LA-UR- 11-03122

*Approved for public release;
distribution is unlimited.*

Title: Emerging Computational Challenges

Author(s): Sarah Michalak

Intended for: posting on Statistical and Applied Mathematical Sciences Institute (SAMSI) website

Los Alamos
NATIONAL LABORATORY
EST. 1943

Emerging Computational Challenges

Sarah Michalak

Real-time and near-real-time processing of massive streaming data presents several challenging computation research problems that statisticians could contribute to. First is the computational environment itself. The development of resilience methods that enable applications to continue running to correct answers in the face of degraded states and faults that could lead to interrupts and silent data corruption are required. Second is the adaptation and development of statistical methods to the streaming setting.

Emerging Computational Challenges

Sarah Michalak

Statistical Sciences Group

Los Alamos National Laboratory

michalak@lanl.gov

Collaborators

- John Bent
- David Bigelow
- Sean Blanchard
- Nathan Brown
- Carolyn Connor
- John Daly
- Nathan DeBardeleben
- Andy DuBois
- Dave DuBois
- Gary Grider
- Andrew Manuzzato
- Dave Modl
- Laura Monroe
- John Morrison
- Heather Quinn
- Bill Rust
- Ruben Salazar
- Curt Storlie
- Scott Vander Wiel
- Joanne Wendelberger
- Geoff Bower
- Scott Brandt
- Andrew Siemion
- Greg Taylor
- Dan Werthimer
- And many others…

With special thanks to Kary Myers for comments on these slides

Two Emerging Computational Challenges

- **Reliability of Computation Itself**
  - Interrupts and Silent Data Corruption
  - Resilience Strategies to Mitigate these Issues
- **Streaming Data**
  - Adapting & Developing Statistical Methods for Streaming Data
  - Leveraging Probability Theory for Fast Computation

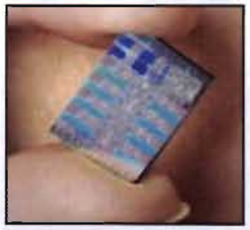
Discuss key research questions in these areas that statisticians can help address

ALMA Image courtesy of NRAO/AUI and Photographer-Kelly Gatlin; Digital composite-Patricia Smiley

Emerging Computational Challenges: Reliability of Computation Itself

- **Large-scale computing is subject to faults which can lead to *job interrupts* and *silent data corruption* (SDC)**
  - "SDC occurs when incorrect data is delivered by a computing system to the user without any error being logged" Cristian Constantinescu (AMD)
  - At scale, rare SDC events can become realities

- ***Resilience strategies* would permit jobs to continue running and produce correct answers despite faults which could cause interrupts and SDC**



References: Michalak et al (2011); DeBardeleben et al (2010); Michalak (2010); Cappello et al (2009); Hong et al (2009); Bairavasundaram et al (2008); Panzer-Steindel (2007); Kola et al (2005); Constantinescu (2000)

SDC Examples

- **CERN File Systems Study (Panzer-Steindel (2007))**
  - Disk Errors: write, read, compare 2 GB file
    - Every 2 hrs for 5 weeks on ~3000 nodes ➔ 500 errors on 100 nodes
  - Recalculate and compare checksum for 33,700 files (~8.7 TB)
    - 22 mismatches ➔ one bad file in 1500

- **LANL Decommissioned HPC Platform Testing (Michalak (2010))**
  - 70 incorrect Linpack calculations; all involve 1 node
  - SDC on two additional platforms

SDC has multiple causes and will likely be more prevalent in new technologies!

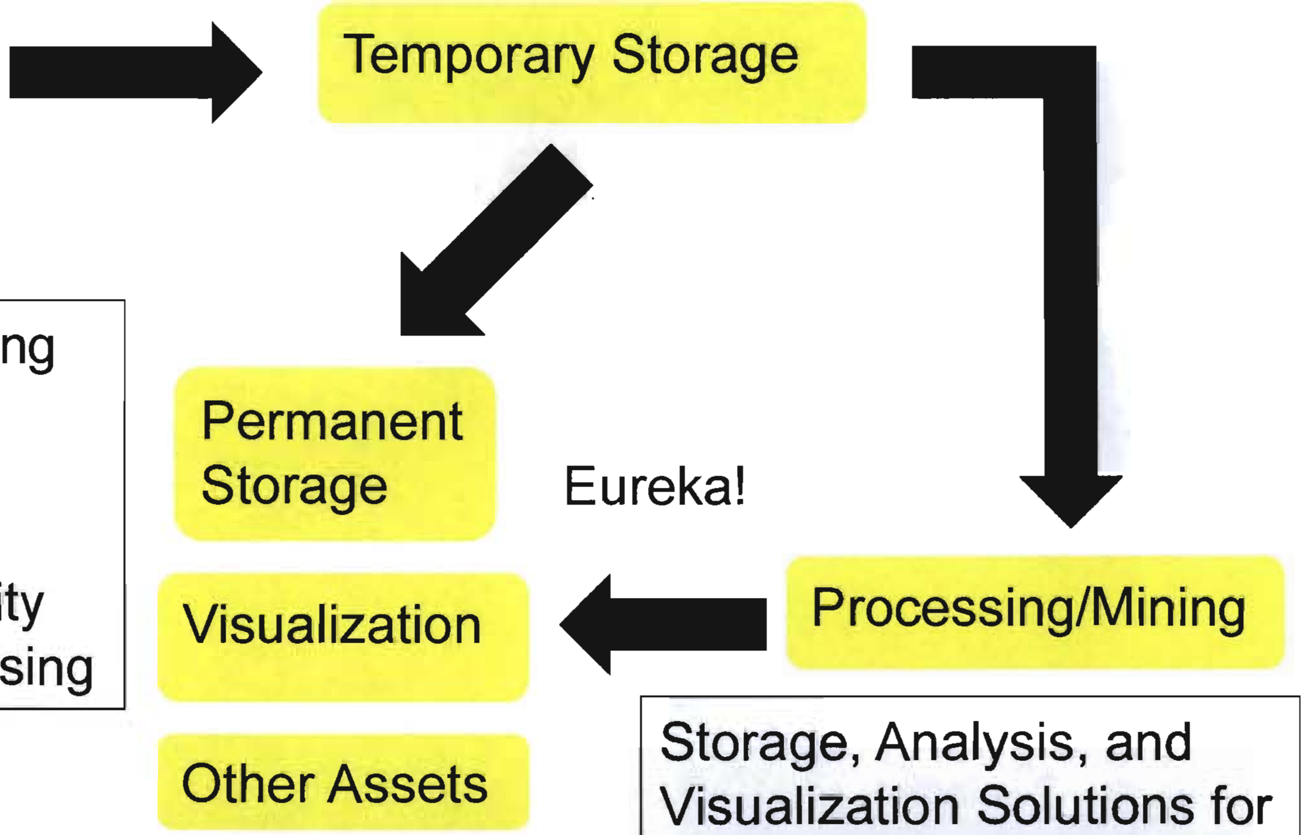
References: Borkar (2009); Constantinescu (2008, 2006); Pan et al (2008)

Open Research Questions in Resilience

- **How can probabilistic computing be leveraged to enable sound results? (Chakrapani et al (2007) & references therein)**
  - Compute using less power via probabilistic switches
    - Power a BIG issue at scale
  - Leverage this strategy for scientific computation
- **Different parts of a computation may need to be calculated with different levels of precision (Feng et al (2010); Li &Yeung (2008, 2006))**
  - What methods can quantify these levels of precision?
  - Can such methods inform the computation so a desired precision is attained?
- **Collaboration with EE/CE/CS researchers & domain scientists required**

Emerging Computational Challenges: Methods for Streaming Data

Example: Real-Time Anomaly Detection: Radioastronomy & Other Fields



Temporary Storage

Permanent Storage

Eureka!

Visualization

Processing/Mining

Other Assets

1) Adapting & Developing Statistical Methods for Streaming Data

2) Leveraging Probability Theory for Fast Processing

Storage, Analysis, and Visualization Solutions for Massive Streaming Data

UNCLASSIFIED

VLA image courtesy of NRAO/AUI

Open Research Questions Related to Streaming Data & Fast Processing

- **Adaptation and development of methods amenable to real-time/near-real-time processing in the massive data regime**
  - Massive is relative to the compute resources available; massive may be small for helicopters and remote locations with limited space and power
- **With variable data flows, some data may need to be dropped**
  - Which data to drop from the processing should be chosen in a sound manner (Babcock et al 2003; Tatbul et al 2003; Chi et al 2003)
- **Need to work jointly with EE/CE/CS & scientific communities to develop sound methods**

Conclusions

- **Statisticians need to collaborate with EE/CE/CS & domain scientists to successfully tackle important research questions**
  - Resilient strategies so computations can use resources efficiently to attain correct answers despite system faults
  - Methods for streaming data
- **Requires statisticians to have basic understanding of key concepts in other disciplines**