SAND2005-7243C

# Ensemble Classification of Disparate Data Sets

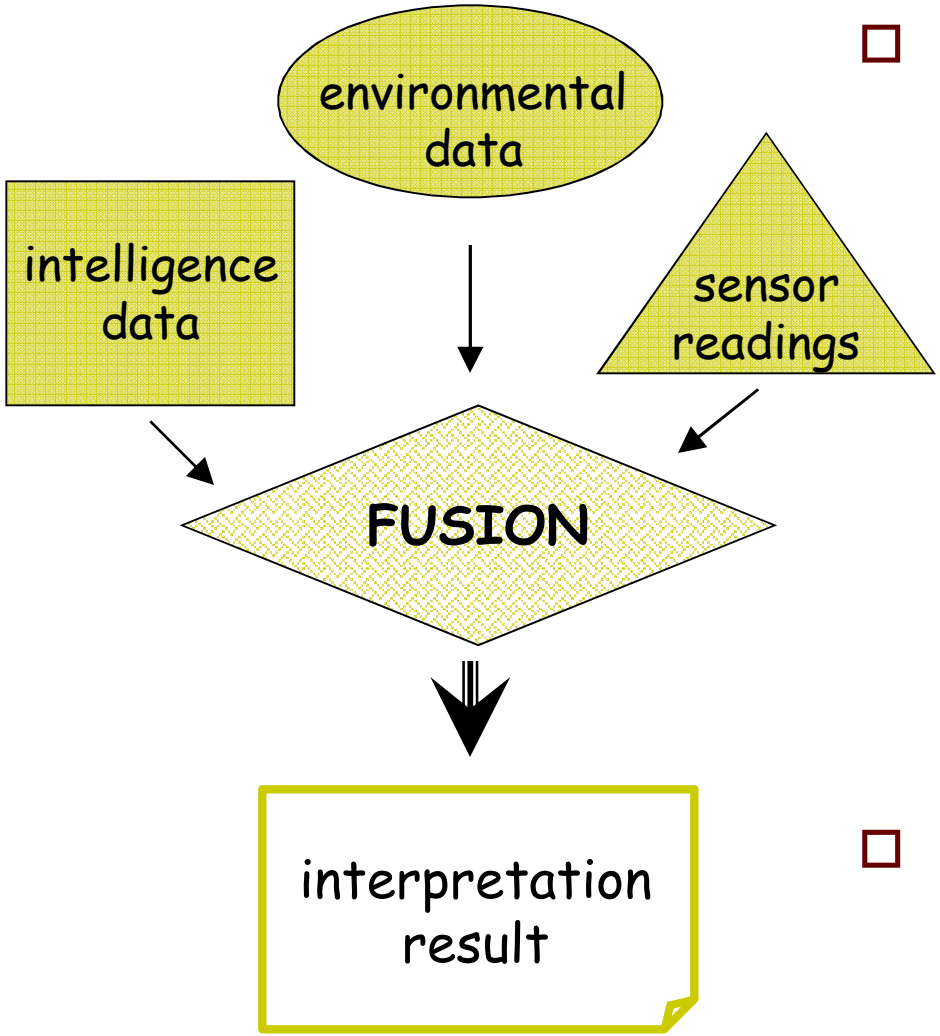
Genetha Gray
Computational Sciences & Mathematics Research
Sandia National Labs, Livermore, CA
gagray@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

# Introduction

- Various types of data can provide different views of the same situation.
  - Multiple camera angles
  - Different experimental protocols
  - Human influence on data collection
- Seemingly dissimilar data can provide complimentary information.
- Data is collected in a wide variety of formats.

# Problem Overview




- Developing detection and assessment systems includes:
  1. Design and deployment of sensors
  2. Methods for simultaneous consideration of sensor data and related info
  3. Techniques for interpreting large data sets generated by sensors
- **Goal**: Develop an algorithm for real-time, interpretive data fusion

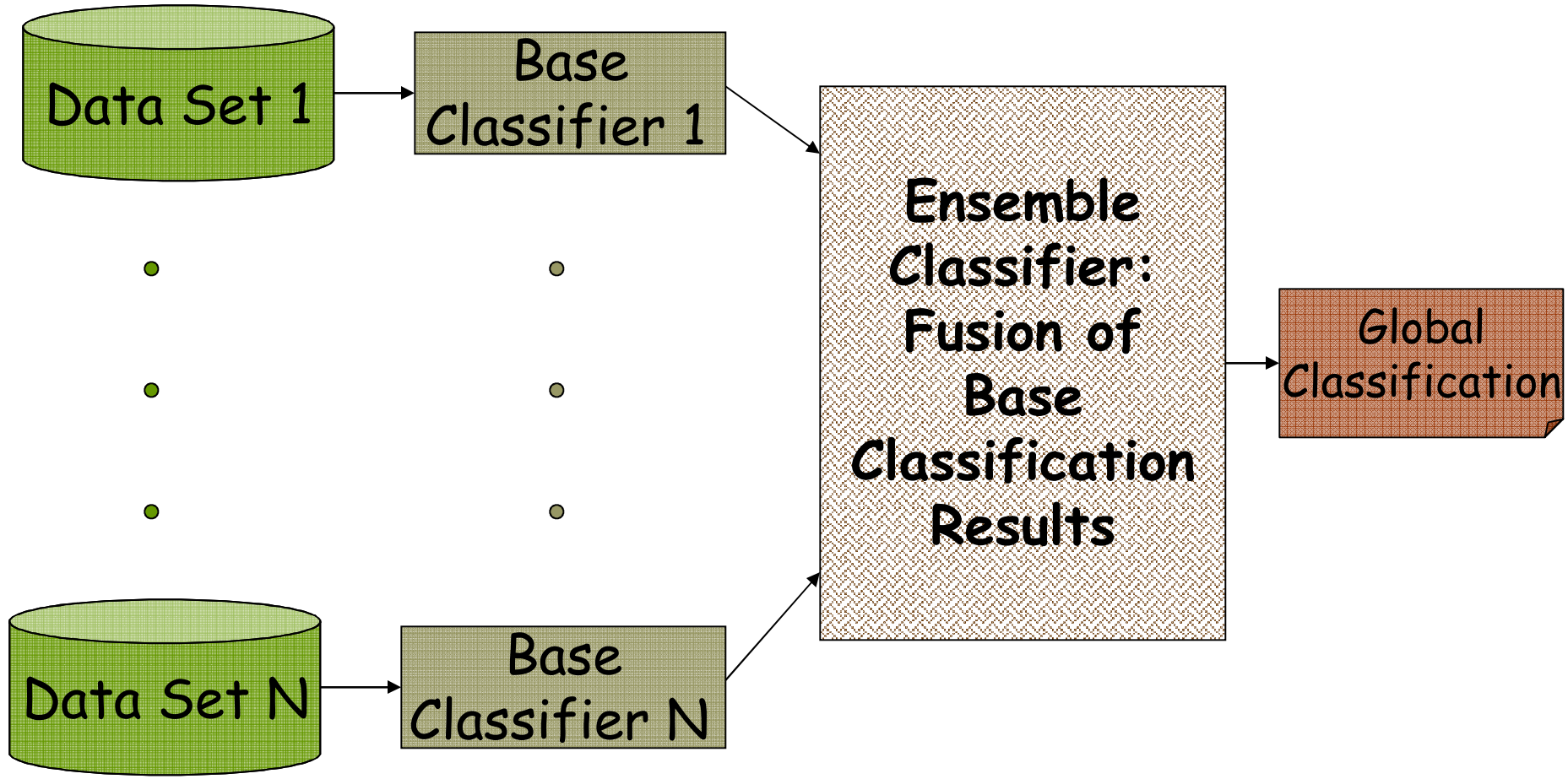
# Ensemble Classification

- Technique for combining the predictions of multiple classifiers into a single classification.
- Typically more accurate than any of the individual classifiers.
- Extracts information from individual data sets and combines the results.
- Inherent parallelism.
- Typically apply different classifiers to the same data set. (Example: bagging, Breiman ‘96)
- Want to extend the idea to disparate data sources.

# Ensemble Classification of Disparate Data

- **Advantages**:
  - Data can exist in separate data bases.
  - Does not require translation of data formats.
  - Saves time and computational resources.
  - Allows for automated fusion and may not require human interaction.
- **Areas of Research**:
  - Need appropriate algorithms to oversee process
  - Development of a provable method of solution
  - Investigation into the applicability of optimization techniques

# Algorithmic Framework



# Framework Components

- Data bases need not be independent.
- Provide a variety of base classifiers; allow users to select from these or provide their own.
- Include more that one fusion algorithm.
- Fusion may occur at more than one level.

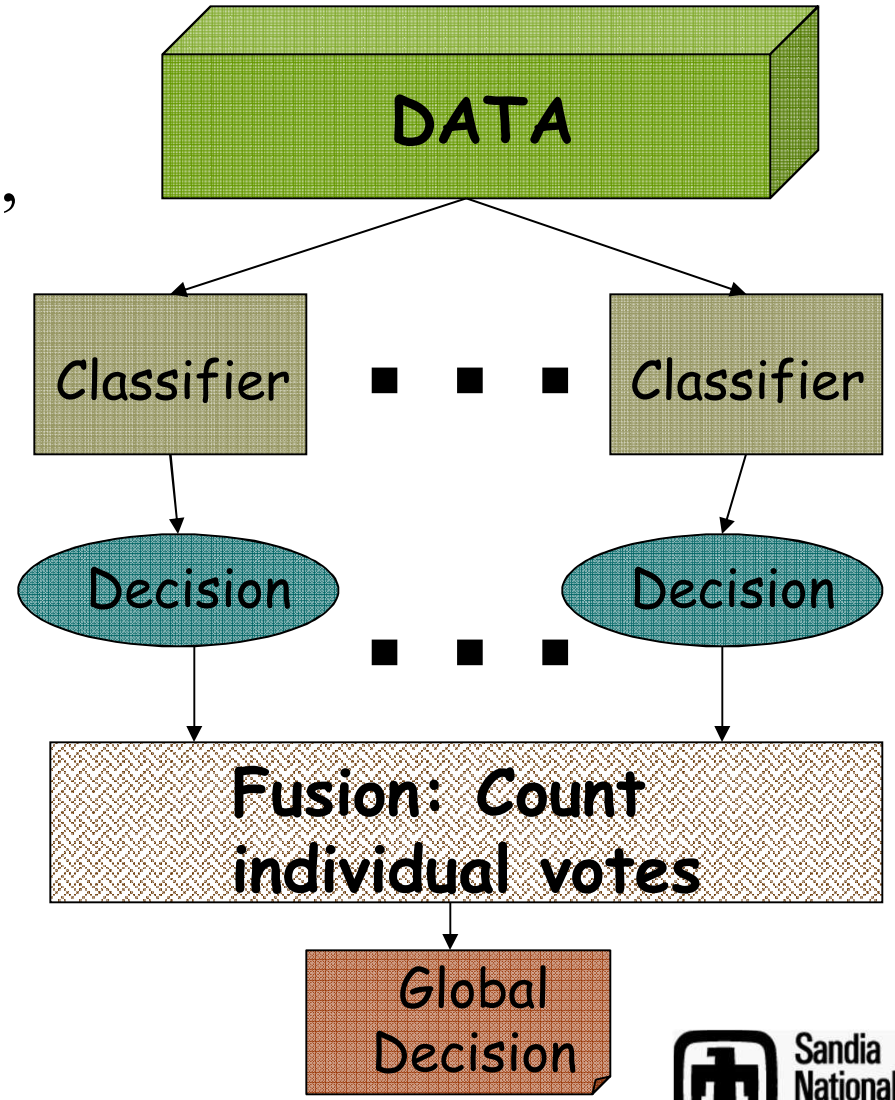
# Traditional Unweighted Voting: One Data Set

- Given *n* data classifications, a decision is accepted if at least *k* of the classifications agree where:

  $k = n/2 + 1$, if $n$ is even

  $k = (n+1)/2$, if $n$ is odd

- Commonly used for comparison

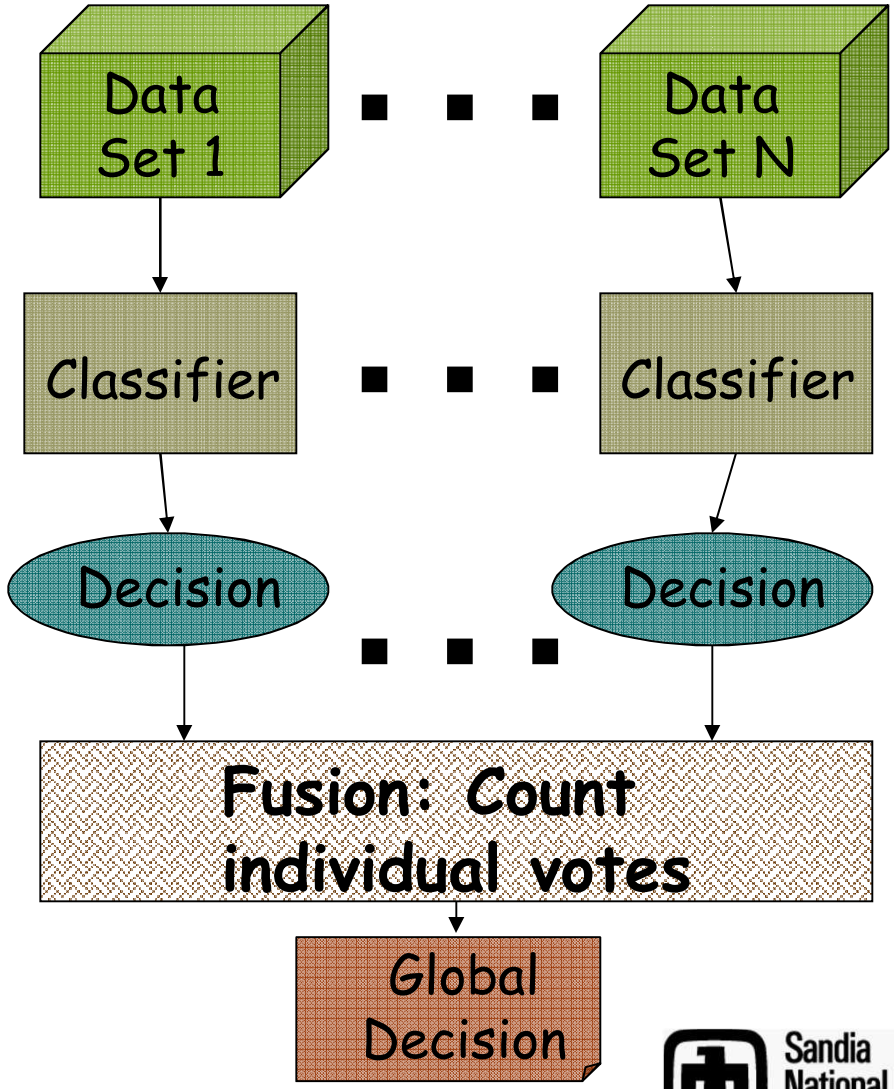


# Unweighted Voting: Disparate Data Sets

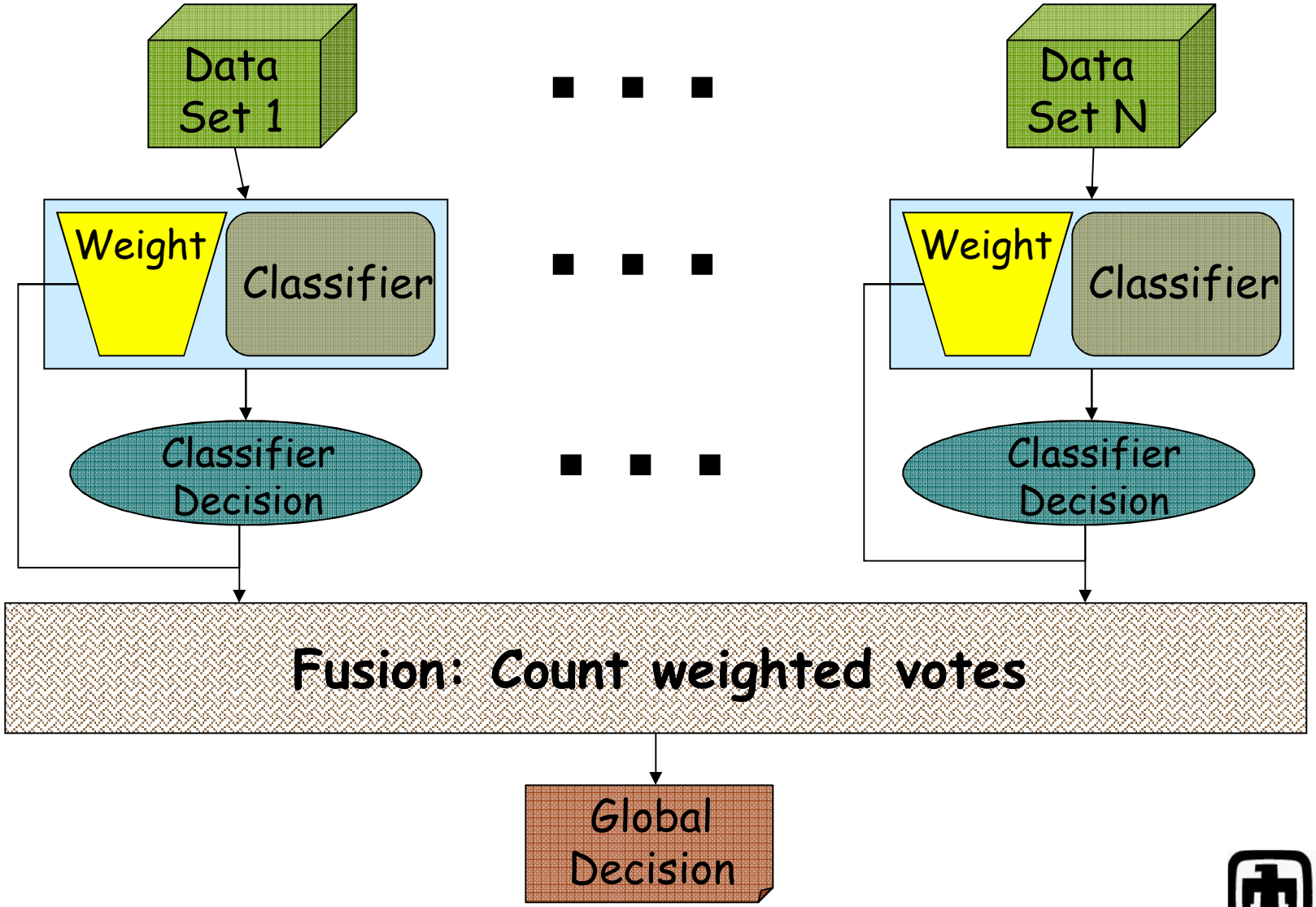
- Given *n* data classifications of *n* individual data sets, a decision is accepted if at least *k* of the classifications agree where:

  $k = n/2 + 1$, if $n$ is even

  $k = (n+1)/2$, if $n$ is odd



# Weighted Voting: Disparate Data Sets



# Determining the Weights

- Incorporate other known information
  - Classification error
  - Relative importance of data
  - Data characteristics- size, number of features, expected error, etc.
- One approach: weights are the solution of an optimization problem

# Optimization: Determine the Weights

$$\min \sum_{i=1}^{m} \left( \sum_{j=1}^{n} E_{ij} x_j \right)^2$$

$$\text{s.t.} \sum_{j=1}^{n} x_j = 1, \quad x_j \geq 0 \; \forall j$$

$n =$ number of classifiers

$m =$ number of observations

$E = m \times n$ matrix where $E_{ij}$ descibes classifier $j$ in terms of observation $i$

# Test 1: Optimized Weights

- Test: SVM light example 1 (Joachims)
- Are Reuters news articles about corporate acquisitions?
- 9947 attributes
- 2 data sets: training (2000 observations) and testing (600 observations)
- Break training data into three subsets
  - Set 1: Eliminate first half of the attributes
  - Set 2: Eliminate “middle” half of the attributes
  - Set 3: Eliminate last half of the attributes

# Test 1 Results

| Classifier | Weights | | | Training Errors (2K) | Errors (600) |
|---|---|---|---|---|---|
| | x1 | x2 | x3 | | |
| Based on set 1 | 1 | 0 | 0 | 11 | 16 * |
| Based on set 2 | 0 | 1 | 0 | 205 | 182 |
| Based on set 3 | 0 | 0 | 1 | 510 | 282 |
| Fusion (E = 0,1) | 0.95 | 0.03 | 0.02 | ------ | 16 * |
| Fusion (E = svm dist.) | 0.76 | 0.02 | 0.22 | ------ | 4 |

* Each of these classifiers made one error that the other did not

# Test 2: Weighted vs. Unweighted

- Data: from the UC Irvine repository
  - BC: breast cancer (Wisconsin)
    - 699 observations, 10 attributes
    - split into 3 overlapping attribute group
  - ION: ionosphere
    - 351 observations, 34 attributes
    - Split into 6 partially overlapping groups
  - SNR: sonar
    - 208 observations, 60 attributes
    - Split into 5 overlapping attribute sets
- Fusion Schemes
  - UW: Unweighted voting
  - W1: Weighting with E = ± 1
  - W2: Weighting with E computed using svm distances

# Test 2 Results

| Fusion Method | Training Data Errors | | | Test Data Errors |
|---|---|---|---|---|
| | BC (699) | ION (200) | SNR (208) | ION (151) |
| UW | 26 | 35 | 44 | 18 |
| W1 | 26 | 34 | 44 | 17 |
| W2 | 22 | 32 | 41 | 14 |

# Test 3: Attribute Splitting

- Pima-indian-diabetes data from UC Irvine repository
  - 768 observations, 8 attributes
  - Split into 3 classifiers based on groups of attributes:
    - CASE 1:
      - G1: 1,2,3,4,5
      - G2: 3,4,5,6,7
      - G3: 5,6,7,8
    - CASE 2:
      - G1: 1,3,5,7
      - G2: 2,4,6,8
      - G3: 1,2,7,8
- Two data splits
  - All 768 for training and testing
  - 500 for training, 268 for testing

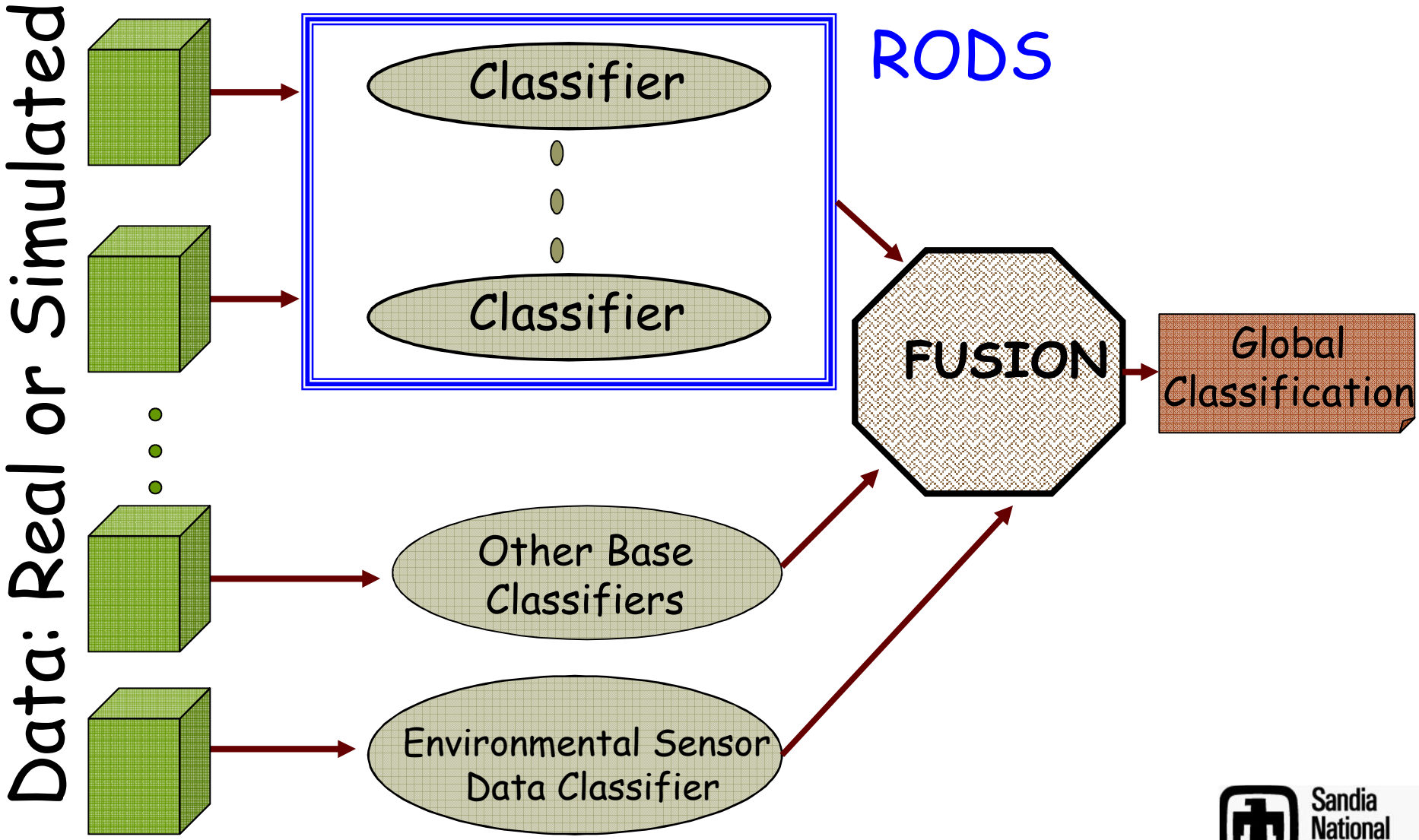
# Test 3 Results

| Fusion Method | Case 1 Train (768) | Case 2 Train (768) | Case 1 | | Case 2 | |
|---|---|---|---|---|---|---|
| | | | Train (500) | Test (268) | Train (500) | Test (268) |
| UW | 266 | 188 | 182 | 86 | 133 | 54 |
| W | 210 | 183 | 137 | 65 | 133 | 58 |

# Conclusions

- Data fusion techniques can improve disparate data classification.
- Fusion techniques must be designed smartly to improve upon best base classification result.
- Weighting can improve classification if the weights are chosen smartly.
- Training is an important consideration.
- Fusion methods should also consider which data types would benefit most from fusion.

# Current Test Problem



# **RODS**: Real-time Outbreak & Disease Surveillance

- Open source set of Java software modules for building public health surveillance systems
- Developed at Pitt & CMU
- Includes classifiers for clinical encounters at hospitals and OTC drug sales
- Additional classifiers will be needed to incorporate additional data types
- Adding new classifiers includes translation of results into input for the fusion algorithm
- Does not include automated fusion

# Current Test Problem: Data

- Real Data
  - OTC drug sales
  - Hospital emergency room visits
  - School absentee rates
  - Environmental sensor
- Simulated Data: Bio-DAC
  - Bio-agent attack decision analysis center
  - Funded at Sandia via the BioNet program (DHS & DTRA)
  - Running with simple models
  - Linked to RODS package

# Current and Future Work

- Investigate a variety of base classifiers
  - Recursive Least Squares - RODS
  - Cumulative Sum (Moore et al.) - RODS
  - Wavelet-based Anomaly Detection (Rizzo et al.) - RODS
  - Support Vector Machines (Boser, Gunyon, Vapnik)
  - Others
- Study and test fusion methods
- Research weighting techniques
- Find appropriate test problems

# Acknowledgements


- Ken Sale
- Pam Williams
- Dave Gay
- Keith Vanderveen
- George Davidson
- Nina Berry
- Funding: Sandia computational & information sciences lab directed research program



Genetha Gray
gagray@sandia.gov


Sandia National Laboratories