Data Mining

Lecture 6

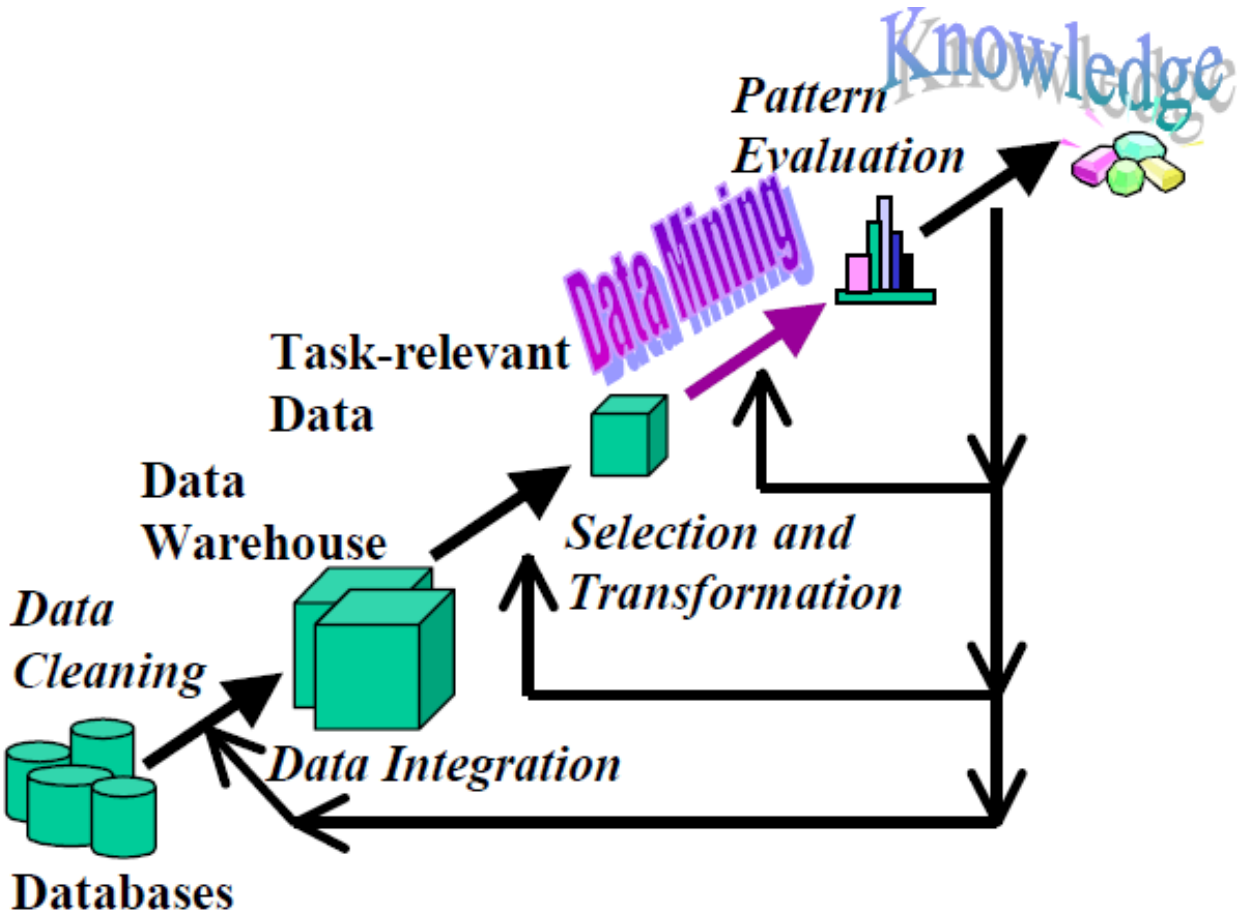
What is Data Mining?

المعروف أيضاً باسم اكتشاف المعرفة في قواعد البيانات

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

يشير إلى الاستخراج غير البديهي للمعلومات الضمنية وغير المعروفة سابقًا والتي قد تكون مفيدة من البيانات الموجودة في قواعد البيانات

The following figure shows data mining as a step in an iterative knowledge discovery process:


Knowledge
Pattern
Evaluation
Data Mining
Task-relevant
Data
Data
Warehouse
Selection and
Transformation
Data
Cleaning
Data Integration
Databases

تتكون عملية اكتشاف المعرفة في قواعد البيانات من بضع خطوات تؤدي من مجموعات البيانات الأولية إلى شكل من أشكال المعرفة الجديدة. تتكون العملية التكرارية من الخطوات التالية

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
  إنها مرحلة يتم فيها إزالة بيانات الضوضاء والبيانات غير ذات الصلة من المجموعة.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
  في هذه المرحلة ، قد يتم دمج مصادربيانات متعددة ، وغالباً ما تكون غير متجانسة ، في مصدرمشترك
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
  في هذه الخطوة ، يتم تحديد البيانات ذات الصلة بالتحليل واسترجاعها من جمع البيانات

تعُرف أيضاً باسم توحيد البيانات ، وهي
مرحلة يتم فيها تحويل البيانات المحددة إلى أشكال مناسبة
لإجراء التعدين

- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

إنها الخطوة الحاسمة التي يتم فيها تطبيق
تقنيات ذكية لاستخراج أنماط يحتمل أن تكون مفيدة.

- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

تقييم النمط

في هذه الخطوة ، يتم تحديد أنماط مثيرة
للاهتمام بشكل صارم تمثل المعرفة بناءً على مقاييس معينة

- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

هي المرحلة الأخيرة التي يتم فيها تمثيل المعرفة المكتشفة بصرياً للمستخدم. تستخدم هذه
الخطوةالأساسية تقنيات التصور لمساعدة المستخدمين على فهم وتفسير نتائج التنقيب عن البيانات

من الشائع الجمع بين بعض هذه الخطوات معا

It is common to combine some of these steps together:

- For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse.

يمكن إجراؤها معاً كمرحلة ما قبل المعالجة لإنشاء مستودع بيانات

- *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

يمكن أيضاً دمجها حيث يكون توحيد البيانات نتيجة للاختيار ، أو ، كما هو الحال بالنسبة لمستودعات البيانات ، يتم الاختيار على البيانات المحولة

مانوع البيانات التي يمكن استخراجها

What kind of Data can be mined?

من حيث المبدأ ، لا يقتصر استخراج البيانات على نوع واحد من الوسائط أو البيانات

- In principle, data mining is not specific to one type of media or data.

يجب أن يكون التنقيب عن البيانات قابلاً للتطبيق على أي نوع من مستودعات المعلومات

- Data mining should be applicable to any kind of information repository.
- However, algorithms and approaches may differ when applied to different types of data.

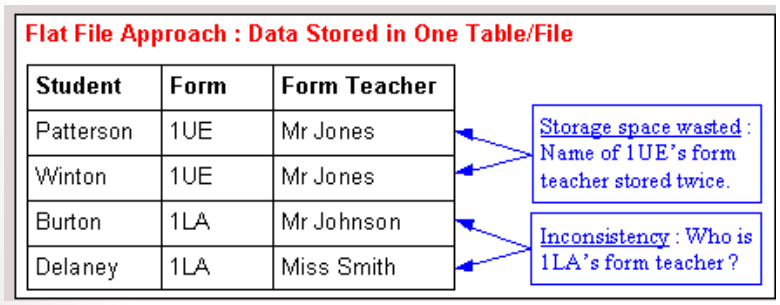
ومع ذلك ، قد تختلف الخوارزميات والنهج عند تطبيقها على أنواع مختلفة من البيانات

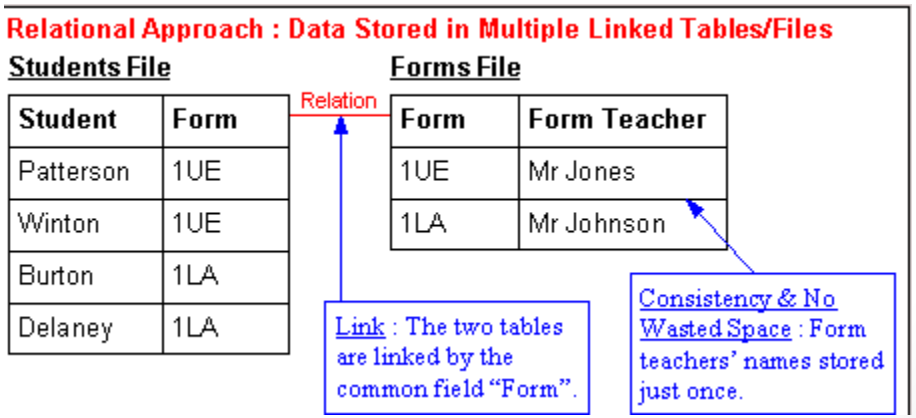
Data mining is being put into use and studied for databases, including:

يتم استخدام التنقيب عن البيانات ودراستها لقواعد البيانات ، بمافي ذلك:

، ملفات مسطحة: الملفات المسطحة هي في الواقع أكثر مصادر البيانات شيوعاً لخوارزميات التنقيب عن البيانات خاصة على مستوى البحث.

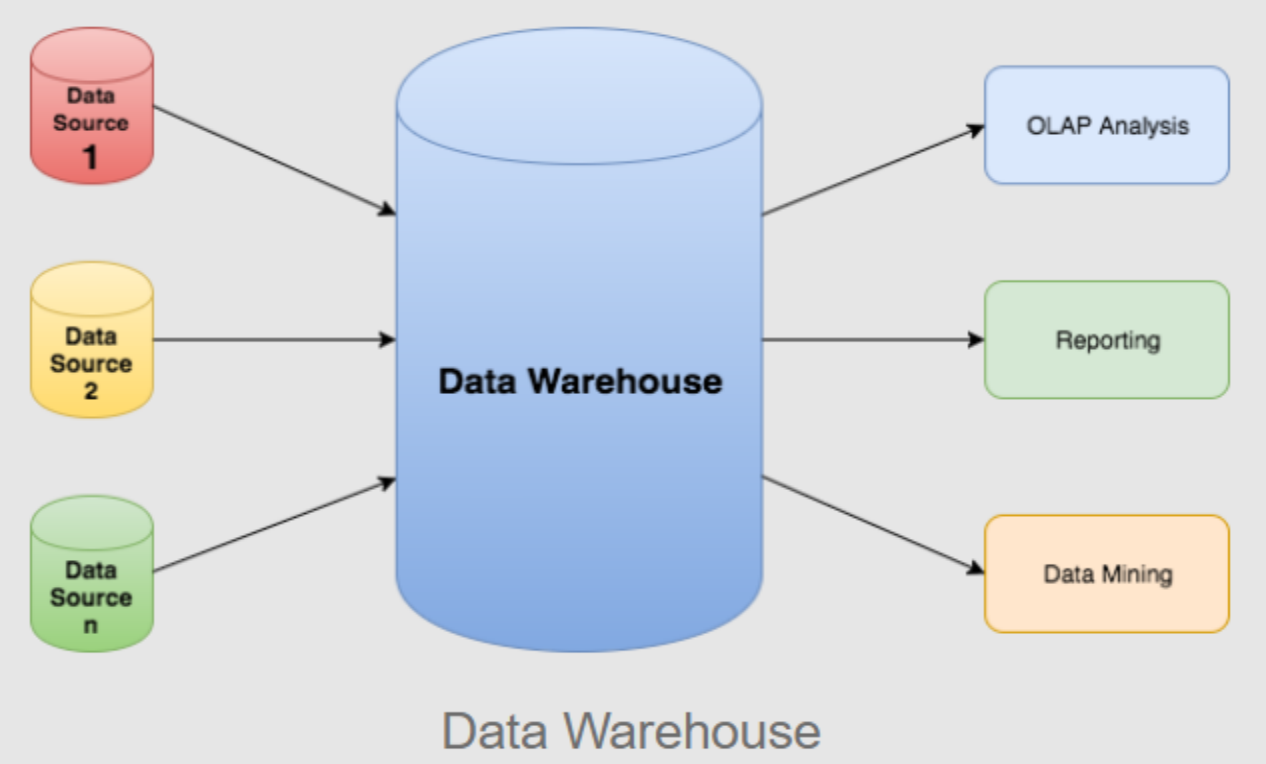
- **Flat files**: Flat files are actually the most common data source for data mining algorithms, especially at the research level. باختصار ، تتكون قاعدة البيانات العلائقية من مجموعة من الجداول التي تحتوي إما على قيم سمات الكيان أو قيم السمات من علاقات الكيانات
- **Relational Databases**: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships.





) مستودعات البيانات: مستودع البيانات كمخزن ، هو مستودع للبيانات التي تم جمعها من مصادر بيانات متعددة غالباً ما تكون غيرمتجانسة(ويقصد استخدامها ككل ضمن نفس المخطط الموحد

- **Data Warehouses**: A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema.



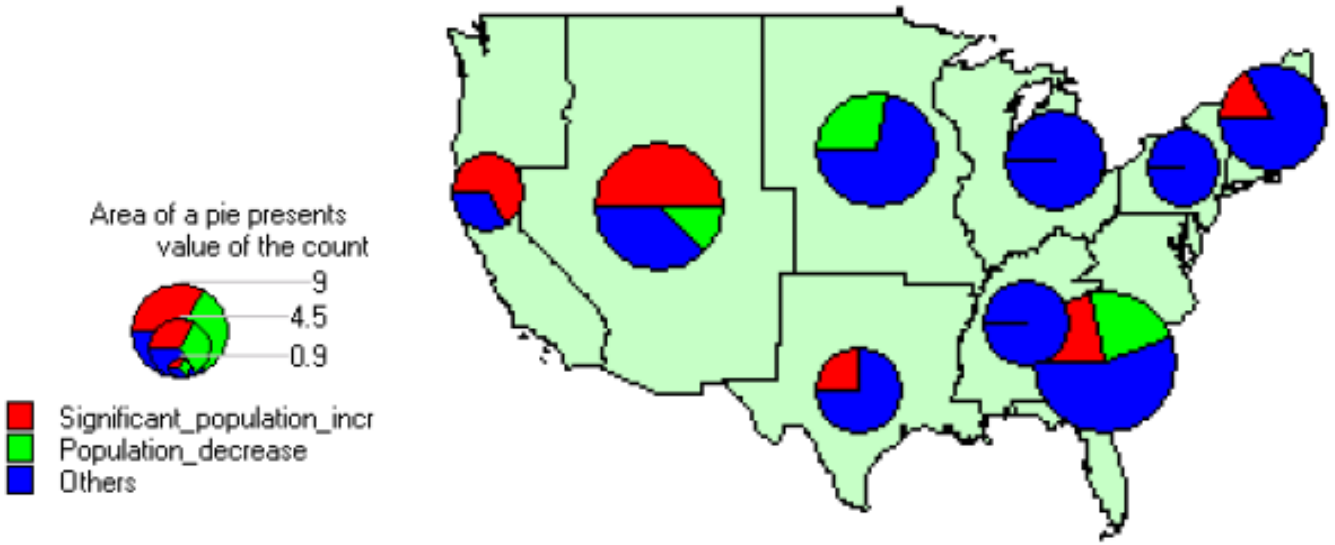
قواعد بيانات المعاملات: قاعدة بيانات المعاملات عبارة عن مجموعة من السجلات التي تمثل المعاملات ، ولكل منها طابع زمني ومعرف ومجموعة من العناصر

- **Transaction Databases**: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items.
- **Multimedia Databases**: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system.

قواعد بيانات الوسائط المتعددة تشمل الفيديو والصور والصوت والوسائط النصية. يمكن تخزينها في قواعد بيانات كائنية أو كائنية موسعة ، أو ببساطة على نظام ملفات

قواعد البيانات المكانية: قواعد البيانات المكانية هي قواعد بيانات تقوم ،بالإضافة إلى البيانات المعتادة ، بتخزين المعلومات الجغرافية مثل الخرائط وتحديد المواقع العالمية أو الإقليمية

- **Spatial Databases**: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning.



قواعد بيانات السلاسل الزمنية: تحتوي قواعد بيانات السلاسل الزمنية على بيانات مرتبطة بالوقت مثل بيانات سوق الأوراق المالية أو الأنشطة المسجلة

- **Time-Series Databases**: Time-series databases contain time related data such stock market data or logged activities.
- **World Wide Web**: The World Wide Web is the most heterogeneous and dynamic repository available. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications.

شبكة الويب العالمية هي المستودع المتاح الأكثرتنوعاً وديناميكية. يتم تنظيم البيانات الموجودة في شبكة الويب العالمية في مستندات مترابطة. يمكن أن تكون هذه المستندات نصية وصوتية وفيديو وبيانات أولية وحتى تطبيقات

- **التوصيف**: توصيف البيانات هو تلخيص للسمات العامة للكائنات في فئة مستهدفة ، وينتج ما يسمى *القواعد المميزة*. **على سبيل المثال**، قد يرغب المرء في تمييز عملاء OurVideoStore الذين يستأجرون بانتظام أكثر من 30 فيلماً سنوياً.

ماالذي يمكن اكتشافه

What can be discovered?

يتم عرض وظائف التنقيب عن البيانات والمعرفة المتنوعة التي يكتشفونها بإيجاز في القائمة التالية

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- **Characterization**: Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules*. **For example**, one may want to characterize the OurVideoStore customers who regularly rent more than 30 movies a year.
- **Discrimination**: Data discrimination produces what are called *discriminant rules* and is basically the comparison of the general features of objects between two classes referred to as the *target class* and the *contrasting class*. **For example**, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5.

تمييز: تمييز البيانات ينتج ما يسمى قواعد تمييزية وهي في الأساس مقارنة الميزات العامة للكائنات بين فئتين يشار إليهما باسم الفئةالمستهدفة و الفئة متناقضة.على سبيل المثال، قد يرغب المرء في مقارنة الخصائص العامة للعملاء الذين استأجروا أكثر من 30فيلماً في العام الماضي مع أولئك الذين يقل حساب تأجيرهم عن5 .

تحليل الرابطة: تحليل الارتباط هو اكتشاف ما يسمى عادة *قواعد الرابطة*. يدرس تواتر العناصر التي تحدث معاً في قواعد بيانات المعاملات ،وعلى أساس عتبة تسمى *يدعم* يحدد مجموعات العناصر المتكررة. **على سبيل المثال**، قد يكون من المفيد لمدير OurVideoStore معرفة الأفلام التي يتم تأجيرها معاً غالباً أو ما إذا كانت هناك علاقة بين استئجار نوع معين من الأفلام وشراء الفشار أو البوب.

- **Association analysis**: Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets**. For example**, it could be useful for the OurVideoStore manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop.
- **Classification**: Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection.

- **تصنيف**: تحليل التصنيف هو تنظيم البيانات في فئات معينة. يعُرف أيضاً باسم *التصنيف الخاضع للإشراف*، يستخدم التصنيف تسميات فئة معينة لترتيب الكائنات في جمع البيانات.

تجمع: مشابه التصنيف ، التجميع هو تنظيم البيانات في الفئات.ومع ذلك ، على عكس التصنيف ، في التجميع ، تسميات الفئات غير معروفة والأمر متروك لخوارزمية التجميع لاكتشاف الفئات المقبولة

تنبؤ: لقد اجتذب التنبؤ اهتماماً كبيراً نظراً للآثار المحتملة للتنبؤ الناجح في سياق الأعمال

- **Prediction**: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context.
- **Clustering**: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.
- **Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as *exceptions* or *surprises*, they are often very important to identify.

تحليل الخارجة: القيم المتطرفة هي عناصر بيانات لا يمكن تجميعها في فئة أو مجموعة معينة. يعُرف أيضاً باسم *استثناءات* أو *مفاجآت*، غالباً ما تكون مهمة جداً لتحديدها.

- **تحليلالتطور والانحراف**: يتعلق تحليل التطور والانحراف بدراسة البياناتالمتعلقة بالوقت والتي تتغير بمرور الوقت. نماذج تحليل التطورالاتجاهات التطورية في البيانات ، والتي توافق على توصيف أو مقارنةأو تصنيف أو تجميع البيانات المتعلقة بالوقت.

  منناحية أخرى ، يأخذ تحليل الانحراف في الاعتبار الاختلافات بين القيمالمقاسة والقيم المتوقعة ، ويحاول العثور على سبب الانحرافات عنالقيم المتوقعة.

- **Evolution and deviation analysis**: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data.

  Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction.

من الشائع ألا يكون لدى المستخدمين فكرة واضحة عن نوع الأنماط التي يمكنهم اكتشافها أو التي يحتاجون إلى اكتشافها من البيانات المتوفرة.لذلك من المهم أن يكون لديك نظام متعدد الاستخدامات وشامل للتنقيب عن البيانات يسمح باكتشاف أنواع مختلفة من المعرفة وعلى مستويات مختلفة من التجريد