**Teaching Science Data Analytics Skills, and the Earth Science Data Scientist**

Submitted by stevek on Mon, 2015-04-27 12:39  Tuesday, July 14, 2015 - 08:30 to 10:00

**Event:** Summer Meeting 2015

**Session Type:** Breakout

**Expertise Level:** Beginner

**Collaboration Area:** Climate Education Working Group

**Education**

**Student Cluster**

**Abstract/Agenda:**

Scientists explore heterogeneous data analysis methodologies to attain information and knowledge. To readily accomplish this and maximize cross-dataset integration and usability, science data analytics techniques need to be advanced and well understood. Data Scientists, specifically science data analysts requiring skills to better understand ambiguous relationships across various datasets, are becoming increasingly appreciated and significant given the expanding amount of available heterogeneous data.

Through presentations given by subject experts, this session provides an opportunity to better understand challenges faced by domain-oriented data scientist and the data analytics skills and the expertise that students must learn to be able to move into high-demand data scientist (data analytics) positions. Session topics are experience oriented and include expertise that is useful, if not required, to solve science data analysis problems. This session also shares and discusses tools and techniques employed to address these needs.

**Notes:**

The goal of this session is to discuss and extract real project data scientist/analytics experience needs, initiated by presentation and discussed by session participants. Of special interest is bringing together people who have needs for data scientists (data analytics) and will be able to articulate those needs by the end of the session...

...and/or; stir ideas: for the use of data analytics in their research or to build tools/services for others.

We have 3 speakers, followed by discussion:

- Peter Fox, Earth & Environmental Sciences, Tetherless World Constellation, Rensselaer Polytechnic Institute
- Wade Bishop, School of Information Sciences, University of Tennessee
- Karen Stocks, Director, Geological Data Center, Scripps Institution of Oceanography

28 ESIP members participated in this session possessing expertise/interests in Science (~5 participants), Data Management (~8), Engineering (~9), Data/Information Science (~4), being a future Data Scientist (~2). Imagine: Last year we may have had 1 or 2 Data Scientist participate, and no one admitting their desire to be a Data Scientist.

As you will see below, we were very fortunate to have three very knowledgeable speakers who provided valuable insights from different points of view. As we discuss this subject, and bring people in to describe their experiences, we are starting to see particular themes begin to surface.

**Data Scientists Are Freaks of Nature but Products of Nature - Peter Fox (RPI)**

- Data Scientists is a new term though data scientists have existed for years
• Peters initial interest (1991) was solar radiation
  ○ Shared a satellite record of total solar irradiance that plotted several instrument on
    the same graphs, that didn't calibrate well
  ○ The project - due to lack of proper tools- took 10 years not just 3 as originally planned
  ○ Results took 724 GB (massive for their time)

• Scientists should be able to access a global distributed knowledge base of scientific data
  ○ But, complications still exists because there is data obtained from multiple sources,
    various protocols, and more
  ○ And data is created in a manner to facilitate its generation NOT its use.

• Data pipelines: we have problems
  ○ Data is coming in faster, in greater volumes and forms and outstripping our ability to
    perform adequate quality control
  ○ Data is being used in new ways and we frequently do not have sufficient information
    on what happened to the data along the processing stages to determine if it is
    suitable for a use we did not envision
  ○ We often fail to capture, represent and propagate manually generated information
    that need to go with the data flows
  ○ Each time we develop a new instrument, we develop a new data ingest procedure
    and collect different metadata and organize it differently. It is then hard to use with
    previous projects
  ○ The task of event determination and feature classification is onerous and we don't do
    it until after we get the data
  ○ And now much of the data is on the Internet/Web (good or bad?)

• A Metaphor for Data Science and Data Analytics:
  ○ Anatomy study of the structure and relationship between body parts
  ○ In Data Science – Anatomy:
    ▪ Technical tools and standards
    ▪ Forms of Analysis, Errors and Uncertainty
    ▪ Data Management and Products
    ▪ Data Life Cycle – Acquisition, Curation and Preservation
  ○ In Data Analytics – Anatomy:
    ▪ Intermediate Skill in parametric and non-parametric statistics
    ▪ Application of a broad spectrum of Data Mining and Machine Learning
      Algorithms
    ▪ Ability to cross-validate and optimize models
    ▪ Application to specific datasets
  ○ Physiology is the study of the function of body parts and the body as a whole.
  ○ In Data Science - Physiology
    ▪ Definition of Science Hypotheses, Guiding Questions
    ▪ Finding and Integrating Datasets
    ▪ Presenting Analyses and Viz.
    ▪ Presenting Conclusions
  ○ In Data Analytics - Physiology (in a group)
    ▪ Definition of Science Hypotheses, with Prediction/ Prescription Goal
    ▪ Cleaning and Preparing Datasets
    ▪ Validating and Verifying Models
    ▪ Presenting Ideas and Results

• For Data Science, ‘math and statistics knowledge’ are essential as a basis for exercising
  ‘Hacking skills’ and ‘Substantive Expertise’

• Managing data should be second nature and as a whole data scientists need to work
  collaboratively

  * Data Scientists should be interdisciplinary from the start (scientists/researchers should be
    data people. Data Science should be taught across all applicable curriculums. Working with
    data becomes second nature.

  * Data Science and Data Analytics – Call to Action: please see Peter’s presentation

  * Data science is the foundation for data analytics. It is not possible to do good analytics
    without good data management!
Developing a Curriculum for their Earth Science Data Science - Wade Bishop (University of TN)

- A DACUM is a job analysis process at a lower cost and time commitment compared to direct observation.
- A DACUM is based on three core principles.
  - Job incumbents know their job better than anyone else! They are currently working in the field and are not necessarily leaders or educators in the field!
  - The best way to define a job is by describing the specific tasks that are performed on the job. Professionals who are actually performing the jobs currently should be best able to clearly explain what those tasks are in terms of task statements.
  - All tasks performed on a job require the use of knowledge, skills and abilities (KSA) that enable successful performance of those tasks.
- Wade is proposing to start a DACUM with ESIP to address what should be taught to promote Data and Information scientists

Educating Data Scientists: a view from the trenches - Karen Stocks (Geological Data Center Scripps Institution of Oceanography)

- We all have a different perspective from academia to government etc.
- People need to have a deep understanding of the instruments and the data that they collect
  - Provided an example of a danger zone where a name of a fish changed leading to an erroneous assessment that a fish population was completely replaced by another population.
- People needs to be a user needs to be good data management - need good data curation principles
- Programs need to introduct best practices for data collection through archive- draw from computer science, statistics, etc., so there is domain knowledge.
  - Need to write code, data handlers
  - Need to understand data, instruments
  - Need to know user needs
  - Need to manage data
- Data analysts should not just be data analysts but they should be domain data analysts (e.g. Geology data analysts, atmospheric data analysts)
- Important to know the importance of the data lifecycle from data collection through analysis all the way to archival - this requires cross domain training
- Information science or data curation must be complemented by background/understanding of what the data says and how to understand the data
- For data scientists to work with domain scientists this can work but it could take a decade and this is not conductive to the NSF 3 year funding cycle
- Skills: Domain expertise, on line library services
  - Library and information science – getting a little long in the tooth

Re-occurring Themes:

- Data is created in a manner to facilitate its generation NOT its use... thus data scientists, via data analytics need to facilitate making data more useful
- Data Scientists should be interdisciplinary from the start.
- Learn your math and statistics.
- Important to know the importance of the data lifecycle from data collection through analysis all the way to archival - this requires cross domain training
Information science or data curation must be complemented by background/understanding of what the data says and how to understand the data.

Recommendation for building better data scientists is for students to do an internship and do it early.

Presenters:
Name: Peter Fox [16]
Organization(s): TWC [17]
Email: pfox@cs.rpi.edu [18]

Name: Wade Bishop [19]
Organization(s): University of Tennessee [20]

Name: Karen Stocks [21]
Organization(s): Geological Data Center [22], Scripps Institution of Oceanography [23]

Notes takers:
Name: Tiffany Mathews [13]
Organization(s): Booz Allen Hamilton [14]
Email: tiffany.j.mathews@nasa.gov [15]

Name: Steve Kempler [11]
Organization(s): Goddard DAAC [12]

Participants:
Victor Zlotnicki, Wade Bishop, Sean Barberie, Nga Quach, Minnie Wong, Peter Fox, Mike Gangl, Arika Virapongse, Emily Northup, Dave Gallaher, Jane Wyngaard, Karen Stocks, Dan pilone, Andrew Pawloski, Jason Gilman, Heather Brown, Dru Clark, Chris Olson, LuAnn Dahlman, Liping Di, Evan Burgess, Charles Thompson, James Ryan, Christian Reilly, Ward Fleri, Chritina DeCesare, Denise Hills

Creative Common License: Creative Commons Attribution 3.0 License

Teaser: What skills do you want your next Data Scientist to have?

Keywords: Data Analytics [24]
Data Scientist [25]
Student [26]
Skills needed [27]

Source URL: http://commons.esipfed.org/node/7999

Links: