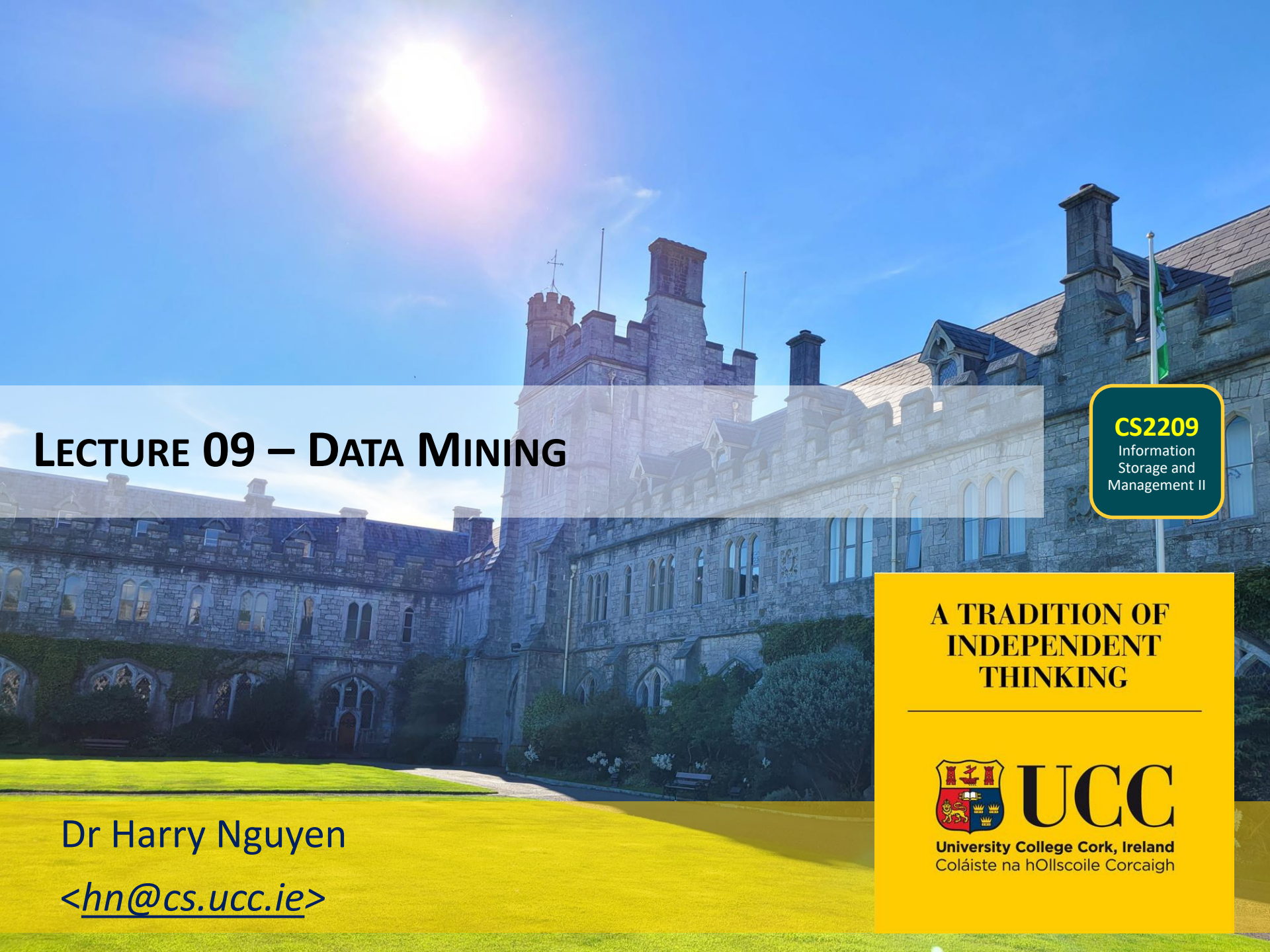


LECTURE 09 – DATA MINING

CS2209
Information Storage and Management II

A TRADITION OF INDEPENDENT THINKING

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Dr Harry Nguyen
<hn@cs.ucc.ie>

What We Saw Last Lecture

Big Data Characteristics

Data Analytics

Data Analysis Examples

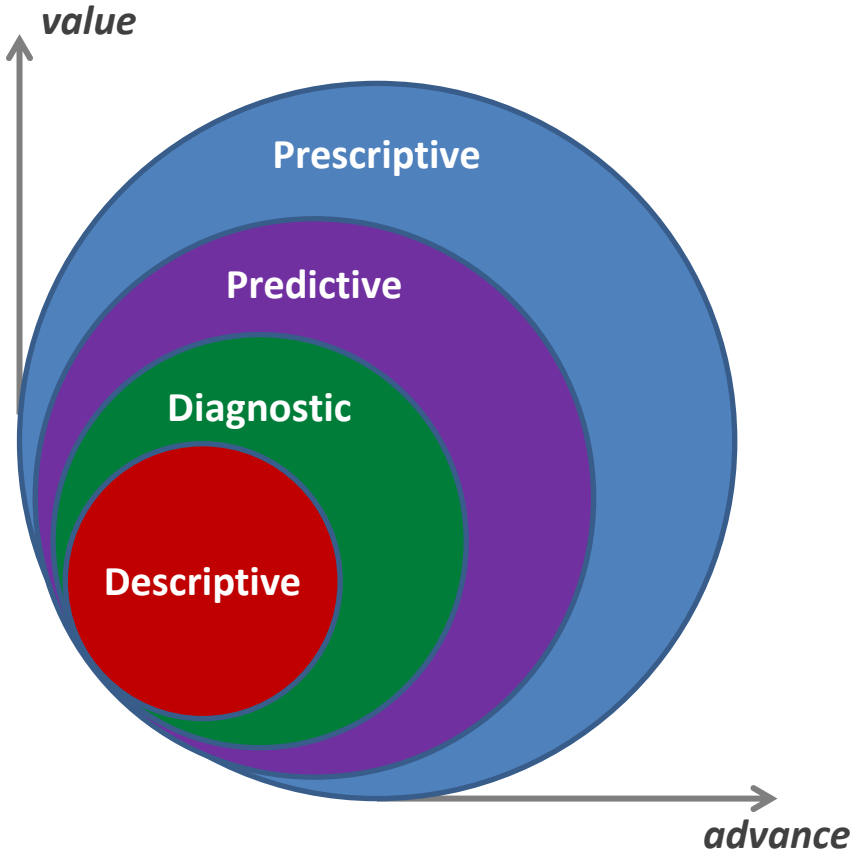
Big Data & 6 Vs

Big Data is data that contains greater variety, arriving in increasing volumes, and with ever-higher velocity

(Gartner, 2001)

Types of Data Analytics

CS2209 Information Storage and Management II



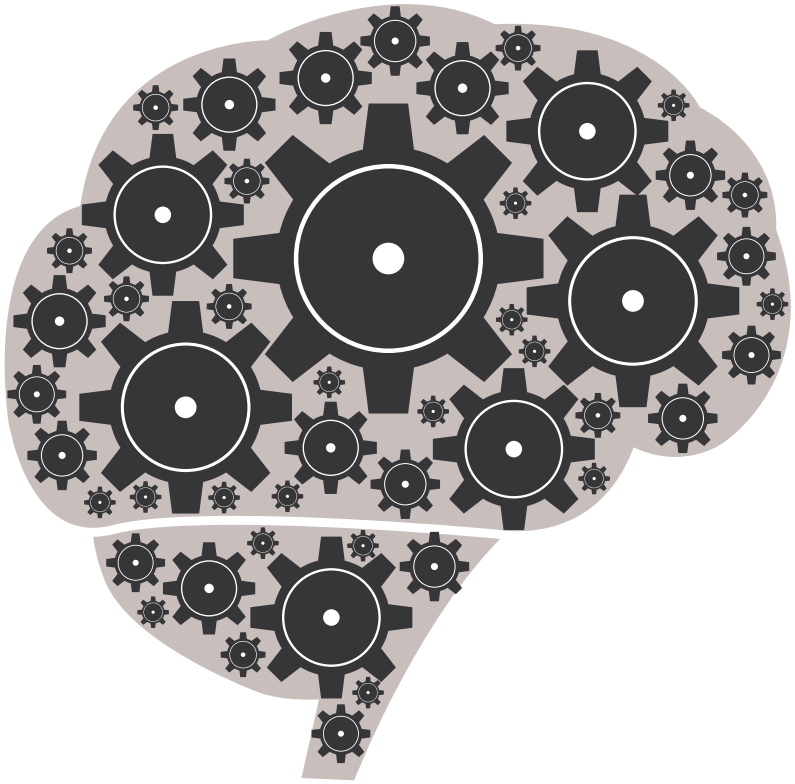
Descriptive Analytics
– What happened?

Diagnostic Analytics
– Why did it happen?

Predictive Analytics
– What will happen?

Prescriptive Analytics
– How to make it happen?

Decision Models



- A model is an abstraction of representation of a real system, idea, or object.
- It captures the most important features.
- Decision Model is a model used to understand, analyze, or facilitate decision making.

Data Mining

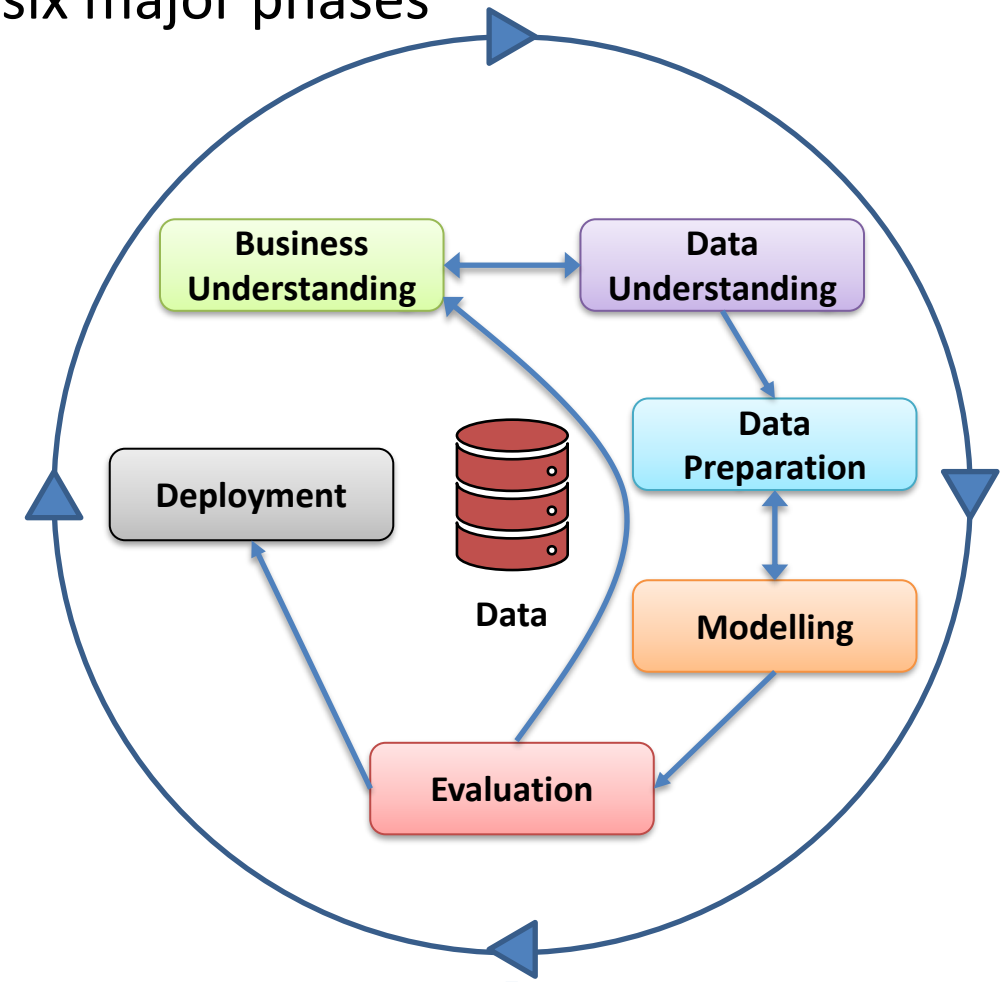
The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make critical business decisions.

This can be based on information retrieval or after further processing

For example: predictive analysis for anomaly detection in electric motors.

CRISP-DM

- Cross-industry standard process for data mining (CRISP-DM) consists of six major phases

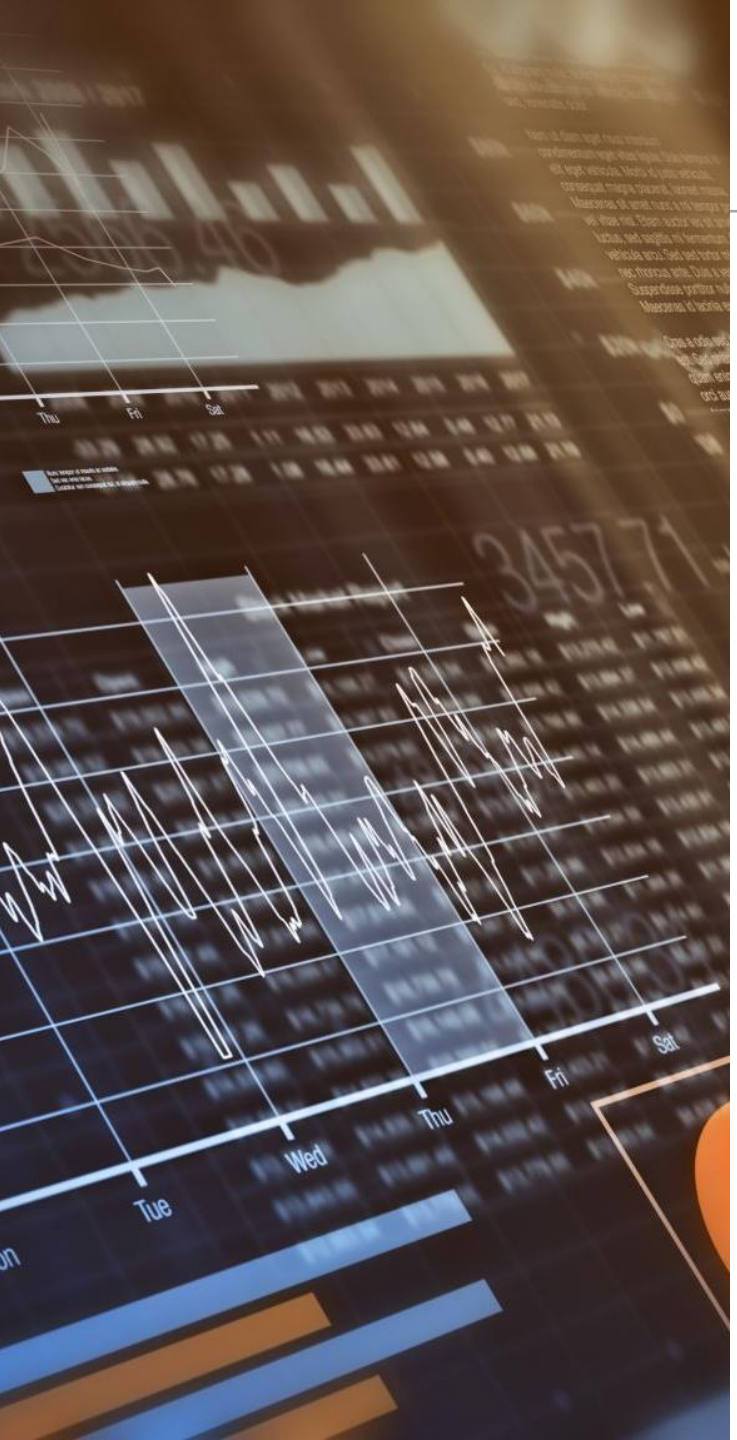


Business Understanding Phase

This initial phase focuses on:

- Understanding project objectives and requirements from a business perspective
- Translating this knowledge into a data mining problem definition and a preliminary plan

A decision model can be used



DATA UNDERSTANDING PHASE

- The data understanding phase starts with:
  - Initial data collection
  - Getting familiar with the data
  - Identifying data quality problems
  - Discovering preliminary insights into the data
  - Generating interesting subsets to form hypotheses for hidden information

Data Preparation Phase

- The data preparation phase covers:
  - Activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data
  - Data preparation tasks are likely to be performed multiple times
    - Table, record, and attribute selection
    - Transformation and cleaning of data

Modeling Phase

- Selection and use of various modeling techniques with the search for optimal parameters
- Some techniques have specific requirements on the data
- Often, stepping back to the data preparation phase

Evaluation Phase

High-quality model(s) from a data analysis perspective

Important to more thoroughly evaluate the model(s)

Reviewing steps executed to construct the model to achieve business objectives

Determining if there is some important business issue that has not been sufficiently considered

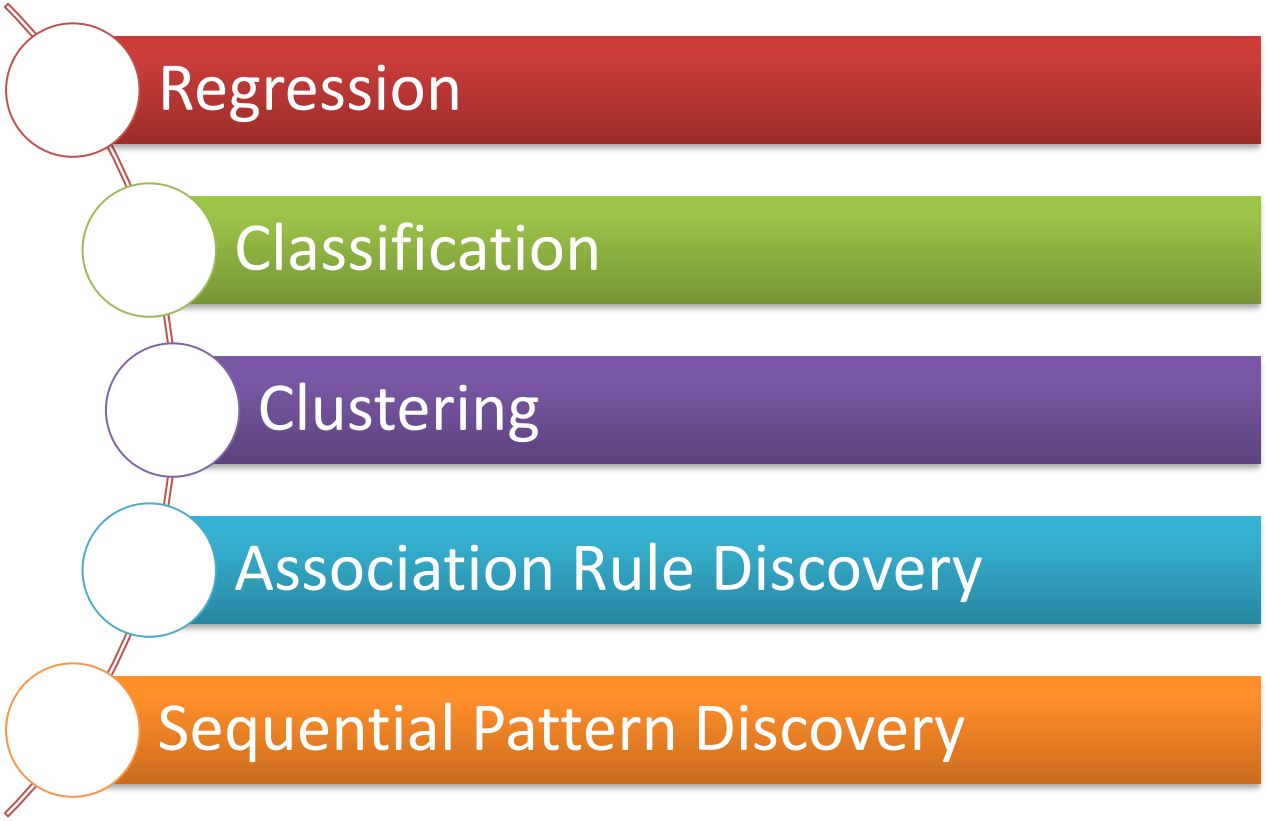
Acceptance and usefulness of data mining results

Deployment Phase

CS2209
Information Storage and Management II

- Knowledge gained will need to be organized and presented in a way that is useful to the customer
- Generating a report to implementing a repeatable data scoring/mining process
- Important for the customer to understand up front the actions needed to actually make use of the created models

Data Mining Tasks



REGRESSION

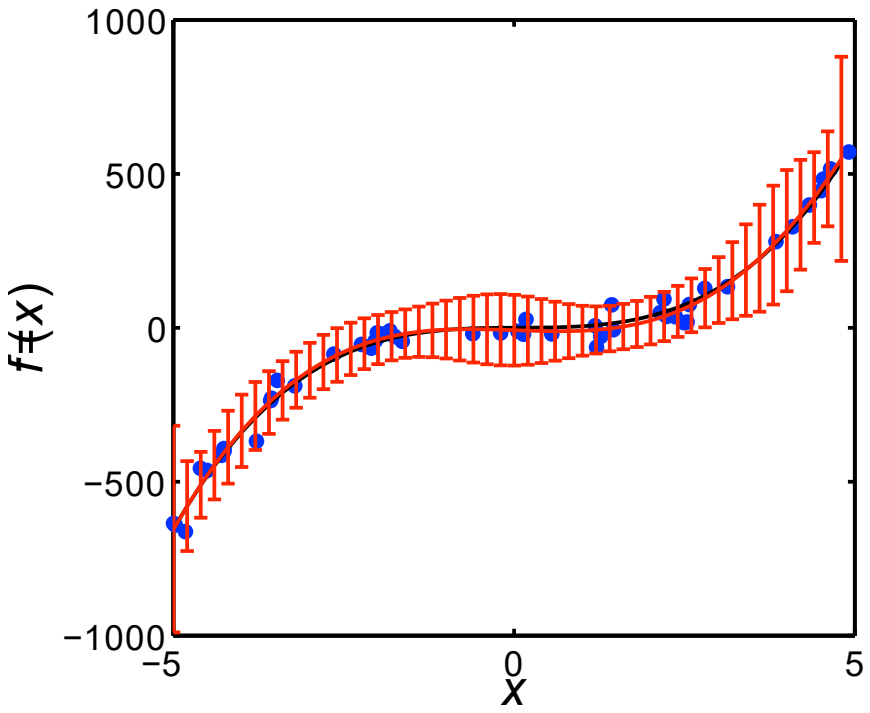
CS2209
Information Storage and Management II

Regression

- Learning a continuous function from a set of examples

Example (s)

- Predicting stock prices (x might be time or some other variable of interest)



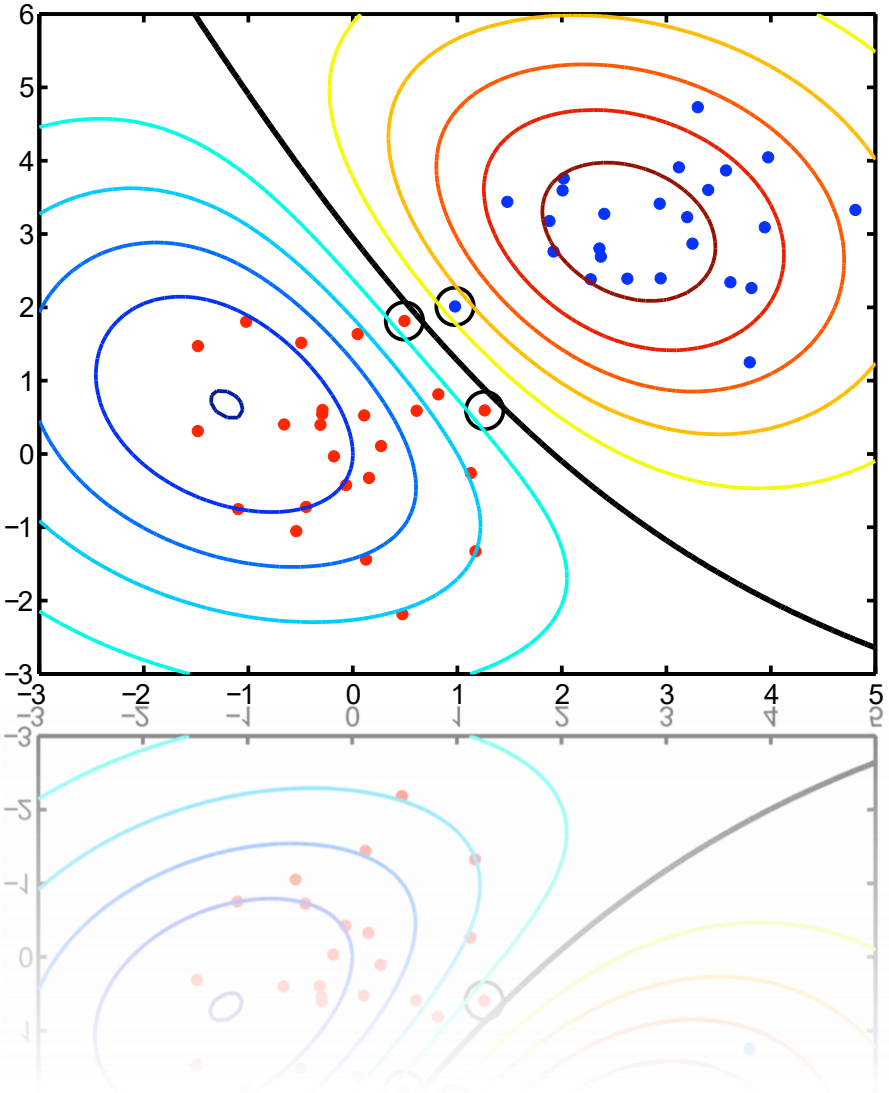
CLASSIFICATION

Classification

- Learning rules that can separate objects of different types from one another

Example(s)

- Disease diagnosis
- Spam email detection



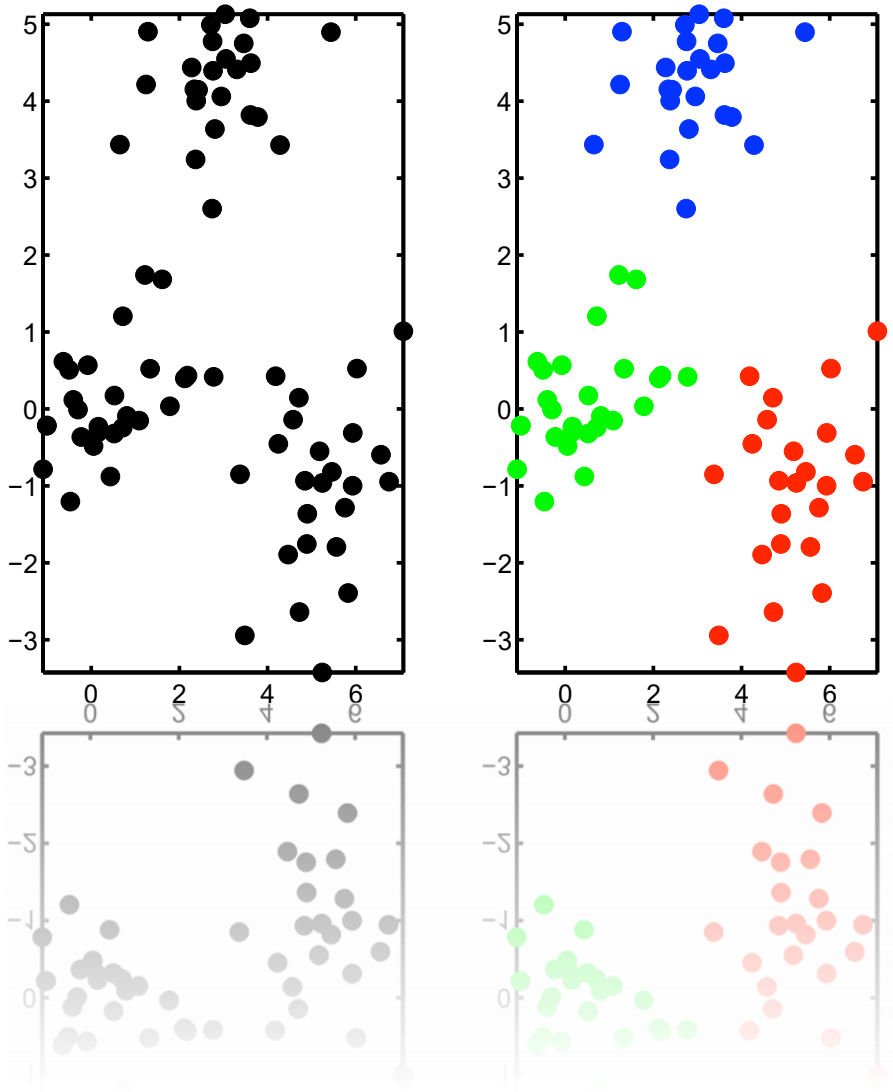
CLUSTERING

Clustering

- Finding groups of similar objects

Example(s)

- People with similar "preferences"
- Genes with similar functions



Example: Views & Sales

Date	Category	Branch	Views	Sales
01/01/2023	Drinks	Cork	171	39
02/01/2023	Drinks	Cork	107	24
03/01/2023	Drinks	Cork	166	37
04/01/2023	Drinks	Cork	208	51
05/01/2023	Drinks	Cork	221	55
06/01/2023	Drinks	Cork	333	83
07/01/2023	Drinks	Cork	235	62
08/01/2023	Drinks	Cork	242	61
09/01/2023	Drinks	Cork	294	79
10/01/2023	Drinks	Cork	300	81
11/01/2023	Drinks	Cork	275	67
12/01/2023	Drinks	Cork	270	65
13/01/2023	Drinks	Cork	319	82
14/01/2023	Drinks	Cork	328	84
15/01/2023	Drinks	Cork	260	70
16/01/2023	Drinks	Cork	344	89
17/01/2023	Drinks	Cork	250	63
18/01/2023	Drinks	Cork	279	67

- Objectives:
  - Investigate social media effects on sales
- Data Mining Problem:
  - Regression
- Data:
  - Social media views (thousands)
  - Sales (thousands)

Social Medial Views & Sales

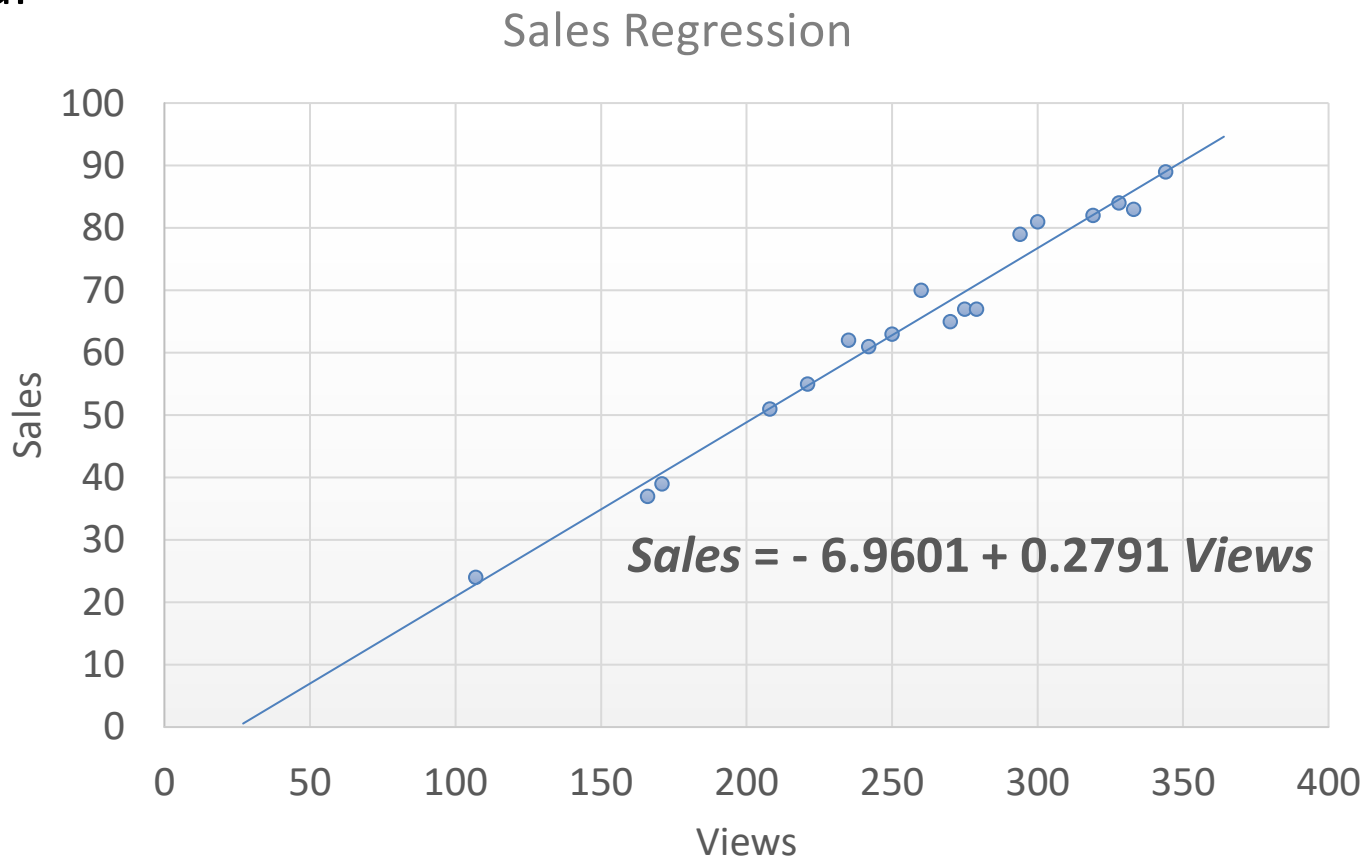
- Assume that there is a linear relationship between Sales and Social Media Views

$$Sales \approx \beta_0 + \beta_1 Views$$

  - Want to see how the Social Media Views affect Sales
  - Estimate β_0 and β_1?

DATA MINING RESULTS

Simple linear regression assumes that there is a single predictor variable X and the relationship between the response Y and X is linear



SUMMARY

- Data Mining Process
- CRISP-DM
- Various Data Mining Techniques