Data Curation

draft-mathis-ippm-model-based-metrics-00.txt

Matt Mathis
mattmathis@google.com

IETF 85 IPPM
6-Nov-2012

The problem

- We have live data archives (e.g. M-Lab, etc)
  - Always growing
  - Must not delete or alter data
- Ongoing data-mining
  - Inference research on hidden patterns in the data
- On going "recalibration" and data requalification
  - Inference research to improve old data
- Scientific method requires rerunning old experiments
  - With or without new added data
  - With or without new calibration

A scenario

- Data is collected
- Researcher A makes claim about a pattern
- (more data is added)
- Researcher B detects a better way to calibrate the data
- (more data is added)
- Researcher C wants to repeat A's experiment
  - With the same data and algorithm as A
  - With A's algorithms and data recalibrated by B
  - With a differential version applied to B's delta
  - A's algorithm on uncalibrated data to date
  - A's algorithm on B's recalibrated data to date

Basic (technical) approach

- Annotate the data by adding "metadata" columns
  - Subject to the same archival rules as the data
    - Never deleted or modified once applied
  - Want to leverage other datasets
    - e.g. Operational logs
    - Must address (meta)data providence
  - Data holder can't be the only repository
    - Requires a distributed join

Approach to advancing the document

- Current document is really just a problem statement
  - The solution is excessively sketchy
- Seeking additional authors
  - Want somebody with Data Mining background
  - There has to be prior art in other fields
    - E.g. Astronomy, Physics, Economics, Meteorology, Earth Sciences, etc
  - Relatively little specific to IP measurement
    - Except perhaps the granularity of the metadata
- Not ready to be a WG work item