

# 1장 현대사회와 데이터 사이언스

## 1.1 데이터 사이언스의 역할

- 빅 데이터 시대와 데이터 사이언스
- 자원으로서의 데이터
- 데이터 사이언티스트

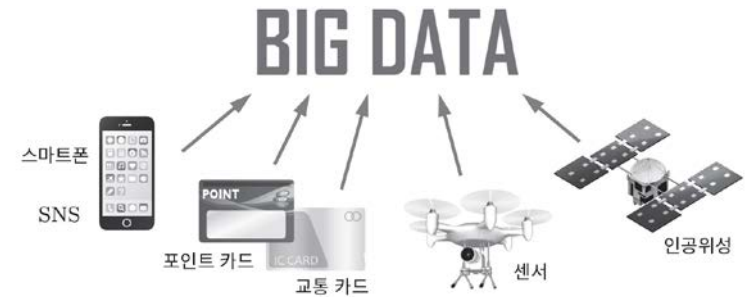
## 1.2 데이터분석을 위한 데이터의 수집과 관리

- 데이터사이언스 프로세스
- 데이터 용량
- 대규모 데이터의 이용
- 데이터의 수집방법
- 데이터의 전처리

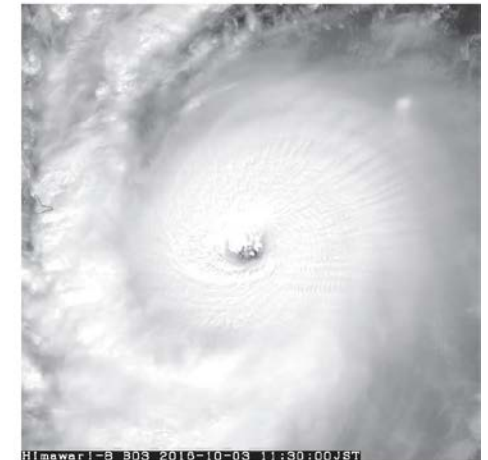
# ● 빅 데이터 시대와 데이터 사이언스

## ● 빅데이터: 다양한 종류의 대량의 데이터

- 스마트폰
  - 개인소유비율이 70%을 넘어설 정도로 보급
  - 30년 전의 슈퍼컴퓨터의 능력에 필적
- 무선통신
  - 지하철에서 스마트폰 사용
- 포인트카드/교통카드
  - 개인의 구매이력 정보를 수집
  - 이동 정보 축적
- 이력
  - SNS 메시지 송수신, 웹 검색, 구매 이력, 행동 이력
  - 관심사, 상품, 서비스 트렌드 분석
- 과학분야
  - 인공위성에서의 기상 관측
- 인공위성을 이용한 GPS(Global Positioning System)
  - 자동차 내비게이션
  - 스마트폰의 위치정보



〈그림 1.1〉 빅데이터의 개념도



〈그림 1.2〉 기상위성 히마와리8호가 촬영한 2016년 태풍18호(일본 기상청 홈페이지 제공)

# ● **자원으로서의 데이터: 21세기의 원유**

---

- 빅데이터: V3 (volume, variety, velocity)의 특성
- 데이터는 "21세기의 원유": 새로운 경제 자원, 데이터 보유가 중요
- 인터넷 관련 거대기업(GAFA: 4대 거대기업, 막대한 데이터 축적)
  - Google
  - Apple
  - Facebook
  - Amazon
- 중국(BAT:3대 기업)
  - 바이두
  - 알리바바
  - 텐센트
- 네트워크 효과: 사용자가 늘어날 수록 편리

# 자원으로서의 데이터: 플랫폼, 인터넷

---

- **플랫폼(Platformer)**

제3자가 비즈니스 또는 정보발신 등을 수행하는 기반(Platform)으로서 이용할 수 있는 서비스 또는 시스템 등을 제공하는 사업자를 뜻함

(ex: Facebook – SNS 플랫폼)

- **인터넷: 플랫폼이 활약하는 기반이 됨**

- 인터넷 자체는 분산형 구조를 갖는다
- 단, 그 기반위에 구축된 서비스에 독점적 경향이 발생하고 있음

# 자원으로서의 데이터: 정보보안

---

- **Facebook 개인정보 유출**

- 2018/4 뉴스보도: 최대 8,700만명의 개인정보 유출되었다
- 유출된 데이터는 Cambridge Analytica라는 데이터분석회사에 전달
- 2016년 미국 대통령 선거에 트럼프 후보진영에 유리하게 사용되었다는 의혹
- Facebook의 개인정보 취급에 대한 비판이 일어남

- **데이터는 21세기의 가장 중요한 자원**

- **데이터를 올바르게 취급하지 않으면 그 영향이 더욱 커지게 됨**

# 자원으로서의 데이터: 데이터 가공/분석기술

---

- 천연자원

- 가공기술 없이 수출하는 것만으로는 선진국으로 도약할 수 없음

- 데이터

- 수집하여 저장해 놓는 것만으로는 가치를 발생시킬 수 없음
- 데이터를 처리/분석하는 기술이 없으면 외국기업이 취득하고 활용
- 데이터를 가공하고 분석하는 기술과 인재가 필요

# 자원으로서의 데이터: 데이터사이언티스트

---

## • 교육의 필요성

- 리터러시(literacy): 문자화된 기록물을 통해 지식/정보를 획득하고 이해할 수 있는 능력
- 데이터 리터러시 (data literacy): 데이터를 목적에 맞게 활용하는 데이터 해석 능력
- 데이터 리터러시 향상이 중요함.
- 인문사회/이공계를 불문하고 모든 분야에서 수리/데이터사이언스 교육
- 데이터사이언스에 전문성이 가진 인재를 조직적으로 육성

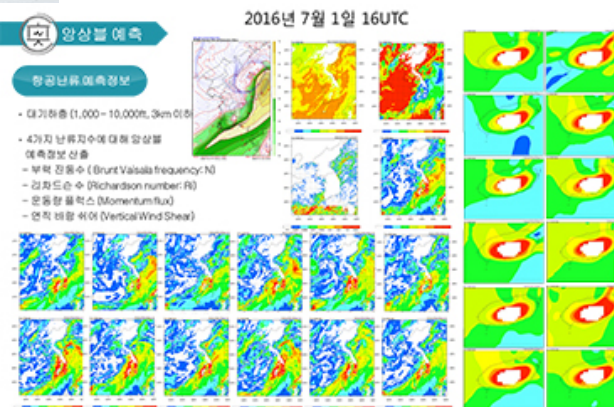
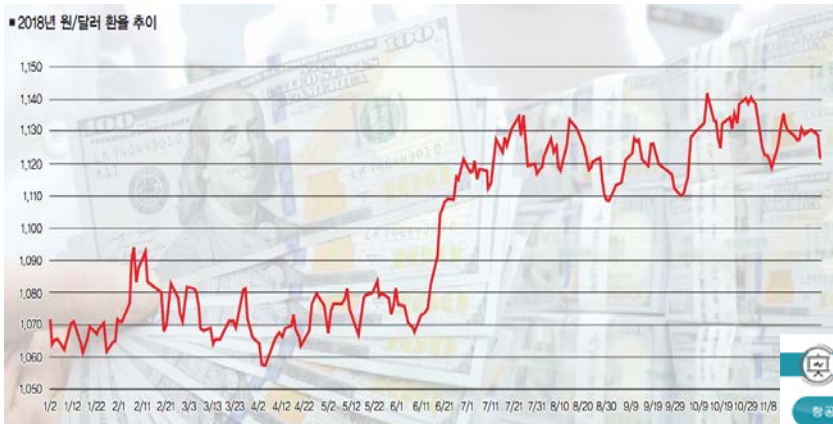
## • 데이터사이언티스트 (Data scientist)

- 데이터를 처리/분석하여 데이터로부터 가치를 만들어 낼 수 있는 전문적인 인재
- 데이터사이언티스트에게 필요한 소양
  - 정보학 또는 컴퓨터과학 분야의 지식
  - 통계학, 기계학습
  - 수학

# 자원으로서의 데이터: 환율 vs 기상 문제

## • 데이터의 관점(환율 vs 기상 예측 문제)

- 환율 예측 문제: 경제 분야
- 기상 예측 문제: 이공계 분야
- 시계열 데이터(time series data)라는 점에서 공통점이 있음
- 데이터를 분석하는 경우에 공통적인 기법을 사용할 수 있음





# 자원으로서의 데이터:

---

## • 인문사회 vs 이공계

- 전공구별은 교육분야에서 문제점으로 부각
- 인문사회 전공 학생들은 이공계 분야 과목, 특히 수학을 기피
- 인문사회 전공의 경영자
  - 숫자에 약함
  - 데이터를 기반으로 의사결정을 하기 보다는 경험이나 감에 의존
- 엔지니어
  - 경력이 기술분야에 한정
  - 기술적인 전문성은 높지만 경영적인 판단을 내리기 어려움
- 기술측면을 알고 있는 경영자, 경영분야를 잘 알고 있는 엔지니어가 요구됨

## • 증거기반 정책수립(EBPM: Evidence Based Policy Making)

- 정부와 지방자치단체들에서 데이터를 토대로 정책입안/평가

# 데이터 사이언스: 종적 지식 vs 횡적 지식

---

## • 종적 지식

- 인문사회계열과 이공계를 구별하는 것은 종적 사회구조를 나타냄
- 대학교의 학부 및 학과의 구성은 대응하는 산업분야에 대한 인재공급을 고려한 형태
  - 법학부 졸업 → 공무원
  - 경제학부 졸업 → 금융기관
  - 공과대학: 학과구성이 각 제조업 분야에 대응

## • 데이터사이언스 (횡적 지식)

- 분야를 불문하고 모든 영역에서 필요
- 범용적이면서 다양한 분야를 횡단하는 융합 기술

# 데이터사이언스란?

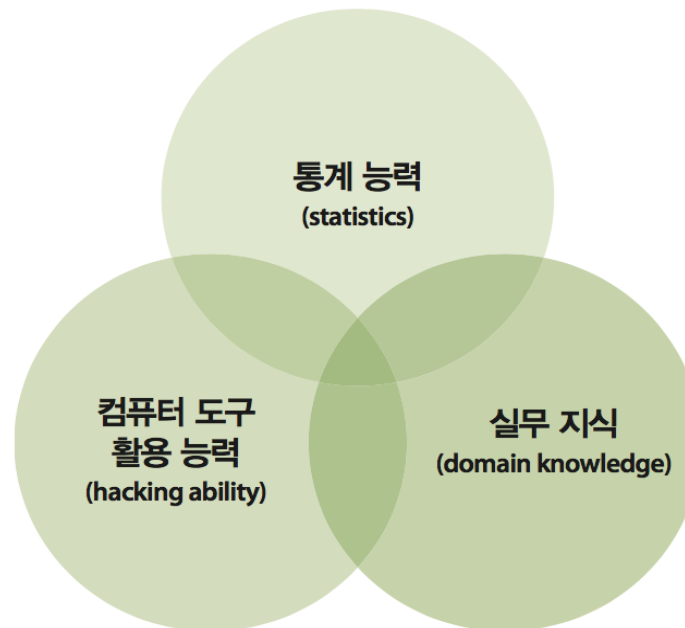
"컴퓨터 도구를 효율적으로 이용하고, 적절한 통계학 방법을 사용하여 실제적인 문제에 답을 내리는 활동"

- 주택 가격을 예측하는 방법은?
- 초등생 자녀의 수학 능력과 상관 관계가 높은 변수는 무엇일까?
- 훌륭한 직원을 뽑는 인터뷰 방법은 무엇일까?
- 괴혈병의 치료법은 무엇일까?
- 산욕열의 원인은 무엇일까?
- 웹사이트를 개선하는 방법은?
- 신약이 혈압을 낮추는 데 효과가 있을까?
- TV 광고가 제품 판매에 얼마만큼의 영향을 줄까?
- 온라인 광고에서 클릭 여부를 예측하는 방법은?
- 집 안에 있는 수영장과 권총 중 어느 것이 어린이에게 더 위험할까?
- 대학 입학에 남녀의 성차별이 있을까?
- 비싼 와인이 더 맛있을까?
- 아버지의 키가 180cm라면 아들의 키는 얼마일까?
- 대학 진학을 할 때 전공이 중요할까, 학교가 중요할까?
- 흡연은 몸에 해로울까?
- 투자신탁과 ETF 인덱스 펀드 중 어느 곳에 투자하는 것이 좋을까?

# 데이터 사이언티스트: 갖추어야 할 능력

---

- 빅데이터라는 용어 2010년부터 사용됨
- How Google works (Grand Central Publishing, 2014)
  - "앞으로 10년간 가장 매력적인 직업은 통계 전문가라고 계속 이야기하고 있다." – 구글의 최고 경제학자 Hal Varian
  - "데이터는 21세기의 검은색이며, 이 검은색을 잘 다룰 수 있는 자가 사무라이이다" – 구글의 수석 부회장 Jonathan Rosenberg,



# 데이터 사이언티스트 육성 현황:

---

- 미국 통계학 및 생물통계학 분야의 학위 수여 건수
  - 학사 학위
    - 2008년: 600명
    - 2015년: 2500명
  - 석사 학위
    - 2008년: 1600명
    - 2015년: 3500명
- 중국
  - 300개 이상의 대학에 통계학부 또는 통계학과가 설치
  - 중국의 IT화는 급속도로 발전
  - BAT와 같은 거대 인터넷 기업들이 많은 데이터사이언티스트를 채용
- 일본
  - 다수의 기업에서 데이터 사이언스 담당부서를 신설
  - 데이터사이언티스트 수요 급증

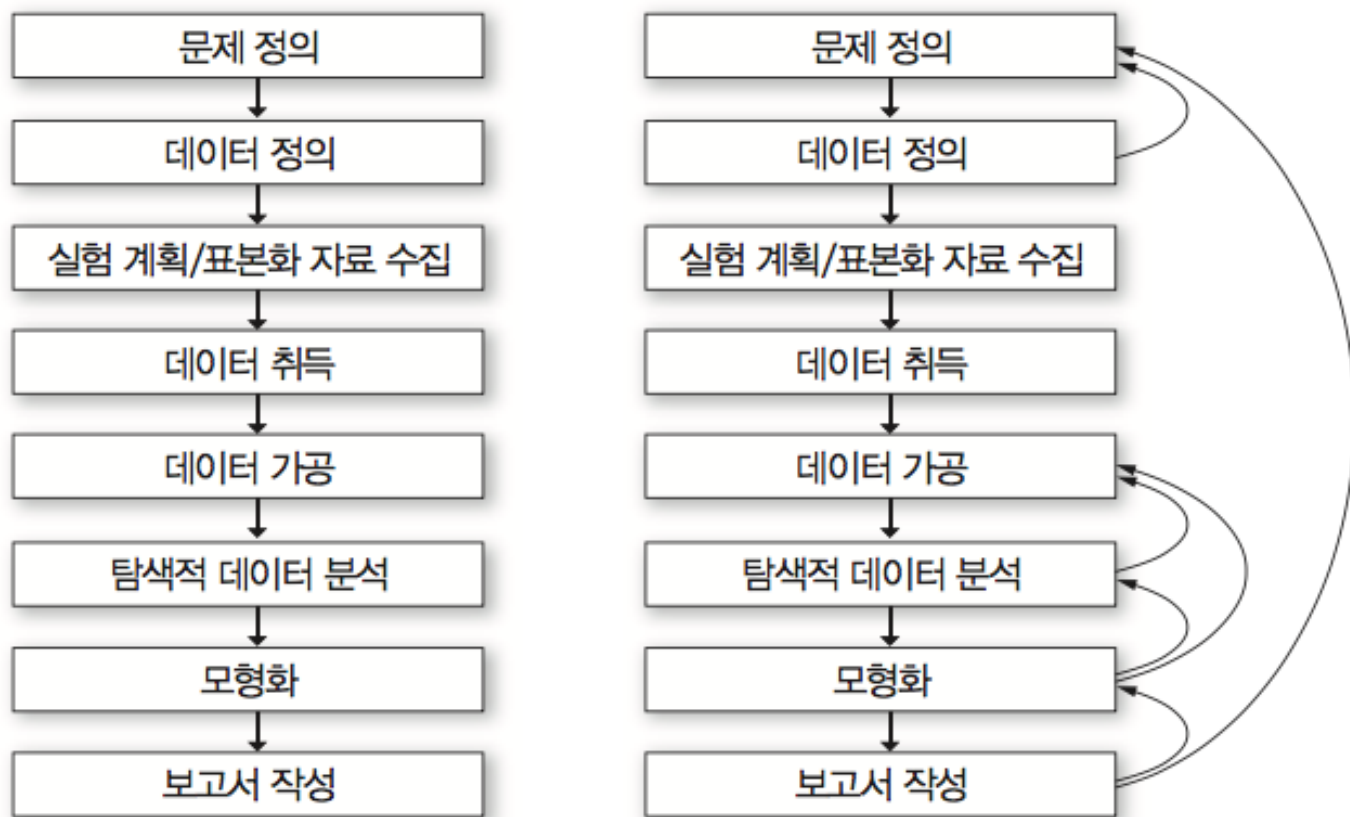
## 1.2 데이터분석을 위한 데이터의 수집과 관리

- 데이터사이언스 프로세스
- 데이터 용량
- 대규모 데이터의 이용
- 데이터의 수집방법
- 데이터의 전처리

# 데이터사이언스 프로세스

---

1. 문제 정의(problem definition)
2. 데이터 정의(data definition)
3. 실험 계획(design of experiment)
4. 데이터 취득(data acquisition)
5. 데이터 가공(data processing, data wrangling)
6. 탐색적 분석과 데이터 시각화(exploratory data analysis, data visualization)
7. 모형화(modeling)
8. 분석 결과 정리(reporting)



데이터 분석 과정에 대한 이상적 관점(왼쪽)과 현실적 관점(오른쪽)



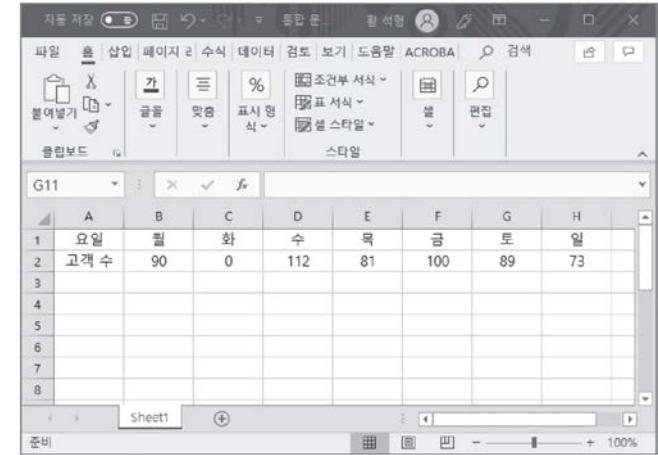
# 데이터의 유형: 디지털데이터

- 수학

- 벡터
- 행렬

- Excel

- 리스트: 동일한 형식을 갖는 데이터를 모아놓은 것
- 표: 리스트를 모아놓은 것



The screenshot shows the Microsoft Excel interface. The active cell is G11. The spreadsheet contains data in the following table:

	A	B	C	D	E	F	G	H
1	요일	월	화	수	목	금	토	일
2	고객 수	90	0	112	81	100	89	73
3								
4								
5								
6								
7								
8								

〈그림 1.4〉 Excel 표

- 프로그래밍(R, Python)

- 배열
- 데이터 프레임

In [1]: C02

Plant	Type	Treatment	conc	uptake
Qn1	Quebec	nonchilled	95	16.0
Qn1	Quebec	nonchilled	175	30.4
Qn1	Quebec	nonchilled	250	34.8

〈그림 1.5〉 데이터 프레임(이산화탄소 연간배출량)

# 데이터 용량: 단위

---

- 비트와 바이트

- 비트(bit)
- 바이트(byte, B)
  - 1바이트 = 8비트
- 한글, 이미지, 음성 등과 같은 데이터도 바이트 또는 비트를 사용하여 용량을 나타냄

- 용량 단위 비교

- 1 Kilo Byte (KB) =  $10^3$  byte
- 1 Mega Byte (MB) =  $10^6$  B
- 1 Giga Byte (GB) =  $10^9$  B
- 1 Tera Byte (TB) =  $10^{12}$  B
- 1 Peta Byte (PB) =  $10^{15}$  B
- 1 Exa Byte (EB) =  $10^{18}$  B
- 1 Zetta Byte (ZB) =  $10^{21}$  B
- 1 Yotta Byte (YB) =  $10^{24}$  B

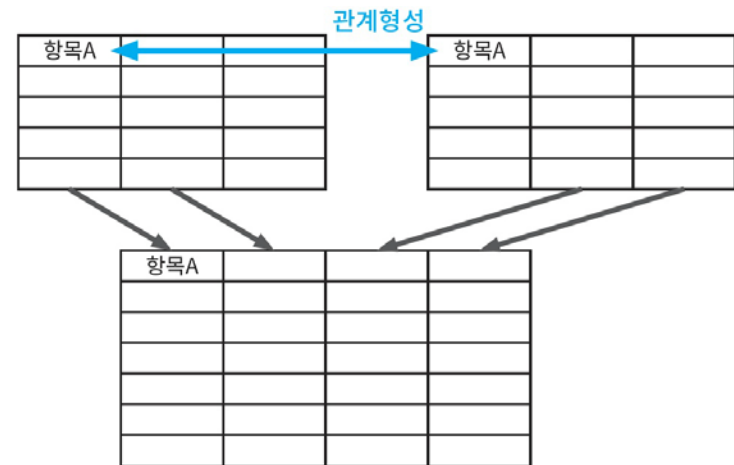
# 데이터 용량: 실제 데이터 크기

---

- 1메가(100만) 바이트: 사진 1장
  - 1기가(10억) 바이트: 영화, 동영상 파일
  - 1테라(1조) 바이트: 개인용 컴퓨터의 하드디스크 용량
  - 1페타 바이트 이상: 빅데이터급
  - 10엑사 바이트: 구글의 데이터 전체용량(2013년)
  - 1제타 바이트: 전세계 데이터 총용량(2018년)
- 
- 빅데이터의 실제 용량을 계측하는 것은 어려움
  - 위에서 설명한 용량은 대략적인 크기를 나타냄에 주의

# 대규모 데이터의 이용: 관계형 데이터베이스

- 관계형 데이터베이스(Relational Database, RDB)
  - 데이터구조를 표 형식으로 다룸
  - 여러 개의 표 사이에서 관계 있는 요소들의 결합이나 참조 수행
  - 키(key)
  - SQL: 표준적인 데이터 질의언어
    - 데이터 조작이나 정의를 수행
    - R, Python 같은 프로그래밍 언어에서도 이용할 수 있음



〈그림 1.6〉 여러 개의 표들로부터 새로운 표를 작성한다.

"거의 모든 데이터 과학자는 언젠가는 SQL을 사용하게 된다. 많은 회사들이 데이터를 SQL을 사용하는 RDBMS에 저장하기 때문이다. 워낙 많은 분석가가 SQL에 익숙하므로 페이스북 등 에 쓰이는 빅데이터를 위한 분산시스템인 하둡의 파일시스템에 저장된 데이터도 SQL 문법을 사용하여 처리하고 추출할 수 있는 하이브(Hive, <https://hive.apache.org/>)가 사용된다 [Apache Hive (2016)]."

표 3-1 R dplyr 문법과 SQL 문법 비교

데이터 처리 작업	R	SQL
1. 행 선택, filter	df %>% filter(x>0)	SELECT * FROM df <b>WHERE</b> x > 0
2. 정렬, arrange	df %>% arrange(x)	SELECT * FROM df <b>ORDER BY</b> x
3. 변수 선택	df %>% select(x)	SELECT x FROM df
4. 변수 변환	df %>% mutate(y=f(x))	SELECT <b>f(x)</b> AS y FROM df
5. 요약 통계량 계산	df %>% summarize(avg_x=mean(x))	SELECT avg(x) AS avg_x FROM df
6. 랜덤 샘플링	df %>% sample_n(100) df %>% sample_frac(0.1)	없음*
7. 유일값 계산	df %>% select(x) %>% distinct()	SELECT <b>DISTINCT</b> (x) FROM df
8. 그룹핑	df %>% group_by(x) %>% summarize(total=n())	SELECT x, count(*) AS total FROM df <b>GROUP BY</b> x
9. 이너 조인(inner join)	inner_join(x, y, by="a")	SELECT * FROM x JOIN y ON x.a = y.a
10. 레프트 조인(left join)	left_join(x, y, by="a")	SELECT * FROM x LEFT JOIN y ON x.a = y.a
11. 풀 조인(full join)	full_join(x, y, by="a")	SELECT * FROM x FULL JOIN y ON x.a = y.a
12. 합집합(union)	union(x, y) union_all(x, y)	SELECT * FROM x UNION SELECT * FROM y

\* Hive SQL에는 제공된다(<https://goo.gl/q2mISm> 참고).

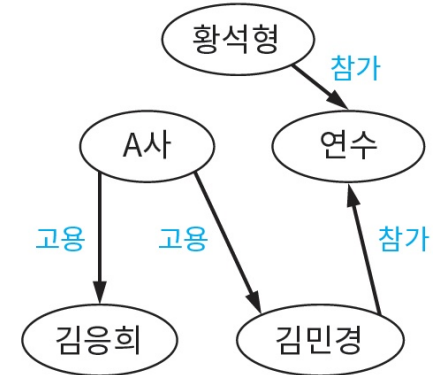
# 대규모 데이터의 이용: 클라우드, NoSQL

- 클라우드

- 저렴하면서도 간편하게 데이터베이스 관리시스템을 이용

- NoSQL (Not Only SQL)

- 빅데이터를 다루기 위해 사용
- 그래프형 데이터베이스
  - NoSQL의 한 가지 형태
  - 데이터들 사이의 관계를 직감적으로 이해
- Hadoop/Spark
  - HDFS(Hadoop Distributed File System)를 사용
  - 파일을 분할하여 여러 대의 컴퓨터에서 관리
  - 페타 바이트급의 데이터를 처리



〈그림 1.7〉 그래프형 데이터베이스

# 데이터의 수집방법: 인터넷에서 이용할 수 있는 데이터

---

- **경진대회용 데이터**

- Kaggle
- SIGNATE

- **금융 데이터:**

- 브라우저를 이용하여 직접 데이터를 복사하여 사용할 수 있음
- 웹페이지를 표현하는 html은 Excel의 표와 유사한 구조를 표현할 수 있음
- Excel 파일이나 csv 파일을 다운로드할 수 있는 사이트도 있음

- **구글 맵**

- API 제공

- **웹 크롤링**

- 여러 개의 웹 사이트들로부터 html의 구조를 갖는 데이터를 찾아내는 기술

- **웹 스크래핑**

- html로부터 필요한 데이터를 수집하는 기술

# 데이터의 수집방법:

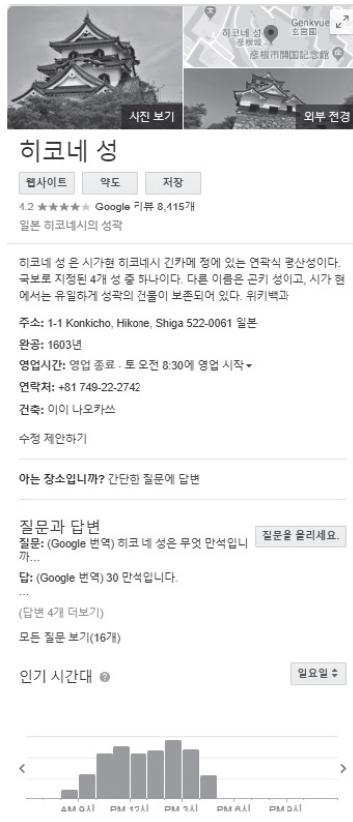
---

1. UCI 머신러닝 리포 [UCI Machine Learning Repository]  
<https://archive.ics.uci.edu/ml/index.php>
2. R에서 제공하는 예제 데이터.
  - a. `help(package='datasets')`
3. 머신러닝/데이터 과학 공유/경연 사이트 캐글  
(<https://www.kaggle.com/>)
4. 위키피디아의 머신러닝 연구를 위한 데이터세트 리스트 (<https://goo.gl/SpCOLK> )
5. 일본정부통계 종합창구 (<http://www.e-stat.go.jp>)



# 데이터의 수집 예: 히코네 성 방문객 수

- 구글 검색 "히코네성"
- 구글 API와 몇줄의 Python 코드를 작성
- csv형식의 파일형태로 입장객 수에 관한 데이터를 수집



〈그림 1.10〉 구글 맵의 데이터 표시

Time	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	11	0	5	0	5	0	0
10	17	0	13	1	15	3	7
11	19	0	21	9	17	13	21
12	17	0	17	11	11	28	34
13	17	0	15	9	15	30	36
14	21	0	17	13	25	36	44
15	26	0	19	19	32	51	69
16	30	0	17	23	34	51	82
17	34	0	15	23	34	34	61
18	44	0	28	42	42	42	55
19	55	0	71	100	59	73	82
20	44	0	73	86	53	61	57
21	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0

〈그림 1.11〉 구글 맵으로부터 수집한 히코네성 입장객수(요일 및 시간별)

# 데이터의 전처리

---

- 결측값 (missing value) 문제: 수집한 데이터에는 결측값 NA(not available) 값이 존재할 수 있다. 결측값은 데이터 중 고의 또는 실수로 누락된 값을 의미한다. 결측값을 그대로 놔둔 채 데이터 가공을 하면 결과값에 오류가 뜨거나 잘못된 연산이 수행될 수 있으므로 정제과정에서 적절한 처리가 필요하다
- 이상값 (outlier) 문제: 수집한 데이터에는 논리적 혹은 통계학적으로 이상한 데이터가 입력되어 있을 수 있다. 이러한 데이터를 이상값이라 한다. 통계학에서 이상값이란 다른 관측값과 멀리 떨어진 관측값을 의미한다.
- 중복과 표기 오류
  - 실수로 인한 오타
  - Data Cleansing: 데이터의 불일치성에 대한 처리
  - 공백문자와 구분기호 유무
- 데이터의 익명화, 개인정보 보호차원의 배려