Campus Irapuato-Salamanca | División de Ingenierías

Data Mining

Topic 1: Data mining and its general process

© blog.tail.digital

August – December 2021

Juan Carlos Gómez Carranza
PhD in Computer Science
jc.gomez@ugto.mx
http://jcgcarranza.wix.com/juancarlosgomez
Office 314

Data Mining

Set of methods to process and analyze data collections with the purpose of discovering **patterns, associations** or interesting findings to **understand** and **explain** such **data**

©creativemarket.com

Pattern

Discernible **regularity** in the world or in a manmade design. The elements of a pattern **repeat** in a **predictable** manner (this makes it useful)

Goal of data mining

© datamensional.com

- The main goal is to **extract information** (patterns, summaries or findings) from a collection of data and **transform** it into an **understandable structure** for a posterior use → **Take decisions (automatic or by humans)**

Data mining related concepts

© datamensional.com

- Data analytics
- Data science
- Knowledge discovery (KDD)
- Knowledge extraction
- Business intelligence
- “Big data”

Data mining examples

Clusters

Find similar users in social media (age, gender, political affiliation, interests, etc.)

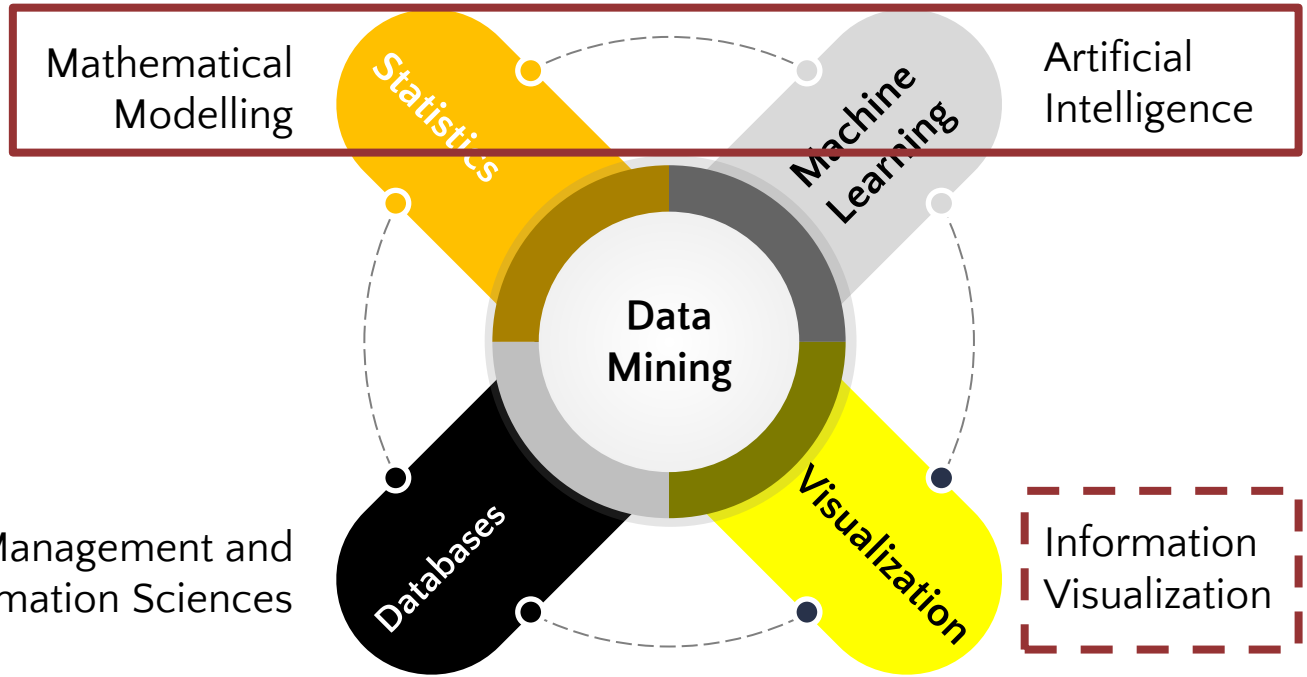
Anomalies

Fraud detection in bank transactions

Rules

Find behavioral patterns in e-comerce

Data mining: confluence of multiple disciplines



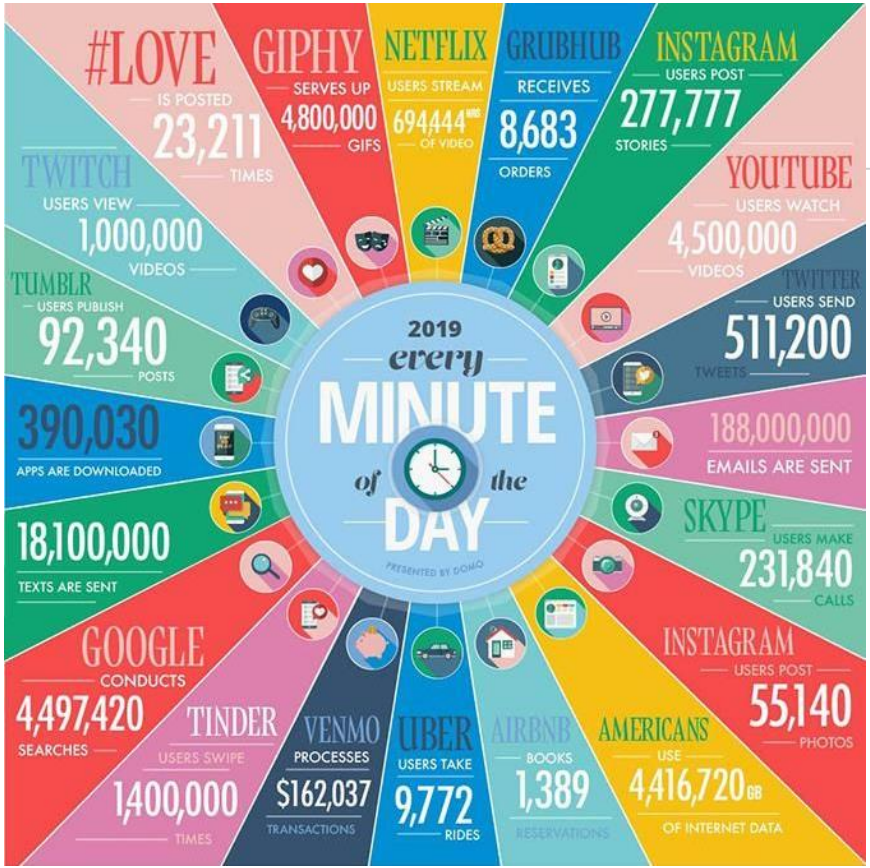
Why Data Mining?

We are living in the era of **Big Data**, where **great amount** of data are daily created, distributed and shared

How big is Big Data?

> **2000TB** of data are created in internet per minute.

~ **3PB** of data are in the database of Google Earth



© domo.com

Datafication

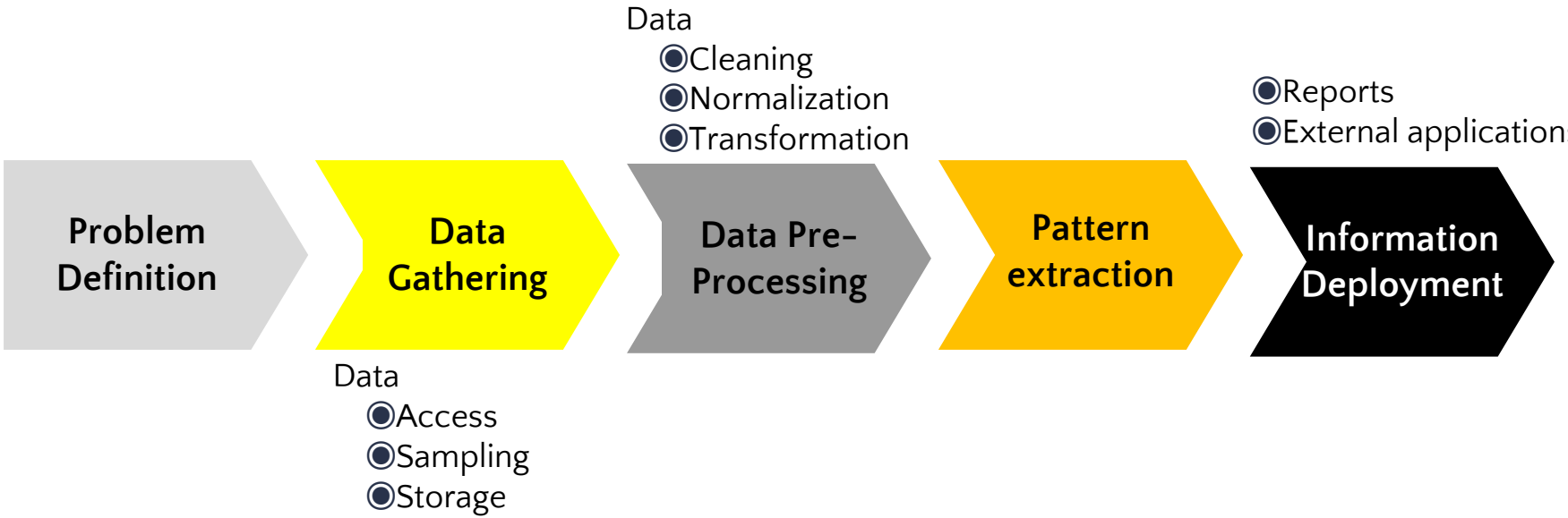
Technological trend **turning** many aspects of our **life** into **data**.

- Mobile devices (phones, watches, tablets)
- Social media
- E-commerce
- Smart homes
- Smart cities
- Smart cars

© promptcloud.com

By 2025 - **2 ZB** of data generated by the Internet of Things

General process of Data Mining



Methods in data minig

Descriptive

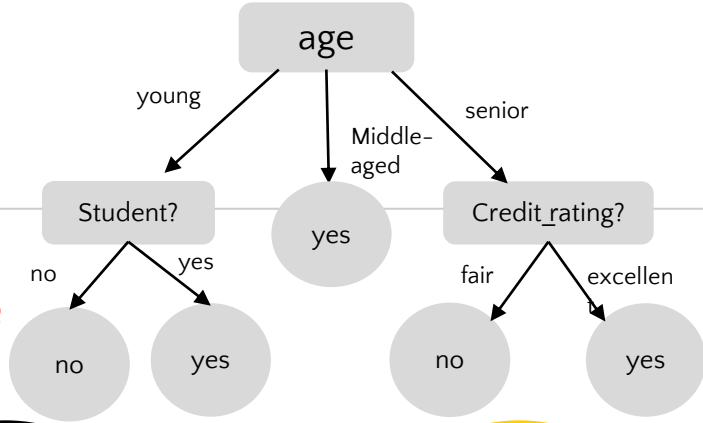
Statistical modeling

Summarize the data in a few relevant features

Clusters

Group data points by similarities

Predictive



Association and hierarchical rules

Predict a variable based on common associations or hierarchical rules

{spoon, dishes} → fork
{cereal} → beer?

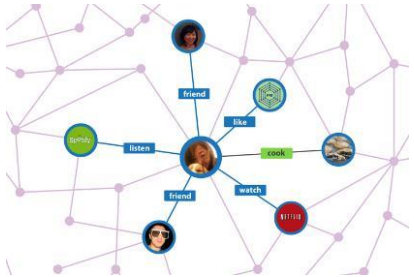
Functions

Predict a variable based on a function

Example of application

Problem Definition

Find tendencies an associations in the data from DICIS students



- Max $\max\{x_i,..n\}$
- Min $\min\{x_i,..n\}$
- Sum $\sum_{i=1}^{n} x_i$
- Mode $\text{mod}\{x_i,..n\}$
- Median $median\{x_i,..n\}$
- Mean $\frac{1}{n}*\sum_{i=1}^{n} x_i$
- Variance $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\mu})^2$
- STD $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$
- Unique $unique\{x_i,..n\}$
- Entropy $P(G_i)=\left(\frac{cell\ count(G_i)}{n}\right)$ $\sum_{i=1}^{k} P(G_i)(-\ln(P(G_i)))$
- Probability $\sum_{i=1}^{k}(P(G_i))^2$

Activity 1

Data Gathering

Data
- Access
- Sampling
- Storage

Individual work: **Ask** to 10 friends for the following information (not common friends and include yourself): *complete name, age, sex (h/m), height (m), weight (kg), semester, # of courses taken up to now (total), pet (dog/cat/other), city of origin, has a personal video games console (y/n)*

- Access → Personal/private
- Sampling → Friend
- Storage → File

End topic 1

Next topics

- Data/information modalities
- Analysis of structured data
- Analysis of text data

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival