Distributed Machine Learning

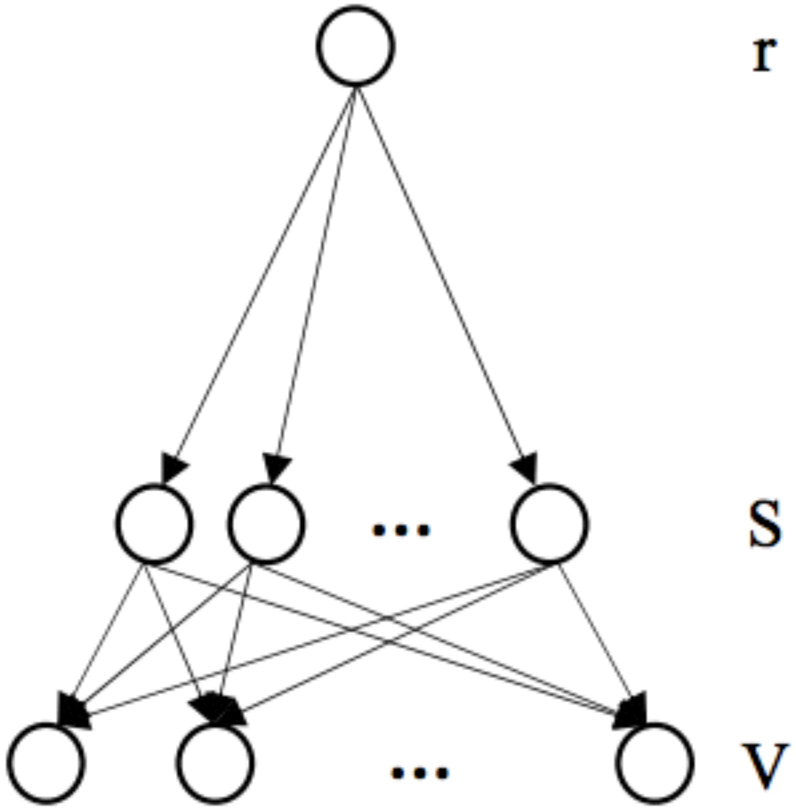
Deep Learning

Yi Wang

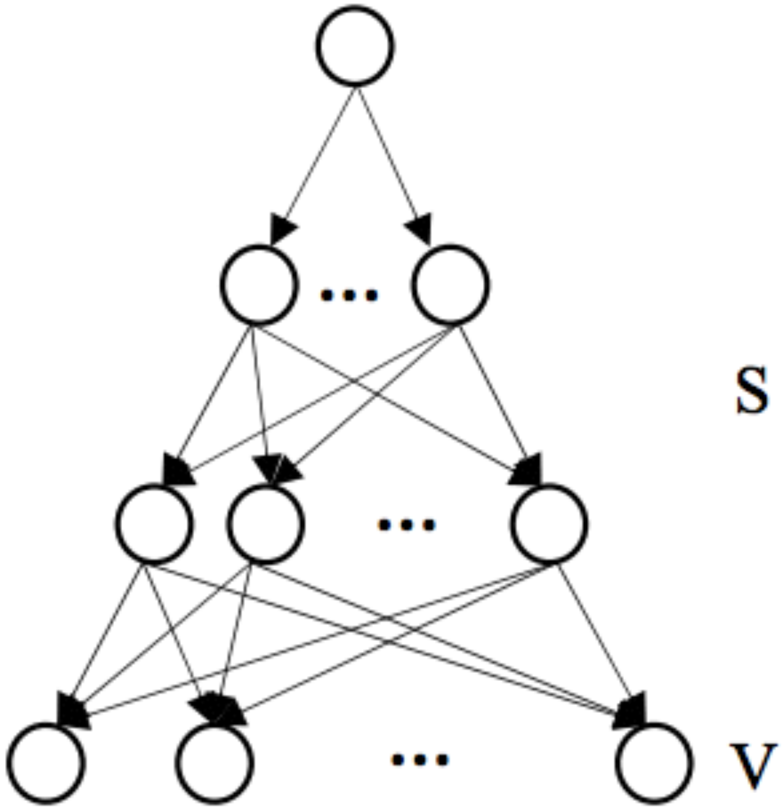
Why Deep Learning?

- Learning the hierarchy of "concepts".
- From the perspective of hierarchical topic modeling.
- Change the building blocks from LDA to RMB.

Hierarchical Topic Models



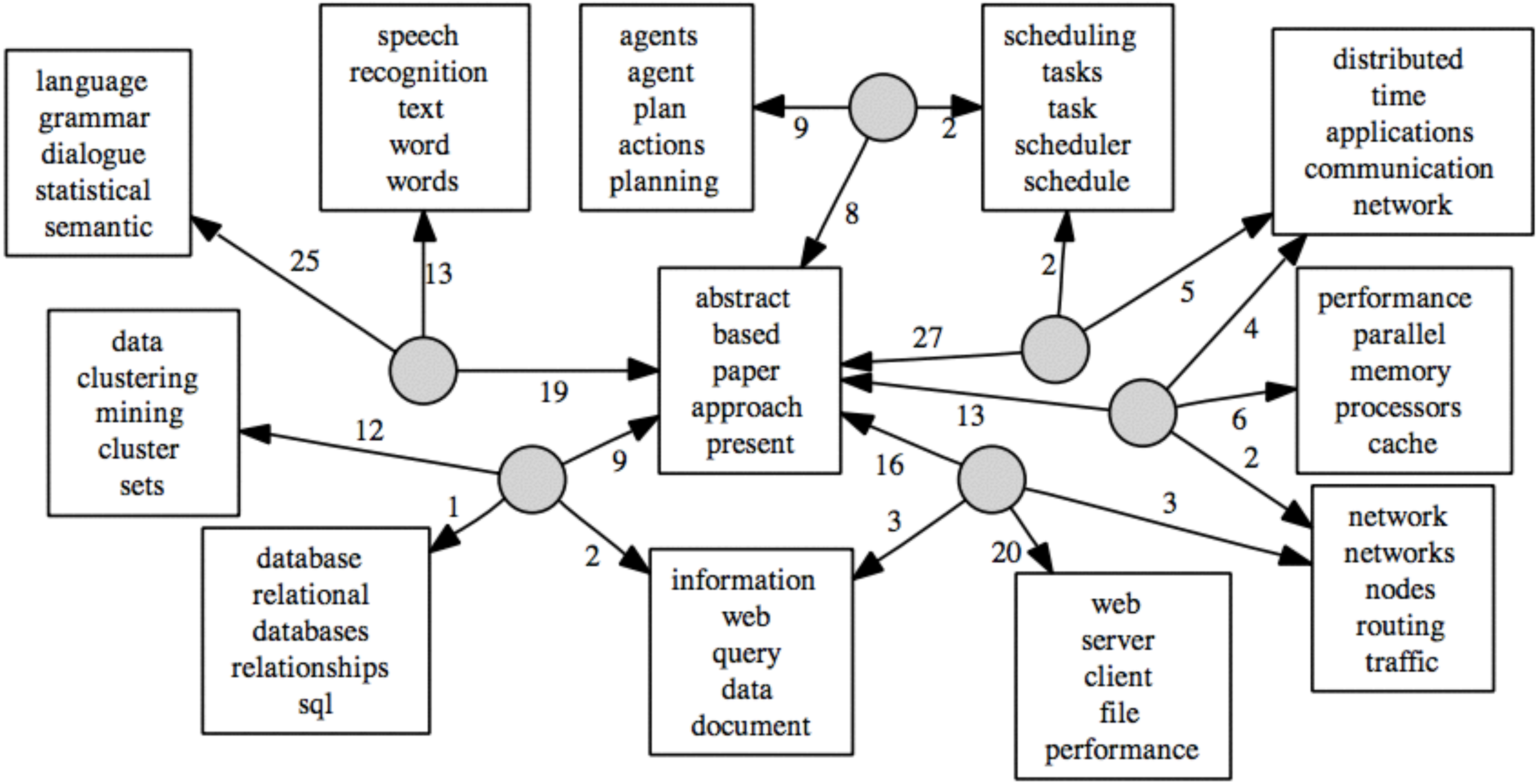
(b) LDA



(c) Four-Level PAM

Learning the hierarchy of concepts.

Concept Hierarchy



Super-topics and topics learned by PAM

http://people.cs.umass.edu/~mccallum/papers/pam-icml06.pdf

Simplification of PAM

- Generating process of LDA
  For each word, there is a latent topic assignment.
- Generating process of PAM
  For each word, there is a latent branch of topics.
- Generating process of Hierarchical-LDA
  For each word, there is a super-topic, then a topic, etc.

Neural Networks

- In 2006, NOCA, comparable with LDA. http://www.cs.berkeley.edu/~jordan/sail/readings/singliar-hauskrecht.pdf
- In 2009, Infinite factor NOCA, comparable with HDP. http://papers.nips.cc/paper/3833-an-infinite-factor-model-hierarchy-via-a-noisy-or-mechanism.pdf
- In 2009, Replicated Softmax, comparable with LDA. http://papers.nips.cc/paper/3856-replicated-softmax-an-undirected-topic-model.pdf

Deep Neural Nets

- Stacking RBM variants over and over.
- Just like stacking LDA over and over.

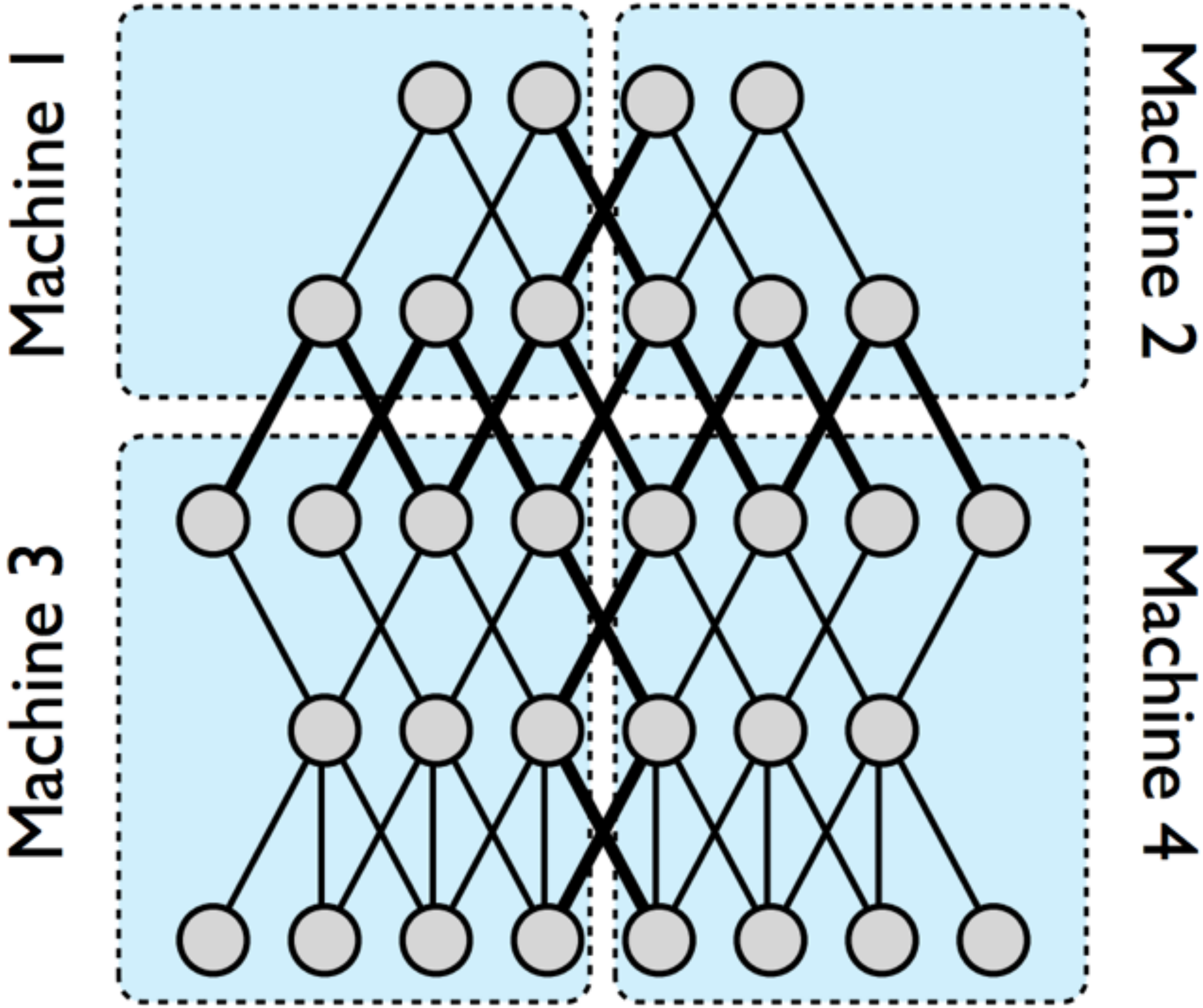
Big Enables Deep

- Stacking models over and over is not a new idea.
  - Deep nets suffers from zero updates in learning.
- Deep nets are reasonable only when data is big.
  - http://arxiv.org/pdf/1003.0358.pdf
  - http://ai.stanford.edu/~ang/papers/nipsdlufl10-AnalysisSingleLayerUnsupervisedFeatureLearning.pdf
  - http://ai.stanford.edu/~ang/papers/icml11-OptimizationForDeepLearning.pdf

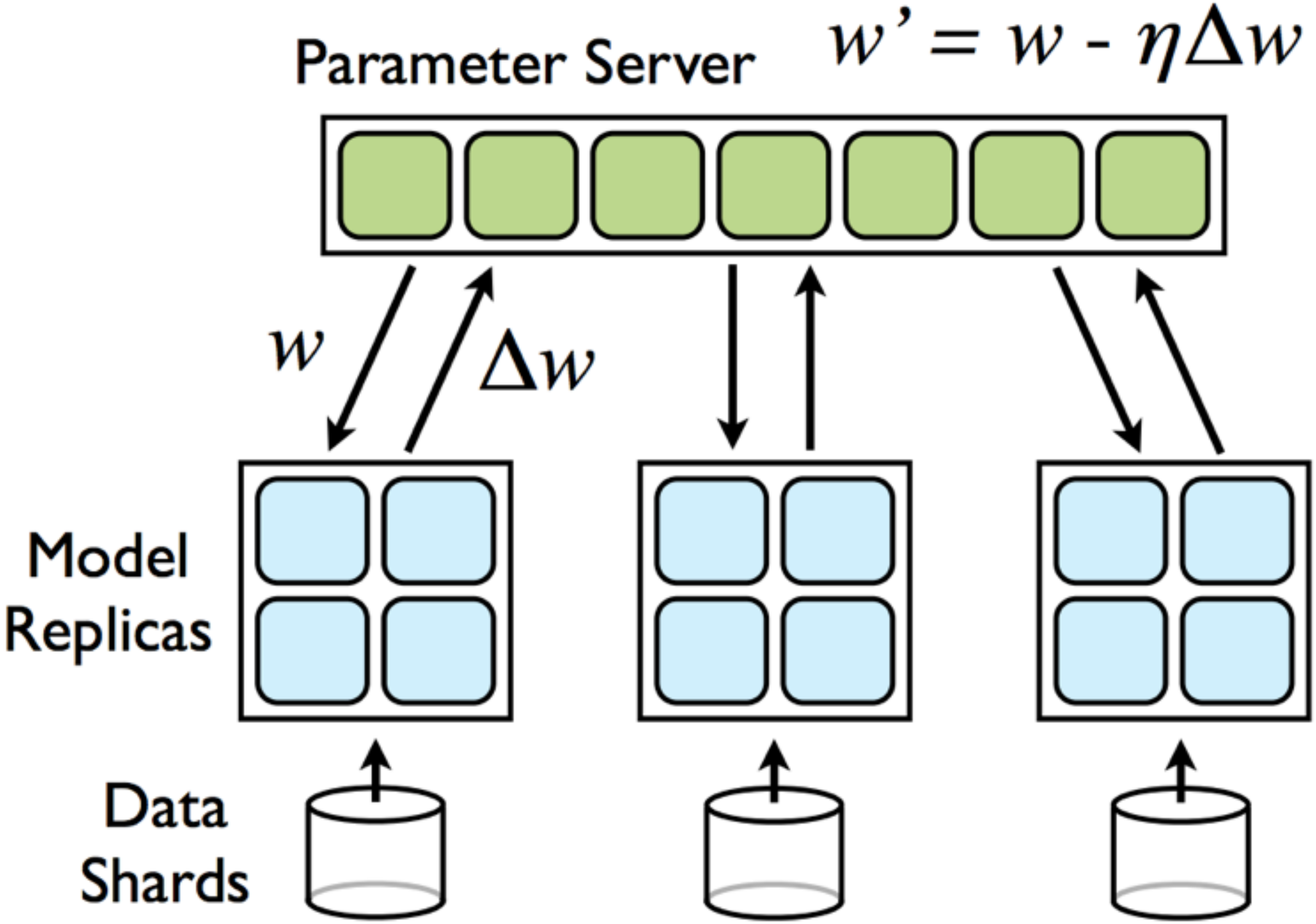
Make It Big

- Parallel Training
  - GPU-based solutions
  - Data less than 6GB-memory of video RAM.
- Distributed Training
  - Google DistBelief
  - http://www.cs.toronto.edu/~ranzato/publications/DistBeliefNIPS2012_withAppendix.pdf

Model Parallelism



Data Parallelism



Asynchronous Update

- DistBelief is good at implementing asynchronous update learning algorithms.
- Asynchronous update differs from math proofs for template algorithms in textbooks. But works better.
- Traditional algorithms are like a bee looking for flowers
- Asynchronous parallel update is like a swamp of bees looking for flowers.
- The swamp is re-gathered irregularly and spread out then to cover a larger area, thus tolerant to local optima.

Future Work

- Learning network topology!
- It is learning the topology of human knowledge.