Introduction to Rapidminer

Amirhassan Monajemi 2023

ABOUT THE LECTURER

Dr. Amirhassan Monajemi (aka Monadjemi) is a Senior Lecturer in AI and Data Science at the School of Computing, the National University of Singapore. Before joining the NUS in 2019, he was with the Faculty of Computer Engineering, University of Isfahan, Iran, where he was serving as a professor of AI and Machine Learning.

Dr. Monajemi has taught diverse computer science courses for years, registered a few patents in the fields of AI, Machine Vision, and Signal Processing applications, published more than a hundred research papers in peer-reviewed, indexed journals, and supervised several related industrial projects in various scales.

AGENDA

Day	Time	Modules	Module Topics
1	9-10:45	Introduction and Environment	1. What is Rapidminer 2. How to install it 3. Applications 4. Environment
	11-13	Data, Preprocessing, and Visualization	1. Data resources in RM 2. Data Preprocessing 3. Basic data visualization in RM
LUNCH			
1	14-15:45	Clustering	1. How to cluster data in RM 2. Clustering example and practice
	15:45-16:30	Classification and Function Estimation	1. How to Estimate a function in RM 2. How to classify data in RM 3. Classification example and practice
	16:30-17:30	Mini Project	

Part 1

Introduction to Machine Learning

WHAT IS MACHINE LEARNING?

Machine learning (ML) is the study of computer algorithms that improve automatically through experience.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks.

Supervised Learning

Reinforcement Learning

Unsupervised Learning

WHAT IS MACHINE LEARNING?

A considerable portion of AI is Machine Learning.

We may design an intelligent machine to do something without the learning ability.

However, if that machine can ‘Learn’, it would be an instance of a learning machine.

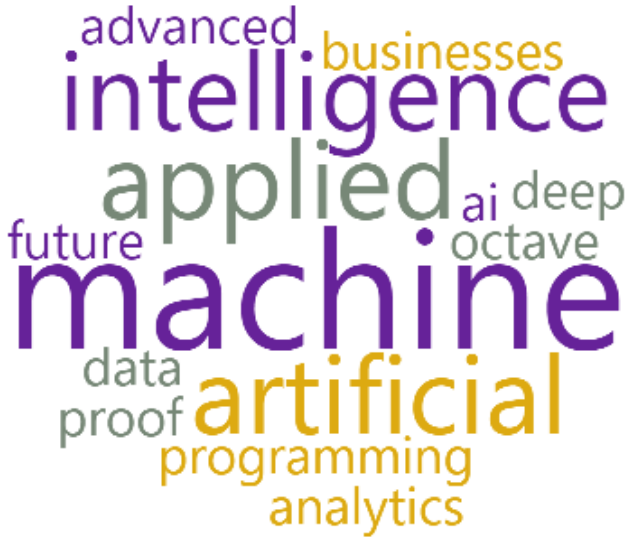
Obviously, the second one has got some privileges:

- It can improve its performance.
- It can learn by examples.
- It can explore and get in rather new duties.
- It would be more like us!

WHAT IS MACHINE LEARNING?

Examples:

- Artificial Neural Networks and Deep Networks
- Regression Alg.
- Support Vector Machines
- Learning Automatons



ARTIFICIAL INTELLIGENCE, A MODERN APPROACH, STEWART RUSELL AND PETER NORVIG, PEARSON PUB, 3RD ED. 2009.

Practice 1

- What are the main differences between data mining and data analytics?
  - Look them up and write them down in a paragraph or two.

Applications of Machine Learning

MACHINE LEARNING APPLICATIONS

Virtual Personal Assistants

Predictions

Video Surveillance

Social Media Services

Email Spam and Malware Filtering

Online Customer Support

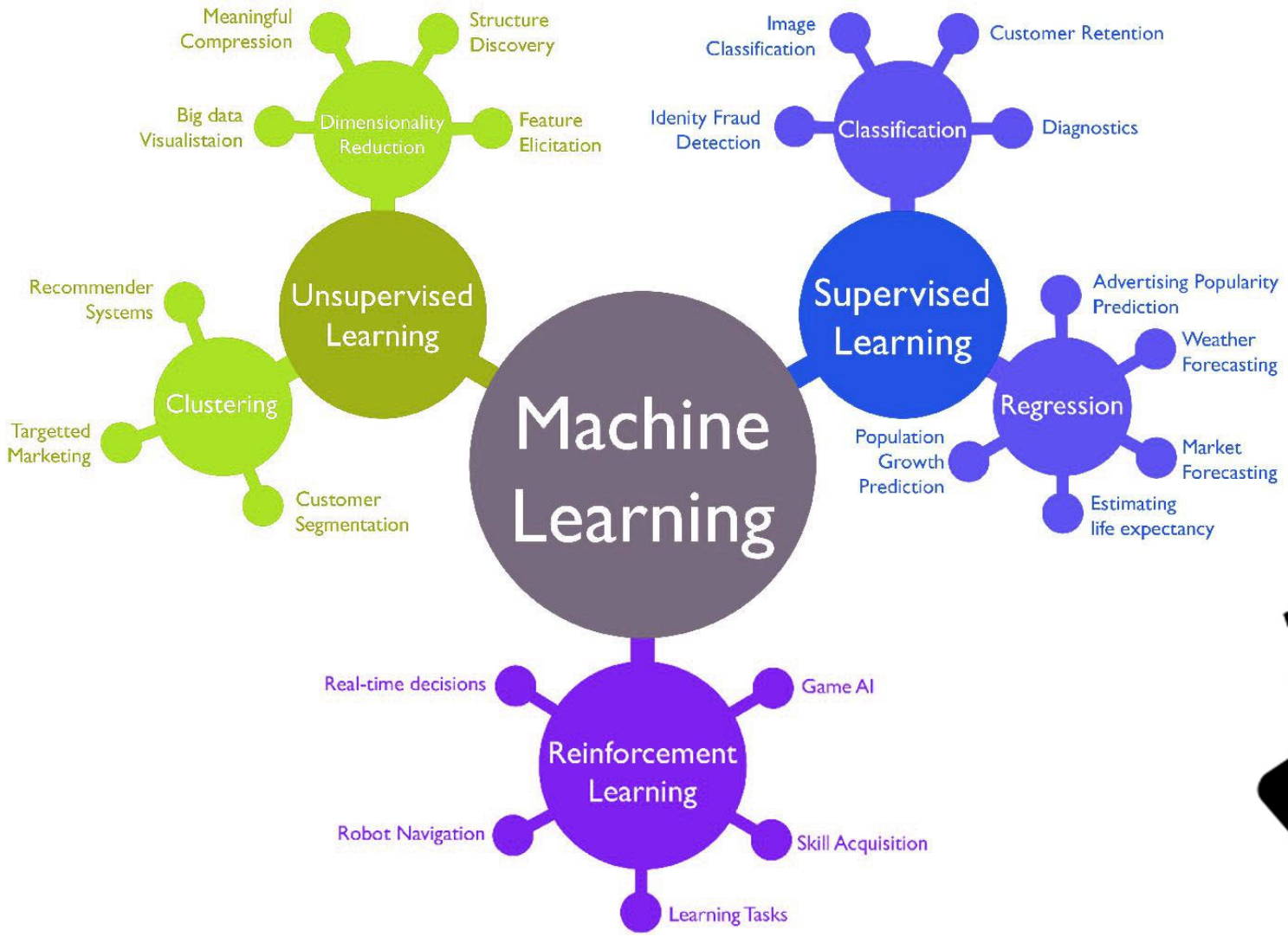
Search Engine Result Refining

Product Recommendations

Natural Language Processing, Machine Translation, Abstraction

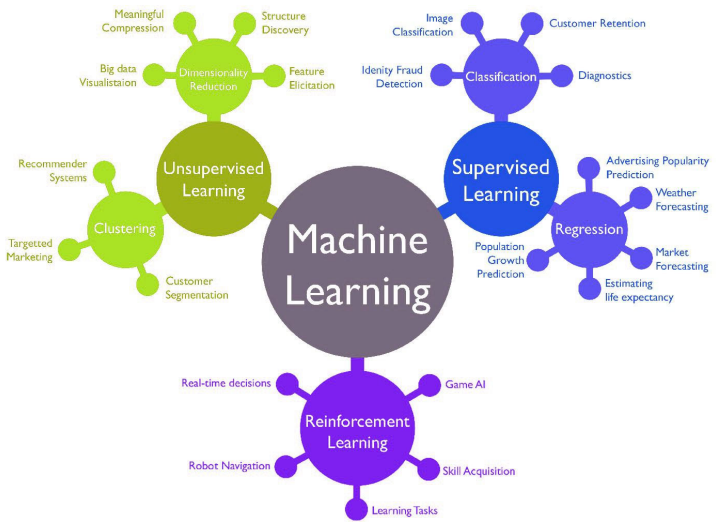
Driverless Vehicles

Machine Learning Approaches



HTTPS://WWW.WORDSTREAM.COM/BLOG/WS/2017/07/28/MACHINE-LEARNING-APPLICATIONS, **ABDUL WAHID**

Supervised Learning



- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- It infers a function from labeled training data consisting of a set of training examples.
- A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

Main Applications

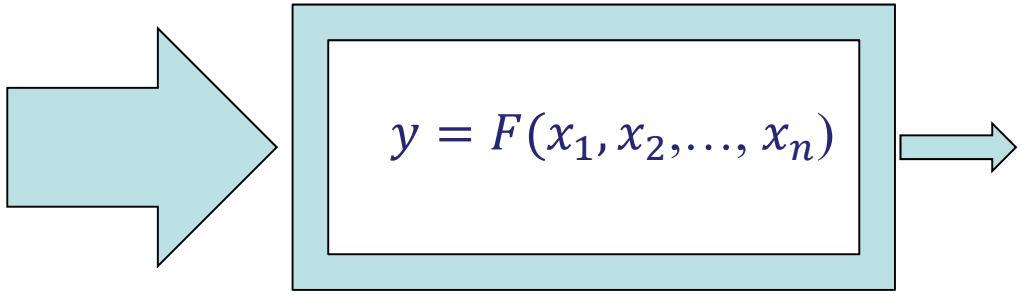
1- Classification

2- Function Estimation (Regression)

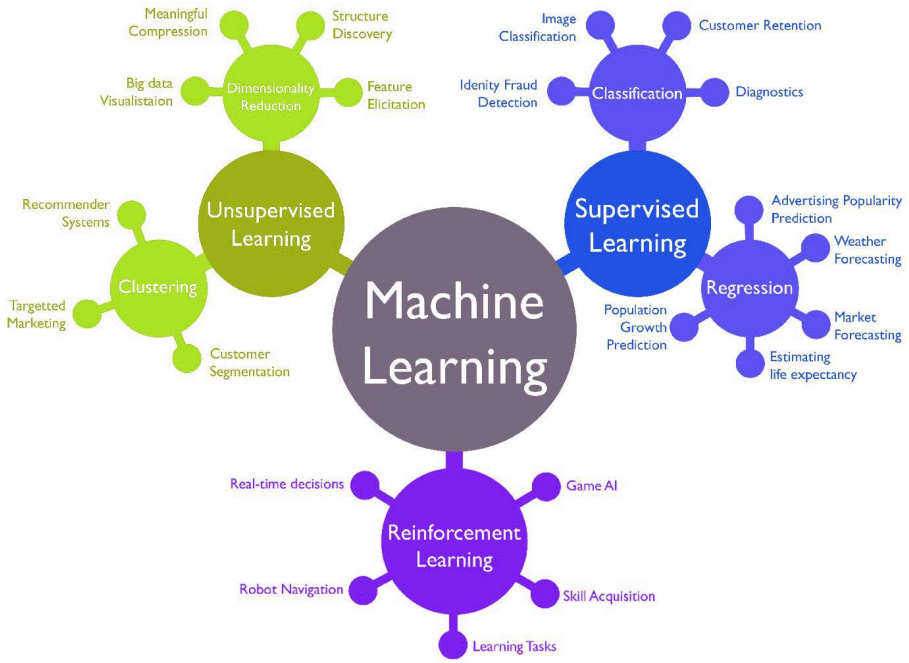
https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications, Abdul Wahid

Supervised Learning

- **To Classify …**
  - Your suppliers into more-reliable and less-reliable classes
  - Celebrities into classic and modern era
  - Loan requests into approvable and risky
- **To Estimate …**
  - Price of a used car based on its features (attributes, characteristics



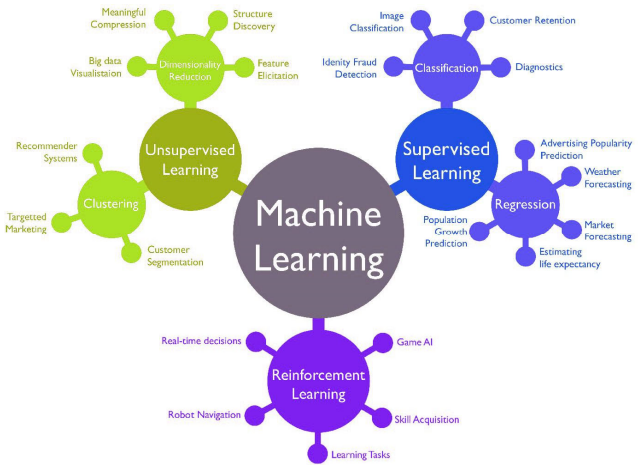
Supervised Learning



- Applications:
  - Text categorization
  - Face Detection
  - Signature recognition
  - Customer discovery
  - Spam detection
  - Weather forecasting
  - Predicting prices
  - Stock price predictions

https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications, Abdul Wahid

Reinforcement Learning

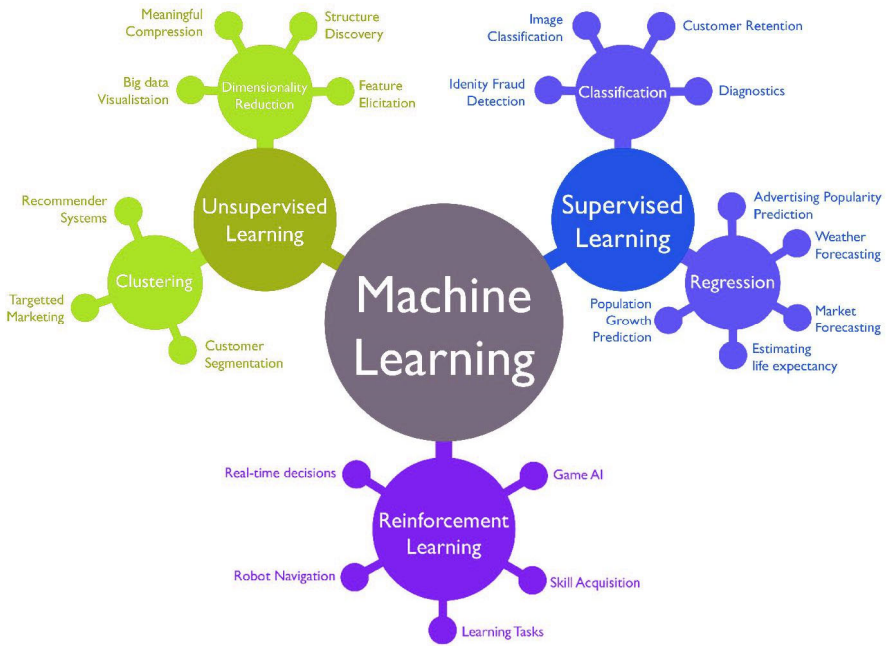


- Reward/Punishment would be applied
- It's good for robot navigation and online decision making

- RL is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.
- Reinforcement learning differs from supervised learning in not needing labelled input/output pairs be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead the focus is on finding a balance between exploration.

https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications, Abdul Wahid

Unsupervised Learning

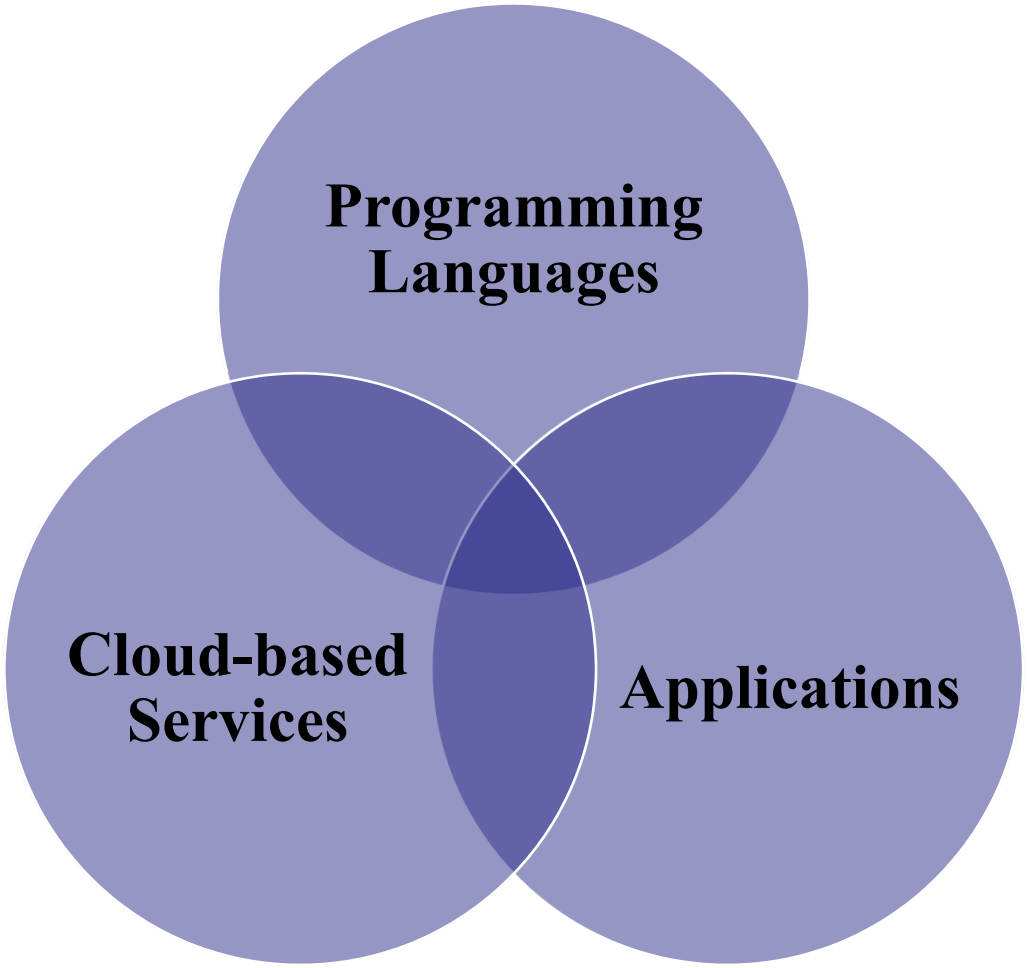


- UL is a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores/outputs for the training data.
- So, UL algorithms must first self-discover any naturally occurring patterns in that training data set.
- Common examples include clustering, where the algorithm automatically groups its training examples into categories with similar features, and principal component analysis, where the algorithm finds ways to compress the training data set by identifying which features are most useful.

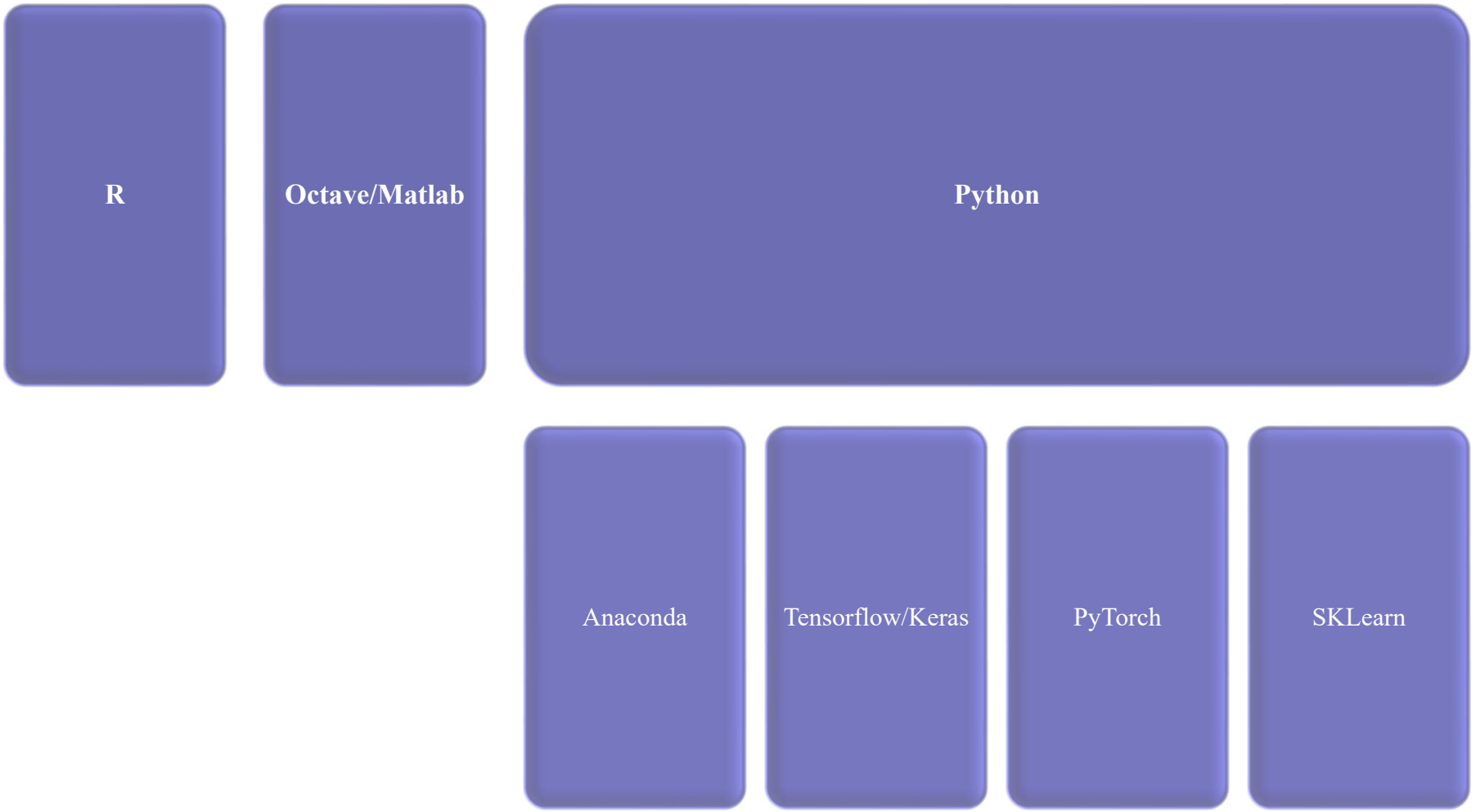
https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications, Abdul Wahid

Tools and Implementation

Approaches


Programming Languages
Cloud-based Services
Applications

Programming Languages



Applications

Orange

Rapidminer

H2O.ai

Cnvrg.io

Spell

Cloud-Based Services

Google Co-laboratory and Cloud AI platform

IBM Watson Machine Learning

Kaggle Kernel

Coclac

Microsoft Azure ML

Amazon AWS Machine Learning

UnSupervised Learning, Clustering

ANIMATIONS

NUS National University of Singapore | School of Computing

UnSupervised Learning, Clustering

Part 3

Rapidminer

What is RapidMiner?

RAPIDMINER

Depth for Data Scientists, Simplified for Everyone Else

Join the 40,000+ global organizations in every industry who use the RapidMiner data science platform to drive revenue, reduce costs, and avoid risk.

RapidMiner provides data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.

RapidMiner provides a GUI to design and execute analytical workflows.

WWW.RAPIDMINER.COM

RAPIDMINER

NUS National University of Singapore | School of Computing

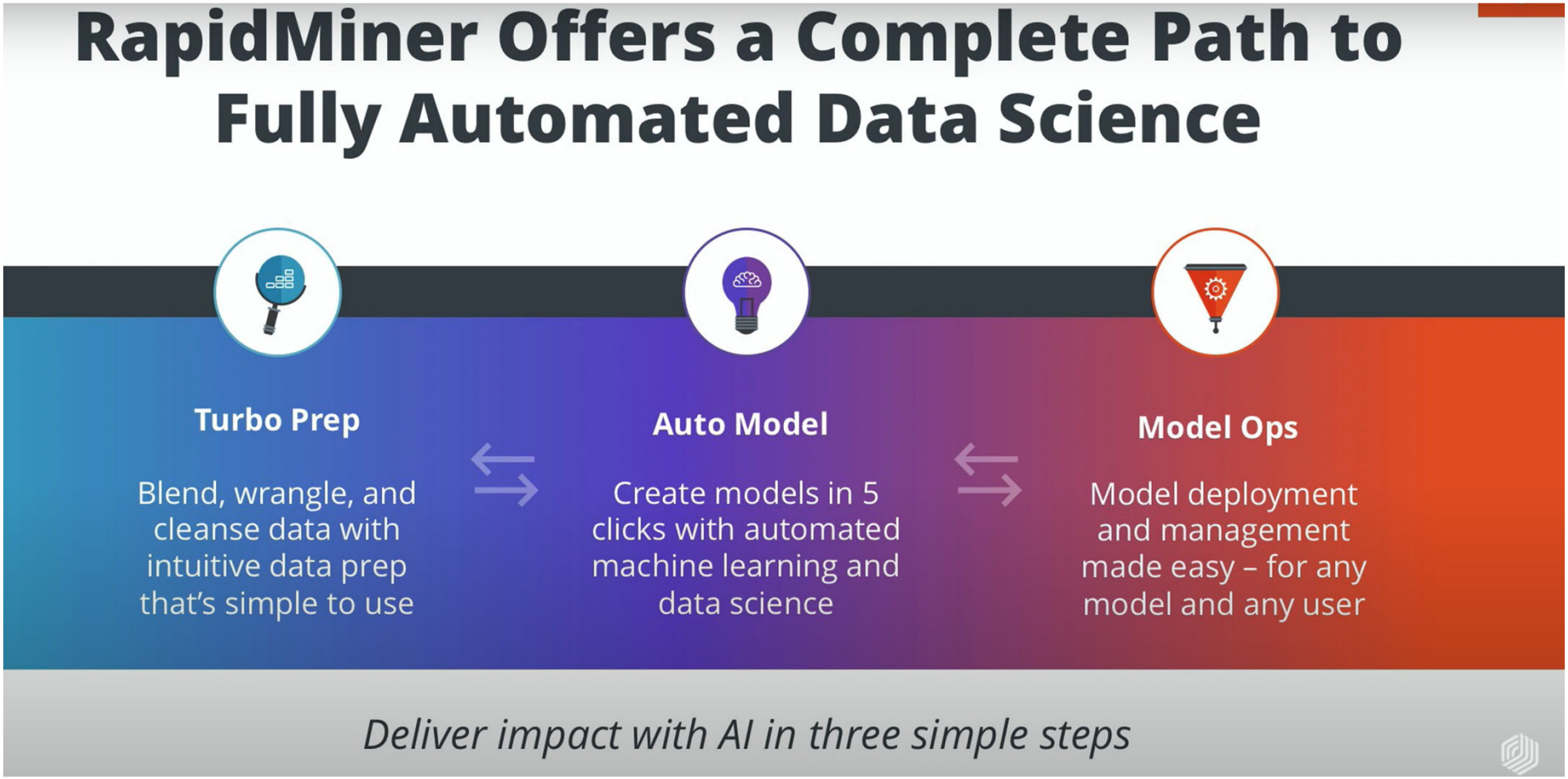
The Hard Way

- Slow to collaborate & transform
- Acquire highly specialized expertise
- Choose complexity or oversimplification
- Opaque, fragmented processes
- Commit to a specific ecosystem

The RapidMiner Way

- Transformational business impact
- Upskill your organization
- Depth for data scientists, simplified for everyone else
- Easy to trust, tune & explain
- Future-proof innovation, portability & extensibility

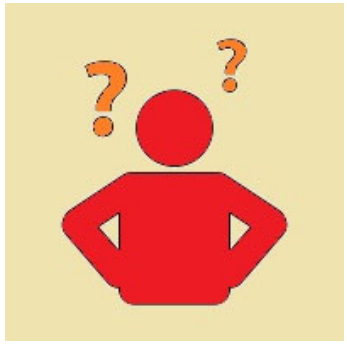
RAPIDMINER



WWW.RAPIDMINER.COM , WWW.YOUTUBE.COM RAPIDMINER

RAPIDMINER

Feedback



Rapidminer

FORRESTER®

RapidMiner is a Leader in The Forrester Wave: Multimodal Predictive Analytics & Machine Learning Solutions, Q3 2020

Gartner

RapidMiner is a Visionary in the 2021 Gartner Magic Quadrant for Data Science and Machine Learning Platforms

RapidMiner is a June 2020 Gartner Peer Insights Customers' Choice for Data Science and Machine Learning Platforms for the third time in a row

G2 CROWD

RapidMiner is the Highest Rated, Easiest to Use Data Science and Machine Learning Platform and was named a Leader in G2's Spring 2021 Report.

RAPIDMINER

Best Machine Learning Software in 2021

Name	Platform	Algorithms or Features
Scikit Learn	Linux, Mac OS, Windows, a Python/ R package	Classification, Regression, Clustering Pre-processing, Model Selection Dimensionality reduction.
PyTorch	Linux, Mac OS, Windows, a Python/ R package	Autograd Module, Optim Module, ANN Module
TensorFlow/ Keras	Linux, Mac OS, Windows, a Python/ R package	Provides a library for dataflow programming, ANN, Deep Learning
Shogun	Windows, Linux, Mac OS	Regression, Classification, Clustering Support vector machines, Dimensionality reduction Online learning etc.
Google Colab and Cloud AI Platform	Cloud Service	Ready to run Python and R programs.
Amazon Web Services (AWS)	Cloud Service	Based on SageMaker got a web-based visual interface, plus engines. Pay-as-you-go pricing
IBM Watson Machine Learning	Cloud Service	Contains Auto AI, Machine Learning Cloud, Machine Learning Server, and Machine Learning on IBM Cloud Pak for Data
MS Azure ML Studio	Cloud Service	An end-to-end solution, Supports open-source tools and frameworks, including PyTorch and TensorFlow

HTTPS://WWW.ITBUSINESSEDGE.COM/

RAPIDMINER

Best Machine Learning Software in 2021

Name	Platform	Algorithms or Features
Weka	Linux, Mac OS, Windows	Data preparation, Classification, Regression Clustering, Visualization, Association rules mining
KNIME	Linux, Mac OS, Windows	Can work with large data volume. Supports text mining & image mining through plugins
Apache Mahout	Cross-platform	Preprocessors, Regression, Clustering Recommenders, Distributed Linear Algebra.
Rapid Miner	Linux, Mac OS, Windows	Data loading & Transformation Data preprocessing & visualization.Regression, ANN, Deep Learning (simple) models. Speed is acceptable.
MATLAB/ Octave	Linux, Mac OS, Windows, A complete PL	Excellent performance in Modelling, Does not support other languages such as Python and R
Python	Linux, Mac OS, Windows, A complete PL	A solution from the scratch, sometimes implementation is lengthy and complicated.

Installation

HOW TO START?

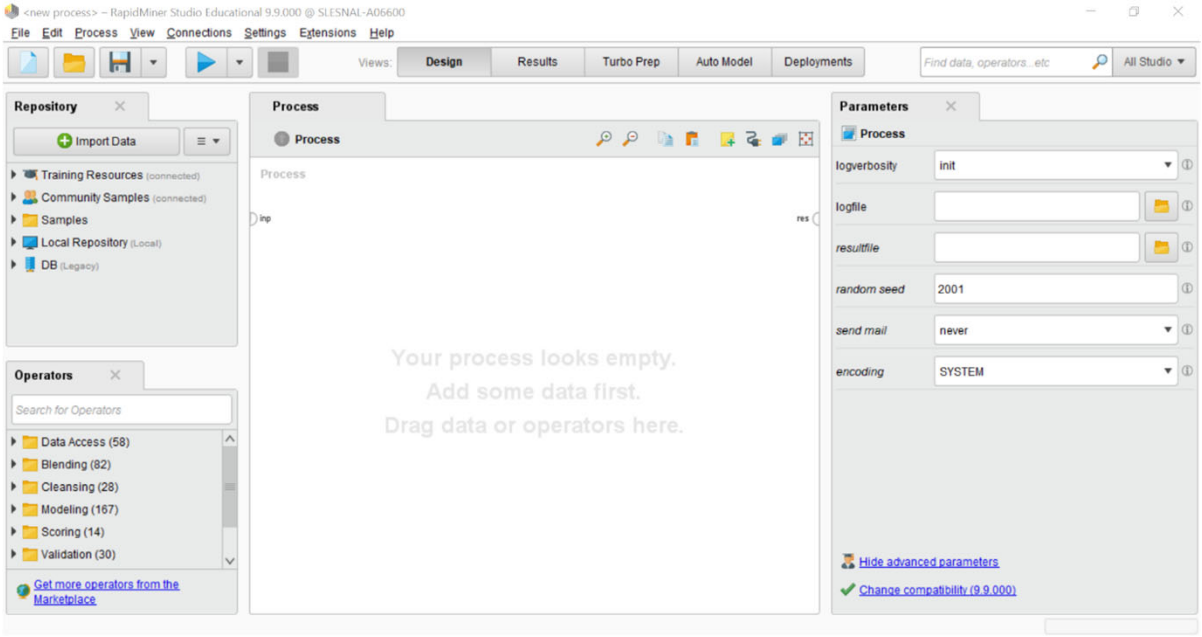
Install it

The basic **RapidMiner Studio** is free.

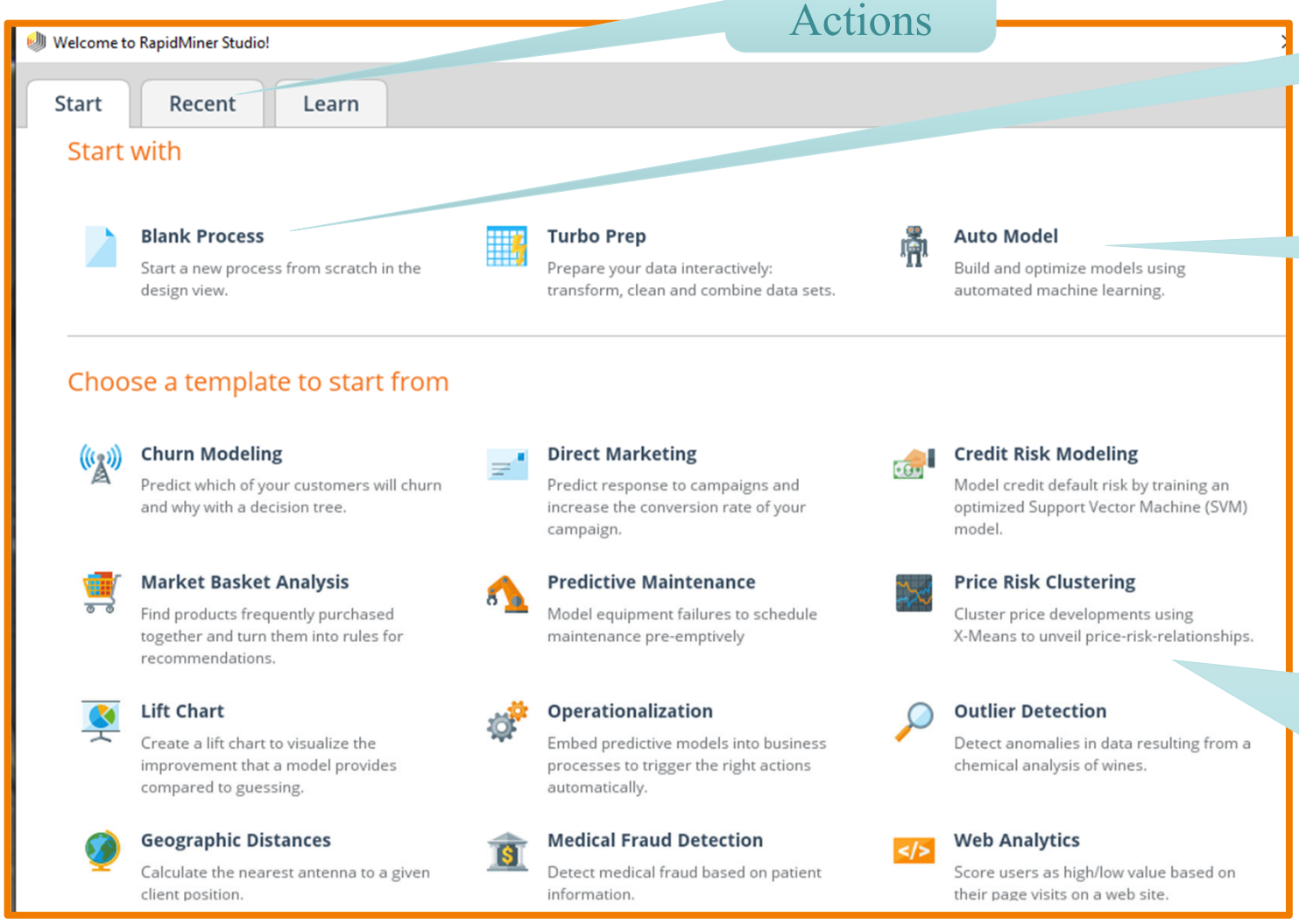
Run it !!!

Mind your OS

www.rapidminer.com



HOW TO START?



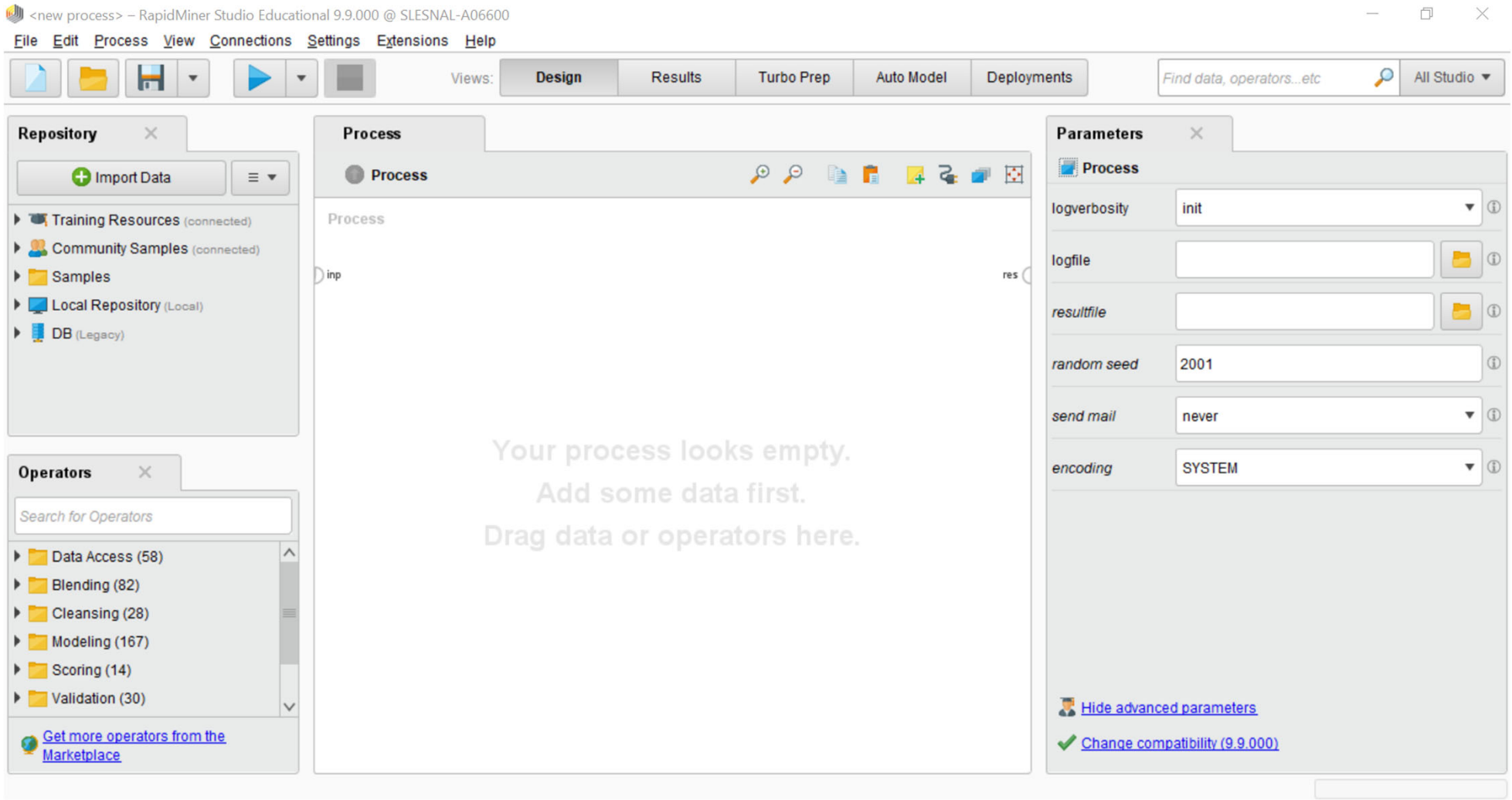
Recent Actions

Start with this

Advanced mode

Ready to use ML and DA Templates. We'll see these later

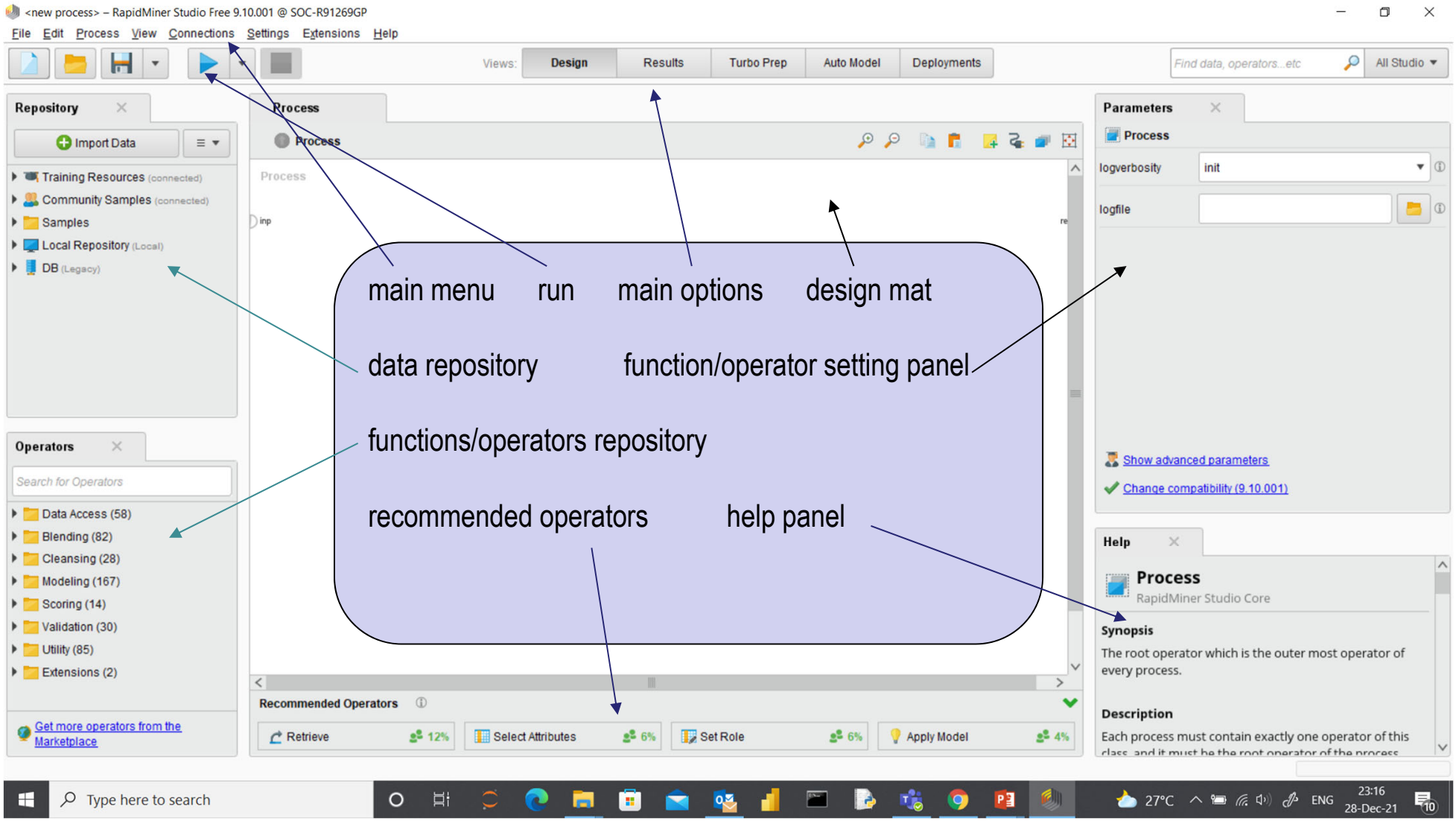
HOW TO START?



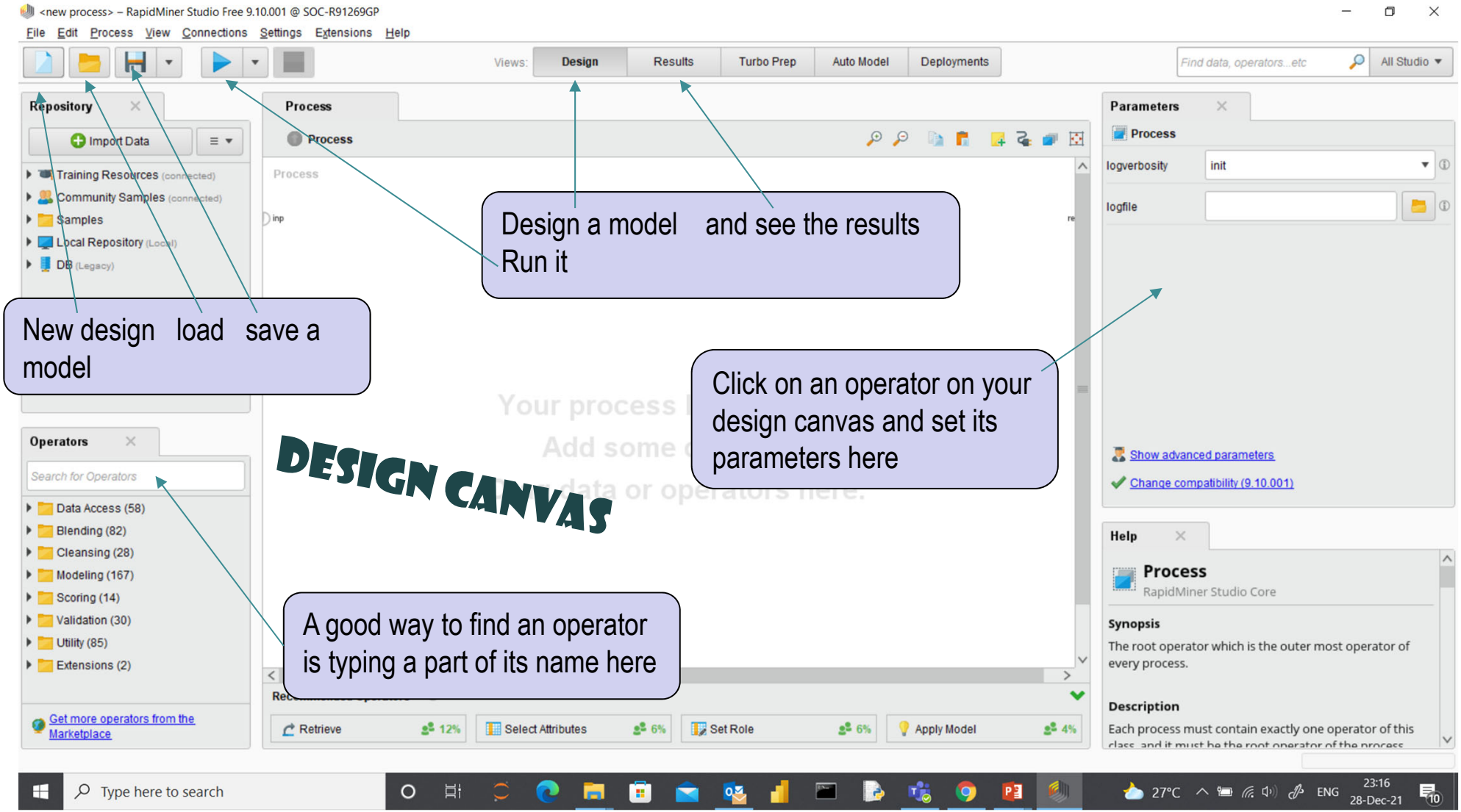
RAPIDMINER MAIN PAGE, WAIT TO SEE THAT.

Environment & Libraries

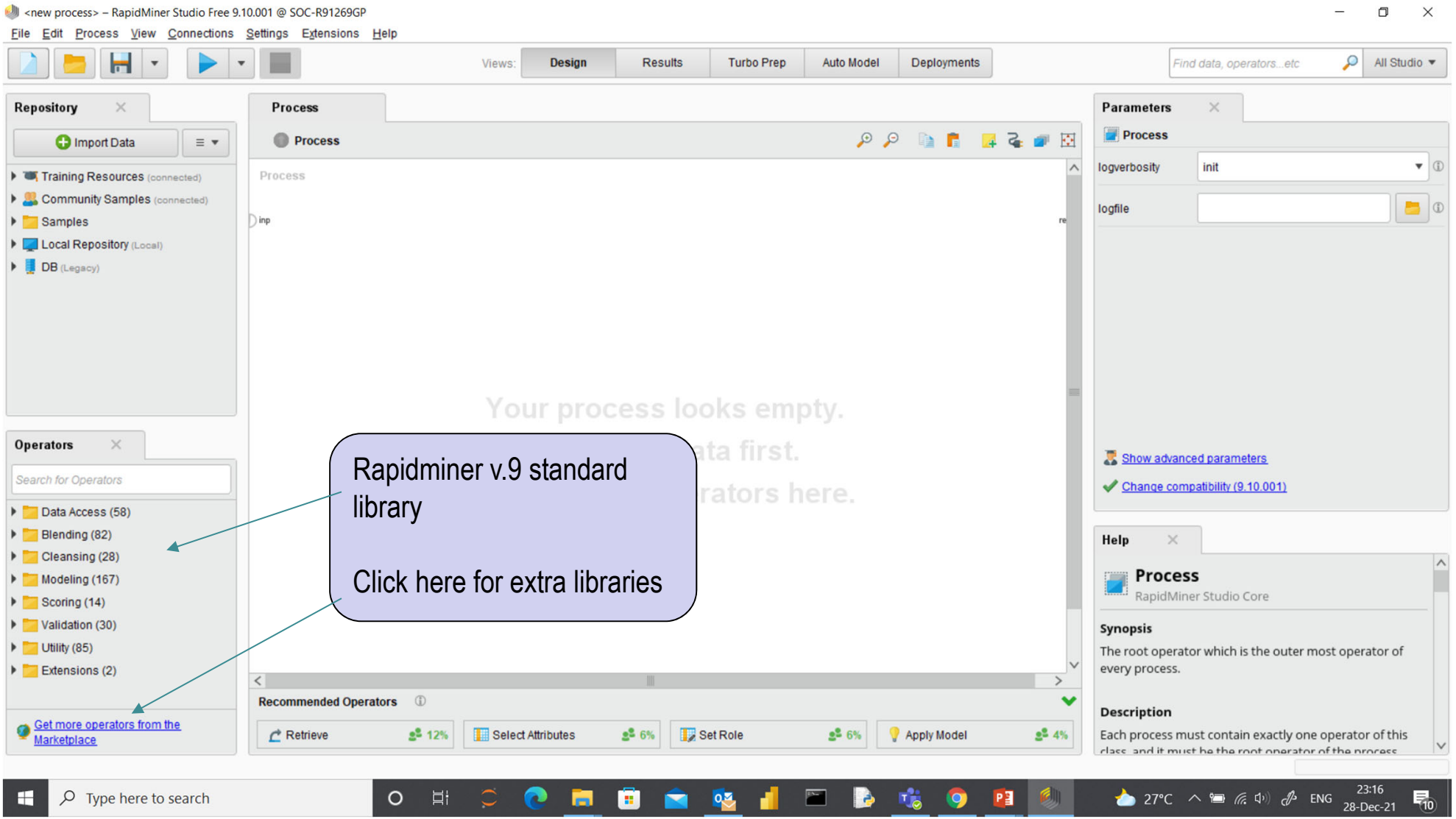
RAPIDMINER MAIN PAGE



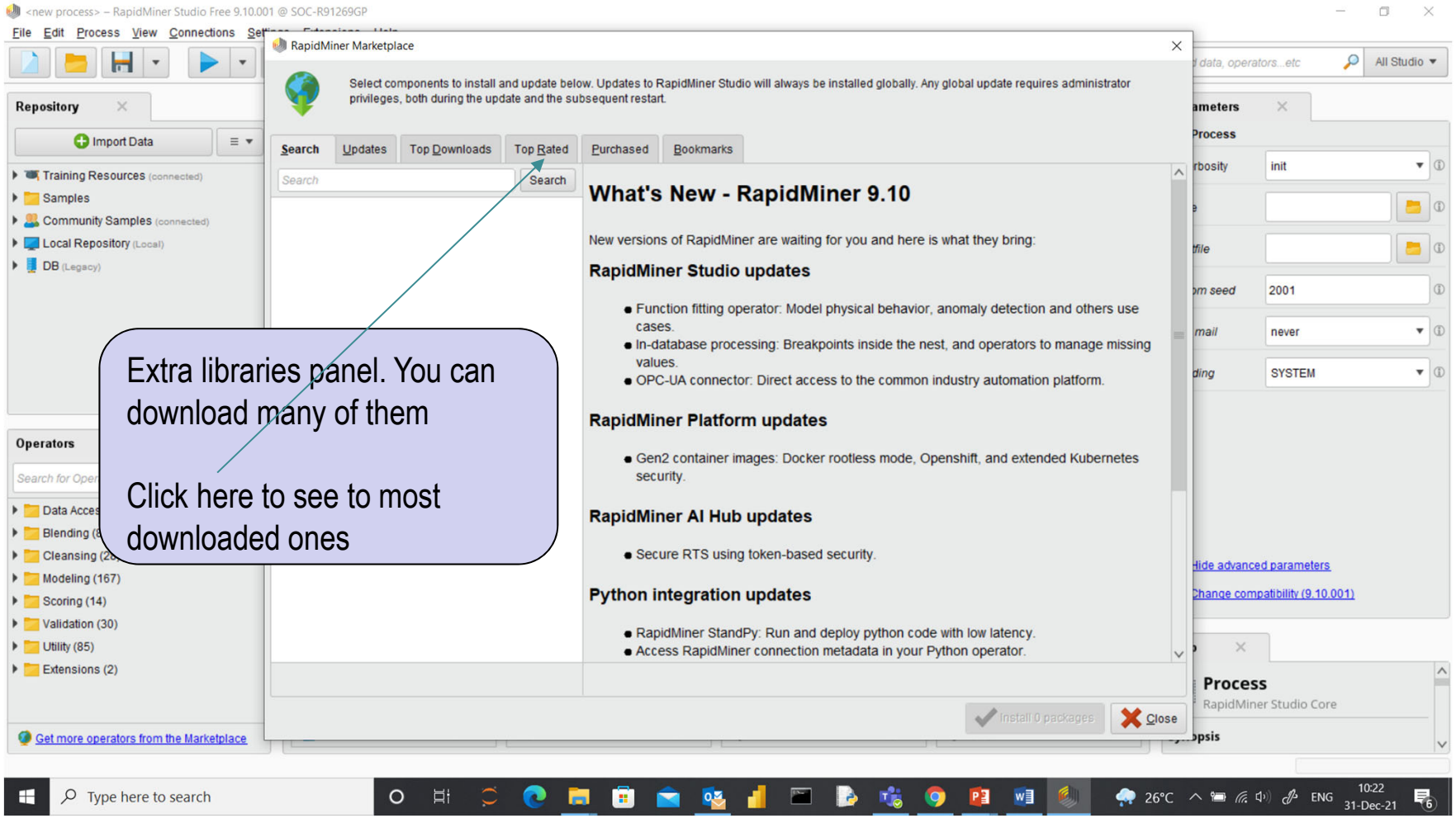
Rapidminer Main page



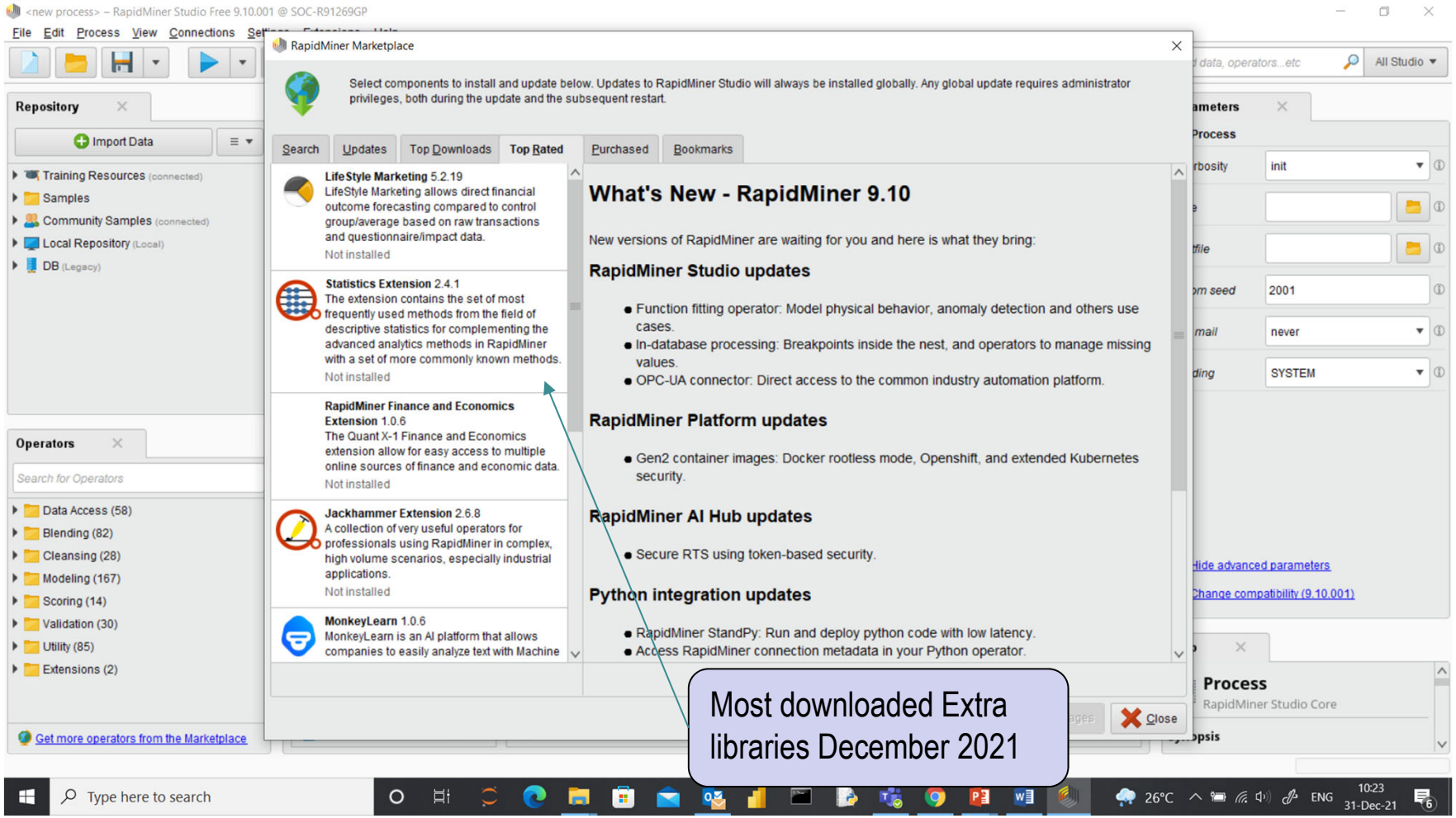
Libraries


Rapidminer v.9 standard library
Click here for extra libraries

Libraries


RapidMiner Marketplace
Select components to install and update below. Updates to RapidMiner Studio will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.
Search
Updates
Top Downloads
Top Rated
Purchased
Bookmarks
What's New - RapidMiner 9.10
New versions of RapidMiner are waiting for you and here is what they bring:
RapidMiner Studio updates
Function fitting operator: Model physical behavior, anomaly detection and others use cases.
In-database processing: Breakpoints inside the nest, and operators to manage missing values.
OPC-UA connector: Direct access to the common industry automation platform.
RapidMiner Platform updates
Gen2 container images: Docker rootless mode, Openshift, and extended Kubernetes security.
RapidMiner AI Hub updates
Secure RTS using token-based security.
Python integration updates
RapidMiner StandPy: Run and deploy python code with low latency.
Access RapidMiner connection metadata in your Python operator.
Install 0 packages
Close
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
Extra libraries panel. You can download many of them
Click here to see to most downloaded ones

Libraries


RapidMiner Marketplace
Select components to install and update below. Updates to RapidMiner Studio will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.
Search
Updates
Top Downloads
Top Rated
Purchased
Bookmarks
LifeStyle Marketing 5.2.19
LifeStyle Marketing allows direct financial outcome forecasting compared to control group/average based on raw transactions and questionnaire/impact data.
Not installed
Statistics Extension 2.4.1
The extension contains the set of most frequently used methods from the field of descriptive statistics for complementing the advanced analytics methods in RapidMiner with a set of more commonly known methods.
Not installed
RapidMiner Finance and Economics Extension 1.0.6
The Quant X-1 Finance and Economics extension allow for easy access to multiple online sources of finance and economic data.
Not installed
Jackhammer Extension 2.6.8
A collection of very useful operators for professionals using RapidMiner in complex, high volume scenarios, especially industrial applications.
Not installed
MonkeyLearn 1.0.6
MonkeyLearn is an AI platform that allows companies to easily analyze text with Machine
What's New - RapidMiner 9.10
New versions of RapidMiner are waiting for you and here is what they bring:
RapidMiner Studio updates
Function fitting operator: Model physical behavior, anomaly detection and others use cases.
In-database processing: Breakpoints inside the nest, and operators to manage missing values.
OPC-UA connector: Direct access to the common industry automation platform.
RapidMiner Platform updates
Gen2 container images: Docker rootless mode, Openshift, and extended Kubernetes security.
RapidMiner AI Hub updates
Secure RTS using token-based security.
Python integration updates
RapidMiner StandPy: Run and deploy python code with low latency.
Access RapidMiner connection metadata in your Python operator.
Close
Most downloaded Extra libraries December 2021
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
All Studio
Process
RapidMiner Studio Core
Type here to search
26°C
ENG
10:23
31-Dec-21

Practice 2, Later at Home

- Install Rapidminer
- Get familiar with it
- Try to test a few operators listed below
  - FP-Growth
  - Apply Model
  - Detect Outliers (Distances)
- Try to install the most popular extra library.

Part 4

Clustering: K-Means

K-Means Clustering

K-MEANS CLUSTERING

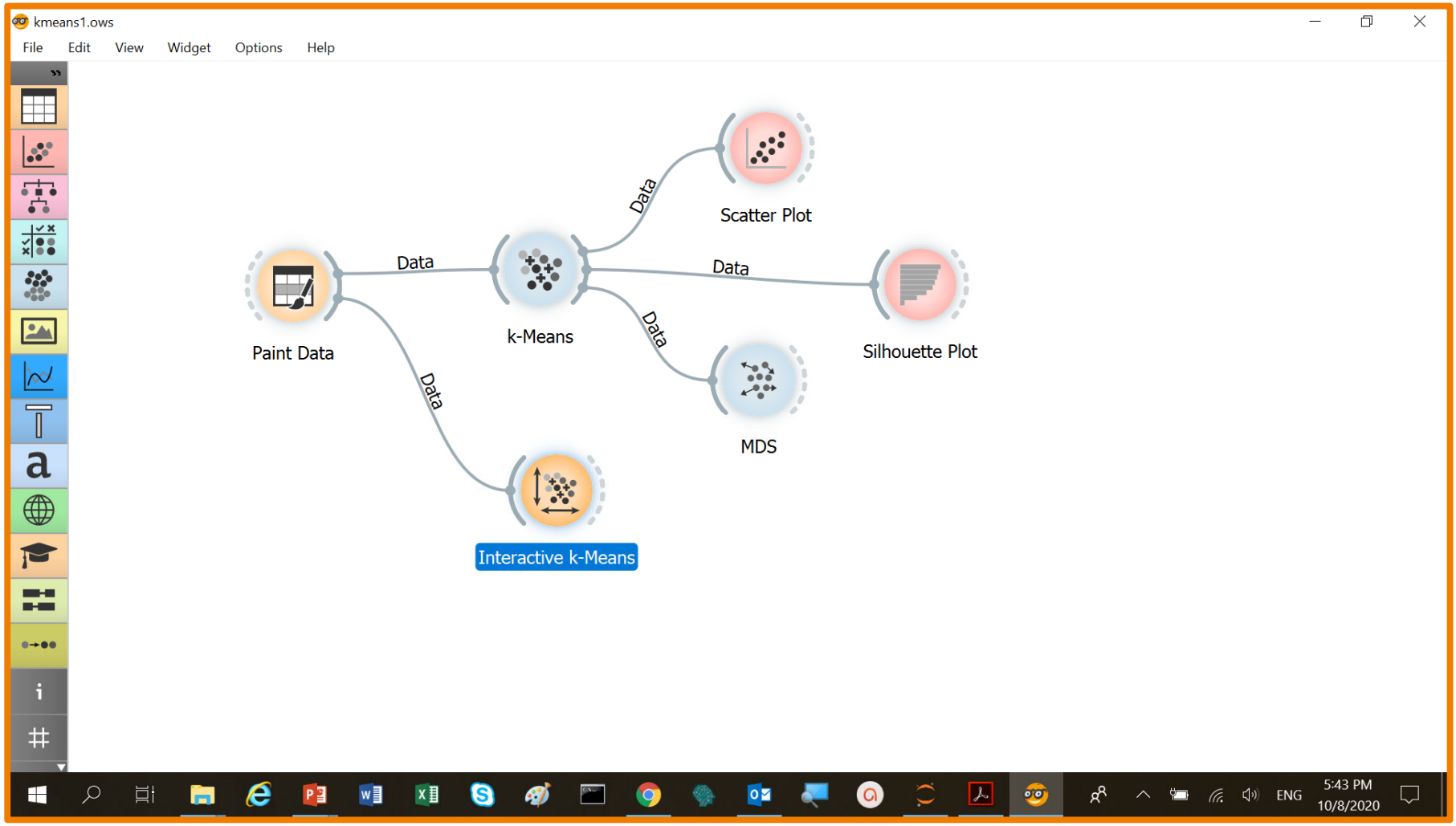
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).
- In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

K-MEANS CLUSTERING

NUS National University of Singapore | School of Computing

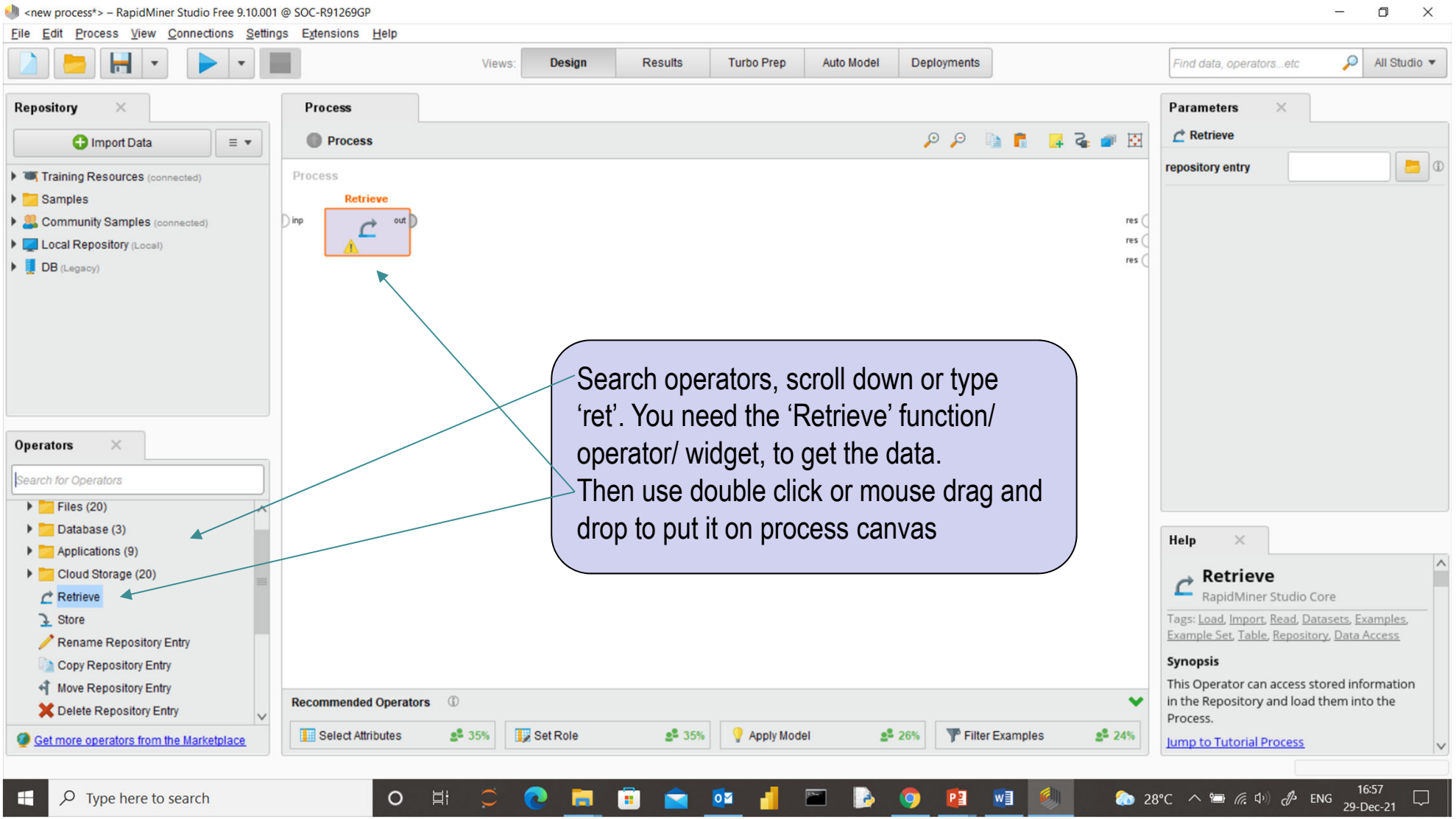
- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.
- K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- It will try to minimize intra-cluster distances, and maximize inter-cluster distances, using a MSD or MAD schemes.

Visualization of K-Means in Orange

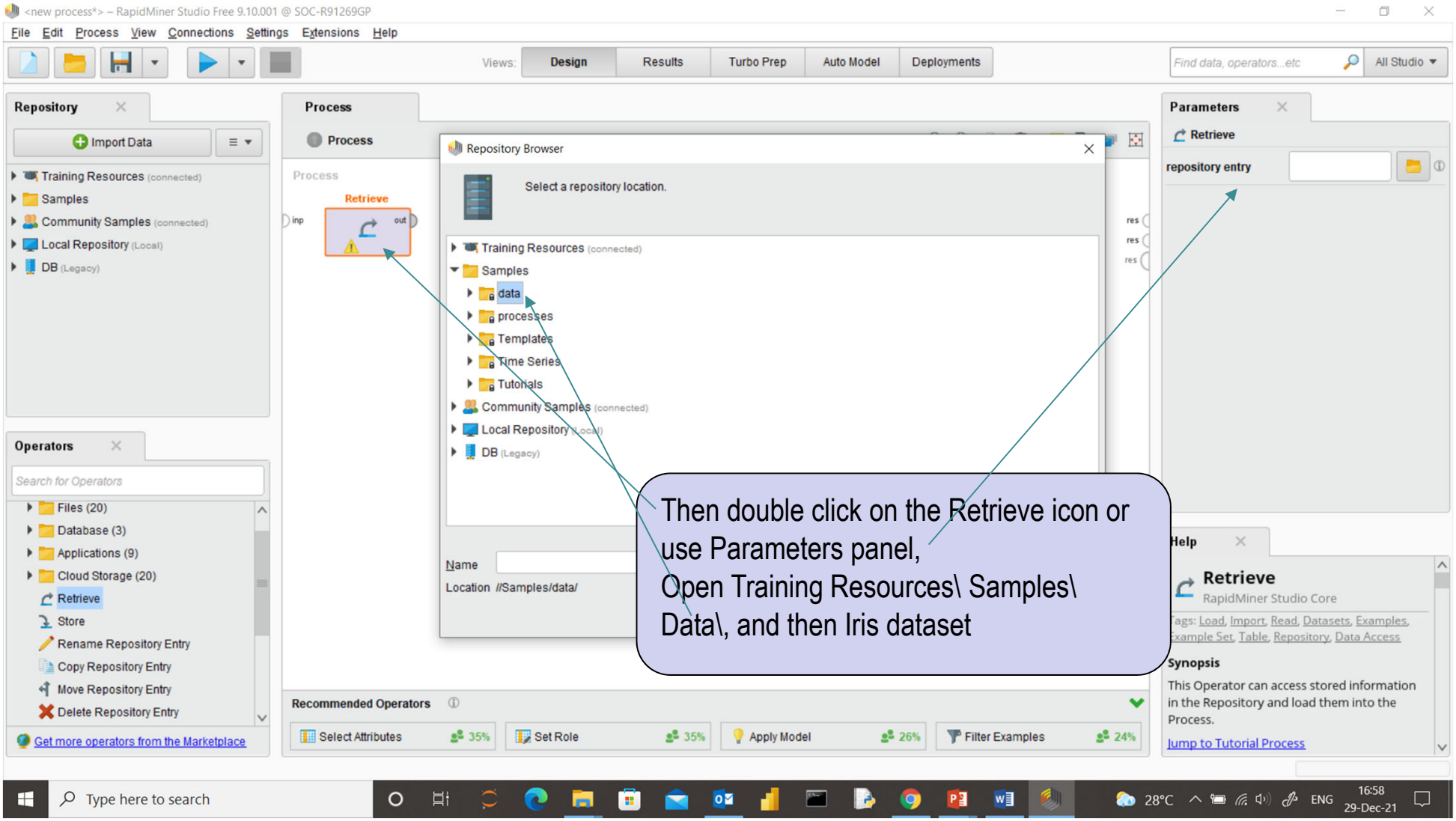


Example 1

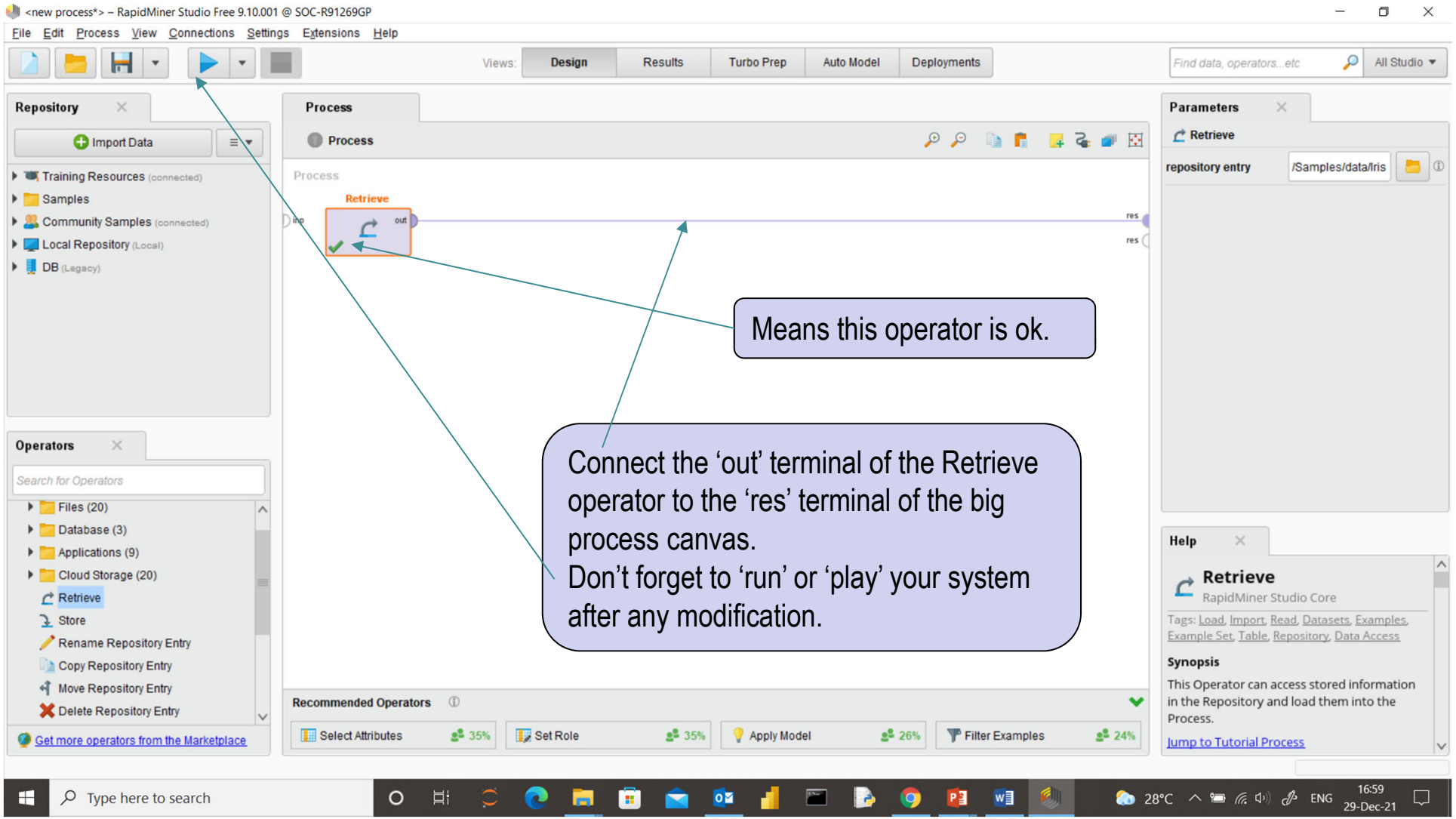
Example 1: Bunch of Iris Flowers



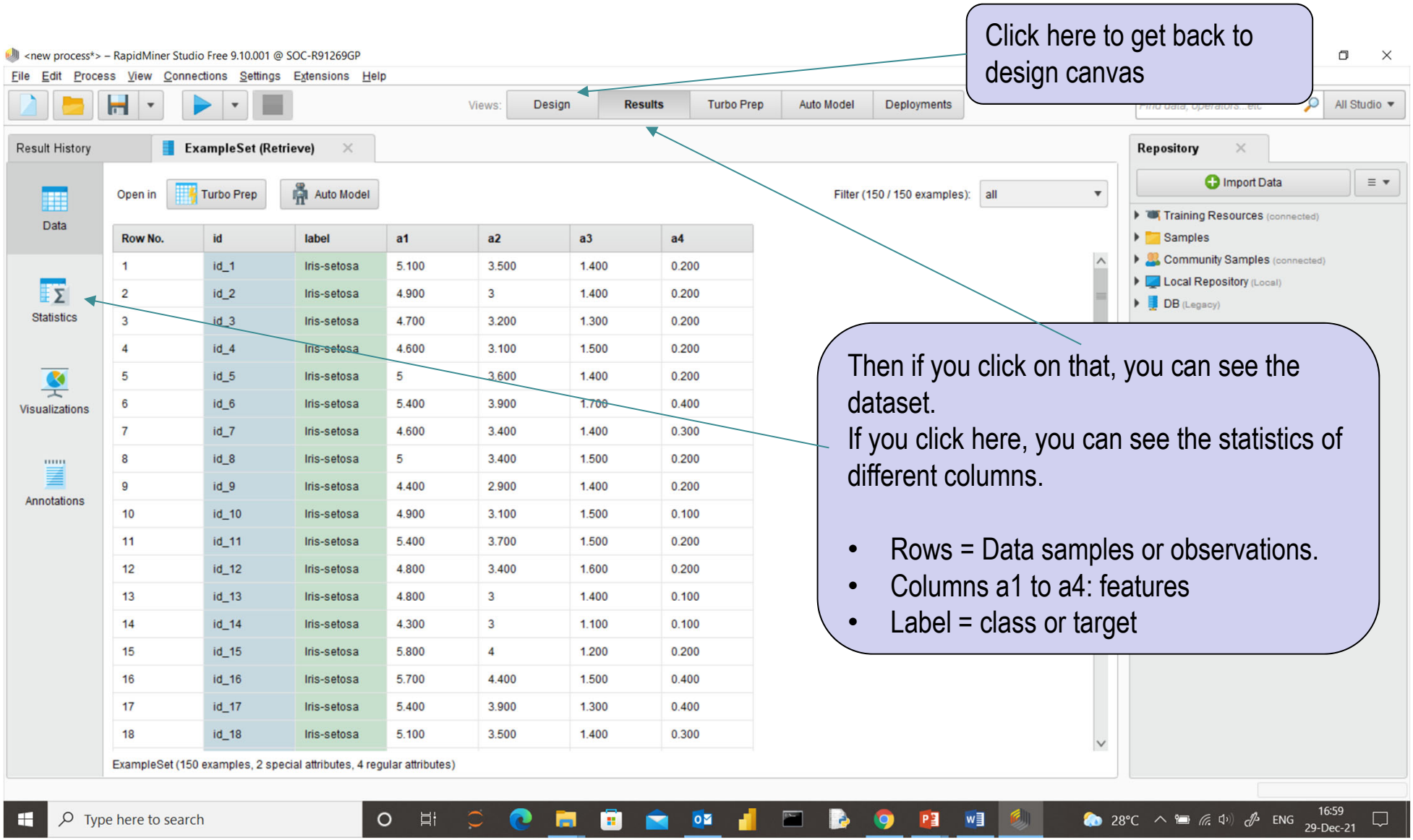
Example 1: Bunch of Iris Flowers



Example 1: Bunch of Iris Flowers



Example 1: Bunch of Iris Flowers

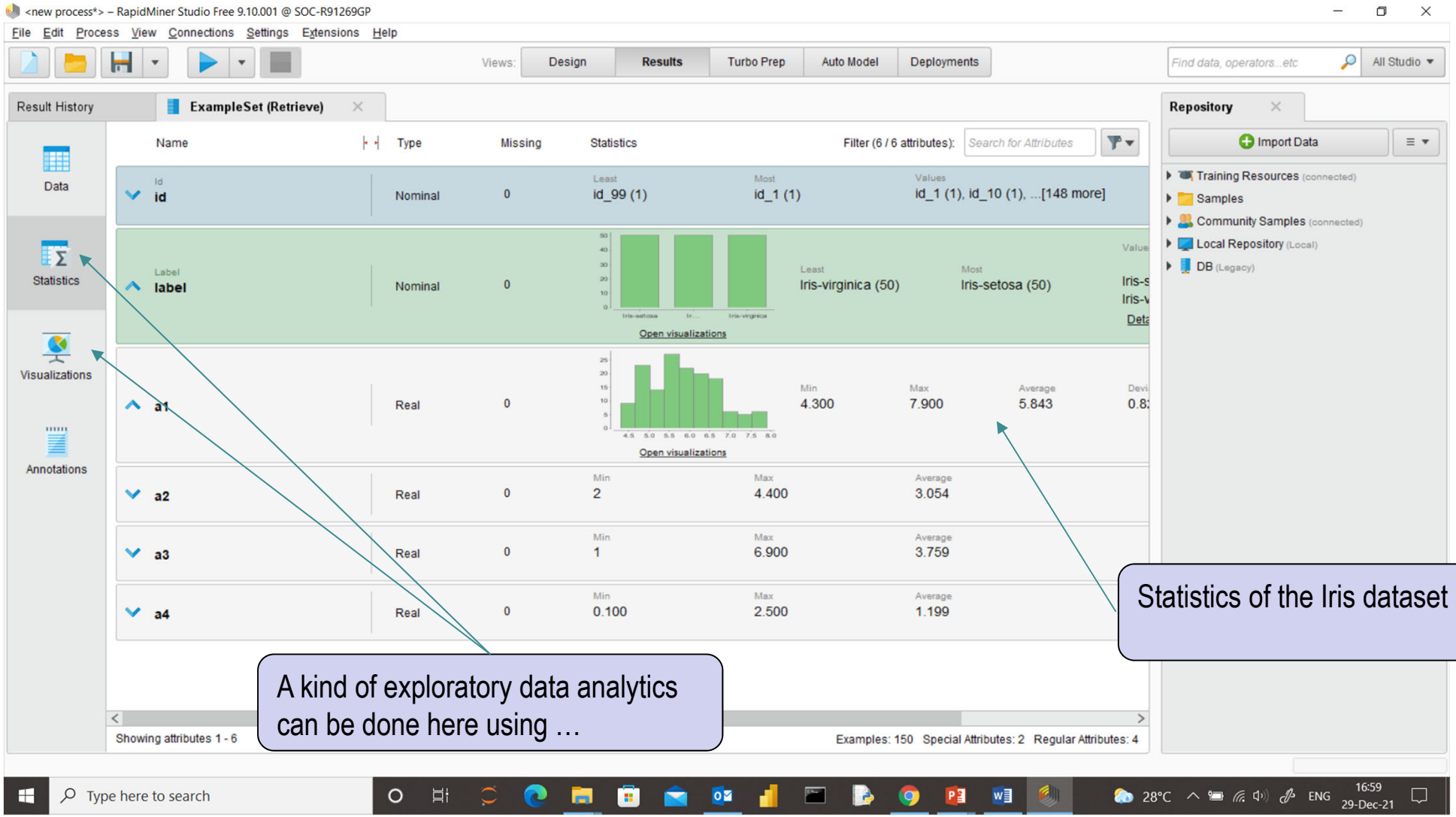


Click here to get back to design canvas

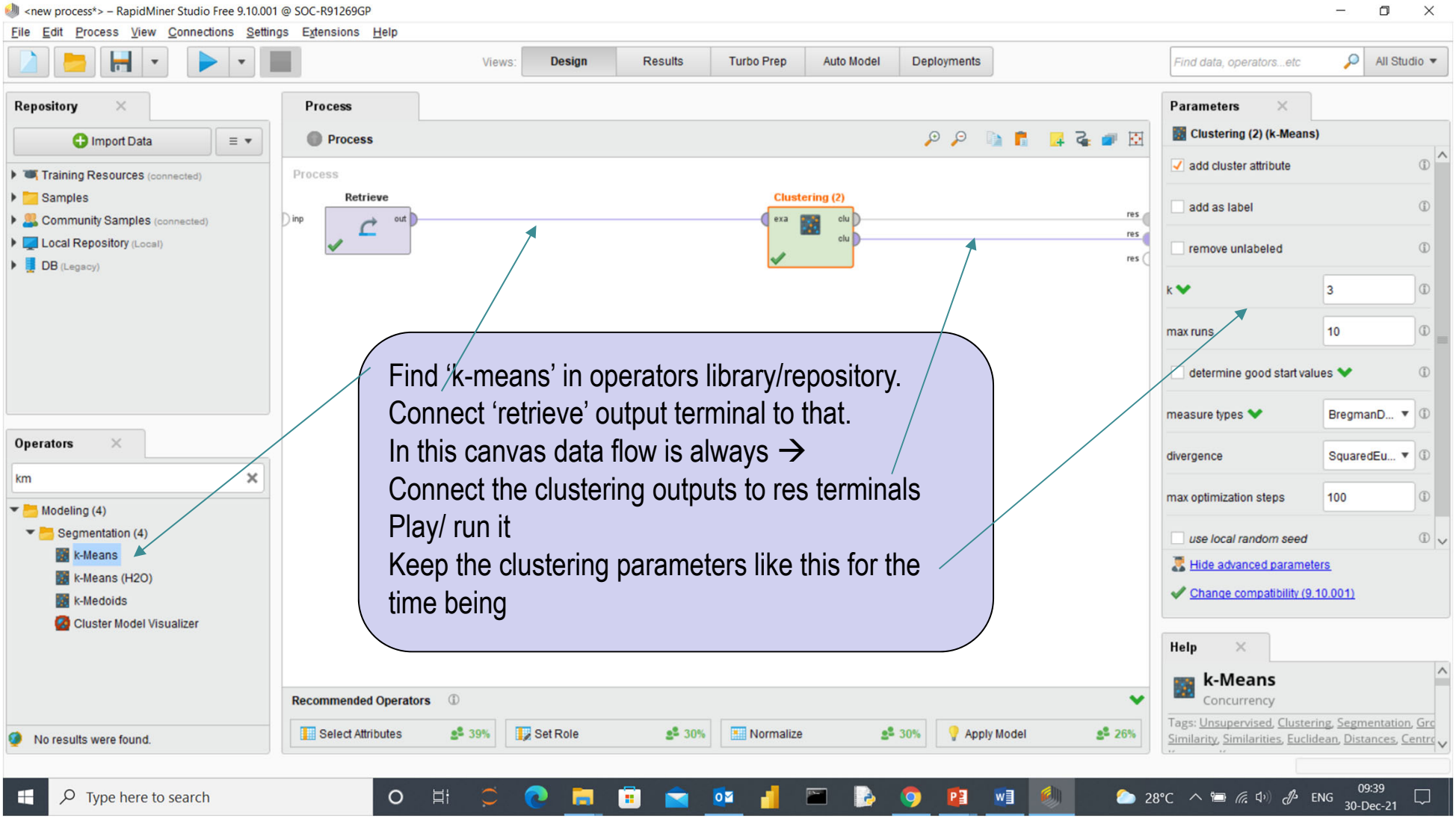
Then if you click on that, you can see the dataset.
If you click here, you can see the statistics of different columns.

- Rows = Data samples or observations.
- Columns a1 to a4: features
- Label = class or target

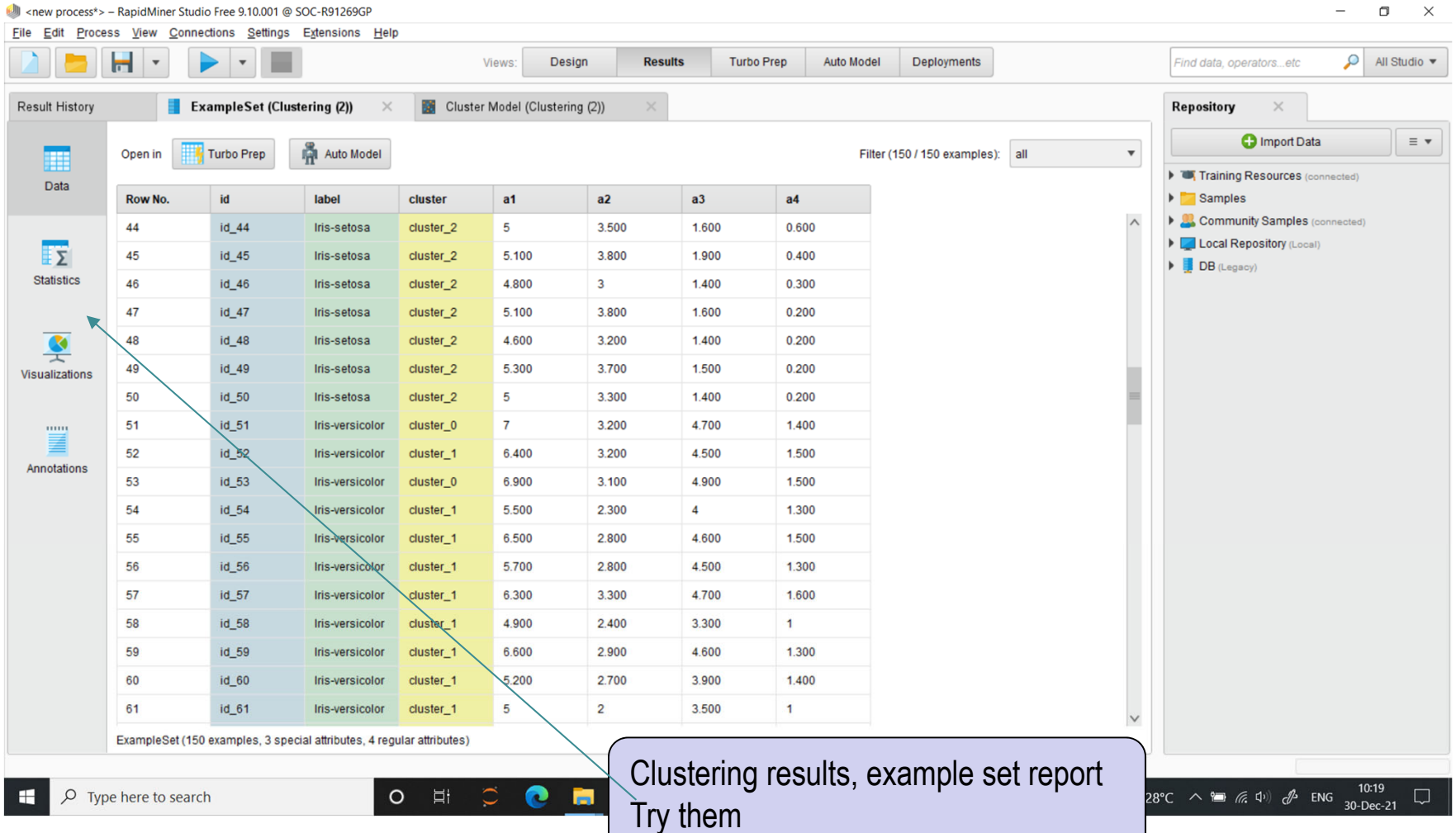
Example 1: Bunch of Iris Flowers



Example 1: Clustering

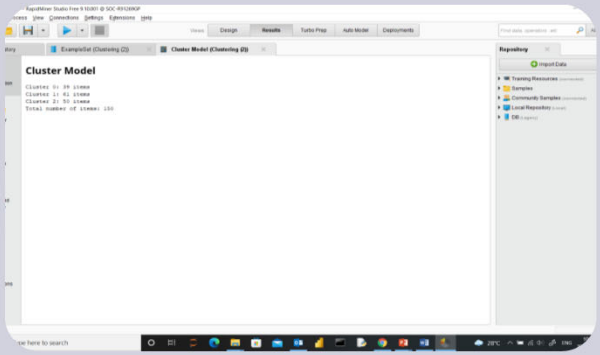


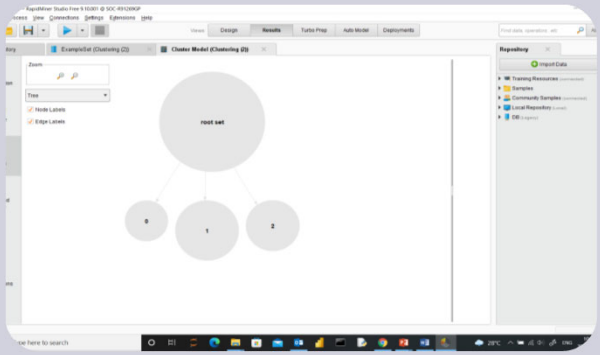
Example 1: Clustering Results

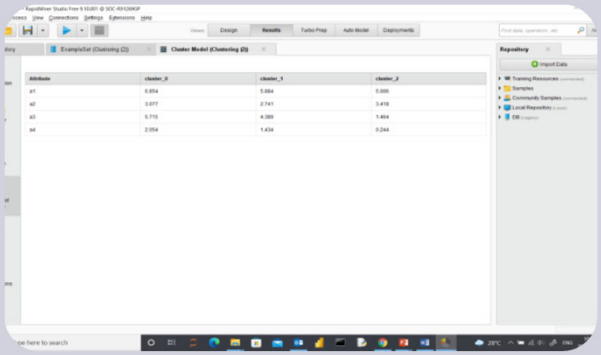


Example 1: Clustering Results

Clustering results, cluster model report







Description: clustering overview

Graph: structure of clusters

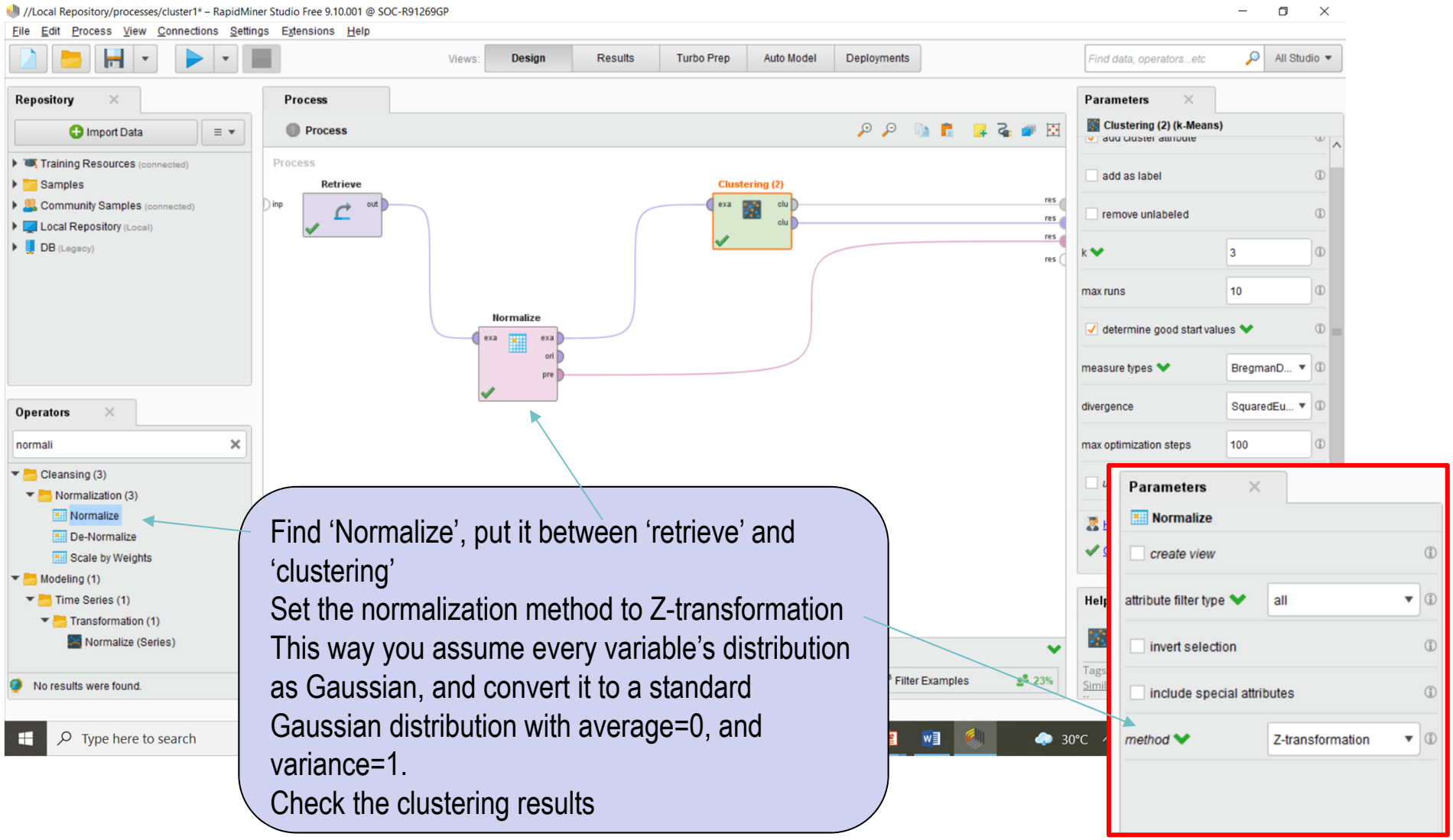
Centroids Table: final coordination of centroids in this 4d feature space

Practice 3

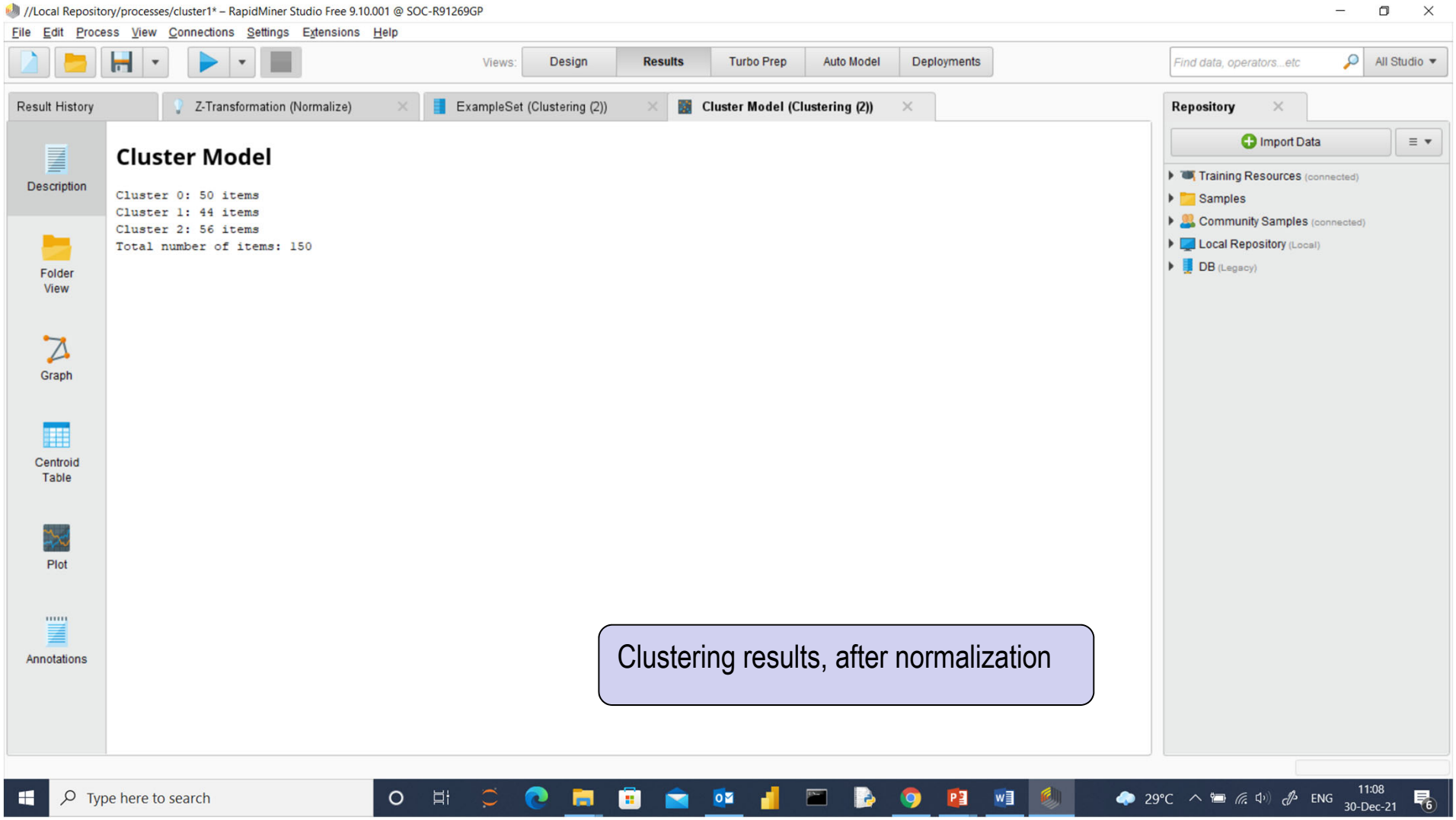
- **Design a k-means clustering system**
  - Use Iris dataset
  - Use another optional dataset
  - Try different k=2, 4, and 6
- **Visualize the results**
- **Discuss the results**
- **Report the results**

Improvement

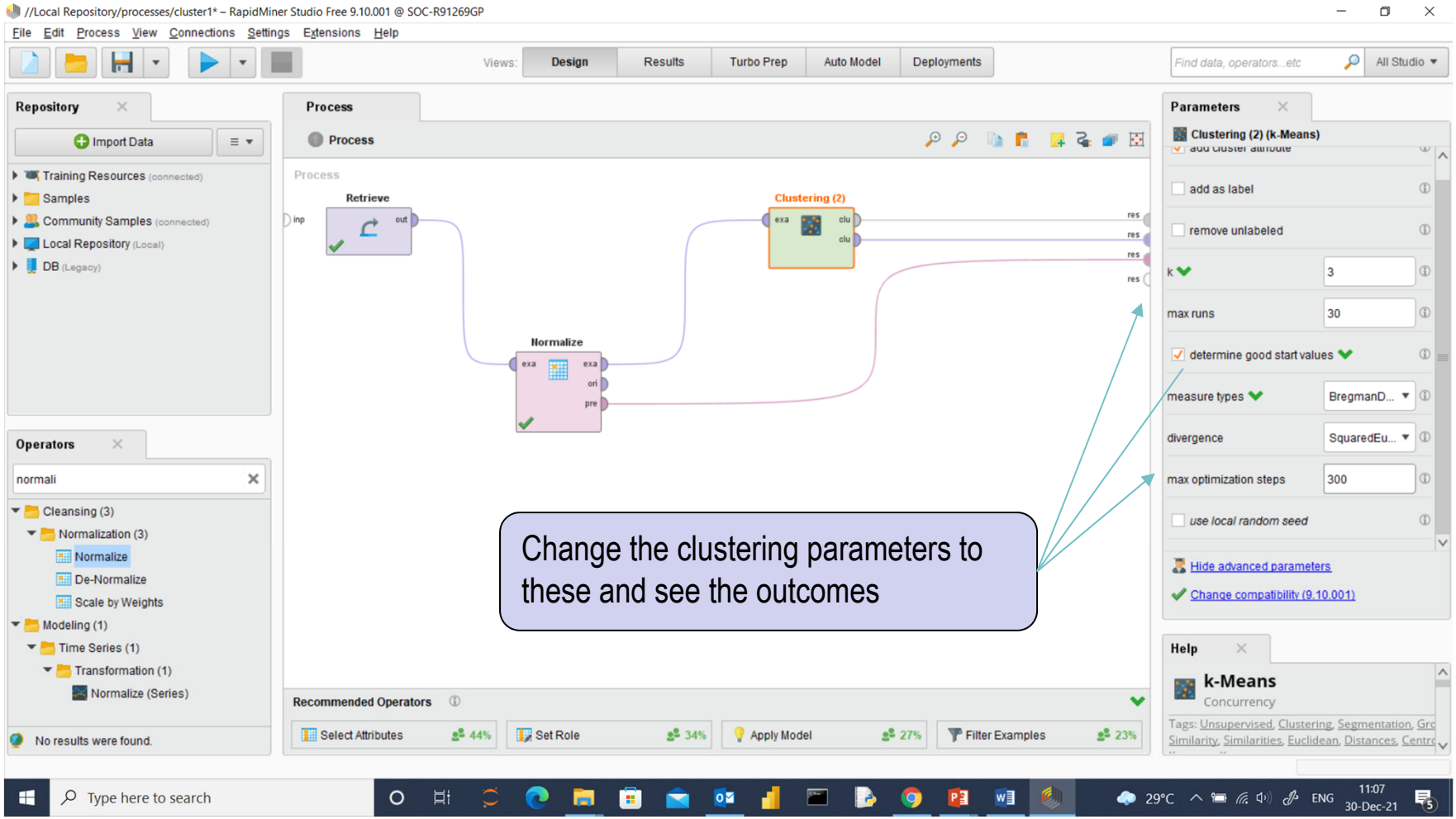
Example 1 with normalization



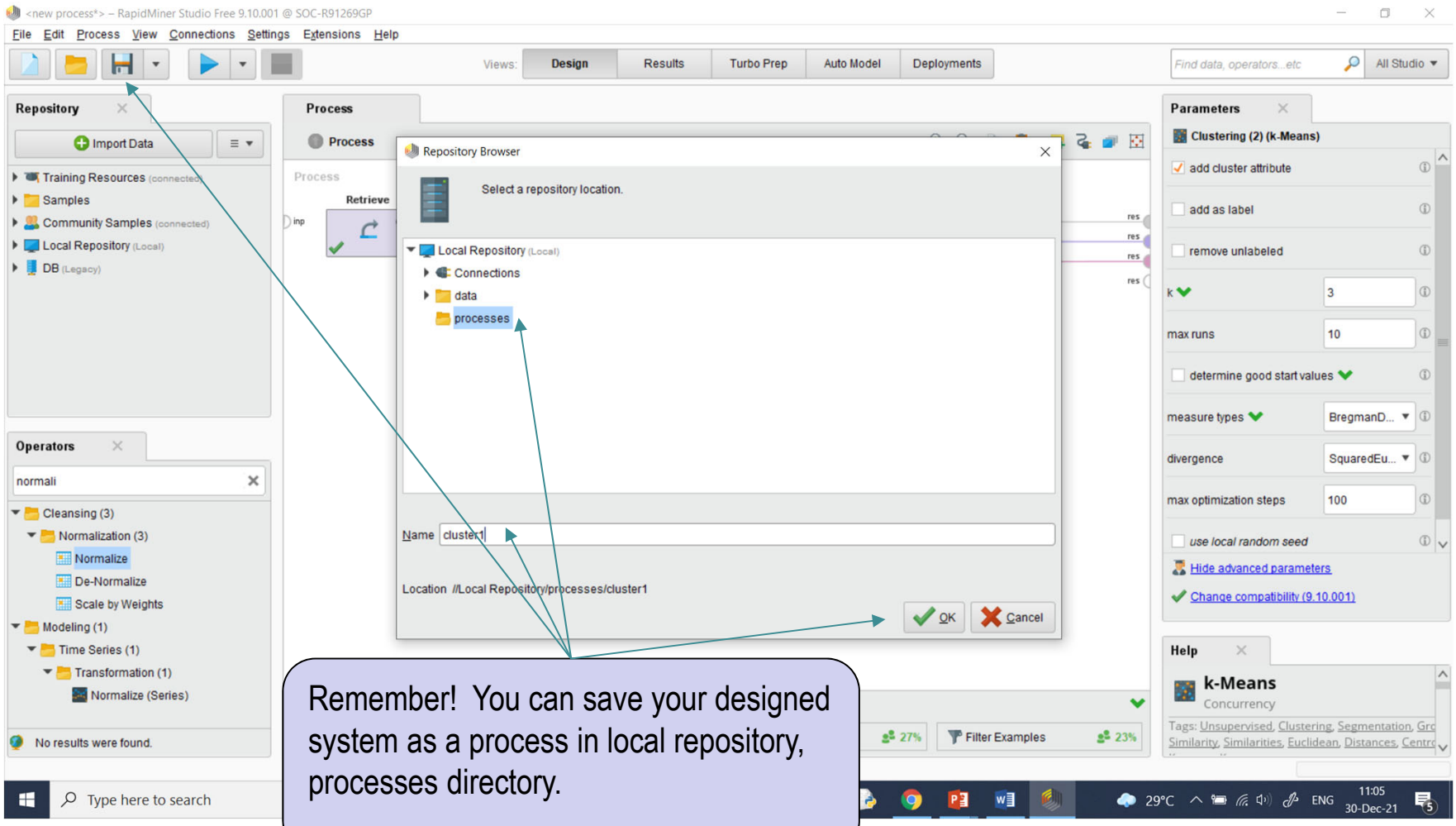
Example 1 with normalization



Example 1, Improvement



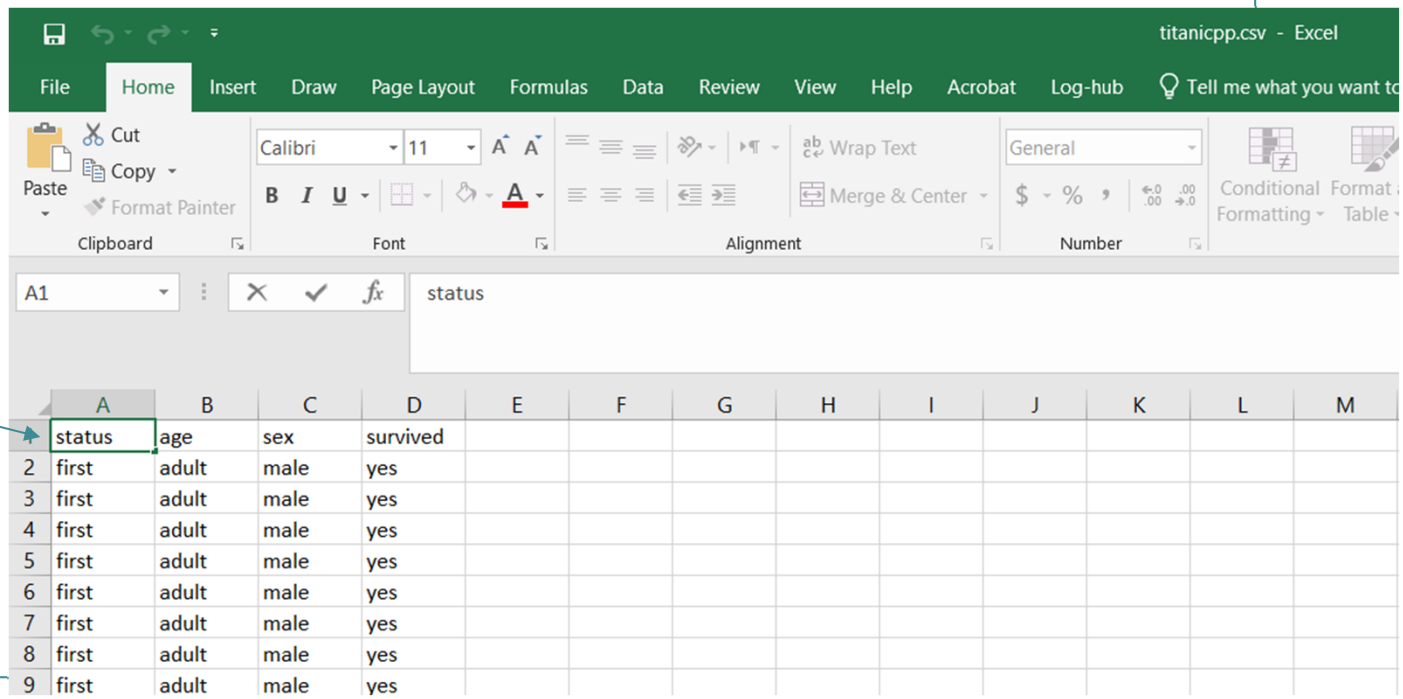
Save your System



Practice 4: Analyze the Titanic Disaster

- You can find titanicpp.csv data file in your examples directory
- This is a comma separated values dataset contains some information about the Titanic passengers and crews
- We have 2201

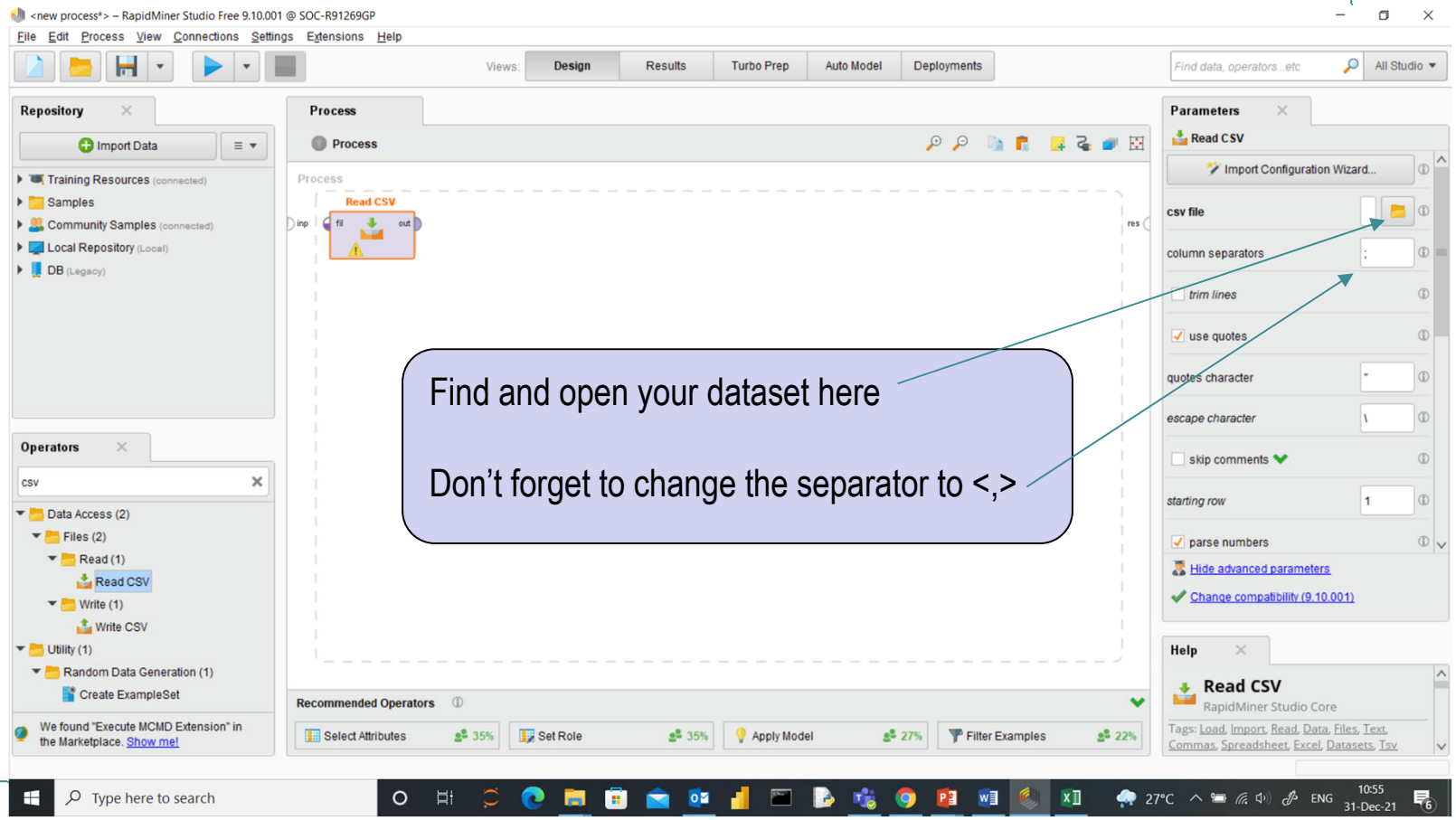
Data samples of Crews and Passengers, and 4 Variables



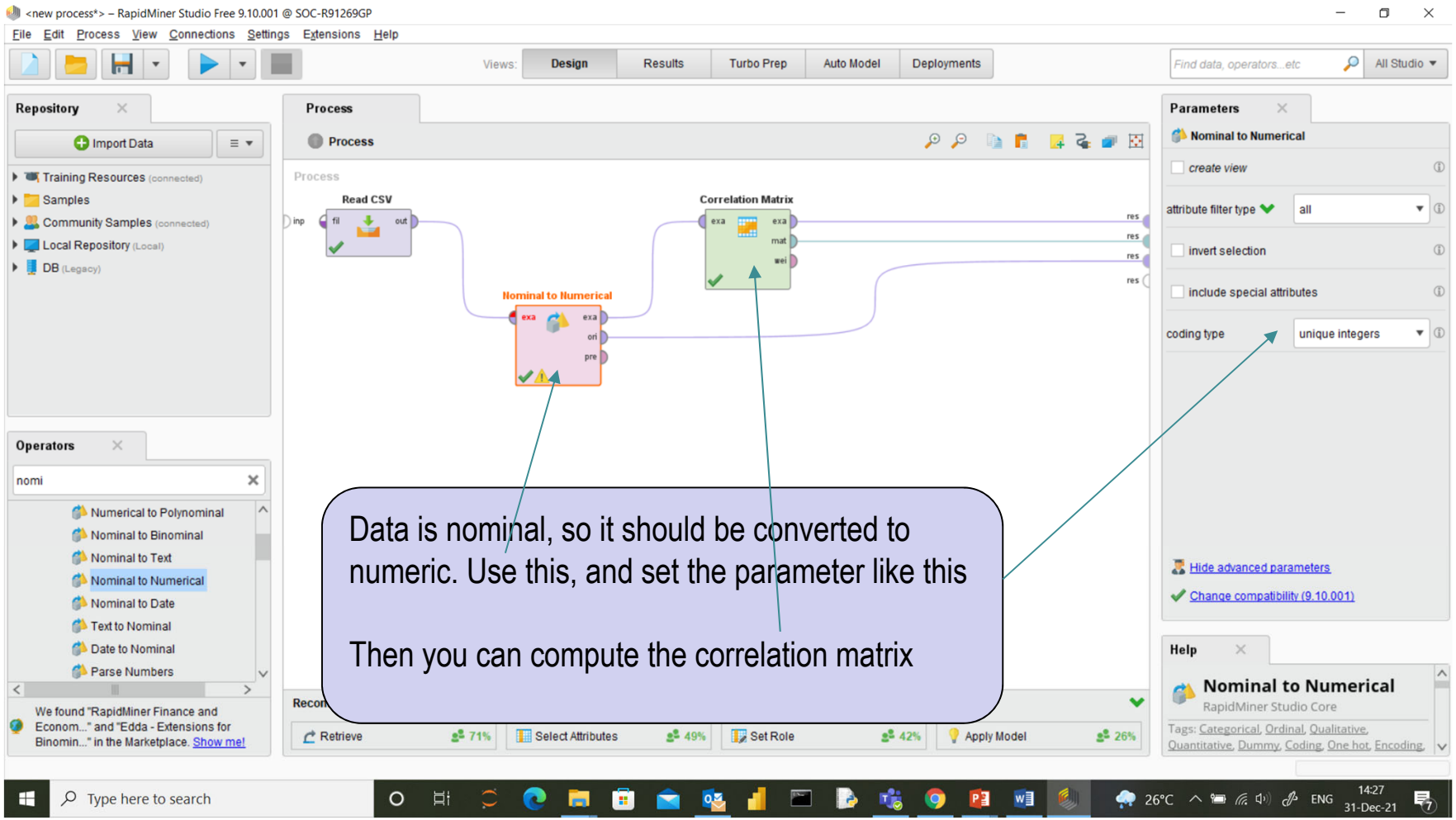
	A	B	C	D
1	status	age	sex	survived
2	first	adult	male	yes
3	first	adult	male	yes
4	first	adult	male	yes
5	first	adult	male	yes
6	first	adult	male	yes
7	first	adult	male	yes
8	first	adult	male	yes
9	first	adult	male	yes

Practice 4 : Analyze the Titanic Disaster

- How to read the dataset? Find and use Read CSV operator.



Practice 4 : Analyze the Titanic Disaster

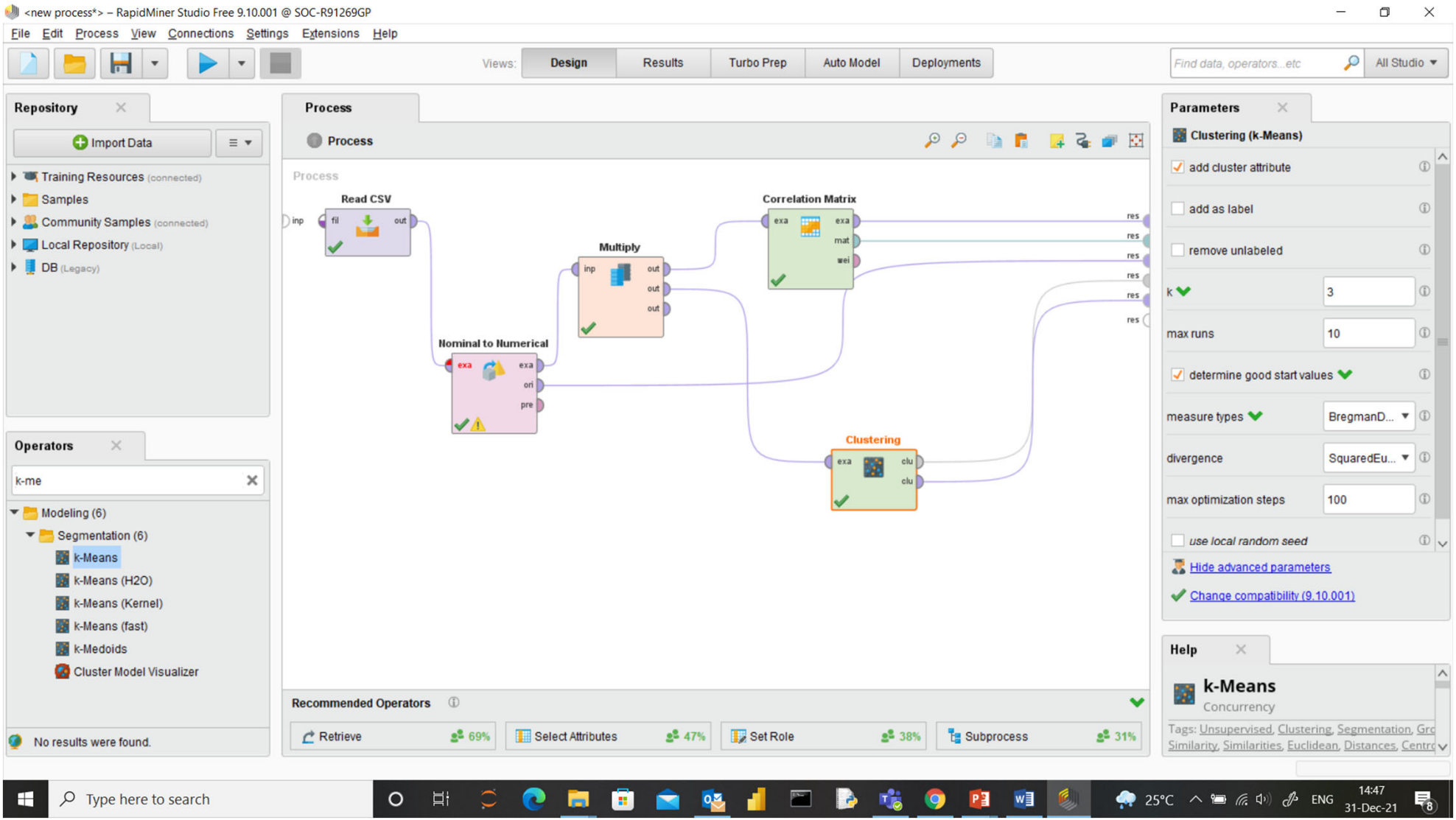


Practice 4 : Analyze the Titanic Disaster

- Try to find any meaningful correlation between variables.
- Try variables statistics to see if you can learn anything from that report too.
- Next, you need to apply k-means clustering. Design something like the next page system.
- Multiply operator gives you more than one copy of a signal, here the output of nominal-to-numeric convertor.
- Try different k=2,3,4, then say what is the major point for each selected k in this clustering task, regarding the clustering outcomes.

NUS National University of Singapore | School of Computing

Practice 4 : Analyze the Titanic Disaster



What is the Hierarchical Clustering Method

Hierarchical Clustering

- In Machine Learning, hierarchical is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:
  - Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  - Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
- In general, the merges and splits are determined in a greedy manner.
- The results of hierarchical clustering are usually presented in a dendrogram.

Hierarchical Clustering

- This is a relatively computationally heavy and time consuming algorithm.
- To decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required, using an appropriate metric and a linkage.
  - Metric: A measure of distance between pairs of observations
  - Linkage criterion: Specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Hierarchical Clustering: Metric

- The choice of metric will influence the shape of the clusters
- Some commonly used metrics for hierarchical clustering are:

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan (or city block) distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maximum distance (or Chebyshev distance)	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the Covariance matrix

- For text or other non-numeric data, metrics such as the Hamming distance or Levenshtein distance are often used.

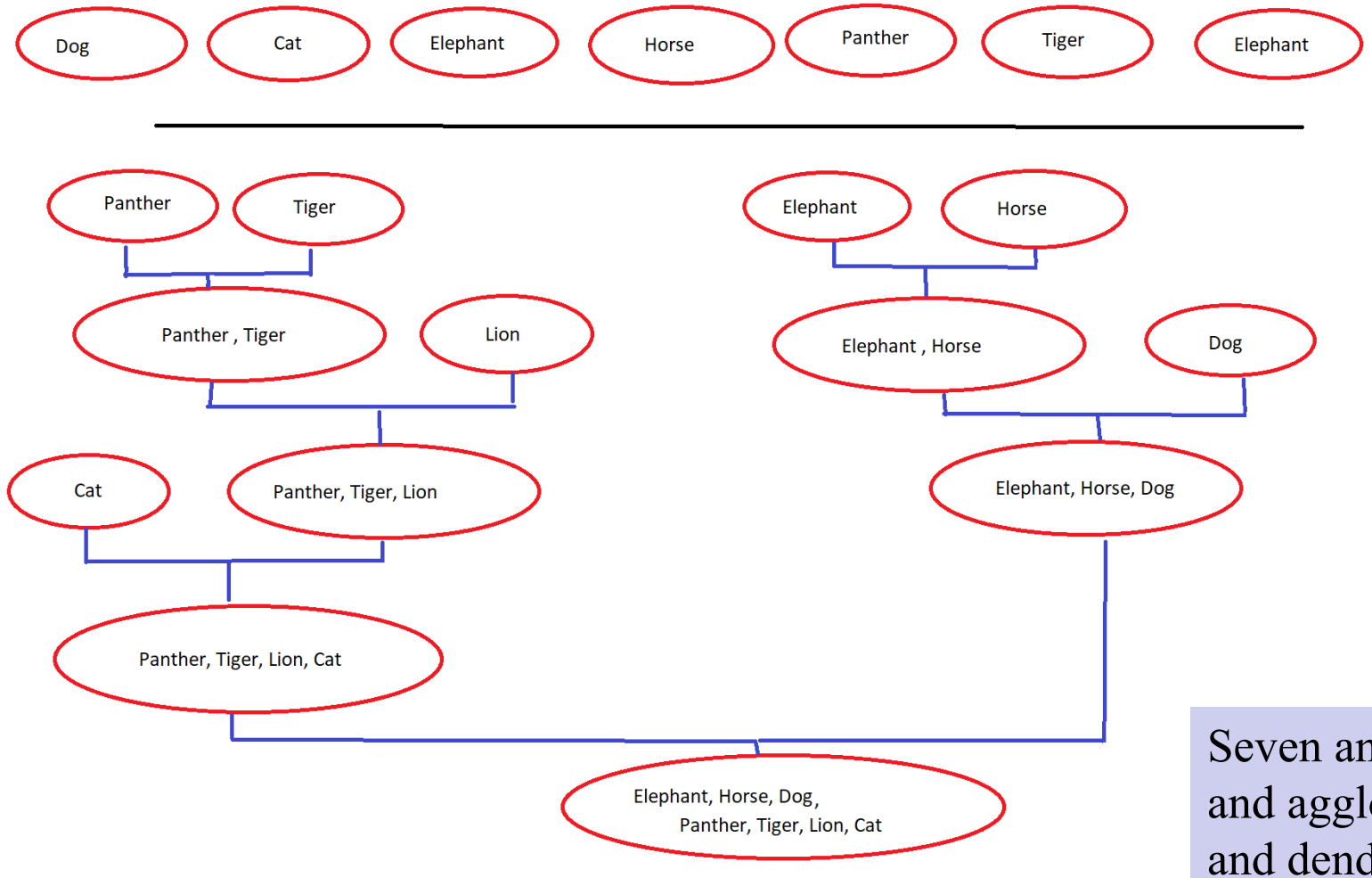
Hierarchical Clustering: LC

- The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.
- Some commonly used linkage criteria between two sets of observations A and B are:

Names	Formula
Maximum or complete-linkage clustering	$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Minimum or single-linkage clustering	$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Unweighted average linkage clustering (or UPGMA)	$\frac{1}{\lvert A\rvert \cdot \lvert B\rvert} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Weighted average linkage clustering (or WPGMA)	$d(i \cup j, k) = \frac{d(i,k) + d(j,k)}{2}.$
Centroid linkage clustering, or UPGMC	$\lVert c_s - c_t \rVert$ where c_s and c_t are the centroids of clusters s and t, respectively.
Minimum energy clustering	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \lVert a_i - b_j \rVert_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} \lVert a_i - a_j \rVert_2 - \frac{1}{m^2} \sum_{i,j=1}^{m} \lVert b_i - b_j \rVert_2$

Basic Algorithm

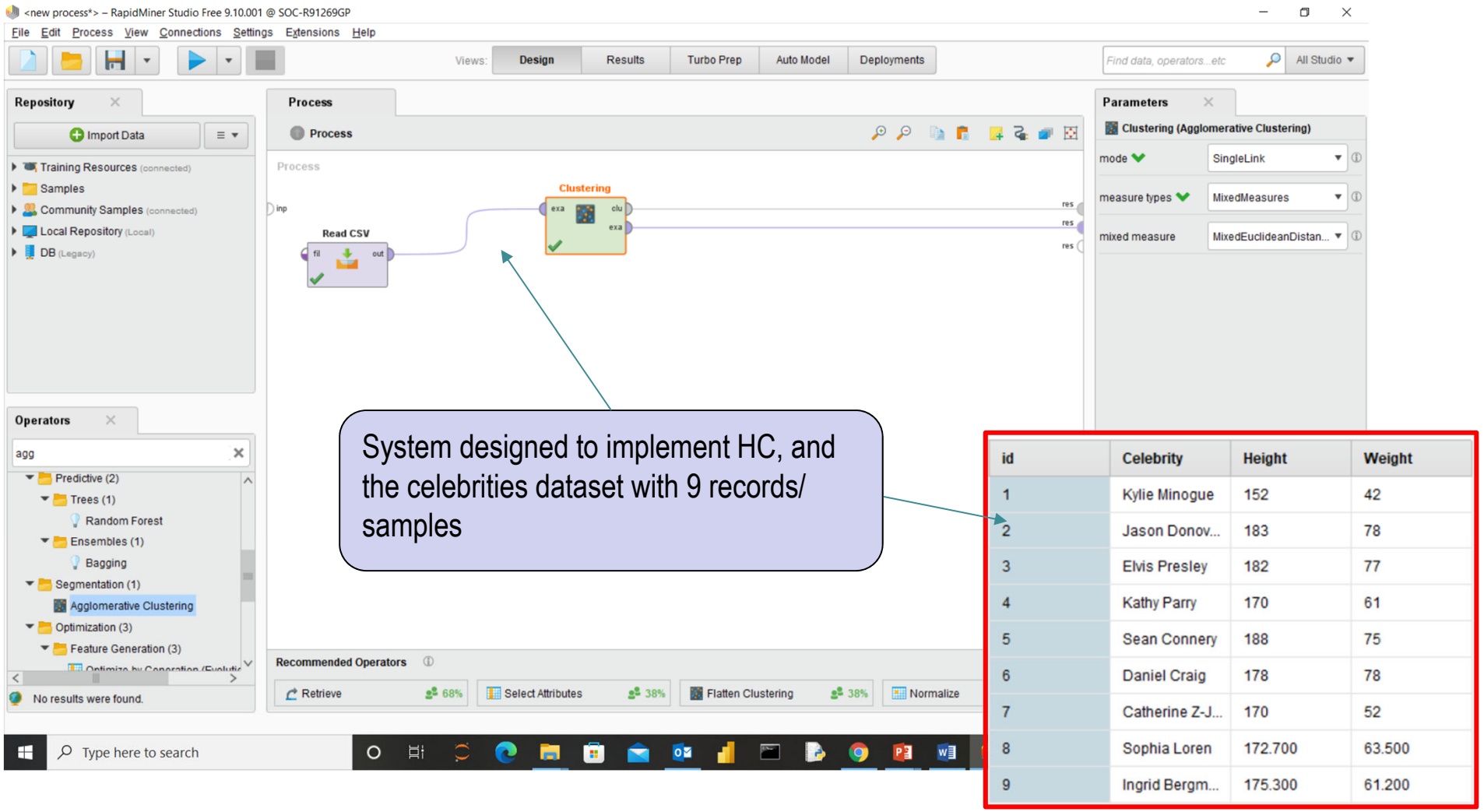
Basic Algorithm



Seven animals, HC and agglomeration, and dendrogram

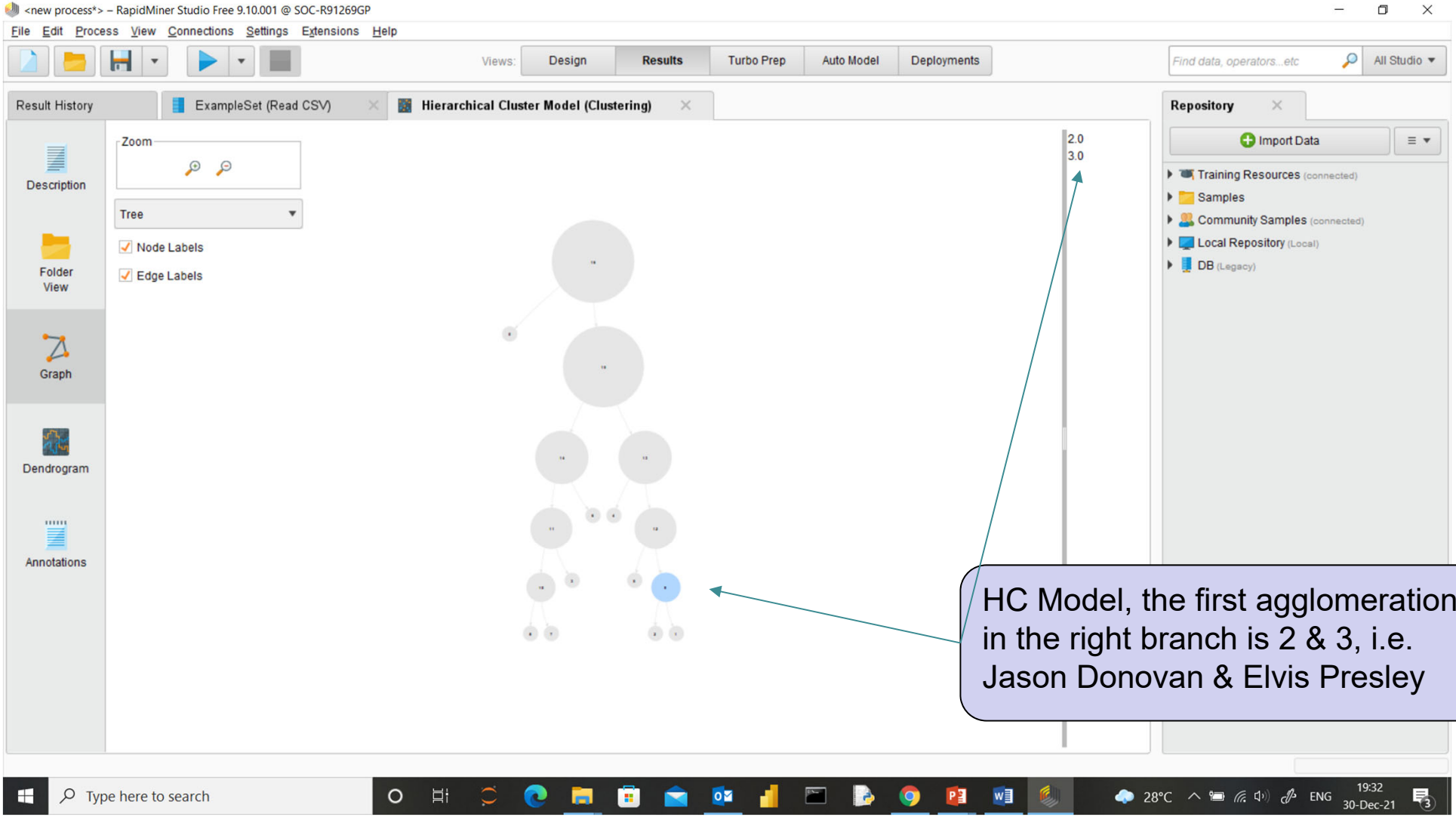
Examples

Example 1: Celebrities, Height, Weight



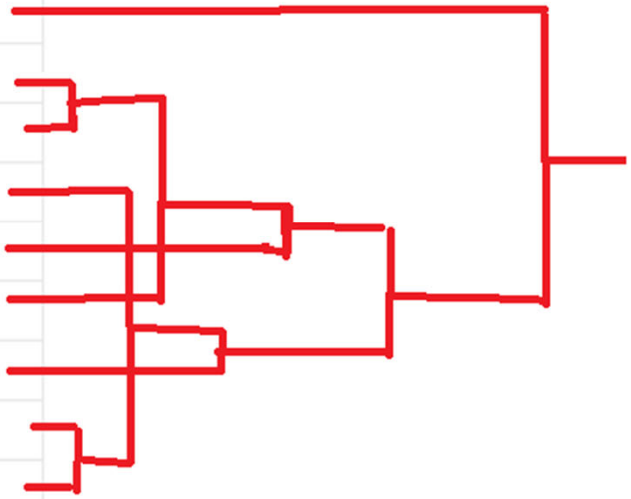
id	Celebrity	Height	Weight
1	Kylie Minogue	152	42
2	Jason Donov...	183	78
3	Elvis Presley	182	77
4	Kathy Parry	170	61
5	Sean Connery	188	75
6	Daniel Craig	178	78
7	Catherine Z-J...	170	52
8	Sophia Loren	172.700	63.500
9	Ingrid Bergm...	175.300	61.200

Example 1: Celebrities, Height, Weight



Example 1: Celebrities, Height, Weight

id	Celebrity	Height	Weight
1	Kylie Minogue	152	42
2	Jason Donov...	183	78
3	Elvis Presley	182	77
4	Kathy Parry	170	61
5	Sean Connery	188	75
6	Daniel Craig	178	78
7	Catherine Z-J...	170	52
8	Sophia Loren	172.700	63.500
9	Ingrid Bergm...	175.300	61.200



Dendrogram generated by HC
Closest together, furthest to the other group.

Mini Project

Mini Project: Customer H Segmentation

- **Dataset Mall_customers_small.csv contains data of 30 customers of a shopping mall. Again you can find it on Examples directory.**
- **Design a system to apply H clustering on that dataset.**
- **One of the variables is nominal, convert it to numeric if you think it's necessary.**
- **Find out how HC cluster the items in the first layers of clustering.**
- **Try different parameters and see the differences.**

Mini Project: Customer H Segmentation

- **Regarding the first agglomerations, which feature seems to be the least relevant/important one?**
- **Try to normalize the data, then re-answer the question above. Can you see any difference? Why?**

Function Estimation Using Linear Regression

Regression

- Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.
- The two basic types of regression are simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis.

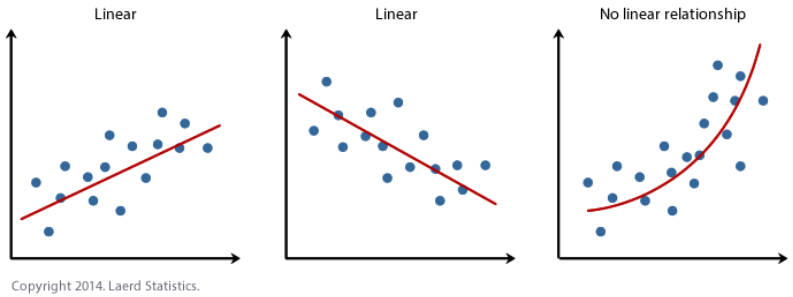
$$Y=f(x)=ax+b$$

Regression

- Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a non-linear (curved) mathematical function.
- The goal of the model is to make the sum of the squares as small as possible.
- Training in regression needs a labeled dataset (inputs/output pairs), and an optimization algorithm.
- That algorithm either uses an iterative MSE minimization scheme, or an inverse matrix scheme.

$$y = ax_1^2 + bx_2^2 + cx_1 + dx_2 + e$$

$$y = ax_1^2 + cx_1 + e$$



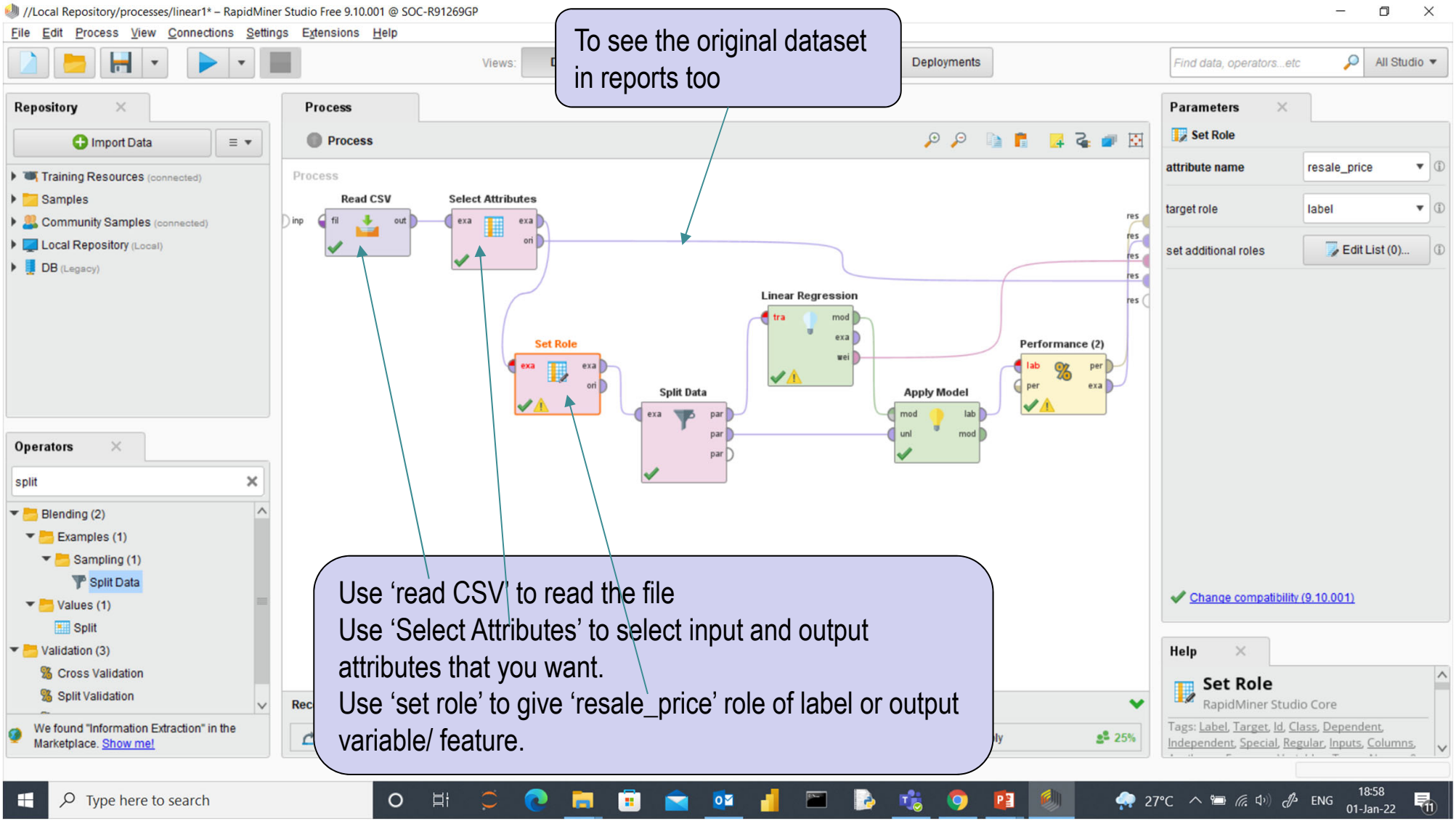
Linear Regression Example

Help your local real estate agent

Scenario:

- Data set: resale-sample.csv, contains 2000 samples of HDBs data, Singapore, there are many features, and an output called, resale-price
- We'd like to use linear egression to estimate resale-price (S$) from flat area (Sq M)
- We will expand this example to involving more features.

HDBs price estimation using LR

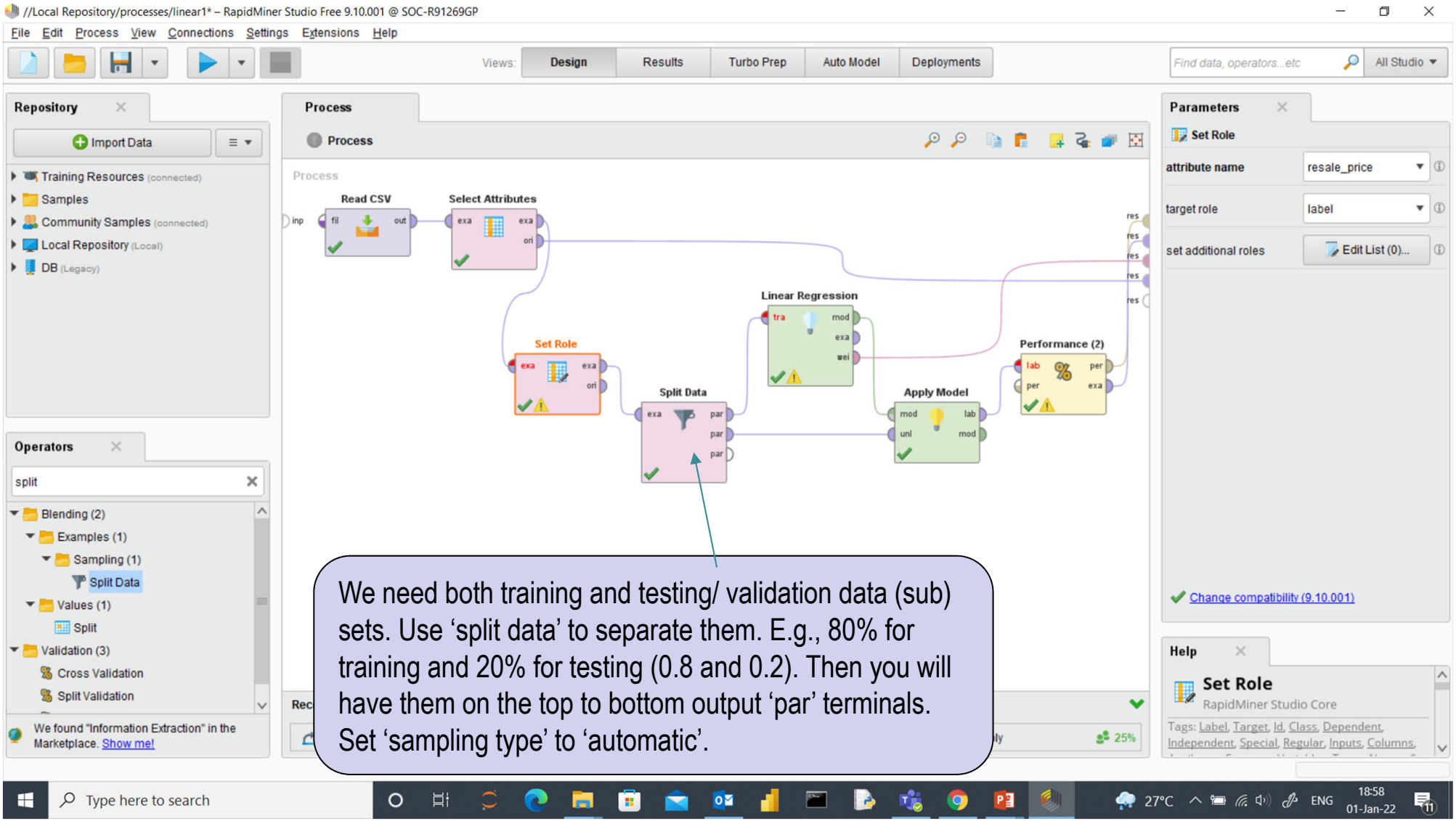


HDBs price estimation using LR

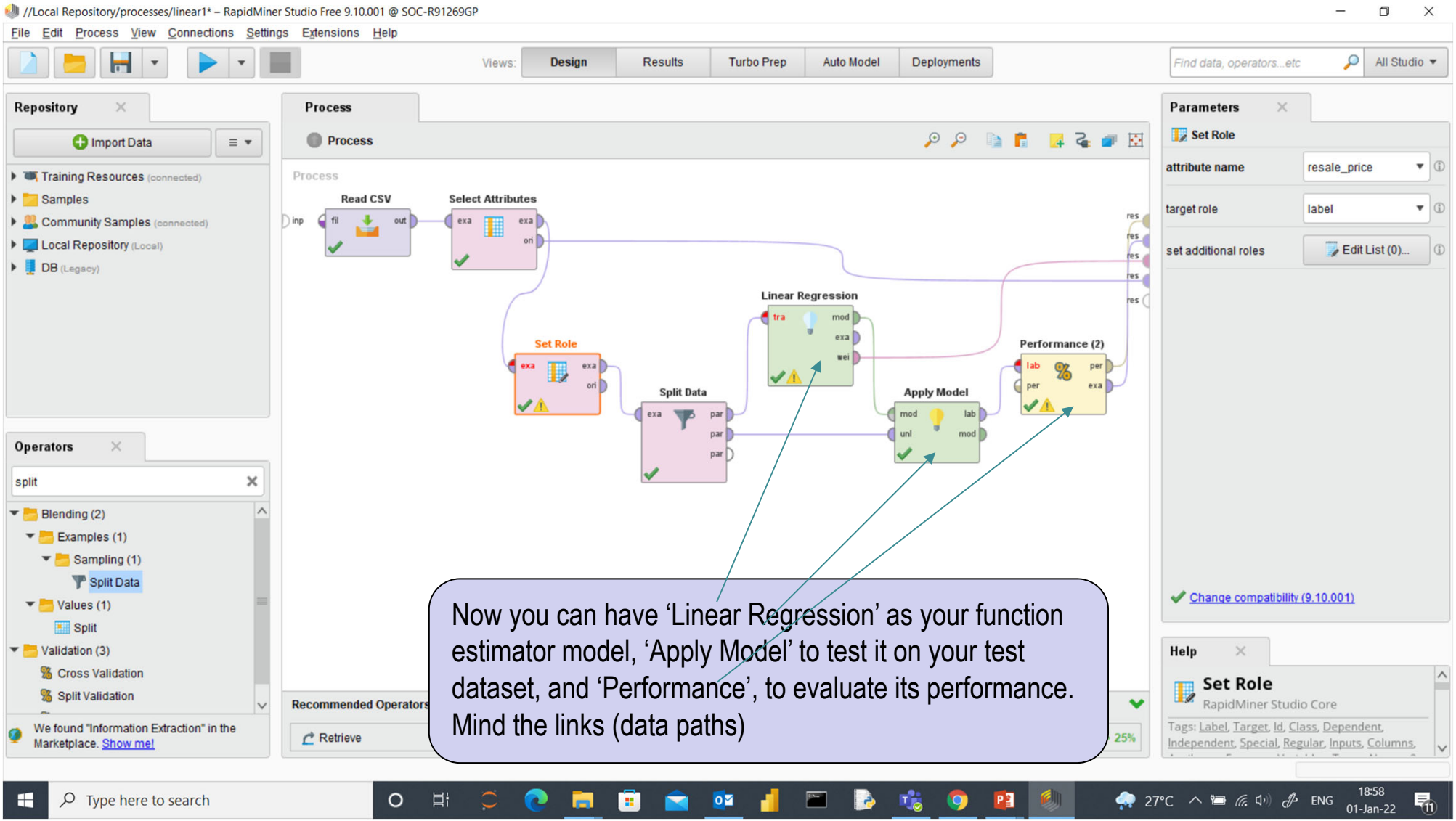
	A	B	C	D	E	F	G	H	I	J	K
1	id	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price
2	13547	2012-09	CHOA CHU	4 ROOM	119	TECK WHYE LA	04 TO 06	104	Model A	1989	400000
3	7200	2012-06	BUKIT MER	3 ROOM	22	HAVELOCK RD	04 TO 06	64	Standard	1969	404000
4	78120	2016-05	JURONG W	EXECUTIVI	906	JURONG WEST	07 TO 09	141	Apartment	1989	422000
5	33138	2013-10	JURONG W	3 ROOM	510	JURONG WEST	13 TO 15	74	Model A	1984	375000
6	57905	2015-04	JURONG EA	4 ROOM	232	JURONG EAST	07 TO 09	95	New Generatio	1982	385000
7	23096	2013-03	JURONG EA	5 ROOM	284	TOH GUAN RD	07 TO 09	120	Improved	1998	655000
8	87705	2016-11	ANG MO KI	4 ROOM	326	ANG MO KIO A	10 TO 12	92	New Generatio	1977	590000
9	95211	2017-04	CENTRAL A	3 ROOM	668	CHANDER RD	10 TO 12	75	Model A	1984	375000
10	86039	2016-10	ANG MO KI	4 ROOM	428	ANG MO KIO A	10 TO 12	92	New Generatio	1978	490000
11	20521	2013-01	GEYLANG	4 ROOM	2	HAIG RD	04 TO 06	92	New Generatio	1976	543000
12	25891	2013-05	BUKIT PAN	4 ROOM	480	SEGAR RD	07 TO 09	94	Premium Apart	2002	435000
13	26437	2013-05	MARINE PA	3 ROOM	33	MARINE CRES	07 TO 09	65	Improved	1975	465000
14	38729	2014-03	HOUGANG	4 ROOM	695	HOUGANG ST	10 TO 12	104	Model A	1987	448000
15	85543	2016-09	SENGKANG	5 ROOM	412B	FERNVALE LIN	04 TO 06	114	Premium Apart	2004	460000
16	49662	2014-11	BEDOK	4 ROOM	103	BEDOK RESER	04 TO 06	93	New Generatio	1985	405000
17	100216	2017-06	WOODLAN	EXECUTIVI	361	WOODLANDS	07 TO 09	145	Apartment	1996	638888
18	48223	2014-10	BUKIT BAT	4 ROOM	338	BT BATOK ST 3	01 TO 03	84	Simplified	1986	353000
19	27999	2013-06	SENGKANG	4 ROOM	319A	ANCHORVALE	01 TO 03	90	Model A	2002	454000
20	95100	2017-04	BUKIT MER	4 ROOM	5	DELTA AVE	04 TO 06	92	New Generatio	1985	638000

resale-sample.csv dataset
Many nominal and date-type input variables/ features
Input variables/ features for our LR model
Output variable/ label

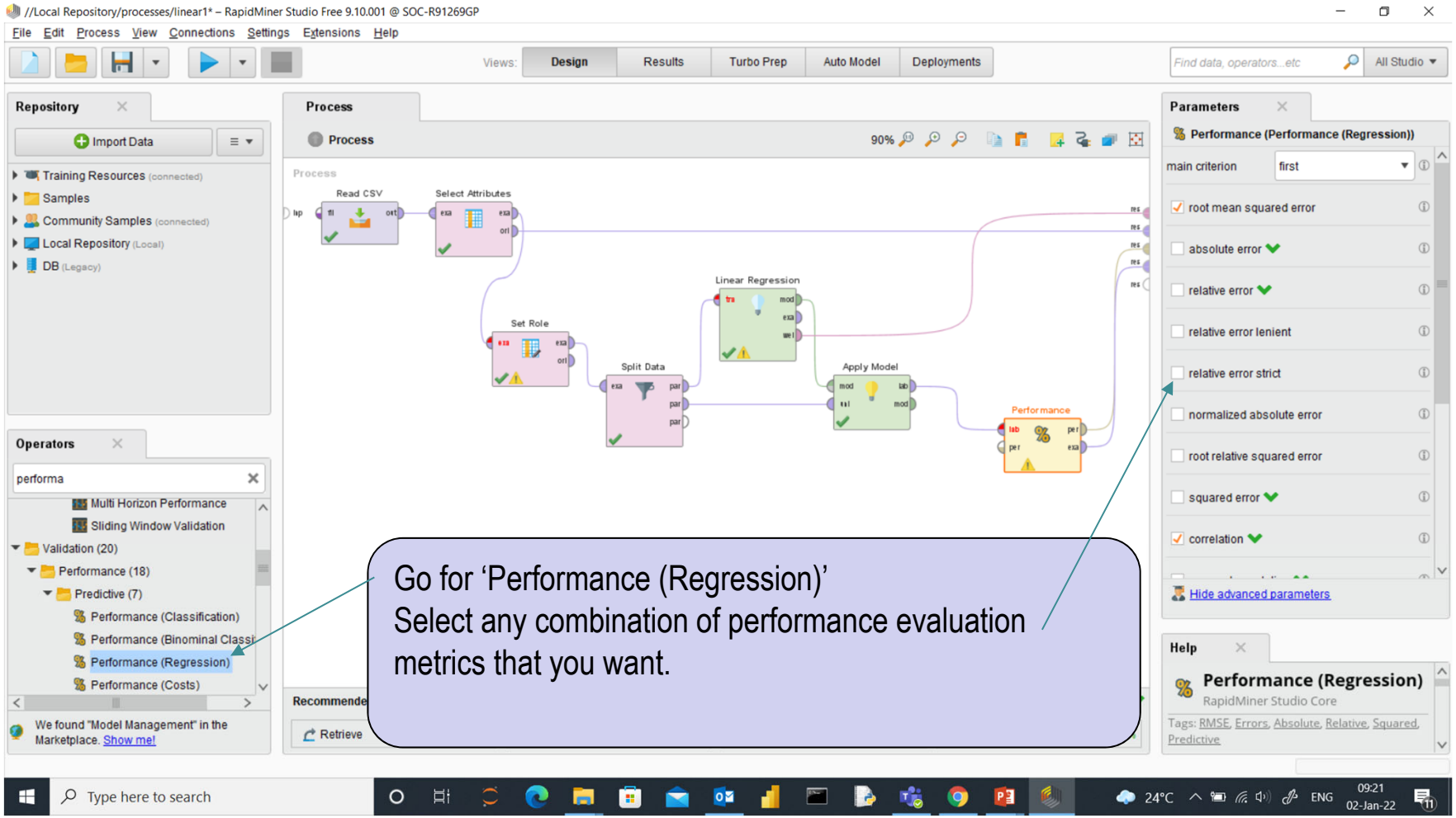
HDBs price estimation using LR



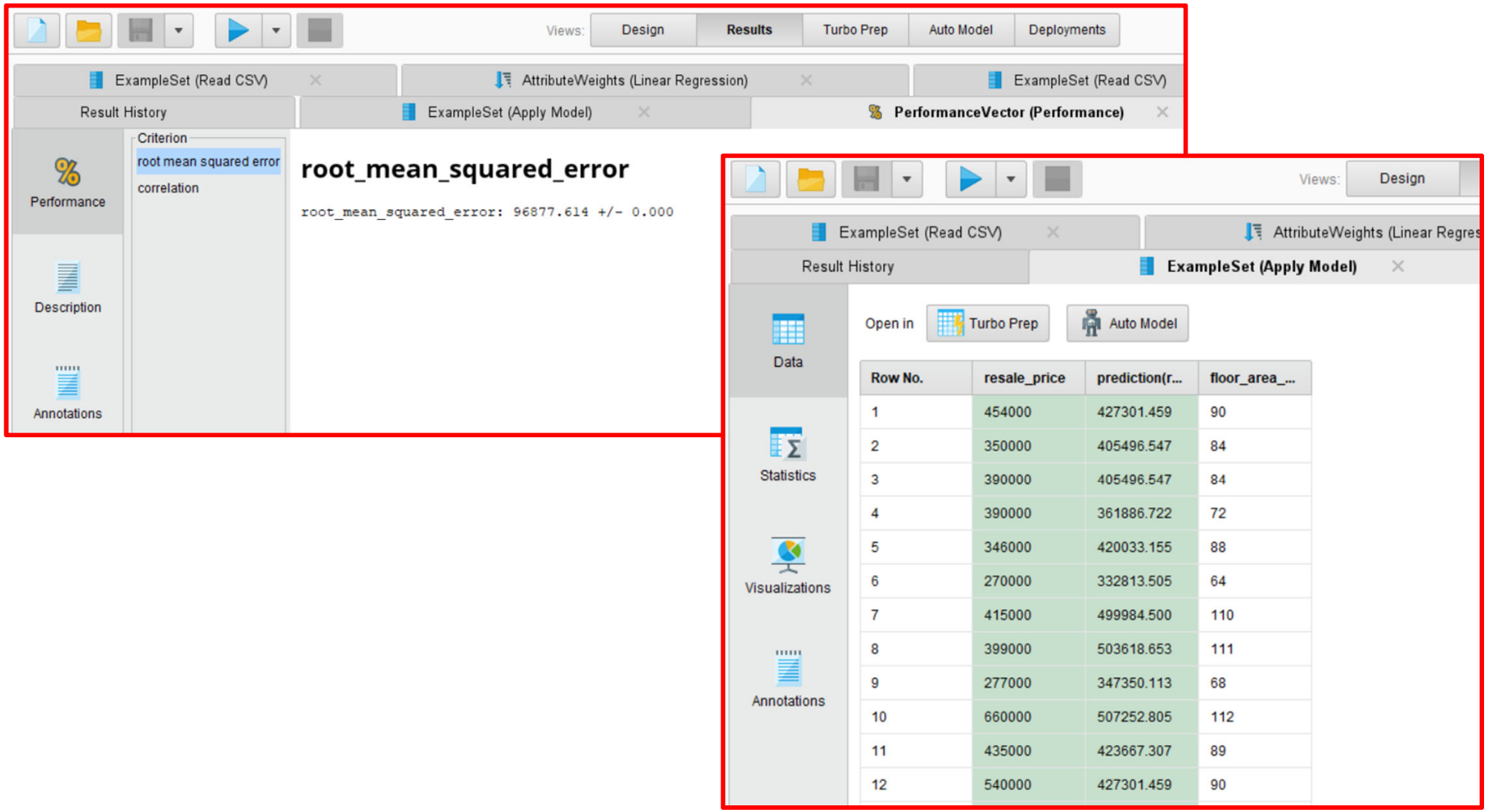
HDBs price estimation using LR



HDBs price estimation using LR



HDBs price estimation using LR, Results



Row No.	resale_price	prediction(r...	floor_area_...
1	454000	427301.459	90
2	350000	405496.547	84
3	390000	405496.547	84
4	390000	361886.722	72
5	346000	420033.155	88
6	270000	332813.505	64
7	415000	499984.500	110
8	399000	503618.653	111
9	277000	347350.113	68
10	660000	507252.805	112
11	435000	423667.307	89
12	540000	427301.459	90

Practice 5

- Design a linear regressor to estimate the HDB prices.
- Use the floor-area-sqm as your input variable and resale-price as the label (output)
- Visualize the results and analyze them
- Add lease-commence-date as another input variable
- Compare the results with the former setting, can we make it better?

Program Evaluation

THE END …


NUS
National University
of Singapore
Advanced Computing for Executives
School of Computing