Overview of data formats

- What's available vs. what I know about
- Open vs. proprietary… not so obvious
  - What formats/tools exist?
  - What tools work with what formats?
  - What do I want to be able to do with the data?
  - How much data?
  - How much cleaning?
  - Who can help me?
  - What might someone else want to do…?

Structured vs. Unstructured

- Requirements: how well-defined?
- Flexibility: emerging requirements vs. a priori
- Requirements horizon: in-project vs. beyond project
- Structuring too soon vs. cost of restructuring
- Structured: spreadsheet, SQL, HDF(v.5)
  - ENLITEN, APAtSCHE, Colby, ???
- Unstructured: NoSQL (non-relational), linked data
  - IDEAL, DM4T, ???
- Translation: meet changing needs of project phases

Processing

- On-line (~real-time) vs. off-line
- Compatibility with storage platform and format
- License conditions
- What is the question?
  - Matlab: e.g. signal processing, visualization
  - R: e.g. stats (!), visualization
  - Python: e.g. matrices, visualization
  - Spreadsheet: macros, Visual Basic; stats + vis
- Who/What is the consumer?

Legacy access and (re-)processing

- Likely to be least considered: SEP!
- Improve prognosis with open, unstructured formats?
- How to make data accessible?
  - Accessible for who/what?
  - Structural description
    - csv is csv, SQL has tables, triples are triples…
  - Semantic description
    - text labels vs. semantic labels… ontologies
- Generating vs. authoring metadata