Introduction To Data Mining

# MoTivation About Data Mining

• Vast amounts of data are collected daily

• knowledge mining from data

• Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

العالم ملئ بالبيانات اللى بتتجمع كل يوم,, البيانات محملة بالمعرفة, بس محتاجين طريقة للبحث داخل البيانات دى عشان اقدر استخلص اللى يهمنى جوه البيانات وهو المعرفة اللى مستخبية أو مضمونة جوه الكم الهائل ده,, عشان كده العلم ده موجود و بيخدم جميع المجالات المختلفة

حاجتنا الى المعرفة فى المجالات هى المحرك لاستخدام هذا العلم

## Data Mining Terms and Names

Data Mining = Knowledge mining from data = Knowledge Extraction = Data/Pattern Analysis = Data Archaeology = Data Dredging

= Knowledge Discovery from Data (KDD)

## Knowledge Discovery Process

مجموعة من العمليات المتاتبعة اللى بنعملها عشان نستخرج المعرفة

* Data Cleaning: remove noise and inconsistent data
* Data Integration: multiple sources combined
* Data Selection: data relevant to analysis task => البيانات اللى مرغوب دراستها
* Data Transformation: احول البيانات لشكل أقدر أتعامل معاه داخل الالجوريزم
* Data Mining: اللى بدور عليهpattern استخدام الجوريزم معين عشان أطلع ال
* Pattern Evaluation: اقيم اد ايه هو مناسب لاتخاذ القرارات
* Knowledge Representation: أعرضه بشكل مفهوم للمستخدم

### What is Data Science? - MAKE ME ANALYSTNot Data Mining **Vs** Data Mining

أوعى تفكر ان حضرتك لو بحثت على النت على طبخة مثلا يبقى ده داتا مينينج ,, ابسلوتلى

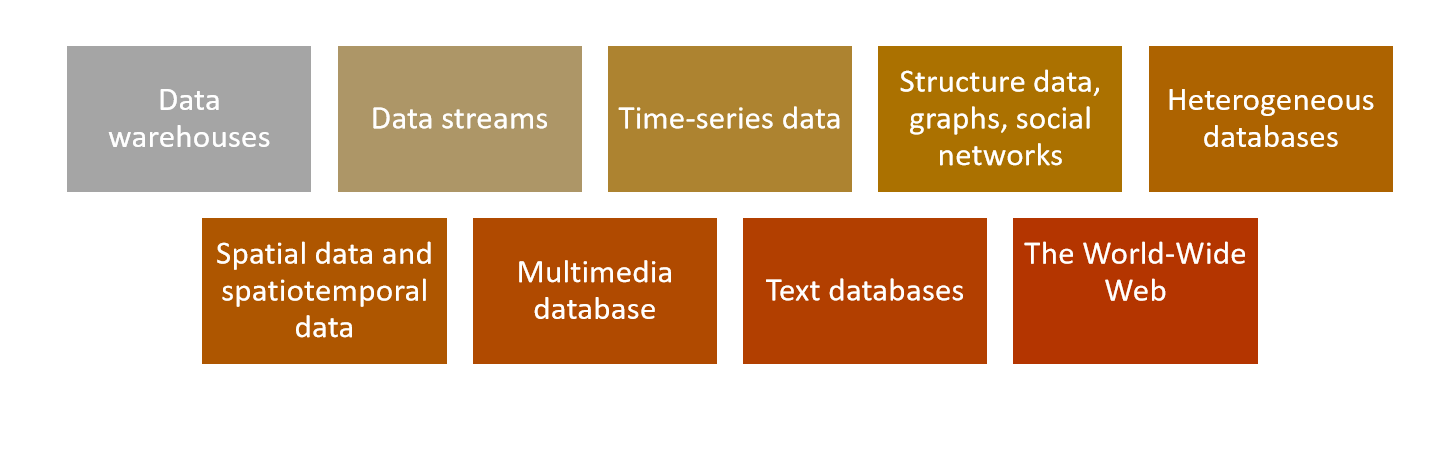


~~Searching for cooking on Google~~ vs. Grouping similar cuisines French, Italian, Arabian …

~~Looking up spa resorts~~ vs. More relevant Spa for curing certain diseases

Data Mining is NOT about ~~searching in a Data~~, but more about Implicit meaningful Information

# What Kinds of Data Can Be Mined?



* Warehouse: منطة لتجميع كمية كبيرة من قواعد البيانات
* Spatial => x,y,z , spatiotemporal = spatial + Time
* Heterogeneous => قواعد بيانات متنوعة ,, يعنى الداتا بتيجى من سكول و نو سكول و هكذا

عرفنا ايه اشكال البيانات اللى منكم نشتغل عليها بس لسه ماعرفناش ايه أنواع الباترن اللى منكم تطلعلنا من العمليات و ازاى منكم تفدنا ؟؟؟

|  |  |
| --- | --- |
| Patterns | How can it help |
| CHARACTERIZATION AND DISCRIMINATION | Business Intelligence and Analytics - From A to Z (Part 4)Characterization => slicing and dicing the data to understand what it is all about.  Discrimination => identifying splitting conditions to partition the data into independent bins. |
| ASSOCIATION / CORRELATION ANALYSIS | بشوف اذا كان فيه ترابط مابين الاعمدة وبعضها و ازاى بتأثر على بعضها يا سلب يا ايجاب يا اما  No Hypotheses => Independent Attributes |
| CLASSIFICATION | و ده نوع من الباترن اللى بيحتاج مودل يدرب البيانات عشان بعد كده يخمن لو جاله داتا بعد كده ياخدله القرار  Supervised Learning => Machine Learning Algorithms |
| CLUSTER ANALYSIS | ده مش محتاج داتا يتدرب عليها ده  Unsupervised learning |
| OUTLIER ANALYSIS | بيشوف ايه البيانات اللى معدية الحدود الطبيعية و منكم تكون مفيدة مثلا فى المجال الطبى للكشف عن الأمراض |
|  |  |

# Terminology

|  |  |
| --- | --- |
|  |  |
| Mining Tasks | **Descriptive Tasks**   * Characterize properties of the data in a target data set. * e.g., (classification, regression, anomalies/outliers detection) * **بتستخدم مجموعة من المتغيرات عشان تتنبأ بالمتغيرات الأخرى**   **Predictive Tasks**   * Perform induction on the current data to make predictions. * (e.g., clustering, association rule discovery, sequential pattern discovery) * **الهدف منه انه يطلعلى باترن يقدر يفهمه المستخدم زى لو جالك بيانات شكلها 1و2و3 خدلها القرار ده** |
| Technologies |  |
| Applications | هيفدنى فى ايه الداتا مينينج ؟ |
| Data Objects | * Data sets are made up of data objects. * Also referred as samples, examples, instances, data points. * e.g. customers, students, patients, books. * Data objects are typically described by attributes   Entity of A DB like Table, Json file, CSV |
| Attributes | * A data field, representing a characteristic or feature of a data object. * attribute, dimension, feature, and variable are often used interchangeably * A customer object can include, for example, customer ID, name, and address. * Observed values for a given attribute are known as observations. |
|  | **Nominal Attributes {B , A , C}**   * The values of a nominal attribute are symbols or names of things * do not have any meaningful order * e.g. hair color, marital status, occupation   **Binary Attributes {A , C}**   * a nominal attribute with only two categories or states: 0 or 1 * symmetric if both of its states carry the same weight (e.g. gender) * Asymmetric like HIV result of medical test, The +ve and -ve values are not in the same weight * هى هياها النومينال بس ماليهاش الا قيمتين بس   **Ordinal Attributes (A, B, C)**   * an attribute with possible values that have a meaningful order or ranking among them * e.g. professional rank, grade, customer satisfaction * Central tendency can be the mode or the median **But the Mean Cannot be Defined (Why?)** |
|  | **Interval-Scaled**   * measured on a scale of equal-size units. * does not have a true zero-point. * e.g. temperature, neither 0C nor 0F indicates “no temperature.”   **Ratio-Scaled**   * a numeric attribute with an inherent zero-point * e.g. years of experience |
| Statistical Descriptions of Data | For data **preprocessing** to be successful, it is **essential** to have an **overall picture** of your data.   * Measuring Central Tendency => Mean , Median , Mode, MidRange * Measuring the Dispersion of Data => Ranges, Quartiles, five number summaries, SD, Variance * Basic Graphs => Histogram, Scatter Plots |
| Measuring Central Tendency | **Mean**   * Simple Mean =      * Weighted mean = * A major problem with the mean is its **sensitivity to extreme** (e.g., outlier) values. * we can instead use the trimmed mean, which is the mean obtained after **chopping off** values at the high and low extremes.   **Median**   * the middle value in a set of ordered data values. * It is the value that separates the higher half of a data set from the lower half. * The median is **expensive to compute** when we have many observations   **Mode & Mid Range**   * the value that occurs most frequently in the set * it can be determined for qualitative and quantitative attributes. * Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. * Mid Range: is the average of the largest and smallest values in the set. |
| Measuring the Dispersion of Data | The dispersion or spread of numeric data is useful in identifying **outliers**.  **Variance and Standard Deviation**   * **low** standard deviation means that the data observations tend to be **very close** to the **mean**, * **high** standard deviation indicates that the data are **spread out** over a large range of values   **Range**   * the difference between the largest (max()) and smallest (min()) values   **Quantiles**   * are points taken at regular intervals of a data distribution, dividing it into equal size consecutive sets. * For example ,3 quantiles shown to the Right. * The distance between the first and third quartiles is **interquartile range (IQR)**   **Five-Number Summary, Boxplots, and Outliers**   * a standardized way of displaying the distribution of data based on a five numbers summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). |

The Practical about this lecture: <https://github.com/AhmedKhalil777/DataScience.Learning/blob/master/Resources/Lectures_Contripution/Lecture1.contribution.ipynb>