

필요한 관련 서적 ...



데이터 과학자의 역할

# 빅 데이터의 수집, 분석, 활용

## 학습목표

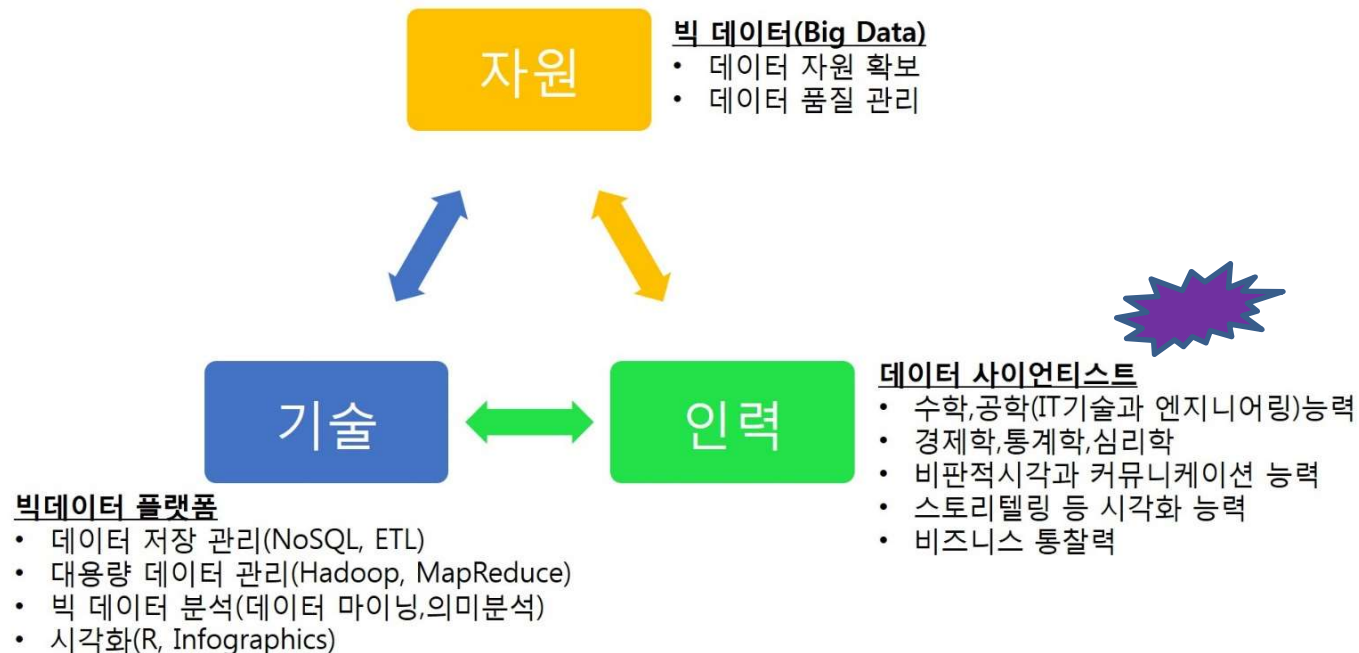
1. 데이터 과학자의 역할
2. 데이터 마이닝의 주요 단계
3. 분석을 위한 샘플 데이터
4. 데이터 관리의 중요성

# 1. 데이터 과학자

## ◆ 빅데이터 시대

- ✓ 빠른 속도(velocity)로 생산되는 다양(variety)한 대용량(volume) 데이터를 체계적으로 관리하고 분석하는 기술, 능력이 중요한 시대

## ◆ 빅데이터 시대의 주요 3 요소

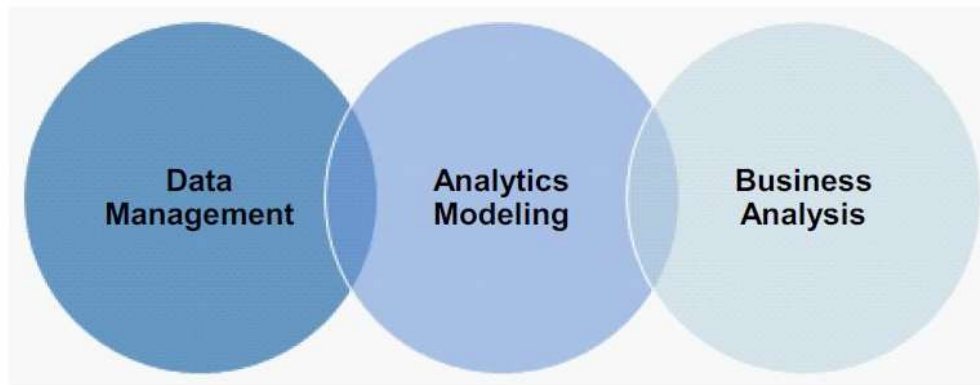


# 1. 데이터 과학자

## ◆ 데이터 과학자

- ✓ 빅 데이터 관련 기술을 이용하여 데이터(data)로부터 중요한 정보 (information), 지식(knowledge)을 **발굴**하는 전문가

## ◆ 데이터 과학자의 필요 역량



### Understand Data

- Integrate
- Manipulate
- QA
- Prep.

### Know Analytics

- Appropriate techniques
- Interpret data and diagnose models
- Meet business requirements

### Focus on the Business

- Goals
- Constraints
- Decisions
- Communication of results

QA = quality assurance



## 2. 데이터 마이닝

### ◆ 데이터 마이닝의 정의

- ✓ 매우 큰 데이터 더미로부터 사전에 알려지지 않은 유용한 정보를 추출하는 지식 발견 방법

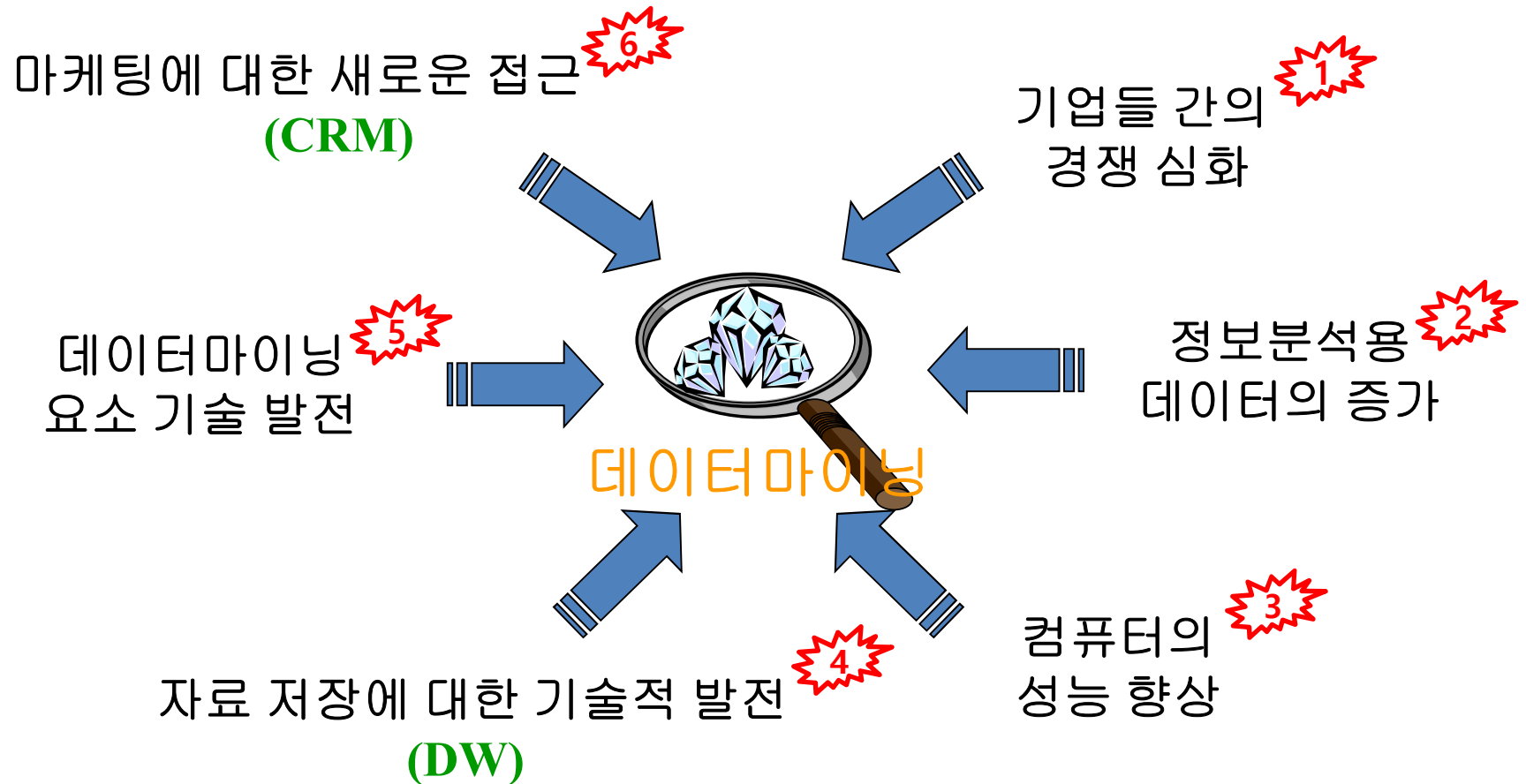


### ◆ 데이터 마이닝의 특성

- ✓ 비즈니스 문제의 이해에서부터 정보기술을 적용하는 포괄적인 과정
- ✓ 데이터 관리 기술, 통계, 머신러닝/딥러닝, 정보검색 등 다양한 지식 필요
- ✓ KDD(Knowledge Discovery in Database)

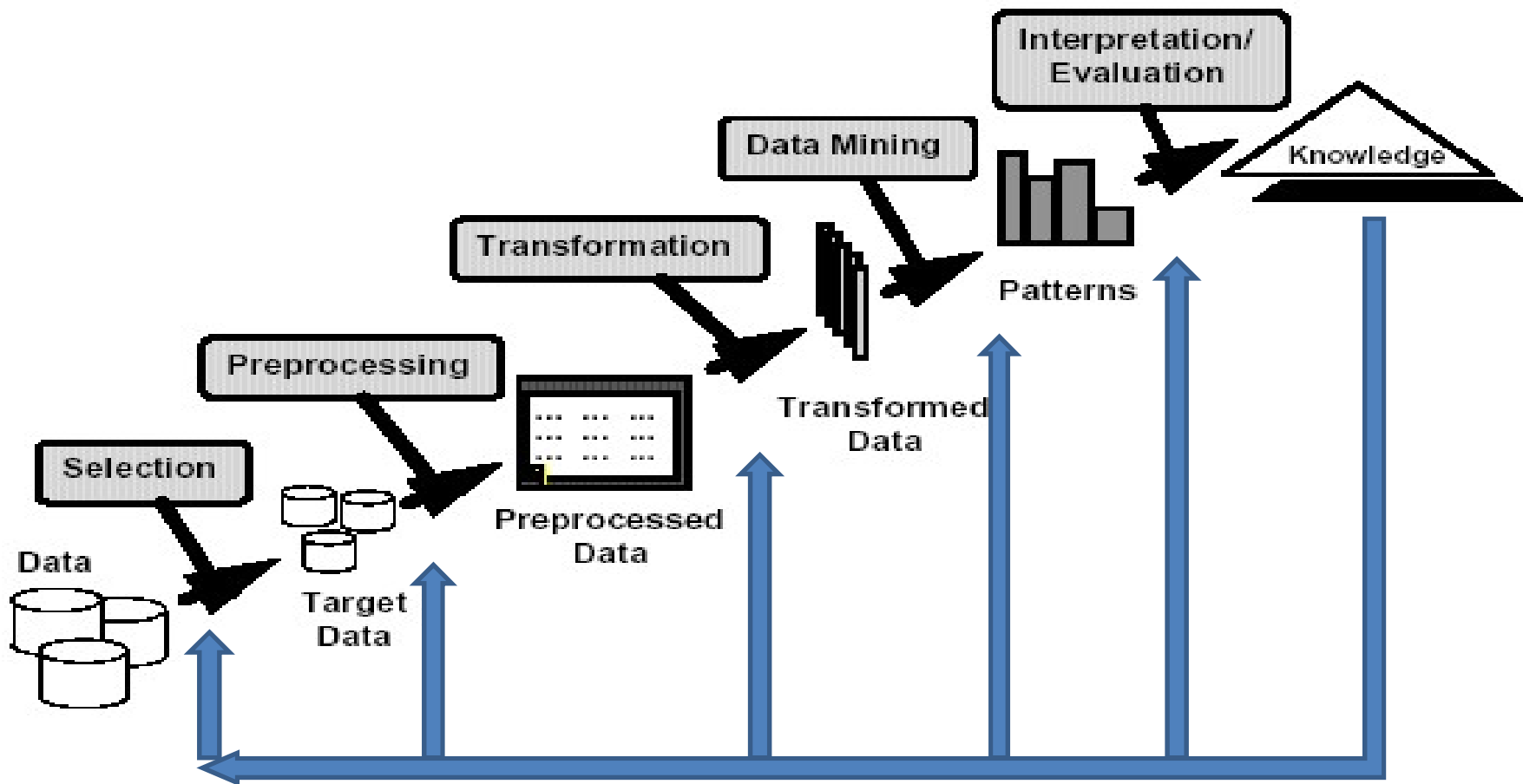
## 2. 데이터 마이닝

### ◆ 데이터 마이닝의 발전 배경



## 2. 데이터 마이닝

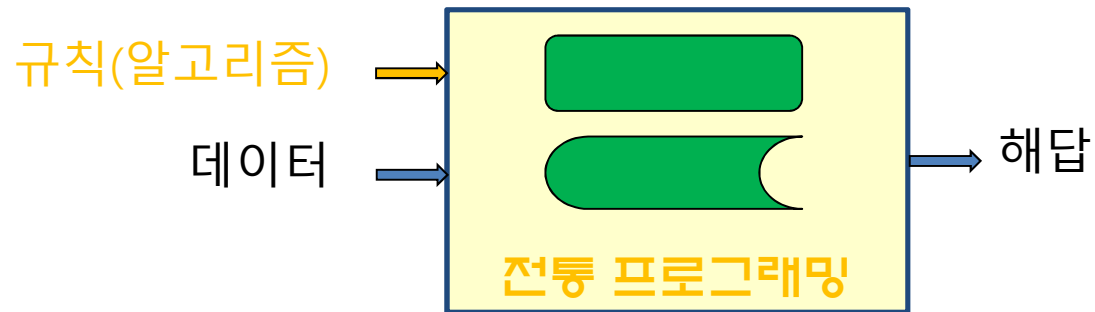
### ◆ 데이터 마이닝의 절차



## 2. 데이터 마이닝

### ◆ 문제 해결 기법의 비교

✓ 전통 프로그래밍 패러다임



✓ 머신 러닝 패러다임




- 찾은 규칙을 새로운 데이터에 적용하여 창의적 해답 구함



### 3. 분석을 위한 샘플 데이터

## ◆ 머신 러닝을 위한 샘플 데이터 저장소

- ✓ 데이터 마이닝의 중요 과정인 머신 러닝을 위한 대표적인 데이터를 모은 저장소 (<http://archive.ics.uci.edu/ml/datasets.php>)













**Machine Learning Repository**  
Center for Machine Learning and Intelligent Systems

[About](#)
[Citation Policy](#)
[Donate a Data Set](#)
[Contact](#)

[View ALL Data Sets](#)

Browse Through: **559 Data Sets**

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 <a href="#">Abalone</a>	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
 <a href="#">Adult</a>	Multivariate	Classification	Categorical, Integer	48842	14	1996
 <a href="#">Annealing</a>	Multivariate	Classification	Categorical, Integer, Real	798	38	
 <a href="#">Anonymous Microsoft Web Data</a>		Recommender-Systems	Categorical	37711	294	1998
 <a href="#">Arrhythmia</a>	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
 <a href="#">Artificial Characters</a>	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
 <a href="#">Audiology (Original)</a>	Multivariate	Classification	Categorical	226		1987
 <a href="#">Audiology (Standardized)</a>	Multivariate	Classification	Categorical	226	69	1992
 <a href="#">Auto MPG</a>	Multivariate	Regression	Categorical, Real	398	8	1993
 <a href="#">Automobile</a>	Multivariate	Regression	Categorical, Integer, Real	205	26	1987

### 3. 분석을 위한 샘플 데이터

#### ◆ 온라인 판매점 데이터

- ✓ 2010년 1월 12일부터 2011년 9월 12일까지의 영국의 온라인 판매점의 거래 데이터 모음

#### Online Retail Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

<b>Data Set Characteristics:</b>	Multivariate, Sequential, Time-Series	<b>Number of Instances:</b>	541909	<b>Area:</b>	Business
<b>Attribute Characteristics:</b>	Integer, Real	<b>Number of Attributes:</b>	8	<b>Date Donated</b>	2015-11-06
<b>Associated Tasks:</b>	Classification, Clustering	<b>Missing Values?</b>	N/A	<b>Number of Web Hits:</b>	542451

### 3. 분석을 위한 샘플 데이터

---

#### ◆ 온라인 판매점 데이터의 추가 정보

- ✓ 데이터를 이해하기 위한 정보 및 속성 구성 정보도 함께 제공

#### Data Set Information:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

#### Attribute Information:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.



### 3. 분석을 위한 샘플 데이터

#### ◆ 온라인 판매점 데이터의 샘플 (전체 541,909개 중 일부)

	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	2010-12-01 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	2010-12-01 8:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 8:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	2010-12-01 8:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	2010-12-01 8:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	2010-12-01 8:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	2010-12-01 8:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	2010-12-01 8:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	2010-12-01 8:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	2010-12-01 8:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	2010-12-01 8:34	4.95	13047	United Kingdom

## 4. 데이터 관리의 중요성

---

### ◆ 제공된 샘플 데이터의 외형적 특성

- ✓ 분석을 위해 데이터 수집, 코드화, 테이블 형식으로 정리
- ✓ 현재 제공된 테이블에는 정보 불일치 등 존재
- ✓ 업체에서는 거래 데이터베이스 형태로 관리
- ✓ 거래 데이터베이스에는 상품 정보, 고객 정보, 거래 정보 등으로 구성

### ◆ 데이터베이스로부터 분석용 데이터 취득을 위한 필요 지식

- ✓ 데이터베이스 기본 지식
- ✓ 사용 데이터베이스 관리 시스템에 대한 지식
- ✓ 데이터베이스 언어 표준 SQL
- ✓ 분석이 용이한 형태의 데이터 전처리 및 변환