DATA MINING-INTRODUCTION

B.Sc. 6th Sem (Paper Code: DSE4)
Paulami Basu Ray
Assistant Professor,Department of Computer Science and Applications
Prabhat Kumar College, Contai

PLACE IN B.SC. SYLLABUS

- Computer Science Hons. (COSH), under Vidyasagar University (V.U.)

- B.Sc. Semester VI (C.B.C.S.) Discipline Specific Elective (DSE4)
- Module 1- Overview

CONTENTS

- Why Data Mining?
- Major Sources of Abundant Data
- What is Data Mining?
- Why not Traditional Data Analysis?
- Process of Knowledge Discovery from Data

WHY DATA MINING?

- We live in a world where vast amounts of data are collected daily. Analysing such data is an important need.

- By studying this subject we aim to see how data mining can meet this need by providing tools to discover knowledge from data.

MAJOR SOURCES OF ABUNDANT DATA : WEB

A search engine (e.g. Google) receives hundreds of millions of queries everyday.

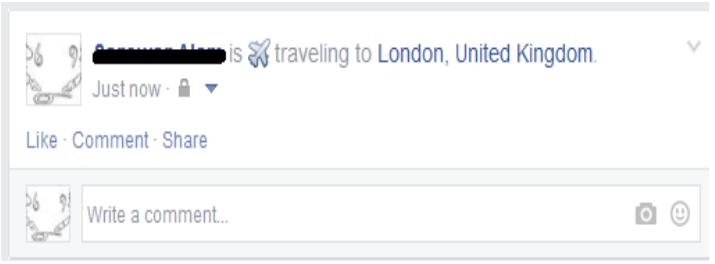
Inference: If I am searching Data Mining maybe it is trending or this information can be used for marketing Data Mining books/ e-contents to me and so on…..

MAJOR SOURCES OF ABUNDANT DATA : E-COMMERCE

Millions of transactions are recorded daily in e-commerce sites.

This Recommendation System works with analysing buying patterns of users of similar interests

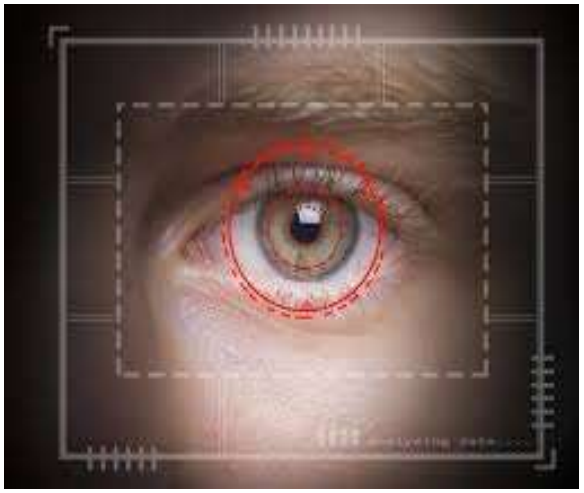
MAJOR SOURCES OF ABUNDANT DATA : SOCIAL NETWORKING



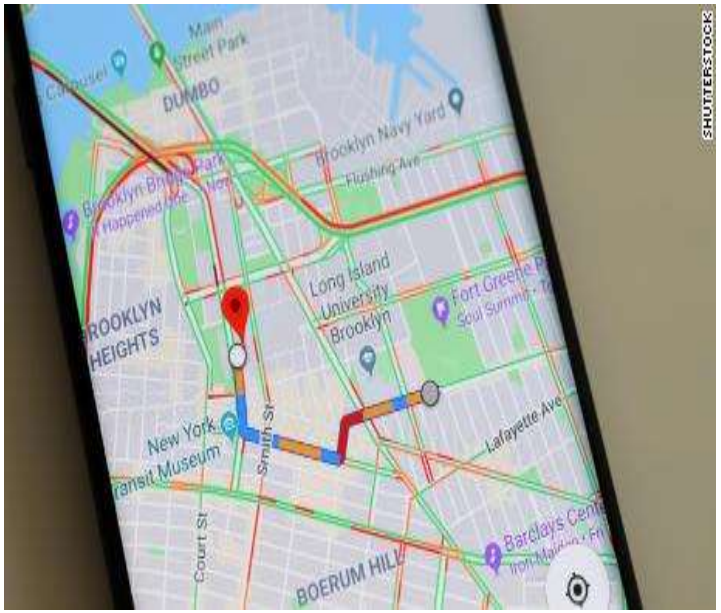
Our Travelling to.. Or Check-In information may attract some wanted/unwanted attention!

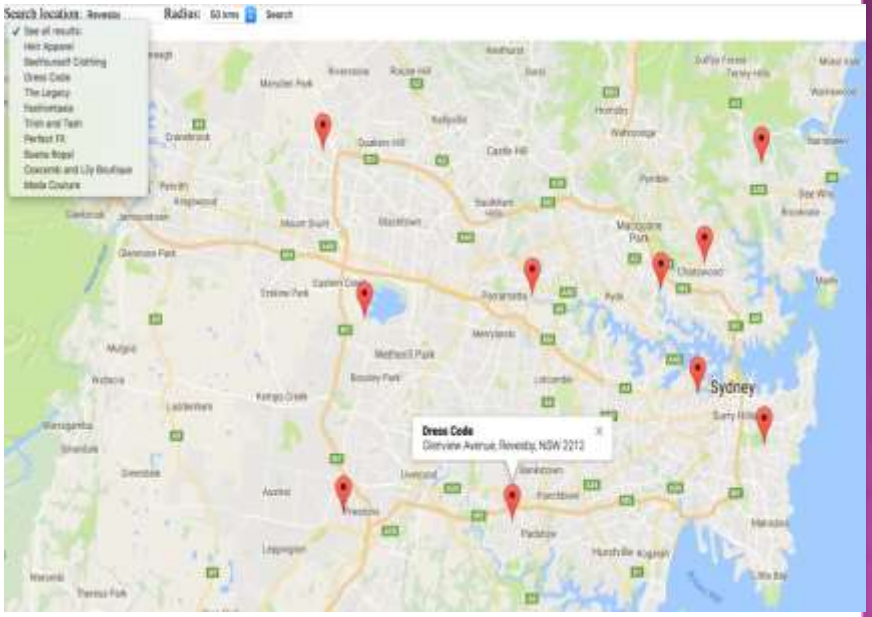
MAJOR SOURCES OF ABUNDANT DATA : BIOMETRICS

Used generally for identification and access control.



MAJOR SOURCES OF ABUNDANT DATA : GPS DATA





WHAT IS DATA MINING?

- Informal Definition: Data Mining is the extraction of strategic/actionable information from data.
- Formal Definition: Data Mining(Knowledge Discovery from Data - KDD) is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

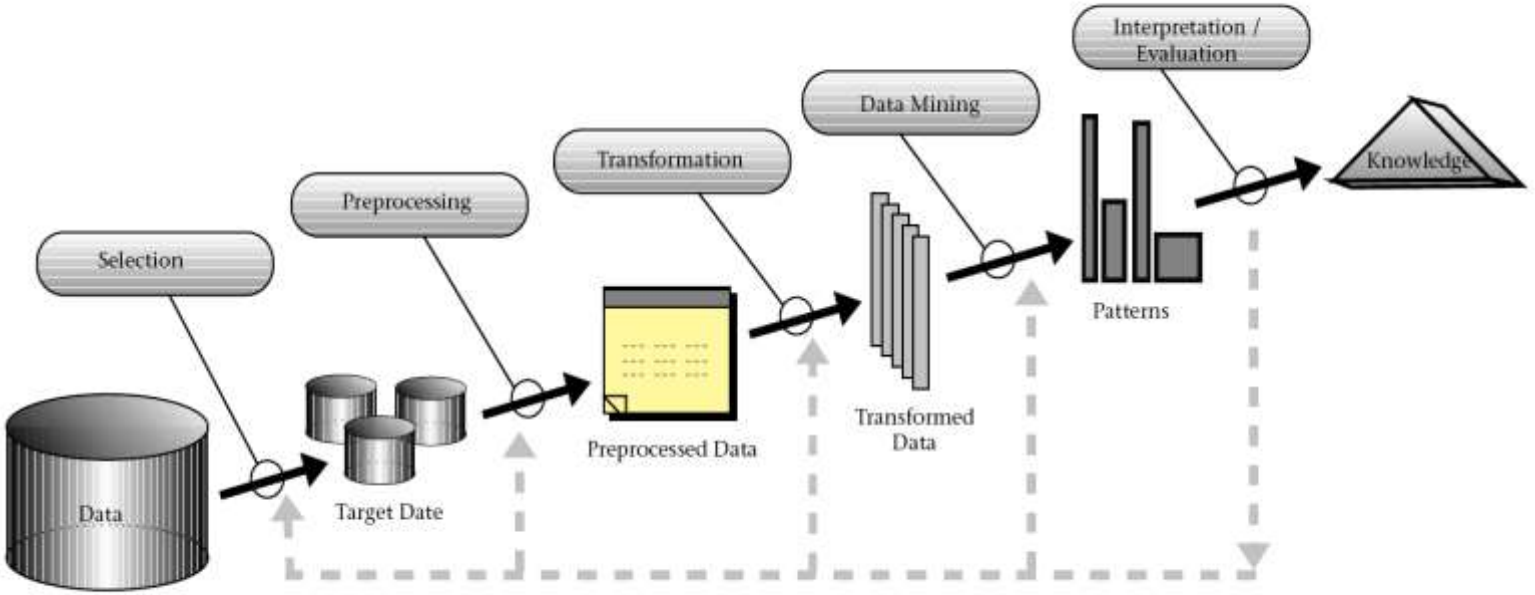
WHY NOT TRADITIONAL DATA ANALYSIS

- Tremendous amount of data
  - Algorithms must be highly scalable to handle tera-bytes/peta-bytes of data
- High-dimensionality of data
  - Data may have tens of thousands of dimensions
- High Complexity of data
  - Spatial, multimedia, text, Web data
  - Graphs, social networks, multi-linked data
  - Temporal, sequence data
  - Data streams and Sensor data

PROCESS OF KNOWLEDGE DISCOVERY FROM DATA

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

PROCESS OF KNOWLEDGE DISCOVERY FROM DATA



SOME OF THE DATA MINING METHODS…

- Class/Concept Description: Characterization and Discrimination
- Mining Frequent Patterns, Associations and Correlations
- Classification and Regression
- Cluster Analysis
- Outlier Analysis