Week 1
Data Mining Overview

Seokho Chi
Assistant Professor I Ph.D.
SNU Construction Innovation Lab

서울대학교 건설혁신연구실
Construction Innovation Lab SNU

Course Information

- Title: 457.658 Construction IT and Automation
- Timetable
  - Monday 4-7pm @ 35-317
- Instructor: Prof. Seokho Chi
  - shchi@snu.ac.kr, 35-304
  - TA: Yoonjung Shin, nicky@snu.ac.kr, 35-429

Course Information

- *Yourself?*
- *Why are you taking? What do you want to learn?*

Course Objectives

- Understand the fundamentals of data mining and knowledge discovery in database
- Apply data management techniques for data classification, prediction, clustering, and mining association rules
- Demonstrate how knowledge discovery in database can be used to support construction management
- Recognize the design, analysis, and implementation issues for data management in civil engineering

Course Materials

- Required
  - Lecture slides and handouts
  - eTL: Update correct contact info

- References
  - Tan, P., Steinback, M., and Kumar, V. (2005) Introduction to Data Mining, Addison-Wesley

Note

- English Lecture, Presentation, and Assignment
- Group Assignment
  - Teamwork is important.
  - Active participation is required.
- Cheating and Plagiarism
  - 0% for the given assessment item without any excuse
  - Penalty by SNU's regulations

Assessment

Item	Weight	Due
Attendance	10%	
Group Assignment		
Interim Report	15%	Week 8 (10/20)
Final Report	20%	Week 15 (12/8)
Final Presentation	5%	Week 15 (12/8)
Individual Assignment	20%	
Final Exam	30%	Week 14 (12/1)
TOTAL	100%	

Course Schedule (1)

Week	Date	Contents
1	9.1	Course Introduction Data Mining Overview
2	9.8	**No Class**
3	9.15	Data Types Data Pre-Processing Data Exploration
4	9.22	Data Visualization Classification
5	9.29	Classification
6	10.6	Computer Lab (1)

Course Schedule (2)

Week	Date	Contents
7	10.13	Classification Prediction
8	10.20	**Interim Group Presentation**
9	10.27	Computer Lab (2)
10	11.3	Cluster Analysis
11	11.10	Mining Association Rules
12	11.17	Computer Lab (3)
13	11.24	Mining Complex Data Types Trends and Construction Applications
14	12.1	**Final Exam**
15	12.8	**Final Group Presentation**

Group Project Brief

- For this project, each group will mine a database to analyze/solve a construction engineering problem. Each group must identify a data set for this project.
- Examples include: productivity, safety performance, pavement management, environmental remediation, project disputes, soil characterization, structural monitoring, schedule control, property appraisals, quality control, among others.

- On Phase I, each team must submit a project proposal. The proposal must describe the problem that will be investigated, justify the need to conduct a data mining study to analyze/solve this problem, provide a short background review on related topics, specify the specific project objectives and scope, identify the target data set, and describe the proposed data mining approaches.
- Each team should perform **at least two** data mining tasks (e.g., classification and clustering) and use **at least three** different algorithms/methods (e.g., decision tree, neural network, and naïve bayes).

Group Project Brief

- On the Final Phase, each team must submit a project report, including the results, discussion, conclusions, and recommendations.

- Each group must meet **at least two** times with me until the end of the course to discuss about the project proposals, progress, and results → Each group should meet **at least once** before the due data of each deliverable. Groups should contact me to schedule these meetings.

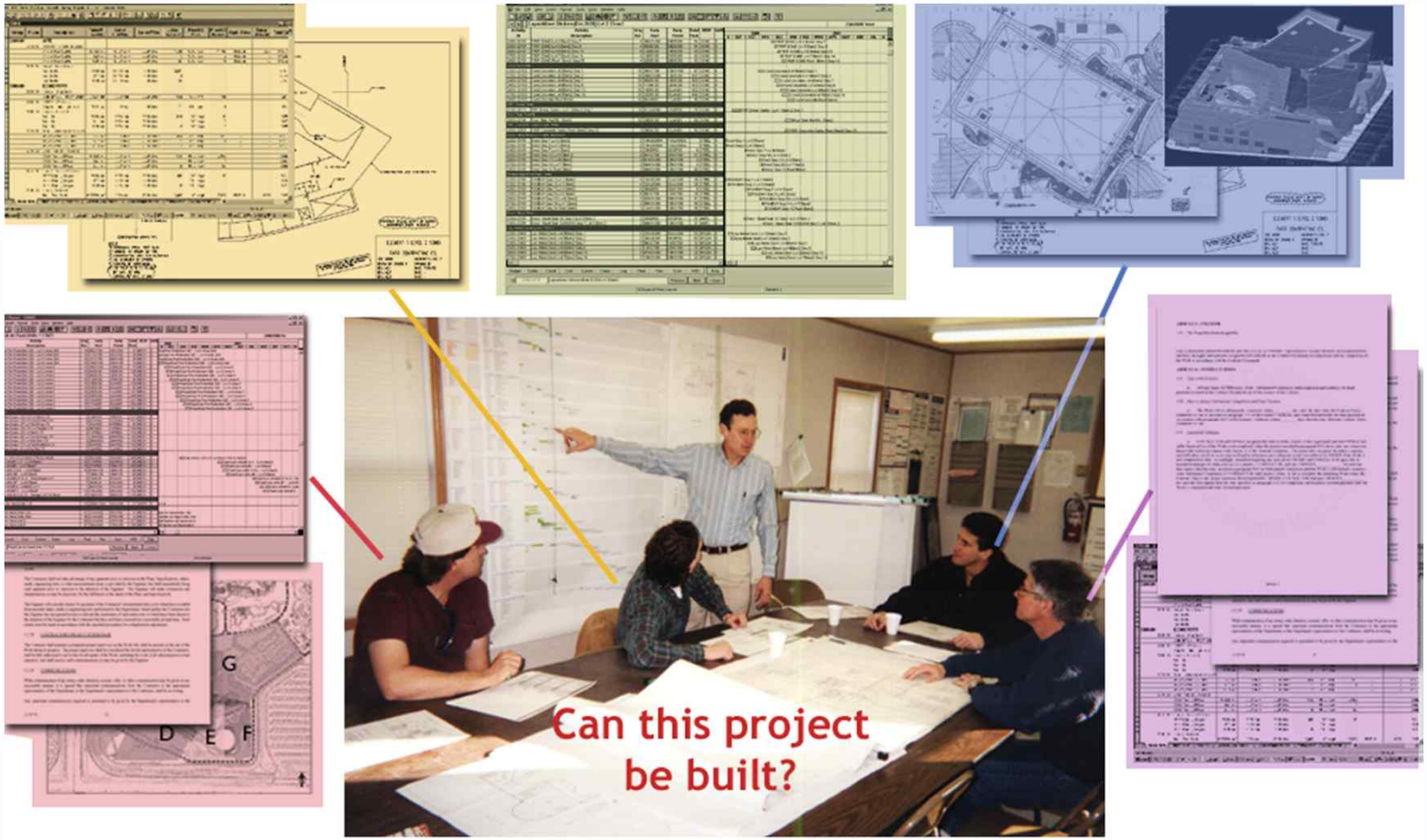
- The data mining should be conducted using WEKA, SAS or other software of your choice.

Group Project Brief

- **DELIVERABLES**
  - Deliverable 1 (10/20) – Project Proposal
    - Problem definition, background, need, objectives, scope, target data set, and proposed data mining approaches
  - Deliverable 2 (12/8) – Project Report
    - Summary of items included on deliverable 1, final results, discussion, conclusions, and recommendations.
- **PRESENTATIONS**
  - Phase 1 (10/20) – Deliverable 1
  - Final (12/8) – Deliverable 2

Data Mining Overview

Data on Construction Projects

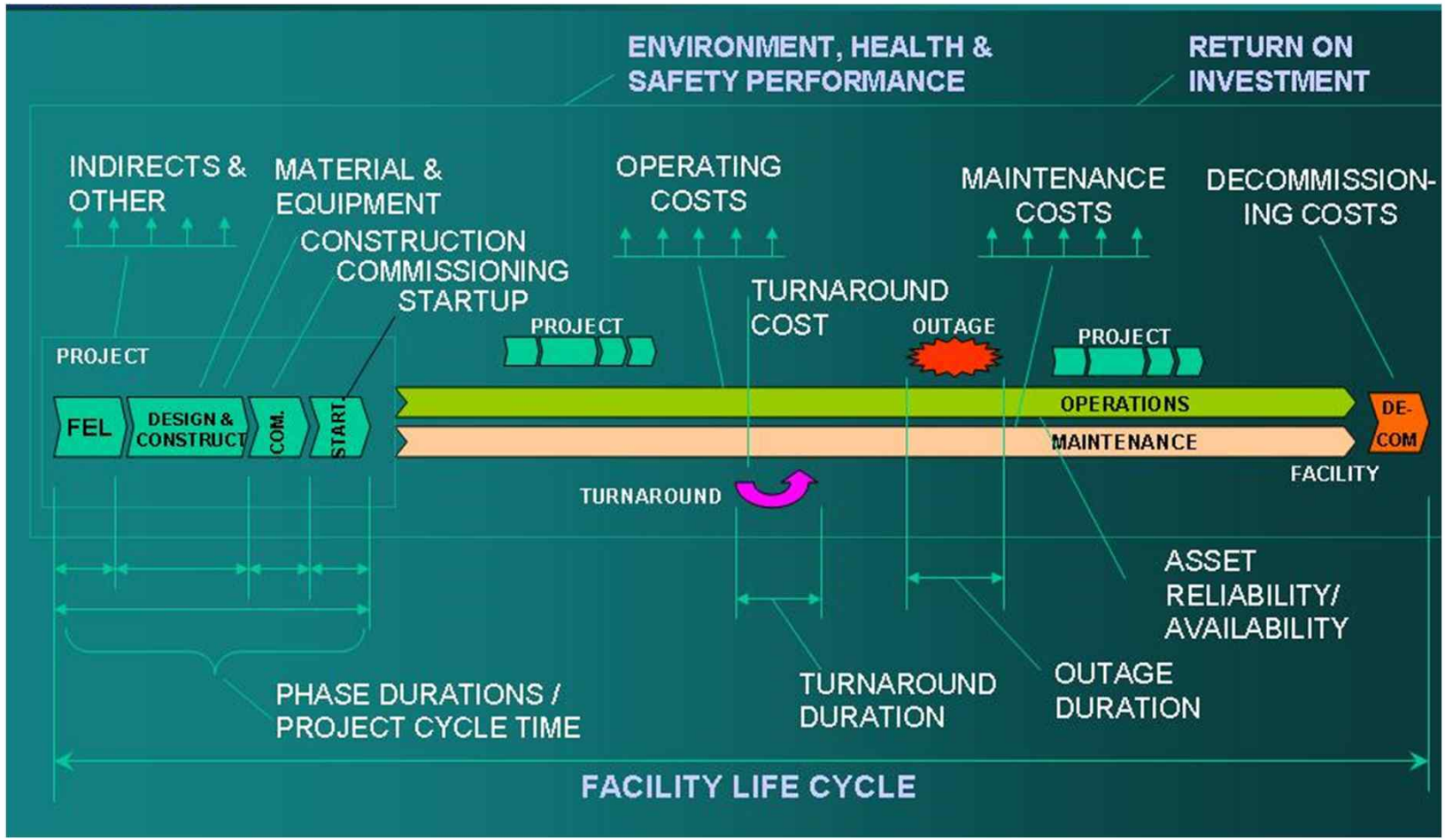


Amount of Project Information from $10M and over Construction Projects (in 2004)

420 Stakeholders, 850 People, 50 Document Types, 56,000 Pages

Data on Construction Projects

- **Information Flood through Life-Cycle of a Project**



Information Technology

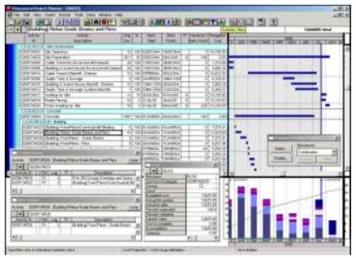
CONSTRUCTION PROJECT

Planning → **Design** → **Construction** → **Operation & Maintenance**

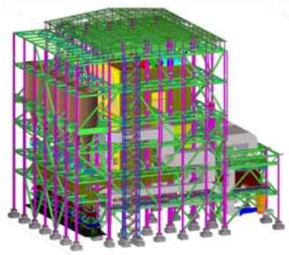
(PMBOK, 2000)

Scope Management	Procurement Management
HR Management	Decision/Risk Management
Time Management	Cost Management
Quality Management	Safety Management

Provide Information, Support Communication, and Strengthen Knowledge



PM Software

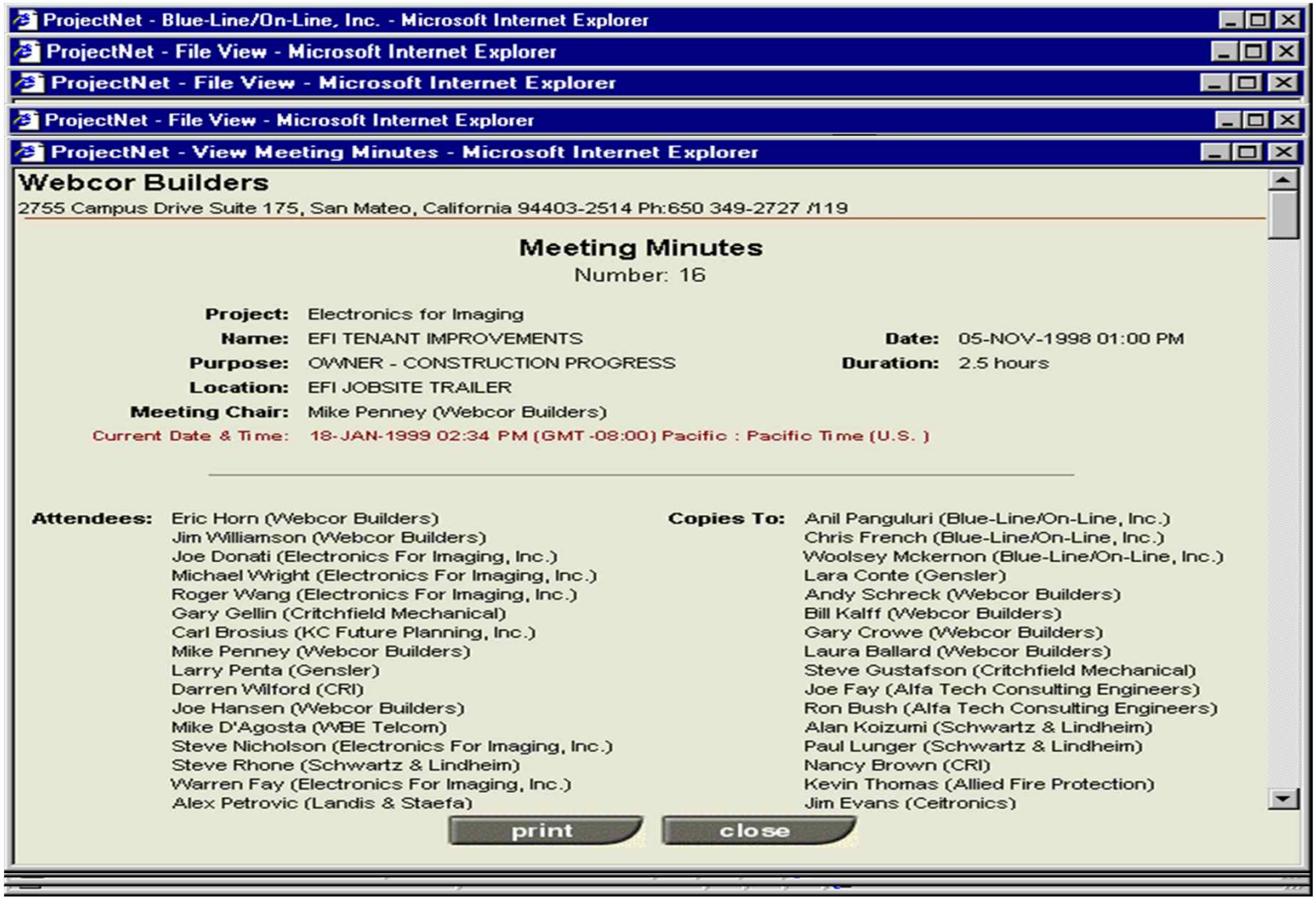


3D Design

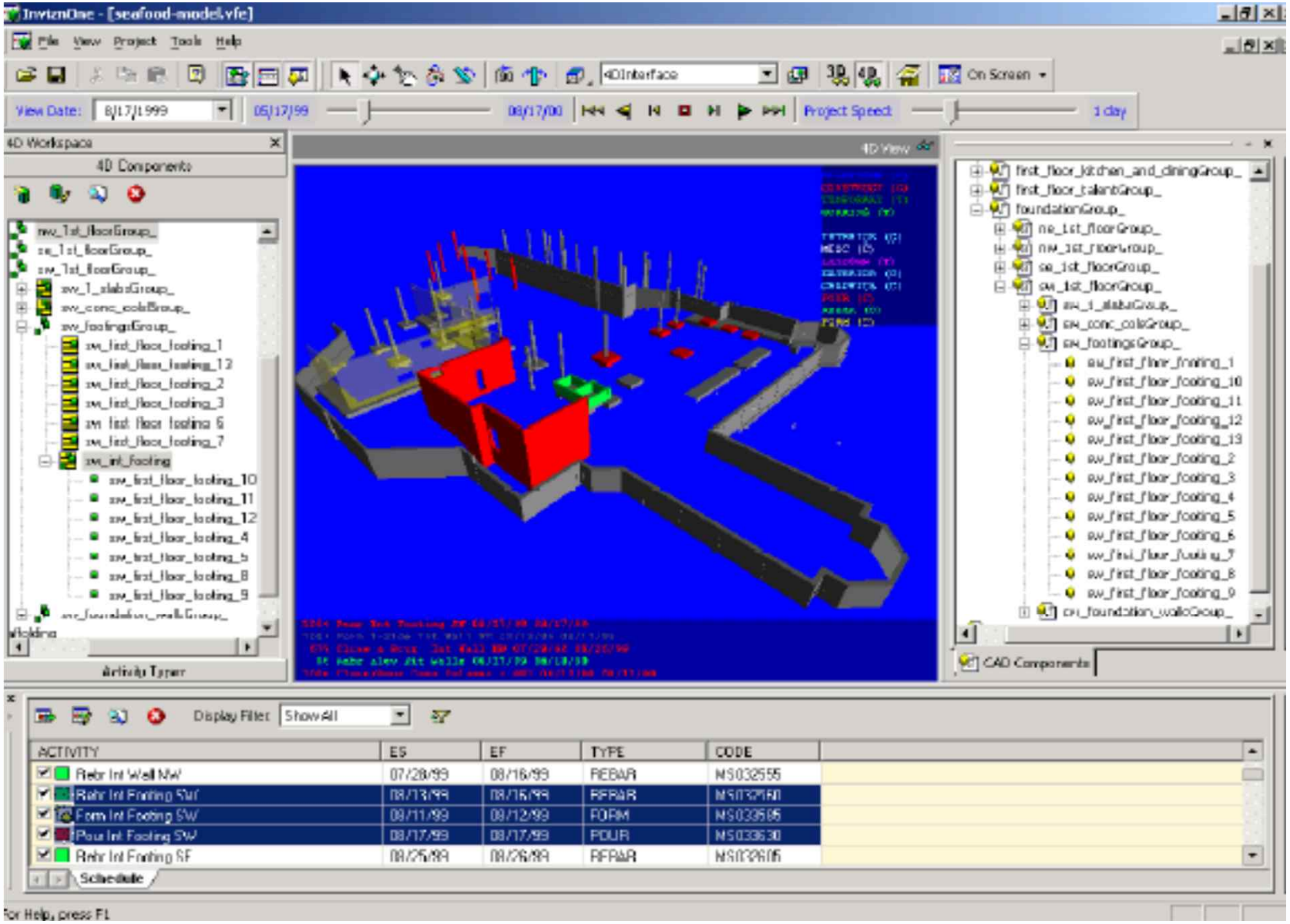
Mobile Computing

Sensing/GPS/Tags

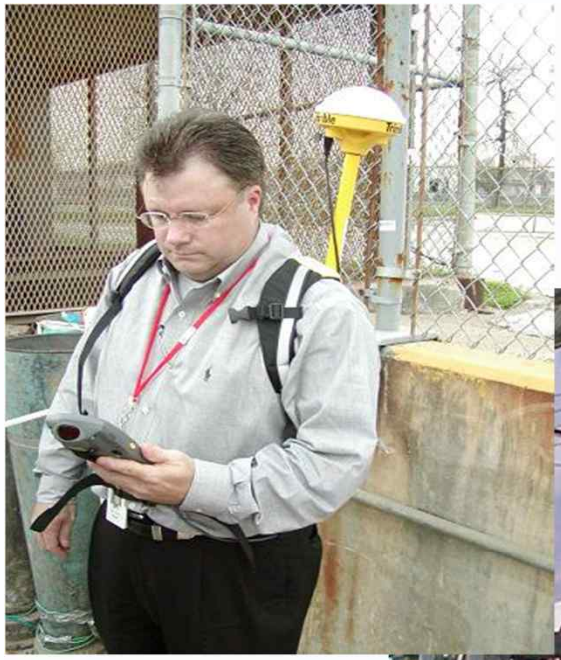
Project Data Sources (1)

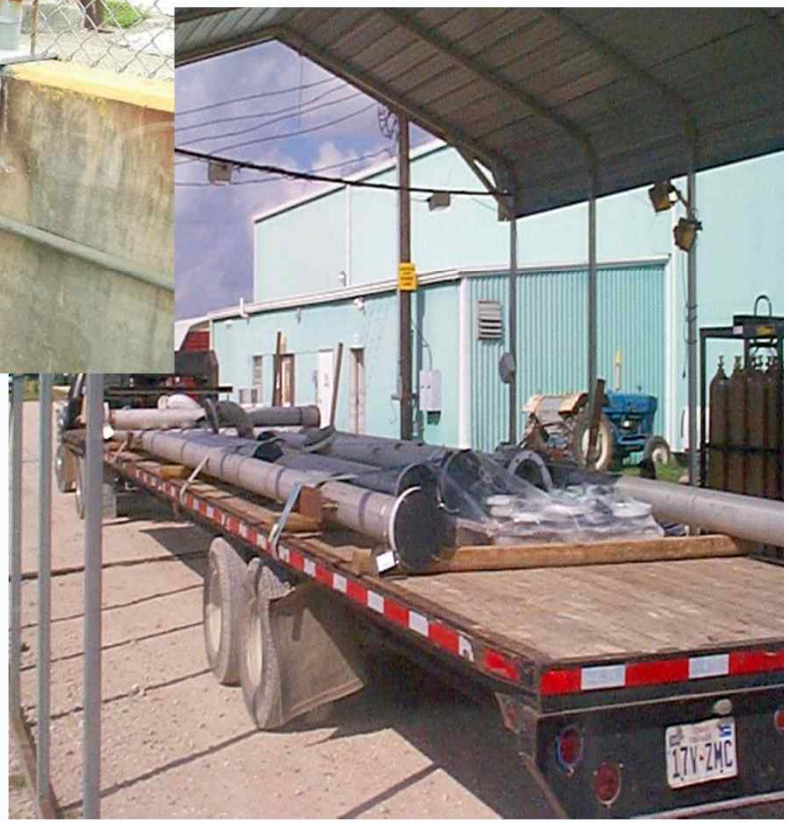


Project Data Sources (2)



Project Data Sources (3)





SMART CHIPS - PILOT PROJECT ON PIPING MARKED PI

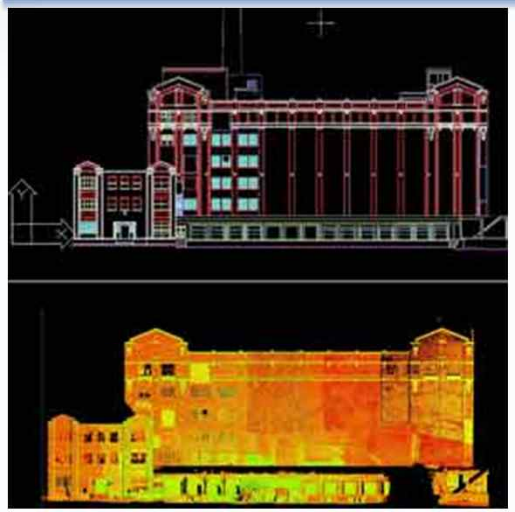
Test 8

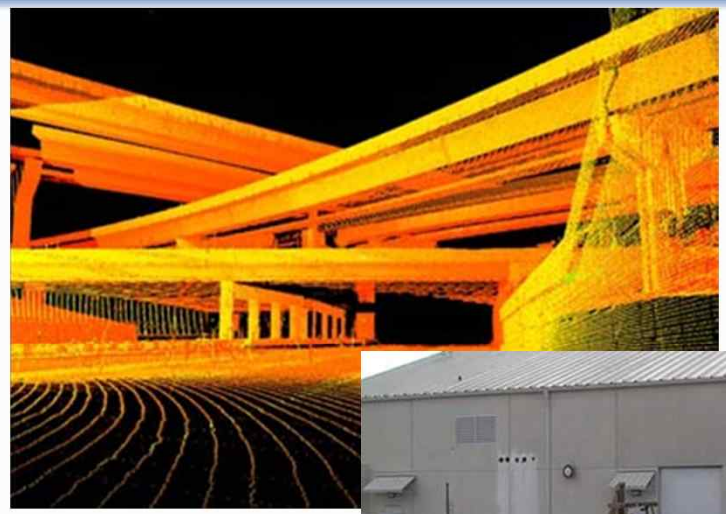
Test Conditions:

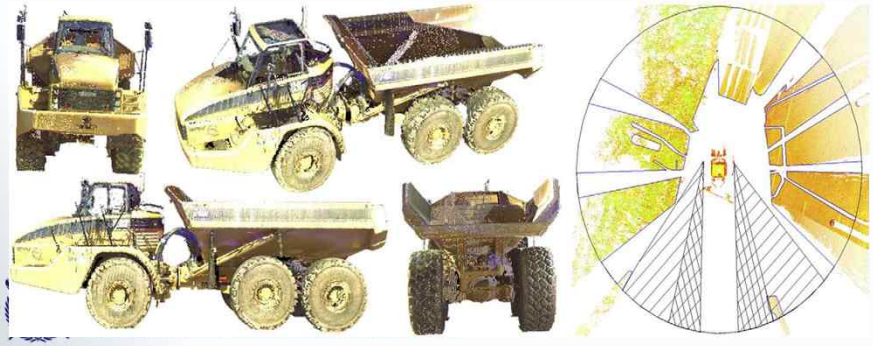
LOCATION SHAW GROUP, HOUSTON

TAG ID	BAR CODE	SPOOL #	SKETCH NO	COMPANY	P.O. NUMBE	MK. PC. #	LOCATION ON SPOO
CW00000016	1900039728	31	66184270	DOW STCHR	31704542	1252077-08-Q	
CW00000017	1900039730	33	66184270	DOW STCHR	31704542	1252077-08-S	
CW00000018	1900039748	51	66184270	DOW STCHR	31704542	1252077-11-K1	
CW00000019	1900039718	21	66184270	DOW STCHR	31704542	1252077-07-E	
CW00000020	1900039712	15	66184270	DOW STCHR	31704542	1252077-06-G	
CW00000021	1900039710	13	66184270	DOW STCHR	31704542	1252077-06-E	
CW00000022	1900039734	37	66184270	DOW STCHR	31704542	1252077-09-W	
CW00000023	1900039726	29	66184270	DOW STCHR	31704542	1252077-08-N	
CW00000024	1900039751	54	66184270	DOW STCHR	31704542	1252077-11-N1	
CW00000025	1900039739	42	66184270	DOW STCHR	31704542	1252077-09-B1	
CW00000026	1900039742	45	66184270	DOW STCHR	31704542	1252077-10-D1	
CW00000027	1900039745	48	66184270	DOW STCHR	31704542	1252077-10-G1	
CW00000028	1900039762	1	66184070	DOW STCHR	31704542	1252077-03-A	
CW00000029	1900039740	43	66184270	DOW STCHR	31704542	1252077-09-C1	
CW00000030	1900039736	39	66184270	DOW STCHR	31704542	1252077-09-Y	
CW00000031	1900039725	28	66184270	DOW STCHR	31704542	1252077-08-M	
CW00000032	1900039735	38	66184270	DOW STCHR	31704542	1252077-09-X	
CW00000033	1900039752	55	66184270	DOW STCHR	31704542	1252077-11-P1	
CW00000034	1900039715	18	66184270	DOW STCHR	31704542	1252077-07-B	

Project Data Sources (4)





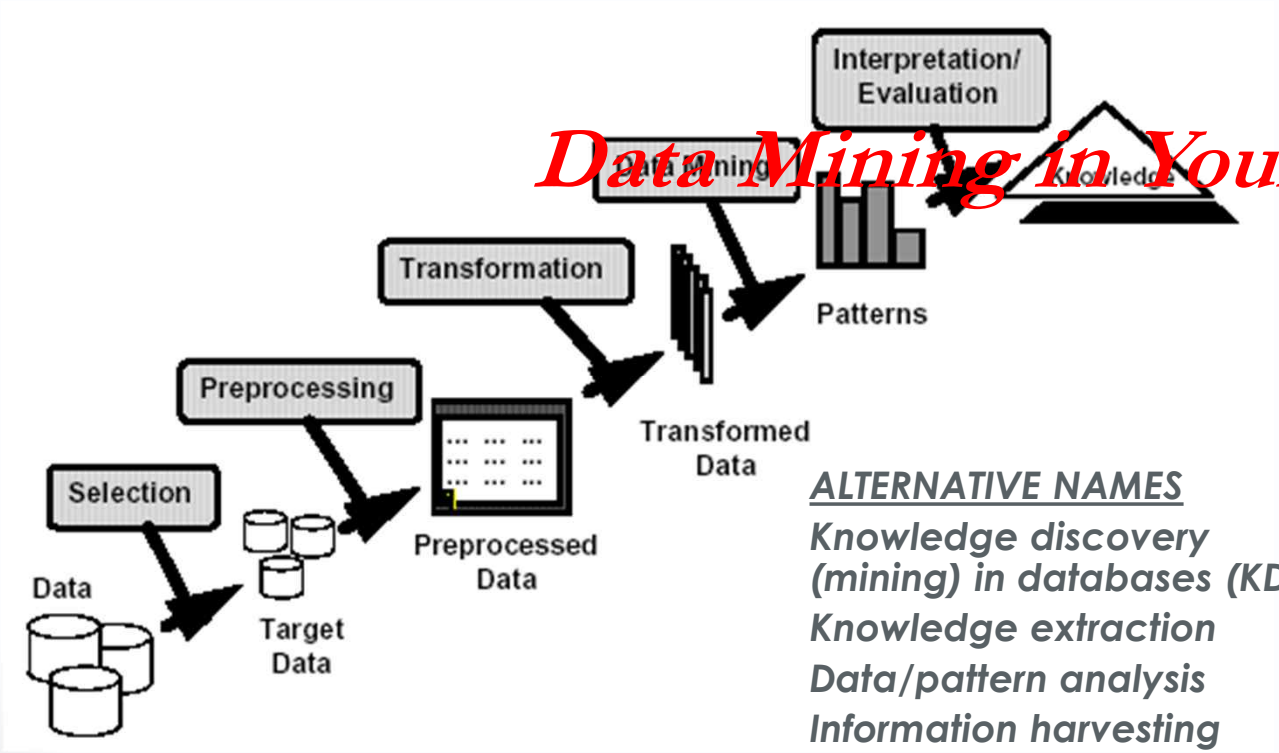


Motivation

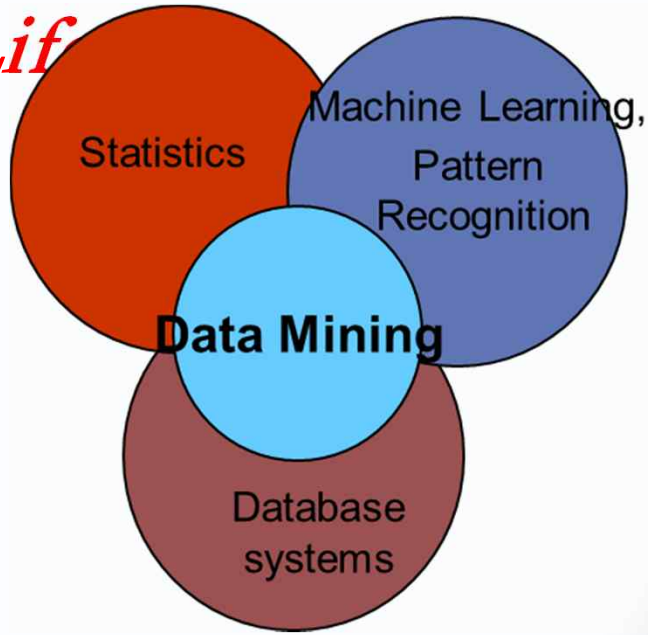
- The information explosion has us drowning in data but often starved of knowledge. Many companies that gather huge amounts of electronic data have now begun applying data mining techniques to their databases to discover and extract pieces of information useful for making smart business decisions.

Data Mining

- **What is Data Mining?**
  - Knowledge discovery from data
  - Extraction of interesting patterns or knowledge from huge amount of data



Data Mining in Your Life



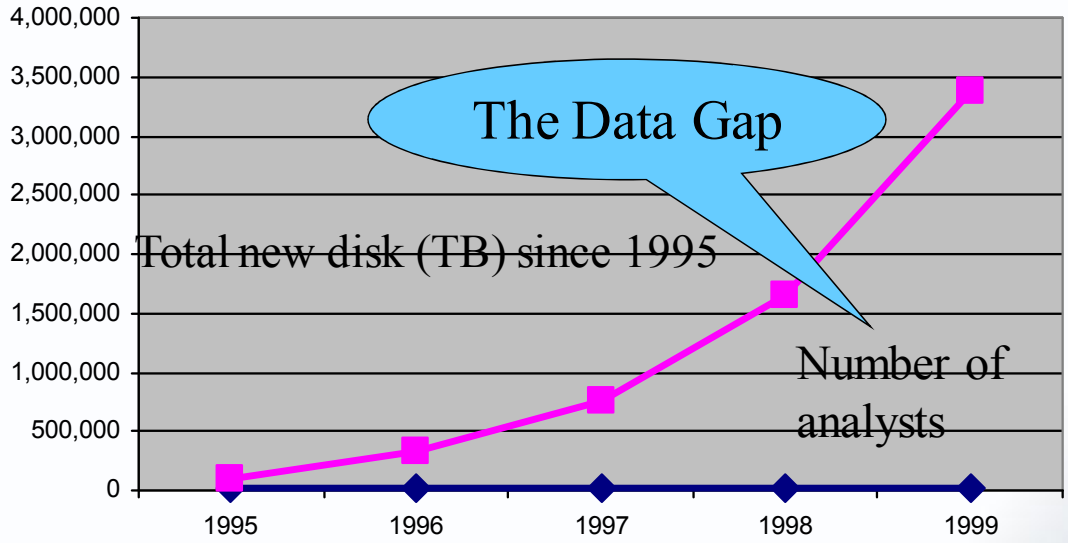
ALTERNATIVE NAMES

Knowledge discovery (mining) in databases (KDD)
Knowledge extraction
Data/pattern analysis
Information harvesting

Data Mining

- **Information Flood**
  - Purchases at department/ grocery stores
  - Bank/Credit card transactions
  - Web data, e-commerce (text mining)
  - Remote sensors on a satellite
  - Forecasting, risk analysis and management

- **PM Data Mining**
  - A lot of data available
  - Human analysts may take weeks to discover useful information
  - Much of the data is never analysed at all



Purpose of Data Mining: Data → Info. → Knowledge

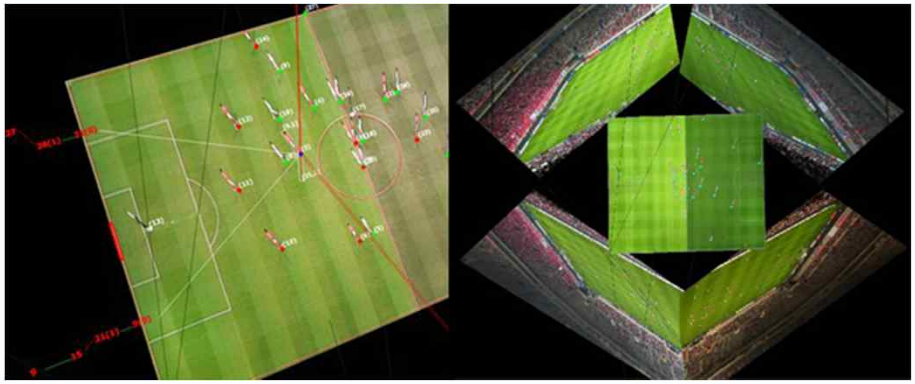
- **Data**
  - Raw description of things, events, activities and transactions that are recorded, but alone do not convey any specific meaning (e.g. 400,000)

- **Information**
  - Data that have been organized so that they have meaning and value to the recipient (e.g. Current $400,000 house price)

- **Knowledge**
  - Information that has been organized and processed to convey understanding experience and expertise as they apply to a current problem or activity (e.g. The current $400,000 house price is cheaper than the last year's price. The property market may be deflated.)

Soccer Player Analysis Using Image Processing

- **Raw Data: Recorded Video**
- **Information: Position Identification & Tracking**
  - Travel distance, movement, position, time, etc.
- **Knowledge: Performance Analysis**
  - Efficient/Inefficient movement, pass accuracy, reasons for poor performance, etc.
  - Planned vs. actual, team formation analysis, set play analysis, etc.
  - Use knowledge for team training

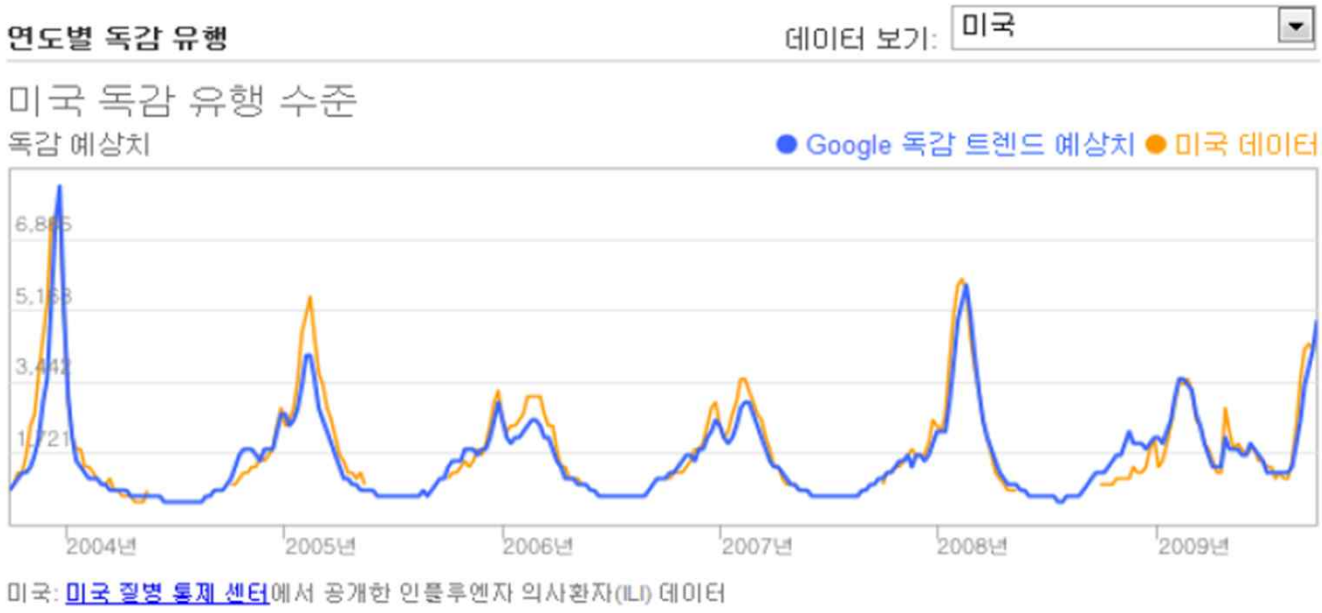


Big Data Cases

Case 1: Google Flue

지역별 독감에 관련된 키워드의 검색패턴을 실시간으로 분석하여 독감 확산 여부를 의료 당국 조사보다 빠르고 정확하게 파악

- 미국, 독일, 일본, 남아프리카공화국 등 국가별 연도별 독감 유행 자료를 구글 독감 트렌드(http://www.google.org/flutrends)를 통해 제공
- 독감과 함께, 뎅기열 유행을 실시간 파악하여 구글 뎅기열트렌드 (http://www.google.org/denguetrends)를 통해 제공



Case 2: SNS Analysis

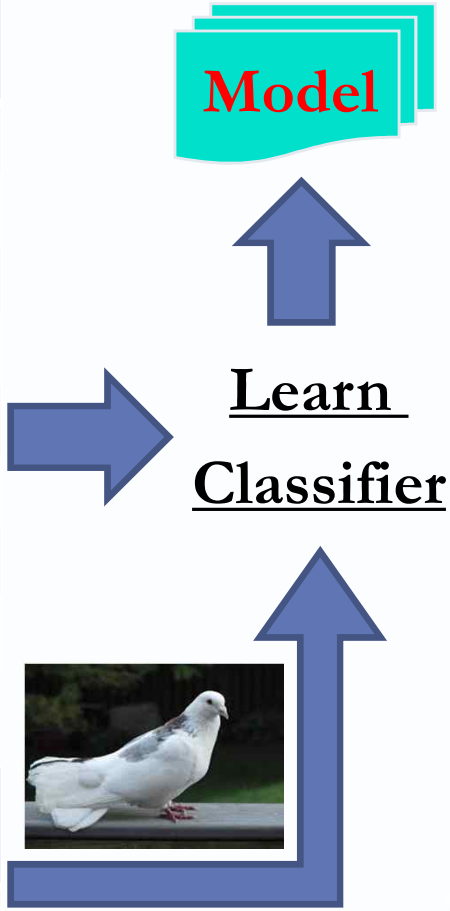
- 최근 2년동안 미국과 아일랜드에서 인터넷 채팅, 블로그, 페이스북, 트위터 등 소셜미디어 데이터의 기분이나 정서를 분석[13)]
  - 미국에서 '우울하다', '열 받는다'와 같은 채팅이 늘어나면 4개월 뒤 실업률이 폭등함을 확인
  - 아일랜드에서는 실업률 증가 5개월 전 '불안하다'는 분위기가 퍼져나갔고, 2개월 전에는 '확신한다'는 채팅이 크게 감소

실험 방법은 간단했다. 실험군과 비교군으로 나누고 실험군 68만9003명의 뉴스피드를 조작했다. 결론부터 얘기하자. 페이스북에서도 감정 전이 현상이 나타났다. 긍정적인 게시물이 줄어들면 사용자는 긍정적인 표현을 줄이고 부정적인 게시물을 더 많이 올렸다. 반대로 뉴스피드에 나타나는 부정적인 게시물이 줄어들면 사용자는 긍정적인 게시물을 더 많이 올렸다. 친구와 직접 교류하는 게 아니고 뉴스피드만 봐도 페이스북 사용자가 감정에 영향을 받았다는 뜻이다. 잘 사는 친구 게

(1) Classification (e.g., Classifying Mammals)

TRAINING SET

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
Human	Warm-blooded	Y	N	N	Y
Elephant	Warm-blooded	Y	Y	N	Y
Leopard shark	Cold-blooded	Y	N	N	N
Turtle	Cold-blooded	N	Y	N	N
Penguin	Cold-blooded	N	N	N	N
Eel	Warm-blooded	N	N	N	N
Dolphin	Warm-blooded	Y	N	N	Y
Spiny anteater	Cold-blooded	N	Y	Y	Y



TESTING SET

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
Pigeon	Warm-blooded	N	N	N	?

(1) Classification

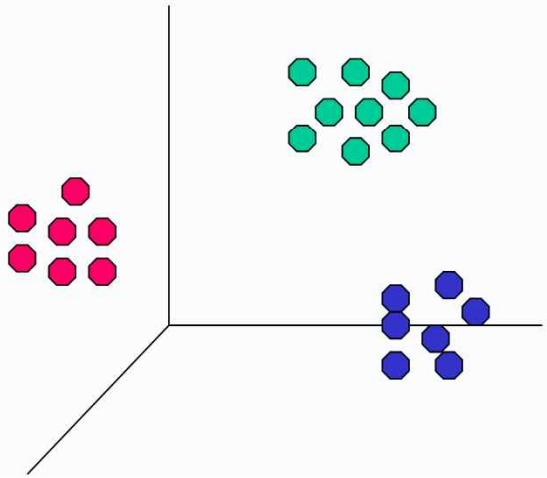
- Direct Marketing
  - Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product
  - Approach: Use the data for a similar product introduced before → We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decisions forms the class attribute → Collect various demographic, lifestyle, and company-interaction related information about all such customers (e.g., type of business, where they stay, earning, etc.) → Use this information as input attributes to learn a classifier model

(2) Clustering

- Given a set of data points, each having a set of attributes and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another

Intracluster distances are minimized

Intercluster distances are maximized



(2) Clustering: Document Clustering

- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document, form a similarity measure based on the frequencies of different terms and use it to cluster
- Gain: Information retrieval can utilize the clusters to relate a new document or search terms to clustered documents

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

(3) Association Rule Discovery

- Given a set of records each of which contain some number of items from a given collection
- Produce **dependency rules** which will predict occurrence of an item based on occurrences of other items

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Coke, Diaper, Milk
5	Beer, Diaper, Milk

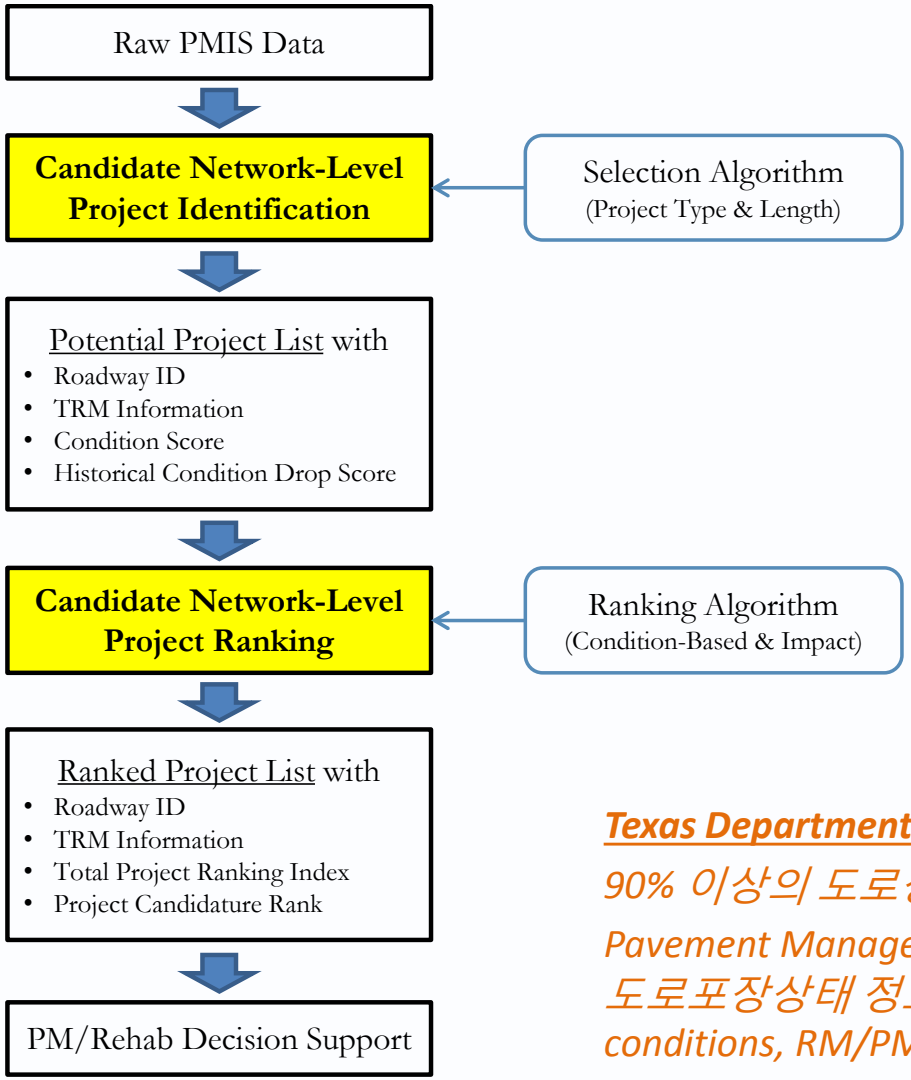
Rules Discovered:
{Milk} → {Coke} X
{Diaper, Milk} → {Beer} O

- *Marketing and sales promotion*
- *Supermarket shelf management*
- *Inventory management*

(4) Sequential Pattern Discovery

- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.
- Association rule: Concurrent events
- Examples: Computer bookstore: Intro to C++ → MFC using C++
  Shoes → Racket, Racketball → Sports Jacket

Data Mining for Pavement Management



CS	CS_Drop	Roadway	TRM	TRM_DISP
18	-24	BI0035LK	422	1.5
58	33	BI0035LK	424	0
53	5	BI0035LK	424	0.5
56	26	BU0079BK	456	0.2
52	15	FM0112 K	556	0
20	-5	FM0112 K	556	0.5
20	1	FM0112 K	556	1
47	31	FM0112 K	556	1.5
58	24	FM0112 K	558	1.5
52	27	FM0112 K	562	1
35	-55	FM0112 K	562	1.5
53	11	FM0112 K	564	0.5
47	7	FM0112 K	564	1
48	16	FM0112 K	566	1.5
27	7	FM0112 K	568	1
28	-6	FM0112 K	568	1.5

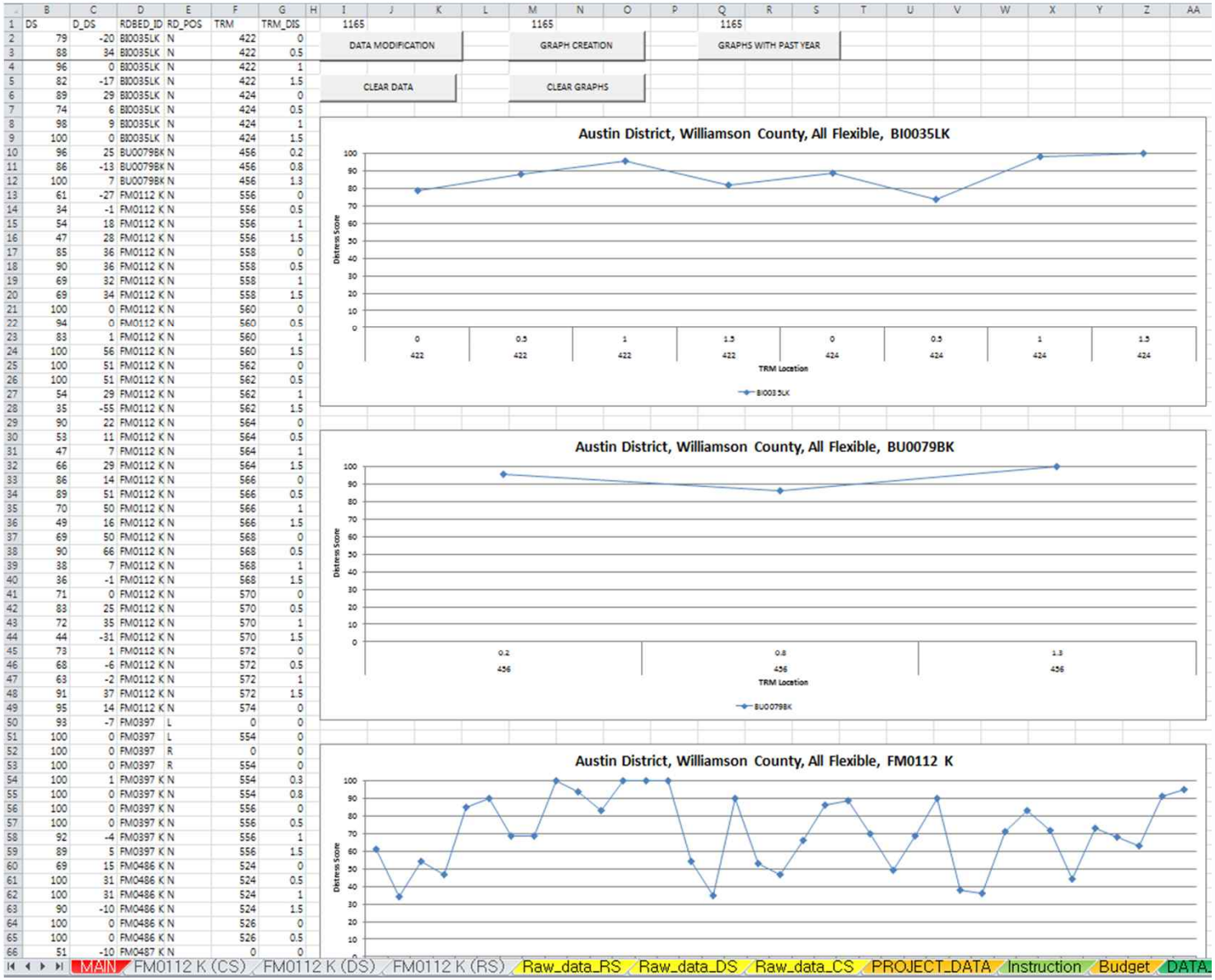
Texas Department of Transportation

90% 이상의 도로상태를 "Good or Better"로 목표

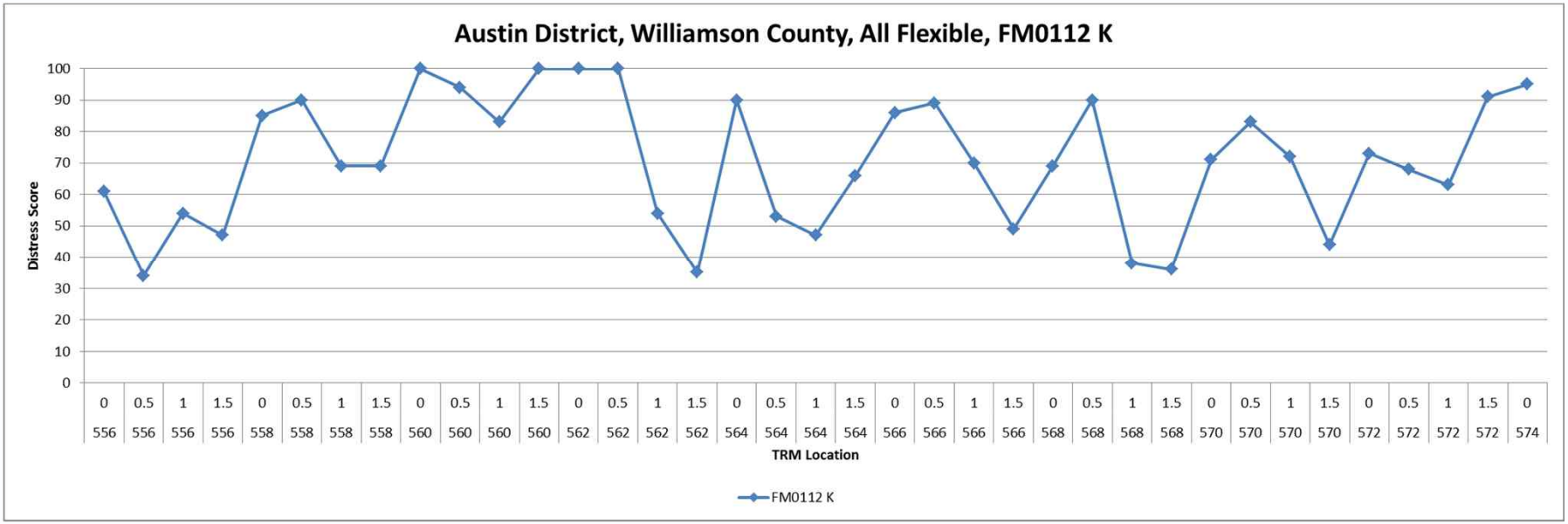
Pavement Management Information Systems (PMIS): 4개월에 걸친 도로포장상태 정보수집 (distress score, ride score, actual distress conditions, RM/PM/Rehab 예산관련정보 등)

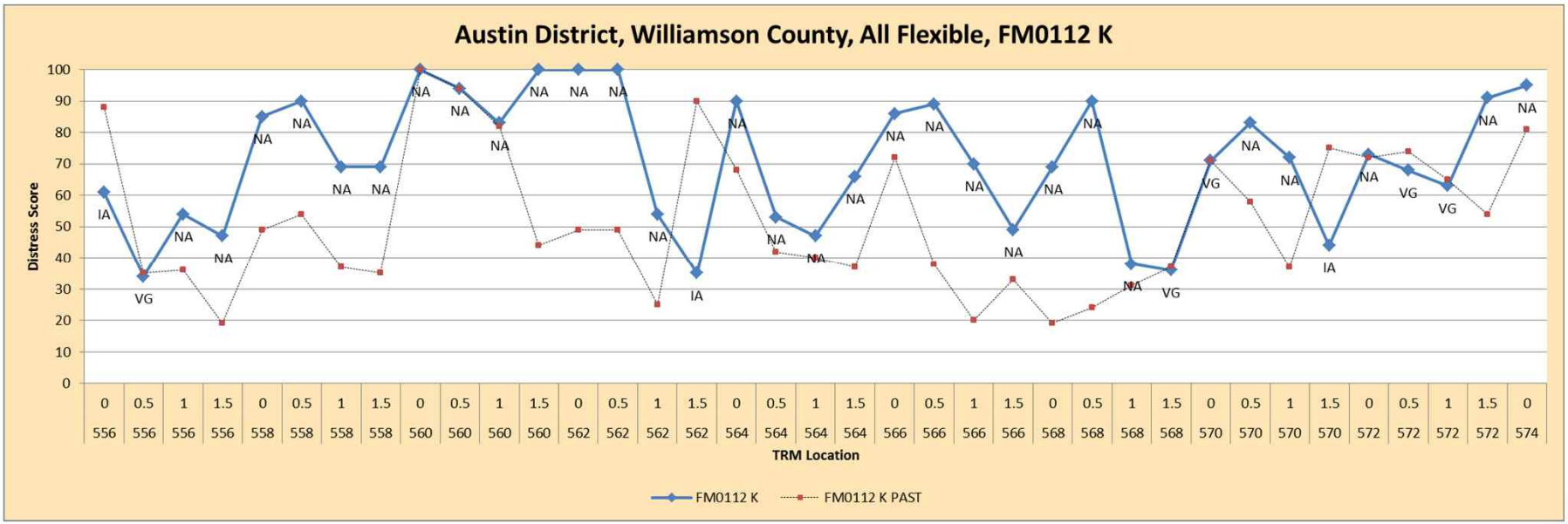
Ref.: Total road in Texas: 314,000 km (2008) VS Korea: 106,414km (2013)

Data Mining for Pavement Management (Processing)



Data Mining for Pavement Management (Processing)





Data Mining for Pavement Management (Rules)

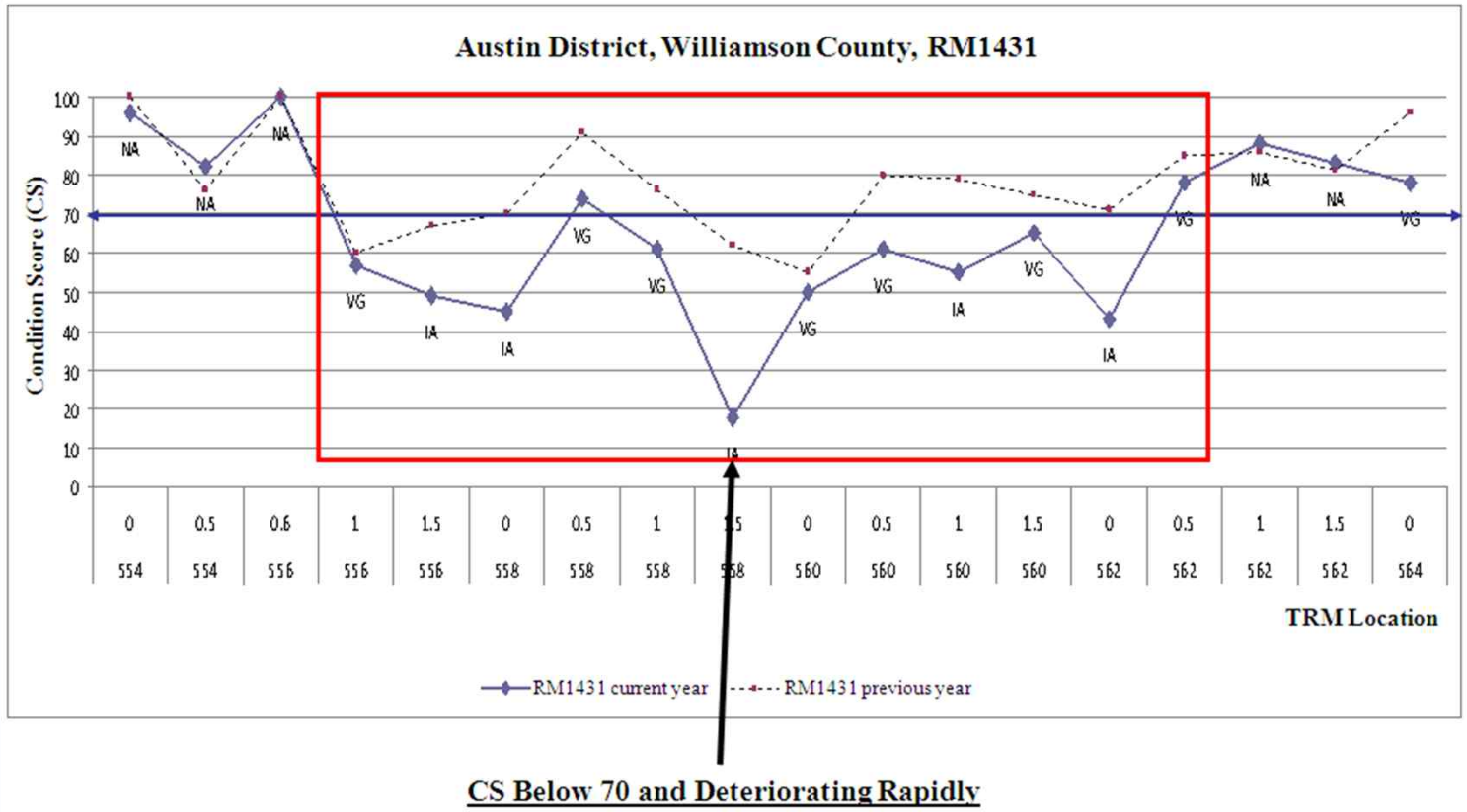


Figure 3 An example Condition Score and Condition Score change graph produced by the first version of the algorithm

Data Mining for Pavement Management (Rules)

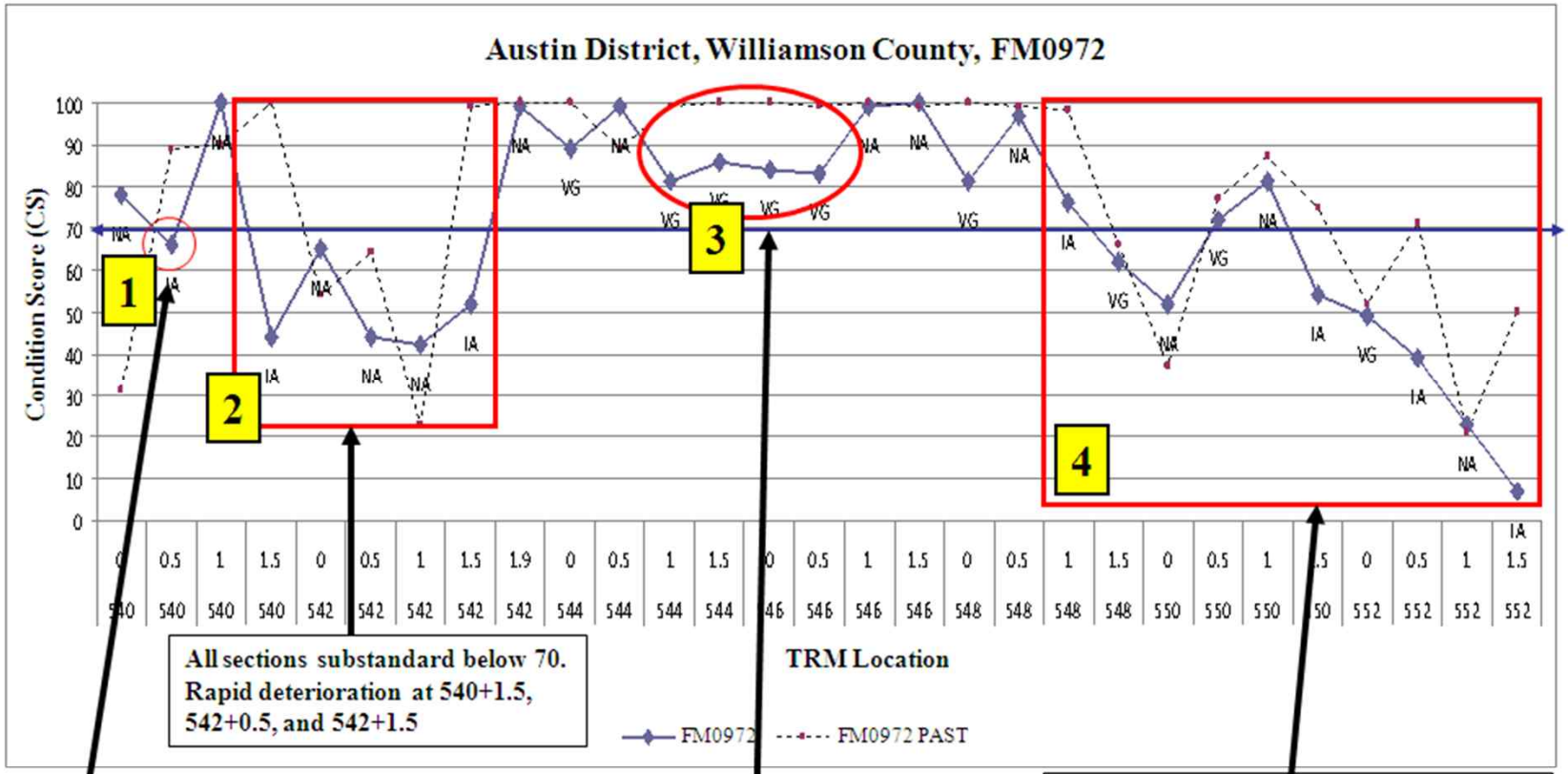


Figure 4 An example of input provided by the District Pavement Engineer (DPE) using a Condition Score graph

Data Mining for Pavement Management (Data Mining)

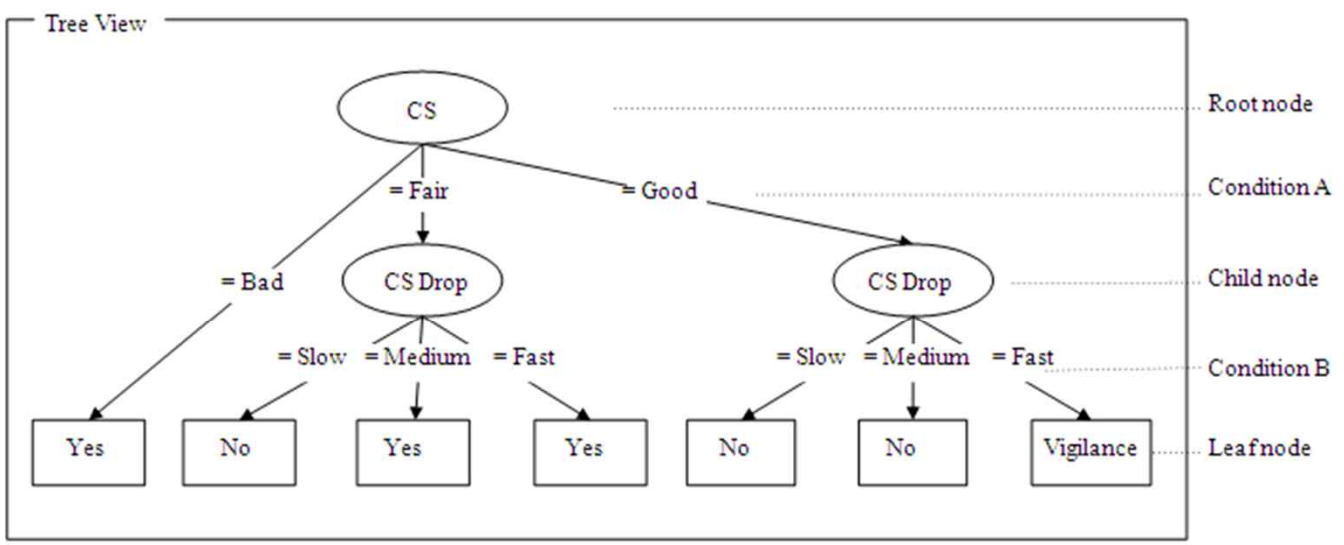
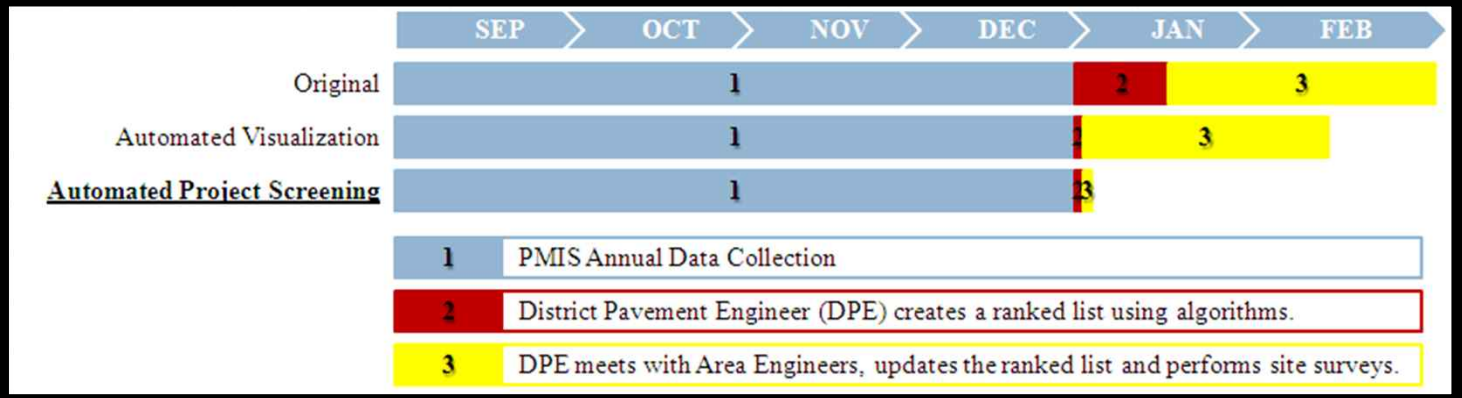


Figure 6 Visualization of the final decision tree generated by J48

Classification Accuracy: 93%

Data Mining for Pavement Management

AREA 1							AREA 2								AREA 3			AREA 4			AREA 5		AREA 6							
Project's PMIS Data							Condition Score (CS)				Condition Score Drop (CSD)				Weighting Factor			0.5	VS	0.5	Final Result		Distress Summation							
							0.6				0.4																			
Project Number	Roadway ID	Beginning Reference Marker	Displace-ment	Ending Reference Marker	Displace-ment	Weighting factor	0.4	0.3	0.2	0.1	0.1	0.2	0.3	0.4	Total Weighted CS	Total Weighted CS Drop	Total Condition (TC)	Rank by TC	Project Length (PL, Sections)	Rank by PL	Final Score (FS)	Final Rank by FS	Shallow Rut	Deep Rut	Patching	Failure	Block	Alligator	Longitudinal	Transverse
						Score Range	0≦CS<30	30≦CS<50	50≦CS<70	70≦CS	-10≦CSD	-20≦CSD<-10	-30≦CSD<-20	CSD<-30																
7	FM0812 K	548	0.5	560	0	Num of Sections	3	10	4	0	13	4	0	0	0.29	0.12	0.23	3	17	1	2	1	10	3	194	1	0	487	1674	0
1	FM0020 K	568	0.5	578	0.5	Num of Sections	3	6	5	1	11	4	0	0	0.27	0.13	0.21	8	15	2	5	2	36	5	309	2	0	25	69	0
19	US0290 K	614																								2	0	59	0	0
8	FM1100 K	560																								2	0	41	125	1
18	SL0150 K	560																								0	0	34	15	1
14	SH0021 L	564																								0	0	11	1422	4
6	FM0696 K	566																								1	0	2	0	0
17	SL0109 K	434																								0	0	26	50	1
15	SH0021 R	568																								0	0	0	769	3
13	SH0021 K	580																								2	0	26	613	0
3	FM0535 K	552																								0	0	30	275	0
12	FM3000 K	560																								1	0	11	0	0
11	FM2336 K	438																								2	0	15	10	0
20	US0290 K	626																								0	0	0	0	0
5	FM0672 K	556	0	556	0.5	Num of Sections	0	0	2	0	1	1	0	0	0.20	0.15	0.18	13	2	13	13	14	7	7	10	2	0	7	0	0
10	FM2104 K	452	0.5	454	0	Num of Sections	0	0	0	3	0	1	2	0	0.10	0.27	0.17	17	3	10	13.5	16	11	1	48	0	0	0	12	0
9	FM2104 K	446	0.5	448	0.5	Num of Sections	0	0	2	1	1	2	0	0	0.17	0.17	0.17	18	3	10	14	17	8	2	144	0	0	2	0	0
2	FM0535 K	546	0	546	0.5	Num of Sections	0	0	1	1	0	2	0	0	0.15	0.20	0.17	16	2	13	14.5	18	1	0	42	0	0	0	0	0
4	FM0535 K	574	0	574	1	Num of Sections	0	0	2	0	2	0	0	0	0.20	0.10	0.16	19	2	13	16	19	4	0	74	0	0	15	0	0
16	SH0071 R	590	1	592	0	Num of Sections	0	0	0	2	0	1	1	0	0.10	0.25	0.16	19	2	13	16	19	0	0	5	0	0	15	15	0



Step-by-Step Approach to DM

Source: Professor Sungjoon Cho (SNU Industrial Engineering, 2013)

(1) Education

- Understand data mining principles
- Formulate data-oriented thinking habit

(2) Brainstorming

- Discover topics from various disciplines (e.g., 20 data mining projects)

(3) Feasibility Analysis

- Data? Business impact? Analysis level?

(4) Data Collection and Analysis

(5) Result Review and Update

- Additional data required
- Review and correct data collection approaches
- Improve data quality
- UI/UX development for better implementation

Issues on Data Mining

Source: Professor Sungjoon Cho (SNU Industrial Engineering, 2013)

- Very New?
- 100% Accurate?
- Possible only with Data, HW/SW Infra?
- What is the good model?

- Challenges of Data Mining
  - Scalability, Dimensionality
  - Complex and heterogeneous data
  - Data quality
  - Data ownership and distribution
  - Privacy preservation