Solicitation Number:

PROPOSAL COVER PAGE

1. SUBMIT TO:
Director
U.S. Army Research Office
ATTN: AMSRL-RO-RI
P.O. Box 12211
Research Triangle Park, NC 27709-2211

2. For consideration by:
☐ Biology/Life Sci ☐ Materials
☐ Chemistry ☒ Mathematics/Statistics
☒ Computer Science ☐ Physics
☐ Electronics ☐ Comp & Info Sci
☐ Mechanical ☐ Wepons & Mtls Sci
☐ Environmental Sciences ☐ Human Rsch & Eng
☐ Sensors & Electron Dev ☐ Surv/Lethality

3. Is this proposal being submitted to another Federal Agency?
☒ No ☐ Yes If Yes, list the agency:

4. Is applicant delinquent on any Federal Debt?
☐ Yes (Attach explanation) ☒ No

5. Proposal Valid Until (min of 6 mos):
12 Months

6. Entity Identification Number (EIN) or Taxpayer Identification Number (TIN)
54-0836354

7. Data Universal Numbering System (DUNS No.):
07-781-7450

8. Commercial and Government Entity (CAGE) Code:
7X764

9. Name of organization to which award should be made:
George Mason University
4400 University Drive, MS 4C6
Fairfax, VA 22030

10. Administrative Address of Organization (if different):

11. Branch/Campus/Other Component (where work is performed, if different): Farifax, VA

12. Submitting Organization's Contract/Grant Administration Office:
George Mason University
Office Sponsored Programs
4400 University Drive, MS 4C6
Fairfax, VA 220303

13. Submitting Organization's Audit Office:
George Mason University, Interal Audit Office
4400 University Drive MSN 1A2
Fairfax, Virginia 22030-4444

14. Submitting Organization: (Check all that apply)
☐ For Profit: ☐ Large ☐ Small ☐ Disadvantaged ☐ 8a ☐ Women-Owned ☐ Foreign ☐ Individual
☒ Educational: ☐ HBCU ☐ Minority Institution ☐ Hispanic ☐ Indian Tribal ☒ State ☐ Private ☐ Foreign ☐ FDP
☐ Hospital: ☐ Public ☐ Private ☐ Nonprofit ☐ For Profit
☐ Nonprofit
☐ Not-For-Profit
☐ Other (Specify)

15. Check appropriate box(es) if this proposal includes any of the items listed below:
☐ Human Subjects ☐ Biosafety Level (BL) 1-4 Facility
☐ Vertebrate Animals ☐ Genetically Engineered Organisms
☐ National Environment Policy Act ☐ Limited Rights Data
☐ Disclosure of Lobbying Activities ☐ Unlimited Rights
☐ Historical Places ☐ Govt Purpose Rights Data
☐ GFE ☐ GFD ☐ Proprietary Data
☐ GFI ☐ GFP ☐ Ozone Depleting Substances

16. Proposed Amount:
$100,000.00

17. Proposed Duration (1-60 mos):
12 months

18. Proposed Start Date:
9/1/06

19. Type of Award Proposed:
☒ Single Investigator
☐ Young Investigator Program
☐ Short Term Innovation Rsch
☐ Research Instrumentation
☐ Conference/Symposia
☒ Other (Specify):
Grant

20. Title of Proposed Project: Adaptive Multi-modal Data Mining & Fusion for Autonomous Intelligence Discovery

21. Principal Investigator (PI)/Project Director (PD) Department and Postal Address:
Edward Wegman, Ph.D.
George Mason University, College of Science
4400 University Drive, Fairfax, VA 22030

22. Year PI's degree conferred
1968

23. Scientific discipline of PI's degree
Statistics/Computer Science

TYPED NAMES	TELEPHONE NUMBER	FACSIMILE NUMBER	ELECTRONIC MAIL ADDRESS
24. PI/PD Edward Wegman, Ph.D.	703-993-1691	703-993-1700	ewegman@gmu.edu
25. CO-PI/PD			
26 a. Primary Administrative representative Authorized to Conduct Negotiations: Ann T. McGuigan, Ph.D.	703-993-2988	703-993-2296	amcguiga@gmu.edu
26 b. Alternate Administrative Representative Authorized to Conduct Negotiations: Karen G. Cohn	703-993-4104	703-993-2296	kcohn@gmu.edu

27 a. Authorized Representative Signing for Applicant Organization: Karen L. Cohen	27 c. By signing and submitting this proposal, the Offeror is providing the certifications contained in this BAA.
27 b. Title: Associate Director, Office of Sponsored Programs	27 d. Signature Date:

Form 51 (REV Jun 05)

Adaptive Multi-modal Data Mining and Fusion For Autonomous Intelligence Discovery

My proposal addresses the challenges of autonomous discovery and triage of contextually relevant information in massive, complex, dynamic text and imagery streams. I will develop a prototype system to mine, filter and fuse multi-modal data streams and dynamically interact with the analysts to improve their efficiency through feedbacks and autonomous adaptation of the algorithms. I plan to implement four core capabilities:

- Text and image mining for feature extraction
- Multi-modal data fusion
- Agent-based adaptive information filtering
- Cognitively friendly information visualization.

Together, these will enhance the capabilities of the analysts to discover, assess, and act on embedded intelligence in near real-time. Figure 1 shows an overview of the architecture of the prototype I plan to develop.

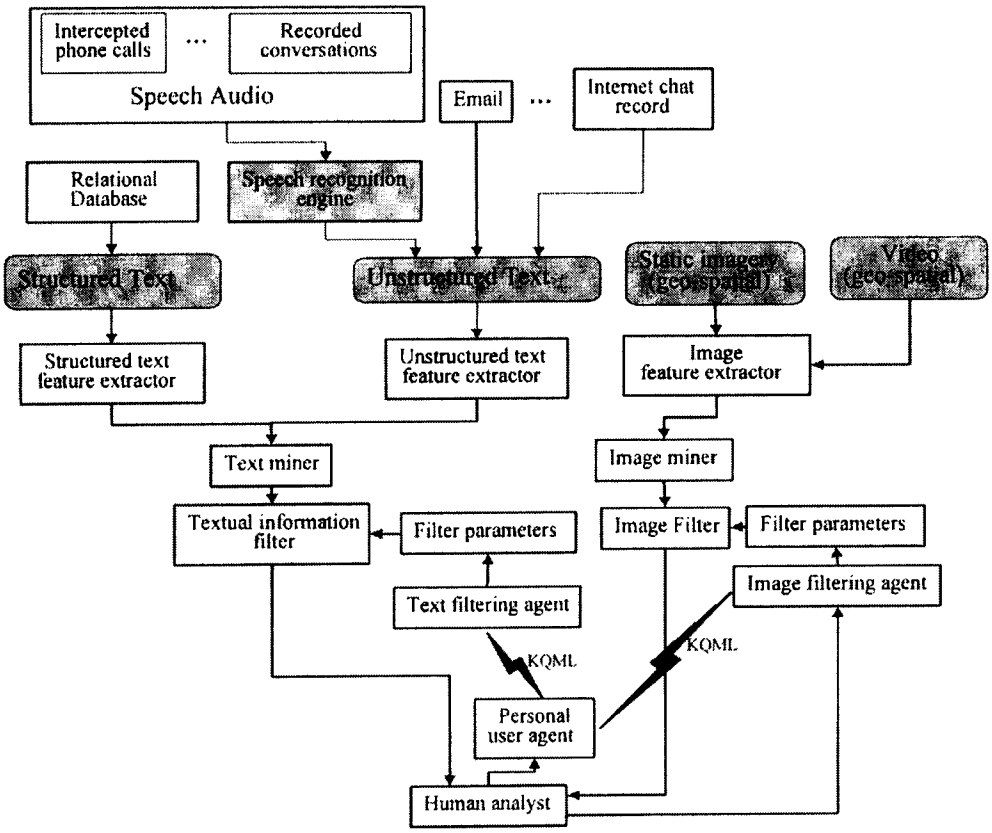

Figure 1. System architecture of the proposed prototype.

Research Objectives

I have two broad classes of research objectives. The first is development of automatic data mining, reduction, fusion and visualization of data streams. The second is development of a co-operative, multi-agent framework to autonomously direct the former in response to analyst behavior. This mirrors the flow of data from fine-grained, massive, raw input requiring the most automation, to the most highly processed, cognitively friendly output presented to analysts as visualizations. The agents filter results so that human analysts are presented with contextually relevant information, and provide a feedback to the data mining algorithms so they can adapt appropriately.

Data Mining, Reduction, Fusion and Visualization

First, I will use basic data mining of text and imagery separately to discover hidden patterns and unusual occurrences. I will extend this capability to determine and extract summary feature vectors from text documents and images, which will be accumulated into probability distribution estimates describing likelihoods of various phenomena. In turn, these provide standards against which to judge how unusual future phenomena are. Distributions also reduce data volume and constitute meta-features describing whole text or image streams, rather than their individual components. This hierarchy, with massive, raw data at the bottom linked to reduced distributions of stream features above, is useful because it allows selective drill-down analysis of raw data in response to associations or patterns discovered at the top. Moreover, I will research and implement prototype methods for fusing text and image stream distributions at the top level, and performing data mining there. Fusion of distributions is new, fundamental research that leverages the hierarchical paradigm, while mining at the top level expands on work currently being done at JPL to study climate phenomena in remote sensing data.

The research described above relies on solid data mining methods at bottom level. Here I will expand on previous work at GMU in both text and image mining. For text, the basic approach is to create feature vectors for each document or message that reflects semantic content. I first de-noise the text by deleting words that do not convey information, and stemming words by removing suffixes. Then, the document or message is summarized by accumulating the numbers of occurrences of word pairs (bigrams) or triplets (trigrams). These have some significant potential for capturing semantic content because they capture noun-verb pairs or adjective-noun-verb triplets. (See Martinez and Wegman, 2002 and 2003, Martinez et al. 2004, Solka et al. 2005). Our feature vector is the counts of bigram and trigram occurrences, called bigram or trigram proximity matrices (BPMs or TPMs). De-noising, stemming and summarizing can be done in a streaming fashion, which is crucial when dealing with high-volume, high-velocity text streams. I have shown (see the references above) that documents residing in separate corpora but having similar semantic content have similar feature vectors. A particular goal in this project is to identify and extract information about location and time of relevant events from message content. Note that speech audio data can be converted to text through commercially available speech-to-text software and then can be handled by the above text mining techniques.

A similar approach is commonly applied to image data. The so-called grey level co-occurrence matrix (GLCM) is analogous to the BPM. The idea is to look at pairs of pixels (each member assuming one of 256 grey levels) and create a 256 by 256 matrix to count the number of occurrences of grey levels pairs. Images that have similar GLCMs are expected to be similar with respect to characteristics implied by the geo-spatial relationship used to define the pair. A major research question is to determine pair definition functions that correspond to objects or phenomena of interest. This is quite analogous to determining which word pairs constitute bigrams of interest in text, but is more complex. A special case of interest to NGA is that found in persistent

surveillance. A sequence of images (video) of the same scene gives rise to a GLCM in time in which one can accumulate occurrence statistics (pixel by pixel if necessary) over different time durations. As with text, statistically significant changes can be identified using formal hypothesis tests with thresholds that adapt in response to analyst feedback.

References

Martinez, Angel R. and Wegman, Edward J. (2002) "A text stream transformation for semantic based clustering," *Computing Science and Statistics*, 34, 184-203.

Martinez, Angel R. and Wegman, Edward J. (2003) "Encoding of Text to Preserve 'Meaning'," *Proceedings of the Eighth U.S. Army Conference on Applied Statistics*, ACAS02/MartinezAngel/MartinezAngel.pdf.

Martinez, A. R., Wegman, E. J., and Martinez, W. L. (2004) "Using weights with a text proximity matrix," in *COMPSTAT 2004*, (Antoch, J., ed.), Berlin: Physica-Verlag, 327-338.

Solka, J. L., Bryant, A. C., and Wegman, E. J. (2005) "Text data mining with minimal spanning tree," *Handbook of Statistics: Data Mining and Data Visualization*, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 133-170.