Earth Science Data Analytics: Preparing for Extracting Knowledge from Information

Steve Kempler[1], Lindsay Barbieri[2]
[1] NASA Goddard Space Flight Center, [2] The University of Vermont

Earth Science Data Analytics/Science Skills Needed: Overall Experiences /Operational Needs

Data Analytics / Data Science

- **Data scientist** studies methods of analyzing data, ways of storing it, and ways of presenting it
- **Data analytics** is performed by the practitioners who applies tools and techniques to co-analyze data.
- **Both**, data science and data analytics **require very similar skill sets**.
- Once acquired, it becomes up to the individual to **determine how best to use these skills, based on their interest and aptitude**

General Experiences

- Need **skills in:** mathematics, numerical modeling, statistics, software engineering and the ability to integrate data across multiple domains.
- Need **expertise in tools and techniques**: rule learning, classification, cluster analysis, data fusion, machine learning, neural networks, anomaly detection, modeling, time series analysis, visualization
- Need **knowledge in particular science domains** where data analytics can advance our understanding of science
- **The role is a hybrid one**… skills to support domain scientists with data and computational needs to communicate across domains.

Operational Needs

- Need to facilitate **making data more useful**
- Should **be interdisciplinary** from the start.
- Learn your **math and statistics**.
- Know the importance of the **data lifecycle**
- Understand **what the data says** and how to understand the data
- **Know the territory**: What information is available, Where to get it, How it is generated, How to use it, How it can be used
- Understand **data, metadata, and data integration**
- **Know how to apply the techniques to the discipline**
- Learn through **internships**

Excerpts from: Mobasher, Woodstock, Evans, Fox, Stocks (ref. provided upon request)

A Student's Journey into Earth Science Data & Information

Natural Resources

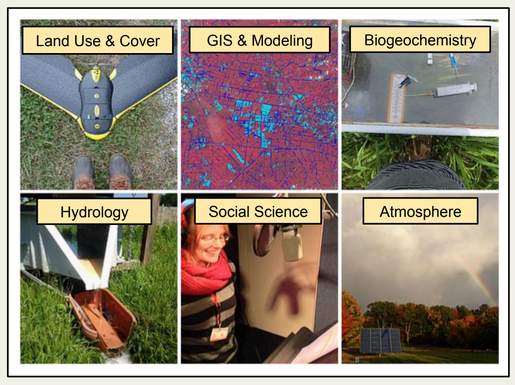
"…understand, **nurture, and enrich** the **interdependence** of **people** with healthy **ecological systems**"

Economics

"…**integrate natural & social sciences** to **understand** the **interactions** between **people and nature** and to help build a **sustainable future**"

Complex Systems

"…**complex systems** {physical, biological, social, engineered} may be **studied, modeled and understood** using similar approaches, **regardless of the application domain**"



Earth Science = lots of interdisciplinary, heterogenous data!

- Scientists spend 30% of time programming, but 90% are self-taught
- Unlike laboratory and field equipment, software is often not carefully validated
- Computing errors can have disproportionate impacts on scientific process

But there are fairly easy things to implement!

my perspective on what Earth Science Needs (from Data Science)

- Repeated **Exposure**
- Sharing of **Vocabulary**
- How, Where, When and Who to Find/Ask for **Resources & Help**
- Sharing of **Community**: "both ways" **Communication**

Summary Table of Best Practices

from **Data Science, Dr. James Bagrow**

- Write programs for people, not computers
- Naming conventions: descriptive non confusing variables
- Make code style and formatting consistent
  - see style guides (Python - PEP8)
- Let the computer do the work
  - Repete computational tasks: make computer do it
  - Save recent commands for re-use in a text file
    - Ex: Shell Scripts
  - Use a tool to automate workflow
- Reproducibility
  - provenance of data, track chain of custody of data
- Agile Development
  - Work in small steps,frequent feedback and corrections
- Keep track of changes: version control (git, dropbox)
- DRY principle: Don't Repeat Yourself
  - Modularize, don't copy and paste
  - every piece of data has a SINGLE authoritative representation (but not unique - "one is none, two is one")
- Optimize software only after it works correctly
  - first focus on what it's supposed to do
- Code commenting
  - interfaces and reasons for the code, NOT what the code is doing.
- Embed documentation for software in software
  - docstrings - have inside the code itself
- Collaborating
  - code reviews, git repository, pair programming: two people writing together

What the Universities Offer (July, 2016 study and comparison with 2013 Study)

Question:

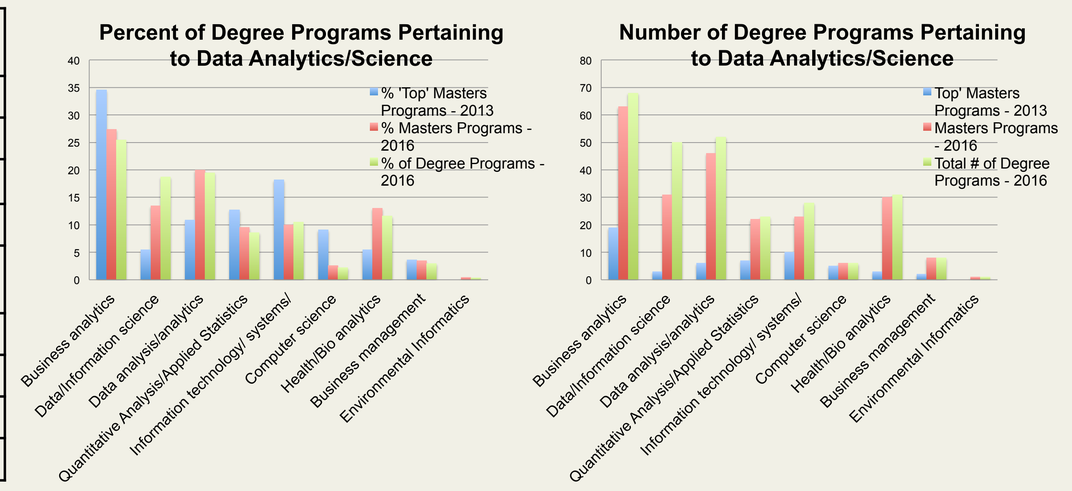
What do university level Data Analytics/Science Programs focus on?

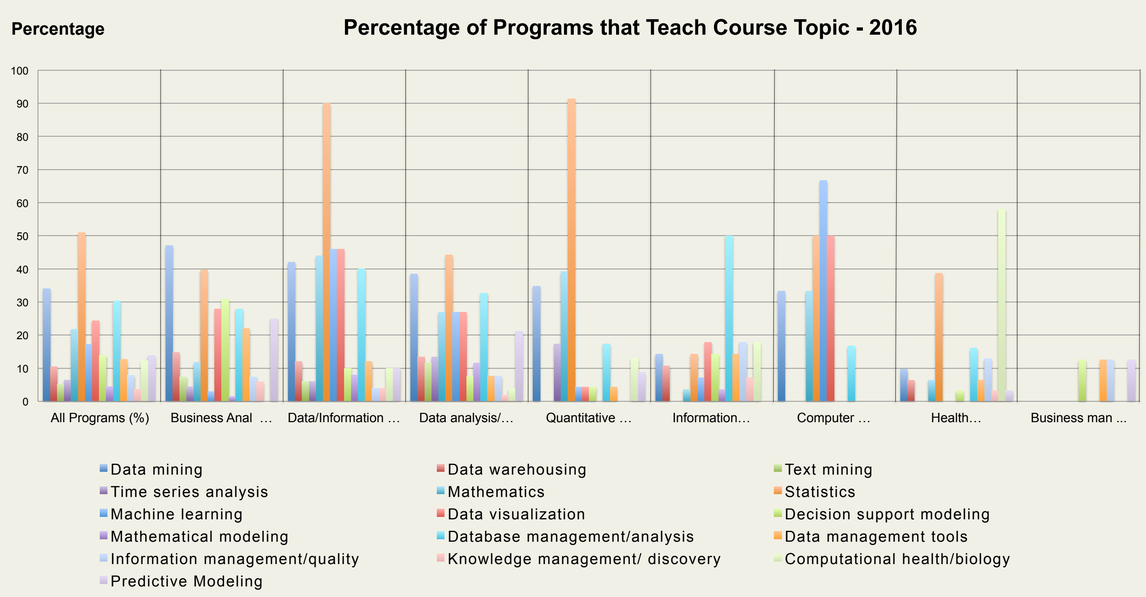
Methodology:

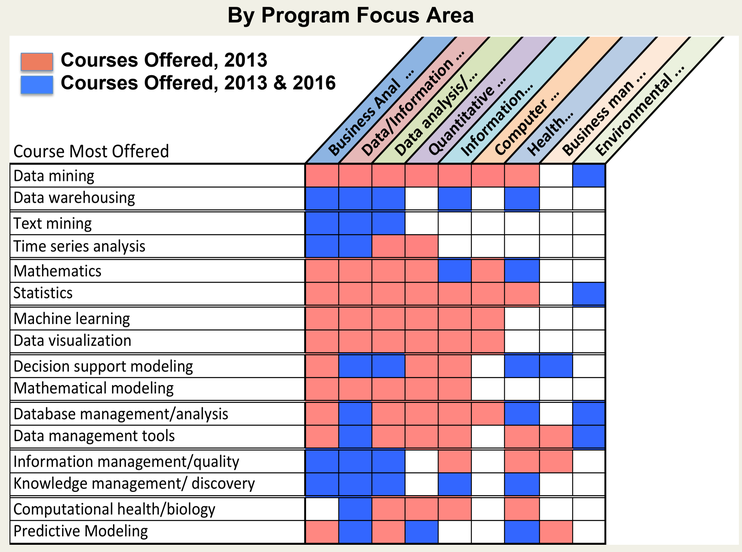
1. Surveyed the curriculums of 167 universities offering 267 'Data Analytics' or Data Science' degree programs
2. Categorized program based on focus area, by degree and on-line/on-campus
3. Surveyed all curriculums to determine course topics providing specific training, relevant to Earth science
4. Charted course topics against program focus area to see what is emphasized per program
5. Compare changes to 2013 survey
6. Compare to 'real world' needs

Broad coarses not included:
Data Science
Data Analytics
Informatiom Technology

Program Focus Areas	B on-line	B campus	M on-line	M campus	PhD on-line	PhD campus
Business analytics		5	22	41		
Data/Information science		15	10	21		4
Data analysis/analytics	1	4	12	34		1
Quantitative Analysis/Applied Statistics			5	17		1
Information technology/ systems/ management			4	19		5
Computer science			1	5		
Health/Bio analytics			15	15		1
Business management			3	5		
Environmental Informatics			1			







References:
- http://datascience.community/colleges
- http://101.datascience.community/2012/04/09/colleges-with-data-science-degrees/
- http://www.mastersindatascience.org/schools/

What Else Universities Should Consider Offering

What Universities are Offering (Findings)

In 3 years, the percentage of degree programs that claim to pertain to Data Analytics/Data Science:

1. Decreased for Business programs.
2. Increased in general Data Analytics/Science programs
3. Decreased in Information technology/systems and Computer Science programs
4. AND… a program called 'Environmental Informatics' shows up

- Some programs are interdisciplinary with other departments
- A few universities listed below, are very in tune with professional needs
- Many programs are introductory, offering 'generic' courses, e.g., 'Introduction to Data Science'
- Some really good Bachelor degree programs have appeared
- PhD programs are obviously more research than course work in nature.

In 3 years, programs have increased in number, some more interdisciplinary and specialized, and much more refined in providing a broader range of relevant courses.

STANDOUT DATA ANALYTICS/SCIENCE/INTERDISCIPLANARY w/ SCIENCE PROGRAMS (2016, no particular order)		
University of Michigan	Regis University	University of Arkansas Little Rock
University of San Francisco	University of St. Thomas, Minnesota	University of Oklahoma Norman
Univeristy of Cal- Irvine	North Carolina State, Raleigh	(Data Science)
University of Rochester (data science)	Columbia University (machine learning)	Tennessee Technological University
Georgetown University	Cornell (Applied Stats, Opt.2)	(Environmental Informatics)
Georgia State Univeristy	Rensselaer Polytechnic Institute	George Mason University
Indiana University Bloomington	St. John's University-New York	

Program Pertaining to Data Science/Data Analytics: Course Topics Most Offered

Blue techniques, skills practiced are 'taught'

Overall:
- Statistics, Data Mining, Database Management/ Analysis

Data Science, Data Analytics, and Computer Science:
- Data Mining, Mathematics, Statistics, Machine Learning, Data Visualization

Data Science, Data Analytics, Information Systems:
- Database Management/Analysis

Quantitative Analysis:
- Data Mining, Mathematics, Statistics

Other Relevant Courses Offered:
- Programming, Neural Networks, Data Analysis, Artificial Intelligence, Clustering, Time Series, Data Warehousing, Pattern Recognition, GIS, Remote Sensing, Text Mining, Information/Knowledge Management

Data Analytics/Data Science Techniques Practiced

- Anomaly Detection
- Artificial Intelligence
- Classification
- Cluster analysis
- Data Compression
- Data Engineering
- Data Fusion
- Data Mining
- Data Warehousing
- Database Management
- Machine Learning
- Mathematics
- Modeling
- Neural networks
- Pattern Recognition
- Rule learning
- Signal Processing
- Statistics
- Time series
- Visualization

Skills Practiced

- Ability to integrate data across multiple domains
- Support domain scientists with data and computational needs to communicate across domains (be interdisciplinary)
- Knowledge of data life cycle
- Software engineering - Programming

Every Earth science program should contain training in Data analytics/science and Programming (Fox, others)