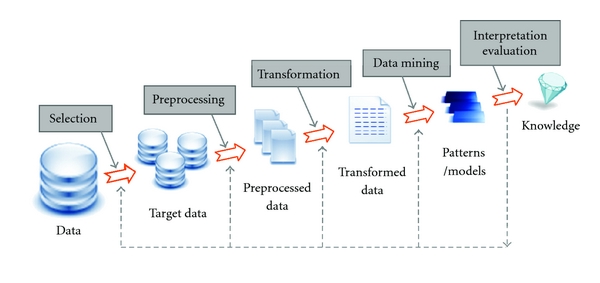
# Data Science Tutorial

# Data Mining: Concepts and Techniques

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.



فرآیند داده‌کاوی شامل چندین گام است. این فرآیند از داده‌های خام آغاز می‌شود و تا شکل‌دهی دانش جدید ادامه دارد. فرآیند بازگشتی داده‌کاوی شامل گام‌های زیر است:

* پاک‌سازی داده (Data Cleaning)
* یکپارچه‌سازی داده (Data Integration)
* انتخاب داده (Data Selection)
* تبدیل داده (Data Transformation)
* کاوش داده (Data Mining)
* ارزیابی الگو (Pattern Evaluation)
* ارائه دانش (Knowledge Representation)

پاک‌سازی داده

در این فاز «نویز» (نوفه) از مجموعه حذف و تدابیری برای داده‌های ناموجود (Missing Values) اندیشیده می‌شود.

یکپارچه‌سازی داده‌ها

در اغلب مسائل داده‌کاوی، داده‌ها از منابع داده گوناگون باید به یکباره مورد تحلیل قرار بگیرند. مثال خوبی از این مورد پایگاه داده‌های شعب مختلف یک فروشگاه زنجیره‌ای در شهرها و کشورهای گوناگون جهان است. برای تحلیل این داده‌ها باید آن‌ها را به صورت یکپارچه در یک انبار داده (Data Warehouse) گردآوری کرد، این کار در فاز یکپارچه‌سازی انجام می‌شود.

انتخاب داده

در فاز انتخاب داده، باید داده‌های مرتبط با تحلیل انتخاب، و از مجموعه داده برای انجام تحلیل‌ها بازیابی شوند.

تبدیل داده

تبدیل داده یک روش تثبیت داده نیز هست. در این فاز، داده‌های انتخاب شده به فرم دیگری تبدیل می‌شوند. این کار به ساده‌تر شدن، بهبود صحت و دقت فرآیند کاوش کمک می‌کند.

داده‌کاوی

در این فاز از روش‌های هوشمندانه برای استخراج الگوهای مهم و جالب توجه از میان داده‌ها استفاده می‌شود.

ارزیابی الگو

در این فرآیند، الگوهای حاصل شده در گام قبل، از جنبه‌های گوناگونی شامل دقت، صحت و قابلیت تعمیم و دیگر موارد مورد ارزیابی قرار می‌گیرند.

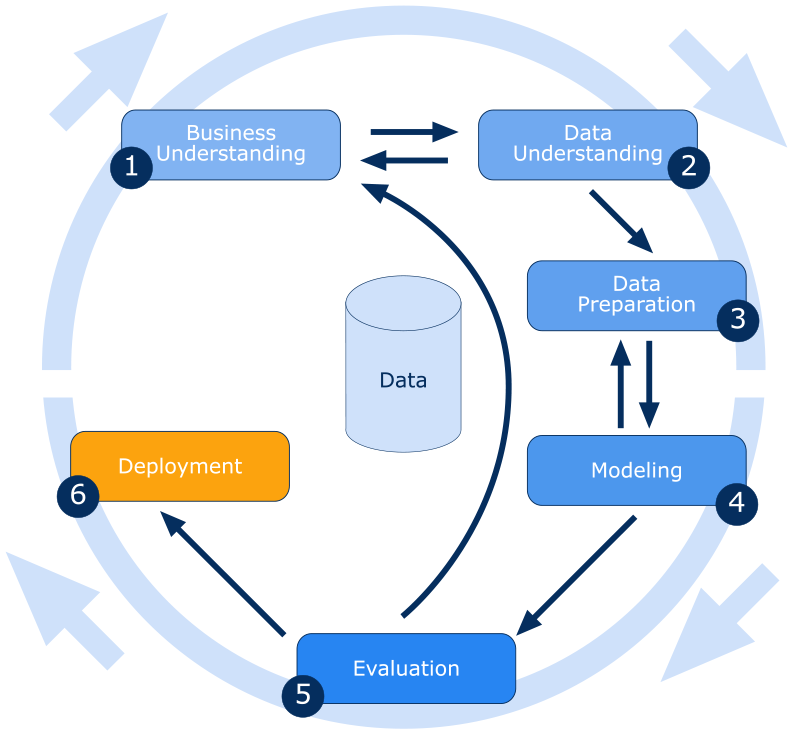
ارائه دانش

ارائه دانش فاز نهایی فرآیند داده‌کاوی است. در این فاز، دانش کشف شده به شیوه قابل درک به کاربر ارائه می‌شود. در این گام حیاتی و بسیار مهم، روش‌های بصری‌سازی مورد استفاده قرار می‌گیرند.

## CRISP (cross industry standard process)

فرآیند استاندارد صنعتی متقاطع برای داده‌کاوی یا به صورت مصطلح کریسپ، یک مدل فرآیندی استاندارد باز است که رویکردهای عمومی متخصصان داده‌کاوی را تشریح می‌کند، این روش‌شناسی پرکاربردترین مدل تحلیلی می‌باشد.

کریسپ یک مدل فرآیند داده‌کاوی است که راهبردهای معمولی که توسط داده‌کاوان خبره برای غلبه بر مشکلات داده‌کاوی استفاده می‌شود را شرح می‌دهد.

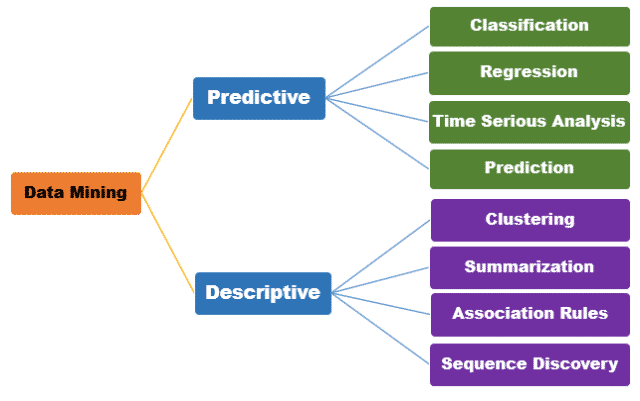


* درک فضای کسب و کار
* درک داده ها
* آماده سازی داده ها
* مدلسازی
* ارزیابی
* بکارگیری داده ها

## Types of Data Mining

The Data Mining Analysis can be divided in two basic parts. They are:

* Predictive Data Mining Analysis (تجزیه و تحلیل پیش بینی داده کاوی)
* Descriptive Data Mining Analysis (تجزیه و تحلیل توصیفی داده کاوی)



مدل سازی توصیفی:

آشکارسازی موارد مشابه یا گروه‌های مشترک در داده‌های موجود، با هدف تشخیص دلایل موفقیت یا شکست؛ از قبیل دسته‌بندی مشتریان بر اساس ترجیحات محصول یا احساسات آنها.

مدل سازی پیش بینی:

این مدل به گونه‌ای عمیق‌تر، به دسته‌بندی رویدادها در آینده می‌پردازد و می‌کوشد نتایج ناشناخته را پیشاپیش برآورد کند. به عنوان مثال، با استفاده از امتیاز اعتباری یک فرد، میزان احتمال بازپرداخت اقساط وام او را تخمین می‌زند.

# Preparation

## Data

داده چیست؟

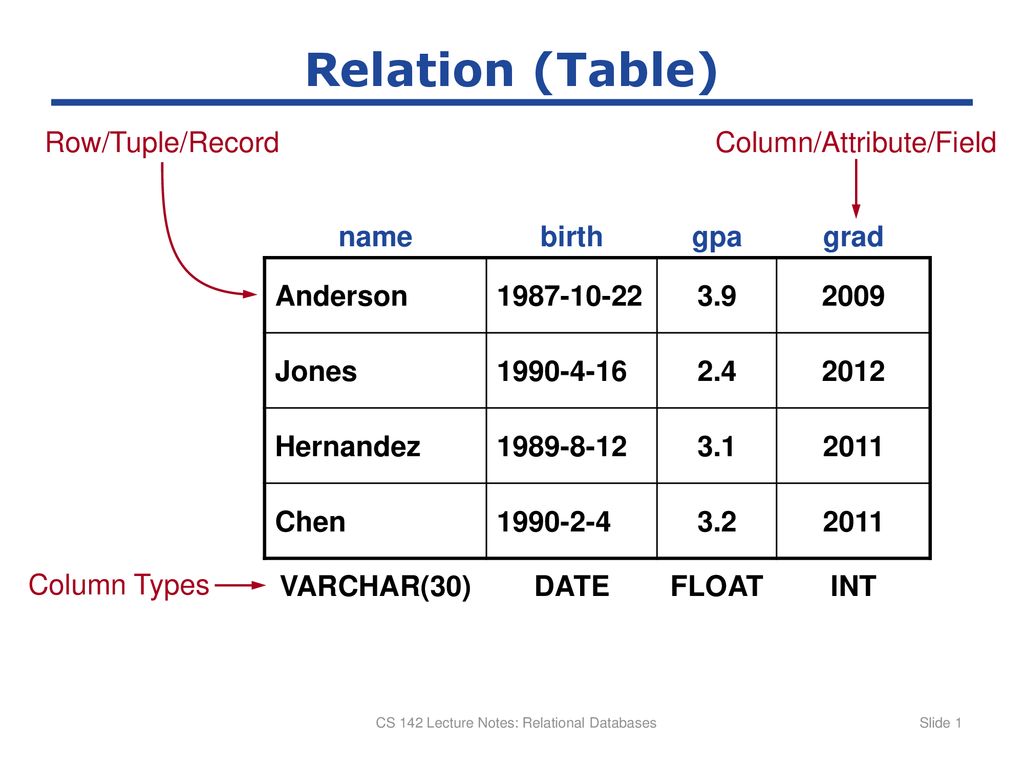
مجموعه ای از اشیا داده (Data Objective) و صفت های آن ها.

صفت (Attribute): یک ویژگی یا مشخصه یک شی (Object) می باشد.

* مثل: نام، نام خانوادگی، سن، رنگ چشم
* صفت با نام های دیگیری نیز شناخته می شود مثل: خصوصیت (Feature) و یا زمینه (Field)

شی (Object): مجموعه ای از صفات را یک شی می نامیم.

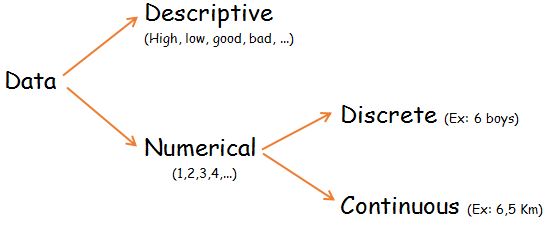
* شی با نام های دیگری نیز شناخته می شود مثل: رکورد (Record) و یا نمونه (Sample)



## Data Type

Data is defined as the facts & collected for the analysis purpose. Data is divided into qualitative data (descriptive) & quantitative data.

داده ها به داده های کیفی (توصیفی) و داده های کمی تقسیم می شوند.



داده های کیفی را نمی توان از نظر تعداد اندازه گیری کرد.

از طرف دیگر ، داده های کمی اطلاعاتی هستند که مقادیر عددی و محدوده استفاده می کنند. این داده ها به داده های گسسته (Discrete) و داده های پیوسته (Continuous) تقسیم می شوند.

داده های گسسته حاوی مقادیر محدود هستند و هیچ ارتباطی با یکدیگر ندارند. از طرف دیگر ، داده های پیوسته حاوی داده هایی است که می تواند اندازه گیری شود ، از جمله اعشار و کسری.