

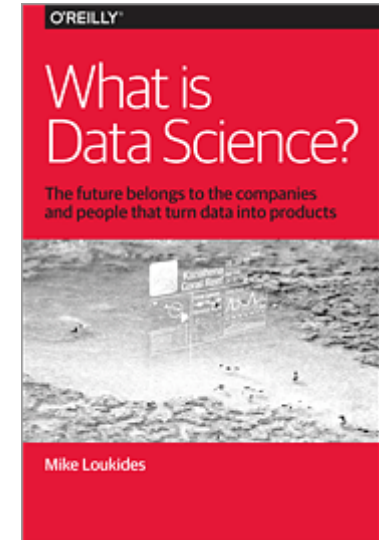
Project Pipeline of Data Science



SIT22009Data Science
Presented by Hyebyong Choi

DATA SCIENCE

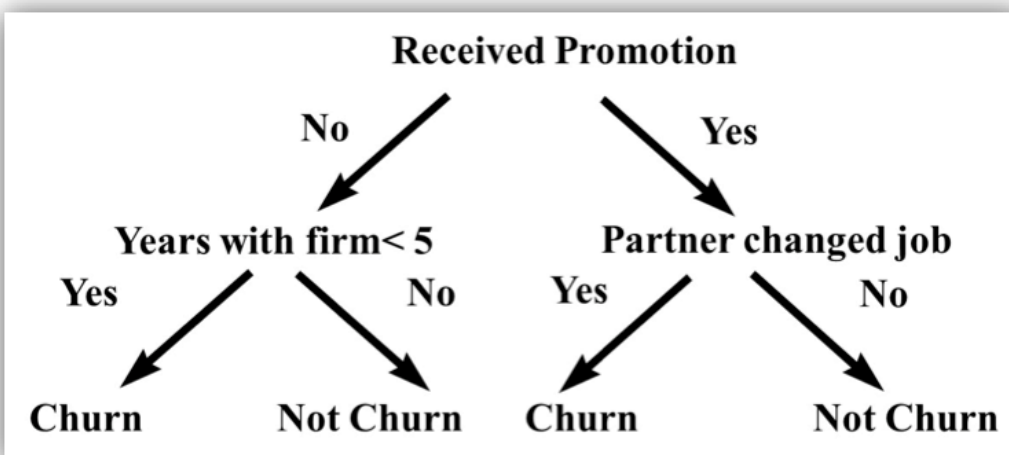
- ❑ **Data Science** aims to derive **knowledge** from big data, efficiently and intelligently
- ❑ **Data Science** encompasses the set of activities, tools, and methods that enable data-driven activities in science, business, medicine, and government
- ❑ **Machine Learning**(or **Data Mining**) is again one of the core technologies that enables **Data Science**



DATA SCIENCE PROJECTS

❑ Churn Prediction

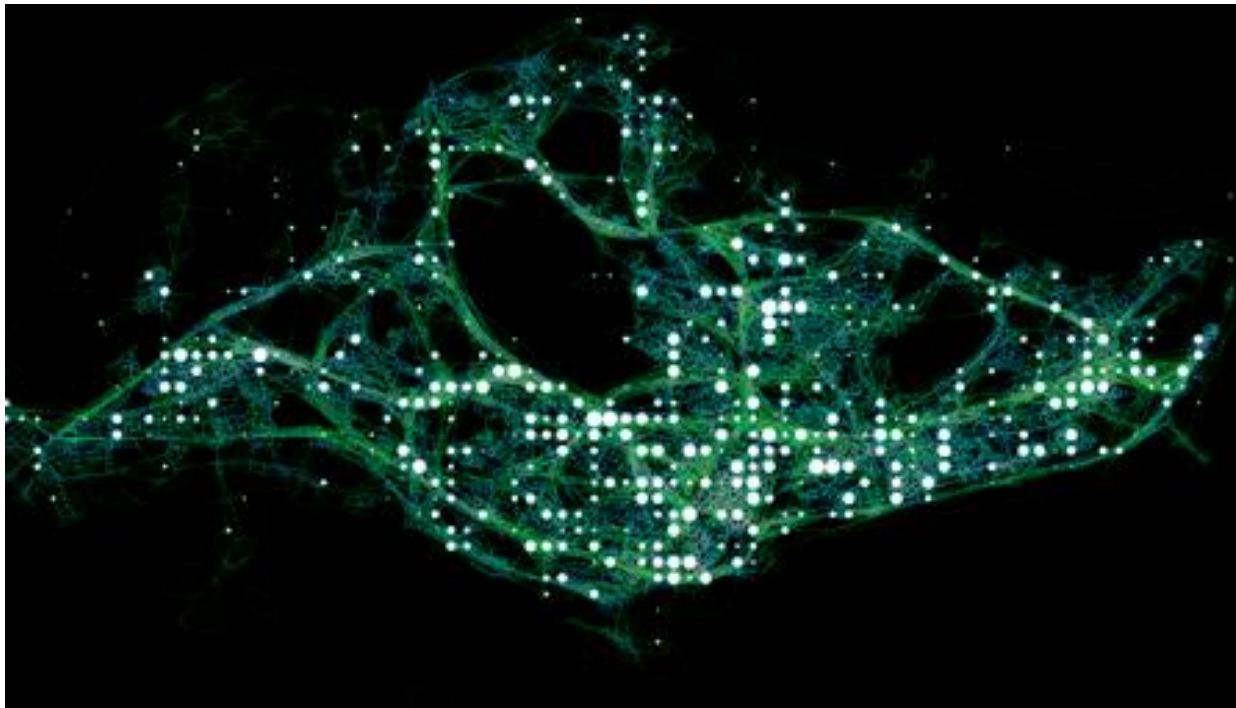
- ❑ Keeping customer loyal and not escaping is crucial as much as observing new customers especially for Telco. company, Credit card company
- ❑ Churn: Customers is just about to leave or finish the relationship and subscription the company provides
- ❑ Goal: Reduce Churn Rate by churn prediction
- ❑ escaping customer / incoming customers



DATA SCIENCE PROJECTS

□ Data Visualization

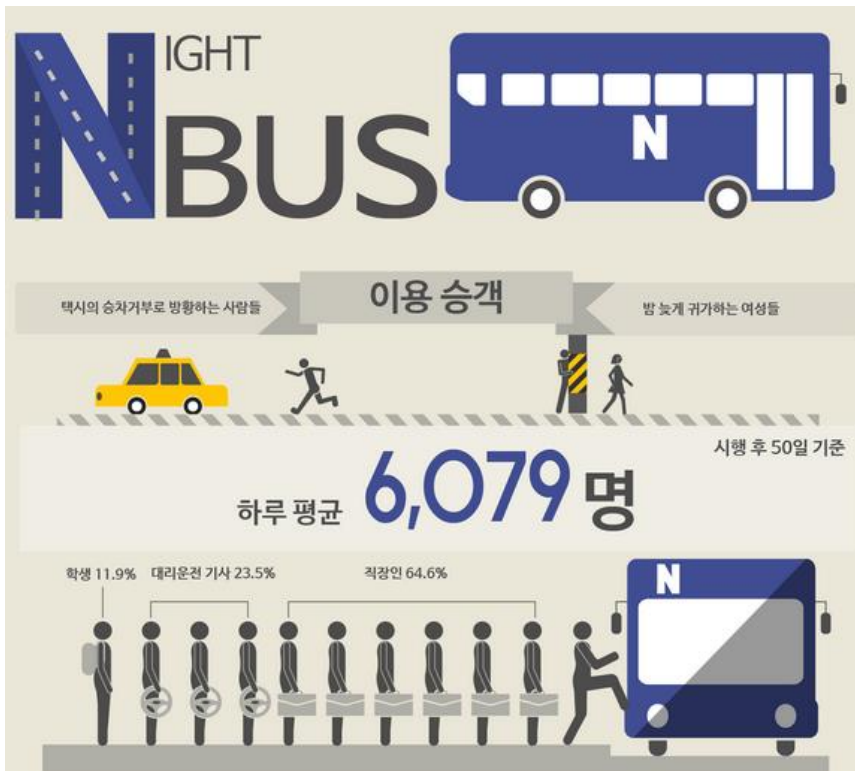
- For authorities, monitoring real-time complex data at a glance would help them make prompt and right decision
- Singapore's Traffic monitoring system, Subway, bus, taxi, express-way info.



DATA SCIENCE PROJECTS

❑ Seoul City-Government, Night-time bus project

- ❑ Before launching the service, to design the bus route that maximize utilization and citizen's satisfaction given limited budget
- ❑ The government analysed night-time floating population derived from 5 million taxi-on/off data, and 3 billion phone-calls (KT)

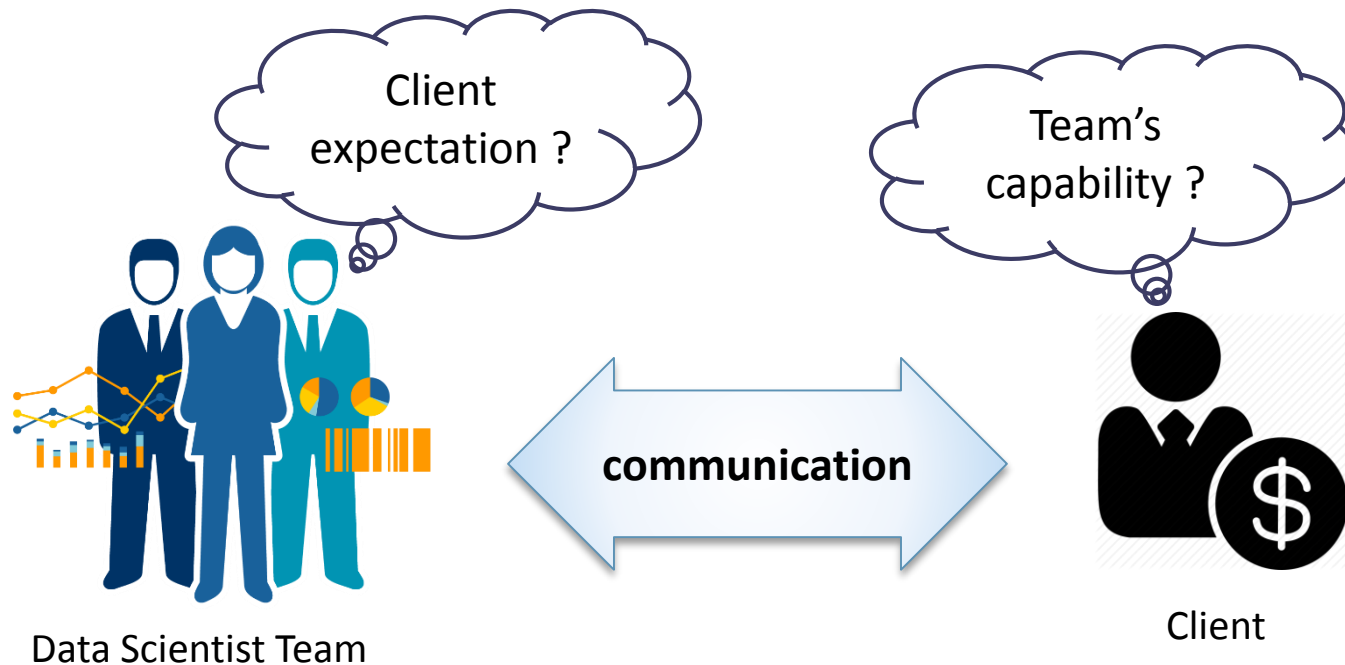


Project Pipeline of Data Science

- Typical Data Science Process
 1. Goal setup
 2. Data Collection
 3. Data Preprocessing (verification, cleaning)
 4. Data Analysis
 5. Evaluation and Revision
 6. Documentation and Presentation of Result
 7. Deployment

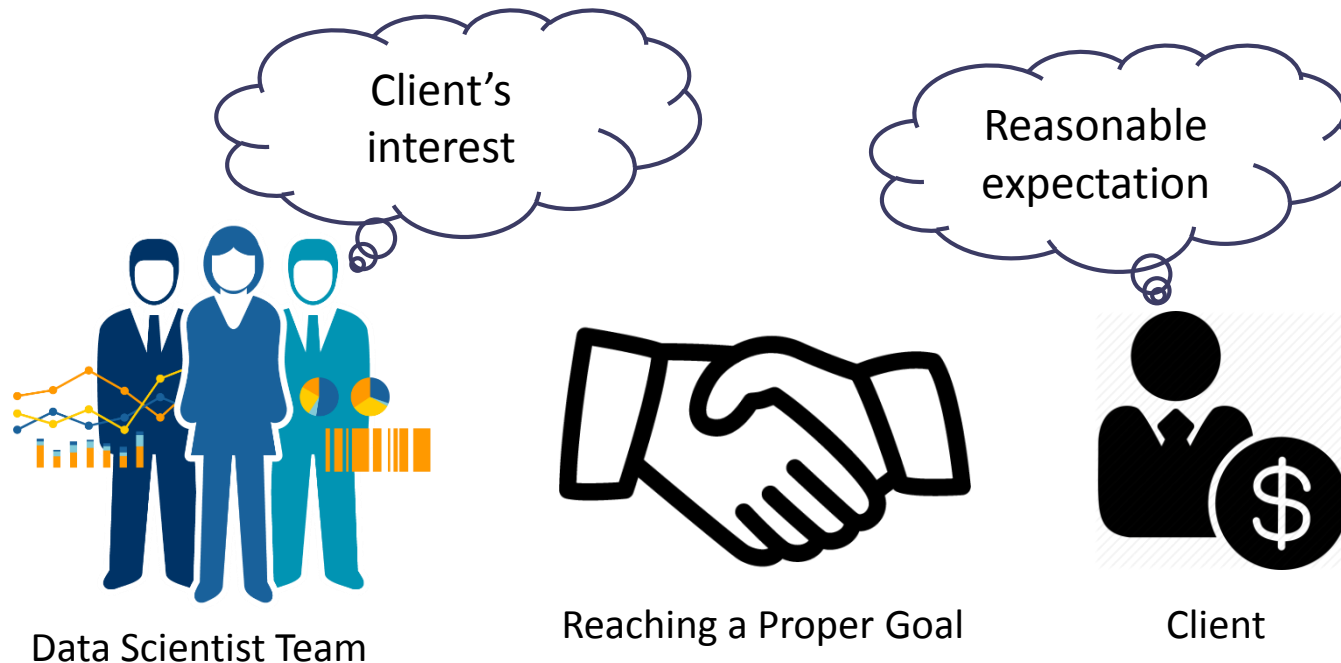
Goal Setup

- Setting up a proper goal is the key to the successful project



Goal Setup

- Setting up a proper goal is the key to the successful project
 - understanding of client's domain knowledge and culture
 - setting up proper measures along with the goal (baseline model)



Goal Setup

- To Goal should be
 - reasonable
 - considering the capability of DS team and resource (time, men-power, ...)
 - considering available data (quality, timeliness, ...)
 - clear
 - should be able to state success or failure of the project
 - Something **good enough** is never **clear enough**
 - e.g. raise click-through-rate of online advertisement by 10%
 - e.g. improve accuracy of existing detecting system by 5%
 - Measurable
 - How to evaluate the performance of the project
 - Testing scheme

Data Collection

- Client may provide the data
 - More important is data verification
 - close communication is needed
 - May require additional data collection process
- To check license, privacy, and confidentiality issue

Data Collection

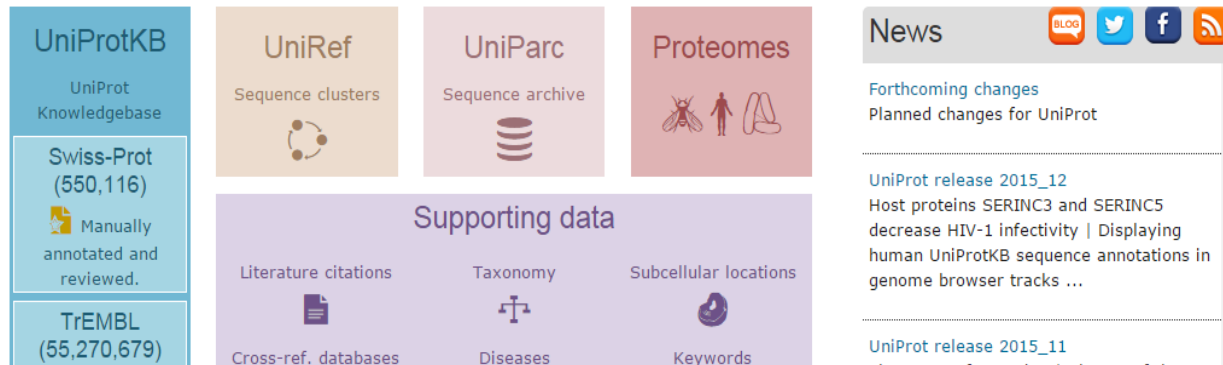
- Open Data Project

- Uniprot

- provide protein sequence and functional information
 - <http://www.uniprot.org/>



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.



Data Collection

- Open Data Project
 - GDELT Project
 - Daily world-wide events
 - <http://gdeltproject.org/>



Data Collection

- Open Data Project

- Public Government Data

- data.go.kr

- data.seoul.go.kr

The screenshot displays the homepage of the data.go.kr portal. At the top, there is a header with a login bar on the left and navigation links on the right. The main navigation bar includes three primary categories: FILE DATA, OPEN API, and STANDARD DATA. Below this, a grid of 16 icons represents various government sectors such as Education, National Administration, Public Administration, Finance, Industry, Social Welfare, Food Safety, Cultural Heritage, Healthcare, Disaster Management, Transportation, Environment, Science, Agriculture, and Law. The bottom section features three featured content blocks: '국가 중점개방 데이터' (National Key Open Data), '공공데이터 활용사례' (Public Data Usage Cases) with a 'My Treasure' example, and a 'WEB AWARD 32nd WINNER' announcement for the 12th Web Award in Korea.

일반사용자

로그인 | 회원가입 | 사이트맵 | 마이페이지

정부 DATA .GO.KR 공공데이터포털

데이터셋 | 활용사례 | 참여마당 | 정보공유

FILE DATA

OPEN API

STANDARD DATA

교육

국토관리

공공행정

재정금융

산업고용

사회복지

식품건강

문화관광

보건의료

재난안전

교통물류

환경기상

과학기술

농축수산

통일외교안보

법률

국가 중점개방 데이터

국민의 손으로 직접 선정한 '국가 중점개방 데이터' 36대 분야를 대용량 데이터로 개방합니다.

공공데이터 활용사례

나의 보물 - My Tr...

★ MyTreasure 아기 예방접종 관리 ★ 아기 정보를 등록하면 국가 기본 예방접종 및 기...

WEB AWARD 32nd WINNER

제12회 웹어워드코리아

공공/의료부문 통합대상 | 공공데이터포털

행정자치부 | NIA 한국정보화진흥원

Data Collection

- Open Data Project

- EU Open Data Portal

- <https://open-data.europa.eu/en/data/>

The screenshot shows the European Union Open Data Portal. At the top, there's a navigation bar with links like Sitemap, Legal notice, Contact, and a language selector. Below this is the portal's logo and name. A secondary navigation bar includes links for Data, Applications, Linked Data, Developers' corner, and About, along with a 'Data provider's area' link. A search bar is prominently displayed with the text 'Find datasets...'. Below the search bar, there are options to 'Show results with' (all of these words, any of these words, the exact phrase) and a count of 'Total datasets available: 7894'. To the right of the search bar is a 'Suggest a dataset' section with a question and a link to 'Please request the dataset>>'. The main content area is divided into two sections: 'Most viewed datasets' on the left and 'Browse datasets by subject or groups' on the right. The 'Most viewed datasets' section lists five datasets with their view counts. The 'Browse datasets by subject or groups' section features a grid of icons representing various subjects: Employment and working conditions, Social questions, Economics, Finance, Trade, Industry, Education and communications, and Production, technology and research. A link for 'more subjects' is at the bottom right.

European Union Open Data Portal

EUROPA > Open Data Portal > Data

Data Applications Linked Data Developers' corner About

Find datasets...

Show results with: ☒ all of these words | ☐ any of these words | ☐ the exact phrase (?)

Total datasets available: 7894

Suggest a dataset
Is there a dataset from the EU that you could not find in this portal?
[Please request the dataset>>](#)

Most viewed datasets [view all >](#)

- «I DGT–Translation Memory (13073 views)
- «I Elevation map of Europe (9901 views)
- «I CORDIS – EU research projects under Horizon 2020 (2014–2020) (9273 views)
- «I EuroVoc (5799 views)
- «I CORDIS – EU research projects under FP7 (2007–2013) (5489 views)

Browse datasets by subject or groups

- Employment and working conditions
- Social questions
- Economics
- Finance
- Trade
- Industry
- Education and communications
- Production, technology and research

[more subjects >](#)

Data Collection


- Open Data Project

- DBLP

- Computer science publications, authors, citations, etc.
 - <http://dblp.uni-trier.de/>

We are currently collecting **comments, criticisms, and testimonials** for the evaluation of our efforts by our public funders. If you want to share any thoughts that we may use in our report, please feel free to send us your comments!

home | browse | search | about





search dblp

[+] Welcome to dblp
[-]

> Home

■ **browse authors | editors**
 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
 ■ **browse journals**
 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z by publisher
 ■ **browse conferences | workshops**
 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
 ■ **browse series**
 CoRR LNCS CEUR-WS LNEE IFIP LNI EPTCS LIPICS other
 ■ **browse monographs**
 books & theses reference works edited collections

[+] About dblp

This service provides open bibliographic information on major computer science journals and proceedings. dblp is a joint service of the  University of Trier and  t3z Schloss Dagstuhl. For more information check out our F.A.Q.

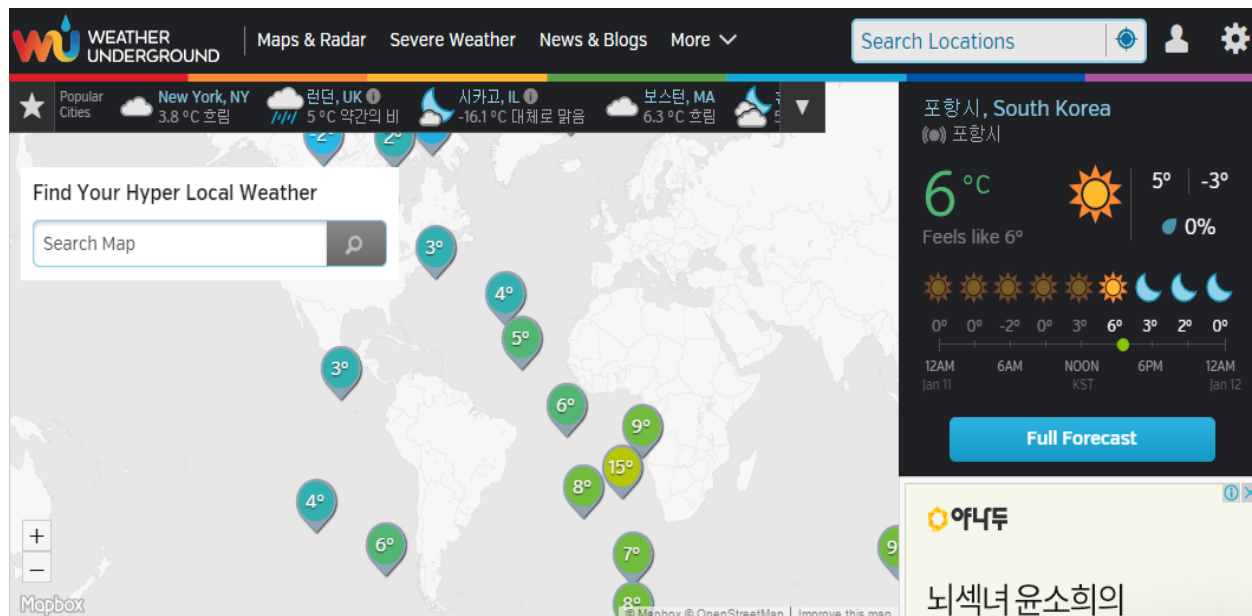
[+] dblp statistics

- # of publications: 3,210,226
- # of authors: 1,667,358
- # of conferences: 4,563
- # of journals: 1,447

[+] News and announcements

Data Collection

- Open Data Project
 - Weather Underground
 - Real-time and Historical World-wide Weather information
 - <http://www.wunderground.com/>



Data Collection


- Open Data Repository








- Welcome to the UC Irvine Machine Learning Repository!
 - 360 data sets as a service to the machine learning community.
 - <http://archive.ics.uci.edu/ml/>



Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 360 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. Our [old web site](#) is still available, for those who prefer the old format. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#). We have also set up a [mirror site](#) for the Repository.

Supported By:  In Collaboration With: 

Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
<p>2013-04-04: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!</p> <p>2010-03-01: Note from donor regarding Netflix data</p> <p>2009-10-16: Two new data sets have been added.</p> <p>2009-09-14: Several data sets have been added.</p> <p>2008-07-23: Repository mirror has been set up.</p> <p>2008-03-24: New data sets have been added!</p> <p>2007-06-25: Two new data sets have been added: UJI Box Characters, MAGIC Gamma</p>	<p>2016-11-23:  NIPS Conference Papers 1987-2015</p> <p>2016-11-16:  Amazon book reviews</p> <p>2016-08-14:  Dota2 Games Results</p>	<p>1267835:  Iris</p> <p>860767:  Adult</p> <p>653079:  Wine</p> <p>562437:  Car Evaluation</p>

Data Verification and Preprocessing

- Data Exploration

- To understand the outline of data
 - number of examples and variables
 - types of variables
 - distribution of each variable, etc.

- Data Verification

- To check consistency and quality
 - errors, outliers, missing values
- How data was collected, by handwriting? sensor?
 - measured value, temperature, humidity, ...
 - calculated value, heat index, discomfort index, ... derived from measured values

Data Verification and Preprocessing

- Data Preprocessing
 - Data Cleaning
 - remove outliers
 - handling missing values
 - remove irrelevant variables
 - Data Processing
 - Joining data
 - Feature extractions

Data Analysis

- Perform the actual Data Analysis
 - Supervised Method
 - Classification, Regression, prediction, fraud detection, recommendation, ...
 - Unsupervised Method
 - Clustering, Dimensionality reduction, ...
- To choose an appropriate method for the **project goal**

Evaluation and Revision

- Check if the result meets the objective set up in the earlier stage
- Internal review
 - inside project team, weekly or bi-weekly basis
- External review
 - with project client
 - Early stages such as goal setup, data verification, frequently

Documentation and Presentation

- To prove that the project achieve the goal
- Result should be reproducible by client side team
 - Documentation should be clear and sufficient
 - Minimizing additional effort for project maintenance

References

- Practical Data Science with R, by Nina Zumel and John Mount
- R을 이용한 데이터 분석 실무, 서민구, 길벗