

TEXT MINING for BEGINNER

Part 1. 텍스트 마이닝 개요 알아보기

전병진 FINGEREDMAN (fingeredman@gmail.com)

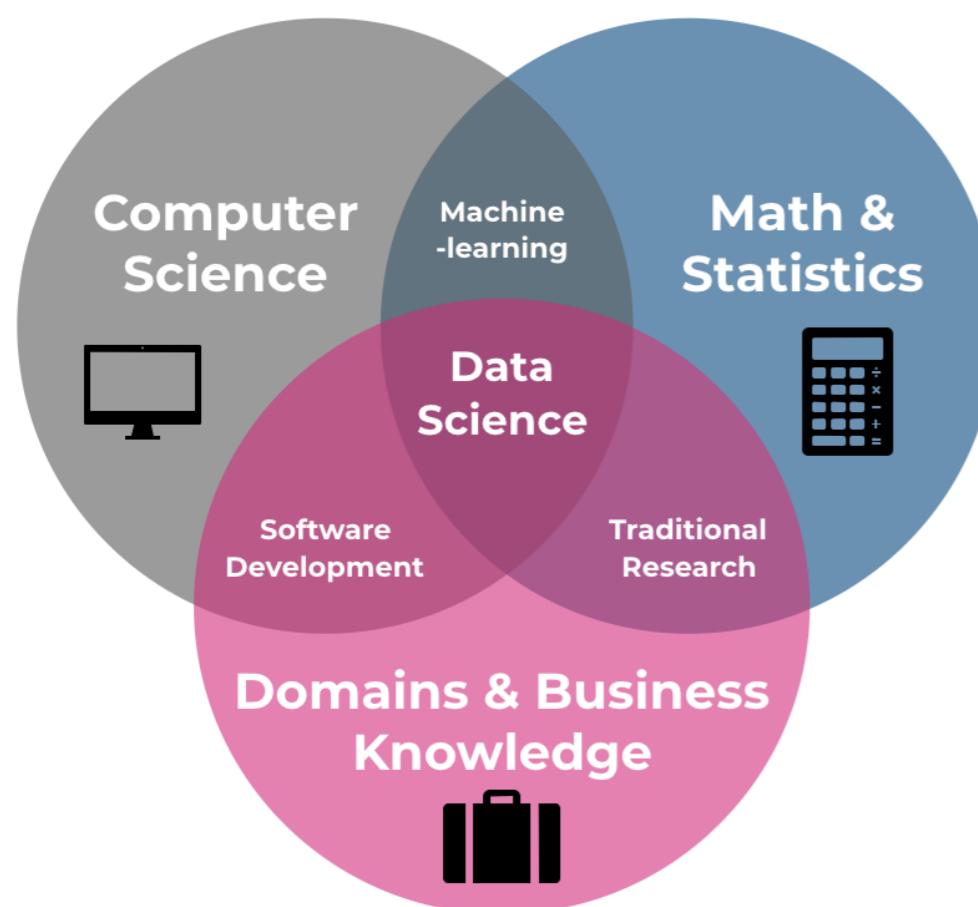
Part 1.

텍스트 마이닝 개요 알아보기

Data Scientist

데이터 과학자 (Data Scientist)가 하는 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
Multivariable Calculus & Linear Algebra	L	H	L	M

* Importance : H > M > L



Data Scientist

01

Generalists

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

02

Industry specialists

- ▶ 데이터 분석적 사고를 통해 문제를 해결하고자 하는 도메인 전문가

03

Deep specialists

- ▶ 특정 데이터 분야에 전문지식을 가진 사람
- ▶ 컴퓨터 사이언스에 대한 이해 필요

04

Analytics developers

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

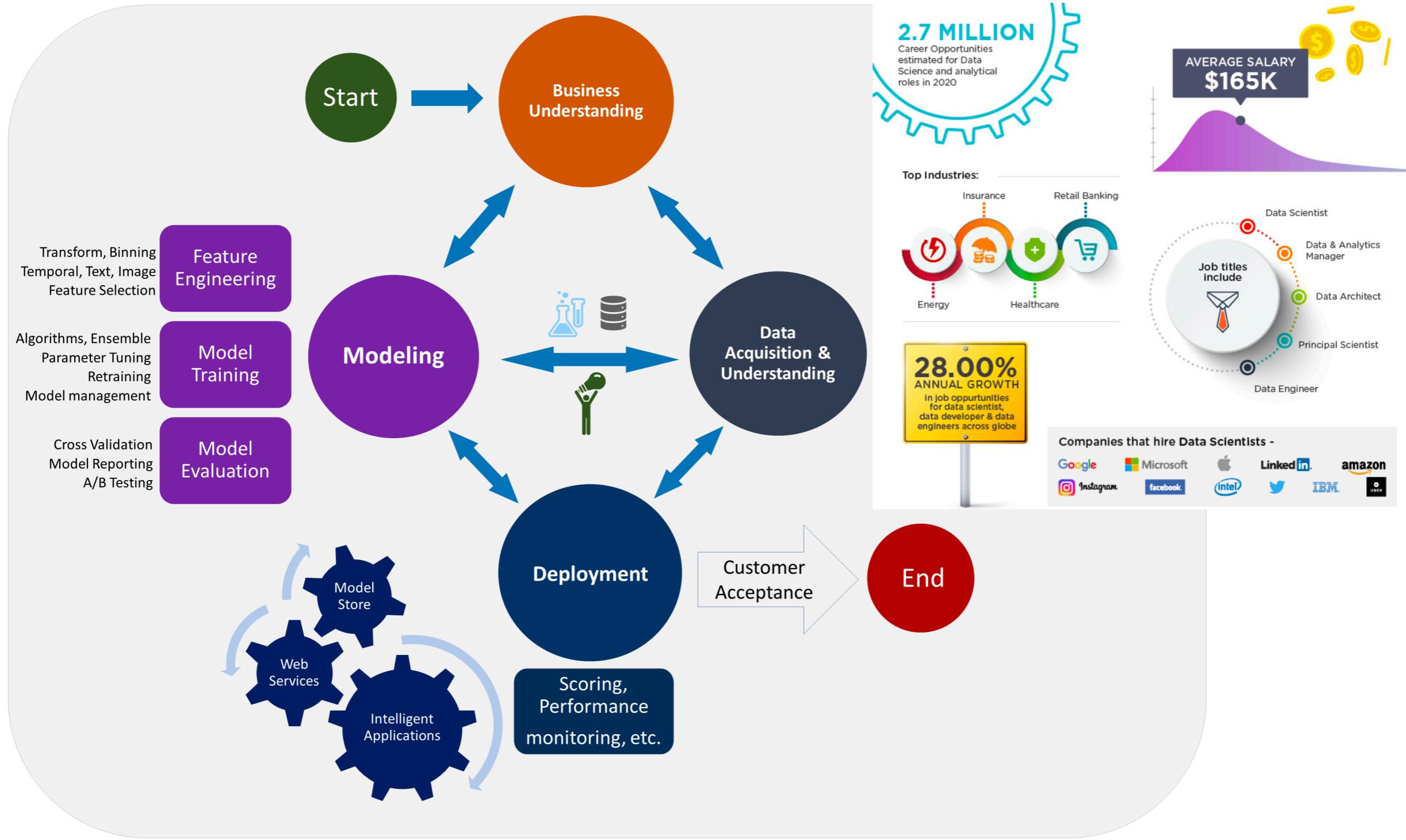
05

Data engineers

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

*Source : Nate Oostendorp, Radical Change Is Coming To Data Science Jobs, 2019.3.1., <https://www.forbes.com/sites/forbestechcouncil/2019/03/01/radical-change-is-coming-to-data-science-jobs/>.

Data Scientist Lifecycle



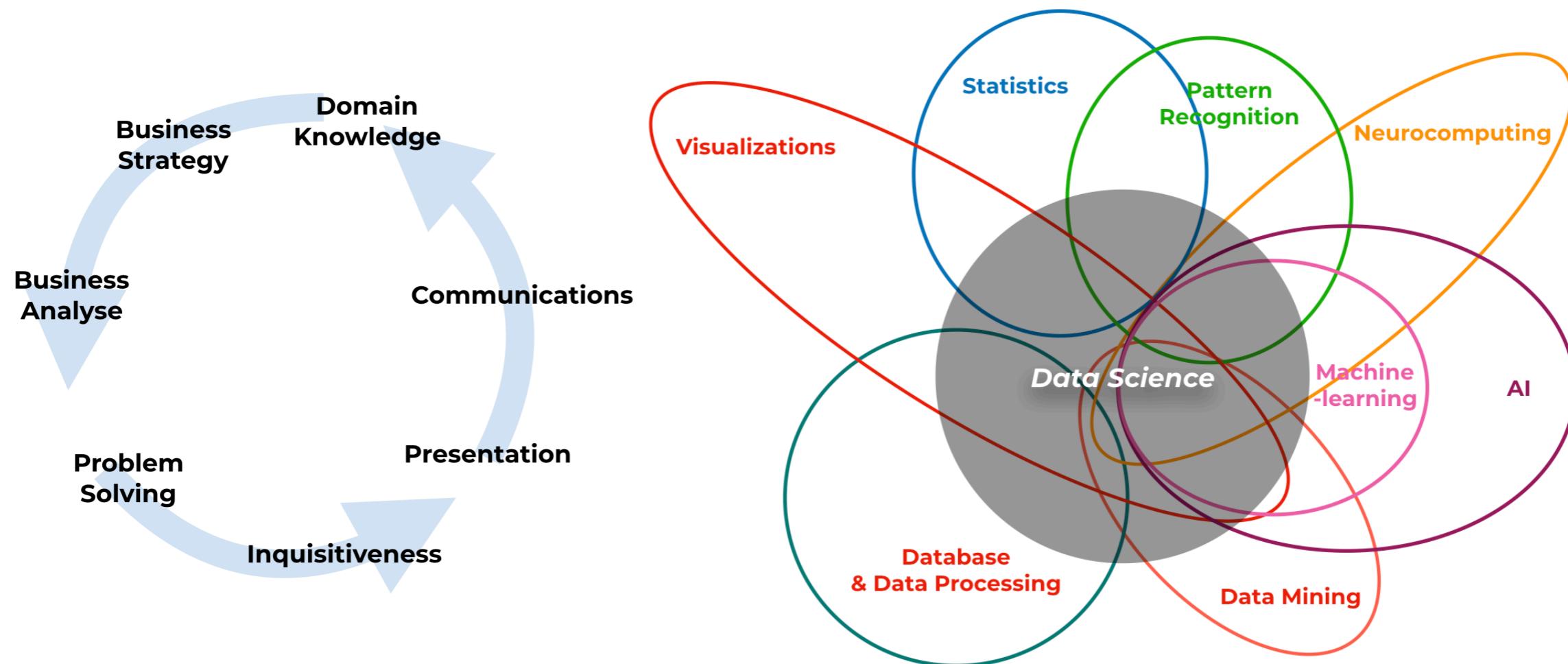
*Source : Gary Ericson et al., What is the Team Data Science Process?, 2017.10.20., <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview/>.

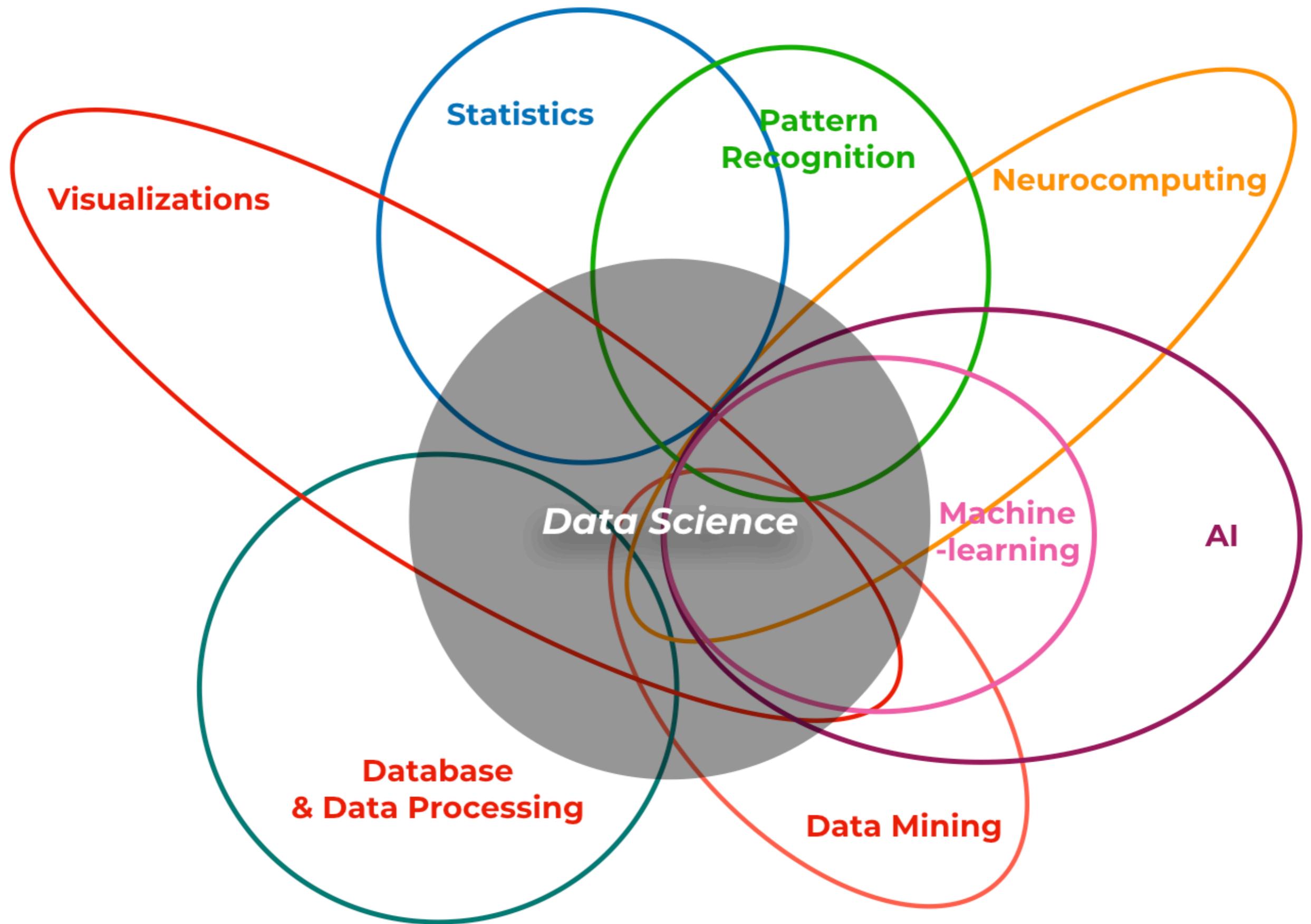
**Source : Simplilearn Solutions, About the program, <https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training/>.

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”

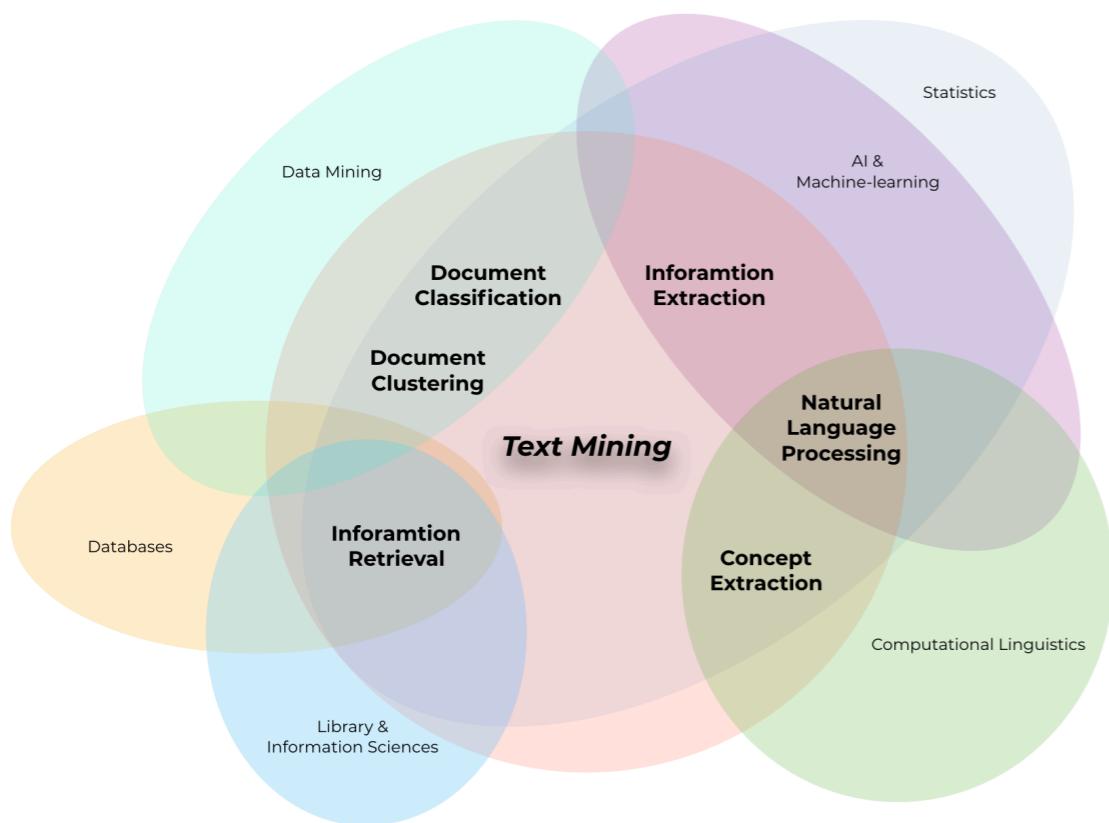




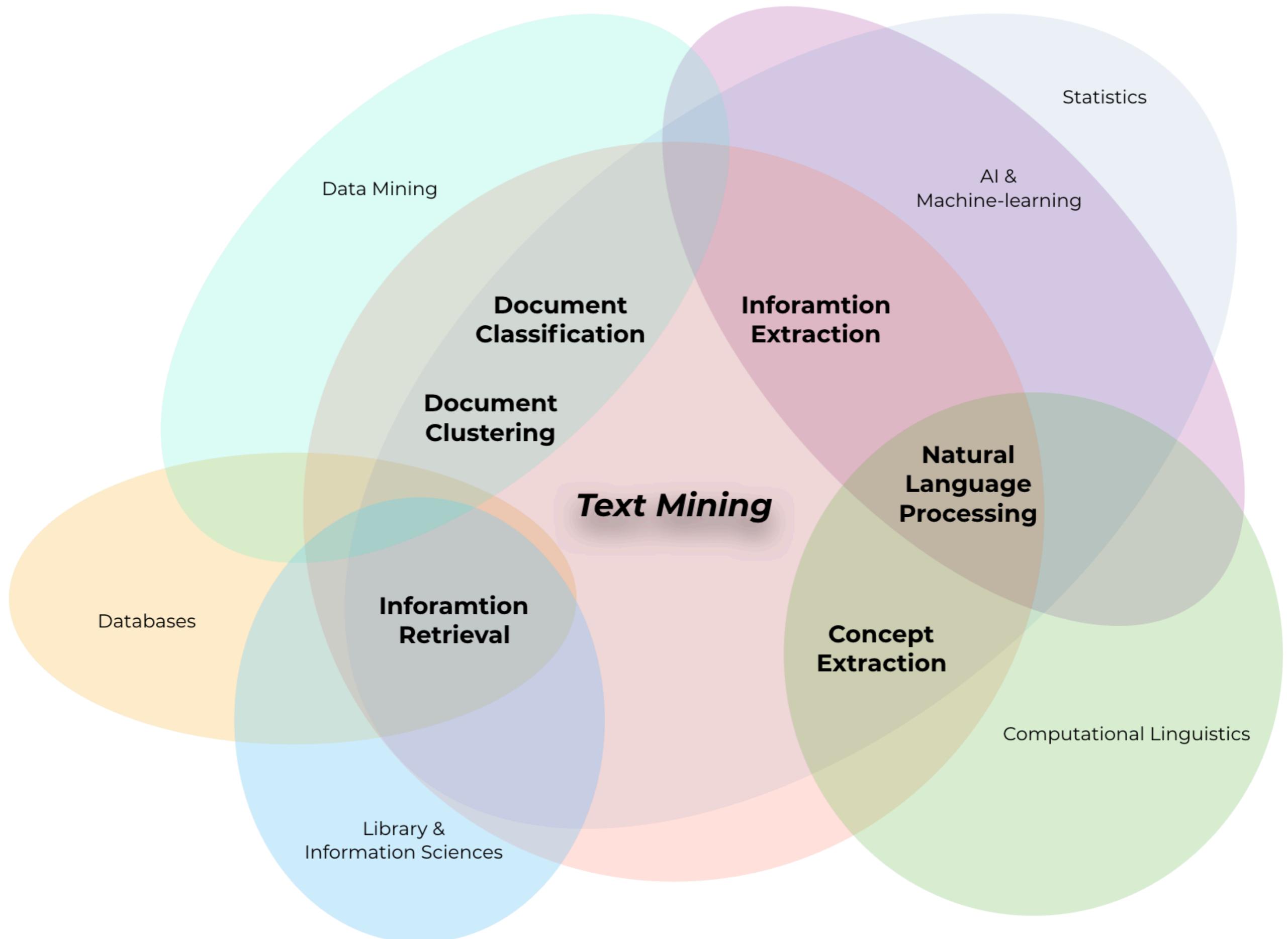
텍스트 마이닝

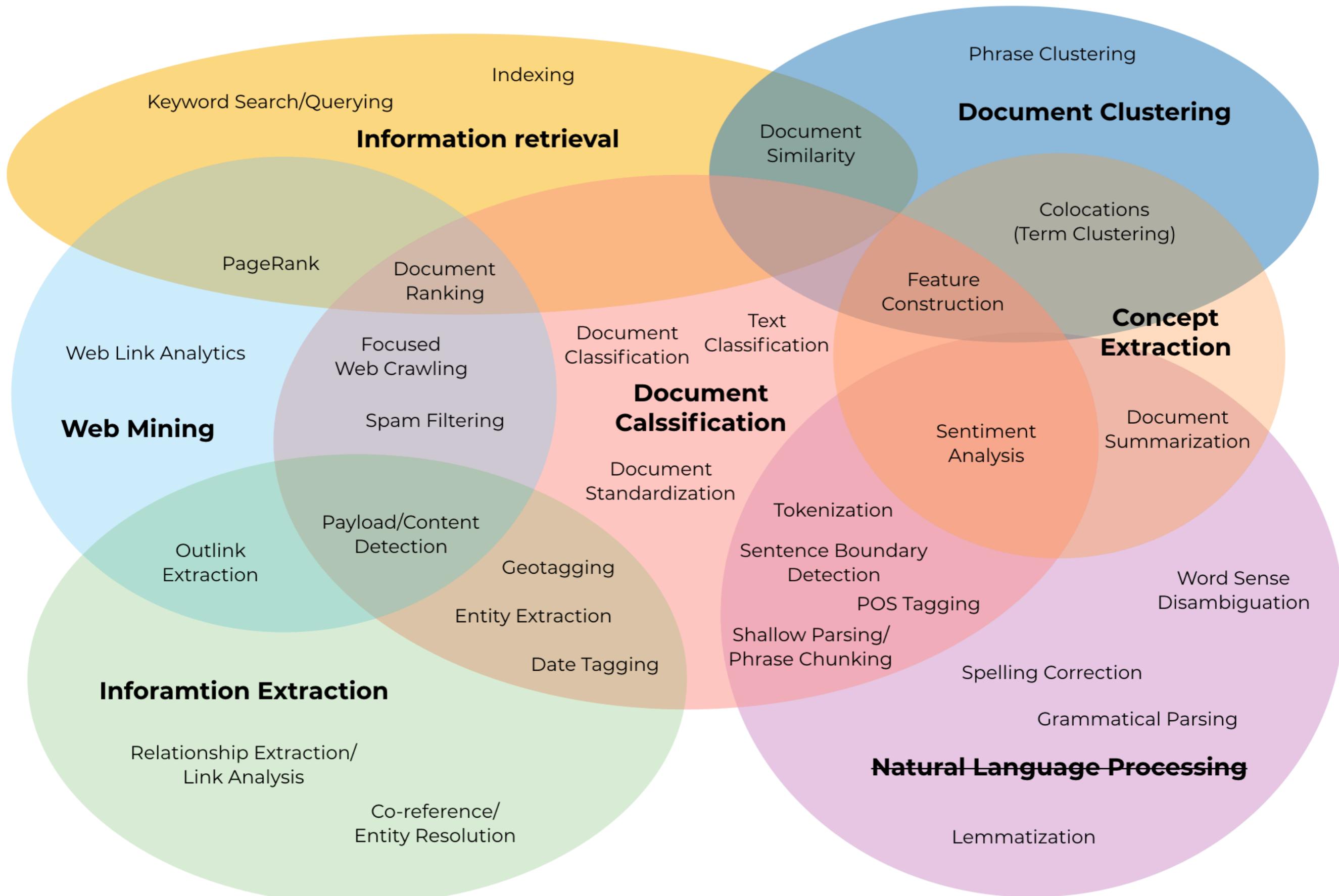
Text Mining 이란?

- ▶ 텍스트 데이터를 이용하여 자연어처리 (Natural Language Processing, NLP) 기술을 바탕으로 문서 속의 유의미한 패턴 또는 유용한 지식을 추출하는 과정
- ▶ 초기에는 언어학과 통계 기반에서 머신러닝을 통해 기계가 언어의 언어학적, 통계적 특징을 학습하는 형태로 발전하여 활용되고 있음
- ▶ Text Mining의 분류
 - Descriptive mining : 텍스트 집합에 있는 의미나 개념을 찾아내거나 이해를 돋는 형태 (문서분류, 정보검색, 단어빈도분석)
 - Predictive mining : 텍스트에 내포된 정보를 의사결정에 활용하는 형태 (질문 자동답변, 재구매자 예측)



텍스트 마이닝 유형	활용분야	
	실무	연구
<ul style="list-style-type: none">• 검색 (Information Retrieval)• 분류 (Classification)• 군집화 (Clustering)• 웹마이닝 (Web Mining)• 정보추출 (Information Extraction)• 개념추출 (Concept Extraction)• 자연어처리 (NLP)	<ul style="list-style-type: none">• 스팸 필터링• 이슈 검출/트래킹• 정보검색• 자살률 예측• 주가 예측• 소비자 인식 조사• 경쟁사 분석	<ul style="list-style-type: none">• 사회동향 분석• 이슈 트래킹• 온라인 행동 분석• 연구분야 탐색• 질병관계 예측• 정책전략 수립





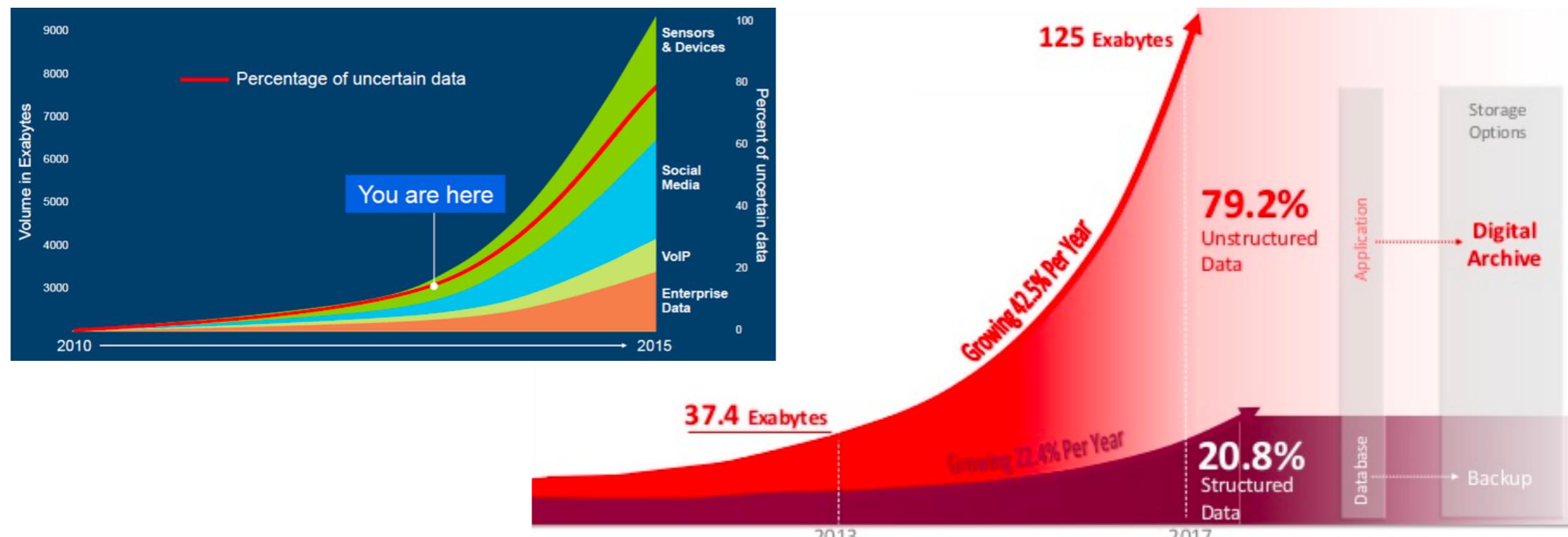
텍스트 마이닝이 중요한 이유

비정형 데이터의 폭발적 증가

- ▶ 미래 관련 잠재적 가치를 갖고 있는 비정형 데이터 (unstructured data)들이 대규모로 생성됨과 동시에 비정형 데이터 속에서 미래의 의사결정에 관련된 유용한 정보를 찾아내어 활용하는 작업이 매우 중요해짐
- ▶ 생산되는 비정형 데이터의 70~80%가 비정형 데이터(기사, 블로그, 문서, 보고서 등)

텍스트 데이터의 폭발적 증가

- ▶ 인터넷과 소셜 네트워크 서비스 (Social Network Service, SNS)를 통한 온라인 양방향 커뮤니케이션의 활성화
- ▶ 4차산업혁명과 사물인터넷 (Internet of Things, IoT) 등 빅데이터 관련 기술의 급진적인 발전



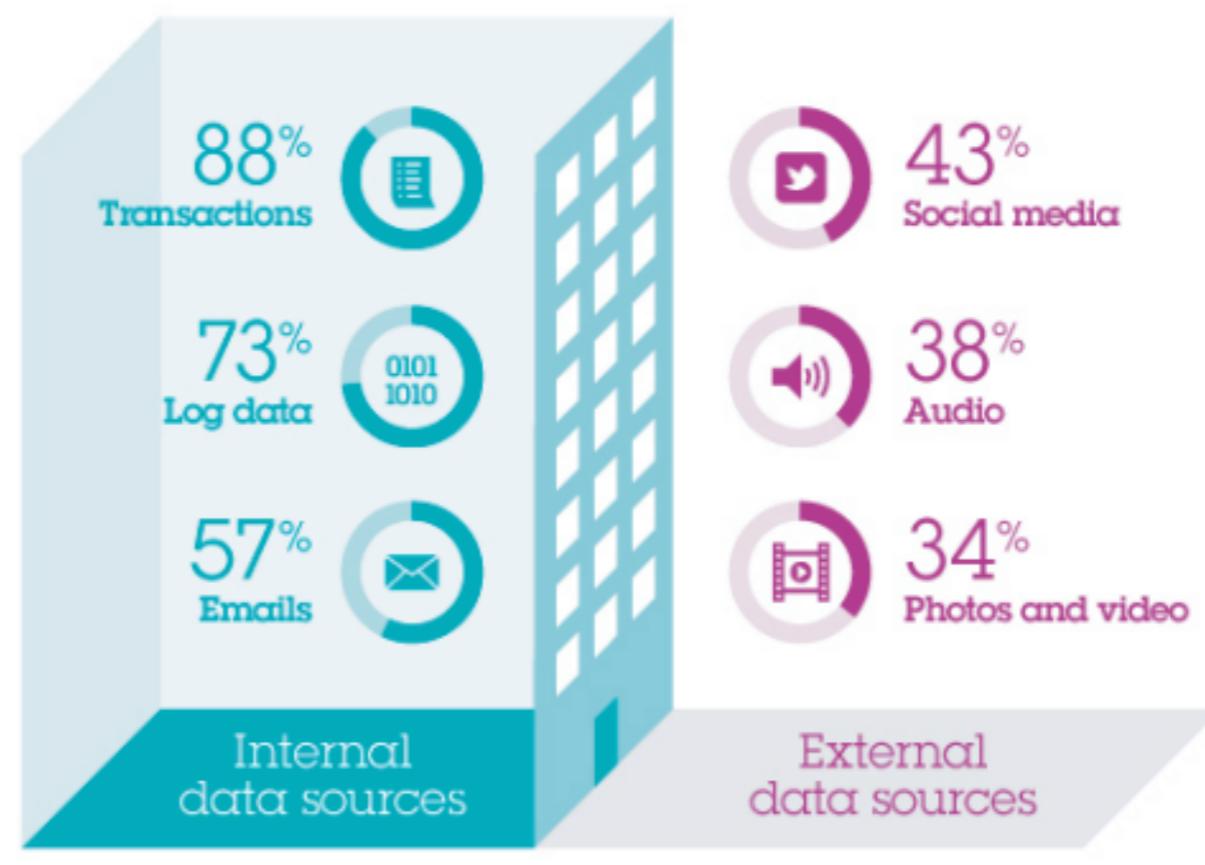
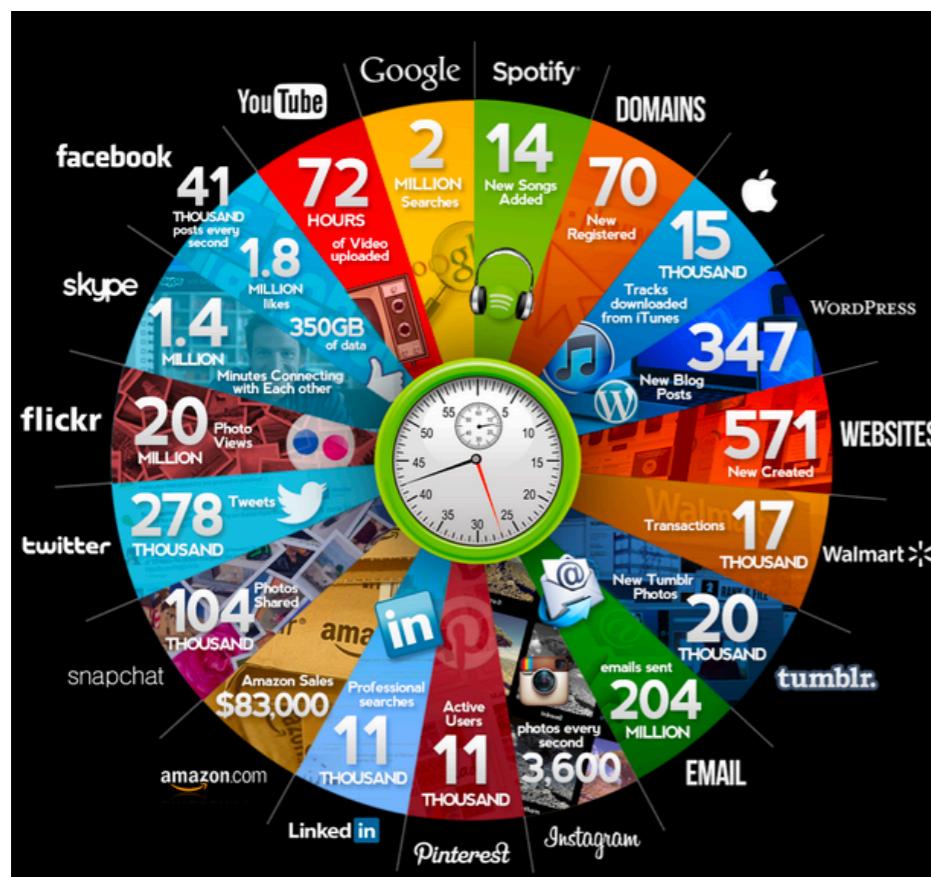
*Source : Nadkarni, A., and Yezhkova, N., Structured versus unstructured data: The balance of power continues to shift, IDC (Industry Development and Models), 2014.3.17., <https://issuu.com/reportlinker/docs/structuredversusunstructureddatathebalanceofpower/>.

**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 마이닝이 중요한 이유

가장 흔하고 접하기 쉬운 데이터

- ▶ 텍스트가 없는 곳은 없으며, 다양한 서비스에 수많은 텍스트가 존재
- ▶ 온라인 비정형 텍스트 데이터의 많은 부분이 SNS에서 발생 (Twitter, Facebook, Instagram, YouTube, 블로그, 커뮤니티 등)
- ▶ 웹에서 사용자들이 컨텐츠를 생성하고 의사소통하는 것은 대체로 텍스트 데이터임
- ▶ 다양한 형태의 비정형 데이터(오디오, 비디오)가 텍스트 형태로 변형되어 활용 (Speech to Text, STT)
- ▶ Web Crawling, Open API 등의 활성화로 텍스트 데이터 수집 및 확보가 용이해짐



*Source : Ashley Feinberg, How Much Happens on the Internet Every 60 Seconds?, 2013.7.29., <https://gizmodo.com/how-much-happens-on-the-internet-every-60-seconds-950463150/>.

**Source : IBM, Where does big data come from Infographic, 2012.10.17., <https://www-03.ibm.com/press/us/en/photo/39145.wss/>.

텍스트 마이닝이 어려운 이유

가장 분석하기 어려운 데이터

▶ 언어적 한계점

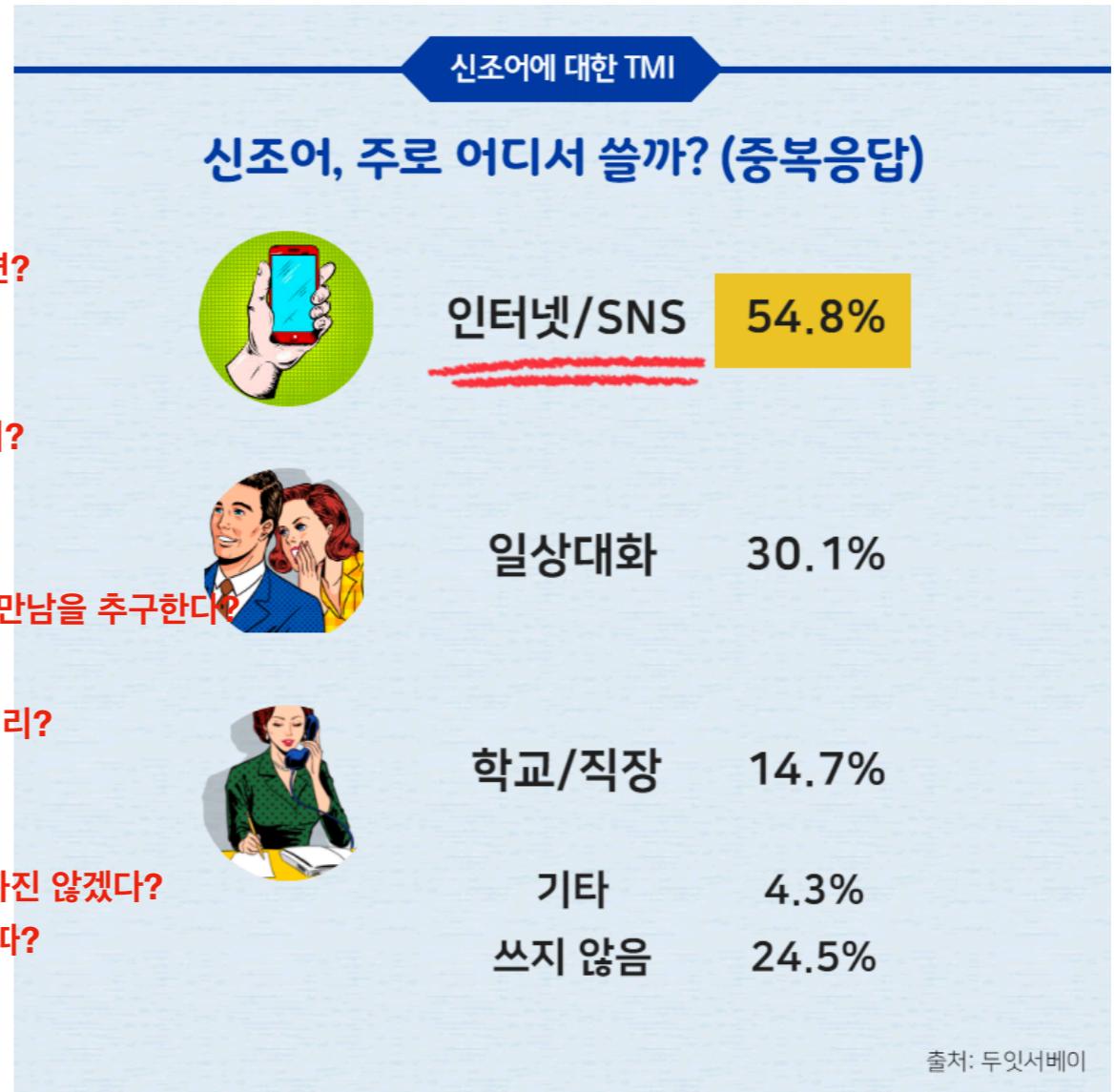
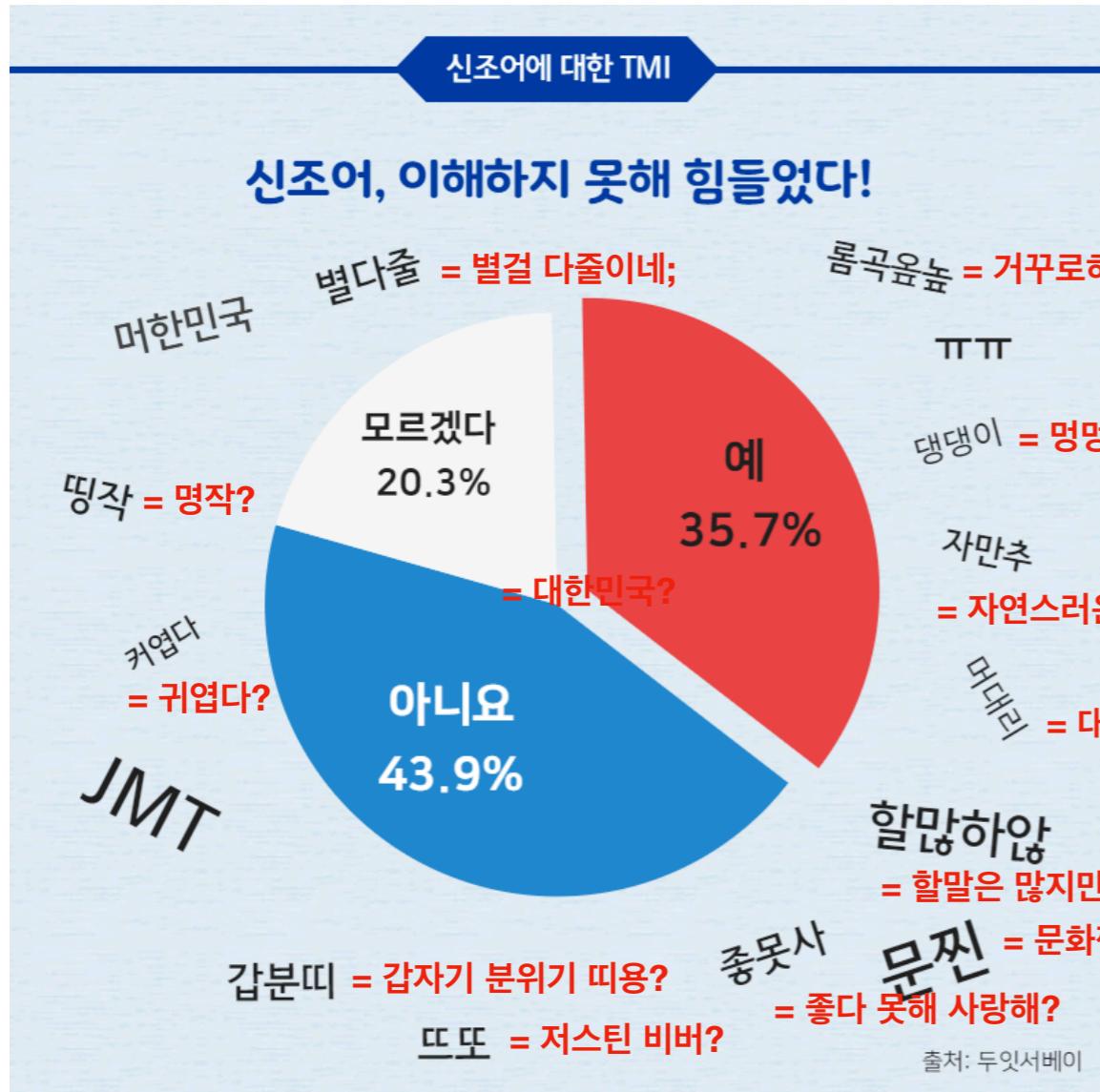
- 사람들이 작성한 문장은 맞춤법과 철자가 틀리고, 단어를 섞어 쓰고, 축약되고, 구두점을 아무데나 찍히는 등 규칙을 지키지 않음
- 동의어, 동형(동음) 이의어가 포함되거나 약어의 의미가 분야별로 다를 수 있음
- 문맥에 따라서 의미가 많이 달라지며, 애매한 표현이 많이 나타남 → 추상적 개념의 모호함

▶ 데이터적 한계점

- 텍스트는 비정형 데이터로서, 일반적인 필드와 레코드 구조를 가지고 있지 않음 → 전처리가 상당히 많이 필요함
- 텍스트의 형태와 특징에 따라 전처리 과정과 분석방법에 대한 접근을 다르게 고려해야함
- 자연어처리에 대한 이해가 필요하고, 분석에 오랜 시간이 소요되어 큰 잠재적 가치에도 불구하고 충분히 활용하지 못하고 있음
- 방대한 양, 데이터의 규모 증가, 그리고 그 형태의 비정형성으로 인하여 그 분석과 활용이 어려움

한계점	예시
오탈자	<ul style="list-style-type: none">• “헝거게임 잼잇써요완전 대신 이전편꼭바여”• “슬까 타노스 보석 하나도 못구했을때 다들 머했음? 3개 얻었을때도 그렇게 안쌔 봄더만...”
동의어, 동음이의어	<ul style="list-style-type: none">• 한혜진 : 1. 모델 한혜진(달심), 2. 배우 한혜진(기성용 부인), 3. 가수 한혜진(트로트 가수)• Close : 1. Opposite of open, 2. A preposition meaning not far• IS : 1. Information System, 2. Islamic State, 3. International Standard
모호한 문장	<ul style="list-style-type: none">• “이 영화의 1부는 2부보다 훨씬 좋다. 연기는 빈약하고 마지막에서는 통제할 수 없을 정도가 되며, ...”
전처리	<ul style="list-style-type: none">• 분석 데이터의 언어를 파악하고 언어의 특징(교착어, 굴절어 등)에 맞는 전처리 작업 진행• 댓글 단위로 분석할지, 문장 단위로 분석할지에 따라서 데이터 분리작업 진행
정보추출	<ul style="list-style-type: none">• 해시 태그(Hash Tag)의 추출 : '#' + (문자)• 핸드폰 번호 추출 '010' - (4자리 숫자) - (4자리 숫자)

텍스트 마이닝이 어려운 이유



텍스트 마이닝이 어려운 이유

비싸고 어려운 상용 애플리케이션/솔루션

- ▶ 상용 애플리케이션, 솔루션 또는 프로그래밍 언어 활용 중 선택이 필요함
- ▶ 애플리케이션 : SAS Text Miner, IBM Watson Explorer (WEX), ...
- ▶ 솔루션 : 소셜메트릭스, 버즈 인사이트, 트루스토리, ...
- ▶ 프로그래밍 언어 : Python, JAVA, R, ...

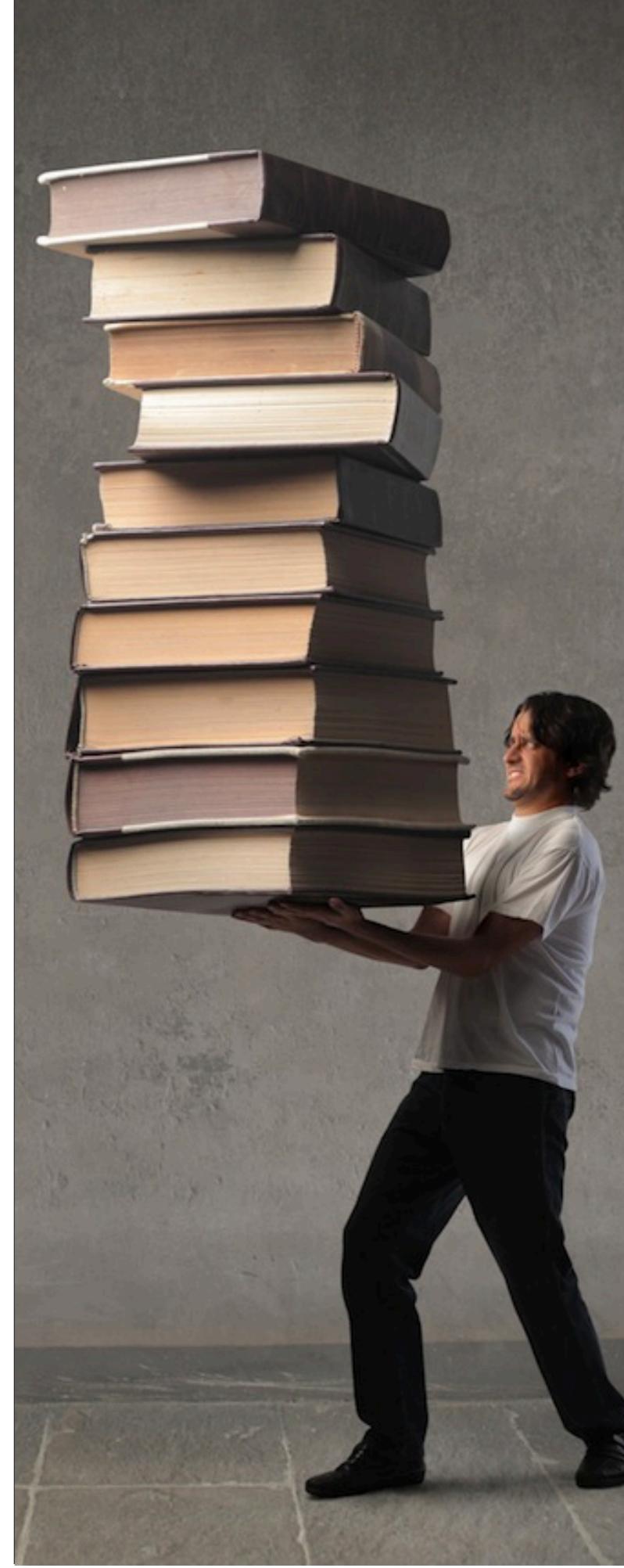
구분	Social Matrix	Watson Explorer (WEX)	SPSS Text Miner 17	Text Miner 12.1
개발사	다음소프트	IBM	IBM	SAS
제품유형	웹 솔루션	윈도우 애플리케이션	윈도우 애플리케이션	윈도우 애플리케이션
클라우드 환경 지원	○	○	-	X
데이터 소스	입력 데이터 제한	기간별 샘플에 한함	없음	없음
	데이터 소스 지원	웹 검색, 뉴스기사, Instagram, Twitter, Facebook	-	-
	데이터 수집 지원	X	X	X
분석	분석방법	<ul style="list-style-type: none">• 연관어 맵• 연관어 추이• 긍부정 연관어	<ul style="list-style-type: none">• 단어 빈도 및 상관관계• 트렌드 분석• 감성분석	<ul style="list-style-type: none">• 범주 및 개념 추출• 군집화• 텍스트 링크
활용 난이도	낮음	보통	높음	높음

텍스트 마이닝이 어려운 이유

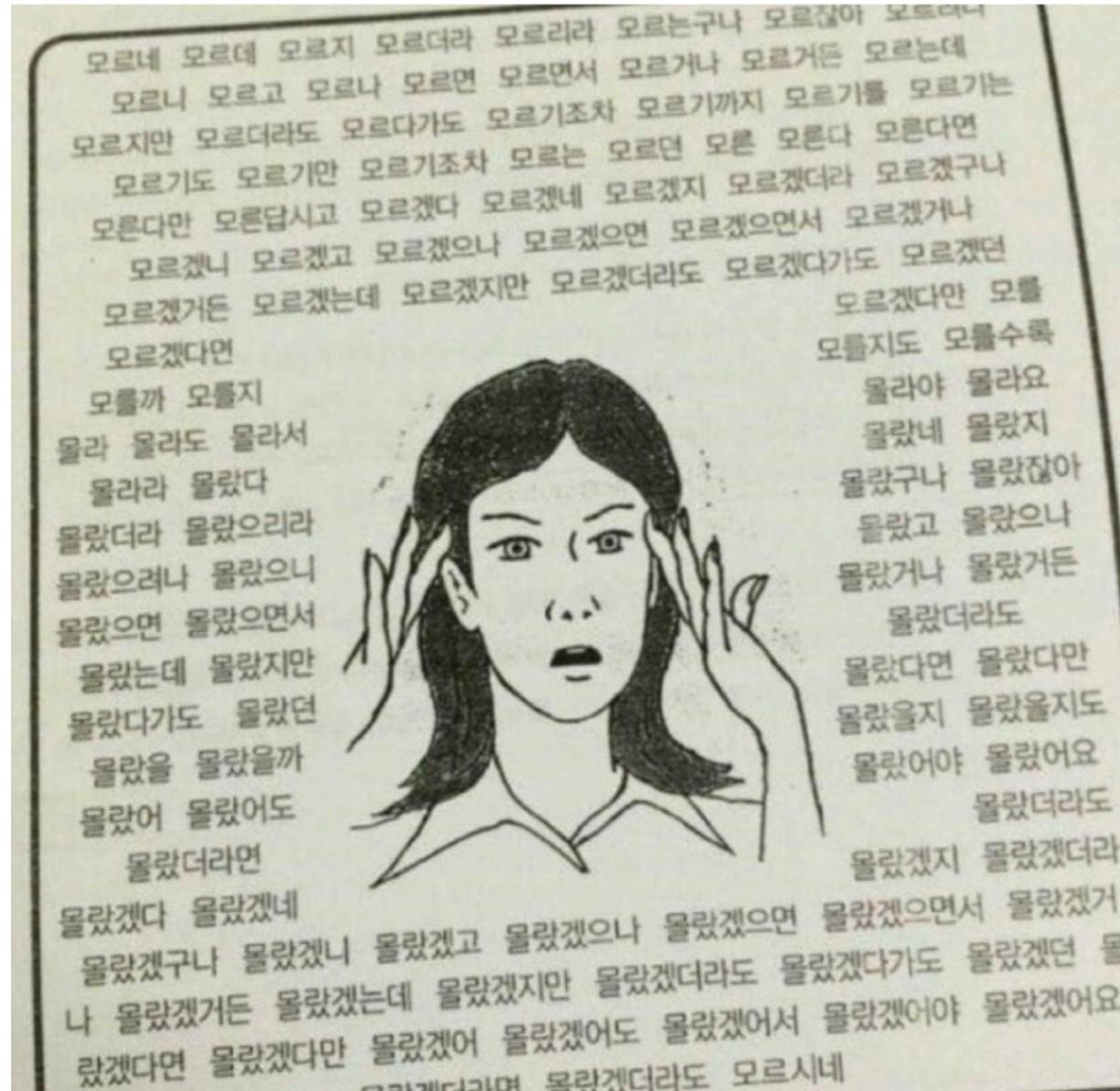
복잡한 한국어 텍스트 분석

- ▶ 한국어 텍스트 분석은 한국어의 언어학적 특성으로 인해 전처리와 분석과정이 까다로움
- ▶ 용언이 변하는 경우의 수가 매우 많고 그 과정이 하나의 개념으로 확인하기가 어려움
- ▶ 형태소 분석기의 한계 (한국어는 90% 이상 정확도를 가진 형태소 분석기를 만들기 어려움)
- ▶ 미비한 어휘 사전 구축 (신조어, 미등록어, 새로운 용어의 조합)

한계점	예시
용언의 변형	<ul style="list-style-type: none">• 모르다 → 모르네, 모르데, 모르지, 모르더라, 모르리라, 모르는구나, 모르니, 모르고, ...
형태소 분석	<ul style="list-style-type: none">• “비비크림 빠빠빠~립스틱을 마마마” → 비/NNG + 비/NNG + 크림/NNG + 빠빠빠/UN + ~/SO + 립스틱/NNG + 을/JX + 마/NNG + 마마/NNG• “황민현에게 트렌치코트는 정말 존멋♥” → 황민/NNG + 현/NNG + 에게/JKM + 트렌치/NNG + 코트/NNG + 는/JX + 정말/MAG + 줄/VV + ㄴ/ETD + 멋/NNG + ♥/SW• “자다가 퇴근했음 좋게따 외냐면 내일 사랑니 째러 가야 되니까는...” → 자/VV + 다가/ECD + 퇴근/NNG + 하/XSV + 었/EPT + 음/ETN + 좋/VA + 게/ECD + 따/VV + 아/ECS + 외/NNG + 이/VCP + 냐/EFQ + 면/NNG + 게/ECD + 사랑니/NNG + 째/VV + 러/ECD + 가/VV + 아야/ECD + 되/VV + ...
신조어 출현	<ul style="list-style-type: none">• 지카 바이러스, 오백따리, 트둥이, 울와(우리 트와이스), 팬코(팬 코스프레 유저)



텍스트 마이닝이 어려운 이유



I just got here.

상기 문장은 영어로 "나 막 도착했어" 가 된다. 자연스럽게 위 문장을 바꿀 수 있는 경우는

I have just arrived 하나 정도다.

한국어에서 저 just라는 표현은 대체 수십 가지로 가능하다.

나 막 왔어.

나 방금 왔어.

나 지금 왔어.

나 금방 왔어.

나 온 지 조금/좀 됐어. (조금에 강세)

나 온 지 별로/얼마 안 됐어.

나 이제 왔어.

나 바로 막 왔어.

게다가 위의 모든 표현의 '왔어'를 "도착했어"로 바꿔도 말이 된다.

- 시발ㅋ, 시발 ㅋㅋ : 웃김
- 오 시발 : 놀라움
- 아 시발 : 아쉬움
- 시발... : 슬픔
- 시발! : 분노
- 시발;; : 어이없음
- 시발ㅜㅜ : 격한슬픔
- 시발;;;; : 당황스러움
- 시바리 : 급함
- 시ㅂ : 더욱 급함
- tlqkf : 정말로 급함

*Source : 디지털이슈팀(조선일보), 외국인들이 꼽은 한국어가 어려운 이유는?, 2017.6.25., http://news.chosun.com/site/data/html_dir/2017/06/25/2017062500228.html.

**Source : 파란불(티스토리), [유머] 한국어를 배우기 어려운 이유, 2014.2.1., <http://bluelight.tistory.com/298/>.

***Source : 라연(네이버 블로그), 한국어가 정말 어려운 이유, 2015.11.4., <https://m.blog.naver.com/PostView.nhn?blogId= savedigi&logNo=220528855064&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F/>.

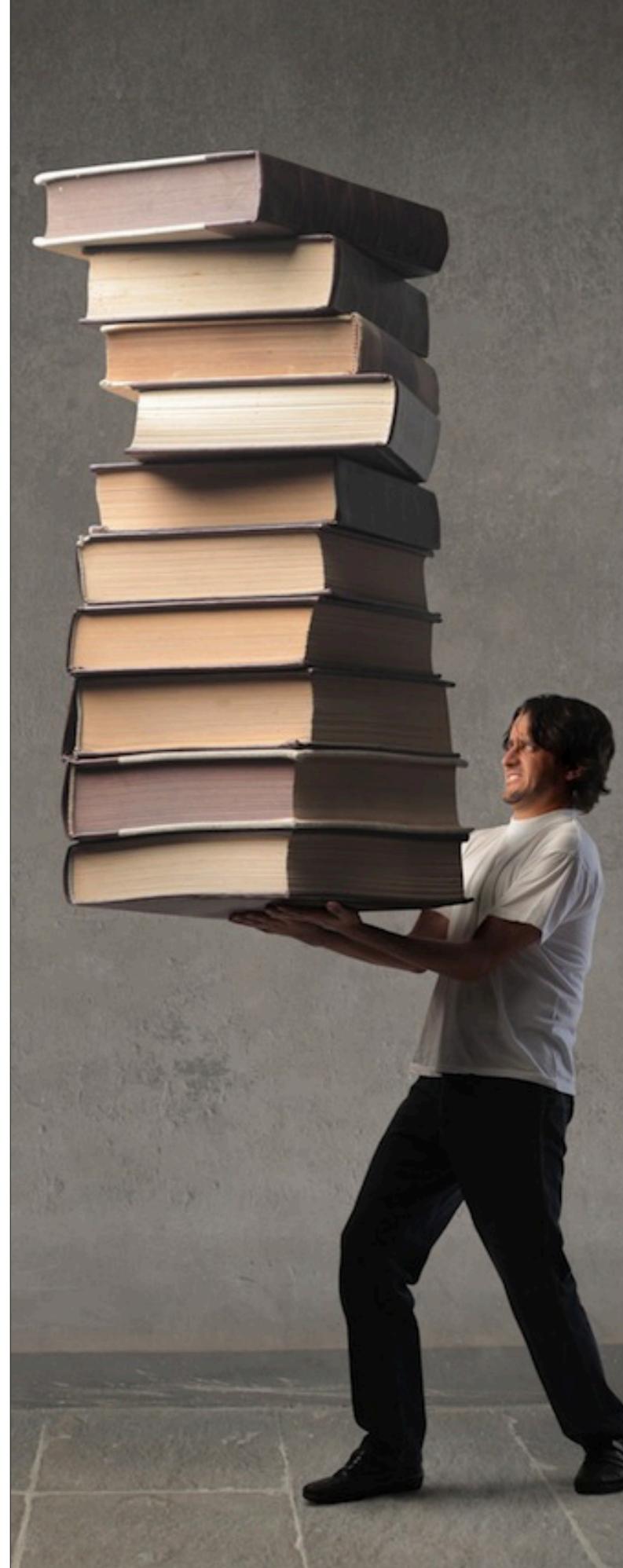
텍스트 마이닝이 어려운 이유

사생활 침해와 보안 (Privacy)

- ▶ 트위터, 페이스북, 블로그 등의 텍스트는 개인의 정보와 생각을 그대로 반영
- ▶ 텍스트 마이닝이 그 유용성에만 주목한 나머지 자칫 무분별한 개인정보의 수집 및 활용으로 사생활 침해 등의 문제를 야기 할 수 있음
- ▶ 데이터 분석 전 데이터에서 반드시 개인정보를 제거하는 전처리 과정이 필요함

정확도 측정 (Accuracy) & 평가 (Validation)

- ▶ 정확성과 신빙성을 정량적으로 평가하기 어려움
- ▶ 데이터 마이닝에 비하여 정보추출 능력이나 정확성이 떨어지는 경향이 있음
- ▶ 오피니언 리더들의 영향력이 과도하게 작용해 분석결과가 편향될 가능성이 있음



E.O.D