



문화 A0007

데이터사이언스입문

김 태 완

kimtwan21@dongduk.ac.kr

데이터 사이언스?

두 가지 이야기



#1 마트로 쳐들어간 아버지



TARGET

TARGET 은 알고 있었다.



향 있는 로션



향 없는 로션

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

"[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date."

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August.

So Target started sending coupons for baby items to customers according to their pregnancy

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

2/16/2012 @ 11:02AM

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had

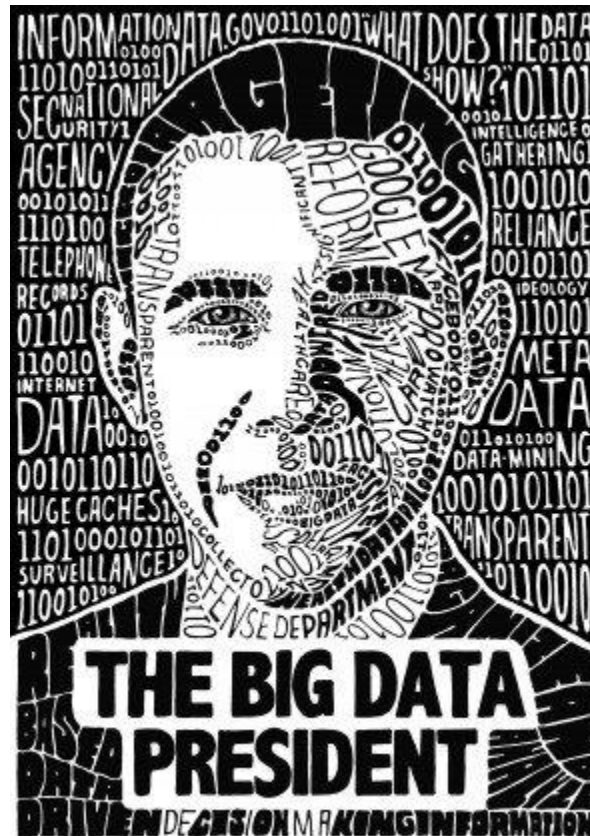


Target has got you in its aim

signed up for Target baby registries in the past:

[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of

#2 버락 오바마 대통령



선거자금 차이 2배 이상



약 2억6544만	총액	약 9665만
약 3070만	4월 모금액	약 1831만
약 4655만	보유현금 총액	약 2398만
약 204만	부채	약 96만
4060만	정당별 보유현금 총액	440만

자료: www.opensecrets.org, 가디언

미대선 후보별 선거자금 모금 현황

※2008년 10월 15일 미 연방선관위 신고 기준

정치활동위(PAC) 후원



VS



자료: 책임정치센터(CRP)

빅데이터 분석 팀 운영



Help build this campaign

What we do right now will determine the course of this election.
It's that simple.

Make a donation today and build the grassroots organization it will take to win.

How much would you like to donate today?

Secure

Select amount

Amount

Name

Payment

Employment

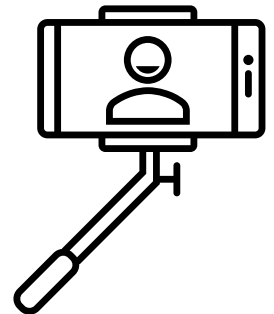
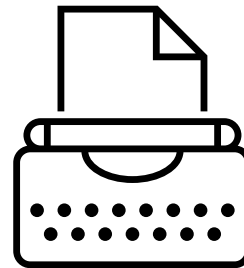
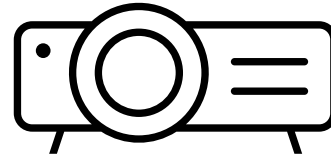
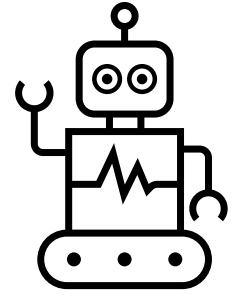
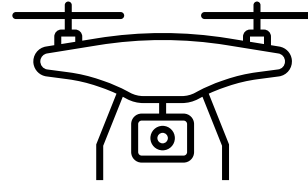


시대가 흐르며

경쟁력의 핵심이 바뀜


기술

기술



기술 → **가격**

기술 → 가격



2,430만 화소 Wi-Fi

최저가/최고가
1,410,000원 ~ 2,449,000원 가격비교

소형물별 최저가 99개의 가격비교상품 더보기

☐ 조건연계
 ☒ 정품-본체

☐ 정품-본체+28~70mm렌즈

☒ 1,410,000 무료배송 카드발입

인터파크	1,411,350	무료배송	<input type="button" value="사라가기"/>
G마켓	1,415,880	무료배송	<input type="button" value="사라가기"/>
옥션	1,415,880	무료배송	<input type="button" value="사라가기"/>
미리리샵	1,450,000	무료배송	<input type="button" value="사라가기"/>
BSC디점	1,470,000	무료배송	<input type="button" value="사라가기"/>
현대Hmall	1,527,590	무료배송 카드발입	<input type="button" value="사라가기"/>
롯데	1,537,420	무료배송	<input type="button" value="사라가기"/>

風林火山

제품정보 2013.10 더보기

제조사/브랜드 : 소니/소니

화소 2430만화소, 센서크기 1:1, 화소셔터속도 1/8000초, 최대연속촬영속도 58매/무게 416g, 1:1 풀프레임, 라이브뷰, HD동영상, 방진, 방습, 무선전송, 휴파인더, 파노라마, NFC, 최대동영상프레임 60프레임, 최대 동영상 크기 1920x1080, 타입 미러리스, 저장매체 SD, 저장매



최저가 **649,000원**

618,500원 2,500원
 현금 647,180원 2,500원

프로	현금 618,500원	2,500원	
ITENJOY	현금 647,180원	2,500원	
AUCTION	최저가 649,000원	무료배송	최대 20개월
로인디지털	649,000원	무료배송	최대 6개월
emart	649,000원	무료배송	최대 12개월
SSG.COM	649,000원	무료배송	최대 12개월
11번가	649,000원	무료배송	최대 12개월
G마켓	649,000원	무료배송	최대 20개월
11번가	649,000원	무료배송	최대 20개월
인터파크	738,240원	무료배송	최대 24개월

현대(SUPER CLUB) 카드 최저가 **603,570원** 무료배송

alphascan 브랜드로그 바로가기

등록일: 2013.05 | 제조사: 알파스캔 | 이미지출처: 제조사제공

기술 → **가격** → **기회**



사람들의 관심이 가장 소중한 자원

눈길을 받는 법이 필요

눈길을 받았을 때 원하는 것을 제시

기술 → 가격 → 기회



Data

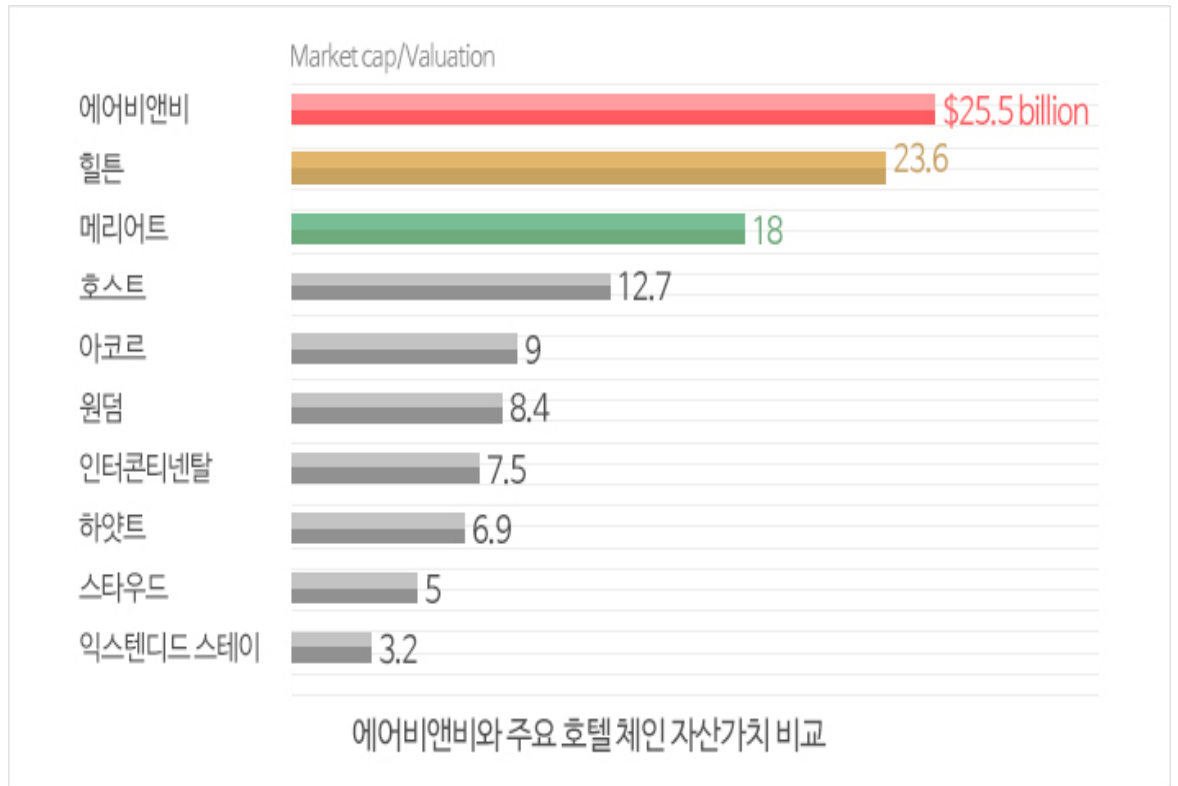
사람들의 관심이 가장 소중한 자원

눈길을 받는 법이 필요

눈길을 받았을 때 원하는 것을 제시

facebook

The New York Times





DATA Driven ...

데이터 기반으로 의사결정

데이터

1. 감정

오늘 춥다.
대존맛
이 어플 킹받네

3. 사실

코로나 시국
4초 남은 횡단보도 초록불
3월 2일은 개강

2. 분류

ESTJ
갤럭시 Z플립3
SKT, KT, LGU+

4. 소통

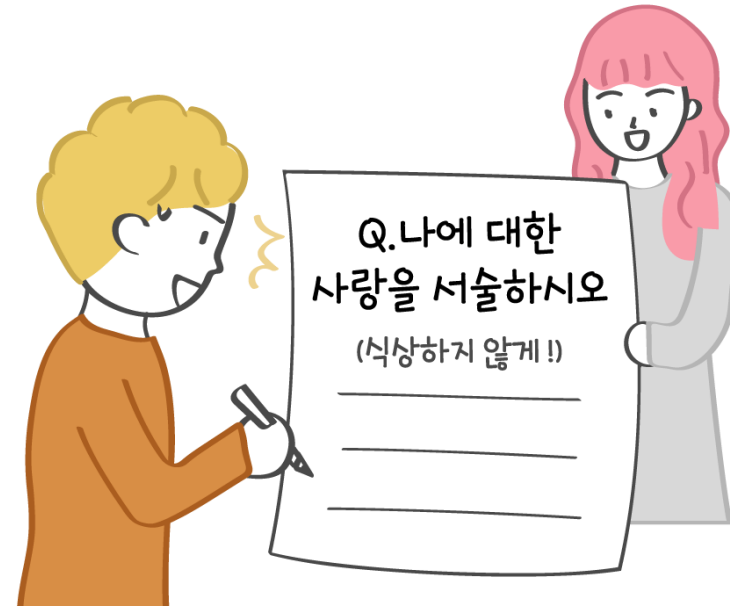
우리 내일 만날까?
파이썬으로 이게 가능해?
너 나한테 왜 그래?



데이터는 우리가 남기는 모든 흔적

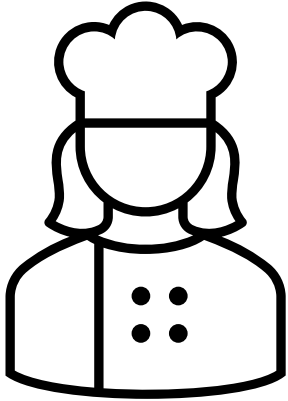
데이터

- 얼마나? 정도의 차이



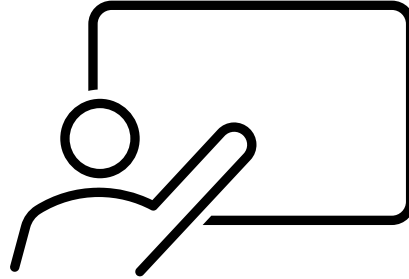
데이터

- 얼마나? 정도의 차이



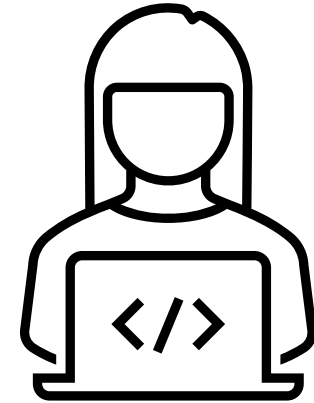
라면 얼마나 끓여야 하지?

적당히 끓이세요



얼마나 수업 할까?

조금만 하세요



얼마나 빠르게 성장할까?

충분히 빠르게 성장하고 있어요

- “그저 알고 있다.”와 “어느 정도까지 알고 있다.”의 차이를 아는 것이 중요
- 데이터를 통해 정도의 차이를 파악하고 의사결정 하자

데이터

- 인지력에 영향을 주는 데이터
 - CASE 1: 페이퍼 타월 절약 문구
 - 10 장이면 거의 100원 ? → 화폐 단위로 환산된 가치를 인식하게 함
 - 동일한 메시지여도 근거의 구체성 (데이터 정량화)에 따라 메시지의 의미 전달력이 달라짐



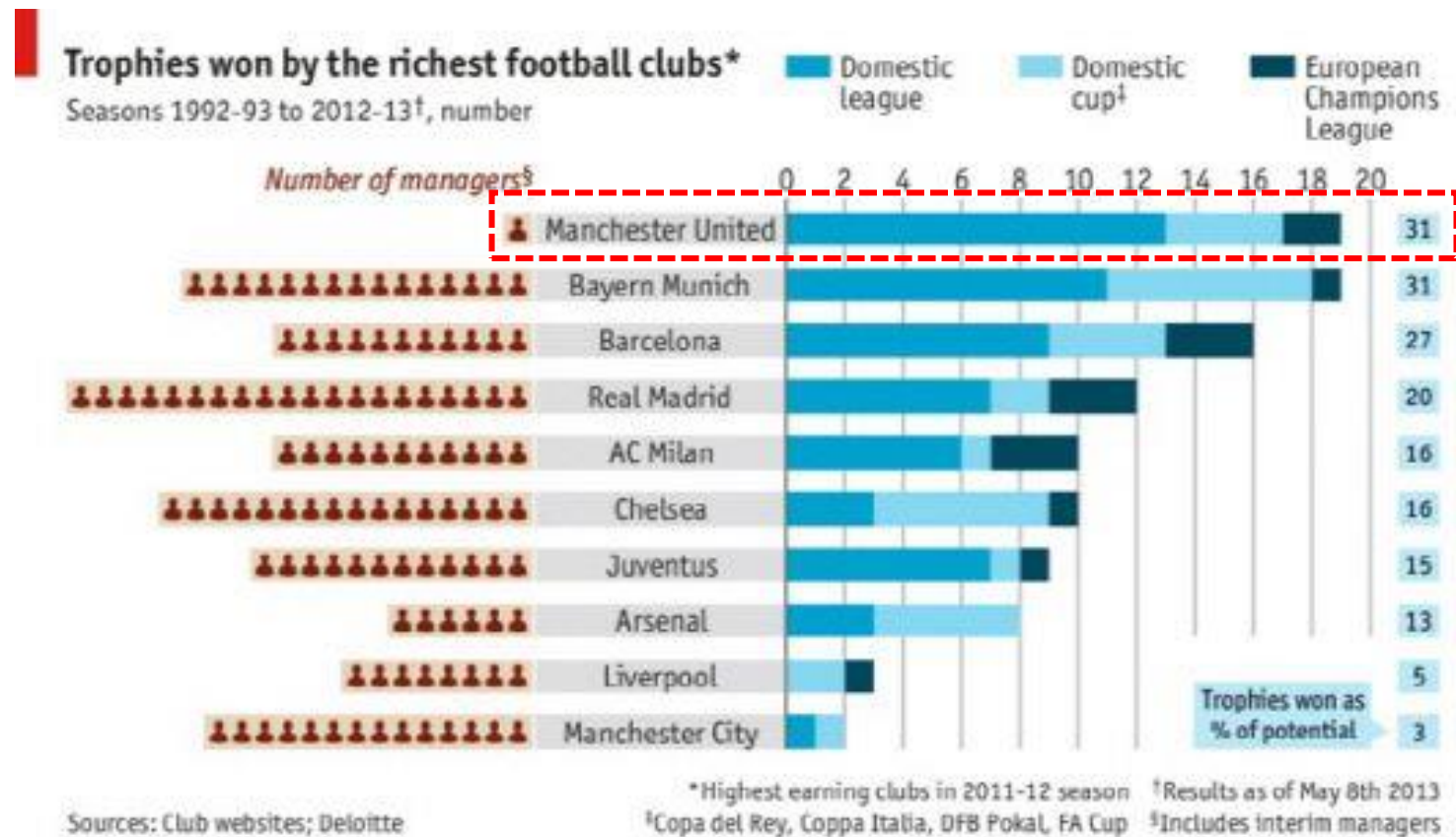
데이터

- 인지력에 영향을 주는 데이터
 - CASE II : 계단 공익광고
 - 한 층을 걸어 올라갈 때 마다 소비되는 칼로리의 정량화 수치 때문에 “얼마나” 인지의 효과 극대화
 - 데이터를 활용한 구체적인 표현은 상대방의 인식 변화 및 강화에 효과적



데이터

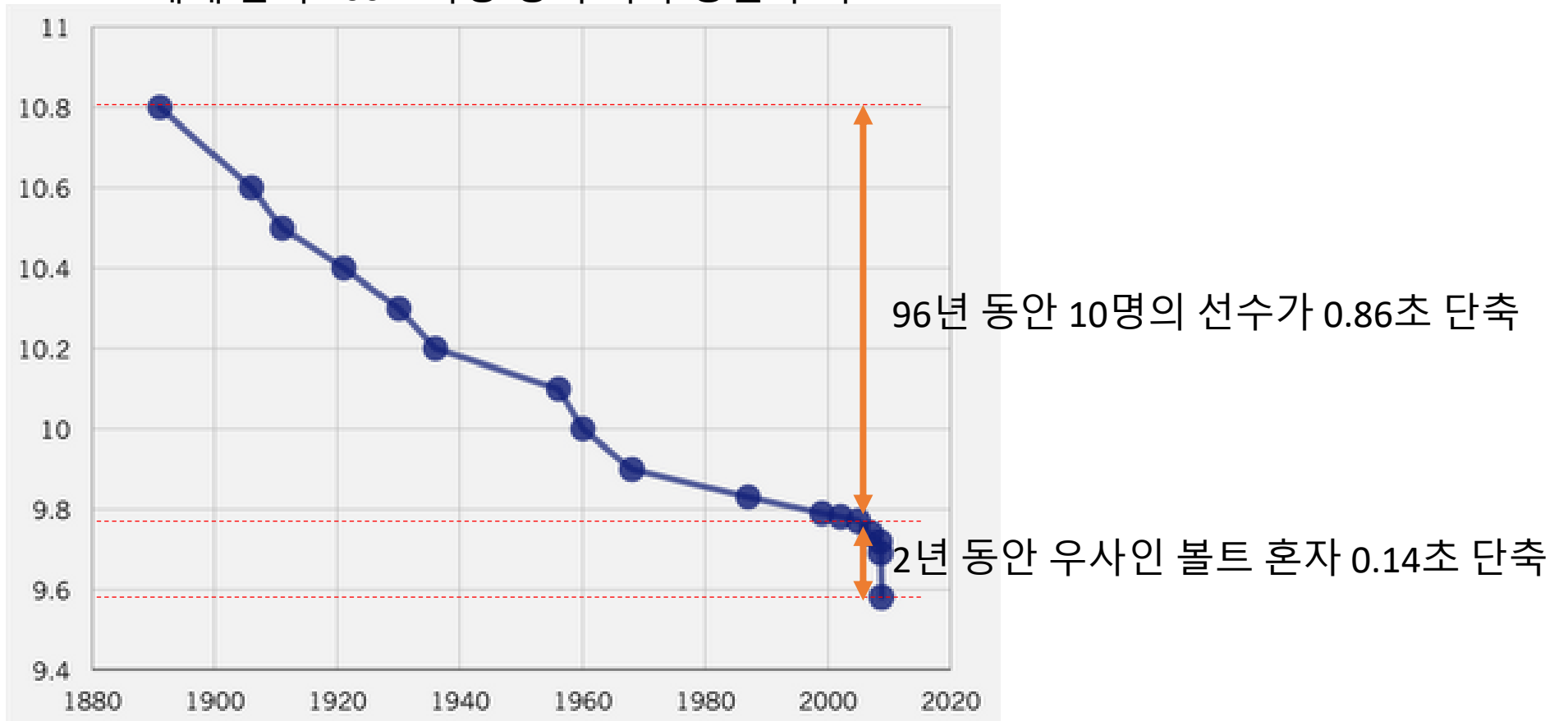
- 인지력에 영향을 주는 데이터
 - CASE III : 알렉스 퍼거슨 감독 (전 맨체스터 유나이티드 FC 감독)
 - “이 감독이 은퇴한 게 그렇게 큰 일이야?”
 - 구체적인 데이터를 활용한 주장은 정도의 차이를 명확하게 인식



데이터

- 인지력에 영향을 주는 데이터
 - CASE IV : 우사인 볼트의 업적
 - 우사인 볼트가 어마어마하게 빠름을 설명해 줌 → 우사인 볼트에 대한 명확한 인식

세계 남자 100m 육상 경기 기록 갱신 추이



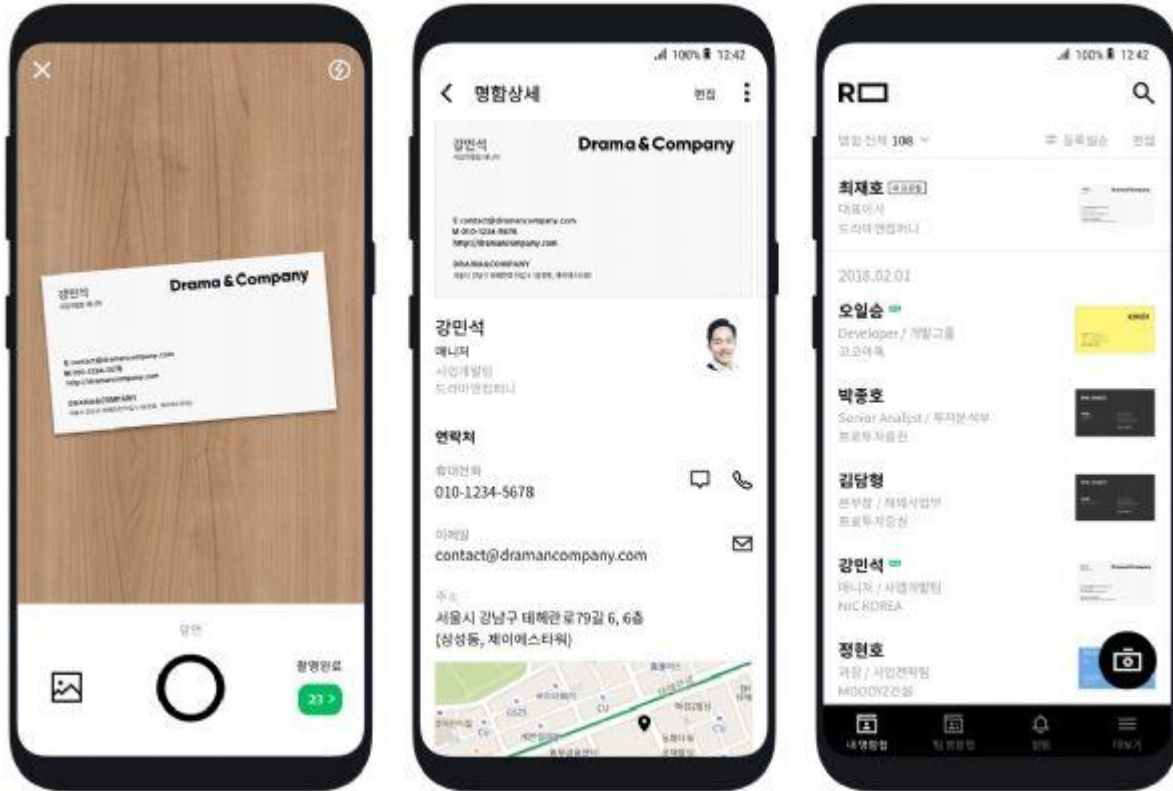
데이터

- 인지력에 영향을 주는 데이터
 - CASE IV : 우사인 볼트의 업적
 - 인간은 그렇게 빠른 동물은 아니다. → 우사인 볼트의 기록을 통해 속도 비교 가능
 - 애매한 것들의 정도 차이를 명확히 하면 명확한 비교를 통한 새로운 인식이 가능



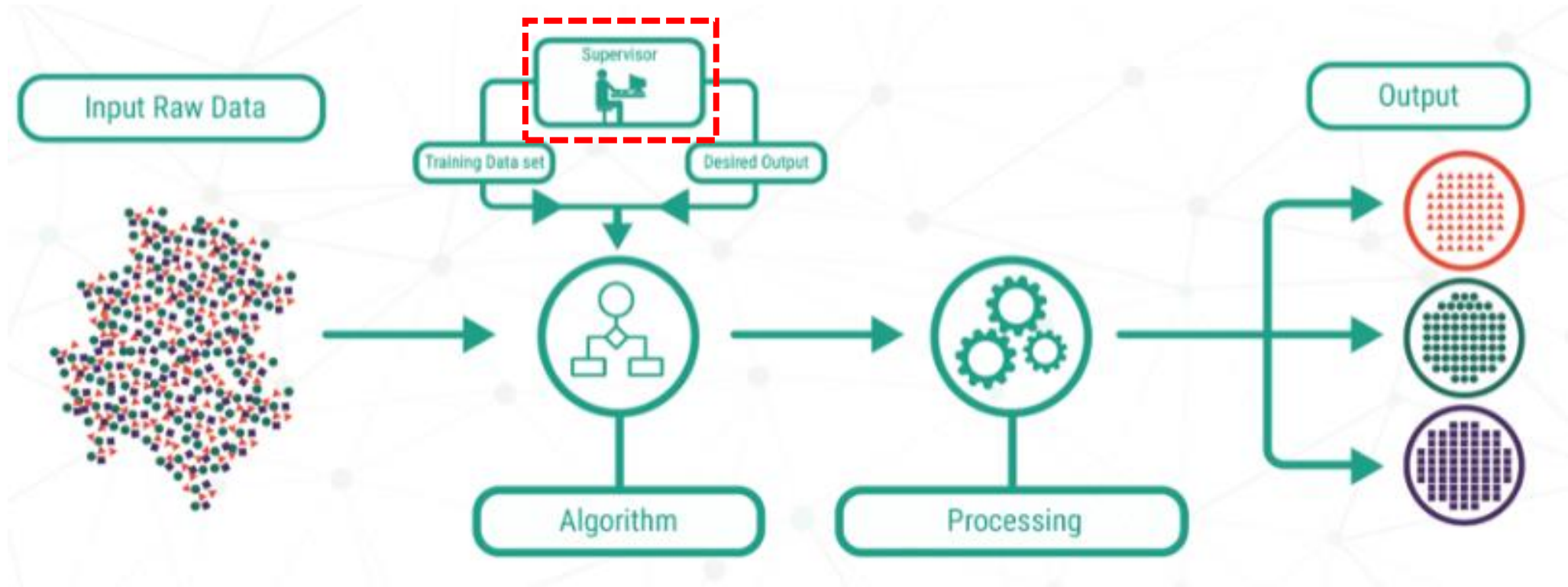
데이터

- 데이터는 수단인가? 아니면 목적인가?
 - 일부 기업은 좋은 데이터를 확보하기 위함을 목적으로 서비스를 판매



데이터

- 성능 좋은 AI 모델도 중요하지만 더 중요한 것은 양질의 많은 데이터 수집이 더 중요



데이터 사이언스

- Computer Science & Statistics
 - Once upon a time ...
 - There were computer science and statistics, as two separate communities
 - Analyzing data was relatively simple (statistics was enough)
 - There was no 'data science'
 - As computer and Internet become widespread, data has been generated and delivered from more Various, numerous entities
 - Data became too big to be processed by and stored in a computer

데이터 사이언스

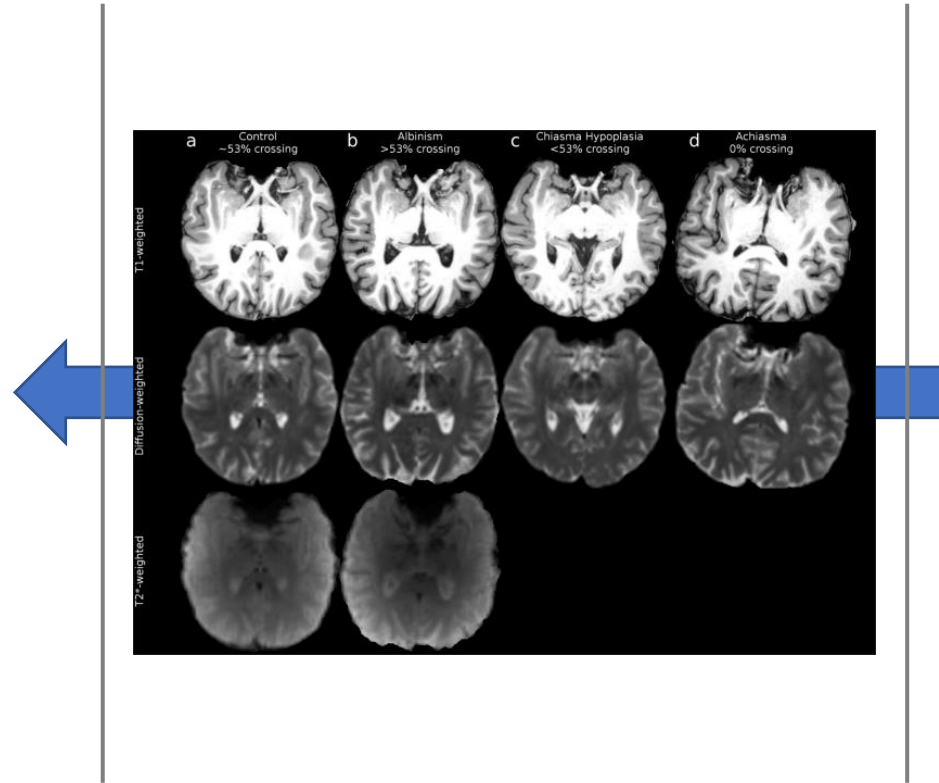
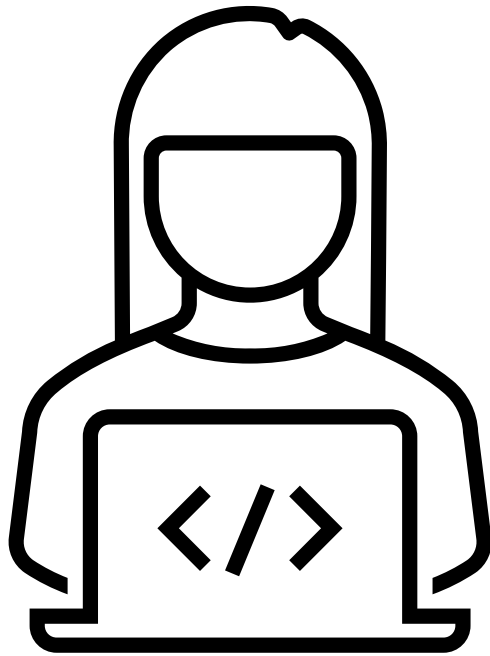
- To handle and analyze Big data, we need people who can do both
 - **Computing**: Working with large amounts of data
 - Collecting, storing, transforming and organizing, ...
 - **Statistics**: Drawing conclusions from data
- There were not a lot people that could do both...
 - They made a new term to describe what they do : **Data Science**
- Data Science
 - Drawing **meaningful conclusions and insights** from **data** using **computation**

데이터 사이언스

- Computer Science & Statistics & Applications
 - Data science is driven by applications (not by technologies)
 - Problems come from an application
 - Data comes from an application
 - Conclusions driven by data are used for the application
 - Purely technical data scientists don't know
 - What data is important, what pattern is interesting, what conclusion is useful...
 - Every data-driven subject brings new challenges
 - Different applications need different techniques

데이터 사이언스

- 데이터 사이언스의 효과를 보려면 서로 협업을 반드시 틀을 깨고 협업을 해야 함



데이터 사이언스

- Looking forward...
 - We would need people who can do all the three, CS, Statistics and Applications
 - At least enough knowledge to collaborate with other experts
- Every single domain is an application area of data science
 - Medicine, Engineering, Laws, Economics, Education, Music, Art, ...
- This is why it may be useful for YOU to take this course, especially for freshman

데이터사이언티스트

- 데이터 사이언티스트가 갖추어야 할 역량
 - 수학을 잘 할 필요는 없다. 그러나 확률과 통계는 알아야 한다.
 - AI 연구/개발자로 진로를 고민 중이라면 선형 대수와 미분학은 추가로 알아야 한다.
- 알고리즘은 코딩이 아니라 논리적으로 생각하는 힘이다.
 - 빌딩 내 N개의 엘리베이터 CASE
- 스토리텔링 기법을 활용하여 커뮤니케이션 능력을 키워라.
 - 아무리 많이 알고 코딩을 잘해도 다른 사람에게 정확하고 효과적으로 전달하는 능력도 매우 중요하다.
- 모든 문제를 데이터 기반으로 해결하려고 하고, 컴퓨팅 사고에 대한 이해가 필요하다.

데이터 사이언스 : 추론

- 추론 (Inference , Reasoning)
 - 알고 있는 (명시적) 지식을 바탕으로 새로운 (암묵적) 지식을 이끌어 냄
 - 전제 (Premise)와 결론 (Conclusion)으로 구성됨
 - 연역적 추론 (deduction) / 귀납적 추론 (induction)

데이터사이언스 : 연역법 vs 귀납법

- 규칙 : 소개팅을 할 때 A속성의 사람들은 왼쪽에 앉고 B속성의 사람들은 오른쪽에 앉는다.
- A와 B는 무엇일까? **남자는 왼쪽에 앉고 여자는 오른쪽에 앉는다.**



데이터사이언스 : 연역법 vs 귀납법

- 규칙 : 소개팅을 할 때 A속성의 사람들은 왼쪽에 앉고 B속성의 사람들은 오른쪽에 앉는다.
- A와 B는 무엇일까? **남자는 왼쪽에 앉고 여자는 오른쪽에 앉는다.**



데이터사이언스 : 연역법 vs 귀납법

- 규칙 : 소개팅을 할 때 A속성의 사람들은 왼쪽에 앉고 B속성의 사람들은 오른쪽에 앉는다.
- A와 B는 무엇일까? 어두운 색상의 상의를 입은 사람이 왼쪽에 앉고, 상대적으로 밝은 색상의 상의를 입은 사람이 오른쪽에 앉는다.



데이터사이언스 : 연역법 vs 귀납법

- 규칙 : 소개팅을 할 때 A속성의 사람들은 왼쪽에 앉고 B속성의 사람들은 오른쪽에 앉는다.
- A와 B는 무엇일까? 어두운 색상의 상의를 입은 사람이 왼쪽에 앉고, 상대적으로 밝은 색상의 상의를 입은 사람이 오른쪽에 앉는다.



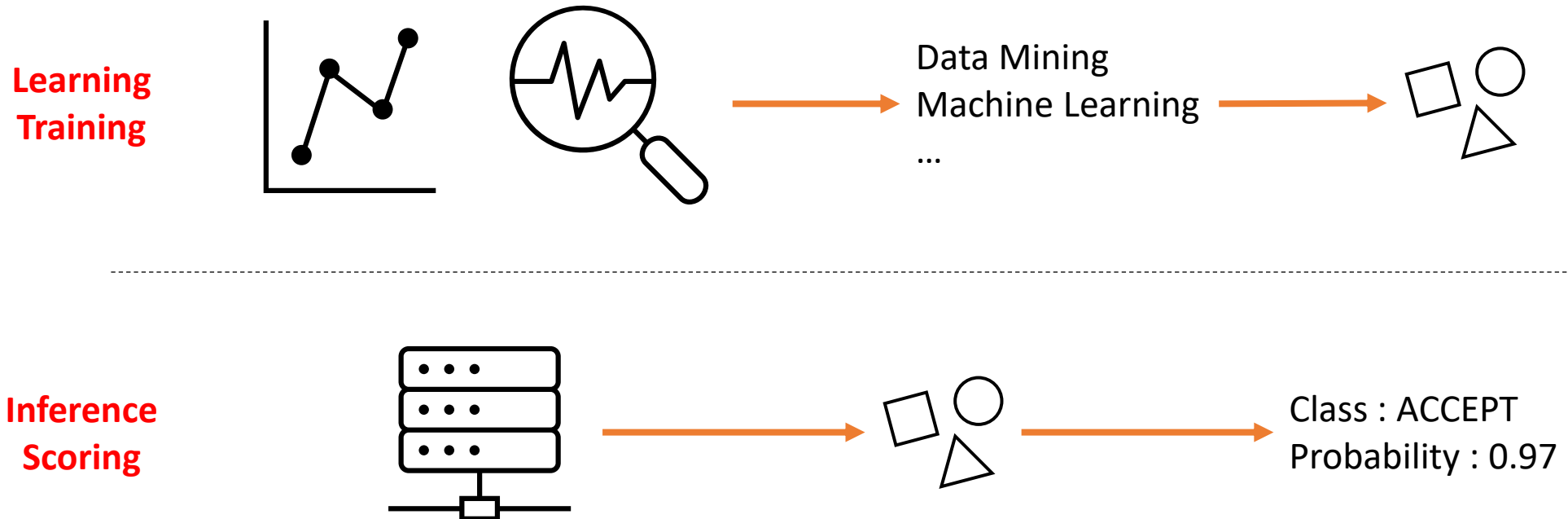
데이터사이언스 : 연역법 vs 귀납법

- 귀납법 : 다양한 관찰된 데이터 (사실)를 바탕으로 여기에 내포되어 있는 일반적인 원리를 추론하는 방법
- 연역법 : 일반적인 가설이나 이론을 전제로 하여 개별적인 특수한 사실이나 원리를 결론으로 추리하는 방법



데이터 사이언스 : 추론

- 데이터 사이언스에서의 추론
 - 주어진 데이터로부터 패턴을 찾아냄 (귀납적 추론)
 - 통계적 확률에 기반
 - 귀납의 오류 발생 가능성 존재
 - 객관적 지표를 통해 추론의 정당성을 검증해야 함

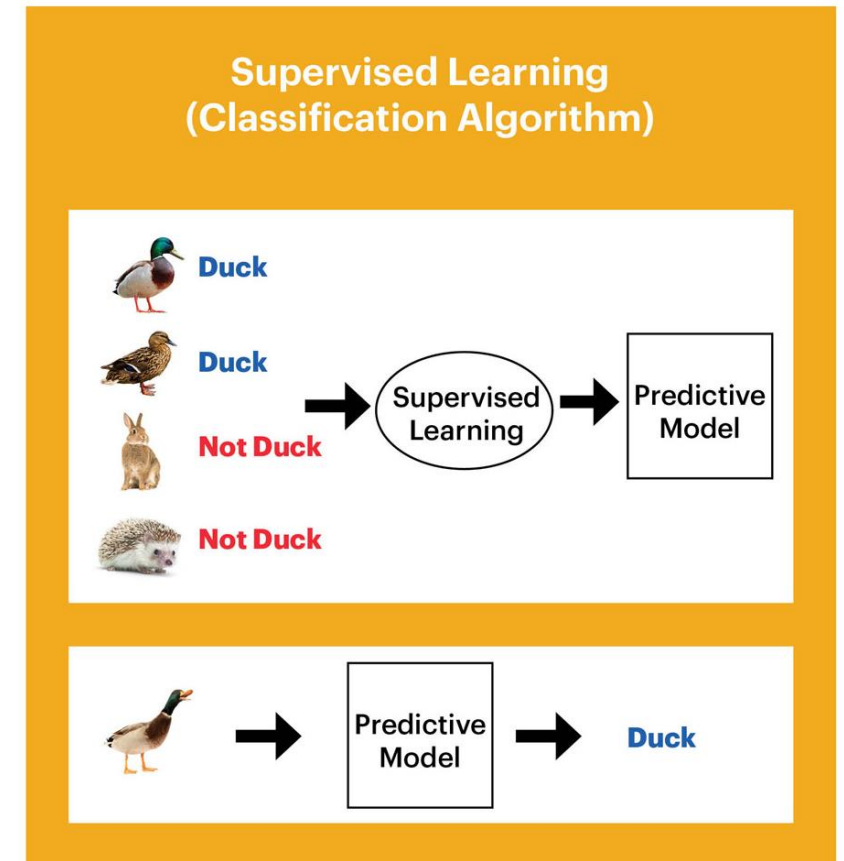


데이터 사이언스 : 학습

- 데이터 사이언스의 학습 (Learning / Training)
 - 지도학습 (Supervised learning)
 - 학습 데이터마다 레이블(Label)을 가짐
 - 비지도학습 (Unsupervised learning)
 - 학습 데이터가 레이블을 가지고 있지 않음
 - 준 지도학습 (Semi-supervised learning)
 - 학습 데이터가 약간의 레이블을 가지고 있음

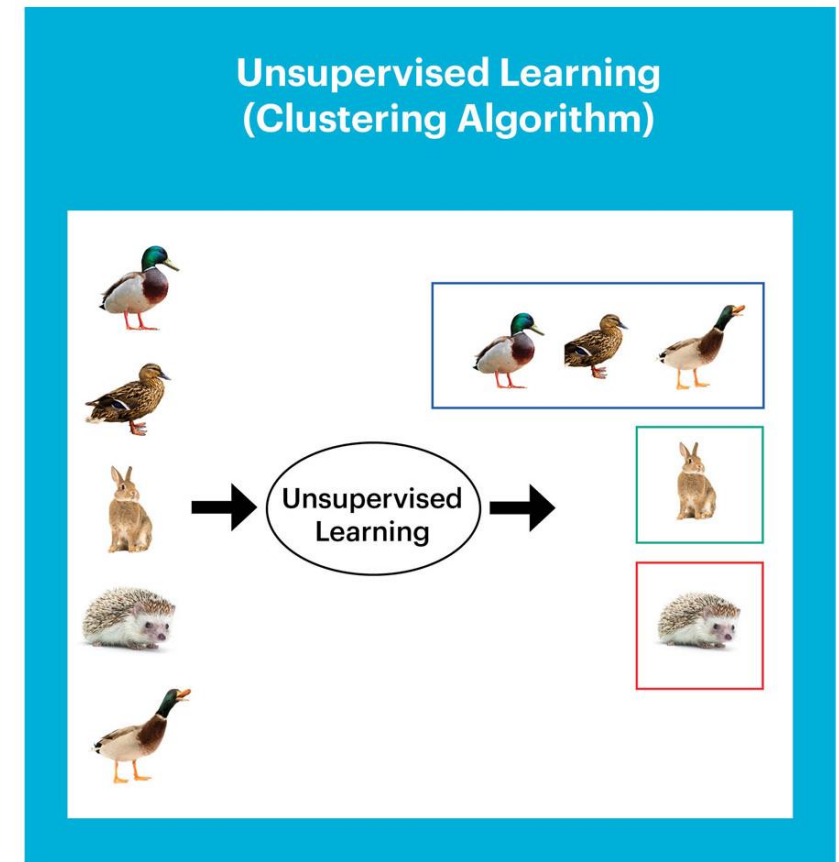
데이터 사이언스 : 학습

- 데이터 사이언스의 학습 (Learning / Training)
 - 지도학습 (Supervised learning) : 정답 (ground - truth)이 주어진 데이터에서 가능
 - 주어진 입력-레이블 쌍들을 맵핑하는 함수를 학습
 - $D=\{X,Y\}$ 로부터 $F(X)=Y$ 를 만족시키는 함수 F 학습
 - 새로운 입력 x' 의 출력을 예측
 - 1) Classification 문제
 - $F(X)$ 가 이산적 (Discrete)
 - 2) Regression 문제
 - $F(X)$ 가 연속적 (Continuous)
 - 3) Estimation 문제
 - $F(X)$ 가 X 의 확률 $P(X)$
 - 장점: 비교적 정확한 학습이 가능
 - 단점: 사용 가능한 데이터의 한계/부족

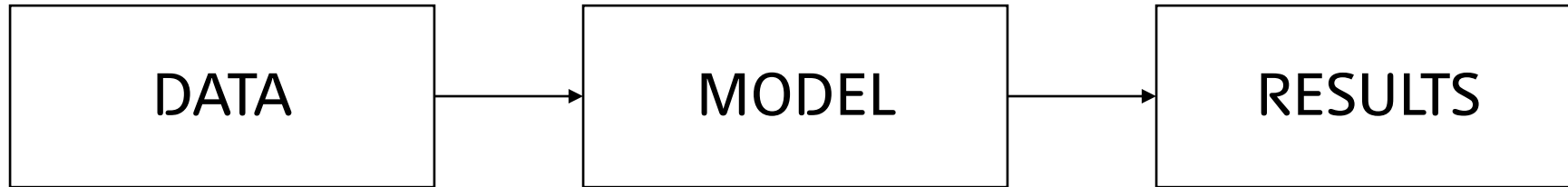


데이터 사이언스 : 학습

- 데이터 사이언스의 학습 (Learning / Training)
 - 비지도학습 (Supervised learning) : 정답 (ground – truth)이 주어진 데이터에서 가능
 - 입력만 있고 레이블은 없는 상태에서 학습
 - $D=\{X\}$ 로부터 $F(X)=X$ 를 만족시키는 함수 F 학습
 - 데이터에 내재된 고유의 특징을 탐색
 - Ex) YouTube 비디오 자동 항목 분류
 - Clustering
 - 비슷한 데이터끼리 묶음
- 지도학습 대비 학습의 난이도 높음
- 대부분의 데이터는 레이블이 없음
- 현재 ML 연구의 방향



데이터사이언스 주요 개념

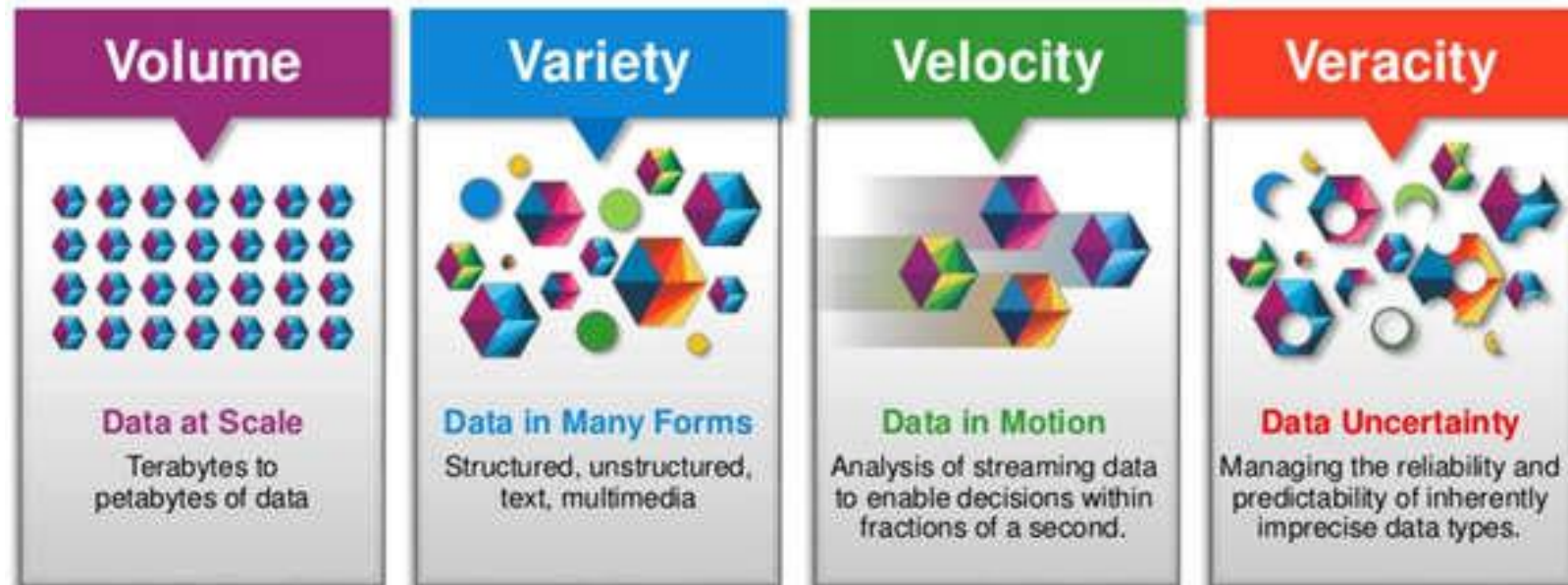


데이터사이언스 : 빅데이터



- 빅데이터

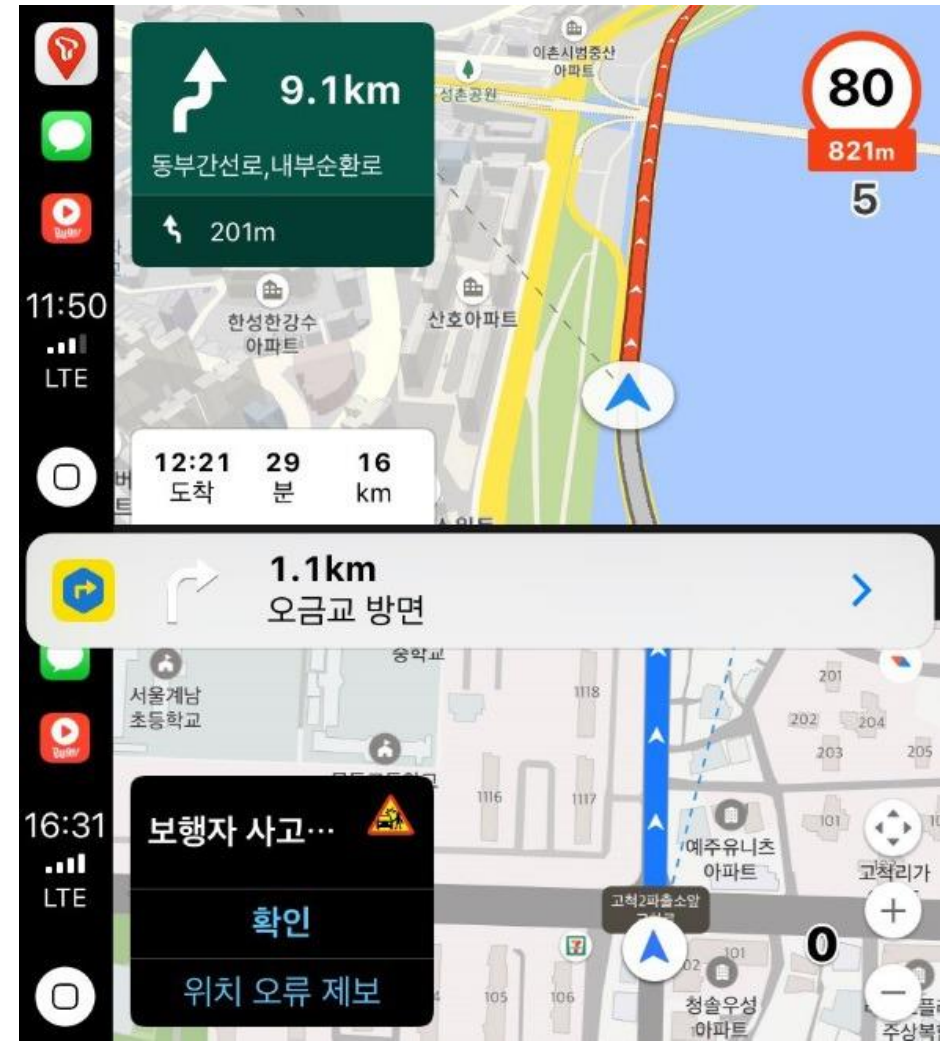
- 위키피디아 정의 : 기존 데이터베이스 관리도구의 능력을 넘어서는 대량(수십 TB)의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술
- 4V : 방대한 양 (Volume), 빠른 데이터 생성 및 처리 속도 (Velocity), 다양한 형태 (Variety), 진실성 (Veracity)
- 7V : 4V + 정확성 (Validity), 휘발성 (Volatility), 잠재 가치 (Value)



데이터사이언스 : 빅데이터

- 빅데이터

- 빅데이터를 처리할 수 있는 시스템이나 알고리즘도 중요하지만 데이터 그 자체가 매우 큰 가치가 있음
- 예시 : 휴대폰 네비게이션 어플리케이션 사용을 차량 자체 탑재된 네비게이션보다 더 많이 사용



데이터사이언스 : 빅데이터

- 빅데이터
 - 빅데이터를 처리할 수 있는 시스템이나 알고리즘도 중요하지만 데이터 그 자체가 매우 큰 가치가 있음
 - 예시 : 수험생 회원 수가 많은 인터넷 교육 사이트의 등급 컷 점수가 더 정확

고3 등급컷

배치표

<

이투스

대성마이맥

진학사

유웨이

종로학원하늘교육

메가스터디

스카이에듀

비

>

과목	1등급			2등급			3등급		
	원점수	표준점수	백분위	원점수	표준점수	백분위	원점수	표준점수	백분위
국어	94	130	97	87	124	89	79	118	77
수학가	91	128	97	84	122	88	77	117	76
수학나	75	142	96	62	130	89	46	114	77
영어	90	-	-	80	-	-	70	-	-
경제	45	75	96	35	66	89	22	55	77
법과정치	47	69	96	42	65	88	34	59	76
사회문화	42	67	96	38	63	88	33	59	77

데이터사이언스 : 빅데이터



- 빅데이터의 세분화 필요성 대두
 - 이제는 빅데이터 라고 하기 보다는 데이터 성격에 따라 세분화해야 보다 깊이 있는 통찰력을 얻을 수 있음
- 4가지 종류의 세분화 데이터
 - 스몰 데이터
 - 패스트 데이터
 - 다크 데이터
 - 스마트 데이터

데이터사이언스 : 스몰데이터



- 스몰 데이터

- 개인의 취향과 라이프스타일 등 사소한 행동에서 나오는 개인화 데이터
- 빅데이터는 한 방향에서 고객을 바라볼 수 있다면 스몰데이터는 모든 방향에서 고객을 살필 수 있음
- 초개인화 트렌드와 데이터 3법 통과에 따라 스몰데이터의 중요성 증가

초개인화
트렌드

데이터
3법 통과

데이터사이언스 : 스몰데이터

- 스몰 데이터

- 예시 : 기존 카드사

- 기존에는 카드사가 보유중인 고객데이터 분석
 - 성별/연령을 통한 세그먼트 마케팅 (segment marketing)
- 최근에는 모든 개인 금융 데이터를 수집하여 고객데이터 분석
 - 개인 맞춤형 재무 설계
 - 개인 최적화 금융상품 추천
- 초 개인화 서비스 가능



데이터사이언스 : 패스트데이터



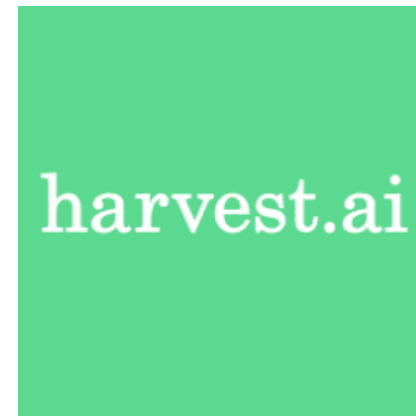
- 패스트 데이터
 - 동영상 재생에 대한 시청자의 반응과 같이 실시간으로 유입이 이뤄지는 데이터
 - 간단하게 빅데이터의 실시간 분석을 강조한 개념
 - 수명이 짧은 페스트 데이터를 실시간으로 분석하여 유의미한 정보 도출
 - 빠른 경영 판단에 기여
 - 예시 : 운송 분야의 화물추적 시스템

데이터사이언스 : 다크데이터



- 다크 데이터

- 활동 과정에서 수집, 처리 저장되었지만 유용하지 않은 정보로 판단되어 활용되지 않는 정보 자산
- 다양한 활동을 통해 일단 저장되기는 하지만 사용 목적이 명확하지 않은 데이터
- 검색 로그 기록, 사용하지 않는 문서, 백업 데이터 등
- 전세계 데이터 중 80% 이상이 다크데이터
- 최근 AI 등의 기술로 방대한 데이터에서 유의미한 정보를 찾아내 분석이 가능해짐에 따라 다크 데이터를 분석, 활용하고자 하는 움직임 증가
- 예시 : 아마존이 harvest.ai 인수 (harvest.ai : 다크데이터를 분석하고 사이버 보안 위험 예방 기술을 보유)



데이터사이언스 : 스마트 데이터



- 스마트 데이터
 - 빅데이터에 비해 용량은 작으나 바로 실질적인 분석을 할 수 있는 양질의 데이터
 - 기업에 주는 의미가 분명하고 바로 활용할 수 있는 데이터
 - 단순 수집된 빅데이터에서 정확하고 의미 있는 정보를 추출하여 조직에서 바로 분석하고 활용할 수 있도록 신속하게 제공하는 기술을 통해 스마트 데이터 획득 가능
 - 예시 : 핀테크 계열의 활용
 - 빅데이터 : 기존 은행들에게 불필요한 자료 모두 제공
 - 스마트 데이터 : 데이터 내 패턴 및 동향을 파악해 핵심 데이터를 제공
- 정리
 - 스몰 데이터 : 개인 행동 데이터
 - 패스트 데이터 : 실시간 처리와 분석이 필요한 데이터
 - 다크 데이터 : 수면 아래 잠겨 있는 데이터
 - 스마트 데이터 : 양보다는 질을 갖춘 데이터

데이터사이언스 : 데이터마이닝

- 데이터 마이닝

- Data + Mining

- 대량의 데이터로부터 의미 있는 규칙이나 패턴을 추출하는 일련의 행동



데이터사이언스 : 데이터마이닝



- 데이터 마이닝

- “대량의 데이터 집합으로 부터 유용한 정보를 추출하는 것” (Han et al 2001)
- “데이터 마이닝이란 의미 있는 패턴과 규칙을 발견하기 위해서 자동화 되거나 반자동화 된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정 (Berry and Linoff, 1977, 2000)
- “데이터 마이닝은 통계 및 수학적 기술 뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정” (Gartner Group 2004)

데이터사이언스 : 데이터마이닝



- 데이터 마이닝 예시

- 휴대폰 요금제를 수정할 경우 어느 연령대 고객이 반응할 가능성이 높은가?
- 홈쇼핑을 시청하고 있는 고객 중 전화가 아닌 휴대폰 어플리케이션을 통해 구매하는 고객은 누구 인가?
- 구독 서비스를 3개월 이내 탈퇴할 것 같은 고객들은 누구 인가?
- 마트에서 동시에 구매하는 물품은 어떤 것인가?

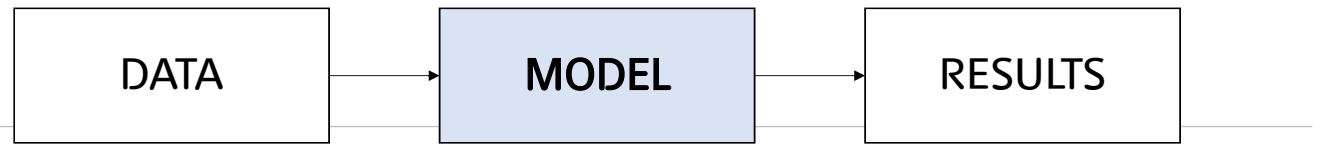


데이터사이언스 : 데이터마이닝

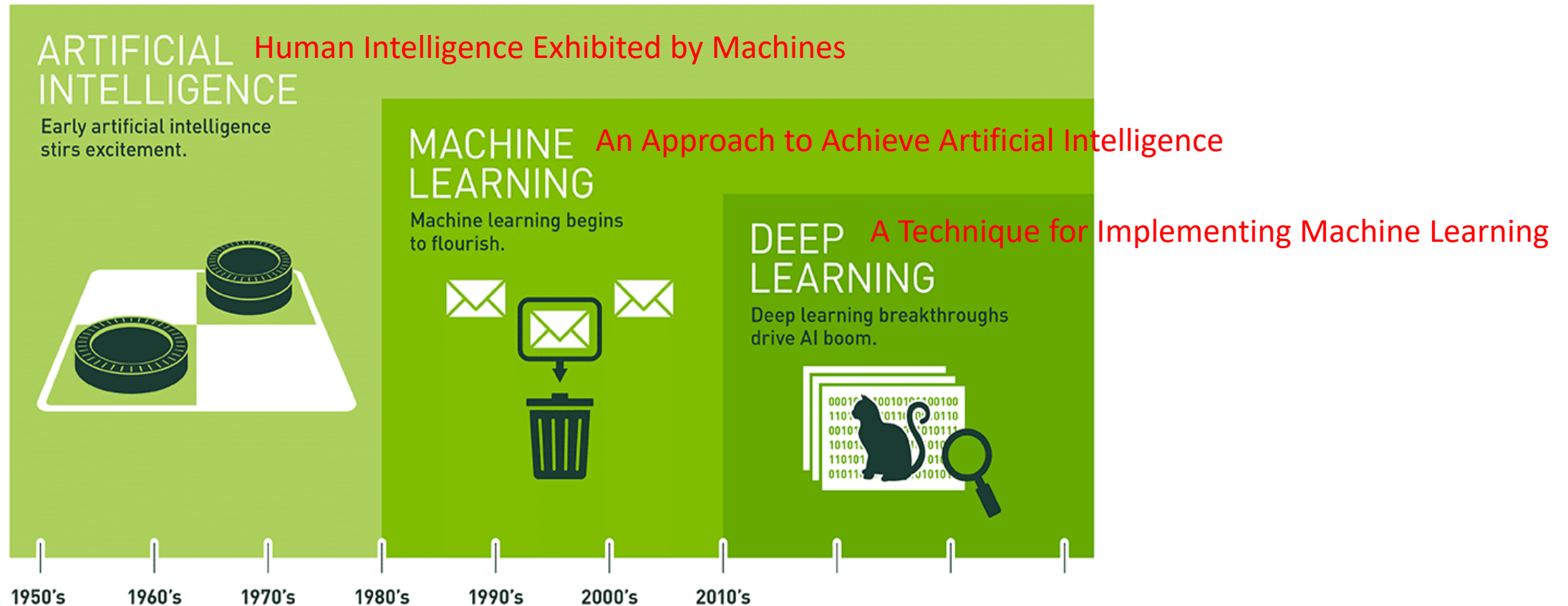


- 데이터 마이닝의 활용
 - 연관 규칙 (association)
 - 데이터 마이닝의 가장 대표적인 기술로 쇼핑몰 등에서 한 번에 구매한 상품들에 관한 연관규칙을 발견
 - 분류 (classification)
 - 새로운 데이터가 있을 때, 이것이 기존의 어떤 유형의 집합에 속하게 될 것인지 예측
 - 예) c카드사는 도난이나 분실카드의 경우 같은 시간대에 여러 번의 결제 행위가 일어난다는 사실을 발견하고 이와 같은 결제시도가 있을 시 가맹점에 전화를 걸어 본인 여부를 확인하는 방법으로 도난이나 분실카드로 인한 손실을 연간 30%이상 줄임
 - 군집 (clustering)
 - 군집화 기술은 전체 데이터의 분포 상태나 패턴 등을 찾아내는 데 유용하게 사용
 - 분류와 다른 점은 각 집합에 해당되는 특징 등과 같은 정보가 제공되지 않음

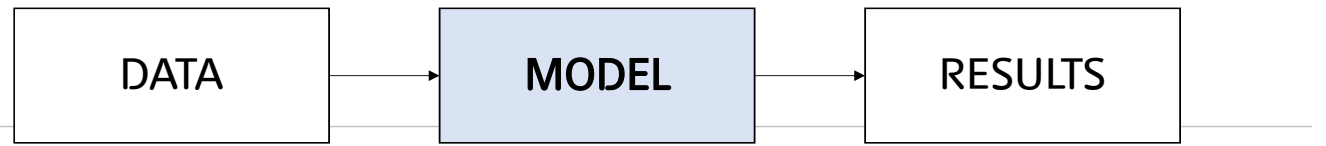
데이터사이언스 : 머신러닝



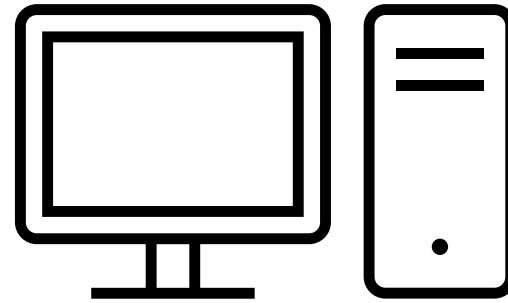
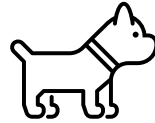
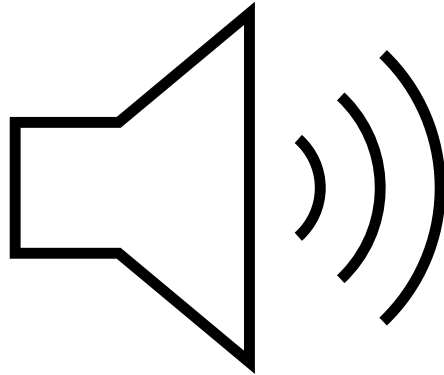
- 인공지능 (Artificial Intelligence)
 - 인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램으로 실현한 기술을 의미
 - 인공지능은 인간의 지능으로 할 수 있는 사고, 학습, 자기 계발 등을 컴퓨터가 스스로 수행할 수 있도록 방법을 연구하는 컴퓨터 공학 및 정보기술의 한 전문지식 분야임



데이터사이언스 : 머신러닝



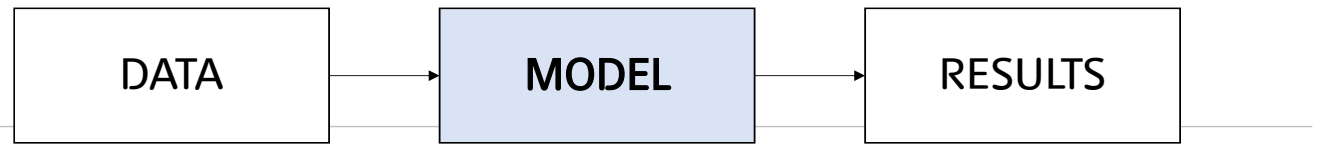
- 인공지능과 머신러닝
 - 기존의 인공 지능 프로그램 : 규칙과 데이터를 입력하면 답을 내놓는 방식



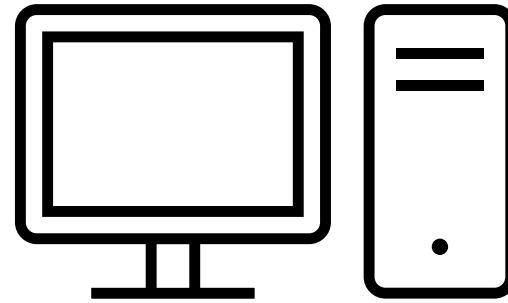
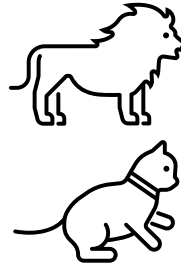
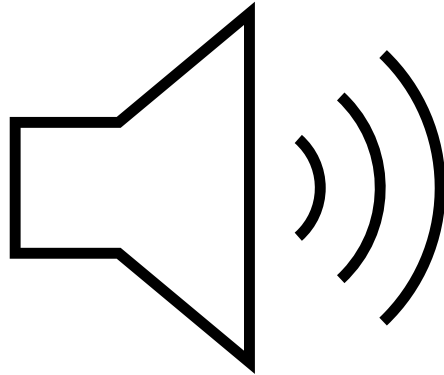
네 발이 달리고 꼬리가 있고,
털이 있고 땡땡하게 생긴 것은
강아지야.

강아지네

데이터사이언스 : 머신러닝



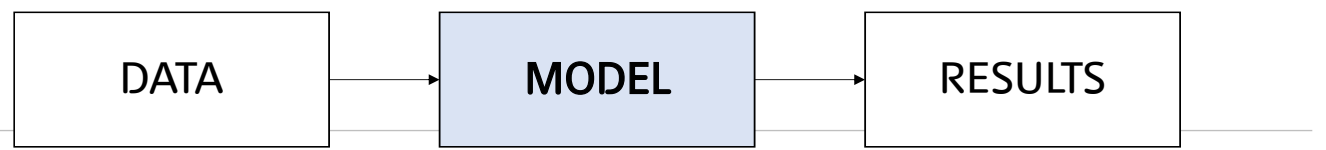
- 인공지능과 머신러닝
 - 기존의 인공 지능 프로그램 : 규칙과 데이터를 입력하면 답을 내놓는 방식



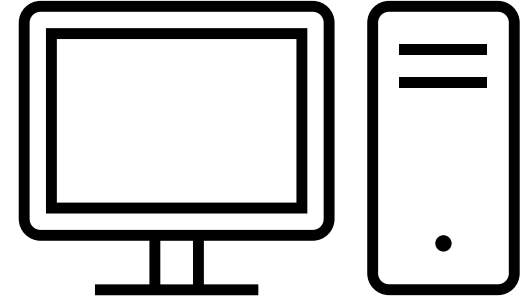
네 발이 달리고 꼬리가 있고,
털이 있고 땡땡하게 생긴 것은
강아지야.

이것도 강아지네

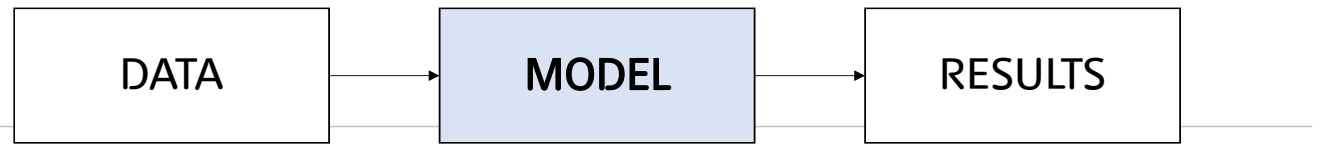
데이터사이언스 : 머신러닝



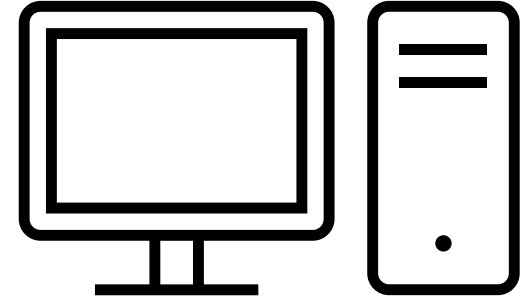
- 인공지능과 머신러닝
 - 최근의 머신러닝은 많은 데이터를 학습하면서 스스로 패턴을 찾음



데이터사이언스 : 머신러닝

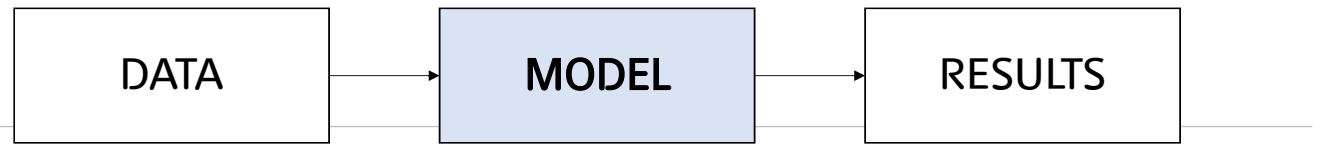


- 인공지능과 머신러닝
 - 최근의 머신러닝은 많은 데이터를 학습하면서 스스로 패턴을 찾음



강아지네

데이터사이언스 : 머신러닝

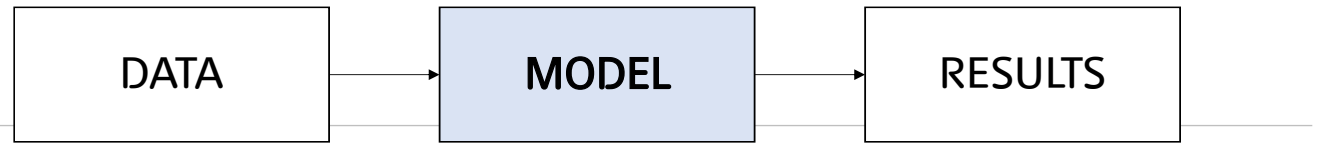


- 인공지능과 머신러닝
 - 최근의 머신러닝은 많은 데이터를 학습하면서 스스로 패턴을 찾음



강아지 = 네 발
털 있음
눈 두 개
댕댕함

데이터사이언스 : 머신러닝



- 인공지능과 머신러닝
 - 최근의 머신러닝은 많은 데이터를 학습하면서 스스로 패턴을 찾음



네 발 **O**

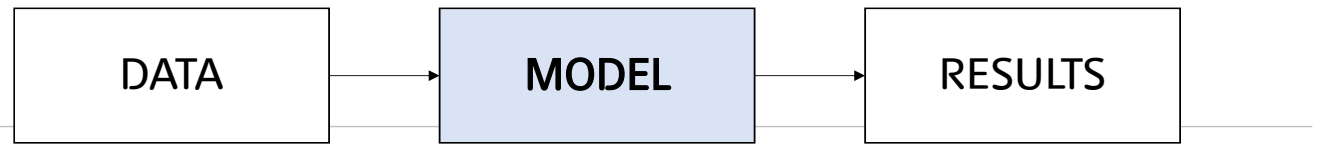
털 있음 **O**

눈 두 개 **O**

댕댕함 ? **X**

강아지 아니네

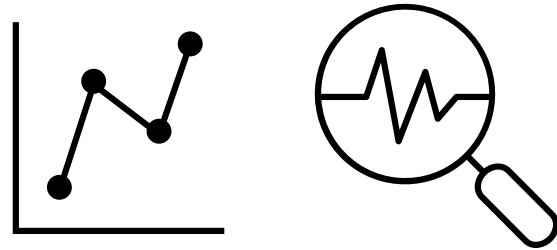
데이터사이언스 : 머신러닝



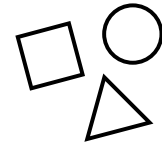
- 머신러닝 (기계 학습)

- 기계가 명시적으로 코딩되지 않은 동작을 스스로 학습해 수행하게 하는 연구 분야
- 인공지능의 한 분야로 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기술과 기법을 의미
- 데이터 마이닝은 데이터 간의 미처 몰랐던 속성을 발견하는 것에 집중하는 반면, 머신러닝은 훈련 데이터를 통해 학습된 알려진 속성을 기반으로 예측하는 데 초점

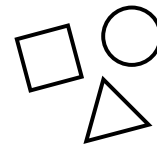
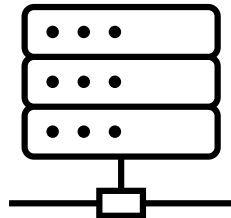
**Learning
Training**



Data Mining
Machine Learning
...

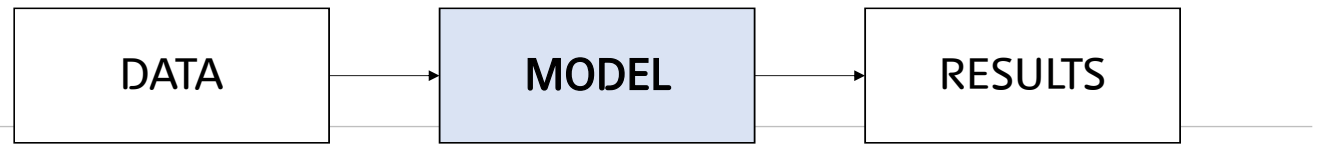


**Inference
Scoring**



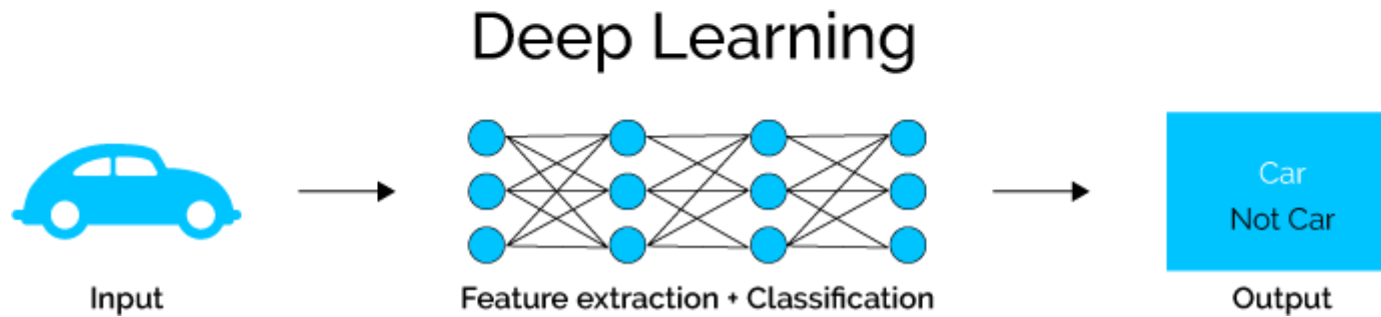
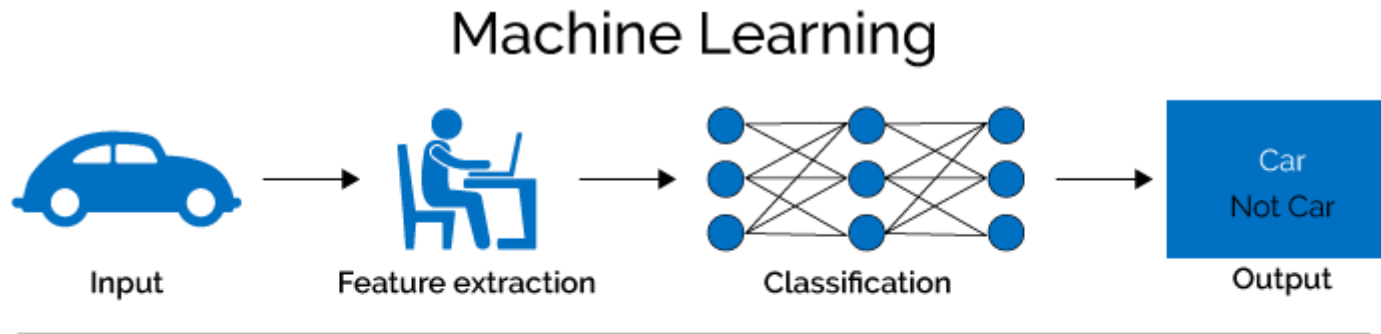
Class : ACCEPT
Probability : 0.97

데이터사이언스 : 딥러닝



- 딥러닝

- 인공 신경망을 잇는 기계 학습법으로 사물이나 데이터를 군집화 하거나 분류하는데 사용하는 기술
- 구글은 음성인식과 번역을 비롯해 로봇의 인공지능 시스템 개발에도 딥러닝 기술을 도입



데이터사이언스 : 사물인터넷 (Internet of Things (IoT))

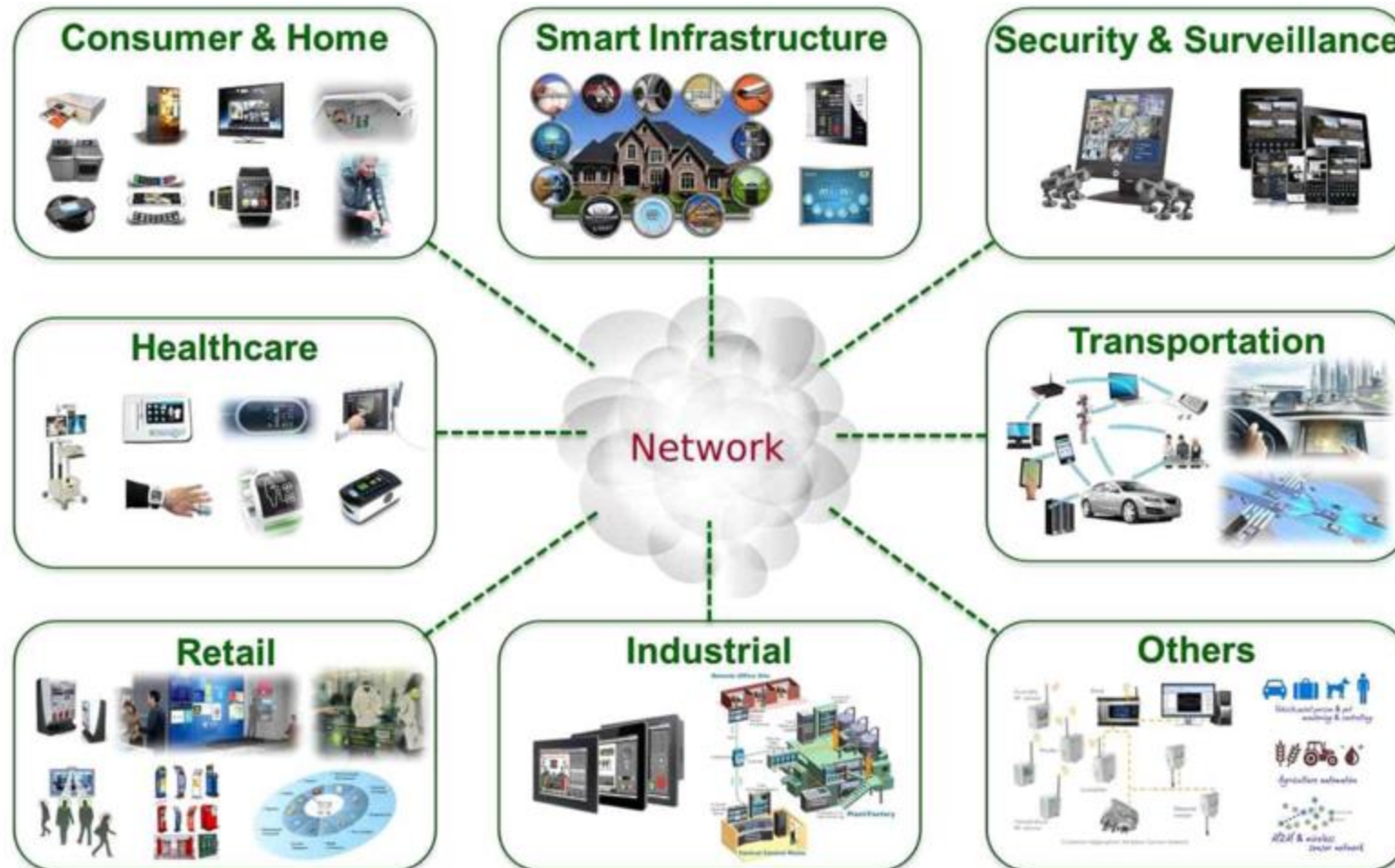
Connected
Things

DATA

MODEL

RESULTS

- 사물 인터넷 : 각종 사물에 센서와 통신 기능을 내장하여 인터넷에 연결하는 기술
 - 다양한 사물에서 얻어진 다량의 데이터를 수집, 처리, 분석하는 업무 필요
 - 2개 이상의 사물들이 서로 연결됨으로써 개별적인 사물들이 하지 못했던 새로운 기능들을 제공할 수 있게 됨



감사합니다

kimtwan21@dongduk.ac.kr

김 태 완