

Data Management for Sensor Data

NSF Spectrum Measurements Workshop
IIT, Chicago, IL

Paul G. Brown
Paradigm4

April 6th, 2016

Why Data Management Matters

- **Data Gathering is Expensive**

- Investment in tools for storage, sharing and analyzing data maximizes ROI
- Effective analysis integrates multiple lines of evidence: geographic, temporal, non-signal data sources
- Sharing data and methods among users with varying analytic questions

General Observations on State of Play

- **Tremendous Variety of DBMS Technologies**

- SQL DBMS, Hadoop, HPC, NoSQL, Array DBMSs ...
- “One size doesn’t fit all”
- General Agreement: SQL DBMSs lay an egg on machine generated data
- More Opinionated: File-systems + programming about as bad

Experience and Direction

- **General Signal Processing Applications**

- General focus on other kinds of machine generated signal data
- Satellite Images, RADAR, Astronomy,
- Emphasis on quantitative analytic functionality: linear algebra, image processing, geo-spatial information management.

- **What To Work On**

- Integration with a variety of client analytic tools: 'R', Python, Matlab, etc
- Server side functionality: GPU based algorithms.
- Integration with File Formats: HDF5
- Scale, Performance, Reliability

What I Hope to Convince You Of ...

- **The Utility of a Benchmark**

- Focus thinking on data management requirements
- Identify eco-system issues: external dependencies, requirements

- **Benchmark Design Questions**

- What Data? (Schema, Size, Throughput)
- Representative Analytic Workflow
- Quality of Service Expectations