APPLICATIONS OF VISUAL ANALYTICS – TEXT ANALYSIS WITH IN-SPIRE AND STARLIGHT

Thomas Dang, Andrew Wade, Victoria Lemieux, Ron Rensink, Brian Fisher, Chris Rogers, Lonnie Hastings, Kyle Melnick, Karl Eckler (UofWashington)
Media and Graphics Interdisciplinary Centre (MAGIC)
School of Library, Archival, and Information Sciences (SLAIS)
Center for the Investigation of Financial Electronic Records (CIFER)

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

COMMON METHODOLOGY

WHAT IS IN-SPIRE AND STARLIGHT?

❑ Unstructured text VA software initially for national security, commercialized through PNNL and Future Point systems

❑ Proprietary "black box" text mining and clustering engines

❑ Scalable to tens of thousands of records

DATA NORMALIZATION

❑ Remove duplicate and empty records

❑ Change plural nouns to singular to avoid double count

❑ Remove irrelevant grammatical constructs - e.g. remove adjectives and adverbs if looking mainly for themes)

❑ Remove repeating boiler-plate phrases

❑ When possible, use Text Wrangler or RegExp scripts to delimit sections within the records

DATA FORMATTING & IMPORT

❑ Format data to program-specific XML, ideally

❑ Starlight XML Engineering Environment

❑ In-Spire Dataset Editor

SORT AND QUERY DATA BY TIME

❑ Use Excel to pre-sort data by time if time is available as an attribute of each record

❑ Use In-Spire Time Slicer and Starlight Time Series to graph these subsets and query them

THEME-CLUSTER VISUALIZATION (ITERATIVE)

❑ Galaxy / Theme view in In-Spire, Topic view in Starlight

❑ Cluster by "theme", i.e. contributions of keywords

❑ Systematically remove themes too general or specific

❑ Distance denotes differences between records and between clusters

PAIR ANALYTICS

❑ Technology (Visual Analytics) experts drive the tools

❑ Subject matter experts verify the visualizations and findings

TRIANGULATION

❑ Use more than one engines to verify if the visualizations converges semantically.

❑ Use multiple datasets of the same theme but different record lengths.

VERIFY WITH MANUAL CODING

❑ Randomly sample 4 articles per cluster, 3 near the centre and 1 at the edge

❑ Verify that the articles at the centre are very coherent in themes, and even the one at the edge is vaguely coherent

LIMITATIONS

❑ Proprietary text mining and visualization engines

❑ High licensing cost of In-Spire and Starlight

❑ Difficulty in obtaining and normalizing unstructured text data

INFORMATION ASSURANCE PROJECT

MOTIVATION & DATA SET INTRODUCTION

❑ Information Assurance (IA) is a new multidisciplinary field bridging information theory, technology and risk management. These complexities lead to disagreement over current responsibilities and future goals.

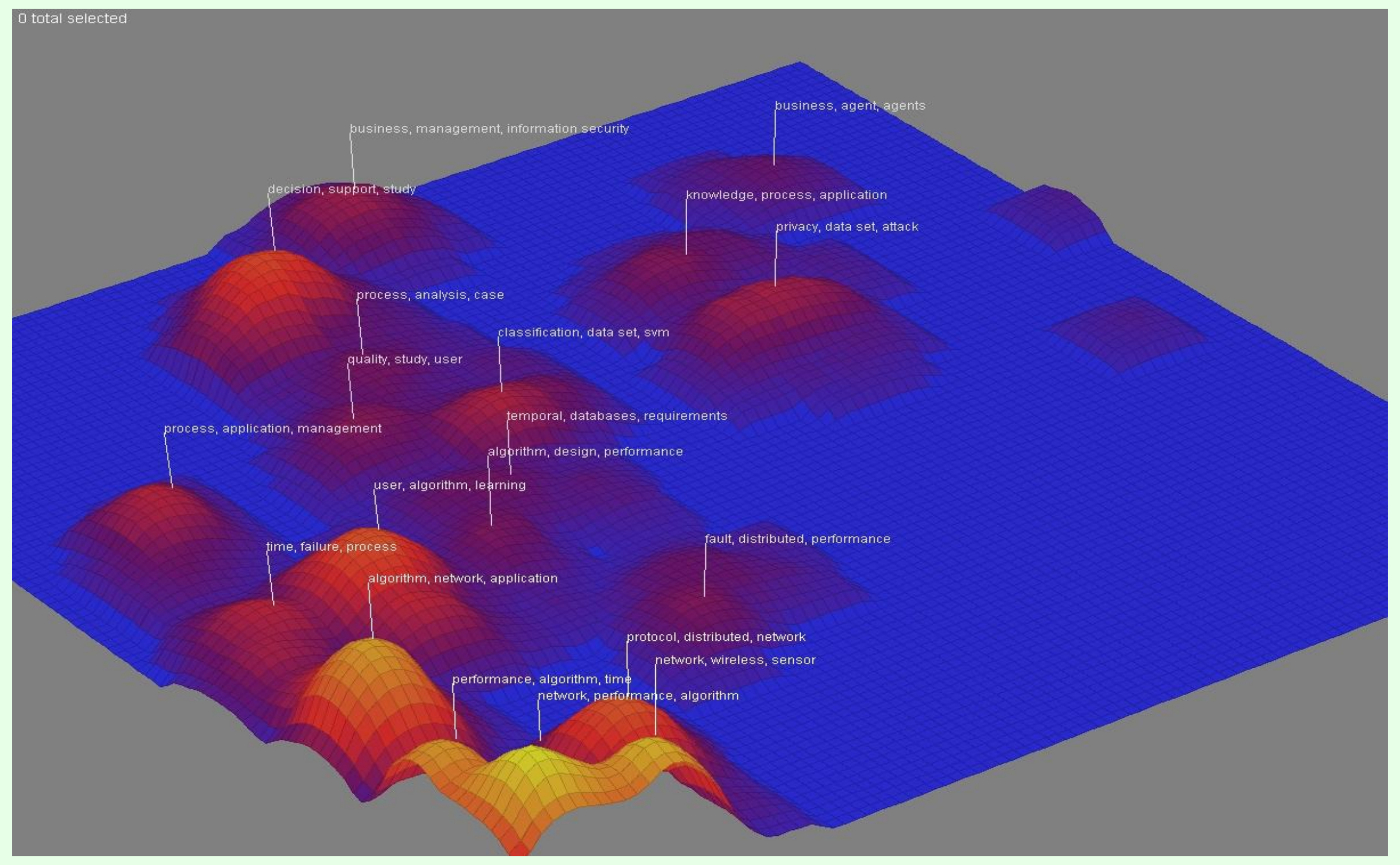
❑ VA is used to enable a large-scale literature survey and taxonomy building

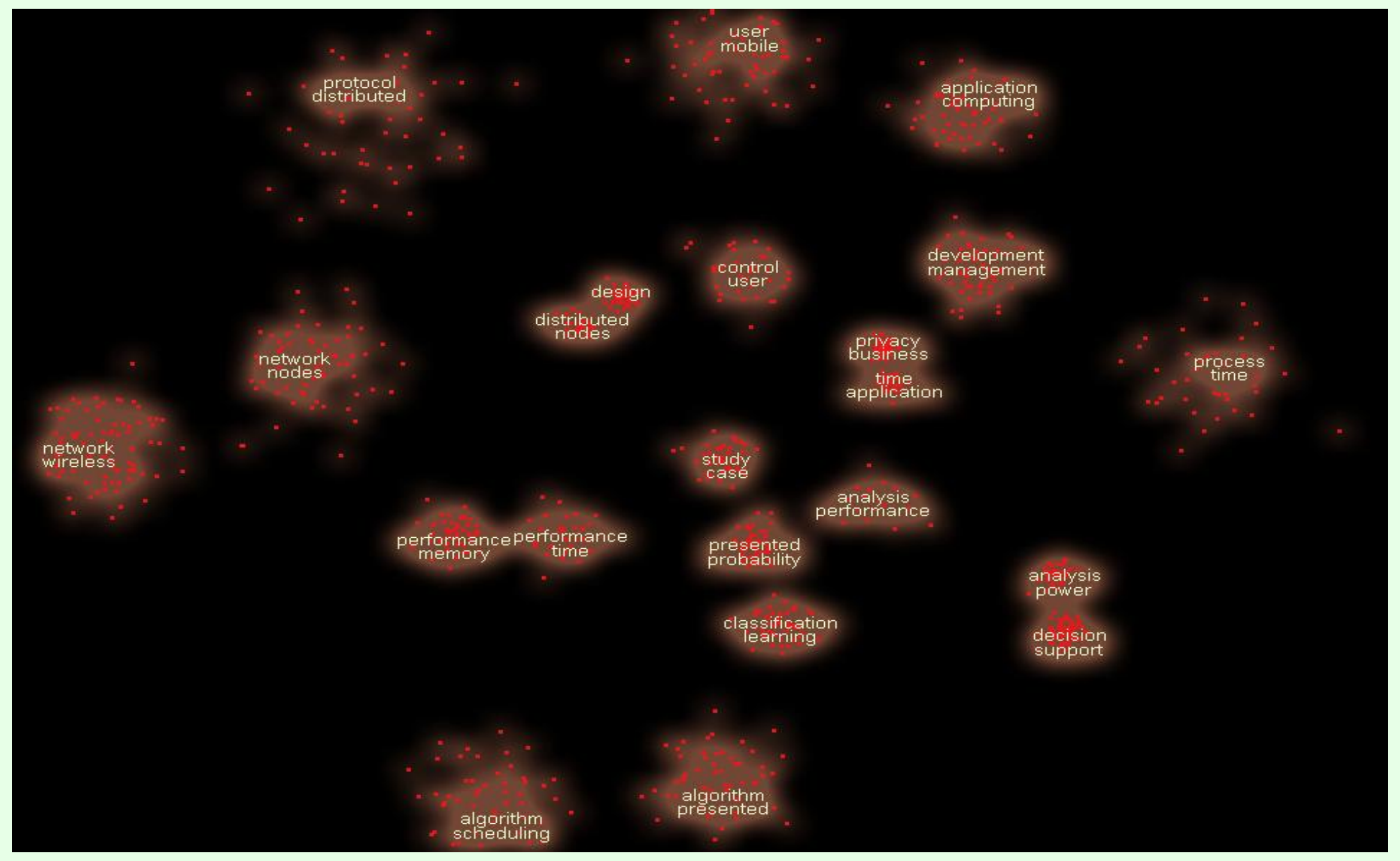
❑ Dataset: 1000+ articles, 10-20x the number practical for manual coding

FINDINGS

❑ Inordinate concentration of research on technological solutions and problems to the detriment of legal, managerial and training issues.

❑ VA enables large-scale bibliographic analysis and coding





FUTURE WORK

❑ Sort and visualize the data by source journals to see if we are biased to certain source.

❑ Visualize and compare side-by-side the subsets of data by time (using the Time Slicer / Time Series visualization)

❑ Taxonomy creation by an IA researcher

ACKNOWLEDGEMENT

❑ The National Security Agency (NSA)

❑ The Boeing Company

FINANCIAL DATA LOSS PROJECT

MOTIVATION & DATA SET INTRODUCTION

❑ Data loss risks cost the financial industry millions of dollars in monetary value and an unquantifiable amount in reputation

❑ Data loss regulations are becoming more aggressive and strict: firms can be held liable for hypothetical losses, not just confirmed losses

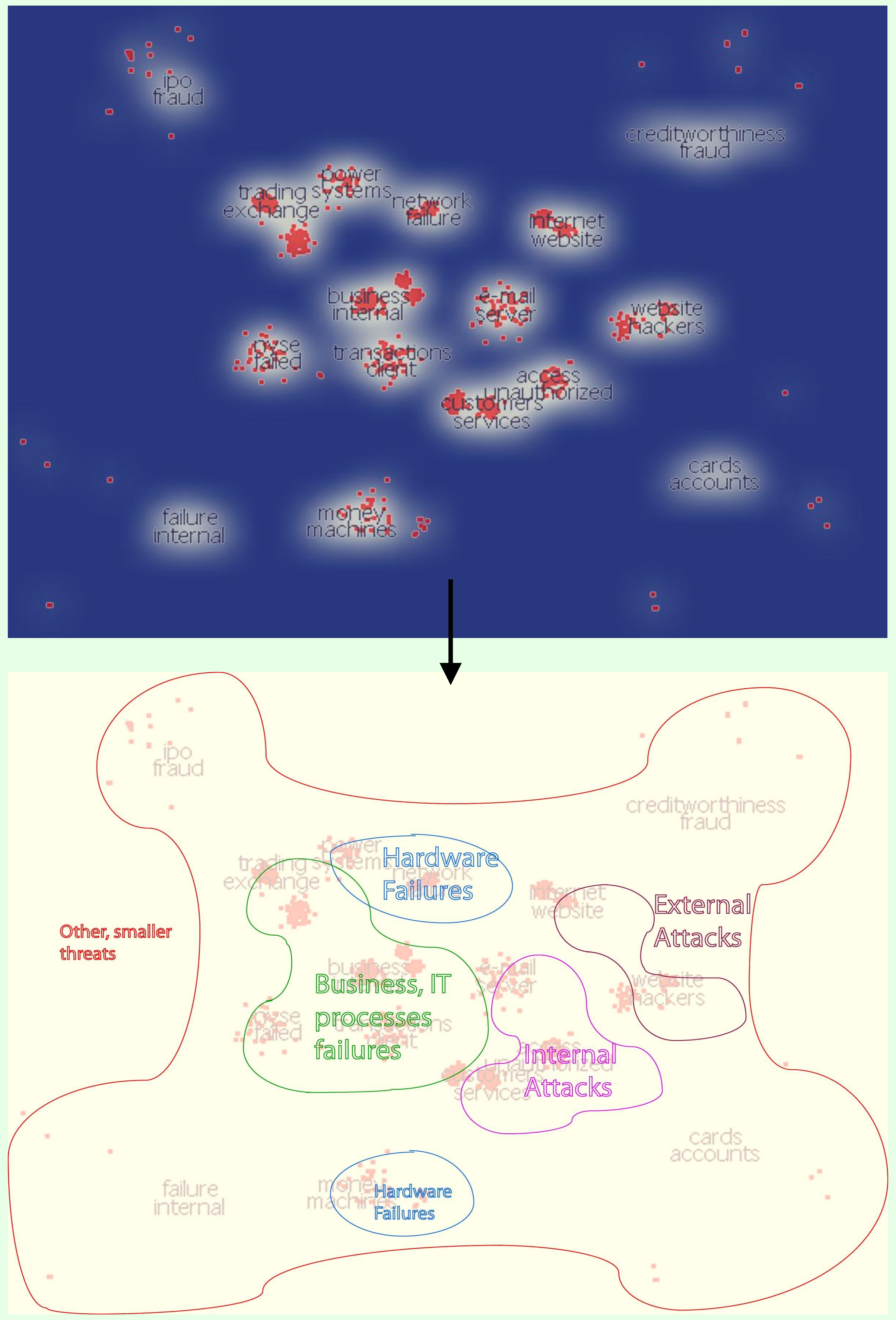
❑ Dataset: 1200+ descriptions of cases in 100+ firms from 30+ countries (confidential internal data donated to CiFER).

FINDINGS

❑ Internal threats >= External threats

❑ Accidental >= Deliberate

❑ Hardware risks are significant!



FUTURE WORK

❑ Refine our Visualizations further to reveal finer, smaller groupings of cases, where the risks involved are more specific

ACKNOWLEDGEMENT

❑ CiFER (for liaison work with firms to gather data)

BOEING BIRD STRIKE PROJECT

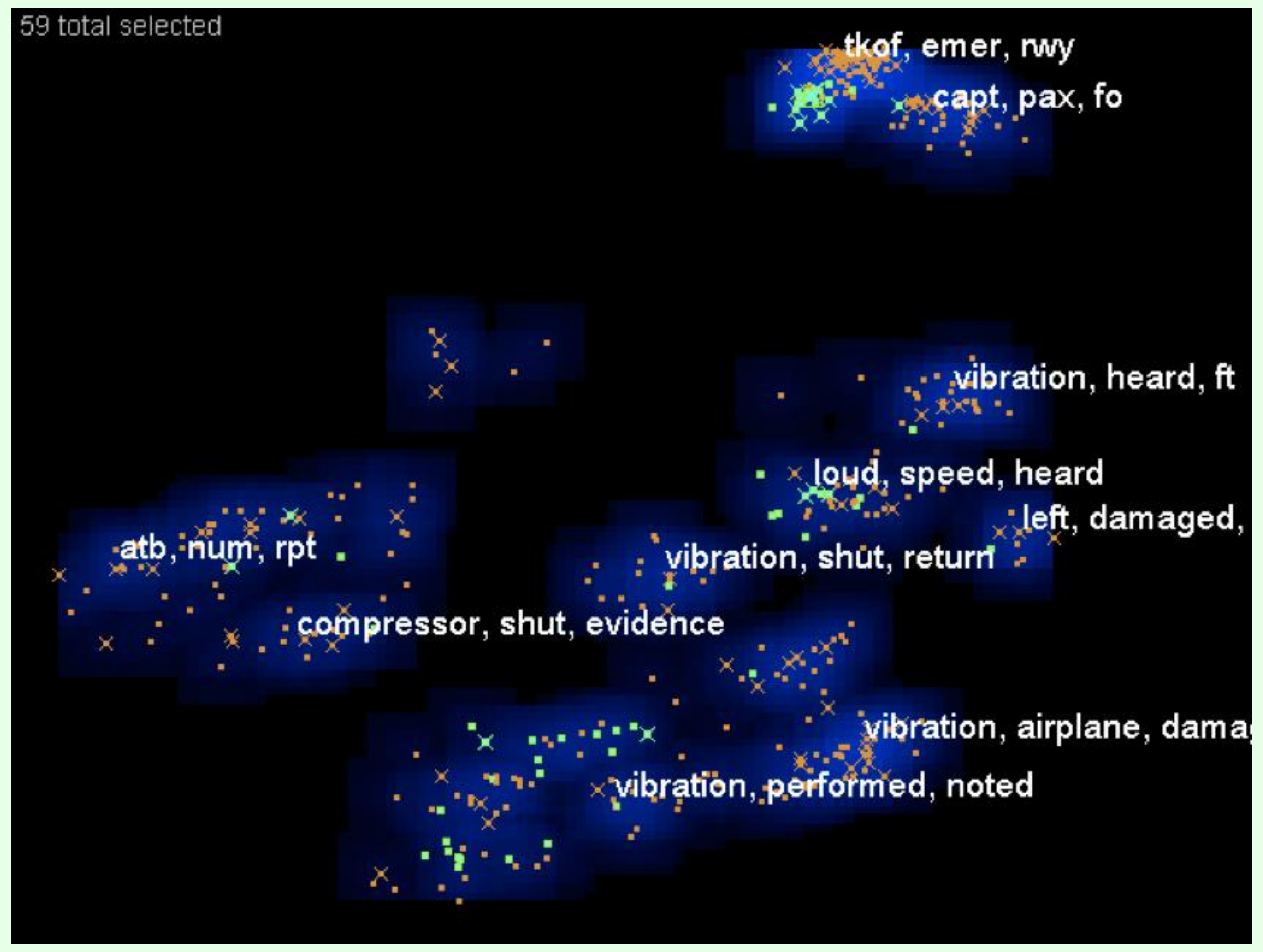
MOTIVATION & DATASET INTRODUCTION

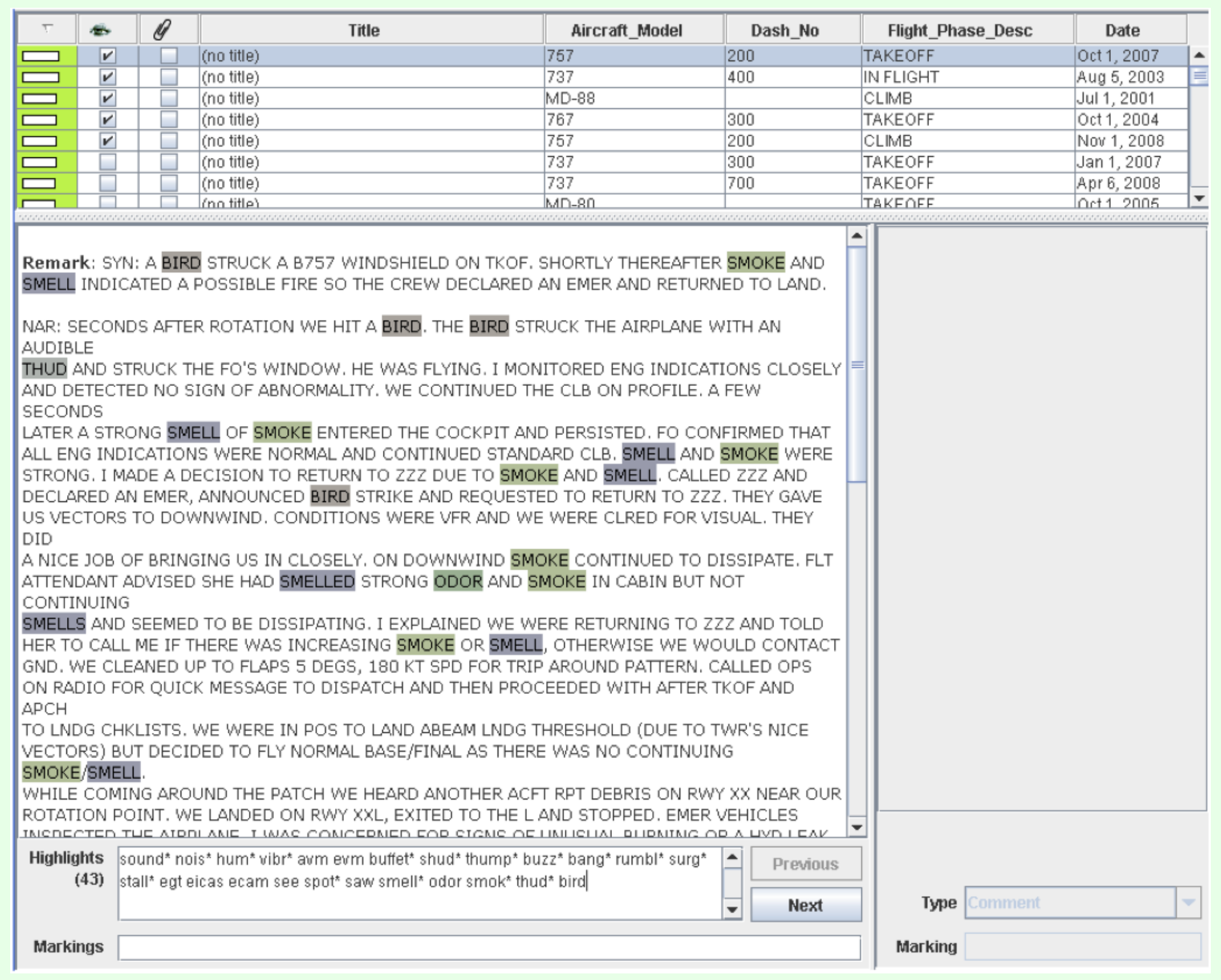
❑ Bird strikes cost the aviation industry millions of dollars a year as well as endanger the lives of pilots and passengers. Applying VA techniques to bird strike data may provide new insights into safety.

❑ ~10000 unstructured text descriptions from the FAA

FINDINGS

❑ VA techniques allowed for more advanced text analysis of pilot responses to bird strikes by greatly reducing the amount of analysis time, and resulted in recommendations for pilot behaviour.





			Title	Aircraft_Model	Dash_No	Flight_Phase_Desc	Date
	☑	☐	(no title)	757	200	TAKEOFF	Oct 1, 2007
	☑	☐	(no title)	737	400	IN FLIGHT	Aug 5, 2003
	☑	☐	(no title)	MD-88		CLIMB	Jul 1, 2001
	☑	☐	(no title)	767	300	TAKEOFF	Oct 1, 2004
	☑	☐	(no title)	757	200	CLIMB	Nov 1, 2008
	☐	☐	(no title)	737	300	TAKEOFF	Jan 1, 2007
	☐	☐	(no title)	737	700	TAKEOFF	Apr 6, 2008
	☐	☐	(no title)	MD-80		TAKEOFF	Oct 1, 2005

Remark: SYN: A BIRD STRUCK A B757 WINDSHIELD ON TKOF. SHORTLY THEREAFTER SMOKE AND SMELL INDICATED A POSSIBLE FIRE SO THE CREW DECLARED AN EMER AND RETURNED TO LAND.

NAR: SECONDS AFTER ROTATION WE HIT A BIRD. THE BIRD STRUCK THE AIRPLANE WITH AN AUDIBLE
THUD AND STRUCK THE FO'S WINDOW. HE WAS FLYING. I MONITORED ENG INDICATIONS CLOSELY AND DETECTED NO SIGN OF ABNORMALITY. WE CONTINUED THE CLB ON PROFILE. A FEW SECONDS
LATER A STRONG SMELL OF SMOKE ENTERED THE COCKPIT AND PERSISTED. FO CONFIRMED THAT ALL ENG INDICATIONS WERE NORMAL AND CONTINUED STANDARD CLB. SMELL AND SMOKE WERE STRONG. I MADE A DECISION TO RETURN TO ZZZ DUE TO SMOKE AND SMELL. CALLED ZZZ AND DECLARED AN EMER, ANNOUNCED BIRD STRIKE AND REQUESTED TO RETURN TO ZZZ. THEY GAVE US VECTORS TO DOWNWIND. CONDITIONS WERE VFR AND WE WERE CLRED FOR VISUAL. THEY DID
A NICE JOB OF BRINGING US IN CLOSELY. ON DOWNWIND SMOKE CONTINUED TO DISSIPATE. FLT ATTENDANT ADVISED SHE HAD SMELLED STRONG ODOR AND SMOKE IN CABIN BUT NOT CONTINUING
SMELLS AND SEEMED TO BE DISSIPATING. I EXPLAINED WE WERE RETURNING TO ZZZ AND TOLD HER TO CALL ME IF THERE WAS INCREASING SMOKE OR SMELL, OTHERWISE WE WOULD CONTACT GND. WE CLEANED UP TO FLAPS 5 DEGS, 180 KT SPD FOR TRIP AROUND PATTERN. CALLED OPS ON RADIO FOR QUICK MESSAGE TO DISPATCH AND THEN PROCEEDED WITH AFTER TKOF AND APCH
TO LNDG CHKLISTS. WE WERE IN POS TO LAND ABEAM LNDG THRESHOLD (DUE TO TWR'S NICE VECTORS) BUT DECIDED TO FLY NORMAL BASE/FINAL AS THERE WAS NO CONTINUING SMOKE/SMELL.
WHILE COMING AROUND THE PATCH WE HEARD ANOTHER ACFT RPT DEBRIS ON RWY XX NEAR OUR ROTATION POINT. WE LANDED ON RWY XXL, EXITED TO THE L AND STOPPED. EMER VEHICLES INSPECTED THE AIRPLANE. I WAS CONCERNED FOR SIGNS OF UNUSUAL BURNING OR A HYD LEAK

Highlights (43): sound* nois* hum* vibr* avm evm buffet* shud* thump* buzz* bang* rumbl* surg* stall* egt eicas ecam see spot* saw smell* odor smok* thud* bird

FUTURE WORK

❑ Compare the use of a wider array of text analytic tools on the dataset in order to evaluate the performance of In-Spire and Starlight as VA text analytic tools.

ACKNOWLEDGEMENT

❑ Roger Nicholson, Boeing Aviation Safety Associate

❑ The Boeing Company