



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

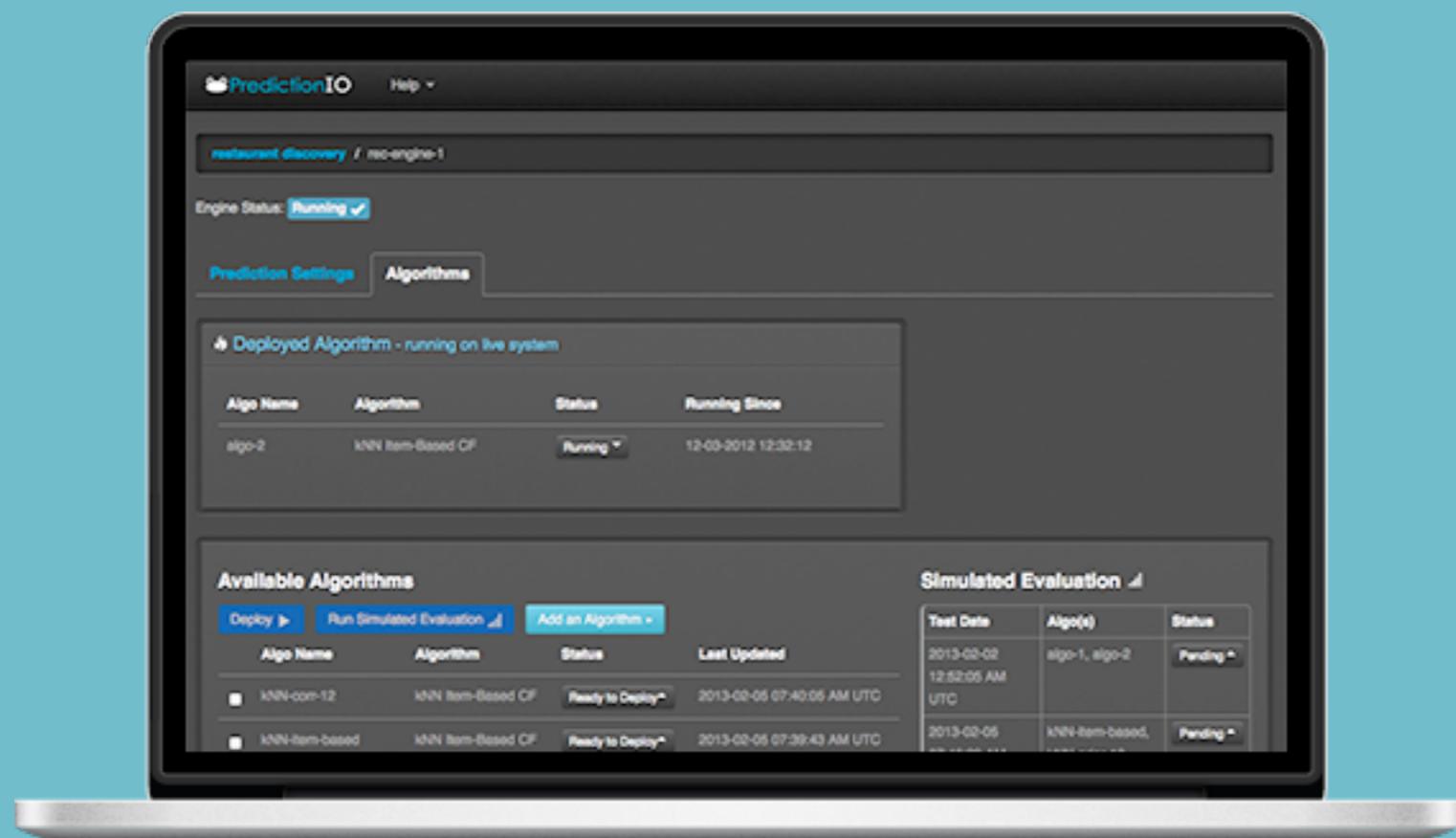




Existing data tools are like raw material

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

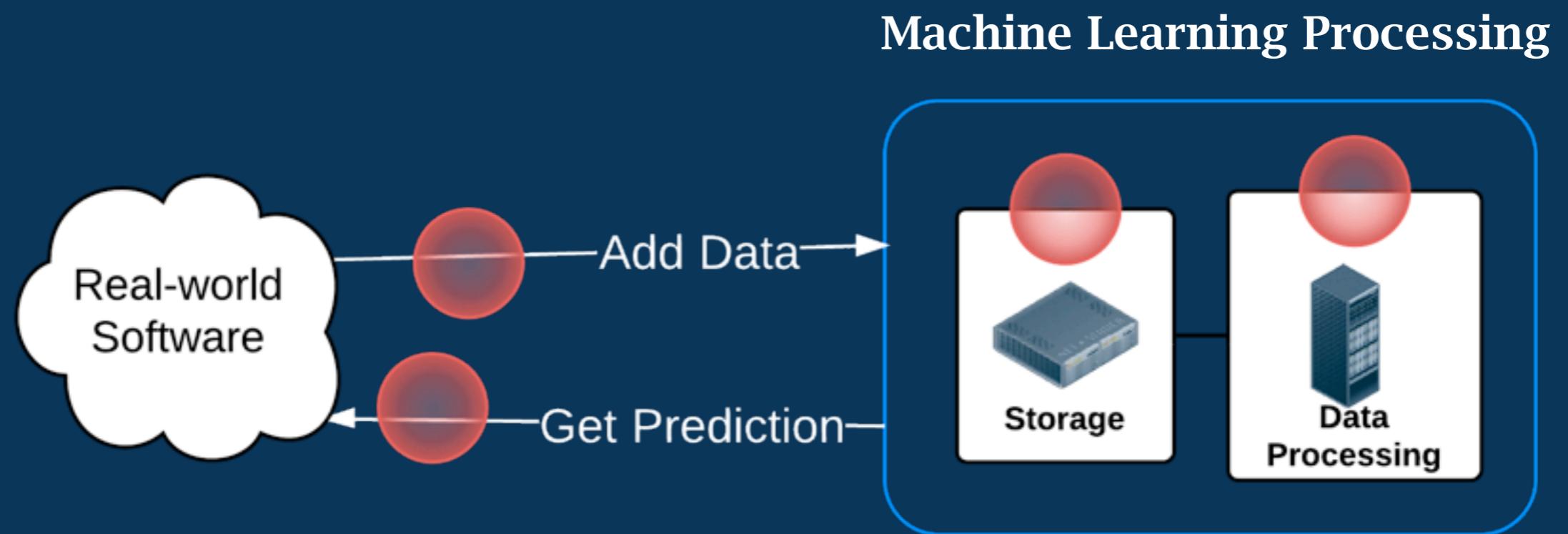


For Production

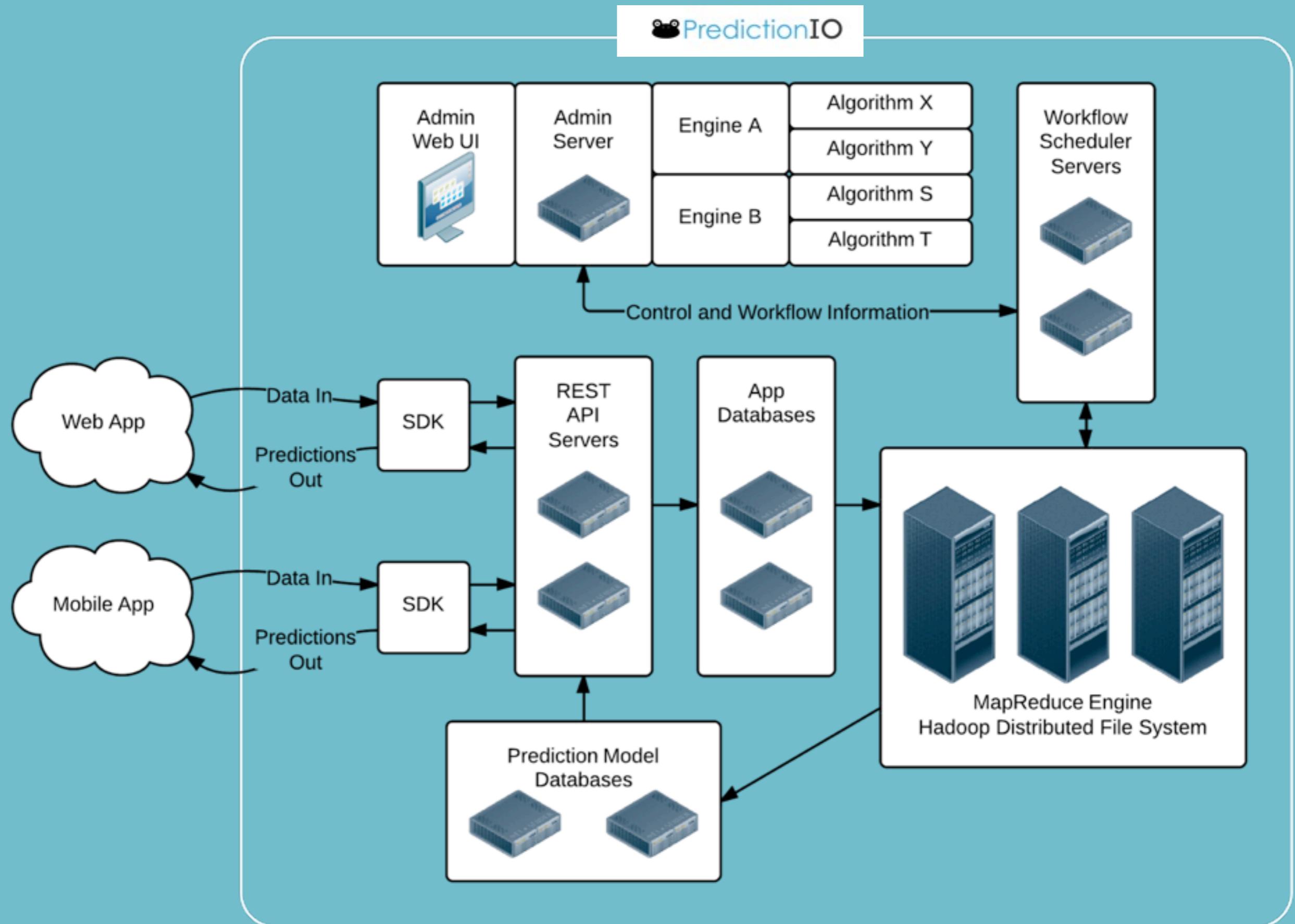
challenge #1

Scalability

Big Data Bottlenecks



PredictionIO has a
horizontally scalable
architecture



challenge #2

Productivity

Conventional Routine

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

Algorithms

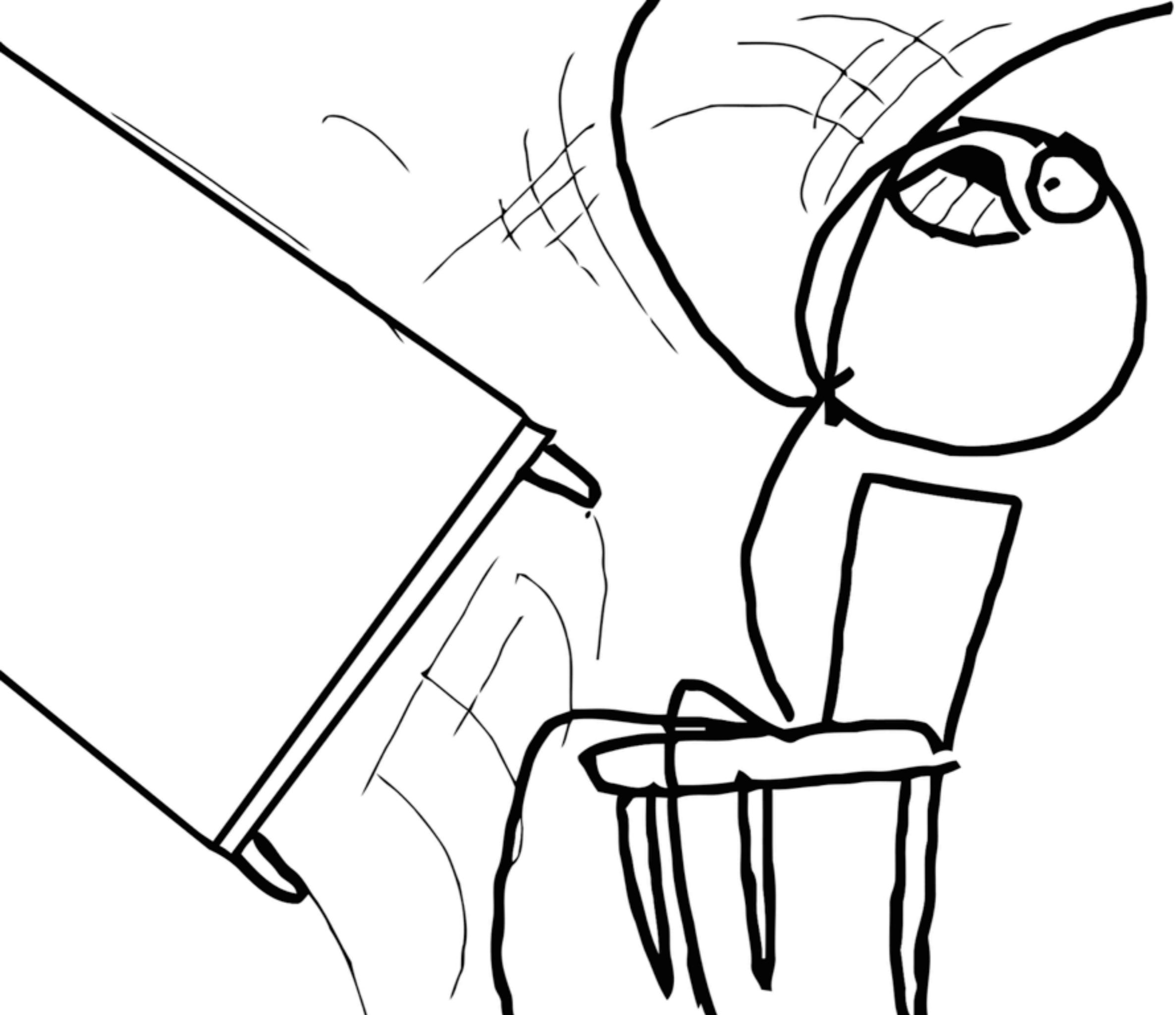
- ▶ Item Similarity
 - ▶ Collaborative Filtering (CF) ???
- ▶ Item Recommendation
 - ▶ kNN Item-based CF ???
 - ▶ Threshold Item-based CF ???
 - ▶ Alternating Least Squares ???
 - ▶ SlopeOne ???
 - ▶ Singular Value Decomposition ???

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 (!)

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 i 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 i 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 i 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 i 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 i 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 i 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 i 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 and Rate.	type algo i Add

PredictionIO Admin Web

PredictionIO Admin Web cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#algoSettings/mahout-itemsimcf/Default-A Google Kenneth

PredictionIO Help ▾

Engine Status: **Not Running: Please deploy an algorithm.**

Engine Algorithms **Algorithm Settings: Default-Algo** Engine Control ▾

Parameter Settings

Item Similarity Measurement

Distance Function **Co-occurrence**

Advanced Parameters

Boolean Data **False** Treat input data as having no preference values.

Numeric Parameters

* Auto Tuning is an experimental option.

Manual Tuning **Auto Tuning**

Threshold **4.9E-** Discard item pairs with a similarity value below this.

Max Num of Preferences per User **1000** Maximum number of preferences considered per user in final recommendation phase.

Min Num of Preferences per User **1** Ignore users with less preferences than this.

This screenshot shows the 'Algorithm Settings' page for a 'Default-Algo' engine. The top navigation bar includes tabs for 'Engine' and 'Algorithms', and a dropdown for 'Engine Control'. A prominent orange message box at the top left states 'Not Running: Please deploy an algorithm.' Below this, the 'Parameter Settings' section is divided into 'Item Similarity Measurement' and 'Advanced Parameters'. Under 'Item Similarity Measurement', the 'Distance Function' is set to 'Co-occurrence'. In the 'Advanced Parameters' section, 'Boolean Data' is set to 'False', with a note explaining it treats input data as having no preference values. The 'Numeric Parameters' section contains three settings: 'Threshold' (set to '4.9E-'), 'Max Num of Preferences per User' (set to '1000'), and 'Min Num of Preferences per User' (set to '1'). A note next to the threshold setting explains it discards item pairs with a similarity value below this threshold. The 'Auto Tuning' tab is currently selected. The bottom right corner features the PredictionIO logo.

Simulated Evaluation

Algorithms to be Evaluated

- mahout-itembased
- latest

Metrics

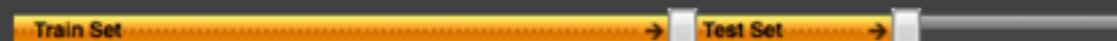
The above algorithms will be evaluated against the following metrics.

- 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: **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

**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

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

$$R_1 = [3, 5, 1, 4]$$

$$R_3 = [3, 4, 2, 5]$$

$$\text{Corr}(R_1, R_3) = 0.8$$

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

$$\begin{aligned}(4 * 0.8 + 2 * 0.9) \\ /(0.8 + 0.9) \\ = 5 / 1.7 \\ = 2.9\end{aligned}$$

	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	5	2	5		2	2			1

PredictionIO In Action

AngelList

https://angel.co/finder

Reader

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

Join Log In

AngelList

Search

HOME STARTUPS SYNDICATES JOBS ACTIVITY MORE

Finder

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 71,362 companies

StarStreet Real money fantasy sports games Boston - Games Joined Jul '10 Followers 236 Signal Path

BackType Acquired by Twitter San Francisco Joined Aug '10 Followers 143 Signal Path

Pinterest A Universal Social Catalog San Francisco - Social Media Joined Aug '10 Followers 2545 Signal Path

Ostrovok Travel booking platform for Russia Russia - E-Commerce Joined Sep '10 Followers 362 Signal Path

Clear All

Company	Joined	Followers	Signal	Path
StarStreet	Jul '10	236	Signal	Path
BackType	Aug '10	143	Signal	Path
Pinterest	Aug '10	2545	Signal	Path
Ostrovok	Sep '10	362	Signal	Path

Skillshare | AngelList

https://angel.co/skillshare

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

Reader

Join Log In

Search

Log in or sign up to see your connections to Skillshare

 Skillshare
Skillshare is a global learning community for classes.
New York City · Education · Marketplaces

Follow

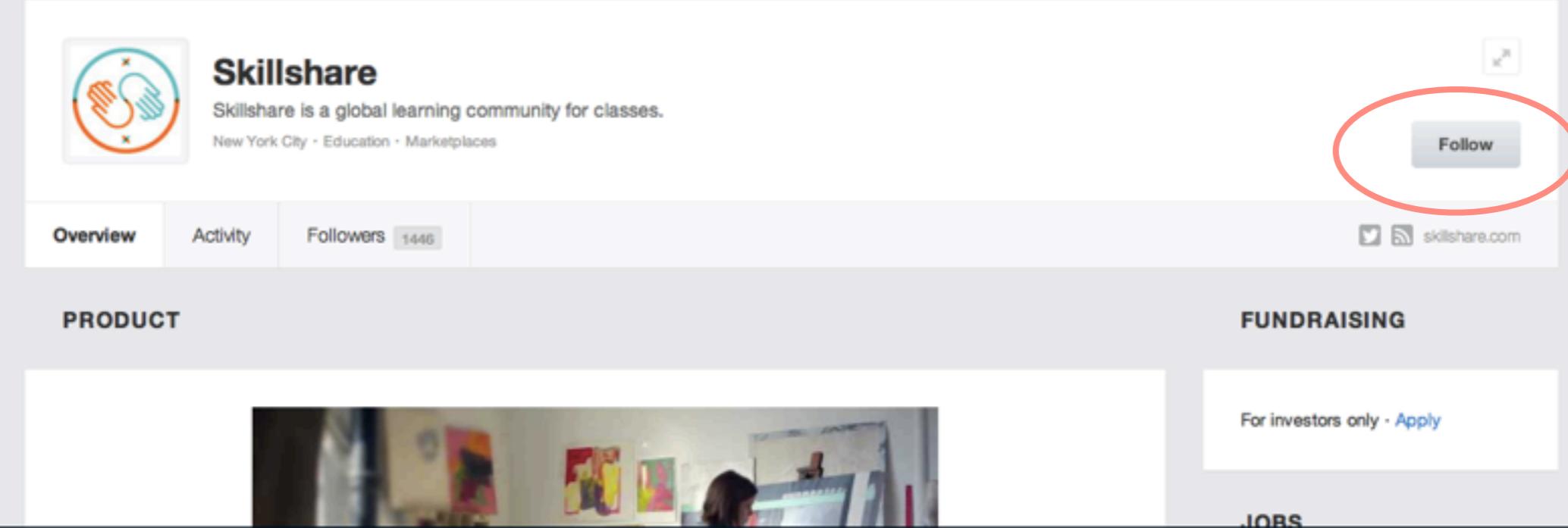
Overview Activity Followers 1446 skilshare.com

PRODUCT

FUNDRAISING

JOBS

For investors only - [Apply](#)



Something is missing here...

Skillshare | AngelList

https://angel.co/skillshare

Apple iCloud Facebook Twitter Wikipedia Yahoo! News Popular

Reader

Join Log In

Search

Log in or sign up to see your connections to Skillshare

 Skillshare

Skillshare is a global learning community for classes.
New York City • Education • Marketplaces

Follow

Overview Activity Followers 1446

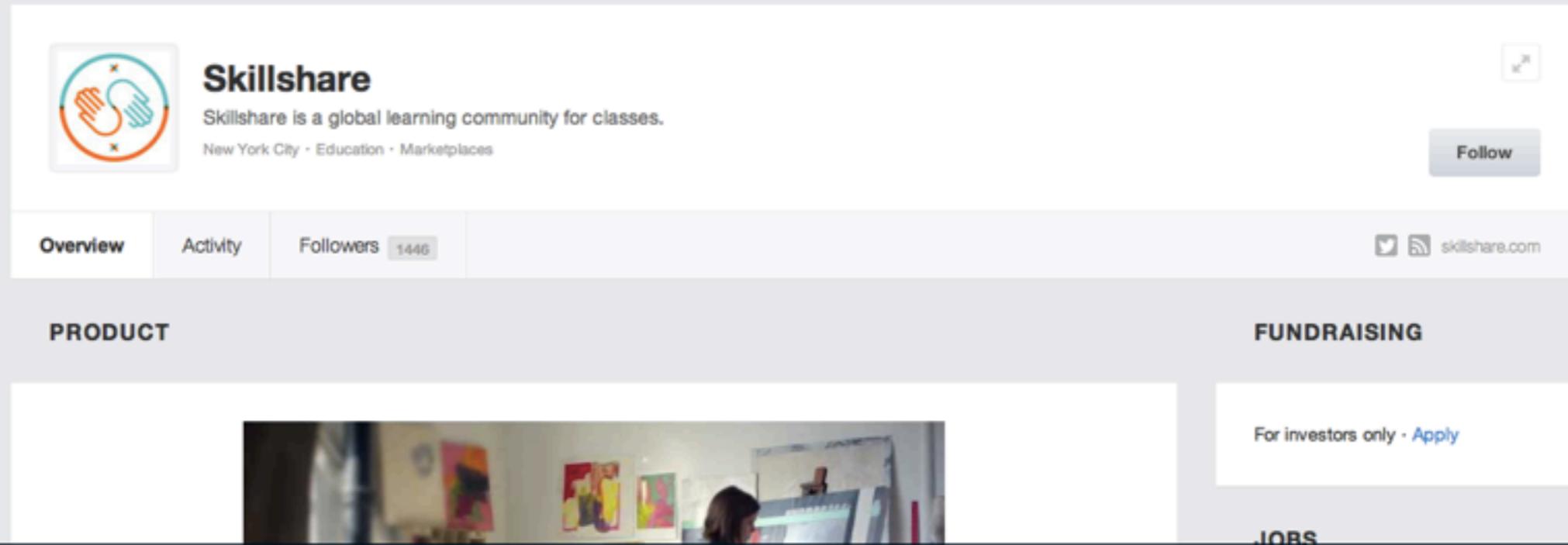
skilshare.com

PRODUCT

FUNDRAISING

For investors only - Apply

JOB



“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)

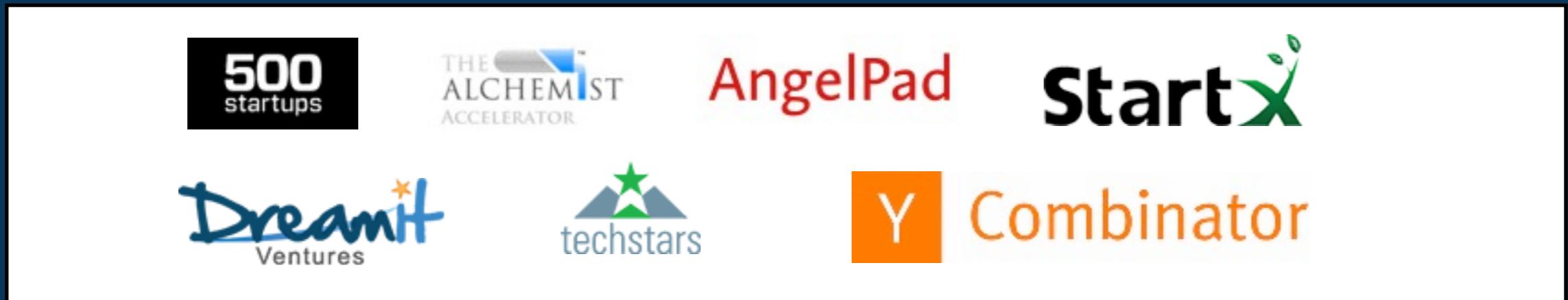
Items - the startups

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>

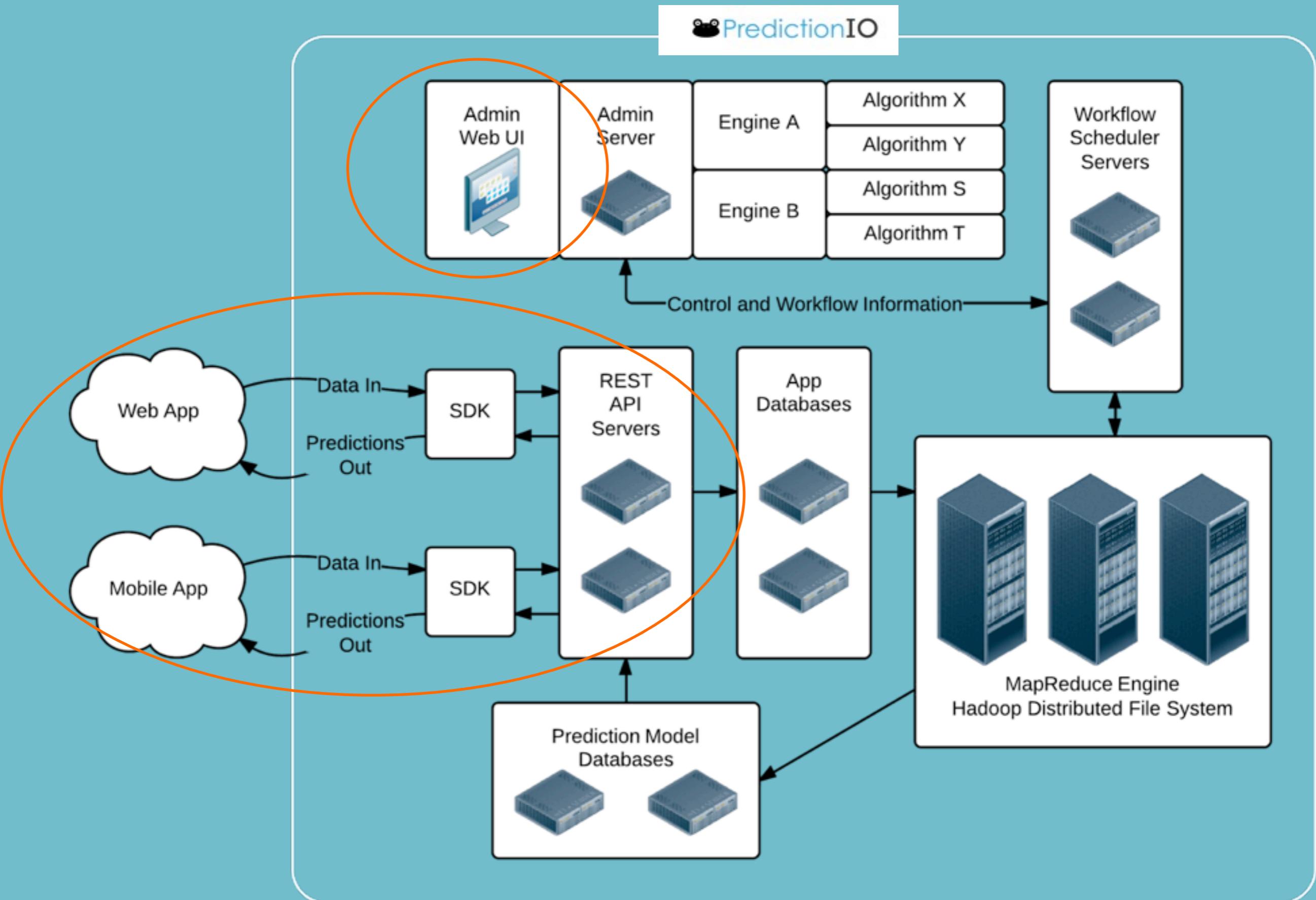
startup_id_name_url_incubator.csv - list of startups with ids, names, AngelList URL and incubators

<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?



How to integrate with PredictionIO?

The screenshot shows the PredictionIO Admin Web interface. The title bar says "PredictionIO Admin Web". The address bar shows "cloudera1:9123/web/#appsDashboard". The top navigation bar includes "PredictionIO", "Help", and a user dropdown for "Kenneth".

The main content area displays the following information for the app "angellist2":

- Last Updated: 2013-10-16 07:32:18 PM UTC
- Number of Users: 16382
- Number of Items: 993
- Number of U2I Actions: 204176

An orange circle highlights the "App Key" section, which contains the value: uM30qTyyXw0lGxugWv6V4eu4QV6cVgJWQbr3J44KvebRgufkSPgkCQAfBDVhicJV.

Below this, the "Prediction Engines" section lists one engine:

Engine Name	Type
itemsim	itemsim

Buttons for "Add an Engine" and "Manage" are present.

At the bottom, there are links for "[Reset and Erase All Data]" and "[Remove this App]".

Input fields for "Name your app. E.g. MyMobileApp" and "Add" are located at the bottom left. A "Add an App" button is at the bottom right.

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?

The screenshot shows the PredictionIO Admin Web interface running in a web browser. The title bar says "PredictionIO Admin Web". The address bar shows "cloudera1:9123/web/#appsDashboard". The top navigation bar includes the PredictionIO logo, a "Help" dropdown, and a user profile for "Kenneth".

The main content area displays the following information for the app "angellist2":

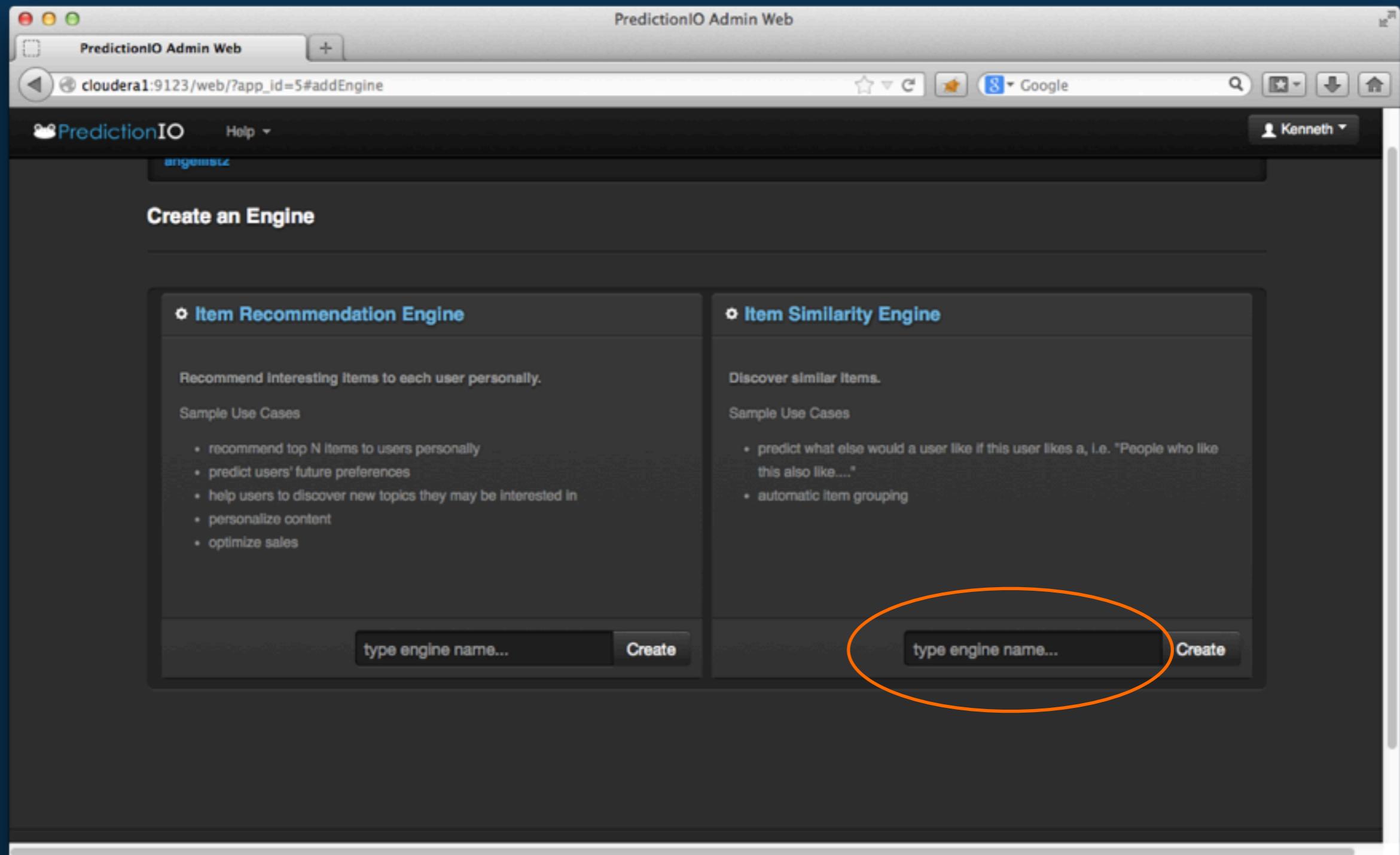
- Last Updated: 2013-10-16 07:32:18 PM UTC
- Number of Users: 16382
- Number of Items: 993
- Number of U2I Actions: 204176

Below this, there is an "App Key" section containing a long string of characters: uM30qTyyXw0lGxugWv6V4eu4QV6cVgJWQbr3J44KvebRgufkSPgkCQAfBDVhicJV.

The "Prediction Engines" section lists one engine named "itemsim" of type "itemsim", with a "Manage" button next to it. An orange oval highlights the "Add an Engine >" button located above the table.

At the bottom of the page, there are links for "[Reset and Erase All Data]" and "[Remove this App]". There is also a text input field for "Name your app. E.g. MyMobileApp" and an "Add" button. On the right side, there is a "Add an App" button.

How to integrate with PredictionIO?



How to integrate with PredictionIO?

The screenshot shows the PredictionIO Admin Web interface running in a web browser. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabAlgorithms".

The interface has a dark theme with blue highlights. At the top, there's a navigation bar with "PredictionIO" and "Help". On the right, there's a user profile for "Kenneth".

The main content area has two tabs: "Engine" (selected) and "Algorithms". The "Engine" tab shows "Engine Status: Running" with a green checkmark. The "Algorithms" tab shows a table of deployed algorithms.

The "Algorithms" table has columns: "Algo Name", "Algorithm", and "Status". It contains one row:

Algo Name	Algorithm	Status
Default-Algo	Mahout's Item Similarity Collaborative Filtering	Running

Two specific elements are highlighted with orange circles:

- The "Running" status button in the Engine status section.
- The "Default-Algo" row in the Algorithms table.

Below the "Algorithms" table, there are two sections: "Available Algorithms" and "Simulated Evaluation".

Available Algorithms

Algo Name	Algorithm	Status	Last Updated
mahout-itemsim	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 04:49:59 AM UTC
mahout-itemsim-likelihood	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 10:02:31 PM UTC

Simulated Evaluation

Created Date	Algo(s)	Actions
2013-10-15 04:17:32 AM UTC	Default-Algo	Actions
2013-10-15 04:52:47 AM UTC	Default-Algo, mahout-itemsim, pio-itemsim, random	Actions
2013-10-15 10:02:31 PM UTC	mahout-itemsim	Actions

The bottom of the page shows the URL "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabAlgorithms".

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)
```

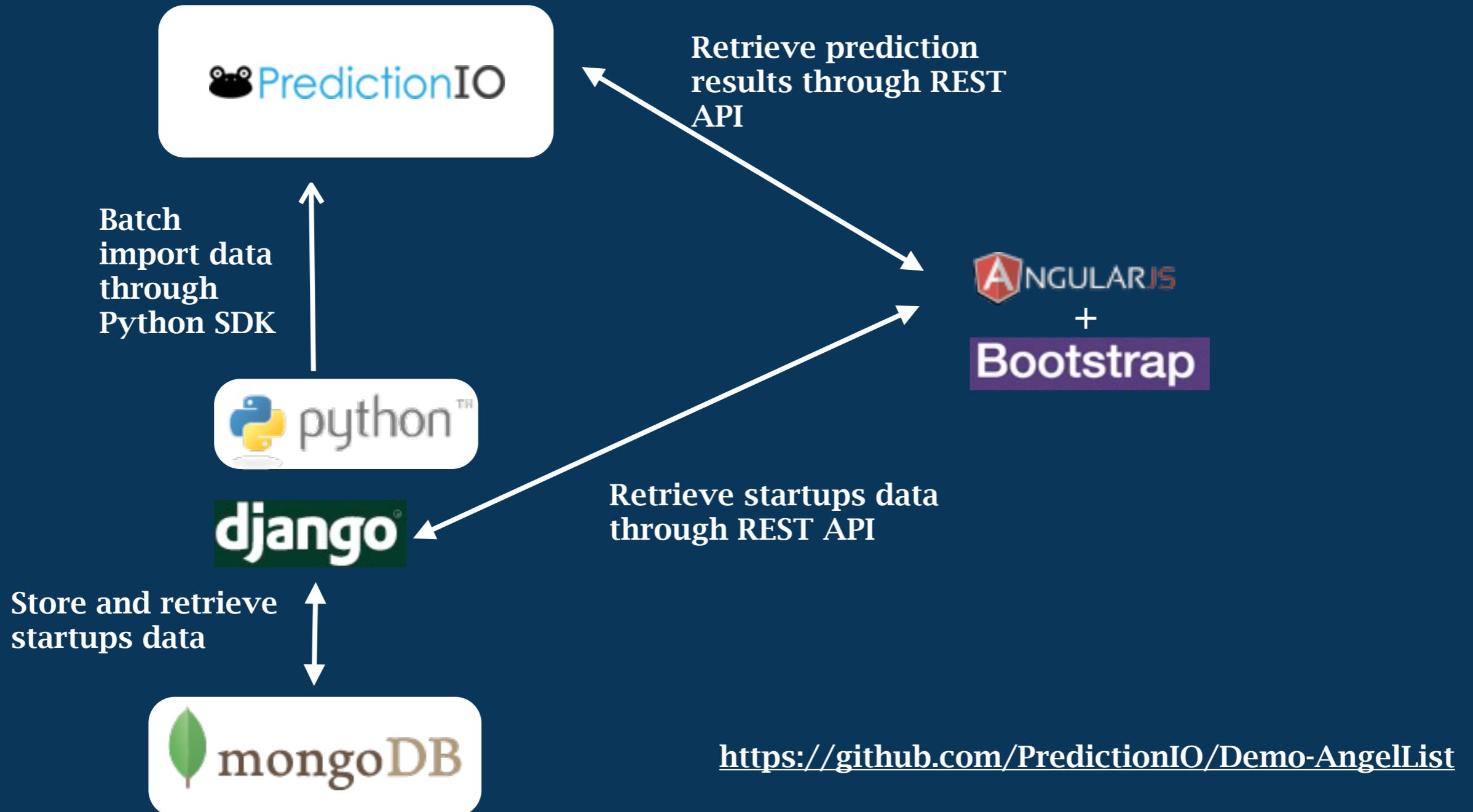
```
# Asynchronous version
```

```
req = client.aget_itemsim_topn("engine_name", "startup_id", num)
```

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

```
rec = client.aresp(req)
```

Demo Setup



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](#))

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](#))

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)
Hipset (AngelList)
Balanced (AngelList)
Yobble (AngelList)

Analytics platform for mobile & web.

E-commerce across social, mobile, and web

Data analysis tool

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)
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**

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)

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**

How to try different algorithms?

The screenshot shows the PredictionIO Admin Web interface running in a web browser. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabAlgorithms".

The main content area has a dark background. At the top, there are two tabs: "Engine" (which is selected) and "Algorithms". To the right of the tabs is a "Engine Control" dropdown menu.

Below the tabs, there is a section titled "Deployed Algorithm - running on live system". It contains a table with one row:

Algo Name	Algorithm	Status
Default-Algo	Mahout's Item Similarity Collaborative Filtering	Running

At the bottom of the page, there is a large panel divided into two sections: "Available Algorithms" on the left and "Simulated Evaluation" on the right.

Available Algorithms: This section has a heading "Available Algorithms" and a "Run Simulated Evaluation" button. Below is a table:

Algo Name	Algorithm	Status	Last Updated
mahout-itemsim	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 04:49:59 AM UTC
mahout-itemsim-likelihood	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 10:02:31 PM UTC

A blue oval highlights the "Add an Algorithm" button in the "Available Algorithms" section.

Simulated Evaluation: This section has a heading "Simulated Evaluation" and a table:

Created Date	Algo(s)	Actions
2013-10-15 04:17:32 AM UTC	Default-Algo	Actions
2013-10-15 04:52:47 AM UTC	Default-Algo, mahout-itemsim, pio-itemsim, random	Actions
2013-10-15 10:02:31 PM UTC	mahout-itemsim	Actions

How to try different algorithms?

The screenshot shows the PredictionIO Admin Web interface. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineAddAlgorithm". The top navigation bar includes "PredictionIO", "Help", and a user profile for "Kenneth". Below the navigation is a breadcrumb trail: "angellist2 / itemsim". A message "Engine Status: Not Running: Please deploy an algorithm." is displayed. The main menu has tabs for "Engine", "Algorithms", and "Add an Algorithm". The "Algorithms" tab is selected. A sub-section titled "Available Algorithms for Item Similarity Engine" lists four algorithms:

Algorithm	Description	Requirement	Data Requirement	Action
Mahout's Item Similarity Collaborative Filtering	This algorithm predicts similar items which the user may also like.	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	<input type="text" value="type algo name"/> Add
Item Similarity Collaborative Filtering	This algorithm predicts similar items which the user may also like.	Hadoop	Users, Items, and U2I Actions such as Like, Buy and Rate.	<input type="text" value="type algo name"/> Add
Latest Rank	Consider latest items as most similar.	Hadoop	Items with starttime.	<input type="text" value="type algo name"/> Add
Random Rank	Predict item similarities randomly.	Hadoop	Items.	<input type="text" value="type algo name"/> Add

An orange circle highlights the "Add" button in the first row of the table.

How to try different algorithms?

The screenshot shows the PredictionIO Admin Web interface running in a web browser. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabAlgorithms".

The main navigation bar includes "PredictionIO", "Help", and a user profile for "Kenneth". Below the navigation, the current location is "angellist2 / itemsim".

The top navigation bar has tabs for "Engine" and "Algorithms", with "Algorithms" being the active tab. A "Engine Control" dropdown is also present.

A message box displays "Deployed Algorithm - running on live system" with a table:

Algo Name	Algorithm	Status
Default-Algo	Mahout's Item Similarity Collaborative Filtering	Running

Below this, there are two main sections: "Available Algorithms" and "Simulated Evaluation".

Available Algorithms: This section contains a table with columns: "Algo Name", "Algorithm", "Status", and "Last Updated". One row is highlighted with a red circle around the "Status" button, which says "Ready to Deploy".

Algo Name	Algorithm	Status	Last Updated
mahout-itemsim	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 04:49:59 AM UTC
mahout-itemsim-likelihood	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 10:02:31 PM UTC

Simulated Evaluation: This section contains a table with columns: "Created Date", "Algo(s)", and "Actions".

Created Date	Algo(s)	Actions
2013-10-15 04:17:32 AM UTC	Default-Algo	Actions
2013-10-15 04:52:47 AM UTC	Default-Algo, mahout-itemsim, pio-itemsim, random	Actions
2013-10-15 10:02:31 PM UTC	mahout-itemsim	Actions

How to change algorithm parameters?

The screenshot shows the PredictionIO Admin Web interface. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#algoSettings/mahout-itemsimcf/Default-A". The top navigation bar has tabs for "Engine", "Algorithms", and "Algorithm Settings: Default-Algo" (which is selected). On the right, there's an "Engine Control" dropdown and a user profile for "Kenneth". A message at the top says "Engine Status: Not Running: Please deploy an algorithm.".

Parameter Settings

Item Similarity Measurement

Distance Function: Co-occurrence

Advanced Parameters

Boolean Data: False (Treat input data as having no preference values.)

Numeric Parameters

* Auto Tuning is an experimental option.

Manual Tuning (selected) Auto Tuning

Threshold: 4.9E- Discard item pairs with a similarity value below this.

Max Num of Preferences per User: 1000 (Maximum number of preferences considered per user in final recommendation phase.)

Min Num of Preferences per User: 1 (Ignore users with less preferences than this.)

How to change algorithm parameters?

The screenshot shows the "User Actions Representation Settings" page in the PredictionIO Admin Web interface. The title bar reads "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#algoSettings/mahout-itemsimcf/Default-A". The top navigation bar includes the PredictionIO logo, a "Help" dropdown, and a user profile for "Kenneth".

The main content area is titled "User Actions Representation Settings" and contains a section for "User Action Scores". A note states: "Define the preference score represented by each user action from 1 to 5. 5 is the most preferred, 1 is the least preferred. 3 is neutral." Below this are four input fields:

Action	Score
View	3
Like	5
Dislike	1
Buy	4

Below the scores is a section titled "Overriding" with the note: "When there are conflicting actions, e.g. a user gives an item a rating 5 but later dislikes it, determine which action will be considered as final preference." A dropdown menu labeled "Override Priority" is set to "Use the latest action".

At the bottom right are "Cancel" and "Change" buttons.

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):

item4

item10

item6

item15

item7

Relevant items
(actual
behavior):

item7

item10

item25

Total relevant items: 3

MAP@k (Mean Average Precision at k)

For this user A...

Retrieve Top 5
items
(prediction):

item4

item10

item6

item15

item7

Relevant items
(actual
behavior):

item7

item10

item25

Total relevant items: 3

How good is this prediction?

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	$AP@5 = (1/2 + 2/5) / \min(5,3)$
item10	2	1	1/2	
item6	3	1	0	$MAP@5 = \text{average of } AP@5 \text{ of all users}$
item15	4	1	0	
item7	5	2	2/5	

Total relevant items: 3

ISMAP@k

- Extension to MAP@k for Item Similarity Engine

MAP@k

For each user....

Retrieve Top 5
items
(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@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

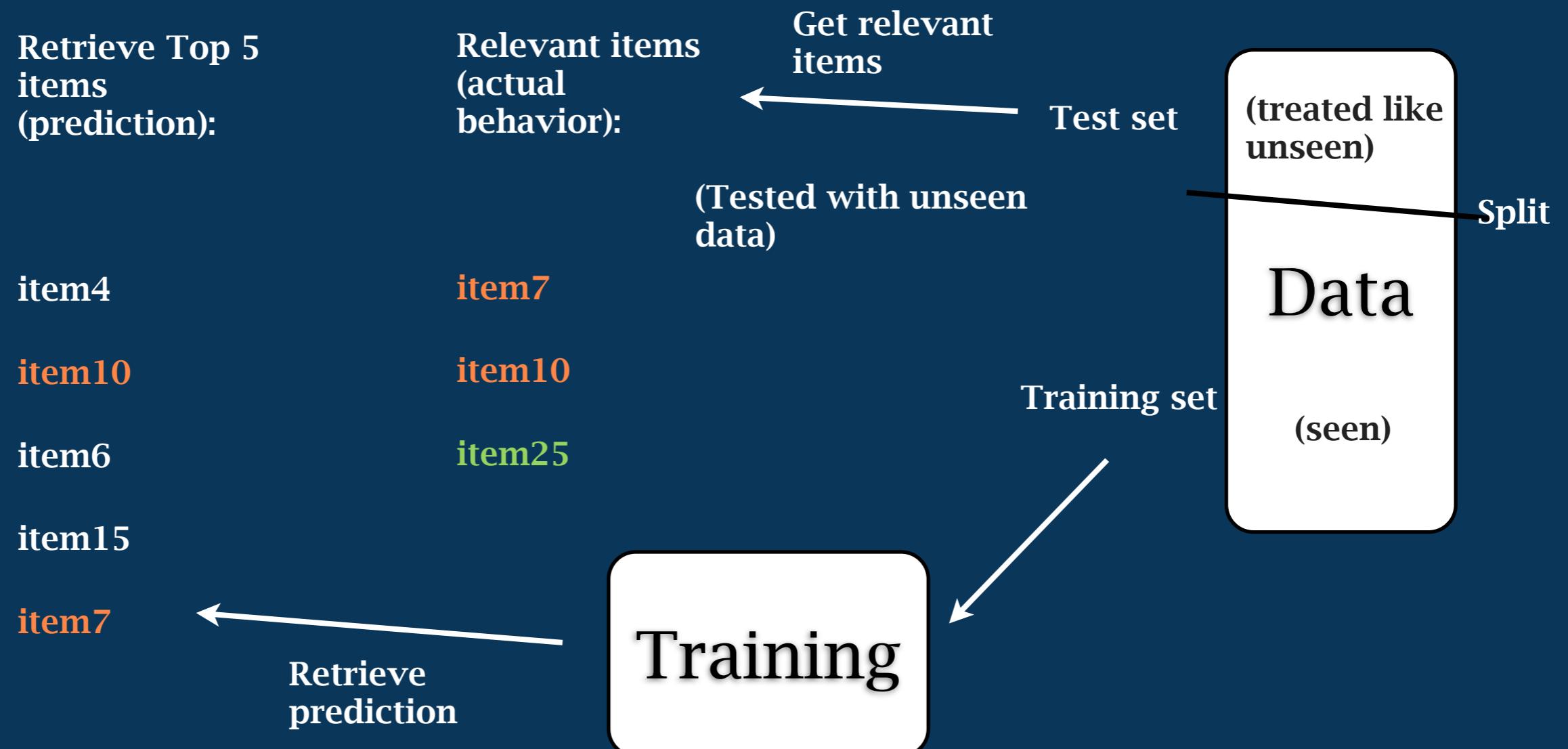
ISMAP@5 = average of MAP@5 of all items

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?

The screenshot shows the PredictionIO Admin Web interface running in a web browser on a Mac OS X system. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabSettings". The top navigation bar includes "PredictionIO", "Help", and a user profile for "Kenneth".

The main content area has a breadcrumb navigation "angellist2 / itemsim". Below it, an orange warning message says "Engine Status: Not Running: Please deploy an algorithm." A red circle highlights the "Engine" tab in the navigation bar, which is currently selected. To its right is the "Algorithms" tab. On the far right of the navigation bar is a "Engine Control" dropdown menu.

The main content area is titled "Prediction Settings". It contains two sections: "Item Types Settings" and "Prediction Preferences".

Item Types Settings: This section explains that one engine should handle only one item type or a set of related item types. It includes a checkbox "Include ALL item types" which is checked, and a "Selected Item Types" table with a "Add" button. The table currently has one row: "Item type name".

Prediction Preferences: This section allows adjusting parameters using sliders. One visible parameter is "Freshness: 0" with a note "preference for newer items".

The bottom of the page shows the same URL as the address bar: "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabSettings".

How to set the goal?

The screenshot shows the PredictionIO Admin Web interface. At the top, the title bar reads "PredictionIO Admin Web". The address bar shows the URL "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engine". The main content area has a dark background.

Prediction Preferences

You could adjust the following parameters using the sliders. Higher value means more important to your App.

Freshness: 0 preference for newer items

Serendipity: 0 preference for surprising discovery

Number of Similar Items

500 Number of similar items to be generated for each item.

Prediction Goal

Please define the goal to be maximized. Algorithms will be evaluated based on the goal defined here.

Similar items that users will likely like

A red oval highlights the input field "like" under the "Prediction Goal" section.

How do we evaluate how good the prediction results are?

The screenshot shows the PredictionIO Admin Web interface running in a web browser. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#engineTabAlgorithms". The top navigation bar includes "PredictionIO", "Help", and a user profile for "Kenneth".

The main content area has tabs for "Engine" (selected) and "Algorithms". On the right, there's an "Engine Control" dropdown. Below the tabs, a message says "⚡ Deployed Algorithm - running on live system". A table shows one algorithm:

Algo Name	Algorithm	Status
Default-Algo	Mahout's Item Similarity Collaborative Filtering	Running

At the bottom left, a section titled "Available Algorithms" contains a "Run Simulated Evaluation" button (circled in orange), an "Add an Algorithm" link, and a table:

Algo Name	Algorithm	Status	Last Updated
mahout-itemsim	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 04:49:59 AM UTC
mahout-itemsim-likelihood	Mahout's Item Similarity Collaborative Filtering	Ready to Deploy	2013-10-15 10:02:31 PM UTC

On the right, a "Simulated Evaluation" table lists evaluations:

Created Date	Algo(s)	Actions
2013-10-15 04:17:32 AM UTC	Default-Algo	Actions
2013-10-15 04:52:47 AM UTC	Default-Algo, mahout-itemsim, pic-itemsim, random	Actions
2013-10-15 10:02:31 PM UTC	mahout-itemsim	Actions

The bottom status bar shows the URL "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#".

How do we evaluate how good the prediction results are?

Screenshot of the PredictionIO Admin Web interface showing configuration for evaluation settings.

Metrics
The above algorithms will be evaluated against the following metrics.
• Mean Average Precision (MAP@k) [remove]

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

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

Iteration
You may want to repeat the evaluation for a few times with different sampling. An average score will be reported.
Number of Iteration:

Sample evaluation scores

The screenshot shows the PredictionIO Admin Web interface. The title bar says "PredictionIO Admin Web". The URL in the address bar is "cloudera1:9123/web/?app_id=5&engine_id=8&enginetype_id=itemsim#simEvalReport/53". The top navigation bar includes "PredictionIO", "Help", and a user profile for "Kenneth".

Algorithms

Settings of algorithms at the time of evaluation.

- Default-Algo
(Mahout's Item Similarity Collaborative Filtering: Boolean Data = false, Max Num of Preferences per User = 1000, Min Num of Preferences per User = 1, Distance Function = SIMILARITY_COOCURRENCE, Threshold = 4.9E-324, View Score = 3, Like Score = 5, Dislike Score = 1, Buy Score = 4, Override = latest)
- mahout-itemsim
(Mahout's Item Similarity Collaborative Filtering: Boolean Data = true, Max Num of Preferences per User = 1000, Min Num of Preferences per User = 1, Distance Function = SIMILARITY_COOCURRENCE, Threshold = 4.9E-324, View Score = 3, Like Score = 5, Dislike Score = 1, Buy Score = 4, Override = latest)
- pio-itemsim
(Item Similarity Collaborative Filtering: Distance Function = jaccard, Virtual Count = 20, Prior Correlation = 0.0, Minimum Number of Raters = 1, Maximum Number of Raters = 10000, Minimum Intersection = 1, View Score = 3, Like Score = 5, Dislike Score = 1, Buy Score = 4, Override = latest)
- random
(Random Rank:)

Average Results

Algorithm ID	ISMAP@k (k = 20)
Default-Algo	0.033570791287767
mahout-itemsim	0.03356520333101245
pio-itemsim	0.02940258592825931
random	0.0039033623469763

Predict User Preference

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

Getting Started!



@PredictionIO



prediction.io



github.com/predictionio