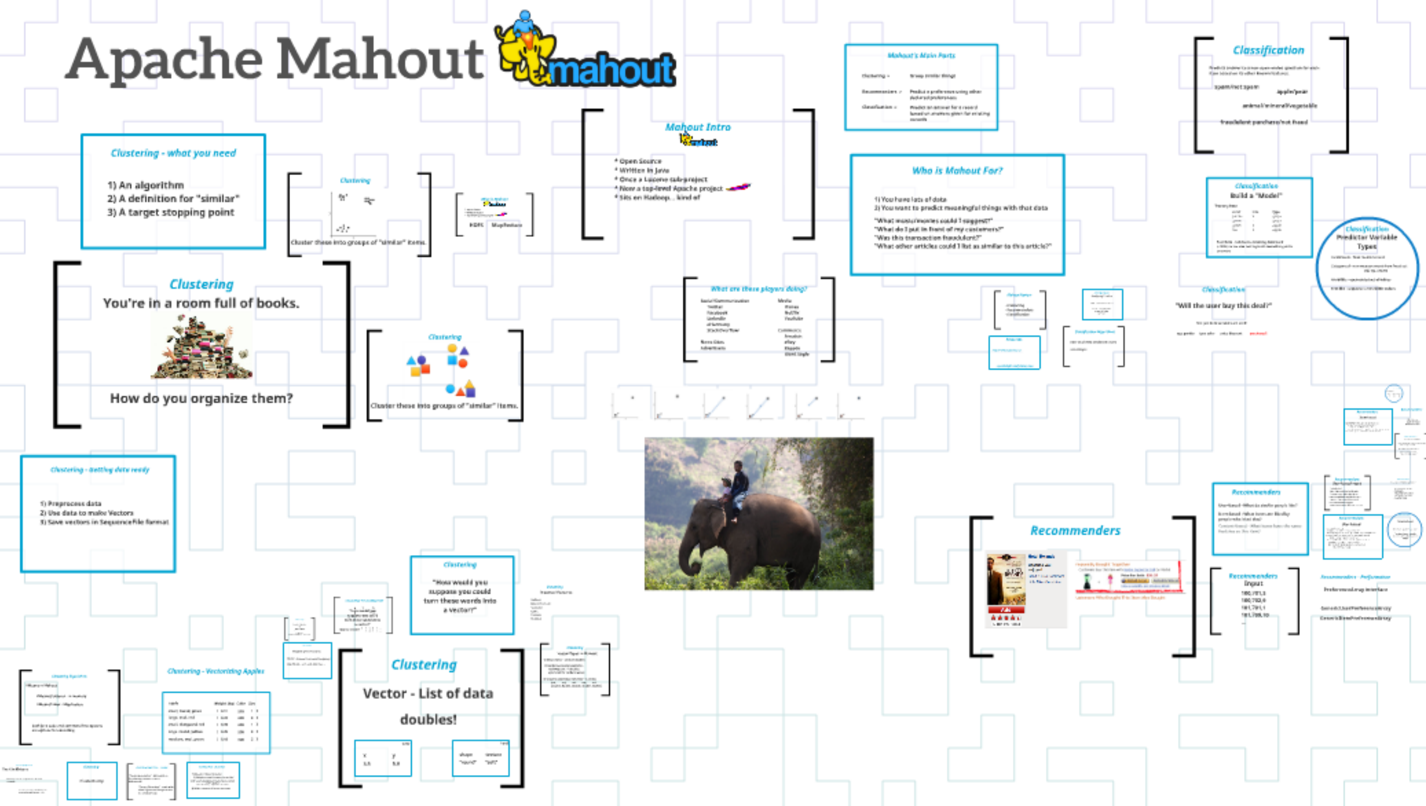

Apache Mahout
mahout
Clustering - what you need
1) An algorithm
2) A definition for "similar"
3) A target stopping point
Clustering
You're in a room full of books.
How do you organize them?
Clustering
Cluster these into groups of "similar" items.
Clustering - Getting data ready
1) Preprocess data
2) Use data to make Vectors
3) Save vectors in SequenceFile format
Clustering - Vectorizing Apples
Clustering
"How would you suppose you could turn these words into a vector?"
Clustering
Vector - List of data
doubles!
Mahout Intro
* Open Source
* Written in Java
* Once a Lucene sub-project
* Now a top-level Apache project
What are these players doing?
Mahout's Main Parts
Who is Mahout For?
1) You have lots of data
2) You want to predict meaningful things with that data
"What music/movies could I suggest?"
"What do I put in front of my customers?"
"Was this transaction fraudulent?"
"What other articles could I list as similar to this article?"
Classification
spam/not spam
apple/pear
animal/mineral/vegetable
fraudulent purchase/not fraud
Classification
Build a "Model"
Classification
Predictor Variable Types
Classification
"Will the user buy this deal?"
Recommenders
Recommenders
Recommenders
Input

Apache Mahout

Clustering - what you need

1) An algorithm
2) A definition for "similar"
3) A target stopping point

Clustering

Cluster these into groups of "similar" items.

Clustering

You're in a room full of books.

How do you organize them?

Clustering

Cluster these into groups of "similar" items.

Clustering - Getting data ready

1) Preprocess data
2) Use data to make Vectors
3) Save vectors in SequenceFile format

Clustering - Vectorizing Apples

Clustering

"How would you suppose you could turn these words into a vector?"

Clustering

Vector - List of data doubles!

x	y
3.5	5.9

shape	texture
"round"	"soft"

Mahout Intro

* Open Source
* Written in Java
* Now a top-level Apache project
* Sits on Hadoop... kind of

Mahout's Main Parts

Clustering >	Group similar things
Recommenders >	Predict a preference using other declared preferences
Classification >	Predict an answer for a record based on answers given for existing records

Who is Mahout For?

1) You have lots of data
2) You want to predict meaningful things with that data

"What music/movies could I suggest?"
"What do I put in front of my customers?"
"Was this transaction fraudulent?"
"What other articles could I list as similar to this article?"

What are these players doing?

Social/Communication
Twitter
Facebook
LinkedIn
eHarmony
StackOverflow

News Sites
Yahoo!Users

Media
iTunes
Netflix
YouTube

Commerce
Amazon
eBay
Zappos
Giant Eagle

Recommenders

Classification

Predicts answer to a nonsupervised question for each item based on its other known features.

spam/not spam

apple/pear

animal/mineral/vegetable

fraudulent purchase/not fraud

Classification

Build a "Model"

Classification

Predictor Variable Types

Classification

"Will the user buy this deal?"

Recommenders

User-Based - What do similar people like?
Item-Based - What items are liked by people who liked this?
Content-based - What items have the same features as this item?

Recommenders

Input

100,701,3
100,702,9
101,701,1
101,709,10
...

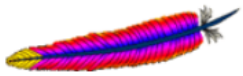
Recommenders - Performance

PreferenceArray interface

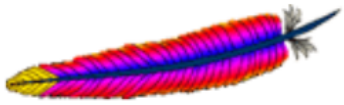
GenericUserPreferenceArray

GenericItemPreferenceArray

Mahout Intro

* Open Source
* Written in Java
* Once a Lucene sub-project
* Now a top-level Apache project 
* Sits on Hadoop... kind of

What is Hadoop?

* Open Source
* Written in Java
* Top-level Apache project 

HDFS **MapReduce**

Who is Mahout For?

1) You have lots of data
2) You want to predict meaningful things with that data

"What music/movies could I suggest?"
"What do I put in front of my customers?"
"Was this transaction fraudulent?"
"What other articles could I list as similar to this article?"

Mahout's Main Parts

Clustering >	**Group similar things**
Recommenders >	**Predict a preference using other declared preferences**
Classification >	**Predict an answer for a record based on answers given for existing records**

What are these players doing?

Social/Communication

- Twitter
- Facebook
- LinkedIn
- eHarmony
- StackOverflow

News Sites

Advertisers

Media

- iTunes
- Netflix
- YouTube

Commerce

- Amazon
- eBay
- Zappos
- Giant Eagle

Cluster these into gr

Clustering

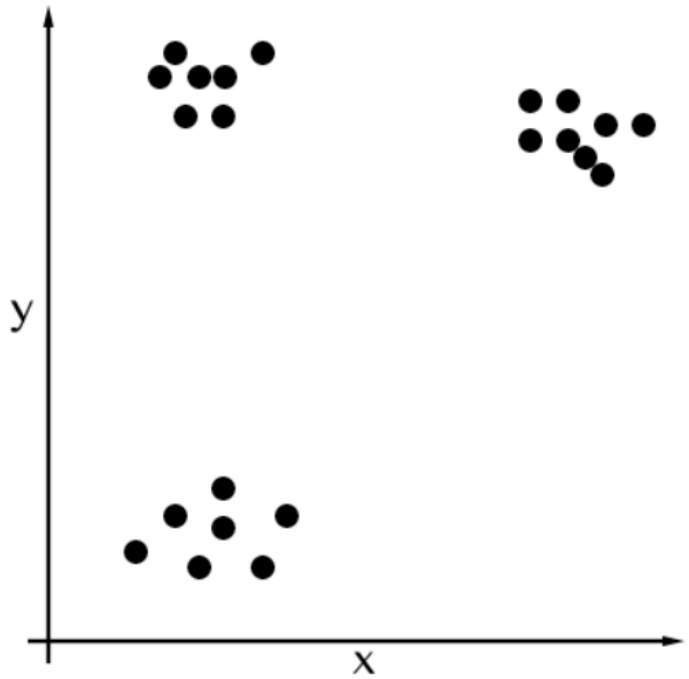
You're in a room full of books.

How do you organize them?

Clustering - what you need

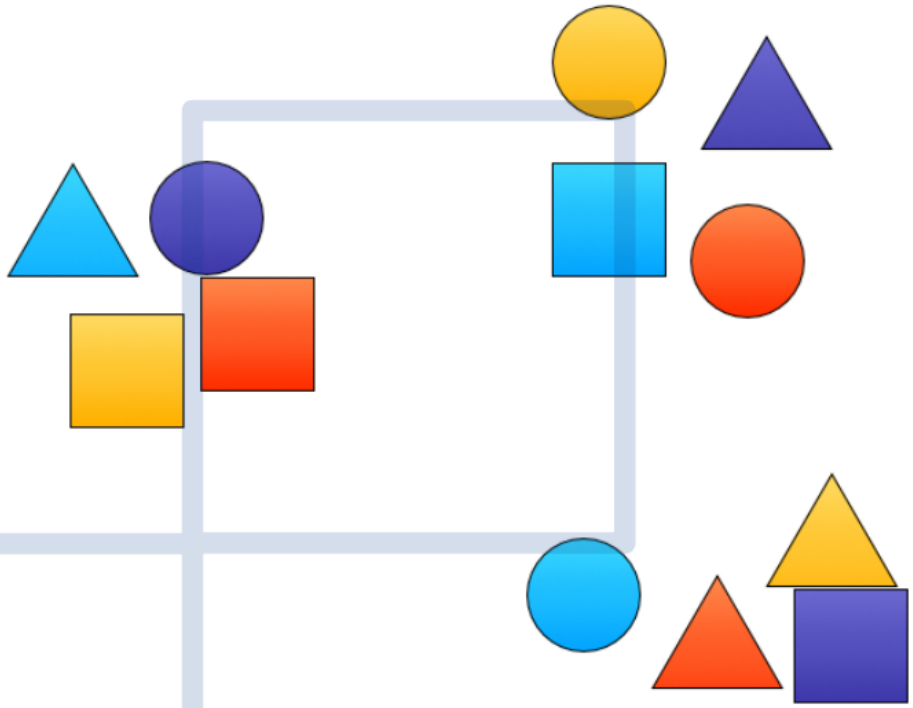
1) An algorithm
2) A definition for "similar"
3) A target stopping point

Clustering



Cluster these into groups of "similar" items.

Clustering



Cluster these into groups of "similar" items.

Clustering - Getting data ready

1) Preprocess data
2) Use data to make Vectors
3) Save vectors in SequenceFile format

Clustering

Vector - List of data doubles!

easy

x	y
3.5	5.9

hard

shape	texture
"round"	"soft"

Clustering - Vectorizing Apples

Apple	Weight (kg)	Color	Size
small, round, green	[0.11	510	1]
large, oval, red	[0.23	650	3]
small, elongated, red	[0.09	630	1]
large, round, yellow	[0.25	590	3]
medium, oval, green	[0.18	520	2]

Clustering

"How would you suppose you could turn these words into a vector?"

Vector Types in Mahout

1) DenseVector - array of doubles

2) RandomAccessSparseVector -
HashMap {int => double}
optimized for random access

3) SequentialAcessSparseVector - 2 Arrays
[int, int, int, int, int]
[double, double, double, double, double]

Clustering

Distance Measures

Euclidean

Squared Euclidean

Manhatten

Cosine

Tanimoto

Weighted

Clustering - Vectorizing Text

0 1 2
"How would you
3 4
suppose you could
…
turn these words into
a vector?"

Sparse Vector

0	1	2	3	4	…
1	1	2	1	1	…

Weighting Text Features

TF-IDF - Inverse Document Frequency

Stop Words - a, if, and, but, the, ...

Clustering

Michael Jordan
vs
Jordan River

n-grams
DictionaryVectorizer class

Clustering Algorithms

KMeans in Mahout

KMeansClusterer - in memory

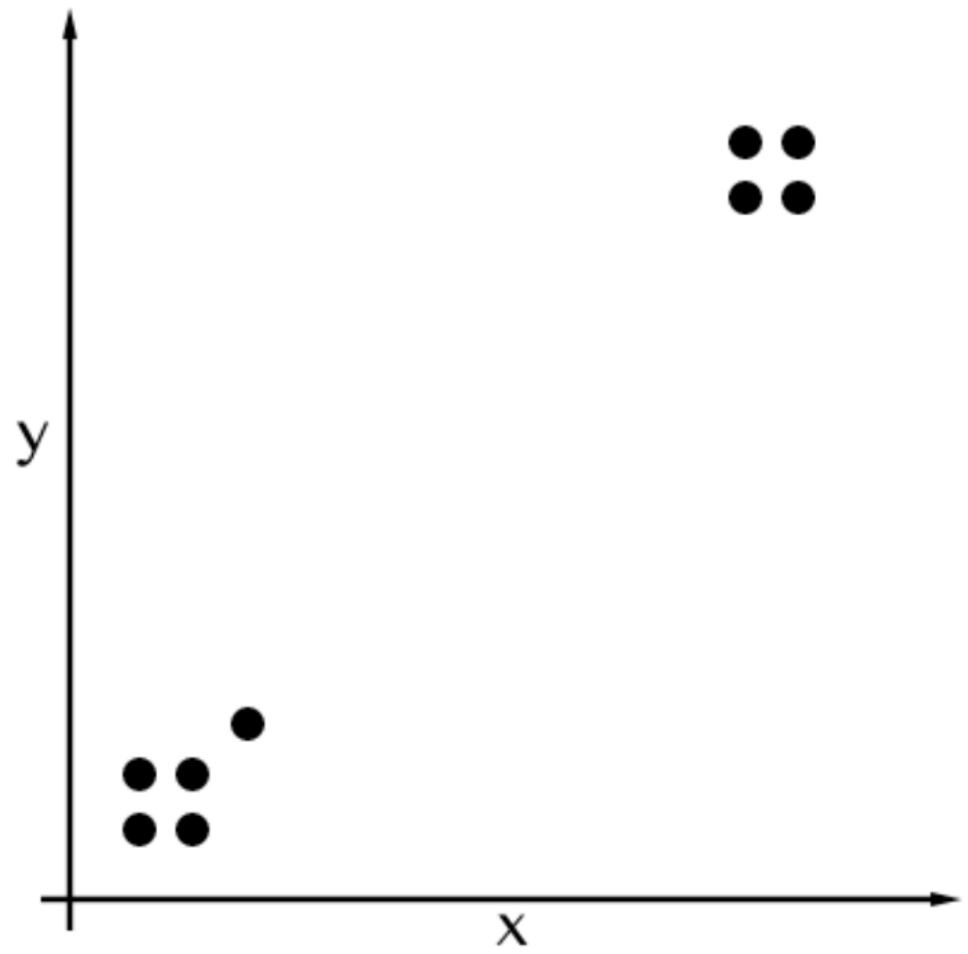
KMeansDriver - MapReduce

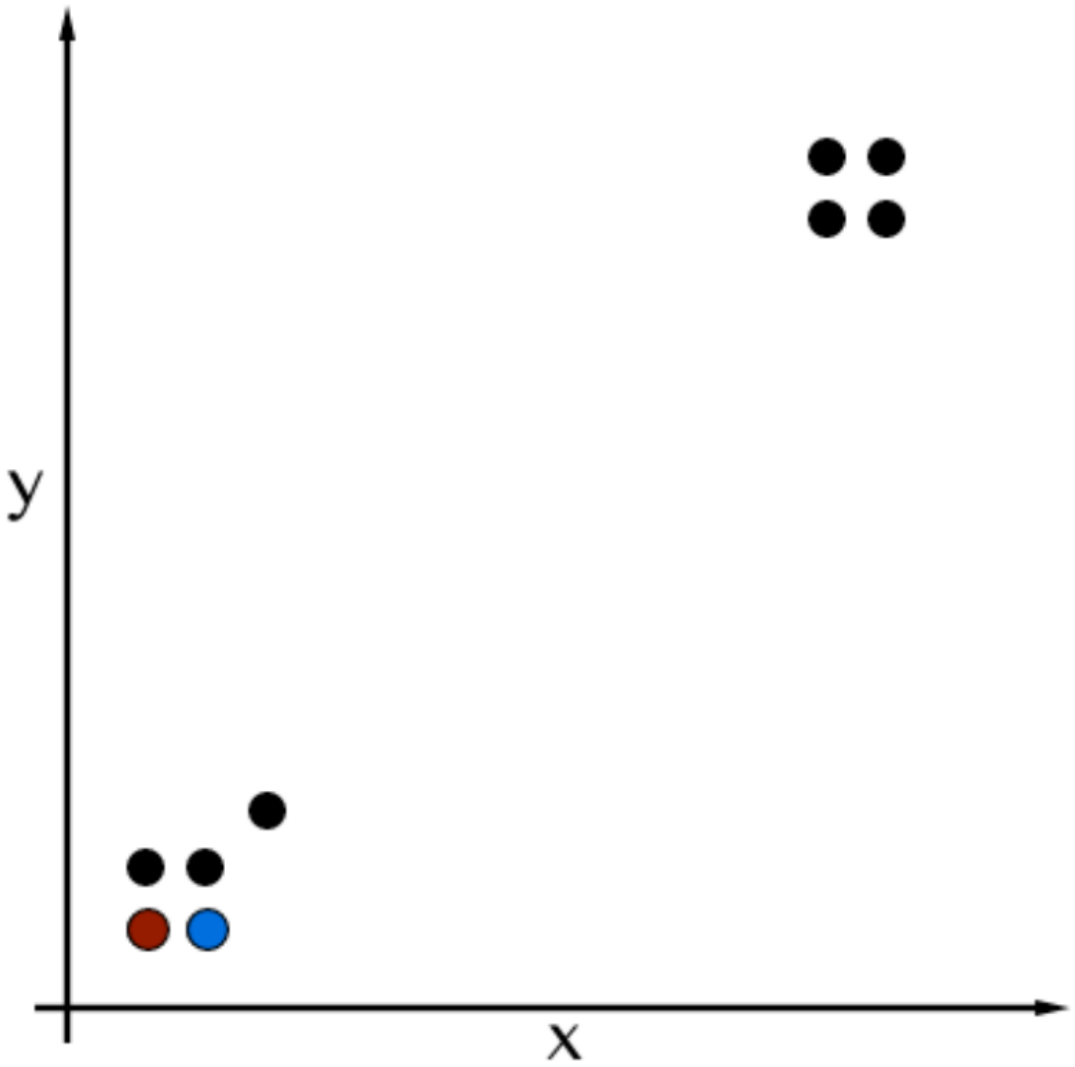
both Java code and command line options are options for executing

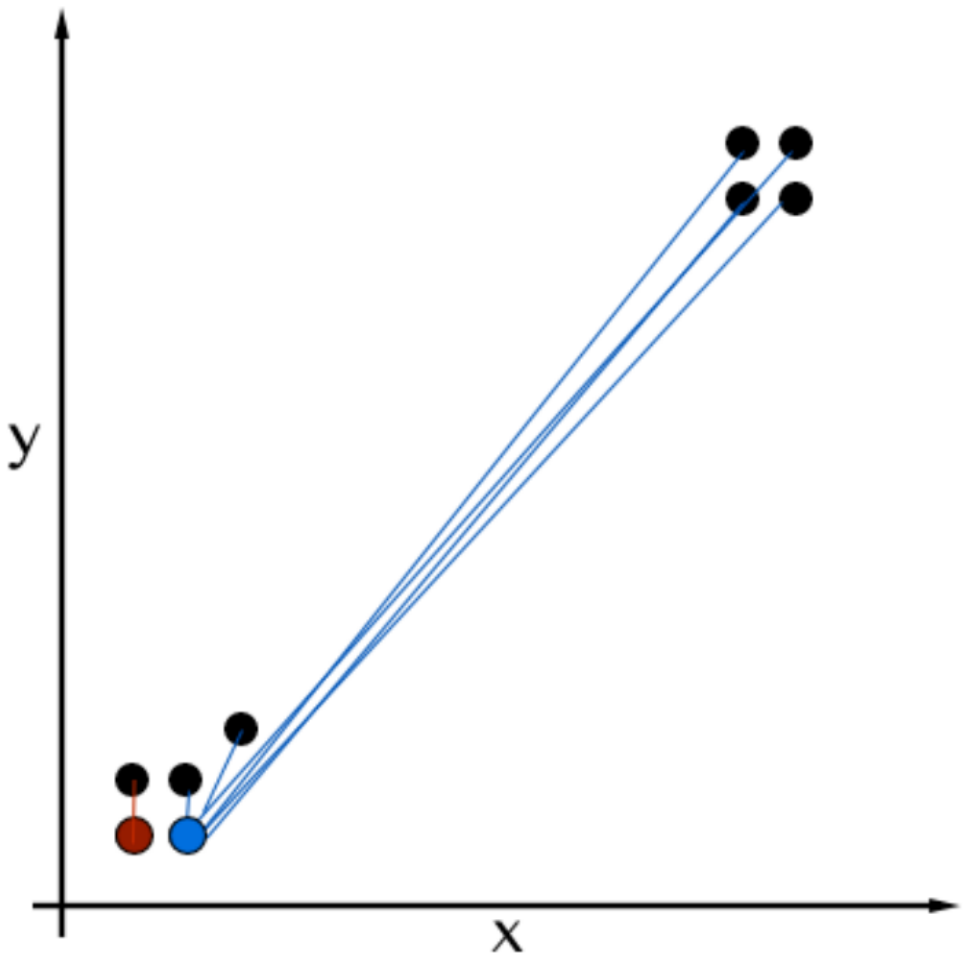
The K in KMeans

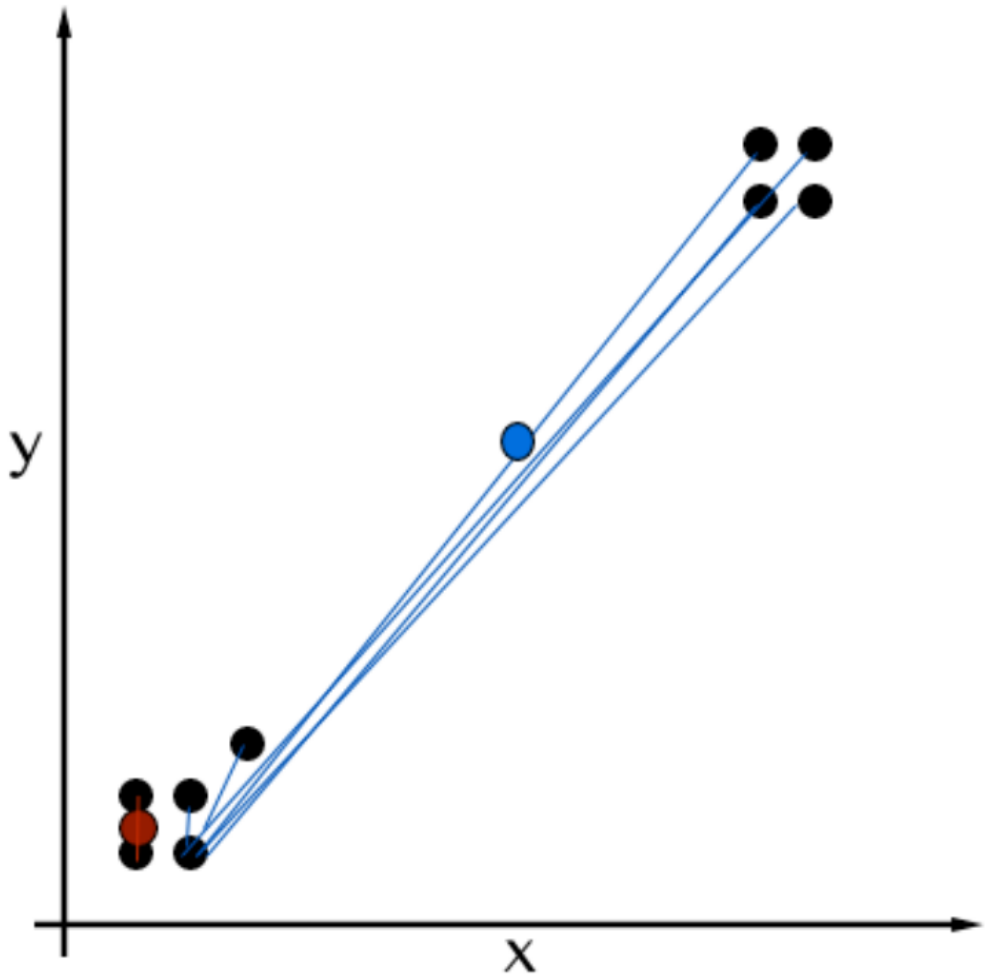
You must provide a SequenceFile of initial centroids

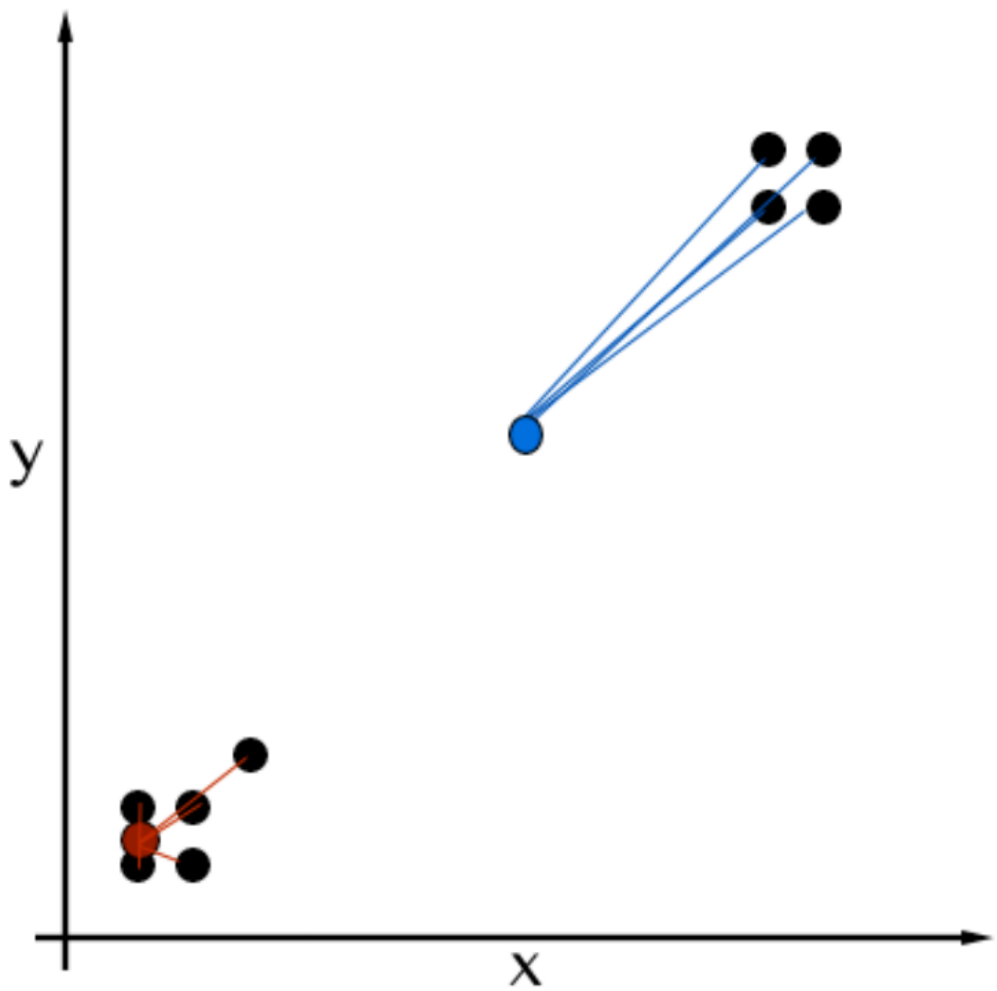
or you can specify for them to be selected at random with -k 20

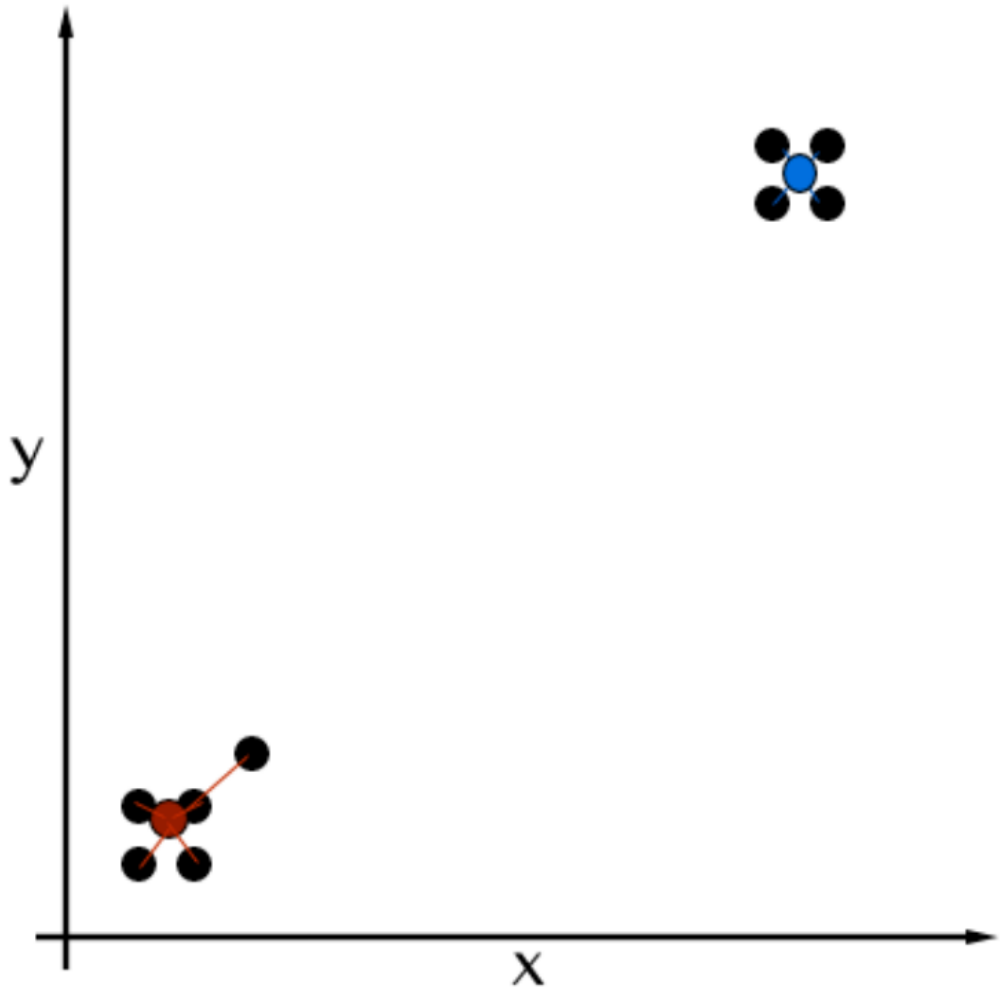

y
x


y
x


y
x


y
x


y
x


y
x

Clustering

ClusterDump

Clustering Algorithms - Canopy

"Canopy Generation" - just calculate the starting k centroids for a KMeans run

"Canopy Clustering" - settle with what it gives and assign vectors to canopy groups

Refining Your Clustering

1) Measure cluster distance

2) Weight vector features (normalize)

3) Improve document vector generation

TokenStreams for

LowerCaseFilter, LengthFilter, StopFilter

4) Write a custom distance measure

Recommenders

Hotel Rwanda

Because you enjoyed:

Bend It Like Beckham

Little Miss Sunshine

Add

Not Interested

requently Bought Together

Customers buy this item with Barbie Superstar Doll by Mattel


+


Price For Both: $26.37

Add both to Cart Add both to Wish List

Show availability and shipping details

ustomers Who Bought This Item Also Bought

Recommenders

User-based - What do *similar* people like?

Item-based - What items are liked by people who liked this?

Content-based - What items have the same features as this item?

Recommenders

Input

100,701,3

100,702,9

101,701,1

101,709,10

...

Recommenders - Performance

PreferenceArray interface

GenericUserPreferenceArray

GenericItemPreferenceArray

Recommenders

User-based

```
for every other user w
   compute a similarity s between u and w
   retain the top users, ranked by similarity, as neighborhood n
for every item i that some user in n has a preference for,
        but that u has no preference for yet
   for every other user v in n that has a preference for i
      compute a similarity s between u and v
      incorporate v's preference for i, weighted by s,
            into a running average
```

Recommenders

User-based input

- **DataModel**
- **User-User Similarity Metric**
  PearsonCorrelationSimilarity
- **UserNeighborhood definition**
  NearestNUserNeighborhood
- **Recommender engine**
  GenericUserBasedRecommender

User-based Similarity Metrics

- Pearson Correlation
- Euclidean
- Spearman correlation
- Tanimoto coefficient
- Log-likelihood

Neighborhood

Fixed size - provide # of users

Threshold based - provide distance and grab all users in range

Recommenders

Item-based

for every item i that u has no preference for yet
  for every item j that u has a preference for
    compute a similarity s between i and j
    add u's preference for j, weighted by s, to a running average
return the top items, ranked by weighted average

Recommenders

Item-based engines

GenericItemBasedRecommender

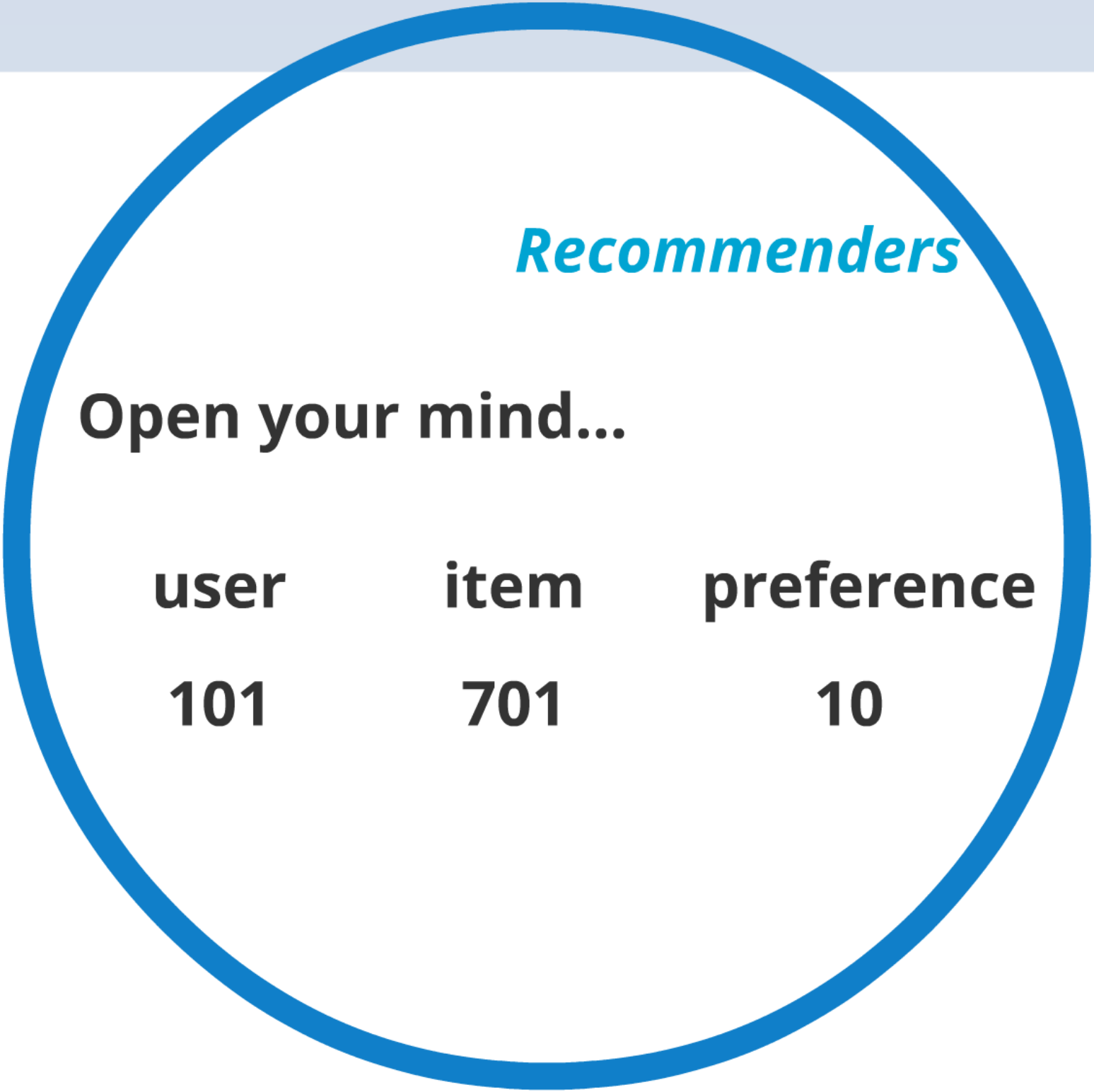
* no neighborhood

SlopeOneRecommender

* no neighborhood
* no similarity metric

Recommenders

What do you recommend to a brand new user?


Recommenders
Open your mind...
user
item
preference
101
701
10

Classification

Predicts answer to a non-open-ended question for each item based on its other known features.

spam/not spam

apple/pear

animal/mineral/vegetable

fraudulent purchase/not fraud

Classification

Build a "Model"

Training Data

color	size	type
purple	3	grape
green	1	grape
green	6	apple
red	5	apple

Test Data - hold some training data back (~20%) so we can test against something with answers

Classification

Predictor Variable Types

Continuous - float measurement

Categorical - non-measurement from fixed set (id, zip, enum)

Word-like - open-ended set of values

Text-like - sequence of word-like values

Classification

"Will the user buy this deal?"

Not just item variables are used!

age gender **type color** **price discount** **purchased?**

Classification Algorithms

SGD - Stochastic Gradient Descent

naive Bayes

Classification

Analyzing Results

80% test accuracy is great!

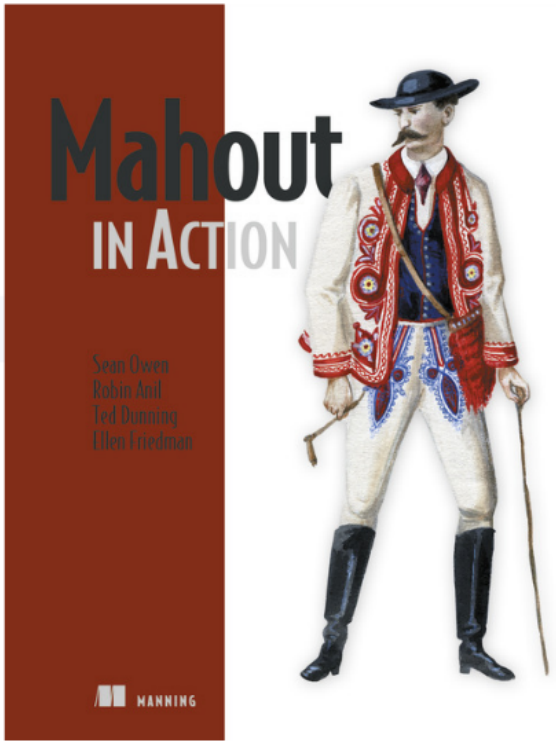
Target Leak - when the answer is in the question

Mahout Review

- Clustering
- Recommenders
- Classification

Resources

http://mahout.apache.org/



mpeabody@manifestcorp.com

Apache Mahout

Clustering - what you need

1) An algorithm
2) A definition for "similar"
3) A target stopping point

Clustering

Cluster these into groups of "similar" items.

Clustering

You're in a room full of books.

How do you organize them?

Clustering

Cluster these into groups of "similar" items.

Clustering - Getting data ready

1) Preprocess data
2) Use data to make Vectors
3) Save vectors in SequenceFile format

Clustering - Vectorizing Apples

Clustering

"How would you suppose you could turn these words into a vector?"

Clustering

Vector - List of data doubles!

x	y
3.5	5.9

shape	texture
"round"	"soft"

Mahout Intro

* Open Source
* Written in Java
* Now a top-level Apache project
* Sits on Hadoop... kind of

Mahout's Main Parts

Clustering >	Group similar things
Recommenders >	Predict a preference using other declared preferences
Classification >	Predict an answer for a record based on answers given for existing records

Who is Mahout For?

1) You have lots of data
2) You want to predict meaningful things with that data

"What music/movies could I suggest?"
"What do I put in front of my customers?"
"Was this transaction fraudulent?"
"What other articles could I list as similar to this article?"

What are these players doing?

Social/Communication
Twitter
Facebook
LinkedIn
eHarmony
StackOverflow

News Sites
Yahoo!Users

Media
iTunes
Netflix
YouTube

Commerce
Amazon
eBay
Zappos
Giant Eagle

Recommenders

Classification

Predicts answer to a nonsupervised question for each item based on its other known features.

spam/not spam

apple/pear

animal/mineral/vegetable

fraudulent purchase/not fraud

Classification

Build a "Model"

Classification

Predictor Variable Types

Classification

"Will the user buy this deal?"

Recommenders

User-Based - What do similar people like?
Item-Based - What items are liked by people who liked this?
Content-based - What items have the same features as this item?

Recommenders

Input

100,701,3
100,702,9
101,701,1
101,709,10
...

Recommenders - Performance

PreferenceArray interface

GenericUserPreferenceArray

GenericItemPreferenceArray