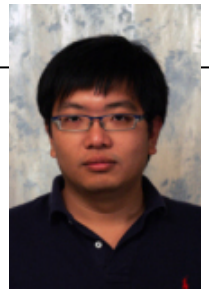
Parallel Machine Learning using GraphLab

Carnegie Mellon

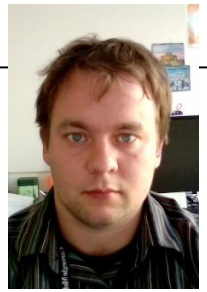
Danny Bickson

Joint work with

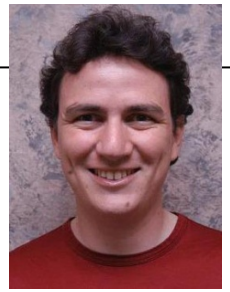
Joseph Gonzalez



Yucheng Low



Aapo Kyrola



Carlos Guestrin

Joe Hellerstein

geekSessions

Machine learning is everywhere

Home | Shop | Answers | Radar: News & Commentary | Safari Books Online | Conferences | Training | School of

O'REILLY radar

Insight, analysis, and research about emerging technologies

Data | Gov 2.0 | Mobile | Programming | Publishing | Web 2.0

The quiet rise of machine learning

Print

Listen

Alasdair Allan on how machine learning is taking over the mainstream.

by Jenn Webb | @JennWebb | Comment | 11 April 2011

Tweet 169 Like 14

The concept of machine learning was brought to the forefront for the general masses when IBM's Watson computer appeared on Jeopardy and wiped the floor with humanity. For those same masses, machine learning quickly faded from view as Watson moved out of the spotlight ... or so they may think.

PANDORA internet radio

Milo Find it local. Get it now!

TAGGED

sense learn act

The challenge: handling big data

PANDORA internet radio

80 million users
8 billion ratings

Milo
Find it local. Get it now!

1.6 million
products

KDD CUP from YAHOO! LABS

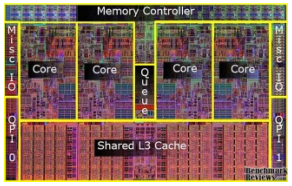
300 million
music ratings

TAGGED™

100 million
users

Problems no longer fit into memory of a single computing node

GPUs



Multicore

Clusters

SuperComputers

amazon web services™

Clouds

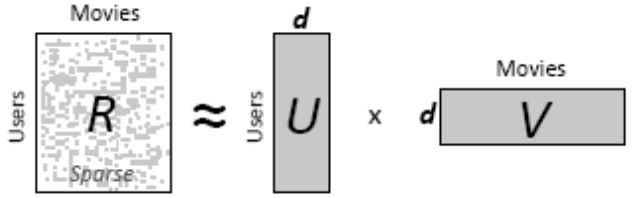
- A natural solution: use 

- Does Hadoop work well for iterative algos?

sense learn act

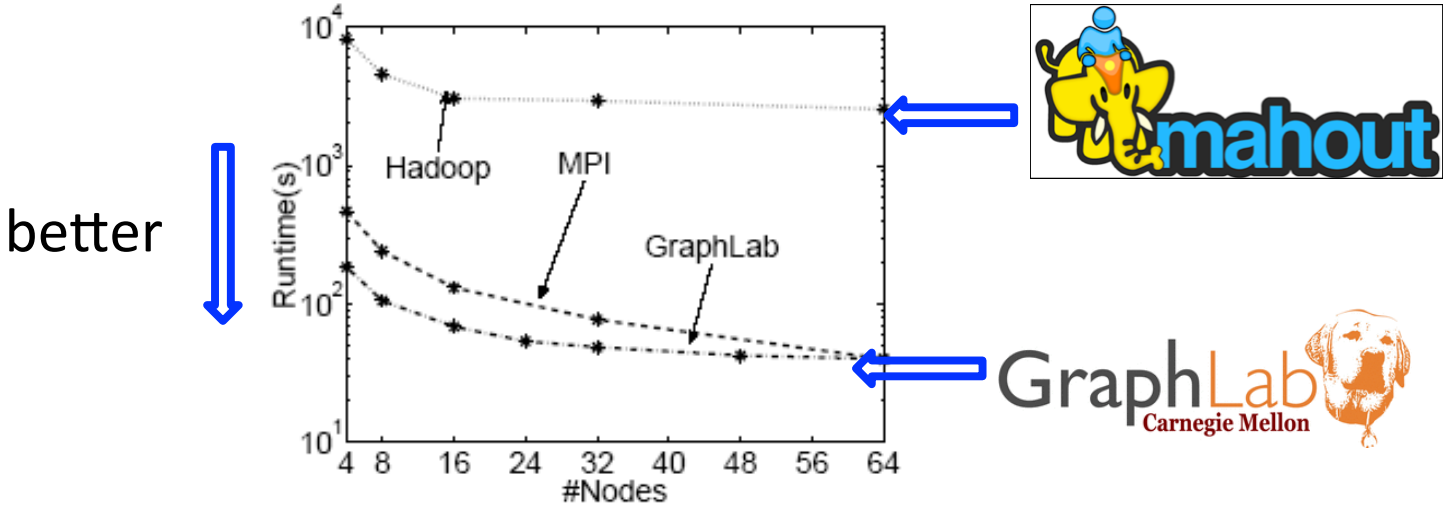
Case study: collaborative filtering

- Computing a linear model for data



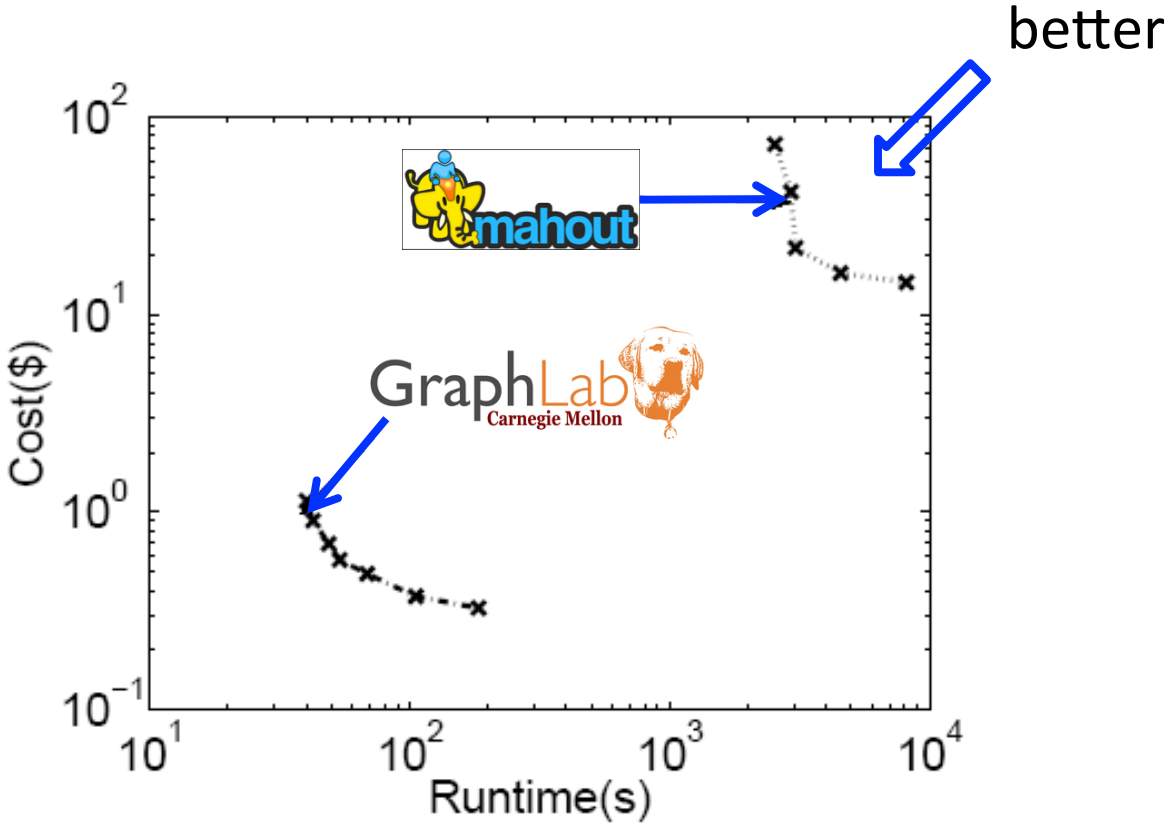
- Implemented alternating least square using GraphLab
- Amazon EC2 runtime results using Netflix data

  (sparse matrix with 100M non-zeros)



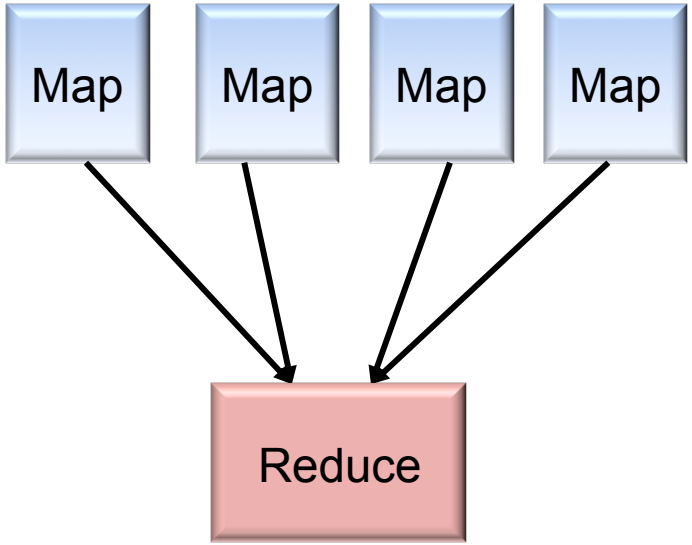
Why should we care?

- Because we are paying!



Map-Reduce

- Composed of two phases
  - Map: computed in parallel
  - Reduce: aggregates the results

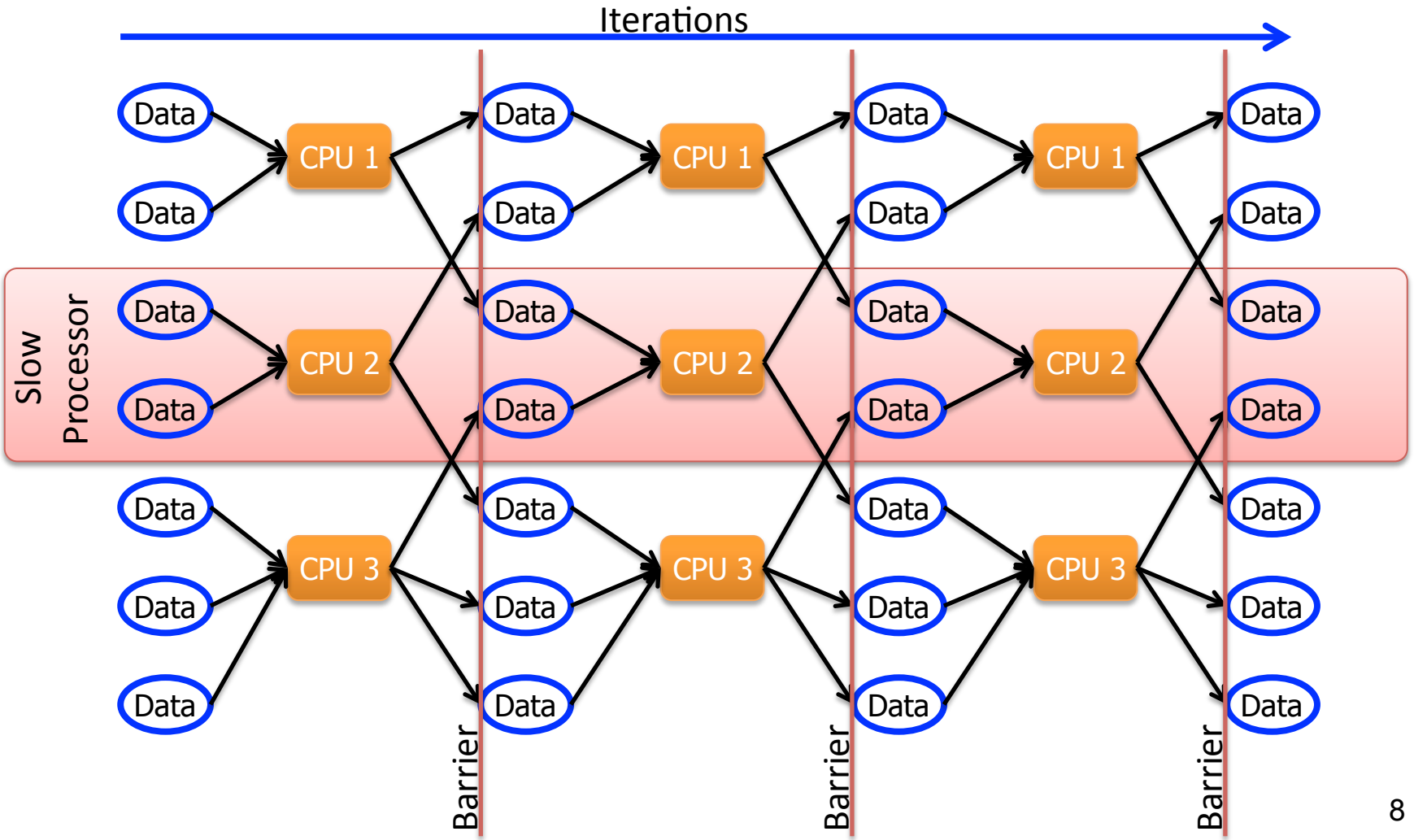


- Pros:
  - Simple framework
  - Fault tolerant

- Cons:
  - Suitable for “embarrassingly parallel” applications
  - Considered inefficient
  - Not suitable for iterative algorithms

Iterative Algorithms?

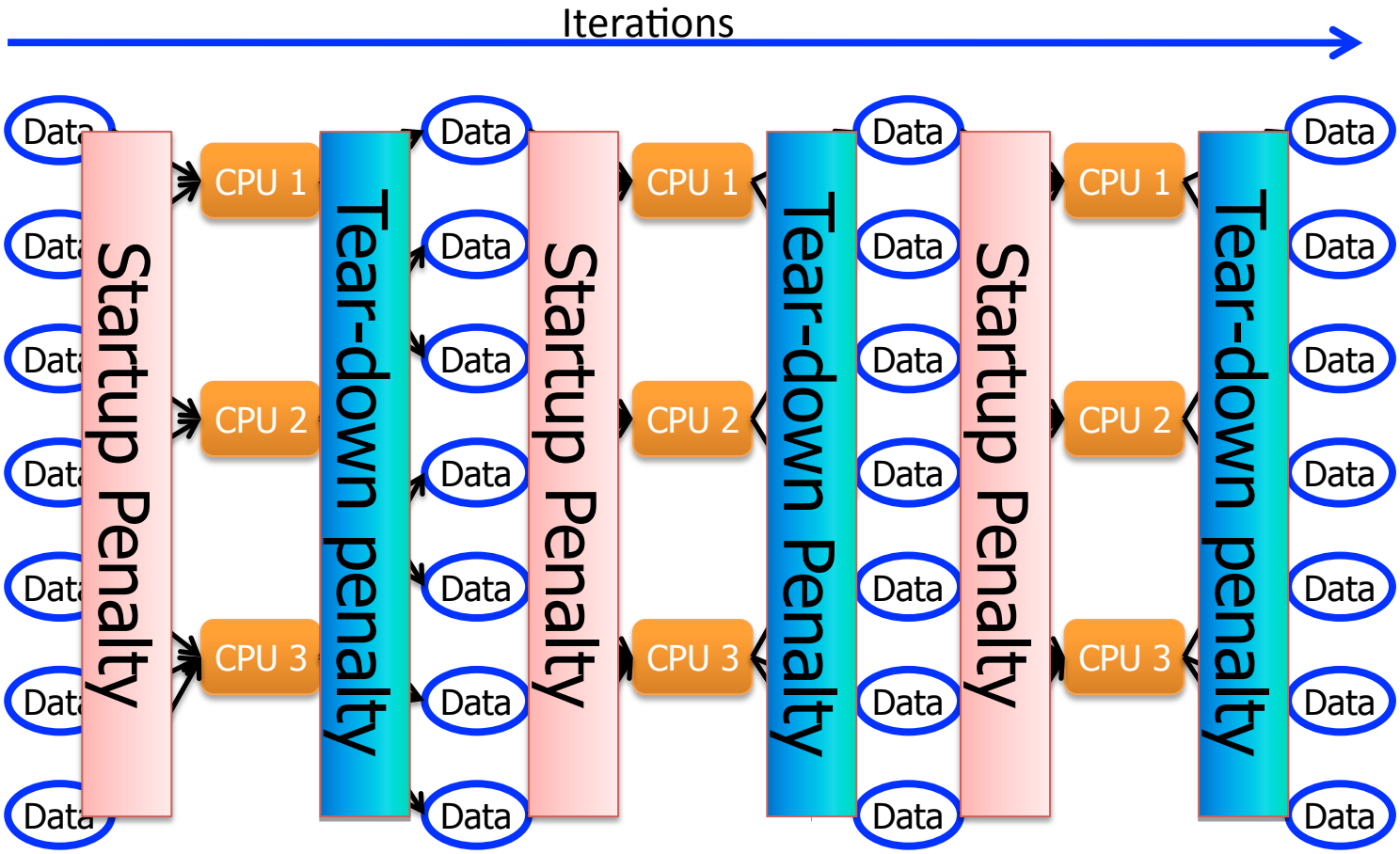
- We can implement iterative algorithms in MapReduce:



sense learn act

Iterative MapReduce

- System is not optimized for iteration:

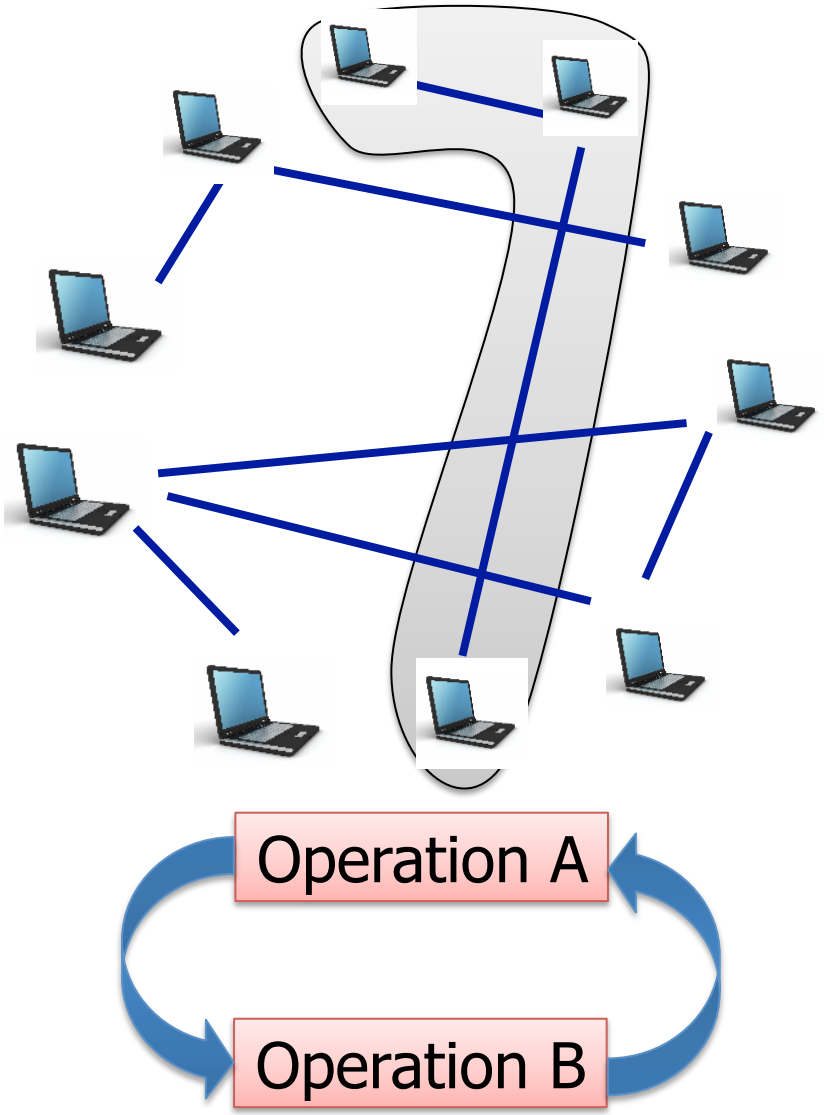


GraphLab targets

1) Sparse Data/Parameter Dependencies

2) Local Computation

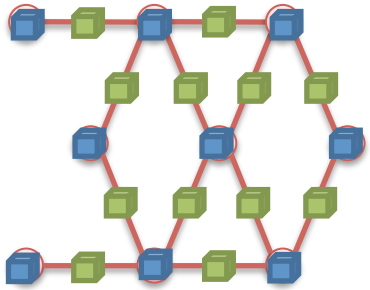
3) Iterative Asynchronous Computation



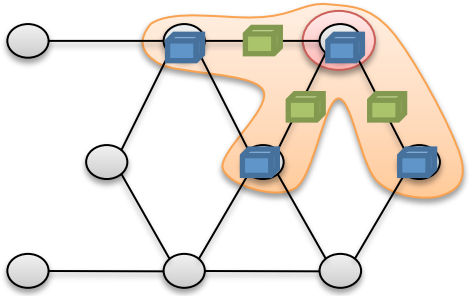
sense learn act

GraphLab Components

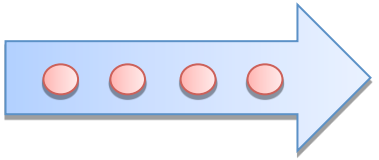
1) Data Graph



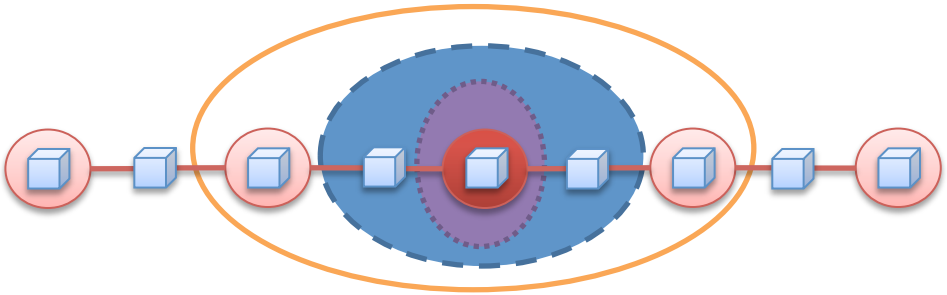
2) Update Functions



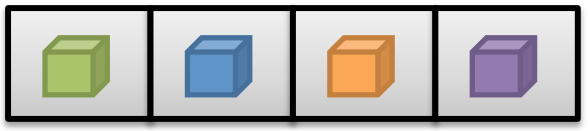
3) Scheduling



4) Consistency Model



5) Shared Data Table



Checkout **GraphLab** Today

Documentation… Code… Tutorials…Applications…

http://graphlab.ml.cmu.edu