

# 1 데이터 분석 단계별 목적 이해하기

[Open in Colab](#)

"데이터분석이 뭐길래?"

- 사람이 비이성적, 비효율적으로 처리하는 방식을 정량적, 체계적, 합리적 의사결정 하도록 다양한 도구들을 활용하는 것

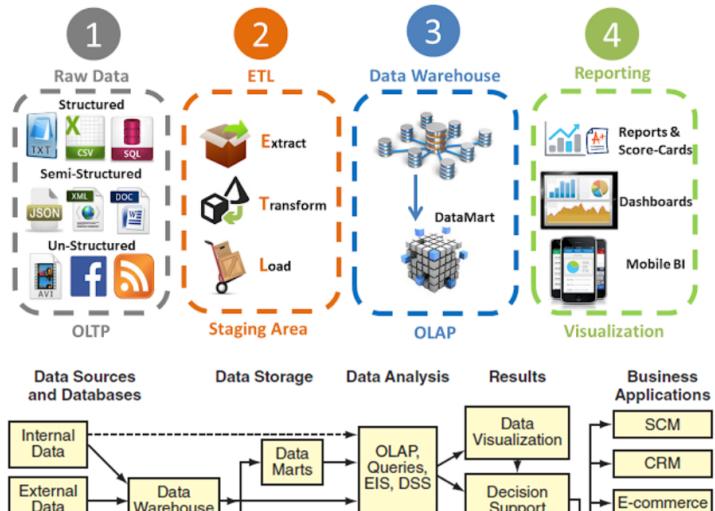
비즈니스/사람 일상 반복	사람	분석	분석도구
문제/호기심	어떤 현상을 보고	어떤 현상을 보고 알고 싶은 것을 정의	
관찰	왜 그런일이 벌어지는지 경험/직감/인터넷으로 짧게 공들이고	왜 그런일이 벌어지는지 정량적 및 체계적으로 대량 수집	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등
주장/의사결정	그 이유를 대충 근거를 들어 주장한 후	통계적 검정을 통해 주장한 후	통계추론, 확률적 검정
검증	나와 의견이 같으면 좋은 사람 아니면 유탐	주장이 충분히 합리적이면, 알고리즘을 이용하여 그 현상을 재현해내며 주장을 설명	확률통계, 기계학습, 딥러닝, 지도학습, 비지도학습, 강화학습, 검증지표 등
해결	???	알고 싶은 것을 해결하고 미래에도 사용	

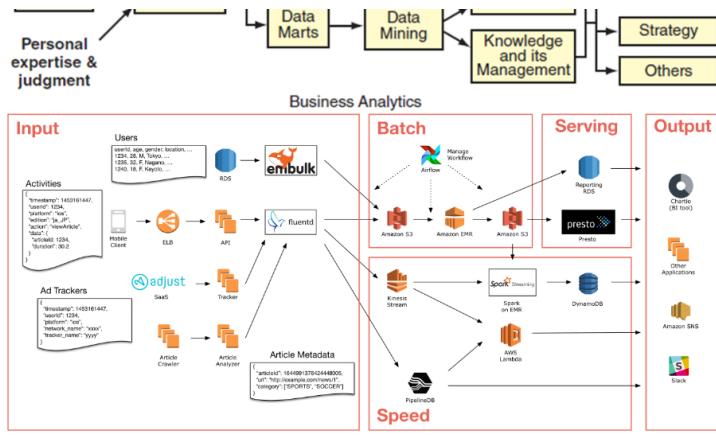
## 1.1 데이터 사이언티스트? 애널리스트? 엔지니어? 고객?

### 1) 분석 단계별 역할:

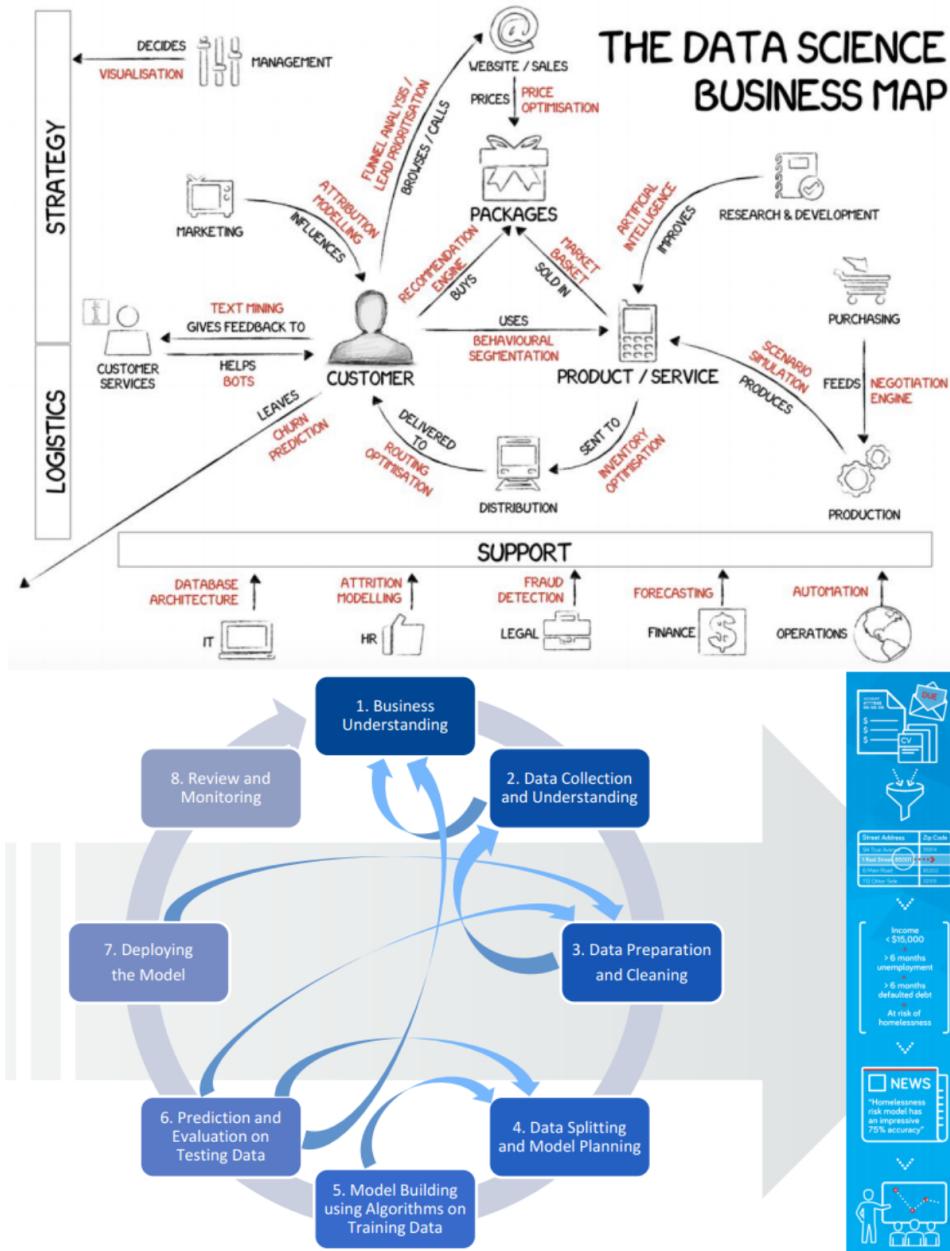
분석 단계	분석 역할	분석 도구
(0) 문제 정의:	분석할 현실 문제를 데이터문제로 변환 및 기획(Planning)	
(1) 데이터 수집:	소스별 데이터 추출 및 저장>Loading)	데이터집계
(2) 데이터 전처리:	불이기(Merge)  지우기(Remove) 및 채우기(Fill)  필터(Filter) 및 변경하기(Transform)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등
	기초통계(Descriptive Statistics)	기술적 분석, 데이터 집계, 상관분석 등
	시각화(Visualization/Dashboard)	데이터 시각화, 상관분석 등
(3) 데이터 정리:	데이터 한곳에 담기(Data Warehouse)  바꾸기 및 정리(Data Mart)  데이터 분리(Data Split)	데이터집계
		데이터집계
(4) 모델링 및 검증:	기초통계(Descriptive Statistics)  모델링(Algorithm)  검증(Evaluation)  에러분석(Error Analysis)	기술적 분석, 데이터 집계, 상관분석 등  통계추론, 확률적 검정, 지도학습, 비지도학습, 강화학습, 기계학습, 딥러닝 등  성능측정, 검증지표 등  잔차추정, 검증지표 등
(5) 결과 활용:	시각화(Visualization/Dashboard)  의사결정(Decision Support)  지식화(Knowledge) 및 공유(Reporting)	데이터 시각화, 상관분석 등  의사결정(Decision Support)  지식화(Knowledge) 및 공유(Reporting)

### 2) 분석 단계 시각화:





## 1.2 데이터 분석단계의 현실



1) 분석현실 요약:





“데이터 분석 설계(1단계)는 모델링(2단계) 보다 훨씬 중요”

“분석 종료(3단계)는 또 다른 새로운 시작”

• 예시1:

퇴사할 사람 찾기 vs 입사할 사람 찾기

• 예시2:

삼성전자 주식을 사야 할까 말아야 할까? vs 내일 삼성전자 주식은 얼마 일까?

## 2) 분석단계별 현실예시:

- 아이폰 고객은 왜 갤럭시 고객보다 충성도가 높지? 라고 질문을 다 듣기도/이해하기도/생각하기도 전에 프로젝트가 시작
- AI를 활용해서 생산공정의 이상을 조기 탐지하고 비용을 줄여봐~ 라고 질문을 다 듣기도/이해하기도/생각하기도 전에 프로젝트가 시작됩니다
- 타겟 마케팅을 하기 위해 누구한테 프로모션을 해야하는지 알려줘~ 라고 질문을 다 듣기도/이해하기도/생각하기도 전에 프로젝트가 시작됩니다

(0) 문제 정의:

분석 단계	분석 역할	분석 도구
<b>(0) 문제 정의:</b> 분석할 현실 문제를 데이터문제로 변환 및 기획(Planning)		

- 무엇을 분석할지 각자 생각이 모두 다름 (솔직히 아무도 모른다)
- 무엇을 분석할지 모르지만 일단 도구(R? Python? 플랫폼? 아마존? 외주?)부터 마련
- 무엇을 분석할지 모르지만 완료일정과 계획이 준비
- 어쨌건 있다고 생각하고 시작

⇒ 이미 데이터분석 프로젝트는 착수 보고 완료

(1) 데이터 수집:

분석 단계	분석 역할	분석 도구
<b>(1) 데이터 수집:</b> 소스별 데이터 추출 및 저장>Loading)		

- 데이터가 PC에 있는줄 알았는데 A4용지에 있어서 시킬 사람 탐색
- 데이터를 구했는데 빅데이터는 아니고 그냥 엑셀 파일 몇개
- 데이터 파일을 열었더니 다 빙판이고 딱봐도 오타 투성
- 데이터가 충분한지 아무도 모르지만 (있는줄/충분한줄 알았는데) 어쨌든 진행

⇒ 빅데이터 기반으로 한 데이터 수집 이 완료 보고

(2) 데이터 전처리:

분석 단계	분석 역할	분석 도구
<b>(2) 데이터 전처리:</b> 병이기(Merge)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
지우기(Remove) 및 채우기(Fill)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
필터(Filter) 및 변경하기(Transform)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
기초통계(Descriptive Statistics)	기술적 분석, 데이터 집계, 상관분석 등	
시각화(Visualization/Dashboard)	데이터 시각화, 상관분석 등	

- 무엇을 분석할지 모르고 데이터는 없지만 전처리 돌입
- 일단 이상해 보이는 데이터를 다 삭제 (남는게 없다..)
- 임의로 데이터를 채움/수정 (어짜피 아무도 모르니까..)
- 할게 많을 줄 알았는데 별로 할게 없음을 깨달음

⇒ 데이터가 무결점으로 잘 준비 완료 보고

(3) 데이터 정리:

분석 단계	분석 역할	분석 도구
<b>(3) 데이터 정리:</b> 데이터 한곳에 담기(Data Warehouse)		
바꾸기 및 정리(Data Mart)	데이터집계	
데이터 분리(Data Split)	데이터집계	

- (대부분 개인PC로 충분 하겠지만) 일단 서버/플랫폼에 데이터를 업로드
- (뭔가 중요한걸 해야 할 것 같은데..) 서버/플랫폼 사용정도 체크

⇒ 데이터 플랫폼에 데이터가 이관 되고 곧 분석 착수 보고

#### (4) 모델링 및 검증:

분석 단계	분석 역할	분석 도구
(4) 모델링 및 검증:	기초통계(Descriptive Statistics) 모델링(Algorithm) 검증(Evaluation)	기술적 분석, 데이터 집계, 상관분석 등 통계추론, 확률적 검정, 지도학습, 비지도학습, 강화학습, 기계학습, 딥러닝 등 성능측정, 검증지표 등
		예러분석(Error Analysis)
		잔차추정, 검증지표 등

- 무엇을 분석하고 무슨 데이터를 사용해야 되는지 모르지만 분석을 시작
- 기초통계는 사람수? 클릭수? 등 횟수(count)면 충분
- 도구/사람을(R? Python? 플랫폼? 아마존? 외주?) 조아서 제일 최신 알고리즘을 적용
- (뭔가 안되면..) 우선 1차 화귀분석? 상관관계? 어디서 들어본걸 실행 후 보고위한 그림부터 생성
- (뭔가 중요한 단계인것 같은데..) 더 이상 할 수 있는게 없을 때 깨달음

⇒ 분석이 완료되어 인사이트가 곧 쓰아질 것이라 보고

#### (5) 결과 활용:

분석 단계	분석 역할	분석 도구
(5) 결과 활용:	시각화(Visualization/Dashboard) 의사결정(Decision Support)	데이터 시각화, 상관분석 등
	지식화(Knowledge) 및 공유(Reporting)	

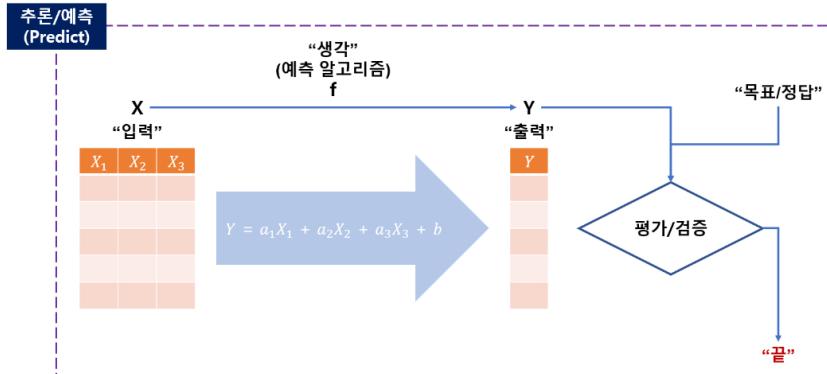
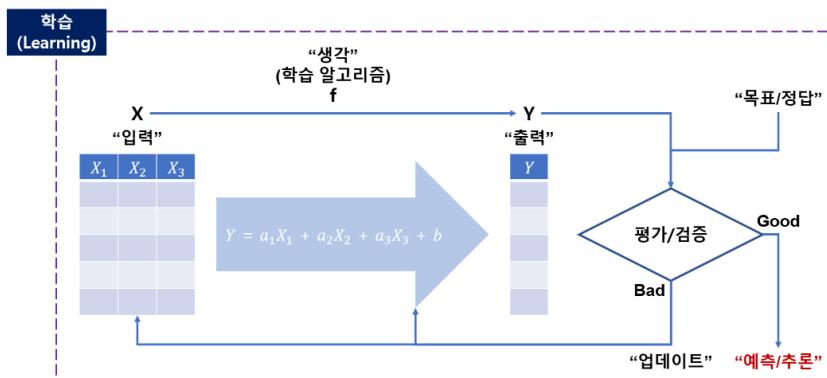
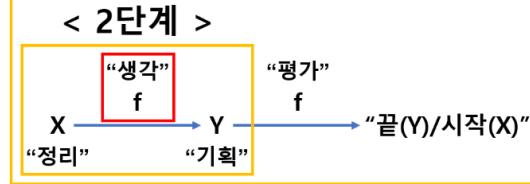
- 무엇을 분석하고 무슨 데이터를 사용하고 무슨 결과가 있는지 모르겠지만 결과를 정리
- (완료일정이 내일이라 퇴사/퇴학이 필요한게 아닌지 잠시 오지 않음)
- (신기 한건 모든 단계는 작동/구현 되었고 각 단계 개발자들은 성과 보고)
- (Kaggle과 현실은 다른데 알게됨)

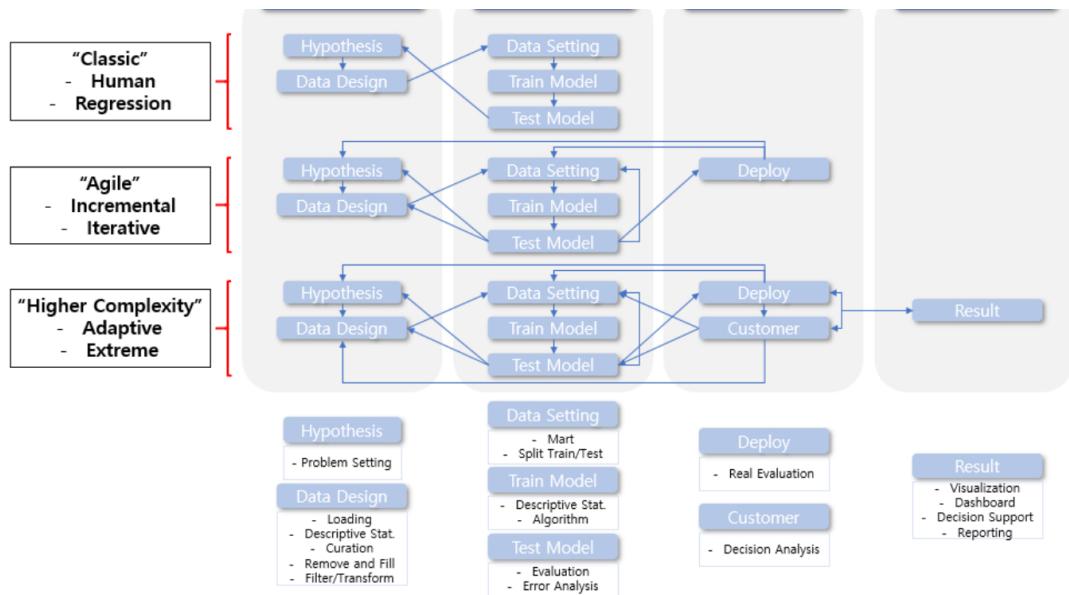
### 1.3 데이터 사이언티스트 단계별 방향

#### < 1단계 >



#### < 3단계 >





#### (0) 문제 정의:

분석 단계	분석 역할	분석 도구
<b>(0) 문제 정의:</b> 분석할 현실 문제를 데이터문제로 변환 및 기획(Planning)		

⇒ 이미 데이터분석 프로젝트는 착수 보고 완료

"문제정의가 없으면 분석은 시작할 필요가 없기 때문에 문제정의에 많은 고민을 해야 한다"

"문제정의에 모든 구성원이 동의 할 수 있도록 끊임없이 커뮤니케이션 해야 한다"

"1회성이 문제정의가 아니라 필요시 끊임없이 진화/변경 시켜야 한다"

#### (1) 데이터 수집:

분석 단계	분석 역할	분석 도구
<b>(1) 데이터 수집:</b> 소스별 데이터 추출 및 저장(Load) 데이터집계		

⇒ 빅데이터 기반으로 한 데이터 수집 이 완료 보고

"데이터가 없으면 분석은 시작할 필요가 없고, 알고리즘/기술보다 데이터 수집부터 시작 해야 한다"

"알고리즘/기술보다 데이터 수집부터 시작 해야 한다"

"데이터는 많을수록 좋지만 양보다(Row) 질(Column)을 늘려야 분석 의미가 생긴다"

"정답 후보(보기)부터 만들고 분석을 시작해야 하며 (어떤 연령이 TV를 보는지 알고 싶은데 데이터에 연령이 없으면 불가능에 가까움), 보기가 데이터에 없으면 문제정의부터 새롭게 수정 해야 한다"

#### (2) 데이터 전처리:

분석 단계	분석 역할	분석 도구
<b>(2) 데이터 전처리:</b> 불이기(Merge)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
지우기(Remove) 및 채우기(Fill)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
필터(Filter) 및 변경하기(Transform)	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등	
기초통계(Descriptive Statistics)	기술적 분석, 데이터 집계, 상관분석 등	
시각화(Visualization/Dashboard)	데이터 시각화, 상관분석 등	

⇒ 데이터가 무결점으로 잘 준비 완료 보고

- **Merge 목적:** 각 소스별 데이터를 하나의 Database로 병합
- **Remove & Fill 목적:** 데이터 오류를 제거하거나 비어있는 데이터를 채움
- **Filter 목적:** 분석범위에 관련된 보기(Feature)들만을 추려냄
- **Transform 목적:** 사람이 이해가 가능한 방식으로 데이터 자체를 변경함
- **Descriptive Statistics 목적:** 데이터 특성을 확인하고 전처리 방향과 완료정도 체크
- **Visualization/Dashboard 목적:** 데이터 특성을 확인하고 전처리 방향과 완료정도 체크

#### (3) 데이터 정리:

분석 단계	분석 역할	분석 도구
<b>(3) 데이터 정리:</b> 데이터 한곳에 담기(Data Warehouse) 데이터집계		
바꾸기 및 정리(Data Mart)	데이터집계	
데이터 분리(Data Split)	데이터집계	

⇒ 데이터 플랫폼에 데이터가 이관 되고 곧 분석 착수 보고

"1회성 수집/전처리/정리로 끝나지 않고 끊임 없이 업데이트하고 전화 시켜야 한다(알고리즘이 해주지 않는다)"

"데이터 수집/전처리/정리 까지 전체 업무의 80% 이상 을 차지한다"

- **Data Warehouse 목적:** 전처리 단계를 거친 1개의 Database를 주로 보관 및 무결점 유지 목적
- **Data Mart 목적:** Warehouse를 변경하지 않고 복사하여 조금 더 목적에 맞게 전처리
- **Data Split 목적:** 주로 과거(Train Data)와 미래(Test Data)를 구분하여 저장 및 알고리즘에 활용

#### (4) 모델링 및 검증:

분석 단계	분석 역할	분석 도구
(4) 모델링 및 검증:	기초통계(Descriptive Statistics) 모델링(Algorithm) 검증(Evaluation) 에러분석(Error Analysis)	기술적 분석, 데이터 집계, 상관분석 등 통계추론, 확률적 검정, 지도학습, 비지도학습, 강화학습, 기계학습, 딥러닝 등 성능측정, 검증지표 등 잔차추정, 검증지표 등

⇒ 분석이 완료 되어 인사이트 가 곧 '쏟아질 것이라 보고'

"수학적으로 어려울 수 있지만 수동적 으로 대응/활용 하는 것이며, 알고리즘(또는 기계)은 정해진 검증방법을 따를 뿐이다"

"각 알고리즘의 사용 목적에 대한 명확한 이해와 결과해석에 집중해야 한다"

"어떤 알고리즘 성능이 뛰어난지 검증(Evaluation)은 결국 사람 이 하기 때문에 많은 고민을 해야 한다"

"알고리즘 적용이 중요한게 아니라 언제 분석과정을 끝내야 하는지 고민 해야 한다"

- **Descriptive Statistics 목적:** 어떤 분석 알고리즘 선정 할지 또는 Input/Output 형태를 결정하는 기준으로 활용
- **Algorithm 목적:** Input/Output의 형태 또는 분석목적에 따라 정해짐
- **Evaluation 목적:** 현 알고리즘 성능 확인 및 다음 업데이트를 위한 기준 설정
- **Error Analysis 목적:** 모든 데이터의 패턴/특징을 알고리즘이 반영하고 있음을 평가

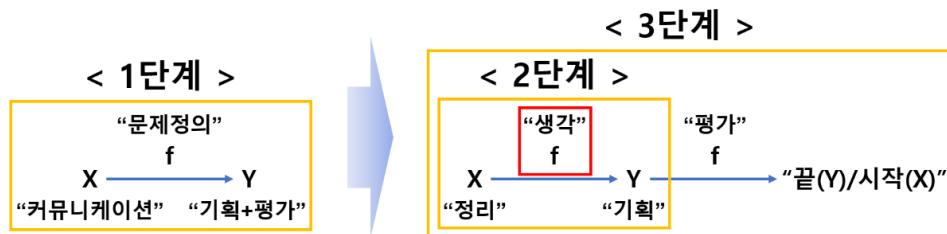
#### (5) 결과 활용:

분석 단계	분석 역할	분석 도구
(5) 결과 활용:	시각화(Visualization/Dashboard) 의사결정(Decision Support)	데이터 시각화, 상관분석 등

"0~4 단계를 무한대로 반복 및 각 단계를 업데이트하며 인사이트를 뽑아낼 수 있어야 한다"

- **Visualization / Dashboard / Decision / Knowledge / Reporting 목적:** 주로 고객에 및 충 설명력을 제공하기 위함으로 일반화된 방향은 없음

## 1.4 데이터 사이언티스트 스킬셋 3종



### 1) 1단계에서의 스킬셋

[현실] 생각하는 창의성 및 힘드는 수고

- 예능보단 예술: 생각을 안해야 즐거움을 얻을 수 있는 예능적 관심 보다 집중해서 생각하여 즐거움을 얻고자하는 예술적 의지
- 커뮤니케이션: 작은 나의 생각을 통한 즐거움을 세상의 큰 생각으로 연결시키려는 의지
- 질문 및 정리: 의지 실현을 위한 세상에 던지는 질문 + 정답(과거경험)이 아닌 보기(미래 가능성)를 만드는 정리

### 2) 2단계에서의 스킬셋

[이론] 데이터분석 관련지식

"순수 인문학 또는 예술과 관련된 학문이 아니면 대부분 수학과 컴퓨터는 모든 학문에서 적극 활용되는 도구(최근엔 순수학문들도..)"

- 수학
- 컴퓨터공학
- 통계학
- 경영과학
- 계량경제학
- 금융공학
- 물리학
- 각종 엔지니어링

"엑셀과 통계분석 툴로도 가능하지만, 모든 상황에 유연한 대응과 자동화를 위해 최소 1개이상의 언어능력 필요"

### 3) 3단계에서의 스킬셋

#### [현실] 설득 및 설명능력

- 인문학적 소양: 데이터 해석 + 인사이트 발견 및 지식화 + 상대에게 공유 및 설득
- 시각화: 기술과 예술의 중간단계로 미적감각 + 창의성 + 컴퓨터 활용능력
- 프리젠테이션: 내가 하고자 하는 말의 전달/주장 보다, 설명 능력으로 상대가 듣고 싶은 주제파악 + 표현문구 창의적 연구 + 스토리 라인 창의적 구현 + 컴퓨터 활용능력