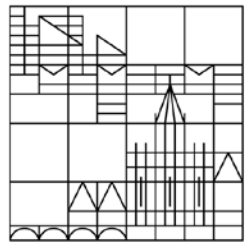
Nycomed Chair for
Bioinformatics and Information Mining

Universität
Konstanz



The New Iris Data – Modular Data Generators

Iris Adä
Michael R. Berthold

[Iris.Adae@Uni-Konstanz.de]

ACM SIGKDD 2010, Washington

Outline

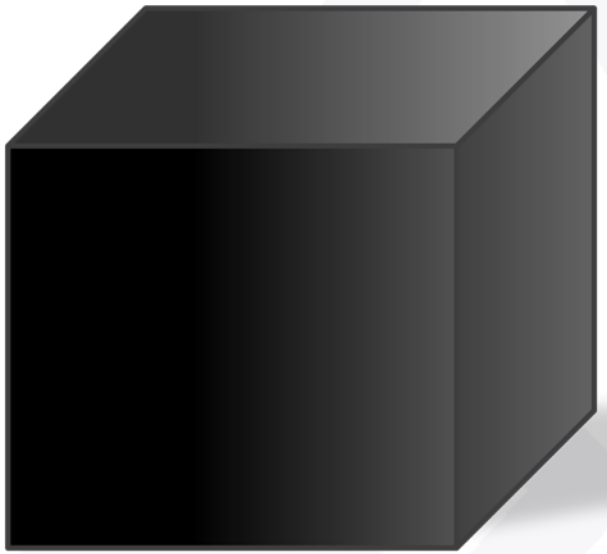
- Motivation
  - Why generate data?
  - What do current generators?
  - Why use modular generators?
- Modular Data Generation
  - Creating clusters (Demo)
  - Creating association rules
  - Combining two motifs
- Outlook

Motivation

- Reasons to generate artificial data
  - Teaching
  - Method Development, Validation, Testing
  - Demonstration

Motivation

- **Current tools**
  - Generate data for one method/purpose
    (cluster, association rules,
    own algorithm evaluating)
  - Difficult to
    - extend,
    - modify or
    - reuse for other purposes

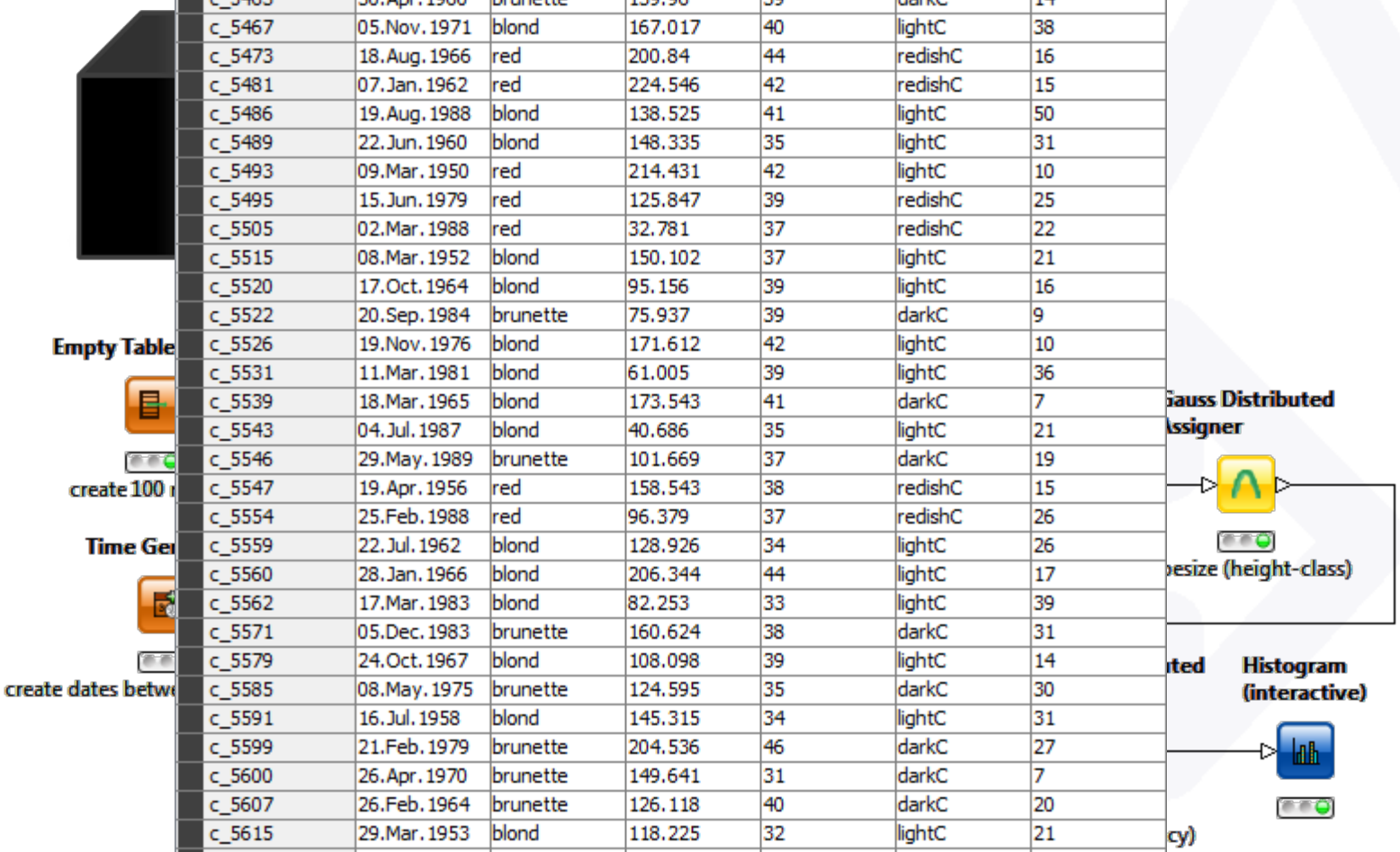


Modular Data Generators

- Splitting t

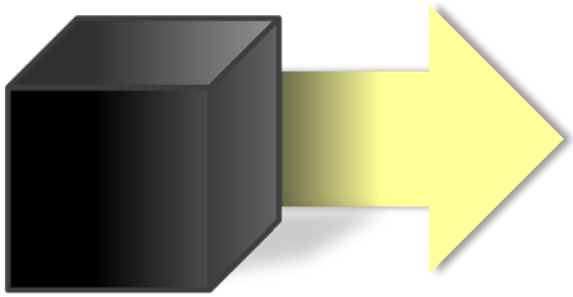
Table "default" - Rows: 3000 | Spec - Columns: 6 | Properties

Row ID	Date a...	haircolor	height	shoesize	agency	nrofjobs
c_5455	05.Jul.1969	blond	125.317	30	lightC	39
c_5458	02.Aug.1964	red	230.497	44	redishC	25
c_5463	30.Apr.1960	brunette	139.96	39	darkC	14
c_5467	05.Nov.1971	blond	167.017	40	lightC	38
c_5473	18.Aug.1966	red	200.84	44	redishC	16
c_5481	07.Jan.1962	red	224.546	42	redishC	15
c_5486	19.Aug.1988	blond	138.525	41	lightC	50
c_5489	22.Jun.1960	blond	148.335	35	lightC	31
c_5493	09.Mar.1950	red	214.431	42	lightC	10
c_5495	15.Jun.1979	red	125.847	39	redishC	25
c_5505	02.Mar.1988	red	32.781	37	redishC	22
c_5515	08.Mar.1952	blond	150.102	37	lightC	21
c_5520	17.Oct.1964	blond	95.156	39	lightC	16
c_5522	20.Sep.1984	brunette	75.937	39	darkC	9
c_5526	19.Nov.1976	blond	171.612	42	lightC	10
c_5531	11.Mar.1981	blond	61.005	39	lightC	36
c_5539	18.Mar.1965	blond	173.543	41	darkC	7
c_5543	04.Jul.1987	blond	40.686	35	lightC	21
c_5546	29.May.1989	brunette	101.669	37	darkC	19
c_5547	19.Apr.1956	red	158.543	38	redishC	15
c_5554	25.Feb.1988	red	96.379	37	redishC	26
c_5559	22.Jul.1962	blond	128.926	34	lightC	26
c_5560	28.Jan.1966	blond	206.344	44	lightC	17
c_5562	17.Mar.1983	blond	82.253	33	lightC	39
c_5571	05.Dec.1983	brunette	160.624	38	darkC	31
c_5579	24.Oct.1967	blond	108.098	39	lightC	14
c_5585	08.May.1975	brunette	124.595	35	darkC	30
c_5591	16.Jul.1958	blond	145.315	34	lightC	31
c_5599	21.Feb.1979	brunette	204.536	46	darkC	27
c_5600	26.Apr.1970	brunette	149.641	31	darkC	7
c_5607	26.Feb.1964	brunette	126.118	40	darkC	20
c_5615	29.Mar.1953	blond	118.225	32	lightC	21
c_5619	30.Jan.1961	brunette	149.374	36	darkC	35
c_5627	11.Sep.1989	red	217.861	39	redishC	14
c_5636	26.Oct.1961	red	54.03	39	redishC	28



Modular Data Generators

- Splitting the generation into small modules



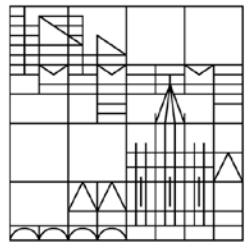
Modular
Data
Generators

- Each module can then be modified, exchanged and/or extended to meet new requirements

- Integrated in an existing data analysis platform (open source, free, reusable, ...)

Nycomed Chair for
Bioinformatics and Information Mining

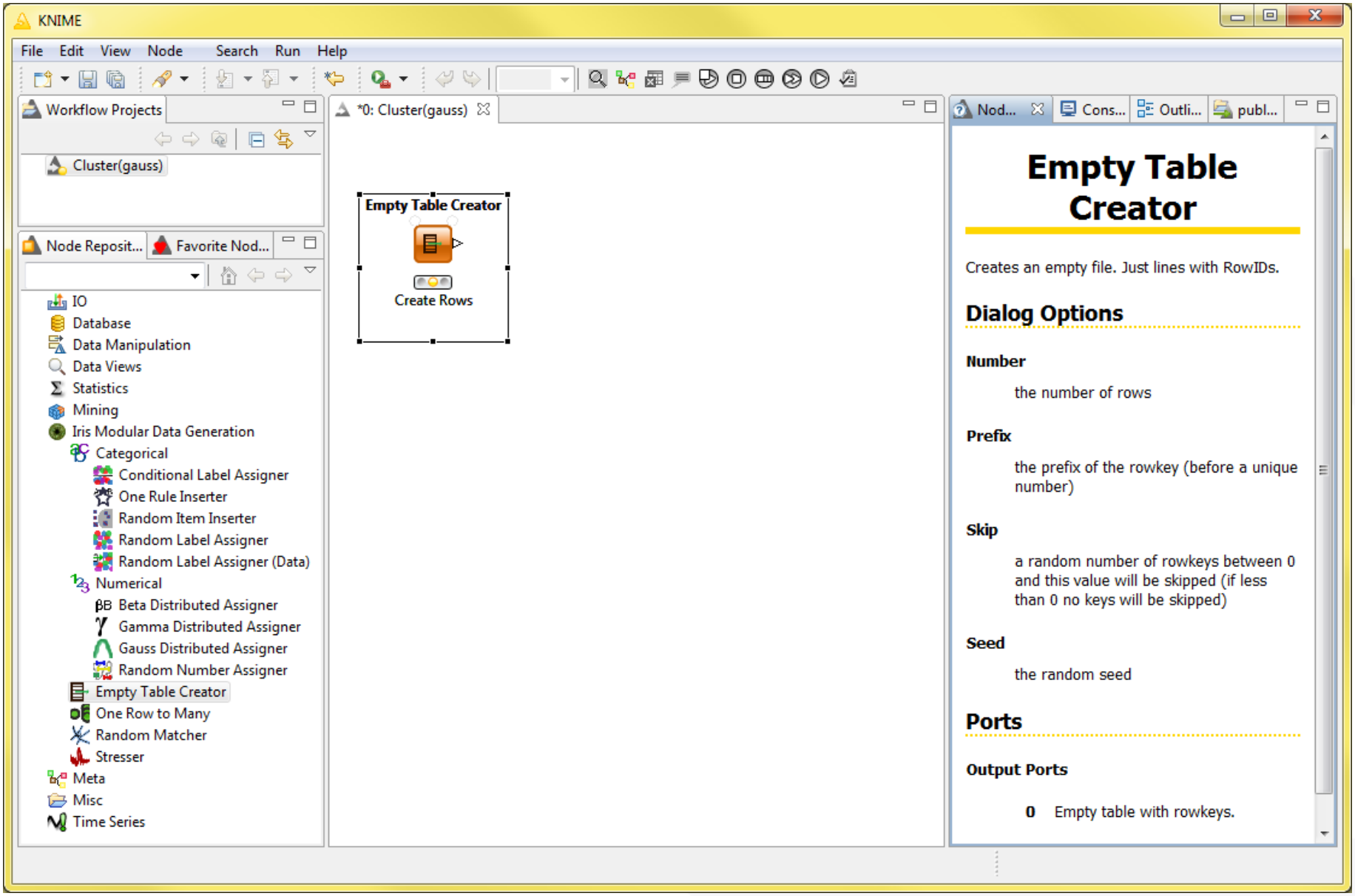
Universität
Konstanz



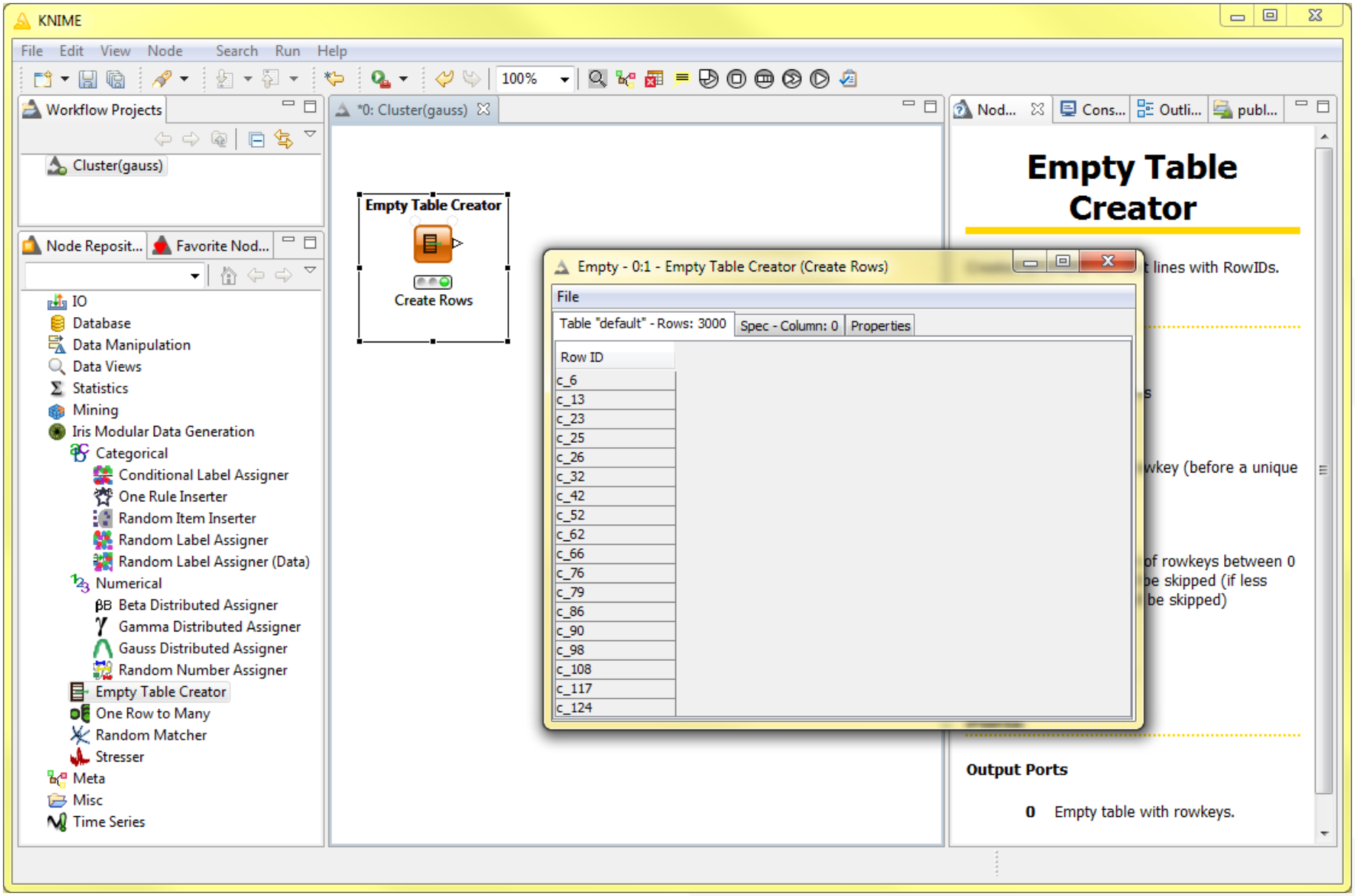
Demo : Creating cluster with MDG's

The New Iris Data – Modular Data Generators

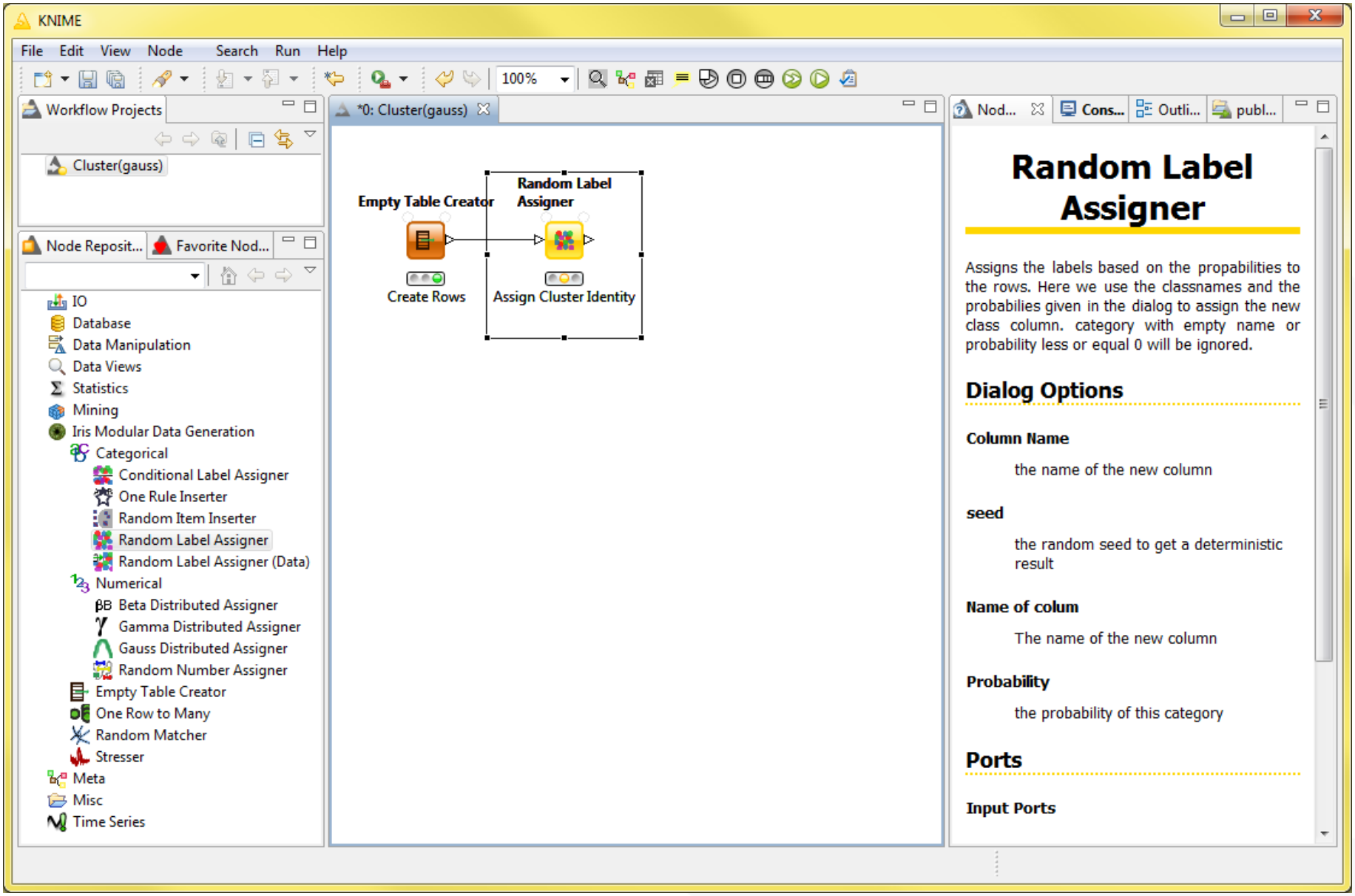
Creating cluster using MDG's



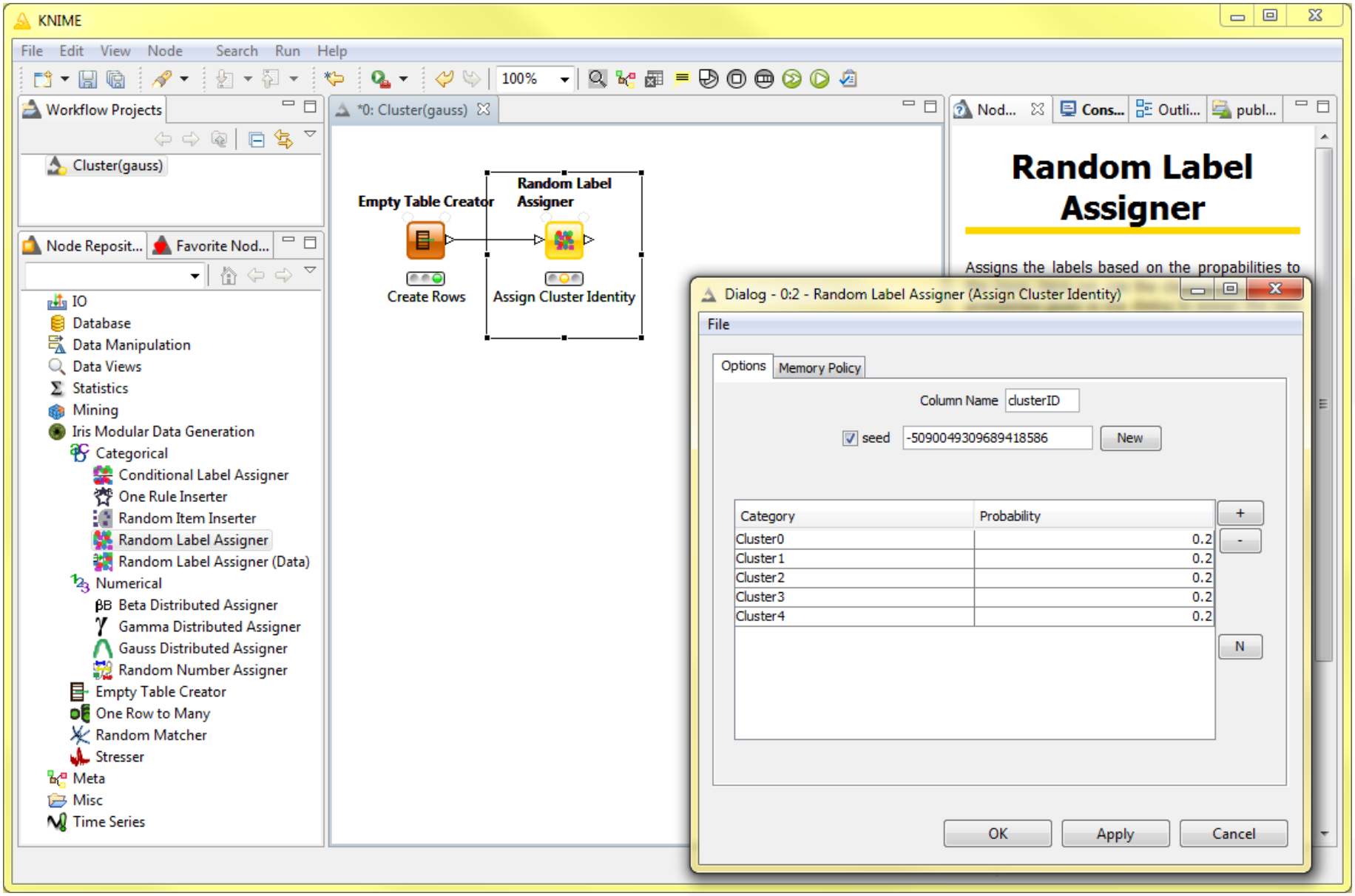
Creating cluster using MDG's



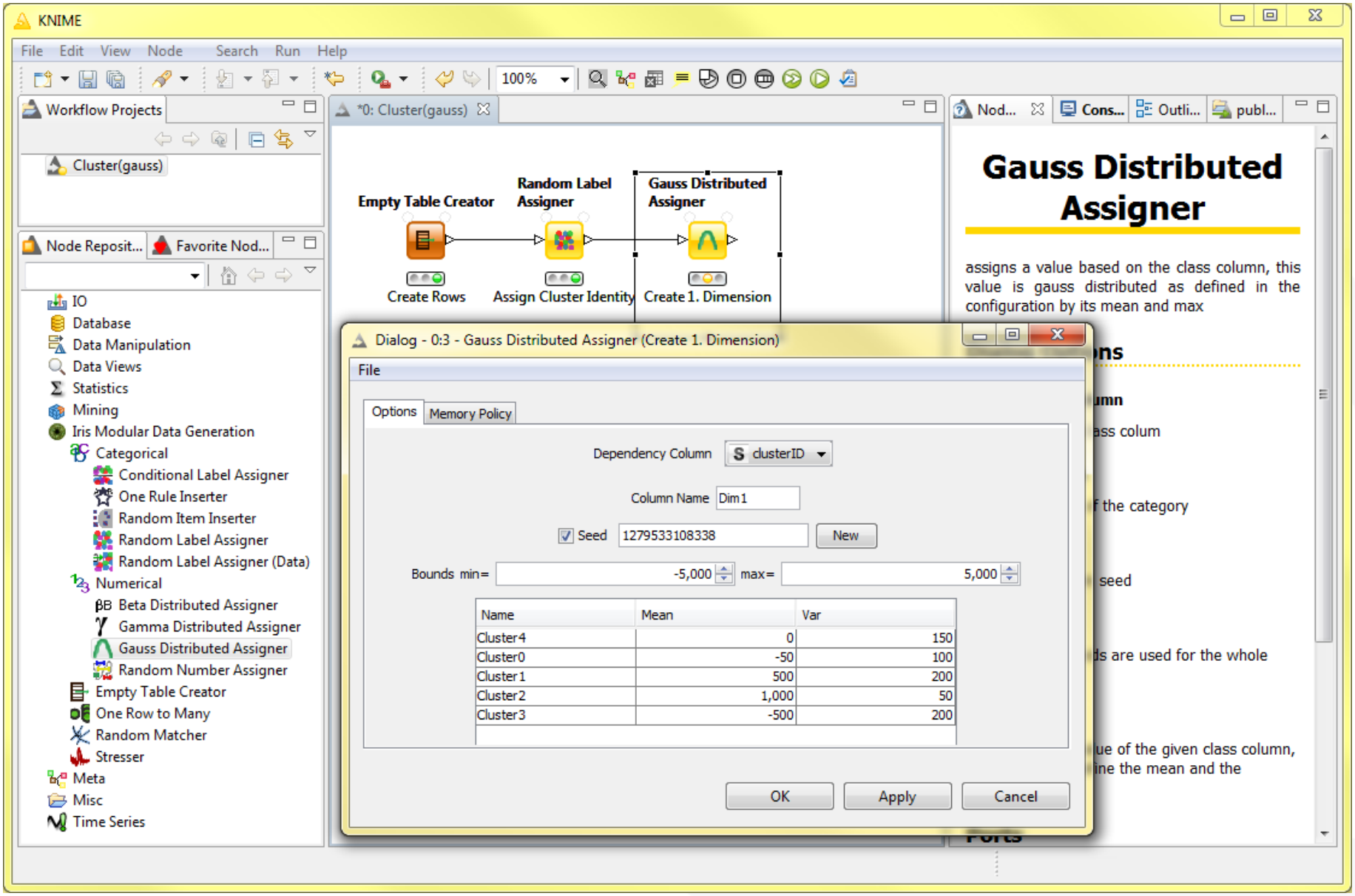
Creating cluster using MDG's



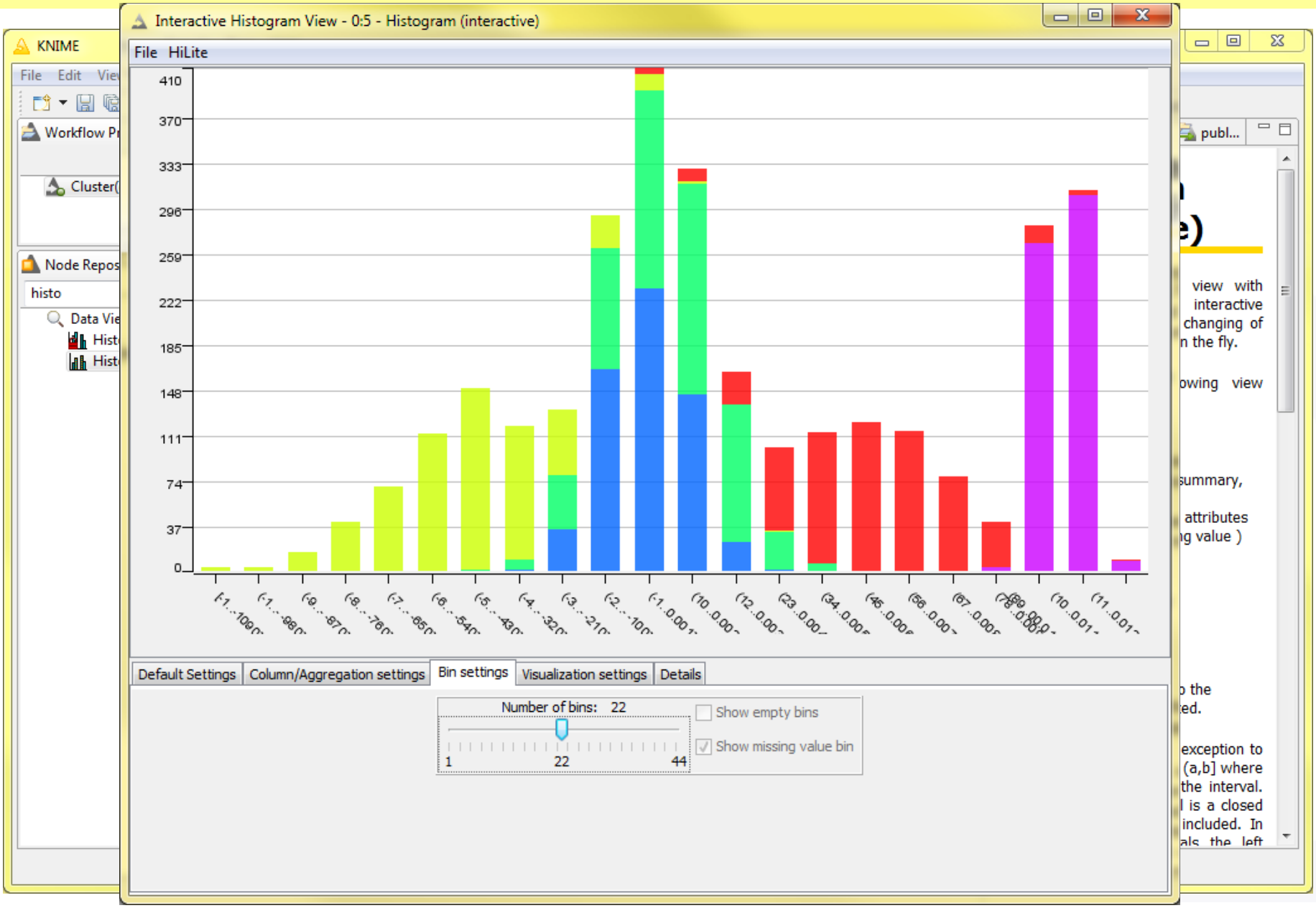
Creating cluster using MDG's



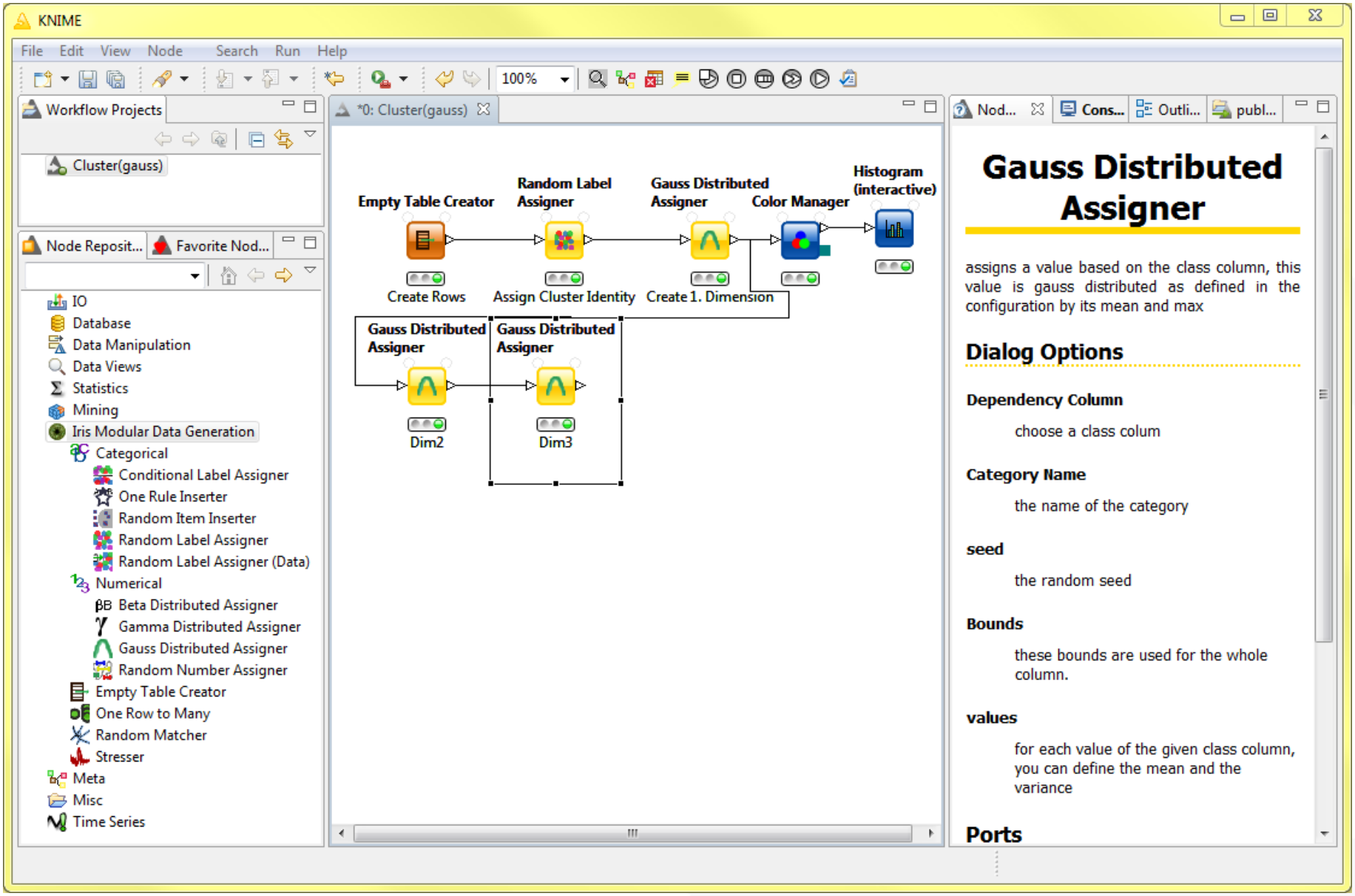
Creating cluster using MDG's



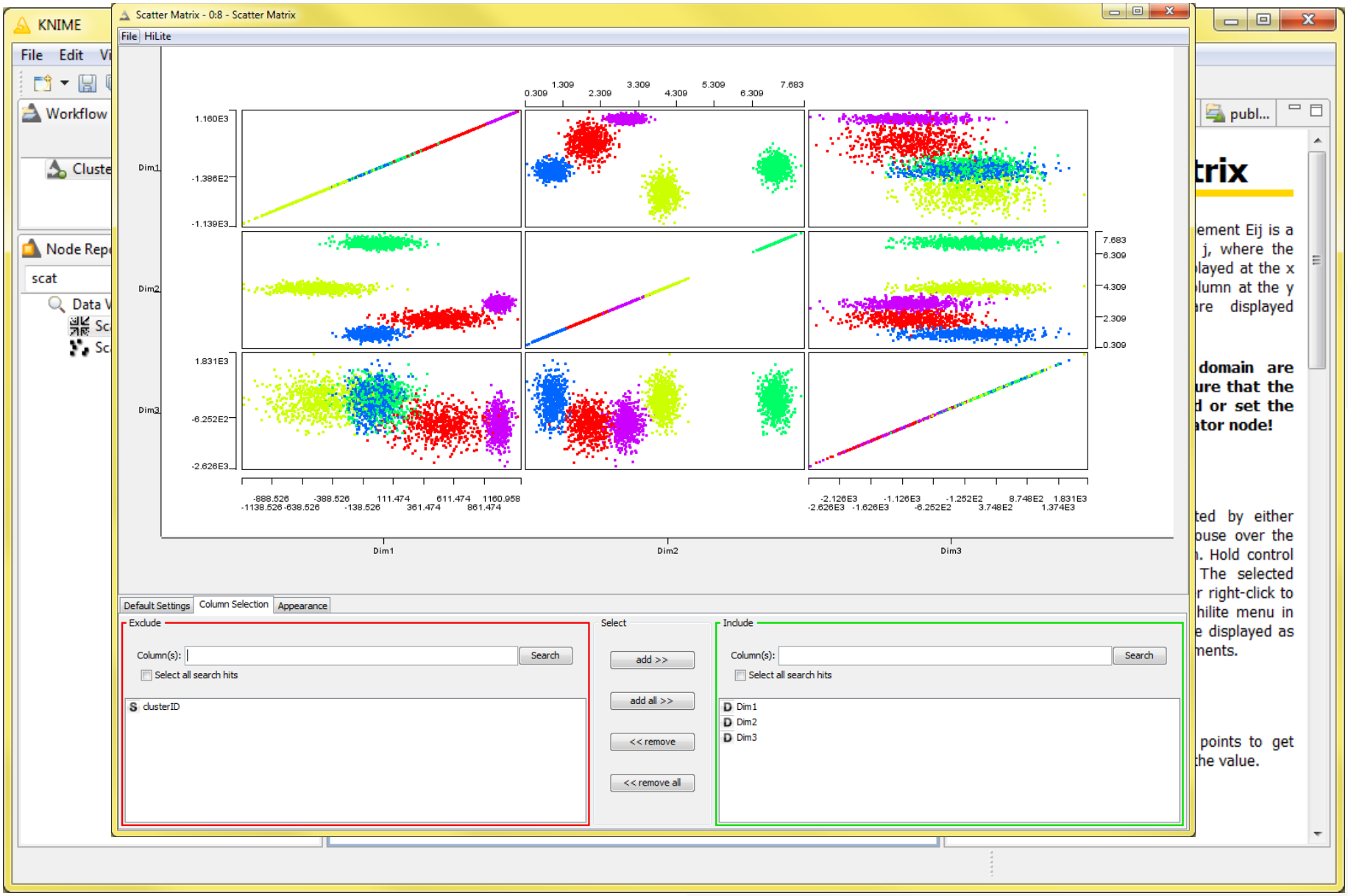
Creating cluster using MDG's



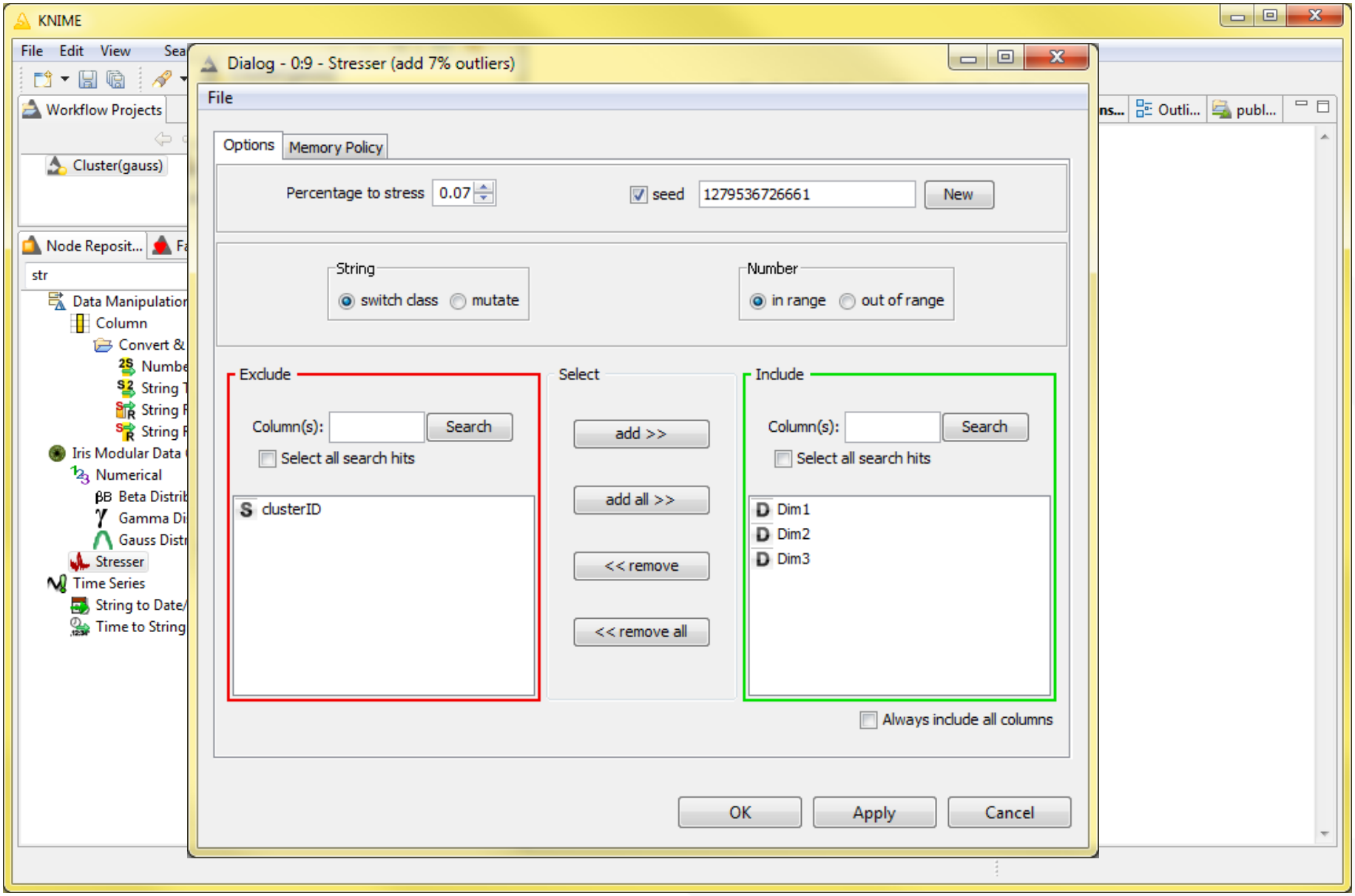
Creating cluster using MDG's



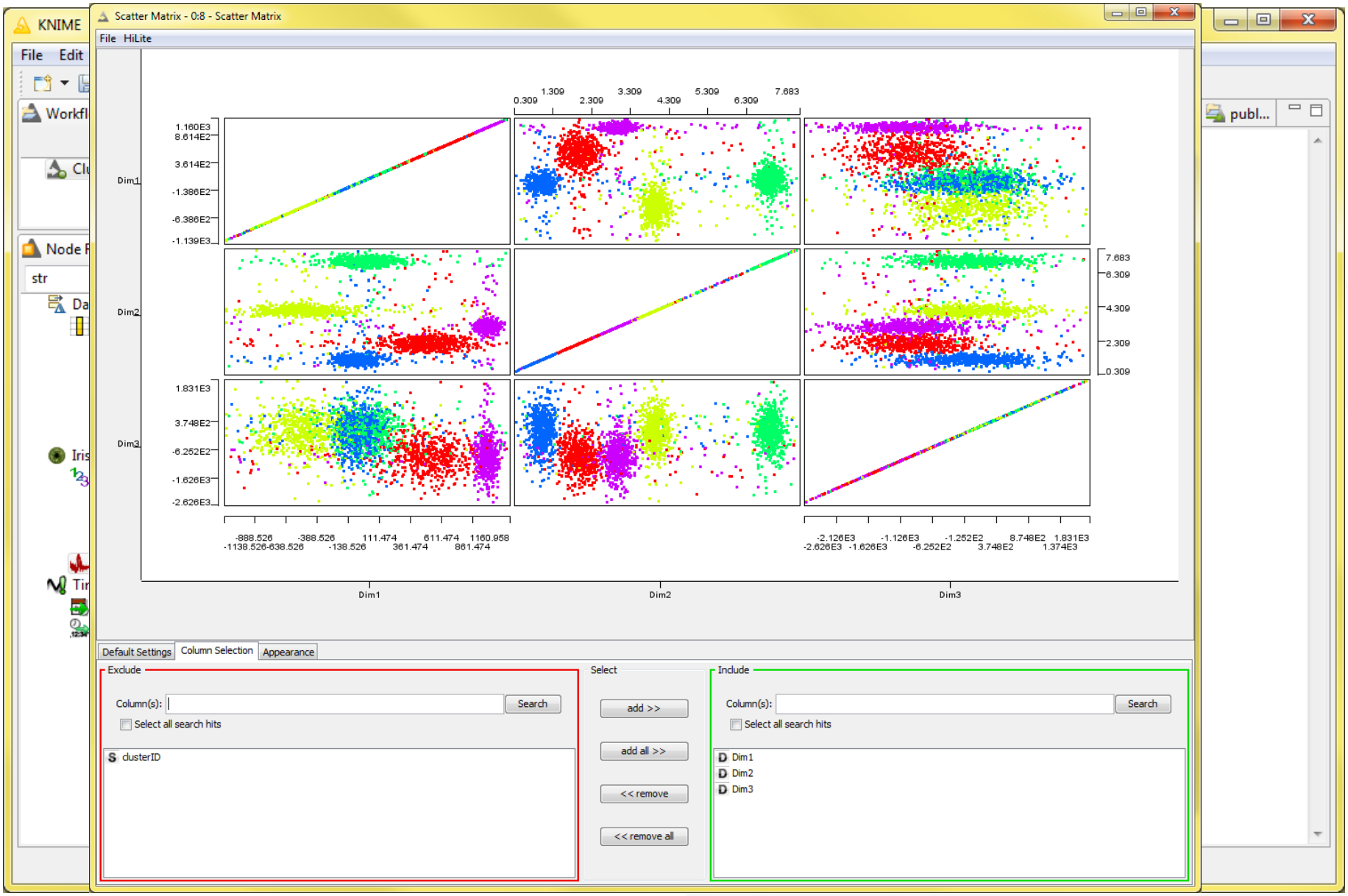
Creating cluster using MDG's



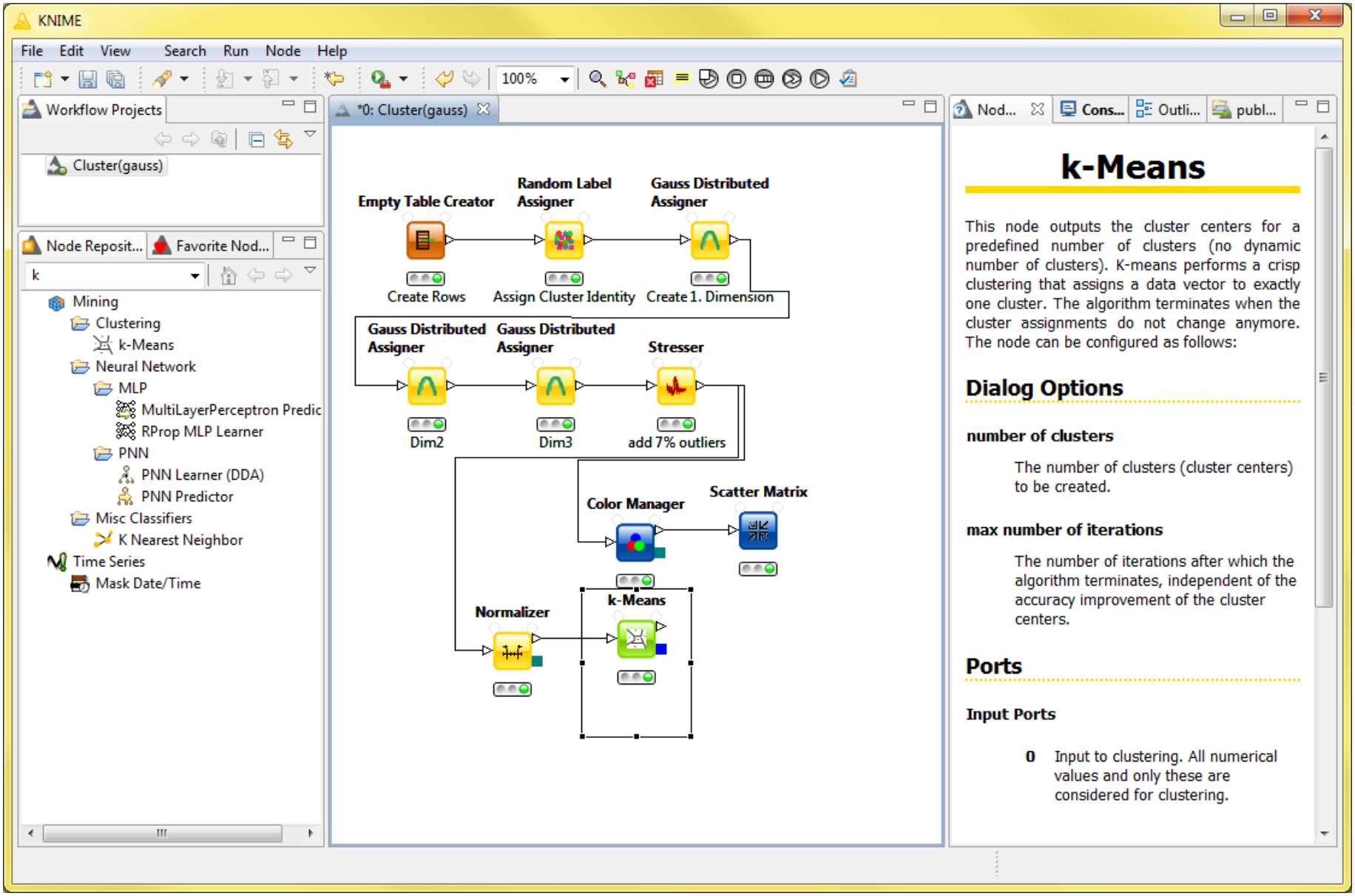
Creating cluster using MDG's



Creating cluster using MDG's



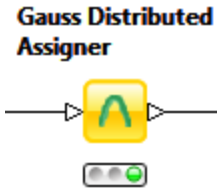
Creating cluster using MDG's



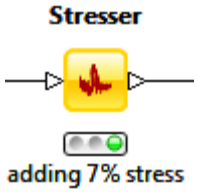
Used Modules

- Creates starting table
  (No columns, only rows with identifiers)

- Assigns random labels.
  (based on specified labels / probabilities)

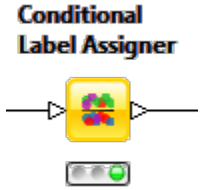


- Assigns random numbers.
  (based on Gaussian distribution)

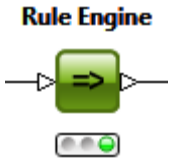


- Adds noise to a column
  (numerical or categorical outliers)

Other Modules



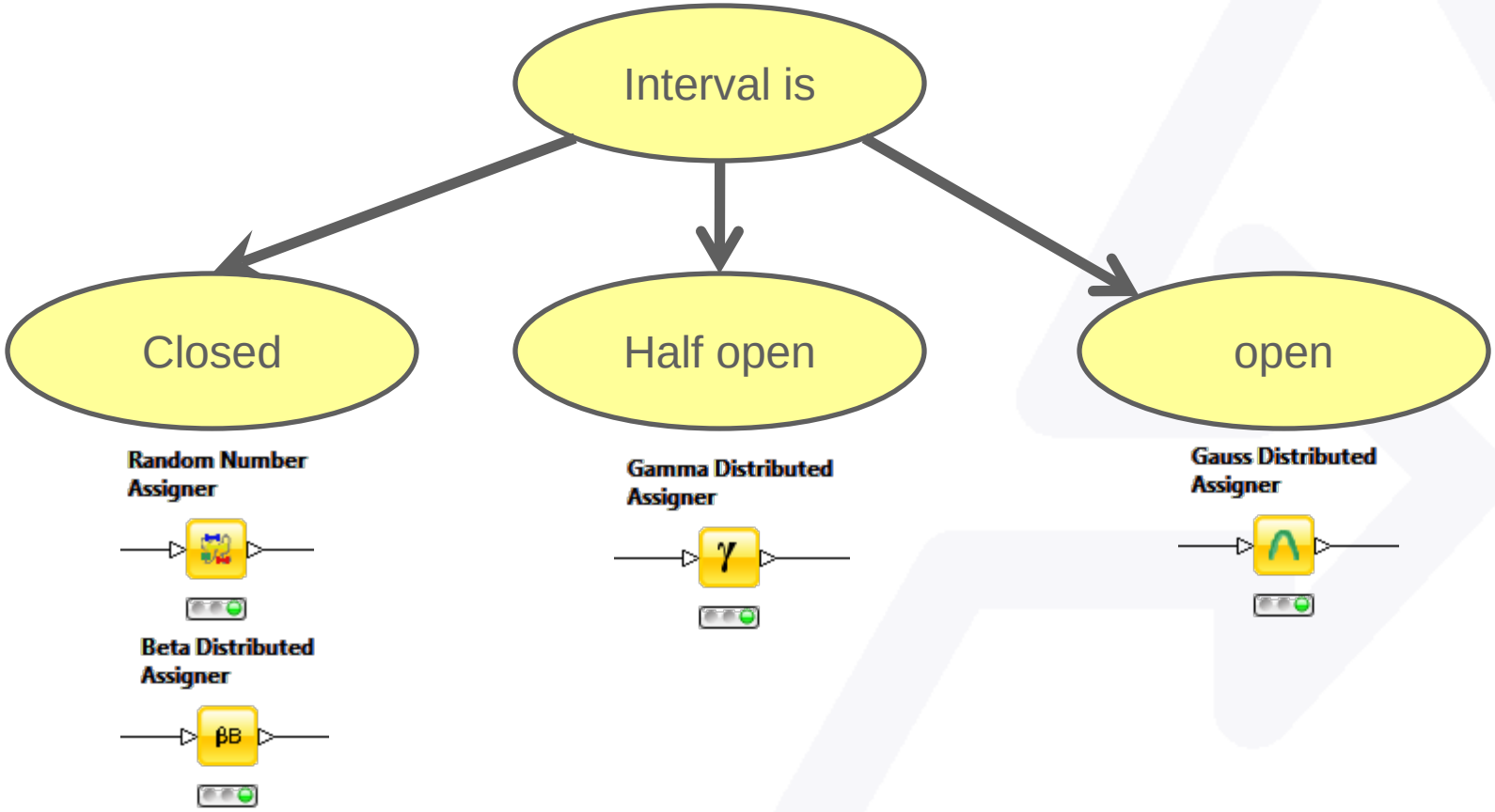
- Assigns conditional random labels (based on values of another column)



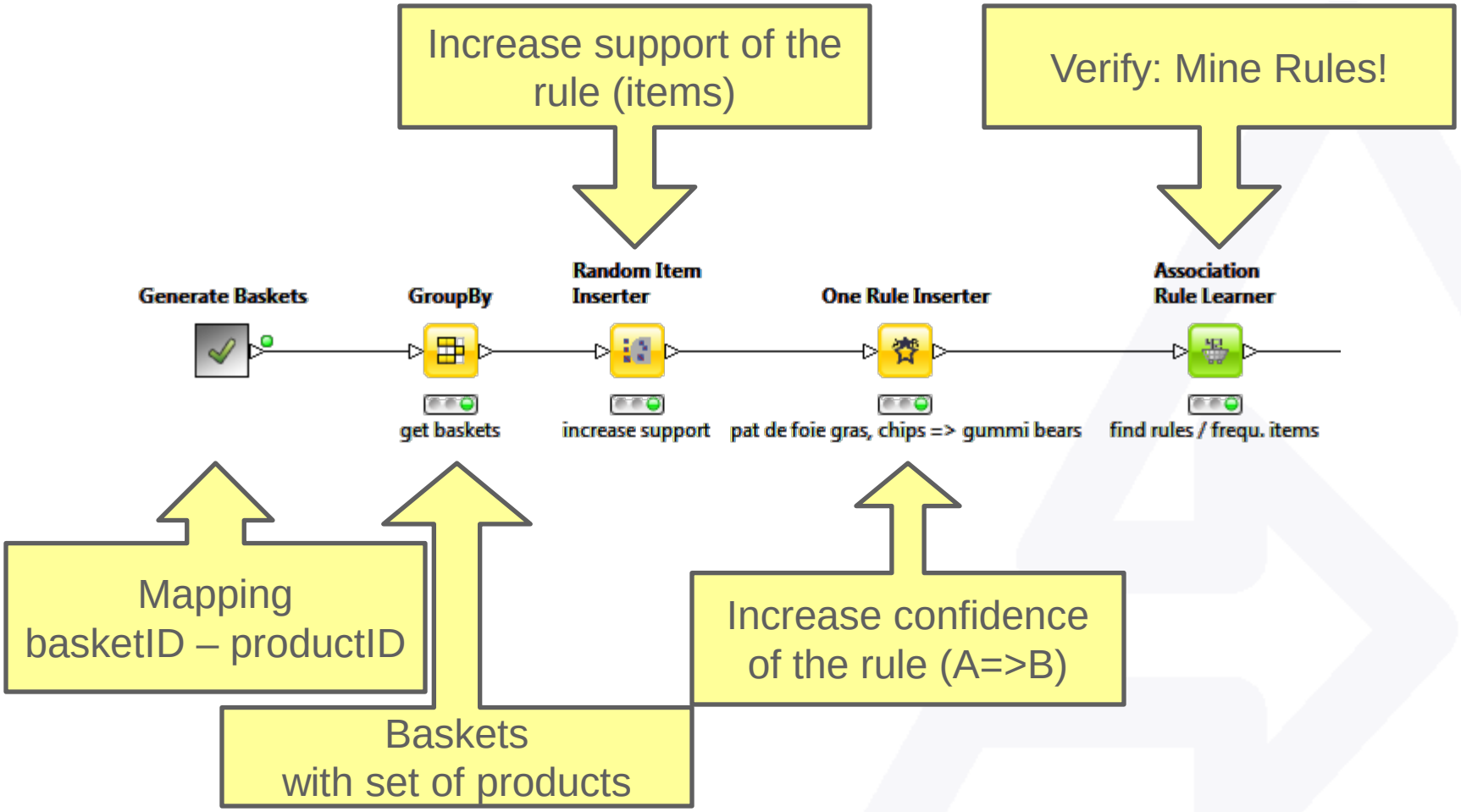
- Inserts various kind of rules

Other Modules for Random Numbers

- Independent: one distribution for entire column
- Dependent: distribution for each value of a second column



Association Rules



Combining tables randomly

- E.g. combine the clusters and the association rules

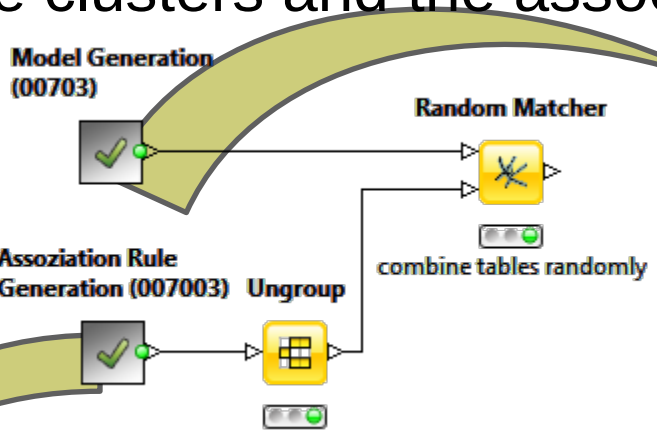


Table "default" - Rows: 2822 | Spec - Colu ns: 2 | Properties | Flow Variables

Row ID	BasketID	(...) pr
Row270	c_117BID_11	[Bio Coke,Ch sandwich,Pinot Noir(Red wine),Ranc
Row271	c_117BID_12	[Bio Coke,Celery,Cran Durian,Egg sandwich,Okra,Ravioli,Tu
Row272	c_117BID_13	[gummi bears]
Row273	c_117BID_14	[Apple cake,Bio Coke,Champagne Esprit du Siècle,Donuts,chips,H
Row274	c_117BID_15	[Chinese Food,Coke,gummi bears,Mouton Rothschild 1945,T-bon
Row275	c_117BID_16	[Broccoli,Champagne Esprit du Siècle,Choclate cake,Dornfelder(R
Row276	c_117BID_17	[Avocado,Bordeaux (Red Wine),Celery,chips,gummi bears,Jackfr
Row277	c_117BID_18	[Broccoli,Cherimoya,Chinese Food,Coconut,Coke,Ice Tea,Ketchu
Row278	c_117BID_19	[chips,Bordeaux (Red Wine),Broccoli,Chicory,Corn,Ravioli,Sauce
Row279	c_117BID_2	[American cheese,Bio Coke,Bio Coke,Brezels,Chinese Food,Frenc
Row280	c_117BID_20	[Beet,Bordeaux (Red Wine),Choclate cake,Donuts,Hamburger,To
Row281	c_117BID_21	[Artichoke,Beet,gummi bears,Bio Coke,apple,Mouton Rothschild 1
Row282	c_117BID_22	[gummi bears]
Row283	c_117BID_23	[Avocado,Avocado,Hamburger,gummi bears,alkopops(wodka/citr
Row284	c_117BID_24	[Egg sandwich,Ham sandwich,apple,chips,clam chowder,duck]
Row285	c_117BID_25	[Apple cake,Okra,Potatoes,Ranch Dressing Mix,Sauvignon Blanc(
Row286	c_117BID_26	[American cheese,Apple cake,chips,Coconut,Egg sandwich,Frenc
Row287	c_117BID_27	[Apple cake,Apricot,Beet,Bio Coke,Clementine,French fries,apple

Table "default" - Rows: 3000 | Spec - Columns | rties | Flow Variables

Row ID	Date a...	haircolor	D h...	sh...	agency	I
c_1916	07.May.1965	red	133.64	35	redishC	14
c_1924	21.Feb.1967	brunette	135.81	40	lightC	27
c_1926	01.Aug.1981	red	211.936	41	redishC	14
c_1932	23.Apr.1972	blond	185.083	41	darkC	18
c_1942	09.May.1956	brunette	155.215	38	darkC	13
c_1950	15.Feb.1974	brunette	34.312	37	darkC	12
c_1960	25.Aug.1982	brunette	234.306	44	lightC	18
c_1966	21.Oct.1980	red	199.217	46	redishC	14
c_1970	03.Jun.1974	blond	105.104	39	lightC	11
c_1979	03.Mar.1960	brunette	104.911	36	darkC	17
c_1982	06.Aug.1973	brunette	219.129	39	darkC	25
c_1986	23.Dec.1960	brunette	196.091	43	redishC	17
c_1988	11.Sep.1981	brunette	231.792	41	darkC	15
c_1994	24.Mar.1951	blond	246.774	46	redishC	21
c_1995	11.Mar.1978	blond	247.151	44	lightC	17
c_2003	03.Aug.1966	red	114.957	38	redishC	21
c_2008	26.Nov.1971	blond	241.238	39	darkC	19
c_2014	17.Jun.1970	blond	203.722	40	lightC	28

Combining tables randomly

- E.g. combine the clusters and the association rules

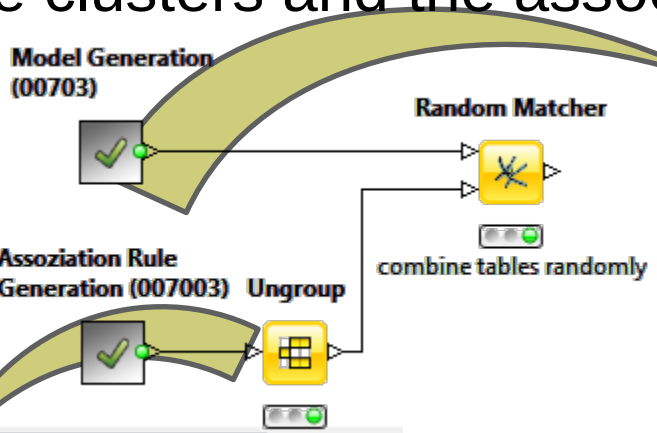


Table "default" - Rows: 46383 | Spec - Columns: 2 | Properties | Flow Variables

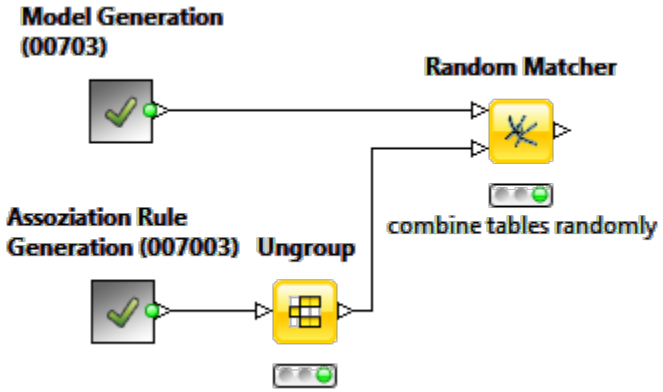
Row ID	BasketID	product
Row301_2	c_117BID_4	Cherry ke
Row301_3	c_117BID_4	apple
Row301_4	c_117BID_4	Papay
Row301_5	c_117BID_4	apples
Row301_6	c_117BID_4	omlett
Row301_7	c_117BID_4	wodka
Row302_1	c_117BID_40	apple
Row302_2	c_117BID_40	Clementine
Row302_3	c_117BID_40	Ham sandwich
Row302_4	c_117BID_40	chips
Row302_5	c_117BID_40	Plum
Row302_6	c_117BID_40	glenfiddich whisky
Row302_7	c_117BID_40	pork haunch
Row302_8	c_117BID_40	white truffles
Row302_9	c_117BID_40	wodka
Row303_1	c_117BID_5	Melon
Row303_2	c_117BID_5	Toast
Row303_3	c_117BID_5	Tuna

Table "default" - Rows: 3000 | Spec - Columns: 6 | Prop | Flow Variables

Row ID	Date a...	haircolor	height	...	agency	nrofj
c_1916	07.May.1965	red	133.64		redishC	14
c_1924	21.Feb.1967	brunette	135.81		lightC	27
c_1926	01.Aug.1981	red	211.936	41	redishC	14
c_1932	23.Apr.1972	blond	185.083	41	darkC	18
c_1942	09.May.1956	brunette	155.215	38	darkC	13
c_1950	15.Feb.1974	brunette	34.312	37	darkC	12
c_1960	25.Aug.1982	brunette	234.306	44	lightC	18
c_1966	21.Oct.1980	red	199.217	46	redishC	14
c_1970	03.Jun.1974	blond	105.104	39	lightC	11
c_1979	03.Mar.1960	brunette	104.911	36	darkC	17
c_1982	06.Aug.1973	brunette	219.129	39	darkC	25
c_1986	23.Dec.1960	brunette	196.091	43	redishC	17
c_1988	11.Sep.1981	brunette	231.792	41	darkC	15
c_1994	24.Mar.1951	blond	246.774	46	redishC	21
c_1995	11.Mar.1978	blond	247.151	44	lightC	17
c_2003	03.Aug.1966	red	114.957	38	redishC	21
c_2008	26.Nov.1971	blond	241.238	39	darkC	19
c_2014	17.Jun.1970	blond	203.722	40	lightC	28

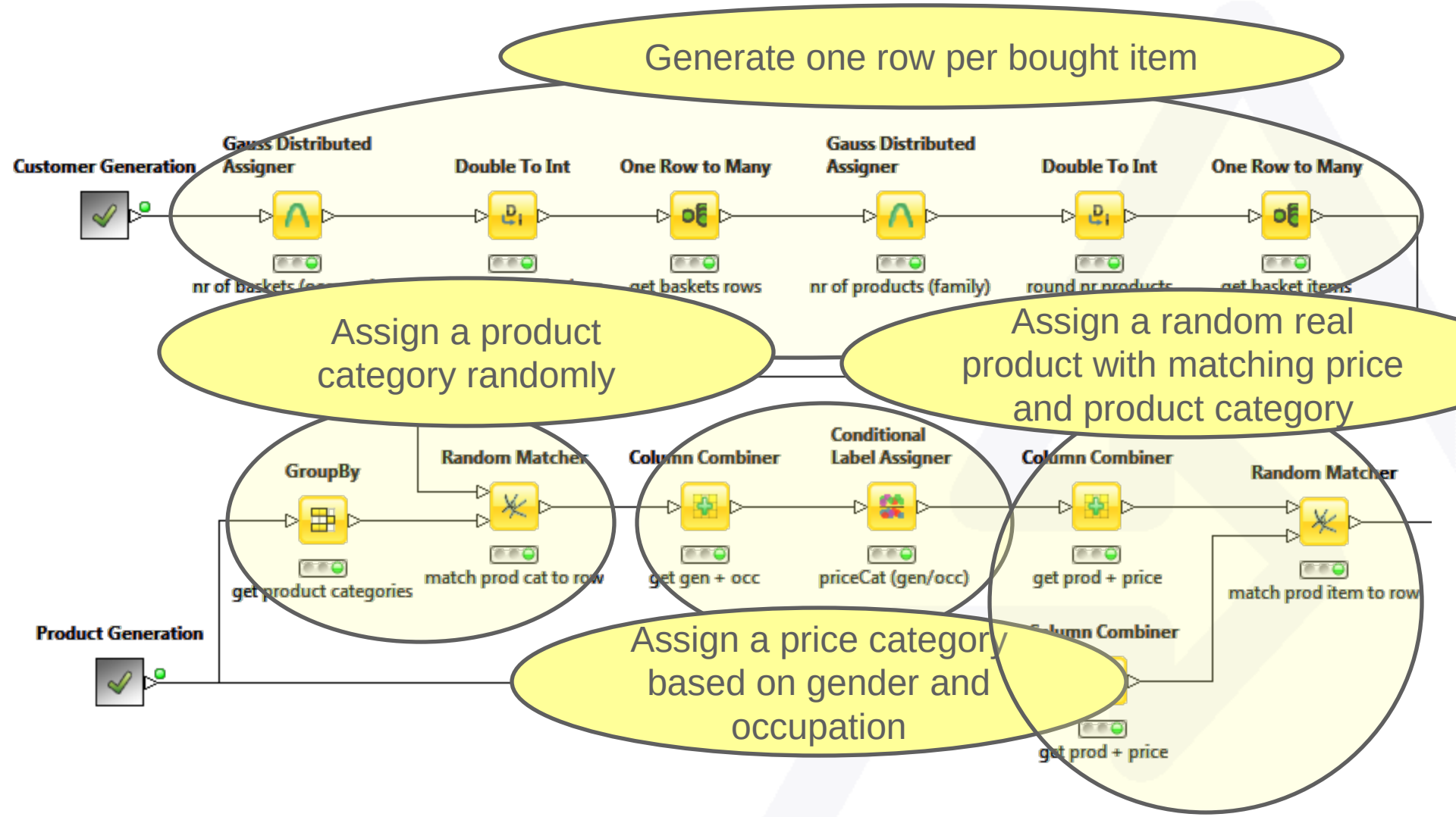
Combining tables randomly

- E.g. combine the clusters and the association rules



Row ID	Date a...	haircolor	height	shoesize	agency	nrofjobs	BasketID	product
c_1916	07.May.1965	red	133.64	35	redishC	14	c_735BID_5	turkey
c_1924	21.Feb.1967	brunette	135.81	40	lightC	27	c_536BID_4	Jackfruit
c_1926	01.Aug.1981	red	211.936	41	redishC	14	c_868BID_6	Coke
c_1932	23.Apr.1972	blond	185.083	41	darkC	18	c_613BID_21	Sharon Fruit (Persimmon)
c_1942	09.May.1956	brunette	155.215	38	darkC	13	c_673BID_10	Persimmon
c_1950	15.Feb.1974	brunette	34.312	37	darkC	12	c_649BID_2	chips
c_1960	25.Aug.1982	brunette	234.306	44	lightC	18	c_1005BID_3	lamm careé
c_1966	21.Oct.1980	red	199.217	46	redishC	14	c_284BID_2	liquorice
c_1970	03.Jun.1974	blond	105.104	39	lightC	11	c_162BID_12	gummibears
c_1979	03.Mar.1960	brunette	104.911	36	darkC	17	c_692BID_2	Apple Juice
c_1982	06.Aug.1973	brunette	219.129	39	darkC	25	c_179BID_2	mussels
c_1986	23.Dec.1960	brunette	196.091	43	redishC	17	c_974BID_11	ostrich eggs
c_1988	11.Sep.1981	brunette	231.792	41	darkC	15	c_986BID_9	chicken
c_1994	24.Mar.1951	blond	246.774	46	redishC	21	c_507BID_5	milk lactosefree
c_1995	11.Mar.1978	blond	247.151	44	lightC	17	c_692BID_49	Cherry coke
c_2003	03.Aug.1966	red	114.957	38	redishC	21	c_59BID_1	wodka
c_2008	26.Nov.1971	blond	241.238	39	darkC	19	c_260BID_2	pork
c_2014	17.Jun.1970	blond	203.722	40	lightC	28	c_437BID_10	Riesling(White wine)
c_2018	07.Apr.1983	blond	217.181	43	lightC	47	c_394BID_4	Jackfruit

Generate Customer Data



- All generators use (user specifiable) random seeds:
  - Deterministic results
  - Allows to generate different data sets following the same pattern.
- KNIME Workflow variables allow to control global settings
- Batch Execution for automatic data generation

Summary / Outlook

Modular Data Generation in KNIME:

- Visual documentation of complex data generation processes
- Repeatability
- Extendibility

Ongoing Work:

- Add functionality to generate time-dependent data

Available now (nodes and workflows):

http://www.knime.org/datageneration

(demo tonight at the poster)