Network Analytics in Claims Level Predictive Modelling

Marcela Granados, Satraajeet Mukherjee

"R in Insurance" conference

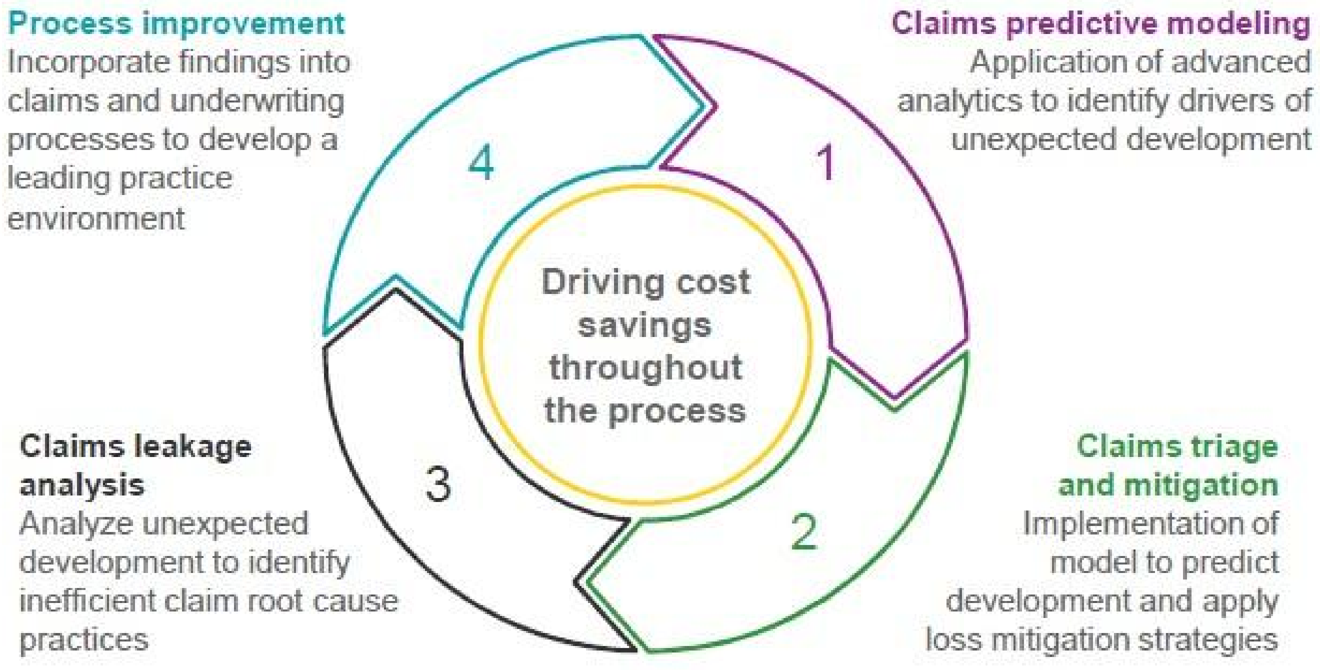
Paris, France

June 8th, 2017

Building a better
working world

Interaction Between Claims and Analytics

- What drives adverse development?
  - *Adverse development is disproportionately driven by specific types of claims*
  - *It can be difficult to quantify the preponderance of factors that drive claims development*
  - *Early identification of these claims allows for proactive claims handling and real cost savings*



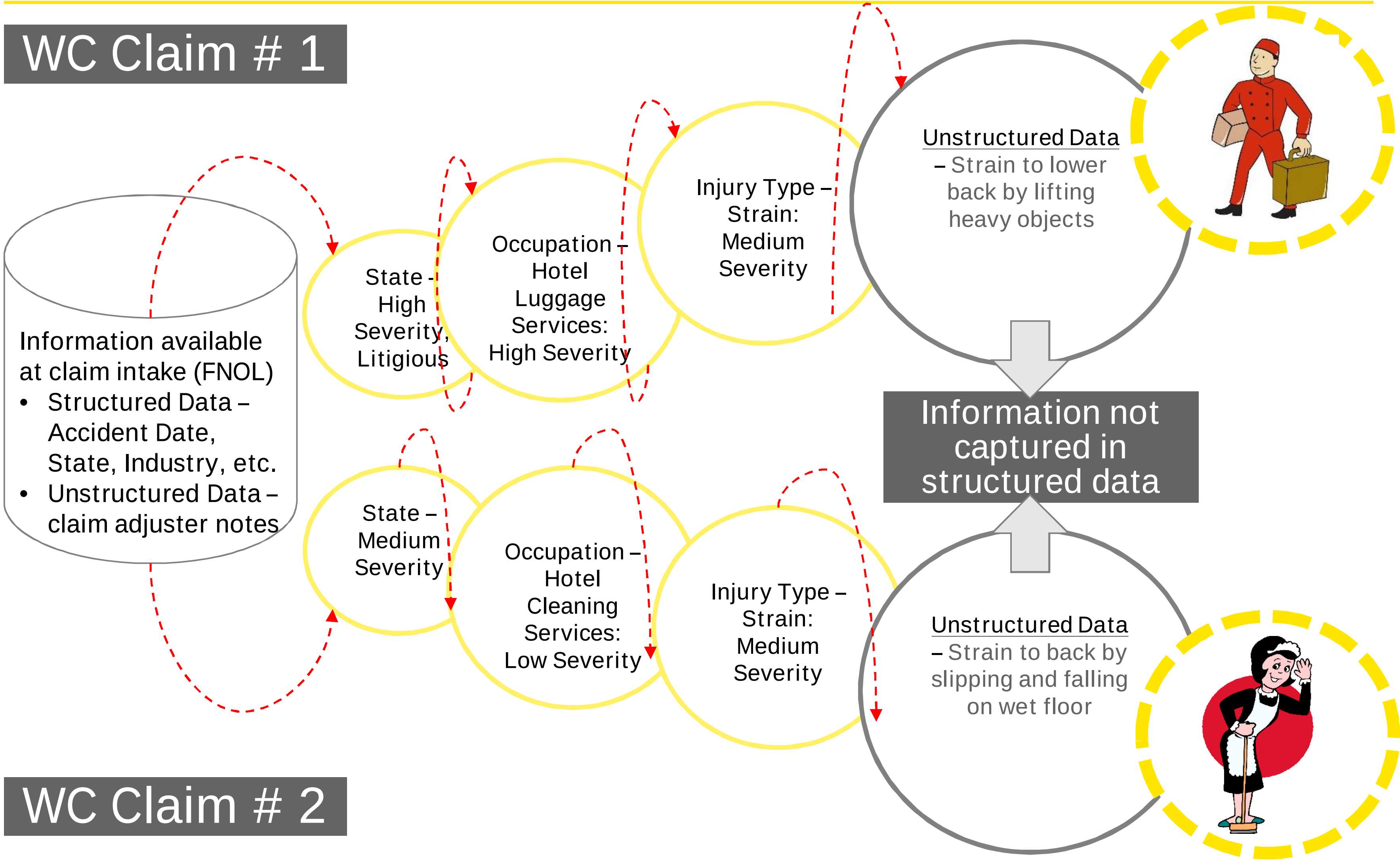
Process improvement
Incorporate findings into claims and underwriting processes to develop a leading practice environment

Claims predictive modeling
Application of advanced analytics to identify drivers of unexpected development

Claims triage and mitigation
Implementation of model to predict development and apply loss mitigation strategies

Claims leakage analysis
Analyze unexpected development to identify inefficient claim root cause practices

Claims Life Cycle



WC Claim # 1

Information available at claim intake (FNOL)

- Structured Data – Accident Date, State, Industry, etc.
- Unstructured Data – claim adjuster notes

State - High Severity, Litigious

Occupation – Hotel Luggage Services: High Severity

Injury Type – Strain: Medium Severity

Unstructured Data – Strain to lower back by lifting heavy objects

Information not captured in structured data

State – Medium Severity

Occupation – Hotel Cleaning Services: Low Severity

Injury Type – Strain: Medium Severity

Unstructured Data – Strain to back by slipping and falling on wet floor

WC Claim # 2

NLP is used to clean unstructured data, then network analytics is used to identify predictors

Tm – for cleaning
SnowballC - stemming

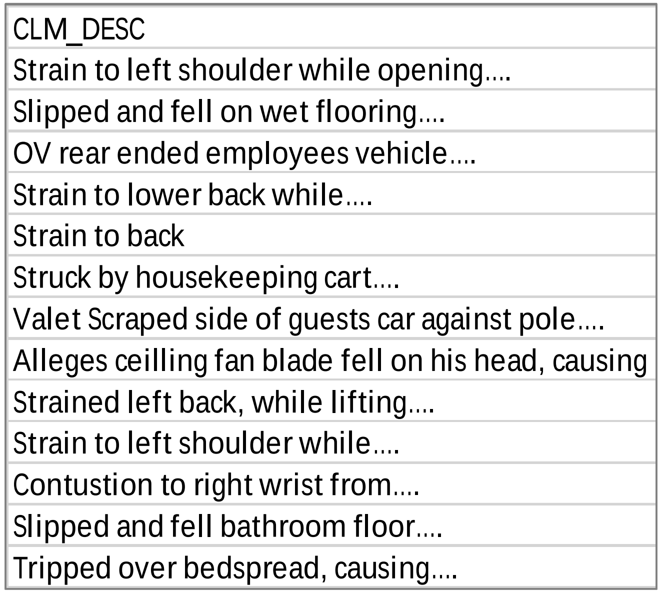
Data Cleaning steps:

1. Remove stop words (e.g. ‘to’, ‘and’, etc.)
2. Stemming (identify words that have the same meaning/root)

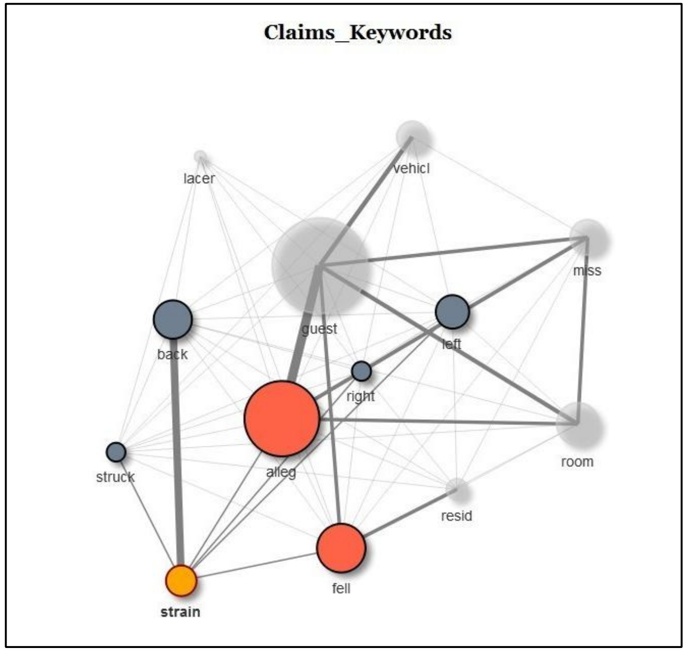
Create network graph:

1. Node size (circle) shows the frequency of words
2. Width of lines between nodes shows frequency of words occurring together

Visnetwork
Rgraphviz



CLM_DESC
Strain to left shoulder while opening....
Slipped and fell on wet flooring....
OV rear ended employees vehicle....
Strain to lower back while....
Strain to back
Struck by housekeeping cart....
Valet Scraped side of guests car against pole....
Alleges ceilling fan blade fell on his head, causing
Strained left back, while lifting....
Strain to left shoulder while....
Contustion to right wrist from....
Slipped and fell bathroom floor....
Tripped over bedspread, causing....



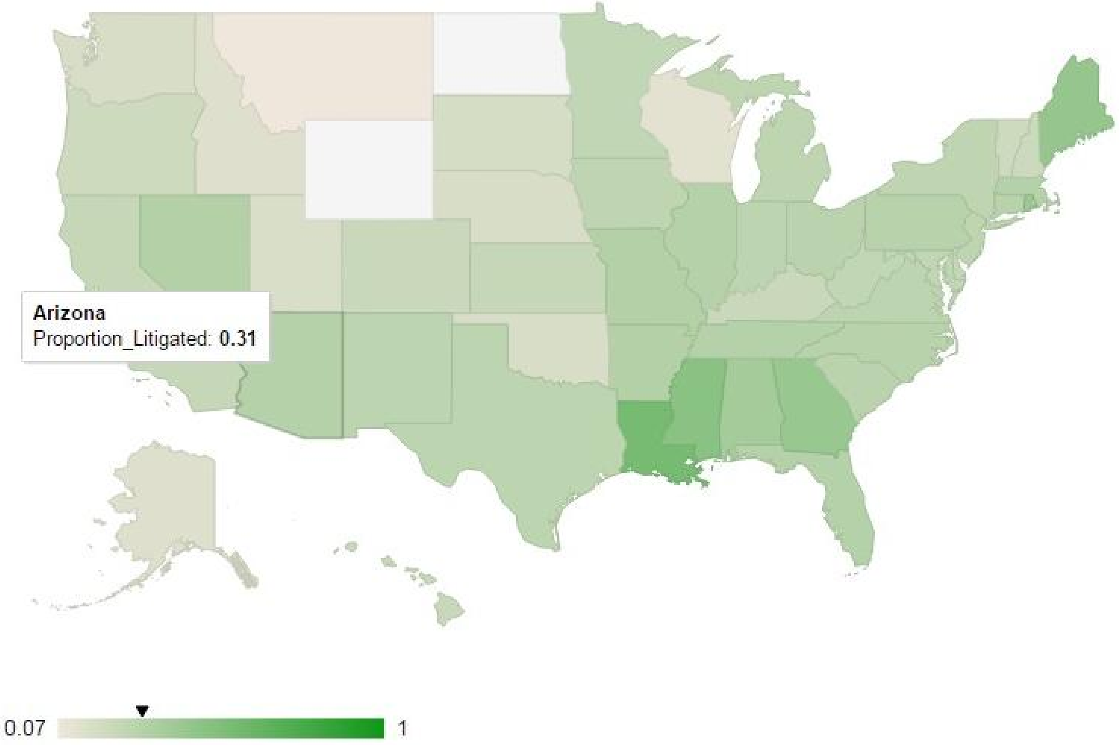
Raw Unstructured Claims Data

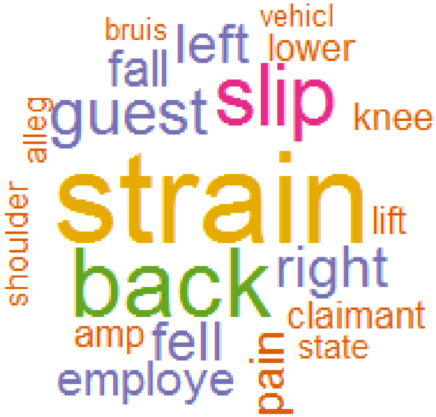
R Packages for Machine Learning – randomforest, caret

Inputs for Predictive Models

Use of GeoChart – preliminary visual analytics

- Powerful visual analytics tools such as Geochart and wordcloud can be used to analyze structured and unstructured data to identify the most predictive variables

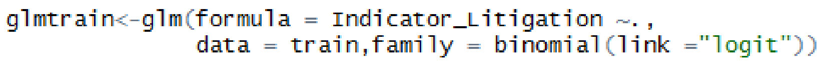




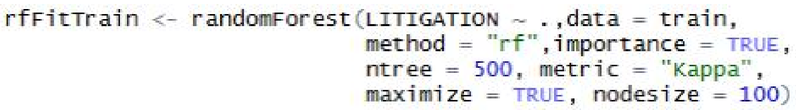
Darker Green indicates higher probability of litigation

Larger words indicate higher frequency of words

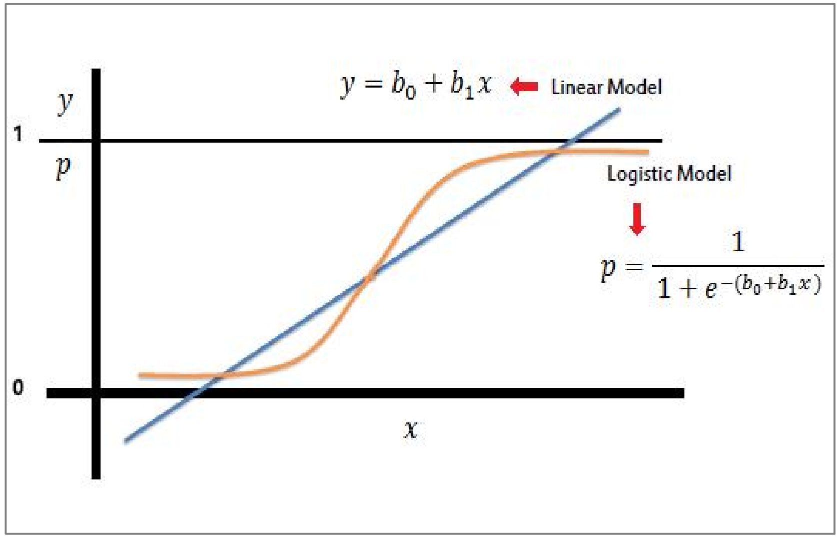
Machine Learning – Model Evolution

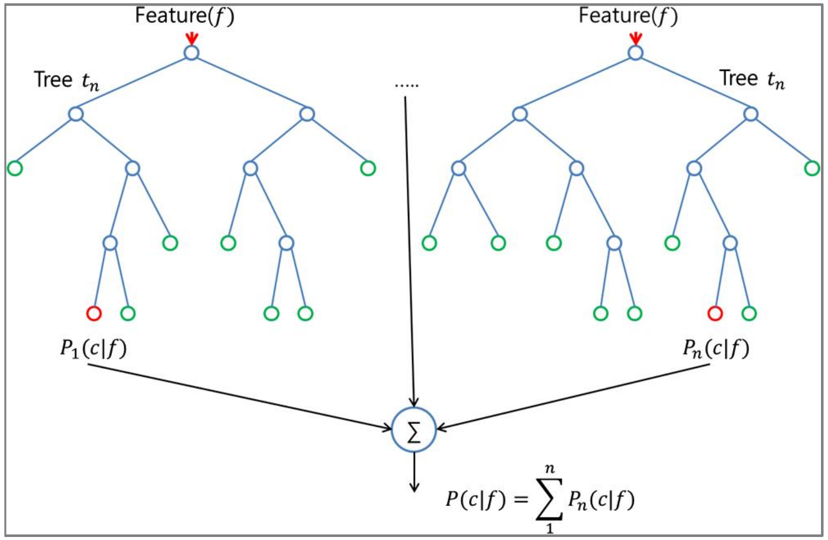


```
glmtrain<-glm(formula = Indicator_Litigation ~.,
              data = train,family = binomial(link ="logit"))
```



```
rfFitTrain <- randomForest(LITIGATION ~ .,data = train,
                           method = "rf",importance = TRUE,
                           ntree = 500, metric = "Kappa",
                           maximize = TRUE, nodesize = 100)
```





Parameter Estimation in Logistic Regression

Criteria for Random Forest Modelling Splits

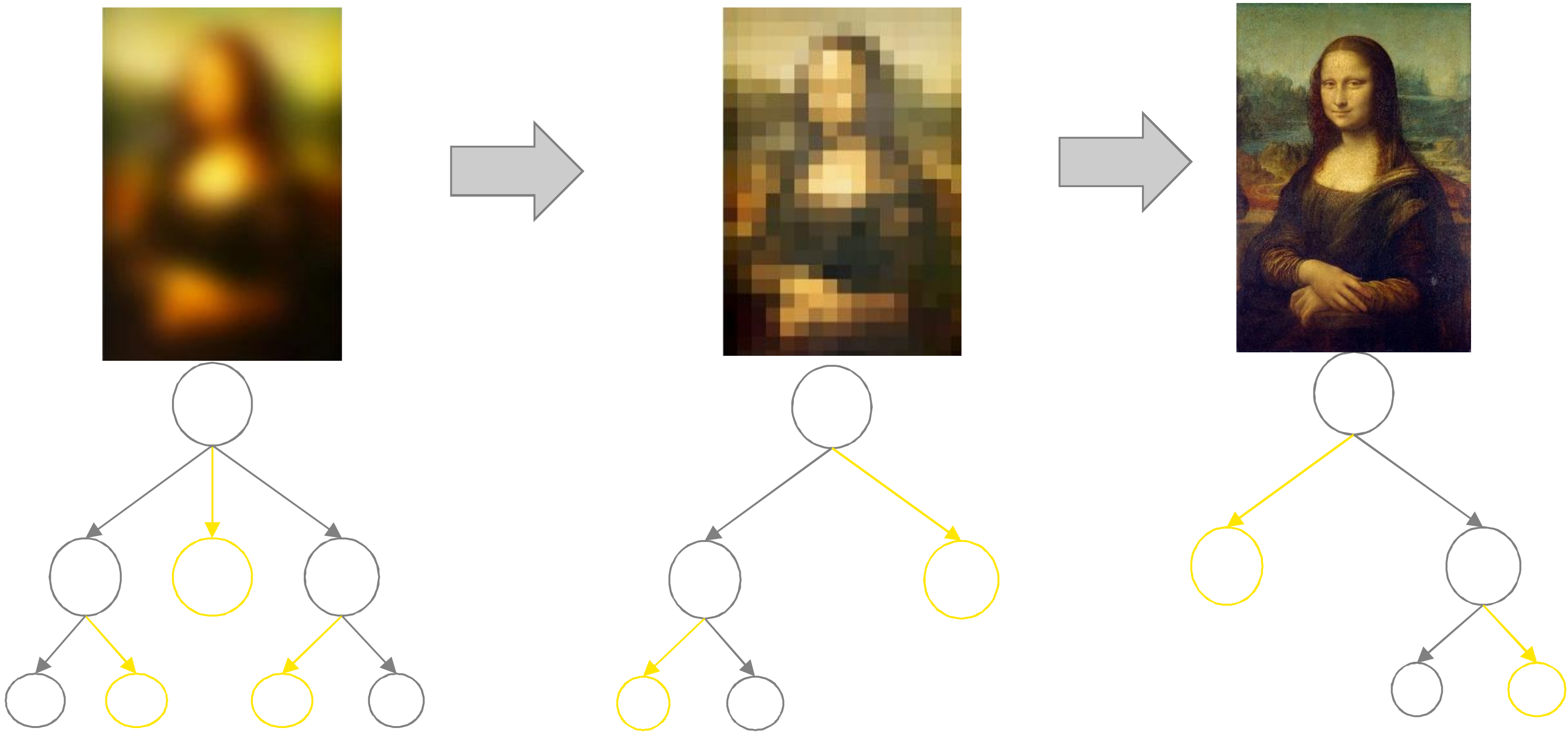
$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$$

$$Weighted\ Gini\ = \sum_i \frac{n_i}{N}\ ({p_i}^2 + {q_i}^2)$$

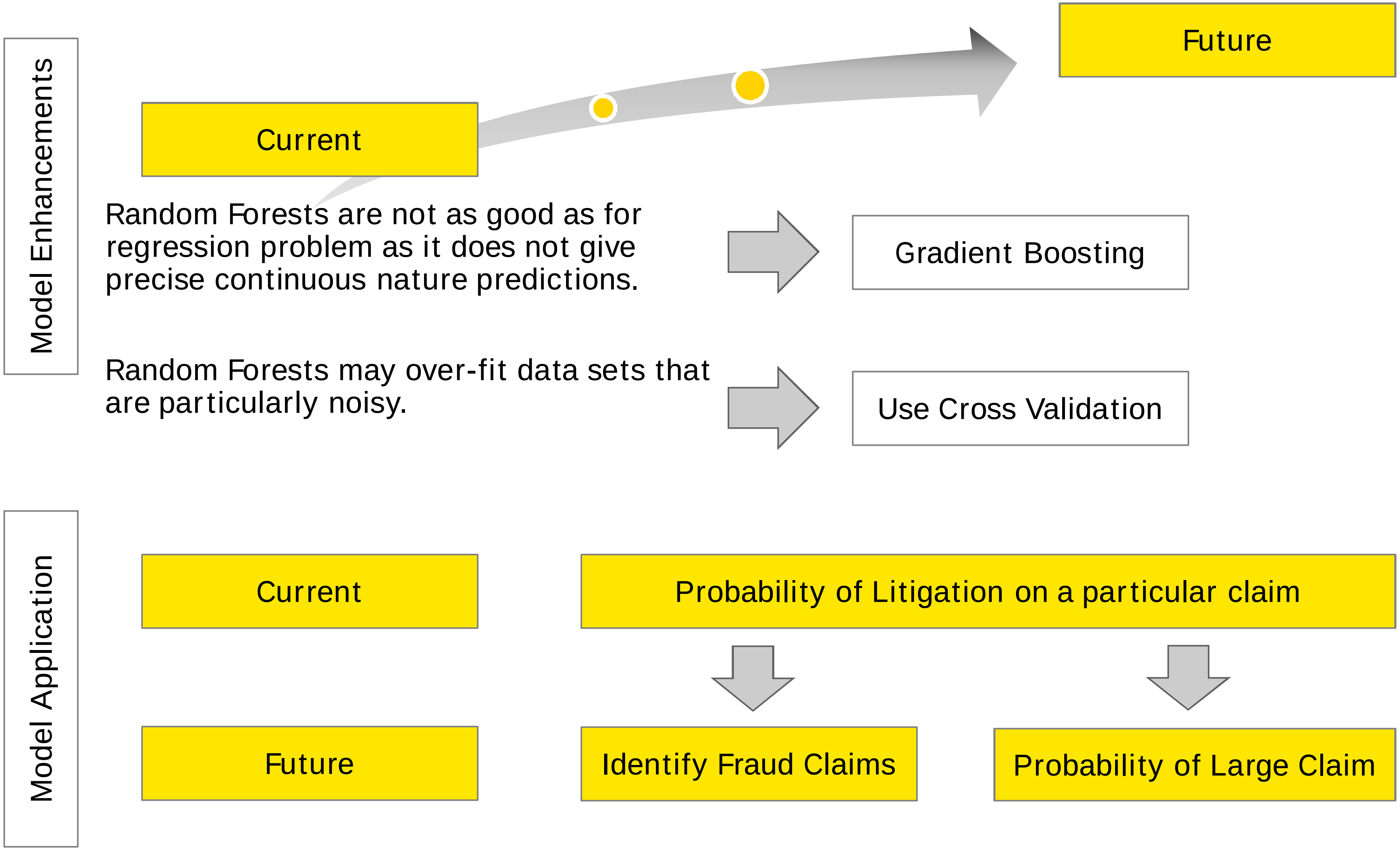
$$Entropy = -plog_2p - qlog_2q$$

Use of Machine Learning – Random Forests

- Random forests build upon the concept of asking the classification question to multiple people who think differently, such that the end answer is truly unbiased
- So instead of relying upon a single decision tree and dataset, the algorithm builds an ensemble of decision trees using bootstrapped versions of the original dataset



Next Steps



Model Enhancements

Current

Future

Random Forests are not as good as for regression problem as it does not give precise continuous nature predictions.

Gradient Boosting

Random Forests may over-fit data sets that are particularly noisy.

Use Cross Validation

Model Application

Current

Probability of Litigation on a particular claim

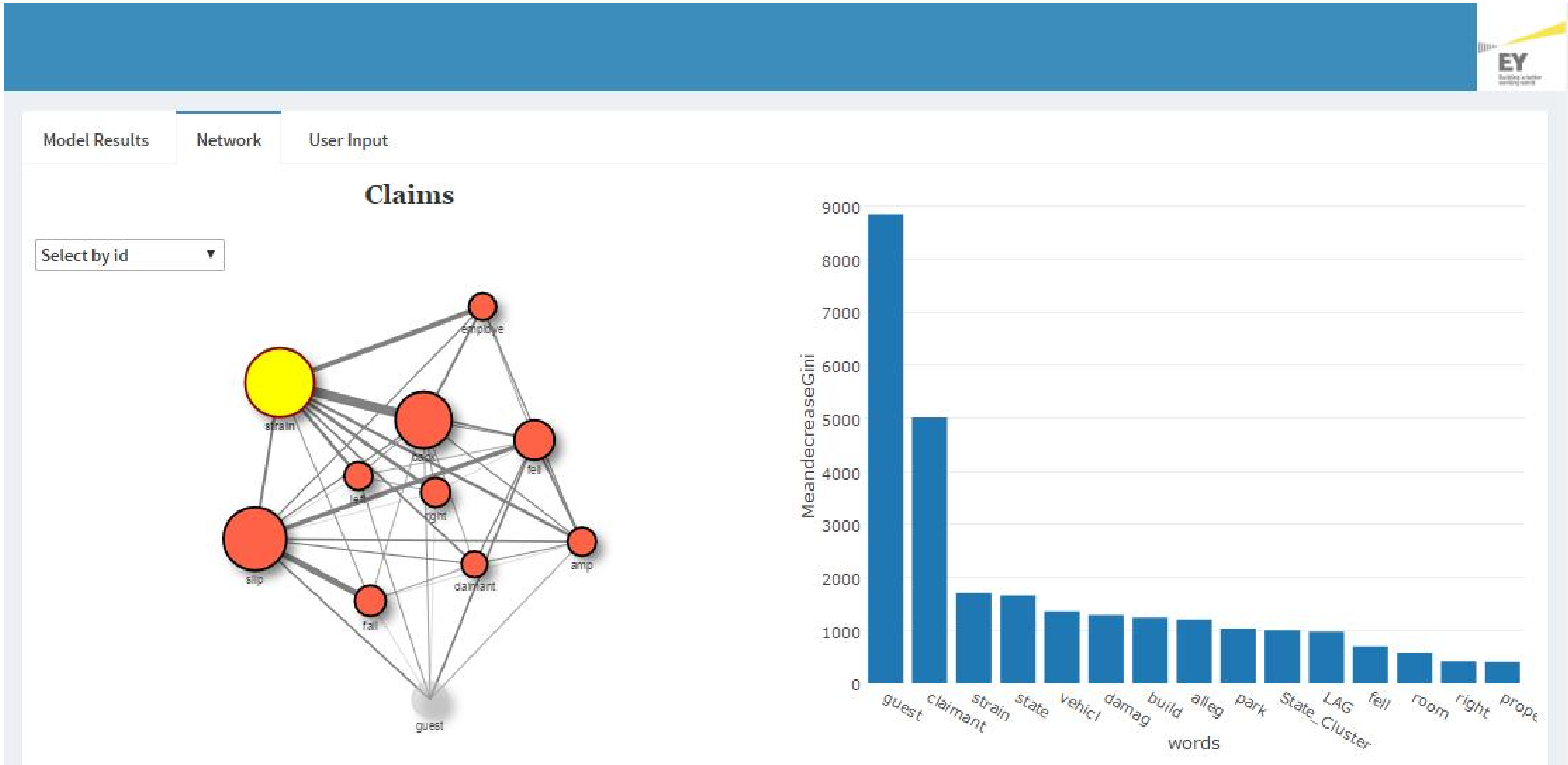
Future

Identify Fraud Claims

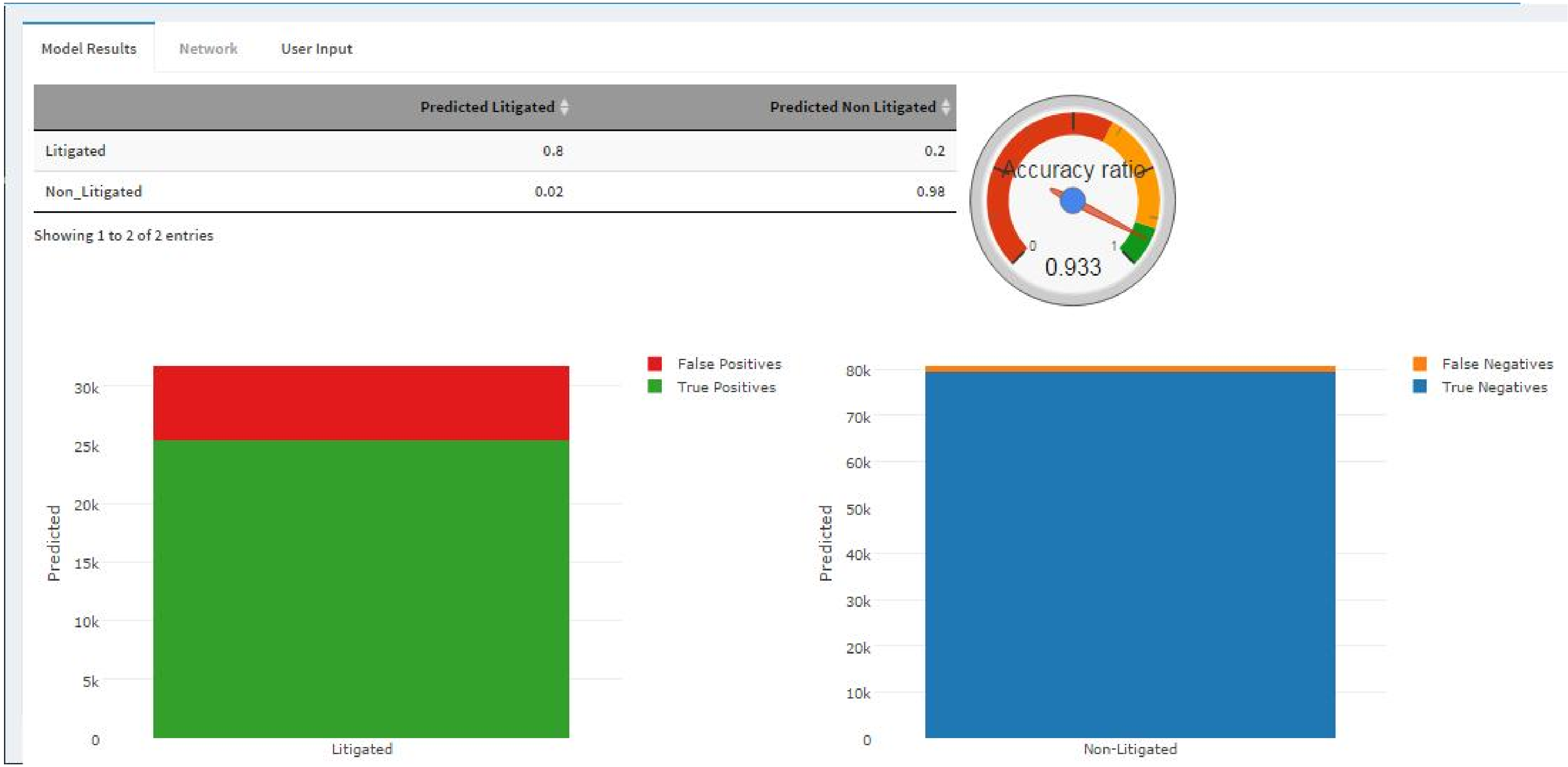
Probability of Large Claim

Appendix

Shiny Interface – Network Graph



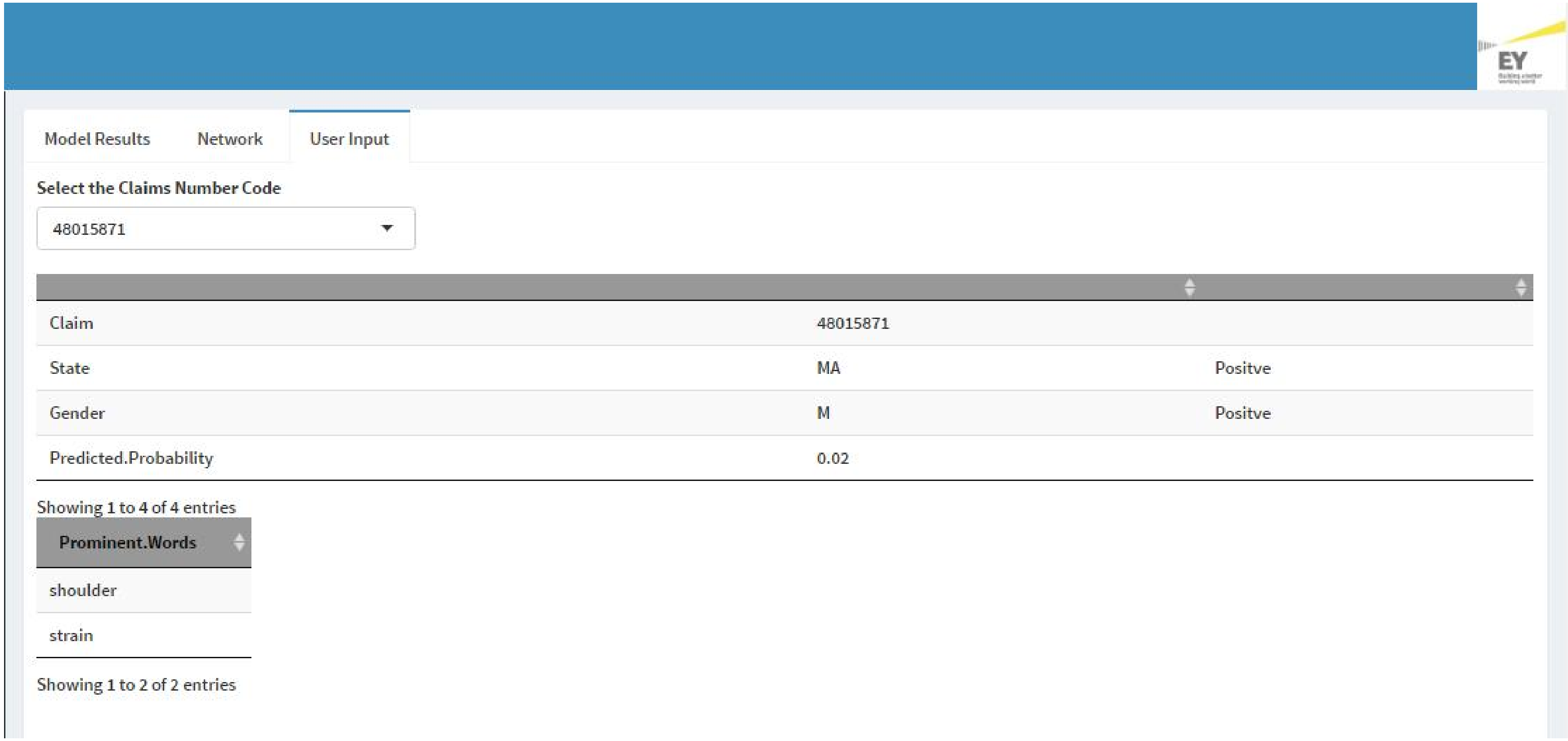
Shiny Interface – Model Results



	Predicted Litigated	Predicted Non Litigated
Litigated	0.8	0.2
Non_Litigated	0.02	0.98

Showing 1 to 2 of 2 entries

Shiny Interface – Claims Adjuster’s tool



Select the Claims Number Code

48015871

Claim	48015871	
State	MA	Positve
Gender	M	Positve
Predicted.Probability	0.02	

Showing 1 to 4 of 4 entries

Prominent.Words
shoulder
strain

Showing 1 to 2 of 2 entries