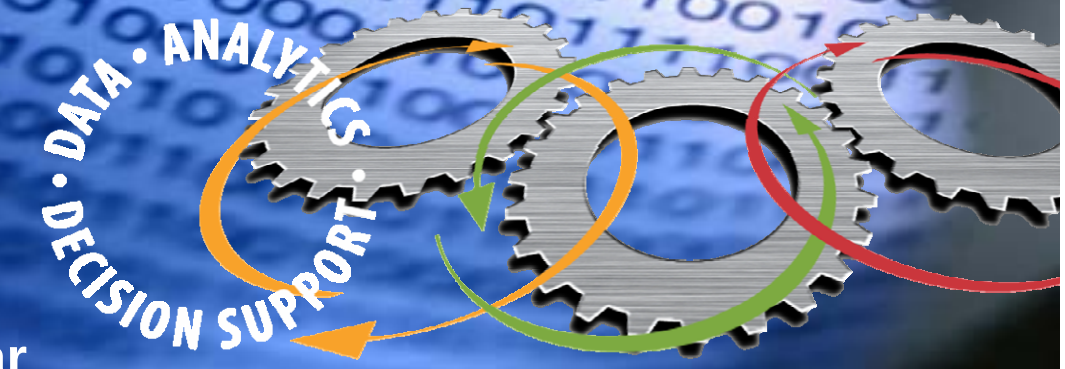
Practical Text Mining in Insurance

Marty Ellingsworth and Karthik Balakrishnan

ISO Innovative Analytics



Predictive Modeling Seminar

San Diego, 6 Oct 2008

Importance and Relevance of Text

Accident: 170824130 - Employee Injured In Fall From Second-Floor Decking

Inspection	Open Date	SIC	Establishment Name
127366367	07/29/1996	1521	

Employee #1 was atop of the second floor decking of a newly constructed home, connecting frame work for a wall. He fell 18 ft 6 in., sustaining injuries that required hospitalization. Employee #1 was not tied off, nor were any other means of fall protection in use. He had not been trained in working from an elevated work surface, the company did not have a written safety program, and regular inspections were not performed.

Keywords:	**decking, fall, tie-off, untrained, work rules, fall protection, construction**

	Inspection	Age	Sex	Degree	Nature	Occupation
1	127366367	29	M	Hospitalized injuries	Cut/Laceration	Carpenters

Source: U.S. Department of Labor Occupational Safety & Health Administration

Accident Report Detail Accident Investigation Summaries (OSHA-170 form) which result from OSHA accident inspections.

Importance and Relevance of Text

Accident: 170824130 - Employee Injured In Fall From Second-Floor Decking

Inspection	Open Date	SIC	Establishment Name
127366367	**07/29/1996**	**1521**	

not tied off, nor were any other means of fall protection in use**.**

He had not been trained in working from elevated work surface

the company did not have a written safety program, **and**

regular inspections were not performed.

Keywords:	**decking, fall, tie-off, untrained, work rules, fall protection, construction**

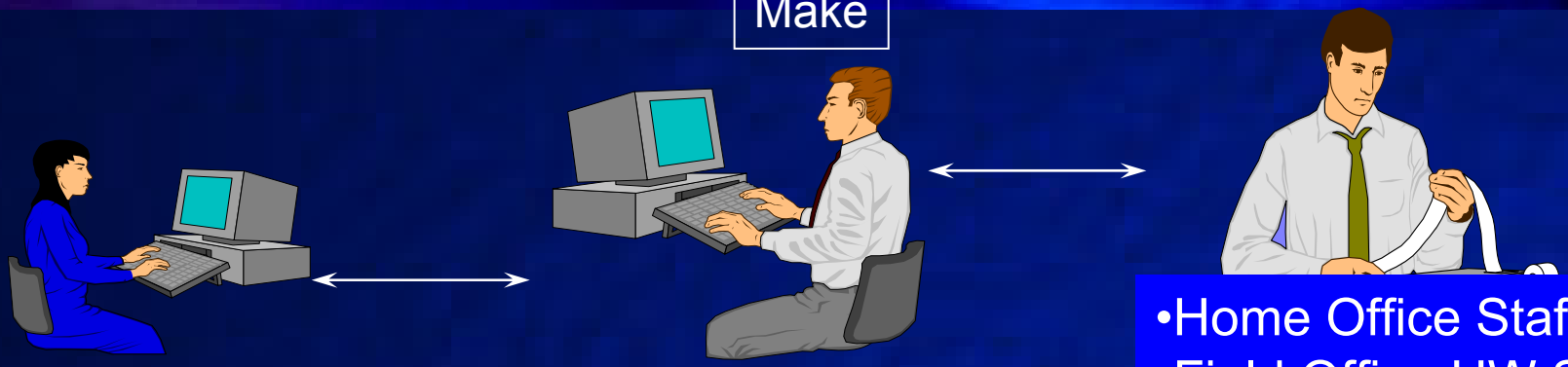
	Inspection	Age	Sex	Degree	Nature	Occupation
1	127366367	29	M	Hospitalized injuries	Cut/Laceration	Carpenters

Source: U.S. Department of Labor Occupational Safety & Health Administration

Accident Report Detail Accident Investigation Summaries (OSHA-170 form) which result from OSHA accident inspections.

Policy Processing
Underwriting Notes and Diaries

Make



- D&B Data
- ISO Data
- Application information
- Claim loss runs
- Hazard mappings
- Concentrations of Staff
- Premium Auditors
- Renewal processing
- Legal Staff
- …others

DESK UNDERWRITER

- Home Office Staff
- Field Office UW Staff
- Insured Risk Manager
- Agent or Broker

- Diary forward – “call Agency next week”
- Business Rule – large loss review
- System Reminder – update renewal pricing
- Correspondence Tracking – legal letter sent

Customer Management
Contact Notes and Diaries

Sell

- Voice of the Customer
- Customer Feedback
- Call Center Notes
- Agent Contacts
- Billing Systems
- Deductible Processing
- Premium Auditors
- Renewal processing

ACCOUNT MANAGER

- Company-wide Sales Staff
- Product Manager
- Insured Risk Manager
- Agent or Broker

- Diary forward – “call Mr Jones tonight”
- Business Rule – DOI Complaint handling
- System Reminder – Visit with Client
- Correspondence Tracking – legal letter sent

Claims Processing
Progress Notes and Diaries

Service

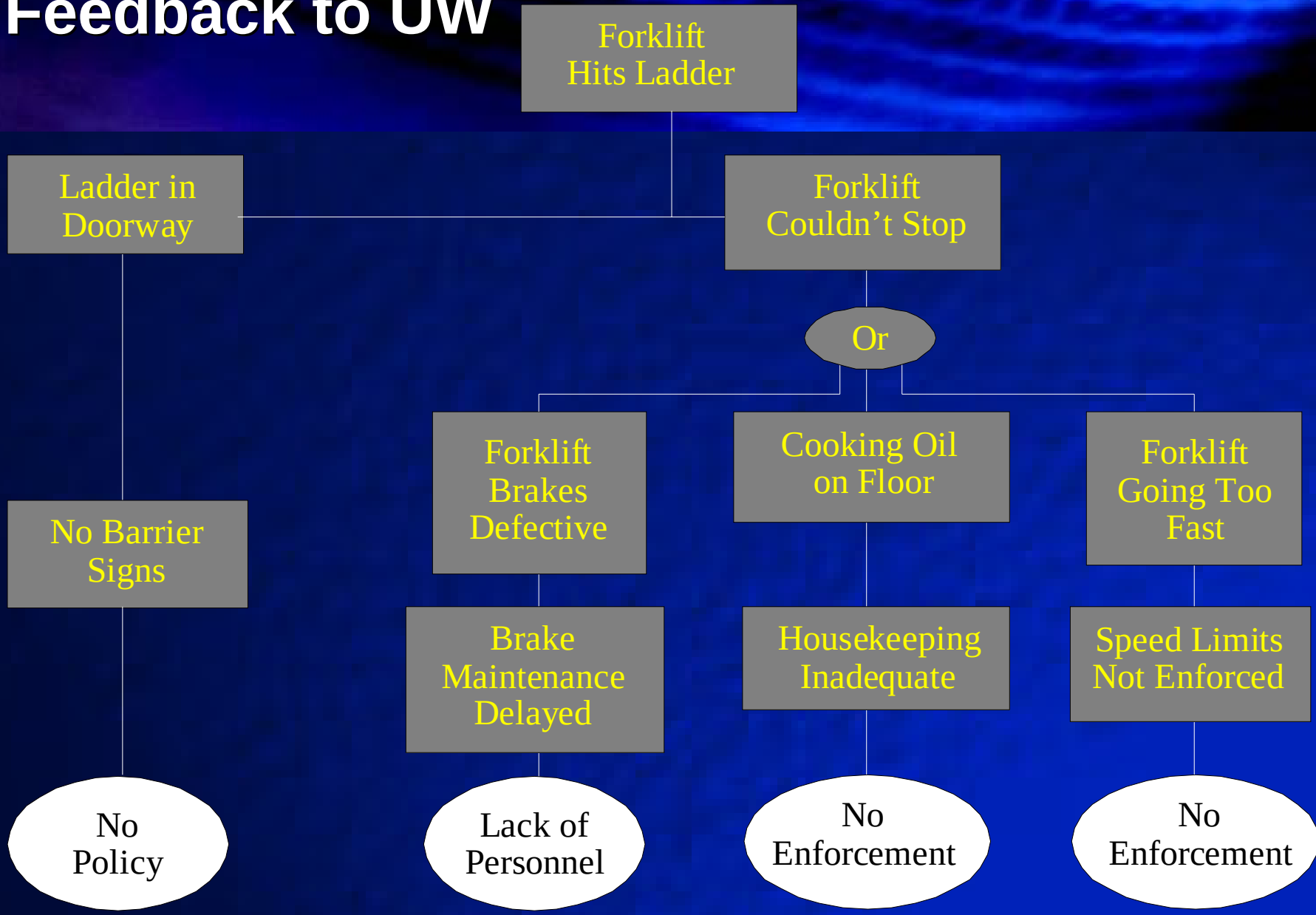
- Medical Management Staff
- Special Investigation Unit
- NICB
- Vendor Management
- Consulting Engineers
- Hearing Representative
- Structured Settlement Unit
- Recovery Staff
- Legal Staff

CLAIMS ADJUSTER

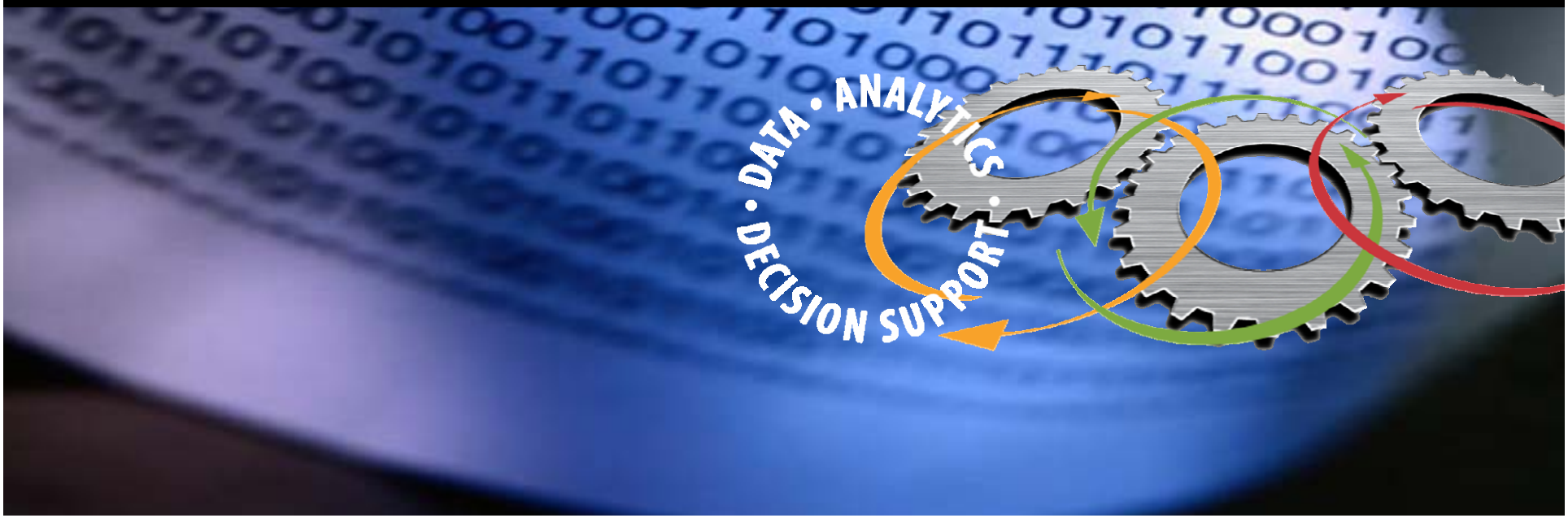
- Home Office Staff
- Field Office Claim Staff
- Insured Risk Manager
- Agent or Broker

- Diary forward – “call Dr Jones next week”
- Business Rule – large loss review
- System Reminder – update case reserves
- Correspondence Tracking – legal letter sent

Feedback to UW



Text Mining in Action



Play the SIU Triage Game –

IT APPEARS THAT THIS WAS A LOW IMPACT COLLISION WHERE THE INSURED'S FOOT SLIPPED OFF THE BRAKE AND SHE ROLLED INTO THE REAR OF THE CLAIMANT. THIS IS CONSSTENT WITH THE FACT THAT THERE WAS NO PROPERTY DAMAGE CLAIM MADE TO THE CLAIMANT VEHICLE. UNDER THESE CIRCUMSTANCES, HOW THE CLAIMANT COULD HAVE SUSTAINED SUCH SEVERE SHOULDER INJURIES AS A RESTRAINED DRIVER APPEARS RATHER SUSPECT.

NO PROP DMG FOR INS AND CLMT AS COLL IMPACT WAS LOW. CLMT CLAIMS INJ FROM AX AND TREATED WITH CP AND PT EXTENSIVELY. TX APPEARS EXAGGERATED

INSURED WAS RUBBER-NECKING AND DID NOT REALIZE TRAFFIC HAT STOPPED. HE RAN INTO JOHN AT 50-60 MPH, CAUSING THE CLAIMANT FORD FESTIVA TO COMPLETELY BUCKLE IN. JOHN HAD SERIOUS WHIPLASH INJ AND WAS AMBULANCED TO A HOSP ALONG WITH THE INSURED.

CLAIMANT WAS VISITED BY THREE SPECIALISTS, WHICH IS NOT EXCESSIVE FOR THIS TYPE OF INJURY.

Congratulations! How did you do it?

NO PROP DMG FOR INS AND CLMT AS COLL IMPACT WAS LOW. CLMT CLAIMS INJ FROM AX AND TREATED WITH CP AND PT EXTENSIVELY. TX APPEARS EXAGGERATED

1. **Read and parse sentences into words**
2. **Knew meanings of words and phrases, in context** подозрительный
3. **Made intelligent guesses on abbreviations and typos**
4. **Identified “concepts” and their relevance to Fraud/Suspicion detection**
5. **Flagged claims containing certain combinations of concepts**

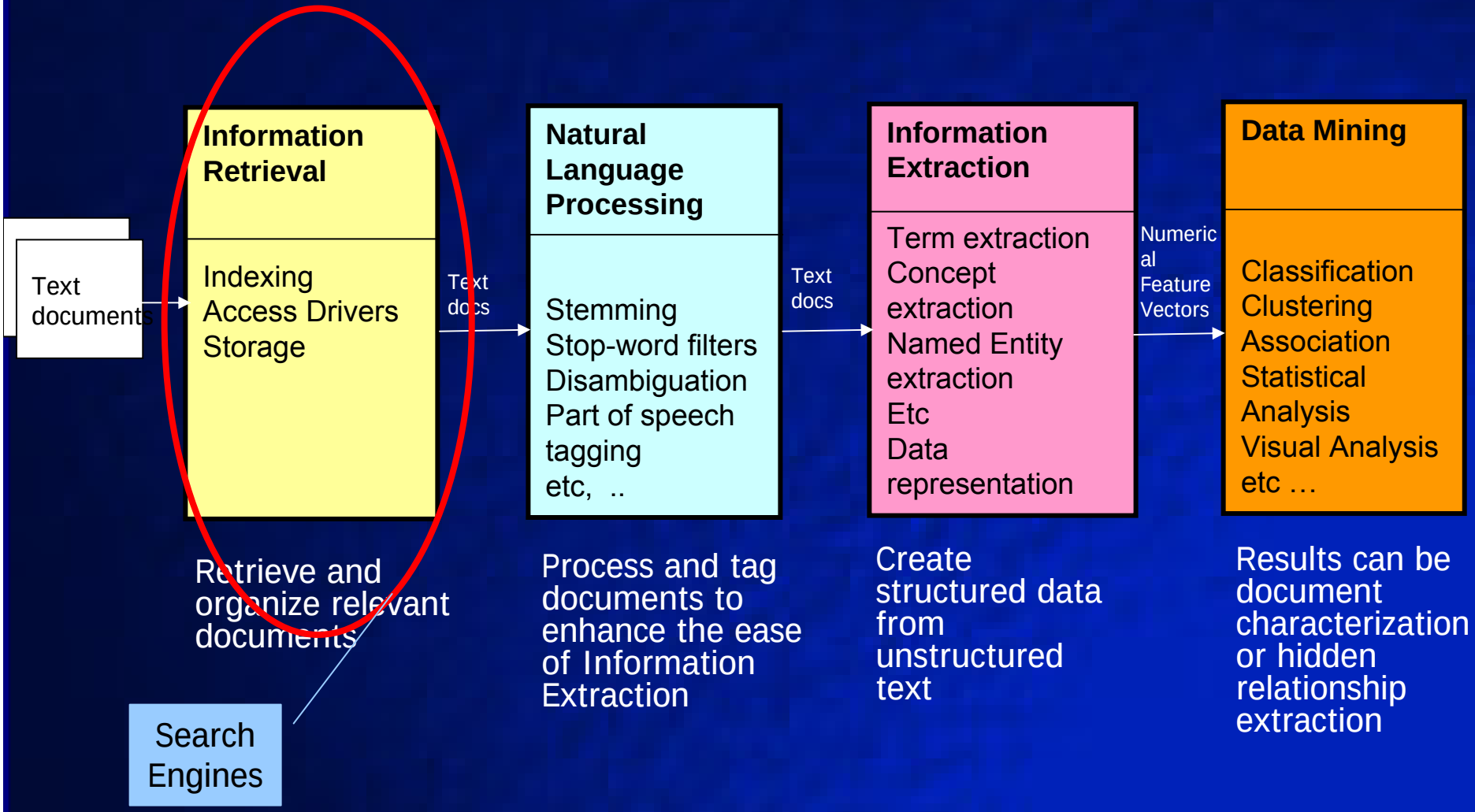
State of Text Mining Technology

1. **Read and parse sentences into words and components**
   - Language-based parsers and tokenizers
   - Stemming –
     - suspicious, suspiciously, suspicion, suspiciousness → suspicion
   - Stop-word removal – to, a, an, of, etc.
2. **Generate meanings of words and phrases, in context**
   - Dictionaries and thesauri
   - Word disambiguation based on context
   - Natural Language Processing (NLP) technology
     - Part of speech tagging (e.g., nouns, verbs, etc.)
3. **Make intelligent guesses on abbreviations and typos**
   - Valid word lists, abbreviation lists, etc. (domain dependent)

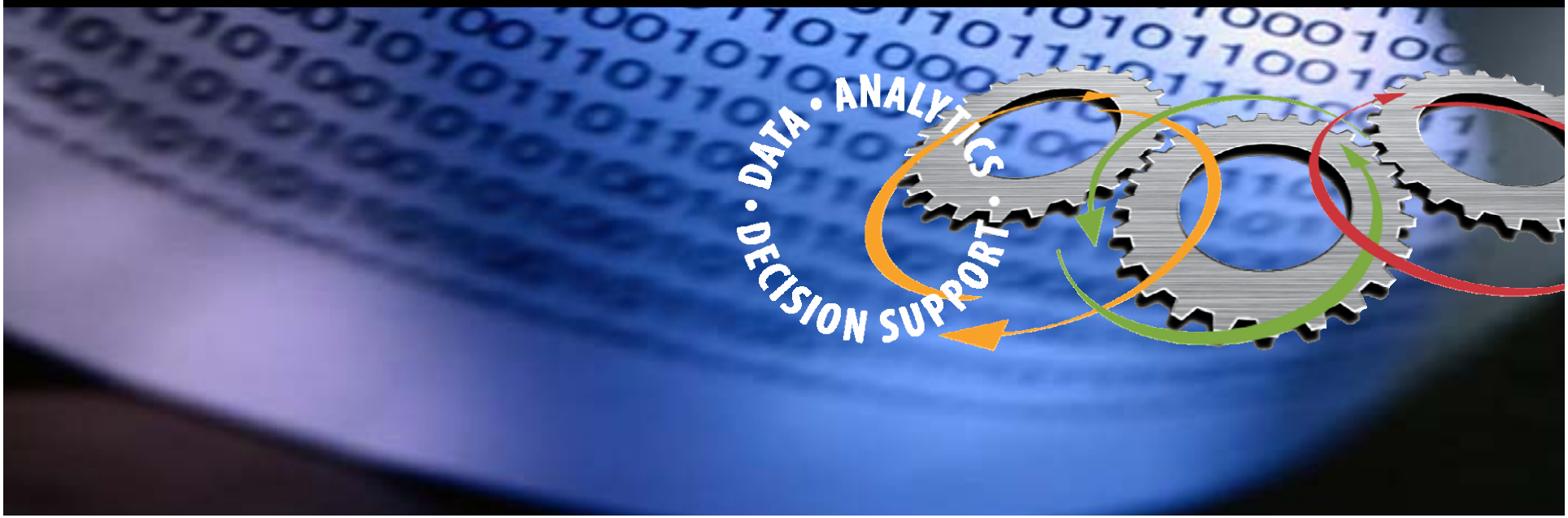
State of Text Mining Technology

4. **Identify "concepts" and their relevance to Problem**
   - Domain-knowledge driven
   - Inductive – semi-automated learning based on labeled examples
5. **Represent concepts in a form suitable for analysis**
   - Vectors of terms, concepts, etc.
   - Typically numeric or flags
6. **Build/discover interesting combinations of concepts**
   - Miscellaneous predictive, descriptive and analytical methodologies

Components of Text Mining



Simple, Practical Text Mining



FIRE Engine Algorithm
Fine-grained Information Retrieval and Evaluation

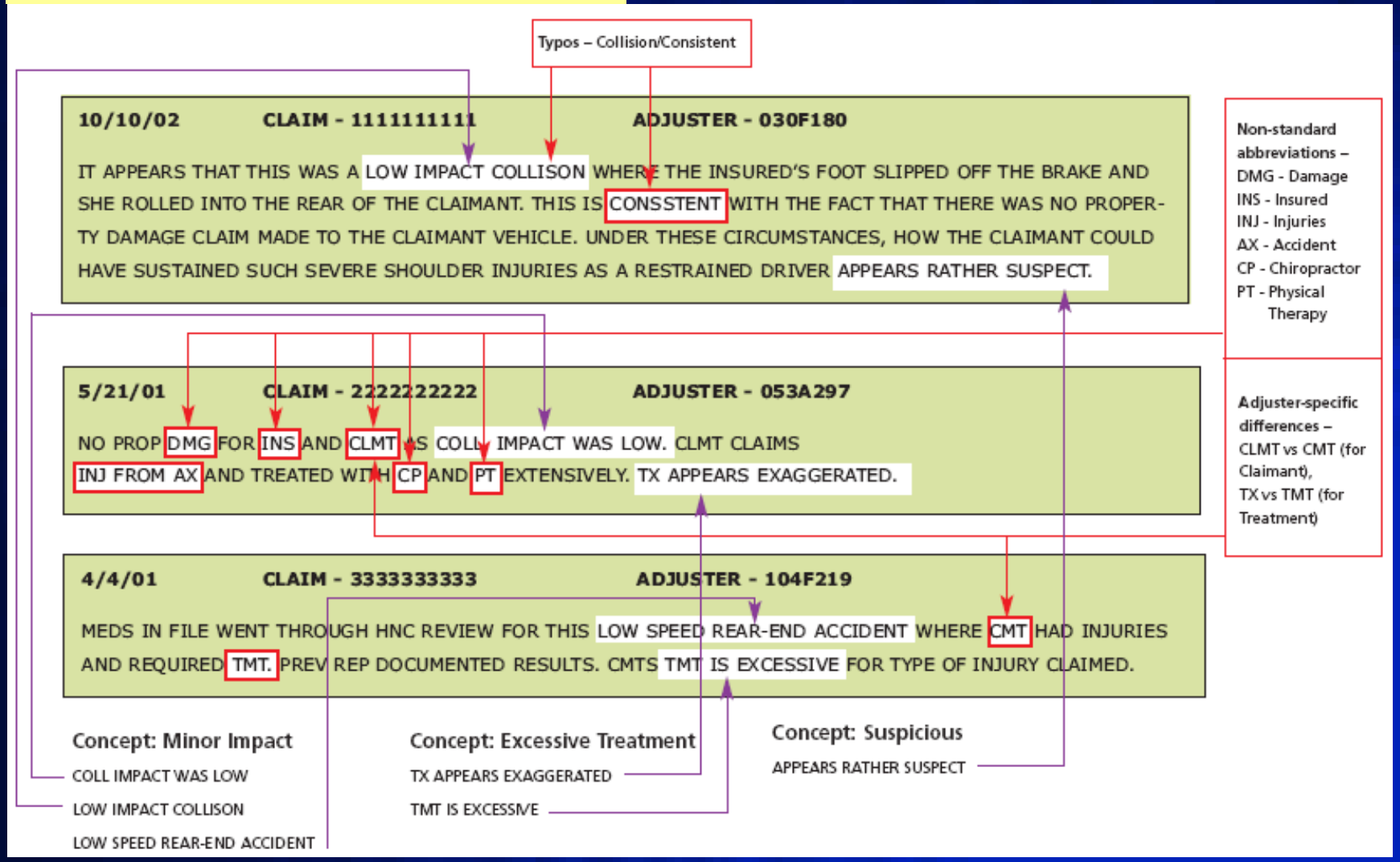
Steps	
1. Determine the Goal	This is the business problem that we would like to "structurize" for
2. Goal-target Labeling	Label each document in the corpus with a Target value corresponding to the Goal. For a binary classification problem, the values are Target=1 and Target=0.
3. Phrase Extraction and Labeling	Extract one-, two-, three-, etc. word phrases from the corpus and label them with Precision, Recall and F-Measure statistics – a) Precision – (# of Target=1 documents containing the phrase)/Total # of documents containing the phase
	b) Recall – (# of Target=1 documents containing phrase)/Total # of Target=1 documents
	c) F-measure – (β^2+1).Precision.Recall/ $(\beta^2$.Precision + Recall)
4. Seed List Generation	Domain experts can provide a *seed list* of words/phrases that are typically associated with the given goal
5. Context-Driven Phrase Extraction and Augmentation	a) Identify *accurate* one-, two-, three-, etc. word phrases from the phrase list that contain elements of the seed list (this brings out the various contexts in which specific phrases appear)
	b) Augment seed list with novel words and phrases identified in a)
	c) Repeat these steps until no more accurate novel words and phrases can be found
6. Generalization or Phrase Pruning	Prune the extracted phrases, retaining shorter (more general) phrases of similar precision but higher recall, where possible
7. Semanticization	Group remaining phrases into "semantic" categories or CONCEPTS (possibly involving domain experts)
8. Structurization	Create a structured data element to represent each concept, driven by the various syntactic flavors of the identified words and phrases

Balakrishnan et al. "Enhancing Knowledge Discovery Using Text Mining"

American Marketing Association – Advanced Research Techniques Forum, 2002

Unstructured Data Challenges

Problem – Fraud/Suspicion Detection



Target Labeling and Phrase Extraction

Label each document with its corresponding Target value, i.e., fraud or non-fraud (1 or 0)

Claim	Target
1111111111	1
2222222222	1
3333333333	0
...	
6666666666	
...	

Labeled 1/2/3 word phrases with their Precision, Recall and F-Measure (Relative Strength) statistics

Universe of Labeled 1/2/3 word Phrases

	Precision	Recall	Relative Strength
LOW	10.6%	63.9%	0.107
IMPACT	5.9%	89.3%	0.060
CP	84.0%	0.6%	0.354
PT	51.0%	2.0%	0.410
PT EXTENSIVELY	67.0%	1.0%	0.405
EXTENSIVELY TX	49.1%	3.1%	0.428
TX APPEARS	37.3%	2.8%	0.332
APPEARS EXAGGERATED	81.4%	4.1%	0.686
NO PROP DMG	58.9%	13.7%	0.570
IMPACT WAS LOW	54.3%	12.7%	0.526
TREATED WITH CP	92.7%	0.9%	0.459
CP AND PT	96.4%	0.3%	0.231
TX APPEARS EXAGGERATED	89.7%	1.9%	0.615
EXCESSIVE TREATMENT	79.6%	3.9%	0.668
INFLATING BILL	94.2%	1.7%	0.612
MED BUILDUP	95.2%	0.9%	0.467
BUILD UP CASE	88.7%	0.4%	0.278
QUESTIONABLE INJURY	72.9%	4.5%	0.634
QUESTIONABLE TREATMENT	78.5%	3.8%	0.657
EXCESSIVE TX	84.5%	1.1%	0.483
EXCESSIVE TMT	81.7%	1.0%	0.452
QUESTIONABLE TRMNT	82.4%	2.1%	0.598

Note: Higher Relative Strength is better
We use $\beta = 0.25$

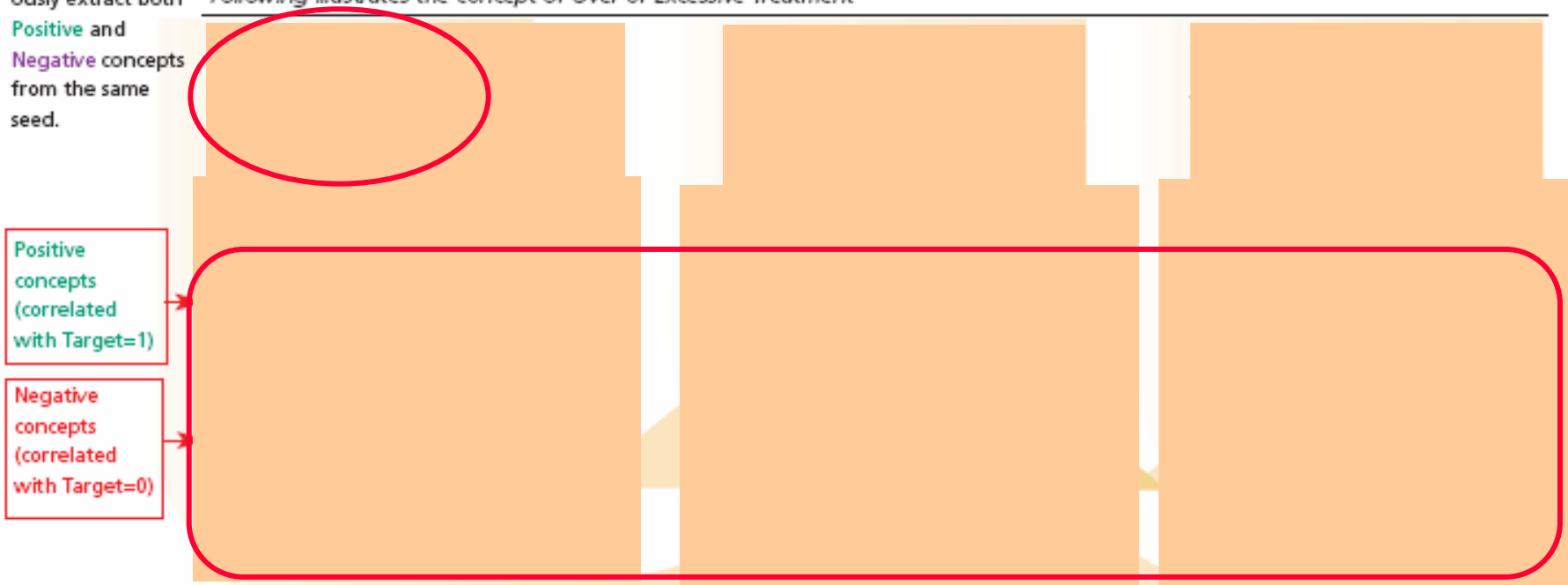
Context-Driven Phrase Extraction and Augmentation Using Seed Lists

Begin with seed list (if available) provided by domain experts and iteratively augment and discover novel phrases of predictive value

Note - By using (1 - Precision) and a Recall/F-measure for Target=0, we can simultaneously extract both Positive and Negative concepts from the same seed.

Seed Concepts *(From Domain Expert, if available)*
Over or excessive treatment
Minor impact
Soft tissue injuries, etc

Following illustrates the concept of Over or Excessive Treatment



Novel words/phrases shown in purple

Generalization or Phrase Pruning

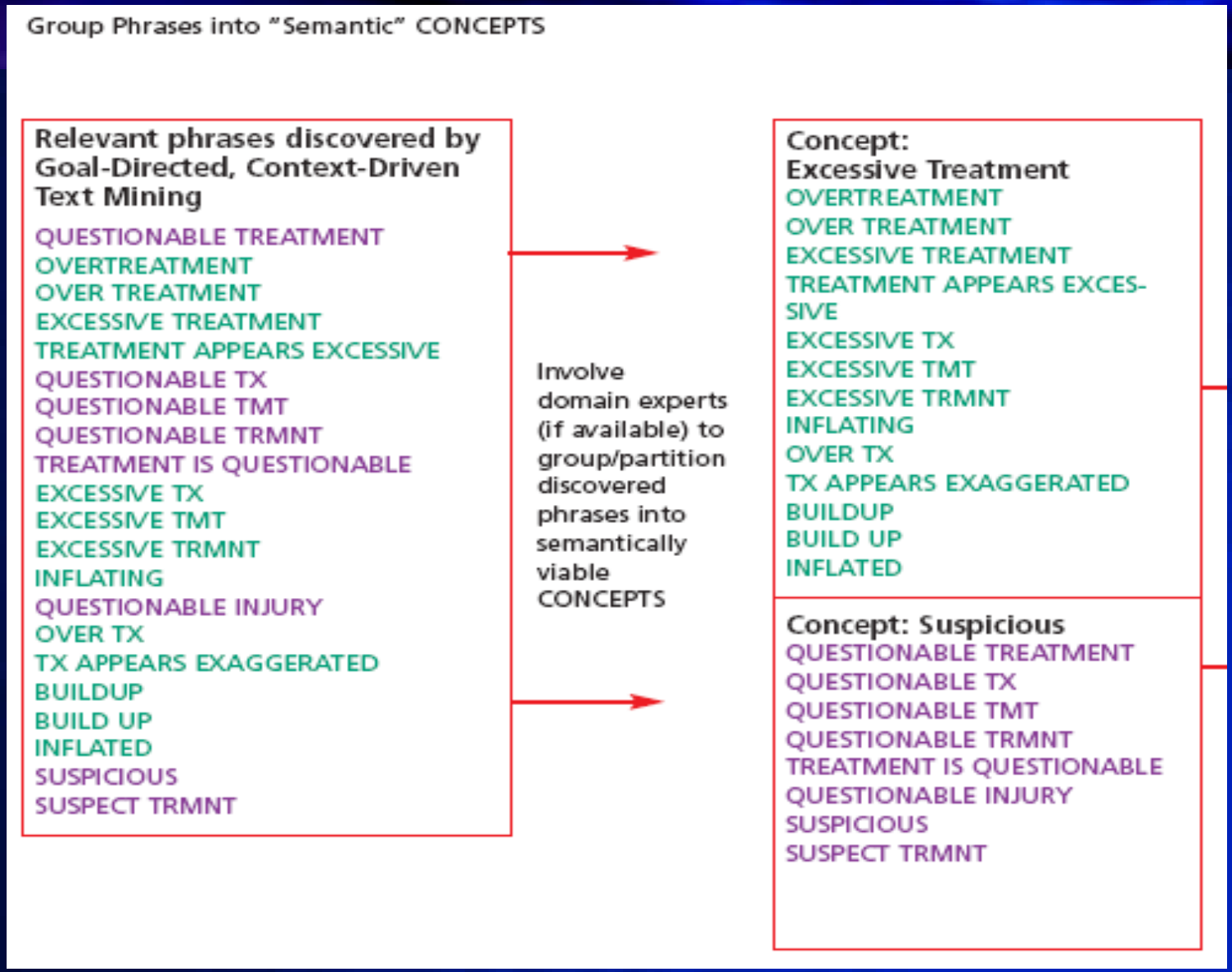
Retain shorter (hence more general) phrases of similar precision but higher recall, where possible

Table A

QUESTIONABLE TREATMENT
VERY QUESTIONABLE TREATMENT
WAS QUESTIONABLE TREATMENT
IS QUESTIONABLE TREATMENT
QUESTIONABLE TREATMENT EXISTS
QUESTIONABLE TREATMENT IN
QUESTIONABLE TREATMENT ON
QUESTIONABLE TREATMENT OF
QUESTIONABLE TREATMENT THAT
QUESTIONABLE TREATMENT FROM
FOR QUESTIONABLE TREATMENT
OF QUESTIONABLE TREATMENT
IN QUESTIONABLE TREATMENT
ON QUESTIONABLE TREATMENT

This "reduced" phrase is more "general" and covers all the phrases in Table A

Semanticization



Structurization

Embed Discovered Phrases into Text Matching Rules to Produce a Structured Representation of the Concept

Concept: Excessive Treatment

EXCESSIVE_TREATMENT = ?
IF document contains any of the following phrases – "OVERTREATMENT" "OVER TREATMENT" "EXCESSIVE TREATMENT" "TREATMENT APPEARS EXCESSIVE" "EXCESSIVE TX" "EXCESSIVE TMT" "EXCESSIVE TRMNT" "INFLATING" "OVER TX" "TX APPEARS EXAGGERATED" "BUILDUP" "BUILD UP" "INFLATED"
THEN EXCESSIVE_TREATMENT = 'YES'

IF document contains any of the following phrases – "NO OVERTREATMENT" "NO OVER TREATMENT" "NO EXCESSIVE TREATMENT" "EXCESSIVE TREATMENT NOT" "TREATMENT AS EXPECTED" "TX BILLS VALID" "TMT AS EXPECTED" "TMT IN LINE WITH" "NO BUILDUP"
THEN EXCESSIVE_TREATMENT = 'NO'

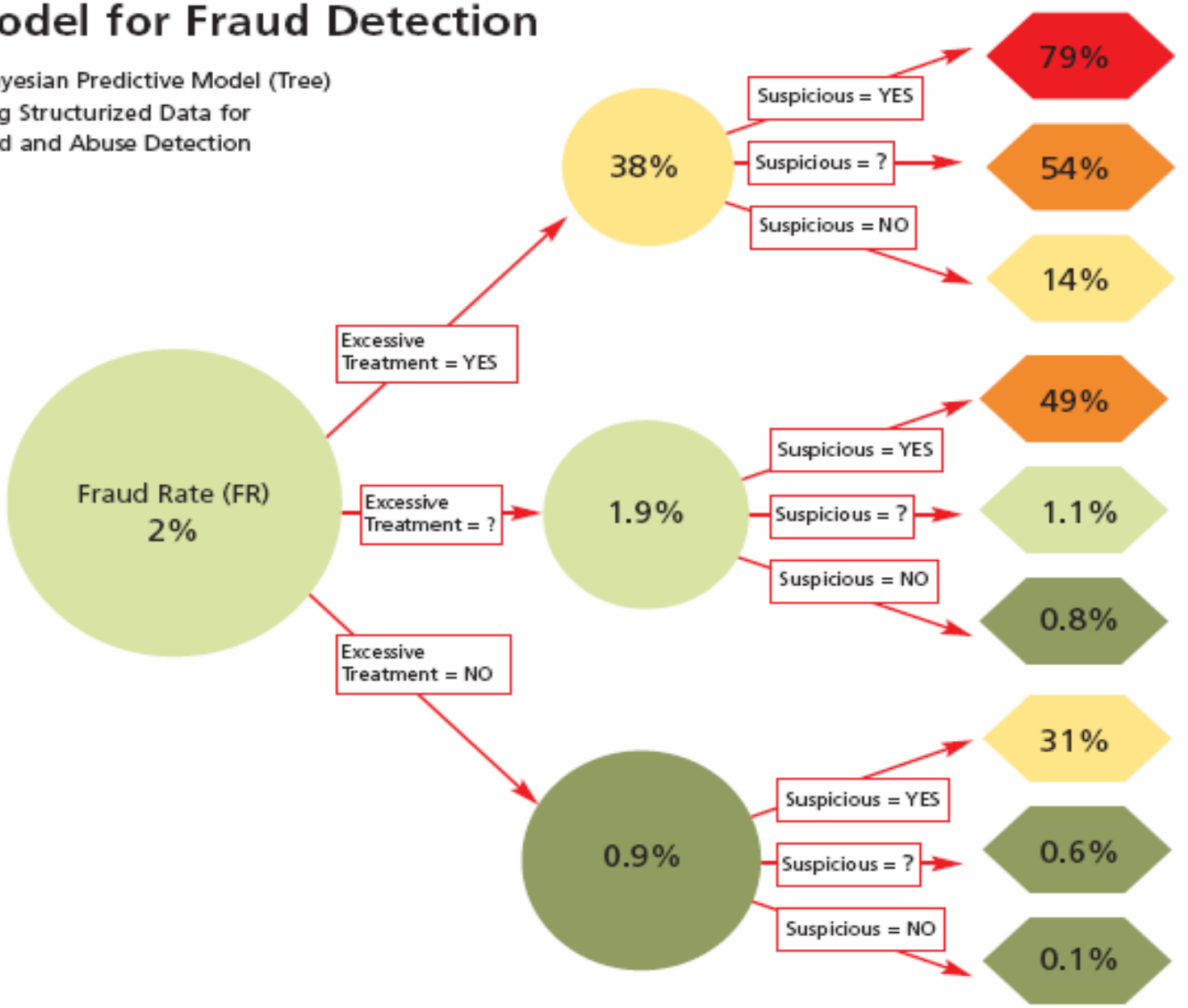
Concept: Suspicious

SUSPICIOUS = ?
IF document contains any of the following phrases – "QUESTIONABLE TREATMENT" "QUESTIONABLE TX" "QUESTIONABLE TMT" "QUESTIONABLE TRMNT" "TREATMENT IS QUESTIONABLE" "QUESTIONABLE INJURY" "SUSPICIOUS" "SUSPECT TRMNT" THEN SUSPICIOUS = 'YES'

IF document contains any of the following phrases – "NO QUESTIONABLE TREATMENT" "NO QUESTIONABLE TX" "QUESTIONABLE TMT NOT" "QUESTIONABLE TRMNT NOT" "TREATMENT IS NOT QUESTIONABLE" "NO QUESTIONABLE INJURY" "NOTHING SUSPICIOUS" "VALID TRMNT" THEN SUSPICIOUS = 'NO'

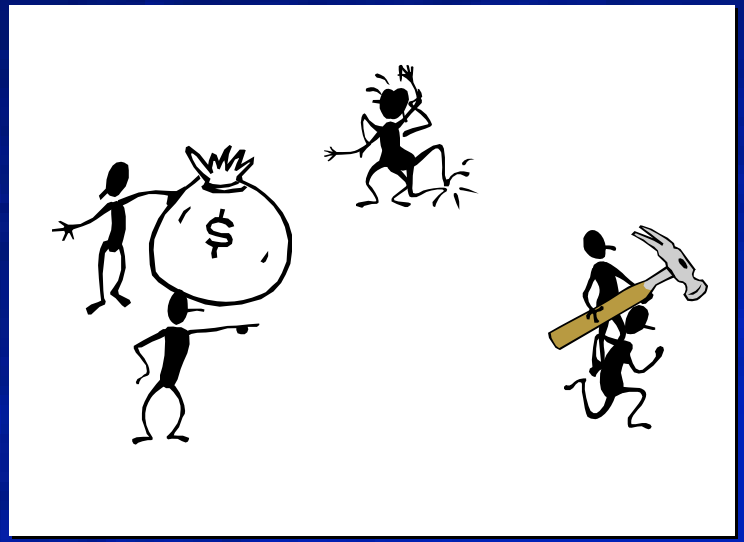
Model for Fraud Detection

A Bayesian Predictive Model (Tree)
Using Structurized Data for
Fraud and Abuse Detection



Subrogation Opportunity Identification

- **What is Subrogation?**
  - Insured suffers a loss
  - Insurance Company settles loss
  - Another party responsible/liable for the loss (or part of the loss)
  - Insurance Company subrogates against the Other Party/Carrier



Subrogation Concept – OP Unidentified

- **If ANY of the following phrases occur, set the "concept"**
  - OP_Unidentified = 1
  - Otherwise OP_Unidentified = 0

NO SUSPECTS		SUSPECTS UNK	
UNK SUSPECTS		UNKNOWN SUSPECT	
SUSPECTS NOT		NO KNOWN SUSPECT	
UNIDENTIFIED SUSPECT		NO IDENTIFIABLE SUSPECT	
I/D UNK	NO I/D	NO ID	UNK PER
UNKNOWN BROKE		UNK STOLE	
UNKNOWNS BROKE		TORTFEASOR UNKNOWN	
HIT AND RUN		SOMEONE BROKE	…

Subrogation Concept – OP Identified

- **If ANY of the following phrases occur, set "concept"**
  - OP_Identified = 1
  - Otherwise, OP_Identified = 0

SUSPECTS APPREHENDED	**SUSPECTS KNOWN**
KNOWN SUSPECTS	**ARRESTS SUSPECTS**
SUSPECTS ARREST	**SUSPECTS CAUGHT**
SUSPECTS ID'D	**ID'D SUSPECTS**
ID'ED SUSPECTS	**SUSPECTS LOCATED**
SUSPECTS NAMED	**IDENTIFIED SUSPECTS**
SUSPECTS IDENTI	**SUSPECTS CHARGED**
TF/CARRIER ID …	

Creating Subrogation "Stories"

- Seven key concepts
- Each concept is represented by a binary flag (1=present, 0=otherwise)

Insured At Fault	OP Unidenti fied	Subro Ruled Out	Pure Loss	OP At Fault	OP Identified	Subro Exists

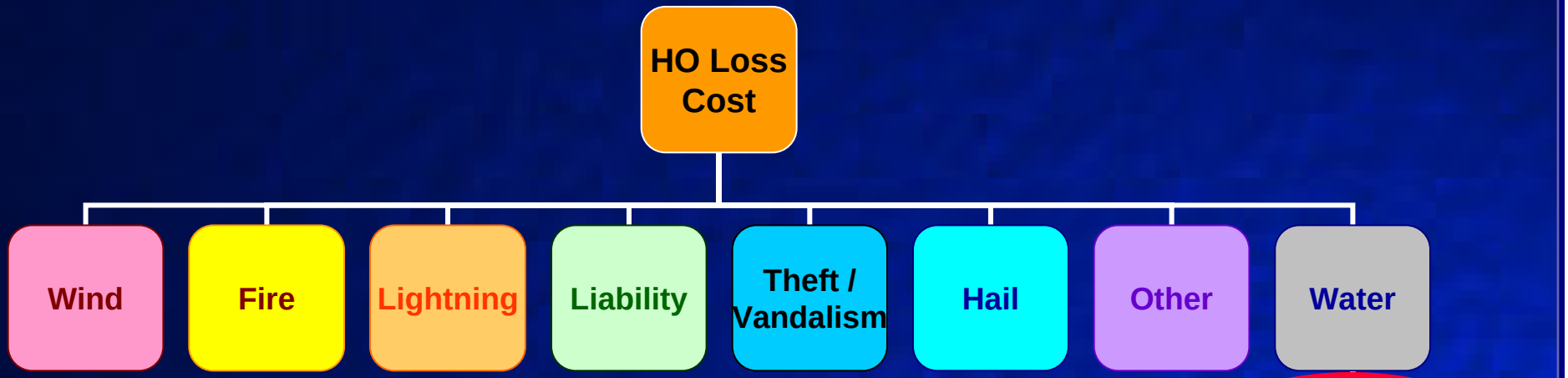
- Each vector state is a Subrogation "Story". E.g.,
  - 1010000 = Insured At Fault and Adjuster Ruled out Subro
  - 0000111 = OP At Fault, OP Identified, and Adjuster assessed Subro
    - But never referred the claim to the Subro Recovery Unit!

Referral Using Subrogation Stories

- **Determine Subrogation Story for a new claim**
- **If Story has HIGH historical Recognition/Hit rates, refer to Recovery Unit**

	SUBROGATION STORY	CLAIMS WITH THE STORY	RECOGNITION RATE	HIT RATE	$ LOSS PAID	$ RECOVERY
LOW	0001000	12,558	0.8	5.2	$81,809,336	$2,836
	0000000	11,790	1.0	17.4	$83,504,471	-$6,438
	0010000	12,740	1.6	4.9	$69,062,230	-$8,014
	0011000	30,006	2.1	1.8	$167,154,325	$11,015
	0111000	21,364	3.6	4.2	$220,820,511	$47,225
	We did a good job of capturing the "concept" of Subro Ruled Out					
HIGH	0001011	1,422	93.8	35.1	$76,902,215	-$2,873,929
	0000111	1,912	98.9	66.9	$73,035,092	-$8,118,833
	0001111	1,425	98.9	53.2	$102,310,964	-$6,785,945
	The "concept" of Subro Exists when Other Party is Identified, was also well captured					

Text Mining in U/W



Decompose HO losses and model by Peril

Will result in "tighter" models

Text Mining for Cause-Of-Loss

- Rich information buried in Unstructured data, such as Loss Descriptions or Adjuster Notes
- E.g., Extracting the “Type of Loss” from the Loss Description

Questions?

- **Contacts @ ISO Innovative Analytics**
  - Marty Ellingsworth
    - President
    - mellingsworth@iso.com
  - Karthik Balakrishnan
    - Vice President, Analytics
    - kbalakrishnan@iso.com