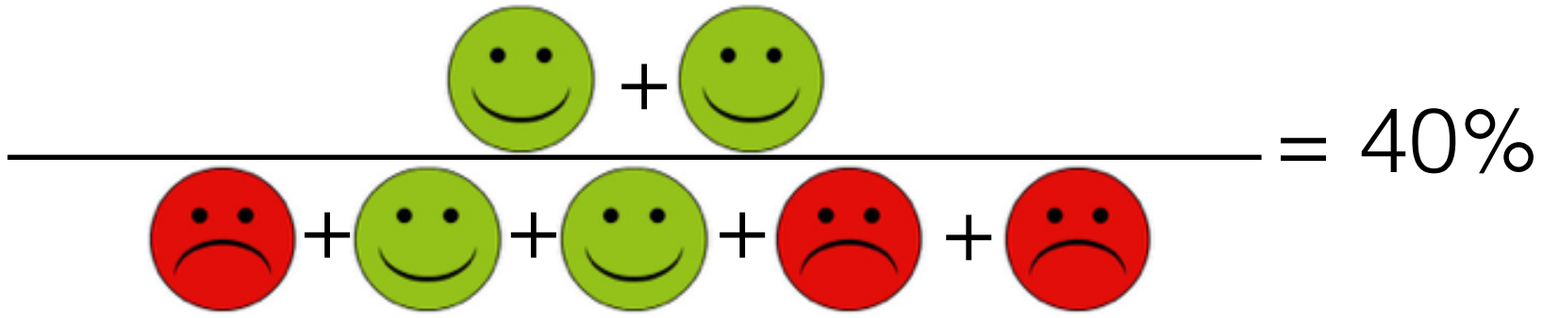
or

Lemon

or

Orange

Accuracy



#	actual	predicted	status
1			
2			
3			
4			
5			

Accuracy Paradox

prediction

true	yes	no
yes	1	5
no	100	10000

accuracy

$$\frac{1 + 10000}{1 + 5 + 100 + 10000} = 98.96\%$$

prediction

true	yes	no
yes	0	6
no	0	10100

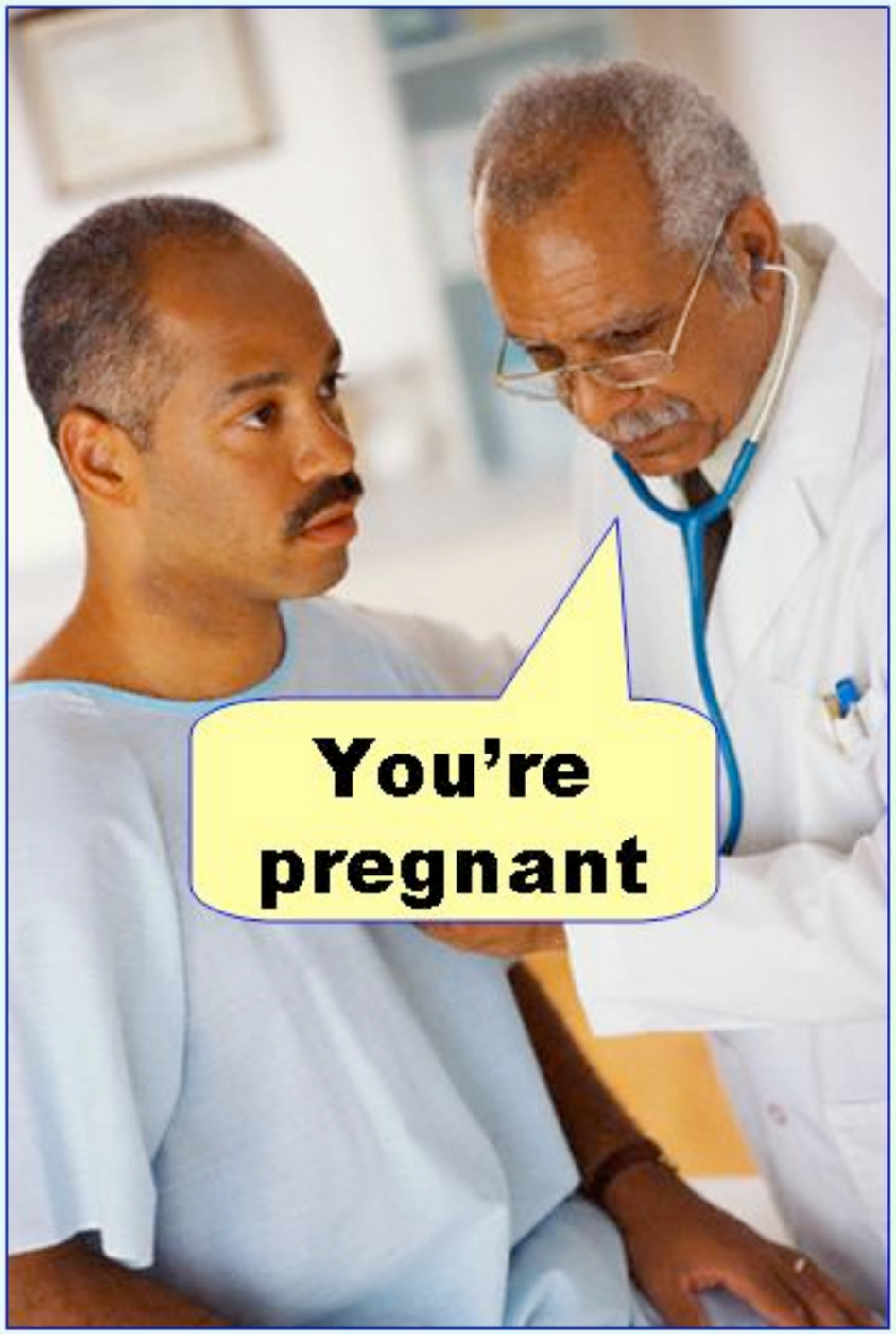
accuracy

$$\frac{10100}{10100 + 6} = 99.94\%$$

https://en.wikipedia.org/wiki/Accuracy_paradox

Type I error

(false positive)



Type II error

(false negative)

nfusion Ma

Confusion Matrix		**Model**			
		Positive	Negative		
TRUE	Positive	a	b	Sensitiviy, (recall, TPR)	b/(a+b)
	Negative	c	d	FPR	c/(c+d)
		Precision	False omission rate	Accuracy = (a+d) / (a+b+c+d)	
		a/(a+c)	b/(b+d)		

ROC Curves

Receiver Operating Characteristic

```
sklearn.metrics.*
```

Royal **O**bserver **C**orps

Post Instrument - prior to the introduction of radar, were commonly used to spot and report aircraft positions.

https://en.wikipedia.org/wiki/Post_Instrument

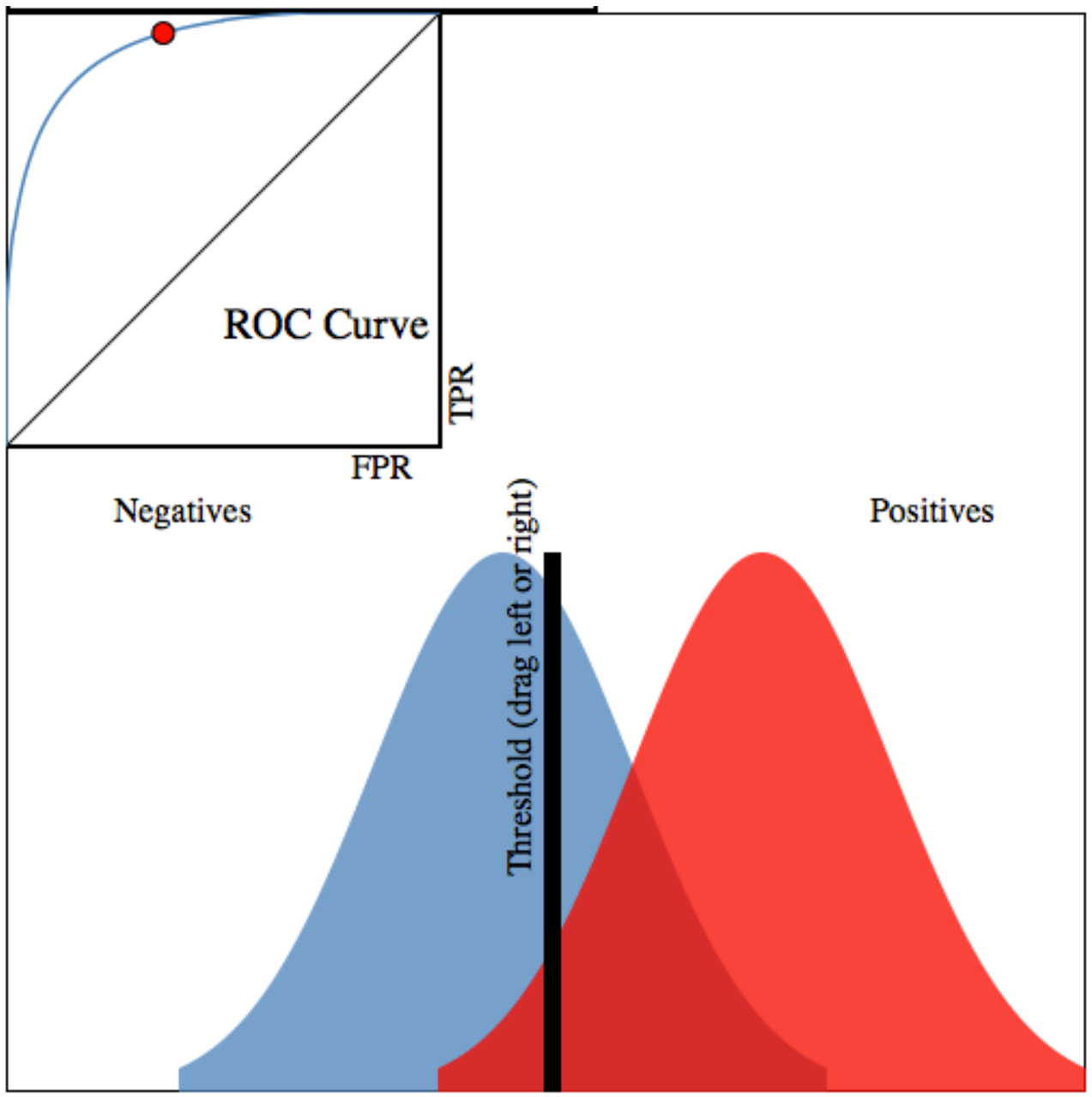
Receiver Operating Characteristic

First* Great Britain Radar

** at least one of them*

For a first radar, was a challenge to distinguish **bird** and **plane**

ROC curves



http://navan.name/roc/

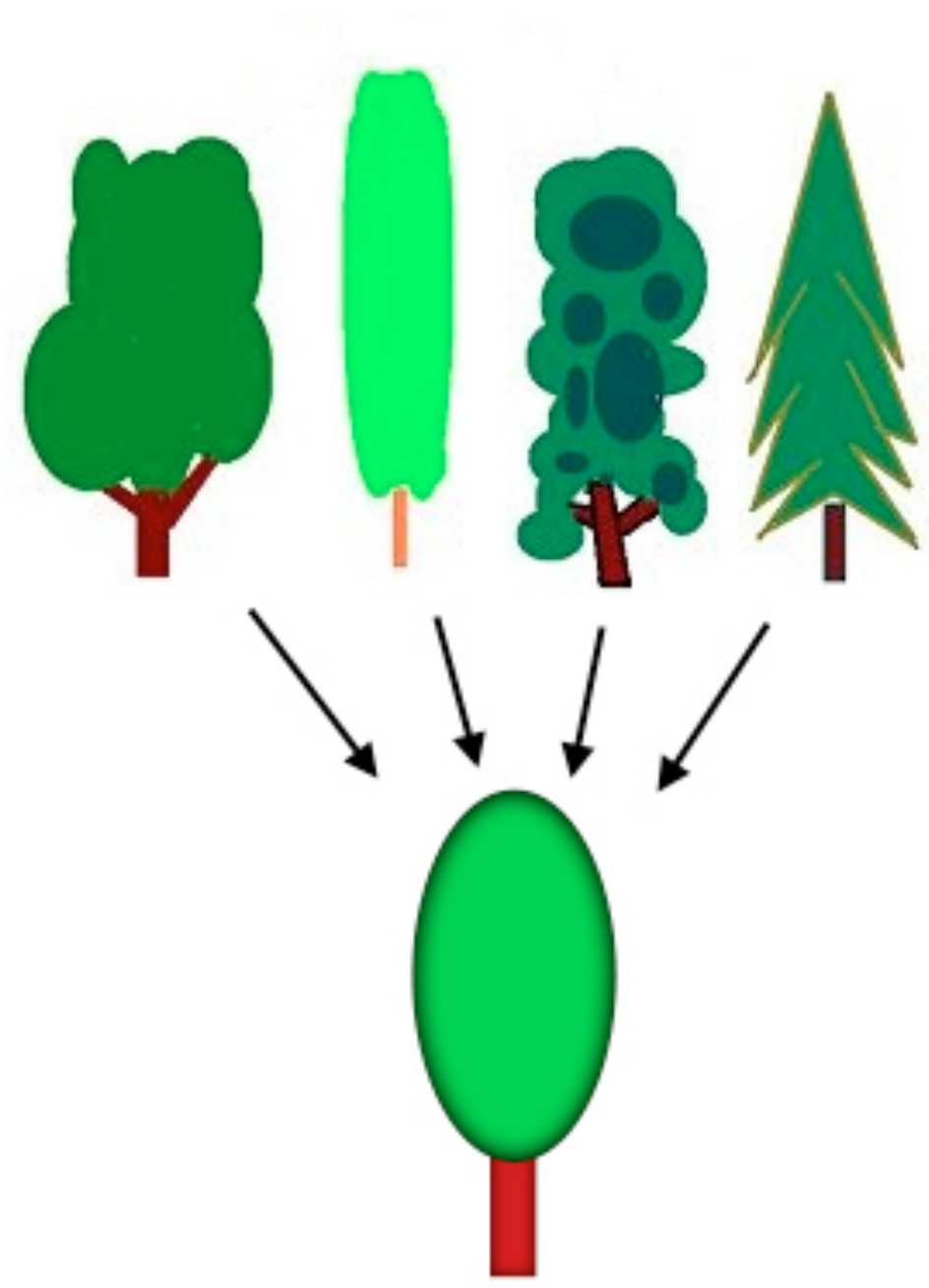

The Boy Who Cried Wolf
Ian Healey

Validation

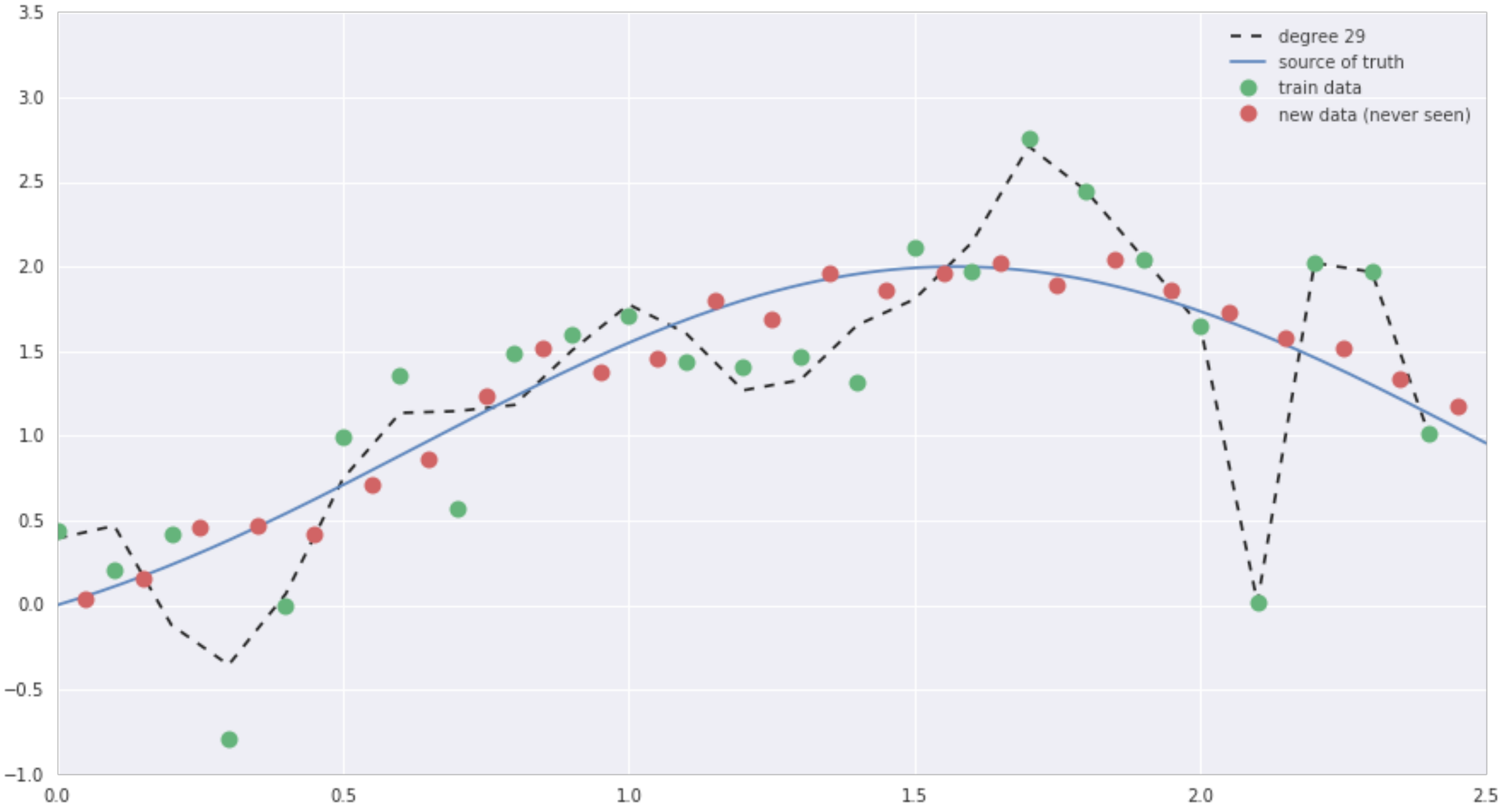
cross-validation (k-fold, nested), bootstraping

```
sklearn.cross_validation.*
```

Generalization



Overfitting



Hold Out Validation

Split for two datasets

```
sklearn.cross_validation.train_test_split
```

Bootstraping

Age	Gender	Height
20	male	170
30	female	163
24	male	198
28	female	169

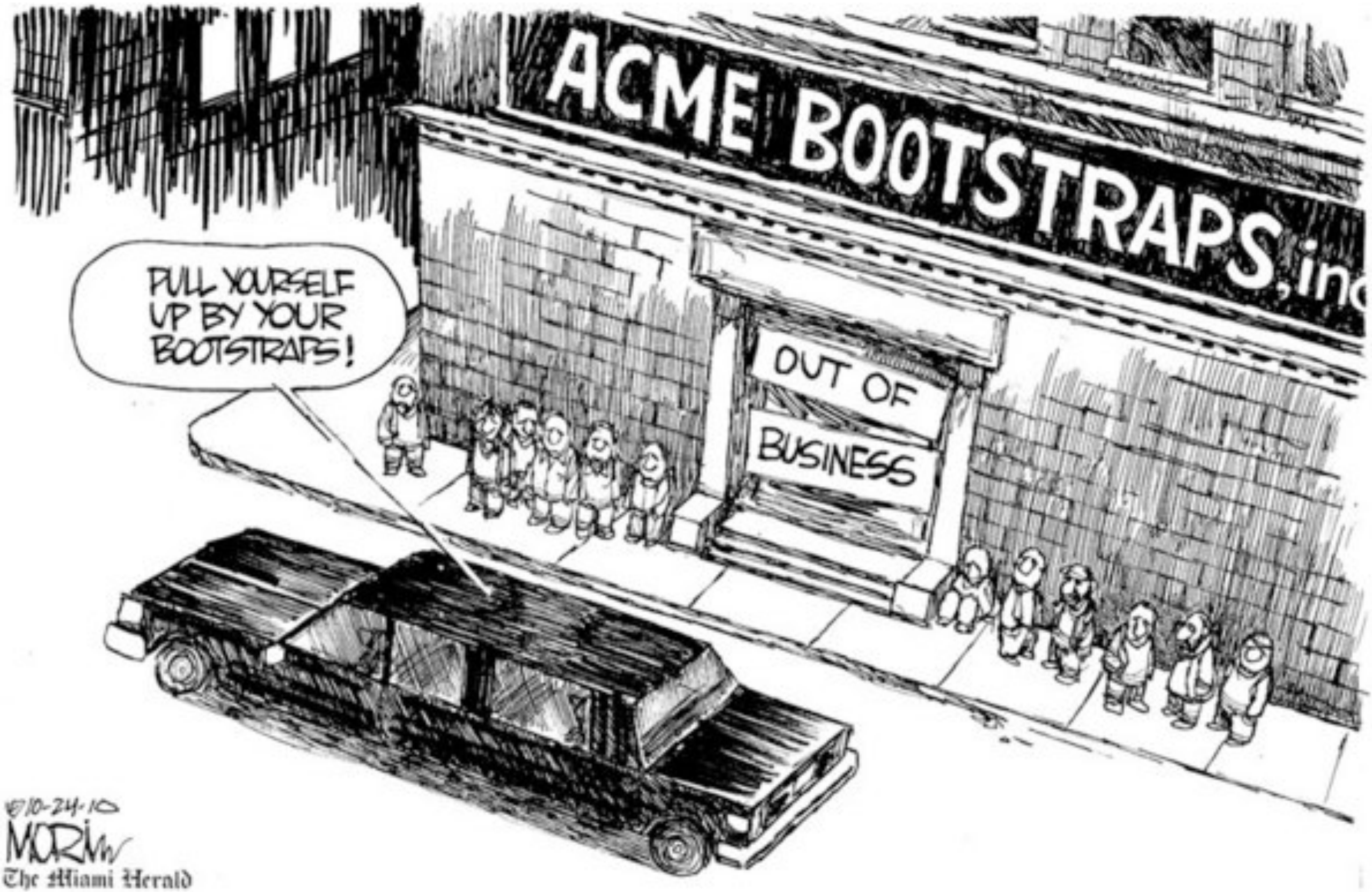
Age	Gender	Height
20	male	170
30	female	163
24	male	198

Age	Gender	Height
28	female	169

Bootstraping

resampling

```
sklearn.utils.resample
```


ACME BOOTSTRAPS, inc
PULL YOURSELF UP BY YOUR BOOTSTRAPS!
OUT OF BUSINESS
10-24-10
MORIN
The Miami Herald

Bootstraping

Age	Gender	Height
20	male	170
30	female	163
24	male	198
28	female	169

Age	Gender	Height
20	male	170
24	male	198
28	female	169

Age	Gender	Height
20	male	170
30	female	163
30	female	163

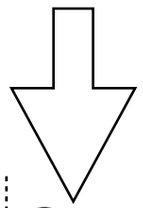
Age	Gender	Height
30	female	163
24	male	198
28	female	169

Cross validation

k-fold, leave-one-out, stratified

Cross-Validation: 3-fold

Age	Gender	Height
20	male	170
30	female	163
24	male	198
28	female	169



Age	Gender	Height
20	male	170
30	female	163
24	male	198

Age	Gender	Height
28	female	169
20	male	170
30	female	163

Age	Gender	Height
24	male	198
28	female	169
20	male	170

Age	Gender	Height
28	female	169

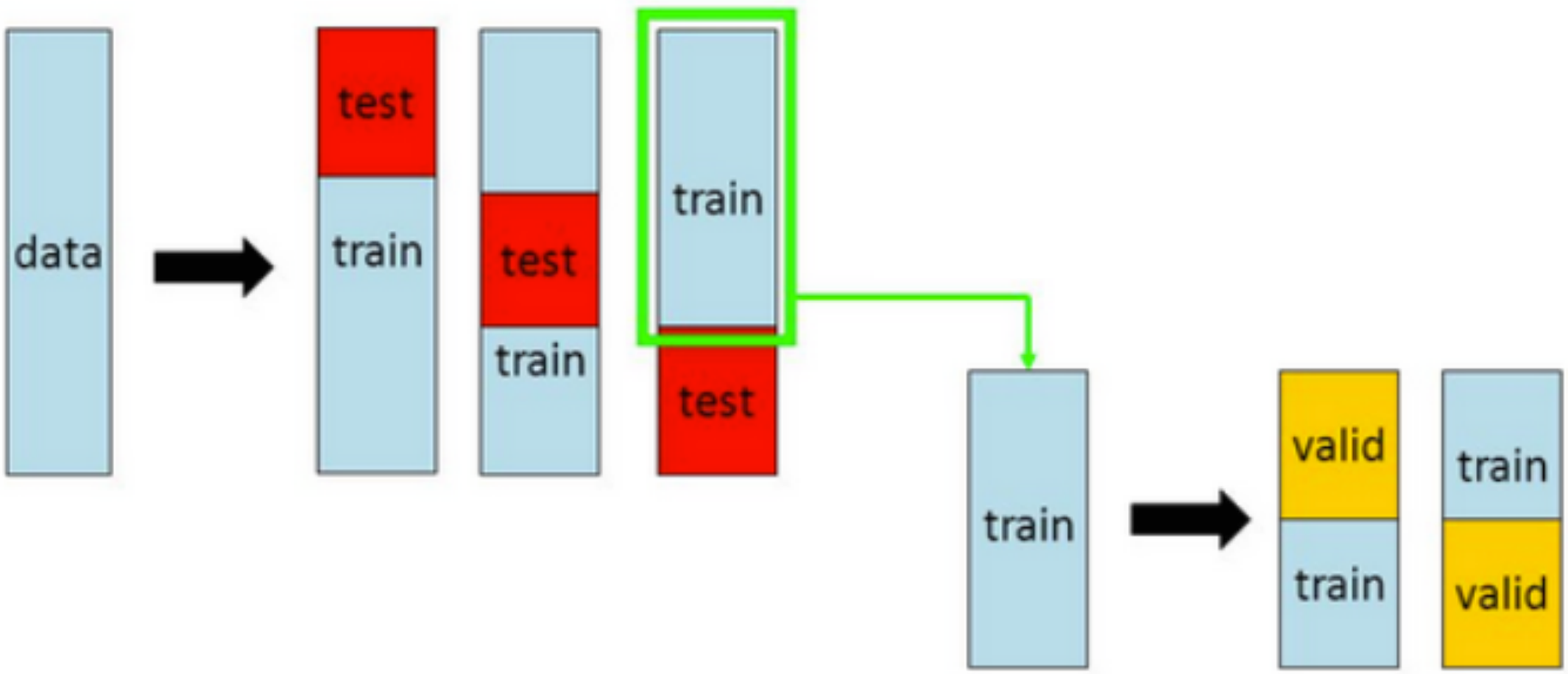
Age	Gender	Height
24	male	198

Age	Gender	Height
30	female	163

Nested Cross Validation

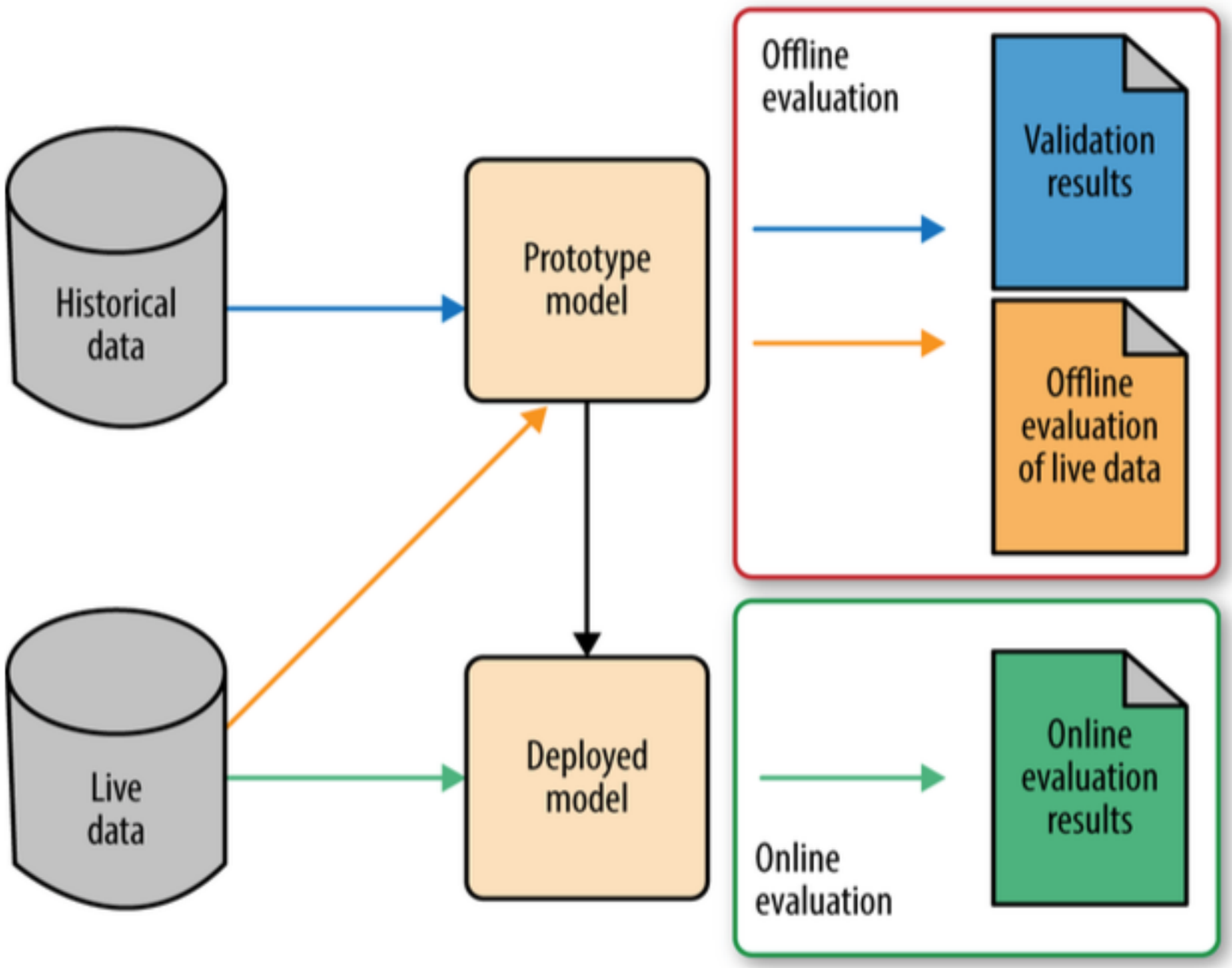
More advanced approach

Nested Cross Validation



Summary

Development & evaluation model



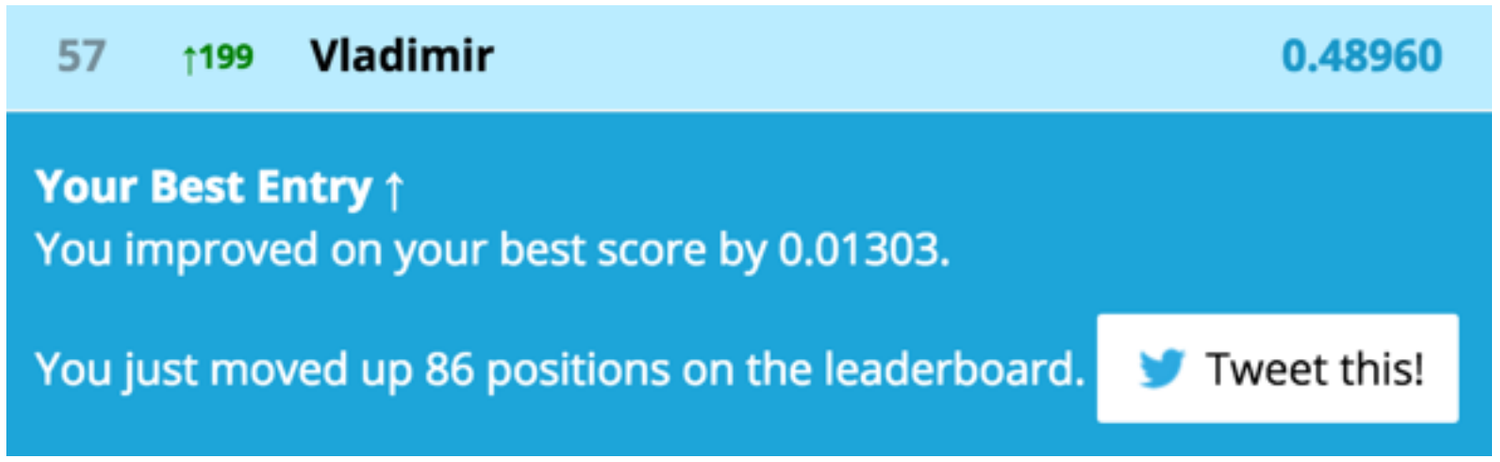
Three things

if you can remember only three…

- Understand expected success by metric[s]
- Never mix training data and evaluation data
- The distribution of data changes over time

New challenge:

Maximize sales and minimize returns of bakery goods



Are you ready to join?

Thank you