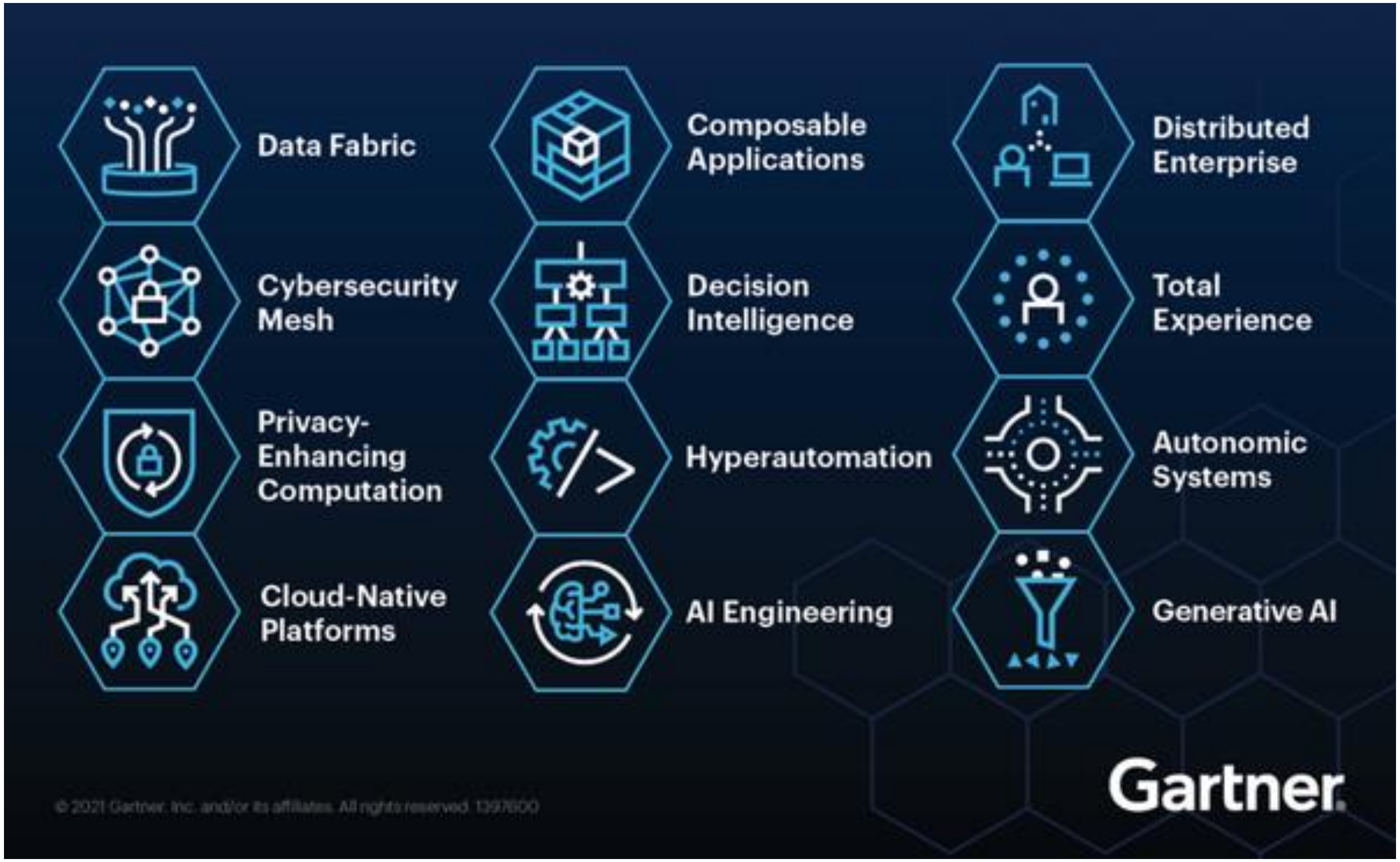
동국대학교
dongguk university

Lecture 1

데이터 애널리틱스 기초

Top Strategic Technology Trends for 2022



Data Fabric

Cybersecurity Mesh

Privacy-Enhancing Computation

Cloud-Native Platforms

Composable Applications

Decision Intelligence

Hyperautomation

AI Engineering

Distributed Enterprise

Total Experience

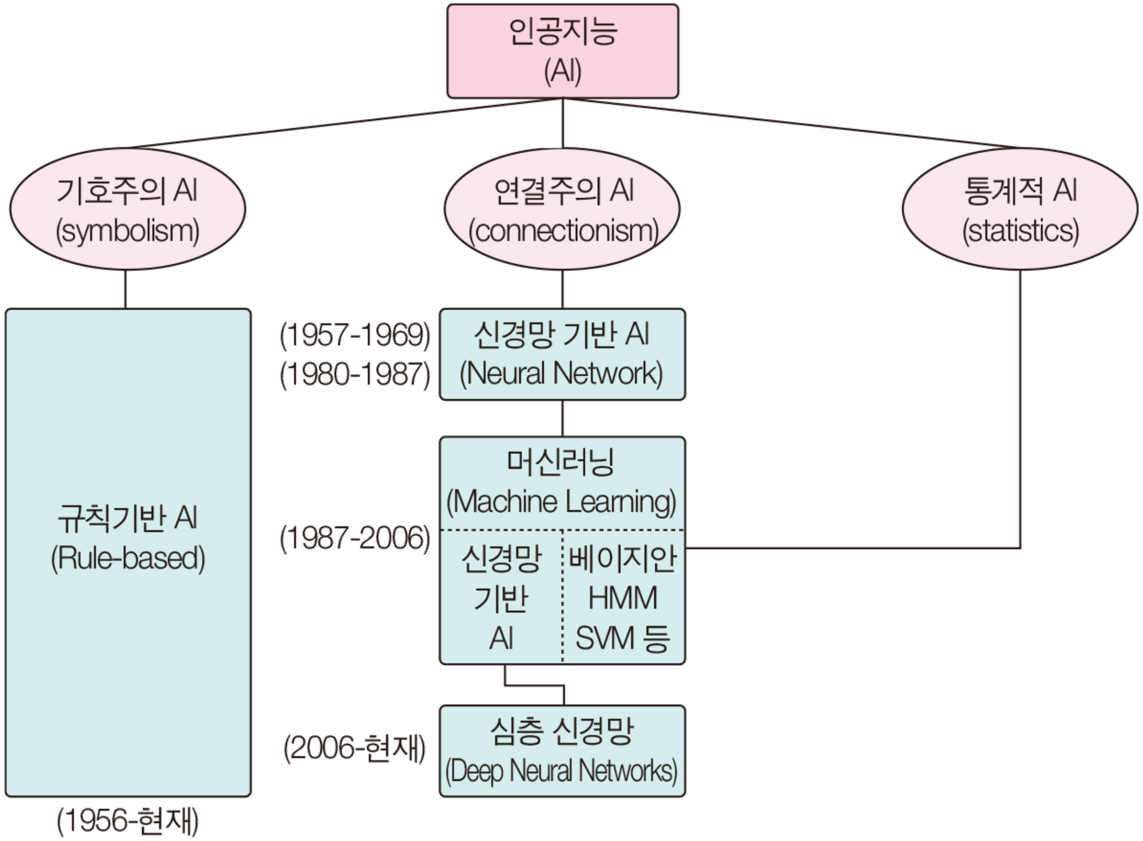
Autonomic Systems

Generative AI

© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. 1397600

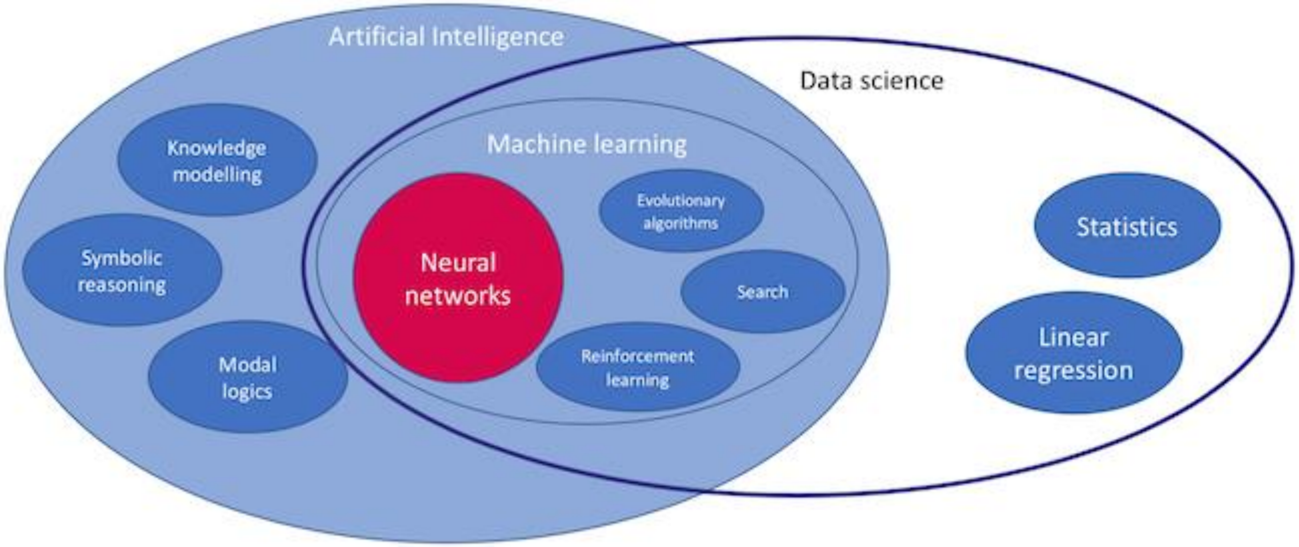
Gartner

인공지능의 분류 체계



자료원: 처음 만나는 인공지능, 김대수, 2020, 생능출판

인공지능과 머신러닝



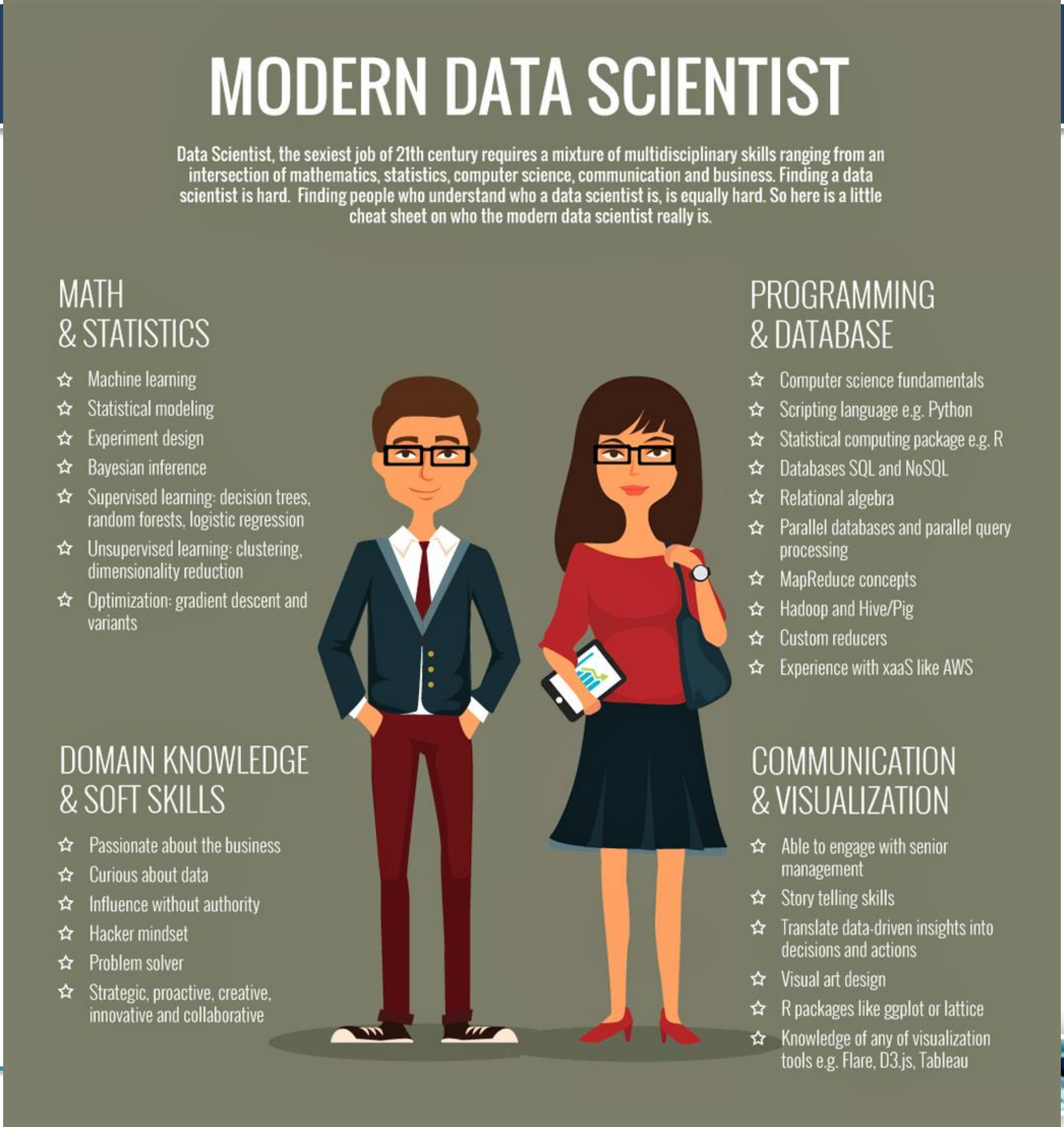
자료원: https://ictinstitute.nl/ai-machine-learning-and-neural-networks-explained/

Data Scientist: *The Sexiest Job of the 21st Century*

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

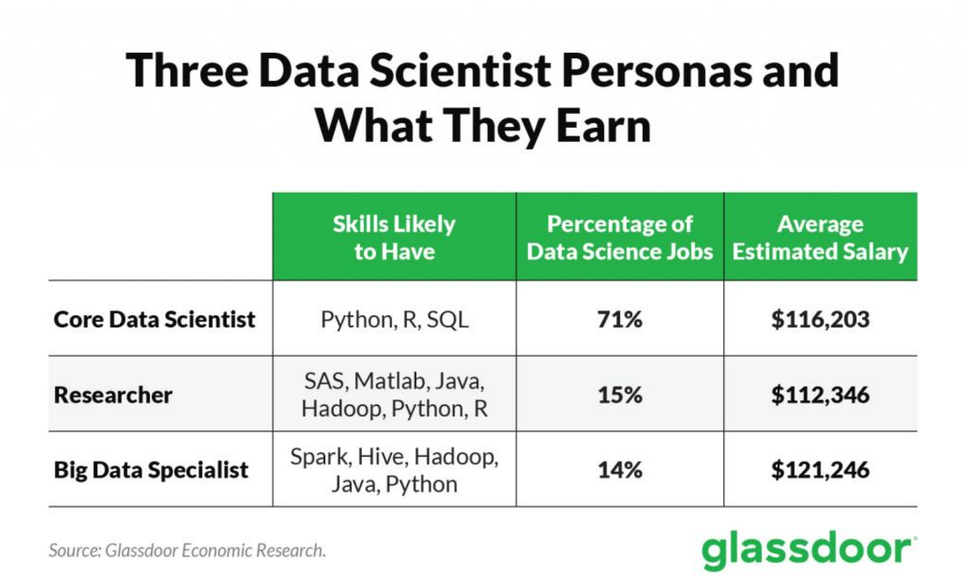
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Three Data Scientist Personas and What They Earn

	Skills Likely to Have	Percentage of Data Science Jobs	Average Estimated Salary
Core Data Scientist	Python, R, SQL	**71%**	**$116,203**
Researcher	SAS, Matlab, Java, Hadoop, Python, R	**15%**	**$112,346**
Big Data Specialist	Spark, Hive, Hadoop, Java, Python	**14%**	**$121,246**

Source: Glassdoor Economic Research.

glassdoor

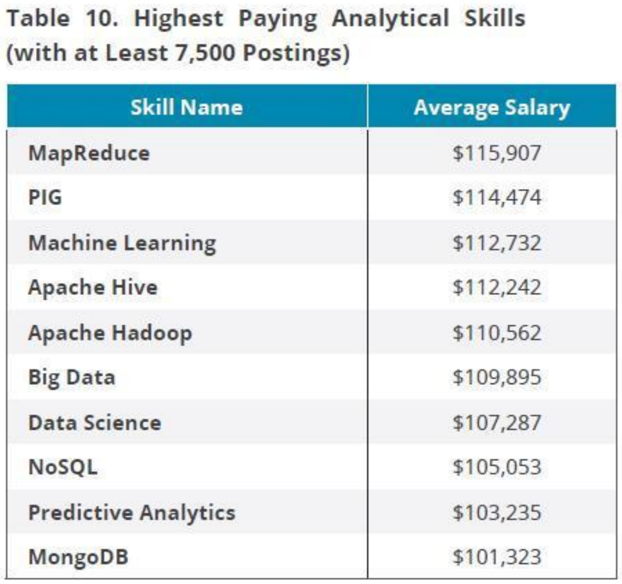


Table 10. Highest Paying Analytical Skills (with at Least 7,500 Postings)

Skill Name	Average Salary
MapReduce	$115,907
PIG	$114,474
Machine Learning	$112,732
Apache Hive	$112,242
Apache Hadoop	$110,562
Big Data	$109,895
Data Science	$107,287
NoSQL	$105,053
Predictive Analytics	$103,235
MongoDB	$101,323

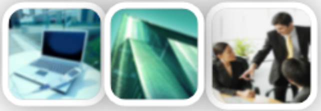


Figure 2. DSA Jobs Matrix

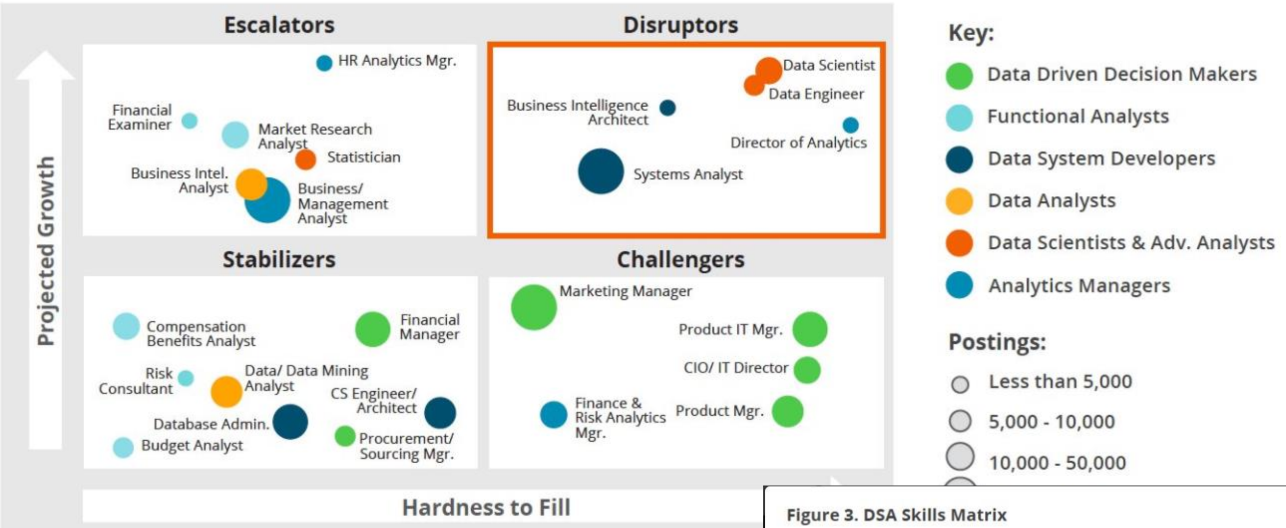
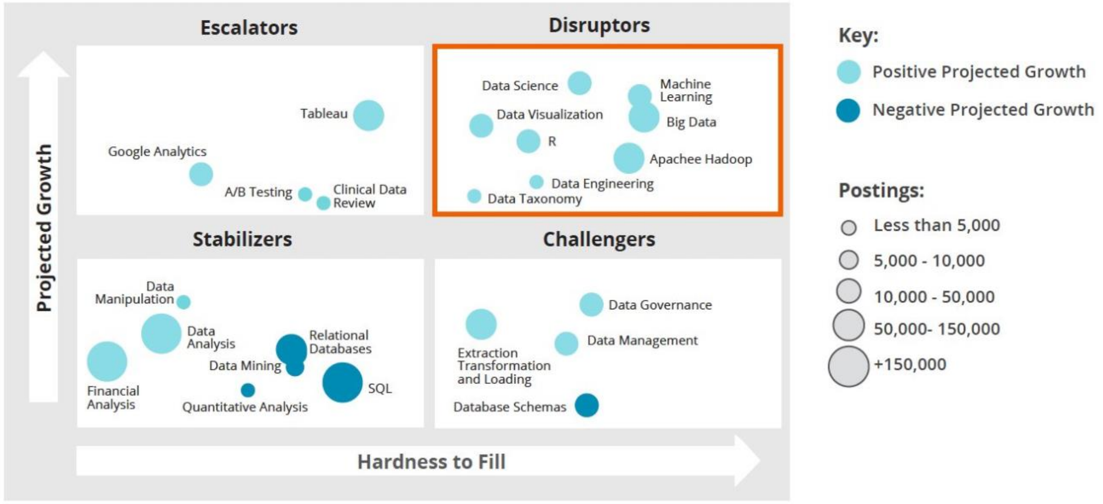
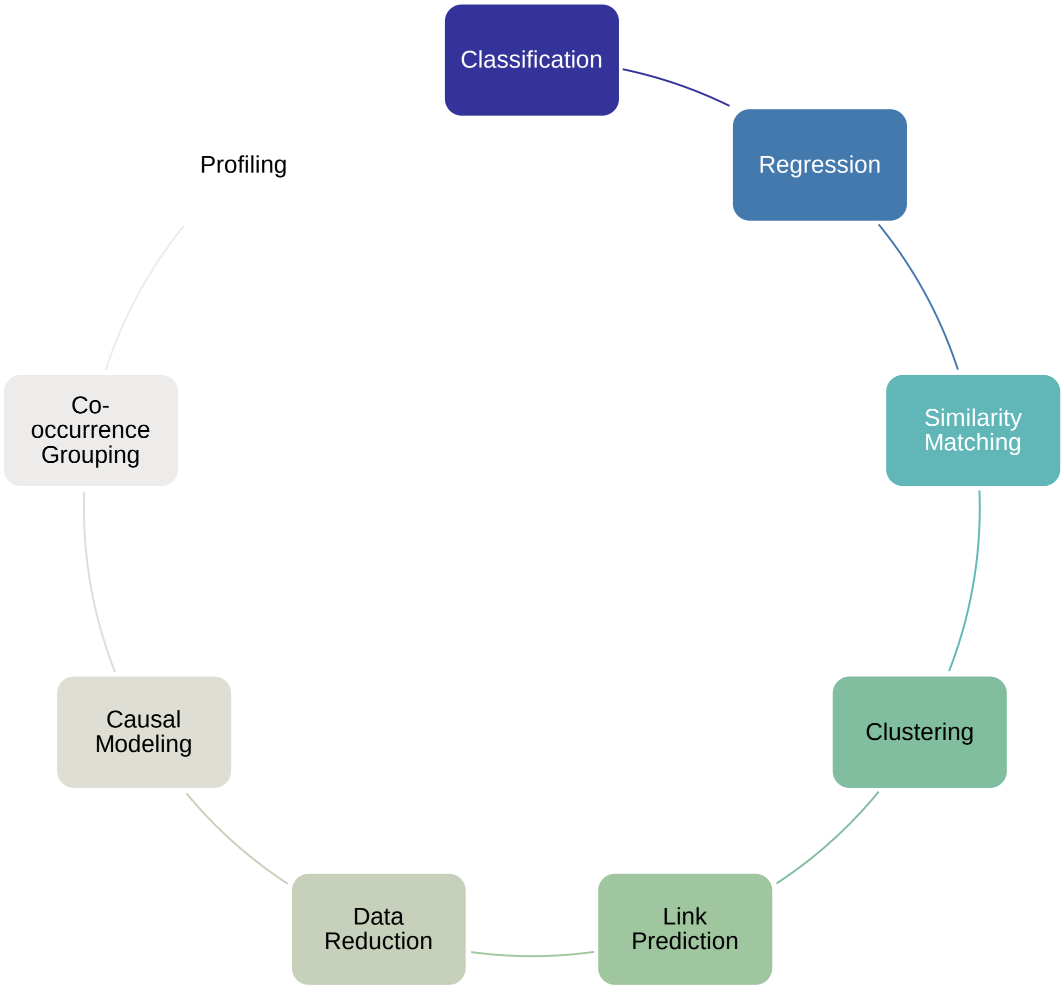


Figure 3. DSA Skills Matrix



Analytics Tasks



Analytics Tasks

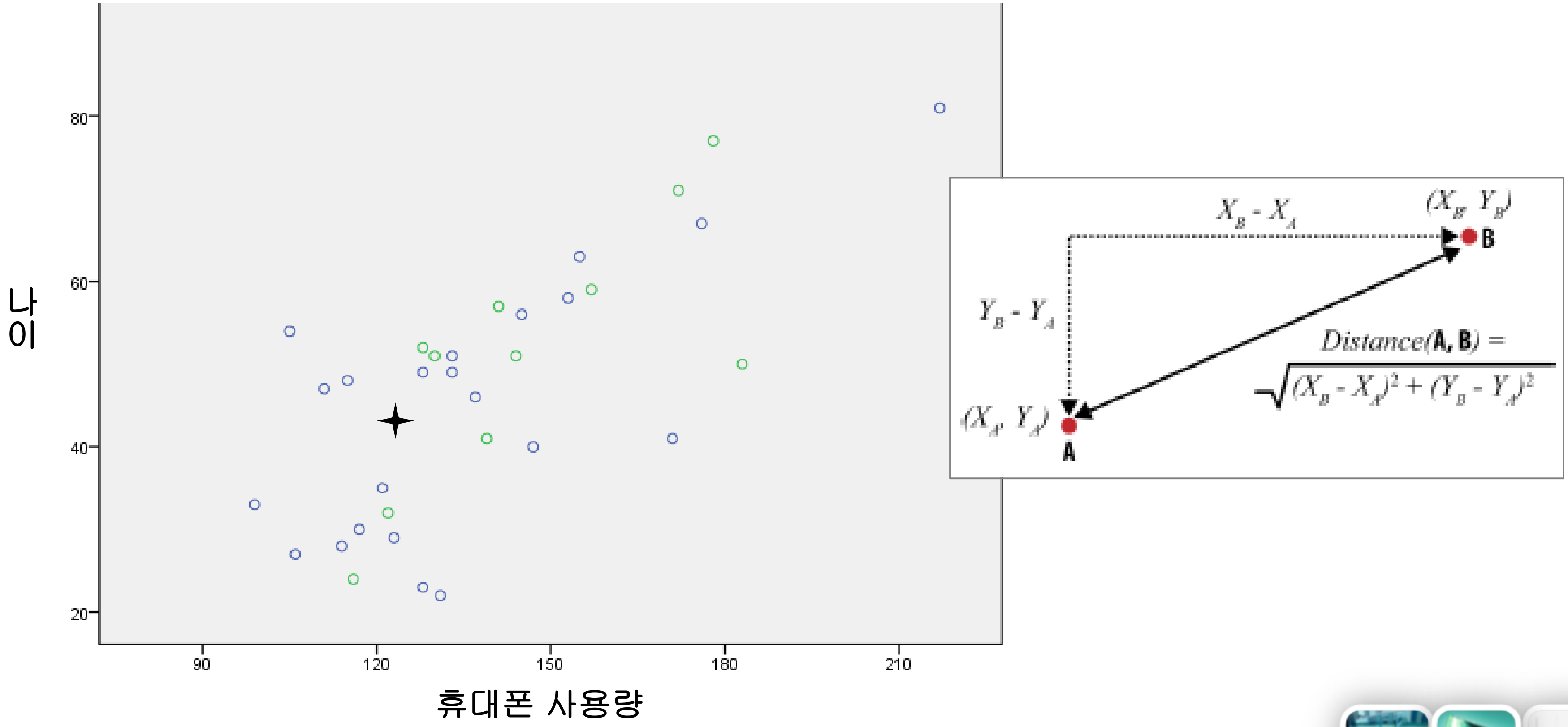
- **Classification: For each individual in a population, identify a (small) set of classes to which that individual belongs.**
  - **Class probability estimation (or scoring) – What is the probability/score that the individual belongs to each class?**

Analytics Tasks

- Regression ("value estimation") is used to estimate or predict, for each individual, the numerical value of some variable.

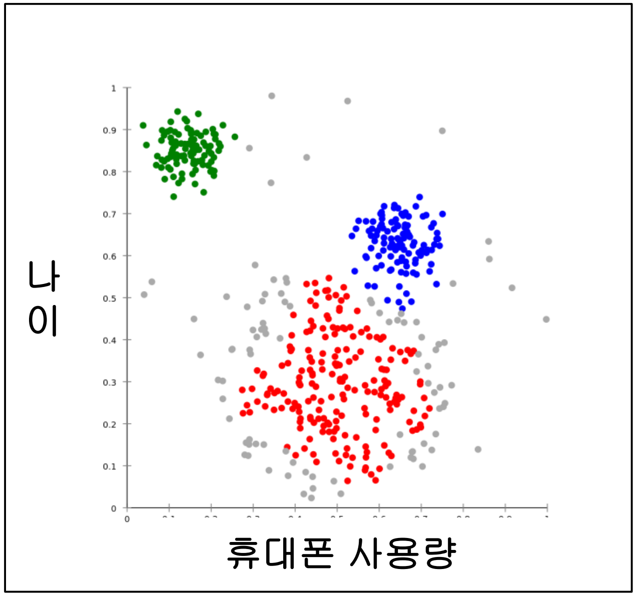
Analytics Tasks

- Similarity matching is used to identify similar individuals based on data known about them. Similarity matching can be used directly to find similar entities.



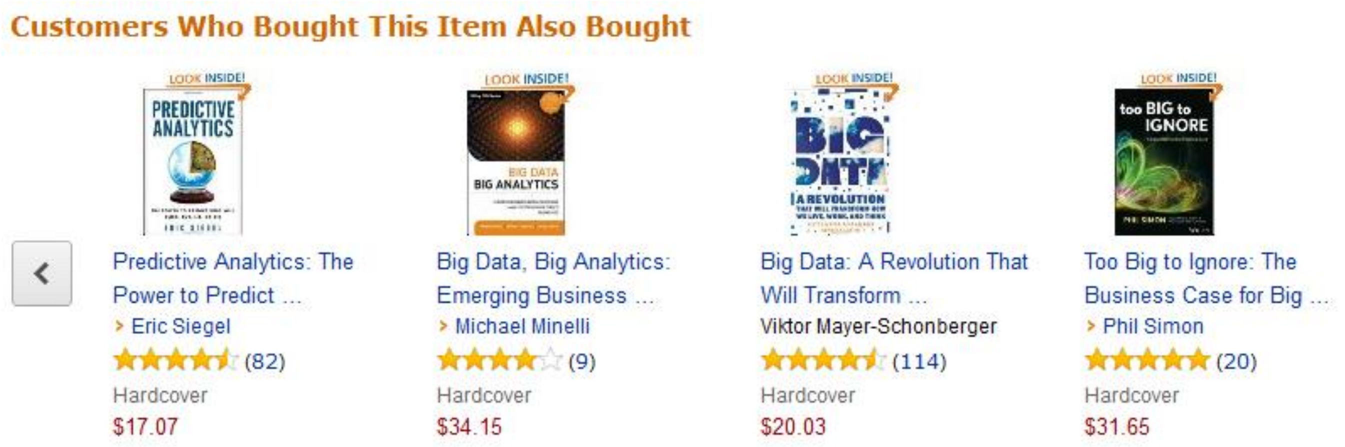
Analytics Tasks

- Clustering is used to group individuals in a population together by their similarity, but not driven by any specific purpose (example or variable).



Analytics Tasks

- Co-occurrence Grouping (also known as frequent itemset mining, association rule discovery, or market-basket analysis) is used to find associations between entities based on the transactions they are involved in.



Customers Who Bought This Item Also Bought

Predictive Analytics: The Power to Predict ...
› Eric Siegel
★★★★½ (82)
Hardcover
$17.07

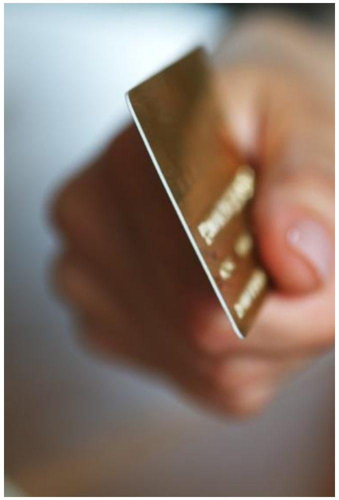
Big Data, Big Analytics: Emerging Business ...
› Michael Minelli
★★★★☆ (9)
Hardcover
$34.15

Big Data: A Revolution That Will Transform ...
Viktor Mayer-Schonberger
★★★★½ (114)
Hardcover
$20.03

Too Big to Ignore: The Business Case for Big ...
› Phil Simon
★★★★★ (20)
Hardcover
$31.65

Analytics Tasks

- Profiling (also known as behavior description) is used to characterize the typical behavior of an individual, group, or population.
  - A sample profiling question is: *What is the typical cellphone usage of this customer segment?*



Analytics Tasks

- **Link Prediction is used to predict connections between data items, usually by suggesting that a link should exist, and possibly also estimating the strength of the link.**
    - ***Since you and Karen have 10 friends in common, maybe you'd like to be Karen's friend?***

Analytics Tasks

- Data Reduction is used to replace a large set of data with a smaller set of data that contains much of the important information in the larger set.

Analytics Tasks

- Causal modeling attempts to help us understand what events or actions actually influence others.
  - For example, consider that we use predictive modeling to target advertisements to consumers, and we observe that indeed the targeted consumers purchase at a higher rate subsequent to having been targeted.

Analytics Tasks

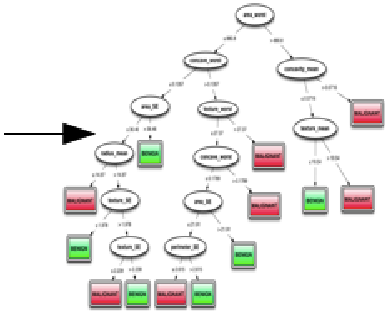
Historical data

Training data have all values specified

X	Y	Z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Data mining

Model



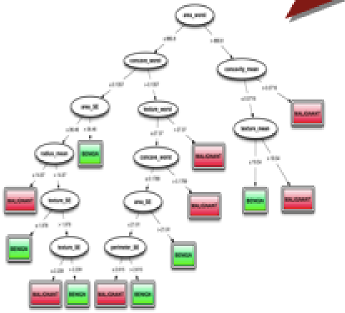
Mining

Model is deployed

Use

(training)

Model



New data item

X	Y	Z	class
30	False	Red	?

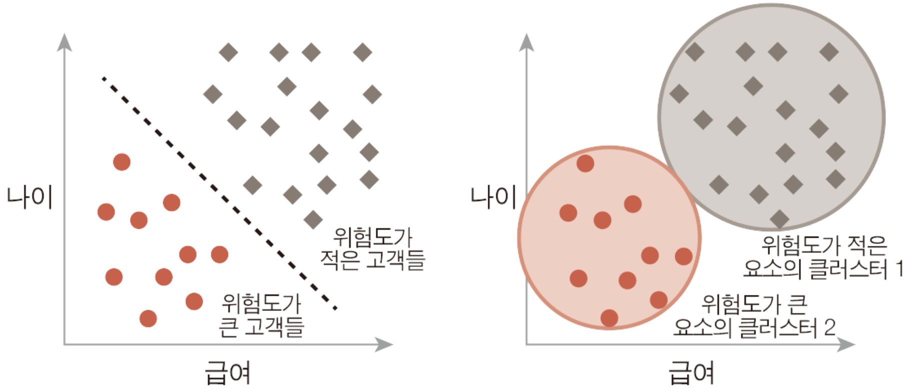
New data item has class value **unknown** (e.g., *will customer accept?*)

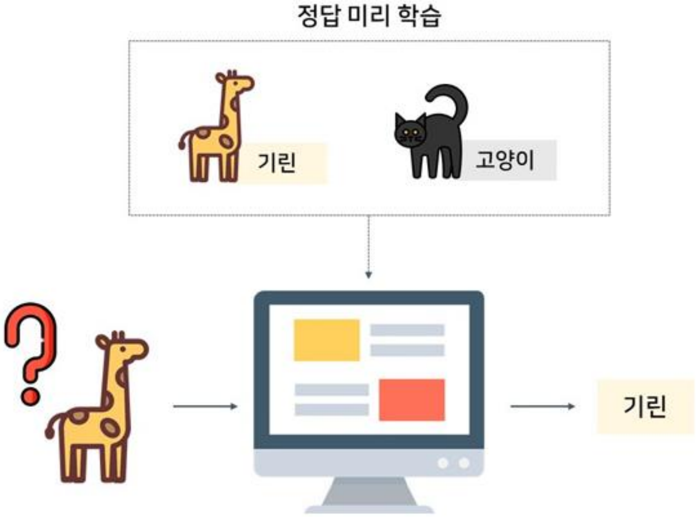
Class: accepted,
Probability: 0.88

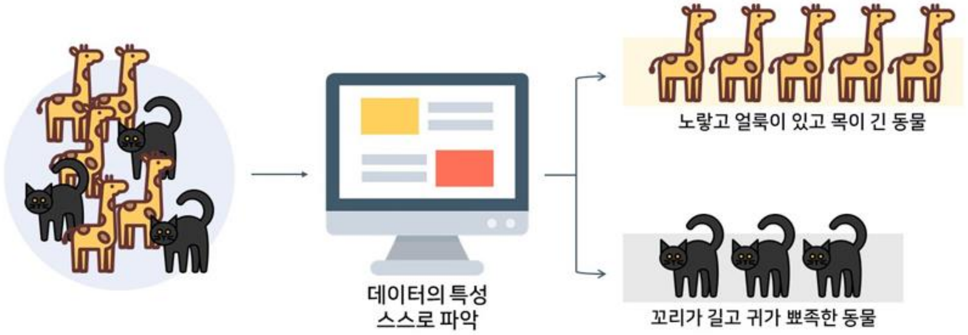
Image from "Data Science for Business", Provost and Fawcett, 2013

분류와 클러스터링

분류 VS 클러스터링





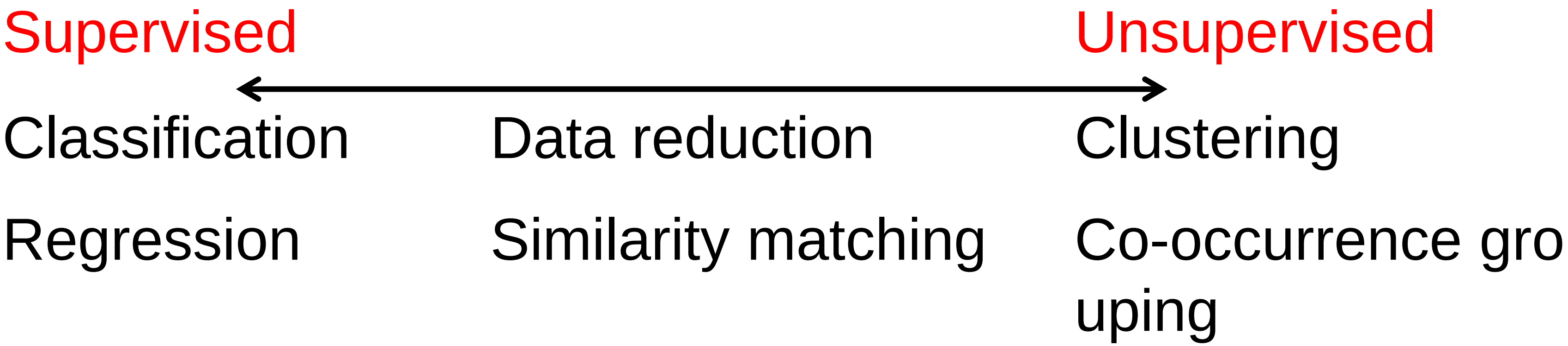


자료원: https://www.samsungsds.com/kr/insights/Generative-adversarial-network-AI.html

감독 vs. 무감독 학습

Key Questions:

- Is there a specific target variable?
- (Are data on this target variable available?)



Supervised

Unsupervised

Classification

Data reduction

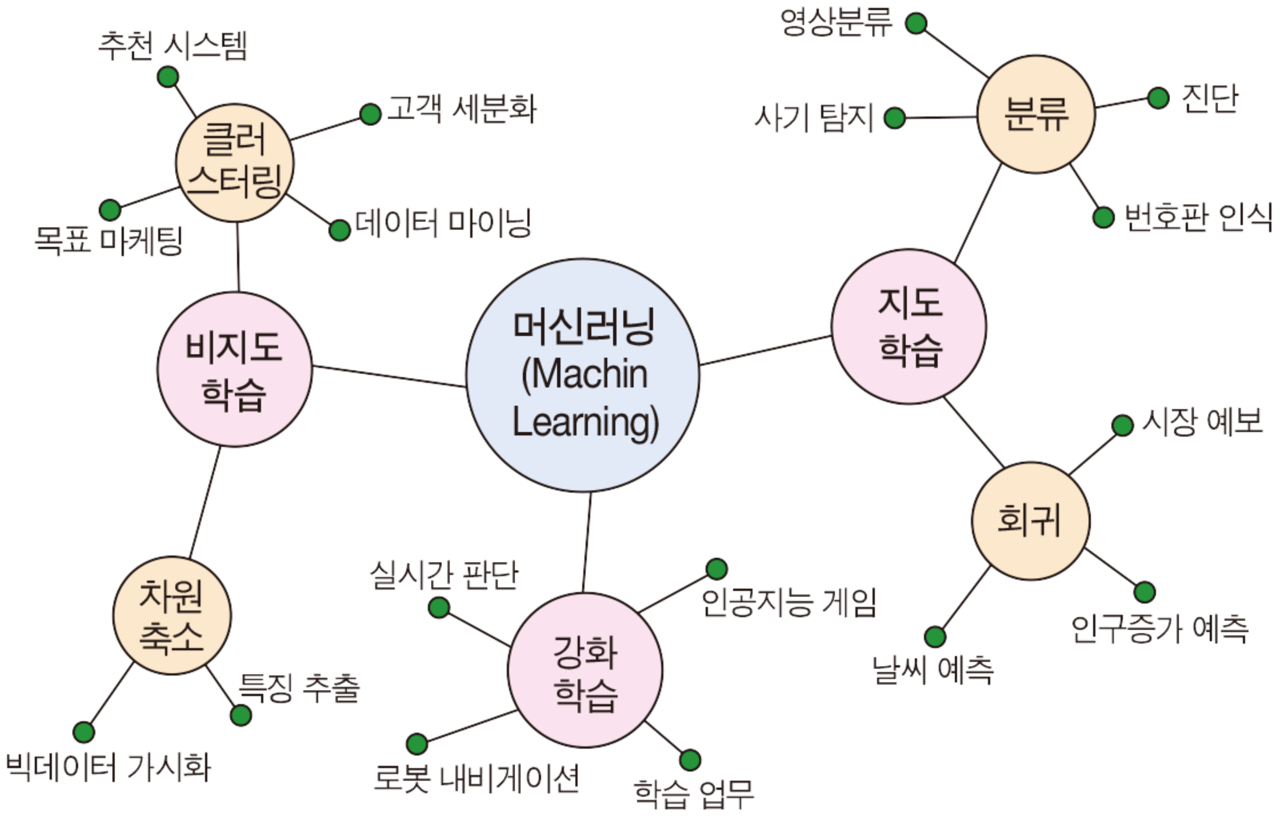
Clustering

Regression

Similarity matching

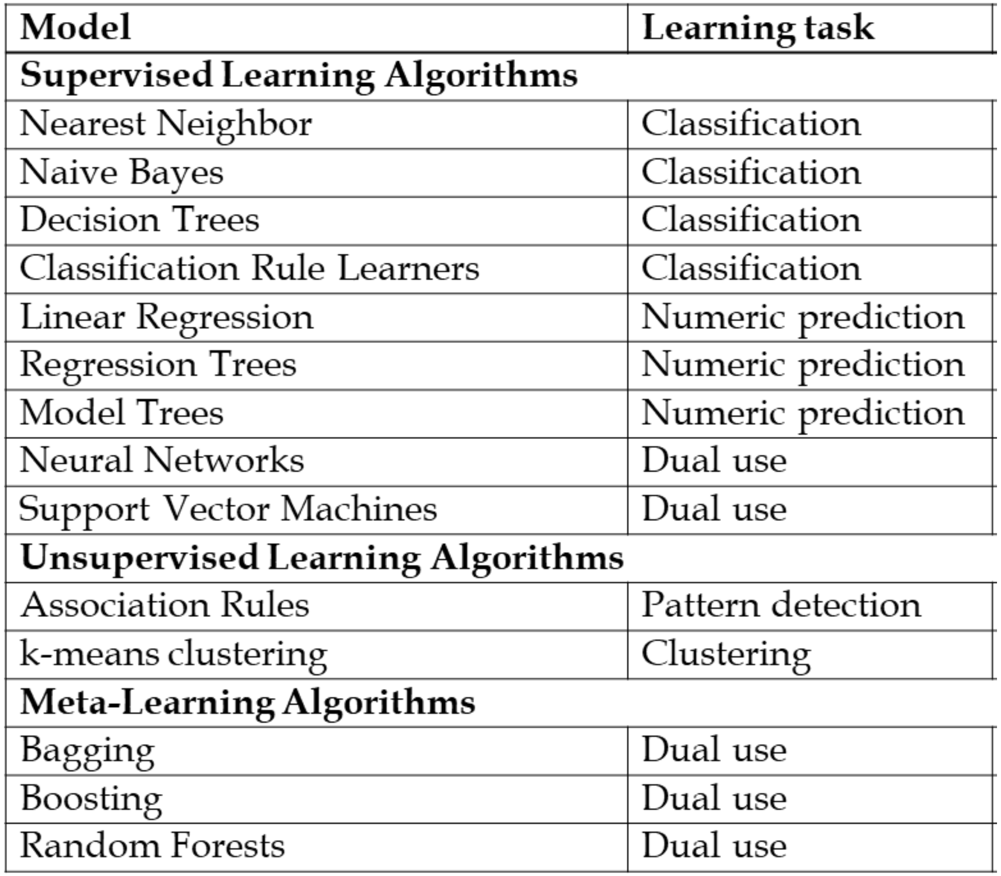
Co-occurrence gro
uping

분석기법의 분류와 주요 활용분야



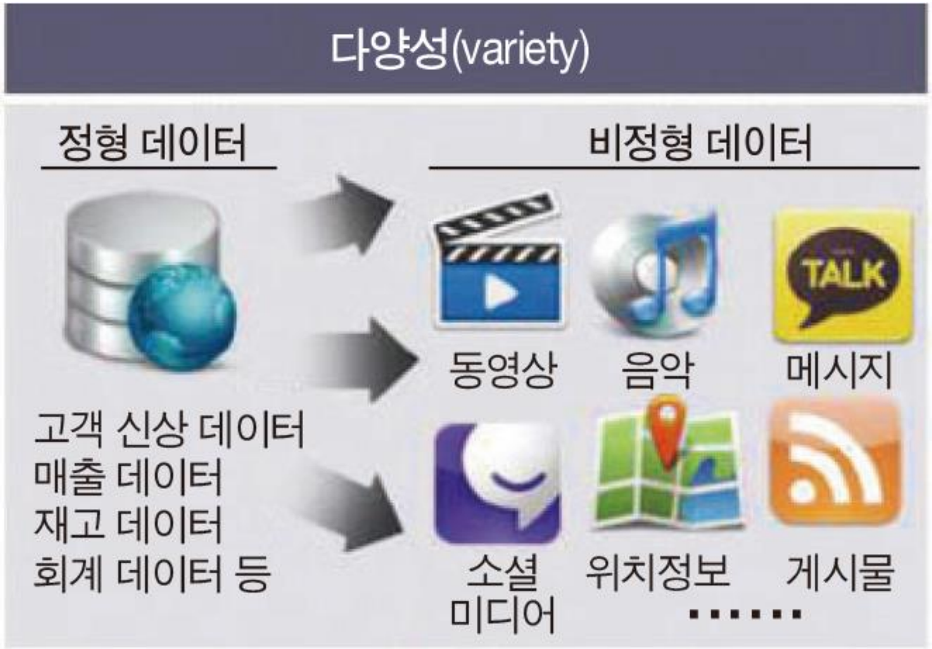
자료원: 처음 만나는 인공지능, 김대수, 2020, 생능출판

분석 알고리즘의 선택



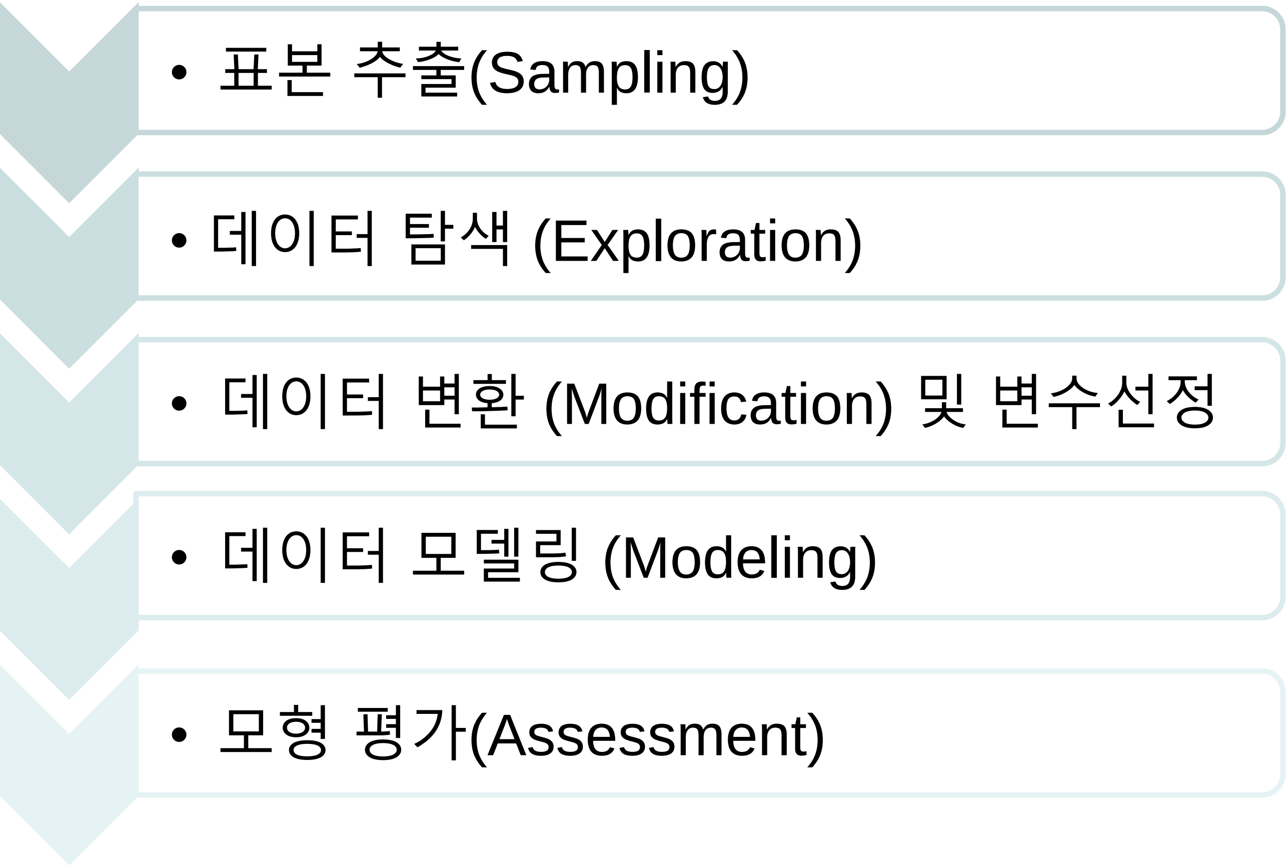
Model	Learning task
Supervised Learning Algorithms	
Nearest Neighbor	Classification
Naive Bayes	Classification
Decision Trees	Classification
Classification Rule Learners	Classification
Linear Regression	Numeric prediction
Regression Trees	Numeric prediction
Model Trees	Numeric prediction
Neural Networks	Dual use
Support Vector Machines	Dual use
Unsupervised Learning Algorithms	
Association Rules	Pattern detection
k-means clustering	Clustering
Meta-Learning Algorithms	
Bagging	Dual use
Boosting	Dual use
Random Forests	Dual use

정형 vs. 비정형 데이터



자료원: 처음 만나는 인공지능, 김대수, 2020, 생능출판

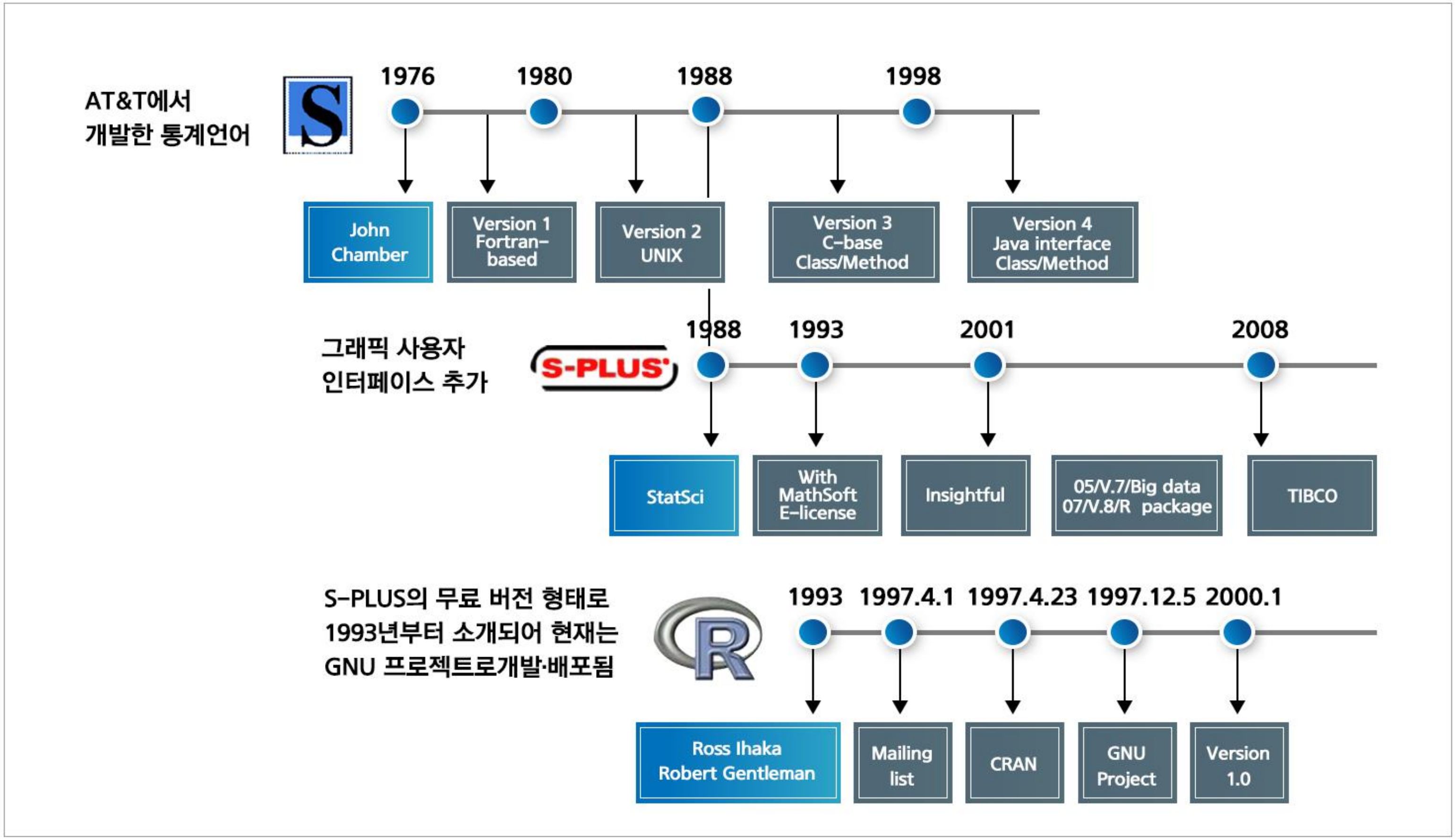
데이터분석 프로세스



- 표본 추출(Sampling)
- 데이터 탐색 (Exploration)
- 데이터 변환 (Modification) 및 변수선정
- 데이터 모델링 (Modeling)
- 모형 평가(Assessment)

R의 탄생과 발전과정

- R은 데이터 분석을 위한 통계분석 기법과 알고리즘, 시각화 기능을 지원하는 오픈 소프트웨어 도구임



AT&T에서 개발한 통계언어

1976 – John Chamber
1980 – Version 1 Fortran-based
1988 – Version 2 UNIX
Version 3 C-base Class/Method
1998 – Version 4 Java interface Class/Method

그래픽 사용자 인터페이스 추가

1988 – StatSci
1993 – With MathSoft E-license
2001 – Insightful
05/V.7/Big data 07/V.8/R package
2008 – TIBCO

S-PLUS의 무료 버전 형태로 1993년부터 소개되어 현재는 GNU 프로젝트로개발·배포됨

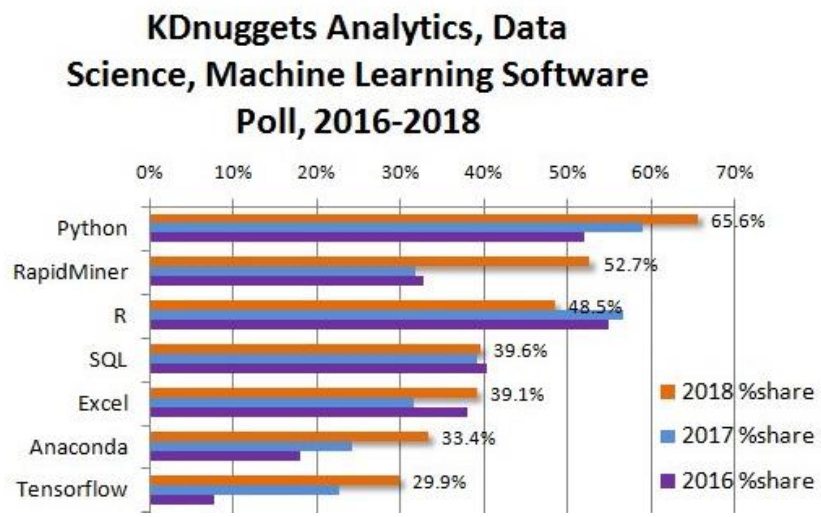
1993 – Ross Ihaka Robert Gentleman
1997.4.1 – Mailing list
1997.4.23 – CRAN
1997.12.5 – GNU Project
2000.1 – Version 1.0

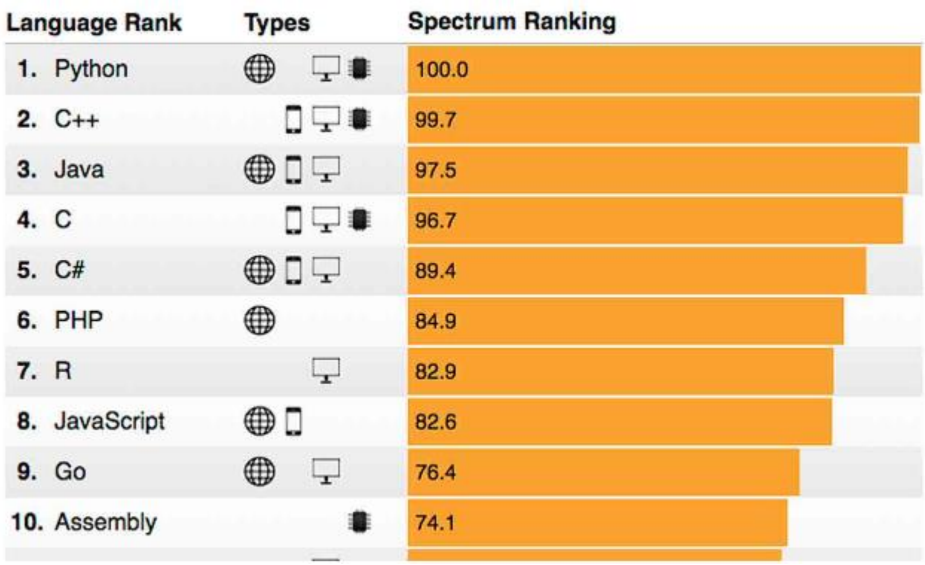
* 출처 : 데이터 오픈소스 플랫폼 기술세미나(2012)

R 소개

출처 : Kdnuggets 2018, IEEE 2018

❖ R 소개





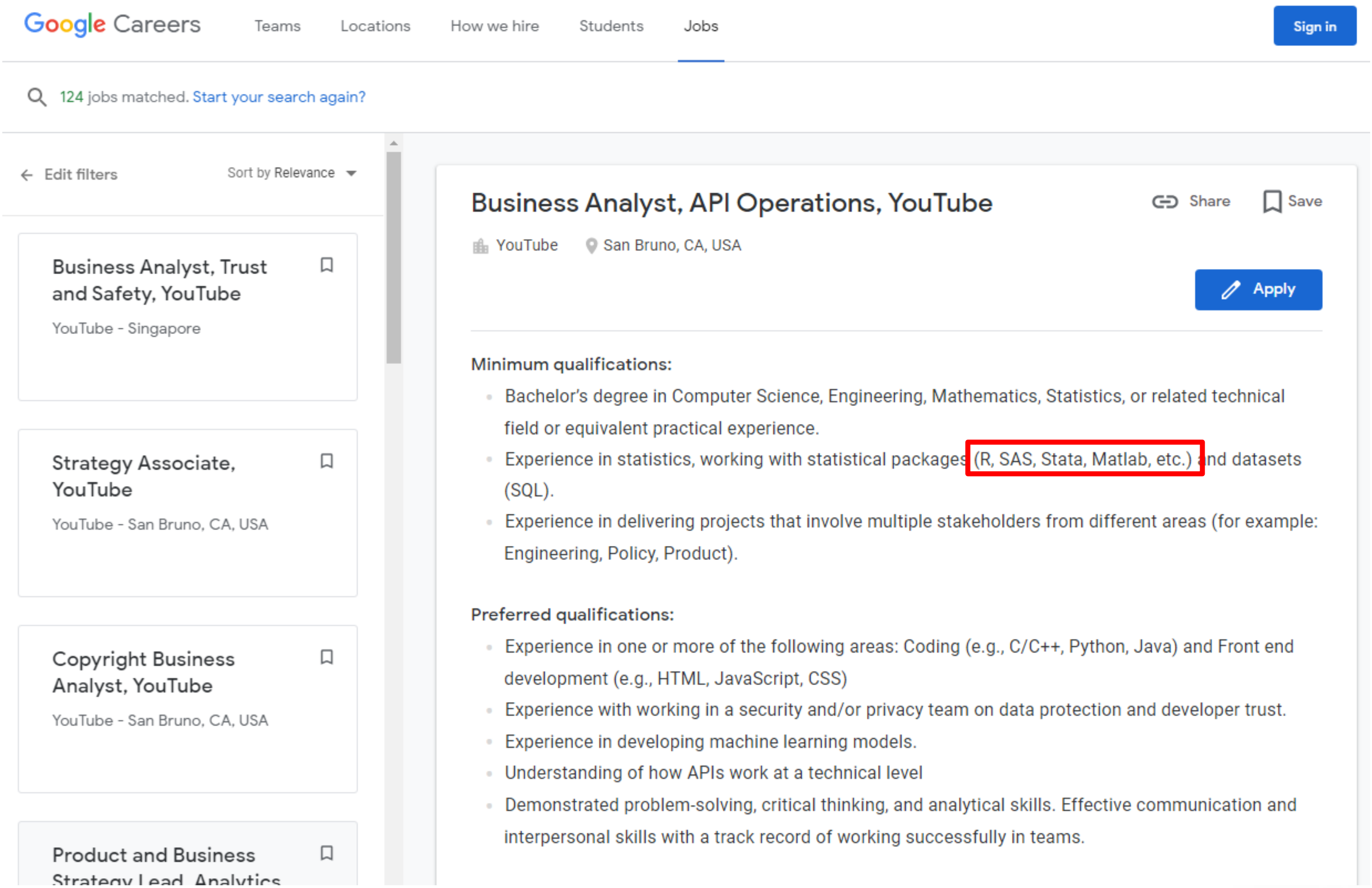
Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥 ▣	100.0
2. C++	📱 🖥 ▣	99.7
3. Java	🌐 📱 🖥	97.5
4. C	📱 🖥 ▣	96.7
5. C#	🌐 📱 🖥	89.4
6. PHP	🌐	84.9
7. R	🖥	82.9
8. JavaScript	🌐 📱	82.6
9. Go	🌐 🖥	76.4
10. Assembly	▣	74.1

- **R**은 다른 언어보다 분석 하기 자유롭고, 내게 알맞게 코딩할 수 있다는 장점 때문에 많은 사람들이 사용한다.

R 소개

❖ R 소개

출처 : Google careers (2020.03.기준)



124 jobs matched. Start your search again?

Business Analyst, Trust and Safety, YouTube
YouTube - Singapore

Strategy Associate, YouTube
YouTube - San Bruno, CA, USA

Copyright Business Analyst, YouTube
YouTube - San Bruno, CA, USA

Product and Business Strategy Lead, Analytics

Business Analyst, API Operations, YouTube

YouTube San Bruno, CA, USA

Minimum qualifications:

- Bachelor's degree in Computer Science, Engineering, Mathematics, Statistics, or related technical field or equivalent practical experience.
- Experience in statistics, working with statistical packages (R, SAS, Stata, Matlab, etc.) and datasets (SQL).
- Experience in delivering projects that involve multiple stakeholders from different areas (for example: Engineering, Policy, Product).

Preferred qualifications:

- Experience in one or more of the following areas: Coding (e.g., C/C++, Python, Java) and Front end development (e.g., HTML, JavaScript, CSS)
- Experience with working in a security and/or privacy team on data protection and developer trust.
- Experience in developing machine learning models.
- Understanding of how APIs work at a technical level
- Demonstrated problem-solving, critical thinking, and analytical skills. Effective communication and interpersonal skills with a track record of working successfully in teams.

준비해야할 분석 환경

R 콘솔프로그램

- R Foundation에서 배포하는 R 기본 패키지
- R GUI 콘솔창을 통해서 필요한 패키지를 다운·설치하고, 다양한 분석작업을 수행할 수 있음

RStudio 통합분석도구

- R 콘솔에 비해 보다 편리한 IDE (Integrated Development Environment) 라는 통합분석개발환경을 제공함
- 4개로 분할된 레이아웃 창을 통해서 R스크립트 작성, R코드 실행결과 확인, 메모리 상황관리, 그래프 구현, 패키지·도움말·파일 관리 등을 편리하게 사용

자바실행환경

- R패키지 중에서 자바언어로 개발된 패키지 실행을 위한 프로그램
- 오라클의 자바다운로드 사이트에서 자바실행환경(JRE: Java Runtime Environment)를 다운받아 설치함

* 출처: 한국정보화진흥원 빅데이터 센터