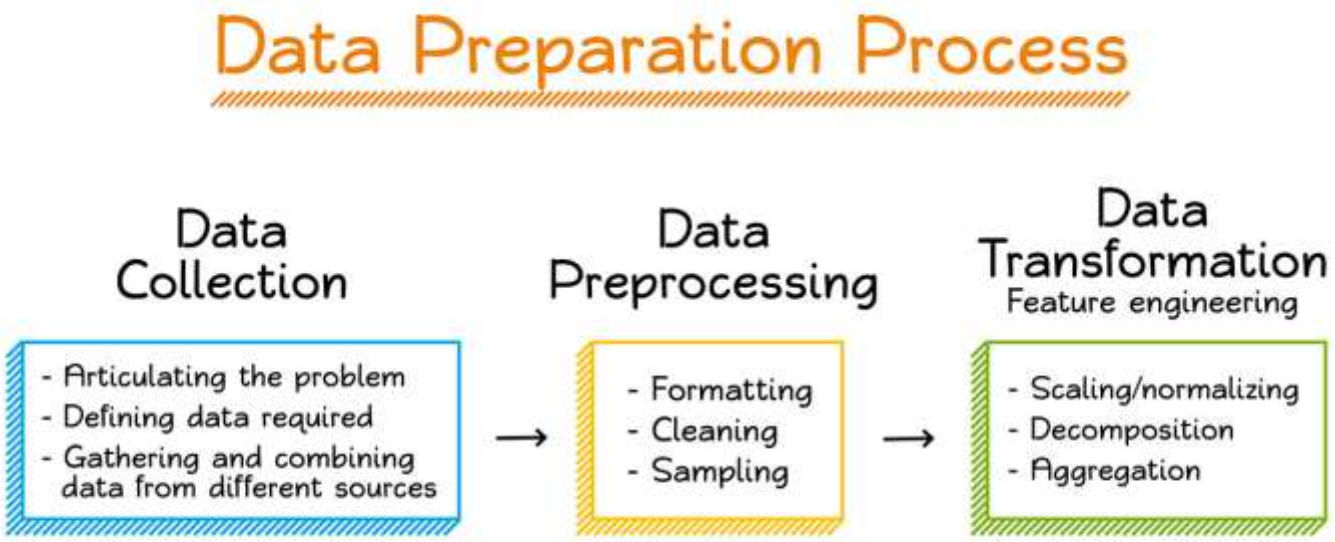
DATA MINING

PERTEMUAN Ke-5

Persiapan Data

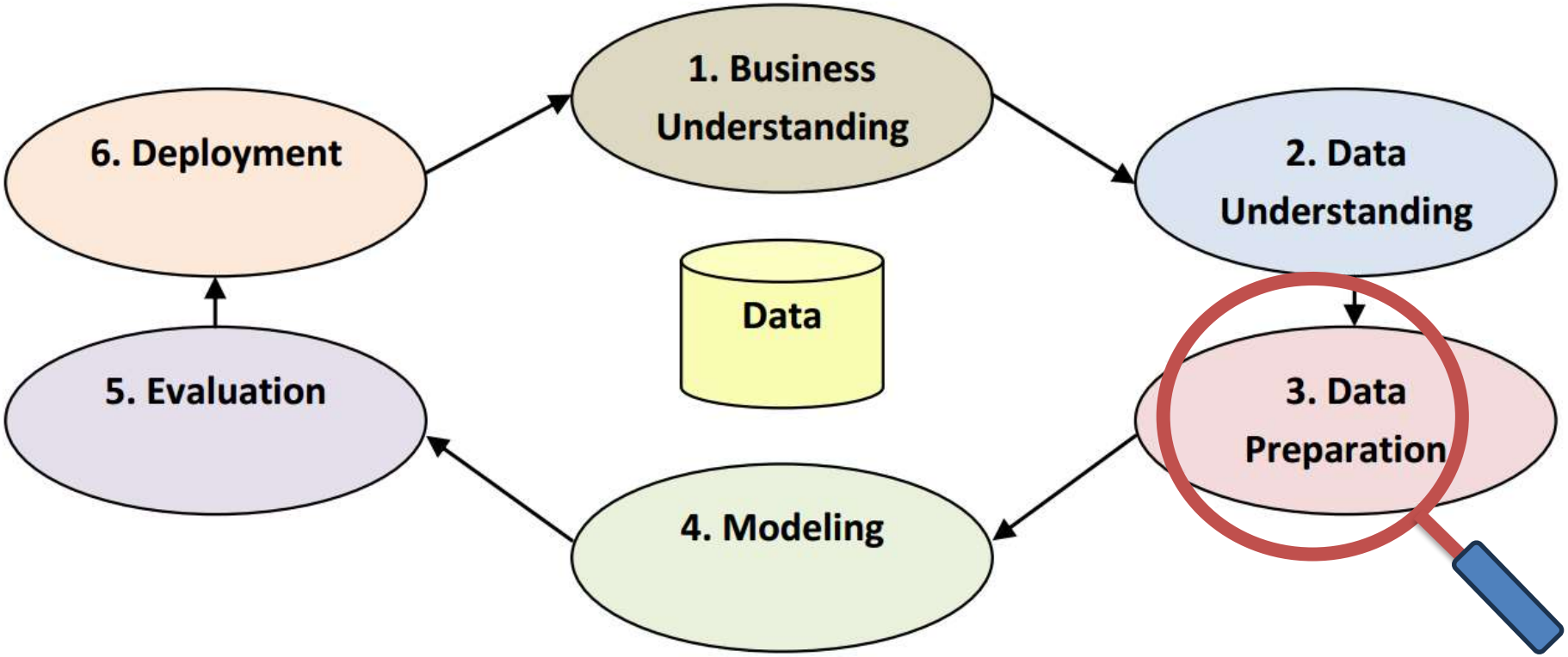
9. Persiapan Data

1 Data Cleaning
2 Data Reduction
- Data Transformation and Data Discretization
- Data Integration



I.3

CRISP-DM



Versi : 01

Why Preprocess the Data?

Pengukuran kualitas data: Sudut Pandang multidimensi

- Akurasi: benar atau salah, akurat atau tidak
- Kelengkapan: tidak tercatat, tidak tersedia,…
- Konsistensi: ada yang diubah tetapi ada yang tidak,…
- Ketepatan waktu: pembaruan tepat waktu?
- Kepercayaan: seberapa dapat dipercaya kebenaran datanya?
- Interpretabilitas: seberapa mudah data dapat dipahami?

I.5 Major Tasks in Data Preprocessing

- Pembersihan data
  - Isi nilai yang hilang
  - Menghaluskan data nois
  - Identifikasi atau hapus outlier
  - Selesaikan inkonsistensi
- Reduksi data
  - Pengurangan dimensi
  - Pengurangan angka
  - Kompresi data
- Transformasi data dan diskritisasi data
  - Normalisasi
  - Pembuatan konsep hierarki
- Integrasi data
  - Integrasi beberapa database atau file

Data Preparation Law (Data Mining Law 3)

"Persiapan data lebih dari separuh proses data Mining"

- Sebagian besar upaya dalam proyek data mining dihabiskan untuk akuisisi dan persiapan data, dan perkiraan informal bervariasi dari 50 hingga 80 persen

Tujuan penyiapan data adalah:

- Untuk memasukkan data ke dalam bentuk di mana pertanyaan data mining dapat diajukan
- Untuk memudahkan teknik analisis (seperti algoritma data mining) dalam menjawabnya

1. Data Cleaning

- Data di Dunia Nyata Itu Kotor: Banyak data yang berpotensi salah, misalnya kesalahan instrumen, kesalahan manusia atau komputer, kesalahan transmisi

Misal:

- **Tidak lengkap**: tidak memiliki nilai atribut, tidak memiliki atribut tertentu yang menarik, atau hanya berisi data agregat
- misalnya, Pekerjaan=“ ” (data hilang)
- **Noisy**: mengandung noise, error, atau outlier
- misalnya, Gaji=“−10” (kesalahan)
- **Inconsistent**: mengandung ketidaksesuaian kode atau nama
- misalnya, Usia= “42”, Ulang Tahun= “03/07/2010”
- Tadinya diberi peringkat “1, 2, 3”, sekarang diberi peringkat “A, B, C”
- Perbedaan antara catatan duplikat
- Disengaja (misalnya, data hilang yang disamarkan)
- 1 Januari sebagai hari ulang tahun semua orang?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Incomplete (Missing) Data

- **Data tidak selalu tersedia**
- Misalnya, banyak tupel yang tidak memiliki nilai tercatat untuk beberapa atribut, seperti pendapatan pelanggan dalam data penjualan
- Data yang hilang mungkin disebabkan oleh
- kerusakan peralatan
- **tidak konsisten** dengan data tercatat lainnya dan karenanya dihapus
- **data tidak dimasukkan** karena kesalahpahaman
- **data tertentu mungkin tidak dianggap penting** pada saat masuk
- tidak mendaftarkan riwayat atau perubahan data
- **Data yang hilang** mungkin perlu disimpulkan

Contoh Missing Data

- Dataset: MissingDataSet.csv

ExampleSet (11 examples, 0 special attributes, 15 regular attributes) View Filter (11 / 11): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_I...	Hours_Per...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Ga...	Facebook	Twitter	Other_Soci...
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
3	F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
5	M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y	Y	N	Y	N	?
6	M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y	N	Y	N	N	?
7	M	African Amer	1981	D	11	2	Firefox	Google	Yahoo	?	Y	?	Y	Y	LinkedIn
8	M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
9	F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?	Y	Y	N	MySpace
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

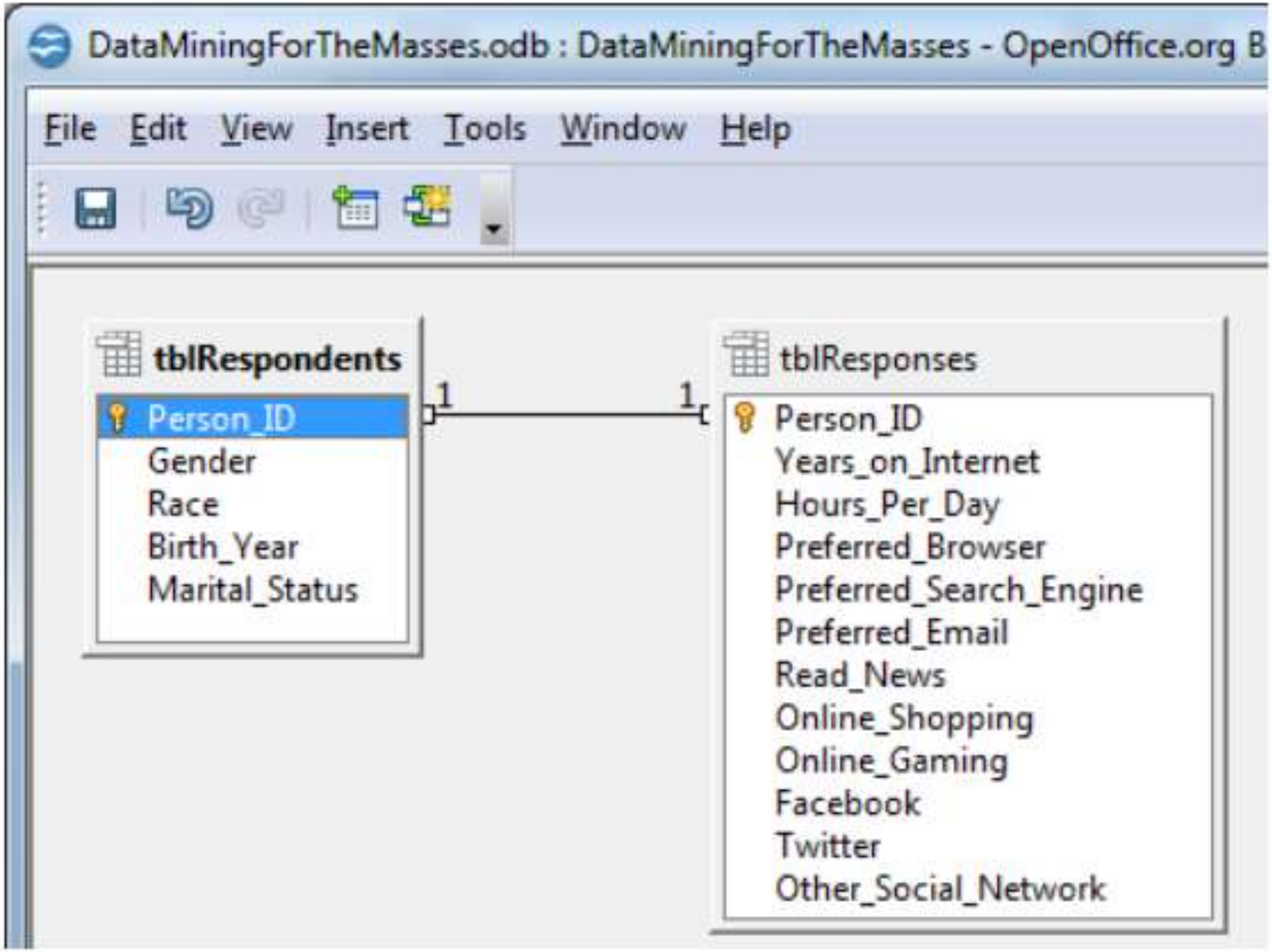
MissingDataSet.csv (1)

- Jerry adalah manajer pemasaran untuk sebuah perusahaan desain dan periklanan Internet kecil
- Bos Jerry memintanya untuk mengembangkan kumpulan data yang berisi informasi tentang pengguna Internet
- Perusahaan akan menggunakan data ini untuk menentukan jenis orang yang menggunakan Internet dan bagaimana perusahaan dapat memasarkan layanan mereka kepada kelompok pengguna ini.

MissingDataSet.csv (2)

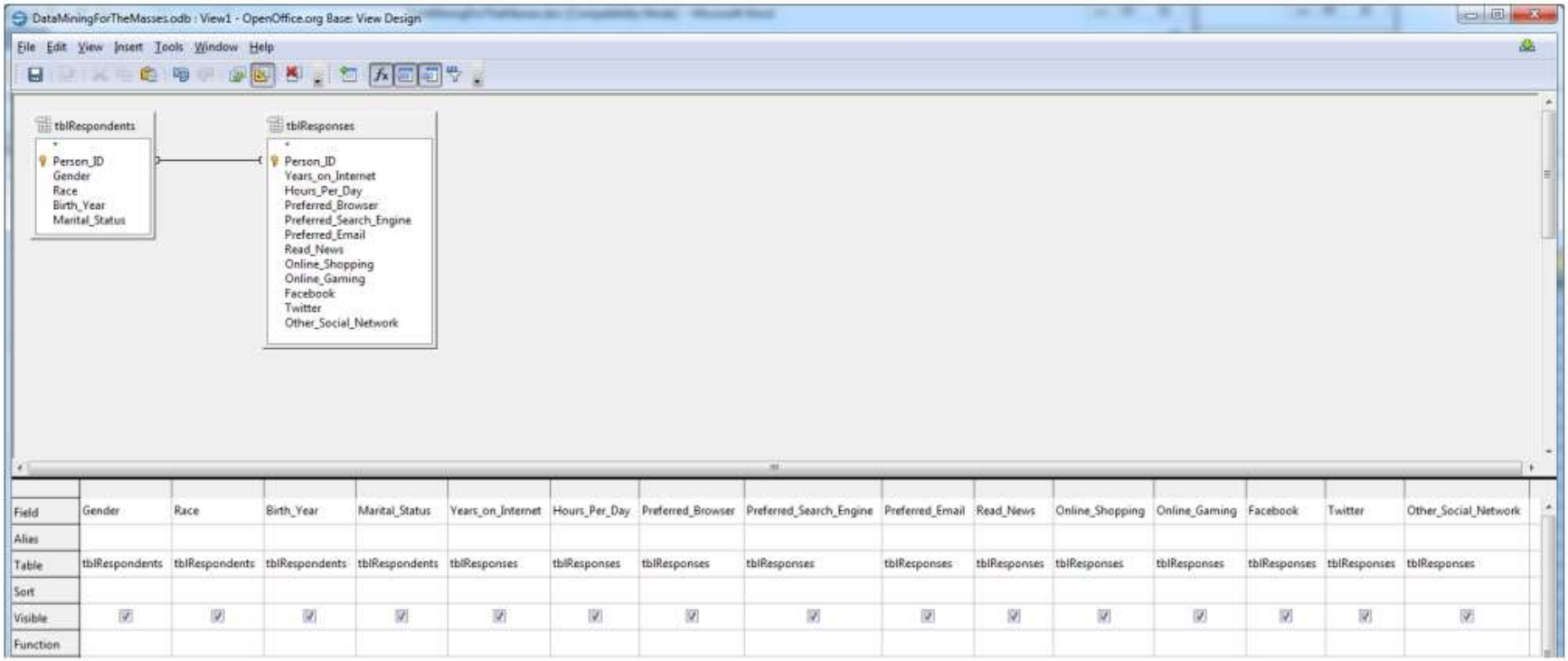
- Untuk menyelesaikan tugasnya, Jerry membuat survei online dan menempatkan link ke survei tersebut di beberapa situs Web populer
- Dalam waktu dua minggu, Jerry telah mengumpulkan cukup data untuk memulai analisis, namun ia mendapati bahwa datanya perlu didenormalisasi
- Dia juga mencatat bahwa beberapa pengamatan dalam kumpulan tersebut ada nilai yang hilang atau tampaknya mengandung nilai yang tidak valid
- Jerry menyadari bahwa beberapa pekerjaan tambahan pada data perlu dilakukan sebelum analisis dimulai.

Relational Data



I.14

View of Data (Denormalized Data)



Field	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
Alias															
Table	tblRespondents	tblRespondents	tblRespondents	tblRespondents	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses
Sort															
Visible	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Function															

vwInternetUser - DataMiningForTheMasses - OpenOffice.org Base: Table Data View

File Edit View Insert Tools Window Help

Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y	N	
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	
M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y		Y	Y	LinkedIn
M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	
M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y	N	MySpace
F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

Versi : 01

Contoh Missing Data

- Dataset: MissingDataSet.csv

ExampleSet (11 examples, 0 special attributes, 15 regular attributes) View Filter (11 / 11): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_I...	Hours_Per...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Ga...	Facebook	Twitter	Other_Soci...
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
3	F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
5	M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y	Y	N	Y	N	?
6	M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y	N	Y	N	N	?
7	M	African Amer	1981	D	11	2	Firefox	Google	Yahoo	?	Y	?	Y	Y	LinkedIn
8	M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
9	F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?	Y	Y	N	MySpace
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

How to Handle Missing Data?

- Ignore the tuple:
  - Usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually:
  - Tedious + infeasible?
- Fill in it automatically with
  - A global constant: e.g., "unknown", a new class?!
  - The attribute mean
  - The attribute mean for all samples belonging to the same class: smarter
  - The most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- **Abaikan tupel:**
  - Biasanya dilakukan ketika label kelas hilang (saat melakukan klasifikasi)—tidak efektif ketika % nilai yang hilang per atribut sangat bervariasi
- **Isi nilai yang hilang secara manual**:
  - Membosankan + tidak mungkin?
- **Isi secara otomatis dengan**
  - Konstanta global: misalnya, “tidak diketahui”, kelas baru?!
  - Arti atributnya
  - Arti atribut untuk semua sampel yang termasuk dalam kelas yang sama: lebih pintar
  - Nilai yang paling mungkin: berbasis inferensi seperti rumus Bayesian atau pohon keputusan

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 3 Data Preparation
  1. Handling Missing Data, pp. 34-48 (*replace*)
  2. Data Reduction, pp. 48-51 (*delete/filter*)

- Dataset: MissingDataSet.csv

- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut?

I.19

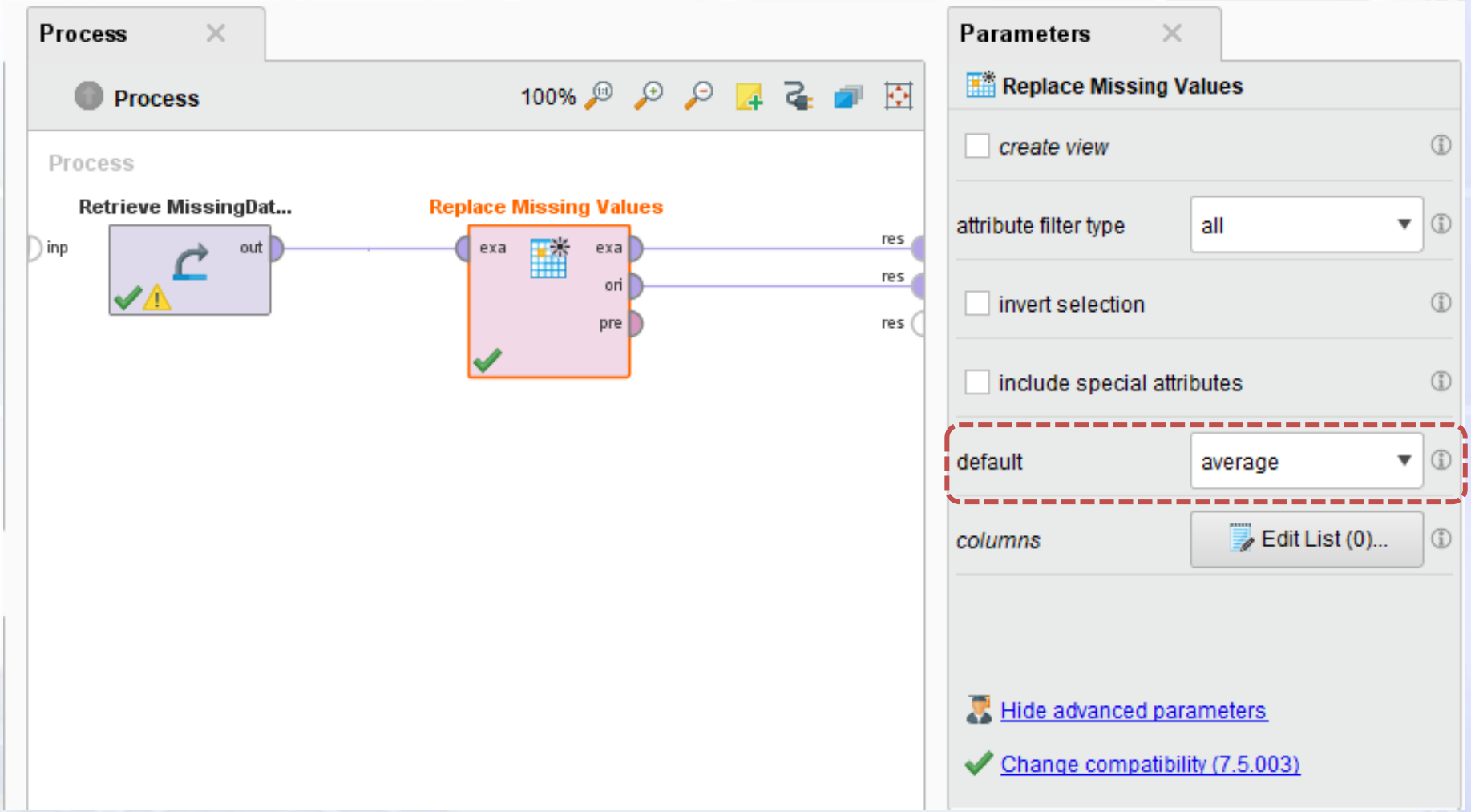
Missing Value Detection

Name	Type	Missing	Filter (15 / 15 attributes): Search for Attributes	
			Open chart	
Read_News	Polynominal	1	Least N (2)	Most Y (8)
Online_Shopping	Polynominal	2	Least N (4)	Most Y (5)
Online_Gaming	Polynominal	3	Least Y (2)	Most N (6)
Facebook	Polynominal	0	Least N (3)	Most Y (8)
Twitter	Polynominal	0	Least 99 (1)	Most N (8)
Other_Social_Network	Polynominal	7	Least MySpace (1)	Most LinkedIn (2)

versi : 01

I.20

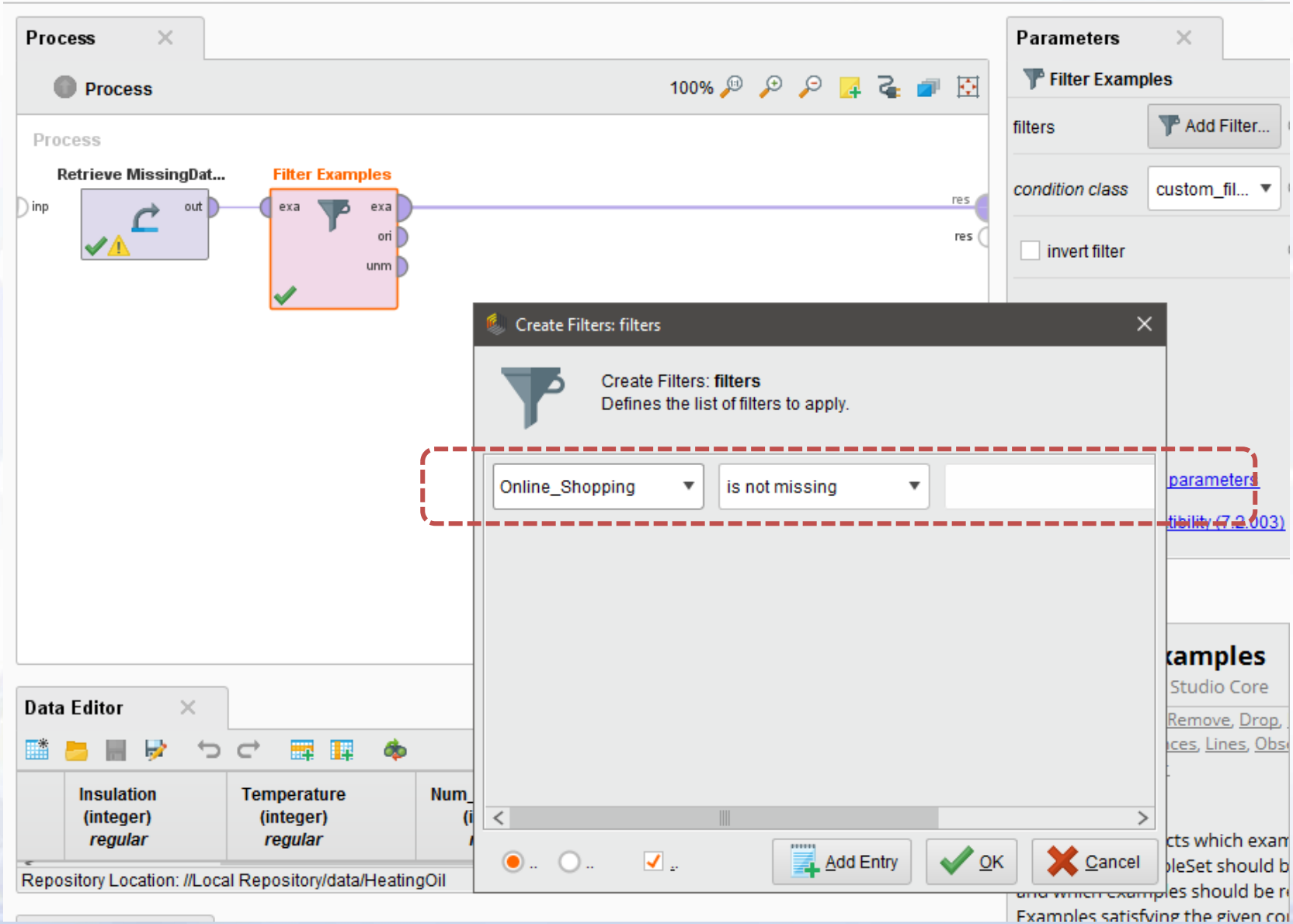
Missing Value Replace



Versi : 01

I.21

Missing Value Filtering



Versi : 01

Noisy Data

- **Noise**: kesalahan acak atau varians dalam variabel yang diukur
- **Nilai atribut yang salah** mungkin disebabkan oleh
  - Instrumen pengumpulan data yang salah
  - Masalah entri data
  - Masalah transmisi data
  - Keterbatasan teknologi
  - Inkonsistensi dalam konvensi penamaan
- **Masalah data lainnya yang memerlukan pembersihan data**
  - Catatan duplikat
  - Data tidak lengkap
  - Data yang tidak konsisten

How to Handle Noisy Data?

- **Binning**
  - Pertama, sortir data dan partisi ke dalam wadah (frekuensi yang sama).
  - Kemudian seseorang dapat menghaluskan dengan cara bin, menghaluskan dengan median bin, menghaluskan dengan batas bin, dll.
- **Regression**
  - Menghaluskan dengan memasukkan data ke dalam fungsi regresi
- **Clustering**
  - Deteksi dan hapus outlier
- **Gabungan inspeksi komputer dan manusia**
  - Mendeteksi nilai-nilai yang mencurigakan dan memeriksanya oleh manusia (misalnya, menangani kemungkinan outlier)

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

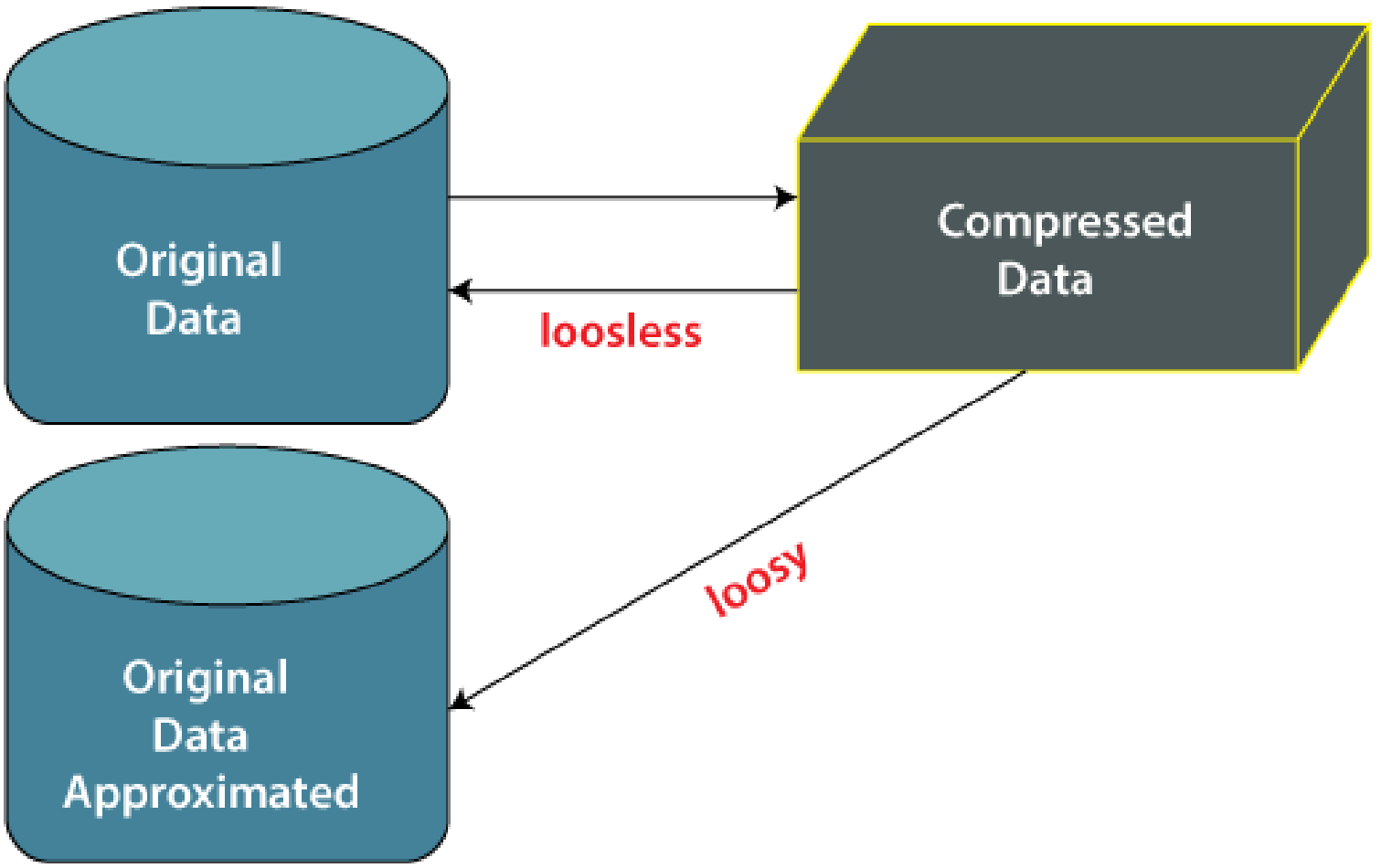
Data Cleaning as a Process (1)

- Deteksi discrepancy detection (perbedaan data)
  - Gunakan metadata (misalnya domain, rentang, ketergantungan, distribusi)
  - Periksa kelebihan muatan di lapangan
  - Periksa aturan keunikan, aturan berurutan, dan aturan nol
  - Gunakan alat komersial
    - Data scrubbing : menggunakan pengetahuan domain sederhana (misalnya, kode pos, pemeriksaan ejaan) untuk mendeteksi kesalahan dan melakukan koreksi
    - Data auditing : dengan menganalisis data untuk menemukan aturan dan hubungan untuk mendeteksi pelanggar (misalnya korelasi dan pengelompokan untuk menemukan outlier)

Data Cleaning as a Process (1)

- **Migrasi dan integrasi data**
  - Alat migrasi data: memungkinkan transformasi ditentukan
  - Alat ETL (Ekstraksi/Transformasi/Pemuatan): memungkinkan pengguna menentukan transformasi melalui antarmuka pengguna grafis
- **Dua proses Integrasi**
  - Iteratif dan interaktif (misalnya, Roda Potter)

2. DATA REDUCTION



Versi : 01

Data Reduction Methods

- Data Reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results
- Why Data Reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis take a very long time to run on the complete dataset

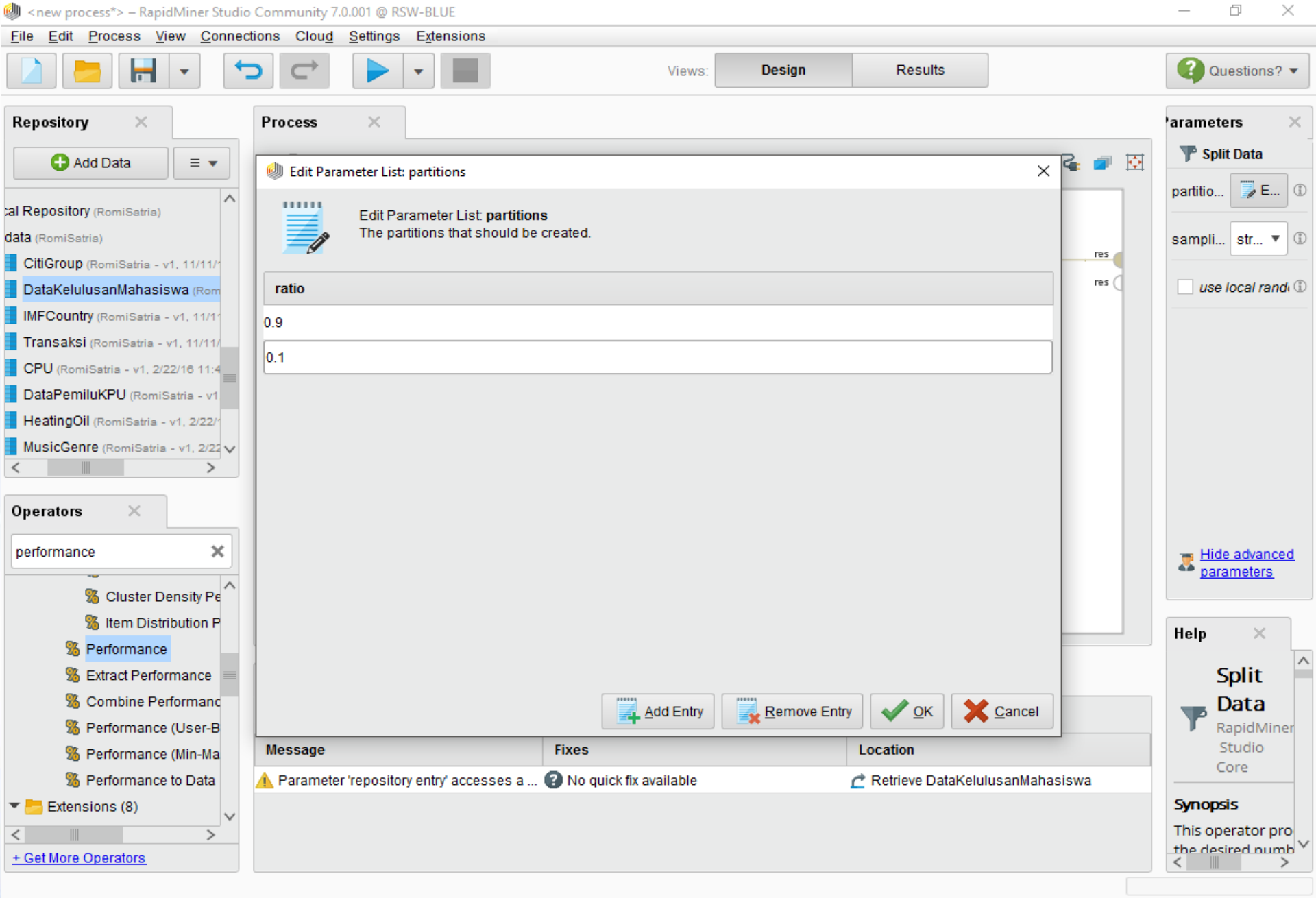
- Data Reduction Methods
  1. Dimensionality Reduction
     1. Feature Extraction
     2. Feature Selection
        1. Filter Approach
        2. Wrapper Approach
        3. Embedded Approach
  2. Numerosity Reduction (Data Reduction)
     - Regression and Log-Linear Models
     - Histograms, clustering, sampling

Data Reduction Methods(1)

- **Data Reduction**
  - Dapatkan representasi tereduksi dari kumpulan data yang volumenya jauh lebih kecil namun menghasilkan hasil analisis yang sama
- **Why Data Reduction?**
  - Basis data/gudang data dapat menyimpan data berukuran terabyte
  - Analisis data yang kompleks membutuhkan waktu yang sangat lama untuk dijalankan pada kumpulan data yang lengkap

Data Reduction Methods(1)

- Data Reduction Methods
  - A. Pengurangan Dimensi
    1. Feature Extraction
    2. Feature Selection
       - Filter Approach
       - Wrapper Approach
       - Embedded Approach
  - B. Numerosity Reduction (Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling


<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
Process
Parameters
Split Data
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.9
0.1
Add Entry
Remove Entry
OK
Cancel
Operators
performance
Cluster Density Pe
Item Distribution P
Performance
Extract Performance
Combine Performanc
Performance (User-B
Performance (Min-Ma
Performance to Data
Extensions (8)
+ Get More Operators
Hide advanced parameters
Help
Split Data
RapidMiner Studio Core
Synopsis
This operator pro
Message
Fixes
Location
Parameter 'repository entry' accesses a ...
No quick fix available
Retrieve DataKelulusanMahasiswa

Versi : 01

1. Dimensionality Reduction

- Curse of dimensionality
  - Ketika dimensi meningkat, data menjadi semakin jarang
  - Kepadatan dan jarak antar titik, yang penting untuk pengelompokan, analisis outlier menjadi kurang bermakna
  - Kemungkinan kombinasi subruang akan bertambah secara eksponensial
- Dimensionality reduction
  - Hindari dimensionality reduction
  - Membantu menghilangkan fitur yang tidak relevan dan mengurangi noise
  - Mengurangi waktu dan ruang yang dibutuhkan dalam data mining
  - Permudah visualisasi

Dimensionality Reduction Methods

1. Feature Extraction: Wavelet transforms, Principal Component Analysis (PCA)
2. Feature Selection: Filter, Wrapper, Embedded

Analisis Komponen Utama (Langkah-1)

- Diketahui N vektor data dari n dimensi, carilah $k \leq n$ vektor ortogonal (komponen utama) yang paling baik digunakan untuk merepresentasikan data

1. Normalisasikan data masukan: Setiap atribut berada dalam rentang yang sama
2. Hitung k vektor ortonormal (satuan), yaitu komponen utama
3. Setiap data masukan (vektor) merupakan kombinasi linier dari k vektor komponen utama

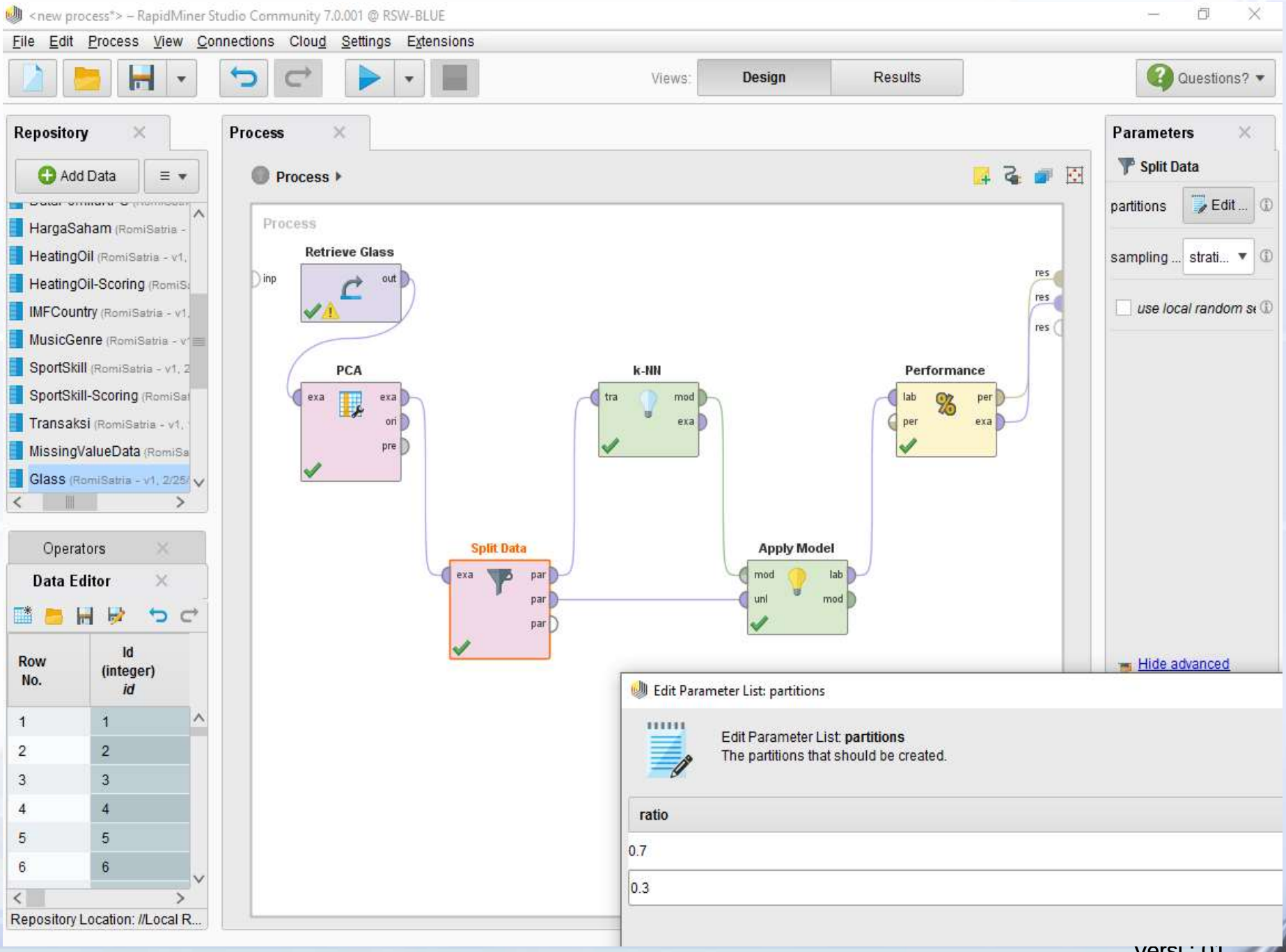
Analisis Komponen Utama (Langkah-2)

4. Komponen utama diurutkan berdasarkan "signifikansi" atau kekuatannya yang semakin menurun
5. Karena komponen-komponen telah diurutkan, ukuran data dapat diperkecil dengan menghilangkan komponen-komponen yang lemah, yaitu komponen-komponen yang variansinya rendah.

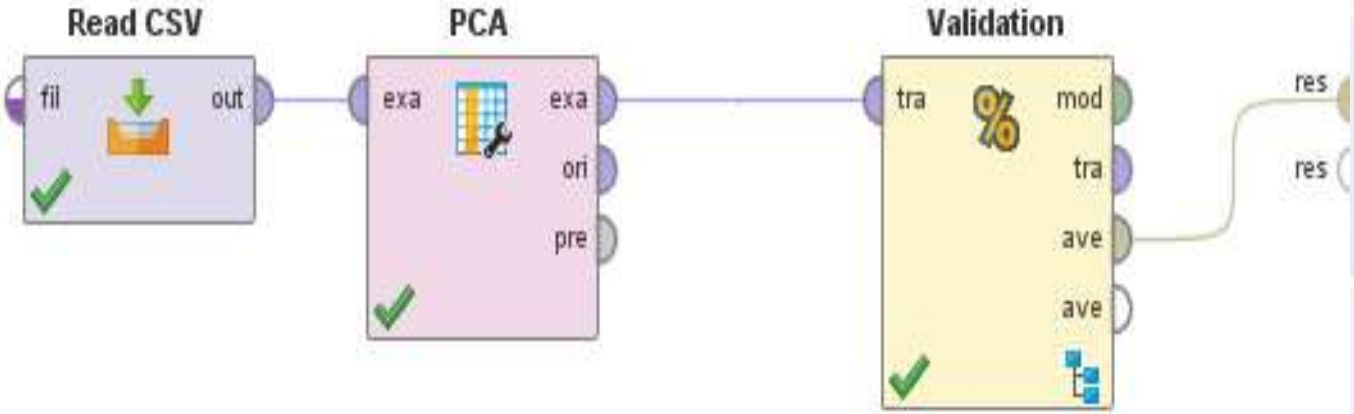
- Hanya berfungsi untuk data numerik

Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 4 (k-Nearest Neighbor Classification II) pp. 45-51
- Dataset: glass.data
- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!
- Bandingkan akurasi dari k-NN dan PCA+k-NN


<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
HargaSaham (RomiSatria -
HeatingOil (RomiSatria - v1,
HeatingOil-Scoring (RomiS
IMFCountry (RomiSatria - v1,
MusicGenre (RomiSatria - v
SportSkill (RomiSatria - v1, 2
SportSkill-Scoring (RomiSat
Transaksi (RomiSatria - v1,
MissingValueData (RomiSa
Glass (RomiSatria - v1, 2/25
Operators
Data Editor
Row No.
Id (integer) id
1 1
2 2
3 3
4 4
5 5
6 6
Repository Location: //Local R...
Process
Process
Process
Retrieve Glass
inp
out
PCA
exa
exa
ori
pre
k-NN
tra
mod
exa
Performance
lab
per
per
exa
res
res
res
Split Data
exa
par
par
par
Apply Model
mod
unl
lab
mod
Parameters
Split Data
partitions
Edit ...
sampling ...
strati...
use local random se
Hide advanced
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
0.3

I.38


Read CSV
fil
out
PCA
exa
exa
ori
pre
Validation
tra
mod
tra
ave
ave
res
res

Versi : 01

Data Awal Sebelum PCA

Result History | ExampleSet (Retrieve glass)

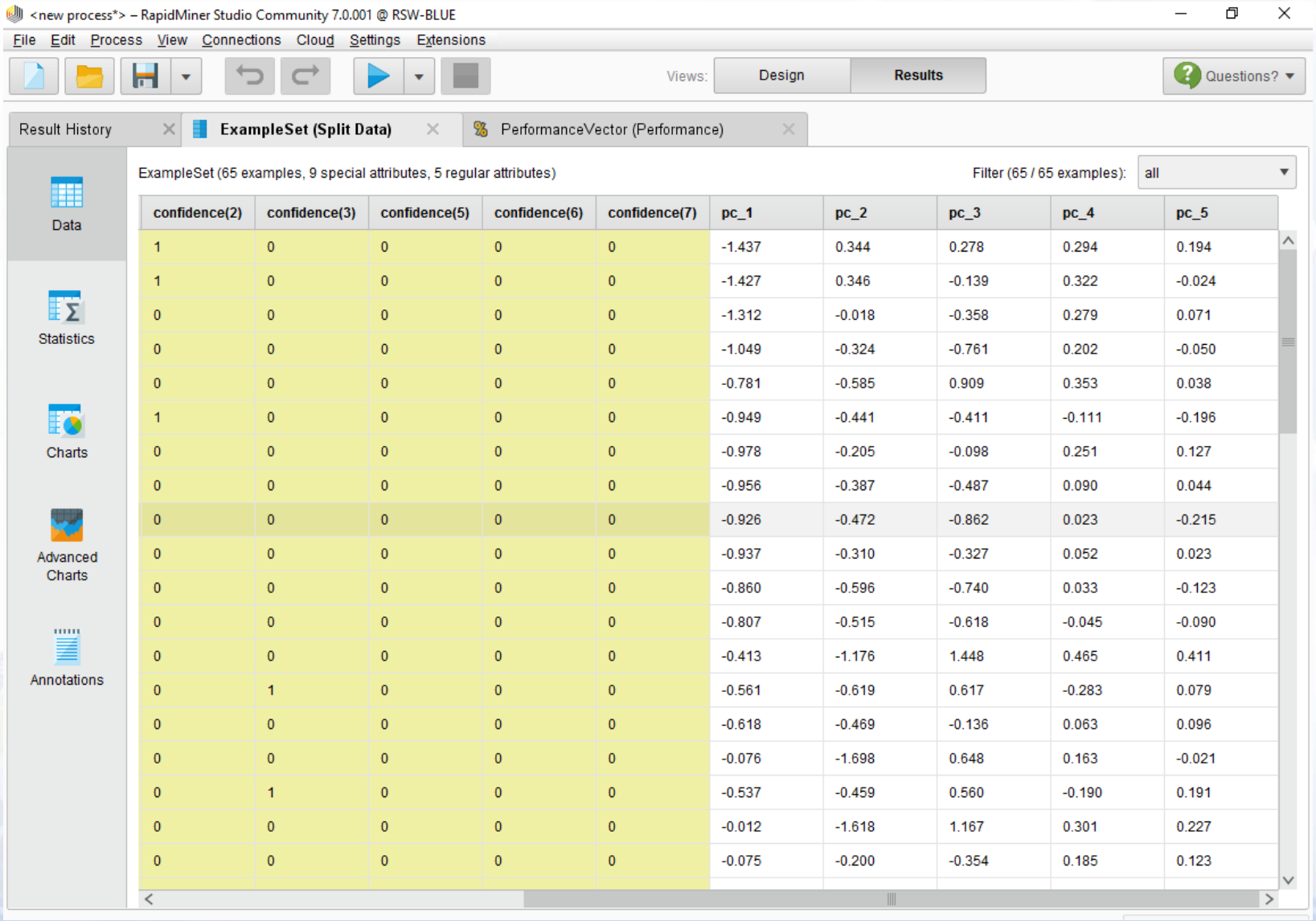
Data | Statistics | Charts | Advanced Charts | Annotations

ExampleSet (214 examples, 2 special attributes, 9 regular attributes) Filter (214 / 214 examples): all

Type	RI	Na	Mg	Al	Si	K	Ca	Ba
1	1.521	13.640	4.490	1.100	71.780	0.060	8.750	0
1	1.518	13.890	3.600	1.360	72.730	0.480	7.830	0
1	1.516	13.530	3.550	1.540	72.990	0.390	7.780	0
1	1.518	13.210	3.690	1.290	72.610	0.570	8.220	0
1	1.517	13.270	3.620	1.240	73.080	0.550	8.070	0
1	1.516	12.790	3.610	1.620	72.970	0.640	8.070	0
1	1.517	13.300	3.600	1.140	73.090	0.580	8.170	0
1	1.518	13.150	3.610	1.050	73.240	0.570	8.240	0
1	1.519	14.040	3.580	1.370	72.080	0.560	8.300	0
1	1.518	13	3.600	1.360	72.990	0.570	8.400	0
1	1.516	12.720	3.460	1.560	73.200	0.670	8.090	0
1	1.518	12.800	3.660	1.270	73.010	0.600	8.560	0
1	1.516	12.880	3.430	1.400	73.280	0.690	8.050	0
1	1.517	12.860	3.560	1.270	73.210	0.540	8.380	0
1	1.518	12.610	3.590	1.310	73.290	0.580	8.500	0
1	1.518	12.810	3.540	1.230	73.240	0.580	8.390	0
1	1.518	12.680	3.670	1.160	73.110	0.610	8.700	0
1	1.522	14.360	3.850	0.890	71.360	0.150	9.150	0
1	1.519	13.900	3.730	1.180	72.120	0.060	8.890	0

Versi : 01

Data Setelah PCA

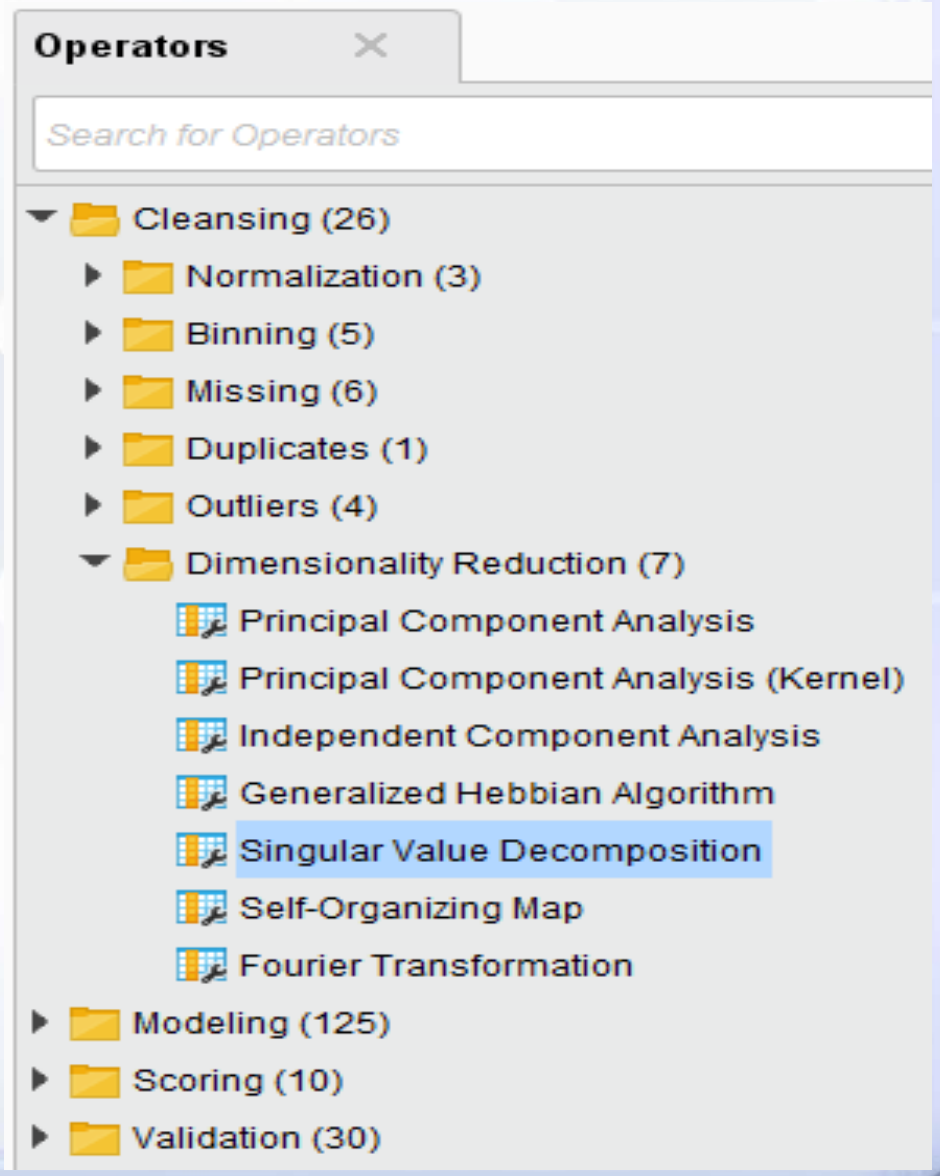


confidence(2)	confidence(3)	confidence(5)	confidence(6)	confidence(7)	pc_1	pc_2	pc_3	pc_4	pc_5
1	0	0	0	0	-1.437	0.344	0.278	0.294	0.194
1	0	0	0	0	-1.427	0.346	-0.139	0.322	-0.024
0	0	0	0	0	-1.312	-0.018	-0.358	0.279	0.071
0	0	0	0	0	-1.049	-0.324	-0.761	0.202	-0.050
0	0	0	0	0	-0.781	-0.585	0.909	0.353	0.038
1	0	0	0	0	-0.949	-0.441	-0.411	-0.111	-0.196
0	0	0	0	0	-0.978	-0.205	-0.098	0.251	0.127
0	0	0	0	0	-0.956	-0.387	-0.487	0.090	0.044
0	0	0	0	0	-0.926	-0.472	-0.862	0.023	-0.215
0	0	0	0	0	-0.937	-0.310	-0.327	0.052	0.023
0	0	0	0	0	-0.860	-0.596	-0.740	0.033	-0.123
0	0	0	0	0	-0.807	-0.515	-0.618	-0.045	-0.090
0	0	0	0	0	-0.413	-1.176	1.448	0.465	0.411
0	1	0	0	0	-0.561	-0.619	0.617	-0.283	0.079
0	0	0	0	0	-0.618	-0.469	-0.136	0.063	0.096
0	0	0	0	0	-0.076	-1.698	0.648	0.163	-0.021
0	1	0	0	0	-0.537	-0.459	0.560	-0.190	0.191
0	0	0	0	0	-0.012	-1.618	1.167	0.301	0.227
0	0	0	0	0	-0.075	-0.200	-0.354	0.185	0.123

Versi : 01

Latihan

- Review operator apa saja yang bisa digunakan untuk *feature extraction*
- Ganti PCA dengan metode *feature extraction* yang lain
- Lakukan komparasi dan tentukan mana metode feature extraction terbaik untuk data Glass.data, gunakan 10-fold cross validation



Feature/Attribute Selection

Cara lain untuk mengurangi dimensi data

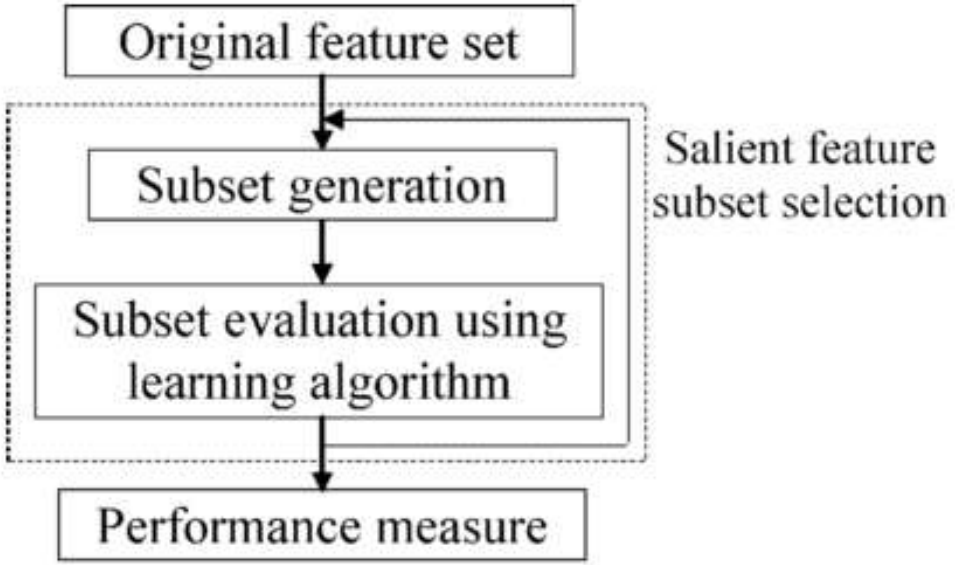
- Atribut yang berlebihan
  - Menduplikasi sebagian besar atau seluruh informasi yang terkandung dalam satu atau lebih atribut lainnya
  - Misalnya, harga pembelian suatu produk dan jumlah pajak penjualan yang dibayarkan
- Atribut yang tidak relevan
  - Tidak berisi informasi yang berguna untuk tugas penambangan data yang ada
  - Misalnya, ID siswa seringkali tidak relevan dengan tugas memprediksi IPK siswa

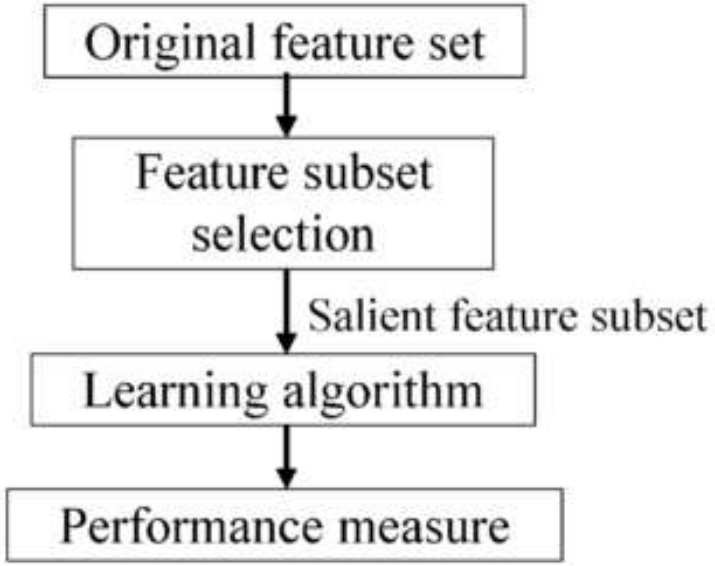
Feature Selection Approach

Sejumlah pendekatan yang diusulkan untuk pemilihan fitur secara umum dapat dikategorikan ke dalam tiga klasifikasi berikut: wrapper, filter, dan embedded (Liu & Tu, 2004)

1. Dalam filter approach, analisis statistik terhadap kumpulan fitur diperlukan, tanpa menggunakan model pembelajaran apa pun (Dash & Liu, 1997)
2. Dalam wrapper approach, model pembelajaran diasumsikan telah ditentukan, dimana fitur-fitur dipilih yang membenarkan kinerja pembelajaran model pembelajaran tertentu (Guyon & Elisseeff, 2003)
3. Embedded approach berupaya memanfaatkan kekuatan yang saling melengkapi dari pendekatan pembungkus dan filter (Huang, Cai, & Xu, 2007)

Wrapper Approach vs Filter Approach



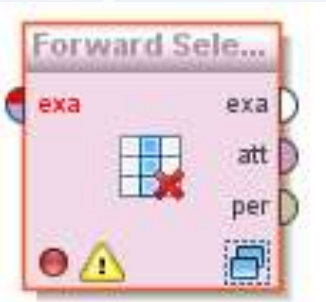


Wrapper Approach

Filter Approach

Feature Selection Approach

1. Filter Approach:
   - information gain
   - chi square
   - log likehood rasio
   - etc
2. Wrapper Approach:
   - forward selection
   - backward elimination
   - randomized hill climbing
   - etc
3. Embedded Approach:
   - decision tree
   - weighted naïve bayes
   - etc



2. Numerosity Reduction

Kurangi volume data dengan memilih bentuk representasi data alternatif yang lebih kecil

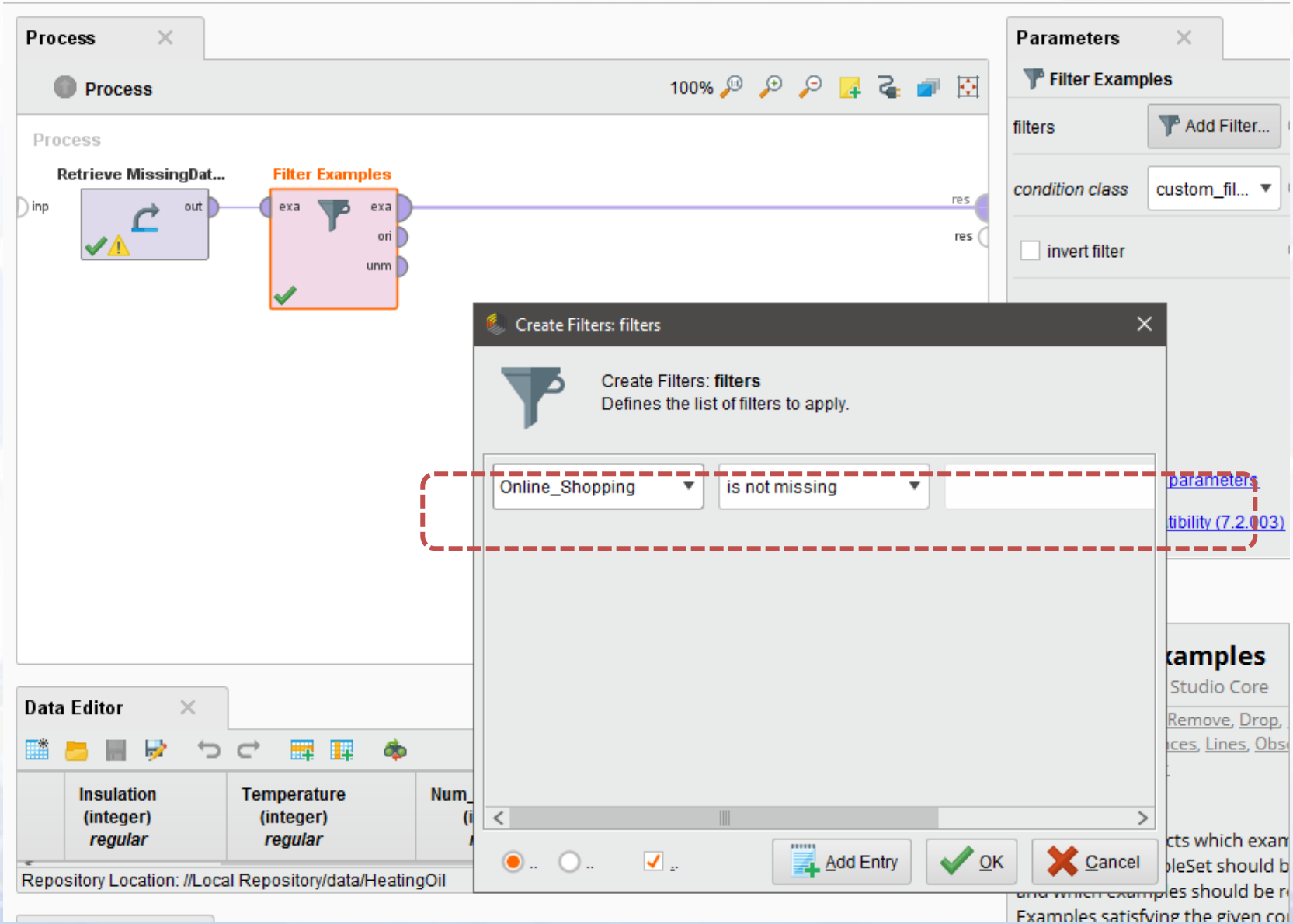
1. Metode parametrik (misalnya regresi)

- Asumsikan data cocok dengan beberapa model, perkirakan parameter model, simpan hanya parameternya, dan buang datanya (kecuali kemungkinan outlier)
- Contoh: Model log-linear—mendapatkan nilai pada suatu titik dalam ruang m-D sebagai hasil kali pada subruang marjinal yang sesuai

2. Metode non-parametrik

- Jangan berasumsi model
- Keluarga besar: histogram, pengelompokan, pengambilan sampel,…

Numerosity Reduction



Versi : 01

Parametric Data Reduction: Regression and Log-Linear Models

- **Regresi linier**
  - Data dimodelkan agar sesuai dengan garis lurus
  - Seringkali menggunakan metode kuadrat terkecil untuk menyesuaikan garis
- **Multiple regression**
  - Mengizinkan variabel respons Y dimodelkan sebagai fungsi linier vektor fitur multidimensi
- **Model log-linier**
  - Perkiraan distribusi probabilitas multidimensi diskrit

Review dan Latihan

☺ END ☺