Towards a Big Data Taxonomy: Batch Analytics Process

Bill Mandrick, PhD

Data Tactics

Version 16_September_2013

Scientific Taxonomies Represent

- Types of Processes
- Types of Objects
  - Physical Objects
  - Information Artifacts
- Types of Characteristics
  - Qualities
  - Roles
- Relationships
  - Between Processes
  - Between Objects
  - Between Characteristics

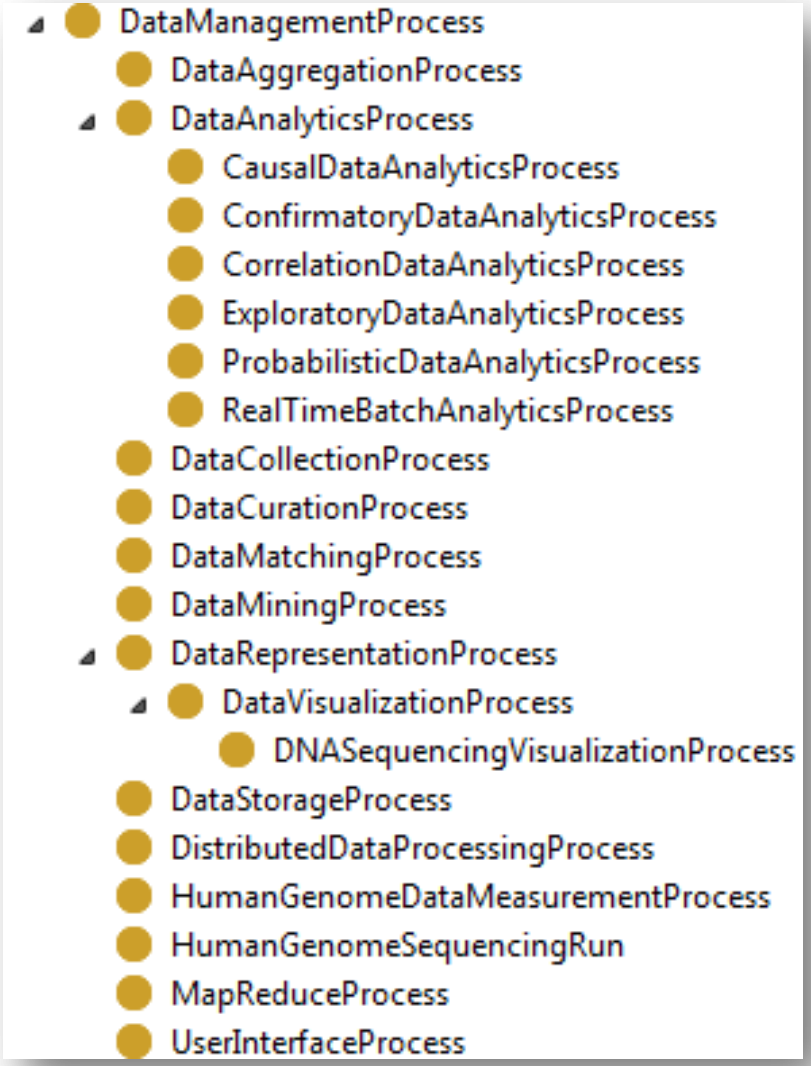
Relations Between Processes

- Processes A <relation> Processes B
  - Complex Process <has part> Sub-Process
  - Sub-Process <part of> Complex Process
  - Process A <precedes> Process B
  - Process A <follows> Process B

Examples:

Data Curation Process <has part> Data Selection Process
<has part> Data Collection Process
<has part> Data Archiving Process

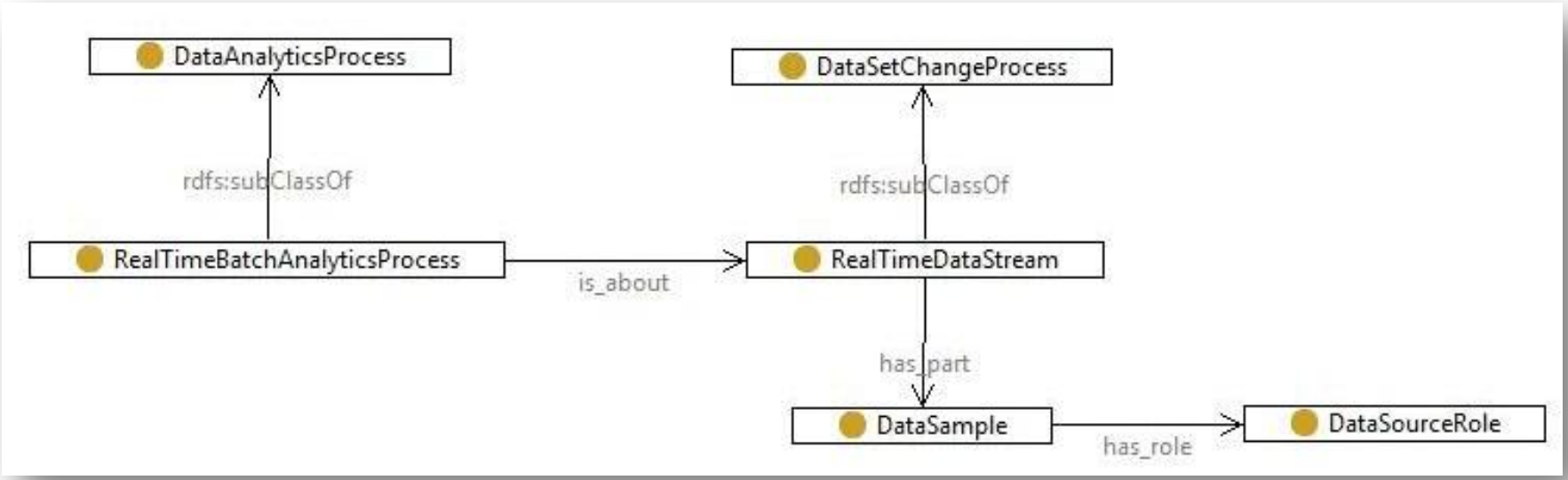
Data Management Processes



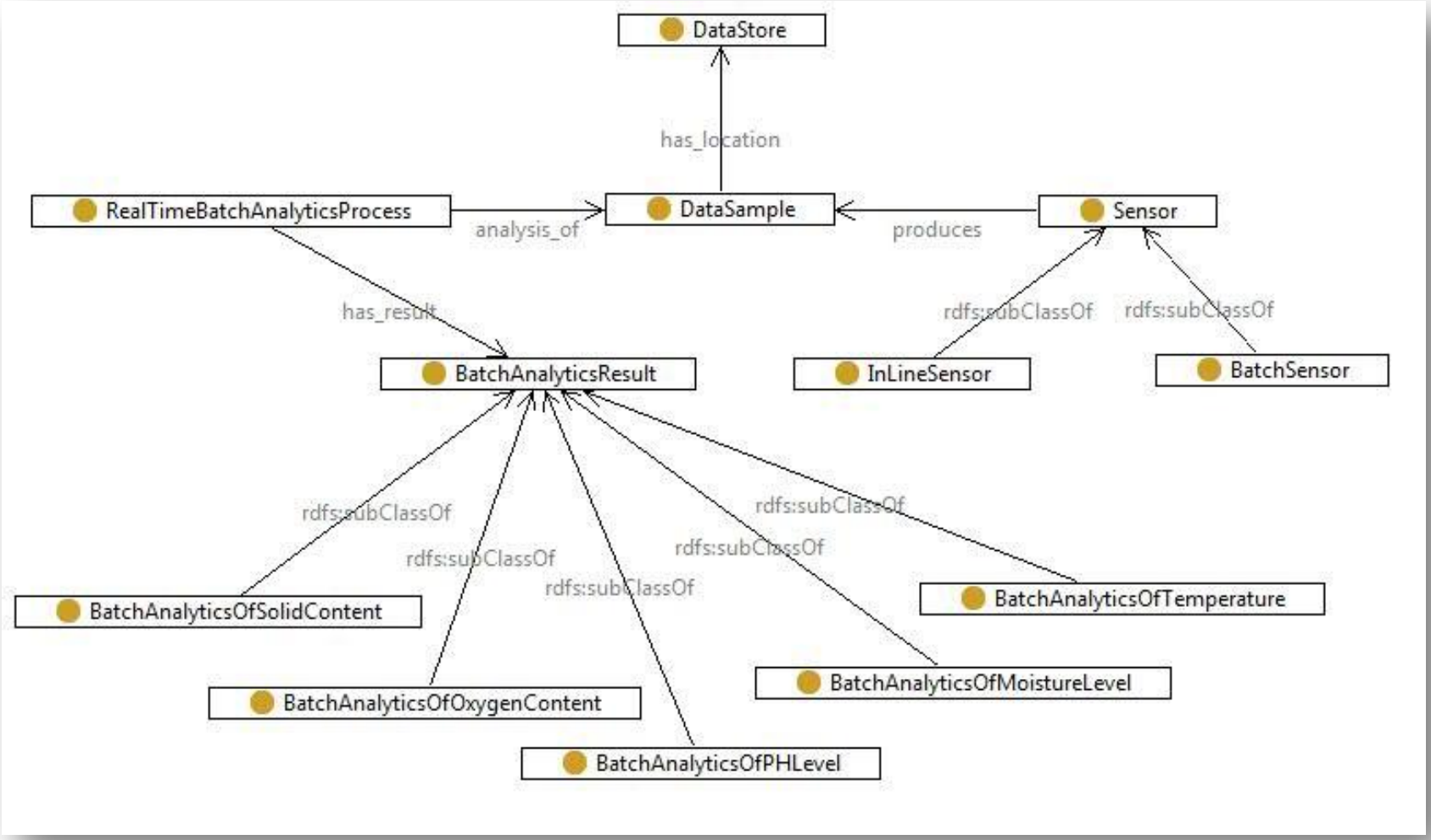
Data Management Processes can be decomposed and related to other (sub)processes

…as well as to their outputs (Information Artifacts).

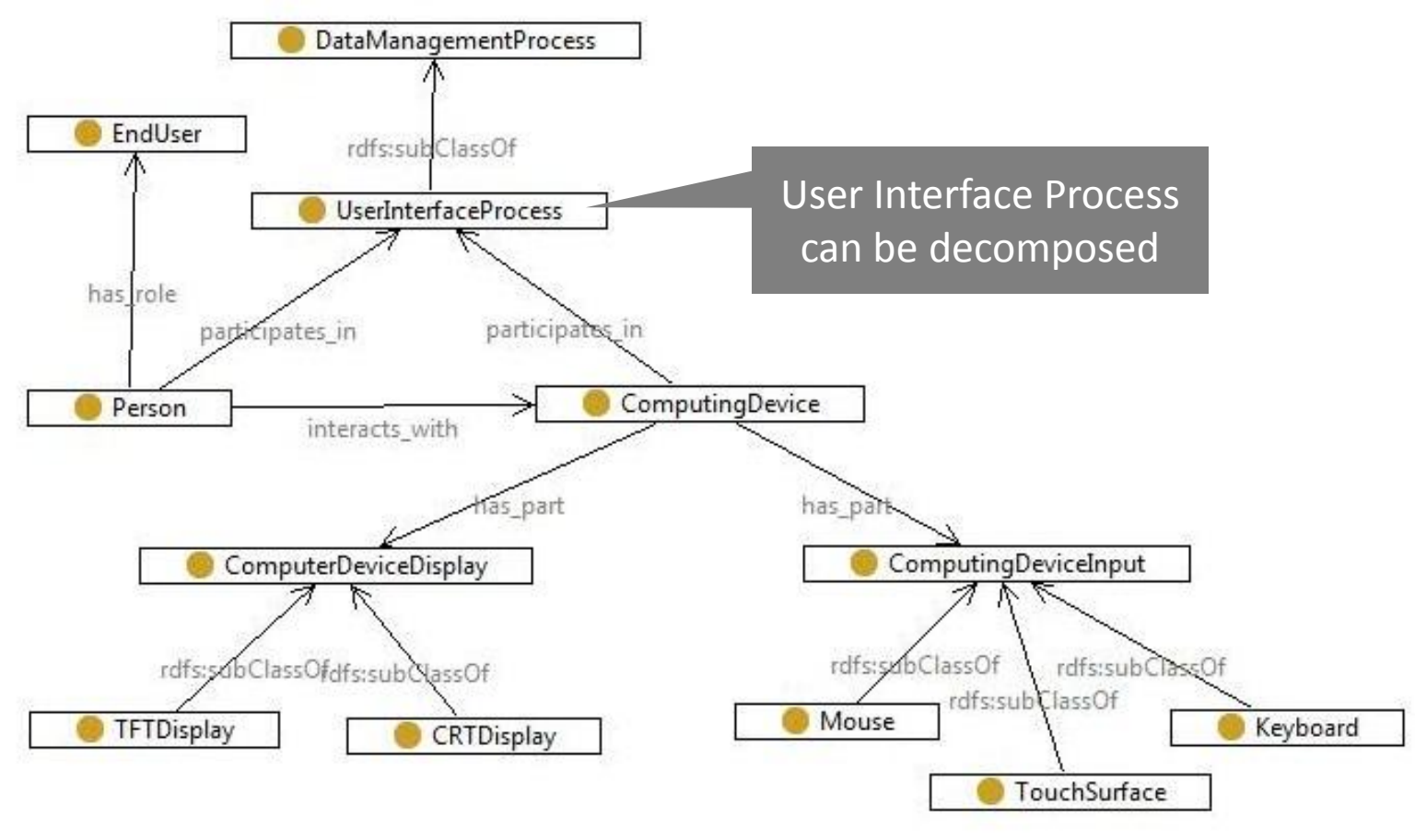
Real Time Batch Analytics Process



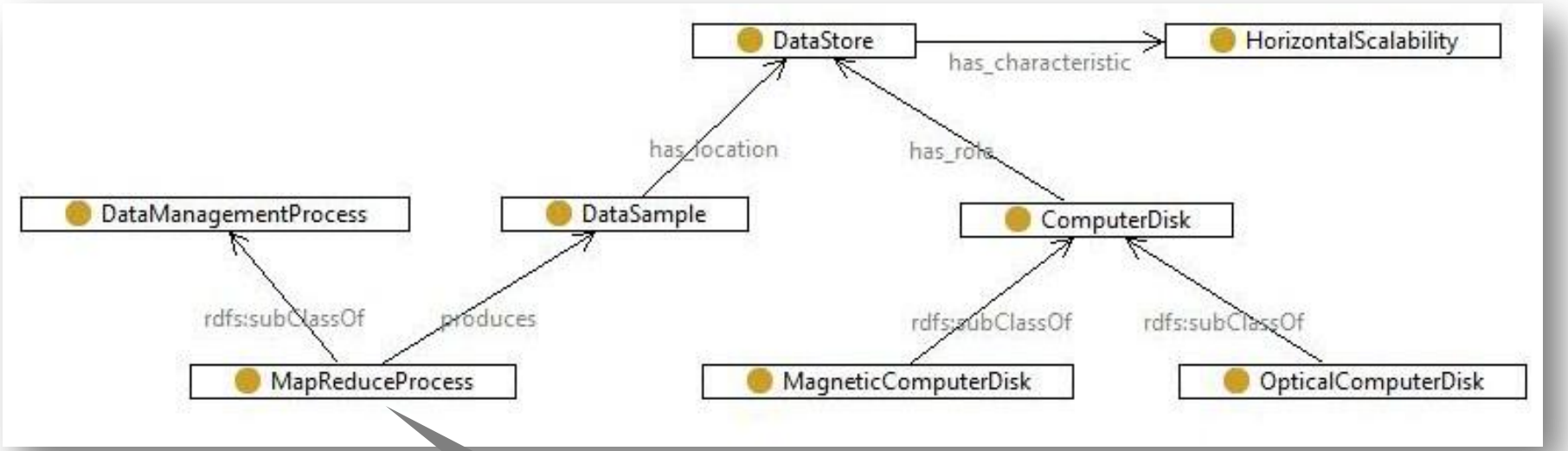
Real Time Batch Analytics Process



User Interface Process



Map Reduce Process



Map Reduce Process can be further decomposed into its constituent parts

Conclusion

- This method can be done for any part of the Big Data Taxonomy
- Need SME input for various areas/domains
- Need to add definitions in owl
- Need to expand set of standardized relations
- Link ***instances*** to the taxonomy (e.g. actual data sets, batch analytics data samples, etc.)