RICE George R. Brown School of Engineering Computer Science

# Big Data and Data Science: Current State and a Call to Arms

Chris Jermaine

1.30.2014

## Big Data At Rice: Summary

- We've got a lot of high-quality human resources in the area...

## Big Data At Rice: Summary

- We've got a lot of high-quality human resources in the area...
- We've got natural advantages: proximity to energy ind. & TMC



## Big Data At Rice: Summary

- We've got a lot of high-quality human resources in the area...
- We've got natural advantages: proximity to energy ind. & TMC
- And some interesting projects underway...

## Big Data At Rice: Summary

- We've got a lot of high-quality human resources in the area...
- We've got natural advantages: proximity to energy ind. & TMC
- And some interesting projects underway...
- But the sum of "Big Data" research at Rice is less than the parts...



## Big Data At Rice: Summary

- We've got a lot of high-quality human resources in the area...
- We've got natural advantages: proximity to energy ind. & TMC
- And some interesting projects underway...
- But the sum of "Big Data" research at Rice is less than the parts...
- What can we do about this going forward?

## Human Resources: A Strength



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
  - **All are relevant to big data!**



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
  - **All are relevant to big data!**
  - Can cast a wider net within engineering as a whole...

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
- Rich Baraniuk (ECE, machine learning; a "NIPS guy")



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
- Rich Baraniuk (ECE, machine learning; a "NIPS guy")
- Genevera Allen (Stats, machine learning esp over medical data)

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
- Rich Baraniuk (ECE, machine learning; a "NIPS guy")
- Genevera Allen (Stats, machine learning esp over medical data)
- Erzsébet Merényi (Stats, ECE, machine learning, remote sensing)



## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
- Rich Baraniuk (ECE, machine learning; a "NIPS guy")
- Genevera Allen (Stats, machine learning esp over medical data)
- Erzsébet Merényi (Stats, ECE, machine learning, remote sensing)
- Bill Symes (CAAM, very large PDE systems, seismology)

## Who Does Big Data at Rice?

- Me (Chris Jermaine; data analytics, I'm a "SIGMOD guy")
- Eugene Ng (networked systems, esp. for data center and cloud)
- Alan Cox (distributed systems)
- Devika Subramanian (machine learning)
- Vivek Sarkar (programming systems for dist/parallel comp)
- Rich Baraniuk (ECE, machine learning; a "NIPS guy")
- Genevera Allen (Stats, machine learning esp over medical data)
- Erzsébet Merényi (Stats, ECE, machine learning, remote sensing)
- Bill Symes (CAAM, very large PDE systems, seismology)
  - — Could add many more names...



## And We Have Interesting Ideas

## Ultra-Low Power BD Computing

- The problem: Power costs at least 50% of data center
  - — Half of those (cooling and infrastructure) invariant to energy prices
  - — Not to mention $CO_2$ and other environmental costs



## Ultra-Low Power BD Computing

- The problem: Power costs at least 50% of data center
  - — Half of those (cooling and infrastructure) invariant to energy prices
  - — Not to mention $CO_2$ and other environmental costs
- Proposed solution: Ultra-low-power "probabilistic" data centers
  - — Run chips fast at really low power, radically reduce energy cost and use
  - — Problem: you introduce errors (ex: bits get flipped randomly)
  - — But many BD analytics computations are randomized/tolerant to error
  - — Re-build the stack from network thru OS thru application layer to support this

## Ultra-Low Power BD Computing

- The problem: Power costs at least 50% of data center
  - — Half of those (cooling and infrastructure) invariant to energy prices
  - — Not to mention $CO_2$ and other environmental costs
- Proposed solution: Ultra-low-power "probabilistic" data centers
  - — Run chips fast at really low power, radically reduce energy cost and use
  - — Problem: you introduce errors (ex: bits get flipped randomly)
  - — But many BD analytics computations are randomized/tolerant to error
  - — Re-build the stack from network thru OS thru application layer to support this
- Impact
  - — Some significant fraction of BD can move out of traditional data center
  - — To alternative low-cost, low-power data center



## BOLD Project

- The problem: Packet-switched networks sub-optimal for BD
  - — Don't support broad/multi-cast well
  - — Limited point-to-point bandwidth
  - — High energy

# BOLD Project

- The problem: Packet-switched networks sub-optimal for BD
  - — Don't support broad/multi-cast well
  - — Limited point-to-point bandwidth
  - — High energy
- Proposed solution: Have an optical "co-network"
  - — Allow app to use it for BD computations
  - — Very low power
  - — Infinite bandwidth broad/multi-cast
  - — Are physically building this now!



# BOLD Project

- The problem: Packet-switched networks sub-optimal for BD
  - — Don't support broad/multi-cast well
  - — Limited point-to-point bandwidth
  - — High energy
- Proposed solution: Have an optical "co-network"
  - — Allow app to use it for BD computations
  - — Very low power
  - — Infinite bandwidth broad/multi-cast
  - — Are physically building this now!
- Impact
  - — Greatly expand the size/scope of BD computations possible in a data center

# SimSQL Project

- The problem: Data-oriented code, once-written, lives forever
  - — Ex: BNYM: Runs 375 million lines of Cobol
  - — Locks you into a particular hardware stack data layout
  - — Much of systems-oriented efforts in BD (ex: ML on GPUs) forgets this



# SimSQL Project

- The problem: Data-oriented code, once-written, lives forever
  - — Ex: BNYM: Runs 375 million lines of Cobol
  - — Locks you into a particular hardware stack data layout
  - — Much of systems-oriented efforts in BD (ex: ML on GPUs) forgets this
- Proposed solution: Data Independence
  - — Program logically/declaratively: data/hardware changes, code stays
  - — Core idea in relational model: Got EF Codd a Turing Award
  - — Can recycle and extend ideas from parallel RDBMS systems for this space

## SimSQL Project

- The problem: Data-oriented code, once-written, lives forever
  - — Ex: BNYM: Runs 375 million lines of Cobol
  - — Locks you into a particular hardware stack data layout
  - — Much of systems-oriented efforts in BD (ex: ML on GPUs) forgets this
- Proposed solution: Data Independence
  - — Program logically/declaratively: data/hardware changes, code stays
  - — Core idea in relational model: Got EF Codd a Turing Award
  - — Can recycle and extend ideas from parallel RDBMS systems for this space
- Impact
  - — Can have a small number of ML codes for most common models
  - — Will be "future proof", just drop in new system backend



## Sooo... Have Some Nice Pieces

- Excellent human resources
- Interesting research

## Sooo... Have Some Nice Pieces

- Excellent human resources
- Interesting research
- But somehow we punch a bit below our weight
  - — Not even a website!
  - — And there are real costs to this
  - — Ex: Berkeley AmpLab has attracted $7M in industrial BigData funding



## Sooo... Have Some Nice Pieces

- Excellent human resources
- Interesting research
- But somehow we punch a bit below our weight
  - — Not even a website!
  - — And there are real costs to this
  - — Ex: Berkeley AmpLab has attracted $7M in industrial BigData funding
  - — True, Berkeley is adjacent to world's tech epicenter
  - — But we're adjacent to world's energy epicenter, and they have many BD problems
  - — And we're adjacent to the TMC

## Raising the Profile: Potential Approaches



## Raising the Profile: Potential Approaches

- Invest $100M in Big Data, a la Rochester

## Raising the Profile: Potential Approaches

- Invest $100M in Big Data, a la Rochester
  - — Awesome idea! Should be on the table...
  - — But are there other, more modest options?



## Other Options...

## Other Options...

- What we **must** do:
  - — Develop a web presence
  - — Make it clear to external world that Rice CS/Engineering works in Big Data



## Other Options...

- What we **must** do:
  - — Develop a web presence
  - — Make it clear to external world that Rice CS/Engineering works in Big Data
- What we **should** do:
  - — Create an organizational umbrella, call it a Big Data “center”
  - — Bring people together periodically for technical interactions
  - — Issue: not going to happen without external help (K2I?)

## Other Options...

- What we **must** do:
  - — Develop a web presence
  - — Make it clear to external world that Rice CS/Engineering works in Big Data
- What we **should** do:
  - — Create an organizational umbrella, call it a Big Data "center"
  - — Bring people together periodically for technical interactions
  - — Issue: not going to happen without external help (K2I?)
- What we **might** do:
  - — Coordinate more with other eng. departments (stats, ECE, applied math)
  - — Pool resources (esp. faculty positions)
  - — Issue: cultural/political barriers, going to require buy-in from admin



## And of Course...

- Get more faculty positions!
  - — Issue: this is what everyone asks for out of a review like this...
  - — Need to be creative

## And of Course…

- Get more faculty positions!
  - — Issue: this is what everyone asks for out of a review like this...
  - — Need to be creative
- One idea: look to external funding sources
  - — Ex: CPRIT: state gives $5M over five years for senior hire (NAE level)
  - — $4M over five years for junior hire
  - — Only compete within Texas
  - — Issue: need to "oncify" the potential recruit
  - — Need buy-in from the administration



## Summary

- We have excellent human resources
- And great projects
- But no (current) overall strategy or organization
- Abdicating the Big Data/Data Science space is not an option!