Ⅰ. 금융 빅데이터와 데이터 분석 (Big data in finance & financial time series data analysis)

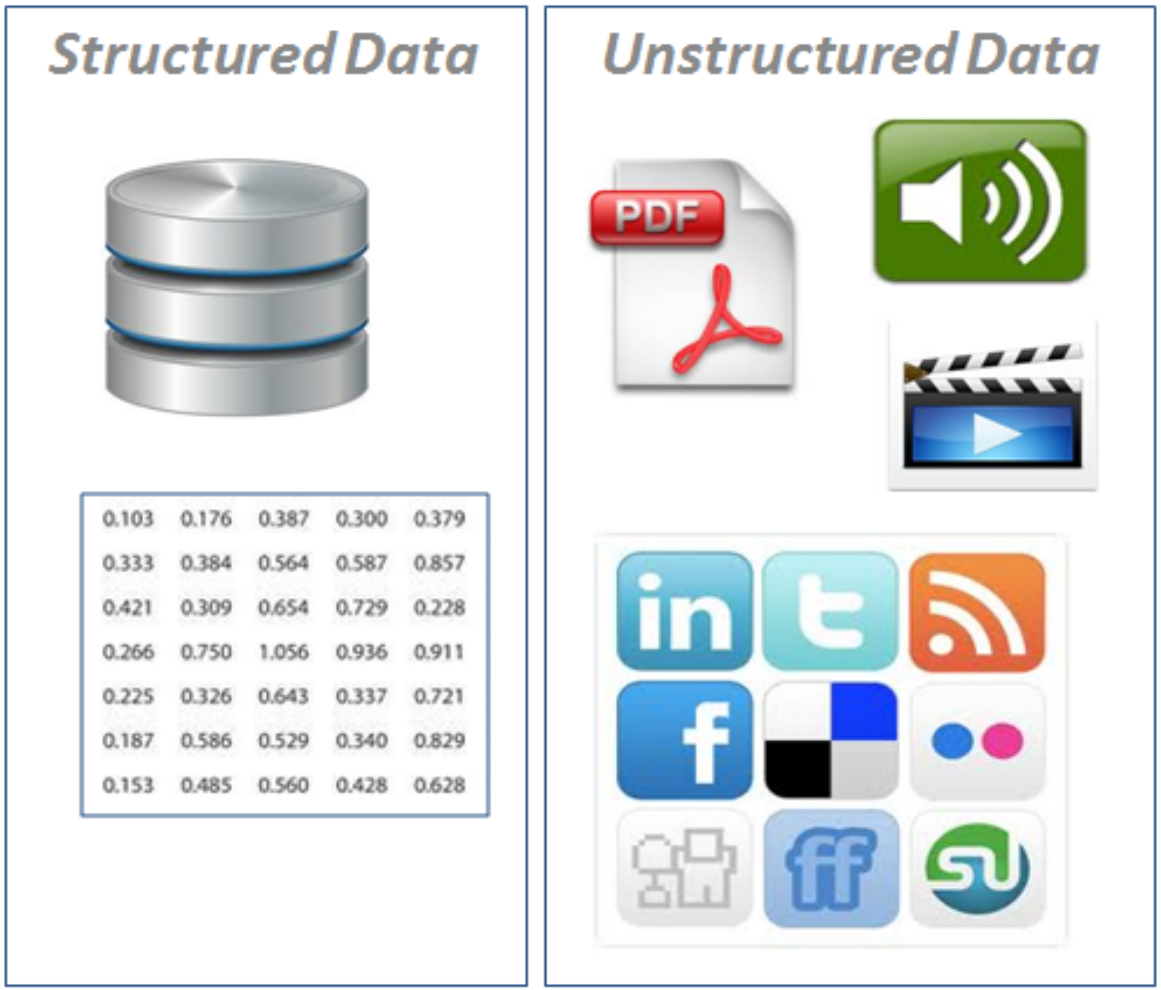
❷ 데이터마이닝 기본 개념

데이터, 정보, 지식

- **데이터** : 과거에 발생한 상황에 대한 사실적 자료들의 집합체
- **정보** : 데이터로부터 의사 결정에 관련된 사항들을 추출하여 상호 연관성, 상관성, 패턴 등을 분석하여 요약 및 정리한 것
- **지식** : 객관적 사실에 입각한 정보들을, 추론하고 일반화 시켜, 의사 결정에 유용한 형태로 변환한 것으로, 일반화된 지식을 바탕으로 몰랐던 것을 알게 되고 (Inferring unknowns from knowns), 미래의 결과를 예측 (prediction) 해 볼 수 있음. (prediction은 forecasting보다 estimation 혹은 inference 쪽에 더 가까움)

데이터(자료; data)의 분류

- **데이터(자료; data)**는 정형자료(Structured data)와 비정형자료(Unstructured data)로 나뉨



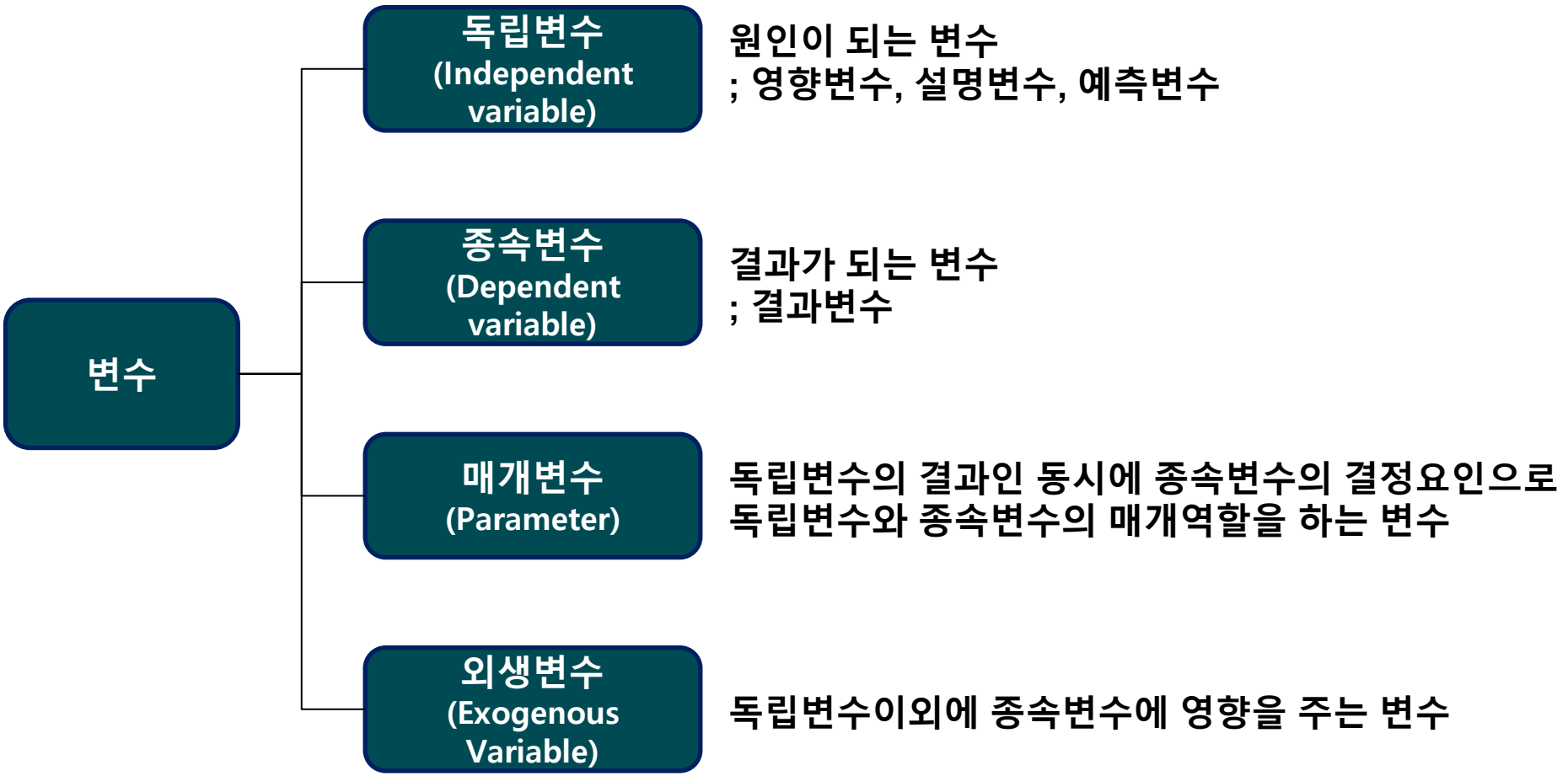
- 우리가 일상적으로 접하는 자료는 대개 정형자료이나 전체 자료 중 정형자료의 비중은 20%에 불과하며, 나머지 80%가 비정형자료로 비정형자료를 의미 있게 분석하는 것이 매우 중요함

변수(Variable)란 무엇인가?

- 비즈니스 애널리틱스(BA)에서 **변수(Variable)**란 **모형에 전달되는 정보**나 **그 밖의 상황에 따라 바뀔 수 있는 값**을 의미
- **변수(Variable) = 개체의 속성(Feature)**

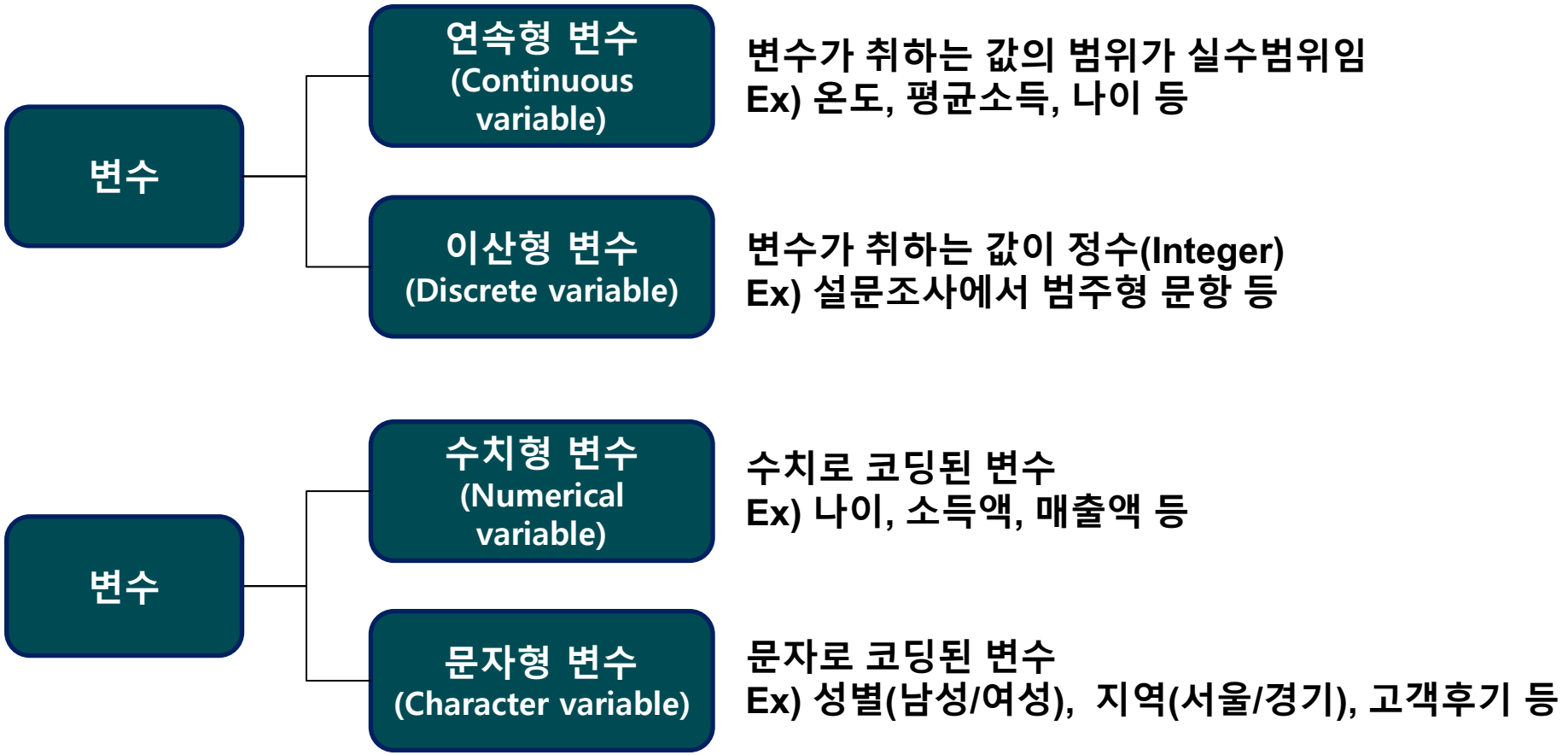
기능에 따른 변수(Variable) 분류

- 변수는 기능적 역할에 따라 독립변수, 종속변수, 매개변수 및 외생변수로 나뉨



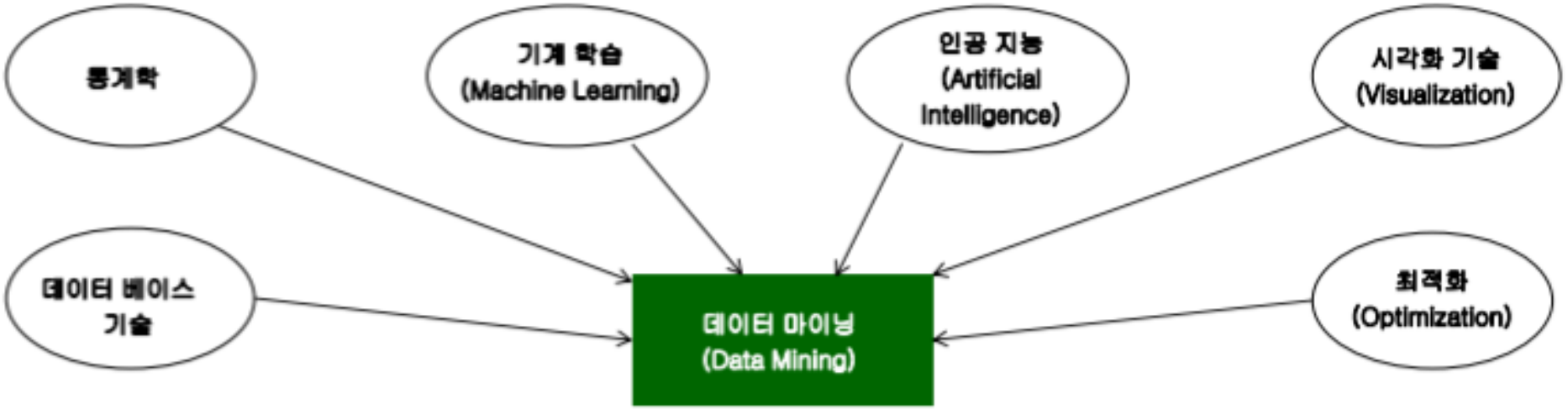
속성에 따른 변수(Variable) 분류

- 변수는 Type에 따라 연속형 변수 / 이산형 변수 혹은 숫자형 변수 / 문자형 변수 등으로 나뉨



데이터 마이닝이란

- **데이터 마이닝**이란 방대한 데이터 (Data)로부터 유용한 정보 (Information)를 추출하고, 의사 결정을 위한 지식 (Knowledge)을 얻는 일련의 과정을 의미한다.
- 데이터 마이닝은 새로운 기술은 아니며 기존의 기술들과 융합된 것이다.



데이터 마이닝의 필요성

1. 회사들이 데이터베이스 시스템에 넣는 데이터의 양은 해마다 끊임없이 증가
2. 인터넷과 전자상거래가 급속하게 보급되면서 소비자와 구매에 관련된 많은 양의 데이터가 자동으로 데이터베이스에 축적됨
3. 과거에는 가능하지 않았던 거대한 양의 데이터를 우리 주변에서 쉽게 찾아볼 수 있는 시대가 도래

- 하지만 이렇게 축적된 데이터로부터 아주 유용한 정보를 찾아내 마케팅이나 회사의 이익을 효율적으로 증대하기 위해 사용하긴 아직도 어려움이 많다.
- 그 이유 중 하나는 정보가 아주 많은 양의 데이터 안에 함축적으로 숨어 있어 사람의 눈으로 일일이 조사하는 것이 불가능하기 때문이다.
- 데이터 마이닝 분야에서 개발된 기술을 통해 이러한 데이터로부터 유용하고 값진 정보를 효과적으로 찾아내 회사의 이익 뿐만 아니라 개인의 일상생활의 편의도 증대시키는 곳에 적용된다.

데이터 마이닝의 활용 I

- **백화점**에서 물건을 진열할 때 고객의 움직임을 줄여 주기 위해 활용할 수 있음. 고객의 구매 패턴을 보고 유용한 패턴을 찾아내 소비자가 살 물건을 미리 예측하고, 쿠폰을 발행해 관심을 유발함으로써 판매를 촉진할 수 있음

- **보험 회사**에서는 고객이 다른 회사로 옮기는 것을 방지하거나 고객의 위험성에 따라 보험료를 차등화해 제공하는 데 사용할 수 있음

- **신용카드 회사**에서는 훔친 신용카드를 사용하는 경우를 발견해 더 이상의 불법 사용을 막거나 새로운 고객이 신용카드를 신청할 경우에 카드 발급 결정에 사용할 수 있음

데이터 마이닝의 활용II

- **전자상거래를 위한 웹 서버인 경우**에는 소비자가 방문한 웹 페이지와 구매한 물건과 소비자의 특징을 보관하고 있기 때문에 이 데이터를 분석하면 각각의 사용자에 맞는 웹 페이지를 동적으로 그때 그때 생성해주거나, 웹 페이지의 캐싱(Caching), 프리페칭(Prefetching), 스와핑(Swapping)을 효율적으로 제공할 수 있어 성능을 높이고 수행속도를 빠르게 할 수 있다. 더욱이 웹 액세스의 다차원 웹 로그 분석을 이용한 트렌드 분석을 통해 웹에서 어떤 일이 일어나고 있는지에 대해서도 대략적인 정보를 제공해줄 수 있다. 또한 모든 소비자에게 동일한 웹 페이지를 제공하는 것이 아니라 소비자의 관심에 따라 다른 웹 페이지를 동적으로 만들어 제공하는 개인화(personalization) 서비스를 가능하게 할 수도 있다.

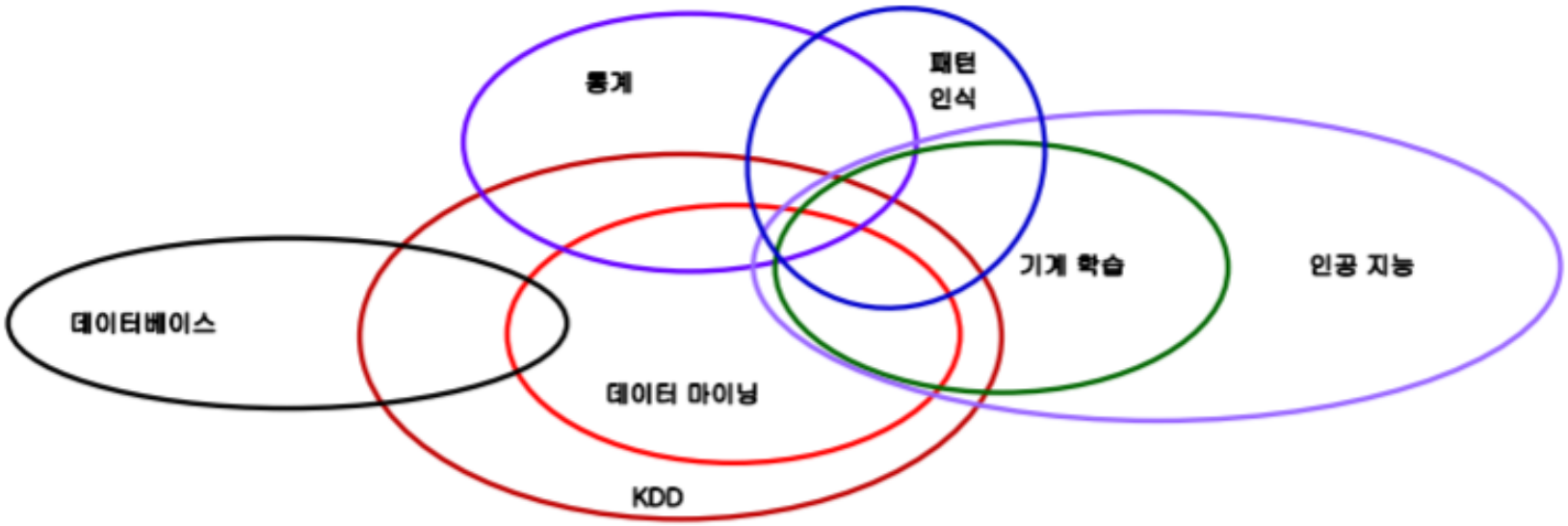
데이터 마이닝의 활용ㅍ

- **네트워크 분야**에서는 네트워크에 이상이 생기기 전에 과거의 네트워크에 관련된 데이터를 이용해 앞으로 몇 시간 안에 네트워크에 생길지도 모르는 문제를 미리 예측해낼 수도 있다. 피자헛 가게를 새로운 장소에 개점할 경우에 과거의 다른 피자헛 가게가 세워진 곳에 관련된 정보로부터 새로 세우는 장소에서 성공할지를 예측하는 데도 사용할 수 있다.

- 교차 판매(cross-selling)나 상승 판매(up-selling) 등을 통해 **회사의 판매 실적**을 더 높일 수도 있다. 교차 판매란 서로 다른 부류에 속하는 상품이지만 서로 연관되어 고객들이 구매하는 경우를 찾아 연관된 상품을 고객에게 추천해 판매하는 것을 뜻한다. 예를 들어 장난감을 사는 고객이 생명보험에 들 가능성이 많다면 장난감을 사는 고객에게 생명보험에 관한 정보도 제공해 보험에 가입할 수 있도록 만드는 것을 말한다. 상승 판매란 1억원의 생명보험을 가입하려는 고객에 대한 정보를 분석해보고 만일 그 고객이 2억원짜리 보험에 가입할 가능성이 많은 고객이라면 2억원의 보험에 대해 같이 소개하고 추천해 더 비싼 보험을 들도록 유도하는 것을 말한다. 이 밖에도 여러 분야에서 데이터 마이닝 기술을 유용하게 사용할 수 있다.

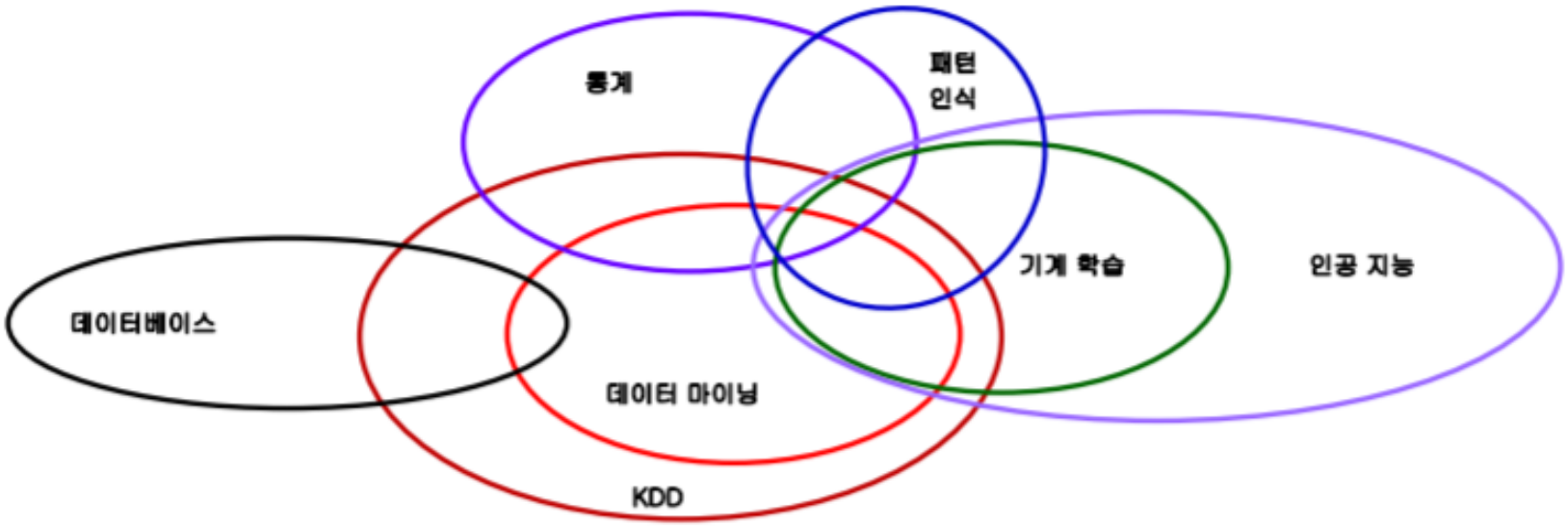
지식 탐사와 데이터 마이닝 (Knowledge Discovery in Database : KDD, Data Mining : DM)

- **지식 탐사(KDD)**도 데이터로부터 유용한 지식을 얻는 과정이라는 점에서 데이터 마이닝과 유사한 의미로 사용됨
- 지식 탐사(KDD)는 1989년 첫 번째 워크샵에서 지식이 데이터에서 발견된 최종 산출물이라는 것을 강조하기 위해 사용된 것으로 초기에는 KDD와 데이터 마이닝이 혼용되어 사용됨



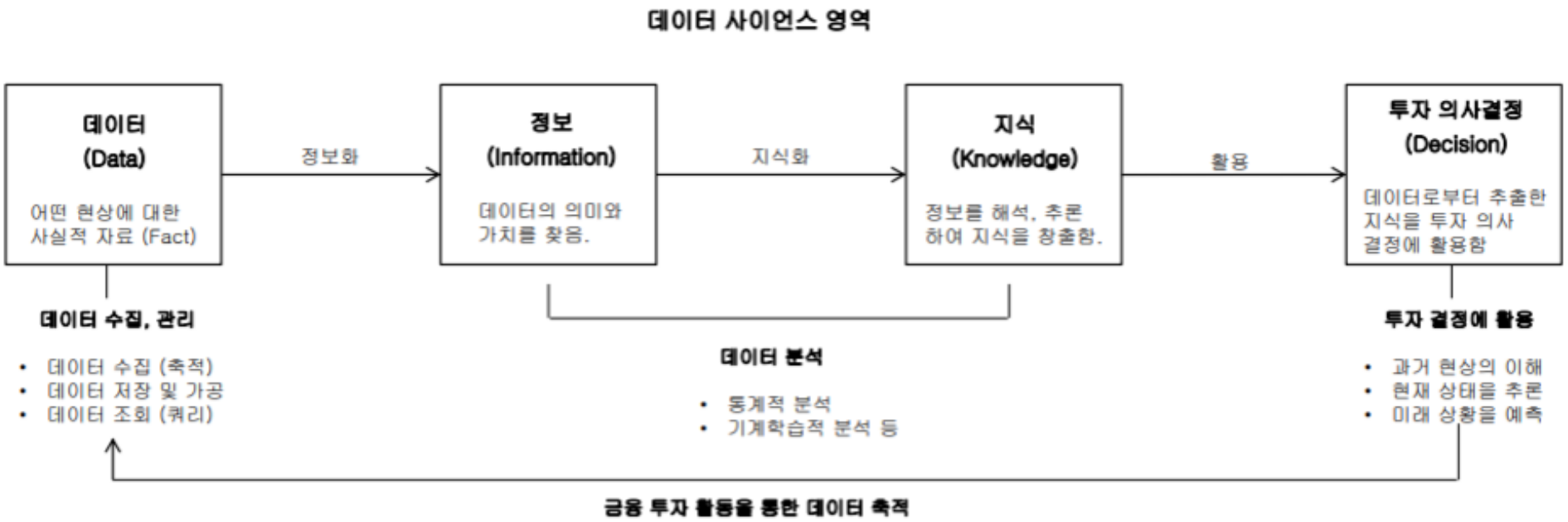
지식 탐사와 데이터 마이닝 (Knowledge Discovery in Database : KDD, Data Mining : DM)

- 1995년 캐나다 몬트리올에서 개최된 "The first international conference on knowledge discovery & data mining" 에서 KDD는 데이터로부터 지식을 발견하는 전체적인 프로세스로 정의함 -> 데이터 마이닝보다 넓은 개념
- **KDD**는 지식 발견을 위한 포괄적인 프로세스를 의미하고, **데이터 마이닝**은 특별한 알고리즘이나 기술을 적용하는 KDD의 한 단계로 간주됨



지식 탐사 과정

- **데이터 분석**이란 수집/축적된 데이터를 기반으로 어떤 현상에 대한 사실적 정보를 추출하고 그 정보로부터 유용한 지식을 얻어 (추론적 정보) 과거 현상에 대해 깊이 이해하거나, 미래 상황을 예측하여 투자 의사결정에 활용하기 위한 과정임.



지식 탐사 과정

- **데이터 분석의 목적**은 사실 (Fact)에 근거한 의사결정을 추구함에 있음 -> Data-Driven Decision making (DDD)
- 축적된 데이터는 정보화, 지식화 과정을 거치면서 그 가치가 커지고 비지니스 활동을 통해 발생하는 데이터가 다시 축적되어 데이터로부터 얻을 수 있는 지식이 점차 증가함