Francis Analytics

Actuarial Data Mining Services

Introduction to Text Mining

Insightful Users Conference
October 27, 2005.

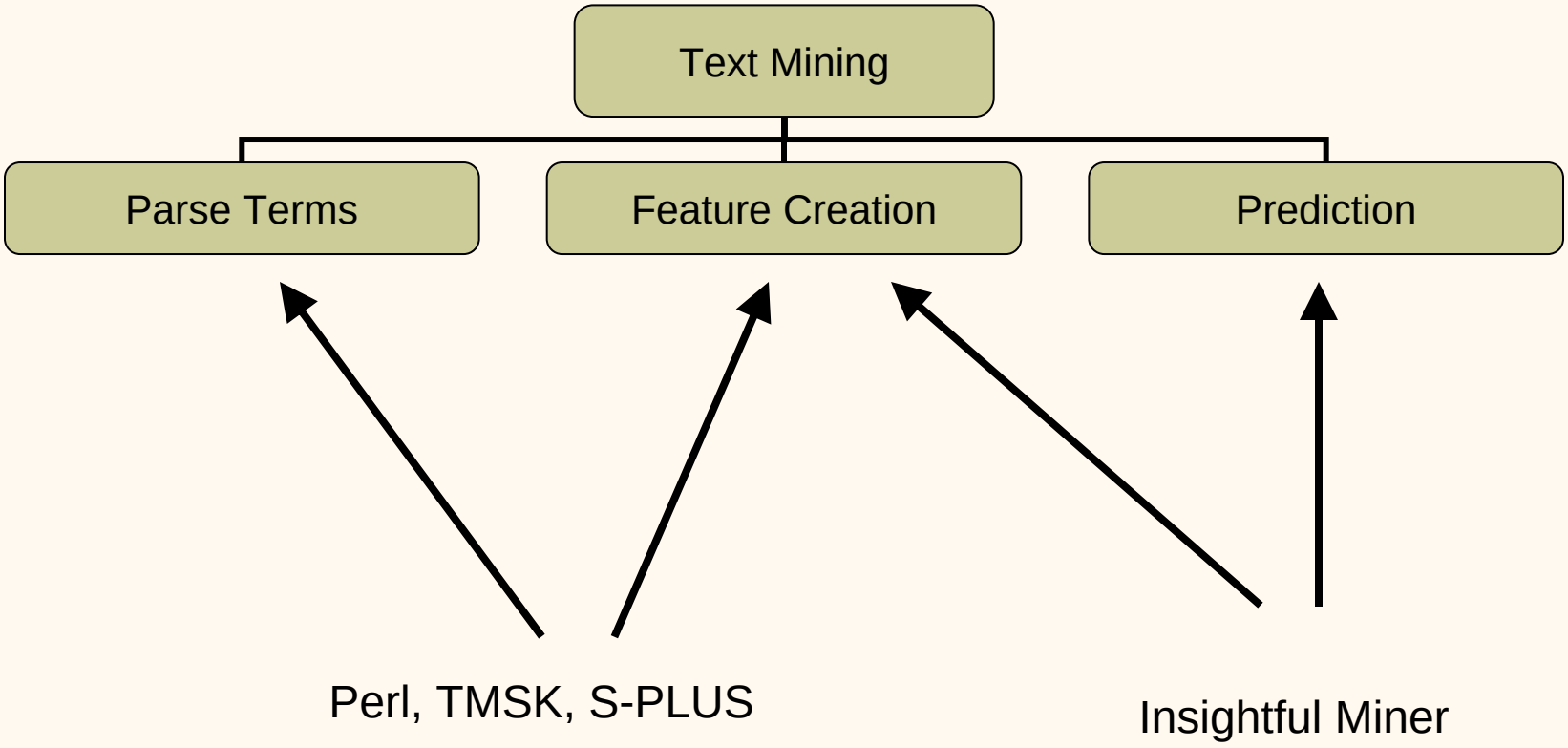
Prepared by: Louise Francis, MCAS, FAAA
Louise_francis@data-mines.com

Francis Analytics and Actuarial Data Mining, Inc
I

Objectives

- Gentle introduction to Text Mining
  - Give insights into how it works
- Illustrate a simple application to insurance claims data
- Show examples using free software (Perl, TMSK) as well as S-PLUS and Insightful Miner

Text Mining Process



Parse Text Into Terms

- Separate free form text into words
- "BROKEN ANKLE AND SPRAINED WRIST" →
  - BROKEN
  - ANKLE
  - AND
  - SPRAINED
  - WRIST

Parsing Text

- Separate words from spaces and punctuation
- Clean up
- Remove redundant words
- Remove words with no content

String Functions

- Use substring function in S-PLUS to find spaces

```
# Initialize
charcount<-nchar(Description)
# number of records of text
Linecount<-length(Description)
Num<-Linecount*6
# Array to hold location of spaces
Position<-rep(0,Num)
dim(Position)<-c(Linecount,6)
# Array for Terms
Terms<-rep(“”,Num)
dim(Terms)<-c(Linecount,6)
wordcount<-rep(0,Linecount)
```

Search for Spaces

```
for (i in 1:Linecount)
{
n<-charcount[i]
k<-1
for (j in 1:n)
{
                Char<-substring(Description[i],j,j)
                  if (is.all.white(Char)) {Position[i,k]<-j; k<-k+1}
                  wordcount[i]<-k
}
}
```

Get Words

```
# parse out terms
for (i in 1:Linecount)
{
        # first word
        if (Position[i,1]==0) Terms[i,1]<-Description[i]  else if (Position[i,1]>0)
        Terms[i,1]<-substring(Description[i],1,Position[i,j]-1)
        for (j in 1:wordcount)
        {
        if (Position[i,j]>0)
        {
        Terms[i,j]<-substring(Description[i],Position[i,j-1]+1,Position[i,j]-1)
        }
        }
}
```

Perl

- Free open source programming language
- www.perl.org
- Used a lot for text processing
- *Perl for Dummies* gives a good introduction

Perl Functions for Parsing

- $TheFile ="GLClaims.txt";
- $Linelength=length($TheFile);
- open(INFILE, $TheFile) or die "File not found";
- # Initialize variables
- $Linecount=0;
- @alllines=();
- while(<INFILE>){
- $Theline=$_;
- chomp($Theline);
- $Linecount = $Linecount+1;
- $Linelength=length($Theline);
- @Newitems = **split(/ /,$Theline);**
- print "@Newitems \n";
- push(@alllines, [@Newitems]);
- } # end while

Stopwords and Stemwords

- Stopwords – frequently occurring words with little real content: a, the , to ,and
  - Eliminate from list of terms
- Stemwords – singular and plural forms of same word, and synonym
  - Knee, knees
  - Fracture, broken
  - Replace multiple variants with just one word

Vectorization

- End result of parsing, stopwords and stemming is a matrix of binary indicator variables

back	contusion	head	knee	strain	unknown	laceration	leg
1	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	1	0	1	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0

TMSK Text Mining Software

- Does parsing, stopwords, stemwords, vectorization and statistical analysis
- Available free if you buy *Text Mining* by Weiss et al., Springer 2005
- Must have java on your computer
  - Download from Sun Microsystems web site at www.sun.com
- It only reads xml files
  - Use Excel or Acrobat to save as xml

Feature creation: Dimension reduction

- Next step: Create a new feature (variable) in the data that is an injury code and can be used to predict outcomes of interest
- Cluster records with similar injuries or injury terms with similar meaning together: these get the same injury code
- Use unsupervised learning or dimension reduction to do this
- From Insightful Miner use k-means clustering and Principal Components

Dimension Reduction: Column-wise and Row-wise

CLAIM NUMBER	DATE OF LOSS	STATUS	INCURRED LOSS
1998001	09/15/97	C	407.81
1998002	09/25/97	C	0.00
1998003	09/26/97	C	0.00
1998004	09/29/97	C	8,247.16
1998005	09/29/97	C	0.00
1998006	10/02/97	C	0.00
1998007	10/10/97	C	0.00
1998008	10/24/97	C	0.00
1998009	10/29/97	C	21,211.66
1998010	10/29/97	C	0.00
1998011	11/03/97	C	0.00
1998012	11/03/97	C	0.00
1998013	11/04/97	C	451.66
1998014	11/04/97	C	0.00
1998015	11/04/97	C	0.00
1998016	11/06/97	C	15,903.66
1998017	11/11/97	C	465.10

K-Means (4)

Clustering with K-Means

Locations of Scaled Centers

Center	back	contusion	head	knee	strain	unknown	laceration	leg	arm	foot	hand	ankle
1	0.00	0.17	0.18	0.37	0.19	0.00	0.00	0.21	0.00	0.00	0.00	4.65
2	0.34	0.00	2.31	0.39	0.00	0.00	0.00	0.45	0.48	0.00	0.57	0.00
3	0.22	1.91	1.00	0.00	0.13	0.00	0.07	0.44	1.48	0.63	0.00	0.00
4	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	1.02	0.29	0.00	0.00
5	0.78	0.00	0.21	0.04	0.54	0.69	0.00	0.31	0.00	0.20	0.00	0.00
6	0.04	0.49	0.33	1.52	0.24	0.00	1.51	0.27	0.00	0.20	0.00	0.00
7	0.00	0.52	0.00	0.37	0.19	0.00	1.13	0.21	0.45	0.00	4.53	0.00

Principal Components (5)

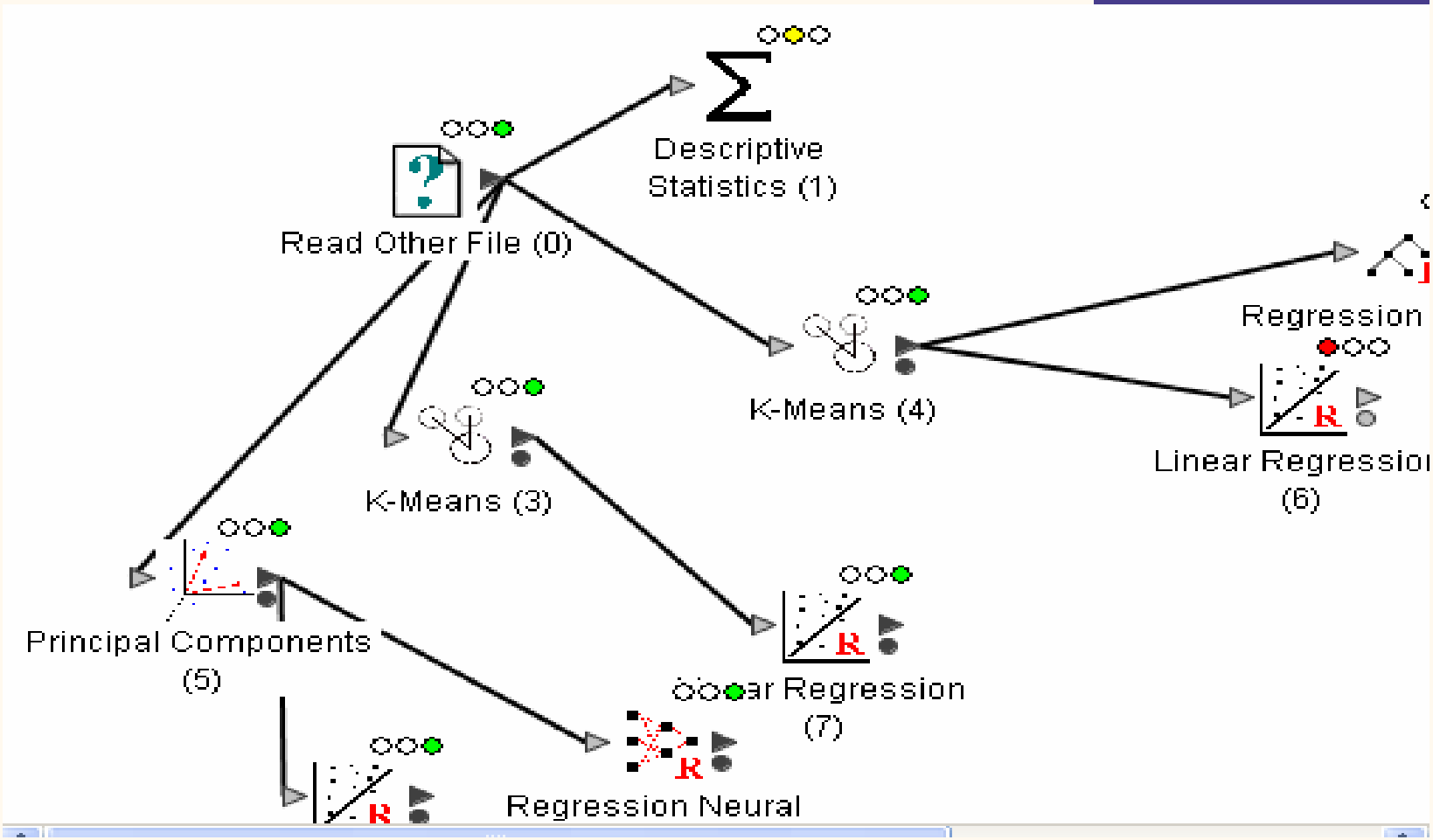
Loadings

	Variance	Cumulative %	back	contusion	head	knee	strain	unknown	laceration	leg	arm	foot
Component1	2.03	9.64	0.56	-0.18	-0.05	-0.09	0.51	-0.07	-0.13	-0.05	-0.05	-0.10
Component2	1.66	17.55	5.81E-4	0.15	0.28	-0.20	-0.10	0.17	0.35	-0.01	-0.31	0.09
Component3	1.45	24.46	-0.02	0.09	0.12	0.02	0.01	-0.25	0.18	0.19	0.27	0.02
Component4	1.37	30.99	0.11	0.53	0.18	0.39	-0.05	-0.26	-0.27	-0.23	-0.11	0.34
Component5	1.30	37.17	-0.03	-0.19	-0.22	-0.04	-0.16	0.64	-0.30	-0.17	0.05	0.08
Component6	1.26	43.19	2.02E-4	0.19	0.49	-0.42	-0.13	0.02	-0.26	0.08	0.34	-0.09
Component7	1.15	48.67	-0.01	-0.26	0.02	-0.29	-0.13	-0.17	0.29	-0.24	0.11	0.36
Component8	1.12	53.99	0.21	-0.10	-0.07	-0.18	-0.09	-0.18	0.04	0.13	0.04	0.27
Component9	1.08	59.13	0.05	-0.06	0.29	0.27	-0.10	0.02	-0.01	-0.51	-0.03	-0.48
Component10	1.07	64.22	-0.05	0.10	-0.01	-0.04	0.12	-0.01	-0.15	0.24	-0.25	0.16
Component11	1.00	69.00	-0.04	0.24	-0.20	-0.15	0.11	0.11	-0.09	-0.51	0.12	0.40
Component12	0.93	73.42	0.01	0.17	-0.18	0.08	0.05	-0.01	-0.09	0.19	0.46	-0.16

Use New Features for Prediction

- Identify serious claims
  - Use predictor variables including new features to score claims on liklihood they are serious
  - Apply more claims department resources to claims with high score
- Predict claim severity
  - Also use as a score to claims as to how serious they are likely to be

Insightful Pallet

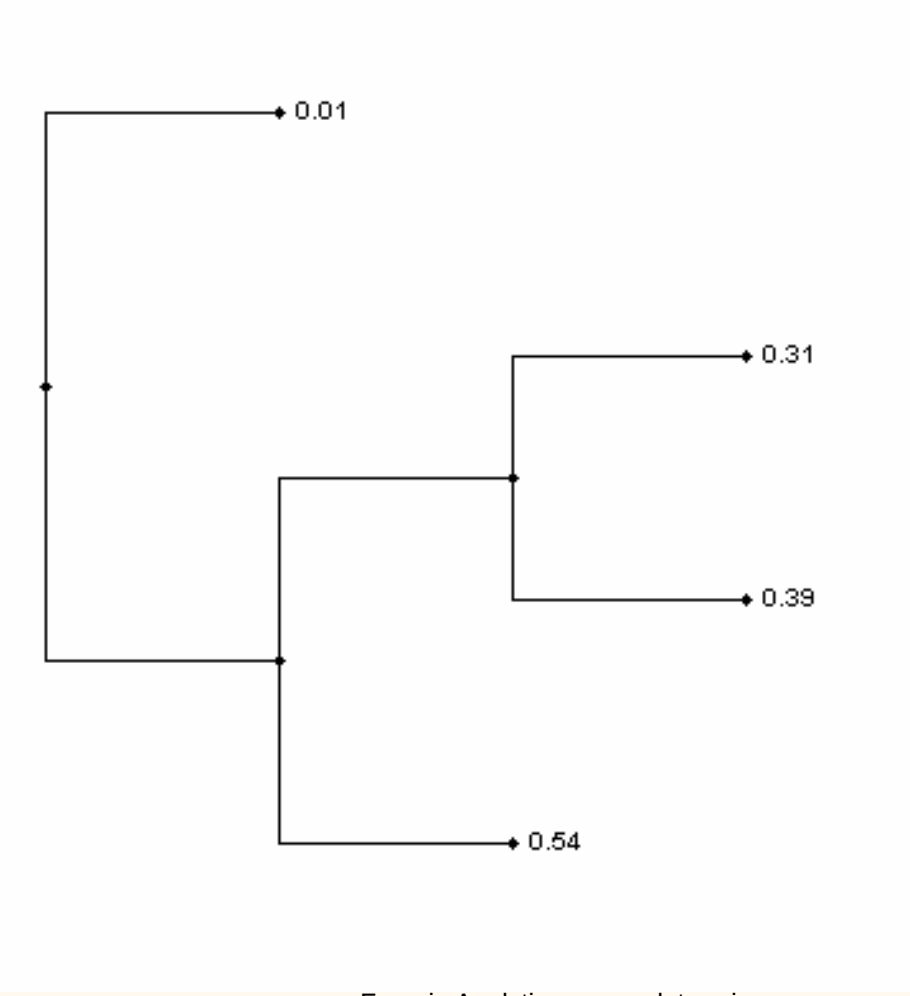


Regression of Components on Severity

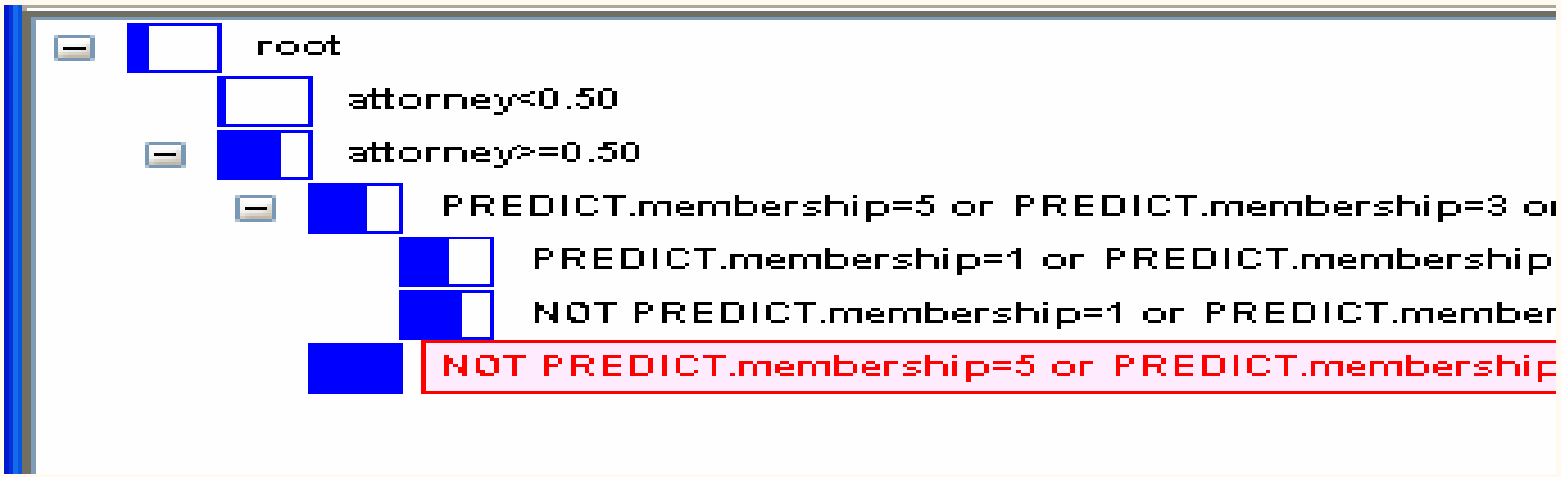
Term Importance

Source	Sum of Squares	F	Pr(F)
attorney	1,202,210,993,904.27	175.02	0.00
Component9:attorney	56,571,829,276.83	8.24	4.15E-3
Component3	3,537,350,761.59	0.51	0.47
Component5:attorney	1,439,804,339.05	0.21	0.65
Component5	85,216,267.44	0.01	0.91
Component9	18,320,032.84	2.67E-3	0.96

Tree of Serious Claims vs Injury, Attorney

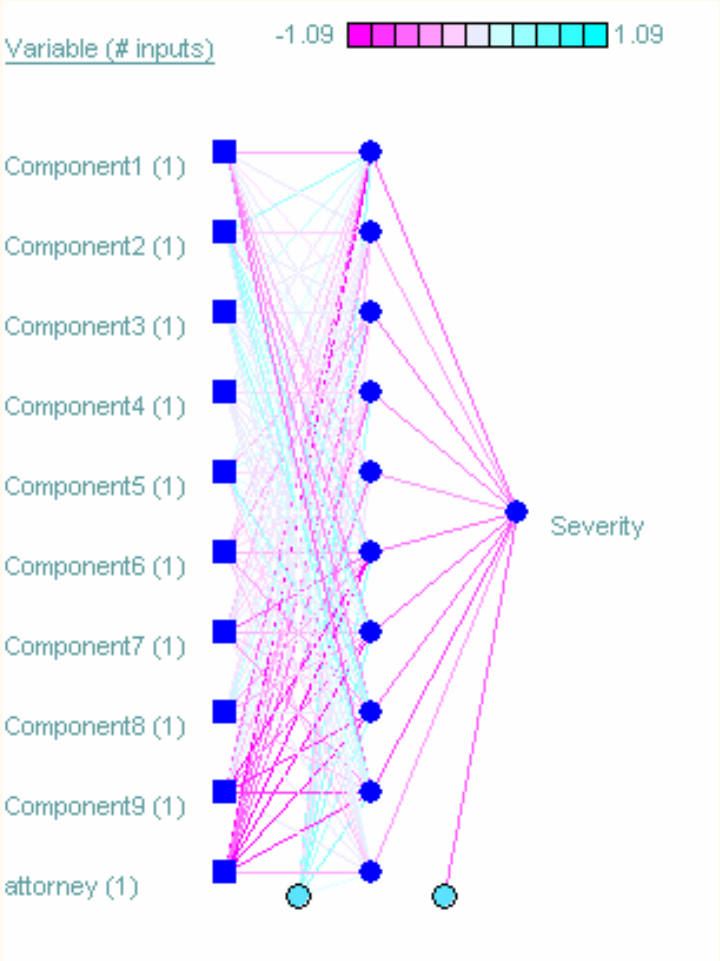


Splits on Attorney and Injury Cluster



- Injuries split on cluster2 vs all other
- Note group 2 has high % trauma claims

Neural Network for Claim Severity



Questions?