

Data Scientist로 살아가는 Big Data 時代의 散策

최 용 석 崔 容 碩(부산대학교 통계학과)

yschoi.pusan.ac.kr

2018.05.23



Multivariate Statistics &
Data Science Lab.

LG 창원



Contents

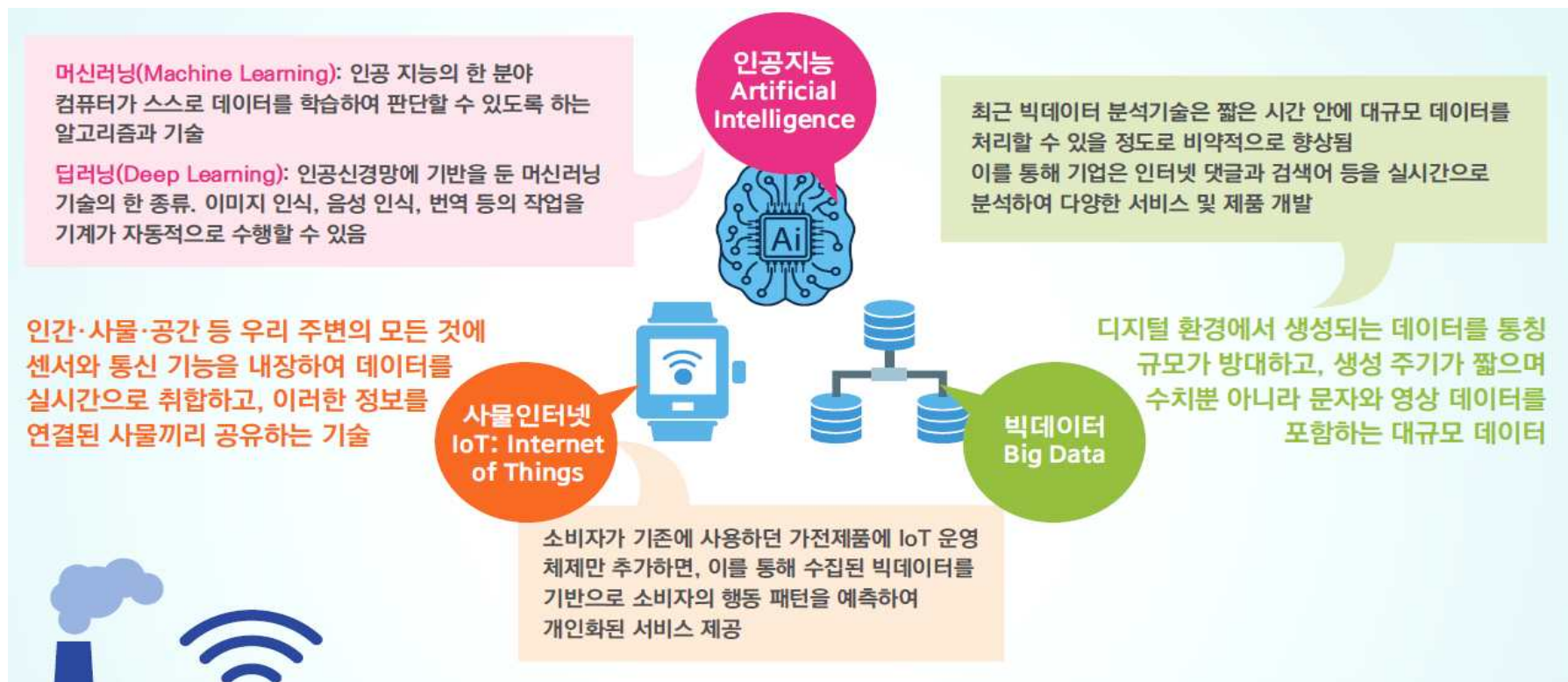
Prolog : 산책 : AI, IoT와 Big Data가 화두가 되는 4차 산업혁명

- ✓ 4차 산업혁명의 본격화 : 역사로 보는 산업혁명
- ✓ Walk on Data Science Age : Data Science/Data Scientist
- ✓ Qualification and Aptitude for Data Scientist :
8 Skills/Aptitude/Environments
- ✓ 우리 과학의 현실/과학에서의 조작
- ✓ My Thoughts about Big Data and Data Science
- ✓ Quantification : 혁명적 발상의 전환/우리나라(조선)의 수량화
- ✓ Big Data : Unstructured Text Data

Epilog : 대한민국 기술발전에서 인간에 대한 깊은 이해를 가진 통찰적 지식인을 기대한다.

Prolog: 산책 散策: (휴식을 취하거나 건강을 위해서) 천천히 걷는 일.

AI, IoT와 Big Data가 화두가 되는 4차 산업혁명



한국의 현실: 인터넷 속도(1위), 전자정부수준(3위), 인프라(세계 최고)

- ✓ 미래 성장동력 '4차 산업혁명' 분야 AI, IoT, Big Data, 3D프린팅, 지능형 로봇 중 4개에서 미/중/일에 비해 기술수준(기술격차) 가장 낮다.
- ✓ 빅데이터 활용 후진국(56위, 자료: 국제경영개발대학원)
- ✓ 개인정보 규제 아시아 최고(OECD 최고) : 여러 부처·법률 복잡하게 얽힘
개인정보보호법(행안부), 통신비밀보호법(과기정통부·법무부), 신용정보의 이용 및
보호법(금융위), 정보통신망 이용촉진 및 정보보호법(과기정통부·방통위) 등 <- 11월
참여연대 등은 신용정보원, 한국신용정보원 등 4개 전문기관과 카드·보험사 등 20개
기업을 개인정보보호법 위반 등 혐의로 검찰에 고발
- ✓ EU: 개인정보, 가명 처리해 활용/미국: 기업이 사전동의 없이 개인정보 수집·분석·활용

한국은 명확한 기준도 없고, 법적으로 모호 (중앙일보, 2018.05.18.)

4차 산업혁명의 본격화 엄청난 빠른 환경 변화

- 2017년 1월 스위스 다보스포럼: 세계경제포럼(WEF)의 연례 총회
주제: **소통과 책임의 리더십(Responsive and Responsible Leadership)**

3차 산업혁명(디지털 혁명)

⇒ 4차 산업(모바일 인터넷, 클라우드 기술, 빅데이터, IoT, AI 등) : 제조업과 정보통신의 융합시대

⇒ 고용 기회 빠른 감소/빈부격차 확대/사회불안 가중/저성장 공고



클라우드 슈바프 세계경제포럼(WEF) 창립자/회장

출처 : <http://www.sedaily.com/NewsView/1OATUOLBKX>



책임을 지지 않는 리더들이 대중인기에 영합하기 위해 포퓰리즘 정책을 남발한다. 경제정책에도 사회적 책임이 수반되어야 한다

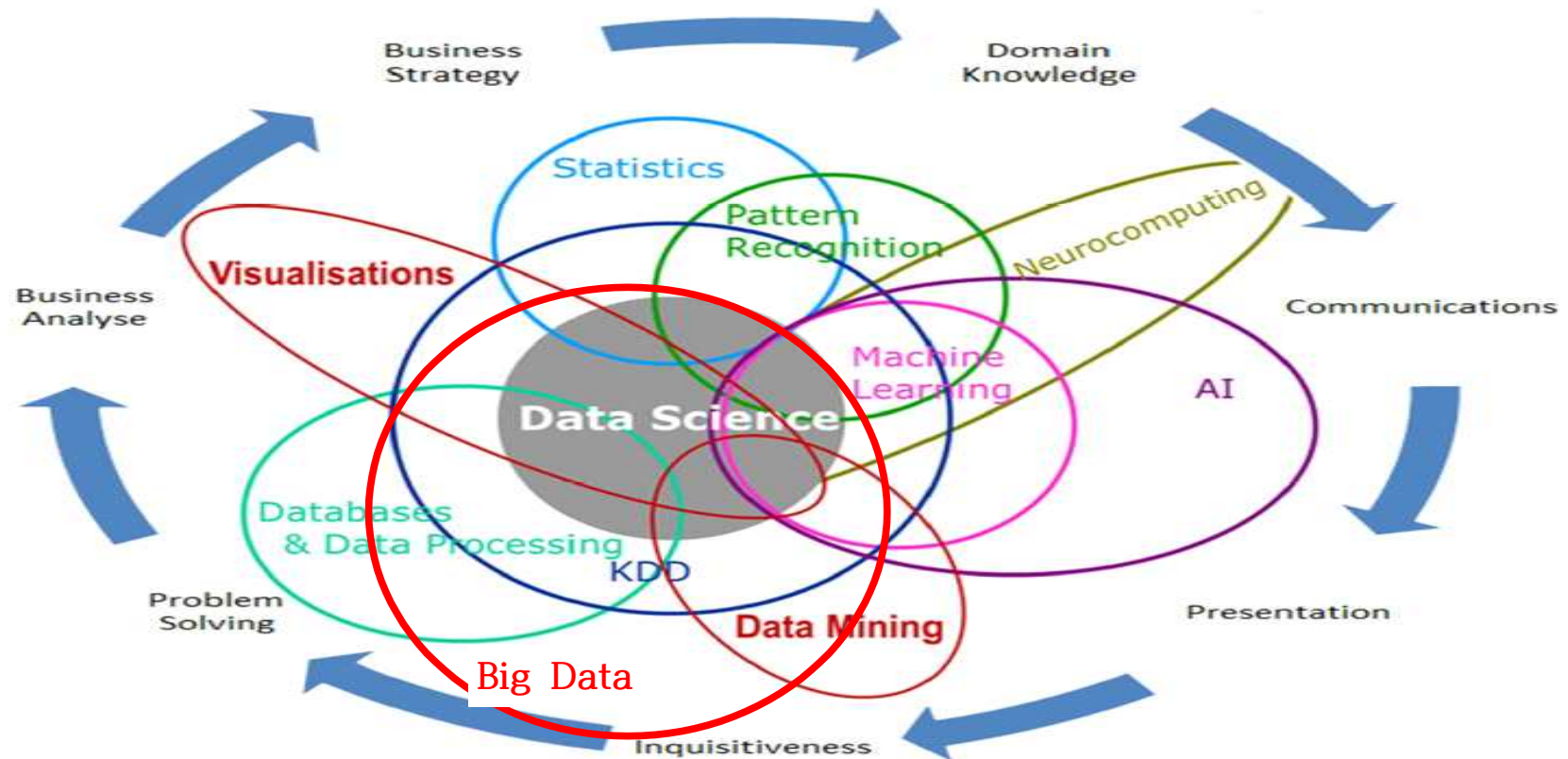
역사로 보는 산업혁명

토인비: 1960~1840 큰 변화의 흐름을 말함

1차 산업혁명 (18세기) 증기기관을 통한 국가 내 연결성 증가	기계발명/자동화/공장생산 체제 섬유산업(영국)/전신 편지 1830년 철도개통(리버풀-맨체스터)	
- 농업사회 -> 공업사회-> 도시화/생산력/소비력 팽창-> 탈 유럽 -> 식민지화: 서양의 힘 - 동양의 패배(1840) -> 1773년 동인도회사(영국): 아편 밀무역 -> 아편전쟁 -> 난징조약(1842) - 1820 산업혁명의 안정기 및 대분기 점: 부국 vs 빈국의 격차 - 일본/대만/한국 : 20세기 기적의 분기점		
4차 산업혁명 (2020년 이후) AI/자동화 연결성 극대화	능동적 판단 작업 수행/다품종 소량생산 알파고-3천만건 기보/1200대 컴퓨터 인터넷 연결/이세돌과 바둑게임 네트워크/Big Data/자가 학습/판단	
정부의 청사진(4차산업혁명위원회): 세계적인 수준의 인공지능(AI) 기술력을 확보 - 투자: AI&뇌연구 : 2조2천억원(18-22년)/AI 반도체개발 : 1조원(20-29년) - AI 대학원 6곳 신설(인재 5000명을 양성)/ 고급인력 : 1370(22년 목표) 문제점: 구체적인 AI산업 발전계획 없음 (한국대학신문, 2018.05.15.)		

Walk on Data Science Age

Multidisciplinary Data Science



<http://www.oralitics.com/2012/06/data-scienceis-multidisciplinary.html>

Data Science

coined at the beginning of the 21C



Cleveland, W.S.
Purdue University

Cleveland, W. S. (2001). "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics". International Statistical Review / Revue Internationale de Statistique 69 (1).



Hayashi, C.
1998

Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data.

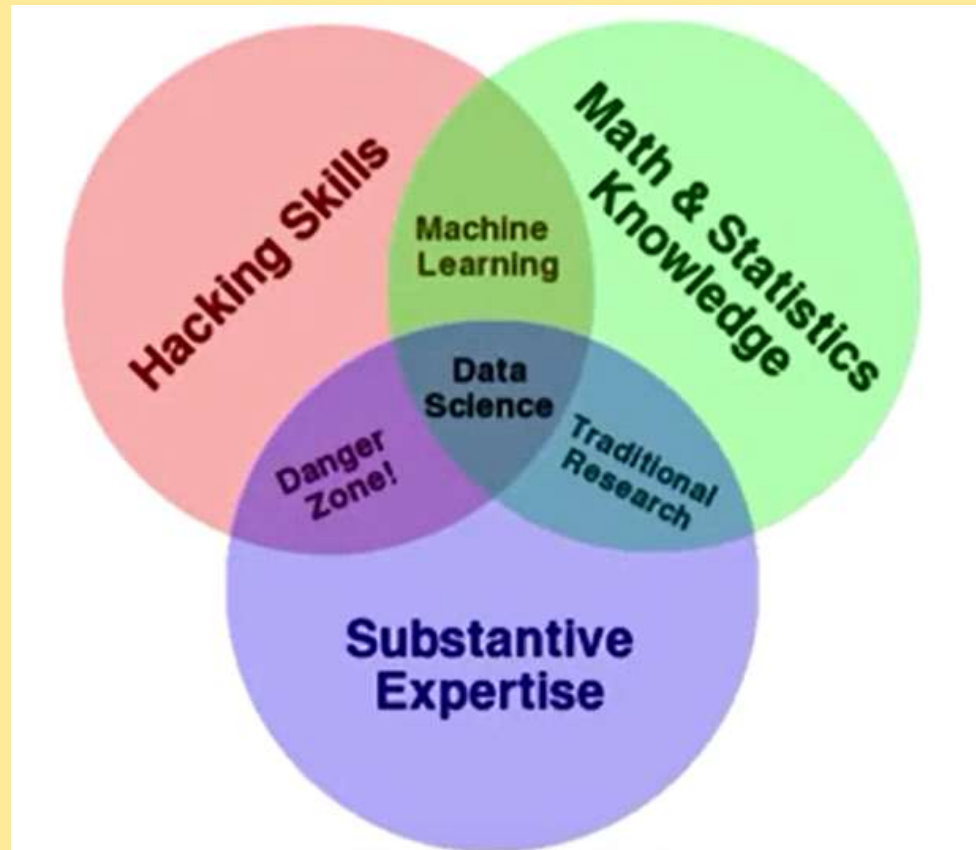
- Quantification I, II, III, IV in 1950s are widely used, especially in Japan, in various fields such as social and marketing surveys, psychological research, medical research, etc.
- Quantification I : Regression Analysis, II : Discriminant Analysis
III : Correspondence Analysis, IV : Multidimensional Scaling

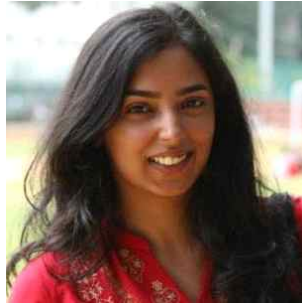
"Data Science" is not Data, it is Science



Leek, J.
2013

Professor, Biostatistics
Johns Hopkins
Bloomberg
School of Public
Health





Monnappa, A,
2018

Lead, Marketing
Content,
Simplilearn,

- Dealing with unstructured and structured data
- Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.
- Data Science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing and aligning the data.

Data Scientist

Coined in 2008



The Sexiest Job of the 21st Century

Davenport and Patil(2012)

Harvard Business Review

- Data scientist is a high-ranking professional with the training and curiosity to make discoveries in the world of big data.
- What data scientists do is make discoveries while swimming in data.
- Think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser.

Google Data Science

전체 이미지 뉴스 동영상 도서 더보기 설정 도구

검색결과 약 19,900,000개 (0.53초)

Azure 빅데이터 분석 | 관계형 데이터 분석 및 시각화
[\[광고\] azure.microsoft.com/ko-kr](https://azure.microsoft.com/ko-kr)
 합리적인 비용, 효율적인 데이터 수집으로 실시간 고급분석까지 지금 바로 무료체험!
 1개월 무료 평가판 사용 - 애저 머신러닝
 SQL 데이터베이스 · Microsoft Azure란? · 다양한 개발 언어 사용

Online Data Science Course | 15,130 People Already Enrolled
[\[광고\] www.udemy.com/Data-Science/Online-Course](https://www.udemy.com/Data-Science/Online-Course)
 Use Real-Life Exercises To Learn **Data Science**. Enroll Today & Get 20% Off!
 On-Demand Courses · Lifetime Access · 30-Day Guarantee
 Courses: Finance, Entrepreneurship, Management, Sales, Data & Analytics, Strategy, Operations,...
 Top Sales Courses · Entrepreneurship Courses · Data & Analytics · All Strategy Courses

Data science - Wikipedia
https://en.wikipedia.org/wiki/Data_science 이 페이지 번역하기
Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, ...
 이 페이지를 18.5.7에 방문했습니다.

Google Big data

전체 이미지 뉴스 동영상 도서 더보기 설정 도구

검색결과 약 42,600,000개 (0.49초)

도움말: **한국어** 검색결과만 검색합니다. 환경설정에서 검색 언어를 지정할 수 있습니다.

빅 데이터 - 위키백과, 우리 모두의 백과사전
https://ko.wikipedia.org/wiki/빅_데이터
 빅 데이터(영어: **big data**)란 기존 데이터베이스 관리도구의 능력을 넘어서는 대량(수십 테라바이트)의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 ...
 활용사례 및 의의 · 정치 · 문화 · 과학기술 및 활용

Big data - Wikipedia
https://en.wikipedia.org/wiki/Big_data 이 페이지 번역하기
Big data is data sets that are so voluminous and complex that traditional data-processing application software are inadequate to deal with them. **Big data** ...
 Big data · Big Data (band) · List of big data companies · Big Data Maturity Model

What Is Big Data? | SAS US
https://www.sas.com/Home/SAS_Insights/Big_Data_Insights 이 페이지 번역하기
Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the ...

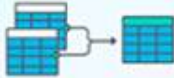
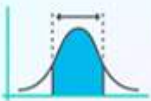
약 2,000만 건

기준: 2018.05.10. 5:07 PM

약 4,260만 건

Qualification and Aptitude for Data Scientist

For the Data Scientist : 8 Skills(Qualification)

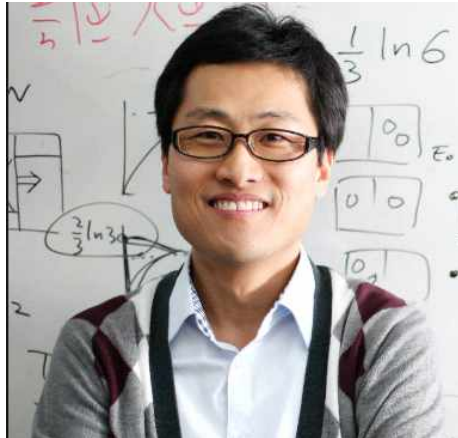
	Programming Skills - Python/Hadoop/MapReduce - SQL : 데이터베이스(dB) 쿼리 - Text data : Xpdf		Data Wrangling - Outliers/Missing values - Inconsistent string formatting - Cleaning& filtering
	Statistics - Multivariate Statistics - EDA		Data Visualization & Communication - ggplot2/Lattice - Biplot/Correspondence Analysis
	Machine Learning - Algorithm 중심 통계학 - Supervised/Unsupervised Learning - Support Vector Machine		Software Engineering - Handling a lot of data - Data-driven products - SAS/R
	Mathematics & Matrix Algebra - 대용량 자료의 표현(행렬)과 연산 - 미분 적분학		Data Intuition - Data-driven problem-solver

Aptitude for the Data Scientist

- 데이터 마인드
- 알고리즘적 사고(수리적 사고/계산적 사고)
- 능동적 태도
- 문제해결능력/창의력
- 사회적 트렌드/과학 기술에 대한 흥미
- 자기학습 및 탐구
- 스토리텔링/인문학적 소양
- 사회적 친화력/협업 소통력/정보교류 : 읽기 말하기 듣기 능력

Environments for the Data Scientist

- 단기적 성과 주의 : 유행의 문제 CRM ⇒ Data Mining ⇒ Big Data
⇒ Data Science (?)
- 용역꾼
- 소통 공간 : 다양한 구성원의 워크숍
- 직장의 이동성
- Big Data와 Small Data의 역할구분
- 법/제도



우리 과학의 현실:

특정 저널에 논문을 실어야 교수가 되거나 연구비를 받을 수 있다. 몇 편 이상의 SCI 논문을 써야 성과급을 준다고 규칙을 정할 수 있다. 하지만 일단 규칙이 만들어지면 과학자들은 수컷 공작과 같이 행동하기 시작한다. 그렇다면 과학자가 거대한 꼬리를 갖는 공작이 되었다고 해서 그를 비난 할 수 있을까?

과학에서의 조작 :

얀 헨드릭 쉰 데이터조작(2000/2001):

네이처, 사이언스 13편 논문 -박사학위박탈(독일 콘스탄츠대학)

황우석 사건(2004) : 줄기세포/사이언스 조작/노벨상 원함

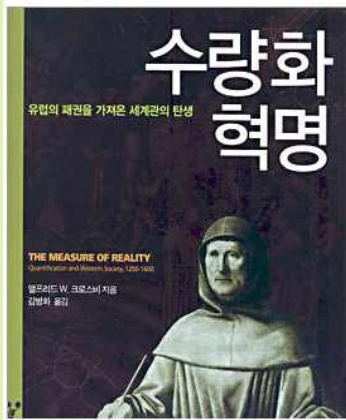
미국의 심장에 태극기를 꽂고 왔다(언론)

올해 초 <네이처> 평가: 한국이 줄기세포 분야에서 약고 평가했다.

Quantification 수량화 : 유럽의 암흑기/지식의 위기

유럽 전 도시는 값 비싼 시계설치 경쟁에 뛰어 들었다(과도한 재정 지출을 감내하고).

시간의 수치화 : 도시의 클래스를 보여 주는 기준(시 재정에 문제가 될 정도).



The Measure of Reality

유럽의 패권을 가져온 세계관의 탄생 Quantification and Western Society, 1250-1600.
앨프리드 W. 크로스비(2005)

- 시간 개념과 기계시계의 등장
시간을 매끄럽게 흐르는 연속체
해시계/물시계/모래시계

- 시간이 돈: 이자가 해와 달의
흐름에 따라 붙어난다.

- 우주의 중심에 지구를 끄집어
내고 태양을 집어넣음(지동설)
(프톨레마이오스: 천동설)

- 원근법: 다빈치와 라파엘로



폴란드 크라쿠프 코페르니쿠스동상

- 금성의 차고 기움 증명: 70년
- 항성의 시차: 3 세기



프라하 시계탑 1410년

혁명적 발상의 전환 : 타 문화 표기 사용

I	1	XXI	21	XLI	41	LXI	61	LXXXI	81
II	2	XXII	22	XLII	42	LXII	62	LXXXII	82
III	3	XXIII	23	XLIII	43	LXIII	63	LXXXIII	83
IV	4	XXIV	24	XLIV	44	LXIV	64	LXXXIV	84
V	5	XXV	25	XLV	45	LXV	65	LXXXV	85
VI	6	XXVI	26	XLVI	46	LXVI	66	LXXXVI	86
VII	7	XXVII	27	XLVII	47	LXVII	67	LXXXVII	87
VIII	8	XXVIII	28	XLVIII	48	LXVIII	68	LXXXVIII	88
IX	9	XXIX	29	XLIX	49	LXIX	69	LXXXIX	89
X	10	XXX	30	L	50	LXX	70	XC	90
XI	11	XXXI	31	LI	51	LXXI	71	XCI	91
XII	12	XXXII	32	LII	52	LXXII	72	XCII	92
XIII	13	XXXIII	33	LIII	53	LXXIII	73	XCIII	93
XIV	14	XXXIV	34	LIV	54	LXXIV	74	XCIV	94
XV	15	XXXV	35	LV	55	LXXV	75	XCV	95
XVI	16	XXXVI	36	LVI	56	LXXVI	76	XCVI	96
XVII	17	XXXVII	37	LVII	57	LXXVII	77	XCVII	97
XVIII	18	XXXVIII	38	LVIII	58	LXXVIII	78	XCVIII	98
XIX	19	XXXIX	39	LIX	59	LXXIX	79	XCIX	99
XX	20	XL	40	LX	60	LXXX	80	C	100
								D	500
								M	1000

- 16세기 유치한 로마식 숫자 표기
⇒ 인도-아라비아 숫자 부기

- 1549 : Mccccxxxviiiij

- 50년 걸림

메디치가(이탈리아어: Medici) :

- 13세기-17세기 피렌체에서 강력한 영향력 가문.
- 세 명의 교황(레오 10세, 클레멘스 7세, 레오 11세)과 통치자(로렌초: 르네상스 예술의 후원자)를 배출
- 혼인을 통해 프랑스와 영국 왕실의 일원
- 도시 정부 피렌체를 지배:
- 예술과 인문주의가 융성한 환경

우리나라의(조선)의 수량화

1) 세제 개혁 여론조사

- 세종12년(1443년 3월) : 전국 17만 2,806명의 양반을 대상으로 찬반 여부를 조사.
- 찬성: 9만 8,657명(57%) vs. 반대 : 7만 4,149명(43%).
- 출처 : 박시형(1940년대) 이조 전세제도의 성립과정, 진단학회.

	찬성	반대
경중(내각)	711	540
한성부(서울)	1123	71
경기도	17105	241
평안도	1332	28510
황해도	4471	15618
충청도	7017	14041
강원도	944	6898
함길도	78	7402
경상도	36317	393
전라도	29547	289

2) 강수량

- 측우기(1441): 세계 최초-장영실-주철, 청동
세종24년 음5월8일 세종실록
- 우량계(1639): 이탈리아 베네데토 카스텔리(benedetto castelli)



건륭경인오월조(乾隆庚寅五月造) 영조 46년(1770)

My Thoughts about (Big) Data and Data Science

- Data = Information = Knowledge
- Big Data = Data : structured and unstructured
- Statistics = Basic Tool of Data Science
- Data Science = Science
- Information and Knowledge from Big Data : Cue and Prior Data
- Self-examination and Innovation

이론적 내용의 습득 또는 지식을 축적하기보다는 사유의 활동 혹은 사유의 높이라는 차원에서 접근

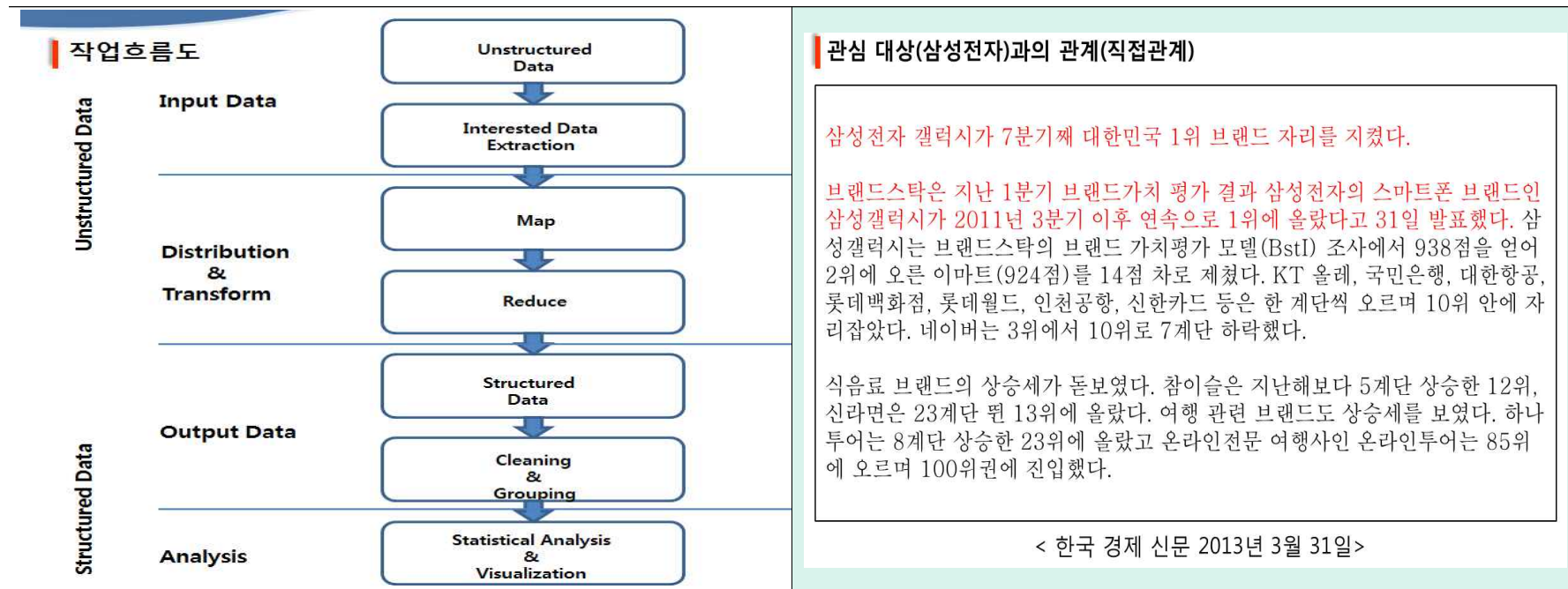
Big Data :: Unstructured Text Data

- 여러 분야의 기관 보고서나 인터넷 상의 기사 문서에 대해 문자로 된 수집된 비정형화자료(unstructured data)이다.
- 이러한 자료는 문서-용어(document-term) 정보를 측정하며 자료는 빈도(frequency)로 이루어진다.
 - 용어의 시각적 연관성 분석 : Word Cloud/SCA/MCA
- 통계분석을 위해서는 정형화자료(structured data)로 전환이 필요
 - 근접성 데이터(proximity data)
 - 군집화를 위한 시각화 기법 : MDS
 - 군집분석 : K-means CA
 - 판별분석 : SVM(Support Vector Machine)

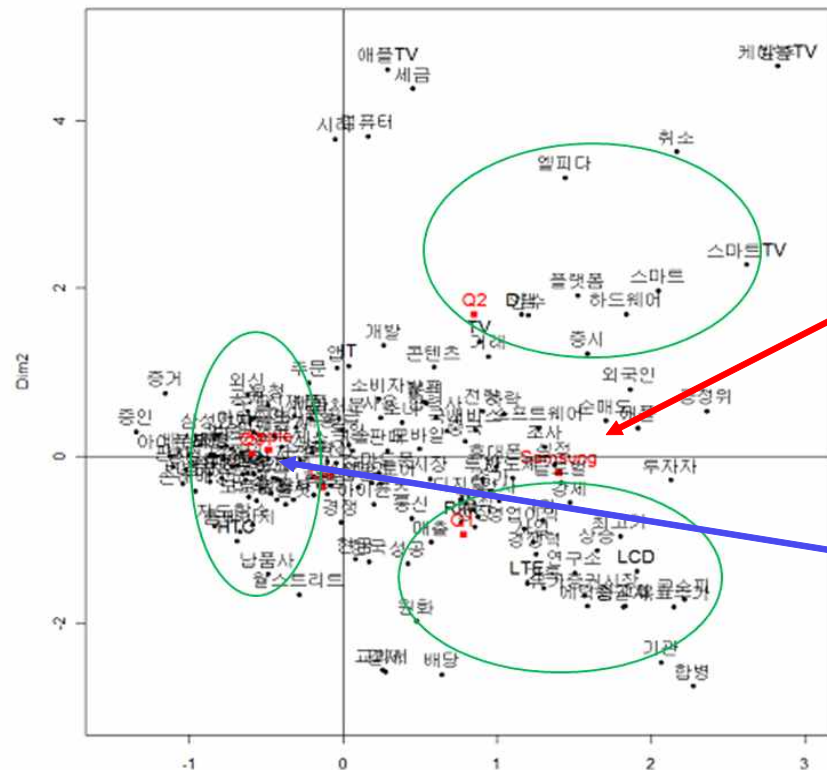
[자료 1] 삼성과 애플의 기사의 맵리듀스와 MCA를 활용한 동향 분석

Choi, J. and Choi, Y.S. (2014). Standardizing Unstructured Big Data and Visual Interpretation Using MapReduce and Correspondence Analysis, *The Korean Journal of Applied Statistics*, 27(2), 169-183.

1) 단어 세분화 및 군집화 (Structured Data) : 두 기업의 분기별 동향과 이슈(2012.01-2012.12, 한국경제신문)



2) 대응분석(Correspondence Analysis)



<삼성전자와 애플의 기사¹⁾ 단어의 다중대응분석도>

기업	분기	사분면	단어	의미
Samsung (삼성전자)	Q1(1분기)	제4사분면	매출, 원화, 성공, 상승, 최고가, 영업이익, 1위, 목표주가, 교과서, 경쟁력, 코스피, 한국, 현금 등	성장
	Q2(2분기)	제1사분면	D램, 하드웨어, 스마트, 스마트TV, 애플TV, 콘텐츠, 개발, 모바일, 소프트웨어 등	제품, 사업확장
Apple (애플)	Q3(3분기)	제2사분면	외신, 증거, 증인, 삼성전자, 요청, 제소, 소송, 판매금지, 배심원 등	특허 소송
	Q4(4분기)	제4사분면	지도, 홈페이지, 손해배상, 변호, 제기, 경쟁사, 모토로라, 판결, 경쟁 등	특허 소송

[자료2]경제 · 인문사회연구소소속10개정부출연연구기관의문서의분류

- 정민지(2017). A Study on Clustering Methods for Proximity Data in Text Mining.
- 정호영(2018). Comparison of Term Weighting Techniques for Text Classification.

[텍스트 군집/분류 제안 프로세스]

[1단계]	비정형 자료 정제화	
	문장부호, 특수문자, 불용어 제거	
[2단계]	문서-용어 빈도행렬 생성(document-term frequency matrix)	
	텍스트 자료 정형화	
[3단계]	문서-용어 가중행렬 생성(document-term weighted matrix)	
	용어 중요도의 수치화를 위한 가중치 부여	
[4단계]	문서-핵심어 가중행렬 생성(document-keyword weighted matrix)	
	불필요한 용어 필터링	
[5단계]	군집화	분류
	MDS와 k-means CA	SVM을 이용하여 개체 분류

[1단계] 비정형 자료의 정제화

크롤링	스크래핑	형태소 분석
암호화되지 않은 PDF 파일	PDF 파일의 문자열 추출	명사 용어 추출
		 <pre> * 데이터 분석 프로그램 : R > m <- tags_matrix("F:/과학기술정책연구원_ tags_5.txt") > m[1:5,] V1 V2 1 창업 1413 2 수 467 3 여성 409 4 것 387 5 기업 344 </pre>
R: X-pdf => 파이썬(Python): KoNLPy(한국어자연어처리),		꼬꼬마(Kkma) 형태소 분석기

[2단계] 문서-용어 빈도행렬 생성(document-term frequency matrix)

2016년도 경제인문사회 연구소 출연 연구기관 문서-용어 빈도행렬

개체 : 경제·인문사회 연구소 소속 10개 정부출연연구기관

문서 : 기관에서 2016년 한 해 동안 발간한 정기간행물 343개 PDF 파일

개 체	ID	용어 문서	개체간 관계								휴양객	흡냄새
			가가	가게	가격	가격대	가공	가능	...			
과학기술정책연구원	1	A01	0	0	0	0	4	3	...	0	0	
대외경제정책연구원	2	A02	0	0	0	0	0	2	...	0	0	
산업연구원	3	A03	0	0	3	0	0	5	...	0	0	
육아정책연구소	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	
정보통신정책연구원	23	A23	0	0	1	0	2	8	...	0	0	
통일연구원	24	B01	0	0	0	0	0	3	...	0	0	
한국개발연구원	25	B02	0	0	3	0	0	3	...	0	0	
한국교육개발원	26	B03	0	0	2	0	0	12	...	0	0	
한국교육과정평가원	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	
한국교통연구원	155	B132	0	0	0	0	0	2	...	0	0	
	156	C01	0	0	0	0	0	1	...	0	0	
	157	C02	0	0	0	0	0	2	...	0	0	
	158	C03	1	0	0	0	0	0	...	0	0	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	

[3단계] 문서-용어 가중행렬 생성(document-term weighted matrix)

TF-IDF(Term Frequency-Inverse Document Frequency)

- 대표적 용어 가중치 기법 알고리즘
- 특정 문서 내에서 용어 빈도가 높을수록, 전체 문서들에는 그 용어를 포함한 문서가 적을수록 TF-IDF 값이 높아짐

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

- $tf_{x,y}$: y 번째 문서에서 x 번째 용어에 대한 빈도
- df_x : x 번째 용어가 포함된 문서의 수, N : 총 문서의 수

ID	문서 \ 용어	가가	가게	가격	가격대	가공	가능	...	휴양객	흙냄새
1	A01	0	0	0	0	8.041	0.573	...	0	0
2	A02	0	0	0	0	0	0.383	...	0	0
3	A03	0	0	3.951	0	0	0.956	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
23	A23	0	0	1.317	0	4.021	1.530	...	0	0
24	B01	0	0	0	0	0	0.573	...	0	0
25	B02	0	0	3.951	0	0	0.573	...	0	0
26	B03	0	0	2.634	0	0	2.295	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
155	B132	0	0	0	0	0	0.383	...	0	0
156	C01	0	0	0	0	0	0.191	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

```

makeTFIDF <- function(data){
  n=nrow(data)
  p=ncol(data)
  TFIDF <- matrix(NA, n, p)
  rownames(TFIDF) <- rownames(data)
  colnames(TFIDF) <- colnames(data)
  for(i in 1:n){
    for(j in 1:p){
      TFIDF[i,j] <- data[i,j]*(log(nrow(data)/sum(data[,j]!=0)))
    }
  }
  TFIDF[is.nan(TFIDF)] <- 0
  return(TFIDF)
}

```

[4단계] 문서-핵심어 가중행렬 $\mathbf{X} = (x_{rit})$, $t = 1, \dots, q$ 생성

[illegible]

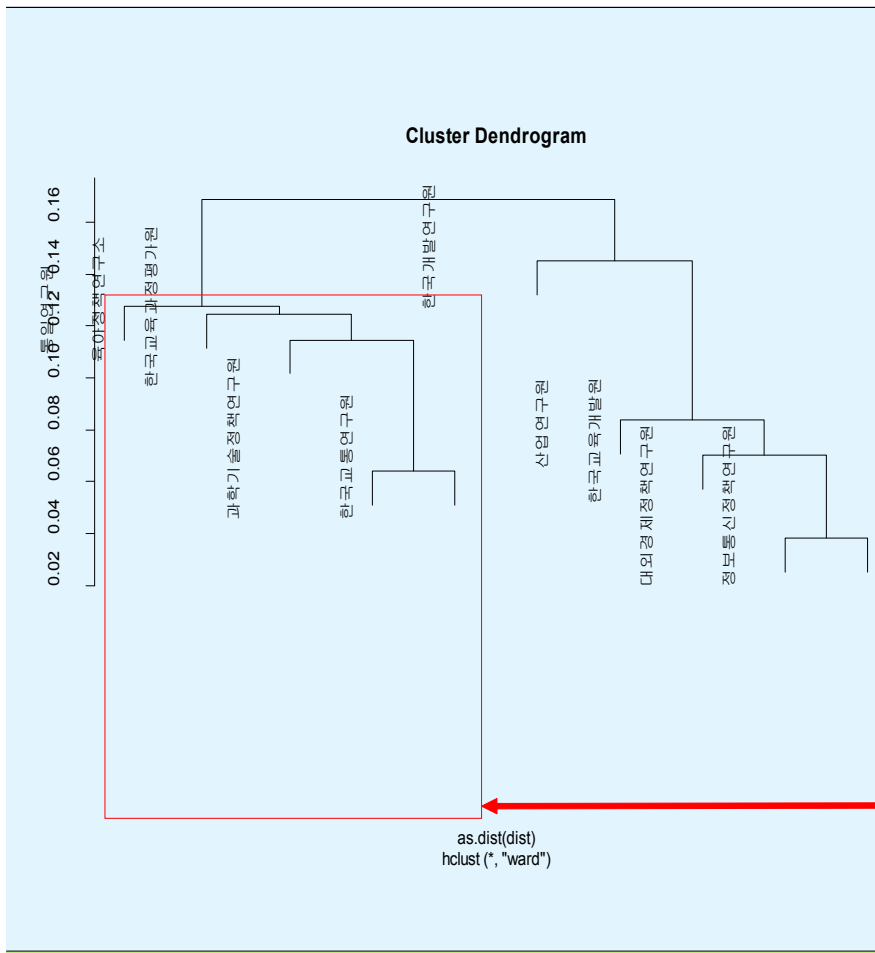
[5단계] 개체 군집화 : Hierarchical CA(Ward Linkage)/Metric MDS (최종 용어 117,820 -> 26,352)

> m3Data_CD	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	B01	B02	B03	B04	B05	B06
A01	0.000000	3.206905	3.446756	3.135831	3.112932	3.425774	3.164416	3.239197	3.120057	3.624077	3.458488	3.243560	5.951523	6.060868	4.809896	6.102394	5.427850	6.851673
A02	3.206905	0.000000	3.447395	3.136805	3.249183	3.372051	3.121716	3.138945	3.078453	3.590831	3.373510	3.224884	5.888912	5.973620	4.765283	6.103320	5.337483	6.658822
A03	3.446756	3.447395	0.000000	3.235597	3.445325	3.479209	3.208754	3.371477	3.282020	3.629488	3.534736	3.306393	6.007731	6.026458	4.875460	6.174279	5.452535	6.920198
A04	3.135831	3.136805	3.235597	0.000000	3.150504	3.212510	2.943465	3.014069	2.964491	3.429519	3.248745	3.072847	5.837448	5.951662	4.716090	6.067601	5.327308	6.807162
A05	3.112932	3.249183	3.445325	3.150504	0.000000	3.330516	3.093575	3.207166	3.151760	3.604940	3.474772	3.318196	5.925739	6.015682	4.785269	6.002871	5.379061	6.837892
A06	3.425774	3.372051	3.479209	3.212510	3.330516	0.000000	3.148407	3.304958	3.220470	3.636024	3.513455	3.370602	5.984741	6.015938	4.869386	6.094700	5.445845	6.894796
A07	3.164416	3.121716	3.208754	2.943465	3.093575	3.148407	0.000000	3.041731	2.938896	3.408736	3.221334	3.112927	5.833370	5.936283	4.673042	6.040274	5.309467	6.765517
A08	3.239197	3.138945	3.371477	3.014069	3.207166	3.304958	3.041731	0.000000	3.029635	3.520874	3.372225	3.181459	5.840920	5.963889	4.647631	6.089577	5.389903	6.832247
A09	3.120057	3.078453	3.282020	2.964491	3.151760	3.220470	2.938896	3.029635	0.000000	3.390363	3.224326	3.002332	5.826901	5.899427	4.664486	6.042246	5.285818	6.761967
A10	3.624077	3.590831	3.629488	3.429519	3.604940	3.636024	3.408736	3.520874	3.390363	0.000000	3.671097	3.573772	6.115112	6.240692	4.957642	6.304374	5.650009	7.017427
A11	3.458488	3.373510	3.534736	3.248745	3.474772	3.513455	3.221334	3.372225	3.224326	3.671097	0.000000	3.407937	6.020923	6.114658	4.902031	6.188610	5.500348	6.898208
A12	3.243560	3.224884	3.306393	3.072847	3.318196	3.370602	3.112927	3.181459	3.002332	3.573772	3.407937	0.000000	5.959055	6.041648	4.763086	6.119602	5.441414	6.852891
B01	5.951523	5.888912	6.007731	5.837448	5.925739	5.984741	5.833370	5.840920	5.826901	6.115112	6.020923	5.959055	0.000000	7.623281	6.747794	7.731151	7.163997	8.347855
B02	6.060868	5.973620	6.026458	5.951662	6.015682	6.015938	5.936283	5.963889	5.899427	6.240692	6.114658	6.041648	7.623281	0.000000	6.779784	7.627627	6.545571	8.325572
B03	4.809896	4.765283	4.875460	4.716090	4.785269	4.869386	4.673042	4.647631	4.664486	4.957642	4.902031	4.763086	6.747794	6.779784	0.000000	6.711388	6.295994	7.588896
B04	6.102394	6.103320	6.174279	6.067601	6.002871	6.094700	6.040274	6.089577	6.042246	6.304374	6.188610	6.119602	7.731151	7.627627	6.711388	0.000000	7.068456	8.491877
B05	5.427850	5.337483	5.452535	5.327308	5.379061	5.445845	5.309467	5.389903	5.285818	5.650009	5.500348	5.441414	7.163997	6.545571	6.295994	7.068456	0.000000	7.975529
B06	6.851673	6.658822	6.920198	6.807162	6.837892	6.894796	6.765517	6.832247	6.761967	7.017427	6.898208	6.852891	8.347855	8.325572	7.588896	8.491877	7.975529	0.000000

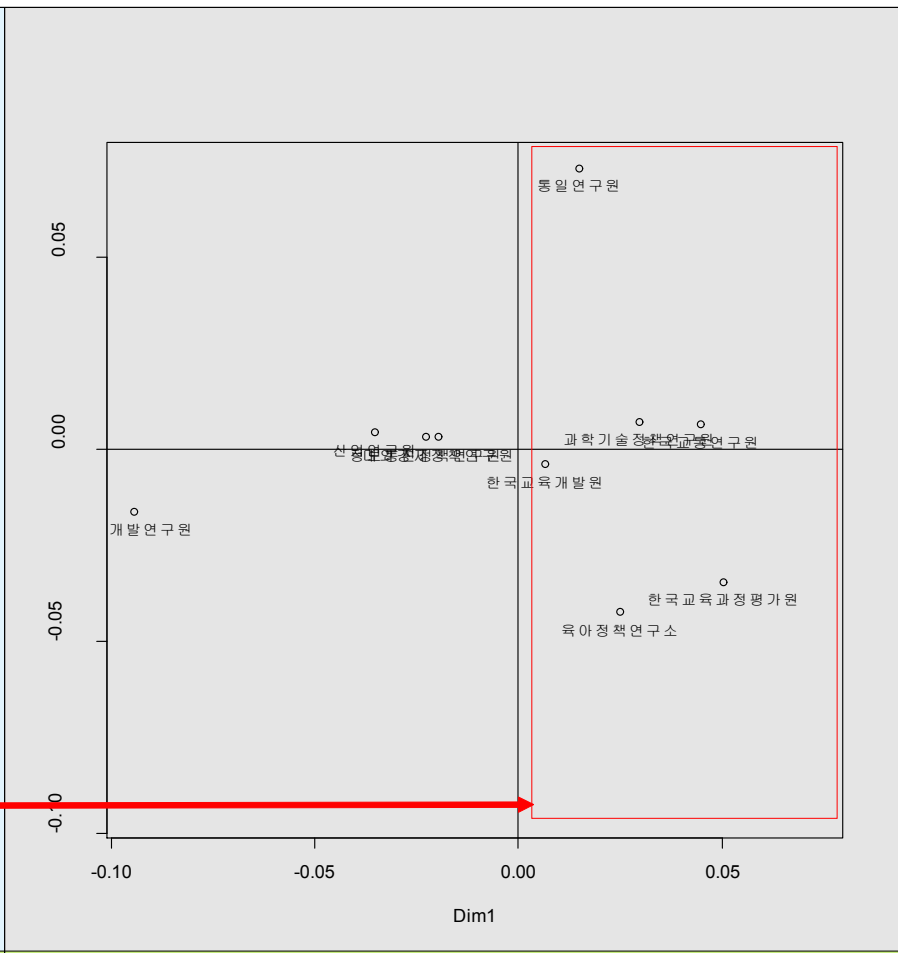
간행물 간의 유사성 : 유클리드거리, 가중유클리드거리

> round(dist,3)	과학기술정책연구원	대외경제정책연구원	산업연구원	육아정책연구소	정보통신정책연구원	통일연구원	한국개발연구원	한국교육개발원	한국교육과정평가원	한국교통연구원
과학기술정책연구원	0.000	0.077	0.100	0.102	0.084	0.099	0.140	0.066	0.099	0.064
대외경제정책연구원	0.077	0.000	0.063	0.101	0.038	0.097	0.107	0.061	0.107	0.097
산업연구원	0.100	0.063	0.000	0.122	0.068	0.121	0.121	0.090	0.132	0.118
육아정책연구소	0.102	0.101	0.122	0.000	0.105	0.130	0.152	0.086	0.125	0.111
정보통신정책연구원	0.084	0.038	0.068	0.105	0.000	0.103	0.111	0.064	0.114	0.103
통일연구원	0.099	0.097	0.121	0.130	0.103	0.000	0.150	0.095	0.132	0.110
한국개발연구원	0.140	0.107	0.121	0.152	0.111	0.150	0.000	0.120	0.161	0.152
한국교육개발원	0.066	0.061	0.090	0.086	0.064	0.095	0.120	0.000	0.100	0.081
한국교육과정평가원	0.099	0.107	0.132	0.125	0.114	0.132	0.161	0.100	0.000	0.105
한국교통연구원	0.064	0.097	0.118	0.111	0.103	0.110	0.152	0.081	0.105	0.000

기관별 유사성 거리 : 기관 내와 기관 간의 거리 평균



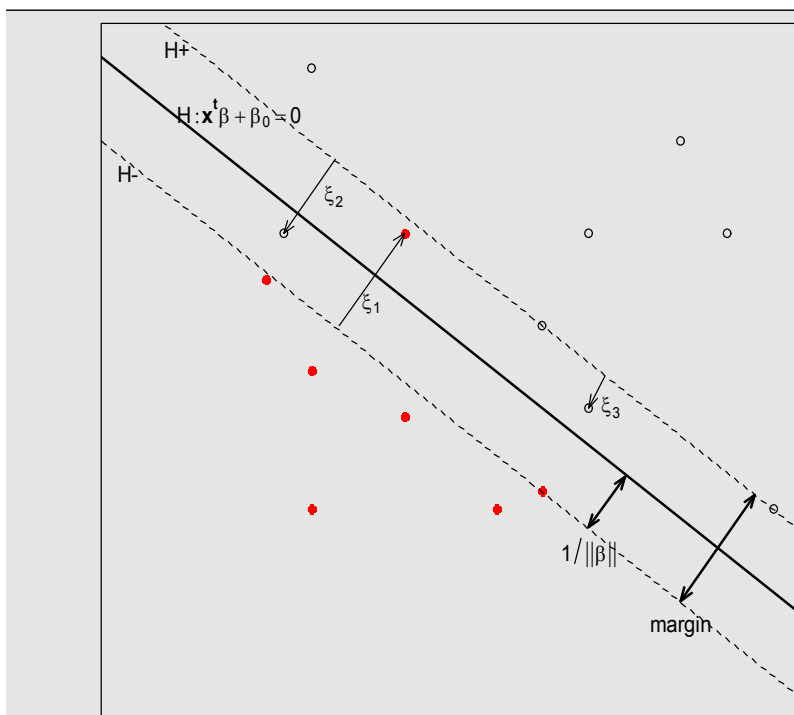
Dendrogram of Hierarchical CA: Ward Linkage



Metric Multidimensional Scaling

[5단계] 개체 분류 : SVM을 활용한 판별분석 (최종 용어 27,203 -> 13,980)

- 텍스트 분류 분석에서 자주 이용되는 machine learning 중 지도학습(supervised learning) 알고리즘 SVM
- 학습 시간이 데이터 개수와 무관, 기존의 분류 방법인 판별분석과 로지스틱 회귀분석보다 예측력이 높음
- 마진을 최대로 하는 초평면 H 을 찾아 범주를 나누는 이진 분류기



$$\mathcal{L} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}, \text{ s.t } \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$$

$$H = \{\mathbf{x} : f(\mathbf{x}) = \beta_0 + \mathbf{x}^t \boldsymbol{\beta}\} : \text{초평면}$$

$$\mathbf{x} = (x_1, \dots, x_p)^t : p\text{차원 변수벡터}$$

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t : \text{가중치 모수 벡터}, \beta_0 : \text{bias 항}$$

$$\text{분류기준: } \hat{f}(\mathbf{x}) > 0, y = +1; \hat{f}(\mathbf{x}) < 0, y = -1$$

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^n \xi_i, \text{ s.t } \xi_i \geq 0, y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i$$

$$, i = 1, \dots, n$$

선형적으로 완전히 분리가 되지 않는 경우

[4 단계]의 문서-핵심어 가중행렬의 SVM을 이용한 10개 기관 문서의 텍스트 분류

- training : test = 80:20의 비율로 기관별 문서에 대해 층화추출

예측	실제									
	A	B	C	D	E	F	G	H	I	J
A	3	0	0	0	0	0	0	0	0	0
B	0	4	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0	0
E	0	0	0	0	3	0	0	0	0	0
F	0	0	0	0	0	2	0	0	0	0
G	0	0	0	0	0	0	3	0	0	0
H	0	1	0	0	0	0	0	26	0	1
I	0	0	0	0	0	0	0	0	8	0
J	0	0	0	0	0	0	0	0	0	5

Accuracy : 0.9677= 60/62

SVM에 의한 10개 기관 문서의 분류표

Epilog :

대한민국의과학기술발전에서인간에대한깊은이해를가진통찰적지식인을기대한다.

최용석(2013) 과학과 기술

나는 오랫동안 수학과 통계학 컴퓨터 알고리즘에 익숙한 삶을 살아왔다. 특히, 난 통계학을 전공하는 학자(Statistician)이며 동시에 난 자연과학자(Natural Scientist)이다. 이대로 이런 지식의 굴레에서 생을 마감 할 것인가를 끊임없이 물어 왔다.

존 브록만(John Brockman)이 저술한 The New Humanist-Science at the Edge(과학의 최전선에서 인문학을 만나다)를 보면 이탈리아의 두 천재 레오나르도 다빈치(1452~1519)와 미켈란젤로(1475~1564)가 언급되고

있다. 다빈치가 살았던 삶을 보면 루브르 박물관의 모나리자를 그린 화가이기도 하지만, 비행기, 헬리콥터, 로봇 등을 발명한 과학자, 도시설계, 공작기계 등을 제안한 공학자였다. 또한 미켈란젤로는 바티칸 박물관 시스티나 성당의 천지창조를 그린 화가이자 조각가 그리고 건축가 또는 시인이기도 하였다.

이들에게 공통적으로 발견하는 것은 모든 분야의 지식에 정통하여 지적 네트워크가 발달한 지식인이라는 점이다. 단순히 이들의 지식이 자신의 업적을 남기고 나타내는 데 사용되지 않고 이들과 같이 살아가는 이웃에게 열려져 있었다는 점이다. 이들에게는 인간이 화두이면서 어떻게 하면 인간이 안전하고 편리하게 그들에게 편안함을 줄 것인가에 그 초점이 모아져 있었다.

이들과 비슷한 삶을 살았던 조선의 레오나르도 다빈치로 평가받는 **다산 정약용(1762~1836)**도 수원화성을 축조한 건축가, 거중기를 발명한 과학자, 목민심서와 같은 시류를 읽는 탁월한 눈을 가진 실학자였다. 그의 업적도 인간에 대한 깊은 이해가 바탕이 된 것이다. 즉, 수원성을 건설하는 데 백성들이 힘들지 않도록 간편함과 편리성을 추구한 거중기와 탐관오리들에게 수탈당하고 굶주려 죽어가는 백성들의 비참한 모습을 보고 목민관이 지켜야 할 사항을 정리한 목민심서인 것이다.

오늘날 전공분야가 전문화, 세분화되어 있어 앞서 언급한 통찰적 지식인을 기대하기는 힘들 수 있다. 그러나 조선, 반도체, 휴대폰에 집중되어 있는 대한민국의 과학기술발전을 보면서 **장기적으로는 과학기술을 다양화하고 경쟁력을 가지기 위한 원천 지식이 요구되며 이를 통찰하고 조정하는 지식인이 필요하다고 여겨진다.** 특히, **의사, 변호사 등 서비스 직종을 선호**

하는 사회적 풍토를 극복하고 이공계 특히 기초연구 분야의 주가 되는 자연과학에 대한 관심을 기울여야 한다.

자연과학 안에는 인간과 밀접한 관계를 맺고 있는 환경, 물리, 화학, 생물, 분석에 관련된 원천적 과학기술을 가진 작은 지식인들이 많다. 선택과 집중을 통한 투자를 늘리기 보다는 이들이 창의적이면서 인간에 대한 깊은 성찰을 통해 편리성과 다양성을 지닌 과학기술을 가져다 줄 진정한 통찰력을 지닌 지식인이 되도록 돕는 것이 대한민국의 과학기술발전을 위한 초석이 될 것으로 기대한다.

나는 진정한 Data Scientist인가 ?