**The Power of Choice in Data-aware Cluster Scheduling**  
Shivaram Venkataraman, Aurojit Panda, Ganesh Ananthanarayanan, Michael J. Franklin, Ion Stoica

### MOTIVATION
Growing data volumes $\rightarrow$ Need for data-aware scheduling
For timely results, applications process a *subset* of inputs
Examples:
- Approximate Query Processing (Minitable, BlinkDB)
- Machine learning algorithms (SGD)

#### Combinatorial choices!

![Combinatorial choices diagram]

**KMN SCHEDULER**
Choice-aware scheduler
Use “late binding” i.e., choose the subset of data dynamically depending on state of the cluster
Extend benefits across stages using small number of additional tasks

**HOW MUCH LOCALITY?**
Memory locality $\rightarrow$ Orders of magnitude faster
“All or Nothing” implies all K tasks need locality
Hard to achieve on shared clusters with higher utilization
Analysis using uniform slot-utilization model

**Locality vs. Utilization when running K = 100 tasks**

![Locality vs. Utilization graph]

### INTERMEDIATE STAGES
Cross-rack skew slows down network transfers
Insight: Run extra tasks ($M > K$)
Spread out the K tasks chosen to reduce skew

### EVALUATION
Cluster setup: 100 EC2 machines, m2.4xlarge
Workload: Replay of Facebook trace
Baseline: Pre-select random subset of inputs

#### Overall improvements from KMN

![Overall improvements graph]

#### Effect of varying M/K

![Effect of varying M/K graph]

### ALSO IN THE PAPER
Straggle mitigation using extra tasks
Placing reduce tasks to minimize network traffic
Evaluation using Conviva SQL queries and ML algorithms