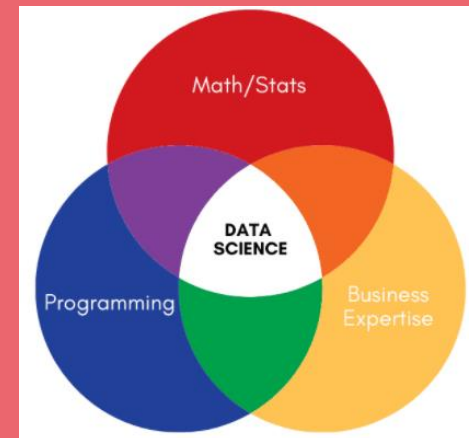


빅데이터, 스몰 데이터

인하대학교
데이터사이언스 학과
김 승 환

swkim4610@inha.ac.kr





- 뉴욕증권거래소 : 1일에 1테라 바이트의 거래 데이터생성
- Facebook : 100억장의 사진, 수 페타바이트의 스토리지
- 통신사 : 시간당 10G 이상의 통화 데이터, 1일240G 생성, 월 생성 데이터의 크기 200T 이상

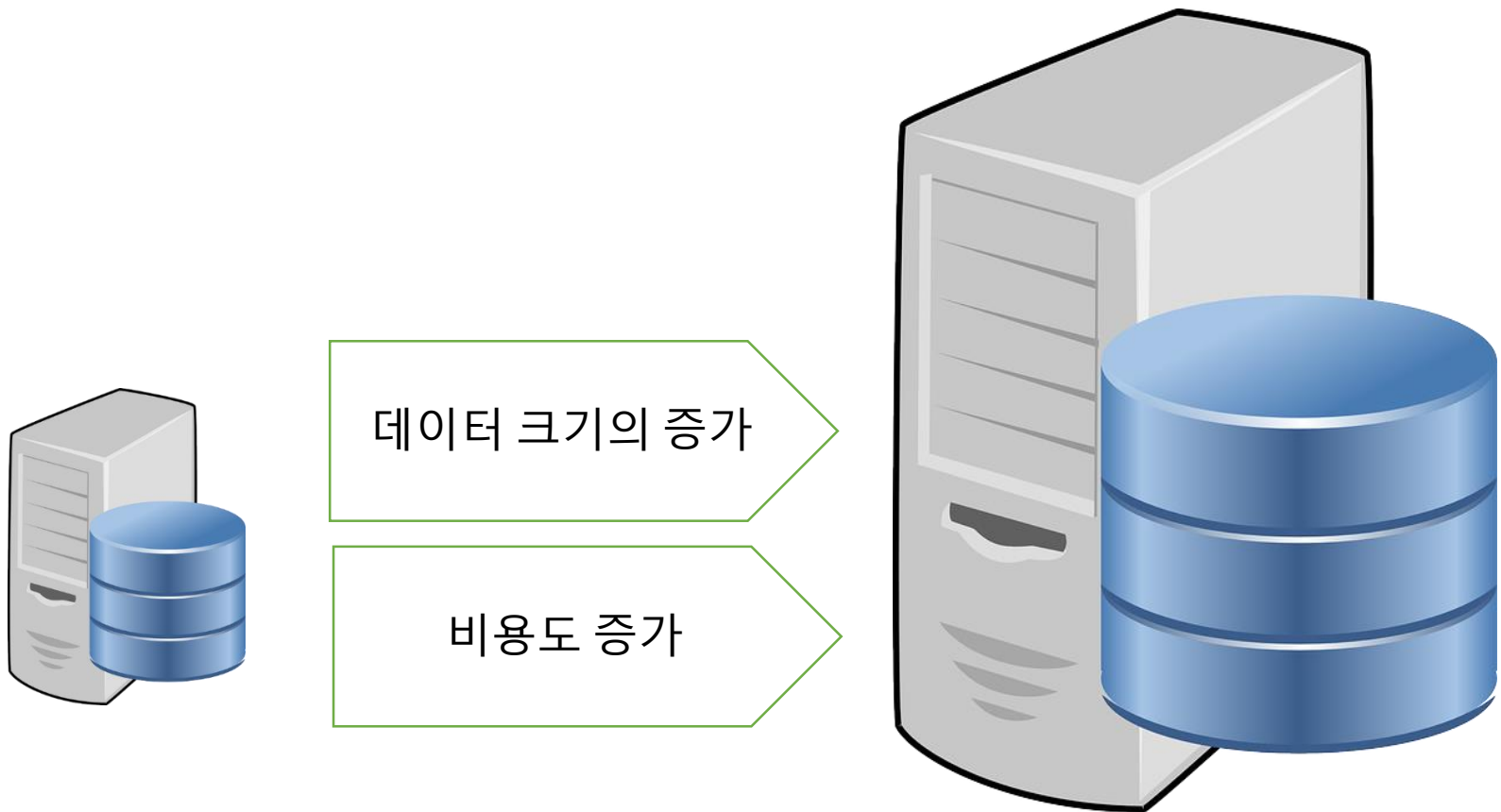


컴퓨터|SW|게임>데스크탑 클릭 <http://www.hmall.com/CS/handler/hmall/kr/View3DepthSect-Start?SectID=1020>



공기청정기 검색

- ✓ 2000년 초반부터 기업은 데이터베이스에 고객정보를 축적하기 시작함
- ✓ DBM/CRM을 통해 Acquisition, Retention, Up Sell, Cross Sell을 유도하기 위해 데이터를 수집
- ✓ 또한, 신속한 경영성과 분석 및 집계를 위함
- ✓ 2010년 이후, 본격적인 모바일 시대가 열리면서 기업 환경도 변화됨
- ✓ 오프라인 매출이 감소하고 온라인, 모바일 매출이 늘어나면서 온라인 모바일 채널의 중요성 부각
- ✓ 초반에는 온라인 모바일 채널에서 고객정보 역시, 카드 정보처럼 매출정보만 관리하였음
- ✓ 빅데이터 기술의 등장으로 매출정보 뿐 아니라 행동로그 정보를 축적하기 시작함
- ✓ 온라인 모바일 채널은 개인화된 간접 광고 혹은 프로모션이 용이함
- ✓ 채널에 머무는 시간 동안 어떤 Action을 할 것인가를 의사결정하기 위해 A.I 를 활용하게 됨
- ✓ 차에 연료가 필요하듯이 A.I 알고리즘의 연료로 양질의 빅데이터와 정제 기술이 필요함



▶ 빅데이터 개념의 시초

컴퓨터 입장에서는 데이터가 커져도 기술적으로 달라질 것이 없음
하지만, 데이터 사이즈가 기술적, 물리적으로 단일 컴퓨터, 단일 데이터베이스로
처리하는데 무리가 되는 크기의 데이터가 있다면 ... 혹은 비용 이슈가 있다면 ...

예: 구글, 네이버 검색 정보, 유튜브의 보유 동영상, 페이스북 보유 자료 등
대용량 컴퓨터로 처리하는 것이 비용 면에서 불리한 크기의 데이터가 발생함.

대용량 고사양 컴퓨터 한대 vs 저용량 저사양 컴퓨터 군단

- ✓ 이에 대한 해결책으로 ... → 분산 처리 Computing 방법론 탄생
- ✓ 거대한 디스크를 가진 고사양 컴퓨터가 아닌 PC 계열 리눅스 OS 탑재한 컴퓨터를 병렬로 연결
- ✓ 저사양 컴퓨터 군단을 병렬로 제어하는 구글 파일 시스템 탄생
- ✓ 빅데이터를 저장하는 비용이 획기적으로 감소하여 빅데이터를 저장하게 되었음
- ✓ 하지만, 이렇게 저장된 빅데이터를 활용하는 분야는 검색, 집계 등 단순 분야였음



빅데이터 개념

Northwind_Customers - Microsoft Excel

ID	Company	Last Name	First Name	E-mail	Address	City	State/Province	ZIP/Postal Code	Country
1	Company A	Bedeck	Anna		(12)555-0 123 1st Street	Seattle	WA	98000	USA
2	Company B	Grassano	Antonio		(12)555-0 123 2nd Street	Boston	MA	98000	USA
3	Company C	Avian	Thomas		(12)555-0 123 3rd Street	Los Angeles	CA	98000	USA
4	Company D	Lee	Christina		(12)555-0 123 4th Street	New York	NY	98000	USA
5	Company E	O'Donnell	Marlin		(12)555-0 123 5th Street	Minneapolis	MN	98000	USA
6	Company F	Perez-Ota	Francisco		(12)555-0 123 6th Street	Milwaukee	WI	98000	USA
7	Company G	Xie	Ming-Yang		(12)555-0 123 7th Street	Boise	ID	98000	USA
8	Company H	Andersen	Elizabeth		(12)555-0 123 8th Street	Portland	OR	98000	USA
9	Company I	Mortensen	Sven		(12)555-0 123 9th Street	Salt Lake City	UT	98000	USA
10	Company J	Wacker	Rufeld		(12)555-0 123 10th Street	Chicago	IL	98000	USA
11	Company K	Kirschne	Peter		(12)555-0 123 11th Street	Miami	FL	98000	USA
12	Company L	Edwards	John		(12)555-0 123 12th Street	Las Vegas	NV	98000	USA
13	Company M	Ludick	Andre		(12)555-0 456 13th Street	Memphis	TN	98000	USA
14	Company N	Gillo	Carlos		(12)555-0 456 14th Street	Denver	CO	98000	USA
15	Company O	Kapkovits	Helena		(12)555-0 456 15th Street	Honolulu	HI	98000	USA
16	Company P	Goldschm	Daniel		(12)555-0 456 16th Street	San Francisco	CA	98000	USA
17	Company Q	Beguel	Jean Philippe		(12)555-0 456 17th Street	Seattle	WA	98000	USA
18	Company R	Austerlitz	Micla		(12)555-0 456 18th Street	Boston	MA	98000	USA
19	Company S	Eggerer	Alexander		(12)555-0 789 19th Street	Los Angeles	CA	98000	USA
20	Company T	Li	George		(12)555-0 789 20th Street	New York	NY	98000	USA
21	Company U	Thom	Bernard		(12)555-0 789 21th Street	Minneapolis	MN	98000	USA
22	Company V	Ramos	Luciana		(12)555-0 789 22th Street	Milwaukee	WI	98000	USA
23	Company W	Entin	Michael		(12)555-0 789 23th Street	Portland	OR	98000	USA
24	Company X	Hasselberg	Jones		(12)555-0 789 24th Street	Salt Lake City	UT	98000	USA
25	Company Y	Rodman	John		(12)555-0 789 25th Street	Chicago	IL	98000	USA
26	Company Z	Liu	Pun		(12)555-0 789 26th Street	Miami	FL	98000	USA
27	Company AA	Toh	Karen		(12)555-0 789 27th Street	Las Vegas	NV	98000	USA
28	Company BB	Reigher	Amirhash		(12)555-0 789 28th Street	Memphis	TN	98000	USA
29	Company CC	Lee	Soo Jung		(12)555-0 789 29th Street	Denver	CO	98000	USA

정형 Data

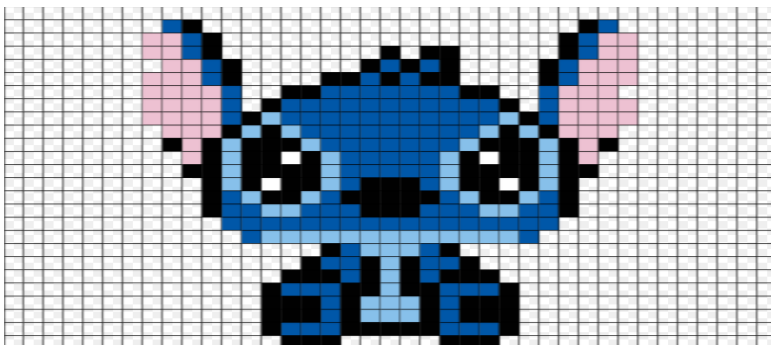


Image Data

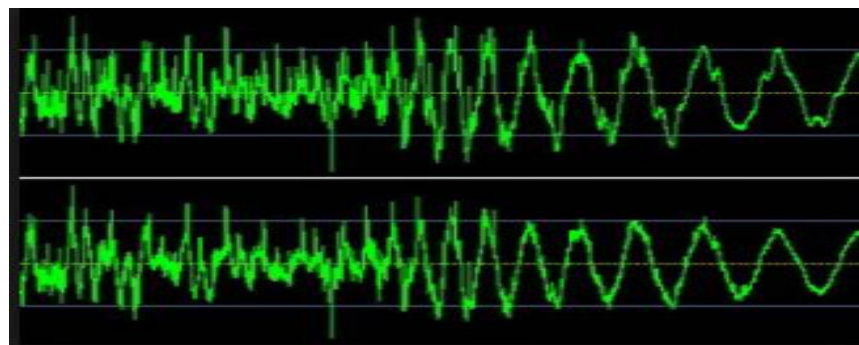
samplelog.log

```

1 #Software: Microsoft Internet Information Services X.X-
2 #Version: X-
3 #Date: 2010-03-24 07:00:01-
4 #Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username
5 2010-03-24 07:00:01 ZZZZC941948879 RUFFLES 222.222.222.222 GET / - 80 - 220.181.7.113 HTTP/1.1
6 2010-03-24 07:00:23 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/12/im_not_mean_im_just_ar
7 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-blank.gif - 80 - 217.2
8 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /grep-options.gif - 80 - 217.2
9 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-cat.gif - 80 - 217.2
10 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-pwd-cd.gif - 80 - 217
11 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /robots.txt - 80 - 95.55.207.95
12 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-short.xml - 80 - 173.45.2
13 2010-03-24 07:00:43 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/08/22-things-you-dont-kno
14 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /screen.css - 80 - 98.88.35.13
15 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/rss-header-red.gif - 80 -
16 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/logo.jpg - 80 - 98.88.35.1
17 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/input-emailsend.jpg - 80 -
18 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /images/cm-ebook-banner.gif - 80 -
19 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg.jpg - 80 - 98.88.35.13
20 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg-top.jpg - 80 - 98.88.35
21 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/checkout-login.gif -
22 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/topnav-contact.jpg - 80 -
23 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/portent-email-sub.gif -
24 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-header.jpg - 80 - 98.88.35

```

Log Data



Sound Data

기업이 수집하는 정보는 크게 속성정보와 로그정보 두 가지로 나눌 수 있음

속성 정보는 개체의 속성에 해당하는 데이터로 시간에 따라 바뀌지 않는 static 데이터를 말함

예: 성별, 연령, 지역, 제조사, 생산일 등

RDB 형태로 저장하는 것이 유리함

Event Log Data는 개체의 상태에 해당하는 데이터로 시간에 따라 바뀌는 데이터를 말함

예: 현 위치, 조회 키워드, 클릭 페이지 등

일반적으로 행은 관측단위, 열은 변수로 지정되는 정형 데이터의 구조에서는

Event Log 데이터를 처리하기 어려움

하지만, 현재 Event Log를 이용한 데이터 처리 수요가 증가하고 있음

Event Log는 하둡 시스템을 저장하는 것이 유리함

예: 구글, 페이스북 보유 텍스트, 센서 데이터 등

- 쿠팡이 보유한 정보 예

거래정보: 누가, 언제, 무엇을 구매했는지에 관한 정보

상품정보: 바코드, 상품분류, 상품명, 제조사, Seller 정보, 가격 등

고객정보: 배송, 과금을 위한 정보

Web, App 사용정보: 접속일자, 링크 클릭 정보, 검색 정보 등

각 정보를 RDB, Hadoop 중 어느 곳에 저장하는 것이 효율적일까요?



데이터수집

저장

집계

지능화

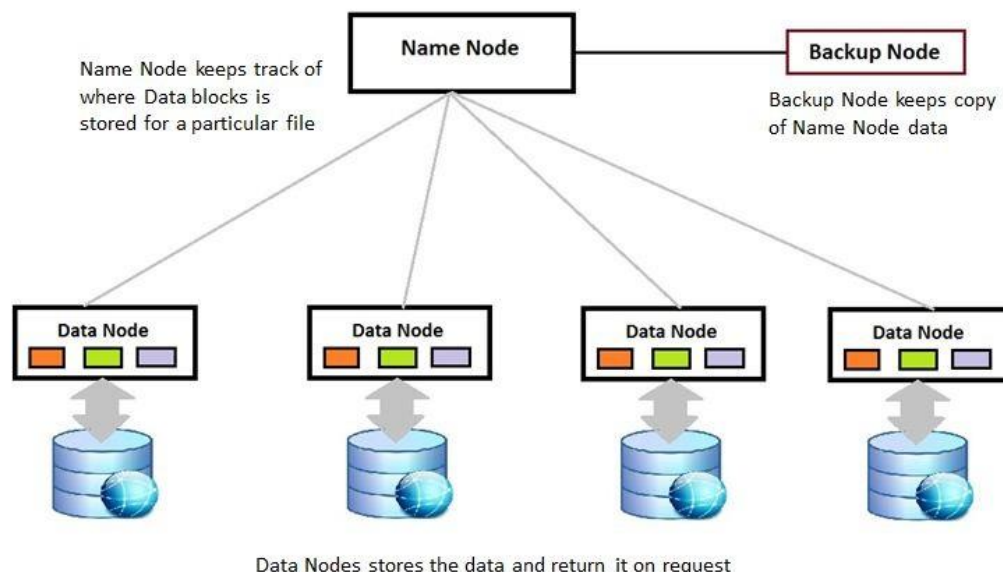
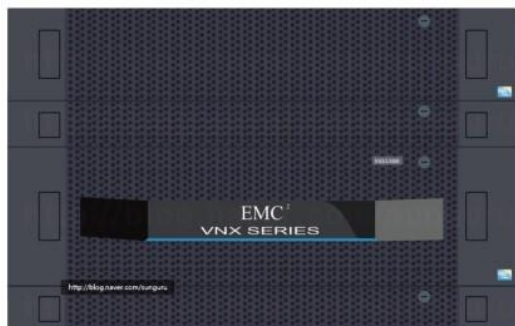


데이터수집

저장

집계

지능화



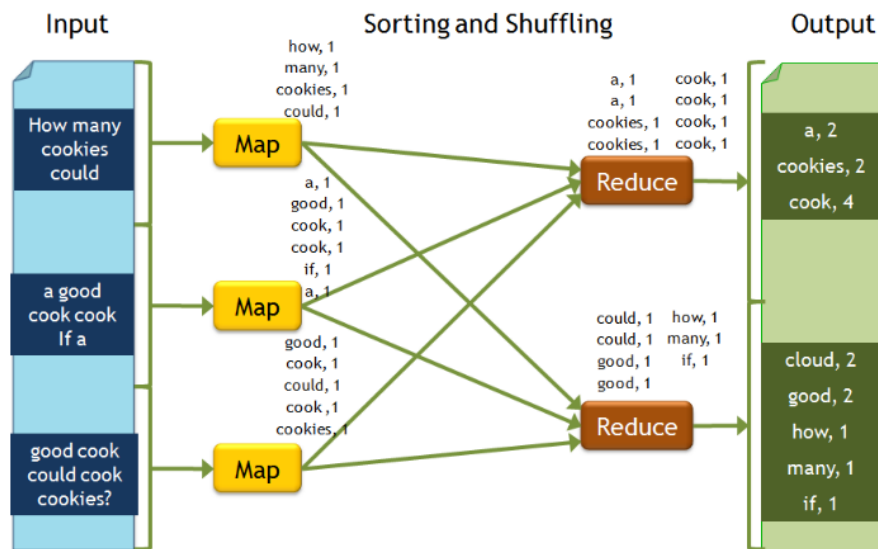
Data Nodes stores the data and return it on request

데이터수집

저장

집계

지능화



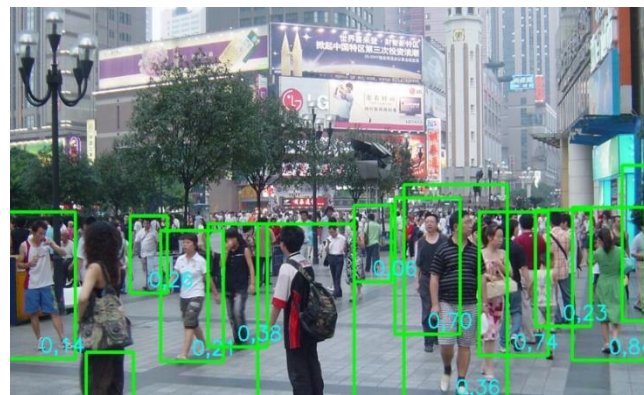
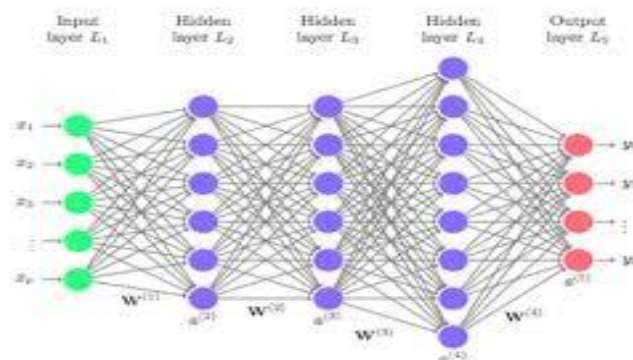
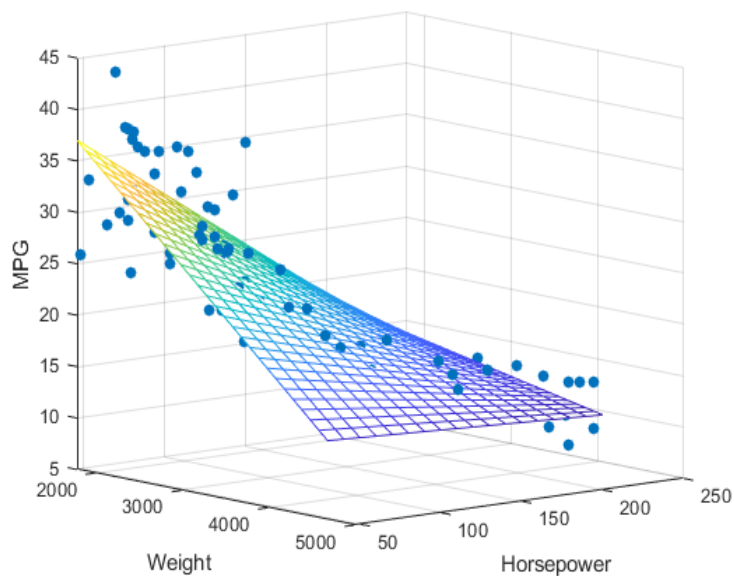
By Manaranjan Pradhan

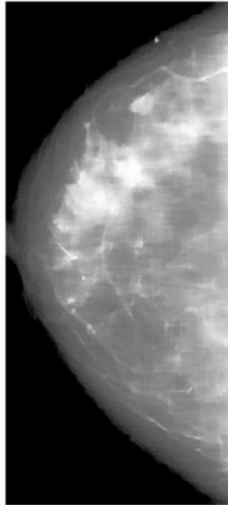
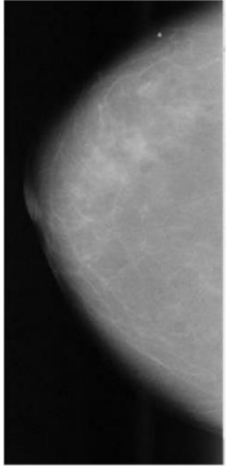
데이터수집

저장

집계

지능화





id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	points_mean	symmetry_mean
1	87139402 B	12.320	12.39	78.85	464.1	0.10280	0.06981	0.039870	0.037000	0.1959
2	8910251 B	10.600	18.95	69.28	346.4	0.09688	0.11470	0.063870	0.026420	0.1922
3	905520 B	11.040	16.83	70.92	373.2	0.10770	0.07804	0.030460	0.024800	0.1714
4	868871 B	11.280	13.39	73.00	384.8	0.11640	0.1360	0.046350	0.047960	0.1771
5	9012568 B	15.190	13.21	97.65	711.8	0.07963	0.14	0.033930	0.026570	0.1721
6	906539 B	11.570	19.04	74.20	409.7	0.08546		0.054850	0.014280	0.2031
7	925291 B	11.510	23.93	74.52	403.5	0.09261		0.111200	0.041050	0.1388
8	87880 M	11.010	23.75	91.56	597.8	0.13230		0.155800	0.091760	0.2251
9	862989 B	11.010	19.29	67.41	336.1					0.2217
10	89827 B	11.010	19.29	67.41	336.1					0.1776
11	91485 M	11.010	19.29	67.41	336.1					0.1848
12	8711003 B	11.010	19.29	67.41	336.1					0.1970
13	9113455 B	11.010	19.29	67.41	336.1					0.1562
14	857810 B	13.030	19.31	82.61	527.2	0.08060	0.03783	0.000692	0.004167	0.1819
15	9111805 M	19.590	25.00	127.70	1191.0	0.10320	0.09871	0.165500	0.090630	0.1663
16	925277 B	14.590	22.68	96.39	657.1	0.08473	0.13300	0.102900	0.037360	0.1454
17	867387 B	15.710	13.93	102.00	761.7	0.09462	0.09462	0.071350	0.059330	0.1816
18	89511502 B	12.670	17.30	81.25	489.9	0.10280	0.07664	0.031930	0.021070	0.1707
19	89263202 M	20.090	23.86	134.70	1247.0	0.10800	0.18380	0.228300	0.128000	0.2249
20	866714 B	12.190	13.29	79.08	455.8	0.10660	0.09509	0.028550	0.028820	0.1880
21	874373 B	11.710	17.19	74.68	420.3	0.09774	0.06141	0.038090	0.032390	0.1516

암 존재 여부

조직의 반지름, 면적, 평평도, 굴곡, 대칭성 등

Showing 1 to 22 of 569 entries

과거에는 우측과 같이 이미지 정보를 요약해서 정형 데이터로 전환시켜 분석에 사용함
 현재에는 우측과 같은 형태로도 사용하지만, 이미지 pixel 값을 데이터로 보고 그대로 분석에 사용하는 기법이 더 좋은 예측력을 보이고 있음
 분석에 정량 데이터 뿐 아니라 신호 데이터까지 처리함

표 1-7 빅데이터 자동 수집 방법 [07]

방법	설명
로그 수집기	내부에 있는 웹 서버의 로그를 수집. 즉, 웹 로그, 트랜잭션 로그, 클릭 로그, DB의 로그 데이터 등 수집
크롤링	주로 웹 로봇으로 거미줄처럼 얽혀 있는 인터넷 링크를 따라다니며 방문한 웹 사이트의 웹 페이지라든가 소셜 데이터 등 인터넷에 공개되어 있는 데이터 수집
센싱	각종 센서로 데이터 수집
RSS 리더/오픈 API	데이터의 생산·공유·참여 환경인 웹 2.0을 구현하는 기술로 필요한 데이터를 프로그래밍으로 수집
ETL Extraction, Transformation, and Loading	데이터의 추출, 변환, 적재의 약자로, 다양한 소스 데이터를 취합해 데이터를 추출하고 하나의 공통된 형식으로 변환하여 데이터웨어하우스에 적재하는 과정 지원

실시간 데이터 처리가 필요한 응용분야의 등장

과거 데이터를 축적하고 이를 분석해 지식을 얻는 방법도 있지만,
실시간 처리가 중요한 일도 많다.

예: 부정고객 탐지, 해킹 등 악의적 접속 시도 탐지, 운전자 비상 상황 감지 등

사물인터넷 개념과 결합되면서 자동으로 행동 데이터 수집 / 분석 / 의사결정

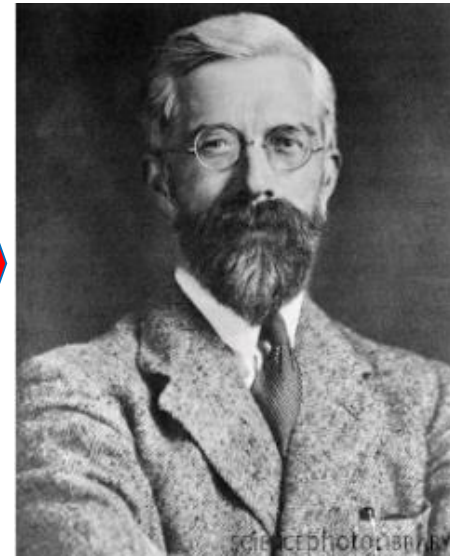
컴퓨터에 제한적 정보에 의한 의사결정의 단점을 보완하고자

빅데이터를 탑재한 중앙서버와의 통신을 통해 의사 결정하는 방식으로 진화

고객에 대한 이해 → 행동 Event 관계를 이해하는 방향으로 발전함

즉, 고객의 속성 정보 모형화에서 로그 등 행동정보의 관계를 모형화 하는 방향으로 발전

통계학: 표본으로 부터 모집단의 정보를 추정하거나 모집단의 상태를 추측하는 과학



상관/회귀 기본개념
제시

상관/회귀분석 완성
중심극한정리

표본이론 정립
현대 추측 통계학 정립

통계학 관점에서 빅데이터가 필요한가?

간단한 문제: 인구 중에 키 185cm 이상인 사람의 비율은?

솔루션 1: 1,000명의 키를 검사하여 185cm 이상인 사람의 비율로 추정

솔루션 2: 185cm 이상인 사람이 나올 때까지 검사하여 비율은 $1/n$ 으로 추정

솔루션 3: 30명의 키를 측정하여 평균과 표준편차를 구해 정규분포 가정 하에 계산함

가장 정확한
방법은?



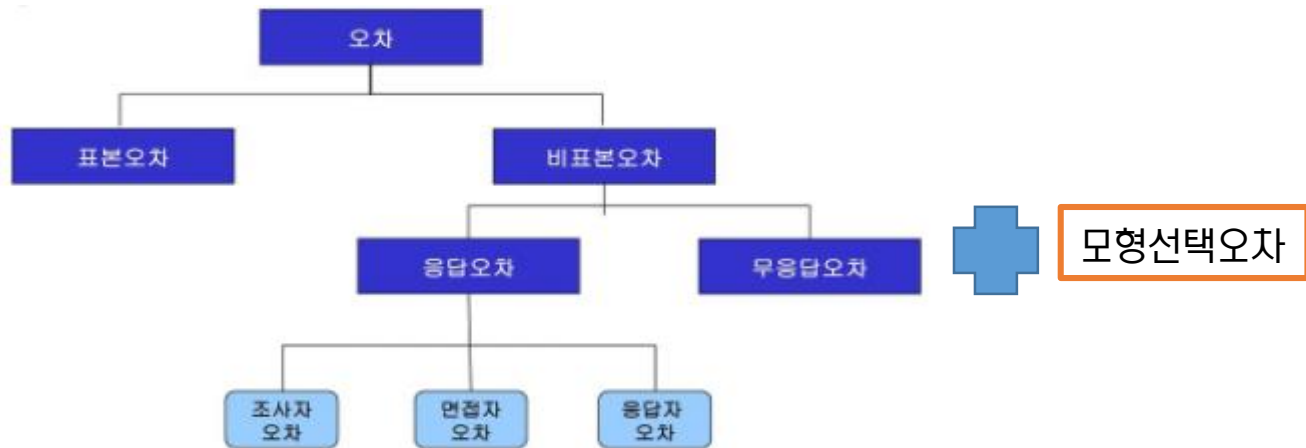
가장 효율적인
방법은?

기본적으로 표본 수를 최소화하고자 함(표본 수 증가는 비용의 증가)

표본오차는 표본 수가 증가하면 줄어 들지만, 비표본오차는 과학적으로 접근이 어려워 계산이 불가능함

비표본 오차에 의해 참값과 통계적 추정값이 다른 결과를 줄 가능성이 존재함

비표본 오차 중에 모형선택오차도 있는데 통계학에서는 모형을 가정하는 경우가 많음(가정이 틀렸다면?)



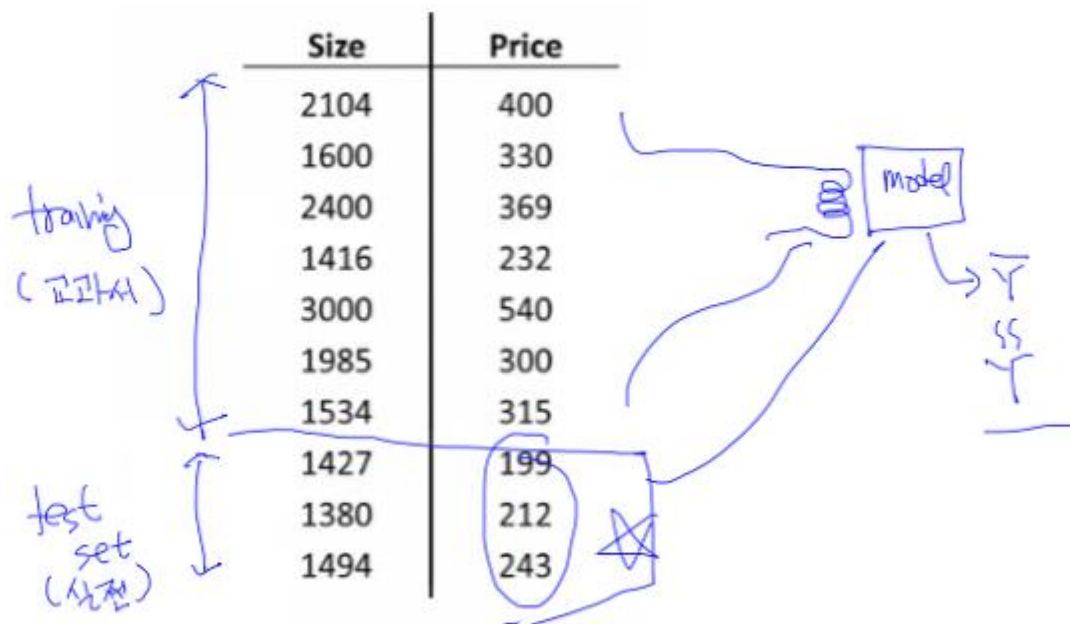
Small Data의 경우, 분석 정확도를 위해 모든 정보를 사용해 추정 혹은 의사결정 수행함

분포나 모형 가정을 통해 재현성을 검정하고자 함

재현성: 다른 표본에서도 우리의 추정이나 의사결정이 재현됨을 수학적으로 증명

Big Data의 경우는 풍부한 데이터 셋을 가지고 있기 때문에 재현성을 수학적으로 증명 하지 않고 데이터를 이용해 실증적으로 재현함

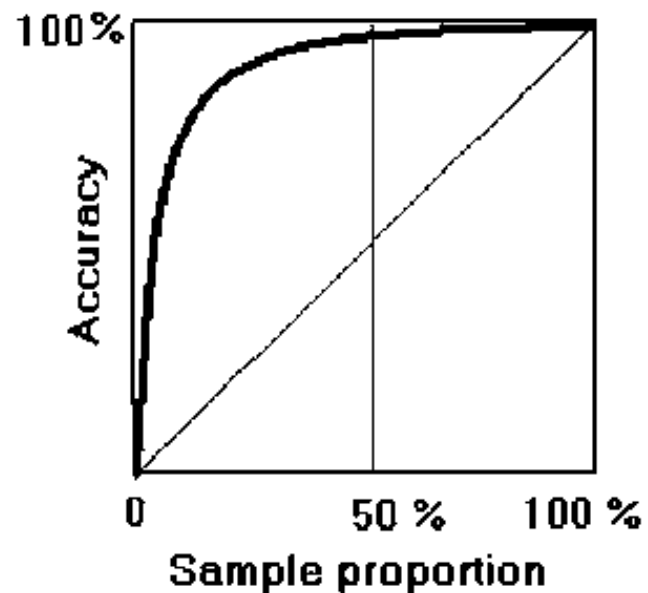
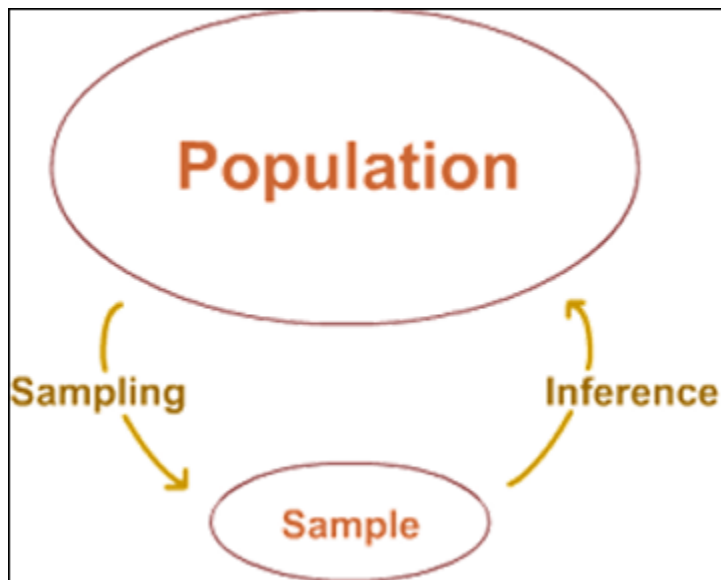
Training and test sets



	Size	Price
training (교과서)	2104	400
	1600	330
	2400	369
	1416	232
	3000	540
	1985	300
	1534	315
test set (시험)	1427	199
	1380	212
	1494	243

http://www.holehouse.org/mlclass/10_Advice_for_applying_machine_learning.html

재현성: 모델을 만드는데 모든 데이터를 사용하지 않고 모델 검증을 위한 Test-Set을 남겨 둘 수 있는 여유가 존재한다.



빅데이터가 존재해도 추정 혹은 의사결정의 Accuracy는 크게 증가하지 않는다.

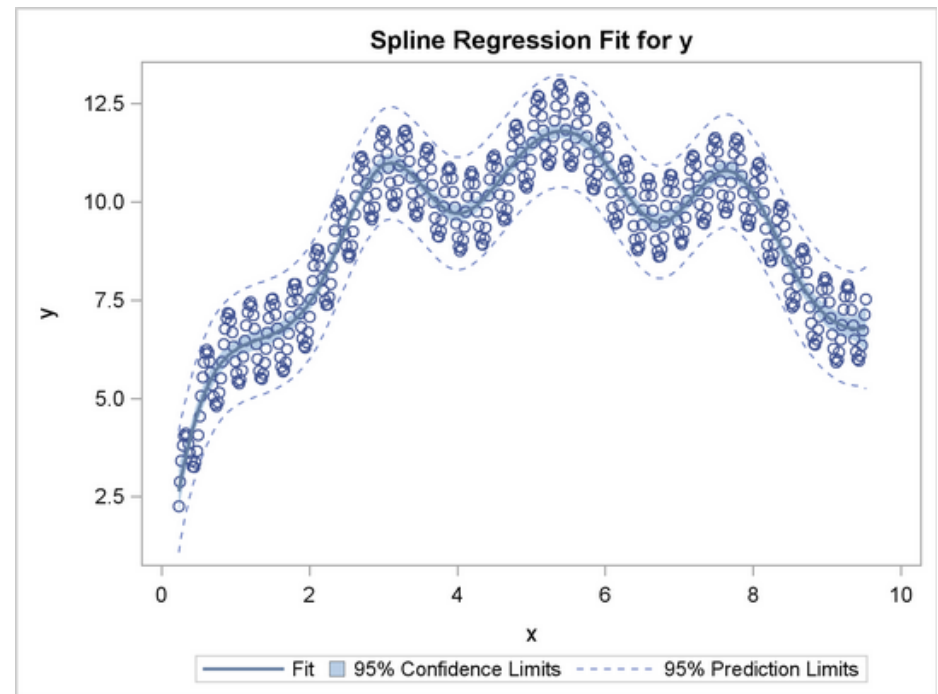
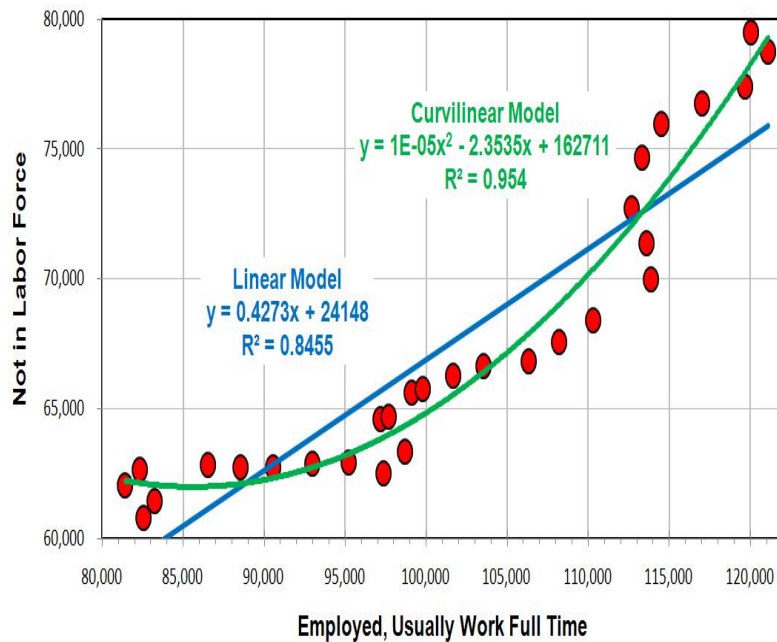
1900년대의 데이터 분석과 2000년대의 데이터 분석은 달라졌음

과거: 자신의 연구가설을 증명(주로 학문적 요구)

현재: Prediction/Forecasting, Decision Making 등 다양한 Biz. 요구 등장

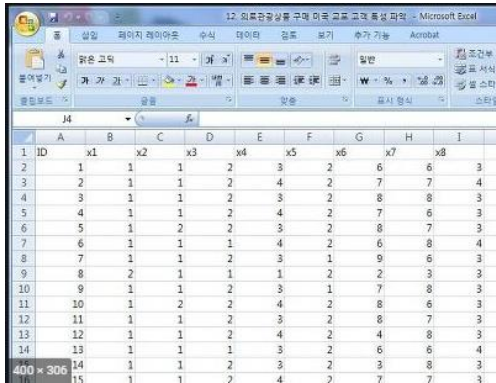
	데이터	관점	모형	모수절약
전통적 통계학	분석을 위해 필요 최소 데이터를 수집함	모집단의 구조파악 을 통한 추론 (보수적 관점)	Linear 중심	설명변수의 수를 최소화하여 분석의 자유도를 확보 하고자 함
빅데이터 접근	기존에 축적된 자료 를 통해 분석함 (자료수집비용 無)	모집단의 구조보다 예측에 초점 (적극적 관점)	Non- Linear 로 확장	데이터가 많으므로 변수의 수에 구애 받지 않음

Linear Regression and Non-Linear Regression Model

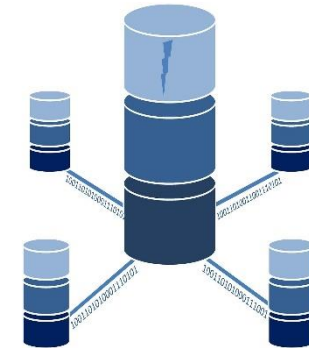


스몰 데이터에서는 Non-Linear 모델을 사용할 수 있으나 Over-fitting 위험이 커진다.

빅데이터에서는 Non-Linear 모델을 사용해도 Over-fitting 위험이 덜하다.



ID	x1	x2	x3	x4	x5	x6	x7	x8
1	1	1	1	2	3	2	6	6
2	2	1	1	2	4	2	7	7
3	3	1	1	2	3	2	8	8
4	4	1	1	2	4	2	7	6
5	5	1	2	2	3	2	8	7
6	6	1	1	1	4	2	6	8
7	7	1	1	2	3	1	9	6
8	8	2	1	1	1	2	2	3
9	9	1	1	2	3	1	7	8
10	10	1	2	2	4	2	8	6
11	11	1	1	2	3	2	8	7
12	12	1	1	2	4	2	4	8
13	13	1	1	1	3	2	6	6
14	14	1	1	2	3	2	3	8



정량화
선형 모형
모형 구조 규명



이벤트 데이터
비선형 모형
추측/예측 자동화

감사합니다

인하대학교
데이터사이언스 학과
김 승 환

swkim4610@inha.ac.kr