



UNIVERSITY OF SANTO TOMAS



# Data Preparation using Altair® AI Studio





# UNIVERSITY OF SANTO TOMAS

Make sure that you have the two files:



CustomerDetails.xls



OrderDetails.xlsx

	A	B	C	D	E	F	G
1	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitude
2		2 No	JEAN	SMITH	female	14.55780655	121.0791117
3		3 No	JULIA	CARRERA	female	14.50814636	121.1537586
4		6 No	H	MACK	male	14.72748316	121.0204381
5		9 No	VIVIAN	GAULDEN	female	14.60740905	120.9738968
6		10 No	PAMELA	WRIGHT	female	14.63883519	121.0540112
7		11 No	MARIA	GONZALEZ	FEMALE	14.54300548	120.9336061
8		12 No	WANDA	MAYBERRY	f	14.51419589	121.1002092
9		14 No	KATHLEEN	KIRBACH	female	14.46343411	121.017539
10		15 Yes	BERNADETTE	MCKALE	female	14.59736204	120.9634643
11		16 No	JEANNINE	QUINLAN	female	14.65834207	121.0245017
12		17	CYNTHIA	LOUY	female	14.5625494	121.068564
13		18 No	JENNIFER	CAMPBELL	female	14.6264478	121.0170642

	A	B	C	D	E	F	G	H	I
1	Custom	Order_ID	Order_Dat	Store Number	Product_ID	Unit_Price	Discount	Quantity	Response
2	101	88209	4/26/14	100	16474	145.45	0.09	13	1
3	102	16710	4/26/14	103	16474	145.45	0.09	53	1
4	129	86698	11/30/15	101	16124	111.03	0.10	43	1
5	180	90784	11/4/15	100	16238	291.73	0.09	21	1
6	180	90785	12/14/15	100	16566	140.98	0.07	21	1
7	1025	89013	11/9/14	108	16182	284.98	0.08	20	1
8	1027	89016	12/29/15	103	15956	449.99	0.08	20	1
9	1030	89622	4/22/14	100	15653	175.99	0.09	11	1
10	1043	87851	9/8/15	100	16083	270.97	0.06	12	1
11	1056	90215	1/15/13	106	16244	113.98	0.06	12	1
12	1060	58628	3/25/12	105	15872	138.75	0.09	23	1
13	1074	86424	1/27/14	107	16474	145.45	0.06	11	1
14	1106	45824	12/8/12	103	16253	140.81	0.08	81	1
15	1118	86773	11/15/13	102	16361	209.84	0.09	26	1
16	1123	87016	12/22/12	100	15675	175.99	0.09	22	1
17	1129	32037	2/20/12	107	15948	699.99	0.07	15	1
18	1135	87942	12/11/13	106	16197	500.98	0.09	14	1
19	1136	87940	7/4/12	104	16083	270.97	0.09	15	1
20	1138	86577	2/4/13	105	16676	107.53	0.07	12	1
21	1139	86585	5/8/15	108	15775	152.48	0.10	12	1
22	1143	86583	8/23/14	103	16194	300.98	0.10	25	1
23	1182	86916	3/6/14	107	16886	125.99	0.10	11	1
24	1211	88601	10/25/13	109	15694	195.99	0.06	23	1
25	1211	88608	7/3/15	109	16149	165.98	0.10	17	1
26	1217	54595	5/1/12	101	16454	130.98	0.09	41	1
27	1220	23164	2/2/12	104	16342	206.10	0.10	22	1

## DATA CLEANING



the process of preparing data for analysis by removing or modifying **incorrect, incomplete, irrelevant, duplicated, or improperly formatted data.**





UNIVERSITY OF SANTO TOMAS

Open the Altair® AI Studio app installed on your computer.



**Altair® AI Studio**

2024.0

Copyright © 2001 - 2024 Altair Engineering Inc. All rights reserved.  
Reading Configuration Files



## Terminologies used in Altair AI Studio:



### attributes

variables  
(represented by columns)  
of a given data set

### examples

cases or observations  
(represented by rows)  
of a given data set

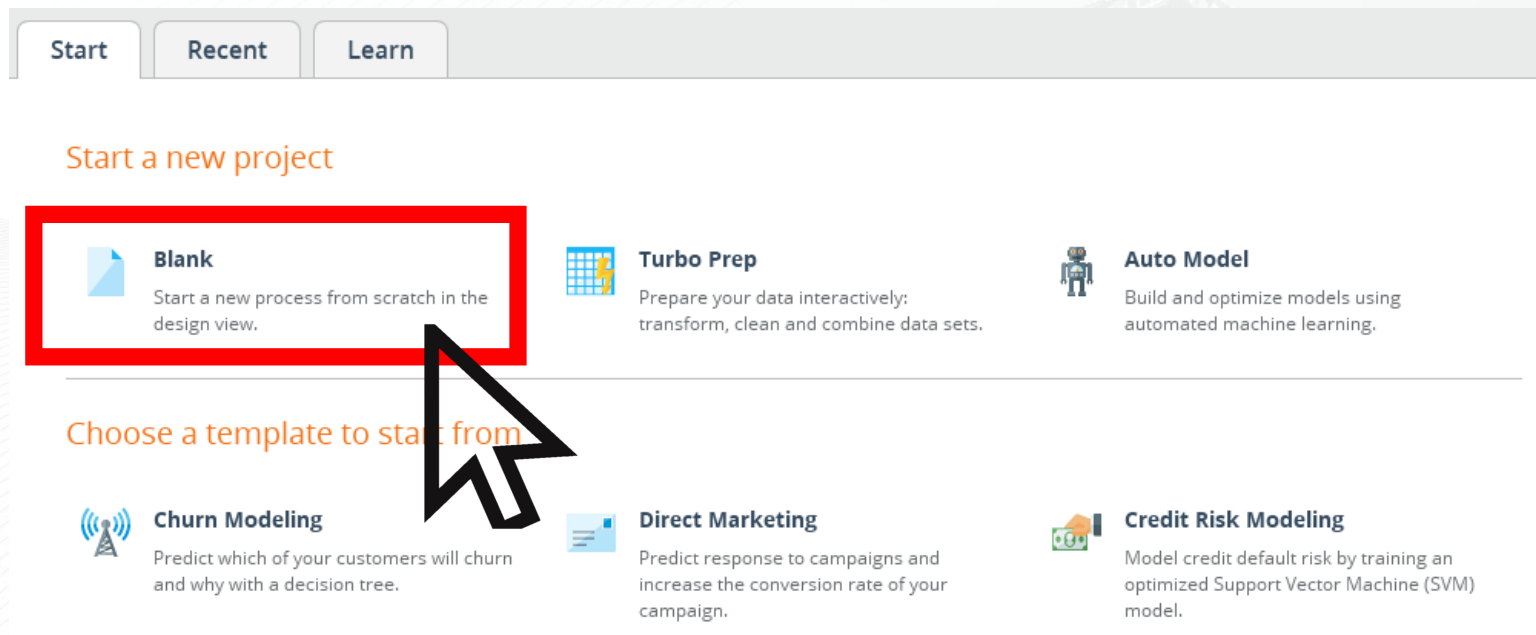
### operators

functions or  
building blocks  
that creates  
processes








Open the **Altair® AI Studio** app installed on your computer.






The screenshot shows the Altair AI Studio interface. At the top, there are three tabs: 'Start', 'Recent', and 'Learn'. Below the tabs, the 'Start a new project' section is highlighted. It contains three options: 'Blank', 'Turbo Prep', and 'Auto Model'. The 'Blank' option is highlighted with a red border and a mouse cursor. Below this, the 'Choose a template to start from' section is visible, containing three options: 'Churn Modeling', 'Direct Marketing', and 'Credit Risk Modeling'.

**Start** **Recent** **Learn**

### Start a new project

-  **Blank**  
Start a new process from scratch in the design view.
-  **Turbo Prep**  
Prepare your data interactively: transform, clean and combine data sets.
-  **Auto Model**  
Build and optimize models using automated machine learning.

### Choose a template to start from

-  **Churn Modeling**  
Predict which of your customers will churn and why with a decision tree.
-  **Direct Marketing**  
Predict response to campaigns and increase the conversion rate of your campaign.
-  **Credit Risk Modeling**  
Model credit default risk by training an optimized Support Vector Machine (SVM) model.



# UNIVERSITY OF SANTO TOMAS

The screenshot shows the RapidMiner Studio interface with several key components highlighted by orange callouts:

- Views:** A horizontal bar at the top with tabs for Design, Results, Turbo Prep, Auto Model, and Deployments.
- Repository:** A sidebar on the left containing a tree view of data sources like Training Resources, Samples, and Community Samples.
- Process:** The central workspace, currently empty, with a toolbar at the top.
- Parameters:** A panel on the right showing configuration options for the selected process, such as loggers, logfile, and random seed.
- Operators:** A sidebar on the left with a search bar and a list of operator categories like Data Access, Blending, and Cleansing.
- Help:** A panel at the bottom right providing context-sensitive help for the selected operator.

Annotations with arrows point to these components, providing brief descriptions of their functions.

**Views**

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Repository

Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- DB (Legacy)
- Local Repository
- Temporary Repository
- Training RapidMiner

**Repository**

storage within RapidMiner Studio for data and RapidMiner processes

Process work area for accessing specific functionality

Process 100%

**Parameters**

settings that modify operator behavior

Parameters

- Process
- loggers
- logfile
- logfile
- logfile
- random seed: 2001
- send mail: never
- encoding: SYSTEM
- Hide advanced parameters
- Change compatibility (9.5.001)

Operators

- Data Access (53)
- Blending (81)
- Cleansing (29)
- Modeling (160)
- Scoring (14)
- Validation (30)
- Get more operators from the Marketplace

**Operators**

building blocks used to create RapidMiner processes.

**Canvas or the Process Panel**

working area for building processes

**Help**

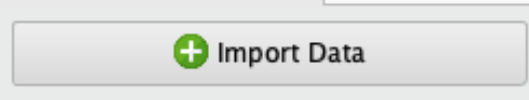
context-sensitive help for selected operator

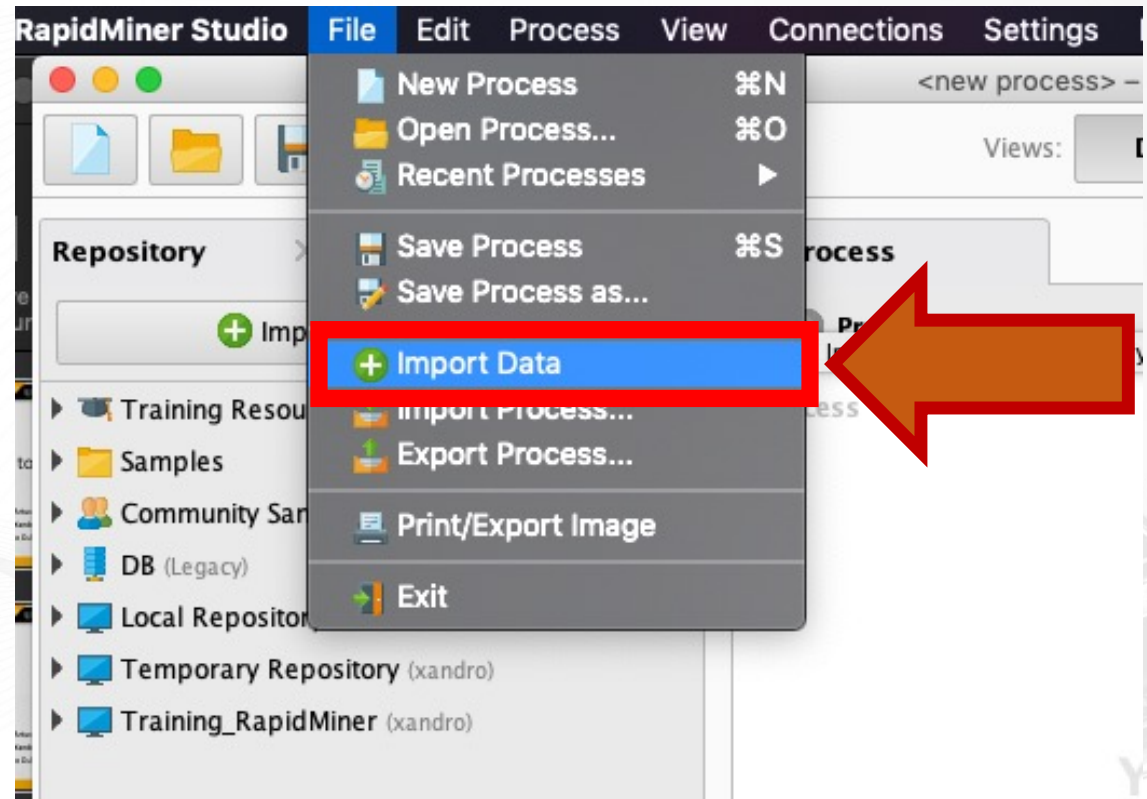
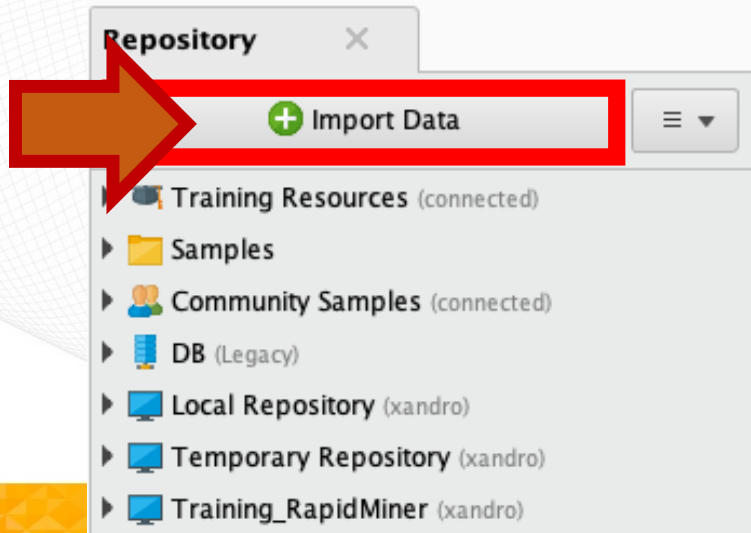
Help

- Process
- RapidMiner Studio Core
- Synopsis



## Importing Data

- Click **File** then **Import Data**,  
or click  in the Repository tab

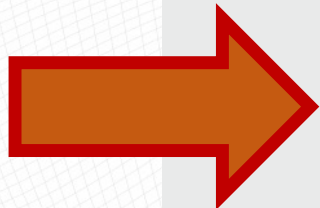






# Importing Data

- Choose the source of your data set

Where is your data?



 My Computer

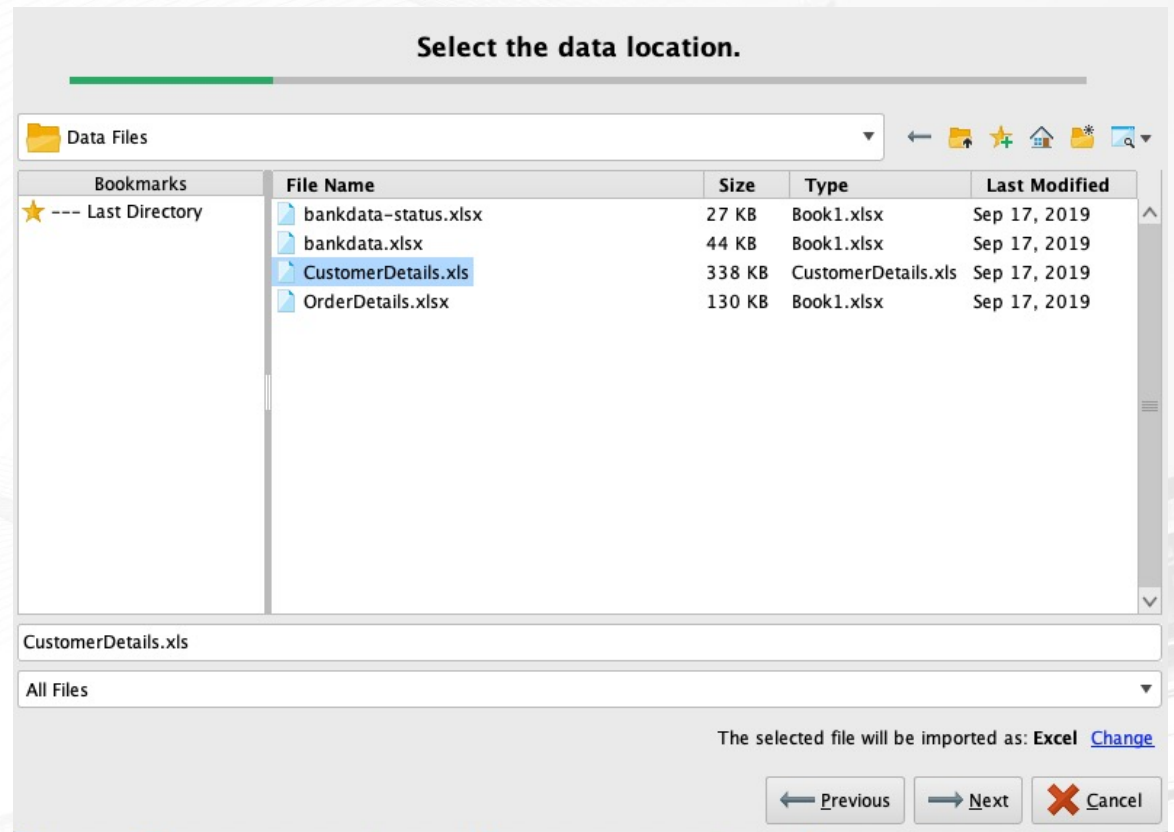
 Database



## Importing Data

- Locate the data then click **Next**.

*For this lecture,  
choose **CustomerDetails.xls**.*





## Importing Data

- Verify the cells you want to import and click **Next**.

Select the cells to import.

Sheet:  Cell range:    Define header row:

	A	B	C	D	E	F	G
1	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitude
2	2	No	JEAN	SMITH	female	14.558	121.079
3	3	No	JULIA	CARRERA	female	14.508	121.154
4	6	No	H	MACK	male	14.727	121.02
5	9	No	VIVIAN	GAULDEN	female	14.607	120.974
6	10	No	PAMELA	WRIGHT	female	14.639	121.054
7	11	No	MARIA	GONZALEZ	FEMALE	14.543	120.934
8	12	No	WANDA	MAYBERRY	f	14.514	121.1
9	14	No	KATHLEEN	KIRBACH	female	14.463	121.018
10	15	Yes	BERNADETTE	MCKALE	female	14.597	120.963
11	16	No	JEANNINE	QUINLAN	female	14.658	121.025
12	17		CYNTHIA	LOUY	female	14.563	121.069
13	19	No	JENNIFER	CAMPBELL	female	14.626	121.017
14	20	No	KRISTA	FLOREZ	Male	14.542	121.049
15	21	Yes	AVIVA	HEIFETS	female	14.612	120.993

# Importing Data

- Format the columns with your specifications.

**Format your columns.**

Replace errors with missing values ⓘ

	Customer ID * <i>integer</i>	Respond * <i>polynomial</i>	polynomial * <i>polynomial</i>	Last Name * <i>polynomial</i>	Sex * <i>polynomial</i>	Z9_Latitude * <i>real</i>	Z9 <i>re</i>
1	2	No	JEAN	SMITH	female	14.558	1
2	3	No	JULIA	CARRERA	female	14.508	1
3	6	No	H	MACK	male	14.727	1
4	9	No	VIVIAN	GAULDEN	female	14.607	1
5	10	No	PAMELA	WRIGHT	female	14.639	1
6	11	No	MARIA	GONZALEZ	FEMALE	14.543	1
7	12	No	WANDA	MAYBERRY	f	14.514	1
8	14	No	KATHLEEN	KIRBACH	female	14.463	1
9	15	Yes	BERNADETTE	MCKALE	female	14.597	1
10	16	No	JEANNINE	QUINLAN	female	14.658	1
11	17	?	CYNTHIA	LOUY	female	14.563	1
12	19	No	JENNIFER	CAMPBELL	female	14.626	1
13	20	No	KRISTA	FLOREZ	Male	14.543	1

no problems.

← Previous    → Next    ✖ Cancel

# Importing Data

- Format the columns with your specifications.
- You may change the type, role, and name of each attribute.

Responder	First Name	Last Name	Sex
<i>polynomial</i>			<i>polynom</i>
No			female
No			female
No			male
No	VIVIAN		female
No	PAMELA		female
No	MARIA		FEMALE
No	WANDA	MAYBERRY	f
No	KATHLEEN	KIRBACH	female



## Importing Data

- Types:

### polynomial

many different string values (for example: red, green, blue, yellow)

### binomial

exactly two values (for example: true/false, yes/no)

### real

a fractional number (for example: 11.23 or -0.0001).

### integer

a whole number (for example: 23, -5, or 11,024,768).

### date\_time

- both date and time (for example: 23.12.2014 17:59).

Responder	First Name	Last Name	Sex
<i>polynomial</i>			<i>polynomial</i>
No			female
No			female
No			male
No	VIVIAN		female
No	PAMELA		female
No	MARIA		FEMALE
No	WANDA	MAYBERRY	f
No	KATHLEEN	KIRBACH	female

### date

date without time (for example 23.12.2014).

### time

time without date (for example 17:59).

## Importing Data

- Format the columns with your specifications.
- You may change the type, role, and name of each attribute.
- Click **Next**.

Responder	First Name	Last Name	Sex
<i>polynomial</i>			<i>polynom</i>
No			female
No			female
No			male
No	VIVIAN		female
No	PAMELA		female
No	MARIA		FEMALE
No	WANDA	MAYBERRY	f
No	KATHLEEN	KIRBACH	female

Change Type ▶

- polynomial
- binominal
- real
- integer
- date\_time
- date
- time

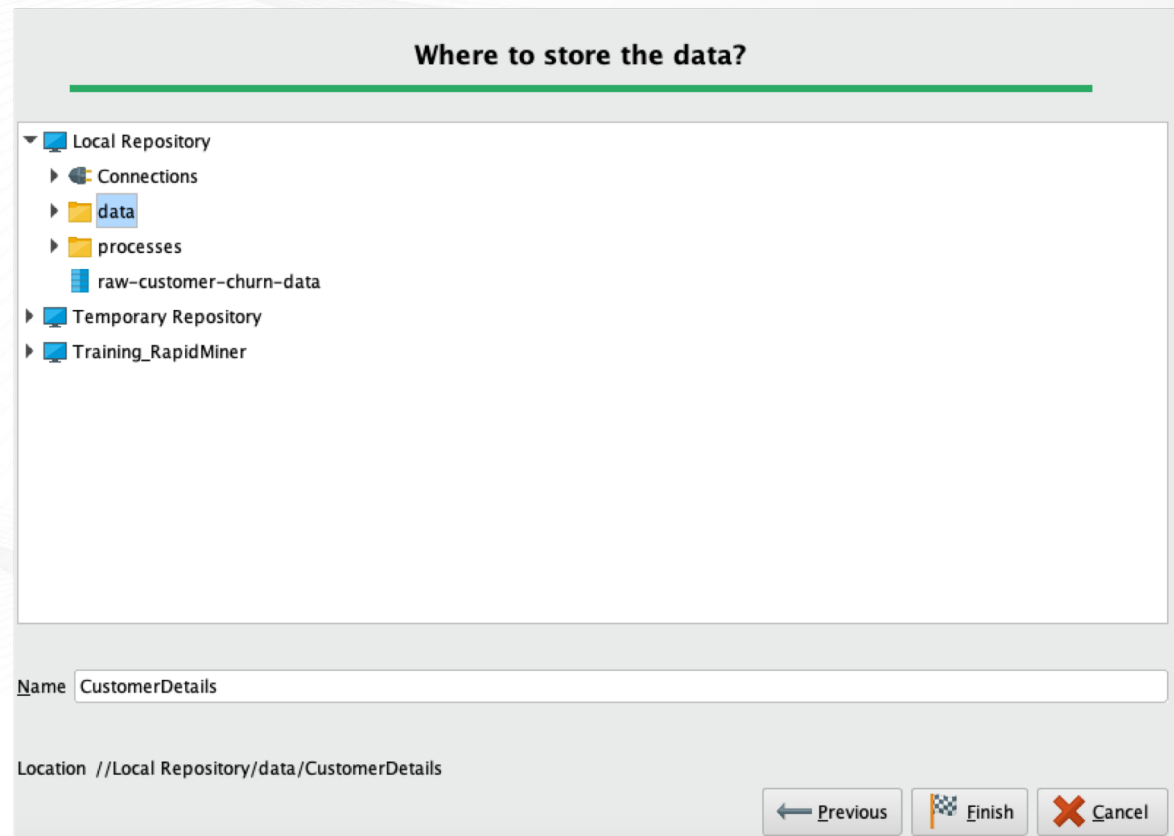
Change Role

Rename column

Exclude column

## Importing Data

- Choose the folder where the data will be stored.
- Type the file name.
- Click **Finish**.
- The data will appear in the result view.





## Importing Data

- The data will appear in the **Results** tab.

The screenshot shows a software interface with a top navigation bar containing 'Design', 'Results', 'Turbo Prep', and 'Auto Model' tabs. The 'Results' tab is highlighted with a red box. Below the navigation bar, there is a 'Result History' section and a main data view area. The data view area has a toolbar with 'Turbo Prep' and 'Auto Model' buttons, and a filter dropdown set to 'all'. The main area displays a table with 14 rows of customer data. The table columns are: Row No., Customer ID, Responder, First Name, Last Name, Sex, Z9\_Latitude, and Z9\_Longitude. The data is as follows:

Row No.	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitu...
1	2	No	JEAN	SMITH	female	14.558	121.079
2	3	No	JULIA	CARRERA	female	14.508	121.154
3	6	No	H	MACK	male	14.727	121.020
4	9	No	VIVIAN	GAULDEN	female	14.607	120.974
5	10	No	PAMELA	WRIGHT	female	14.639	121.054
6	11	No	MARIA	GONZALEZ	FEMALE	14.543	120.934
7	12	No	WANDA	MAYBERRY	f	14.514	121.100
8	14	No	KATHLEEN	KIRBACH	female	14.463	121.018
9	15	Yes	BERNADETTE	MCKALE	female	14.597	120.963
10	16	No	JEANNINE	QUINLAN	female	14.658	121.025
11	17	?	CYNTHIA	LOUY	female	14.563	121.069
12	19	No	JENNIFER	CAMPBELL	female	14.626	121.017
13	20	No	KRISTA	FLOREZ	Male	14.542	121.049
14	21	Yes	AVIVA	HEIFETS	female	14.612	120.993

Below the table, it says 'ExampleSet (2,267 examples, 0 special attributes, 7 regular attributes)'. On the right side, there is a 'Repository' panel showing a tree view of data sources.



UNIVERSITY OF SANTO TOMAS

One  
More  
Time...?

**This time, using a RapidMiner operator.**



## Importing Data

- In the Views tab, click **Design**.

The screenshot shows a software interface with a top navigation bar containing buttons for 'Design', 'Turbo Prep', and 'Auto Model'. The 'Design' button is highlighted with a red box and a red arrow. Below the navigation bar, there is a 'Result History' section with a tab for 'ExampleSet (//Local Repository/data/Customers)'. The main area displays a table of customer data with columns: Row No., Customer ID, Responder, First Name, Last Name, Sex, Z9\_Latitude, and Z9\_Longitude. The table contains 14 rows of data. On the right side, there is a 'Repository' panel showing a tree view of data sources.

Row No.	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitu...
1	2	No	JEAN	SMITH	female	14.558	121.079
2	3	No	JULIA	CARRERA	female	14.508	121.154
3	6	No	H	MACK	male	14.727	121.020
4	9	No	VIVIAN	GAULDEN	female	14.607	120.974
5	10	No	PAMELA	WRIGHT	female	14.639	121.054
6	11	No	MARIA	GONZALEZ	FEMALE	14.543	120.934
7	12	No	WANDA	MAYBERRY	f	14.514	121.100
8	14	No	KATHLEEN	KIRBACH	female	14.463	121.018
9	15	Yes	BERNADETTE	MCKALE	female	14.597	120.963
10	16	No	JEANNINE	QUINLAN	female	14.658	121.025
11	17	?	CYNTHIA	LOUY	female	14.563	121.069
12	19	No	JENNIFER	CAMPBELL	female	14.626	121.017
13	20	No	KRISTA	FLOREZ	Male	14.542	121.049
14	21	Yes	AVIVA	HEIFETS	female	14.612	120.993

ExampleSet (2,267 examples, 0 special attributes, 7 regular attributes)



## Importing Data

- Search for **Read Excel** in the operator tab.

The screenshot displays a software interface for data integration. It is divided into three main sections:

- Repository:** Shows a tree view of data sources. Under 'Local Repository (xandro)', the 'data (xandro)' folder is expanded, showing files like 'credit (xandro - v1, 8/23/19 1:31 PM - 47 kB)' and 'credit cleaned and converted to nume'.
- Operators:** A search bar contains the text 'read excel'. Below it, a list of operators is shown under 'Data Access (4)' > 'Files (4)' > 'Read (3)'. The 'Read Excel' operator is highlighted. A red arrow points from the search bar to this operator. At the bottom of this panel, a message states: 'We found "Spreadsheet Table Extraction" in the Marketplace. [Show me!](#)'
- Process:** Shows a 'Process' tab with a 'Process' button and an 'inp' input field.

At the bottom right, a 'Recommended Operators' section lists 'Retrieve' (12%) and 'Select Attributes' (6%).

Your process looks empty  
Add some data first  
Drag data or operators



## Importing Data

- Search for **Read Excel** in the operator tab.
- Drag and drop it to the canvas.

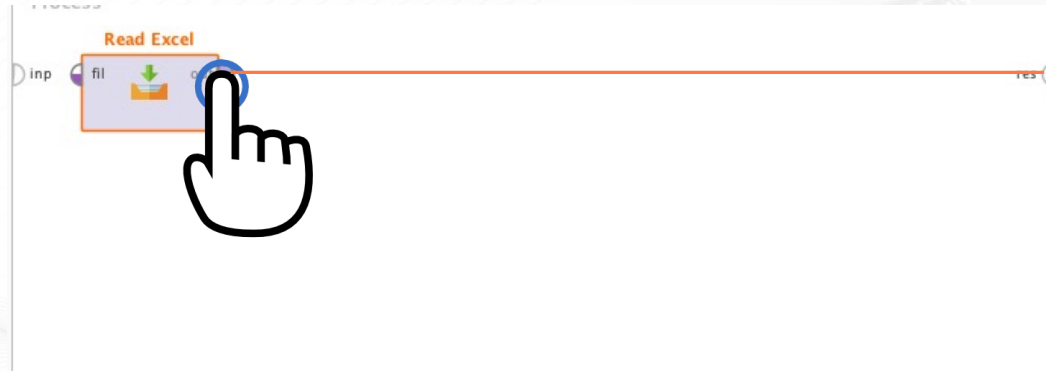
The screenshot displays a software interface for data integration, divided into several panels:

- Repository:** A tree view showing data sources. The 'Local Repository (xandro)' is expanded to show a folder named 'data (xandro)', which contains several files, including 'credit (xandro - v1, 8/23/19 1:31 PM - 47 kB)'.
- Operators:** A search bar contains the text 'read excel'. Below it, a list of operators is shown. The 'Read Excel' operator is highlighted with a red rectangular box. A green arrow points from this box to the 'Read Excel' operator on the canvas.
- Process:** A canvas area where a 'Read Excel' operator is being added. The operator is represented by a box with 'inp' and 'out' ports and a yellow warning icon. A green arrow points from the 'Read Excel' operator in the Operators panel to this operator on the canvas.
- Recommended Operators:** A section at the bottom right showing suggested operators: 'Select Attributes' (31%) and 'Set Role' (27%).

## Data Preparation

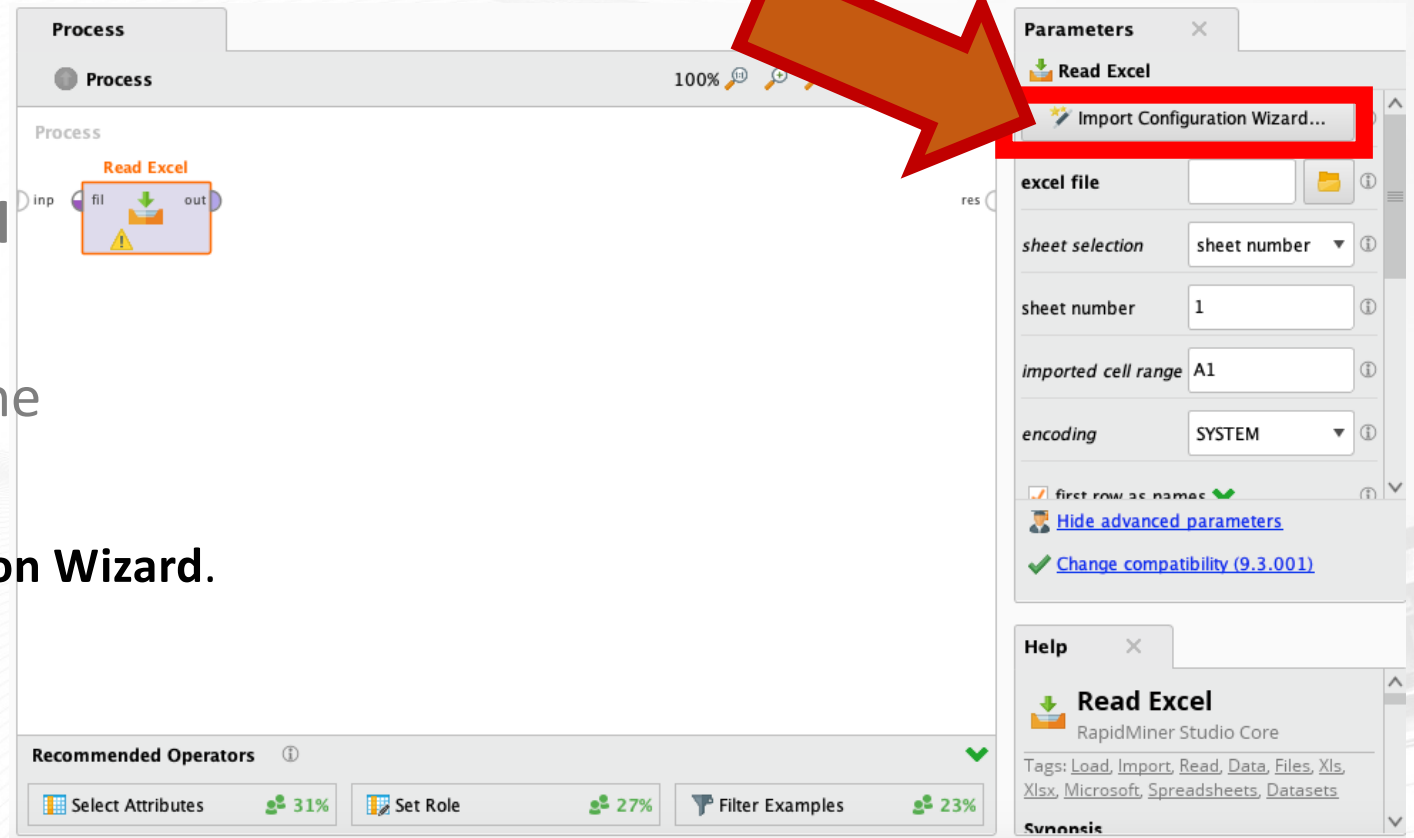
Connect the **Out** node of the **Read Excel** operator and **res** of the result knob.

This will put the **output** to the **results** tab.



## Importing Data

- Search for **Read Excel** in the operator tab.
- Drag and drop it to the canvas.
- Click **Import Configuration Wizard**.



The screenshot displays the RapidMiner Studio interface. The main canvas shows a 'Process' view with a 'Read Excel' operator. A red arrow points to the 'Import Configuration Wizard...' button in the 'Parameters' panel. The 'Parameters' panel is open, showing configuration options for the 'Read Excel' operator. The 'Import Configuration Wizard...' button is highlighted with a red box. The 'Recommended Operators' panel at the bottom shows 'Select Attributes' (31%), 'Set Role' (27%), and 'Filter Examples' (23%).

**Parameters**

- Read Excel
- Import Configuration Wizard...

excel file

sheet selection sheet number

sheet number 1

imported cell range A1

encoding SYSTEM

first row as names

[Hide advanced parameters](#)

[Change compatibility \(9.3.001\)](#)

**Help**

**Read Excel**  
RapidMiner Studio Core

Tags: [Load](#), [Import](#), [Read](#), [Data](#), [Files](#), [Xls](#), [Xlsx](#), [Microsoft](#), [Spreadsheets](#), [Datasets](#)

Synopsis

**Recommended Operators**

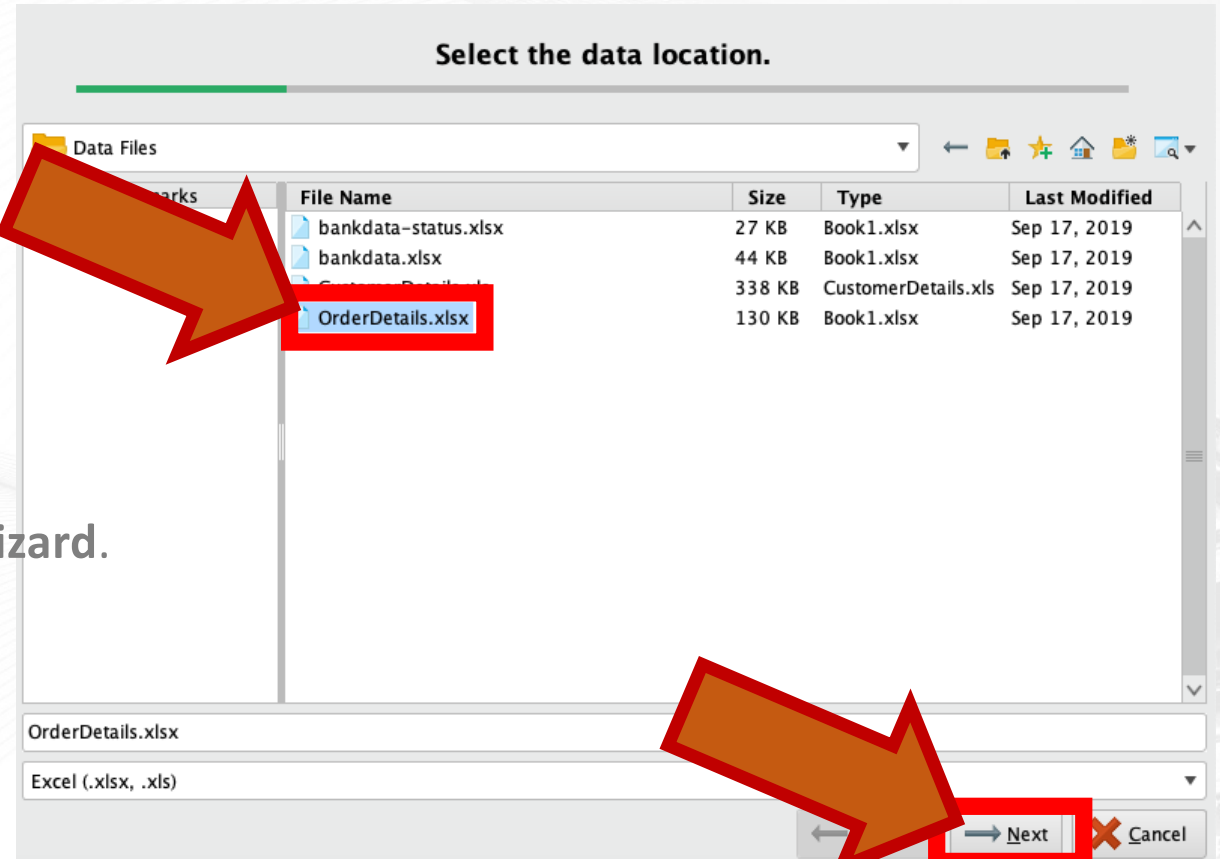
- Select Attributes 31%
- Set Role 27%
- Filter Examples 23%

## Importing Data

- Search for Read Excel in the operator tab.
- Drag and drop it to the canvas.
- Click **Import Configuration Wizard**.
- Locate and open the file.

*For this lecture, choose **OrderDetails.xls**.*

- Click **Next, Next, and Finish**.





## Data Preparation

Click the **Play** button to execute the process.

The screenshot shows a software interface for data preparation. At the top, there is a toolbar with several icons: a document, a folder, a save icon, a play button (highlighted with a red box and a red arrow), and a stop icon. To the right of the toolbar, there are tabs for 'Views: Design' and 'Result'. Below the toolbar, there are two main panels. The left panel is titled 'Repository' and contains a tree view of data sources. It includes an 'Import Data' button at the top. The tree view shows a hierarchy: 'Community Samples (connected)', 'DB (Legacy)', 'Local Repository (xandro)', 'Connections (xandro)', and 'data (xandro)'. Under 'data (xandro)', there are three items: 'credit (xandro - v1, 8/23/19 1:31 PM - 47 kB)', 'credit cleaned and converted to nume', and 'credit cleaned and converted to nume'. The right panel is titled 'Process' and shows a 'Process' tab. Below it, there is a 'Process' section with a 'Read Excel' operator. The operator is represented by a box with a green arrow pointing down into a folder icon, and it has 'inp' and 'out' ports. A blue line connects the 'out' port to the right.



## Importing Data



Views:

Design

**Results**

Turbo Prep

Auto Model

Result History

ExampleSet (Read Excel) ×

ExampleSet (//Local Repository/data/CustomerDetails) ×



Data

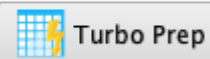


Statistics



Visualizations

Open in



Turbo Prep



Auto Model

Filter (2,268 / 2,268 examples):

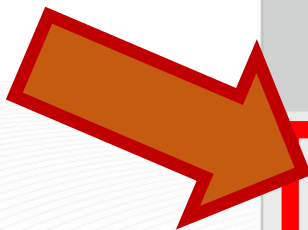
all

Row No.	Customer_I...	Order_ID	Order_Date	Store Num...	Product_ID	Unit_Price	Discount	Quantity
1	101	88209	Apr 26, 2014	100	16474	145.450	0.090	13
2	102	16710	Apr 26, 2014	103	16474	145.450	0.090	53
3	129	86698	Nov 30, 2015	101	16124	111.030	0.100	43
4	180	90784	Nov 4, 2015	100	16238	291.730	0.090	21
5	180	90785	Dec 14, 2015	100	16566	140.980	0.070	21
6	1025	89013	Nov 9, 2014	108	16182	284.980	0.080	20
7	1027	89016	Dec 29, 2015	103	15956	449.990	0.080	20
8	1030	89622	Apr 22, 2014	100	15653	175.990	0.090	11



# Exploratory Analysis

- To find the basic statistics of each attributes, click **Statistics**.



Views: Design Results Turb

Result History ExampleSet (Read Excel) ExampleSet (//Local Repository/data/Cus

Open in Turbo Prep Auto Model Filter (2

Row No.	Customer_I...	Order_ID	Order_Date	Store Num...	Product_ID
1	101	88209	Apr 26, 2014	100	16474
2	102	16710	Apr 26, 2014	103	16474
3	129	86698	Nov 30, 2015	101	16124
4	180	90784	Nov 4, 2015	100	16238
5	180	90785	Dec 14, 2015	100	16566
6	1025	89013	Nov 9, 2014	108	16182
7	1027	89016	Dec 29, 2015	103	15956
8	1030	89622	Apr 22, 2014	100	15653
9	1043	87851	Sep 8, 2015	100	16083
10	1055	88215	Nov 15, 2013	100	16244

Data

**Statistics**

Visualizations

Annotations



# Exploratory Analysis

- To find the basic statistics of each attribute, click **Statistics**.

The screenshot shows a data analysis tool interface with a sidebar on the left containing icons for Data, Statistics, Visualizations, and Annotations. The main area displays a table of attributes with their respective statistics. The 'Order\_ID' row is highlighted, and a bar chart visualization is shown for it. The table includes columns for Name, Type, Missing, Statistics, and Filter (9 / 9 attributes).

Name	Type	Missing	Statistics	Filter (9 / 9 attributes):
Customer_ID	Integer	0	Min 10, Max 1866, Average 1332.0	Search for Attribute.
Order_ID	Integer	0	Min 193, Max 91577	
Order_Date	Date	0	Earliest date Jan 2, 2012, Latest date Jan 2, 2016, Duration 1461 c	
Store Number	Integer	0	Min 100, Max 109, Average 104.23	
Product_ID	Integer	0	Min 15635, Max 16893, Average 16273	
Unit_Price	Real	0	Min 1.140, Max 3502.140, Average 83.789	



# Data Preparation

Go back to **Design** view.



Software interface showing a toolbar with icons for file operations and execution. Below the toolbar is a 'Result History' section with two tabs: 'ExampleSet (Read Excel)' and 'ExampleSet (//Local Repository/c'. The main area displays a table with columns: Name, Type, Missing, and Statistics.

	Name	Type	Missing	Statistics
Data	Customer_ID	Integer	0	Min 10
Statistics	Order_ID	Integer	0	Min 193
Visualizations	Order_Date	Date	0	Earliest date Jan 2, 2012
Annotations	Store Number	Integer	0	Min 100
	Product_ID	Integer	0	Min 15635
	Unit_Price	Real	0	Min 1.140



UNIVERSITY OF SANTO TOMAS

# Data Filtering

using

## Altair® AI Studio

the process performed to decide which examples are **kept** and which are removed





# Data Preparation

## 1. Filtering cases.

- In the operator tab, search for **Filter Examples**, then drag and drop on the line connecting the *Read Excel* and the *res* knob.

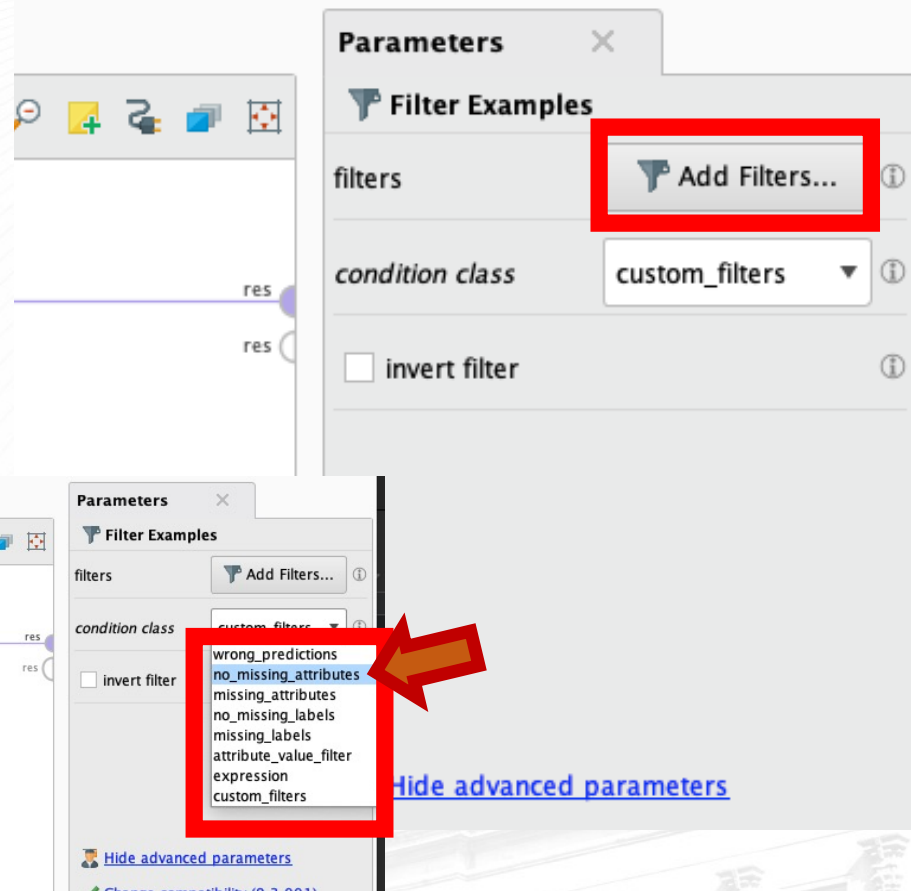
The screenshot displays the Alteryx software interface. On the left, the **Repository** pane shows a tree view with 'Local Repository (xandro)' expanded to 'data (xandro)', which contains files like 'credit' and 'credit cleaned and converted to nume'. Below it, the **Operators** pane shows a search for 'filter examples' with a red box highlighting the 'Filter Examples' operator under 'Filter (1)'. On the right, the **Process** pane shows a workflow with a 'Read Excel' operator connected to a 'Filter Examples' operator. A green arrow points from the highlighted 'Filter Examples' operator in the Operators pane to the 'Filter Examples' operator in the Process pane.

## Data Preparation

### 1. Filtering cases.

- In the parameter tab, choose **Add Filter** in the condition class.

Note: Instead of filtering, you may **remove all cases with missing values** using the condition class, instead of Add Filters.





# Data Preparation

## 1. Filtering cases.

- Choose the attribute's filtering criteria.



Create Filters: **filters**  
Defines the list of filters to apply.

Customer\_ID|

=








## Data Preparation

### 1. Filtering cases.

- Choose the attribute's filtering criteria.
- Example, retaining only the orders before 2016.

 Create Filters: **filters**  
Defines the list of filters to apply.

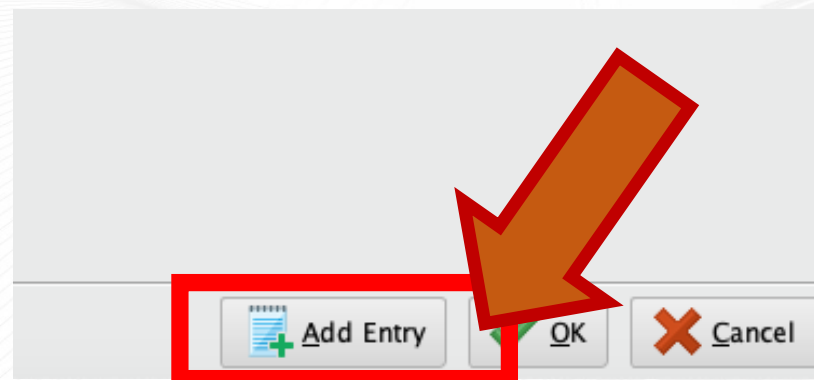
Order_Date	▼	<	▼	01/01/2016		
------------	---	---	---	------------	--	--

This will remove case(s) ordered from 2016 and beyond.

## Data Preparation

### 1. Filtering cases.

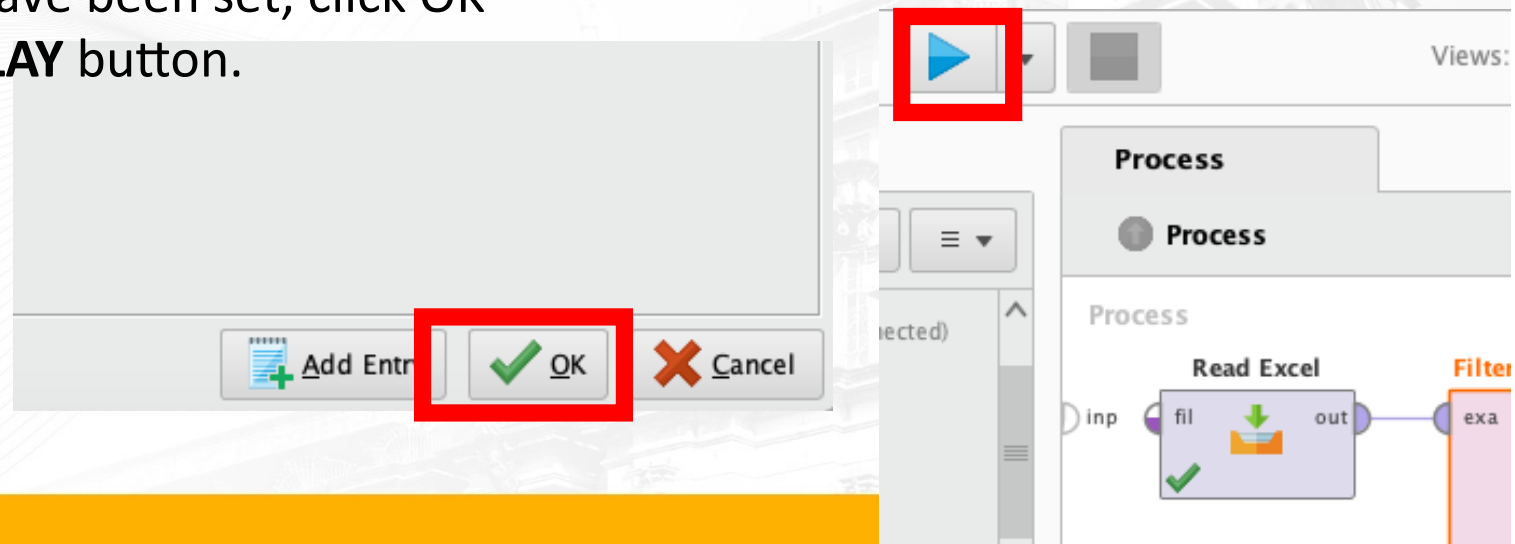
- Choose the attribute's filtering criteria.
- Example, retaining only the orders before 2016.
- You may add more criteria by clicking **Add Entry**.



## Data Preparation

### 1. Filtering cases.

- Choose the attribute's filtering criteria.
- Example, retaining only the orders before 2016.
- You may add more criteria by clicking **Add Entry**.
- Once all criteria have been set, click OK then click the **PLAY** button.

A screenshot of a data preparation software interface. The interface shows a workflow with a 'Read Excel' process connected to a 'Filter' process. The 'Filter' process is highlighted in orange. A modal dialog box is open in the foreground, containing three buttons: 'Add Entry' (with a plus icon), 'OK' (with a green checkmark icon), and 'Cancel' (with a red X icon). The 'OK' button is highlighted with a red square. In the background, a blue play button icon is also highlighted with a red square. The interface includes a 'Views' dropdown menu and a 'Process' section with a 'Process' button.



# Data Preparation

RapidMiner removed 1 case, an order taken from 2016 onwards.

Design Results Turbo Prep Auto Model Find

Set (//Local Repository/data/CustomerDetails) × SampleSet (//Local Repository/data/CustomerDetails) ×

Filter (2,268 / 2,268 examples): a Filter (2,267 / 2,267 examples): All

Store Num...	Product_ID	Unit_Price	Discount	Quantity
100	16474	145.450	0.090	13
103	16474	145.450	0.090	53
101	16124	111.030	0.100	43
100	16238	291.730	0.090	21
100	16566	140.980	0.070	21

Store Num...	Product_ID	Unit_Price	Discount	Quantity
100	16474	145.450	0.090	13
103	16474	145.450	0.090	53
101	16124	111.030	0.100	43
100	16238	291.730	0.090	21
100	16566	140.980	0.070	21
108	16182	284.980	0.080	20



UNIVERSITY OF SANTO TOMAS

# Missing Value Imputation

using

**Altair® AI Studio**

replaces missing values by the attribute's  
minimum, maximum or average value;  
zero may also be a replacement





# Data Preparation

As seen in the statistics of the data, **199 examples** have missing values in the **Discount** attribute.

Name	Type	Missing	Status
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earliest Jan
Store Number	Integer	0	Min 100
Product_ID	Integer	0	Min 156
Unit_Price	Real	0	Min 1.14
Discount	Real	199	Min 0.01



# Data Preparation

Go back to **Design** view.

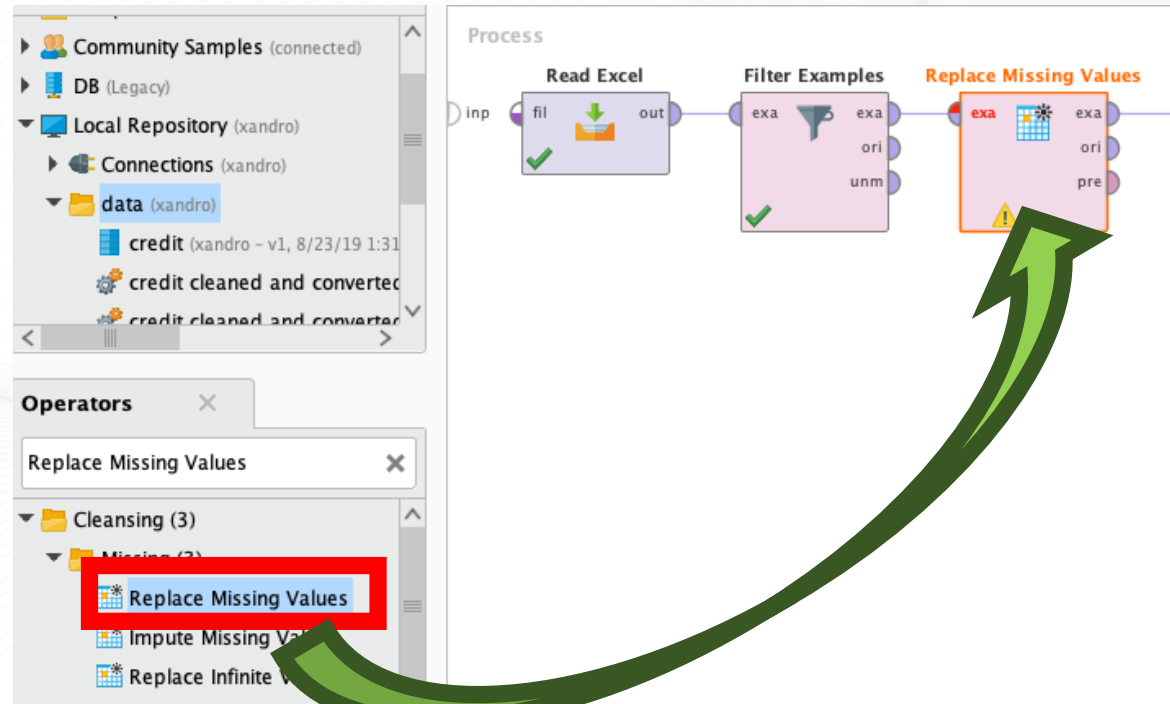
The screenshot shows a software interface for data preparation. At the top, there are icons for file operations and a 'Views' dropdown menu with 'Design' and 'Results' options. The 'Design' option is highlighted with a red box. Below the toolbar, there are tabs for 'Result History', 'ExampleSet (Filter Examples)', and 'ExampleSet (//Local Reposi...'. On the left side, there are three main sections: 'Data' (with a grid icon), 'Statistics' (with a sigma icon), and 'Visualizations' (with a pie chart icon). The main area displays a table with the following data:

Name	Type	Missing	Statistic
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earliest Jan 2,
Store Number	Integer	0	Min 100

## Data Preparation

### 2. Imputing Missing Data

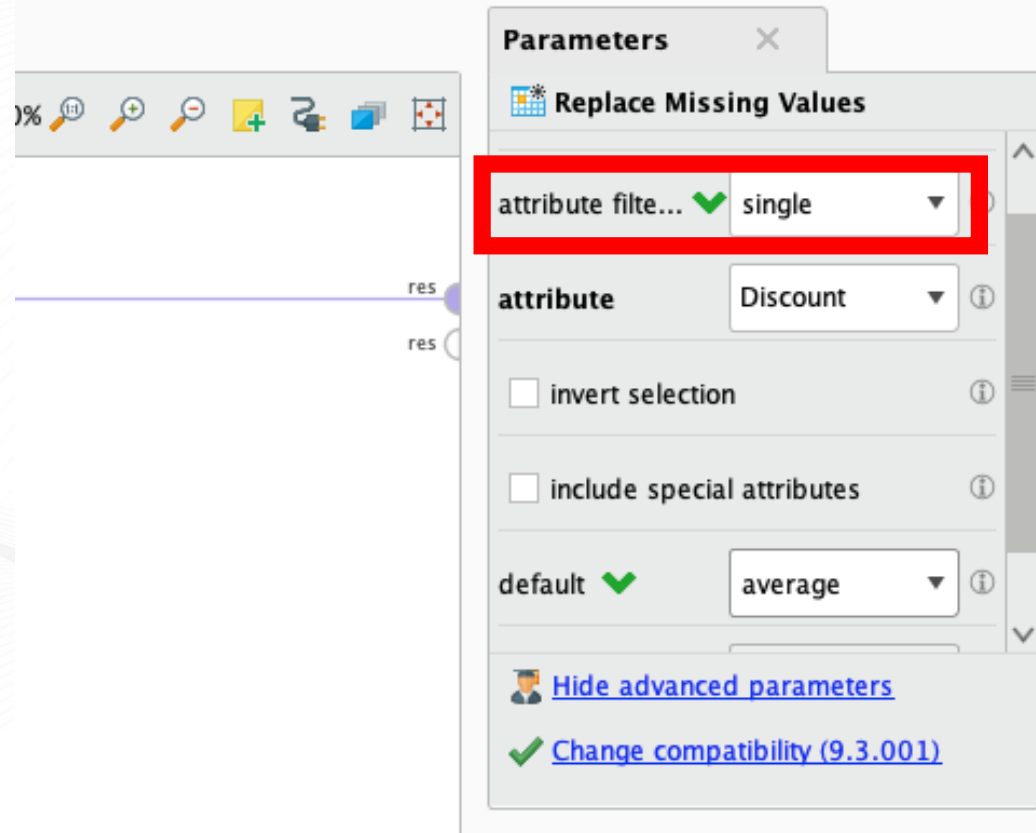
- In the operator tab, search for **Replace Missing Values**, then drag and drop on the line connecting the *Filtering Examples* and the *res* knob.



# Data Preparation

## 2. Imputing Missing Data

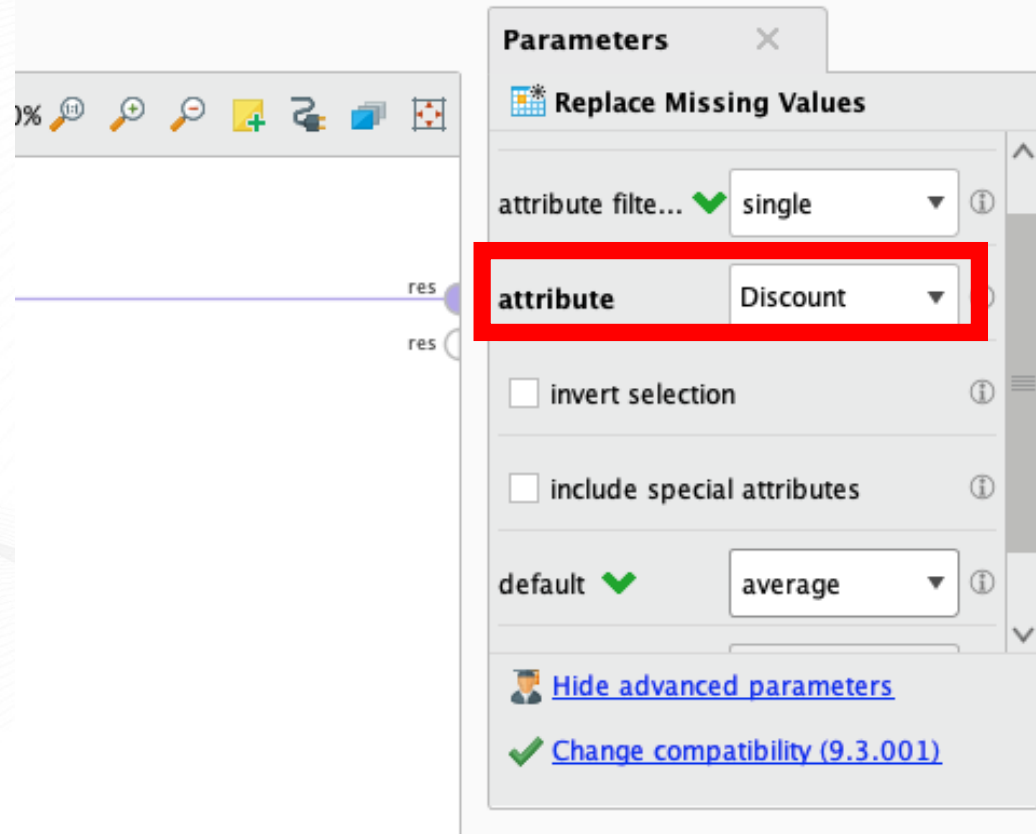
- In the parameter tab, select how many **attribute filter**. Choose **single** if the imputation will apply to a single attribute.



# Data Preparation

## 2. Imputing Missing Data

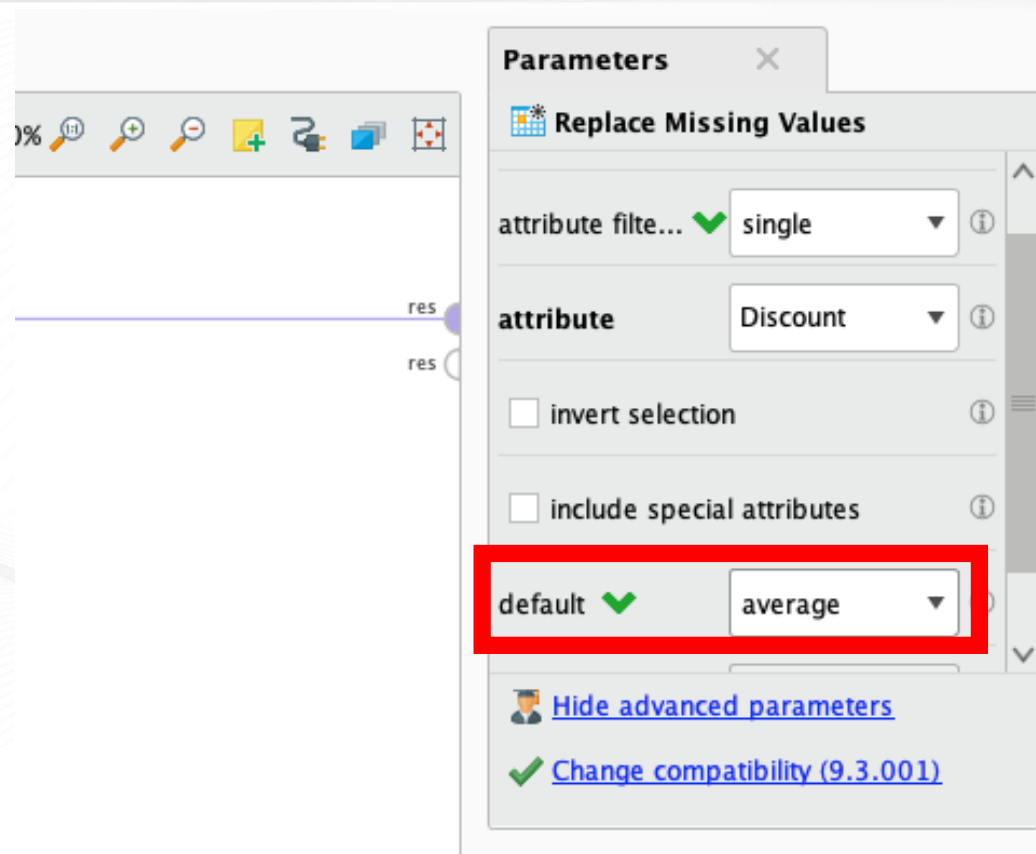
- In the parameter tab, select how many **attribute filter**. Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.



# Data Preparation

## 2. Imputing Missing Data

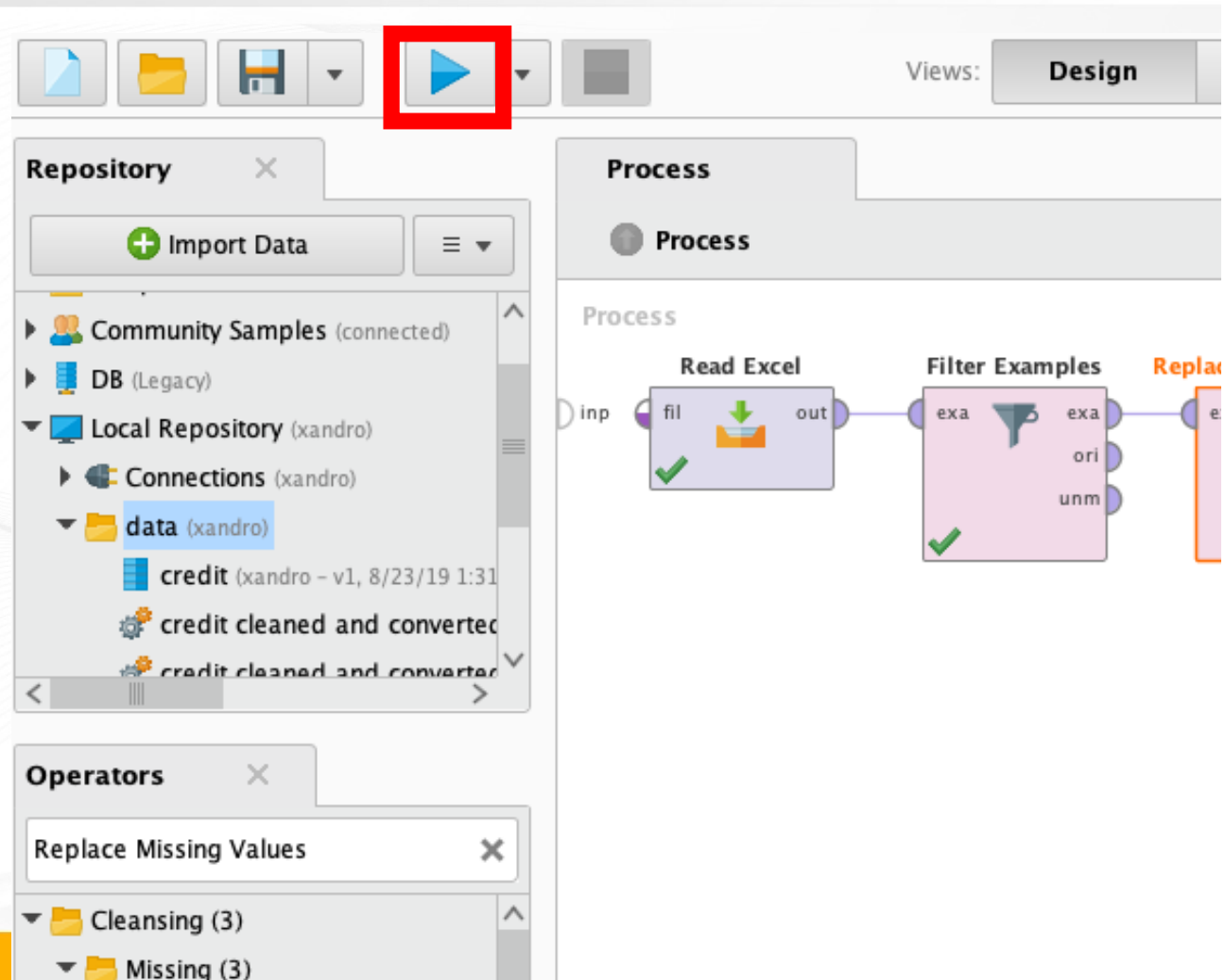
- In the parameter tab, select how many **attribute filter**. Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.
- Select the imputation method in the **Default**.



## Data Preparation

### 2. Imputing Missing Data

- In the parameter tab, select how many **attribute filter**. Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.
- Select the imputation method in the **Default**.
- Click the **Play** button to see result.



The screenshot displays a data preparation software interface. At the top, a toolbar contains several icons, with a blue play button icon highlighted by a red square. Below the toolbar, the interface is divided into several panels:

- Repository:** A tree view showing data sources. The 'data' folder is expanded, showing files like 'credit' and 'credit cleaned and converted'.
- Process:** A workflow diagram showing a sequence of operators: 'Read Excel' (with a green checkmark) and 'Filter Examples' (with a green checkmark). The 'Filter Examples' operator has input ports labeled 'exa' and 'ori', and an output port labeled 'unm'.
- Operators:** A list of available operators, with 'Replace Missing Values' selected.

The 'Views' dropdown at the top right is set to 'Design'.