Datasets and Use Cases for Data Science Research

Distilled Discussion Notes

Patrick Grother, Marion Le Bras, Brian Antonishek

Domains

Big, in the sense of people

- Epidemiology
  - Detection of disease, discovery of cause (Example: localized disease)
- Health records
  - Mining
  - Discovery
  - Statistical models
- Genomics, proteomics, traits
- Geospatial (standards exist)

Big, in the sense of petabytes

- Terabyte images
  - Near Earth Object detection
  - Brain Initiative

Big, in the sense of "growing" or "ongoing"

- Longitudinal tracking
  - Biomarkers as predictors of disease

NIST might usefully…

- Leverage its application agnosticism which is a desirable property (to industry, USG)
- Support abstraction of the generic use-case-defining properties from specific examples
  - c.f. Design patterns
- Construct best-practices (or standards) for
  - Dissemination of data
  - Collection + construction
  - IP documentation, approval
  - PII
  - Anonymisation
  - Vocabulary around Big Data

- Support experimenter's initial question: "Will this data be useful to address my problem"
  - A semantic manifest
  - Triage – quick look
- Support publication of a "story" i.e. vignette examples of what a dataset can do.
- Support visualization of data
- Support re-purposing of data
- Support collaboration
  - Wiki, tools, working groups
- Support ontology development

Best Practices (or standards?)

For a new Big Data set

- Documentation
  - Traditional + “story” (what can be done with data) + schemas
- Dissemination
- Collection + Preparation
  - Binding raw + meta data
- Subset definitions (for tractable experimentation)

For publishers, curators

- Sustainability
- Extensibility
  - e.g. Cancer case files + molecular diagnostics or genomics, or proteomics
  - Merge with other (similar) datasets
- Affording discovery
- APIs

NIST should help prevent

- Balkanization
  - A dataset is too large, so gets sliced in varying, non-portable, poorly documented ways
- Poor quality datasets
  - Incomplete
  - Erroneous (meta)data
  - Unevenly sampled

- Scientific fraud
  - Traceability
  - Reproducibility
  - Integrity
- Manipulation of policy
- Duplication of records
  - Identical data received by end-users via two suppliers

Value = F(Data, Analysis)

Data → People

- (Free, wide) dissemination of data
- Open source data
  - Unanticipated benefits
  - Supports R&D in trade and industry
- Static vs. Streaming
  - Value in offline (analysis)
  - Values in real-time
- Synthetic data
  - Less value

People → Data

- Apply analyst's software to sequestered data
- Sequestered data
  - Exists
  - Is rightly sequestered
  - Is often valuable
    - For mining, analysis
    - For evaluation of software capability

Organizations + Standards

- Research Data Alliance
  - https://rd-alliance.org/
- Federal Big Data WG
  - http://www.nitrd.gov/ (?)
- Data.gov
- Open Science Data Cloud (U. Chicago + others)
  - https://www.opensciencedatacloud.org/publicdata/
- PCAST
- NIEM www.niem.gov
  - An unbrella standard for information exchange
  - Add new domains
- Linguistic data consortium
- CRISP-DM
  - Cross Industry Standard Process for Data Mining

One citation