INTRODUCTION TO SAS® TEXT MINER™

THE POWER TO KNOW.

CUSTOMER LOYALTY TEAM • Support You Can Count On

TODAY’S AGENDA

INTRODUCTION TO SAS® TEXT MINER™

- Define data mining
- Overview of SAS ® Enterprise Miner™
- Describe text analytics and define text data mining
- Text Mining Process
- SAS® Text Miner™
- Illustrate text mining by example

- Q&A

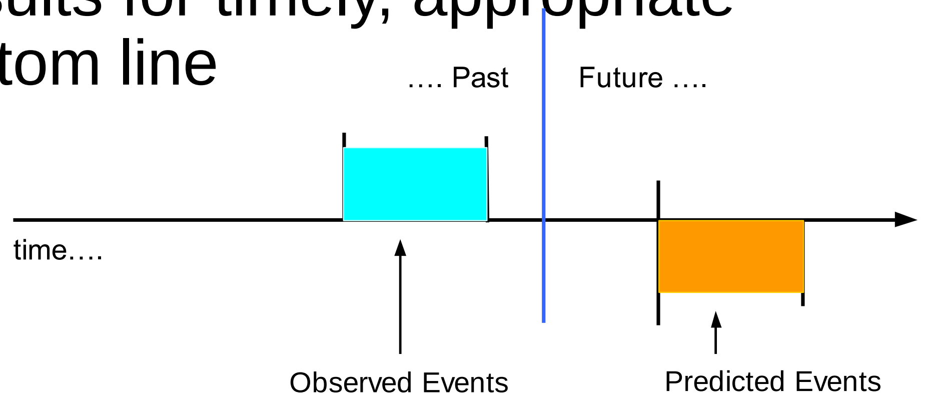
A QUICK INTRODUCTION TO DATA MINING

WHAT IS DATA MINING?

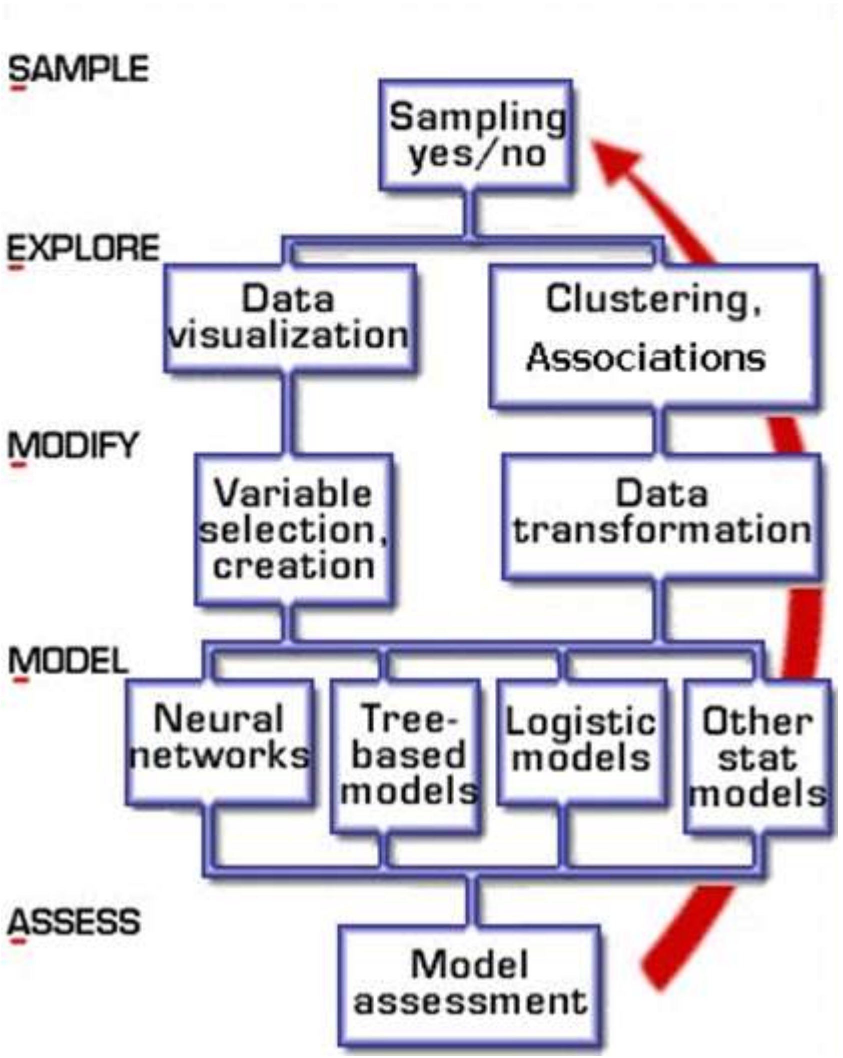
Turning increasing amounts of raw data into useful information

DATA MINING IS:

- Discovering patterns, trends and relationships represented in data
- Developing models to understand and describe characteristics and activity based on these patterns
- Using insights to help evaluate future options and take fact-based decisions
- Deploying scores and results for timely, appropriate action that affects the bottom line



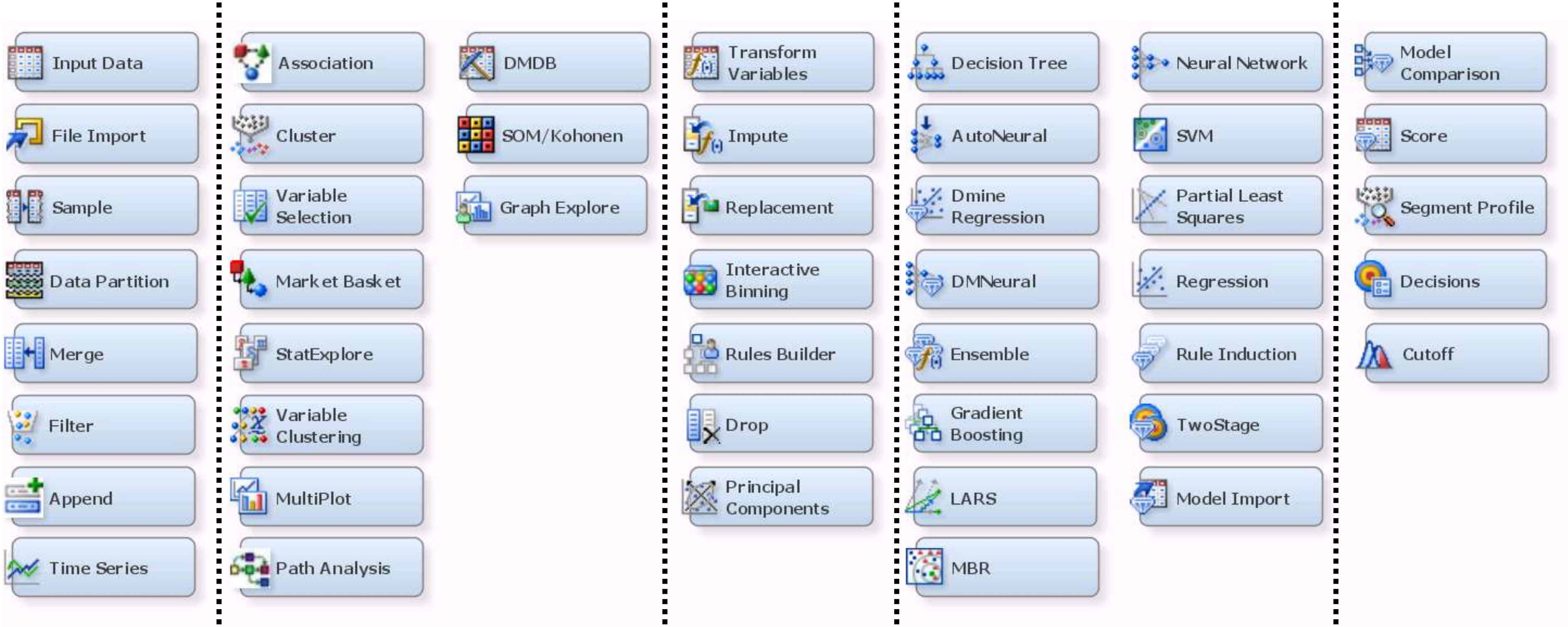
SEMMA DATA MINING WITH SAS® ENTERPRISE MINER™



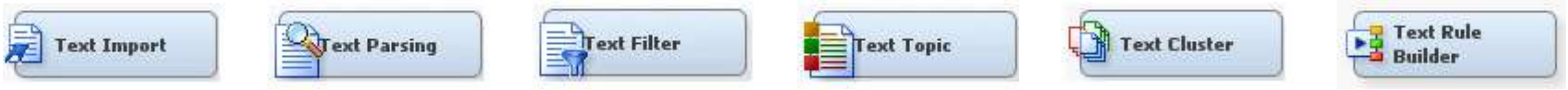
SAS® ENTERPRISE MINER™

MODEL DEVELOPMENT PROCESS

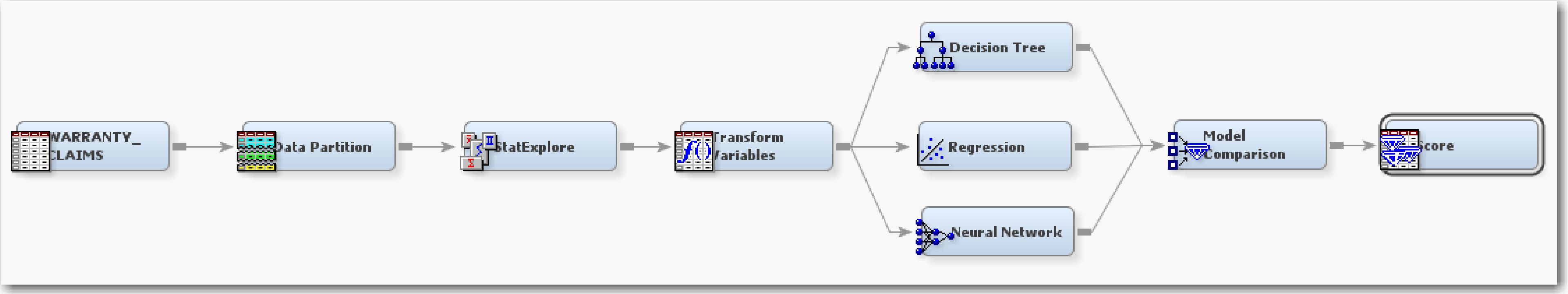
Sample Explore Modify Model Assess



Text



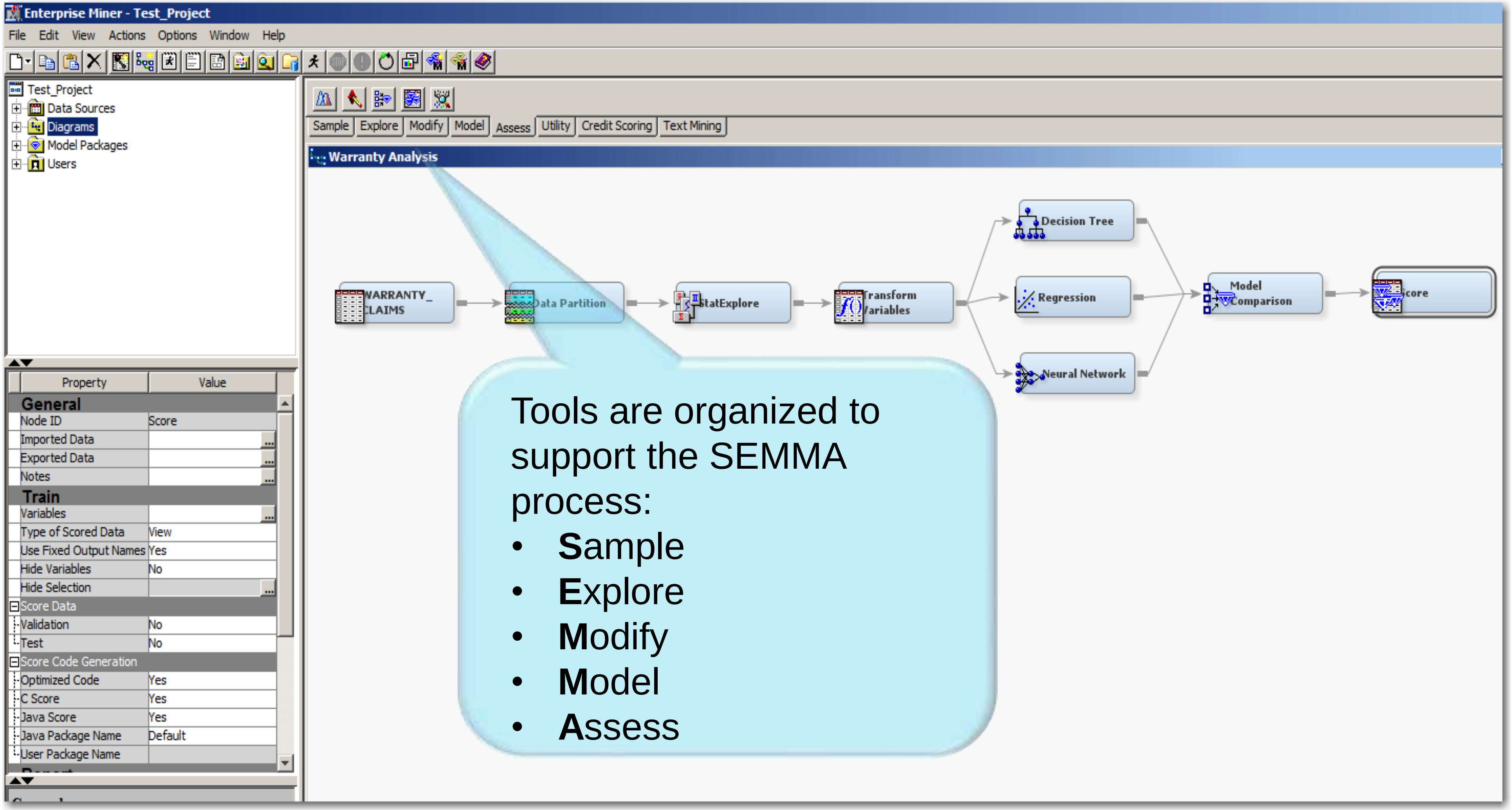
SEMMA A GLIMPSE OF SAS® ENTERPRISE MINER™



SAMPLE EXPLORE MODIFY MODEL ASSESS and SCORE

Sample Explore Modify Model Assess

SEMMA | A GLIMPSE OF SAS® ENTERPRISE MINER™

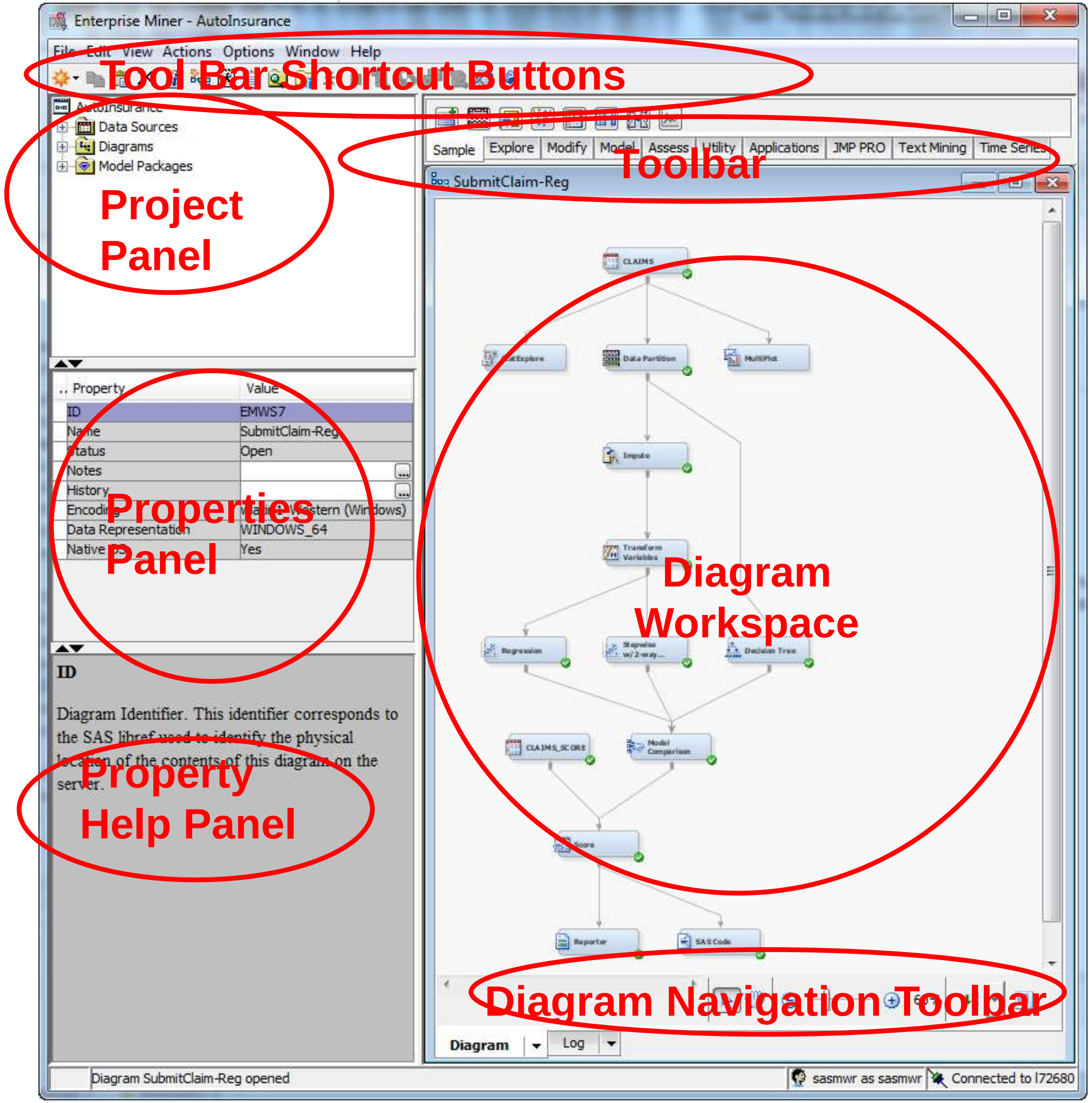


Property	Value
General	
Node ID	Score
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Type of Scored Data	View
Use Fixed Output Names	Yes
Hide Variables	No
Hide Selection	...
Score Data	
Validation	No
Test	No
Score Code Generation	
Optimized Code	Yes
C Score	Yes
Java Score	Yes
Java Package Name	Default
User Package Name	

Tools are organized to support the SEMMA process:

- **S**ample
- **E**xplore
- **M**odify
- **M**odel
- **A**ssess

SAS Enterprise Miner GUI



Property	Value
ID	EMWS7
Name	SubmitClaim-Reg
Status	Open
Notes	
History	
Encoding	Wlatin1 Western (Windows)
Data Representation	WINDOWS_64
Native OS	Yes

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram SubmitClaim-Reg opened

sasmwr as sasmwr

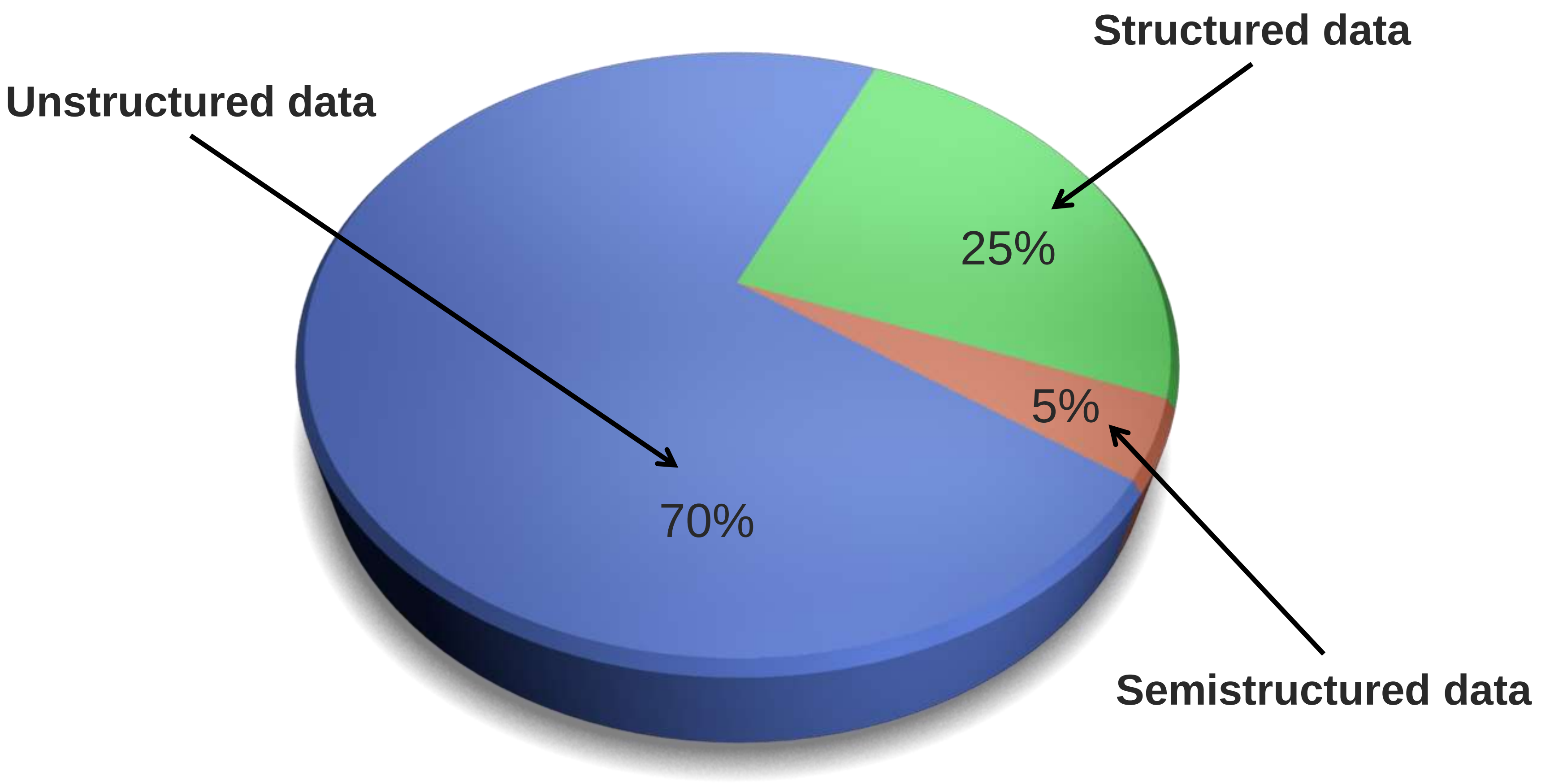
Connected to l72680

SAS® TEXT MINER™

WHY MINE TEXT?

IS VALUABLE INFORMATION “LOCKED AWAY” IN UNSTRUCTURED DATA?

UBIQUITY UNSTRUCTURED AND SEMI-STRUCTURED DATA



EXAMPLE | DATA

Structured Data

- Age Group = 60+
- Satisfaction = Not Very
- Rewards Customer= No
- Total Hold Time = 8

Unstructured Data

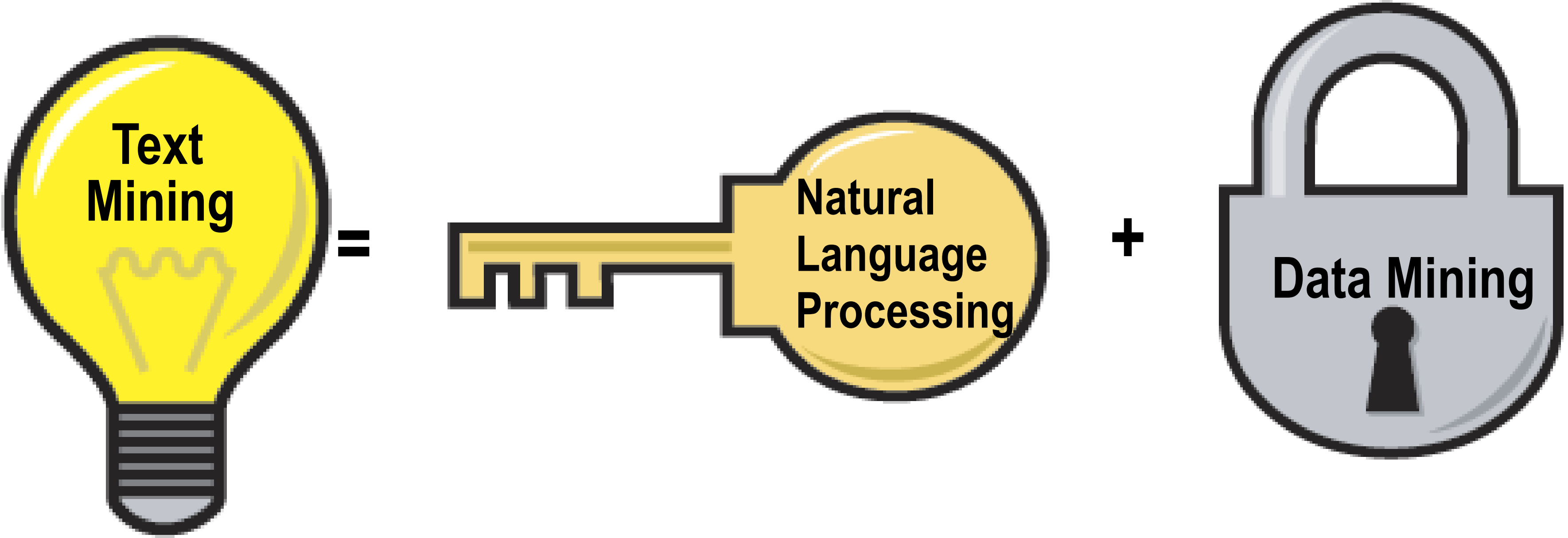
- they called me so i returned their call because it was cut off in the middle of the conversation. every time they call me, they're cut off.

WHY MINE TEXT? WHAT CAN BE LEARNED FROM UNSTRUCTURED DATA?

- Are any of these documents related to one another based on their contents and the characteristics of their contents?
- What are the key topics, themes or concepts being discussed?
- Are there emerging issues?
- Do the documents contain potentially valuable information that could improve predictive models?

TEXT MINING DEFINED

The process of discovering and extracting meaningful patterns and relationships from text collections



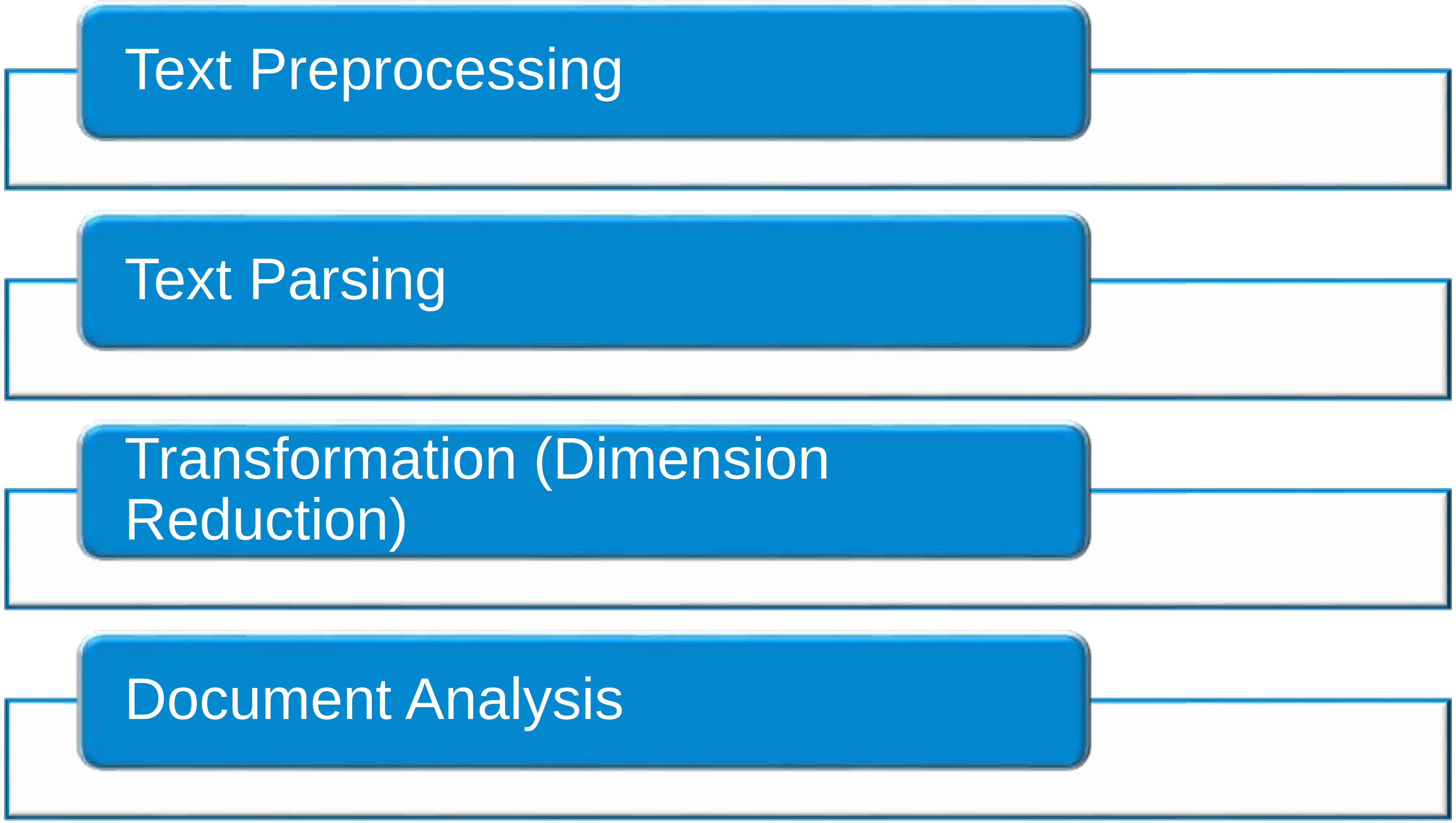
TEXT MINING TWO GENERAL GOALS

1. Pattern Discovery (Unsupervised Learning)
2. Prediction (Supervised Learning)

These are the same general goals of data mining.

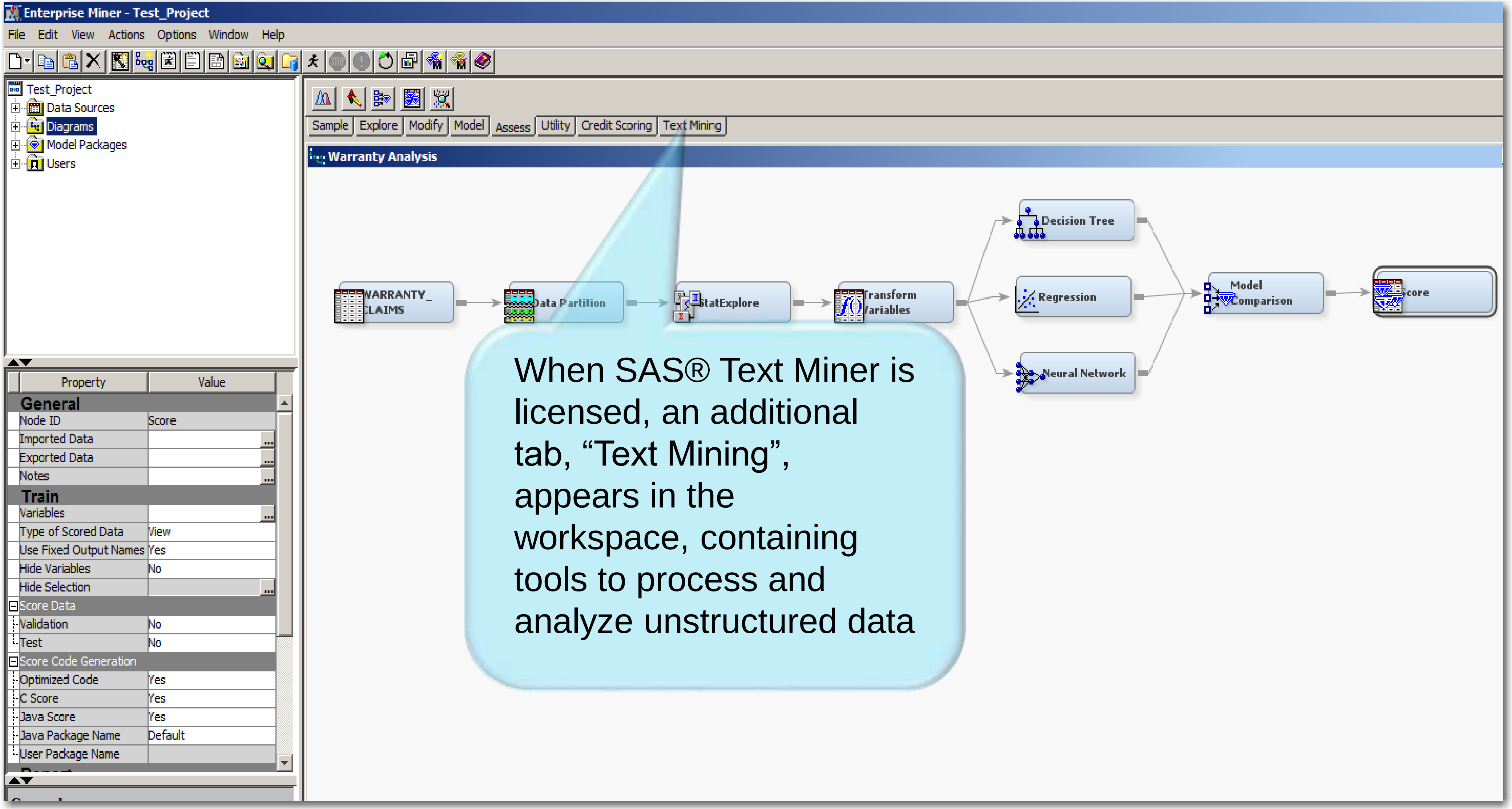
WHAT IS THE TEXT MINING PROCESS?

TEXT MINING | THE PROCESS



- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis

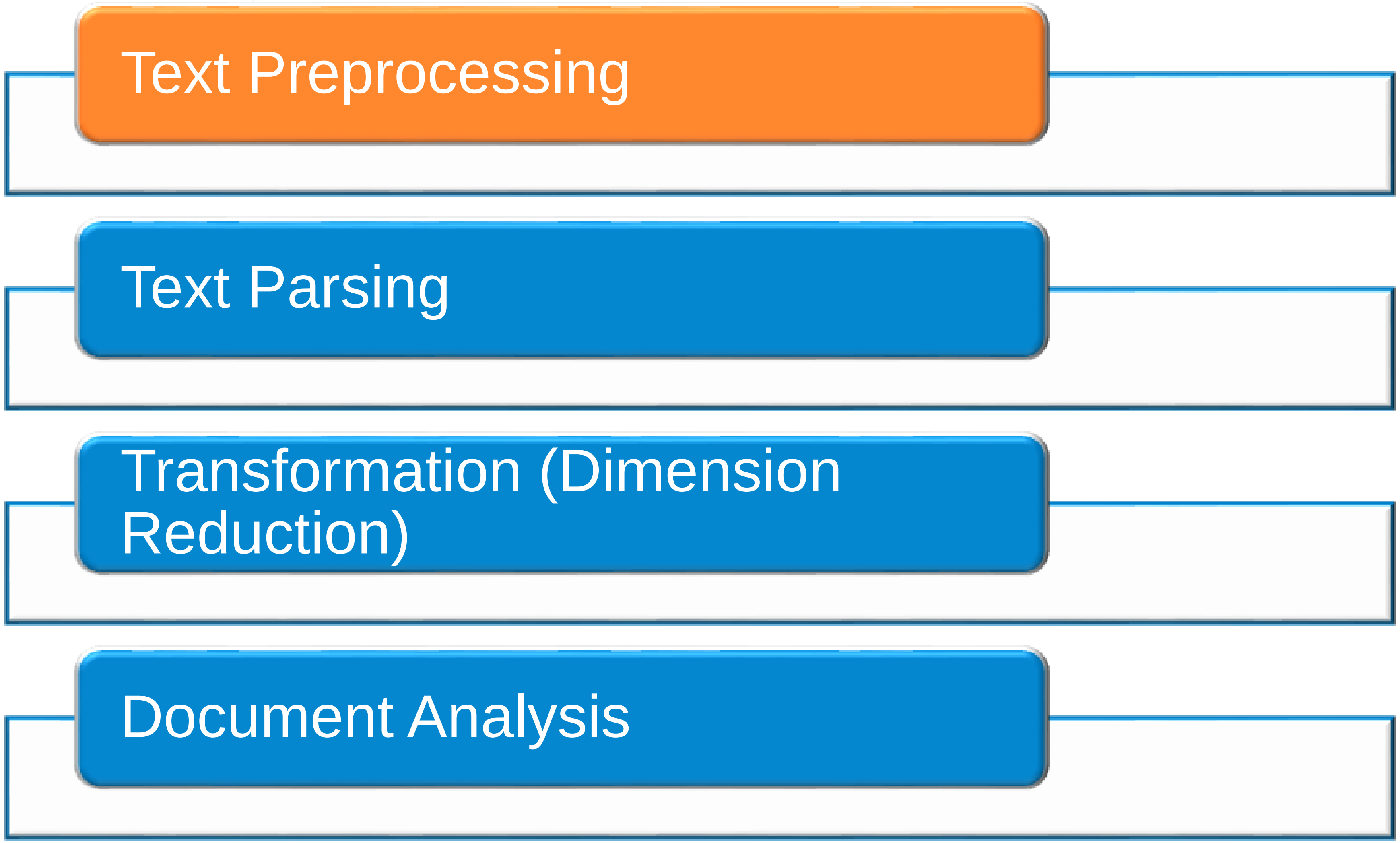
SAS® TEXT MINER ADD-ON



Property	Value
General	
Node ID	Score
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Type of Scored Data	View
Use Fixed Output Names	Yes
Hide Variables	No
Hide Selection	...
Score Data	
Validation	No
Test	No
Score Code Generation	
Optimized Code	Yes
C Score	Yes
Java Score	Yes
Java Package Name	Default
User Package Name	

When SAS® Text Miner is licensed, an additional tab, “Text Mining”, appears in the workspace, containing tools to process and analyze unstructured data

TEXT MINING | THE PROCESS



Text Preprocessing

Text Parsing

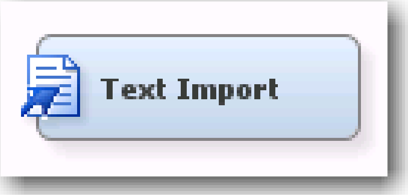
Transformation (Dimension Reduction)

Document Analysis

INPUT DATA | TEXT MINER

- The expected SAS data set for text mining should have the following characteristics:
  - One row per document
  - A document ID (suggested)
  - A “text” column
- The “text” column can be either:
  - The actual full text of the document, up to 32,000 characters
  - A pointer to a text file (*.txt, *.html) located on the file system
- The SAS data set can also have structured data and a target variable (dependent variable, response variable)

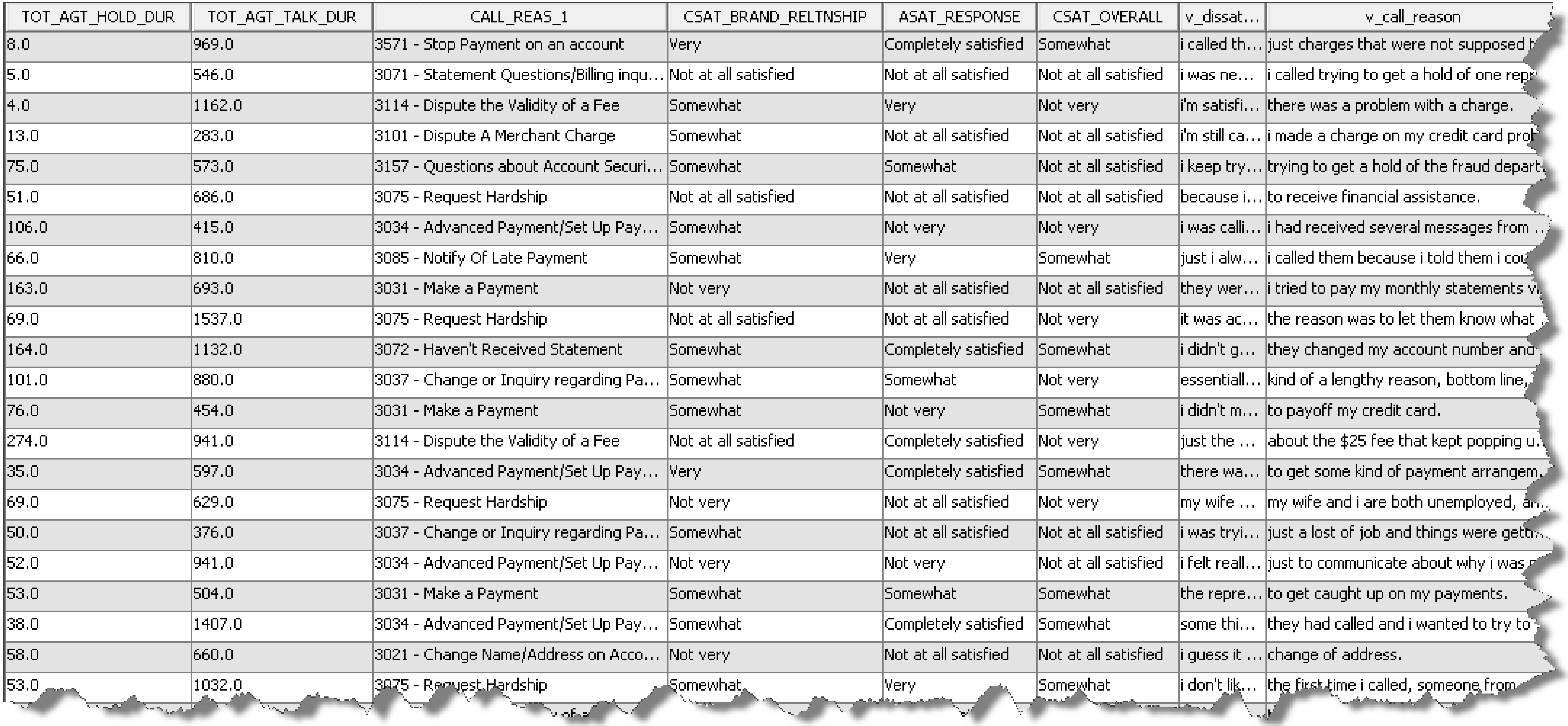
TEXT IMPORT NODE



Text Import

- Enables you to create data sets dynamically from files contained in a directory or from the Web.
- Takes an import directory containing text files in potentially proprietary formats such as MS Word and PDF files as input.
- Extracts the text from the files, places a copy of the text in a plain text file, and a snippet (or possibly even all) of the text in a SAS data set.
- If a URL is specified, the node will crawl Web sites and retrieve files from the Web
- The output of a **Text Import** node is a data set that can be imported into the **Text Parsing** node.

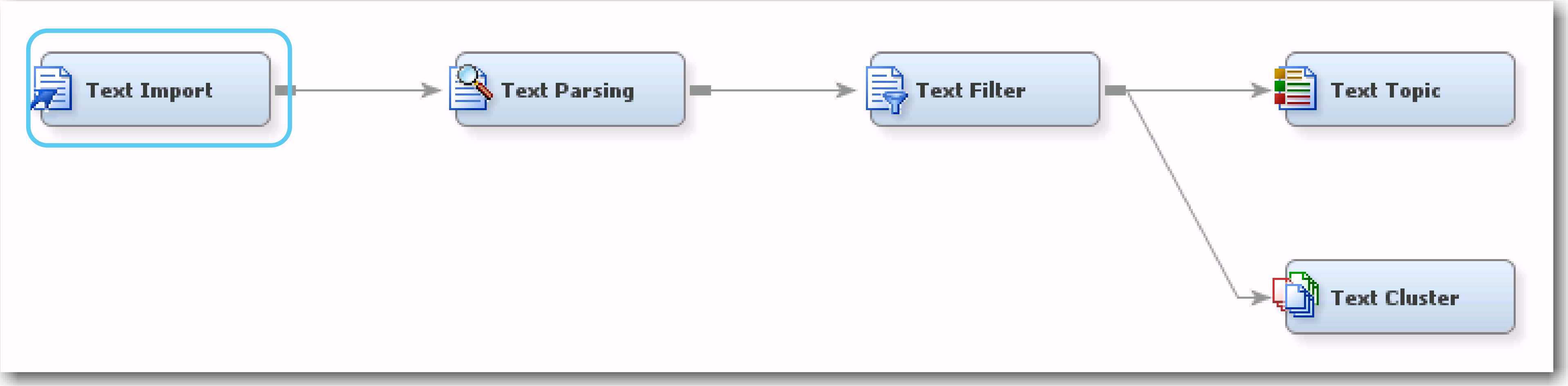
EXAMPLE INPUT DATA



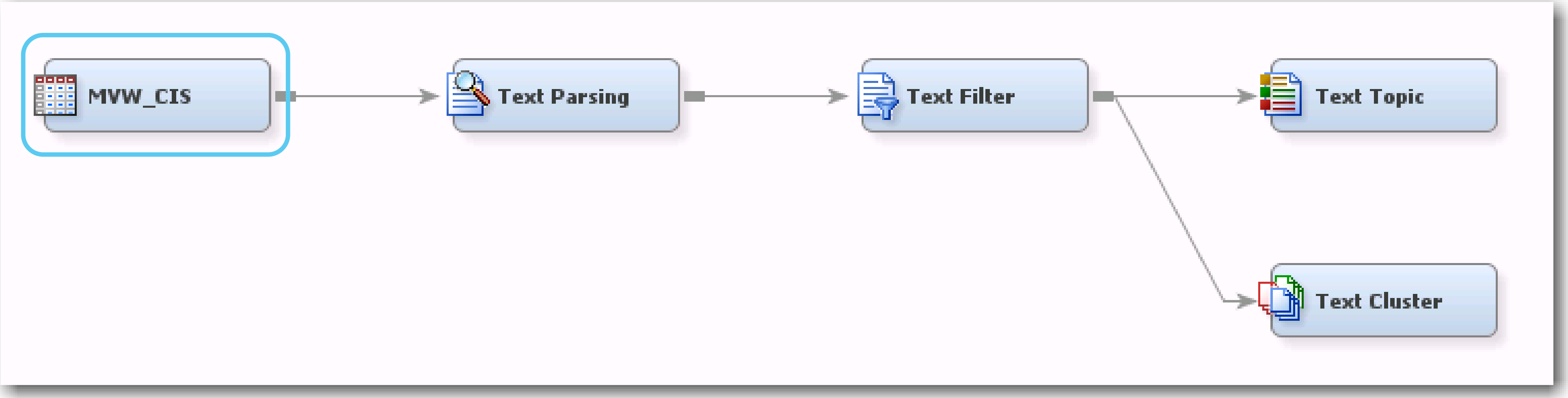
TOT_AGT_HOLD_DUR	TOT_AGT_TALK_DUR	CALL_REAS_1	CSAT_BRAND_RELTNSHIP	ASAT_RESPONSE	CSAT_OVERALL	v_dissat…	v_call_reason
8.0	969.0	3571 - Stop Payment on an account	Very	Completely satisfied	Somewhat	i called th…	just charges that were not supposed
5.0	546.0	3071 - Statement Questions/Billing inqu…	Not at all satisfied	Not at all satisfied	Not at all satisfied	i was ne…	i called trying to get a hold of one rep
4.0	1162.0	3114 - Dispute the Validity of a Fee	Somewhat	Very	Not very	i'm satisfi…	there was a problem with a charge.
13.0	283.0	3101 - Dispute A Merchant Charge	Somewhat	Not at all satisfied	Not at all satisfied	i'm still ca…	i made a charge on my credit card pro
75.0	573.0	3157 - Questions about Account Securi…	Somewhat	Somewhat	Not at all satisfied	i keep try…	trying to get a hold of the fraud depart
51.0	686.0	3075 - Request Hardship	Not at all satisfied	Not at all satisfied	Not at all satisfied	because i…	to receive financial assistance.
106.0	415.0	3034 - Advanced Payment/Set Up Pay…	Somewhat	Not very	Not very	i was calli…	i had received several messages from …
66.0	810.0	3085 - Notify Of Late Payment	Somewhat	Very	Somewhat	just i alw…	i called them because i told them i cou
163.0	693.0	3031 - Make a Payment	Not very	Not at all satisfied	Not at all satisfied	they wer…	i tried to pay my monthly statements v
69.0	1537.0	3075 - Request Hardship	Not at all satisfied	Not at all satisfied	Not very	it was ac…	the reason was to let them know what
164.0	1132.0	3072 - Haven't Received Statement	Somewhat	Completely satisfied	Somewhat	i didn't g…	they changed my account number and
101.0	880.0	3037 - Change or Inquiry regarding Pa…	Somewhat	Somewhat	Not very	essentiall…	kind of a lengthy reason, bottom line,
76.0	454.0	3031 - Make a Payment	Somewhat	Not very	Somewhat	i didn't m…	to payoff my credit card.
274.0	941.0	3114 - Dispute the Validity of a Fee	Not at all satisfied	Completely satisfied	Not very	just the …	about the $25 fee that kept popping u…
35.0	597.0	3034 - Advanced Payment/Set Up Pay…	Very	Completely satisfied	Somewhat	there wa…	to get some kind of payment arrangem
69.0	629.0	3075 - Request Hardship	Not very	Not at all satisfied	Not very	my wife …	my wife and i are both unemployed, a
50.0	376.0	3037 - Change or Inquiry regarding Pa…	Somewhat	Not at all satisfied	Not at all satisfied	i was tryi…	just a lost of job and things were gett
52.0	941.0	3034 - Advanced Payment/Set Up Pay…	Not very	Not very	Not at all satisfied	i felt reall…	just to communicate about why i was
53.0	504.0	3031 - Make a Payment	Somewhat	Somewhat	Somewhat	the repre…	to get caught up on my payments.
38.0	1407.0	3034 - Advanced Payment/Set Up Pay…	Somewhat	Completely satisfied	Somewhat	some thi…	they had called and i wanted to try to
58.0	660.0	3021 - Change Name/Address on Acco…	Not very	Not at all satisfied	Not at all satisfied	i guess it …	change of address.
53.0	1032.0	3075 - Request Hardship	Somewhat	Very	Somewhat	i don't lik…	the first time i called, someone from

EXAMPLE

TEXT MINING PROCESS FLOWS

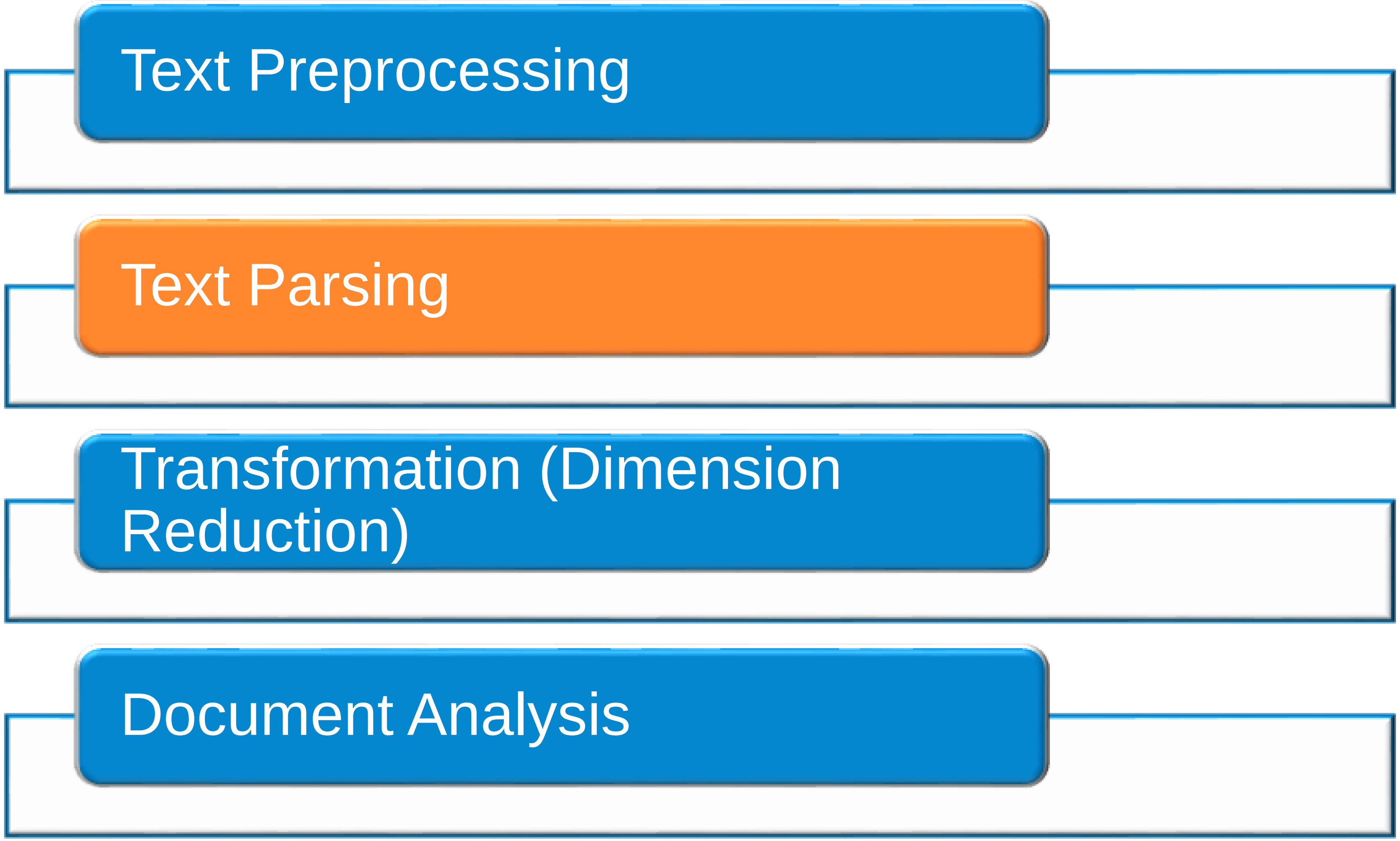


External Documents



Text in Column or Document location in column

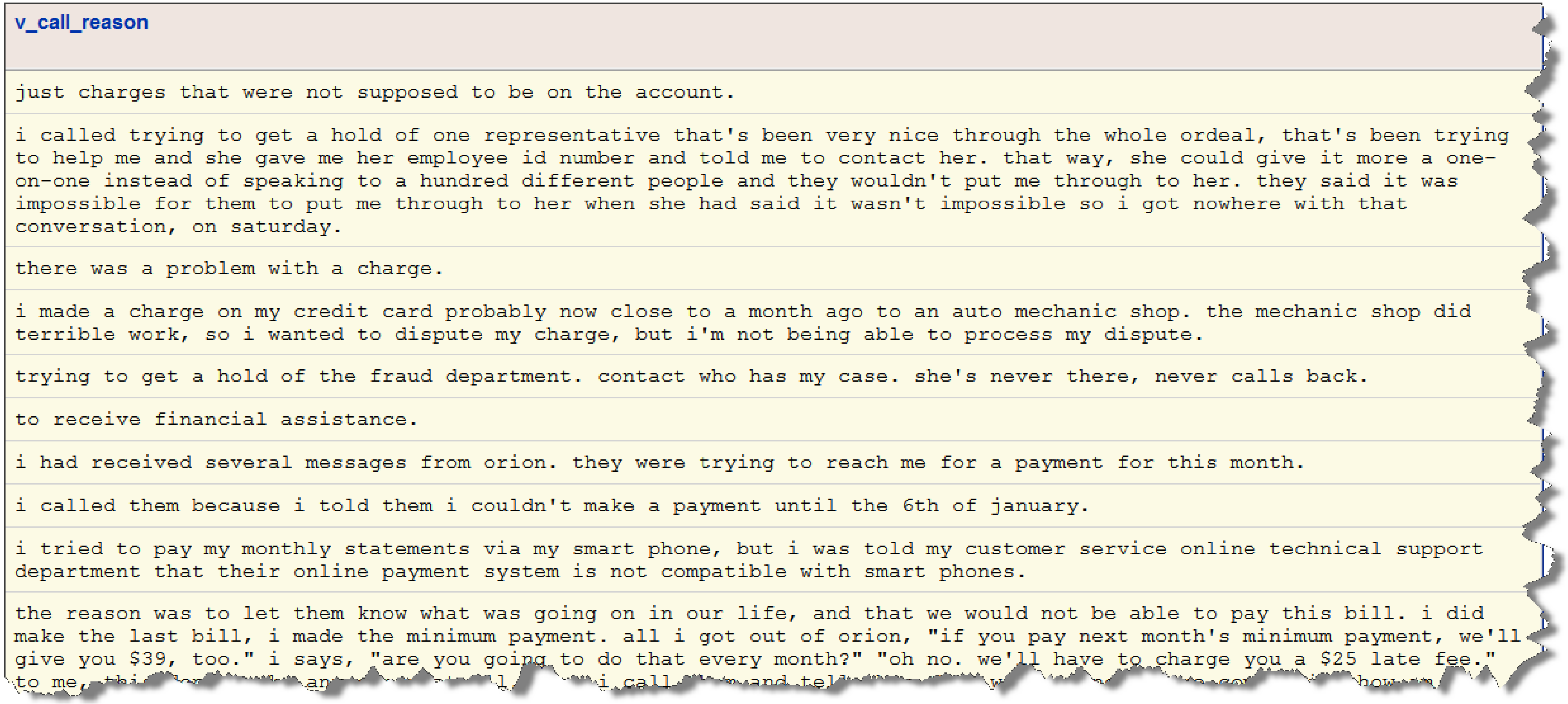
TEXT MINING PROCESS



- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis

TEXT PARSING

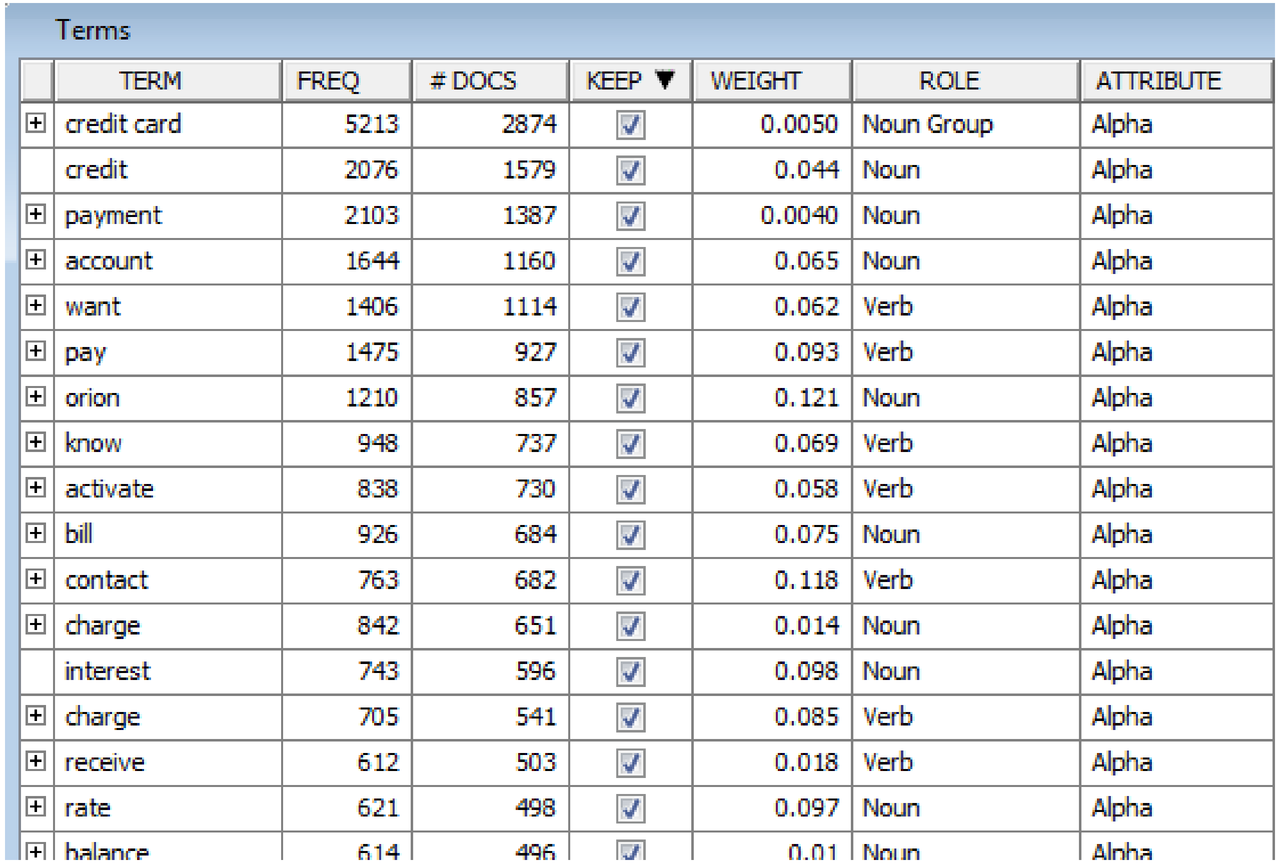
- Text parsing decomposes textual data and generates a quantitative representation suitable for data mining purposes.
- It transforms this:



v_call_reason
just charges that were not supposed to be on the account.
i called trying to get a hold of one representative that's been very nice through the whole ordeal, that's been trying to help me and she gave me her employee id number and told me to contact her. that way, she could give it more a one-on-one instead of speaking to a hundred different people and they wouldn't put me through to her. they said it was impossible for them to put me through to her when she had said it wasn't impossible so i got nowhere with that conversation, on saturday.
there was a problem with a charge.
i made a charge on my credit card probably now close to a month ago to an auto mechanic shop. the mechanic shop did terrible work, so i wanted to dispute my charge, but i'm not being able to process my dispute.
trying to get a hold of the fraud department. contact who has my case. she's never there, never calls back.
to receive financial assistance.
i had received several messages from orion. they were trying to reach me for a payment for this month.
i called them because i told them i couldn't make a payment until the 6th of january.
i tried to pay my monthly statements via my smart phone, but i was told my customer service online technical support department that their online payment system is not compatible with smart phones.
the reason was to let them know what was going on in our life, and that we would not be able to pay this bill. i did make the last bill, i made the minimum payment. all i got out of orion, "if you pay next month's minimum payment, we'll give you $39, too." i says, "are you going to do that every month?" "oh no. we'll have to charge you a $25 late fee." to me, thi[illegible] an[illegible] i call [illegible] and tel[illegible]

TEXT PARSING

- Text parsing decomposes textual data and generates a quantitative representation suitable for data mining purposes.
- … into this:

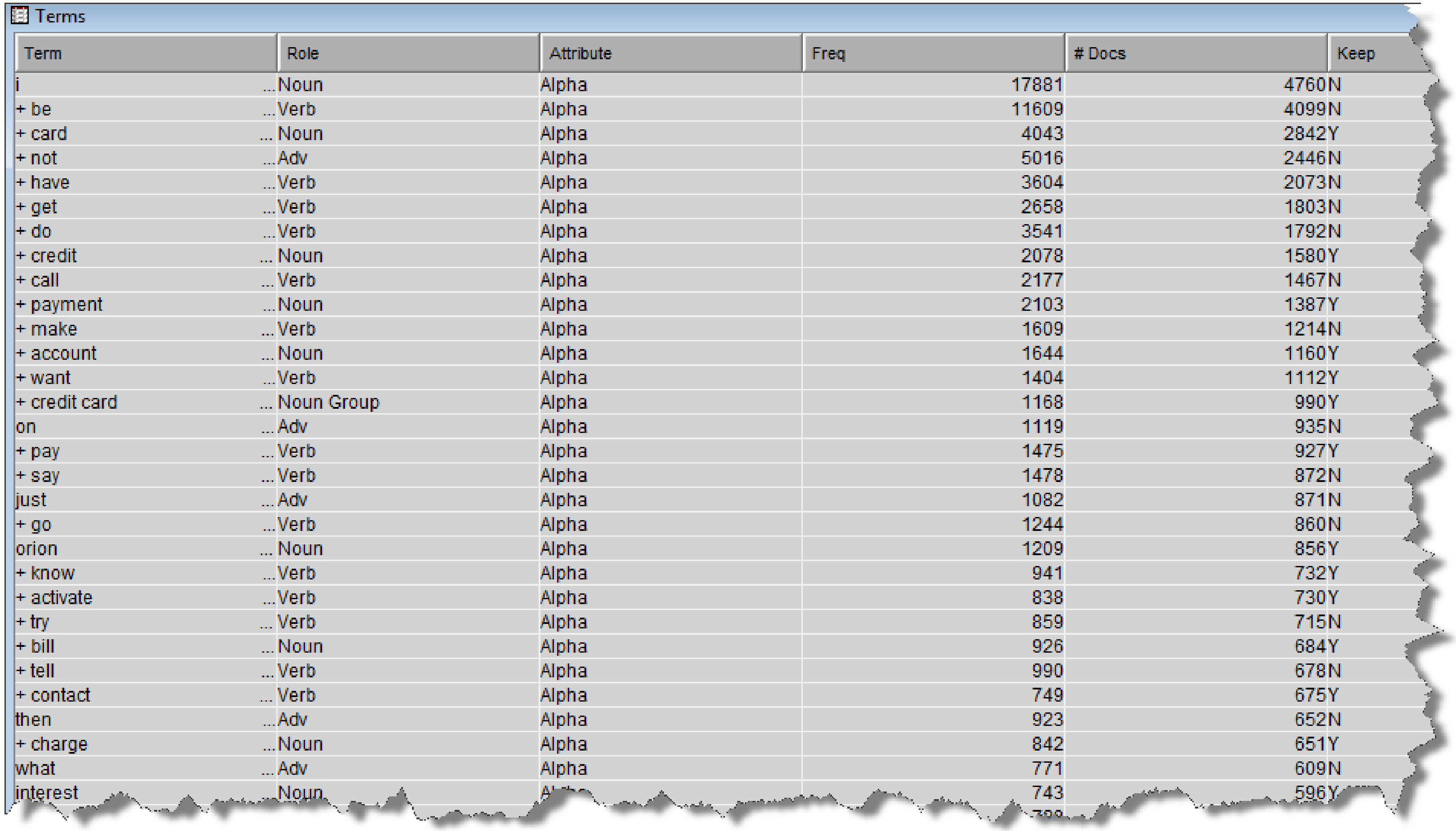


Terms

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
⊞	credit card	5213	2874	☑	0.0050	Noun Group	Alpha
	credit	2076	1579	☑	0.044	Noun	Alpha
⊞	payment	2103	1387	☑	0.0040	Noun	Alpha
⊞	account	1644	1160	☑	0.065	Noun	Alpha
⊞	want	1406	1114	☑	0.062	Verb	Alpha
⊞	pay	1475	927	☑	0.093	Verb	Alpha
⊞	orion	1210	857	☑	0.121	Noun	Alpha
⊞	know	948	737	☑	0.069	Verb	Alpha
⊞	activate	838	730	☑	0.058	Verb	Alpha
⊞	bill	926	684	☑	0.075	Noun	Alpha
⊞	contact	763	682	☑	0.118	Verb	Alpha
⊞	charge	842	651	☑	0.014	Noun	Alpha
	interest	743	596	☑	0.098	Noun	Alpha
⊞	charge	705	541	☑	0.085	Verb	Alpha
⊞	receive	612	503	☑	0.018	Verb	Alpha
⊞	rate	621	498	☑	0.097	Noun	Alpha
⊞	balance	614	496	☑	0.01	Noun	Alpha

TEXT PARSING

- Documents are represented internally in SAS® Text Miner by a vector that contains the frequency of how many times each term occurs in each document.

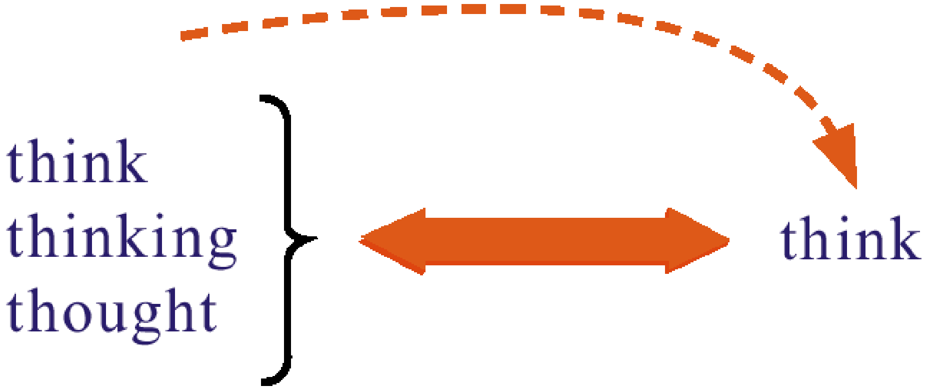


Terms

Term	Role	Attribute	Freq	# Docs	Keep
i ...	Noun	Alpha	17881	4760	N
+ be ...	Verb	Alpha	11609	4099	N
+ card ...	Noun	Alpha	4043	2842	Y
+ not ...	Adv	Alpha	5016	2446	N
+ have ...	Verb	Alpha	3604	2073	N
+ get ...	Verb	Alpha	2658	1803	N
+ do ...	Verb	Alpha	3541	1792	N
+ credit ...	Noun	Alpha	2078	1580	Y
+ call ...	Verb	Alpha	2177	1467	N
+ payment ...	Noun	Alpha	2103	1387	Y
+ make ...	Verb	Alpha	1609	1214	N
+ account ...	Noun	Alpha	1644	1160	Y
+ want ...	Verb	Alpha	1404	1112	Y
+ credit card ...	Noun Group	Alpha	1168	990	Y
on ...	Adv	Alpha	1119	935	N
+ pay ...	Verb	Alpha	1475	927	Y
+ say ...	Verb	Alpha	1478	872	N
just ...	Adv	Alpha	1082	871	N
+ go ...	Verb	Alpha	1244	860	N
orion ...	Noun	Alpha	1209	856	Y
+ know ...	Verb	Alpha	941	732	Y
+ activate ...	Verb	Alpha	838	730	Y
+ try ...	Verb	Alpha	859	715	N
+ bill ...	Noun	Alpha	926	684	Y
+ tell ...	Verb	Alpha	990	678	N
+ contact ...	Verb	Alpha	749	675	Y
then ...	Adv	Alpha	923	652	N
+ charge ...	Noun	Alpha	842	651	Y
what ...	Adv	Alpha	771	609	N
interest ...	Noun	Alpha	743	596	Y

TEXT PARSING

STEMMING

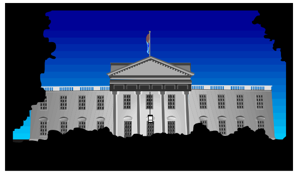


PART OF SPEECH

- Determines if the word is a common noun, verb, adjective, proper noun, adverb, etc.
- Disambiguate parts of speech when a word is used in a different context,
  - *I wish that my bank did not have a service charge for using other vendor ATM's.*
  - *You can bank on either Germany or England winning the world cup next year.*

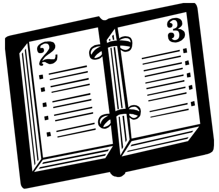
ENTITY EXTRACTION

Places
White House



People's Names
James H. Goodnight

Dates



PARTS OF SPEECH IN SAS® TEXT MINER

- Abbr (abbreviation)
- Adj (adjective)
- Adv (adverb)
- Aux (auxuliary or modal)
- Conj (conjunction)
- Det (determiner)
- Interj (interjection)
- Noun (noun)
- Num (number or numeric expression)
- Part (infinitive marker, negative participle, or possessive marker)
- Pref (prefix)
- Prep (preposition)
- Pron (pronoun)
- Prop (proper noun)
- Punct (punctuation)
- Verb (verb)
- VerbAdj (verb adjective)

ENTITIES IN SAS® TEXT MINER

STANDARD ENTITIES (IDENTIFIED OUT-OF-THE-BOX)

- Address
- Company
- Currency
- Date
- Internet
- Location
- Measure
- Organization
- Percent

- Person
- Phone
- Prop_Misc (proper noun – ambiguous classification)
- SSN (U. S. Social Security Number)
- Time
- Time_Period
- Title
- Vehicle (motor vehicle)

ADDITIONAL PARSING STEPS

- Specify Start/Stop/Synonym Lists
  - Filtering out low information words such as
    - articles (e.g. the, a, this)
    - prepositions (e.g. of, from, by)
    - conjunctions (e.g. and, but, or)
  - Consider document subject matter as well as domain-specific language and acronymns

- Vertical dictionaries
  - Automatically generate synonyms appropriate to the data

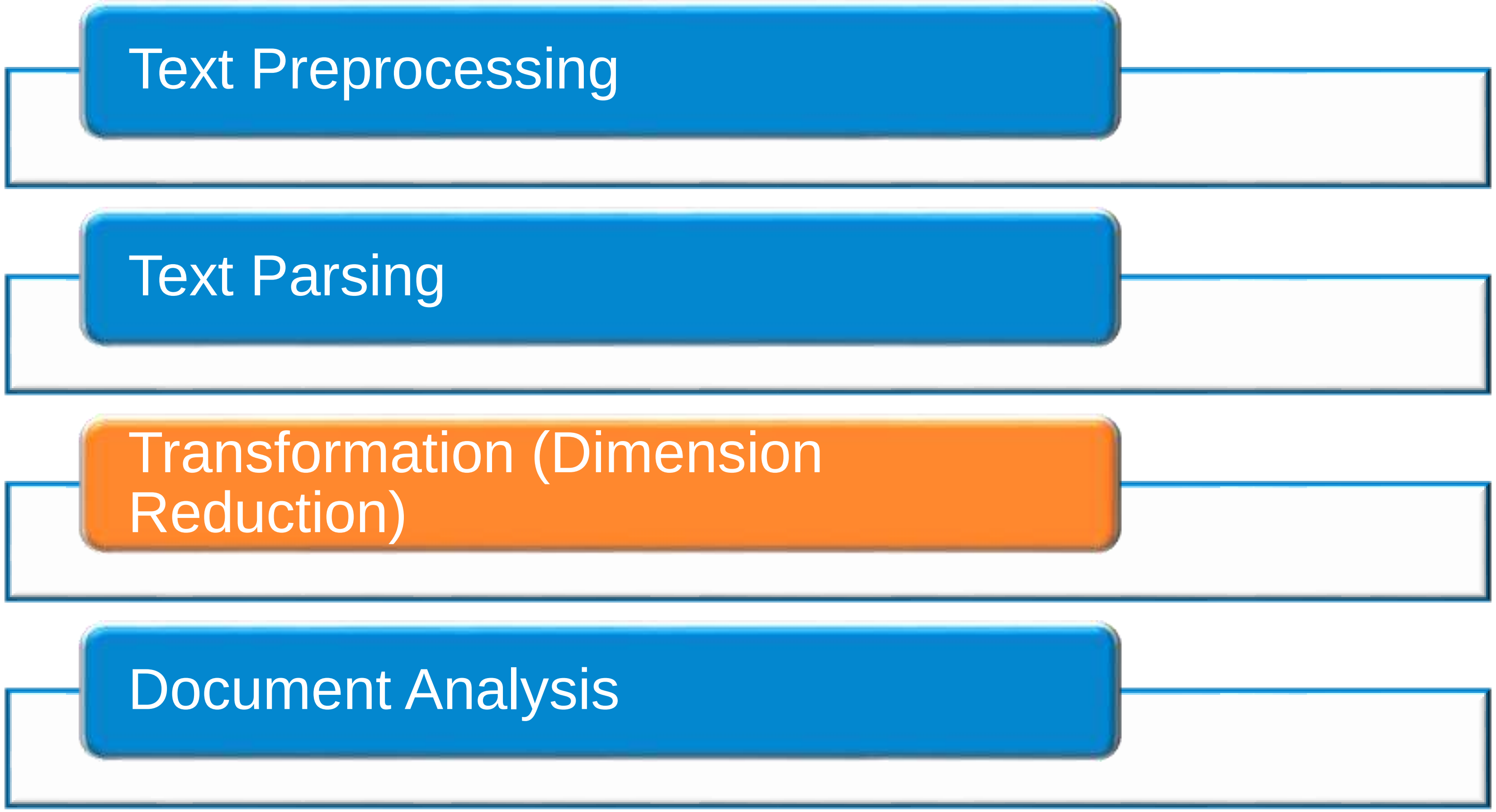
ADDITIONAL DATA PREPARATION

- Remove "*boilerplate*" language common to most or all documents
  - Headers and footers
  - Common qualifiers
  - Disclaimers
- Parse created data
  - Convert abbreviations
  - Correct misspellings
- Use term frequency filtering to assist with the creation of a stop list

ADDITIONAL DATA PREPARATION

- Recommendation: create subsets of documents by language. For example, all English documents in one corpus, all German documents in another corpus, etc.
- SAS includes extremely robust and sophisticated data manipulation capabilities, including character functions and regular expressions.

TEXT MINING PROCESS



Text Preprocessing

Text Parsing

Transformation (Dimension Reduction)

Document Analysis

TEXT TRANSFORMATION

- Also referred to as “Dimension Reduction”
- Transforms the quantitative representation into a compact and informative format
- Can also be used to further refine the data to be analyzed. For example, you can reduce the total number of parsed terms or documents that are analyzed.
- Eliminates extraneous information so that only the most valuable information or information that relates to a particular area of interest is considered.

TEXT FILTER NODE

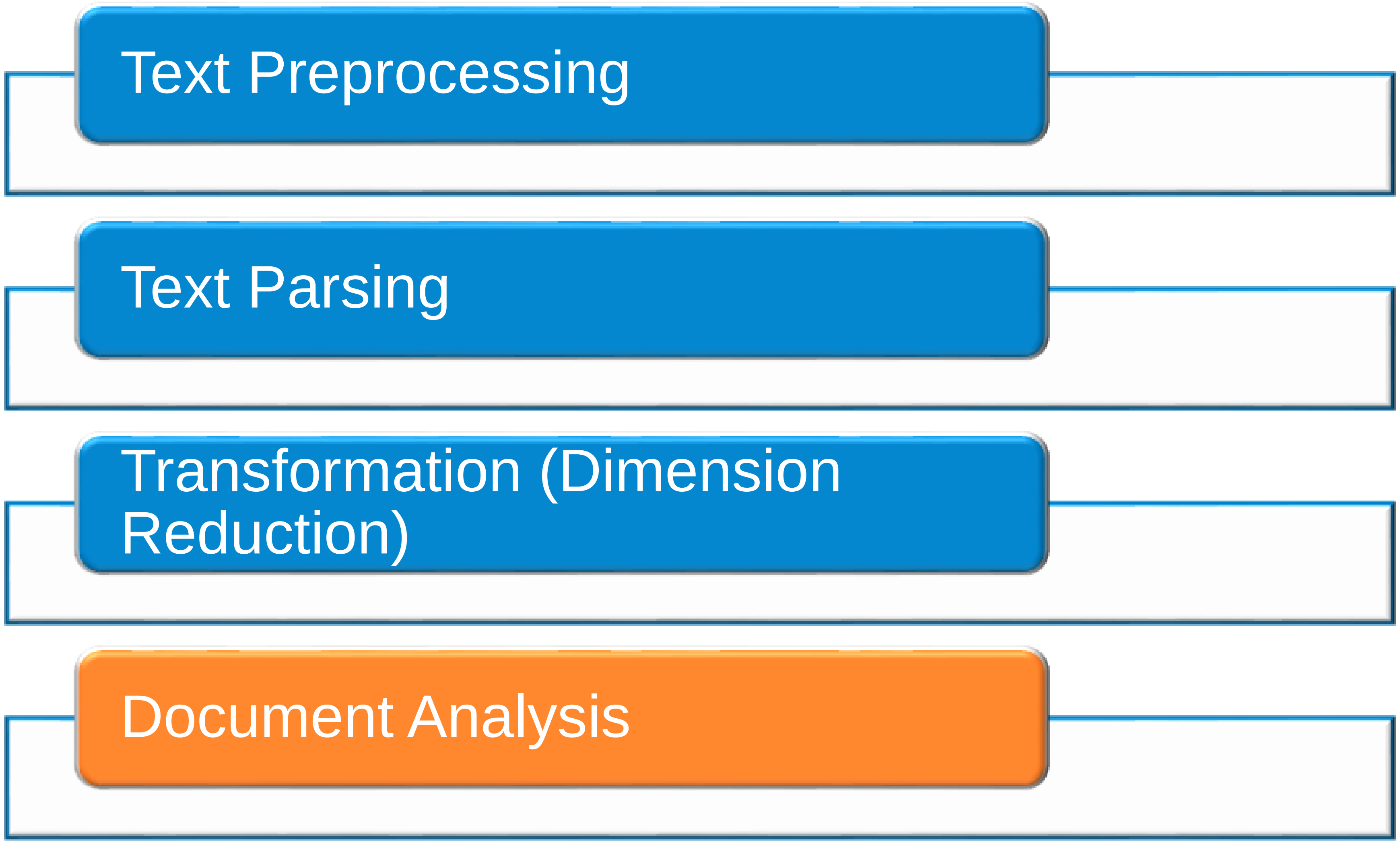
Text Filter

- Spell checking
- Concept Linking
- Full text search
- Define additional synonyms
- Sub-setting management of terms and documents that are passed to subsequent nodes

DIMENSION REDUCTION TECHNIQUES

- Singular value decomposition (SVD)
- Roll up terms
- Combination of both approaches

TEXT MINING PROCESS



Text Preprocessing

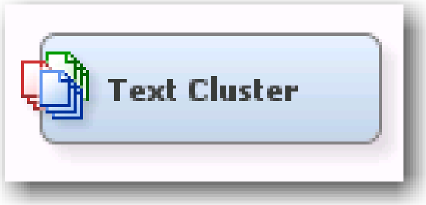
Text Parsing

Transformation (Dimension Reduction)

Document Analysis

DOCUMENT ANALYSIS

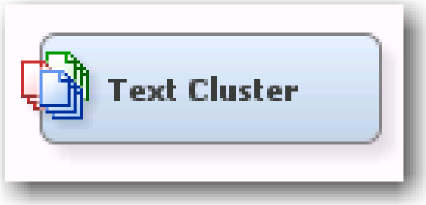
CLUSTERING



- Expectation Maximization Clustering
  - Generates groups of similar documents from output of SVD
  - Fast clustering of many documents
- Hierarchical Clustering
  - Great for creating document taxonomies

DOCUMENT ANALYSIS

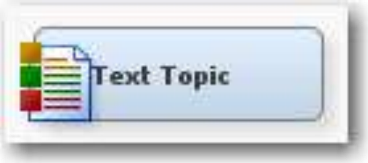
CLUSTERING



- Note: each document is assigned to a single cluster
- Optionally, use unsupervised data mining methods like self organizing maps or clustering after building text mining clusters, using the text mining cluster segment identifiers as inputs in the subsequent analysis

DOCUMENT ANALYSIS

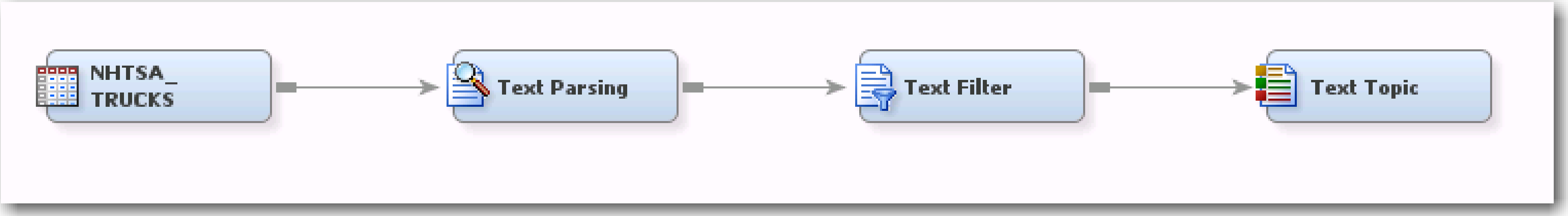
TOPICS



- Discovers topics in document collection
- Allows automatic creation of single and multi-word topics
- User defined topics and editing of automatic topics
- Multiple topics per document
  - Soft clustering using rotated SVD (PROC SVD followed by PROC FACTOR)

SAS® TEXT MINER PROCESS

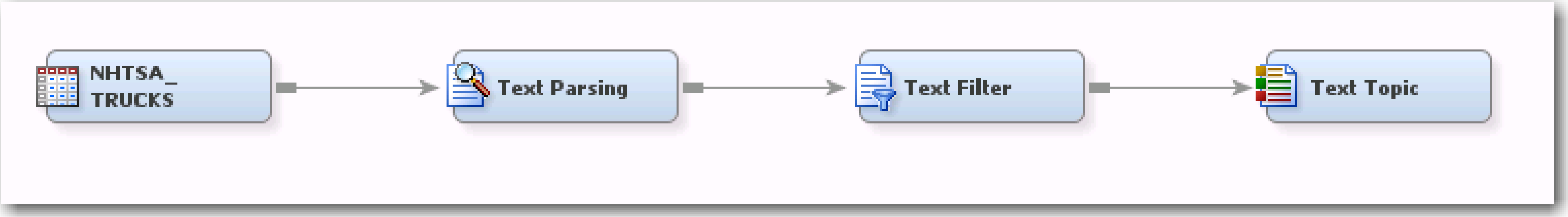
EXAMPLE TEXT MINING PROCESS FLOWS



Start with a table that contains either:

- Documents saved as a variable (column)
- A column that points to physical text files

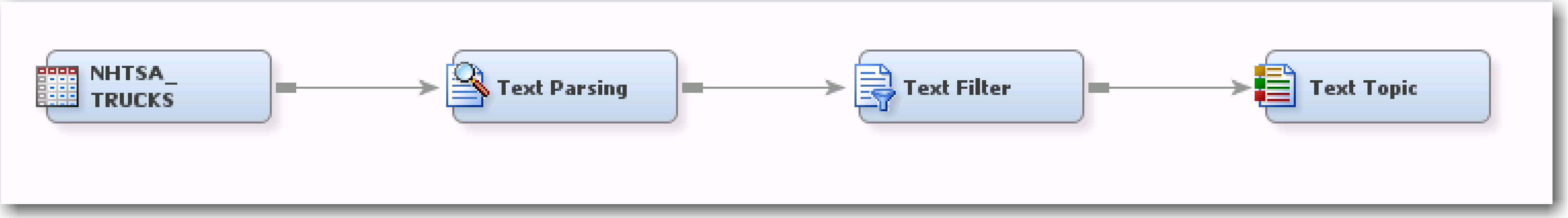
EXAMPLE TEXT MINING PROCESS FLOWS



Apply natural language processing algorithms to **parse the documents** and **quantify information** about the terms in the corpus.

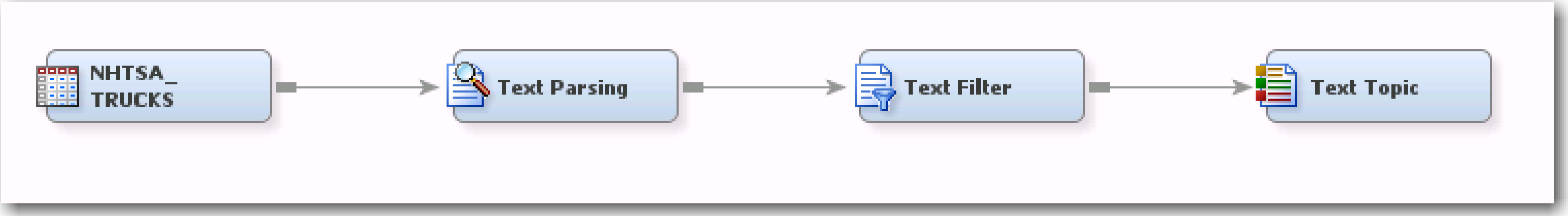
- Determine parts of speech (noun, verb, etc.)
- Perform stemming (run, runs, running, ran, etc.)
- Identify entities (names, places, etc.)

EXAMPLE TEXT MINING PROCESS FLOW



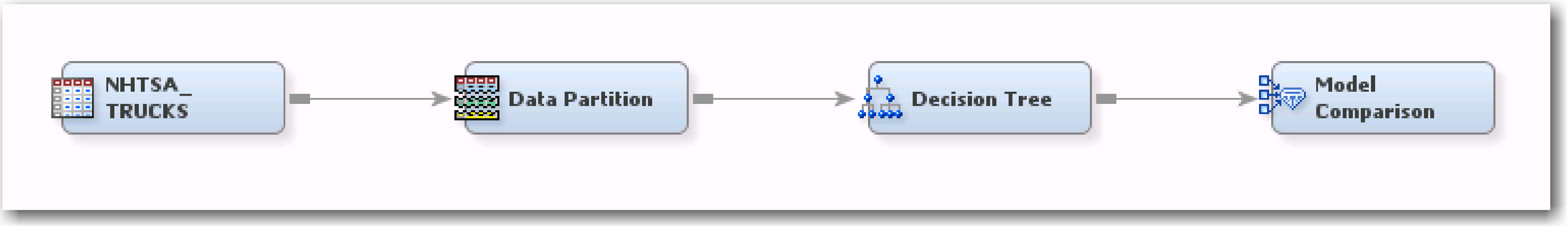
Optionally, filter the terms or documents that will be analyzed.
Can also perform spell-checking, full text searches, and analyze and view with Concept Linking

EXAMPLE TEXT MINING PROCESS FLOWS

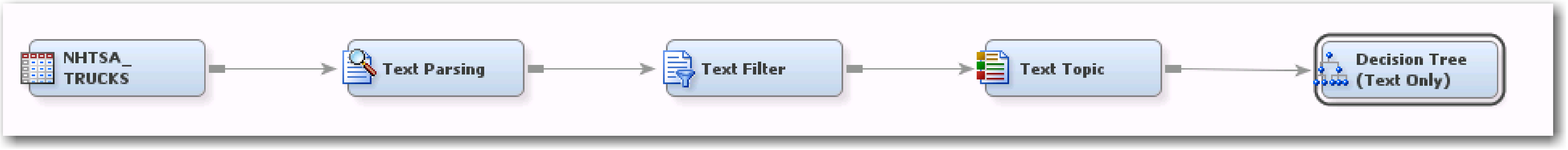


Analyze the documents to **create topics** and assign each document to one or more topics. In addition to derived topics, users can add their own topic definitions.

EXAMPLE DATA AND TEXT MINING PROCESS FLOW

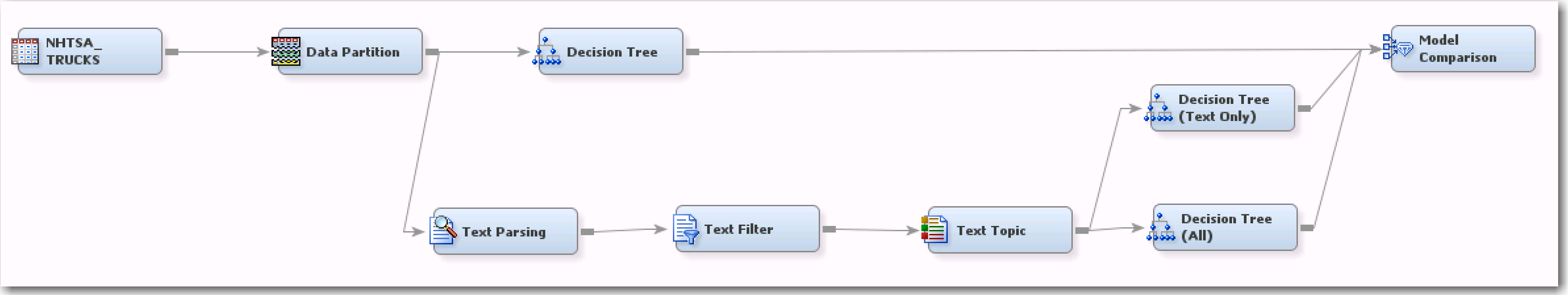


Mining Structured Data



Mining Unstructured (Text) Data

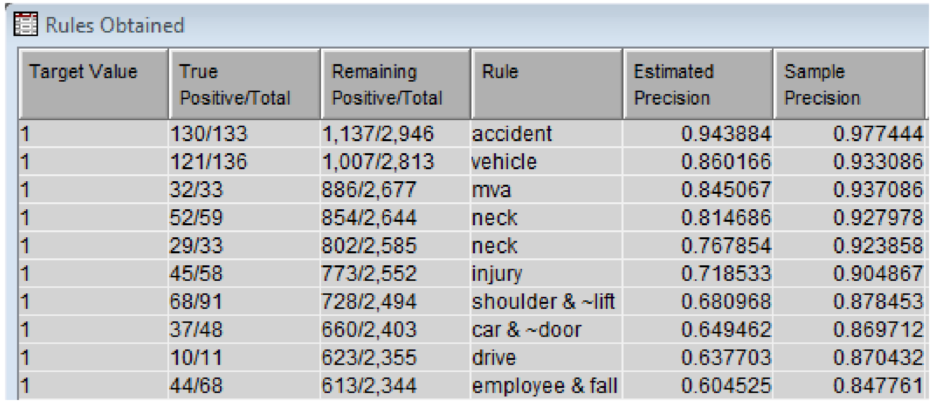
EXAMPLE DATA AND TEXT MINING PROCESS FLOW



Mining *ALL* Data:
either Structured, or Unstructured or Both

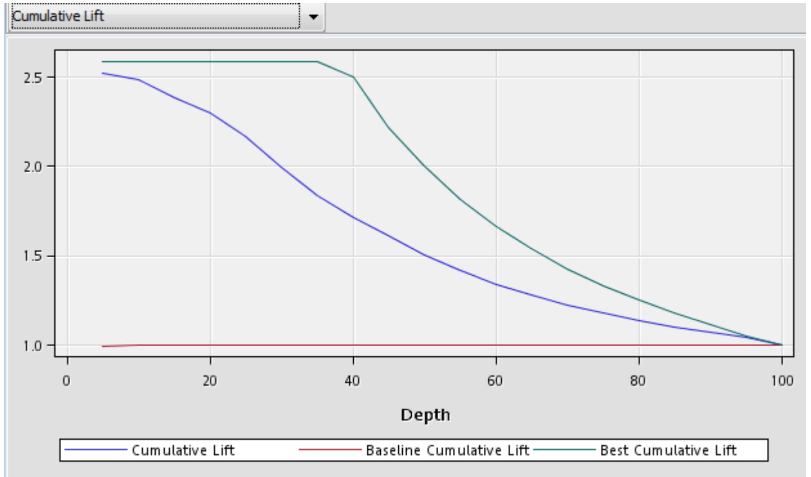
TEXT RULE BUILDER

- The **Text Rule Builder** node generates an ordered set of rules that together are useful in describing and predicting a target variable.
- Each rule in the set is associated with a specific target category, consisting of a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)).
- A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3.



Rules Obtained

Target Value	True Positive/Total	Remaining Positive/Total	Rule	Estimated Precision	Sample Precision
1	130/133	1,137/2,946	accident	0.943884	0.977444
1	121/136	1,007/2,813	vehicle	0.860166	0.933086
1	32/33	886/2,677	mva	0.845067	0.937086
1	52/59	854/2,644	neck	0.814686	0.927978
1	29/33	802/2,585	neck	0.767854	0.923858
1	45/58	773/2,552	injury	0.718533	0.904867
1	68/91	728/2,494	shoulder & ~lift	0.680968	0.878453
1	37/48	660/2,403	car & ~door	0.649462	0.869712
1	10/11	623/2,355	drive	0.637703	0.870432
1	44/68	613/2,344	employee & fall	0.604525	0.847761



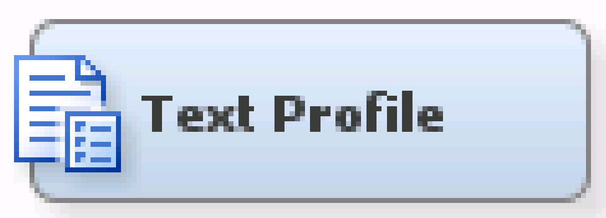
NEW NODE TEXT PROFILE NODE

A tool providing a supervised approach to discovering and reporting the terms that best **profile** a set of documents associated with each level of a target variable.

- Uses a “new” procedure, Proc TMBelief, to determine the descriptive terms.
- Useful for binary, nominal, ordinal and date target variables.
- Internally we bin date variables to day, month, year etc. and map to ordinal.
- Note: User can bin interval target variables and then analyze as nominal or ordinal.

NEW NODE | TEXT PROFILE NODE

- How are men’s and women’s attitudes different toward my product?
- How has the answer to survey question #5 varied over the last 4 years?
- What is going on in the twitter feed over the last few months?
- Is there a difference in what people are talking about in different regions of the country?



SAS® TEXT MINER™ DEMO

SAS® TEXT MINER™
WHERE TO LEARN MORE

FOR SELF-STUDY

- Visit http://support.sas.com/documentation/onlinedoc/txtminer/index.html
- Download “Getting Started with SAS Text Miner” (How to Guide) (Available for multiple versions)
- Download “Getting Started Examples (Zip)”
- Work to complete the examples.

SAS® TEXT MINER RESOURCES

SAS Text Miner Product Web Site

http://www.sas.com/text-analytics/text-miner/index.html

SAS Text Miner Technical Support Web Site

http://support.sas.com/software/products/txtminer/index.html

SAS Text Miner Technical Forum (Join Today!)

Data Mining and Text Mining Community SAS Training

Data Miner Training Path:
http://support.sas.com/training/us/paths/dm.html
Courses for SAS® Text Miner:
https://support.sas.com/edu/prodcourses.html?code=TM&ctry=US

YOUTUBE VIDEOS

- SASSoftware YouTube Channel
  - http://www.youtube.com/user/SASsoftware?feature=watch
- Manage All Unstructured Data with SAS® Text Analytics
  - http://www.youtube.com/watch?v=NHAq8jG4FX4&list=PL8BD07CC2C164FC40&index=4&feature=plpp_video
- SAS® Text Analytics Software Demo
  - http://www.youtube.com/watch?v=l1rYdrRCZJ4&feature=BFa&list=PL8BD07CC2C164FC40

Connect with me:
LinkedIn: https://www.linkedin.com/in/melodierush
Twitter: @Melodie_Rush

QUESTIONS?

Thank you for your time and attention!

CUSTOMER LOYALTY TEAM • Support You Can Count On