



elalabs

Big Data 시장에서의 Data Scientist의 역할과 비전

유 충 현

elalabs 수석 연구원
(R Tech Center 고문)

2014. 10. 8

1. Big Data 시장의 전망
2. Big Data 핵심요소
3. Data Scientist의 필요 역량
4. Data Scientist의 비전
5. Big Data 프로젝트 성공 요소
6. Data Driven

IT 패러다임은 솔루션과 벤더, 그리고 컨설팅 그룹에서 이끌고 있음
→ 수요가 아닌 공급에서의 Leading

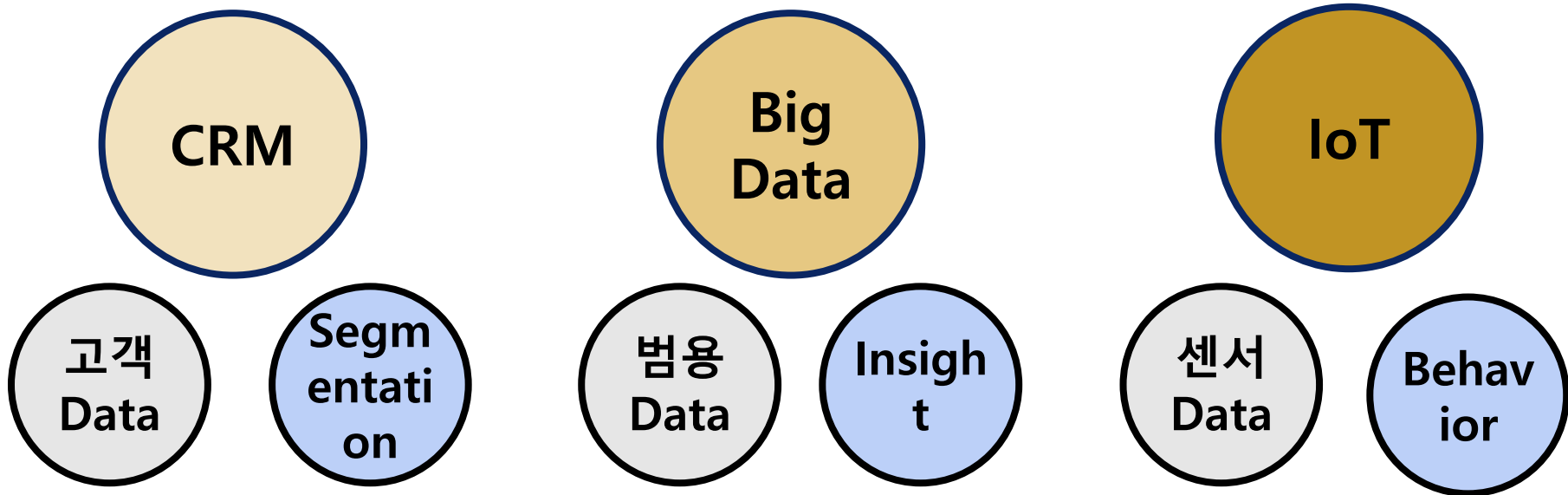


- 구체적이고 팬시한 용어
- 욕구 층만의 Top-Down 시행
- 시행착오의 시련기 극복

- 모호한 개념과 춘추전국
- 여전히 Top-Down 시행
- 추수가 아닌 파종기

- 또 다른 언어 수사인가
- 다분히 Bottom-Up 수행
- 새로운 융합시대의 태동기

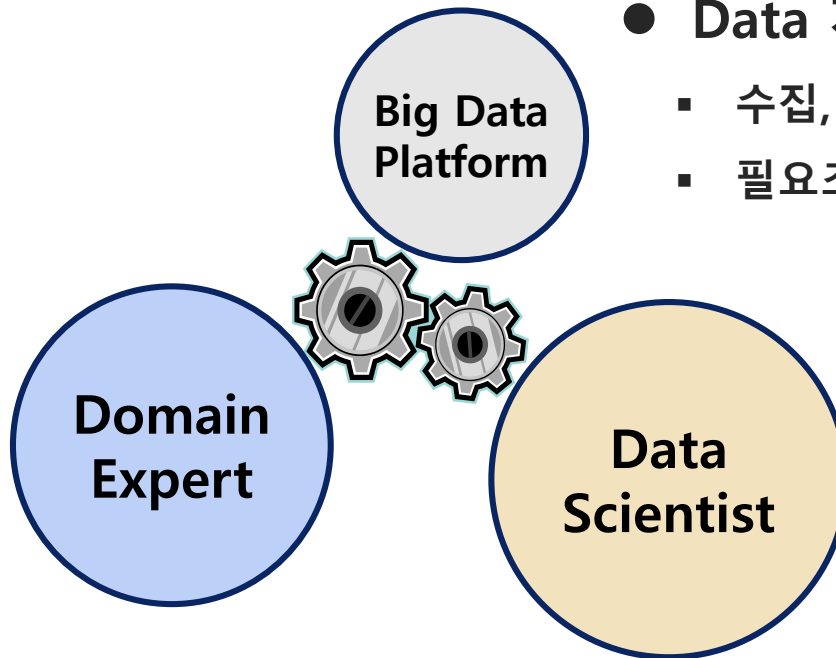
IT 패러다임의 핵심에 Data와 Data Analytics가 자리잡고 있음



- CRM 전략
- 고객의 행태 Data
- MASS to Target

- Big Data Analytics 전략
- 가능한 모든 Data
- Insight

- 융합 전략
- 센서 계측 Data
- Interactive Insight

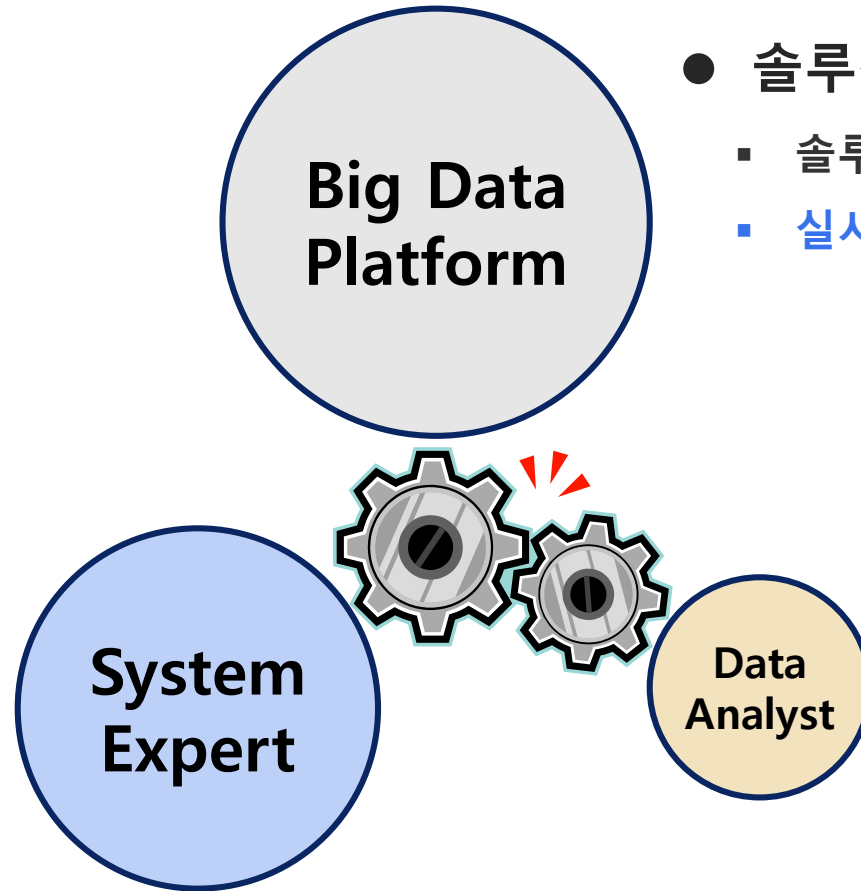


- Data 처리 한계의 극복
 - 수집, 가공, 저장, 분석 기능
 - 필요조건, **충분조건이 아님**

- 비즈니스 모델, 방향성 제시
 - **업무의 이해, Data의 이해**
 - 현업 담당자 + 시스템 담당자

- 분석 모델 제시, Insight 도출
 - Data분석 방법론의 이해
 - **Data 전처리, 분석, 경험**

그러나 현실은,



- 솔루션의 종합 선물세트

- 솔루션은 수단이지 목적이 아님
- 실시간 처리에 과도한 집착

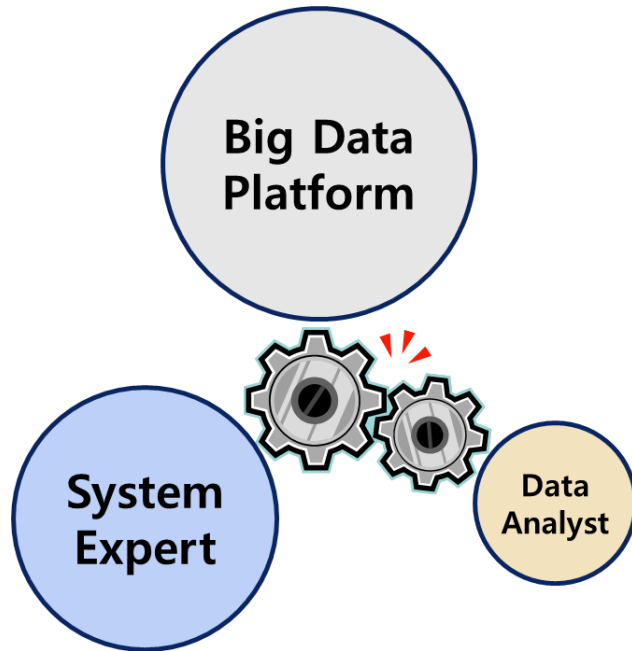
- Data Scientist 희소성

- 특정영역, 분석패턴 의존성
- Data 전처리 능력 부족

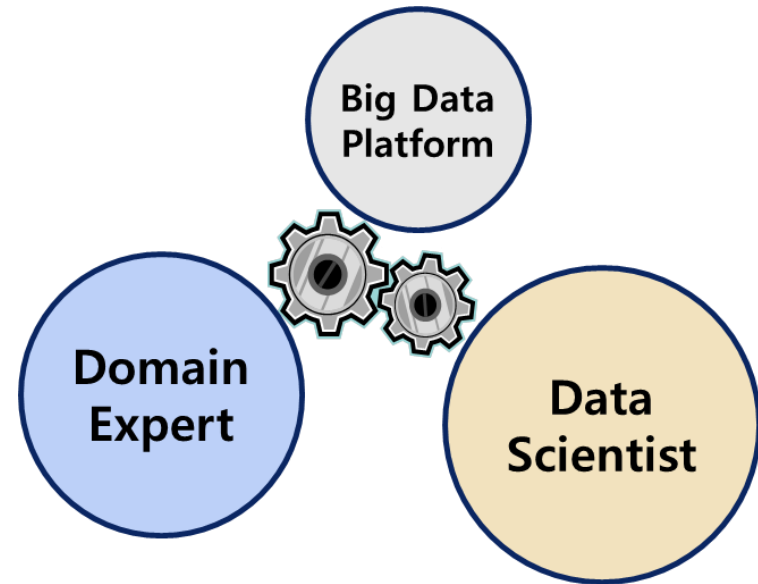
- 분석의 목적은 무엇인가?

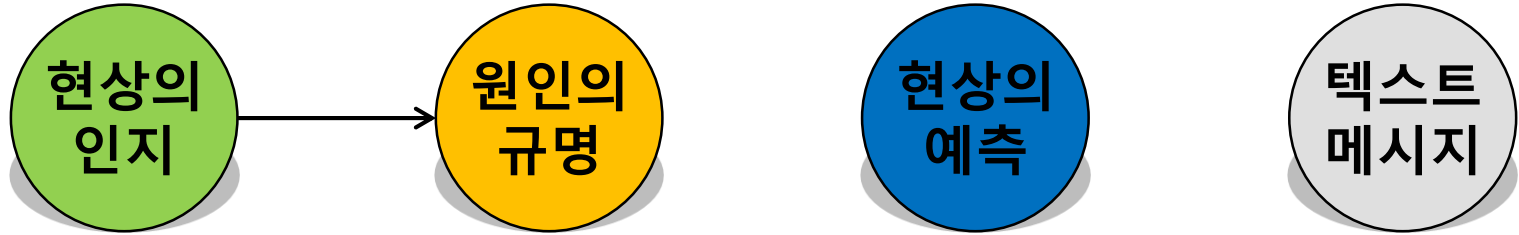
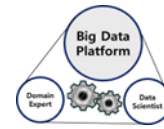
- 현업 업무 담당자의 부재
- IT 주도의 빅Data 실행 조직

실패로 가는 환경

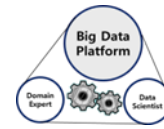


성공으로 가는 환경

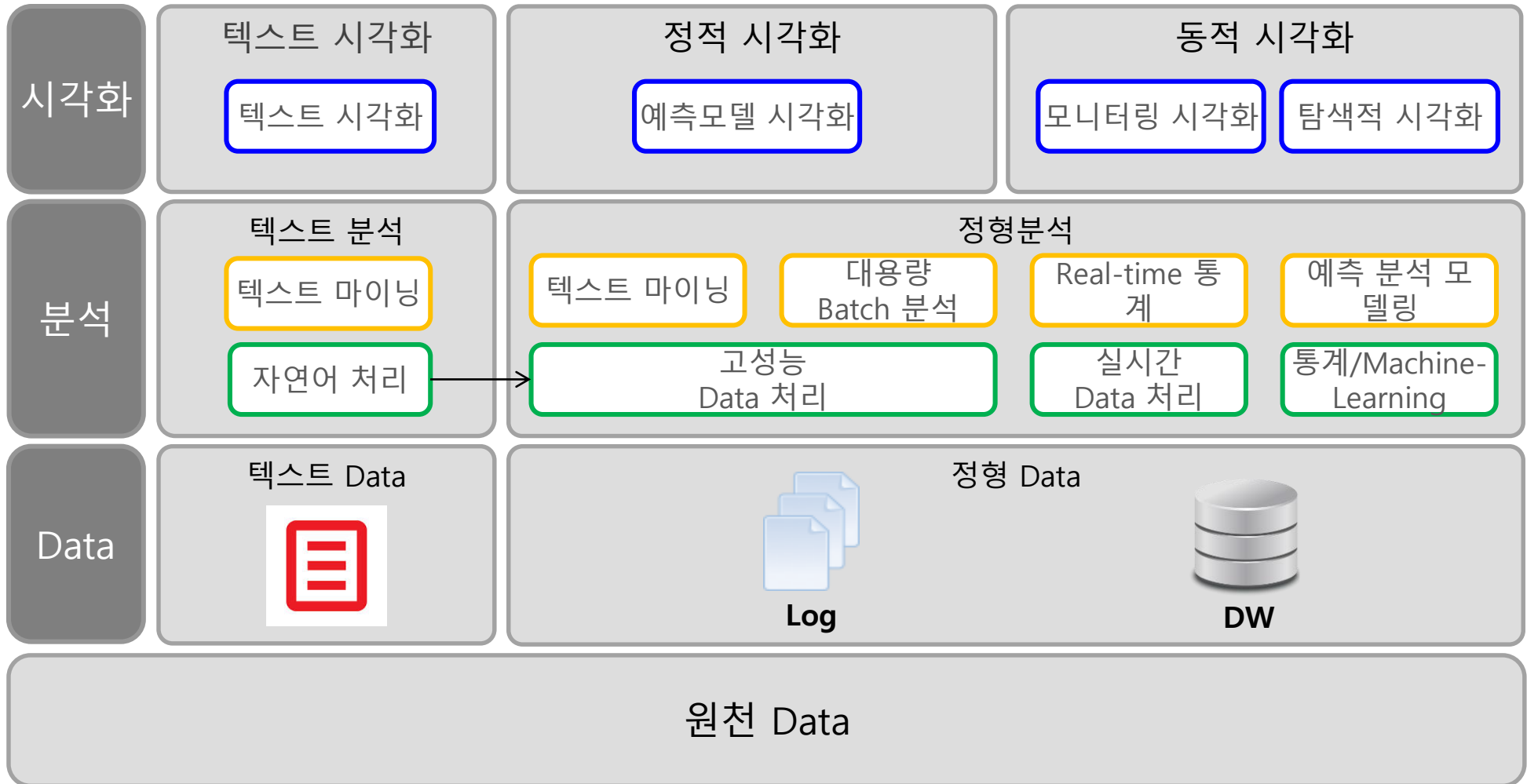




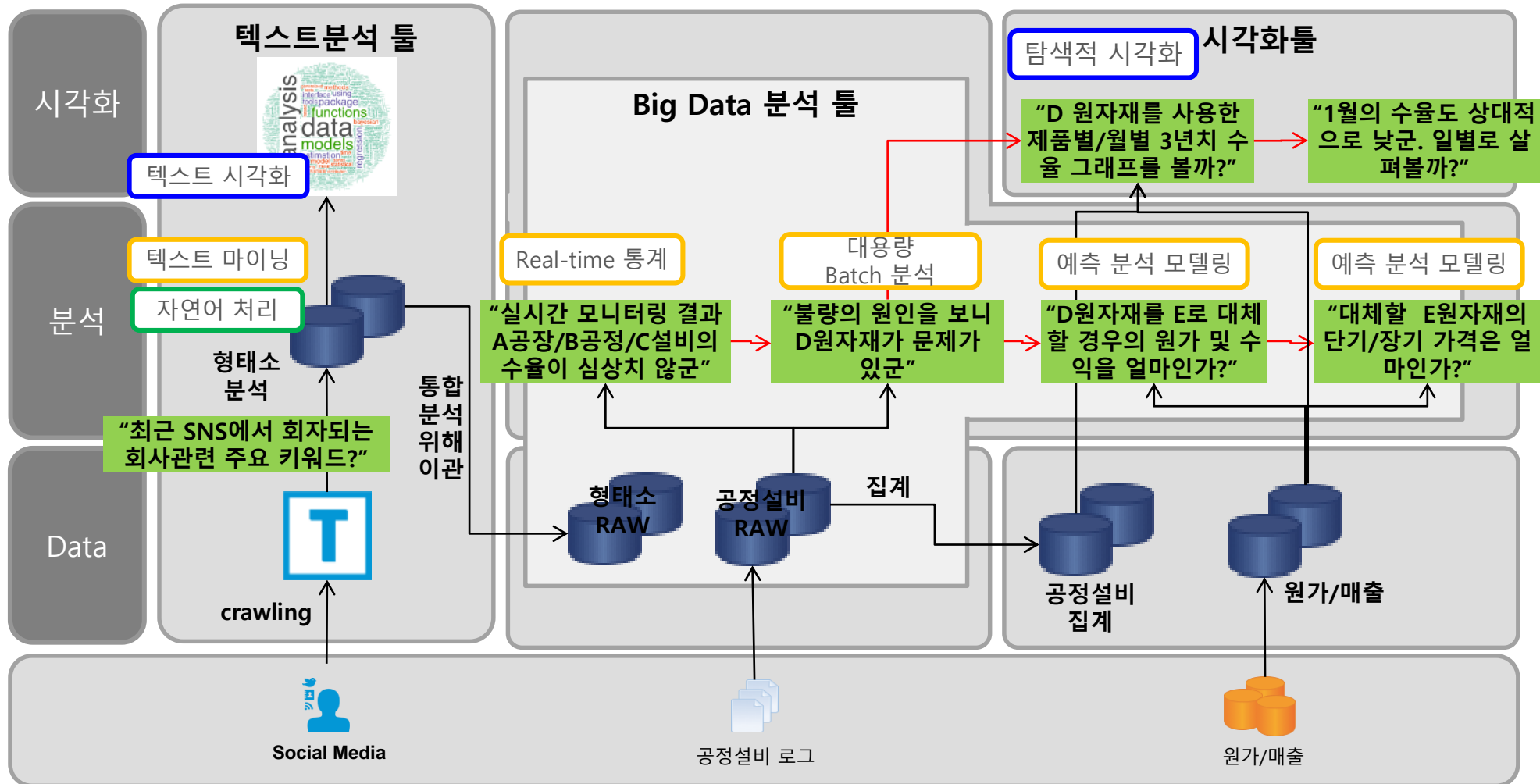
문제 인식	문제가 발생하고 있는 공정/설비는 무엇인가?	왜 불량률이 높아지고 있는 것일까?	석 달 뒤의 원자재의 가격은 얼마나 될까?	텍스트 안에 숨겨진 의미가 있을 것이다.
분석 형태	모니터링	인과관계 규명	예측	텍스트 분석
분석 방법	Real-time 통계	대용량 Batch 분석	예측 분석 모델링	텍스트 마이닝
분석 기술	실시간 Data 처리	고성능 Data 처리	통계/Machine-Learning	자연어 처리
시각화 기술	모니터링 시각화	탐색적 시각화	예측모델 시각화	텍스트 시각화



효율적인 분석/시각화 구성의 제안을 위해서 Data 분석의 기술 요소를 Data/분석/시각화라는 3개의 Layer에 유기적인 관계를 도식하면 다음의 그림과 같음



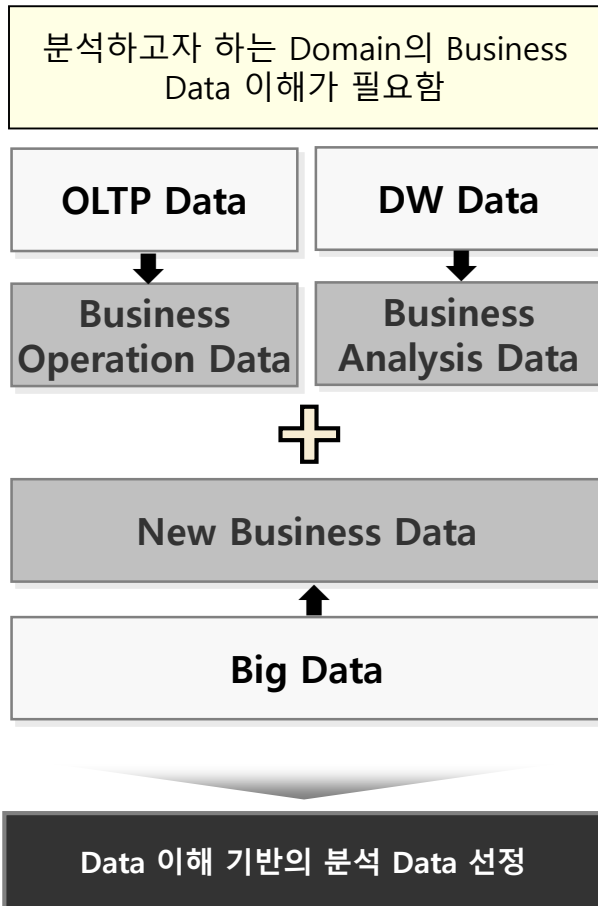
Seamless하게 연결된 각 솔루션들의 분석결과로서의 Insight는 유기적으로 연결되어 마치 하나의 분석 도구를 통한 결과처럼 통합된 Insight로 활용되어야 함





Business를 이해하는 것부터 Data 분석이 시작되며, Domain Expert의 참여나 해당 Domain에서의 유사한 Data 분석의 경험은 Big Data 분석의 실패 위험을 제거해 줌

Business Data 이해



Business 이해



Business Analytics 이해

분석하고자 하는 Domain에 최적화된 Data 분석 방법론의 이해 필요

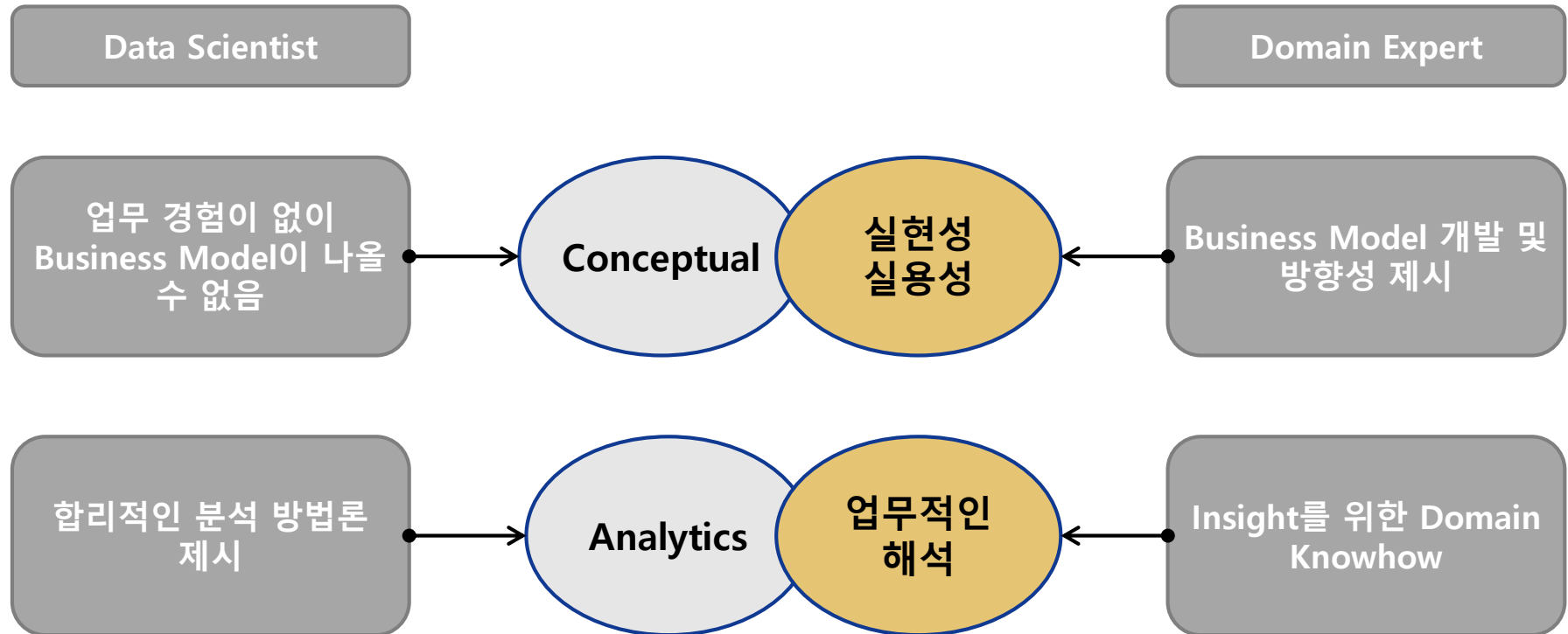
Domain	목적	대표적 분석 방법
Social / Internet	서비스	<ul style="list-style-type: none"> SNA Text Mining
Consumer	CRM	<ul style="list-style-type: none"> Churn Analysis RFM Customer Segmentation
Manufacture	수율 개선	<ul style="list-style-type: none"> SPC Optimization
Finance	Risk Management	<ul style="list-style-type: none"> Forecasting Portfolio Analysis CSS, BSS

Domain 특성에 기반한 분석 모델 선정



Big Data를 활용한 Business Model을 제시해 달라.

→ Business Model은 Data Scientist 보다 Domain Expert가 주도해야 함



히딩크 감독의 Multi-Player 전략과 Data Scientist의 공통점은?

→ 단편적인 기술이 아닌 입체적인 기술을 요구함

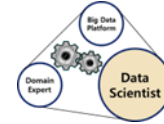


“공격할 땐 수비를, 수비할 때는 공격을 생각하라”

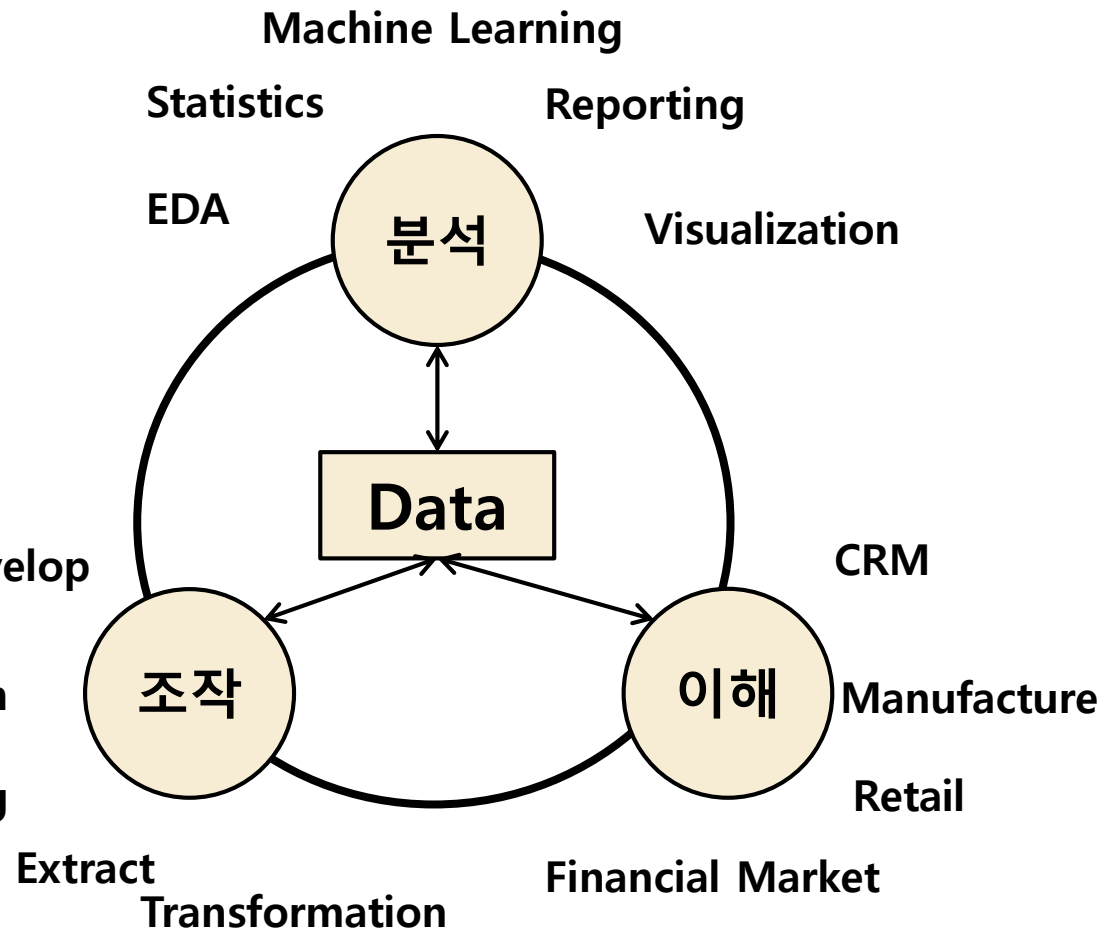
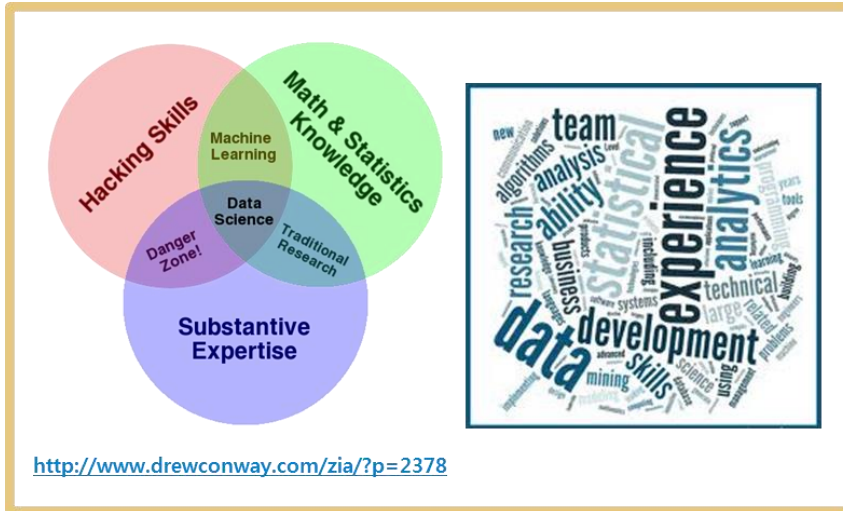
공격수
(수비수의 체력을)

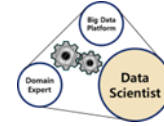
수비수
(공격수의 골 결정력)

기초 체력



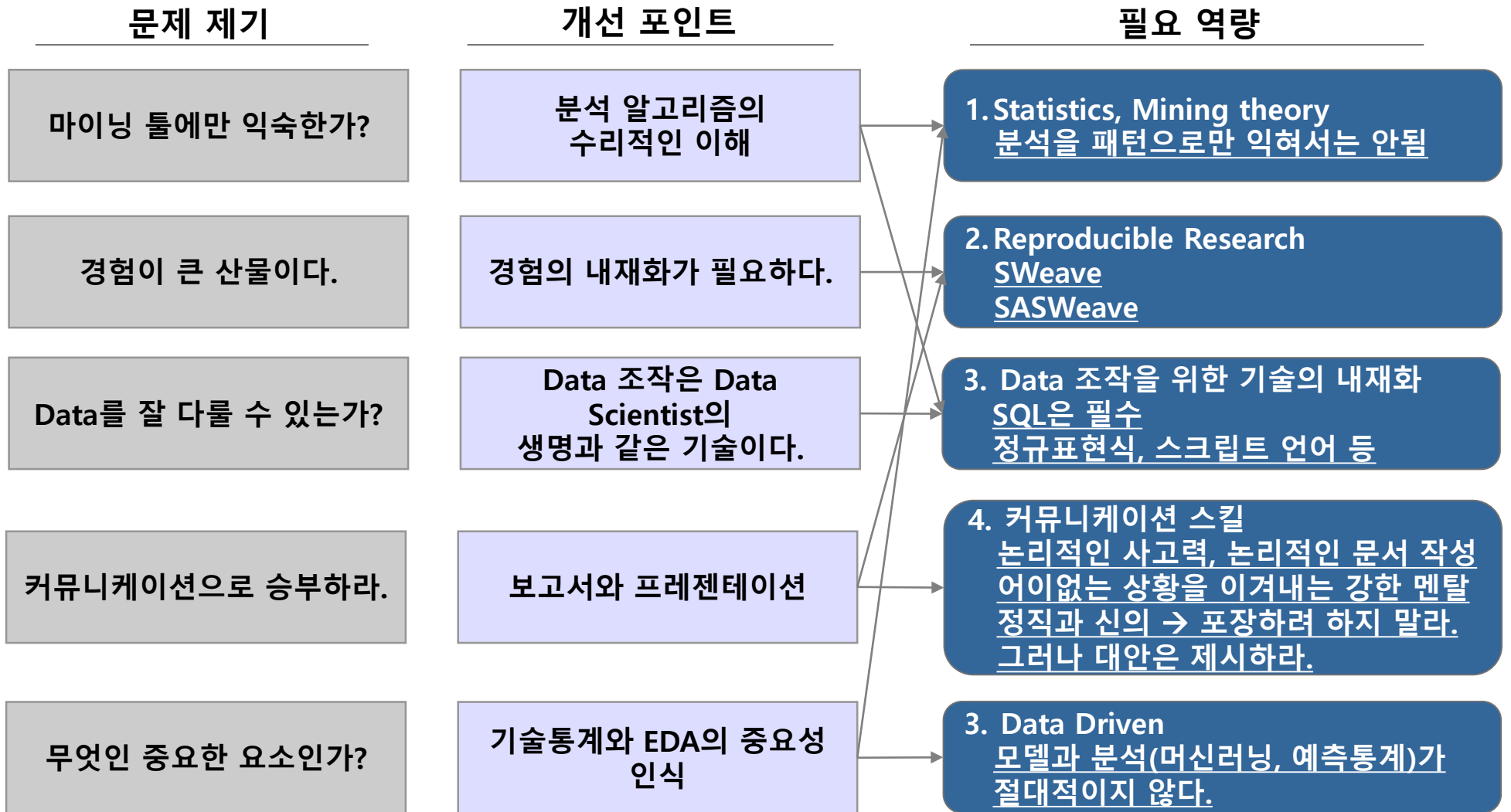
“Data 분석을 위한 수리적인 이론, Data 조작을 위한 컴퓨터 엔지니어링, Data 이해를 위한 업무지식과 경험을 갖춘 **Multi-player**”





Data를 조작하고 분석하며, 이해하는 일련의 Data Service 과정의 작업 수행

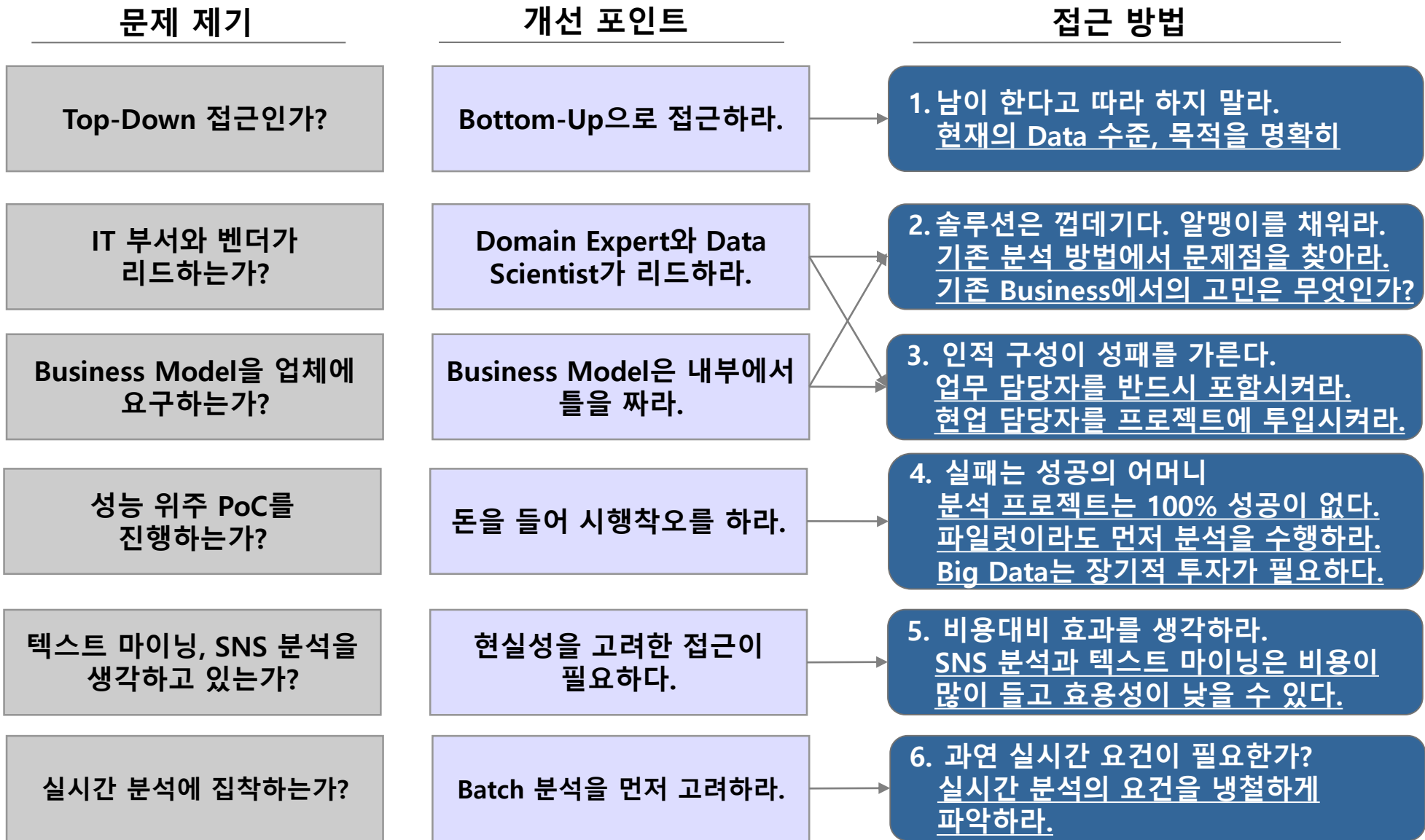
	Input Data 준비	Process Data 분석	Output 분석 결과의 정보화
Roles	<ul style="list-style-type: none"> 조작 	<ul style="list-style-type: none"> 분석 	<ul style="list-style-type: none"> 보고
Activity	<ul style="list-style-type: none"> Data 획득 Data 정제 Data 변형 	<ul style="list-style-type: none"> Data 분석 	<ul style="list-style-type: none"> 분석 결과 해석 보고서 작성 분석 결과 보고
Technics	<ul style="list-style-type: none"> Computer Science Data Manipulation 	<ul style="list-style-type: none"> Mathematics, Statistics Data Mining Visualization 	<ul style="list-style-type: none"> Reporting Domain Knowledge Graphics Design Presentation
Who	<ul style="list-style-type: none"> Data Geeks 	<ul style="list-style-type: none"> Statistician 	<ul style="list-style-type: none"> Domain Expert



구분	정의	특징	현실적 문제
<p>SM형 Data Scientist</p>	<ul style="list-style-type: none"> ▪ 회사 내부의 Data 분석을 목적으로 조직된 부서 소속의 Data Scientist ▪ 규모가 큰 대기업에 근무 ▪ 게임, 포털 업체 등 	<ul style="list-style-type: none"> ▪ 근무환경이 안정적임 ▪ 업무 전문가를 겸하며, 분석 주제가 다양하지 않음 ▪ 수요가 많지 않음 	<ul style="list-style-type: none"> ▪ 반복되는 유사한 분석으로 매너리즘에 빠질 수 있음
<p>SI형 Data Scientist</p>	<ul style="list-style-type: none"> ▪ Data 분석 용역 수행을 위해서 프로젝트에 파견되어는 Data Scientist ▪ Data 분석 용역을 수행하는 회사 소속이거나 프리랜서 	<ul style="list-style-type: none"> ▪ 프로젝트 사이트 파견으로 근무환경이 안정적이지 않고 열악함 ▪ 다양한 Domain, 다양한 Data, 다양한 분석 방법을 사용함 ▪ 공급이 많지 않음. 기피현상? 	<ul style="list-style-type: none"> ▪ 필드에서의 분석가는 환상적이지 않다. → 개발자보다는 낫다. ▪ 프로젝트를 인연으로 SM Data Scientist로 이직하는 경우 많음

그러면 비전은 무엇인가?

- 수요와 공급의 불균형 → 그런데 왜 몸값은 오르지 않는가?
- 재미?, 자기만족?, 명예?, 회사의 기여도?
- 연구 학습과 실무가 병행 가능하다.



기존의 Data 분석 답습을 버리고 Data 중심으로 재 도전하라.

핵심은 모델이 아닌 파생Data

“기존에 안 다루려던 Data를 찾다 보니 VoC, SNS, 웹 로그 등을 생각한다.”

Cleansing이 관건이다.

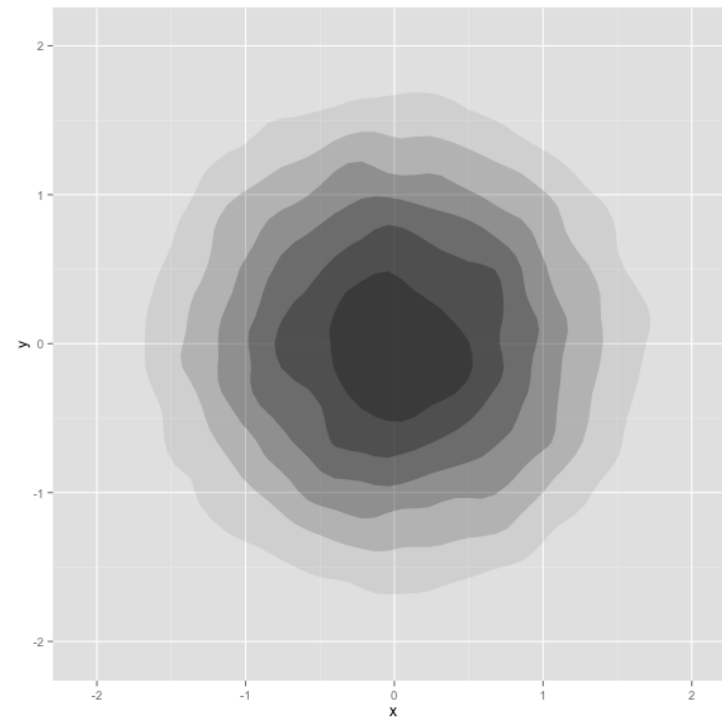
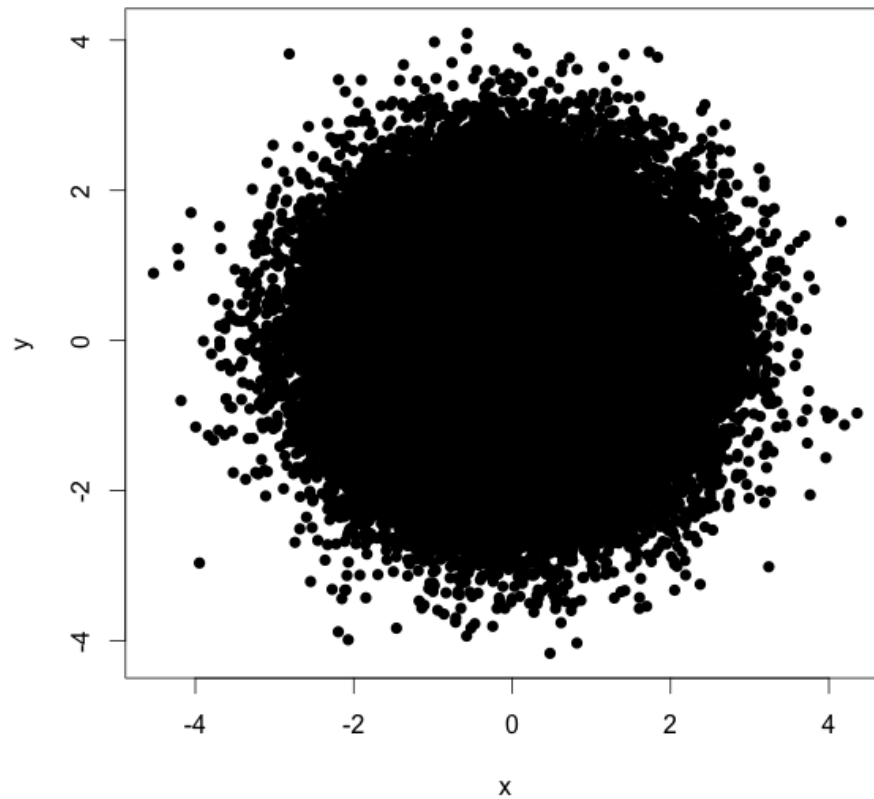
Data Driven

간단한 집계로도 Insight가 도출된다.

“의외로 Data의 정합성이 떨어지고, Spec대로 만들어지지 않고 있다.”

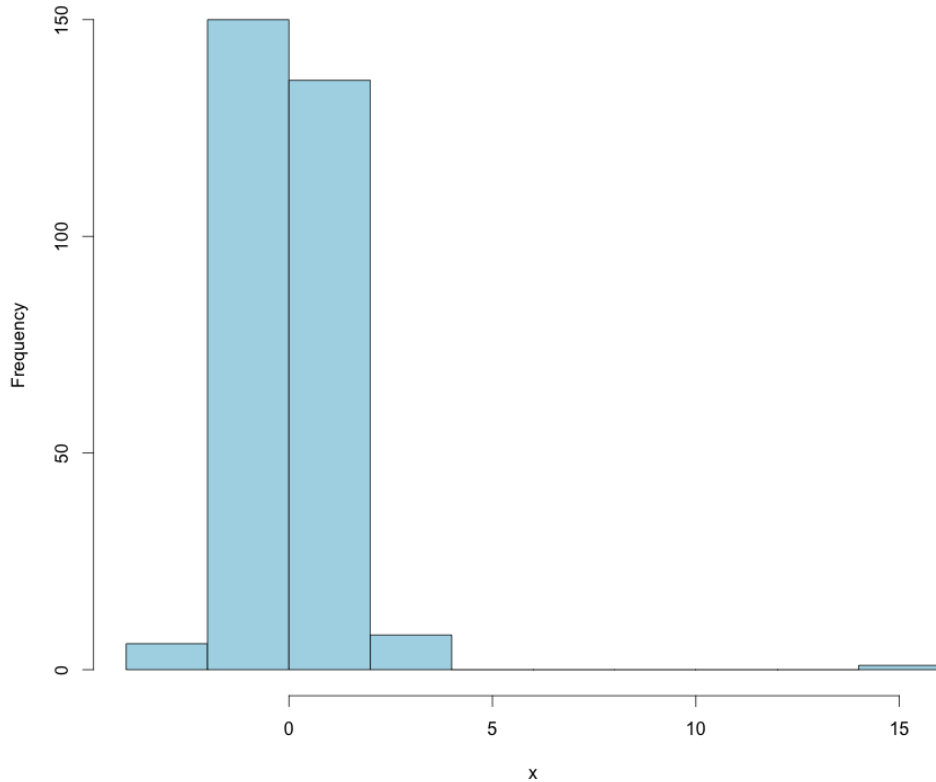
Data 거버넌스를 선행하라. Garbage in, Garbage Out

그 많은 크기의 Data를 어떻게 집계할 것인가? 집계를 통한 패턴의 인식이 필요함

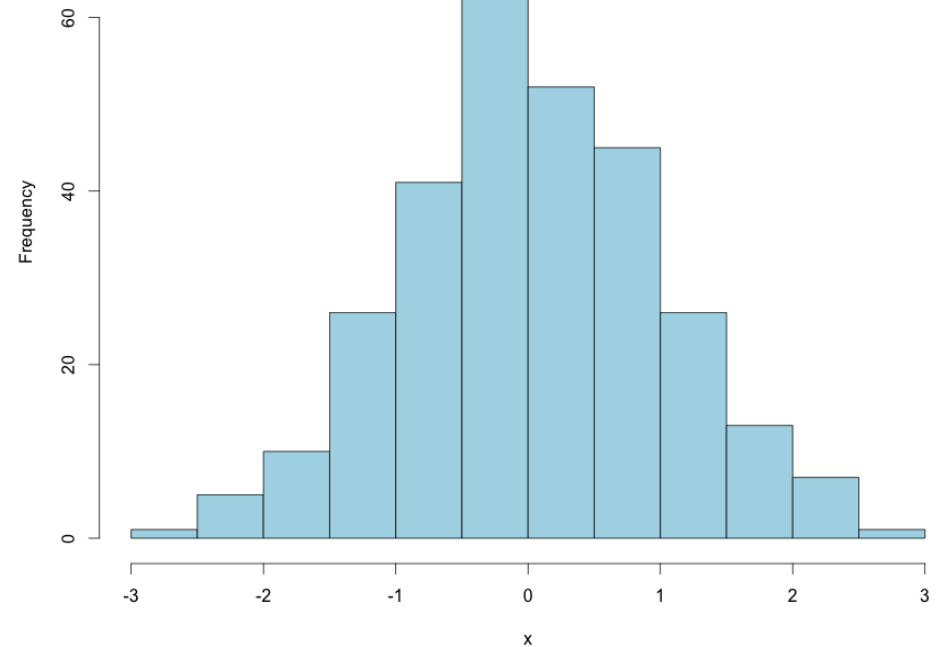


관리 되지 않고, 분석되지 않던 Data라서 Cleansing과 Discovery의 공수가 따라온다.

Histogram of x



Histogram of x



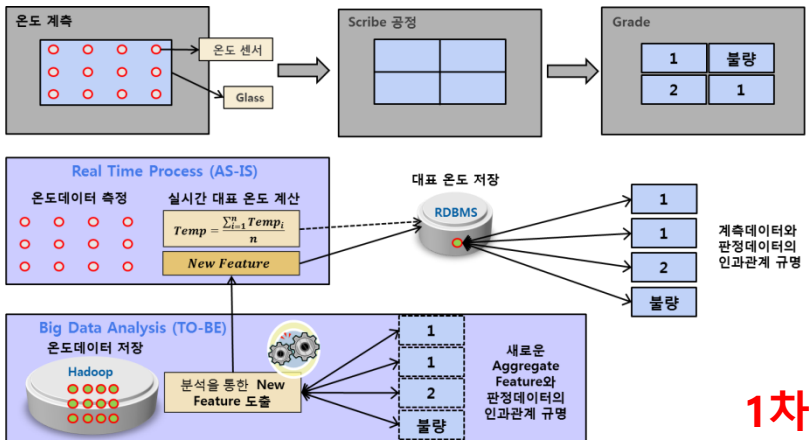
OOOO의 문제제기

“센서 Data가 어마어마하게 쏟아지지만 저장의 한계로 단위 시간 기준 산술평균으로 가공하여 분석을 하는데 이게 맞는 방법인가요?” - 2006년 LCD 공정 분석 담당자 (Big Data 플랫폼 이전의 환경)

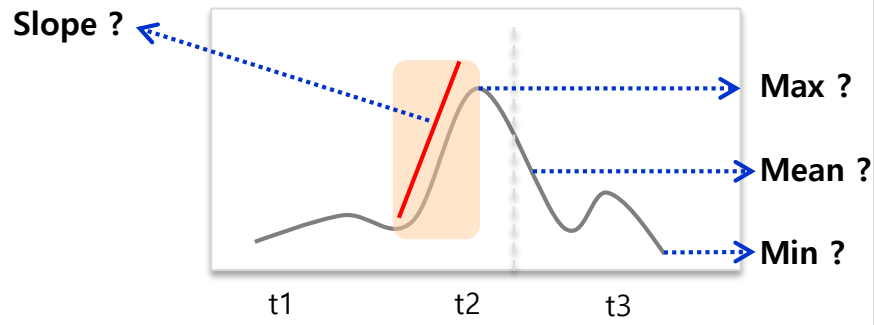
Data의 질적 향상

“이미 수많은 수율분석 프로젝트를 수행했기 때문에 새로운 모델로 수율 향상을 꾀하기는 어렵습니다. 일단 계속 Data를 가감 없이 담아보세요. 먼저 원천 Data에서 Data를 집계하는 방법부터 Clinic해야 합니다.”

위치적 집계의 문제점



시간적 집계의 문제점



1차 분석을 통해서 유용한 Measure 도출 필요

감사합니다.

antony.ryu@elalabs.com
bdboy@rtechcenter.com