CSCI 5408 Data Analytics: DM and DW Tech
- Learning actions for Week 8-9

- Ass4 Due: Mar 14
  - Brightspace: Assignment 4, Tutorial slides, etc.
  - Help hours: Wed/Fri, 1:00-2:30PM, CS 134/CS 233
- Write answers for review questions
  - Final Exam: Apr 20, 3:30-5:30 PM
- Read:
  - Lectures: 11-12
  - Text: Ch1, 2, 4 for 3rd edition, or Ch1-3 for 2nd edition

Ch2. Data Preprocessing (DP)

(Textbook: Ch3 of 3rd, or Ch2 of 2nd)

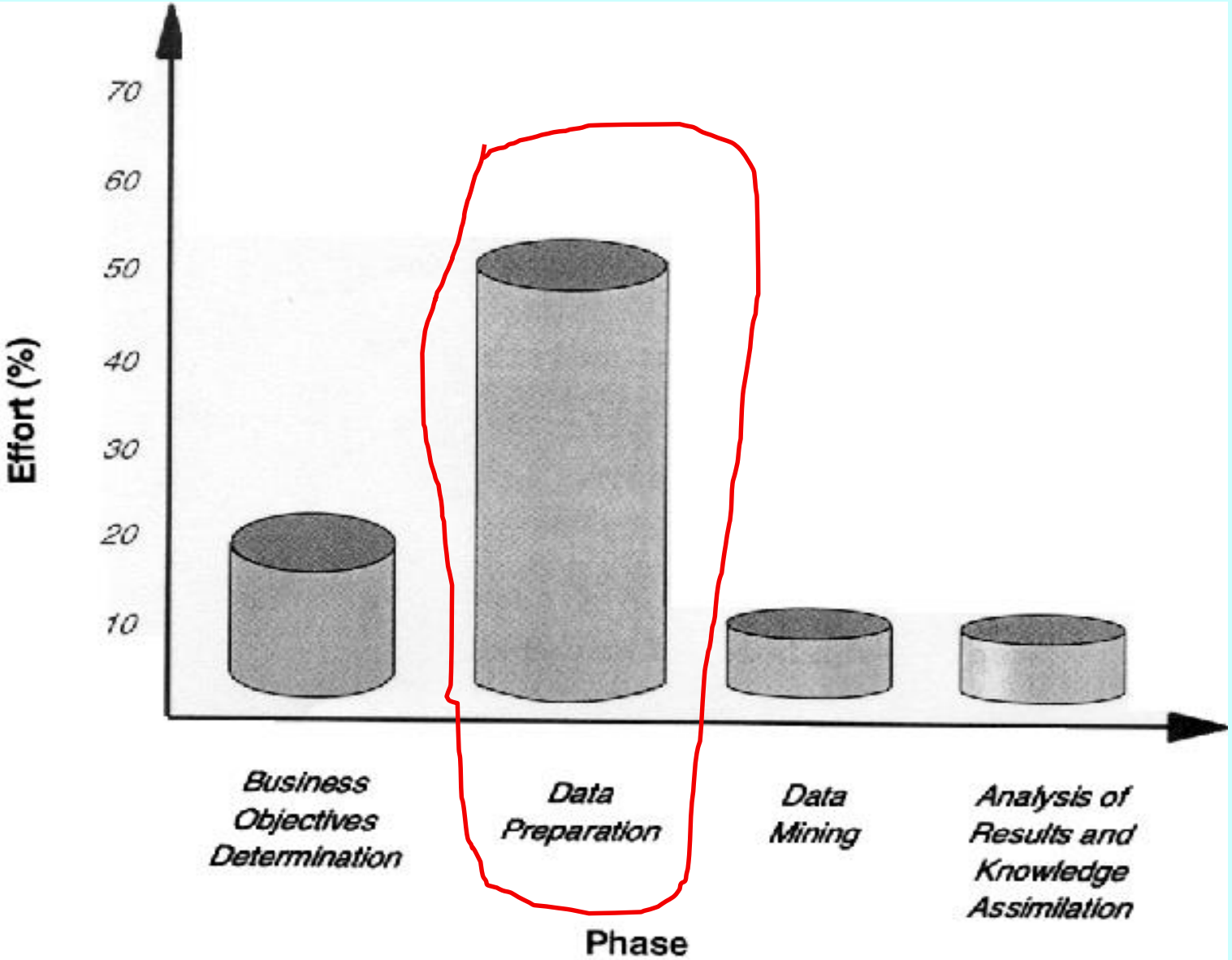
- Why DP is important for DM & DW
- Typical tasks of DP
- A case study on DP
- DP examples (past research projects)

Why Preprocess Data?

- No quality data, no quality DW/DM results!
  - Quality decisions must be based on quality analytical information and knowledge which can only be derived from quality data.
- What properties should quality data have?
  - Enriched, i.e. Integrated from different sources
  - Relevant
  - Clean
  - Consistent
  - In right format and type
- How to get quality data for DM/DW?

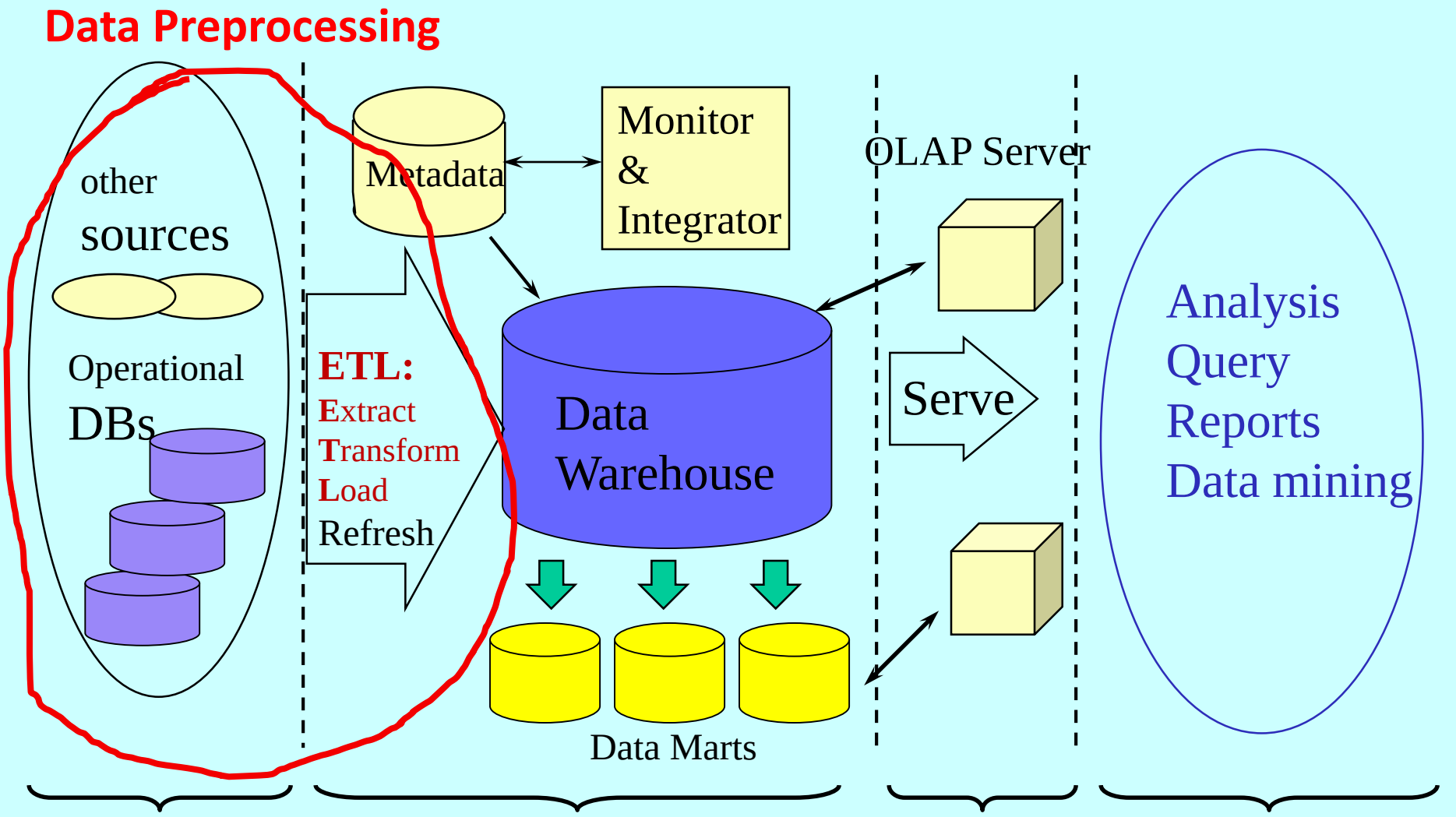
DP based on ETL Notion

- In DP, **Extract, Transform, Load** (**ETL**) refers to a general notion of DP to identify/prepare/move data from sources to a target data form. The ETL process became a popular concept since 1970s
  - **Data extraction** is where data is extracted from homogeneous or heterogeneous data sources
  - **Data transformation** where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis;
  - **Data loading** where the data is loaded into the final target database, more specifically, an operational data store, or data warehouse.

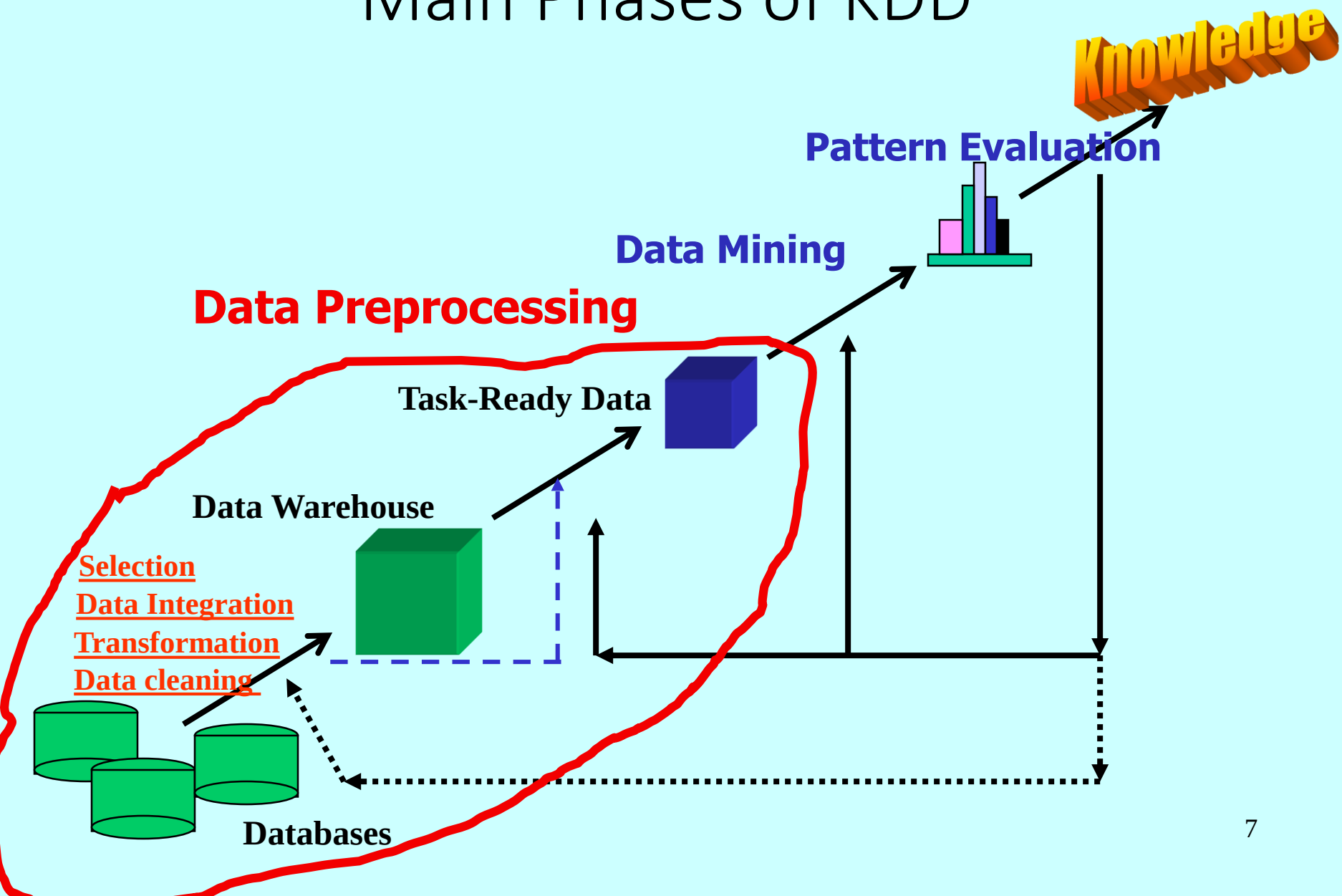


Effort Required for Each Data Mining Process Step

Multi-Tiered Architecture of OLAP System, e.g. SQL Server 2014

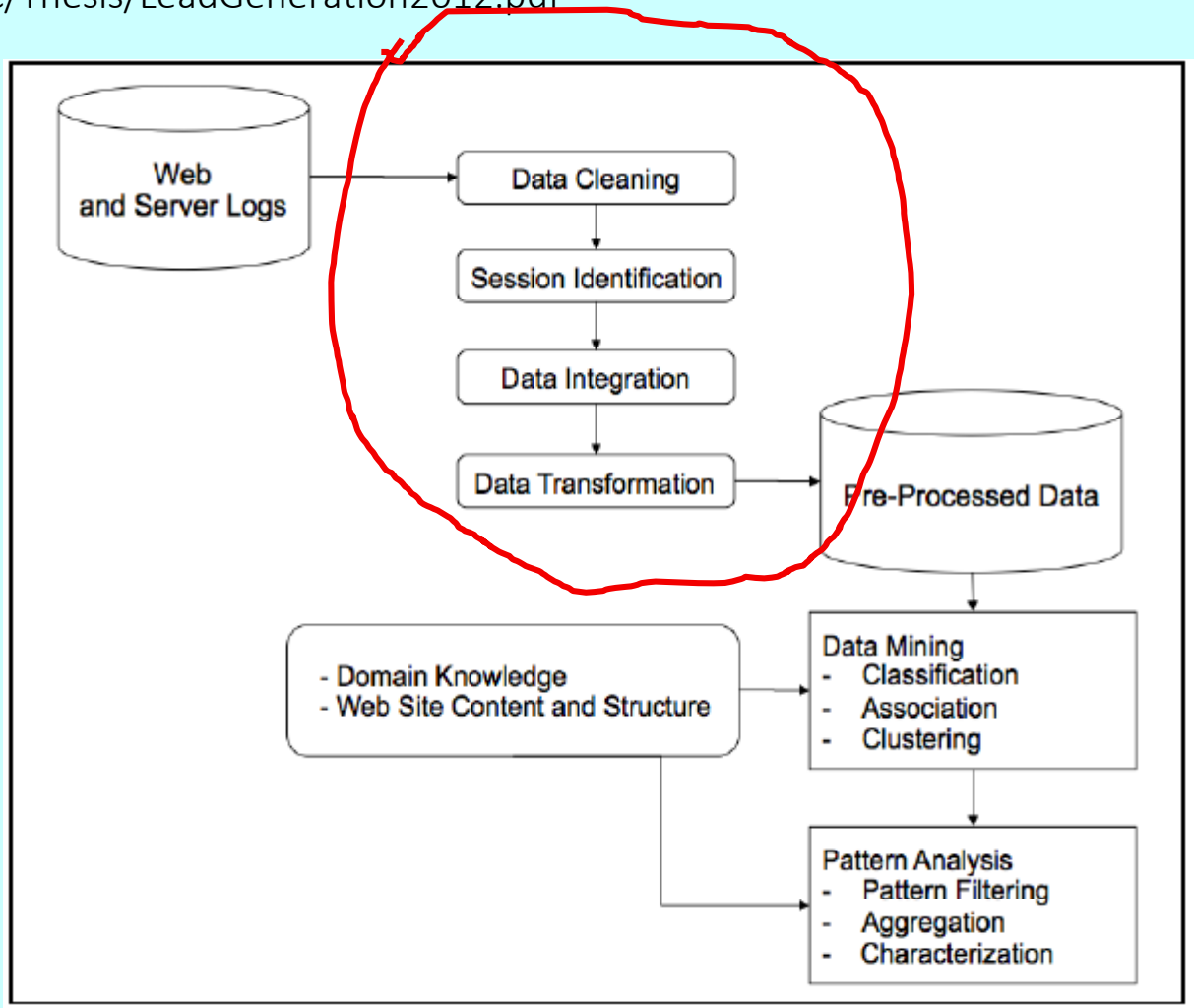


Main Phases of KDD



E.g.1, "Customer Predictive-leads Pattern Discovery by Mining Integrated Web-click Stream and Pageview Content Data" (NSERC Engage Project)

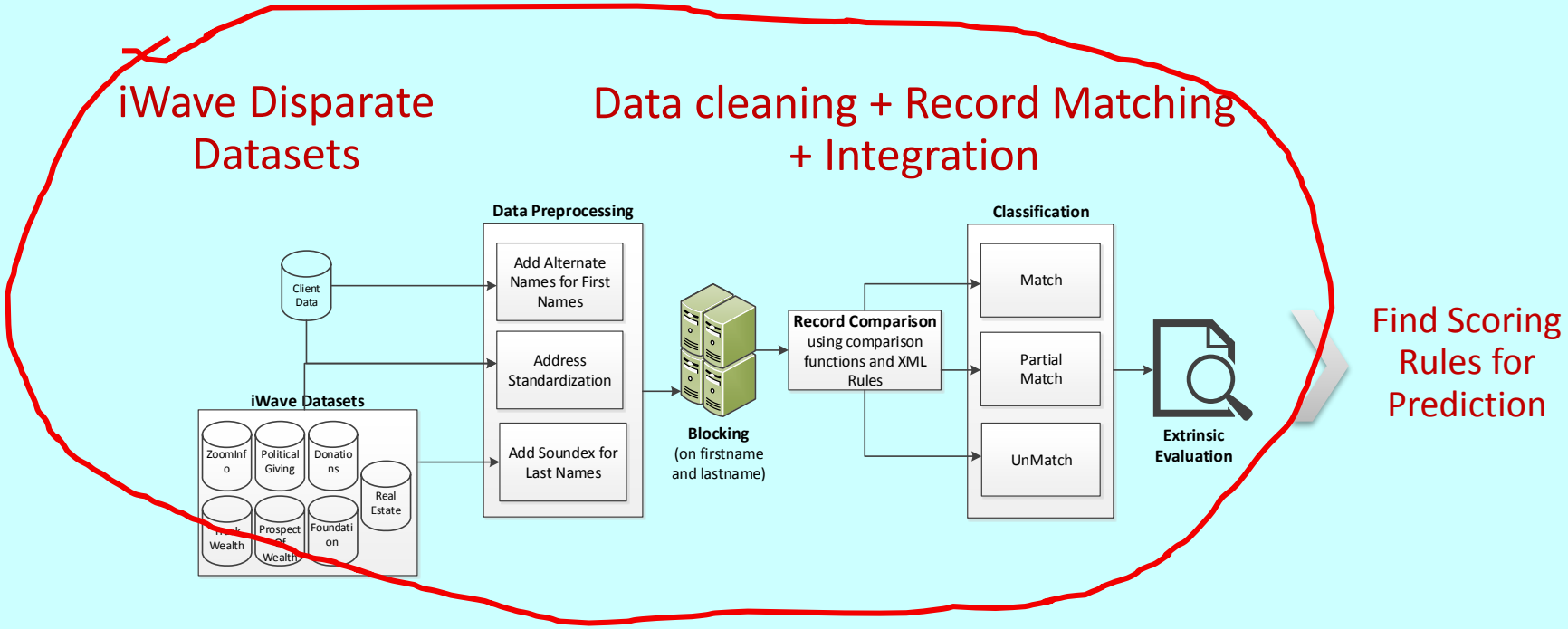
- Application website: http://www.homezilla.ca
- prof5408: Doc/Thesis/LeadGeneration2012.pdf



E.g. 2, “A framework for logic based information integration and knowledge discovery for prospect prediction and rating”

(NSERC Engage Project)

- Application website: https://www.iwave.com/
- Doc/Thesis/iWaveEngageProjectReport2015.pdf



Why Data Preprocessing?

- Data in the real world is dirty
  - **Incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - **Noisy:** containing errors or outliers
  - **Inconsistent:** containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - DM & DW need consistent integration of quality data

Why Data Preprocessing (cont)

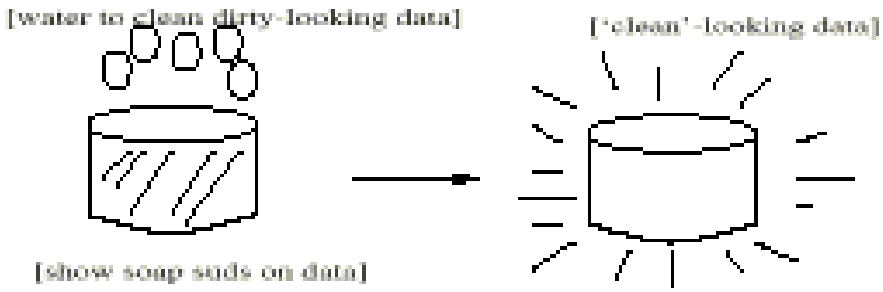
- What properties should quality data have?
  - **Relevant**
  - **Clean**
  - **Consistent**
  - **Enriched (Integrated)**
  - **Normalized and in right format and type**
- How to get quality data for DM/DW?

Major Tasks in Data Preprocessing

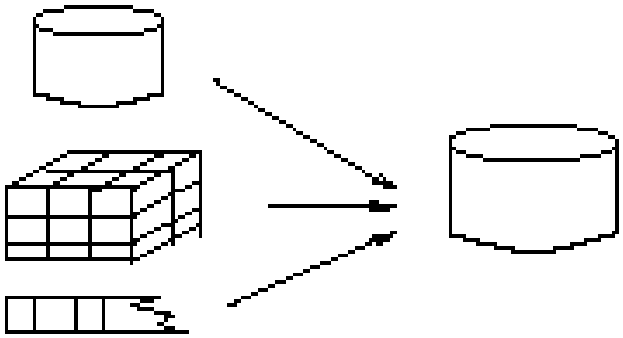
- **Data cleaning:**
  - Fill in missing values, correct errors, smooth noisy data/identify or remove outliers, and resolve inconsistencies.
- **Data integration:**
  - Integration of multiple databases, data cubes, or files.
- **Data transformation:**
  - Normalization and aggregation.
- **Data reduction:**
  - Obtains reduced representation in volume but produces the same or similar analytical results, including feature selection.
- Data discretization:
  - Part of data reduction but with particular importance, especially for numerical data.

Data Preprocessing: Major Tasks

Data Cleaning



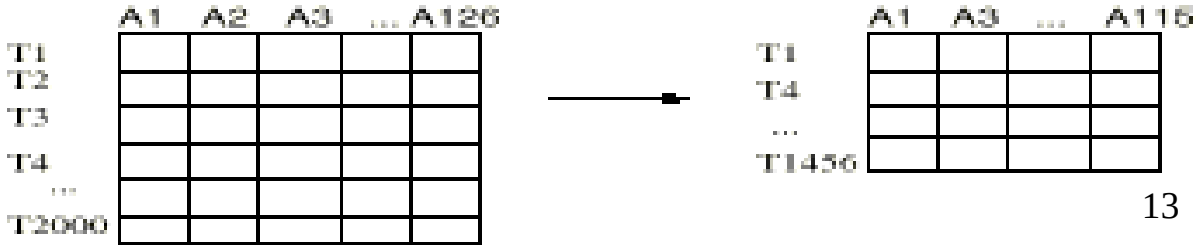
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



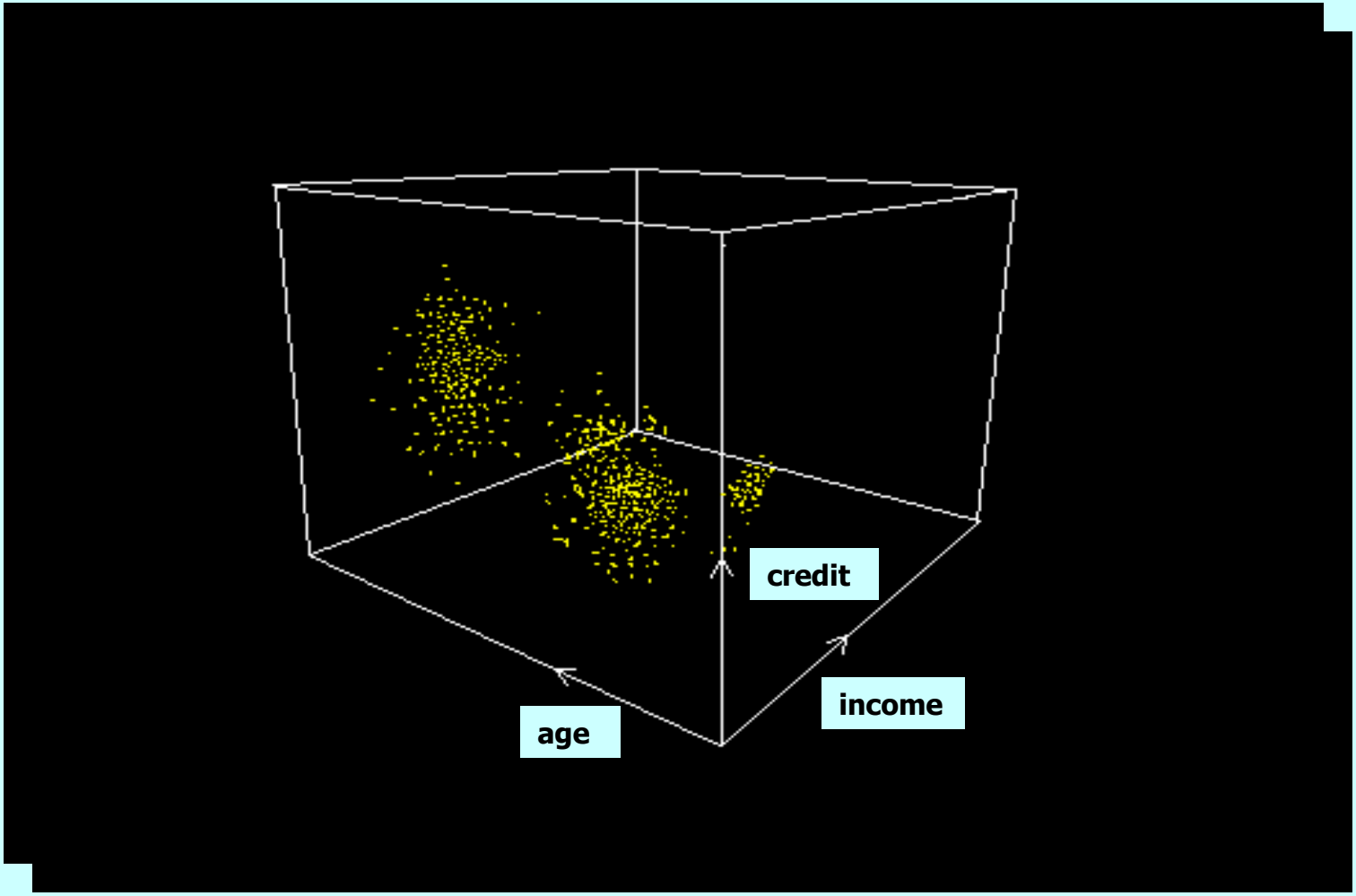
Data Cleaning

- **Data cleaning** or **data scrubbing** is the act of <u>detecting</u> and <u>correcting</u> (or removing) corrupt or inaccurate data records from a data set.
- The actions may include 1) find/fill in missing values, 2) identify and correct errors, 3) find and remove duplications, 4) identify and remove outliers, and 5) find and resolve inconsistencies.
- Examples:
  - **Incomplete: missing** attribute values: E.g., some records with *occupation= “ ”.*
  - **Noisy:** containing errors or outliers
    - E.g., *Salary= “-10”.*
    - E.g., The irrelevant part of an email for text mining.
  - **Inconsistent:** containing discrepancies in codes or names
    - E.g., *Age=“26”* vs. *Age (Birthday)=“03/07/1990”.*
    - E.g., *Rating= {“1, 2, 3”},* vs. *Rating= {“A, B, C”}.*

Data type/form Transformation

- Data type and form must be appropriate for the task
  - E.g., Text words are transformed into numerical weights for clustering mining.
  - E.g., Age (continuous) values are discredited into categorical labels, such as children, adults, seniors, etc.
  - E.g., Data normalization for clustering.

E.g., Clustering a 3D customer DB by measuring the distances between data points: any issue?



Data Normalization

E.g., Clustering mining for a customer database:

DB (Age, Income, Credit)

The distance (d) between to data points P1(x1,y1,z1) and P2(x2,y2,z2):

$$d(P1, P2)=\sqrt{(x2 - x1)^2 + (y2 - y1)^2 + (z2 - z1)^2}$$

	Age	**Income**	**Credit**
Customer1:	32	40,000	10,000
- Customer2:	24	30,000	2,000
	8	10,000	8,000
Normalized:	*1	*1/1000	*1/1000
	8	**10**	**8**
		(rescaled)	(rescaled)

If we scale all the attributes to the same order of magnitude we obtain <u>reliable distance measure</u> between the different records.

Data Integration

- Integration of multiple data sets, databases, data cubes, or other data files
  - E.g.3: “Web-based OLAP and data mining for CS student database” (Doc/Thesis/MACSprojOu03.pdf):
    - Built “Grade” data cube by integrating 11 datafiles: 3 Students files, 3 courses files, 3 Grades files, and other 2 registration files.
  - E.g.4: “Integrated data cube for Web content accessing pattern analysis” (Doc/Thesis/MACSprojTayyaba06.pdf):
    - Built “Count” data cube by integrating log data with web content data (the result of web pages clustering).
  - E.g.5: “An Integrated CRM Data Mining Method for Predicting Best Next Offer” (Doc/Thesis/MCSthesisWu05.pdf):
    - Integrate customer data from different databases and different tables into customer data set.

Data Reduction

- The process of reducing data size and making the prepared data more relevant. This includes choosing proper subsets from the original data, remove irrelevant attributes and object records, feature selection, etc.
  - All the projects presented in Doc/Thesis include this task in the data preparations stage.
  - E.g. 5: “An Integrated CRM Data Mining Method for Predicting Best Next Offer” (Doc/Thesis/MCSthesisWu05.pdf):
    - How to decide what features should be used for customer classification (i.e. choose the most relevant features to the classification target).

A Case Study: DP for discovering customer profiles to promote business sales

- **Business Background:**
  - **Business content data:** a publishing company sells magazines on cars, houses, sports, music, and comics.
  - **Typical information queries:**
    - *What is the typical profile of a reader of a car magazine?*
    - *Is there any correlation between an interest in cars and an interest in comics?" ...*
  - **Business DSS model:** finding clusters of clients and the profiles in order to set up a marketing exercise.
- **Data mining task:**
  - Mining clusters of clients & association analysis.
- **Data preparation for clustering:** Source data –(ETL)→ Target data
  - Data cleaning, data integration, data reduction, data transformation, etc.

Data Extraction: selection

- The company may have multiple data sources but the relevant datasets may include:
  - The subscription invoice dataset
  - The customer dataset
- The relevant data need to be determined including the tables and the attributes
  - The selected data need to be copied to a separated table

Table 1: Selected invoice data

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	01-01-01	comic
23013	King	3 High Road	02-30-95	sports
23019	Jonson	1 Downing Street	01-01-01	house

- The records Johnson and Jonson have different client numbers but the same address, which is a strong indication that they are the same person.
- Of course, we can never be sure of this, but a de-duplication algorithm using pattern analysis techniques could identify the situation and present it to a user to make a decision.

Data cleaning: Remove duplications

- **Record duplications**
  - Some clients may be represented by several records, some possible causes may include:
    - the result of negligence, such as people making typing errors
    - clients moving from on place to another without notifying change of the address
    - the cases in which people deliberately spell their names incorrectly or give incorrect information about themselves for avoiding a negative decision ...
  - This type of pollution will give a company the impression that it has more clients than the actual fact.

- **De-duplication**
  - The duplicated records may be identified by a pattern recognition algorithm, such as de-duplication algorithm, and then corrected.

1. Invoice data

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	01-01-01	comic
23013	King	3 High Road	02-30-95	sports
23019	*Jonson*	1 Downing Street	01-01-01	house

2. Data cleaning: De-duplication

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	01-01-01	comic
23013	King	3 High Road	02-30-95	sports
23003	Johnson	1 Downing Street	01-01-01	house

Data cleaning: correct domain inconsistency

- **Data inconsistency**
  - **Pollution may be caused by wrong domain values which are not consistent with the definitions**
    - E.g. When worked with a database in 1999, in the example table, date 01-01-01 means 1 January 1901 (the company did not even exist at that time).
  - **E.g. In some databases, analysis shows an unexpected high number of people born on 11 November:**
    - When people were forced to fill in a birth date on a screen and they either do not know or do not want to divulge it, they were inclined to type in `11-11-11‘.
    - This kind of untrue random values can be disastrous in a data mining context.
    - If information is unknown (NULL), it should be represented as such in the database.

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	01-01-01	comic
23013	King	3 High Road	02-30-95	sports
23003	Johnson	1 Downing Street	01-01-01	house

3. Table with corrected domain values in consistency

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	NULL	comic
23013	King	3 High Road	02-30-95	sports
23003	Johnson	1 Downing Street	12-20-94	house

Data enrichment (integration)

- The costumer database should be integrated together with the cleaned invoice data table

4. The additional “Costumer” dataset is available for integration

Client name	Date of birth	Income	Credit	Car owner	House owner
Johnson Clinton	04-13-76 10-20-71	$18,500 $36,000	$17,800 $26,600	no yes	no no

Integration: natural join operation of the two tables

Client number	Name	Address	Date purchase	Magazine purchased
23003	Johnson	1 Downing Street	04-15-94	car
23003	Johnson	1 Downing Street	06-21-93	music
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	NULL	comic
23013	King	3 High Road	02-30-95	sports
23003	Johnson	1 Downing Street	12-20-94	house

Client name	Date of birth	Income	Credit	Car owner	House owner
Johnson	04-13-76	$18,500	$17,800	no	no
Clinton	10-20-71	$36,000	$26,600	Yes	No
King	NULL	NULL	NULL	NULL	NULL

How to integrate the data from the two data sources?

5. Enriched data table

Credit number	Name	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	Clinton	10-20-11	$36,000	$26.600	yes	no	2 Boulevard	NULL	comic
23013	King	NULL	NULL	NULL	NULL	NULL	3 High Road	02-30-95	sports
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

Any irrelevant data should be removed?

5. Enriched table

Credit number	Name	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	Clinton	10-20-11	$36,000	$26.600	yes	no	2 Boulevard	NULL	comic
23013	King	NULL	NULL	NULL	NULL	NULL	NULL	02-30-95	sports
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

Data deduction

- **Remove the columns and rows which are not valuable to the DM process**
  - The column NAME and the row with multiple NULL values should be removed from the database.
  - In a real life data, maybe some of the missing data can be retrieved. However the records with missing data can not be retrieved should not participate a DM process (if the attributes are relevant).

* In some cases, especially fraud detection, lack of information can be a valuable indication of interesting patterns. Up to this point, the process phase has consisted of mainly simple SQL operations

5. Enriched table

Credit number	Name	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	Clinton	10-20-11	$36,000	$26.600	yes	no	2 Boulevard	NULL	comic
23013	King	NULL	NULL	NULL	NULL	NULL	NULL	02-30-95	sports
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

6. Table with column and row removed

Credit number	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	10-20-11	$36,000	$26.600	yes	no	2 Boulevard	NULL	comic
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

Data transformation

- For most of databases, the information provided is much too detailed to be used as input of data mining algorithms, such as Table 6.

- Apply the following transformation steps:
  1. Label address to region.
  2. Convert birth date to age.
  3. Divide income by 1000.
  4. Divide credit by 1000.
  5. Convert cars: yes-no to 1-0.
  6. Convert purchase date to month numbers starting from 1990.

6. Table with column and row removed

Credit number	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	10-20-11	$36,000	$26.600	yes	no	2 Boulevard	NULL	comic
23003	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

7. An intermediate coding stage

Credit number	Age	Income	Credit	Car owner	House owner	Region	Month of purchase	Magazine purchased
23003	20	18.5	17.8	0	0	1	52	car
23003	20	18.5	17.8	0	0	1	42	music
23003	20	18.5	17.8	0	0	1	29	comic
23009	25	36.0	26.6	1	0	1	NULL	comic
23003	20	18.5	17.8	0	0	1	48	house

Make each record represents a single customer:

Credit number	Age	Income	Credit	Car owner	House owner	Region	Month of purchase	Magazine purchased
23003	20	18.5	17.8	0	0	1	52	car
23003	20	18.5	17.8	0	0	1	42	music
23003	20	18.5	17.8	0	0	1	29	comic
23009	25	36.0	26.6	1	0	1	NULL	comic
23003	20	18.5	17.8	0	0	1	48	house

8. The final table

Credit number	Age	Income	Credit	Car owner	House owner	Region	Car magazine	House	Sports	Music	Comic
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

Edit View Favorites Tools Help

Back Search Favorites History

http://www.kdnuggets.com/software/dataxform.html

nuggets : Software : Data Transformation

oftware for Data Transformation and Cleaning

mmercial:

- Ab Initio, provides high-performance software library and graphical environment for data transformation
- NEW! AMADEA, data Extraction, Transformation, and Real Time Reporting software
- BioComp iManageData(tm), Accesses, cleans, filters, converts and transforms data from files, Excel, Oracle, SQL Server, process control systems and more.
- BrainMine, SAS add-ons for variable transformation in risk/response modeling.
- Data Manager, Visual Basic GUI application for data transformation for Win95/Win98.
- Data-Audit, diagnostic tool for market research and database evaluation
- Datagration, data cleansing tools that find and use data patterns.
- Dataskope, department-level tools to map, transform, alarm, output and view high volumes of binary or ASCII input data.
- Datawash, provides business to business marketers online access to prospect data & data enhancement services that clean and dedupe databases for increased deliverability and professional presentation.
- DEXTR's Data EXTRactor - mine, convert and transfer your data with one powerful, and easy-to-use software package for Win 95/98/NT and 3.1
- Digital Archaeology's Digital Excavator(TM), simplifies/speeds the preparation of data for modeling, using an intuitive GUI for data assembly, exploration, and the transformation of data from multiple sources into a single analysis set.
- NEW! DQnow, profiling, cleansing, and dedup tools, providing a clear view of the data
- GritBot, for identifying anomalies in data (compatible with See5 and Cubist).
- Genio

Internet

Summary

- DP is a big issue, and the most time cost process for DM and DW projects.
- Main tasks of data preparation include Data cleaning, integration, transformation, and reduction.
- Many tools have been developed for supporting data preprocessing.
- It is still an active research area because of the big effort needed for getting ready for DM/DW, such as feature selection, etc.
  - Research e.g., PhD thesis: "Data Reduction Techniques in Classification Processes" (2007)

(http://www.tdx.cat/bitstream/handle/10803/10479/lozano.pdf?sequence=1)

E.g.6, DP for Email Clustering

- **The application**:
  "Automatic text categorization and text retrieval for PPML archive“, Doc/Thesis/MCSthesisGuo00.pdf.
- **Data cleaning:**

  - Data deduction by taking out unnecessary headers: 13.7M->8.4M

  (only leave the message subject, author, date)

  - Remove the irrelevant files:

  The files with the following headings, such as unsubscribe, un-delivered message, ... Deduction: 8.47M → 7M.
- **Document normalization**
- **Generate domain based term dictionary**
- **Term weighting, …**

PPML raw data - Email messages

Header

Return-path: <Drhbg@aol.com>
Received: from DIRECTORY-DAEMON by SYSWRK.UCIS.DAL.CA (PMDF V4.3-13 #6307)
 id <01J615F5VHLS00BCUD@SYSWRK.UCIS.DAL.CA>; Fri, 01 Jan 1999 15:42:13 -0400
Received: from imo23.mx.aol.com by SYSWRK.UCIS.DAL.CA (PMDF V4.3-13 #6307)
 id <01J615F12Y6O00CSAS@SYSWRK.UCIS.DAL.CA>; Fri, 01 Jan 1999 15:42:07 -0400
Received: from Drhbg@aol.com by imo23.mx.aol.com (IMOv18.1)
 id NVXFa07005 for <pediatric-pain@ac.dal.ca>; Fri,
 1 Jan 1999 14:41:54 -0500 (EST)
Date: Fri, 01 Jan 1999 14:41:54 -0500 (EST)
From: Drhbg@aol.com
Subject: Re: Management of nerve injury
To: pediatric-pain@ac.dal.ca
Message-id: <7deeafb8.368d2502@aol.com>
MIME-version: 1.0
x-Mailer: AOL 2.5 for Windows
Content-type: text/plain; charset=US-ASCII
Content-transfer-encoding: 7bit

Body (Content)

I agree with William Fenton. I think mexiletine should be used as a second line drug.

I ordinarily treat patients with chronic neuropathic pain. However, on a number of occasions, I have treated patients with acute neuropathic pains such as sciatica or brachial plexopathies. I have prescribed gabapentin at the outset of the pain, and have found that patients have responded extremely well. They often require lower than anticipated dosages of opioid analgesics. I doubt there is any data on the benefits of early use of anticonvulsants, but a case-control study would be of value.

E.g., A sample thread

- Thread 1: opioids and meningitis
- **Date: Wed, 04 Jan 1995 16:54:48 -0500 (EST)From: posterSubject: opioids and meningitis**X is a 13 month (9.8kg) old boy suffering from acute meningitis (pneumocoque) treated with IV cefotaxime; at day three, I have been called as pediatric pain consultant to assess X; I have discovered an extreme painfull state: one could not handle or touch him without producing screaming. The child was unable to move spontaneously he looked paralysed by pain and hypertonia ; he also presented a neurological complication : ptosis at the right side.The pain treatment was IV acetaminophen. The first day I have prescribed IV Nalbuphine (weak opioid u antagonist and agonist) 11mg/24h after a loading dose of 1.4 mg; Pain at rest has been succesfully relieved but not the mobilisation pain; the dose has been increased at 14 mg/day wihout relieving the pain associated with moving; he has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4 days later. Neurological examination and CT scan have been still normal (except ptosis) during this period. No opioid's side effects have been observed.What do you think of this case ?Have you any experience with opioids and acute meningitis ?Dr Poster, Pediatric pain unit, Poster Hospital

  Date: Wed, 04 Jan 1995 17:27:25 -0500 (EST)From: first replySubject: re: opioids and meningitis Is there any periosteal involvement? If so an NSAID (ibuprofen or naproxen) may be much more effective than even opioid.

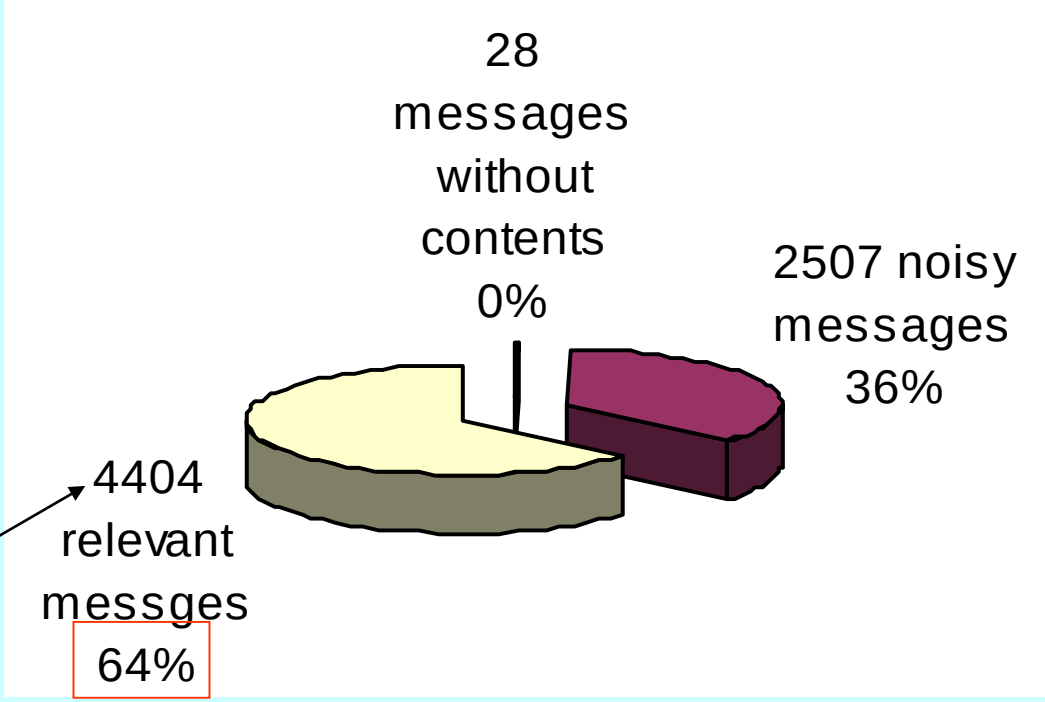
  Date: Wed, 04 Jan 1995 19:06:32 -0400From: second replySubject: Re: opioids and meningitis Poster writes:>X is a 13 month (9.8kg) old boy suffering from acute meningitis...>extreme painfull state: one could not handle or touch him without>producing screaming....>The first day I have prescribed IV Nalbuphine ...>succesfully relieved but not the mobilisation pain;...>has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4>days later. Neurological examination and CT scan have been still normal...I have used IV morphine for similar severe meningitis pain, with success. I wouldn't hesitate to use a pure opioid agonist (in conjunction with acetaminophen, NSAID, and/or tricyclics). However, it sounds like you have the situation under control.Second Reply, Associate Professor, Dept and University

  Date: Thu, 05 Jan 1995 18:58:32 -0800 (PST)From: Third ReplySubject: Re: opioids and meningitisI wonder if the problem is not due to severe arachnoiditis that is secondary to the inflammation. I would suggest a trial of steroids in this patient, perhaps in combination with a benzodiazepine to reduce the spasm. Narcotics may reduce the pain but I would not like to keep X on them for too long. Good luck Third Reply

doc-term matrix (vector space)

term / doc	t_1	t_2	…	t_i	…	t_t
d_1	w_{11}	w_{12}	…	w_{1i}	…	w_{1t}
d_2	w_{21}	w_{22}	…	w_{2i}	…	w_{2t}
…	…	…	…	…	…	…
d_j	w_{j1}	w_{j2}	…	w_{ji}	…	w_{jt}
…	…	…	…	…	…	…
d_m	w_{m1}	w_{m2}	…	w_{mi}	…	W_{mt}

Data Preparation - Data Cleaning



337duplicates

Composition of Raw PPML Data

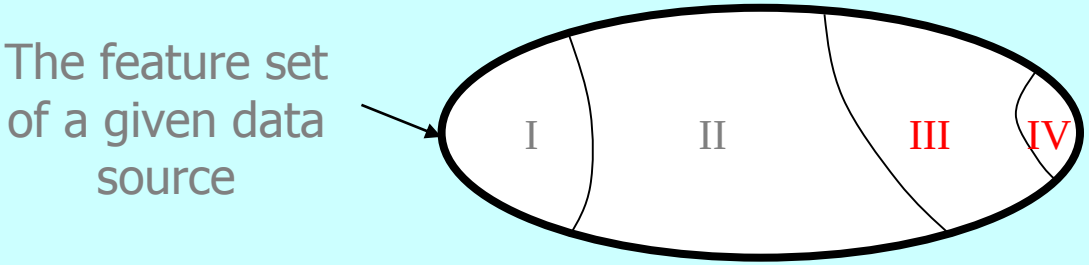
E.g.5, DP on Feature Selection for Classification

- **The application:** "An Integrated CRM Data Mining Method for Predicting Best Next Offer" (Thesis-Project/MCSthesisWu05.pdf)

- A feature is good for classification if it is
  - Relevant to the target and
  - Not redundant

Feature Selection for Classification

Feature subsets for classification



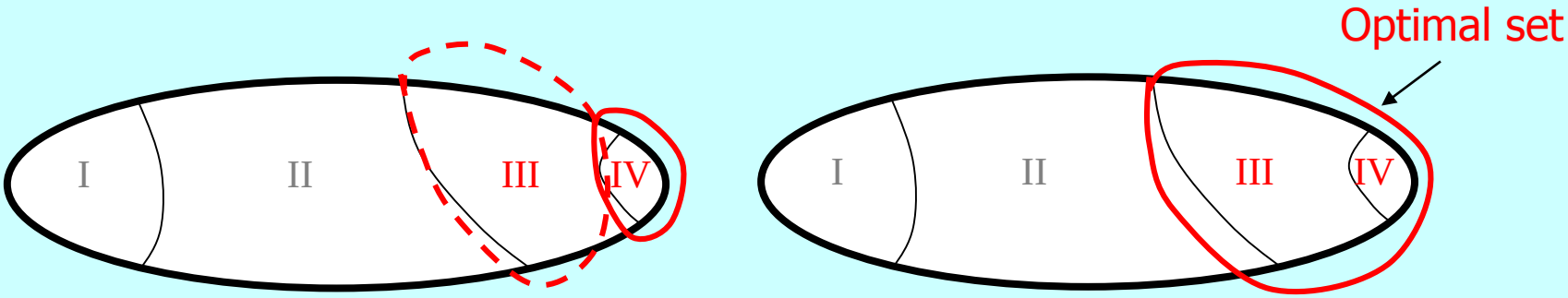
- I: Irrelevant features
- II: Weakly relevant and redundant features
- III: Weakly relevant but non-redundant features
- IV: Strongly relevant but non-redundant features
- III + IV: Optimal subset

Feature Selection (Cont)

- Feature selection strategies:
  - Evaluate features group by group (Wrapper based methods): <u>more accurate for a given dataset, high cost, over-fitting issue (over specific)</u>
  - Evaluate features one by one (Filter based methods): <u>very efficient, less accurate for the given dataset but more general</u>

Problem of Fast Correlation Based Filter (FCBF)

- FCBF may filter out too many useful features of subset III
  - E.g., 95% features of the test data sets (from a real financial data source) were filtered out by FBCF.



Evaluation of the Extended FCBF (EFCBF) Feature Selection

Data Set	#Instances	#Original Features	#Selected Features			
			EFCBF	FCBF	CFS	ReliefF
SV	3566	43	4	2	1	4
ML	3776	102	7	2	2	7
TD	4052	67	7	2	2	7
OD	4597	94	16	4	9	16
CL	5993	86	12	2	2	12
CQ	6575	89	14	4	2	14

Each of the following DM/DW project has a devoted chapter on DP
(Doc/Thesis/…)

- Text data analysis
  1. MECthesisWan15.pdf: An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis
  2. MCSthesisXu04.pdf: Text Sentiment Analysis on Survey Comments
  3. MACSprojSamp05.pdf: Detecting Semantic Orientation of Opinions Using Knowledge Based Approach
  4. MCSthesisGuo99.pdf: Text Clustering and Retrieval For PPML Archive

- Web data analysis
  5. MCSthesisJGuo04.pdf: Integrating Automatic Web Page Clustering Into Web Log association mining
  6. MACSprojTayyaba06.pdf: Integrated Data Cube for Web Content Accessing Pattern Analysis

- CRM for financial institutions
  7. EMCthsisWang03.pdf: Customer Profiling - Descriptive Data Mining for Financial Institutions
  8. MCSthesisWu05.pdf: An Integrated CRM Data Mining Method for Predicting Best Next Offer

- DW and OLAP applications
  10. MACSprojReportOu04.pdf: Web-based OLAP and Data Mining for CS Student Database
  11. MHINthesisNariman05.pdf: Designing a Framework of Intelligent Information Process on Dentistry Administration Data

- DM system for small busines
  12. HBthesisPothier99.pdf: Data mining in the small enterprise

Review Questions

1. Give three reasons why data need to be processed for DW and DM tasks.
2. What good properties should “quality data” have before conducting DM?
3. What are the typical tasks for DP (provide a brief description for each)?
4. Choose a DM/DW project as example, such as from Dec/Thesis or from the Internet, to examine & explain about its DP tasks.