# ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

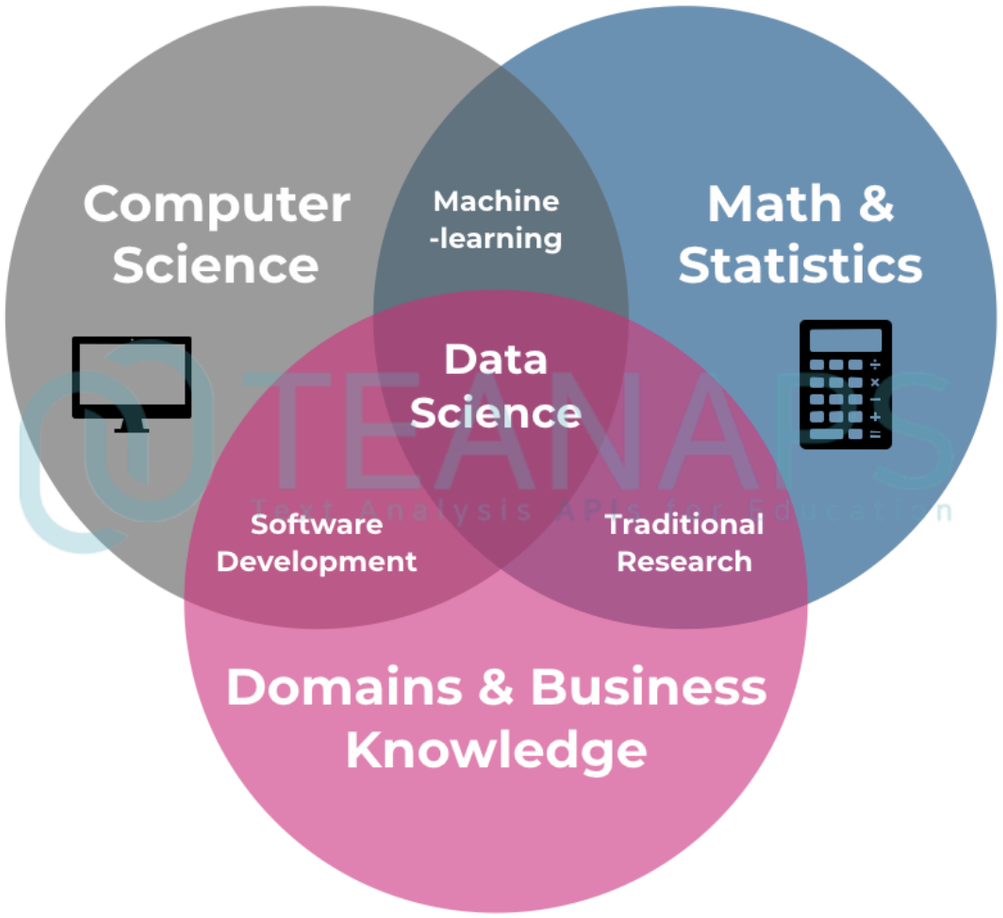
WEEK 01

# Introduction

# What Do Data Scientists Do?

## 데이터 과학자 (Data Scientist)

- 데이터 분석을 통해 비즈니스 패턴을 기회로 활용하고, 문제점에 대한 혜안(solution)을 제시하는 직업군
- **데이터 분석가** (Data Analyst) : 데이터 분석을 통해 비즈니스와 결합하여 실행 가능한 통찰력(insight)을 제공하는 직업
- **데이터 엔지니어** (Data Engineer) : 데이터 분석을 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

* Michael Barber, Data science concepts you need to know! Part 1, 2018.1.14., https://towardsdatascience.com/introduction-to-statistics-e9d72d818745/.
** Dave Holtz, What is Data Science? 8 Skills That Will Get You Hired in Data, 2014.11.7., https://blog.udacity.com/2014/11/data-science-job-skills.html/.
*** references

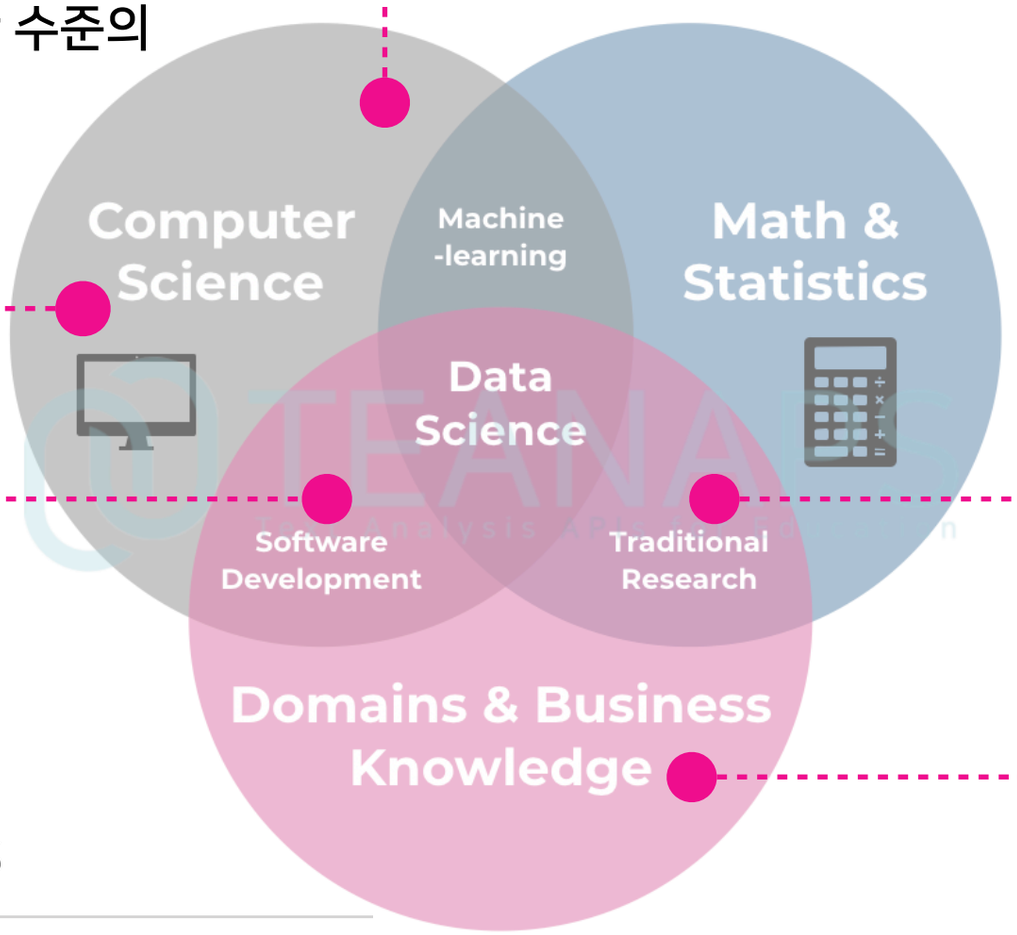
# What Do Data Scientists Do?

## Analytics Developers

- 데이터 분석 전문지식과 함께 이를 S/W로 구현 가능한 전문가
- 알고리즘 구현이 가능한 수준의 코딩능력이 필수

## Data Engineers

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




## Generalists

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

## Deep Specialists

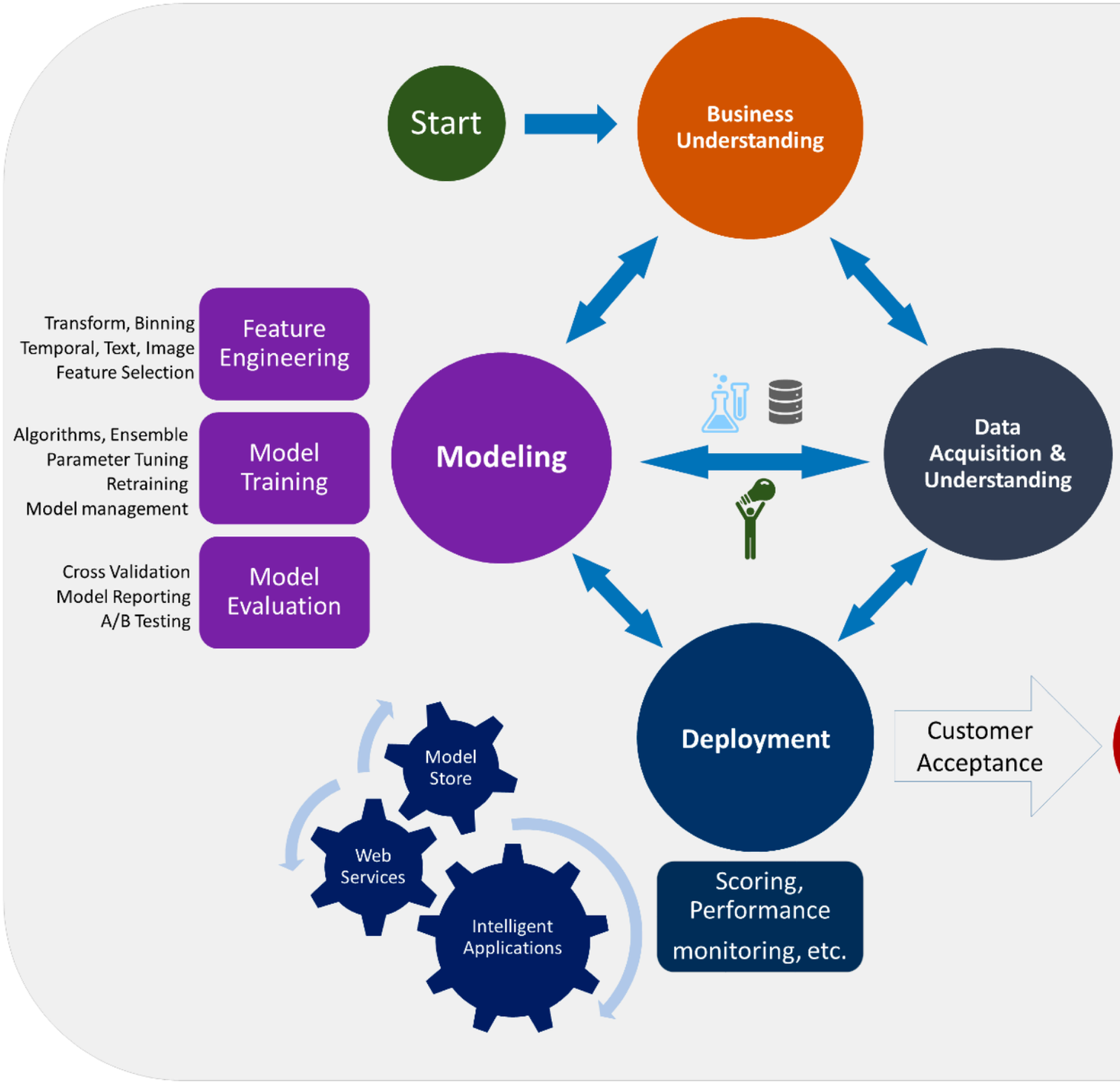
- 특정 데이터 분야에 전문지식을 가진 사람
- 컴퓨터 사이언스에 대한 이해 필요

## Industry Specialists

- 데이터 분석적 사고를 통해 문제를 해결하고자 하는 도메인 전문가

* Michael Barber, Data science concepts you need to know! Part 1, 2018.1.14., https://towardsdatascience.com/introduction-to-statistics-e9d72d818745/.
** 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/.
*** references

# Data Scientist Lifecycle



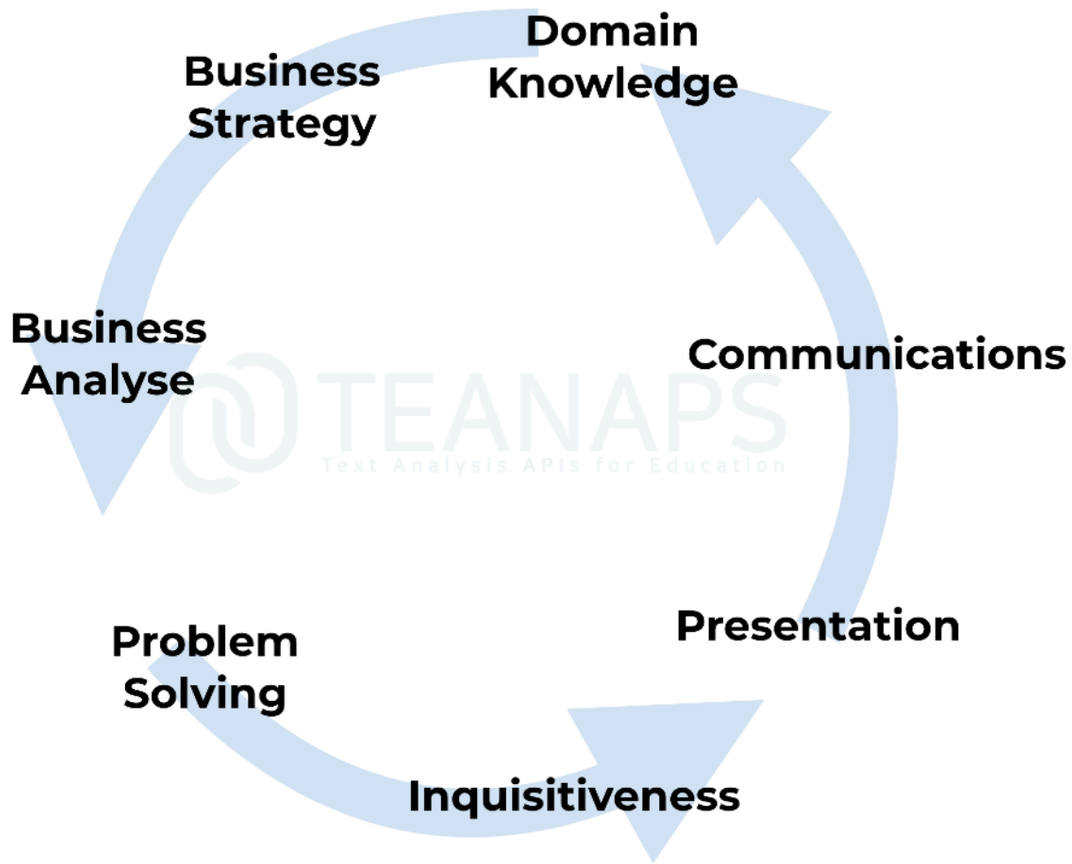


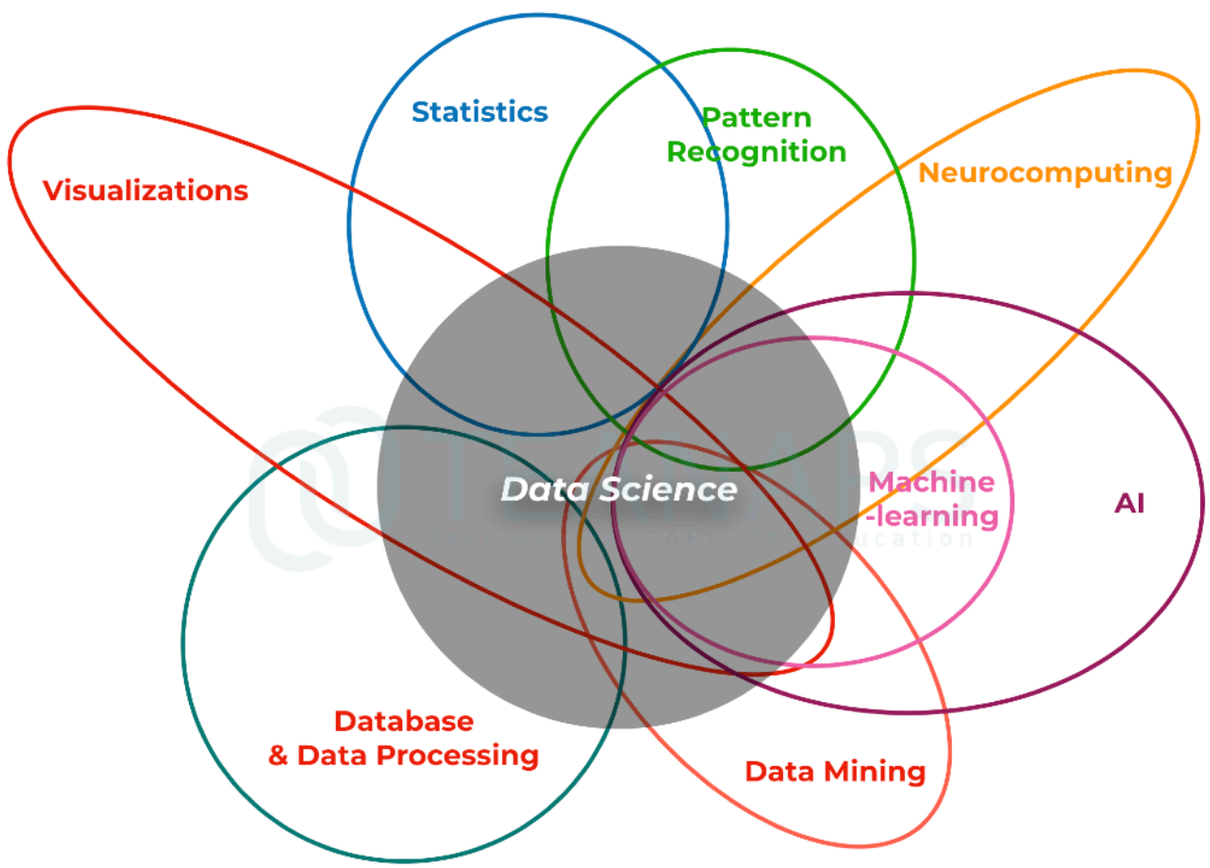
* Gary Ericson et al., What is the Team Data Science Process?, 2017.10.20., https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview/.
** Simplilearn Solutions, About the program, https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training/.
*** references

# What is Data Mining

## 데이터 마이닝 (Data Mining)

- 데이터 속의 유용한 패턴 또는 지식을 찾아내는 것
- 중요한 의사결정을 위해 데이터에서 유효하고(valid), 새롭고(novel), 잠재적으로 유용(potentially useful)하면서, 의미있는 패턴(pattern)이나 관계(relationship)을 파악해 가는 프로세스
- “Its goal is to develop knowledge of some phenomena.”



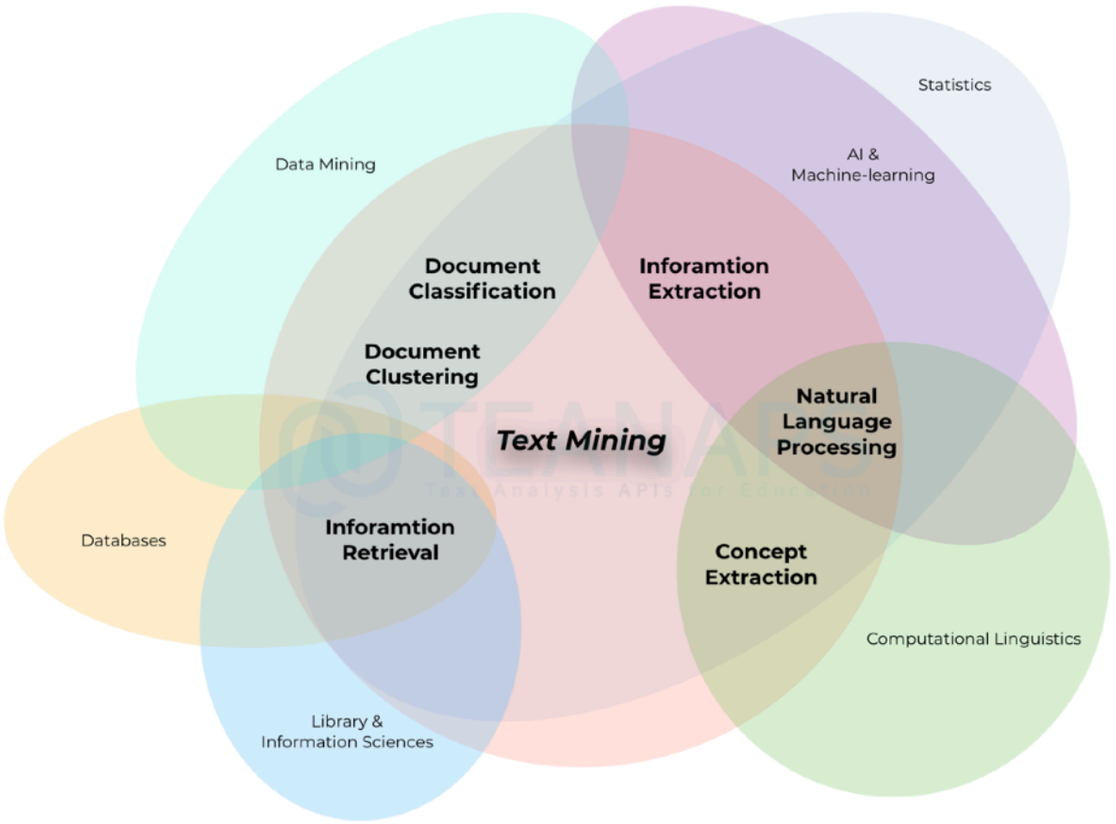




* Leonard Heiler, Difference of Data Science, Machine Learning and Data Mining, 2017.3.20., https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining/.
** references
*** references

# What is Text Mining

## 텍스트 마이닝 (Text Mining)

- 텍스트 데이터에서 자연어처리(Natural Language Processing, NLP) 기술을 바탕으로 유의미한 패턴 또는 지식을 추출하는 과정
- 언어학과 통계 기반에서 머신러닝을 통해 기계가 언어학적, 통계적 특징을 학습하는 형태로 발전하여 활용됨
- 텍스트 마이닝 유형
  1) **설명적 마이닝**(descriptive mining) : 텍스트 집합에 있는 의미나 개념을 찾아내거나 이해를 돕는 형태 (분류, 검색, 여론조사 등)
  2) **예측적 마이닝**(predictive mining) : 텍스트에 내포된 정보를 의사결정에 활용하는 형태 (질문 자동답변, 구매 예측, 주가예측, 스팸분류 등)




| 텍스트 마이닝 유형 | 활용분야 | |
|---|---|---|
| | 실무 | 연구 |
| 검색 (Information Retrieval) | 스팸 필터링 | 사회동향 분석 |
| 분류 (Classification) | 이슈 검출/트래킹 | 소셜미디어 분석 |
| 군집화 (Clustering) | 정보검색 | 이슈 트래킹 |
| 웹마이닝 (Web Mining) | 자살률 예측 | 온라인 행동 분석 |
| 정보추출 (Information Extraction) | 주가 예측 | 연구분야 탐색 |
| 개념추출 (Concept Extraction) | 소비자 인식 조사 | 질병관계 예측 |
| ~~자연어처리 (NLP)~~ | 경쟁사 분석 | 정책전략 수립 |

* Gary Miner, John Elder, Andrew Fast, Thomas Hill, Robert Nisbet, and Dursun Delen, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications 1st Edition, 2012.1.25., Academic Press.
** references
*** references

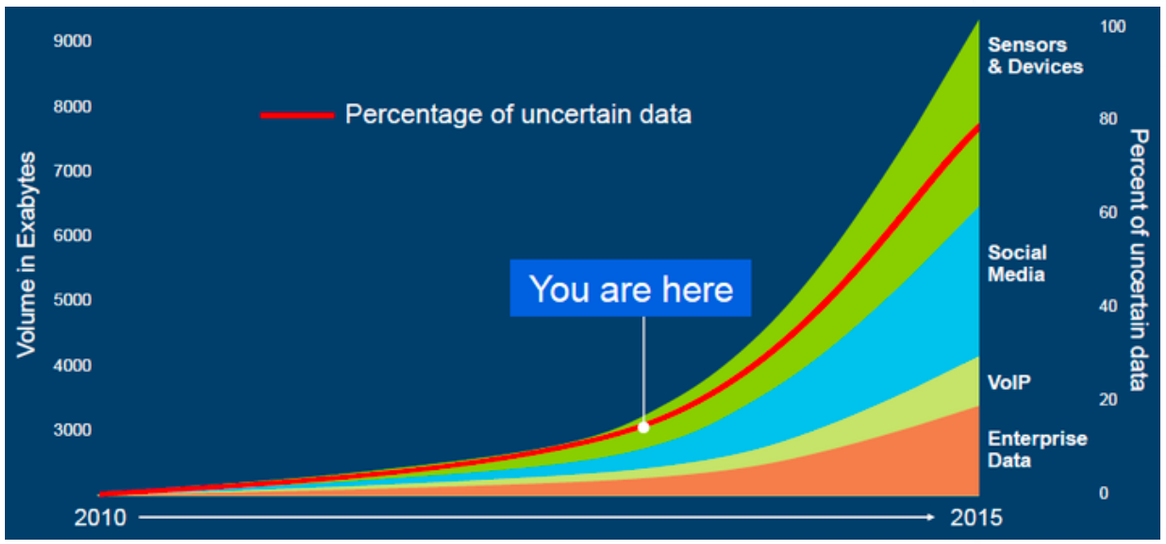
# 텍스트 마이닝이 중요한 이유

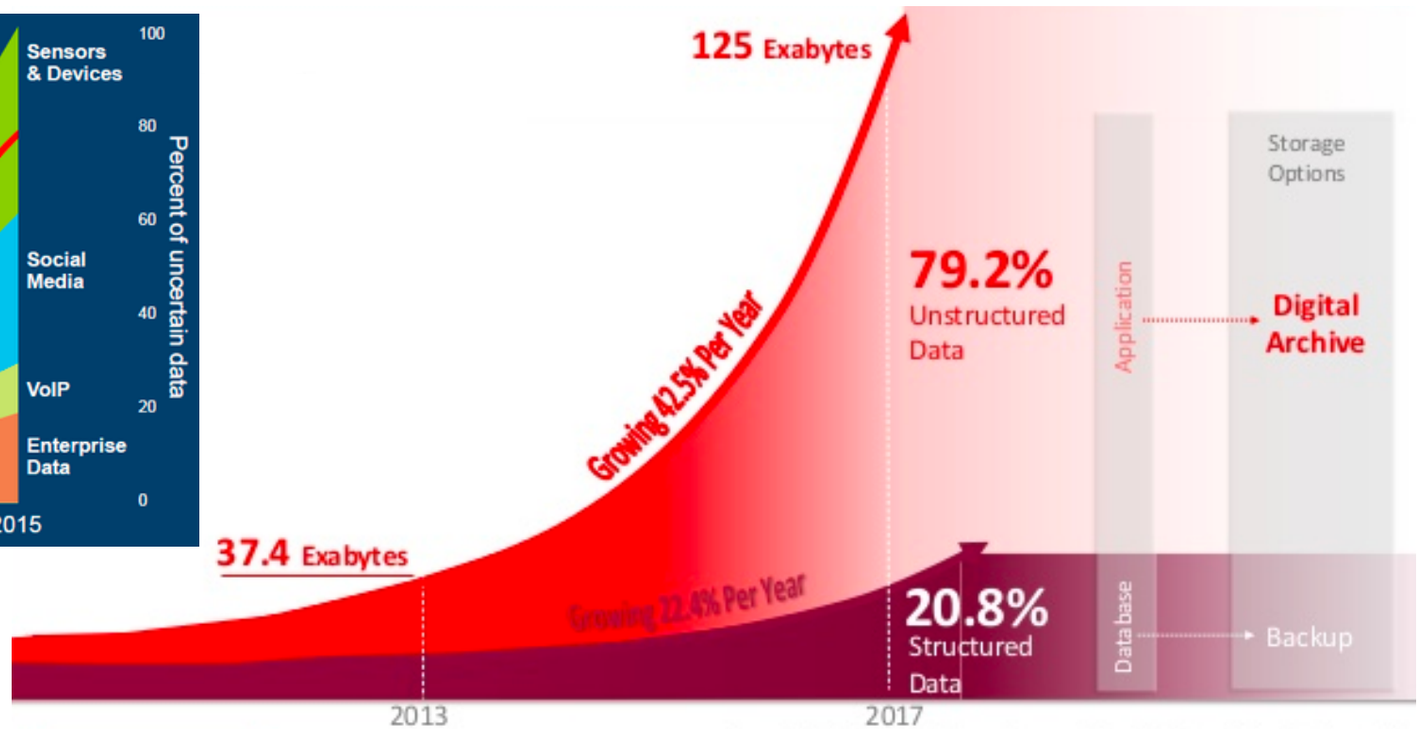
## 비정형 데이터의 폭발적 증가

- 잠재적 가치를 포함하는 비정형 데이터(unstructured data)가 대규모로 생성됨과 동시에 비정형 데이터 속에서 미래의 의사결정에 관련된 유용한 정보를 찾아내어 활용하는 작업이 매우 중요해짐
- 실제로 생산되는 데이터의 70~80%는 비정형 테이터에 해당함 (기사, 블로그, 문서, 보고서 등)

## 텍스트 데이터의 폭발적 증가

- 소셜 네트워크 서비스(Social Network Service, SNS)를 통한 온라인 양방향 커뮤니케이션이 활성화됨
- 4차산업혁명과 사물 인터넷(Internet of Things, IoT) 등 빅데이터 관련 기술이 급진적으로 발전함



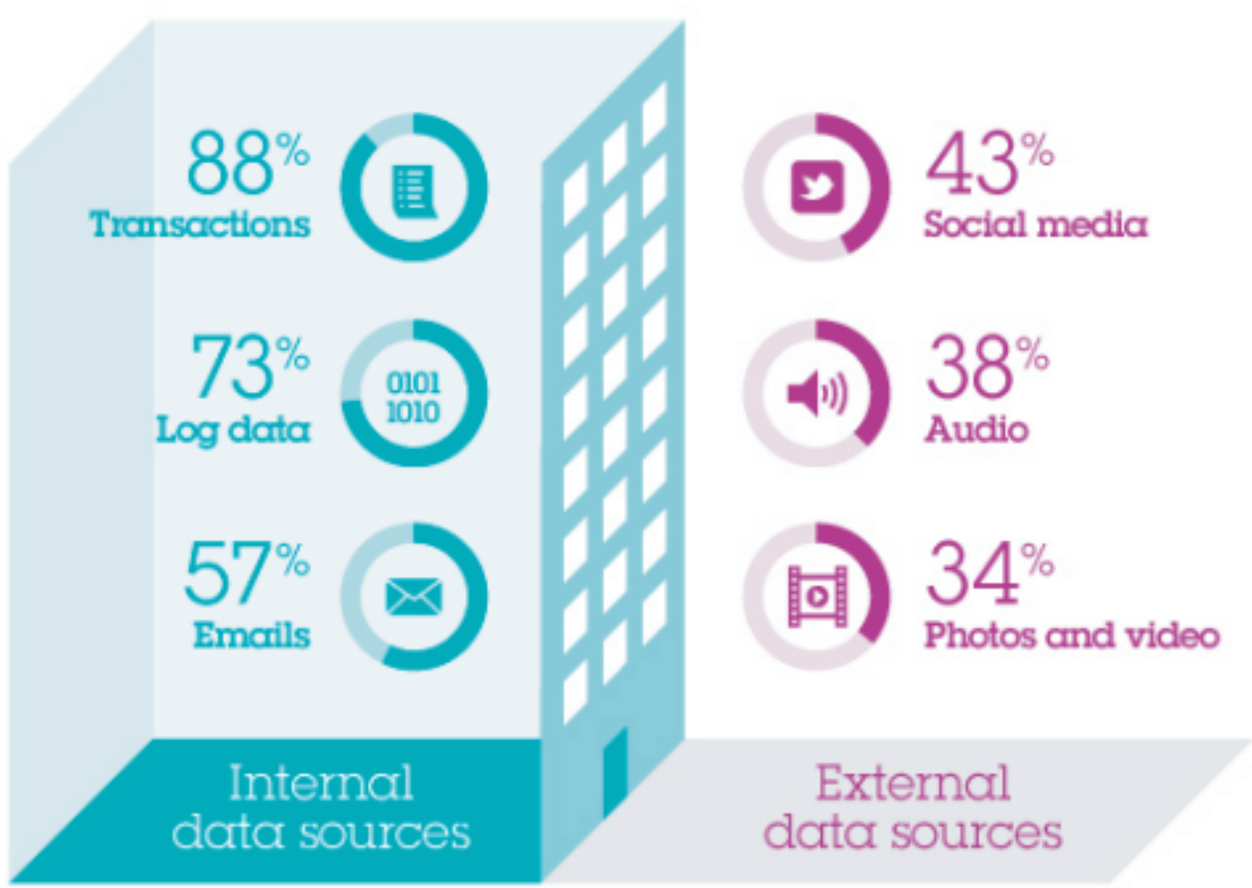




* Nadkarni, A., and Yezhkova, N., Structured versus unstructured data: The balance of power continues to shift, IDC (Industry Development and Models), 2014.3.17., https://issuu.com/reportlinker/docs/structuredversusunstructureddatathebalanceofpower/.
** Larry Dignan, IBM eyes China, South America, Africa 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/.
*** references

# 텍스트 마이닝이 중요한 이유

## 가장 흔하고 접하기 쉬운 데이터

- 텍스트가 존재하지 않는 곳은 없으며, 다양한 서비스를 통해 수많은 텍스트 데이터가 생산됨
- 온라인 비정형 텍스트 데이터의 대부분이 SNS에서 발생함 (Twitter, Facebook, YouTube, 블로그, 커뮤니티 등)
- 웹에서 사용자들은 주로 텍스트 데이터를 활용해 콘텐츠를 생성하고 의사소통함
- 다양한 형태의 비정형 데이터(오디오, 비디오, 각종 센서 등)가 텍스트 형태로 변형되어 활용됨 → 음성-텍스트 변환(Speech to Text, STT)
- 웹 크롤링(web crawling), Open API(Application Programming Interface) 등의 활성화로 텍스트 데이터 수집 및 확보가 용이해짐


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

# 텍스트 마이닝이 어려운 이유

## 언어적 한계점

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

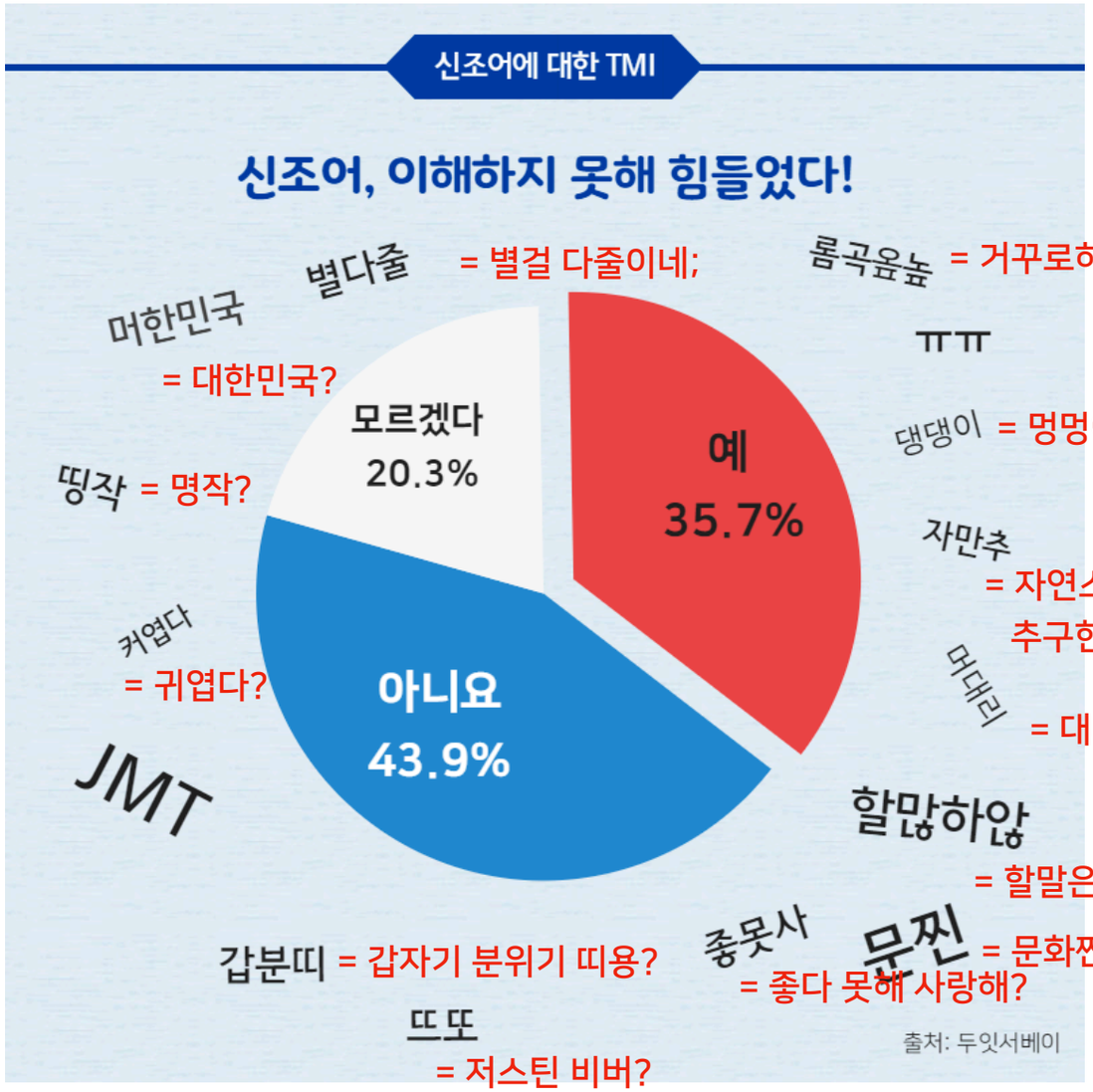
## 데이터적 한계점

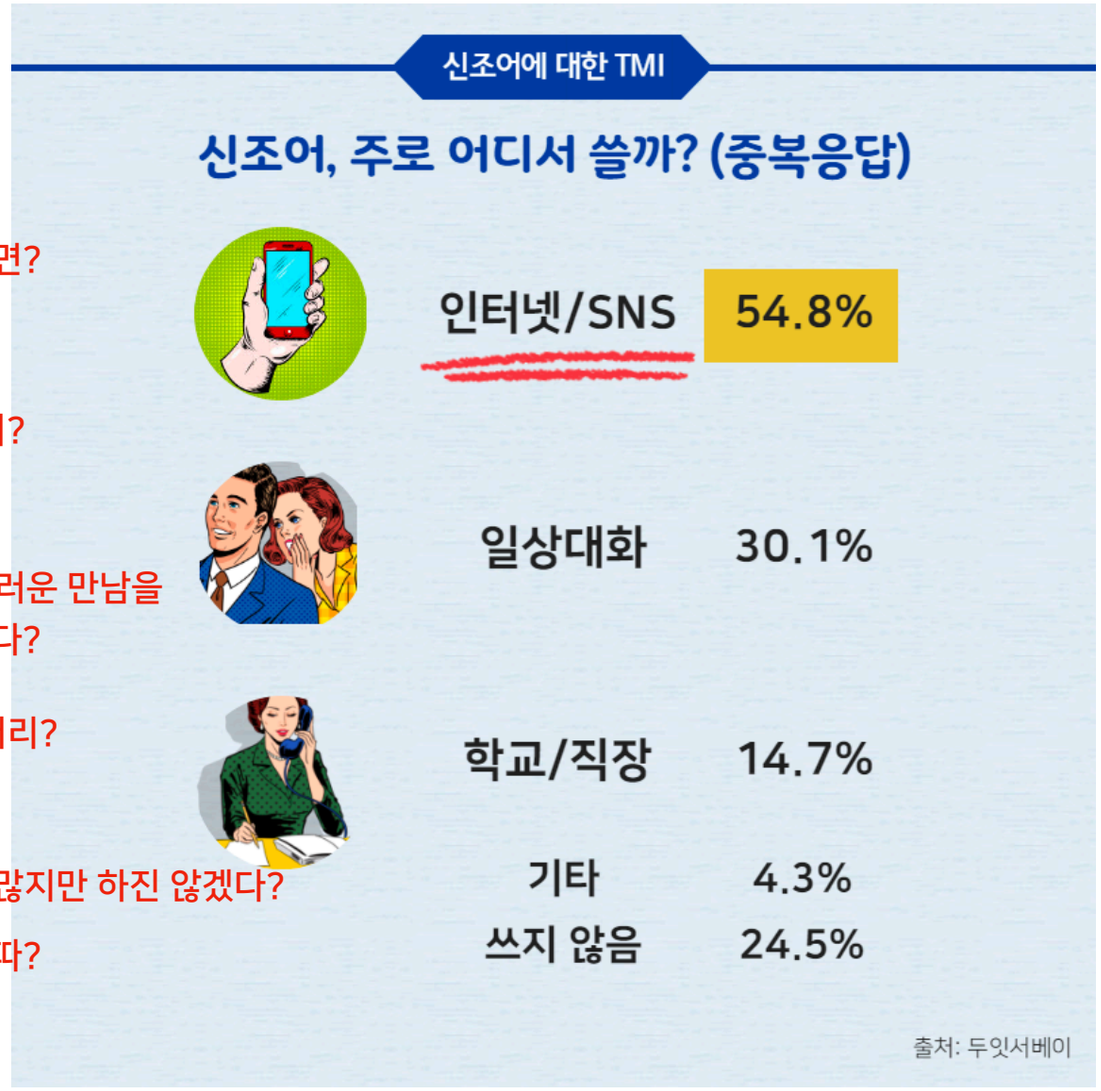
- 텍스트는 비정형 데이터로서, 일반적인 필드와 레코드 구조를 가지고 있지 않음 → 전처리 과정이 복잡하고 어려움
- 텍스트의 형태와 특징에 따라 전처리 과정과 분석방법에 대한 접근을 다르게 고려해야함
- 자연어처리에 대한 이해가 필요하고, 분석시간이 길어 잠재적 가치에도 불구하고 충분히 활용하지 못하고 있음
- 방대한 양, 데이터의 규모 증가, 그리고 그 형태의 비정형성으로 인하여 그 분석과 활용이 어려움

| 구분 | 내용 |
|---|---|
| 오탈자 | “헝거게임 잼잇써요완전 대신 이전편꼭바여”<br>“솔까 타노스 보석 하나도 못구했을때 다들 머했음? 3개 얻었을때도 그렇게 안쌔 뵈더만...” |
| 동의어,<br>동음이의어 | 한혜진 : 1. 모델 한혜진 (달심), 2. 배우 한혜진 (기성용 부인), 3. 가수 한혜진 (트로트 가수)<br>Close : 1. Opposite of open, 2. A preposition meaning not far<br>IS : 1. Information System, 2. Islamic State, 3. International Standard |
| 전처리 | 분석 데이터의 언어를 파악하고 언어의 특징 (교착어, 굴절어 등)에 맞는 전처리 작업 진행<br>댓글 단위로 분석할지, 문장 단위로 분석할지에 따라서 데이터 분리작업 진행 |
| 정보추출 | 해시 태그 (hash tag) 추출 : ‘#’ + (문자)<br>핸드폰 번호 추출 ‘010’ - (4자리 숫자) - (4자리 숫자) |

* 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/.
** 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/.
*** references

# 텍스트 마이닝이 어려운 이유




* 통계청, 별다줄! 신조어에 대한 TMI (Too Much Information), 2018.8.14., https://blog.naver.com/hi_nso/221338158877/.
** references
*** references

# 텍스트 마이닝이 어려운 이유

## 복잡한 한국어 텍스트 분석

- 한국어 텍스트 분석은 언어학적 특성으로 인해 전처리와 분석과정이 까다로움
  1) 초성, 중성, 종성의 조합이 하나의 음절을 형성함
  2) 첨가어로 첨용과 활용의 케이스가 매우 많음 → 조사와 접사가 붙어 문법적 관계를 형성함
- 용언이 변하는 경우의 수가 매우 많고 그 과정이 하나의 개념으로 확인하기가 어려움
- 형태소 분석기의 한계 및 미비한 어휘사전 → 신조어, 미등록어, 새로운 용어의 조합을 반영하기 어려움

| 한계점 | 예시 |
|---|---|
| **용언의 변형** | 모르다<br>→ 모르네, 모르데, 모르지, 모르더라, 모르리라, 모르는구나, 모르니, 모르고, … |
| **형태소 분석** | “비비크림 빠빠빠~립스틱을 마마마”<br>→ 비/NNG + 비/NNG + 크림/NNG + 빠빠빠/UN + ~/SO + 립스틱/NNG + 을/JX + 마/NNG + 마마/NNG<br>“황민현에게 트렌치코트는 정말 존멋♥”<br>→ 황민/NNG + 현/NNG + 에게/JKM + 트렌치/NNG + 코트/NNG + 는/JX + 정말/MAG + 졸/VV + ㄴ/ETD + 멋/NNG + ♥/SW<br>“자다가 퇴근했음 좋게따 외냐면 내일 사랑니 째러 가야 되니까는…”<br>→ 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** | 지카 바이러스, 오백따리, 트둥이, 울와(우리 트와이스), 팬코(팬 코스프레 유저) |

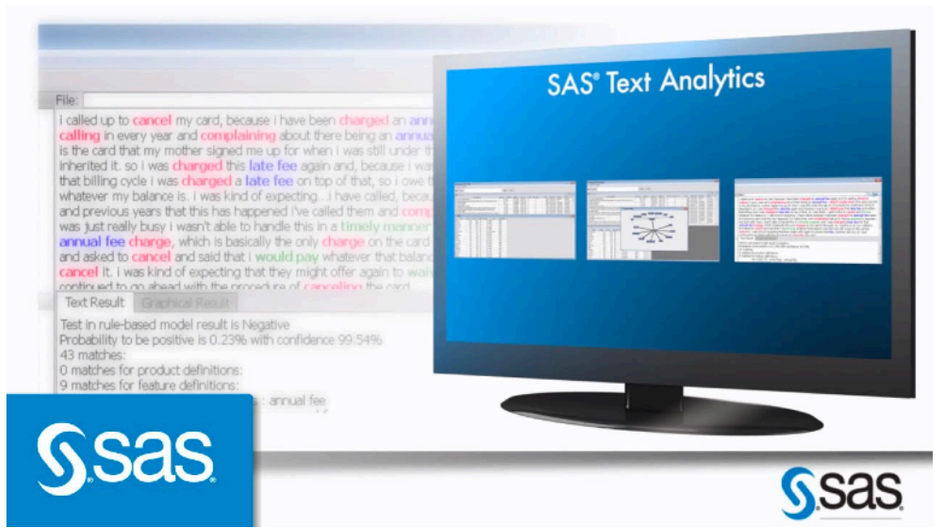
* references
** references
*** references

# 텍스트 마이닝이 어려운 이유

## 비싼 상용 애플리케이션 & 솔루션, 어려운 프로그래밍 언어

- 텍스트 마이닝의 목적과 특성에 따라 상용 애플리케이션, 솔루션 또는  프로그래밍 언어 활용 중 선택이 필요함
- **애플리케이션**(Application) : SAS Text Miner, IBM Watson Explorer(WEX), …
- **솔루션**(Solution) : 소셜메트릭스(Social Matrix, Sometrend), …
- **프로그래밍 언어**(Programming Language) : Python, JAVA, R, …




* references
** references
*** references

# 텍스트 마이닝이 어려운 이유

## 사생활 침해와 보안 (Privacy)

- 트위터, 페이스북, 블로그 등의 텍스트는 개인의 정보와 생각을 그대로 반영
- 자칫 무분별한 개인정보의 수집 및 활용으로 사생활 침해 등의 문제를 야기 할 수 있음
- 데이터 분석 전 데이터에서 반드시 개인정보를 제거 또는 마스킹(masking) 처리하는 전처리 과정이 필요함

## 정확도 측정과 평가 (Accuracy & Validation)

- 정확도, 신뢰도 등 분석결과를 정량적으로 평가하기 어려움
- 정형 데이터를 활용하는 데이터 마이닝에 비해 분석결과가 충분하지 않거나 정확성이 떨어지는 경향이 있음
- 오피니언 리더(opinion leader)들의 영향력이 과도하게 작용해 분석결과가 편향될 가능성이 있음

* references
** references
*** references

# E.O.D

Contact
http://www.teanaps.com
fingeredman@gmail.com