Insurance Claims Frauds & Text Analytics

Adam Green
Enzo Martoglio

Presentation Overview

- ➔ A short introduction to fraud in insurance
- ➔ The development of new models of fraud analysis
- ➔ Some challenges in modelling fraud
- ➔ Examples of advanced analysis approaches
- ➔ Text analytics with R
- ➔ An example: Using the LIWC dictionary to detect fraud
- ➔ The business case for R and enhanced analysis

A short introduction to fraud in insurance (1)

The scale of the problem is large, and potentially growing

- ➔ Every week a staggering 2,670 fraudulent insurance claims worth £19 million are being uncovered as UK insurers intensify their crackdown on insurance cheats (ABI 9/2012)

UK Insurers detected 139,000 bogus or exaggerated insurance claims, up 5% on 2010. The value of savings on these frauds was nearly £1 billion - £983million - a rise of 7% on 2010.

Home insurance frauds were the most common with 71,000 dishonest claims, valued at £106 million, detected.

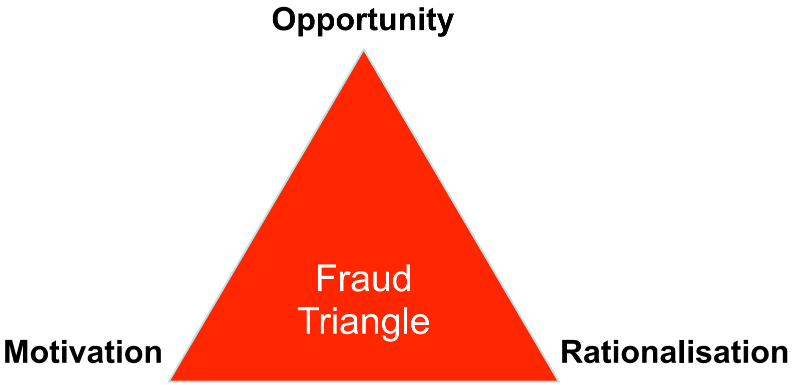
Dishonest motor insurance claims were the most expensive with savings of £541 million made from the 45,000 bogus claims uncovered. Fraudulent whiplash claims was the main factor for the rise.

- ➔ Undetected US general insurance claims fraud total £2.1billion a year adding on average £50 to the annual costs individual policyholders face, on average, each year (www.insurancefraudbureau.org)
- ➔ Fraud can impact many different areas of the value chain, from dishonest policy inception to staged events, to exaggerated claims, supplier claims padding etc.

A short introduction to fraud in insurance (2)

The nature of fraud covers a range of events and issues

- The term fraud here refers to dishonest activities which could reduce an organisation's profits (without necessarily leading to formal responses, e.g. legal)
  - Hard frauds
    - Clear and wilful act / planned / coordinated
    - Proscribed by law
    - Obtaining money or value
    - Under false pretences
    - Penal / police matter
  - Softer frauds / 'abuse'
    - Unwanted, unintended, unnecessary claims
    - Disputable damages
    - Civil matters

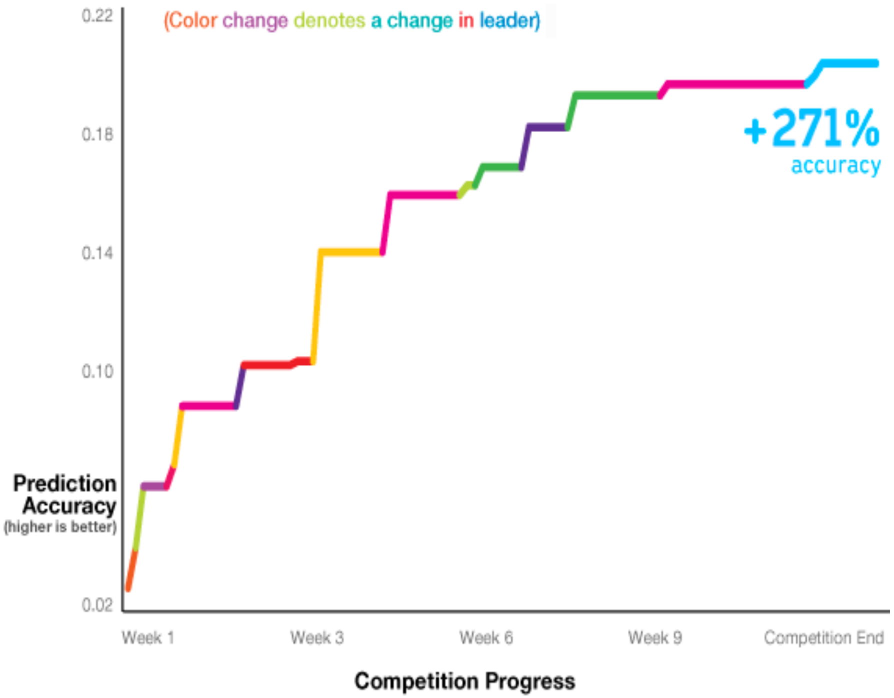


- The incidence of 'hard' frauds is relatively small in many sectors
- However, softer fraud and 'abuse' may be quite prevalent (between ¼ to 1/3 of all claims in motor and household?)
- Important issues for insurances, honest policy holders and, increasingly, for regulators

Development of New Models of Fraud Analysis

Rapidly growing sophistication and maturity

- ➔ AllState launched a competition on the analysis site Kaggle in 2011
- ➔ Competition: Predicting claims liability for injury from car accidents
- ➔ Allstate provided anonymised training data and analysts submited multiple predictive models
- ➔ The winning entry was **271%** more accurate than AllState existing method for predicting claims based on vehicle characteristics
- ➔ AllState is ranked as #4 in the Insurance (Property & Casualty) sector in US



Fraud Modelling Challenges

Data often not shaped for traditional approaches

- Traditional regression analysis requires the availability of a dependent variable (e.g. Fraud Yes / No)
  - y = a + b1x1 + b2x2 + …+ bnxn
  - Fraud (Yes / No – or Fraud "Propensity Score") → f(predictor variables)
  - Requires datasets where where claims have been properly determined to be fraudulent or legitimate (avoiding risks of under-reporting; analysis bias; censored data etc.)
- Collecting and managing these data sets can prove complex and expensive
- 3rd Party or public data is rare, often not shared, not in the required form or in general not a good match for the type of analysis planned
- Good data sets often take time to build and clean, and may be costly to keep up-to-date possibly missing newer fraud types

Examples of advanced analysis approaches (1)

Moving from traditional regression analysis

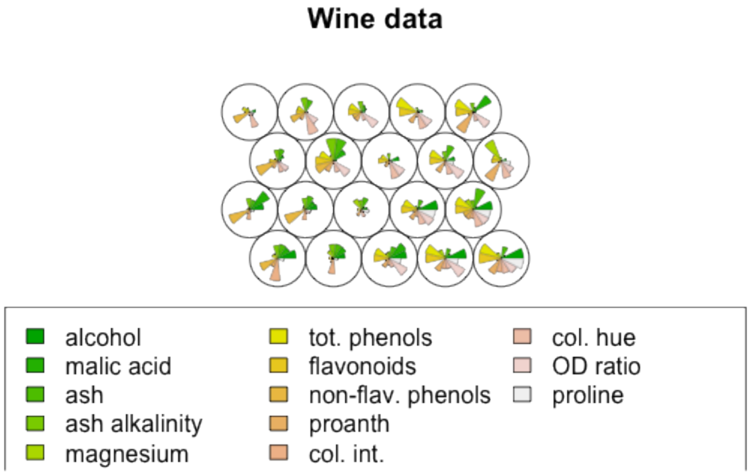
- ➔ “Unsupervised learning” (machine learning)
  - No dependent variable required
  - Approaches available to work with category or continuous data
  - Use Information Augmentation to identify similar characteristics through different groups
    - Geocoding
    - Entity Extraction (text analytics)

- ➔ Some useful methods:
  - Cluster Analysis
    - Records are grouped into categories with similar values (and differentiation)
    - An example – Kohonen’s Self Organising Maps (“SOMs”)
  - Unsupervised, nonparametric aggregation techniques
    - An example – PRIDIT (Principal Component Analysis of RIDITs)

Examples of advanced analysis approaches (2)

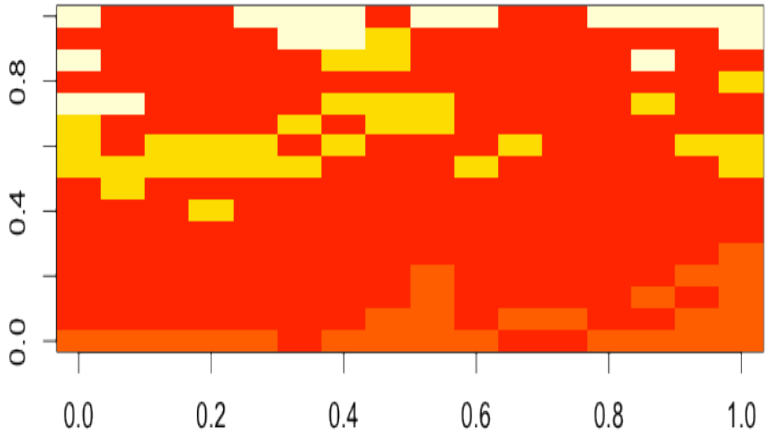
Kohonen’s SOMs – a useful ML tool

Package: Kohonen



Package: RSNNS

(The R Stuttgart Neural Network Simulator)



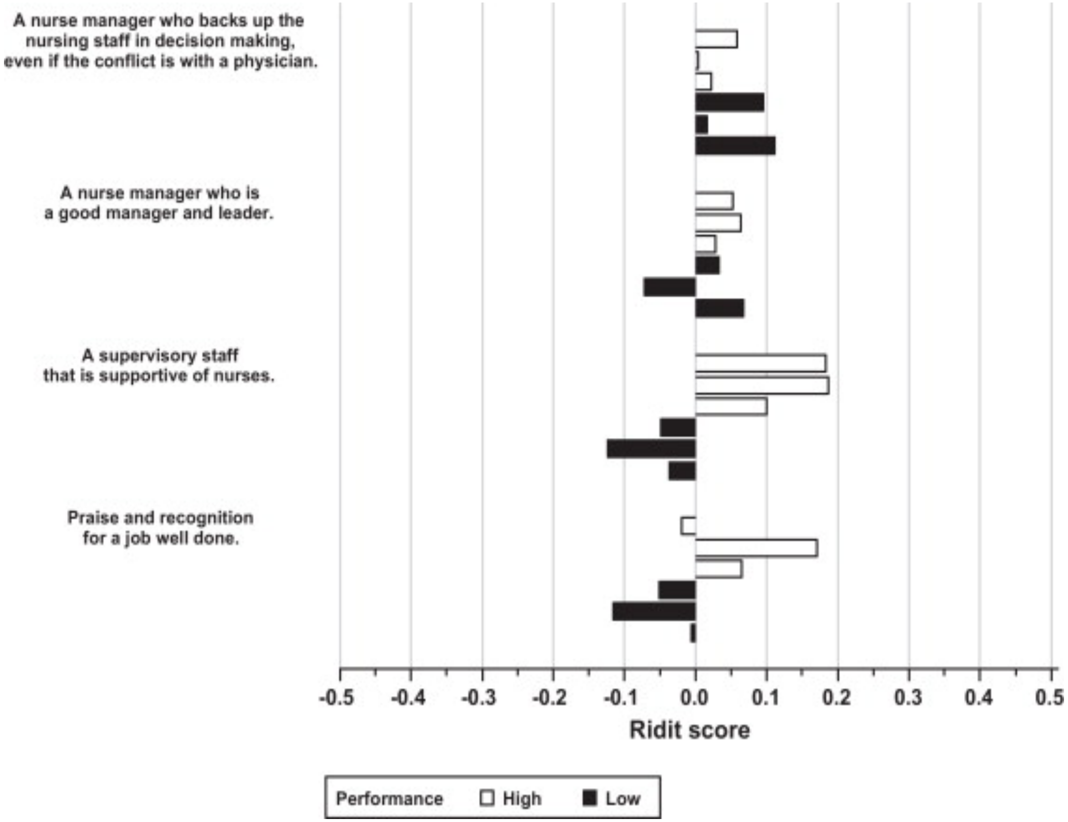
Kohonen’s SOMs also available in Packages: SOM and WCCSOM

Examples of advanced analysis approaches (3)

RIDIT Scoring – A method for ordered qualitative measurements

Package: RIDIT

- ➔ Analysis output is a score (not a cluster mapping)
- ➔ the qualitative responses of each claim on every variable are transformed into numerical variable scores, without altering the ranked nature of the data
- ➔ RIDIT analysis is used for data which are ordered (e.g. injury categories), but are not on an interval scale



PRIDIT as a fraud detection method

- Brockett and colleagues (Journal of Risk and Insurance, 2002)
- Use RIDIT scores – best for categorical data

- PRIDIT—Principal Component Analysis (PCA) on RIDIT scores
  - Take binary, categorical, and continuous data (a unified PRIDIT method has been proposed to incorporate both categorical and continuous predictors)
  - Empirical cumulative distribution function on variables
  - Transform and normalize using RIDIT scoring

- These variables proxy for an unobserved latent characteristic (i.e. fraud)
  - Use weightings and scores to determine likelihood of latent characteristic
  - Use PCA to assess variance and covariance of variables
  - Those that account for the most of the variation get the highest weighting

Examples of advanced analysis approaches (4)

Different approaches for different jobs

Kohonen SOM

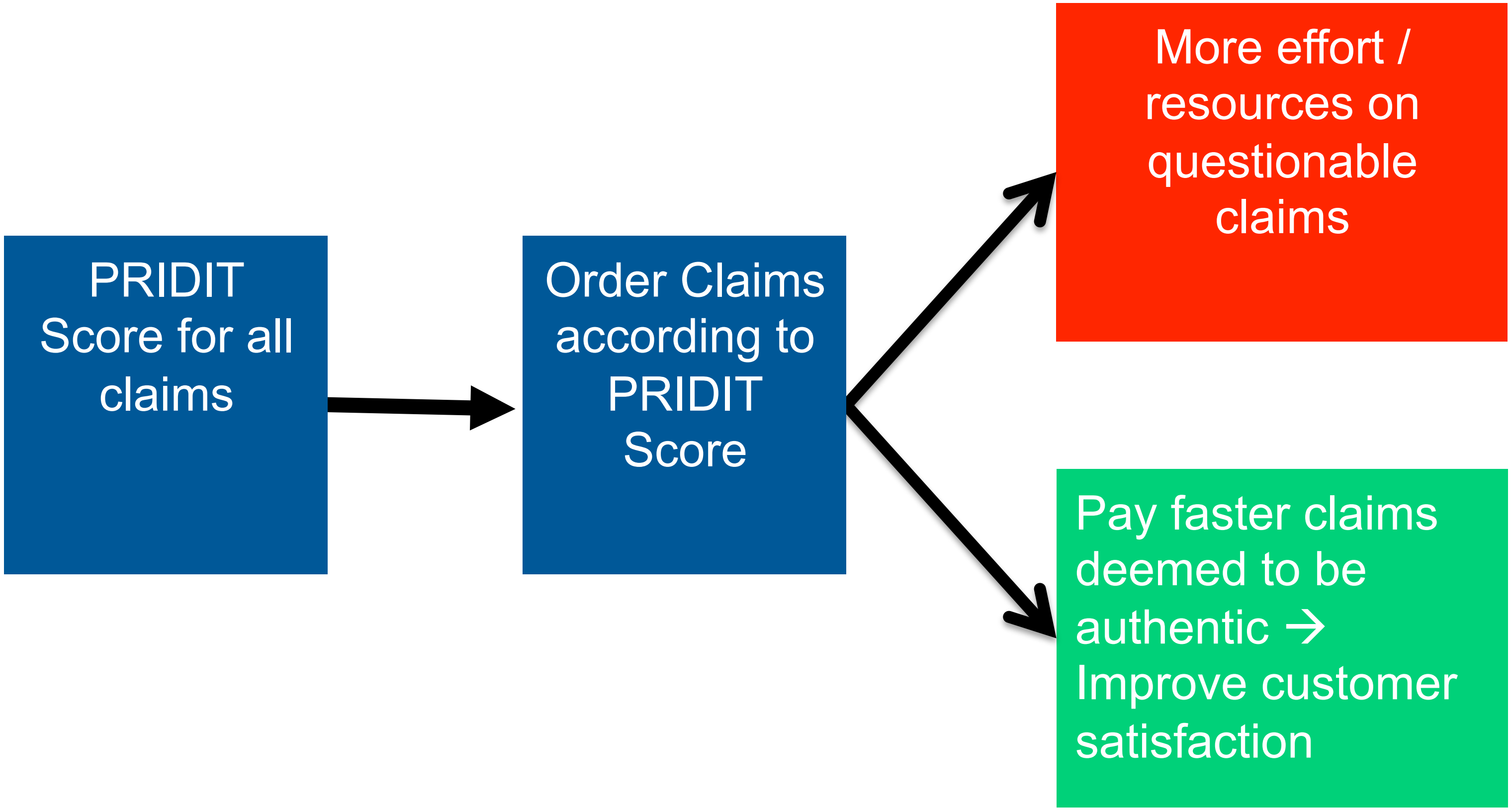
- Clustering has been shown to be effective in separating legitimate from fraudulent or abusive claims
- Kohonen SOM helps to measure similarity / dissimilarity for unordered categorical variables with many values (i.e., injury type) variables
- SOM is visually attractive but sometimes hard to interpret
- It tend to create border problems

PRIDIT / (PCA + RIDIT)

- Studies have indicated a strong relationship between PRIDIT scores and expert suspicion that claim is fraudulent or abusive
- Score offer a clear to understand way to analyse a progression of claims
- The relative importance of variables may be very different from supervised regressions
- No out-of-the-box R Package for PRIDIT

Fraudulent Claim Scoring Workflow

Assigning a “**suspicious level**” score to each claim



PRIDIT Score for all claims

Order Claims according to PRIDIT Score

More effort / resources on questionable claims

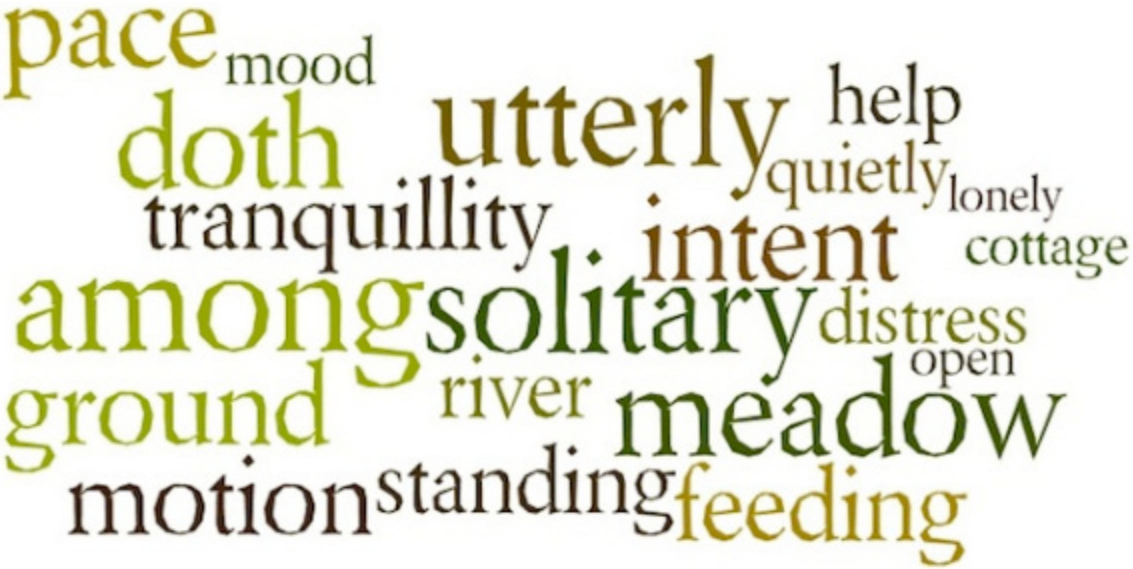
Pay faster claims deemed to be authentic → Improve customer satisfaction

Using Text Analytics & R
to
Detect Fraudulent Claims

Text analysis using R (1)

Everyone is mining text... and discovering new knowledge

Words that are consistently more common in works by William Wordsworth than in other poets from 1780 to 1850.

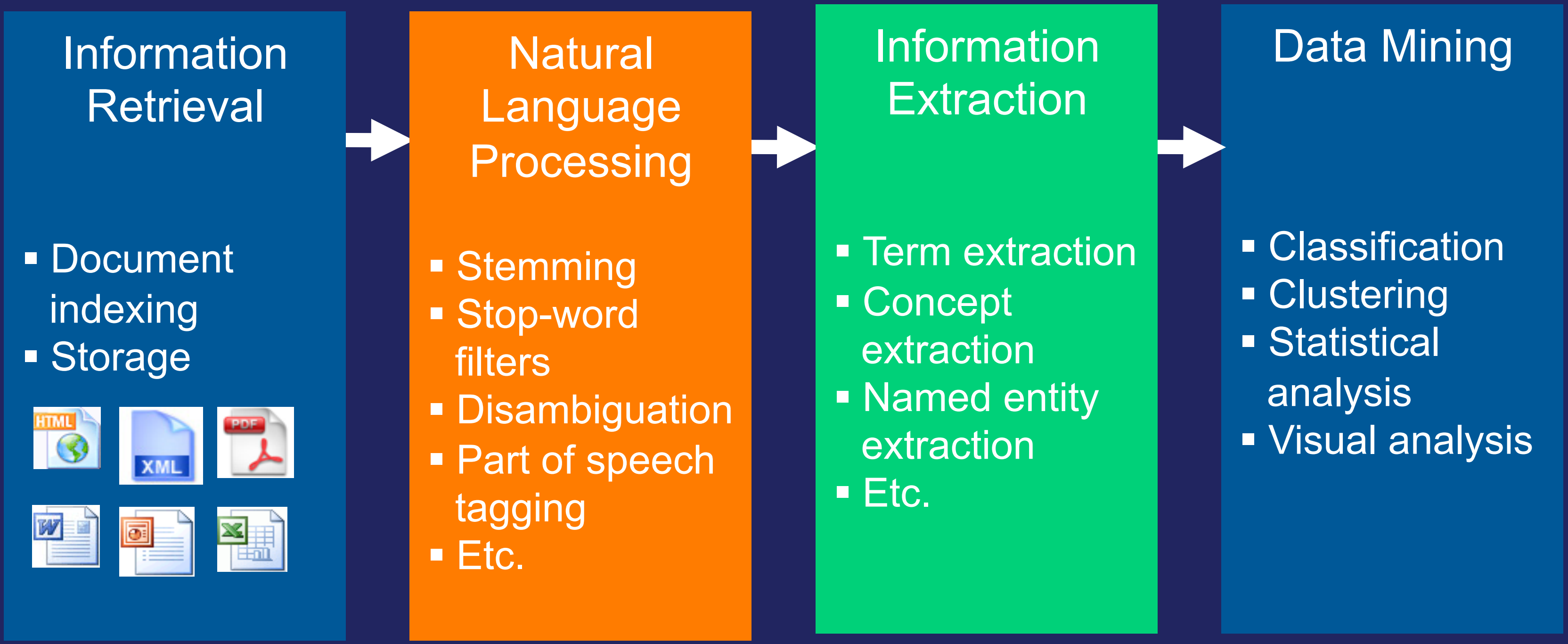


Analysis by Ted Underwood, Associate Professor of English, University of Illinois

Graphic using Wordle's display tool - R used to run a Mann-Whitney test (measures over-representation relative to a context)

Text analysis using R (2)

A basic text analysis workflow



Information Retrieval

- Document indexing
- Storage

Natural Language Processing

- Stemming
- Stop-word filters
- Disambiguation
- Part of speech tagging
- Etc.

Information Extraction

- Term extraction
- Concept extraction
- Named entity extraction
- Etc.

Data Mining

- Classification
- Clustering
- Statistical analysis
- Visual analysis

Retrieve & organise documents

Process & tag documents to ease information extraction

Create Structured data from Unstructured text

Outcome is gather new knowledge

Claim Narrative Entity Extraction

1. Use a tokenizer for entity extraction
   - Wide choice in R: tm or Apache OpenNLP (preferred)
2. Compare with specific “claims’ family” bag-of-words (dictionaries)
   - Auto claims dictionary
   - Health claims dictionary
   - Property claims dictionary
3. Enrich the claims analysis / scoring

➔ Examples:

- Study on auto pre-accident conditions e.g. road works etc.
- Extraction of any relevant medical terms used in the narrative
- Identify potential excessive treatments in health claims
- Potential subrogation opportunities
- More fine-grained peril causes
- Cause-of-loss: mine claimant / adjustor narratives

Deception – A Definition

- Deception
  - Deliberate choice to mislead a target without notification
  - Often to gain some advantage
  - Excludes:
    - Self-deception
    - Theatre, etc.
    - Falsehoods due to ignorance/error
    - Pathological behaviours

Using the LIWC Dictionary to detect fraud (1)

Deceptive Narratives in Claims Linguistic Inquiry and Word Count (LIWC)

- ➔ Professor James Pennebaker is psychology professor & psychology department chair at University of Texas - Austin
- ➔ Developed a dictionary / SW tool known as **Linguistic Inquiry and Word Count (LIWC).**
- ➔ LIWC works with natural speech or writing to predict whether someone is lying
- ➔ Studies confirm that people trying to deceive others often give out valuable clues

•**Fewer I-words:** Liars avoid statements of ownership, distance themselves from their stories and avoid taking responsibility for their behaviour

•**More negative emotion:** words, such as “hate”, “worthless” and “sad” are. Liars are generally more anxious and sometimes feel guilty.

•**Fewer exclusionary words (i.e. less complex speech):** Including “except”, “but” or “nor”--words that indicate that writers distinguish what they did from what they did not do. Liars seem to have a problem with this complexity, and it shows in how they express themselves.

Using the LIWC Dictionary to detect fraud (2)

Why use LIWC?

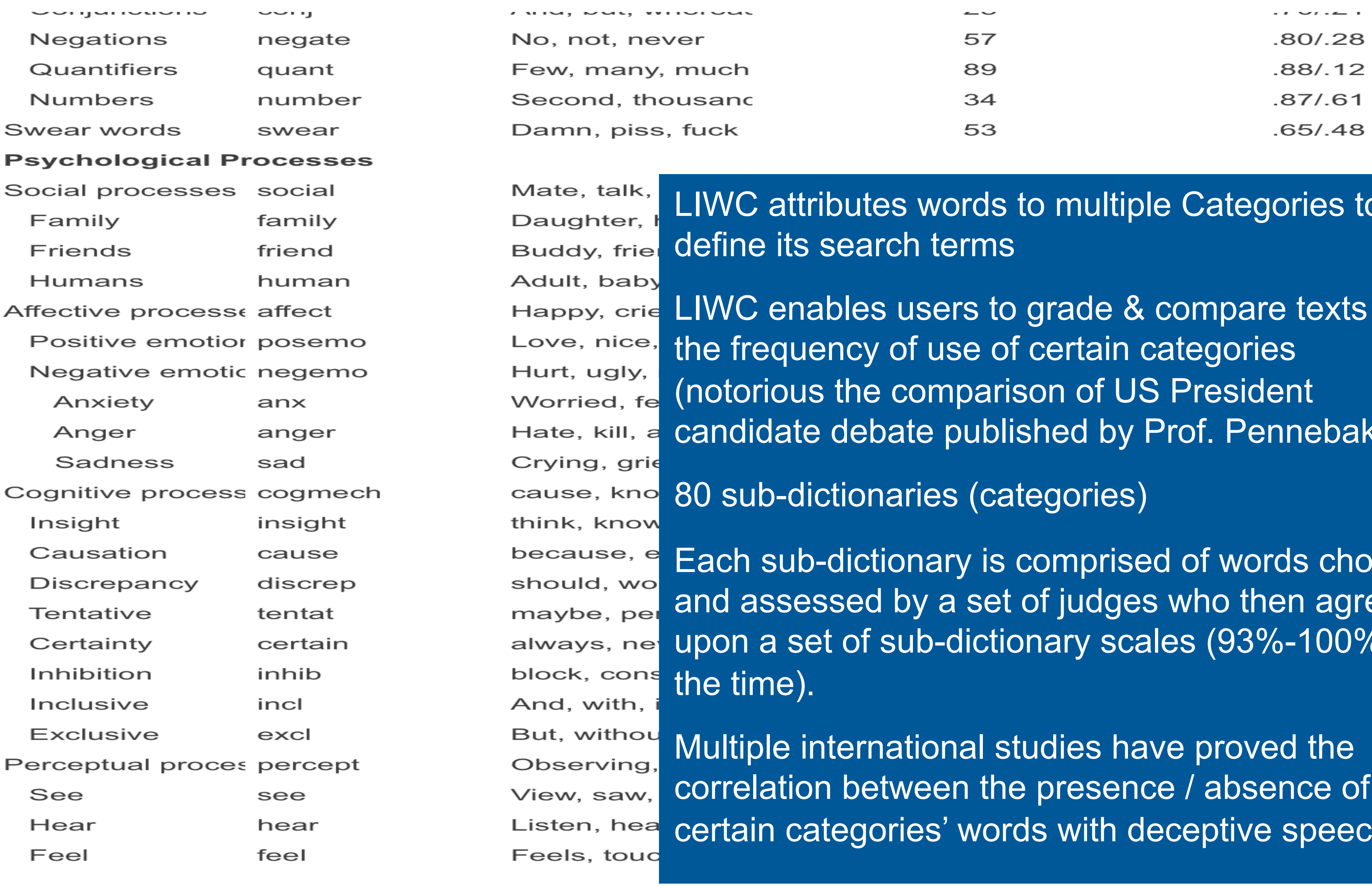
- ➔ Exploit the claim's narrative as part of the claim analytics
- ➔ Add another "dimension" to the claims scoring
- ➔ Extend the analysis to witness statements and any other written / verbal exchange with the claimants (if recorded and digitised)
- ➔ Associated to claimant's attributes (e.g. age; education etc.) promise to offer higher % of success

LIWC has been used in different contexts (and different languages), consistently identifying liars in ~70% of cases

Note – commercial use of LIWC requires licence

Using the LIWC Dictionary to detect fraud (3)

An extract from LIWC



Negations	negate	No, not, never	57	.80/.28
Quantifiers	quant	Few, many, much	89	.88/.12
Numbers	number	Second, thousanc	34	.87/.61
Swear words	swear	Damn, piss, fuck	53	.65/.48
Psychological Processes				
Social processes	social	Mate, talk,		
Family	family	Daughter, h		
Friends	friend	Buddy, frie		
Humans	human	Adult, baby		
Affective processe	affect	Happy, crie		
Positive emotior	posemo	Love, nice,		
Negative emotic	negemo	Hurt, ugly,		
Anxiety	anx	Worried, fe		
Anger	anger	Hate, kill, a		
Sadness	sad	Crying, grie		
Cognitive process	cogmech	cause, kno		
Insight	insight	think, know		
Causation	cause	because, e		
Discrepancy	discrep	should, wo		
Tentative	tentat	maybe, pe		
Certainty	certain	always, ne		
Inhibition	inhib	block, cons		
Inclusive	incl	And, with, i		
Exclusive	excl	But, withou		
Perceptual proces	percept	Observing,		
See	see	View, saw,		
Hear	hear	Listen, hea		
Feel	feel	Feels, touc		

LIWC attributes words to multiple Categories to define its search terms

LIWC enables users to grade & compare texts on the frequency of use of certain categories (notorious the comparison of US President candidate debate published by Prof. Pennebaker)

80 sub-dictionaries (categories)

Each sub-dictionary is comprised of words chosen and assessed by a set of judges who then agreed upon a set of sub-dictionary scales (93%-100% of the time).

Multiple international studies have proved the correlation between the presence / absence of certain categories' words with deceptive speech

LIWC Workflow

1. Store Claim text in a corpus
2. Tokenize Claim narrative
3. Calculate LIWC scoring

➔ As LIWC proposes many categories it is necessary to identify the ones most relevant to the class of claims / types of narrative examined

➔ This is done using a propositional rule learner:

- Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (proposed by William W. Cohen as an optimized version of IREP)
- RIPPER is available in WEKA & accessible in R through RWEKA

The business case for R and enhanced analysis (1)

The business case for R and enhanced fraud analysis

A recent market survey of insurers* in the US found that there was a common recognition of the growing importance of anti fraud activities and technology, and:

88% are employing anti-fraud technology (with under half currently using it for non-claims areas, such as underwriting)

40% are using some form of text analysis for fraud detection

The top two anti fraud areas insurers are looking to invest in are **predictive modelling** and **text mining**

“We are an industry that could be accused of being data rich and knowledge poor” Murli Buluswar, Chief Science Officer at Chartis **

R is perfectly suited to developing advanced text analysis “proof of concepts” with limited resource, which shows whether the cost benefit can be met

* Source: SAS, The State of Insurance Fraud Technology, September 2012

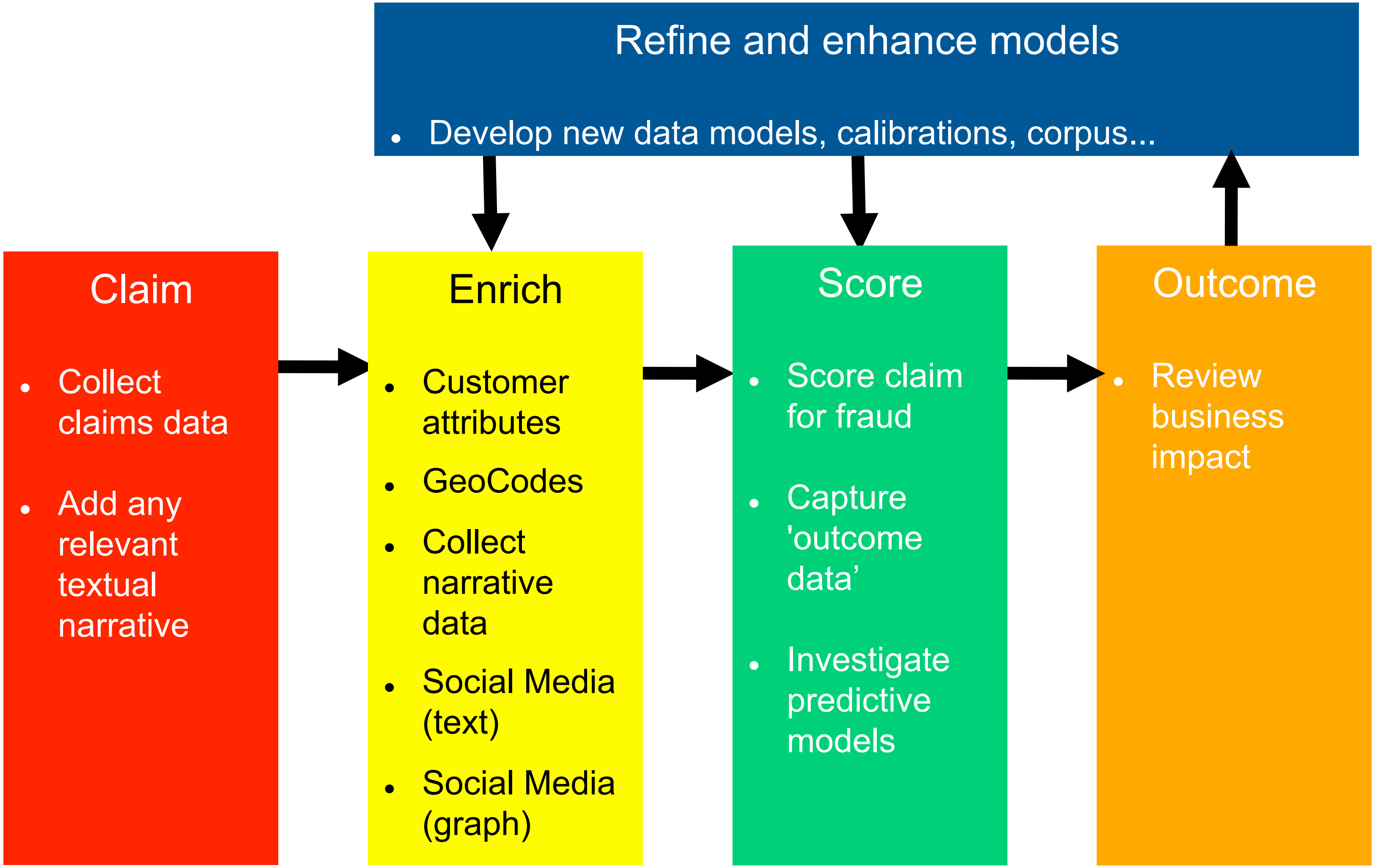
The business case for R and enhanced analysis (2)

Where does R fit and what are the alternatives?

- ➔ Text mining is a growing and increasingly part of insurer’s anti-fraud analysis
- ➔ R is very well suited to help you develop powerful text analytics
- ➔ Text mining is simple to perform to a basic level, but requires careful thought, analysis and domain knowledge to achieve advanced results

- ➔ R (OSS Version) is great for:
  - Rapid analysis and developing custom analysis approaches
  - Building prototype workflows (demonstrating repeatable automation)
  - Text analytics
- ➔ R (OSS version) is not so good for:
  - Large scale text analytics (but hadoop could be easily integrated)
  - Live industrial grade process
  - Working with non-technical users
- ➔ Other flavours of R may enable effective large scale text mining
  - Revolution R, Oracle R, IBM R; SAP R & TIBCO

Summary
Enhancement of the claims workflow



Refine and enhance models

- Develop new data models, calibrations, corpus...

Claim

- Collect claims data
- Add any relevant textual narrative

Enrich

- Customer attributes
- GeoCodes
- Collect narrative data
- Social Media (text)
- Social Media (graph)

Score

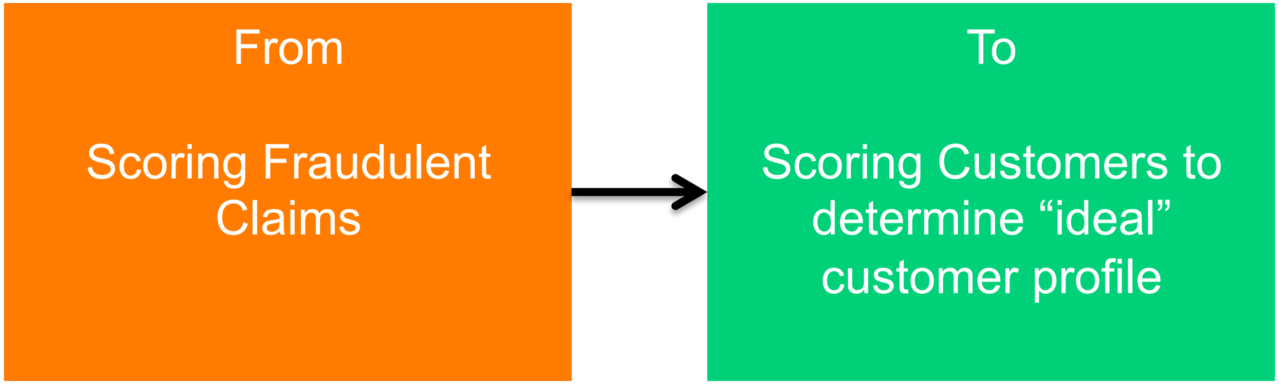
- Score claim for fraud
- Capture 'outcome data’
- Investigate predictive models

Outcome

- Review business impact

What Next?

- Predictive Analytics of fraudulent claims is just one step in the journey



From

Scoring Fraudulent Claims

To

Scoring Customers to determine “ideal” customer profile

- Scoring customers by:
  - “Risk profile” / Risk segmentation
  - Propensity to fraud
- Address historically low post-claim retention rate