

TEXT MINING for PRACTICE

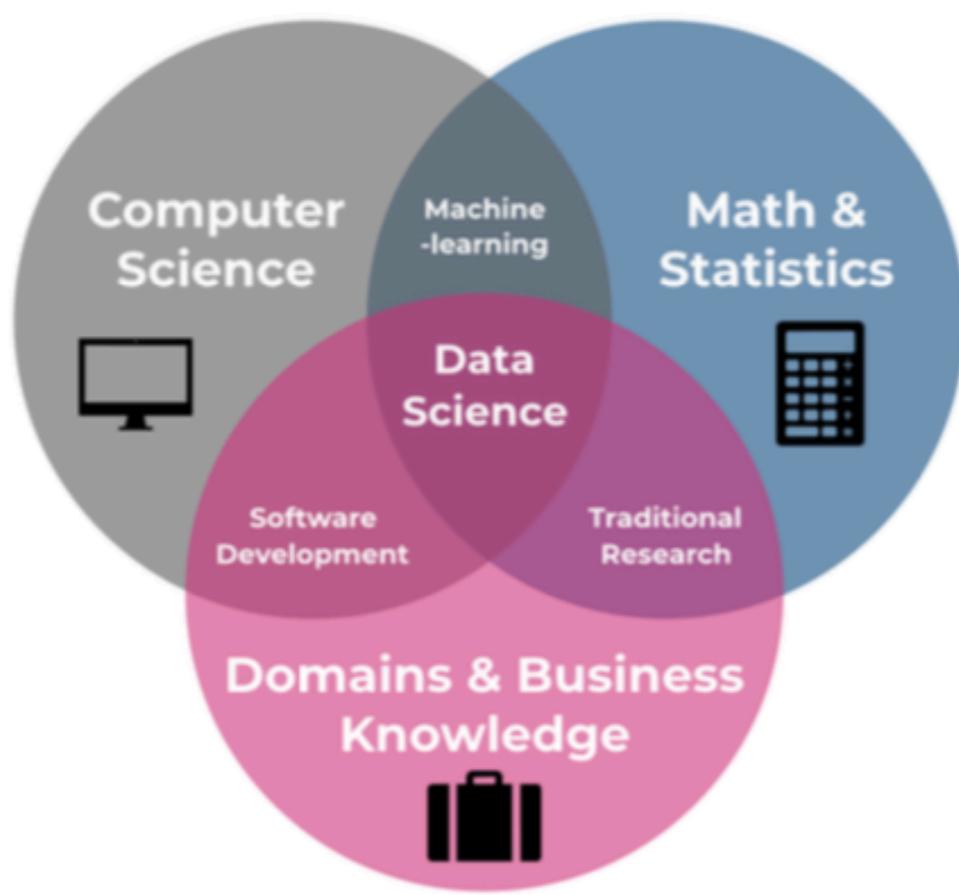
Python을 활용한 비정형 데이터 분석 - WEEK 01
Introduction

연세대학교 | 서중원

Data Scientist

데이터 과학자의 주요 업무는?

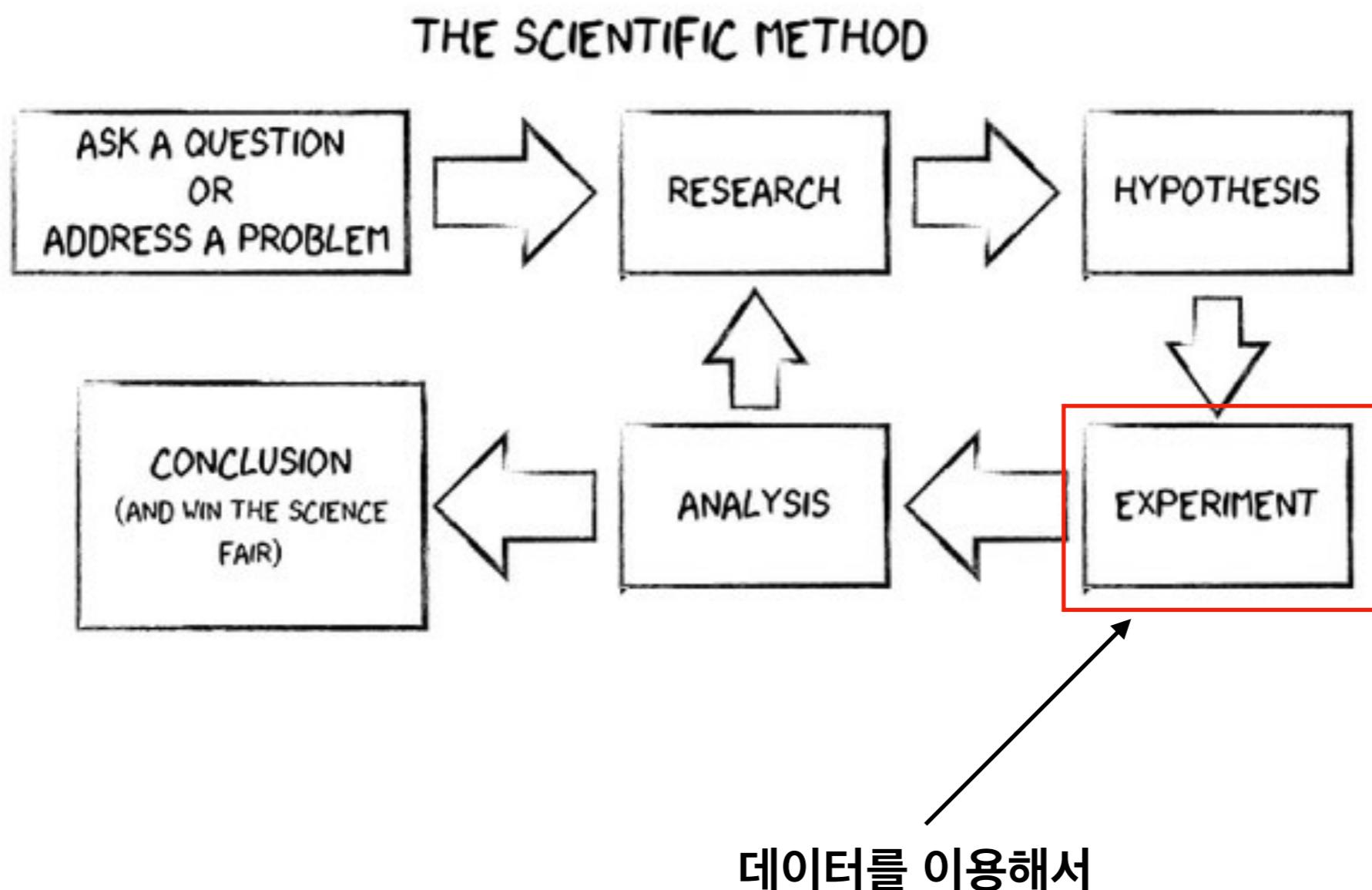
- ▶ 데이터 분석을 기반으로 비즈니스 패턴을 기회로 활용하거나, 문제점을 도출하여 해방안을 제시하는 직업군
- ▶ Data Analyst : 비즈니스와 결합한 실행 가능한 통찰력 (insight)을 제공하는 직업
- ▶ Data Engineer : 데이터 분석 산출물을 위한 소프트웨어 (software) 환경을 설계하고 구현하는 직업



	Data Analyst	Machine-learning Engineer	Data Engineer	Data Scientist
Programming Tools	H	H	H	H
Data Visualization & Communication	H	M	M	H
Data Intuition	M	H	M	H
Statistics	M	H	M	H
Data Wrangling	L	L	H	H
Machine Learning	L	H	L	H
Software Engineering	L	M	H	M
Multivariabile Calculus & Linear Algebra	L	H	L	M

* Importance : H > M > L

왜 Scientist인가?



Data Scientist

1. Generalists

- ▶ 데이터 분석적 사고가 가능한 사람
- ▶ 데이터 분석에 대한 이해와 사고능력을 가진 사람

2. Industry specialists

- ▶ 데이터 분석적 사고를 통해 문제를 해결하고자 하는 도메인 전문가

3. Deep specialists

- ▶ 특정 데이터 분야에 전문지식을 가진 사람

4. Analytics developers

- ▶ 데이터 분석 전문지식과 함께 이를 S/W로 구현 가능한 전문가
- ▶ 알고리즘 구현을 포함해 코딩능력이 필수

5. Data engineers

- ▶ 데이터 분석의 전 과정을 파이프라인으로 구축하고 자동화할 수 있는 능력을 가진 전문가

왜 Data Scientist가 되야하나?

A.I.가 우리 직업을 대체할 것이기 때문에?

자동화 위험이 높은 상위 20대 직업			자동화 위험에 따른 대체 확률		
분류코드	직업명	대체확률	분류코드	직업명	대체확률
5302	통신서비스 판매원	0.990	2440	영어 전문 번역자	0.009
5303	텔레마케터	0.990	2411	장학금 관리자	0.009
5304	인터넷 판매원	0.990	2591	장학금 관리자	0.009
8922	사진인화 및 현상기 조작원	0.990	1312	교육 프로그램 개발자	0.009
2714	관세사	0.985	1331	보건 관리자	0.009
3125	무역 사무원	0.985	2521	중고차 판매원	0.009
3142	전산 자료 입력원 및 사무 보조원	0.980	2545	학습지 및 망문 교사	0.009
3132	경리 사무원				
5220	상품 대여원				
8212	표백 및 염색 관련 조작원				
8222	신발제조기 조작원 및 조립원				
8324	고무 및 플라스틱 제품 조립원				
8912	가구조립원				
8919	기타 목재 및 종이 관련 기술자				
9991	구두 미화원				
3201	출납창구 사무원	0.965	2392	섬유공학 기술자	0.018
3126	운송 사무원	0.960	2393	가스에너지 관리자	0.018
8211	섬유제조 기계조작원	0.960	1311	연구 관리자	
2712	회계사	0.957	2311	건축가 및 건축공학 기술자	0.018
2713	세무사	0.957	2341	환경공학 기술자 및 연구원	0.018

[Today글로벌뉴스] 영국 교육부, AI로 사라질 직업군 재교육 계획 발표



[이코노믹리뷰=홍석윤 기자] 인공지능(AI)과 로봇이 인간을 대체하는 무인화 현상에 대비하는 정책을 펼치는 나라들이 늘고 있다고 영국 BBC가 17일(현지시간) 보도했다.

AI 상담사 ‘콜센터 취업, 성공적’…노동시장 중장기 변화 불가피

미래일자리 | AI

이원갑 기자 | 기사작성 : 2019-07-10 17:00

시사 > 전체기사

잘 나가는 AI, 일자리가 떨고 있다



AI가 대체할 일에 대한 사회적 합의 우선

입력 : 2018-09-15 04:01

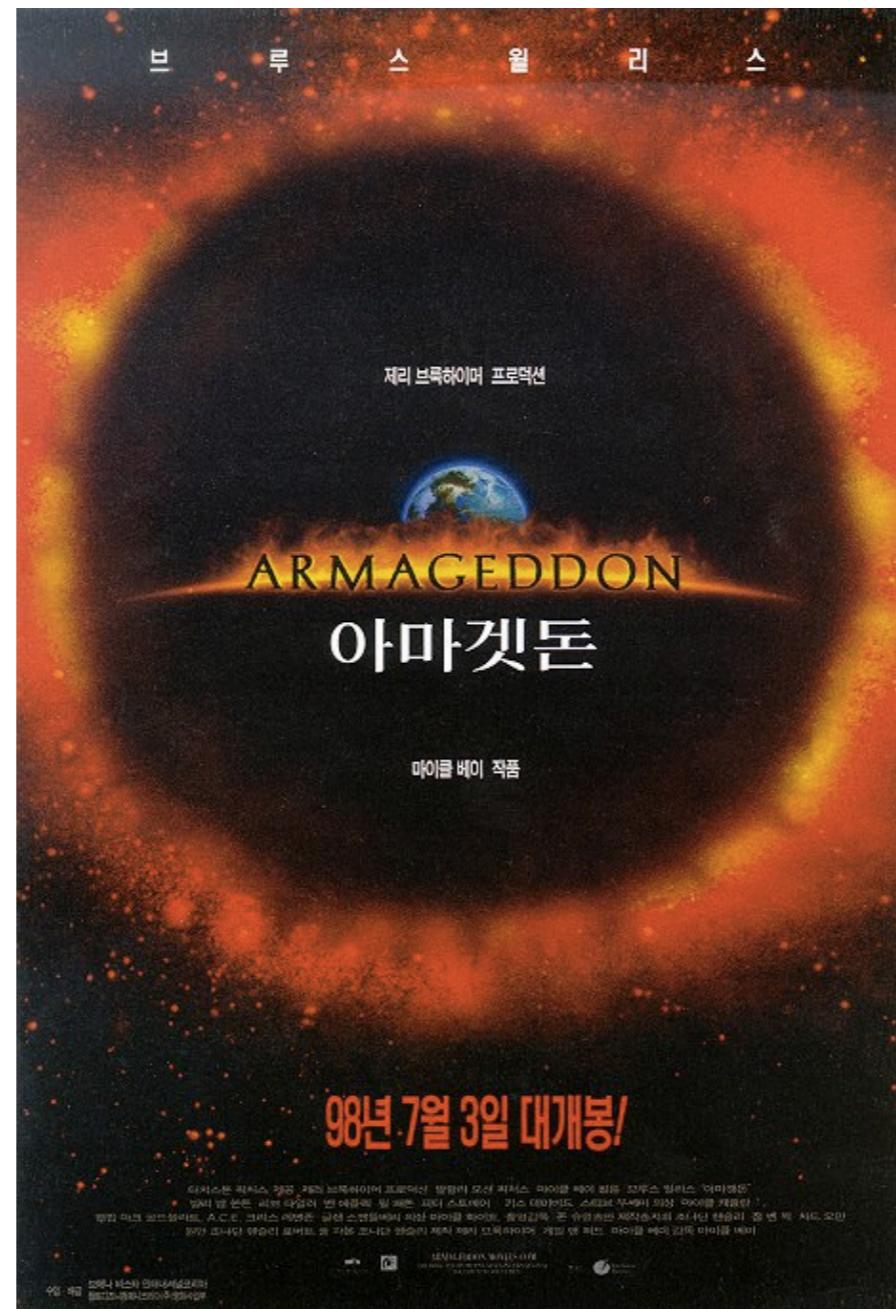
자료 : Frey & Osborne(2013), LG경제연구원

* Source: 홍석윤, [Today글로벌뉴스] 영국 교육부, AI로 사라질 직업군 재교육 계획 발표, 2019.07.18, <http://www.econovill.com/news/articleView.html?idxno=367857>.

** Source: 이원갑, AI 상담사 ‘콜센터 취업, 성공적’…노동시장 중장기 변화 불가피, 2019.07.10, <http://www.news2day.co.kr/131572>.

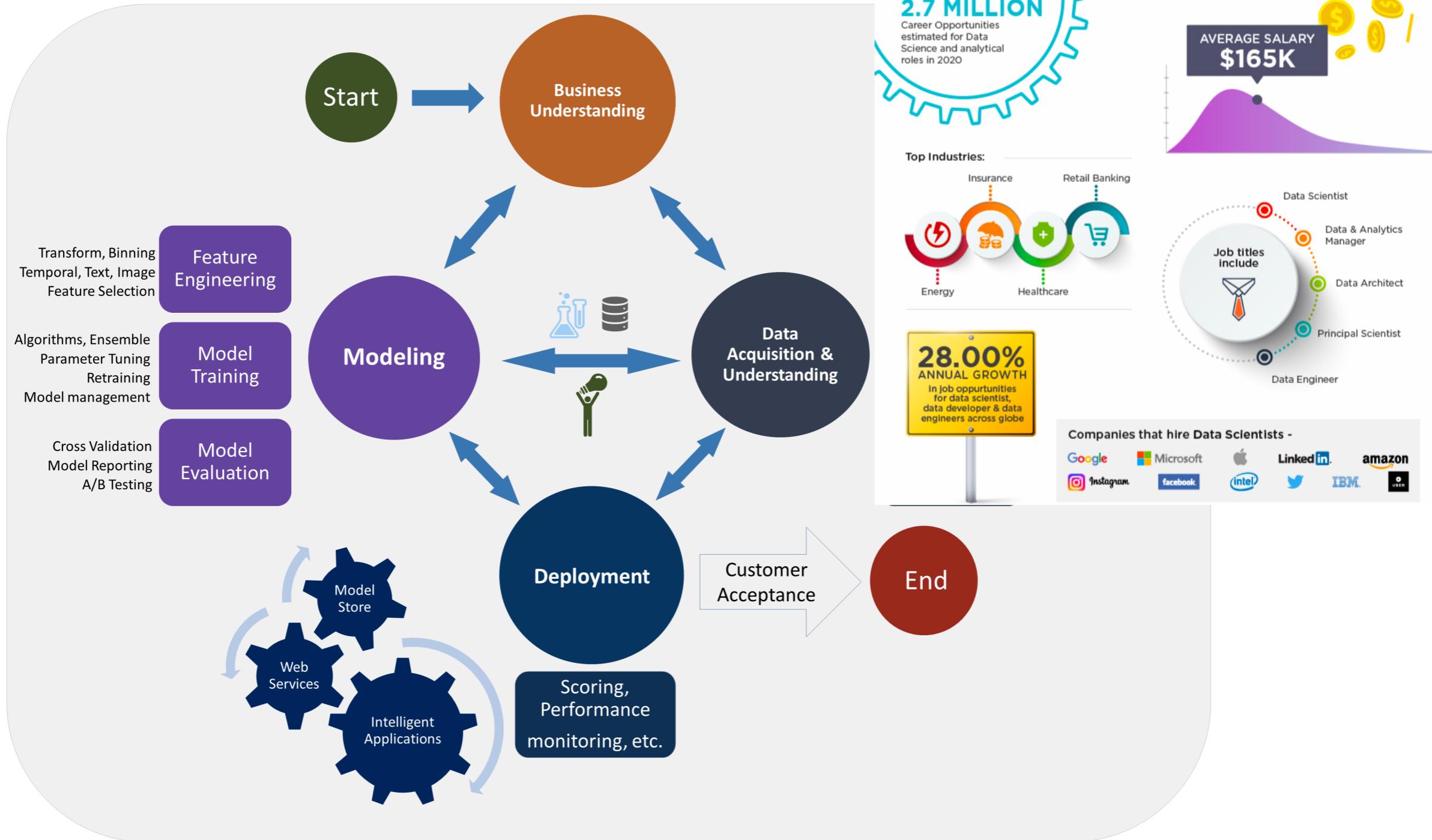
*** Source: 강창욱, 잘 나가는 AI, 일자리가 떨고 있다, 2018.09.15, <http://news.kmib.co.kr/article/view.asp?arcid=0924006455>.

이 영화 기억 하시나요?



Data Science Lifecycle

Data Science Lifecycle



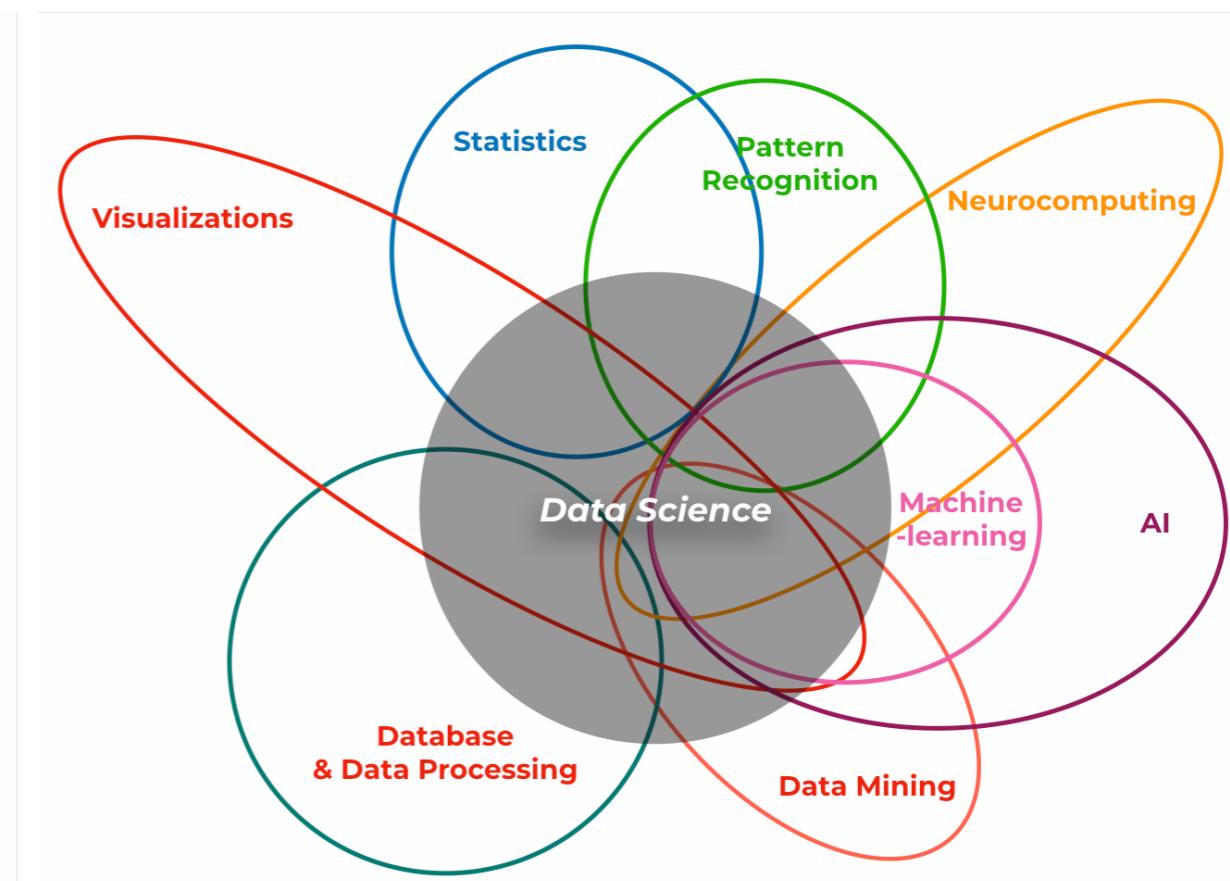
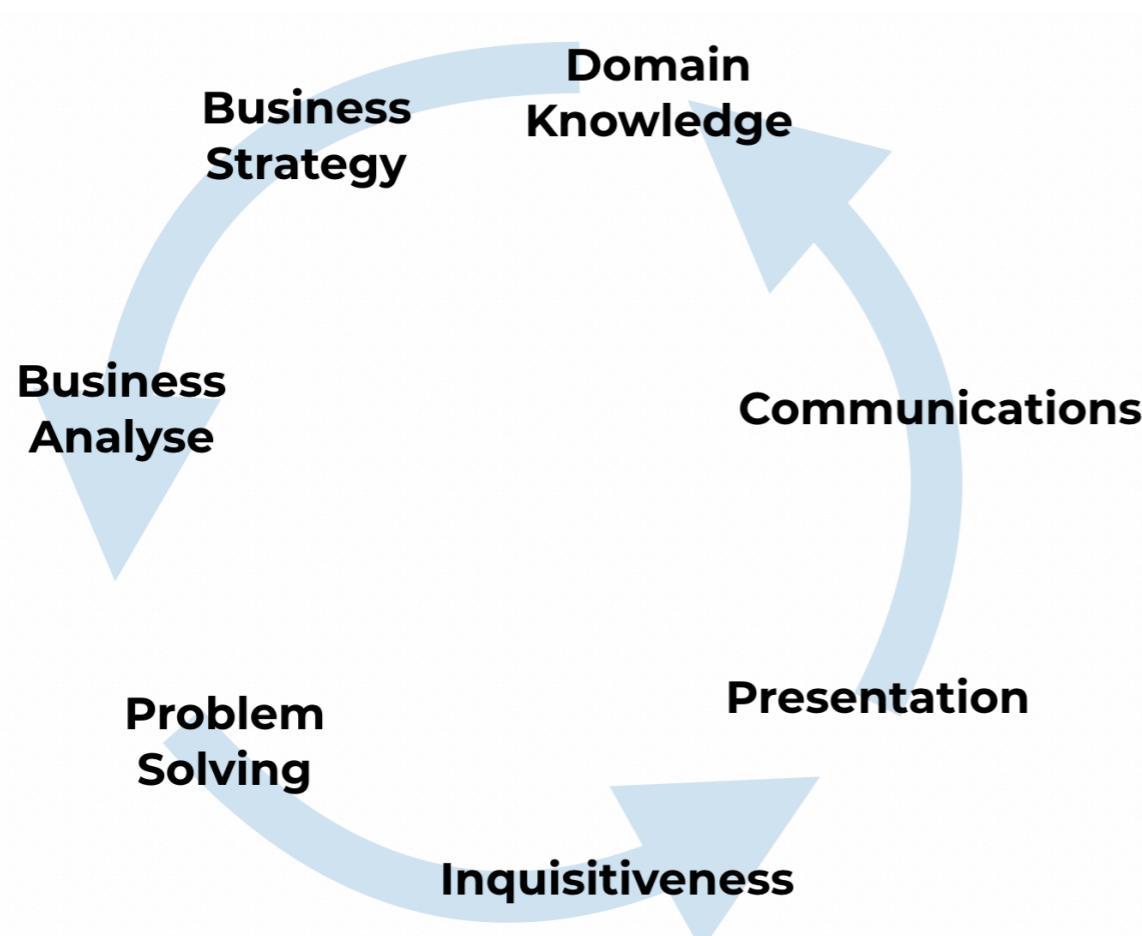
*Source : Gary Ericson et al., What is the Team Data Science Process?, 2017.10.20., <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview/>.

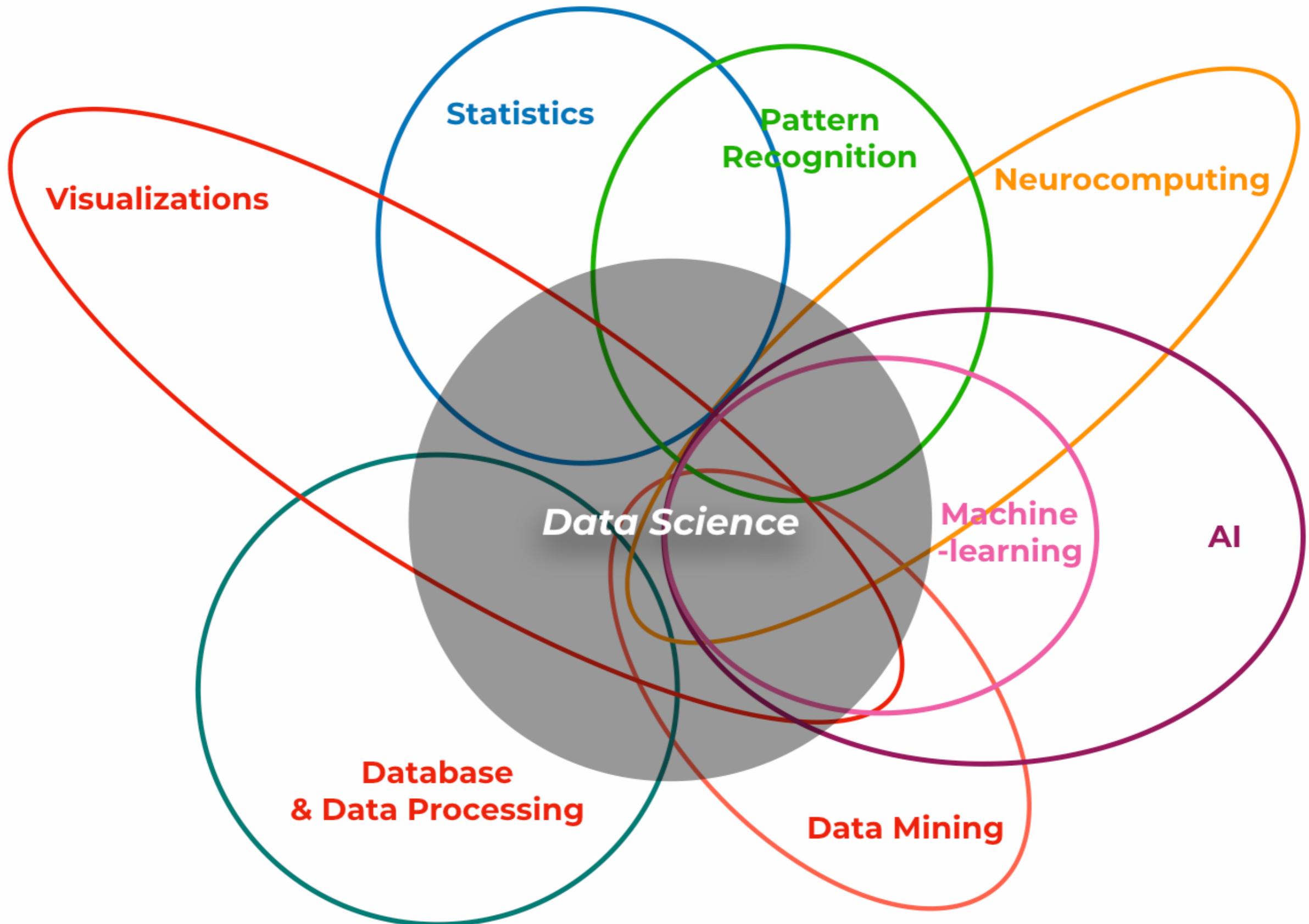
**Source : Simplilearn Solutions, About the program, <https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training/>.

Data Mining

Data mining 이란?

- ▶ 데이터 속의 유용한 패턴(지식)을 찾아내는 것
- ▶ 지식발견은 중요한 의사결정을 위해 데이터에서 유효하고 (valid), 새롭고 (novel), 잠재적으로 유용 (potentially useful)하면서, 궁극적으로 이해할 수 있는 패턴 (pattern)이나 관계 (relationship)을 파악해 가는 프로세스
- ▶ “Its goal is to develop knowledge of some phenomena”

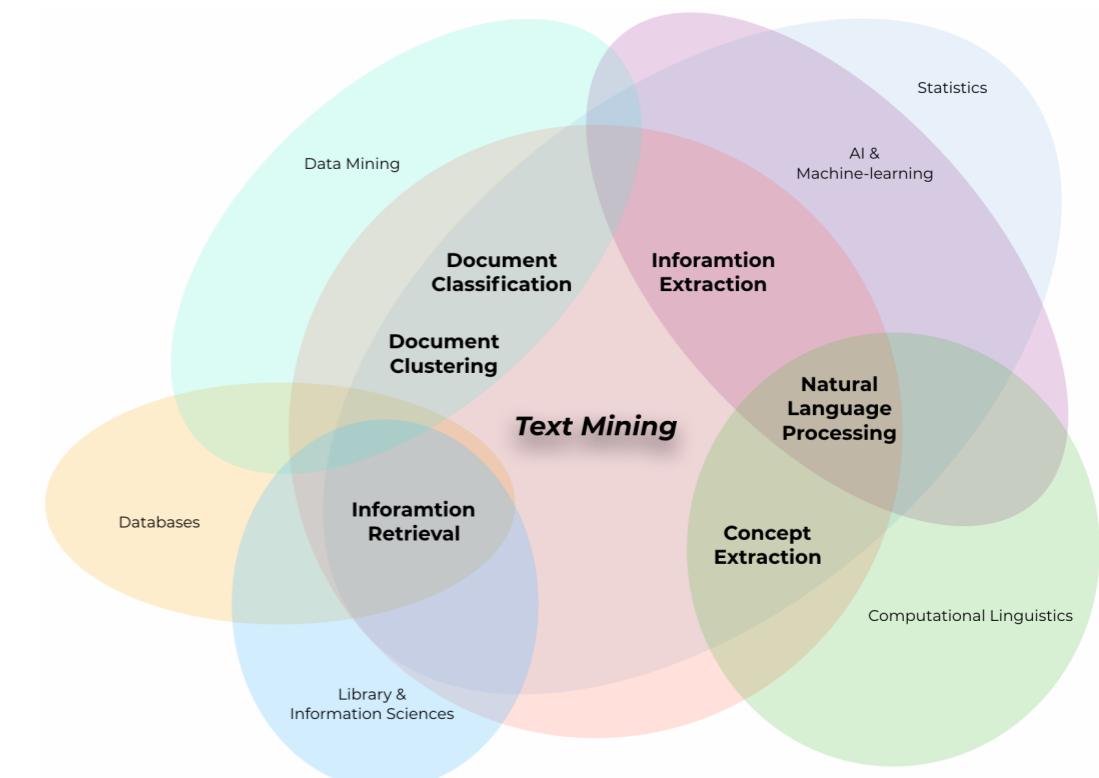


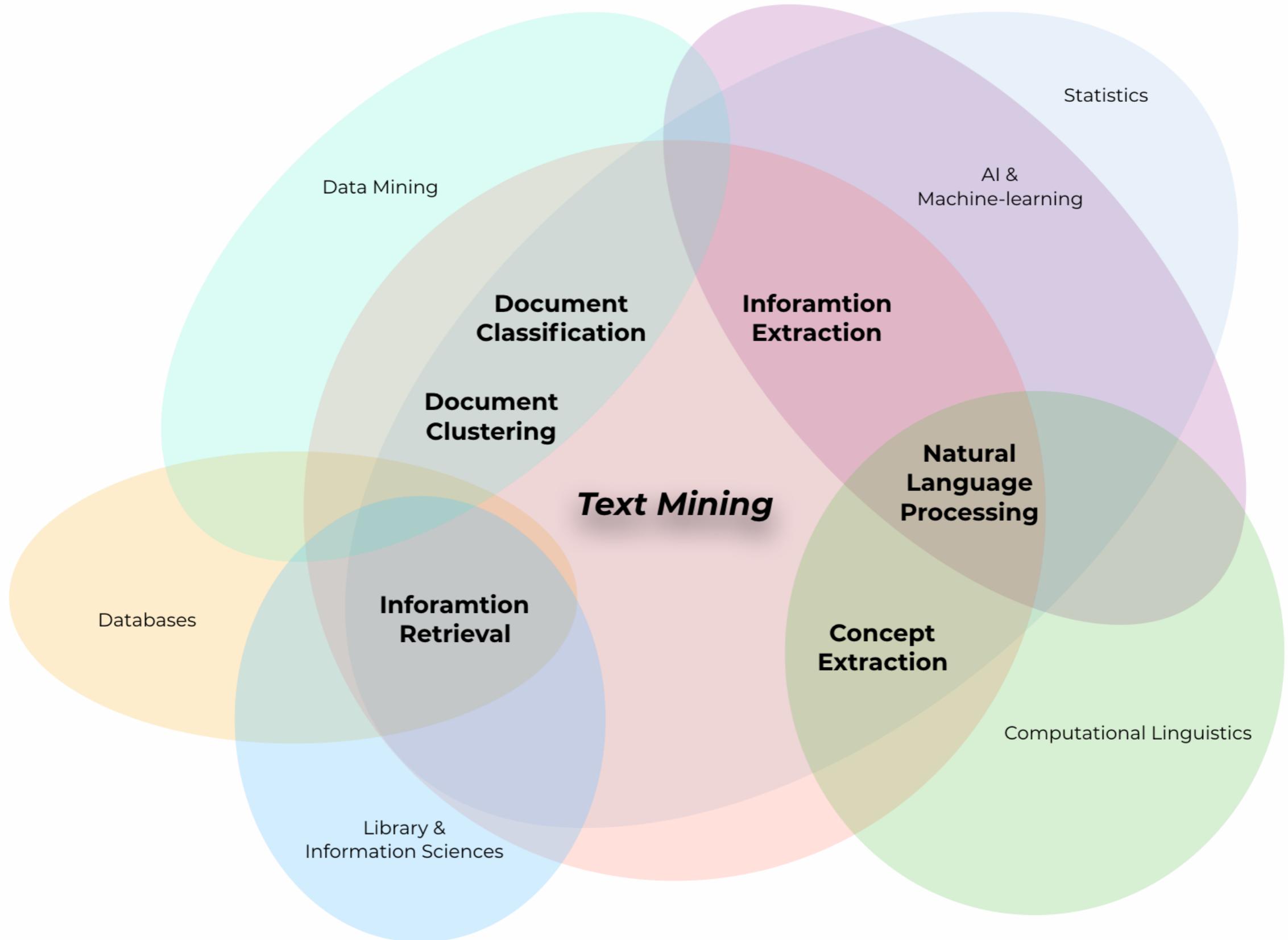


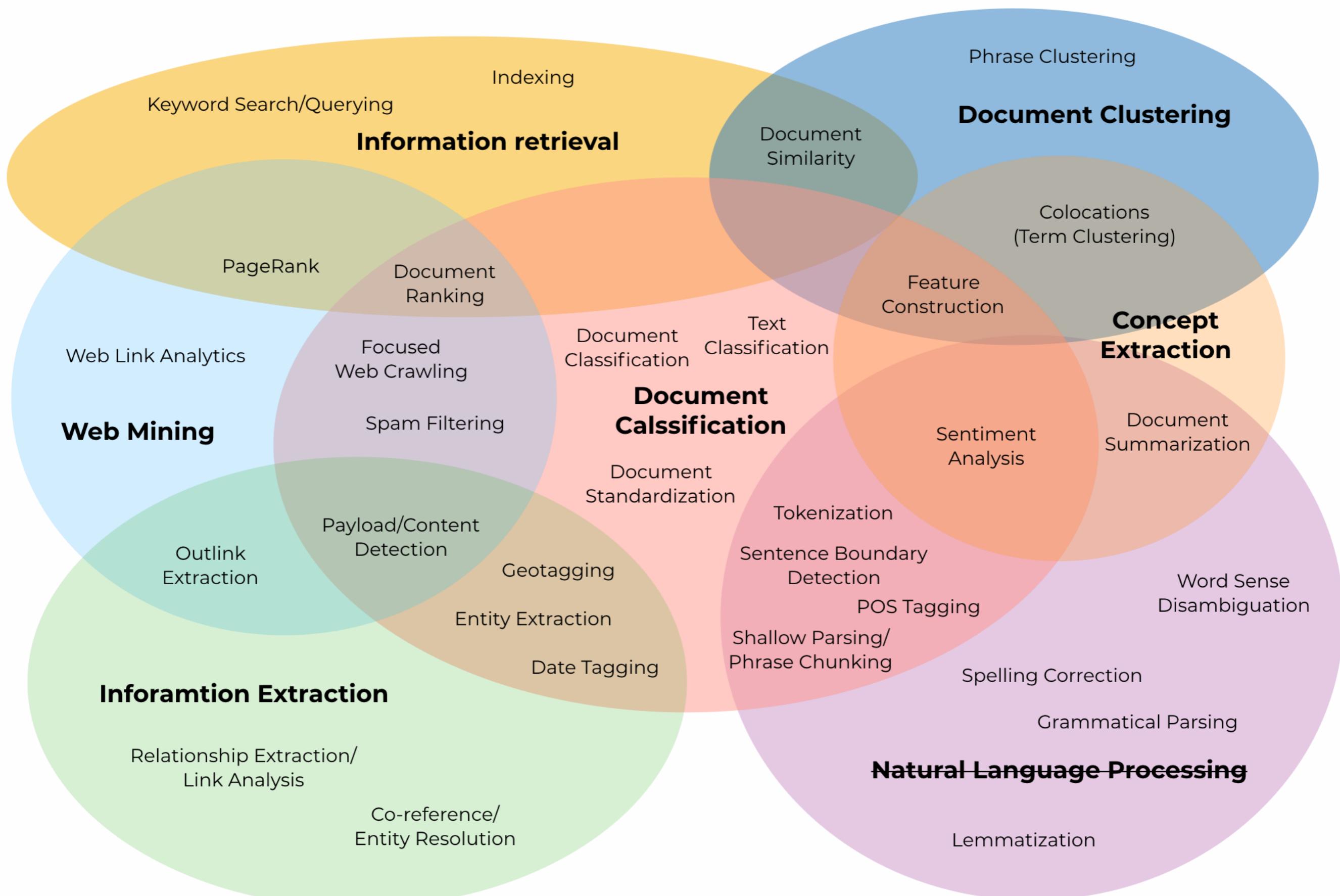
텍스트 마이닝

Text mining 이란?

- ▶ 대량의 텍스트 데이터셋에서 **흥미로운 규칙들을** 찾아내는 것 (Usama Fayyad)
- ▶ 문자로 된 자료들로부터 자동으로 정보를 추출하는, 이전에 알려지지 않은 **새로운 정보의 발견** (Marti Hearst)
- ▶ 텍스트 데이터를 이용하여 자연어 처리 (Natural Language Processing, NLP) 기술을 바탕으로 문서 속의 **유의미한 패턴** 또는 **유용한 지식**을 추출하는 과정
- ▶ 초기에는 언어학과 통계 기반에서 머신러닝을 통해 기계가 언어의 언어학적, 통계적 특징을 학습하는 형태로 발전하여 활용되고 있음



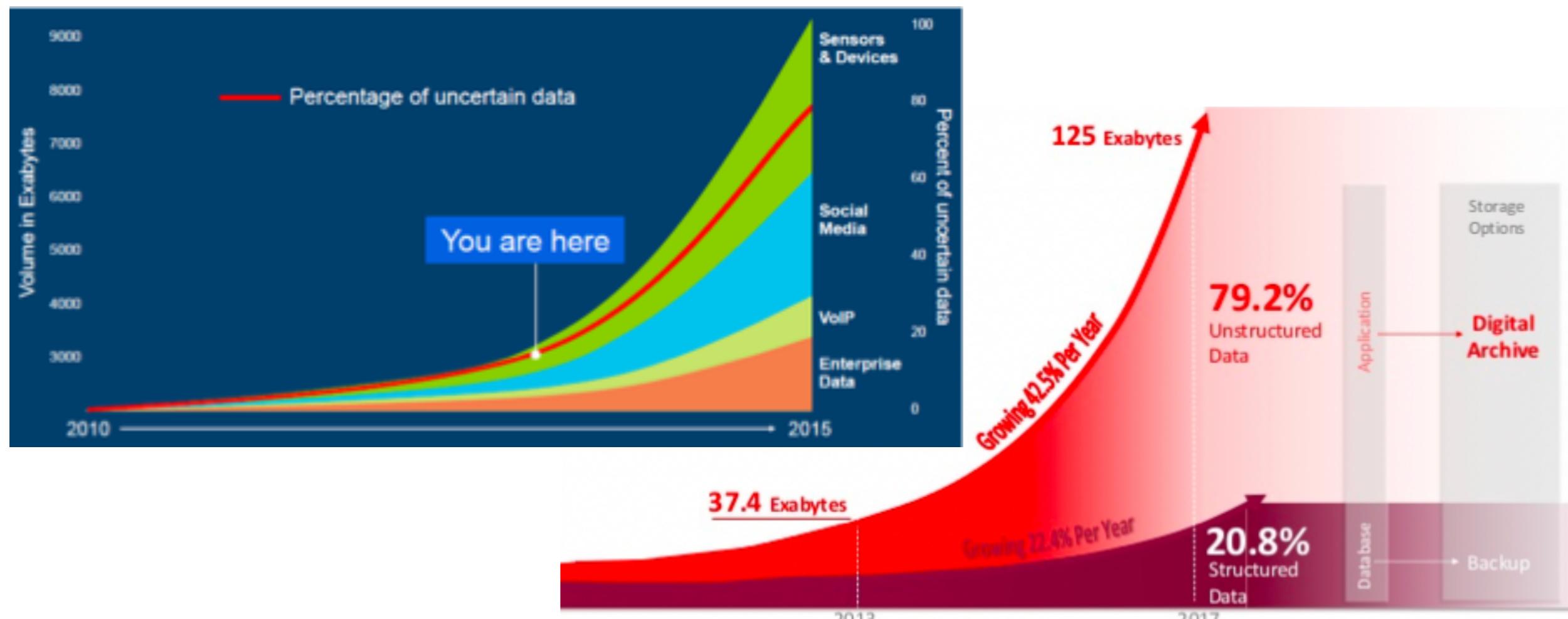




텍스트 마이닝이 중요한 이유

비정형 데이터의 증가와 분석 기술 발전

- ▶ 생산되는 전체 데이터의 70~80%가 비정형 데이터
- ▶ 스마트폰, SNS의 발전으로 인한 데이터 생성속도 증가
- ▶ 4차 산업혁명 관련 기술 (인공지능, 5G, 빅데이터 등)의 발전으로 인한 분석기술 향상
- ▶ 미래 관련 잠재적 가치를 갖고 있는 데이터의 분석을 통해 유용한 정보를 추출/활용하는 작업이 중요



* Source : Nadkarni, A., and Yezhkova, N., Structured versus unstructured data: The balance of power continues to shift, IDC (Industry Development and Models), 2014.3.17., https://issuu.com/reportlinker/docs_structuredversusunstructureddatathebalanceofpower/.

** Source : Larry Dignan, IBM eyes China, South America, Africa and big data for 2015 growth, 2013.2.28., <http://www.zdnet.com/article/ibm-eyes-china-south-america-africa-and-big-data-for-2015-growth/>.

텍스트 마이닝 현 주소

거의 해결됨 : **Easy**

- ▶ 스팸메일 분류 - Spam Detection
- ▶ 품사 태깅 - Part-of-speech (POS) tagging
- ▶ 개체명 인식 - Named entity recognition (NER)

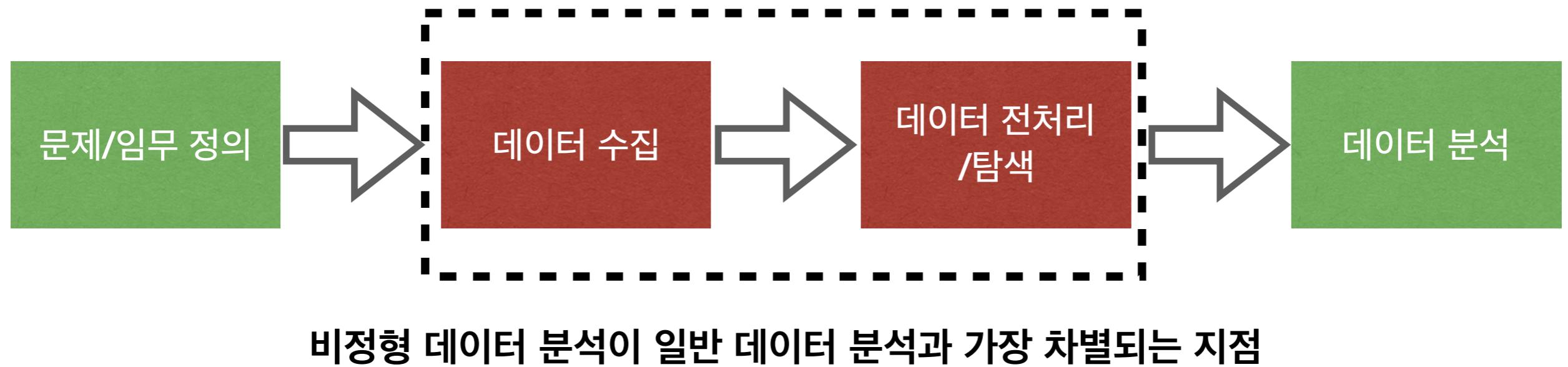
여전히 어려움 : **Hard**

- ▶ 질의응답 - Question answering (QA)
- ▶ 바꿔쓰기 - Paraphrase
- ▶ 요약 - Summarization
- ▶ 대화 - Dialog

좋은 성과를 보여주고 있음 : **Okay**

- ▶ 감성분석 - Sentiment analysis
- ▶ 동일 지시어 분석 - Coreference resolution
- ▶ 단어 의미 중의성 해소 - Word sense disambiguation (WSD)
- ▶ 파싱 - Parsing
- ▶ 기계번역 - Machine translation (MT)
- ▶ 정보추출 - Information extraction (IE)

Data Mining vs. Text Mining



텍스트 데이터 수집

텍스트 데이터 수집 Overview

- ▶ 이미 수집된 데이터 다운로드
- ▶ API를 통한 데이터 획득
- ▶ 웹 크롤링/스크래핑을 이용한 데이터 획득
- ▶ 오프라인에서 직접 획득
- ▶ 음성에서 텍스트 추출 (speech-to-text)

텍스트 데이터 전처리

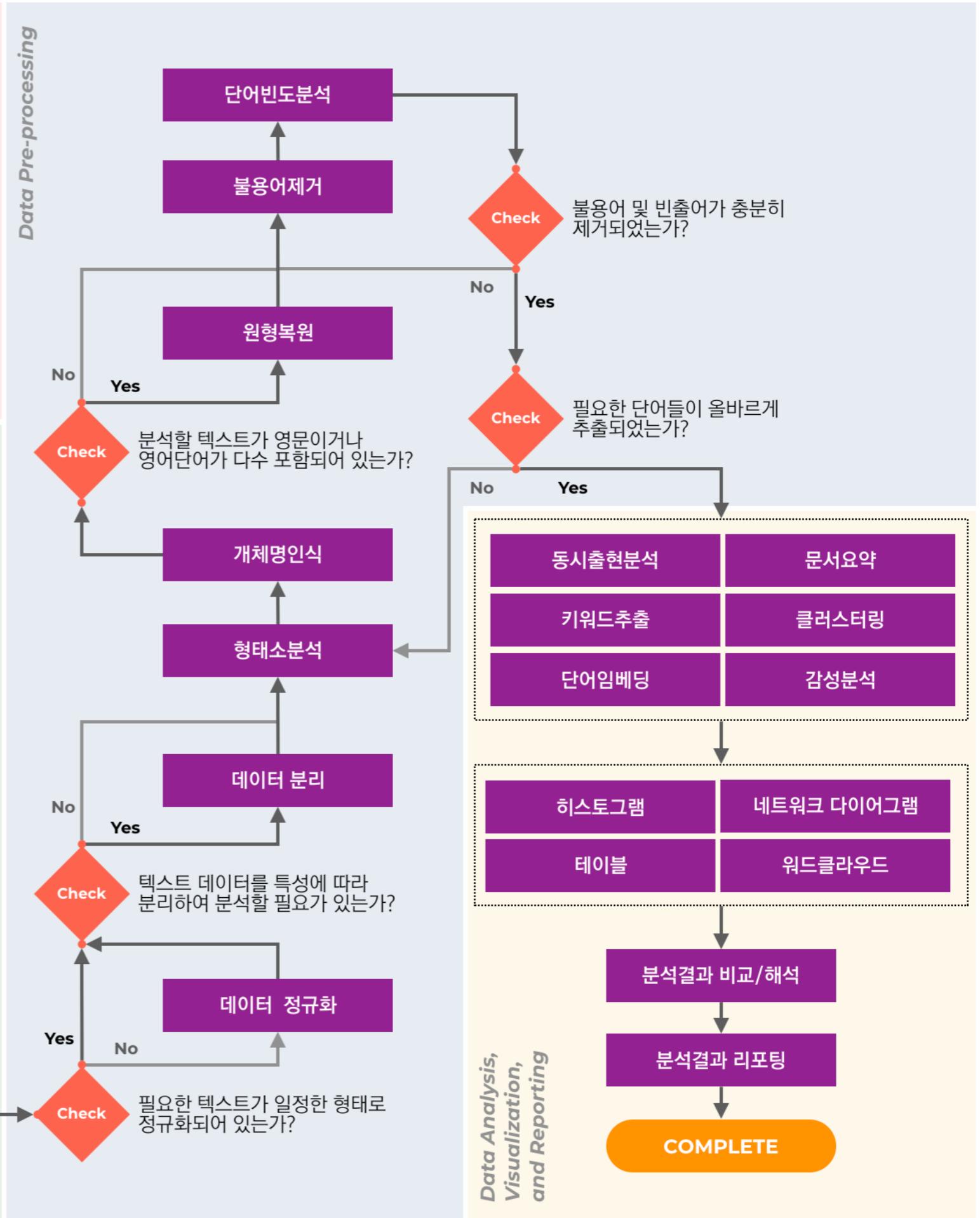
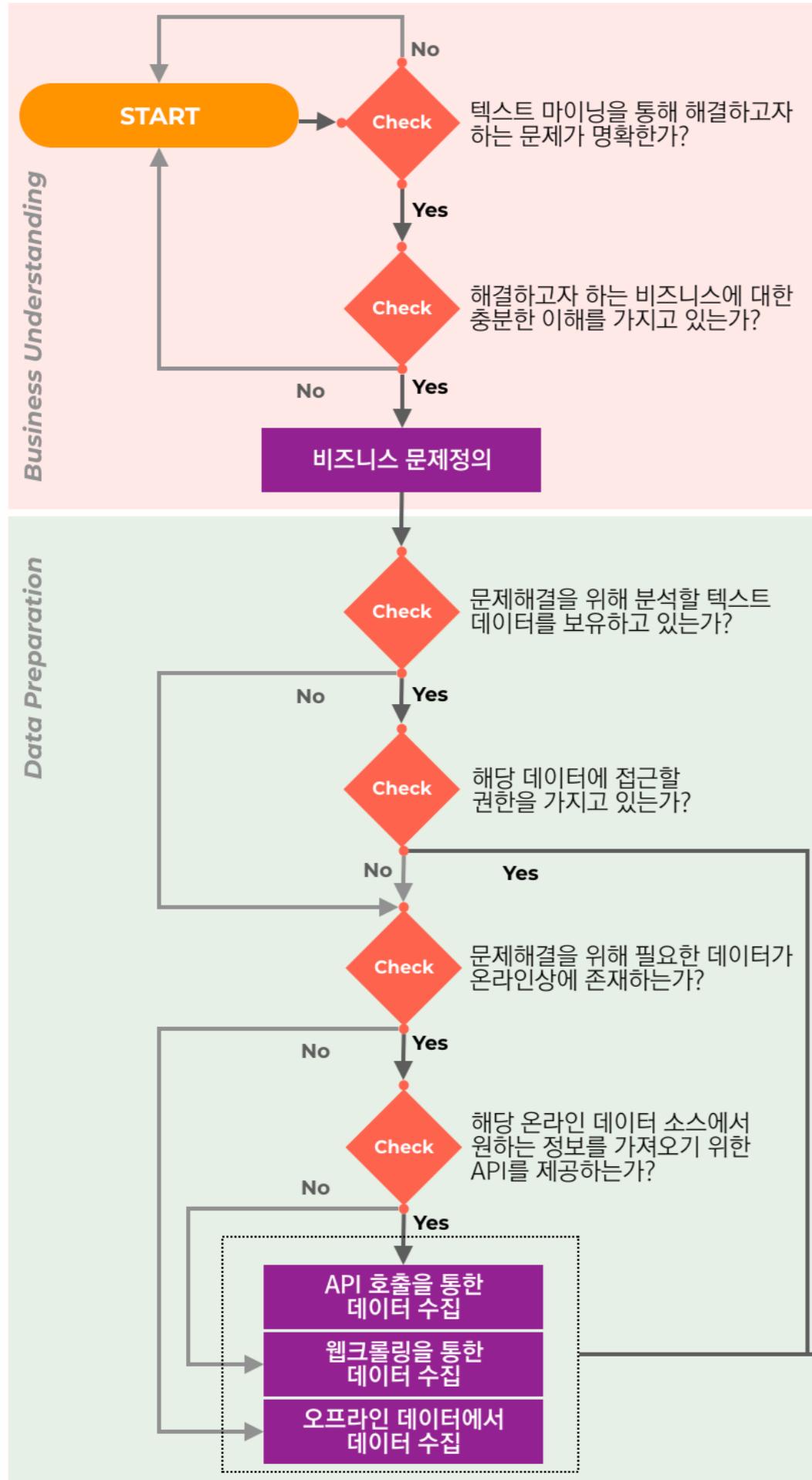
텍스트 데이터 전처리 Overview

- ▶ 형태소 분석
- ▶ 개체명 인식
- ▶ 정규화 (Normalization)
- ▶ 토큰화 (Tokenization)
- ▶ 어간 추출 (Stemming)
- ▶ 표제어 추출 (Lemmatization)
- ▶ 불용어 제거(Stopword removal)

텍스트 마이닝 기법

텍스트 마이닝 기법 Overview

- ▶ 자연 언어 이해 (Natural Language Understanding)
- ▶ 토픽 모델링(Topic Modeling)
- ▶ 감성 분석(Sentiment Analysis)
- ▶ 문서 분류(Document Classification)
- ▶ 집단화(Clustering)



E.O.D