SAND2016-5170D

Multimodal Data Integration Under Uncertainty

Sandia National Laboratories
D.J. Stracuzzi, M.G. Peterson, M.G. Chen, S.M. Dauphin, and R. Riley
Sandia National Laboratories, New Mexico 87185

Problem

Background:
- Platforms for sensor data collection are becoming increasingly ubiquitous.
- People use data for decision making in many mission domains including defense, intelligence, and security.
- Accounting for uncertainty plays a critical role in high-consequence decision making.

Problem: Existing data integration methods often ignore uncertainty, leaving important information locked in the data.

Goal: Improve information extraction in multi-source settings by analyzing uncertainty in data.

Questions:
1. *Generality:* Reference distributions differ from sensor to sensor. What interlingua will support integration of uncertainty across data sources?
2. *Efficiency:* What is the minimum information required to estimate uncertainty distributions for a single source?
3. *Propagation:* No sensor is definitive; how combine and propagate information through layers of analysis?
4. *Value of Information:* How does uncertainty in analytic outputs relate to decision-making value? When should we seek more data and which data will help most?

Approach

1. Preprocessing
a. Co-register samples from each source by geolocation.
b. Calculate local gradients from lidar.
c. Decompose SAR into surface, dihedral, volumetric, and helical responses.

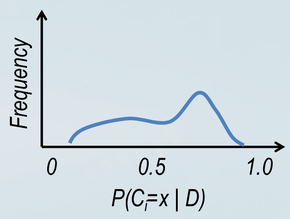
2. Data Analysis
a. Separate *n* pixels into *k* clusters via Gaussian mixture model (or similar)

$$\mathcal{L}_{MIX}(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i|\theta_k)$$

b. Perform hierarchical agglomeration to obtain classifications for labels *l*

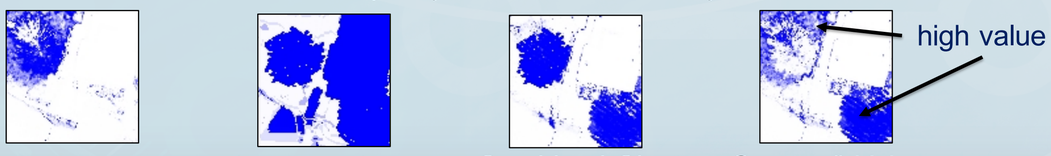
$$\max \mathcal{L}_{CL}(\boldsymbol{\theta}, \mathbf{l}|\mathbf{y}) = \prod_{i=1}^{n} f_{l_i}(\mathbf{y}_i|\theta_{l_i})$$

c. Apply EM to estimate model parameters
d. Use Bayesian Information Criterion to select model.
e. Bootstrap steps a. – d. to establish non-parametric uncertainty distributions.



3. Integration (forthcoming)
a. Use supervised data to establish conditional probability relationships among unsupervised classes and semantic labels.
b. Marginalize across sensors to integrate multiple sources.

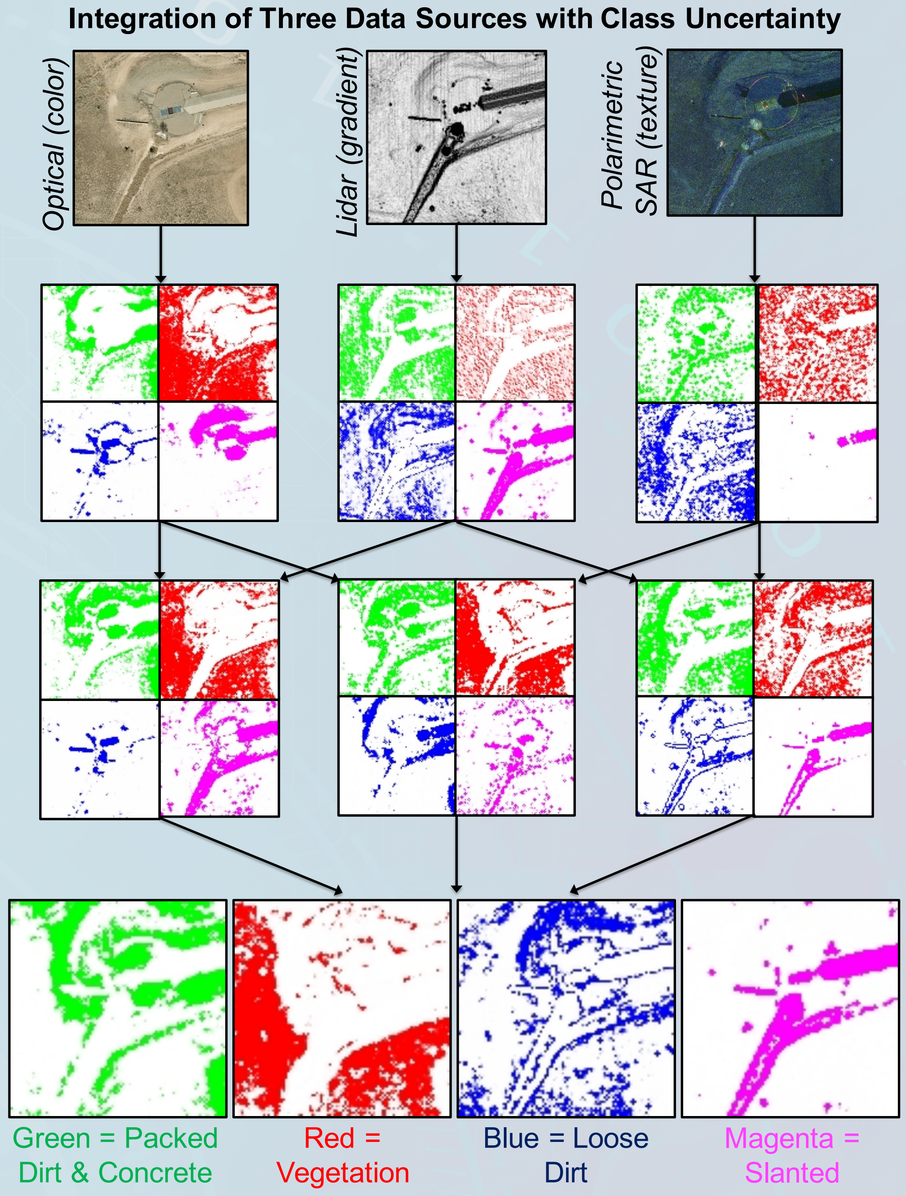
4. Value of Information (forthcoming)
- Defined in terms of data's capacity to reduce uncertainty and alter labels.



Source 1 Class *Source 2 Class* *Combined Class* *Source 2 Value*

Preliminary Results

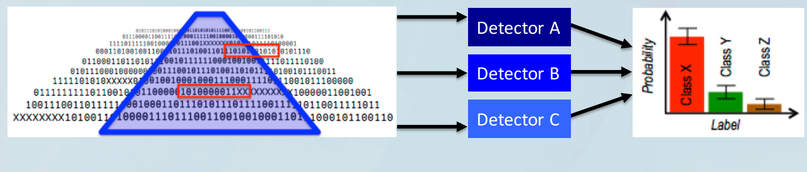
Integration of Three Data Sources with Class Uncertainty



- Color indicates class label, intensity indicates probability
- Classes are unsupervised; Semantic labels applied post hoc
- Semantic associations apply to bottom row only

Significance

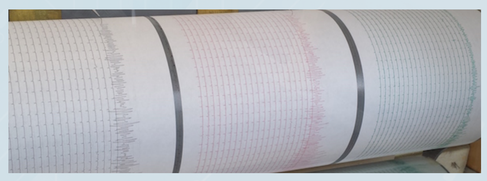
- **Shift in Analytic Philosophy**
  - From data-driven to decision-driven processing
- **Broad Mission Application –** converting many weak indicators into one strong indicator
  - Cyber Security



  - Remote Sensing

	Position Uncertainty	Characterization Uncertainty	Temporal Resolution	Geospatial Resolution
Sensor A	Small	Detailed Classifiers	Low	High
Sensor B	Medium	Change classifiers	Medium	Medium
Sensor C	Large	Broad classifiers	High	Low

  - Seismic Monitoring



- **Provides basis for resource tasking under uncertainty**
  - Relax assumption that desired information is obtained from a collect.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND 2016-xxxxP

Sandia National Laboratories