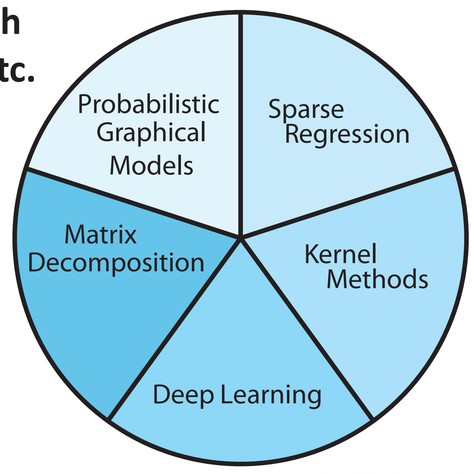
- Sub-bullet
  - sub-sub-bullet

Big Machine Learning: Needs and Directions

Qirong Ho, Jim Cipar, Wei Dai, Jinliang Wei, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B. Gibbons*,
Gregory R. Ganger, Garth Gibson, Eric P. Xing (CMU, *Intel)

BIG PICTURE

- Many ML algorithms, but grouped into families
  - › Five major families (ML Pie)
  - › Each family uses math and techniques
- ML researchers keep implementing from scratch
  - › Effort duplicated on MPI, distributed state, etc.
  - › Little code reuse between research groups
- Need system support for Big ML that:
  - › Handles computation and data partitioning
  - › Addresses unique aspects of ML families
  - › Facilitates future systems-ML research

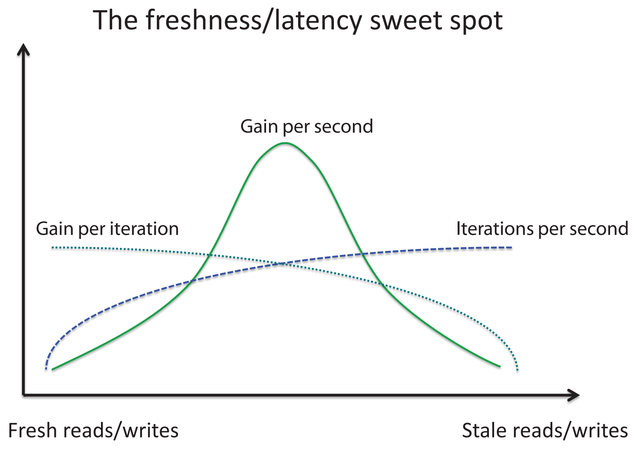


WHAT'S AN ML ALGORITHM?

- Most "ML algorithms" are combo of two things:
  - › Mathematical/statistical model
  - › Algorithmic technique to solve the model
- This ML pie is model-centric
  - › Family members share mathematical properties
  - › Techniques may be used in more than one family

STALENESS

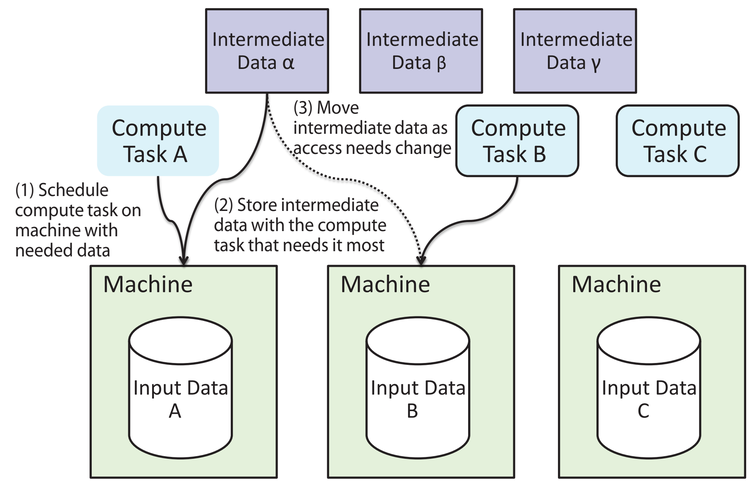
- ML algorithms iterate until convergence
  - › Minor errors in intermediate data induce more iterations, but don't prevent convergence
- Each iteration reads/writes shared intermediate data
  - › Locking quickly becomes a bottleneck
  - › Limited network bandwidth a secondary bottleneck
- Big idea: let threads work on stale data
  - › More iterations, but often much faster
- See LazyTables poster for more info!



LOCALITY

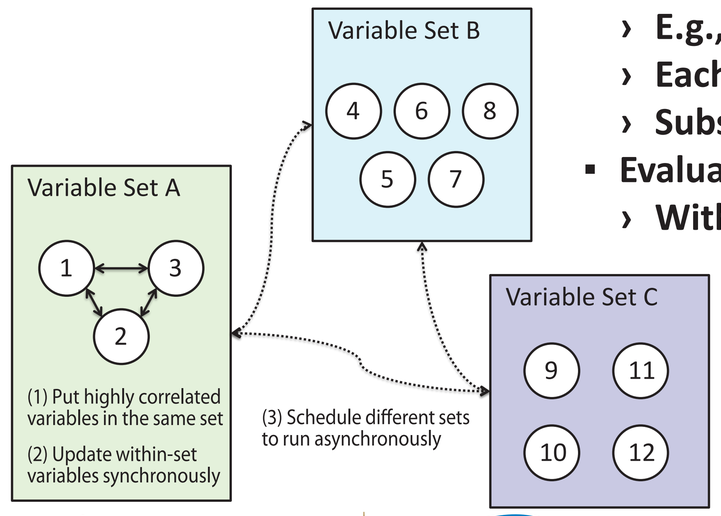
- Intermediate data is often large and distributed
  - › Computation should be near intermediate data being used
- Intermediate data usage patterns change over time
  - › May shift focus within intermediate data
  - › Adaptive placement required to maximize performance
- Staleness can help with locality
  - › LazyTables caches stale copies near each thread

Locality: Input Data, Intermediate Data, and Computation



ADAPTIVE SYNCHRONIZATION

- Synchronization among threads can help and hurt
  - › Slows iteration rate and often not needed
  - › But, highly correlated variables converge slowly without it
- Big idea: dynamic variable sets
  - › Within a set, update are synchronized
  - › Between sets, they are not
  - › Adaptively formed by measuring correlation



ONGOING EXPLORATIONS

- Refining the ML pie
  - › Useful for identifying systems opportunities
- Evaluating existing platforms
  - › E.g., GraphLab, Spark, Piccolo
  - › Each has strengths and weaknesses
  - › Substantial slices of pie served by none
- Evaluating new systems support ideas
  - › With real ML algorithm implementations

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington