Overarching Themes in this Area

Recent Successes (last 3 years)

Major Obstacles Impeding More Rapid Progress

Areas of Neglect

Strategic Priorities & Investments That Will Advance Innovation

RAW NOTES BELOW

Q: What are the overarching themes in this area>
- How should we define *algorithms for big data?*
+ Multi-scale
    + Multi-physics
- What does it mean by *structured vs unstructured* data?
    + How to find appropriate structure in structured data?
    + How should we begin to analyzing unstructured data? This is data about which we know nothing -- no provenance.
- Are the previous two questions encompassed under the question: What is the *right model of computing* on big data?
    + This model should capture intentions and working of batch systems like map reduce, as well as streaming systems.
    + It should also gives us ``data structures'' to think about big data.
    + Muthu: What will be the one data strcuture that will become popular? Will it be graph? Or Tensors? Every model has external interfaces -- how does it take input, how does it produce its output. Maybe thinking about these interfaces might help in getting an answer.
- How should we proceed in any task in this area? Ask research questions first, analyze data later or vice versa.
- Most algorithms here require giving up correctness
    + One reason is because problem definitions/questions are fuzzy.
- Other theme is all algorithms deal with some tradeoff of resources. These resources can be:
    + storage
    + communication
    + accuracy of result
    + data collection
    + etc.

Q: Recent succcesses that can be considered foundational (and popular)
- Phase field modeling of mechanics
- "Large scale" linear algebra methods
- Sketching methods (incl. for graph oriented data)
- Iso-geometric methods
- Stochastic approximation techniques

- Tensor methods? (Not many people in the audience aware of this area)
- **Interaction between (machine) learning and optimization**
    + Representation learning is one of the biggest successes of ML.
- New algorithms to optimize read/write performance for a variety of memory hierarchies

Q: Strategic priorities to advance innovation
- We understand how to analyze data with high n (number of data points) and low d (dimension of data). What about the case when n is very small (in the hundreds) and d is very large (in the millions)? Most existing methods fail.
    + Arises in biology (drug testing? cancer research?)
    + Is "extreme learning" related to this?
- Large scale non-convex optimization
- Analysis of feedback systems, multi armed bandits.
- Broader applicability of ideas.
    + Need discussion about applicability & accessiblility both.
    + With every idea, explain the domain that drives the idea (or problem) and explain how will a solution to the problem change the domain (for the better)?

Q: Obstacles
- Gap (time, resources) between data scientists (users) and computer scientists (researchers)
- Access to datasets? Raised in every conference/meeting. No one knows how to solve this.
- Theory community should not restrict itself to picking up problems from the "Internet" compaines only. Look at other communities (cancer research, genomics), take time to develop new abstractions and problems.
- Train students to think about foundational problems.