The Power of Words

How Quantitative Text Analysis Can Support ML Fraud Detection

Erik Halvorson – 303-901-3070 – erik.halvorson@archimedeseval.com

1

Introduction

Anti-Fraud Program Manager for New Funding Endeavors (steward approximately $500B for OI) with 13 years as a SA and 18 years as a LEO. Ranging from nuclear security to Military Police Patrol/Operations to white collar fraud cases and proactive fraud detection design.

Experienced in domestic and international law enforcement. Assigned to DOJ's ICCTF in Saudi Arabia and joint constabulary interface in UK. Assigned to 6 countries and based in 4 states. Worked cases across US federal jurisdictions ranging from Colorado to SD to WY to MO to DOJ Main Justice to local District Attorneys to foreign governments. NIJ Law Enforcement Advancing Data and Science Scholar.

Holds degrees in Criminal Justice, Accounting, Applied Analytics, and currently finishing a PhD in Research Methods and Statistics with a focus in Analytics and Program Evaluation.

2

Learning Objectives:

1. Understand research underpinning text and deceit.
2. Gain knowledge surrounding how text can be used in analytics.
3. Understand the basic process to analyze text generally.

3

Topics:

Part 1: Research
Part 2: Basic Text Analysis
Part 3: Predictive Analysis
Part 4: Topic Modeling
Part 5: Implications

4

Part 1: Research:

Pure

Applied

AI

ML

Where:

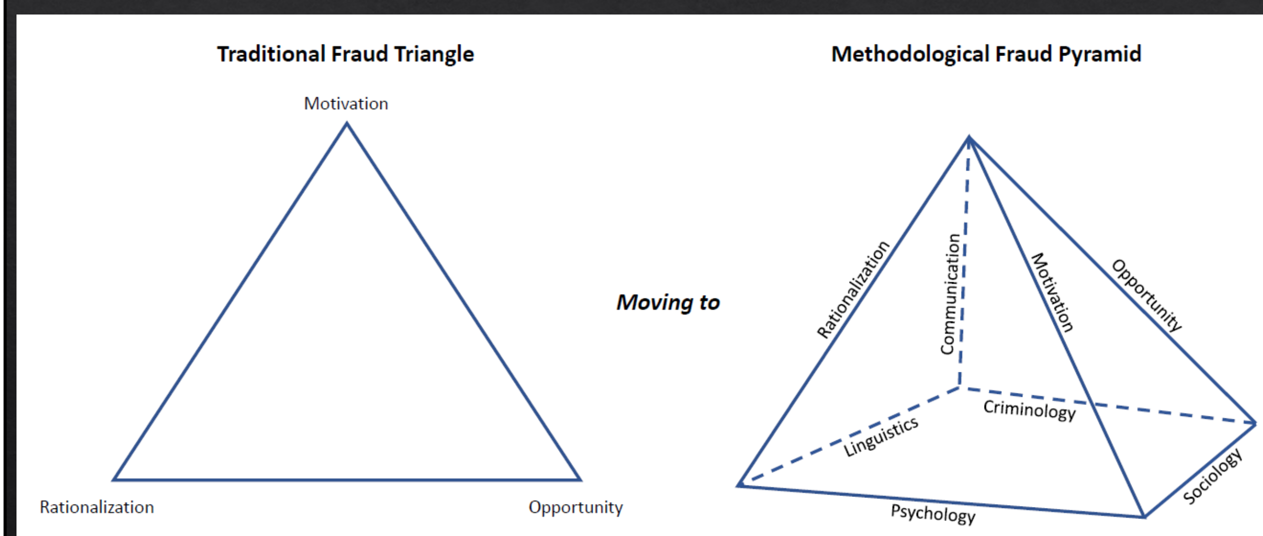
Machine Learning is a subset of AI through applied learning where computers learn and adapt to draw inferences from patterns in data.

And:

Modeling is the use of math and stats to analyze data to make predictions about the real world.

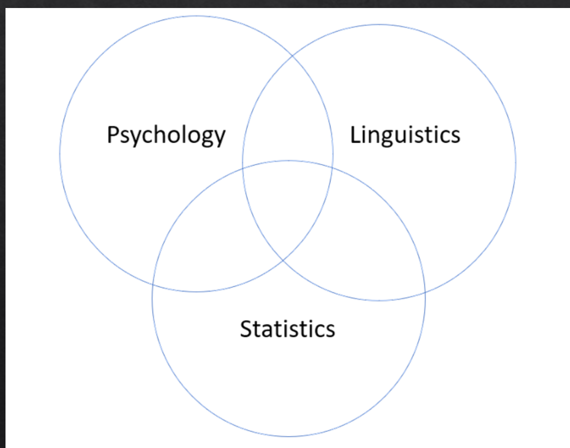
5

Research Premise:



6

Research Premise:



Why Text? Why words?

Text exists all around us but it is not often utilized in many analyses.

Lifeblood of government is reports. What if we could automate some of that?

They say the eyes are the window to the soul but I would say for our jobs, words are the window into a bad actor's intent.

7

Previous Research:

- Most research is on poorly built numerical models with bad results (Koreff et al, 2021; Ramamoorti & Curtis, 2003; Phillips, & Lanclos, 2014).
- Although many published works fail to provide accuracy rates, a systematic review of fraud detection methodologies within healthcare shows that attempts to use numerical based models have generally created high false positive rates and low accuracy rates (Ai et al, 2022).
- This combination discourages government agencies from using modern machine learning predictive models to detect and prevent fraud.
- But, behavioral psychology found when deception was employed by subjects that word choice and text-based content change (they shrink) (Adams, 2002)
- These studies ranged from analyzing between 5 documents to 100 documents or studies with up to128 students (Craig et al, 2013; Clatworthy and Jones, 2220; Caso et al, 2005).
- Due to these small samples, previous research was generally relegated to manual qualitative reviews and limited statistical techniques.

8

Part 2: Basic Text Analysis:

- The SBIR program gives money to small businesses to conduct R/R&D.
- Awarded more than 179,000 grants for over $54B since 1982.
  - Also, had 3 GAO reports to congress about fraud within the program.
- Using press releases we can identify known fraud. In our case we identified more than 700 awards with known fraud:
- Once we have known fraud examples, we can extract the text we want to analyze.
- If our sample is big enough, we can use basic trend analysis to look for recurring words that might help identify program risk areas.

9

Basic Text Analysis:

Reports & Documents

Social Media

Internet Pages

Extracting Blocks of Text:

To enable the future needs of efficient computing for edge data centers, Flexnode and its partners, the University of Maryland (UMD), Boeing, Iceotope, SHoP Architects, and Arup, propose to develop a prefabricated modular micro data center with unprecedented energy efficiency and power density.

Cleaned and formatted:

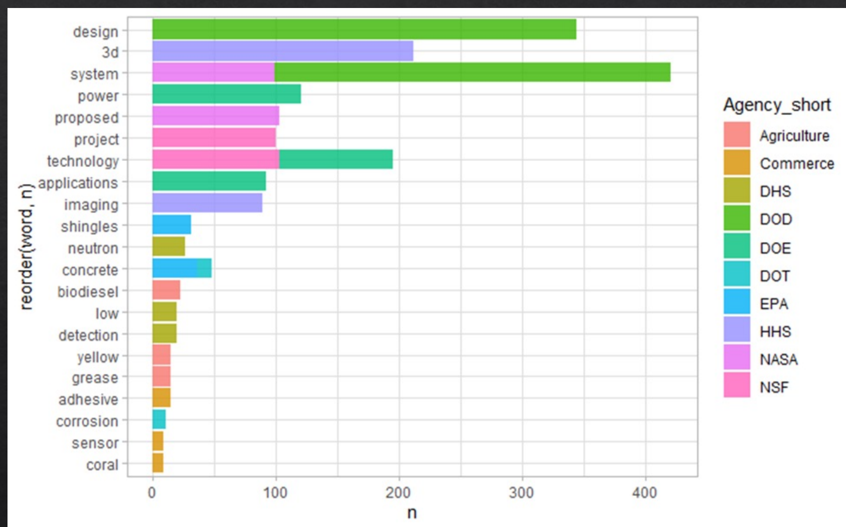
enable future needs efficient computing edge data centers flexnode partners, university maryland boeing iceotope shop architects arup propose develop prefabricated modular micro data center unprecedented energy efficiency power density

Aggregated:

Words are scored, categorized, and topic areas are understood

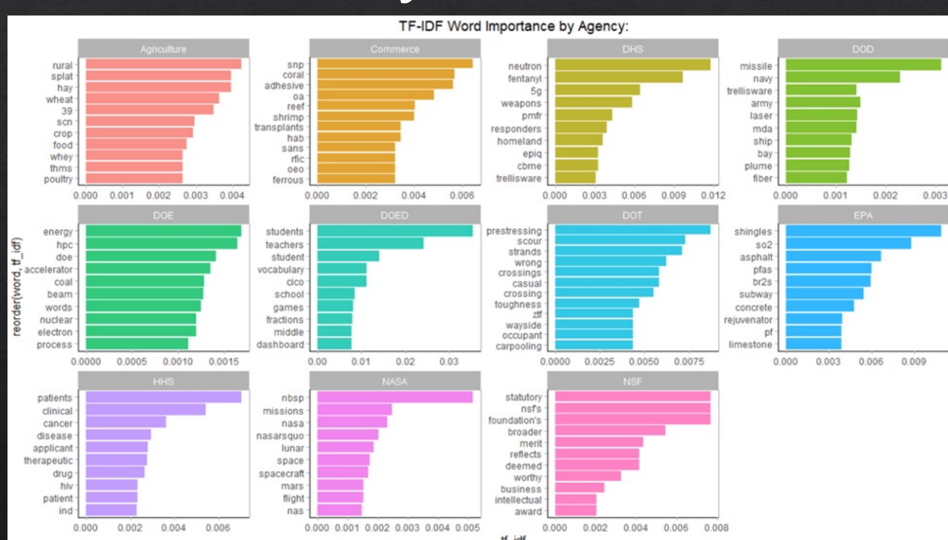
10

Basic Text Analysis:



11

Basic Text Analysis:



12

Part 3: Predictive Analysis:

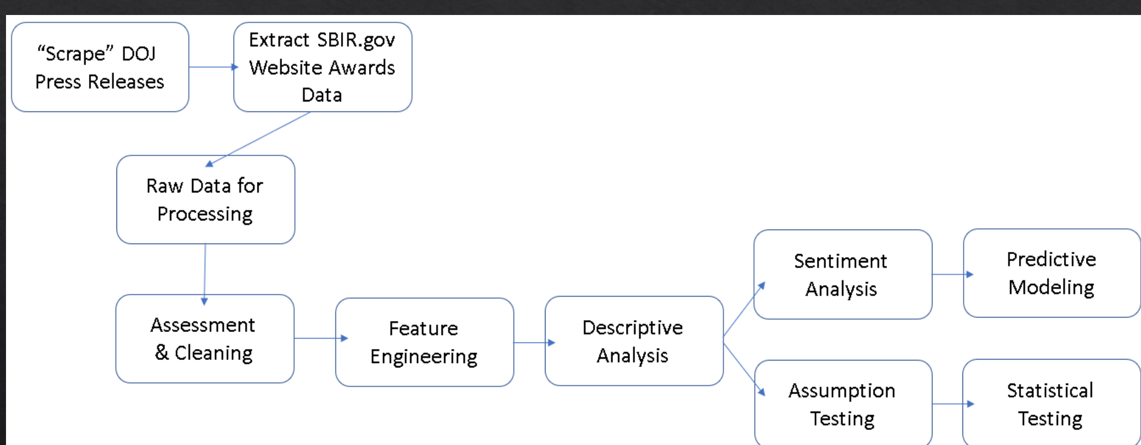
Lets take a specific scenario within the same program:

- Data was cleaned, this resulted in final numbers: 747 known fraud, 3,227 not known fraud for a total of 3,974 awards.
- Used feature engineering to create 14 variables related to the Abstract itself.
- Also, used two numerical variables already present in the data (# of employees and award amount) and three binary self certifications (Woman Owned Business, HUBZone, and Socially or Economically Disadvantaged).
- The inclusion of text-based variables and more standard variables will allow us to compare the performance between the two types.

Base Accuracy: 69.4% & Null Model Accuracy: 81.1%

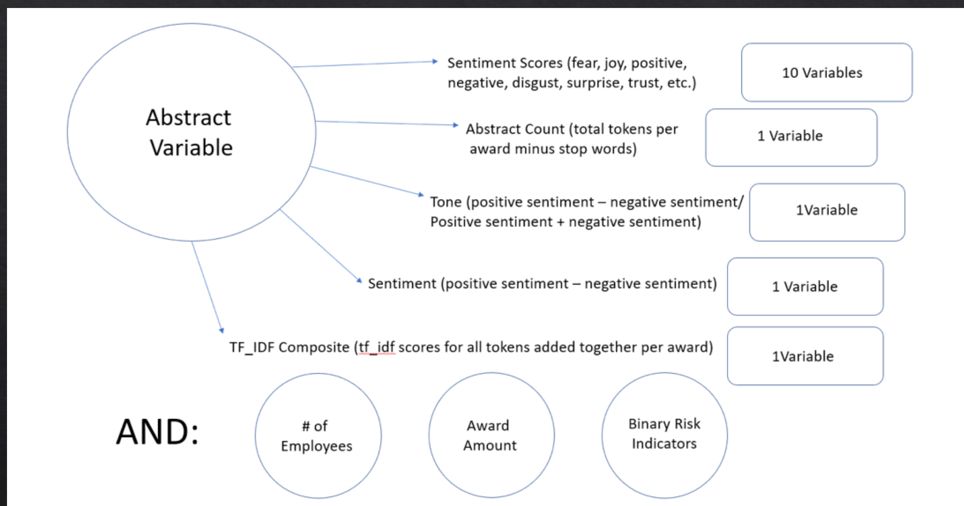
13

Cleaning and Modeling:



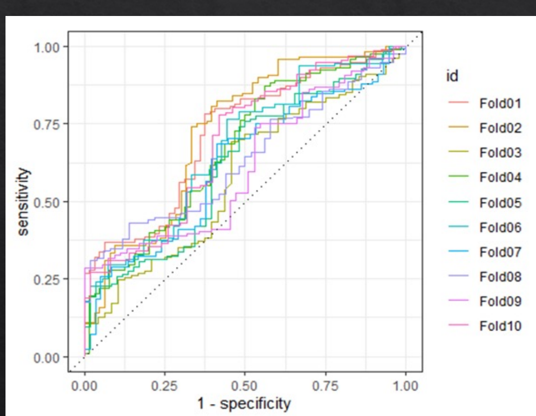
14

Variable Building:



15

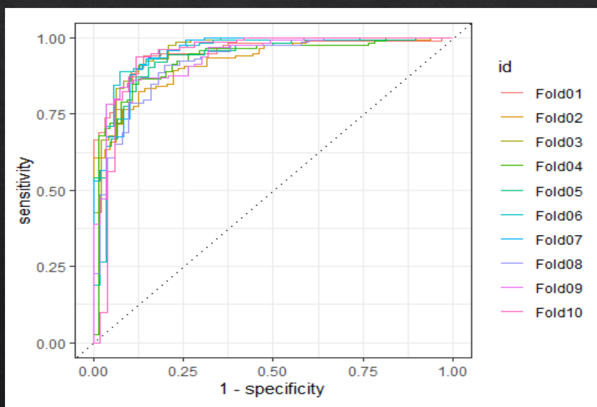
Initial (Non-Text) Model:



Assessment	How did it do?
Accuracy	69%
ROC_AUC	66%

16

Text Inclusive Model:



Assessment	How did it do?
Accuracy	90.3%
ROC_AUC	91.5%

Accuracy showed a 31% increase and ROC_AUC showed a 39% increase!

17

Statistical Analysis:

Fraud vs Not Known Fraud Awards

Variable Titles	Fraud		Not Known Fraud		t	p	Cohen's d
	M	SD	M	SD			
Abstract Count	2.045	0.211	2.144	0.246	11.13	<0.001	0.411
Sent. - Anger	0.340	0.289	0.431	0.336	7.52	<0.001	0.279
Sent. - Anticipation	0.634	0.277	0.784	0.314	13.00	<0.001	0.489
Sent. - Disgust	0.201	0.261	0.324	0.339	10.92	<0.001	0.378
Sent. - Fear	0.471	0.319	0.596	0.376	9.35	<0.001	0.343
Sent. - Joy	0.382	0.269	0.497	0.314	10.25	<0.001	0.378
Sent. - Negative	0.632	0.327	0.764	0.375	9.70	<0.001	0.362
Sent. - Positive	1.130	0.251	1.235	0.268	10.20	<0.001	0.398
Sent. - Sadness	0.298	0.275	0.449	0.373	12.57	<0.001	0.423
Sent. - Sentiment	0.969	0.316	1.046	0.328	5.84	<0.001	0.236
Sent. - surprise	0.271	0.262	0.329	0.279	5.13	<0.001	0.208
TF-IDF Composite	0.662	0.033	0.682	0.037	14.52	<0.001	0.542
Tone	0.179	0.095	0.167	0.098	-3.21	.002	-0.127
Sent. - Trust	0.853	0.279	0.958	0.282	9.21	<0.001	0.374
Employee Number	53.666	76.730	41.697	78.953	-3.82	<0.001	-0.152
Award Amount	$336,938	350,651	$513,297	638,156	10.34	<0.001	0.296

18

Part 4: Topic Modeling:

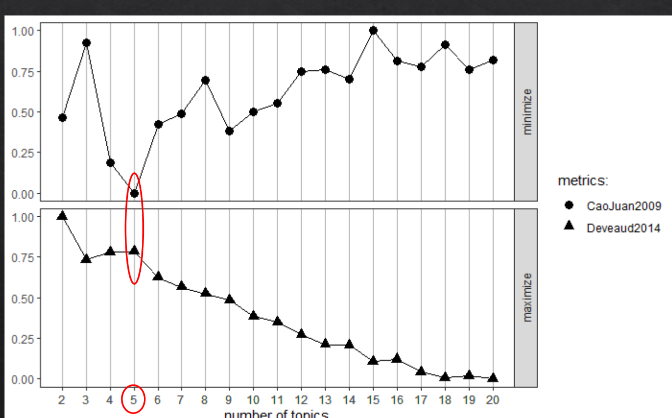
Topic Modeling is a flexible (statistical) analysis technique that identifies:

- Topics within individual documents
- Themes that occur across a series of documents
- Within the topics and themes we can extract key words that covary together

So, imagine you could import thousands of pages or reports, website data, budgets, etc. and analyze the data in a few minutes to identify areas of risk.

19

Topic Modeling:



The first step of a topic model is to find your data sources. Once you have downloaded your files:

- .pdfs
- word/text documents
- websites
- excel files with text
- social media
- tables

Analyze the optimal number of topics based on word correlation.

20

Topic Modeling:

Say we want to compare the CHIPS and Science Act Requirements across a budget to see what requirements are being funded versus potential areas of risk.

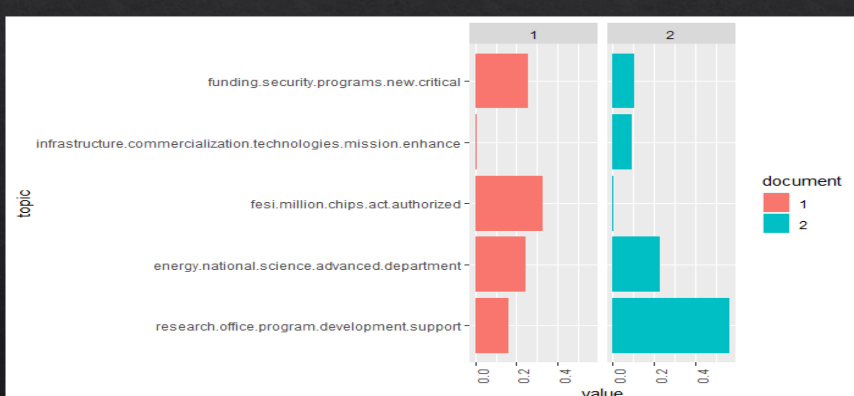
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Science	Funding	Security	Research	FESI
Energy	Millions	Technologies	Development	Infrastructure
Techology	Innovation	Accelerate	National	Commercialization
Mission	New	Computing	Support	National
Department	Critical	Microelectronics	Programs	Private

Taking what we know about areas of risk we can extract key words associated with Topic 3: Security we find the following risk factors: *universities, non-profits, funding/financial, fabrication facilities, foreign, collaboration, and advanced computing.*

Or if we look at Topic 4: Research we see the following program areas and potential risk indicators: n*uclear, fusion, carbon, quantum (computing), foreign, investments/funding/grants, chips, non-profits, academic, research.*

21

Topic Modeling:



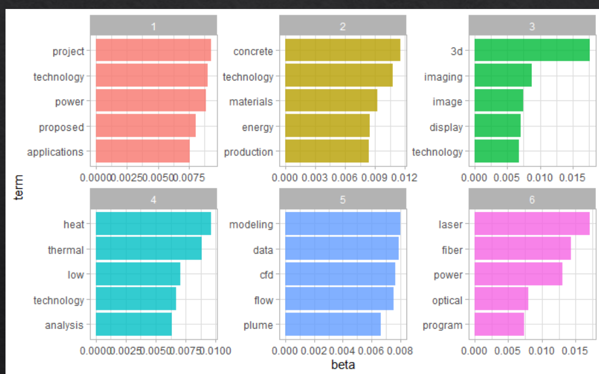
Compare document 1 (Chips Act Law) vs Expenditures:

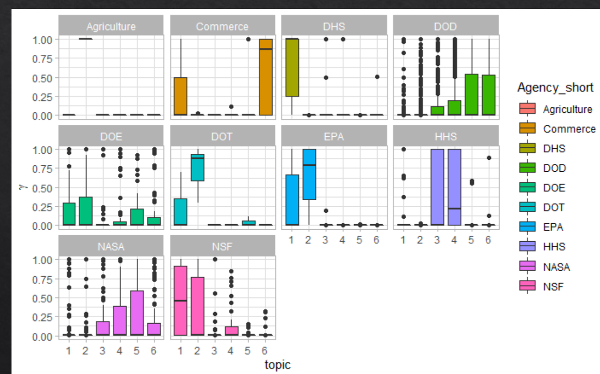
Law talks about Fundings, Security, Programs 35% of the time but expenditures only address that same function 10% of the time.

Infrastructure, commercialization, technologies, etc. 1% law vs 10% funded.

22

Topic Modeling:





23

Topic Modeling:

Category Title	# of Reports	# of Pages	Description of Reports
GAO Reports	6	948	GAO reports concerning oversight and risk with Agency based programs
Audit Reports	10	411	OIG reports concerning fraud risks to Agency programs
OIG SARs	11	1,145	Semi-Annual Reports to congress from FY 2018 to Current
Misc. Reports	8	300	Miscellaneous reports regarding fraud risks that have a nexus to Agency's specific mission
DOJ Press Releases	2*	162*	DOJ press releases for prosecutions related to agency programs and cases
Total:	**37**	**2,966**	

24

Topic Modeling:

Category Title	Document n*	Themes
GAO Reports	4,069	Grantees, disaster fraud, recovery, applicant eligibility, subrecipients, corruption
OIG Audit Reports	4,775	CARES Act, contractor, grantee, kickbacks/corruption, benefits fraud, ineligible payment, fictitious
OIG SARs	7,028	Applications, grants, community block grants, lead-based paint
Misc. Reports	2,709	COVID, payment schemes, program management, eligibility fraud, officials/corruption, loan, rental
DOJ Press Releases	503	Sexual harassment, civil fraud, discrimination, public authority, etc.
Total:	19,084	23

25

Topic Modeling:

Meta-theme #	Title	Document Themes Comprising:
1	Grant & Loan Fraud	Grantees, disaster fraud funding, Program fraud, community block grants, grants
2	Eligibility & Benefits Fraud	Subrecipients, applicant eligibility, contractor, benefits fraud, ineligible payment, fictitious (eligibility)
3	Pandemic Fraud	CARES Act, Recovery Act, applications, PPP, COVID, grants
4	Program Fraud	Program Areas, payment schemes, program management, theft, embezzlement
5	Employee Corruption Facilitated through Program Fraud	Kickbacks/corruption, officials/corruption, program (mis)management, fictitious (businesses and payments), subrecipients, contractors

Note. Some document level themes map onto multiple meta-themes. This is considered a dual code and supports the holistic approach identifying overlap between the themes.

26

Part 5: Implications:

We see that text based variable scores do support the literature that content and context shrink/change when stress from deceit is introduced.

- This is important for our anti-fraud work The applications of this type of analysis are broad We already have the data… lets use it!

There is a dramatic increase in predictive power when text-based variables are included.

- Using a LASSO/Elasticnet introduces some bias but predicts across groups better

There are more applications for scheme detection and vulnerability analysis in areas like cluster, LDA/Topic Modeling where we can analyze thousands of cases but lose some of the granularity of traditional qual methods.

There is no silver bullet. Lets be creative and solve problems.

27

Questions?

Follow up? Questions? Training? Cool project or research ideas?

Erik Halvorson – 303-901-3070 – erik.halvorson@archimedeseval.com

28