



Digital Asset and Real Applications of Data Science

Contents

- I. 데이터 비즈니스란 무엇인가?
- II. 삼성의 데이터 비즈니스 필요성과 사례
- III. Platform of AI Center

Kyungwon Kim

Assistant Professor
Department of International Trade
College of Global Political Science and Economics
Incheon National University

September 1, 2021

Contents

I. 데이터 비즈니스란 무엇인가?

- 1) 데이터 분석 프로세스
- 2) (과거) 프로젝트 리스트
- 3) 현실적인 데이터 분석

II. 삼성의 데이터 비즈니스 필요성과 사례

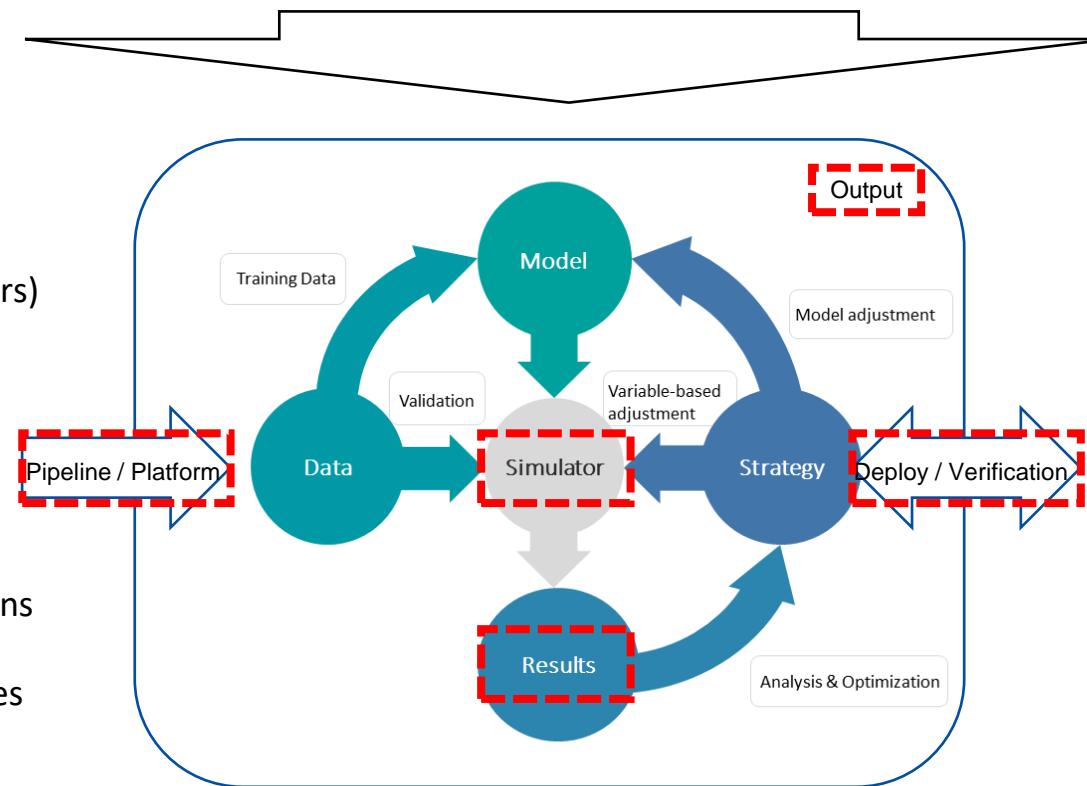
- 1) 삼성전자 구성원
- 2) 데이터 분석 협업 예시
- 3) 데이터 분석 목표
- 4) 데이터 현황
- 5) 데이터 분석/개발 이슈
- 6) 현재 프로젝트 리스트
- 7) Data Analytics Lab 추진방향
- 8) Data Analytics Lab 선행 기술

III. Platform of AI Center

- 1) 대내외 환경
- 2) AI Center

Leverage every possible data asset in the Samsung Ecosystem
to drive the next business innovation breakthrough
as well as to deliver engagement and customer royalty

- ✓ Market conditions (including competitors)
- ✓ Economic indicators (nonfarm payroll, unemployment, CPI, exchange..)
- ✓ Sales and campaigns
- ✓ Retailer and supply
- ✓ Marketing/Promotions
- ✓ Enterprise
- ✓ Customer preferences and demographics

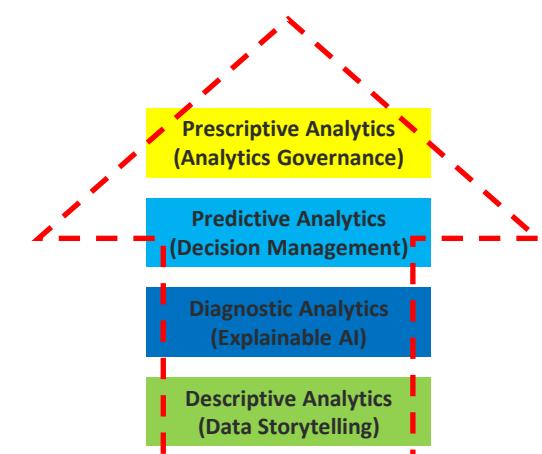


- Financial Management
- Strategy and Product Planning
- Enterprise
- HR and Operation
- Sales and Marketing
- Legal and Regulatory
- Supply Chain
- Quality Management
- R&D and IT
- Manufacturing

프로젝트 리스트

➤ Projects: “Macroscopic Thinking” + “Communication” + “Not Practice, But Production”

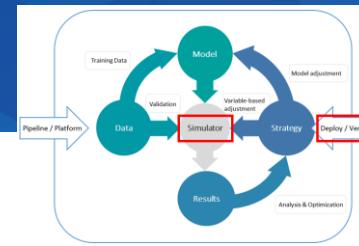
- Dataset : Customer / Service / Sales / Marketing / Ads. / HR / IoT / Products / Quality
- Project @ AI Center of Samsung Research (2017.09 ~) :
 - 1) 실시간 채널 별 광고효과 추론 및 최적 입찰(Bidding) 가격 예측 (2020.01.~)
 - 2) 마케팅 프로모션 효과 증대를 위한 광고추천분석 및 매출기여 효과검증 (2019.05.~2020.02.)
 - 3) 제품 별 부품수급 비용 최적화 및 부진재고 최소화를 위한 공정 최적화 (2019.01.~2019.12.)
 - 4) 빅데이터 기반 글로벌 마케팅 브랜드/프로모션/홍보/광고/영업/소셜/경쟁사동향 등의 매출기여도 효과 분석 및 실시간 최적 투자 포트폴리오 전략 (2018.01.~2019.06.)
 - 5) 매출 최대화 및 비용 최소화를 위한 제품 판매 수요 중장기 예측 (2018.01.~2018.09.)
 - 6) 동남아 시장 약 4억명 고객 대상 프로모션 효과 및 ROI 향상 예측 (2018.01.~2018.09.)
 - 7) 모바일 생산 자동화를 위한 베트남 공장 데이터 플랫폼 기반 최적화 (2017.09.~2018.06.)
- Project @ BigData Lab of Samsung Electronics (2014.04~ 2017.08) :
 - 1) 고객 불만 사전대응 및 감소를 위한 고객voc 경보시스템 구축 (2017.01.~2017.12.)
 - 2) 모바일 불량 원인 파악을 위한 생산라인 이상기기 파악 및 수명 예측 (2016.12.~17.04.)
 - 3) 개인화 맞춤 프로모션을 위한 마케팅 전략 (2016.10.~2016.12.)
 - 4) 16년 기능/앱 사용성 향상을 위한 15년 사용성 분석 (2016.01.~2016.12.)
 - 5) 제품 기능/앱 사용유사성 기반 고객 프로모션 전략기획 (2015.03.~2015.12.)
 - 6) 광고효과 증대를 위한 소비자 나이/성별 예측 및 컨텐츠 추천 (2014.10.~2015.06.)
 - 7) 기존 제품 사용성 개선을 위한 기능/앱 이탈자 파악 (2014.05.~2014.12.)



과거 프로젝트 리스트

1. Forecasting of price strategy of competitors to expand product sales
(with Corporate SCM Group(CSCM), Feb.2017~Aug.2017)
2. Forecasting of customer's VOC growth for reducing complaints in advance
(with Global CS Center(GCS), Jan.2017~Aug.2017)
3. Confirm and improve confidence interval in product quality of manufactured goods to identify the cause of defects
(with Procurement Planning Group[Visual Display], Dec.2016~Mar.2017)
4. Planning of market strategy for personalized promotion of customers
(with Online Marketing Group(Visual Display), Oct.2016~Jun.2017)
5. Analysis of 15-year usability for improving 16-year app and function of devices
(with Service Business Team(Visual Display), Jan.2016~Dec.2016)
6. Strategy planning for multiple purchasing of customers based on similar usage patterns of app/function of products
(with BigData Analytics Group(Mobile), Mar.2015~Dec.2015)
7. Extract user profiling using Machine Learning: age/gender/income
(with BigData Lab(Visual Display), Oct.2014~Jun.2015)
8. Identify the churn customers in app/function usage of product to improve existing environment
(with BigData Analytics Group(Mobile), May.2014~Dec.2014)

과거 프로젝트 리스트



1. Forecasting of price strategy of competitors to expand product sales

(with Corporate SCM Group(CSCM), Feb.2017~Aug.2017)

Q. M/S is lowered, sales is lowered, cost is increased, should TV business be continuing?

A. (Well..) It may be a temporary market impact, but it is not easy!



Q. So, come with your new business.

A. (For now..) The manager is pioneering new display market!

(ex. Digital Signage)



Q. The new business takes a long time. What can we do in existing business?

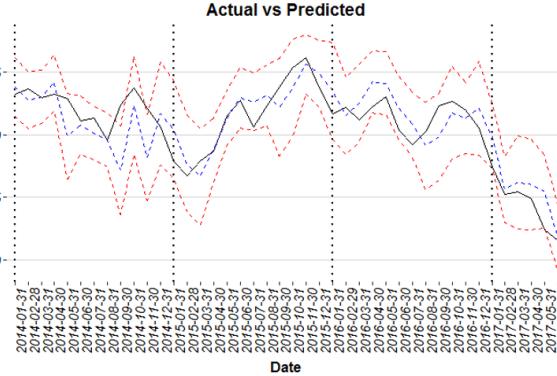
A. (For the moment..) It seems to be a way to advance business by data analysis!

5. Analysis of 15-year usability for improving 16-year app and function of devices

Significant Level on Intuition: 95%

"Difference and (Accuracy %): 0.0527 (83.7362 %) ~ 1.303e-06 (99.9995 %)"

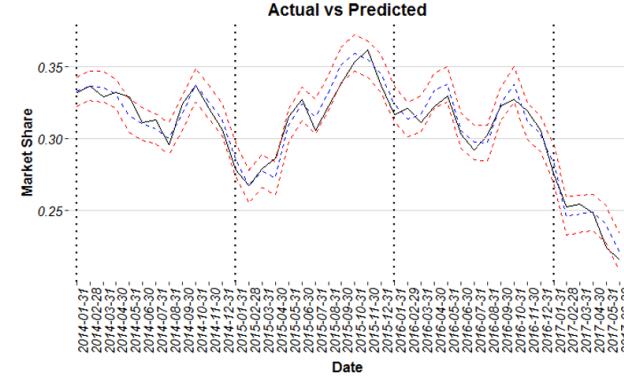
"Averaged Difference (Accuracy %): 0.0135 (95.5520 %)"



Significant Level on Automation: 70%

"Max & Min Difference (Accuracy %): 0.0161 (92.7770 %) ~ 0.000054 (99.9837 %)"

"Averaged Difference (Accuracy %): 0.00605 (97.9871 %)"

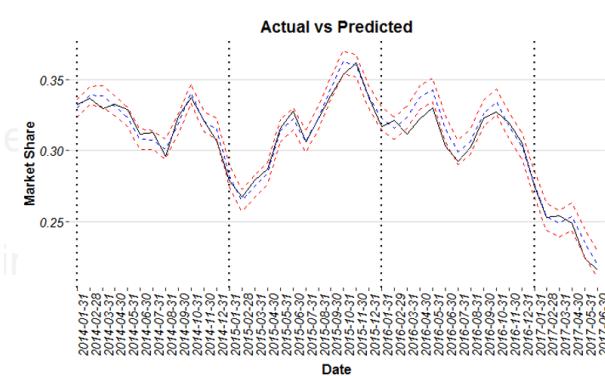


With Google Trend (auto sl: 60%)

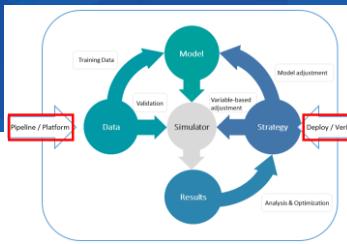
: just use "TV" keyword of each brands'

"Max & Min Difference (Accuracy %): 0.0151 (93.2819 %) ~ 0.000026 (99.9163 %)"

"Averaged Difference (Accuracy %): 0.00486 (98.3931 %)"



과거 프로젝트 리스트



2. Forecasting of customer's VOC growth for reducing complaints in advance (with Global CS Center(GCS), Jan.2017~Aug.2017)

A. (No) The A/S engineer does not reflect the data for the reason of VOC!

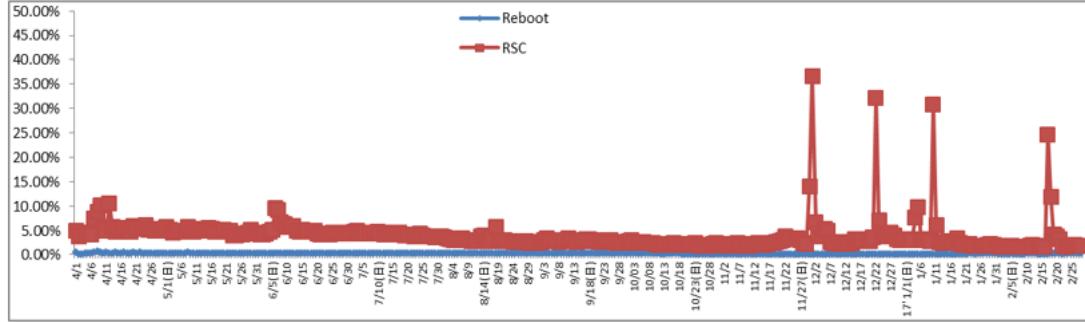


Q. Then, reduce the cost of A/S engineer for fixing VOC.

A. (For a while..) I'll let you know when you should be focusing a specific defect!

3. Confirm and improve confidence interval in product quality of manufactured goods to identify the cause of defects (with Procurement Planning Group, Mar.2017~May.2017)

5. Analysis of 15-year (with Service Business)



6. Strategy planning for app/function of products (with BigData Analytics)

Kruskal-Wallis one-way analysis of variance test

Date	2016-12-16	2016-12-17	2016-12-18	2016-12-19	2016-12-20	2016-12-21
Hour	18	18	18	18	18	18
1Hour Ago	[1] "Stable"					
5Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Level2Warn"	[1] "Stable"	[1] "Stable"	[1] "Stable"
10Hour Ago	[1] "Stable"	[1] "Level2Warn"	[1] "Level3Warn"	[1] "Stable"	[1] "Level2Warn"	[1] "Stable"
15Hour Ago	[1] "Stable"	[1] "Level2Warn"	[1] "Level3Warn"	[1] "Stable"	[1] "Level2Warn"	[1] "Stable"
20Hour Ago	[1] "Stable"	[1] "Level2Warn"	[1] "Level3Warn"	[1] "Stable"	[1] "Level2Warn"	[1] "Stable"
Yesterday	[1] "Stable"	[1] "Level2Warn"	[1] "Level3Warn"	[1] "Stable"	[1] "Level3Warn"	[1] "Stable"
During past 5 days	[1] "Stable"	[1] "Level3Warn"	[1] "Level3Warn"	[1] "Stable"	[1] "Level3Warn"	[1] "Stable"
During past 10 days	[1] "Level2Warn"	[1] "Level3Warn"	[1] "Level3Warn"	[1] "Level1Warn"	[1] "Level3Warn"	[1] "Level3Warn"
During past 15 days	[1] "Level3Warn"					
During past 20 days	[1] "Level3Warn"					

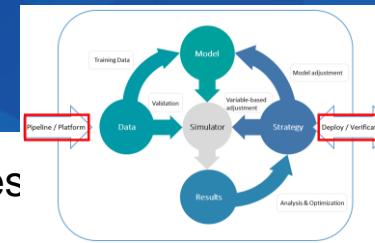
Mann-Whitney U or Wilcoxon rank sum test

Date	2016-12-16	2016-12-17	2016-12-18	2016-12-19	2016-12-20	2016-12-21
Hour	18	18	18	18	18	18
1Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Level1Warn"	[1] "Stable"	[1] "Stable"
5Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
10Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
15Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
20Hour Ago	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
Yesterday	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
During past 5 days	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
During past 10 days	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
During past 15 days	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"
During past 20 days	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"	[1] "Stable"

Modified Cross Method

Date	2016-12-16	2016-12-17	2016-12-18	2016-12-19	2016-12-20	2016-12-21
Hour	18	18	18	18	18	18
1Hour Ago	[1] "Stable"					
5Hour Ago	[1] "Stable"					
10Hour Ago	[1] "Stable"					
15Hour Ago	[1] "Stable"					
20Hour Ago	[1] "Stable"					
Yesterday	[1] "Stable"					
During past 5 days	[1] "Stable"					
During past 10 days	[1] "Stable"					
During past 15 days	[1] "Stable"					
During past 20 days	[1] "Stable"					

과거 프로젝트 리스트



5. Analysis of 15-year usability for improving 16-year app and function of devices (with Service Business Team(Visual Display), Jan.2016~Dec.2016)

Q. Bring the results of the 15-year service effect to allocate 16-year service development costs. A. Customers does not recognize the most services!

(with BigData Analytics Group(Mobile), Mar.2015~Dec.2015)

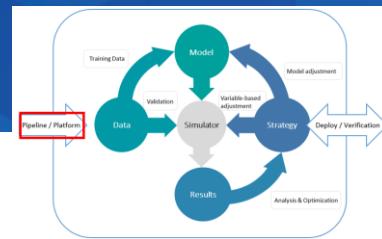
시점	15년1분기		15년2분기		15년3분기		15년4분기		16년1분기		16년2분기	
	KPI항목	순위	KPI스코어	순위								
14_firstscreen_OLD	3	3.05	1	100.00	1	100.00	1	100.00	1	100.00	1	100.00
15_wireless	14	0.00	15	0.00	15	0.00	2	8.00	2	7.72	2	7.79
15_apps	6	0.51	3	2.72	2	4.79	3	3.17	3	1.65	3	1.30
player	1	100.00	2	30.36	3	3.16	4	1.16	4	0.60	4	0.57
15_webbrowser	9	0.09	7	0.96	5	0.61	5	0.64	5	0.56	5	0.50
mycontent	7	0.36	9	0.23	7	0.23	7	0.53	6	0.41	6	0.33
15_tvmenu	4	0.65	5	2.55	4	1.14	6	0.64	7	0.27	7	0.15
15_allsearch	11	0.03	8	0.31	8	0.18	9	0.12	8	0.08	8	0.05
15_sportemode	13	0.00	10	0.08	9	0.05	10	0.03	10	0.02	9	0.02
15_tvnpn	5	0.53	6	1.38	6	0.33	8	0.14	9	0.02	10	0.01
15_voice	8	0.19	11	0.06	11	0.03	11	0.01	11	0.01	11	0.01
15_mls	10	0.04	13	0.03	13	0.02	14	0.01	14	0.00	12	0.00
15_extratweet	14	0.00	14	0.01	14	0.01	13	0.01	12	0.01	13	0.00
15_game	12	0.01	12	0.03	12	0.02	12	0.01	13	0.00	14	0.00
15_motion	2	6.41	4	2.66	10	0.03	15	0.00	15	0.00	15	0.00

시점 KPI항목	전년동기대비편차		16년1분기 순위	KPI스코어	16년2분기 순위	KPI스코어
	16년1분기	16년2분기				
15_wireless	12	-7.720323121	13	-7.79043497		
mycontent	1	-0.051204131	3	-0.098802012		
15_webbrowser	4	-0.472638819	2	0.463814316		
15_sportemode	3	-0.016327198	1	0.064943572		
15_mls	-4	0.040982522	1	0.023559244		
15_extratweet	2	-0.006995801	1	0.008349987		
14_firstscreen_OLD	2	-96.954669893	0	0		
15_apps	3	-1.142720455	0	1.424669424		
15_allsearch	3	-0.047484084	0	0.25787682		
15_voice	-3	0.185098321	0	0.055834864		
15_partymode	-2	-8.78083E-06	-1	-5.17756E-06		
player	-3	99.3989757	-2	29.79495573		
15_tvmenu	-3	0.378135728	-2	2.401857627		
15_game	-1	0.00546263	-2	0.024890103		
networkspeaker	-3	0	-2	0		
14_webbrowser	-3	0	-2	0		
15_url	-3	0	-2	0		
14_sso	-3	0	-2	0		

- ✓ 스마트 TV 고객들의 80%는 회사에서 제공하는 서비스의 13.17%만 사용
 - 전체 평균보다 많이 사용하는 서비스: 14_firstscreen_old(22.32%), 15_wireless (14.39%)
 - 소비자 90%이상의 관심도가 20% 미만인 서비스들은 소비자들이 존재 여부 모를 수 있음
- ✓ 프로모션이 효과를 볼수 있는 TV 서비스가 있는가?
 - 사용확대를 위한 프로모션: 14_firstscreen_old, 15_wireless, 15_remocon
 - 존재성 홍보를 위한 프로모션: 위 3 개를 제외한 모든 서비스

서비스명	하위10%	하위20%	하위30%	하위40%	하위50%	하위60%	하위70%	하위80%	하위90%	전체100%
14_firstscreen_old	0.01%	0.19%	0.72%	1.82%	3.80%	7.15%	12.74%	22.32%	40.80%	100.00%
15_wireless	0.00%	0.00%	0.00%	0.00%	0.29%	1.85%	5.84%	14.39%	33.21%	100.00%
전체서비스	0.00%	0.03%	0.10%	0.30%	0.84%	2.27%	5.70%	13.17%	29.95%	100.00%
15_remocon	0.00%	0.00%	0.00%	0.00%	0.03%	0.42%	1.76%	5.12%	12.55%	29.32%
15_apps	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.19%	1.41%	5.81%	18.37%
15_webbrowser	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.20%	1.62%	8.39%
15_tvmenu	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.77%	7.24%
mycontent	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.14%	5.06%
15_soccer	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
15_sportemode	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.93%
15_smarthub	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
15_voice	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
15_smarthub	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
14_allsearch	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
15_mls	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
15_partymode	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
smartview_window	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
15_virtualchannel	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
smartview_mobile	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

과거 프로젝트 리스트



7. Extract user profiling using Machine Learning: age/gender/income (with BigData Lab(Visual Display), Oct.2014~Jun.2015)

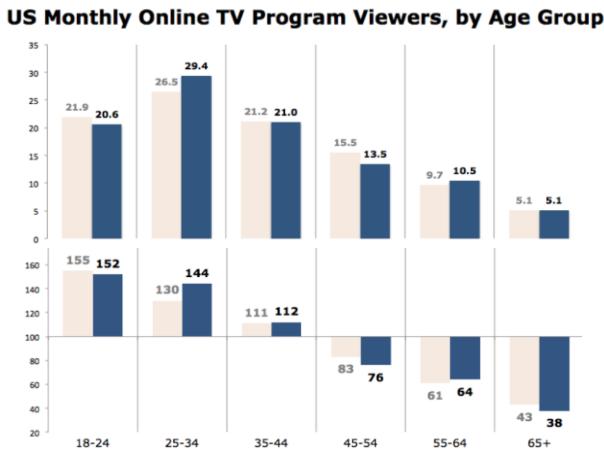
**Q. Let's put an ads to Smart TV and get paid from advertisers.
A. (No) We cannot know whether the people are seeing the ad!**

Q. What do advertisers want?

A. (Well) They want to improve the efficiency of advertising!

Q. How do we get paid from advertisers?

A. If we tell them who are looking at specific ads, then it'll improve the efficiency!



- Nielsen과 ACR의 유사성

- 일일 가구당 평균 TV시청 시간 : Nielsen(5.03 hours) vs ACR(5.54 hours)
- 일일 가구당 평균 TV시청 프로그램 수: Nielsen(8.74 units) vs ACR(8.01 units)

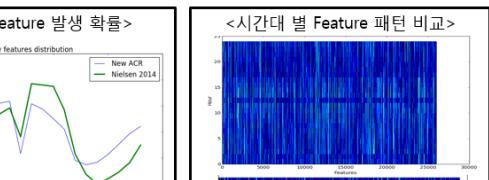
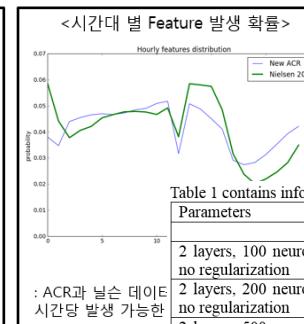
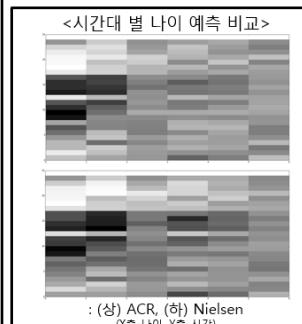


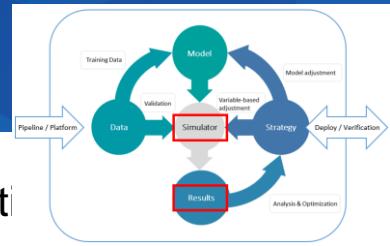
Table 1 contains information about NN architectures and their results:

Parameters	Patent NN		NN	
	Accuracy	F1	Accuracy	F1
2 layers, 100 neurons in each layer, no regularization	53%	79.4%	52.6%	79.6%
2 layers, 200 neurons in each layer, no regularization	54%	80.4%	53.6%	80.3%
2 layers, 500 neurons in each layer, no regularization	54.7%	81%	54%	80.4%
2 layers, 1000 neurons in each layer, no regularization	54.8%	81.1%	53.8%	80.1%
2 layers, 2000 neurons in each layer, no regularization	55%	81.2%	53.6%	80%

In addition hyper parameter optimization was done, and table 2 shows them:

Parameters	Patent NN			NN		
	Accuracy	F1	Micro-F1	Accuracy	F1	Micro-F1
3 layers, 2000 neurons in each layer	57.1%	82.3%	83.1%	56.4%	82%	82.9%
Regularization:						
L1: 8.5×10^{-6}						
L2: 3.5×10^{-6}						
Input dropout: 18 %						
Layer dropout: 54.6%						

과거 프로젝트 리스트



8. Identify the churn customers in app/function usage of product to improve exist
(with BigData Analytics Group(Mobile), May.2014~Dec.2014)

Q. What is the reason of transfer to another one (iPhone, LG, Xiaomi, etc.)?

A. (W don't know) Is it broken? New product effect? Feeling? Low price?



Q. Who frequently change the phone?

A. (Do not know) Young people? Men or women? Having a lot of money?



Q. How do our customers use our phone?

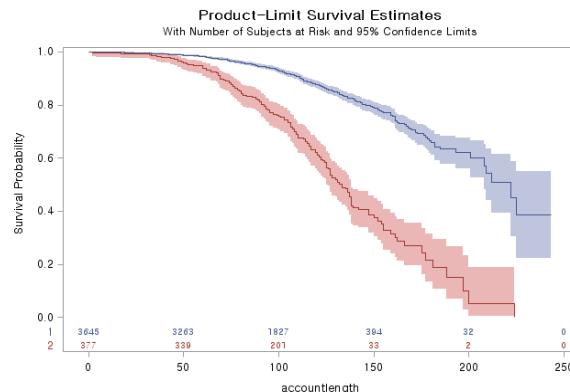
A. He/she turned off the power several times, pressed a few times on the Internet!



Q. How long do our customers use it?

!. A half of them usually changes it after 6 months based on SKT!

Analysis of Maximum Likelihood Estimates (Stepwise Result)							
Parameter	DF	Cox PH			Weibull PH		
		Parameter Estimate	Pr > ChiSq	Hazard Ratio	DF	Parameter Estimate	Pr > ChiSq
internationalplan	1	1.33215	<.0001	3.789	1	-0.4843	<.0001
numbercustomerservic	1	0.27982	<.0001	1.323	1	-0.1012	<.0001
totaldayminutes	1	0.00839	<.0001	1.008	1	-0.0031	<.0001
voicemailplan	1	-2.2239	<.0001	0.108	1	0.8059	<.0001



➤ The Real World Data Science is not a Kaggle Competition

- It can be worthwhile to step back a little and realize what exactly your **ultimate goal** is.
- The **best performance** might not be equivalent to a model yielding the **best score** in real.

Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems

New to Data Science? Get started with a tutorial on our most popular competition for beginners, [Titanic: Machine Learning from Disaster](#).

Build a Model Get the data & use whatever tools or methods you prefer to make predictions.

InClass Prediction Competition

Housing Prices Competition for Kaggle Learn Users

Apply what you learned in the Machine Learning course on Kaggle Learn alongside others in the course.

13,432 teams · 4 months to go

Overview Data Notebooks Discussion Leaderboard Rules Submit Predictions

Overview

Description

Evaluation

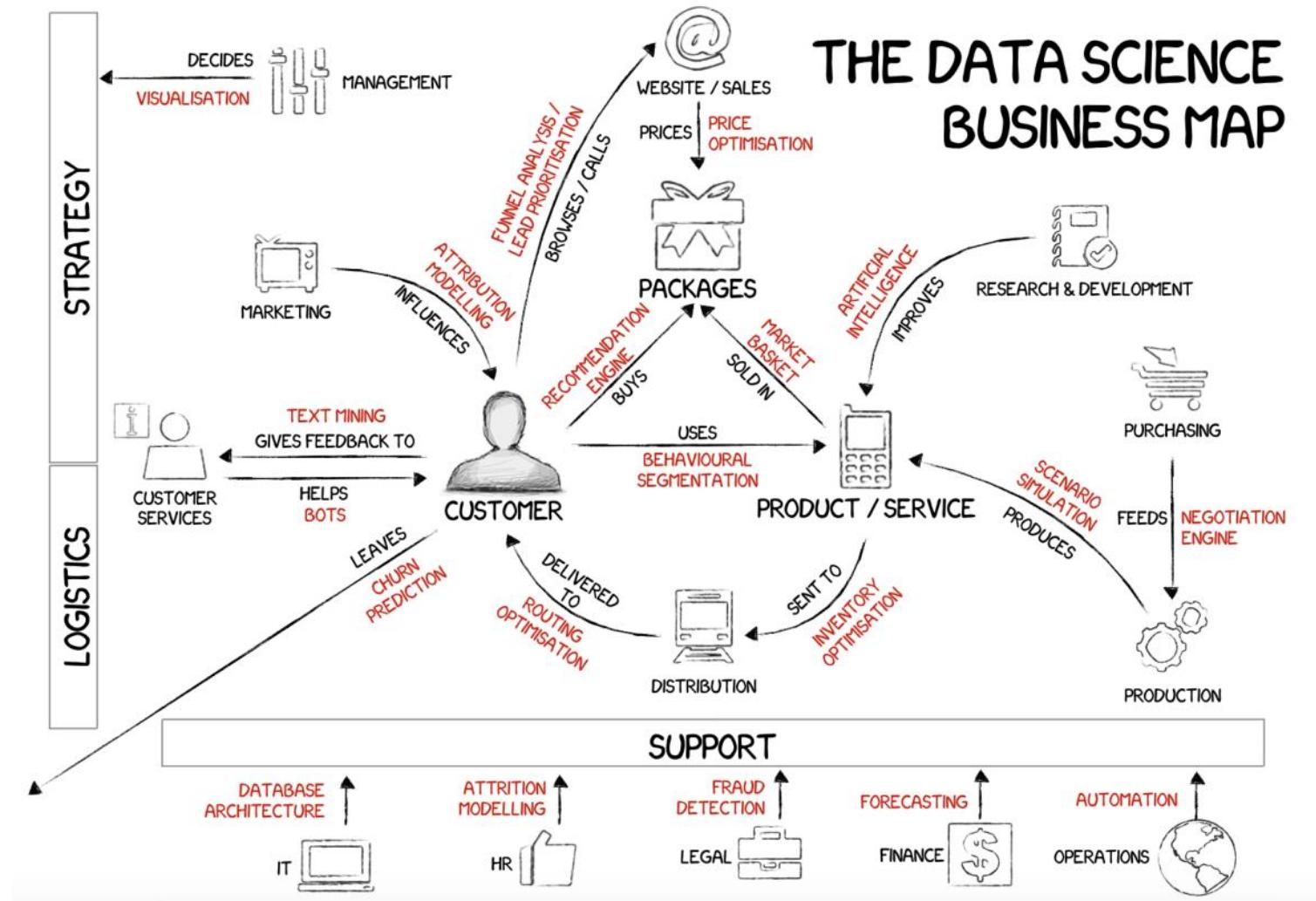
Frequently Asked Questions

What is a Getting Started competition?

Getting Started competitions were created by Kaggle data scientists for people who have little to no machine learning background. They are a great place to begin if you are new to data science or just finished a MOOC and want to get involved in Kaggle.

현실적인 데이터 분석

➤ The Real World Data Science



➤ Typical Collaboration of Data Science Project

- Makes data science teams more productive
- Broad support for open source libraries in various languages



Understand
Business
Objectives

ID Mapping
Procure
Training
Data

Prepare Data
and Build
Features

Train, Tune,
and Test
Models

Deploy and
Operationalize
Models

Update
Models

Contents

I. 데이터 비즈니스란 무엇인가?

- 1) 데이터 분석 프로세스
- 2) (과거) 프로젝트 리스트
- 3) 현실적인 데이터 분석

II. 삼성의 데이터 비즈니스 필요성과 사례

- 1) 삼성전자 구성원
- 2) 데이터 분석 협업 예시
- 3) 데이터분석 목표
- 4) 데이터 현황
- 5) 데이터 분석/개발 이슈
- 6) 현재 프로젝트 리스트
- 7) Data Analytics Lab 추진방향
- 8) Data Analytics Lab 선행 기술

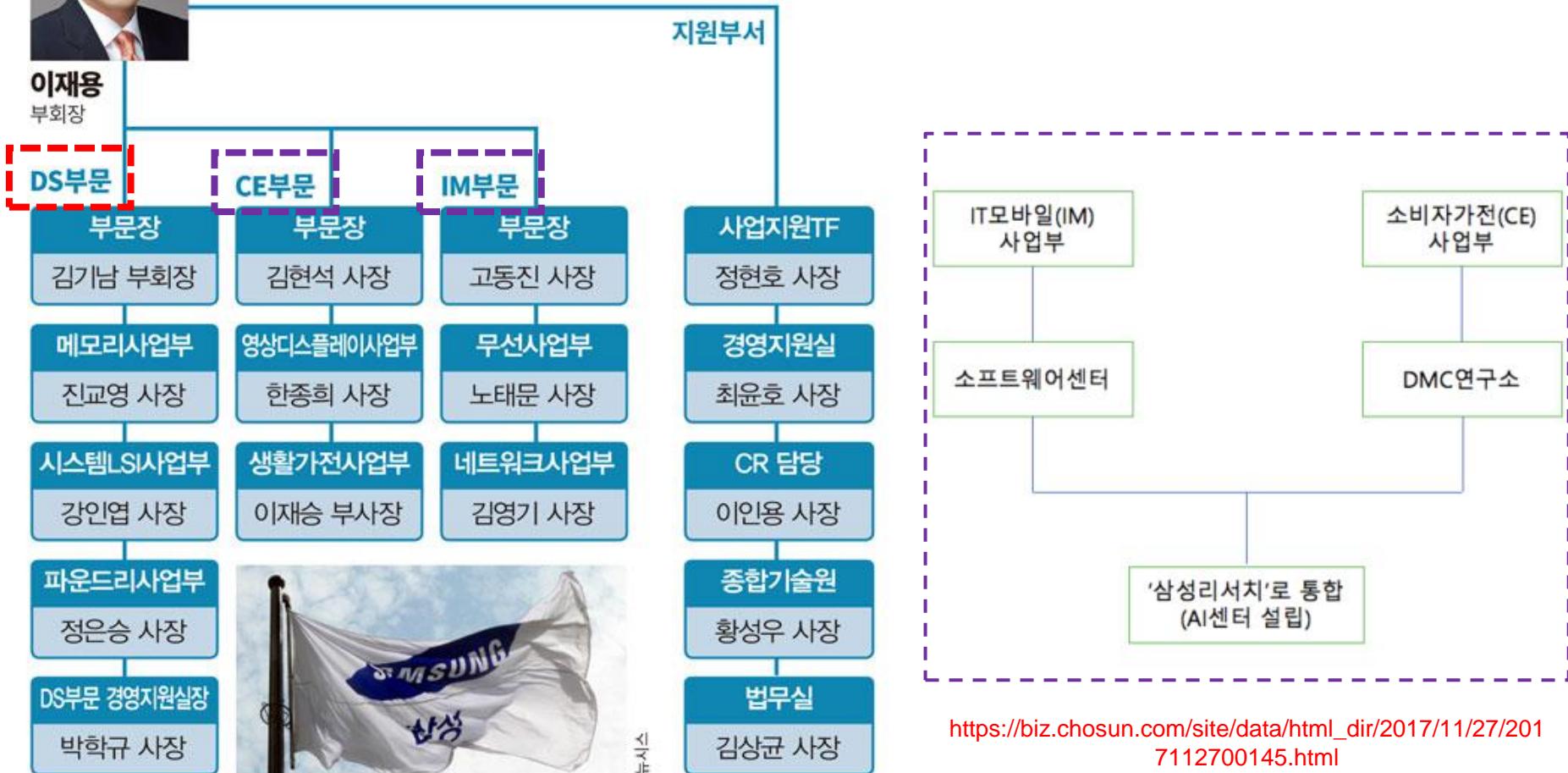
III. Platform of AI Center

- 1) 대내외 환경
- 2) AI Center

삼성전자 구성원



삼성전자 주요임원 조직도

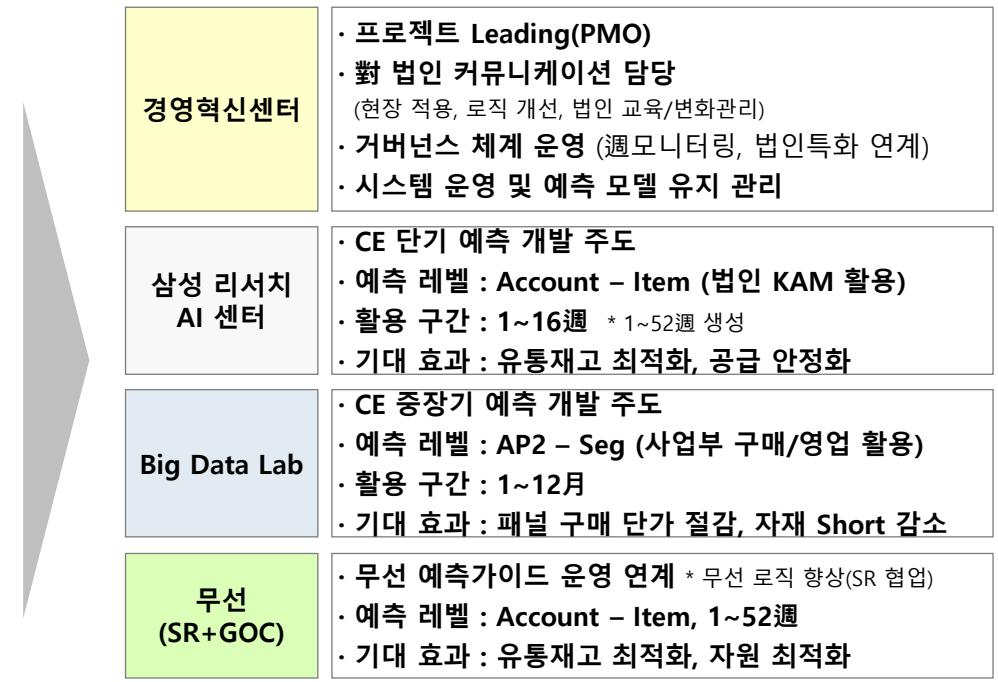
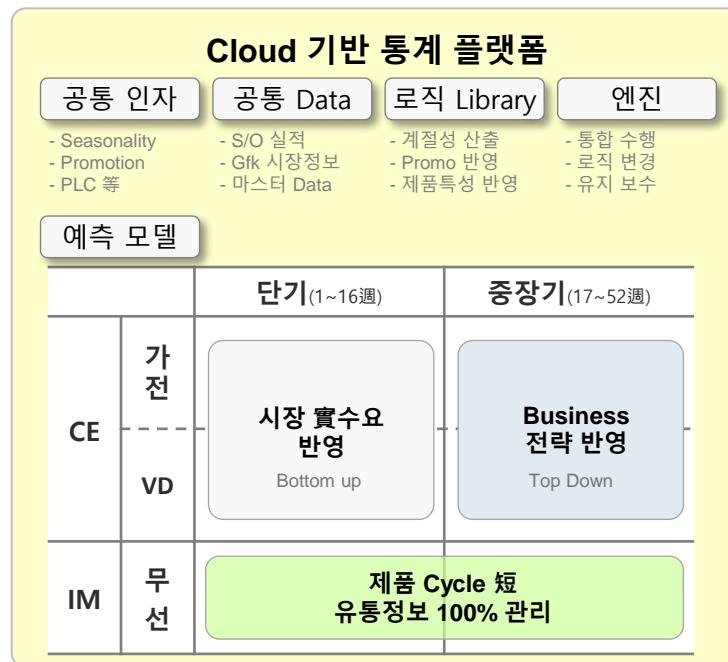


데이터 분석 협업 예시

➤ 제품 및 예측 구간 특성 반영하여 내부 3개의 Sub Logic으로 개발 중

- 삼성리서치 + Big Data Lab + GOC + SDS + 혁신센터 협업체계 구축

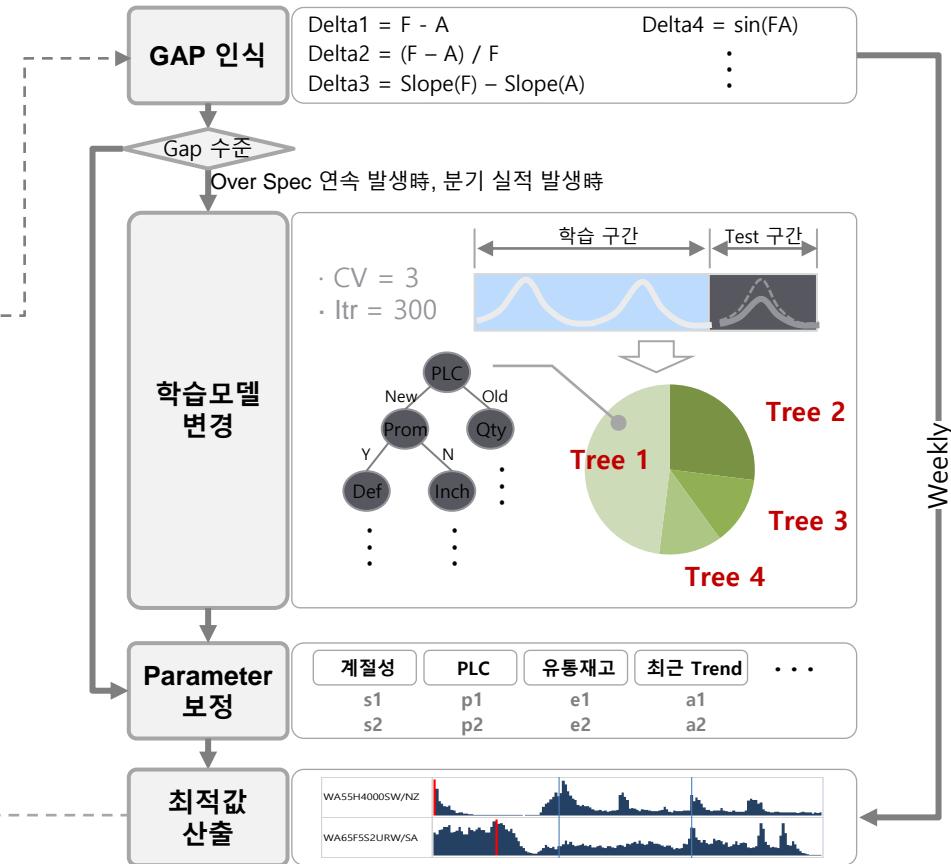
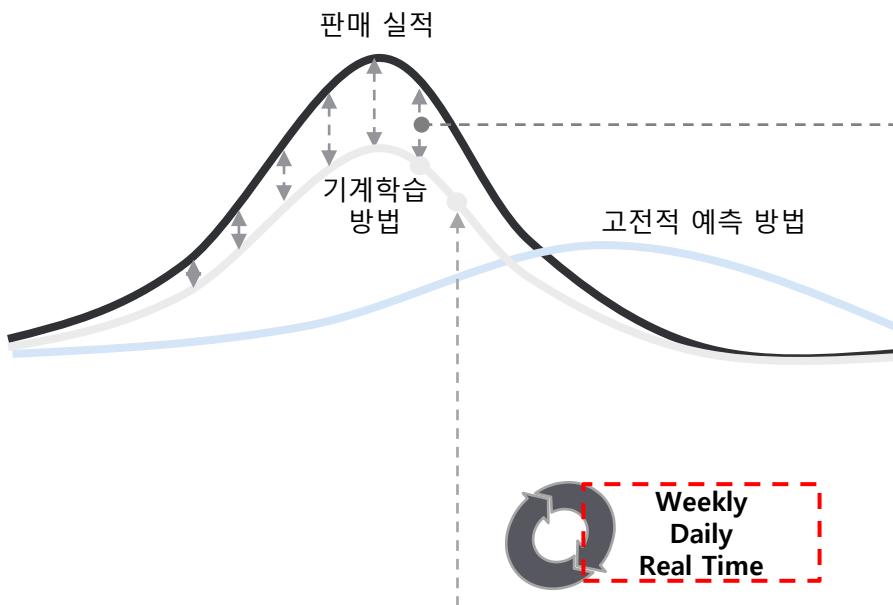
【 전사 표준 통계 플랫폼 개발 현황 】



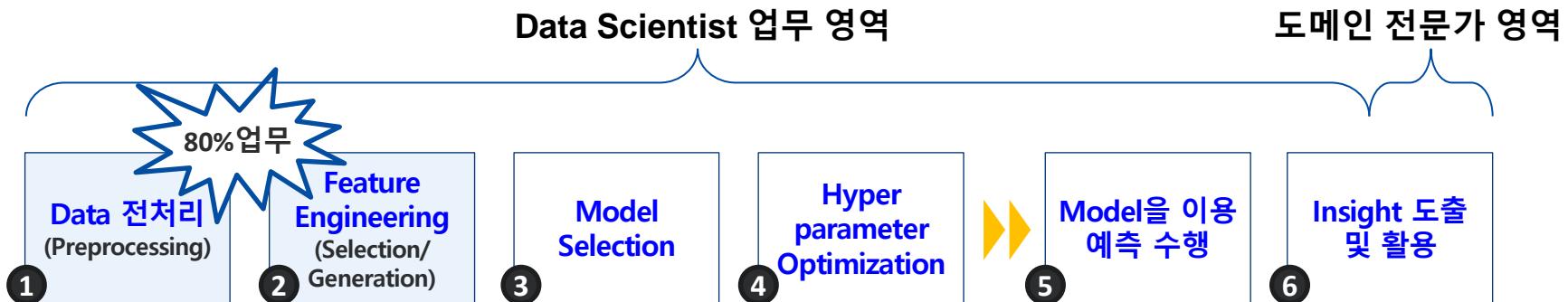
데이터 분석 협업 예시

➤ 인공지능 및 기계학습을 활용한 수요예측

- 고전적 예측 방법 : 사전 정의된 Logic 항목에 한하여 대응
- 기계학습 방법 : 발생 상황에 따라 대응 방법 자율 선택 (정기/비정기적 대응 Rule 업데이트)



데이터 분석 협업 예시



- ① 데이터 전처리를 잘 수행하여 분석되는 데이터를 깔끔하게 정리
- ② 도구와 직관을 활용해서 예측 결과에 가장 영향력이 큰 인자(feature)를 선택
- ③ 원하는 예측에 가장 잘 부합하는 모델을 선택하고
- ④ 최적 분석 결과를 얻기 위해서 Hyper parameter를 최적화
원하는 결과를 얻을 때까지 1, 2, 3, 4를 반복하는 끈기를 발휘하는 것

분석엔진 필요
(M&A/산학/자체)

분석 자동화의 결과로

- Data Scientist가 분석을 더 빠르고 정확히 수행
- Data 분석에 대한 전문 지식이 없는 사람들도 고급 기술을 활용한 Data 분석 가능

데이터분석 목표

➤ AI가 "미래의 전기"라면 데이터는 "미래의 에너지"

- 미래에 AI를 통해 돈버는 회사는 데이터를 보유한 회사
- 오랫동안 축적된 데이터는 경쟁사가 넘볼 수 없는 자산

“분석을 통해 추론을 하고 (데이터를 들여다보며 생각하고),
추론을 통해 예측을 수행 (생각하여 예측함)”

“전사 Data 통합/분석을 통해 Data 기반으로 경영 효율화 추진”

➤ 삼성의 Data 현황

- 5억대 기기/서비스/경영/제조 데이터의 크기와 종류는 세계 최고 수준
- 수집된 데이터의 상당수가 목적없는 다크 데이터이며, Data 파편화, 분석 역량 부족으로 데이터활용 미흡

기기/서비스 Data	서비스 이용 Data	사용자 관계 Data	시청/검색 Data	기기 Data													
	<ul style="list-style-type: none">• App 사용 정보• 구매 정보 (앱, 삼성페이, 패밀리허브, TV+)• Context 정보 (GPS, Time)• 헬스정보 (심박, 혈당)	<ul style="list-style-type: none">• PIMS, 주소록• SNS, 메일/통화, 메시지	<ul style="list-style-type: none">• 컨텐츠 제목/채널/장면• 제품, 주요인물• 음성, 음악• 입력 키워드, 웹 방문기록• 사용자 재/부재	<ul style="list-style-type: none">• 컨텐츠 제목/채널/장면• 제품, 주요인물• 음성, 음악• 입력 키워드, 웹 방문기록• 사용자 재/부재													
사내 시스템 Data	<table border="1"><thead><tr><th>개발</th><th>제조</th><th>유통</th><th>판매</th><th>사용</th><th>A/S</th></tr></thead><tbody><tr><td><ul style="list-style-type: none">• PLM• 설계• 부품</td><td><ul style="list-style-type: none">• 공정• 설비• 에너지</td><td><ul style="list-style-type: none">• 재고• 물류• 공급</td><td><ul style="list-style-type: none">• 매출• 마케팅• 반품</td><td><ul style="list-style-type: none">• 사용자• 사용기록• VoC</td><td><ul style="list-style-type: none">• 증상• 진단• 수리</td></tr></tbody></table>					개발	제조	유통	판매	사용	A/S	<ul style="list-style-type: none">• PLM• 설계• 부품	<ul style="list-style-type: none">• 공정• 설비• 에너지	<ul style="list-style-type: none">• 재고• 물류• 공급	<ul style="list-style-type: none">• 매출• 마케팅• 반품	<ul style="list-style-type: none">• 사용자• 사용기록• VoC	<ul style="list-style-type: none">• 증상• 진단• 수리
개발	제조	유통	판매	사용	A/S												
<ul style="list-style-type: none">• PLM• 설계• 부품	<ul style="list-style-type: none">• 공정• 설비• 에너지	<ul style="list-style-type: none">• 재고• 물류• 공급	<ul style="list-style-type: none">• 매출• 마케팅• 반품	<ul style="list-style-type: none">• 사용자• 사용기록• VoC	<ul style="list-style-type: none">• 증상• 진단• 수리												

데이터 현황

➤ Direction: Efficient management and improvement of product competitiveness

- 업무/사업부별 Data 수집 → Data 파편화
- 글로벌 CS센터 : 품질 Data를 통합 분석하여 시장/개발 품질 혁신 추진 중

개발	제조	유통	판매	사용	A/S
<ul style="list-style-type: none"> • PLM • 설계 • 부품 <p>각 사업부/GCS</p> <ul style="list-style-type: none"> • PLM: 설계표준, 실패사례 • QWEB: 표준서, 내구품질, 강건설계 • CPCX: 기술자료, 도면자료, 관련문서 • SQCI: 수입검사, 협력업체, 출하검사 • E-CIMS: 유해물질, 승인결과, 부품정보 • CIS: BOM정보, 코드정보 • GSCM: P/O, D/O FCST 	<ul style="list-style-type: none"> • 공정 • 설비 • 에너지 <p>GCS</p> <ul style="list-style-type: none"> • MES 2.0: 검사결과 • GMAP: 사양정보, 작업지시, 검사정보 • GSCM: 생산 계획/실적 /CAPA <p>GTC</p> <ul style="list-style-type: none"> • Visual Inspection: 부품, 제품 이미지 • Robot Intelligence: 실시간 위치 추적, 동선, 이탈관리, 배터리 <p>GTC/G-EHS 협력</p> <ul style="list-style-type: none"> • FDMS: 불량 검사 결과/수리 (비정형) • FFMS: 부품마모, 설비 운전 상태 • FEMS: 설비 운전 상태, 소비 전력 <p>Network事</p> <ul style="list-style-type: none"> • Asset Tracking: 자재 위치, 상태, 체류기간 추적 	<ul style="list-style-type: none"> • 재고 • 물류 • 공급 <p>GCS</p> <ul style="list-style-type: none"> • MES 2.0: 출하검사 • GSCM: RTF, 선적정보, BOD <p>(전사) 경영혁신팀</p> <ul style="list-style-type: none"> • GSBN: 거래선 유통정보/ 상품보증 <p>(전사) 지원팀</p> <ul style="list-style-type: none"> • CORES: 프로모션 비율 계획/목표/현황 	<ul style="list-style-type: none"> • 매출 • 마케팅 • 반품 <p>GCS</p> <ul style="list-style-type: none"> • GSCM: 판매, 거래선, 매장 정보 <p>(전사) GMC</p> <ul style="list-style-type: none"> • M-Net: 시장/소비자정보, 구매 요인/경로, 경쟁력지수 • Gfk, NPD: 시장규모, M/S, 경쟁사판매 <p>한국총괄</p> <ul style="list-style-type: none"> • CRM: 고객정보(삼성계정, 구매/결제/배송지), 매장, 제품 	<ul style="list-style-type: none"> • 사용자 • 사용기록 • VoC <p>무선事</p> <ul style="list-style-type: none"> • Rubin: 사용자 기기 사용 정보 • 서비스 데이터: 당사 app별 사용 정보 • S. IoT Cloud: 기기 설정/운용 데이터 • DQA: 개발/제조/검증/시장 데이터 <p>VD事</p> <ul style="list-style-type: none"> • KPI: TV app 서비스/제품 로그 • RM: 기기 데이터 <p>가전事</p> <ul style="list-style-type: none"> • 스마트가전 Cloud: 기기 사용 내역 • HRM: 기기 센서/설정값 • HASS: 기기 데이터, 방문 수리 이력 <p>의료기기</p> <ul style="list-style-type: none"> • S-Detect for Breast: 초음파 임상 데이터 <p>한국총괄</p> <ul style="list-style-type: none"> • CRM: 사용자 demo. 	<ul style="list-style-type: none"> • 증상 • 진단 • 수리 <p>GCS</p> <ul style="list-style-type: none"> • QINGS: 불량증상, 원인 코드, 수리정보 • GCIC: 고객 Call, 상담정보 • GPLS: PL 발생/처리/종결 • MES 2.0: 수리이력 <p>한국총괄</p> <ul style="list-style-type: none"> • CRM: 제품 서비스 정보 (증상, 수리결과 등)

➤ “Data 분석 업무와 SW 개발 업무의 유사성”

- **공통의 표준 도구와 분석 환경의 부재**
→ 각 Data Scientist가 목적에 맞게 개별로 분석 도구를 활용하여 분석 진행
- **분석 코드의 공유/재사용 부재**
→ 전문가의 지식 공유 부족
- **Data Scientist 개인의 역량에 의존하여 분석 결과 좌우**
→ 분석 결과에 대한 균일한 성능 확보가 어려움
→ 분석 결과에 대한 설명이 어려움
- **협업 및 코드 리뷰 환경 부재**
→ 코드/분석의 품질 저하



사내 공통 분석 도구 및 분석 개발 문화 필요

데이터 분석/개발 이슈

- 데이터 Silo 현상으로 인해 통합/연계 분석 어려움
- 데이터 간 존재하는 다양한 Conflict들로 인해 데이터 정비 작업은 많은 시간과 인력 필요

GFK

Subsidiary	Product	GfK Category	Brand	(GFK)Main_Types	Metrics
SER	REFRIGERATOR	REFRIGERATOR	SAMSUNG	1 DOOR >90 CM	
SER	REFRIGERATOR	REFRIGERATOR	SAMSUNG	2 DR FRZ. BTM	
SER	REFRIGERATOR	REFRIGERATOR	SAMSUNG	2 DR FRZ. TOP	
SER	REFRIGERATOR	REFRIGERATOR	SAMSUNG	3+ DOORS	
SER	REFRIGERATOR	REFRIGERATOR	SAMSUNG	SIDE BY SIDE	
SER	REFRIGERATOR	REFRIGERATOR	AEG-	ELECTROL 1 DOOR >90 CM UX	
SER	REFRIGERATOR	REFRIGERATOR	AEG-	ELECTROL 1 DOOR 81 - 90 CM UX	

GFK_한국

DOOR_TYPE	(FREEZE_POSITION)	FREEZE?	FRONT
SINGLE SWING	UNKNOWN	4 STARS	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	FREEZER ON TOP	1 STAR	WHITE
SINGLE SWING	UNKNOWN	4 STARS	WHITE
SINGLE SWING	NO FREEZER	N.A.	WHITE
SINGLE SWING	NO FREEZER	N.A.	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE
SINGLE SWING	NO FREEZER	WITHOUT	WHITE

NPD

GfK Category	Brand	(NPD)Type	Metrics	Sales(Q)
REFRIGERATOR	SAMSUNG	FREEZER ON BOTTOM		1
REFRIGERATOR	SAMSUNG	FREEZER ON TOP		10,195
REFRIGERATOR	SAMSUNG	FRENCH DOORS		2,166
REFRIGERATOR	SAMSUNG	REFRIGERATOR ONLY-NO FREEZER		105
REFRIGERATOR	SAMSUNG	SIDE BY SIDE		585
REFRIGERATOR	ALL OTHERS	COMPACT		3,467
REFRIGERATOR	ALL OTHERS	FREEZER ON BOTTOM		1

Traqline_Aham

PRODUCTGROUP	PRODUCT_TYPE	BRAND
REF	SBS	Frigidaire
REF	SBS	Other Brands
REF	SBS	Maytag
REF	SBS	KitchenAid
REF	SBS	Amana
REF	TMF	Samsung
REF	TMF	LG
REF	TMF	Whirlpool
REF	TMF	General Electric
REF	TMF	Kenmore
REF	TMF	Frigidaire

AC_Nielson

(GFK)Height_In_Cm	(GFK)Main_Types	(GFK)Mounting_System
DOOR		
TMF		

→ 이종 데이터 통합/연계 위한 데이터 Curation 기술 확보 필요

- Scalability through automation
- Non-programmer orientation
- Incremental

데이터 분석/개발 이슈

- 많은 중요한 정보들이 자연어 형태로 저장되어 있어 활용 어려움

PLM 데이터 (VD事)

product	code	symptom	date	status	priority	type	...
UT	P150119	Press guide key -> video distortion happen	2015-01-19 20:24:26	Close	A	S/W	...
UT	P160211	Corrupted images observed on RF Screen when pressing Guide key	2016-02-11 16:36:30	Close	B	S/W	...
UT	P160112	Videos distortion	2016-01-12 19:00:02	Close	B	S/W	...
UT	P160212	가이드 키 누를 때 화면 깨짐	2016-02-12 16:30:13	Close	B	S/W	...

Qings 데이터 (GCS, 에어컨 고객 클레임)

고장구분	처리형태	증상명	부위그룹	부위명	처리명	고장증상
제품외불량	사용설명	보턴/KEY동작안됨	송신기	밧데리 점검/교환	기능설명	전원버튼꽉눌러야켜짐/수동버튼O
제품불량	자재사용수리	기구파손	Cabinet	전면 케비넷트	PL 입수 등급 C급	스탠드 설치 & 선 연결 요청
설치불량	자재 미사용수리	화면무	신호	선로/장비/케이블	외부케이블재연결	전원불/유상
제품외불량	사용설명	보턴/KEY동작안됨	타기기	IPTV	기능설명	거실에서안방으로옮기고벽걸이
제품불량	자재사용수리	기구파손	공통부품	벽걸이프레임	PL 입수 등급 C급	음성 안나옴/셋탑연결중
제품외불량	사용설명	채널 수신안됨	비고장	화면/화질 설명	분해조립	한쪽색상이 갈라져나옴
제품외불량	사용설명	음성무	타기기	SET TOP BOX	기능설명	STB/케이블연결확인뜸
제품불량	자재 미사용수리	변형/변색/도색(코팅벗겨짐)	Cabinet	후면 케비넷트		
제품불량	환불	화면무	PCB ASSY	MAIN 보드	고장성(H/W) 불만	
소비자과실	사용설명	사이드/단차	공통부품	스탠드	기능설명	

→ 자연어 데이터로부터 의미를 추출하여 활용할 수 있는 기술 확보 필요

- Natural Language Processing
- Information Extraction
- Ontology Learning

데이터 분석/개발 이슈

- 과거 데이터 수집 목적과 현재 활용 목적이 다름
- Anomaly Detection/Fault Prediction 時 Unlabeled 데이터 문제
- 고장 데이터 부족으로 인해 감지 및 예측 모델 개발 어려움
- 모든 고장을 재현하여 고장 데이터를 수집하는 것은 현실적으로 불가능

Time	Comp1	Comp2	4Way	Hot Gas1	미터	谱写주파수1	전체주파수1	지시주파수1	谱写주파수2	전체주파수2	고압포화온도	저압포화온도	고압포화온도	저압포화온도	트림온도1	트림온도2	에러코드	날짜/시간
11:54:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	36.7°C		
11:55:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	36.7°C		
11:56:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	36.7°C		
11:57:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	37°C		
11:58:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	37°C		
11:59:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	37°C		
12:00:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.6	25°C	38.4°C	37°C		
12:01:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	36.7°C		
12:02:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	36.7°C		
12:03:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	36.7°C		
12:04:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	36.7°C		
12:05:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.4°C	37°C		
12:06:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.4°C	37°C		
12:07:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.4°C	37°C		
12:08:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.4°C	37°C		
12:09:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.7	25°C	38.6°C	37°C		
12:10:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.7	25°C	38.6°C	37°C		
12:11:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.7	25°C	38.6°C	37°C		
12:12:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.4	25°C	15.7	25°C	38.6°C	37°C		
12:13:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.8°C	37°C		
12:14:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.8°C	37°C		
12:15:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.8°C	37°C		
12:16:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	37.2°C		
12:17:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	37.2°C		
12:18:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	37.2°C		
12:19:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.7	25°C	38.6°C	37.2°C		
12:20:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.8	25°C	38.6°C	37.2°C		
12:21:49	Off	Off	Off	Off	20HP	0	0	0	0	0	15.5	25°C	15.8	25°C	38.6°C	37.2°C		

→ Unlabeled 데이터 활용 가능한 분석 알고리즘 및 데이터 Labeling 기술 확보 필요

- 非지도 학습 (예> Hierarchical Temporal Memory)

데이터 분석/개발 이슈

- 성능 한계 극복을 위한 영향 인자 부족
- 현재 목적에 맞는 영향 인자가 부족한 경우가 많으며, 신규 영향 인자 발굴에 많은 시간 소요
- 도메인 전문가도 완벽하게 영향 인자 파악하지 못하는 경우가 많음

가전事例



IT	Humidity	...	Eva Out	Eva In	OT
32	60%	...	23.4	14.4	43
32	60%	...	23	13.4	37
32	60%	...	12.5	12	32
32	60%	...	12.3	10.1	27
...

Eva Out – Eva In

9.0
9.6
0.5
2.2
...

SCM 그룹 Sellout 예측 例

구분	SCM 그룹	AIC
주 단위	57.6%	63.9%
주 + 매장 단위	45.4%	50.1%

※ 2017년 33주차에 37주차 실 예측 결과
(SCM 그룹 Lasso, AIC DNN 활용)

→ 신규 영향 인자 발굴 및 검증 위한 자동 Feature Generation 기술 확보 필요

- 既보유 Feature 間 조합에 Operation 적용, 외부 데이터 연계

■ 분석을 위한 데이터의 양적 부족

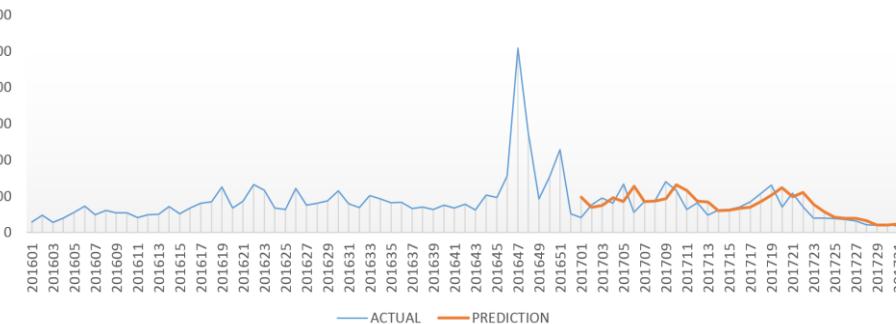
- 신규 제품 출시 時 또는 데이터를 이제 막 수집하기 시작한 경우 분석을 위한 데이터 부족

→ 데이터의 양적 부족을 극복할 수 있는 알고리즘 확보 필요

현재 프로젝트 리스트

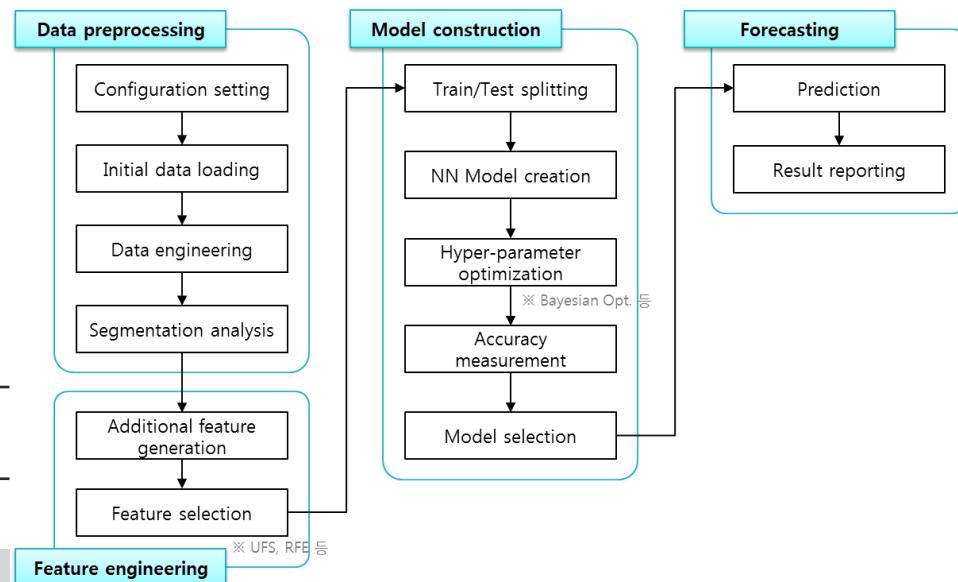
➤ 전사 Sell-Out 예측 고도화를 위한 AI 기반 분석 Core 기술 및 서비스 제공

- 전사 S/O FCST Guide의 정확도 제고 및 예측 중앙화 실현 위한 Advanced 분석 기술 필요
- 예측 분석에 필요한 주요 인자 자동 선정·추출 및 DNN 최적화 분석 Pipeline 개발 적용해 북미 TV S/O 예측 PoC 추진 完
- 지식 기반 인자 추가 및 Seg別 모델 다변화 等 기능 확장해 가전 주요 제품 예측 특화 모델 先수립 후 제품/시장 확대 협업 예정

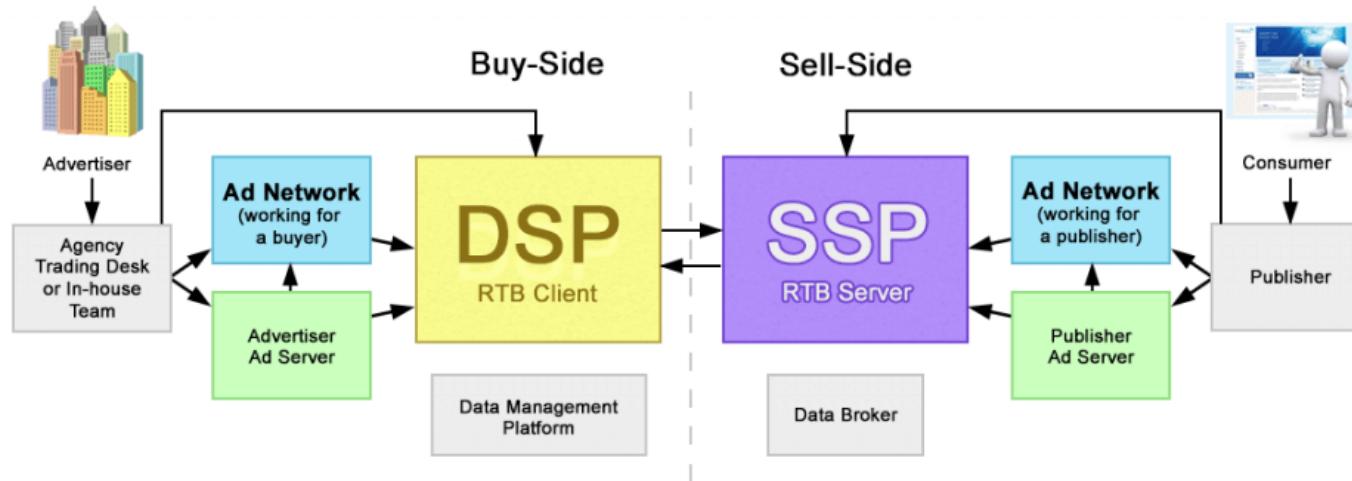


구분	SR 모델		S/O FCST Guide
	1주前	4주前	4주前
주차별 판매량 예측 정확도	79.0%	63.7%	57.6%
주차별 각 매장 판매량 예측 정확도	69.4%	54.3%	45.4%

[DNN 기반 모델 고도화]



➤ Structure of the Real-time Digital Advertising



- **Seller:** Publisher
- **Buyer:** Advertiser
- **Selling Target:** Inventory of the publisher
- **RTB:** Real Time Bidding
- **SSP:** is a platform that holds an auction bargaining with DSP, selling the inventory of the publisher as profitably as possible.
- **DSP:** is a platform to buy advertising at the best price and to show ads to users who exactly matches the need of advertisers.

현재 프로젝트 리스트

➤ TaoBao: Real Advertising Dataset by Alibaba

- **Table1(Ad Feature)**: what advertisements were exposed
- **Table2(User Profile)**: to whom the exposed advertisements were sent
- **Table3(Behavior Log)**: what kind of shopping behavior these users have done in the past
- **Table4(Respond Feature)**: who received and how they respond

Feature Group	Feature	# of Unique / Range
Ad Feature	adgroup_id	1051755
	brand	226392
	campaign_id	836715
	cate_id	38123
	customer	925205
	price	0~3660000
User Profile	age_level	7
	cms_group_id	13
	cms_segid	97
	final_gender_code	2
	new_user_class_level	4
	occupation	2
	pvalue_level	3
	shopping_level	3
Behavior Log	userid	1715676
	brand	226392
	cate_id	38123
	time_stamp	2713061
Respond Feature	user	1715676
	adgroup_id	1051755
	clk	2
	pid	2
	time_stamp	2713061
user	user	1715676

TABLE 1: Taobao data table and it's information

Segment id	Age level	Final gender code	# of Unique users	# of Unique category	# of Unique brand
0	0	1	161	166	392
		2	600		
1	1	2	74,128	1,503	9,147
2	2	2	248,396	2,505	26,319
3	3	2	395,429	3,184	24,458
4	4	2	306,764	3,278	5,325
5	5	2	203,339	3,115	21,050
6	6	2	17,090	1,320	31,385
7	1	1	21,154	1,467	30,630
8	2	1	60,063	2,503	26,836
9	3	1	129,163	3,478	6,115
10	4	1	135,700	3,636	6,352
11	5	1	111,996	3,550	13,637
12	6	1	11,693	1,580	24,522

TABLE 2: Segment data statistics

현재 프로젝트 리스트

➤ Performance Measures: Increasing predicted CTR (Click Through Rate)

1) eCPM (Effective Cost Per Mile): Expected charge amount of 1000 views ($= \frac{Cost}{Impression} \times 1000$)

2) eCPM by Pricing Type

> CPI (Cost Per Impression) = $BA_{impression} \times 1000$

> CPC (Cost Per Click) = $BA_{click} \times [pCTR \times 1000]$

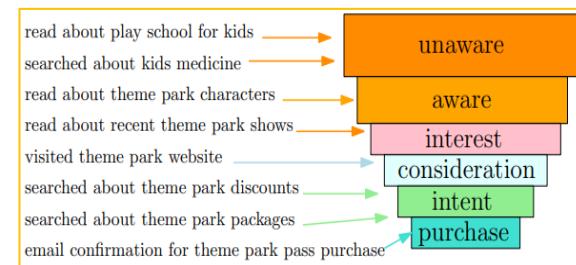
> CPA (Cost Per Action) = $BA_{action} \times [pCTR \times pCVR \times 1000]$

- BA: Bid Amount

- pCTR: predicted Click Through Rate = $Click / Impression$

- pCVR: predicted ConVersion Rate = $Action / Click$

▪ Flow of Customer Behaviors



	Pricing Type	Bid Amount	pCTR	pCVR	eCPM	Rank
Ad1	CPI	15	-	-	15000	4
Ad2	CPC	120	0.13	-	15600	3
Ad3	CPC	140	0.12	-	16800	1
Ad4	CPA	1300	0.14	0.19	16380	2

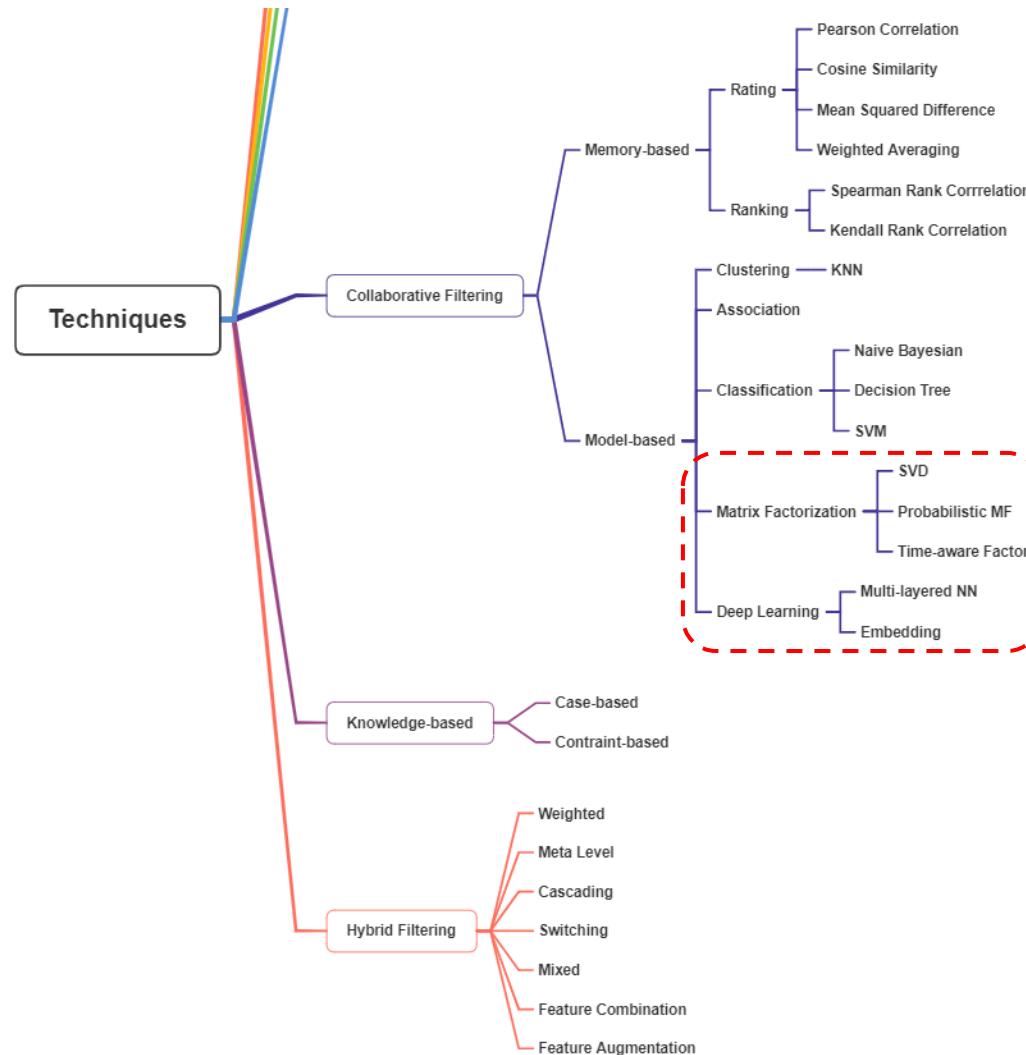
3) CAL (Calibration) = $\frac{\text{sum of } [predicted \ CTR]}{\text{sum of } actual \ CTR}$

> CAL < 1: under prediction → higher ad effect

> CAL > 1: over prediction → lower ad effect

현재 프로젝트 리스트

➤ Algorithms of Recommendation System and Deep Neural Networks with Time Series



현재 프로젝트 리스트

➤ Algorithms of Recommendation System and Deep Neural Networks with Time Series

Model	Journal & Year	Affiliation	Paper	Description
FM	ICDM 2010	Osaka Univ.	Factorization Machines	Sparse Feature의 고차원 상호작용 학습 가능
FsNN	ECIR 2016	Univ. College London	Deep Learning over Multi-field Categorical Data: A Case Study on User Response Prediction	FM을 Pre-train하여 DNN을 적용하는 방식 제시
FFM	RecSys 2016	Criteo Carnegie Mellon Univ.	Field-aware Factorization Machines for CTR Prediction	고차원 상호작용이 발생하는 Feature Bias 줄임
PNN	ICDM 2016	Shanghai Univ. Univ. College London	Product-based neural networks for user response prediction	Embedding Layer와 FC Layer와의 Product Layer 제시
Wide & Deep	DLRS 2016	Google	Wide & Deep Learning for Recommender Systems	고차원과 저차원의 하이브리드 네트워크 구조 제시
Deep & Cross	ADKDD 2017	Stanford Google	Deep & Cross Network for Ad Click Predictions	Wide & Deep의 MLP 부분을 Residual Network 변경
AFM	IJCAI 2017	Singapore National Univ.	Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks	FM의 성능을 향상위해 Time-series or Sequence 반영
NFM	SIGIR 2017	Singapore National Univ.	Neural Factorization Machines for Sparse Predictive Analytics	Sparse Feature를 Deep Structure에서 학습 일반화
DeepFM	IJCAI 2017	Harbin (HIT) Huawei	DeepFM: A Factorization-Machine based Neural Network for CTR Prediction	추천시스템에 Deep Learning을 결합
FwFM	WWW 2018	Yahoo Alibaba	Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising	Difference Feature 상호작용을 반영하여 적은 파라미터로 높은 성능
xDeepFM	KDD 2018	China (UCTC) Microsoft	xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems	Cross Network을 활용 Compressed Interaction Network 제시하여 고차 상호작용 명시적 모델링
DIN	KDD 2018	Alibaba	Deep Interest Network for Click-Through Rate Prediction	다양한 Time-series Behavior나 Historical Interest 반영위해 Local Activation Unit 제시
DIEN	AAAI 2019	Alibaba	Deep Interest Evolution Network for Click-Through Rate Prediction	GRU(Chung et al. 2014)를 사용해 Latent Interest 반영을 위한 Attentional Update GRU 제시
DSIN	IJCAI 2019	Alibaba Zhejiang Univ.	Deep Session Interest Network for Click-Through Rate Prediction	User behavior Sequence의 Session(Segment, Cluster) 정보 반영위한 Self-attention + Bi-LSTM 적용
FiBiNET	RecSys 2019	Weibo	FiBiNET: Combining Feature Importance and Bilinear feature Interaction for Click-Through Rate Prediction	Sequence-and-Excitation Network(SENENET)을 사용하여 Evolving Feature Occupation을 반영

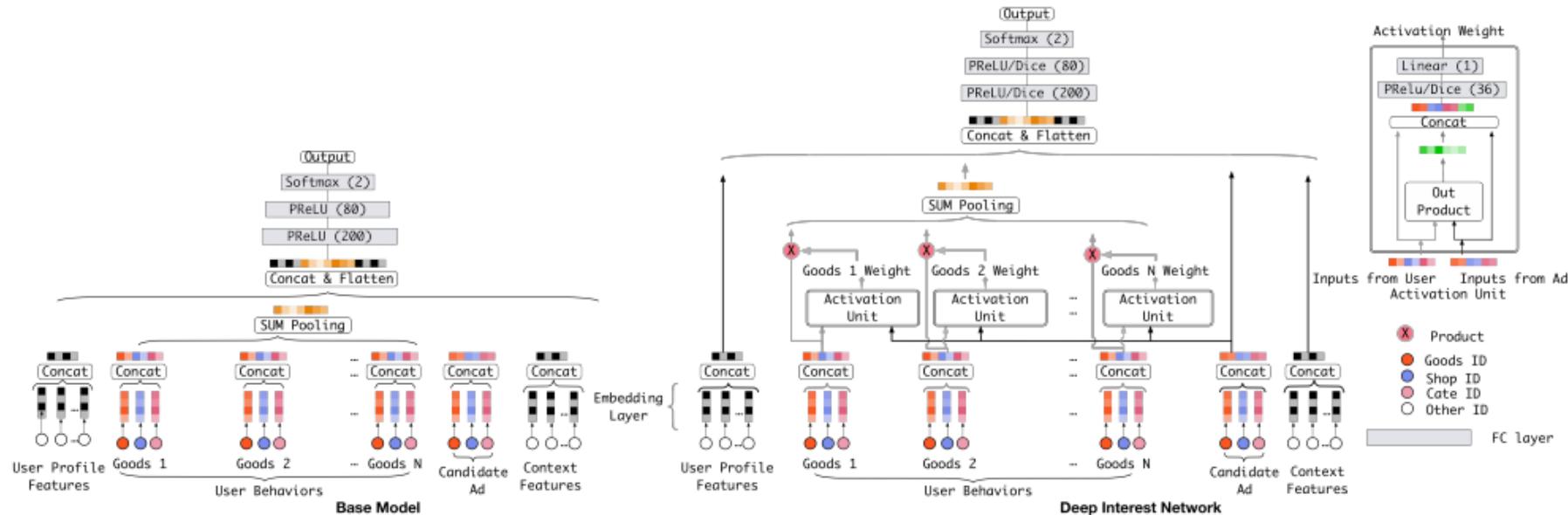
현재 프로젝트 리스트

➤ Algorithms of Recommendation System and Deep Neural Networks with Time Series

▪ Automated Feature Engineering Comparison

	No Pretraining	High-order Features	Low-order Features	No Feature Engineering
FNN	X	O	X	O
PNN	O	O	X	O
Wide & Deep	O	O	O	X
DeepFM	O	O	O	O

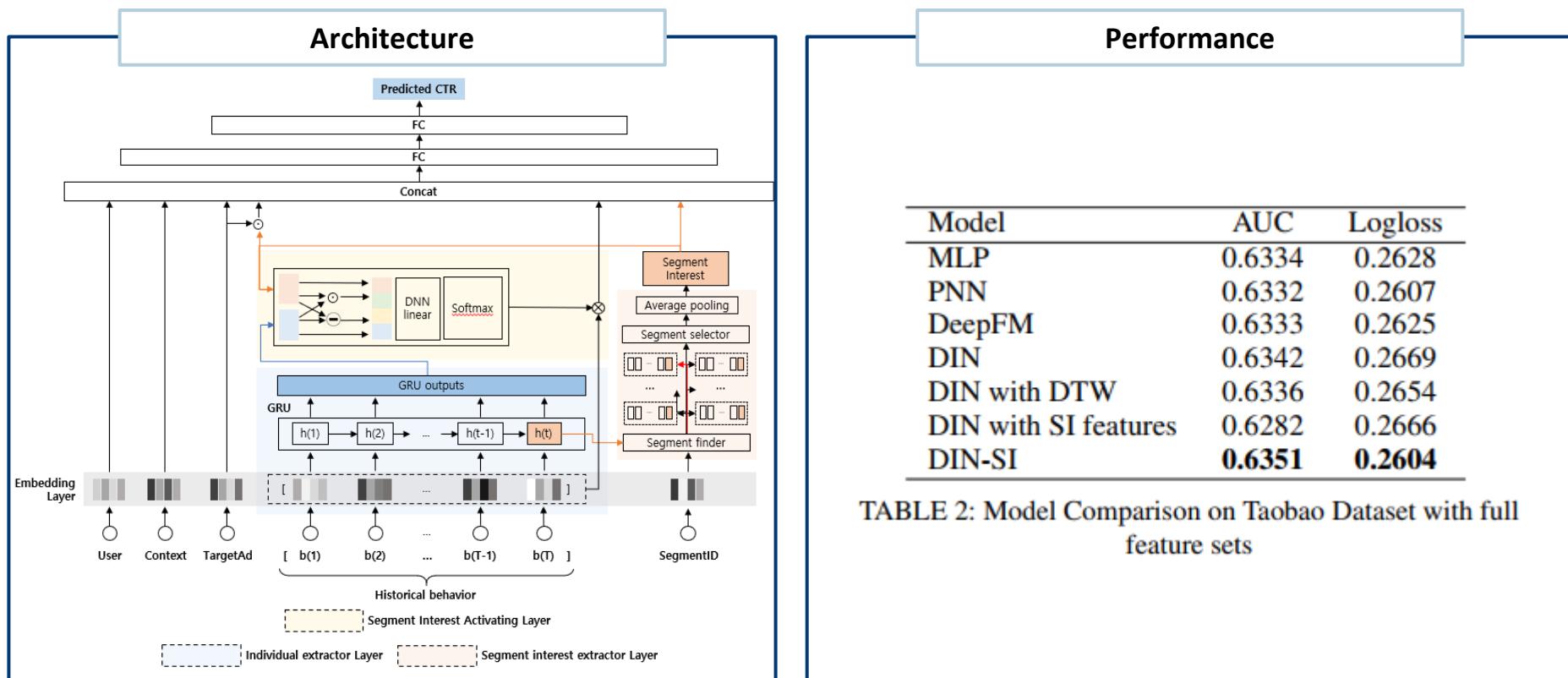
▪ Time-based Latent Sequential Feature Interactions: DIN*, DIEN, DSIN



현재 프로젝트 리스트

➤ Algorithms of Recommendation System and Deep Neural Networks with Time Series

- CTR prediction can estimate the popularity of the advertisements that will be displayed. It is critical to many web applications including web search, recommender systems, sponsored search, and online advertising.
- To predict the probability that a user will click on an item plays an important role in online advertising systems, which not only increases the revenue of the advertising media, but also maximizes users' satisfaction.



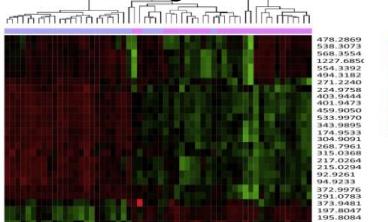
현재 프로젝트 리스트

➤ 광고추천 신규서비스를 위한 고객 개인정보(Profile) 추론

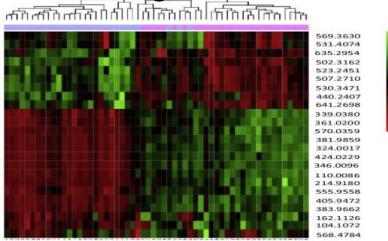
- **방향:** 1) 글로벌 리서치사(N사)의 인구조사와 TV시청률 데이터를 기계에 학습
2) TV 시청패턴 매칭으로 나이/성별/지역/연봉 고객정보 추론
3) 광고 이메일에 반응(Click)하고 구매하는 고객과 그렇지 않은 고객을 학습시켜 종속변수로 사용
- **기술:** SQL, Hadoop, Python, Collaborative Filtering, Alternating Least Square, CNN
- **성과:** 성별과 지역은 90%이상, 나이와 연봉은 각각 65%와 75%정도의 정확성

단계1) 고객 Profile 추론 예시

N사: Training



S사: Testing



Age 추론 정확성

Data Set	Sample Size According to Age Group	Classification Accuracy		Overall Classification Rate
		Size	Rate	
Test Data Set	8-13 - 30	24	80%	74.39%
	14-25 - 81	58	71.60%	
	26-45 - 60	40	66.66%	
	46-60 - 29	23	79.31%	
Training Data Set	8-13 - 45	32	71.11%	79.09%
	14-25 - 50	41	82%	
	26-45 - 55	37	67.27%	
	46-60 - 50	48	96%	

성별 추론 정확성

Type of Data Set	Sample Size	Correctly Classified		Classification Rate
		Size	Rate	
Test Set	Male - 80	68	85%	85.83%
	Female - 120	104	86.66%	
Training Set	Male - 95	84	88.42%	89.92%
	Female - 105	96	91.42%	
Training + Test Set	Male - 90	79	87.77%	87.06%
	Female - 110	95	86.36%	

단계2) 구매 예정고객 추론 예시

Data Sources

[S사]
계정정보
멤버쉽기록
S.Com 구매기록
헬스 등 건강정보
IoT다기기 보유기록

[A사]
지역정보
지점분포
지역별점유율
인구조사 정보

Decision Evaluation

Customer Journey
Attribution Flow
Impression Result
Click Through Rate
Purchase Result
Channel Pricing

6개월간 3주 간격

고정
업데이트

현재 프로젝트 리스트

➤ 마케팅 채널 매출기여도(ROI) 분석 및 포트폴리오 전략 추천을 위한 시뮬레이터 개발

- **방향:** 1) 모든 마케팅 ROI를 설명하기 위해 별도 평가함수(Cost Function)를 반영한 신규 알고리즘 개발
2) 모든 매출기여도를 극단적이지 않고 제한된 범위 내 추정
3) 포트폴리오 최적화 전략에 활용되는 알고리즘으로 시뮬레이터 개발하여 미래 계획 수립
- **기술:** Python, S-curve & Adstock, Time-series Lag Estimation, Decay Effect, Nonparametric Weibull Dist.
- **성과:** 1) 모든 마케팅 채널 중 96%의 매출기여도 추정
2) 신규 마케팅 채널 미래 투자비용 포트폴리오 시뮬레이션 제시
3) 동일매출 발생 위한 비용절감 약16% 가능, 동일비용 기준 최대 매출증가 약23% 가능

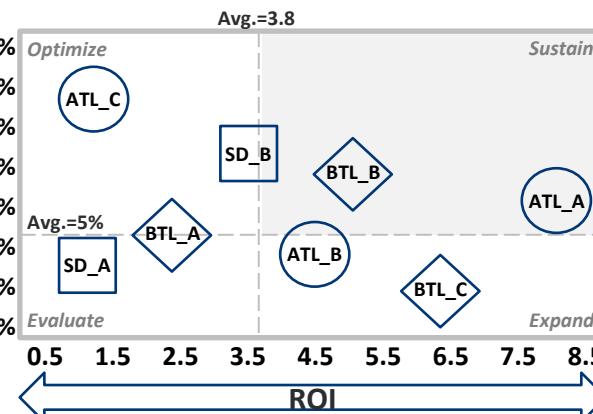
매출기여도 분석 및 시뮬레이터 예시

▪ 비용의 매출발생 도메인 분석

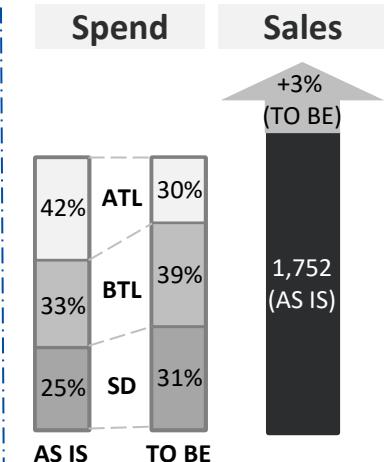
Lag	Retention	S-curve
ATL_A	1.50000	0.86680 321099.30701
ATL_B	2.00000	0.85708 3291.78608
ATL_C	0.75000	0.53617 17512.23333
BTL_A	5.25000	0.90511 2803.75182
BTL_B	3.75000	0.80140 187679.75333
BTL_C	2.75000	0.76067 67506.13376
SD_A	1.00000	0.90722 43.10248
SD_B	0.00000	0.87155 1135.37200

Sales Contribution

▪ 매출기여도 분석



▪ 포트폴리오 시뮬레이터 분석



현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

- Process and Purpose for each step of **Marketing Mix Modeling (MMM) Solution**

Solution	Process	Purpose	Analysis
Marketing Mix Modeling	Step 1: Planning Analytics	Estimation of promotion cost properties	1) Lag Estimation 2) Diminishing Return Trend 3) Carry Over Effect
	Step 2: Descriptive Analytics	Relationship transformation between cost and sales	1) S-curve Fitting
	Step 3: Diagnostic Analytics	ROI estimation of marketing features	1) Mean-shifted Regression
	Step 4: Predictive Analytics	Forecasting of sales and evaluation of parameters	1) Positive Channel Rate(PCR) 2) Effective Spend Rate(ESR) 3) ROI Boundary Rate(RBR) 4) Base Sales Weight(BSW) 5) R-square(RSQ) 6) Mean Absolute Percentage Error(MAPE) 7) Durbin-watson(DW)
	Step 5: Prescriptive Analytics	Simulation of relationship between cost and sales	1) Budget Minimization 2) Sales Maximization

현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

- **Dataset:** Past 5-year (2014 ~ 2019) weekly marketing costs and total sales
- **Marketing Components:** ATL, BTL, SD features
- **Non-marketing Components:** Base features

Input Features: ATL

Category	Feature	Feature (Short Form)
ATL (Above the Line)	Banner	BAN
	Cinema	CINEMA
CONTENT	CONTENT	CONTENT
	Coop TV	COOP
Direct TV	TVC	
Display	DIS	
Google Display	GDIS	
Google Search	GS	
Magazine	MAG	
Mobile	MOB	
Newspaper	NP	
Non-bounced Visit	NBV	
OTHER	OTHER	
Out of Home	OOH	
Programmatic	SOCIAL	
Radio	RAD	
Social Click	SOCIAL	
Total Display	TOTALDIS	
Video	VIDEO	
Youtube	YT	

Input Features: BTL, SD, Base

Category	Feature	Feature (Short Form)
BTL (Below The Line)	Base Sell-In Allowance	BSIA
	Dealer Event Support	DES
	Direct Marketing	DM
	Exhibition & Events	EE
	Indirect Price Discount	IPD
	Packages & Bundles	PBS
	Price Protection	PP
	Promotion Material	PM
	Public Relations	PR
	Sales Staff Incentive	SSI
	Sales Staff Support	SSS
	Shelf Merchandising	SM
	Shop in Shop	SIS
	Sponsorship	SP
SD (Sales Discount)	Direct Price Discount	DPD
	Marketing Research	MR
	SKU Listing	SKULIST
	Volume Rebate	VR
Base	Economic Index	Economic Index
	Launching Date	Launching Date
	Price	Price
	Samsung Brand Equity	Samsung Brand Equity
	Seasonality	Seasonality
	Total Distribution Points	TDP
	Weighted Distribution	WD

현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

▪ Step1: Algorithm of Planning Analysis and Results

Estimation Process

- For the transformed features reflecting memory effect of $\text{Lag}(L)$,

$$A_t = X_{t-L} + \lambda * A_{t-1}$$

- Function representation between Adstock (A_t) and Sales (Y):

$$\begin{aligned}\hat{Y} &= \gamma_0 \cdot A_{j,t}^0 + \gamma_1 \cdot A_{j,t}^1 + \cdots + \gamma_L \cdot A_{j,t}^L + \varepsilon_t \\ &= \sum_{0 \leq i \leq L} \gamma_i \cdot A_{j,t}^i + \varepsilon_t\end{aligned}$$

- where $\gamma_i, i = 0, 1, \dots, L$ is coefficient of lagged adstock.

- Assuming that each lagged adstock is independent of each other, estimating covariance with lagged adstock ($A_{j,t}^l$) at each time point generates $L + 1$ equations:

$$\text{Cov}(\hat{Y}, A_{j,t}^0) = \sum_{0 \leq i \leq L} \gamma_i \cdot \text{Cov}(A_{j,t}^i, A_{j,t}^0) + \text{Cov}(\varepsilon_t, A_{j,t}^0) = \gamma_0 \cdot \text{Cov}(A_{j,t}^0, A_{j,t}^0)$$

$$\text{Cov}(\hat{Y}, A_{j,t}^1) = \sum_{0 \leq i \leq L} \gamma_i \cdot \text{Cov}(A_{j,t}^i, A_{j,t}^1) + \text{Cov}(\varepsilon_t, A_{j,t}^1) = \gamma_1 \cdot \text{Cov}(A_{j,t}^1, A_{j,t}^1)$$

⋮

$$\text{Cov}(\hat{Y}, A_{j,t}^L) = \sum_{0 \leq i \leq L} \gamma_i \cdot \text{Cov}(A_{j,t}^i, A_{j,t}^L) + \text{Cov}(\varepsilon_t, A_{j,t}^L) = \gamma_L \cdot \text{Cov}(A_{j,t}^L, A_{j,t}^L)$$

- As $A_{j,t}^i$ is just shifted series, $\text{Cov}(A_{j,t}^i, A_{j,t}^i) = \text{Var}(A_{j,t}^i)$,

$$\gamma_0 = \frac{\text{Cov}(\hat{Y}, A_{j,t}^0)}{\text{Var}(A_{j,t}^0)}$$

$$\gamma_1 = \frac{\text{Cov}(\hat{Y}, A_{j,t}^1)}{\text{Var}(A_{j,t}^1)}$$

⋮

$$\gamma_L = \frac{\text{Cov}(\hat{Y}, A_{j,t}^L)}{\text{Var}(A_{j,t}^L)}$$

- If i can be fitted as λ_i for some constant λ , we can obtain the lag(L) and the retention rate(λ) at the same time after fitting λ_i as $\lambda^{|i-L|}$.

현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

- Step1: Algorithm of Planning Analysis and Results

Estimation Result of ATL				
Category	Feature	Planning Analysis		Descriptive Analysis
		Lag Time	Retention Rate	Curve Driver
ATL	CINEMA_MODEL	0	0.72	67377.96
	CONTENT_MODEL	2	0.87	321099.31
	CONTENT_Mix	0	0.63	1301.73
	COOP_Mix	0	0.22	57569.6
	COOP_Model	0	0.55	75179.76
	DIS_MODEL	2	0.82	6515.39
	DIS_Mix	2	0.86	3291.79
	GDIS_MODEL	1	0.54	17512.23
	GS_MODEL	4	0.8	187679.75
	GS_Mix	5	0.91	2803.75
	MAG_MODEL	0	0	0
	MAG_Mix	0	0	0
	NBV	3	0.76	67506.13
	OOH_BRAND	0	0	0
	OOH_MODEL	2	0.83	186.12
	OOH_Mix	1	0.91	43.1
	OTHER_MODEL	0	0.87	1135.37
	RAD_MODEL	2	0.66	631.71
	RAD_Mix	0	0	0
	SOCIAL_MODEL	2	0.8	653150.52
	SOCIAL_Mix	0	0.8	274011.56
	TOTALDIS_MODEL	2	0.62	10159.96
	TVC_Mix	1	0.77	99.82
	TVC_Model	0	0.91	231.44
	VIDEO_MODEL	0	0.82	3696.33
	VIDEO_Mix	4	0.7	7543.25
	YT_BRAND	0	0	0
	YT_MODEL	2	0.68	5409236.55
	YT_Mix	0	0	0
Average		1.20	0.58	247171.14

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

- Step2: Relationship Transformation between Cost and Sales
- Step3: ROI Estimation and Sales Decomposition

"S-curve" Fitting & Estimation Setting

- For the estimation of sales decomposition from the total, reflects the nonlinearity between the input and output of the main algorithm: "S-curve Fitting".

$$S(A_{j,t}^L) = S(X_{j,t-L} + \lambda_j * A_{j,t-1}^L)$$

- Depending on the data, the Weibull and Polynomial function showed higher curve fitting performance in this study.
- The function for estimating sales contribution is as follows:

$$\begin{aligned}\hat{Y} &= \beta_1 \cdot S(A_{1,t}^{L_1}) + \beta_2 \cdot S(A_{2,t}^{L_2}) + \cdots + \beta_N \cdot S(A_{N,t}^{L_N}) + \epsilon_t \\ &= \sum_{j \in \text{inputs}} \beta_j \cdot S(A_{j,t}^{L_j}) + \epsilon_t\end{aligned}$$

- where \hat{Y} is total sales, $S(A_{j,t}^{L_j})$ is "S-curve" fitted adstock, and β_j is return on investment (ROI) for each marketing channel.

Mean-variance Adjusted Regression

- Expected mean and variance are reflected in the cost function as prior information of marketing channels:

$$\begin{aligned}\hat{\beta}_{\text{MVAR}} &= \arg \min_{\beta} \left[\frac{1}{T} \sum_{1 \leq t \leq T} (Y_t - \hat{Y}_t)^2 + \alpha \sum_{j \in \text{inputs} \setminus \text{Base}} h(\beta_j) \right] \\ h(\beta_j) &= \left| \frac{\beta_j \cdot S(A_{j,t}^{L_j})}{C_j} - R_j \right|^k \quad \text{or} \quad \exp \left| \frac{\beta_j \cdot S(A_{j,t}^{L_j})}{C_j} - R_j \right|^k\end{aligned}$$

- where $h(\cdot)$ is some function reflects the prior constraints of contribution, and α is a regularizer. C_j is cost, R_j is ROI of each marketing channel, and k is the shape parameter of cost function.
- For the comparison, the cost function of Elastic Net is as follows:

$$\hat{\beta}_{\text{Elastic Net}} = \arg \min_{\beta} \left[\frac{1}{T} \sum_{1 \leq t \leq T} (Y_t - \hat{Y}_t)^2 + \alpha_1 \sum_{j \in \text{inputs}} \beta_j^2 + \alpha_2 \sum_{j \in \text{inputs}} |\beta_j| \right]$$

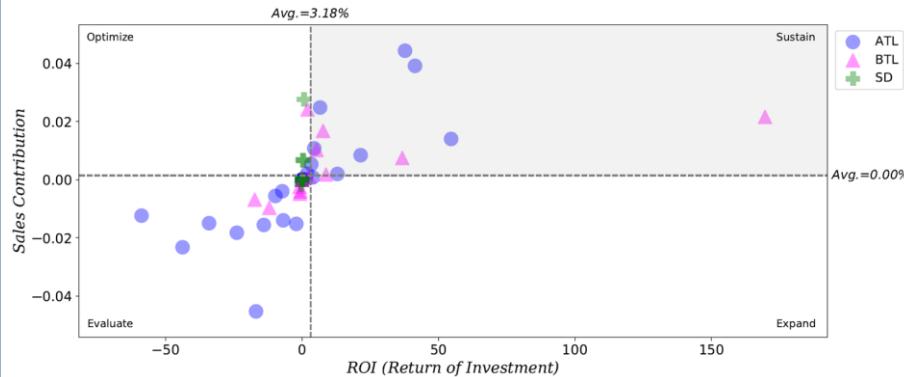
현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

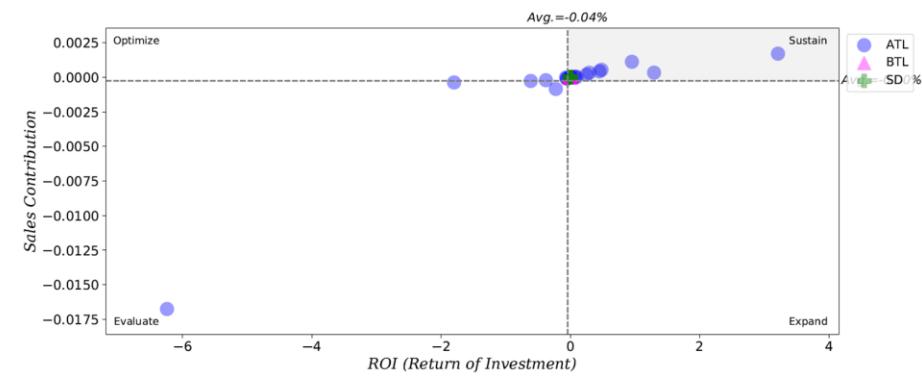
- Mean-variance Adjusted Algorithm estimates much more realistic results (not negative or over 100 of ROI).

Sales Contribution and ROI for each Algorithm

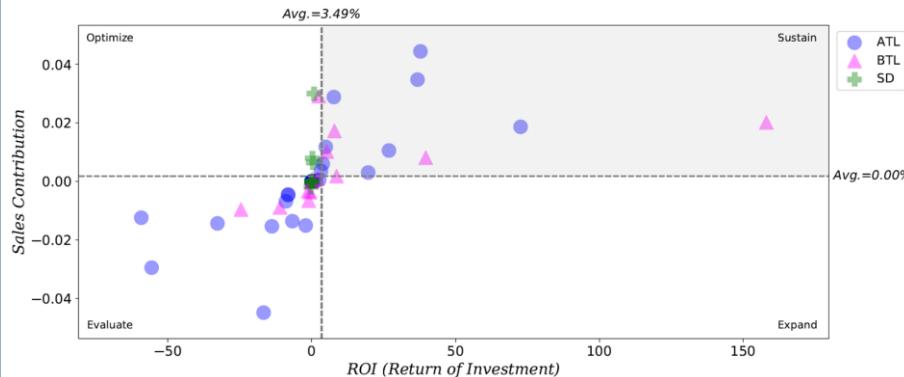
▪ Ridge



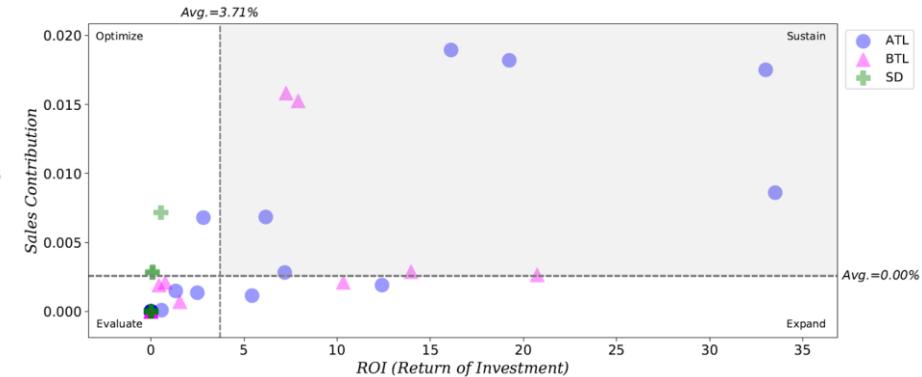
▪ Elastic Net



▪ LASSO



▪ Mean-variance Adjusted Regression



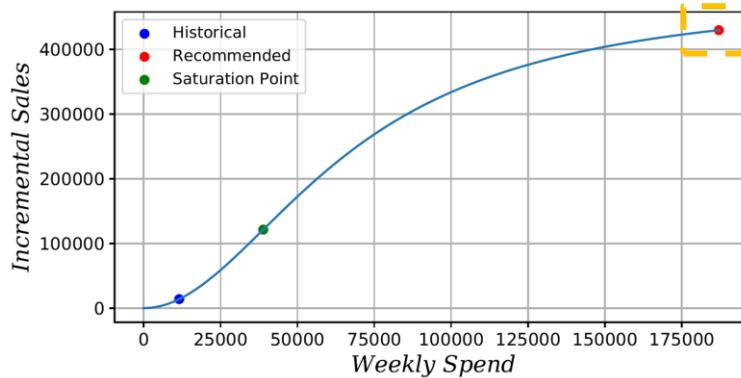
현재 프로젝트 리스트

➤ Estimating Sales Contribution and Prediction-based Portfolio Optimization in Marketing

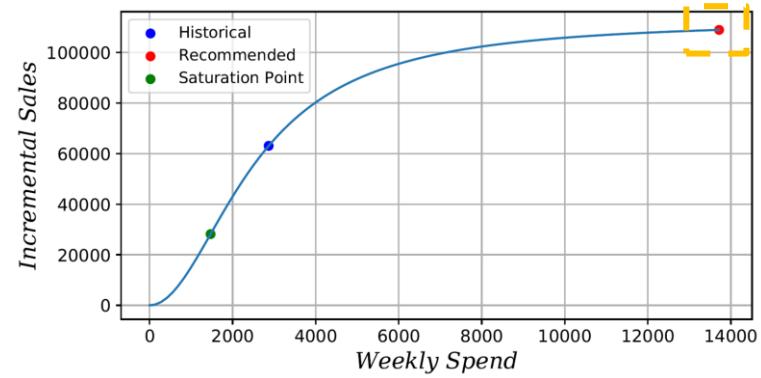
- Digital marketing need to invest more, but traditional things has lower sales contribution and less investment.

Recommended Portfolio

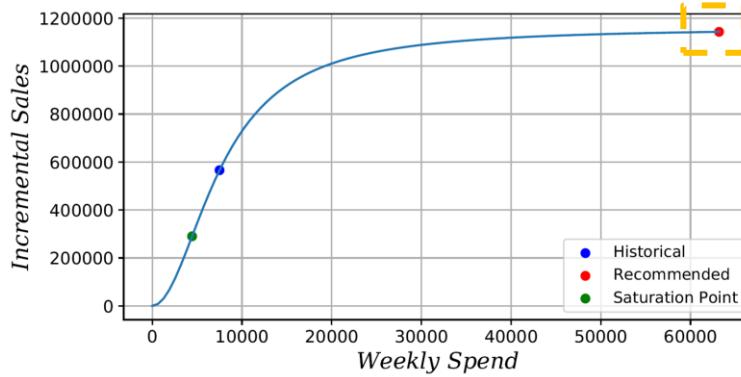
▪ Cinema



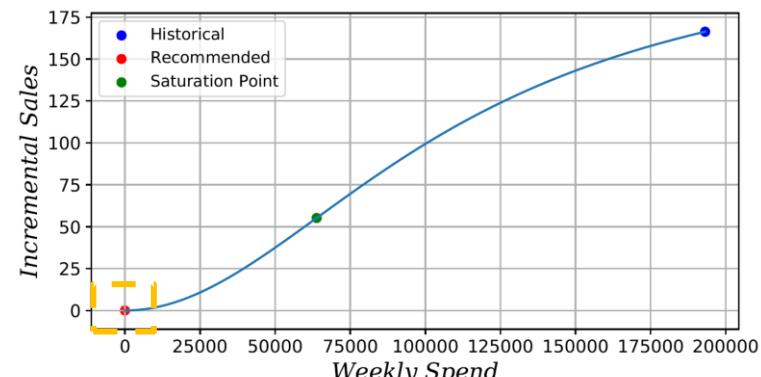
▪ Indirect Price Discount



▪ Direct Marketing



▪ Sales Staff Support



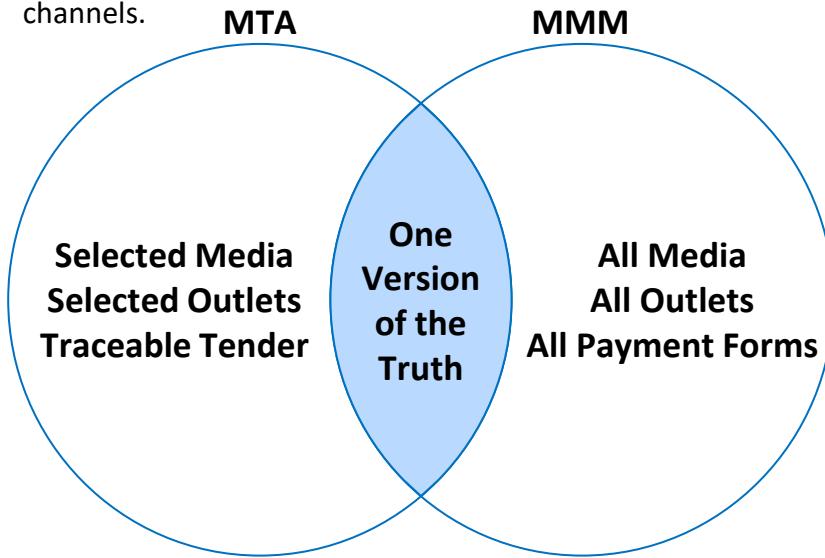
현재 프로젝트 리스트

➤ Multi Touch Attribution of Customer using Attention based Recurrent Neural Network

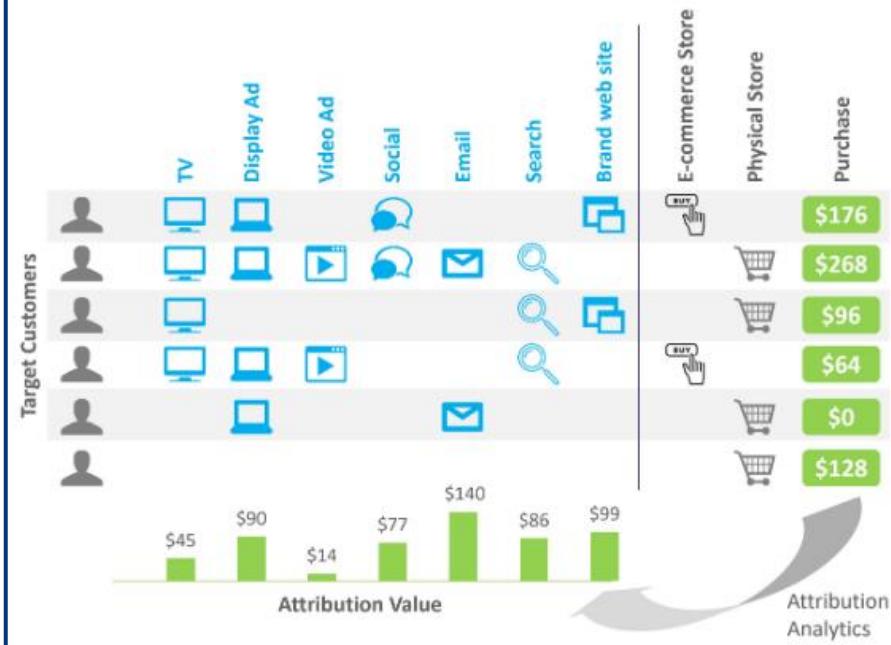
- Multi Touch Attribution (MTA) pertains to the question of **how much the marketing touchpoints a user was exposed to, contributes to an observed action by the consumer.**
- Understanding the contribution of various marketing touchpoints is an input to **good campaign design, to optimal budget allocation** and for understanding the reasons for why one campaign worked and one did not.

MTA vs. MMM

- MTA has very detailed data, and can do better than MMM on very targeted media. But MTA generally can't cover all sales channels.



How MTA Works



현재 프로젝트 리스트

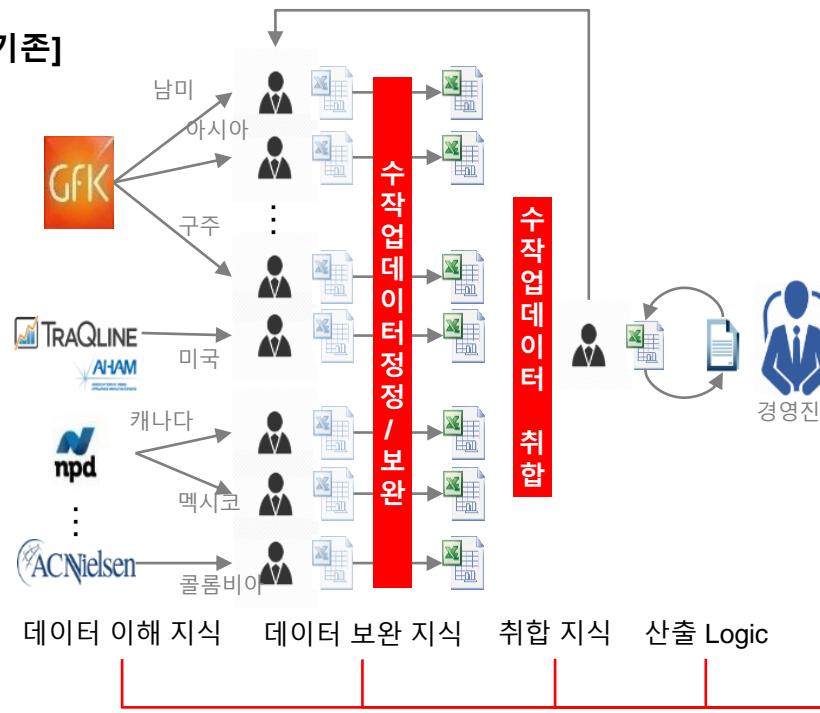
➤ 수작업 기반 M/S 산출 프로세스의 자동화를 위한 데이터 통합 분석 모델 제공

- M/S 산출 위한 기반 Data 표현/Schema 이질성 및 오류 문제로 분석을 위한 전처리에 비효율적 공수가 반복
- 이종 Data 이질성 극복 및 취합/산출全과정의 지식을 Encoding한 Knowledge Graph 구축으로 자동 전처리
- Text mining 기술을 활용한 데이터 내 오류 자동 정정 기능 제안
- → 수 분 내 M/S 자동 산출 (Idea 시스템 연계) 통한 경영진 의사결정 속도 향상 Feasibility 검증 완

제품군/지역별 담당자 다수 관여

(최대 1주일 (30여명), 숙련도에 따라 정확도/소요 시간 가변적)

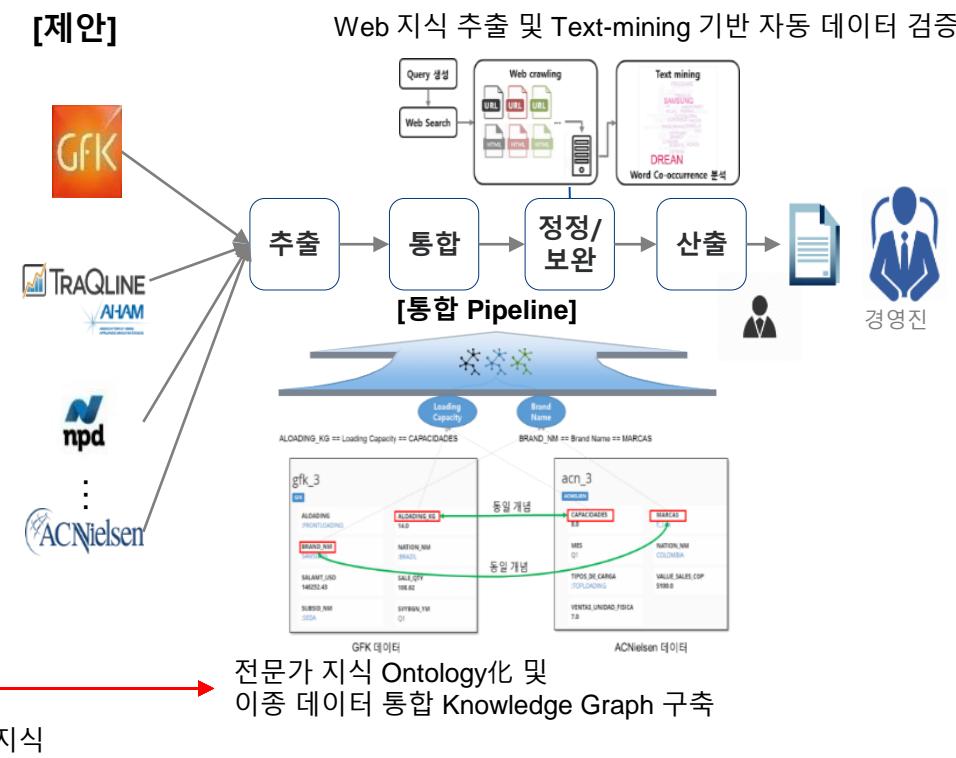
[기준]



Human error 발생 없이 최소 인력 소요

(수 분 내, 최소 시스템 운영 인력 등)

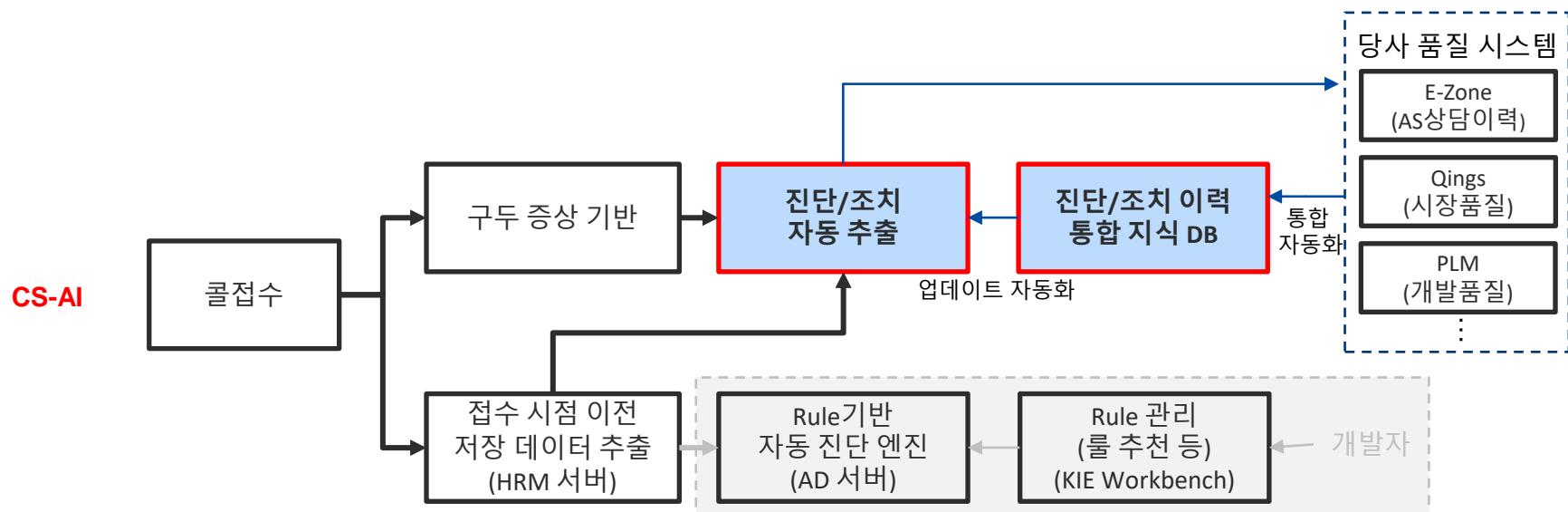
[제안]



현재 프로젝트 리스트

➤ 고객 Claim Data의 불완전성 개선위한 콜센터 자동 고장진단 CS-AI 엔진

- M/S 산출 위한 기반 Data 표현/Schema 이질성 및 오류 문제로 분석을 위한 전처리에 비효율적 공수가 반복
- 이종 Data 이질성 극복 및 취합/산출全과정의 지식을 Encoding한 Knowledge Graph 구축으로 자동 전처리
- Text mining 기술을 활용한 데이터 내 오류 자동 정정 기능 제안
- → 수 분 내 M/S 자동 산출 (Idea 시스템 연계) 통한 경영진 의사결정 속도 향상 Feasibility 검증 완



현재 프로젝트 리스트

조직	분석 목적	결과 및
경영혁신 센터 SCM그룹 RCM그룹	<ul style="list-style-type: none">▪ SCM TV Data 활용한 Sell-Out(S/O) Forecast 정확도 향상<ul style="list-style-type: none">· 프로모션 투입에 따른 판매효과를 예측하여 프로모션 최적화 및 업무 효율화· Data : 학습('16.1~52주), 검증('17.1~32주)	<ul style="list-style-type: none">▪ 판매량 예측 정확도 (최적화된 DNN 예측 모델)<ul style="list-style-type: none">1주 후 : (지역 기준) SCM그룹 55.4%, SR연 79.3%4주 후 : (지역 기준) 전사 S/O Guide 57.6%, SR연 63.9%(지역+매장) 45.4% 50.1%▪ SCM그룹의 S/O Forecast 정확도 고도화 시스템 구축 프로젝트에 SR연 DNN 모델 반영 협의 중
가전사 경영혁신그룹 영업혁신그룹	<ul style="list-style-type: none">▪ 엑셀 수작업을 통한 M/S 산출 업무를 자동화하는 분석 모델 제공<ul style="list-style-type: none">· 기존 30여명의 마케팅 인력이 수작업으로 오류 및 변경 사항을 검토한 후 글로벌 M/S 산출 (1주일 소요)	<ul style="list-style-type: none">▪ Gfk, AC닐슨 등 다양한 이종 Data를 통합 및 Data 오류자동 정정 후 M/S를 계산(5분 이내 소요)<ul style="list-style-type: none">- 이종 Data 통합 기술 (Graph DB)- 모델명 패턴 분석에 의한 모델명 오류 정정→ 경영진 의사결정 속도 향상
가전사	<ul style="list-style-type: none">▪ 시스템 에어컨 운전 Data 분석을 통한 이상 감지 모델<ul style="list-style-type: none">· 비효율 운전 감지 통한 에너지 절감 유도· 비정상 운전 패턴 감지 통한 점검 유도· 진행성 고장 감지 통한 비효율 운전 방지	<ul style="list-style-type: none">▪ 이상 감지 예측 정확도 $F1 > 0.8$ 확보(※ MS사 $F1 > 0.8$)<ul style="list-style-type: none">- 비지도 온라인 러닝을 통해 기기 및 운영 환경에 따라 실시간으로 자동학습 기술 확보▪ 빌딩IoT사업팀에서 기술을 실제 상품화 적용 예정

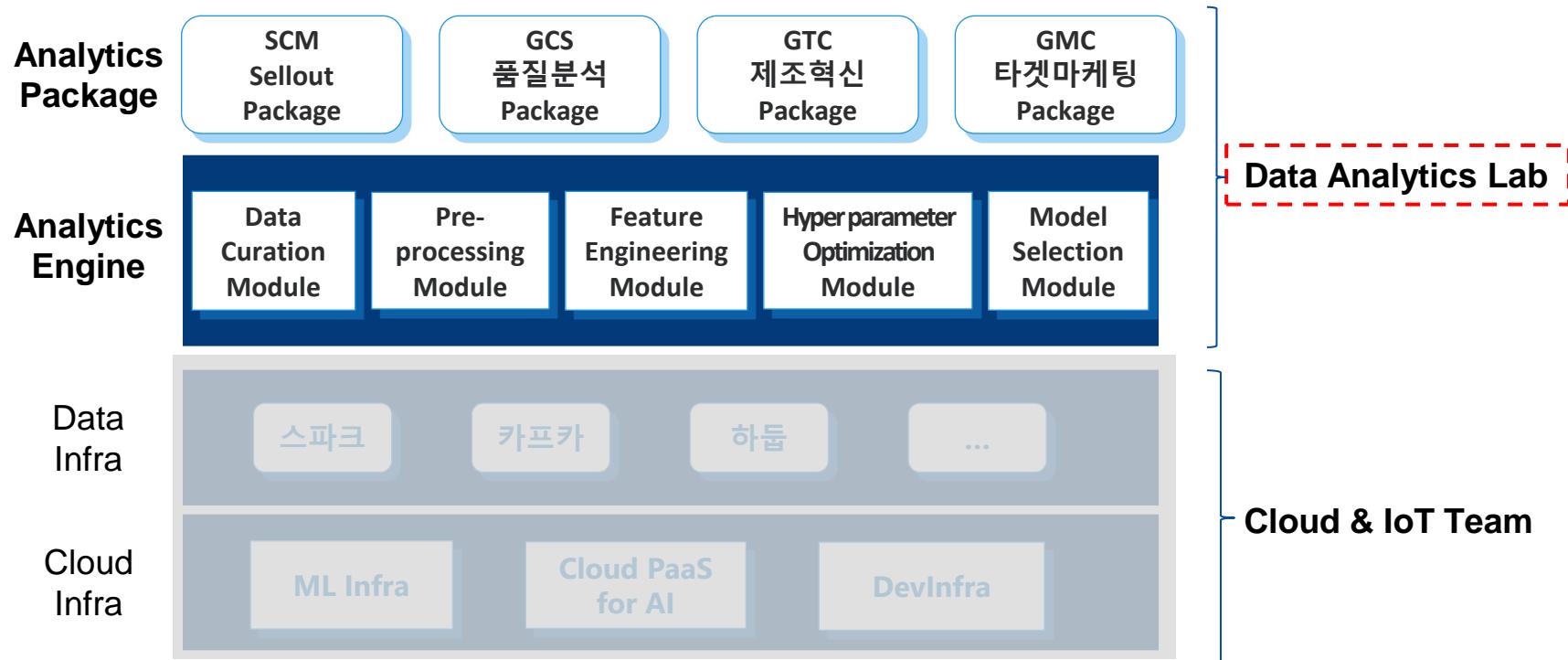
현재 프로젝트 리스트

조직	분석 목적	결과 및 현황
네트워크	<ul style="list-style-type: none">인도 VoLTE망 Call mute 원인 자동 분석 (현재 수작업 분석)	<ul style="list-style-type: none">기지국 로그데이터(VoMA v610) 기반 분석 정확도 93%- 32종 오류 원인에 대한 정확한 판정- 학습 모델 생성 및 분석 툴 제공
SRIN연	<ul style="list-style-type: none">인니 시장 대상으로 S8 판매 및 사용자 정보 Data를 활용하여 N8 구매 가능성을 분석후 타겟 마케팅 효과 제고 추진	<ul style="list-style-type: none">AUC 기준 0.90 정확도 확보- N8 판매 추이를 보면서 지속 검증 중
GCS센터	<ul style="list-style-type: none">QINGS Data를 분석하여 자동 고장 진단 엔진 개발<ul style="list-style-type: none">콜센터 상담 시간 감소정확한 문제 진단으로 엔지니어 방문감소	<ul style="list-style-type: none">시스템에어컨 관련 Data를 우선 활용하여 POC 완료- Bayesian Network 활용하여 자동 고장 진단 및 조치방안 도출 Feasibility 검증
GMC	<ul style="list-style-type: none">버즈 Data 분석을 통한 S7 판매량 예측<ul style="list-style-type: none">제품 출시 4주후 년간 판매량 예측	<ul style="list-style-type: none">Neural Network 활용한 예측 모델 개발- 에러율 2.64달성(97.36% 정확도)

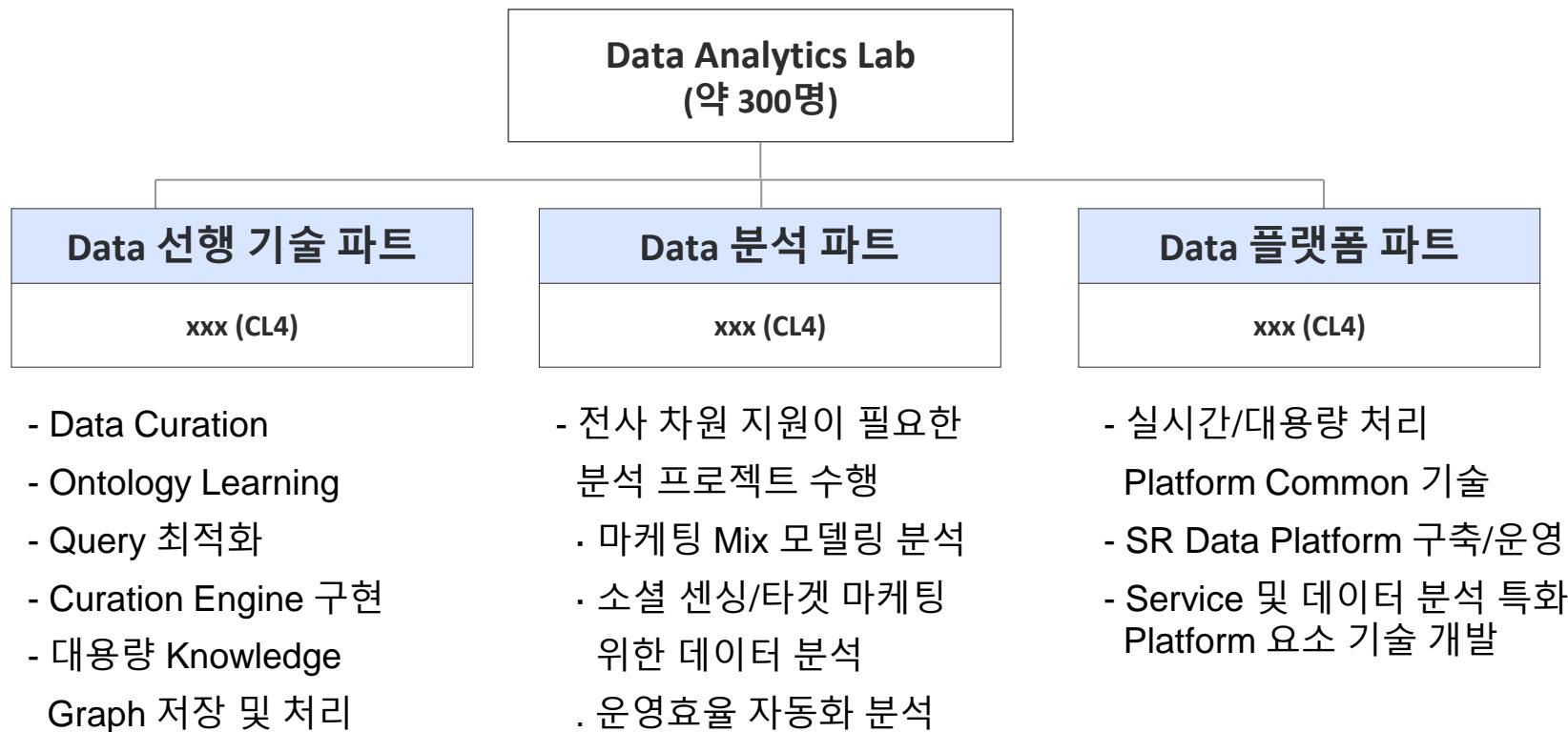
Data Analytics Lab 추진방향

➤ Data Curation 및 자동 Feature Engineering 적용 기존 데이터 분석 엔진 한계 극복

- Analytics Package : 분석 Code(API), Model, Query, Report, Graph(Visualization)

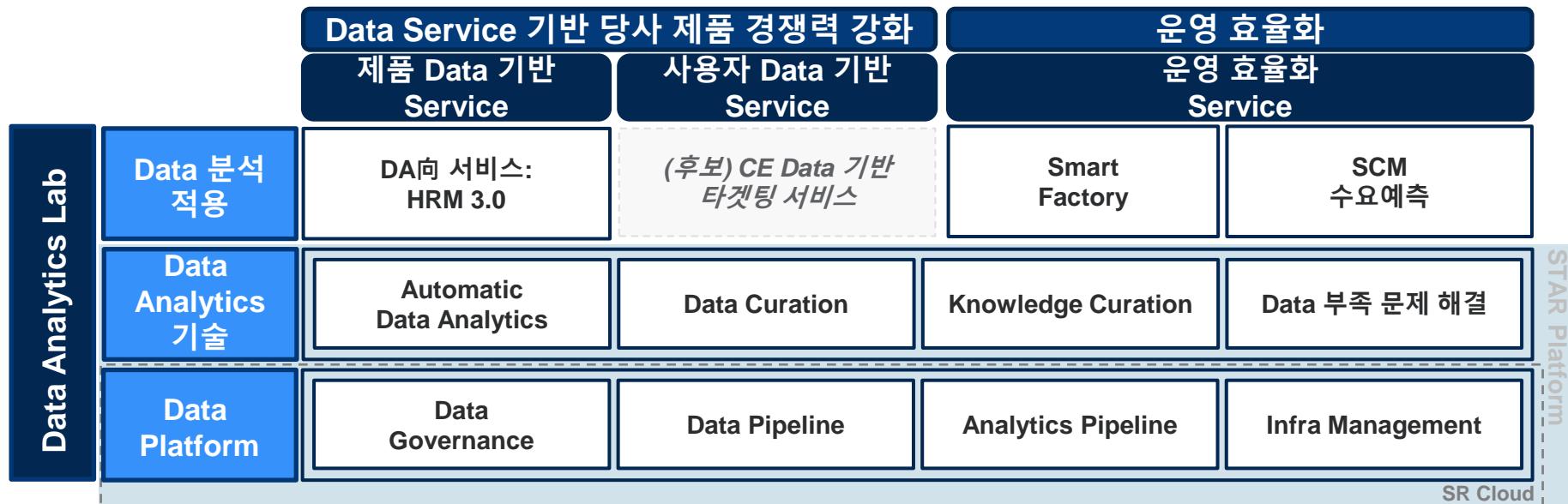


Data Analytics Lab 추진 방향



Data Analytics Lab 추진 방향

당사 데이터 문제점 극복 및 활용성 극대화를 위한 데이터 처리/분석 Core기술과
데이터 수집/분석 플랫폼을 확보하여 제품/서비스 차별화 및 경영 효율화 추진



Data Analytics Lab 선행 기술

➤ 데이터 문제점을 극복하여 활용성 극대화를 위한 데이터 처리/분석 Core 기술 확보

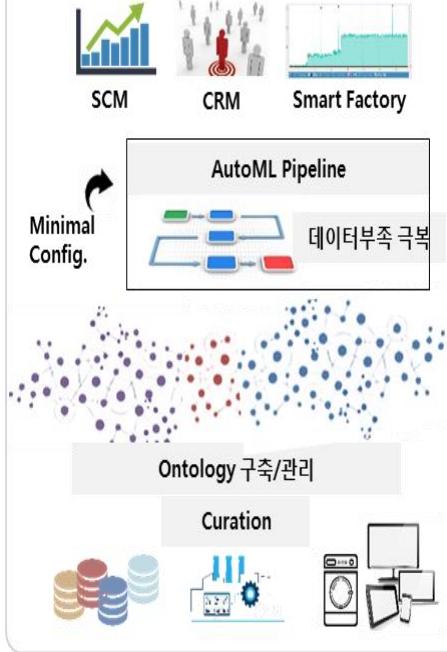
Data Curation

- 이종 데이터 연계: 학습 기반 이종 데이터 레코드 및 필드 간 매핑 자동 추정
- Conflict 해소: 데이터 간 표현 이질성 해소
- NER/관계추출: 비정형 데이터로부터 관계 추출
- Uncertainty 관리: 수집 데이터의 불확실성을 고려한 추론 신뢰도 확보

Feature Engineering & AutoML

- Feature Generation: 데이터 변형/조합 및 외부 유관지식 동원 Feature 확충/선별
- Model Selection: 문제 해결에 최적인 algorithm 및 Hyper-parameter 탐색/추정
- Model Interpretation: 선정 모델 해석력 제공

데이터 처리/분석 Core 기술



Ontology Learning

- Virtual Ontology: 기존의 이종 RDB 시스템 데이터 소스 유지하며 Ontology로 연계
- Ontology Construction: 비정형 데이터로부터 의미를 추출하여 Ontology 구조를 학습/생성
- Ontology Evolution: 신규 데이터 및 외부 지식 반영
- Query rewriting: 대용량 Ontology 질의 처리 및 Reasoning 속도 확보

데이터 부족 극복

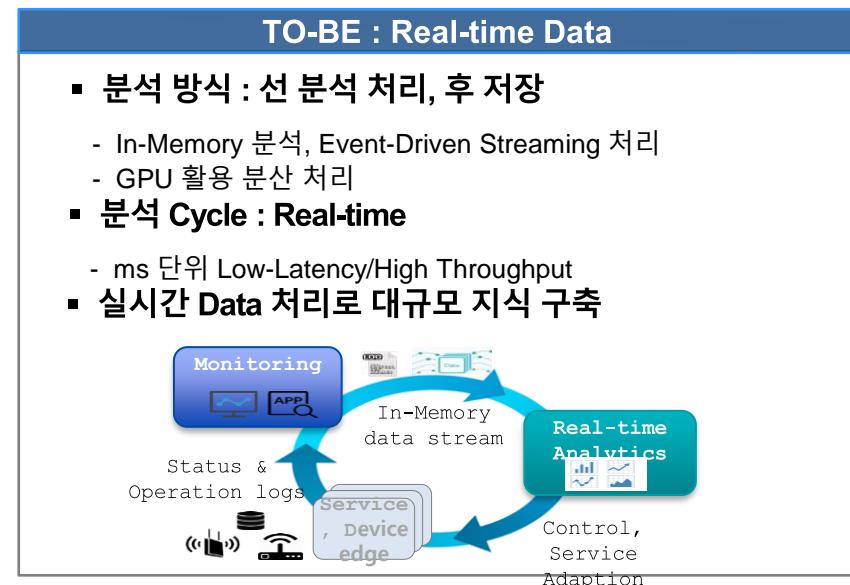
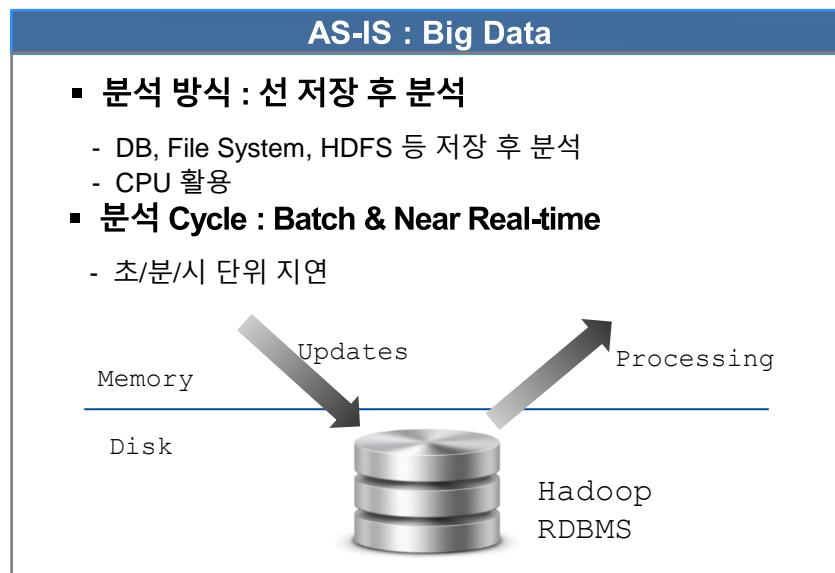
- Bayesian Imputation: 데이터의 시간 의존성 고려한 결측 데이터 estimation
- GAN: 데이터 확충을 위한 생성모델 구축
- One-shot/Transfer BNN: 소량의 데이터로부터 구분 모델 구축

➤ 기술 확보 방안

기술	개발주체	협력선	개발비 (백만원)	개발 기간
Data Curation	자체/사외(업체)	美 Tamr사	200	'18.01~'18.06
Ontology Learning	자체	-	-	'18.01~'19.10
Feature Engineering	자체/해외연/사외(업체)	SRK연, SparkBeyond사 H2O.ai 등	300	'18.01~'18.03 (S사 POC) '18.01~'18.12 (내재화)
데이터 부족 극복	사외(산학)	Cambridge대	전략 산학	'18.01~'20.12

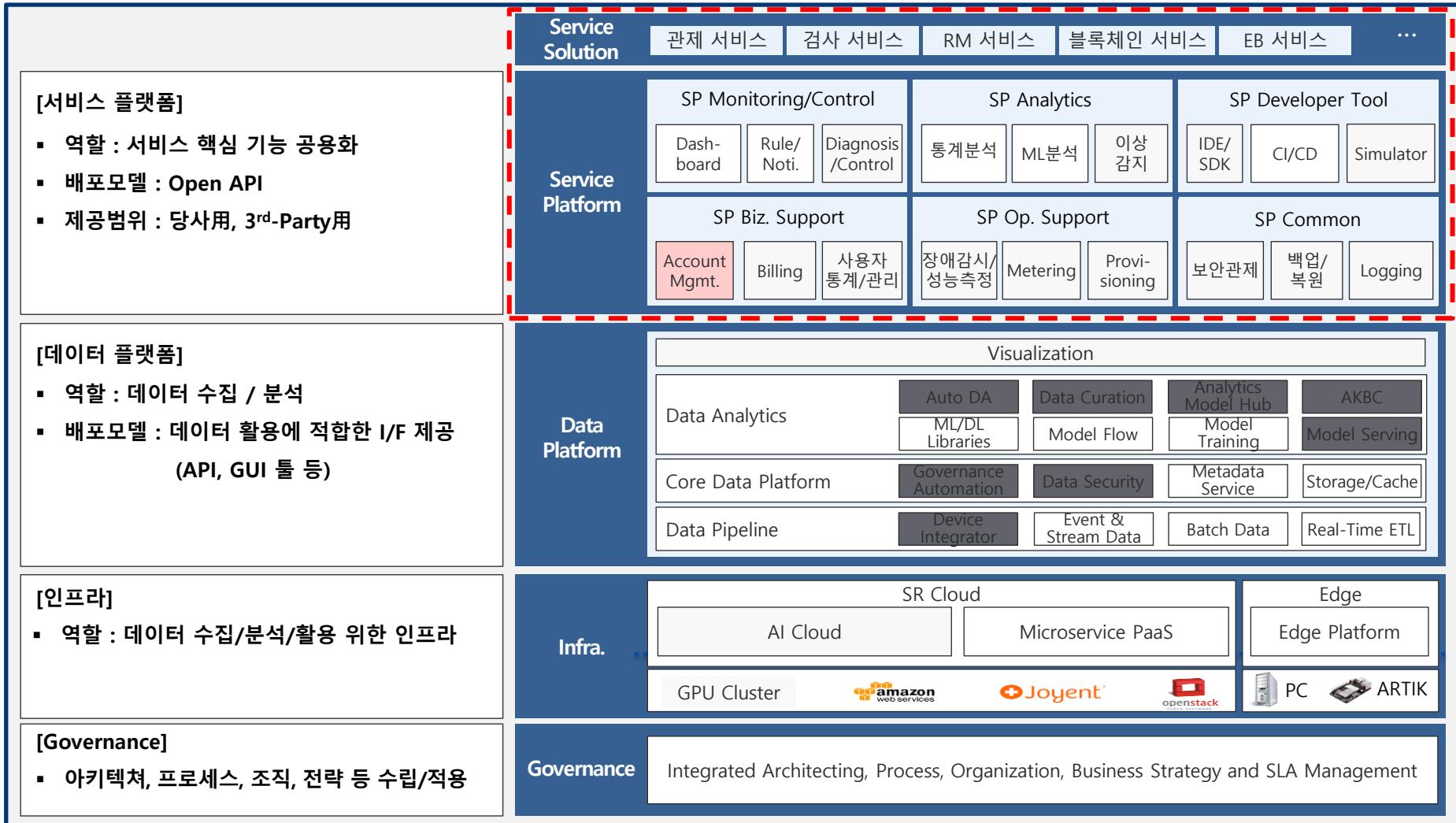
➤ 실시간 Data 처리 플랫폼 및 Knowledge Base 구축

- 실시간 Stream 처리 및 정형/비정형 산업 IoT Data 처리
- 실시간 Data 분석을 지원하는 분산 ML Serving/Training 환경 제공
- Graph DB, Knowledge Base, AKBC (automatic KB construction) 구축



Data Analytics Lab 선행 기술

확보 미확보 무선사



Data Analytics Lab 선행 기술

확보 미확보 무선사

Service Solution	관제 서비스	검사 서비스	RM 서비스	블록체인 서비스	EB 서비스	...
	SP Monitoring/Control		SP Analytics		SP Developer Support	
	Dashboard UI Framework	Rule	Diagnosis	통계분석	Machine Learning	Data Workflow
	Remote Monitoring	Notification	Remote Control	전처리	Model Lifecycle Mgmt.	이상감지
	Real-time Monitoring	Scheduler	Historical Asset Management	시계열 데이터처리	모델 배포	추론예측
Service Platform						
	SP Business Support		SP Operation Support		SP Common	
	Invoicing	Pricing/Rating	SLA Reporting	Provisioning	Service Request Mgmt.	Configuration Mgmt.
	Entitlement/Role Mgmt	Order/Offering Mgmt.	가입자 관리	Inventory	Metering	Service Level Mgmt.
	Account Mgmt.	Billing/정산	사용자 통계/가입/관리 (CRM)	장애감시/성능 측정	Service Activation/Orchestration	Service Catalog

Data Analytics Lab 선행 기술

- 경쟁사는 데이터 분석의 핵심 기능인 Data Curation 및 Automatic Feature Engineering 관련 기술 미비
- H2O, RapidMiner 기술 수준 유사한 것으로 보여 POC 통해 상세 기술력 검증 필요
- IBM Watson Analytics 플랫폼은 타 플랫폼 대비 기술력 우위 : Watson Research Center 통해 최신 기술 연구개발 중

	무선사 빅데이터 분석 플랫폼	SDS Brightics	IBM Watson Analytics	RapidMiner	H2O.ai	MS Azure
목적	데이터 통계 분석 및 시각화 플랫폼	프로그래밍 지식 없이 데이터 분석 가능한 플랫폼	프로그래밍 지식 없이 데이터 분석 가능한 플랫폼	프로그래밍 지식 없이 데이터 분석 가능한 플랫폼	도메인 전문가, DataScientist 위한 데이터 분석 플랫폼	Data Scientist 위한 데이터 분석 플랫폼
Data Curation	★★★ (미제공)	★★★ (미제공)	★★★ (Data Mapping)	★★★ (Data Mapping)	★★★ (미제공)	★★★ (Data Mapping)
Data Preprocessing	(SAS 기반 전처리 제공)	(Sampling 등 기본 전처리)	(Normalization, Transformation 등)	(Transformation, Binning 등)	(Normalization, Transformation 등)	(Normalization, Transformation 등)
Feature Engineering	★★★ (Feature Selection/ Extraction)	★★★ (Feature Selection/ Extraction)	★★★ (Automated Feature Engineering 연구)	★★★ (Automated Feature Engineering)	★★★ (Automated Feature Engineering)	★★★ (Feature Selection/ Extraction)
제공 기능	★★★★ (사용자가 선택)	★★★★ (사용자가 선택)	★★★★ (AUC 기반 모델 선택 연구 등)	★★★★ (자동 모델 선택)	★★★★ (AutoML)	★★★★ (사용자가 선택)
Hyperparameter Optimization	★★★ (미제공)	★★★ (미제공)	★★★ (Derivative-free optimization)	★★★ (Automatic optimization)	★★★ (Random, Grid Search)	★★★ (Grid Search)
Algorithm	(SAS 기반 ML, DL)	(ML, DL)	(ML, DL, Graph 데이터 분석)	(ML, 타 플랫폼 연계 통한 DL)	(ML, DL, Ensemble)	(ML, DL)
Model Evaluation	★★★	★★★	★★★	★★★	★★★	★★★
Data Visualization	★★★	★★★	★★★	★★★	★★★	★★★
UI/UX	★★★	★★★	★★★	★★★	★★★	★★★
API	★★★	★★★	★★★	★★★	★★★	★★★

Contents

I. 데이터 비즈니스란 무엇인가?

- 1) 데이터 분석 프로세스
- 2) (과거) 프로젝트 리스트
- 3) 현실적인 데이터 분석

II. 삼성의 데이터 비즈니스 필요성과 사례

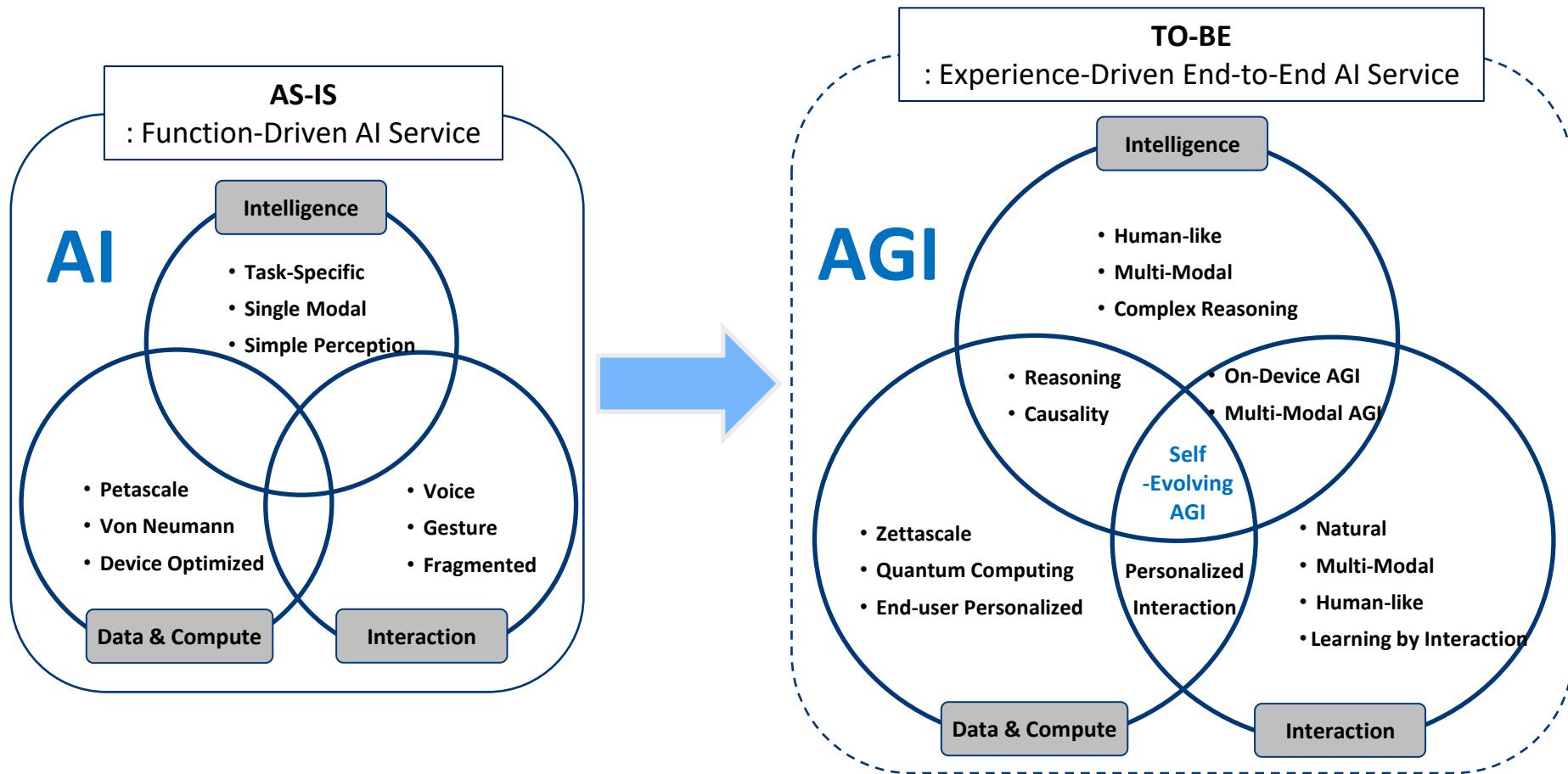
- 1) 삼성전자 구성원
- 2) 데이터 분석 협업 예시
- 3) 데이터분석 목표
- 4) 데이터 현황
- 5) 데이터 분석/개발 이슈
- 6) 현재 프로젝트 리스트
- 7) Data Analytics Lab 추진방향
- 8) Data Analytics Lab 선행 기술

III. Platform of AI Center

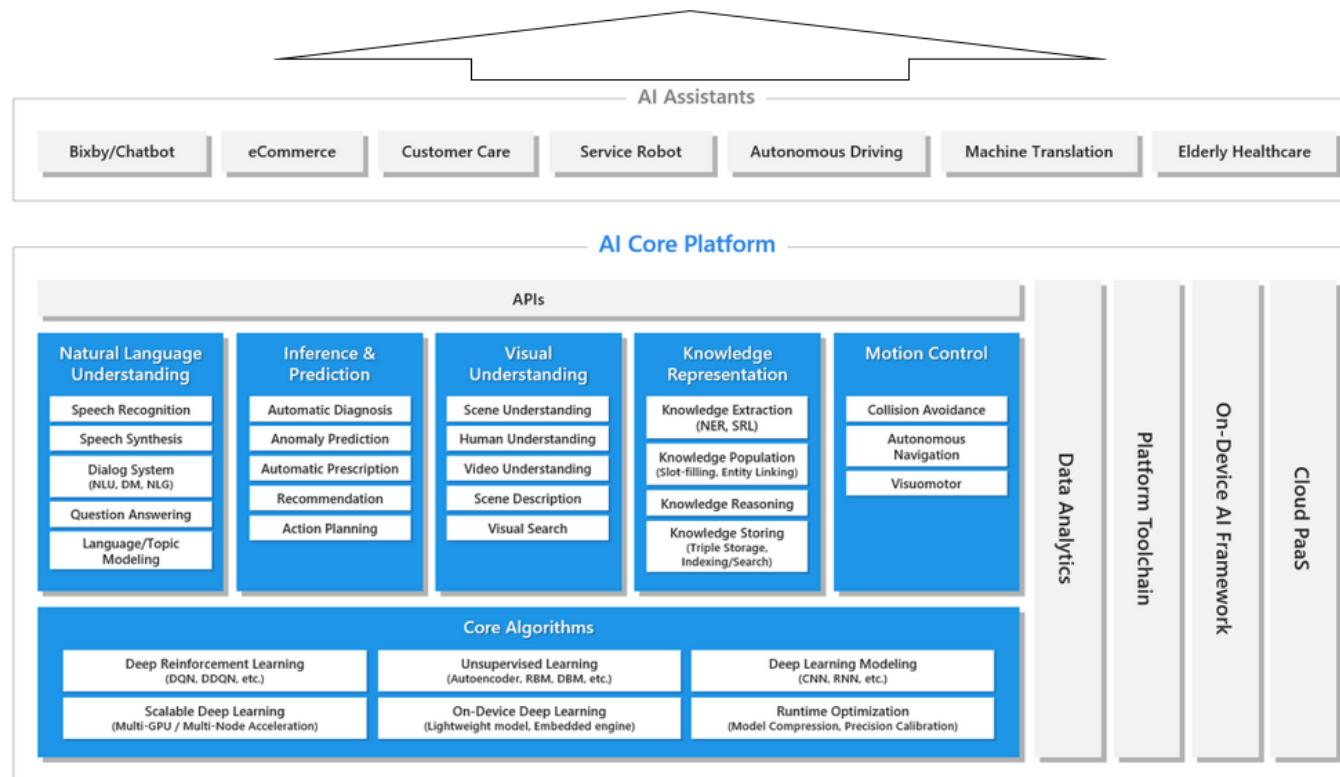
- 1) 대내외 환경
- 2) AI Center

➤ Technology Vision: Beyond AI, Towards Artificial General Intelligence (AGI)

- Enabling new AI technology and **new AI experience** for Data Science Analytics.
- The need for **human-level AI research** in the AL industry including DeepMind, MS and Bengio.



- ✓ Construct and upgrade AI Core Platform (STAR) by combining global capacity
- ✓ Strengthen business competitiveness through Artificial Intelligence



THANK YOU

Q&A