INTEGRATED VEHICLE HEALTH MANAGEMENT

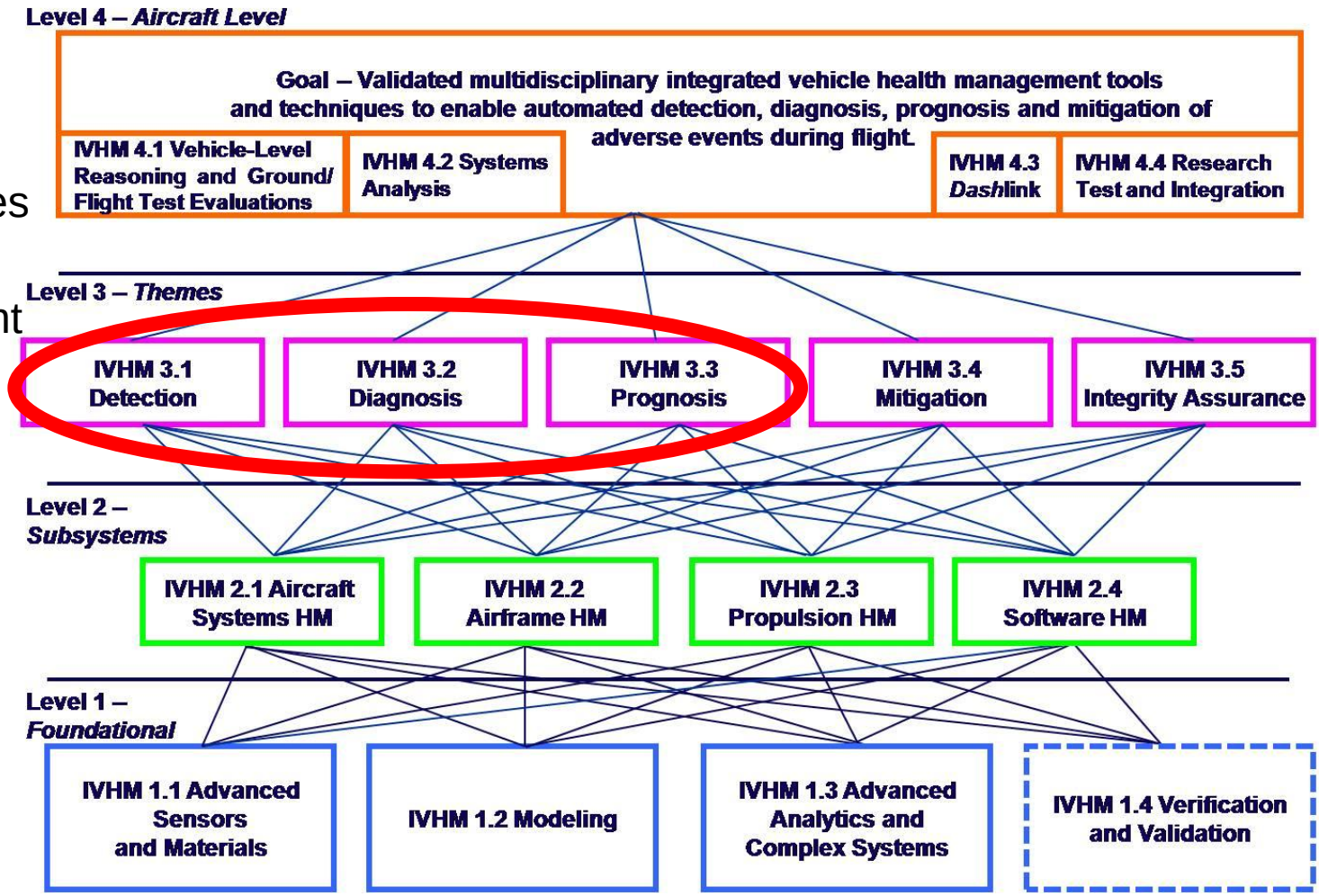
Event Cube: An Organized Approach for Mining and Understanding Anomalous Aviation Events

PI Name: Jiawei Han

Aviation Safety Program Technical Conference
November 17-19, 2009
Washington D.C.

Outline

- Problem Statement
- IVHM milestones addressed
- Approach: Event Cube
- Results
- Conclusions
- Future Plans



Problem Statement

Aviation Safety Reporting System

How to mine ASRS for analyzing patterns and causes of aviation anomalies?

Light	Anomaly	Phase	Report
Night	equipment : critical	intermediate altitude	…we hit what appeared to be a flock of large white birds. Some of them seem to be ducks..
Daylight	equipment : critical	intermediate altitude	…our flight encountered a small flock of ducks…
Daylight	equipment : less severe	ground: parked	…there were no bird strikes noticed while taxiing..
…	…	…	… …

- 40,000 reports in 2006
- 3,300 reports /month
- 160+/working day

- What are possible causes of landing problems? (**cause analysis**)
- Is there any difference in causal factors for landing at night and daylight time? (**comparative analysis**)
- Where did landing problems most often happen? (**ranking**)

….

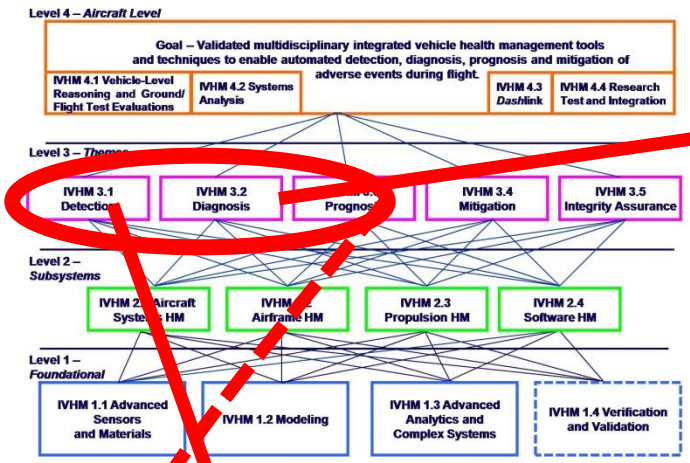
Source: http://asrs.arc.nasa.gov/overview/summary.html

The EventCube Project

- **General mission:** Develop advanced technology for mining text data and interrelated, multidimensional datasets for effective and efficient IVHM

- **Approach:**
  - Develop Event Cube technology for effective analysis of causal and contributing factors in aviation anomalies
  - Develop a set of data mining and text analysis methods for systematic, multidimensional analysis of ASRS database
  - Further extend such event cube and text mining methods for effective analysis of other text data sets of NASA's interest

IVHM Milestones Addressed



3.2.4 **An automated capability to <u>diagnose the causal factors</u> of anomalous operations in real or emulated data of large, fleet-wide or airspace heterogeneous data sources.**

3.1.3 **An anomaly detection method that has the ability to <u>detect</u> at least 3 known anomalies in real or emulated data of large, fleetwide heterogeneous data sources.**

Potentially,

3.3.4 **Forecasting technology that has the ability to <u>predict</u> at least 3 known anomalies in real or emulated data of large, fleetwide heterogeneous data sources.**

Approach: Event Cube

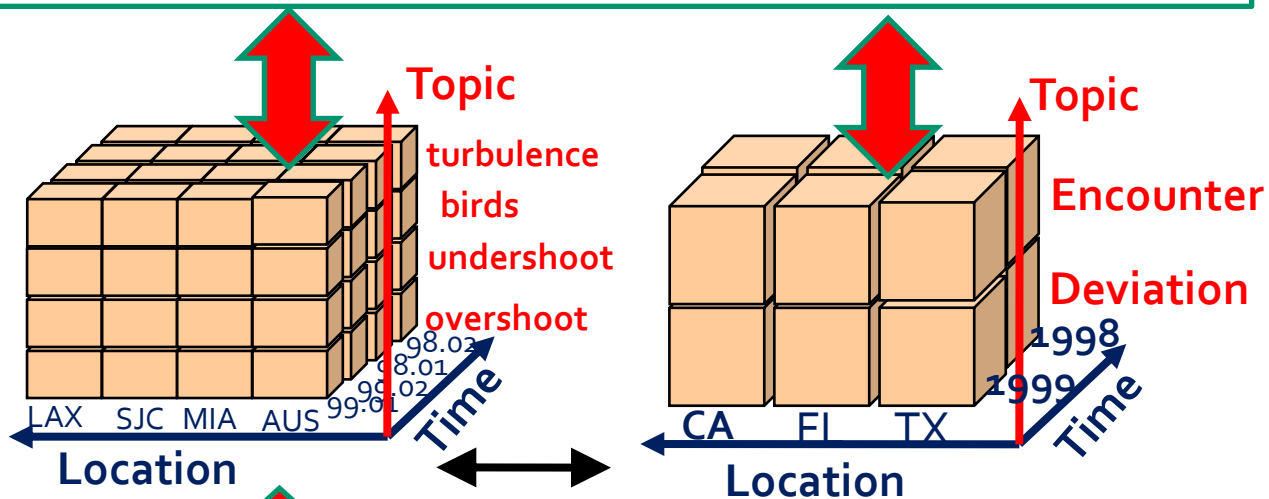
Analyst

. . .

Analysis Support

Multidimensional OLAP, Ranking, Cause Analysis, Topic Summarization/Comparison ……

Event Cube Representation



Multidimensional Text Database

Light	Anomaly	Phase	Report
Night	equipment : critical	intermediate altitude	...we hit what appeared to be a flock of large white birds. Some of them seem to be ducks..
Daylight	equipment : critical	intermediate altitude	...our flight encountered a small flock of ducks...
Daylight	equipment : less severe	ground: parked	...there were no bird strikes noticed while taxiing..
---	---	---	--- ---

Selected Research Results

1. Topic/Text Cube for comparative analysis
2. Cell ranking and summarization
3. Causal factor annotation

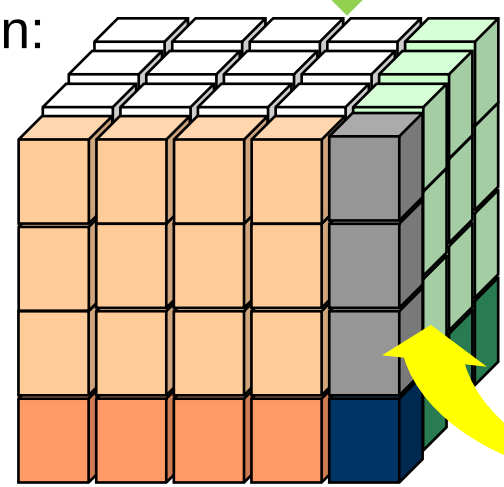
1. Text/Topic Cube: General Idea

- Aggregating text data (ASRS reports) in subspaces
- Heterogeneous: categorical attributes + unstructured text

ACN	Time	Location	Place	Environment	… …	Event	Report
							Text data

- How to combine?
- Our solution:

Cells

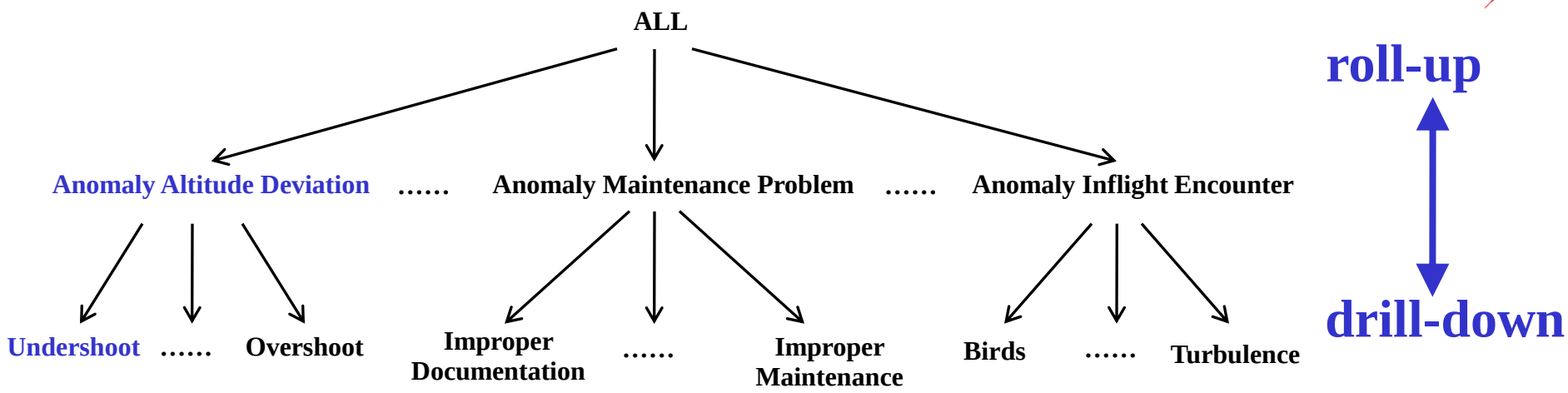


Measure

Text/Topic Model: Unstructured Text

Term/Topic	Weight
altimeter	0.01
leveling	0.008
leveloff	0.007
…	…

Topic Cube: Multidimensional Text Analysis by Topic Modeling



Time	Loc	Env	…	Narrative
98.01	TX	Daylight	…	
98.01	LA	Daylight	…	
98.01	LA	Night	…	
99.02	FL	Night	…	

Altitude	0.03
Ft	0.02
Climb	0.01
…	….

Descent	0.06
Cloud	0.03
Ft	0.01
…	….

Altitude	0.04
Ft	0.03
Instruct	0.01
…	….

Descent	0.05
System	0.02
View	0.01
…	….

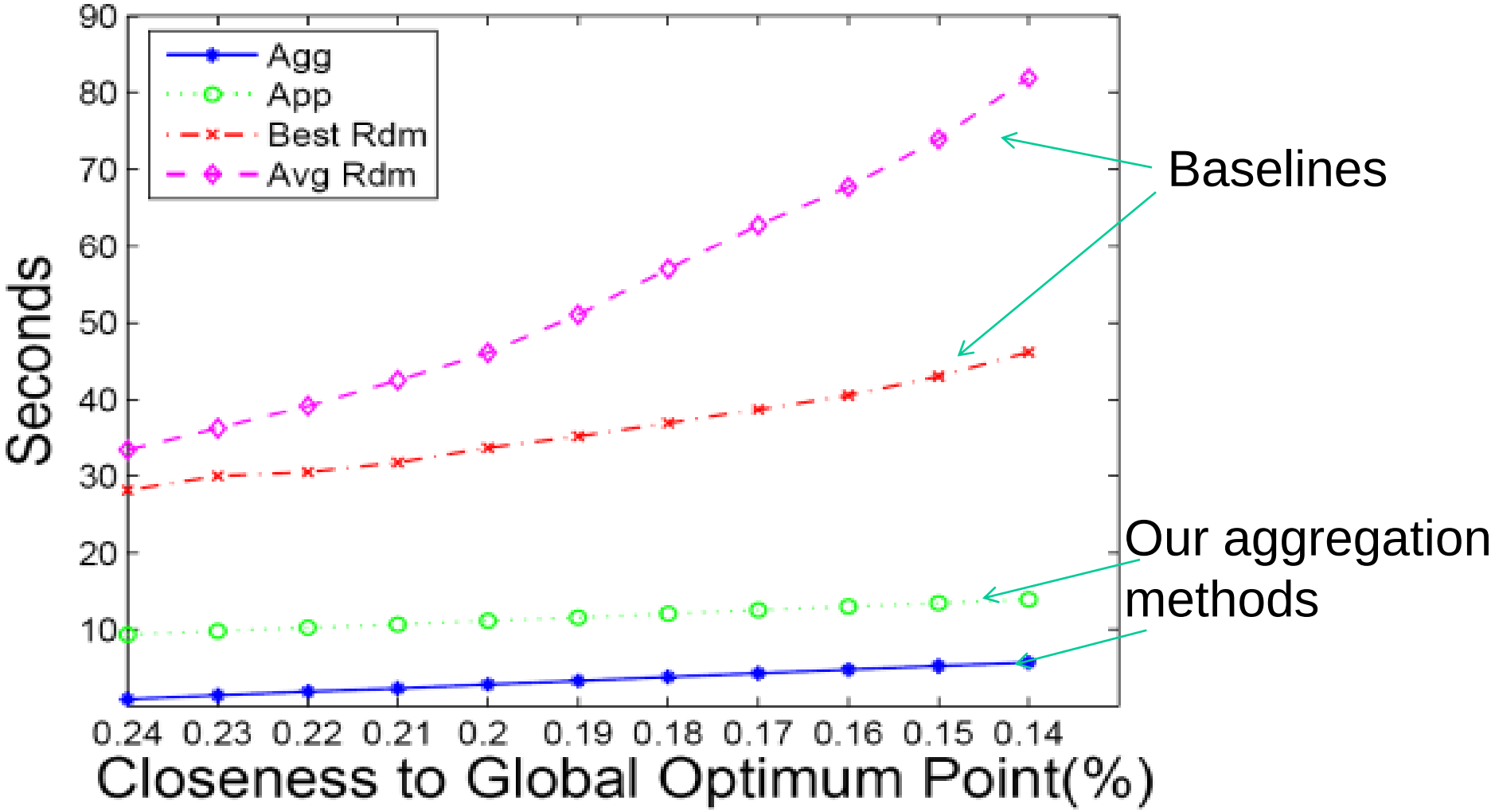
Text/Topic Cube: A Powerful Tool

- Classification, clustering and summarization of each cell
  - Cluster the documents in one cell to reveal aspects of anomalous events
  - Summarization of documents in a cell
- Comparative analysis of anomalous events in contexts, e.g.,
  - In what kind of weather condition, does a flight tend to have the problem of landing without clearance?
  - In Place A and Time B, which kind of anomaly event occurs most often?
- Keyword search in text cube
  - Ranking relevant cells
  - Ranking cuboids (in the order of discriminativeness of keywords)
- Cause analysis based on text/topic cube
  - E.g. what are the main causal factors of each anomalous event in different situations?
- ...

How can we efficiently do all these at large-scale?

Aggregation Speeds Up TopicCube Construction

Time for constructing cell (1999, CA, *) with 629 documents



Sample Result : Topic Cube for Topic Content Comparison

landing without clearance: daylight vs. night

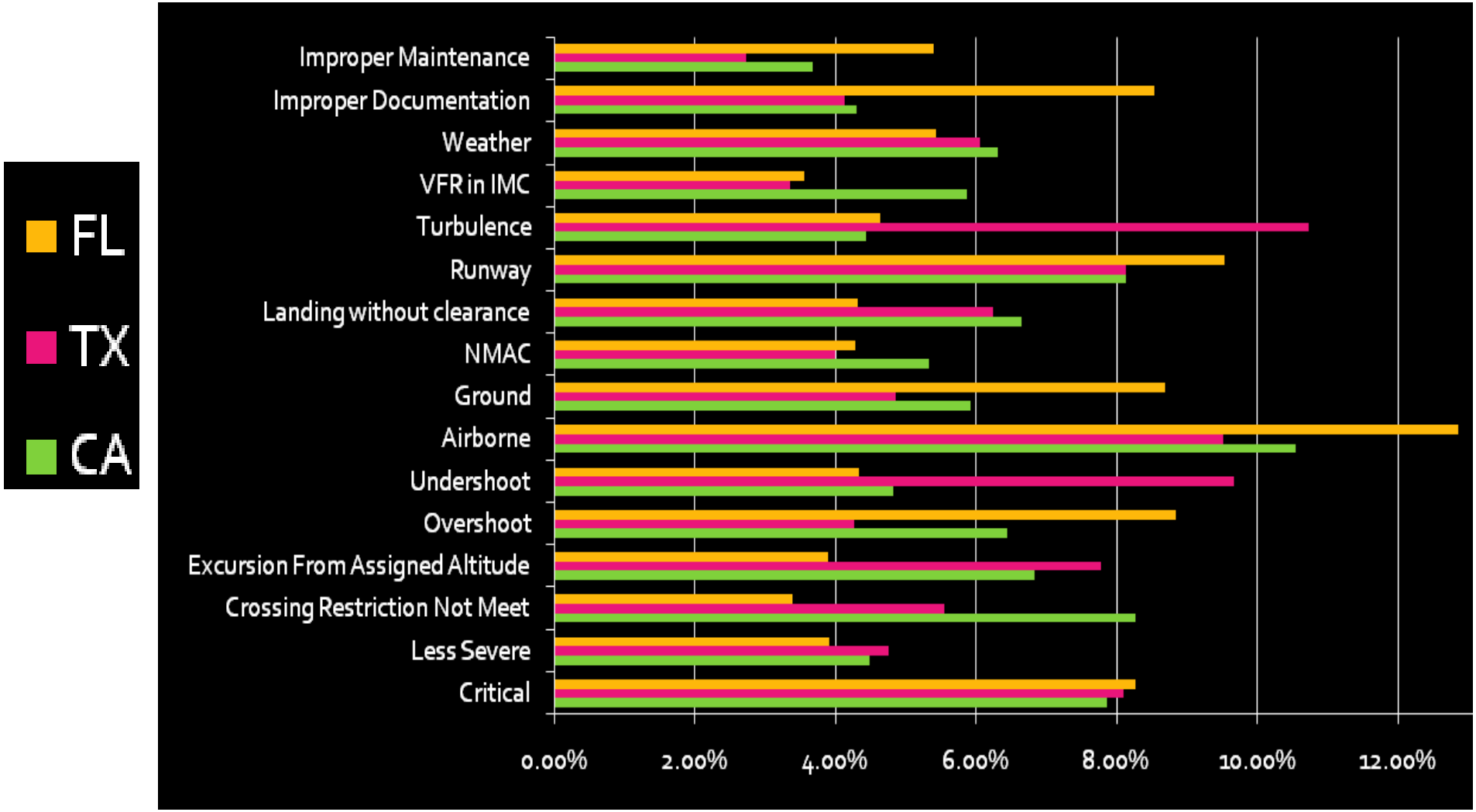
Context	Word	p(w|θ)
daylight	Tower	0.075
	Pattern	0.061
	Final	0.060
	Runway	0.053
	Land	0.052
	Downwind	0.039
night	Tower	0.035
	Runway	0.029
	Light	0.027
	Instrument Landing System	0.015
	Beacon	0.014

…WINDS ALOFT AT PATTERN ALT OF 1000 FT MSL, WERE MUCH STRONGER AND A DIRECT XWIND. NEEDLESS TO SAY, THE PATTERNS AND LNDGS WERE DIFFICULT FOR MY STUDENT AND THERE WAS LIGHT TURB ON THE DOWNWIND…

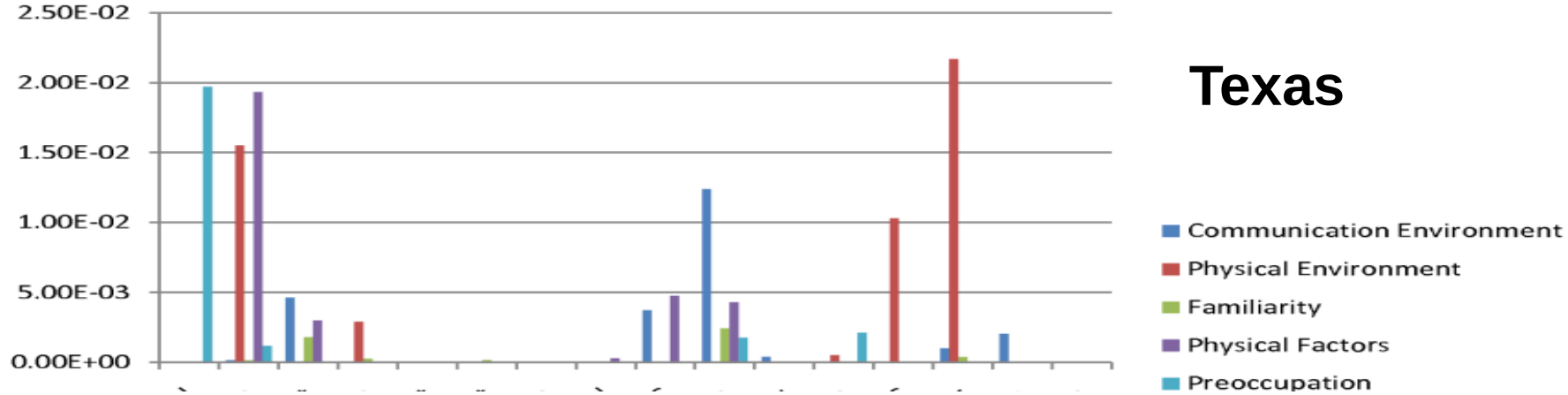
…I LISTENED TO HWD ATIS AND FOUND THE TWR CLOSED AND AN ANNOUNCEMENT THAT THE HIGH INTENSITY LIGHTS FOR RWY 28L WERE INOP. BROADCASTING IN THE BLIND AND LOOKING FOR THE TWR BEACON AND LOW INTENSITY LIGHTS AGAINST A VERY BRIGHT BACKGROUND CLUTTER OF STREET LIGHTS, ETC…

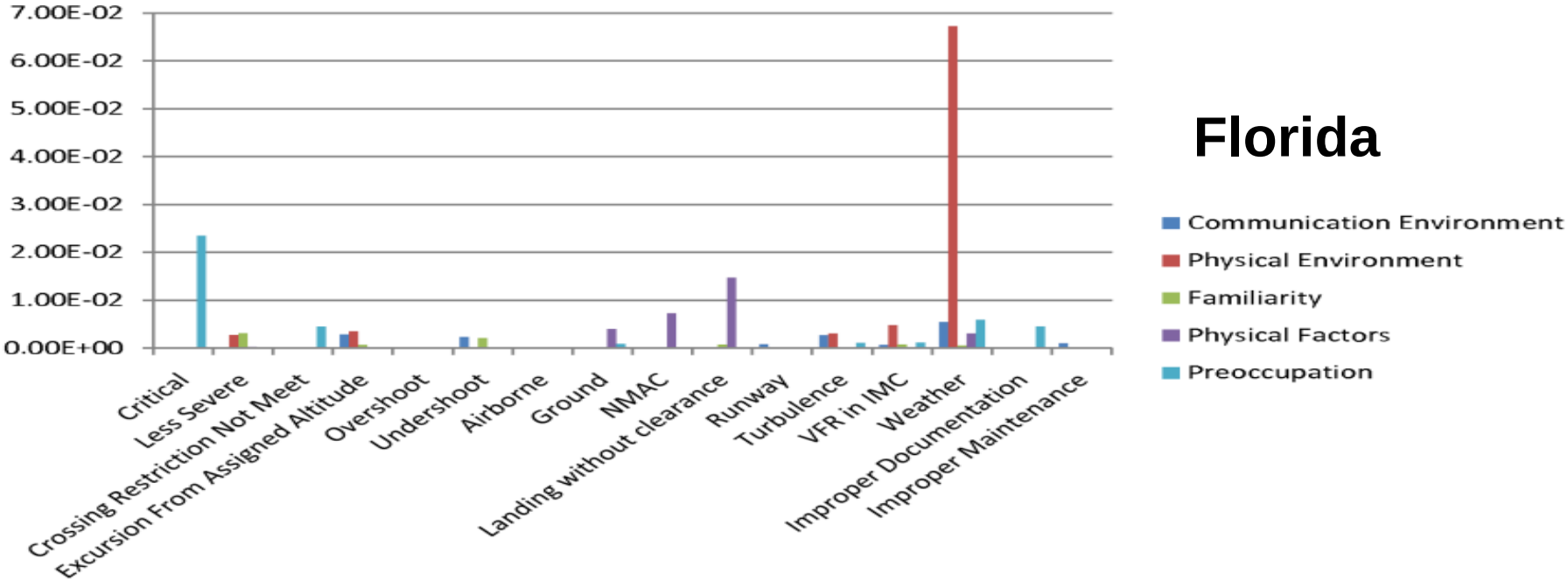
Sample Result: TopicCube for Topic Coverage Comparison

Comparison of distributions of anomalies in FL, TX, and CA



Comparative Analysis of Shaping Factors



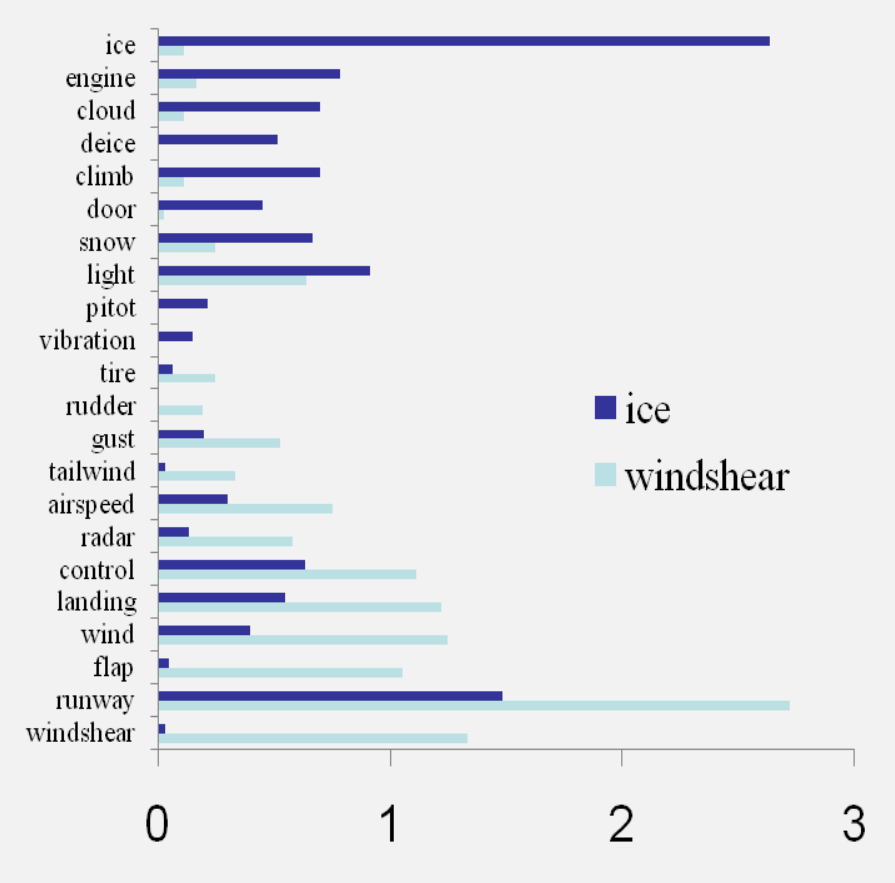


Text Cube for Comparative Analysis of Sub-Events

Comparison of Average TF of Words

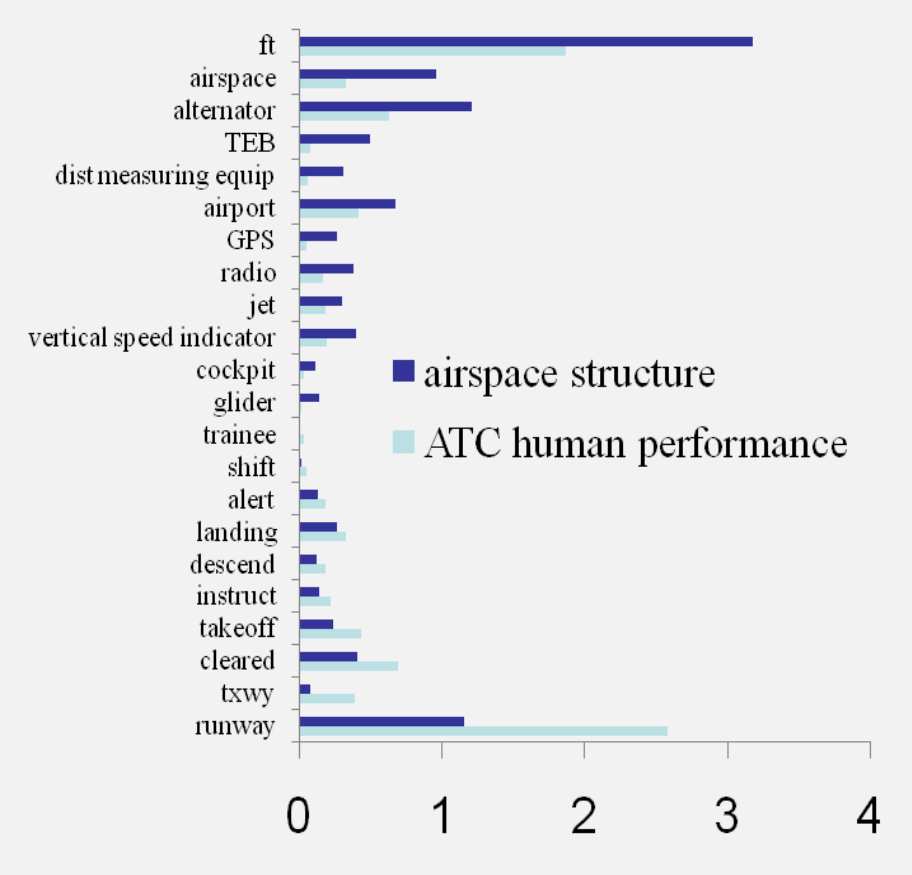
"ice" vs. "windshear" in

"*Environment: Weather Elements*"



"airspace structure" vs. "ATC human performance" in

"*Supplementary: Problem Areas*"



Leverage Sequential Pattern Mining: More Meaningful Units

- Anomaly1 = aircraft equipment problem : critical
- Anomaly2 = inflight encounter : weather
- Anomaly3 = conflict : nmac

Pattern	Support		
	Anomaly1	Anomaly2	Anomaly3
LNDG UNEVENTFUL	11	0	0
LANDED WITHOUT INCIDENT	12	0	0
SHUT DOWN ENG	12	0	0
VISIBILITY FOG	0	13	0
CEILING VISIBILITY	0	15	0
DOWNWIND RWY	0	0	12
SAW OTHER ACFT	0	0	10
CLRED FOR RWY	0	0	44
TOOK EVASIVE ACTION	0	0	44
SUPPLEMENTAL FROM	17	10	31
CALLBACK WITH REVEALED FOLLOWING	37	13	24
CALLBACK WITH REVEALED FOLLOWING HAT	13	0	0

Selected Research Results

1. Topic/Text Cube for comparative analysis
2. Cell ranking and summarization
3. Causal factor annotation

2. Cell Ranking and Summarization

Ranking

- Find critical anomaly and flight phases related to birds by searching flight reports with keyword query: “bird”, “geese”, “duck”

Light	Anomaly	Phase	Report
Night	equipment : critical	intermediate altitude	…we hit what appeared to be a flock of large white **bird**s. Some of them seem to be **duck**s..
Daylight	equipment : critical	intermediate altitude	…our flight encountered a small flock of **duck**s…
Daylight	equipment : less severe	ground: parked	…there were no **bird** strikes noticed while taxiing..
…	…	…	… …

Cell (Light=Daylight)

Cell (Anomaly=equipment : critical, Phase=intermediate altitude)

- **Traditional IR:** Ranking all documents
- **Our Approach**: Ranking all cells in the order of relevance given keyword query

Summarization of Cell Content

- A fundamental challenge in Online Analytical Processing (OLAP) of multidimensional text database is to summarize the content in its text cells.
  1. Neutral Summarization

     Give the most representative documents within a text cell
  2. Topic-biased Summarization

     Give the most relevant documents to a query within a text cell that also cover the content of the text cell well

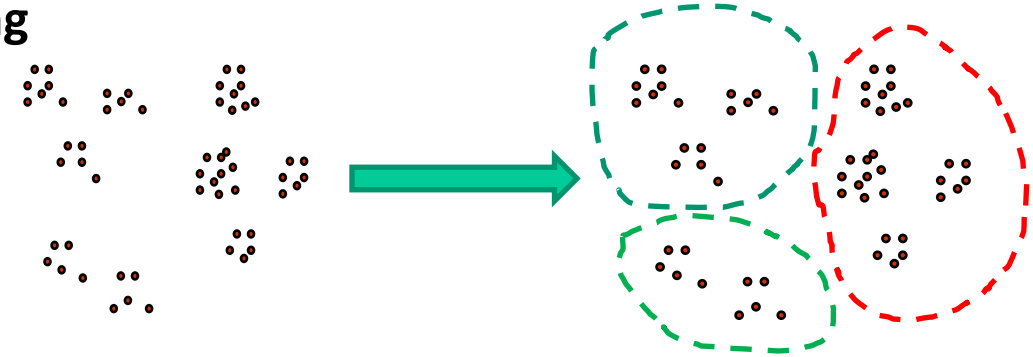
Table 1: An example of text database in ASRS

ACN	Time	Airport	···	Light	Narrative
101285	199901	MSP	···	Daylight	Document 1
101286	199901	CKB	···	Night	Document 2
101291	199902	LAX	···	Dawn	Document 3

- For example:
  1. What are the main points made in those reports about anomalies during night in Jan. 1999? (neutral)
  2. What have the pilots said about landing at LAX in 1999? (topic-biased: topic="landing")

Micro Text Clustering

Micro Clustering



Neutral Summarization

…so that if we saw the ARPT, we could land…

…due to stronger than forecasted winds and weather going…

…resulted in RWY excursion during engine fail…

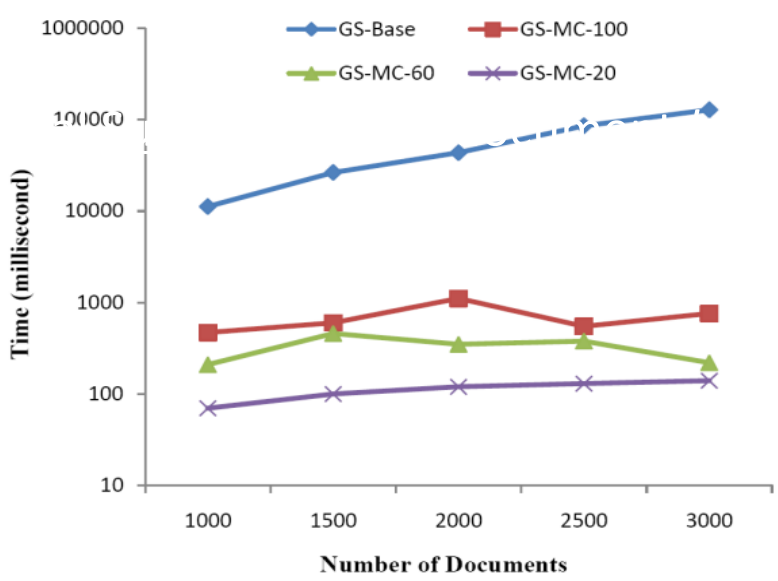
Topic-biased Summarization

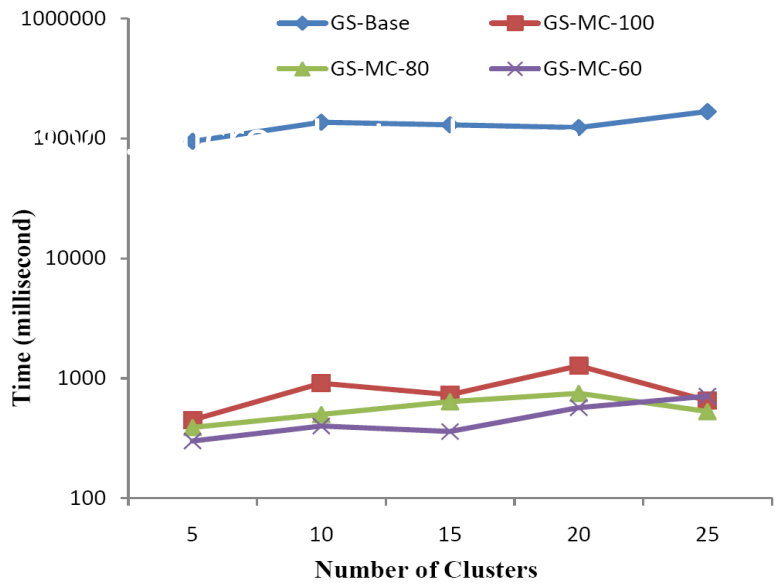
…so that if we saw the ARPT, we could **land**…

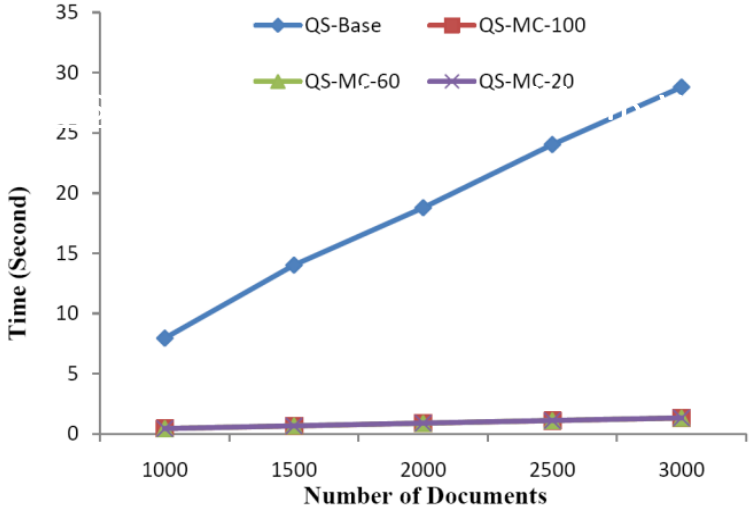
…after an hour, the weather had not much improved which forced us to **land**…

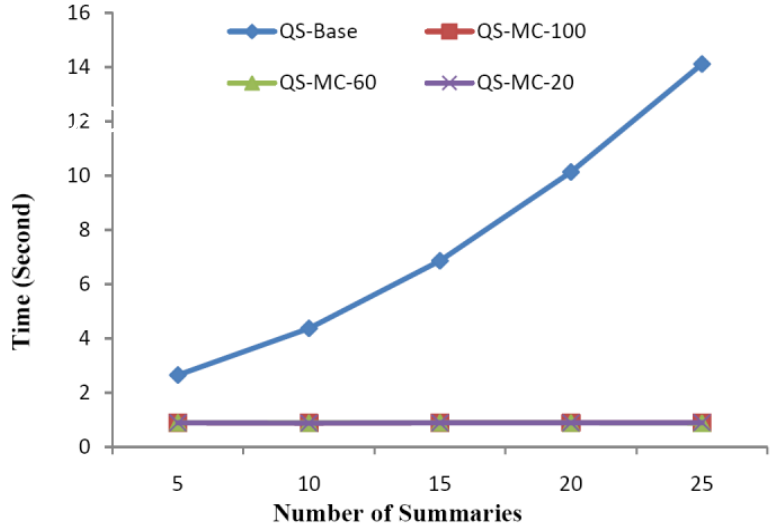
…SMA engine failure, forced **landing** at LGB by instructor…

Micro-Text-Cluster is More Efficient than Direct Summarization









Selected Research Results

1. Topic/Text Cube for comparative analysis
2. Cell ranking and summarization
3. Causal factor annotation

3. Causal Factor Annotation

- Goal
  - Identify the cause(s) of an incident described in an ASRS report narrative from a set of 14 predefined causes, or shaping factors.
- Challenges
  - Reports are informally written (acronyms, abbreviations, and typos)
  - Reports can have multiple labels (shaping factors)
  - Labeled data is scarce
  - Skewed class distributions: 10 of the 14 classes account for 25% of labels

Approach 1: Semantic Lexicon Construction

- Motivation: insufficient annotated data implies that it is difficult to identify relevant words for each shaping factor
- Idea: learn words and phrases related to each shaping factor in a bootstrapping manner.
- Experimental Results:
  - Both rule-based and learning-based approaches outperform the baseline.
  - The learned words are useful features.
  - The best result of 52.7% illustrates the difficulty of the task.

Approach 2: Training Set Expansion

- Motivation: insufficient annotated data
- Idea: develop a bootstrapping algorithm to augment the labeled data in an iterative fashion.
- Repeat the following in each iteration:
  - augment either the positively labeled data or the negatively labeled data, depending on which set is smaller
  - find the word w that best discriminates the positives and the negatives based on the currently labeled data
  - label all unlabeled documents that contain w with the corresponding label
- Experimental Results:
  - This approach is specially effective in improving the F-score for the 10 minority (i.e., under-represented) shaping factors, where positive instances are scarce.

Words Chosen by Bootstrapping Algorithm

Shaper	Positive Expanders	Negative Expanders
Familiarity	unfamiliar, layout, unfamiliarity, rely	
Physical Environment	cloud, snow, ice, wind	
Physical Factors	fatigue, tire, night, rest, hotel, awake, sleep, sick	declare, emergency, advisory, separation
Preoccupation	distract, preoccupied, awareness, situational, task, interrupt, focus, eye	declare, ice, snow, crash, fire, rescue, anti, smoke
Pressure	bad, decision, extend, fuel, calculate, reserve, diversion, alternate	

Conclusions

- EventCube: A framework for mining and analyzing ASRS database to help achieve three IHVM milestones: Detection, Diagnosis, and Prognosis
  - Unification of relational data mining and text mining
  - Integrative analysis of ASRS data
- We proposed algorithms for constructing specific EventCube instances:
  - Text Cube and Topic Cube support OLAP on text dimension and comparative analysis of anomalies
  - Sequential pattern mining enhances analysis with pattern units
- We developed algorithms for interactive analysis:
  - Multidimensional keyword search: TopCells
  - Multidimensional cell summarization: MiTexCluster
- We developed algorithms for automatic analysis
  - Causal factor analysis
  - Data stream classification
- We also developed a prototype system

System Demo: Text Cube and Topic Cube

Different word distributions reflect different content in the two cells

Cell 1: (Weather="fog", *)

Term	Probability
runway	0.022331
approach	0.018184
feet	0.017086
aircraft	0.012505
flight	0.007746
tower	0.007044
weather	0.006380
time	0.006355
mile	0.006087
airport	0.005806
visibility	0.005640

Cell 2: (Weather="Turbulence", *)

Term	Probability
feet	0.015784
aircraft	0.015541
flight	0.012476
flight level	0.009766
turbulence	0.009297
runway	0.009082
approach	0.008561
altitude	0.007906
time	0.006232
air traffic control	0.006059
weather	0.005340

System Demo: Text Cube and Topic Cube

Three sample topics in a cell

Topic #33

Term	Probability
flight	0.014851
turbulence	0.013191
aircraft	0.012837
feet	0.011509
approach	0.009760
runway	0.008078
kts	0.007171
air traffic control	0.007104
weather	0.007016
flight level	0.006197
captain	0.006064
degs	0.005644

Topic #34

Term	Probability
feet	0.023412
visual flight rules	0.017479
flight	0.015342
weather	0.014002
approach	0.013109
instrument flight rules	0.011738
clouds	0.009824
mile	0.009601
zzz	0.009250
conditions	0.008867
airport	0.008325

Topic #35

Term	Probability
wake	0.021715
aircraft	0.019422
runway	0.015918
feet	0.013929
turbulence	0.013453
approach	0.009560
flight	0.009387
degs	0.008435
air traffic control	0.008089
turbulence	0.006618
time	0.006186
kts	0.006056

System Demo: Cell Ranking

Enter key words (space separated) below:

Key Words: excursion [Search]

The top ranked cells are:

Top Cells for keyword query="*excursion*"

Rank	Year	State	Person	Weather	Light	Make/Model	Flight Phase	Primary Area	Event Anomaly	Resolutory Action	Score
1	2000	*	*	Rain	Night	*	landing : roll	*	aircraft equipment problem : critical	*	34.3084299
2	*	*	*	*	Daylight	Boeing	cruise : enroute altitude change	Flight Crew Human Performance	altitude deviation : overshoot	controller : issued advisory	33.3752199
3	*	*	*	*	*	McDonnell Douglas	*	Airport	excursion : taxiway	none taken : anomaly accepted	33.0382225
4	*	OH	*	*	*	*	*	*	non adherence : far	flight crew : returned to original clearance	32.5993384

System Demo

Please visit me!

http://infonetcube.cs.uiuc.edu/nasa/

Future Work

- Further enhance EventCube through
  - Multidimensional cause analysis/classification/prediction
  - Multidimensional data stream management/analysis
  - Construction and analysis of information networks (entities, relations)
  - Natural language processing and text mining for causal analysis
- Closer collaboration with aviation experts to evaluate our prototype system and obtain feedback

Research Publications (Led by UIUC Group)

- Bolin Ding, Bo Zhao, Cindy Xide Lin, Jiawei Han, and Chengxiang Zhai, "*TopCells: Keyword-Based Search of Top-k Aggregated Documents in Text Cube*", to appear in Proc. 2010 Int. Conf. on Data Engineering (ICDE'10).
- Duo Zhang, Chengxiang Zhai, Jiawei Han, Nikunj C. Oza, Ashok N. Srivastava, "*Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases*", Journal version invited to special issue of SDM'09 (to appear)
- Cindy Xide Lin, Bolin Ding, Jiawei Han, Nikunj C. Oza, Ashok N. Srivastava, Bo Zhao, Feida Zhu, "*TextCube: Computing IR Measures for Multidimensional Text Database Analysis*", submitted to IEEE Trans. on Knowledge and Data Engineering, 2009
- Duo Zhang, Chengxiang Zhai and Jiawei Han, "*Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases*", Proc. 2009 SIAM Int. Conf. on Data Mining (SDM'09), Sparks, NV, April 2009. (**Best of SDM'09**)
- Bolin Ding, David Lo, Jiawei Han, and Siau-Cheng Khoo, "Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database", Proc. 2009 Int. Conf. on Data Engineering (ICDE'09), Shanghai, China, Mar. 2009.
- Feida Zhu, Xifeng Yan, Jiawei Han and Philip S. Yu, "*Mining Frequent Approximate Sequential Patterns*", in H. Kargupta, et al., (eds.), Next Generation of Data Mining, Chapman & Hall, 2009.
- Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, and Bo Zhao, "*Text Cube: Computing IR Measures for Multidimensional Text Database Analysis*", Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.
- Luiz Mendes, Bolin Ding, and Jiawei Han, "*Stream Sequential Pattern Mining with Precise Error Bounds*", Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.
- Jing Gao, Bolin Ding, Wei Fan, Jiawei Han, and Philip S. Yu, "*Classifying Data Streams with Skewed Class Distribution and Concept Drifts*", IEEE Internet Computing (Special Issue on Data Stream Management), 12(6): 37-49, 2008
- Chen Chen, Cindy Xide Lin, Xifeng Yan, Jiawei Han, "*On Effective Presentation of Graph Patterns: A Structural Representative Approach*", Proc. 2008 ACM Conf. on Information and Knowledge Management (CIKM'08).

Research Publications (Led by UTD Group)

- Muhammad Arshad Ul Abedin, Vincent Ng, and Latifur Rahman Khan, "*Weakly Supervised Cause Identification from Aviation Reports via Semantic Lexicon Construction*", submitted to Journal of Artificial Intelligence Research, Feb. 2009.
- Isaac Persing and Vincent Ng, "*Semi-Supervised Cause Identification from Aviation Safety Reports*", in Proc. of the Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), Singapore, Aug. 2009.
- Clay Woolam, Latifur Khan, "*Multi-concept Document Classification Using a Perceptron-Like Algorithm*," 2008 IEEE / WIC / ACM Int. Conf. on Web Intelligence (WI'08), Dec. 2008, Sydney, NSW, Australia, Page: 570-574.
- Clay Woolam, Latifur Khan, "*Multi-label large margin hierarchical perceptron*," Int. Journal of Data Mining, Modelling and Management (IJDMMM), 1(1): 5-22 (2008), Interscience Publisher (invited paper).
- Mohammad M Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "*A Multi-Partition Multi-Chunk Ensemble Technique to Classify Concept-Drifting Data Streams*", Proc. 2009 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'09), Bangkok, Thailand, Apr. 2009.
- Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "*A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data*", Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.