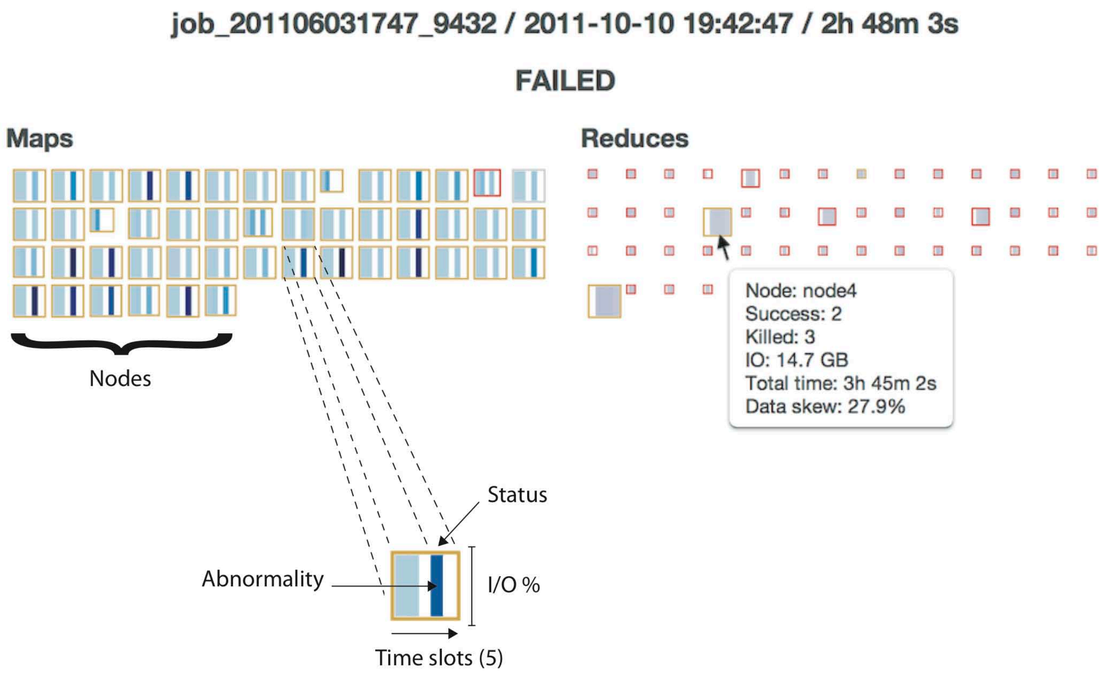
THEIA: VISUAL SIGNATURES FOR HADOOP DIAGNOSIS

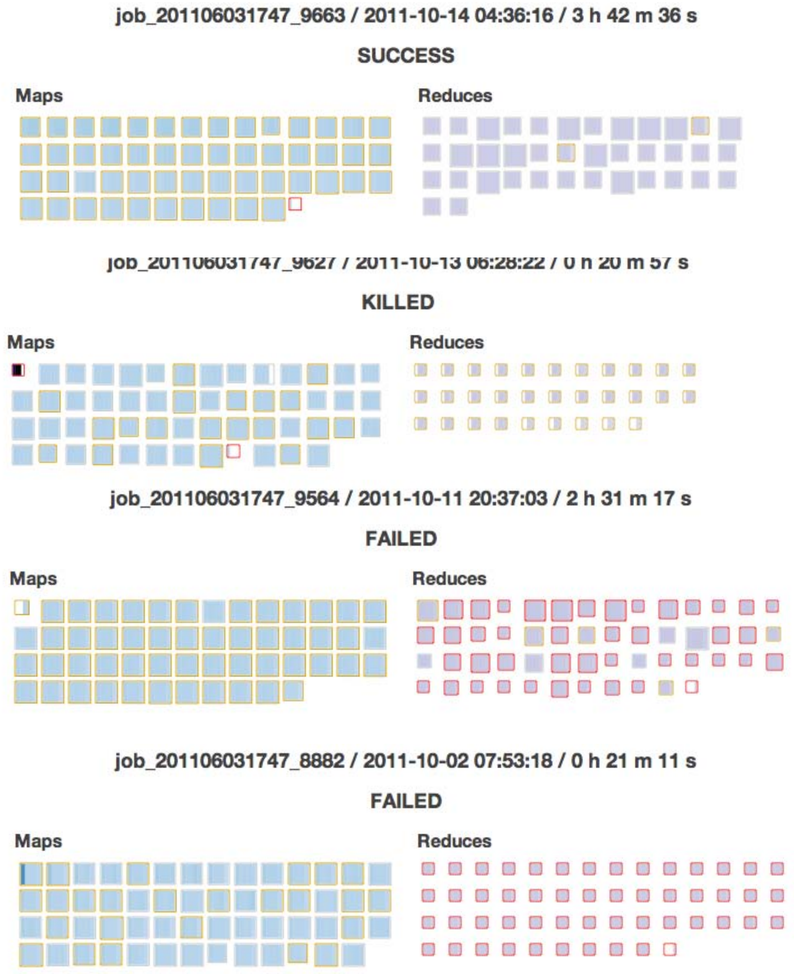
Elmer Garduno (CMU), Soila P. Kavulya (CMU), Rajeev Gandhi (CMU), Priya Narasimhan (CMU)

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

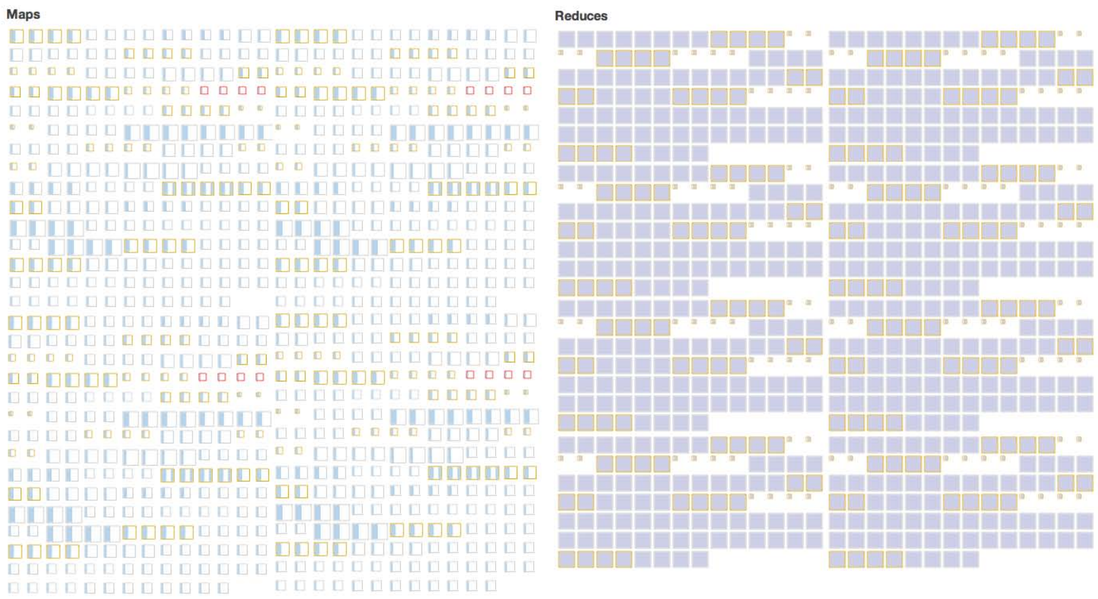


SHOWCASE



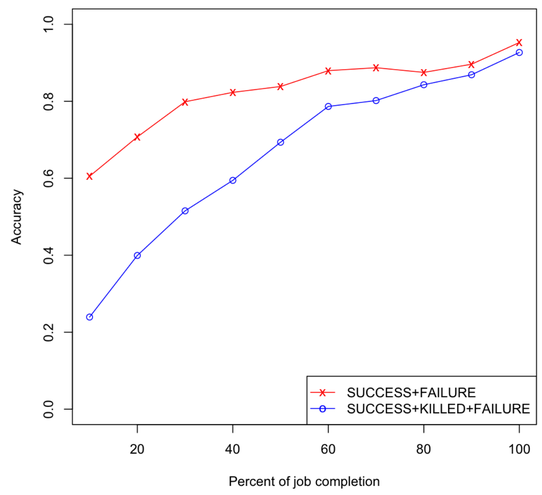
SCALABLE VISUALIZATIONS

- Keep space bounded as N grows
- $Rows \cong O(\sqrt{N})$
- 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
  - success_reduces_ratio > 54% && success_map_ratio > 60% = SUCCESS
  - success_reduces_ratio < 54% && success_map_ratio < 33% = FAILURE
- Classify in-progress jobs
  - Accuracy around 0.8 with 40% of the job completed



Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · PRINCETON UNIVERSITY · UC Berkeley