2.1 정형 데이터

정형 데이터는 데이터 형식이 정해져 있는 데이터이다.
쉽게 말하면 성별, 요일, 날짜 등의 형식이 정해져 있는 데이터를 생각하면 될 듯 하다.
보통 정형 데이터는 공공데이터 포털에서 제공해주는 데이터를 생각하면 될 듯 하다.(반정형 데이터도 있긴하지만, 대부분 정형 데이터가 많은 편이다.)
글쓴이도 최근에 공공데이터를 활용해서 시간별, 역별, 요일별 전철의 복잡도를 분석해서 예측을 해본 적이 있다. 이것도 정형 데이터랑 이따 설명에 나오는 반정형 데이터를 혼합해서 사용한 것이다.

2.2 비정형 데이터

비정형 데이터는 데이터 형식 정해져 있지 않는 데이터라고 생각하면 된다.
즉 SNS 글이나, 멀티미디어 데이터를 생각하면 될 듯 하다.
SNS글은 크롤링을 해서 분석을 하게 되면, 형식이 정해져 있지 않은 데이터라서, 가공하고, 파악하는데 너무나 오래 걸리고, 이미지나 영상 같은 경우도 수많은 패턴을을 딱 뽑아서 해야 되기 때문에 엄청나게 오래 걸리거나, 작업 난이도가 생각보다 높은 편이고, 고성능의 컴퓨팅 기술이 필요하다.
요즘은 TensorFlow나, Keras 등 라이브러리가 생겼고, Hadoop, 맵리듀스 등의 기술이 생겨서 전에보다는 그나마 접근성이 낮아진 작업이 아닐까 싶다. (단 고성능의 컴퓨팅 기술을 못 쓰면 꽝이지만..)

2.3 반정형 데이터

반정형 데이터는 데이터 내부에는 논리적인 형식을 가지고 있으나, 외형상으로는 데이터 포맷이 정형 데이터처럼 완전히 정의되어 있지 않는 데이터라고 생각하면 된다.
즉 무슨 소리이나면, GPS 값, 통화시간 등과 같이 형식을 정해져 있으나, 값이 방대한 데이터라고 생각하면 될 듯 하다.

이렇게 3가지 데이터를 가지고 데이터 분석을 많이 하는 편이다. 은행쪽에서는 주식이나, 이용자가 입,출금 패턴 분석 할 때는 반정형과 정형을 같이 사용하여 분석을 하고, 영상처리나, 이미지 처리 등은 비정형 데이터를 이용해서 처리하면 된다고 생각하면 된다.