Donald Szeto
Kenneth Chan
Simon Chan
support@prediction.io

Big Data TechCon - Oct 17, 2013
Building Applications That Predict User Behavior Through Big Data
Using Open-Source Technologies

Machine Learning is....

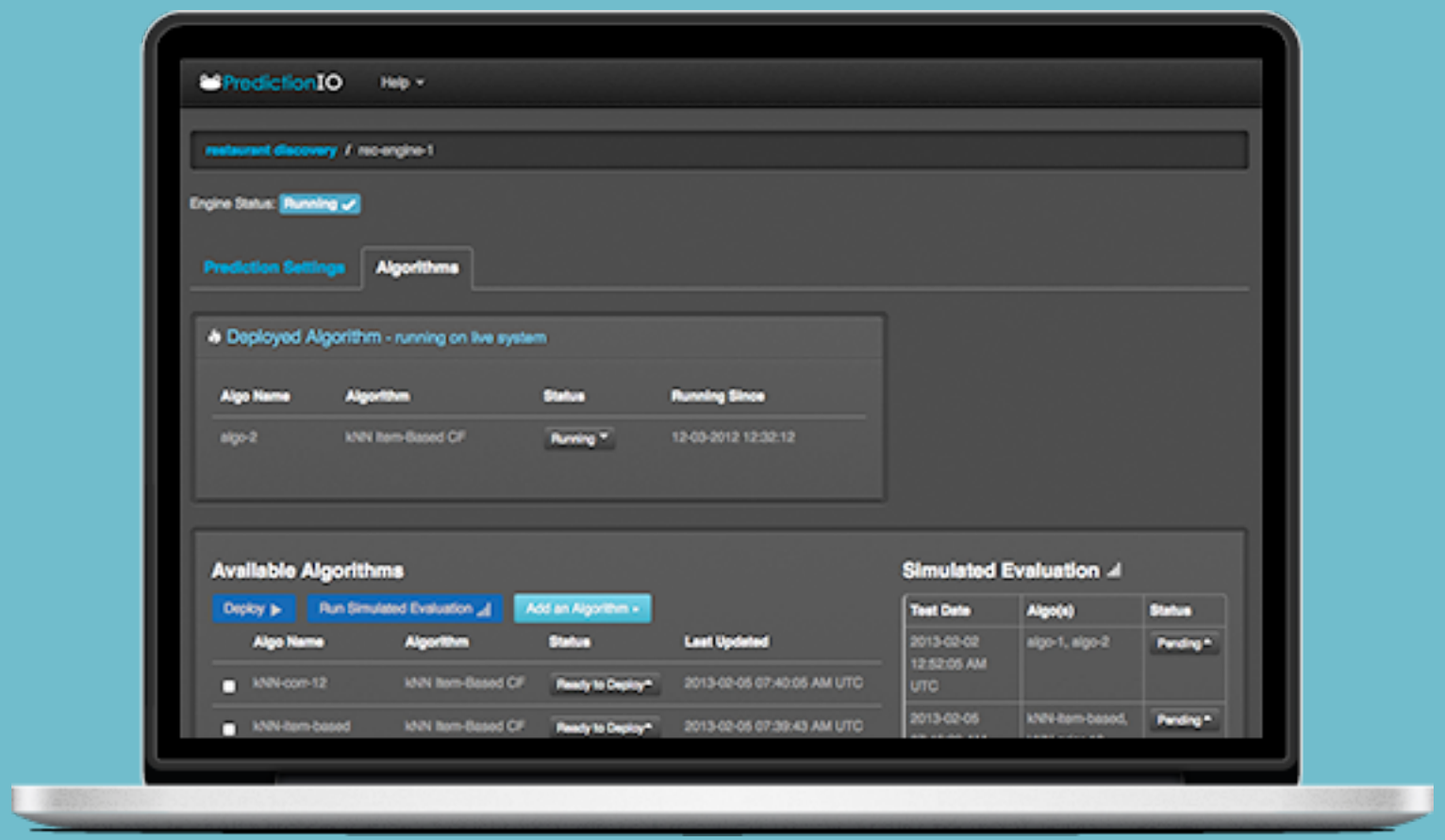
computers learning to predict from data

putting Machine Learning into practice

Existing data tools are like raw material

PredictionIO

3 Product Principles



Love Developers

- By Developers, For Developers
- REST APIs & SDKs
- Streamlined Data Science Process with UIs

Be Open

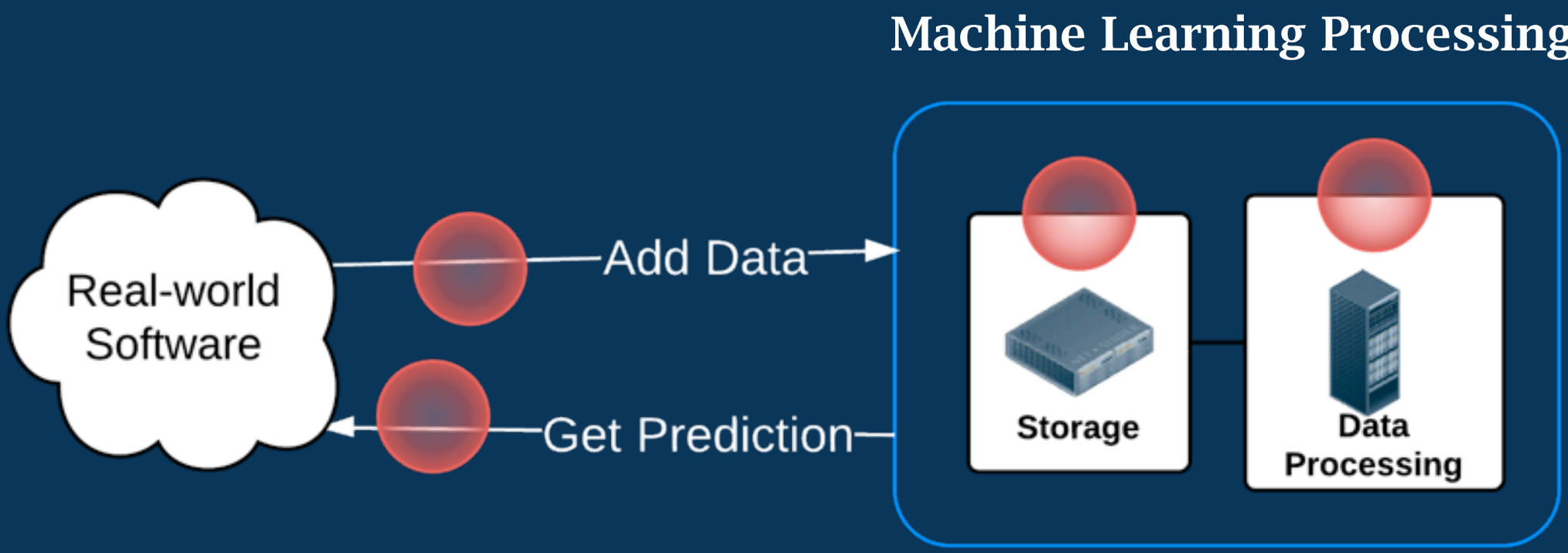
- Open Source
- Github - github.com/predictionio
- Supported by Mozilla WebFWD

For Production

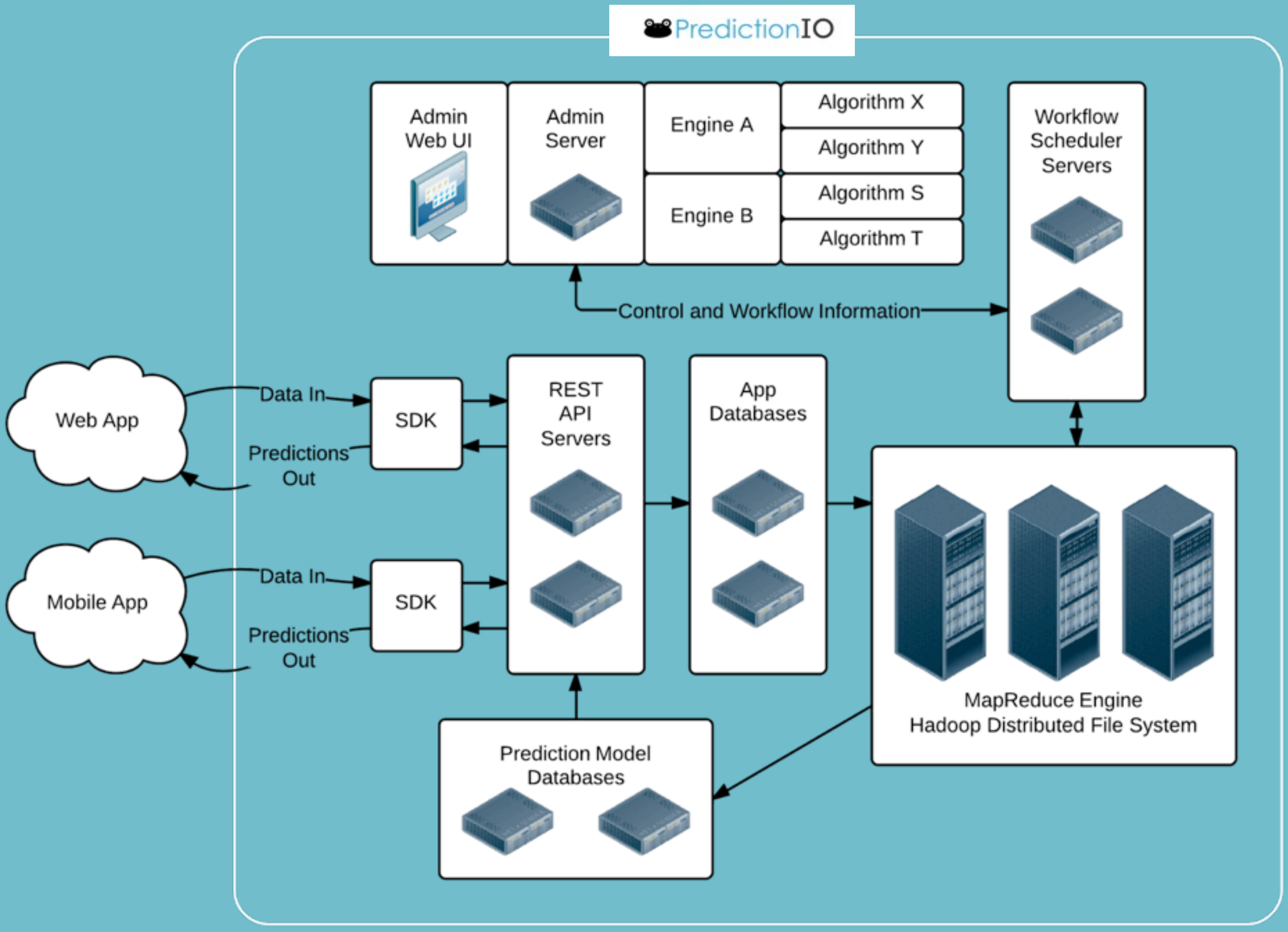
challenge #1

Scalability

Big Data Bottlenecks



PredictionIO has a
horizontally scalable
architecture


PredictionIO
Admin Web UI
Admin Server
Engine A
Algorithm X
Algorithm Y
Engine B
Algorithm S
Algorithm T
Workflow Scheduler Servers
Control and Workflow Information
Web App
Data In
Predictions Out
SDK
REST API Servers
App Databases
Mobile App
Data In
Predictions Out
SDK
MapReduce Engine
Hadoop Distributed File System
Prediction Model Databases

challenge #2

Productivity

Conventional Routine

- Collecting Data
- Picking an Algorithm
- Writing Scalable Code
- Evaluating Results
- Tuning Algorithm Parameters
- Iterate . . .

Algorithms

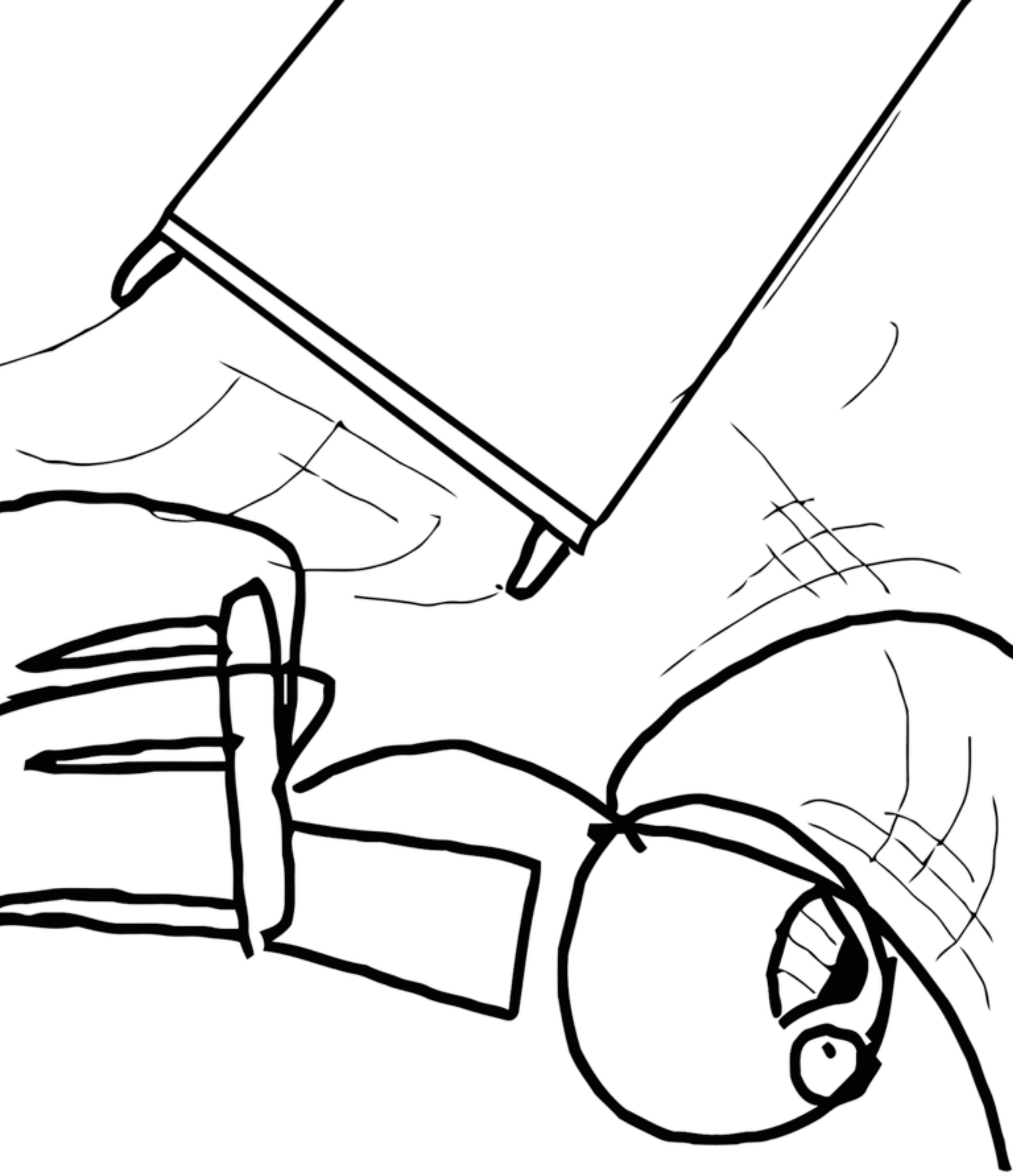
- Item Similarity
  - Collaborative Filtering (CF) ???
- Item Recommendation
  - kNN Item-based CF ???
  - Threshold Item-based CF ???
  - Alternating Least Squares ???
  - SlopeOne ???
  - Singular Value Decomposition ???

MapReduce - Native Java

```
public class WordCount {
 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(LongWritable key, Text value, Context context) throws .....{
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
 }
 public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
      throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) { sum += val.get(); }
        context.write(key, new IntWritable(sum));
    }
 }
 public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.waitForCompletion(true);
 }
}
```

What else?

- Evaluating Results
  - But how?
- Tuning Algorithm Parameters
  - Lambda
  - Regularization
  - # Iterations
  - . . .



Better Routine with PredictionIO

- Collecting Data
  - Using PredictionIO SDKs and API
- Picking an Algorithm ~~& Writing Scalable Code~~
  - Ready-to-use Algorithms
- Evaluating Results & Tuning Algorithm Parameters
  - Ready-to-use Metrics
  - Baseline Algorithms
- All of the Above Driven by a GUI (!)

PredictionIO Admin Web

PredictionIO Admin Web

cloudera1:5585/web/?appid=1&engineid=1&engineinfoid=itemrec#engineAddAlgorithm

Google

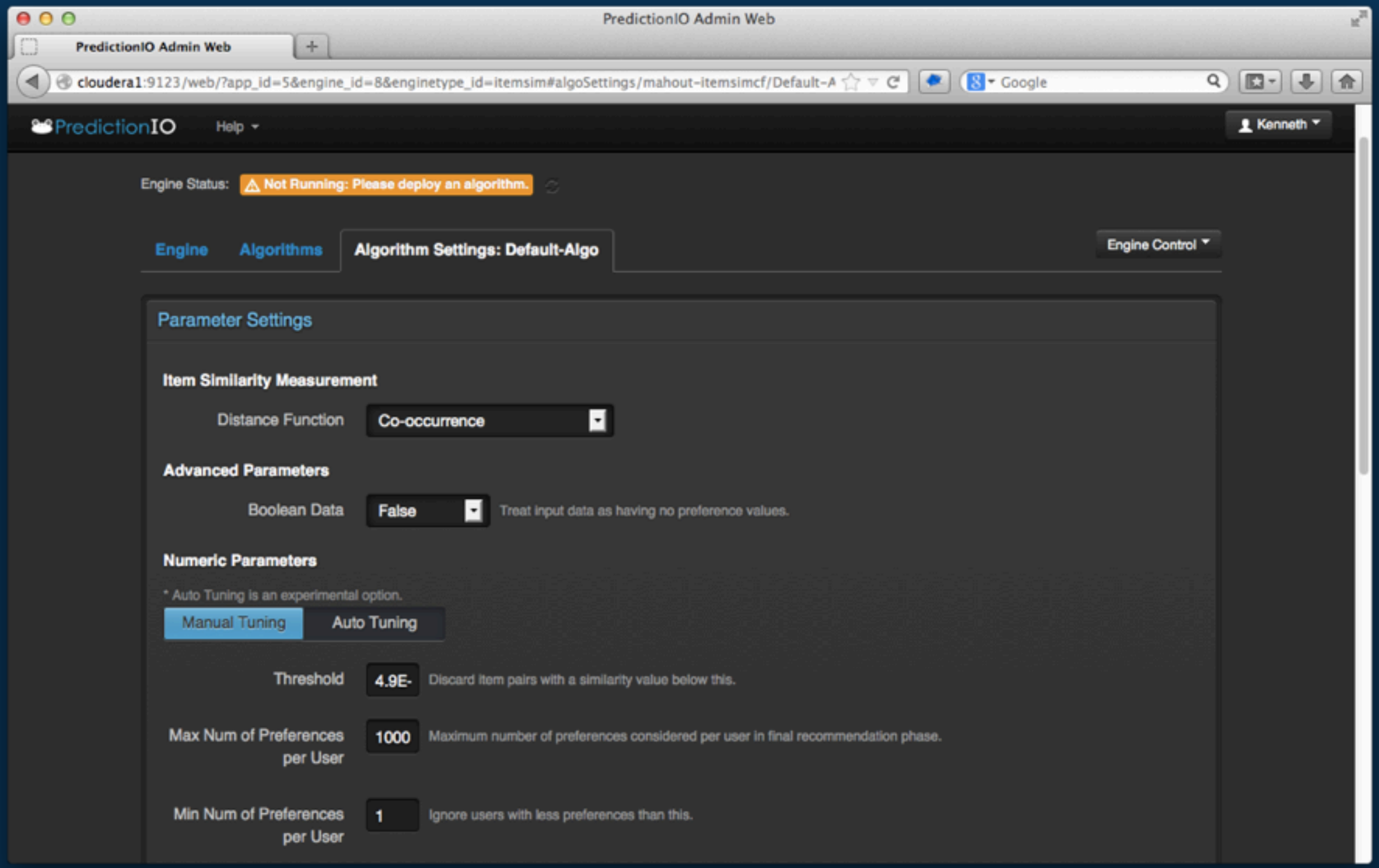
PredictionIO Help Donald Szeto

Engine Algorithms Add an Algorithm Engine Control

Available Algorithms for Item Recommendation Engine

Algorithm	Description	Requirement	Data Requirement	
Mahout's ALS-WR (Non-distributed)	Predict user preferences using matrix factorization (Non-distributed).	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's Threshold Item Based Collaborative Filtering	Predicts user preferences based on previous behaviors of users on similar items.	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's kNN User Based Collaborative Filtering (Non-distributed)	Predicts user preferences based on previous behaviors of users who are the k-nearest neighbors (Non-distributed).	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's Parallel ALS-WR	Predicts user preferences based on previous behaviors of users.	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's SlopeOne Rating Based Collaborative Filtering (Non-distributed)	Predicts user preferences based on average difference in preference values between new items and the items for which the user has indicated preferences (Non-distributed).	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's SVDPlusPlus Recommender (Non-distributed)	Predict user preferences using matrix factorization (Non-distributed).	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's SVD-RatingSGD Recommender (Non-distributed)	Predict user preferences using matrix factorization (Non-distributed).	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	type algo ı Add
Mahout's Threshold User Based Collaborative Filtering	Predicts user preferences based on previous behaviors of users whose similarity meets or exceeds a certain threshold	Hadoop	Users, Items, and U2I Actions such as Like, Buy	type algo ı Add

PredictionIO



PredictionIO

PredictionIO Admin Web

PredictionIO Admin Web

cloudera1:5585/web/?appid=1&engineid=1&engineinfoid=itemrec#simEvalSettings/62,150

PredictionIO Help Donald Szeto

Simulated Evaluation

Algorithms to be Evaluated

- mahout-itembased
- latest

Metrics

The above algorithms will be evaluated against the following metrics.

- Mean Average Precision (MAP@k) k = 20 [remove]

Add Another Metric

Data Split

Adjust the slider to specify how you want to split data into Train Set and Test Set.

Train Set → Test Set →

Train: 60 % Test: 20 %

Data Selection: Random Sampling

Random with Time Order means that data in Test Set is always newer than those in Train Set.

Iteration

PredictionIO

Efficiently tune algorithms
as a developer.

Develop intuition with
how algorithms behave.

(Very) Light Intro to kNN CF

Users and items

Predict user rating on unseen item

Items \ Users	1	2	3	4	5	6	7	8	9	10
1	3			5		4	1		4	1
2		2			2		1			2
3	3		1	4		?	2		5	
4		3		1		1		3	1	
5	5	2		5		2	2			1

k-Nearest Neighbor (kNN)

Find k most similar items to the targeted items

- **Pearson Correlation**
- **Cosine / Adjusted Cosine**
- **Jaccard coefficient**

Weight-sum the known ratings of the targeted users on these items

Users

Items	1	2	3	4	5	6	7	8	9	10
1	3			5		4	1		4	1
2		2			2		1			2
3	3		1	4		?	2		5	
4		3		1		1		3	1	
5	5	2		5		2	2			1

k-Nearest Neighbor (kNN)

Correlation as Item Similarity:

Item 1 & 3:

R_1 = [3, 5, 1, 4]
R_3 = [3, 4, 2, 5]

Corr(R_1, R_3) = 0.8

Users

Items	1	2	3	4	5	6	7	8	9	10
1	3			5		4	1		4	1
2		2			2		1			2
3	3		1	4			2		5	
4		3		1		1		3	1	
5	5	2		5		2	2			1

k-Nearest Neighbor (kNN)

S_{ij}: similarity score of item i and item j

Let S_{13} = 0.8, S_{35} = 0.9, predicted score of User 6 on Item 3 would be:

(4 * 0.8 + 2 * 0.9) /(0.8 + 0.9)
= 5 / 1.7
= 2.9

Users

Items	1	2	3	4	5	6	7	8	9	10
1	3			5		4	1		4	1
2		2			2		1			2
3	3		1	4		?	2		5	
4		3		1		1		3	1	
5	5	2		5		2	2			1

PredictionIO In Action

AngelList

https angel.co/finder

Reader

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

AngelList

Search

Join Log In

HOME | STARTUPS | SYNDICATES | JOBS | ACTIVITY | MORE

Finder

Clear All

All Companies 71,362

Hiring 2,640

Only show companies

- ☐ Publicly raising
- ☐ With traction
- ☑ Claimed
- ☐ Featured

Market

Add Market Add

Location

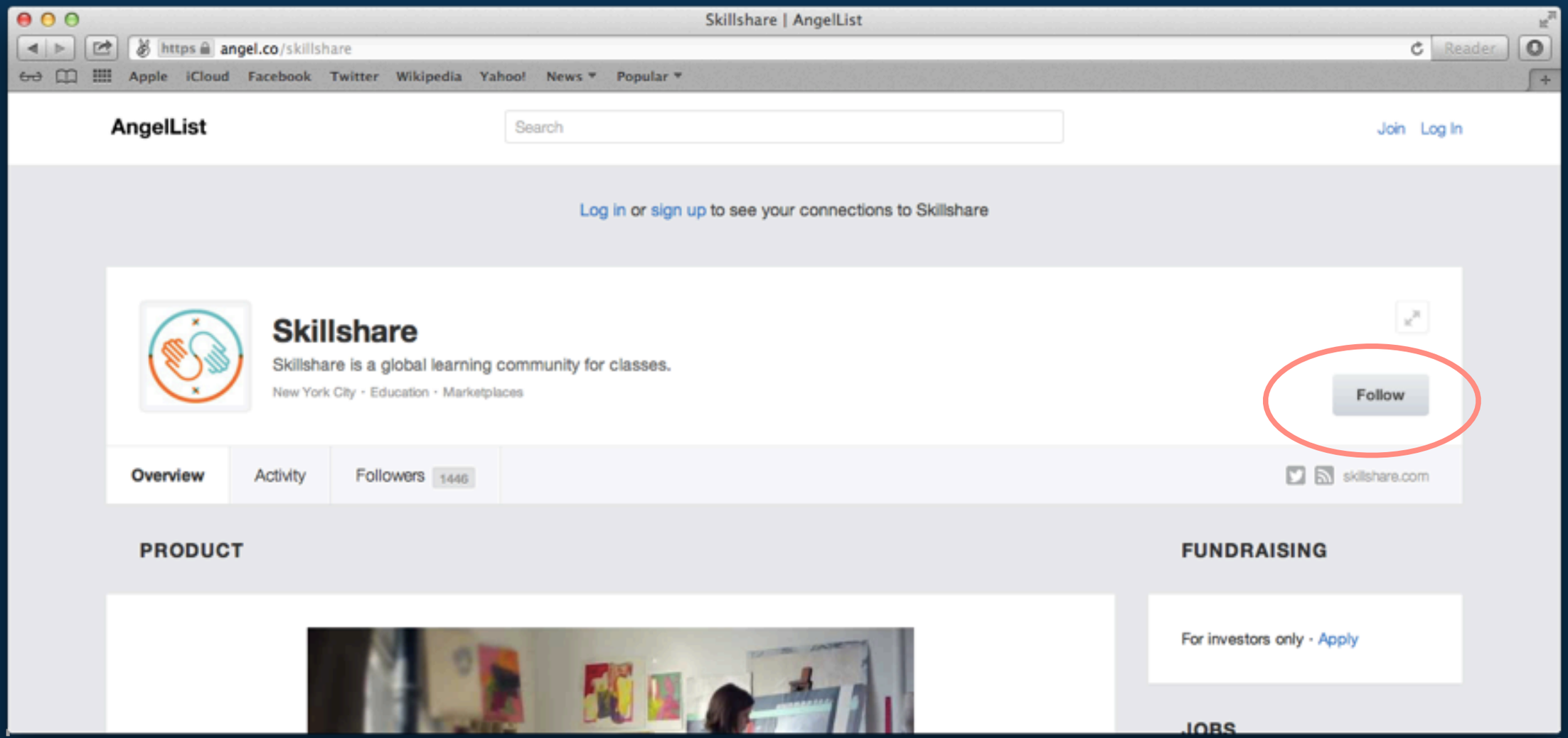
Add Location Add

Team

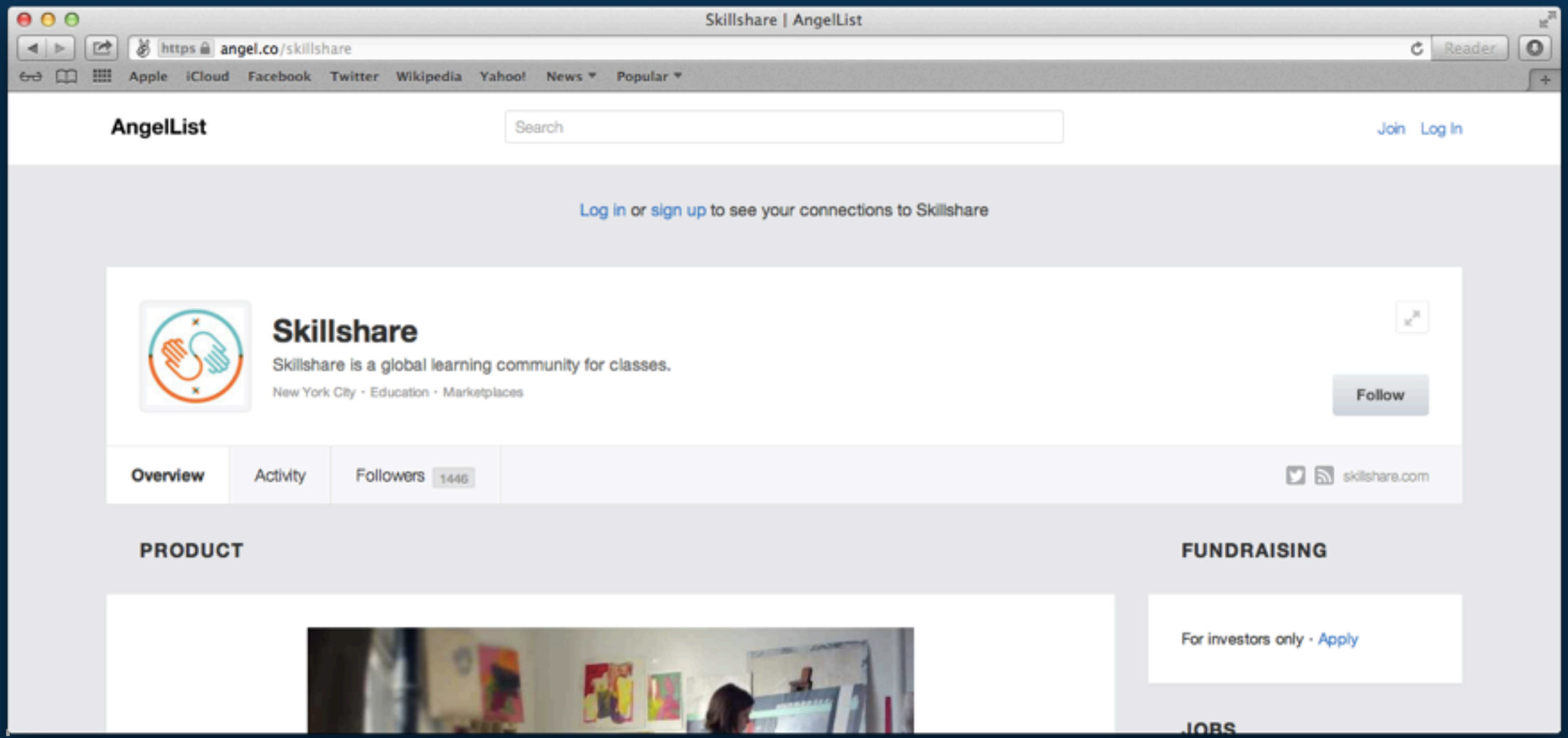
Add people Add

Company	Joined	Followers	Signal	Path
71,362 companies				
StarStreet Real money fantasy sports games Boston · Games	Jul '10	236		
BackType Acquired by Twitter San Francisco	Aug '10	143		
Pinterest A Universal Social Catalog San Francisco · Social Media	Aug '10	2545		
Ostrovok Travel booking platform for Russia Russia · E-Commerce	Sep '10	362		

PredictionIO



Something is missing here...



"If you follow this startup, you may also want to follow these X, Y, Z startups."

How are we going to do this?

Data +  PredictionIO

What data do we need?

Users - the followers (AngelList users)

Items - the startups

Actions - the “follow” actions

What data do we need?

For this demo...

- Use AngelList API (https://angel.co/api)

- Retrieve all startups (Items) belonging to the following accelerators:

- Retrieve all users following these startups (Users and Actions).

- 16382 users. 993 items. 204176 actions.

What data do we need?

Generate these files...

users.csv - list of unique user ids

<user_id>

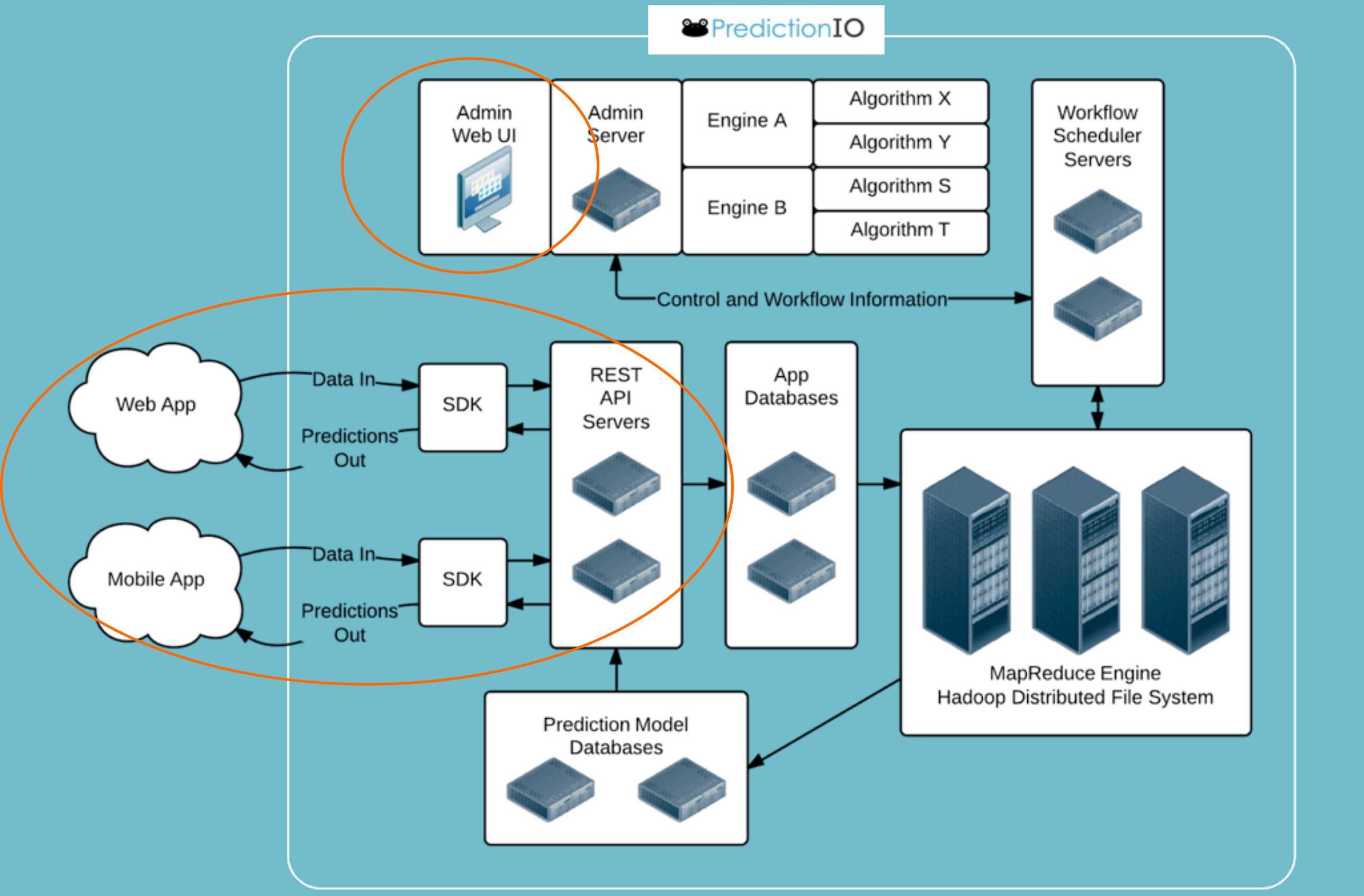
startup_id_name_url_incubator.csv - list of startups with ids, names, AngelList URL and incubators

<startup_id>, <name>, <url>, <incubator>

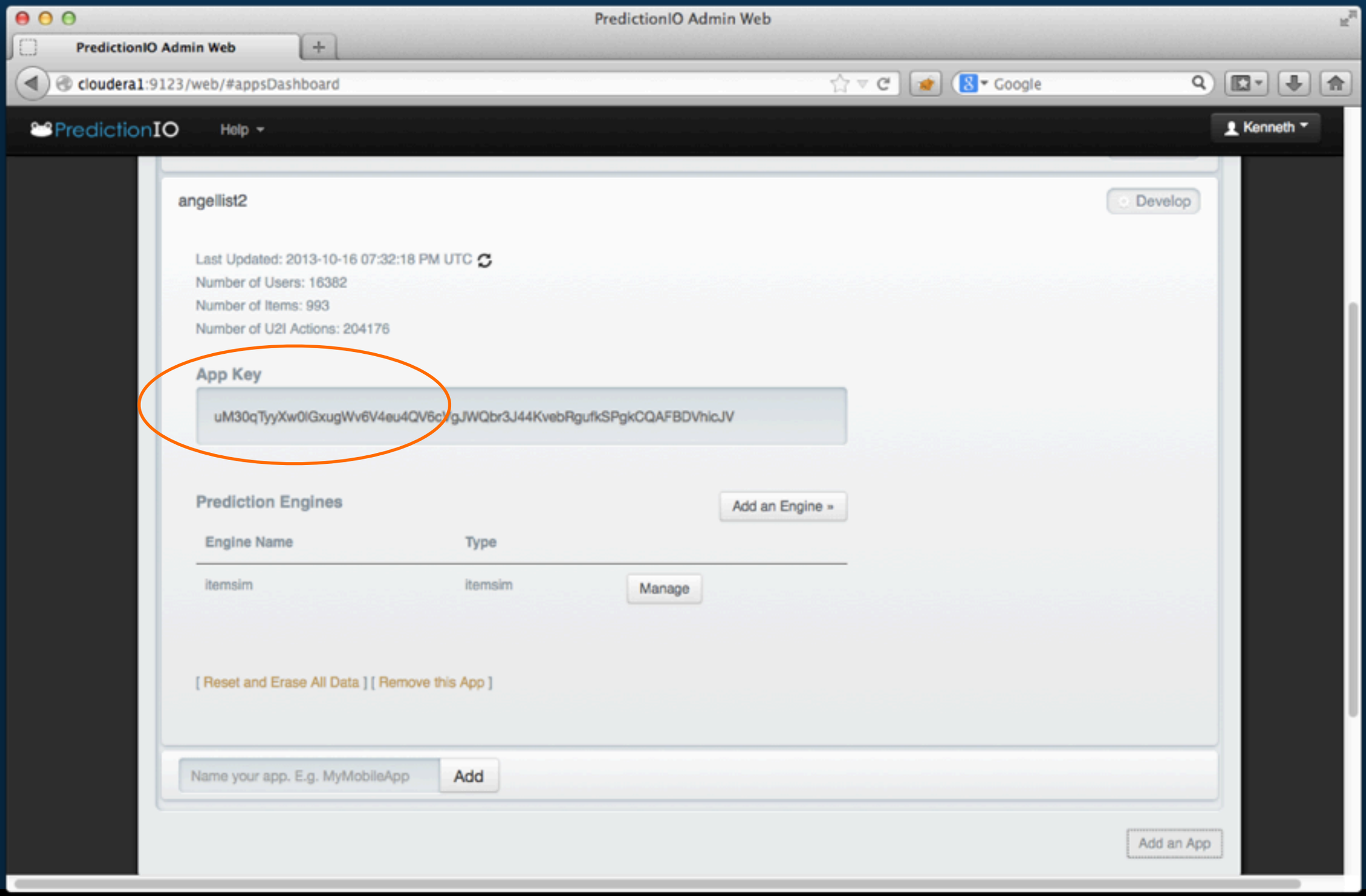
follows.csv - list of ids for startups and users who follow the startup

<startup_id>, <user_id>

How to integrate with PredictionIO?


PredictionIO
Admin Web UI
Admin Server
Engine A
Algorithm X
Algorithm Y
Engine B
Algorithm S
Algorithm T
Workflow Scheduler Servers
Control and Workflow Information
Web App
Data In
Predictions Out
SDK
REST API Servers
App Databases
Mobile App
Data In
Predictions Out
SDK
MapReduce Engine
Hadoop Distributed File System
Prediction Model Databases

How to integrate with PredictionIO?



How to integrate with PredictionIO?

Python SDK is used in this demo (other SDKs are similar)....

```
# PredictionIO Client Object

client = predictionio.Client(“APP_KEY”)


# Import users and items

client.create_user(“user_id”)

client.create_item(“startup_id”, (“item_type_1”,))


# Example

client.create_user(“5”)

client.create_item(“17”, (“startup”, “y-combinator”,))
```

How to integrate with PredictionIO?

Python SDK is used in this demo (other SDKs are similar)....

```
# Import actions

client.identify(“user_id”)

client.record_action_on_item(“action_name”, “startup_id”)
```

PredictionIO comes with the following built-in actions:

“like”, “dislike”, “rate”, “view”, “conversion”

```
# Example

client.identify(“5”)

client.record_action_on_item(“like”, “17”) # use like to represent follow
```

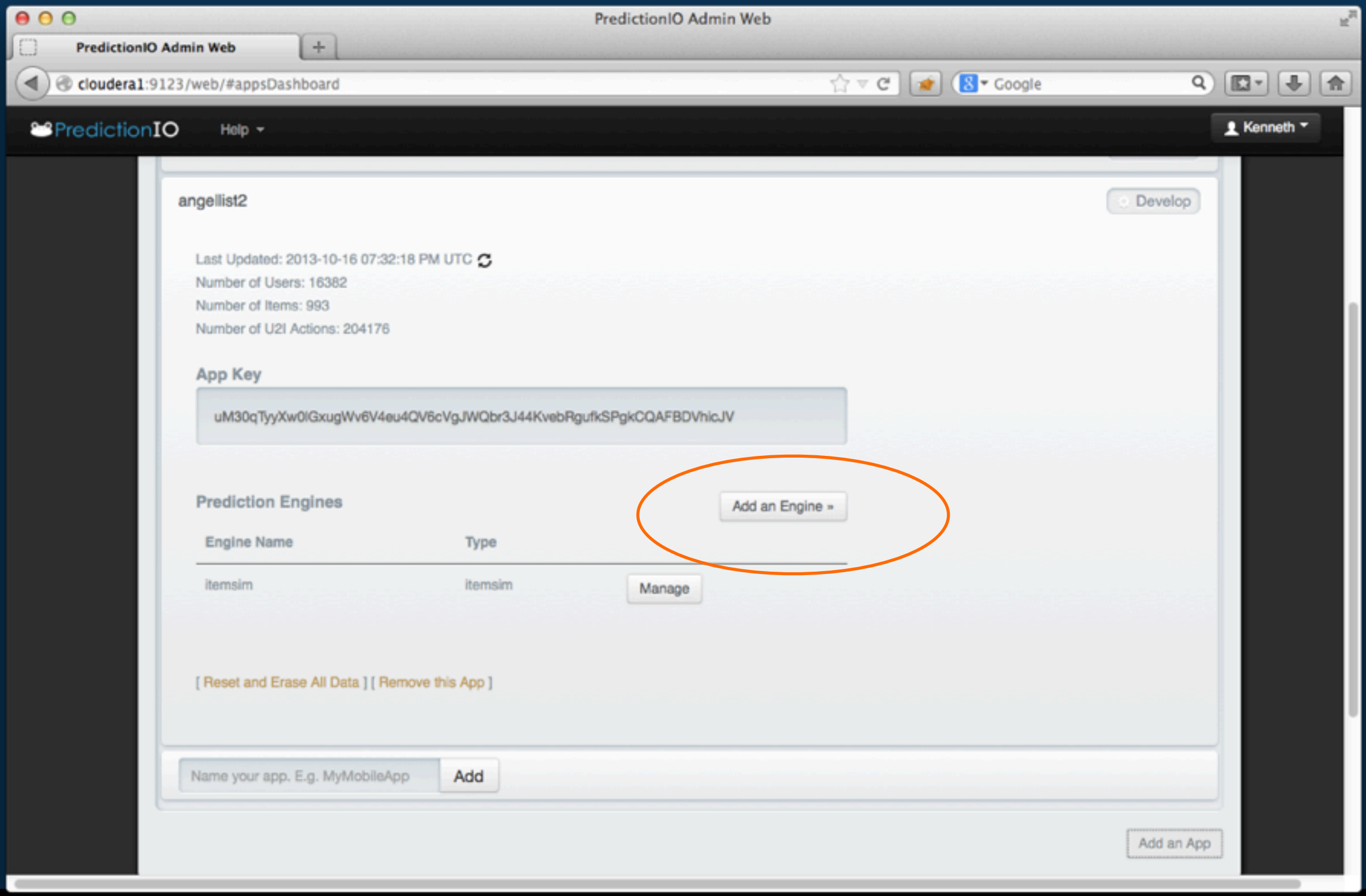
How to integrate with PredictionIO?

Python SDK is used in this demo (other SDKs are similar)....

```
# More advanced usage

client = predictionio.Client(APP_KEY, THREADS, API_URL, qsize=REQUEST_QSIZE)
```

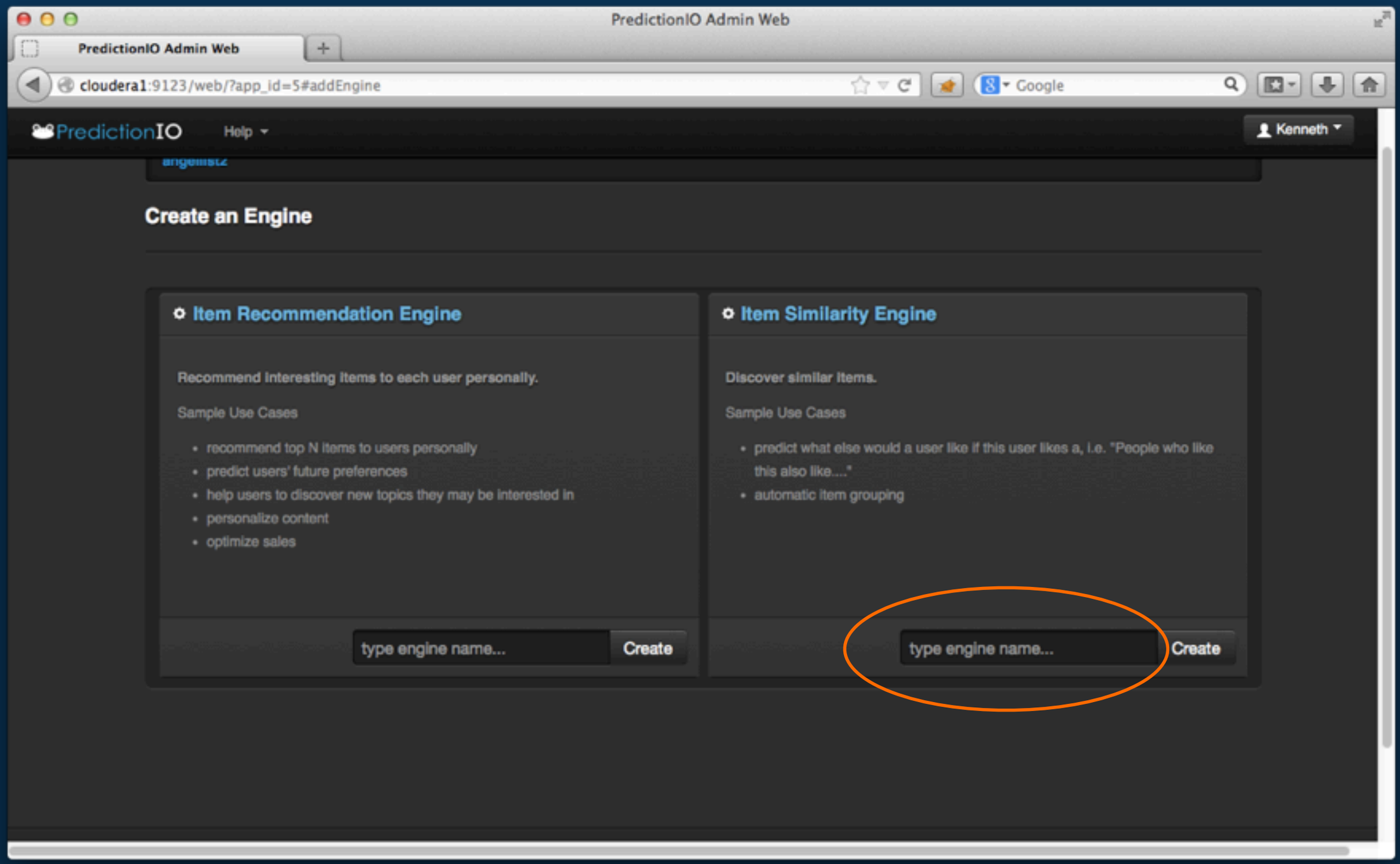
```
# Asynchronous version

client.acreate_user(“user_id”)

client.acreate_item(“item_id”, (“item_type_1”,))

client.arecord_action_on_item(“action_name”, “item_id”)
```

How to integrate with PredictionIO?

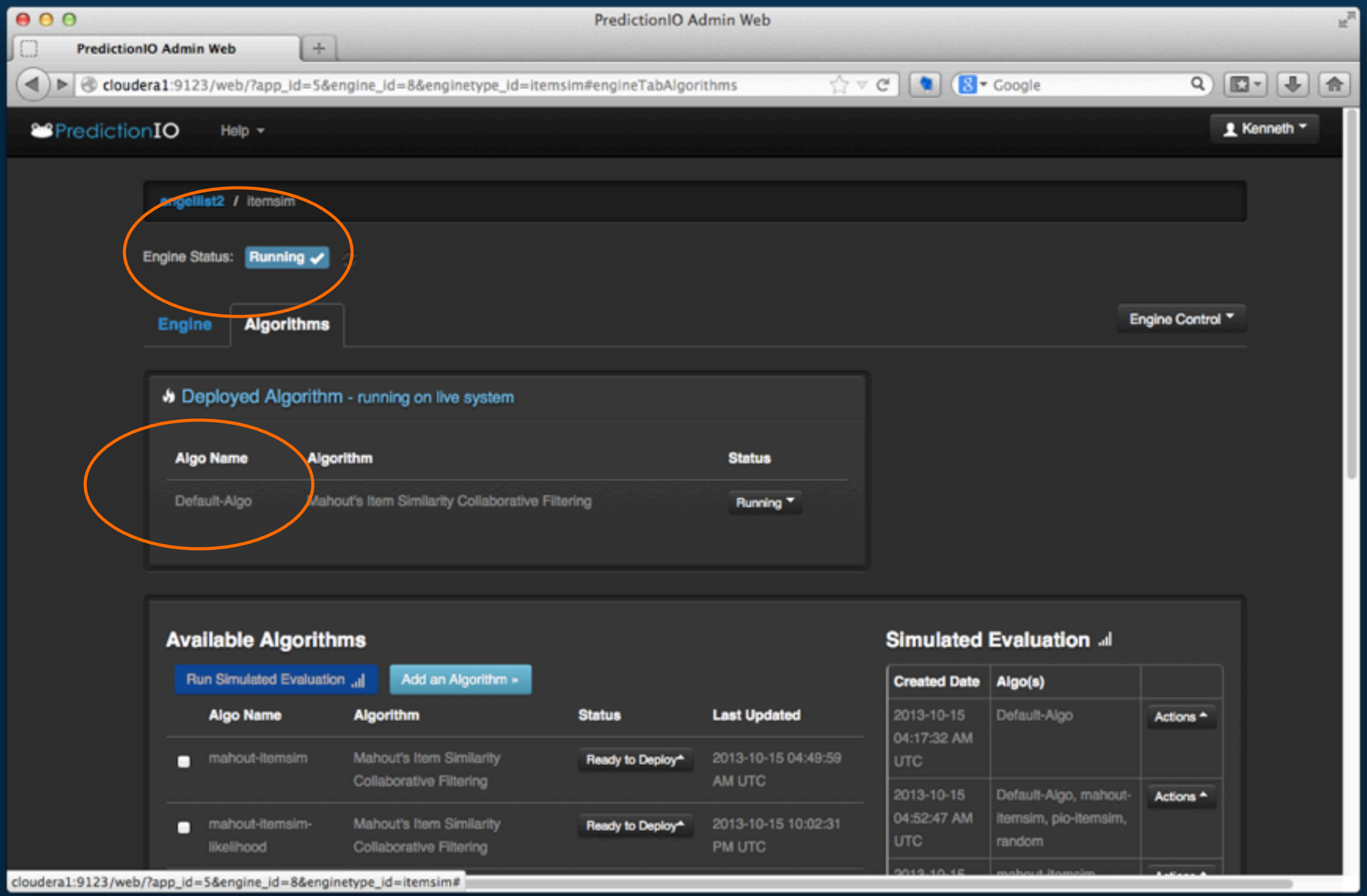


PredictionIO

How to integrate with PredictionIO?



How to integrate with PredictionIO?

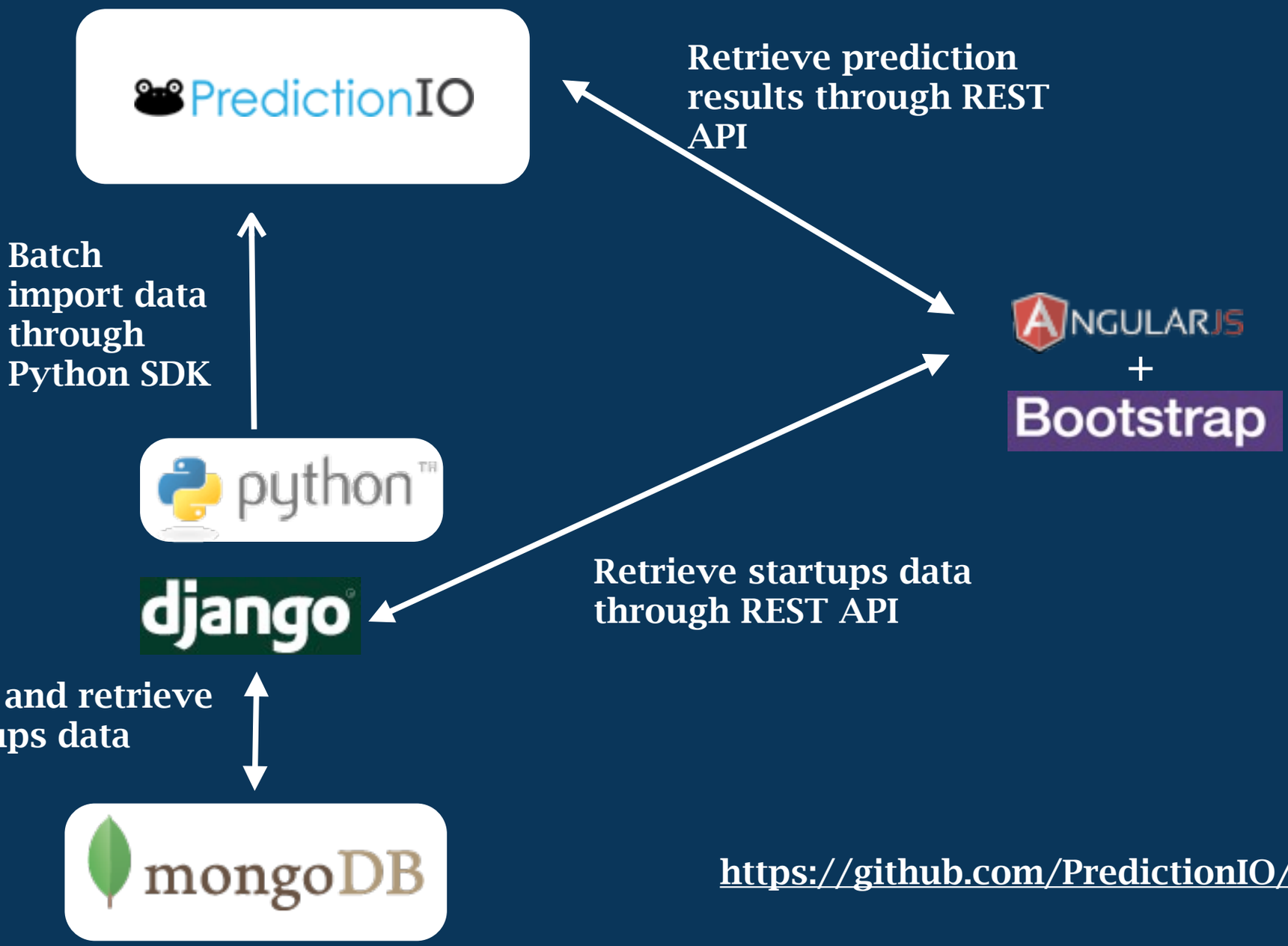


How to retrieve prediction results?

Python SDK is used in this demo (other SDKs are similar)....

```
# retrieve prediction results from engine

rec = client.get_itemsim_topn(“engine_name”, “startup_id”, num)


# Example

rec = client.get_itemsim_topn(“my-engine”, “17”, 10)


# Asynchronous version

req = client.aget_itemsim_topn(“engine_name”, “startup_id”, num)

# can do something else in between.....

rec = client.aresp(req)
```

Demo Setup



https://github.com/PredictionIO/Demo-AngelList

PredictionIO Sample Project with Public AngelList Data

A showcase on how to use PredictionIO to generate predictions and display them.

Visit http://prediction.io for more information on PredictionIO.

If you follow...

Keen IO

Filter by incubators:

All | 500 startups | The Alchemist Accelerator | AngelPad | Dreamit Ventures | StartX | techstars | Y Combinator

Keen IO (AngelList)

First Previous 1 Next Last

You may also follow...

Visually (AngelList)
Crittercism (AngelList)
Mixpanel (AngelList)
Ribbon (AngelList)
Vungle (AngelList)
Statwing (AngelList)
WillCall (AngelList)
Hipset (AngelList)
Balanced (AngelList)
Yobble (AngelList)

PredictionIO Sample: Public AngelList Data

angellist-demo.prediction.io

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

PredictionIO Sample Project with Public AngelList Data

A showcase on how to use PredictionIO to generate predictions and display them.

Visit http://prediction.io for more information on PredictionIO.

If you follow...

Keen IO

Filter by incubators:

All | 500 startups | The Alchemist Accelerator | AngelPad | Dreamit Ventures | StartX | techstars | Y Combinator

Keen IO (AngelList) Analytics Backend as a Service (web + mobile + internet of things)

First Previous 1 Next Last

You may also follow...

Visually (AngelList)
Crittercism (AngelList)
Mixpanel (AngelList) Analytics platform for mobile & web.
Ribbon (AngelList)
Vungle (AngelList) E-commerce across social, mobile, and web
Statwing (AngelList) Data analysis tool
WillCall (AngelList)
Hipset (AngelList)
Balanced (AngelList)
Yobble (AngelList)

PredictionIO

PredictionIO Sample: Public AngelList Data

angellist-demo.prediction.io

Reader

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

PredictionIO Sample Project with Public AngelList Data

A showcase on how to use PredictionIO to generate predictions and display them.

Visit http://prediction.io for more information on PredictionIO.

If you follow...

Keen IO

Filter by incubators:

All | 500 startups | The Alchemist Accelerator | AngelPad | Dreamit Ventures | StartX | techstars | Y Combinator

Keen IO (AngelList) Analytics Backend as a Service (web + mobile + internet of things)

First Previous 1 Next Last

You may also follow...

Visually (AngelList) Production and distribution of visual content
Crittercism (AngelList) Mobile Application Performance Monitoring
Mixpanel (AngelList)
Ribbon (AngelList) Mobile monetization
Vungle (AngelList)
Statwing (AngelList)
WillCall (AngelList)
Hipset (AngelList)
Balanced (AngelList) Payments for marketplaces and crowdfunding
Yobble (AngelList) Xbox Kinect for iPhone

PredictionIO

PredictionIO Sample: Public AngelList Data

angellist-demo.prediction.io

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

PredictionIO Sample Project with Public AngelList Data

A showcase on how to use PredictionIO to generate predictions and display them.

Visit http://prediction.io for more information on PredictionIO.

If you follow...

Keen IO

Filter by incubators:

All | 500 startups | The Alchemist Accelerator | AngelPad | Dreamit Ventures | StartX | techstars | Y Combinator

Keen IO (AngelList) Analytics Backend as a Service (web + mobile + internet of things)

First Previous 1 Next Last

You may also follow...

Visually (AngelList)
Crittercism (AngelList)
Mixpanel (AngelList)
Ribbon (AngelList)
Vungle (AngelList)
Statwing (AngelList)
WillCall (AngelList) Live shows
Hipset (AngelList) Youtube network for Artists
Balanced (AngelList)
Yobble (AngelList)

PredictionIO

PredictionIO Sample: Public AngelList Data

angellist-demo.prediction.io

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

PredictionIO Sample Project with Public AngelList Data

A showcase on how to use PredictionIO to generate predictions and display them.

Visit http://prediction.io for more information on PredictionIO.

If you follow...

Keen IO

Filter by incubators:

All | 500 startups | The Alchemist Accelerator | AngelPad | Dreamit Ventures | StartX | techstars | Y Combinator

Keen IO (AngelList) Analytics Backend as a Service (web + mobile + internet of things)

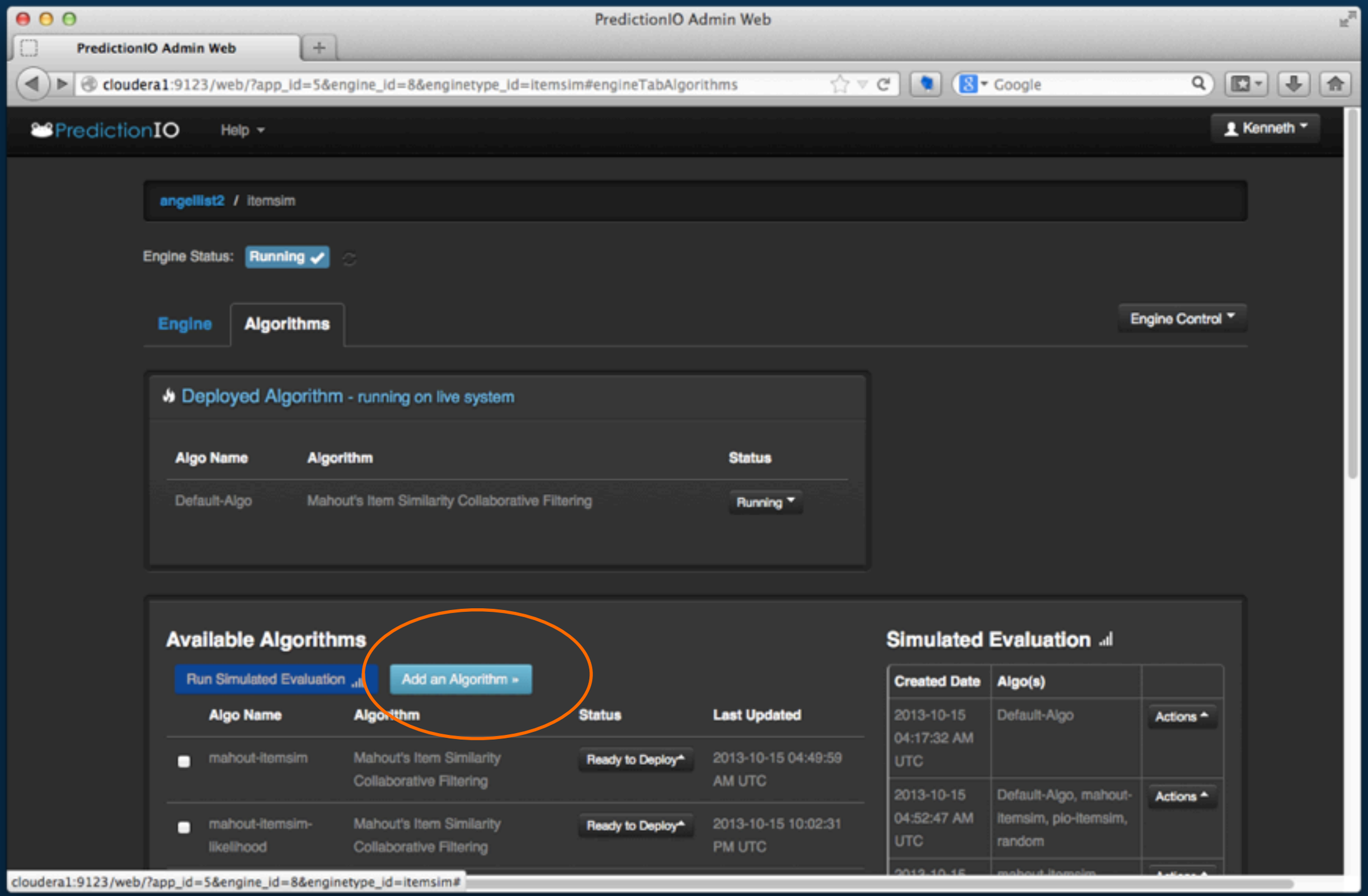
First Previous 1 Next Last

You may also follow...

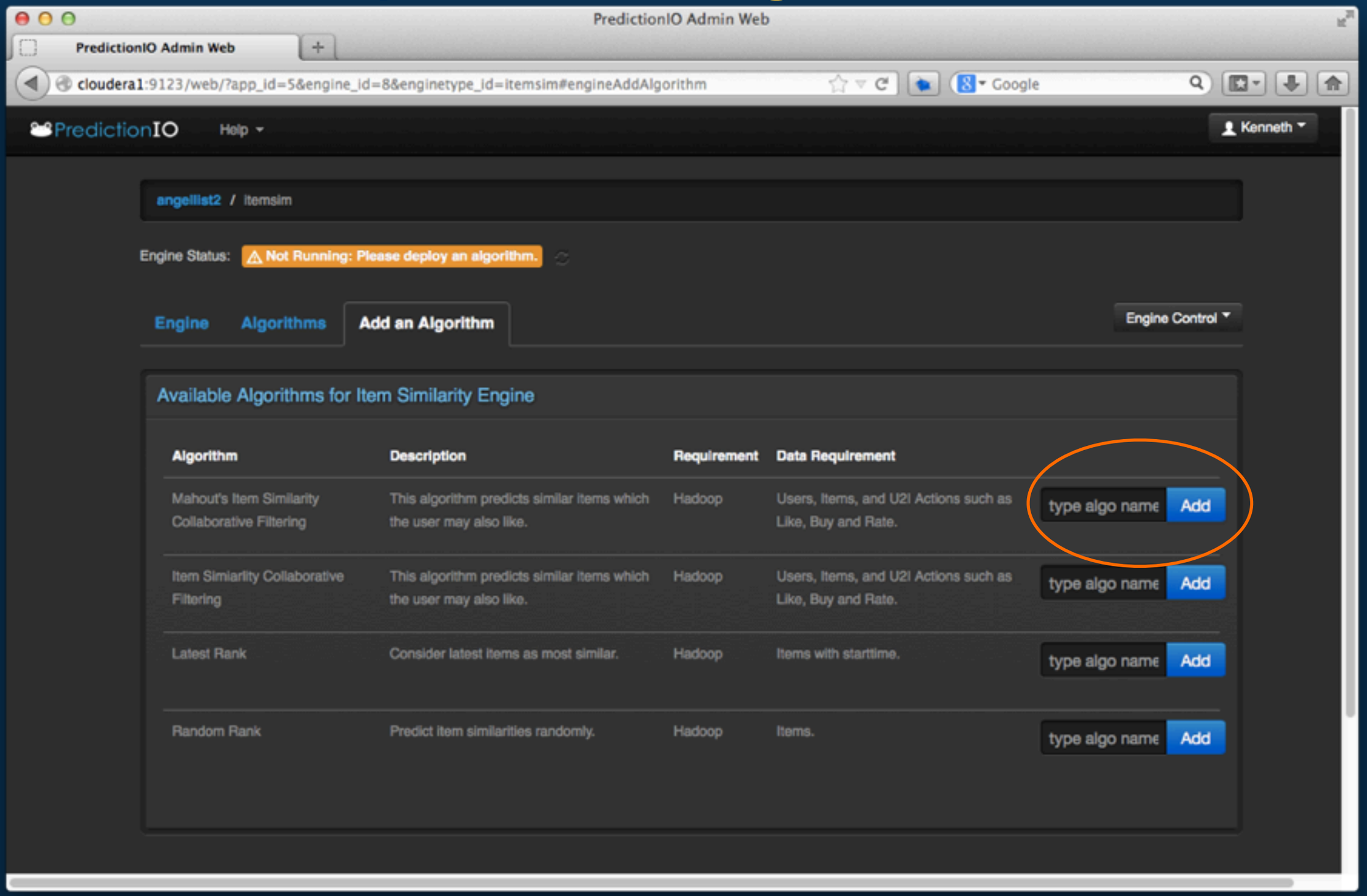
Visually (AngelList) Production and distribution of visual content
Crittercism (AngelList) Mobile Application Performance Monitoring
Mixpanel (AngelList) Analytics platform for mobile & web.
Ribbon (AngelList) Mobile monetization
Vungle (AngelList) E-commerce across social, mobile, and web
Statwing (AngelList) Data analysis tool
WillCall (AngelList) Live shows
Hipset (AngelList) Youtube network for Artists
Balanced (AngelList) Payments for marketplaces and crowdfunding
Yobble (AngelList) Xbox Kinect for iPhone

PredictionIO

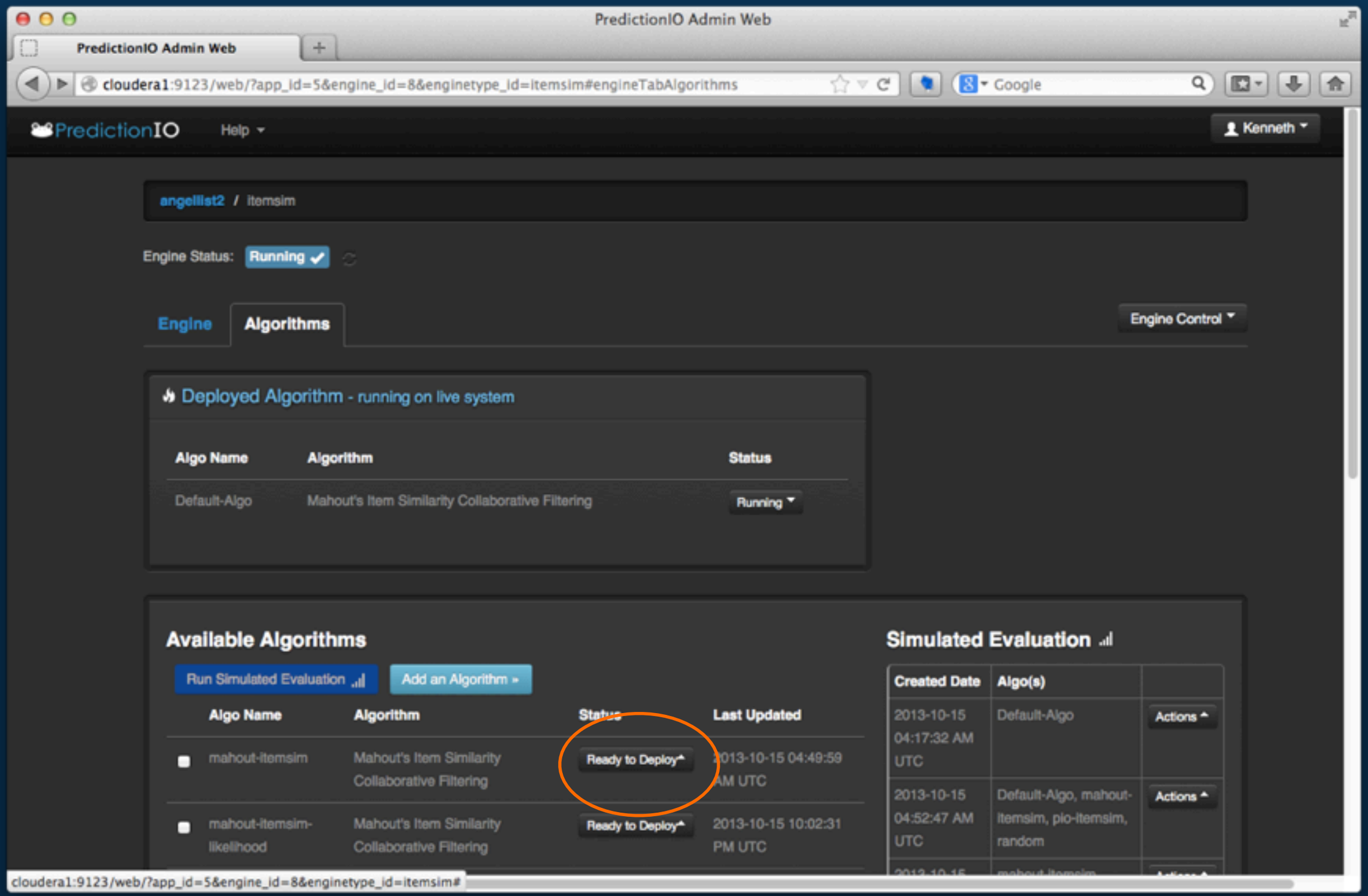
How to try different algorithms?



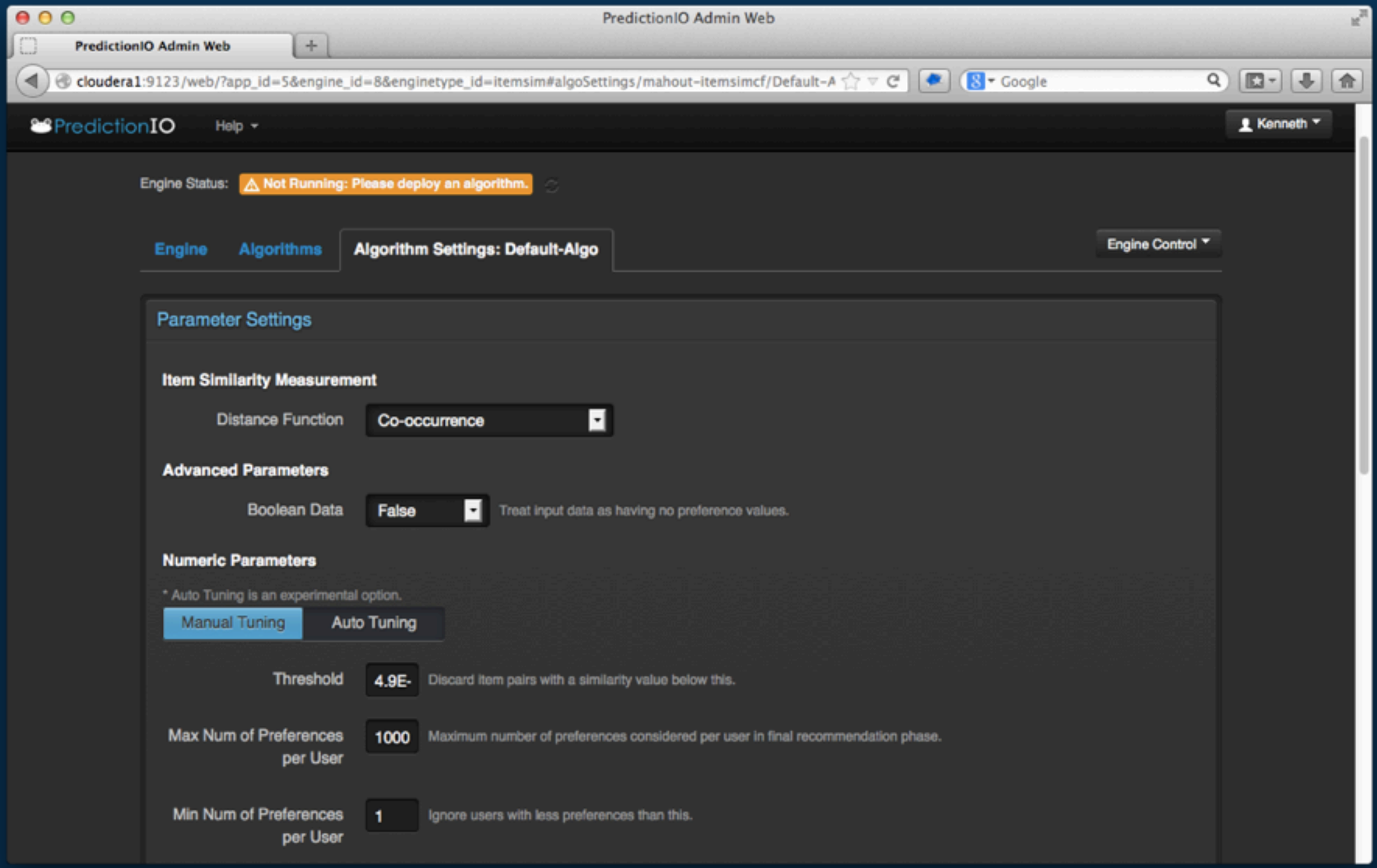
How to try different algorithms?



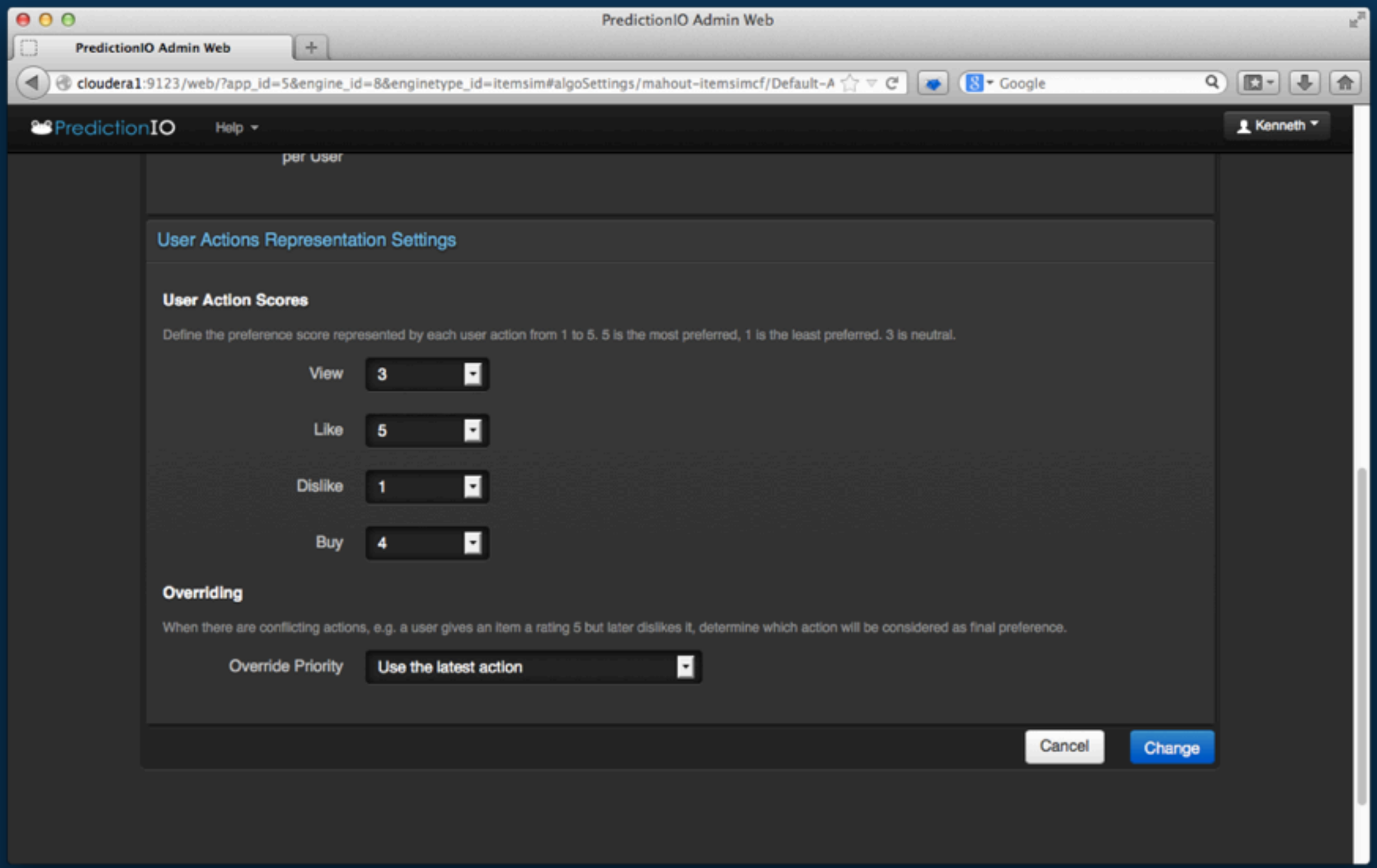
How to try different algorithms?



How to change algorithm parameters?



How to change algorithm parameters?



How do we evaluate how good the prediction results are?

We need metrics!

MAP@k

MAP@k (Mean Average Precision at k)

For this user A...

Retrieve Top 5 items (prediction):	Relevant items (actual behavior):
item4	item7
item10	item10
item6	item25
item15	
item7	

Total relevant items: 3

MAP@k (Mean Average Precision at k)

For this user A...

Retrieve Top 5 items (prediction):	Relevant items (actual behavior):
item4	item7
item10	item10
item6	item25
item15	
item7	

How good is this prediction?

Total relevant items: 3

MAP@k (Mean Average Precision at k)

For this user A...

Retrieve Top 5 items (prediction):	Position:	Number of relevant items up to this position:	P@k:
item4	1	0	0
item10	2	1	1/2
item6	3	1	0
item15	4	1	0
item7	5	2	2/5

Total relevant items: 3

AP@5 = (1/2 + 2/5) / min(5,3)

MAP@5 = average of AP@5 of all users

ISMAP@k

- Extension to MAP@k for Item Similarity Engine

MAP@k

For each user....

Retrieve Top 5 items (prediction):	Relevant items (actual behavior):
item4	item7
item10	item10
item6	item25
item15	
item7	

AP@5 = (1/2 + 2/5) / min(5,3)

MAP@5 = average of AP@5 of all users

ISMAP@k

For each item X...

For each user who "follows" this item X....

Retrieve Top 5 similar items to X that the user may also follow (prediction):

item4

item10

item6

item15

item7

Relevant items (actual behavior):

item7

item10

item25

AP@5 = (1/2 + 2/5) / min(5,3)

MAP@5 = average of AP@5 of all users

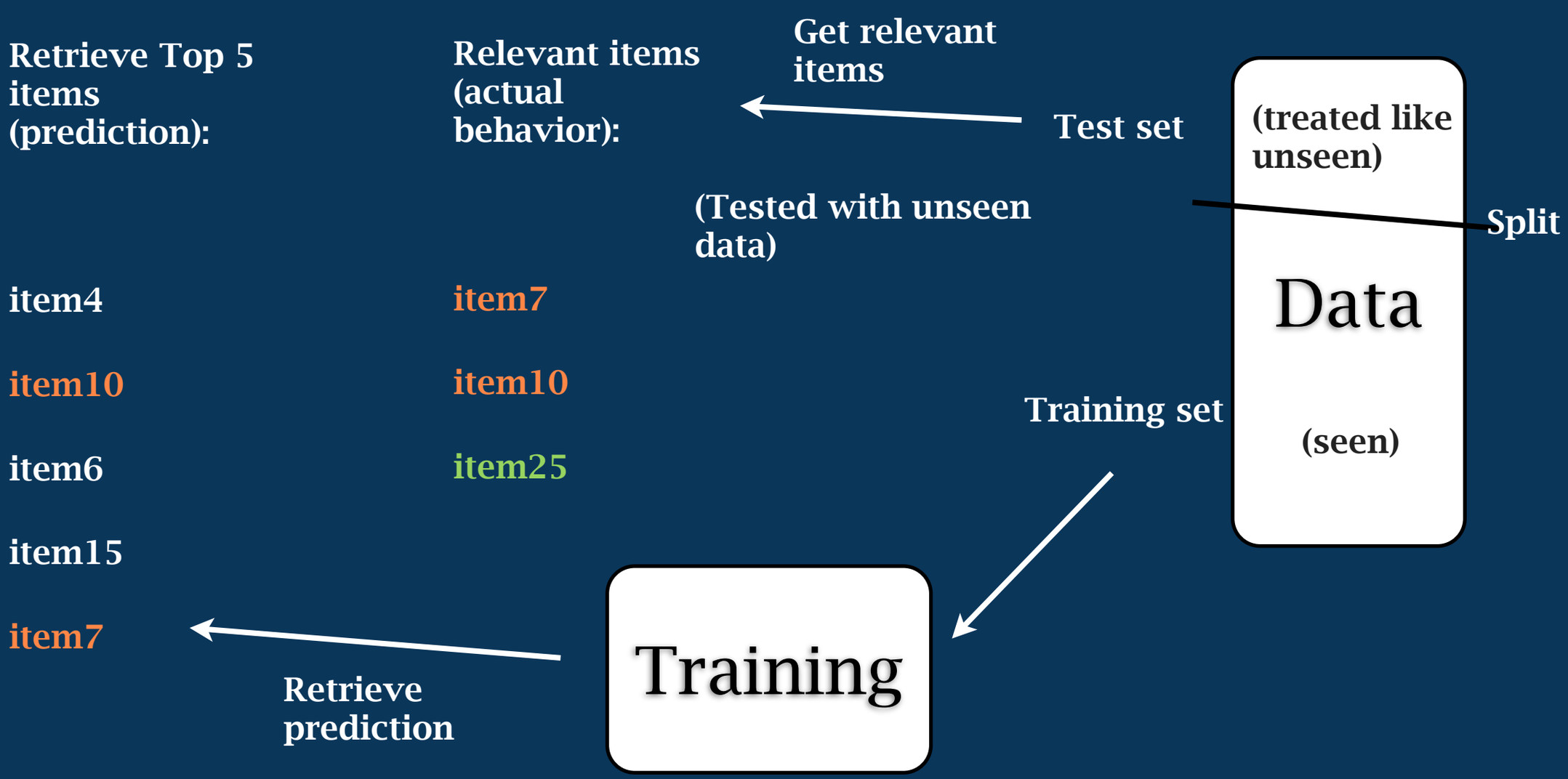
ISMAP@5 = average of MAP@5 of all items

We have metrics, but how to evaluate?

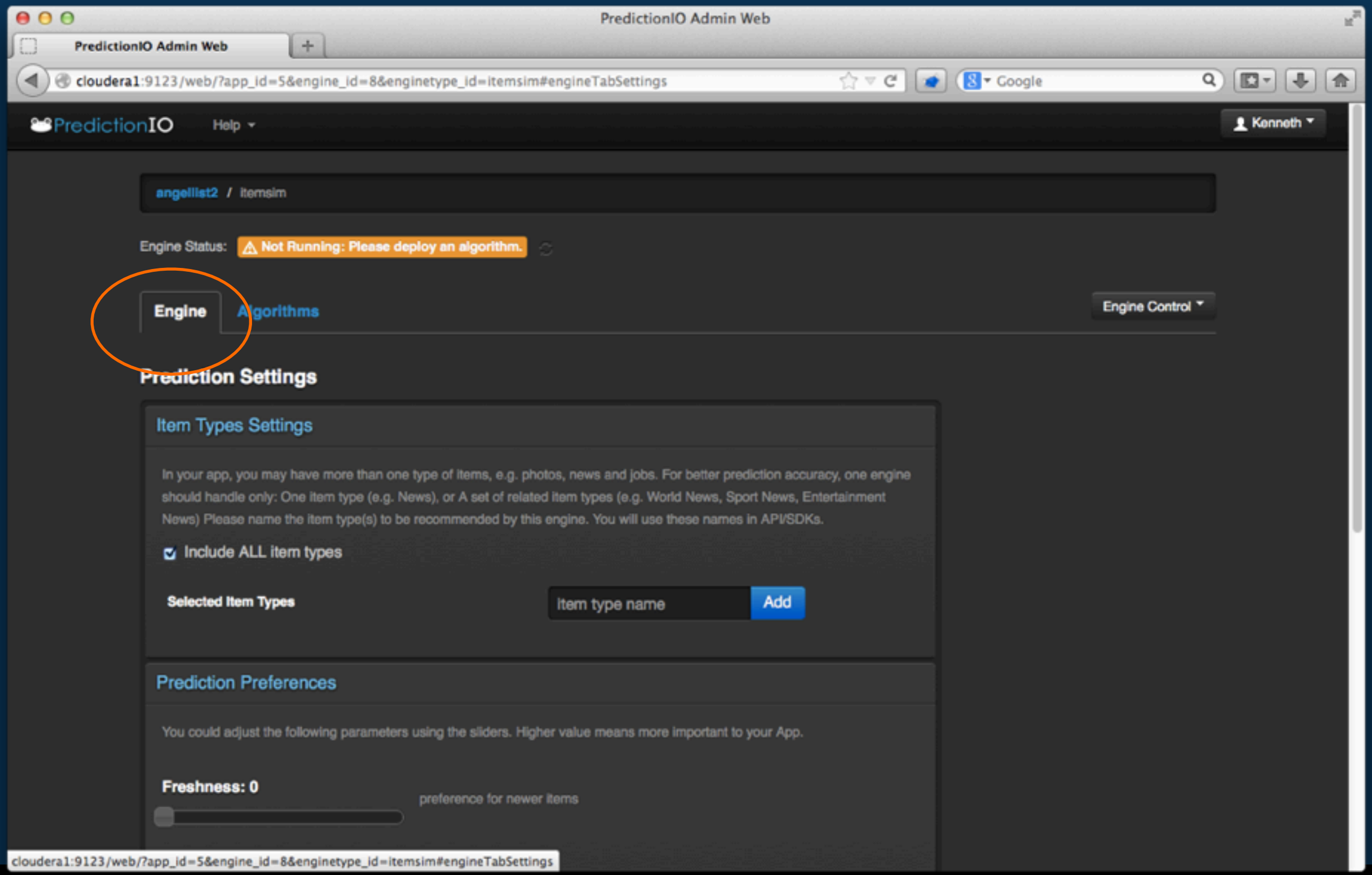
Offline Evaluation

Offline Evaluation

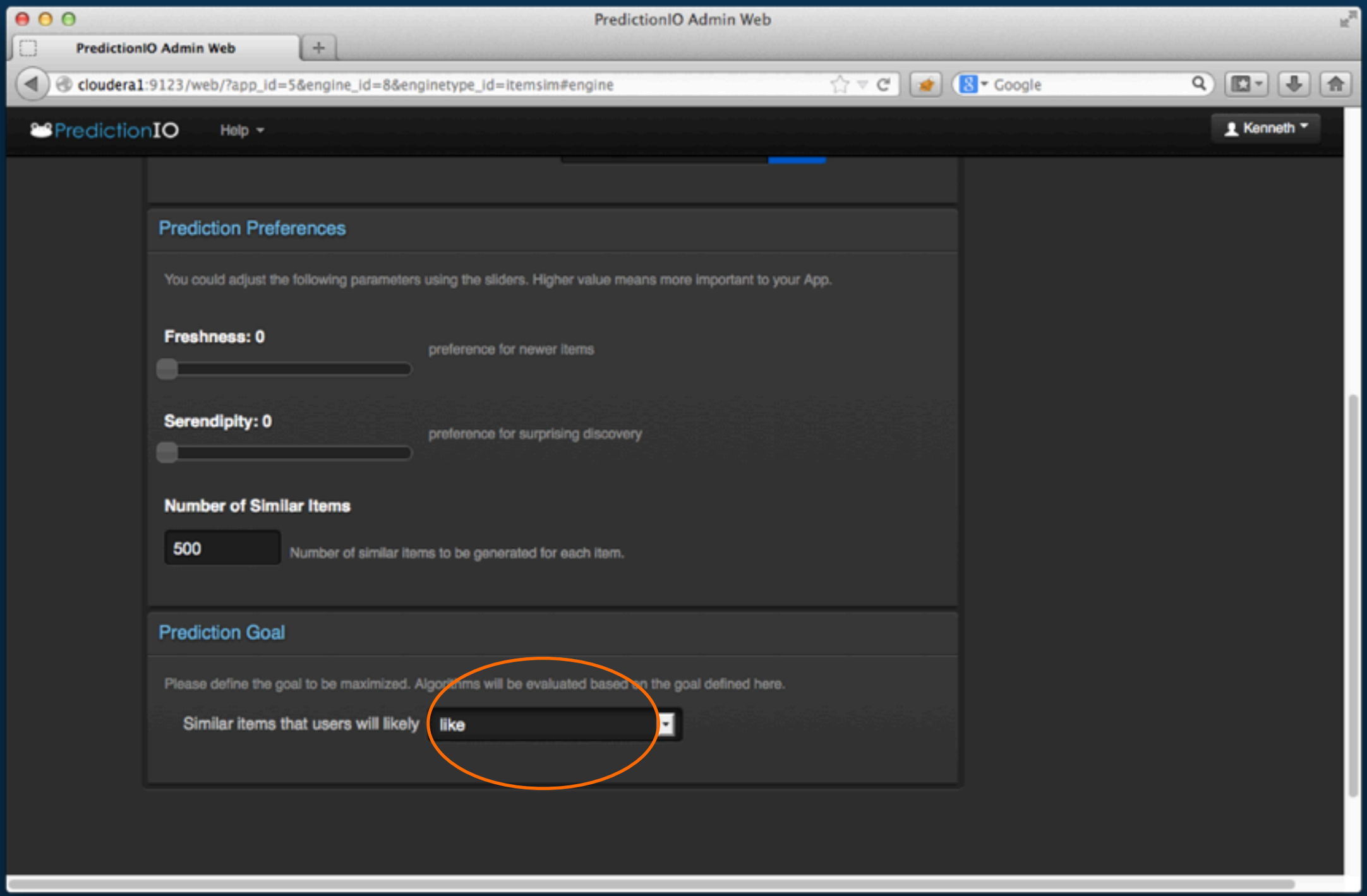
What does it mean by "relevant"? What's the prediction goal?



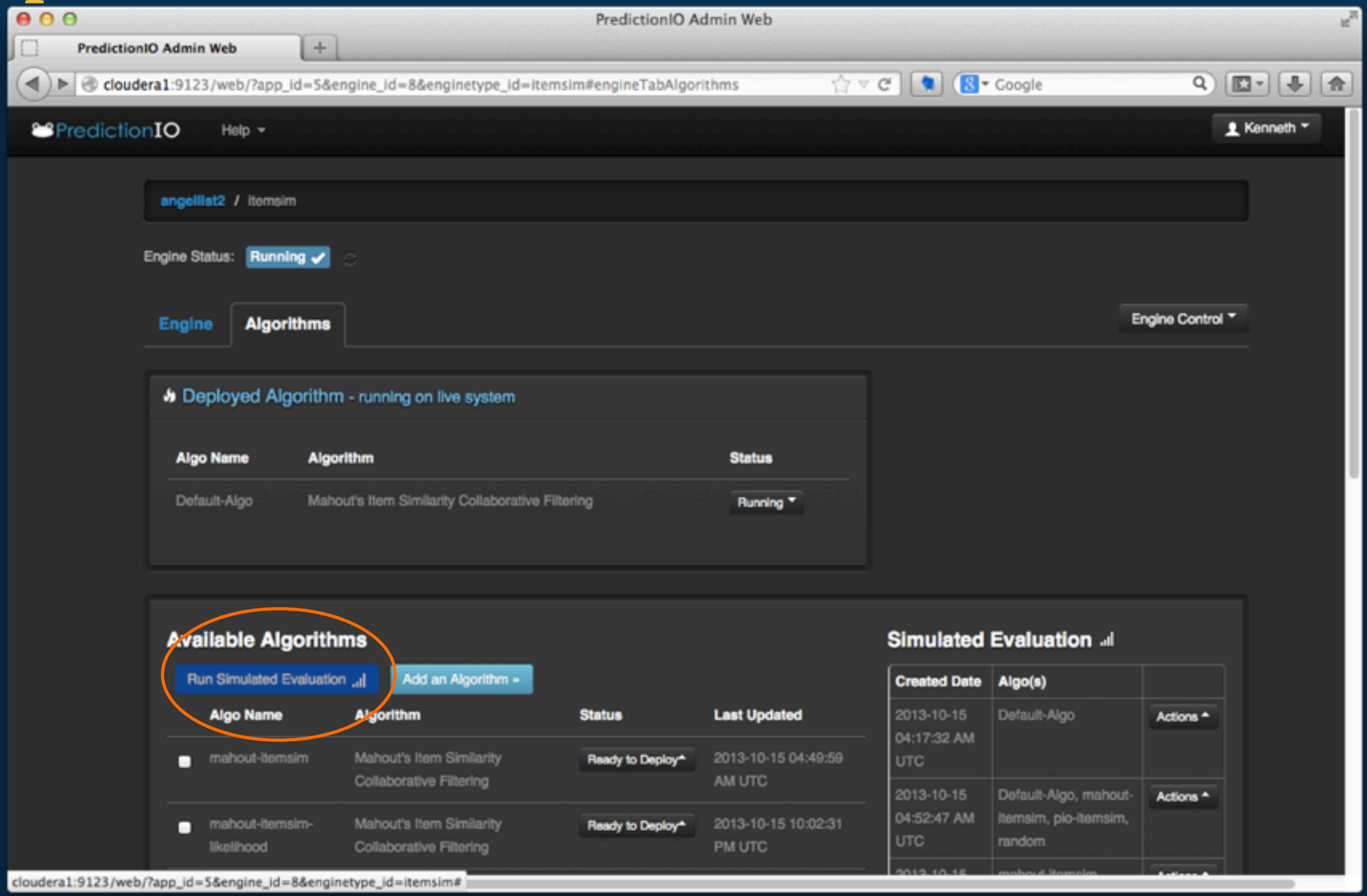
How to set the goal?



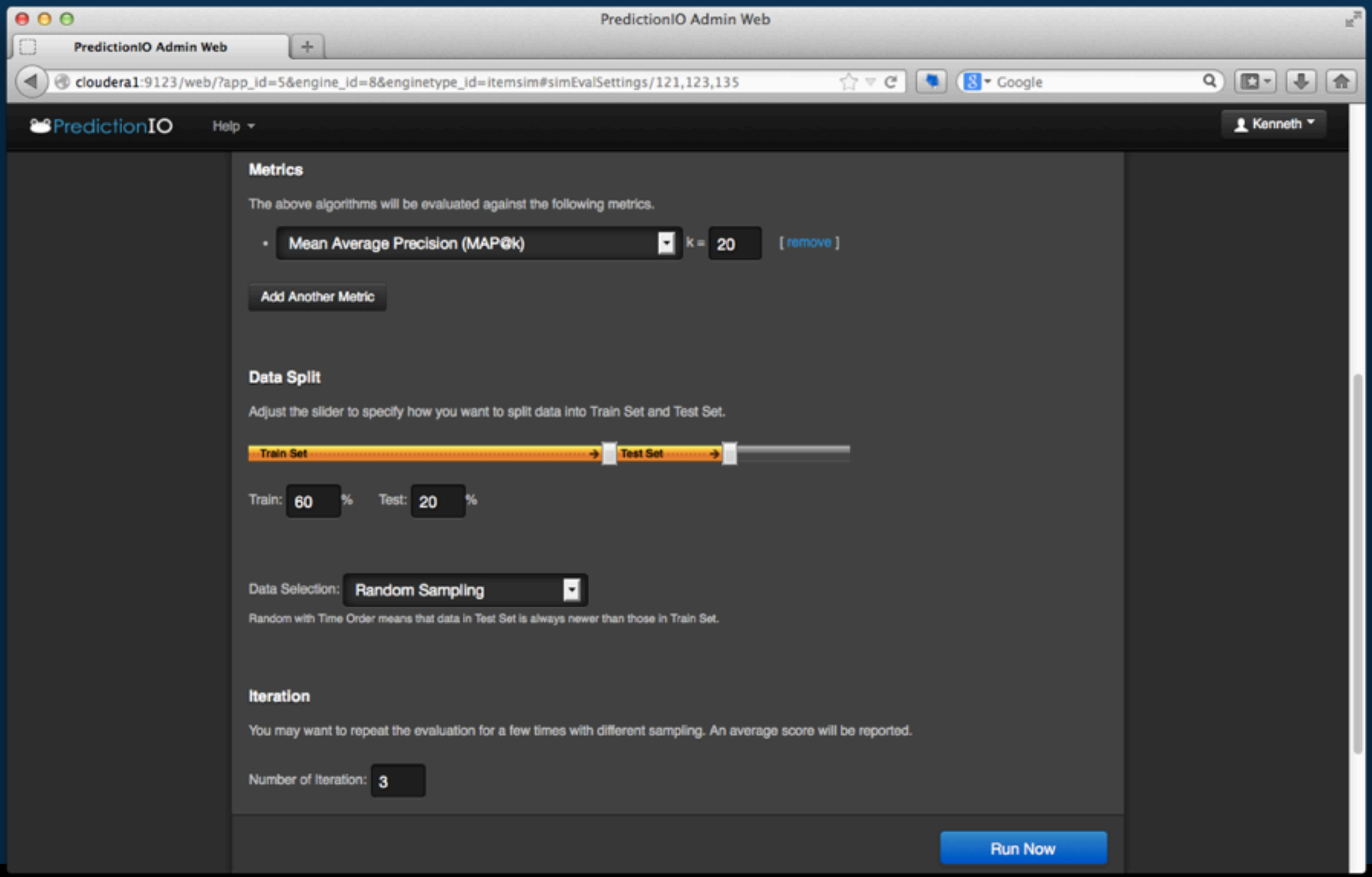
How to set the goal?



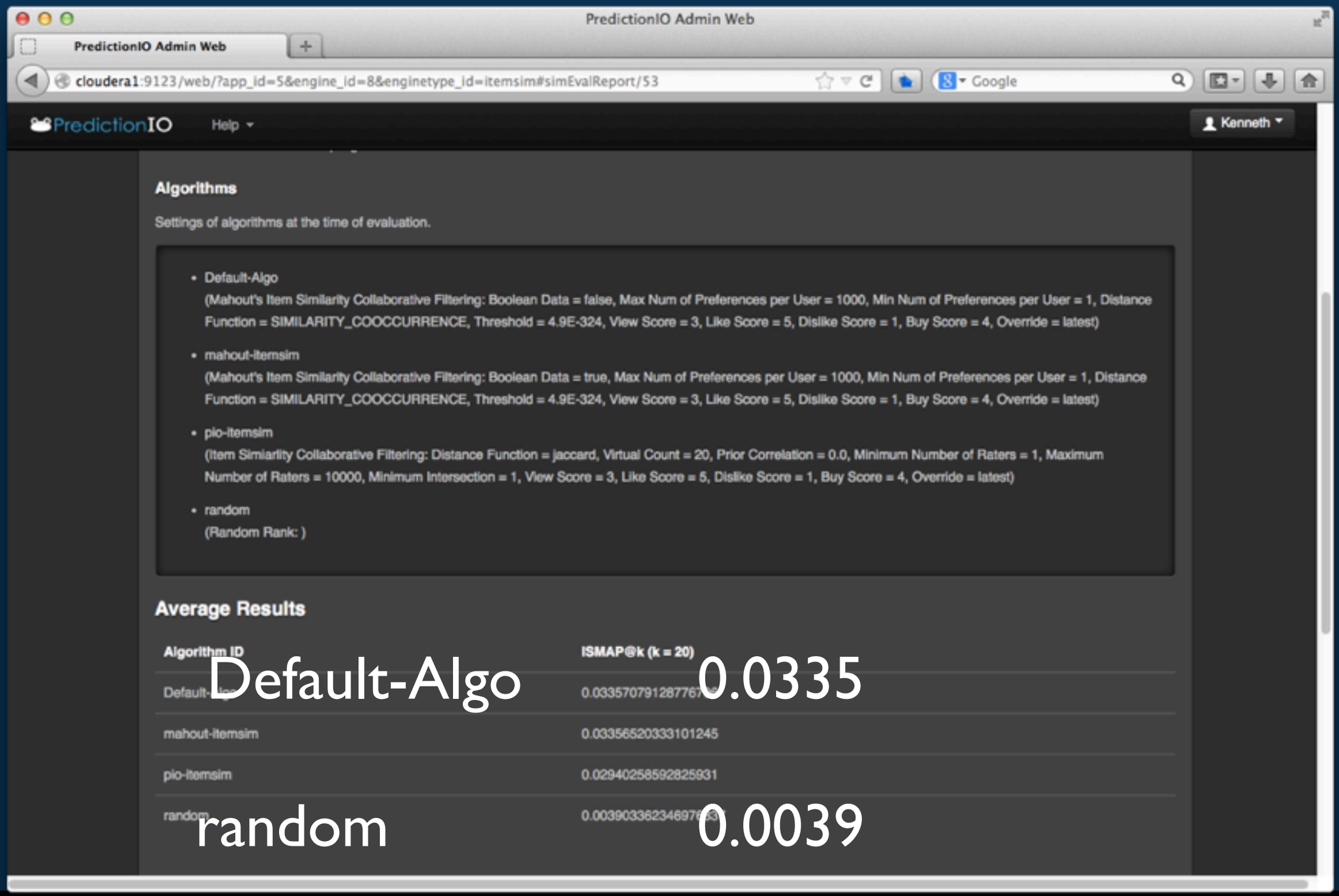
How do we evaluate how good the prediction results are?



How do we evaluate how good the prediction results are?



Sample evaluation scores



Predict User Preference

- Smarter Ride-Sharing Apps
- Personalized Meal Recommendation
- Intelligent Online Course Discovery
- App Discovery
- and many more...

@PredictionIO

Getting Started!

prediction.io

github.com/predictionio