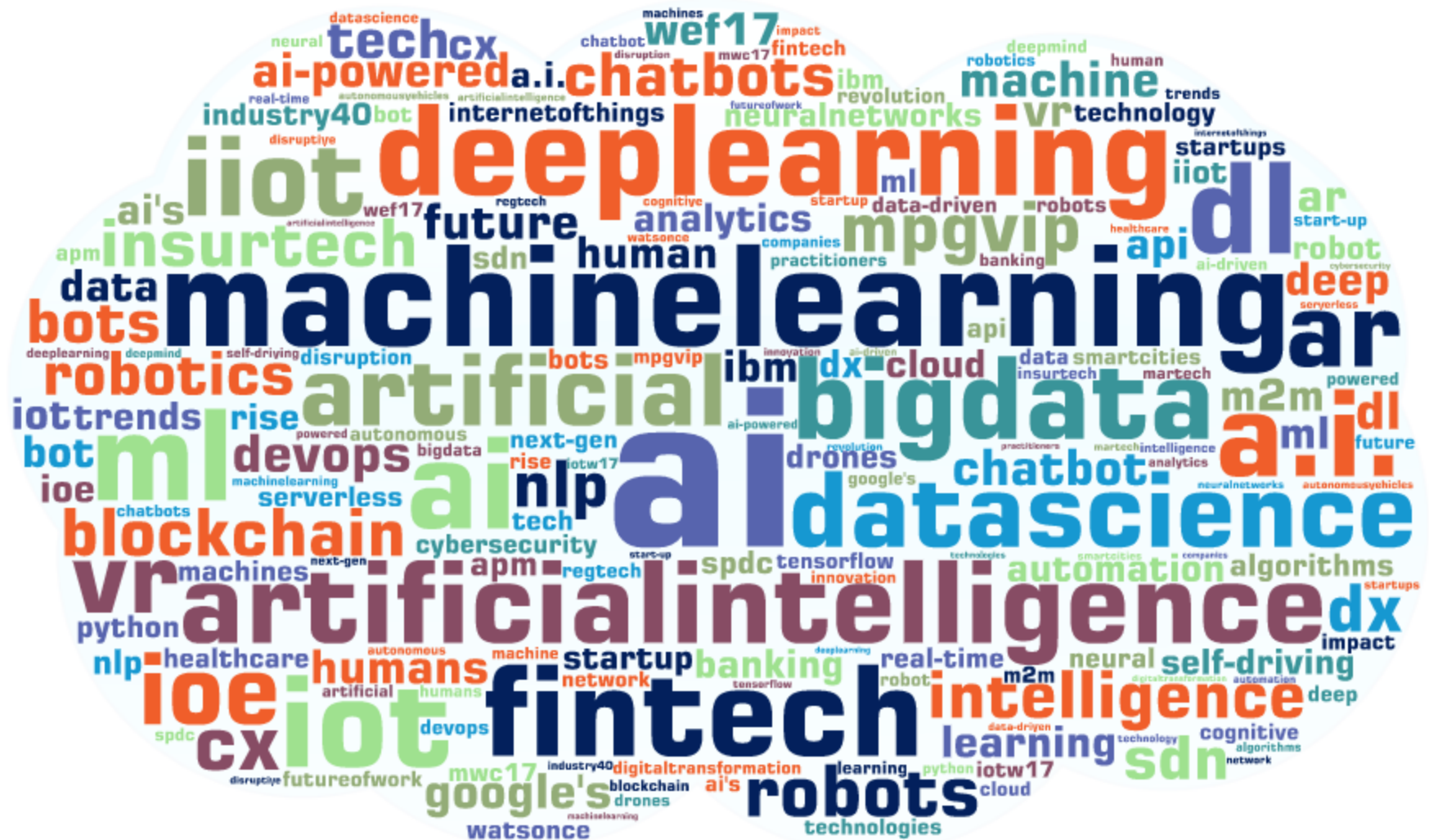


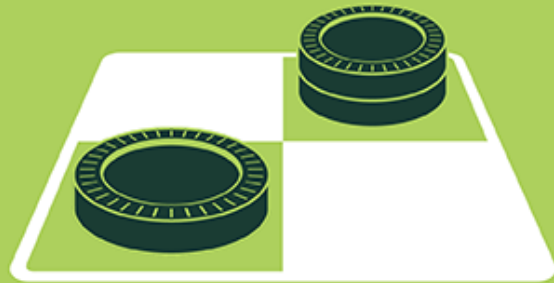
NLP

Artificial Intelligence



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

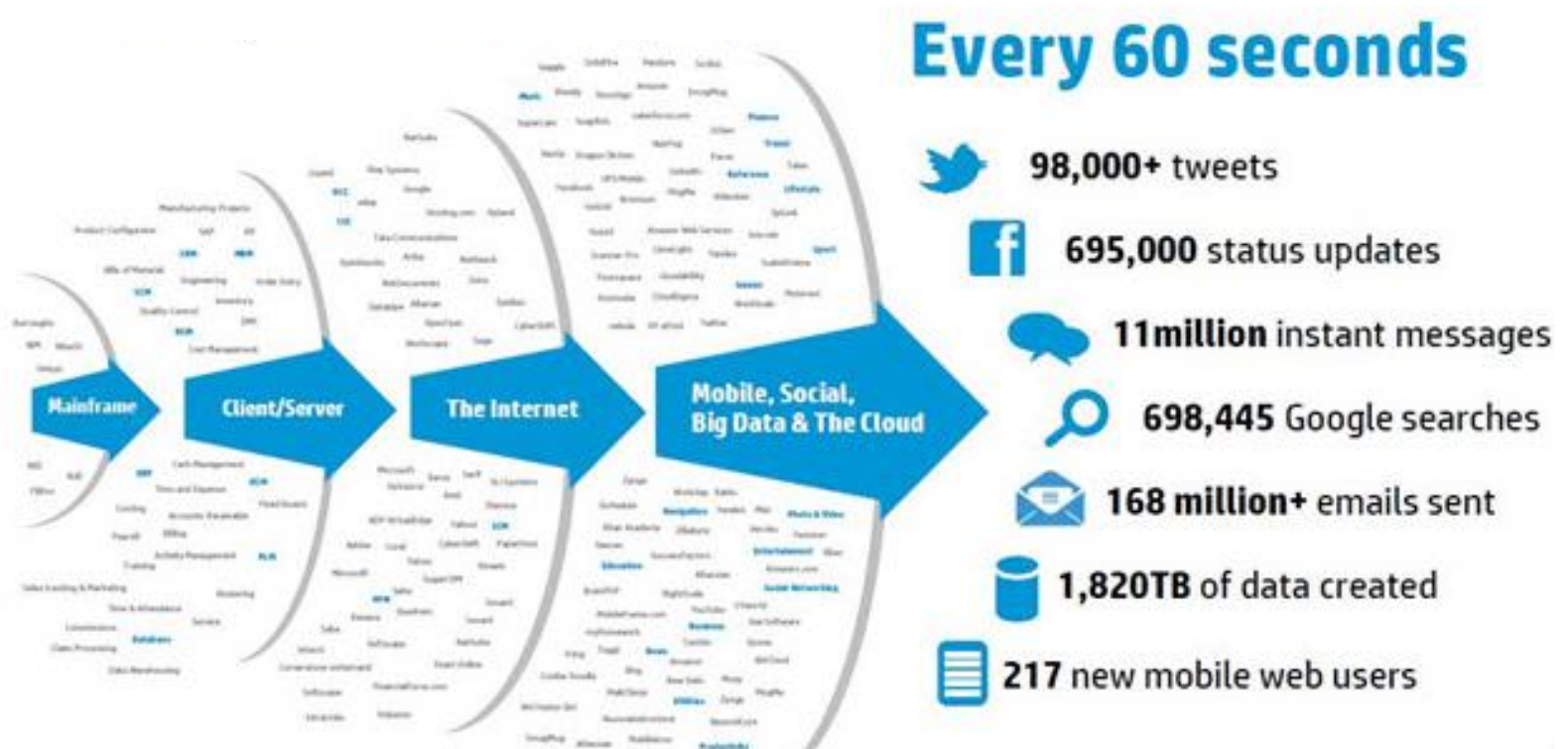
Deep learning breakthroughs drive AI boom.



1950's 1960's 1970's 1980's 1990's 2000's 2010's

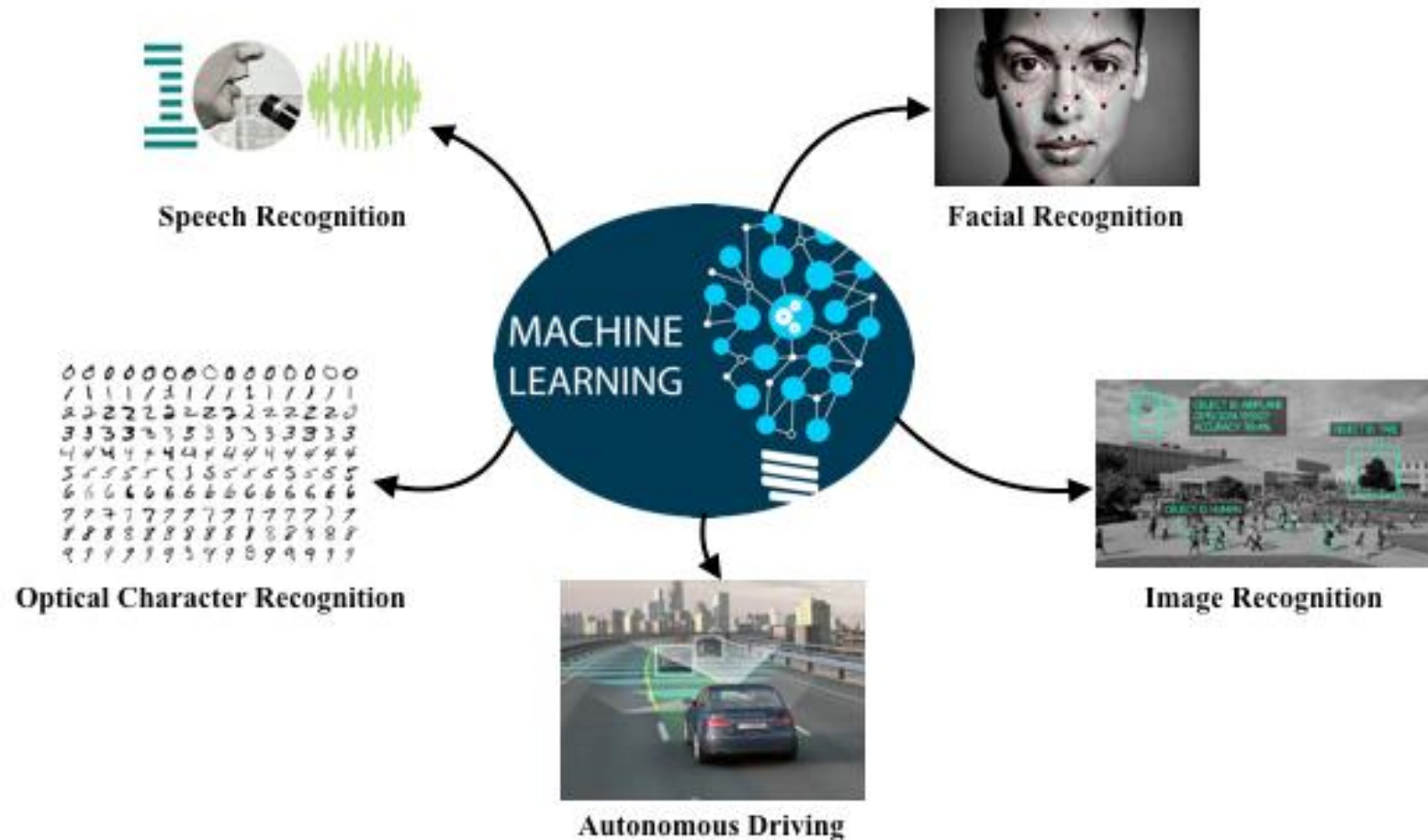
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

abundance of data

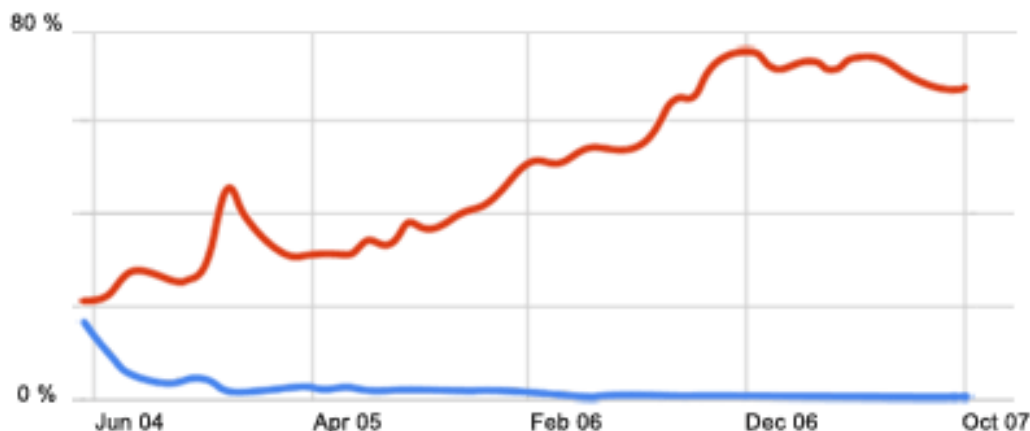


HANA Healthcare Big Data Subject Matter Experts and Analysts, 2017

applications



More spam, but not in Gmail inboxes



■ Spam prevalence: % of all incoming Gmail traffic (before filtering) that is spam

■ Missed spam: % of total spam reported by Gmail users

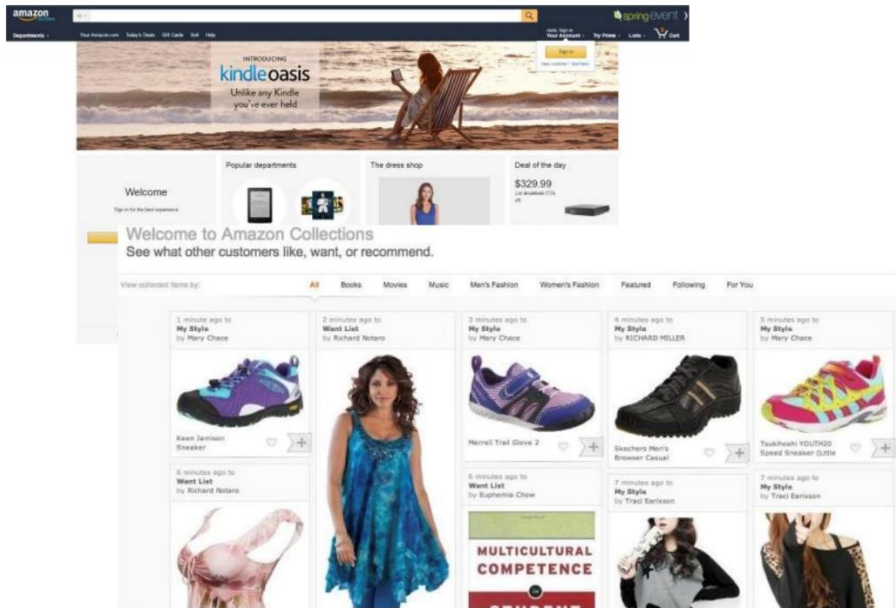
As the amount of spam has increased, Gmail users have received less of it in their inboxes, reporting a rate less than 1%.

The screenshot shows the Gmail web interface. At the top is the Google search bar. Below it is the Gmail header with navigation icons and a search bar. The left sidebar contains the 'COMPOSE' button and a list of folders: 'Less', 'Important', 'Chats', 'All Mail', 'Spam' (highlighted with a black box), and 'Trash'. Below these are 'Categories' and 'Circles'. The main content area shows the 'Primary' tab selected, displaying an email from 'L [redacted] (via Google Do.' with the subject 'Joe Jones Letter.docx - Invitation to e' dated 'Nov 3'. Below the email, it shows '0.01 GB (0%) of 15 GB used' and a 'Manage' link. On the right, it says 'Last account activity: 17 minutes ago' with a 'Details' link. The bottom of the sidebar shows a user profile for 'Laura'.

Interactive Chart for DOLLAR INDEX SPOT (DXY)



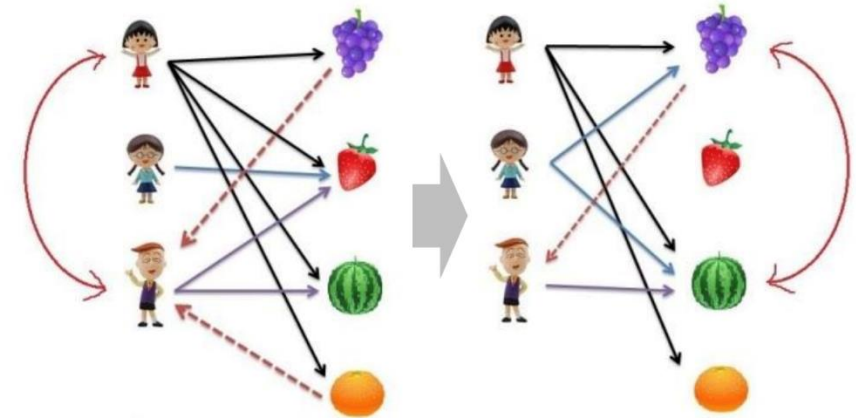
아마존 웹사이트



- ▲ 아마존 메인 화면, 아마존은 빅데이터를 이용한 분석 기술을 토대로 고객에게 맞춤형 추천 상품을 제공합니다.

출처: Amazon

아마존의 Item-to-item collaborative filtering



[유저 기반 필터링]

[아이템 기반 필터링]

- ▲ 기존의 데이터를 통해 상품을 추천해주는 알고리즘으로 고객 수나 아이템 수와 관계없이 대량의 데이터를 처리해서 고품질의 추천을 해줌. 상품 간의 상관관계를 결정하는 아이템 매트릭스를 만든 후 고객의 최신 입력 데이터를 기반으로 고객의 기호를 유추해서 상품을 추천.

즉, 기존의 온라인 쇼핑몰에서 자주 활용하던 소비자의 패턴 분석에 의한 추천 방식이 아닌 구매한 물건 혹은 서치한 물건 중심으로 추천하는 방식.

출처: <http://blog.naver.com/sanny0314/220630765408>

넷플릭스 웹사이트

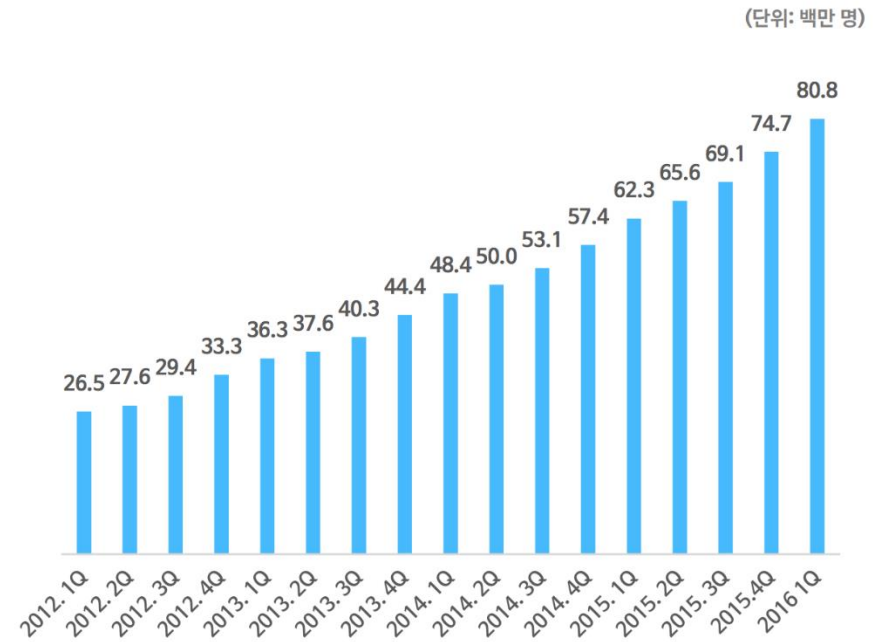


▲ '하우스 오브 카드(House of Cards)'로 큰 성공을 거둔 넷플릭스 웹사이트

보통 방송사, 영화사들은 신규 프로그램을 방영할 경우 막대한 광고를 송출. 반면, 3천 700만 명이 넘는 회원을 확보하고 있는 넷플릭스에서는 광고 대신 빅데이터를 활용. 빅데이터 분석에 기반한 넷플릭스 콘텐츠 추천 서비스는 드라마 제작에 큰 영향을 미침.

출처: 정보통신정책연구원(2014.12), 넷플릭스의 빅데이터

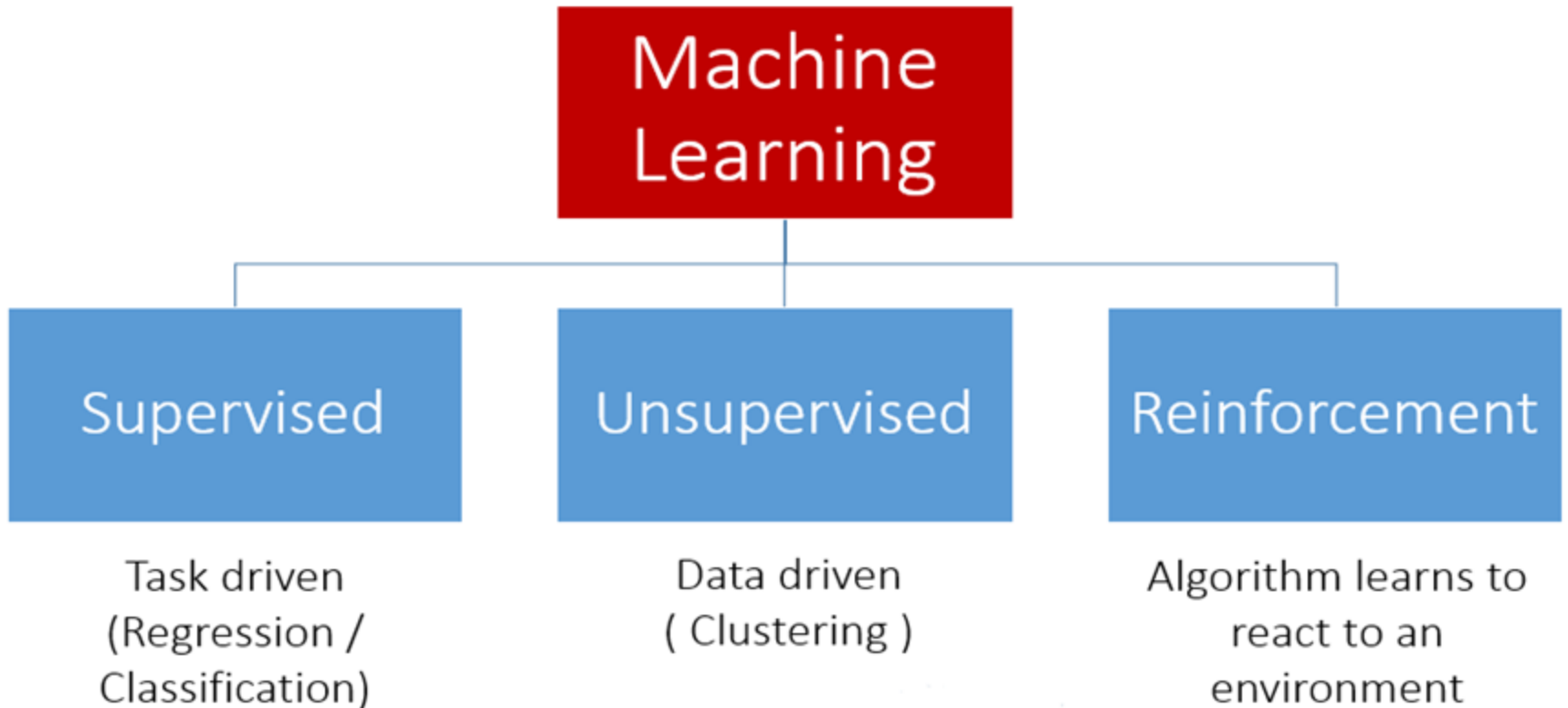
넷플릭스의 분기별 가입자 증가 추이



출처: 2016년은 예상

출처: Netflix

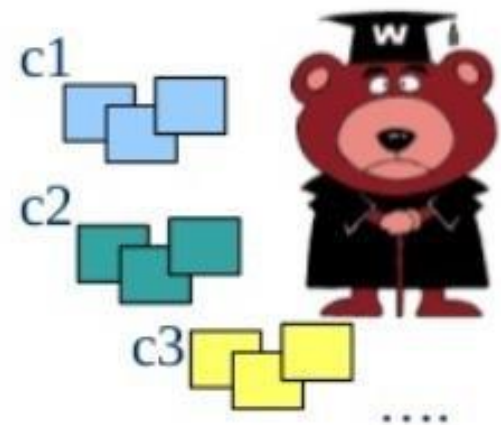
Types of Machine Learning



Supervised Vs. Unsupervised

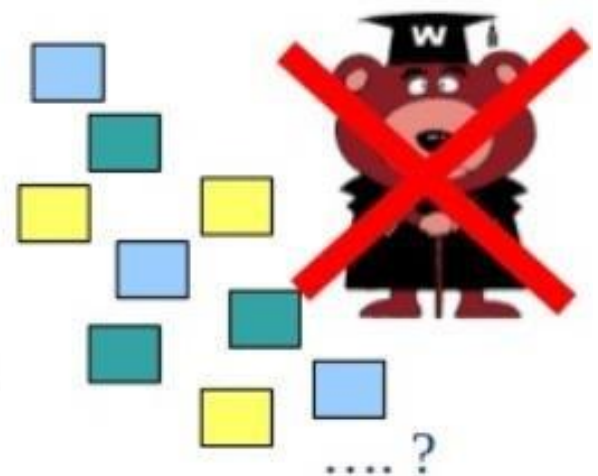
- **Supervised**

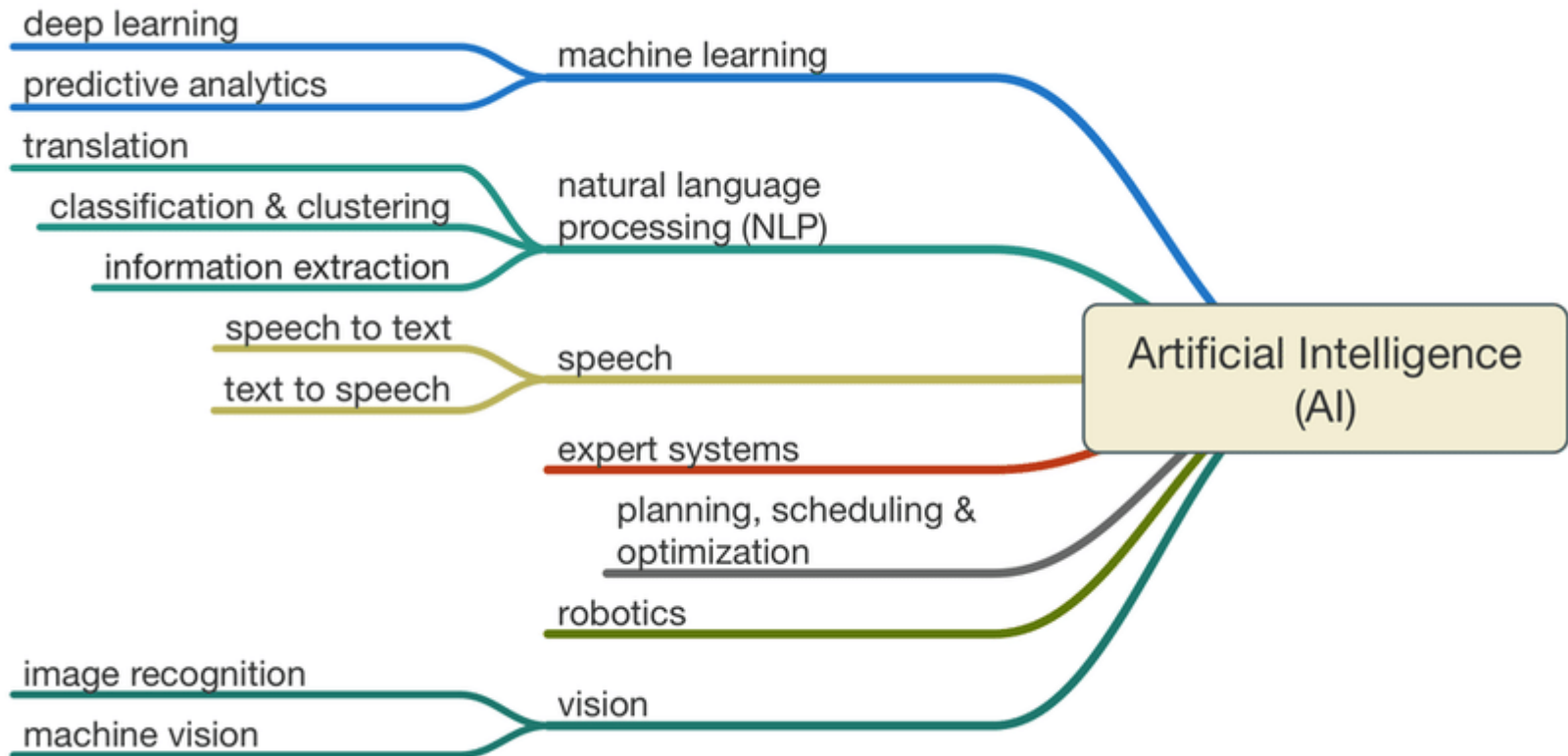
- **knowledge of output** - learning with the presence of an “expert” / teacher
 - data is **labelled** with a class or value
 - **Goal:** predict class or value label
 - e.g. Neural Network, Support Vector Machines, Decision Trees, Bayesian Classifiers



- **Unsupervised**

- **no knowledge of output** class or value
 - data is **unlabelled** or value un-known
 - **Goal:** determine data patterns/groupings
- Self-guided learning algorithm
 - (internal self-evaluation against some criteria)
 - e.g. k-means, genetic algorithms, clustering approaches ...







전문가
시스템

논리/규칙 기반
(1950s~1980s)

- 현실의 문제는 명확하게 정의되지 않는 경우가 대부분
- 세상의 모든 상황과 지식을 알려줄 수 없다는 점



인공
신경망

연결기반
(1980s~1990s)

- 계산이 복잡하고 컴퓨팅 파워가 신경망을 충분히 학습시킬 수준이 안됨
- 훈련시킬 데이터가 부족



기계 학습

통계 기반
(1990s~2010s)

- 자질 설계의 능숙도에 따라 성능이 좌우
- 최적화 방식으로 문제를 해결하기 때문에, 다른 분야에 적용하기 어려움



딥 러닝

딥 러닝 기반
(2010s~)

- 컴퓨터가 스스로 자질을 학습해 설계
- 학습한 데이터를 다른 문제에 재사용 가능

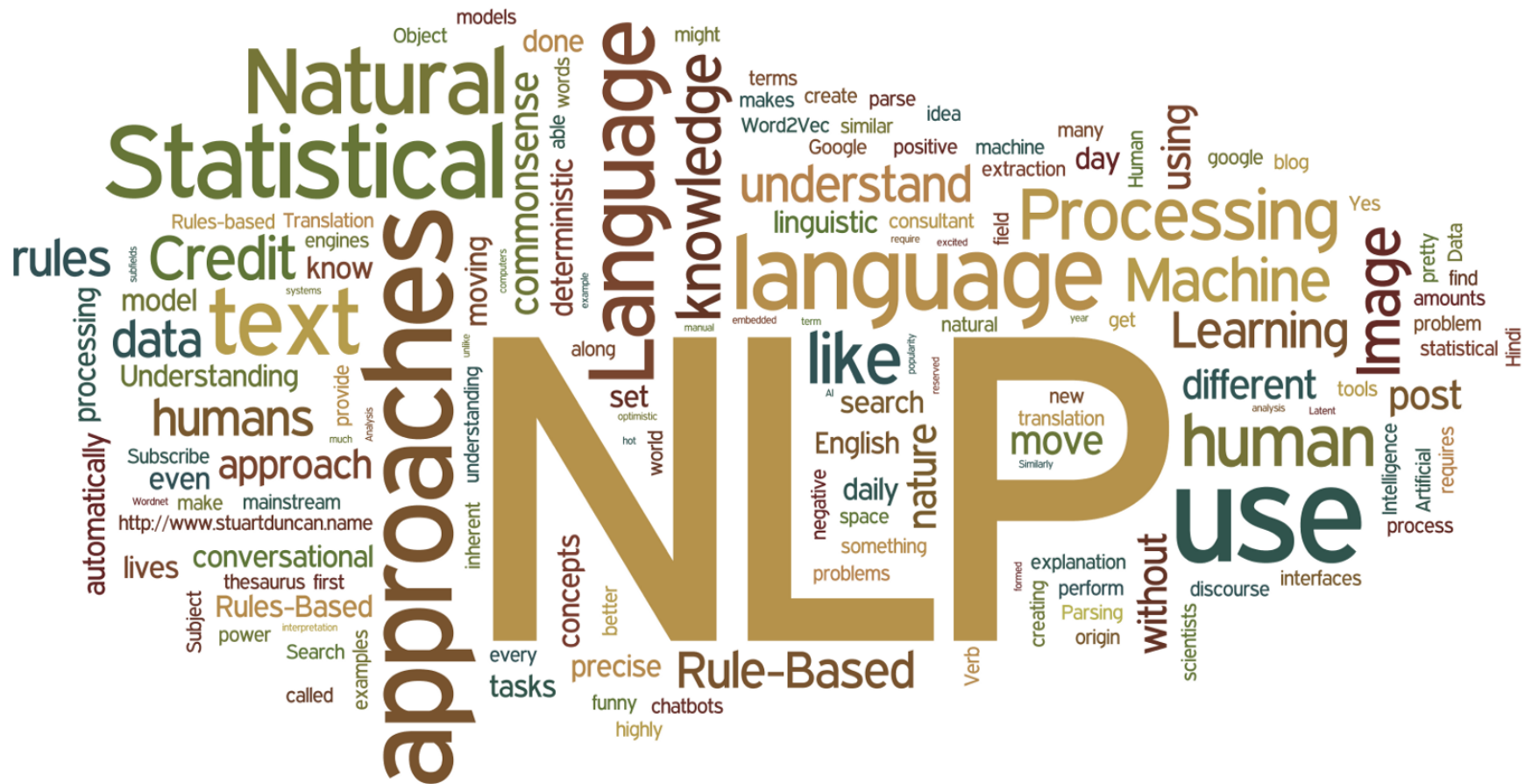
Statistics

likelihood

Probabilities

Bayes theorem

Natural Language Processing



Information Retrieval

Doc A



Doc 1

Doc 2

Doc 3

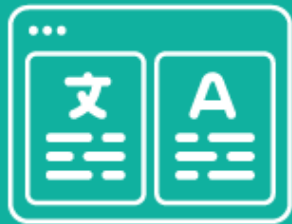
Sentiment Analysis



Information Extraction



Machine Translation



Natural Language Processing

Question Answering



Human: When was Apollo sent to space?



Machine: First flight - AS-201, February 26, 1966

자연언어

- 자연언어란?
 - 정보전달의 수단
 - 인간 고유의 능력
 - 인공언어에 대응되는 개념
 - 특정 집단에서 사용되는 모국어의 집합
 - 한국어, 영어, 불어, 독일어, 스페인어, 일본어, 중국어 등
- 인공언어란?
 - 특정 목적을 위해 인위적으로 만든 언어
 - 자연언어에 비해 엄격한 구문을 가짐
 - 형식언어, 에스페란토어, 프로그래밍 언어



자연언어처리

- 자연언어처리란?
 - 컴퓨터를 통하여 인간의 언어를 처리하고 이용하려는 학문 분야
 - 인간의 언어를 이해하고, 이를 바탕으로 각종 정보처리에 적용함으로써 보다 빠르고 편리한 정보 획득
- 자연언어처리 응용 분야
 - 인간의 언어가 사용되는 실세계의 모든 영역
 - 정보검색, 질의응답 시스템
 - Google, Naver, iphone siri, IBM Watson
 - 기계번역, 자동통역
 - Google 번역기, ETRI 지니톡
 - 문서작성, 문서요약, 문서 분류, 철자 오류 검색 및 수정, 문법 오류 검사 및 수정

형태론적 다양성

- 첨가어
 - 한국어, 일본어, 터키어 등
 - 다수의 형태소가 결합하여 어절 형성
 - 터키어는 평균 7개의 형태소가 결합
- 굴절어
 - 라틴어 (영어, 불어 등은 첨가어와 굴절어의 특징이 모두 있음)
 - 어간이 변함 (영어의 예 : run, ran, run)
- 스와히리어
 - 수(number)를 위한 형태소가 문두에 붙음
 - (예) 사람 : m+tu (단수), wa+tu (복수)
나무 : m+ti (단수), mi+ti (복수)
- 아랍어
 - 자음이 어간이고 모음이 시제, 수 등을 표현
 - (예) ktb(쓰다) kAtAb(능동) KUtlb(수동)
 kttb(쓰게하다) kAttAb(능동)KUttlb(수동)

통사적 다양성

- Postfix 언어 (Head-Final Languages)
 - 동사가 문장의 뒤에 위치
 - 한국어, 일본어 등
- Infix 언어
 - 동사가 문장의 중간에 위치
 - 영어, 불어 등
- Prefix 언어
 - 동사가 문장의 처음에 위치
 - 아일랜드어

자연언어 분석 단계

자연언어문장

형태소 분석
(Morphological Analysis)

구문 분석
(Syntax Analysis)

의미 분석
(Semantic Analysis)

화용 분석
(Pragmatic Analysis)

분석 결과

어절, 단어, 형태소

- 어절

양쪽에 공백을 갖는 띄어쓰기 단위의 문자열

- 단어 / 형태소

단일 품사를 갖는 단위 / 사전에 등록되어 있는 색인어의 집합

예: 나는 책을 읽었다.

파랗파랗한 색이 나는 계절이다.

하늘을 나는 새를 보라.

I tried to go to school.

He tries to pass the exam.

나 + 는

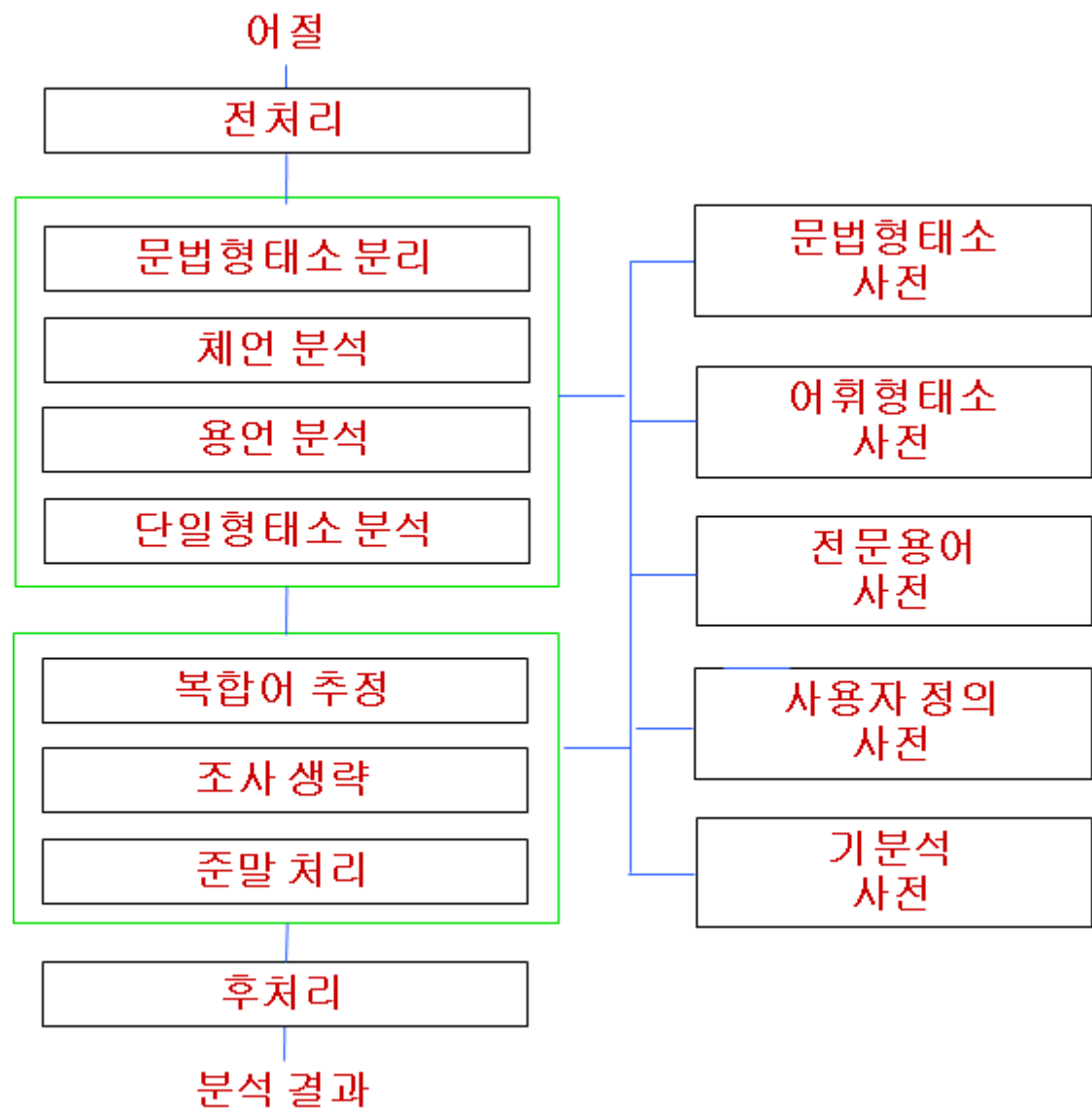
날다 + 는

나다 + 는

형태소 분석 (Morphological Analysis)

- 입력된 문자열을 분석하여
형태소(morpheme)라는 최소 의미 단위로 분리
- 사전 정보와 형태소 결합 정보 이용
- 정규 문법(Regular Grammar)으로 분석 가능
- 언어에 따라 난이도가 다름
 - 영어, 불어 : 쉬움
 - 한국어, 일본어, 아랍어, 터키어 : 어려움

소분류 범주	중분류 범주	대분류 범주
명사	체언	불변화사
대명사		
수사		
관형사	수식언	
부사		
조사		
감탄사		
동사	용언	변화사
형용사		



형태소 분석의 난점

- 중의성 (ambiguity)
 - "감기는"의 분석 결과
 - 감기(명사:cold) + 는(조사)
 - 감(동사 어간) + 기(명사화 어미) + 는 (조사)
 - 감(동사 어간) + 기는(어미)
- 접두사, 접미사 처리
- 고유명사, 사전에 등록되지 않은 단어 처리
 - 한국어, 독일어처럼 복합명사 내의 명사를 띄우지 않거나, 일본어처럼 띄어쓰기가 없으면 더욱 어려워짐
- 한국어 형태소 결합의 예 ("친구에게서였었다라고")
 - 친구(명사) + 에게(조사) + 서(조사) + 이(서술격조사) +
 - 였(과거시제어미) + 었(회상어미) + 다(어말어미) +
 - 라고(인용격조사)

형태소 분석과 품사 태거 예

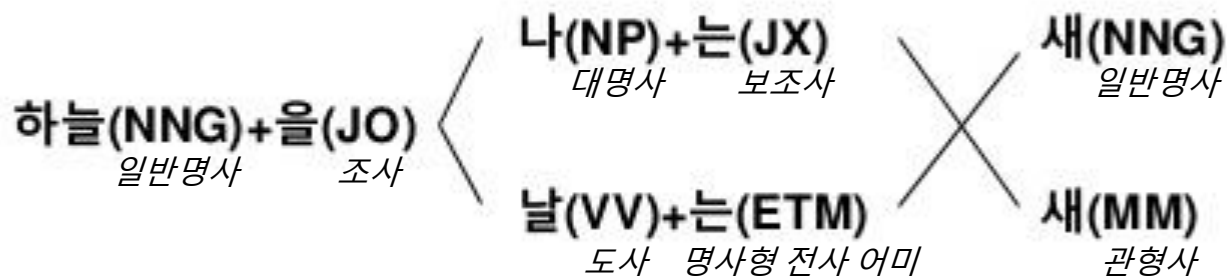
입력 문장: 하늘을 나는 새

하늘을

나는

새

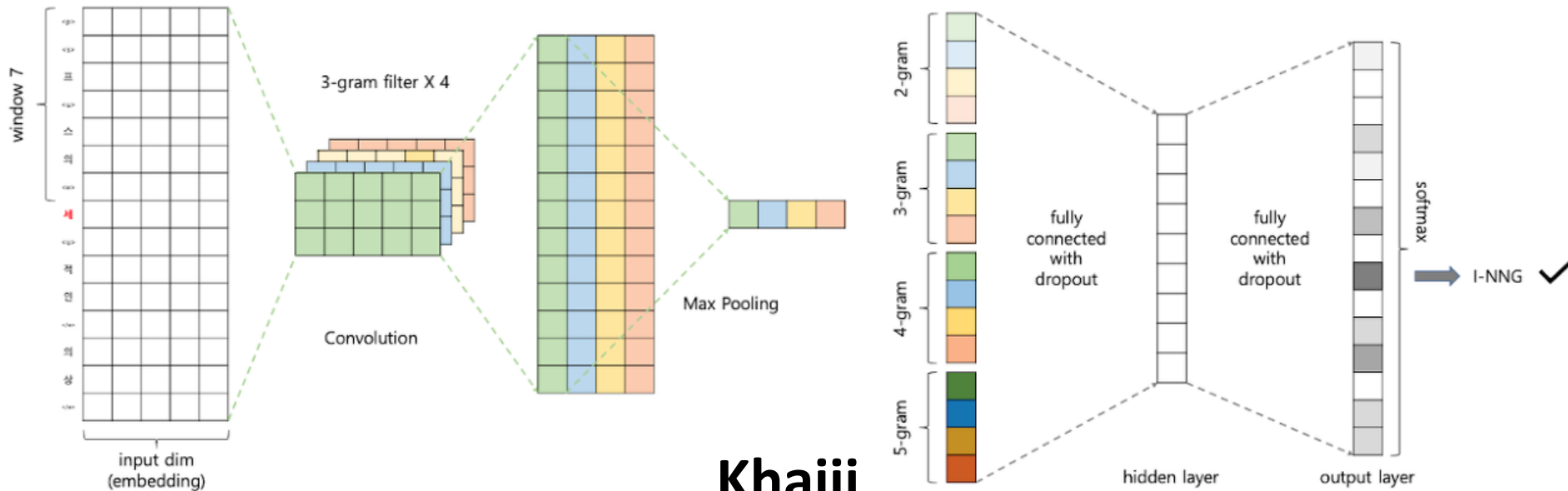
형태소
분석



품사
태거

하늘(NNG)+을(JO) 나(NP)+는(JX) 새(NNG)
하늘(NNG)+을(JO) 나(NP)+는(JX) 새(MM)
하늘(NNG)+을(JO) 날(VV)+는(ETM) 새(NNG)
하늘(NNG)+을(JO) 날(VV)+는(ETM) 새(MM)

kakao

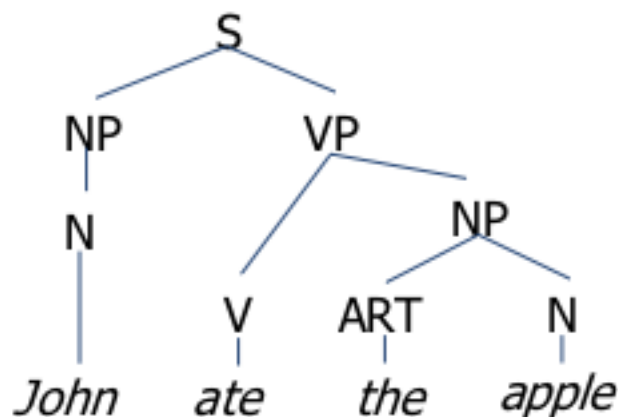


Khaiiii

(Kakao Hangul Analyzer III)

문법, 구문 분석

- 문법 (Grammar) :
 - 문장의 구조적 성질을 규칙으로 표현한 것
- 구문 분석기 (Parser) :
 - 문법을 이용하여 문장의 구조를 찾아내는 process
 - 문장의 구문 구조는 Tree 형태로 표현할 수 있다. 즉, 몇 개의 형태소들이 모여서 구문 요소(구: phrase)를 이루고, 그 구문 요소들간의 결합구조를 Tree형태로써 구문 구조를 이루게 된다.



문법 (Grammars)

- Grammar : a set of rewrite rules

(ex) $S \longrightarrow NP\ VP$
 $NP \longrightarrow ART\ N$
 $NP \longrightarrow N$
 $VP \longrightarrow V\ NP$

- Context Free Grammar (CFG) :
 - 각 rule의 LHS(Left-Hand side)가 하나의 symbol로 이루어진 문법 규칙
- Grammar Rule 을 이용해서 문장(sentence)을 생성할 수도 있고(sentence generation), 분석할 수도 있다(sentence parsing).

Sentence Generation

(ex) By rewrite rule

$S \rightarrow NP \ VP$

$\rightarrow N \ VP$

$\rightarrow John \ VP$

$\rightarrow John \ V \ NP$

$\rightarrow John \ ate \ ART \ N$

$\rightarrow John \ ate \ the \ N$

$\rightarrow John \ ate \ the \ apple.$

Bottom-up Parsing

(ex) *John ate the apple.*

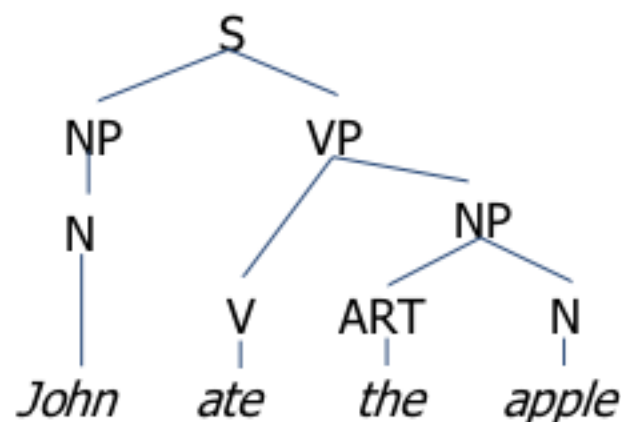
→ N V ART N

→ NP V ART N

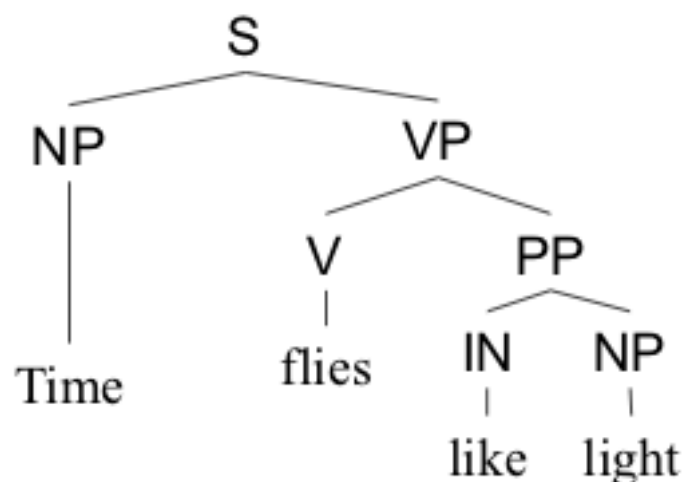
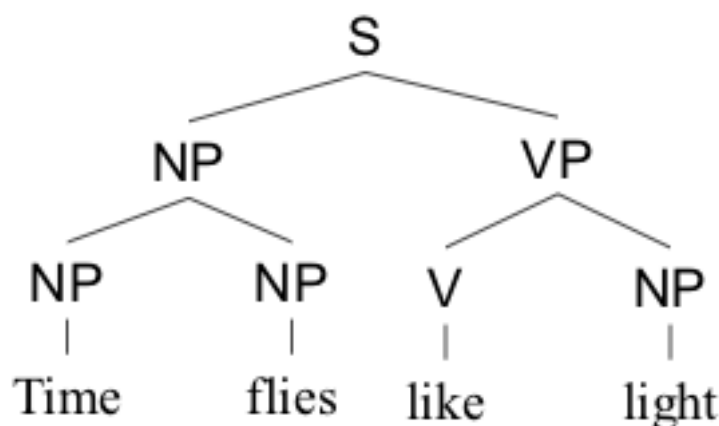
→ NP V NP

→ NP VP

→ S



구문 분석 - Structural Ambiguities



- Structural Ambiguities

- Time flies like light. \Rightarrow 2가지 이상의 구조로 분석됨
 - flies (noun or verb), like(verb or preposition)
- A man see a woman with a telescope on the hill. \Rightarrow 5가지 이상



"I saw the man on the hill with the telescope"

I saw the man. The man was on the hill. I was using a telescope.

I saw the man. I was on the hill. I was using a telescope.

I saw the man. The man was on the hill. The hill had a telescope.

I saw the man. I was on the hill. The hill had a telescope.

I saw the man. The man was on the hill. I saw him using a telescope.

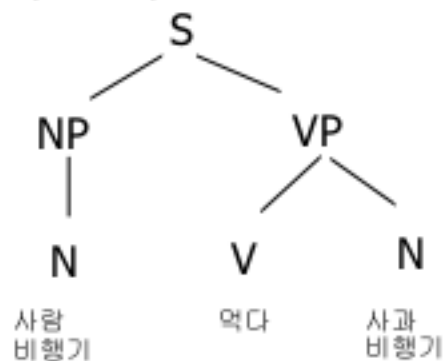
의미 분석 (Semantic Analysis)

- 통사 분석 결과에 해석을 가하여 문장이 가진 의미를 분석
- 형태소가 가진 의미를 표현하는 지식 표현 기법이 요구됨
- 통사적으로 옳으나 의미적으로 틀린 문장이 있을 수 있음
 - 돌이 걸어간다 (cf. 사람이 걸어간다)
 - 바람이 달린다 (cf. 말이 달린다)
- Ambiguity
 - 말이 많다 (horse, speech)

의미 분석 – cont'd

- 문법적으로는 맞지만 의미적으로 틀린 문장들
 - 사람이 사과를 먹는다. (o)
 - 사람이 비행기를 먹는다. (x)
 - 비행기가 사과를 먹는다. (x)

구문 구조



의미적 제약

[먹다

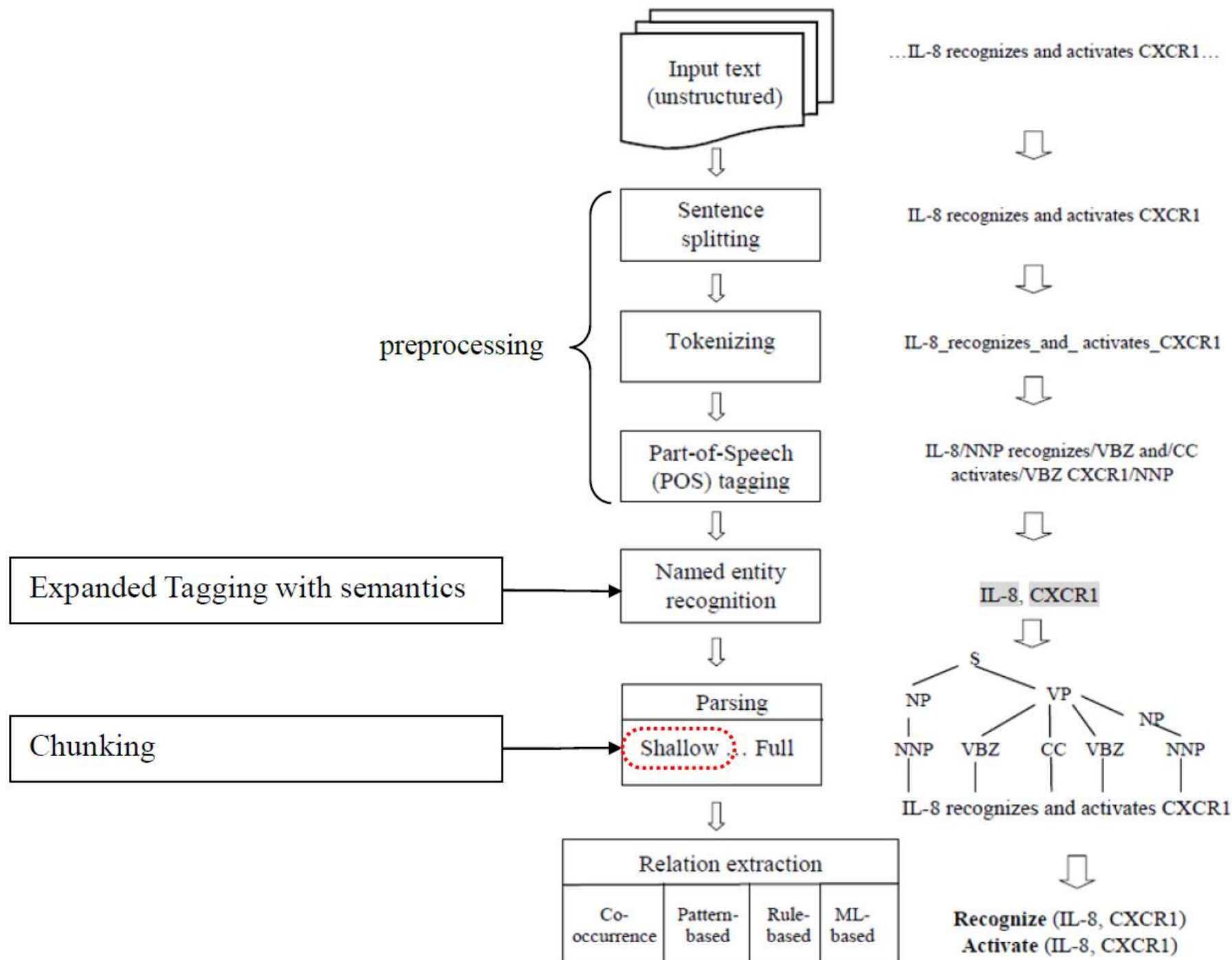
[agent : 먹을 수 있는 주체
object : 먹을 수 있는 대상
....]]

화용 분석 (Pragmatic Analysis)

- 문장이 실세계(real world)와 가지는 연관관계 분석
- 실세계 지식과 상식의 표현이 요구됨
- 지시(anaphora), 간접화법(indirect speech act) 등의 분석
 - Anaphora : 대명사의 지시 대상

The city councilmen refused the women a permit because
(1) *they* feared violence.
(2) *they* advocated revolution.
 - Speech Act : 상대방에게 행동을 요구하는 언어 행위

Can you give me a salt?
Would you mind opening the window?

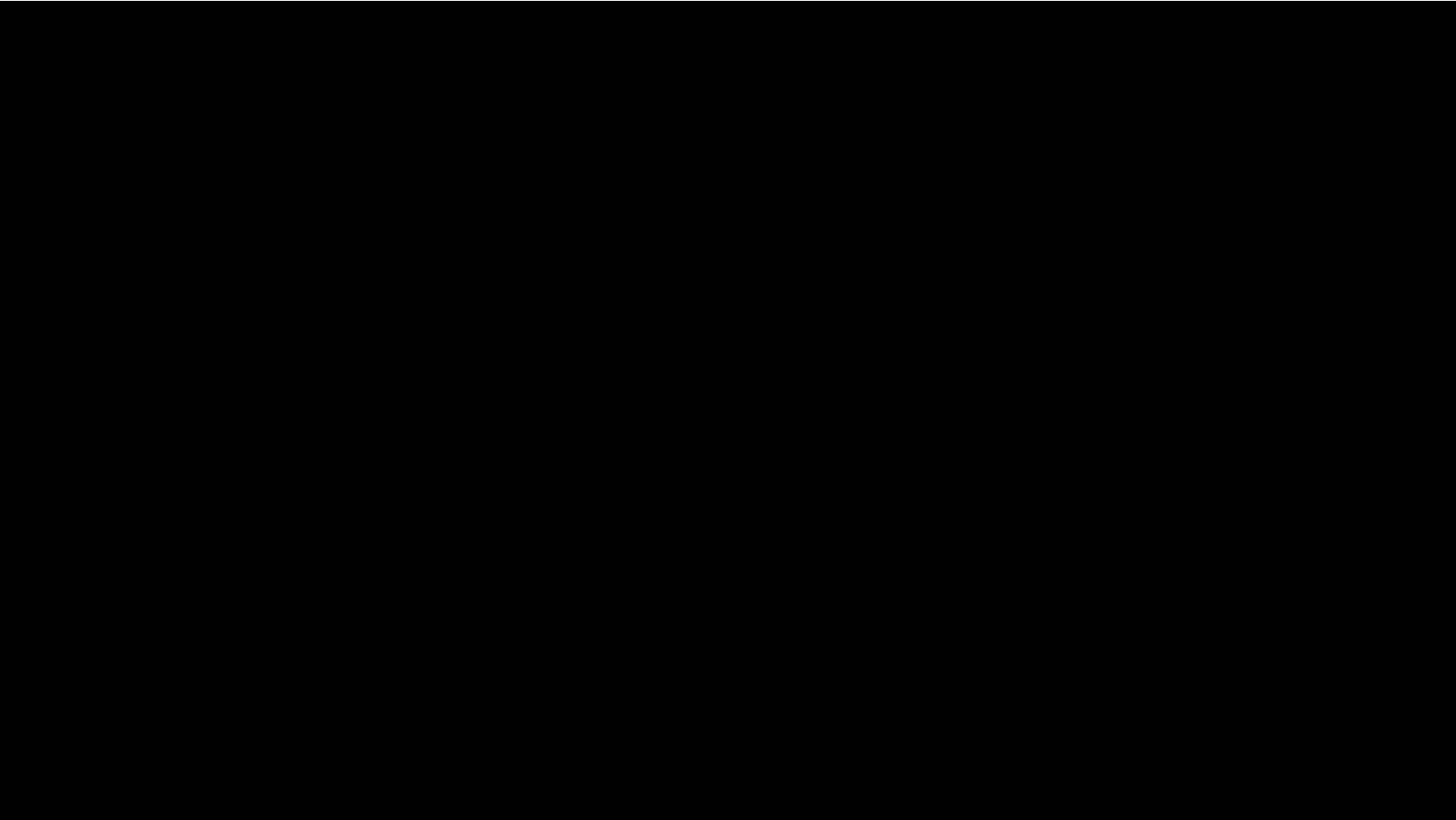


자연어처리, 인공지능, 기계학습

- 인공지능
 - 인공지능은 철학적으로 인간이나 지성을 갖춘 존재, 혹은 시스템에 의해 만들어진 지능을 뜻한다
 - 다양한 연구주제
 - 지식표현, 탐색, 추론, 문제해결, 기계학습, 인지, 행동, 자연언어처리
- 기계학습
 - 기계학습(machine learning)은 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다
 - 가령, 기계학습을 통해서 수신한 이메일이 스팸인지 아닌지를 구분할 수 있도록 훈련할 수 있다

자연어처리, 인공지능, 기계학습 – cont'd

- 인공지능의 연구분야로서의 자연언어처리
 - 음성인식, 형태소분석, 통사분석, 의미분석
 - 언어이해 → 인공지능
- 자연언어처리를 위한 인공지능 기법
 - 형태론, 구문론, 의미론, 화용론적 언어지식 → 지식표현 (WordNet)
 - 자연언어처리 문제 해결 → 기계학습
 - 자연어처리 문제 == 중의성 해소 → 분류문제
 - 품사, 구문, 의미 중의성 해소, 전치사 접속 결정 등
 - 기계학습 알고리즘
 - 결정 트리(decision tree), 선형 분리자(Snow, Perceptron), SVM, Maximum Entropy, HMM, K-NN, 신경망(Neural Network), 딥 러닝(Deep Learning), ...

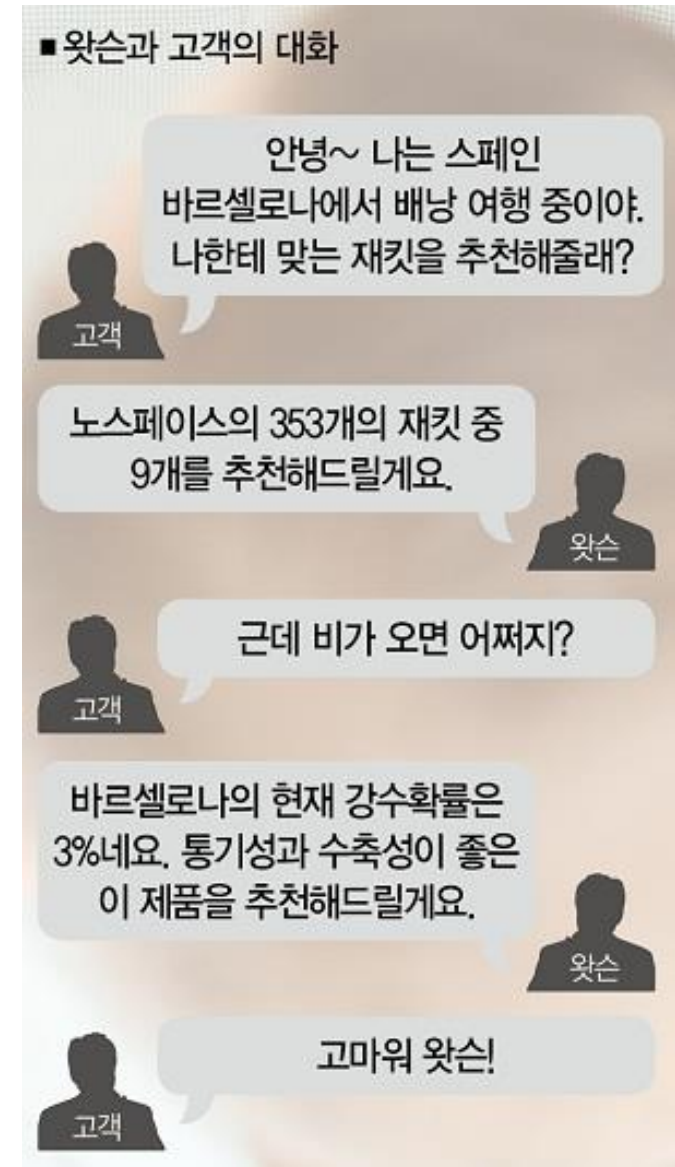


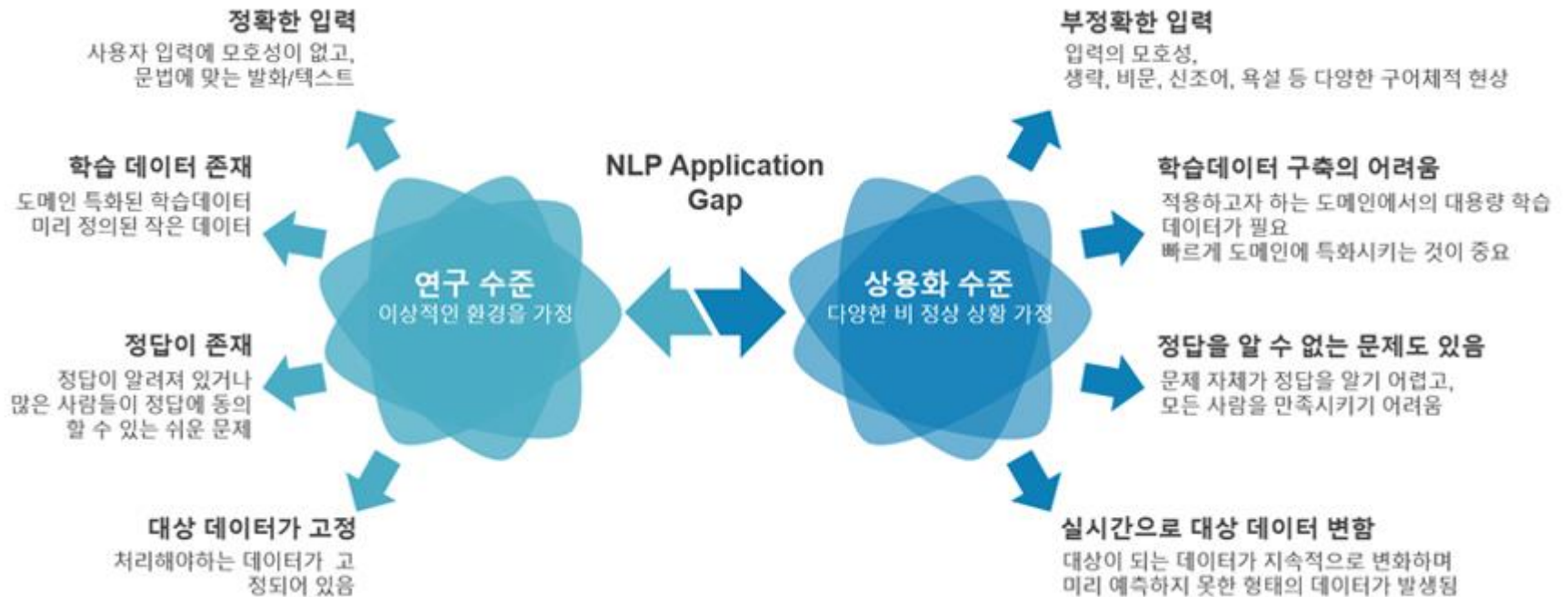


챗봇은 인간의 감성과 언어를 이해하고 학습하기 위해
자연어 처리 (Natural Language Processing),
텍스트 마이닝(Text Mining),
패턴 인식, 상황 인지 (Context-Awareness) 기술을
기반으로 합니다



VS.

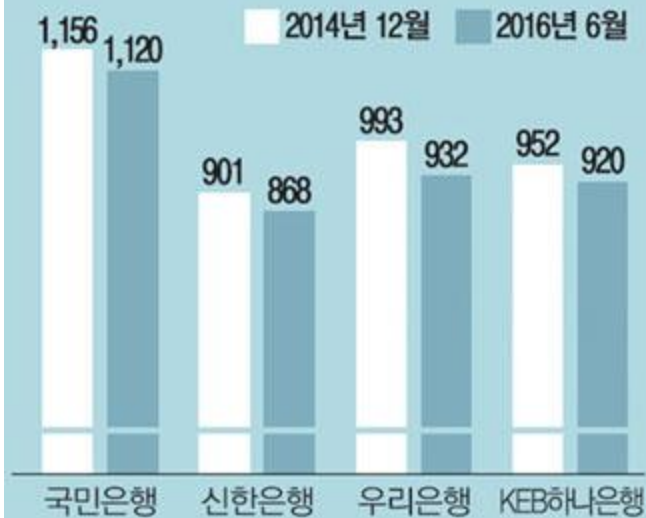




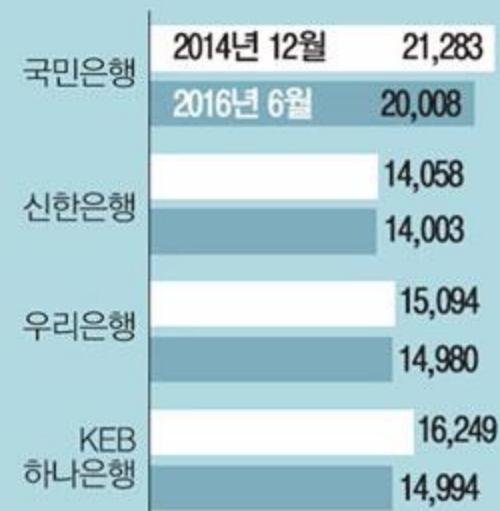
금융권 챗봇 서비스 현황

은행	서비스	내용
NH농협은행	NH농협은행 금융봇	금융상품 안내, 자주 묻는 질문(FAQ), 이벤트 안내 등 고객 상담
신한은행	인공지능 금융 상담봇	상담원 근무 시간 이후 챗봇을 통한 24시간 상담 제공
우리은행	챗봇 헬프데스크	직원들을 대상으로 단말시스템 규정 등을 상담
신한카드	판(FAN)페이봇	카드사용내역 자동 분류, 소비 관리 어드바이스
라이나생명	챗봇 서비스	자주 묻는 질문(FAQ), 가입 상품 안내

은행권 점포도 줄고 (단위=개)



직원도 줄고 (단위=명)





Siri



Bixby

구분 \ 단계	1단계	2단계	3단계
	챗봇 서비스 (Chatbot)	지능형 비서 (Intelligent Assistant)	감성 비서 (Conscious Assistant)
제공 방식	텍스트, 음성	텍스트, 음성, 시각자료	텍스트, 음성, 시각자료, 행동인지
입력 방식	폐쇄형, 개방형 일부	폐쇄형 일부, 개방형	개방형·폐쇄형 복합사용
주요 기술	패턴매칭, 키워드 및 연관어 추출 등	딥러닝·머신러닝, 자연어 처리, 기타 신기술 융합 등	감성인지기술, 데이터 정형화 기술 등
내용	<ul style="list-style-type: none"> - 학습된 내용에 대한 질의응답 - 사용자와 단순한 형태의 소통 - 검색을 통한 결과 제공 	<ul style="list-style-type: none"> - 사용자의 패턴·상황을 고려한 개인 맞춤형 서비스 제공 - 간단한 업무 처리 	<ul style="list-style-type: none"> - 감정 교류를 통한 서비스 및 각종 서비스에 대한 선제적 대응

