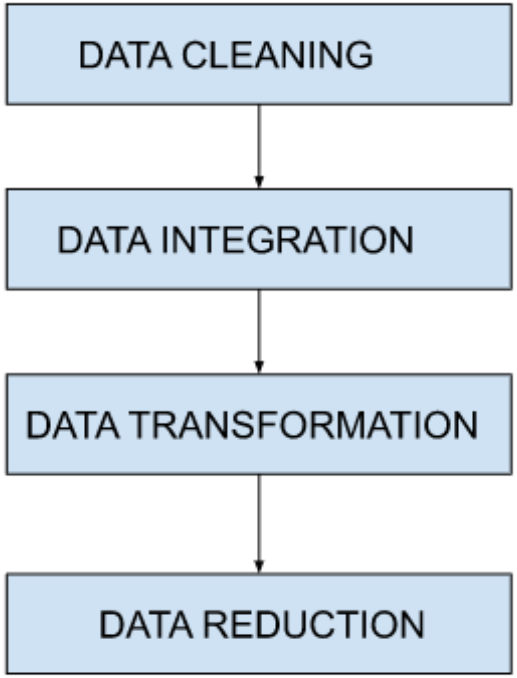
SEEM4630 Tutorial
Weka Demo on Data Preprocessing

Chaojun Wang
cj.wang@link.cuhk.edu.hk

Data Preprocessing

- No quality data, no quality mining results!
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data transformation**
  - Normalization, standardization, discretization
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results



Overview

- ❑ Weka introduction
- ❑ Weka demo on data preprocessing
  - ❑ Data loading
  - ❑ Data preprocessing
    - ❑ Data transformation
    - ❑ Data reduction
  - ❑ Data splitting
- ❑ Assignment 1 introduction

What is Weka?

- **W**aikato **E**nvironment for **K**nowledge **A**nalysis
  - It's a data mining tool developed by Department of Computer Science, University of Waikato, New Zealand.
  - Weka is also a bird found only on the islands of New Zealand.

What is in Weka?

- WEKA provides implementations of data mining algorithms that you can easily apply to your dataset.
  - It also includes a variety of tools for transforming datasets, such as the algorithms for discretization and sampling. You can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance—all without writing any program code at all.

Download and Install Weka

- WEKA is available from http://www.cs.waikato.ac.nz/ml/weka

Project **Software** **Book** **Courses** **Publications** **People** **Related**

Weka 3: Machine Learning Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

We have put together several **free online courses** that teach machine learning and data mining using Weka. The videos for the courses are available **on Youtube.**

Weka supports **deep learning**!

Getting started

- **Requirements**
- **Download**
- **Documentation**
- **FAQ**
- **Getting Help**

Further information

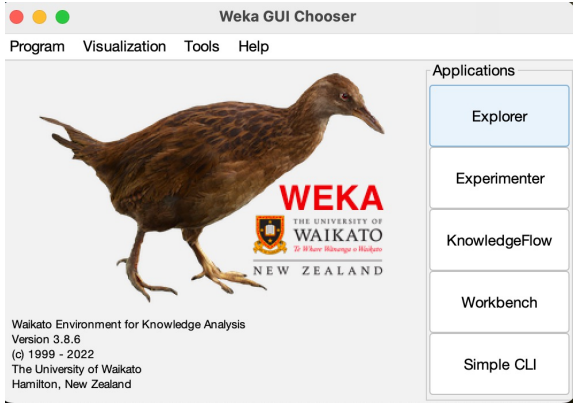
- **Citing Weka**
- **Datasets**
- **Related Projects**
- **Miscellaneous Code**
- **Other Literature**

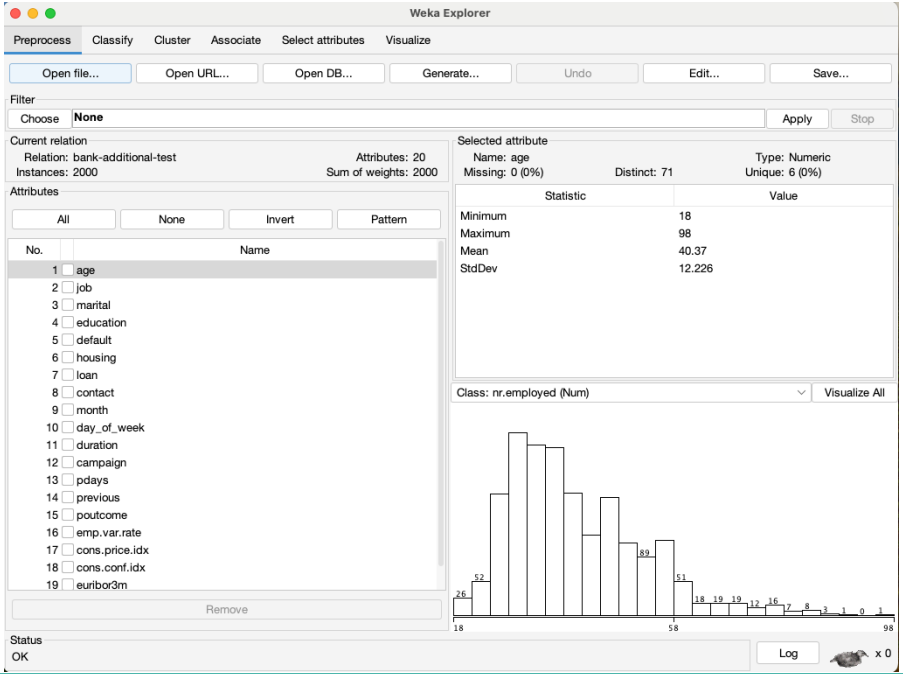
Developers

- **Development**
- **History**
- **Subversion**
- **Contributors**
- **Commercial licenses**

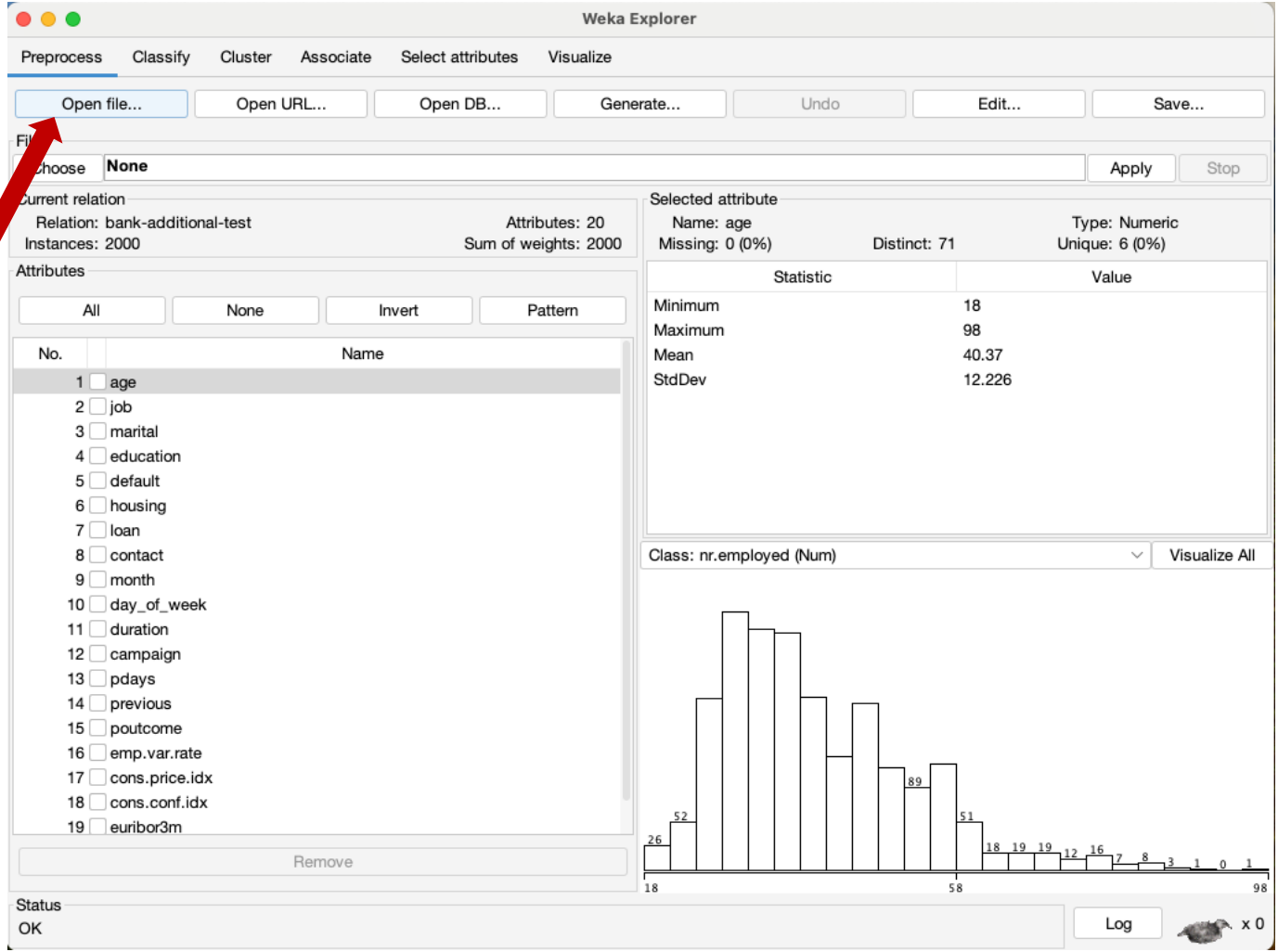
Weka Interface

- WEKA's main graphical user interface, the Explorer, gives access to all its facilities using menu selection and form filling.





Data Loading



Data Loading

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, JSON, etc.
- WEKA's native data storage method is ARFF format
  - ARFF= Attribute Relation File Format
  - ARFF files include two main parts:
    - Specification of the features
    - The actual data
  - When you specify a .csv file
    - it is automatically converted into ARFF format.

Arff data files (*.arff)
Arff data files (*.arff.gz)
C4.5 data files (*.names)
C4.5 data files (*.data)
CSV data files (*.csv)
JSON Instances files (*.json)
JSON Instances files (*.json.gz)
libsvm data files (*.libsvm)
Matlab ASCII files (*.m)
svm light data files (*.dat)
Binary serialized instances (*.bsi)
XRFF data files (*.xrff)
XRFF data files (*.xrff.gz)

Weka ARFF Example

This defines dataset name

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

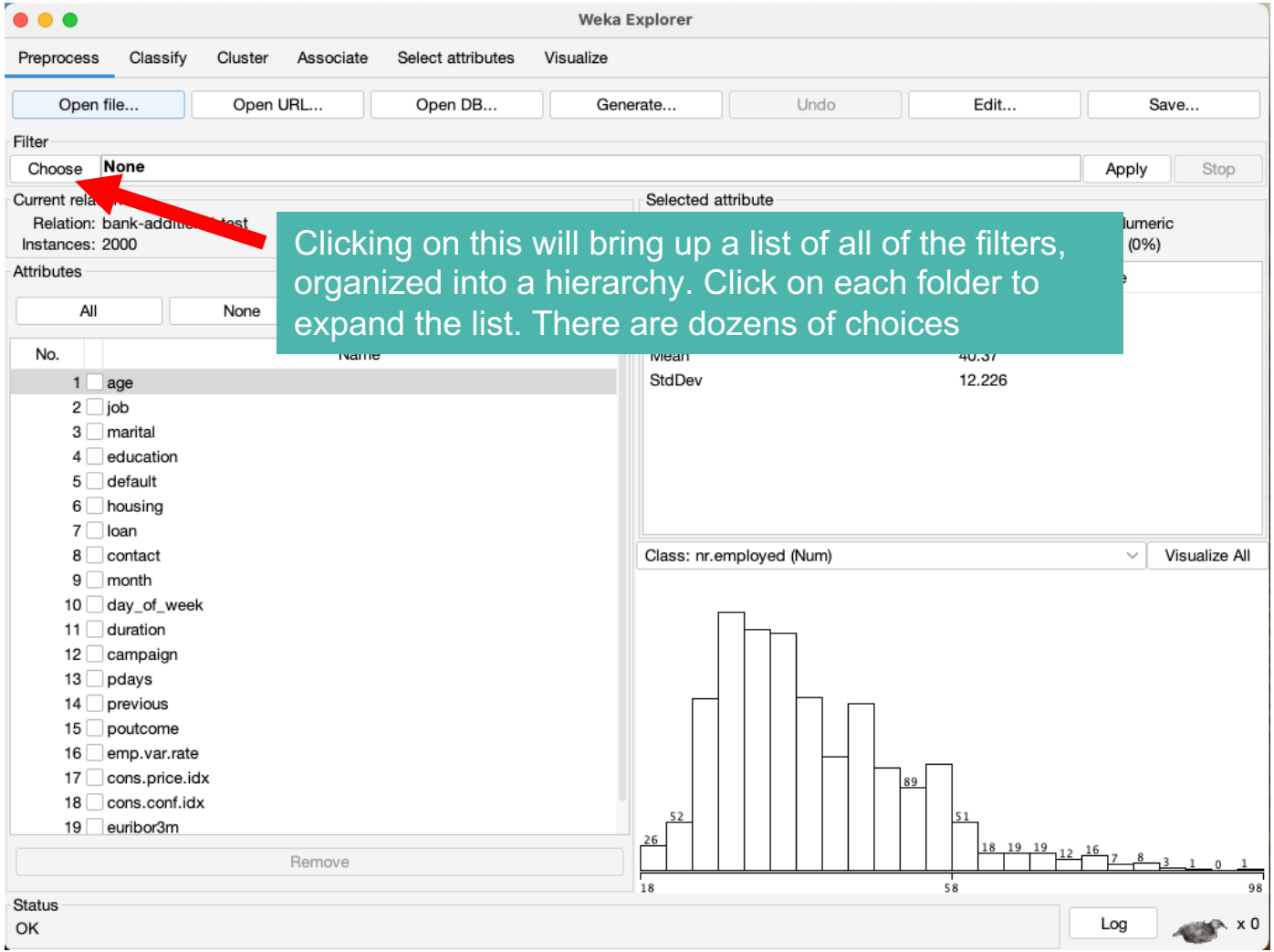
Nominal features must list values

Data Preprocessing

- Data pre-processing tools in WEKA are called **“filters”**
- Supervised vs Unsupervised
  - Supervised Filters: That can be applied but require user control or make use of the class information in some way. Such as rebalancing instances for a class.
  - Unsupervised Filters: That can be applied in an undirected manner. For example, discretize the numerical attributes or rescale all values in the attribution to the range 0 to 1.
  - **Supervised Filters “require a class attribute”, while unsupervised filters do not**

Data Preprocessing

- Data pre-processing tools in WEKA called “filters”
- Attribute vs Instance
  - Attribute Filters: Apply an operation on attributes or one attribute at a time.
  - Instance Filters: Apply an operation on instances or one instance at a time.



Clicking on this will bring up a list of all of the filters, organized into a hierarchy. Click on each folder to expand the list. There are dozens of choices

Data Transformation

- Data transformation techniques
  - Data discretization
  - Data normalization
  - Data standardization
  - Convert nominal attributes to dummy variables

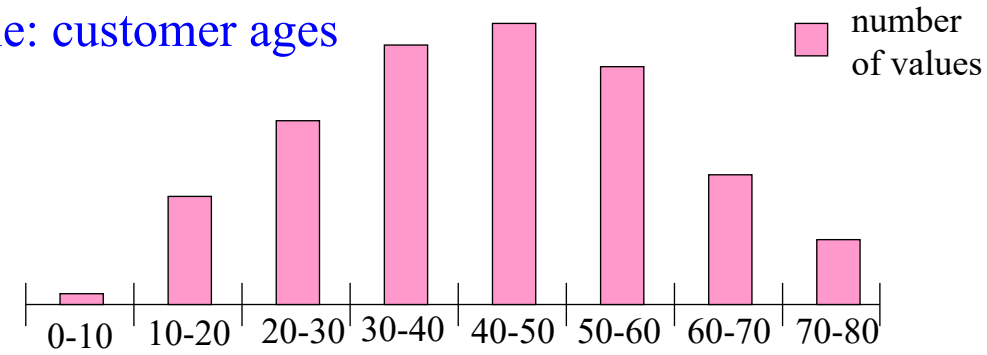
Data Transformation

- **Data discretization**
  - **This technique converts continuous data into discrete data by assigning each observation to a specific category or class.** This can be useful in cases where the relationship between the variables is non-linear or when using algorithms that require categorical inputs, such as decision trees and Naive Bayes classifiers.
  - Common methods:
    - **Equi-width binning**: Divide the data range into equal intervals
    - **Equi-frequency binning**: Divide the data into intervals with equal numbers of observations

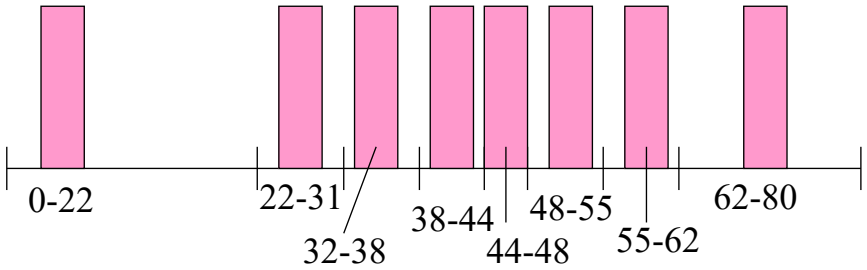
Data Discretization

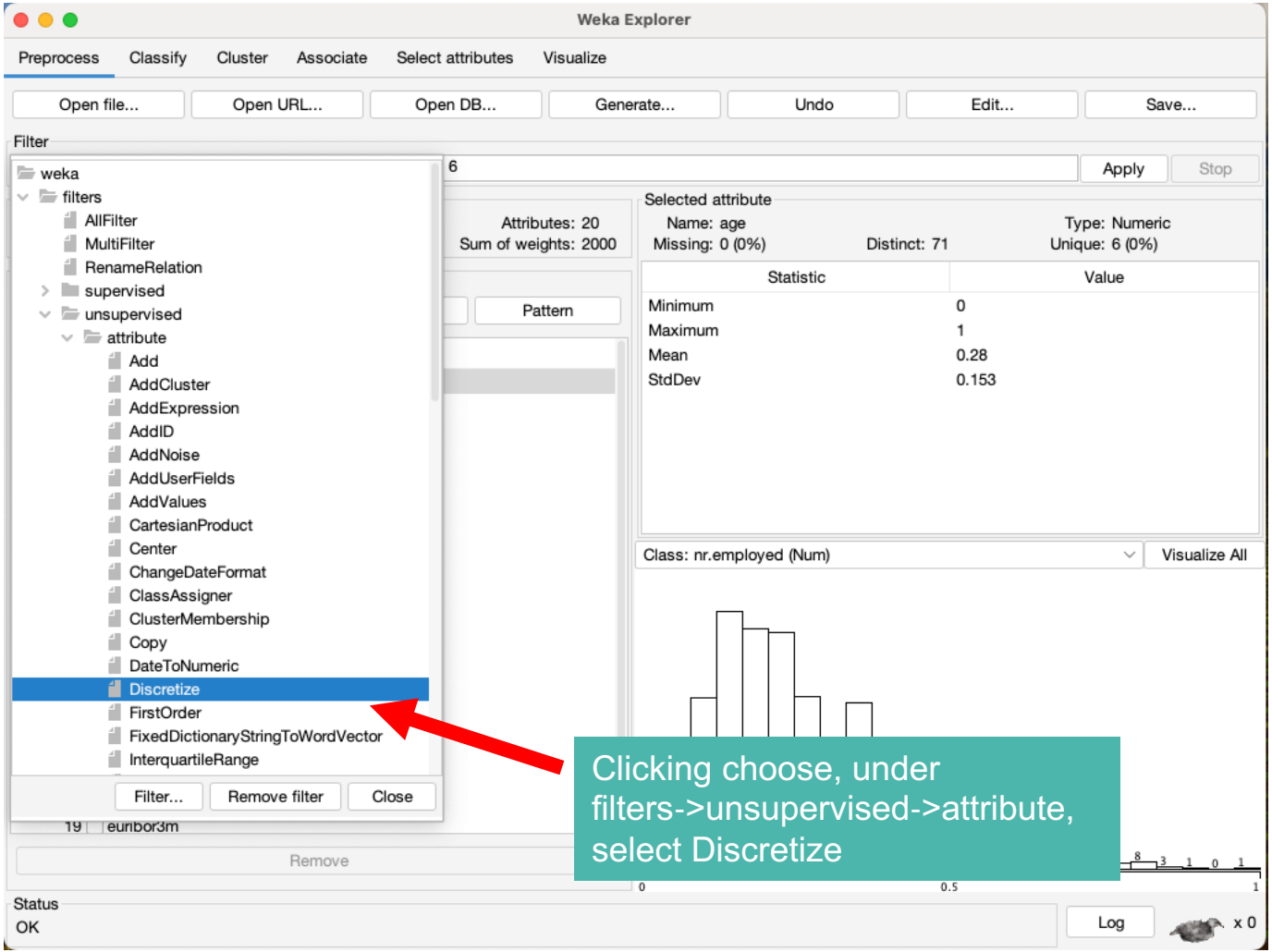
Example: customer ages

Equi-width binning



Equi-frequency binning





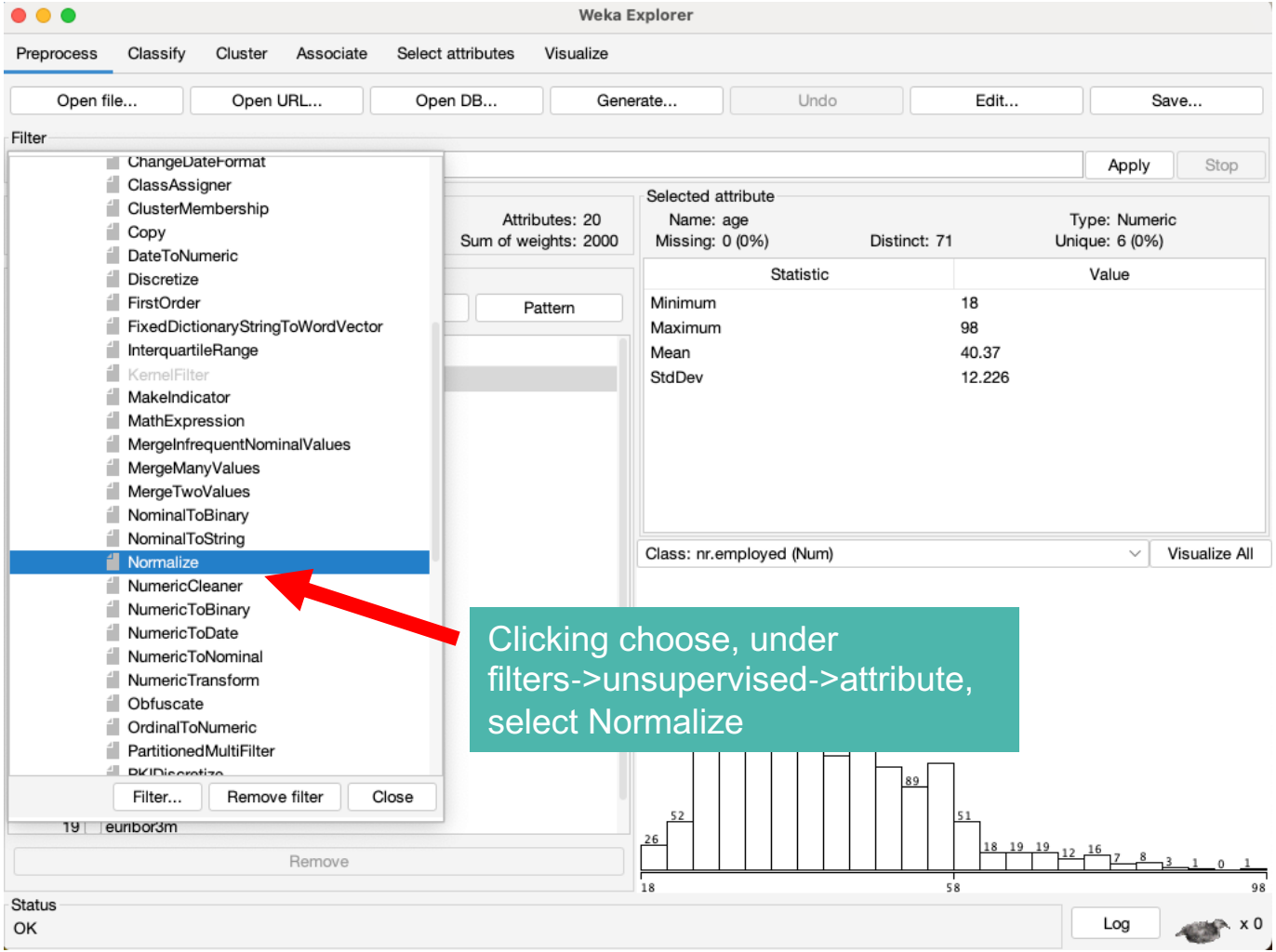
Clicking choose, under filters->unsupervised->attribute, select Discretize

Data Transformation

- **Data normalization**
  - **This technique scales the data to a specific range, usually between 0 and 1**. This helps in reducing the impact of outliers and allows for a better comparison between different variables.
- min-max normalization (usually new_max=1, new_min=0)

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - e.g. convert age=30 to range 0-1, when min=10, max=80. new_age=(30-10) * (1-0) / (80-10) + 0 = 2/7



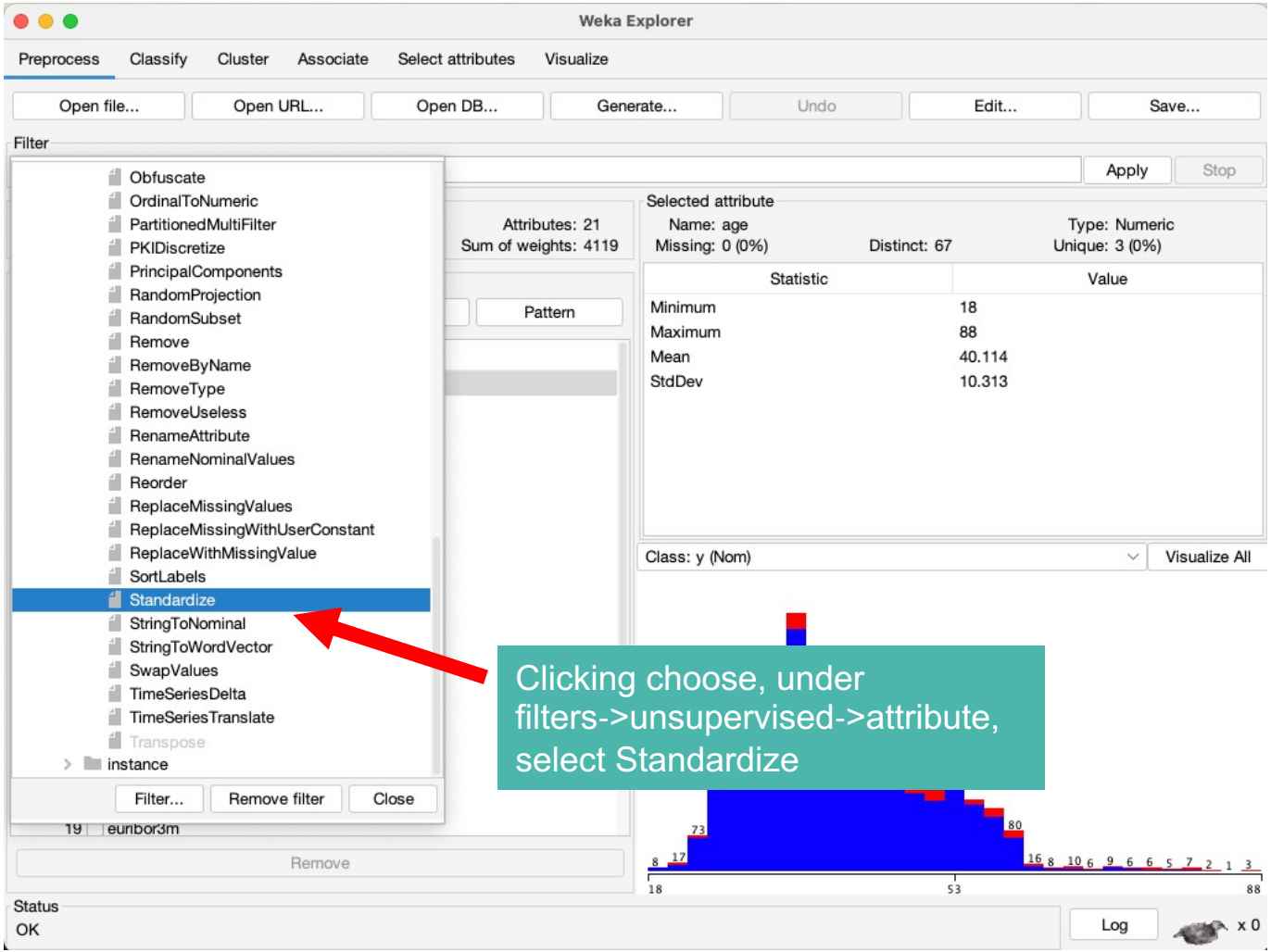
Clicking choose, under filters->unsupervised->attribute, select Normalize

Data Transformation

- Data standardization
  - **This technique transforms the data to have a mean of 0 and a standard deviation of 1.** It helps in comparing features with different units or scales and is particularly useful for algorithms that are sensitive to the scale of input features

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

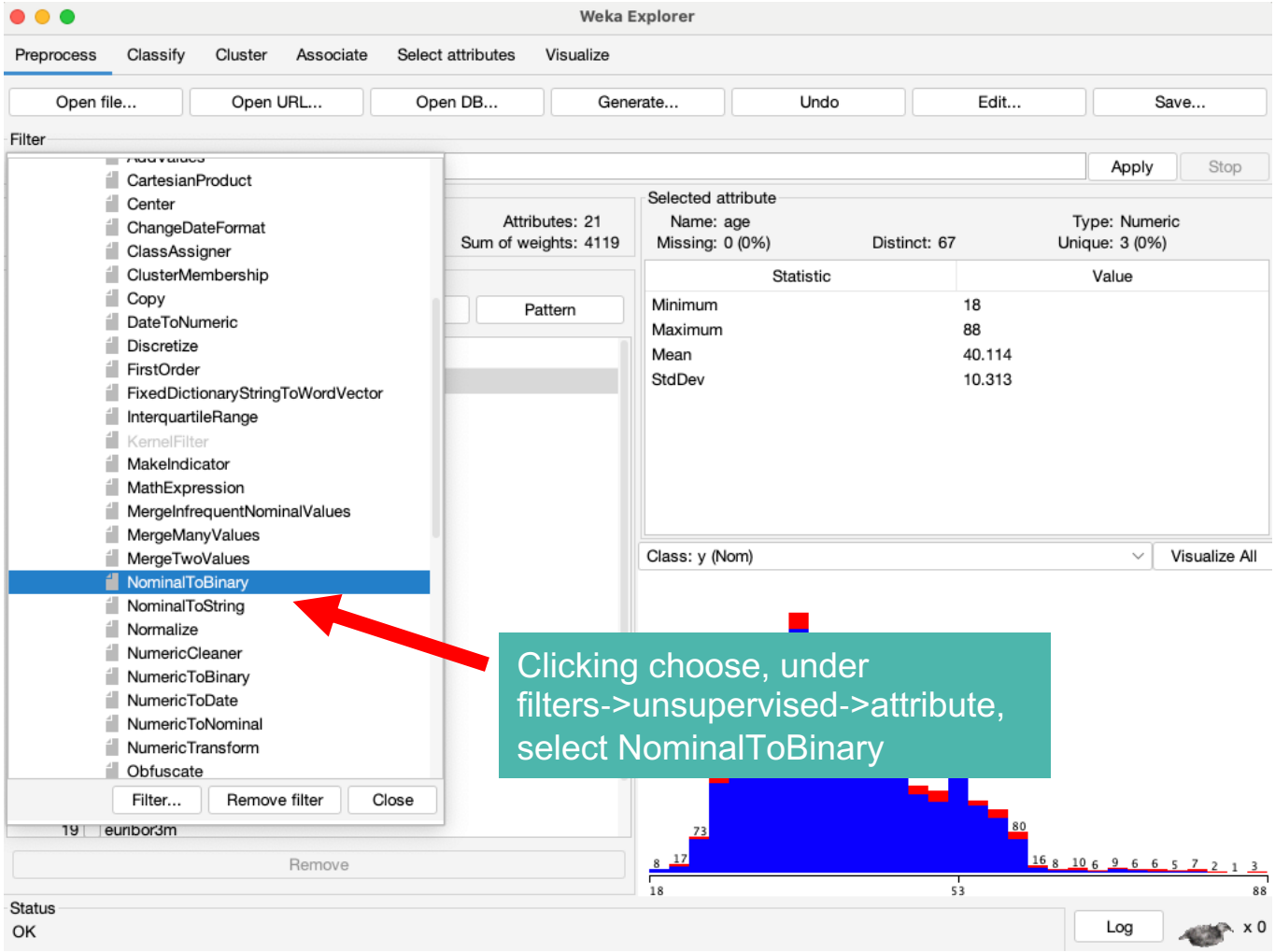
  - e.g. convert age=30, when mean=20, stand_dev=10. new_age=(30-20) / 10 = 1



Clicking choose, under filters->unsupervised->attribute, select Standardize

Data Transformation

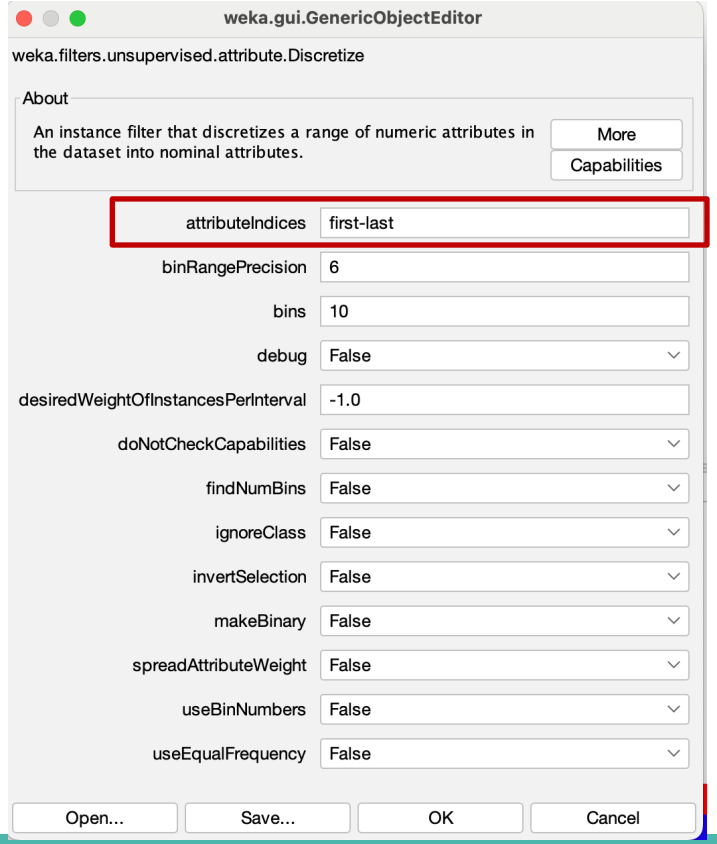
- **Convert Nominal Attributes to Dummy Variables**
  - Some machine learning algorithms prefer to use real valued inputs or do not support nominal or ordinal attributes.
  - Nominal attributes can be converted to real values. **This is done by creating one new binary attribute for each category.** For a given instance that has a category for that value, the binary attribute is set to 1 and the binary attributes for the other categories is set to 0. **This process is called creating dummy variables.**



Clicking choose, under filters->unsupervised->attribute, select NominalToBinary

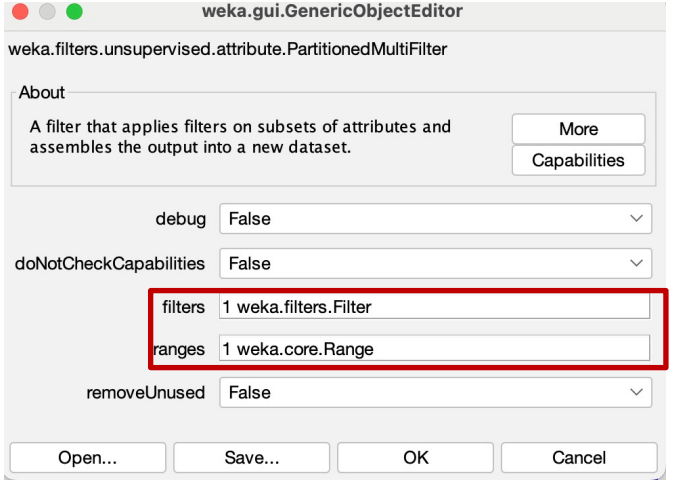
How to apply filters on particular variables?

- Some filters in Weka such as Discretize naturally support this
  - Use the parameter “attributeIndices” to indicate the desired attributes
  - By default, this parameter is set to “first-last”, which means that applying the filter on all variables



How to apply filters on particular variables?

- Some filters in Weka such as Normalize do not support this
  - We can use **unsupervised.attribute.PartitionedMultiFilter** to apply a filter on selected variables!
  - Use the “filters” parameter to indicate the filters used and the “ranges” parameter to indicate the selected variables



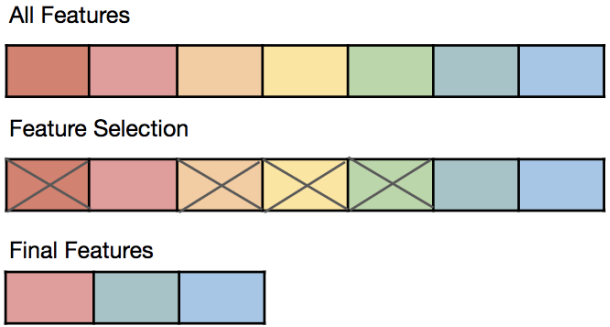
Data Reduction

- **Problem:**
  - Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- **Solution:**
  - Data reduction …

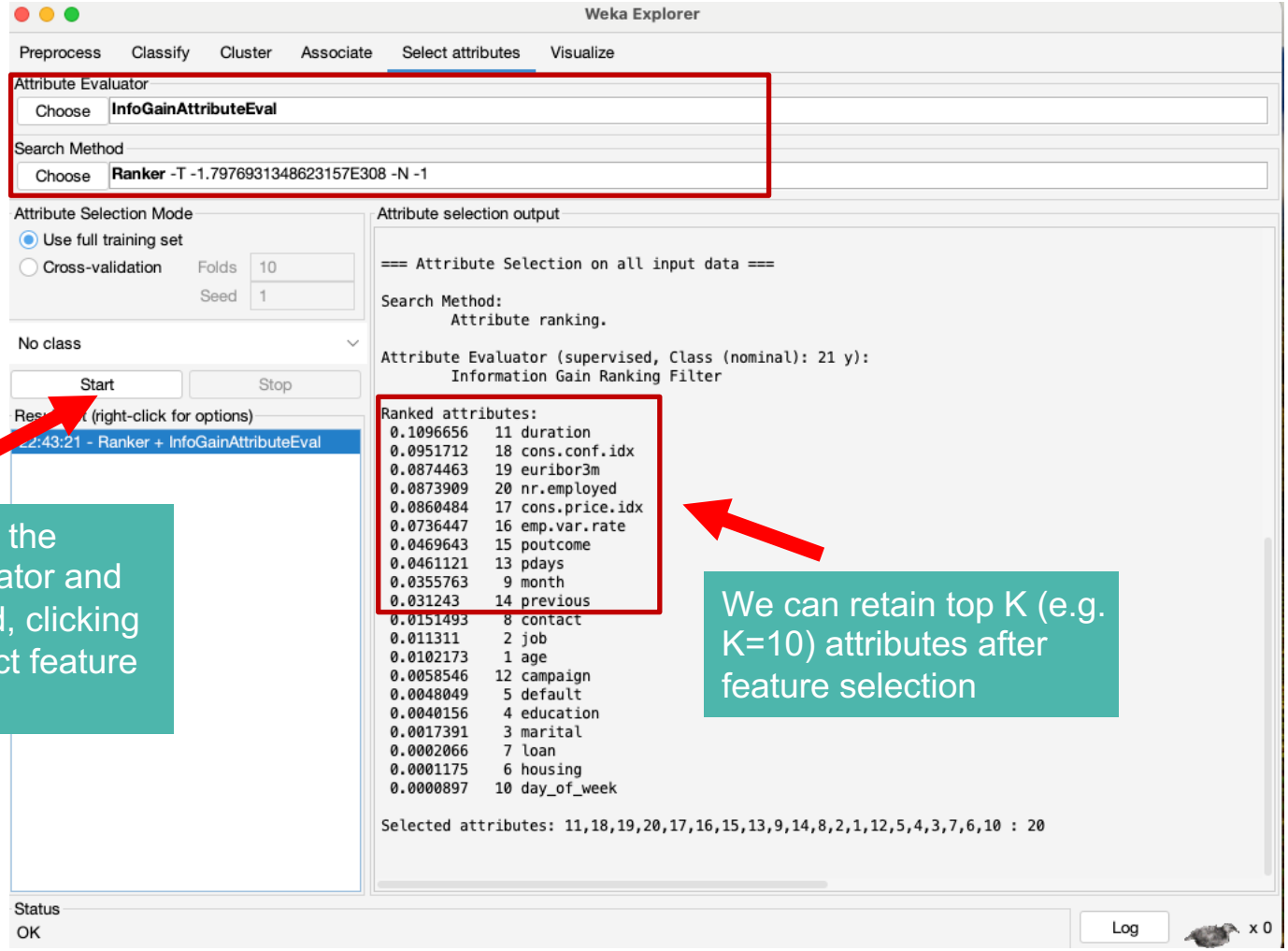
Data Reduction: Feature Selection

- **Feature selection (i.e., attribute subset selection): :**
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)



Feature Selection in Weka

- Feature Selection is divided into two parts
  - Attribute Evaluator: The attribute evaluator is the evaluation method for evaluating each attribute in the dataset based on the class
  - Search Method: The search method is the method for trying different combinations of attributes in the dataset in order to arrive on a short list of chosen variables.
- Some Attribute Evaluator techniques require the use of specific Search Methods.
  - For example, the CorrelationAttributeEval can only be used with a Ranker Search Method, it evaluates each attribute and rank the results.



After choosing the attribute evaluator and search method, clicking Start to conduct feature selection

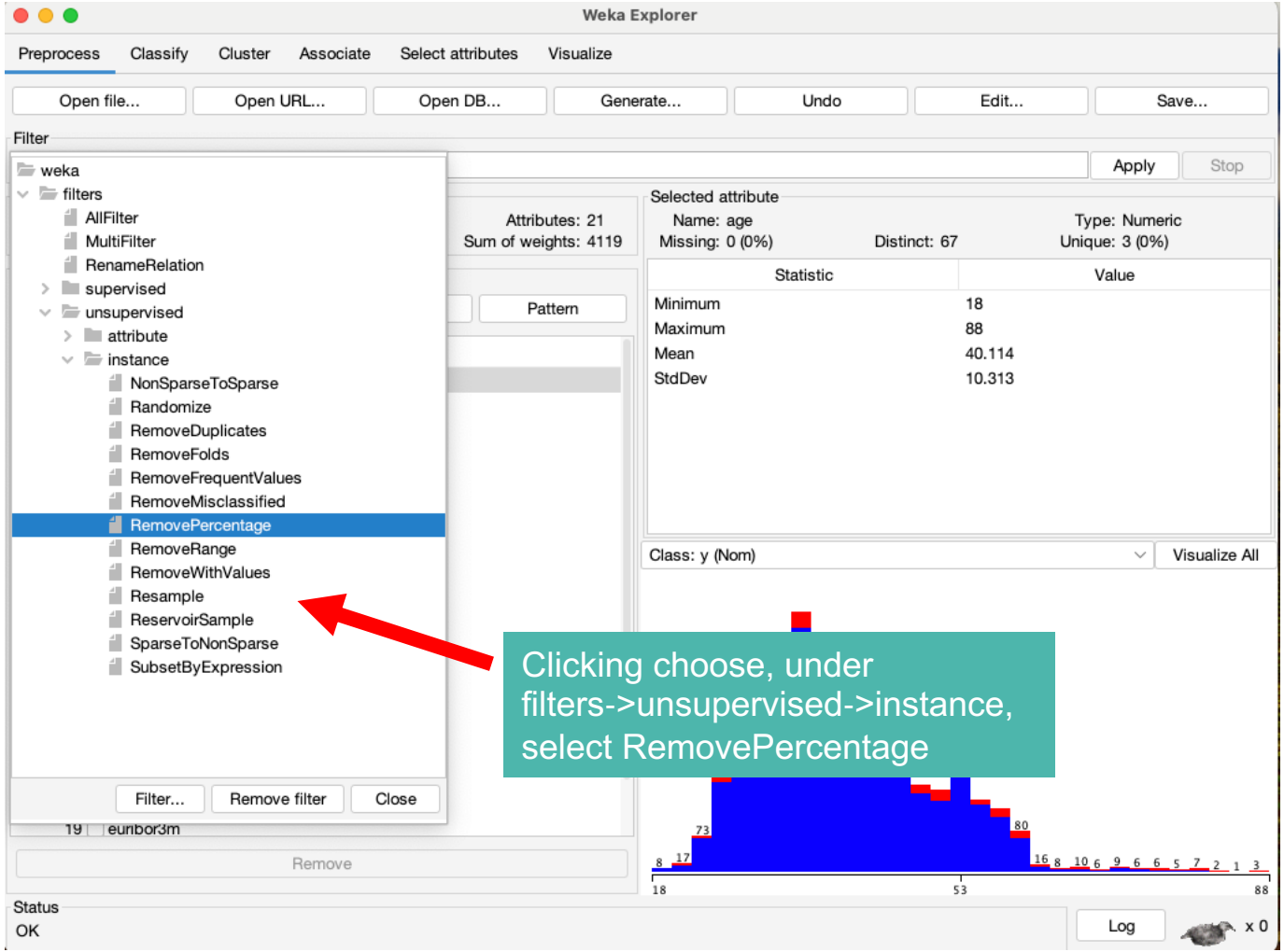
We can retain top K (e.g. K=10) attributes after feature selection

Data Splitting

- For developing models, we need to split the dataset into training and testing sets
  - For example, a training set comprising 80% of the original dataset and a test set consisting of the remaining 20%
- Data splitting using Weka
  - Use the RemovePercentage filter (An instance filter)

Data Splitting

- ❑ Training set
  - ❑ Load the full dataset
  - ❑ select the RemovePercentage filter in the preprocess panel
  - ❑ set the correct percentage for the split
  - ❑ apply the filter
  - ❑ save the generated data as a new file
- ❑ Test set
  - ❑ Load the full dataset (or just use undo to revert the changes to the dataset)
  - ❑ select the RemovePercentage filter if not yet selected
  - ❑ set the invertSelection property to true
  - ❑ apply the filter
  - ❑ save the generated data as a new file



Clicking choose, under filters->unsupervised->instance, select RemovePercentage

Q&A