

Process Innovation using operational Big Data

Hyerim Bae

ML with Customer Data

- Mind mining
 - VOC
 - Mind mining
- Usage mining
 - Customer usage data
- Process mining
 - Process improvement

Contents

Part1. Mind mining

Marketing

Part2. Usage mining

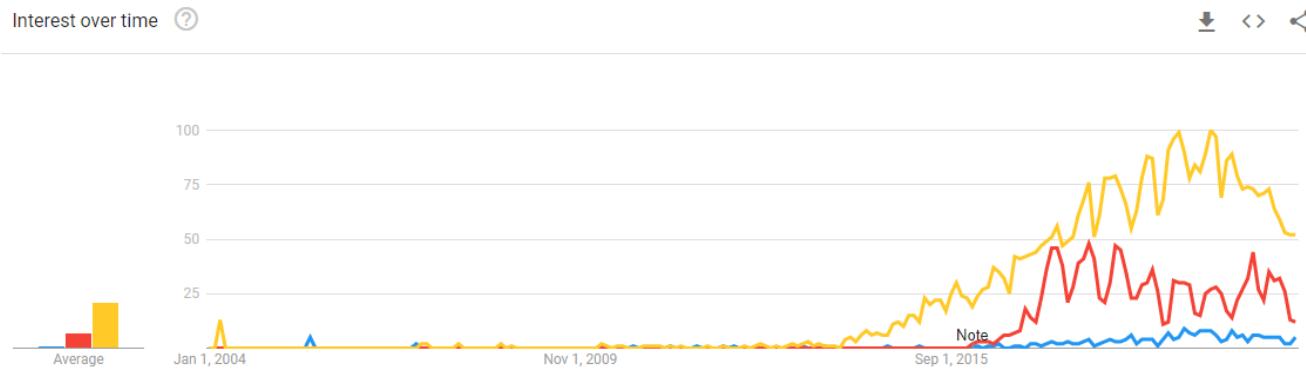
Household electronic appliances

Part3. Process mining

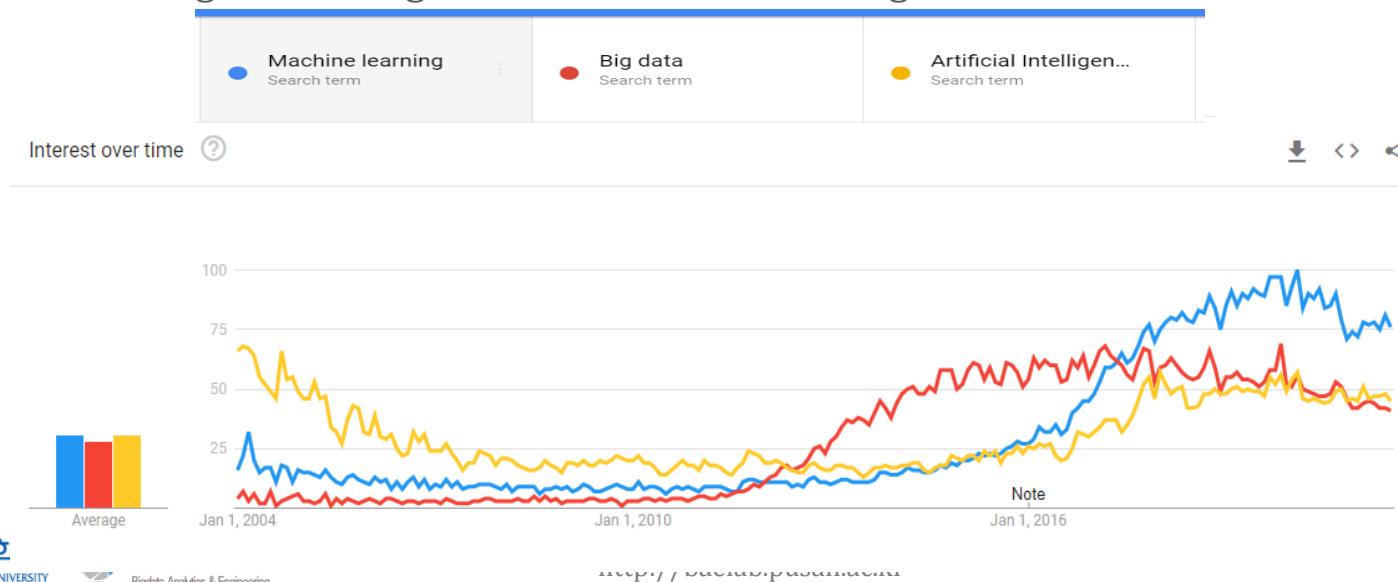
Theory and applications

4차산업혁명 ?

- ‘4차 산업혁명’ vs. ‘The 4th Industrial Revolution’ vs. ‘Industry 4.0’



- Artificial intelligence vs. Big data vs. Machine Learning



Why Big data?

- 속도
 - 트렌드를 즉시 잡아내야 한다.
 - 자료의 생명력
- 커버리지
 - 설문조사: 수십개의 문항으로 응답자의 지난 몇년동안의 변화를 추적할 수 있는가?
 - “엄마가 좋아? 아빠가 좋아?”
- 샘플링
 - 5000명의 생각이 전체 5000만명의 생각과 같은가?
 - “세상에서 가장 높은 산은?” vs. “세상에서 가장 인기 있는 가수는?”



1메가바이트
= 100만 바이트
한 스푼 정도의 모래



1기가바이트
= 10억 바이트
생수통 절반을 채울 정도의 모래



1테라바이트
= 1조 바이트
국민주택(85제곱미터)
아파트에 10센티미터
깊이로 모래 쌓기



1페타바이트
= 1,000테라바이트
해운대 백사장의 모래



1엑시바이트
= 1,000페타바이트
미국 메인 주에서 노스캐롤라이나 주까지 해안선의 모래(한반도 모든 백사장 모래의 합)



1제타바이트
= 1,000엑시바이트
미국 전체 해안선의 모래
(한반도 모든 백사장 모래 합의 1,000배)

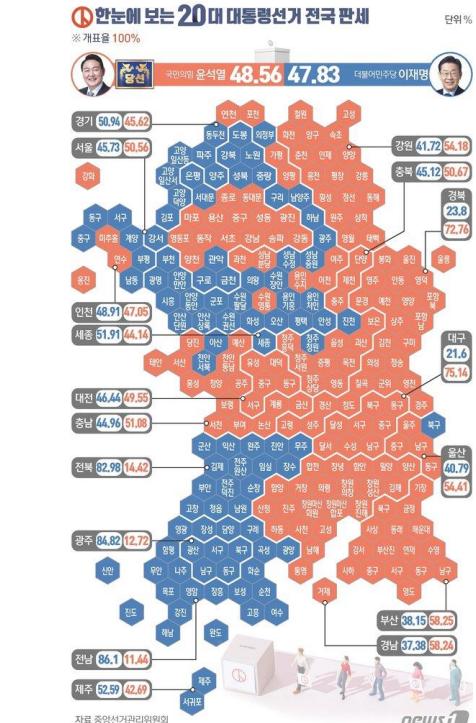
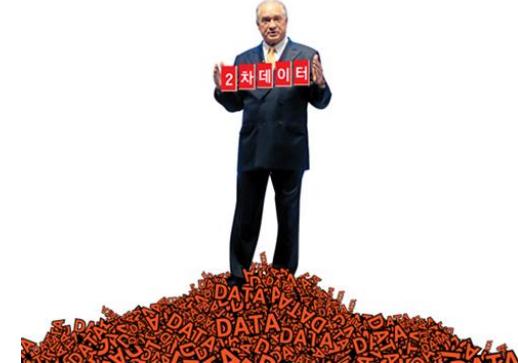


1오타바이트
= 1,000제타바이트
미국 전체를 90미터 깊이로
덮어버릴 모래의 양

출처: 함유근, 채승병, “빅데이터 세상을 바꾸다”

With big data

- Gallup의 실패
 - In 1936,
 - The Literary Digest(천만명) vs. Gallup (5000명)
 - 앨프리드 랜던 or 프랭클린 루즈벨트
 - 샘플링의 승리
 - In 2012
 - 룰니 (52%) or 오바마 (45%)
 - 샘플링의 실패
 - Now, 클리프턴
 - 1차 데이터만의 해석으로는 정확한 예측 불가능
 - 빅데이터는 해석이 중요



Big Data로 할 수 있는 것

(이하 출처: 송길영, “여기에 당신의 욕망이 보인다”)

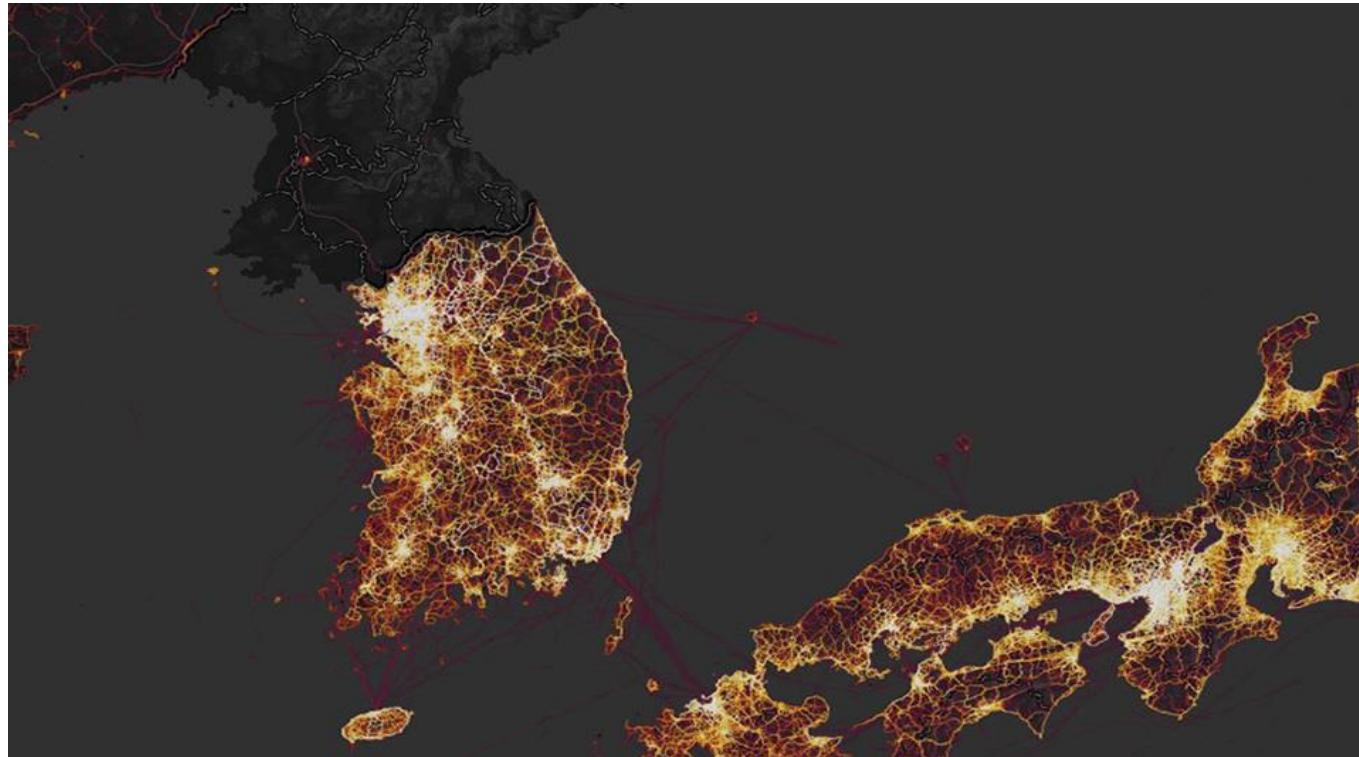
휴가와 기온의 상관관계



98.6% 매출 증가,
2013년 1월 464% 매출 증가
('12.8~'13.1월, 전년 동기 대비)

Big-data

- Why does everybody want to know about big-data?





Usage mining

2021년 조별 프로젝트

프로젝트 배경 - 스마트 가전 선행 연구

〈스마트홈 어플리케이션의 고객반응리뷰분석을 통한 기업별 서비스개선전략에 대한 연구 : 스마트홈 사용성의 기능적 요소와 디자인적 요소 분류 바탕으로〉



클러스터 이름	사용성 가치 효율성	가치 세분화		포함 단어
		시스템/수행 효율성	자원의 효율적 사용	
클러스터 1	안정성	"반응", "응답", "로그인", "종료", "문의", "답변", "로그인", "테이터", "한정 모니터링", "접속"		
클러스터 2	시스템 보안	"해제"		
클러스터 3	오류 허용성	프라이버시 보호 "계정", "수정", 예비 및 회복	오류문제 해결 "업데이트", "연결", "등록", "기능"	
클러스터 4	유효성	인터랙션 안정 "안정", "속도", "시스템", 피드백 제공 "안내"		
클러스터 5	조작 유효성	조작 유효 "권한", "호환"		
클러스터 6	연동성		"연동", "커넥트", "블루투스", "컨트롤", "이상", "원격", "관리", "미리 예약", "없어 죽어"	
클러스터 7	지능화		"예약", "아이콘", "동작", "세부", "사운드", "모니터링", "주시", "조작", "센터", "상담", "매뉴"	
클러스터 8	만족성		"디자인", "위젯", "디바인", "사진", "광고", "인터넷페이스"	
클러스터 9	용이성		"순차", "직관", "목록", "태그", "진단", "처리", "순차", "표시", "모드", "화면", "시작" 등	
클러스터 10	전원의 기능		"알람", "표시", "전원의 기능"	
	비튼 기능		"전원", "버튼", "설정", "제어", "와이파이", "삭제" 등	

스마트 홈 사용자 니즈를 충족시키기 위해 고객이 원하는
사용성 가치에 따라 경쟁우위전략이 필요

〈다항 로짓 모형을 활용한 스마트 공기청정기에 대한 소비자 선호 분석 연구〉

변수 분류	변수	인구통 계학적 변수	다항 로짓 모형 1	다항 로짓 모형 2
			계수값	계수값
경제적 비용	구입가격	-	-0.212***	-0.280***
	필터교체비용		-0.115***	-0.053*
청정 기술	필터등급:H13	-	0.248***	0.191
	필터등급:U15		0.440***	0.312
스마트 제어 기술	공기 청정범위(m ³)	-	0.014***	-0.048***
	소음도(dB)		-0.016***	-0.019***
스마트 제어 기술	스마트 제어 기능: 스마트폰 앱(app)제어	-	-0.018	0.434**
	스마트 제어 기능: 스마트폰 앱(app) 및 AI기반 음성 명령 제어		0.271***	0.591***

다항 로짓 모형 1: 속성 및 속성 수준만 고려한 기본 모형

다항 로짓 모형 2: 인구통계학적 변수와 속성들간 교차항 고려

속성	상대적 중요도	
	구입 가격	39.69%
경제적 비용 관련	필터교체비용	14.35%
	필터등급: H13	6.19%
공기청정 기술	필터등급: U15	10.98%
	소음도(dB)	15.97%
스마트 제어 기술	공기청정범위(m ³)	5.59%
	스마트 제어 기능: 스마트 폰 앱(App)제어	0.44%
	스마트 제어 기능: 스마트 폰 앱(App) 및 AI기반 음성 명령 제어	6.76%

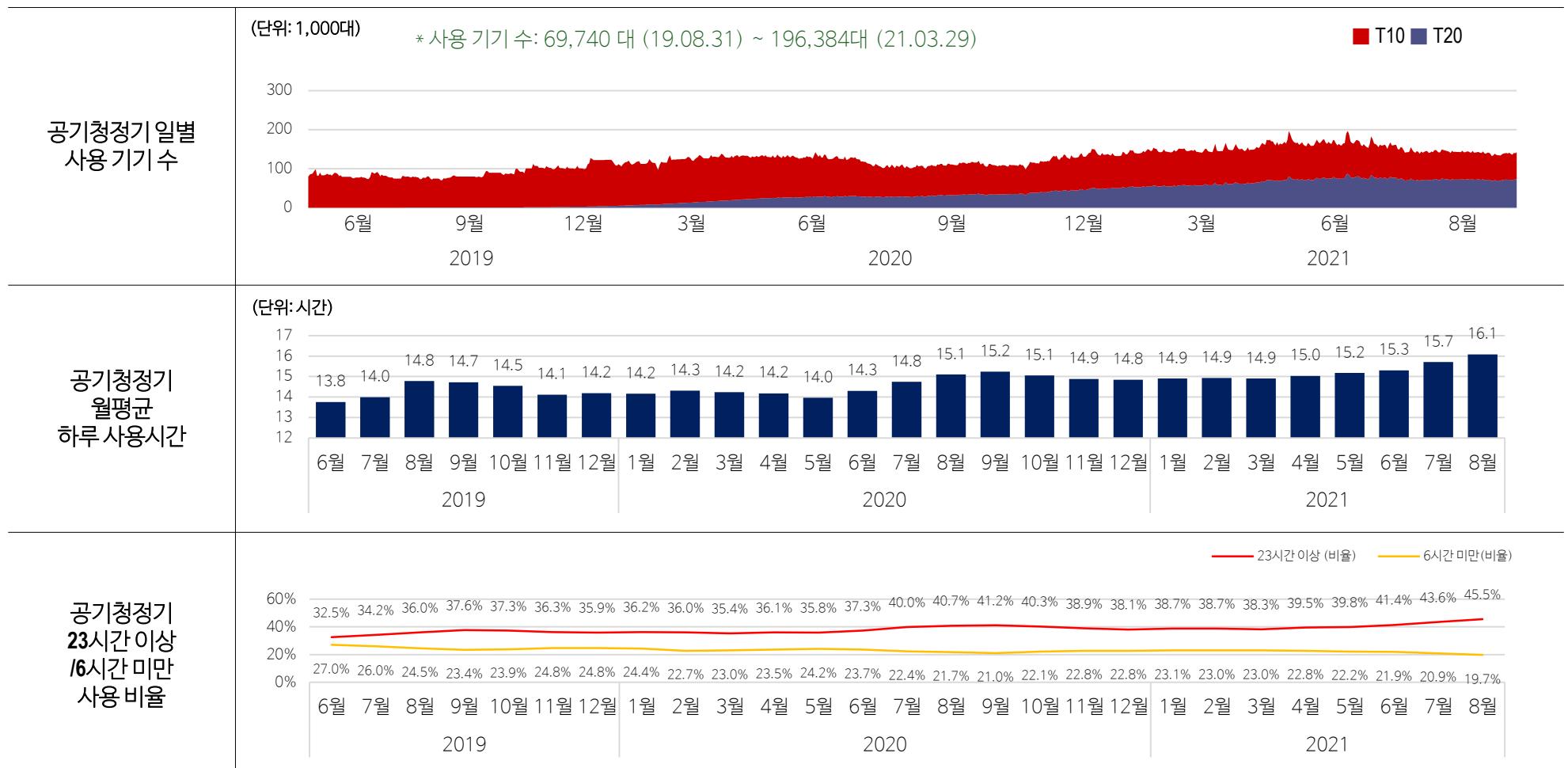
→ 현재 판매되는 공기청정기 상품의 속성들에서 구입 가격, 필터 등급, 소음도, 스마트 제어 기능 등을 주요 속성으로 선정

→ 선정된 속성들 기준으로 소비자 선호 분석 연구 결과, 스마트 제어 기능 탑재에 대한 지불의사가 크며, 스마트가전 시장확대에 따라 제어기술에 대한 중요성 증가

프로젝트 배경 – 공기청정기 사용 Trend 변화

- 사계절 가전으로 자리 잡고 있으며, 사용량은 지속 증가 추세. 특히, 하루종일 사용하는 장시간 사용 고객이 증가하고 있음

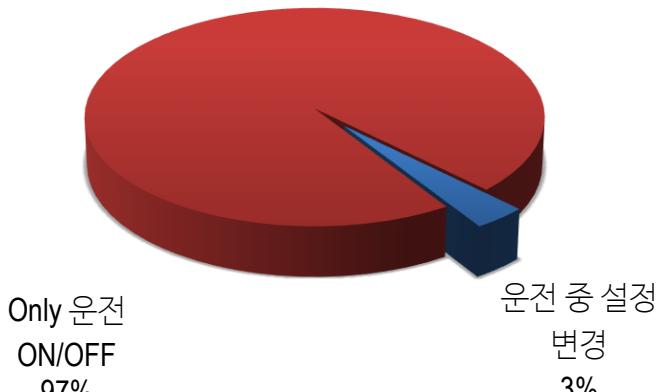
- ✓ 계절성 없이 일정 수준 이상 사용되고 있고, 8월~10월에 사용량이 소폭 상승하는 경향을 보이며, 21년 이후 전체적인 사용량이 증가함이 확인됨
- ✓ 하루 23시간 이상 사용자의 수는 지속 증가하는 추세(32.5% → 45.5%)이며, 6시간 미만 사용자는 20% 초반을 유지하다가 21년부터 감소하는 경향이 뚜렷해짐



* ThinQ 연결된 전체 내수 공기청정기 분석 결과 (차량용 미니 공기청정기 제외)

(1/2)

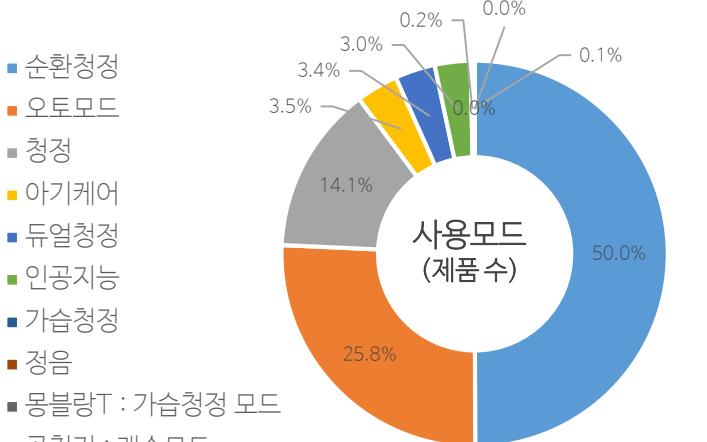
- On/Off 이외에 추가적인 제어를 거의 하지 않으며, 자동 설정의 비중이 높고, 동작의 가시화가 되는 모드를 선호함



<공기청정기 제어 특징 (운전 중 설정 변경 비율)>

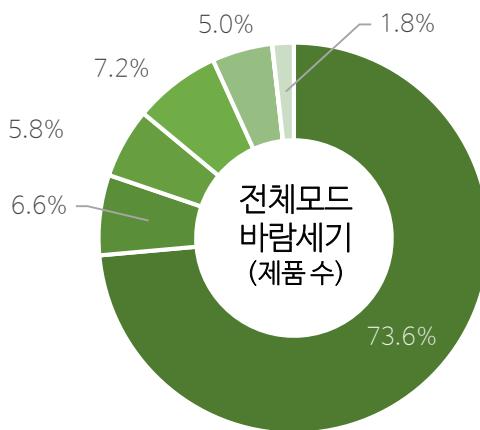
✓ 1대가 1년에 On/Off 이외에 제어하는 평균 횟수 180회

데이터 범위: 국내 / ThinQ2.0 / 공기청정기 전체 (72,041 ea)
데이터 기간: 2021년 3월 12일 (24시간)



<공기청정기 선호 모드 (제품 수 기준)>

■ 자동(A)
■ 강풍(H)
■ 중풍(M)
■ 약풍(L)
■ 파워(P)
■ 미풍(W)
■ 쾌속풍(PD)



<공기청정기 선호 바람세기 (제품 수 기준)>

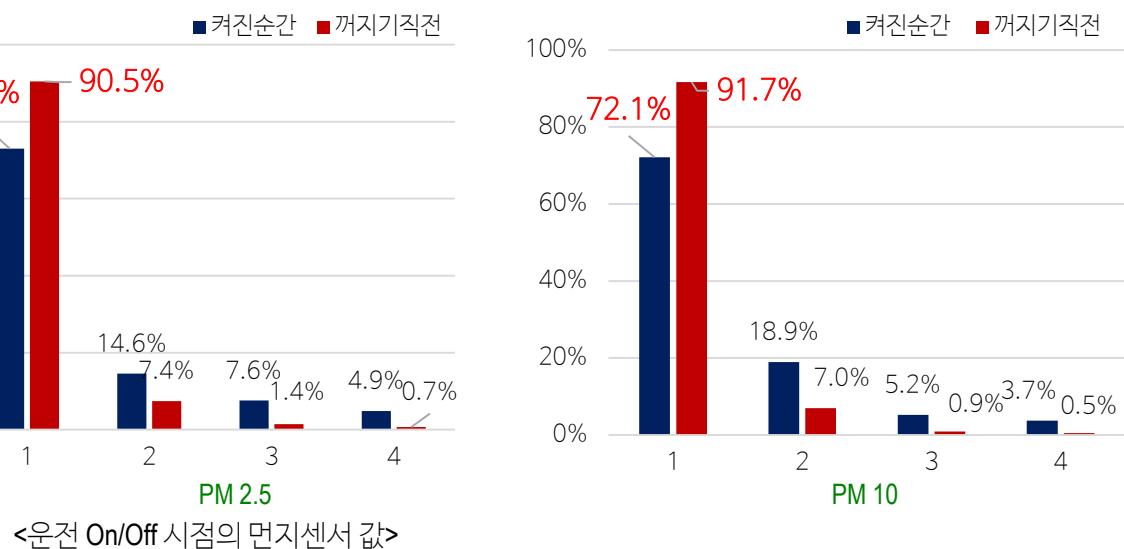
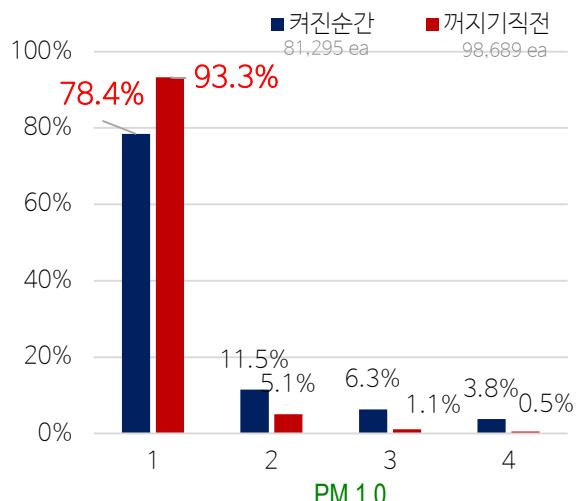
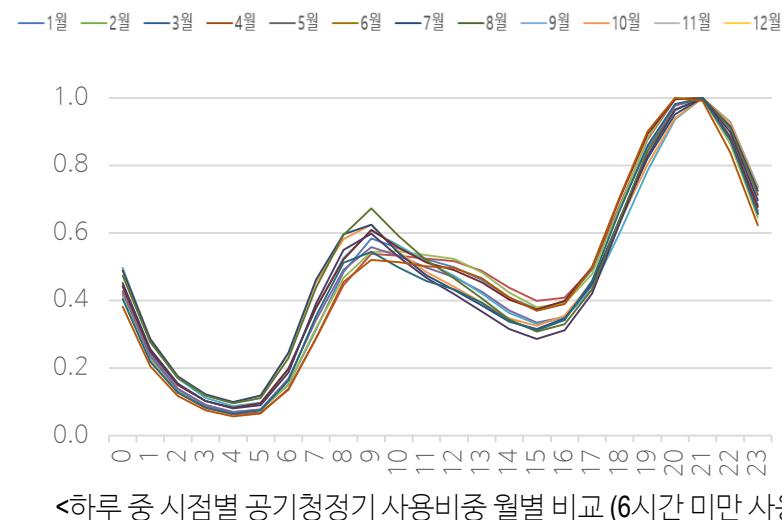
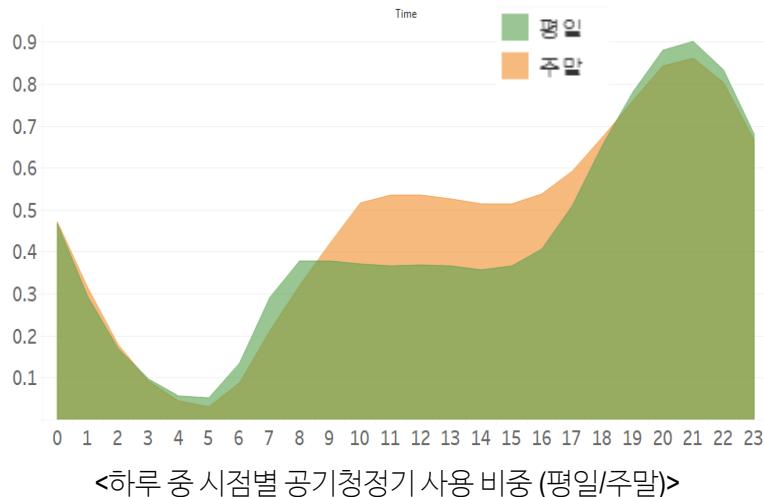
데이터 범위: 국내 / ThinQ2.0 / 공기청정기 전체
데이터 기간: 21년 W27~W34 (8주)

프로젝트 배경 – 공기청정기 사용 특징

(2/2)

- 제품을 사용하는 시점은 공기질의 변화 보다 일상적인 생활패턴에 의해 결정 되는 것으로 추정됨

- 데이터 범위 : 국내 / ThinQ2.0 / 공기청정기 전체
- 데이터 기간 : 20년 1월1일 ~ 12월31일 (1년)



■ 온라인 VOC 분석

- 데이터 출처 : Intellytics

- 기간 : '20.1 ~ '21.11 총 105,625개 VOC 긍정 75.1%, 부정 14.2%

→ 배송, 필터, 소음과 같은 주요 제품 관련 인자 이외에 스마트 제어와 관련된 오토, 자동, 센서 등 keyword 긍정, 부정 평가 분석 수행

Intellytics VOC, '20.1 ~ '21.11, Keyword : 오토, 자동, 센서, 알아서

긍정 평가 예

“자동모드로 해두니 현재 상태를 알려주고 켜야한다고 알림이 와서 좋아요”(자사, 수동으로 모드 변경)
“자동으로 해두니 알아서 파워 조절을 해요”(자사)
“자동으로 정화되어서 좋아요”(자사)
“자동으로 해두면 알아서 파워조절을 해요”(자사)
“핸드폰으로 연동하였더니 공기가 좋지않을 경우 자동으로 알아서 현재 상태를 알려줘 켜야 한다고 알림이 와요. 강추”(자사)
“자동감지센서가 있어서 똑똑해요”(자사)
“자동으로 스스로 모드를 잘 맞춰줘요”(위닉스)
“평일에 자동으로 켜져서 좋아요”(삼성큐브)

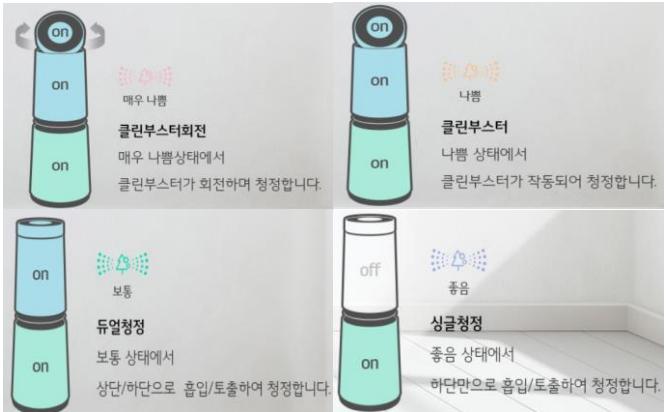
부정 평가 예

“센서가 민감하지않은게 문제예요”(샤오미)
“센서가 약한지 계속 파란불이에요”(위닉스)
“저녁에 불이 환하고 자동으로 꺼지지않네요”(자사, 퓨리케어)
“자동모드가 약한거 같아요”(삼성)
“오토모드가 있으나 전원이 자동으로 켜지지않아요.”(다이슨)
“자동모드가 가끔 오작동해요”(삼성)
“공기가 좋으면 무풍으로 가는줄알았는데 따로 켜야해요”(삼성)
“자동 on/off가 가끔풀려서 불편해요!”(샤오미)
“오토로 해두니 시끄러워서 밤엔 끄고 사용해요”(오아)
“자동모드가 있어 취침모드로 바뀌지만 아침에 취침모드가 off되지않네요”(삼성)

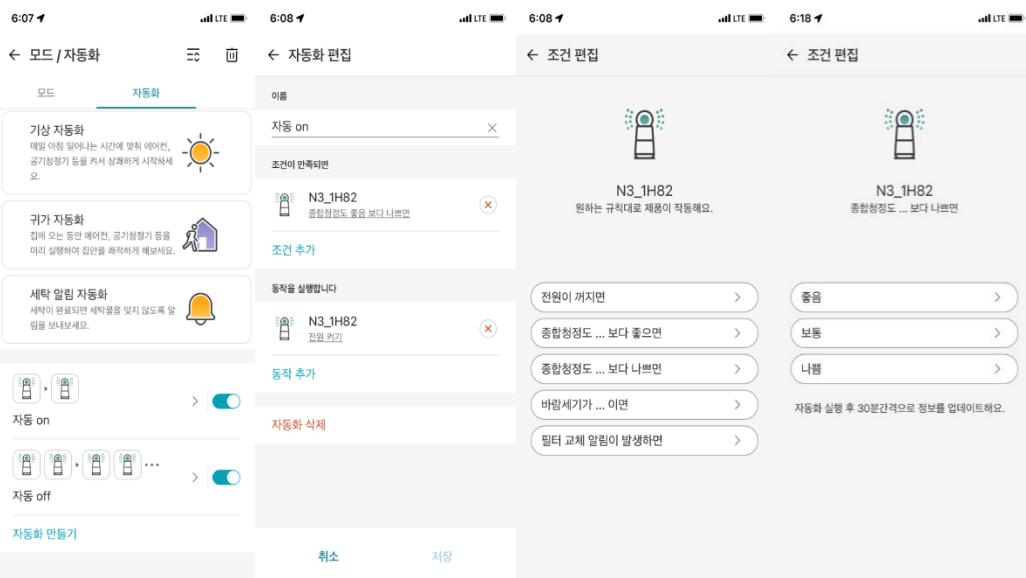
자동모드에 대한 만족감이 크지만, 수동으로 on/off를 이용하면서 불편을 느끼는 고객들이 있음

프로젝트 배경 – 기존 기능 및 고객 Unmet Needs

■ 기존 자동운전 / 루틴 운전



✓ 종합청정도 단계에 따라 운전 조건 자동 전환됨



✓ IFTTT 기반 설정하여 제품 on/off 가능 (30분 단위로 업데이트)

■ 고객 Unmet Needs

▶ 장시간 사용 고객 Unmet Needs

- 공기청정기 사용이 불필요한 시점에 소비전력 loss 존재
- 야간에 소음에 의한 불만이 생길 수 있음
- 주요 부품의 사용 연한이 단축됨
- ➔ 불필요한 시점과 야간에 제품을 완전히 off 하면 소비전력 save 및 소음에 의한 우려를 제거 할 수 있음
- ➔ 제품의 사용 연한을 계속 틀어 놓을 때보다 연장시킬 수 있음

▶ 단시간 사용 고객 Unmet Needs

- 사용하지 않을 때 미세먼지에 노출될 우려가 있음
- 고지 않고 외출하거나 잠든 경우 의도하지 않은 소비전력 낭비
- ➔ 제품을 고객이 동작 시키지 않아도 알아서 공기질이 나쁠 때 운전하면 항상 쾌적한 공기상태 유지 가능
- ➔ 필요한 시점에만 운전하고 일정 조건이 되면 운전 off하면 의도하지 않은 낭비 제거

- ✓ 기존 자동운전은 제품을 완전 off 시키지 않음
- ✓ 루틴운전의 경우 on/off 조건에서 빈번한 on/off 가 발생할 우려가 있음
(현재 임의로 30분 업데이트 주기 반영)

프로젝트 내용 - 제안하는 능동 공기질 관리 모드 컨셉

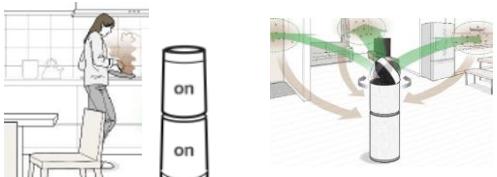


- 고객이 설정한 조건이 되면 운전 On하여 선호 모드로 동작
- 공기질이 안정화되었거나, 사용패턴에 기반한 조건이 되면 자동 운전 Off
- 운전 Off 상태에서도 종합청정도 기반 LED 디스플레이는 유지
- 취침시간에는 스스로 소음을 줄이고, LED 밝기 조절

기능 설정은 리모컨 / 앱을 통한 One Touch 설정 (상세 조건 설정은 부가 기능)

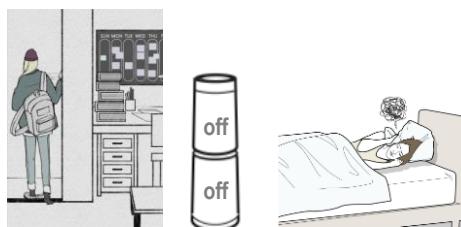
자동 On

- 종합청정도 or 미세먼지 센서 값 기준 만족 시 즉시 제품 운전
- 운전 On 시점 시간정보 / 타기기 취침모드 여부에 따라 선호모드 or 저소음 모드 결정
- 빈집일 때도 오염 감지 시 운전



자동 Off

- 공기질이 좋음 상태에서 안정화되면 제품 off
- 사용자가 주로 외출하는 시간이나 취침하는 시간에 맞춰 제품 off (공기질 좋음 상태에서)



공기질 표시

- 제품 운전 여부와 무관하게 상시 Display
- 주변 조도나 타 기기 취침모드 설정 여부에 맞춰 밝기 조절

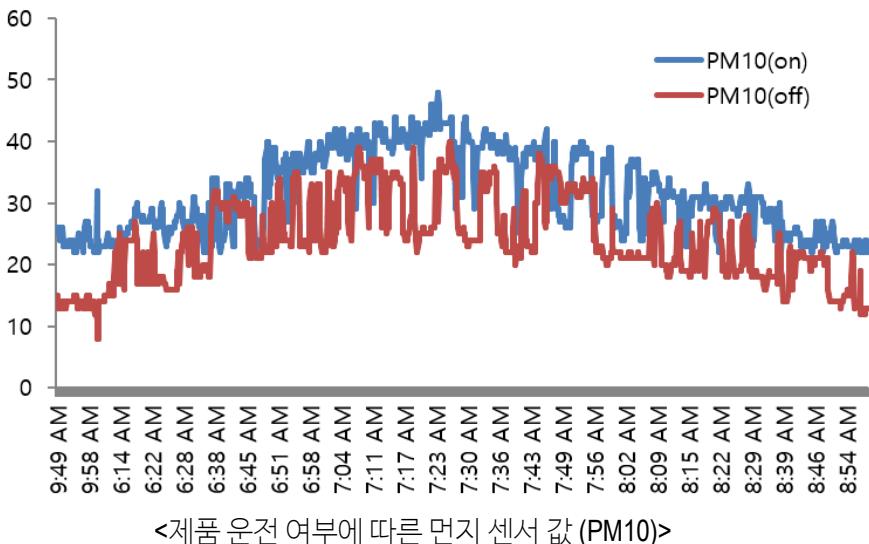


매우 나쁨 나쁨 보통 좋음

프로젝트 내용 - 데이터 기반 문제 해결 방법

▪ On 시점 결정

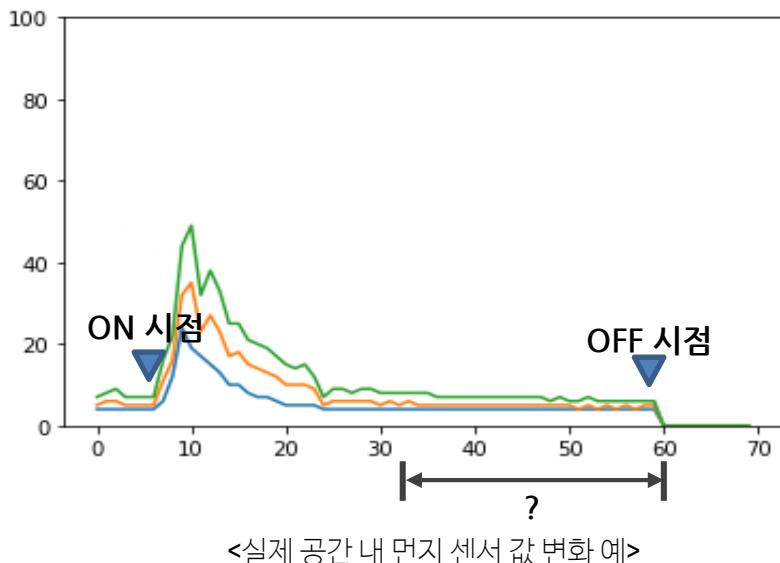
- 먼지의 유입 및 공기청정 패턴을 볼때 일정 기간동안의 데이터를 보고 판단하는 것 보다, 기준값을 넘는게 감지 되는 경우 바로 On해야함
- 제품 Off 상태에서 측정한 센서 데이터와 제품 On 상태에서 측정한 센서 데이터 차이에 대한 보상이 필요



- 동일 조건에서 On상태에서 획득한 센서데이터를 Y값, Off상태에서 획득한 센서데이터를 X값으로 하여 회귀식 도출
- 만들어진 회귀식의 Cost function 구하고, Test Set으로 평가

▪ Off 시점 결정

- 데이터 기반으로 먼지량 변화와 제품 사용간의 상관관계 확인 필요
- 먼지 값 시계열 데이터를 학습하여 Off 시점 예측
(LSTM / CNN 등 다양한 Method 활용)



- 현재 고정 30분으로 설정된 업데이트 주기를 학습된 결과로 변경
 - 사용시간, 사용시점 기준의 고객 분류 필요 예상
- 걱정점: 학습에 필요한 양질에 데이터 셋 분류 어려움
- 사용자별/환경별 차이에 대한 데이터 반영 방안 필요

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(1/6)

- 실제 사무공간에서 획득한 데이터로 Regression 식을 도출하고, Cost function 및 오차를 구함

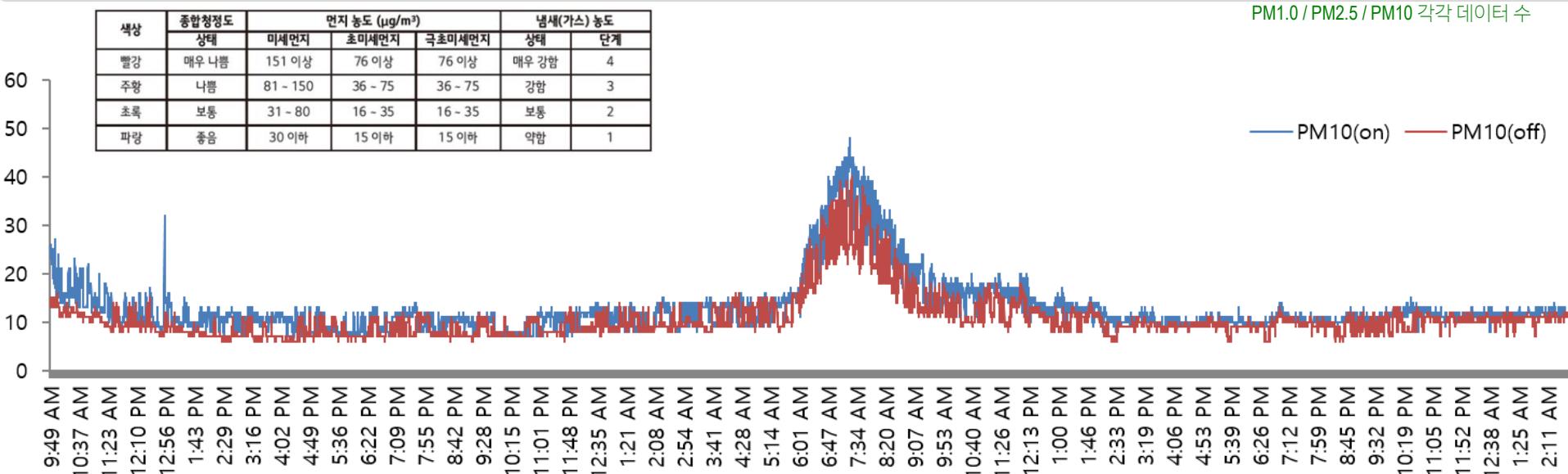
데이터 수집

- 수집 기간: 2021년 11월 15일 09시~17일 03시 (42시간)
- 수집 장소: LG스마트파크2 A2동 3층 사무공간
- 수집 방법: 동일 공간 내 동일한 제품 2대를 설치하고, 1대는 운전 On / 1대는 운전 Off 상태에서 먼지센서 값 (PM1, PM2.5, PM10) 6초마다 저장

데이터 분류

- 수집된 전체 데이터를 8:2비율로 Training Set과 Test Set으로 분류
- 종합청정도 단계별로 데이터를 분류

	전체 데이터 수	Training Set	Test Set
전체 데이터	24,443	19,554	4,889
종합청정도1	22,772	18,218	4,554
종합청정도2	1,671	1,337	334

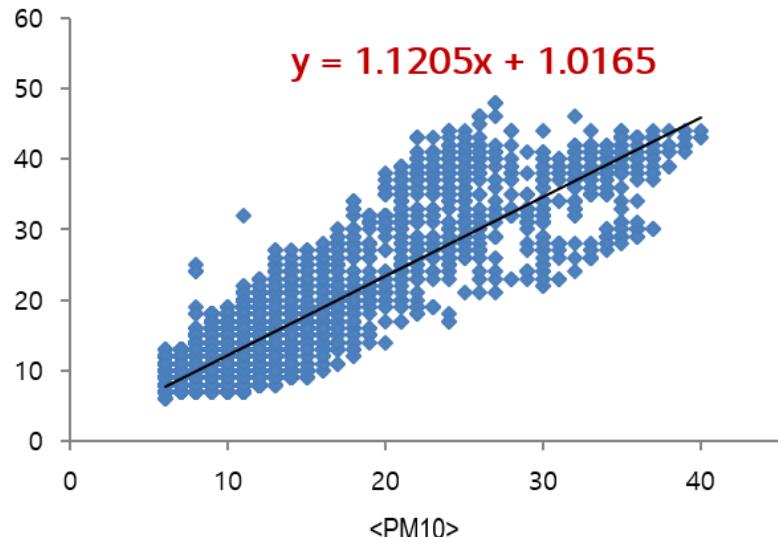
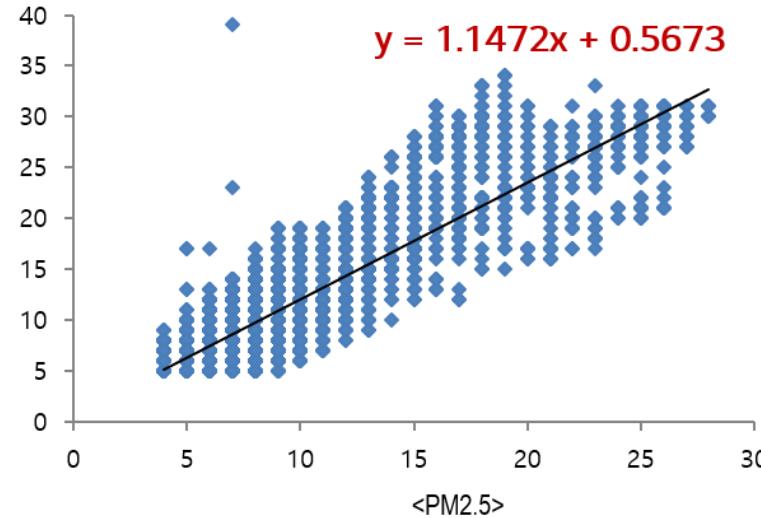
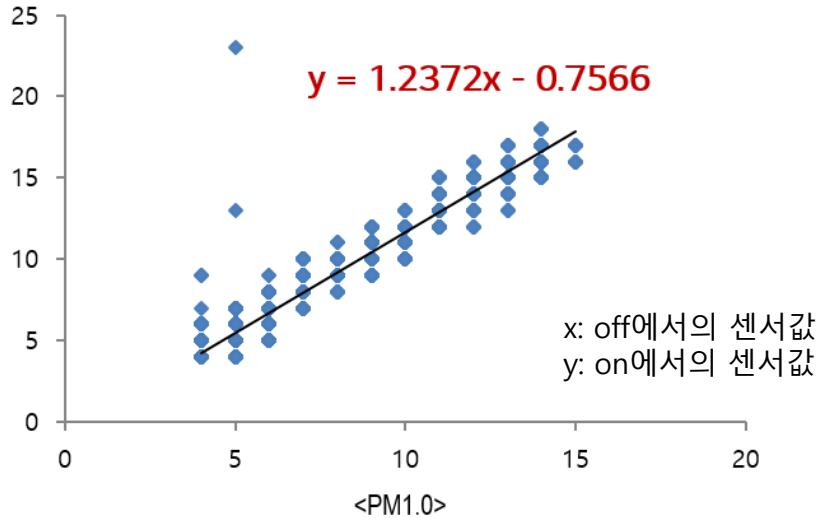


<제품 운전 여부에 따른 먼지 센서 값 (PM10)>

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(2/6)

- 수집한 데이터를 각각의 센서 별로 **Regression** 모델을 이용하여 회귀식을 구하고, 오차를 확인



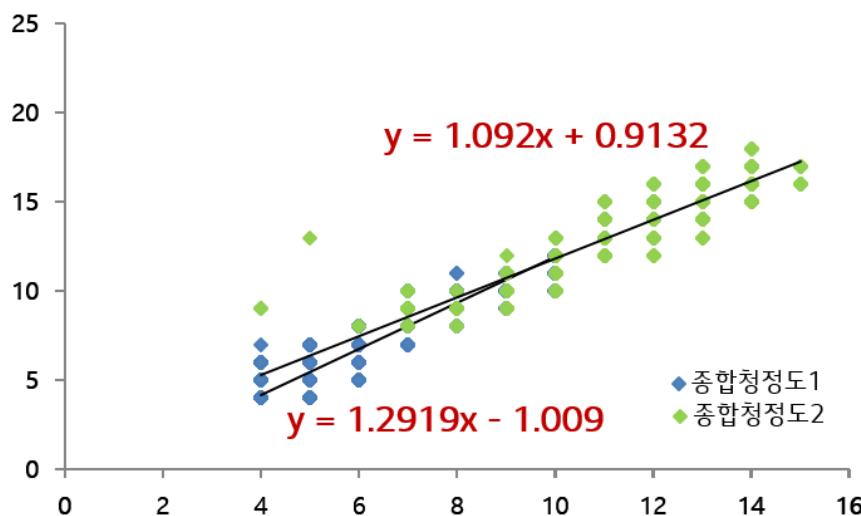
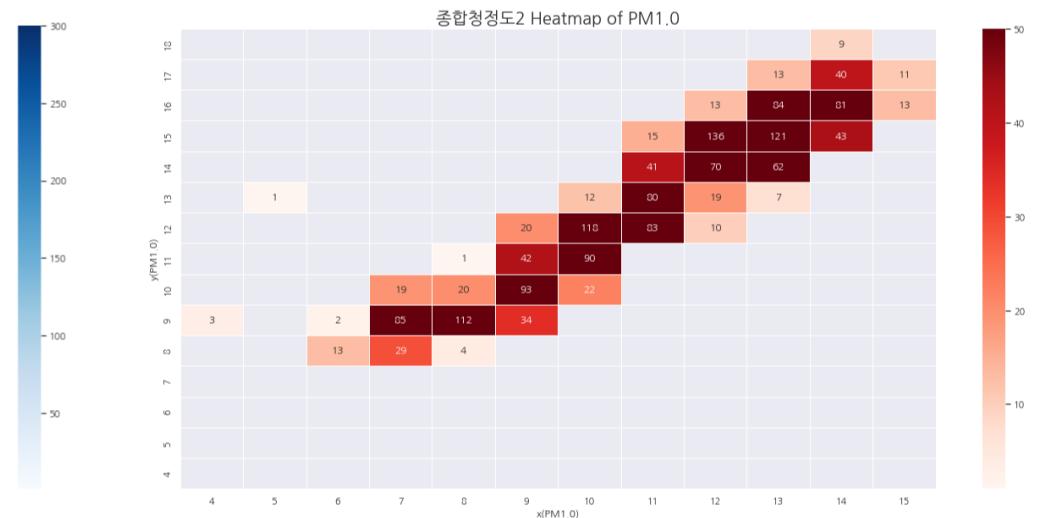
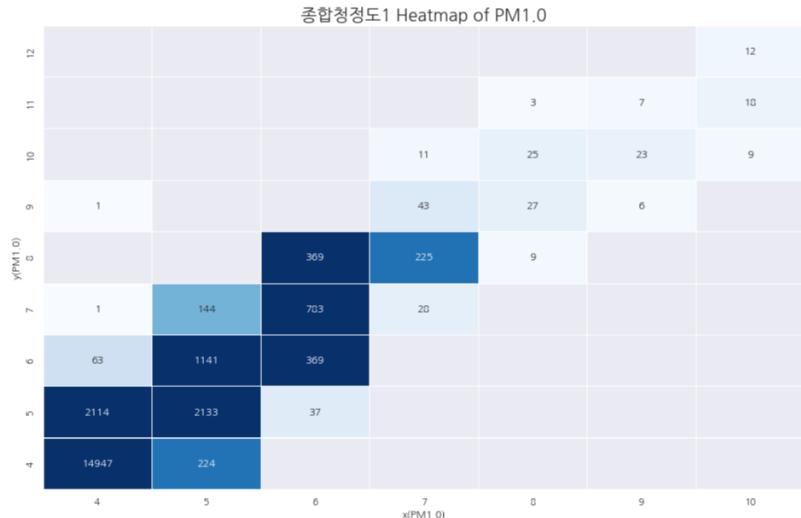
	Cost Function	Training Error	Test Error
PM1.0	$\frac{1}{n} \sum_{i=1}^n t_i - (1.2372x_i - 0.7566) $	0.43	0.21
PM2.5	$\frac{1}{n} \sum_{i=1}^n t_i - (1.1472x_i + 0.5673) $	1.69	1.14
PM10	$\frac{1}{n} \sum_{i=1}^n t_i - (1.1205x_i + 1.0165) $	2.34	1.83

✓ Cost Function은 오차의 직관성을 높이기 위해 MSE가 아닌 MAE로 계산

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(3/6)

- PM1.0 센서 값을 종합청정도 기준으로 구분하여 회귀식 도출 및 오차 계산

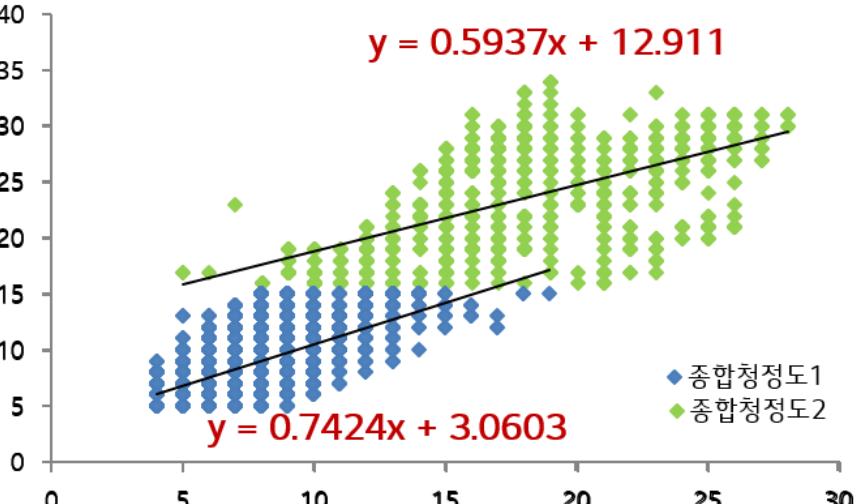
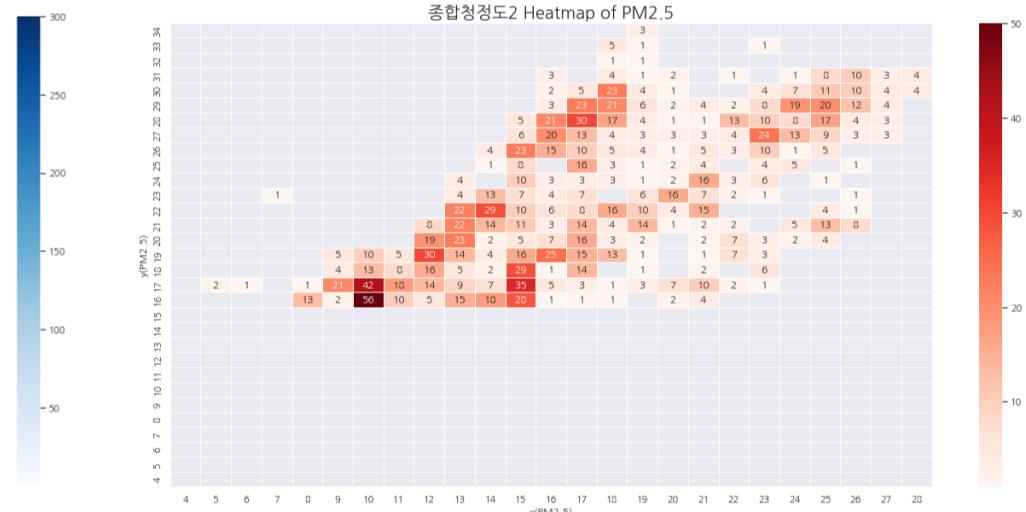
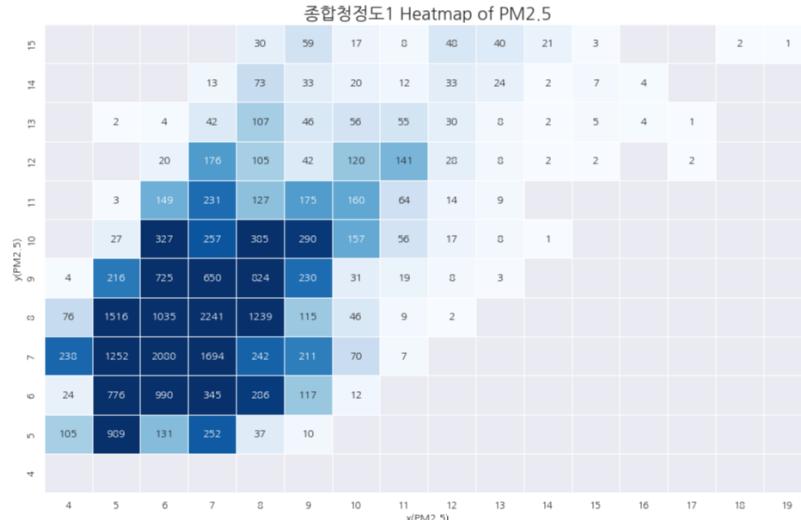


	Cost Function	Training Error	Test Error
종합 청정도1	$\frac{1}{n} \sum_{i=1}^n t_i - (1.2919x_i - 1.009) $	0.37	0.18
종합 청정도2	$\frac{1}{n} \sum_{i=1}^n t_i - (1.092x_i + 0.9132) $	0.73	0.67
전체 데이터	$\frac{1}{n} \sum_{i=1}^n t_i - (1.2372x_i - 0.7566) $	0.43	0.21

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(4/6)

- PM2.5 센서 값을 종합청정도 기준으로 구분하여 회귀식 도출 및 오차 계산

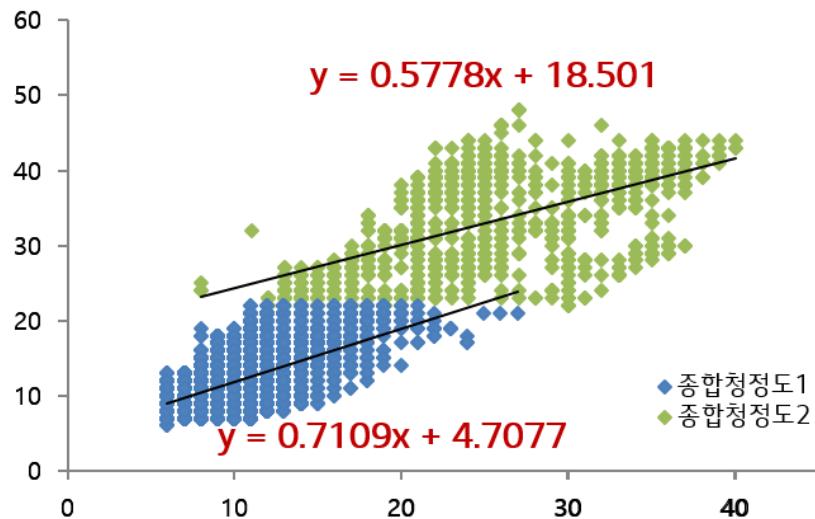
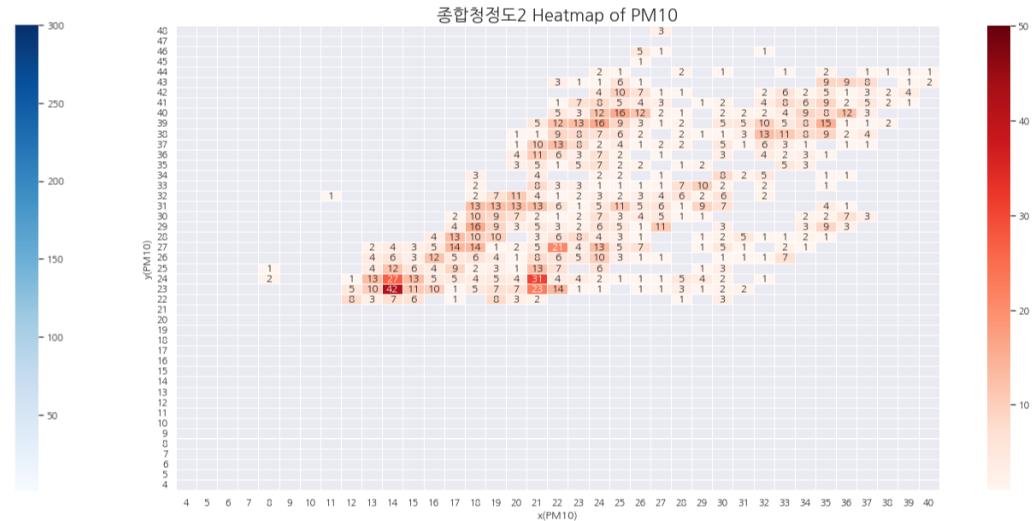
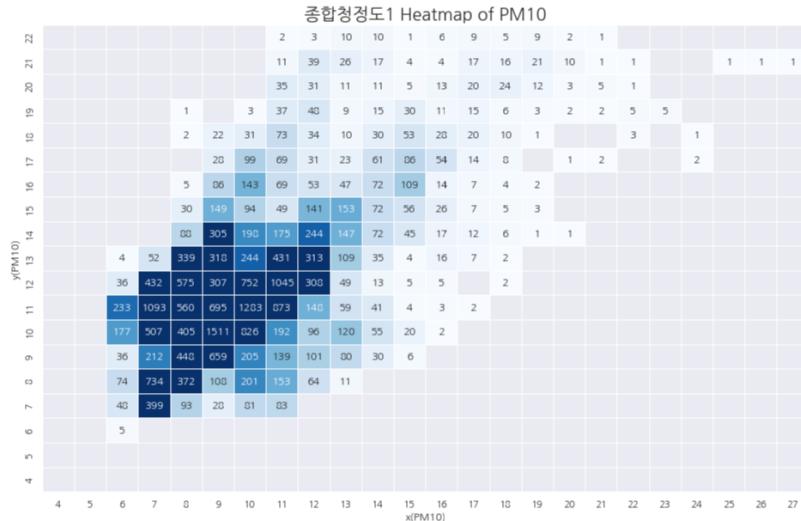


	Cost Function	Training Error	Test Error
종합 청정도1	$\frac{1}{n} \sum_{i=1}^n t_i - (0.7424x_i + 3.0603) $	1.32	0.76
종합 청정도2	$\frac{1}{n} \sum_{i=1}^n t_i - (0.5937x_i + 12.911) $	3.09	2.57
전체 데이터	$\frac{1}{n} \sum_{i=1}^n t_i - (1.1472x_i + 0.5673) $	1.69	1.14

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(5/6)

- PM10 센서 값을 종합청정도 기준으로 구분하여 회귀식 도출 및 오차 계산



	Cost Function	Training Error	Test Error
종합 청정도1	$\frac{1}{n} \sum_{i=1}^n t_i - (0.7109x_i + 4.7077) $	1.89	1.10
종합 청정도2	$\frac{1}{n} \sum_{i=1}^n t_i - (0.5778x_i + 18.501) $	4.32	3.73
전체 데이터	$\frac{1}{n} \sum_{i=1}^n t_i - (1.1205x_i + 1.0165) $	2.34	1.83

프로젝트 내용 – 제품 On 시점 결정 (Regression)

(6/6)

- Regression을 이용하여 제품 off 상태에서 측정된 센서값을 보정하였으며, 추가적인 개선 계획을 수립

✓ 제품 On 시점 결정을 위한 회귀 모델 검증 결과

구분	Cost Function	Training Error	Test Error
PM 1.0	종합 청정도1 $\frac{1}{n} \sum_{i=1}^n t_i - (1.2919x_i - 1.009) $	0.37	0.18
	종합 청정도2 $\frac{1}{n} \sum_{i=1}^n t_i - (1.092x_i + 0.9132) $	0.73	0.67
	전체 데이터 $\frac{1}{n} \sum_{i=1}^n t_i - (1.2372x_i - 0.7566) $	0.43	0.21
PM 2.5	종합 청정도1 $\frac{1}{n} \sum_{i=1}^n t_i - (0.7424x_i + 3.0603) $	1.32	0.76
	종합 청정도2 $\frac{1}{n} \sum_{i=1}^n t_i - (0.5937x_i + 12.911) $	3.09	2.57
	전체 데이터 $\frac{1}{n} \sum_{i=1}^n t_i - (1.1472x_i + 0.5673) $	1.69	1.14
PM 10	종합 청정도1 $\frac{1}{n} \sum_{i=1}^n t_i - (0.7109x_i + 4.7077) $	1.89	1.10
	종합 청정도2 $\frac{1}{n} \sum_{i=1}^n t_i - (0.5778x_i + 18.501) $	4.32	3.73
	전체 데이터 $\frac{1}{n} \sum_{i=1}^n t_i - (1.1205x_i + 1.0165) $	2.34	1.83

✓ 추가 연구 계획

1) 데이터 추가 수집

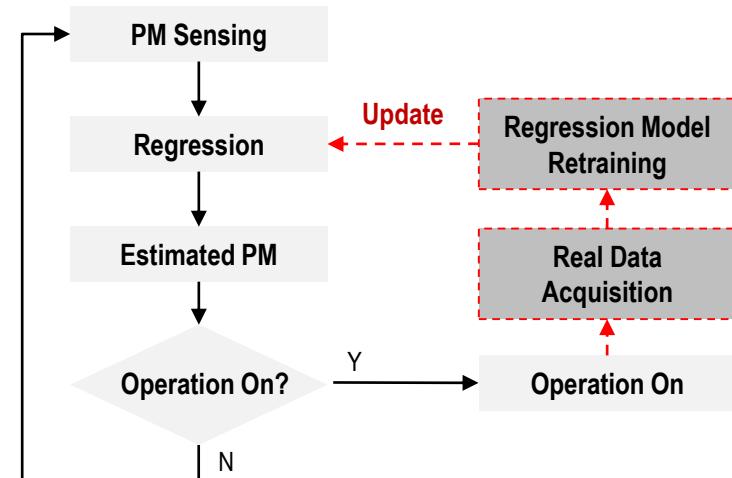
- 다양한 환경의 데이터 추가 확보하여 회귀모델 업데이트
종합청정도별 데이터 양 동등한 수준으로 확보 필요

2) 데이터 재분류

- 종합청정도 기준 분류가 아닌 각 센서별 값을 기준으로 구간을 나누어 회귀모델 만들고 검증, 비선형 회귀모델 적용 추가 검토

3) 현장별 설치 이후 추가 데이터 확보 및 Regression 모델 업데이트

- 고정 회귀식이 아니라 설치 이후 업데이트 되는 구조로 개선



프로젝트 내용 – 제품 Off 시점 결정 (LSTM)

(1/6)

- (t-n) ~ (t) 시점의 먼지센서 값을 Single LSTM에 입력하고, (t+1) 시점의 운전 상태를 학습하여, 예측하는 모델을 설계

데이터 수집

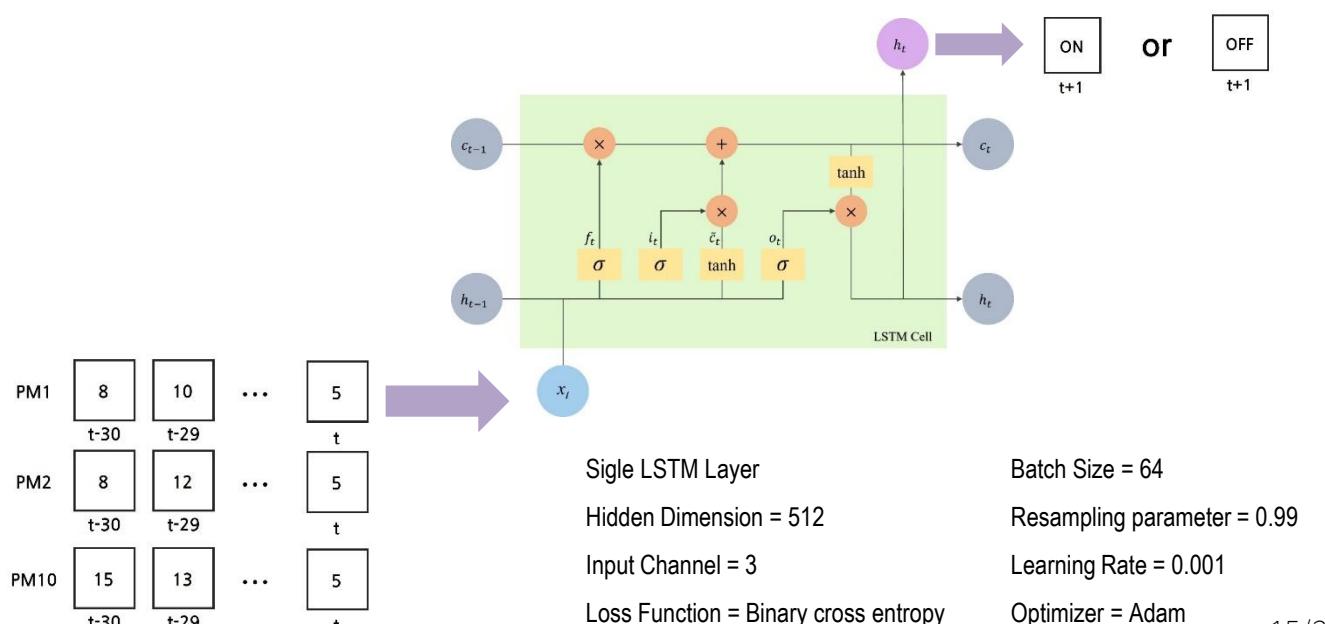
- 수집 기간: 2020년 10월 1일 ~ 10일 (10일)
- 수집 장소: ThinQ2.0서버 기기데이터, 내수 공기청정기 전체 (차량용 미니 공청기 제외)

데이터 분류

- 수집된 전체 데이터를 8:2비율로 Train Set과 Test Set으로 분류
- 하루 사용 시간 기준으로 데이터를 분류

	수집 데이터 수	사용 데이터 수	Train Set	Test Set
하루 0~12시간 사용자	38,307,314	1,000,000	5,792	1,448
하루 0~6시간 사용자	20,689,374	1,000,000	5,805	1,452
하루 6~12시간 사용자	17,617,940	1,000,000	6,060	1,515

device_id_deidentification	create_dt	PM1	PM2	PM10	airState.operation
0	2021-08-01 00:01:00	8	8	15	1
0	2021-08-01 00:06:00	10	12	13	1
0	2021-08-01 00:11:00	8	8	10	1
0	2021-08-01 00:16:00	8	8	9	1
0	2021-08-01 00:21:00	8	8	8	1
0	2021-08-01 00:26:00	8	8	9	1
0	2021-08-01 00:31:00	8	8	10	1
0	2021-08-01 00:36:00	8	8	9	1
0	2021-08-01 00:41:00	8	8	8	1
0	2021-08-01 00:46:00	13	15	16	1
0	2021-08-01 00:51:00	8	8	10	1
0	2021-08-01 00:56:00	8	8	14	1
0	2021-08-01 01:01:00	8	8	8	1

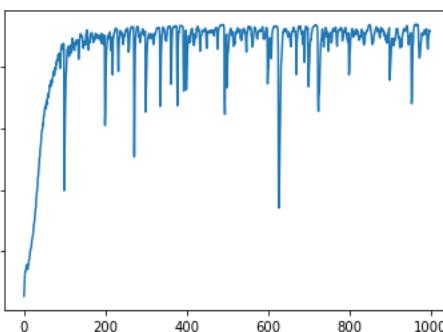
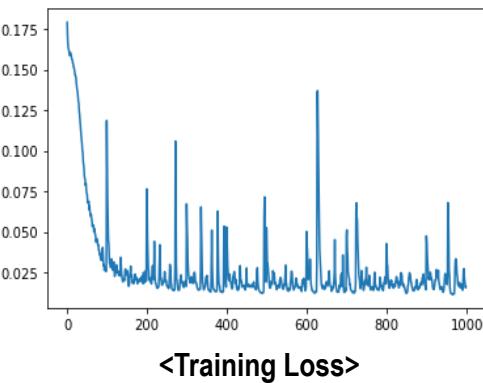


프로젝트 내용 – 제품 Off 시점 결정 (LSTM)

(2/6)

- 200개 제품의 데이터만으로 검증하였을 때는 우수한 성능을 보였으나, 빅데이터로 모델을 학습할 경우 제대로 학습되지 않는 결과 나옴

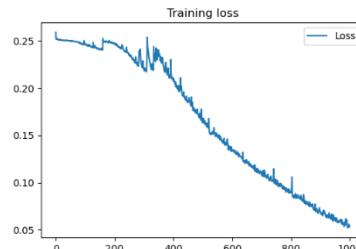
✓ 200 개 제품 데이터로 모델 학습 / 검증한 결과



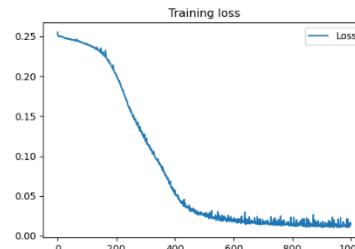
	Precision	Recall	F1-Score
On	0.98	0.95	0.96
Off	0.86	0.94	0.90
Accuracy	0.95		

✓ 빅데이터로 모델 학습 / 검증한 결과

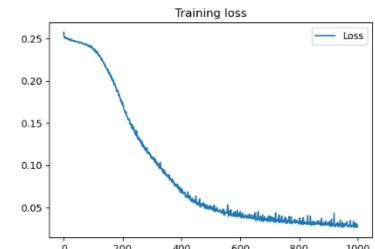
<6시간 미만 사용자>



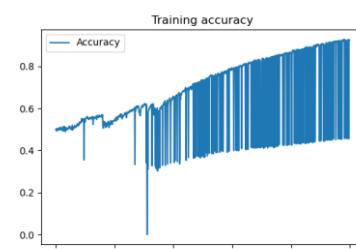
< 6~12시간 사용자>



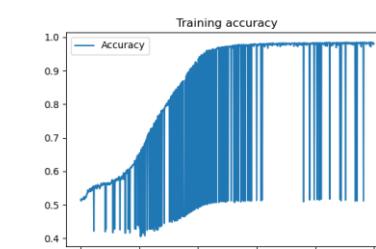
<12시간 미만 사용자 전체>



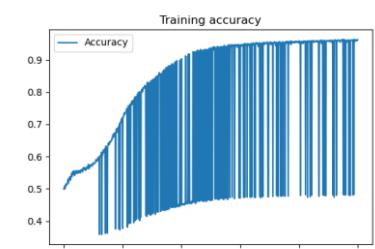
Training accuracy



Training accuracy



Training accuracy



	Precision	Recall	F1-Score		Precision	Recall	F1-Score		Precision	Recall	F1-Score
On	0	0	0	On	0	0	0	Off	0.51	1.00	0.68
Off	0.52	1.00	0.69	Off	0.51	1.00	0.68	Off	0.51	1.00	0.68
Accuracy	0.52			Accuracy	0.51			Accuracy	0.51		

➔ 학습은 된 것처럼 보이나, 일반화하기 어려운 결과로 확인됨

- 3 (PM1.0 / PM2.5 / PM10) x 30 (t ~ t-30) 형태로 먼지센서 값을 1D CNN에 입력하여, (t+1) 시점의 운전 상태를 예측하는 모델 설계

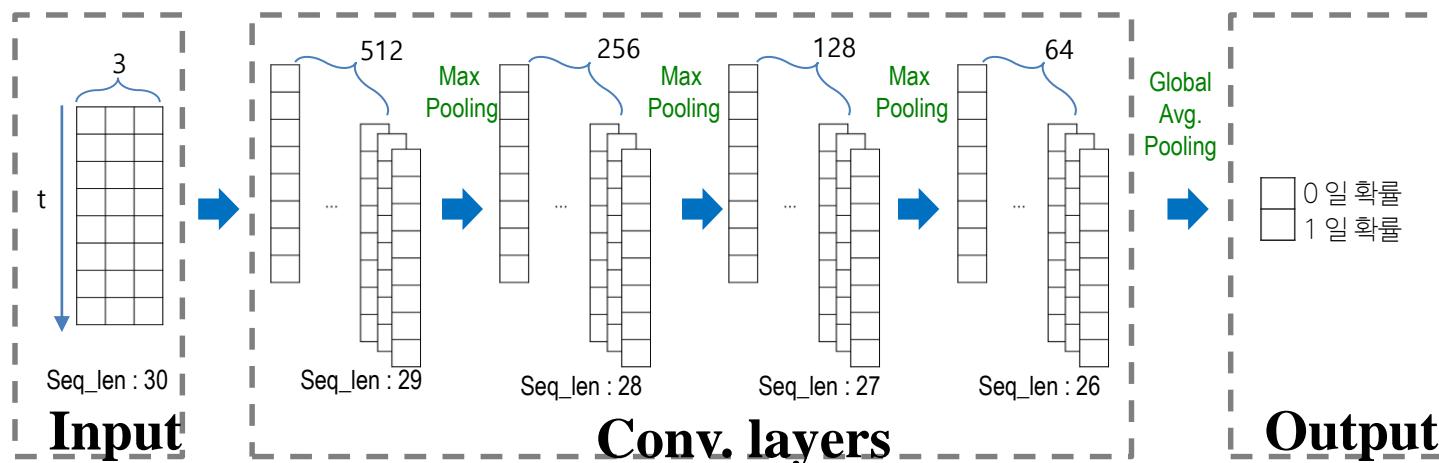
데이터 수집

- 수집 기간: 2020년 10월 1일 ~ 10일 (10일)
- 수집 장소: ThinQ2.0서버 기기데이터, 내수 공기청정기 전체 (차량용 미니 공청기 제외)

데이터 분류

- 수집된 전체 데이터를 8:2비율로 Training Set과 Test Set으로 분류
- 하루 사용 시간 기준으로 데이터를 분류

	수집 데이터 수	사용 데이터 수	Train Set	Test Set
하루 0~12시간 사용자	38,307,314	1,000,000	5,812	1,453
		2,000,000	13,572	3,393
하루 0~6시간 사용자	20,689,374	1,000,000	5,850	1,463
		2,000,000	11,960	2,990
하루 6~12시간 사용자	17,617,940	1,000,000	6,034	1,509
		2,000,000	12,220	3,055



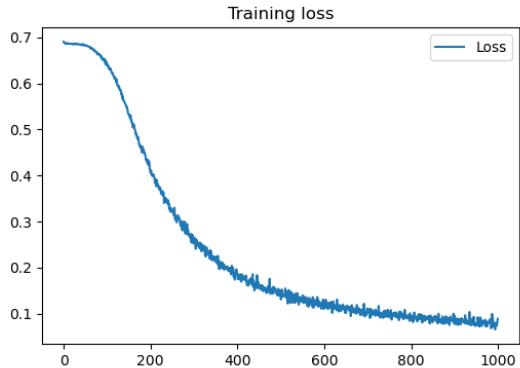
Sequential length: 30
 Loss function: Cross-Entropy Loss
 Batch size: 16
 Learning rate: 0.001
 Optimizer: Adam

프로젝트 내용 – 제품 Off 시점 결정 (CNN)

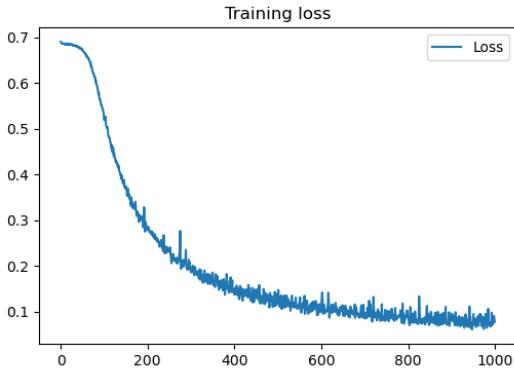
(4/6)

- 100만개 데이터를 학습/평가에 사용한 결과 먼지 센서 값으로 특정 시점 이후의 운전 상태를 예측하는 것은 어려운 것으로 확인됨

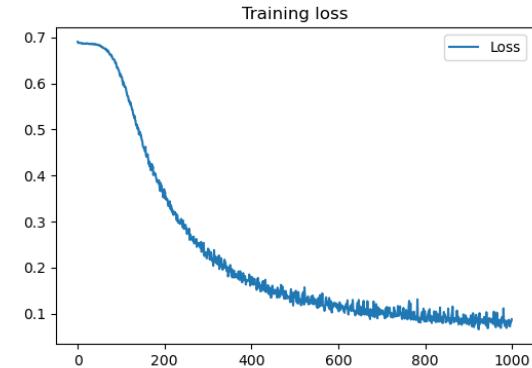
<6시간 미만 사용자>



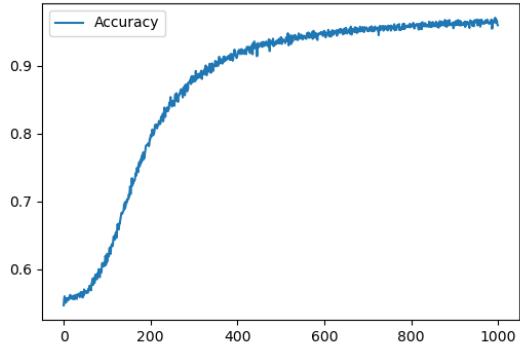
<6~12시간 사용자>



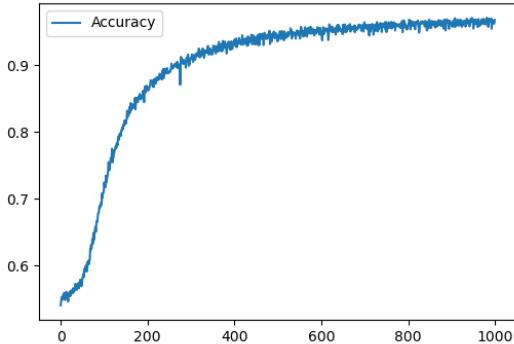
<12시간 미만 사용자 전체>



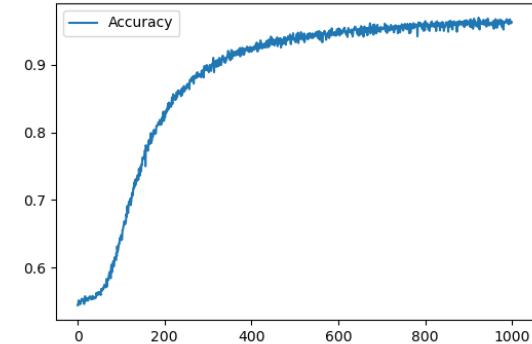
Training accuracy



Training accuracy



Training accuracy



Precision	Recall	F1-Score
-----------	--------	----------

On	0.52	0.51	0.51
Off	0.52	0.52	0.52
Accuracy	0.52		

Precision	Recall	F1-Score
-----------	--------	----------

On	0.52	0.39	0.45
Off	0.52	0.64	0.58
Accuracy	0.52		

Precision	Recall	F1-Score
-----------	--------	----------

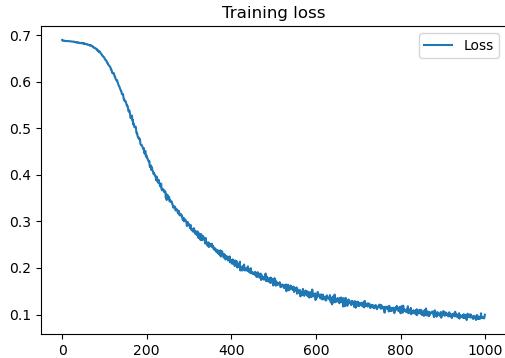
On	0.51	0.43	0.47
Off	0.51	0.59	0.55
Accuracy	0.51		

프로젝트 내용 – 제품 Off 시점 결정 (CNN)

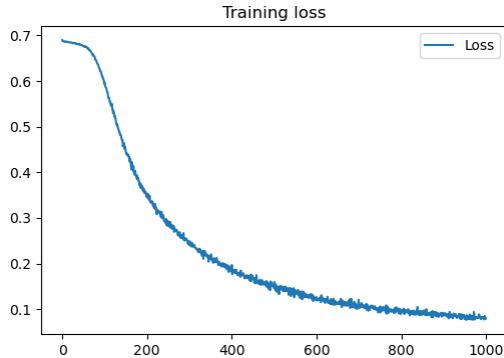
(5/6)

- 200만개 데이터를 학습/평가에 사용한 결과도 100만개 데이터 사용한 결과와 유사함 → 데이터 셋을 세분화하여 별도 학습해야 함

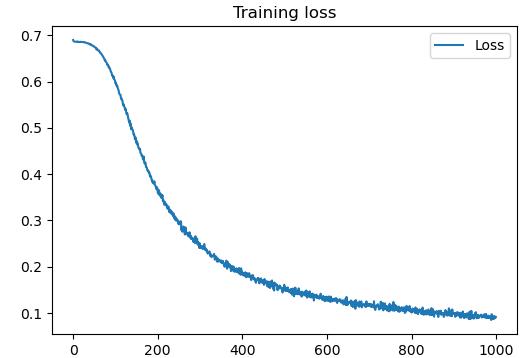
<6시간 미만 사용자>



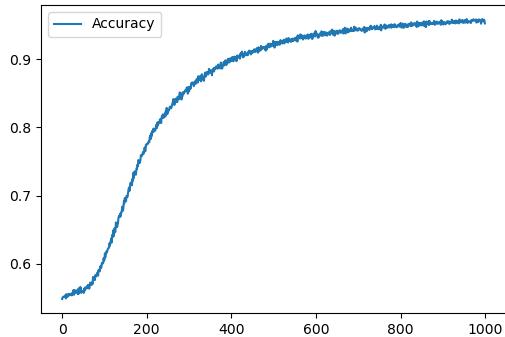
<6~12시간 사용자>



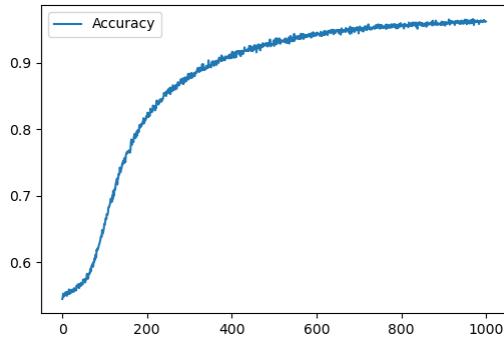
<12시간 미만 사용자 전체>



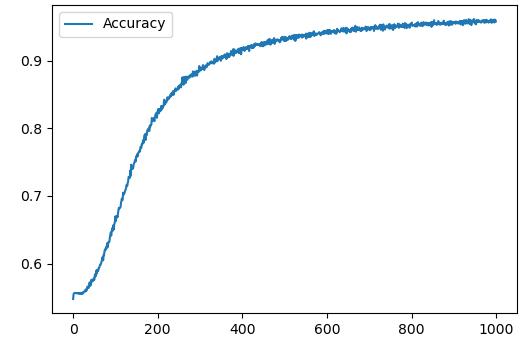
Training accuracy



Training accuracy



Training accuracy



	Precision	Recall	F1-Score
--	-----------	--------	----------

	Precision	Recall	F1-Score
On	0.51	0.48	0.49
Off	0.49	0.53	0.51

	Precision	Recall	F1-Score
--	-----------	--------	----------

	Precision	Recall	F1-Score
Off	0.55	0.44	0.49
Accuracy	0.51		

	Precision	Recall	F1-Score
--	-----------	--------	----------

	Precision	Recall	F1-Score
Off	0.45	0.45	0.45
Accuracy	0.52		

- 사용시점 / 운전시간 등의 사용자 패턴 정보를 추가하고, 각각의 AI 모델 **parameter**를 최적화 할 예정

✓ **제품 Off 시점 결정을 위한 AI 모델 검증 결과**

구분		Precision	Recall	F1-Score
LSTM	12시간 미만 사용자 전체	0.51	1.00	0.68
	6시간 미만 사용자	0.52	1.00	0.69
	6~12시간 사용자	0.51	1.00	0.68
CNN (100 만 개)	12시간 미만 사용자 전체	0.51	0.59	0.55
	6시간 미만 사용자	0.52	0.52	0.52
	6~12시간 사용자	0.52	0.64	0.58
CNN (200 만 개)	12시간 미만 사용자 전체	0.45	0.45	0.45
	6시간 미만 사용자	0.49	0.53	0.51
	6~12시간 사용자	0.55	0.44	0.49

✓ **추가 연구 계획**

1) 데이터 재분류

- 사용시점, 운전시간 기준으로 데이터셋을 재 분류
Ex. 오전시간에 2시간 사용한 사용자의 해당 시간 데이터만 별도로 모아서 학습

2) 모델 **Parameter** 최적화

- Seq. Length를 사용시간 그룹별로 다르게 설정
사용시간이 짧은 사용자에게는 Seq. Length가 짧아야 함

3) 제품 로직으로 AI모델 적용 대상 정의

- 전체 사용자를 대상으로 AI모델을 적용하지 않고, 특징이 명확히
분류되는 고객들만 AI모델을 이용하여 off 시점 결정
설정하는 그룹에 포함되지 않는 고객은 고정 off 조건 설정

1 공기청정기 사용 패턴 분석 및 **Unmet Needs** 파악을 통한 기능 설계

- 스마트 제어 기능에 대한 긍정적인 관심과 지불가치가 확인 되었으며, 공기청정기는 자동운전에 대한 **Needs**가 큰 것을 데이터로 확인
- 현재 제품이 가진 한계를 극복할 수 있는 방법을 **Data Driven**으로 수립

선행 연구 조사 / 기기데이터 분석 / 온라인 VOC파악 등 다양한 데이터를 기반으로 문제 정의

2 운전On 시점 결정을 위한 제품 Off 상태에서의 센서 값 보상을 위한 **Regression** 모델 설계 및 검증

- 기존 제품이 가진 한계점인 제품 off 상태에서의 정확한 먼지량 감지가 안되는 점을 개선 하기위한 모델 설계
- 특정조건(사무공간)에서 획득한 데이터를 기반으로 회귀모델을 만들어 모델의 활용 가능성을 확인

데이터 보강 및 먼지 구간별 회귀모델을 별도로 만들고, 제품 설치 후 참값을 수집하여 모델을 지속 업데이트하는 것으로 개발 방향 수립

3 LSTM / CNN 모델을 이용하여 먼지 센서 값을 학습하여 운전 off 시점을 예측하는 모델 설계 및 검증

- 시계열 데이터로 수집되는 먼지센서 값만으로 학습해서 off 시점을 예측하는 것은 어려움을 확인
먼지상태보다 생활패턴에 따라 제품을 On/Off 함을 데이터 분석 모델 설계 및 검증으로 확인
 - CNN 모델이 LSTM 대비 사용자를 그룹화하여 각각 학습할 경우 개선할 여지가 보이나, 적합한 데이터셋의 재정의 후 평가되어야함
사용시점 / 운전시간 정보를 추가하여 데이터를 분류하여 각각의 그룹별로 별도 학습하고, 로직으로 제약 조건을 설정하여 제품화 개발 예정
- ❖ 이번 PJT를 통해 도출한 접근 방법과 추가 연구를 통해 **23년향 자사 공기청정기 제품에 능동 공기질 관리 모드로 적용!!!!**

Lesson Learned

김*정

프로젝트를 진행하면서 제품의 철저한 이해가 바탕이 되어야 **데이터**의 가치 있는 분석과 활용이 가능하다는 것을 느꼈습니다. 기존에는 엔지니어의 입장에서 제품개발이 이루어졌지만, 앞으로는 고객 **데이터**기반으로 정말 필요한 기능을 개발하는 것이 중요하다 것을 배울 수 있었고, 여러 **데이터** 모델링/분석법이 있지만, 얻고자하는 **Output**에 맞는 **데이터**와 방법을 선택해야 좋은 결과를 얻을 수 있음을 배웠습니다.

김*엽

기존에 AI 관련 개발 시 지도학습 위주의 라벨링 된 **데이터**를 학습시키고 결과 값에 맞춰 제품을 제어하는 로직을 개발했다면, **빅데이터** 분석을 접목하여 사용자 **데이터**를 디바이스와 연결하여 새로운 Pain point를 찾고 개발해 나갈수 있다는 점을 배웠습니다. 특히 **데이터** 가공, **데이터** 처리 중요성에 대해 체감하였고, 지속해서 **데이터**와 연관하는 사고를 가져야 한다는 것을 교육을 통해 배웠습니다.

박*현

제조업에서 디지털 **데이터**를 바탕으로 산업 전반의 효율 향상을 이끌어 낼 수 있는 것을 처음 배우게 됐습니다. 그리고 똑같은 **데이터**라 할 지라도 학습 방법에 따라 학습 결과가 달라 질 수 있기 때문에 정확한 분석을 통해 용도에 맞는 방법을 찾는 것이 엔지니어의 할 일이라는 것을 배울 수 있었습니다.

한*우

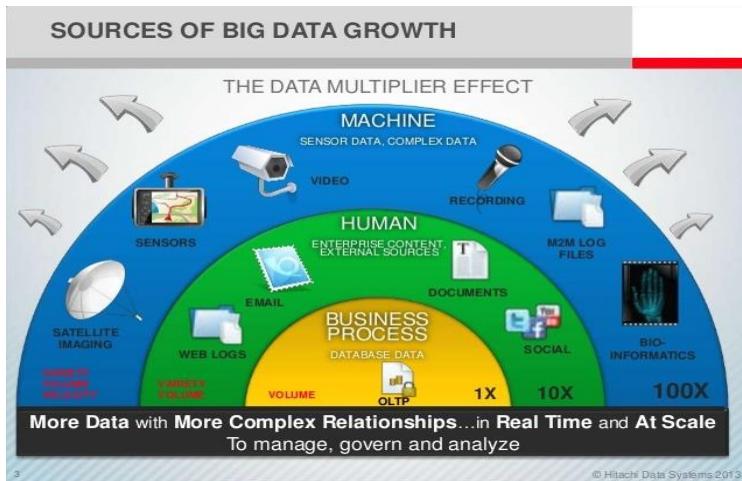
DX관점에서 기능이나 서비스를 개발함에 있어 기존과 달리 고객 **데이터**를 잘 이해하고 해석해야함을 확인하였습니다. 다양한 AI 기술을 적용하기 위해 문제를 잘 정의하고, 적합한 모델과 주요 파라미터 최적화가 필요하며, 무엇 보다 양질의 **데이터** 셋을 확보하는 것이 중요함을 깨달았습니다. **데이터!! 데이터!! 데이터!!**
전체관점에서 **데이터** 기반의 사고 (문제정의 / 해결방안 도출 / 실행 / 계획 수립)를 되짚어보고 역량을 올릴 수 있는 기회였습니다.



Process mining (operational big data)

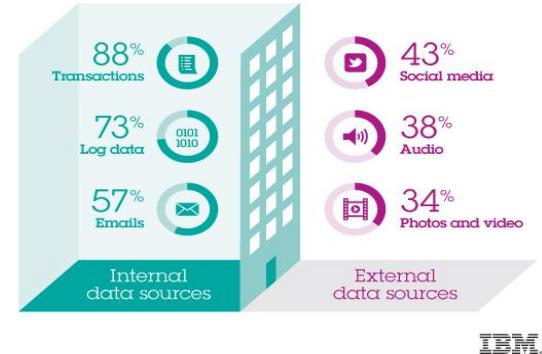
BI vs. BI

- Where do we have Big-data?

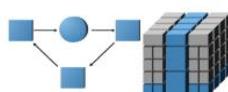


Where does big data come from?

Most big data efforts are currently focused on analyzing internal data to extract insights. Fewer organizations are looking at data outside their firewalls, such as social media.



Source: "Capitalize on Big data through Hitachi Innovation", 2013



Business Application Data

- Relational data, highly structured, based on inflexible schema
- Financial records, multidimensional data, math computation
- Monthly reporting, not for real-time events



Human-generated Data

- Generated by human-to-human interaction
- Includes email, IM, voice, video and text across
- Stored in centralized corporate servers, fileshares and desktops



Machine Data

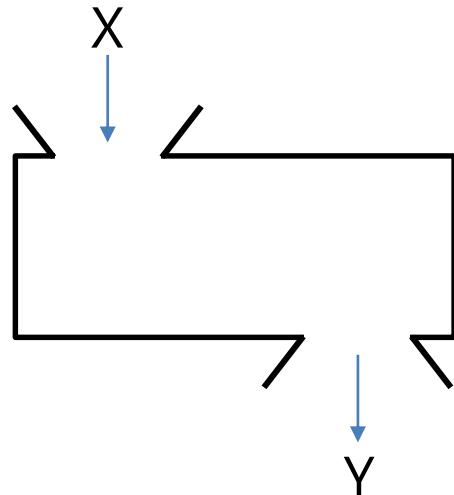
- Time series unstructured data, no predefined schema
- Generated by all IT systems, highly diverse formats
- Massive volume; fast navigation and correlation paramount

Machine Learning

- Finding 'f'

$$Y = f(X)$$

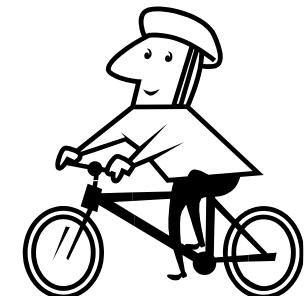
rule
pattern
knowledge



What is learning

- Learning
 - A process that allows an agent to adapt its performance through **instruction** or **experience**
 - Considered fundamental to intelligent behavior
 - May be
 - Simple association task
 - A specific output is required when given some input
 - Acquisition of a skill

changes in a system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively **next time**.



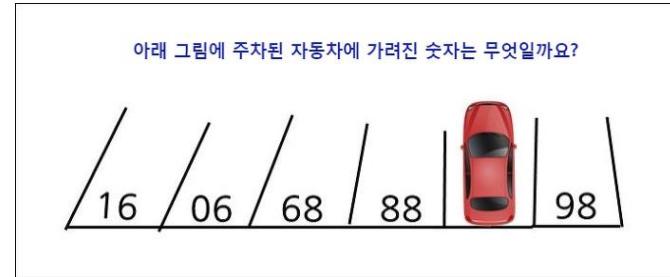
- Why?
 - Very active and large area of AI
 - Biological and cognitive perspective
 - Desire to understand more about ourselves
 - Get machines to perform tasks that serve us in some way

Learning methods

- Different data
- Different methods
- Different usages (purposes)

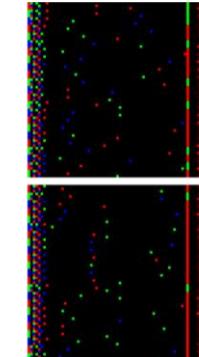
We need data

- 수치형, 범주형
 - (234, 0.327, ...) (토요일, 맑음, 배혜림)
- 연속형, 이산형
 - (0.234, 0.327, ...) (0, 1)
- 정형, 비정형
 - (Table, 벡터, 리스트), (이미지, 음성, 문서)
- 균형, 비균형
 - 양, 불량
- 기계는 모든 유형의 data를 받아 들일 수 있을까요?



Handling categorical variables
Mixed input/output

COL_ID	PAC_CD1	PAC_CD2	SUBST_PAC_CD	THK	WSDHT	WDHT	SDT	EDT	PROD_TIME
180042101111A	PP	PP2	PP2	5.99	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	RC	RC1	RC1	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	CR	CR2	CR2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111A	PP	PP2	PP2	5.99	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	CR	CR2	CR2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111A	HS	HS3	HS3	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111B	PP	PP2	PP2	5.99	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111B	CR	CR2	CR2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111B	HS	HS2	HS2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	PP	PP2	PP2	5.99	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	HS	HS3	HS3	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111A	CR	CR2	CR2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111B	HS	HS3	HS3	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111B	RC	RC1	RC1	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111B	HS	HS3	HS3	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111A	RC	RC1	RC1	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	HS	HS3	HS3	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111A	PP	PP2	PP2	5.99	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111B	CR	CR2	CR2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,400
180042101111B	HS	HS2	HS2	3.5	1.132	22.900	01/09/2018 16:16:00	01/09/2018 16:16:00	1,300
180042101111A	PP	PP1	PP1	1.41	1.068	14.900	01/09/2018 09:00:00	01/09/2018 09:00:00	1,800
180042101111A	SS	SS4	SS4	0.6	300	1.990	01/09/2018 08:00:00	01/09/2018 08:00:00	3,000
180042101111A	HS	HS2	HS2	0.6	300	1.790	01/09/2018 08:00:00	01/09/2018 08:00:00	2,100
180042101111A	RC	RC2	RC2	1.41	1.068	14.900	01/09/2018 09:00:00	01/09/2018 09:00:00	1,800
180042101111A	CR	CR2	CR2	1.41	1.068	14.900	01/09/2018 09:00:00	01/09/2018 09:00:00	1,800
180042101111A	CR	CR2	CR2	1.198	18.000	01/09/2018 09:00:00	01/09/2018 09:00:00	2,400	
180042101111A	CR	CR2	CR2	0.6	300	1.790	01/09/2018 08:00:00	01/09/2018 08:00:00	2,100
180042101111A	AN	AN1	AN1	1.5	1.198	15.000	01/09/2018 13:00:00	01/11/2018 13:00:00	208,600
180042101111A	RC	RC1	RC1	0.51	1.198	15.000	01/09/2018 13:00:00	01/09/2018 13:00:00	2,400
180042101111A	SP	SP2	SP2	0.51	1.198	15.000	01/09/2018 13:00:00	01/09/2018 13:00:00	2,100
180042101111A	HS	HS3	HS3	0.5	1.198	15.000	01/09/2018 13:00:00	01/09/2018 13:00:00	3,000
180042101111A	AN	AN1	AN1	0.51	1.198	15.000	01/09/2018 13:00:00	01/11/2018 13:00:00	207,204

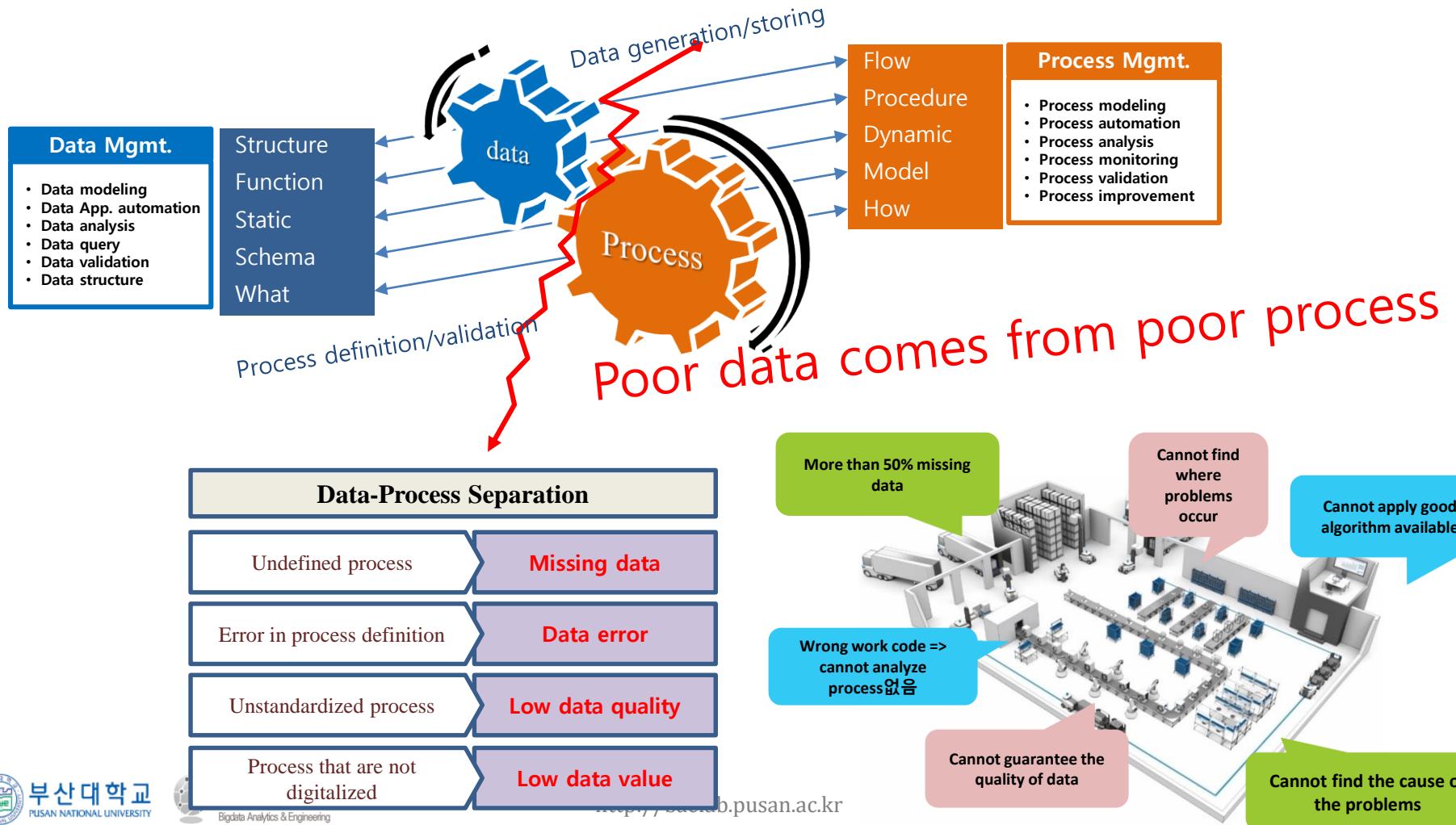


What methods do we need to use?

- Different methods for different data
- Multi-purposed model

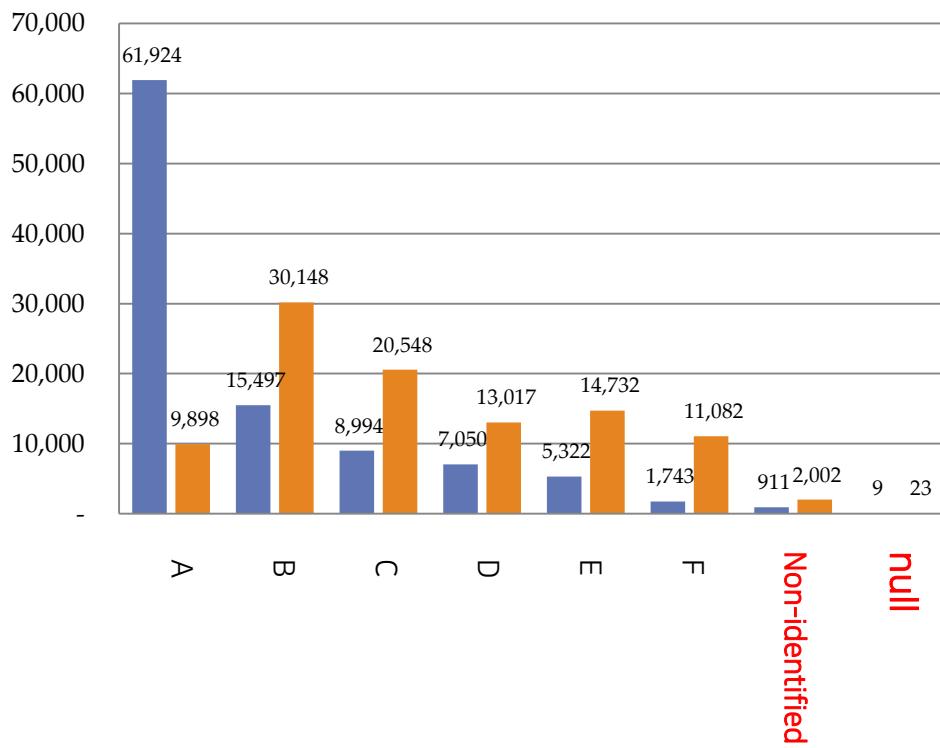
구분	설명을 위한 선형 회귀분석	예측을 위한 선형 회귀분석
목적	독립변수들과 종속변수들 간의 관계를 밝히기 위함	독립변수 값은 존재하나, 종속변수 값이 존재하지 않는 데이터의 종속변수 값을 예측하기 위함
사용 데이터 셋	모집단에서 가정된 관계에 대한 정보가 최대한 반영된 최적의 적합 모델을 추정하기 위해서 전체 데이터 세트를 사용	데이터는 일반적인 학습세트와 검증세트로 나뉘지며, 학습세트는 모델을 추정하는데, 검증세트는 새로운 데이터에 대한 모델의 성능을 평가하는 데 사용
평가	데이터가 모델에 얼마나 잘 적합 하는가	모델이 새로운 사례를 얼마나 잘 예측 하는가

Big-data vs. process improvement

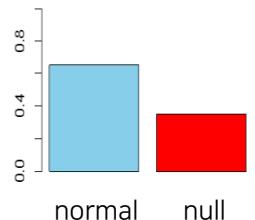


- Is technology the only issue for Industry 4.0?

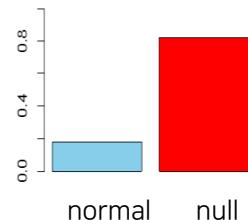
Frequency of activities



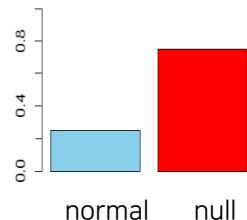
Work code missing



TP loading missing

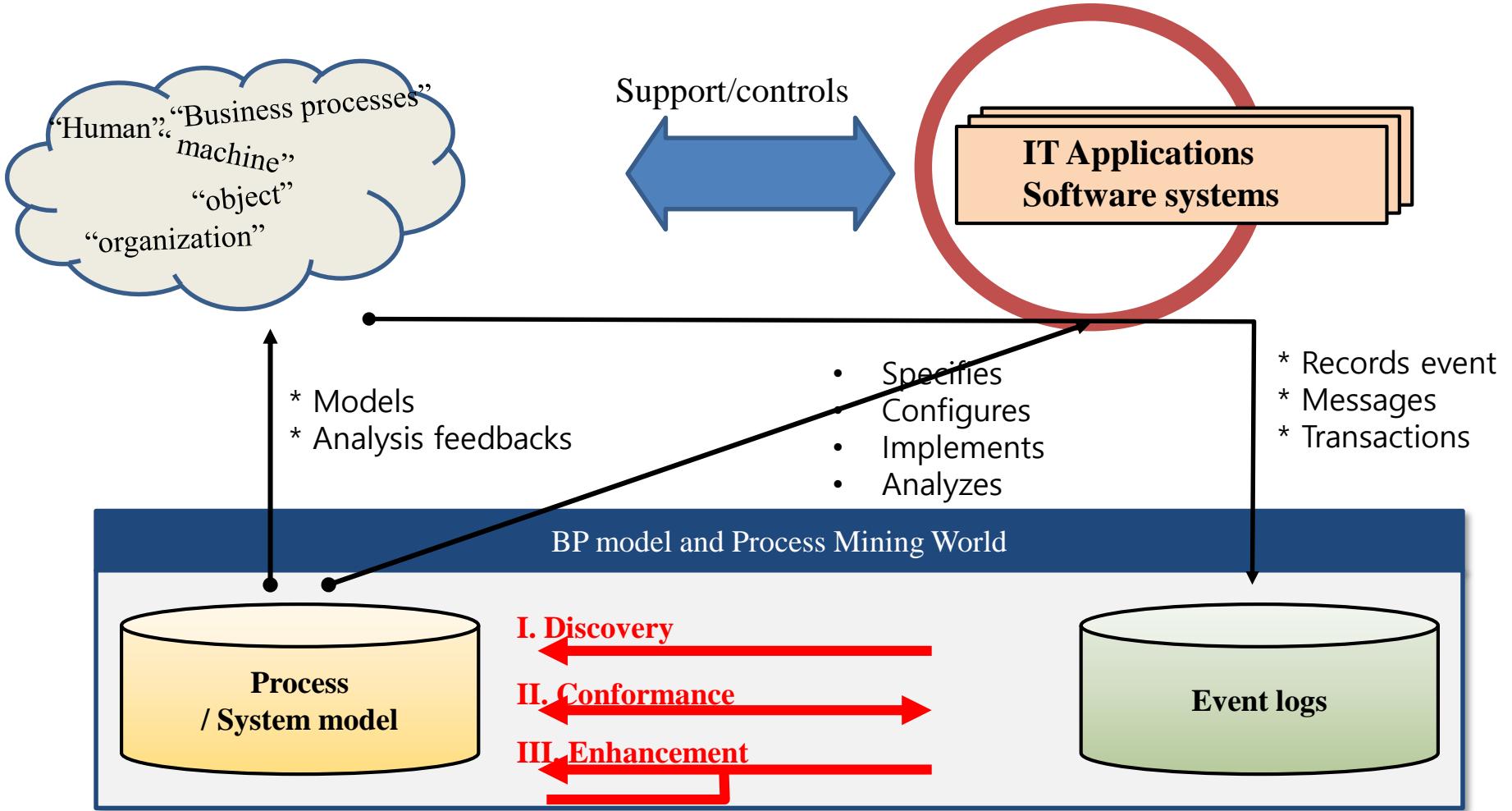


TP discharging missing



After introducing a mobile system

Process Analytics

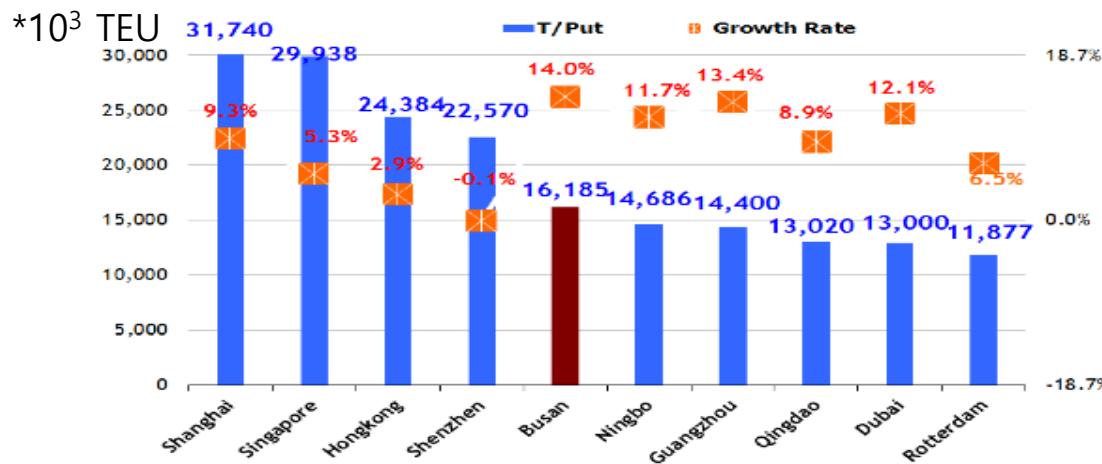


Process Analytics



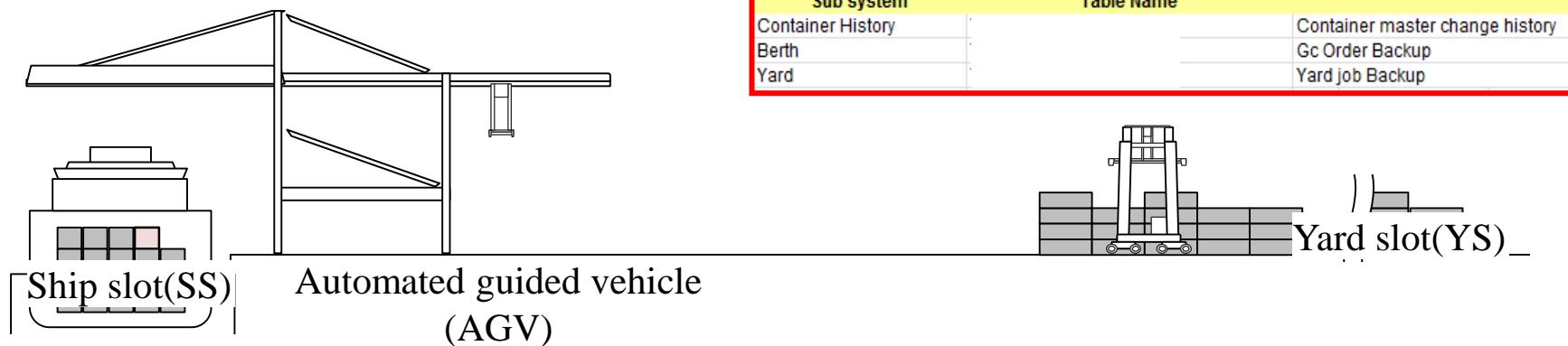
CASE I: Container handling process in a container port

- Busan Port

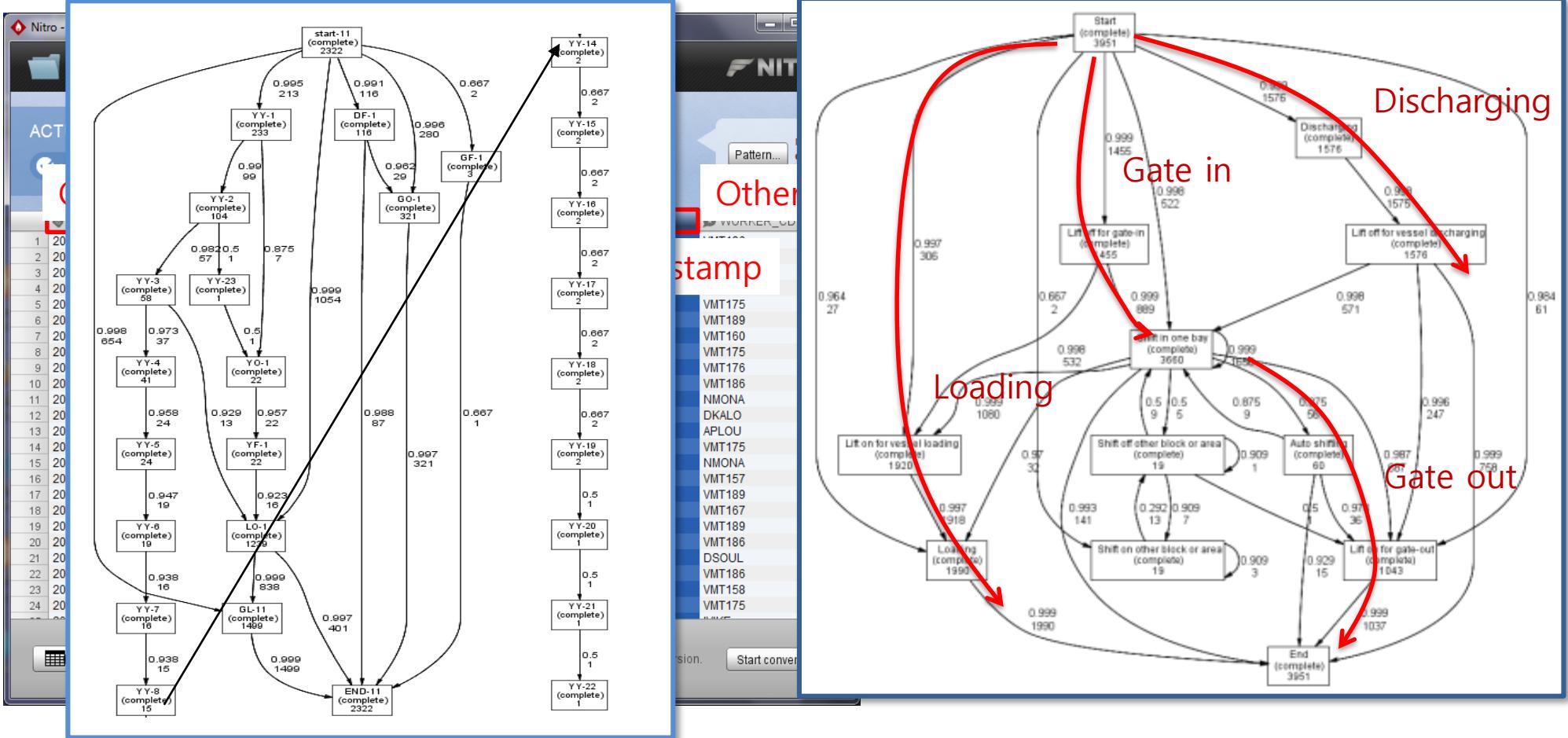


2013년 ranking		
ranking	Port	Container handling
1	Shanghai	3362
2	Singapore	3258
3	Shenzhen	2328
4	Hong Kong	2235
5	Busan	1769
6	Ningbo	1733
7	Qingdao	1552
8	Guangzhou	1531

10 largest container ports in the world (2011), BPA Korea 2012.06



Container handling process discovered



CASE II: Ship Building Process

- Korea is number 1 in ship building industry

Compensated Gross Tonnage, clarkson

Order received, Jan. 2014

Rank	Yard	Location
1	Hyundai H.I.	Ulsan
2	Daewoo SB	Okpo
3	Samsung H.I.	Koje
4	Dalian Shipbd. Ind.	Dalian
5	STX Shipbuild.	Jinhae
6	Sungdong S.B.	Tongyoung
7	Hyundai Samho	Samho



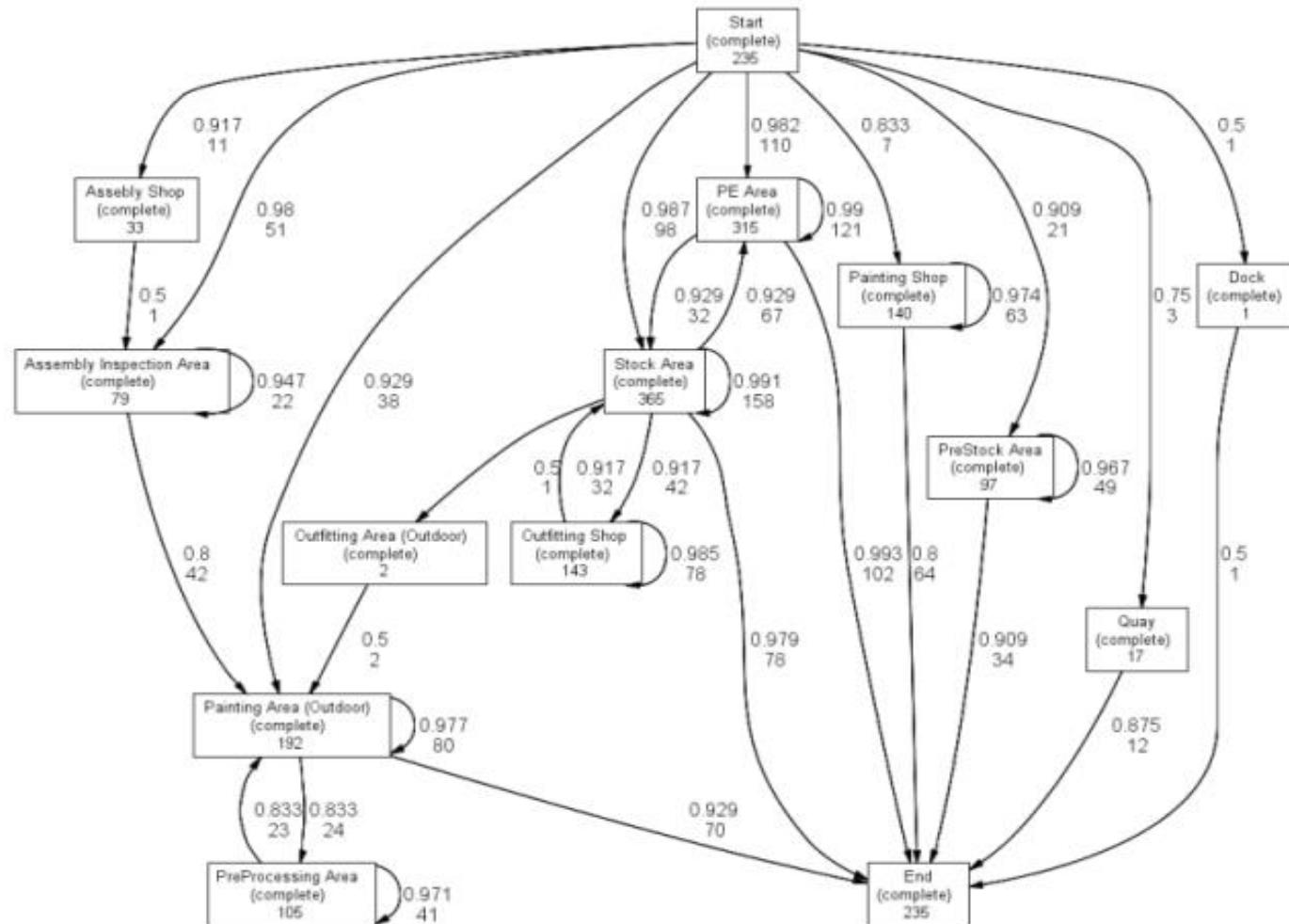
Design	Steel Stock	Cutting & Forming	Assembly	Pre-outfitting Painting	Pre Erection (PE Area)	Erection (Dock)	Quay
Block Division	Unloading Quay	Pretreatment N/C Cutting Forming Press	Component Sub Assembly Unit Assembly Grand Assembly	Pre outfitting Pre painting	G/C	G/C	Outfitting Painting Sea trial
Nesting Plan		Roll			K/L F/L L/C		



Block Assembly Operations

<http://baelab.pusan.ac.kr>

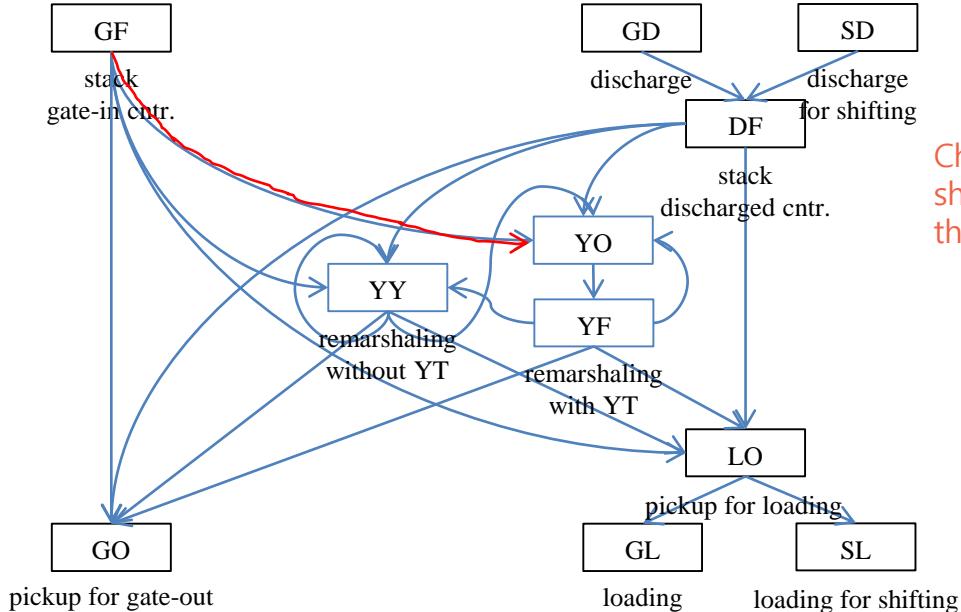
Process model discovered (Block movement)





1. For better understanding what we are doing

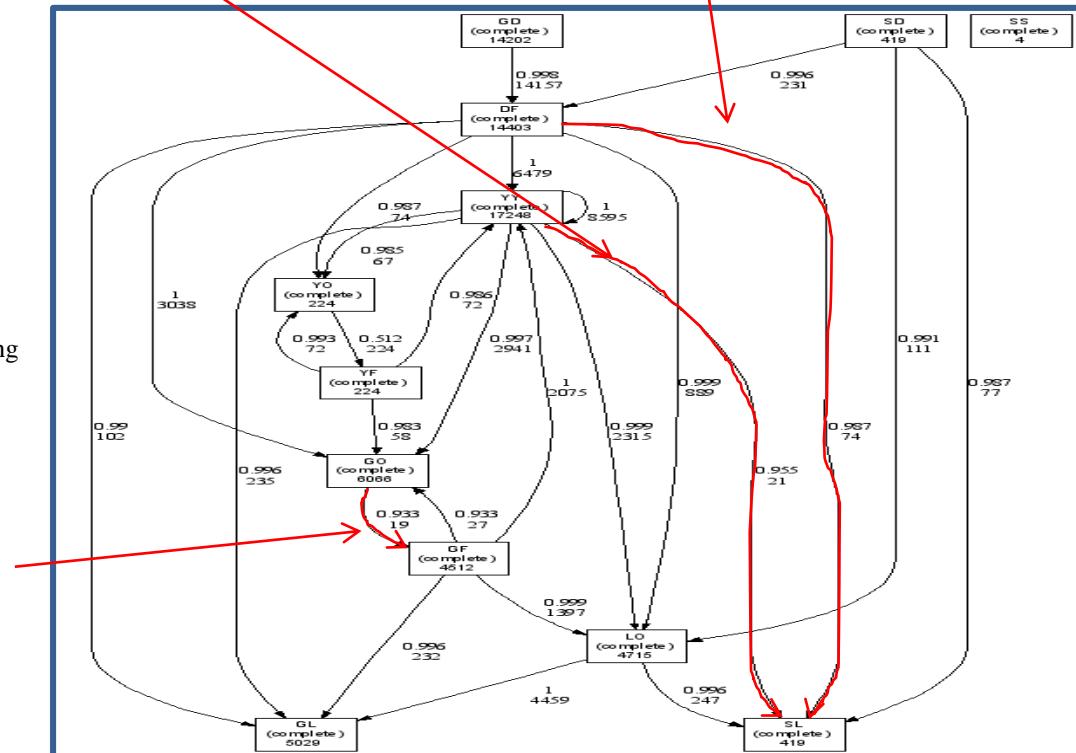
Pre-defined process model vs. Discovered process model



Picked up for Gate out but stacked in the yard again

Change position or loaded onto ship without being picked up in the yard

shifted by QC without being picked up by YC

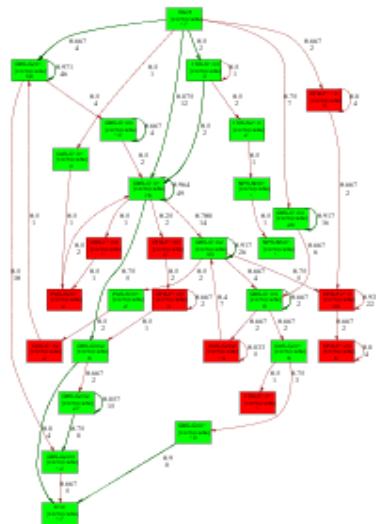


Comparison between two model

- Plan vs. Actual
- As-Is vs. To-Be
- Peer vs. Peer

ITEM	Plan	Actual	A - P
Earliest	2012-05-15 00:00:00	2012-05-11 21:11:00	- 3D 02:49:00
Latest	2012-09-18 00:00:00	2012-10-04 11:57:00	16D 11:57:00
Duration	05-07 09:00:00	05-26 23:46:00	19D 14:46:00
Instances	19	19	0
Events	459	442	-17
Tasks	21 (start, end 포함)	33 (start, end 포함)	12
Fitness	0.375	0.357	-0.018
Cross Fitness	0.118	0.167	0.049
Node Matched	0.95	0.606	-0.344
Arc Matched	0.477	0.288	-0.19

Plan



Actual



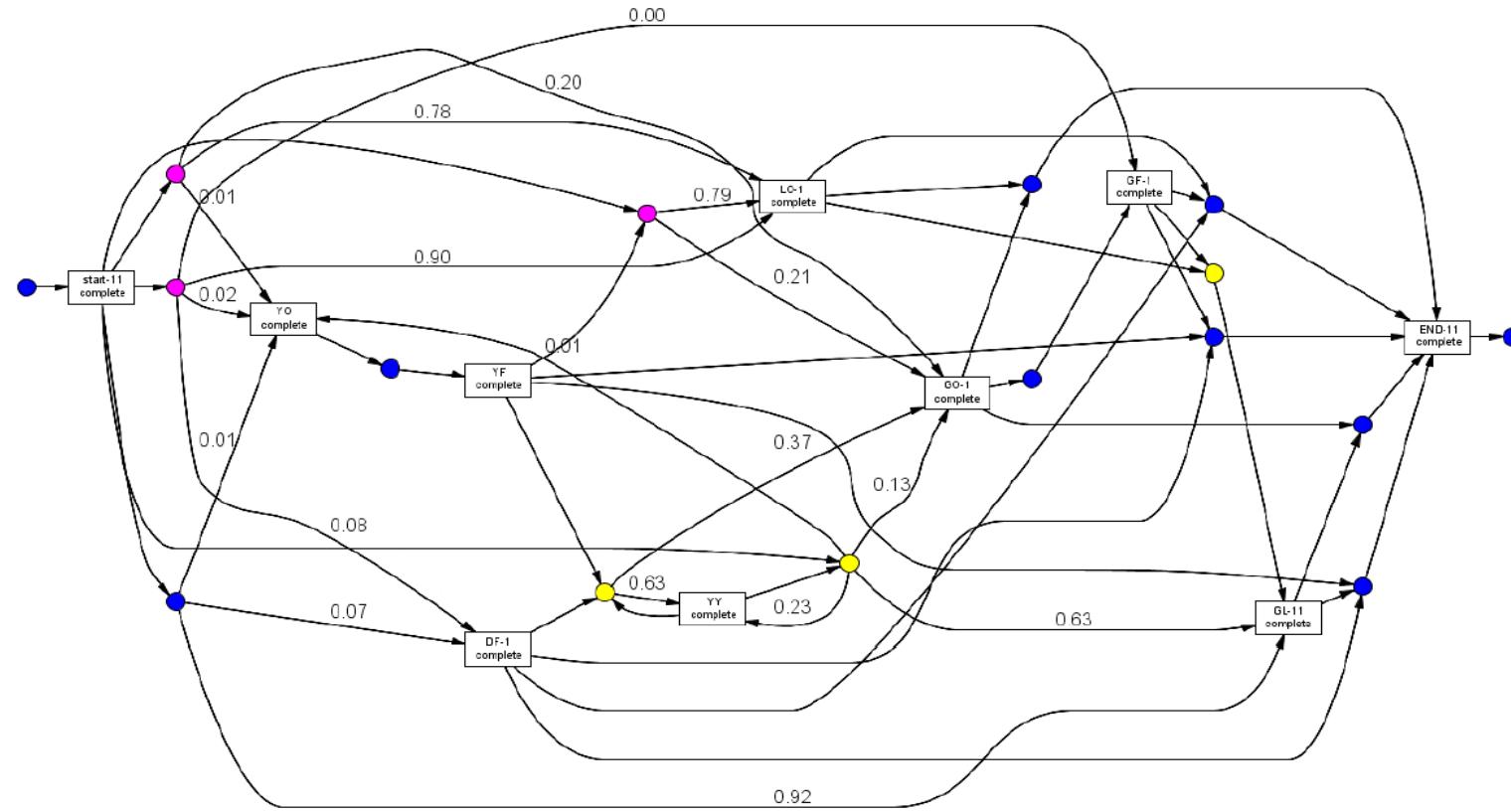
Port example: Better understanding of current situation





2. Finding cause and fixing the problem

What is the bottleneck in the port?



Good flow and Bad flow

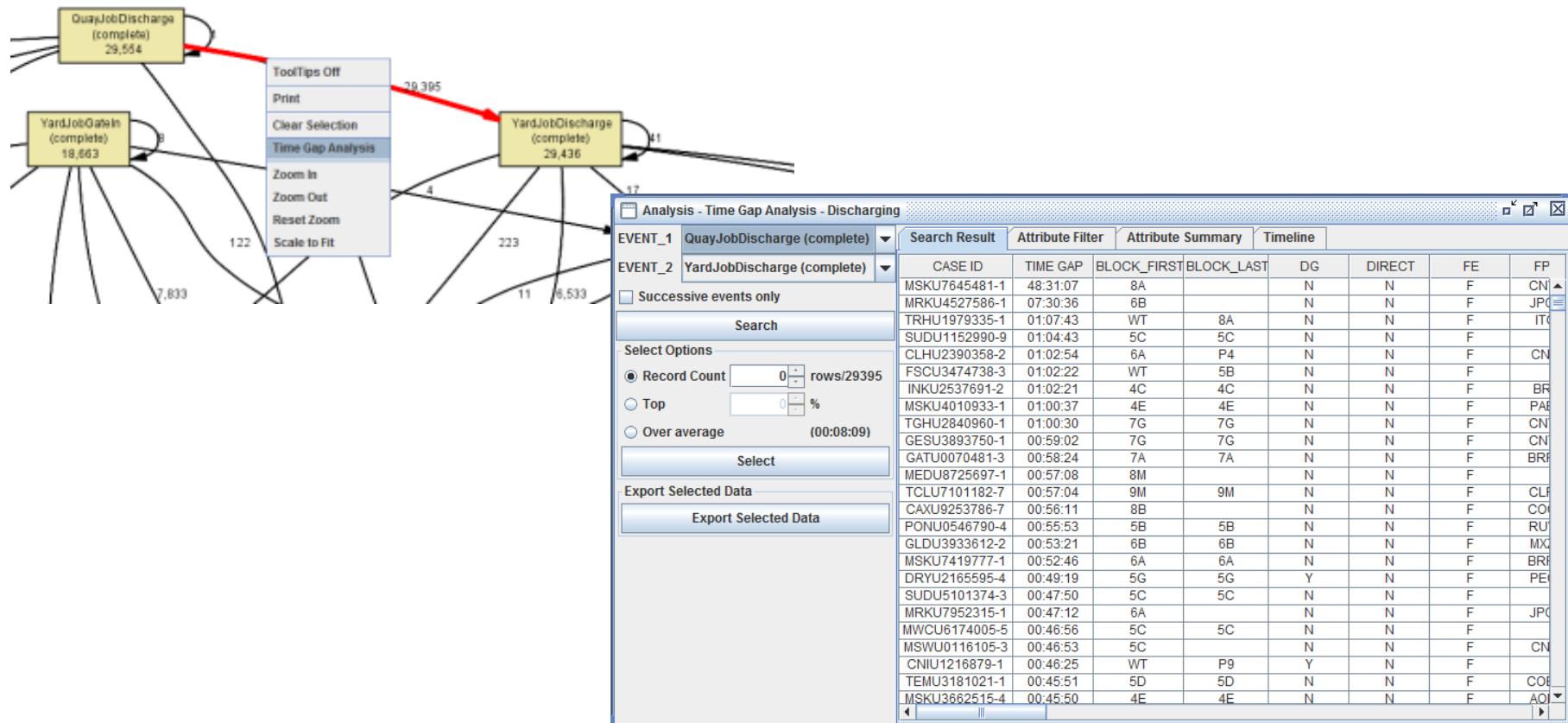
- Process Discovery
 - Good flow and Bad flow

	QC discharge	YC work discharge	YC work gate-in	YC work gate-out	YC work loading	QC loading	Truck Loading	Truck discharging	Refer Plug-in	Refer Plug-out
QC discharge	1	29395	0	12	17	50	26	38	0	0
YC work discharge	0	41	0	6495	6553	0	765	223	781	17
YC work gate-in	0	0	8	1058	7833	1	604	122	385	4
YC work gate-out	0	0	0	4	0	0	0	3	0	18
YC work loading	0	0	0	0	49	30396	0	0	0	0
QC loading	0	0	0	0	7	0	7	10	0	0
Truck Loading	0	0	0	310	301	0	24	3391	19	0
Truck discharging	0	0	0	775	569	0	687	312	27	2
Refer Plug-in	0	0	0	1	0	0	0	0	2	1751
Refer Plug-out	0	0	0	1002	311	0	37	11	612	38

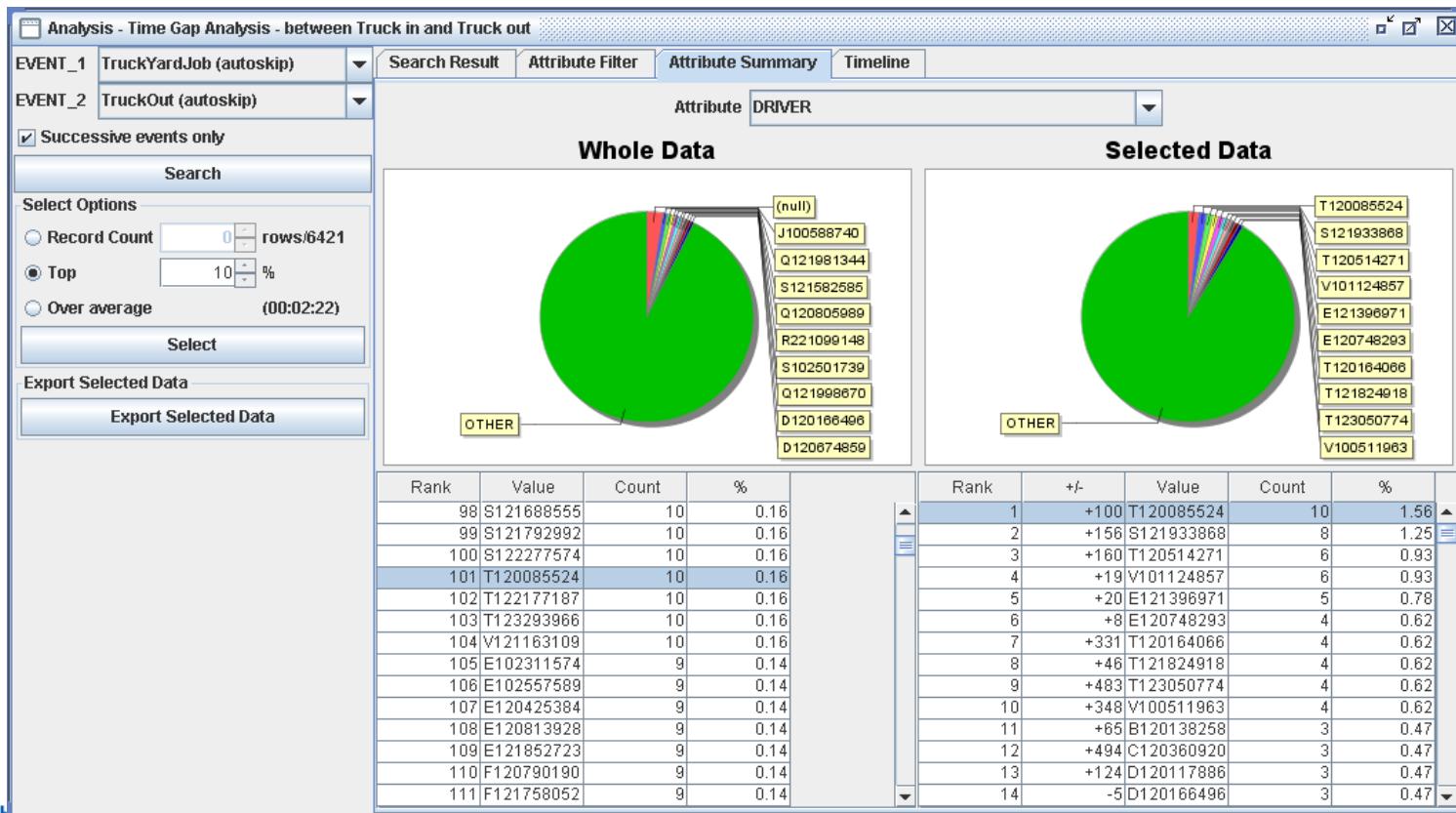
Good flow Irregular flow Bad flow

Single dimensional time gap analysis

- Time Gap between two arbitrary nodes
 - Shows time gap of all cases in a decreasing order



- Time Gap Analysis → Timeline
 - We can know when delayed cases occur in the time line





3. For predicting future result

Bayesian Network

- Bayesian Network (BN)
 - Bayesian network is a useful tool for inference and sensitivity analysis
 - Generating the structure is not an easy task (Chickering et al, 2004)
 - Inference **without** Bayesian network
 - To make one inference, **scan event logs every time**
 - **If we have Bayesian network**

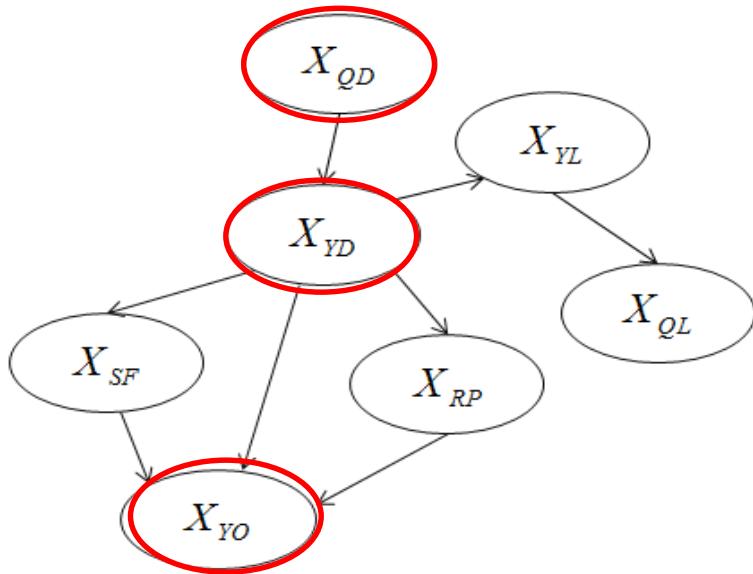
To make one inference, scan event logs every time

```
</AuditTrailEntry>
</ProcessInstance>
<ProcessInstance ID="TCIUS021880-2">
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH112</Attribute>
  </Data>
  <WorkflowModelElement>Start</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T09:59:11.000+08:00</Timestamp>
  <Originator>XH112</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH112</Attribute>
  </Data>
  <WorkflowModelElement>Dispatch</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T10:00:01.000+08:00</Timestamp>
  <Originator>XH112</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH210</Attribute>
  </Data>
  <WorkflowModelElement>Completed</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T10:00:01.000+08:00</Timestamp>
  <Originator>XH210</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH210</Attribute>
  </Data>
  <WorkflowModelElement>DispatchGateOut</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T10:33:11.000+08:00</Timestamp>
  <Originator>XH210</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH210</Attribute>
  </Data>
  <WorkflowModelElement>Complete</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T10:41:05.000+08:00</Timestamp>
  <Originator>XH210</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
  <Data>
    <Attribute name="UC_ID">XH210</Attribute>
  </Data>
  <WorkflowModelElement>End</complete></WorkflowModelElement>
  <Eventtype>complete</Eventtype>
  <Timestamp>2013-08-08T10:41:06.000+08:00</Timestamp>
  <Originator>XH210</Originator>
</AuditTrailEntry>
</AuditTrail>
</ProcessInstance>
</ProcessInstances>
```

A vertical blue downward-pointing arrow, indicating that the following text continues on the next page.

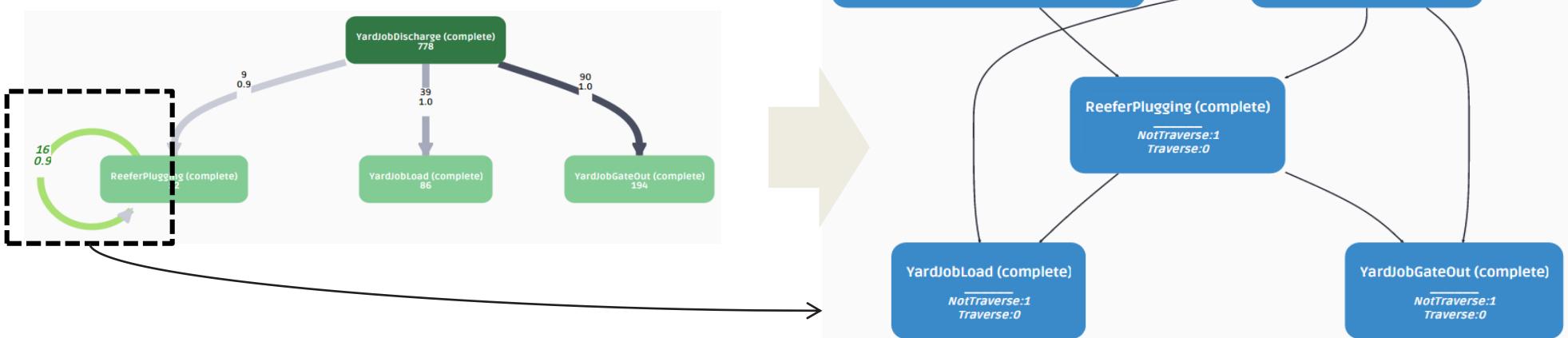
- If we have Bayesian network?

Make Bayesian network **once**, and using
node traversal every time make inference



BN and Process model

- Methodology of generating Bayesian network
 - Decomposition of dependency graph into directed acyclic graph (Sutrisnowati et al., 2012)
 - Learned Bayesian network using dynamic programming with mutual information test (MIT) score (Sutrisnowati et al., 2013)
 - Learned Bayesian network using genetic algorithm (Sutrisnowati et al., 2013)
- Arc in process model discovered by process mining technique
 - It contains causal dependency between nodes
- Using Bayesian Networks (BN):
 - (1) Inference (Causal inference, Prediction)
 - (2) Sensitivity analysis (What if simulation)

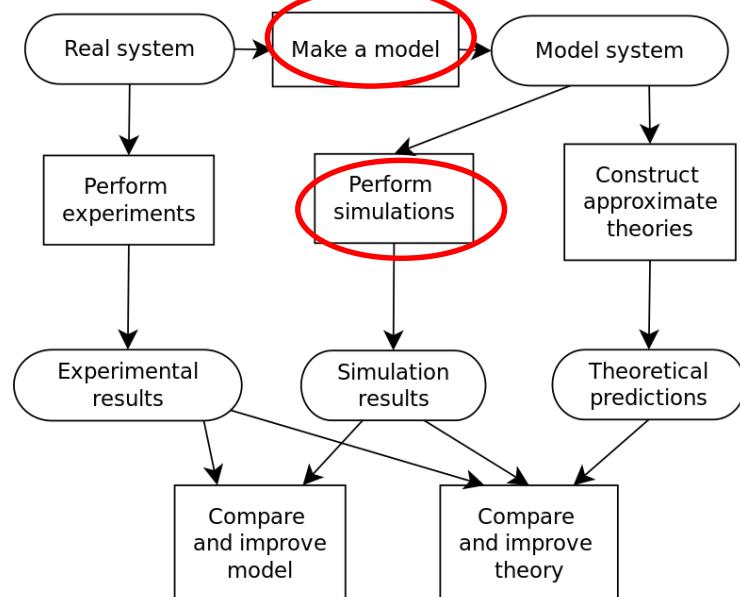




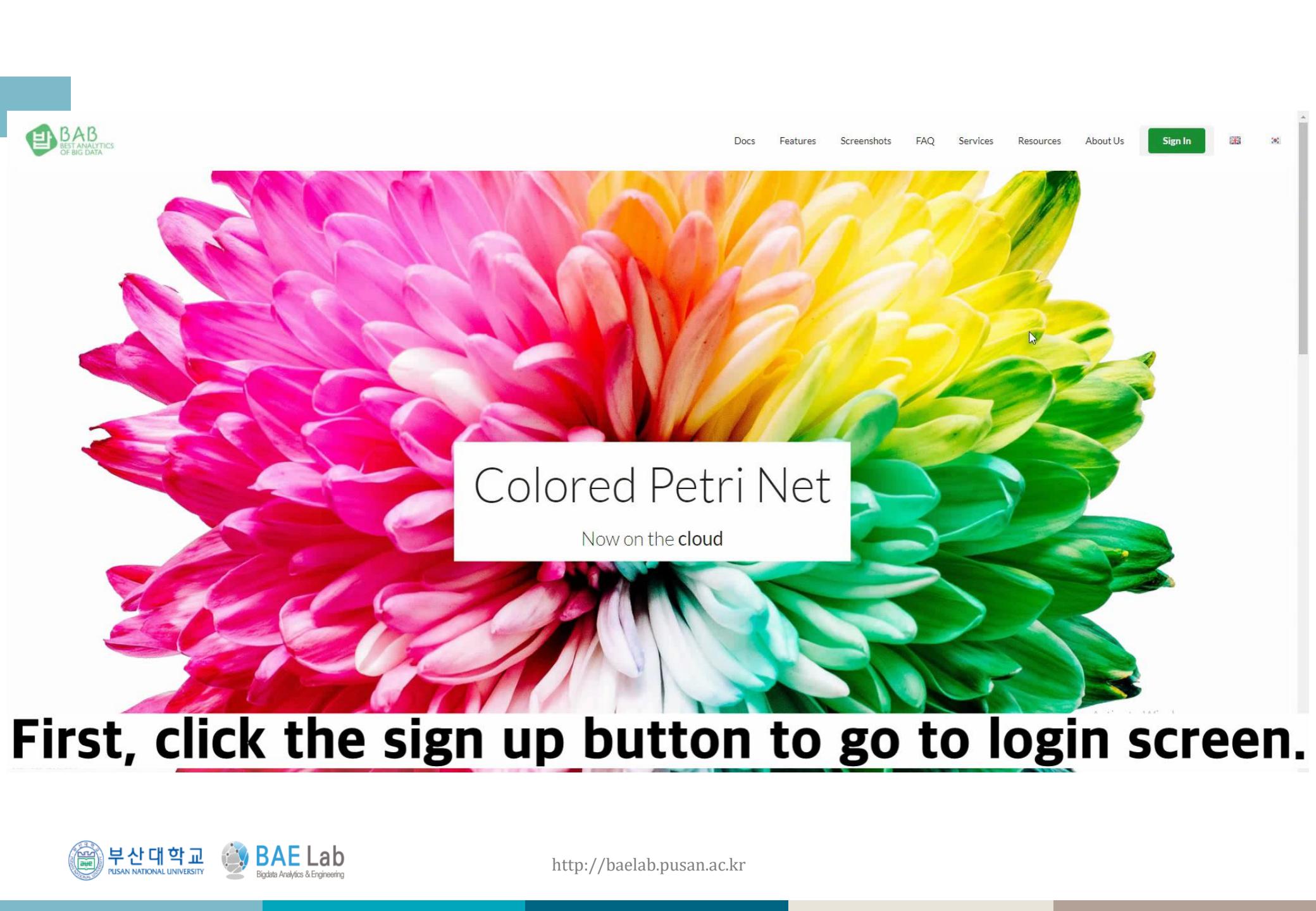
Process simulation

Simulation vs. easy simulation

- To make a model
 - Understand process
- To perform simulation
 - Prepare input data (distribution)



Source: wikipedia



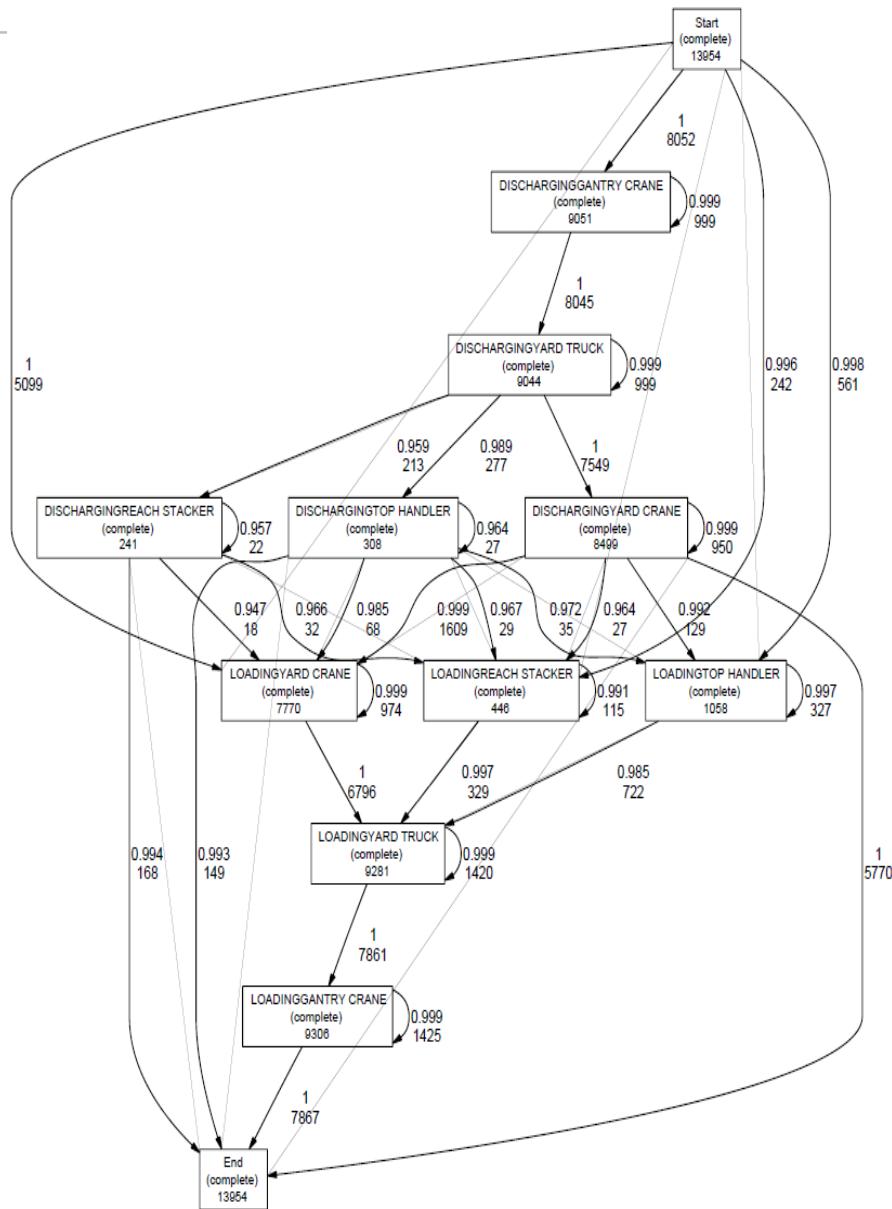
Colored Petri Net

Now on the cloud

First, click the sign up button to go to login screen.

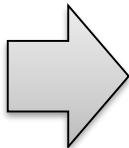
Simulation Analytics

- Process model

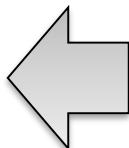
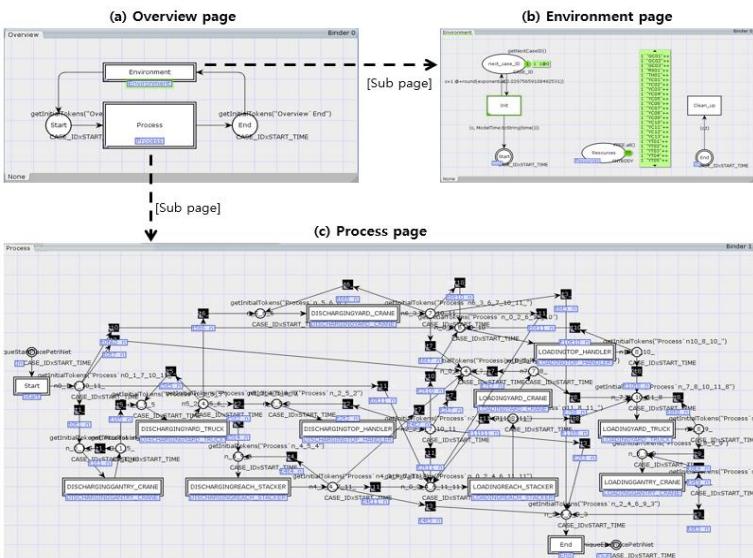


- Data set

Case ID	Activity	Timestamp	Start Time	End Time	Equipment	Attributes
13	Discharging#Gantry Crane	2015-07-07 05:34:00		2015-07-07 5:36:00	GC05	Reefer
13	Discharging#Yard Truck	2015-07-07 5:34:00	2015-07-07 5:41:00		YT06	Reefer
13	Discharging#Yard Crane	2015-07-07 5:41:00		2015-07-07 5:44:00	YC26	Reefer
135	Loading#Reach Staker	2015-07-03 10:36:00		2015-07-03 10:37:00	RS02	General
135	Loading#Yard Truck	2015-07-03 10:19:00	2015-07-03 10:48:00		GC02	General
135	Loading#Gantry Crane	2015-07-03 10:19:00		2015-07-03 10:48:00	YT16	General
...



- Simulation model



Port Logistics Simulation

- Result

- # of gantry crane 4, # of reach stacker 1, # of top handler 1, # of yard crane 13, # of yard truck 37
- Simulation result: Throughput 13,149 containers, (actual :13,954)
 - 5.8% difference
- Application
 - Change the environments and look how the result will be changed
 - Predict the number of equipment required

Number of Equipment	Yard trucks				
	25	30	35	37	40
Gantry cranes	3	9,165	8,754	9,112	8,672
	4	13,351	13,246	13,131	13,149
	5	17,246	17,475	17,224	17,524
	6	18,237	18,246	17,993	18,079
	7	17,917	17,894	18,090	17,964

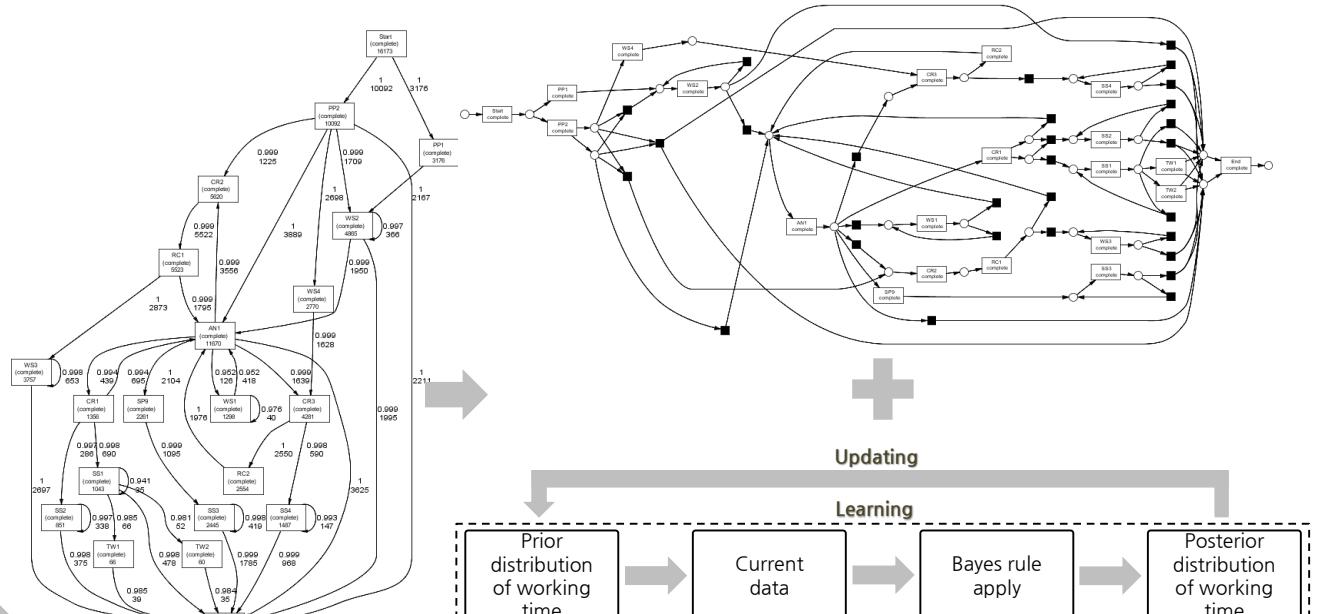
Simulation analytics with Bayesian updating

Event log data(Steel manufacture company data)

Case ID	Activity	Start time	End time
43	CR2	2016-01-11 06:01	2016-01-11 06:01
43	RC1	2016-01-11 06:21	2016-01-11 06:22
43	WS3	2016-01-11 08:06	2016-01-11 08:08
43	SS4	2016-01-11 14:01	2016-01-11 14:06
44	CR2	2016-01-11 06:01	2016-01-11 06:01
44	RC1	2016-01-11 06:21	2016-01-11 06:22
44	WS3	2016-01-11 08:06	2016-01-11 08:08
44	SS4	2016-01-11 14:01	2016-01-11 14:06
...

Heuristic model

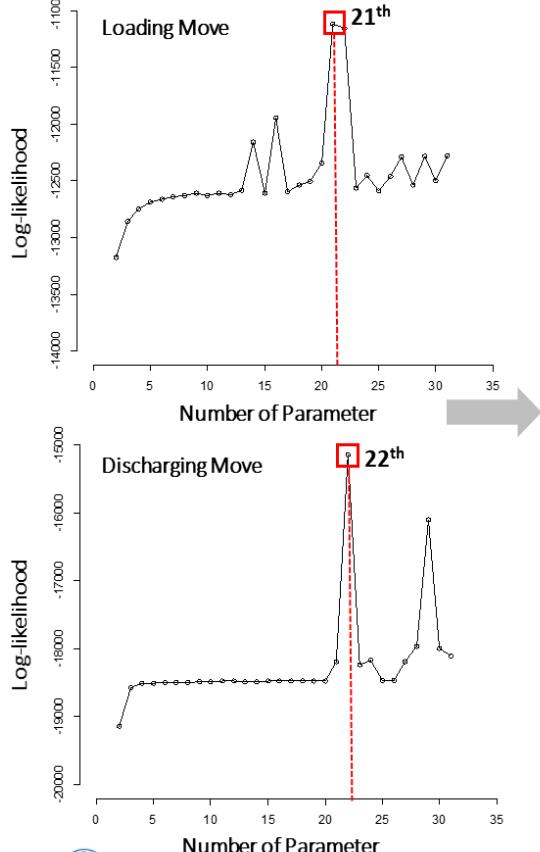
- Dependency threshold : 0.9
- Fitness : 1



Using period	Predict period	Actual Number of coils in data	Simulation throughput		Simulation throughput (with Bayesian inference result)	
			Number of coil	Percentage	Number of coil	Percentage
2016-01-01~2016-01-21	2016-01-01~2016-01-21	17,079	16,843	2.52 %	-	-
2016-01-01~2016-01-21	2016-01-22~2016-01-27	9,343	8,017	14.19 %	10,078	5.86 %

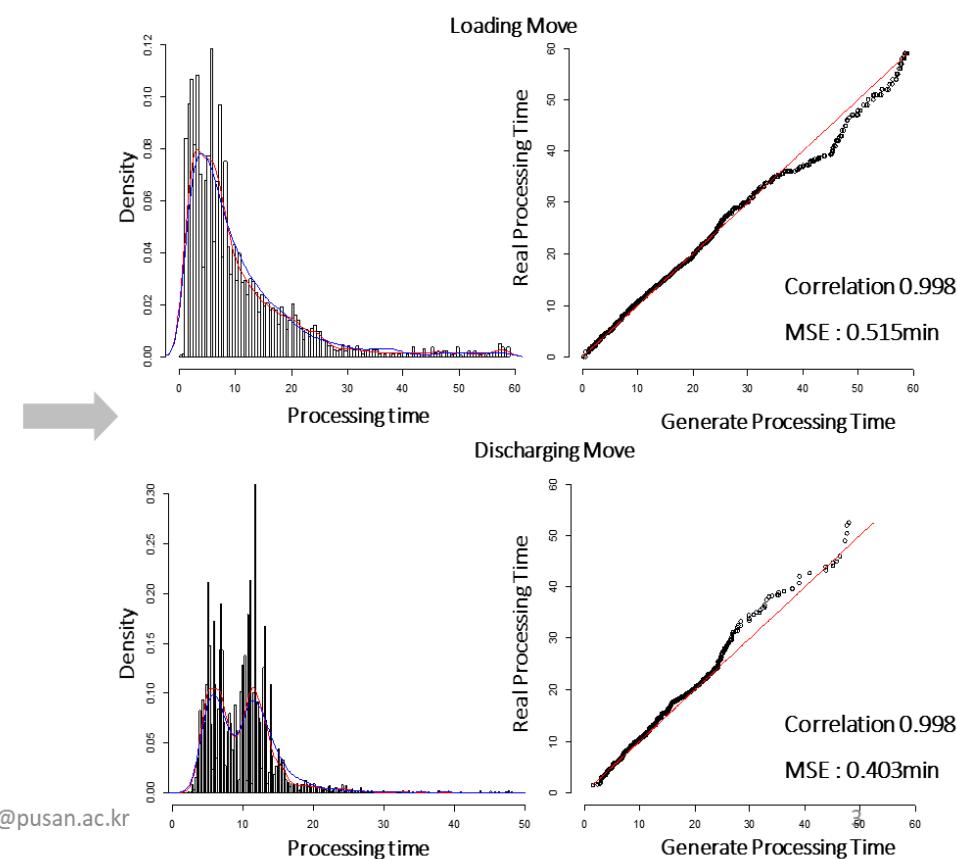
Distribution fitting

Generate parameter using EM algorithm

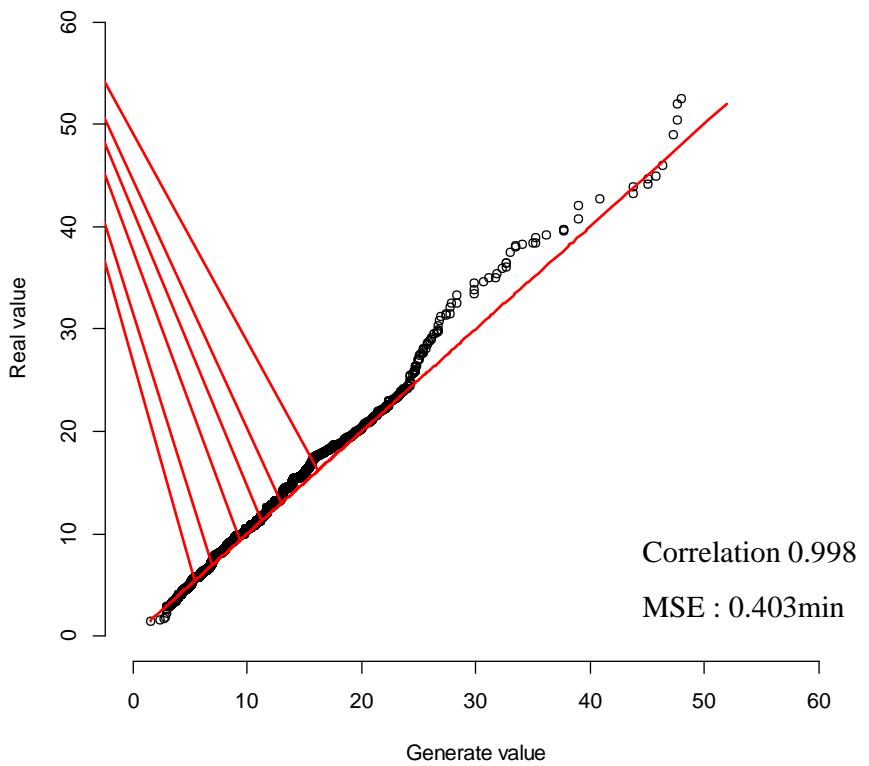
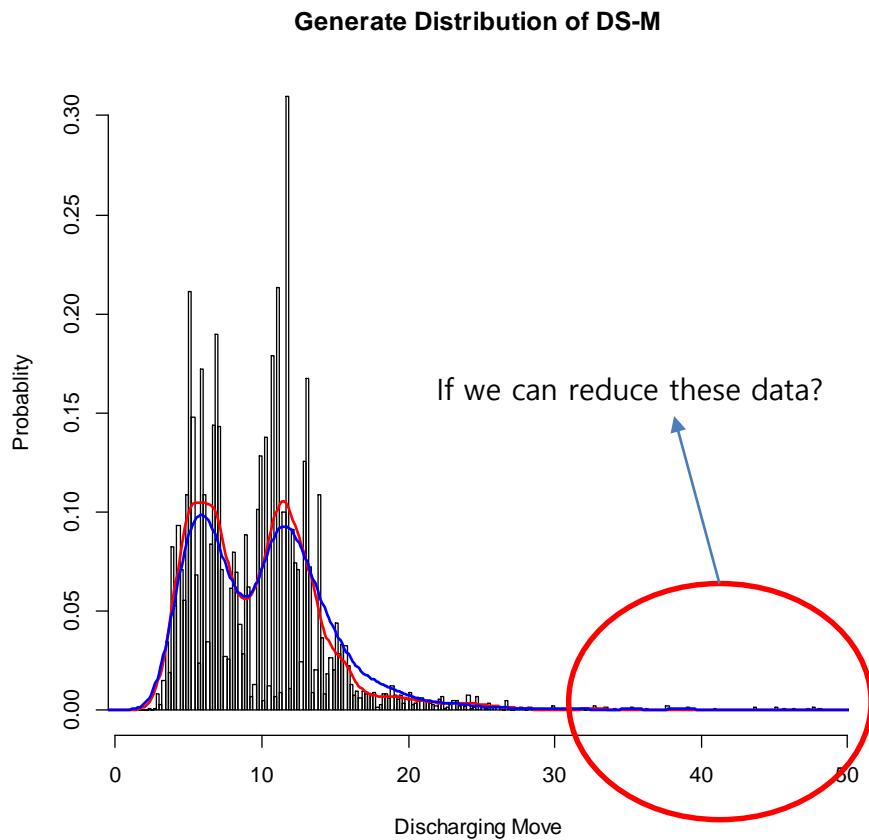


P	Mu	Sd
0.024106	1.596953	0.005459
0.039399	6.358991	0.040802
0.115863	32.68136	114.646
0.111198	10.72172	0.636048
0.023069	15.1706	0.075701
0.027475	16.33176	0.189188
0.080093	4.428641	0.08901
0.043027	12.99287	0.175449
0.039917	18.08647	0.389651
0.013997	19.71049	0.107697
0.09746	8.357846	0.311623
0.035511	1.177455	0.03849
0.052618	2.071511	0.038044
0.050804	7.069218	0.035666
0.035511	5.140146	0.01461
0.119233	3.184928	0.182327
0.053655	5.708535	0.04299
0.001555	1.716667	0.01201
0.009072	20.61524	0.036771
0.004666	13.78426	0.001878
0.021773	14.21548	0.064396

Generate distribution using Metropolis-Hastings Algorithm



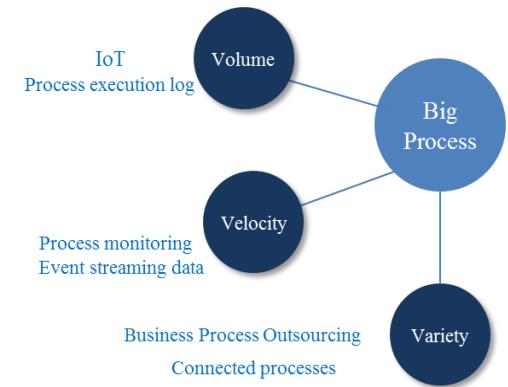
What can we do more?



Conclusions

- Data-process compliance is a beginning of Industry 4.0
- Using operational big-data, we can
 - Better understand process
 - Find where a problem is and what is the cause
 - Predict process result
- Using simulation
 - Forecast KPI
- Using AI
 - We can know what will happen

- Volume
 - Terabyte, petabyte
- Velocity
 - Batch and real-time
- Variety
 - Structured and unstructured





운영빅데이터를 분석하는 최고의 Operational Intelligence 도구

BAB 클라우드 서비스와 함께 운영빅데이터 분석의 맛나는 세상을 만나보세요.

Manufacturing, Logistics, Distribution, Sports, Healthcare, Education, ...

Thank you

