Data Scientist Story for Earth Science Collaboratory

Chris Lynnes

What the _____ are "Data Scientists"?

- Experts in techniques that can be applied to managing or analyzing data, e.g.,
  - Semantic web
  - Data mining
  - Statistics
- Work two sides of the street
  - Using data
  - Providing data
- Ranks are likely to grow
  - University programs in data science
  - "Big Data" hype
  - Not much data bigger than Earth Sciences
    - Also big, complex problems

Key Personae Attributes

- Knowledge about the Earth Science domain can vary considerably from little to broad knowledge
- But highly skilled in a field outside Earth science: math | statistics | machine learning | AI | semantics | informatics etc.
  - May possess awesome programming skills
- Key Strengths include:
  - Adaptability
  - Communications
  - Understanding of methods
  - Technical breadth

Key Needs (User Side)

- Access to Earth science domain knowledge
- Data that are:
  - Clean = spurious, unreliable values removed (or clearly marked)
  - Tidy = well formatted and structured
- Benchmark datasets
  - canonical community datasets for which the “solution” is well known
  - e.g., Land cover classification, DeFries and Townshend (1994)
- Problem datasets
  - Datasets where the solution is not well known or are controversial

Jay the Data Scientist

- Jay works the User Side of the Street
- He is an expert in Video Image Analysis
  - Has developed a new technique for measuring ground level haze using traffic cameras
- Next up: develop a method of combining traffic camera image analysis with satellite AOD to estimate ground-level PM 2.5* in urban areas
- (Well, it *could* happen. In theory.)
- Primary goal of the project is to demonstrate the technique
  - Hopeful that aerosol domain scientists may be interested...

*PM 2.5: Fine particulate matter, less than 2.5 microns

Step 1: What’s Out There?

- Reliable ground station PM 2.5?
  - Search for data with PM 2.5 and information (community sentiment?) on reliability
  - e.g. “AirThen”, quality-controlled archive of AirNow
    - Comments assess reliability, articles describe reliability
    - Perhaps a “Community Benchmark Data Set” designation?
- Clean satellite AOD
  - Artifacts in data clearly marked, with pictures and workflow examples and / or related articles

Step 2 – Process data (safely)

- Jay chooses from reprojection or colocation techniques for Satellite AOD relative to traffic locations and PM2.5 measures
  - Uses community comments on pro and cons of competing techniques
- Posts results with annotations / questions to StackExchange for ESC
  - “What the ___ is *this* spike?”

Step 3: Develop and Test Method

- Add traffic data to ESC
- Add traffic cam analysis algorithm to ESC
  - Image analysis of traffic cams for visibility and traffic flow + colocated satellite AOD
- Construct compound workflow processing satellite data + traffic cam analysis
- Generate results of predicted PM 2.5 with measured (in AirThen)

Step 4: Seek Community Feedback

- Publish (make visible) traffic cam data + processed satellite data + algorithm + workflow + results
- Aerosol scientists check out satellite data results, pointing out artifacts, known issues etc.
- Jay revises, republishes, etc.

Step 5: Publish

- Jay publishes article
  - Acknowledgments section generated automatically:
    - users that answered questions, provided comments that Jay “liked”
    - colocation algorithm author
  - Data citations generated for community datasets used
- Links to algorithms, data, workflows, results, community comments