**Lab 05**

**Topic – Data Pre-Processing Techniques (Part-2)**

**Objective**

* To learn more about Data Pre-processing techniques.
* To apply type conversion
* To apply dimensionality reduction techniques

**Current Lab Learning Outcomes (LLO)**

By completion of the lab the students should be able to use

1. Attribute’s Type Conversion:
   * + Convert data from Numerical to Binominal
2. Dimensionality Reduction by Using:
   * Principal Components Analysis (PCA)
   * Information Gain
3. Sampling
   * + Random Sampling (Simple)
     + Random Sampling (Stratified)

**Lab Requirements**

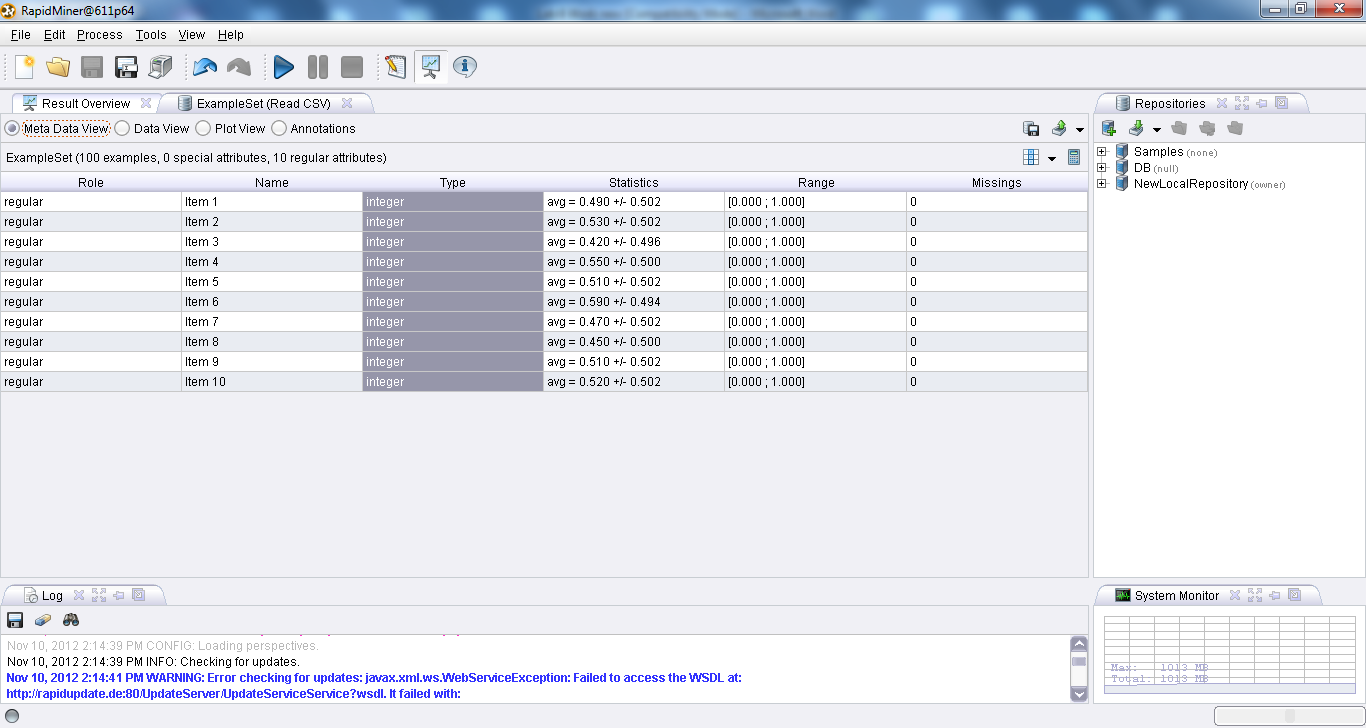
Rapid Miner

**Lab Assessment**

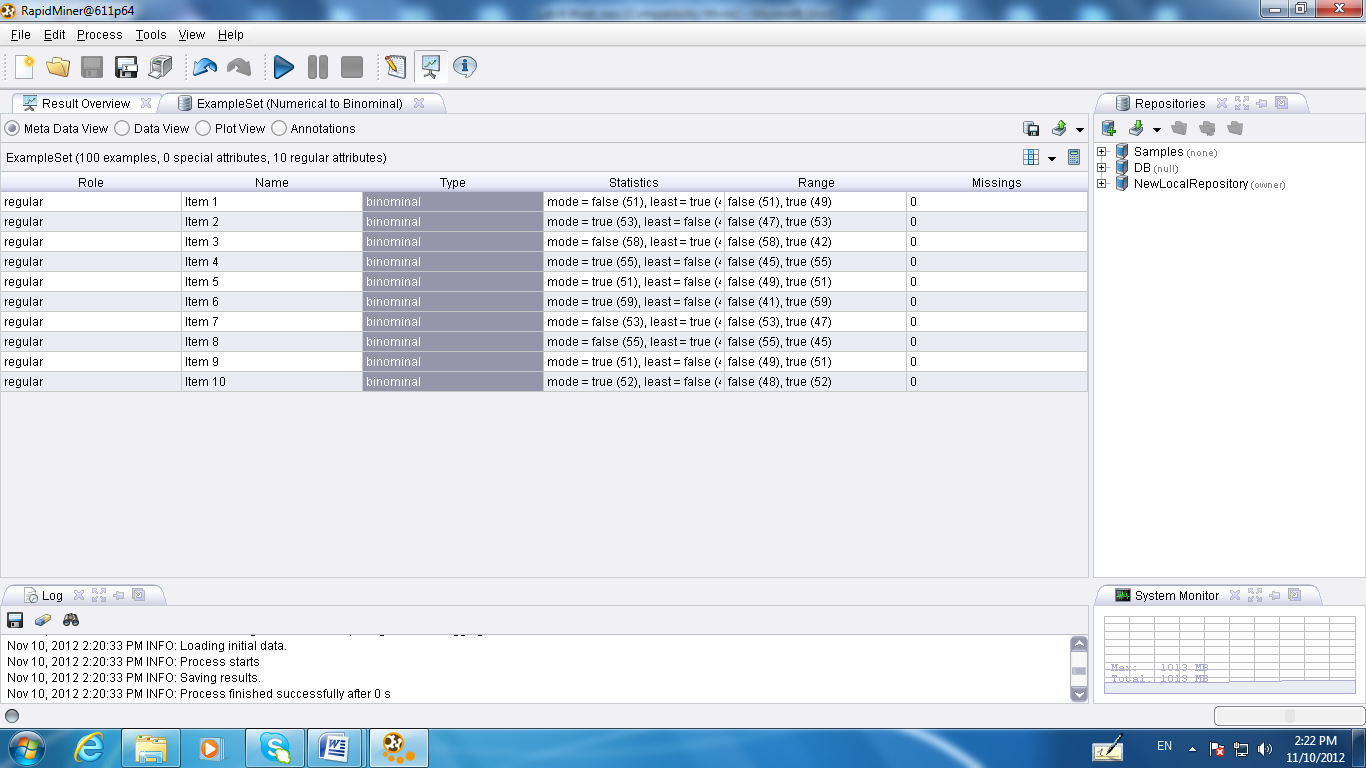
Hands on practice. **Lab Description**

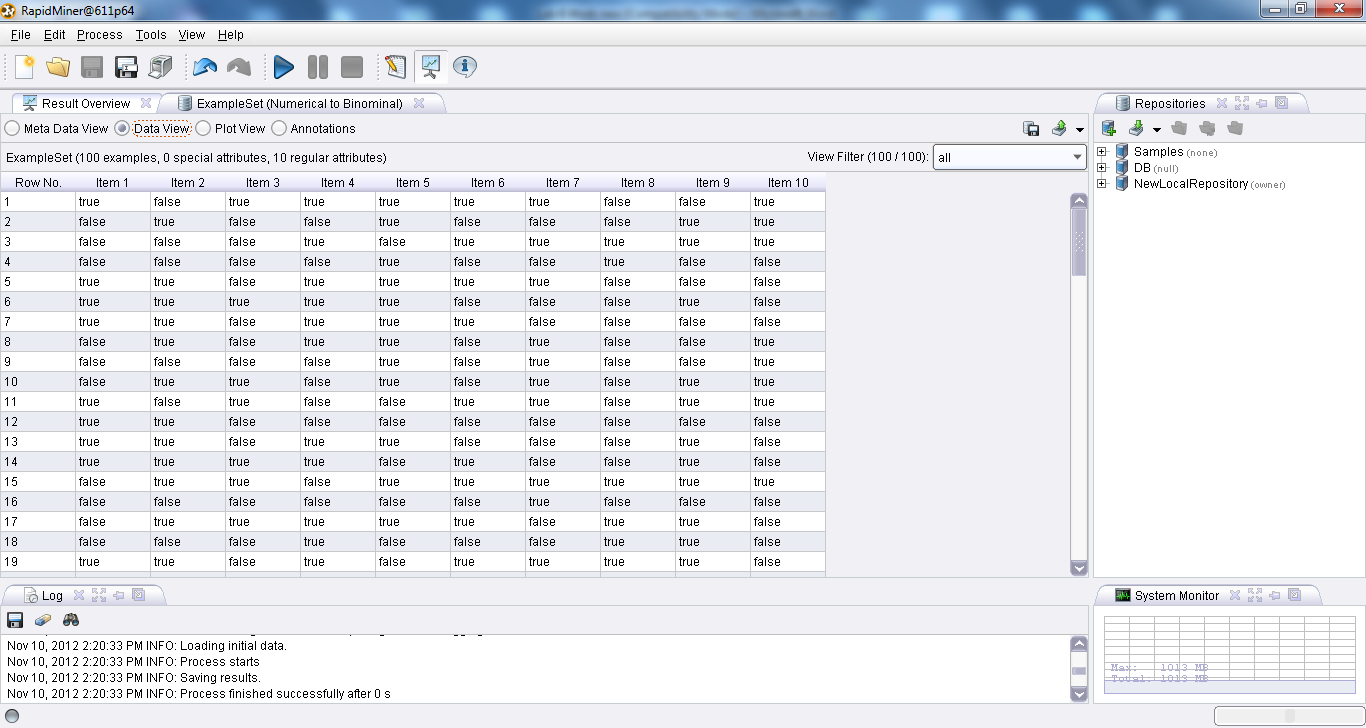
1. Attribute’s Type Conversion:
   * + Convert data from numerical to binominal

* Open Rapidminer and Import File “Lab-4\_Dataset “
* Check the attribute values in meta data view ,



* Using operators expand Data Transformation ,then expand type conversion ,open numerical to binominal .





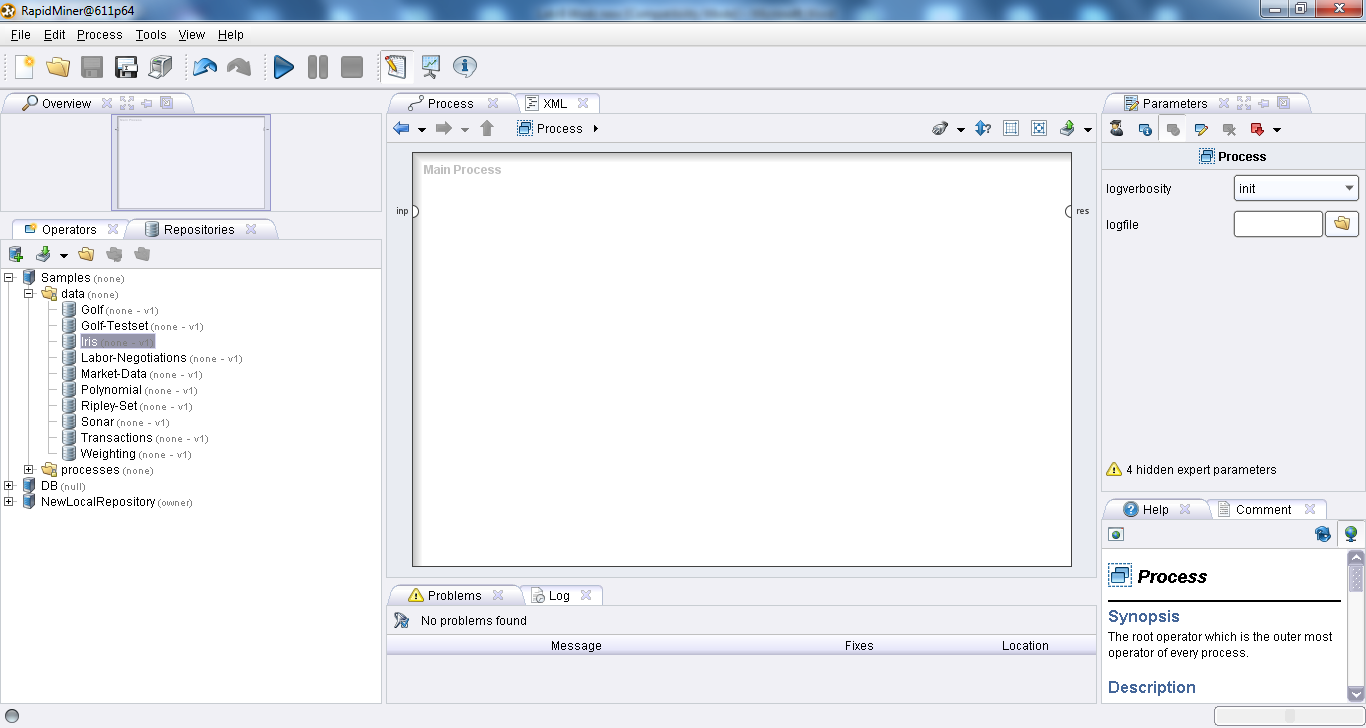
1. Dimensionality Reduction
   1. Principal Component Analysis (PCA)

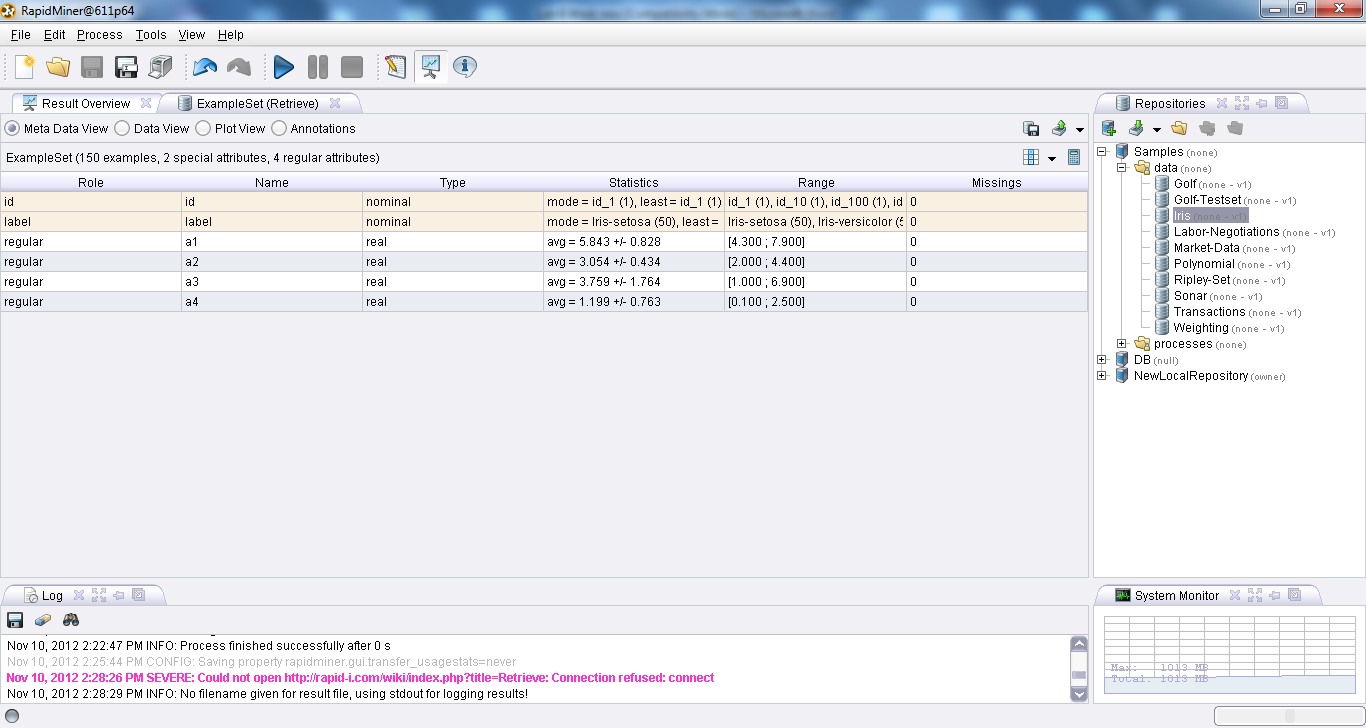
PCA is a dimensionality reduction technique that transforms a set of attributes into another set called principal components. These principal components are sorted by decreasing order of importance using the proportion of variance they explain from the original data. PCA requires all the input attributes to be numerical. If nominal (binominal or polynominal) attributes exist, they must be transformed to numerical. The dataset used for this process is **iris**. The RapidMiner operator used is PCA (Principal Component Analysis). The *dimensionality reduction* parameter value for PCA *is Keep Variance* and the corresponding *variance threshold* is *0.95*. The former parameter defines the criteria for dimensionality reduction while the latter one explains that we need to keep those principal components that combined explains 95% of the variance. In RapidMiner’s result section, check the PCA tab. This tab explains the principal components as PC1, PC2 etc and the proportion of variance they explained. For the iris dataset, the first principal component, PC1, explains 92.5% of variance. This is followed by PC2, with 5.3%, PC3 with 1.7% and PC4 with 0.5%. Since our variance threshold is 0.95 (95%) we only need to keep PC1 and PC2 as they explain 97.8% (92.5+5.3) variance. PC3 and PC4 can be removed from the data, thus reducing the number of dimensions from four to two. This can be seen in the ExampleSet tab in the results section.

Use the sample data set Iris

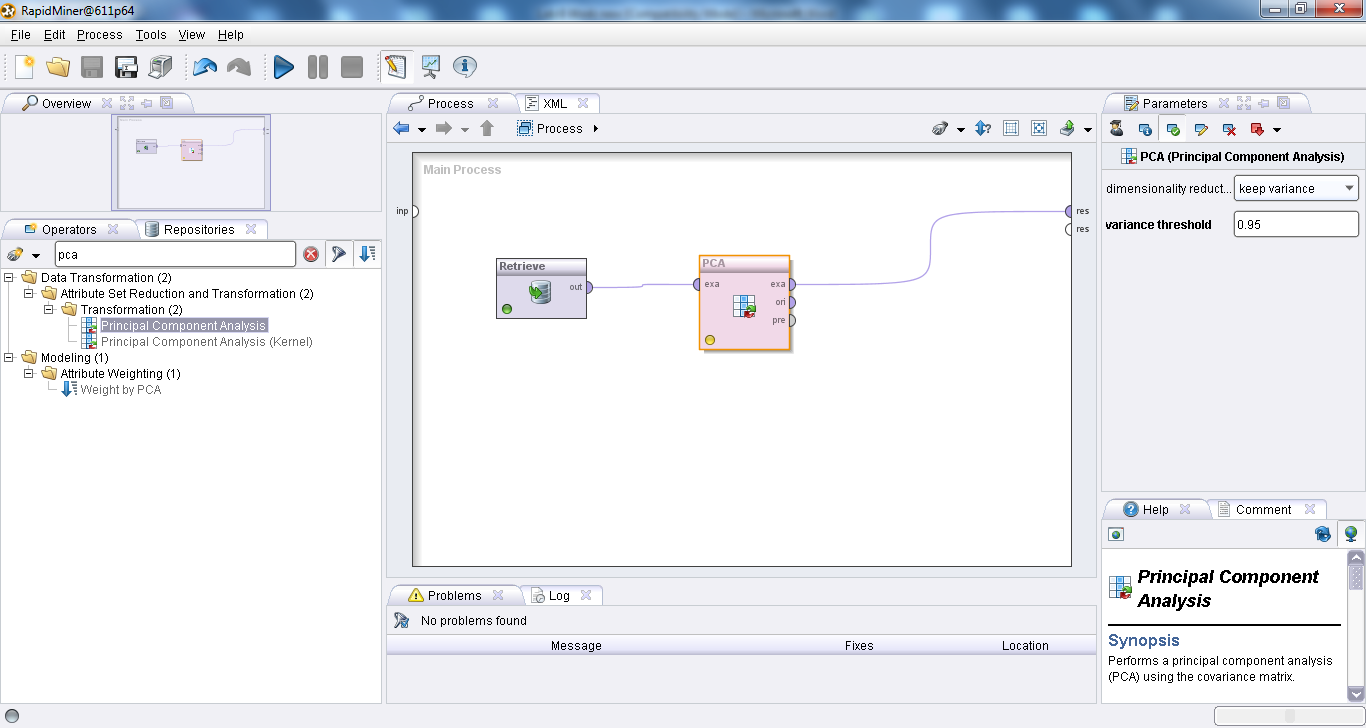
Location:

Repository 🡪 Samples 🡪 Data 🡪 Iris

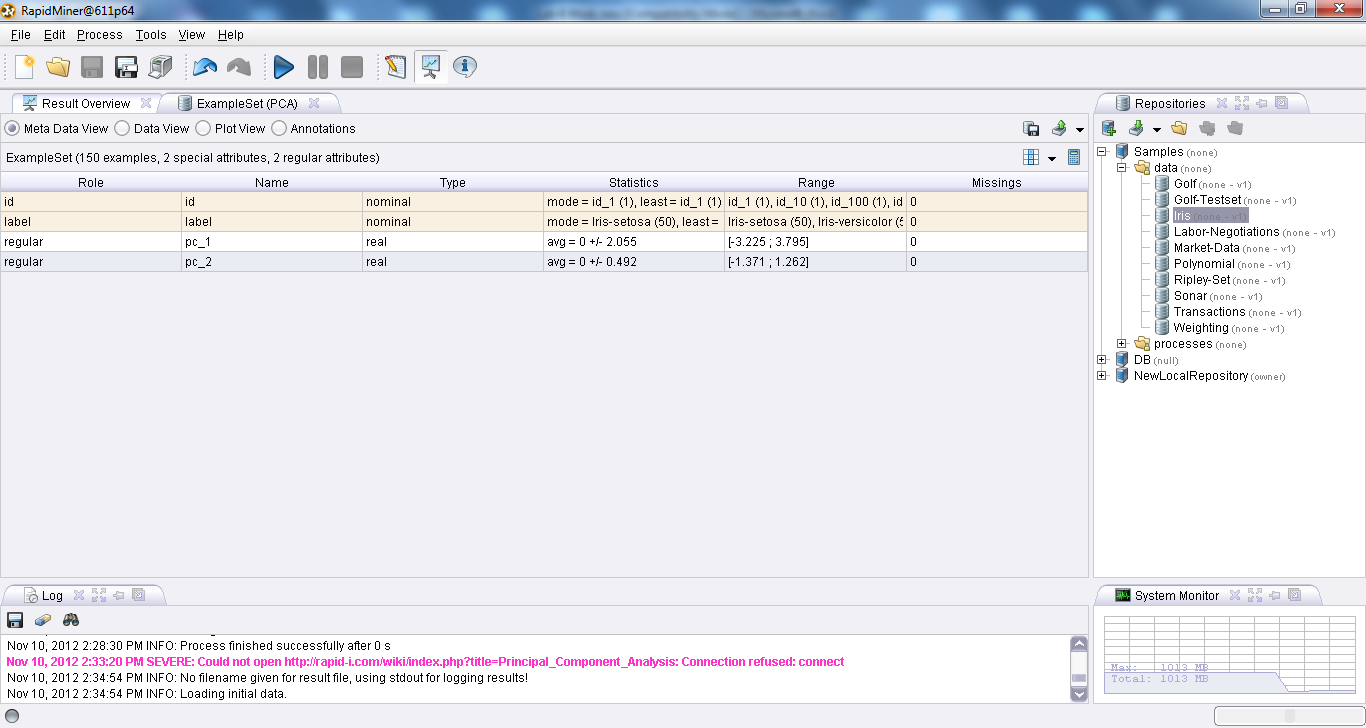


Check the number of attribute 

* Using operators expand Data Transformation ,then expand attribute set reduction and transformation, transformation open principal component analysis
* In principal component analysis check the variance threshold

****

* **Run the program, check number of attribute**

****

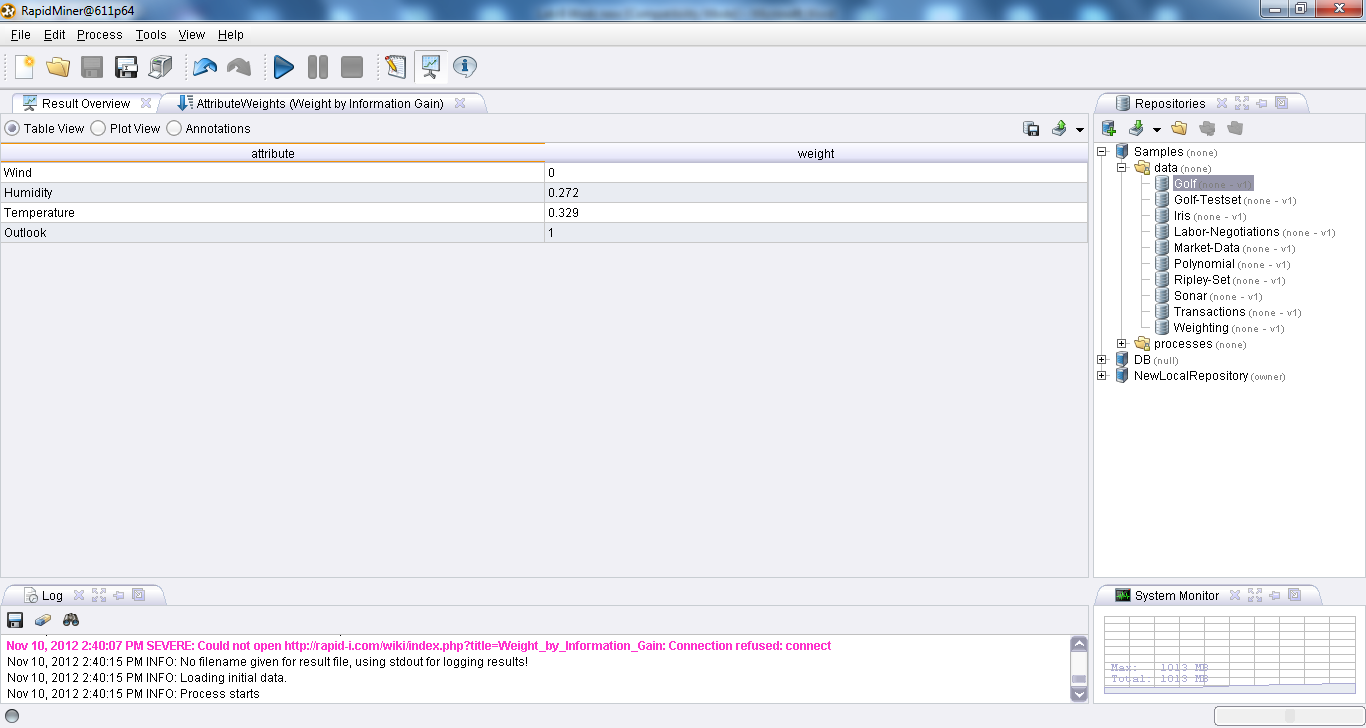
* 1. Information Gain

Information gain is used to select the best attributes for classification, hence a class (label) attribute is required. Information gain for an attribute X with respect to a class label Y is the reduction in uncertainty (gain of information) about Y, if we know the value of X. The dataset used for this process is **golf**. The RapidMiner operator used for this process is *Weight by Information Gain.* In RapidMiner’s results section, the Attribute Weight tab contains the weights of the attribute using information gain. We can set a threshold and the attributes with weights less than the threshold can be removed from the dataset. In the golf dataset, we can safely remove, the attribute wind, whose weight is zero. This can be done using the RapidMiner operator *Select by Weight.* In our example the weight threshold is set to 0.25.

Use the sample data set golf data set

* Repository 🡪 Sample 🡪 Data 🡪 Golf
* Using operators expand modeling ,then expand attribute weighting open weight by information gain **)**

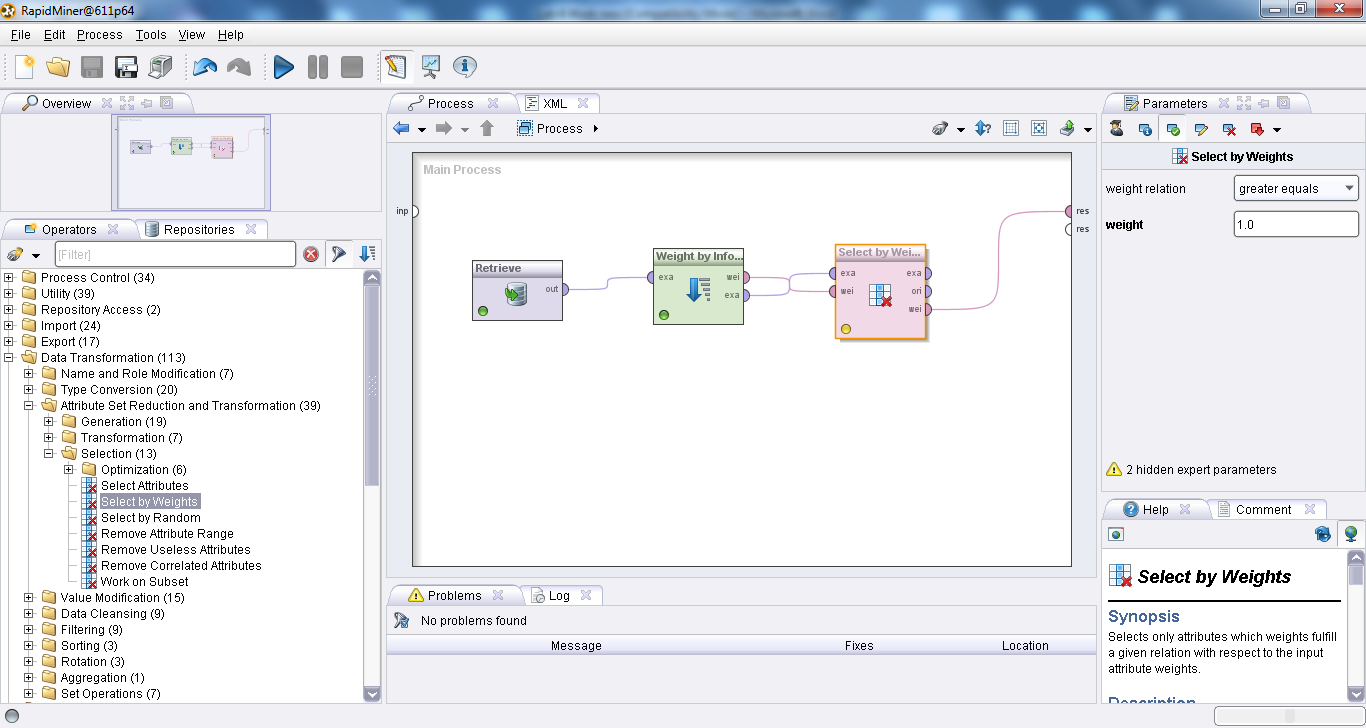
Check the weight for each attribute

****

To select the satisfied attribute

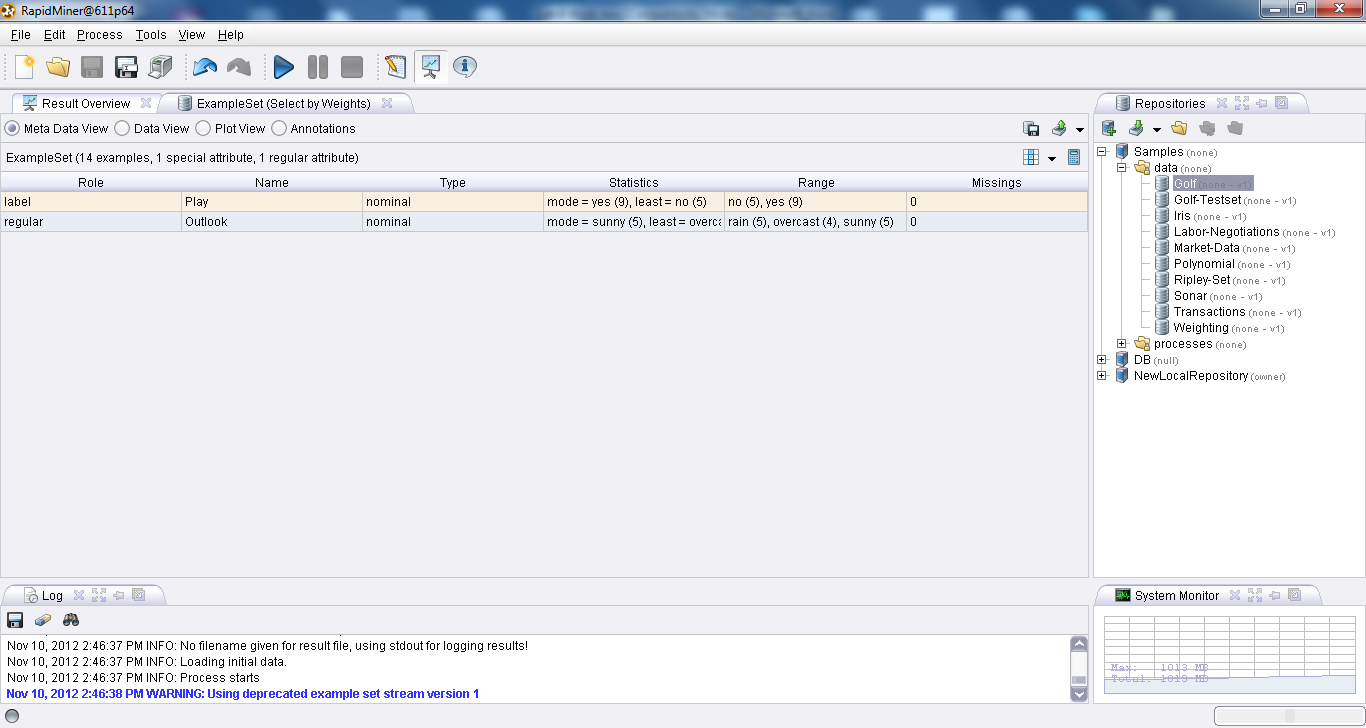
* Using operators expand data transformation ,then expand attribute set reduction and transformation, selection open select by weight **)**

**Chose the proper weight**

****

]

Result Window



1. **Sampling: obtaining a small sample *s* to represent the whole data set *N***

**Simple random sampling:**

There is an equal probability of selecting any particular item

**Sampling without replacement:**

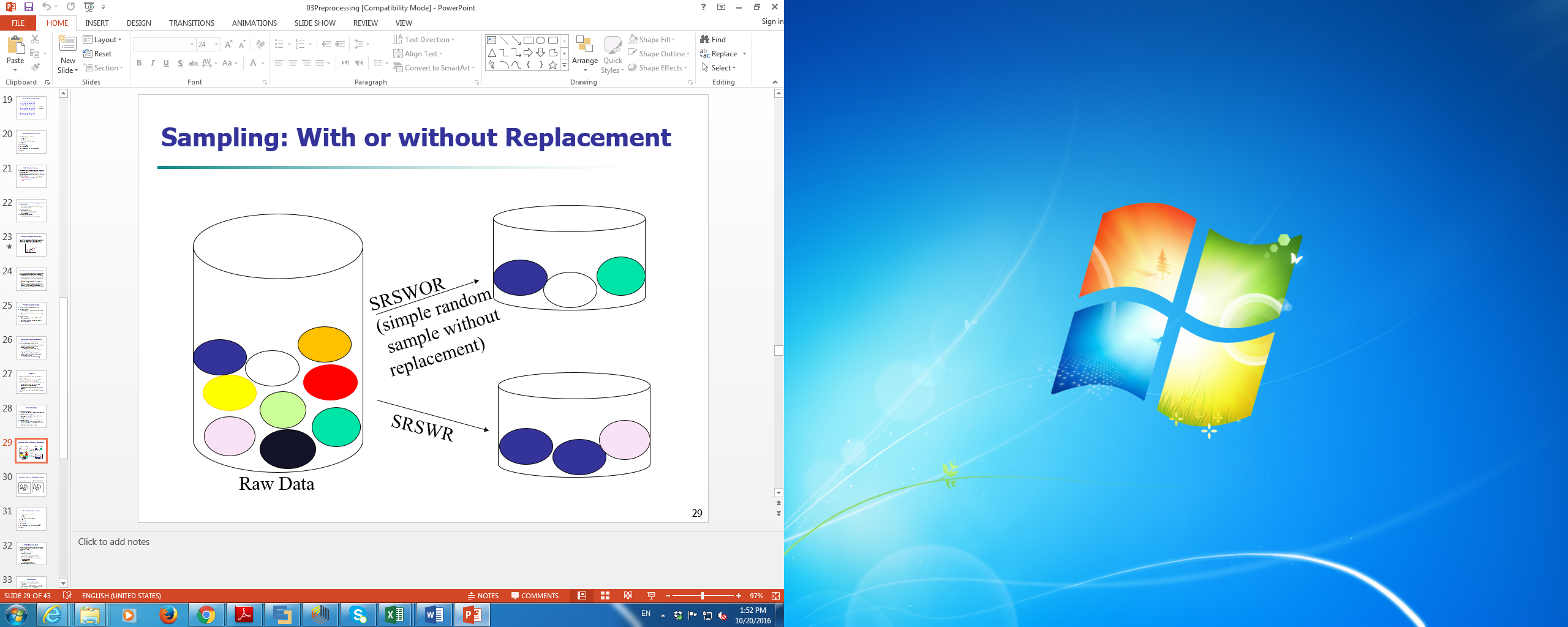
Once an object is selected, it is removed from the population

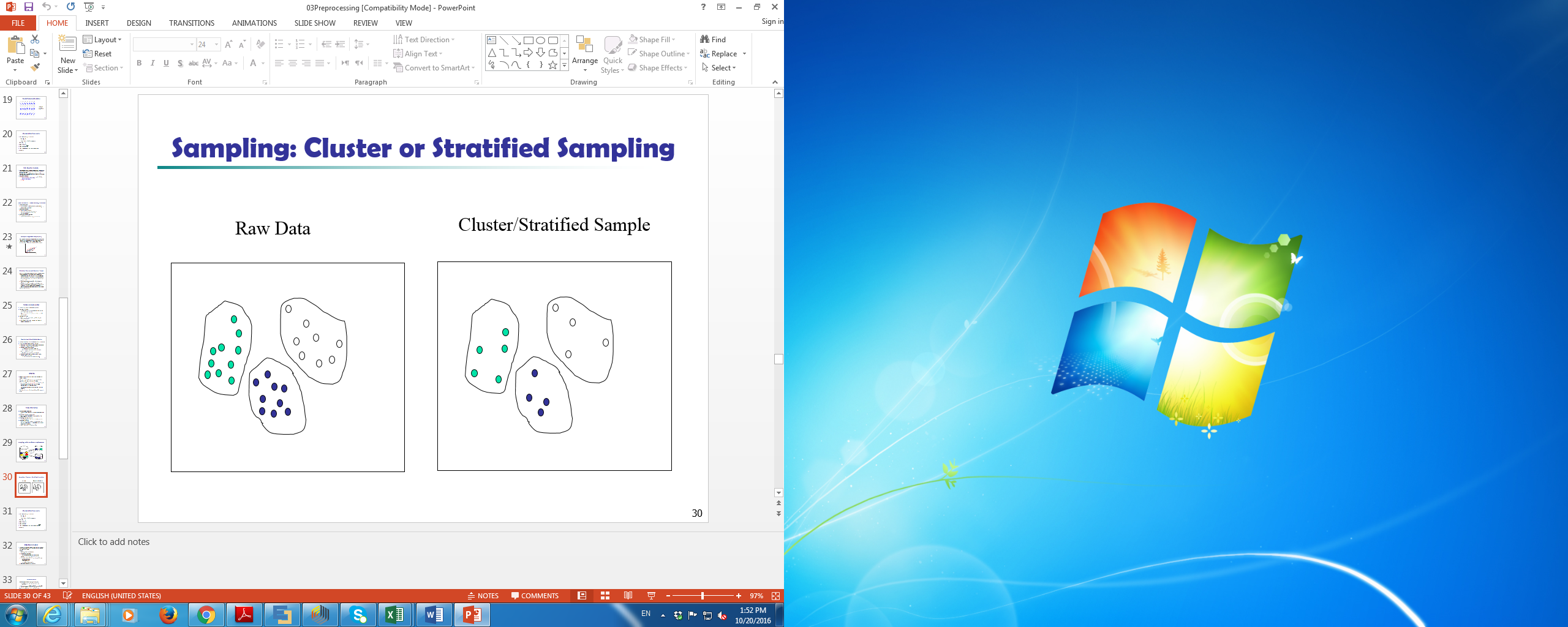
**Sampling with replacement:**

A selected object is not removed from the population

**Stratified sampling:**

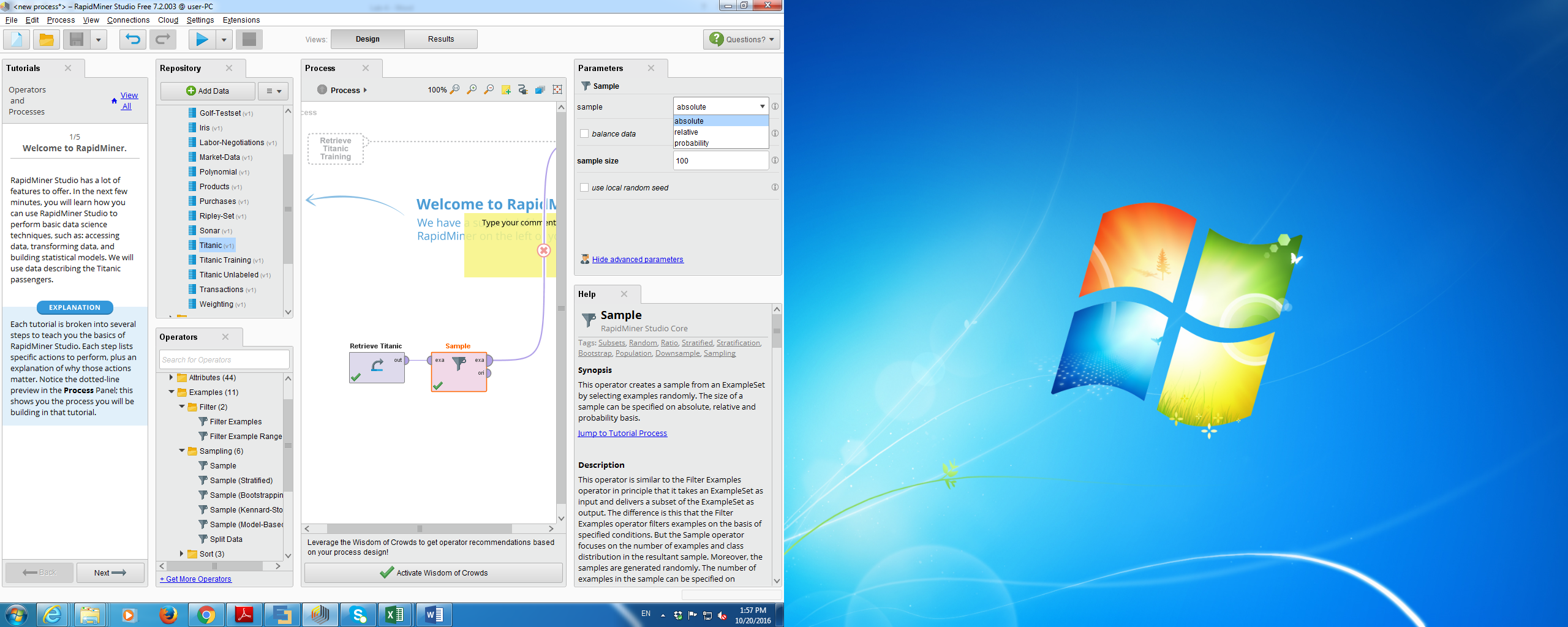
Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data) Used in conjunction with skewed data





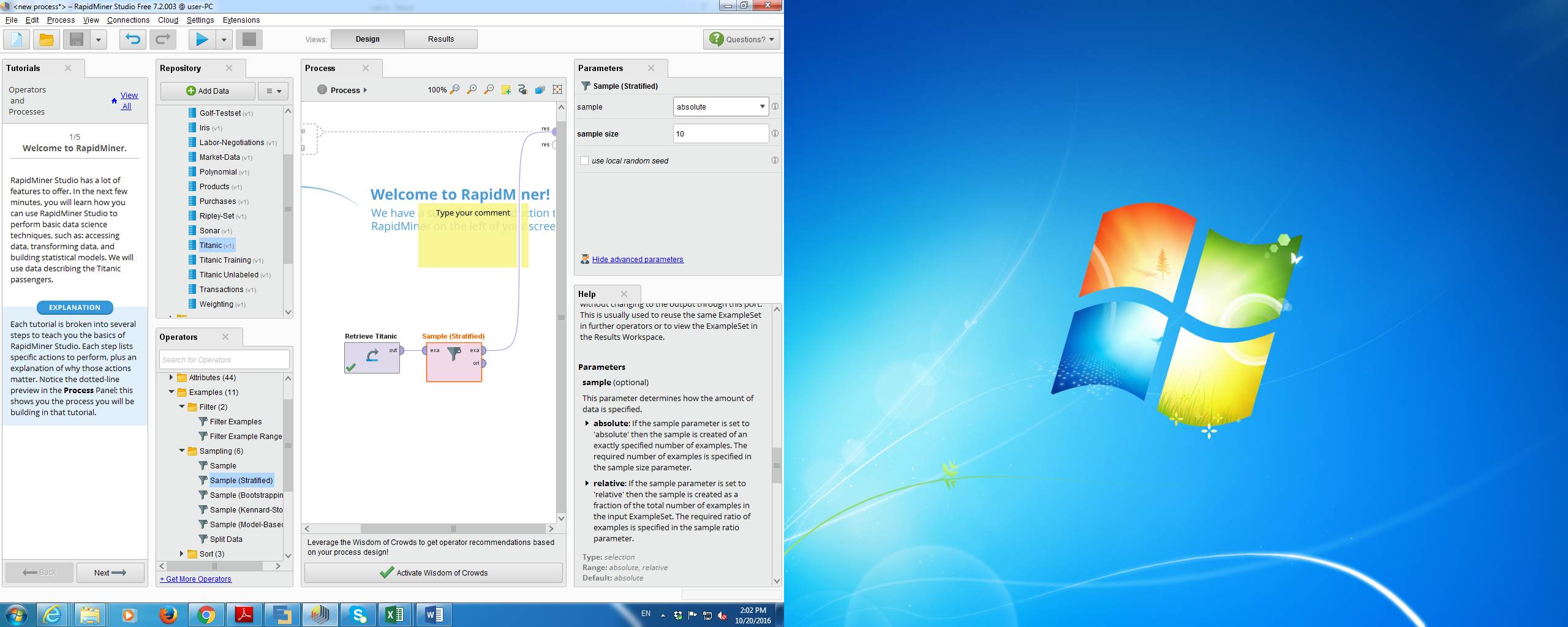
**Simple Sampling:**

In rapid miner, select Titanic dataset, using operator “Sample” from location “operator” 🡪 blending 🡪 example 🡪 sampling 🡪 sample. You can select the type of sample, and **sample size** as available in rapid miner.



**Stratified sampling:**

In rapid miner, select “iris” dataset, using operator “Sample” from location “operator” 🡪 blending 🡪 example 🡪 sampling 🡪 sample (Stratified). You can select the **type of sample** and **sample size** as available in rapid miner.



**Extra (supplementary) Materials** (if any)

Data Files provided through blackboard.