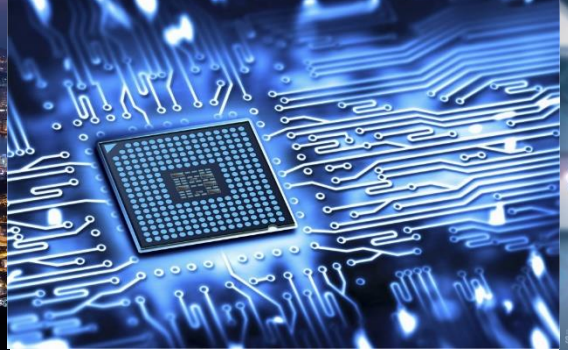




Process Innovation using operational Big Data

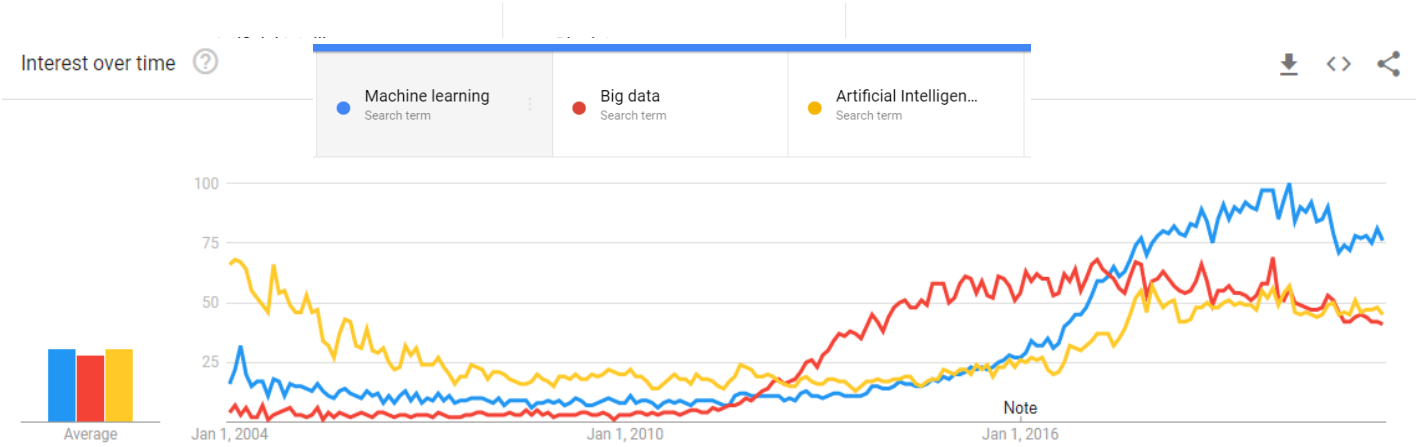
Hyerim Bae

Contents

4차산업혁명 ?

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

- Artificial intelligence vs. Big data vs. Machine Learning



hrbae@pusan.ac.kr

Why Big data?

- 속도
  - 트렌드를 즉시 잡아내야 한다.
  - 자료의 생명력
- 커버리지
  - 설문조사: 수십개의 문항으로 응답자의 지난 몇 년동안의 변화를 추적할 수 있는가?
  - "엄마가 좋아? 아빠가 좋아?"
- 샘플링
  - 5000명의 생각이 전체 5000만명의 생각과 같은가?
    - *"세상에서 가장 높은 산은?"* vs. *"세상에서 가장 인기있는 가수는?"*

출처: 함유근, 채승병, "빅데이터 세상을 바꾸다"

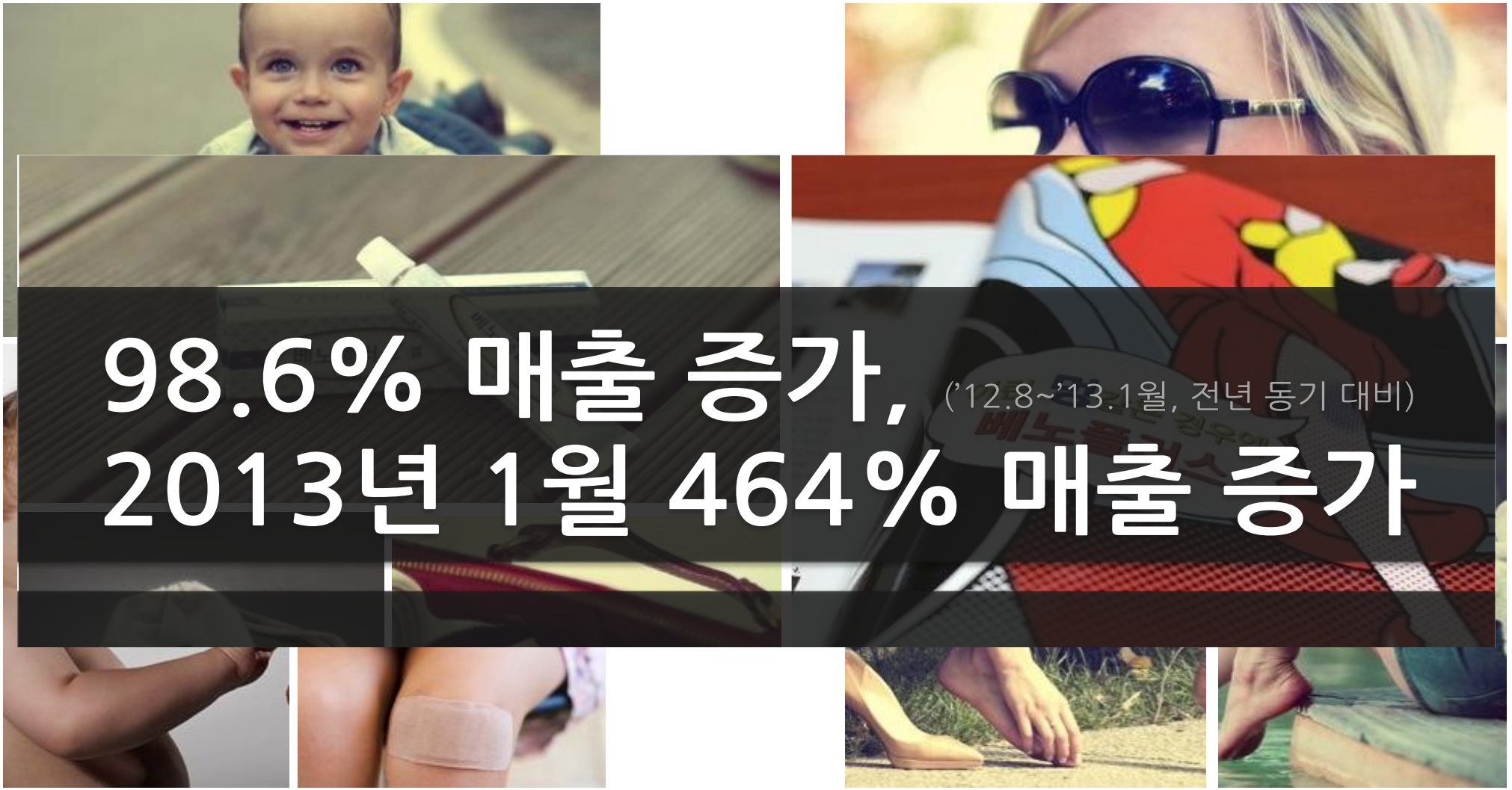
With big data

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

hrbae@pusan.ac.kr

Big Data 로 할 수 있는 것 (이하 출처: 송길영, "여기에 당신의 욕망이 보인다")

휴가와 기온의 상관관계



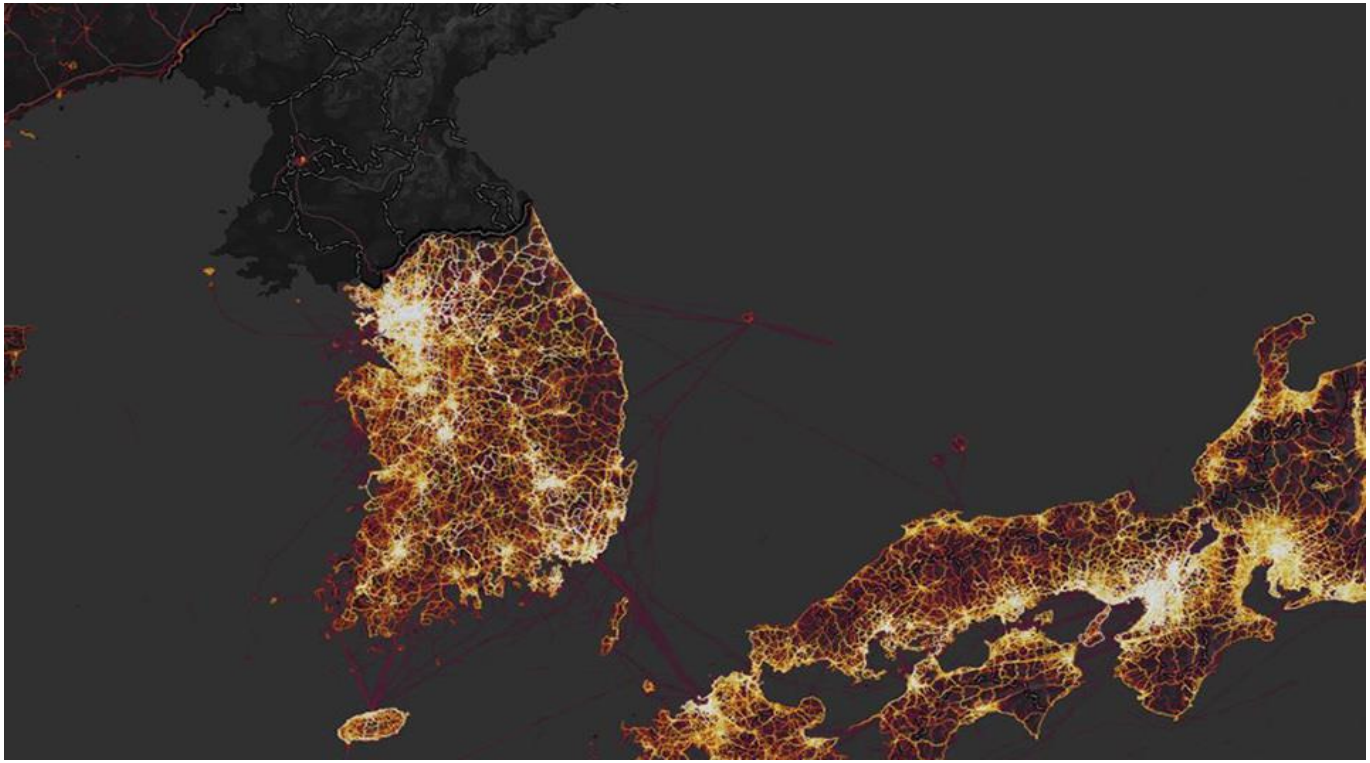
98.6% 매출 증가, 2013년 1월 464% 매출 증가

('12.8~'13.1월, 전년 동기 대비)

hrbae@pusan.ac.kr

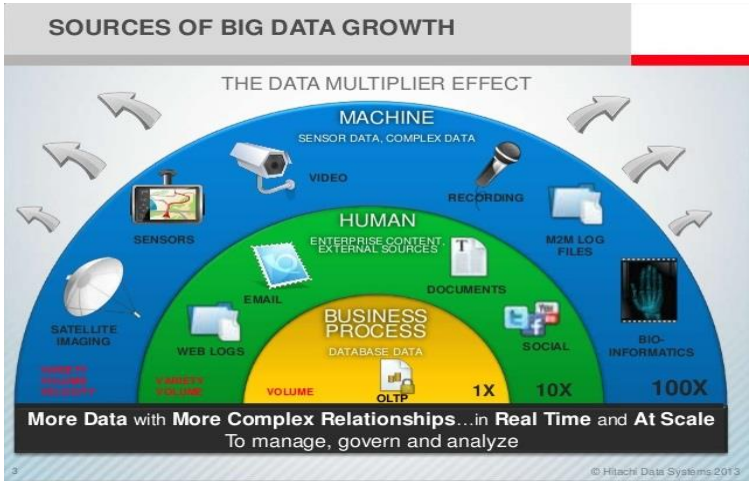
Big-data

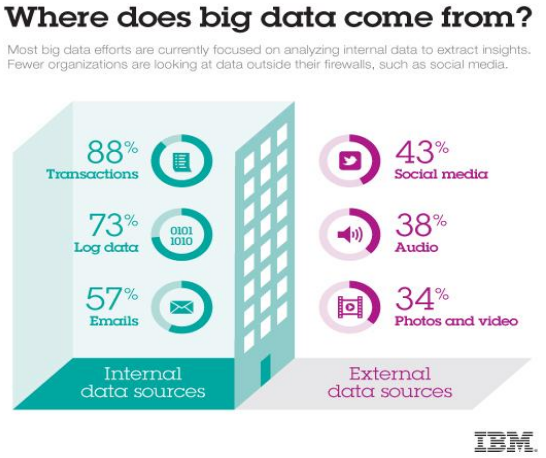
- Why does everybody want to know about big-data?



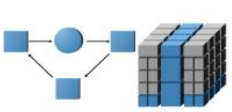
BI vs. OI

- Where do we have Big-data?





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

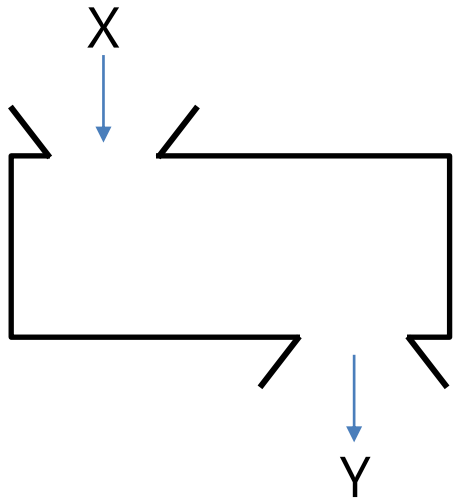
- Time series unstructured data, no predefined schema
- Generated by all IT systems, highly diverse formats
- Massive volume; fast navigation and correlation paramount

Machine Learning

- Finding 'f'

$$Y = f(X)$$

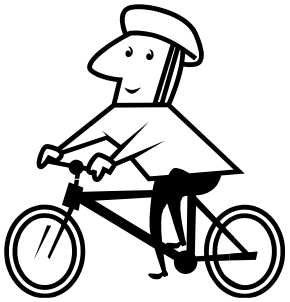
rule
pattern
knowledge



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.



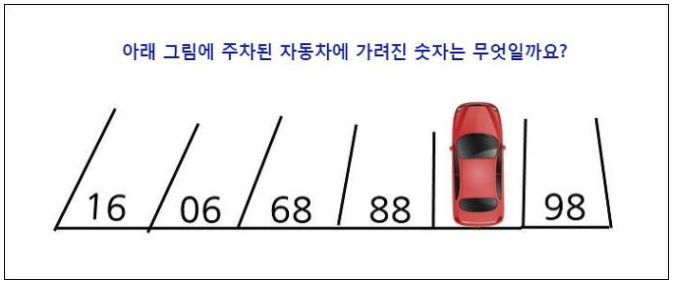
- Why?
  - Very active and large area of AI
  - Biological and cognitive perspective
    - Desire to understand more about our selves
  - Get machines to perform tasks that serve us in some way

Learning methods

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

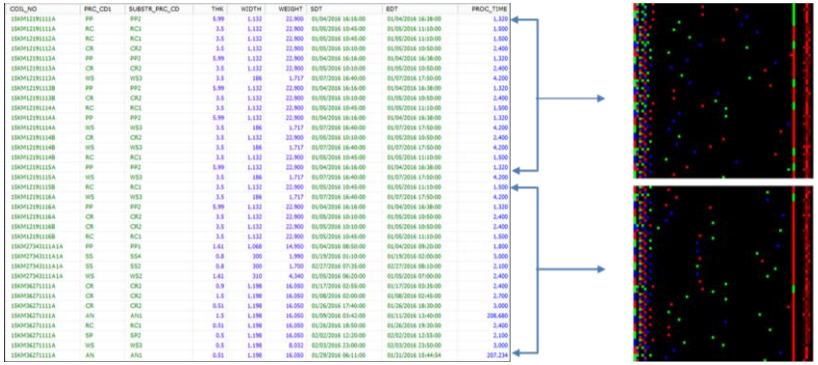
We need data

- 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

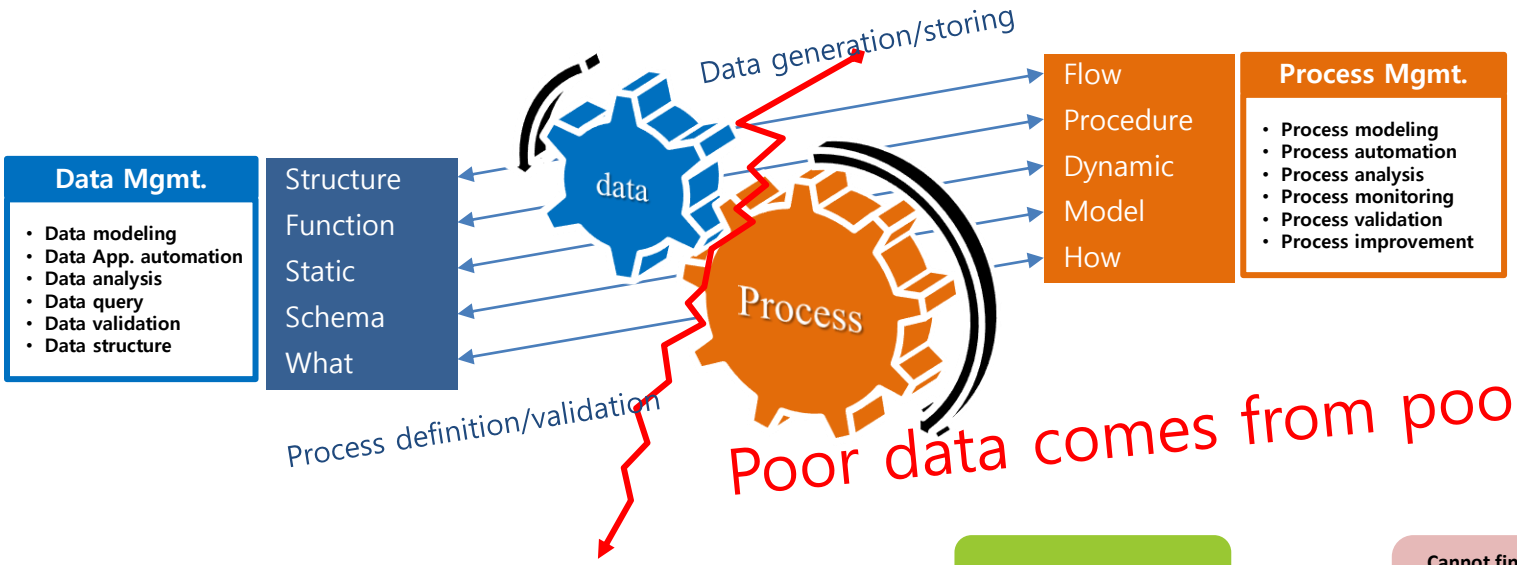


What methods do we need to use?

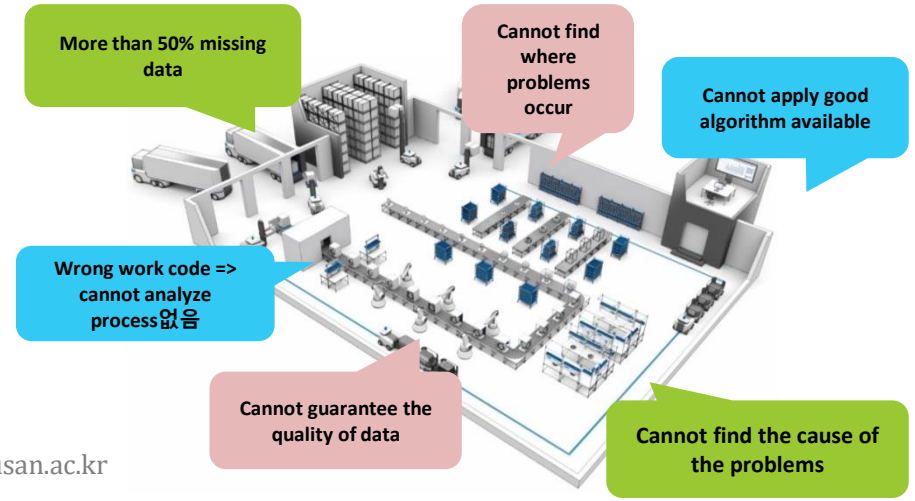
- Different methods for different data
- Multi-purposed model

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 나눠지며, 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

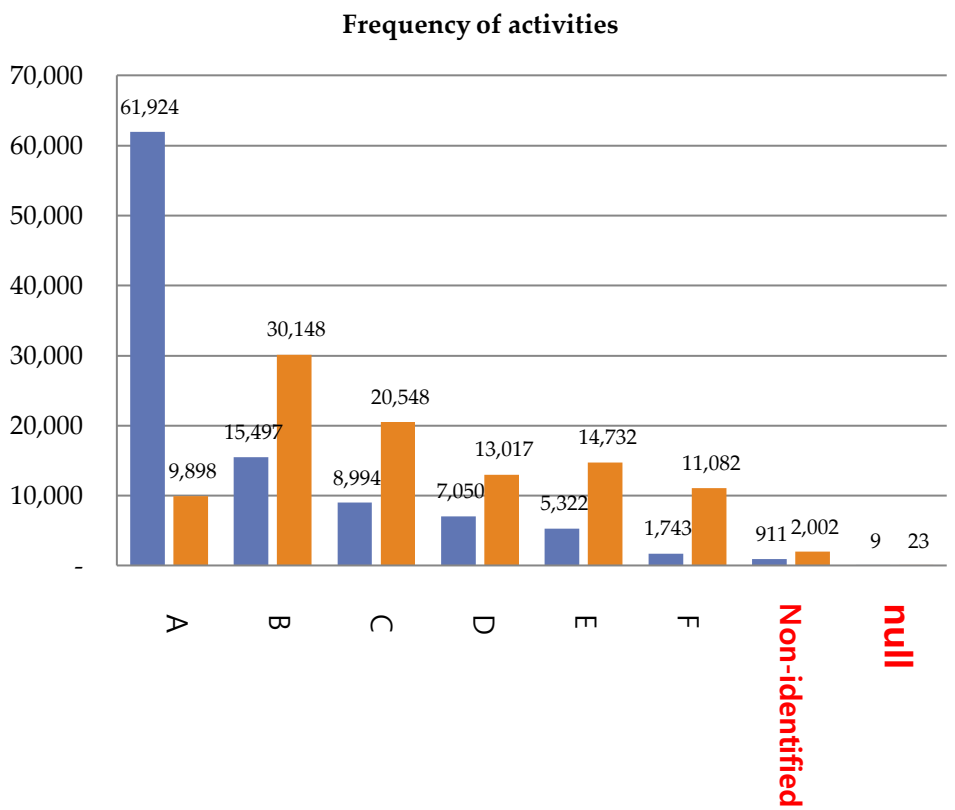
Big-data vs. process improvement

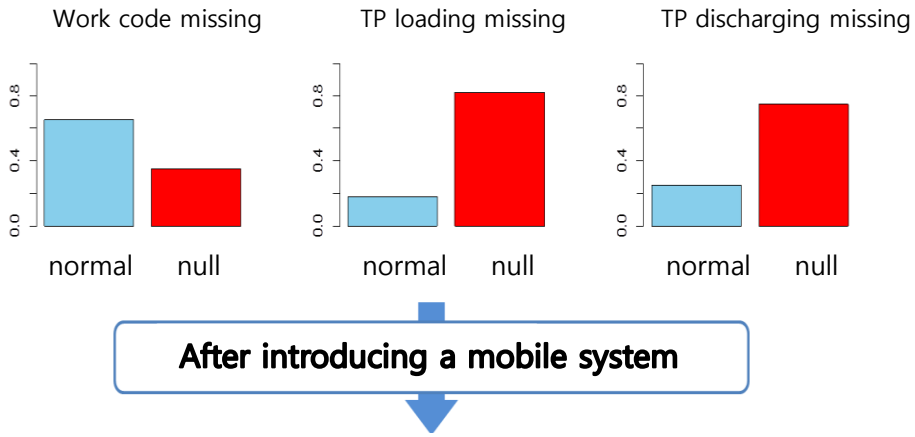


Data-Process Separation	
Undefined process	Missing data
Error in process definition	Data error
Unstandardized process	Low data quality
Process that are not digitalized	Low data value

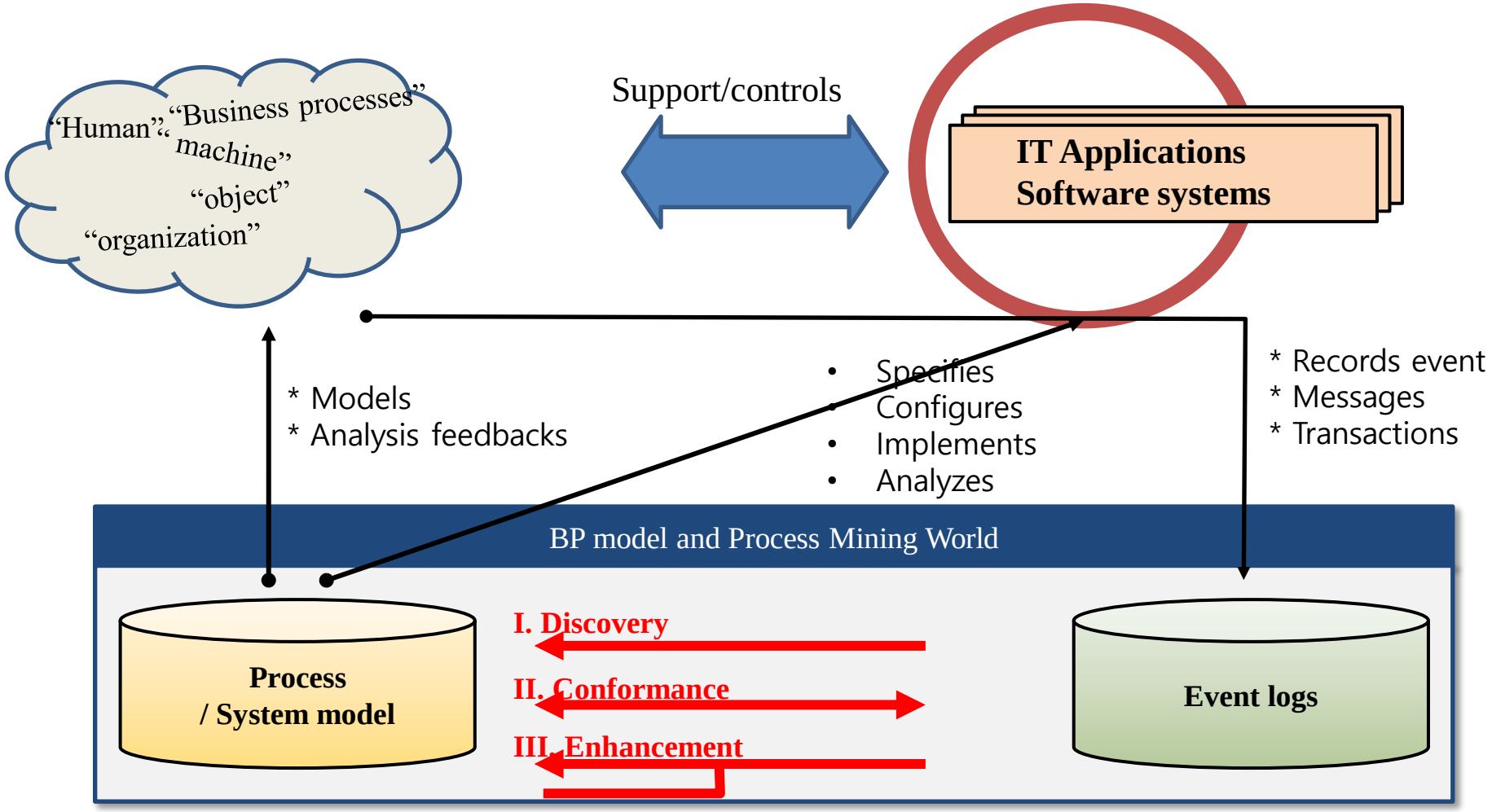


- Is technology the only issue for Industry 4.0?





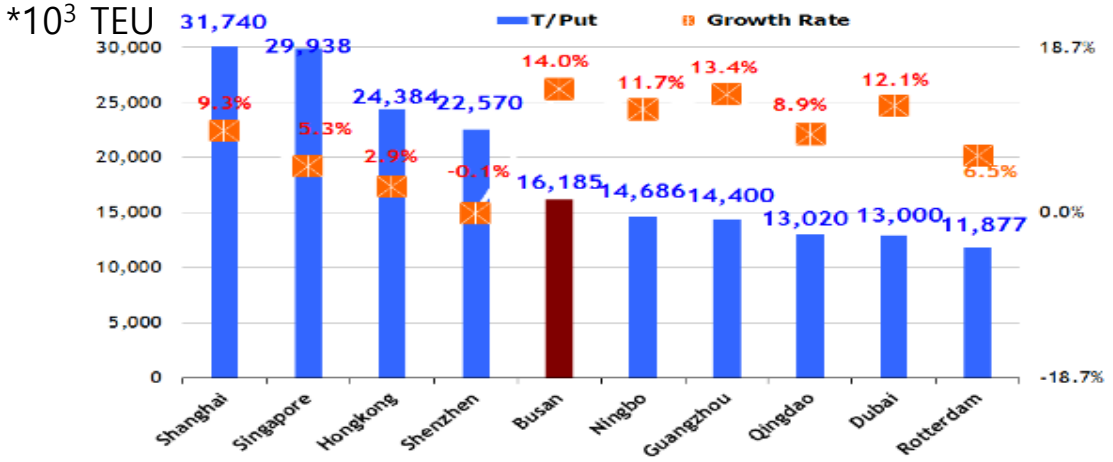
Process Analytics



Process Analytics

CASE I: Container handling process in a container port

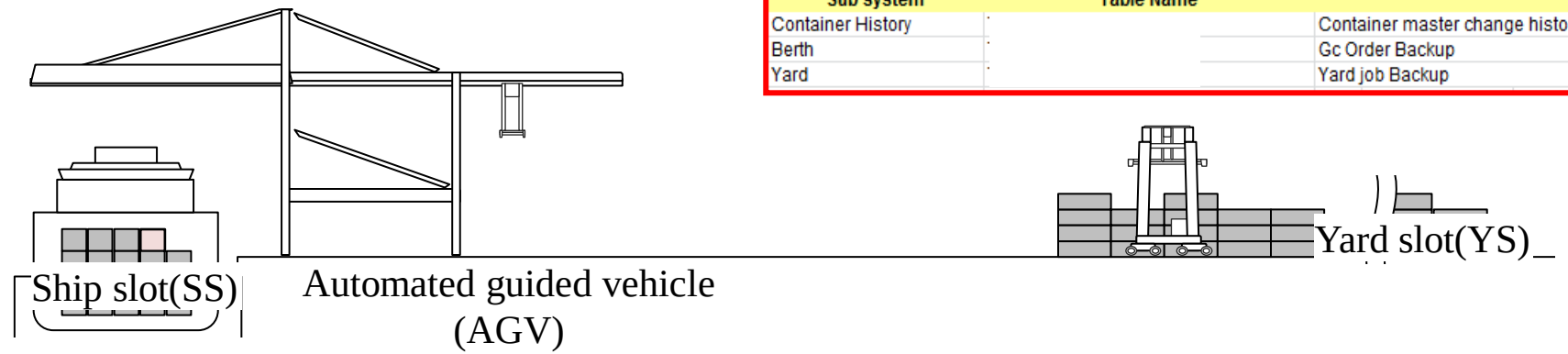
- Busan Port



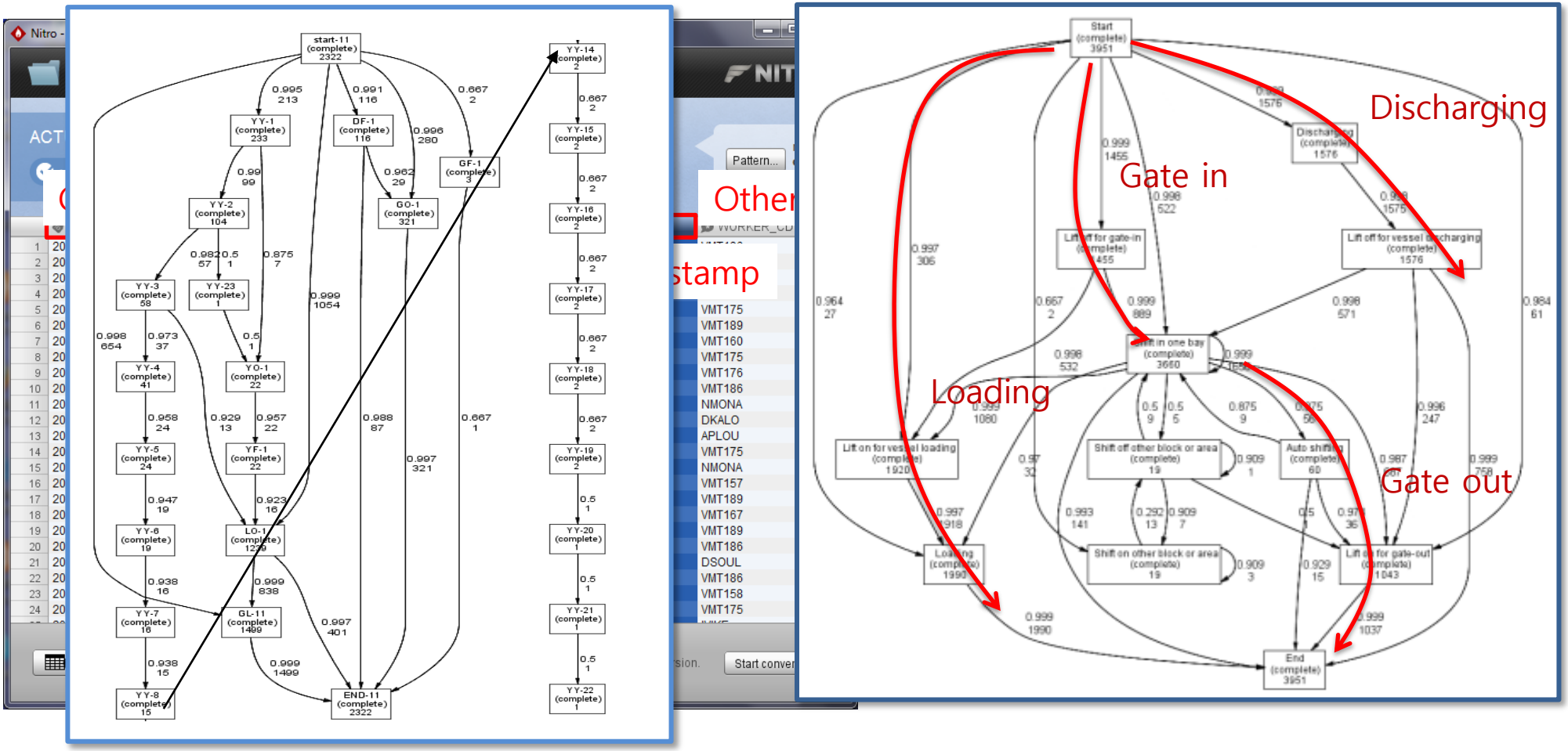
2013년 ranking		
ranking	Port	Container handling
1	Shanghai	3362
2	Singapore	3258
3	Shenzhen	2328
4	Hong Kong	2235
5	**Busan**	1769
6	Ningbo	1733
7	Qingdao	1552
8	Guangzhou	1531

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

Sub system	Table Name	
Container History		Container master change history
Berth		Gc Order Backup
Yard		Yard job Backup



Container handling process discovered



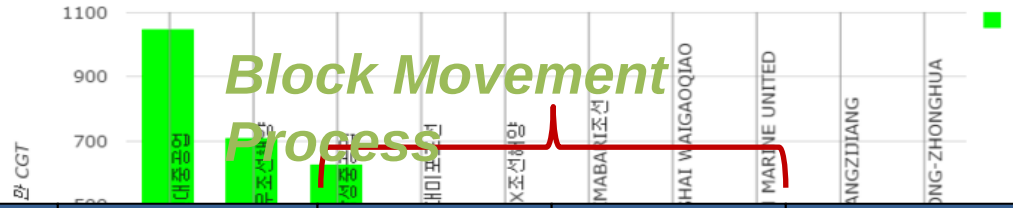
hrbae@pusan.ac.kr

CASE II: Ship Building Process

- Korea is number 1 in ship building in[illegible]

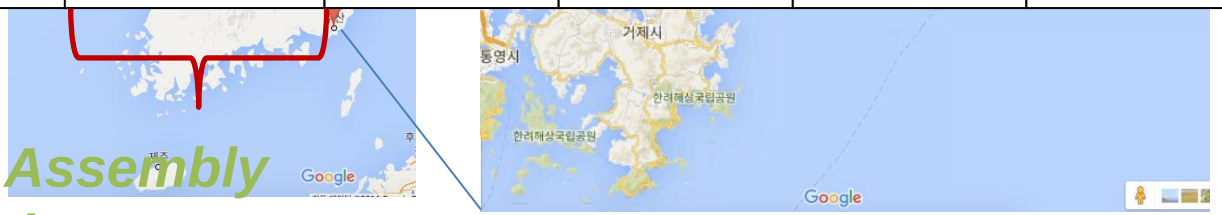
Compensated Gross Tonnage, clarkson

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



Block Movement Process

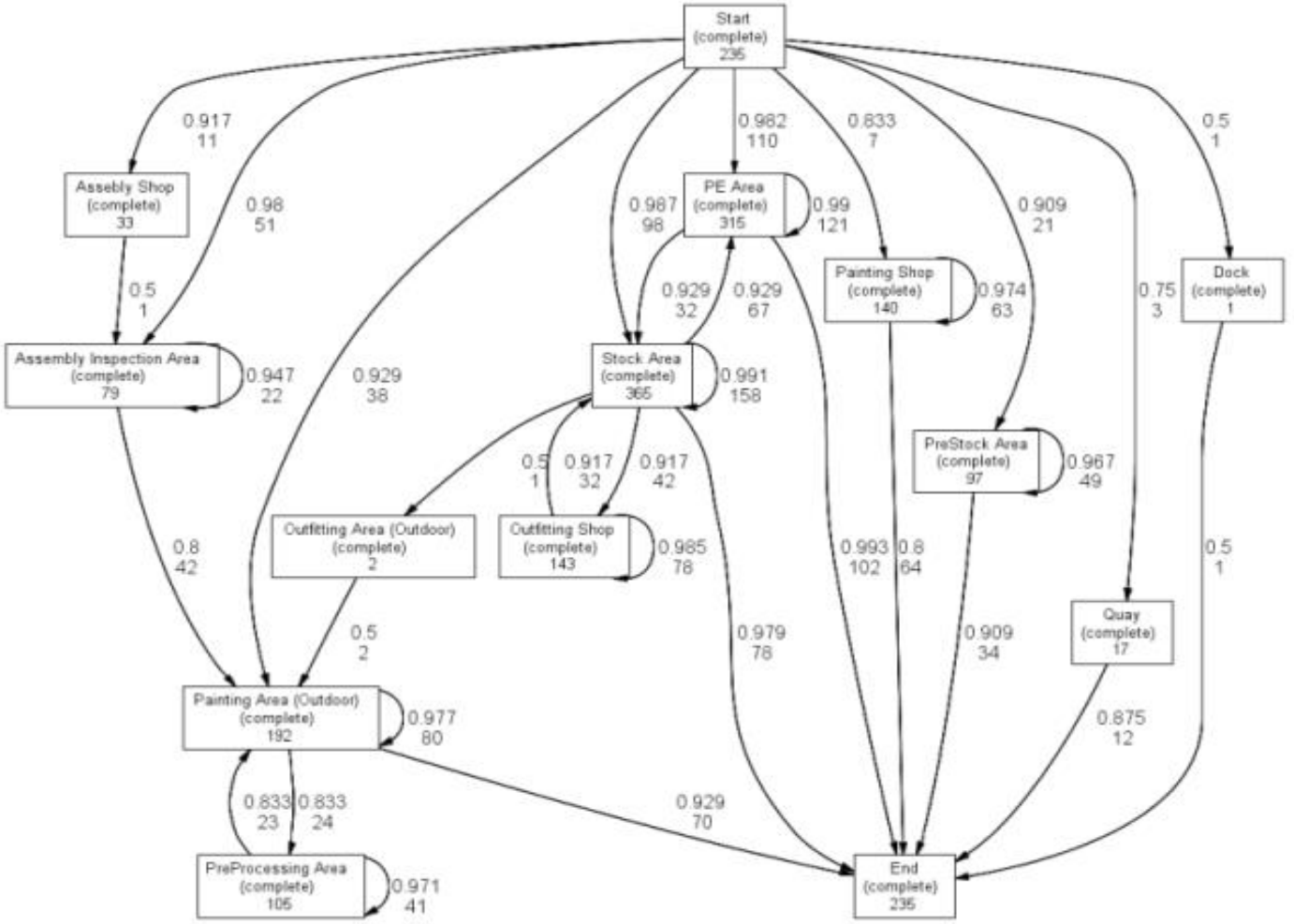
Design	Steel Stock	Cutting & Forming	Assembly	Pre-outfitting Painting	Pre Erection (PE Area)	Erection (Dock)	Quay
Block Division Nesting Plan	Unloading Quay	Pretreatment N/C Cutting Forming Roll Press	Component Sub Assembly Unit Assembly Grand Assembly	Pre outfitting Pre painting	G/C	G/C K/L F/L L/C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



Block Assembly Operations

부산대학교 PUSAN NATIONAL UNIVERSITY
BAE Lab Bigdata Analytics & Engineering
http://baelab.pusan.ac.kr
hrbae@pusan.ac.kr

Process model discovered (Block movement)

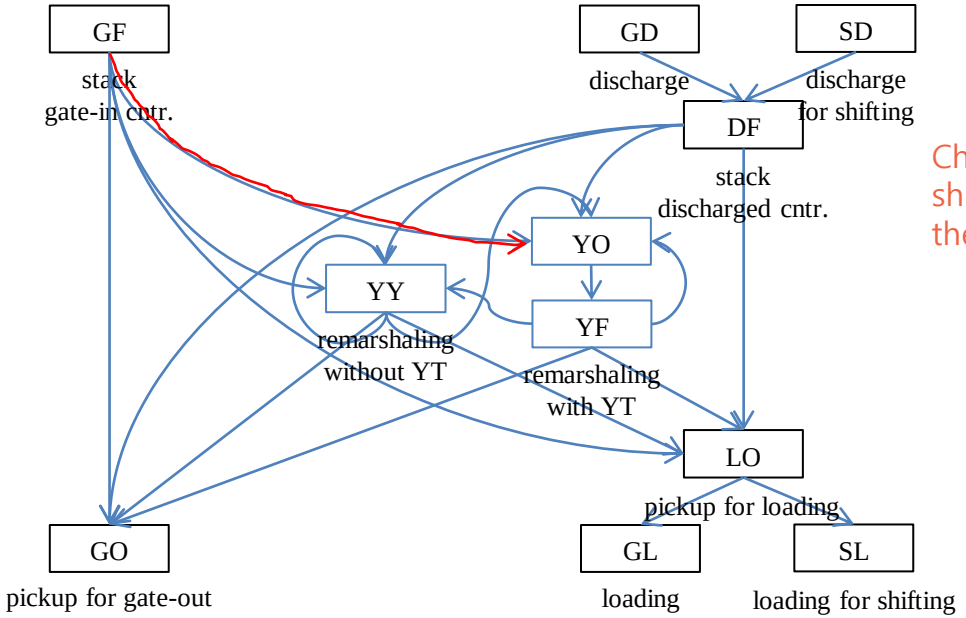


hrbae@pusan.ac.kr

1. For better understanding what we are doing

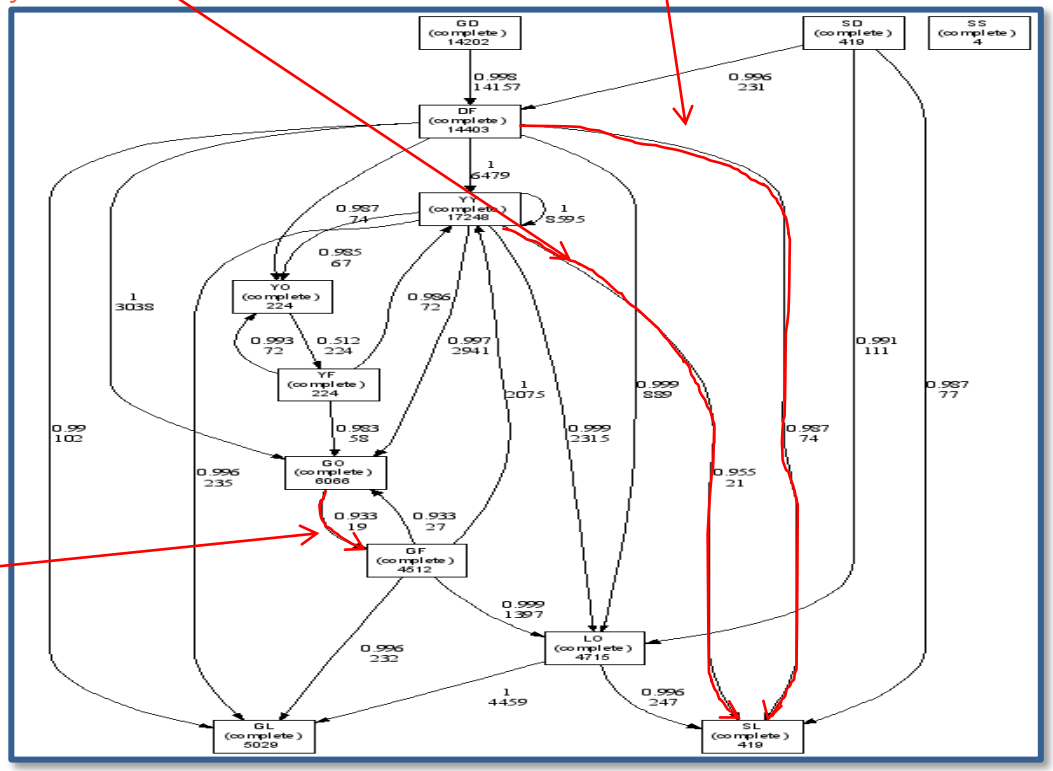
hrbae@pusan.ac.kr

Pre-defined process model vs. Discovered process model



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

shifted by QC without being picked up by YC

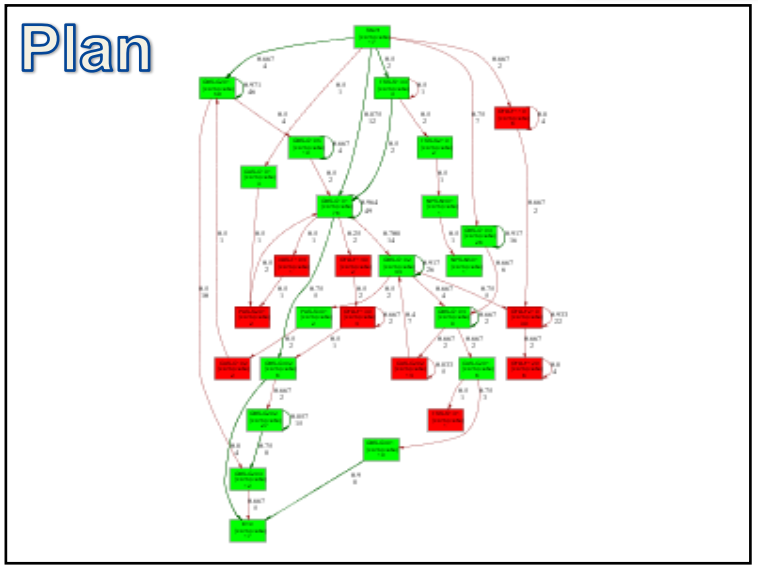


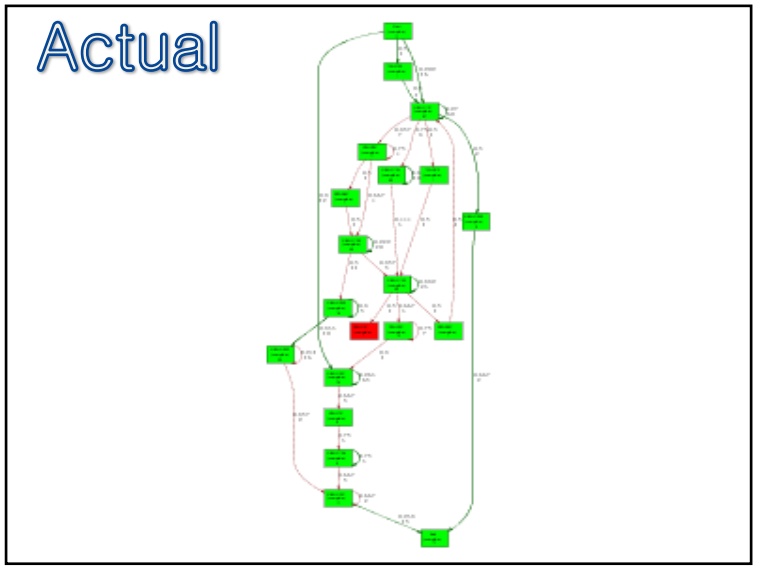
Picked up for Gate out but stacked in the yard again

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



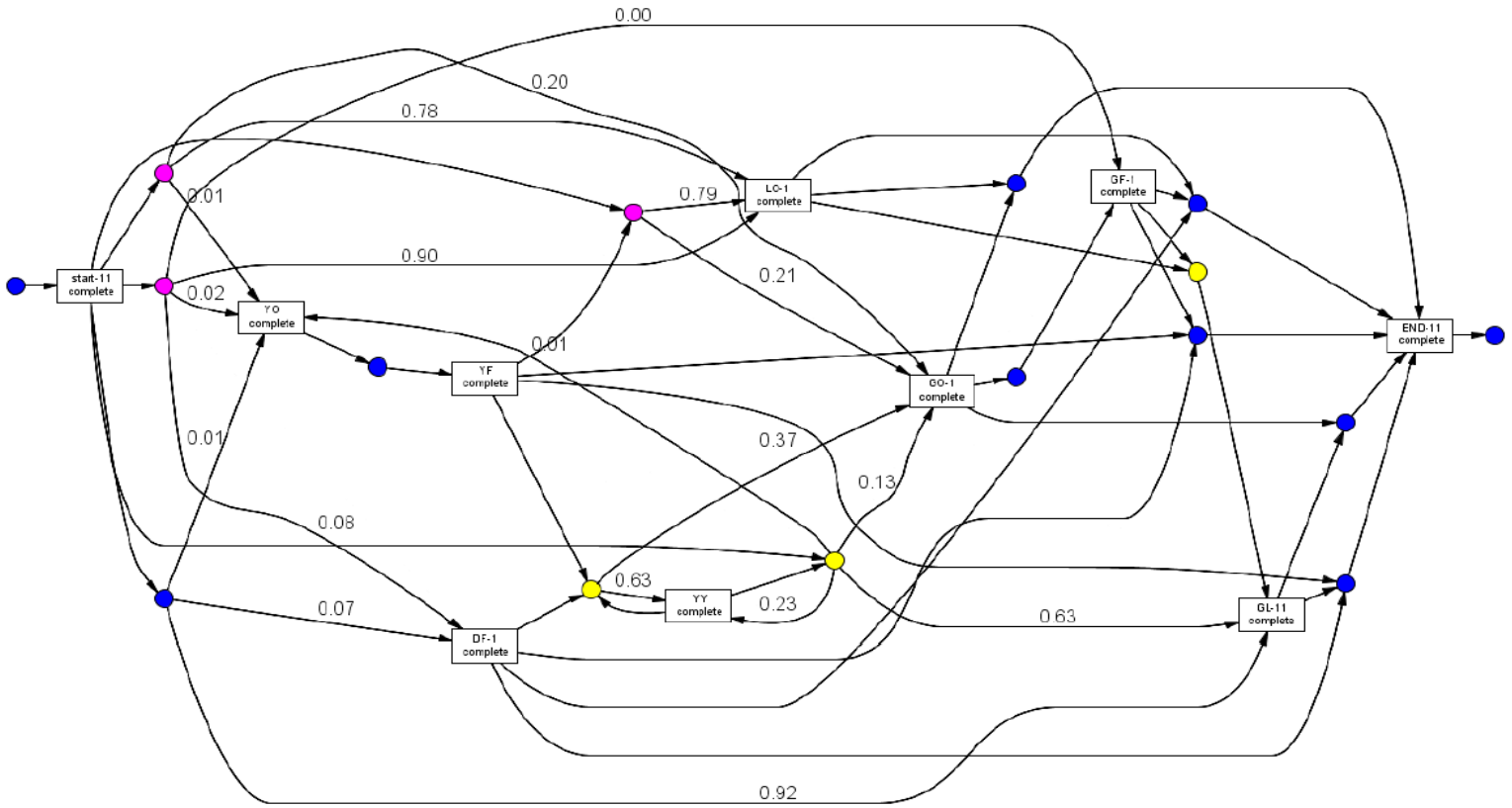


Port example: Better understanding of current situation

2. Finding cause and fixing the problem

hrbae@pusan.ac.kr

What is the bottleneck in the port?



hrbae@pusan.ac.kr

Good flow and Bad flow

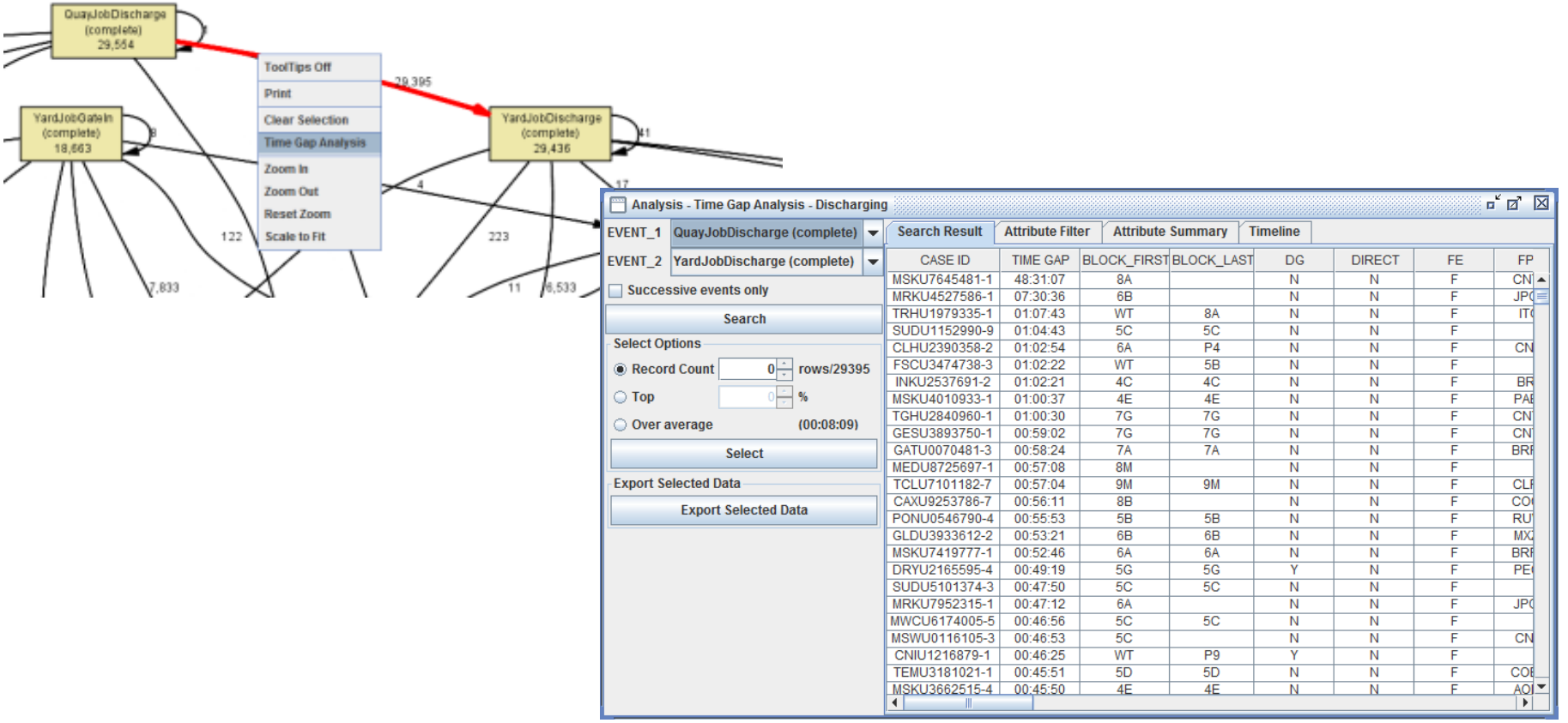
- Process Discovery
  - Good flow and Bad flow

	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

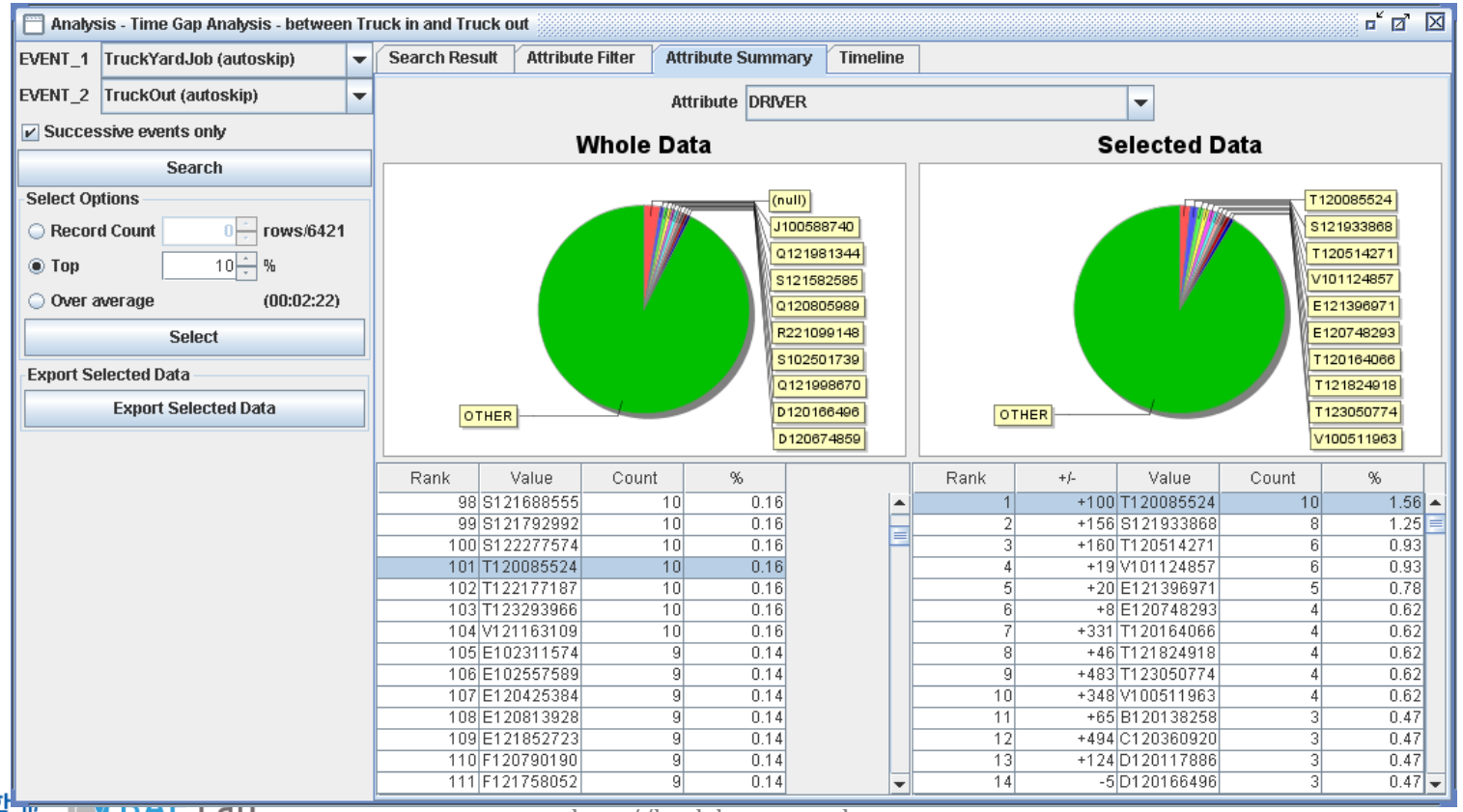
Good flow Irregular flow Bad flow

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



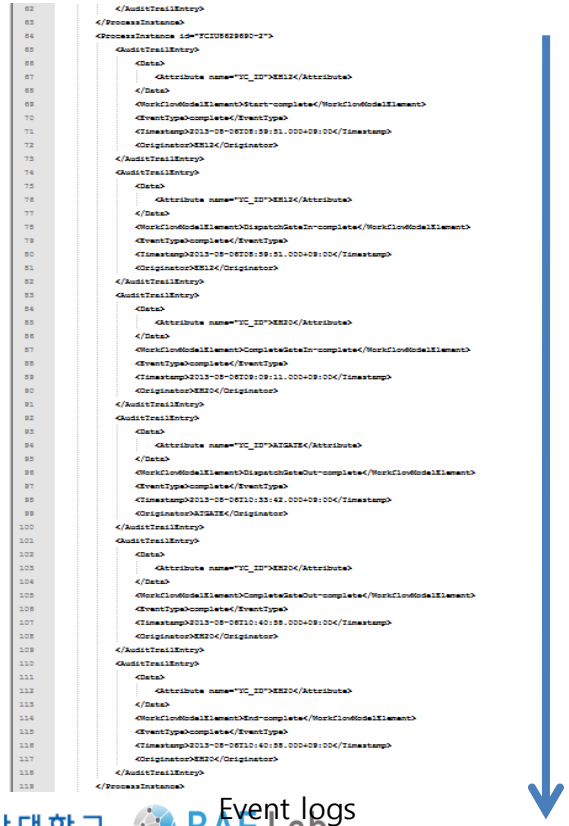
3. For predicting future result

hrbae@pusan.ac.kr

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

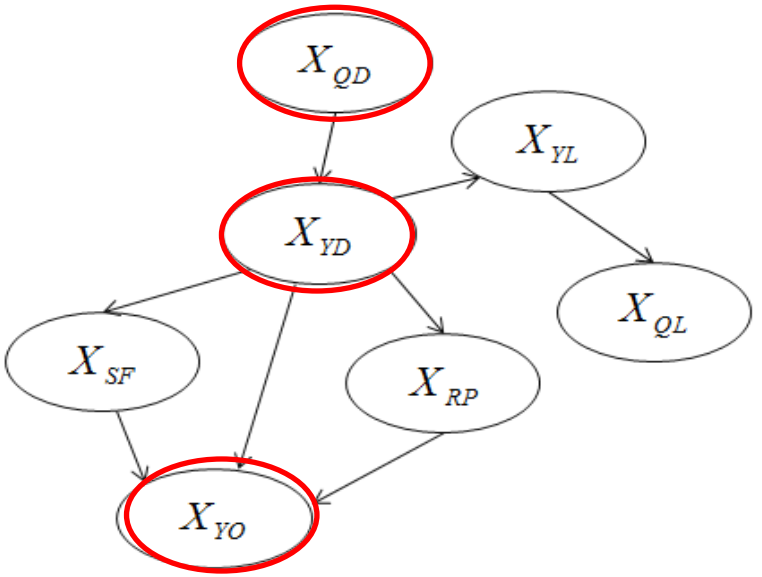
To make one inference, **scan event logs every time**



Event logs

- **If we have Bayesian network?**

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

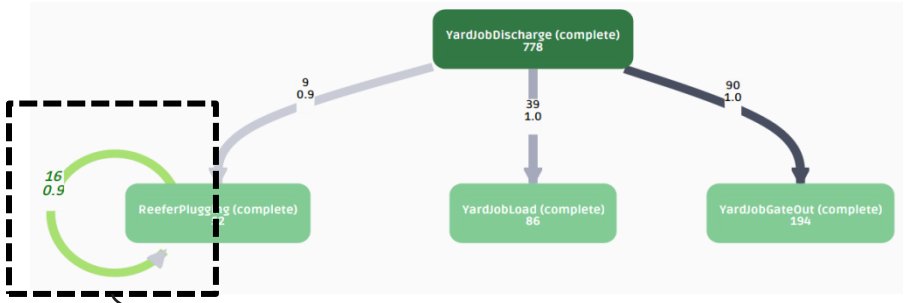


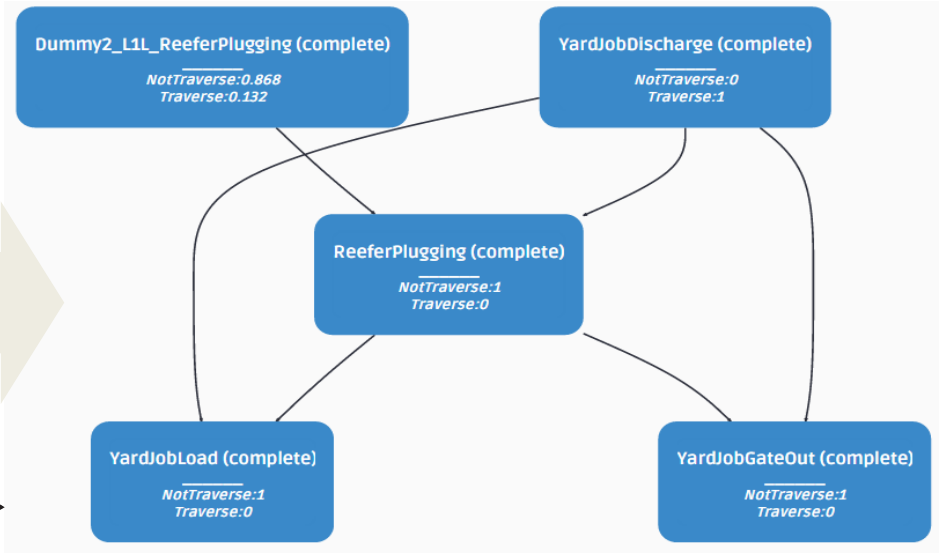
Bayesian networks

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)





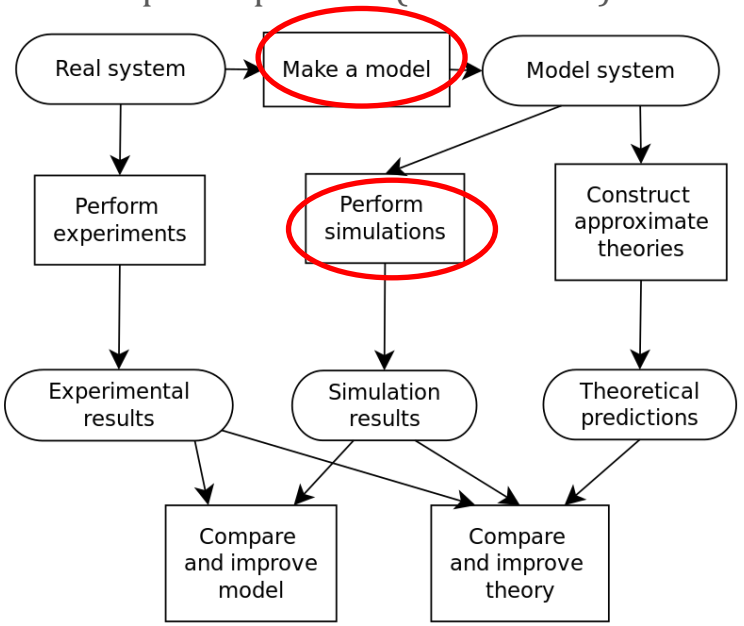
hrbae@pusan.ac.kr

Process simulation

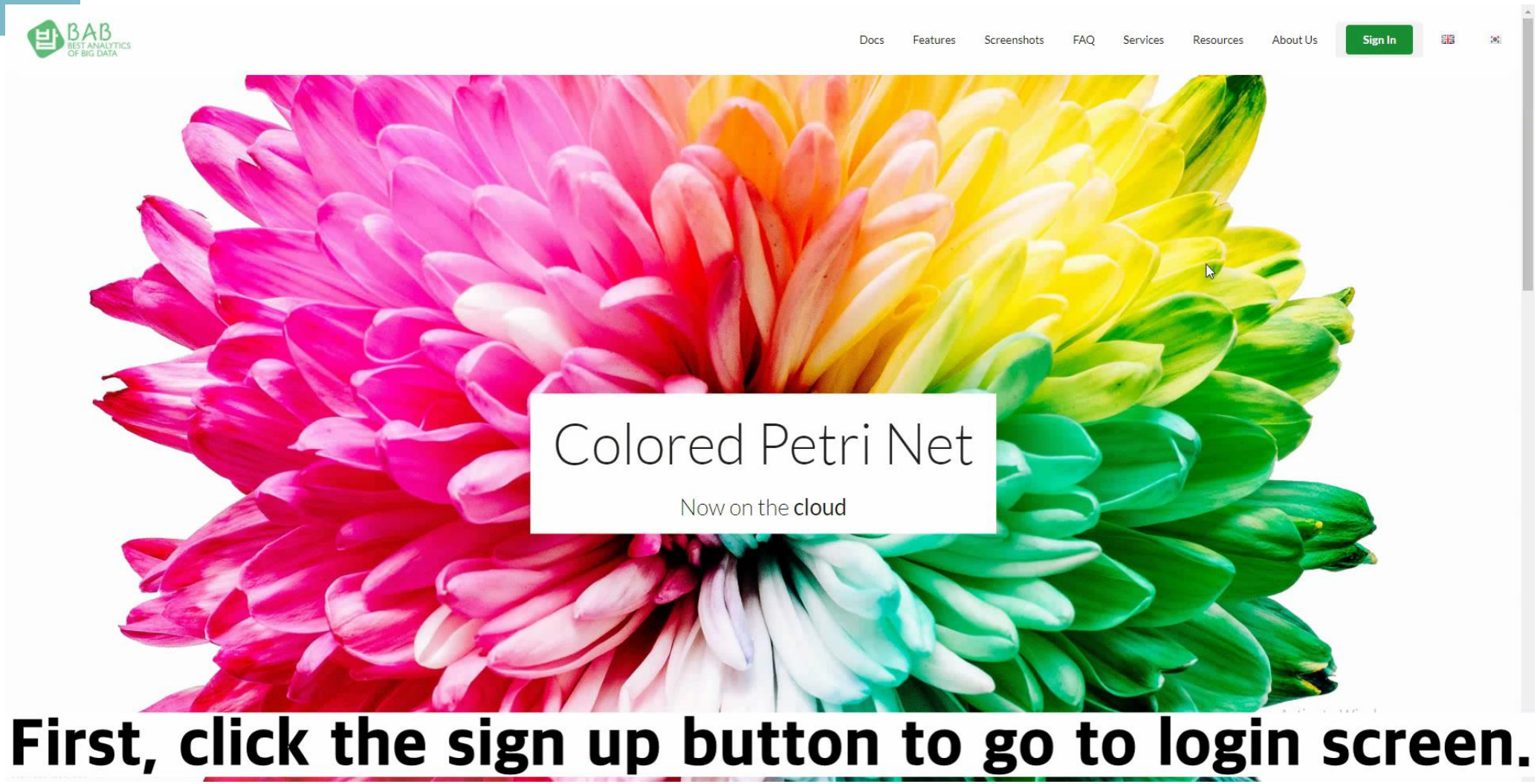
Simulation vs. easy simulation

- To make a model
  - Understand process

- To perform simulation
  - Prepare input data (distribution)



Source: wikipedia

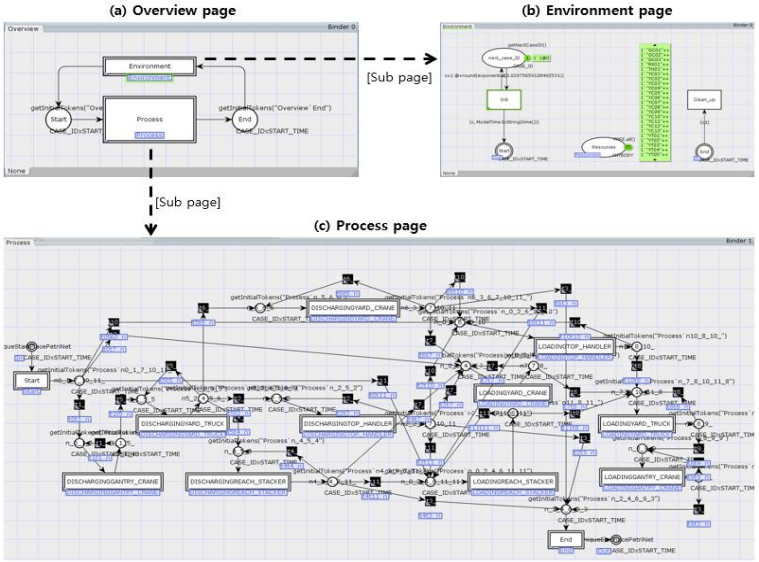

BAB
BEST ANALYTICS OF BIG DATA
Docs
Features
Screenshots
FAQ
Services
Resources
About Us
Sign In
Colored Petri Net
Now on the cloud
First, click the sign up button to go to login screen.

Simulation Analytics

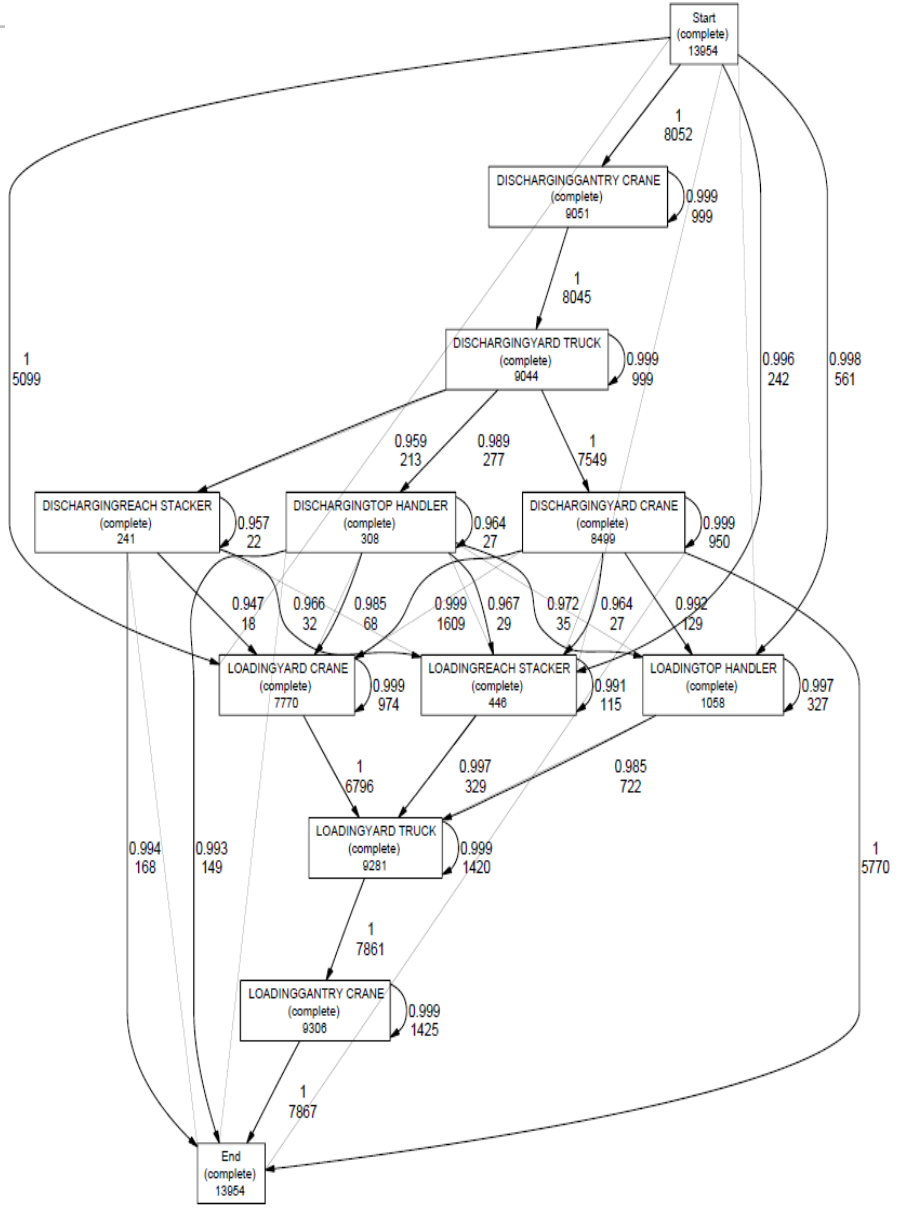
- Data set

Case ID	Activity	Timestamp		Attributes	
		Start Time	End Time	Equipment	Container Type
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



- Process model



Port Logistics Simulation

- Result

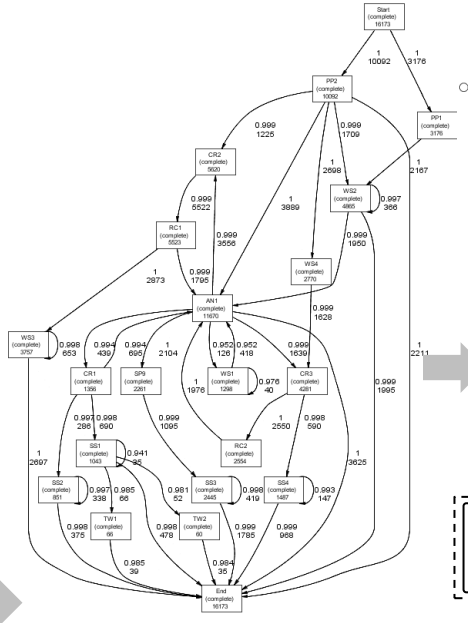
  - # of gantry crane 4, # of reach stacker 1, # of top handler 1, # of yard crane 13, # of yard truck 37
  - Simulation result: Throughput 13,149 containers, (actual :13,954)
    - 5.8% difference
  - Application
    - Change the environments and look how the result will be changed
    - Predict the number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

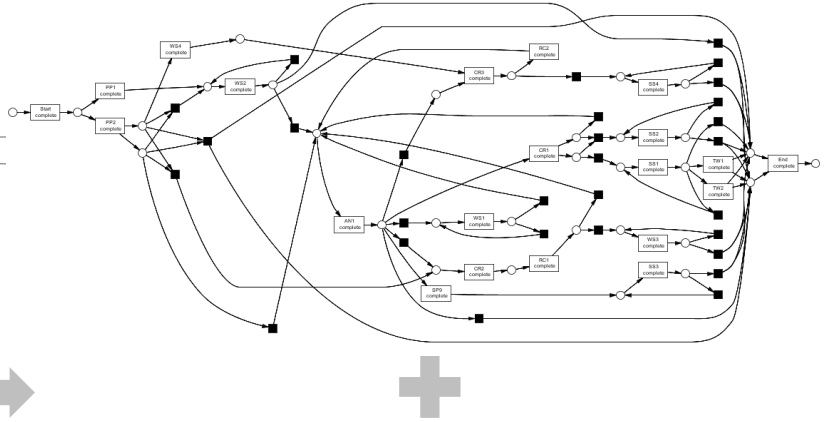
Number of Equipment		Yard trucks				
		25	30	35	37	40
Gantry cranes	3	9,165	8,754	9,112	8,672	9,086
	4	13,351	13,246	13,131	13,149	13,070
	5	17,246	17,475	17,224	17,524	17,201
	6	18,237	**18,246**	17,993	18,079	17,861
	7	17,917	17,894	18,090	17,964	17,886

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
…	…	…	…



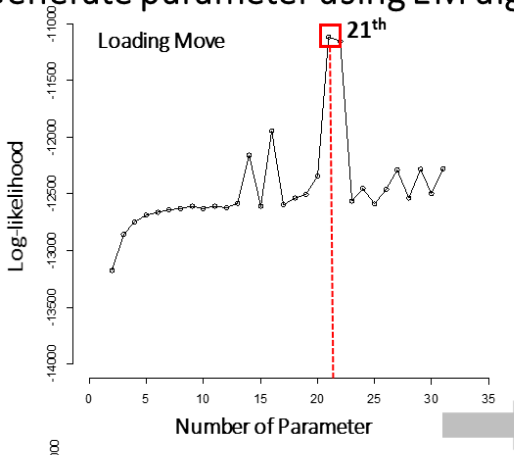


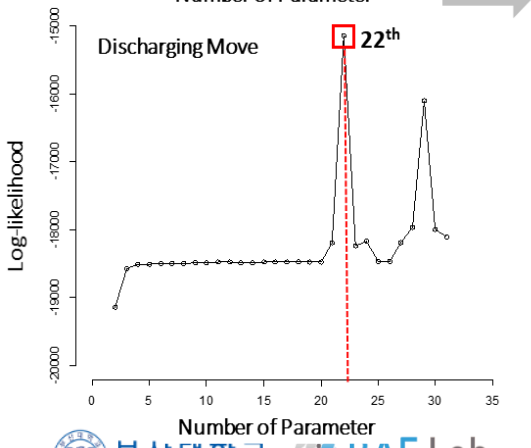
Heuristic model
- Dependency threshold : 0.9
- Fitness : 1

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

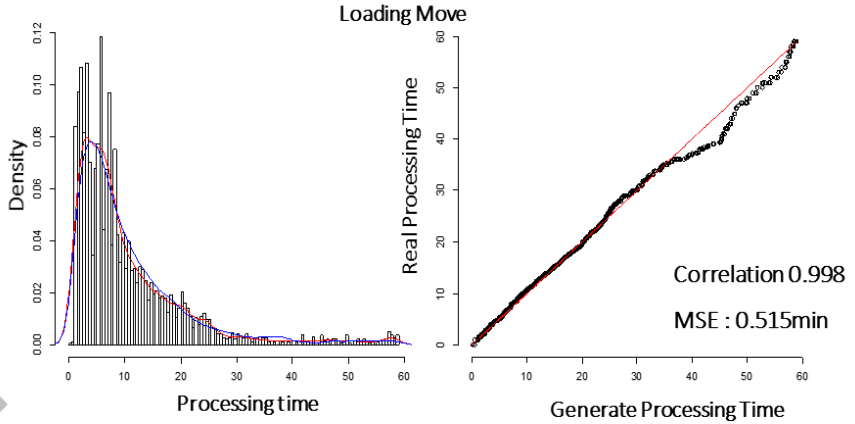


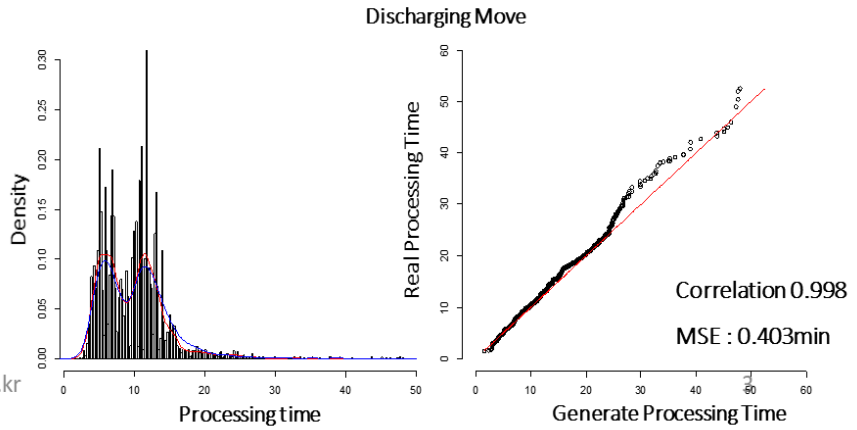


P	Mu	Sd
0.024106	1.596953	0.005459
0.039399	6.358991	0.040802
0.115863	32.68136	114.646
0.111198	10.72172	0.636048
0.023069	15.1706	0.075701
0.027475	16.33176	0.189188
0.080093	4.428641	0.08901
0.043027	12.99287	0.175449
0.039917	18.08647	0.389651
0.013997	19.71049	0.107697
0.09746	8.357846	0.311623
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

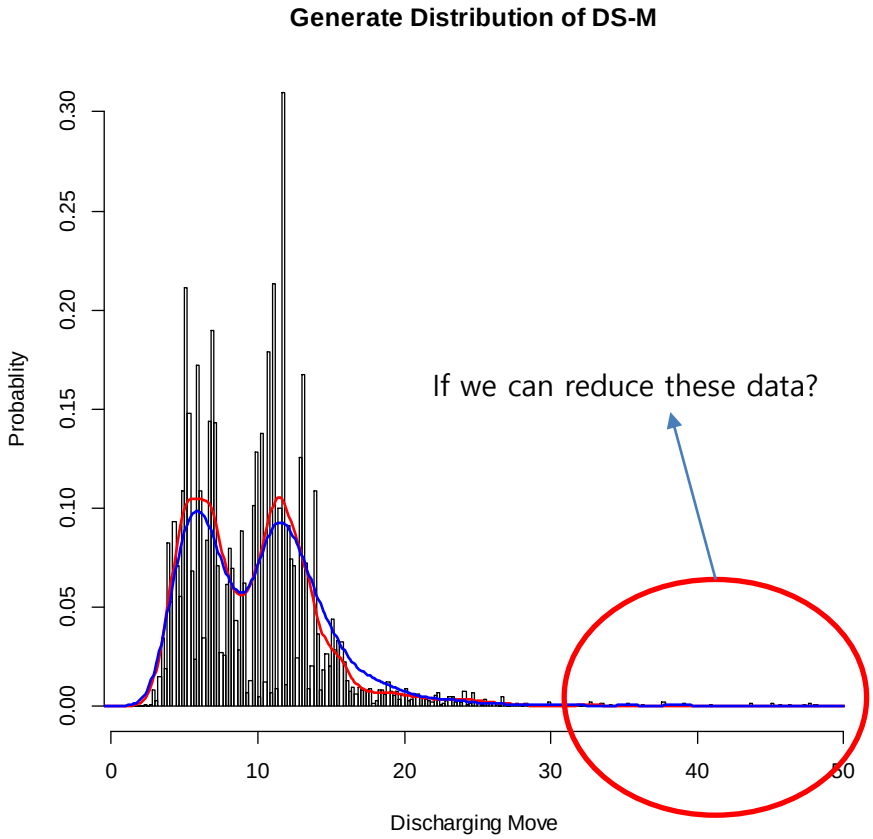
Parameter of Loading Move

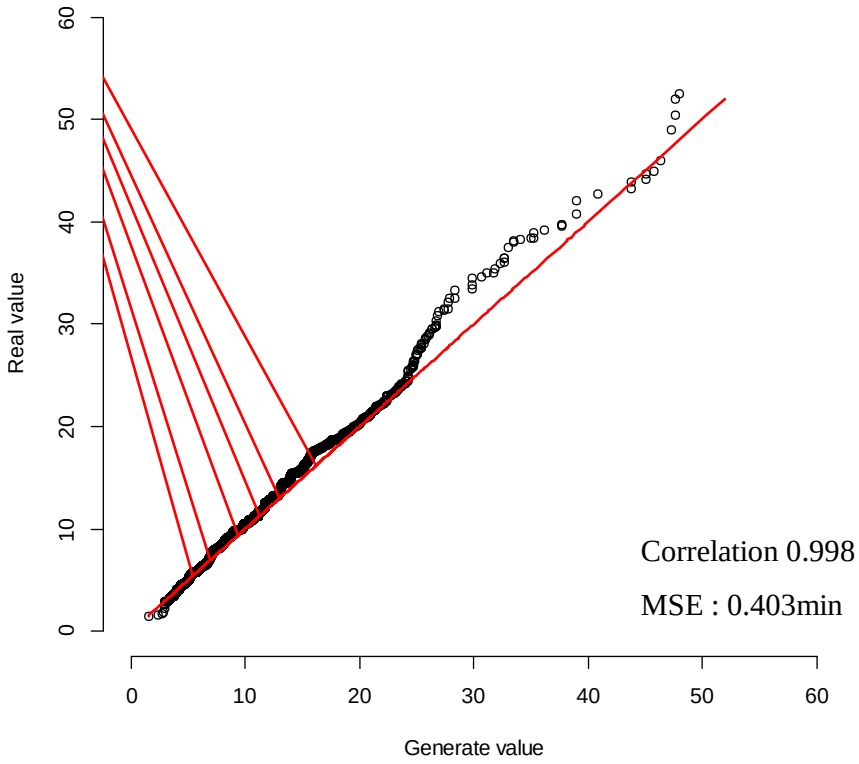
Generate distribution using Metropolis-Hastings Algorithm





What can we do more?

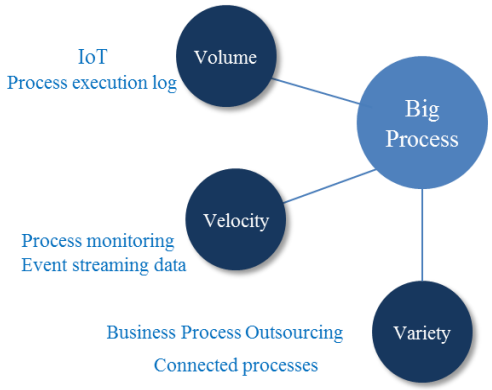




Conclusions

- 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



BSC LAB
Business & Service Computing

BAB
BEST ANALYTICS
OF BIG DATA