Big Data

David Karger
Somebody Else’s Problem

David Karger
What is Big Data?

• Enthusiasm for power of data to provide insight
• Right at limits of what computers can handle
• Apply techniques that are trivial on small data
  – Nearest neighbor
  – Regression
  – Shortest paths
• Everything sacrificed for sake of performance
• Obvious tremendous potential
• Big numbers: interest/research/money
What is the Semantic Web Angle?

Not Clear there is One
All About Performance

• Must manage data locally, not on web
• Performance is maximized by hard-coding
  – Removing unneeded generality from system
  – Anti Semantic Web
• SQL Databases with specific tables will outperform triple stores
  – Orri: taking away the schema costs you 5x
• Specialized ML will outperform general ML
• Analysts will trade ease of use for performance
Data vs. Schema

• Semantic Web strength is mutable schemas
• Focus on richness of schematic structure
• In big data, ratio of schema to data goes to 0
• Arbitrarily difficult schema work (e.g. alignment, understanding) becomes negligible compared to data processing
• So SW addresses the unimportant piece
Unity

• Given critical role of databases, tackling big data outside their community will fragment the work and make it less effective
  – Publish in SIGMOD/VLDB!

• Given critical role of machine learning, tackling text mining outside their community will fragment the work and make it less effective
  – Publish in NAACL/KDE!
Caveats

• What would a SW big-data problem look like?
• Would have to involve massive numbers of schemas
• E.g., the Semantic Web Search Engine