BREAK OUT SESSION

Foundational Big Data Algorithms and Theory

BIG DATA PI Meeting 2016

Overarching Themes in this Area

- **Conceptual models** of data:
  - And how to map data to models (finding structure)
  - Right programming/interface models for interacting with data
- **Tradeoffs** between (new) resources:
  - time, space, communication, approximation (error), randomness, statistical efficiency, verifiability
- **Interpretability**: What is an answer?
- **Workflow:** Hypothesis-driven (question followed by algorithm+data) or exploratory (the reverse)

Recent Successes (last 3 years)

Redefine “3”

Recent Successes (last 3 years)

- **Tradeoffs:**
  - (sublinear resource) computations, especially on graphs
  - Managing multi-memory hierarchies.
  - Programming abstractions
  - Stochastic approximations
- **Representations**:
  - Tensors, graphs and social networks.
  - Nonparametric representations
  - (Hierarchical) representation learning
- **Computation:**
  - Managing high dimensions (projections, hashing)
  - Optimization and learning
  - Submodularity, non-convexity and beyond

Major Obstacles Impeding More Rapid Progress

- "more rapid than what?"
- Tension between day-to-day work and abstractions
- Data access: not just for utility, but to ask the most effective kind of foundational questions.
- Glaring spotlight of "internet problems" in comparison to harder science problems
- Communication:
  - Between groups inside foundations
  - Between disciplines

Strategic Priorities & Investments That Will Advance Innovation

- Still need to understand resource tradeoffs (communication/space/time/samples/estimation/etc)
- Understanding inference and computation with large dimensions, small samples (or small access)
- Inference with lots of labels (personalization/feature learning)
- Non-convex optimization
  - Saddles
  - Non-convex regularization
  - Other convex relations
- Feedback systems ((po)-MDPs, bandits etc): what happens once you act on the prediction?

Areas of Neglect

- Problems of validation with limited access to fresh data: methodologies for evaluation when constrained by amoun t of annotation
- Understanding consequence of complex human-machine interactive systems (e.g. fairness, security)
- Deeper integration of [shape] and [inference]
- Foundational exploration of causality