Analytics and Data Scientists

Session: Earth Science Data Analytics 101

Steve Kempler, Moderator

January 7, 2015

ESIP Federation Meeting
Washington, DC

Earth Science Data Analytics (ESDA) 101

"Large Heterogeneous Datasets" (LHDs) aka "Big Data", "Analytics", "Data Scientist", "Data-Intensive Science" are overused and often abused terms that mean different things to different people.

Purpose of this Session:

To decipher what Earth Science Data Analytics, (etc.) means, and provide exemplary use cases

Agenda – ESDA 101

- Introduction to Earth Science Data Analytics (ESDA) and ESIP's ESDA Cluster – Steve Kempler - ~15 min

  ESDA Use Cases (~10 min each):

- David Bolvin - **From Many, One (or creating one great precipitation data set from many good ones)**
- David Gallaher - **Reconstructing Sea Ice Extent from Early Nimbus Satellites**
- Thomas Hearty - **Sampling Total Precipitable Water Vapor using AIRS and MERRA**
- Radina Soebiyanto - **Using Earth Observations to Understand and Predict Infectious Diseases**
- Promising data analytics technologies – Tiffany Matthews – 15 min
- Panel Discussion: Q&A with audience - remaining time

  Focus: How can Federation information technologists support/facilitate Earth science data analytics oriented research – All

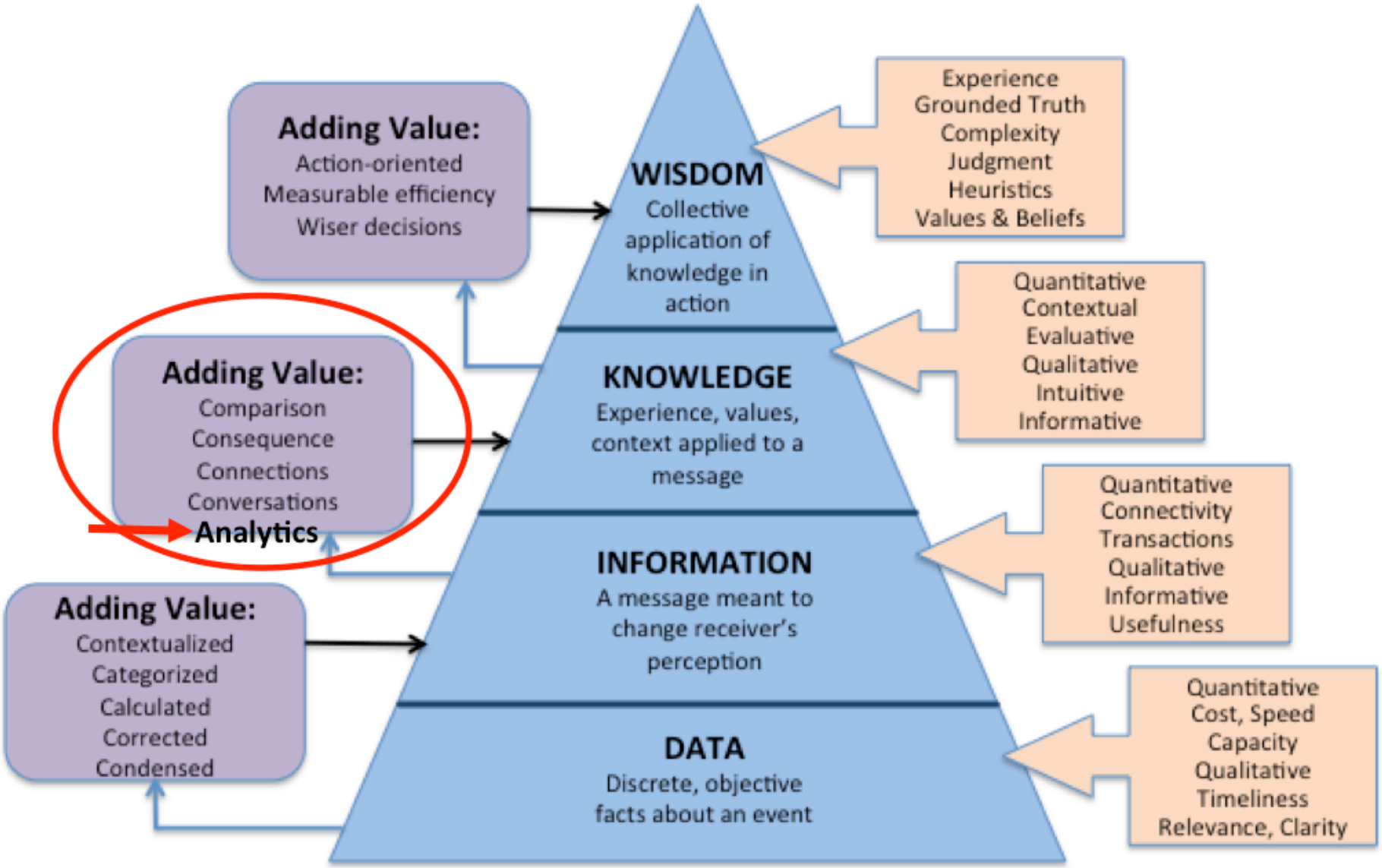
Earth Science Data Analytics Cluster

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



(Adapted from: https://km4meu.wordpress.com/tag/dikw-pyramid/)

From the Fourth Paradigm
(Editors: Hey, Tansley,Tolle)

http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

"We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization" (P. xvii)

From the Fourth Paradigm
(Editors: Hey, Tansley,Tolle)

http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

"Clearly, data-intensive science… must **move beyond data warehouses** and closed systems, striving instead to allow access to data to those outside the main project teams, allow for **greater integration of sources**, and provide interfaces to those who are expert scientists but not experts in data administration and computation." (Fox, P. 147)

Motivation

Increasing Amounts of Heterogeneous Datasets
aka Big Data
… and a lot of people/directives are addressing it

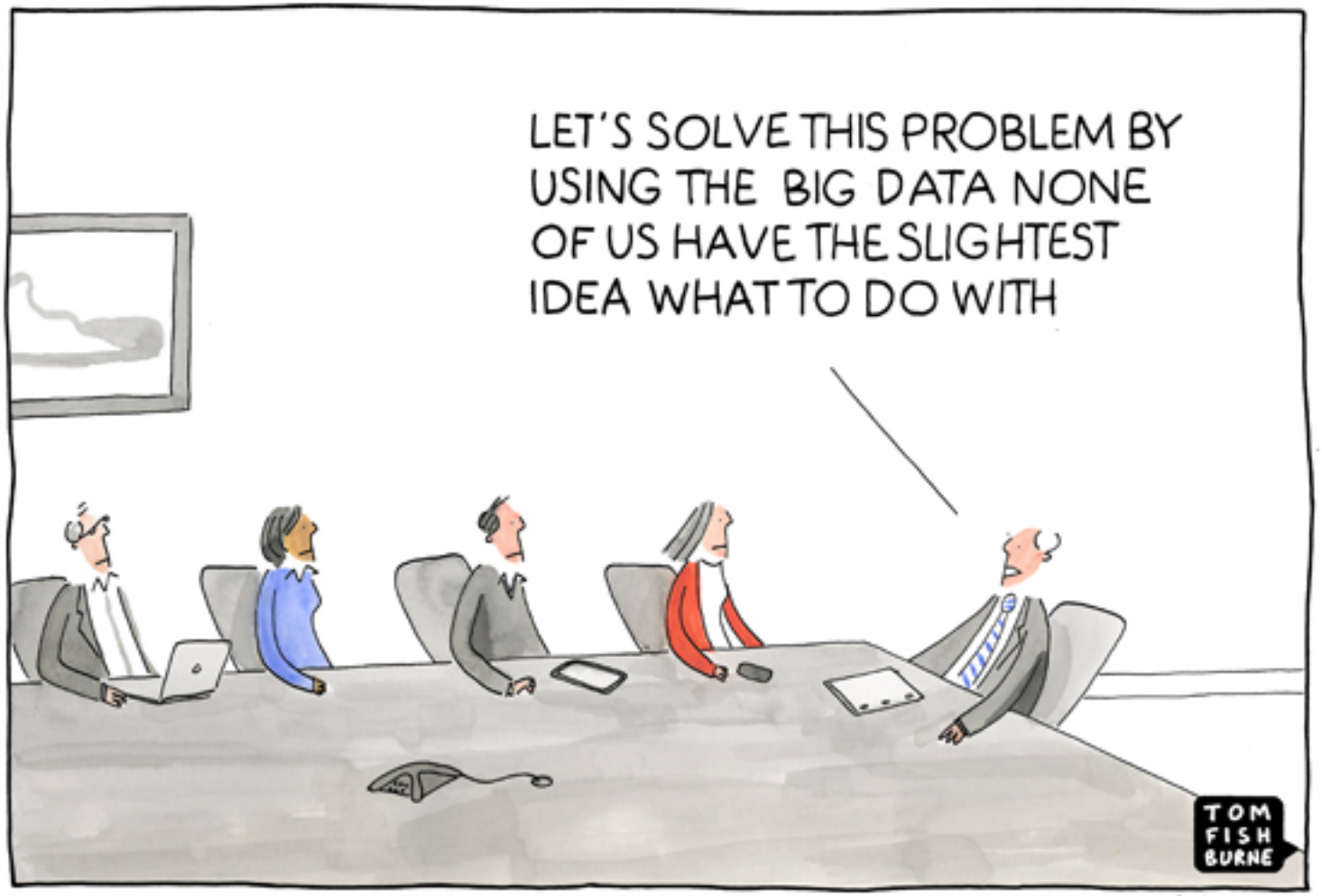
But don't worry… I won't discuss any words that begin with 'v' *.

(If you were at AGU, you've seen them enough)

* I have backup slides for later, if you need a 'v' refesher

Motivation: e.g., NASA Earth Science Data
(EOSDIS FY2013 Annual Metrics Report)

- Need to analyze large amounts of data and information
  - 9.8 PB
  - 9,800,000,000,000,000 bytes
- Demand to quickly access data
  - 22 TB/day distributed, but…
  - … 2.3 TB/day Near Real Time, distributed within hours
- Increasing need to co-analyze heterogeneous data
  - 6,861 Unique Datasets
  - Not to mention non-NASA datasets that can be co-analyzed
- Necessity to know data is reliable and accurate
  - Metadata provides indicators to describe data quality
  - Ongoing data comparison and refinement analysis continuously improves data trustworthiness



© marketoonist.com

Big Data Initiatives and Interest Groups

- Big Earth Data Initiative (BEDI) and Climate Data Initiative (CDI) (actually has funding)
- IEEE International Conference on Big Data
- OGC Big Data WG
- bigdatawg.**nist**.gov
- Research Data Sharing (RDA) BDA (Big Data Analytics)
- Earth Sciences Information Partners (ESIP) ESDA

And many, many more private and public organizations that sponsor special interest groups to promote and share techniques, technologies, and knowledge to address Big Data

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

Gleaning Knowledge about Earth from All Available Data and Information

From a '**to advance science**' point of view:

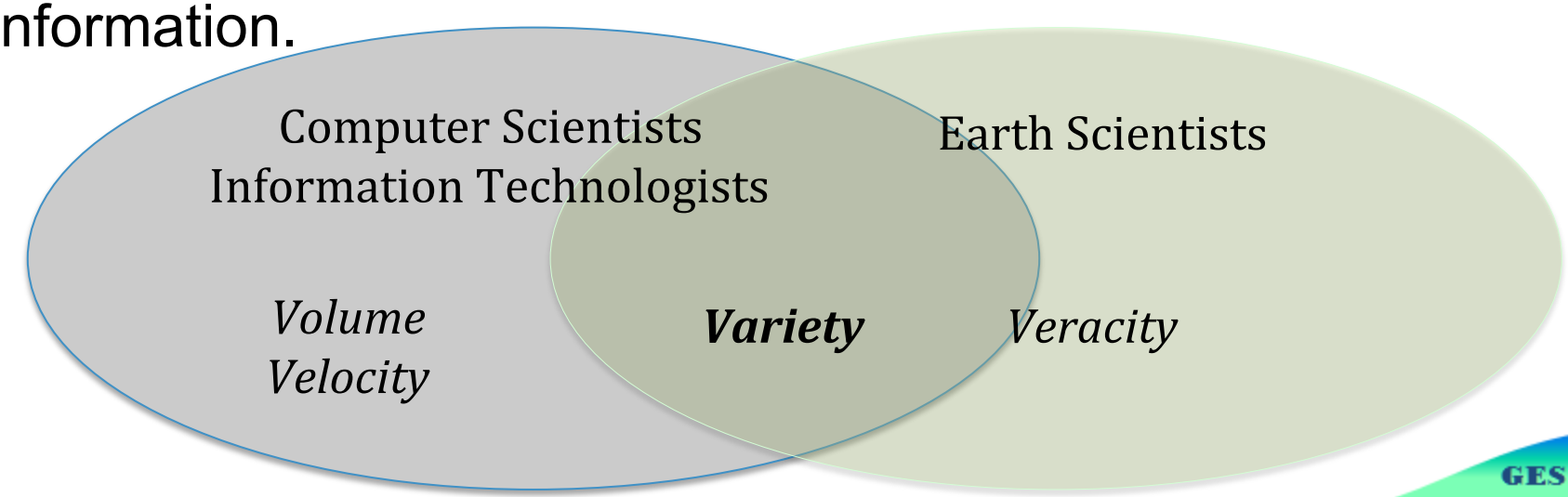
On the continuum of ever evolving data management systems, we need to understand and develop ways that allow for the **variety** of data relationships to be examined, and information to be manipulated, such that knowledge can be enhanced, to facilitate science.

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.



Enter the Data Scientist…

The **data scientist**… analyzes huge volumes of data as well as other data sources that may be left untapped by conventional programs.

(http://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.

(http://searchbusinessanalytics.techtarget.com/definition/Data-scientist)

… And Data Analytics

- **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)

The Punchline

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

Components of Earth Science Data Analytics

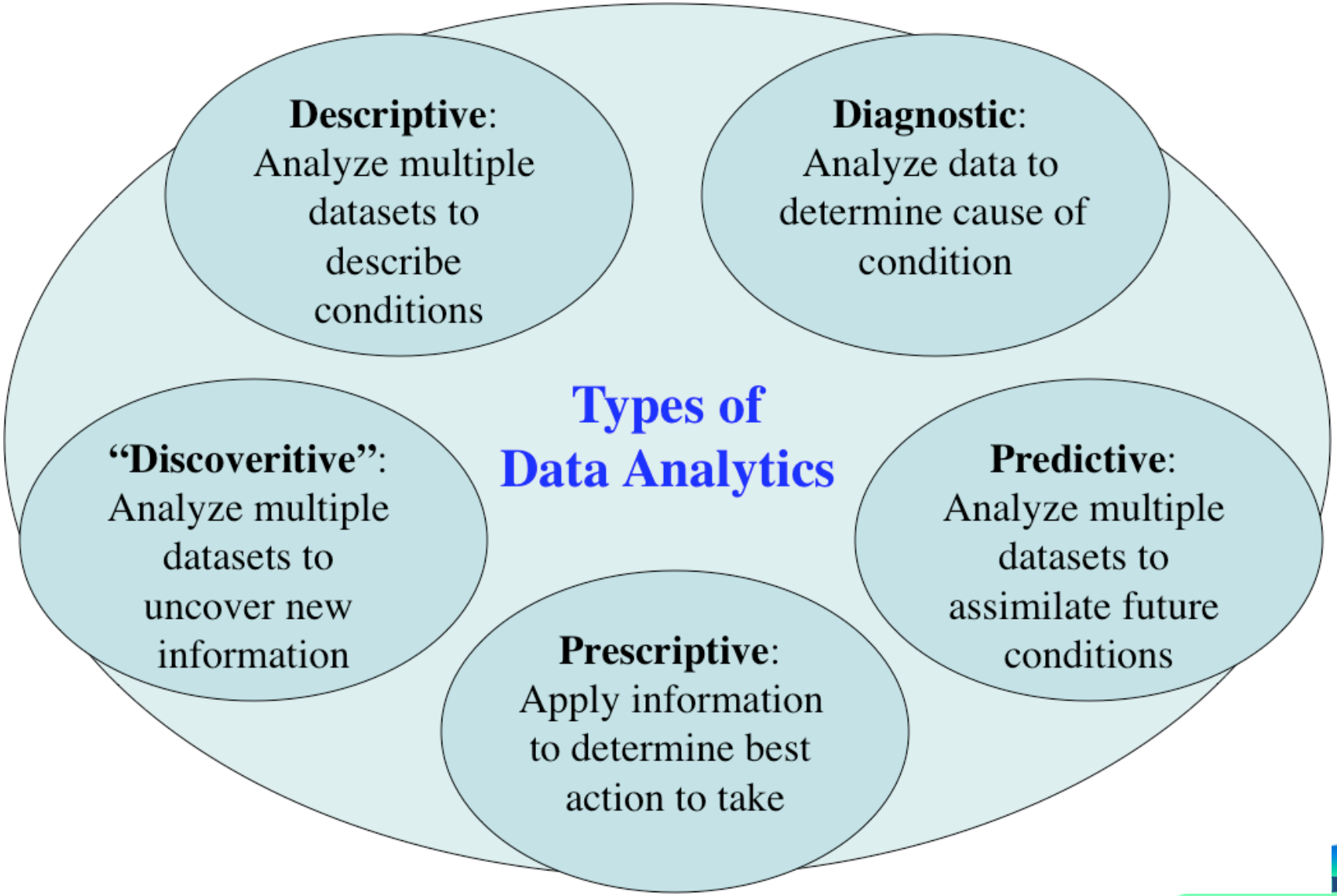
It is our job (Information Technologists) to facilitate Data Analytics through our understanding and implementation of supportive information technologies, in close coordination with the specific data analysis needs of the science community

- **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 (next slide)

The 5 Types of Data Analytics



Data Analytics Today

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 (ESDA) 201

(sneak look ahead)

Purpose of the ESDA 201 Session:

To discuss completing our matrix

	Descriptive	Diagnostic	Discoveritive	Predictive	Prescriptive
Use cases	Bolvin (multi-dataset) Gallaher (single-dataset	Hearty		Soebiyanto	
Techniques					
Tools					
User Types	Look at user matrix marked up				
Skills/Expertise					

To discuss use cases (←next big step)

~ To discuss advancing our efforts

Thank you

BACKUP

NIST Big Data Definitions and Taxonomies, V 0.9

National Institute of Standards and Technology (NIST) Big Data Working Group (NBD-WG)
February, 2014, http://bigdatawg.nist.gov/show_InputDoc.php, M0142

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.*

Open Geospatial Consortium (OGC)
Big Data Working Group

http://external.opengeospatial.org/twiki_public/BigDataDwg/WebHome

*"**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*

IEEE BigData 2014

http://cci.drexel.edu/bigdata/bigdata2014/callforpaper.htm

*… 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*

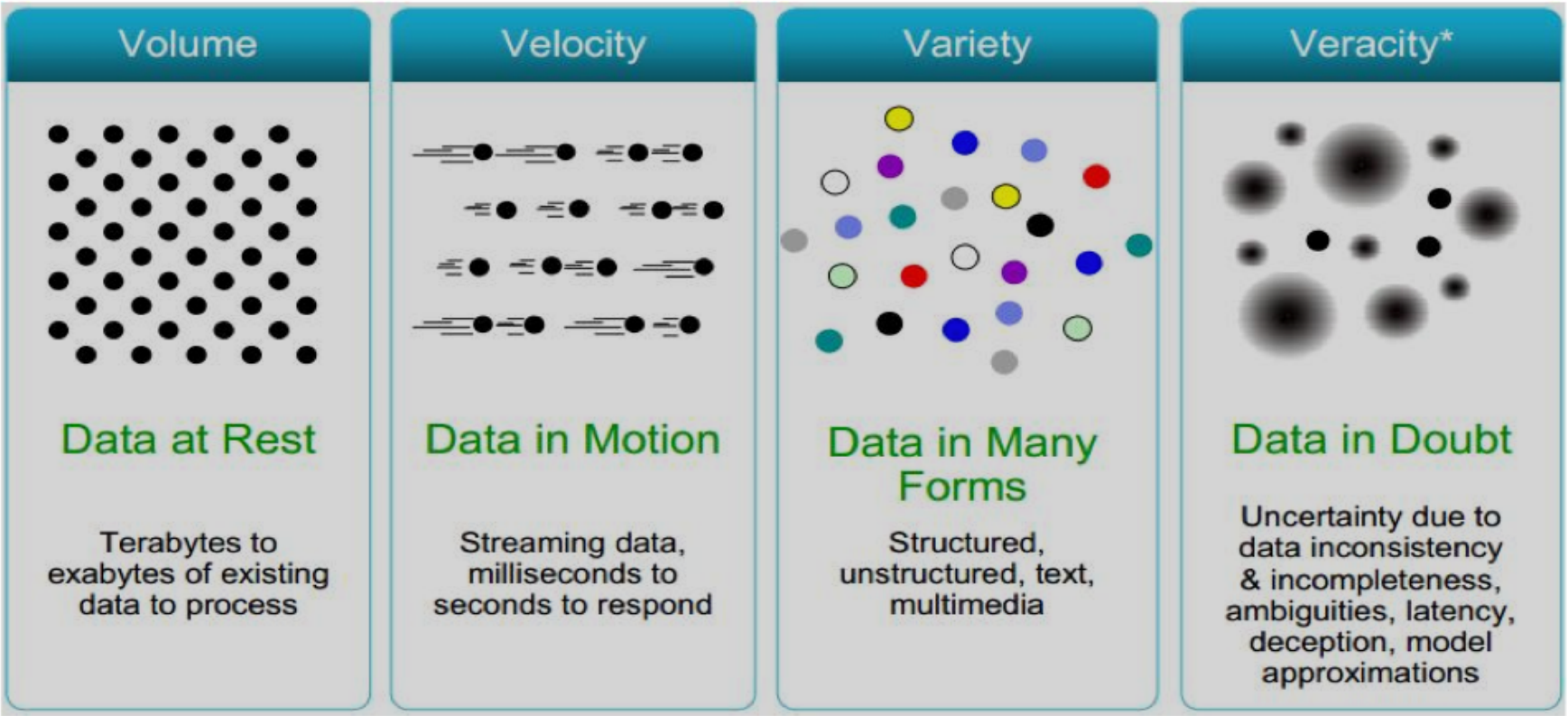
Ruth adds:

Visibility

From: Demystifying Data Science

(Natasha Balac , accessible via: http://bigdatawg.nist.gov/show_InputDoc.php, M0169)

4 V's of Big Data



IBM, 2012

So, Why does Big Data Have Everybody's Attention?

This is an encourager: (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf)

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

Data Scientist in the context of analytics

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

Simulation

- 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

Predictive Analytics

Data Mining

Predictive modeling

- 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.

Descriptive Analytics

Data Modeling

Visualization

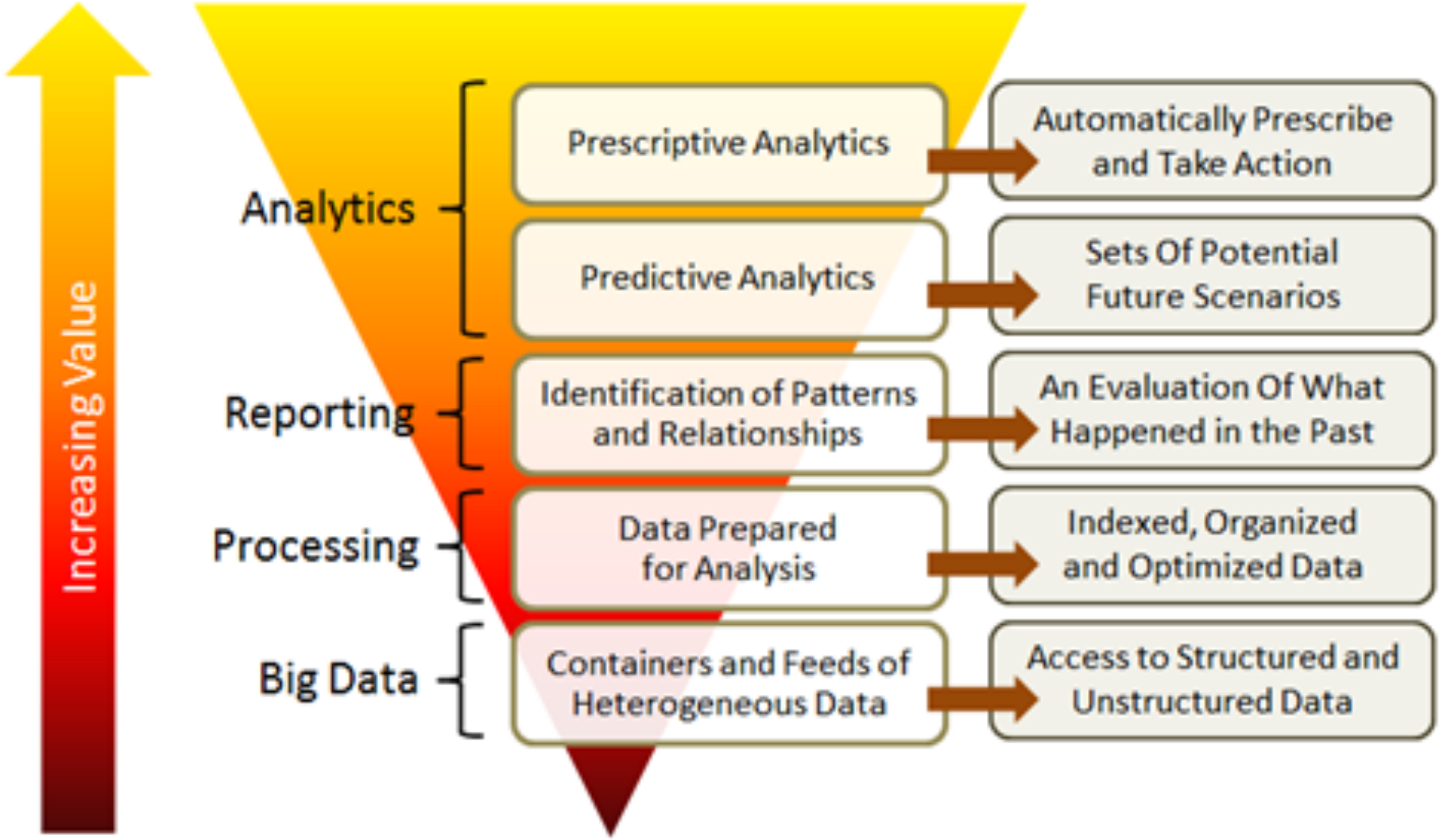
Regression

- 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

Source: Cap Gemini Blog, May 27, 2011

Another look at Analytics

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



2014 IEEE International Conference on Big Data (IEEE BigData 2014) Call for papers in the following (consolidated) areas:	What V's do the call for papers address:			
	Volume	**Velocity**	**Variety**	**Veracity**
1. Big Data Science and Foundations				
a. Novel Theoretical Models for Big Data	√	√	√	√
b. New Computational Models for Big Data	√	√		
c. Data and Information Quality for Big Data	√			√
d. New Data Standards			√	√
2. Big Data Infrastructure				
a. High Performance/Parallel/Cloud/Grid/Stream Computing for Big Data	√	√		
b. Autonomic Computing and Cyber-infrastructure, System Architectures, Design and Deployment	√	√		
c. Programming Models, Techniques, and Environments for Cluster, Cloud, and Grid Computing to Support Big Data	√			
d. Big Data Open Platforms	√	√		
e. New Programming Models and Software Systems for Big Data beyond Hadoop/MapReduce, STORM	√		√	

2014 IEEE International Conference on Big Data (IEEE BigData 2014) Call for papers in the following (consolidated) areas:	What V's do the call for papers address:			
	Volume	**Velocity**	**Variety**	**Veracity**
3. Big Data Management				
a. Algorithms, Architectures, and Systems for Big Data Web Search and Mining of variety of data.		√	√	√
b. Algorithms, Architectures, and Systems for Big Data Distributed Search	√	√	√	
c. Data Acquisition, Integration, Cleaning, and Best Practices			√	√
d. Visualization Analytics for Big Data			√	√
e. Computational Modeling and Data Integration	√		√	
f. Large-scale Recommendation Systems and Social Media Systems			√	√
g. Cloud/Grid/Stream (Semantic-based) Data Mining and Pre-processing- Big Velocity Data		√	√	√
h. Multimedia and Multi-structured Data- Big Variety Data			√	

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.

Analytics Master's Degrees Programs