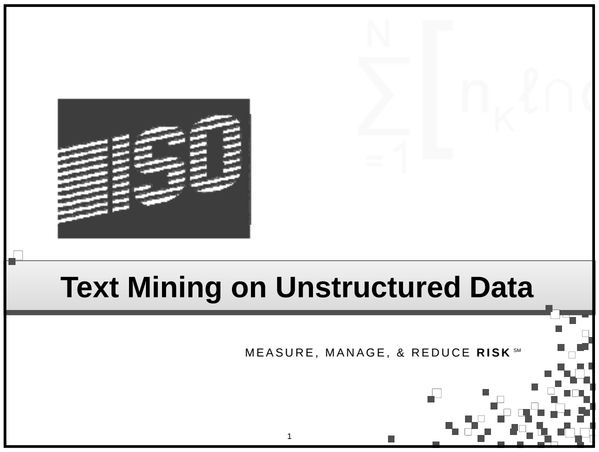
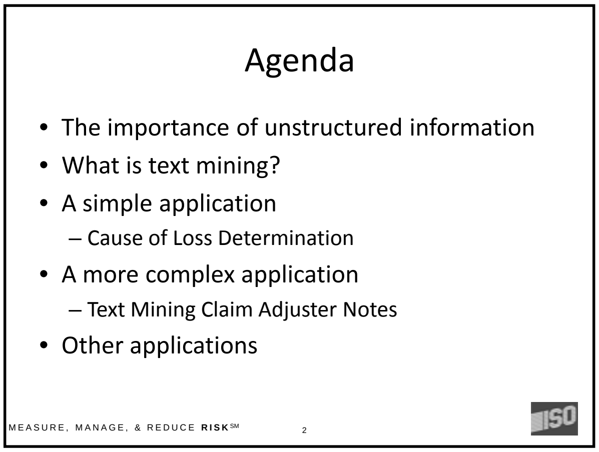


## Text Mining on Unstructured Data

MEASURE, MANAGE, & REDUCE **RISK** SM

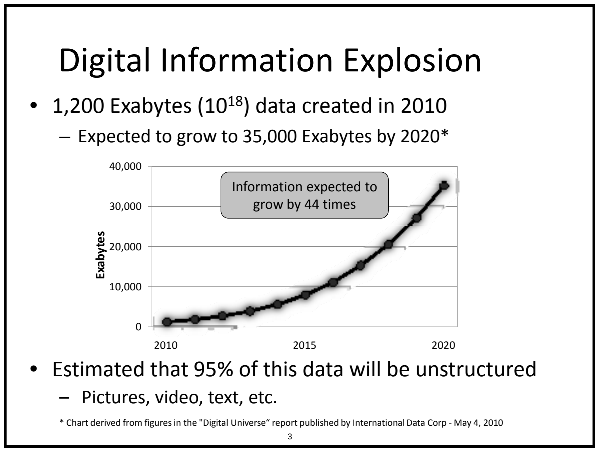
1



## Agenda

- The importance of unstructured information
- What is text mining?
- A simple application
  - – Cause of Loss Determination
- A more complex application
  - – Text Mining Claim Adjuster Notes
- Other applications

MEASURE, MANAGE, & REDUCE **RISK** SM 2



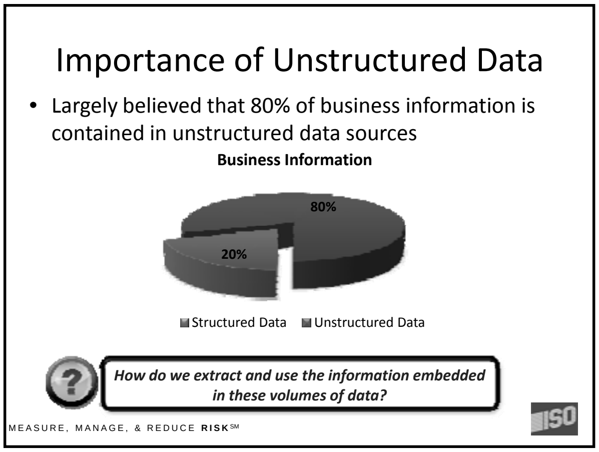
## Digital Information Explosion

- 1,200 Exabytes ($10^{18}$) data created in 2010
  - – Expected to grow to 35,000 Exabytes by 2020*

- Estimated that 95% of this data will be unstructured
  - – Pictures, video, text, etc.

* Chart derived from figures in the "Digital Universe" report published by International Data Corp - May 4, 2010

3



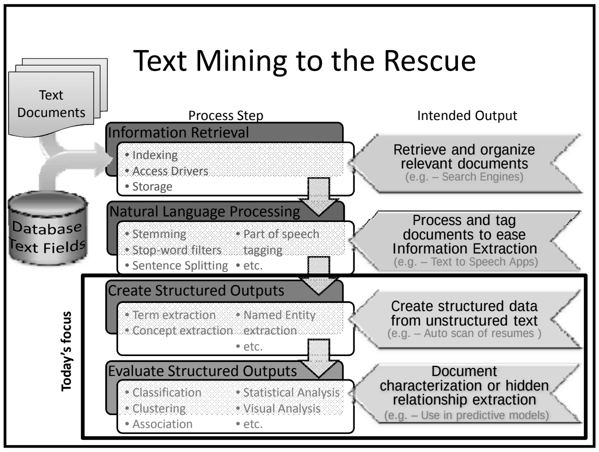
## Importance of Unstructured Data

- Largely believed that 80% of business information is contained in unstructured data sources

**Business Information**

***How do we extract and use the information embedded in these volumes of data?***

MEASURE, MANAGE, & REDUCE **RISK** SM



## Text Mining to the Rescue

Text Documents

Database Text Fields

Process Step

Intended Output

Information Retrieval
- Indexing
- Access Drivers
- Storage

Retrieve and organize relevant documents (e.g. – Search Engines)

Natural Language Processing
- Stemming
- Stop-word filters
- Sentence Splitting
- Part of speech tagging
- etc.

Process and tag documents to ease Information Extraction (e.g. – Text to Speech Apps)

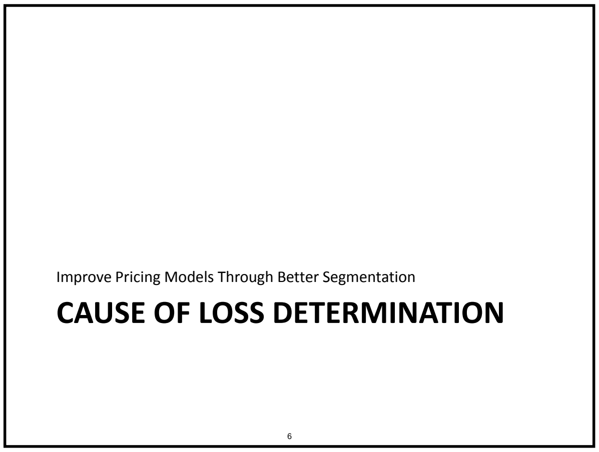
Today's focus

Create Structured Outputs
- Term extraction
- Concept extraction
- Named Entity extraction
- etc.

Create structured data from unstructured text (e.g. – Auto scan of resumes )

Evaluate Structured Outputs
- Classification
- Clustering
- Association
- Statistical Analysis
- Visual Analysis
- etc.

Document characterization or hidden relationship extraction (e.g. – Use in predictive models)

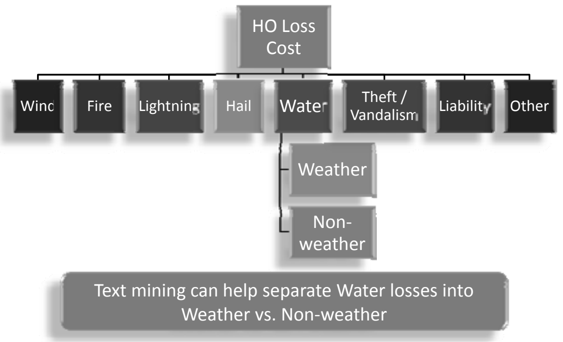


Improve Pricing Models Through Better Segmentation

## CAUSE OF LOSS DETERMINATION

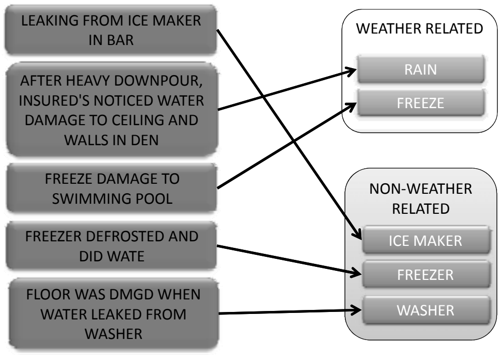
6

## Accurate Models – Build at Peril-Level

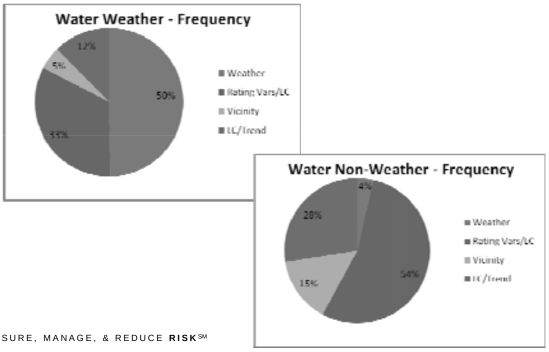


Text mining can help separate Water losses into Weather vs. Non-weather

## Text Mining HO Loss Descriptions


LEAKING FROM ICE MAKER IN BAR
AFTER HEAVY DOWNPOUR, INSURED'S NOTICED WATER DAMAGE TO CEILING AND WALLS IN DEN
FREEZE DAMAGE TO SWIMMING POOL
FREEZER DEFROSTED AND DID WATE
FLOOR WAS DMGD WHEN WATER LEAKED FROM WASHER
WEATHER RELATED
RAIN
FREEZE
NON-WEATHER RELATED
ICE MAKER
FREEZER
WASHER

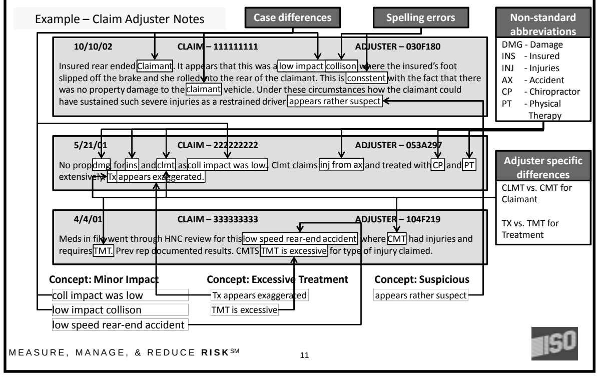
## More Relevant Predictors



Process for Mining Claim Adjuster Notes

## TEXT MINING FOR SUSPICIOUS CLAIMS

## Challenges and Opportunities in Text


Example – Claim Adjuster Notes
Case differences
Spelling errors
Non-standard abbreviations
Adjuster specific differences

10/10/02 CLAIM – 111111111 ADJUSTER – 030F180

Insured rear ended Claimant. It appears that this was a low impact collison 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 to the claimant vehicle. Under these circumstances how the claimant could have sustained such severe injuries as a restrained driver appears rather suspect

5/21/01 CLAIM – 222222222 ADJUSTER – 053A297

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.

4/4/01 CLAIM – 333333333 ADJUSTER – 104F219

Meds in file went through HNC review for this low speed rear-end accident where CMT had injuries and requires TMT. Prev rep documented results. CMTS TMT is excessive for type of injury claimed.

| Non-standard abbreviations | |
|---|---|
| DMG | - Damage |
| INS | - Insured |
| INJ | - Injuries |
| AX | - Accident |
| CP | - Chiropractor |
| PT | - Physical Therapy |

Adjuster specific differences

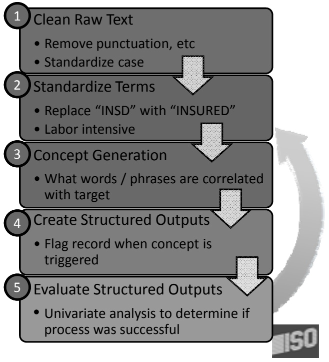
CLMT vs. CMT for Claimant

TX vs. TMT for Treatment

| Concept: Minor Impact | Concept: Excessive Treatment | Concept: Suspicious |
|---|---|---|
| coll impact was low | Tx appears exaggerated | appears rather suspect |
| low impact collison | TMT is excessive | |
| low speed rear-end accident | | |

## Simple & Practical "Text Mining"

- Program your own text extraction engine
  - SAS, Perl, Java, etc.
- Use simple "phrase matching" to identify related "interesting" concepts
- Create concept flags using textual patterns

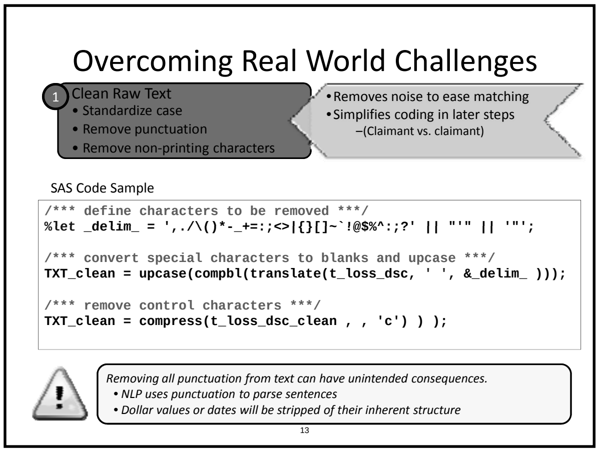

1 Clean Raw Text
- Remove punctuation, etc
- Standardize case

2 Standardize Terms
- Replace "INSD" with "INSURED"
- Labor intensive

3 Concept Generation
- What words / phrases are correlated with target

4 Create Structured Outputs
- Flag record when concept is triggered

5 Evaluate Structured Outputs
- Univariate analysis to determine if process was successful



## Overcoming Real World Challenges

1 Clean Raw Text
- Standardize case
- Remove punctuation
- Remove non-printing characters

- Removes noise to ease matching
- Simplifies coding in later steps
  - (Claimant vs. claimant)

SAS Code Sample

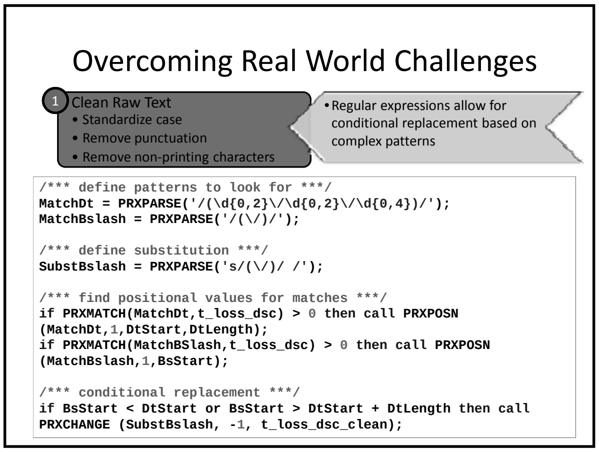
```
/*** define characters to be removed ***/
%let _delim_ = ',./\()*-_+=:;<>|{}[]~`!@$%^:;?' || "'" || '"';

/*** convert special characters to blanks and upcase ***/
TXT_clean = upcase(compbl(translate(t_loss_dsc, ' ', &_delim_ )));

/*** remove control characters ***/
TXT_clean = compress(t_loss_dsc_clean , , 'c') ) );
```

*Removing all punctuation from text can have unintended consequences.*
- *NLP uses punctuation to parse sentences*
- *Dollar values or dates will be stripped of their inherent structure*

13



## Overcoming Real World Challenges

1 Clean Raw Text
- Standardize case
- Remove punctuation
- Remove non-printing characters

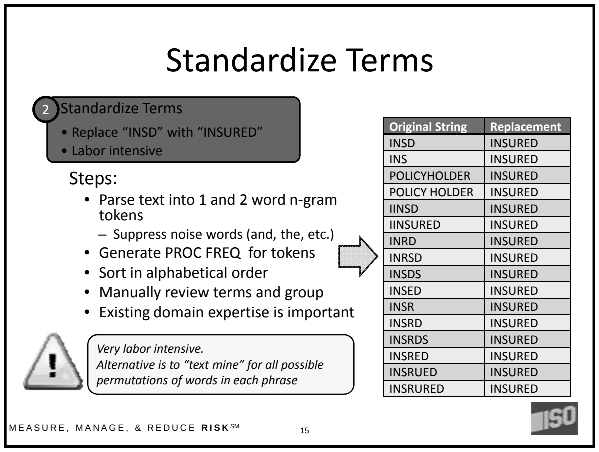
- Regular expressions allow for conditional replacement based on complex patterns

```
/*** define patterns to look for ***/
MatchDt = PRXPARSE('/(\d{0,2}\/\d{0,2}\/\d{0,4})/');
MatchBslash = PRXPARSE('/(\/)/');

/*** define substitution ***/
SubstBslash = PRXPARSE('s/(\/)/ /');

/*** find positional values for matches ***/
if PRXMATCH(MatchDt,t_loss_dsc) > 0 then call PRXPOSN
(MatchDt,1,DtStart,DtLength);
if PRXMATCH(MatchBSlash,t_loss_dsc) > 0 then call PRXPOSN
(MatchBslash,1,BsStart);

/*** conditional replacement ***/
if BsStart < DtStart or BsStart > DtStart + DtLength then call
PRXCHANGE (SubstBslash, -1, t_loss_dsc_clean);
```



## Standardize Terms

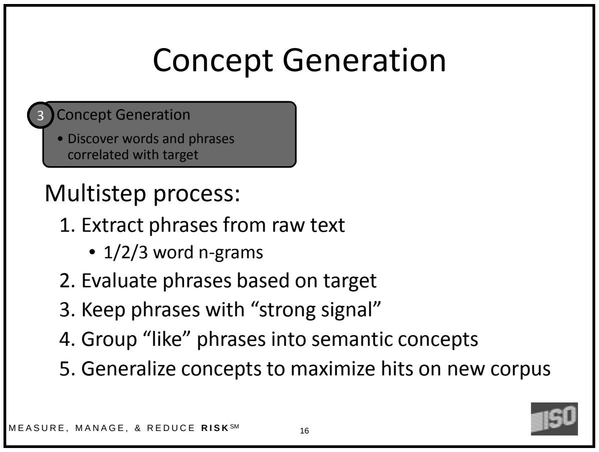
2 Standardize Terms
- Replace “INSD” with “INSURED”
- Labor intensive

Steps:
- Parse text into 1 and 2 word n-gram tokens
  - Suppress noise words (and, the, etc.)
- Generate PROC FREQ for tokens
- Sort in alphabetical order
- Manually review terms and group
- Existing domain expertise is important

| Original String | Replacement |
|---|---|
| INSD | INSURED |
| INS | INSURED |
| POLICYHOLDER | INSURED |
| POLICY HOLDER | INSURED |
| IINSD | INSURED |
| IINSURED | INSURED |
| INRD | INSURED |
| INRSD | INSURED |
| INSDS | INSURED |
| INSED | INSURED |
| INSR | INSURED |
| INSRD | INSURED |
| INSRDS | INSURED |
| INSRED | INSURED |
| INSRUED | INSURED |
| INSRURED | INSURED |

*Very labor intensive.*
*Alternative is to “text mine” for all possible permutations of words in each phrase*

MEASURE, MANAGE, & REDUCE **RISK**℠ 15



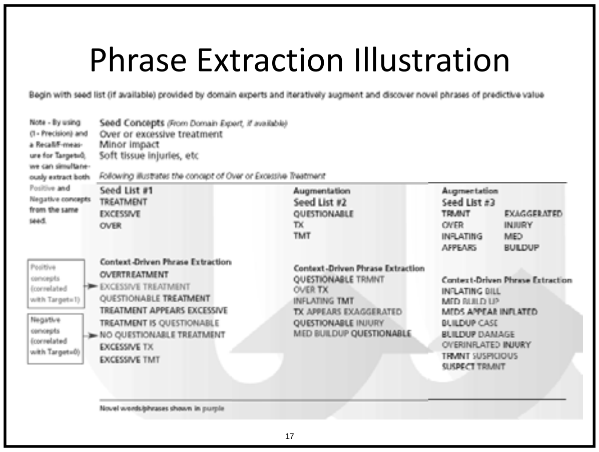
## Concept Generation

3 Concept Generation
- Discover words and phrases correlated with target

Multistep process:

1. Extract phrases from raw text
   - 1/2/3 word n-grams
2. Evaluate phrases based on target
3. Keep phrases with “strong signal”
4. Group “like” phrases into semantic concepts
5. Generalize concepts to maximize hits on new corpus

MEASURE, MANAGE, & REDUCE **RISK**℠ 16



## Phrase Extraction Illustration

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*

**Seed List #1**
TREATMENT
EXCESSIVE
OVER

**Context-Driven Phrase Extraction**
OVERTREATMENT
EXCESSIVE TREATMENT
QUESTIONABLE TREATMENT
TREATMENT APPEARS EXCESSIVE
TREATMENT IS QUESTIONABLE
NO QUESTIONABLE TREATMENT
EXCESSIVE TX
EXCESSIVE TMT

**Augmentation**
**Seed List #2**
QUESTIONABLE
TX
TMT

**Context-Driven Phrase Extraction**
QUESTIONABLE TRMNT
OVER TX
INFLATING TMT
TX APPEARS EXAGGERATED
QUESTIONABLE INJURY
MED BUILDUP QUESTIONABLE

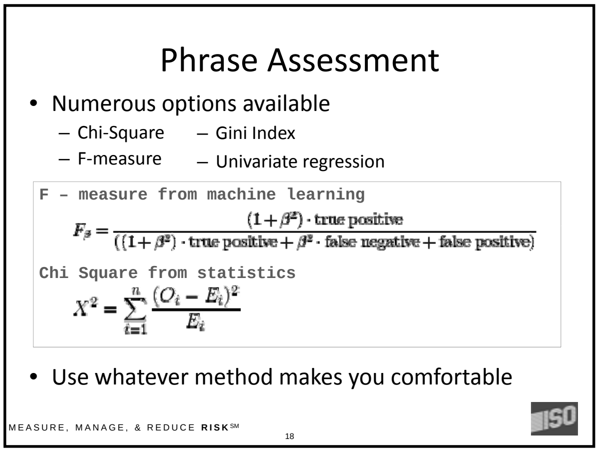
**Augmentation**
**Seed List #3**
TRMNT EXAGGERATED
OVER INJURY
INFLATING MED
APPEARS BUILDUP

**Context-Driven Phrase Extraction**
INFLATING BILL
MED BUILD UP
MEDS APPEAR INFLATED
BUILDUP CASE
BUILDUP DAMAGE
OVERINFLATED INJURY
TRMNT SUSPICIOUS
SUSPECT TRMNT

Positive concepts (correlated with Target=1)
Negative concepts (correlated with Target=0)

Novel words/phrases shown in purple

17



## Phrase Assessment

- Numerous options available
  - Chi-Square
  - F-measure
  - Gini Index
  - Univariate regression

F – measure from machine learning

$$F_\beta = \frac{(1+\beta^2)\cdot \text{true positive}}{((1+\beta^2)\cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive})}$$

Chi Square from statistics

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- Use whatever method makes you comfortable

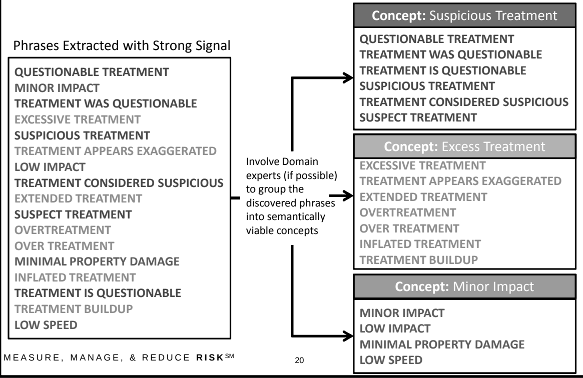
MEASURE, MANAGE, & REDUCE **RISK**℠ 18

## Phrase Assessment

- N-Grams labeled with Precision, Recall and F-measure
  - Higher F-measure is better
  - $\beta$ = .25

| Phrase | Precision | Recall | F- Measure |
|---|---|---|---|
| APPEARS EXAGGERATED | 81.4% | 4.1% | 0.686 |
| EXCESSIVE TREATMENT | 79.6% | 3.9% | 0.668 |
| QUESTIONABLE TREATMENT | 78.5% | 3.8% | 0.657 |
| QUESTIONABLE INJURY | 72.9% | 4.5% | 0.634 |
| TX APPEARS EXAGGERATED | 89.7% | 1.9% | 0.615 |
| INFLATING BILL | 94.2% | 1.7% | 0.612 |
| NO PROP DMG | 58.9% | 13.7% | 0.570 |
| IMAPCT WAS LOW | 54.3% | 12.7% | 0.526 |
| MED BUILDUP | 95.2% | 0.9% | 0.467 |
| TREATED WITH CP | 92.7% | 0.9% | 0.459 |
| EXCESSIVE TMT | 81.7% | 1.0% | 0.452 |
| PT | 51.0% | 2.0% | 0.410 |
| PT EXTENSIVELY | 67.0% | 1.0% | 0.405 |
| CP | 84.0% | 0.6% | 0.354 |
| TX APPEARS | 37.3% | 2.8% | 0.332 |
| BUILD UP CASE | 88.7% | 0.4% | 0.278 |
| CP AND PT | 78.0% | 0.3% | 0.231 |
| LOW | 10.6% | 63.9% | 0.107 |
| IMPACT | 5.9% | 89.3% | 0.060 |

## Concept Generation



Phrases Extracted with Strong Signal

QUESTIONABLE TREATMENT
MINOR IMPACT
TREATMENT WAS QUESTIONABLE
EXCESSIVE TREATMENT
SUSPICIOUS TREATMENT
TREATMENT APPEARS EXAGGERATED
LOW IMPACT
TREATMENT CONSIDERED SUSPICIOUS
EXTENDED TREATMENT
SUSPECT TREATMENT
OVERTREATMENT
OVER TREATMENT
MINIMAL PROPERTY DAMAGE
INFLATED TREATMENT
TREATMENT IS QUESTIONABLE
TREATMENT BUILDUP
LOW SPEED

Involve Domain experts (if possible) to group the discovered phrases into semantically viable concepts

**Concept:** Suspicious Treatment

QUESTIONABLE TREATMENT
TREATMENT WAS QUESTIONABLE
TREATMENT IS QUESTIONABLE
SUSPICIOUS TREATMENT
TREATMENT CONSIDERED SUSPICIOUS
SUSPECT TREATMENT

**Concept:** Excess Treatment

EXCESSIVE TREATMENT
TREATMENT APPEARS EXAGGERATED
EXTENDED TREATMENT
OVERTREATMENT
OVER TREATMENT
INFLATED TREATMENT
TREATMENT BUILDUP

**Concept:** Minor Impact

MINOR IMPACT
LOW IMPACT
MINIMAL PROPERTY DAMAGE
LOW SPEED

MEASURE, MANAGE, & REDUCE **RISK**℠ 20

## Concept Generalization

**Concept:** Suspicious Treatment

QUESTIONABLE TREATMENT
TREATMENT WAS QUESTIONABLE
TREATMENT IS QUESTIONABLE

SUSPICIOUS TREATMENT
TREATMENT CONSIDERED SUSPICIOUS
SUSPECT TREATMENT

Two key patterns among all phrases

- "QUESTION" within a few words of "TREAT"
- Some variation of "SUSPECT" within a few words or "TREAT"

Within 20 characters

```
/** use regular expressions to match generalized patterns **/
Suspect_Tmt1 = PRXPARSE
('/((QUESTION(?:\W+\w+){0,20}?\W+TREAT)|(TREAT(?:\W+\w+){0,20}
?\W+(QUESTION))/');
Suspect_Tmt2 = PRXPARSE ('/((SUSP(?:\W+\w+){0,20}?\W+TREAT)
|(TREAT(?:\W+\w+){0,20}?\W+(SUSP))/');
```

*Keep patterns as simple as possible. If there are five distinct patterns create five regular expressions. This simplifies backend diagnostics.*

21

## Create Structured Outputs

Specify pattern to match

```
/** use regular expressions to match generalized patterns **/
Suspect_Tmt1 = PRXPARSE
('/\b((QUESTION(?:\W+\w+){0,20}?\W+TREAT)|TREAT(?:\W+\w+){0,20}
?\W+(QUESTION))\b/');
Suspect_Tmt2 = PRXPARSE ('/\b((SUSP(?:\W+\w+){0,20}?\W+TREAT)
|TREAT(?:\W+\w+){0,20}?\W+(SUSP))\b/');

/** if pattern matches set flag = 1 **/
F_tmt1 = 0; F_tmt2 = 0;

if PRXMATCH(Suspect_Tmt1, TXT_Clean_subst) > 0 then F_tmt1 = 1;
if PRXMATCH(Suspect_Tmt2, TXT_Clean_subst) > 0 then F_tmt2 = 1;

/** look to see if any "child" concept matched **/
Suspect_Treatment = max (Suspect_Tmt1, Suspect_Tmt2);
```

Matched pattern returns positional value. Set flag.

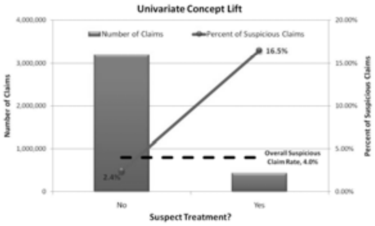
Determine if any pattern matched. Can use **sum** to get total matches.

## Evaluate Univariate Lift

5 Evaluate Structured Outputs

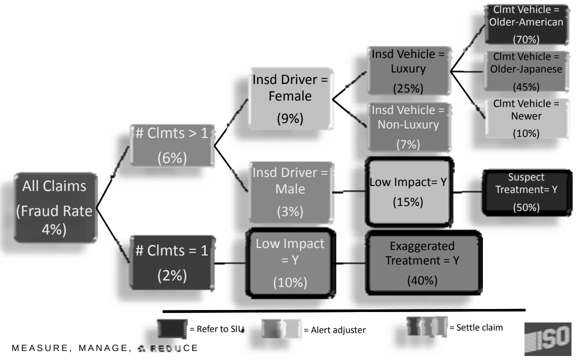
- Univariate analysis to determine if process was successful

- Numerous options available
  - Chi-Square
  - F-measure
  - Gini Index
  - Univariate regression
  - Visual



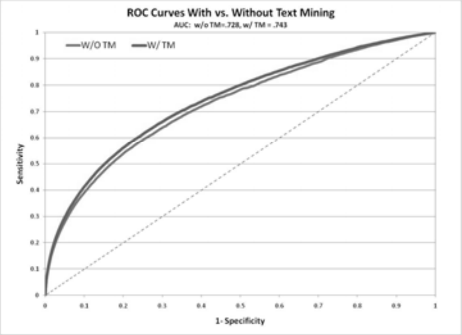
*If "parent" concept is not significant, begin looking at individual "children" concepts to determine which pieces may need refinement*

## Enhanced Detection with Text Mining


All Claims (Fraud Rate 4%)
# Clmts > 1 (6%)
# Clmts = 1 (2%)
Insd Driver = Female (9%)
Insd Driver = Male (3%)
Low Impact = Y (10%)
Insd Vehicle = Luxury (25%)
Insd Vehicle = Non-Luxury (7%)
Low Impact= Y (15%)
Exaggerated Treatment = Y (40%)
Clmt Vehicle = Older-American (70%)
Clmt Vehicle = Older-Japanese (45%)
Clmt Vehicle = Newer (10%)
Suspect Treatment= Y (50%)
= Refer to SIU
= Alert adjuster
= Settle claim

MEASURE, MANAGE, & REDUCE

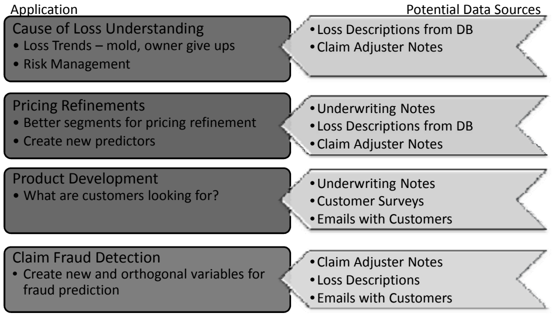
# Model Lift


ROC Curves With vs. Without Text Mining
AUC: w/o TM=.728, w/ TM = .743

What else can I do?

## OTHER P&C APPLICATIONS

# Potential P&C Applications


Application | Potential Data Sources

Cause of Loss Understanding
- Loss Trends – mold, owner give ups
- Risk Management

- Loss Descriptions from DB
- Claim Adjuster Notes

Pricing Refinements
- Better segments for pricing refinement
- Create new predictors

- Underwriting Notes
- Loss Descriptions from DB
- Claim Adjuster Notes

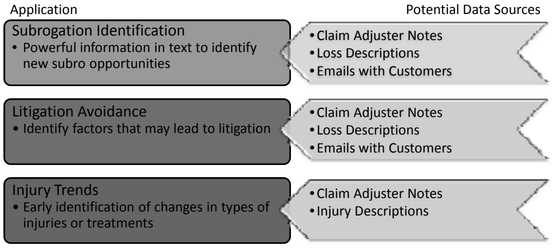
Product Development
- What are customers looking for?

- Underwriting Notes
- Customer Surveys
- Emails with Customers

Claim Fraud Detection
- Create new and orthogonal variables for fraud prediction

- Claim Adjuster Notes
- Loss Descriptions
- Emails with Customers

# Potential P&C Applications


Application | Potential Data Sources

Subrogation Identification
- Powerful information in text to identify new subro opportunities

- Claim Adjuster Notes
- Loss Descriptions
- Emails with Customers

Litigation Avoidance
- Identify factors that may lead to litigation

- Claim Adjuster Notes
- Loss Descriptions
- Emails with Customers

Injury Trends
- Early identification of changes in types of injuries or treatments

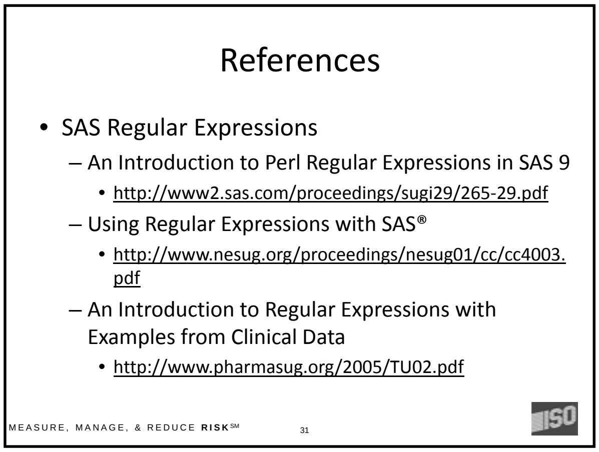
- Claim Adjuster Notes
- Injury Descriptions

# Summary

- Text Mining can release the power buried in unstructured data
  - Many applications to explore in P&C
- Numerous issues exist in real world text that must be addressed to harness this power
- Text mining is an iterative – learn and refinement process
- Programming your own extraction engine is possible with existing tools
  - Good open source tools also exist

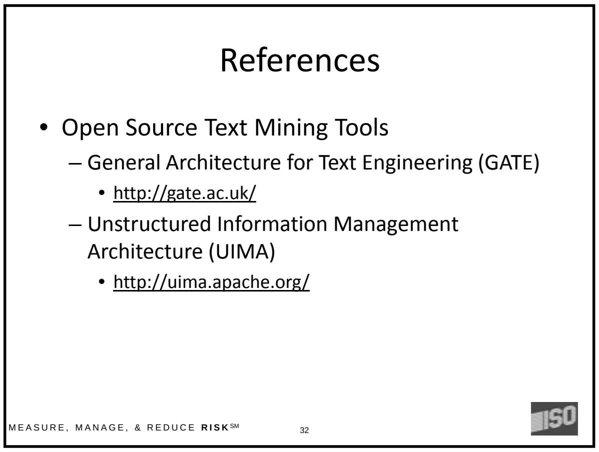
# Feedback and Questions

- Send feedback to:
  - Janine Johnson
  - 415.276.4105
  - e-mail: janine.johnson@iso.com



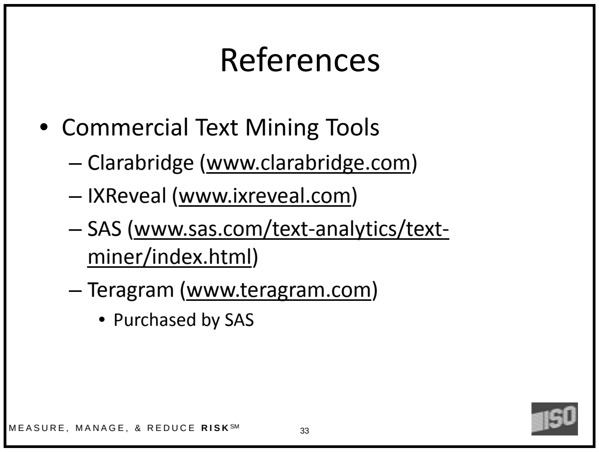
## References

- SAS Regular Expressions
  - An Introduction to Perl Regular Expressions in SAS 9
    - http://www2.sas.com/proceedings/sugi29/265-29.pdf
  - Using Regular Expressions with SAS®
    - http://www.nesug.org/proceedings/nesug01/cc/cc4003.pdf
  - An Introduction to Regular Expressions with Examples from Clinical Data
    - http://www.pharmasug.org/2005/TU02.pdf



## References

- Open Source Text Mining Tools
  - General Architecture for Text Engineering (GATE)
    - http://gate.ac.uk/
  - Unstructured Information Management Architecture (UIMA)
    - http://uima.apache.org/



## References

- Commercial Text Mining Tools
  - Clarabridge (www.clarabridge.com)
  - IXReveal (www.ixreveal.com)
  - SAS (www.sas.com/text-analytics/text-miner/index.html)
  - Teragram (www.teragram.com)
    - Purchased by SAS