1. Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing.

Khai phá dữ liệu là việc nghiên cứu để thu thập, làm sạch, xử lý, phân tích và thu nhận hiểu biết sâu sắc hữu ích từ dữ liệu. Rất nhiều biến đổi tồn tại khi xét đến các lĩnh vực của vấn đề, ứng dụng, việc hình thành công thức và cách trình bày dữ liệu trong thực tế. Vì vậy, khai phá dữ liệu là một thuật ngữ rất rộng dùng để miêu tả những mặt khác nhau này của việc xử lý dữ liệu.

1. Dữ liệu (Data) (chính xác là tập dữ liệu (Data set)) được hiểu như là một tập F gồm hữu hạn các trường hợp (sự kiện), Theo nội dung của phát hiện tri thức trong các CSDL, dữ liệu phải bao gồm nhiều trường hợp.
2. Mẫu (Model): Trong quá trình KDD, người ta sử dụng một ngôn ngữ L để biểu diễn các tập con sự kiện (dữ liệu) thuộc vào tập sự kiện F, theo đó mỗi biểu thức E trong ngôn ngữ L sẽ biểu diễn một tập con FE tương ứng các sự kiện trong F. E được gọi là *mẫu* nếu nó đơn giản hơn (theo một ngữ cảnh nào đó) so với việc liệt kê các sự kiện thuộc FE.
3. Có giá trị (Confidence): Mẫu được phát hiện cần phải *có giá trị* đối với các dữ liệu mới (xuất hiện trong tương lai) theo một mức độ chân thực nào đó. Tính chất *có giá trị* được hiểu theo nghĩa liên quan tới một *độ đo tính có giá trị (chân thực)* (confidence measures) là một hàm C ánh xạ một biểu thức thuộc ngôn ngữ biểu diễn mẫu L tới một không gian đo được (bộ phận hoặc toàn bộ) Mc .
4. Tính mới (Novelty): Mẫu phải là mới trong một miền xem xét nào đó, ít nhất là hệ thống đang được xem xét. Tính mới có thể đo được khi quan tâm tới sự thay đổi trong dữ liệu (bằng việc so sánh giá trị hiện tại với giá trị quá khứ hoặc giá trị kỳ vọng) hoặc tri thức (tri thức mới quan hệ như thế nào với các tri thức đã có). Tổng quát, điều này có thể được đo bằng một hàm N(E,F) hoặc là độ đo về tính mới hoặc là độ đo kỳ vọng.
5. Hữu ích tiềm năng (Utility): Mẫu cần có khả năng chỉ dẫn tới các tác động hữu dụng và được đo bởi một hàm tiện ích. Chẳng hạn, hàm U ánh xạ các biểu thức trong L tới một không gian đo có thứ tự (bộ phận hoặc toàn bộ Mu), theo đó u = U(E,F).
6. Có thể hiểu được: Một mục tiêu của KDD là tạo ra các mẫu mà con người hiểu chúng dễ dàng hơn các dữ liệu nền (dữ liệu sẵn có trong hệ thống). Chính vì lý do tiêu chí này là khó mà đo được một cách chính xác cho nên thường tính chất “có thể hiểu được” được thay bằng một độ đo về sự dễ hiểu. Tồn tại một số độ đo về sự dễ hiểu, các độ đo như vậy được sắp xếp từ cú pháp (tức là cỡ của mẫu theo bit) tới ngữ nghĩa (tức là dễ dàng để con người nhận thức được một tác động nào đó). Bởi lý do đó, chúng ta gia định rằng tính hiểu được là đo được bằng một hàm S ánh xạ biểu thức E trong L tới một không gian đo được có thứ tự (bộ phận hoăc toàn bộ) Ms; theo đó, s = S(E,F).
7. Độ hấp dẫn (interestingness): thường được coi như một độ đo tổng thể về mẫu là sự kết hợp của các tiêu chí giá trị, mới, hữu ích và có thể hiểu được. Một số hệ thống KDD thường sử dụng một hàm hấp dẫn dưới dạng hiển i = I(E, F, C, N, U, S) thực hiện ánh xạ một biểu thức trong L vào một không gian đo được Mi. Một số hệ thống KĐ khác lại có thể xác định giá trị hấp dẫn của mẫu một cách trực tiếp thông qua thứ tự của các mẫu được phát hiện.
8. Tri thức: Một mẫu E L được gọi là tri thức nếu như đối với một lớp người sử dụng nào đó, chỉ ra được một ngưỡng i Mi mà độ hấp dẫn I(E, F, C, N, U, S) > i.
9. Phân lớp (Classification/Categorization) thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo. (học máy giám sát (supervised learning))
10. Phân cụm (Clustering) thực hiện việc nhóm dữ liệu thành các “cụm” (có thể coi là các lớp mới) để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng. Thông thường, mục tiêu định hướng của bài toán phân cụm là cực đại tính tương đồng giữa các phần tử tỏng mỗi cụm và cực tiểu tính tương đồng giữa các phần tử thuộc các cụm khác nhau. (học máy không giám sát (unsupervised learning))