

PROCEEDINGS OF SPIE

# ***Signal Processing, Sensor Fusion, and Target Recognition XXII***

**Ivan Kadar**  
*Editor*

**29 April–2 May 2013**  
**Baltimore, Maryland, United States**

*Sponsored and Published by*  
SPIE

**Volume 8745**

Proceedings of SPIE 0277-786X, V. 8745

SPIE is an international society advancing an interdisciplinary approach to the science and application of light.

Signal Processing, Sensor Fusion, and Target Recognition XXII,  
edited by Ivan Kadar, Proc. of SPIE Vol. 8745, 874501 · © 2013  
SPIE · CCC code: 0277-786X/13/\$18 · doi: 10.1117/12.2031900

Proc. of SPIE Vol. 8745 874501-1

The papers included in this volume were part of the technical conference cited on the cover and title page. Papers were selected and subject to review by the editors and conference program committee. Some conference presentations may not be available for publication. The papers published in these proceedings reflect the work and thoughts of the authors and are published herein as submitted. The publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Please use the following format to cite material from this book:

Author(s), "Title of Paper," in *Signal Processing, Sensor Fusion, and Target Recognition XXII*, edited by Ivan Kadar, Proceedings of SPIE Vol. 8745 (SPIE, Bellingham, WA, 2013) Article CID Number.

ISSN: 0277-786X

ISBN: 9780819495365

Published by

**SPIE**

P.O. Box 10, Bellingham, Washington 98227-0010 USA

Telephone +1 360 676 3290 (Pacific Time) -Fax +1 360 647 1445

SPIE.org

Copyright © 2013, Society of Photo-Optical Instrumentation Engineers.

Copying of material in this book for internal or personal use, or for the internal or personal use of specific clients, beyond the fair use provisions granted by the U.S. Copyright Law is authorized by SPIE subject to payment of copying fees. The Transactional Reporting Service base fee for this volume is \$18.00 per article (or portion thereof), which should be paid directly to the Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923. Payment may also be made electronically through CCC Online at [copyright.com](http://copyright.com). Other copying for republication, resale, advertising or promotion, or any form of systematic or multiple reproduction of any material in this book is prohibited except with permission in writing from the publisher. The CCC fee code is 0277-786X/13/\$18.00.

Printed in the United States of America.

Publication of record for individual papers is online in the SPIE Digital Library.



[SPIDigitalLibrary.org](http://SPIDigitalLibrary.org)

---

**Paper Numbering:** Proceedings of SPIE follow an e-First publication model, with papers published first online and then in print and on CD-ROM. Papers are published as they are submitted and meet publication criteria. A unique, consistent, permanent citation identifier (CID) number is assigned to each article at the time of the first publication. Utilization of CIDs allows articles to be fully citable as soon as they are published online, and connects the same identifier to all online, print, and electronic versions of the publication. SPIE uses a six-digit CID article numbering system in which:

- The first four digits correspond to the SPIE volume number.
- The last two digits indicate publication order within the volume using a Base 36 numbering system employing both numerals and letters. These two-number sets start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B ... 0Z, followed by 10-1Z, 20-2Z, etc.

The CID Number appears on each page of the manuscript. The complete citation is used on the first page, and an abbreviated version on subsequent pages. Numbers in the index correspond to the last two digits of the six-digit CID Number.

# Contents

- ix *Conference Committee*
- xiii *Invited Panel Discussion: Real-World Issues and Challenges in Big Data Processing with Applications to Information Fusion*  
*Participants: S. Kumar, Technology Consultant (United States); R. Mehra, Scientific Systems Co., Inc. (United States); K. Lossau, Sotera Defense Solutions (United States); P. Natarajan, Raytheon BBN Technologies Corp. (United States); S. Das, Machine Analytics (United States); E. Blasch, Air Force Research Lab. (United States); I. Kadar, Interlink Systems Sciences, Inc. (United States)*

---

## **SESSION 1 MULTISENSOR FUSION, MULTITARGET TRACKING, AND RESOURCE MANAGEMENT I**

---

- 8745 03 **Estimability of thrusting trajectories in 3D from a single passive sensor** [8745-1]  
T. Yuan, Y. Bar-Shalom, P. Willett, R. Ben-Dov, S. Pollak, Univ. of Connecticut (United States)
- 8745 04 **Advances in displaying uncertain estimates of multiple targets** [8745-3]  
D. F. Crouse, U.S. Naval Research Lab. (United States)
- 8745 06 **Overview of Dempster-Shafer and belief function tracking methods** [8745-5]  
E. Blasch, Air Force Research Lab. (United States); J. Dezert, B. Pannetier, The French Aerospace Lab. (France)

---

## **SESSION 2 MULTISENSOR FUSION, MULTITARGET TRACKING, AND RESOURCE MANAGEMENT II**

---

- 8745 07 **Decentralized closed-loop collaborative surveillance and tracking performance sensitivity to communications connectivity** [8745-6]  
J. T. DeSena, S. R. Martin, J. C. Clarke, D. A. Dutrow, B. C. Kohan, A. J. Newman, The Johns Hopkins Univ. Applied Physics Lab. (United States)
- 8745 08 **Stochastic context-free grammars for scale-dependent intent inference** [8745-7]  
B. Balaji, Defence Research and Development Canada, Ottawa (Canada); M. Fanaswala, V. Krishnamurthy, The Univ. of British Columbia (Canada)
- 8745 09 **Sensor selection for target localization in a network of proximity sensors and bearing sensors** [8745-8]  
Q. Le, Hampton Univ. (United States); L. M. Kaplan, U.S. Army Research Lab. (United States)
- 8745 0A **Evaluating detection and estimation capabilities of magnetometer-based vehicle sensors** [8745-9]  
D. M. Slater, G. M. Jacyna, The MITRE Corp. (United States)

---

**SESSION 3 MULTISENSOR FUSION METHODOLOGIES AND APPLICATIONS I**

---

- 8745 0B **Urban multitarget tracking via gas-kinetic dynamics models** [8745-11]  
R. Mahler, Lockheed Martin Corp. (United States)
- 8745 0C **Background agnostic CPHD tracking of dim targets in heavy clutter** [8745-12]  
A. I. El-Fallah, A. Zatezalo, Scientific Systems Co., Inc. (United States); R. P. S. Mahler, Lockheed Martin Corp. (United States); R. K. Mehra, Scientific Systems Co., Inc. (United States); W. E. Pereira, Air Force Research Lab. (United States)
- 8745 0D **Tracking, identification, and classification with random finite sets** [8745-13]  
B. T. Vo, B. N. Vo, Curtin Univ. (Australia)
- 8745 0E **PHD filtering with localised target number variance** [8745-15]  
E. Delande, J. Housseineau, D. Clark, Heriot-Watt Univ. (United Kingdom)
- 8745 0F **Divergence detectors for multitarget tracking algorithms** [8745-16]  
R. Mahler, Lockheed Martin Corp. (United States)
- 8745 0G **A Gaussian mixture ensemble transform filter for vector observations** [8745-17]  
S. Nanuru, M. Coates, McGill Univ. (Canada); A. Doucet, Univ. of Oxford (United Kingdom)

---

**SESSION 4 MULTISENSOR FUSION METHODOLOGIES AND APPLICATIONS II**

---

- 8745 0L **High Level Information Fusion (HLIF) with nested fusion loops** [8745-22]  
R. Woodley, M. Gosnell, A. Fischer, 21st Century Systems, Inc. (United States)

---

**SESSION 5 MULTISENSOR FUSION METHODOLOGIES AND APPLICATIONS III**

---

- 8745 0M **A robust technique for semantic annotation of group activities based on recognition of extracted features in video streams** [8745-23]  
V. Elangovan, A. Shirkhodaie, Tennessee State Univ. (United States)
- 8745 0N **Feynman path integral discretization and its applications to nonlinear filtering** [8745-24]  
B. Balaji, Defence Research and Development Canada, Ottawa (Canada)
- 8745 0O **Particle flow inspired by Knothe-Rosenblatt transport for nonlinear filters** [8745-25]  
F. Daum, J. Huang, Raytheon Co. (United States)
- 8745 0P **Particle flow with non-zero diffusion for nonlinear filters** [8745-26]  
F. Daum, J. Huang, Raytheon Co. (United States)
- 8745 0Q **Zero curvature particle flow for nonlinear filters** [8745-27]  
F. Daum, J. Huang, Raytheon Co. (United States)
- 8745 0R **Fourier transform particle flow for nonlinear filters** [8745-28]  
F. Daum, J. Huang, Raytheon Co. (United States)

---

**SESSION 6    MULTISENSOR FUSION METHODOLOGIES AND APPLICATIONS V**

---

- 8745 0S    **Sequential testing over multiple stages and performance analysis of data fusion** [8745-30]  
G. Thakur, The MITRE Corp. (United States)
- 8745 0T    **Multisource information fusion for enhanced simultaneous tracking and recognition**  
[8745-32]  
B. Kahler, SAIC (United States)
- 8745 0U    **Dempster-Shafer theory and connections to information theory (Invited Paper)** [8745-33]  
J. S. J. Peri, The Johns Hopkins Univ. Applied Physics Lab. (United States)

---

**SESSION 7    SIGNAL AND IMAGE PROCESSING, AND INFORMATION FUSION APPLICATIONS I**

---

- 8745 0V    **Object detection and classification using image moment functions in the applied to video and imagery analysis** [8745-34]  
O. Mise, S. Bento, GE Intelligent Platforms (United States)
- 8745 0W    **Multi-parametric data fusion for enhanced object identification and discrimination**  
[8745-35]  
S. Kupiec, V. Markov, Advanced Systems & Technologies, Inc. (United States); J. Chavez, Air Force Research Lab. (United States)
- 8745 0X    **A neuromorphic system for object detection and classification** [8745-36]  
D. Khosla, Y. Chen, K. Kim, S. Y. Cheng, A. L. Honda, L. Zhang, HRL Labs. LLC (United States)
- 8745 0Y    **Machine vision tracking of carrier-deck assets for improved launch safety** [8745-37]  
B. J. Davis, R. W. Kaszeta, R. D. Chambers, B. R. Pilvelait, P. J. Magari, Create Inc. (United States); M. Withers, D. Rossi, Naval Air Warfare Ctr. Aircraft Division (United States)
- 8745 0Z    **A comparison of sensor resolution assessment by human vision versus custom software for Landolt C and triangle resolution targets** [8745-38]  
A. R. Pinkus, D. W. Dommett, Air Force Research Lab. (United States); H. L. Task, Task Consulting (United States)
- 8745 10    **Development of a real-world sensor-aided target acquisition model based on human visual performance with a Landolt C** [8745-39]  
H. L. Task, Task Consulting (United States); A. R. Pinkus, E. Geiselman, Air Force Research Lab. (United States)

---

**SESSION 8    SIGNAL AND IMAGE PROCESSING, AND INFORMATION FUSION APPLICATIONS II**

---

- 8745 11    **Qualitative evaluations and comparisons of six night-vision colorization methods** [8745-40]  
Y. Zheng, Alcorn State Univ. (United States); K. Reese, Univ. of Louisville (United States); E. Blasch, Air Force Research Lab. (United States); P. McManamon, Exciting Technology LLC (United States)

- 8745 12 **Real-time classification of ground from lidar data for helicopter navigation** [8745-41]  
F. Eisenkeil, Univ. Konstanz (Germany); T. Schafhitzel, U. Kühne, Cassidian (Germany);  
O. Deussen, Univ. Konstanz (Germany)
- 8745 13 **High-resolution land cover classification using low resolution global data** [8745-42]  
M. J. Carlotto, General Dynamics Advanced Information Systems (United States)
- 8745 14 **Fusion of multispectral and stereo information for unsupervised target detection in VHR  
airborne data** [8745-43]  
D. C. Borghys, M. Idrissa, M. Shimoni, Royal Belgian Military Academy (Belgium); O. Friman,  
M. Axelsson, M. Lundberg, Swedish Defence Research Agency (Sweden); C. Perneel, Royal  
Belgian Military Academy (Belgium)
- 8745 15 **Combining structured light and ladar for pose tracking in THz sensor management**  
[8745-44]  
P. Engström, M. Axelsson, M. Karlsson, Swedish Defence Research Agency (Sweden)

---

**SESSION 9 SIGNAL AND IMAGE PROCESSING, AND INFORMATION FUSION APPLICATIONS III**

---

- 8745 17 **Human activity recognition based on human shape dynamics** [8745-46]  
Z. Cheng, S. Mosher, Infoscitex (United States); H. Cheng, T. Webb, Air Force Research Lab.  
(United States)
- 8745 18 **Seismic signature analysis for discrimination of people from animals** [8745-47]  
T. Damarla, U.S. Army Research Lab. (United States); A. Mehmood, Air Force Institute of  
Technology (United States); J. M. Sabatier, Univ. of Mississippi (United States)
- 8745 19 **Anomalous human behavior detection: an adaptive approach** [8745-48]  
C. van Leeuwen, TNO (Netherlands); A. Halma, Research Kitchen (Netherlands); K. Schutte,  
TNO (Netherlands)
- 8745 1A **Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations  
of different surveillance operators** [8745-49]  
H. Bouma, J. Vogels, O. Aarts, C. Kruszynski, R. Wijn, G. Burghouts, TNO (Netherlands)
- 8745 1B **Invariant unsupervised segmentation of dismounts in depth images** [8745-50]  
N. S. Butler, R. L. Tutwiler, The Pennsylvania State Univ. (United States)

---

**SESSION 10 SIGNAL AND IMAGE PROCESSING, AND INFORMATION FUSION APPLICATIONS IV**

---

- 8745 1C **Acoustic signature recognition technique for Human-Object Interactions (HOI) in persistent  
surveillance systems** [8745-51]  
A. Alkilani, A. Shirkhodaie, Tennessee State Univ. (United States)
- 8745 1D **Time series prediction of nonlinear and nonstationary process modeling for ATR** [8745-52]  
A. Sokolnikov, Visual Solutions and Applications (United States)
- 8745 1E **A multi-attribute based methodology for vehicle detection and identification** [8745-53]  
V. Elangovan, B. Alsaidi, A. Shirkhodaie, Tennessee State Univ. (United States)

- 8745 1G **A cross-spectral variation of the cross-ambiguity function** [8745-55]  
D. J. Nelson, U.S. Dept. of Defense (United States)
- 8745 1H **Analysis of angle of arrival estimation at HF using an ensemble of structurally integrated antennas** [8745-56]  
C. F. Corbin, G. A. Akers, Air Force Intitute of Technology (United States)

---

**SESSION 11 INVITED SPECIAL SESSION: SOCIAL/CULTURAL MODELING WITH APPLICATION TO INFORMATION FUSION PART 1: ACTIVITY-BASED MODELING**

---

- 8745 1I **Summary of human social, cultural, behavioral (HSCB) modeling for information fusion panel discussion (Invited Paper)** [8745-58]  
E. Blasch, J. Salerno, Air Force Research Lab. (United States); I. Kadar, Interlink Systems Sciences, Inc. (United States); S. J. Yang, Rochester Institute of Technology (United States); L. Fenstermacher, Air Force Research Lab. (United States); M. Endsley, SA Technologies (United States); L. Grewe, California State Univ., East Bay (United States)
- 8745 1K **Pattern of life from WAMI objects tracking based on visual context-aware tracking and infusion network models (Invited Paper)** [8745-60]  
J. Gao, Intelligent Fusion Technology, Inc. (United States); H. Ling, Temple Univ. (United States); E. Blasch, K. Pham, Air Force Research Lab. (United States); Z. Wang, G. Chen, Intelligent Fusion Technology, Inc. (United States)
- 8745 1L **Learning and detecting coordinated multi-entity activities from persistent surveillance (Invited Paper)** [8745-61]  
G. Levchuk, M. Jacobsen, C. Furjanic, A. Bobick, Aptima, Inc. (United States)

---

**PART 2: ACHIEVING HIGHER LEVELS OF FUSION, UNDERSTANDING THE HUMAN ELEMENT**

---

- 8745 1M **Consumer-oriented social data fusion: controlled learning in social environments, social advertising, and more (Invited Paper)** [8745-62]  
L. Grewe, California State Univ., East Bay (United States)
- 8745 1O **Influence versus intent for predictive analytics in situation awareness (Invited Paper)** [8745-64]  
B. Cui, S. J. Yang, Rochester Institute of Technology (United States); I. Kadar, Interlink Systems Sciences, Inc. (United States)

---

**POSTER SESSION**

---

- 8745 1Q **Infrared small target detection technology based on OpenCV** [8745-66]  
L. Liu, Z. Huang, Nanjing Univ. of Science and Technology (China)
- 8745 1R **Simultaneous optimization by simulation of iterative deconvolution and noise removal for non-negative data** [8745-67]  
A. M. Amini, Southern Univ. and A&M College (United States); G. E. Ioup, J. W. Ioup, Univ. of New Orleans (United States)

- 8745 1S **Simultaneous optimization by simulation of iterative deconvolution and noise removal to improve the resolution of impulsive inputs** [8745-68]  
A. M. Amini, Southern Univ. and A&M College (United States); G. E. Ioup, J. W. Ioup, Univ. of New Orleans (United States)
- 8745 1T **Self-adaptive characteristics segmentation optimized algorithm of weld defects based on flooding** [8745-69]  
C. Dang, J. Gao, Z. Wang, F. Chen, Xi'an Jiaotong Univ. (China)
- 8745 1U **Intrusion detection on oil pipeline right of way using monogenic signal representation** [8745-71]  
B. M. Nair, V. Santhaseelan, C. Cui, V. K. Asari, Univ. of Dayton (United States)
- 8745 1V **Optimising the use of hyperspectral and multispectral data for regional crop classification** [8745-72]  
L. Ni, Ctr. for Earth Observation and Digital Earth (China) and Univ. of Chinese Academy of Sciences (China); B. Zhang, L. Gao, S. Li, Y. Wu, Ctr. for Earth Observation and Digital Earth (China)
- 8745 1X **Breast tumor classification via single-frequency microwave imaging** [8745-74]  
C. M. Do, R. Bansal, Univ. of Connecticut (United States)
- 8745 1Z **Stabilizing bidirectional associative memory with Principles in Independent Component Analysis and Null Space (PICANS)** [8745-76]  
J. P. LaRue, JADCO Signals (United States); Y. Luzanov, Air Force Research Lab. (United States)
- 8745 20 **Option pricing formulas and nonlinear filtering: a Feynman path integral perspective** [8745-77]  
B. Balaji, Defence Research and Development Canada, Ottawa (Canada)

*Author Index*



# Conference Committee

## *Symposium Chair*

**Kenneth R. Israel**, Major General (USAF Retired) (United States)

## *Symposium Cochair*

**David A. Whelan**, Boeing Defense, Space, and Security (United States)

## *Conference Chair*

**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)

## *Conference Cochairs*

**Erik P. Blasch**, Air Force Research Laboratory (United States)

**Kenneth Hintz**, George Mason University (United States)

**Thia Kirubarajan**, McMaster University (Canada)

**Ronald P. S. Mahler**, Lockheed Martin Corporation (United States)

## *Conference Program Committee*

**Mark G. Alford**, Air Force Research Laboratory (United States)

**William D. Blair**, Georgia Tech Research Institute (United States)

**Mark J. Carlotto**, General Dynamics Advanced Information Systems (United States)

**Kuo-Chu Chang**, George Mason University (United States)

**Chee-Yee Chong**, BAE Systems (United States)

**Marvin N. Cohen**, Georgia Tech Research Institute (United States)

**Frederick E. Daum**, Raytheon Company (United States)

**Mohammad Farooq**, AA Scientific Consultants Inc. (Canada)

**Charles W. Glover**, Oak Ridge National Laboratory (United States)

**I. R. Goodman**, Consultant (United States)

**Lynne L. Grewe**, California State University, East Bay (United States)

**David L. Hall**, The Pennsylvania State University (United States)

**Michael L. Hinman**, Air Force Research Laboratory (United States)

**Jon S. Jones**, Air Force Research Laboratory (United States)

**Martin E. Liggins II**, The MITRE Corporation (United States)

**James Llinas**, University at Buffalo (United States)

**Raj P. Malhotra**, Air Force Research Laboratory (United States)

**Alastair D. McAulay**, Lehigh University (United States)

**Raman K. Mehra**, Scientific Systems Company, Inc. (United States)

**Harley R. Myler**, Lamar University (United States)

**David Nicholson**, BAE Systems (United Kingdom)  
**Les Novak**, Scientific Systems Company, Inc. (United States)  
**John J. Salerno Jr.**, Air Force Research Laboratory (United States)  
**Andrew G. Tescher**, AGT Associates (United States)  
**Stelios C. A. Thomopoulos**, National Center for Scientific Research  
Demokritos (Greece)  
**Wiley E. Thompson**, New Mexico State University (United States)  
**Pierre Valin**, Defence Research and Development Canada,  
Valcartier (Canada)

#### *Session Chairs*

- 1 Multisensor Fusion, Multitarget Tracking, and Resource Management I  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)  
**Thia Kirubarajan**, McMaster University (Canada)  
**Kenneth Hintz**, George Mason University (United States)
- 2 Multisensor Fusion, Multitarget Tracking, and Resource Management II  
**Kenneth Hintz**, George Mason University (United States)  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)  
**Thia Kirubarajan**, McMaster University (Canada)
- 3 Multisensor Fusion Methodologies and Applications I  
**Ronald P. S. Mahler**, Lockheed Martin Corporation (United States)
- 4 Multisensor Fusion Methodologies and Applications II  
**Kenneth Hintz**, George Mason University (United States)  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)
- 5 Multisensor Fusion Methodologies and Applications III  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)  
**Kenneth Hintz**, George Mason University (United States)
- 6 Multisensor Fusion Methodologies and Applications V  
**Kenneth Hintz**, George Mason University (United States)  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)
- 7 Signal and Image Processing, and Information Fusion Applications I  
**Lynne L. Grewe**, California State University, East Bay (United States)  
**Mark J. Carlotto**, General Dynamics Advanced Information Systems  
(United States)
- 8 Signal and Image Processing, and Information Fusion Applications II  
**Mark J. Carlotto**, General Dynamics Advanced Information Systems  
(United States)  
**Lynne L. Grewe**, California State University, East Bay (United States)

- 9 Signal and Image Processing, and Information Fusion Applications III  
**Lynne L. Grewe**, California State University, East Bay (United States)
- 10 Signal and Image Processing, and Information Fusion Applications IV  
**Lynne L. Grewe**, California State University, East Bay (United States)
- 11 Invited Special Session: Social/Cultural Modeling with Application to Information Fusion  
  
Part 1: Activity-Based Modeling  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)  
**Lynne L. Grewe**, California State University, East Bay (United States)  
  
Part 2: Achieving Higher Levels of Fusion, Understanding the Human Element  
**Lynne L. Grewe**, California State University, East Bay (United States)  
**Ivan Kadar**, Interlink Systems Sciences, Inc. (United States)



## **Invited Panel Discussion**

### **Real World Issues and Challenges in Big Data Processing with Applications to Information Fusion**

The panel addressed real-world issues and challenges highlighting the problem of handling/processing and using big data sources. A number of invited experts discussed current challenges of using big data sources in the fusion process and research to address these challenges. The proliferation of data sources has created an urgent need to manage, collect/retrieve and make sense of "big data". The big data problem is present in diverse areas such as: cybersecurity, financial and health analytics, social media networks, smart cities, digital video, text data, sensor networks, etc. Methods are needed to handle big data feeds from sensors, perform data and information fusion, and provide real-time and near real time information delivery. Processing challenges include machine learning techniques, robust predictive and explanatory analysis of high dimensional structured or unstructured data (e.g., data varying in format and dimensionality such as fusing text and video when video is not annotated and audio if needed), distributed and parallel processing paradigms, distributed fusion techniques to handle distributed data sources/bases, cloud computing, database query processing and model-based techniques. Additional necessary paradigms include structural learning, grouping data/clustering, dimensionality reduction, data mining, feature extraction and selection, statistical inference, regression, predictive modeling, signal processing, association and fusion. For example, in the unstructured data fusion problem mentioned above, if the video is not annotated a human interpreter would not know what parts are of the video and the text is related, thus requiring sophisticated visualization methods and cognitive modeling of user interface. The panelists illustrated parts of the above mentioned areas in many applications and addressed applications to all levels of information fusion. The objective of this panel was to bring to the attention of the fusion community the importance of dealing with, processing and using big data sources, the challenges and potential benefits. Conceptual and real-world related examples associated with the overall complex problem were used by the panel to highlight issues and challenges.

**Ivan Kadar**

**Invited Panel Discussion**  
**Real-World Issues and Challenges in**  
**Big Data Processing with Applications to**  
**Information Fusion**

**Organizers**

Ivan Kadar, Interlink Systems Sciences, Inc.

\*Chee-Yee Chong, BAE Systems Advanced Information  
Technologies

Srikanta Kumar, Technology Consultant

**Moderators**

Srikanta Kumar, Technology Consultant

Ivan Kadar, Interlink Systems Sciences, Inc.

April 29, 2013

SPIE Conference 8745

“Signal Processing, Sensor Fusion and Target Recognition XXII”

\* Note: Unable to attend     Baltimore, MD., 29 April – 2 May 2013

**Invited Panel Discussion**

***Panel Participants:***

\*Dr. Christopher White, DARPA, U.S.A

Dr. Srikanta Kumar, Technology Consultant, U.S.A.

Dr. Raman Mehra, Scientific Systems Co., U.S.A.

Dr. Kathleen Lossau, Sotera Defense Solutions, U.S.A.

Dr. Premkumar Natarajan, Raytheon BBN Technologies,  
Corp. U.S.A.

Dr. Subrata Das, Machine Analytics, U.S.A.

\*Dr. Erik Blasch, Air Force Research Lab., Rome Research  
Site, U.S.A.

Dr. Ivan Kadar, Interlink Systems Sciences, Inc., U.S.A.

\* Note: Unable to attend

**Invited Panel Discussion  
Topics**

**"DARPA XDATA Program"** Overview by  
Dr. Srikanta Kumar, Technology Consultant

**"Advanced Machine Learning & Statistical  
Inference Approaches for Big Data Analytics  
and Information Fusion"**

Dr. Raman K. Mehra , Scientific Systems Co.

**"Data in the Aggregate: Discovering Honest  
Signals and Predictable Patterns Within Ultra  
Large Data Sets"**

Dr. Kathleen Lossau, Sotera Defense Solutions

**"Big Data: A Multimodal Capability-centric  
Perspective"**

Dr. Premkumar Natarajan, Raytheon BBN  
Technologies, Corp.

**Invited Panel Discussion  
Topics**

**"Distributed and Cloud-Based Big Data Analytics  
and Fusion"**

Dr. Subrata Das, Machine Analytics, Inc.

**"Fusion Utility, Search, Index, Obtain, and  
Navigate (FUSION) over Enormous Data "**

Dr. Erik Blasch, AFRL/RIEA, Rome Research Site

**"Perceptual Reasoning Managed Big Data  
Analytics and Information Fusion"**

Dr. Ivan Kadar, Interlink Systems Sciences, Inc.

## **Advanced Machine Learning & Statistical Inference Approaches for Big Data Analytics and Information Fusion**

---

by

**Raman K. Mehra, Avinash Gandhe,  
Vikash Mansinghka (MIT),  
Patrick Shafto (U. of Louisville),  
Dan Lovell, Ssu-Hsin Yu**

**Presented at SPIE Defense and Sensing Conference,  
April 29 – May 3, 2013**

**Panel: Real World Issues and challenges in Big Data Processing with  
Applications to Information Fusion.**

**Acknowledgement:** This work is being supported by the DARPA XDATA Program under contract no. FA8750-12-C-0315



### **Outline**

---

- **Background & Historical Perspective**
- **Automated Bayesian Machine Learning using CrossCat**
- **Applications to XDATA Datasets**
- **Future Research Directions**



2



## What Has Created a Revolution in Big Data Predictive Analytics & Statistical Inference?

---

- Availability of Massive Datasets (Big Data) with 4 Vs (**Volume**, **Variety**, **Velocity**, and **Veracity**)
- Distributed Computing on Clusters
- Nonparametric Bayesian (NB) Inference
- Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo Methods for Probabilistic Inference

The combination of these 4 revolutionary developments has created a “**perfect storm**” for Big Data Analytics and its applications to a very large number of fields.

## Historical Perspective on Data Science, Probabilistic Reasoning and Bayesian Inference

---

- **Cardano (1501-1576)**: Formalized the notion of probability in gambling.
- **Bernoulli (1713) & Laplace (1774)**: Developed mathematical definitions of Probability.
- **Bayes (1763)**: Bayes' theorem (“An Essay toward solving a Problem in the Doctrine of Chance” suggested that probability judgements based on mere hunches should be combined with probabilities based on relative frequencies using Bayes' theorem.)
- **Gauss (1821)**: Principle of Least Squares.
- **Thiele (1880)**: First derivation of Recursive Least Squares or “**Kalman Filter**”.
- **Markov (1906, 1913)**: Markov Chain models.

## Historical Perspective on Data Science, Probabilistic Reasoning and Bayesian Inference (continued)

---

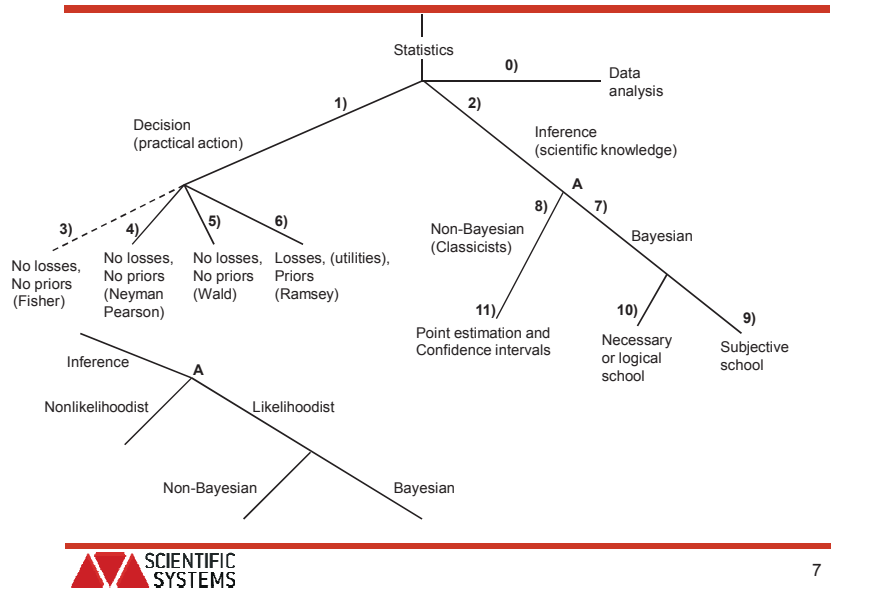
- [Fisher \(1925-1956\)](#): “Father of Statistical Science” (ANOVA, Fisher Information, Likelihood, Design of Experiments, Sufficient Statistics, ...).
- [Ramsey \(1926\), Jeffreys \(1939\), DeFinetti \(1937\), Savage \(1950\)](#): Foundations of Bayesian Decision Theory.
- [Kolmogorov \(1930-1940\)](#): Foundations of Probability; Measure Theory; Stochastic Processes.
- [Wiener \(1940-1950\)](#): Wiener Process; Wiener Filter.
- [Wald \(1945-1950\)](#): Sequential Decision Analysis.
- [Metropolis & Hastings \(1953, 1970\)](#): Monte Carlo Markov Chain Methods.

## Historical Perspective on Data Science, Probabilistic Reasoning and Bayesian Inference (continued)

---

- [Raiffa \(1961\)](#): Bayesian Decision Analysis, Conjugate Priors, Business Applications.
- [Parzen \(1960\)](#): Nonparametric Density Estimation.
- [Kalman \(1960\)](#): Filtering & Prediction for State Space Gauss-Markov Models; Stochastic Realization.
- [Box and Jenkins \(1970\)](#): Time Series Modeling and Forecasting.
- [Akaike \(1973\)](#): Gauss-Markov Model Identification (Canonical variate analysis which led to Subspace System Identification).
- [Ferguson \(1973\)](#): Nonparametric Bayesian Analysis.
- [Pearl \(1988\)](#): Probabilistic Reasoning in Intelligent Systems (AI applications).
- [Gordon, et al \(1993\)](#): Sequential Monte Carlo Methods & Particle Filtering for Nonlinear Estimation Filtering, Smoothing & Prediction.
- [The & Jordan \(2010\)](#): Hierarchical Bayesian Nonparametric Models.
- [Mansinghka, et al \(2013\)](#): Automated Bayesian Machine Learning using CrossCat, Probabilistic Programming.

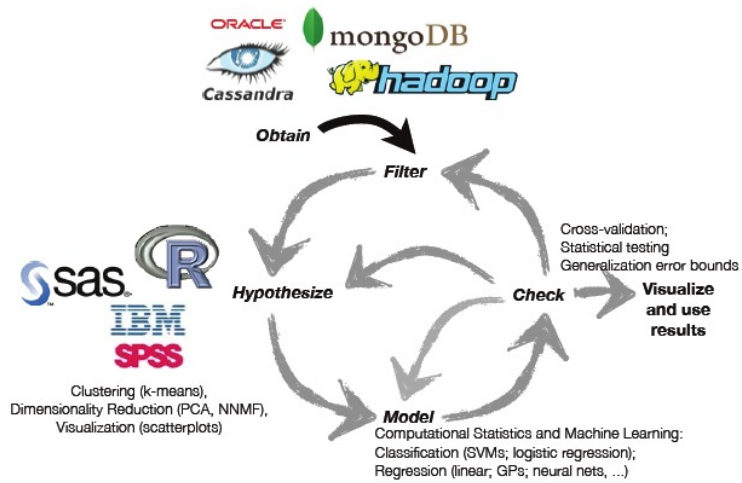
## An Overview of Statistics (Raiffa, 1970)



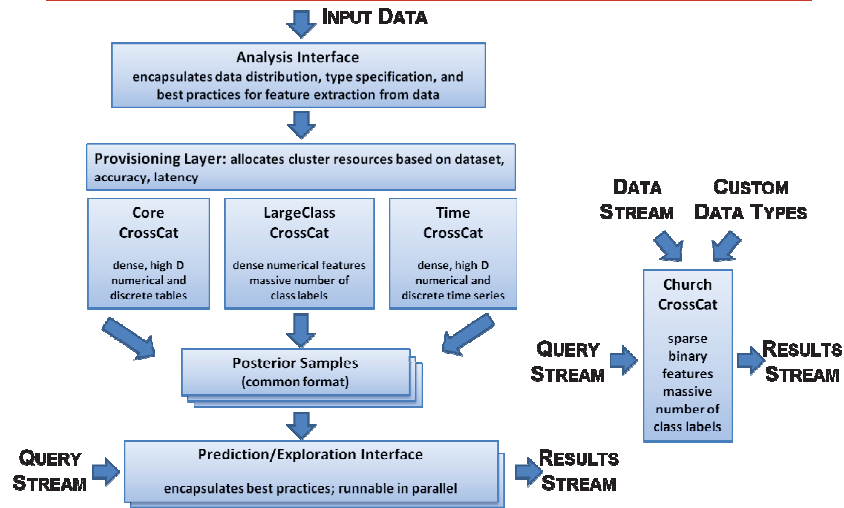
## Challenges for Big Data Predictive Analytics

- Reduce time & cost for analyzing heterogeneous high dimensional Big Data.
- Develop Machine Learning system that can be used by non-experts in Statistics & Data Science.
- Make it easy for users & decision makers to interact with data & provide inputs to guide the analysis process.
- Provide analysis results that are meaningful & can be used reliably by the decision maker. Quantification of uncertainty in Inference is crucial.
- Provide guidance for new data collection & design of experiments to collect relevant data while reducing clutter & noise.
- Develop methods for on-line/real-time analysis of streaming data.

## The "Big Data" Problem: *Productivity and Personnel*



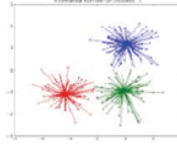
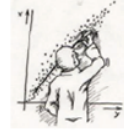
## Automated Bayesian CrossCat Software Family



## Under the Hood: CrossCat, An Estimator for the Full Joint Stochastic Process behind the Table

### Regression/classification:

$$\Pr[\text{target} \mid \text{predictor1}, \text{predictor2}, \dots]$$



### Clustering/Naive Bayes:

$$\Pr[X, Y] = \sum_c \Pr[c] \Pr[X|c] \Pr[Y|c]$$

### CrossCat: simulate and calculate any conditional distribution, and its independencies

$$\Pr[X=x, Y=y, Z=z, A=a, \dots]$$

$$\Pr[X=3|Y=4]$$

$$\Pr[Z=5|A=\text{male}, B=140]$$

$$(x,y) \sim \Pr[X,Y|A=\text{female}]$$



SCIENTIFIC  
SYSTEMS

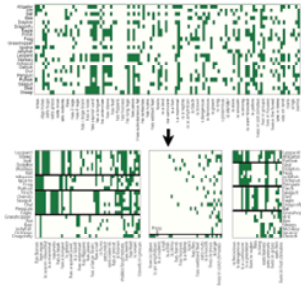
11

## Background: The CrossCat Method

### Learning

**Input:** Messy, high-dimensional heterogeneously typed data table

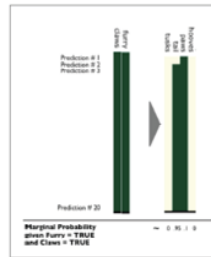
**Output:** Posterior samples, each representing an estimate of the full joint distribution



### Prediction

**Input:** Target variables, and given values

**Output:** Predictive samples, from  $P(\{\text{targets}\} \mid \{\text{givens}\})$



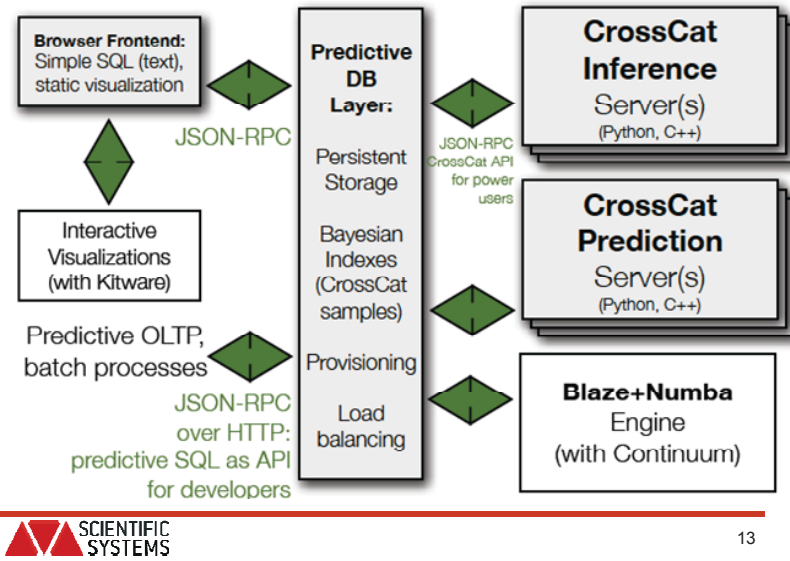
Mansinghka et al., accepted JMLR 2013



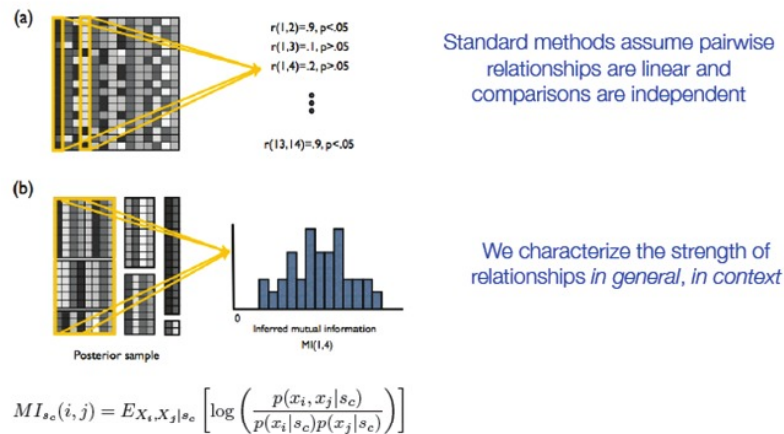
SCIENTIFIC  
SYSTEMS

12

## Predictive DB System Architecture



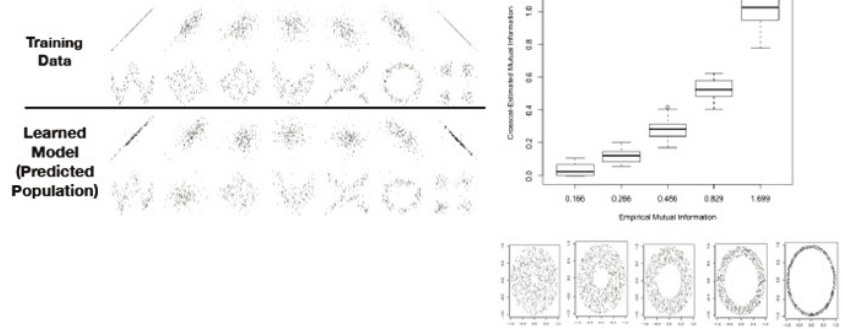
## Predictive Database Roadmap: Quantifying Bivariate Relationships



## Predictive Database Roadmap: Quantifying Bivariate Relationships (continued)

Finds non-linear stochastic structure that correlation cannot...

... and also estimates mutual information



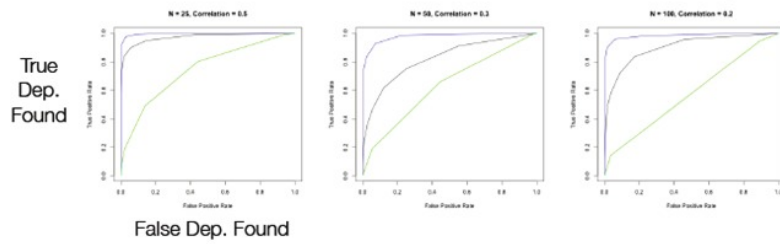
Krafft, Shafto, Mansinghka (in review)



## Predictive Database Roadmap: Multiplicity in High Dimensions

Improved statistical decisions by leveraging context

CrossCat in blue  
Mine in black  
Regression in green



25 datapoints  
0.5 correlation

50 datapoints  
0.3 correlation

100 datapoints  
0.2 correlation



---

## Validation of CrossCat on Simulated & XDATA Datasets:

- Airline & Weather Data
- Kiva Microloan Data

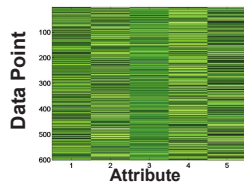
## Simulated Multi-dimensional Data

---

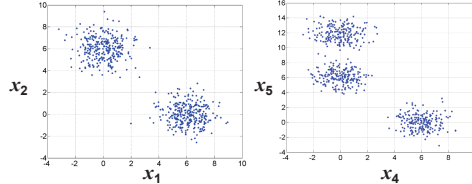
### DATA MODEL – 5 ATTRIBUTES, $x_1, x_2, x_3, x_4$ and $x_5$

$p(x_1, x_2, x_3, x_4, x_5) = p(x_1, x_2)p(x_3)p(x_4, x_5)$		$(x_1, x_2), x_3, (x_4, x_5)$ mutually ind.
$p(x_1, x_2) = \sum_{i=1}^2 \pi_i^{(1)} N(x_1, x_2; m_i^{(1)}, \Lambda_i^{(1)})$	$\pi_i^{(1)} = \frac{1}{2}$	Mixture of 2 Gaussians
$p(x_3) = N(x_3; m_i^{(2)}, \sigma_i^{(2)})$		Single Gaussian
$p(x_4, x_5) = \sum_{i=1}^3 \pi_i N(x_4, x_5; m_i^{(3)}, \Lambda_i^{(3)})$	$\pi_i^{(3)} = \frac{1}{3}$	Mixture of 3 Gaussians
$\Lambda_i^{(1)}, \Lambda_i^{(3)}$ : diagonal matrix		

600 INSTANCES FROM MODEL:



PROJECTIONS OF DATA POINTS:



Can CrossCat capture the independency among variables and the mixtures of Gaussians?



## CrossCat Model

- CrossCat works by stochastic shuffles that give an anytime sampler for the *posterior over joint distributions*

Joint Posterior Distribution

$$P(X, \{\bar{\theta}_c^d\}, \{\bar{y}^v, \alpha_v\}, \{\bar{\lambda}_d\}, \{\bar{z}, \alpha_D\})$$

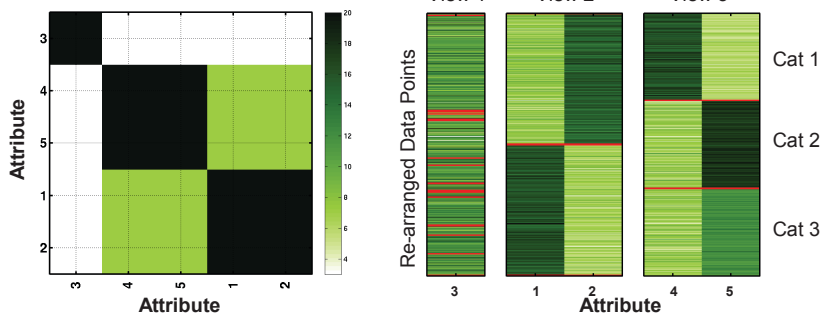
$$= V(\alpha_D) CRP(\bar{z}; \alpha_D) \prod_{v \in \bar{z}} V(\alpha_v) CRP(\bar{y}^v; \alpha_v) \prod_{d \in D} V(\bar{\lambda}_d) \prod_{c \in \bar{y}^v} \prod_{d \in D} \left( M(\bar{\theta}_c^d; \bar{\lambda}_d) \prod_{rec} L(x_{(r,d)}; \bar{\theta}_c^d) \right)$$

Prior prob. of column groupings (views)     
 Prior prob. of data partitions (categories) for each view     
 Hyper-prior for each dimension     
 Parameter Prior for each class, dimension     
 Likelihood for sampled data

## CrossCat Results on Simulated Data

CrossCat successfully found the independency among variables in views and mixtures of Gaussians in categories within views.

- Column Dependency Matrix – dependency of attributes (columns)
- An instance of views and categories within views



## Combined Flight and Weather Open Datasets

- Combining flight and weather data from 2 datasets
- The joined dataset is high-dimensional and heterogeneous
  - 11 categorical (from 2 to 20 cat.) and 23 ordinal attributes

Is there additional information to be found by sharing information among them? Can CrossCat find it?

- A typical data point:

Flight Data		Weather Data	
'DayOfWeek'	'Cancelled'	'Origin Mean Temp'	'Destination Mean Temp'
'CRSDepTime'	'CancellationCode'	'Origin Mean Dew Point'	'Destination Mean Dew Point'
'UniqueCarrier'	'Diverted'	'Origin Mean Visibility'	'Destination Mean Visibility'
'ActualElapsedTime'	'Delayed'	'Origin Mean Wind Speed'	'Destination Mean Wind Speed'
'CRSElapsedTime'	'CarrierDelay'	'Origin Precipitation'	'Destination Precipitation'
'AirTime'	'WeatherDelay'	'Origin Fog Flag'	'Destination Fog Flag'
'ArrDelay'	'NASDelay'	'Origin Rain Flag'	'Destination Rain Flag'
'DepDelay'	'SecurityDelay'	'Origin Snow Flag'	'Destination Snow Flag'
'Distance'	'LateAircraftDelay'		



## Experimental Conditions

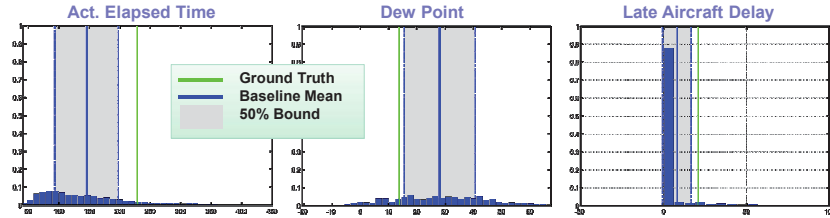
- Experiments use a table of 2000 instances and 34 attributes for each instance with a total of 68,000 cells
  - 20% of the table consist of cancelled flights, 30% delayed flights, and 50% neither cancelled nor delayed
- To conduct imputation experiments, we randomly withheld 1.5%, 5%, 10%, 25%, 50% and 75% of cells from the table.
  - Imputation results were compared to their true values



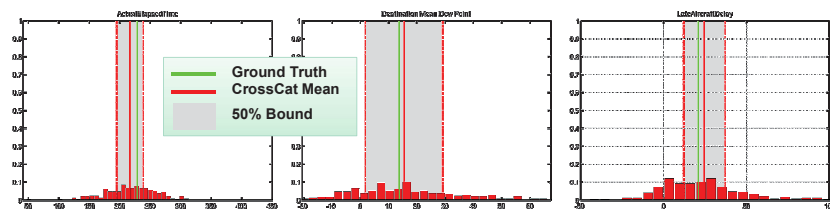


## Imputing Missing Values – Continuous Attributes

- **Baseline** – mean of marginal distribution (blue histogram) of each attribute.



- **CrossCat** – mean of posterior distribution (red hist.) conditional on known attributes.



CrossCat has smaller errors and the errors are within the predicted bounds

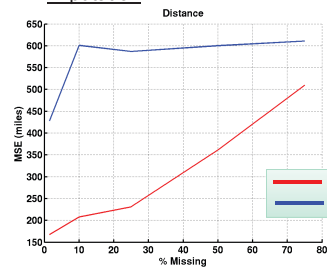


25

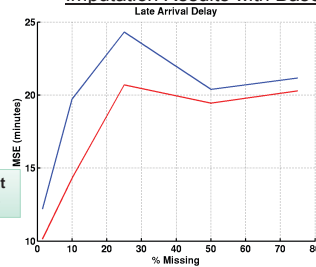
## Overall Imputation Performance Comparison

- For **continuous** attributes, comparing mean squared errors (MSE) of CrossCat and Baseline as a function of % missing cells:
  - CrossCat offers significant improvement over Baseline
  - As % missing increases, CrossCat converges gracefully to marginal distributions of attributes (Baseline)

Attribute Column with Good Imputation

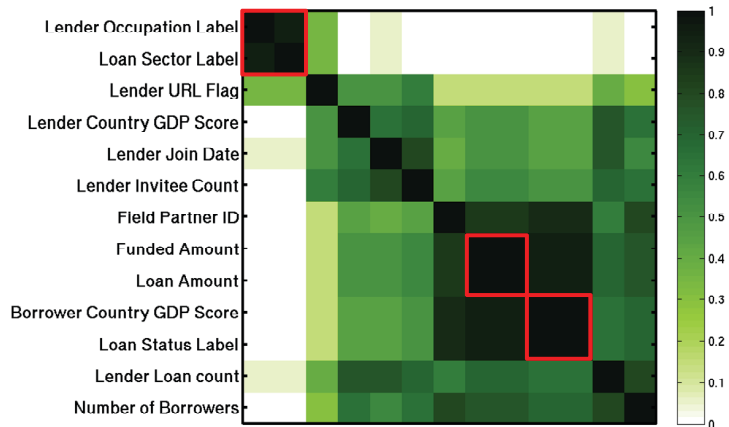


Attribute Column with Comparable Imputation Results with Baseline



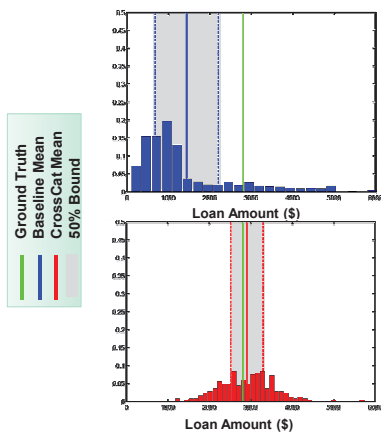
26

## Column Dependency – Kiva Microloan Dataset

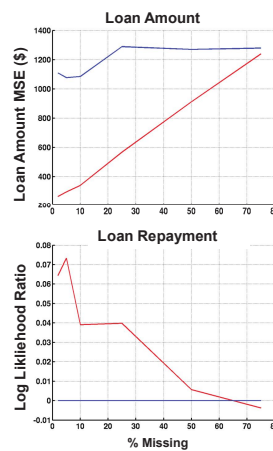


## Imputation Performance on Kiva Microloan Dataset

### Imputation of a single loan amount

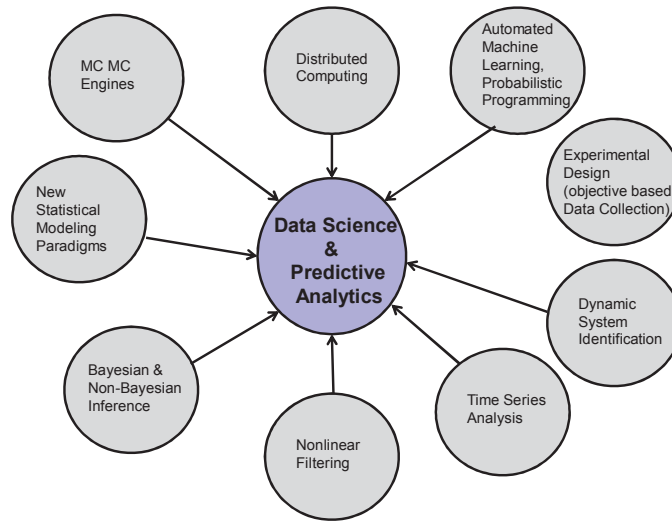


### Imputation Performance for Particular Columns



CrossCat reduces errors in imputation of many values resulting in overall performance improvement compared to baseline imputation

## Future Research Direction: Unification of Concepts and Methods developed in different fields



# Advanced Machine Learning and Statistical Inference Approaches for Big Data Analytics and Information Fusion

Raman K. Mehra<sup>a</sup>, Avinash Gandhe<sup>a</sup>, Vikash Mansinghka<sup>b</sup>, Patrick Shafto<sup>c</sup>, Dan Lovell<sup>a</sup>, Ssu-Hsin Yu<sup>a</sup>

<sup>a</sup>Scientific Systems Company, Inc., 500 West Cummings Park Suite 3000, Woburn MA 01801;

<sup>b</sup>Massachusetts Institute of Technology, Cambridge MA 02139;

<sup>c</sup>University of Louisville, Louisville KY 40202.

## ABSTRACT

A revolution in Big Data Predictive Analytics has been created by the confluence of four major revolutionary technologies, *viz.* (i) availability of massive datasets, (ii) distributed cluster computing (iii) advances in non-parametric Bayesian Inference and (iv) Markov Chain Monte Carlo (MCMC) methods for fast probability calculations and stochastic searches in high dimensions. The paper presents a historical perspective on the seminal breakthrough developments related to probabilistic reasoning and Bayesian Inference leading up to the current state-of-the-art in data science. This is followed by a discussion of challenges in Big Data Analytics and presentation of a method for Automated Bayesian Machine Learning using a recently developed approach called CrossCat. This approach is based on non-parametric Bayesian Inference and efficient use of MCMC numerical algorithms. Under the DARPA XDATA program, SSCI, MIT and the University of Louisville are developing a Predictive Database System which will have an SQL type front-end and CrossCat backend to facilitate the use of sophisticated machine learning methods by non-experts.

**Keywords:** Machine Learning, Big Data, CrossCat, XDATA, MCMC, Predictive Databases, Non-parametric Bayesian

## 1. BACKGROUND AND HISTORICAL PERSPECTIVE

A revolution in Big Data Predictive Analytics has been created by the confluence of four major revolutionary technologies, outlined in **Slide 3** of our accompanying presentation, that have created a perfect storm for Big Data Analytics. **Slides 4 through 6: Historical Perspective on Data Science ...** provide an overview of the major historical contributions starting with the work of Bernoulli, Laplace, Gauss and Bayes during the 18<sup>th</sup> and 19<sup>th</sup> centuries, followed by groundbreaking work in the 20<sup>th</sup> century by Markov, Fisher, Ramsey, Kolmogorov, Wiener and others. **Slide 7: An Overview of Statistics** presents an interesting overview of the field of statistics by Raiffa (1970) and the evolution of Bayesian methods. **Slides 8 and 9** present challenges for Big Data analytics.

## 2. AUTOMATED BAYESIAN MACHINE LEARNING USING CROSSCAT

**Slide 10: Automated Bayesian CrossCat Software Family** presents SSCI's research plan under the DARPA XDATA program for the development of a family of Bayesian machine learning software packages based on CrossCat. **Slides 11 and 12** present the basics of CrossCat, an estimator for the full joint stochastic process behind Big Data observations. **Slide 13: Predictive DB System Architecture** gives the system architecture for a Predictive Database system being developed by SSCI, MIT and Univ. of Louisville in collaboration with Kitware Inc. for visualization and Continuum Analytics for efficient parallel distributed implementation in order to scale CrossCat to high dimensional and large datasets. **Slides 14-16** discuss the use of CrossCat in identifying non-linear relationships in the data and present a comparison with linear correlation methods.

## 3. APPLICATIONS TO XDATA DATASETS

As part of our ongoing work in developing the CrossCat Machine Learning System, we have begun testing the performance of CrossCat on various datasets of interest in the XDATA DARPA program. In this section we examine the performance of CrossCat on two of the datasets but begin by using a synthetic example to introduce the CrossCat terminology and illustrate the type of information that can be obtained from its use.

### 3.1 Synthetic Data

CrossCat operates on data tables, with each row representing a data sample with multiple attributes, i.e., the columns of the table. In order to validate the performance of CrossCat we have designed a synthetic data table which contains simulated multi-dimensional data. The model for the simulated data is shown in **Slide 18: Simulated Multi-dimensional Data**. As shown in the slide, the data comprises 5 attributes of which attributes 1 and 2 are related, as are attributes 4 and 5. For each data sample in the table, Attributes 1 and 2 are drawn from a mixture of 2 Gaussians, Attribute 3 is drawn from a single Gaussian and Attributes 4 and 5 are drawn from a mixture of 3 Gaussians. Six hundred such data samples are drawn to construct the synthetic data table. **Slide 19: CrossCat Model** shows the generative model for CrossCat and the equation used to compute the posterior distribution. CRP stands for the Chinese Restaurant Process based on a Dirichlet Mixture Model. Computations of all distributions are done using an efficient Gibbs MCMC engine.

The results of applying CrossCat to the synthetic data is shown in **Slide 20: CrossCat Results on Simulated Data**. The figure on the right shows a *Column Dependency Matrix*, which captures the level of dependency between the different attributes in the data table. High dependency between attributes is shown by coloring the cell in the matrix with a darker shade and low dependency is indicated by a lighter shade (with black representing the highest dependency and white representing the least). The figure shows that Attributes 1 and 2 are dependent on each other, as are Attributes 4 and 5. Attribute 3 on the other hand is not dependent on any other attribute. These results are reflective of the model from which the data was sampled to begin with and therefore validate CrossCat's ability to find dependencies among attributes. The figure on the right hand side of the slide shows a single sample drawn from the Joint Posterior Distribution that CrossCat has estimated. This figure shows that CrossCat has partitioned the table correctly based on attributes (i.e., column clustering) and has created "Views" based on grouping dependent attributes together. CrossCat further partitions each View based on grouping similar rows together within each view. We refer to each of these groups within a View as a "Category". One can easily verify that the data in the View corresponding to Attributes 1 and 2 is divided into two categories based on the fact that the data samples are drawn from a mixture of two Gaussians. Similarly, the View corresponding to attributes 4 and 5 is partitioned into 3 categories based on the underlying mixture of three Gaussians. Based on these results it can be easily verified that CrossCat successfully found the dependency structure among variables in Views and mixtures of Gaussians in Categories within Views.

### 3.2 Flight and Weather Data

Based on a suggestion from Sotera, Inc. we studied the performance of CrossCat on a combined Flight and Weather database. Our goal is to determine if there is statistically significant information to be found in each of the datasets alone and if additional information be found by combining the two datasets. The flight data comprises flight information and on-time statistics for domestic flights between 1987 and 2008. We have downloaded a listing of flights from 2008, which has roughly 700,000 US domestic flight entries (see [http://www.transtats.bts.gov/Fields.asp?Table\\_ID=236](http://www.transtats.bts.gov/Fields.asp?Table_ID=236) for a description of the data). The weather dataset (from the National Climatic Data Center) contains data continuously collected at global weather stations including temperature, wind information, snow depths, etc. (<http://www.ncdc.noaa.gov/cgi-bin/res40.pl?page=gsod.html>). In order to merge the information in the two databases, we extracted, from the weather database, the weather conditions on the day of the each in the flight database at both the origin and destination airports. The resulting dataset is high-dimensional and heterogeneous with each data sample comprising 11 categorical and 23 ordinal attributes. An example of a typical entry in the data table is shown in **Slide 21: Combined Flight and Weather Open Datasets**. While a list of attributes for each entry is shown in the slide, the attributes can be broadly thought of as belonging to one of three classes: (i) Scheduled Flight Information (Route and Times), (ii) Actual Flight Information and (iii) Weather Information.

As in the previous section on synthetic data, we evaluated CrossCat's ability to describe relationships between attributes and data samples. In order to do so, we ran CrossCat on 2000 data points from flights in January 2008. To ensure diversity in the dataset, approximately 20% of the flights selected were cancelled, 30% of the flights were delayed and the remaining were on-time. **Slide 23: Column Dependency Matrix – Flight + Weather Data** shows the dependencies between various attributes in the dataset. Attributes are not individually identified on the slide due to space constraints but one can see that scheduled flight and airline information attributes, flight delay attributes and weather attributes each cluster together in the Column Dependency Matrix. However, the weather delay attribute is clustered with the weather information, which is intuitive. Another interesting relationship, although not explicitly labeled on the slide, is between Late Aircraft Delay and Scheduled Flight Departure Time. On further investigation, we confirmed that Late Aircraft Delays do aggregate through the day with highest delays in this category occurring during the mid-afternoon hours.



In addition to studying the relationships between attributes, we also studied CrossCat's imputation accuracy, i.e., its accuracy in predicting missing data in the database. To conduct imputation experiments, we randomly withheld 1.5%, 5%, 10%, 25%, 50% and 75% of cells from the table and estimated the data model in each case using CrossCat; imputation results were then compared to their true values to quantify the effectiveness of CrossCat. **Slide 25: Imputing Missing Values – Continuous Attributes** shows the imputation results for several missing continuous attribute cells in the table based on a Joint Posterior Distribution computed with 5% of the cells in the table missing. In each case, CrossCat is compared with a baseline imputation value, i.e., the mean of the marginal distribution of each attribute. The top row on the slide shows the marginal distribution of each attribute with the baseline imputation value (the blue vertical line), along with the true value (green vertical line) of the missing cell. Also shown using the dashed blue lines are the 50% error bounds, i.e., the bounds within which 50% of the distributions samples lie. Below each figure in the top row is shown the corresponding imputation result using CrossCat. Again, the red lines show the predicted value, along with error bounds, and the green line is true value. In each case shown, CrossCat's prediction of the missing value is significantly closer to the true value than the baseline imputation. **Slide 26: Overall Imputation Performance Comparison** aggregates the imputation performance over all the missing cells for a particular attribute and compares the performance of CrossCat to the baseline by comparing their Mean Squared Errors (MSE). One can see that CrossCat offers significant improvement over the baseline for each of the attributes shown (Distance and Late Arrival Delay). Also, as the percentage of missing cells in the table increases, CrossCat converges gracefully to marginal distributions of attributes (Baseline).

### 3.3 Kiva Micro-Loan Dataset

Kiva is a micro-lending organization that allows users to lend money to borrowers through its partnering micro-financing institutions. Kiva provides a publicly available database of its micro-loan activities. We have used the Kiva API to obtain a set of loan *transactions* that contain fields related to Lender information (Lender Id, Country, Occupation, etc.), Borrower information (Name, Country, etc.) and Loan information (Amount, Loan Sector, Loan Status, etc.). We sampled the roughly 150,000 downloaded entries to obtain 2000 data samples on which to perform analysis. Location information is converted from country labels to country GDPs in order to reduce the number of categories with a large number of labels relative to the size of the data table.

**Slide 27: Column Dependency – Kiva Microloan Dataset** shows the relationships between various attributes in the Kiva table. Many of the relationships are intuitive - for example, the funded loan amount is highly related to the requested loan amount. However, also of interest is the fact that the Loan Status (whether a loan has been repaid or defaulted) is related to the Borrowers country. In investigating the database, we have seen that several countries do indeed have higher rates of default than others.

In **Slide 28: Imputation Performance on Kiva Microloan Dataset** we show, in a manner similar to the analysis on the flight and weather dataset, the results of imputation using CrossCat. We have shown the improvement over baseline for a single missing "Loan Amount" cell on the left hand side of the slide. On the right side of the slide, we have shown the overall imputation performance for specific attributes, aggregated over all missing cells in the table from that attribute. Again, for the continuous valued attribute (Loan Amount), the Mean Squared Error in imputation using CrossCat is significantly lower than the MSE when using the baseline technique. For categorical attributes, we use a likelihood ratio (LLR) based metric to compare CrossCat to the baseline; values larger than 0 indicate that CrossCat is providing higher confidence in the correct category than Baseline. Based on this metric, CrossCat shows improvement in the Loan Status (i.e. repaid or defaulted) category when compared to the baseline. As expected, the performance of CrossCat approaches the baseline as the percentage of missing cells in the table increases.

## 4. FUTURE RESEARCH DIRECTIONS

**Slide 29: Future Research Direction** indicates fruitful directions for future research. The main thrust is unification of concepts and methods from the field of Statistics, Machine Learning, Computer Science, Control and Estimation theory, Time Series Analysis, Design of Experiments and MCMC stochastic search engines to greatly advance the state-of-the-art in Big Data Science and Predictive Analytics.

### ACKNOWLEDGMENT

This work is being supported by the DARPA XDATA program under contract no. FA8750-12-C-0315.

**SOTERA**  
DEFENSE SOLUTIONS



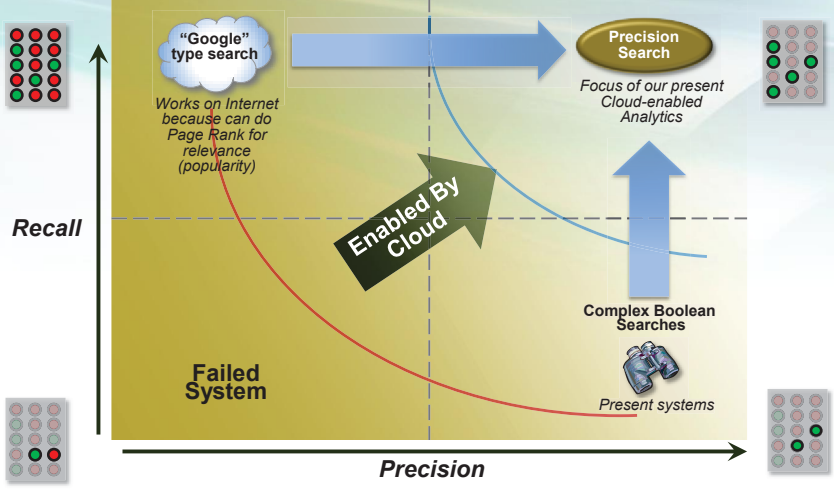
## Data in the Aggregate: Discovering Honest Signals and Predictable Patterns Within Ultra Large Data Sets

Kathleen Lossau and Jonathan Larson

April 2013

**SOTERA**  
DEFENSE SOLUTIONS

## Getting Precision Search from Big Data



Source: Dr. Russell Richardson – Chief Architect & Senior Science Advisor U.S. Army INSCOM

2  
Sotera Defense Solutions -  
Kathleen.Lossau@soteradefense.com

# Why Data Science

- Forbes - How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
- NYT - What Does Your Credit-Card Company Know About You?
  - Canadian Tire store (which sells electronics, sporting goods, kitchen supplies, automotive goods)
  - Credit ratings from purchasing activities
  - Chrome-skull hitch devices
  - Premium birdseed and “snow roof rakes” which helps protect pedestrian that walk by

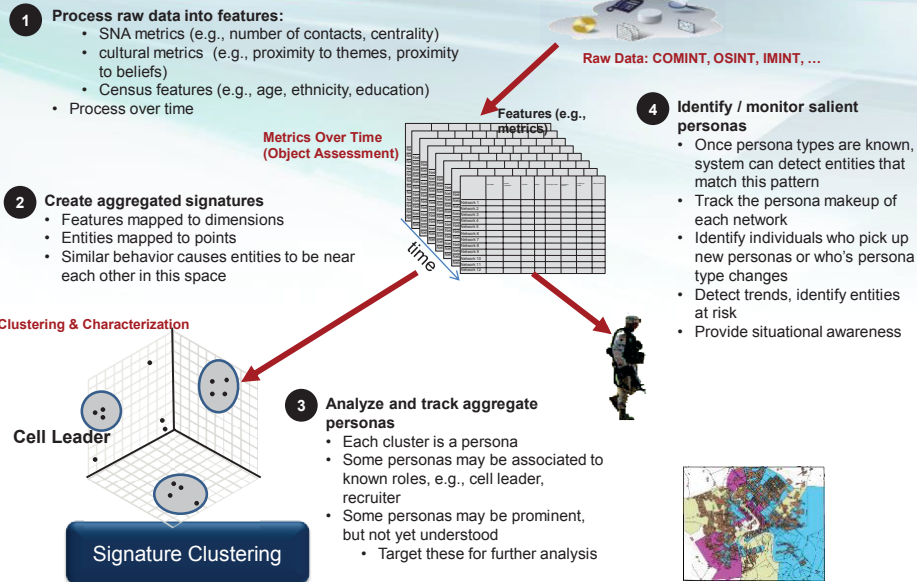
*His data indicated, for instance, that people who bought cheap, generic automotive oil were much more likely to miss a credit-card payment than someone who got the expensive, name-brand stuff. People who bought carbon-monoxide monitors for their homes or those little felt pads that stop chair legs from scratching the floor almost never missed payments. Anyone who purchased a chrome-skull car accessory or a “Mega Thruster Exhaust System” was pretty likely to miss paying his bill eventually.*

<http://www.nytimes.com/2009/05/17/magazine/17credit-t.html?pagewanted=all>

*As Pole’s computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.*

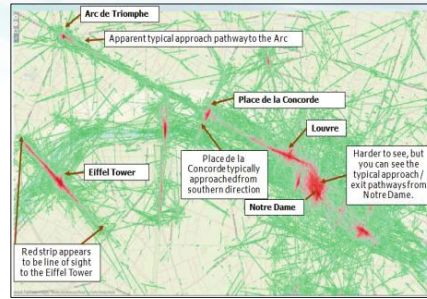
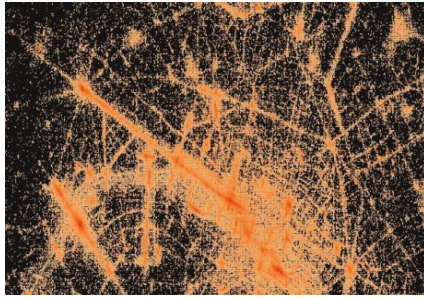
<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

# Analysis Process: How We Do It



## Inferring Movement From Points

- How can we infer movement patterns from vast amounts of what appears to be just point data collected over time and associated with a distinct identifier (e.g., a user ID, bank account number)?
- *Aggregate Micro-paths* - Technique is applicable to Twitter, FourSquare and MANY other sources.



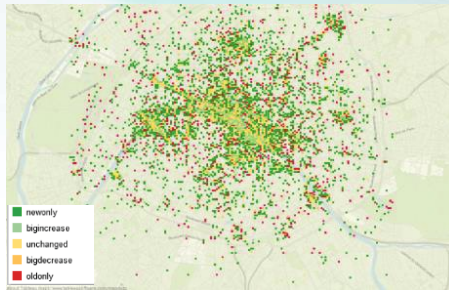
Volume plot of photos binned by area on log scale  
— Paris, France as seen from Flickr over all time.

5

## Tool/Technique:

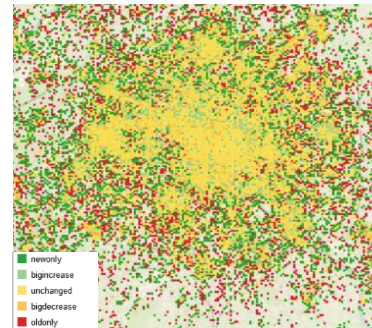
## Temporal Difference Analysis

### Flickr Paris 2004 changes vs 2005



Measuring the relative change of photographic activity year over year. Significant changes between years denoted by shades of green and red.

### Flickr Paris 2011 changes vs 2010



6

## Tool/Technique: 3D Hourly Tripline Blankets



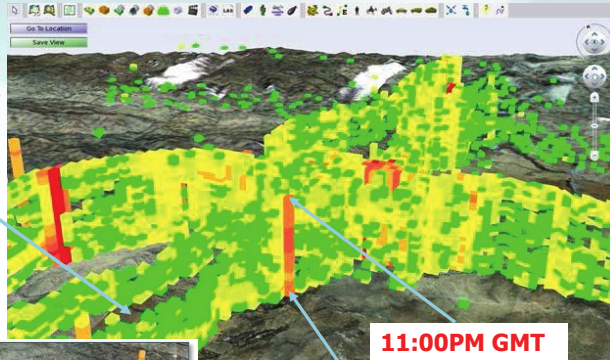
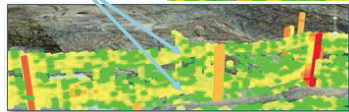
- All pings are assigned to a latitude / longitude cell and aggregated
- Additionally, all pings are binned to an hour of day on the Z-Axis **across all days**

Legend (log scale):



No pings on this road during the middle of the night (shown as absence of reading)

Notice the rush hour



11:00PM GMT

12:00AM GMT

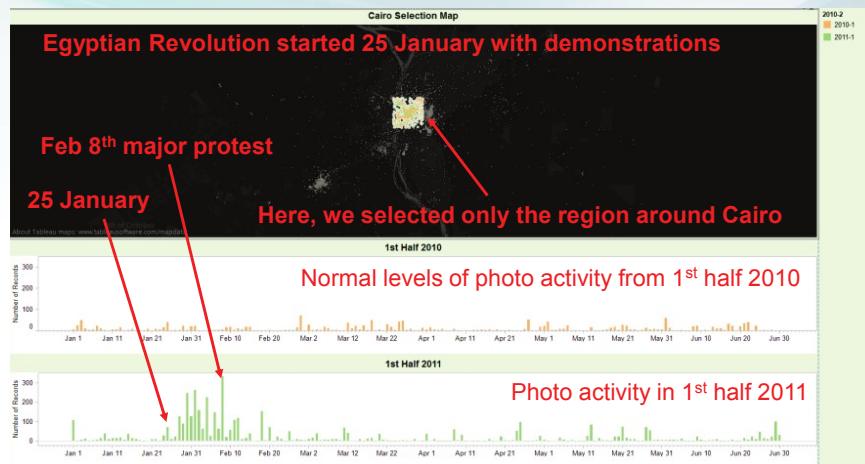
Unclassified

7

## Tool/Technique:



## Time - Event Characterization

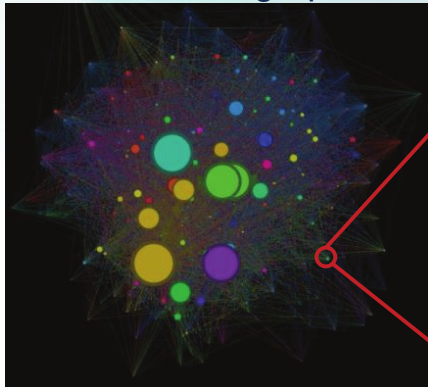


8

Sotera Defense Solutions -  
Kathleen.Lossau@soteradefense.com

## Zooming into a community

- This functionality allows actual browsing of enormous graphs



Zoomed snapshot of sscott5@enron.com's community

9

## Conclusion

- Previously the job was to filter through large sources of data to find specific pieces of information that fused together tell a picture - Now it is the large data itself that is the product
  - Information fused on different levels reveals patterns and trends within a given slice of the data.
  - The challenge is in finding the right people to excavate the relevant dimensions within the data to create meaningful and relevant aggregated data products.
- Analysts will increasingly be looking at aggregated data products consisting of multiple sources of data fused together to provide an understanding of normal patterns of behavior.
  - Look at trends - products can be created on behaviors and other dimensions of communities, regions, large corporations, and ethnic, religious, cultural organizations.
  - Compare incoming (streaming?) data against known patterns and trends to quickly find anomalies

10

# Data in the Aggregate: Discovering Honest Signals and Predictable Patterns within Ultra Large Data Sets

K. Lossau, J. Larson

Sotera Defense Solutions, 1515 S. Capital of Texas Hwy; Austin, TX 78746

## ABSTRACT

Traditionally information fusion has focused on the tactical value of finding and tracing a single needle in a haystack. While this approach provides value, it focuses only on a single person instead of identifying the entire culture, community, and scope of a target organization. Data analysis in the aggregate can provide immense strategic value, especially in identifying honest signals<sup>1</sup> and habits (often unintentional). Aggregation of data through data warehousing has been used on large data sets to enhance query response times by summarizing or partially summarizing the data over various dimensions (e.g. pivot tables) or grouping data based on relationships (e.g. clustering). We continue to explore how to use data aggregations as additional data elements to be further processed, analyzed, and queried. We will discuss several mechanisms for analyzing different types of large data sets including dimensional databases and graph data through the application of cluster computing (both in memory and file based representations). This strategy will employ several information fusion techniques that operate on these aggregations to detect anomalies, discover correlations and present historical patterns within the datasets. Approximation techniques, which can be used to reduce the computational order of complexity, are also discussed.

## 1. ISSUES ANALYSTS FACE TODAY

Data continues to be generated and digitally archived at increasing rates, resulting in vast databases available for search and analysis. Access to these databases has generated new insights through data-driven methods in the commerce, science, and computing sectors. Senior DoD leaders have said the Defense sector is “swimming in sensors and drowning in data.”<sup>2</sup> The so-called “big data” problem has now become a challenge for military operations, both at strategic and tactical levels. The data being brought to bear on operations are growing rapidly in volume and complexity, and are most often imperfect, incomplete, heterogeneous, and consumed by diverse end-users from analysts to field soldiers. Defense applications now have environments where data can sometimes be seen only once, for milliseconds, or can only be stored for a short time before being deleted. The trends are accelerated by the proliferation of various digital devices and the Internet, which are being used by adversaries in all stages of threat production, from planning to logistics to resource movement to operations. Therefore, it is critical to develop fast, scalable, and efficient methods for processing and visualizing data that not only support ingestion and transformation but also enable fast search and analysis. Aggregations and data projections are important not only for analytic execution, but also for visualization and statistical presentation of patterns within the data.

---

<sup>1</sup> “Honest Signals - How They Shape Our World” Alex Pentland, MIT Press, ISBN: 978-0-262-16256-2, 2008

<sup>2</sup> We’re going to find ourselves in the not too distant future swimming in sensors and drowning in data,” said Lt. Gen. David A. Deptula, Air Force deputy chief of staff for intelligence, surveillance and reconnaissance. going to find ourselves in the not too distant future swimming in sensors and drowning in data,”: LTG. David A. Deptula, Air Force deputy chief of staff for intelligence, surveillance and reconnaissance.

Using these base technologies alongside existing feature extraction / enrichment analytics can allow for a flexible platform for the development and implementation of new capabilities that work at scale. In this manner, the raw data is used to build analytic projections recursively, such that each new projection is built off of an existing set of projections. Effectively, this creates a serendipitous ad-hoc development environment, in which many analytics can be tested forensically across historical data, while also enabling a pipeline for pre-defined analytics that are optimized for real-time operations and feedback.

The benefit of these techniques is to provide a rich basis for analysis of the large-scale multiINT environment. The use of a large data framework will have significant advantages:

- Scalability – ability to work at scale in a cloud-type environment by use of map/reduce, bulk synchronous parallel, aggregator trees, and other techniques
- Reliability – ability to partition the data provide quality of service through distributed hardware fault tolerance
- Adaptability – ability to provide new and data aggregated data products
- Modularity – ability to bring new analytic results across a non-static dataspace, and the ability to persist those findings back into an architecture for provisioning and dissemination (and follow-on analysis)

These techniques provide the platform on which analytics can be built to “see what might otherwise not be seen”. The ability to pivot and project data from this stance of deep multi-dimensionality can allow discovery of trends and traits, including those the opponent may not even realize they have.

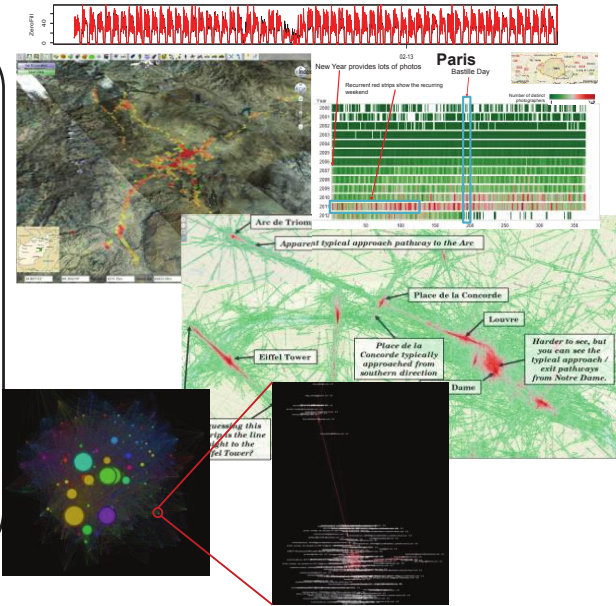
## 2. EMERGING DATA PRODUCTS

Large scale data pose significant challenges in locality, transmission, and the partitioning of the data. Sotera and other organizations have increasing more experience in processing these large data streams by using open source cloud technologies, such as HDFS, MapReduce, Hive, Impala, Shark, Giraph, Storm, and Spark which distribute the computational and storage load into a parallelized and scalable environment. For real time or iterative analytics, usage of in-memory distributed systems such as Spark / Shark are used. For pure real time analytics, Storm and Spark Streaming provide excellent capability. For graph processing, Giraph, Bagel, and Hama each provide for a distributed Bulk Synchronous Parallel implementation. . The result of using open source and Government developed software these systems provide users with sub second response to queries to all of the data. Using the projections of the data and the aggregated data products, the analyst can see all of the information and ad-hoc drill down into the relevant information quickly. In the past, analysts were restricted in their query options and often forced to choose selective criteria. This selective criteria biased the returned dataset as it tossed out information that could potentially have value that the analyst had not considered.

The discovery of honest signals in the aggregated data can be discovered by statistically looking at these patterns in the data. Normalizing the data over a common set of features (e.g. time and location being obvious ones) we can uncover patterns with single or multiple data sources. The figures below demonstrate some of these examples. A time series analysis of flight data, traffic patterns in a region at a specific time, social media displayed on a map, and large graphs aggregated through hierarchical clustering with a drill down capability.



- Large Multi-Int analysis
- Data mining, discovery, advanced search
- Discovery, trend analysis and correlation
- Modeling, entity management and Pattern of Life
- Multi-INT correlation across disparate sources



Drawing on experience and techniques that have been leveraged and refined in the multiple DoD programs, analytics can be built for combinatorial calculations en masse by pair-wise combinations across many different data sources and dimensions at scale. Further, approximations techniques can aid in discovering highly correlated items and greatly expedite algorithms that would otherwise have computational complexity that is intractable. This capabilities together formulate an automated correlation capability that can uncover non-obvious relationships between datasets that would otherwise be left undiscovered. The initial results of these analytics, which range into the trillions of comparisons, have provided groundbreaking results that prove the worthy of applying massive computational power to discover otherwise hidden patterns. Sotera has also developed several classes of statistical anomaly detection algorithms, aggregate movement characterization analysis, and rasterization techniques that are optimized to operate at ultra-large scale and can be applied to new problem sets. These analytics are given a final polish through visualization using a set of multi-dimensional and graph tools tailored to analyst’s needs.

### 3. CONCLUSION

Analysts will increasingly be looking at aggregated data products consisting of multiple sources of data fused together to provide an understanding of normal patterns of behavior. In addition to looking at patterns-of-life over individual people, specific entities, events or locations, patterns-of-life products can be created on behaviors and other dimensions of communities, regions, large corporations, and ethnic, religious, cultural organizations. New information can be compared to known trends to discover anomalies and changing patterns of behavior. Where previously the job was to filter through large sources of data to find specific pieces of information that fused together tell a picture, now it is the large data itself that is the product and when fused on different levels reveals patterns and trends within a given slice of the data. The challenge is in finding the right people to excavate the relevant dimensions within the data to create meaningful and relevant aggregated data products.

The slide features a background with abstract, colorful, swirling patterns in shades of purple, blue, and orange. The text is overlaid on this background. The title 'Raytheon BBN Technologies' is in a large, bold, black font. Below it, the subtitle 'Big Data: A Multimodal Perspective' is in a smaller, black font. The date '29 April 2013' is positioned to the right of the subtitle. The speaker's name and title, 'Dr. Prem Natarajan, Executive Vice President and Principal Scientist', are listed below the date, along with his email address 'pnataraj@bbn.com'. The Raytheon BBN Technologies logo is located in the bottom right corner of the slide.

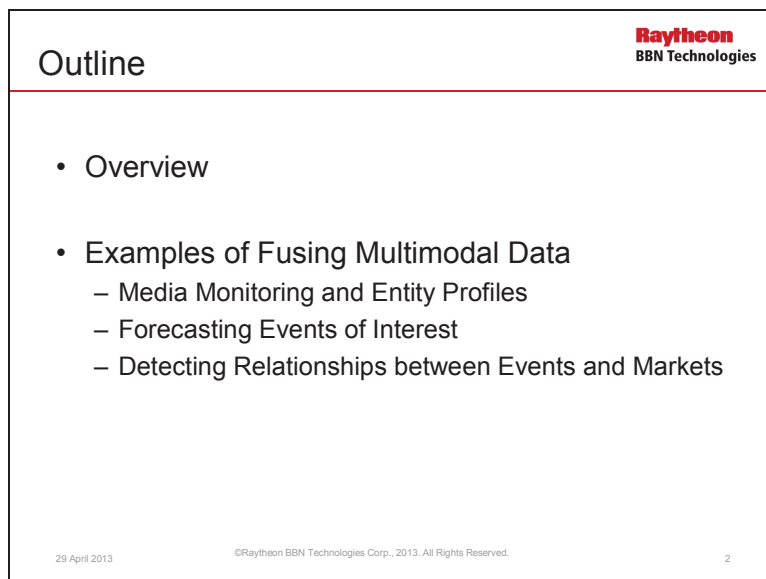
**Raytheon BBN Technologies**

Big Data: A Multimodal Perspective

29 April 2013

Dr. Prem Natarajan  
Executive Vice President  
and Principal Scientist  
[pnataraj@bbn.com](mailto:pnataraj@bbn.com)

**Raytheon**  
BBN Technologies

The slide has a white background with a red horizontal line near the top. The word 'Outline' is written in a large, black font on the left side. The Raytheon BBN Technologies logo is in the top right corner. A bulleted list is centered on the slide, containing three main items: 'Overview', 'Examples of Fusing Multimodal Data', and 'Detecting Relationships between Events and Markets'. The second item has three sub-items: 'Media Monitoring and Entity Profiles', 'Forecasting Events of Interest', and 'Detecting Relationships between Events and Markets'. At the bottom of the slide, there is a small copyright notice and the page number '2'.

**Raytheon**  
BBN Technologies

## Outline

- Overview
- Examples of Fusing Multimodal Data
  - Media Monitoring and Entity Profiles
  - Forecasting Events of Interest
  - Detecting Relationships between Events and Markets

29 April 2013 ©Raytheon BBN Technologies Corp., 2013. All Rights Reserved. 2

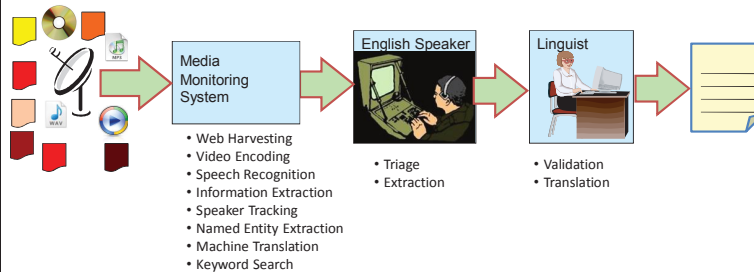
- Big data: Quantity, but also
  - Modality
  - Richness
  - Diversity
- Our focus: Analytics that capitalize on multiple modalities to
  - Fuse information to enable effective summarization
  - Produce robust inferences despite weak signals

## **MEDIA MONITORING & ENTITY PROFILES**

## Integrated Media Monitoring System

**Raytheon**  
BBN Technologies

- 24-7 Collection of broadcast and web media



- Provides tools that break the language barrier
  - English speakers locate and select foreign language content
  - Linguists focus on translation and cultural nuance
- Automatically filters, sorts, monitors, and prioritizes

29 April 2013

©Raytheon BBN Technologies Corp., 2013. All Rights Reserved.

5

## From Search to Summary

**Raytheon**  
BBN Technologies

- Media monitoring system: English search capabilities for multimedia or non-English sources
- As volume grows, simple keyword search becomes a barrier
  - Too many hits are returned for analyst to read
  - Keyword search only identifies the information the analyst knows she/he needs to know
- Automatic summarization of key facts is necessary to reduce information overload and direct analyst to relevant information

29 April 2013

©Raytheon BBN Technologies Corp., 2013. All Rights Reserved.

6

## Big Data and Summarization

- Information Extraction (IE) automatically identifies key facts in text, but with errors
- For some entities, information may:
  - Be present in one modality, but not another
  - Conflict across data streams, and require validation
- Enriching analysis by leveraging multiple data streams will:
  - Improve precision, aggregating all available sources
  - Find information that would otherwise be missed

## FORECASTING EVENTS OF INTEREST

## Forecasting

- Media monitoring system monitors news as it happens
- Even more powerful: predict events before they occur, e.g.: riots, strikes, protests, elections
- Many potential sources, each with opportunities and limitations
  - Twitter and other social media sources
  - Traditional news and online news sources
  - YouTube, Instagram, Flickr

## Example Signals for Forecasting

- Twitter
  - Opportunity: Near real-time, person-centric
  - Limitation: Biased sample (only a subset of the population is involved enough to tweet)
  - Limitation: Language is extremely challenging
- News Media
  - Opportunity: Existing technology for text processing
  - Limitation: Reports events after they happen
    - Need to extract precursors (e.g. negotiations, poll results) rather than forecasted events (e.g. strikes, election results)
  - Limitation: Without aggregators, harvesting is limited
    - National elections, but not city-council elections

## Forecasting: Approach

- Enrich the data stream where possible
  - Extract poll results from news
  - Transcribe audio to identify videos about a topic
  
- Employ machine learning to fuse weak signals into a prediction, for example forecasting election winner from
  - Frequency of
    - Key terms (e.g. candidate names) in Twitter stream
    - YouTube views of candidate-sponsored video
  - Poll results

## DETECTING RELATIONSHIPS BETWEEN EVENTS AND MARKETS

**Raytheon**  
BBN Technologies

## Goal

- Discover novel theories of influence between financial markets and extrinsic world events
- Examples:
  - Regulatory entity restricts an industrial activity
  - Ruling on a merger/acquisition
- Aim for novel, latent, links of influence
  - Manipulation
  - Unintended consequences

29 April 2013 ©Raytheon BBN Technologies Corp., 2013. All Rights Reserved. 13

**Raytheon**  
BBN Technologies

## Scale, Richness and Diversity

- Large scale
  - Petabytes of financial data
  - Terabytes of noisy unstructured text data
- Extrinsic events reflected in textual sources
  - News, blogs, tweets,....
  - Unstructured data
- Financial market indicators
  - Equities, fixed income, commodities,...
  - Structured data
- Multiple data streams; extended time spans
- Requires joint analysis
  - Discover causal and correlation relationships

29 April 2013 ©Raytheon BBN Technologies Corp., 2013. All Rights Reserved. 14



# Big Data: A Multimodal Capability-centric Perspective

Premkumar Natarajan

Raytheon BBN Technologies Corp., 50 Moulton Street, Cambridge, MA, USA 02138

## 1. INTRODUCTION

A dramatic increase in the number and variety of sensors over the past decade has resulted in a proliferation of data sources and in the availability of massive amounts of data in different modalities. The sheer volume and diversity of the available data has inspired the coining of a new term – “Big Data.” Big Data can refer to the scale, richness, and diversity of this data. In recent years, a significant amount of research and development interest has focused on the exciting fusion, inference, and visualization possibilities presented by Big Data.

In this brief paper, we describe some recent efforts in building multimedia analytics for big data by applying state-of-the-art technology to enrich the data stream, for example transcribing speech into text or identifying key events mentioned in an article. Fusing different data sources, especially those that represent different modalities, can enable powerful new ways for humans to understand and respond to key information as it emerges. For example fusing the Twitter stream with the news stream can provide insights into both authoritative and popular sentiment about an issue. Similarly, fusion across modalities affords the opportunity to improve accuracy and identify information that would otherwise be missed. At a more fundamental level, by operating over big, multimodal data-streams, we can able to develop inference mechanisms that are robust and accurate even in the face of weak and noisy signals. In the following sections we provide some examples that illustrate how different data sources have been combined to uncover and organize new information.



Figure 1. Fused and translated French and Arabic search results for an English query.

## 2. CASE STUDY: PROFILES FOR ENTITIES

BBN's Multi Media Monitoring System (M3S) ingests, transcribes, and translates broadcast news, web-news, and blogs. M3S performs robust, 24-7 collection and thus results in a rapidly growing text corpus. In initial versions, M3S was configured to allow English search over audio and non-English text, thus allowing an English speaking analyst query capabilities over broadcast news and non-English sources.

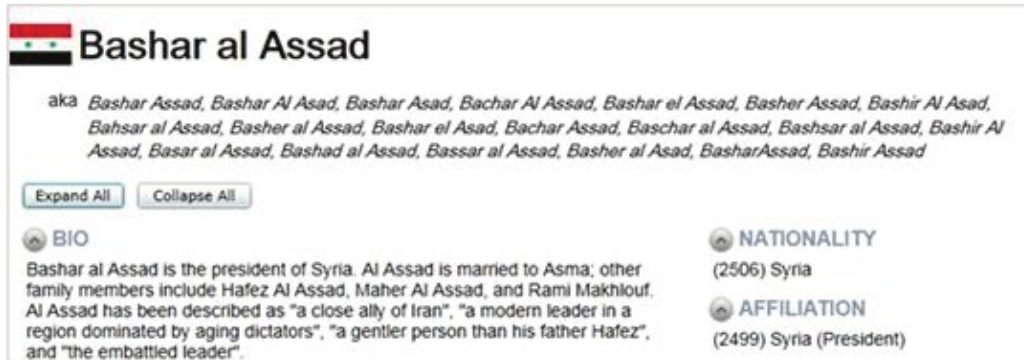


Figure 2. Automatically generated profile of Bashar Al Assad.

As the data stream has grown, more sophisticated analytics have become desirable-- exhaustively reviewing search results becomes infeasible and even thinking of the correct queries may be difficult. BBN's IDX information extraction (IE) system extracts key pieces of information from a document (e.g. *EmployedBy(ABC Petrochemical Corp, John Doe)*) from the text "*The board appointed John Doe to the newly created position of chairman at ABC Petrochemical Corp.*" However, as demonstrated by system-performance in NIST's ACE and TAC-KBP evaluations, even state-of-the-art IE is far from perfect—the top performing system in NIST's 2011 Slot Filling evaluation received an F-score of ~0.3.

Recently BBN has developed and deployed an Entity Profile (EP) capability as a part of M3S. The EP system capitalizes on the large data volume available through M3S to generate accurate, up-to-date, useful profiles for people and organizations of interest despite noisy IE technology. Ongoing and future work in this area will address techniques for fusing information from different modalities – for example, finding photographs and video clips that contain the entity of interest, or attaching audio clips of a person entity talking. The Entity Profiles system updates the profiles as new information comes in, e.g., identifying changes in employment of a person. Future versions will extend the capability to produce Event Profiles in addition to Entity Profiles, with profiles being updated as events unfold.

## 3. CASE STUDY: FORECASTING OF CIVIL UNREST AND ELECTION RESULTS

Traditionally, text processing has focused on finding events as reported in the news. More recently, several groups have been working on methods for fusing large amounts of data from social media and news media sources to predict the occurrence of events of interest. For example, on IARPA's OSI project, we archive of 10% of the tweets in originating from South and Central America to support our research. Our archive currently contains ~5 billion tweets and is constantly updated. Separately we collect news information using M3S augmented with a light weight, rapidly-deployable collection mechanism. Currently we harvest over 100 websites with M3S and an additional 600 using light weight scraping. We also use the YouTube and Google Trends APIs to collect information about web searches, video availability, and video viewing.

These diverse sources provide complementary indicators about the world. Traditional news media provides authoritative, trusted reporting about events that have happened. However such sources tend to be reactive rather than predictive and it is infeasible to collect new sources that cover all areas of a region-- coverage will be rich for national elections, but sparse for elections at the level of a city or town. Twitter and posted videos provide standardized access to the general population, but they will present a biased view—those with the strongest opinions will speak the most; interested parties can manipulate the stream. Search results from Google/YouTube provide are more limited in that they speak to what people are searching for and not what they are saying, but may be more reflective of the general population-- an individual can show interest by viewing a video even if they are not interested enough to tweet about a topic.

Examples of ongoing research tasks include:

- Predicting election results using both the frequency with which videos that mention a candidate's name are viewed on YouTube with polling results as extracted by an Information Extraction System from news sources.
- Predicting the occurrence of a labor strike at a given time in a given place by examining the words appearing in the Twitter stream as well as reports of precursor-events (e.g. labor negotiations) extracted from news sources.

#### **4. CASE STUDY: DETECTING EVENTS IN VIDEO**

BBN is developing approaches that integrate low-level and semantic features in combination with powerful machine learning techniques to detect events of interest in large volumes of consumer-domain videos and to provide a human understandable recounting of the salient attributes/evidences that resulted in the detection of an event. The available event-related information is often noisy – the current state-of-the-art in visual concept detection is far from reliable and the audio stream in web videos is extremely noisy and often contains unrelated/overlaid content.

We address the challenge posed by weak/noisy signals by processing information from multiple modalities including the visual, audio, speech, and videotext streams, and by extracting and combining a large variety of complementary as well as redundant low-level audio and visual features. We also combine these with semantic information from visual object and concept detectors as well as multimodal information from audio event detection, automatic speech recognition and video text recognition. To efficiently integrate all available information, we have developed a multi-stage fusion strategy – first, we develop multiple sub systems using feature-level fusion, and then combine the outputs of these subsystems using score fusion. This strategy has produced consistent performance gains and enabled the BBN VISER team to achieve excellent performance in the TRECVID MED evaluations conducted by NIST.

Looking forward, we are pursuing multiple ideas that attempt to build on our initial success. First, we can rapidly filter irrelevant data by using a cascade strategy where we make a first pass with fast features and then progressively apply more expensive features on smaller video sets. A second idea is to exploit cross-modal context, where high confidence detections in one modality can be used to improve performance in another modality. A third idea is the propagation of tags between similar videos that would minimize the need for human annotation and provide rich textual summaries.

#### **5. CONCLUSIONS**

This position paper provides motivating examples that highlight the need for further research into effective and efficient fusion of multi-source, multimodal data with the goal of performing robust, accurate inference in the face of noisy and weak signals. While there are numerous interesting technical challenges in the area of Big Data processing, the tradecraft associated with effective exploitation of Big Data is still in its infancy. The emerging importance of Big Data provides a unique opportunity for the development of framework that natively combines two key elements – technology and tradecraft – of effective information exploitation. Successful development of such a framework could have a transformative impact on organizational efficiency and mission success.

# Distributed and Cloud-based Big Data Analytics and Fusion

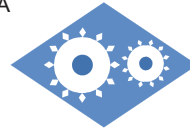
*Real World Issues and Challenges in Big Data Processing with Applications to Information Fusion*

Invited Panel Discussion at SPIE, Baltimore, MD

**Subrata Das**

Machine Analytics, Inc., Belmont, MA  
www.machineanalytics.com

April 29, 2013



## Analytics and Data Fusion

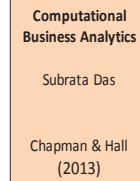


"... study of relationships among objects and events of interest within a dynamic environment." – Das

"... leverage data in a particular functional process (or application) to enable context-specific insight that is actionable. – Gartner"



(CRC Press, 2008)




(forthcoming; not actual cover)

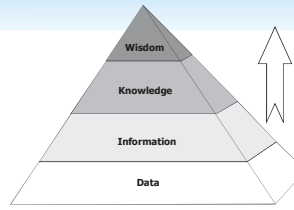
**Two sides of the same coin!**

Machine Analytics, Inc.

2

# Analytics Categorized

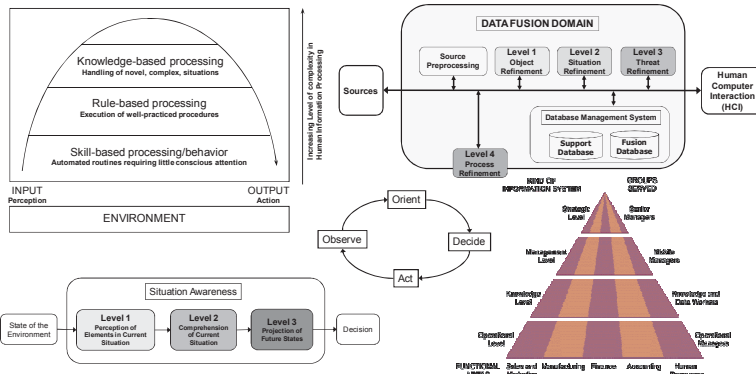
- Descriptive analytics
  - Current and historical look at organizational performance.
- Predictive analytics
  - Predicts future trends, behaviour and events for decision support.
- Prescriptive analytics (a.k.a. decision support)
  - Determines alternative courses of actions or decisions given the historical, current and projected situations, and a set of objectives, requirements, and constraints.
  - Check out the LinkedIn group Prescriptive Analytics 



**Descriptive Analytics:** How have been the monthly sales for the past twelve months? Who are the most valuable customers?  
**Predictive Analytics:** What are the projected sales for the next six months? Who are the customers likely to leave?  
**Prescriptive Analytics:** What actions could be taken to increase the sales? what incentives can be offered to encourage customers to stay/prevent from leaving?



# Many similar architectures!



## Big Data Scenarios



- Inherently distributed and maintained autonomously
- Stored in a centralized data warehouse
- Reside in the cloud

How do we query and perform analytics?

## MapReduce vs Parallel Databases



- Stonebraker et al. (2010) argues that MapReduce (Dean and Ghemawat, 2008) complements parallel DBMSs since databases are not designed for extract-transform-load tasks, which is a MapReduce specialty.
- Techniques for MapReduce query evaluation over large datasets, especially various types of join queries, have been presented in (Blanas et al., 2010; Chandar, 2010).

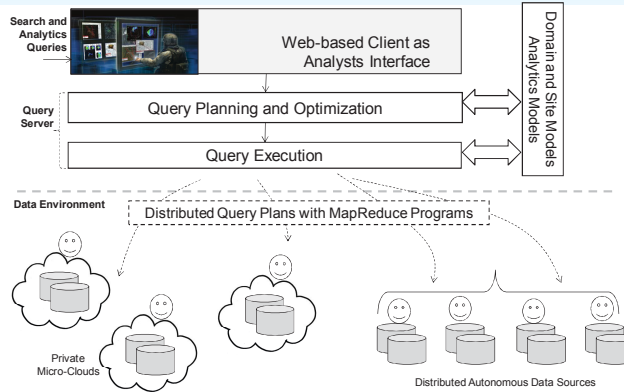
# Hadoop/MapReduce Analytics



- Broadly applicable MapReduce parallel programming paradigm (summation form) applied to many different learning algorithms (Ng et al., 2006).
- Bayesian network parameter learning from incomplete data (Basak et al., 2012).
- RDF closure of a graph (Urbani et al., 2009; Mutharaju, 2010).
- Tree learning (Wu et al., 2009; Panda et al., 2009) and random forest learning (Basilico et al., 2011).
- Search for optimization via simulated annealing (Radenski, 2012).
- Association rule learning (Hammoud, 2011; Woo and Xu, 2011).
- Text processing via graph algorithm (Lin and Dyer, 2010).
- Neural network (NN) training (Liu et al., 2010).

Make use of clustered and sorted (key, value) pairs mapped between multiple cores

# Architecture



## Query and Analytics

- Query execution
  - Join query (broadcast, natural and parallel) is one of the most common operations.
- Analytics for big data on the cloud
  - Distributed search for evidence and centralized evaluation of models.
- Distributed belief propagation
  - Inherently distributed data, models, expertise.

## Broadcast Join

**QUERY:** Select all sensing platforms from which reports have been generated for named areas of interest with no go mobility.

NAI	Mobility
23	No Go
40	No Go
55	No Go
61	No Go
...	...

NAI-Mobility, Mobility = 'No Go'



Memory-based hash table

Map phase splits data by (key, value)

NAI	FROM	ACTIVITY	EQUIPMENT	TIME	SIZE	NAI	FROM	ACTIVITY	EQUIPMENT	TIME	SIZE	NAI	FROM	ACTIVITY	EQUIPMENT	TIME	SIZE
23	JSTARS	Sensing	UAV	18:12	4500	45	JSTARS	Sensing	UAV	18:30	4500	61	JSTARS	Sensing	UAV	18:30	4500
2	UAV	Employed	BMP	18:12	7	65	UAV	Employed	BMP	18:12	7	95	Lt. Infantry	Employed	BMP	18:12	7
4	LRF	Meeting	AK-47	19:30	100-200	81	LRF	Meeting	AK-47	19:30	100-200	98	LRF	Meeting	AK-47	19:30	100-200
23	JSTARS	Sensing	Truck	05:10	1	29	ROKIT	Digging	Truck	05:10	1	47	JSTARS	Digging	Truck	05:10	1

Horizontal Split



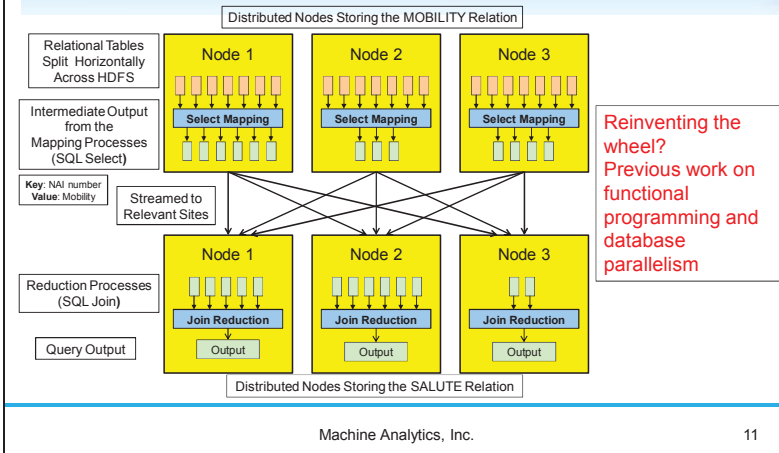
Output

NAI	Mobility
23	JSTARS
40	UAV
55	Lt. Infantry
61	HUMINT
...	...

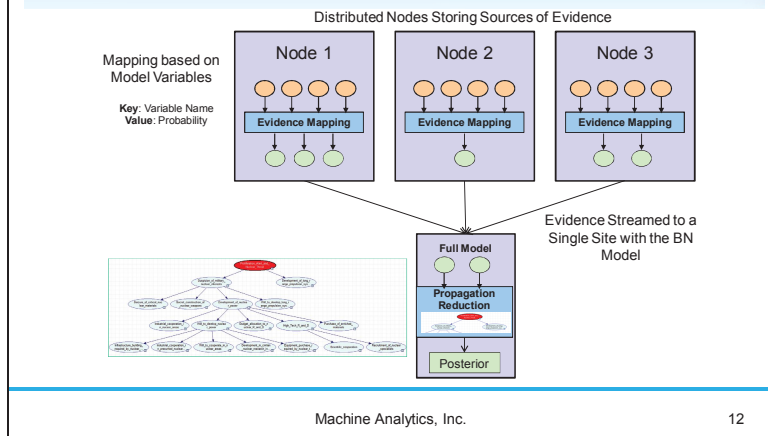
Data for HDFS are in flat files



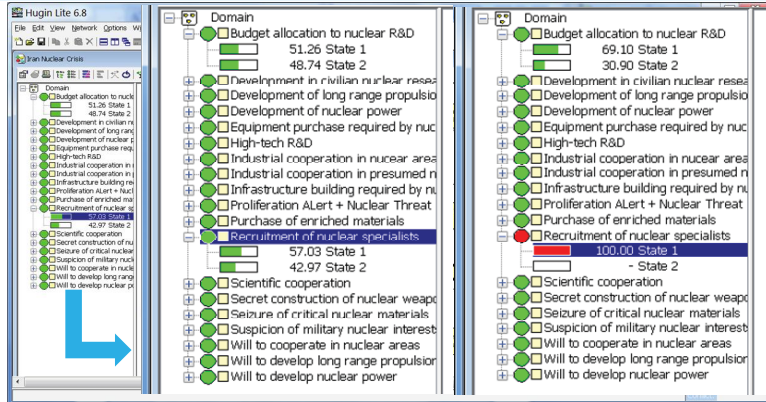
# Parallel Join



# Centralized Evidence Propagation



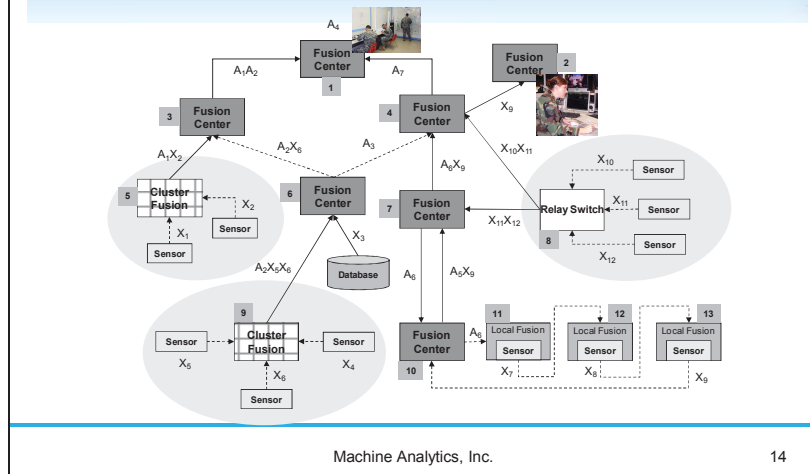
# Centralized Evidence Propagation



Machine Analytics, Inc.

13

# Typical Distributed Fusion Environment



Machine Analytics, Inc.

14

# Distributed Agent-based Search and Distributed Belief Propagation



Das, 2013

Fragment	Parent	Probabilities
Fragment 0	Fragment 0 Parent: Proliferation Alert and Nuclear Threat	(0.769,0.231)
Fragment 1	Fragment 1 Parent: Suspicion of military nuclear interests	(0.68,0.32)
Fragment 2	Fragment 2 Parent: Development of long range propulsion system	(0.537,0.463)
Fragment 3	Fragment 3 Parent: Development of nuclear power	(0.52,0.475)
Fragment 4	Fragment 4 Parent: Will to develop nuclear power	(0.691,0.319)
Fragment 5	Fragment 5 Parent: High Tech R and D	(0.52,0.48)

Posterior probability of proliferation threat after observing evidence on civilian nuclear cooperation.

## Conclusions

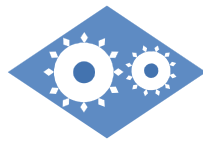


- Cloud-based big data analytics is at an early stage.
- Agent-based approaches are promising for distributed analytics.

## Contact



[sdas@machineanalytics.com](mailto:sdas@machineanalytics.com)



Thanks!

# Distributed and Cloud-based Big Data Analytics and Fusion

*Subrata Das*

Machine Analytics, Belmont, MA

## 1. Abstract

Distributed and parallel processing paradigms provide efficient ways to implement big data analytics and fusion. In this position paper, I will present such approaches to analytics and fusion to process big data that are inherently distributed or residing on the cloud. The underlying foundation consists of traditional database query processing and model-based techniques for high-level fusion and descriptive and predictive analytics.

## 2. Overall Technical Approach

Various defense agencies routinely collect and store large volumes of data on a continuous basis from a variety of disparate and heterogeneous sources. The magnitude and multi-proprietary nature of available data often requires data to be stored in multiple “inherently” independent repositories distributed over a network. Though such distribution is coherent with recent thrusts towards net-centric warfare (DoD, 2001), analysts often face a daunting task when searching for specific data or for series of correlated data residing in distributed sources. One solution is to build a large centralized data storage area in advance, such as a cloud. However, the proprietary nature of some of the sources requires that they operate autonomously. Thus any approach to search and analytics in such a data intensive scenario should be able to seamlessly deal with data both inside and outside the cloud.

The approach we advocate automatically translates a high-level user search or analytics query into a plan of sub-queries to be executed in parallel at both cloud and distributed data sources (Stonebraket et al., 2010). For centralized data stored, our approach fully exploits the in-built distributed file storage and parallel execution paradigm of the cloud to process queries via the map-reduce paradigm (Dean and Ghemawat, 2008). Our approach incorporates distributed micro-clouds as data sources, along with other non-cloud-based distributed data sources. The query planning and optimization component generates optimized programs to be carried as wrappers by agents to micro-clouds. Hence, our framework is essentially an integration of cloud- and agent-based distributed search and analytics tools for the cloud and other external distributed data sources.

The proposed approach supports two fundamental operations, namely search query and complex model-based analytics, on both centralized cloud and distributed data sources. With an assumption that select-join-projection queries to databases provide some form of completeness towards the need for information requirements, and owing to the similarity between the map-reduce paradigm and the select-projection operation, the approach supports queries to the cloud in the general select-join-projection format (Blanas et al., 2010; Chandar, 2010) translated from user natural language queries. For complex analytics, the approach adopts a model-based approach, where Bayesian belief network (Das, 2008) models are built in combination with automated machine learning and subject-matter expert consultation. A model is fragmented into its sub-parts with a view that, like distributed data sources, individual proprietary model parts are to be constructed and maintained by distributed autonomous agencies. The approach employs a distributed fusion algorithm (Das, 2012) on fragmented Bayesian network models by taking into account pedigreed data used as evidence.

## 3. The Architecture

Figure 1 shows the architecture of our envisioned integrated environment adopted from our ongoing effort. The upper half is a client-server setup to process user queries. The lower half of the environment shows a typical data-intensive environment consisting of various cloud- and non-cloud-based autonomous data sources.

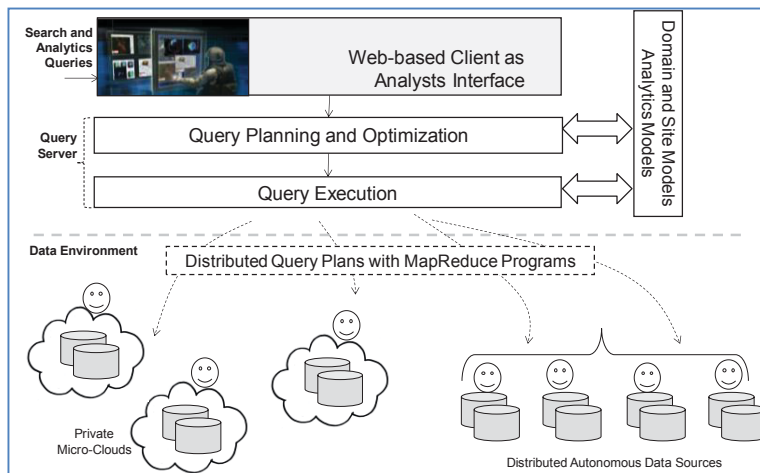
The Web-based Analyst Interface client communicates with a Query Server component over the internet or a secured connection. The interface allows a user to specify search and analytics queries in a declarative manner in a natural-language-like syntax with constrained vocabulary or via a high-level query language such as SQL. One of the major advantages of a web interface, as opposed to running on a client machine, is the increased security of having control over the server code and the agents that it spawns. The Query Server has two primary modules: Query Planning and Optimization, and Query Execution.

A set of sub-queries is generated in the Query Planning and Optimization module corresponding to a high-level search and analytics query posed by a human analyst. The module uses the local Domain and Site Model database that contains data site descriptions and domain models. To maximize retrieval efficiency, the ordering of sub-queries is optimized by the module. An execution plan for the sub-queries is then passed to the Query Execution module, which is responsible for generating and spawning the actual agents. If a part or the whole query is to be executed on

the cloud, the module generates an appropriate program with embedded map and reduce functions following the MapReduce framework. The optimization strategy here is to best exploit the in-built distributed execution and parallelism of the cloud.

The Query Execution module is shown communicating with other (authorized) resident agents accessing several distributed data sources and the clouds. We have the capability of spawning mobile agents in a controlled manner to be sent to the data sources, but this mechanism requires further study to confirm security. A cloud is considered just another data source, and hence more than one cloud can be considered. An agent is communicated for each sub-query retrieving answers to its sub-query from the appropriate data sources, including the cloud.

Our approach to complex analytics is model-based, in the sense that we have the Analytical Model library of complex analytical models for situation assessment. We choose a model depending on the situation assessment task posed by the user, and evidence is then propagated into the model as observed. Since the data sources are distributed, we need to distribute the computations involved in evidence propagation. The Query Planning and Optimization component formulates a plan to be sent to the Query Execution module, which then establishes communications with agents. The Query Execution module employs a distributed belief propagation algorithm.



**Figure 1: Cloud Analytics and Distributed Fusion Environment**

The Query Execution module employs a distributed belief propagation algorithm.

#### 4. Experimental Implementation

The current prototype demonstrated a search query in the context of the vignette involving retrieval of distributed and cloud-based terrain mobility and SALUTE data: *Select all sensing platforms from which reports have been generated for named areas of interest with no go mobility.* The prototype was implemented on a cloud- and agent-based distributed environment in Java on a 3-node experimental Local Area Network setup. The environment makes use of the agent platform Aglet, and Hadoop and MapReduce frameworks. For the demonstration of distributed analytics, we selected a “nuclear proliferation” scenario. We answered the following question in the context of the scenario: *Determine the level of nuclear proliferation threat from a rogue nation attempting to develop nuclear weapons.* We built a distributed Bayesian network computational model to determine, via a distributed fusion algorithm (Das, 2012), the degree of nuclear proliferation by a rogue nation. The model computed the posterior probability of nuclear threat given all the accumulated evidence on the variables of the model. This distributed approach has the potential of integrating results from distributed and disparate analytical frameworks to produce a common operational picture.

#### 5. Selected References

- [1] Blanas, S., et al. (2010). “A comparison of join algorithms for log processing in mapreduce,” Proceedings of the International Conference on Management of Data (SIGMOD), pp. 975–986, New York, NY, ACM.
- [2] Chandar, J. (2010). “Join Algorithms using Map/Reduce,” M.Sc. Thesis, University of Edinburgh, Scotland.
- [3] Das, S. (2008). *High-Level Data Fusion*, Artech House, Norwood, MA.
- [4] Das, S. (2012). “A Framework for Distributed High-Level Fusion,” In *Net-centric Distributed Fusion*, Hall, Llinas, Liggins, and Chee-Chong (eds.), CRC Press/Taylor & Francis.
- [5] Das, S. (2013). “Computational Business Analytics,” CRC Press/Chapman and Hall, *forthcoming*.
- [6] DoD. (2001). *Network Centric Warfare*, Department of Defense, Report to Congress.
- [7] Dean, J. and Ghemawat, S. (2008). “Mapreduce: simplified data processing on large clusters,” *Communication of the ACM*, Vol. 51(1), pp. 107–113.
- [8] Stonebraker, M., Abadi, D., DeWitt, D., Madden, S., Paulson, E., Pavlo, A., and Rasin, A. (2010). “MapReduce and Parallel DBMSs: Friends or Foes?” *Communication of the ACM*, Vol. 53, No. 1.

# Fusion Utility, Search, Index, Obtain, and Navigate (FUSION) over Enormous Data



**Erik Blasch**

AFRL/RIEA

Guna Seetharaman, Alexander J. Aved, and James Nagy

Sponsor: AFRL



## Outline

### Context search and indexing for content retrieval

- **Challenges**
- **Fusion:** data mining and association,
- **Utility:** metrics of data uncertainty for information quality,
- **Search:** an ontology for efficient queries,
- **Index:** through metadata storage techniques,
- **Obtain:** data access and dissemination, and
- **Navigate:** presentation and linking

**Application:** Cloud Computing for Wide-Area Motion  
Imagery Target Tracking and Identification

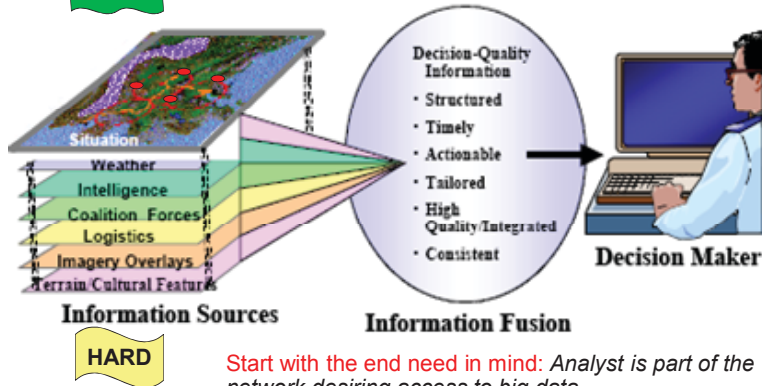


## Navigate: Presentation and Linking

Integration of product information

**SOFT**

• HUMINT Reports



Erik Blasch -SPIE13

3



## Obtain: Data Access and Dissemination

The **user** of information fusion systems **desires** attributes of

- (1) accurate and timely information for situational awareness;
- (2) ability to control and manage resources due to varying conditions, and
- (3) predictive capabilities to cue attention.

**Information fusion common theme**

*all the data (high volume) is sent to a similar location.*

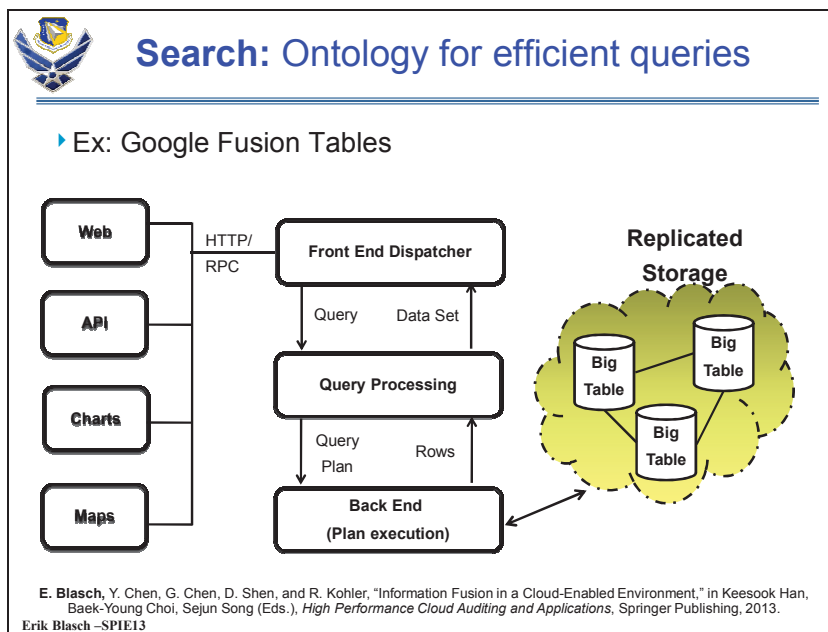
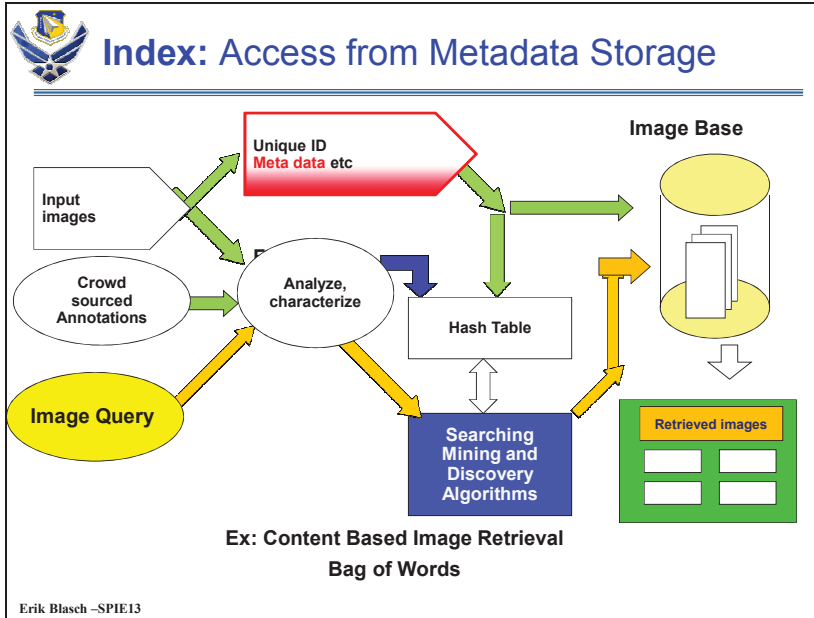


Data access from distributed sources requires methods of information management to enable processing, exploitation and dissemination (PED)

Erik Blasch -SPIE13

4



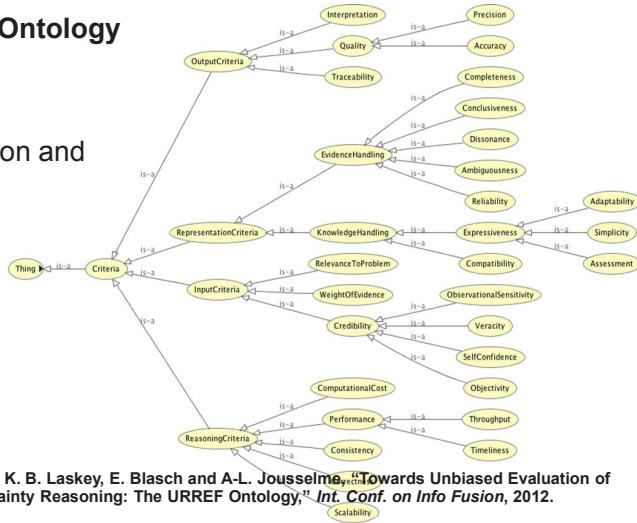




# Utility: Metrics of data uncertainty

## Example of Ontology

Uncertainty  
Representation and  
Reasoning  
Evaluation  
Framework  
(URREF)

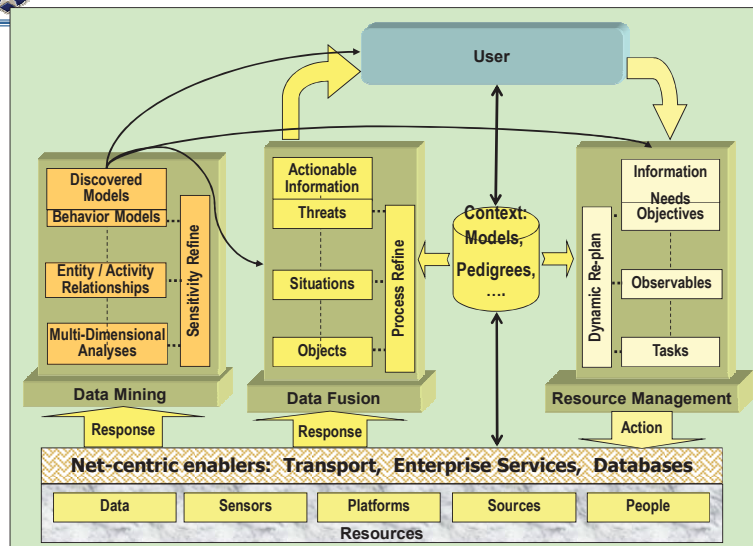


P. C. G. Costa, K. B. Laskey, E. Blasch and A-L. Joussem, *Towards Unbiased Evaluation of Uncertainty Reasoning: The URREF Ontology*, *Int. Conf. on Info Fusion*, 2012.

Erik Blasch –SPIE13



# Fusion: data mining and association



Erik Blasch –SPIE13



## Outline

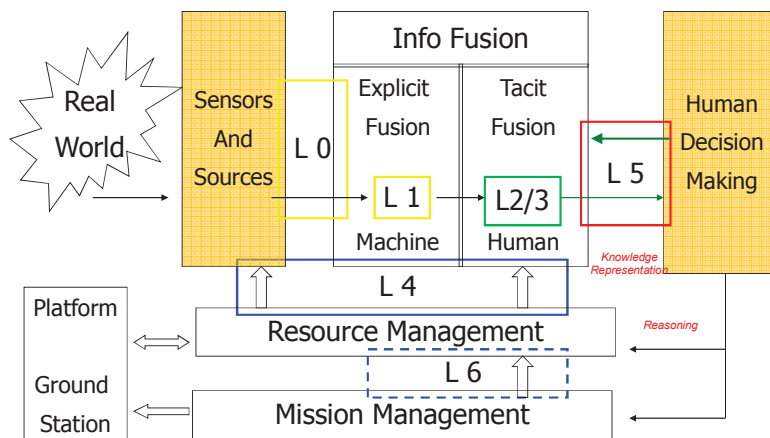
### Context search and indexing for content retrieval

- **Challenges**
- **Fusion:** data mining and association,
- **Utility:** metrics of data uncertainty for information quality,
- **Search:** an ontology for efficient queries,
- **Index:** through metadata storage techniques,
- **Obtain:** data access and dissemination, and
- **Navigate:** presentation and linking

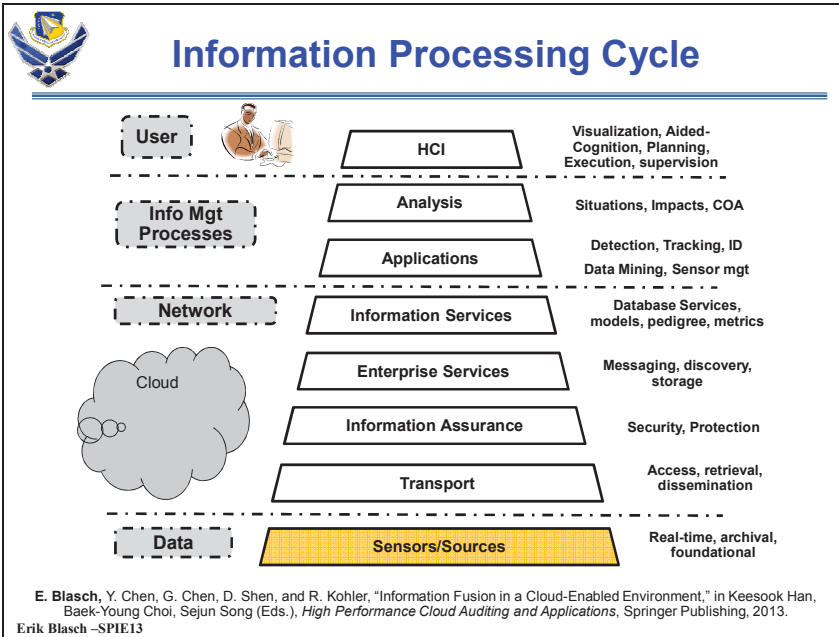
**Application:** Cloud Computing for Wide-Area Motion Imagery Target Tracking and Identification



## DFIG - Fusion Model Target Tracking (L1) to User Refinement (L5)



E. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, G. M. Powell, D. D. Corkill, and E. H. Ruspini, "Issues and challenges of knowledge representation and reasoning methods in situation assessment (Level 2 Fusion)", *J. of Advances in Information Fusion*, Dec. 2006.



**Big Data Example: Tracking Analysis**

- ▶ H. Ling, Y. Wu, E. Blasch, G. Chen, H. Lang, and L. Bai, *Fusion 2011*
- ▶ Used CLIF (Columbus Large Image Format) EO sample data set
  - Used six tracking algorithms for four targets
  - Reported the results (Modified for track confidence over time)
  - Worked with the management of the data

**Test Process**

**Data sets** ↓

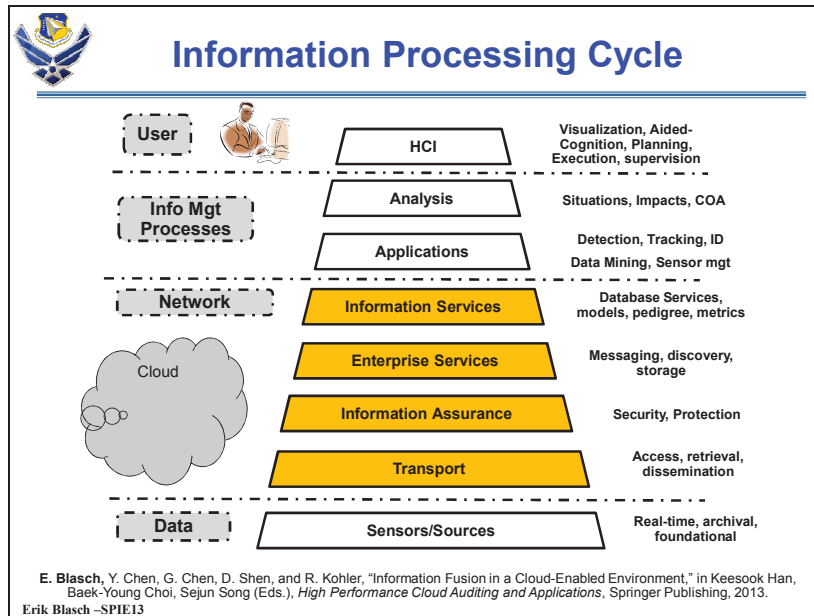
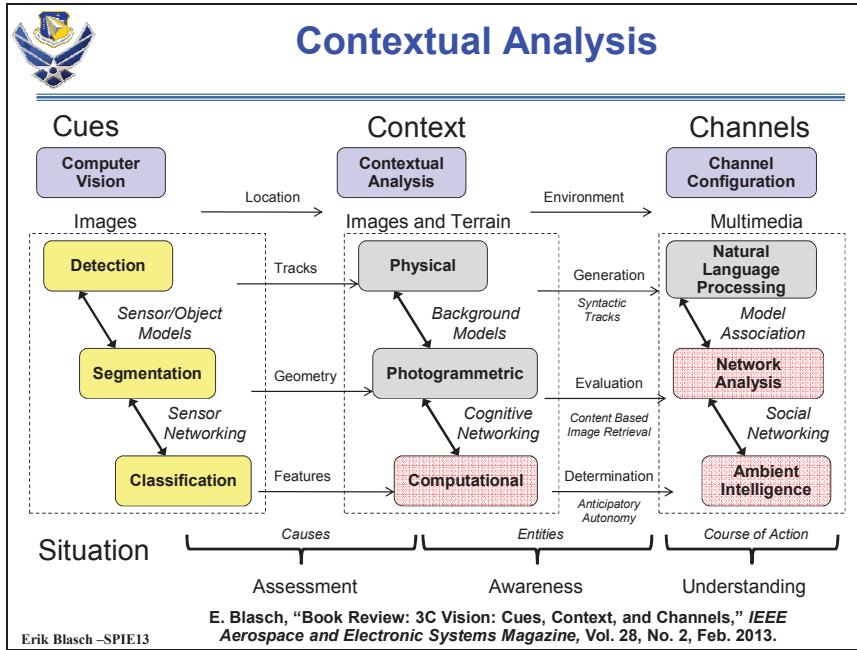
#28   #29   #30

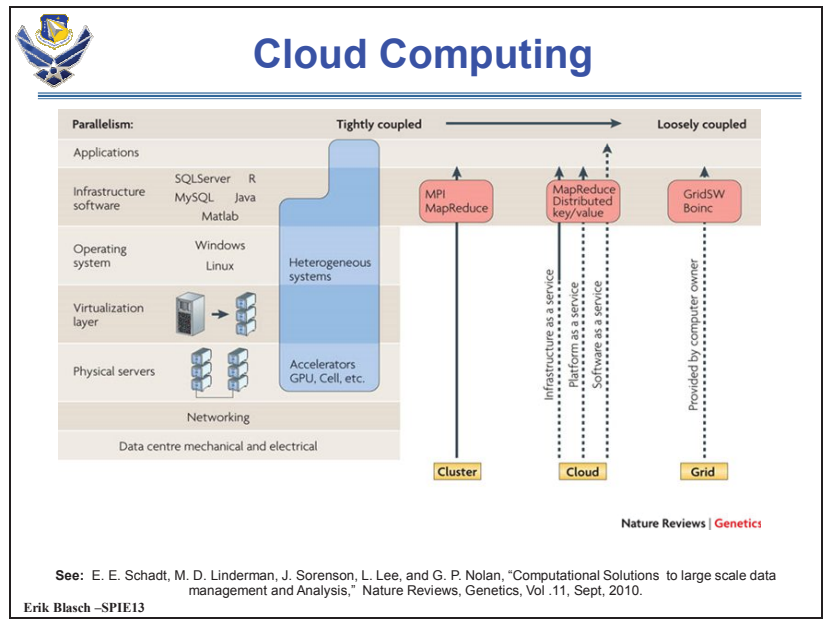
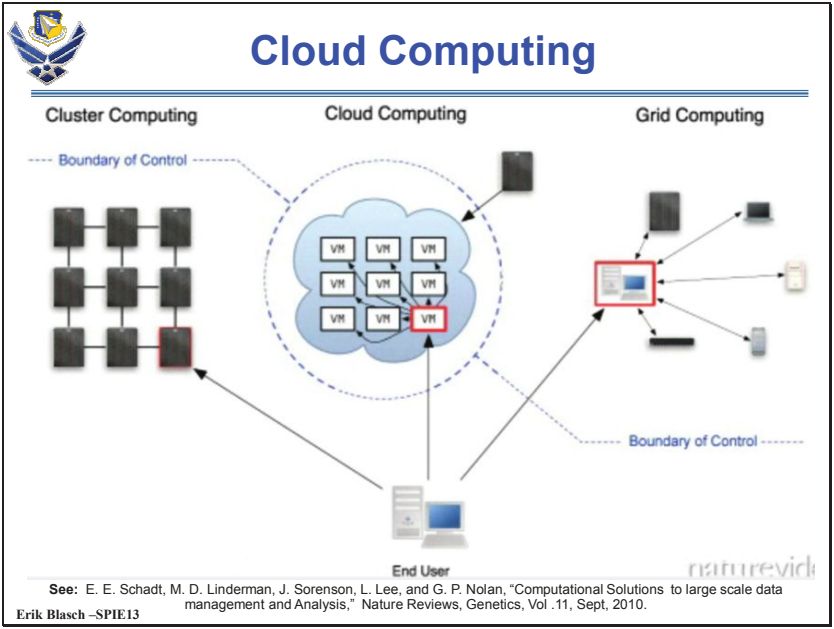
**Evaluation**

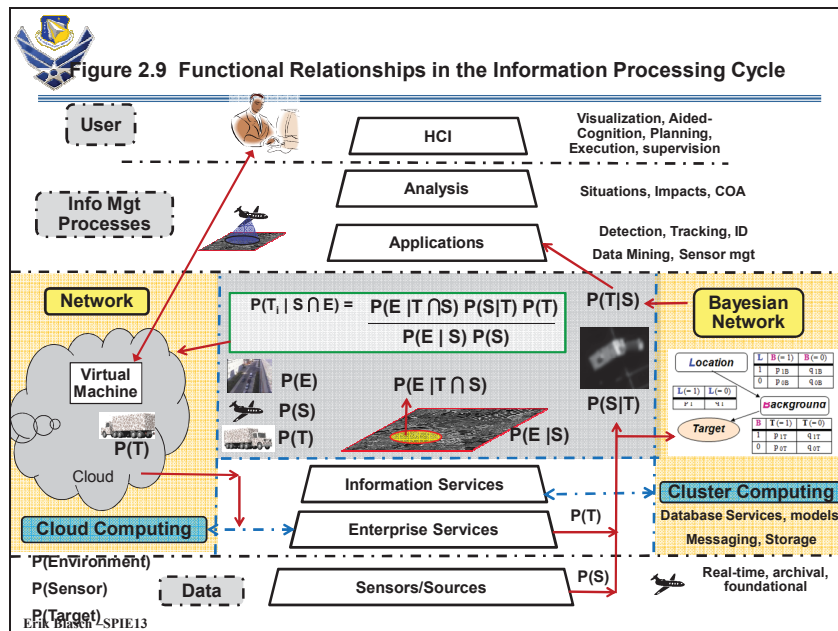
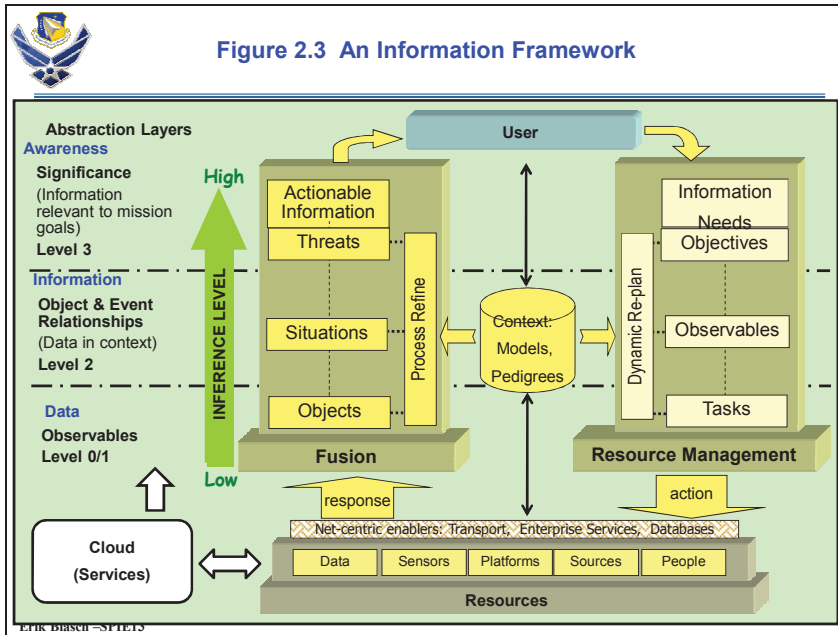
— Annotation — CPF — HPF — MS — MIL — L1-BPR

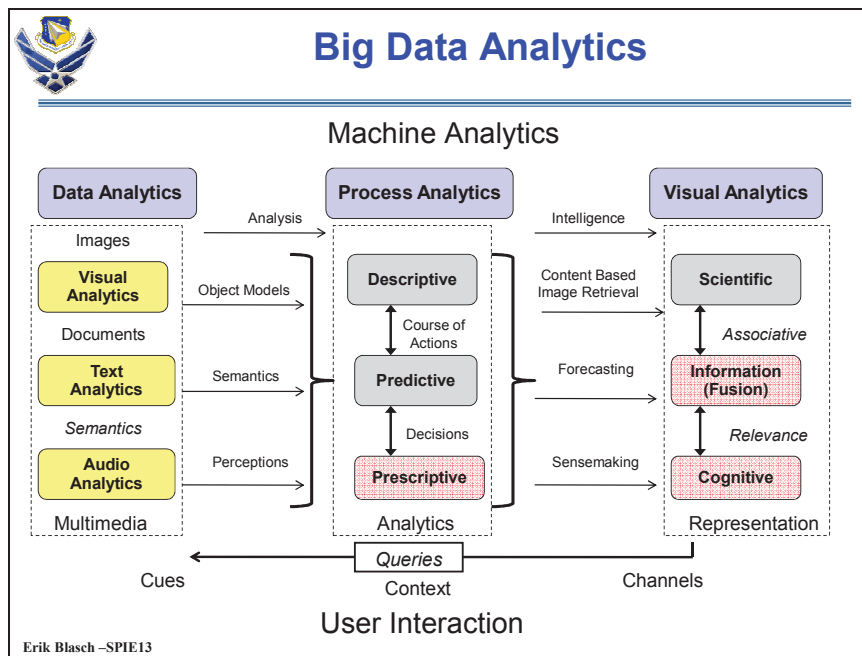
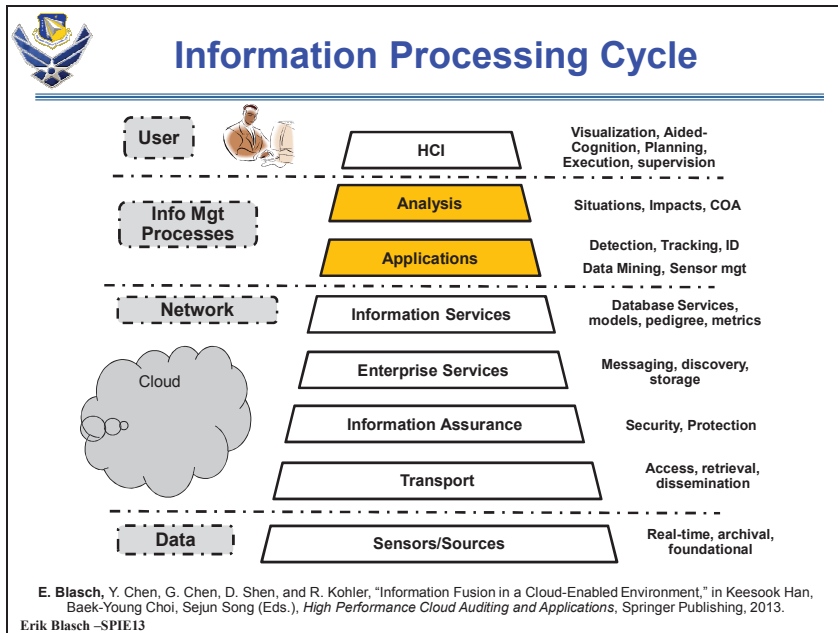
- Annotated feature analysis using different tracking results: covariance-based particle filter (CPF), intensity histogram PF (HPF), mean-shift (MS), multiple instance learning (MIL), and the Bounded Particle Resampling L1 (L1-BPR) tracker and Annotation is the truth.

Erik Blasch –SPIE13      12













## Machine and Visual Analytics

**Machine analytics (MA)** covers the broad spectrum of applications

(1) descriptive, prescriptive, predictive analytics

**Descriptive** means to diagnose the situation;

**Predictive** is to suggest course of actions profile and trending.

**Prescriptive analytics** suggests decision options over the predictions.

**Visual Analytics (VA)** seeks

(2) scientific, information, and cognitive analytics.

**Scientific visualization** deals with data that has a natural geometric structure (e.g., image data),

**Information visualization** handles abstract data structures such as trees or graphs for communication, and

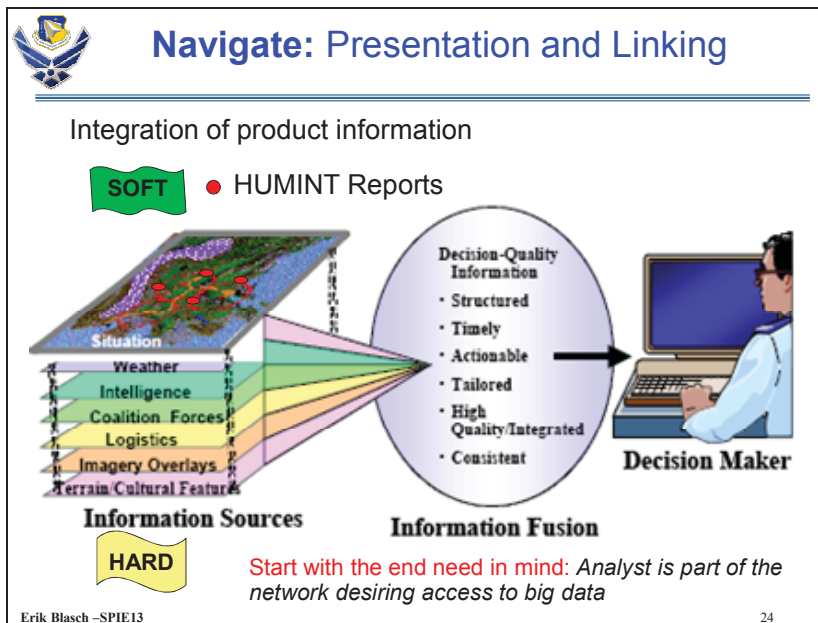
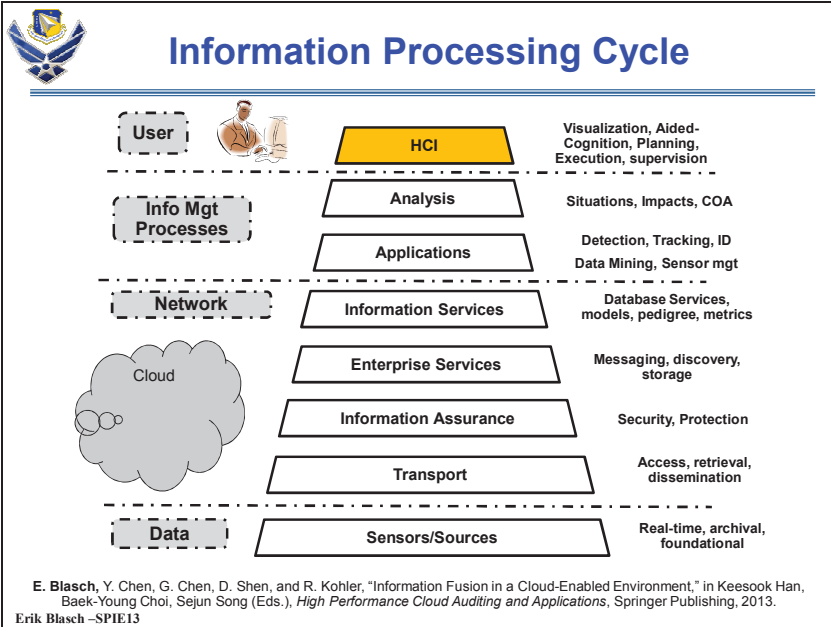
**Cognitive visualization** concerns sensemaking and reasoning.



## Machine and Visual Analytics

### Analytics and Information Fusion.

Fusion	Machine	Concept
Level 0	Scientific	Access to data and pedigree of information and issues of structured/unstructured data
Level 1	Information (Images, text)	Development of graphical methods for data analysis
Level 2	Descriptive	Uses data mining to estimate the current state (i.e. Machine learning) over different reasoning of trends for modeling
Level 3	Predictive	Future options from current estimates
Level 4	Prescriptive	Sequencing of selected actions
Level 5	Visual	Sensing Making and Reasoning
Level 6	Activity-Based Analytics	Policy instantiation of desired outcomes as to a focused mission





## Summary

### Challenges for Big Data Analysis include:

- 1) **Big Data accessibility and Analytic Representations** (context, environments, and processes over text, video, and data) for association information management (data mining, situation awareness);
- 2) **Decision support search and retrieve processes** (reasoning, inference, and explanation of relationships) from queries to answer user's needs;
- 3) **Standardized evaluation methods** (measures of performance/ effectiveness, and empirical case studies) to conduct evaluations over the various analytics (predict, search, extract, and match) using intelligent metadata;
- 4) **Systems design techniques for User Refinement** (scenario-based, user-based, and distributed-agent) to provide reasoning capabilities and the ability for user interaction (e.g., annotations, visualization, tag, label, reporting); and
- 5) **Advances in Big Data Processing** (semantic, knowledge, and complex) for acquisition, relevancy, and processing of data and information to support graphical models for descriptive, predictive, and prescriptive analytics.

# Fusion Utility, Search, Index, Obtain, and Navigate (FUSION) over Enormous Data

Erik P. Blasch, Guna Seetharaman, Alexander J. Aved, and James Nagy  
 Air Force Research Laboratory  
 Rome, NY 13441

## ABSTRACT

It is said that a picture is worth a thousand words, wherein complexity is distilled quickly to a common framework. A picture affords domain registration, signal correlation, entity association, and user appeal; however it is assumed that the context exploitation of the picture is well-characterized for content extraction. With enormous amounts of data available from videos, images, documents, emails, and news reports; there is a need to do context search and indexing for content retrieval. The challenges include (1) Fusion: data mining and association, (2) Utility: metrics analysis of data uncertainty for information quality, (3) Search: an ontology for efficient queries, (4) Index: through metadata storage techniques, (5) Obtain: data access and dissemination, and (6) Navigate: presentation and linking. All of these issues are important for enormous data coordination for a machine or user that seeks to relevant information to answer queries.

**Keywords:** DFIG Fusion Model, Content-base image retrieval, Big Data, Machine Analytics

**INTRODUCTION:** With the advent of web technology, the real-world challenges and issues to information fusion [1] have progressed from low-level information fusion (object assessment of tracking and classification/identification [2, 3, 4]) to high-level information fusion of situation/impact assessment (SA/IA), situation awareness (SAW), and information management (IM), shown in Figure 1, [5]. IM seeks to coordinate the information fusion products with the user's needs (Figure 2) such as objects, activities, events, and relationships among them over geospatial, temporal, and semantic properties [6]. Graphical models [7] expand on Bayes Nets have been applied to many types of data individually (e.g., text and image processing) from which parts are probabilistically related in a common framework and offers scalability. Future challenges are to use methods such as graphical methods over heterogeneous data in a common framework. As an example from computer vision (which has counterparts in text processing); issues are (A) scalability of fusion methods and frame-based processing, (B) efficiency and effectiveness of partitioning of distributed computing models, (C) context search in data space, (D) indexing over primary hazards such as invariants, and (E) retrieval for content characterization and perceptual group. As an example, *content-based image retrieval* (CBIR) over enormous data will require retrieval for fusion of physics-based and human-based information. Future efforts will require quickly gathering both the picture (context exploitation) and the thousand words (content extraction) simultaneously [8].

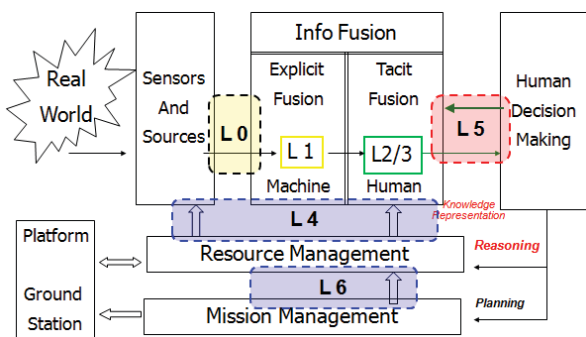


Figure 1 - DFIG Information Fusion model (L = Level)

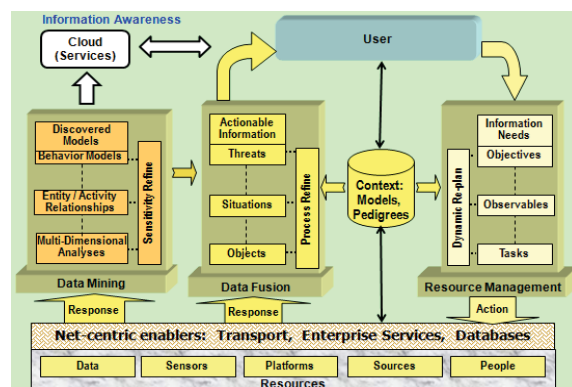


Figure 2 – Information Fusion Enterprise Model [9]

With the enormous amount of data types, distributed locations, and various connections to different applications (e.g. finance to surveillance) resulting from the expansion of the World-Wide Web, new techniques are needed. Related concepts recently emerging are context awareness (Figure 3) and machine, descriptive, prescriptive, predictive, visual, and other analytics (Figure 4). There are three issues of importance *hardware* (e.g., Apache Hadoop data intensive

distributed architecture), *software* (e.g., machine analytics), and *user/domain* applications [10] (e.g. visual analytics, text analytics).

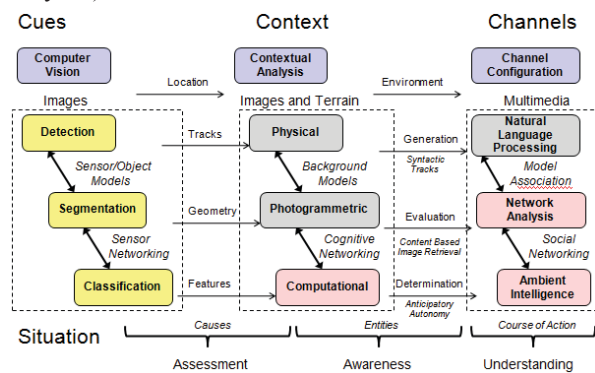


Figure 3 – Contextual Awareness and Understanding [11]

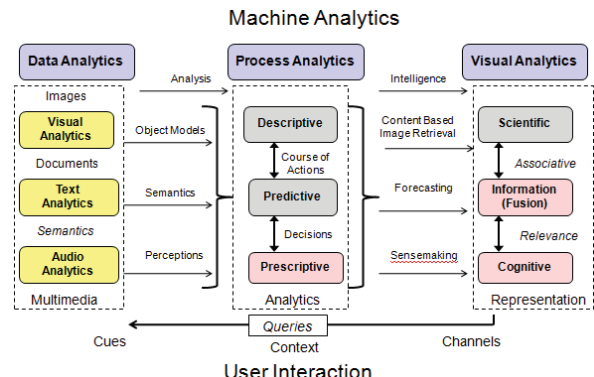


Figure 4 – Big Data Analysis

**Machine analytics (MA)** covers the broad spectrum of applications for data analysis that include: physics-derived sensor (e.g. video), human-derived (e.g. text), and machine (e.g., web files) data. Machine analytics is also based on the processes from which the literature (and business world) discuss man-machine and machine-machine interactions. Inside machine analytics are the emerging concepts of (1) descriptive, prescriptive, predictive analytics [12, 13], and (2) scientific, information, and cognitive analytics. Descriptive means to diagnose the situation; while predictive is to suggest course of actions profile and trending. Using the predictive information, prescriptive analytics suggests decision options over the predictions.

**Visual Analytics (VA)**, as another emerging concept [14], seeks scientific, information, and cognitive representations. Visualization supports planning, and decision making through effective data representations and transformations over physical- and human-derived data (sometimes referred to hard-soft fusion) as well as visualization analytical reasoning. Finally, the interactions of users with machines [15] is important for the collection, exploitation, presentation, and dissemination of data.

- Scientific visualization deals with data that has a natural geometric structure (e.g., image data),
- Information visualization handles abstract data structures such as trees or graphs for communication, and
- Cognitive visualization concerns sensemaking and reasoning.

If we look at “analytics” it mirrors the Data Fusion Information Group model (Figure 1) in having both the “reasoning” (e.g., Bayes) and “management” (i.e., control) functions. Thus, the MA is like reasoning, while VA is about management. Merging analytics functions with FUSION techniques for enormous data is highlighted below:

Fusion	Machine	Concept
Level 0	Scientific	Access to structured/unstructured data and pedigree of information
Level 1	Information (Images, text)	Development of graphical methods for data analysis
Level 2	Descriptive	Uses data mining to estimate the current state (i.e. Machine learning) over different reasoning methods to gather trends for modeling
Level 3	Predictive	Future options from current estimates
Level 4	Prescriptive	Sequencing of selected actions
Level 5	Visual	Sensing Making and Reasoning
Level 6	Activity-Based Analytics	Policy instantiation of desired outcomes as to a focused mission

CBIR, as related to the other techniques including deep learning, cloud computing [16], prescriptive analytics, and graphical methods. There is both a need for *sequential* operations for real time operator query analysis (e.g., Fusion Cell) as well as parallel opportunities to enable multiple screens for big data analysis of information to process data for non-real time analysis over cultural assessments. Key dimensions coordinating these techniques are:

**Big Data Fusion for Contextual Awareness**

- FUSION: Data Mining and Data Fusion (e.g., graphical and probabilistic models)
- UTILITY: Cost functions and metrics (e.g, data uncertainty, quality of service, information quality)

- SEARCHING: DataBase Ontology (e.g., RDF, queries for indexing)
- INDEXING: Storage and metadata (e.g., Google Tables)
- OBTAINING: Information retrieval tools for Human Interaction (e.g., user-defined operating picture)
- NAVIGATION: visualization and efficient processing of enormous data (e.g., through data clouds)

**Issues** for Big Data modeling for Computer Vision currently open are:

- 1) FUSION: Scalability and Frame based processing (real time association, ubiquity in data sets, and pedigree of data collection source)
- 2) UTILITY: Partition of distributed computing models, ontologies of data uncertainty [17]
- 3) SEARCHING: Context (search) in data space (versus time, location, and frequency)
- 4) INDEXING: Primary hazards due to occlusions and lack of invariants requires indexing
- 5) RETRIEVAL: Challenges in content characterization and perceptual grouping (e.g., multiple objects with the same features for inter/intra cluster discernability)

Our brief paper has outlined the need for big data processing in the development of future information fusion systems designs. While the explosion of techniques for data mining, machine analytics, cloud computing [18], and multimedia; future systems will require refinement of techniques to support applications through queries [19] to extract meaningful content, sensemaking, and information relevance to a Dynamic Data Driven Applications System (DDDAS) [20].

**Challenges** for Big Data Analysis include:

- 1) *Big Data accessibility and Analytic Representations* (context, environments, and processes over text, video, and data) for association information management (data mining, situation awareness);
- 2) *Decision support search and retrieve processes* (reasoning, inference, and explanation of relationships) from queries to answer user's needs;
- 3) *Standardized evaluation methods* (measures of performance/ effectiveness, and empirical case studies) to conduct evaluations over the various analytics (predict, search, extract, and match) using intelligent metadata;
- 4) *Systems design techniques for User Refinement* (scenario-based, user-based, and distributed-agent) to provide reasoning capabilities and the ability for user interaction (e.g., annotations, visualization, tag, label, reporting); and
- 5) *Advances in Big Data Processing* (semantic, knowledge, and complex) for acquisition, relevancy, and processing of data and information to support graphical models for descriptive, predictive, and prescriptive analytics.

## REFERENCES

- [1] Blasch, E., Kadar, I., Salerno, J. J., Kokar, M. M., Das, S., Powell, G. M., Corkill, D. D., and Ruspini, E. H., "Issues and Challenges in Situation Assessment (Level 2 Fusion)," *J. of Adv. in Info. Fusion*, Vol. 1, No. 2, pp. 122–139, December 2006.
- [2] Blasch, E., and Kahler, B., "Multi-resolution EO/IR Tracking and Identification" *Int. Conf. on Info Fusion*, (2005).
- [3] Blasch, E., Banas, C., Paul, M., Bussjager, B., and G. Seetharaman, "Pattern Activity Clustering and Evaluation (PACE)," *Proc. SPIE*, Vol. 8402, (2012).
- [4] Blasch, E., Ling, H., Wu, Y., Seetharaman, G., Talbert, M., Bai, L., and Chen, G., "Dismount Tracking and Identification from Electro-Optical Imagery," *Proc. SPIE*, Vol. 8402, (2012).
- [5] Blasch, E., Bosse, E., and Lambert, D.A., [*High-Level Information Fusion Management and Systems Design*], Artech House, (2012).
- [6] Blasch, E., Deignan, Jr. P. B., Dockstader, S. L., Pellechia, M., Palaniappan, K., and Seetharaman, G., "Contemporary Concerns in Geographical/Geospatial Information Systems (GIS) Processing," *Proc. IEEE Nat. Aerospace Electronics Conf.*, (2011).
- [7] Chong, C-Y., and Mori, S., "Graphical Models for Nonlinear Distributed Estimation," *Int. Conf on Information Fusion*, (2004).
- [8] Bowman, L., "Persistent ISR: the social network analysis connection," *Proc. SPIE*, Vol. 8389, (2012).
- [9] Blasch, E., Steinberg, A., Das, S., Llinas, J., Chong, C.-Y., Kessler, O., Waltz, E., and White, F., "Revisiting the JDL model for information Exploitation," *Int'l Conf. on Info Fusion*, (2013).
- [10] Blasch, E., "Level 5 (User Refinement) issues supporting Information Fusion Management" *Int. Conf. on Info Fusion*, (2006).
- [11] Blasch, E., "Book Review: 3C Vision: Cues, Context, and Channels," *IEEE Aerospace and Electronic Systems Mag.*, Vol. 28, No. 2, Feb. (2013).
- [12] [http://en.wikipedia.org/wiki/Prescriptive\\_analytics](http://en.wikipedia.org/wiki/Prescriptive_analytics)
- [13] Das, S., [*Computational Business Analytics*], CRC Press/Chapman & Hall (2013).
- [14] [http://en.wikipedia.org/wiki/Visual\\_analytics](http://en.wikipedia.org/wiki/Visual_analytics)
- [15] Blasch, E. and Plano, S., "JDL Level 5 Fusion model 'user refinement' issues and applications in group Tracking," *Proc. SPIE*, Vol. 4729, 2002.
- [16] Blasch, E., Chen, Y., Chen, G., Shen, D., and Kohler, R., "Information Fusion in a Cloud-Enabled Environment," in K. Han, B.-Y Choi, S. Song (Eds.), [*High Performance Cloud Auditing and Applications*], Springer Publishing, (2013).
- [17] Costa, P. C. G., Laskey, K. B., Blasch, E., and Jousselme, A-L. "Towards Unbiased Evaluation of Uncertainty Reasoning: The URREF Ontology," *Int. Conf. on Info Fusion*, (2012).
- [18] Liu, B., Chen, Y., Blasch, E., Pham, K., Shen, D. and Chen, G., "A Holistic Cloud-Enabled Robotics System for Real-Time Video Tracking Application," *Int'l Conf. on Info Fusion*, (2013).
- [19] Aved, A. J, [*Scene Understanding for Real Time Processing of Queries over Big Data Streaming Video*], PhD Dissertation, Univ. of Central Florida, (2013).
- [20] Blasch, E., Seetharaman, G., and Reinhardt, K., "Dynamic Data Driven Applications System concept for Information Fusion," *Int'l Conf. on Computational Science*, (2013).

## **Perceptual Reasoning Managed Big Data Analytics and Information Fusion**

Ivan Kadar  
Interlink Systems Sciences, Inc.  
Lake Success, NY, USA  
29 April 2013

Invited Panel Discussion on “Real-World Issues and  
Challenges in Big Data Processing with Applications to  
Information Fusion”

SPIE Conference 8745 “Signal Processing, Sensor Fusion and Target  
Recognition XXI”, 29 April 29 – 2 May 2013, Baltimore, MD

### **Outline**

- **Challenges:** The Problem Setting – Importance of the analyst end-user being the system manager & modeling the analyst’s cognitive functions -
- Relationship of Big Data Analytics and Information Fusion
- Visual Analytics processing components
- Cognitive Models of Intent - Perception and Perceptual System to model human thinking and thought processes (cognition)/Analyst Modeling
- The Perceptual Reasoning Machine (PRM) paradigm - emulate/model/aid Analyst
- Fusion Levels Interaction via PRM Mapping – PRM elements information flow and relationship to the Joint Director of Laboratories (JDL) and Data Fusion Information Group (DFIG) models
- Cognitive Models of Intent (SA/TA) in Social Networks over enormous data
- Application of Information Process Model System (PMS) to Big Data Analytics and Information Fusion
- Other Applications: Complex Adaptive Systems and Networks





## Visual Analytics Processing Components

- Problem Specific Issues and Challenges in BDA and Information Fusion:
- Tasks in BDA Processing Chain in Visualization Processing, subsequent to data mining/machine learning is “**Visual Analytics**” [1] providing processing steps necessary for data analysis using visualization techniques, which can include:
  1. Computational Data Analysis: modeling and transformation of process that prepares purely numerical data for the visualization process.
  2. Interactive Visual Interfaces: methods for explanatory analysis even to examine unstructured text and to find the key information hidden within
  3. *Analytical Reasoning*: focuses on people how they think, how they generate hypotheses to reach decisions, that is the human cognitive reasoning process.
- Methods to aid **Analytical Reasoning** to facilitate the cognitive function include:
  - (a) **Bayesian Analysis in Visual Analytics (BAVA) [1a]**: data transformation from deterministic to probabilistic Bayesian methods, enabling analysts to quantify data uncertainty, include expert judgment into analyses, rapidly generate and test new hypotheses, and allow multi-source and multi-scale data to contribute to one data display.
  - (b) **Multi-Source Visual Analytics [1b]**: dimensionality reduction step to visualize the data for data analysis, such as searching, clustering, and the detection of outliers. Method is based on multiple kernel learning (MKL) using unsupervised learning and allows fusing a multitude of heterogeneous independently collected data.

[1] Georgia Tech's "Foundations of Data and Visual Analytics (FODAVA)" Program Outline in "Horizons, Vol. 30, No.1, Fall 2012 –Winter 2013" and Research Topics [1a, 1b] [www.fodava.gatech.edu](http://www.fodava.gatech.edu) "

## Components of BDA Processing Involving the Human's Role

- **Predictive Analytics – Data Mining**
- Components of **Visual Analytics** : *Interactive Visual Interfaces* and
- “**Analytical Reasoning**” - the “*cognitive model*” (modeling how people think: learn, anticipate, plan, predict, fuse information, generate hypotheses, manage resources and reach decisions, *as done by PRM*)
- **Case-Based Reasoning** [1] - uses a library of past experiences. For a new problem the system searches the case library to match the new problem with similar problems (*alike to associative recall by humans, used in PRM*). Once the closest match is found, its solution is adapted and reused to solve the current problem.

## Example: Cognitive Models of Intent – Issues and Challenges

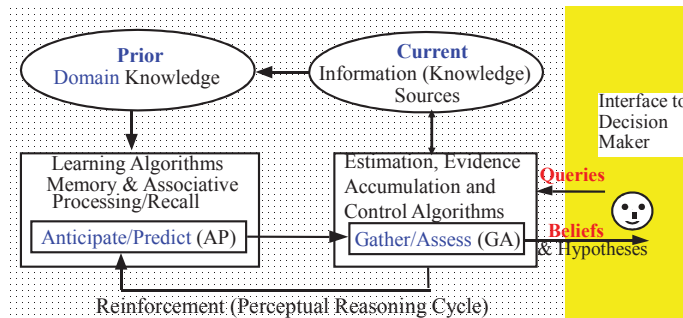
- The capability to “sense/observe, mine/access” data, learn, *associate*, recall, anticipate and predict/act” are key ingredients of **human perceptual reasoning**. These attributes are necessary constructs in **cognitive modeling**.  
**The key ingredient is timely information access in modeling intent\*\*.**
  - Cognitive models** are imbedded in large family of methods called **Predictive Analytics/Modeling** (techniques to predict future entities) including:
    - sensing/collecting, data mining, sorting, organizing, aligning, associating, fusing, and using a-priori and learned, SME based and current data
  - Predictive models using algorithms, such as:** machine adaptive learning, SVMs, regression, neural and abductive networks, classifiers, feature selection, distance measures, Bayes-nets/ influence diagrams, logic, decision making under uncertainty,...., *have been used in intent modeling, but did not use a cognition framework, which uses above methods/algorithms, the cognitive PRM paradigm:*
    - The **Perceptual Reasoning Machine (PRM)** [2-4]: a “meta-level information management system”, for adaptive information gathering/assessment, learning, anticipation, and prediction – **emulating/modeling the analyst**
- Objectives of models?** : minimize uncertainty and maximize the value of deduced information to **detect/identify potential intent**, and to act in a real-time environment with time constraints – (by modeling/aiding analyst by PRM)

[2] I. Kadar, “Perceptual Reasoning In Adaptive Fusion Processing”, *Signal Processing, Sensor Fusion and Target Recognition XI*, Ivan Kadar Editor, Proc. SPIE 4729, Orlando, FL., April 2002

[3] E. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, G. M. Powell, D. D. Corkill, and E. H. Ruspini, “Issues and challenges of knowledge representation and reasoning methods in situation assessment (Level 2 Fusion)”, *J. of Advances in Information Fusion*, 2006

[4] \*\* I. Kadar, “Issues and Challenges in Intent Modeling in Social/Cultural Networking Domain”, presented at Invited Panel Discussion: Real-World Issues & Challenges in Social/Cultural Modeling with Applications to Information Fusion, Co-Organizers John Salerno & Ivan Kadar, *Signal Processing, Sensor Fusion and Target Recognition XXI*, Ivan Kadar Editor, Proc. SPIE Vol. 8392, April 2012.

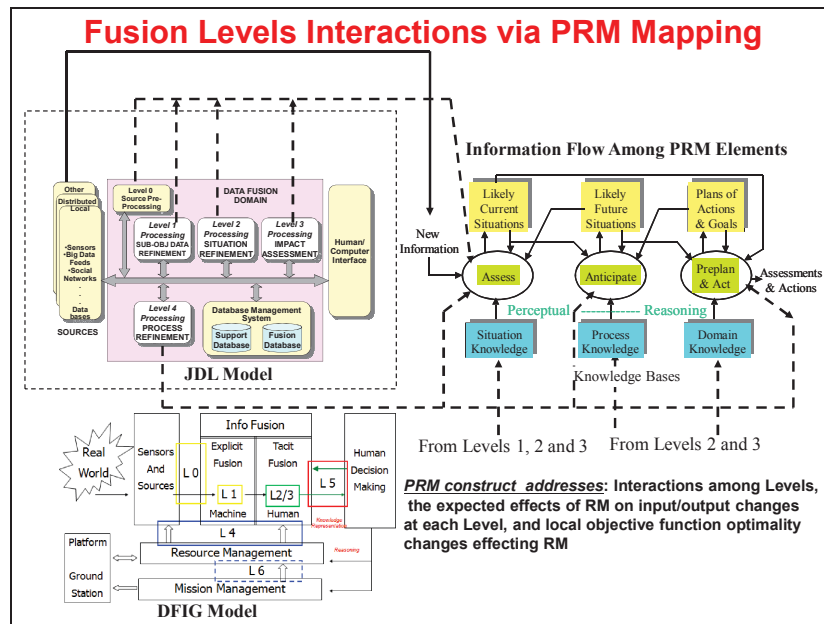
## Perceptual Reasoning Machine (PRM) Providing Adaptive Information Gathering, Assessment, Prediction and Control



**PRM construct provides: adaptive information gathering, assessment, prediction & control** - by taking into account interactions among system components, the expected effects of input/output changes of local optimality at each component, and global optimality changes effecting the control itself in a human-like perceptual reasoning framework: viz., - *In order to perceive one needs to:*

(1) *sense and deliver stimuli to the “system”.*

(2) *the “system” when properly “stimulated” delivers a feedback “reinforcement” to the “system” in order to modify its output and optimize objectives.*



### Cognitive Models of Intent – Issues and Challenges in Social Networking Over Enormous Data

- What is the Role of Social Networking in the PRM intent model framework?
  - Social Networks (SNs) **provide access to real-time information exchange** [derived\* **context (e.g., sentiments, emotions...)**], extracted from cultural/social interactions - messages with location and time stamped data] to be used as input to the model

Potential Issues and Challenges:

1. Is the extracted “big data” based on consensus of the population or only from “outliers”? (“outliers” can exert **\*\*influence**, coalesce and become significant intent indicators). Furthermore, how to handle potential data sparsity (per individual) vs. enormity (web) of data; and contextual validity into emotional aspects?
2. Is information exchange restricted globally by particular entities? (potential direct intent indicators)
3. How to “*associate*” massive information from multiple SNs as input to PRM ?

\* Note: The preprocessing of linguistic messages to learn, classify and group various context is assumed a given herein.

\*\*W. Pan, W. Dong, M. Cebrian, T. Kim, J. H. Fowler and A. (Sandy) Pentland, “Modeling Dynamical Influence in Human Interaction”, *IEEE Signal Processing Magazine*, March 2012.

## Cognitive Models of Intent – Issues and Challenges in Social Networking Over Enormous Data (Cont'd)

- **Information access** is crucial as an input both for real-time assessment, prediction and to data bases (learning) & for message rate “change detection” impending intent?

- Example: Ben Zimmer, “Twitterology: A New Science?”, *The New York Times*, October 30, 2011. The article illustrates the degree of relevant real-time information that can be derived from social/cultural interactions expressed in Twitter: (e.g., monitoring tweets to track on-the-ground *sentiment* over the course of the Arab Spring in Egypt & Libya to detect changes in *sentiments*)

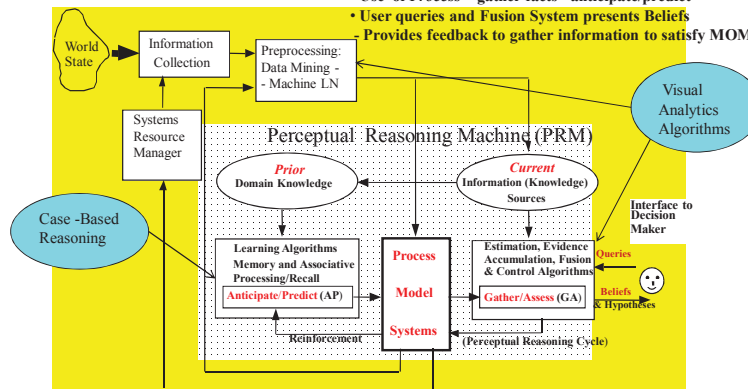
- The Machine Learning & linguistically extracted Twitter *context* information from messages can be used as **input to intent modeling: sentiments, emotions - moods, opinions\* etc. extracted context data (including locations, time, consensus types, groups and number of constituting elements or computed probabilities) used as input with other data sources to detect/ID potential intent via the cognitive PRM model –emulates/models interface to and role of Analyst**

\*A.Pak, and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

## Information “Process Model System” (PMS)

### Issues addressed

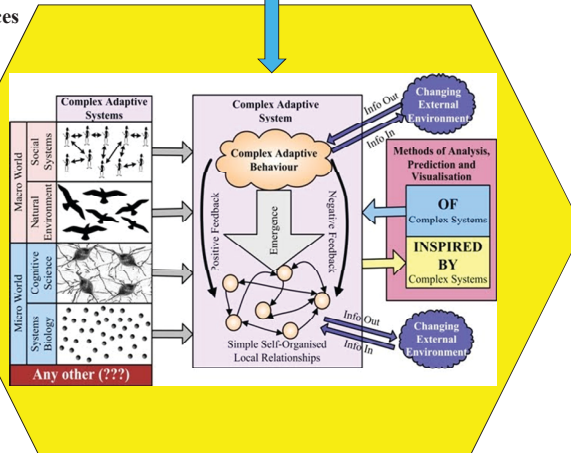
- Use of Knowledge – Prior, Learned and Current
- Use of Process – gather facts - anticipate/predict
- User queries and Fusion System presents Beliefs
- Provides feedback to gather information to satisfy MOMs



## Complex Adaptive Networks and Systems (CANSs)\*

Apps of PRM – Analyst Control - Perceptual Reasoning

- Big Data Sources
- Weather Data
- Flickr data
- Change Detect
- Sensors
- Big Data/Feeds
- Data Mining
- Visual Analytics
- Analysts Interaction Capability



\*<https://sites.google.com/site/cnskcl/interdisciplinary-Complex-Adaptive-Networks-&Systems@KCL>

### Summary

- There are many issues and challenges remaining requiring research, implementation, testing to validate the proposed methods. Questions?
- **Addressed:**  
  - Challenges: The Problem Setting – Importance of the analyst end-user being the system manager & modeling the analyst's cognitive functions -
  - Relationship of Big Data Analytics and Information Fusion
  - Visual Analytics processing components
  - Cognitive Models of Intent - Perception and Perceptual System to model human thinking and thought processes (cognition)/Analyst Modeling
  - The Perceptual Reasoning Machine (PRM) paradigm –emulate/model/aid Analyst
  - Fusion Levels Interaction via PRM Mapping – PRM elements information flow and relationship to the Joint Director of Laboratories (JDL) and Data Fusion Information Group (DFIG) models
  - Intent (SA/TA) Modeling in Social Networks over enormous data
  - Application of Information Process Model System (PMS) to Big Data Analytics and Information Fusion
  - Other Applications: Complex Adaptive Systems and Networks

# Perceptual Reasoning Managed Big Data Analytics and Information Fusion

Ivan Kadar  
Interlink Systems Sciences, Inc.  
1979 Marcus Avenue, Lake Success, NY 11042

## 1. INTRODUCTION

This succinct position paper, coupled with the associated viewgraphs, highlights real-world issues and challenges of the problem of handling, processing and using “Big Data” [1] sources coupled with the fusion process, with specific focus on modeling the human’s (analyst’s) cognitive role as the systems manager. Paraphrasing part of the description of panel discussion in the conference program, “the proliferation of data sources has created an urgent need to manage, collect/retrieve and make sense of “big data”. The big data problem is present in diverse areas such as: cybersecurity, financial and health analytics, smart cities, social media networks, digital video, text data, sensor networks, etc. Methods are needed to handle big data feeds, perform data mining/machine learning (components of Big Data Analytics “BDA” [1]) and information fusion, and provide real-time and near real-time information delivery for accurate decision making.

The focus and the challenge herein is to model an adaptive big data processing and information fusion system emulating the human perceptual reasoning/cognitive functions to facilitate two-way interactions between the visual data display and the analysts. This process is shown addressable in part by using Visual Analytics processing algorithms [2]. The proposed system is to enable the analyst to both refine the displayed data and via Resource Manager (RM) tasking elicit additional information in order to optimize confidence in decision making by feedback control. The analyst’s cognitive functions are modeled/emulated/aided by the Perceptual Reasoning Machine (PRM) paradigm [3-7], a meta-level information management system. Part of the PRM uses associative recall, which is implementable via Case-Based Reasoning [8] is described. The issues and challenges in the BDA and Information fusion system chain coupled with the PRM paradigm imbedded within the associated information Process Module System (PMS) [3] are addressed. The associated interaction between PRM (which also serves as model for high level fusion processing [5]) and the fusion levels is illustrated. Applications include Cognitive Intent Modeling (JDL Levels 2/3) [7] from Social Networks (SNs) enormous “tweets” data feeds is addressed. Another application of the PRM/PMS is to address the challenges of the new UK project, entitled “Complex Adaptive Systems and Networks” [9].

## 2. PROBLEM SETTING AND CHALLENGES

Common challenges and issues BDA and Information Fusion include: Applications independent modeling the human perceptual reasoning/cognitive functions and making the “analyst” end-user the “manager” thereby creating an adaptive and automated system with: sensing, learning, prediction/anticipation, hypothesis management, control/feedback capabilities yielding near real-time accurate information via cognitive decision making based on visual analysis processing; Big Data conversion/preprocessing for interactive display visualization and human - display interaction. In the processing chain, the slide entitled “Big Data and Fusion Processing” depicts the high level processing functions used for complex high volume heterogeneous “Big Data” information sources. A form of “preprocessing” is used to extract/learn relevant information, consisting of Data Mining/Machine Learning, followed by Visual Analytics - Analytical Reasoning [8] chain providing analyst/display interaction capability. At the display level the data is coupled with data association/fusion processed data for interaction/feedback from the user/analyst at high fusion levels via the RM. The preprocessed data is also fed, as needed, to the RM from fusion Levels (0 and 1).

## 3. CHALLENGES IN VISUAL ANALYTICS PROCESSING

It turns out that Georgia Institute of Technology is the lead research team on the project called “FODAVA” (Foundations of Data Analysis and Visualization) [8] including machine learning and computational statistics, information visualization, massive-dataset algorithms and data structures, and optimization theory. That is, it appears that elements of FODAVA could provide components needed to implement the adaptive cognitive processing system and address a part of the challenges. As stated before, it is the visualization process coupled with fusion processing, where cognitive modeling of the human’s role becomes important. The cognitive processing is referred to “Analytical

Reasoning,” [8] which is a component of “Visual Analytics” [2, 8] representing data analysis using visualization techniques. Analytical Reasoning is dependent in part on modeling the human perceptual reasoning/cognitive process. This is exactly where the adaptive PRM paradigm fits in providing needed processing elements. Tasks in BDA Processing Chain in Visualization Processing, subsequent to data mining/machine learning are: “Visual Analytics” [2, 8] providing processing steps necessary for data analysis using visualization techniques, which are detailed in accompanying slide, include: (1) Computational Data Analysis; (2) Interactive Visual Interfaces; and (3) *Analytical Reasoning*. Methods of *Analytical Reasoning* to facilitate the cognitive function include: (a) “Bayesian Analysis in Visual Analytics (BAVA)” - probabilistic data mapping [8a]; and “Multi-Source Visual Analytics” - dimensionality reduction [8b]. Components of BDA processing involving the human’s role are shown: (1) *Predictive Analytics – Data Mining*; (2) *Visual Analytics* component of the HCI interface – “Analytical Reasoning” – the “*cognitive model*” (modeling how people think: learn, anticipate, plan, predict, fuse information, generate hypotheses, manage resources and reach decisions); and Case-Based Reasoning, which uses a library of past experiences to associate a new problem with similar learned problems to solve it (*alike to associative recall by humans, used in PRM*).

#### 4. COGNITIVE MODELS OF INTENT (CMI) – ISSUES AND CHALLENGES

As described in the accompanying slides, the capability to “sense/observe, mine/access” data, associate, learn, recall, anticipate and predict/act” are key ingredients of *human perceptual reasoning*. These attributes are necessary constructs in *cognitive modeling*. *The key ingredient is timely information access in modeling intent*. *Cognitive models* are imbedded in large family of methods called *Predictive Analytics/Modeling* [10] (techniques to predict future entities) including: sensing/collecting, data mining, sorting, organizing, aligning, associating, fusing,; and using a-priori and learned, SME based and current data. *Predictive models have been used in intent modeling, but did not use a cognition framework, which includes many well known algorithms (please see viewgraphs for additional details)* including the *cognitive PRM paradigm*: the cognitive *Perceptual Reasoning Machine (PRM)* [3-7]: a “meta-level information management system”, for adaptive information gathering/assessment, learning, anticipation, and *prediction*. Objectives of models are to minimize uncertainty and maximize the value of deduced information to *identify potential intent*, and to act in a real-time environment with time constraints - (by emulating/modeling/aiding analyst by PRM) [3-7]. As a matter of fact, the godfather of the Internet and knowledge representation, Vannevar Bush [11] in his famous 1945 essay, “As We May Think” stated, op. cit., “The human mind does not work that way hierarchically. It operates by association.” Spatial and temporal associations are key ingredients of PRM.

#### 5. CMI IN SOCIAL NETWORKING OVER ENORMOUS DATA

Social Networks (SNs) provide basis for information exchange, in a social-cultural setting, allowing exchange and expression of information, and enabling extraction of context, such as: ideas, concerns, sentiments, emotions, and opinions. What is the Role of Social Networking in the PRM intent model framework? Social Networks (SNs) provide access to information exchange [derived\* context (e.g., sentiments, emotions), extracted from cultural/social interactions - messages with location and time stamped data] to be used as input to the model [7] as shown in the slides. Potential Issues and Challenges: (1) Is the extracted data based on consensus of the population or only from “outliers”? (“Outliers” can exert *influence* [12], coalesce and become significant intent indicators). Furthermore, how to handle potential data sparsity (per individual) vs. enormity (web) of data; and contextual validity into emotional aspects? (2) Is information exchange restricted globally by particular entities? - (represents potential intent); (3) How to “*associate*” massive information from multiple SNs as input to PRM? Information access is crucial as an input both for real-time assessment, prediction and to data bases (learning) & for *message rate “change detection”* - impending intent? An example: Ben Zimmer, “Twitterology: A New Science?”, *The New York Times*, October 30, 2011 [13], the article illustrates the degree of relevant real-time information that can be derived from social/cultural interactions expressed in Twitter. That is, Twitter extracted information from tweets can be used as input to intent modeling: - such as sentiments, emotions, moods, opinions [14], etc., (including locations, time, consensus types, groups and number of constituting elements or computed probabilities) as input to assess potential intent via the cognitive PRM model and associated PMS. (\* Note above: The preprocessing of linguistic messages to learn, classify and group various context is assumed a given herein).

## 6. THE PERCEPTUAL REASONING MACHINE PARADIGM

Viewed as a “meta-level information management system”, PRM consists of a feedback planning/resource management system whose interacting elements are: “assess”, “anticipate” and “preplan/act” [3-7]. That is:

- *Gather/Assess* current, *Anticipate* future (hypotheses by learning), and *Preplan/Act* (predict) on information requirements as well as likely intent and threats,
- *Anticipate/Predict (Plan)* the allocation of information/sensor/system resources and acquisition of data through the control of a specific distributed multisource sensors/systems resource manager (RM),
- *Interpret and Act* (shared by above functions) on acquired (sensor, spatial and contextual) data in light of the overall situation by interpreting conflicting/misleading information to either identify or rule out the potential or existence of intent.

The elements of the fundamental PRM construct are shown in the slide “Perceptual Reasoning Machine” depicting the interrelations among the constituting elements described above, providing adaptive information gathering (e.g., fusion) learning, anticipation, assessment, prediction and control. The PRM information flow management elements and their relationships are depicted in the viewgraph entitled “Fusion Levels Interactions via PRM Mapping” along with the knowledge requirements for each PRM function. Note both the JDL and the DFIG (human user perspective) [15] fusion models are shown. It should be noted that the *current information* noted can be derived from processed information collection which is can be controlled by a systems/sensors resource manager by feedback from the PRM [3-7].

This function is illustrated in Figure 1, depicting the Information “Process Model System” PMS application of the PRM to Big Data Processing. Process modeling is defined as a set of procedures and algorithms that capture the functional and required (temporal and spatial) dependency relationships of tasks (e. g., needed for intent/threat assessment) and/or processes, which are being modeled. Referring to the PRM slide, or within Figure 1, the “assess module,” responding to dynamically managed and received multisource information, uses additional information from its associated knowledge base and from the “anticipate module” to form a database of “likely current situations” which include potential intents/threats. The “anticipate module” provides information on “likely future situations” that are used for short- and long-duration planning. This planning is based on the “likely current situations” from the “assess” module; prior, learned, process and tactical/planned knowledge and associated hypotheses. The “likely current situation” information is fed back to the “predict module”, which provides “plans of actions and goals”.The “assess module” also provides current situations information to the “predict module” which, along with its knowledge base and likely future situations information from the “anticipate module”, (based in part on associated process knowledge), issues assessments, identifies potential intents/threats, and as needed, request actions from the resource manager for additional information to confirm or negate conflicting hypotheses thus closing the outer loop via the systems/sensors manager.

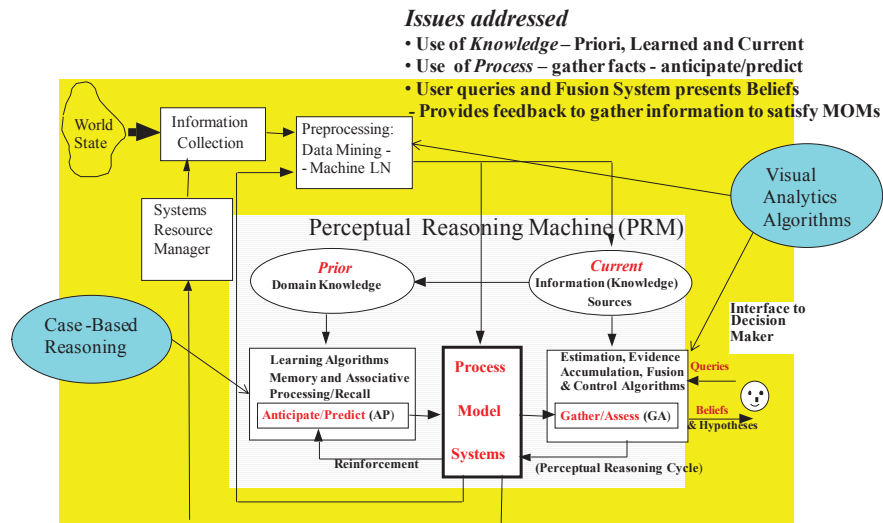


Figure1- Information “Process Model System” Application to Big Data Processing

The application of PRM/PMS to Complex Adaptive Networks and Systems (CANS) [9], shown in the CANS slide, is identical to the processing depicted above for CANS’ “advanced modeling, analysis, prediction and visualization.”



## SUMMARY

The focus and the challenge herein was to model an adaptive big data processing and information fusion system emulating the human perceptual reasoning/cognitive functions to facilitate two-way interactions between the visual data display and the analysts. Methods were presented with proposed solutions, issues and challenges using a combination of Data Mining, Visual Analytics and Perceptual Reasoning/Cognitive processing algorithms. The proposed systems are to enable the analyst to both refine the displayed data and via Resource Manager (RM) tasking elicit additional information in order to optimize confidence in decision making by feedback control. The analyst's cognitive functions were modeled/emulated/aided by the Perceptual Reasoning Machine (PRM) paradigm, a meta-level information management system. Part of the PRM uses associative recall, which was shown implementable via Case-Based Reasoning. The issues and challenges in the BDA and Information fusion system chain coupled with the PRM paradigm imbedded within the associated information Process Module System (PMS) were addressed. The associated interaction between PRM (which also serves as model for high level fusion processing) and the fusion levels was illustrated. Applications include Cognitive Intent Modeling (JDL Levels 2/3) from Social Networks (SNs) enormous "tweets" data feeds was addressed. Another application of the PRM/PMS was illustrated to address the challenges of the new project, entitled "Complex Adaptive Systems and Networks". There are many issues and challenges remaining requiring research, implementation and testing of the proposed methods.

## REFERENCES

- [1] [http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data)
- [2] [http://en.wikipedia.org/wiki/Visual\\_analytics](http://en.wikipedia.org/wiki/Visual_analytics)
- [3] I. Kadar, "Perceptual Reasoning in Adaptive Fusion Processing" *Proceedings of the Signal Processing, Sensor Fusion and Target Recognition Conference XI*, Ivan Kadar, Editor, Proc. SPIE Vol. 4729, Orlando, FL 2002.
- [4] E. P. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, G. M. Powell, D. D. Corkill, E. H. Ruspini, "Issues and Challenges in Situation Assessment (Level 2 Fusion)", *Journal on Advances in Information Fusion (JAIF)*, Vol. 1, No. 2, December 2006
- [5] I. Kadar, "Results from Levels 2/3 Fusion Implementations: Issues, Challenges, Retrospectives and Perspectives for the Future - An Annotated View", Invited Panel Session on "Results from Levels 2/3 Fusion Implementations: Issues, Challenges, Retrospectives and Perspectives for the Future", Organizer: Ivan Kadar; Moderators: Ivan Kadar and John Salerno, *Proceedings of the 10<sup>th</sup> International Conference on Information Fusion*, 9-12 July 2007, Quebec City, Canada.
- [6] J. Salerno, S. J. Yang, I. Kadar, M. Sudit, G. P. Tadda, J. Holsopple, "Issues and Challenges in Higher Level Fusion: Threat/Impact Assessment and Intent Modeling (A Panel Summary)" *13<sup>th</sup> International Conference on Information Fusion*, Edinburgh, Scotland, 26-29 July 2010.
- [7] I. Kadar, "Issues and Challenges in Social/Cultural Modeling with Applications to Information Fusion", Presented at Invited Panel Discussion, Co-Organizers : John Salerno and Ivan Kadar, *Signal Processing, Sensor Fusion and Target Recognition XXI*, Ivan Kadar, Editor, Proc. SPIE Vol. 8392 Baltimore MD., April 2012.
- [8] Georgia Tech's "Foundations of Data and Visual Analytics (FODAVA)" Program Outline in "*Horizons, Vol. 30, No.1, Fall 2012 - Winter 2013*" and Research Topics [8a, 8b] [www.fodava.gatech.edu](http://www.fodava.gatech.edu)
- [9] <https://sites.google.com/site/cnskcl/home>
- [10] [http://en.wikipedia.org/wiki/Predictive\\_analytics](http://en.wikipedia.org/wiki/Predictive_analytics)
- [11] Vannevar Bush, "As We May Think", *The Atlantic Monthly*, July 1945
- [12] W. Pan, W. Dong, M. Cebrian, T. Kim, J. H. Fowler and A. (Sandy) Pentland, "Modeling Dynamical Influence in Human Interaction", *IEEE Signal Processing Magazine*, March 2012.
- [13] B. Zimmer, "Twitterology: A New Science?", *The New York Times*, October 30, 2011
- [14] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta
- [15] Blasch, E., Deignan, Jr, P. B., Dockstader, S. L., Pellechia, M., Palaniappan, K., and Seetharaman, G., "Contemporary Concerns in Geographical/Geospatial Information Systems (GIS) Processing," *Proc. IEEE Nat. Aerospace Electronics Conf.*, (2011).

