Data Science Benchmarking & Performance Measurement

Data Science Symposium
Breakout Session

March 5, 2014

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Measurement Needs for Data Science

- Focus Areas
  - **Scalability, resource utilization , speed, and accuracy** of analytic components, end-to-end systems
  - **Propagation of error and uncertainty** through the system
  - **Visualization, user interfaces, usability**, and systems with **humans in the loop**
- Questions
  - What are the current approaches?
  - What are the solved problems?
  - What are the challenges and gaps?
  - What forms of measurement are needed to:
    - accelerate research?
    - effectively field data analytics systems?

Scalability, Resource Utilization, Speed, and Accuracy

- Scalability
  - Scalability of system vs. scalability of data
    - The effect of scalability on application speed and errors
    - The effect of scalability on accuracy of output of analysis (underexplored area)
  - Performance consistency of evolving datasets
    - Challenge: Repeatable results on evolving datasets
- Repurposing data
  - Challenge: understanding the suitability of data to meet needs of a different problem space
- Measurement focus:
  - Technical metrics (system researcher or developer)
  - Operational metrics (value to the user)

Scalability, Resource Utilization, Speed, and Accuracy

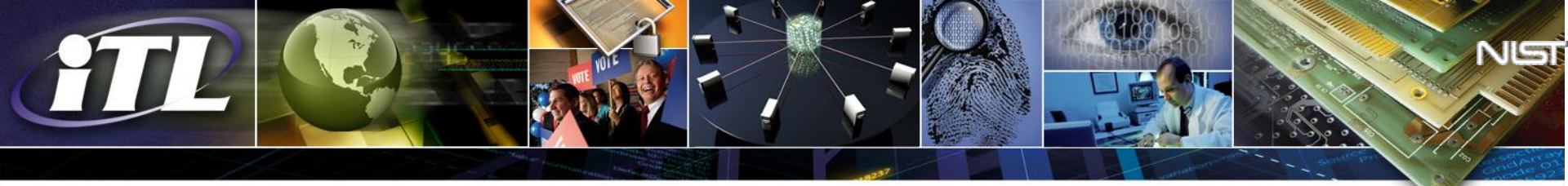
- Accuracy
  - Data accuracy vs. accuracy of system output
  - Challenge: need techniques for validation without ground truth
    - Data validation and system validation
  - Challenge: determining accuracy of data
    - How to measure gaps in data?
- Quality of data
  - Assumption: compensate quality for quantity
    - Does quantity drown out need for quality
- Data characterization
  - Include in data set's metadata?
    - How to incorporate metadata into analytic workflow
  - Provenance (record of data source, context, and transformation)
    - Helps determine future use / allows leverage of data
  - Research Area: automated vs human in the loop

Scalability, Resource Utilization, Speed, and Accuracy

- Data Infrastructure
  - What is needed for a national scale measurement infrastructure for research
  - Cataloging datasets
    - "Consumer Reports" measure of availability, usefulness, and quality
      - (e.g. red, yellow, green)

Propagation of error and uncertainty

- Challenge
  - Heterogeneous data (different modalities)
    - How to measure the holistic error & uncertainty when combining data of different modalities
  - Uncertainty due to human bias
    - Understanding human bias & measuring its impact
  - Research Area: Measuring accuracy & uncertainty of results as a function of time
- Current Approaches
  - Information theoretic measures useful for characterizing propagation of error and uncertainty

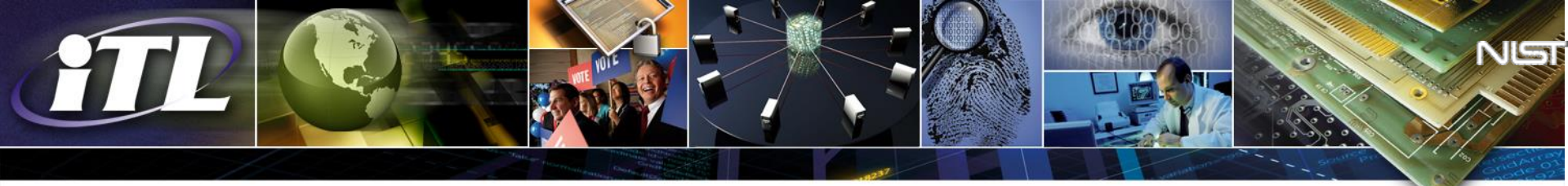


Visualization, user interfaces, usability, and systems with humans in the loop

- Users representative of the end-user population
  - User testing should be done on the appropriate user group
    - Challenge: finding “expert” users with spare time
- User testing participants depends on goal of testing
  - E.g. small scale testing with expert users
  - Challenge: How to structure tests to be more broadly applicable

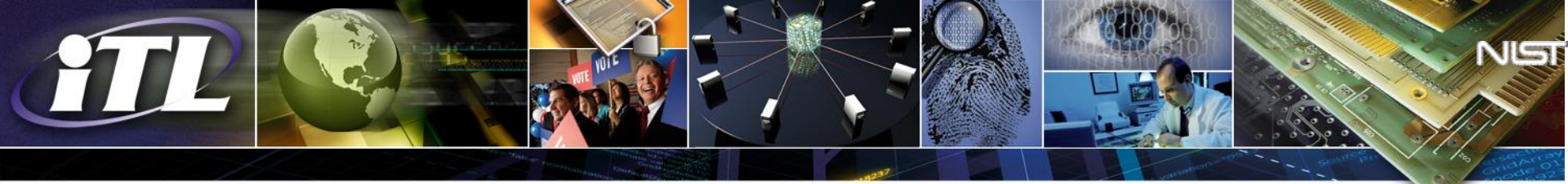
Measurement Methods for Data Science

- Analytic Objectives
  - Challenge: Are there common metrics across analytic objectives?
    - Beneficial for reproducibility
    - Research Area: tradeoff between evolving data and reproducibility of benchmark testing
- Data Representations
  - Challenge: understanding what algorithms can be ported across classes of data



Final Thoughts

- Challenge: Interpreting results from different data sources
  - E.g. Questionnaire vs. physical measurement
- Open question: is speed and scalability easier to quantify than accuracy?
- Characterizing source data was a recurring theme
  - E.g. quality, metadata, provenance, reusability
- Can we learn interesting things from the outcome of the analysis as the data is scaled up?



Thanks!