

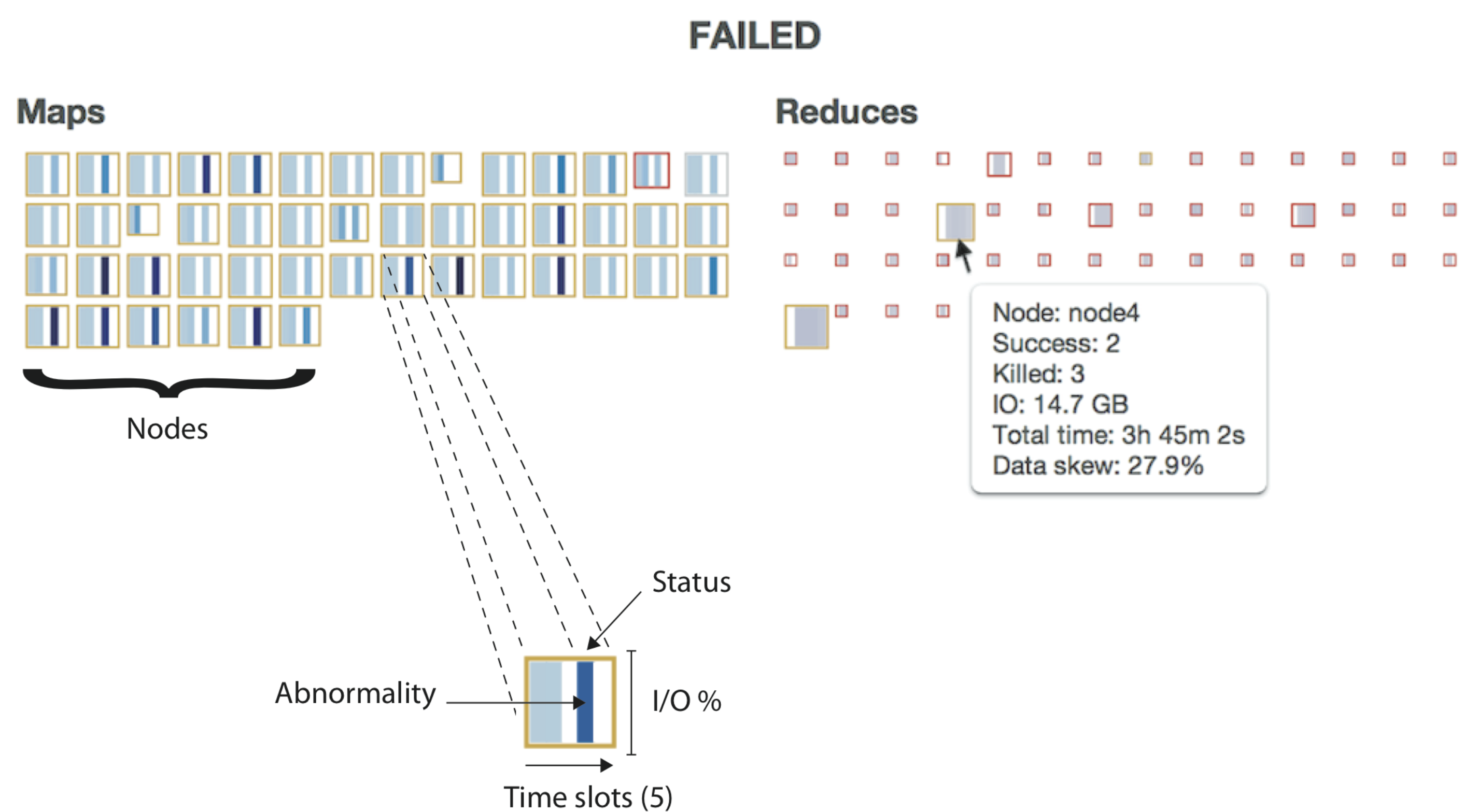
VISUAL SIGNATURES FOR HADOOP DIAGNOSIS

Elmer Garduno, Soila P. Kavulya, Jiaqi Tan, Rajeev Gandhi, Priya Narasimhan (Carnegie Mellon University)

VISUAL SIGNATURES

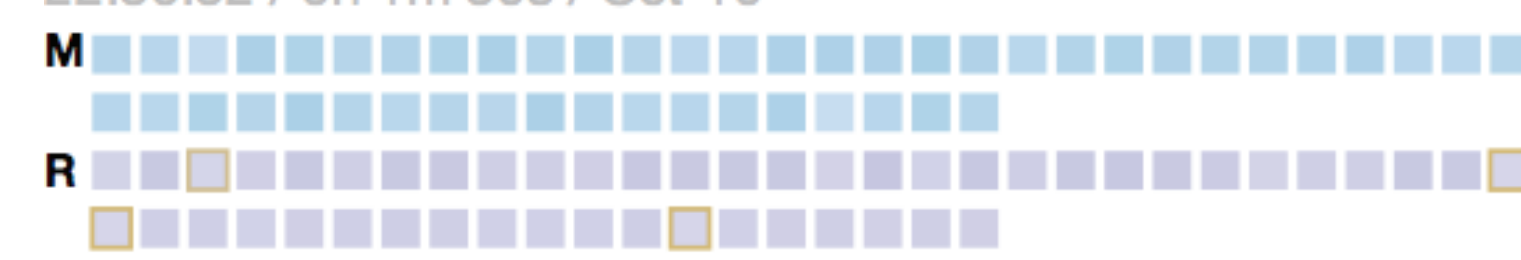
- Signatures of execution for Hadoop jobs
- Compact representation of informative variables
- Discriminate between two types of problems:
 - User centric: bogus jobs, data skew
 - Infrastructure centric: node contention, cluster degradation

job_201106031747_9432 / 2011-10-10 19:42:47 / 2h 48m 3s



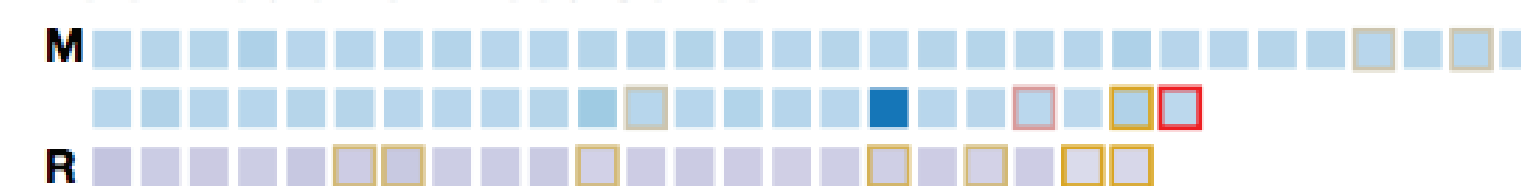
SHOWCASE

job_201106031747_9451 / HALFCAM1_kmeans_iteration_4 -- SUCCESS
22:36:32 / 0h 1m 30s / Oct-10



Success // normal conditions

job_201106031747_8985 / mad_hadoop -- SUCCESS
18:52:20 / 0h 9m 43s / Oct-05



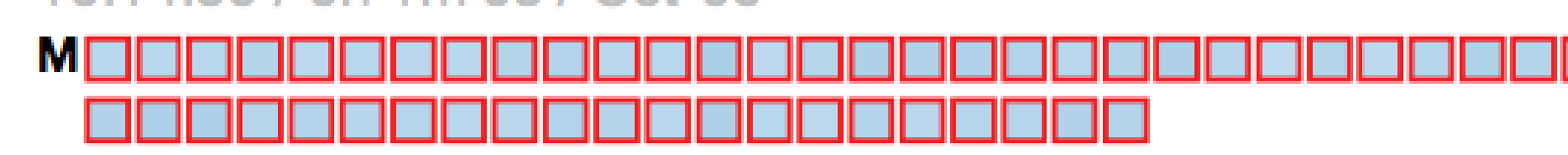
Success // single node failure

job_201106031747_9146 / DocPerLineSentenceExtractor -- SUCCESS
20:25:39 / 0h 17m 36s / Oct-07



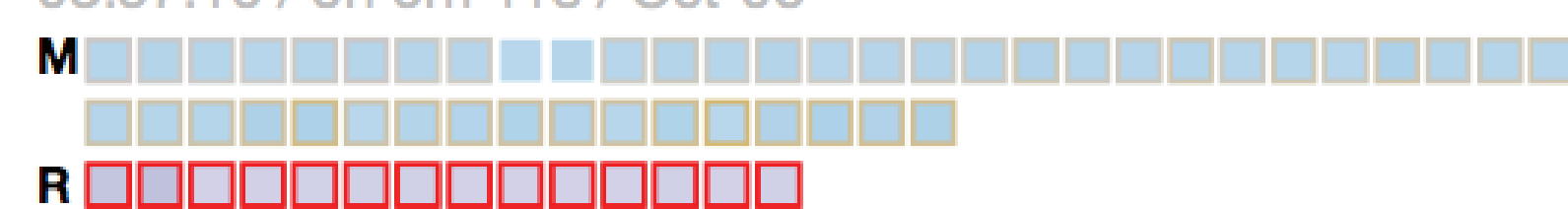
Success // data skew

job_201106031747_8977 / node_lookup_hadoop -- FAILED
10:14:59 / 0h 1m 0s / Oct-05



Failed // bogus map

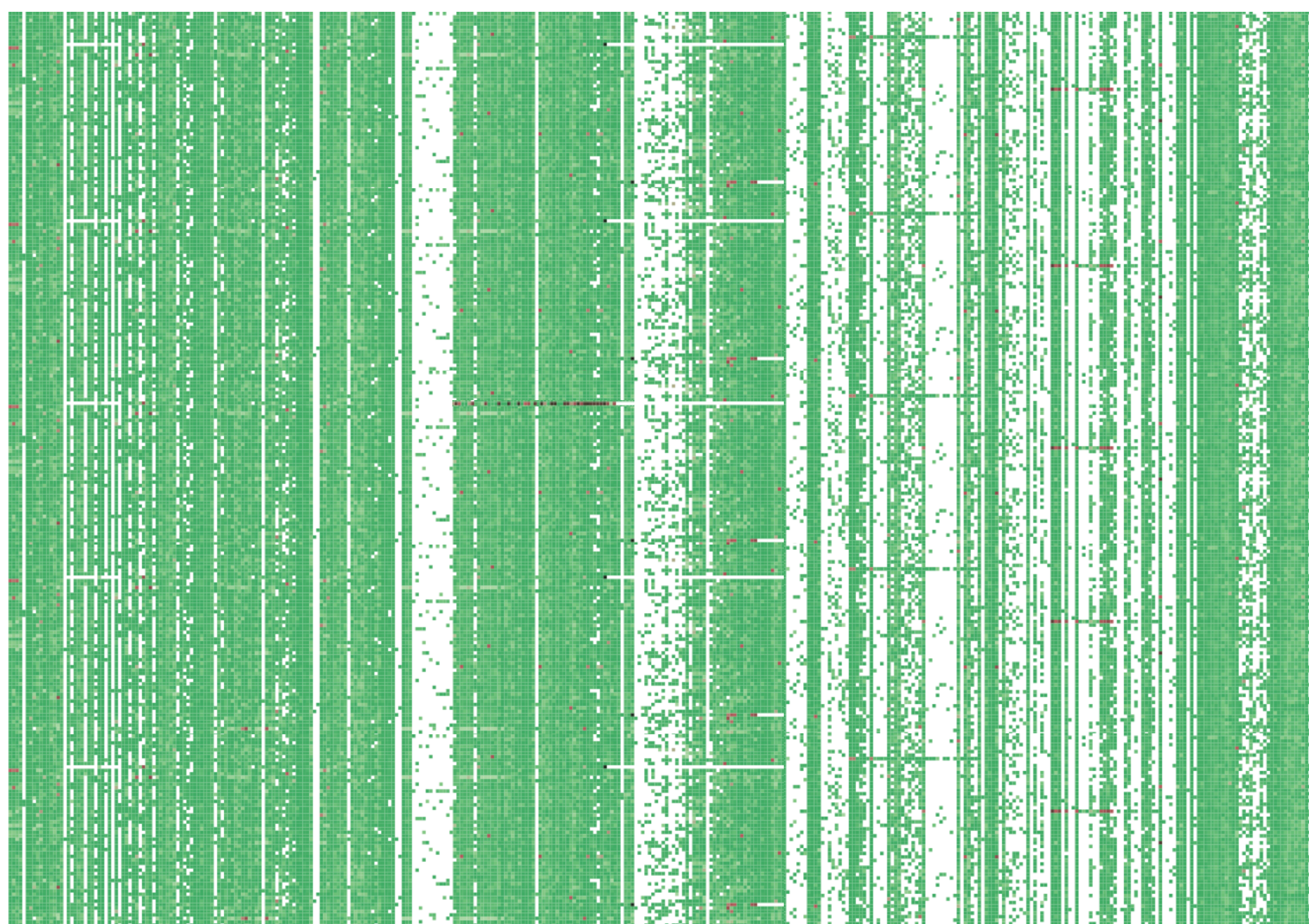
job_201106031747_9035 / depParseFeatExtract -- FAILED
08:57:16 / 0h 6m 41s / Oct-06



Failed // bogus reduce

SCALABLE VISUALIZATIONS

- Data density = num. entries / area of display
- Using 2x2 pixel squares per job/node
- Approximately 2,900 numbers per sq. inch
- 700 nodes x 1200 jobs fit on a 27" display



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

