COLUMBIA BUSINESS SCHOOL

CS@CU

Data Science and Technology Entrepreneurship

Data to Clusters
Startup Story
Entrepreneurship Resources

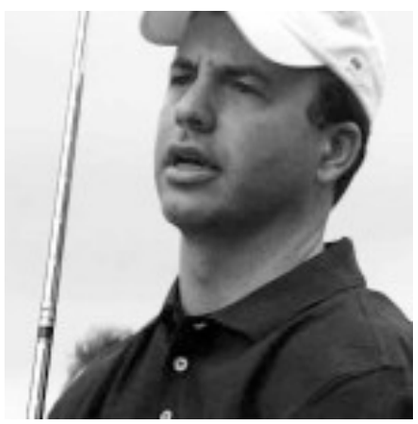
Week 12

Announcements

- Friday Open Office hours + Co-work with teams
  - Come and work together
  - Ask questions
  - Dates :
    - April 19 - 4:30 - 6:30 pm
    - April 26 - 4:30 - 6:30 pm
    - May 3 - 4:30 - 6:30 pm
  - Location
    - CS conference room
- Assignment 4 will be available on April 18th

Guest Lectures

- Paul Tumpowsky
  - Chairman InSITE
  - startup topics, inSITE startups

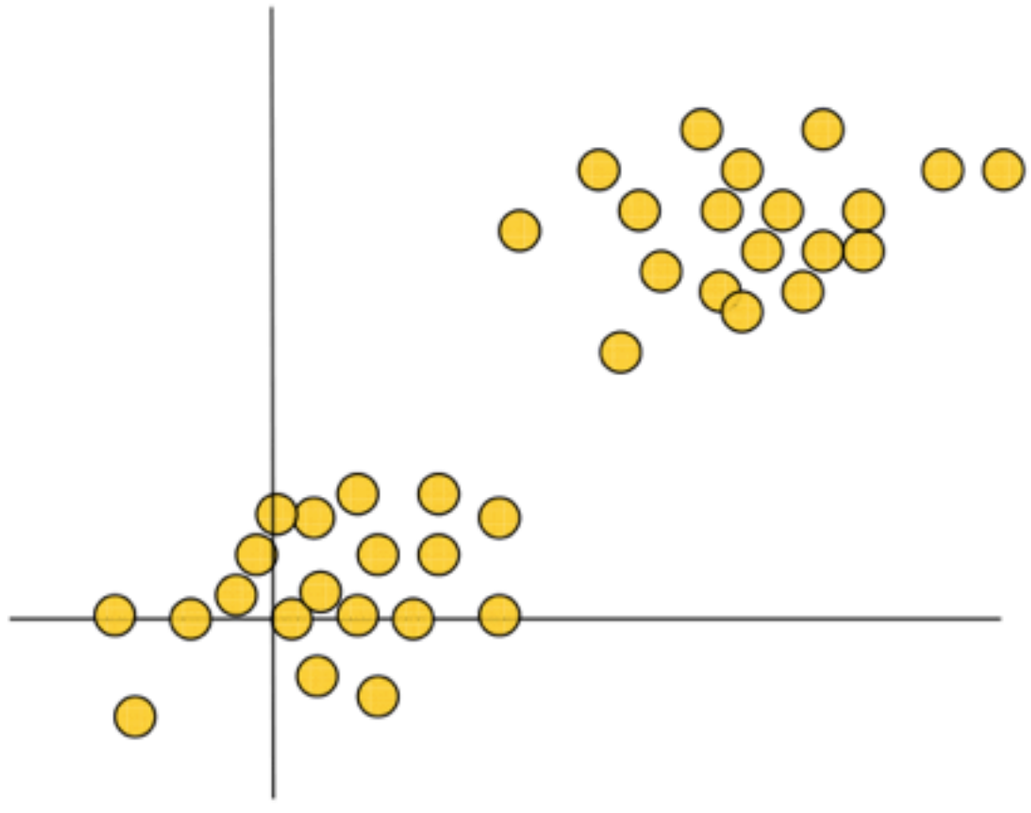


- Shari Coulter Ford
  - Executive Director, NYC TechConnect
  - Entrepreneurial resources

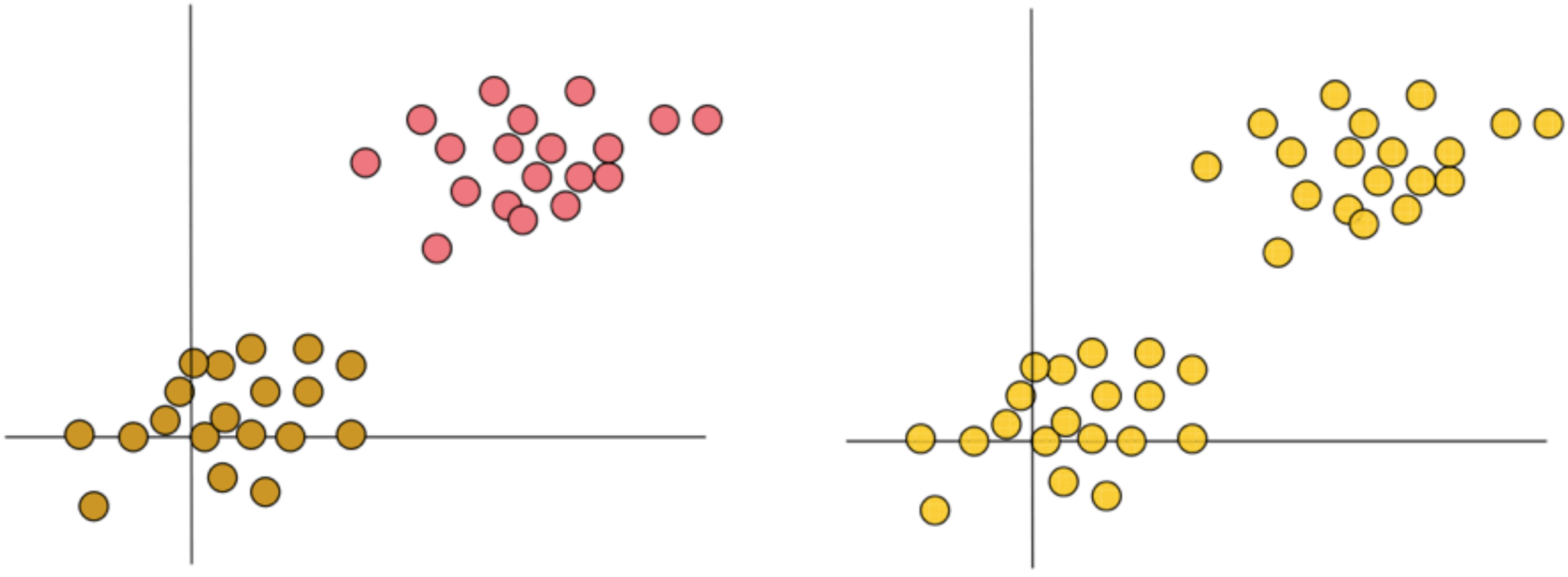
Topics for Today

- Data to Clusters
- Startups, inSITE, Marketing - Paul
- Entrepreneurship Resources - NY - Shari

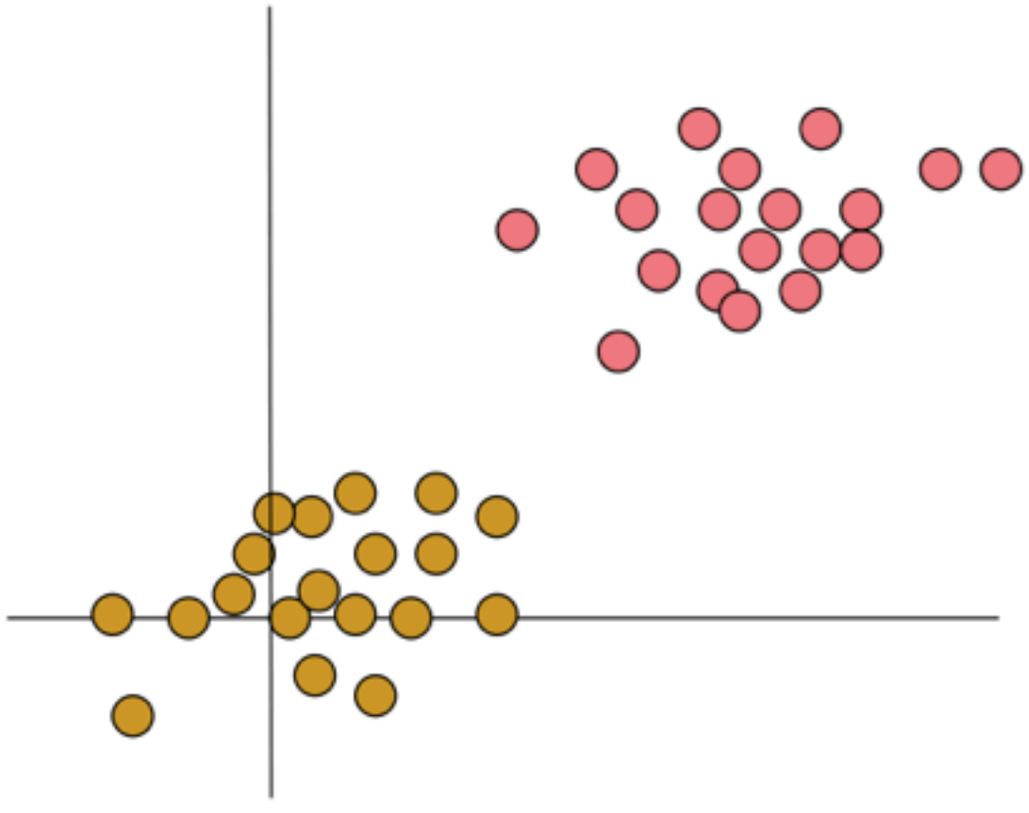
Data to Clusters



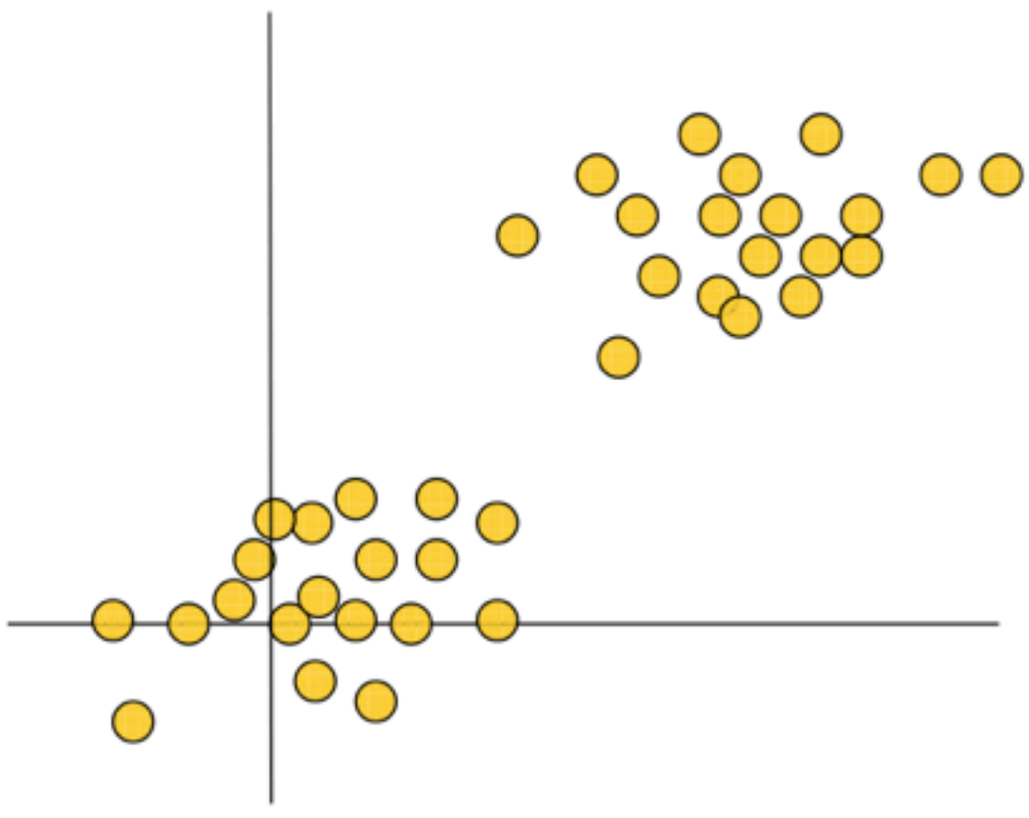
Data to Clusters



Data to Clusters



Supervised Training
of Classification Algorithm



Unsupervised Training
of Clustering Algorithm

Hypothetical Example

Nike wants to target new model of shoes to a particular customer segment?

Hypothetical Example

Assume

Users login online and fill out forms

Wants to target to a particular customer segment?

Hypothetical Example

Assume

Users login online and fill out forms

Wants to target to a particular customer segment?

Shoes Targeted to Diabetic Patients?

- Assume new innovation in shoes particularly useful for diabetic patients
- Can Nike target diabetic patients with online ads about diabetic patients
- or mail ads to their homes
- or send free samples

Hypothetical Example

- Nike wants to target new kind of shoes for diabetic customers

Hypothetical Example

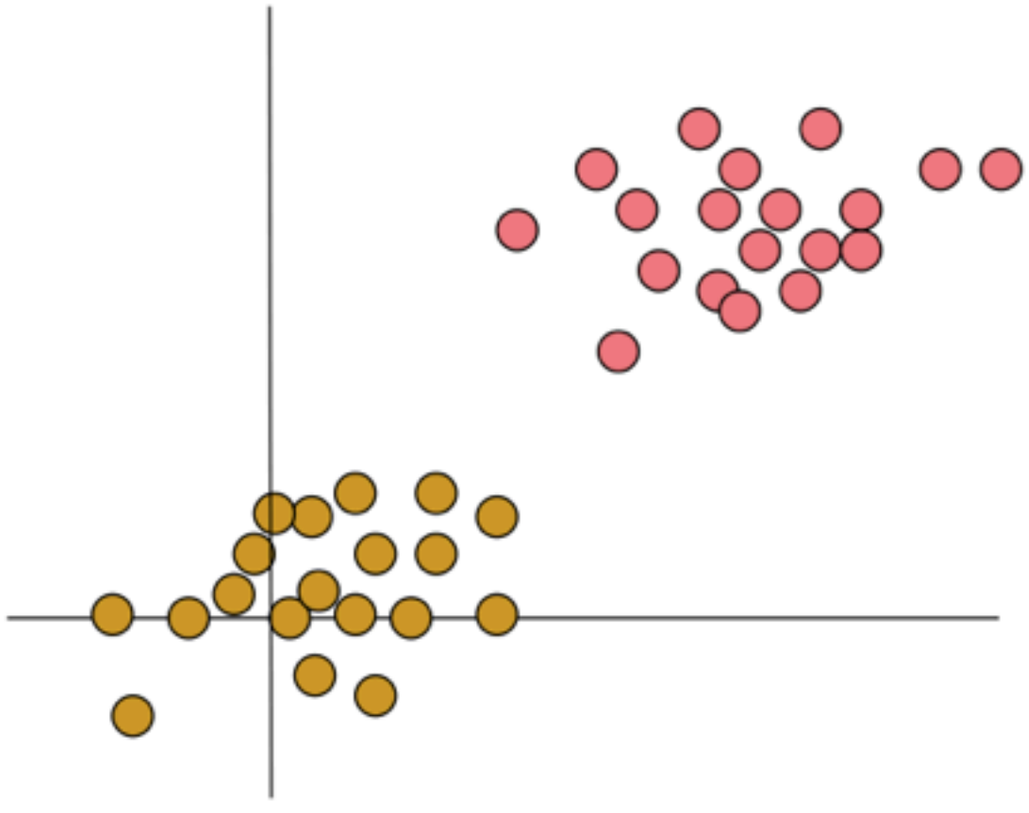
- Nike wants to target new kind of shoes for diabetic customers

```
%    2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
%    3. Diastolic blood pressure (mm Hg)
%    4. Triceps skin fold thickness (mm)
%    5. 2-Hour serum insulin (mu U/ml)
%    6. Body mass index (weight in kg/(height in m)^2)
%    7. Diabetes pedigree function
%    8. Age (years)
```

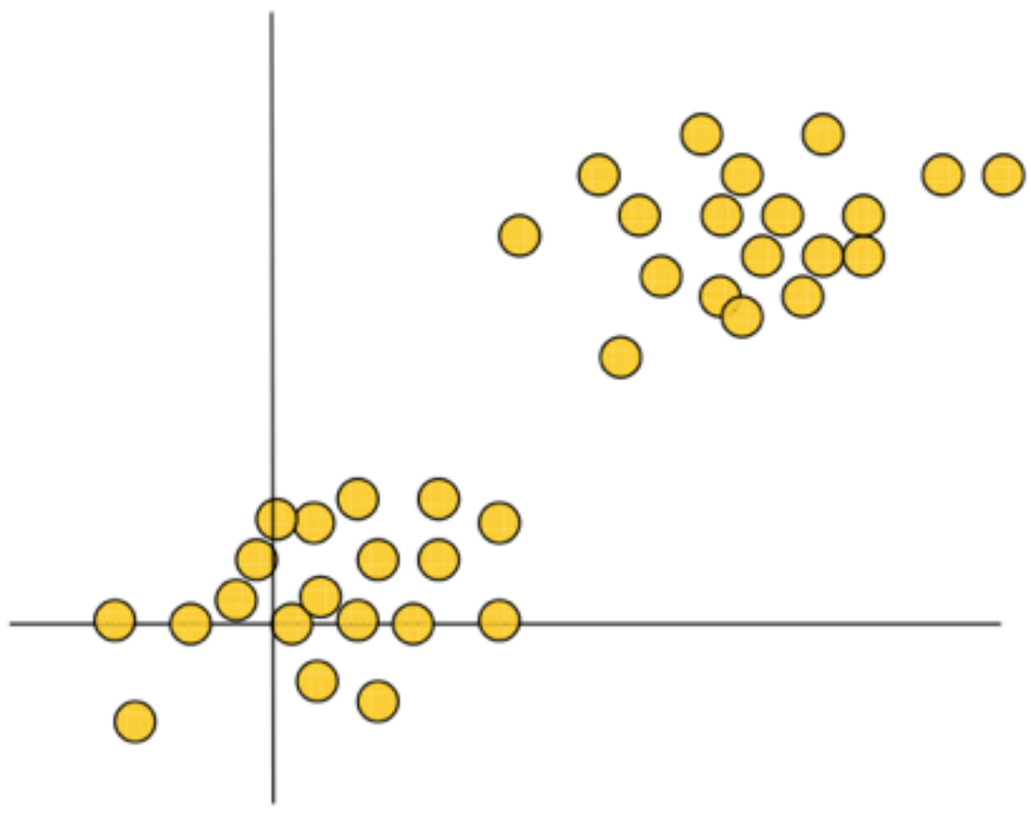
Data Sample

- 6,148,72,35,0,33.6,0.627,50,tested_positive
- 1,85,66,29,0,26.6,0.351,31,tested_negative
- 8,183,64,0,0,23.3,0.672,32,tested_positive
- 1,89,66,23,94,28.1,0.167,21,tested_negative
- 0,137,40,35,168,43.1,2.288,33,tested_positive
- 5,116,74,0,0,25.6,0.201,30,tested_negative
- 3,78,50,32,88,31,0.248,26,tested_positive
- 10,115,0,0,0,35.3,0.134,29,tested_negative

Supervised Classification vs. Clustering



Supervised Training
of Classification Algorithm



Unsupervised Training
of Clustering Algorithm

Data Sample

- 6,148,72,35,0,33.6,0.627,50,
- 1,85,66,29,0,26.6,0.351,31,
- 8,183,64,0,0,23.3,0.672,32,
- 1,89,66,23,94,28.1,0.167,21,
- 0,137,40,35,168,43.1,2.288,33,
- 5,116,74,0,0,25.6,0.201,30,
- 3,78,50,32,88,31,0.248,26,
- 10,115,0,0,0,35.3,0.134,29,

Who is Diabetic?

- Can Nike target diabetic patients with online ads about diabetic patients
- or mail ads to their homes
- or send free samples

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Clusterer

Choose **EM** -I 20 -N 2 -M 1.0E-6 -S 100

Cluster mode

- (•) Use training set
- ( ) Supplied test set — Set...
- ( ) Percentage split — % 66
- ( ) Classes to clusters evaluation

(Nom) class

- [x] Store clusters for visualization

Ignore attributes

Start | Stop

Result list (right-click for options)

11:59:49 - EM
12:02:06 - EM
12:05:43 - EM
12:06:41 - EM

Clusterer output

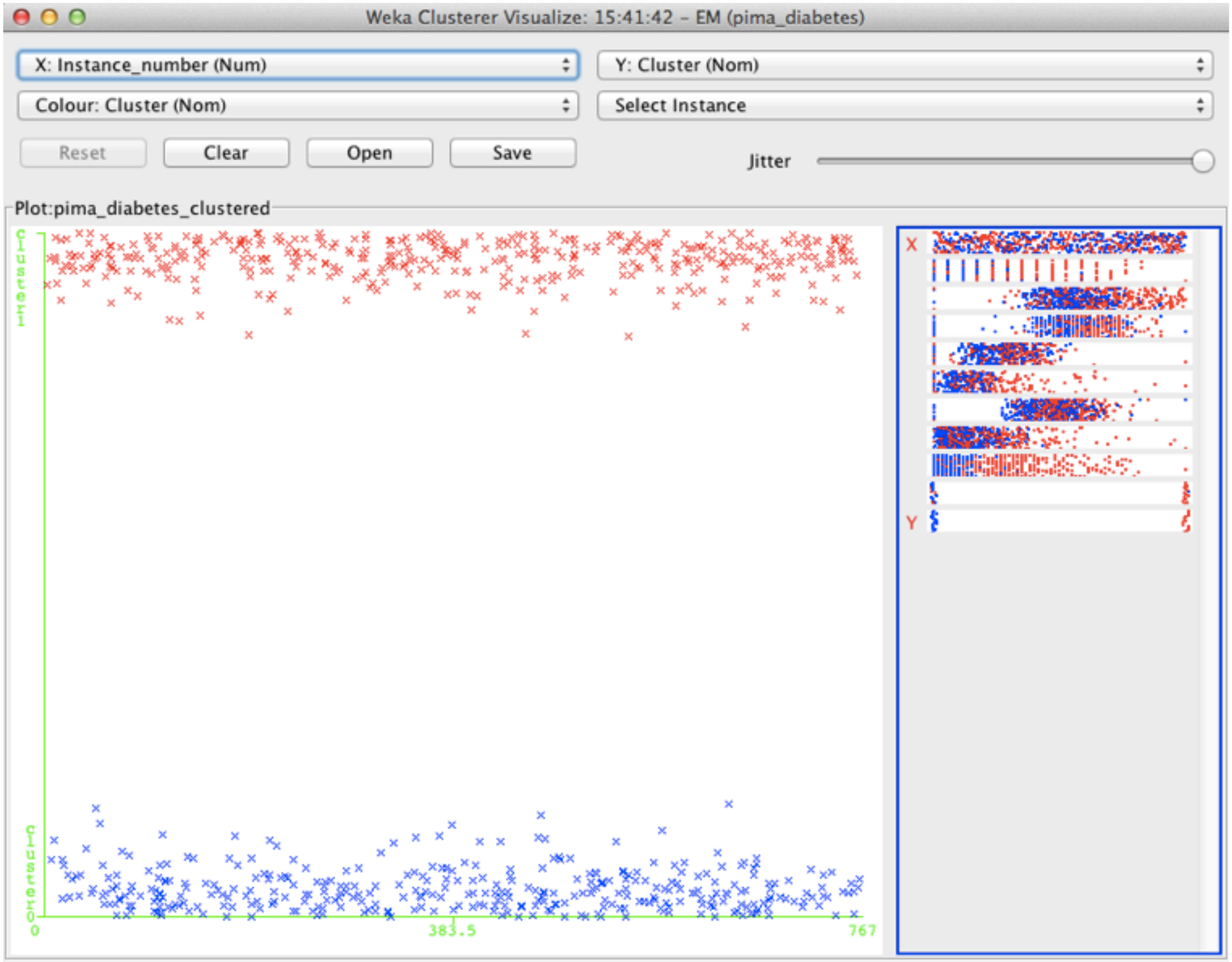
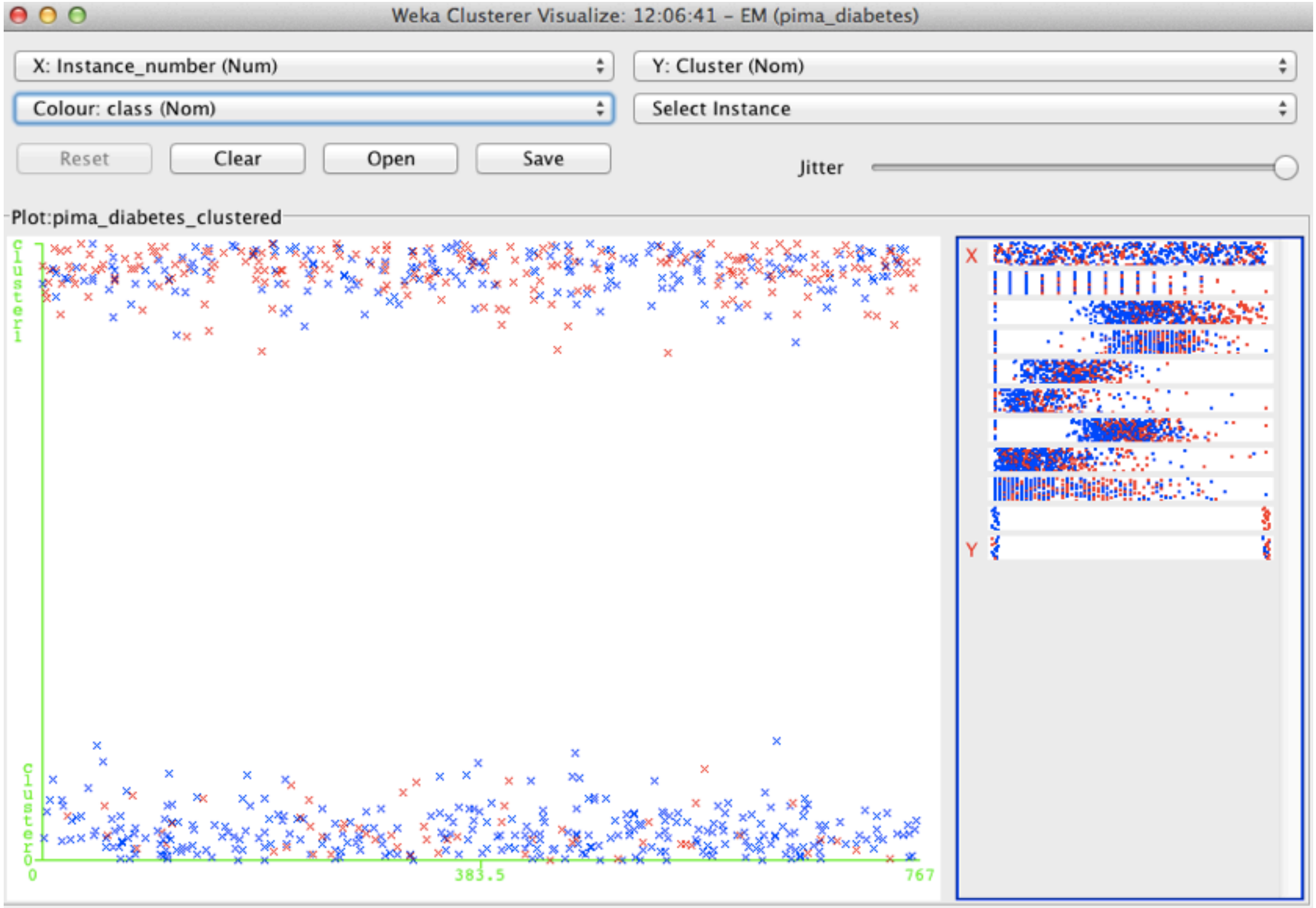
```
plas
  mean        107.6108 131.8415
  std. dev.    23.978  33.5139

pres
  mean         61.8736  75.0652
  std. dev.   22.4743  13.6982

skin
  mean         19.2371  21.6073
  std. dev.   14.4439  17.0033

insu
  mean         50.9599 103.5658
  std. dev.   59.0167 141.6826

mass
  mean         30.2672  33.4144
  std. dev.    8.3375   7.1747

pedi
  mean          0.3816   0.5463
  std. dev.     0.2046   0.3915

age
  mean         24.9518  40.0718
  std. dev.    3.5731  11.7534


Time taken to build model (full training data) : 3.38 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       357 ( 46%)
1       411 ( 54%)


Log likelihood: -29.12845
```

Status

OK

Log

Clustering Data



Clustering Data