Let's start **solve** ~~problems~~ **challenges** on **Kaggle**

Vladimir Alekseichenko

About me

Vladimir Alekseichenko

Love analyze data

Architect Search Platform

slon1024
slon1024
vova@vova.me

Data Science

Mainstream

Deep Learning

Pattern recognition

Machine Learning

Data Science

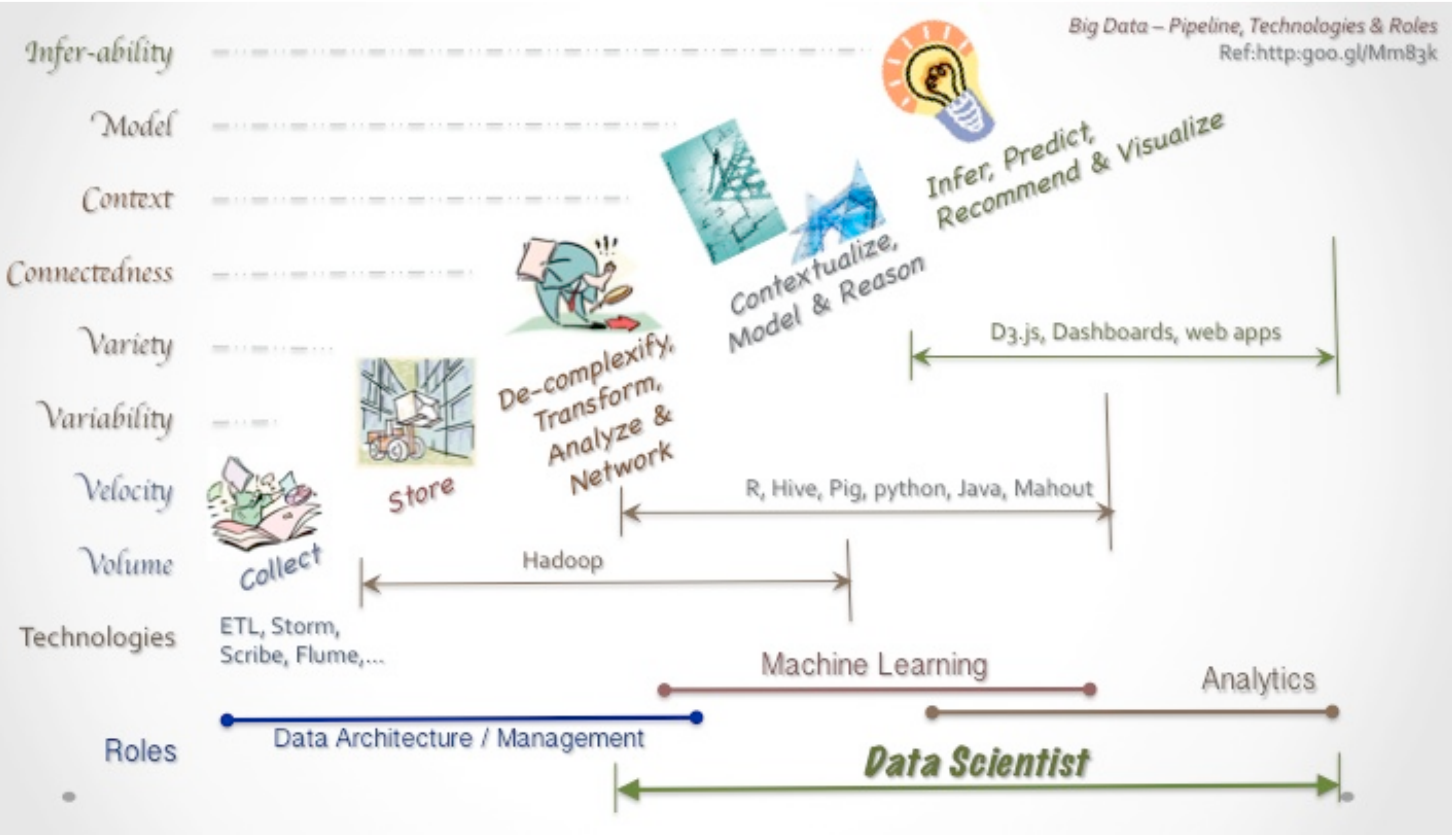
Big Data

Data Mining

Statistics

GET IN LOSERS
WE'RE GOING TO
DO SCIENCE.

Big Data – Pipeline, Technologies & Roles
Ref:http:goo.gl/Mm83k



Infer-ability

Model

Context

Connectedness

Variety

Variability

Velocity

Volume

Technologies

Roles

Collect

Store

De-complexify, Transform, Analyze & Network

Contextualize, Model & Reason

Infer, Predict, Recommend & Visualize

ETL, Storm, Scribe, Flume,...

Hadoop

R, Hive, Pig, python, Java, Mahout

D3.js, Dashboards, web apps

Machine Learning

Analytics

Data Architecture / Management

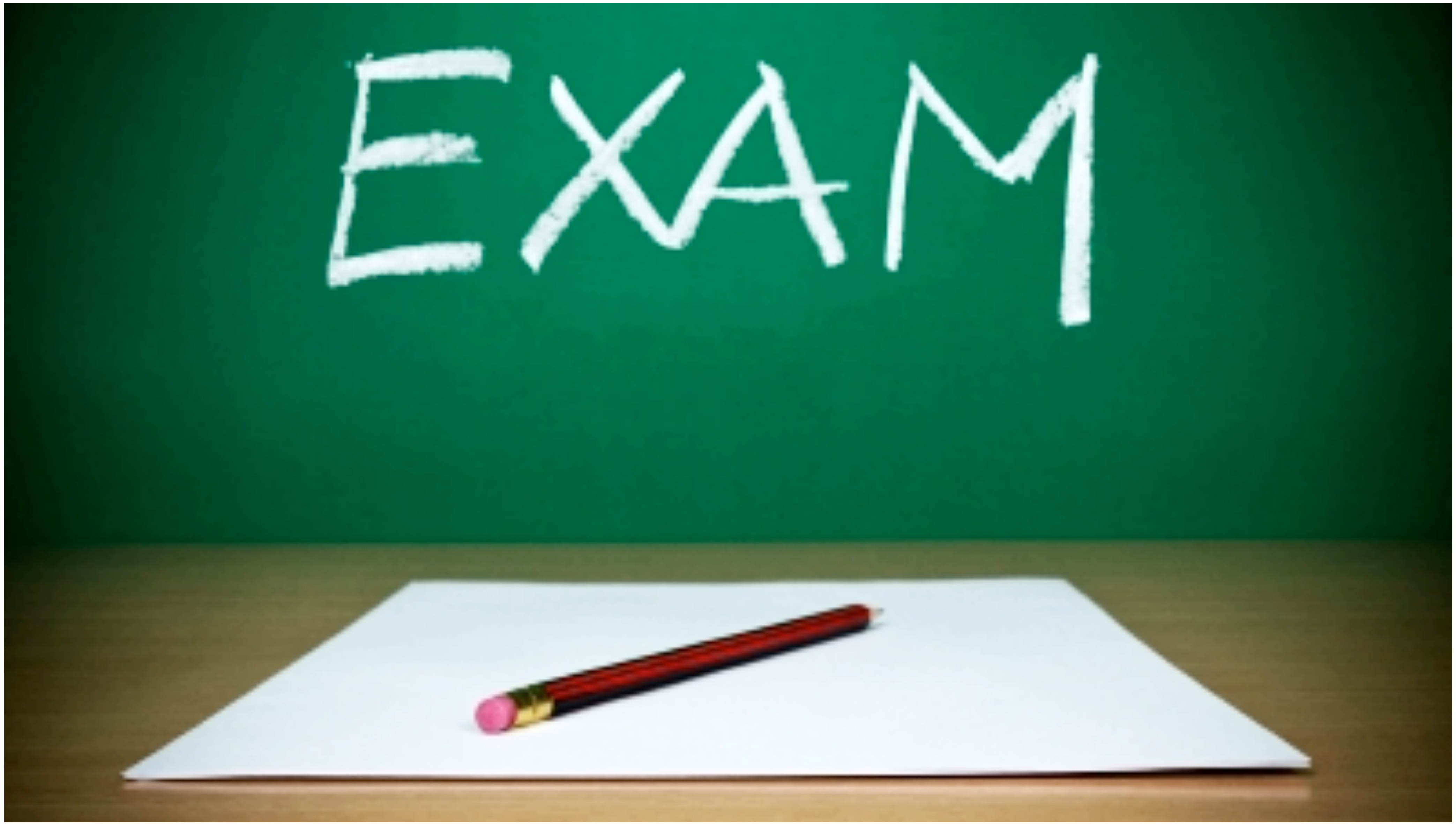
Data Scientist

Let’s focus on

Machine Learning

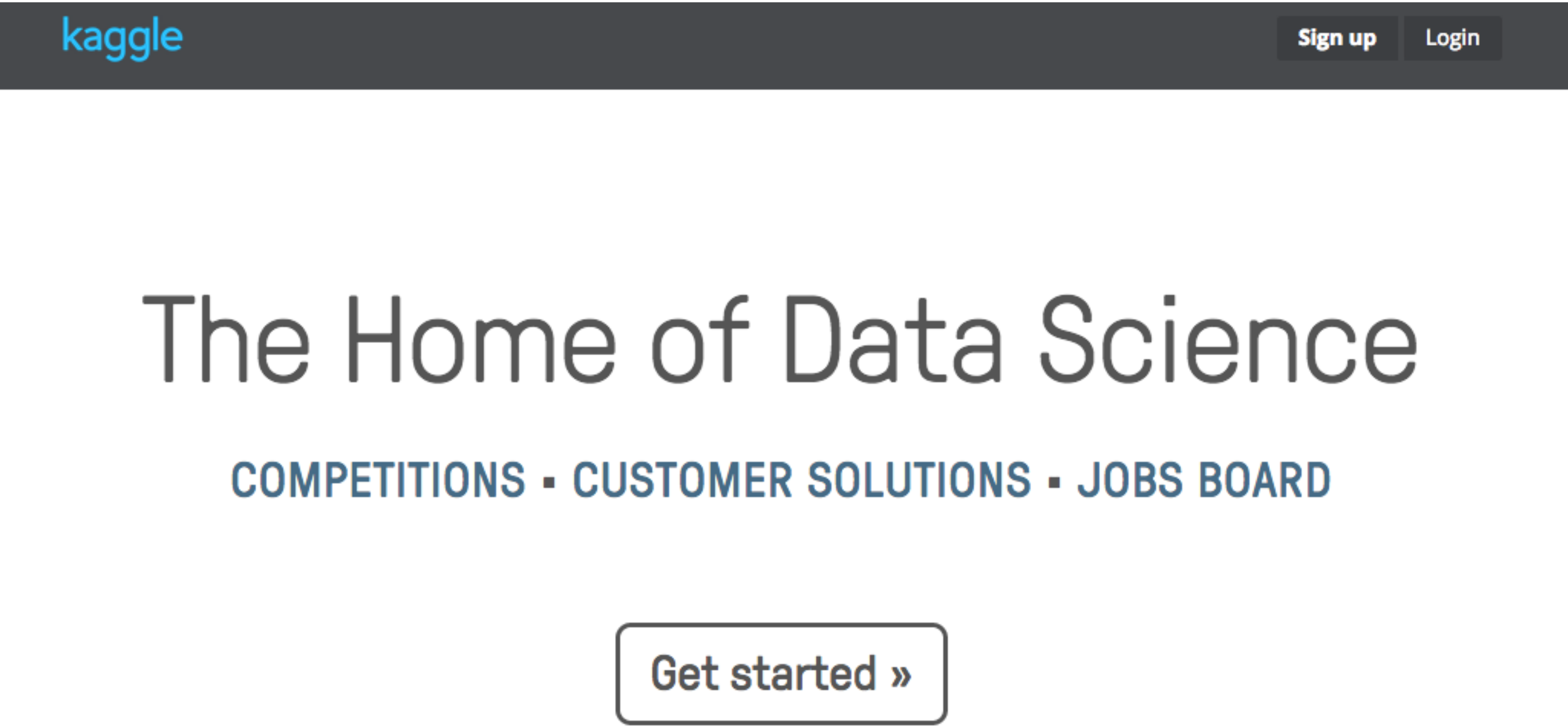
Learning by doing

What is learning?



Algorithm for any exam

- Prepare to exam (train phase)
- Prepare answers (predict phase)
- Check answers (evaluation phase)



The Home of Data Science

COMPETITIONS ▪ CUSTOMER SOLUTIONS ▪ JOBS BOARD

kaggle.com

Bike Sharing Demand

kaggle.com/c/bike-sharing-demand

Tools

IPython notebook

pandas
numpy

scikit-learn

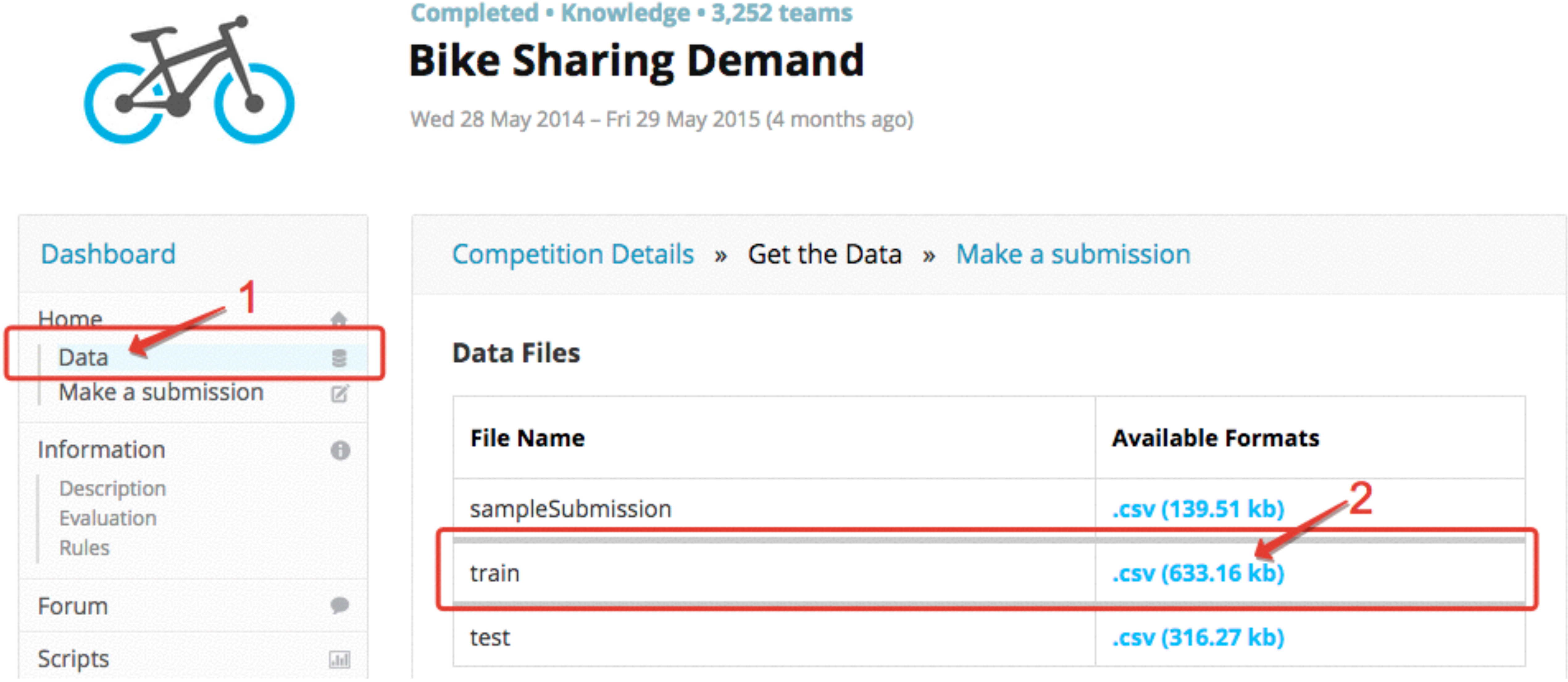
matplotlib
ggplot
seaborn

My Solution

bit.ly/1LlGD9U

Please download :)

Data



Completed • Knowledge • 3,252 teams

Bike Sharing Demand

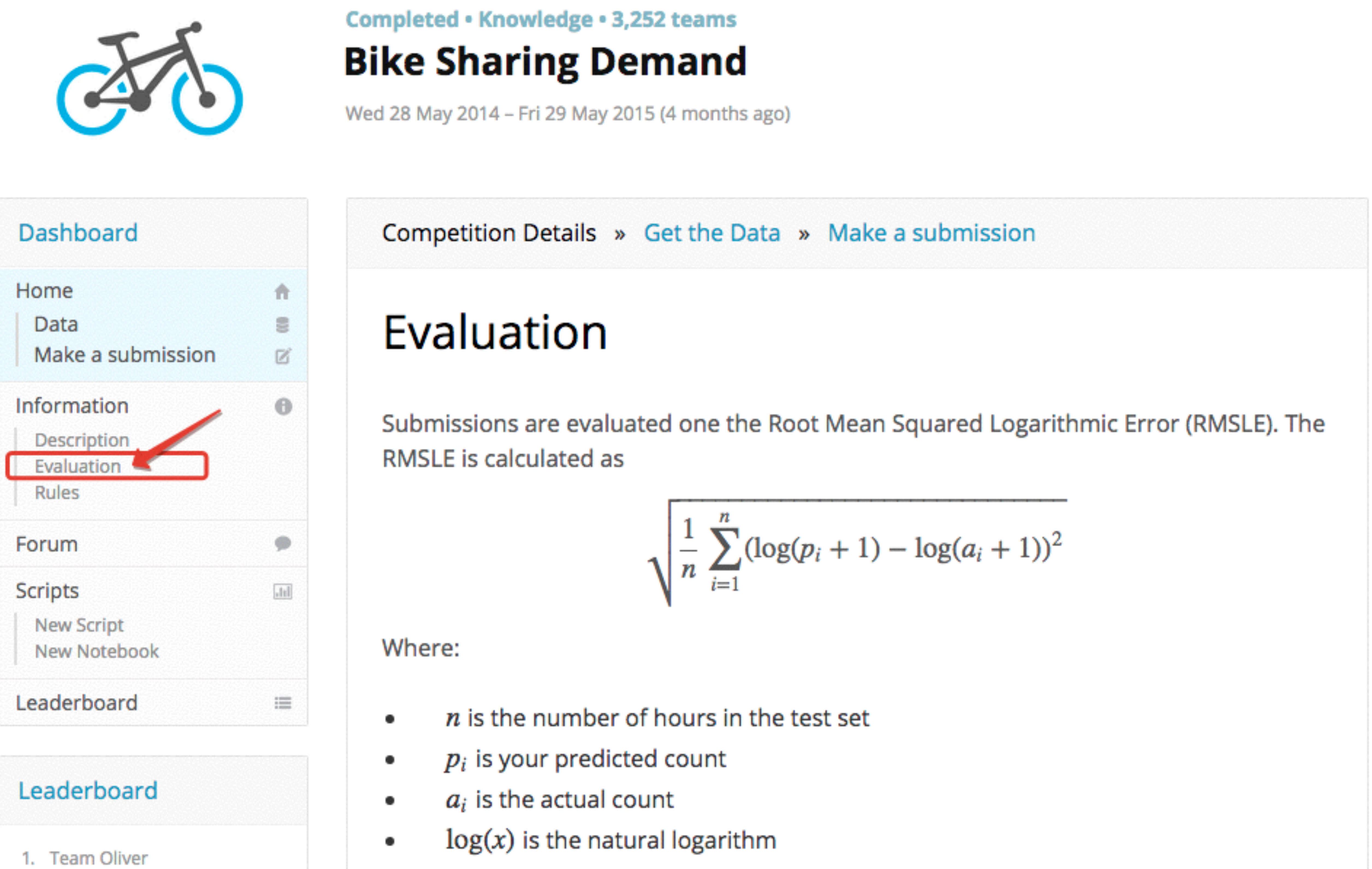
Wed 28 May 2014 – Fri 29 May 2015 (4 months ago)

Data Files

File Name	Available Formats
sampleSubmission	.csv (139.51 kb)
train	.csv (633.16 kb)
test	.csv (316.27 kb)

or use this temporary link: bit.ly/**1MTq4FM**

Evaluation



Completed • Knowledge • 3,252 teams

Bike Sharing Demand

Wed 28 May 2014 – Fri 29 May 2015 (4 months ago)

Evaluation

Submissions are evaluated one the Root Mean Squared Logarithmic Error (RMSLE). The RMSLE is calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Where:

- n is the number of hours in the test set
- p_i is your predicted count
- a_i is the actual count
- $\log(x)$ is the natural logarithm

Predict is it **bike** or not?

Machine learning on intuitive level

Data

Prepare data

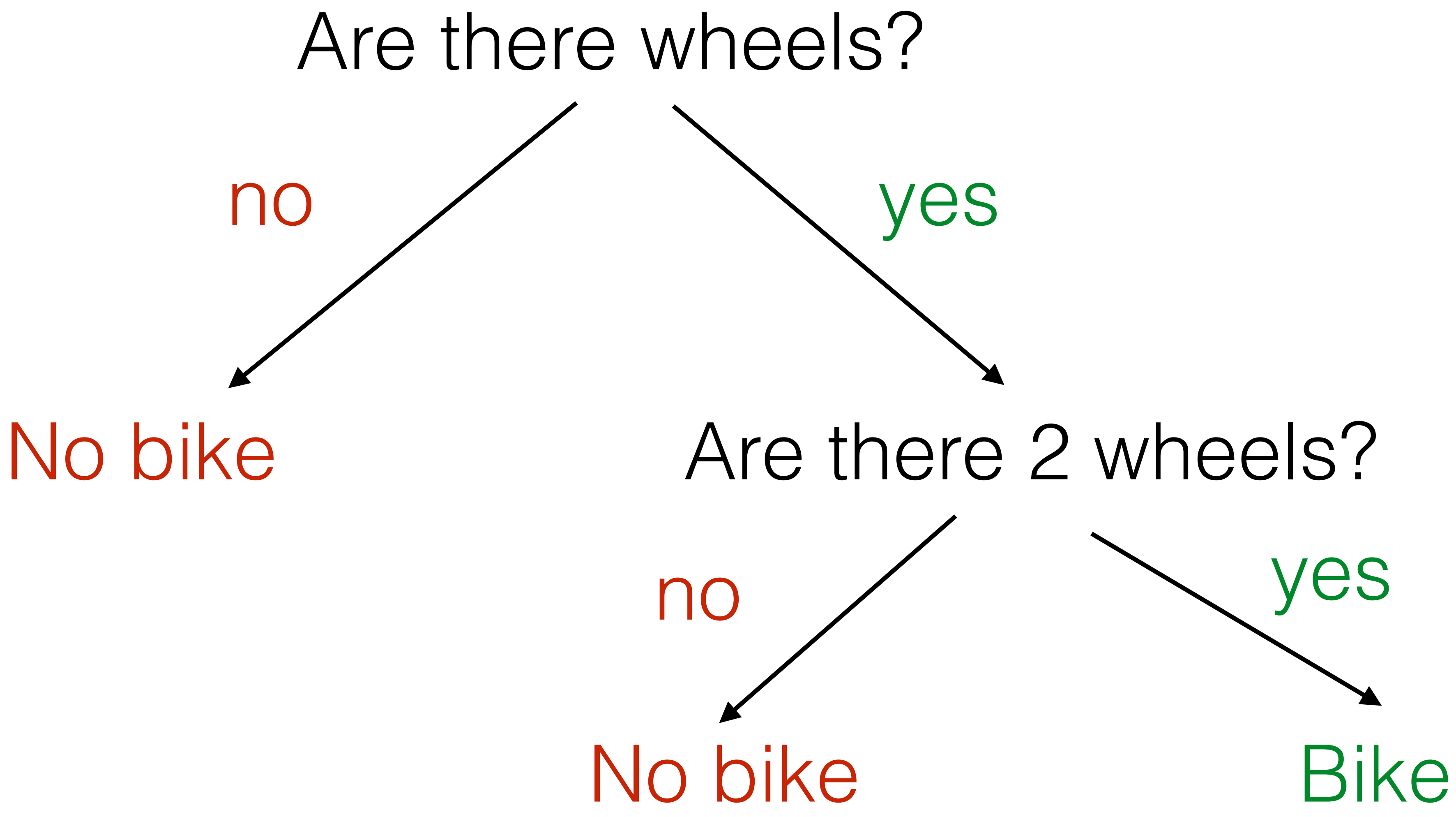
Feature engineering

Features

object	numbers of wheels	shape of wheels	...
	2	circle	
	4	circle	
	2	circle	
	0	-	
	2	circle	

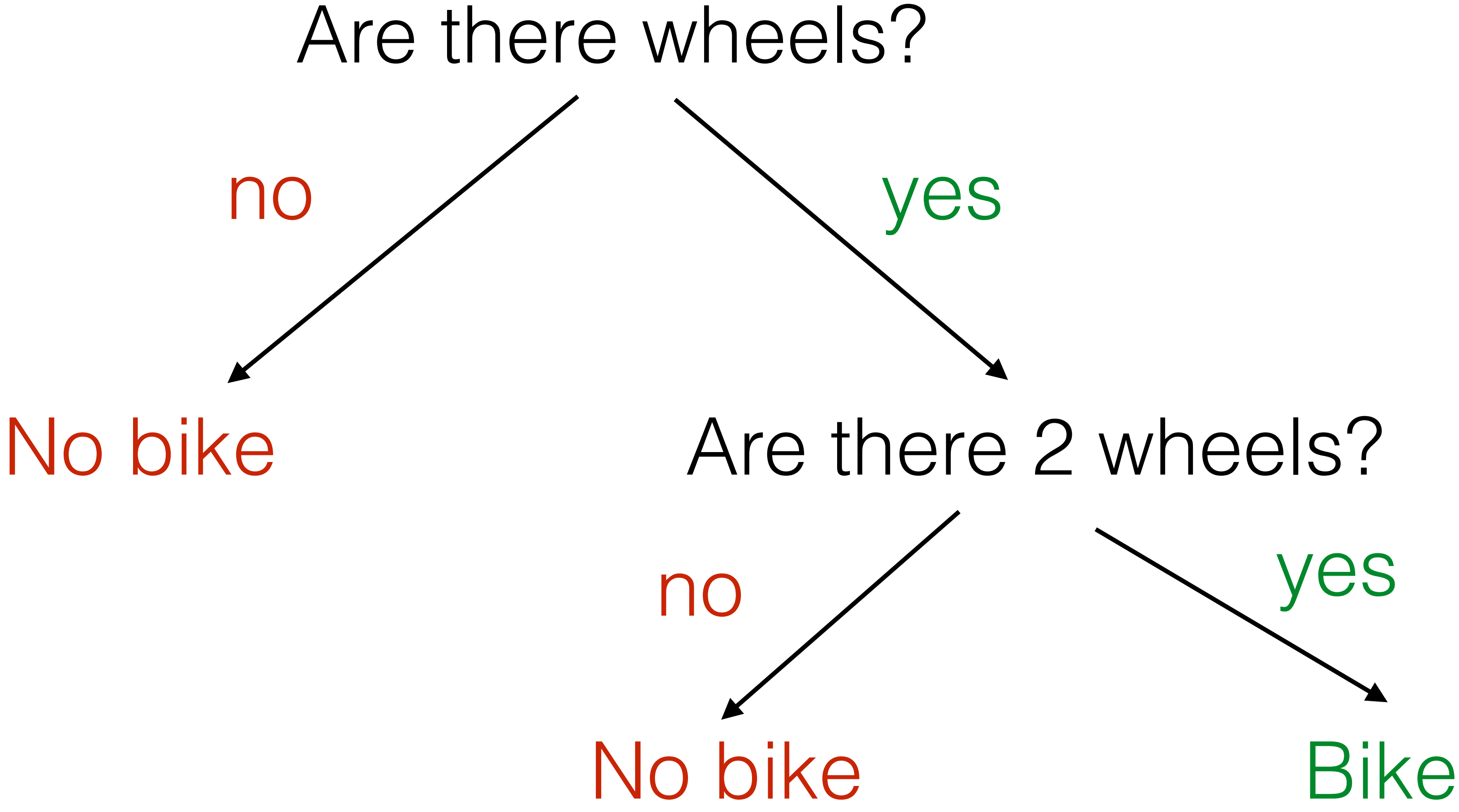
Build a model

Model


Are there wheels?
no
yes
No bike
Are there 2 wheels?
no
yes
No bike
Bike

Evaluate
(quality checking)

Model Evaluation



Are there wheels?

no

No bike

yes

Are there 2 wheels?

no

No bike

yes

Bike

Success

… or not?

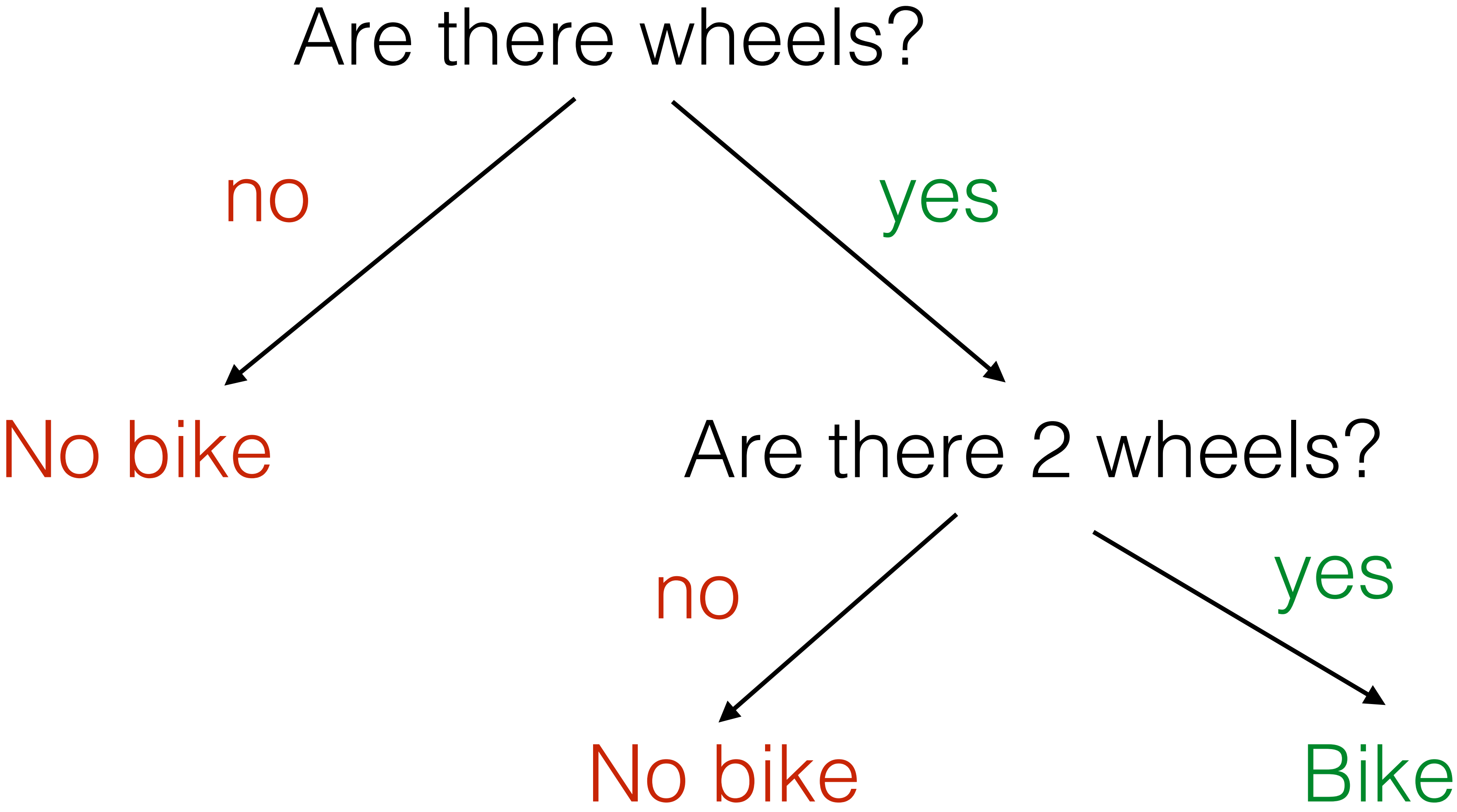
What about this?

Lil' Giant
GIANT
GIANT

SURLY
SURLY
SURLY
SURLY

FAZER

Model Evaluation



Are there wheels?

no

No bike

yes

Are there 2 wheels?

no

No bike

yes

Bike

Start looks good, but…

In summary

- Understand your success metrics (evaluation)
- Understand your data
- Do a lot of experiments

Thank you!