

Data Mining I

# Optimization with Rapidminer



# Outline

1. Introduction
2. Parameter Optimization
3. Attribute Selection Optimization

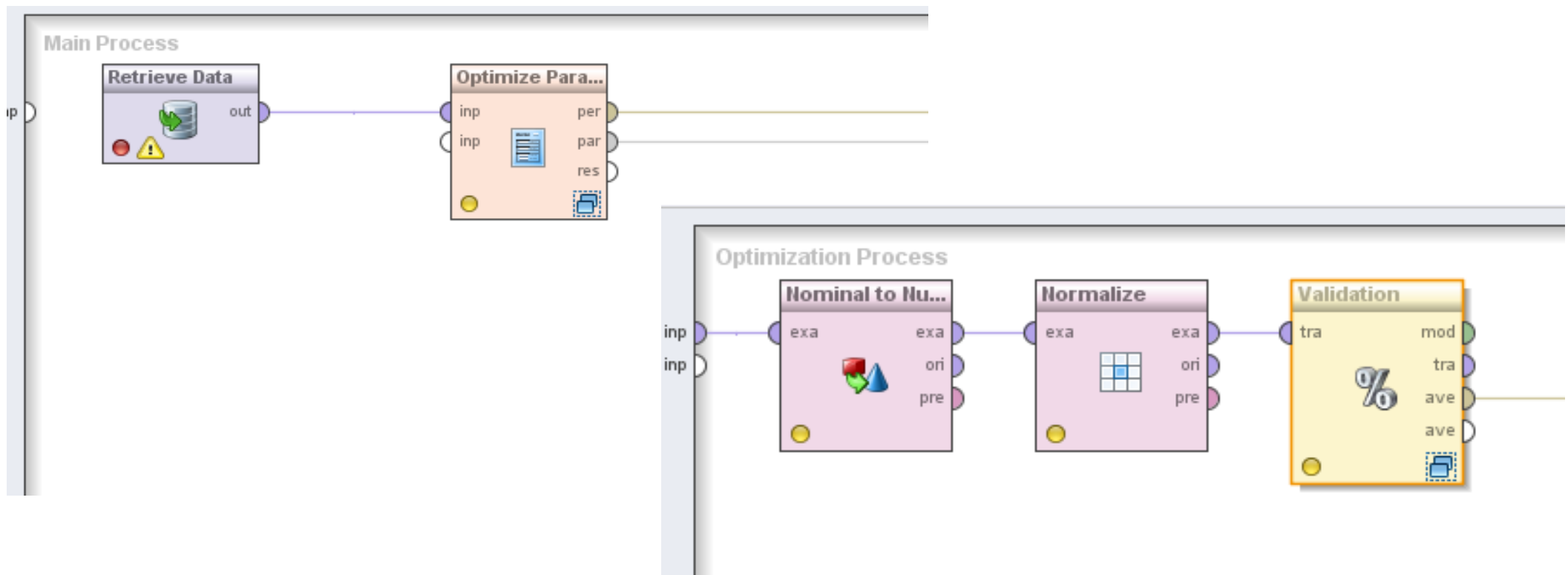
# Introduction

- Why should we optimize?
  - Default configuration does not work always best
  - Operators come with a (large) set of different parameters
  - Dataset attributes might be useful or useless
- Why not manual?
  - Testing all combinations of kernel type and svm type (of LibSVM) means rerunning the process 25 times.
  - Playing around with epsilon in addition (5 different values) leads to 125 possible set-ups
  - But, yes you can do it manually, if you are patient
- What can be optimized?
  - Parameters of Operators
  - Selection of Attributes from the Dataset

# Parameter Optimization

Main Idea:

Let Rapidminer test possible operator parameter combinations for you.



# Parameter Optimization

The screenshot shows the 'Select Parameters: configure operator' dialog box. It is divided into several sections:

- Operators:** A list of operators including 'Nominal to Numerical (Nominal to Numerical)', 'Normalize (Normalize)', 'Validation (X-Validation)', 'SVM (Support Vector Machine (LibSVM))', 'Apply Model (Apply Model)', and 'Performance (Performance (Classifier))'. The 'SVM' operator is selected.
- Parameters:** A list of parameters for the selected operator, including 'degree', 'gamma', 'C', 'nu', 'cache\_size', 'epsilon', 'p', and 'class\_weights'.
- Selected Parameters:** A list of parameters to be optimized, including 'SVM.svm\_type', 'SVM.kernel\_type', and 'SVM.coef0'.
- Grid/Range:** A section for defining the range and steps for the parameters. It includes fields for 'Min' (1), 'Max' (10), 'Steps' (9), and 'Scale' (linear).
- Value List:** A section for defining a list of values to test. It includes a list of values (1 through 9) and a 'Selection setup' section.

Annotations point to specific parts of the dialog:

- List of nested operators:** Points to the 'Operators' list.
- List of parameters of selected operator:** Points to the 'Parameters' list.
- Parameters to Optimize:** Points to the 'Selected Parameters' list.
- Selection setup: Range to Test Values to Test:** Points to the 'Grid/Range' section.
- Final number of combinations!:** Points to the status bar at the bottom, which indicates '3 parameters / 250 combinations selected'.

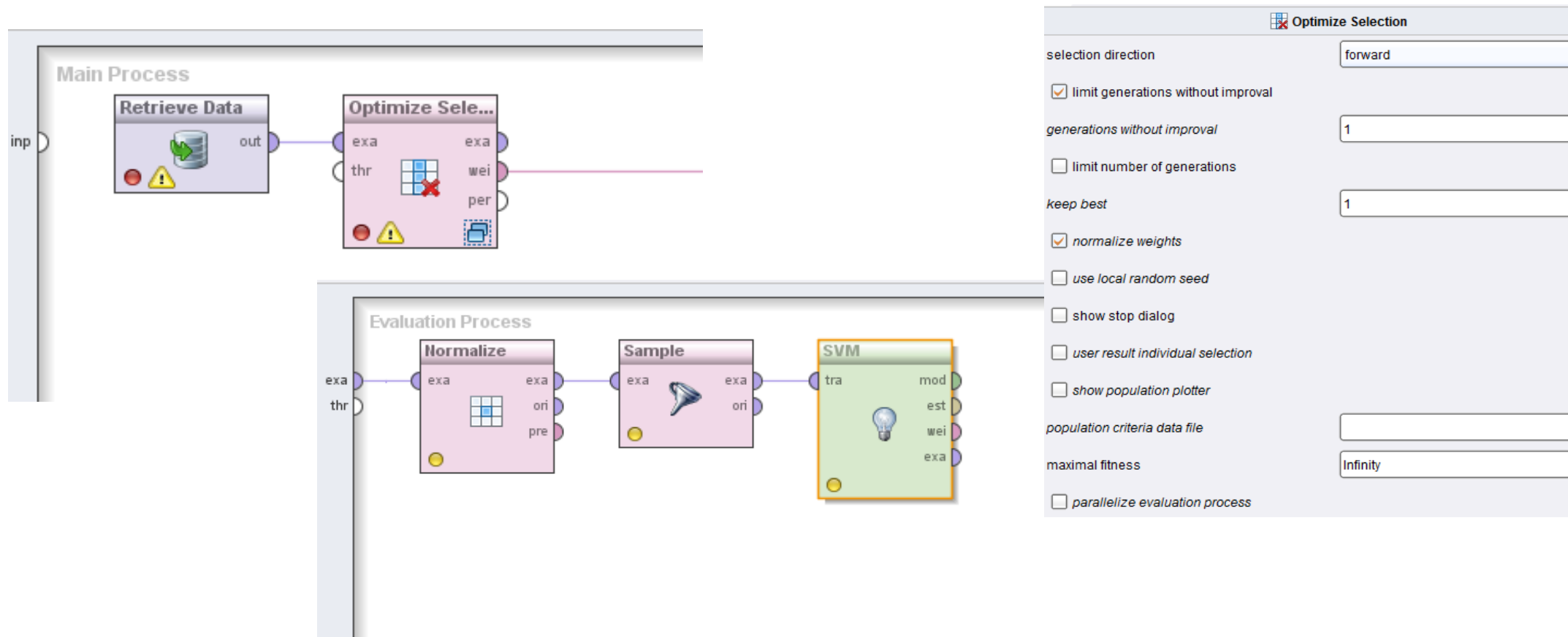
The status bar at the bottom shows '3 parameters / 250 combinations selected' and 'OK' / 'Cancel' buttons.



# Attribute Selection Optimization

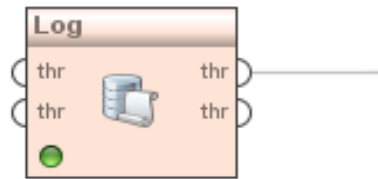
Main Idea:

Let Rapidminer select optimal configuration of attributes for the process and the selected classifier/clustering algorithm



# Interim Results

- Using the *Log* Operator



Edit Parameter List: log

List of key value pairs where the key is the column name and the value specifies the process value to log.

column name			value
Shuffle	Neural Net	parameter	shuffle
Training	Neural Net	parameter	training_cycles
Accuracy	Validation	value	performance

Column Name

Operator

parameter  
or port

Selection of  
output

Log (22 rows, 3 columns)

Shuffle	Training	Accuracy
true	1	0.567
false	1	0.500
true	11	0.708
false	11	0.722
true	21	0.722
false	21	0.723
true	31	0.707
false	31	0.723
true	41	0.682

# Need more information?

- Parameter Optimization YouTube Video:
  - <http://www.youtube.com/watch?v=R5vPrTLMzng>
- Attribute Selection Optimization YouTube Video:
  - Part 1: <http://www.youtube.com/watch?v=7IC3IQEdWxA>
  - Part 2: <http://www.youtube.com/watch?v=j5vhwbLIZWg>



# Questions?

