THEIA: VISUAL SIGNATURES FOR HADOOP DIAGNOSIS
Elmer Garduno (CMU), Soila P. Kavulya (CMU), Rajeev Gandhi (CMU), Priya Narasimhan (CMU)

VISUAL SIGNATURES
- Signatures of execution for Hadoop jobs
- Compact representation of informative variables
- Discriminate between two types of problems:
  - User centric: bogus jobs, data skew
  - Infrastructure centric: node contention, cluster degradation

SHOWCASE

SCALABLE VISUALIZATIONS
- Keep space bounded as N grows
- \( \text{Rows} \equiv O(\sqrt{N}) \)
- Preserve attribute information
- 1000 nodes can easily fit on standard displays

ON-LINE DIAGNOSIS + VISUALIZATION
- Use spotted patterns to predict failures
- Provide visual feedback / Human-in-the-loop
- Relevant features
  - Success, failed and killed ratio
  - Proportion of total bytes written and read
  - Variance on abnormality
- Use classification trees to find rules
  - \( \text{success reduces ratio} > 54\% \) \& \( \text{success map ratio} > 60\% = \text{SUCCESS} \)
  - \( \text{success reduces ratio} < 54\% \) \& \( \text{success map ratio} < 33\% = \text{FAILURE} \)
- Classify in-progress jobs
  - Accuracy around 0.8 with 40\% of the job completed