



Statistical Discovery™ From SAS.

TEXT ANALYTICS – LEARNING FROM UNSTRUCTURED DATA

MORS-Talk
August 21, 2020

Tom Donnelly
JMP Defense & Aerospace Team
SAS Institute Inc.
tom.donnelly@jmp.com



Heath Rushing
Co-Founder & Principal Consultant
Adsurgo, LLC
heath.rushing@adsurgo.com

AGENDA

- Prepared questions for panelists from Susan Reardon, MORS CEO
 - MORS Symposium Abstracts (& NTSB Incident Reports)
 - NSF Abstracts
- Questions from online attendees
- Resources Provided
 - Slides from Text Mining tutorial
 - Links to papers, books, and bibliography
 - Links to recordings

EXAMPLES OF TEXT DATA

USE TEXT MINING WHEN YOU CAN'T READ THEM ALL

- Documents – Word, PDF, TXT, ...
 - Abstracts/Proposals – MORS, NSF
 - Reports – NAVSAFECEN, PNSY (PDF images – arghh!)
- Open ended questions in surveys - Army SHARP
- Tweets – Russian ambassador assassination
- Emails – Enron energy scandal
- Web pages – DHS monitoring WMD
- Voice-to-text – recorded service calls - Oshkosh Corporation

Text mining revolutionizes a 24/7 customer support network

Oshkosh Corporation uses free-form service call records to systematize technical support operations and prioritize engineering improvements



Oshkosh Corporation

Challenge Optimize equipment performance and minimize machinery downtime by providing rapid, proficient technical support over the phone to customers around the world.

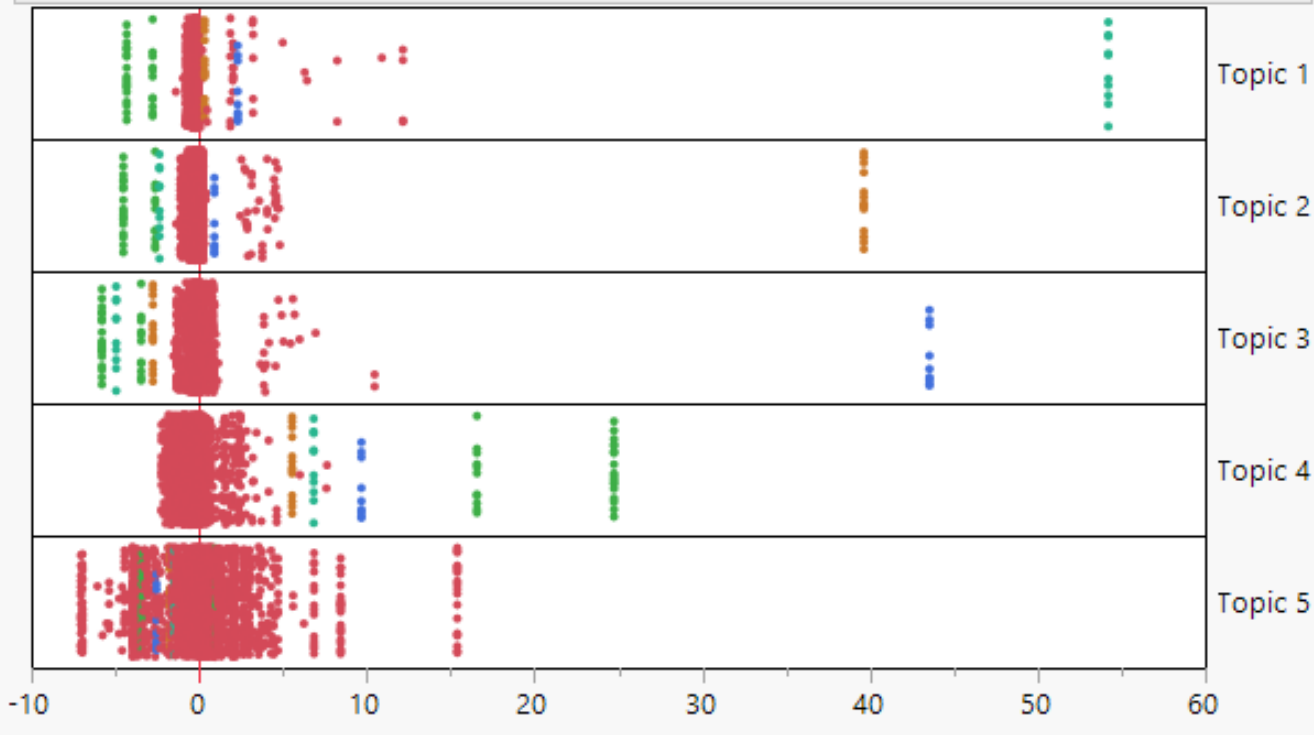
Solution Use JMP® Pro to mine unstructured text data from support calls, identifying common problems and matching sources of malfunction with verified solutions that have proven successful in the past. With an ever-growing repository of maintenance data that can be deployed to troubleshoot recurring issues, Oshkosh's 24/7 global service providers are now able to proactively resolve service calls as they come in.

Results A dramatic reduction in incident resolution time has not only saved on labor costs and improved customer satisfaction, insights gained from data analysis now guide future engineering efforts to preempt mechanical problems in products currently in development.

TWITTER CONTENT

RUSSIAN AMBASSADOR TO TURKEY ASSASSINATED IN ANKARA, DEC 19, 2016

Topic Scores Plots



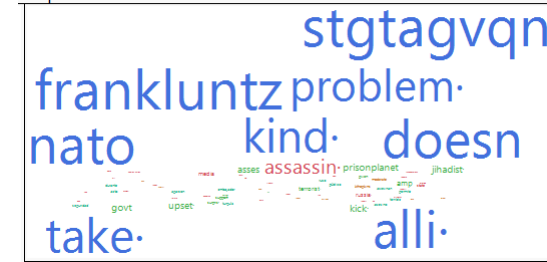
Topic 1



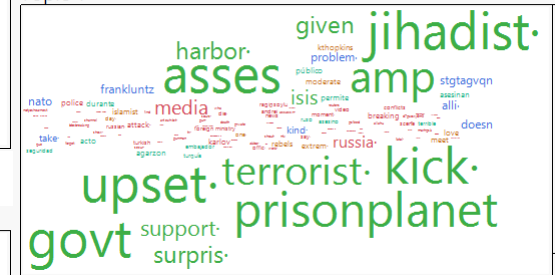
Topic 2



Topic 3



Topic 4



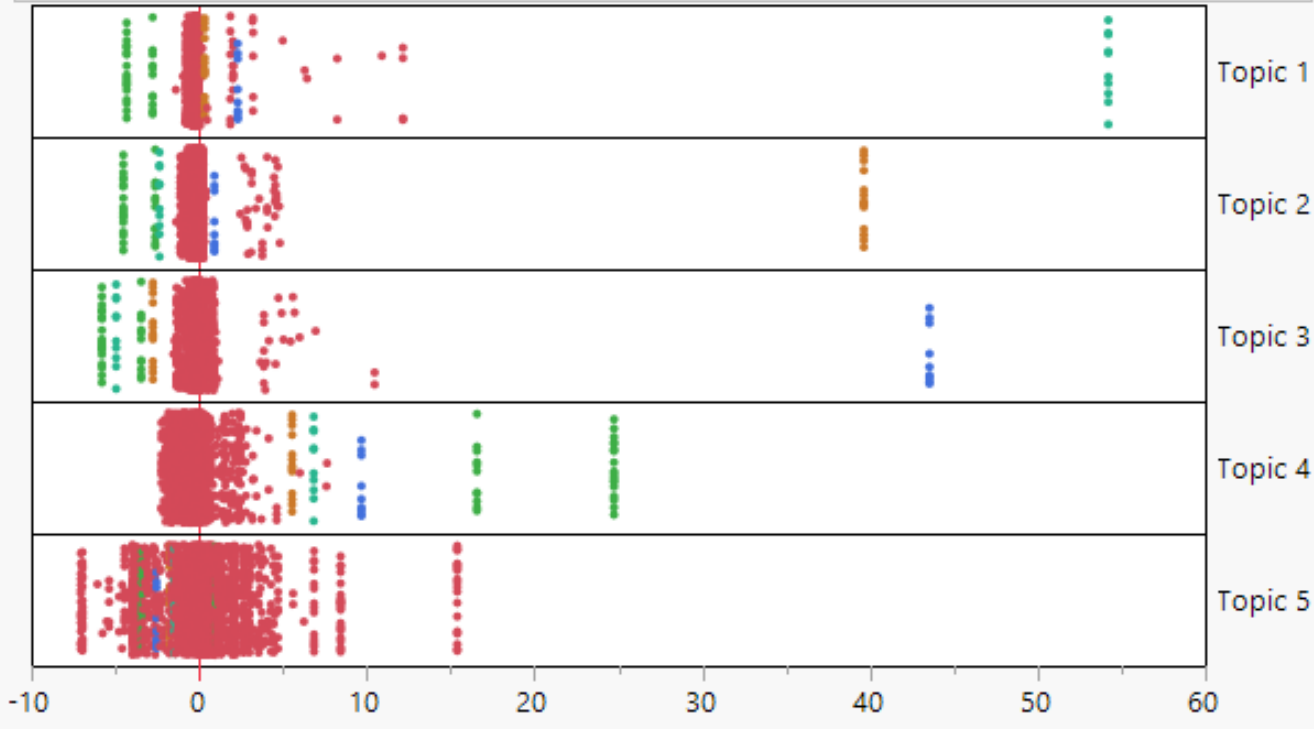
Topic 5



TWITTER CONTENT

RUSSIAN AMBASSADOR TO TURKEY ASSASSINATED IN ANKARA, DEC 19, 2016

Topic Scores Plots



RT @agarzon: Asesinan al embajador ruso en Turquía durante un acto público. Y terrible la “seguridad”, que permite al asesino da...

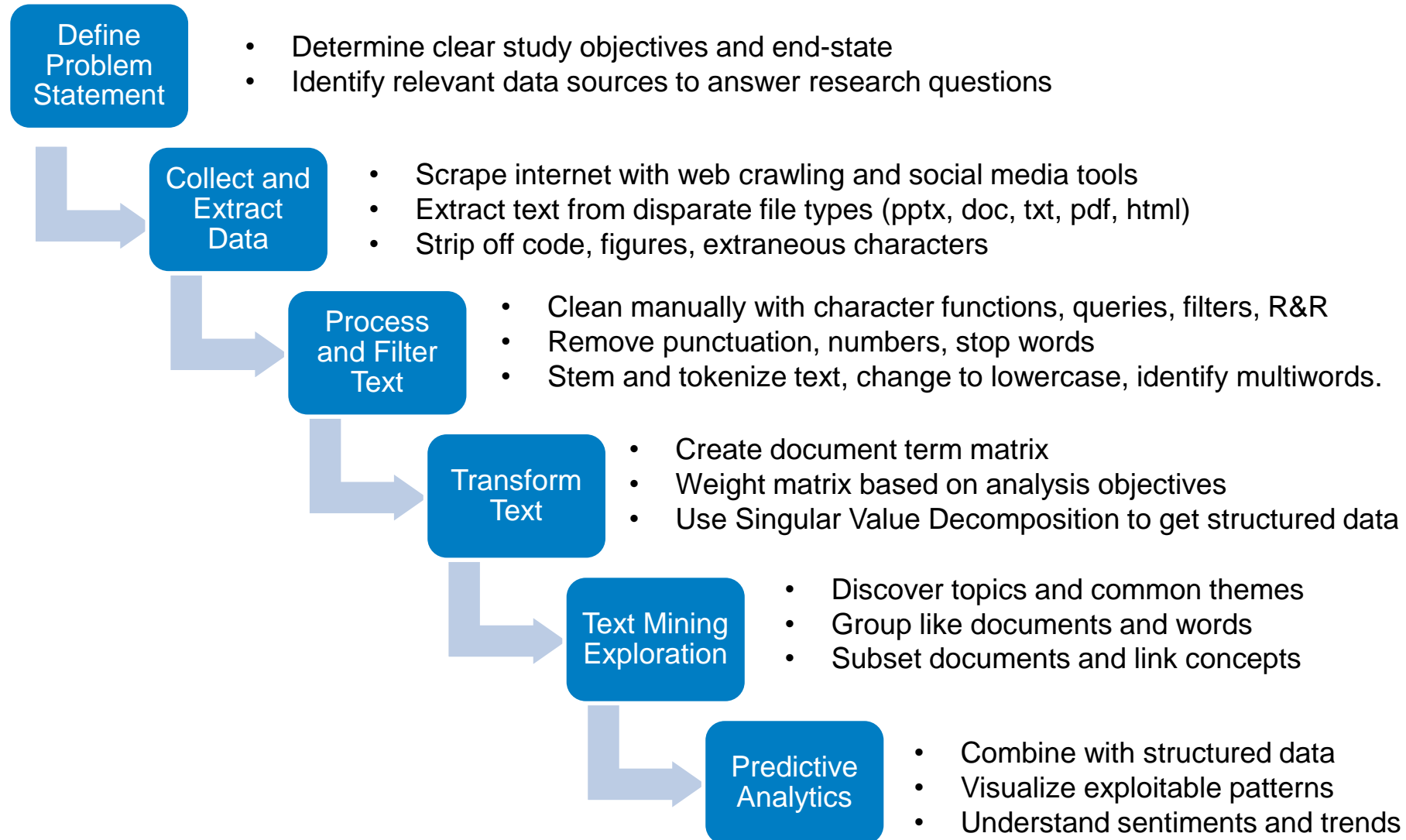
RT @KTHopkins: Russian ambassador shot dead by 'moderate Islamist rebels'. Would love to meet the extreme ones some day. #Turkey [2004]

RT @FrankLuntz: This is a problem. Turkey is a @NATO ally, and Russia doesn't take kindly to assassinations. <https://t.co/5sTGtAGvQN> [937]

RT @PrisonPlanet: The terrorist who shot the Russian ambassador was upset at Russia kicking the asses of the jihadists our media & gov. su... [1206]

RT @NotJoshEarnest: The man that shot the Russian Ambassador to Turkey screamed "Allahu Akbar" before firing. We are trying to find out wha... [508]

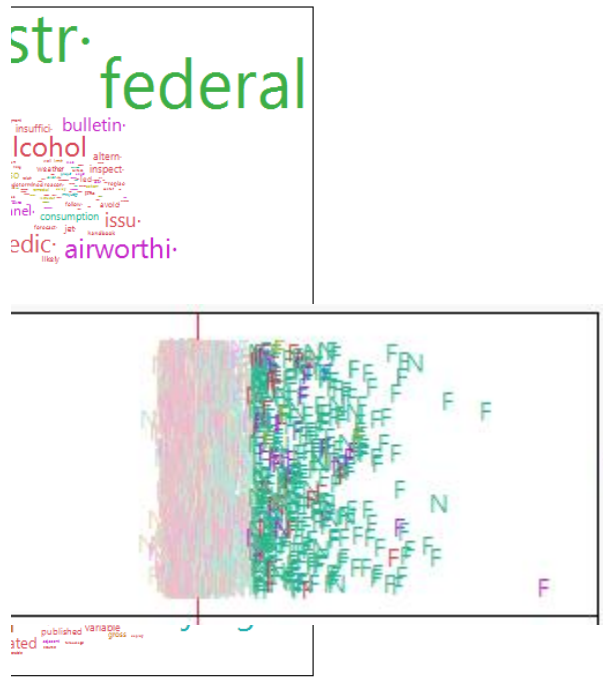
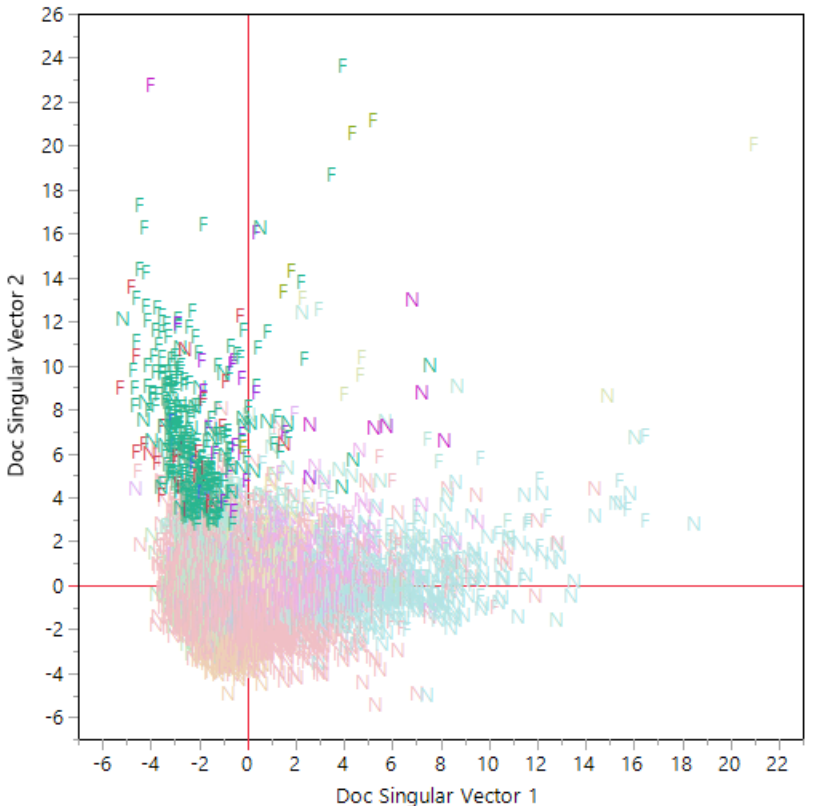
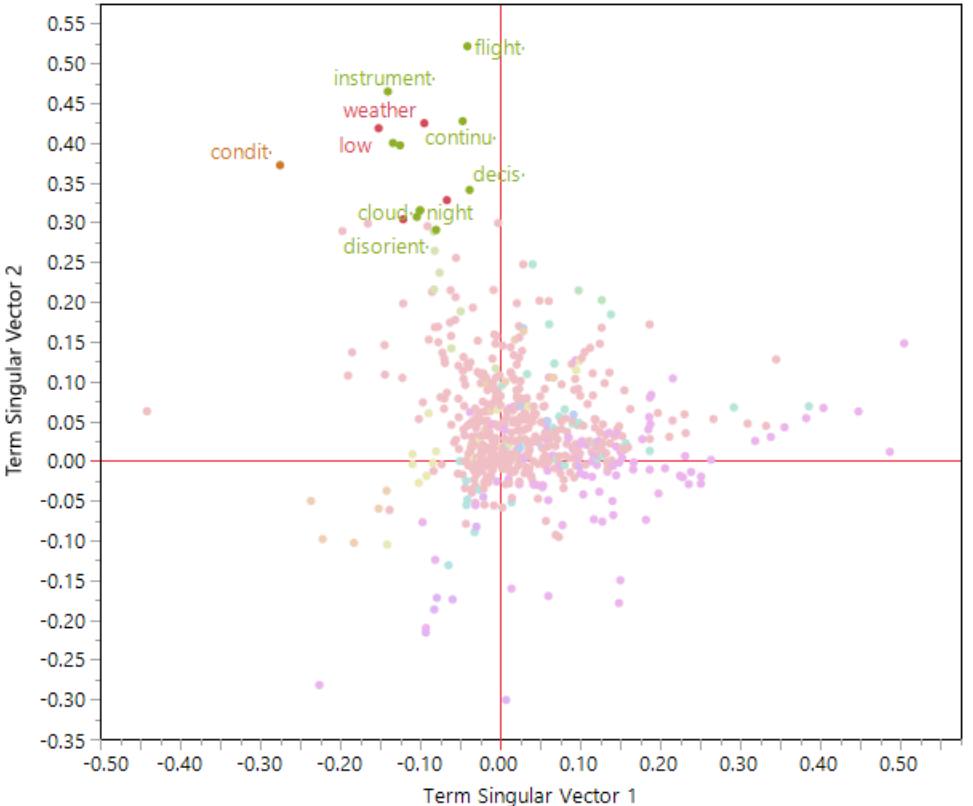
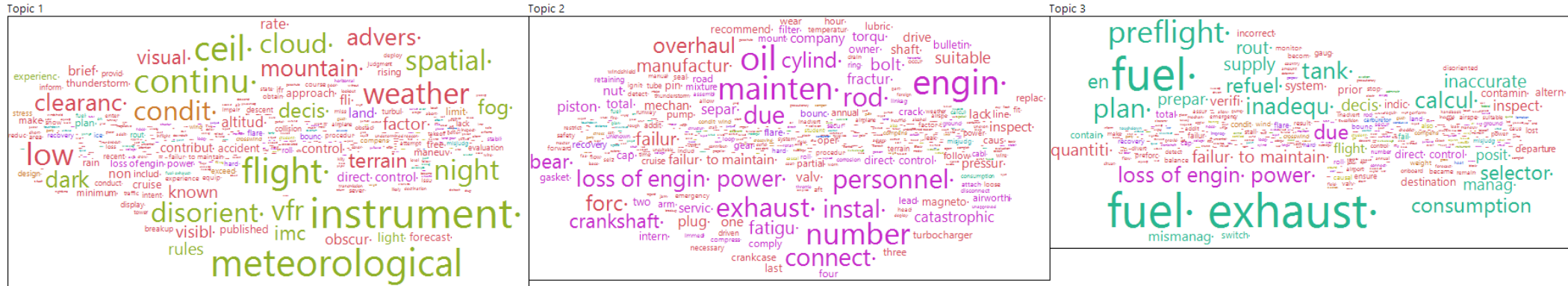
TEXT ANALYTICS FLOW



AIRCRAFT INCIDENTS DATA FROM NTSB

DIMENSION REDUCTION OF SPARSE DOCUMENT TERM MATRIX INTO DOCUMENT AND TERM VECTORS – ALSO CLUSTERING OF DOCUMENTS AND TOPICS

Word Clouds by Topic



FILTERED WORD CLOUDS

COLORED BY PROPORTION FATAL AND FILTERED BY METEOROLOGICAL CONDITIONS

Visual Meteorological Conditions

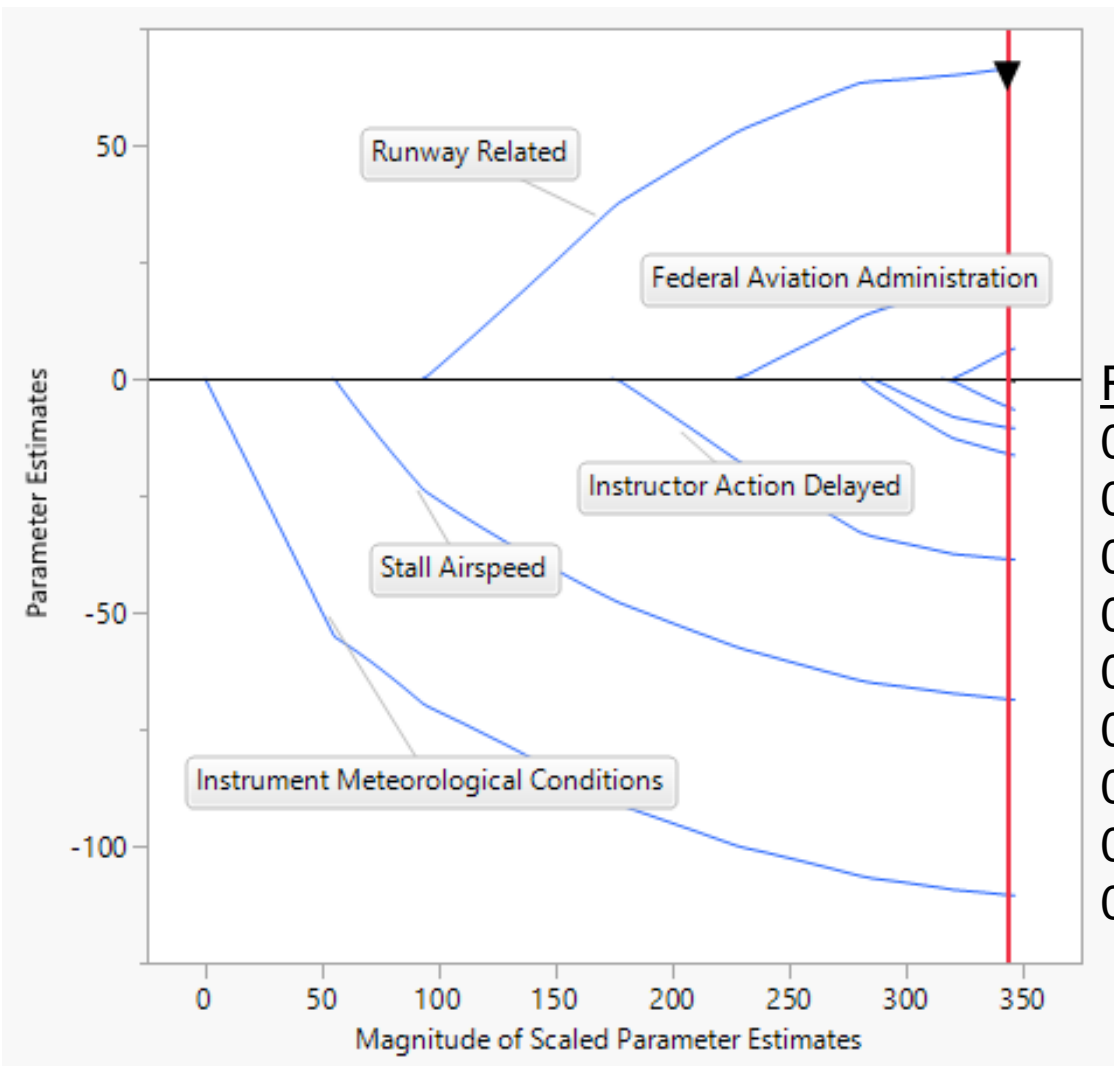
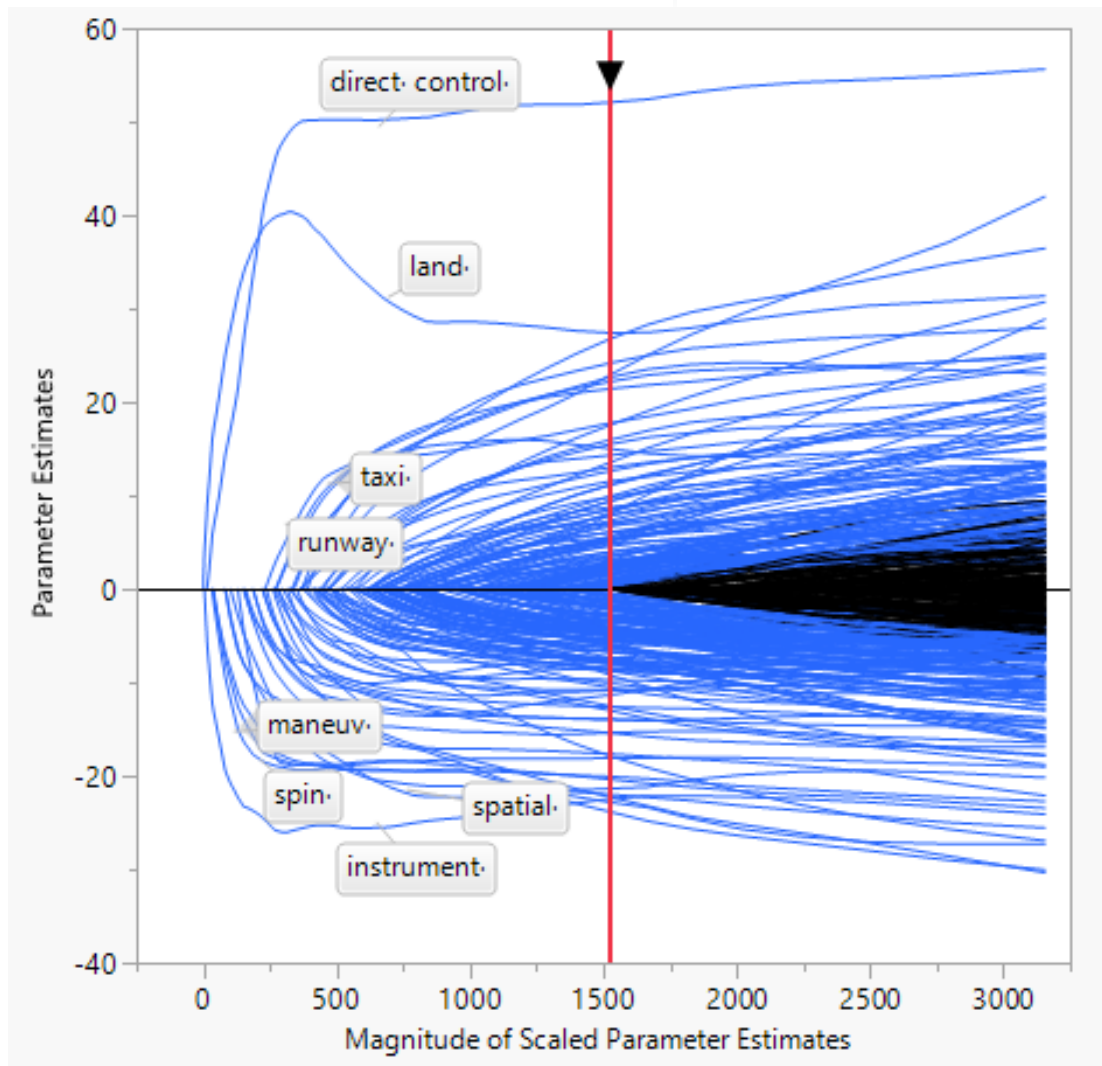


Instrument Meteorological Conditions



AIRCRAFT INCIDENTS DATA FROM NTSB

MODEL "FATAL" USING THE 630 TERMS IN THE DOCUMENT TERM MATRIX (LEFT)
AND THE TOP TEN TOPIC VECTORS (RIGHT)



R^2	
0.00	0%
0.25	53%
0.34	71%
0.43	90%
0.46	95%
0.47	98%
0.47	99%
0.48	100%
0.48	100%

A practical guide to text mining with topic extraction

Andrew Karl, James Wisnowski* and W. Heath Rushing

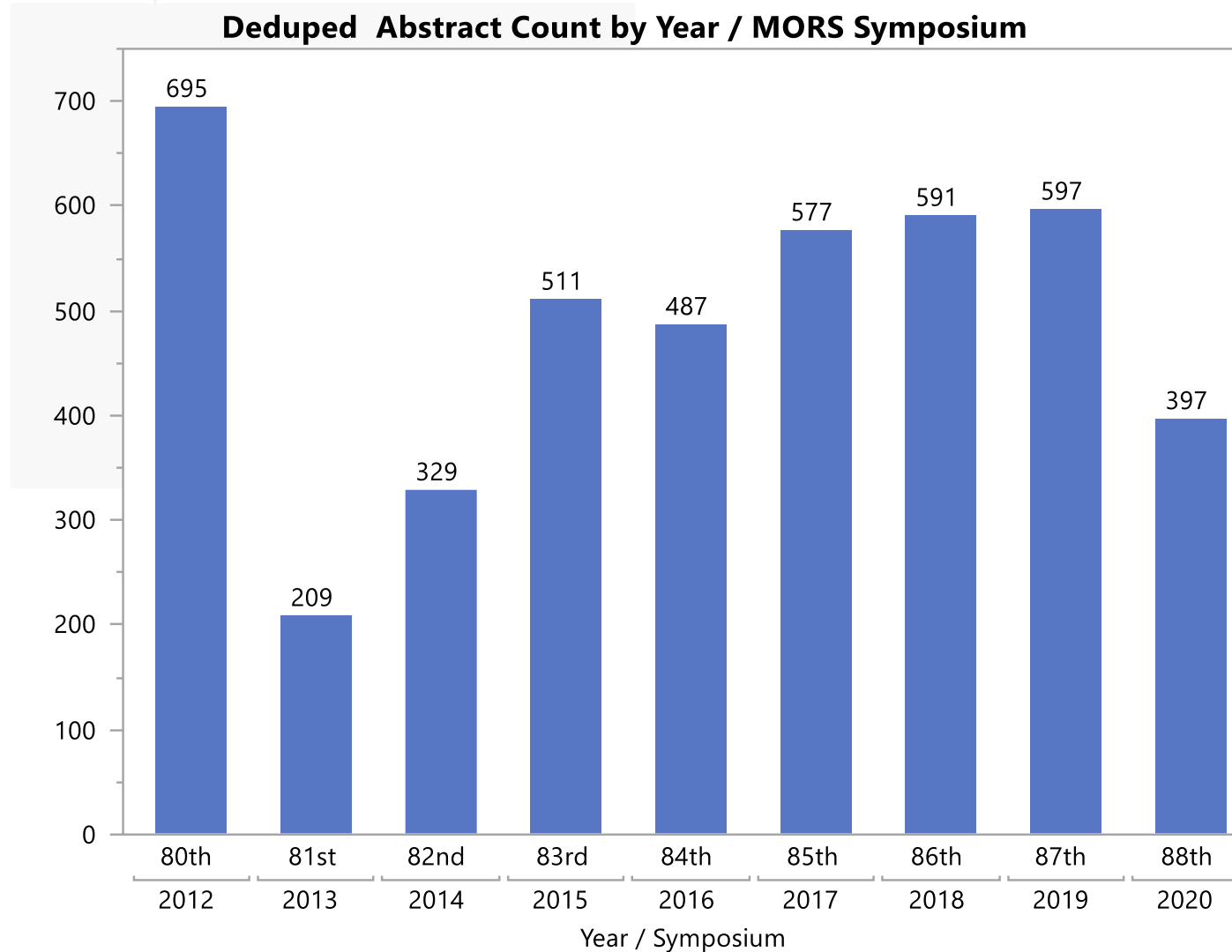


Text analytics continue to proliferate as mass volumes of unstructured but highly useful data are generated at unbounded rates. Vector space models for text data—in which documents are represented by rows and words by columns—provide a translation of this unstructured data into a format that may be analyzed with statistical and machine learning techniques. This approach gives excellent results in revealing common themes, clustering documents, clustering words, and in translating unstructured text fields (such as an open-ended survey response) to usable input variables for predictive modeling. After discussing the collection and processing of text, we explore properties and transformations of the document-term matrix (DTM). We show how the singular value decomposition may be used to drastically reduce the size of the document space while also setting the stage for automatic topic extraction, courtesy of the varimax rotation. This latent semantic analysis (LSA) approach produces factors that are compatible with graphical exploration and advanced analytics. We also explore Latent Dirichlet Allocation for topic analysis. We reference published R packages to implement the methods and conclude with a summary of other popular open-source and commercial software packages.

© 2015 Wiley Periodicals, Inc.

MORSS CORPUS OF DOCUMENTS

UNIQUE ABSTRACTS USED PER YEAR – JUST UNDER 4400 TOTAL OVER 9 YEARS



UNSTEMMED TERMS

Term and Phrase Lists

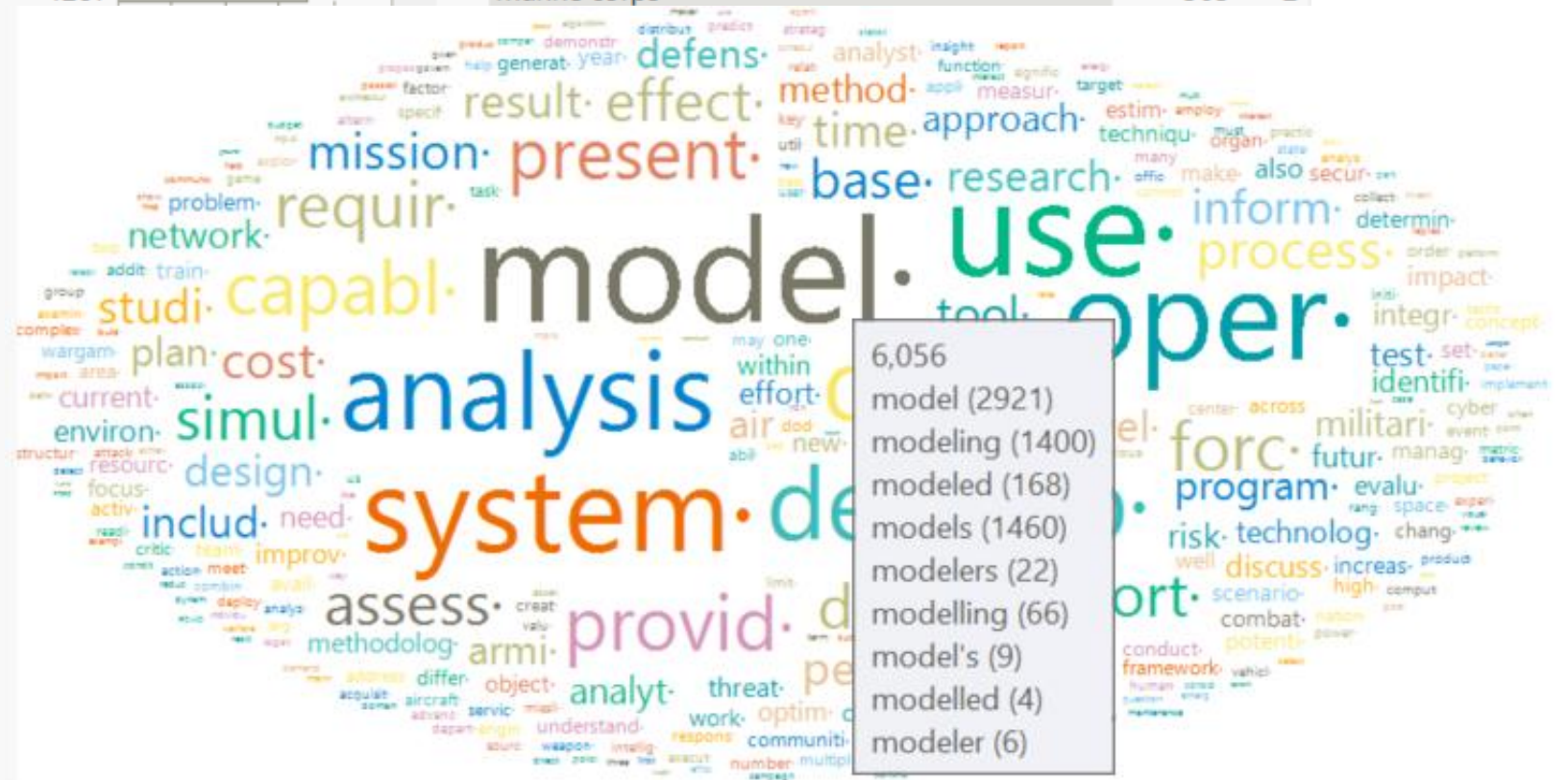
Term	Count		Phrase	Count	N
analysis	4525		air force	701	2
data	4286		decision making	423	2
model	2921		operations research	389	2
system	2474		decision makers	377	2
systems	2383		united states	340	2
operations	2149		marine corps	303	2
force					
support					
operational					
based					
used					
presentation					
research					
simulation					
mission					
using					
air					
cost					
decision					
army					
process					
development					
military					
time					
defense					



STEMMED TERMS

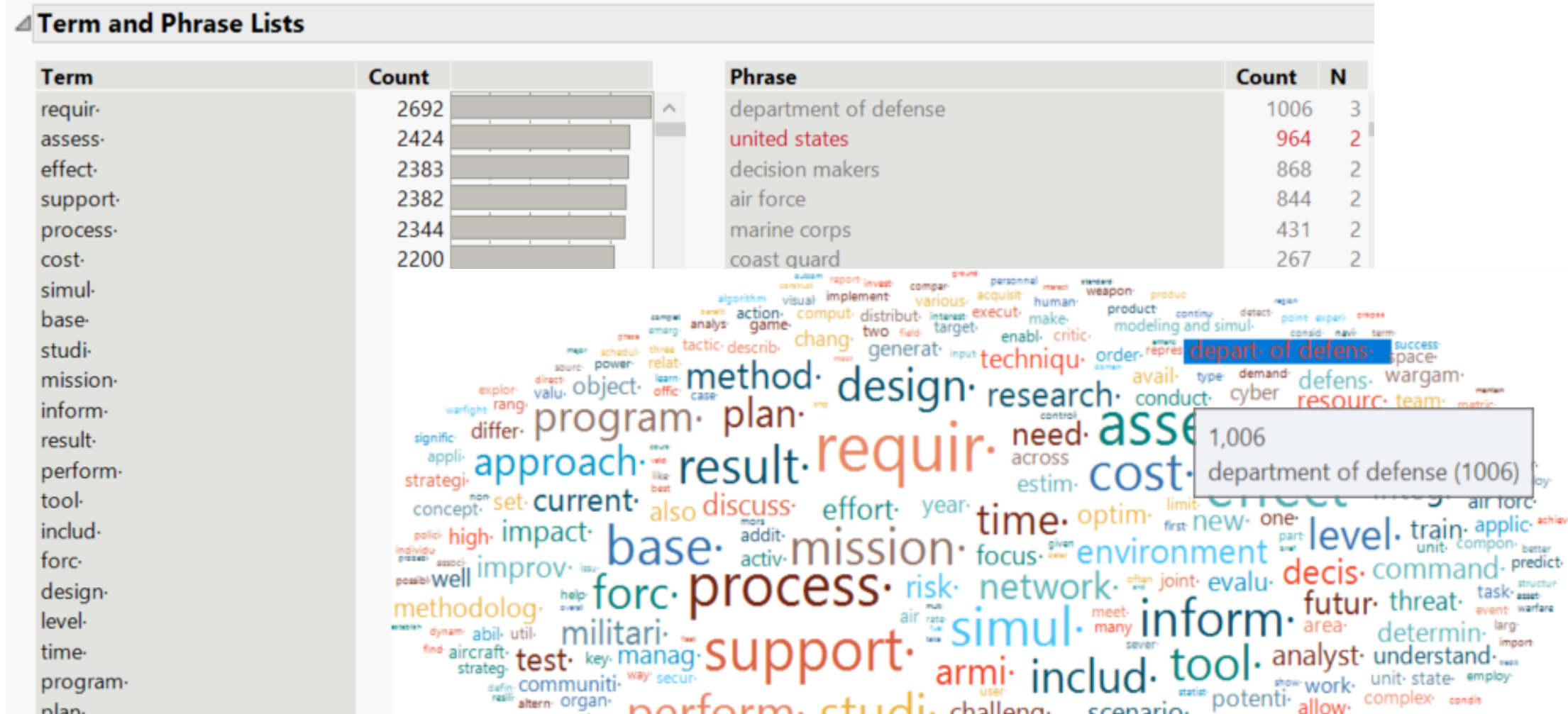
Term and Phrase Lists

Term	Count	Phrase	Count	N
model-	6056	air force	701	2
use-	5615	decision making	423	2
oper-	5297	operations research	389	2
system-	4896	decision makers	377	2
analysis	4525	united states	340	2
data-	4287	marine corps	303	2



STEMMED TERMS

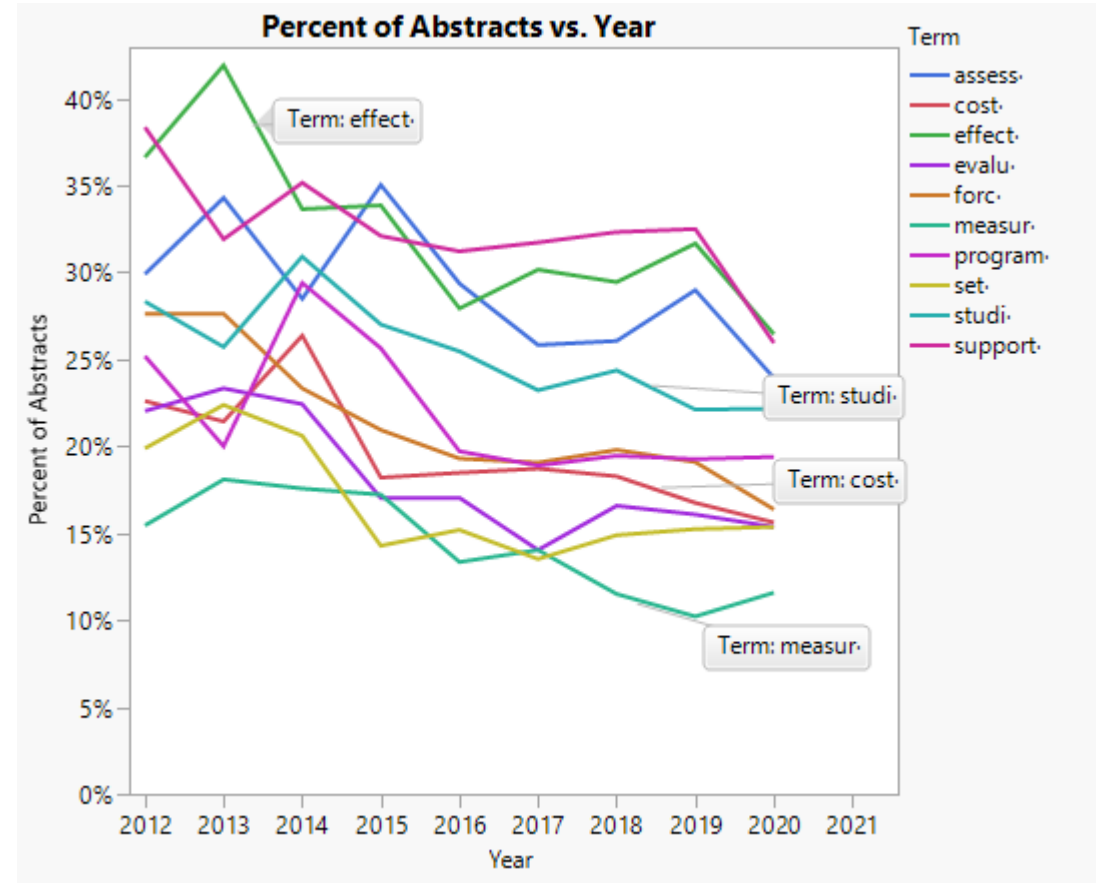
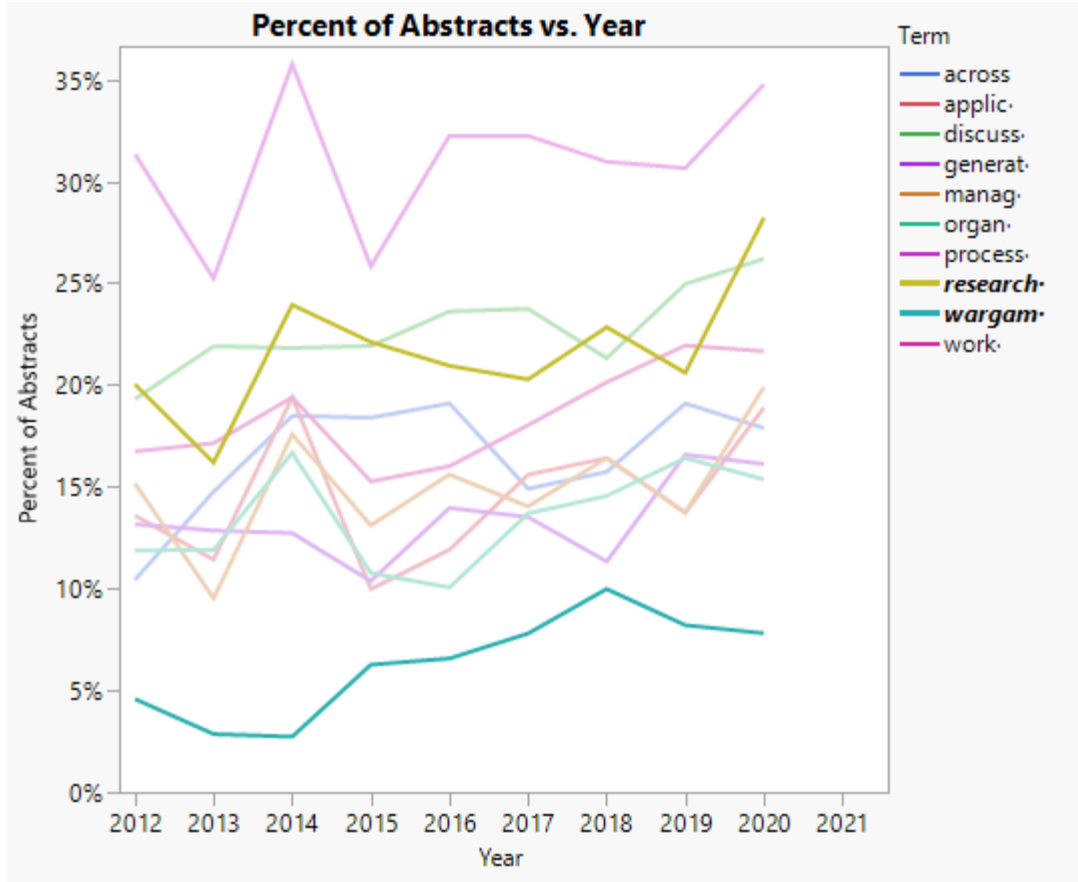
STOP WORDS REMOVED AND PHRASES ADDED TO TERMS LIST
 RECODING OF LIKE TERMS: ALL VARIANTS OF DOD = DEPARTMENT OF DEFENSE



This is the most time-consuming part of text mining, requiring subject matter expertise on choosing what terms and phrases should be analyzed.

TERMS OVER TIME

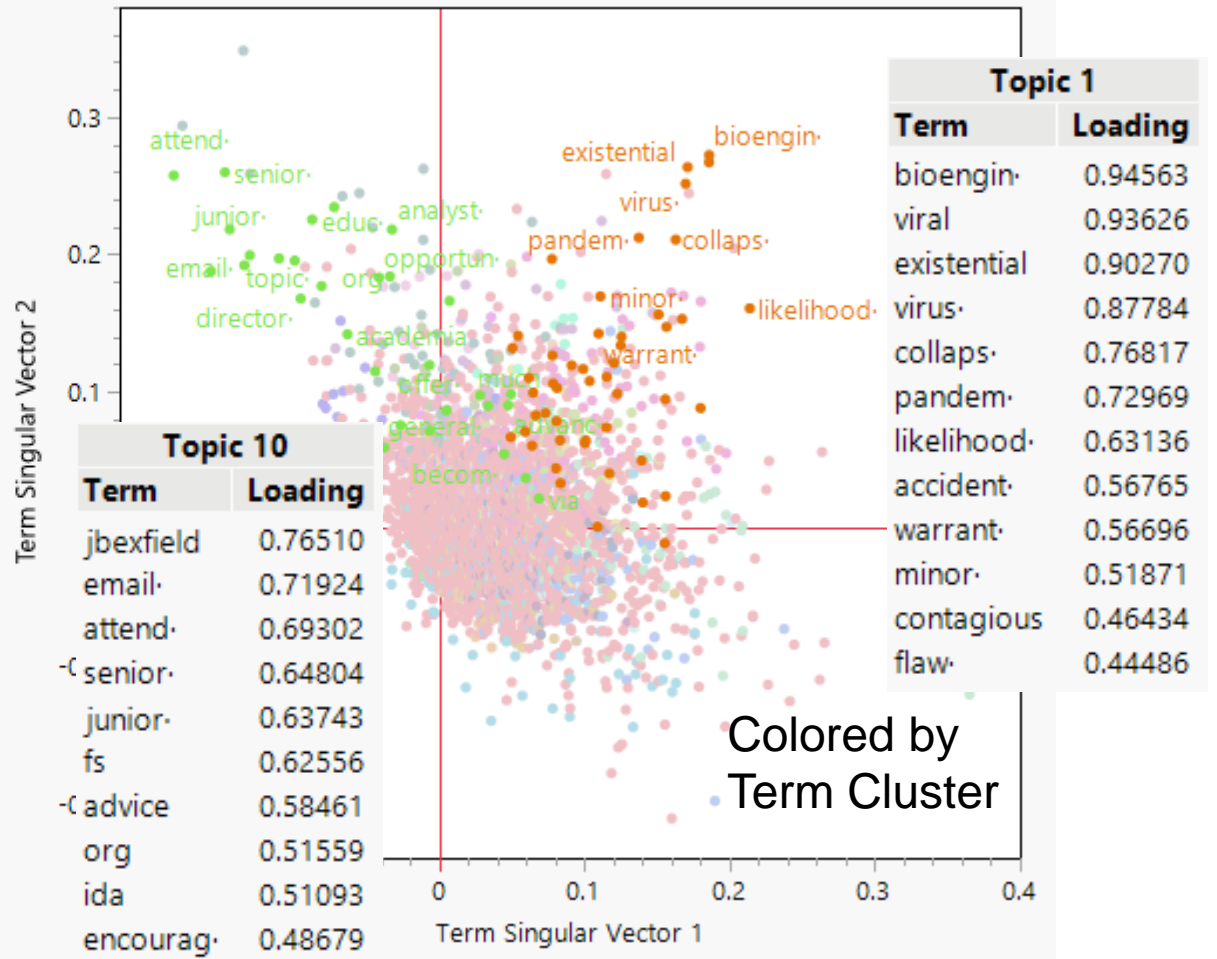
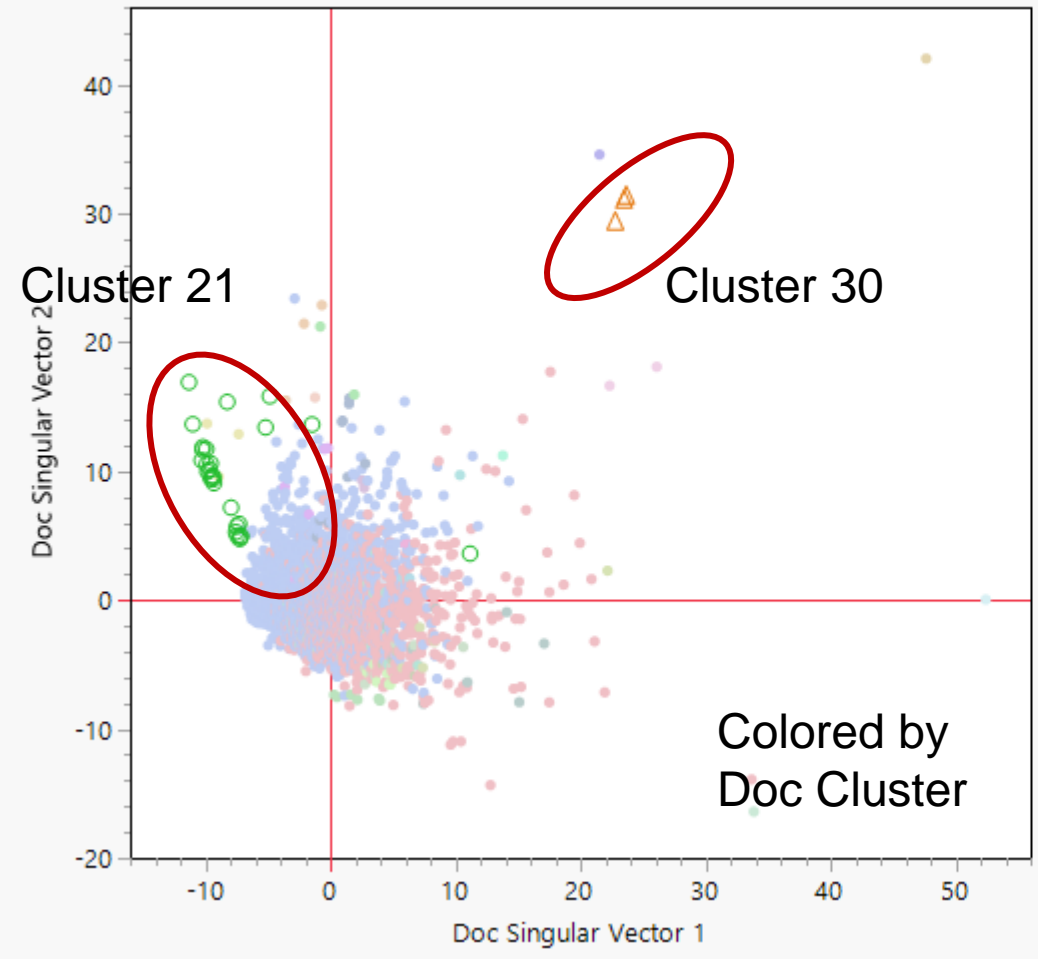
PERCENT OF ABSTRACTS USING TERMS VERSUS YEAR



MORSS ABSTRACTS DOCUMENT VECTORS & TERM VECTORS

DIMENSION REDUCTION - SINGULAR VALUE DECOMPOSITION (SVD IS LIKE PCA)

SVD Plots

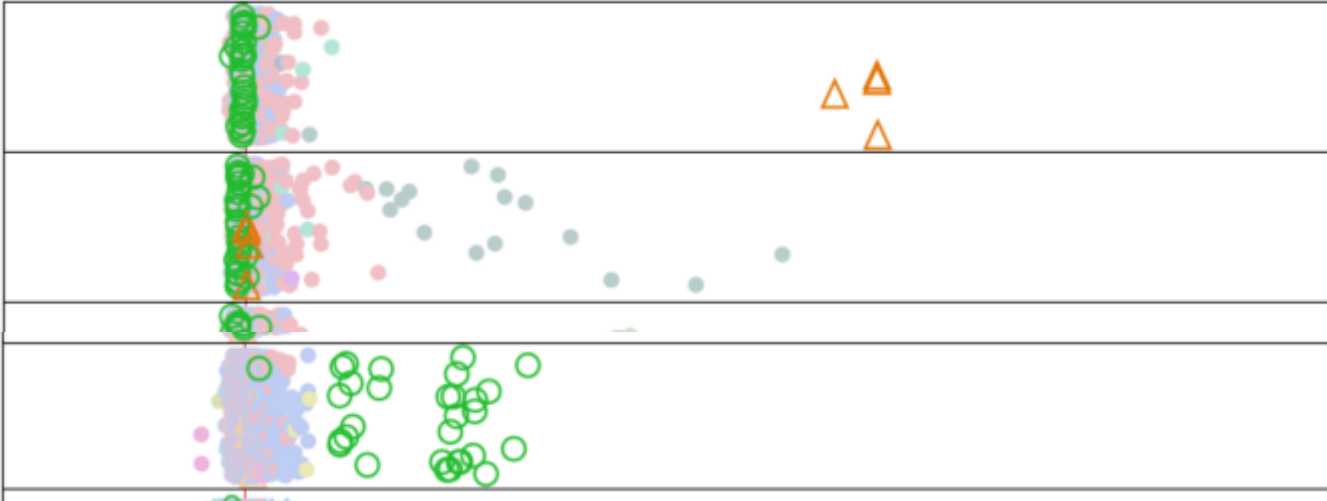


TOPIC SCORES ASSOCIATED TOPIC WORD CLOUD

Topic Analysis for 30 topics

Topic Scores Plots

[Show Text](#)



Topic 1

Topic 2

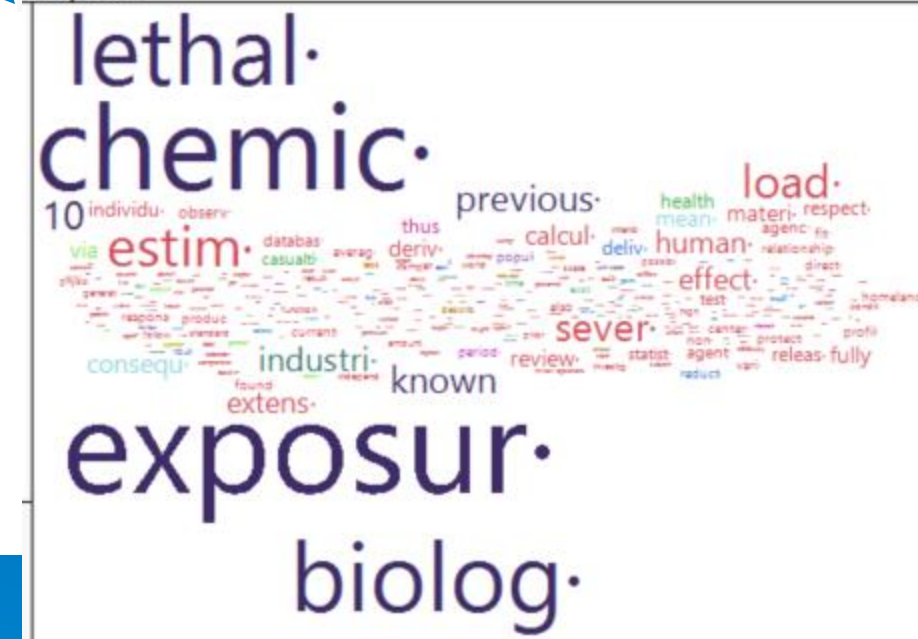
Topic 10

Word Clouds by Topic

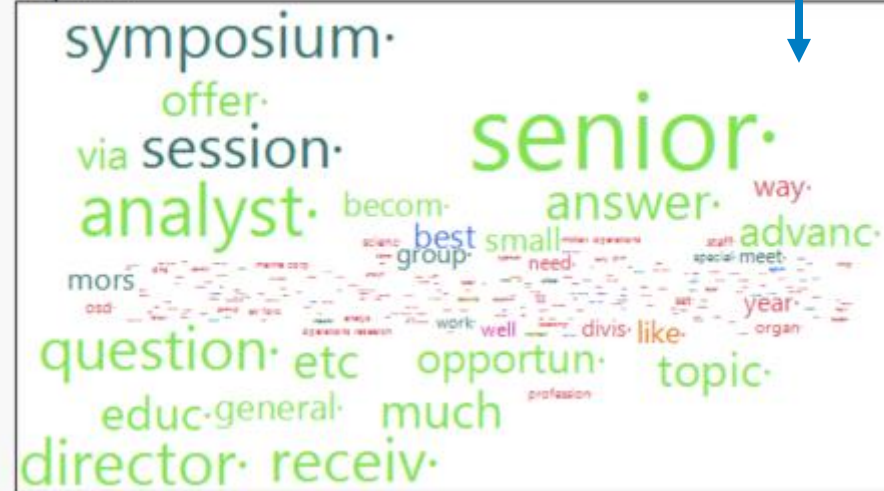
Topic 1



Topic 2



Topic 10

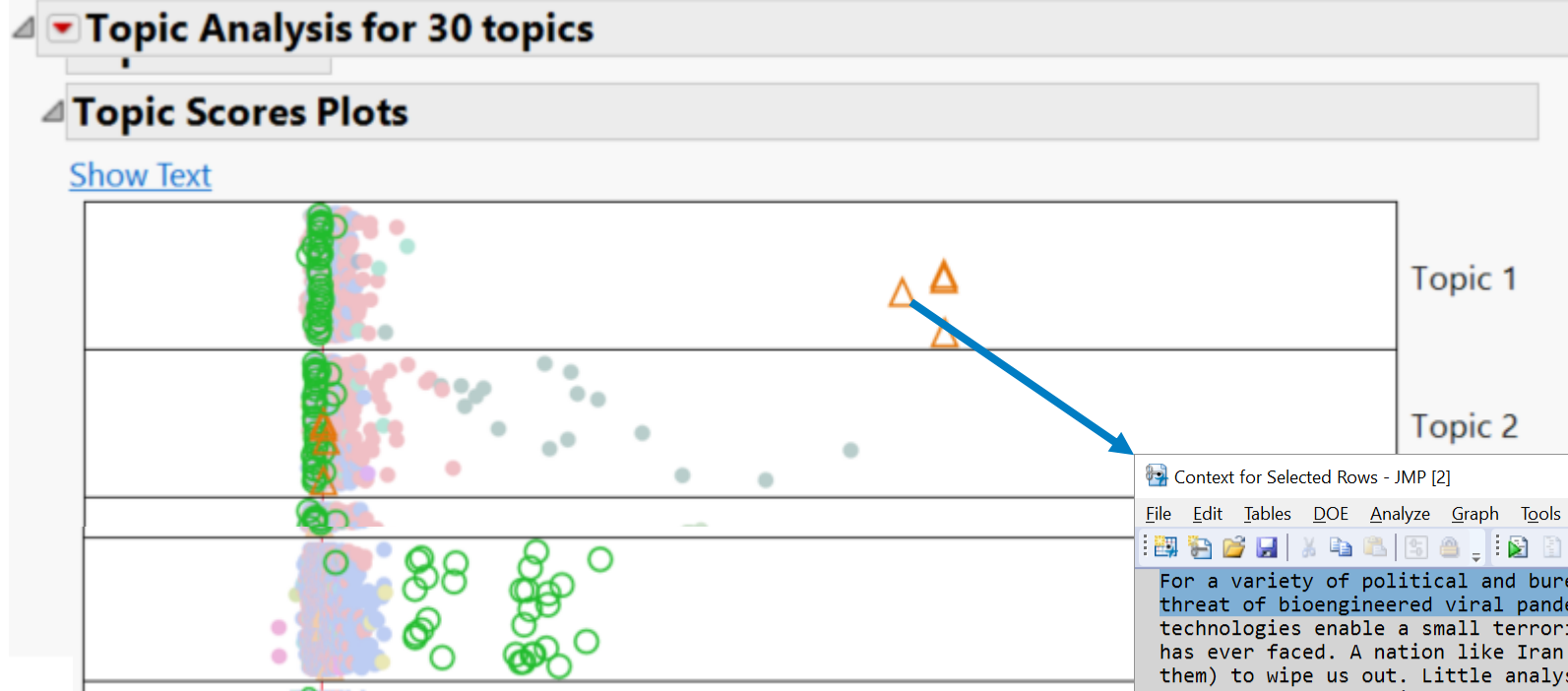


TOPIC SCORES

SELECT ANY DOCUMENT(S) AND “SHOW TEXT”

For a variety of political and bureaucratic reasons, we are not adequately preparing to deal with the threat of bioengineered viral pandemics.

(80th MORSS in 2012)



Context for Selected Rows - JMP [2]

File Edit Tables DOE Analyze Graph Tools Add-Ins View Window Help

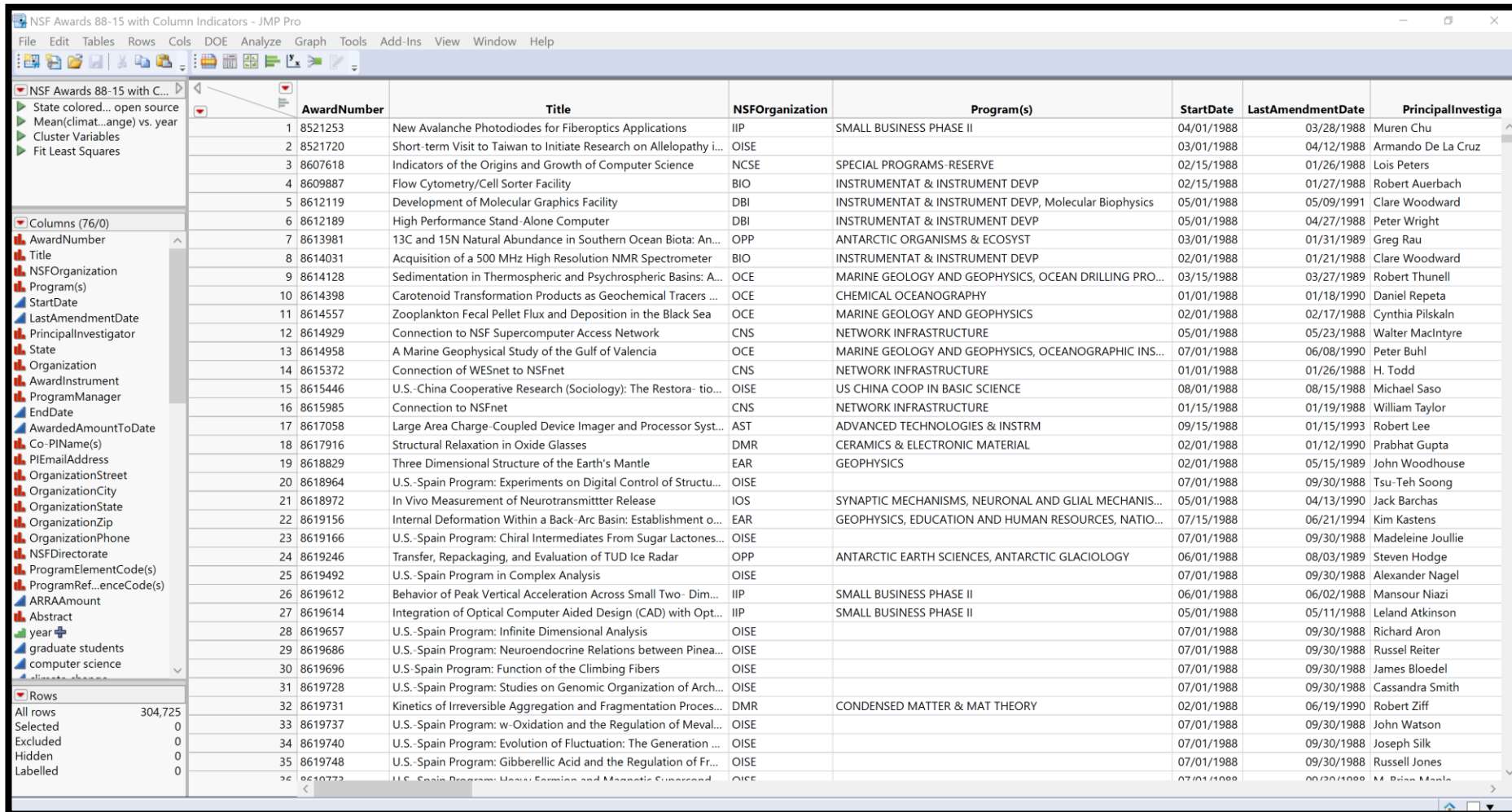
80th-88th Symposium AI

For a variety of political and bureaucratic reasons, we are not adequately preparing to deal with the threat of bioengineered viral pandemics. Rapid advances and spread of DNA manipulation and bioengineering technologies enable a small terrorist group to produce a virus that could pose the worst threat our nation has ever faced. A nation like Iran could bioengineer a highly contagious, lethal virus (and vaccine for them) to wipe us out. Little analysis is available on both the likelihood of a bioengineered virus attack or how our complex, interdependent and fragile “just in time” economy and our increasingly dependent, less resilient population will react. Economic activity may quickly cease with people not showing up for work. Lessons from Katrina and the UK riots in Aug 2011 suggest that even relatively minor disasters can lead to loss of law & order. A “Probability of Collapse Model” is used to look at the bioengineered viral pandemic threat and how it may or may not lead to a chain reaction of cascading effects and a severe collapse in the economy and domestic security.

While the likelihood of a deliberate or accidental release of a bioengineered virus is clearly rising, it is best to following Nassim Taleb’s advice from The Black Swan: the Impact of the Highly Improbable and not try to pretend we can reliably estimate the likelihood of disastrous events. Per Taleb, we are better off focusing on the consequences and how to prepare for them rather than deluding ourselves into thinking you can accurately estimate the probability of occurrence. Thus the traditional two-dimension risk model of focusing on higher consequence and higher likelihood risks is flawed. Low probability (or more accurately, unknowable probability) risks that pose a catastrophic, “existential” threat do warrant attention. The last segment of the briefing examines low cost options the Dept of Defense might take to more effectively access and use the Nat’l Guard, Active and Retired Reserves, and auxiliary forces to help restore order. A multi-criteria decision support system is used to compare these options for improving the nation’s capacity to recover from existential threats like viral pandemics and EMP attack. The author has published several articles using multi-criteria DSS including “Improving Cost Effectiveness in the Department of Defense,” Air and Space Power Journal, Spg 2010 and “Inside the Detention Camps: A New Campaign In Iraq,” Joint Force Q, Jan 2009. [3953]

NSF Abstracts

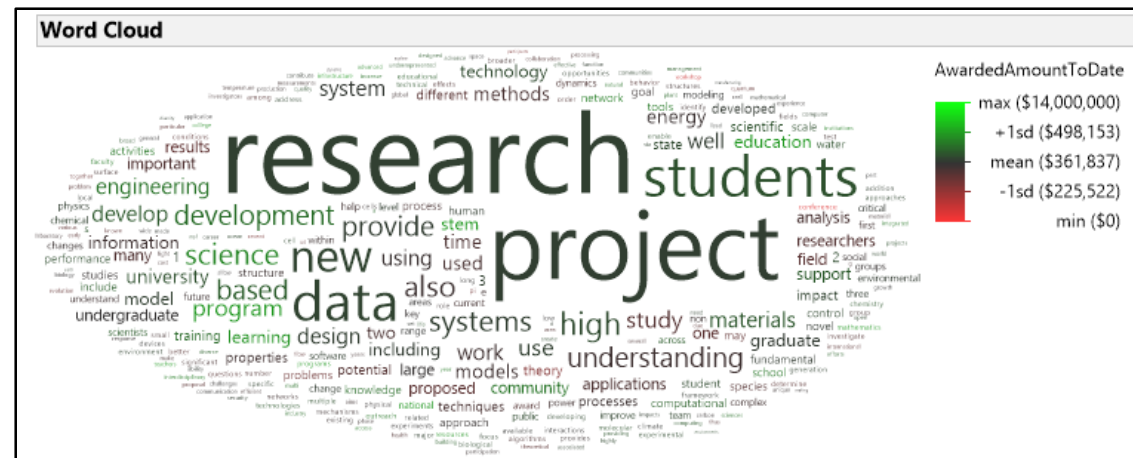
- This corpus is abstracts from the 304,725 submitted NSF grant proposals from 1988 – 2015.



AwardNumber	Title	NSFOrganization	Program(s)	StartDate	LastAmendmentDate	PrincipalInvestigator
1 8521253	New Avalanche Photodiodes for Fiberoptics Applications	IIP	SMALL BUSINESS PHASE II	04/01/1988	03/28/1988	Muren Chu
2 8521720	Short-term Visit to Taiwan to Initiate Research on Allelopathy i...	OISE		03/01/1988	04/12/1988	Armando De La Cruz
3 8607618	Indicators of the Origins and Growth of Computer Science	NCSE	SPECIAL PROGRAMS- RESERVE	02/15/1988	01/26/1988	Lois Peters
4 8609887	Flow Cytometry/Cell Sorter Facility	BIO	INSTRUMENTAT & INSTRUMENT DEVP	02/15/1988	01/27/1988	Robert Auerbach
5 8612119	Development of Molecular Graphics Facility	DBI	INSTRUMENTAT & INSTRUMENT DEVP, Molecular Biophysics	05/01/1988	05/09/1991	Clare Woodward
6 8612189	High Performance Stand-Alone Computer	DBI	INSTRUMENTAT & INSTRUMENT DEVP	05/01/1988	04/27/1988	Peter Wright
7 8613981	13C and 15N Natural Abundance in Southern Ocean Biota: An...	OPP	ANTARCTIC ORGANISMS & ECOSYST	03/01/1988	01/31/1989	Greg Rau
8 8614031	Acquisition of a 500 MHz High Resolution NMR Spectrometer	BIO	INSTRUMENTAT & INSTRUMENT DEVP	02/01/1988	01/21/1988	Clare Woodward
9 8614128	Sedimentation in Thermospheric and Psychrospheric Basins: A...	OCE	MARINE GEOLOGY AND GEOPHYSICS, OCEAN DRILLING PRO...	03/15/1988	03/27/1989	Robert Thunell
10 8614398	Carotenoid Transformation Products as Geochemical Tracers ...	OCE	CHEMICAL OCEANOGRAPHY	01/01/1988	01/18/1990	Daniel Repeta
11 8614557	Zooplankton Fecal Pellet Flux and Deposition in the Black Sea	OCE	MARINE GEOLOGY AND GEOPHYSICS	02/01/1988	02/17/1988	Cynthia Pilskaln
12 8614929	Connection to NSF Supercomputer Access Network	CNS	NETWORK INFRASTRUCTURE	05/01/1988	05/23/1988	Walter MacIntyre
13 8614958	A Marine Geophysical Study of the Gulf of Valencia	OCE	MARINE GEOLOGY AND GEOPHYSICS, OCEANOGRAPHIC INS...	07/01/1988	06/08/1990	Peter Buhl
14 8615372	Connection of WESnet to NSFnet	CNS	NETWORK INFRASTRUCTURE	01/01/1988	01/26/1988	H. Todd
15 8615446	U.S.- China Cooperative Research (Sociology): The Restora- tio...	OISE	US CHINA COOP IN BASIC SCIENCE	08/01/1988	08/15/1988	Michael Saso
16 8615985	Connection to NSFnet	CNS	NETWORK INFRASTRUCTURE	01/15/1988	01/19/1988	William Taylor
17 8617058	Large Area Charge- Coupled Device Imager and Processor Syst...	AST	ADVANCED TECHNOLOGIES & INSTRM	09/15/1988	01/15/1993	Robert Lee
18 8617916	Structural Relaxation in Oxide Glasses	DMR	CERAMICS & ELECTRONIC MATERIAL	02/01/1988	01/12/1990	Prabhat Gupta
19 8618829	Three Dimensional Structure of the Earth's Mantle	EAR	GEOPHYSICS	02/01/1988	05/15/1989	John Woodhouse
20 8618964	U.S.-Spain Program: Experiments on Digital Control of Structu...	OISE		07/01/1988	09/30/1988	Tsu-Teh Soong
21 8618972	In Vivo Measurement of Neurotransmitter Release	IOS	SYNAPTIC MECHANISMS, NEURONAL AND GLIAL MECHANIS...	05/01/1988	04/13/1990	Jack Barchas
22 8619156	Internal Deformation Within a Back-Arc Basin: Establishment o...	EAR	GEOPHYSICS, EDUCATION AND HUMAN RESOURCES, NATIO...	07/15/1988	06/21/1994	Kim Kastens
23 8619166	U.S.-Spain Program: Chiral Intermediates From Sugar Lactones...	OISE		07/01/1988	09/30/1988	Madeleine Joulle
24 8619246	Transfer, Repackaging, and Evaluation of TUD Ice Radar	OPP	ANTARCTIC EARTH SCIENCES, ANTARCTIC GLACIOLOGY	06/01/1988	08/03/1989	Steven Hodge
25 8619492	U.S.-Spain Program in Complex Analysis	OISE		07/01/1988	09/30/1988	Alexander Nagel
26 8619612	Behavior of Peak Vertical Acceleration Across Small Two- Dim...	IIP	SMALL BUSINESS PHASE II	06/01/1988	06/02/1988	Mansour Niazi
27 8619614	Integration of Optical Computer Aided Design (CAD) with Opt...	IIP	SMALL BUSINESS PHASE II	05/01/1988	05/11/1988	Leland Atkinson
28 8619657	U.S.-Spain Program: Infinite Dimensional Analysis	OISE		07/01/1988	09/30/1988	Richard Aron
29 8619686	U.S.-Spain Program: Neuroendocrine Relations between Pine...	OISE		07/01/1988	09/30/1988	Russel Reiter
30 8619696	U.S.-Spain Program: Function of the Climbing Fibers	OISE		07/01/1988	09/30/1988	James Bloedel
31 8619728	U.S.-Spain Program: Studies on Genomic Organization of Arch...	OISE		07/01/1988	09/30/1988	Cassandra Smith
32 8619731	Kinetics of Irreversible Aggregation and Fragmentation Proces...	DMR	CONDENSED MATTER & MAT THEORY	02/01/1988	06/19/1990	Robert Ziff
33 8619737	U.S.-Spain Program: w-Oxidation and the Regulation of Meval...	OISE		07/01/1988	09/30/1988	John Watson
34 8619740	U.S.-Spain Program: Evolution of Fluctuation: The Generation ...	OISE		07/01/1988	09/30/1988	Joseph Silk
35 8619748	U.S.-Spain Program: Gibberellic Acid and the Regulation of Fr...	OISE		07/01/1988	09/30/1988	Russell Jones
36 8619772	U.S.-Spain Program: Heavy Fermion and Magnetic Supercond...	OISE		07/01/1988	09/30/1988	M. Brian Maple

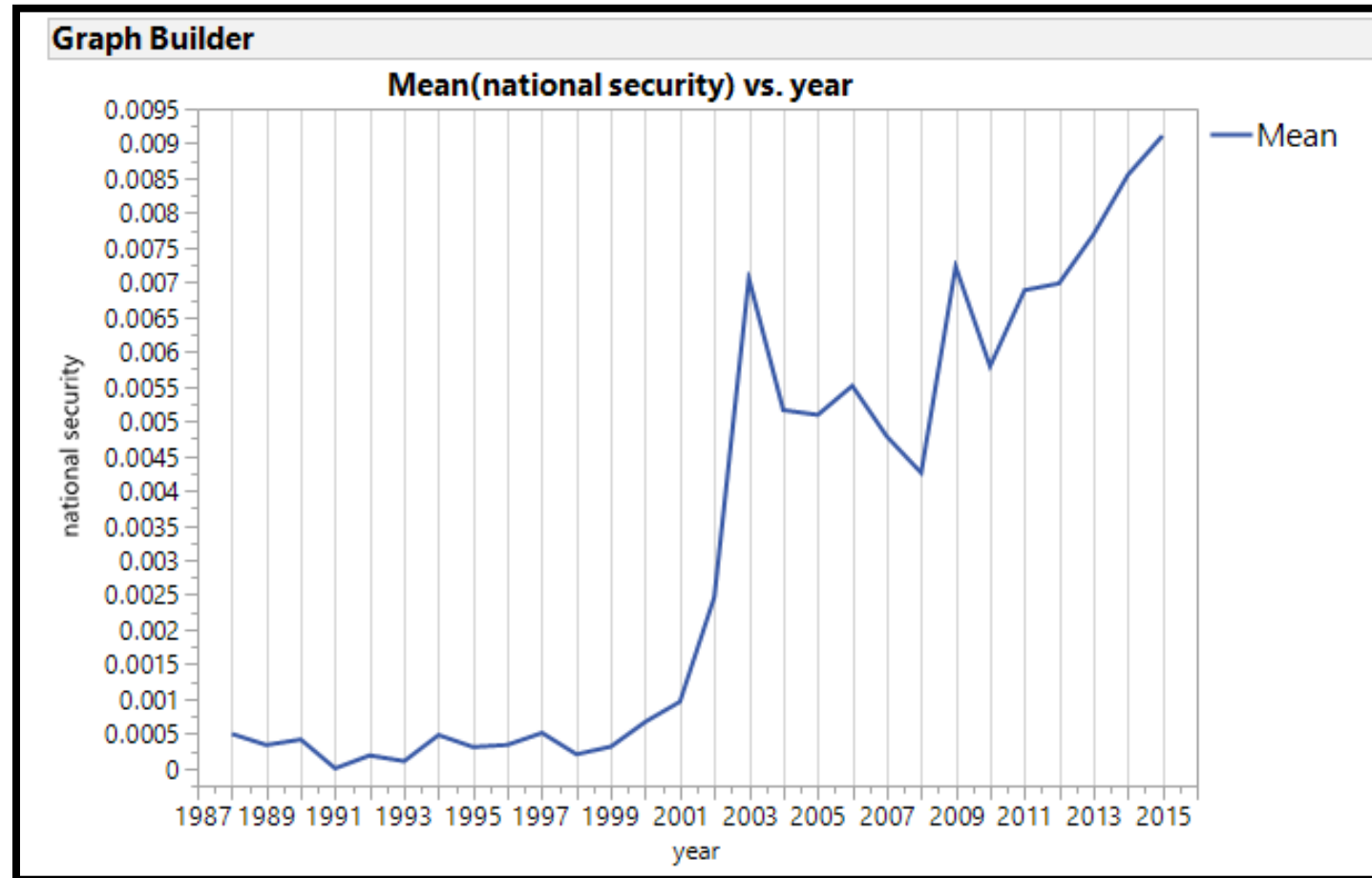
- Frequent terms and phrases:

Term and Phrase Lists				
Term	Count	Phrase	Count	N
research	34410	graduate students	2590	2
project	29465	high school	2036	2
students	17662	e g	2030	2
data	17123	undergraduate students	1773	2
new	14391	u s	1735	2
high	9942	large scale	1367	2
systems	9628	long term	1338	2
science	9157	proposed research	1320	2
also	9112	research project	1225	2
provide	8678	next generation	1222	2
understanding	8674	3 d	1143	2



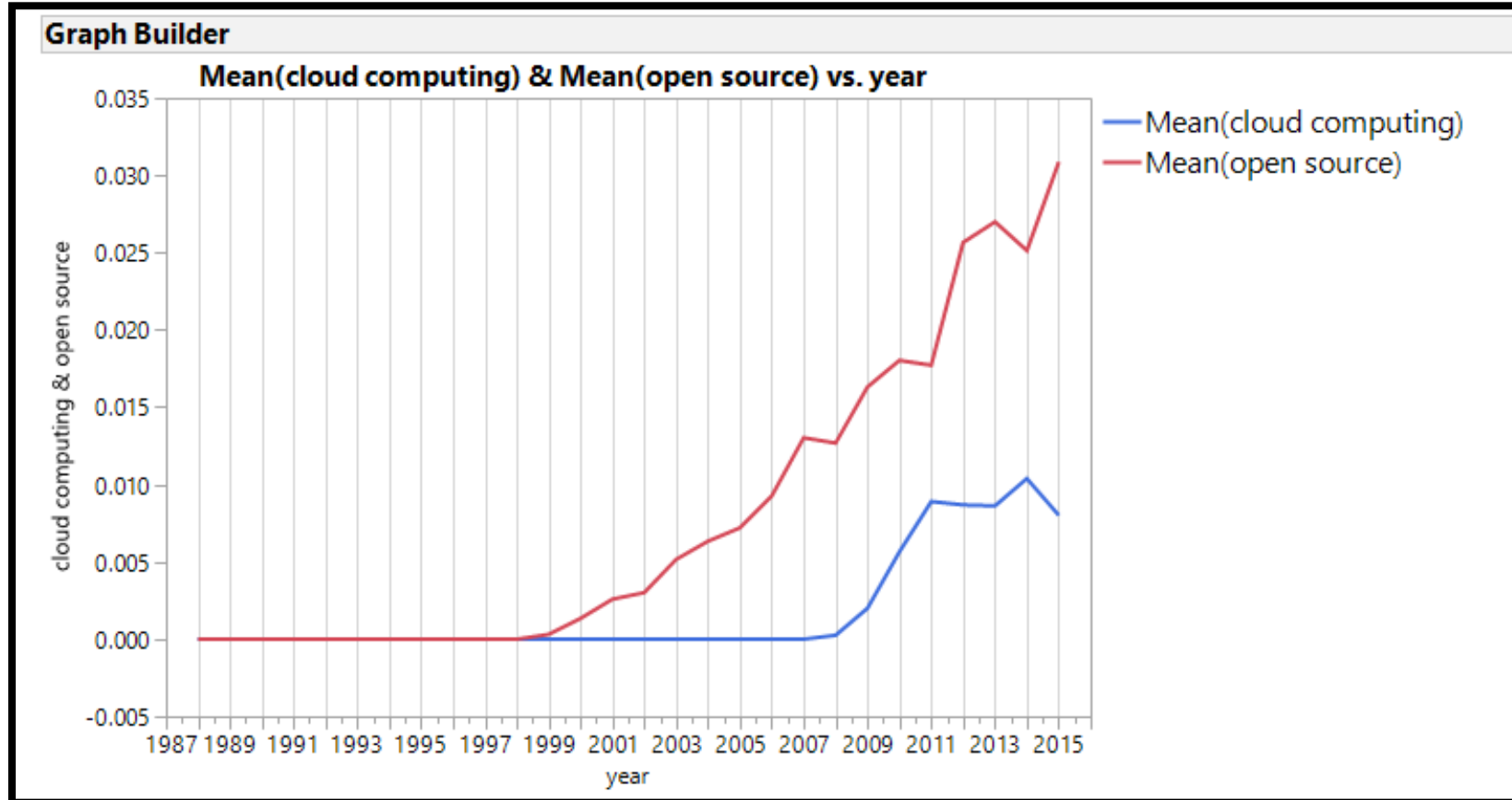
NSF Abstracts

- Certain phrases become much more frequent over time.



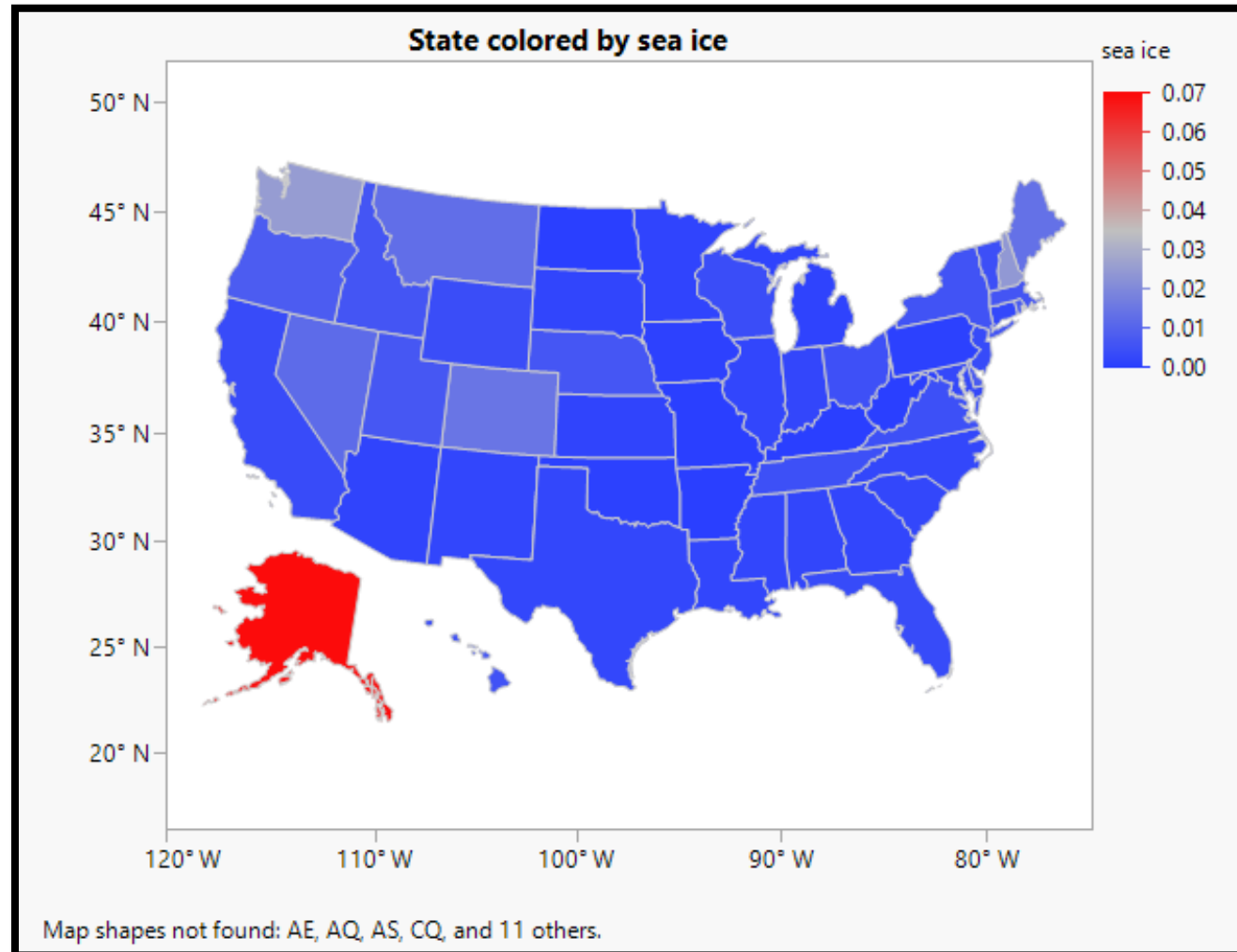
NSF Abstracts

- While some phrases are newly minted over time.



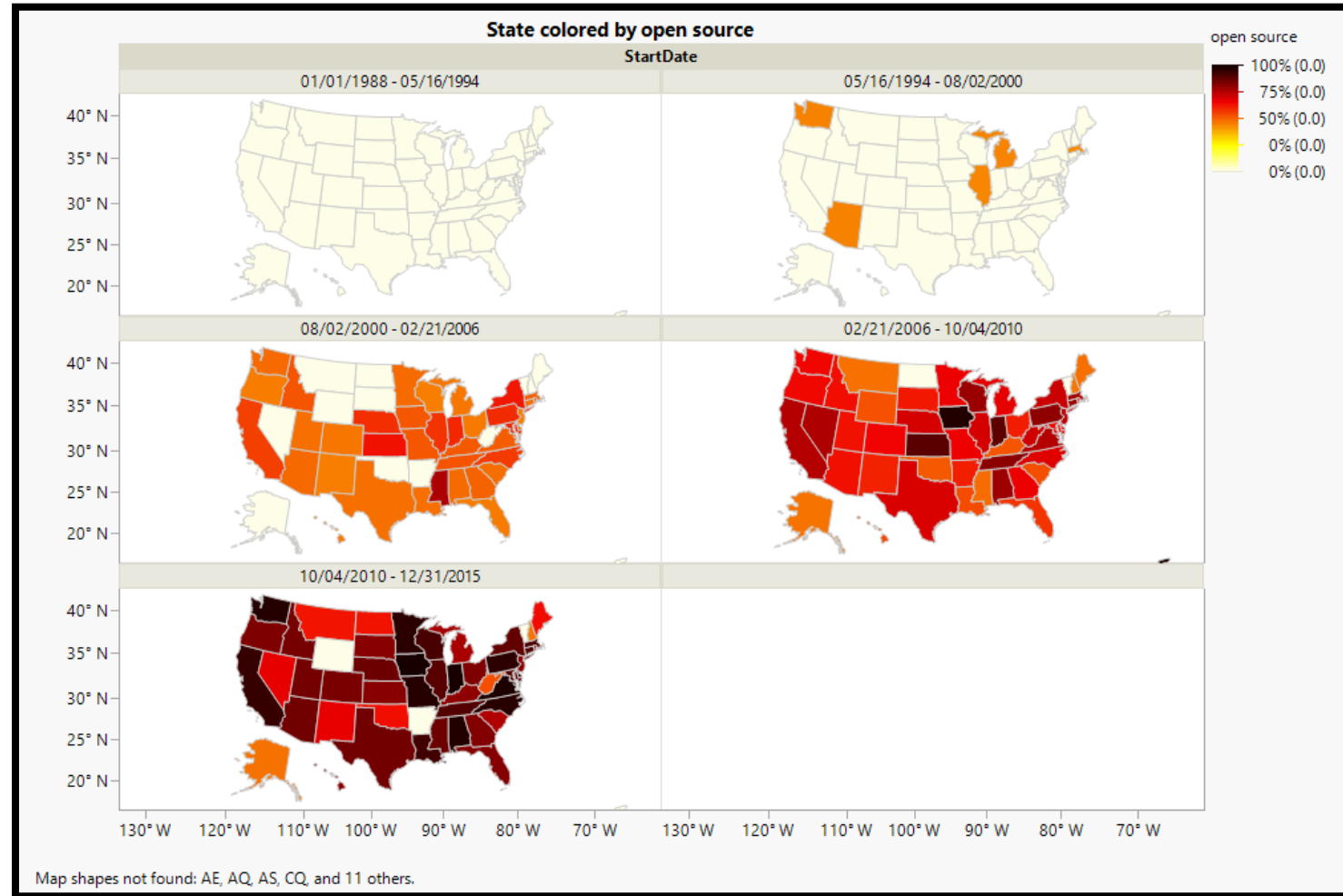
NSF Abstracts

- Certain phrases are highly associated with certain states.



NSF Abstracts

- Visually display the frequency of phrases over time by state.



- Evaluate if the Amount Awarded is associated with a certain phrase.

Response AwardedAmountToDate

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Prob > F
Error	304674	9.8863e+18	3.245e+13	
C. Total	304724	9.8915e+18		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	400317.3	12167.22	32.90	<.0001*
sea ice	1427033.2	153718.1	9.28	<.0001*
high performance	291956.53	71206.05	4.10	<.0001*
differential equations	-234832.6	82463.2	-2.85	0.0044*
graduate students	90212.107	34836.3	2.59	0.0096*
dark matter	373844.71	186260.4	2.01	0.0447*
climate change	126987.51	64661.26	1.96	0.0495*
national security	270166.94	172367	1.57	0.1170
data mining	229988.36	148201.6	1.55	0.1207
magnetic resonance	184363.2	119903.6	1.54	0.1241
open source	162765.21	110676.2	1.47	0.1414
number theory	-176996.8	123009.2	-1.44	0.1502
algebraic geometry	-171012.4	124658.3	-1.37	0.1701
black holes	256320.16	189328.2	1.35	0.1758

These 9 videos cover predictive analytics (including text exploration) and data visualization.

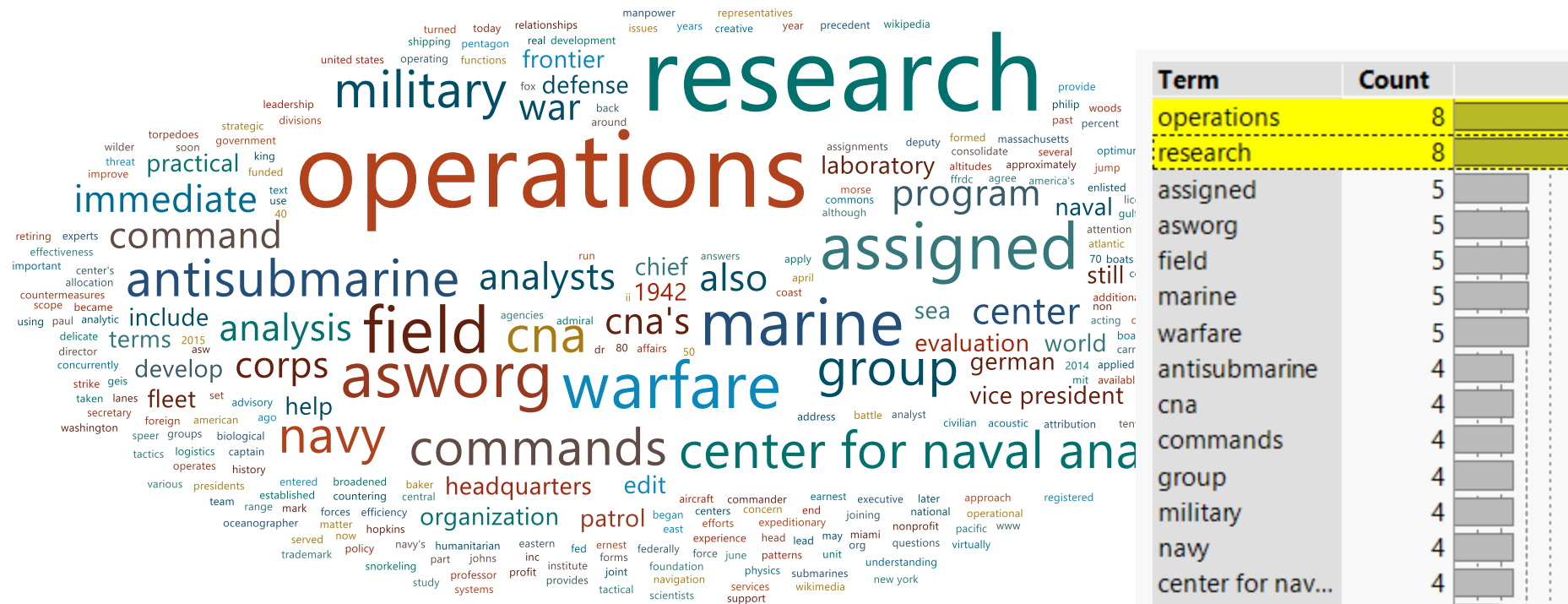
Building Better Models Overview and Use of Honest Assessment	Neural Networks - Single Layer, Dual Layer, Boosted	All Graphs are Wrong - Some are Useful - Or view Xan Gregg's Original 2015 Discovery Summit Presentation
Regression Linear, Stepwise, Logistic, & All Possible	Generalized Regression Near Machine Learning Accuracy – More Explainable Model	What's New in JMP 14? JMP Learning Resources
Decision Trees Simple Partition, Bootstrap Forest, & Boosted Tree	Text Exploration Analyze Unstructured Free Text	Functional Data Explorer Modeling a “Stream” of Data – New in JMP 14

Summary

- Data is growing exponentially across DOD
- Much of this is unstructured text data
- Text mining takes statistical tools and converts the text into meaningful mathematical expressions
- Example uses of text mining
 - Concept extraction
 - Grouping like documents or records together
 - Grouping terms together
 - Creating structured variables that represent the text fields to use in predictive analytics
- Relatively short learning curve to perform powerful text analytics using open source software and commercial solutions

CNA Example

- Unstructured text
- Can you get idea what's going on from word frequencies?
- Do word clouds help?

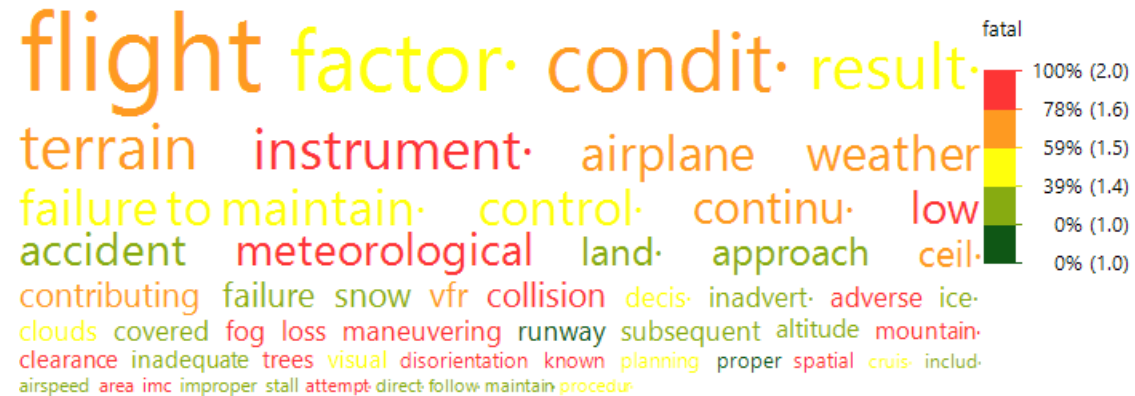


Word Cloud Colored by Proportion Fatal and Filtered by Meteorological Conditions

Visual
Meteorological
Conditions

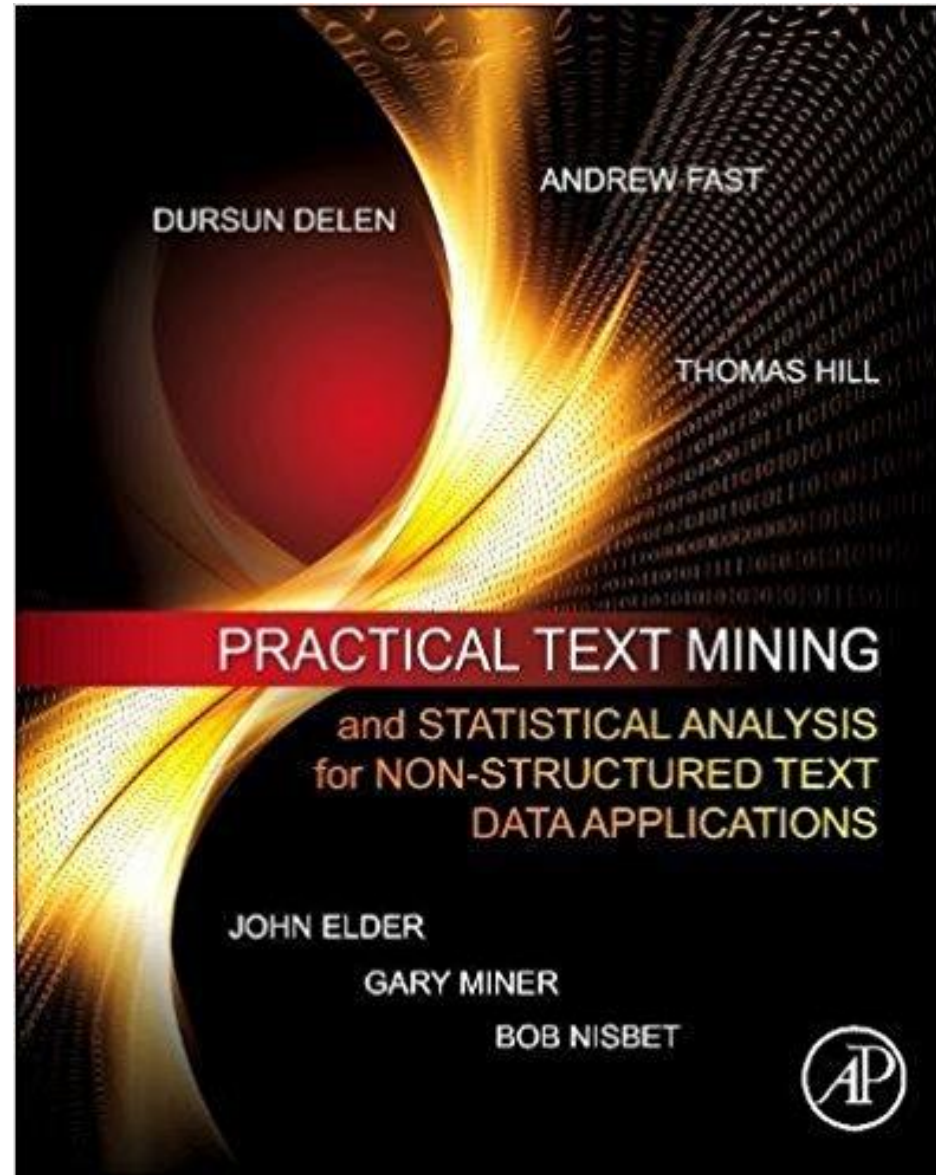


Instrument
Meteorological
Conditions



What is Text Mining?

- Text mining: semi-automated process of detecting patterns (useful information and knowledge) from large amounts of *unstructured* data sources
- Text analysis: an examination of structure, composition, and meaning that provides insight to advance some purpose...that characterize and describe a text itself. Analysis may be heuristic, informal, and/or qualitative.
- Text analytics: methods used for intelligent analyses of textual data; a larger set of activities around inference steps of discovering information, grouping documents, summarizing information, etc.
 - Systematic application of numerical and statistical methods that derive and deliver quantitative information, whether in the form of indicators, tables, or visualizations. Analytics is formal and repeatable.



A practical guide to text mining with topic extraction

Andrew Karl, James Wisnowski* and W. Heath Rushing



Text analytics continue to proliferate as mass volumes of unstructured but highly useful data are generated at unbounded rates. Vector space models for text data—in which documents are represented by rows and words by columns—provide a translation of this unstructured data into a format that may be analyzed with statistical and machine learning techniques. This approach gives excellent results in revealing common themes, clustering documents, clustering words, and in translating unstructured text fields (such as an open-ended survey response) to usable input variables for predictive modeling. After discussing the collection and processing of text, we explore properties and transformations of the document-term matrix (DTM). We show how the singular value decomposition may be used to drastically reduce the size of the document space while also setting the stage for automatic topic extraction, courtesy of the varimax rotation. This latent semantic analysis (LSA) approach produces factors that are compatible with graphical exploration and advanced analytics. We also explore Latent Dirichlet Allocation for topic analysis. We reference published R packages to implement the methods and conclude with a summary of other popular open-source and commercial software packages.

© 2015 Wiley Periodicals, Inc.

<http://onlinelibrary.wiley.com/doi/10.1002/wics.1361/abstract>

A practical guide to text mining with topic extraction

By: Andrew Karl, James Wisnowski and W. Heath Rushing

<http://onlinelibrary.wiley.com/doi/10.1002/wics.1361/abstract>

REFERENCES

1. Miner G, Elder J, Hill T, Nisbet R, Dursun D, Fast A. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Oxford: Academic Press; 2012.
2. Chakraborty G, Pagolu M, Garla S. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. Cary, NC: SAS Institute Inc; 2013.
3. Ingersoll G, Morton T, Farris A. *Taming Text*. Shelter Island, NY: Manning; 2013.
4. Grun B, Hornik K. Topic models: an R package for fitting topic models. *J Stat Softw* 2011, 40:1–30.
5. The Apache Software Foundation. OpenNLP. 2015. Available at: <https://opennlp.apache.org/> (Accessed March 24, 2015).
6. Lichman M. UCI machine learning repository. University of California Irving, School of Information and Computer Science. 2013. Available at: <http://archive.ics.uci.edu/ml> (Accessed February 13, 2015).
7. Scrapinghub S. A fast and powerful web crawling framework. 2015. Available at: <http://scrapy.org/> (Accessed March 10, 2015).
8. Temple Lang D. RCurl: general network (HTTP/FTP/...) client interface for R. CRAN. 2015. Available at: <http://cran.r-project.org/web/packages/RCurl/index.html> (Accessed March 10, 2015).
9. Barbera P. streamR: access to Twitter streaming API via R. 2014. Available at: <http://CRAN.R-project.org/package=streamR> (Accessed March 10, 2015).
10. Bilisoly R. *Practical Text Mining with Perl*. Hoboken, NJ: John Wiley & Sons; 2008.
11. Porter MF. An algorithm for suffix stripping. *Program* 1980, 14:130–137.
12. Yang C. Who's afraid of George Kingsley Zipf? Or: do children and chimps have language? *Significance* 2013, 10:29–34.
13. Bates D, Maechler, M. Matrix: sparse and dense matrix classes and methods. CRAN. 2015. Available at:

- Heath Rushing's *Technically Speaking JMP*[®]:
Tackling Unstructured Data With Text Exploration
https://www.jmp.com/en_us/events/ondemand/technically-speaking/tackling-unstructured-data-with-text-exploration.html
- Brady Brady's *Advanced Mastering JMP*[®]:
Retrieving and Organizing Text for Analysis
https://www.jmp.com/en_us/events/ondemand/mastering-jmp/text-explorer.html

Text Analytics: Trending Upward

- In 2015, SAS text mined 7,000 data scientist job descriptions for top skills required
- Text analytics is fastest growing area
- **~80% of data* is unstructured and is waiting to be analyzed**

Top 10 Types of Analysis Mentioned	
ML_NEURAL_NETS_SVM	52%
OPTIMIZATION	23%
TEXT_ANALYTICS	17%
TIME_SERIES	17%
DATA_WRANGLING	16%
CLUSTERING	16%
LINEAR_REGRESSION	15%
DATA_VISUALIZATION	13%
MATLAB	12%
DESIGN_OF_EXPERIMENTS_AB	10%

n=7027

<http://blogs.sas.com/content/text-mining/page/2/>

* 83.92% of all statistics are made up. (Prof. Dick DeVeaux, Williams College)

Social Media Analytics—Twitter

Sentiment Analysis

- Real-time feed of social media provides intelligence opportunity
- Example: Trump
 - 2 minutes of all Tweets at 6:30AM on 19 June 2017
- Sentiment analysis/opinion with text mining tabulates the number of positive terms and number of negative terms (Harvard IV dictionary) from all Tweets



	Negative	Positive
1132	liable	pleasantry
1133	liar	please
1134	lie	pleased
1135	lifeless	pleasurable
1136	limit	pleasure
1137	limitation	pledge
1138	limp	plentiful
1139	liquidate	plenty
1140	liquidation	poetic
1141	litter	poignant
1142	load	poise
1143	lone	polish
1144	loneliness	polite
1145	lonely	politeness
1146	loner	pomp
1147	lonesome	popular
1148	loom	popularity
1149	lose	populous
1150	loser	portable

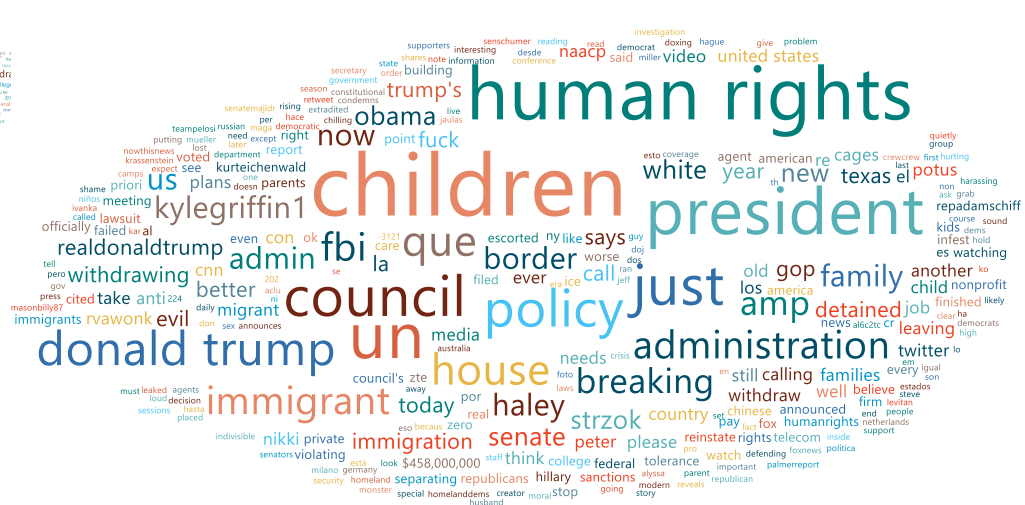
	Positive	Negative
	Sum	Sum
	801	833

Social Media Analytics—Twitter Sentiment Analysis

- Real-time feed of social media provides intelligence opportunity
- Example: Trump
 - 2 minutes of all Tweets at 3:30 PM PDT on 19 June 2018
- Sentiment analysis/opinion without sentiment analysis

trump

Add “Trump” to
Stop Word List

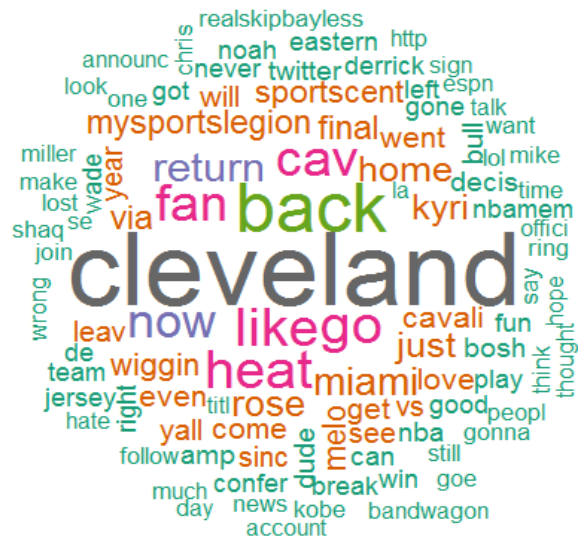


Term	Count
trump	91
children	20
human rights	17
president	15
un	15

Social Media Analytics—Twitter

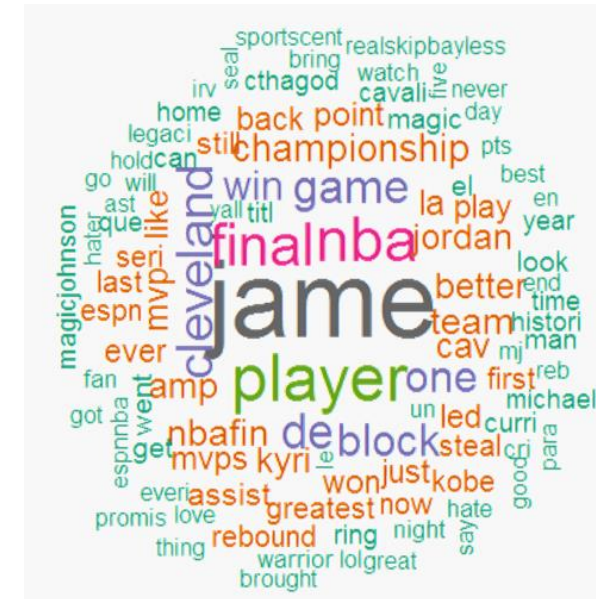
Sentiment Analysis

- Real-time feed of social media provides intelligence opportunity
- Example: LeBron James after announcement he is going home
 - All Tweets for 5 mins the day LeBron made his statement
 - All Tweets for 5 mins the day he won the NBA Championship



Positive	Sum	3533
Negative	Sum	1626

	Negative	Positive
1132	liable	pleasantry
1133	liar	please
1134	lie	pleased
1135	lifeless	pleasurable
1136	limit	pleasure
1137	limitation	pledge
1138	limp	plentiful
1139	liquidate	plenty
1140	liquidation	poetic
1141	litter	poignant
1142	load	poise
1143	lone	polish
1144	loneliness	polite
1145	lonely	politeness
1146	loner	pomp
1147	lonesome	popular
1148	loom	popularity
1149	lose	populous
1150	loser	portable

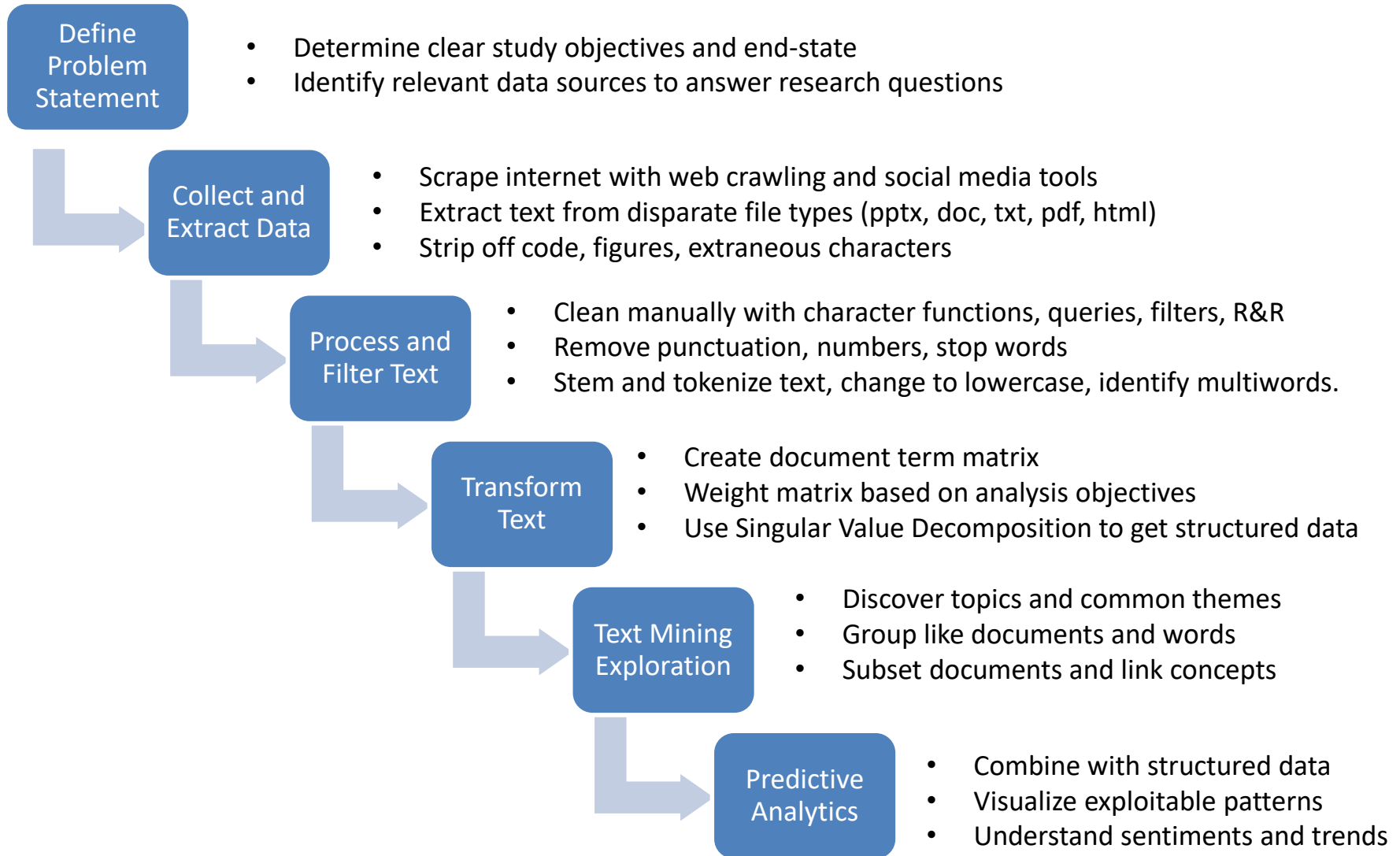


Positive	Negative
Sum	Sum
2722	1257

Some DOD Applications of Text Mining

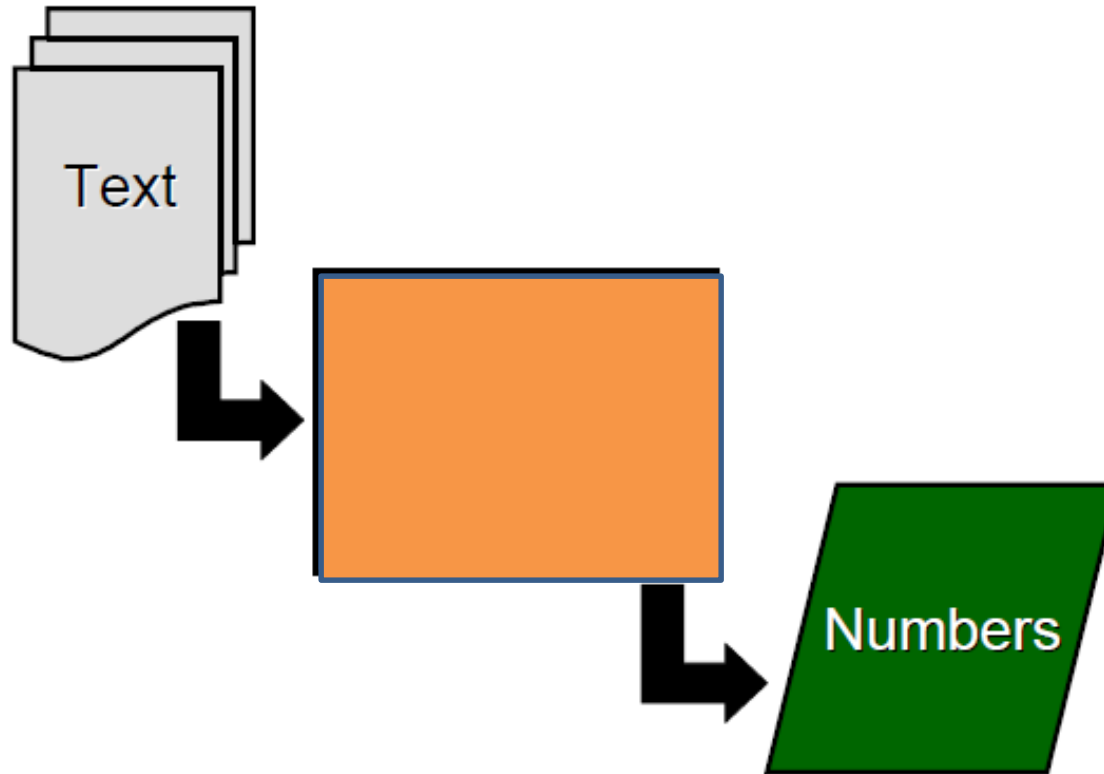
- Performance report analysis—powerful words and phrases, grouping like reports, combining with structured data
- Analysis of text fields in system maintenance reports or software trouble reports
- Using mission report text fields to help predict IED locations
- **DHS/ECBC monitoring [www](#) for discussions of chemical/biological agents**
- Boosting survey insights with deeper analysis of free text responses
- Contract performance reports
- Summarizing volumes of MIL-STDS; searching for specific related topics
- Finding patterns in voice-to-text translations of communications during operations
- Evaluating sentiment about quality of life from social media posts
- Sentiment analysis applied to intelligence; pattern-of-life
- **Electronic health record analytics**
- R&D: what are the most common themes in the abstracts at reliability conference? What are topics from a large group of technical journal articles?

Text Analytics Flow

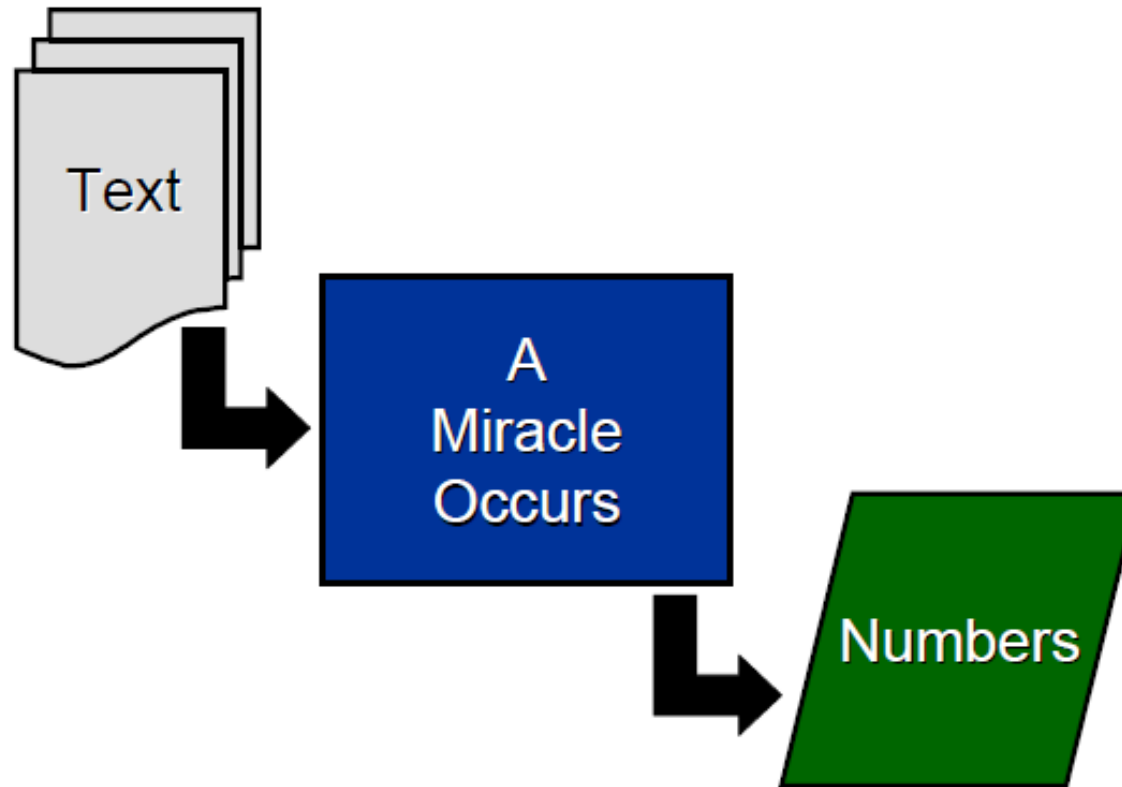


ORSA: How Exactly Does TM Work?

Another View of Text Mining



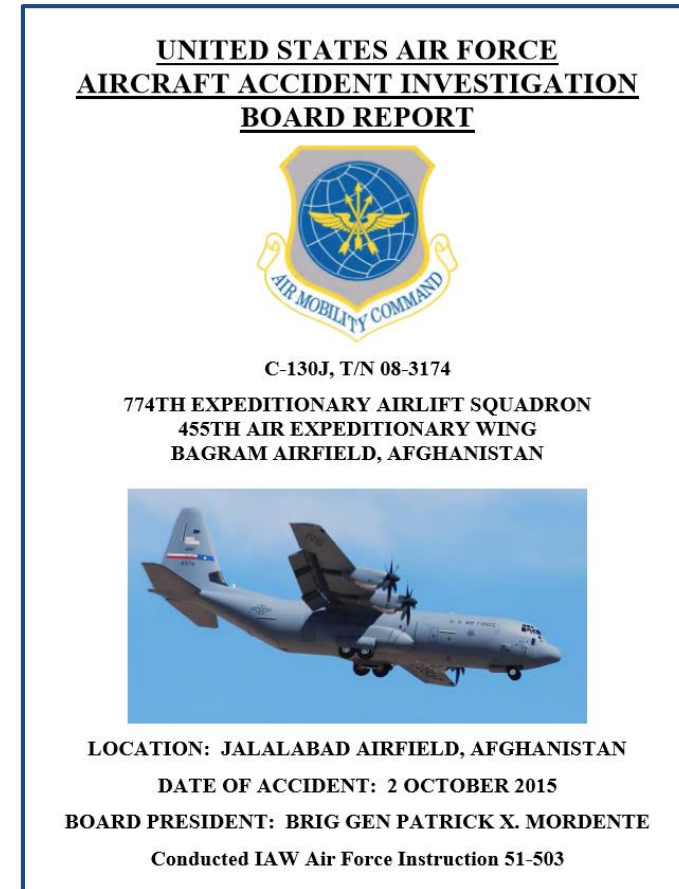
Another View of Text Mining



Unclassified Text Mining Example

Aircraft Accident Reports

- Data: Air Force Accident Investigation Board Reports
- Objectives:
 - what are common themes?
 - what factors contribute to fatal accidents?
 - what words group together?
 - what reports group together?
 - how can we link structured and unstructured data fields?



Let's First Revisit the ORSA Question

Key Quantity: Document Term Matrix

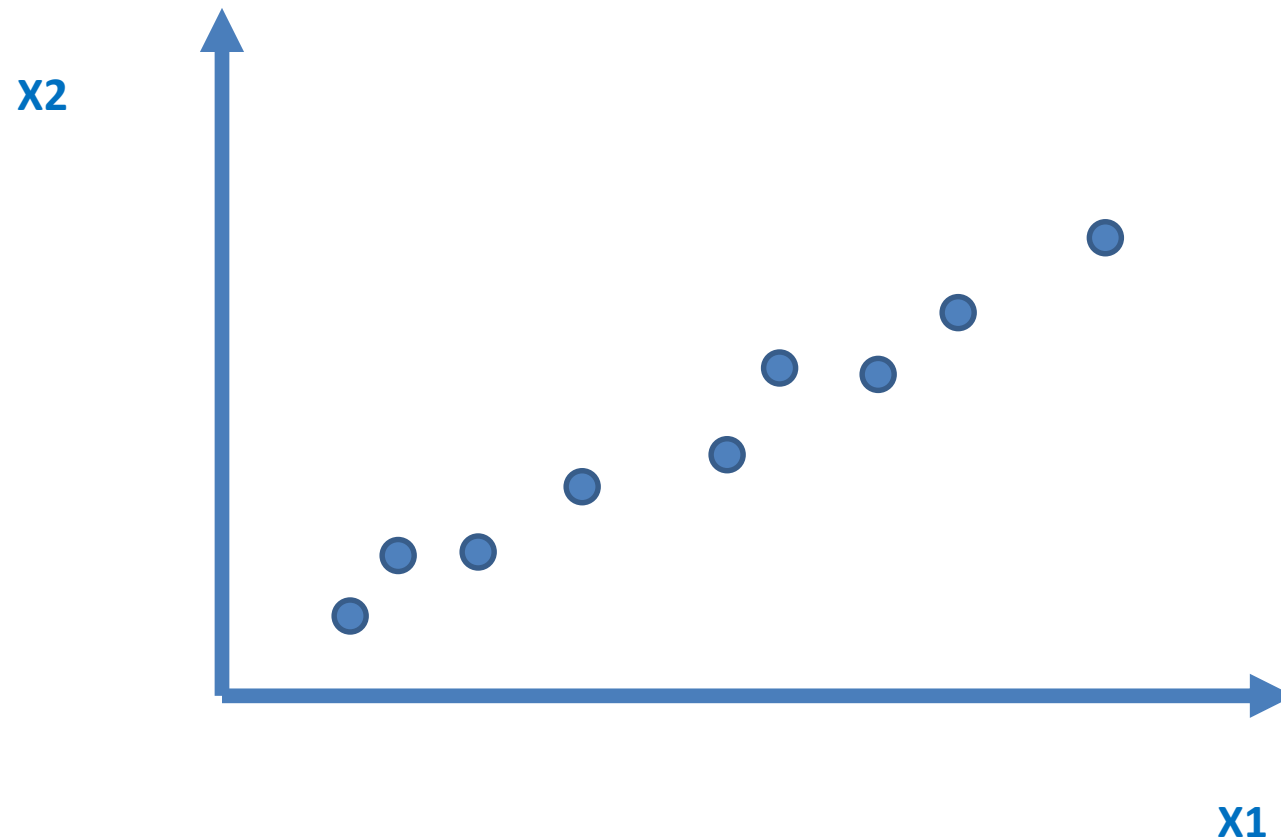
- DTM is a sparse matrix with documents as rows and terms as columns=> 3,200 rows (accident reports) by 800 cols (words)
- Tallies the word counts for each document
 - Mostly 0's
 - Can use binary, term frequency, inverse document frequency, and other weighting schemes
- DTM itself is helpful (i.e could do correlation analysis on columns), but need to take it a step further with Latent Semantic Analysis

narr_cause	adequ	adjac	adjust	advers	advisori
the pilot's failure to maintain directional control. Factors were the crosswind, th...	0	1	0	0	0
The flight instructor's failure to ensure (supervision) the student had an adequat...	1	0	0	0	0
The pilot's inflight decision to continued visual flight into instrument meteorolo...	0	0	0	0	0
the loss of power to both engines for undetermined reasons during approach. ...	0	0	0	0	0
The pilots decision not to fly to the alternate airport, his decision to continue th...	0	0	0	1	0

- The reduced-rank singular value decomposition (SVD) provides us with a dimensionality reduction technique.
- The SVD reduces the DTM to a (dense) matrix with fewer columns. The new (orthogonal) columns are linear combinations of the rows in the original DTM, selected to preserve as much of the structure of the original DTM as possible.

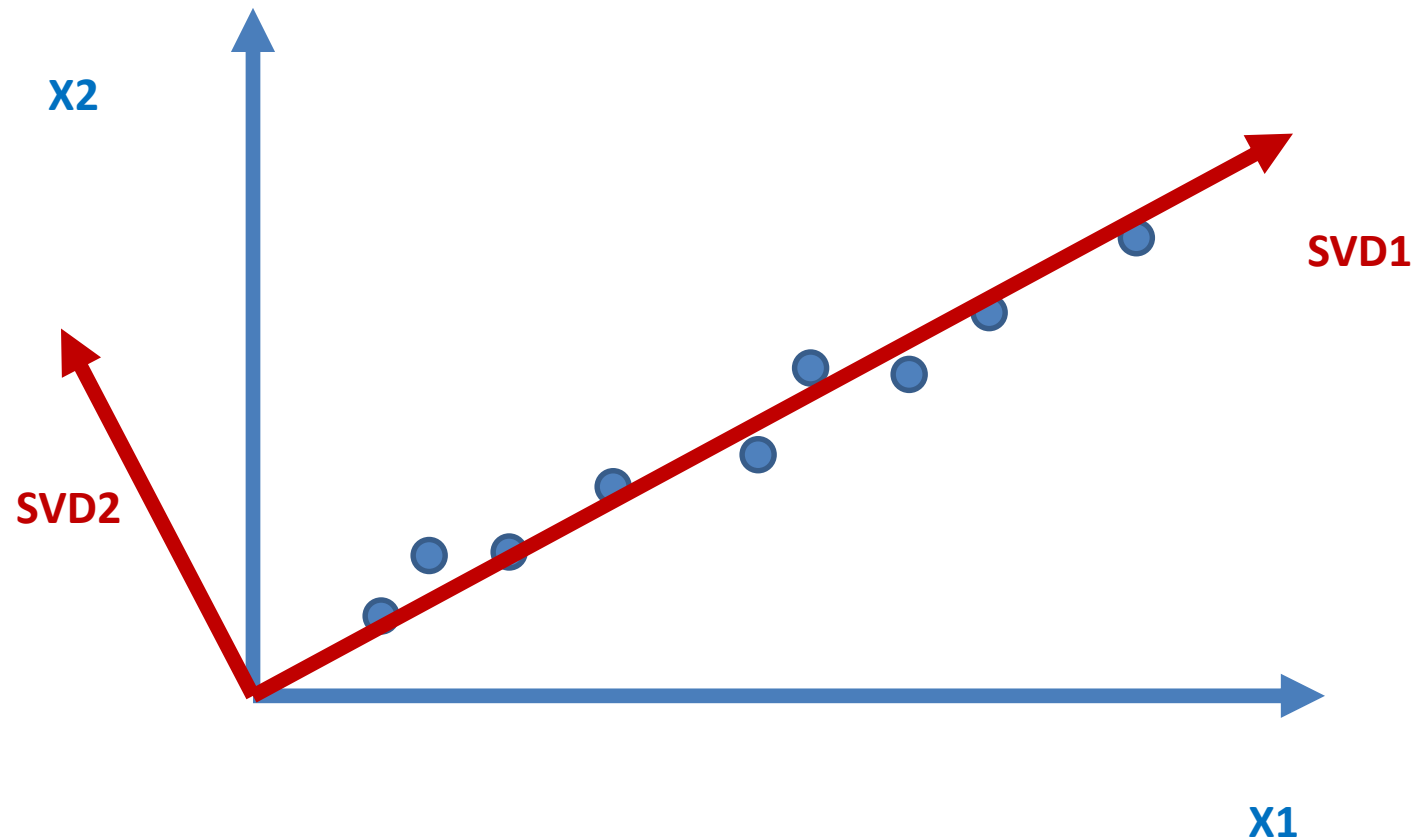
SVD Example

X1 and X2 describe the location of these points.
However, they appear to fall mostly along a line.



SVD Example

Roughly, the SVD finds a new set of orthogonal basis vectors such that each additional dimension accounts for as much of the variation of the data as possible.



Singular Value Decomposition

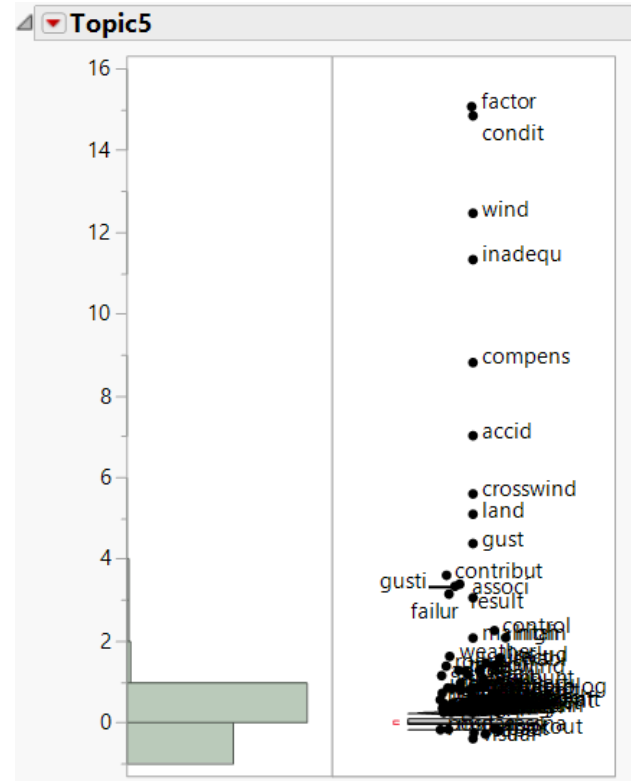
- For a DTM X , the SVD factorization is

$$X \approx UDV^t,$$

where

- U is a dense d by s orthogonal matrix **U gives us a new rank-reduced description of documents**
- D is a diagonal matrix with nonnegative entries (the singular values).
- V^t is a dense s by w orthogonal matrix, where s is the rank of the SVD factorization ($s=1, \dots, \min(d, w)$), and the superscript t indicates “transpose.” **V gives us a new rank-reduced description of terms.**
- d is the number of documents
- w is the number of words
- s is the rank of the SVD factorization ($s=1, \dots, \min(d, w)$).

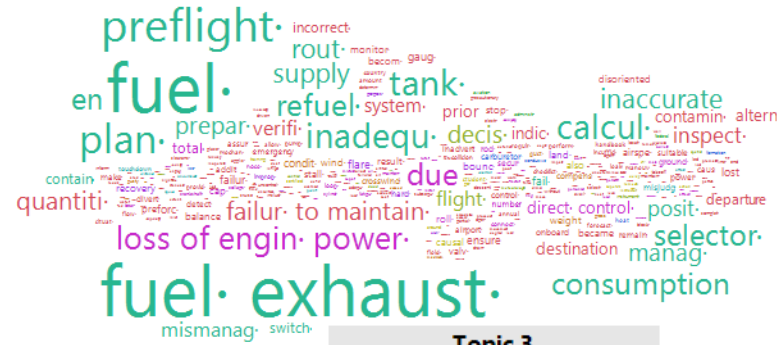
Topic Extraction With SVD V Matrix



- The 5th eigenvector from the V matrix loads on inadequate compensation for crosswinds

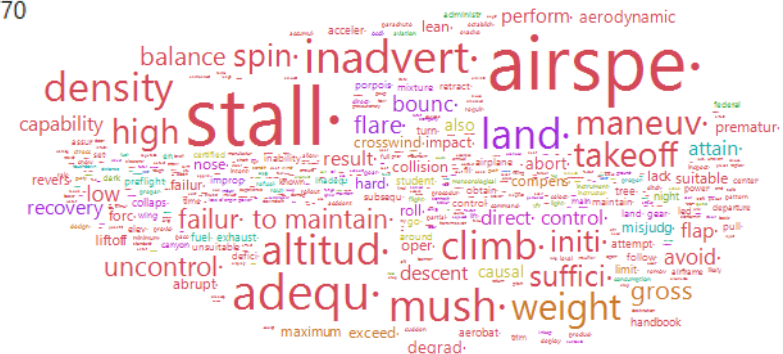
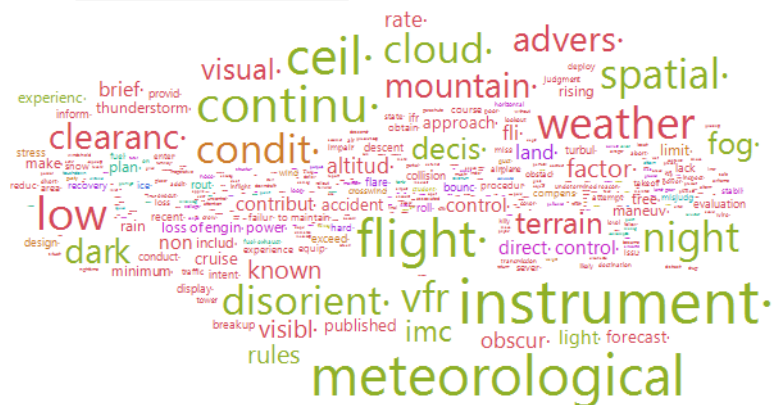
Topic Term Loading, Word Cloud Filtered by Topic, and Colored by Topic Cluster

Topic 1	
Term	Loading
instrument	0.60037
meteorological	0.53265
ceil	0.51157
flight	0.50531
continu	0.48495
low	0.44785
weather	0.43053
vfr	0.41082
night	0.40613
spatial	0.40265
disorient	0.40086
condit	0.39957
cloud	0.39942
advers	0.35134
clearanc	0.34990



Topic 3	
Term	Loading
fuel·exhaust	0.71220
fuel	0.69099
preflight	0.47071
plan	0.42553
tank	0.37138
inadequ	0.34722
calcul	0.34694
loss of engine·power	0.34295
consumption	0.32196
due	0.31965
refuel	0.31666
en	0.29369
selector	0.29298
inaccurate	0.28770

Topic 5	
Term	Loading
stall	0.6717
airspe	0.6344
adequ	0.4250
inadvert	0.3899
mush	0.3747
altitud	0.3658
land	-0.3565
density	0.3551
climb	0.3297
weight	0.3148
maneu	0.3061
high	0.2923
takeoff	0.2891
spin	0.2730

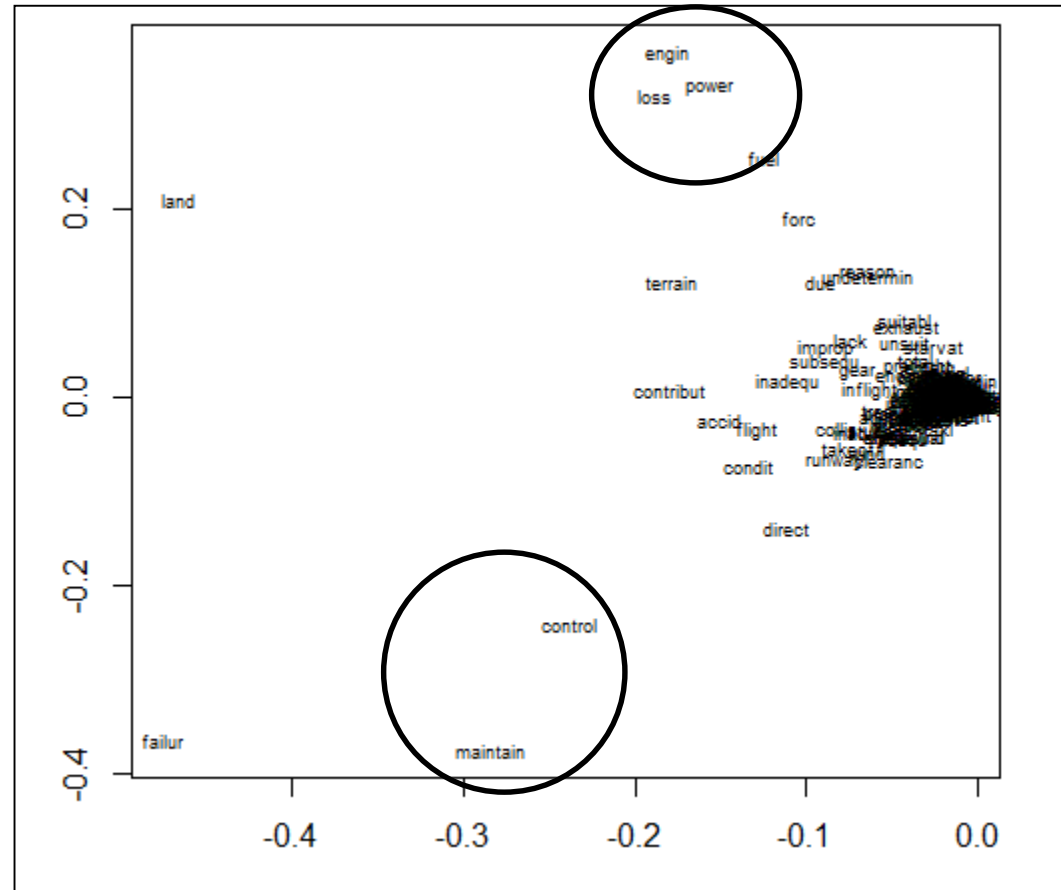


Finding Documents Related to Topic 5: SVD U Matrix

	narr_cause	SVD5
1	The pilot's inadequate compensation for gusty wind conditions. Factors associated with the accident were the pilot's inadequate ...	2.5480873522
2	The pilot's inadequate compensation for wind conditions during the landing roll, which resulted in an inadvertent ground loop/sw...	2.5178671713
3	The pilot's inadequate compensation for the winds conditions which resulted in the failure to maintain directional control of the ai...	2.4973844764
4	The pilot's inadequate compensation for the wind conditions that resulted in directional control not being maintained during the l...	2.4872052169
5	The pilot's inadequate compensation for wind and his failure to maintain directional control of the airplane which resulted in a gro...	2.482312475
6	The pilot's inadequate compensation for wind conditions during the landing roll, resulting in a nose over. A factor associated wit...	2.4742882397
7	The pilot's inadequate compensation for the wind conditions which resulted in the failure to maintain directional control of the air...	2.4740669529
8	The pilot's inadequate compensation for wind conditions during the takeoff roll. A factor associated with the accident was a cross...	2.4421888586
9	The pilot's inadequate compensation for wind conditions while on approach, and the inadvertent stall of the airplane while attem...	2.4388091067
10	The pilot failed to maintain directional control of the airplane due to inadequate compensation for the wind conditions. Factors a...	2.4363088679
11	The pilot's inadequate compensation for wind conditions. Factors associated in the accident were a crosswind, and a worn tailwhe...	2.4288790043
12	The pilot's inadequate compensation for wind conditions. A factor associated with the accident was a crosswind.	2.4158701461
13	the inadequate rotation speed and compensation for wind conditions by the pilot. Contributing factors were the crosswind and g...	2.3758763124
14	The pilot's inadequate compensation for wind conditions during the takeoff run, which resulted in a loss of control and subsequen...	2.3750758348
15	the pilot's inadequate compensation for wind conditions. Factors were the crosswind and the gusts.	2.3677401171
16	The pilot displayed inadequate compensation for the wind conditions that existed at the time of the accident and directional contr...	2.3549840887
17	the pilot's inadequate compensation for the wind conditions which resulted in a loss of control during landing. A contributing fac...	2.3503264913
18	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control during an aborted landing...	2.3333072329
19	The pilot's inadequate compensation for wind conditions which resulted in an in-flight collision with trees. A factor related to the ...	2.3188632769
20	Inadequate preflight planning and inadequate compensation for the wind conditions which resulted in a failure to maintain directi...	2.3182120659
21	The pilot's inadequate compensation for wind conditions while landing. Factors associated with the accident were the pilot's inad...	2.3039591872
22	The pilot's inadequate compensation for wind conditions during initial climb, which resulted in an in-flight collision with trees. A f...	2.2971297018
23	The pilot's inadequate compensation for wind conditions during takeoff, which resulted in an in-flight collision with trees. A facto...	2.2900786845
24	The student pilots inadequate compensation for wind conditions which resulted in an off-field landing, and the CFI's improper dec...	2.2833687328
25	The pilot's inadequate compensation for wind conditions during takeoff. Factors associated with the accident were trees and a va...	2.2826089858
26	The pilot's inadequate compensation for wind conditions during takeoff. Factors associated with the accident were variable winds...	2.2826089858
27	The pilot's inadequate compensation for wind conditions, resulting in an inadvertent ground loop. Factors include wind gusts duri...	2.2729439036
28	The pilot's failure to adequately compensate for wind conditions after encountering a crosswind gust during the landing roll. Fact...	2.2675327857
29	The pilot's inadequate compensation for wind conditions, and the excessive use of the airplane's brakes, which resulted in a nose ...	2.2661909279
30	The pilot's inadequate compensation for wind conditions. Factors associated with the accident are variable winds, and a downdraft.	2.2660439108
31	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control. Contributing factors were...	2.2648282967
32	The pilot's inadequate compensation for wind conditions during takeoff. A factor associated with the accident was a variable wind.	2.2579465664
33	The pilot's inadequate compensation for wind conditions during the landing roll. A crosswind was a factor.	2.2515447851
34	The pilot's inadvertent stall of the airplane during takeoff. Factors associated with the accident were the pilot's inadequate weath...	2.250859567
35	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control during the landing roll. Fac...	2.2497623998
36	the student pilot's inadequate compensation for the crosswind during landing roll. A contributing factor was the cross wind weat...	2.2386981287
37	The pilot's inadequate compensation for wind conditions. A factor associated with the accident was a sudden wind shift.	2.2340296545
38	the pilot's inadequate compensation for the gusting and shifting wind conditions, which resulted in a failure to maintain direction...	2.2275070311

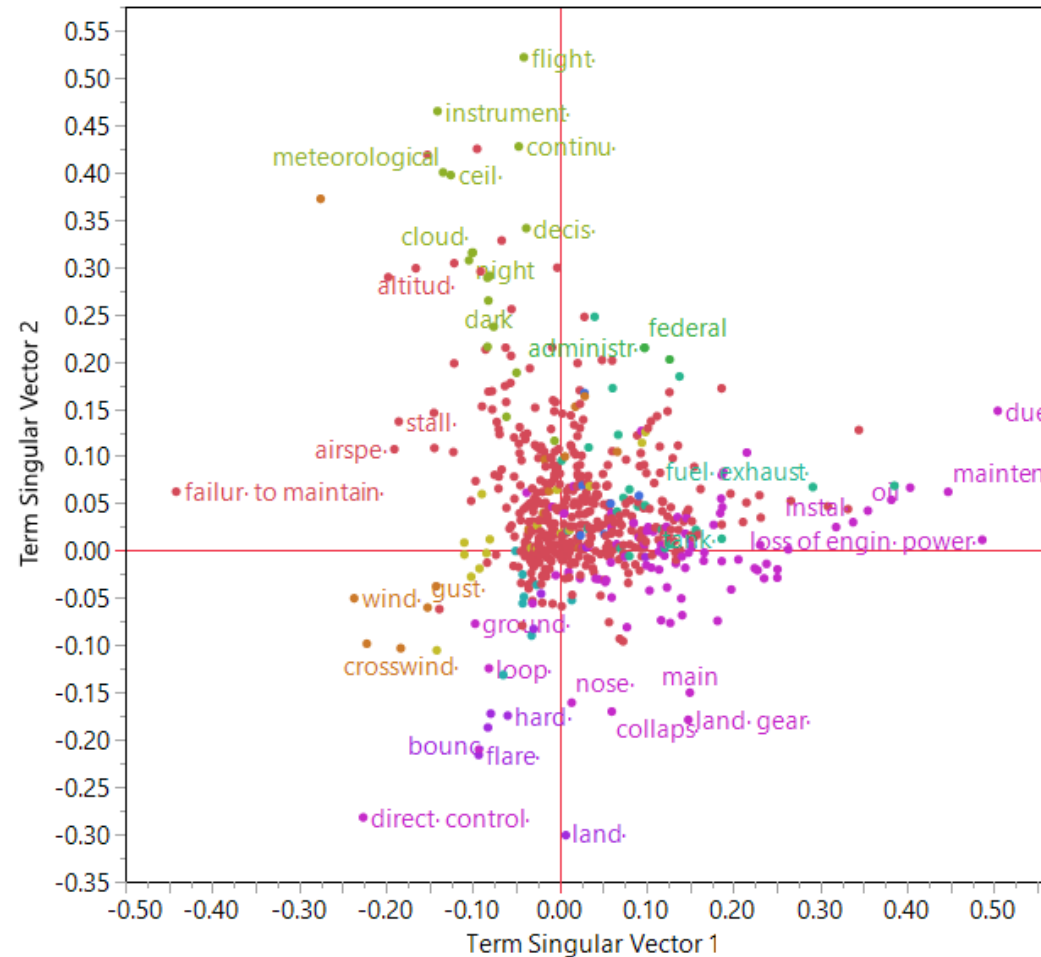
- If we sort the corresponding U matrix on SVD 5 descending, we see the reports with inadequate wind compensation rise to the top

SVD1 vs. SVD2



- Plotting first two eigenvectors is often helpful and a recommended first step

Term SVD1 vs. SVD2



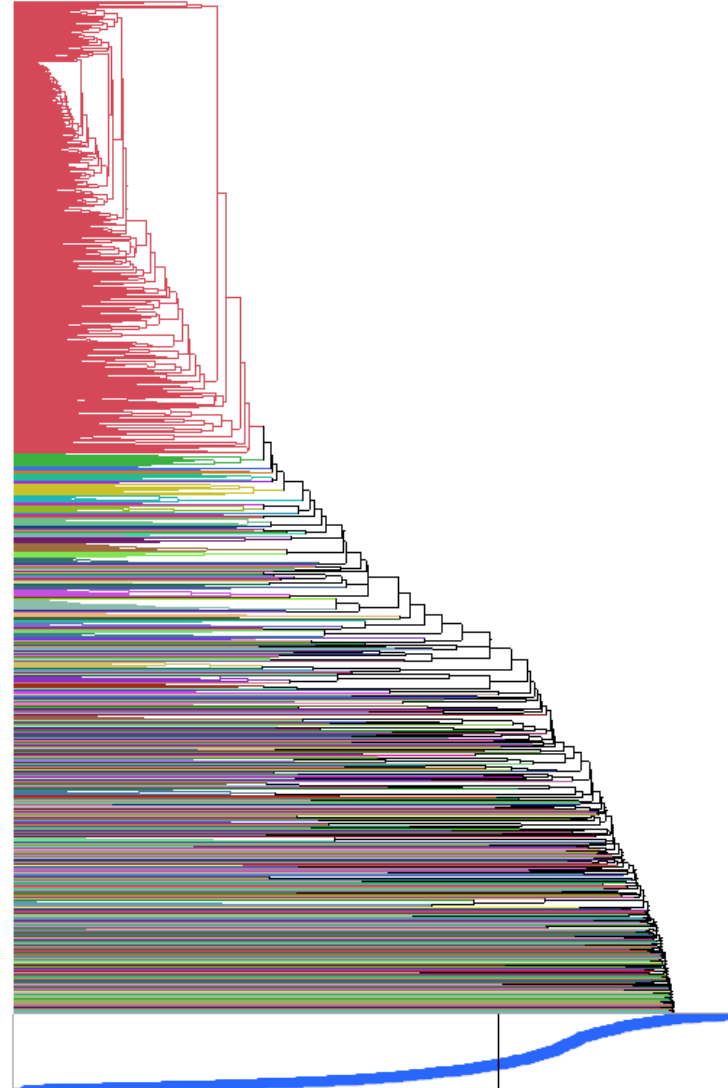
- Plotting first two eigenvectors is often helpful and a recommended first step

Clustering Terms

Hierarchical Clustering

Method = Ward

Dendrogram



- Often, there will be a large cluster (seen at right) of unimportant terms

Clustering Terms

- Here we can see with a large number of clusters, words typically associated with one another are in the same cluster
- The distance matrix allows us to see the closest terms to any specified word

		Label	Cluster
•	1	reason	441
•	2	undetermin	441
•	3	attent	332
•	4	divert	332
•	5	disorient	311
•	6	spatial	311
•	7	pattern	305
•	8	traffic	305
•	9	defici	284
•	10	known	284
•	11	rough	275
•	12	uneven	275
•	13	pole	262
•	14	util	262
•	15	hing	233
•	16	spring	233
•	17	stud	233
•	18	tab	233
•	19	tension	233
•	20	worn	233

Clustering Accident Reports

- It is possible to cluster the eigenvectors of the U matrix to group accident reports that have common themes
- We can cross-tabulate the clusters with the proportion that were fatal
- These clusters were had about 35 records each and all were non-fatal

Cluster	N(NO)	Row %(NO)	N(YES)	Row %(YES)
188	38	100.00%	0	0.00%
169	35	100.00%	0	0.00%
189	33	100.00%	0	0.00%
236	31	100.00%	0	0.00%
193	29	100.00%	0	0.00%
176	26	100.00%	0	0.00%
180	26	100.00%	0	0.00%
192	26	100.00%	0	0.00%
60	25	100.00%	0	0.00%

Bounced Landings: Not Fatal

Cluster	fatal	narr_cause
188	NO	The loss of control on landing due to the student's improper recovery from a bounced landing, and the resulting nose over on the grass runway.
188	NO	The student's failure to maintain control of the aircraft during landing due to his improper landing flare height and improper recovery from a bounced landing.
188	NO	the student pilot's failure to recover from a bounced landing, which resulted in porpoising and subsequently a nose over.
188	NO	The pilot's premature flare, which resulted in an inadvertent stall and a bounced landing. A factor was the improper recovery from a bounced landing.
188	NO	the student pilot's improper recovery from a bounced landing.
188	NO	The pilot's improper flare, and improper recovery from a bounced landing.
188	NO	The pilot's improper flare and his improper recovery from a bounced landing.
188	NO	The pilot's improper recovery from a bounced landing.
188	NO	The student pilot's failure to maintain aircraft control during the landing, her failure to recover from the bounced landing, and the nose gear overload.
188	NO	The student pilots improper flare, and improper recovery from a bounced landing. A factor was the student pilot's lack of total experience.
188	NO	The pilot's inadequate recovery from a bounced landing. A factor associated with the accident was a crosswind.
188	NO	The pilot's improper recovery from a bounced landing.
188	NO	An inoperative airspeed indicator and the pilot's improper recovery from the bounced landing.
188	NO	The pilot's improper flare, and improper recovery from a bounced landing.
188	NO	The student pilot's improper flare and recovery from a bounced landing.
188	NO	The pilot's inadequate recovery from a bounced landing which resulted in a hard contact with the runway. Factors associated with the accident were t
188	NO	The pilot's improper recovery from a bounced landing. A factor in the accident was the pilot improper flare.
188	NO	The student pilot's improper flare and failure to recover from a bounced landing resulting in the subsequent collapse of the nose gear during the landing
188	NO	The pilot's improper recovery from a bounced landing. A factor was the pilot's failure to flare during initial touchdown.
188	NO	The pilot's misjudgment of distance, his subsequent improper recovery from a bounced landing, and the failure to maintain airspeed which resulted in
188	NO	the student pilot's failure to properly recover from a bounced landing which resulted in the airplane porpoising. A contributing factor was the student pi

Soft Terrain: Not Fatal

Cluster	fatal	narr_cause
169	NO	The pilot's failure to maintain a proper glidepath during final approach. A factor associated with the accident was soft terrain.
169	NO	The pilot's failure to maintain directional control during the landing roll. Factors were the crosswind and soft terrain condition.
169	NO	The pilot's inadequate preflight planning/preparation, and his selection of unsuitable terrain for landing. A factor in the accident was snow-covered terrain.
169	NO	The pilot's selection of unsuitable terrain for takeoff, and his inadequate preflight planning/preparation resulting in a collision with trees during the initial climb.
169	NO	The pilot's inadvertent stall while maneuvering. A factor associated with the accident was soft, snow-covered terrain.
169	NO	The pilot's selection of unsuitable terrain for landing and subsequent nose over during the landing flare. Factors in the accident were soft, snow-covered terrain.
169	NO	the pilot's failure to maintain directional control during the takeoff initial climb. Contributory factors were the pilot's lack of experience with the aircraft and the soft terrain.
169	NO	the rocker assembly failure during low level maneuvering. Factors were the soft and sandy terrain and the unsuitable terrain the pilot encountered during the landing.
169	NO	A soft area in the turf runway, which resulted in a loss of directional control during the landing rollout.
169	NO	The pilot's selection of unsuitable terrain for landing. Factors in the accident were a soft area of runway, and sun glare.
169	NO	The pilot's selection of unsuitable terrain for takeoff. Factors in the accident were soft terrain, and the pilot's delay in aborting the takeoff.
169	NO	The pilot's selection of an unsuitable landing area. A factor associated with the accident was soft terrain.
169	NO	The selection by the pilot of an unsuitable precautionary landing site on soft, uneven terrain, which resulted in a rollover.
169	NO	The inadequate preflight planning by the pilot, the pilot initiating the flight with an inadequate fuel supply, and the unsuitable terrain encountered during the flight.
169	NO	The inadequate fuel supply for the flight which resulted in fuel exhaustion. A factor associated with the accident was the low altitude and the soft terrain.
169	NO	The pilots failure to maintain directional control during the landing. Factors were the crosswind and the soft terrain.
169	NO	the pilot's failure to maintain directional control during the landing roll, which resulted in the airplane departing the runway, impacting with a windsock, and the soft terrain.
169	NO	the unsuitable terrain for landing selected by the pilot. A factor was the soft terrain.
169	NO	the pilot's improper rotation and failure to maintain directional control during takeoff. Additional factors were the crosswind and the soft terrain.
169	NO	A loss of engine power for undetermined reasons, which resulted in a forced landing and subsequent nose over during landing roll. A factor was the soft terrain.
169	NO	The student pilot's failure to maintain directional control of the airplane during the landing roll. A contributing factor was the soft terrain.
169	NO	The improper planning/decision in runway selection. The soft runway condition and wet snow were contributing factors.
169	NO	Loss of engine power for undetermined reasons. Soft terrain was a factor.
169	NO	The pilot's decision to continue the takeoff. A factor in the accident was the soft wet runway.
169	NO	The pilot's use of unsuitable terrain (landing surface) at his privately owned landing site. A contributing factor was the soft area which the aircraft's right main landing gear contacted.
169	NO	The pilot did not maintain directional control and executed improper use of the brakes. A factor associated with the accident was the soft terrain.

Concentration of Fatal Accidents in Clusters

Cluster	N(NO)	Row %(NO)	N(YES)	Row %(YES)
130	3	42.00%	4	57.14%
27	2	40.00%	3	60.00%
139	7	36.84%	12	63.16%
82	11	35.48%	20	64.52%
155	1	33.33%	2	66.67%
208	7	31.82%	15	68.18%
206	3	30.00%	7	70.00%
81	2	28.57%	5	71.43%
83	2	28.57%	5	71.43%
77	4	25.00%	12	75.00%
50	1	20.00%	4	80.00%
187	7	19.44%	29	80.56%
53	2	18.18%	9	81.82%
205	2	10.00%	18	90.00%
222	0	0.00%	1	100.00%

- The clusters at the bottom of this table have a higher concentration of fatal accidents.

Spatial Disorientation: Fatal

Cluster	fatal	narr_cause
205	NO	Improper weather evaluation by both the pilot and pilot/passenger, and the pilot's inadvertent VFR flight into IMC resulting in his spatial disorientation.
205	YES	The pilots decision not to fly to the alternate airport, his decision to continue the flight in known adverse weather conditions, spatial disorientation by tl
205	YES	The pilot's failure to maintain control due to spatial disorientation.
205	YES	The pilot flying at an altitude insufficient to clear surrounding terrain. Contributing factors were the pilot becoming lost/disoriented, his subsequent spat
205	YES	The pilot's spatial disorientation due to a night visual illusion. A factor was the dark night condition.
205	YES	the pilot's spatial disorientation, which led to his failure to maintain aircraft control. A contributing factor was the pilot's decision to intentionally fly into
205	YES	the pilot's continued VFR flight into IMC, which resulted in spatial disorientation and the ensuing loss of aircraft control while in cruise flight. Contribut
205	YES	the pilot's VFR flight into IMC, which resulted in spatial disorientation and a loss of aircraft control. A contributing factor to the accident was the pilot's
205	YES	The pilot experienced spatial disorientation, which resulted in an in-flight loss of control and subsequent collision with trees and terrain. A factor was tl
205	YES	The pilot's failure to maintain a proper climb rate while taking off at night, which was a result of spatial disorientation. Factors in the accident were the
205	YES	The pilot's loss of control in flight due to spatial disorientation, and his subsequent overstress of the airplane during a recovery attempt. A factor in the
205	YES	The pilot initiated a VFR flight into known IMC conditions which resulted in a loss of control of the airplane due to spatial disorientation. Factors were t
205	NO	Pilot's failure to maintain adequate separation from terrain during the initial climb. Factors include spatial disorientation and a dark moonless night.
205	YES	The pilot's becoming lost and disoriented and his failure to maintain control of the airplane while flying over an unpopulated area on a dark night, which
205	YES	the pilot's failure to maintain aircraft control and his inadvertent flight into known adverse weather conditions. Factors relating to this accident were the
205	YES	The pilot experienced spatial disorientation that resulted in the loss of control.
205	YES	Flight into known adverse weather conditions by the pilot and the spatial disorientation of pilot. Contributing factors were the lack of certification by th
205	YES	The pilot's failure to follow operating procedures and, experienced spatial disorientation while attempting a night landing to an offshore platform. A fact
205	YES	The pilot's spatial disorientation, which resulted in his subsequent loss of control of the airplane. A factor was the dark night, over water visual conditi
205	YES	The pilot's spatial disorientation during a missed approach, which resulted in a loss of control, and the airplane's subsequent impact with water. Fact

Drugs: Fatal

Cluster fatal narr_cause

- 53 YES The airplane flightcrew's failure to maintain adequate distance/altitude from mountainous terrain during a departure climb to cruise flight, and the captain's impairment from drugs. Factors in
- 53 YES The pilot's inadequate altitude clearance above water while conducting low level flight maneuvers. A factor related to the accident was the pilot's impairment of judgment due to alcohol cons
- 53 YES The pilot's failure to maintain aircraft control during takeoff. A factor was the pilot's impairment due to a narcotic painkiller and antihistamine.
- 53 YES The pilot's failure to maintain aircraft control. A factor in the accident was the physiological impairment of the pilot due to the consumption of alcohol.
- 53 YES The pilot's unsuccessful recovery from an intentional aerobatic stall/spin maneuver. Contributing to the accident were the pilot's impairment (alcohol), and his psychological condition.
- 53 YES The pilot inadvertently stalled the airplane. A factor was the impairment due to marihuana.
- 53 YES The inadvertent flat spin of the airplane by the flightcrew resulting from the flight instructor's inadequate supervision. A contributing factor was the impairment (drugs) of the private pilot.
- 53 YES The pilot's unsuccessful corrective action (recovery) from an inverted spin. A contributing factor was the pilot's encounter with the inverted spin maneuver.
- 53 NO The pilot's failure to maintain control of the aircraft which resulted in an uncontrolled descent and an in flight collision with water. Contributing to the accident was the impairment of the pilot
- 53 YES The pilot's failure to maintain adequate airspeed which resulted in an inadvertent stall, and subsequent collision with terrain. A contributing factor was the pilot's impairment from the effects
- 53 NO The pilot's physical impairment due to a previous head injury which resulted in his becoming disoriented. A contributing factor was the lack of suitable terrain for the precautionary landing.

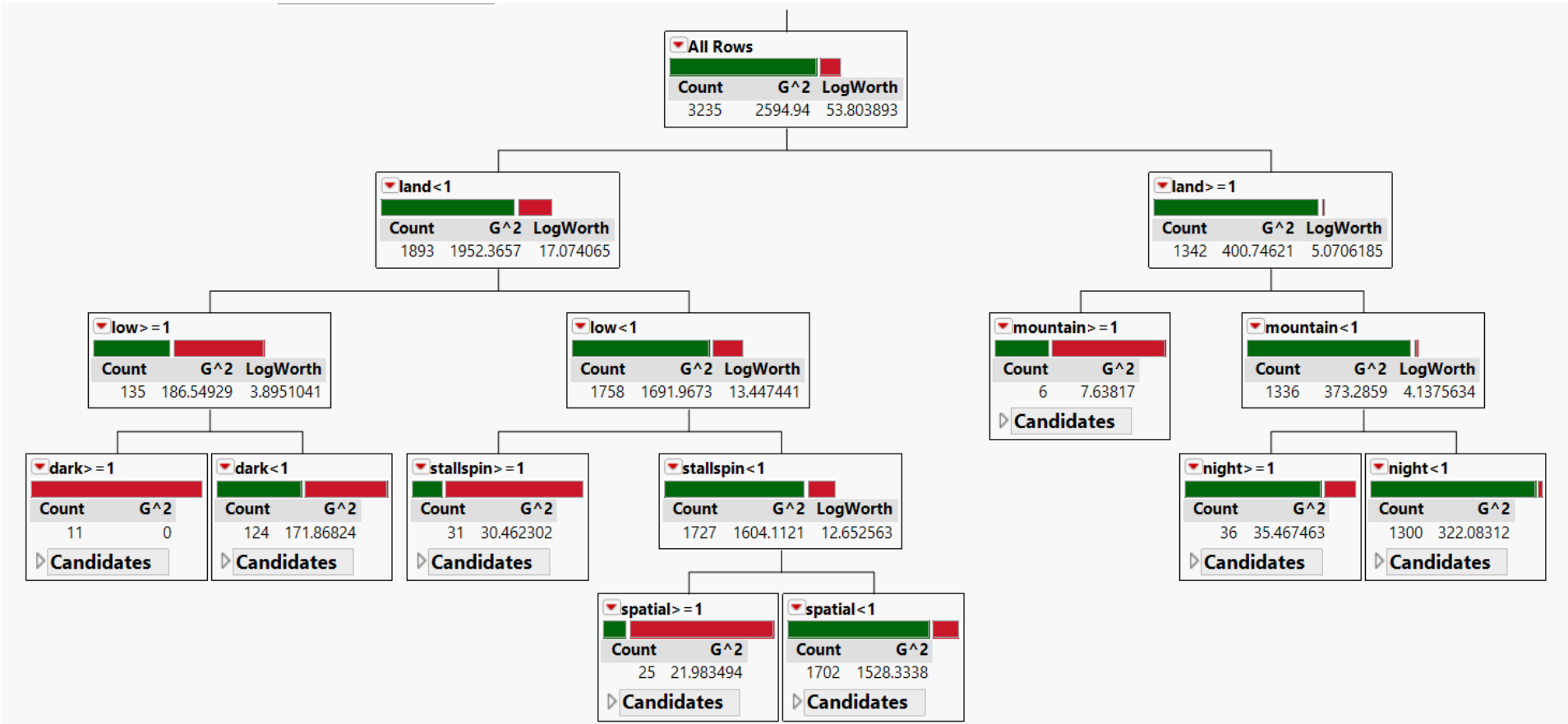
Failure to Maintain Airspeed: Fatal

Cluster fatal narr_cause

- 187 YES The pilot's failure to maintain airspeed, which resulted in an inadvertent stall/spin while on base leg.
- 187 YES the inadvertent stall/spin. Additional factors included the aerobatic maneuvers, low altitude, and the procedures not followed.
- 187 YES the pilot in command inadvertently allowing the airplane to stall/spin. Contributing factors were the pilot's total lack of experience in airplane make/model, and l
- 187 YES The pilot not maintaining aircraft control during the initial climb after takeoff and the inadvertent stall/spin. A factor to the accident was the pilot's total lack of e
- 187 YES the pilot's failure to maintain aircraft control due to his incapacitation for an undetermined reason. A contributing factor was the subsequent inadvertent stall/sp
- 187 YES the pilot's failure to maintain aircraft control following a loss of engine power while maneuvering, which resulted in an inadvertent stall/spin. Contributing factors
- 187 YES The student's failure to maintain adequate airspeed during the crosswind climb that resulted in a stall/spin at low altitude and the airplane's subsequent ground
- 187 YES The pilot's failure to maintain control of the airplane resulting in the inadvertent stall/spin. A factor was the pilot's unfamiliarity with the airplane.
- 187 YES The pilot's failure to maintain airspeed during an aerobatic maneuver, which resulted in an inadvertent inverted spin.
- 187 YES The pilot's improper use of the flight controls while turning to base, which resulted in a stall/spin and subsequent impact with the ground.
- 187 YES the failure of the pilot to maintain airspeed, which resulted in an inadvertent stall/spin, and subsequent impact with the terrain.
- 187 YES The pilot not performing an aborted takeoff and the inadvertent stall he encountered on his inadvertent initial climb. Factors were his inadvertent lift-off, the repo
- 187 YES the failure of the pilot to maintain airspeed, while attempting a forced landing following a loss of engine power for undetermined reasons, which resulted in an in
- 187 YES The inadvertent stall/spin by the pilot.
- 187 YES the pilot's failure to maintain aircraft control during the base turn, which resulted in an inadvertant stall/spin.
- 187 NO loss of engine power due to both piston rings failing, and the subsequent inadvertent stall/spin during the attempted forced landing. A contributing factor was tl
- 187 NO The inadvertent stall/spin encountered by the pilot during a slow flight maneuver. Factors relating to this accident were the low airspeed and the trees.
- 187 YES the failure of the pilot to maintain airspeed, which resulted in an inadvertent stall/spin, and subsequent impact with trees, while at a low altitude.
- 187 YES The pilot's failure to maintain airspeed during a low-altitude aerobatic maneuver, which resulted in an inadvertent stall/spin and subsequent uncontrolled descen
- 187 YES The loss of engine power for undetermined reasons, and the pilot's failure to maintain airspeed which resulted in an inadvertent stall/spin.
- 187 YES The pilots failure to maintain airspeed while maneuvering in instrument flight conditions resulting in an inadvertence stall/spin (vertical descent) and subsequent
- 187 YES the pilot's failure to maintain control of the airplane while maneuvering resulting in an inadvertent stall/spin.
- 187 YES The pilot's failure to maintain airspeed after a loss of engine power, which resulted in an inadvertent stall/spin. Also causal, was the loss of engine power for ur
- 187 YES The pilot's failure to maintain adequate airspeed during the turn to final, which resulted in an inadvertent stall/spin. Factors included low ceilings and night light
- 187 YES the pilot's failure to maintain control of the airplane resulting in the airplane entering a flat spin from which the pilot did not recover.
- 187 YES The pilots' failure to maintain airspeed, which resulted in an inadvertent stall/spin. The continued spin to the ground was a result of the pilots' failure to deploy t

Decision Tree for *Fatal* Using DTM

- We can use all 800 columns (each a word) in the Document Term Matrix as input variables to predict whether or not an accident will be fatal
- If land is in the narrative, it will almost surely not be fatal



Most Useful Words to Predict *Fatal*

Column Contributions			
Term	Number of Splits	G ²	Portion
land	1	241.828087	0.1725
low	2	80.4145073	0.0574
mountain	2	66.080032	0.0471
stallspin	1	57.3928079	0.0409
stall	2	57.2399443	0.0408
spatial	1	53.7948553	0.0384
loss	3	47.7425258	0.0341
control	4	47.5295539	0.0339
maneuver	2	34.322482	0.0245
inflight	1	33.8711685	0.0242
maintain	3	33.2827629	0.0237
intent	1	32.7165376	0.0233
fog	1	32.3386054	0.0231
failur	5	25.9087836	0.0185
night	3	25.6879536	0.0183
undetermin	1	25.1220351	0.0179
collis	2	25.0195528	0.0179
direct	1	24.1305318	0.0172
vfr	1	21.7555089	0.0155
dark	2	19.8593295	0.0142

Word Cloud Colored by Proportion Fatal and Filtered by Meteorological Conditions

Visual
Meteorological
Conditions



Instrument
Meteorological
Conditions



Summary

- Data is growing exponentially across DOD
- Much of this is unstructured text data
- Text mining takes statistical tools and converts the text into meaningful mathematical expressions
- Example uses of text mining
 - Concept extraction
 - Grouping like documents or records together
 - Grouping terms together
 - Creating structured variables that represent the text fields to use in predictive analytics
- Relatively short learning curve to perform powerful text analytics using open source software and commercial solutions

Questions?

Thank You!