

데이터분석 프로세스

2020년 1학기

강윤희

차례

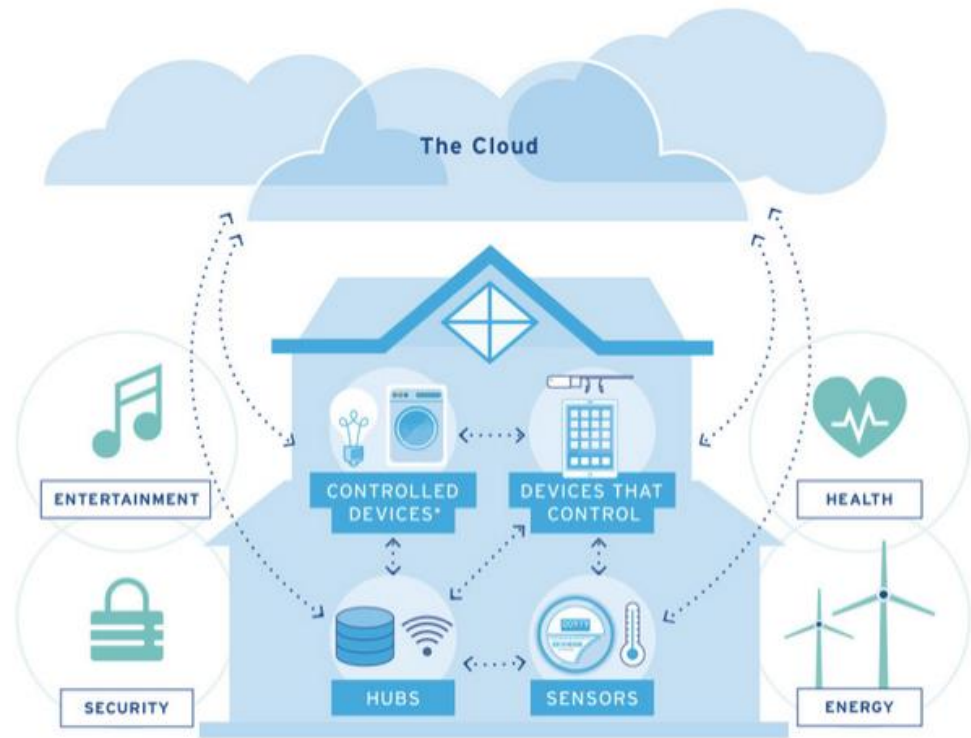
- 주요 IT 기술 변화
- 빅데이터 시대
- 데이터 과학자
- 데이터 분석 필요와 수업의 이해
- 요약

주요 IT 기술 변화

- 모든 과학기술 분야에서의 데이터 홍수(Data Deluge)
- 멀티코어(Multicore) 에 따른 병렬 처리의 중요성
 - 클러속도에서 추가 코어를 통한 성능 개선
 - GPU 을 사용한 처리 능력 및 빠른 계산
- 클라우드
 - 상업적 데이터 센터 모델
 - 소유에서 대여로, 서비스 계산자원 이용
- 경량 클라이언트
 - 센서, 스마트 폰 및 테플릿
 - 백엔드 서비스에 대한 필요성
- 기업 주도의 기술 혁신과 기술 도입



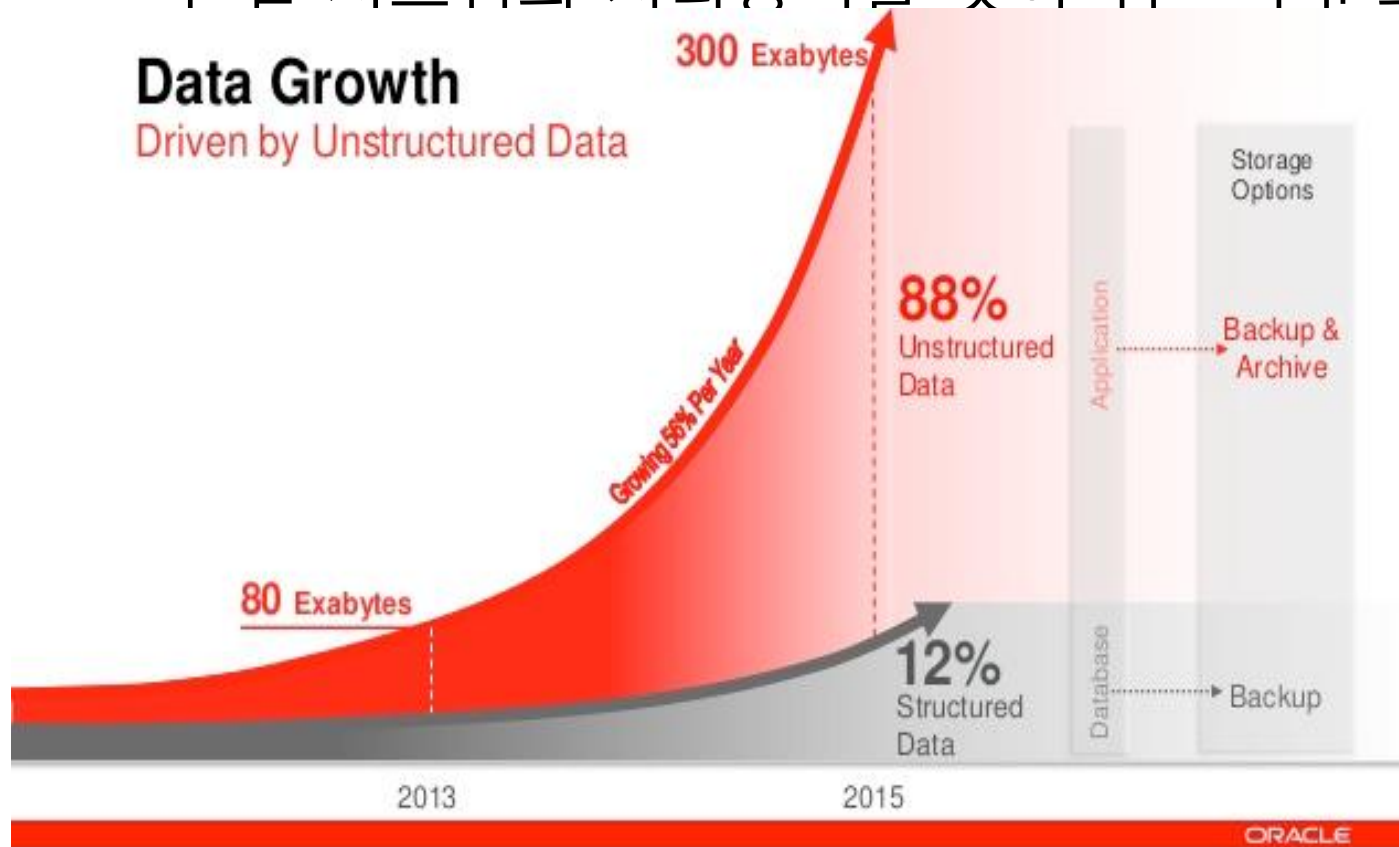
주요 IT 기술 변화



빅데이터와 사물인터넷

빅데이터 시대

- 빅데이터란
 - 단일 시스템의 처리능력을 벗어나는 수준의 데이터셋



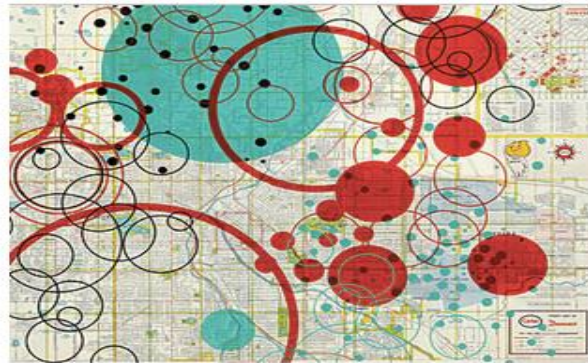
빅데이터 시대

Big Data: The Management Revolution

by Andrew McAfee and Erik Brynjolfsson

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT \$8.95 BUY COPIES



ARTWORK: TAMAR COHEN, HAPPY MOTORING, 2010, SILK SCREEN ON VINTAGE ROAD MAP, 26" X 18"

“You can't manage what you don't measure.”

There's much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker, and it explains why the recent explosion of digital data is so important. Simply put, because of big data, managers can measure, and hence know,

radically more about their businesses, and directly translate that knowledge into improved decision making and performance.

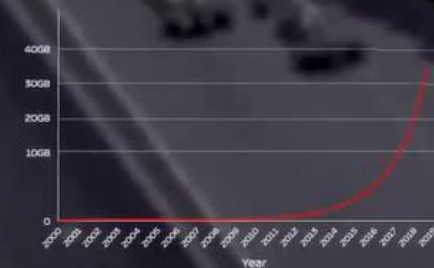
200 sensors
Millions of Data points
Fully dynamic modelling

200 SENSORS



DATA GROWTH

DATA PER A LAP (MB) **2019**
28714 ↗



데이터 과학자

데이터 과학(data science)이란, 데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다.^[1]

데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.

데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다. 사용되는 기술은 여러분야에 걸쳐있으며 수학, 통계학, 계산기과학, 정보공학, 패턴인식, 기계학습, 데이터마이닝, 데이터베이스 등과 관련이 있다. 데이터 과학을 연구하는 사람을 데이터 과학자라고 한다.

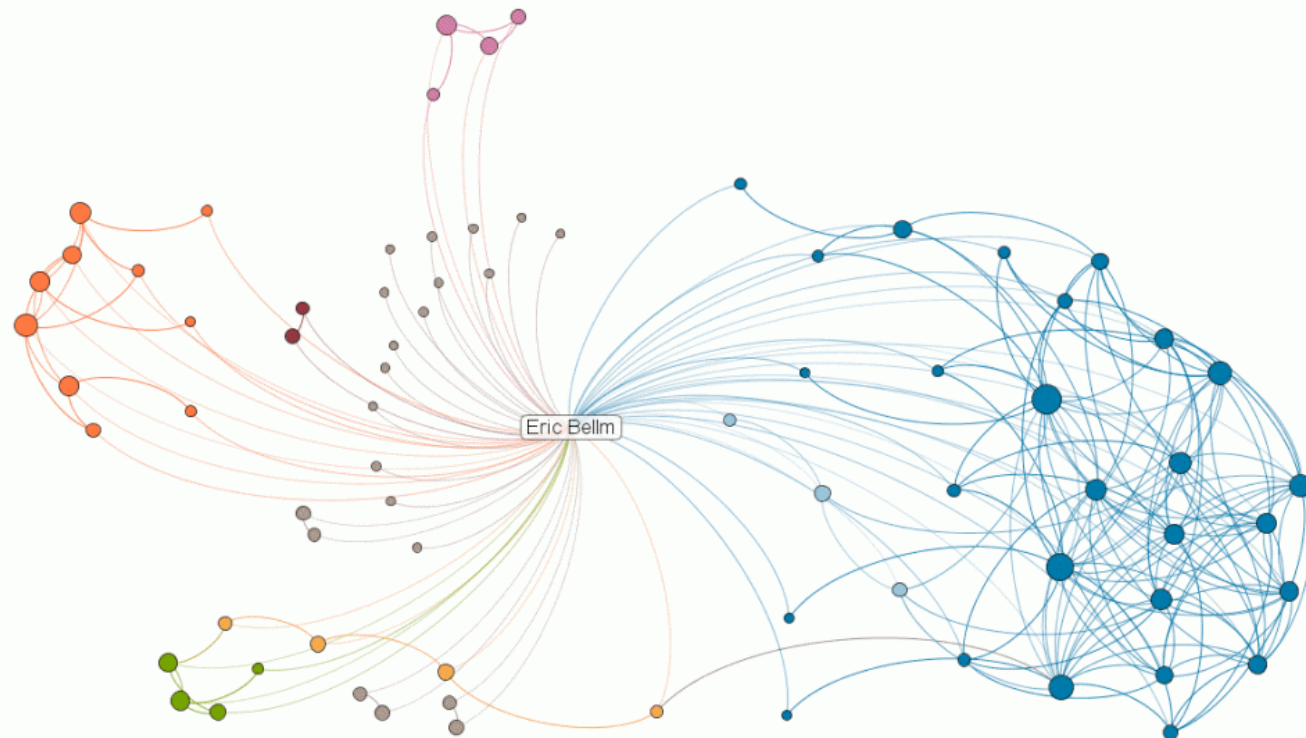
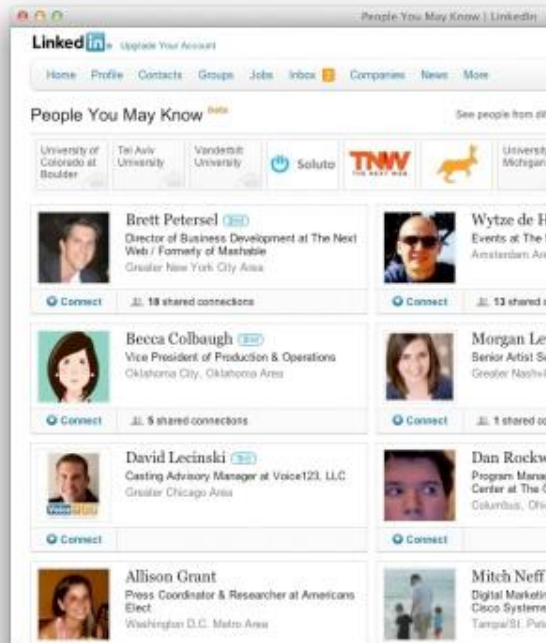
튜링상을 수상한 짐그레이(Jim Gray) 박사는 데이터 과학은 과학의 네번째 패러다임으로 정의하고 과학(경험, 이론, 계산, 그리고 이제 데이터)에 관한 모든 것이 바뀌고 있는데 이유는 정보 기술과 데이터 범람(data deluge) 때문이라고 주장했다.

데이터 과학은 생물학, 의학, 공학, 사회학, 인문과학 등의 여러 분야에 응용되고 있다.

데이터 과학자

How Target Figured Out A Teen Girl Was

LinkedIn Maps Eric Bellm's Professional Network
as of May 30, 2011



데이터 과학자



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

XML 이론 및 실습



데이터분석 방법 및 실습

데이터 분석 필요와 수업의 이해

iceman67 / -Python

[Code](#)

[Issues 0](#)

[Pull requests 0](#)

[Actions](#)

[Projects 0](#)

[Wiki](#)

[Security](#)

No description, website, or topics provided.

[Manage topics](#)

[29 commits](#)

[1 branch](#)

[0 packages](#)

[Branch: master](#)

[New pull request](#)

[Create new branch](#)

[iceman67](#) [Create README](#)

pet_ftp	Add files via upload
1thread.py	Add files via upload
2thread.py	Add files via upload
Average Var StdDev .ipynb	Add files via upload

데이터 분석 필요와 수업의 이해

← → ↺

github.com/iceman67/Web-Scraping

Apps ★ Bookmarks 새 탭 Science Cloud Sum... k yunheekanghomep... ICECECE 2013 - Int... CS545: Machine Le...

Web Scraping with Python

[Manage topics](#)

🕒 7 commits


🔑 1 branch








📦 0 packages

Branch: master ▼

New pull request

Create new

 **iceman67** Update README

 Getting started Web scraping.ipynb	Add files via upload
 README	Update README
 Web Scraping.ipynb	Add files via upload
 Worldcup 2018 player.ipynb	Add files via upload
 books_and_authors.csv	Add files via upload
 data visualization with python.ipynb	Add files via upload
 simple_web_scraper.py	Add files via upload

데이터 분석 필요와 수업의 이해

[←](#) [→](#) [↻](#) [github.com/iceman67/Public_OpenAPI](#)

[Apps](#) [★ Bookmarks](#) [🌐 새 탭](#) [Science Cloud Sum...](#) [k](#) [yunheekanghomep...](#) [ICECECE 2013 - Int...](#) [CS CS545: Machine Le...](#) [🔍](#) [🔍](#)

[↔ Code](#) [! Issues 0](#) [🔗 Pull requests 0](#) [▶ Actions](#) [📊 Projects 0](#) [📖 Wiki](#) [🛡️ Secu](#)

No description, website, or topics provided.

[Manage topics](#)

🔒 5 commits

🔗 1 branch

📦 0 packages

Branch: master ▼

New pull request

Create new

🚧 iceman67 Update README

📄 [02-1-XML-응용-openAPI.pdf](#)

Add files via upload

📄 [README](#)

Update README

📄 [WeatherSightSeeing.py](#)

Add files via upload

📄 [weather_app.py](#)

Add files via upload

데이터 분석 필요와 수업의 이해

Branch: master ▾

Data-Visualization / Geo Starbucks.ipynb



iceman67 Add files via upload

1 contributor

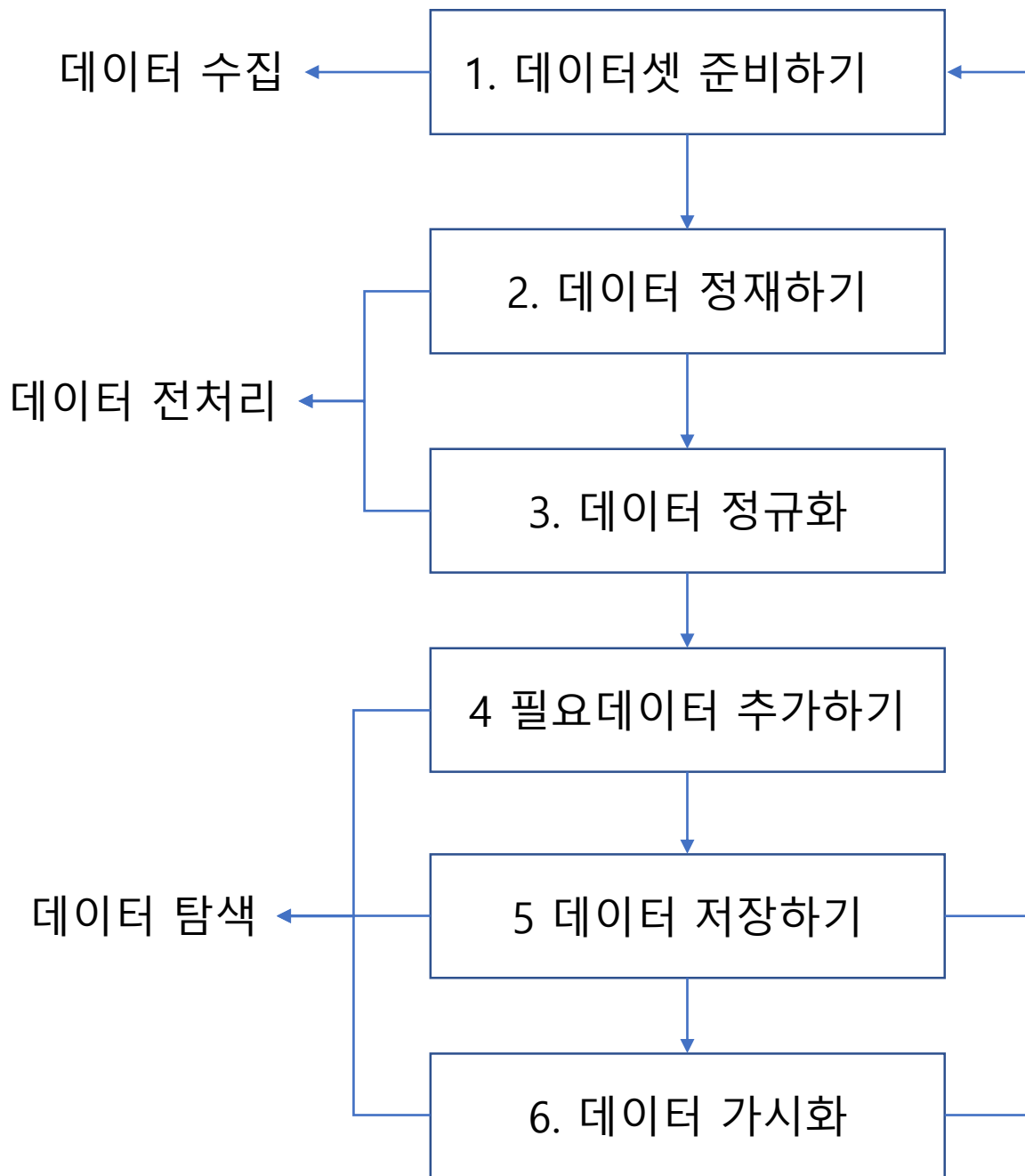
401 lines (401 sloc) | 61.4 KB

```
In [1]: import requests
import json
import pandas as pd
from pandas.io.json import json_normalize
```

```
In [2]: data = {
    'ins_lat': '37.56682', # 지정한 위도와 경도에서 가까운 순으로 나열
    'ins_lng': '126.97865',
    'p_sido_cd': '01', # 01=서울시, 08=경기 ... 16=제주
    'p_gugun_cd': '', # 세부지역 (지정하지 않으면 시/도 전체)
    'in_biz_cd': '',
    'set_date': '',
    'ins_cd': '1000'
```


데이터 파이프라인

재사용 가능한 코드 구축



In [7]: `df=pd.read_csv('season.csv')`

`# This is our dataframe`
`df`

Out[7]:

	dates	day	temp	wind-speed
0	2/1/2012	sunny	45.0	12
1	3/1/2012	rainy	46.0	34
2	4/1/2012	hot	47.0	45
3	5/1/2012	NaN	NaN	56
4	6/1/2012	hot	49.0	Not available
5	7/1/2012	NaN	NaN	Not available
6	8/1/2012	hot	12.0	45
7	9/1/2012	rainy	23.0	41
8	10/1/2012	NaN	NaN	NaN
9	11/1/2012	NaN	NaN	NaN

수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정

데이터 수집

1. 데이터셋 준비하기



원시데이터 수집방법

- 웹(데이터) 스크래핑(scraping)
- 파일
- API
- DBMS

원시데이터 종류 별 특징

- 자연어로 작성된 비정형 텍스트 (한국어, 영어, 중국어...)
- 정형데이터
 - CSV
 - JSON
 - HTML/XML 마이크업 자료
 - 데이터베이스에 저장된 테이블 자료

오픈데이터 활용하기



여기,
서울시의
모든 공공데이
모아서 개방합

오픈 API란?
개방된 공공데이터를 누구나 사용할 수 있도록 공개된 API(application program interface)를 말하며, 시민이 직접 응용 프로그램과 서비스를 개발할 수 있는 원천데이터를 제공하고 있습니다

[오픈 API 사용방법](#) [오픈 API 목록보기](#) [데이터셋 목록보기](#)

▲ ▼

목록

답장

전체회신



전달

삭제

헤더보기

저장

프린트

보낸사람	memberadmin@seoul.go.kr	 수신거부	 주소록추가
받는사람	강윤희		
보낸날짜	2015-06-28 13:51:11		
제 목	서울시 홈페이지 회원 가입을 축하드립니다.		
강윤희님의 회원 가입을 진심으로 축하 드립니다. 강윤희님은 2015년 06월 28일 서울시 홈페이지 회원으로 등록 하셨습니다, 등록하신 아이디는 입니다. 앞으로 많은 이용 부탁드립니다.			
Character Set :UTF-8			

<http://data.seoul.go.kr/>

오픈데이터 활용하기

• 주제 선정하기

홍대입구	20029844
잠실	21946587
고속터미널	24677365
서울역	28492186
강남	28903020

지하철 호선별 역별 시간대별 승객 현황 [상세정보 보기](#)

교통카드(선후불교통카드 및 1회용 교통카드)를 이용한 지하철 호선별 역별(1~9호선, 서울시 관할 운송기관에 한함) 시간대별 승하차인원을 나타내는 정보입니다.

제공기관 : 서울특별시

등록일 : 2015.02.17

[SHEET](#) [Open API](#)

지하철 호선별 역별 유/무임 승객 현황 [상세정보 보기](#)

교통카드(선후불교통카드 및 1회용 교통카드)를 이용한 지하철 호선별 역별(1~9호선, 서울시 관할 운송기관에 한함) 유/무임차인원을 나타내는 정보입니다.

제공기관 : 서울특별시

등록일 : 2015.02.17

[SHEET](#) [Open API](#)

지하철역별 승하차인원 [상세정보 보기](#)

교통카드(선후불교통카드 및 1회용 교통카드)를 이용한 지하철역별(서울메트로, 도시철도공사, 한국철도공사, 공항철도, 9호선) 하차인원을 나타내는 정보입니다.

제공기관 : 서울특별시

등록일 : 2013.08.01

[SHEET](#) [Open API](#)

오픈API란?

오픈API는 다양한 서비스와 데이터를 좀더 쉽게 이용할 수 있도록 공개한 개발자를 위한 인터페이스입니다.

오픈API 이용방법

01. 오픈API를 사용하기 위해서는 ID와 인증키가 필요합니다.
02. 오픈API 인증키를 신청해 주세요.
03. 인증키는 타인에게 양도할 수 없습니다.
04. 오픈API별 개발 명세서에 설명되어 있는 포맷에 따라 목적에 맞게 활용하세요.
05. 오픈API를 통해 개발 한 어플리케이션을 이용자들이 활용할 수 있도록 열린데이터 광장에 등록해주세요.


01 열린데이터광장
포털사이트 접속

02 오픈API
인증키신청

03 오픈API검색 및
이용방법확인

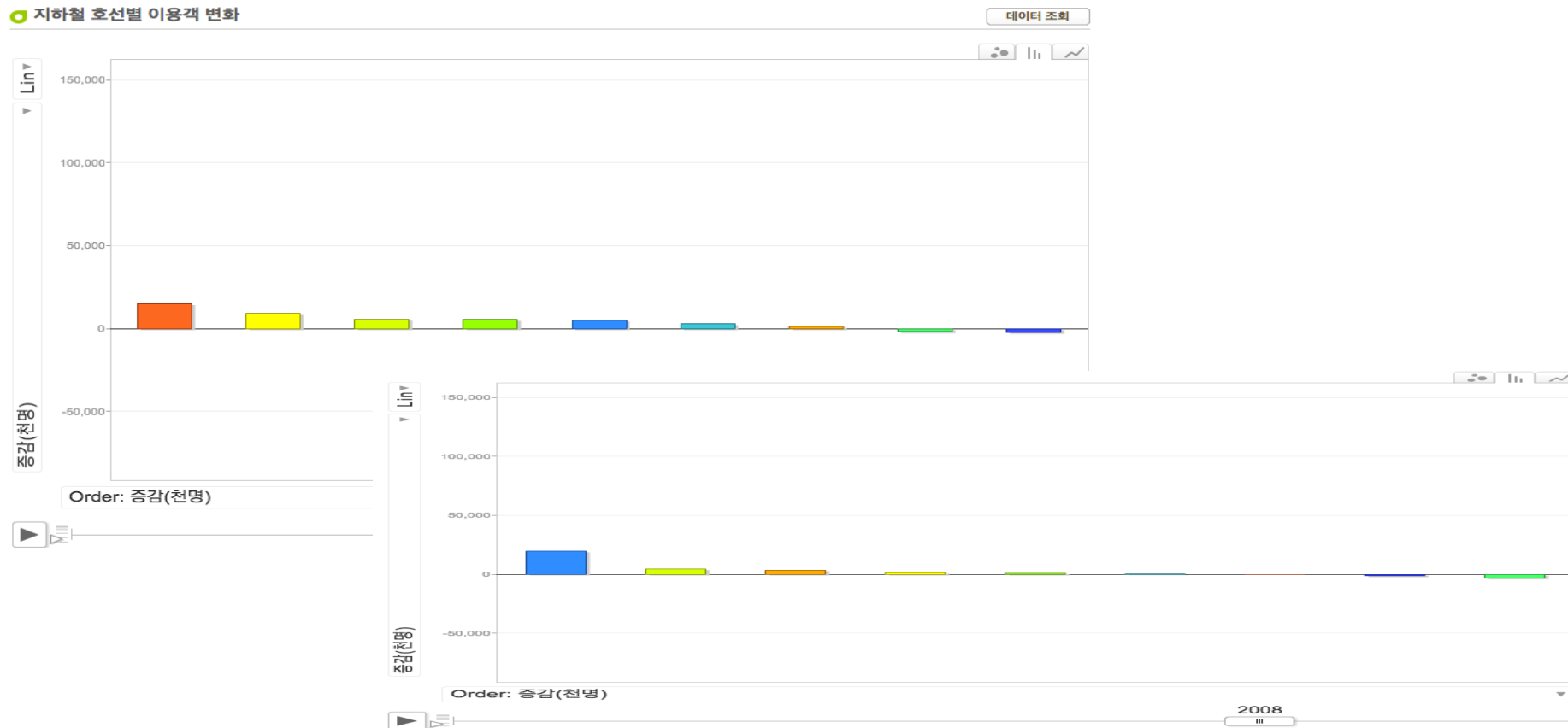
04 오픈API를 이용
어플리케이션제작

05 어플리케이션
등록

 서울 열린데이터 광장
SEOUL OPEN DATA PLAZA

오픈데이터 활용하기

- 지하철 이용정보 가시화 하기



요약

- 데이터 생산에 사물이 참여로 빠르게 이루어지고 있음
- 데이터 활용을 위한 인공지능(기계학습) 활용이 다양한 분야에서 이루어지고 있음
- 데이터 과학(자)에 대한 사회적 요구가 증가하고 있음
- 데이터를 다루기 위해 기반 수집, 전처리, 탐색 기술에 대한 학습을 진행