

Data Preparation for Data Mining

Lesson 3

Lesson 3 Overview

- Basic Data Preparation
 - Introductory example
 - Obtaining the Data
 - Data Characterization
 - Data Assembly
 - Example



Introduction

- Just as manufacturing and refining are about transformation of raw materials into finished products, so too is the data to be used for data mining
- ECTL – extraction, clean, transform, load – is the common terminology used to describe preparing data for data mining
- The goal: ideal DM environment

What the Data Should Look Like

- All data mining algorithms want their input in tabular form – rows & columns as in a spreadsheet or database table

Reference	Time	Location	Count	Total Sales
1234	6/1/2000	San Diego	5	1700
1235	6/1/2000	San Diego	2	500
1236	6/1/2000	San Diego	3	1000
1234	6/2/2000	San Diego	5	1700
1235	6/2/2000	San Diego	2	500
1236	6/2/2000	San Francisco	3	1000
1240	6/2/2000	San Francisco	5	1500
1235	6/3/2000	San Francisco	2	500
1236	6/3/2000	San Francisco	3	1000
1240	6/3/2000	New York	10	3000
1235	6/3/2000	New York	2	500
1236	6/6/2000	New York	3	1000

What the Data Should Look Like

● Example: Customer Signature Data

- Next slide
- Continuous “snapshot” of customer behavior
- Each row represents the customer and whatever might be useful for data mining

This column is an id field where the value is different in every column. It gets ignored for data mining purposes.

This column is from the customer information file.

This column is the target, what we want to predict.

2610000101	010377	14		A	19.1		14 Spring ...	TRUE
2610000102	103188	7		A	19.1		NULL	TRUE
2610000105	041598	1		B	21.2		71 W. 19 St	FALSE
2610000171	040296	1		S	38.3		3562 Oak, .	FALSE
2610000182	061990	22		C	54.1		9672 W. 142	FALSE
2610000183	111192	45		C	56.1		NULL	TRUE
2620000107	080891	6		A	19.1		P.O. Box 11	FALSE
2620000108	120398	3		D	10.0		560 Robson	TRUE
2620000220	022797	2		S	38.3		222 E. 11th	FALSE
2620000221	021797	3		A	19.1		10122 SW 9	FALSE
2620000230	060899	1		S	38.3		NULL	TRUE
2620000231	062099	10		S	38.3		RR 1729	TRUE
2620000300	002894	7		B	21.2		1920 S. 14th	FALSE

These rows have invalid customer ids, so they are ignored.


This column is summarized from transaction data.

This column is a text field with unique values. It gets ignored (although it may be used for some derived variables).

What the Data Should Look Like

● The columns

- Contain data that describe aspects of the customer (e.g., sales \$ and quantity for each of product A, B, C)
- Contain the results of calculations referred to as *derived variables* (e.g., total sales \$)



Num	Unit Price	Quantity	Total Price
12345	10.99	50	549.50
24357	21.95	7	153.65
87921	39.95	25	998.75

What the Data Looks Like

1. Columns with One Value - Often not very useful
2. Columns with Almost Only One Value
3. Columns with Unique Values
4. Columns Correlated with Target Variable (synonyms with the target variable)

0
0
0
0
0

1.

9
0
9
9
9

2.

ABC Co.
XYZ Co.
MMM Co.
RRR Co.
TTT Co.

3.

What the Data Should Look Like

● Columns have important Model Roles in Data Mining:

- Input columns – input into the model
- Target column(s) – used only for predictive models – the values are created by the algorithm
- Ignored columns – not used in a particular data mining analysis

What the Data Should Look Like

● Variable Measures

- Categorical variables (e.g., CA, AZ, UT...)
- Ordered variables (e.g., course grades)
- Interval variables (e.g., temperatures)
- True numeric variables (e.g., money)

What the Data Looks Like

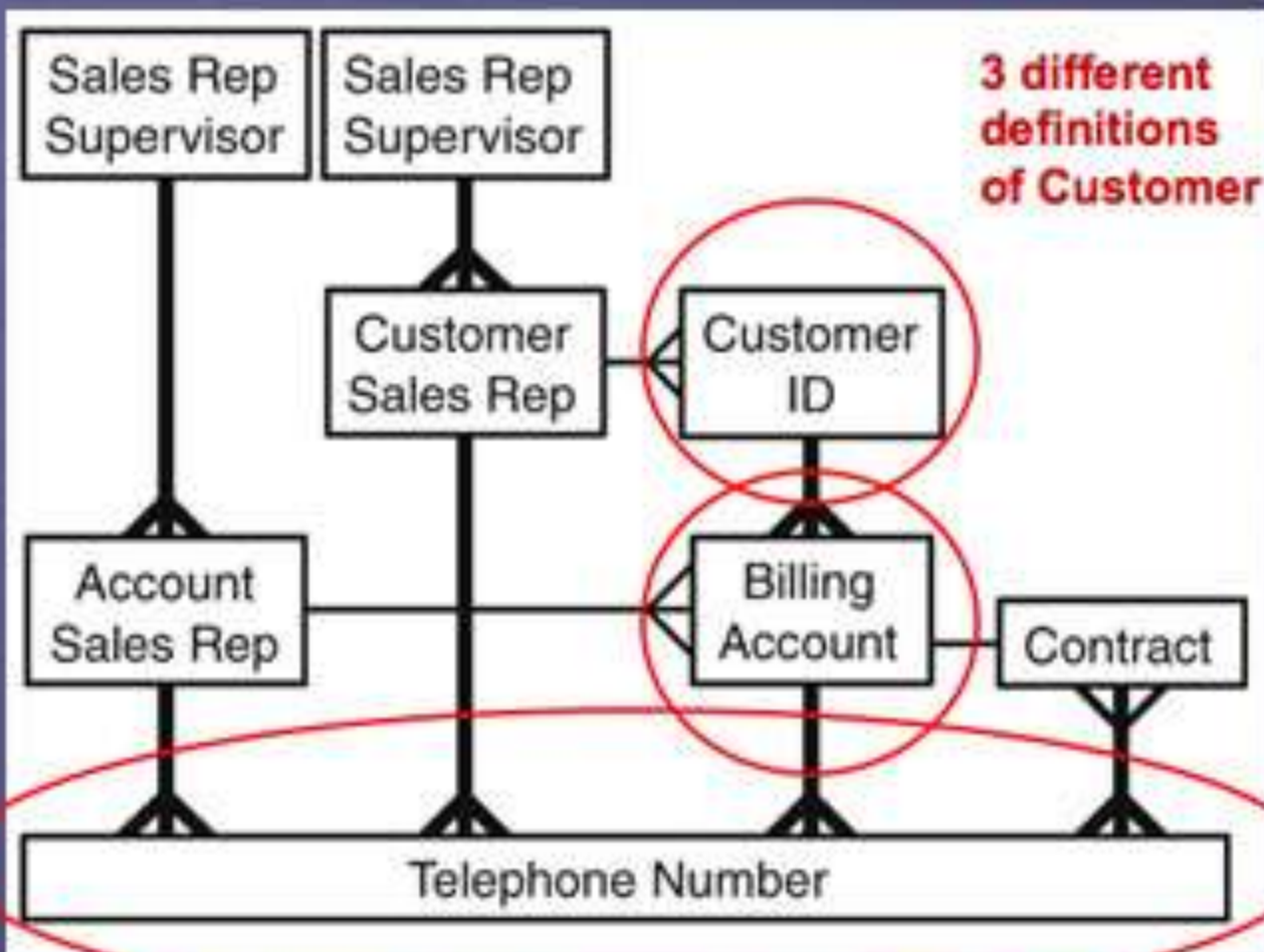
- Dates & Times
- Fixed-Length Character Strings (e.g., Zip Codes)
- IDs and Keys – used for linkage to other data in other tables
- Names (e.g., Company Names)
- Addresses
- Free Text (e.g., annotations, comments, memos, email)
- Binary Data (e.g., audio, images)

What the Data Should Look Like

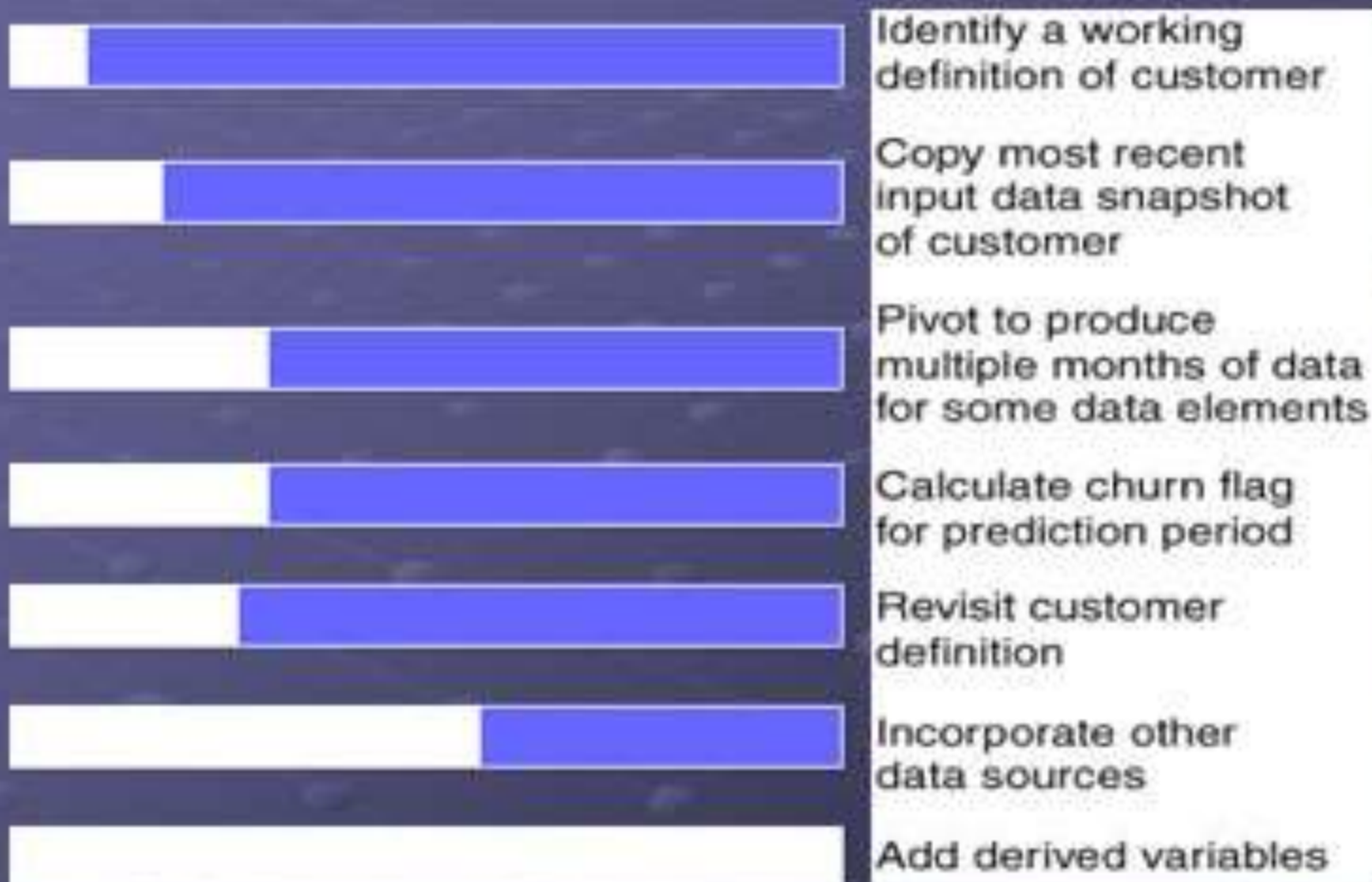
● Data Format Expectations for Data Mining

- All data in a single table (rows/columns)
- Each row corresponds to an entity (customer)
- Columns with single value should be ignored
- Columns with unique values should be ignored
- Target column identified for predictive DM

Typical Customer Model



Constructing the Customer Signature



The “Dark Side” of Data

- ◆ Missing values (nulls = empty or something else)
- ◆ Dirty data (erroneous zip codes, etc.)
- ◆ Inconsistent values (different revisions)
- ◆ Hidden data “traps”

Basic Preparation In Short

- Finding data for mining
- Creating and understanding of the quality of data
- Manipulating data to remove inaccuracies and redundancies
- Making a row-column (text) file of the data

Assessing Data - "Data assay"

• Data Discovery

- Discovering and locating data to be used.
Coping with the bureaucrats and data hiders.

• Data Characterization

- What is it, the data found? Does it contain the information needed or is it mere garbage?
- Domain experts

• Data Set Assembly

- Making a "flat file" – a (ascii) table combining the data coming from different sources

Outcome of data assay

Discover
Data

Study
Data

Assemble

- Detailed knowledge in the form of a report detailing:
 - quality, problems, shortcomings, and suitability of the data for mining.
- Tacit knowledge of the database
 - The miner has a perception of data suitability

Discover
Data

Study
Data

Assemble

Data Discovery

Discover
Data

Study
Data

Assemble

- Input to data mining is a row-column text file
 - Original source of the data may be in the form of various databases, warehouses, flat measurement data files, Excel files, binary data files etc.

Data Discovery

Discover
Data

Study
Data

Assemble

● Data Access Issues

- Overcome accessibility challenges, like legal issues, cross-departmental access limitations, company politics
- Overcome technical challenges, like data format incompatibilities, data storage mediums, database architecture incompatibilities, measurement concurrency issues
- Internal/external source and the cost of data

Data characterization

Discover
Data

Study
Data

Assemble

- Characterize the nature of data sources
 - Study the nature of variables and usefulness for modeling
 - Look into frequency distributions and cross-tabs
 - Find useless columns
 - Avoiding GIGO

Characterization: Granularity

Discover
Data

Study
Data

Assemble

- Variables fall within a spectrum of very detailed to very aggregated
- The level of aggregation is a continuum
- General rule: detailed is preferred over aggregated for mining

Characterization: Granularity

Discover
Data

Study
Data

Assemble

- Level of aggregation determines the accuracy of the model.
 - One level of aggregation less in input compared to requirement of output
 - Model outputs weekly variance, use daily measurements for modeling.

Characterization: Consistency

Discover
Data

Study
Data

Assemble

- Undiscovered inconsistency in data stream leads to Garbage Out model.
 - If car model is stored as M-B, Mercedes & M-Benz, it is impossible to detect cross relations between person characteristics and the model of car owned.
- Labeling of variables is dependent on the system producing variable data.
 - Employee means different thing for HR department and to Payroll system in the presence of contractors
 - So, how many employees do we have?

Characterization: Pollution

Discover
Data

Study
Data

Assemble

- Data is polluted if variable label does not reveal the meaning of variable data
- Typical sources of pollution
 - Misuse of record field
 - B to signify "Business" in gender field of credit card holders -> How do you do statistical analysis based on gender then ?
 - Data transfer unsuccessful (misinterpreted fields while copying (commas in addresses in .csv?))
 - Human resistance (Car sales example)

Characterization: Objects

Discover
Data

Study
Data

Assemble

- The precise nature of object measured needs to be known
 - Consumer spending vs. consumer buying patterns (total \$ vs. product types)
- Data miner needs to understand why information was captured in the first place
 - Perspective may color data

Characterization: Relationships

Discover
Data

Study
Data

Assemble

- Data mining needs a row-column text file for input - This file is created from multiple data streams
- Data streams may be difficult to merge
 - There must be some sort of a key that is common to each stream
 - Example: different customer ID values in different databases.
 - Key may be inconsistent, polluted or difficult to get access; there may be duplicates etc.

Characterization: Domain

Discover
Data

Study
Data

Assemble

- Variable values must be within permissible range of values
- Summary statistics and frequency counts will reveal out-of-bounds values.
- Conditional domains:
 - Medical data (diagnosis bound to gender)
 - Business rules, like fraud investigation for claims of > \$1k
- Automated tools to find unknown business rules (e.g. WizRule)

Characterization: Defaults

Discover
Data

Study
Data

Assemble

- Default values in data may cause problems
- Conditional defaults dependent on other entries may create fake patterns
 - but really it is the question of lack of data
 - may be useful patterns, but often of limited use

Characterization: Integrity


Discover
Data

Study
Data

Assemble

- Checking the possible/permitted relationships between variables
 - Many cars perhaps, but one spouse
- Acceptable range
 - Outlier may actually be the data we are looking for
 - Fraud looks often like outlying data because majority of claims are not fraudulent.

Characterization: Concurrency



Discover
Data

Study
Data

Assemble

- Data capture may be of different epochs
 - Thus streams may not be comparable at all
 - Example: Last years tax report and current income/possessions may not match

Characterization: Duplicates/ Redundancies



Discover
Data

Study
Data

Assemble

- Different data streams may involve redundant data - even one source may have redundancies
 - like DOB and age, or
 - price_per_unit - number_purchased - total_price
- Removing redundancies may increase modeling speed
- Some algorithms may crash if two variables are identical
 - Tip: if two variables are almost collinear use difference

Data Set Assembly



- Data is assembled from different data streams to row-column text file
- Then data assessment continues from this file

Data Set Assembly: Reverse Pivoting

Discover
Data

Study
Data

Assemble

- Feature extraction by sorting data by one key from transactions and deriving new fields
- Example on the next slide: from transaction data to customer profile

Date	Acc #	Balance	Branch	Product
1-1-00	1	500	5	3
1-1-00	2	700	3	3
2-1-00	1	700	3	2
2-1-00	3	1000	4	3
2-1-00	4	1500	3	2
3-1-00	2	1000	4	3
3-1-00	1	400	3	2
4-1-00	2	1500	5	3
4-1-00	3	1600	3	2

Acc #						ATM P1	ATM P2	ATM P3	ATM P4					
1						\$1400	\$1000	\$1500	\$500					
2						\$200	\$500	\$700	\$200					
3						\$2500	\$1200	\$3500	\$1000					
4						\$100	\$100	\$300	\$200					

Data Set Assembly: Feature Extraction

Discover
Data

Study
Data

Assemble

- Choice of variables to extract determines how data is presented to data mining tool
 - Miner must judge which features are predictive
 - Choice cannot be automated but actual extraction of features can.

Data Set Assembly: Feature Extraction

Discover
Data

Study
Data

Assemble

- Reverse pivot is not the only way extract features
 - Source variables may be replaced by derived variables
 - Physical models: flat most of time - take only sequences where there are rapid changes

Data Set Assembly:

Explanatory Structure

Discover
Data

Study
Data

Assemble

- Data miner needs to have an idea how data set can address problem area
 - It is called the explanatory structure of data set
 - Explains how variables are expected to relate to each other
 - How data set relates to solving the problem

Data Set Assembly:

Explanatory Structure

Discover
Data

Study
Data

Assemble

- **Sanity check: Last phase of data assay**
 - Checking that explanatory structure actually holds as expected
 - Many tools like OLAP

Data Set Assembly: Enhancement/ Enrichment



- Assembled data set may not be sufficient
- Data set enrichment: adding external data
- Data enhancement
 - expanding data set without external data
 - feature extraction
 - adding bias (e.g. remove non-responders from data)
 - data multiplication
 - Generate rare events (add some noise)

Data Set Assembly: Sampling Bias

Discover
Data

Study
Data

Assemble

- Undetected sampling bias may ruin the model
 - US Census: cannot find poorest segment of the society - no home, no address
 - Telephone polls: have to own a telephone, have to be willing to share opinions over phone lines
- At this phase - the end of data assay the miner needs to realize existence of possible bias and explain it

What does a data assay look like in practice?

- The purpose of data assay: check that the data is coherent, sufficient, can be assembled into a needed format, and makes sense within the proposed framework
- Real life example: CREDIT data for FNBA bank

FNBA

- Data comes in the form of credit histories purchased from credit bureaus
- During solicitation campaign, FNBA contact the targeted market by phone/mail
- Prospective cc users either apply or do not respond
- Input stream is a flag: responded or not
- Predictive Model: profile of people who are most likely to respond

Example 1: CREDIT

- Study of data source report
 - to find out integrity of variables
 - to find out expected relationships between variables for integrity assessment
- Tools for single variable integrity study
- Tools for cross correlation analysis

Example 1: CREDIT

- Study of data source report
- Tools for single variable integrity study
 - Status report for Credit file
 - Complete Content Report
 - Leads to removing some variables
- Tools for cross correlation analysis

Example 1: CREDIT

- Study of data source report
- Tools for single variable integrity study
- Tools for cross correlation analysis
 - KnowledgeSeeker - chi-square analysis
 - Weka's attribute selection tools
 - Checking that expected relationships are there

CREDIT data

- Campaign results: various data streams are assembled into a table
- Look under Class Resources: CREDIT.DAT
- 41 fields
- 13996 rows

Inputs

- BEACON, DAS, CRITERIA, ROPEN, RBALNO, LST_R_OPEN, RBAL, RLIMIT, TOPEN, TBALNO, MOF, RBAL_LIMIT, EQLIMIT, EQBAL, EQHIGHBAL, EQCURBAL, BCLIMIT, BCBAL, IHIGHBAL, ICURBAL, UNSECLIMIT, UNSECBAL, MTHIGHBAL, MTCURBAL, BCOPEN, YEARS_RES, CHILDREN, EST_INC, OWN_HOME, HOME_VALUE, HOME_INC, HOME_ED, PRCNT_WHIT, PRCNT_PROF, DOB_MONTH, DOB_YEAR, AGE_INFERR, SEX, MARRIED, BUYER

How to assay this data?

- Start looking at the basic statistics for the file
- Measurements in a statistics file one would use are:
 - MAX, MIN, DISTINCT, EMPTY
 - CONF, REQ, VAR
 - LIN
 - VAR TYPE

FIELD	MAX	MIN	DISTINCT	EMPTY	CONF	REQ	VAR	LIN	VAR TYPE
AGE_INFERR	57	35	3	0	0.96	280	0.8	0.9	N
BCOPEN	0.0	0.0	1	59	0.95	59	0.0	0.0	E
BEACON_C	804.0	670.0	124	0	0.95	545	1.6	1.0	N
CRITERIA	1.0	1.0	1	0	0.95	60	0.0	0.0	N
EQBAL	67950	0.0	80	73	0.95	75	0.0	1.0	E
DOB_MONTH	12.0	0.0	14	8912	0.95	9697	0.3	0.6	N
HOME_VALUE	531.0	0.0	191	0	0.95	870	2.6	0.9	N
HOME_ED	160	0.0	8	0	0.95	853	3.5	0.7	N
PRCNT_PROF	86	0	66	0	0.95	579	0.8	1.0	N

- BEACON_C lin>0.98
- CRITERIA: constant
- EQBAL: empty, distinct values?

Example 1: CREDIT: Single-variable status

FIELD	CONTENT	CCOUNT
DOB_MONTH		8912
DOB_MONTH	00	646
DOB_MONTH	01	12
DOB_MONTH	02	7
DOB_MONTH	03	10
DOB_MONTH	04	9
DOB_MONTH	05	15
DOB_MONTH	06	14
DOB_MONTH	07	11
DOB_MONTH	08	10
DOB_MONTH	09	13
DOB_MONTH	10	10
DOB_MONTH	11	15
DOB_MONTH	12	13

Conclusions

- DOB month:
sparse, 14 values?

Example 1: CREDIT: Single-variable status

FIELD	CONTENT	CCOUNT
HOME_VALUE	000	284
HOME_VALUE	027	3
HOME_VALUE	028	3
HOME_VALUE	029	3
HOME_VALUE	030	3
HOME_VALUE	031	2
HOME_VALUE	032	5

Conclusions

- HOME VALUE:min 0.0? Rent/own?

Statistics File Summary

- Each field, each variable raises various Qs
- Is the range of values reasonable?
- Is the distribution reasonable?
- Should the variable be kept or removed?
- Just the basic report of frequencies can point to a number of questions

Next Step

- Need to consider the relationships between the variables
- Tools:
 - Chi-square analysis: compares the variable of interest with only one other variable at a time
 - Weka, KnowledgeSeeker

Example 1: Relationships

- The most highly interactive variable for AGE_INFERR? As expected, it correlates with DOB_YEAR
 - Right, it does - data seems ok
 - Do we need both? Remove other? Depends.
- HOME_ED correlates with PRCNT_PROF
 - Right, it does - data seems ok

Example 1: Relationships

- Talk about bias

- Introducing bias for e.g. increase number of child-bearing families to study marketing of child-related products.

Data Assay

- Assessment of quality of data for mining
- Leads to assembly of data sources to one file.
- How to get data and does it suit the purpose

Data Assay

- Main goal: miner understands where the data come from, what is there, and what remains to be done.
 - It is helpful to make a report on the state of data
- It involves miner directly - rather than using automated tools
 - After assay rest can be carried out with tools

Next

- Sampling, variability and confidence
- Dealing with specific types of variables