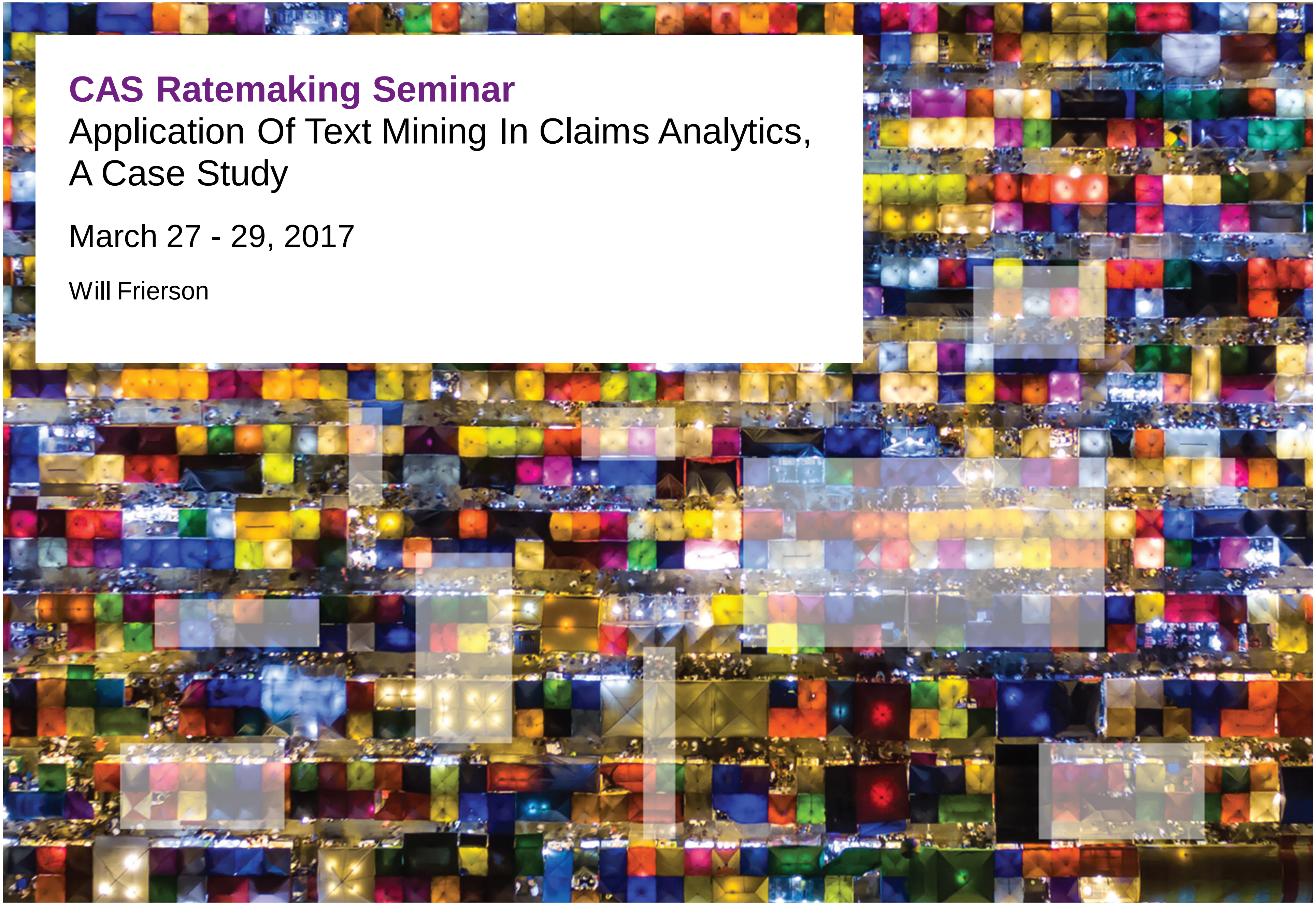
CAS Ratemaking Seminar

Application Of Text Mining In Claims Analytics, A Case Study

March 27 - 29, 2017

Will Frierson



Willis Towers Watson

Table of Contents

- Overview
- Background on Text Mining
- Topic Modeling
- Application of Text Mining and Topic Modeling
- Questions

Overview

- Most insurers have **text data** related to their risks:
  - Loss adjuster notes
  - UW notes
  - Customer feedback
  - Agent notes

- Text data is often **unstructured**, meaning you cannot easily or accurately restate its content as a codified data field
  - When information is extracted from unstructured data, meaning is lost

- Information in unstructured data can **add significant value** to insurance applications
  - More value is added when text data includes detailed descriptions of underlying risks which are not already reflected in existing data fields

Overview

- Traditionally, if an insurer wants to systematically summarize information within text documents, then **word indicators** are used:

Adjuster's Note for Claim 123

Employee injured lower back by **lifting** a heavy box

Restated Adjuster's Note via Word Indicators

Claim #	Surgery Ind	Lift Ind
123	0	1

- Word indicators ignore relationships among words, and so part of a document's **meaning is lost**

Unused Words in Claim 123

Employee injured lower back by ~~lifting~~ a heavy box

Overview

- Advanced text mining techniques like **Topic Modeling** can capture the content and meaning encoded in your text documents by **restating them as a blend of common topics or themes** inferred from your collection of documents

Adjuster's Note for Claim 123

Employee injured lower back by lifting a heavy box

Restated Adjuster's Note via Topic Modeling

Claim #	% of Topic 1	% of Topic 2
123	0.32	0.15

- Topic modeling can **create structured data** from text documents without significant loss of meaning

Overview

- **A topic is a set of co-occurring of words** which can describe specific events or ideas

- E.g., topic describing recurring sci-fi ideas
  - future, technology, space, aliens, science

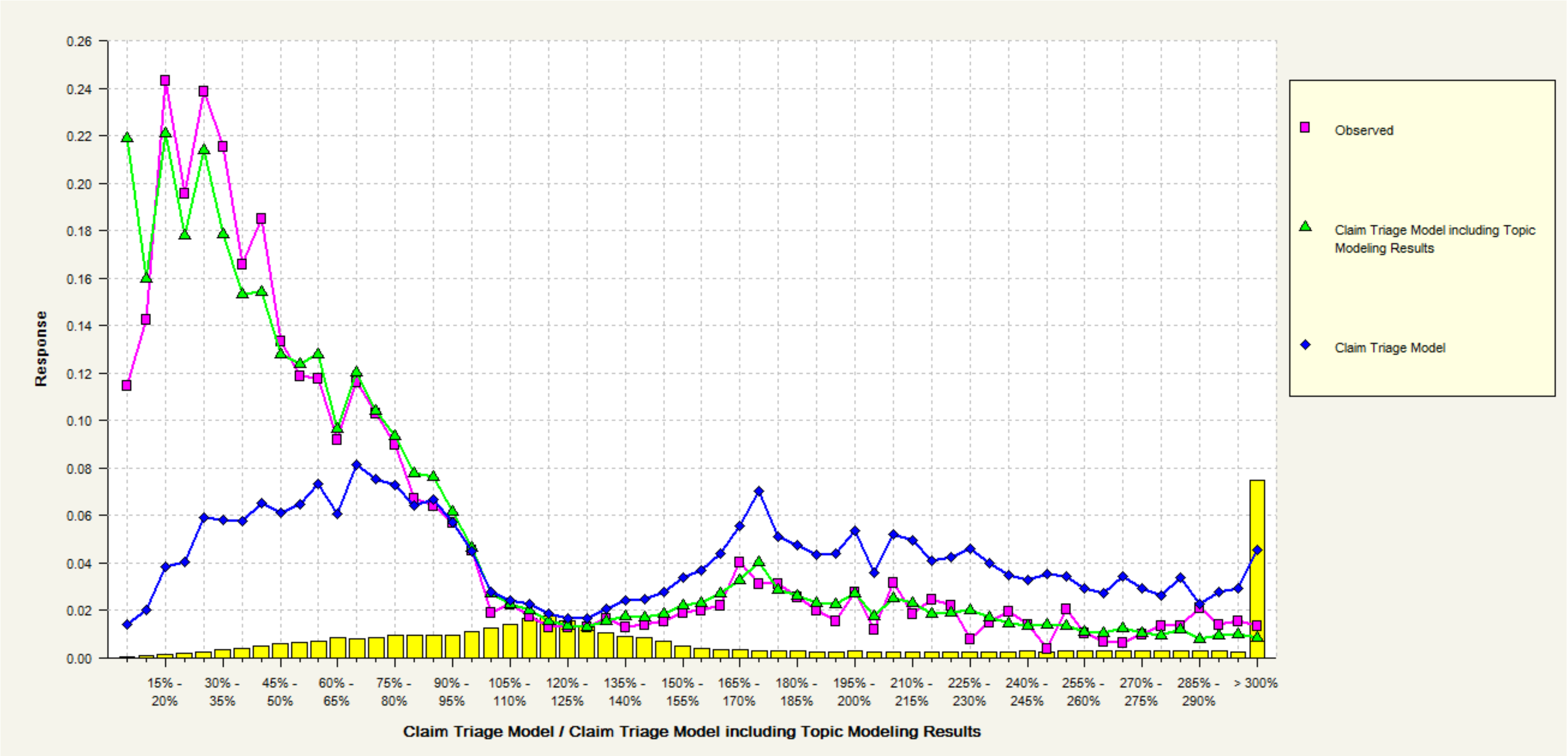
- In an insurance context, topics represent **common events related to the insurance process**

- For loss adjuster notes, topics reflect how:
  - An adjuster handles a claim
  - A claimant recovers from the loss/injury

- For UW notes, topics reflect how:
  - A policyholder relates to, manages, or cares for the insured item
  - An insured item was reviewed and documented

Overview

- Although Topic Modeling is an intricate machine learning algorithm with mathematics not common to actuaries, its results can be used to **build better predictive models**



Classical Text Mining

Background

- **Text mining** is a process of extracting high quality information from unstructured text, e.g., patterns in digitized documents
- Throughout this section, we will examine text from **claim adjuster notes** to understand how text mining can be used in insurance

Background

- **Example of unstructured text**
- PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report back sprain injury to his supervisor until D/L NLT…

Background

- **Example of unstructured text**

- PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report back sprain injury to his supervisor until D/L NLT…

- Problems with unstructured data
  - Junk words, numbers, and formatting

Background

- **Example of unstructured text**

- **PC** to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his **back** and **strain** his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report **back** sprain injury to his supervisor until D/L NLT…

- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)

Background

- **Example of unstructured text**

- PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and **strain** his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report back **sprain** injury to his supervisor until D/L NLT…

- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)
  - Many words with the same meaning (synonymy)

Background

- **Example of unstructured text**
- PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did **not** report back sprain injury to his supervisor until D/L NLT…
- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)
  - Many words with the same meaning (synonymy)
  - Negation

Background

- **Example of unstructured text**
- **PC** to Jane Doe/**insd**: DOI: 01/01/16 **Clmt** was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing **clmt** to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith **Clmt** did not report back sprain injury to his supervisor until **D/L NLT**…
- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)
  - Many words with the same meaning (synonymy)
  - Negation
  - Abbreviations

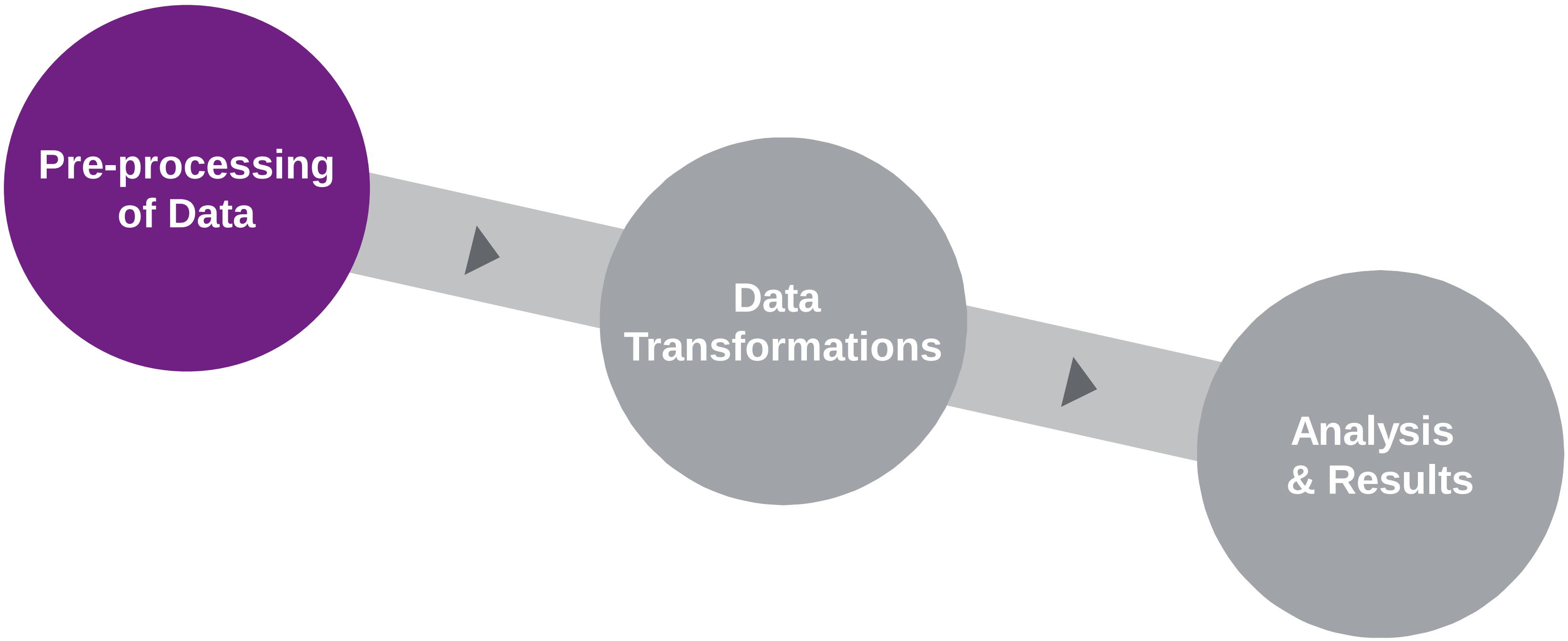
Background

- **Example of unstructured text**

- PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report back sprain injury to his supervisor until D/L NLT…

- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)
  - Many words with the same meaning (synonymy)
  - Negation
  - Abbreviations
  - Curse of dimensionality
    - This one claim has >1100 words
    - Number of unique words for a set of claims is massive!

Steps in Text Mining



Pre-processing

Overview

- **Primary purpose of pre-processing is to "clean" text data**

- **Reduce complexity**
  - If there are 20K distinct words used in a set of documents and an average document contains 1K words, then there are $\mathbf{1K^{20K} = 10^{60K}}$ possible documents
  - Number of atoms in the observable universe, ~$\mathbf{10^{80}}$
- **Enhance core relationships**
  - Combine words, phrases, or acronyms that are assumed to have the same meaning for your application

- **Secondary purpose is to standardize text data**

Pre-processing

Examples

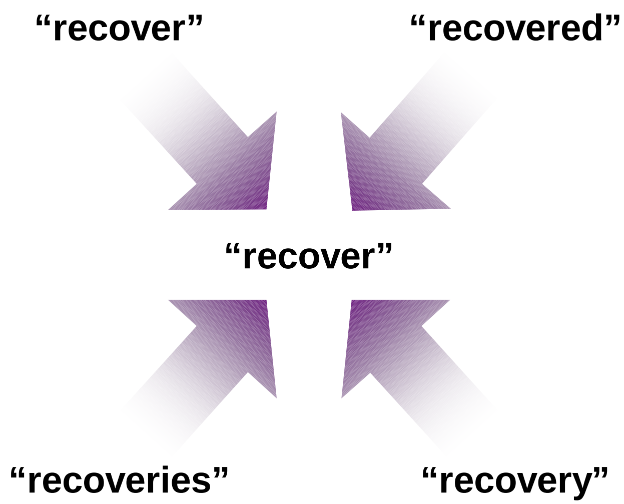
- **Remove characters, words, and phrases you assume to be irrelevant for your analysis**
  - "Stop words" (extremely frequent words that carry little meaning, e.g., "the")
  - Generic insurance words, e.g., "claim" in claim adjuster notes
  - Common first and last names (to prevent over-fitting)
  - Punctuation, numbers, whitespace, etc.
- **Make lower case for consistency**
- **Remove short and rare words**

carrying drywall steps worker reached top stair steps started walk faster causing twisted back strain shoulder pain mid back shoulder incident witnessed report sprain injury supervisor…

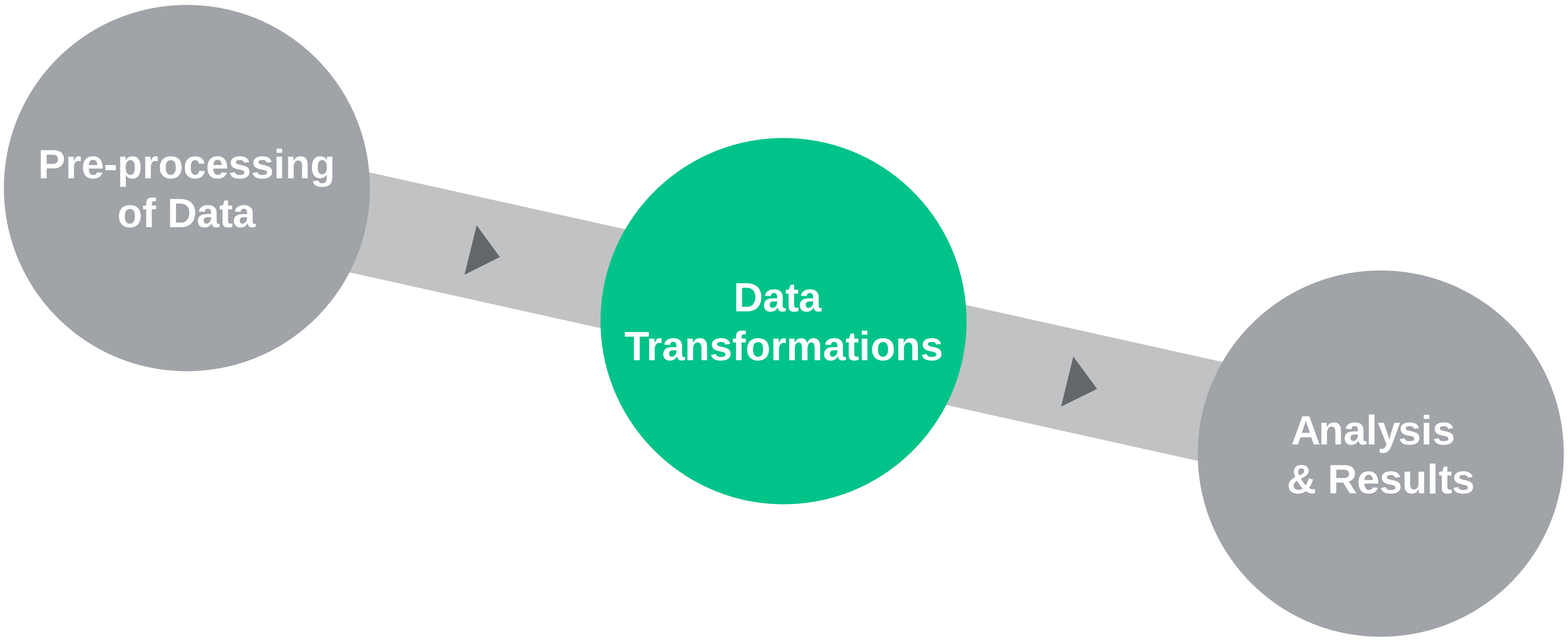
Pre-processing

Examples

- Apply a **stemming** procedure
  - Map conjugations and declensions to their root word



Steps in Text Mining



Data Transformations

- Create a matrix that contains selected information about the relationships among terms and documents from your collection
  - Often called "**Bag of Words**" or "**Document-Term Matrix**"
    - Using vectors and matrices allows use of linear algebra, which aids in computational efficiency
    - Phrases can also be examined, either with single-word terms or in isolation

	Claim 1	**Claim 2**
"claim"		
"surgery"		

Data Transformations

- Relationship among terms and documents can be stored in different ways
  - **Binary**
    - Whether a given term is present at all in a document (1 or 0)

Binary	Claim 1	Claim 2
“claim”	1	1
“surgery”	1	0

Data Transformations

- Relationship among terms and documents can be stored in different ways
  - Binary
  - **Term frequency (Tf)**
    - How often a given term appears in a document
    - Can be in absolute counts or relative to number of terms in a document

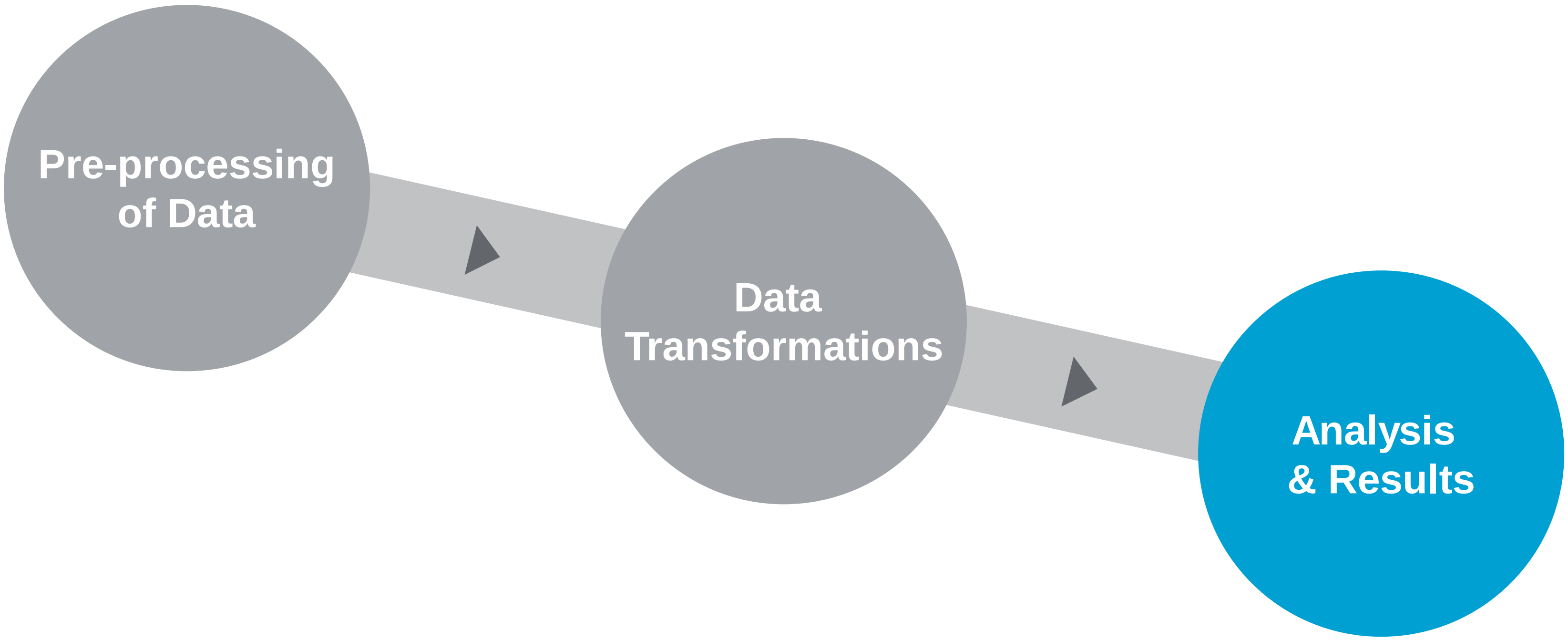
Tf	**Claim 1**	**Claim 2**
“claim”	25	50
“surgery”	3	0

Data Transformations

- Relationship among terms and documents can be stored in different ways
  - Binary
  - Term frequency (Tf)
  - **Term frequency – Inverse document frequency (Tf-Idf)**
    - A statistic intended to reflect a word's importance in a document
    - Term frequency is multiplied by the negative log of % documents that contain a given term. This factor increases frequencies for rare words in a document set.

Tf-Idf	Claim 1	Claim 2	% in All Documents
"claim"	25 x –log(0.99) = **0.25**	50 x –log(0.99) = **0.50**	0.99
"surgery"	3 x –log(0.05) = **9**	0 x –log(0.05) = **0**	0.05

Steps in Text Mining



Analysis & Results

- Methods for exploring document-term matrix
  - Create **Word Cloud**, where the most important words are displayed and scaled by their relative importance

Analysis & Results

Advanced text mining methods

- Use a variable reduction method to select important words or phrases to use in a predictive model
  - Pros: Easy to automate and relatively quick to complete
  - Cons: Cannot account for relationships among words, i.e., **topics**

Analysis & Results

Advanced text mining methods

- Use a variable reduction method to select important words or phrases to use in a predictive model

- Linear algebra based methods
  - Latent Semantic Analysis (i.e., rank reduction of document-term matrix via PCA)
  - Non-negative matrix factorization (like LSA but components are positive)
    - Pros: Easy way to reduce dimensionality. Word combinations can have semantic meaning
    - Cons: Does not scale easily. Word combinations often have no semantic meaning

Analysis & Results

Advanced text mining methods

- Use a variable reduction method to select important words or phrases to use in a predictive model

- Linear algebra based methods
  - Latent Semantic Analysis (i.e., rank reduction of document-term matrix via PCA)
  - Non-negative matrix factorization (like LSA but components are positive)

- Probability based methods
  - Probabilistic Latent Semantic Analysis (model word co-occurrence under a probabilistic framework)
    - Pros: Results have statistical meaning and can be more important than those from LSA
    - Cons: Number of parameters grows with number of documents. Provides no practical application for new documents

Analysis & Results

Advanced text mining methods

- Use a variable reduction method to select important words or phrases to use in a predictive model
- Linear algebra based methods
  - Latent Semantic Analysis (i.e., rank reduction of document-term matrix via PCA)
  - Non-negative matrix factorization (like LSA but components are positive)
- Probability based methods
  - Probabilistic Latent Semantic Analysis (model word co-occurrence under a probabilistic framework)
  - Topic Modeling

Topic Modeling

Big Picture of Topic Modeling

- Goal of topic modeling is to discover the **hidden thematic structure** in a large set of documents using posterior inference
- Documents are assumed to exhibit traits from multiple topics with **different topic proportions**,
  i.e., *mixed-membership model*
- Topic modeling:
  - Automates the annotation of a set of documents
  - Does not require any prior annotation or labeling of documents, i.e., *unsupervised*
- Topic modeling represents **a core idea with many different versions**
  - Like Regression, different versions include OLS, GLM, Ridge, Lasso, and Elastic Nets
  - Like CART, different versions include Gradient Boosting and Random Forests

Latent Dirichlet Allocation, D. Blei et al. 2003

What is a topic?

- A topic is a **probability distribution over a fixed vocabulary**

	Topic 1	Topic 2
claim	0.05	0.05
arm	0.30	0.01
leg	0.01	0.40
…	…	…

- We can understand a topic by examining its **most likely words** (as well as other methods discussed later)

Topic	laceration	sutures	removal	hospital	stitches	feet	issued	wound	complete	injuring

Latent Dirichlet Allocation, D. Blei et al. 2003

What is a topic?

- Topics are not guaranteed to be constructed in a meaningful way. Unless certain operations are performed, topics can be impacted by the following issues:

  - **Randomness**: no semantic coherence among likely words, i.e., “junk”
    - box, reserve, car, wrote, worker

  - **Word Chains**: likely words are related through pair-wise word chains
    - box, lift, cutter, container

  - **Word Intrusion**: likely words are semantically coherent and reasonable, except for a few likely words which appear to have no relationship with the other likely words
    - laceration, sutures, removal, banana

Optimizing Semantic Coherence in Topic Models, D. Mimno et al. 2011

Core Topic Modeling Algorithm

Motivation

- To find topics, we need to **identify sets of co-occurring words**

carrying drywall steps worker reached top stair steps started walk faster causing twisted back strain shoulder pain mid back shoulder incident witnessed report sprain injury supervisor…

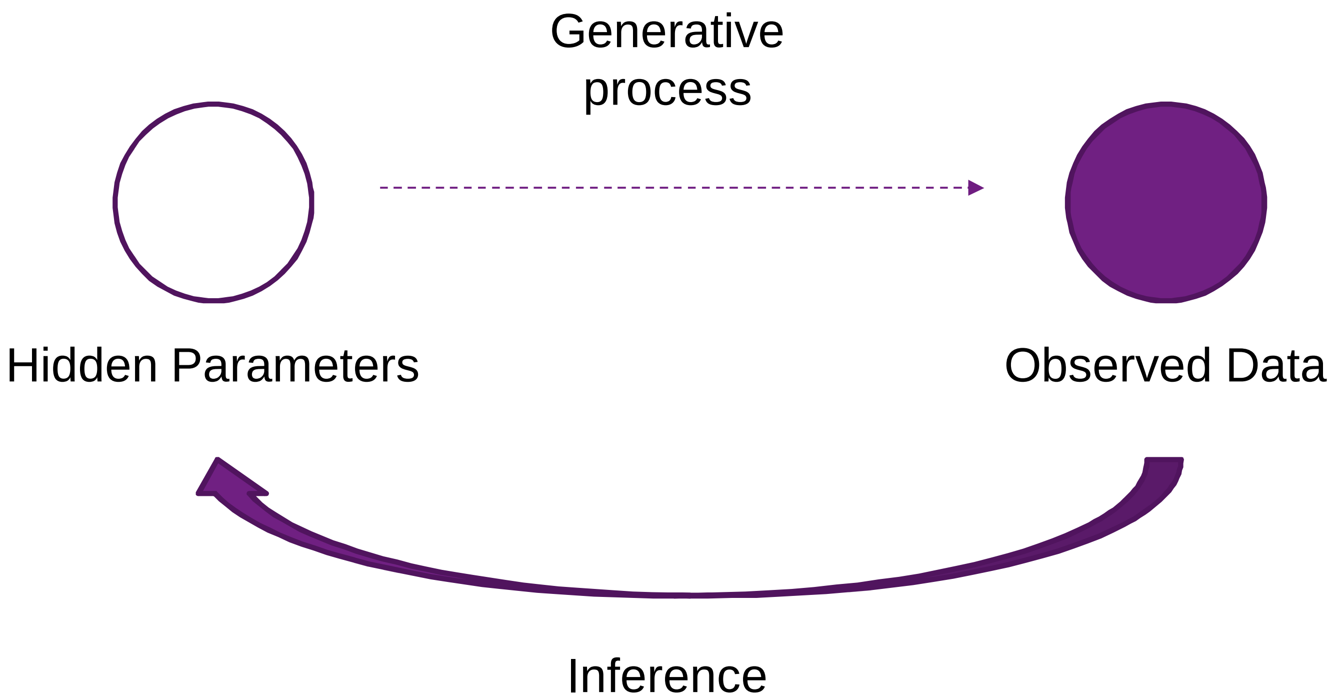
- Assuming specific relationships and structures gives a more practical framework to search for topics

- Compare with GLMs:
  - Assume a link function and an error structure
  - Construct a likelihood function
  - Use numerical methods to estimate parameter values

Latent Dirichlet Allocation, D. Blei et al. 2003

Core Topic Modeling Algorithm
Motivation

- Topic modeling is a **generative probabilistic model** for a set of documents

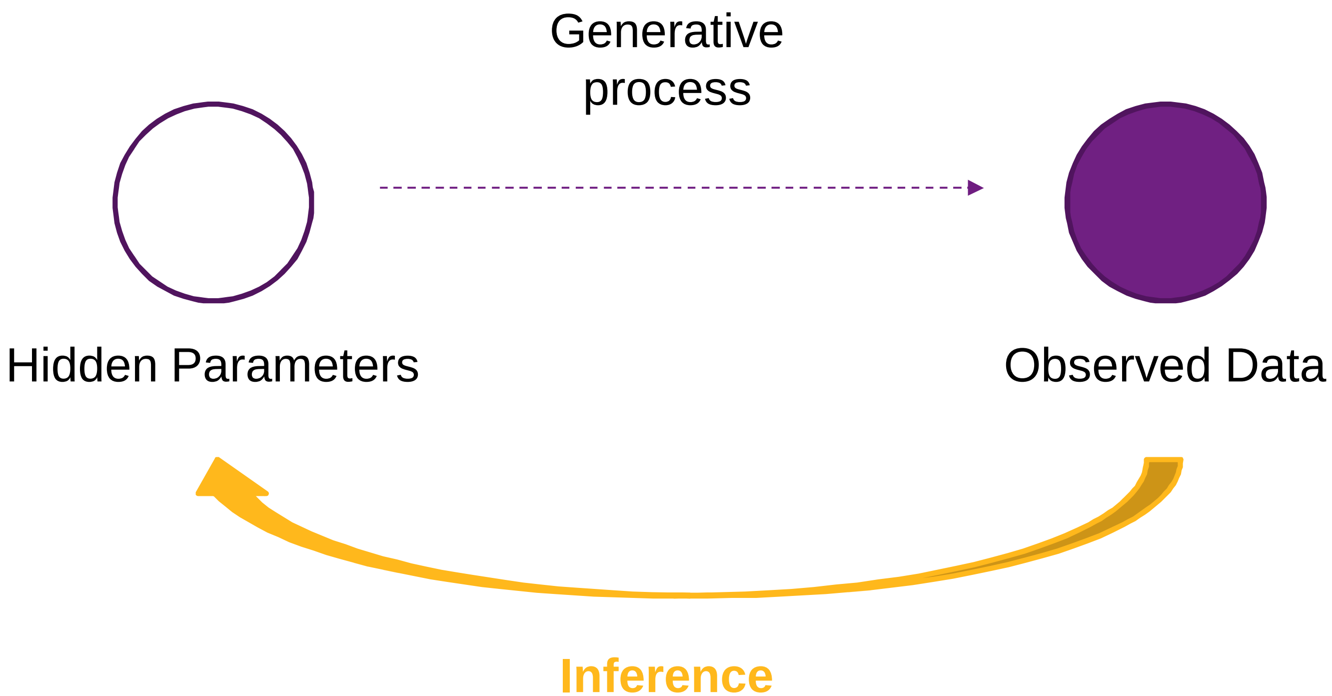


Latent Dirichlet Allocation, D. Blei et al. 2003

Core Topic Modeling Algorithm

Motivation

- Topic modeling is a **generative probabilistic model** for a set of documents



Latent Dirichlet Allocation, D. Blei et al. 2003

Core Topic Modeling Algorithm

- **Latent Dirichlet Allocation (LDA)** is commonly treated as the core topic modeling algorithm

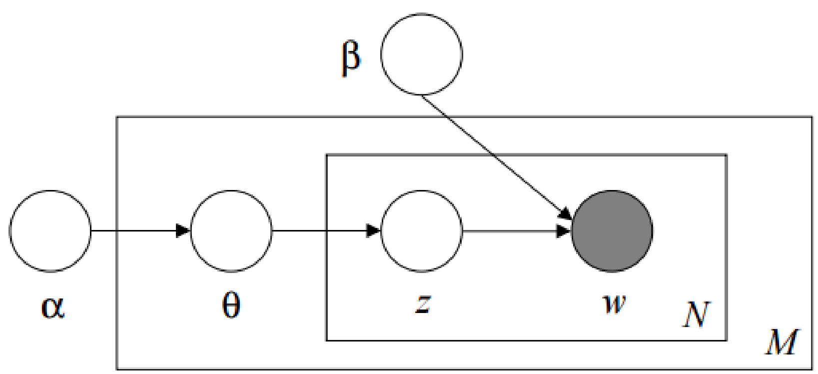
- The only input from the user is the number of topics, **k**

- Other relevant parameters include:
  - Vocabulary of words used across documents, **V**
  - Number of documents, **M**
  - Number of words in each document, **N**
  - Matrix of document-topic proportions, $\boldsymbol{\theta}_{\mathbf{M \times k}}$
  - Matrix of topic-word proportions, $\boldsymbol{\beta}_{\mathbf{k \times V}}$
  - Smoothing parameters, $\boldsymbol{\alpha}$ (for $\boldsymbol{\theta}$) and $\boldsymbol{\eta}$ (for $\boldsymbol{\beta}$)

Latent Dirichlet Allocation, D. Blei et al. 2003

Core Topic Modeling Algorithm

- LDA is a **generative probabilistic model** for a set of documents
- We assume each document is generated as follows:
  - Fix topic-word proportions, **β** ~ Dirichlet(η)
  - Fix document-topic proportions, **θ** ~ Dirichlet(α)
  - For each of the N words in a document:
    - Choose a topic z_N ~ Multinomial(**θ**)
    - Choose a word w_N from a multinomial probability distribution conditioned on the topic z_N and the topic-word proportions



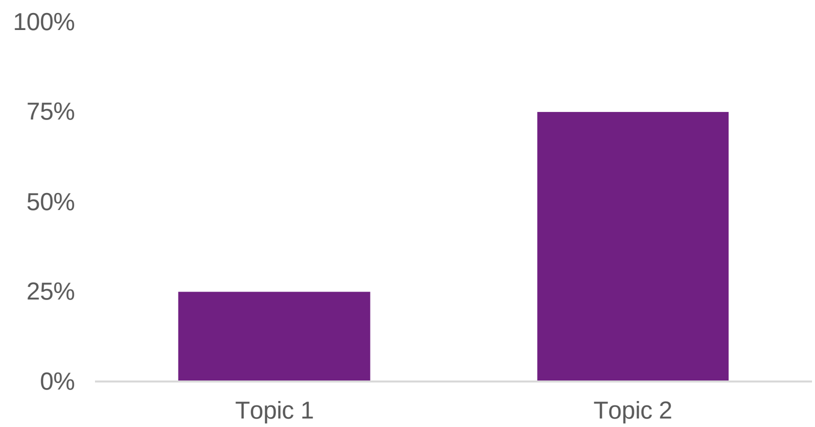
Latent Dirichlet Allocation, D. Blei et al. 2003

Example Generative Process from LDA

Dirichlet Process or "Draw from two hats"

- For Document X:
  - We assume we are given **β** and **θ** values
- Word 1:
  - **Topic Selection**:
    Take 1 sample from
    $z_N \sim$ Multinomial(θ**)**
    - Result is Topic 1
  - **Word Selection**:
    Take 1 sample from multinomial conditioned on z_N and **β**
    - Result is "arm"
- Word 1 is "**arm**"
  - P["arm" | topic = 1, β] x P[topic = 1 | θ_X]
  - 25% x 30% = 7.5%

Topic Proportions for Document X, **θ**



Topic-word Proportions, **β**

	Topic 1	Topic 2
claim	0.05	0.05
arm	0.30	0.01
leg	0.01	0.40
…	…	…

Latent Dirichlet Allocation, D. Blei et al. 2003

Extensions to LDA

- Depending on your application, topic modeling results are of higher quality when using extensions to LDA

- **FREX scoring** for individual words
  - Calculate measure of **fr**equency and **ex**clusivity for a given word
  - Helps prevent topics from being too similar, i.e., more efficient results
- **Semantic coherence**
  - A measure to prevent word chains, intrusions, and random topics
  - Similar concept as Tf-Idf
- **Correlated Topic Models (CTM)**
  - LDA assumes there are no correlations among topics
  - However, we expect certain claim types to appear in the same claim, e.g., lower back injury (topic 1) preceding litigation against employer (topic 2)
  - CTM allows latent topic correlations to be inferred, yielding higher quality results
- **Structural Topic Models (STM)**
  - Uses CTM framework and allows document metadata to influence the document-topic proportions (i.e., prevalence model), topic-word proportions (i.e., content model), or both simultaneously

Considerations for Topic Modeling

- **Topic modeling is not a simple exercise.** Many of its components require careful consideration depending on your application.

- **Considerations**:
  - Topic modeling results can be obtained via different numerical methods. Each has its own pros and cons.
    - Variational inference, collapsed Gibbs sampling, semi-collapsed variational EM

  - Posterior probability is intractable, and so the likelihood function must be approximated

  - CTM and STM fit times can take *days* to converge, depending on your data set, number of topics **k**, STM covariates, etc.

  - Difficult to find optimal number of topics, **k**. Automated methods exist, but there are no guarantee that the resulting **k** value is appropriate for your application, e.g., whether the results are credible enough for actuarial use

Case Study: Claims Triaging Model

Claims Triaging Model

- We built a claims triaging GLM for a workers compensation LOB to estimate the propensity of claims to develop adversely as of a given day after first notice of loss
  - i.e., provide estimates on whether a claim would "blow up" given the information available at a certain point in time
- The data set had >> 10K claims over nearly a decade of experience
- This model examined possible predictors from many data sources such as:
  - **Claim-level data**: claim status, subrogation flag, insurer's internal description of claim (i.e., their guess at topics)
  - **Claimant-level data**: claimant age & location
  - **Medical-level data**: # prescribed drugs, # procedures, flags for specific medical issues
  - **Text mining results**: word indicators for specific issues/events described in claim adjuster notes

Fitting Topic Models

- To test the predictive power of topic modeling, I fit various correlated and structural topic models
  - k = 100 or 200 topics
    - CTM
    - STM with prevalence model using the target of the GLM
    - STM with content model using the target of the GLM
    - STM with both prevalence and content models using the target of the GLM

- Using the GLM's target as a covariate in STM effectively makes a supervised topic model

Topic Modeling Results

Incorporating results into claims triaging model

- Fitted document-topic proportions **θ** were extracted from all topic models and joined back to the experience data

	Topic 1	Topic 2
Claim 1	0.05	0.05
Claim 2	0.30	0.01
Claim 3	0.01	0.40
…	…	…

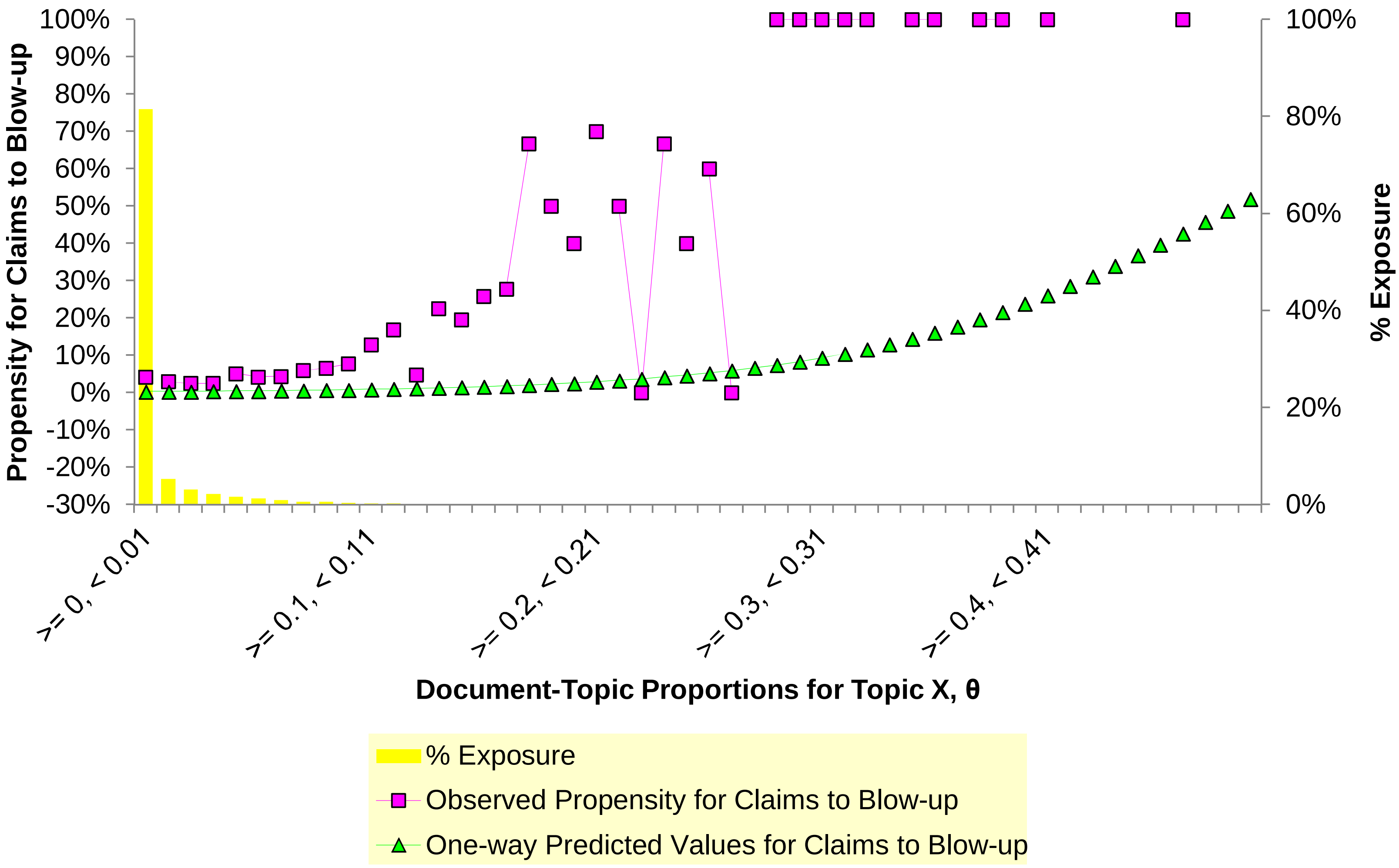
- New GLMs were built by keeping all prior predictors and creating new topic predictors reflecting the fitted document-topic proportions

- Variable reduction and manual methods were used to find important and predictive topic model factors

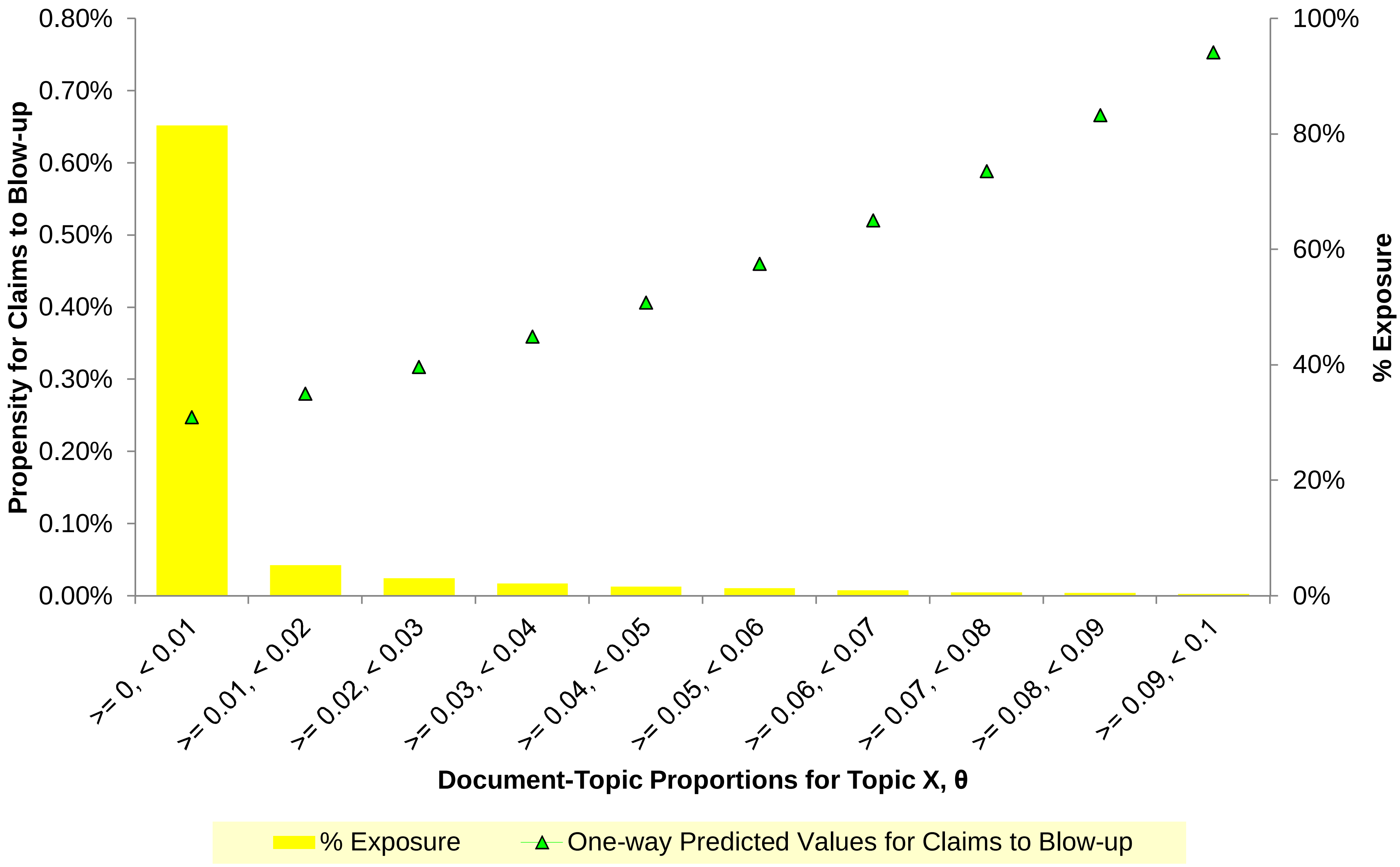
Topic Modeling Results

Topic 1	dentist	tooth	dental	teeth	lip	rebar	patent	crown	jaw
Topic 2	lifting	felt	muscle	heavier	pulled	weighs	lbs	strain	pop
Topic 3	herniated	esi	disc	stenosis	epidural	spine	neuro surgeon	bulge	fusion

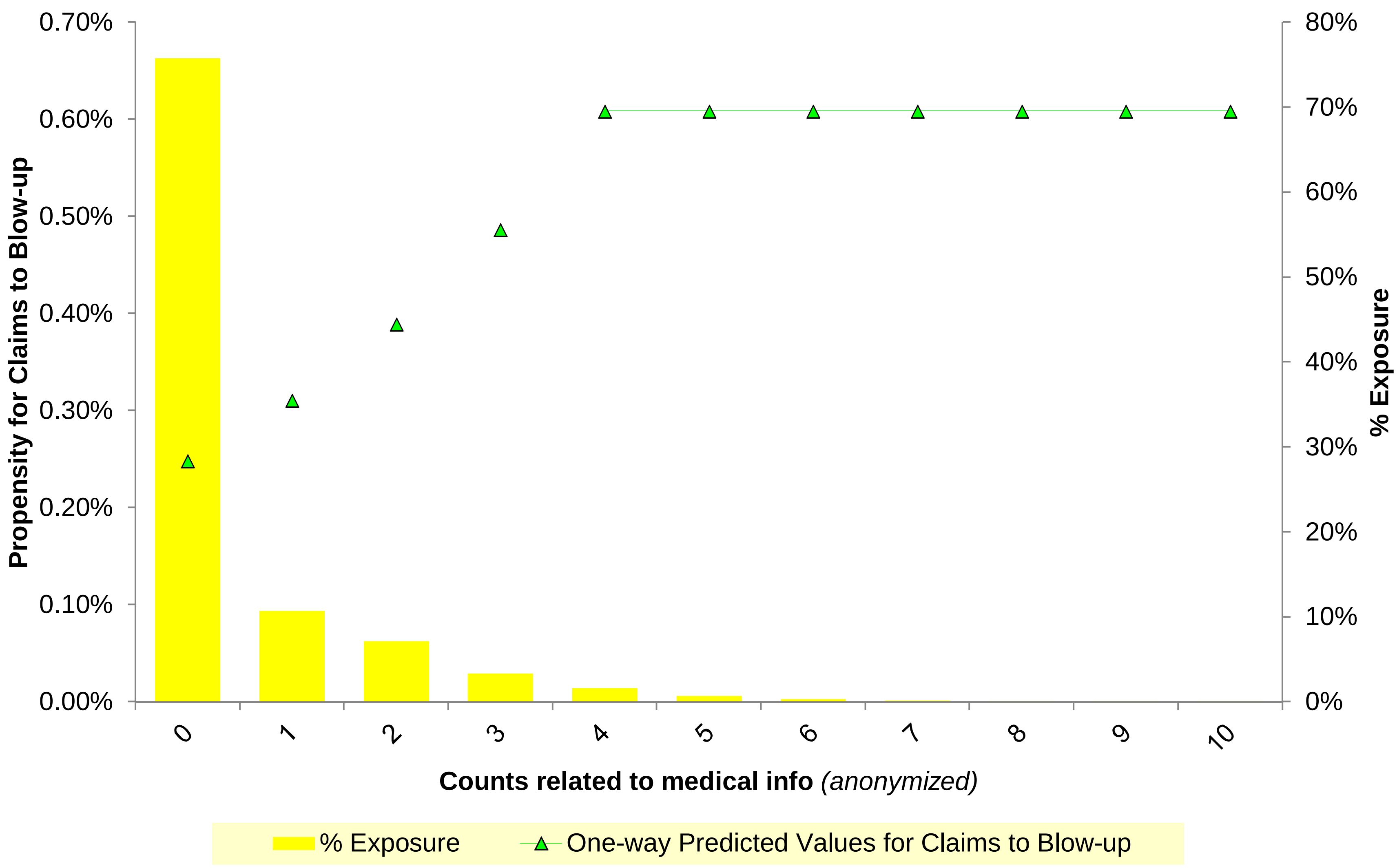
Topic Modeling Results



Topic Modeling Results



Topic Modeling Results



Topic Modeling Results

Topic Model used in GLM	Number of Topics
CTM	100
STM, C	100
STM, P	100
STM, P&C	100
CTM	200
STM, C	200
STM, P	200
STM, P&C	200

- “P” indicates a prevalence model was used in STM
- “C” indicates a content model was used in STM
- All included topic factors were time consistent

Topic Modeling Results

Topic Model used in GLM	Number of Topics	Number of Important and Included Topic Predictors
CTM	100	6
STM, C	100	7
STM, P	100	8
STM, P&C	100	6
CTM	200	1
STM, C	200	5
STM, P	200	2
STM, P&C	200	7

- "P" indicates a prevalence model was used in STM
- "C" indicates a content model was used in STM
- All included topic factors were time consistent

Topic Modeling Results

Topic Model used in GLM	Number of Topics	Number of Important and Included Topic Predictors	% of Included Topic Factors in Top 10 Most Important Factors (via Backwards Regression AIC)
CTM	100	6	17%
STM, C	100	7	86%
STM, P	100	8	50%
STM, P&C	100	6	83%
CTM	200	1	0%
STM, C	200	5	60%
STM, P	200	2	100%
STM, P&C	200	7	71%

- “P” indicates a prevalence model was used in STM
- “C” indicates a content model was used in STM
- All included topic factors were time consistent

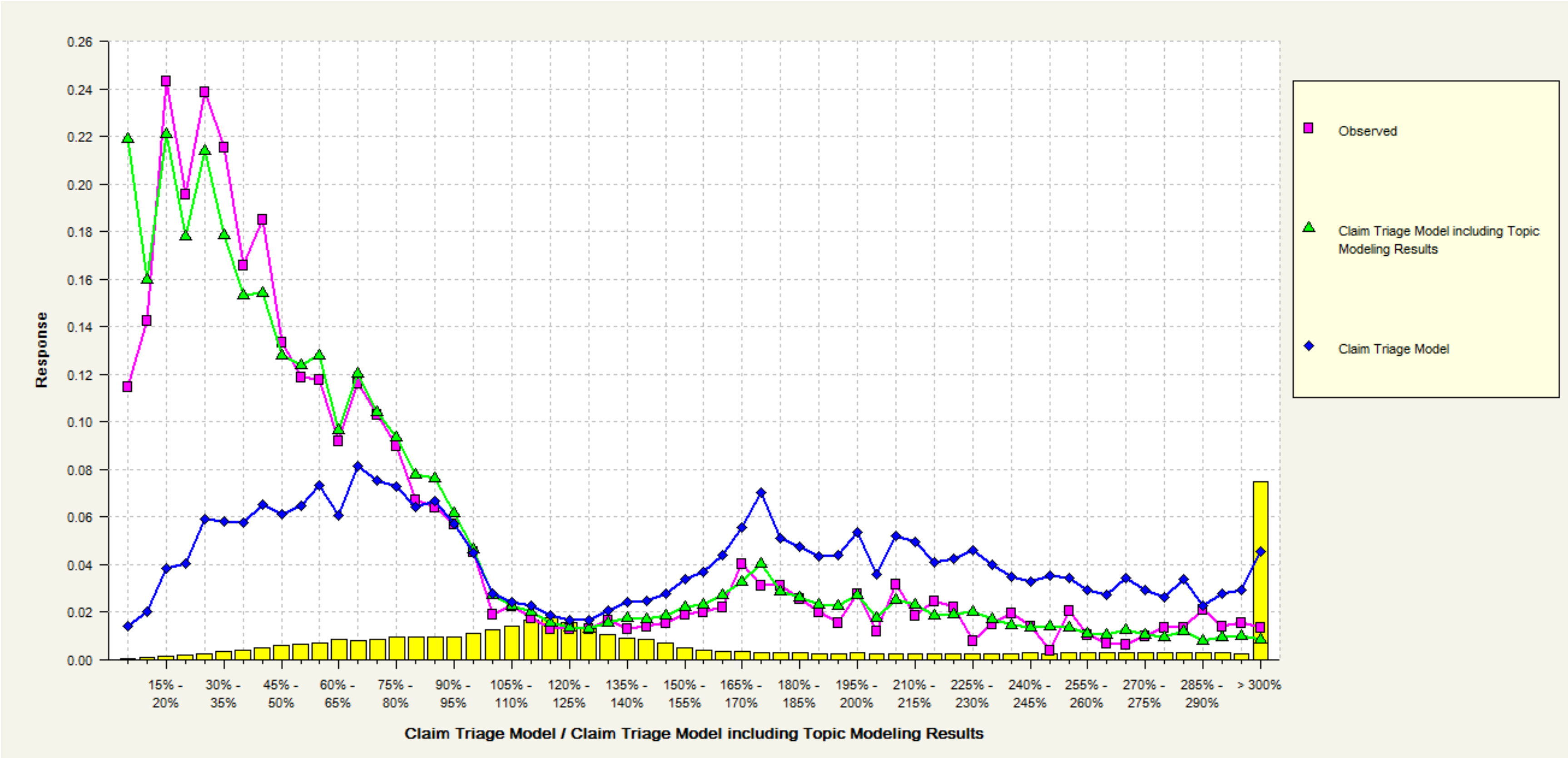
Topic Modeling Results

Topic Model used in GLM	Number of Topics	Number of Important and Included Topic Predictors	% of Included Topic Factors in Top 10 Most Important Factors (via Backwards Regression AIC)
CTM	100	6	17%
STM, C	100	7	86%
STM, P	100	8	50%
STM, P&C	100	6	83%
CTM	200	1	0%
STM, C	200	5	60%
STM, P	200	2	100%
STM, P&C	200	7	71%

- **Summary**:
  - The included topic factors were statistically significant, time consistent, semantically coherent & reasonable for this application, **as well as more important than many factors already in the model (including word indicators)**

Topic Modeling Results

- Double lift chart example from GLM including results from 100-topic STM with (supervised) content covariate



Summary

- Many insurers have text data containing valuable information not already reflected in standard insurance databases
- Advanced text mining techniques like topic modeling can restate unstructured text data as structured numeric data without a significant loss of meaning
- Topic modeling results can provide significant lift to predictive models and other insurance applications

Questions

Thank you