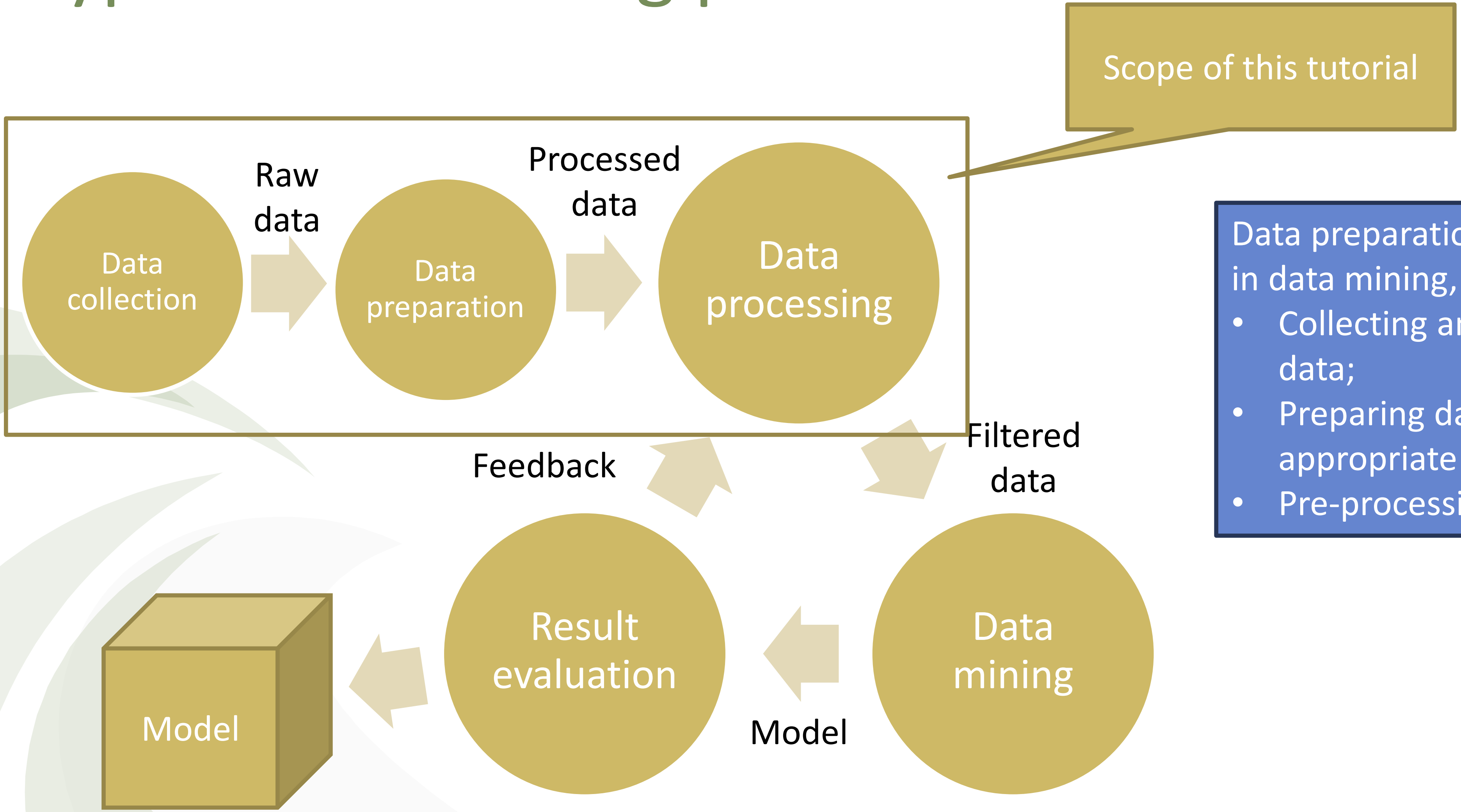
COMP7103 Data Mining

Tutorial 1

Data Preparation / Weka

Typical data mining process



Data preparation is an important step in data mining, it includes:

- Collecting and processing of raw data;
- Preparing data file in an appropriate format;
- Pre-processing of data.

Data preparation tools

An open-source data mining software in Java
https://www.cs.waikato.ac.nz/ml/weka/
(Download version 3.8 or above)

- **Weka**
  - Allow you to apply filters to existing data set
  - Not useful if raw data is in a format not readable by Weka
- **Spreadsheets (Excel, Google Sheets, etc.)**
  - Good for small data set, especially if you are familiar with the formulas
  - Explore potion of dataset before action
  - Cannot process large amount of data
- **Own program**
  - Require the most effort
  - Most customizable

In any cases, data needs to be converted into a format appropriate for the data mining exercise depending on the choice of tools.

Example dataset for this tutorial

- 150 Iris data is collected in a CSV file
  - https://archive.ics.uci.edu/ml/datasets/iris
- There are 4 attributes
  - ***Sepal*** **length** and **width**
  - ***Petal*** **length** and **width**
- Three species of iris

Available on Moodle:
`iris.csv`

Iris Setosa

Iris Versicolor

Iris Virginica

Image source: http://en.wikipedia.org/wiki/Iris_flower_data_set

Understanding the data

- Read carefully the information in the data source.
  - https://archive.ics.uci.edu/ml/datasets/iris

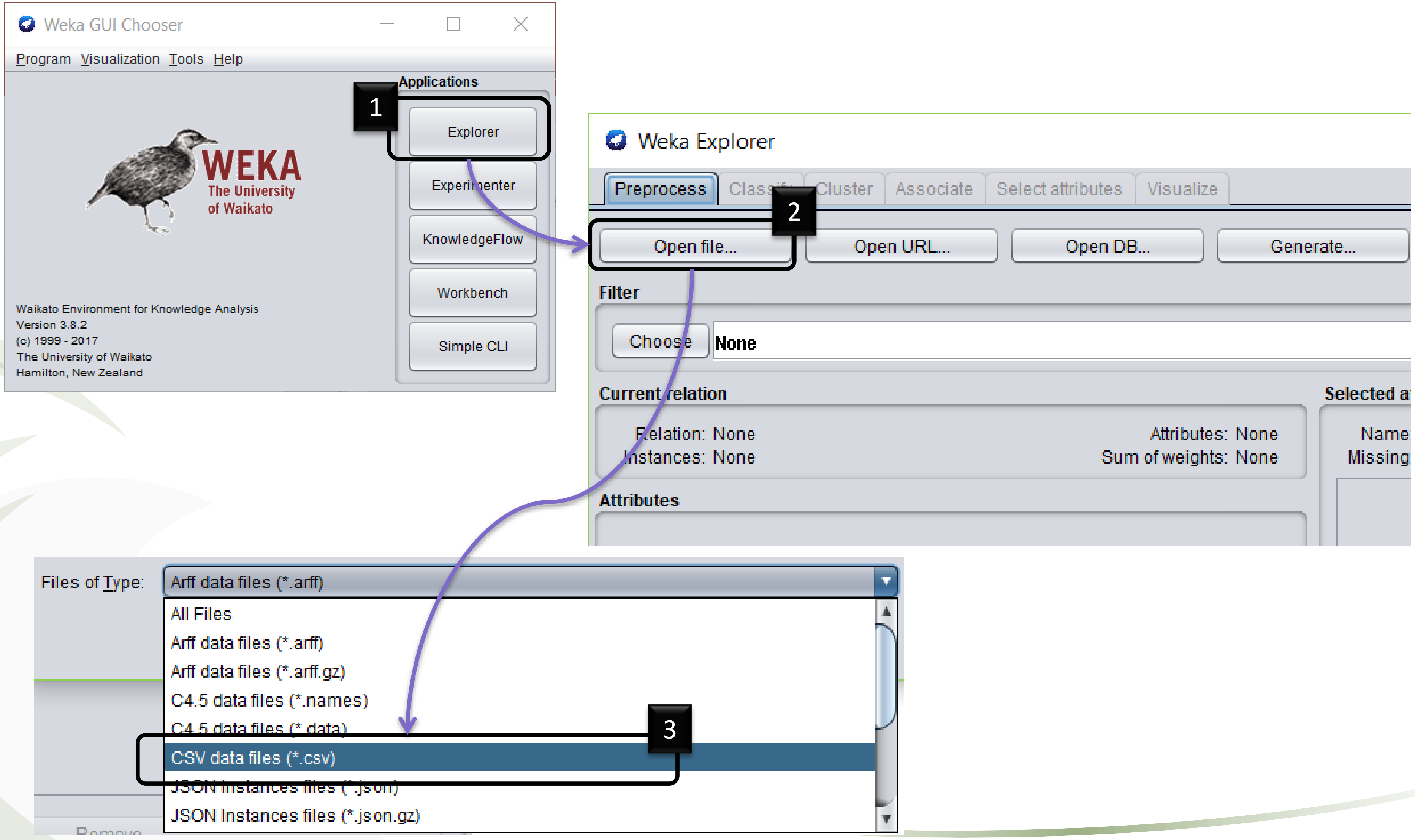
> “…The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.”

- Plan the preprocessing steps:
  - Fix data as described in the data source
  - Remove attributes that may not be useful
  - Save data for future use

Pre-processing CSV file in Weka

Data preparation

Weka Explorer – opening CSV file



Weka GUI Chooser

Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Weka Explorer

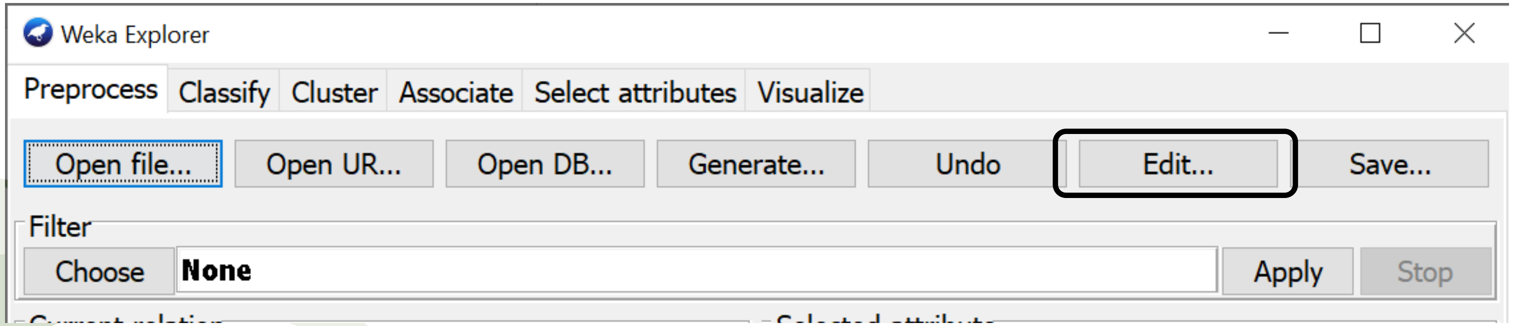
Relation: None
Instances: None
Attributes: None
Sum of weights: None

Files of Type: Arff data files (*.arff)

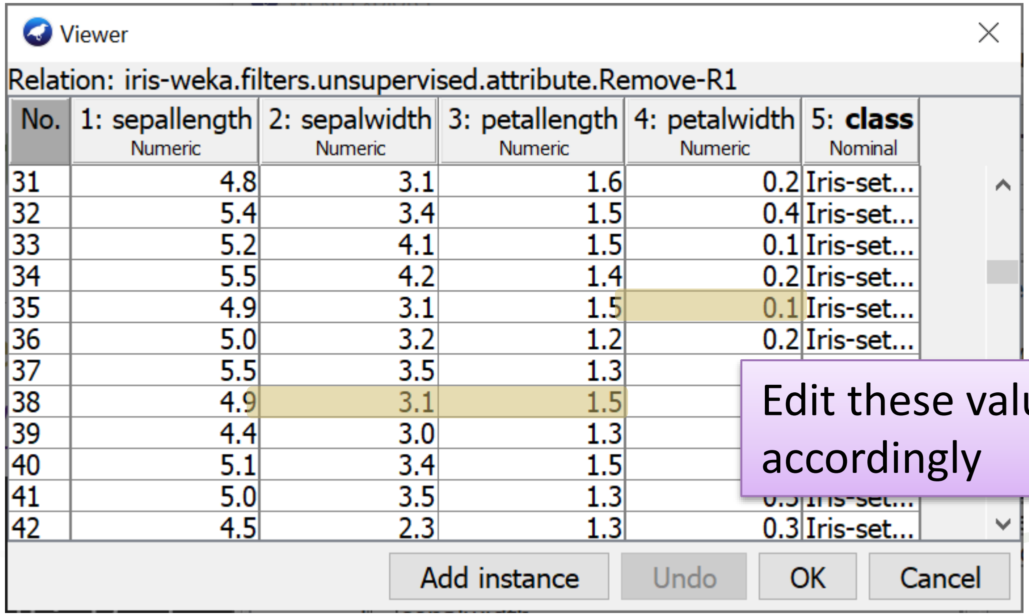
Edit data

“…The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.”

- Open the data in Weka and edit the data.



Filter: None

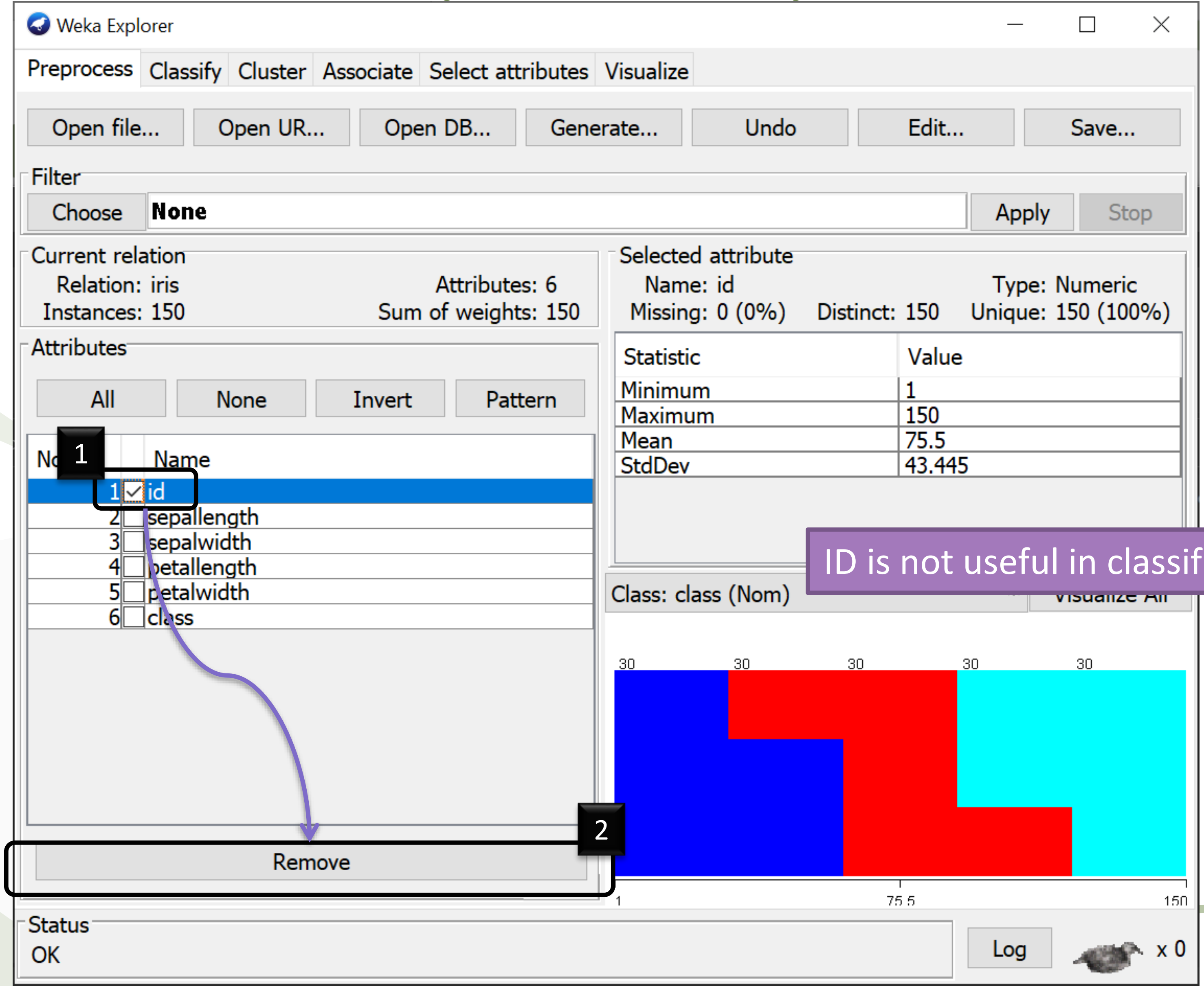


Relation: iris-weka.filters.unsupervised.attribute.Remove-R1

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal
31	4.8	3.1	1.6	0.2	Iris-set...
32	5.4	3.4	1.5	0.4	Iris-set...
33	5.2	4.1	1.5	0.1	Iris-set...
34	5.5	4.2	1.4	0.2	Iris-set...
35	4.9	3.1	1.5	0.1	Iris-set...
36	5.0	3.2	1.2	0.2	Iris-set...
37	5.5	3.5	1.3		
38	4.9	3.1	1.5		
39	4.4	3.0	1.3		
40	5.1	3.4	1.5		
41	5.0	3.5	1.3	0.3	Iris-set...
42	4.5	2.3	1.3	0.3	Iris-set...

Edit these values accordingly

Remove attribute (if needed)



Filter: None

Current relation

Relation: iris
Attributes: 6
Instances: 150
Sum of weights: 150

Selected attribute

Name: id
Type: Numeric
Missing: 0 (0%)
Distinct: 150
Unique: 150 (100%)

Statistic	Value
Minimum	1
Maximum	150
Mean	75.5
StdDev	43.445

Attributes

No.		Name
1	☑	id
2	□	sepallength
3	□	sepalwidth
4	□	petallength
5	□	petalwidth
6	□	class

Status

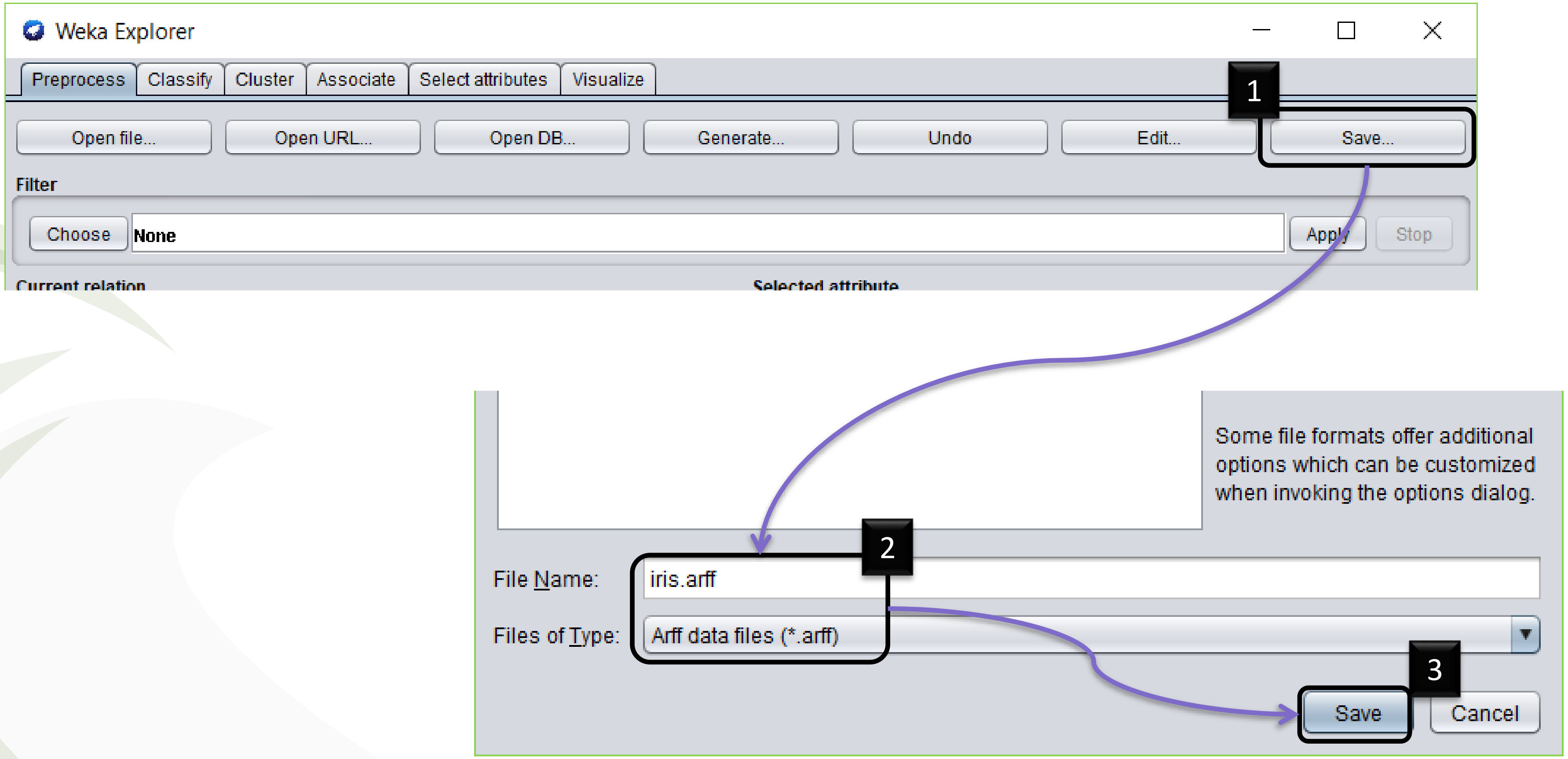
OK

ID is not useful in classification, we can remove it

Save data

ARFF is the default data file type in Weka

- Then we can save the data into an ARFF file



Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: iris.arff

Files of Type: Arff data files (*.arff)

ARFF file

- An ARFF file is a plain-text file with a specific format:
  - https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/
- For example, this is the iris dataset:

```
% comments
@RELATION relation_name

@ATTRIBUTE attribute_name attribute_type
@ATTRIBUTE attribute_name attribute_type
...

@DATA
comma-separated values
```

```
% Iris Plants Database
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth  NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth  NUMERIC
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iros-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
...
```

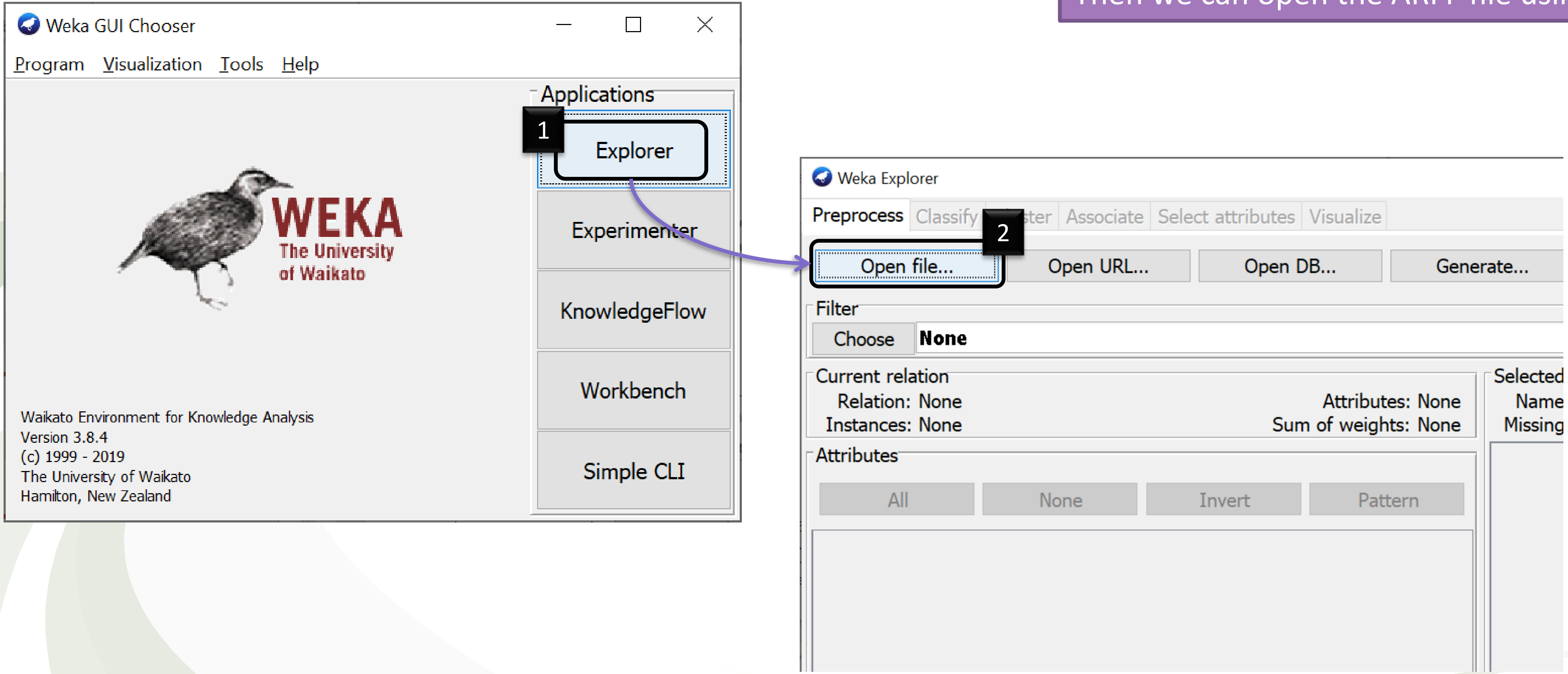
Data-type can be either:

- `numeric`
- `{list,of,normal,values}`
- `string`
- `date [<date-format>]`

After this line, the data is presented in CSV format

Weka Explorer – opening ARFF file

Then we can open the ARFF file using Weka



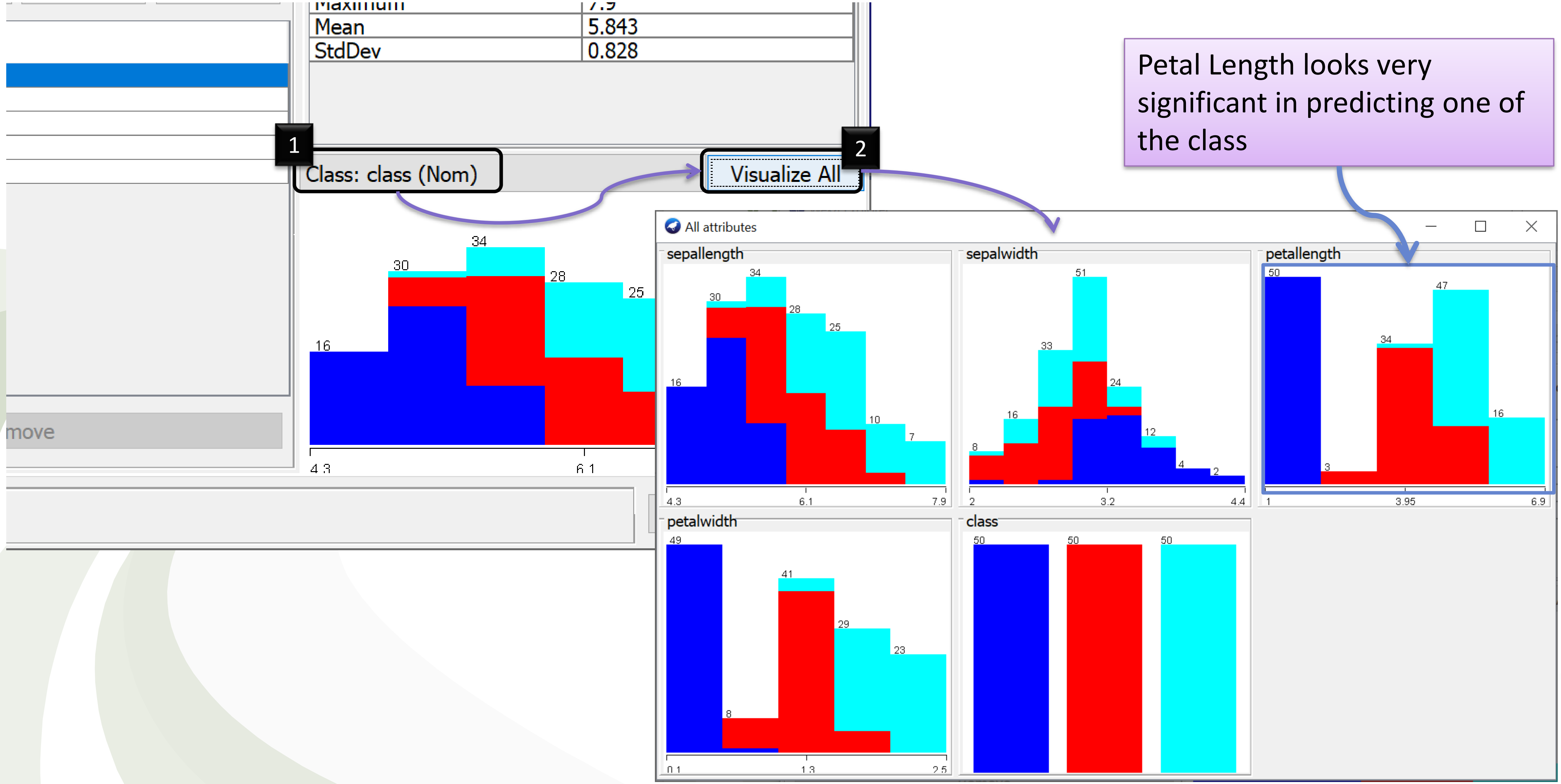
Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name
Missing

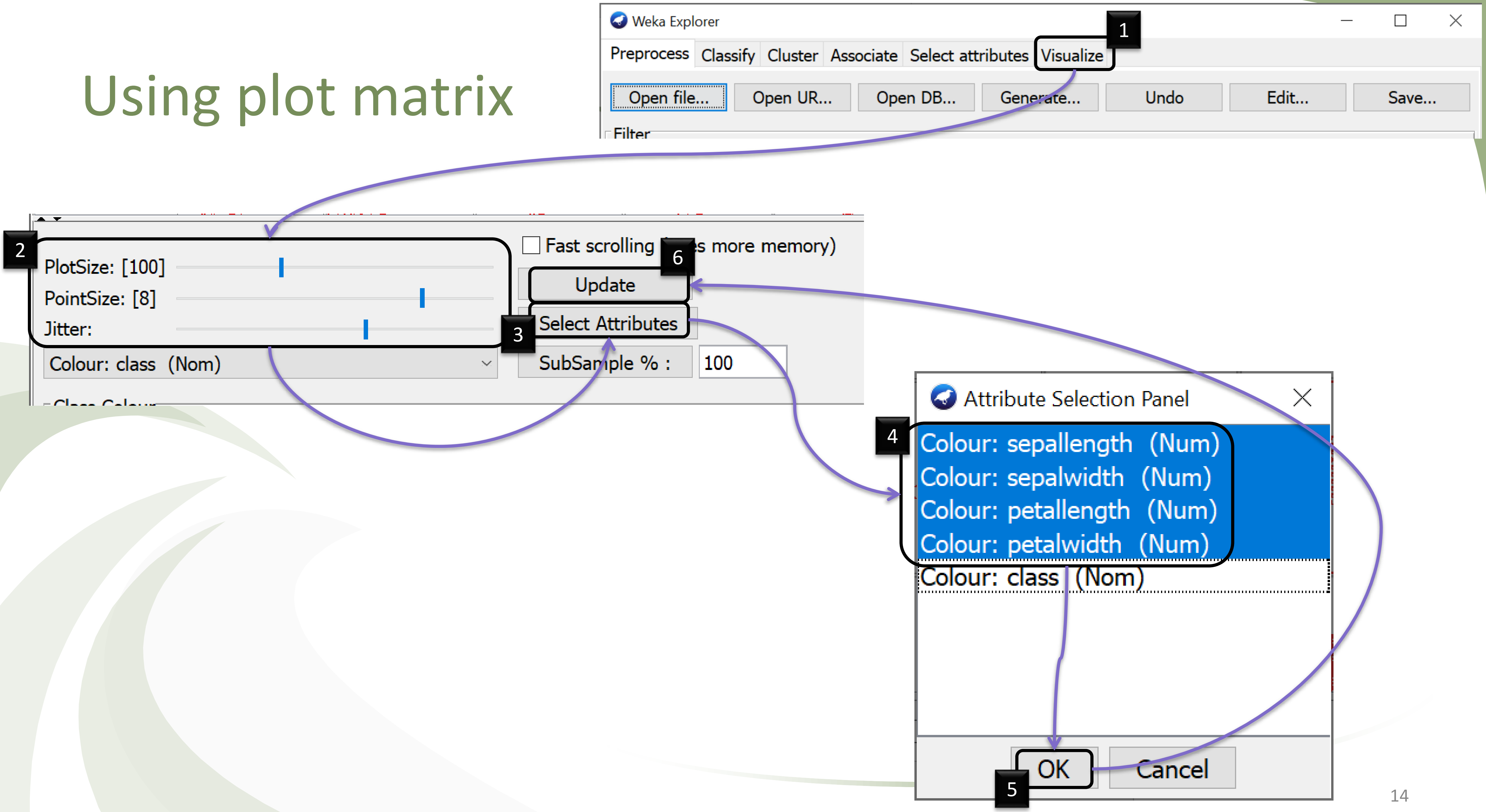
Visualize one attribute



Maximum	7.9
Mean	5.843
StdDev	0.828

Petal Length looks very significant in predicting one of the class

Using plot matrix



PlotSize: [100]

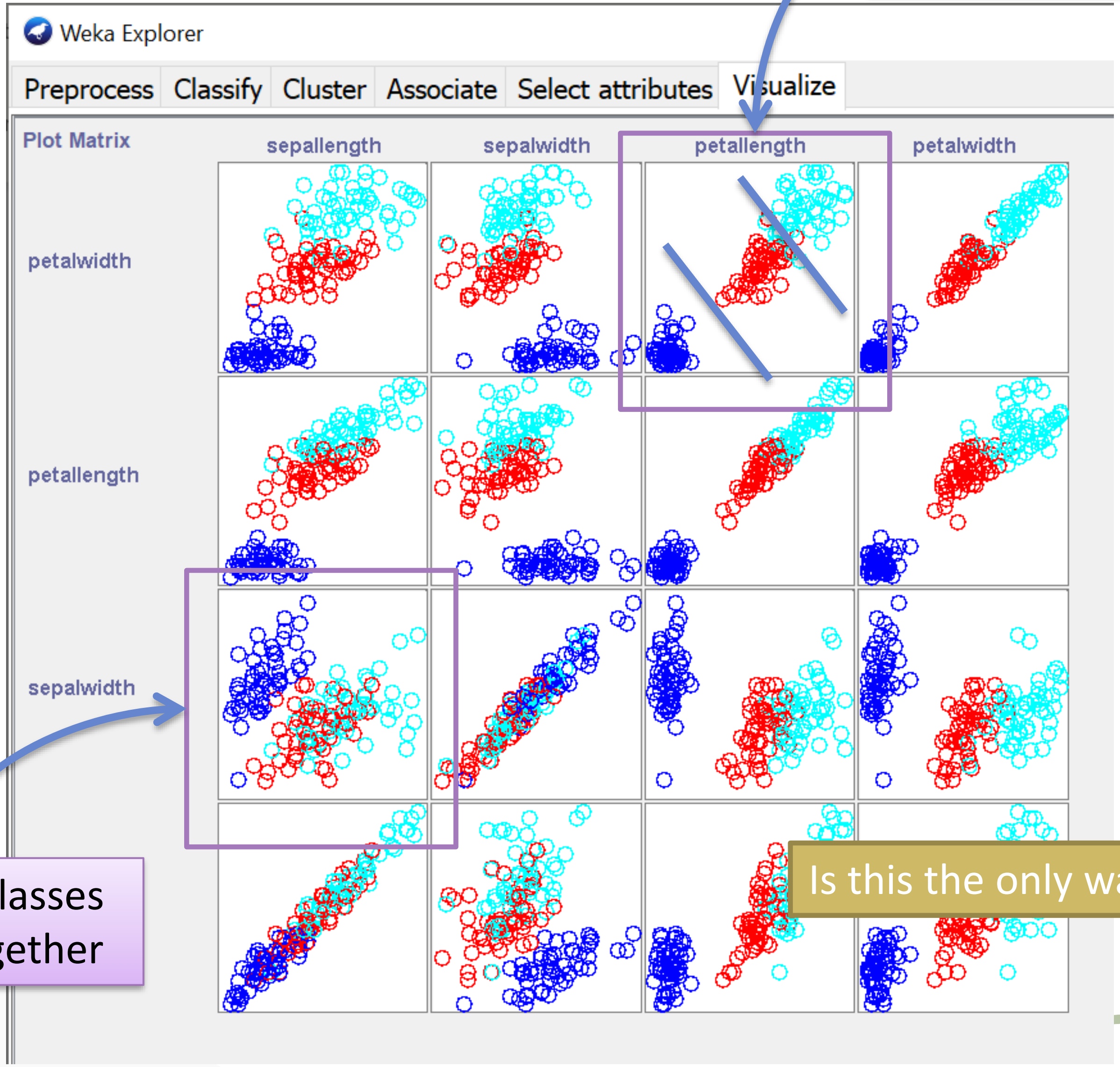
PointSize: [8]

Jitter:

Colour: class (Nom)

SubSample % : 100

Interpreting the matrix

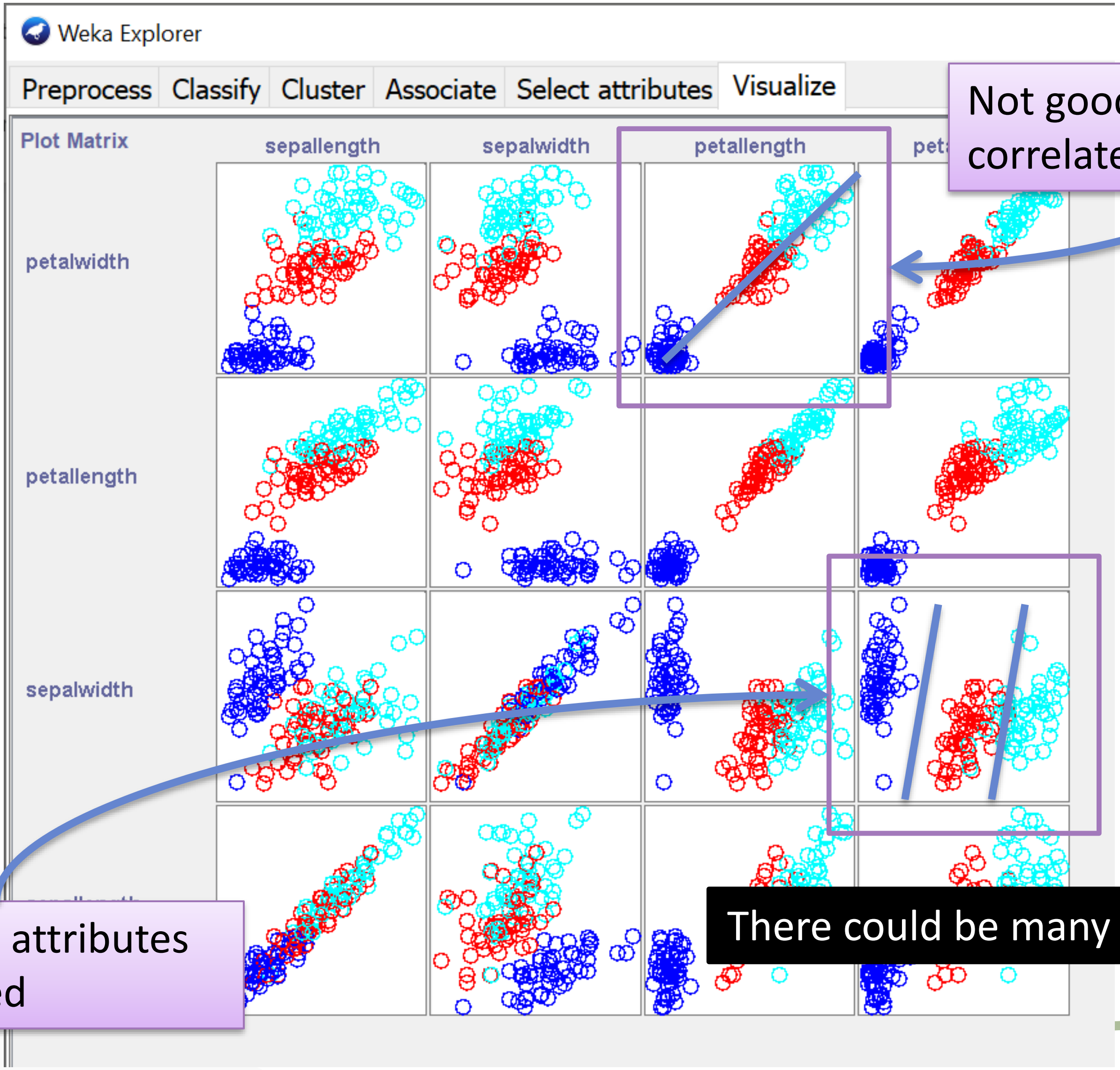


Looks good: classes are separated

Looks poor: classes are mixed together

Is this the only way to interpret the matrix?

Interpreting the matrix again



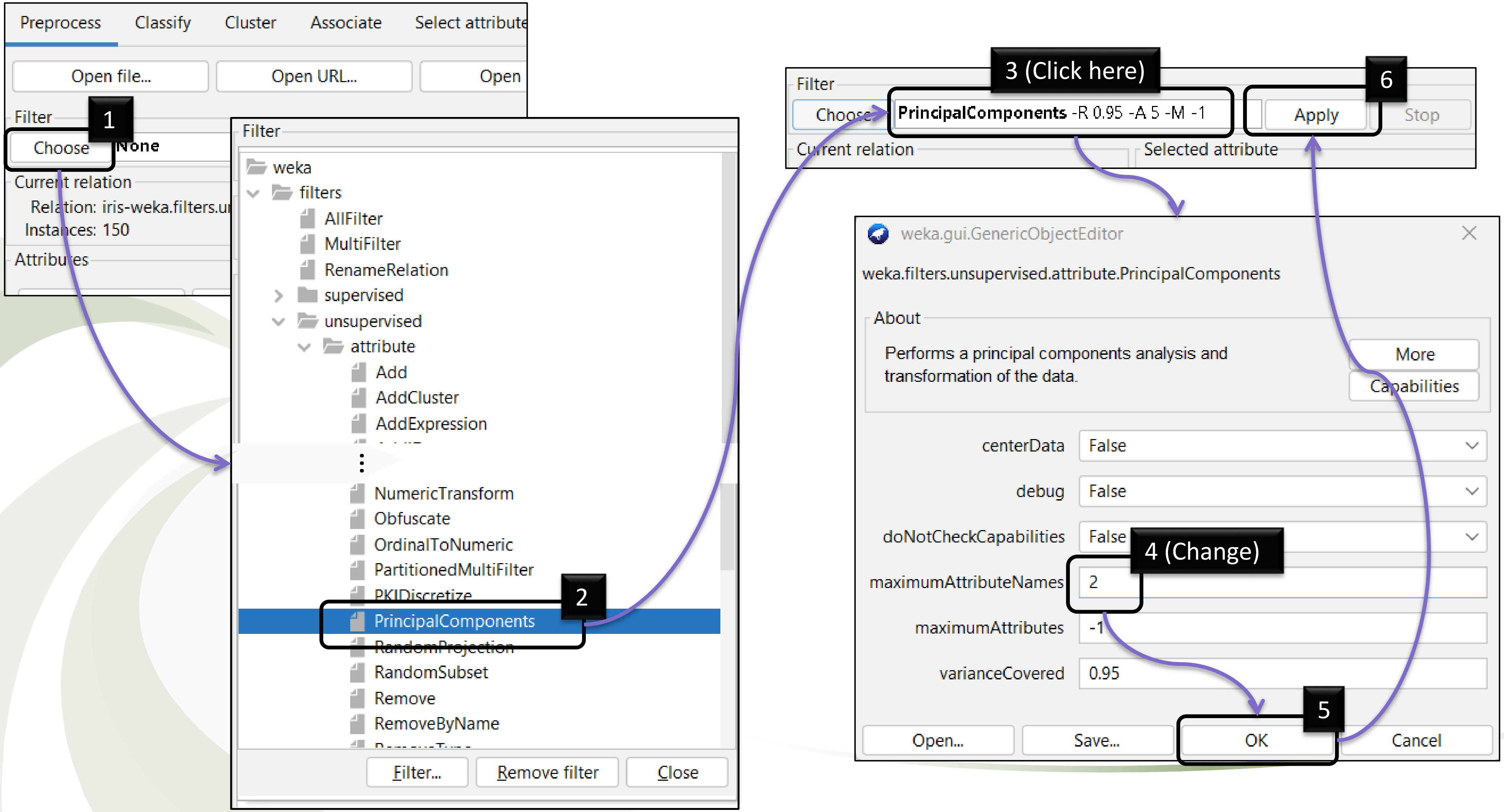
Not good, attributes are correlated

Seems good: attributes not correlated

There could be many ways to interpret the data!

Applying filter

- Applying filter under the “Pre-process” tab of Weka is one of the easiest way to preprocess the data.



Relation: iris-weka.filters.u

Instances: 150

PrincipalComponents -R 0.95 -A 5 -M -1

weka.filters.unsupervised.attribute.PrincipalComponents

Performs a principal components analysis and transformation of the data.

centerData: False

debug: False

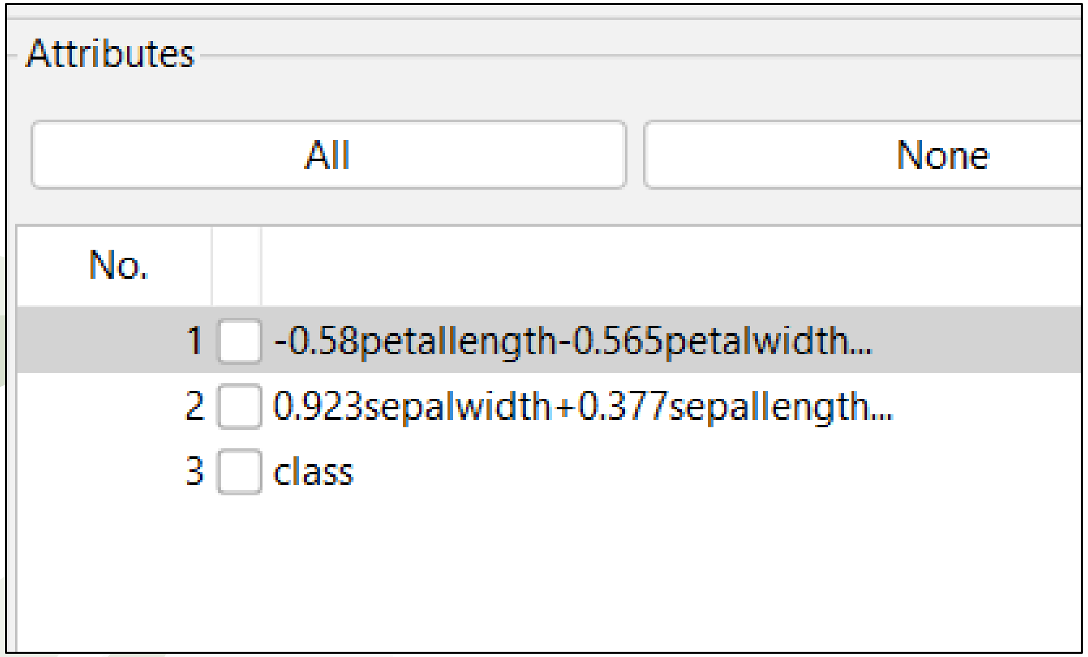
doNotCheckCapabilities: False

maximumAttributeNames: 2

maximumAttributes: -1

varianceCovered: 0.95

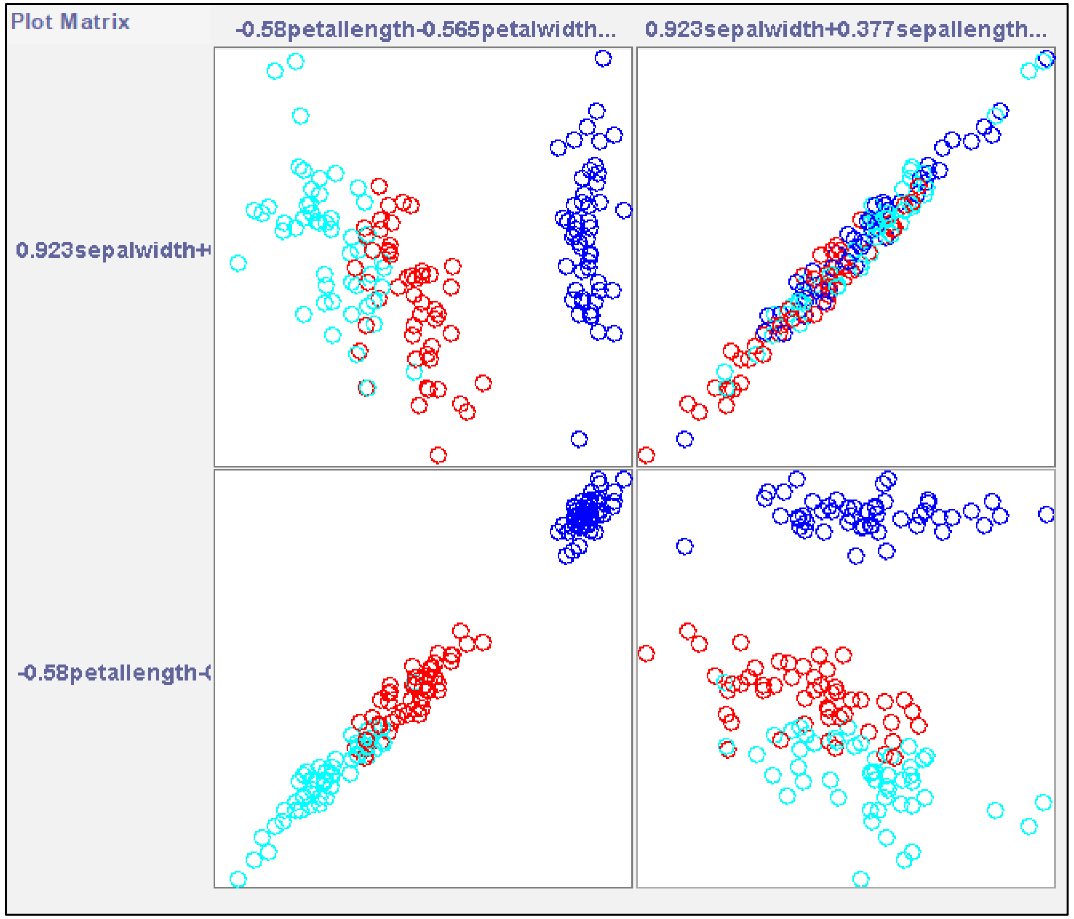
Result



Attributes

All None

No.		
1	□	-0.58petallength-0.565petalwidth...
2	□	0.923sepalwidth+0.377sepallength...
3	□	class



Data Denormalization

Pre-processing Raw data

Pre-processing raw data

- In this demonstration, we are using data from the UCI Machine Learning Repository.
  - https://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

```
I,4,"www.microsoft.com","created by getlog.pl"
T,1,"VRoot",0,0,"VRoot"
N,0,"0"
N,1,"1"
T,2,"Hide1",0,0,"Hide"
N,0,"0"
N,1,"1"
A,1287,1,"International AutoRoute","/autoroute"
A,1288,1,"library","/library"
A,1289,1,"Master Chef Product Information","/masterchef"
A,1297,1,"Central America","/centroam"
A,1215,1,"For Developers Only Info","/developer”
...
```

Available on Moodle:
`webdata.csv`

Understanding the raw data

- The raw data is in CSV form
- For each row, the first column indicate the type of the data:
  - ‘A’ for attribute (a page);
  - ‘C’ for case (a user);
  - ‘V’ for vote (a visit of a page);
  - All others are ignored in our case.

```
...
A,1008,1,"Free Downloads","/msdownload"
...
A,1046,1,"IE Support","/iesupport"
...
A,1034,1,"Internet Explorer","/ie"
...
C,10027,10027
V,1008,1
V,1046,1
V,1034,1
...
```

- So basically, we can create a set of relational data from the data set:

user_id
…
10027
…

Users

user_id	page_id
10027	1008
10027	1034
10027	1046

Visits

page_id	path
1008	/msdownload
1034	/ie
1046	/iesupport

Pages

What do we need?

Users
user_id

Visits
user_id, page_id

Pages
page_id, url

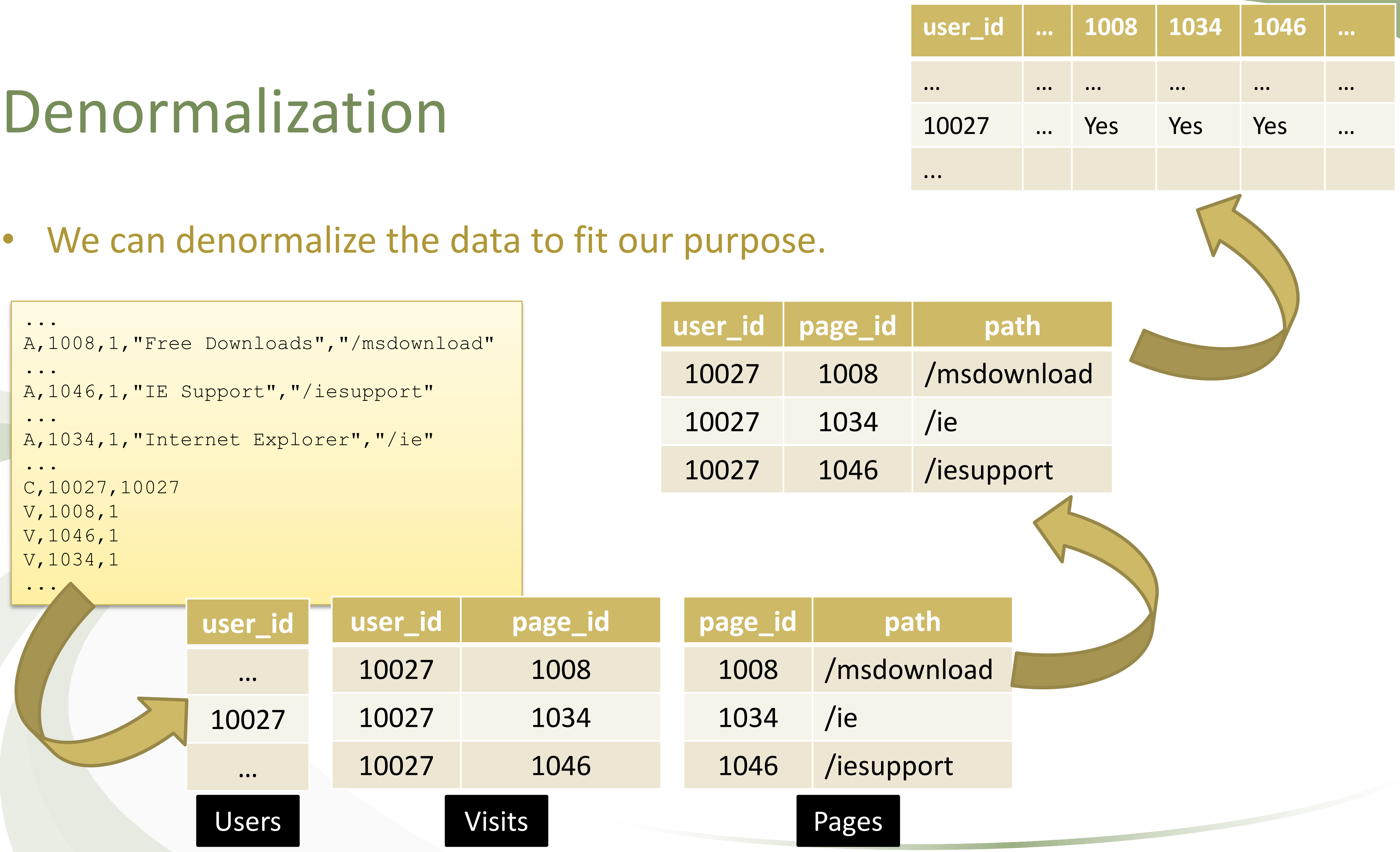
- What is the purpose of our data mining exercise?
- What is the data mining exercise that we are going to do?
- Suppose, we only focus on classifying whether a user will visit a page, we need a data set like this:

User_id	page 1	page 2	page 3	page 4	page 5	…
1	Yes	No	Yes	No	No	…
2	No	No	Yes	No	No	…
…						

What can we do?

Denormalization

- We can denormalize the data to fit our purpose.



```
...
A,1008,1,"Free Downloads","/msdownload"
...
A,1046,1,"IE Support","/iesupport"
...
A,1034,1,"Internet Explorer","/ie"
...
C,10027,10027
V,1008,1
V,1046,1
V,1034,1
...
```

user_id
…
10027
…

Users

user_id	page_id
10027	1008
10027	1034
10027	1046

Visits

page_id	path
1008	/msdownload
1034	/ie
1046	/iesupport

Pages

user_id	page_id	path
10027	1008	/msdownload
10027	1034	/ie
10027	1046	/iesupport

user_id	…	1008	1034	1046	…
…	…	…	…	…	…
10027	…	Yes	Yes	Yes	…
…					

Creating `.arff` file for Weka

- By implementing a custom program to process the raw data, a CSV file is generated:

```
activeplatform,activex,athome,corpinfo,education,exchange, …
N,N,Y,N,N,N,...
N,N,N,N,N,N,...
N,N,N,N,N,N,...
```

Available on Moodle:
`webdata.processed.csv`

- One can easily convert it to `.arff` file by converting the first row (the labels) into ARFF header.

```
@relation web_log_data

@attribute activeplatform {Y,N}
@attribute activex        {Y,N}
...

@data
N,N,Y,N,N,N,...
N,N,N,N,N,N,...
...
```

Auto-conversion by Weka

- If you open the CSV file in Weka and save it as ARFF file, you may need to check the nominal attributes for compatibility.

```
@relation web.data.processed

@attribute activeplatform {N,Y}
@attribute activex {N,Y}
@attribute athome {Y,N}
...

@data
N,N,Y,N,N,N,...
N,N,N,N,N,N,...
...
```

The two attributes will be considered incompatible in Weka