Earth Science Data Analytics (ESDA)  
Telecon XX  (that would be 20)  
Earth Science Data Analytics Cluster  

Steve Kempler, Moderator  
February 18, 2016
Agenda

- In The Beginning
- After the Beginning
- A New Beginning - ESIP
- After the New Beginning – ESDA
- The Beginning of ESDA Cluster
- After the Beginning of ESDA Cluster
- Beginning to Better Understand ESDA
Earth Science Data Analytics (ESDA) Cluster Goal:

To understand where, when, and how ESDA is used in science and applications research through speakers and use cases, and determine what Federation Partners can do to further advance technical solutions that address ESDA needs. Then do it.

**Ultimate Goal:**

To Glean Knowledge about Earth from All Available Data and Information
"You just stood there and let the glacier run over your foot?"
In The Beginning
In The Beginning
In The Beginning
In The Beginning
In The Beginning

Serpent River
WEATHER STATION

IF THE ROCK IS WET...  It's Raining
IF THE ROCK IS SWAYING... It's Windy
IF THE ROCK IS HOT...  It's Sunny
IF THE ROCK IS COOL...  It's Overcast
IF THE ROCK IS WHITE... It's Snowing
IF THE ROCK IS BLUE...  It's Cold
IF THE ROCK IS GONE...  TORNADO
In The Beginning

Hmm, the weather rock says rain with a chance of flooding...

Whoa! These things are good!
After The Beginning

Science Data Collecting…
A New Beginning - ESIP

ESIP, ushering in an era of:

- Increasingly advancing information technologies
- Attracting the Best of the Best people in creating innovative solutions to preserve Earth science data and serve Earth science research, applications and education
- Directly responding to the data access and usage needs of Earth science data users…
- … Such as the increasing interest in exploiting all available information in new ways, singularly, but more importantly, in combination with other information
ESIP Mission, Vision & Values (let’s be reminded)

**Mission:** To support the networking and data dissemination needs of our members and the global community by linking the functional sectors of observation, research, application, education and ultimate use of Earth science.

**Vision:** To be a leader in promoting the collection, stewardship and use of Earth science data, information and knowledge that is responsive to societal needs.

**Values:**
- Innovative
- Neutral
- Open
- Participatory
- Voluntary
Current ESIP Collaborative Areas. ESIP can not cover *everything*, but where there is interest, that gets covered **really** good:

<table>
<thead>
<tr>
<th>Agriculture &amp; Climate</th>
<th>Information Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Education</td>
<td>Information Technology &amp;</td>
</tr>
<tr>
<td>Cloud Computing</td>
<td>Interoperability</td>
</tr>
<tr>
<td>Data Stewardship*</td>
<td>Libraries</td>
</tr>
<tr>
<td>Disasters</td>
<td>Products and Services*</td>
</tr>
<tr>
<td>Discovery</td>
<td>Science Communication</td>
</tr>
<tr>
<td>Documentation</td>
<td>Science Software</td>
</tr>
<tr>
<td>Drones</td>
<td>Semantic Web</td>
</tr>
<tr>
<td>Drupal</td>
<td>Sustainable Data Mgmt</td>
</tr>
<tr>
<td>Earth Science Data Analytics</td>
<td>Usability</td>
</tr>
<tr>
<td>Education</td>
<td>Visioneers</td>
</tr>
<tr>
<td>Energy and Climate</td>
<td>Web Services</td>
</tr>
<tr>
<td>EnviroSensing</td>
<td></td>
</tr>
</tbody>
</table>
This Big Data thing started to pop up everywhere
Let's solve this problem by using the Big Data. None of us have the slightest idea what to do with.
Big Data consists of extensive datasets, primarily in the characteristics of **volume**, **velocity** and/or **variety**, that require a scalable architecture for efficient storage, manipulation, and analysis.
“Big Data” is an umbrella term coined by Doug McLaney and IBM several years ago to denote data posing problems, summarized as the four Vs:

• **Volume** – the sheer size of “data at rest”
• **Velocity** – the speed of new data arriving (“data at move”)
• **Variety** – the manifold different
• **Veracity** – trustworthiness and issues of provenance
... in any aspect of **Big Data** with emphasis on **5Vs** (*Volume, Velocity, Variety, Value and Veracity*) relevant to variety of data (scientific and engineering, social, ...) that contribute to the Big Data challenges

Ruth adds:

*Visibility*
4 V’s of Big Data

Volume: Data at Rest
- Terabytes to exabytes of existing data to process

Velocity: Data in Motion
- Streaming data, milliseconds to seconds to respond

Variety: Data in Many Forms
- Structured, unstructured, text, multimedia

Veracity*: Data in Doubt
- Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

IBM, 2012
So… What’s the Big Deal about Big Data

If you just look at the ‘Big Data’ problem, it can indeed be overwhelming.

But, what’s new?... what’s different?... what’s the problem?

- We have been managing large volumes of heterogeneous datasets for a long time
- Researchers have been analyzing this data for a long time
- Technology is accommodating our needs

What is new is the need to grow and implement the ability to efficiently analyze data and information in order to extract knowledge
Increasing Amounts of Heterogeneous Datasets being made available to advance science research

... and a lot of people/directives are addressing it

Thus, it is not necessarily about Big Data, itself.

It is about the ability to examine large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information.

That is:

To glean knowledge from data and information
From a ‘to advance science’ point of view:

On the continuum of ever evolving data management systems, we need to understand and develop ways that allow for the variety of data relationships to be examined, and information to be manipulated, such that knowledge can be enhanced, to facilitate science.

In short, we have a lot of heterogeneous data that we really have not provided opportunity for users to holistically ‘mine’.

It’s new… and it ain’t easy…
Tackling Variety, Because…

• …it’s new…Information technology is just beginning to provide the tools for advancing the analysis of heterogeneous datasets in a ‘big’ way, thus, providing opportunity to discover unobvious scientific relationships, previously invisible to the science eye.

• … it ain’t easy… It takes individuals, or teams of individuals, with just the right combination of skills to understand the data and develop the methods to glean knowledge out of data and information.
Adding Value:
- Action-oriented
- Measurable efficiency
- Wiser decisions

Adding Value:
- Comparison
- Consequence
- Connections
- Conversations
- Analytics

Adding Value:
- Contextualized
- Categorized
- Calculated
- Corrected
- Condensed

WISDOM
Collective application of knowledge in action

KNOWLEDGE
Experience, values, context applied to a message

INFORMATION
A message meant to change receiver’s perception

DATA
Discrete, objective facts about an event

Experience
Grounded Truth
Complexity
Judgment
Heuristics
Values & Beliefs

Quantitative
Contextual
Evaluative
Qualitative
Intuitive
Informative

Quantitative
Connectivity
Transactions
Qualitative
Informative
Usefulness

Quantitative
Cost, Speed
Capacity
Qualitative
Timeliness
Relevance, Clarity

(Adapted from: https://km4meu.wordpress.com/tag/dikw-pyramid/)
Earth Science Data Analytics (ESDA) Cluster Goal:

To understand where, when, and how ESDA is used in science and applications research through speakers and use cases, and determine what Federation Partners can do to further advance technical solutions that address ESDA needs. Then do it.

Ultimate Goal:

To Glean Knowledge about Earth from All Available Data and Information
Mission:

- To promote a common understanding of the usefulness of, and activities that pertain to, Data Analytics and more broadly, the Data Scientist
- Facilitate collaborations between organizations that seek new ways to better understand the cross usage of heterogeneous datasets and organizations/individuals who can provide accommodating data analytics expertise, now and as the needs evolve into the future
- Identify gaps that, once filled, will further collaborative activities.
Objectives:

- Provide a forum for ‘Academic’ discussions that allow ESIP members to be better educated and on the same page in understanding the various aspects of Data Analytics.
- Bring in guest speakers to describe overviews of external efforts and further teach us about the broader use of Data Analytics.
- Perform activities that:
  - Compile use cases generated from specific community needs to cross analyze heterogeneous data (could be ESIP members or external).
  - Compile experience sources on the use of analytics tools, in particular, to satisfy the needs of the above data users (also, could be ESIP members or external).
  - Examine gaps between needs and expertise.
  - Document the specific data analytics expertise needed in above collaborations.
- Seek graduate data analytics/Data Science student internship opportunities.
ESDA Cluster – Highlights

- 19 Telecons
- 7 face-to-face sessions
- 16 ‘guest’ presentations
- Created the ESDA specific use case template
- Gathered 18 use Cases, and counting
- Defined Earth Science Data Analytics (submitted for ESIP adoption)
- Specified 3 types of ESDA definition types
- Defined 10 Earth science data analytics goals (submitted for ESIP adoption)
- Commenced ESDA Tools/Techniques requirements analysis
  - Began gathering and describing known tools/techniques
  - Began analyzing use case ESDA tools/techniques usage/needs
- Held sessions on teaching Earth science data analytics skills
- Presented our work at AGU
- Followed by 156 members (no they are not all active)
It did not take long to realize that ESDA was going to be a different kind of Cluster:

- The subject is new… unlike most other clusters that come with an existing knowledge base.

- When the ESDA Cluster was launched (early 2014), the literature was virtually absent of Earth science data analytics information. Bauman *(Big Data Analytics for Earth Sciences: the EarthServer approach)*, but few others, have since made in roads

- Data Analytics… what does that even mean. ‘I think it is something I should know about’

- What is the Cluster going to deliver? Software? An infrastructure? ‘I am not sure how to contribute’

- ‘How is this different from ______’ – fill in the blank
As a result:

- Although we felt that our ultimate goal would be to identify, for ESIP members to implement, gaps between Earth science multi-data usage and available facilitating tools and techniques …
- We realized we were going to be a cluster, initially, academic in nature, to help ESIP participants better understand the practical implementations of data analytics.
- Our telecon/meeting attendance have thus far attracted over 150 people.
- When surveying the audience at face-to-face meetings, ~80% still continue to attend to learn (note, however, that our scientist/data scientist attendance triple over 2 years… from 2 to 6).
- We brought in guest speakers.
- We studied the literature to better understand data analytics.
- We acquired use cases to help us scope data analytics.
- We kept our eye on the need to nurture young Data Scientists.
Guest Speakers – Telecons and Face-to-Face

- **Wo Chang:** NIST Big Data Public Working Group & Standardization Activities - 2/20/14
- **Brand Niemann:** Sorting out Data Science and Data Analytics - 3/20/14
- **John Schnase:** MERRA Analytic Services (MERRA/AS) - 3/20/14
- **Bamshad Mobasher:** Data Analytics Masters Program at DePaul University Overview - 3/20/14
- **Joan Aron:** Data Analytics Needs Scenario - 4/17/14
- **Rudy Husar:** User-Oriented Data Analytics and Tools using the Federated Data System DataFed - 4/17/14
- **Tiffany Mathews:** Atmospheric Science Data Center Sample Analytics Use Cases - 4/17/14
- **Peter Fox:** Data Science and Analytics Curriculum development at Rensselaer (and the Tetherless World Constellation) - 7/10/14
- **Steve Kempler:** Analytics and Data Scientists, Earth Science Data Analytics 101 - 1/7/15
- **Dave Bolvin:** From Many, One (or creating one great precipitation data set from many good ones) - 1/7/15
- **David Gallaher:** Reconstructing Sea Ice Extent from Early Nimbus Satellites - 1/7/15
- **Thomas Hearty:** Sampling Total Precipitable Water Vapor using AIRS and MERRA - 1/7/15
Guest Speakers – Telecons and Face-to-Face

- Radina Soebiyanto: Using Earth Observations to Understand and Predict Infectious Diseases - 1/7/15
- Tiffany Mathews: Promising data analytics technologies - 1/7/15
- Peter Fox: Data Scientists Are Freaks of Nurture but Products of Nature - 7/14/15
- Wade Bishop: Developing a Curriculum for the Earth Science Data Scientist - 7/14/15
- Karen Stocks: Educating Data Scientists: a view from the trenches - 7/14/15
- Steve Kempler: The Need for Earth Science Data Analytics to Facilitate Community Resilience (and other applications) - 7/16/15
- Shea Caspersen: MaxEnt: Modeling Terrestrial Ecology Under Climate Change - 1/8/16
The **data scientist**… analyzes huge volumes of data as well as other data sources that may be left untapped by conventional programs.

(https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics)

A data scientist possesses a combination of **analytic**, machine learning, data mining and statistical skills, typically related to a discipline domain.

(https://searchbusinessanalytics.techtarget.com/definition/Data-scientist)
Data Analytics: The process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information.

Analytics uses descriptive and predictive models to gain valuable knowledge from data...

Thus, analytics is not so much concerned with individual analyses or analysis steps, **but with the entire methodology**.

(http://en.wikipedia.org/wiki/Analytics)
Why is it important to identify Data Analytics Types

To better identify key needs that tools/techniques can be developed to address.

Basically, once we can categorize different types of Data Analytics, we can better associate existing and future Data Analytics tools and techniques that will help solve particular problems.
The 5 Types of Data Analytics

**Descriptive:** Analyze multiple datasets to describe conditions

**Diagnostic:** Analyze data to determine cause of condition

**“Discoveritive”:** Analyze multiple datasets to uncover new information

**Prescriptive:** Apply information to determine best action to take

**Predictive:** Analyze multiple datasets to assimilate future conditions

What the Literature Told Us
New analysis techniques and methods are being initiated to address large volumes of heterogeneous data that provide opportunities to examine data as we never did before. Growing computer capabilities facilitate this.

- **Business and healthcare applications** have jumped on advancing Data Analytics. Of the top Data Scientist/Analytic graduate programs:
  - ~80% focus on business applications
  - ~10% focus on health related applications
  - ~50% provide coursework that can also be applied to Earth science applications (However, do not necessarily include Earth science applications as part of their curriculum)

- In addition, specific applications have been performing Data Analytics... forever: e.g., Forensics, Crime solving

- **In Earth Science Research**, new data analytics techniques and methods, education, and tools are beginning to be formulated
Earth Science Data Analytics

It is our job (Information Technologists) to facilitate Data Analytics through our understanding and implementation of supportive information technologies, in close coordination with the specific data analysis needs of the science community

• **Data Preparation** – Making heterogeneous data so that they can ‘play’ together

• **Data Reduction** – Smartly removing data that do not fit research criteria

• **Data Analysis** – Applying techniques/methods to derive results

*Tools/Services for: Preparation are fairly generic; Reduction, and especially Analysis, are very specific research dependent (and, thus difficult for us to address without science domain expertise)*

Each component is required to some degree for each type of Data Analytics ... so we felt.
# First Earth Science Data Analytics Use Case Analysis Attempt

[http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics](http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics)

<table>
<thead>
<tr>
<th>Use cases</th>
<th>Descriptive</th>
<th>Diagnostic</th>
<th>Discoverive</th>
<th>Predictive</th>
<th>Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolvin (multi-dataset)</td>
<td></td>
<td>Hearty</td>
<td></td>
<td>Soebiyanto</td>
<td></td>
</tr>
<tr>
<td>Gallaher (single-dataset)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Techniques**

**Tools**

**User Types**

Look at user matrix marked up

**Skills/Expertise**
We discovered:

- Use Case types should be Goal Oriented

- Our Use Cases did not always fit cleanly into the 5 types of Data Analytics identified

- Maybe, the 5 types of data analytics, appropriate for the business world, do not accommodate goal oriented Earth science data analytics.

That is, in Earth science, we do not necessarily come up with the answers, but typically come up with discoveries that explain, at least for now, an answer.
Our studies and discussions began to focus on the uniqueness of Earth science data analytics
The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

This encompasses:

- **Data Preparation** – Preparing heterogeneous data so that they can be jointly analyzed
- **Data Reduction** – Correcting, ordering and simplifying data in support of analytic objectives
- **Data Analysis** – Applying techniques/methods to derive results
Data Analytics Goals

Why is it important to identify Data Analytics Goals

To better identify key needs that tools/techniques can be developed to address.

Basically, once we can categorize different goals of Data Analytics, we can better associate existing and future Data Analytics tools and techniques that will help solve particular problems.
Earth Science Data Analytics Goals
(currently being discussed for adoption by ESIP)

(read: Earth science data analytics needed ...)

1. To calibrate data
2. To validate data (note it does not have to be via data intercomparison)
3. To assess data quality
4. To perform coarse data preparation (e.g., subsetting data, mining data, transforming data, recovering data)
5. To intercompare datasets (i.e., any data intercomparison; Could be used to better define validation/quality)
6. To tease out information from data
7. To glean knowledge from data and information
8. To forecast/predict/model phenomena (i.e., Special kind of conclusion)
9. To derive conclusions (i.e., that do not easily fall into another type)
10. To derive new analytics tools
ESDA Use Case Template

- Use Case Title
- Author/Company/Email
- Actors/Stakeholders/Project URL and their roles and responsibilities
- Use Case Goal
- Use Case Description
- Current technical considerations to take into account that may impact needed data analytics.
- Data Analytics tools applied
- Data Analytics Challenges (Gaps)
- Type of User
- Research Areas
- Societal Benefit Areas
- Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)
- More Information and relevant URLs (e.g. who to contact or where to go for more information)
<table>
<thead>
<tr>
<th>Use Cases</th>
<th>Earth Science Data Analyitics Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MERRA Analytics Services: Climate Analytics-as-a-Service</td>
<td>1 2 3 4 5 6 7 8 9 10</td>
</tr>
<tr>
<td>2 MUSTANG QA: Ability to detect seismic instrumentation problems</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>3 Inter-calibrations among datasets</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>4 Inter-comparisons between multiple model or data products</td>
<td>✓</td>
</tr>
<tr>
<td>5 Sampling Total Precipitable Water Vapor using AIRS and MERRA</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>6 Using Earth Observations to Understand and Predict Infectious Diseases</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>7 CREATE-IP - Collaborative REAnalysis Technical Environment - Intercomparison Project</td>
<td>✓</td>
</tr>
<tr>
<td>8 The GSSTF Project (MEaSUREs-2006)</td>
<td>✓</td>
</tr>
<tr>
<td>9 Science- and Event-based Advanced Data Service Framework at GES DISC</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>10 Risk analysis for environmental issues</td>
<td>✓</td>
</tr>
<tr>
<td>11 Aerosol Characterization</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>12 Creating One Great Precipitation Data Set From Many Good Ones</td>
<td>✓</td>
</tr>
<tr>
<td>13 Reconstructing Sea Ice Extent from Early Nimbus Satellites</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>14 DOE-BER AmeriFlux and FLUXNET Networks *</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>15 DOE-BER Subsurface Biogeochemistry Scientific Focus Area *</td>
<td>✓</td>
</tr>
<tr>
<td>16 Climate Studies using the Community Earth System Model at DOE’s NERSC center *</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>17 Radar Data Analysis for CReSIS *</td>
<td>✓</td>
</tr>
<tr>
<td>18 UAVSAR Data Processing, Data Product Delivery, and Data Service *</td>
<td>✓</td>
</tr>
</tbody>
</table>

* - Borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php]
Goal oriented Earth Science Data Analytics (ESDA) reveal requirements for needed data analytics tools/techniques

**Motivation**
How can we maximize the usability of large heterogeneous datasets to glean knowledge out of the data?

**Earth Science Data Analytics: Definition**
The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

**Earth Science Data Analytics: Goals**
- To derive new analytics tools
- To derive conclusions
- To forecast/predict/model
- To glean knowledge
- To tease out information
- To intercompare datasets
- To perform coarse data preparation
- To assess data quality
- To validate data
- To calibrate data
- To forecast/ predict/model
- To derive new analytics tools

**Earth Science Data Analytics: Initial Requirements**
- Ingest from various sources; High speed processing; Math functions
- Access large datasets; Assess erroneous data; Detect data anomalies

**Earth Science Data Analytics: Exemplary Tools, Techniques, Integrated Systems**

### Types of Analytics

- **Data Preparation**
  - R, SAS, Python, Java, C++
  - SPSS, MATLAB, Minitab
  - CPLEX, GAMS, Gauss
  - Tableau, Spotfire
  - VBA, Excel, MySQL
  - JavaScript, Perl, PHP
  - Open Source Databases
  - PBO, NCL, Parallel NetCDF
  - AWS, Cloud Solutions, Hadoop
  - MPI, GIS, ROB-PAC, GDAL

- **Data Reduction**
  - Statistics functions
  - Machine Learning
  - Data Mining
  - Natural Language Processing
  - Linear/Non-linear Regression
  - Logical Regression
  - Time Series Models
  - Clustering
  - Decision Tree
  - Factor Analysis
  - Principal Component Analysis
  - Neural Networks
  - Bayesian Techniques
  - Text Analytics
  - Graph Analytics
  - Visual Analytics
  - Map Reduce

### Integrated Systems

- EarthServer (http://www.earthserver.eu)
- NASA Earth Exchange (https://nex.nasa.gov/nex/)
- EDEN (http://cda.ornl.gov/projects/eden/)
- EARTHDATA (https://earthdata.nasa.gov)
- Giovanni (http://giovanni.gsfc.nasa.gov/giovanni/)

**Methodology**
Categorize/Analyze ESDA use cases; derive data analytics requirements; associate tools/techniques; perform gap analysis

**Compiled from:** [practicalanalytics.co/predictive-analytics-101/](http://practicalanalytics.co/predictive-analytics-101/) and [cda.ornl.gov/research.shtml](http://cda.ornl.gov/research.shtml)

---

**The good news…**

**Earth Science Data Analytics: Preparing for the Future**

Central England NERC Training Alliance  
**CENTA**

Big data analysis to fuel environmental research at Reading University

2nd Annual Graduate Workshop on Environmental Data Analytics  
May 28-29, 2015

Hosted by the National Center for Atmospheric Research in Boulder, CO

- ... offering degrees in Data Science
- ... summer school on Big Data Analytics
- ... online master’s degree in data analytics

**Earth Science Data Analytics: Looking Ahead**

- Complete Gap Analysis between ESDA requirements and current tools/technologies
- Continue to evolve tools/techniques to address growing scope of the ‘Internet of Things’

- IoT
- Big Data
- Analytics
- Analysis and Remote Sensing
- computer models on remote sensing... 

- Two main themes:  
  - big data and data analytics "fueled" by Earth observation  
  - data integration and interoperability among Earth sciences in the IOT
Deriving Earth Science Data Analytics Requirements

Earth Science Data Analytics: Definition
The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

Earth Science Data Analytics: Goals
- Data Preparation
  - To validate data
  - To perform coarse data preparation
  - To calibrate data
  - To assess data quality

- Data Reduction
  - To intercompare datasets
  - To tease out information

- Data Analysis
  - To glean knowledge
  - To forecast/predict/model
  - To derive conclusions
  - To derive new analytics tools

Earth Science Data Analytics: Initial Requirements
- Ingest from various sources; Homogenize data; Visualization; Sampling; Gridding
- Access large datasets; High speed processing; Subsetting, mining, machine learning
- Homogenize data; Intercomparison statistics; Pattern recognition
- Seek heterogeneous data relationships; Ingest from various sources; Image processing
- Looking for Community input
- Data exploration; Filter, mine, fuse, interpolate data; Manage custom code
- Data exploration; Neural networks; Math/Stat modeling; Near Real Time data
- Access very large datasets; homogenize data; visualization
## Earth Science Data Analytics
### Exemplary Tools, Techniques, Integrated Systems

<table>
<thead>
<tr>
<th>Types of Analytics</th>
<th>Tools</th>
<th>Techniques</th>
<th>Integrated Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Preparation</td>
<td>R, SAS, Python, Java, C++</td>
<td>Statistics functions</td>
<td>EarthServer (<a href="http://www.earthserver.eu">http://www.earthserver.eu</a>)</td>
</tr>
<tr>
<td></td>
<td>SPSS, MATLAB, Minitab</td>
<td>Machine Learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPLEX, GAMS, Gauss</td>
<td>Data Mining</td>
<td>NASA Earth Exchange (<a href="https://nex.nasa.gov/nex/">https://nex.nasa.gov/nex/</a>)</td>
</tr>
<tr>
<td>Data Reduction</td>
<td>Tableau, Spotfire</td>
<td>Natural Language Processing</td>
<td>EDEN (<a href="http://cda.ornl.gov/projects/eden/#">http://cda.ornl.gov/projects/eden/#</a>)</td>
</tr>
<tr>
<td></td>
<td>VBA, Excel, MySQL</td>
<td>Linear/Non-linear Regression</td>
<td>EARTHDATA (<a href="https://earthdata.nasa.gov">https://earthdata.nasa.gov</a>)</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>Javascript, Perl, PHP</td>
<td>Logical Regression</td>
<td>Giovanni (<a href="http://giovanni.gsfc.nasa.gov/giovanni/">http://giovanni.gsfc.nasa.gov/giovanni/</a>)</td>
</tr>
<tr>
<td></td>
<td>Open Source Databases</td>
<td>Time Series Models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PIO, NCL, Parallel NetCDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AWS, Cloud Solutions, Hadoop</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPI, GIS, ROI-PAC, GDAL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Techniques**
- Factor Analysis
- Principal Component Analysis
- Neural Networks
- Bayesian Techniques
- Graph Analytics
- Visual Analytics
- Map Reduce
- Decision Tree
<table>
<thead>
<tr>
<th>Tool/Technique/Integrated System</th>
<th>Description</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. (Wikipedia)</td>
<td>Steve</td>
</tr>
<tr>
<td>SAS</td>
<td>SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics. SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated. (Wikipedia)</td>
<td>Steve</td>
</tr>
<tr>
<td>Python</td>
<td>Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. (Wikipedia)</td>
<td>Sean</td>
</tr>
<tr>
<td>Java</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>C++</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>SPSS</td>
<td></td>
<td>Sean</td>
</tr>
<tr>
<td>MATLAB</td>
<td></td>
<td>Sean</td>
</tr>
<tr>
<td>Mintab</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>CPLEX</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>GAMS</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>Gauss</td>
<td></td>
<td>Steve</td>
</tr>
<tr>
<td>Tableau</td>
<td>A tool that enables data visualization using a drag and drop interface.</td>
<td>Thomas</td>
</tr>
<tr>
<td>Spotfire</td>
<td>A tool that enables data mining and visualization of very large data sets. Similar to Excel but apparently easier to use for large data sets.</td>
<td>Thomas</td>
</tr>
<tr>
<td>VBA</td>
<td>(Visual Basic for Applications) An implementation of Visual Basic that enables user defined functions and interaction with Windows API and libraries.</td>
<td>Thomas</td>
</tr>
<tr>
<td>Excel</td>
<td>A spreadsheet program created by Microsoft that enables data analysis and visualization. It includes VBA.</td>
<td>Thomas</td>
</tr>
</tbody>
</table>
We Began Describing Identified Tools/Techniques/Integrated Systems

<table>
<thead>
<tr>
<th><strong>MySQL</strong></th>
<th>Thomas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Javascript</td>
<td>Thomas</td>
</tr>
<tr>
<td>Perl</td>
<td>Thomas</td>
</tr>
<tr>
<td>PHP</td>
<td>Thomas</td>
</tr>
<tr>
<td>Open Source Databases</td>
<td>Steve</td>
</tr>
<tr>
<td>PIO</td>
<td>Steve</td>
</tr>
<tr>
<td>NCL</td>
<td>Steve</td>
</tr>
<tr>
<td>Parallel NetCDF</td>
<td>Steve</td>
</tr>
<tr>
<td>AWS</td>
<td>Steve</td>
</tr>
<tr>
<td>Cloud Solutions</td>
<td>Steve</td>
</tr>
<tr>
<td>Statistics functions</td>
<td>-</td>
</tr>
<tr>
<td><strong>Machine Learning</strong></td>
<td>Chung-Lin</td>
</tr>
<tr>
<td>Data Mining</td>
<td>Chung-Lin</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>Chung-Lin</td>
</tr>
</tbody>
</table>

- Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.

- Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

- Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.
We Began Describing Identified Tools/Techniques/Integrated Systems

| Linear/Non-linear Regression | In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable Y (e.g., a sounding temperature) and one or more explanatory variables (or independent variables) denoted X, (or X1, X2...) (e.g., the satellite retrieved temperature(s)). The case of one explanatory variable is called simple linear regression. In statistics, nonlinear regression is a form of regression analysis in which observational data (e.g., Y) are modeled by a function which is a nonlinear combination of the model parameters (e.g., aX + bX2 +...) and depends on one or more independent variables (e.g., X or X1, X2,...). The data are fitted by a method of successive approximations. | Chung-Lin |
| Logical Regression          | Bob |
| Time Series Models          | Time Series Models are used to represent trends, often graphically, by applying temporal measurements within a sequence. | Bob |
| Clustering                  | Clustering is an approach to organize objects into a classification and can be accomplished utilizing various methods, including statical techniques. | Bob |
| Decision Tree               | A Decision Tree is a graphical representation of the sequence of decisions to be completed when answering a particular question. | Bob |
(Thanks Ethan)

- This opened our eyes to a great resource that associates computational techniques to specific data science ‘stages’:
  - Describe, Discover, Predict, Advise
    - These stages are described in terms of increasing maturity
    - Interpreted for Earth science, each stage would have independent maturity levels. We would call them ‘goals’, albeit at a different level
    - However, these ‘stages’ provide organization towards the utilization of techniques and tools to achieve analytics goals

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics
Data Science:

- **Describe**
  - Processing
    - Filtering, Imputation, Dimensionality Reduction, Normalization/Transformation
  - Aggregation
  - Enrichment

- **Discover**
  - Clustering
  - Regression
  - Hypothesis Testing

- **Predict**
  - Regression
  - Recommendation

- **Advise**
  - Local reasoning
  - Optimization
  - Simulation
Also described are the different classes of techniques:

Transforming
Learning
Predictive
These classes pretty much correspond to ESDA types:

- Transforming → Data Preparation, Data Reduction
- Learning → Data Analysis
- Predictive → Data Analysis
What we have to do

- Review/Understand technology descriptions
- Categorize them by ESDA types
- Determine what goals they can support
Then We Went to AGU …

Our Analytics Session

- “Geophysical Science Data Analytics Use Case Scenarios”
- 12 Posters
- Will be acquiring additional Use Cases
- Analytics methodologies highlighted include: Decision Trees, Machine Learning, Data Mining, Decision Tree
But also, at the AGU ...

Visited science posters to better understand research methodologies…

- Looked for presentations that discussed the co-analysis of multiple datasets
- Looked for presentations that described methodology techniques employed
- ‘Scanned’ 100’s of posters, identifying presentations (and through discussion with authors) that provide sought after information
  - 31 Atmospheric Science research projects identified
  - 12 Hydrology Science research projects identified
  - (Don’t read into the numbers, this is just as far as we got)
- Science research methodology techniques being used …
Science research methodology techniques being used (AGU findings)

- In Atmospheric Research (study of gases):
  - Correlation Analysis; Bias Correlation
  - Regression Analysis; Bivariant Regression
  - Decision Tree
  - Machine Learning
  - Data Mining
  - Data Fusion
  - Computational Tools
  - Constrained Variational Analysis
  - Model Simulations
  - Ratios
  - Time Series Analysis
  - Spectral Analysis
  - Temporal Trending; Trend Analysis
  - Spatial Interpolation
  - Revised Averaging Scheme
  - Forward Modeling; Inverse Modeling
  - Radiative Transfer Model
  - Baysian Synthesis Inversion
  - Temporal Stability
  - Gaussian Distribution
  - Exponential Differentiation

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics
Science research methodology techniques being used (AGU findings)

- In Hydrology Research (study of liquid):
  - Linear Regression
  - Monte Carlo
  - Darcy Equation
  - Poisson Regression
  - Multi-variate time series analysis
  - BUDYKO formula
  - Smoothing (Gaussian)
  - Filtering (Destriping)
  - MESH Model

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics
## Framework for Putting it All Together

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics

### ESDA Goals

<table>
<thead>
<tr>
<th>ESDA Goals</th>
<th>Data Preparation</th>
<th>Data Reduction</th>
<th>Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. To calibrate data</td>
<td>ESDA Requirements</td>
<td>ESDA Tools/Techniques</td>
<td>ESDA Requirements</td>
</tr>
<tr>
<td>2. To validate data (note it does not have to be via data intercomparison)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. To assess data quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. To perform coarse data preparation (e.g., subsetting data, mining data, transforming data, recovering data)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. To intercompare datasets (i.e., any data intercomparison; Could be used to better define validation/quality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. To tease out information from data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. To glean knowledge from data and information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. To forecast/predict/model phenomena (i.e., Special kind of conclusion)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. To derive conclusions (i.e., that do not easily fall into another type)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. To derive new analytics tools</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Framework for Putting it All Together

<table>
<thead>
<tr>
<th>ESDA Goals</th>
<th>Data Preparation</th>
<th>Data Reduction</th>
<th>Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ESDA Requirements</td>
<td>ESDA Tools/Techniques</td>
<td>ESDA Requirements</td>
</tr>
<tr>
<td>1. To calibrate data</td>
<td>Ingest from various sources</td>
<td></td>
<td>High speed processing; Math functions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. To validate data (note it does not have to be via data intercomparison)</td>
<td>Ingest from various sources; Homogenize data</td>
<td>Sampling</td>
<td>Visualization; Gridding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. To assess data quality</td>
<td>Access large datasets</td>
<td></td>
<td>Assess erroneous data; Detect data anomalies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. To perform coarse data preparation (e.g., subsetting data, mining data,</td>
<td>Access large datasets</td>
<td>Subsetting, mining, machine learning</td>
<td>High speed processing</td>
</tr>
<tr>
<td>transforming data, recovering data)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. To intercompare datasets (i.e., any data intercomparison; Could be used</td>
<td>Homogenize data</td>
<td></td>
<td>Intercomparison on statistics; Pattern recognition</td>
</tr>
<tr>
<td>to better define validation/quality)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. To tease out information from data</td>
<td>Seek heterogeneous data relationships; Ingest from</td>
<td></td>
<td>Seek data relationships; Image processing</td>
</tr>
<tr>
<td></td>
<td>various sources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. To glean knowledge from data and information</td>
<td>Looking for Community input</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. To forecast/predict/model phenomena (i.e., Special kind of conclusion)</td>
<td>Data exploration; Near Real Time data</td>
<td>Neural networks</td>
<td>Math/Stat modeling</td>
</tr>
<tr>
<td>9. To derive conclusions (i.e., that do not easily fall into another type)</td>
<td>Data exploration; code</td>
<td>Filter, mine, fuse, interpolate data</td>
<td>Manage custom code</td>
</tr>
<tr>
<td>10. To derive new analytics tools</td>
<td>Access very large datasets; homogenize data</td>
<td></td>
<td>Visualization</td>
</tr>
</tbody>
</table>
Beginning to Better Understand ESDA

- Looking for more use cases.....

- Getting our arms around the usefulness of existing tools and techniques for ESDA data preparation, data reduction, data analysis… our current effort

- We see that analytics useful for Data Analysis is the most difficult to develop an approach for:
  - Science research/analysis is very individual
  - Libraries of mathematical tools already exist
  - The plethora of specific research models are unique, and well understand by the researcher, rendering us no real opportunity to add value for large groups of users per model.
Beginning to Better Understand ESDA

We see that heterogeneous Data Preparation is where the most pain points are, and tools/techniques that target heterogeneous data preparation should be targeted first.

- Addressing Data Preparation needs will directly help Data Analysis
- To help, we need to invite more scientists to our cluster to provide more insights to their experiences and needs regarding the co-analysis of heterogeneous data.
- Institute a 'science advisory board' (maybe ESIP would/will)??
- Include applications researchers
- Engage young data analytics scientists

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics
Thank you

… and thanks to all who have peeked in to the cluster…

… and the ‘regulars’:

Joan Aron, The Barberie Twins, Rob Casey, Robert Downs, Beth Huffer, Ethan McMahon, Erin Robinson, Chung-Lin Shie, and of course, Tiffany Mathews

… any everyone in between

http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics
BACKUP
Data Scientist

A data scientist possesses a combination of analytic, machine learning, data mining and statistical skills as well as experience with algorithms and statistical skills as well as experience with algorithms and coding. Perhaps the most important skill a data scientist possesses, however, is the ability to explain the significance of data in a way that can be easily understood by others. (Source: http://searchbusinessanalytics.techtarget.com/definition/Data-scientist)

Rising alongside the relatively new technology of big data is the new job title data scientist. While not tied exclusively to big data projects, the data scientist role does complement them because of the increased breadth and depth of data being examined, as compared to traditional roles. (Source: http://www-01.ibm.com/software/data/infosphere/data-scientist/)
Analytics

(http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/)

**Categories of Analytics**

**Prescriptive Analytics**
- **Optimization**
  - Focus on decision making and efficiency
  - Optimization in a problem solving technique where situations and constraints are modeled to arrive at the most optimal solution
  - Simulation is used to analyze complex system to gain insight in to the system’s behavior and identify issues

- **Simulation**

**Predictive Analytics**
- **Data Mining**
  - Focus on prediction of Model probabilities and verify
  - Data Mining is the method of extracting patterns from large data sets in order to provide insight and future forecasts
  - Predictive Modeling uses statistical techniques such a linear and logistic regression to understand Model and predict future outcomes.

- **Predictive modeling**

**Descriptive Analytics**
- **Data Modeling**
  - Analytics involved in preparing data for advanced analysis or for general day-to-day business intelligence
  - Data Modeling is used to collect, store and cut the data in an efficient way
  - Visualization looks at the creation of reports and presenting information in a thoughtful fashion
  - Regression is used to find simple trends in the data

- **Visualization**
- **Regression**

*Source: Cap Gemini Blog, May 27, 2011*
Another look at Analytics
(http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/)
A 2011 McKinsey report suggests suitable technologies include...

(http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

...A/B testing, association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modelling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualisation.