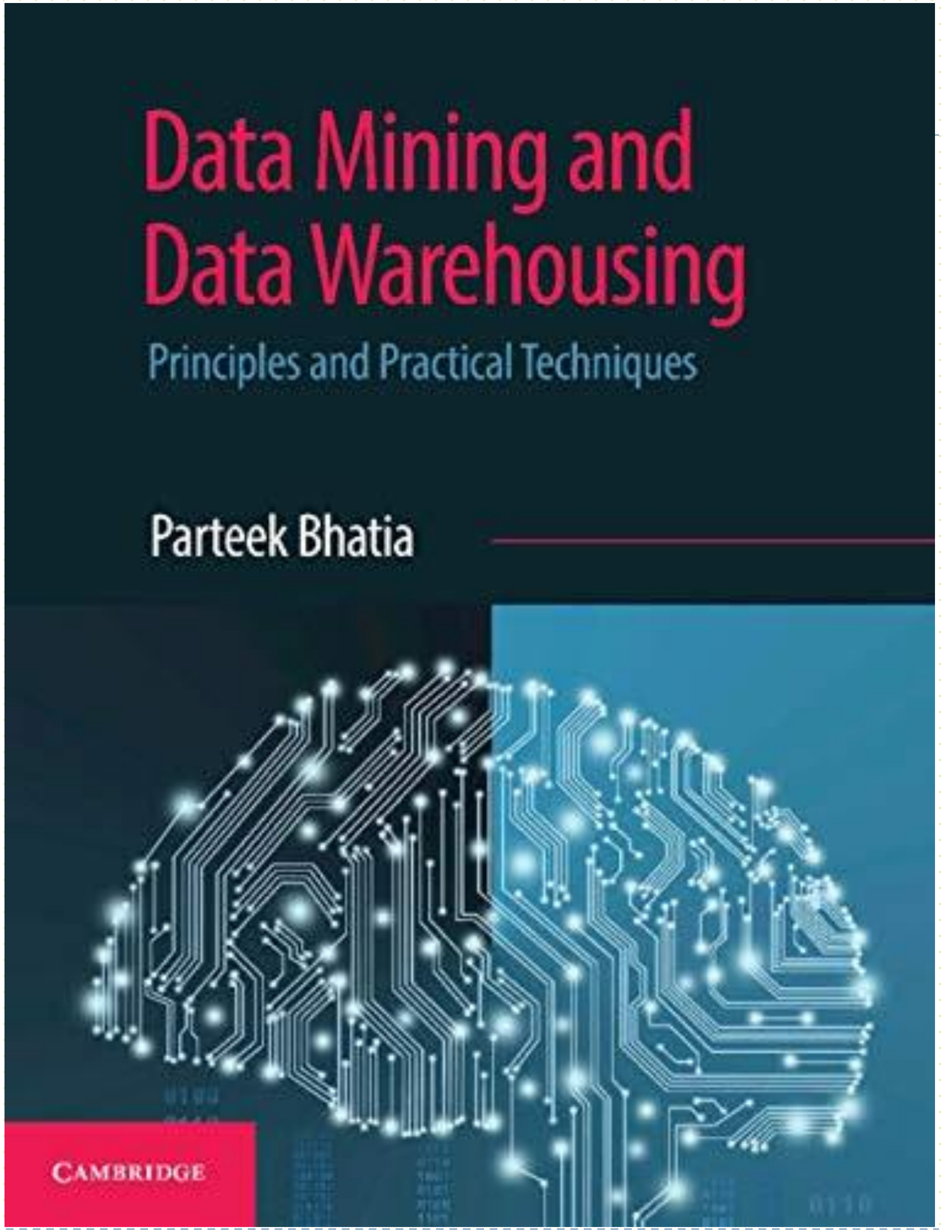
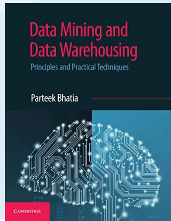
Chapter 4
Data Preprocessing



CHAPTER OBJECTIVES

1. To understand the need for data preprocessing.
2. To identify different phases of data preprocessing such as data cleaning, data integration,
3. Data transformation and data reduction



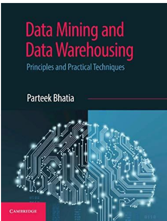
Data Mining and Data Warehousing

Principles and Practical Techniques

Parteek Bhatia

Need for Data Processing

- Data preprocessing is a data mining technique that involves transformation of raw data into an understandable format, because real world data can often be incomplete, inconsistent or even erroneous in nature.
- Data preprocessing resolves such issues. Data preprocessing ensures that further data mining process are free from errors. It is a prerequisite preparation for data mining, it prepares raw data for the core processes.



Data Pre-processing example

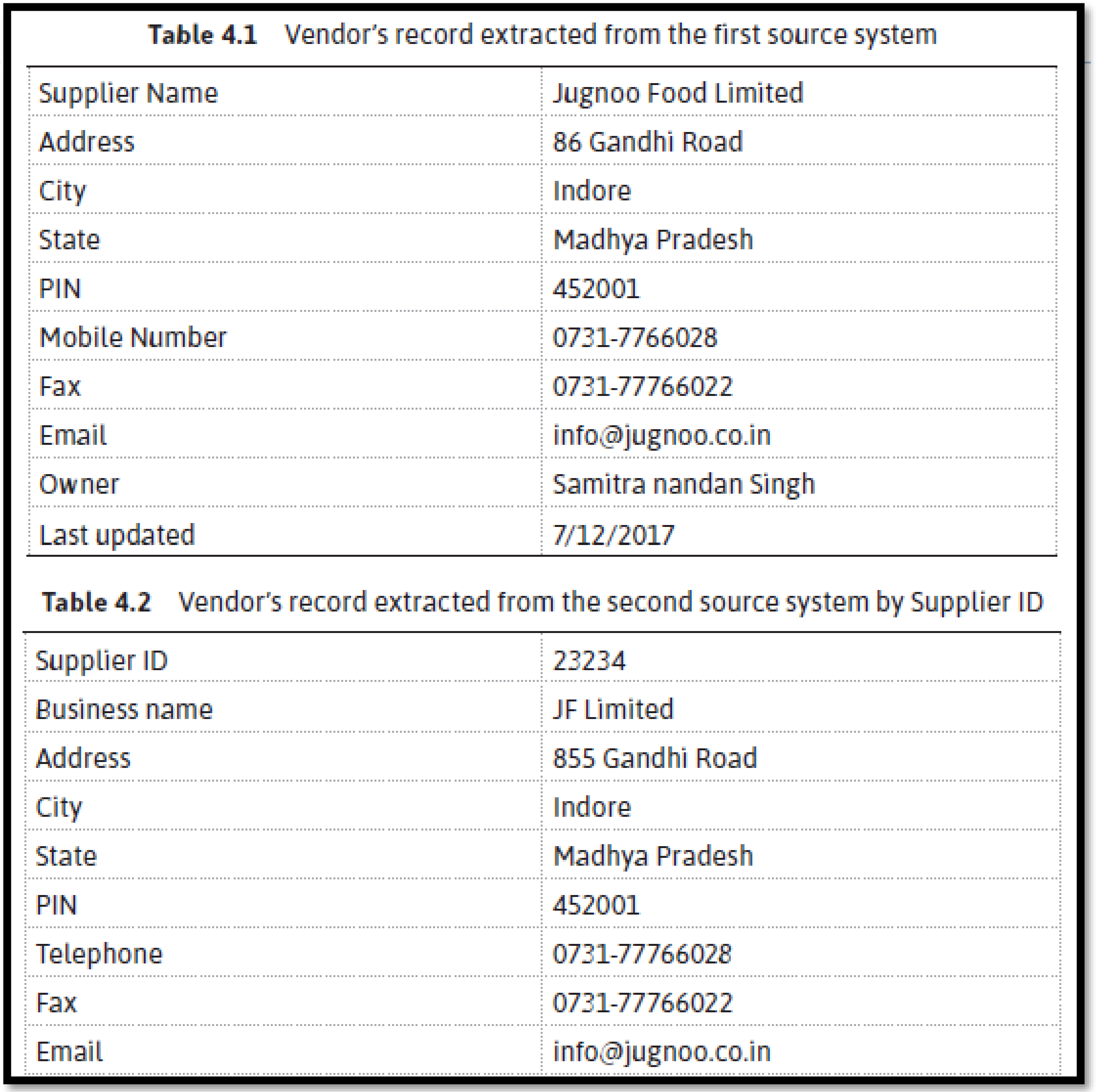


Table 4.1 Vendor's record extracted from the first source system

Supplier Name	Jugnoo Food Limited
Address	86 Gandhi Road
City	Indore
State	Madhya Pradesh
PIN	452001
Mobile Number	0731-7766028
Fax	0731-77766022
Email	info@jugnoo.co.in
Owner	Samitra nandan Singh
Last updated	7/12/2017

Table 4.2 Vendor's record extracted from the second source system by Supplier ID

Supplier ID	23234
Business name	JF Limited
Address	855 Gandhi Road
City	Indore
State	Madhya Pradesh
PIN	452001
Telephone	0731-77766028
Fax	0731-77766022
Email	info@jugnoo.co.in

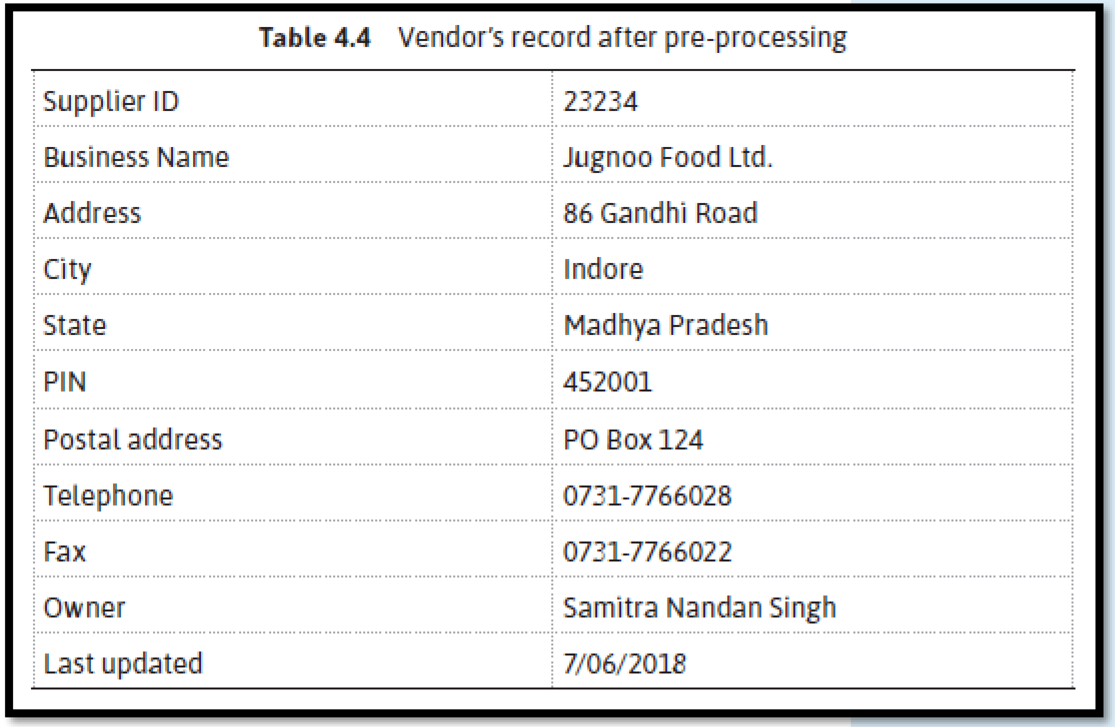
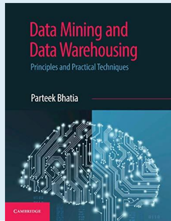
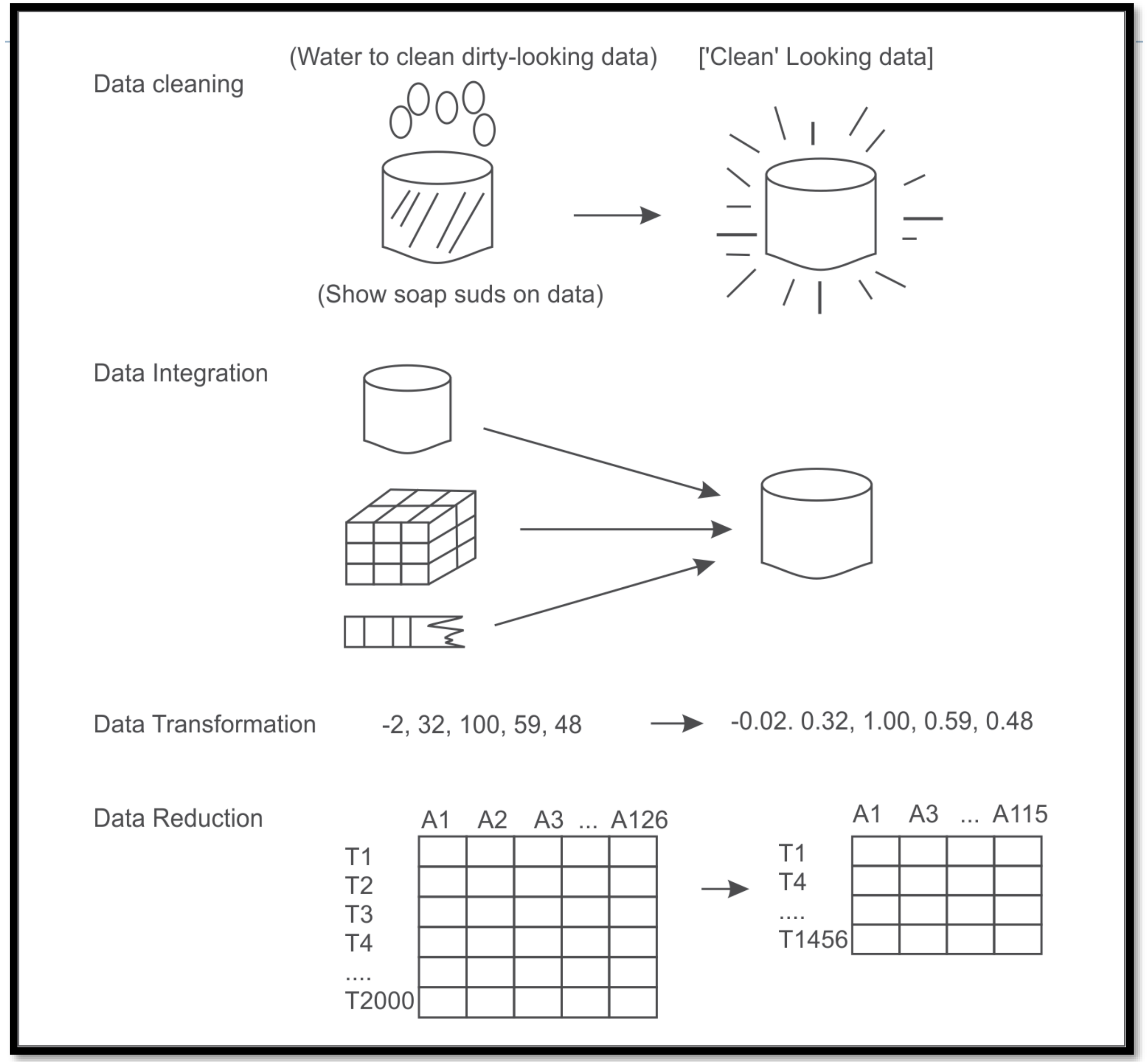


Table 4.4 Vendor's record after pre-processing

Supplier ID	23234
Business Name	Jugnoo Food Ltd.
Address	86 Gandhi Road
City	Indore
State	Madhya Pradesh
PIN	452001
Postal address	PO Box 124
Telephone	0731-7766028
Fax	0731-7766022
Owner	Samitra Nandan Singh
Last updated	7/06/2018



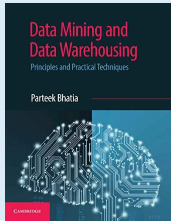
Data Pre-processing Methods



Raw data is highly vulnerable to missing values, noise and inconsistency and the quality of data affects the data mining results.

The various stages in which data preprocessing is performed:

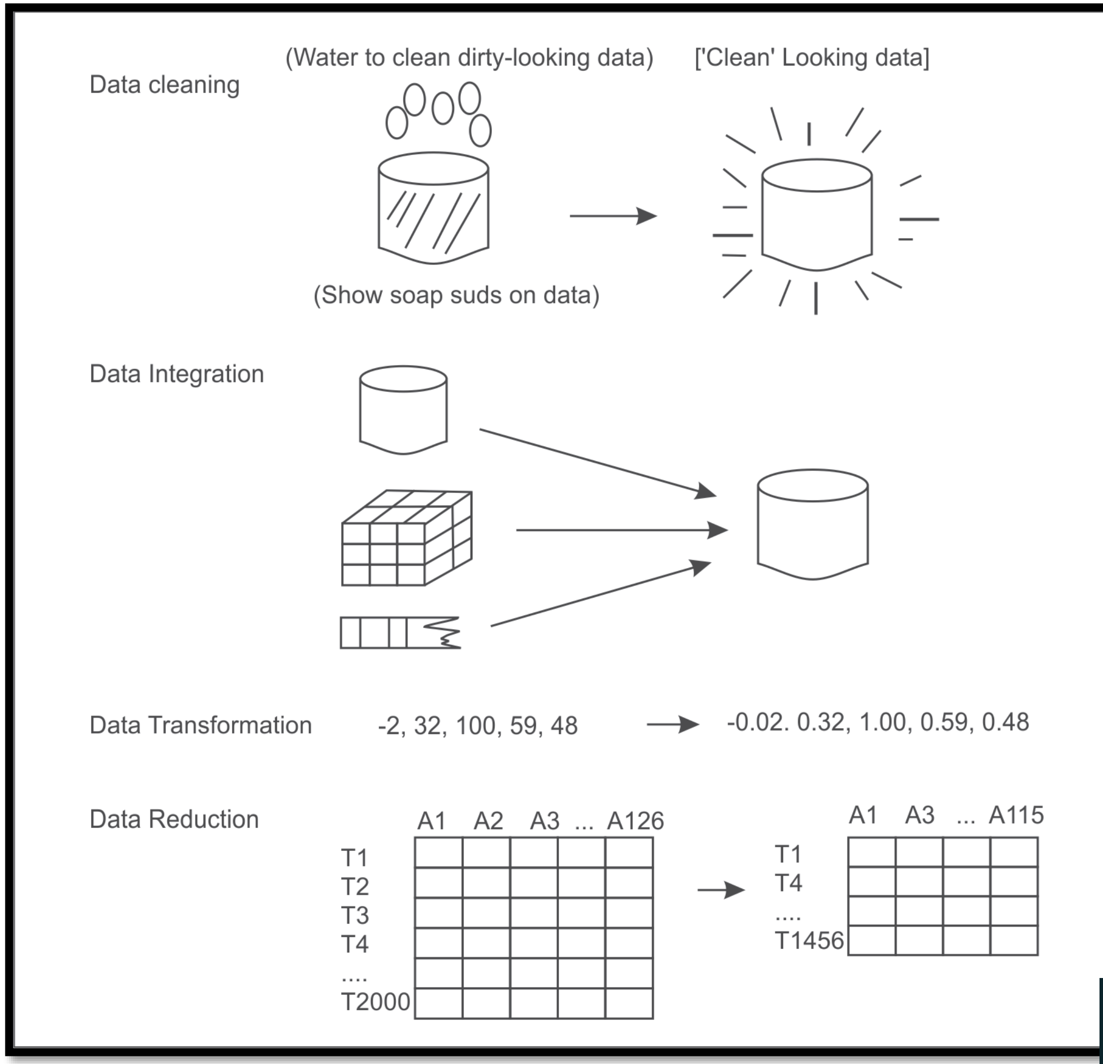
1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction


Data Mining and Data Warehousing
Principles and Practical Techniques
Parteek Bhatia

Data Pre-processing Methods

Data Cleaning

- Raw data or noisy data goes through the process of cleansing first. In Data cleansing missing values are filled, noisy data is smoothened, inconsistencies are resolved, outliers are identified and removed in order to clean the data.



Data cleaning

(Water to clean dirty-looking data)

['Clean' Looking data]

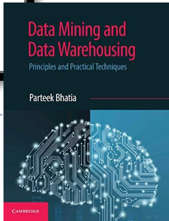
(Show soap suds on data)

Data Integration

Data Transformation

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

Data Reduction



Data Mining and Data Warehousing

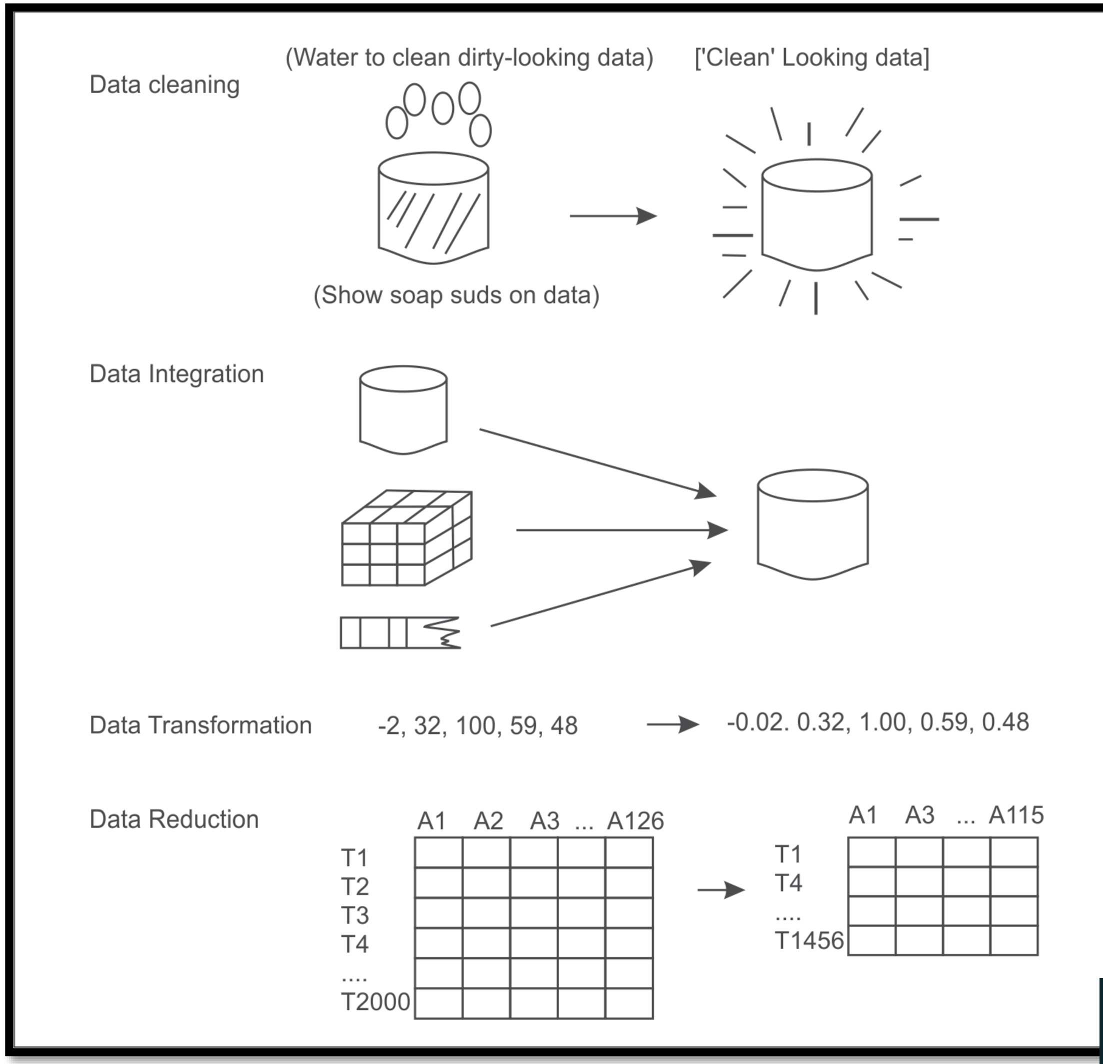
Principles and Practical Techniques

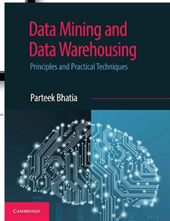
Parteek Bhatia

Data Pre-processing Methods

Operations Performed during Data Cleaning

- Handling of Missing Values
- Handling of Noisy Data
- Handling of Inconsistent data

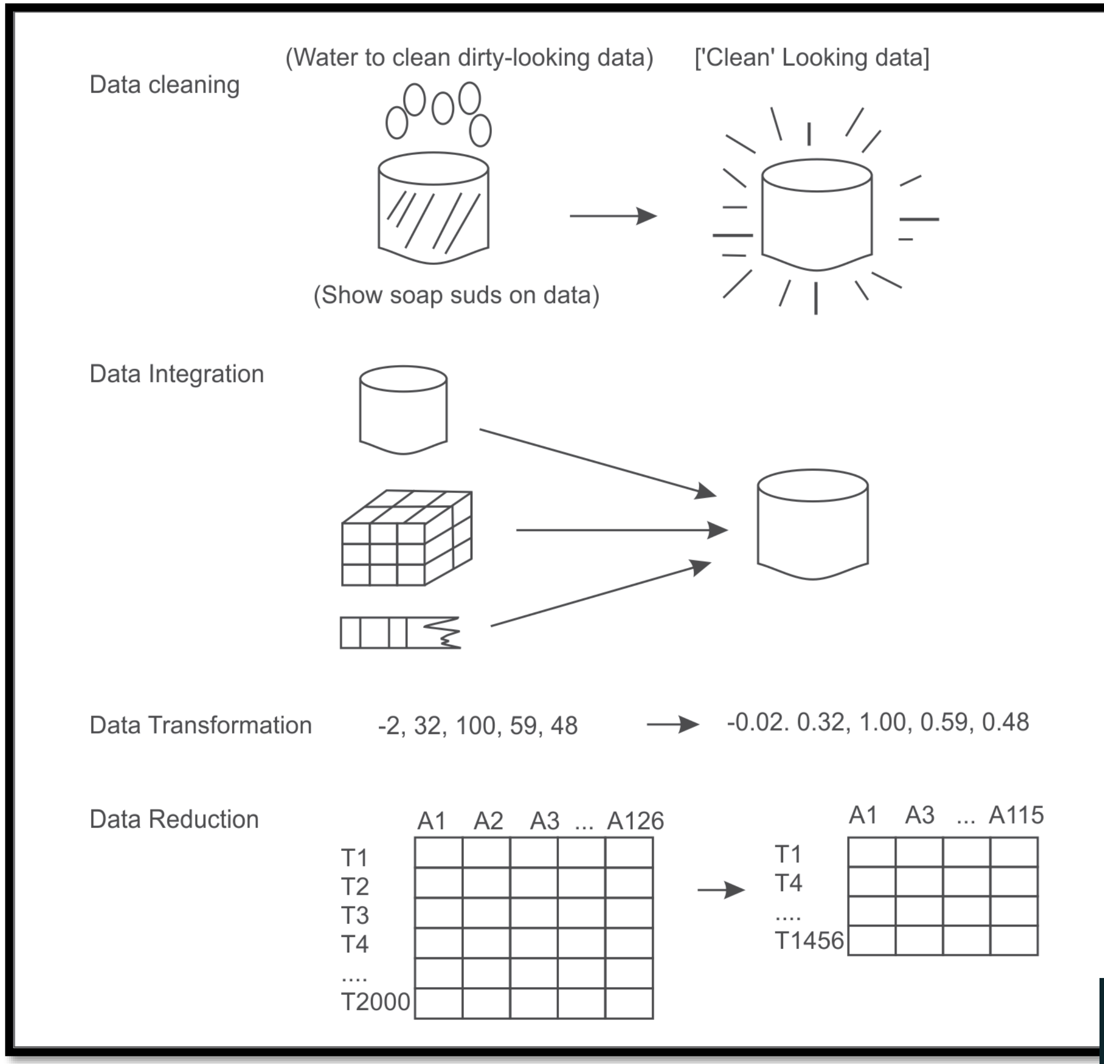


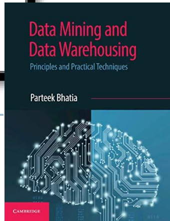

Data Mining and Data Warehousing
Principles and Practical Techniques
Parteek Bhatia

Data Pre-processing Methods

Data Integration

- One of the most necessary steps taken during the data analysis is Data Integration. Data integration is a method which combines data from plethora of sources (such as multiple databases, flat files or data cubes) into a unified data store.

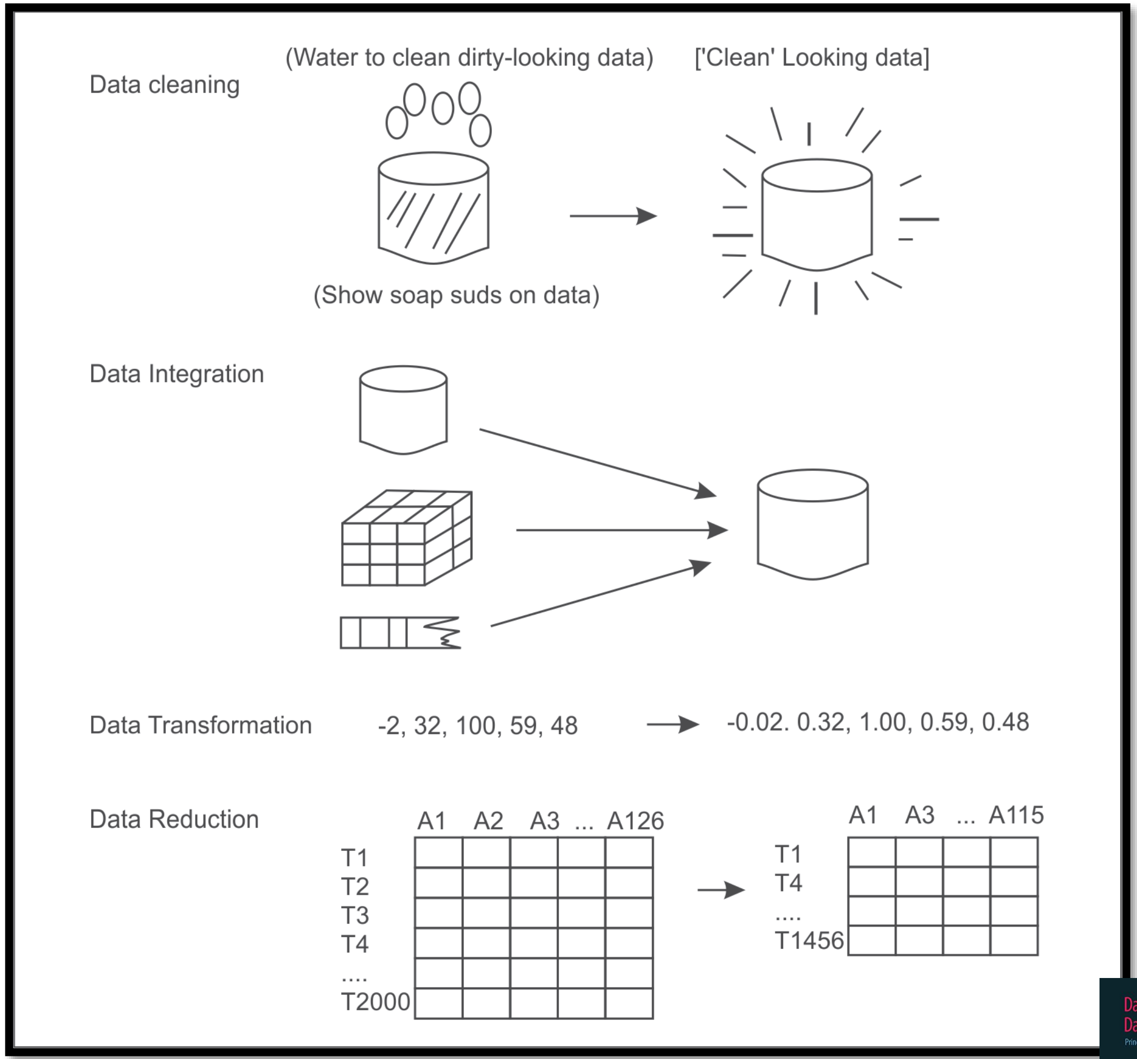


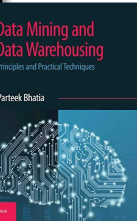

Data Mining and Data Warehousing
Principles and Practical Techniques
Parteek Bhatia

Data Pre-processing Methods

Data Transformation

- Sometimes, the value of one attribute may be small as compared to other attributes, then in this scenario, that attribute will not have much influence on mining of information, since the values of this attribute were smaller than other attributes and the variation within the attribute will also be small.
- Normalization and Standardization are most popular and widely used Data Transformation methods.

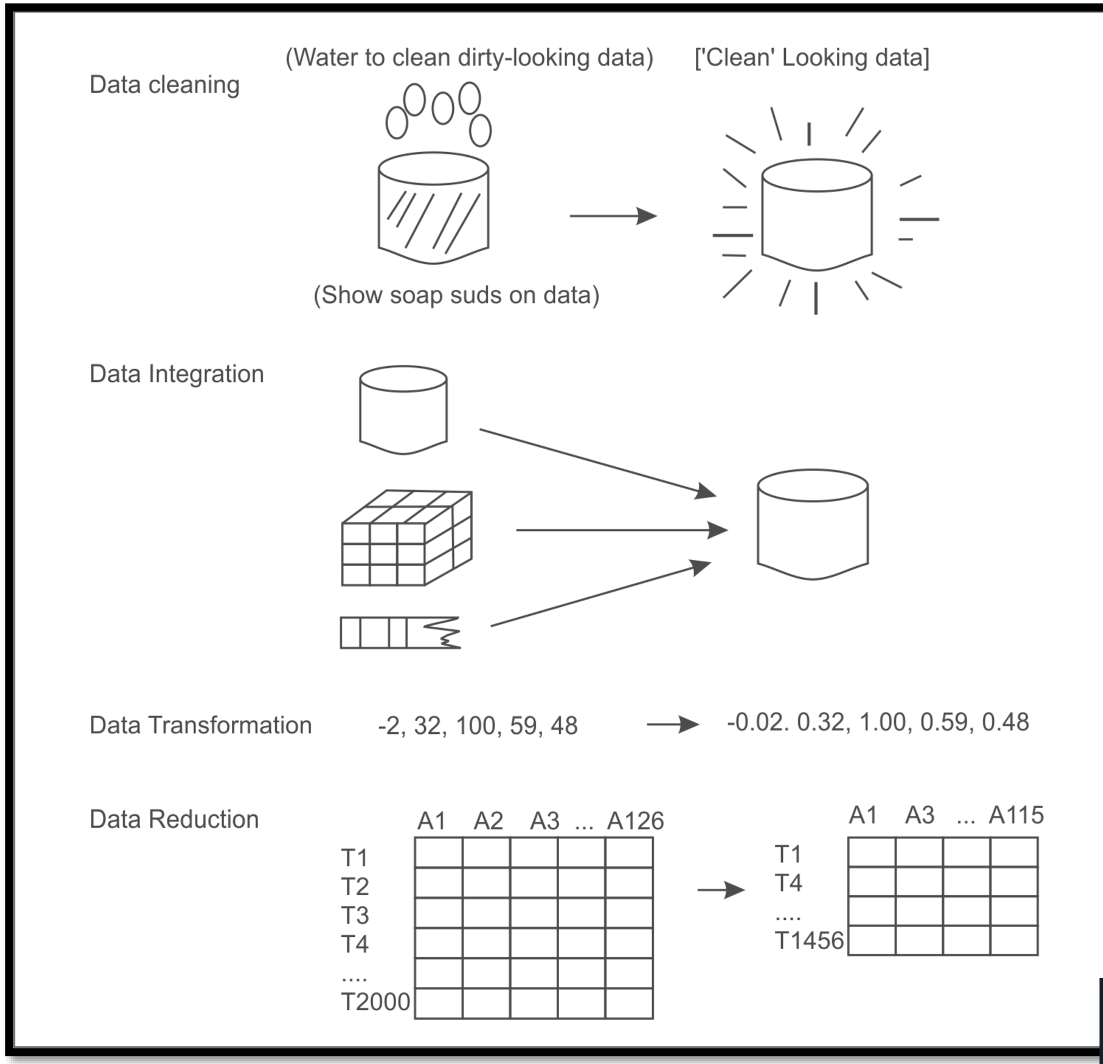



Data Mining and Data Warehousing
Principles and Practical Techniques
Parteek Bhatia

Data Pre-processing Methods

Data Reduction

- It is often seen that, when the complex data analysis and mining processes are carried out over humongous data sets, they cost a very long time, henceforth making the whole data mining or analysis process unsuitable (or not feasible). Data reduction techniques come for the rescue in such scenarios. Using data reduction techniques dataset could be represented in a reduced manner without actually compromising the integrity of original data.



Data cleaning

(Water to clean dirty-looking data)

['Clean' Looking data]

(Show soap suds on data)

Data Integration

Data Transformation

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

Data Reduction

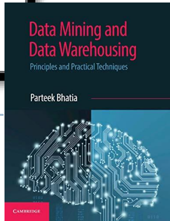
A1 A2 A3 ... A126

T1
T2
T3
T4
....
T2000

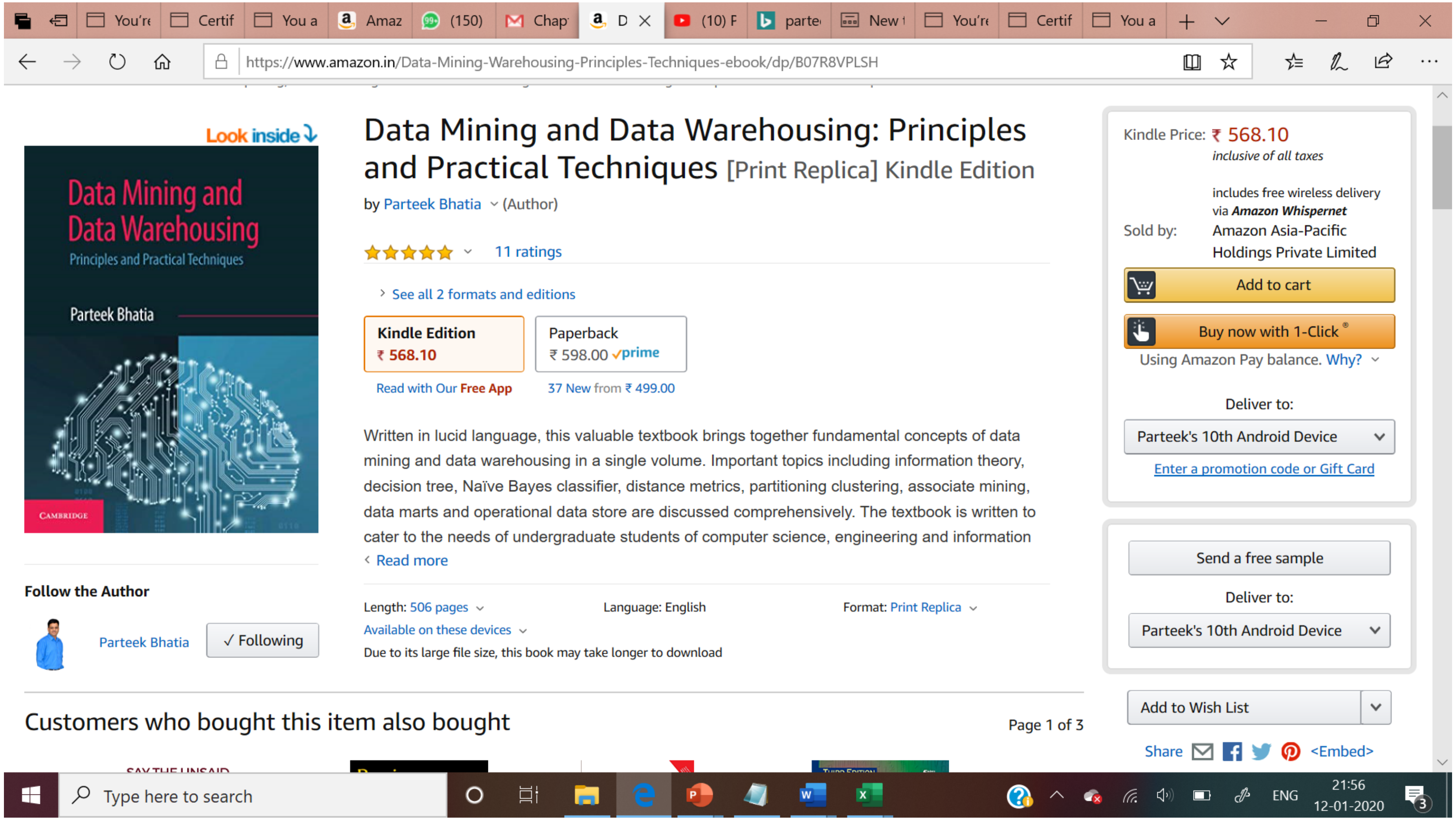
→

A1 A3 ... A115

T1
T4
....
T1456


Data Mining and Data Warehousing
Principles and Practical Techniques
Parteek Bhatia

Reference



https://www.amazon.in/Data-Mining-Warehousing-Principles-Techniques-ebook/dp/B07R8VPLSH

Data Mining and Data Warehousing: Principles and Practical Techniques [Print Replica] Kindle Edition

by Parteek Bhatia (Author)

★★★★★ 11 ratings

See all 2 formats and editions

Kindle Edition	Paperback
₹ 568.10	₹ 598.00 ✓prime
Read with Our Free App	37 New from ₹ 499.00

Written in lucid language, this valuable textbook brings together fundamental concepts of data mining and data warehousing in a single volume. Important topics including information theory, decision tree, Naïve Bayes classifier, distance metrics, partitioning clustering, associate mining, data marts and operational data store are discussed comprehensively. The textbook is written to cater to the needs of undergraduate students of computer science, engineering and information

Read more

Length: 506 pages Language: English Format: Print Replica

Available on these devices

Due to its large file size, this book may take longer to download

Follow the Author

Parteek Bhatia ✓ Following

Customers who bought this item also bought

Page 1 of 3

Kindle Price: ₹ 568.10
inclusive of all taxes

includes free wireless delivery via ***Amazon Whispernet***

Sold by: Amazon Asia-Pacific Holdings Private Limited

Add to cart

Buy now with 1-Click®

Using Amazon Pay balance. Why?

Deliver to:

Parteek's 10th Android Device

Enter a promotion code or Gift Card

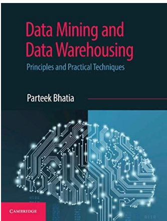
Send a free sample

Deliver to:

Parteek's 10th Android Device

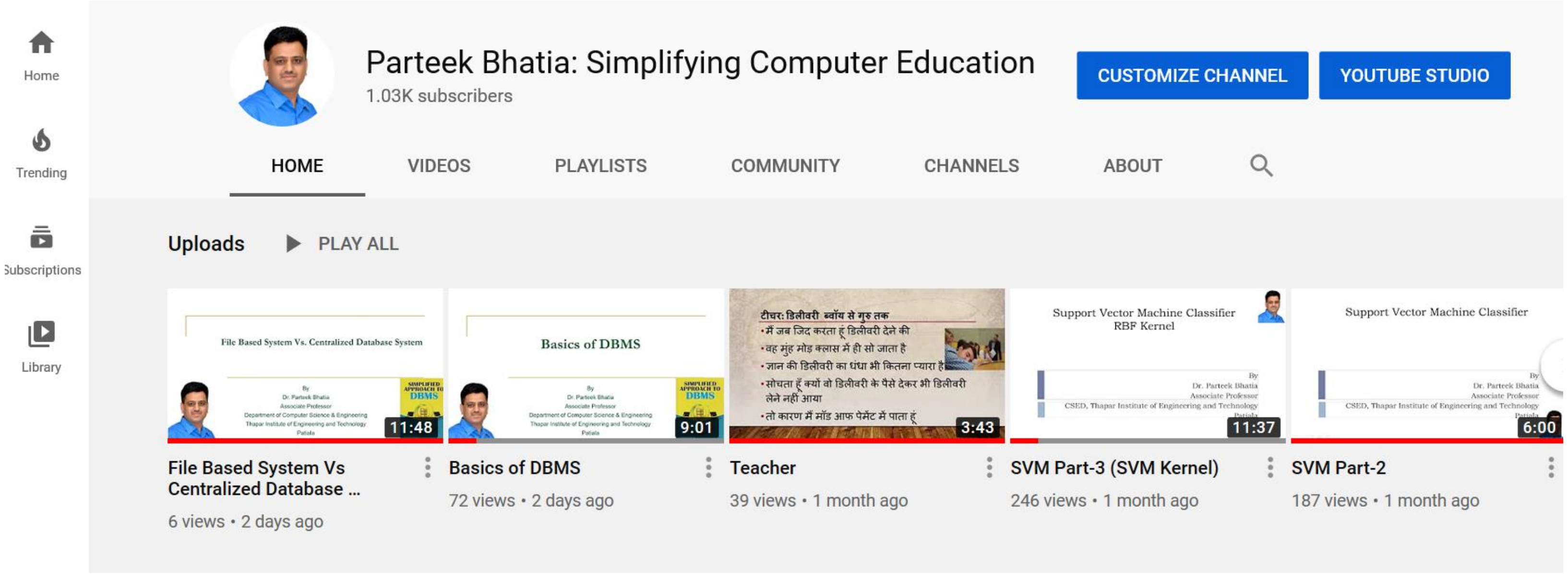
Add to Wish List

Share <Embed>



For more information

- Subscribe to YouTube Channel from the Author
  - To receive latest video tutorials on Data Mining, Machine Learning, DBMS, Big Data, NoSQL and many more.
- https://www.youtube.com/user/parteekbhatia



Parteek Bhatia: Simplifying Computer Education

1.03K subscribers

Uploads

File Based System Vs Centralized Database ...
6 views • 2 days ago

Basics of DBMS
72 views • 2 days ago

Teacher
39 views • 1 month ago

SVM Part-3 (SVM Kernel)
246 views • 1 month ago

SVM Part-2
187 views • 1 month ago

Free Online on SQL at Udemy

Udemy

ROYAL ACADEMY OF ENGINEERING Funded

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY (Deemed to be University)

Simplified Approach to SQL

QUICK START
Ace Interviews and College Exams in 1 month

BI - LINGUAL
Explanation of each of the topic in English and Hindi

INTERACTIVE
Discussion forums and direct messaging to instructor

RELAXING
Study the course at your own pace

STUDY ANYWHERE
Study on your smartphone with udemy app

COURSE CONTENT

Week 1 : *Introduction to SQL and performing basic operations with SQL.*
Week 2 : *Creation of Tables with Integrity constraints.*
Week 3 : *Table Alterations and Joins.*
Week 4 : *Grouping of Data*

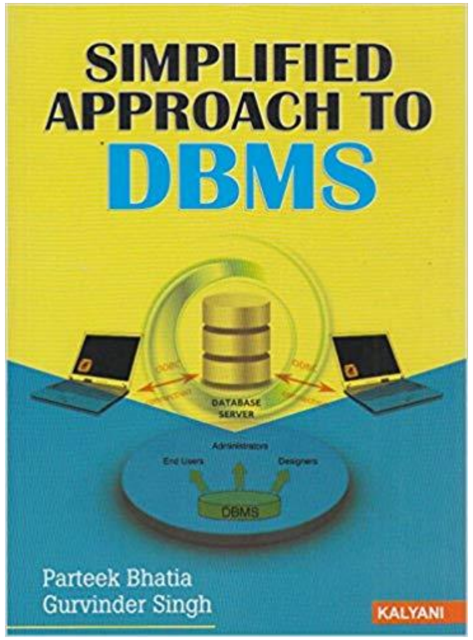
ABOUT THE INSTRUCTOR

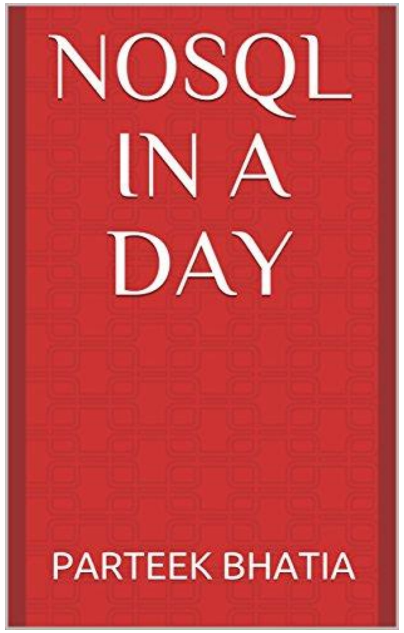
Dr. Parteek Bhatia is Associate Professor in the Department of Computer Science and Engineering at Thapar Institute of Engineering and Technology, Patiala. He has more than 18 years of academic experience.He has authored several books in various areas of computer science. His book - Simplified approach to DBMS is one of the bestseller. Currently, he is working on plethora of Projects which are funded by Department of Science and Technology, CSIR and other funding agencies of India.

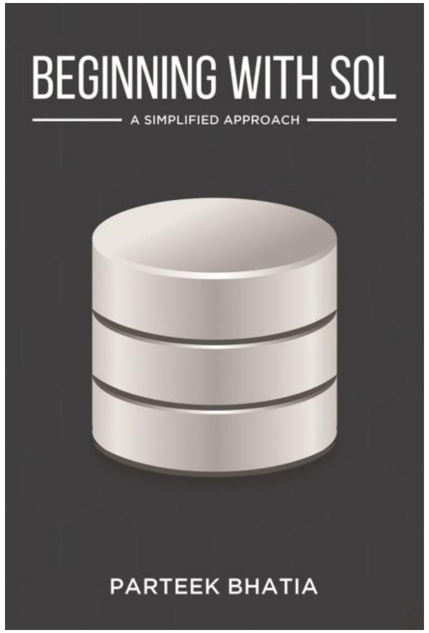
QR CODE

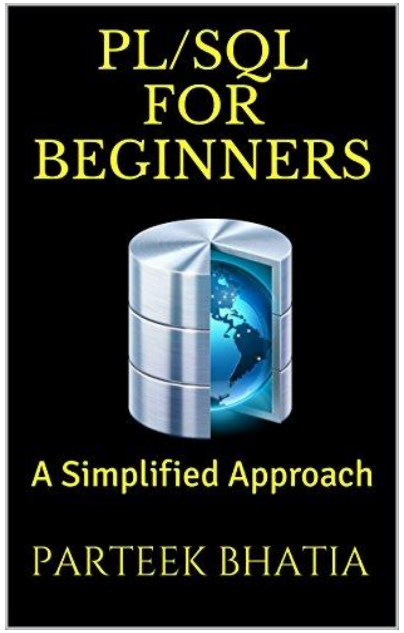
COURSE LINK: https://www.udemy.com/learn-sql-in-a-simplified-manner/

Books from the Same Author

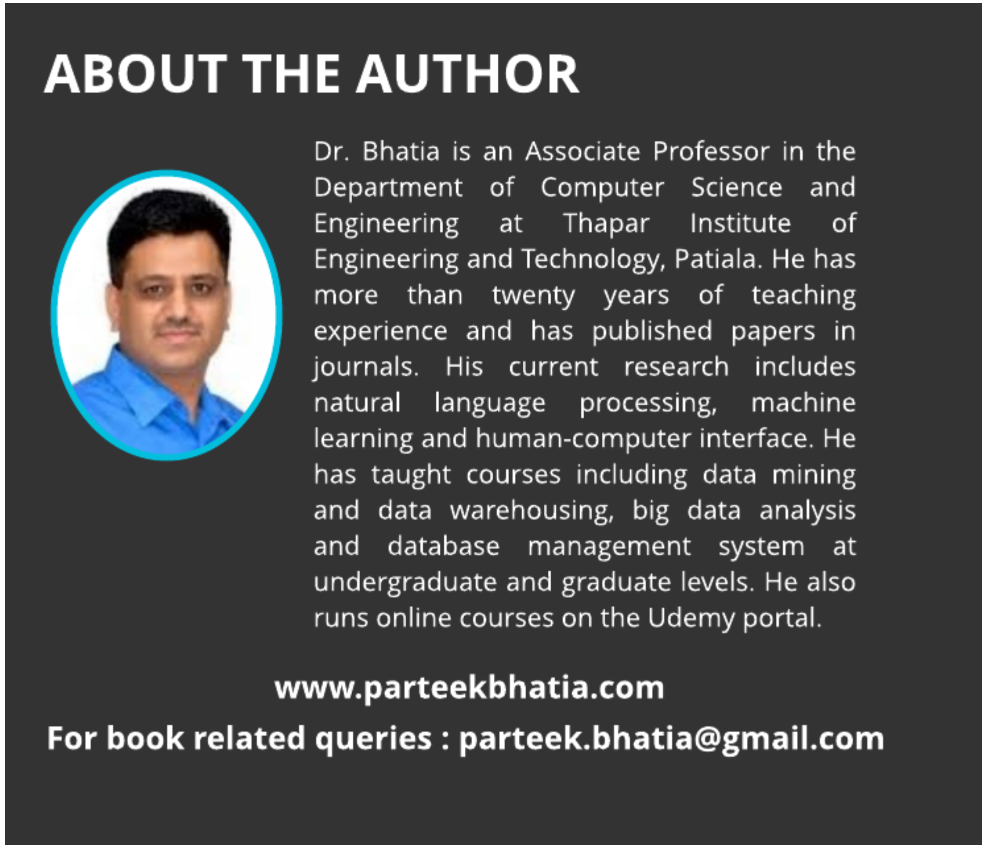








For more information visit: www.parteekbhatia.com

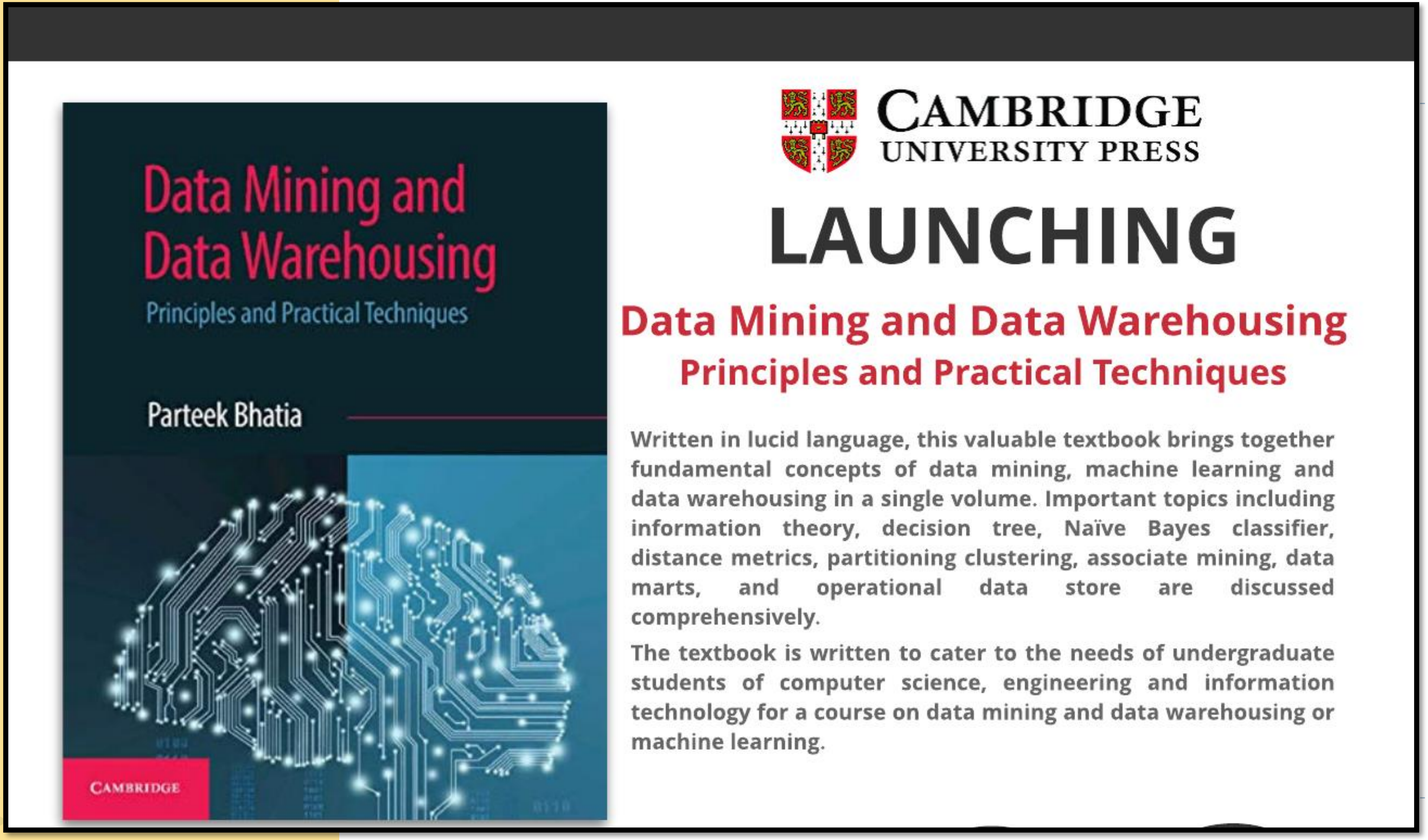


ABOUT THE AUTHOR

Dr. Bhatia is an Associate Professor in the Department of Computer Science and Engineering at Thapar Institute of Engineering and Technology, Patiala. He has more than twenty years of teaching experience and has published papers in journals. His current research includes natural language processing, machine learning and human-computer interface. He has taught courses including data mining and data warehousing, big data analysis and database management system at undergraduate and graduate levels. He also runs online courses on the Udemy portal.

www.parteekbhatia.com

For book related queries : parteek.bhatia@gmail.com



Data Mining and Data Warehousing

Principles and Practical Techniques

Parteek Bhatia

CAMBRIDGE

CAMBRIDGE
UNIVERSITY PRESS

LAUNCHING

Data Mining and Data Warehousing

Principles and Practical Techniques

Written in lucid language, this valuable textbook brings together fundamental concepts of data mining, machine learning and data warehousing in a single volume. Important topics including information theory, decision tree, Naïve Bayes classifier, distance metrics, partitioning clustering, associate mining, data marts, and operational data store are discussed comprehensively.

The textbook is written to cater to the needs of undergraduate students of computer science, engineering and information technology for a course on data mining and data warehousing or machine learning.

ABOUT THE AUTHOR

Dr. Bhatia is an Associate Professor in the Department of Computer Science and Engineering at Thapar Institute of Engineering and Technology, Patiala. He has more than twenty years of teaching experience and has published papers in journals. His current research includes natural language processing, machine learning and human-computer interface. He has taught courses including data mining and data warehousing, big data analysis and database management system at undergraduate and graduate levels. He also runs online courses on the Udemy portal.

www.parteekbhatia.com

For book related queries : parteek.bhatia@gmail.com