01.

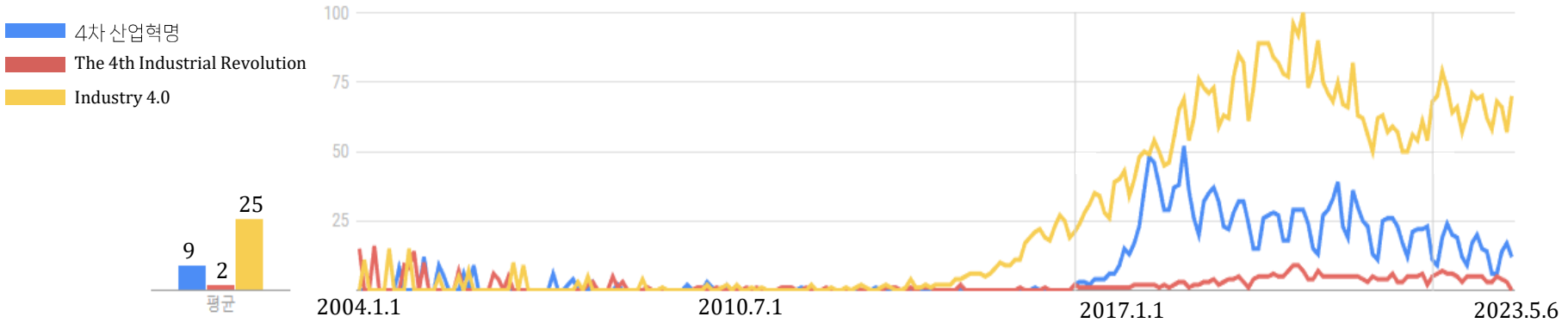
Operational Bigdata

소 속 : 부산대학교 산업공학과
이 름 : 배 혜 림 교수
이메일 : hrbae@pusan.ac.kr

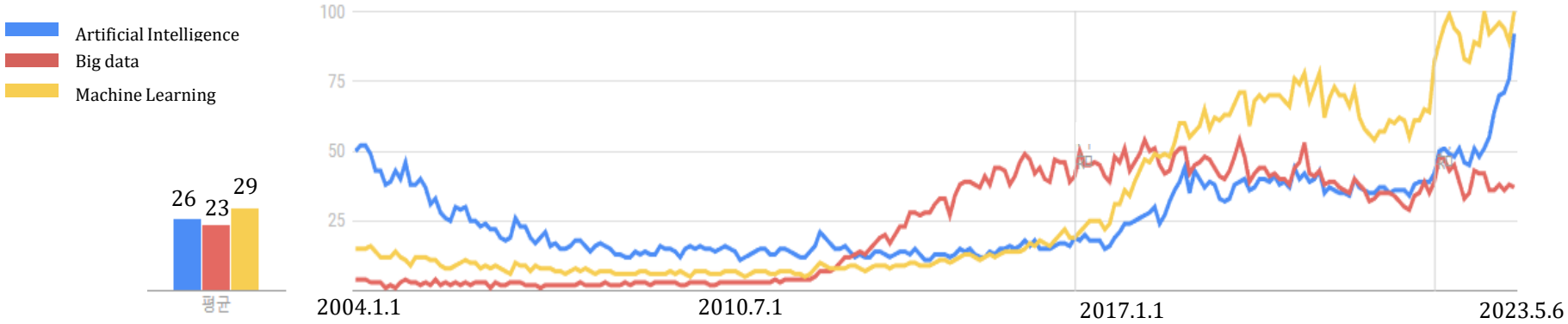
Contents

1.1 Big Data

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



Artificial Intelligence vs. Big data vs. Machine Learning

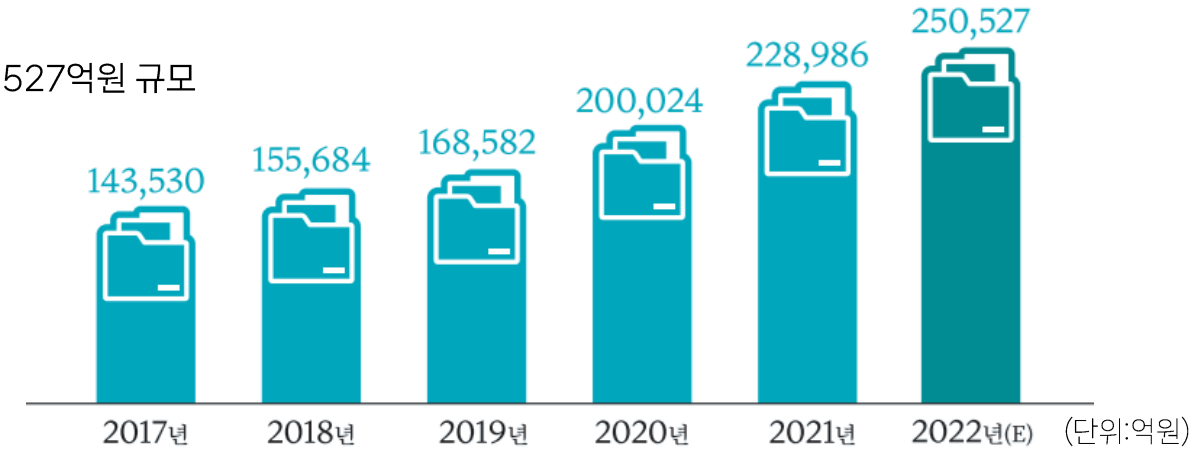


https://trends.google.co.kr/trends/explore?hl=ko

1.1 Big Data

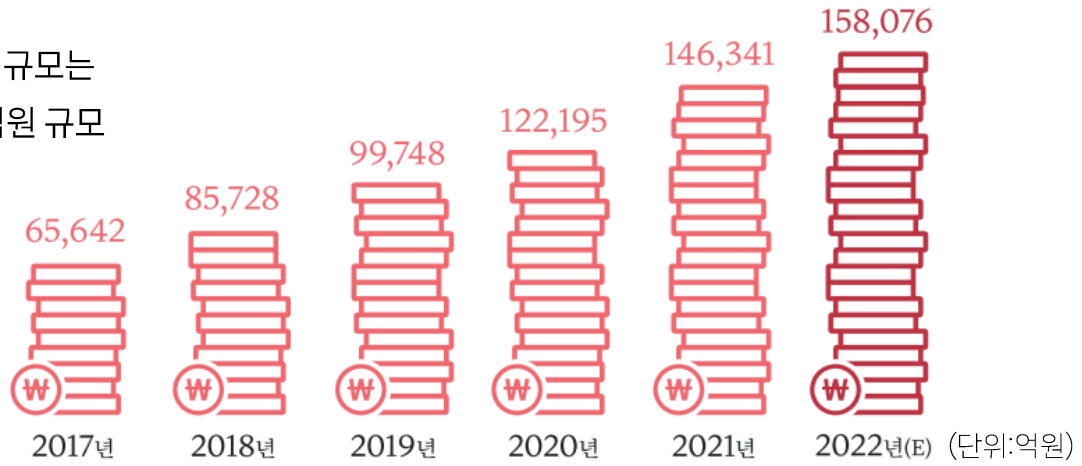
데이터산업 시장 규모

2022년 기준 데이터산업 시장은
2021년 대비 9.4% 성장한 25조 527억원 규모



데이터산업 직접매출 규모

2022년 기준 데이터산업 시장의 직접매출 규모는
2021년 대비 8.0% 성장한 15조 8,076억원 규모

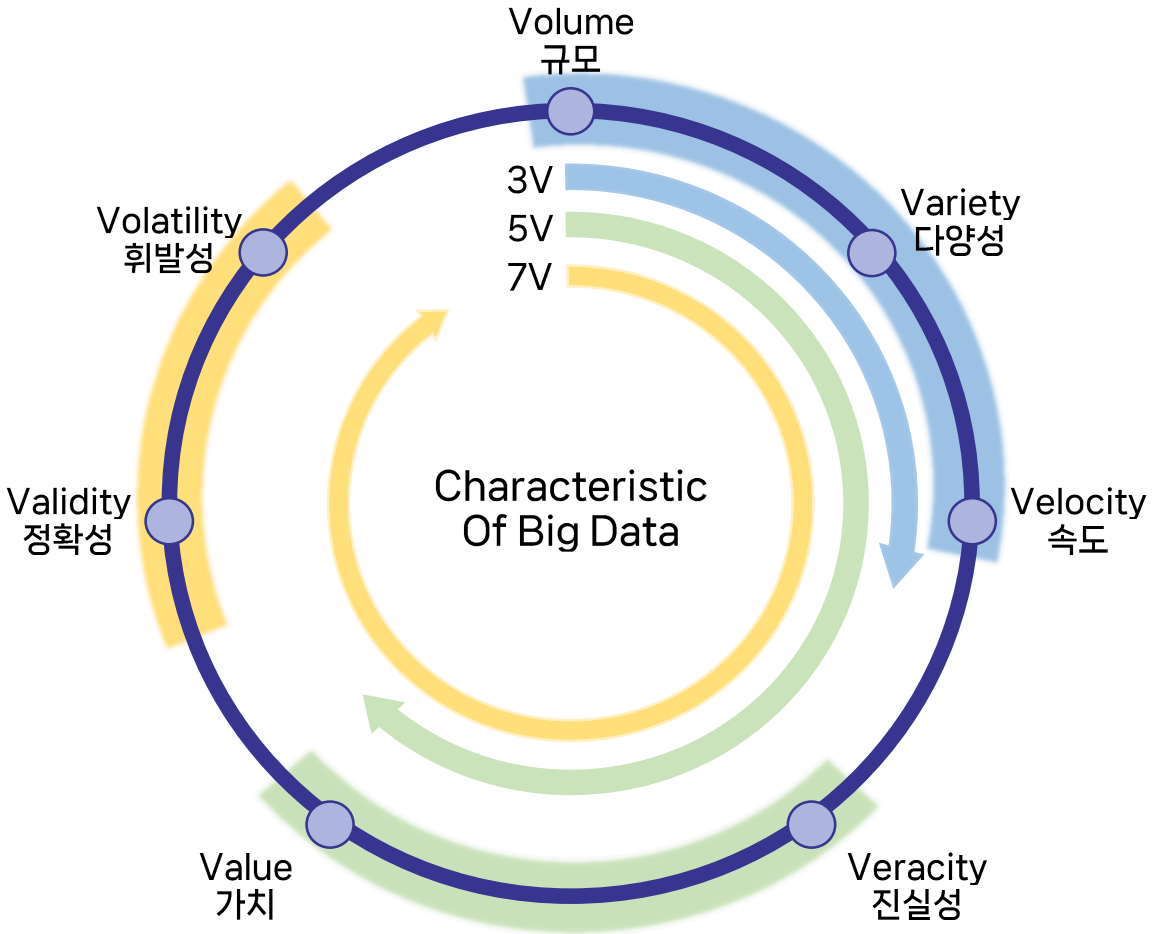


https://www.kdata.or.kr/kr/board/info_01/boardView.do?pageIndex=1&bbsIdx=33688&searchCondition=all&searchKeyword=

1.1 Big Data

What is Big data and 3V ,5V, 7V?

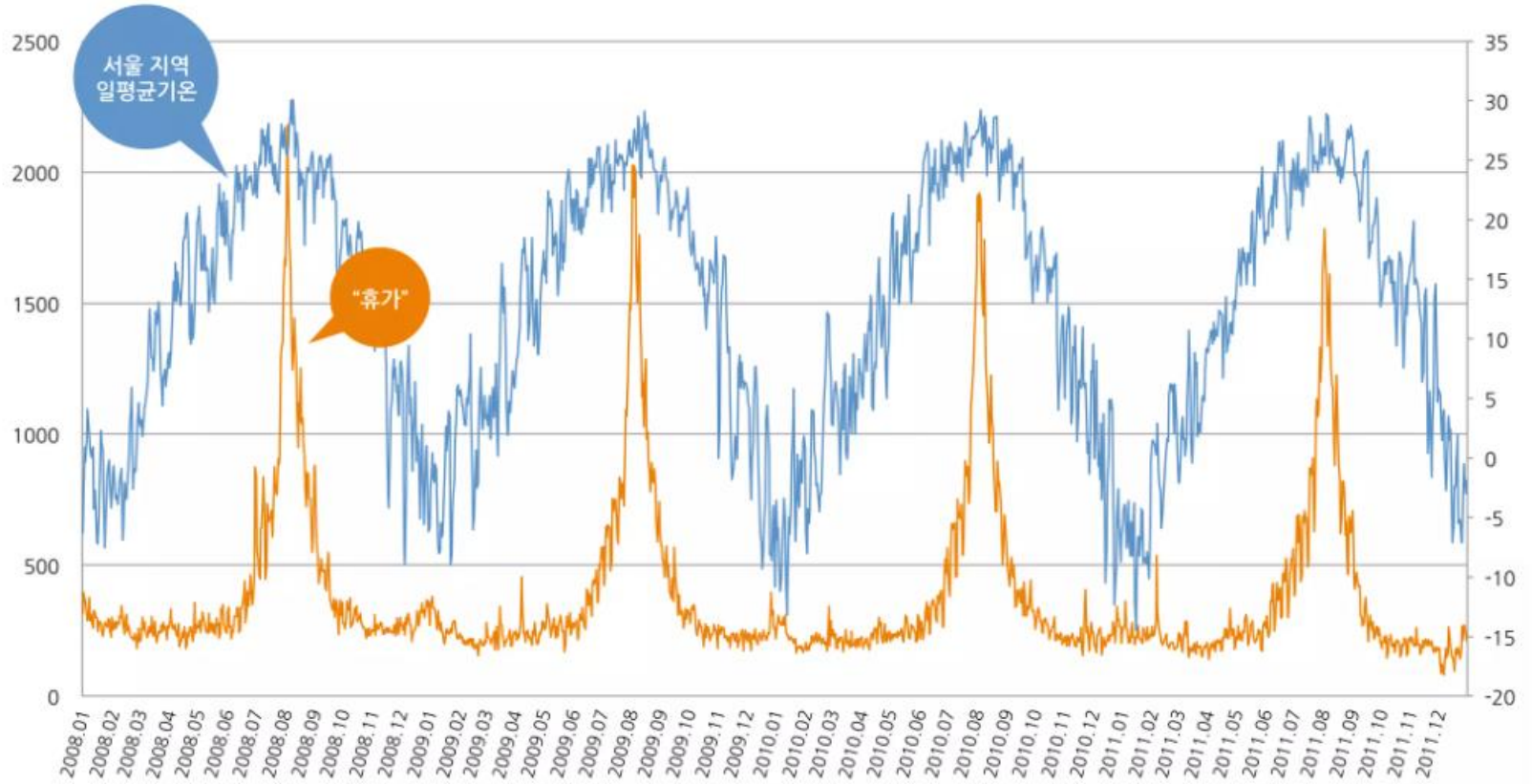
기존 데이터베이스를 넘어서는 대량 데이터, 대량 데이터를 추출하고 결과를 분석하는 기술



1.1 Big Data

Example of Big Data Utilization

① 휴가와 기온의 상관관계



(이하 출처 : 송길영, "여기에 당신의 욕망이 보인다")

1.1 Big Data

Example of Big Data Utilization

① 휴가와 기온의 상관관계

F-test 결과 유의수준 0.05 이하에서 차이가 있음.

(이하 출처 : 송길영, "여기에 당신의 욕망이 보인다")

1.1 Big Data

Example of Big Data Utilization

② 약 장수

"찢어지지 않은 모든 상처에"

1.1 Big Data

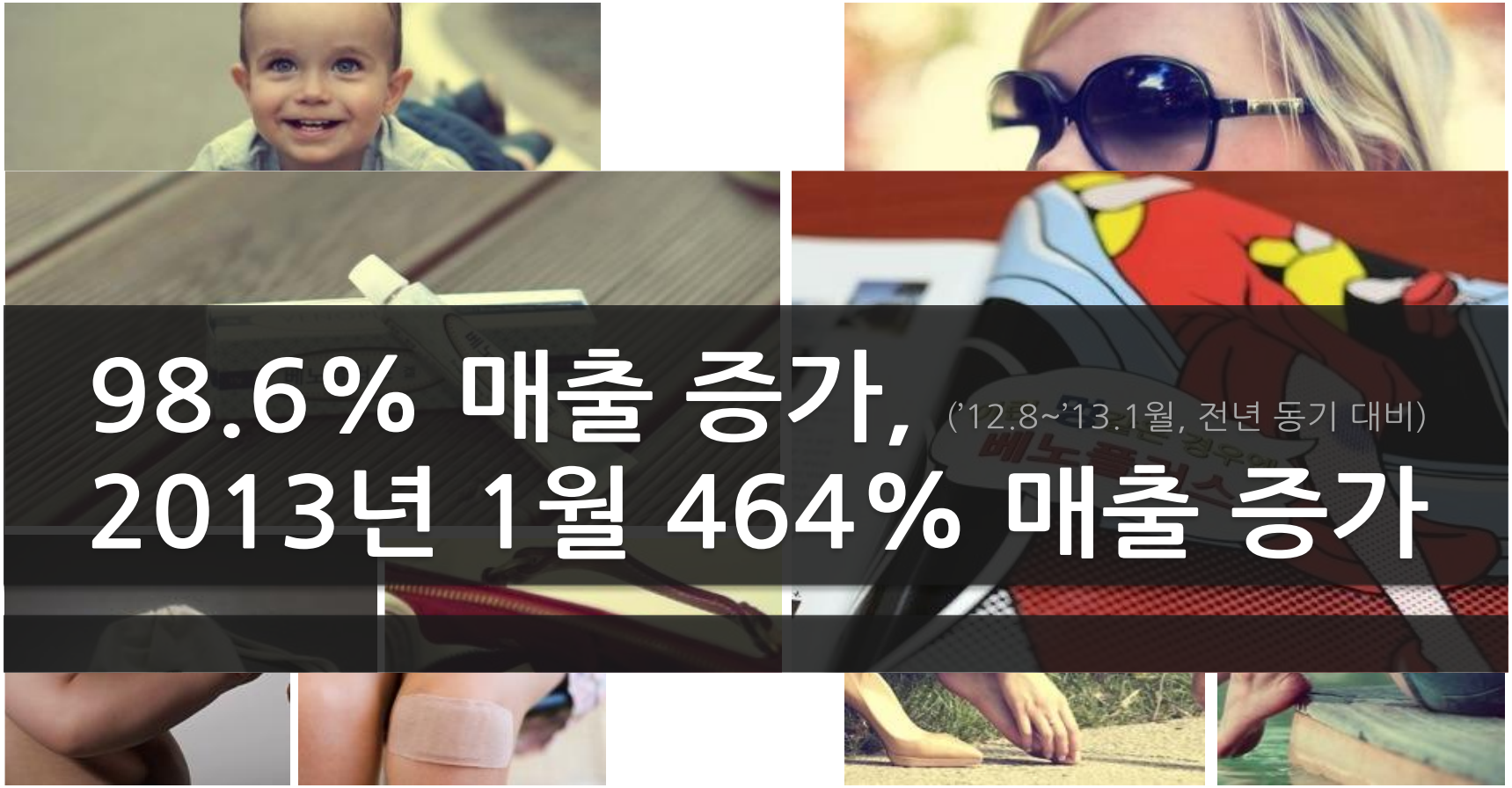
Example of Big Data Utilization

② 약 장수

1.1 Big Data

Example of Big Data Utilization

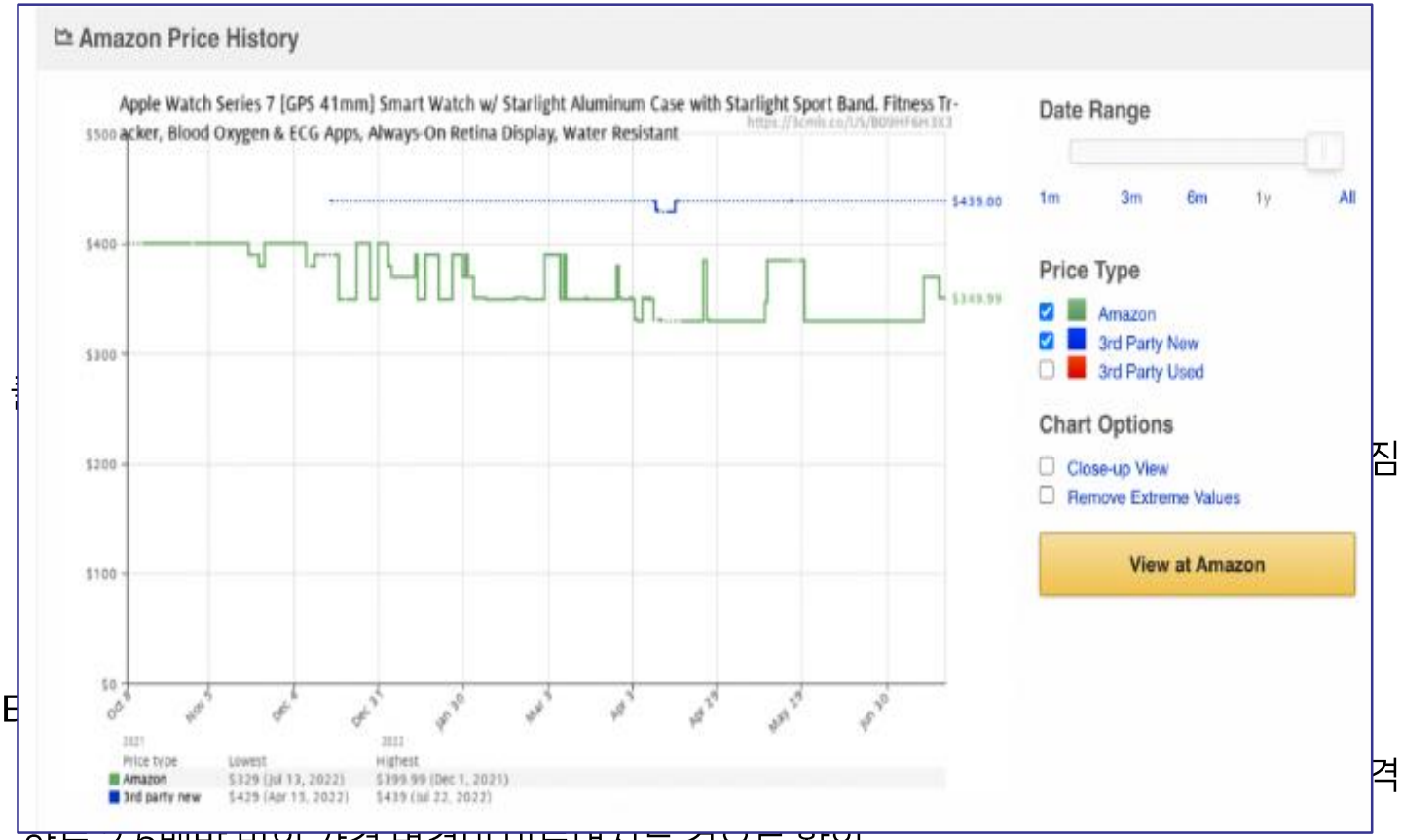
② 약 장수



1.1 Big Data

Example of Big Data Utilization

③ Amazon

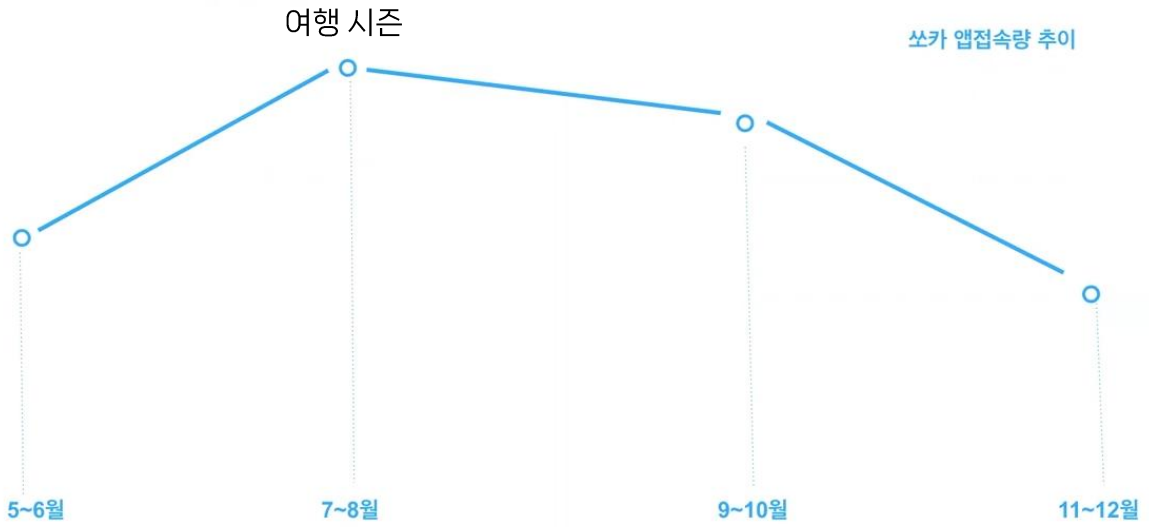


- 고객
  - → 짐을 확인
  - →
  - →
- 데이터
  - → 격 책정
  - → 하루 2.5백만 번의 가격 변경이 이루어지는 것으로 확인
  - → 매출 25% 증가

1.1 Big Data

Example of Big Data Utilization

④ 쏘카(SOCAR)



여행시즌이 끝난 후 놀고 있는 차량이 발생

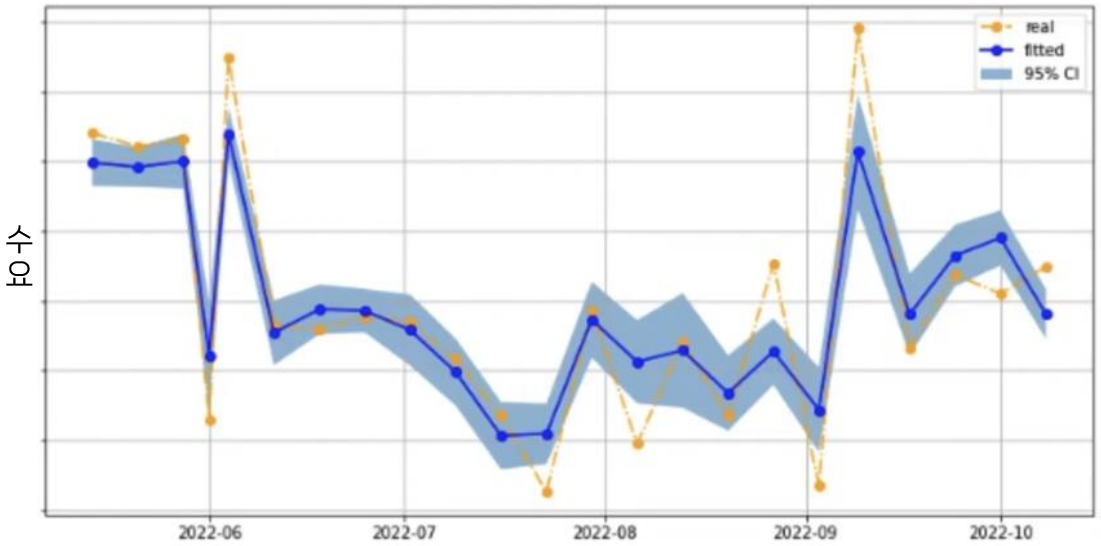
수요가 적은 시 → 할인
수요가 많은 시기 → 할증

https://www.youtube.com/watch?v=3Iy6lyzN71w&t=906s

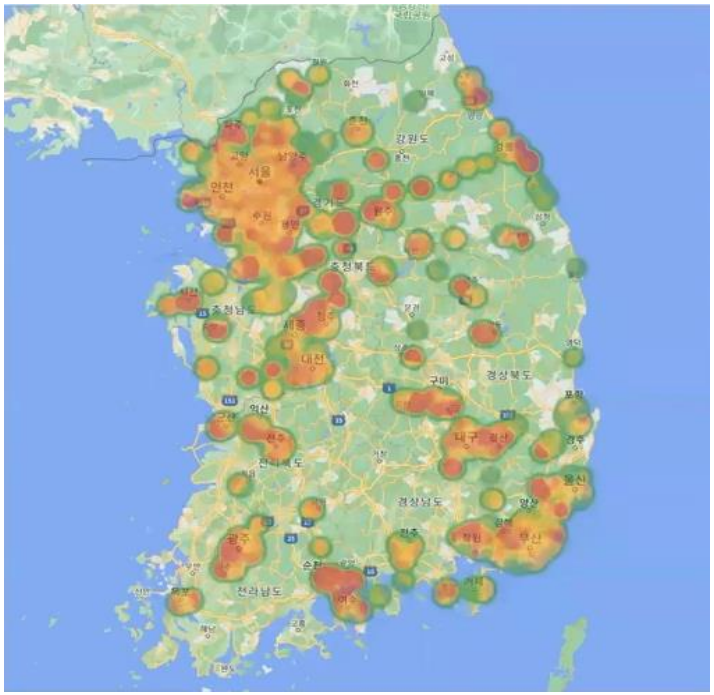
1.1 Big Data

Example of Big Data Utilization

④ 쏘카(SOCAR)



시기별 수요 예측



지역별 수요 예측

시기와 지역에 따른 수요예측을 통해 가격 최적화

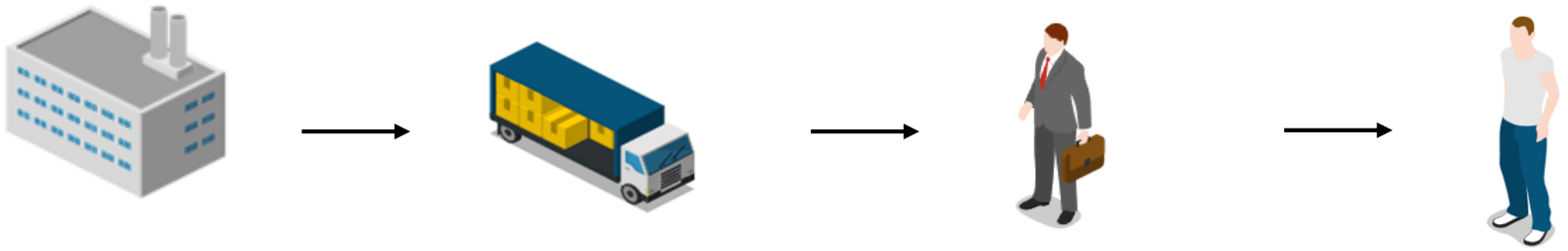
연 총 매출 30억 증대 추정

https://www.youtube.com/watch?v=3Iy6lyzN71w&t=906s

1.1 Big Data

Example of Big Data Utilization

⑤ Another usable area

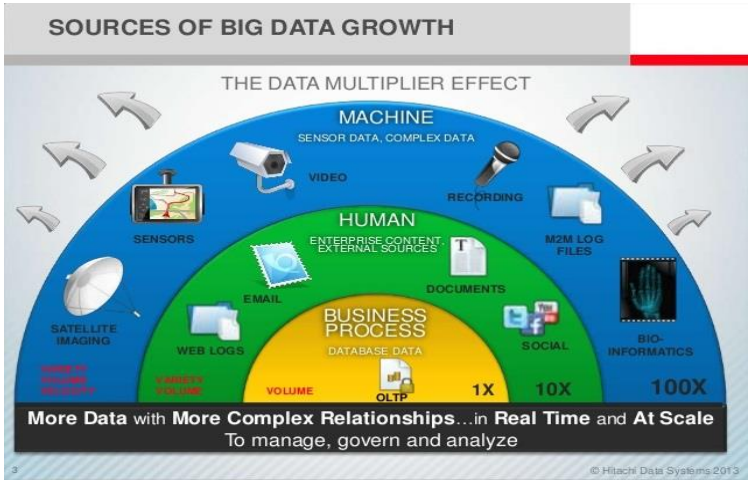


Manufacturer	Distributor	Marketing & Sales	Customer
• 제품 설계 개선 • 생산 라인의 효율성 개선 • 인공지능 기술 개발	• 유통망 관리	• 마케팅 및 판매활동 개선	• 고객 경험 개선 • 제품 수명주기 관리 • 예측 유지보수

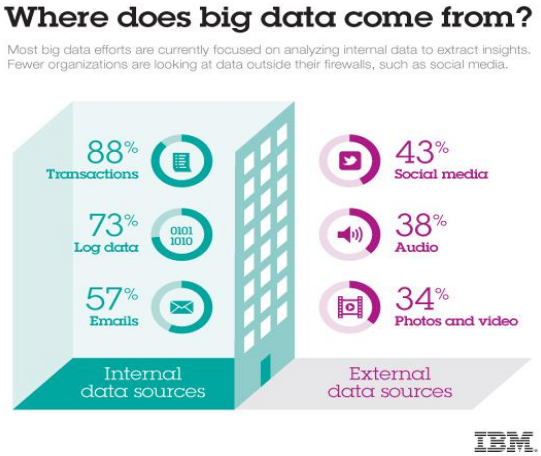
일반적인 제조업에서 고객에게 물건이 전달되는 과정을 예로 살펴보면,
Big Data 분석은 **공정부터 고객에게 이어지는 공급사슬망의**
전반적인 부분에 기여 가능함

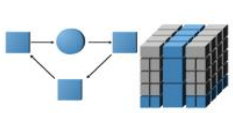
1.1 Big Data

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

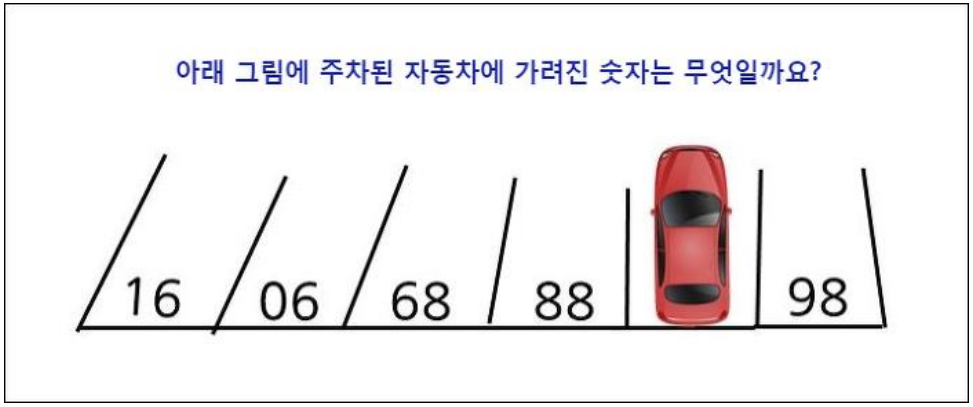
Machine Data

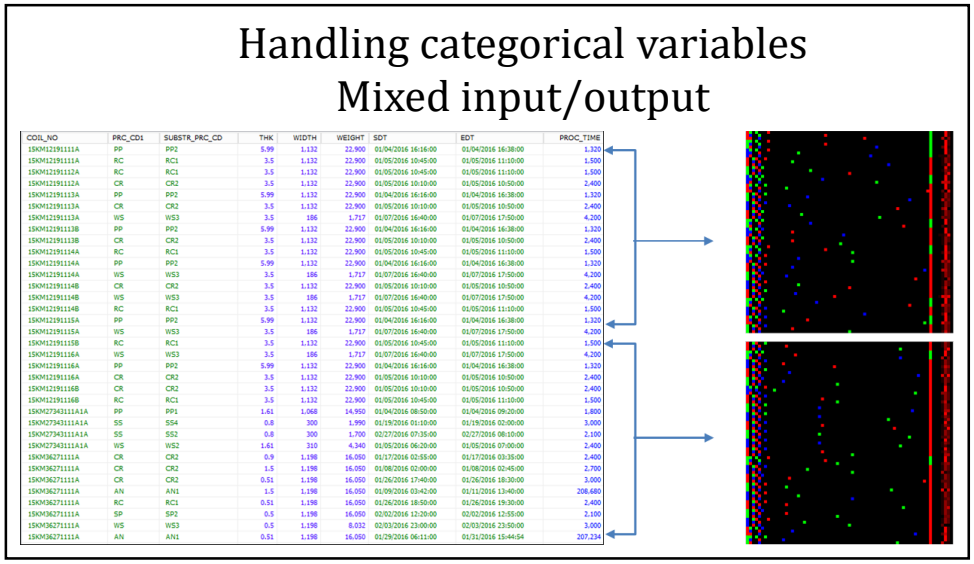
- Time series unstructured data, no predefined schema
- Generated by all IT systems, highly diverse formats
- Massive volume; fast navigation and correlation paramount

1.2 Data Mining & Process Mining

Type of Data

- 수치형, 범주형
  - (234, 0.327, …), (토요일, 맑음, 배혜림)
- 연속형, 이산형
  - (234, 0.327, …), (0, 1)
- 정형, 비정형
  - (Table, Vector, List), (이미지, 음성, 문서)
- 균형, 불균형
  - 제조 불량품 탐지(정상/불량)
  - 신용 카드 이상 거래 탐지
- 기계는 모든 유형의 **data**를 받아 들일 수 있을까요?





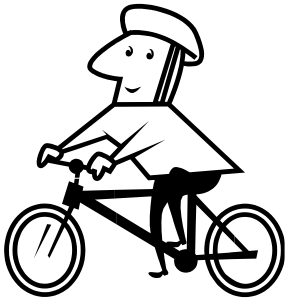
1.2 Data Mining & Process Mining

What is Learning?

Learning

- 어떤 객체로 하여금 가르침이나 경험을 통해 자신의 능력을 향상시킬 수 있도록 하는 일련의 과정
- 지적행위의 가장 기초적인 개념
- Example)
  - Simple association task
  - 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

1.2 Data Mining & Process Mining

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

1.2 Data Mining & Process Mining

What is Machine Learning?

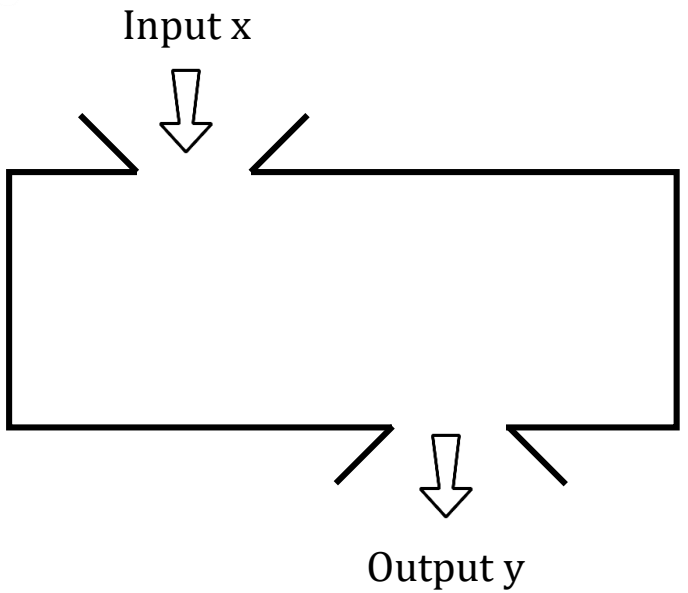
Finding ‘f’

$$Y = f(X)$$

rule

pattern

knowledge



1.2 Data Mining & Process Mining

Data Mining

1. What is Data Mining?

대규모 데이터에서 존재하는 관계, 패턴, 규칙, 인사이트 등 탐색하고 모형화함으로써 **유용한 정보**를 추출하는 과정

2. Why use Data Mining

경영의사결정 지원 및 수익을 증가시키며 시장 동향을 예측하는 등 다양한 이점을 제공

3. Data Mining Techniques

- 회귀 모형(Regression Analysis)
- 모형 평가(Model Assessment)
- 의사결정나무(Decision Tree)
- 군집분석(Cluster Analysis)
- etc.

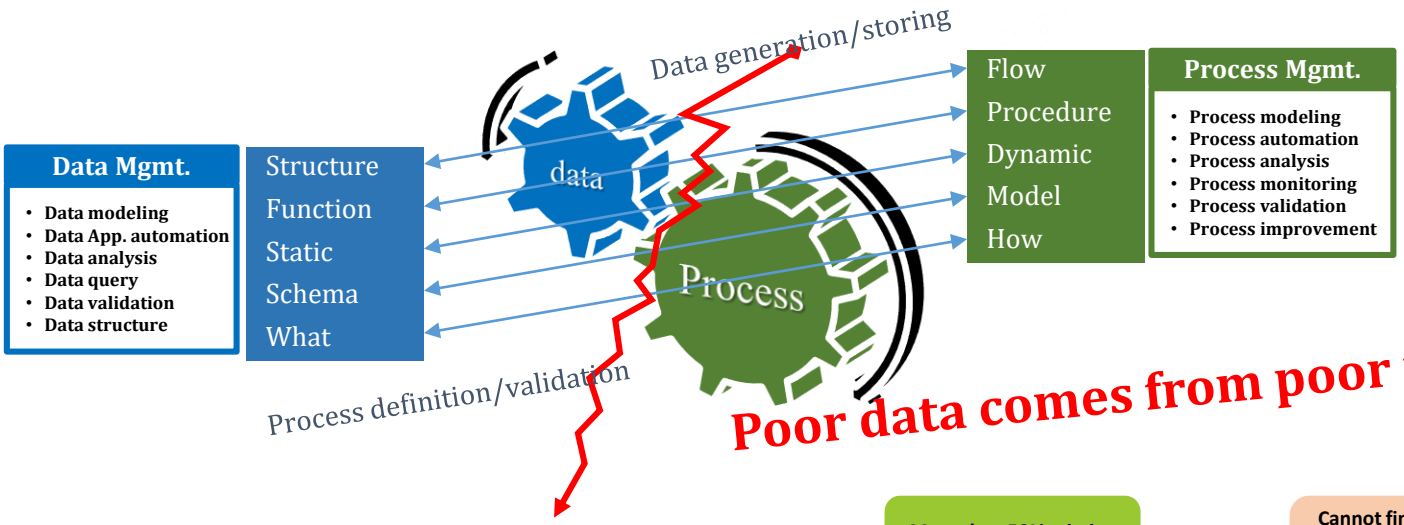
Differences Between Data Mining and Machine Learning

- 데이터 마이닝은 데이터로부터 알려지지 않은 특성(unknown properties)을 '발견'하는 데 초점
- 머신러닝은 학습한 알려진 특성(known properties)을 통해 어떤 '예측'을 하는 데 초점

https://dataonair.or.kr/db-tech-reference/d-guide/data-practical/?mod=document&uid=409

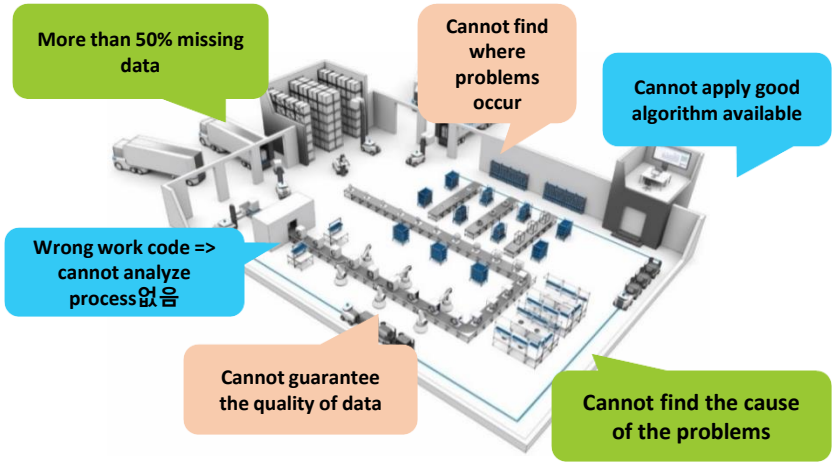
1.2 Data Mining & Process Mining

Big-data vs. process improvement



Poor data comes from poor process

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**



1.2 Data Mining & Process Mining

Difference between Data Mining & Process Mining

- **Field of Application**

  프로세스 마이닝의 데이터 마이닝의 한 종류이지만 프로세스 그 자체에 집중

- 제조 공정에서의 예시

  - **Data Mining(≈M/L)**

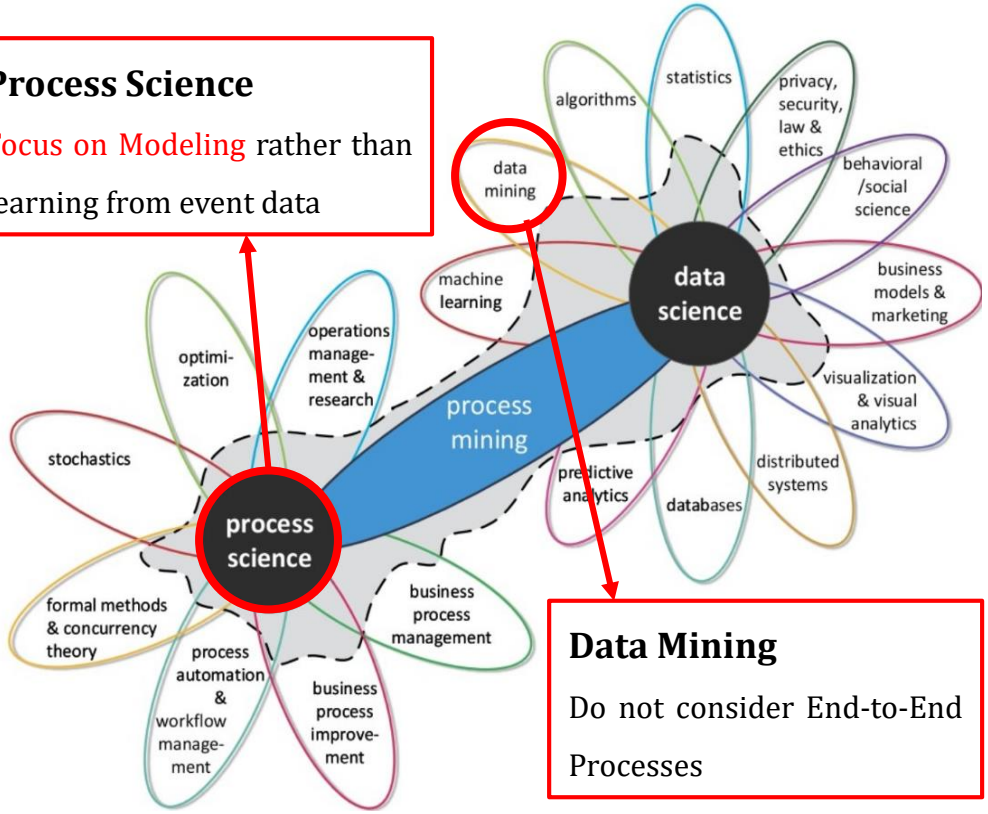
    제조 공정에서 어떤 결과물이 나왔는지, 어떤 결과물이 나올 것인지, 이러한 결과물이 왜 나왔는지에 대해 진행

  - **Process Mining**

    제조 공정 전체를 분석하고 시각화하여 제조 공정 자체를 효율적으로 개선하고 발전시킬 수 있는지에 대해 진행

Process Science

Focus on Modeling rather than learning from event data



Data Mining

Do not consider End-to-End Processes

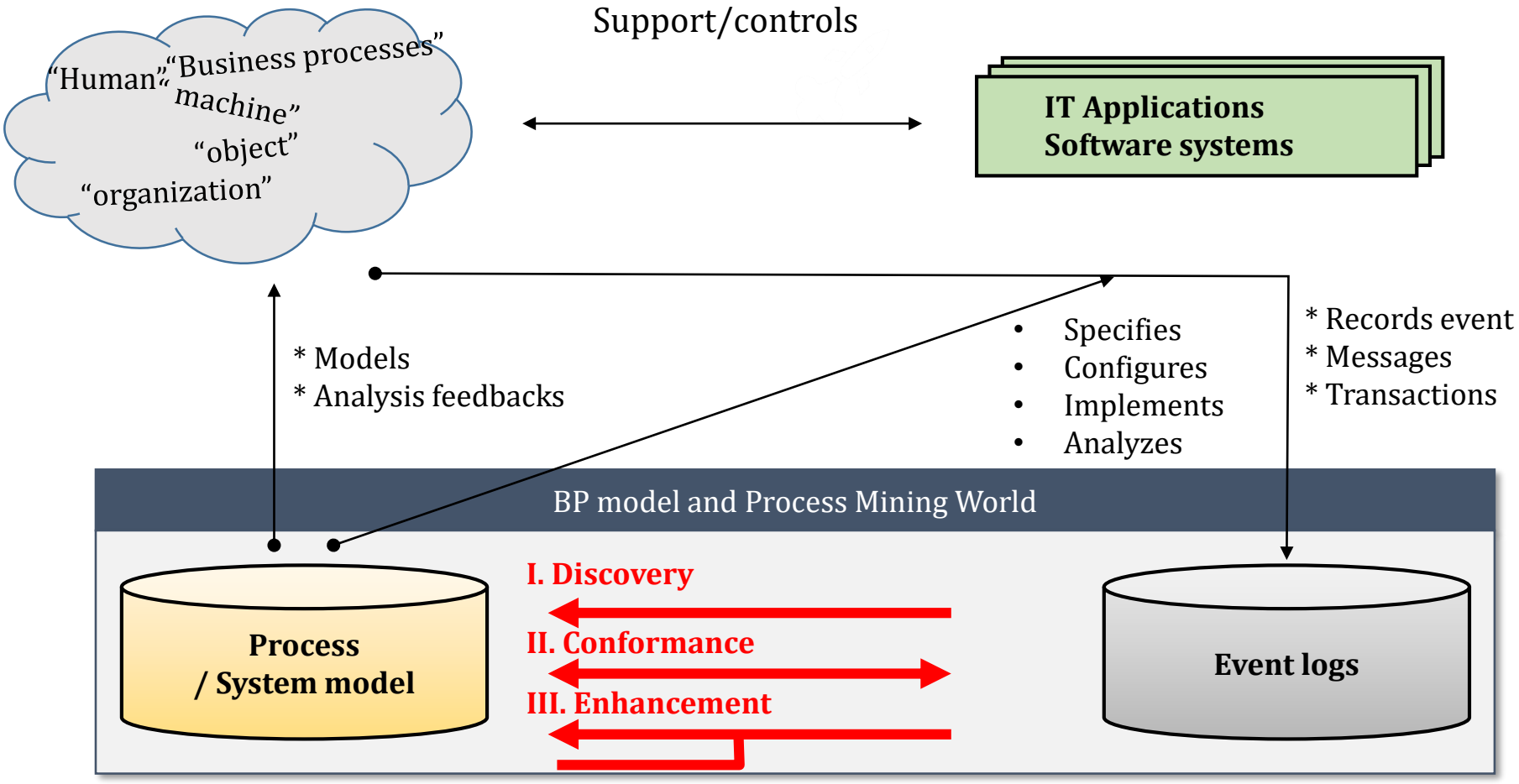
프로세스 마이닝은 모든 도메인에서 적용이 가능하며

Process Science와 Data Science의 연결고리이다

Process Mining : Data Science in Action (Second Edition), Springer 2016

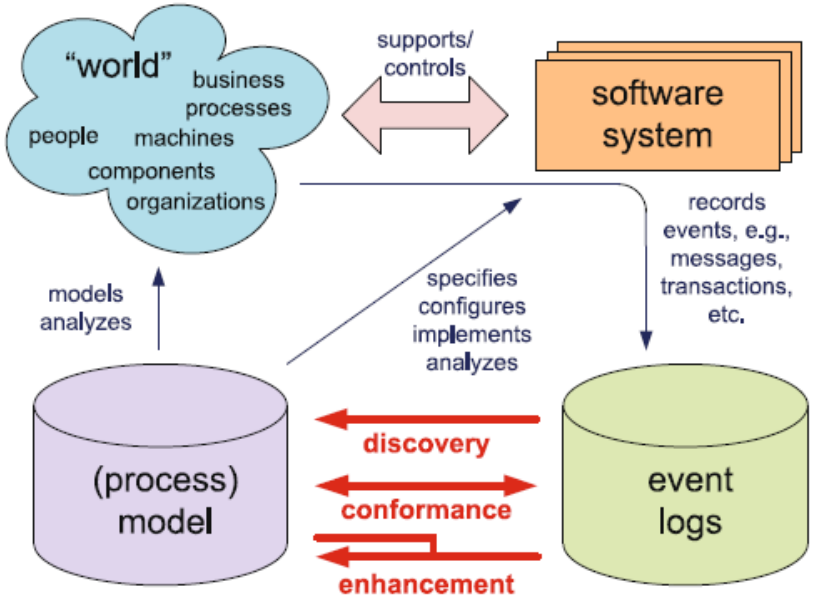
1.3 Concepts and Benefits of Process Mining

Process Analytics

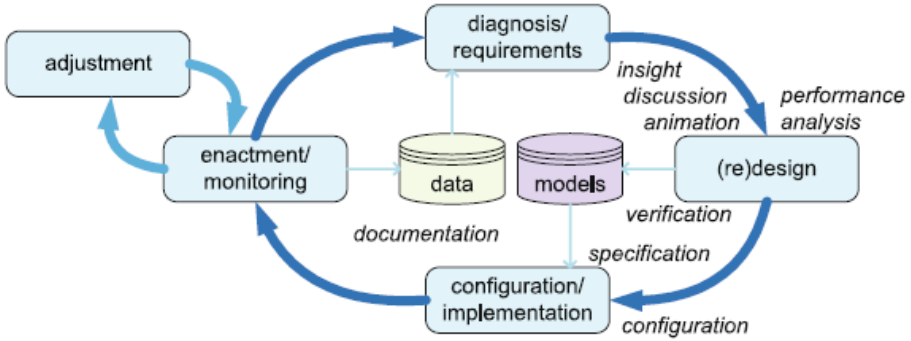


1.3 Concepts and Benefits of Process Mining

Three main types of process mining



PM Life Cycle



1.3 Concepts and Benefits of Process Mining

Advantage of Process Mining

1.Understanding

For better
understanding
what we are do

2.Finding

Finding cause
and fixing the problem

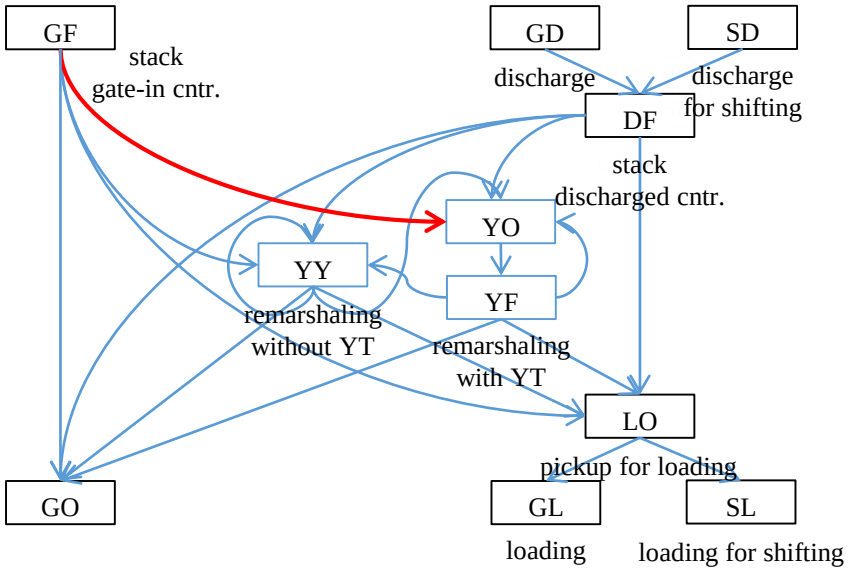
3.Predicting

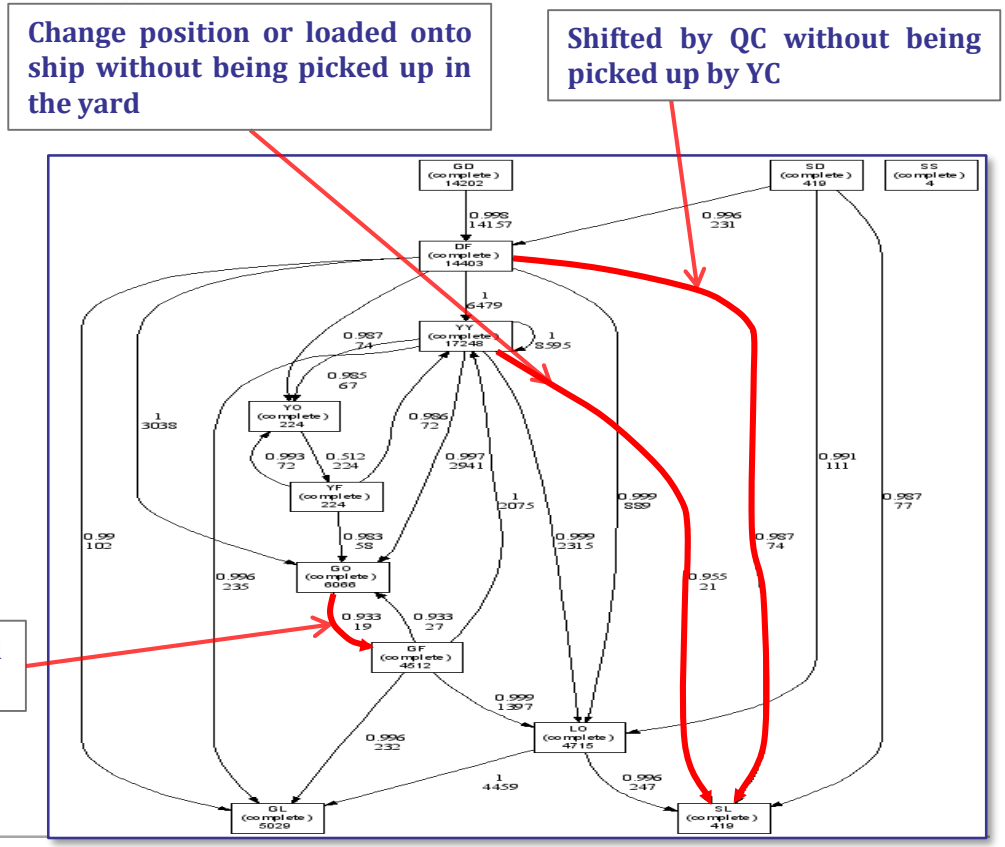
For predicting
future result

1.3 Concepts and Benefits of Process Mining

1. For better understanding what we are do

Pre-defined process model vs. Discovered process model





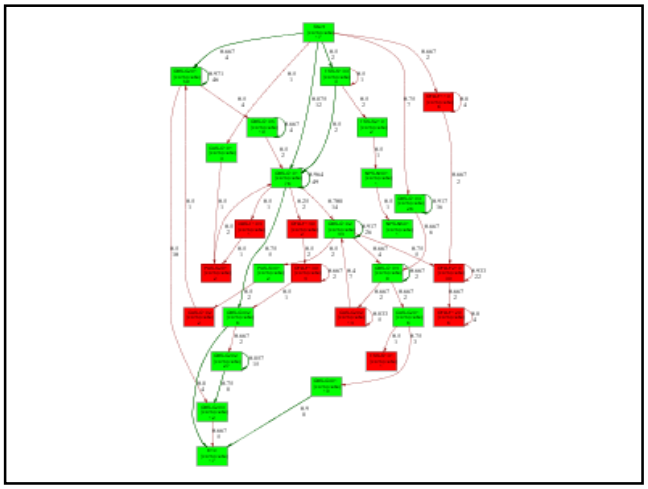
1.3 Concepts and Benefits of Process Mining

1. For better understanding what we are do

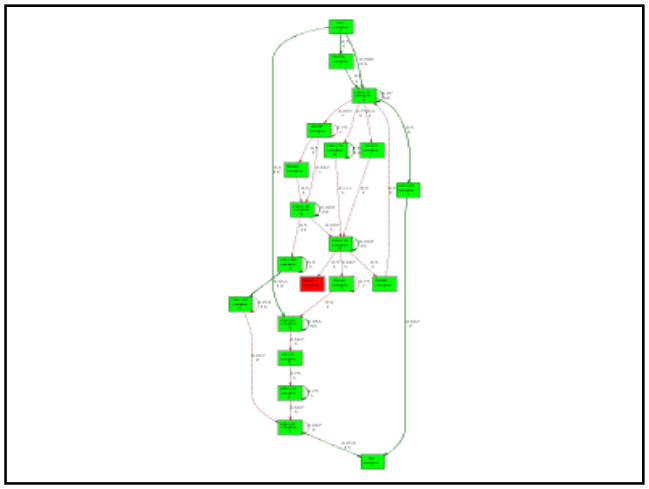
- Plan vs. Actual
- As-Is vs. To-Be
- Peer vs. Peer

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



1.3 Concepts and Benefits of Process Mining

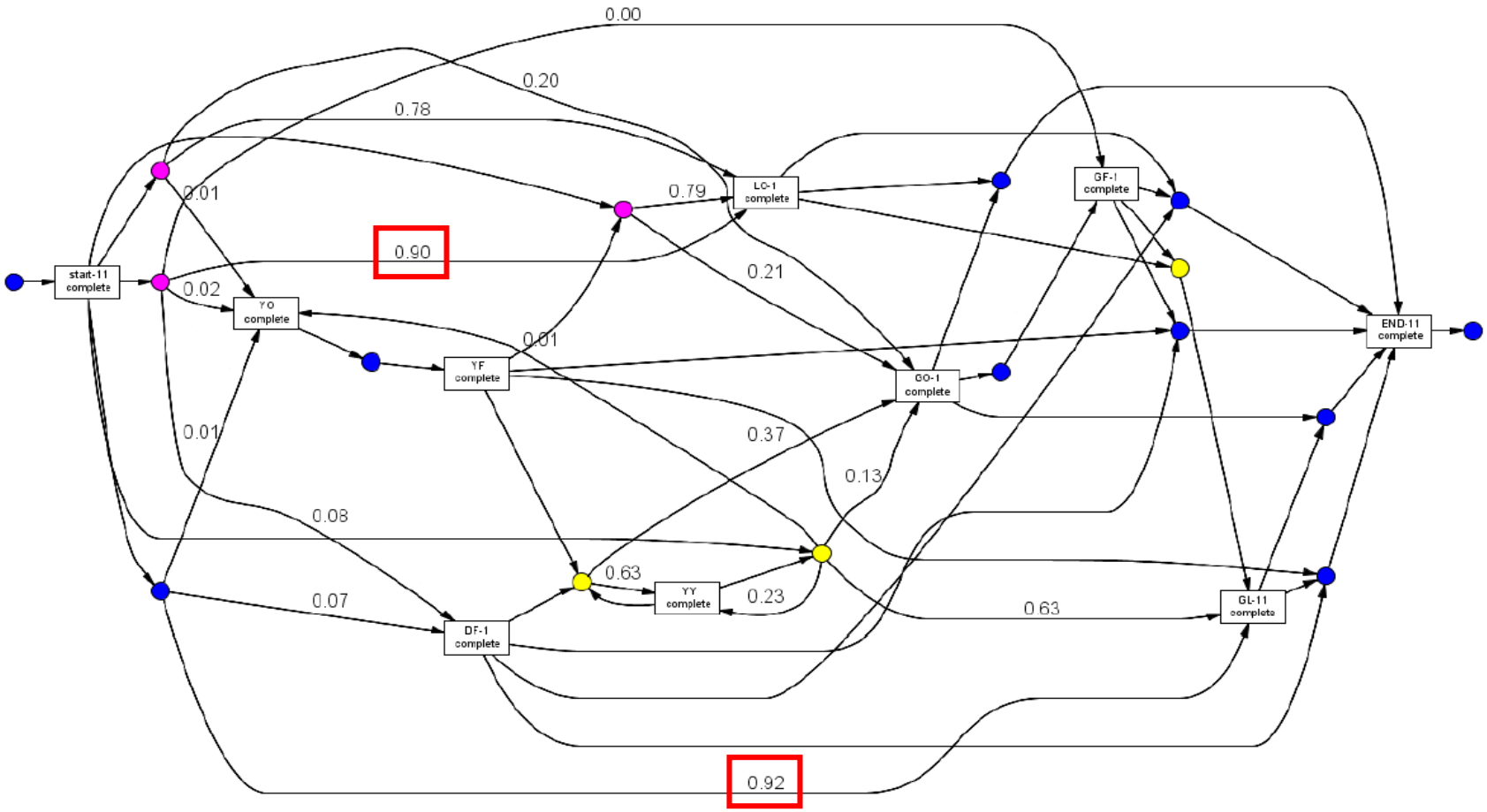
1. For better understanding what we are do

Port example : Better understanding of current situation

1.3 Concepts and Benefits of Process Mining

2. Finding cause and fixing the problem

What is the bottleneck in the port?



1.3 Concepts and Benefits of Process Mining

2. Finding cause and fixing the problem

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

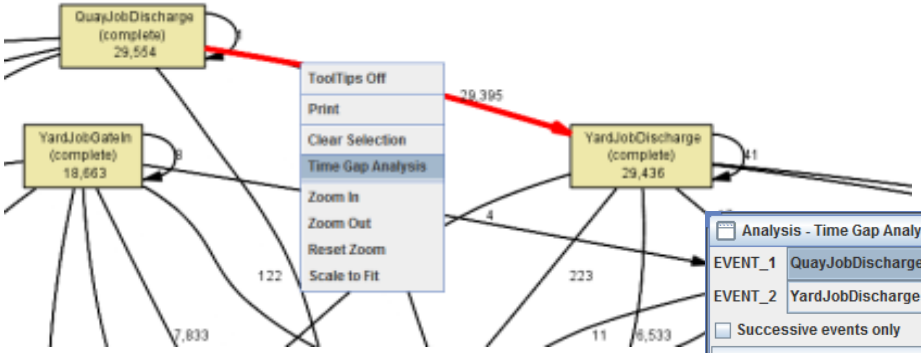
1.3 Concepts and Benefits of Process Mining

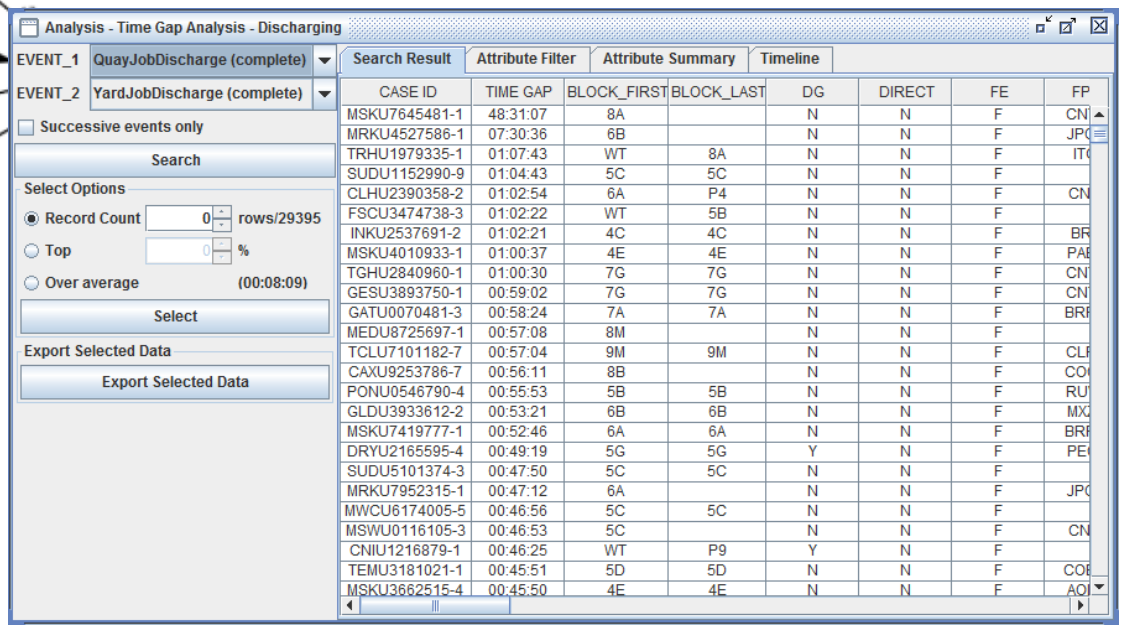
2. Finding cause and fixing the problem

Single dimensional time gap analysis

- Time Gap between two arbitrary nodes

  - Shows time gap of all cases in a decreasing order



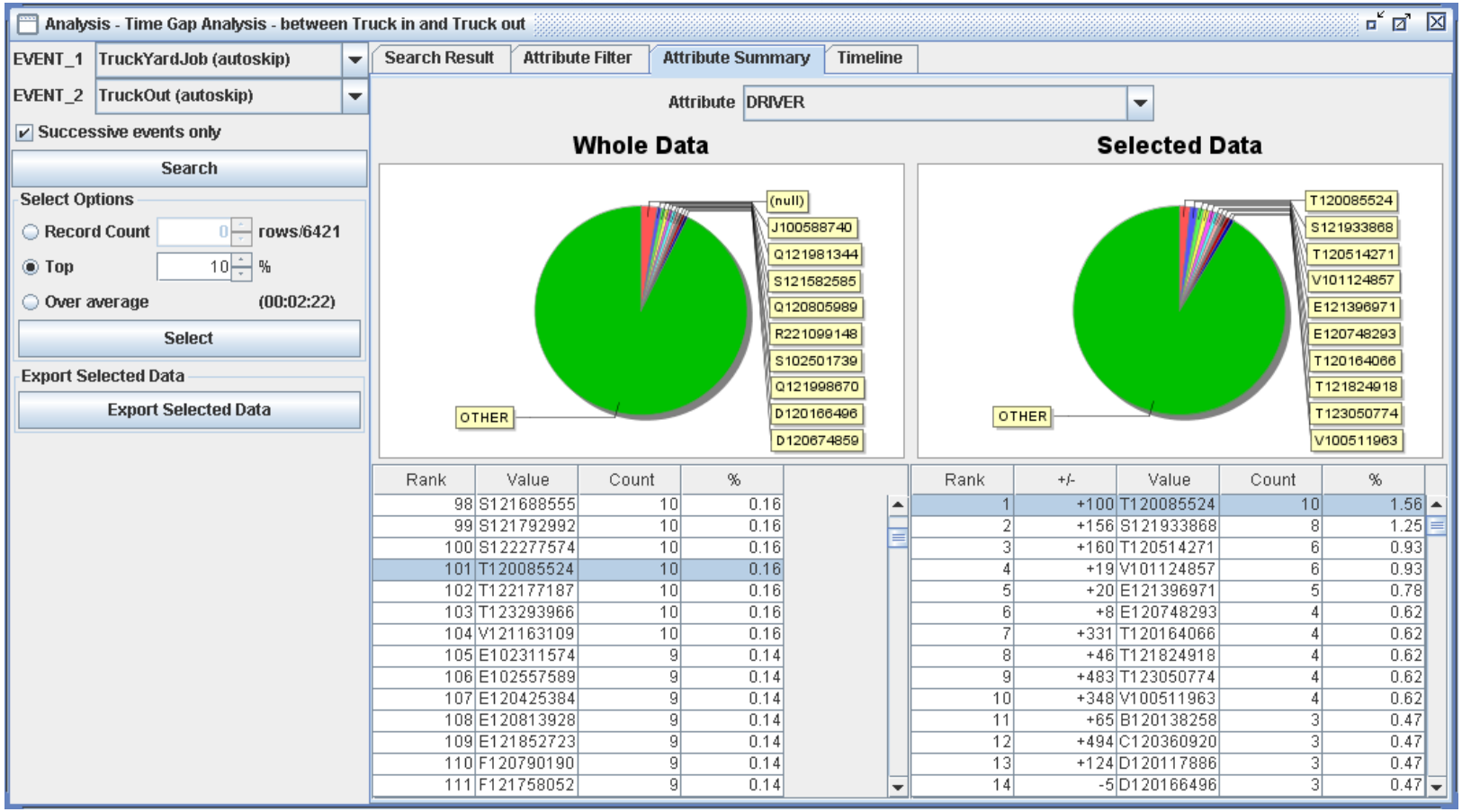


1.3 Concepts and Benefits of Process Mining

2. Finding cause and fixing the problem

Time Gap Analysis → Timeline

- We can know when delayed cases occur in the timeline



Rank	Value	Count	%
98	S121688555	10	0.16
99	S121792992	10	0.16
100	S122277574	10	0.16
101	T120085524	10	0.16
102	T122177187	10	0.16
103	T123293966	10	0.16
104	V121163109	10	0.16
105	E102311574	9	0.14
106	E102557589	9	0.14
107	E120425384	9	0.14
108	E120813928	9	0.14
109	E121852723	9	0.14
110	F120790190	9	0.14
111	F121758052	9	0.14

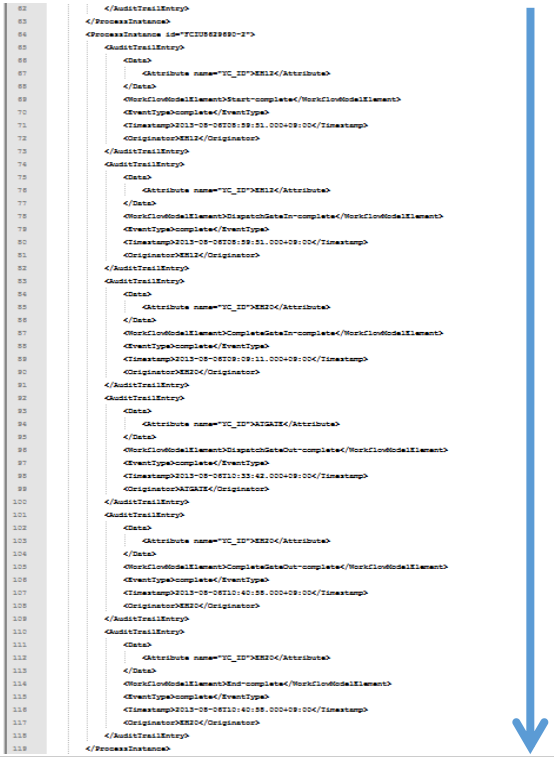
Rank	+/-	Value	Count	%
1	+100	T120085524	10	1.56
2	+156	S121933868	8	1.25
3	+160	T120514271	6	0.93
4	+19	V101124857	6	0.93
5	+20	E121396971	5	0.78
6	+8	E120748293	4	0.62
7	+331	T120164066	4	0.62
8	+46	T121824918	4	0.62
9	+483	T123050774	4	0.62
10	+348	V100511963	4	0.62
11	+65	B120138258	3	0.47
12	+494	C120360920	3	0.47
13	+124	D120117886	3	0.47
14	-5	D120166496	3	0.47

1.3 Concepts and Benefits of Process Mining

3. For predicting future result

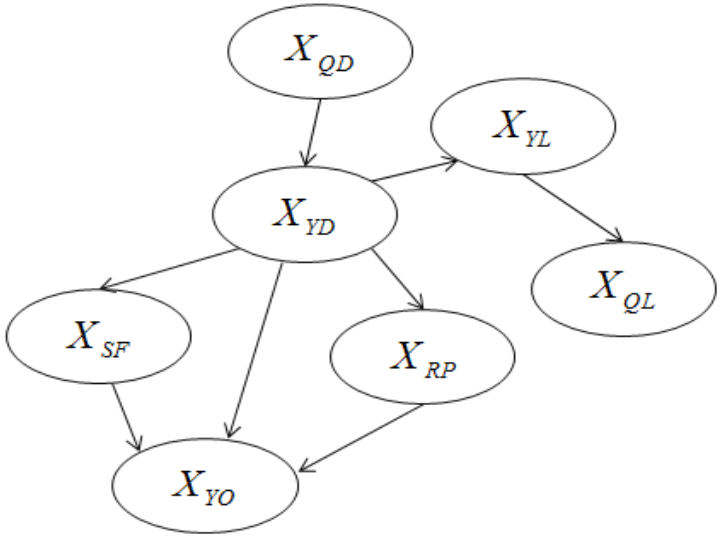
- Bayesian Network (BN)
  - Bayesian network is a useful tool for inference and sensitivity analysis
  - Generating the structure is not an easy task (Chickering et al, 2004)
  - Inference **without** Bayesian network

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



- **If we have Bayesian network?**

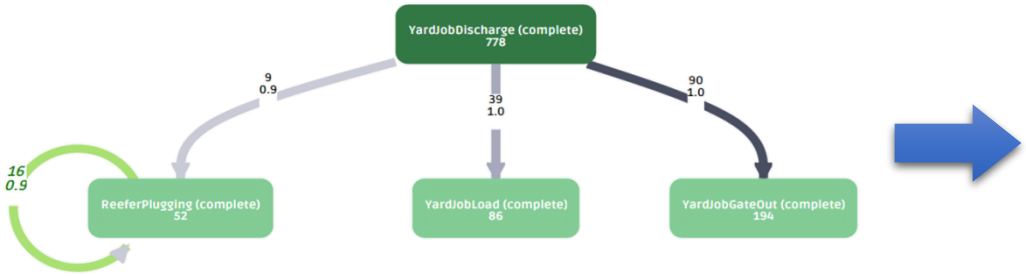
Make Bayesian network **once**, and using
node traversal every time make inference

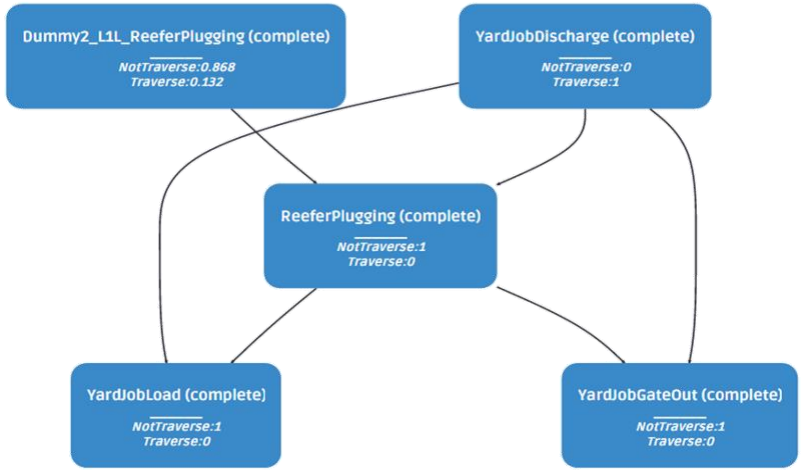


1.3 Concepts and Benefits of Process Mining

3. For predicting future result

- Methodology for 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 Network(BN)
  ① Inference (Causal inference, Prediction)
  ② Sensitivity, analysis (What if simulation)

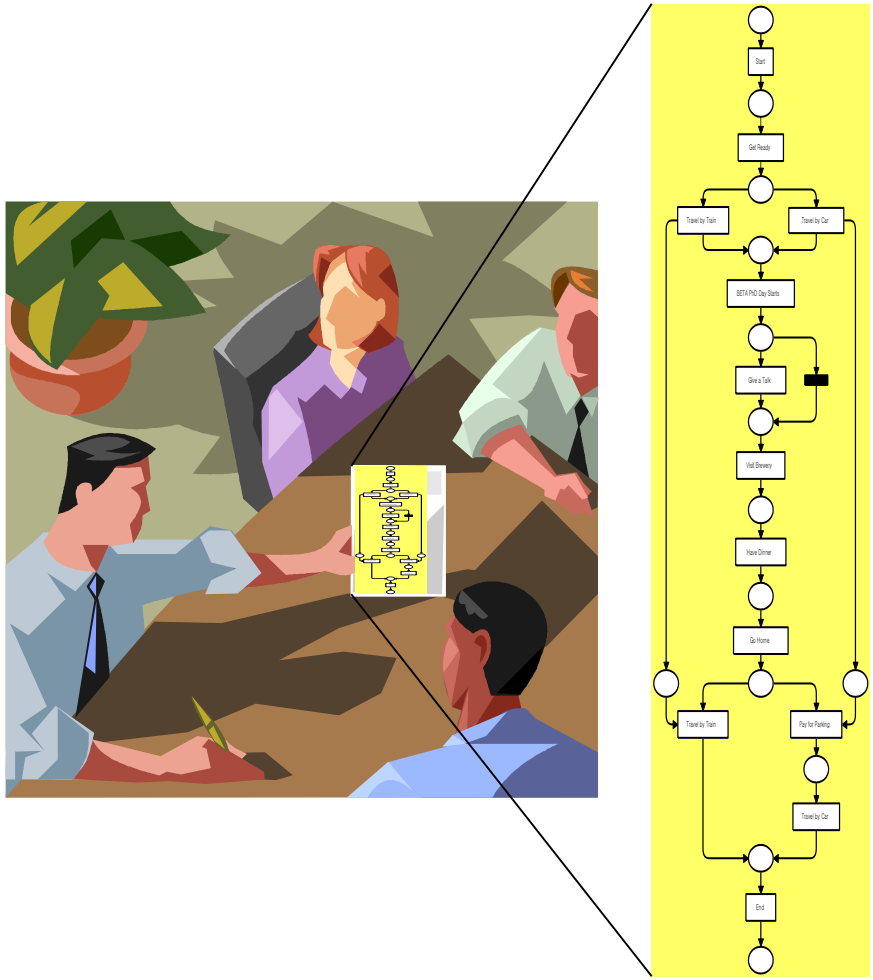




1.3 Concepts and Benefits of Process Mining

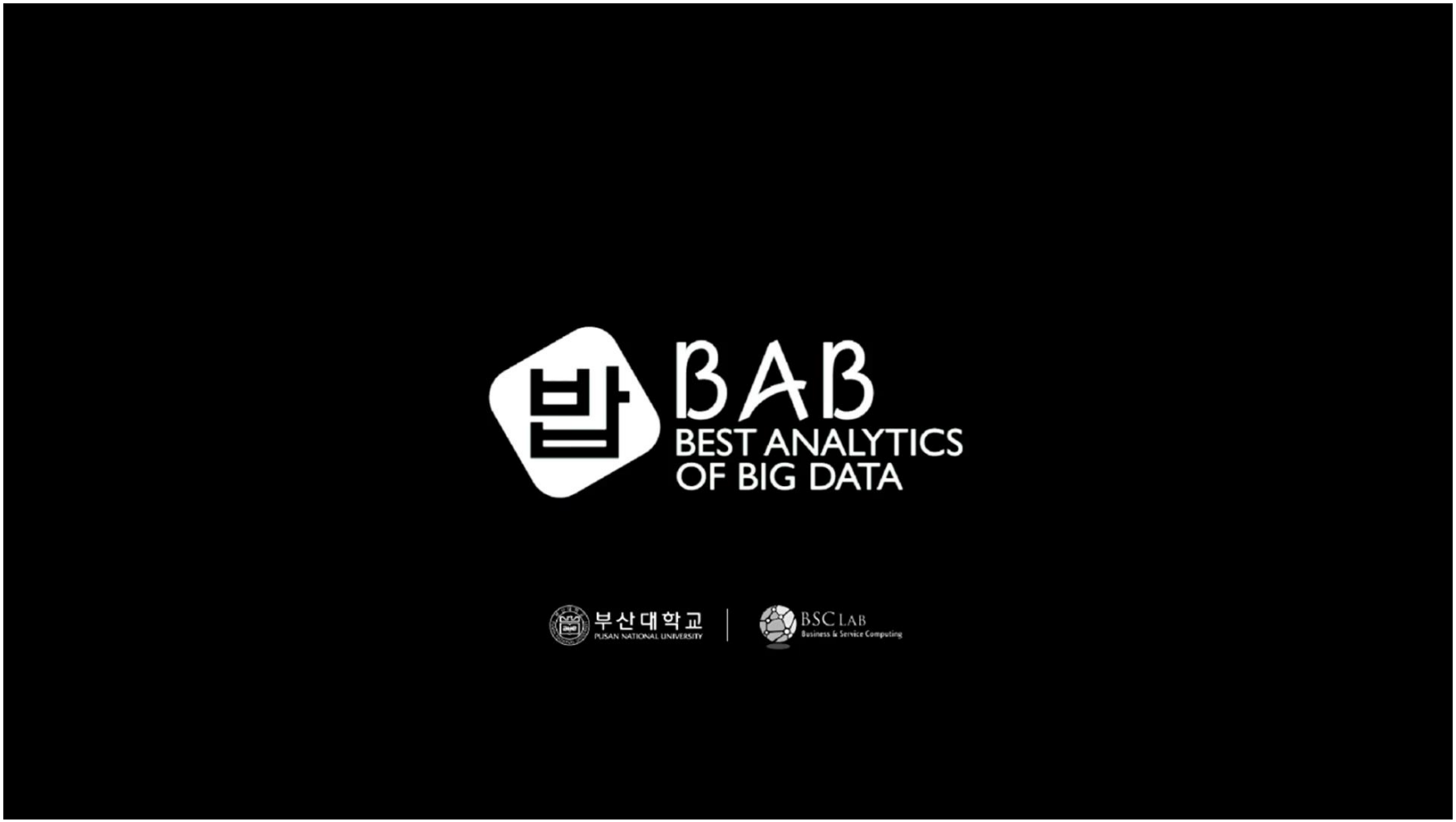
Role of Process Mining

- 프로세스 마이닝 적용 전
  - Objective picture of how the process has been executed
- 프로세스 마이닝 적용 후
  - Feedback mechanism
  - Process Improvement



1.3 Concepts and Benefits of Process Mining

Process Analytics Example



1.3 Concepts and Benefits of Process Mining

Process Mining

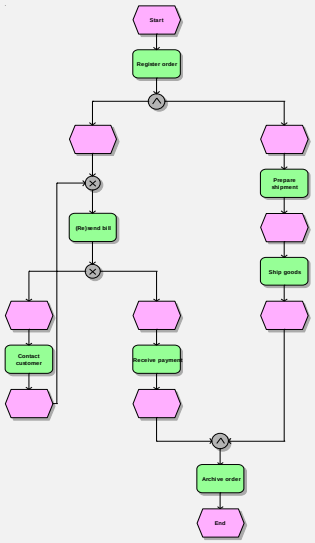
Event Log

Activity	Timestamp	Customer
Create Purchase Order item	15-2-2018@10.30	Kevin
Vendor creates invoice	15-2-2018@10.27	Ann
Vendor creates invoice	15-2-2018@10.49	Jade
Record Goods Receipt	15-2-2018@11.10	Pole

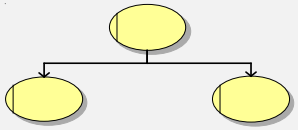
Mining Techniques

Mined Models

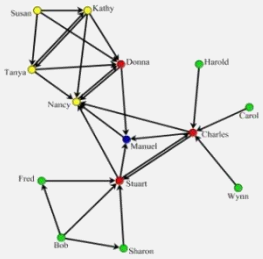
Process Model



Organizational Model



Social Network



Performance Analysis

Auditing/Security

1.4 Data & Model

Data for Process Mining

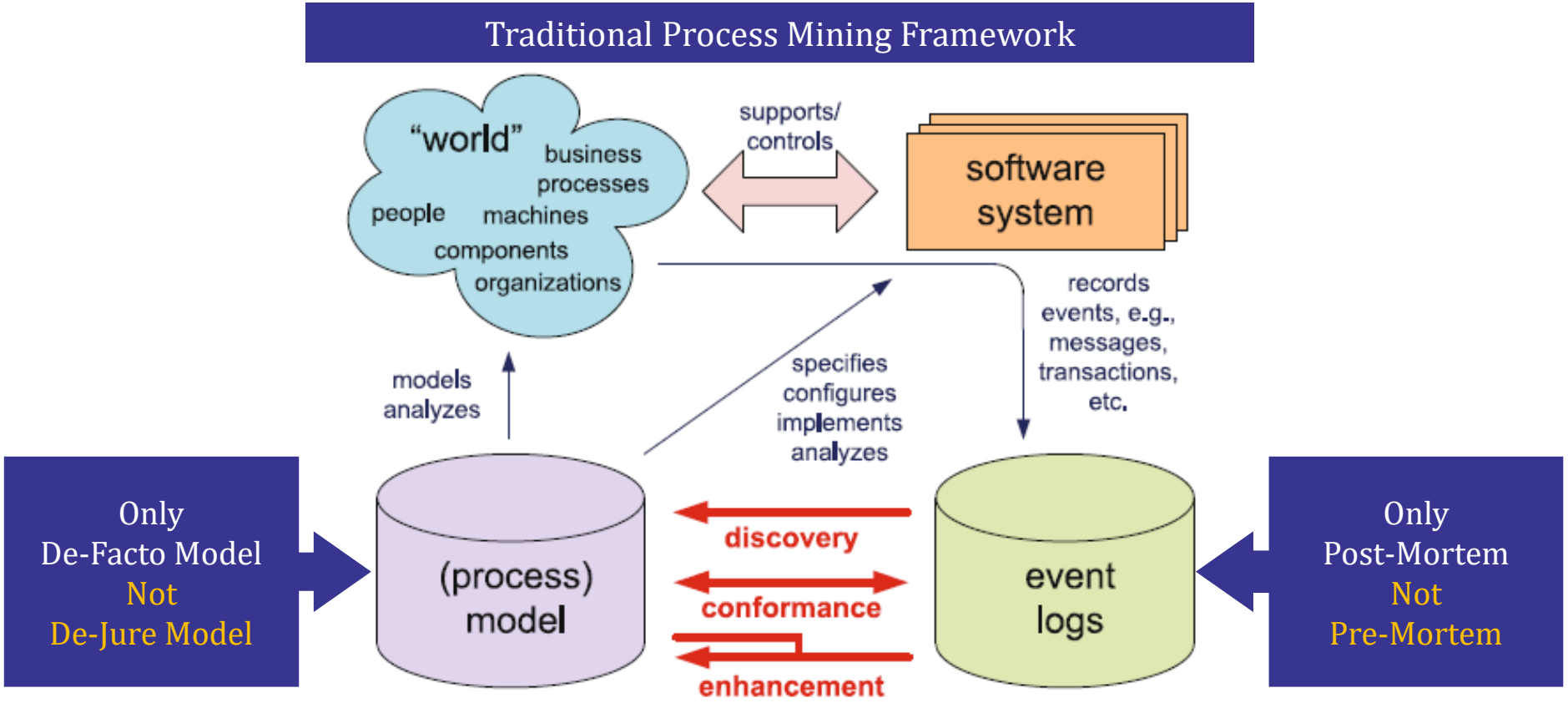
Main Purpose of Process Mining

"Process Mining" is not only about creating a process model or evaluation and improving the performance of the model based on accumulated historical data, but it also enables us to provide direction and make predictions for current unfinished data.

<Data>	<Model>
Post-Mortem (=Historic)	***De-facto Model (=Descriptive)***
• 이미 완료된 케이스의 데이터 • 프로세스 도출 및 기존 프로세스 향상에 활용 가능 • 과거 데이터로 인해 해당 케이스에 대한 예측 등의 변화 제공 등이 제한	• 현재 프로세스가 어떻게 작동하고 있는지를 표현하는 모델 • 실제 프로세스 상태를 정확히 표현하는 것을 목표
Pre-Mortem (=Current)	***De-jure Model (=Normative)***
• 아직 완료되지 않은 케이스의 데이터 • 케이스에 대한 추천이나 다음 행동을 예측하는 등의 변화를 줄 수 있는 데이터	• 목표로 하고 있는 이상적인 프로세스 모델 • 프로세스 모델이 나아가야 할 방향을 표현하는 모델

1.5 Type of Process Mining

Bridge between Traditional and Refined



1.5 Type of Process Mining

1. Traditional Process mining

Process Discovery

Discover model from event logs without any prior context

Conformance Checking

Check the conformance of an existing model using data

Enhancement

Enhance Pre-defined model based on new event logs to new improved model

1.5 Type of Process Mining

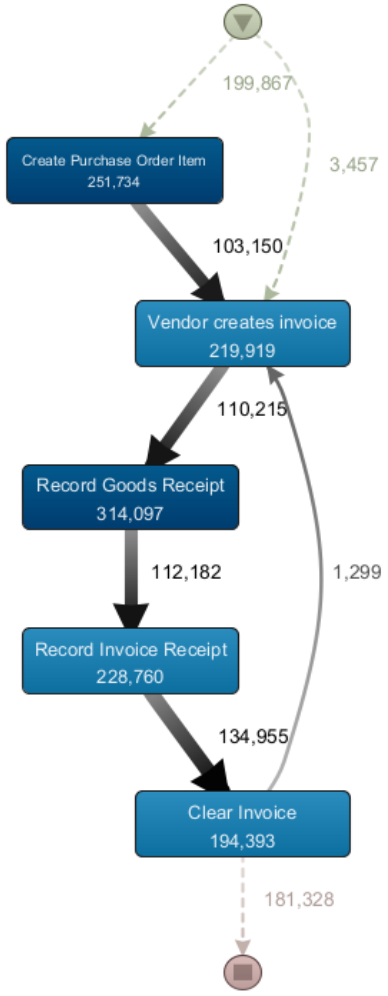
1. Traditional – Process Discovery

- How?
  - Alpha Miner
  - Heuristic Miner
  - ILP Miner
  - Inductive Miner

Activity	Timestamp	Customer
Create Purchase Order item	15-2-2018@10.30	Kevin
Vendor creates invoice	15-2-2018@10.27	Ann
Vendor creates invoice	15-2-2018@10.49	Jade
Record Goods Receipt	15-2-2018@11.10	Pole
Record Invoice Receipt	15-2-2018@12.34	Rom
Clear Invoice	15-2-2018@12.41	Lussy

Event logs

Discovery

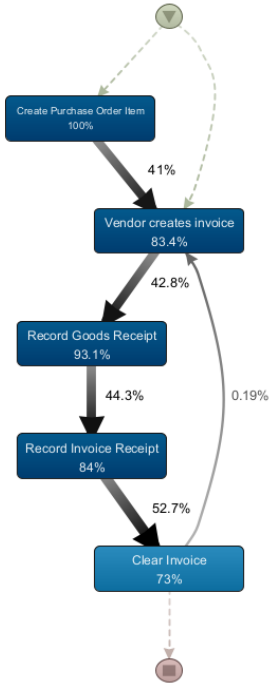


Model

1.5 Type of Process Mining

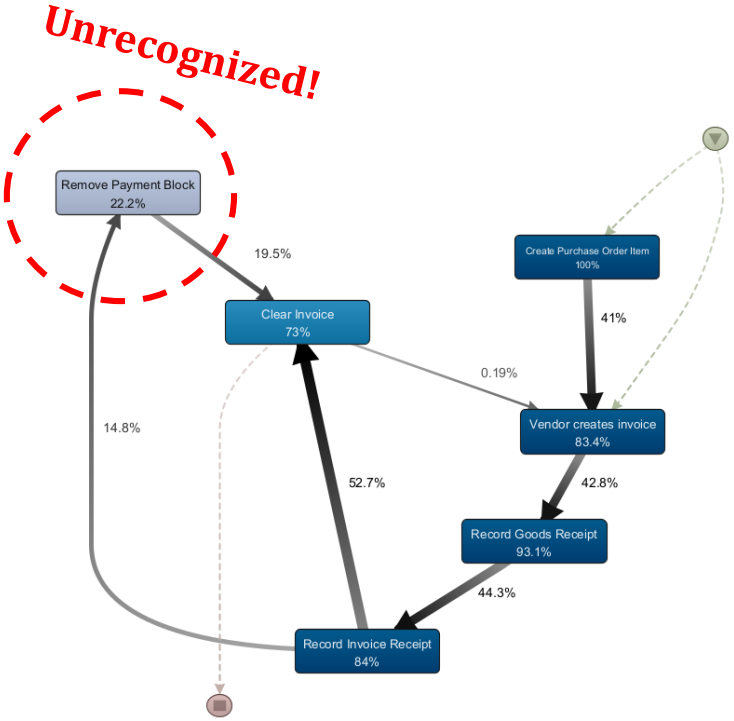
1. Traditional – Conformance Checking

- How?
  - Casual footprint
  - Token-Based Replay
  - Synchronous Product Net



Pre-defined model

Conformance check

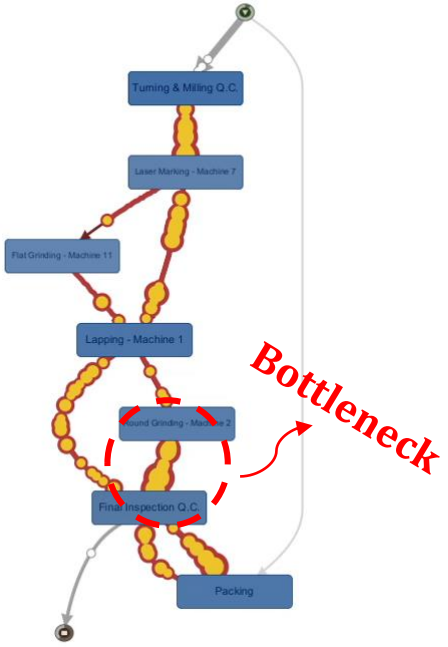


Real model

1.5 Type of Process Mining

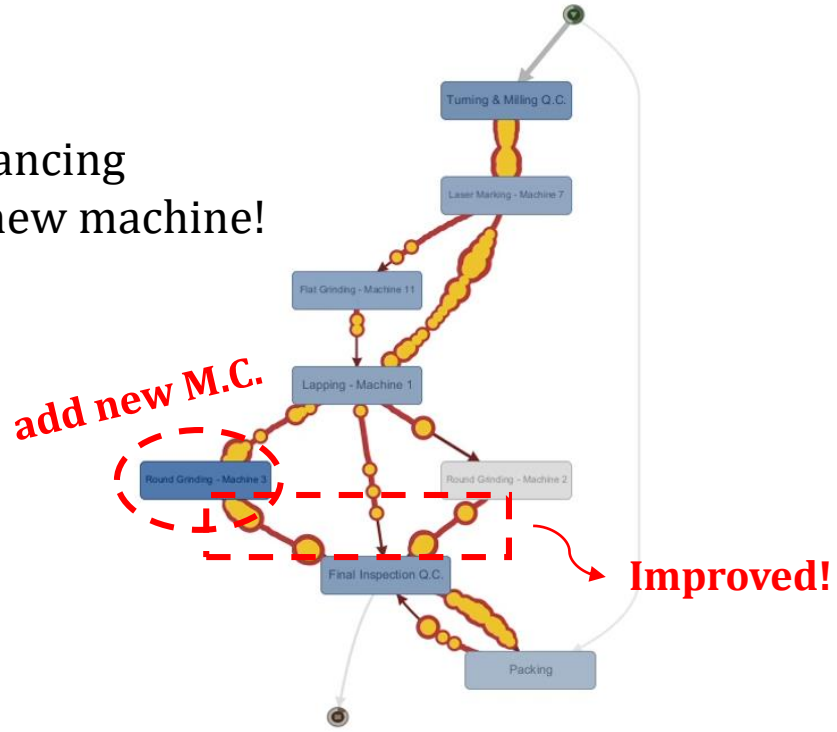
1. Traditional – Enhancement

- Enhanced model can provide more information such as…
  - Bottleneck, frequency, duration



Enhancing
by adding new machine!

Enhance

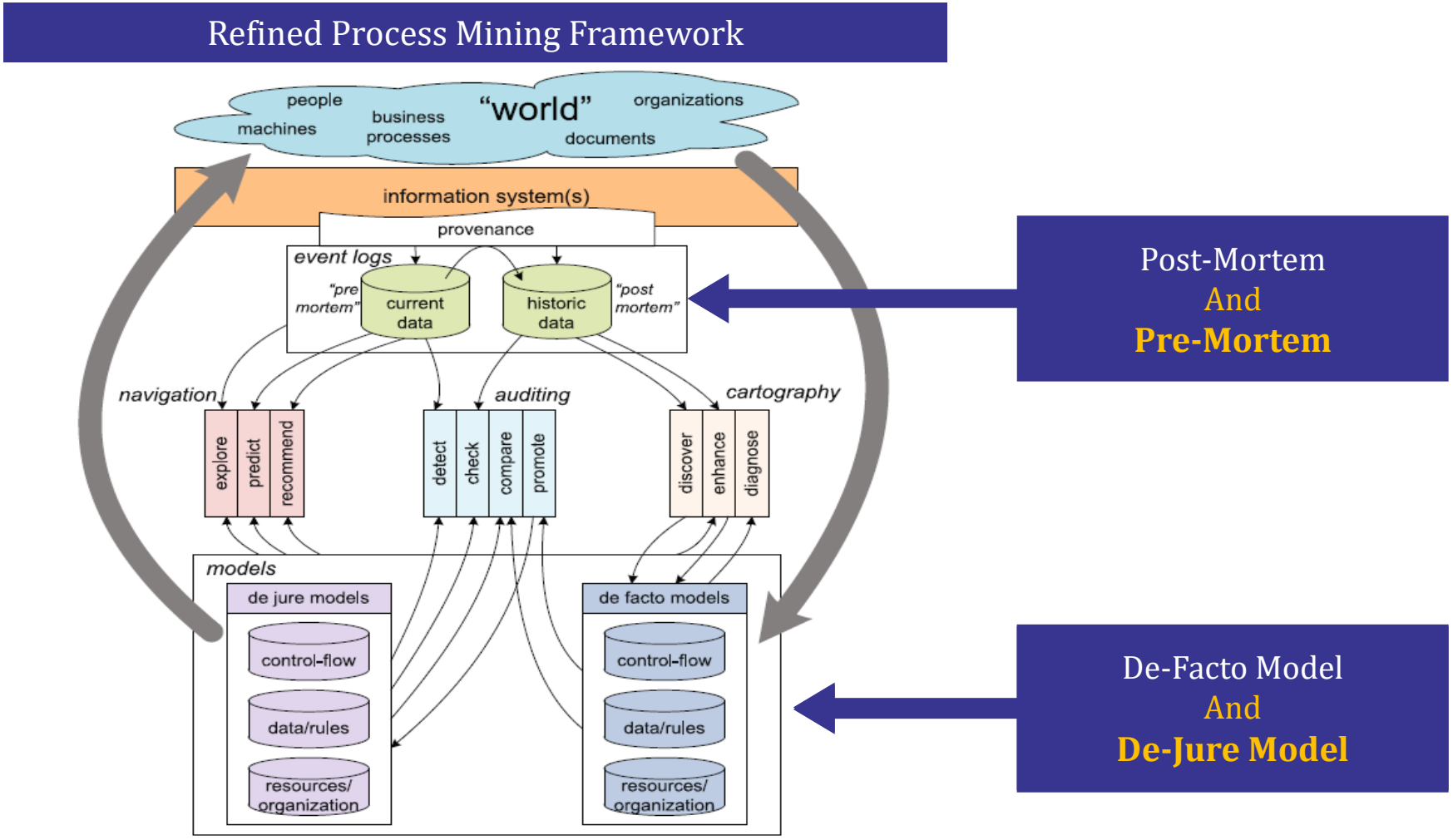


Pre-defined model

Enhanced model

1.5 Type of Process Mining

Bridge between Traditional and Refined



1.5 Type of Process Mining

2. Refined – Cartography, Auditing, Navigation

Cartography

Making Process Model
about Real World
Process

Auditing

Ensuring
the process is flowing in
the desired direction

Navigation

Giving Directions
to happen in the future

1.5 Type of Process Mining

2. Refined – Cartography, Auditing, Navigation

Cartography	**Discover**	Traditional Process Discovery
	Enhance	Developing a Process Model
	Diagnose	Diagnosing the Process Model itself
Auditing	**Detect**	Analyzing Process flow & Finding deviations
	Check	The Model complies with the Data
	Compare	Current Model VS Ideal Model
	Promote	Developing Model by "Compare"
Navigation	**Explore**	Find 'process' Run-Time data is in the Model
	Predict	Predict Info by Running-case & Model
	Recommend	Recommend Way by Running-case & Model