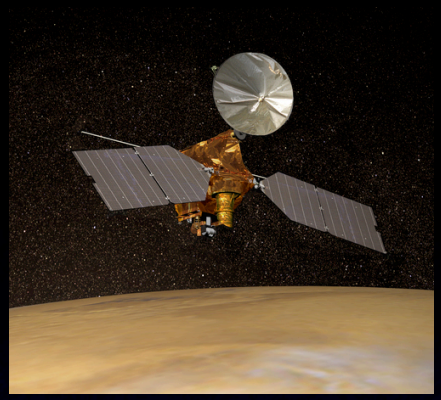


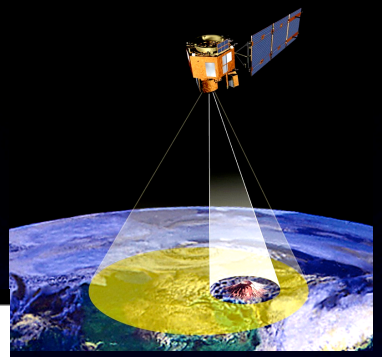
# K-means in Space: A Radiation Sensitivity Evaluation

Kiri L. Wagstaff and Benjamin Bornstein
Jet Propulsion Laboratory, California Institute of Technology

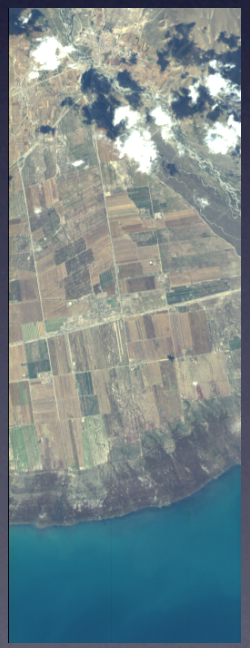
June 16, 2009

# Motivation

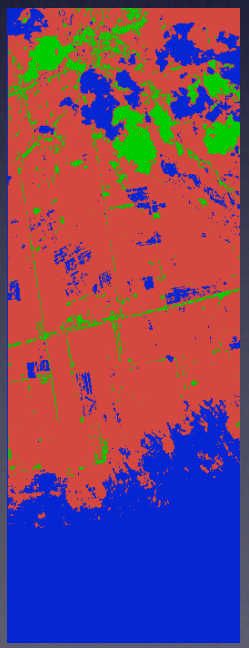
- Machine learning in space
  - SVMs on EO-1: classify hyperspectral pixels
  - Dust devil detection on Mars Exploration Rovers



- Space: a computing frontier
  - Limited CPU speed (20-133 MHz)
  - Limited memory size (128+ MB)
  - High-radiation environment



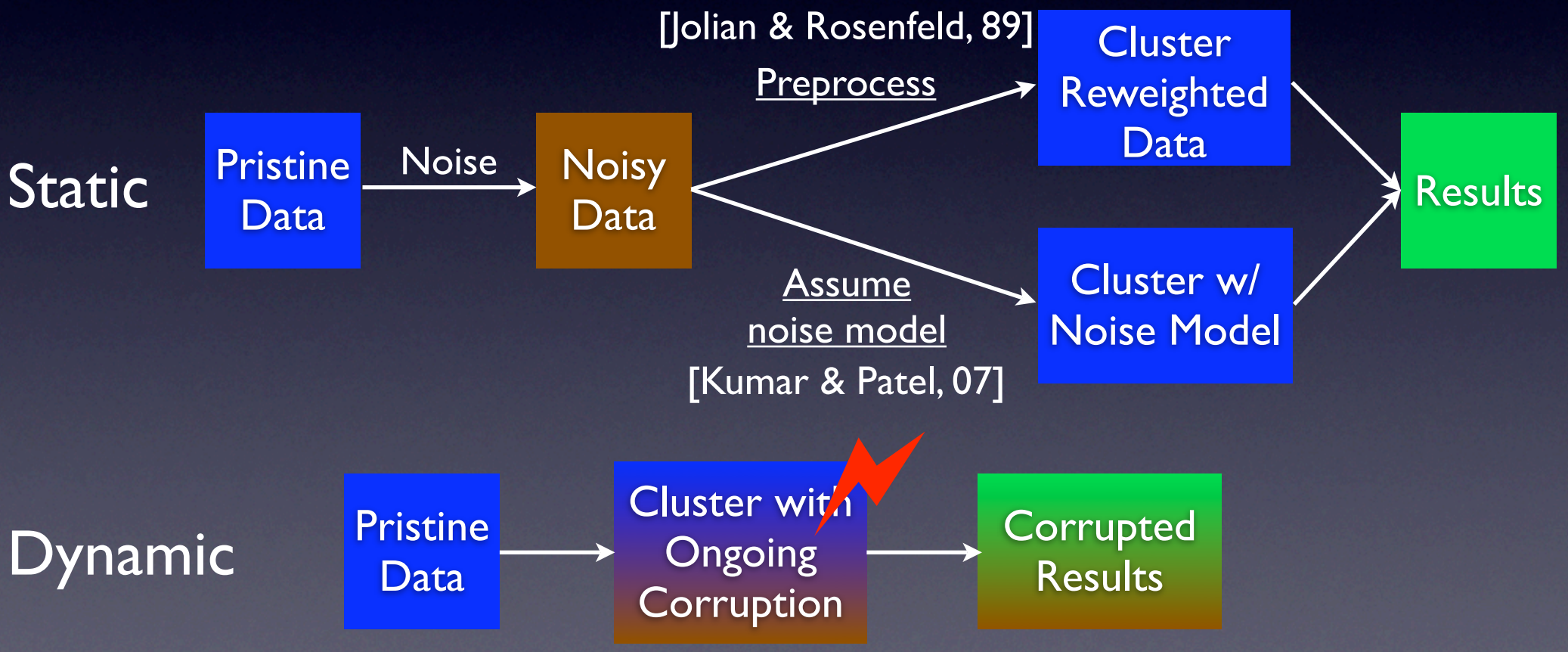
Hyperion (EO-1)



K-means (k=3)

# Radiation

- Data corruption *during* analysis
- Data structures also corrupted during analysis
- Static noise models (and methods) don't apply



# Goals

- This talk
  - Challenge assumption of incorruptible memory
  - Quantify the magnitude of the problem (how sensitive is a given algorithm?)
- Future work
  - Devise solutions to reduce sensitivity

# Radiation Simulation

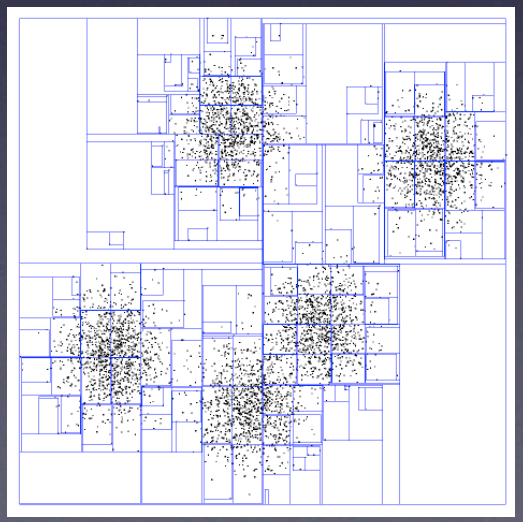
- SEU: Single-Event Upset (bit-flip)
- BITFLIPS: Basic Instrumentation Tool for Fault Localized Injection of Probabilistic SEUs
  - Extension to Valgrind debugging/profiling tool
  - Injects SEUs at user-specified rate (SEUs/kB/sec)
  - Can specify what memory to expose (input data, algorithm data structures)
- Report Adjusted Rand Index (pairwise agreement)

Increasing cost

| | SEUs/bit/day |
|---|---|
| Commercial | $10^{-5}$ |
| Rad-tolerant | $10^{-7}$ to $10^{-8}$ |
| Rad-hard | $10^{-10}$ to $10^{-12}$ |

# K-means Algorithms

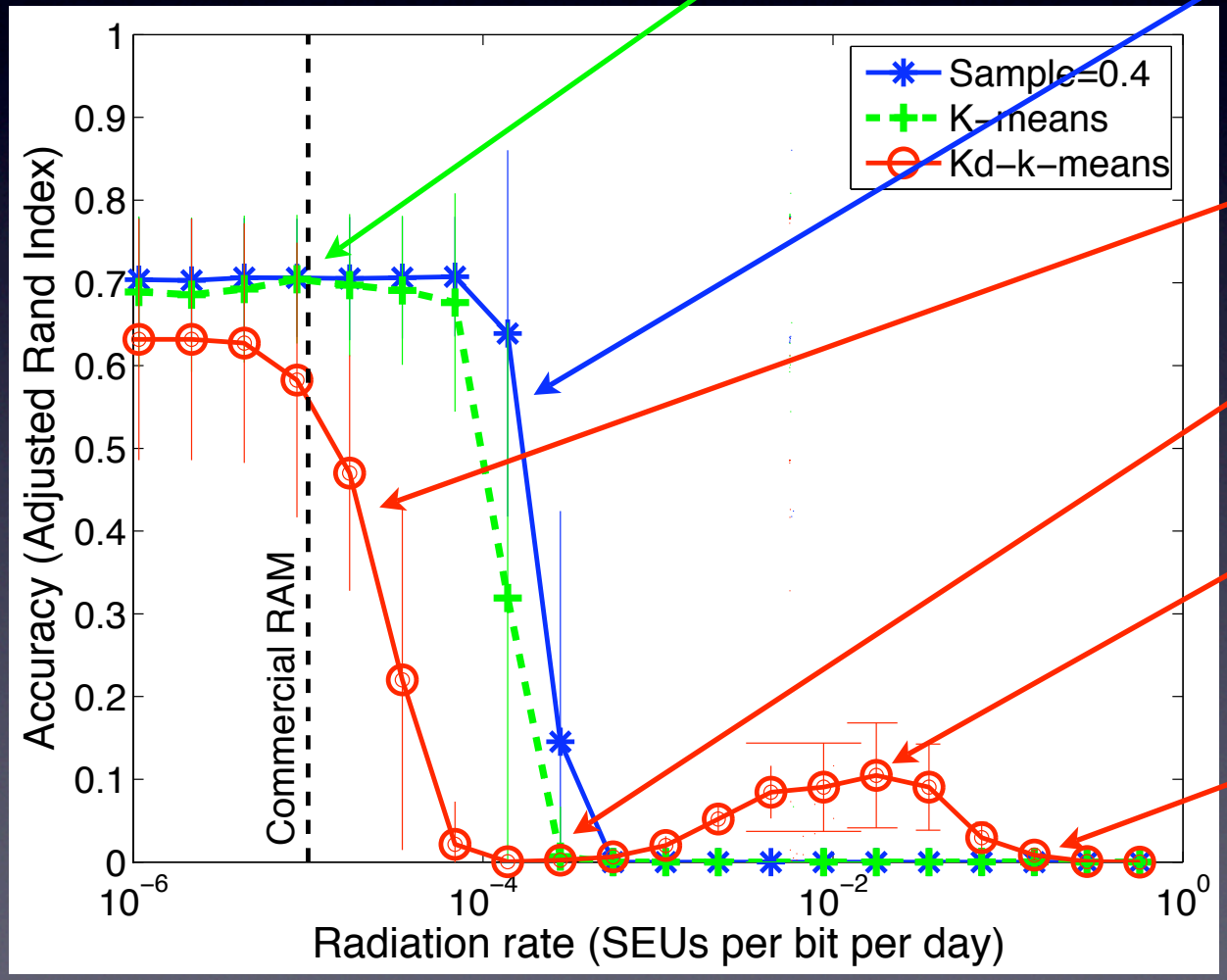
- K-means [MacQueen, 67]
  - Randomly initialize *k* clusters; assign all items to cluster 0
  - Until no assignments change:
    1. Assign each item to its closest cluster center
    2. Update each cluster to be the mean of its items
- K-means with subsampling [Bradley & Fayyad, 98]
  - C = clusters obtained from k-means with x% of the data
  - Run k-means with full data, initializing clusters to C
- Kd-k-means [Alsabti et al., 98; Kanugo et al., 99; Pelleg & Moore, 99]
  - Build kd-tree on data
  - Assign tree cells to clusters with recursive traversal



[Dan Pelleg]

# Results: Iris

Iris (UCI): 3 classes, 4 features, 150 items




SEUs tolerated

Subsampling helps a little

Kd-k-means hurts a lot

All items in same (valid) cluster

Items in valid and invalid clusters

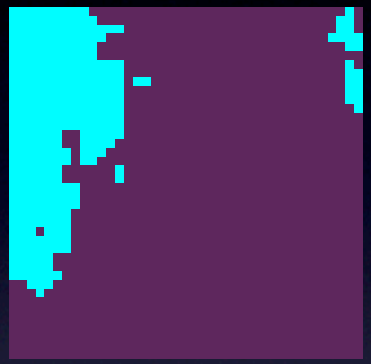
All items to invalid clusters

# Results: Satellite Data

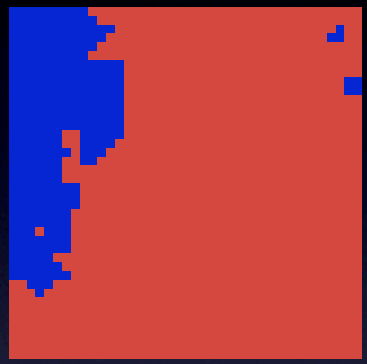
## China: 2 classes, 11 features, 1600 pixels

RGB

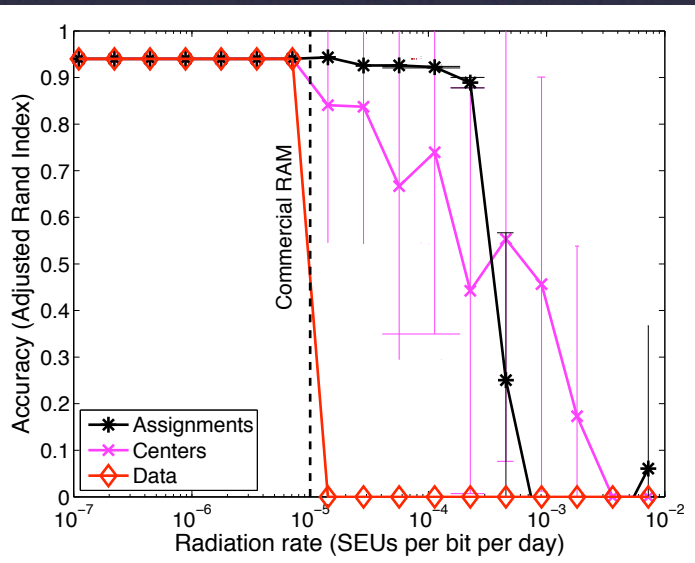
Labels



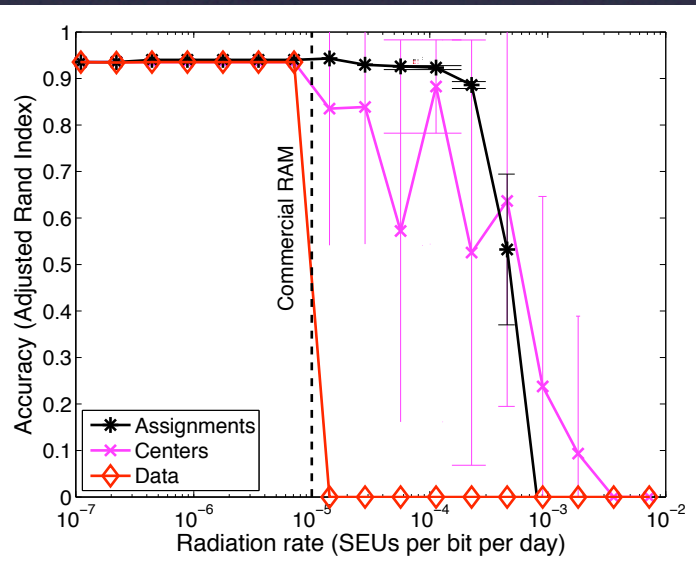
K-means



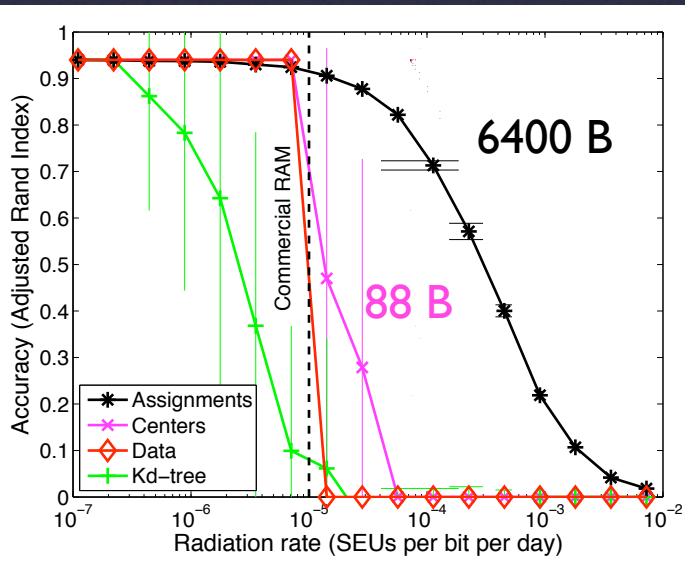
K-means




10% Subsampling




Kd-k-means



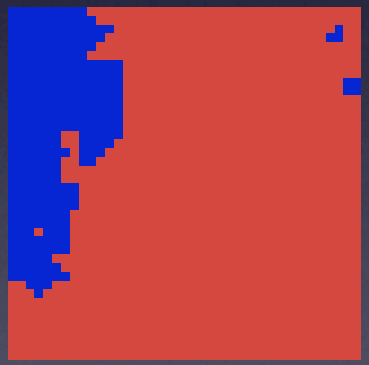
# Results: Satellite Data

- Rad-tolerant RAM needed
- Perturbed solutions of interest

K-means
with assignments exposed

RGB

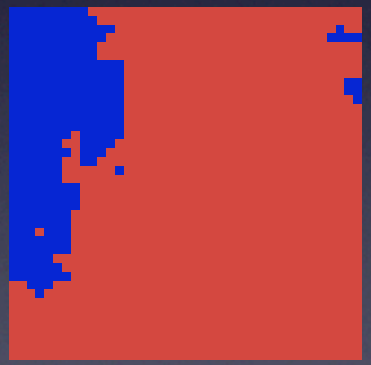
K-means



ARI = 0.940

Mild

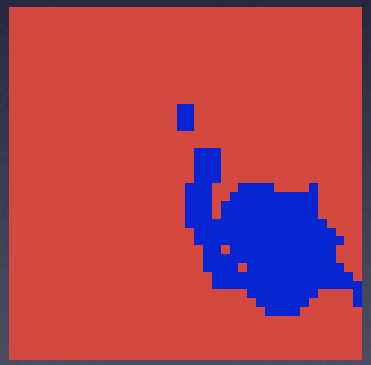
$1.3 \times 10^{-6}$



ARI = 0.946,
higher var.

Severe

$6.7 \times 10^{-4}$



ARI = -0.102,
extreme var.

# Conclusions and Future Work

- First quantitative ML radiation sensitivity results
- Subsampling helps k-means somewhat;
  kd-k-means is much more sensitive
- Data structures differ in sensitivity (function, not size)
- Sensitivity results can inform selection of RAM type
- Next steps:
  - How to decrease sensitivity?
  - Sensitivity of other ML algorithms, like SVMs
  - Application to other domains (not just space)

Experiments used the JPL Supercomputing Facility, which is funded by the JPL Office of the Chief Information Officer. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.