Preparing Earth Data Scientists for 'The Sexiest Job of the 21st Century'*

*'Data Scientist: The Sexiest Job of the 21st Century', Thomas Davenport, Harvard Business Review, October, 2012

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

What does it take to be an Earth Data Scientist and apply Earth science data analytics to glean knowledge from data

Steven Kempler
NASA Goddard Earth Science Data and Information Services Center(GES DISC)
Steven.J.Kempler@nasa.gov

Evolution of Data Analysis

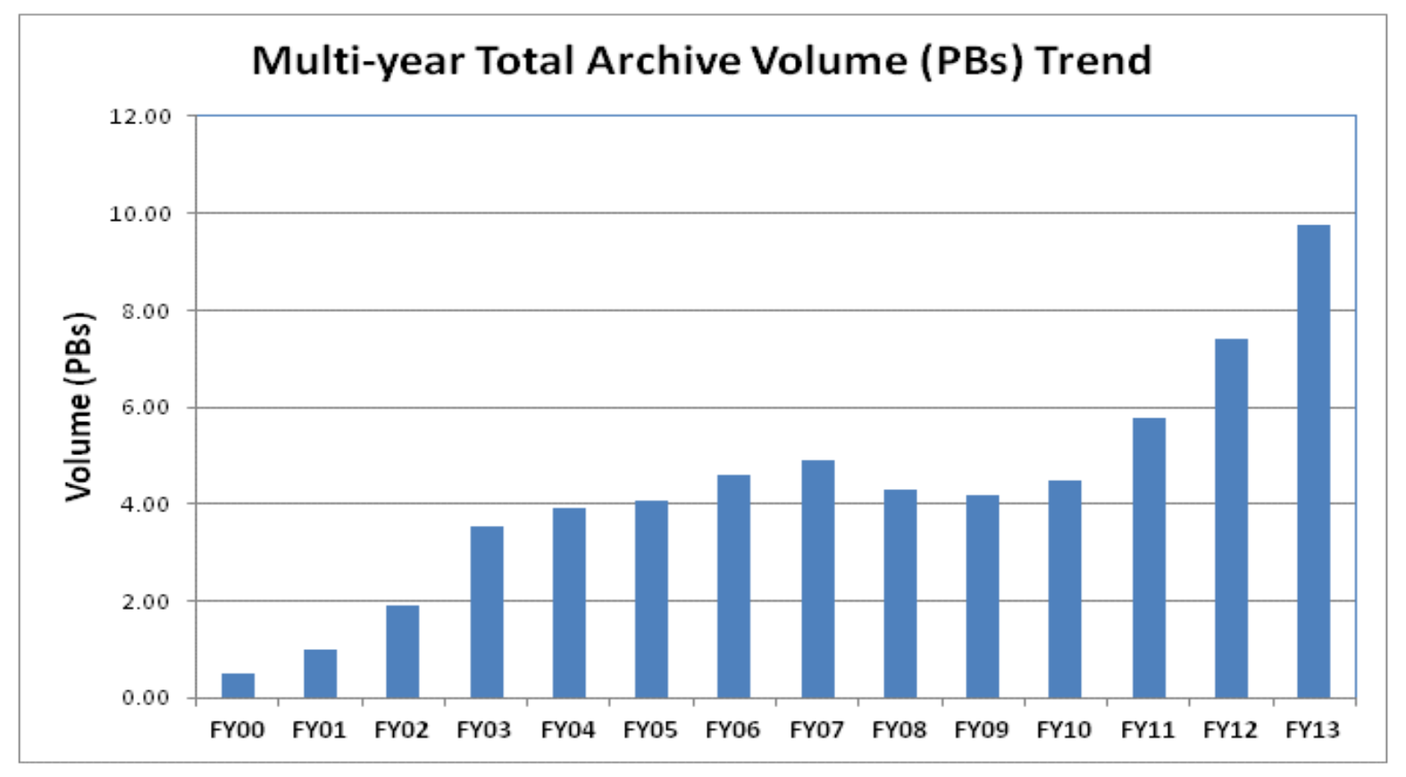
- Thousand years ago:
  - science was **empirical;** *describing natural phenomena*
- Last few hundred years:
  - **theoretical** branch; *using models, generalizations*
- Last few decades:
  - a **computational** branch; *simulating complex phenomena*
- Today:
  - **data exploration;** *unify theory, experiment, and simulation*
    - Data captured by instruments or generated by simulator
    - Processed by software
    - Information/knowledge stored in computer
    - Scientist analyzes database/files using data management and statistics

(The Fourth Paradigm: Data-Intensive Scientific Discovery, Edited by: Tony Hey, et al, 2009)

A Growing Profession

"Computer and Information Research Scientists is projected to grow 15 percent from 2012 to 2022. Computer scientists are likely to enjoy excellent job prospects, because many companies report difficulties finding these highly skilled workers".
(http://www.bls.gov/ooh/computer-and-Information-technology/computer-and-information-research-scientists.htm)

NASA Earth Science Archive Growth



Data Analytics Use Cases

TRMM 3B42 Precipitation Product

The 3B42 algorithm produces TRMM-adjusted merged-infrared (IR) precipitation and root-mean-square (RMS) precipitation-error estimates. (product developer: George Huffman)

Step 1 – Uses TRMM VIRS and TMI orbit data and monthly TMI/TRMM Combined Instrument (TCI) calibration parameters to produce monthly IR calibration parameters.

Step 2 – Uses these derived monthly IR calibration parameters to adjust the **merged-IR precipitation data, which consists of GMS, GOES-E, GOES-W, Meteosat-7, Meteosat-5, and NOAA-12 data**.

The final gridded, adjusted merged-IR product and RMS estimates:

- *Daily temporal resolution; 0.25 by 0.25-degree spatial resolution; Spatial coverage extends from 50 south to 50 north latitude.*

(http://disc.sci.gsfc.nasa.gov/precipitation/documentation/TRMM_README/TRMM_3B42_readme.shtml)

Producing a New Global Particulate Dataset

From David Lary's AGU 2013 Presentation: 'Using Multiple Big Datasets and Machine Learning to Produce a New Global Particulate Dataset'

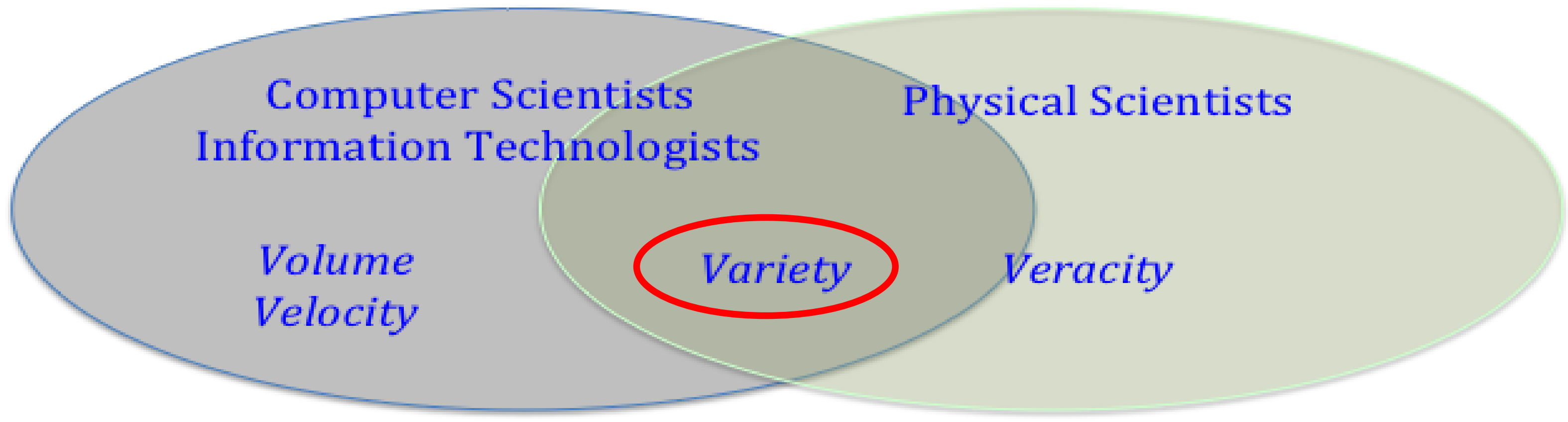
- Over 40 datasets from several satellites, high-resolution global meteorological data, social media and in-situ **observations are combined using machine learning on a distributed cluster using an automated workflow.**
- The global particulate dataset is relevant to global public health studies and would not be possible to produce without the use of the multiple big datasets, in-situ data and **machine learning**.

(Source: AGU Fall Meeting 2013, IN21D-06 Abstract)

*"Big Data" is an umbrella term coined by Doug McLaney and IBM to denote data posing problems, summarized as the **four Vs** (exemplified by **NASA Earth Science data metrics**):*

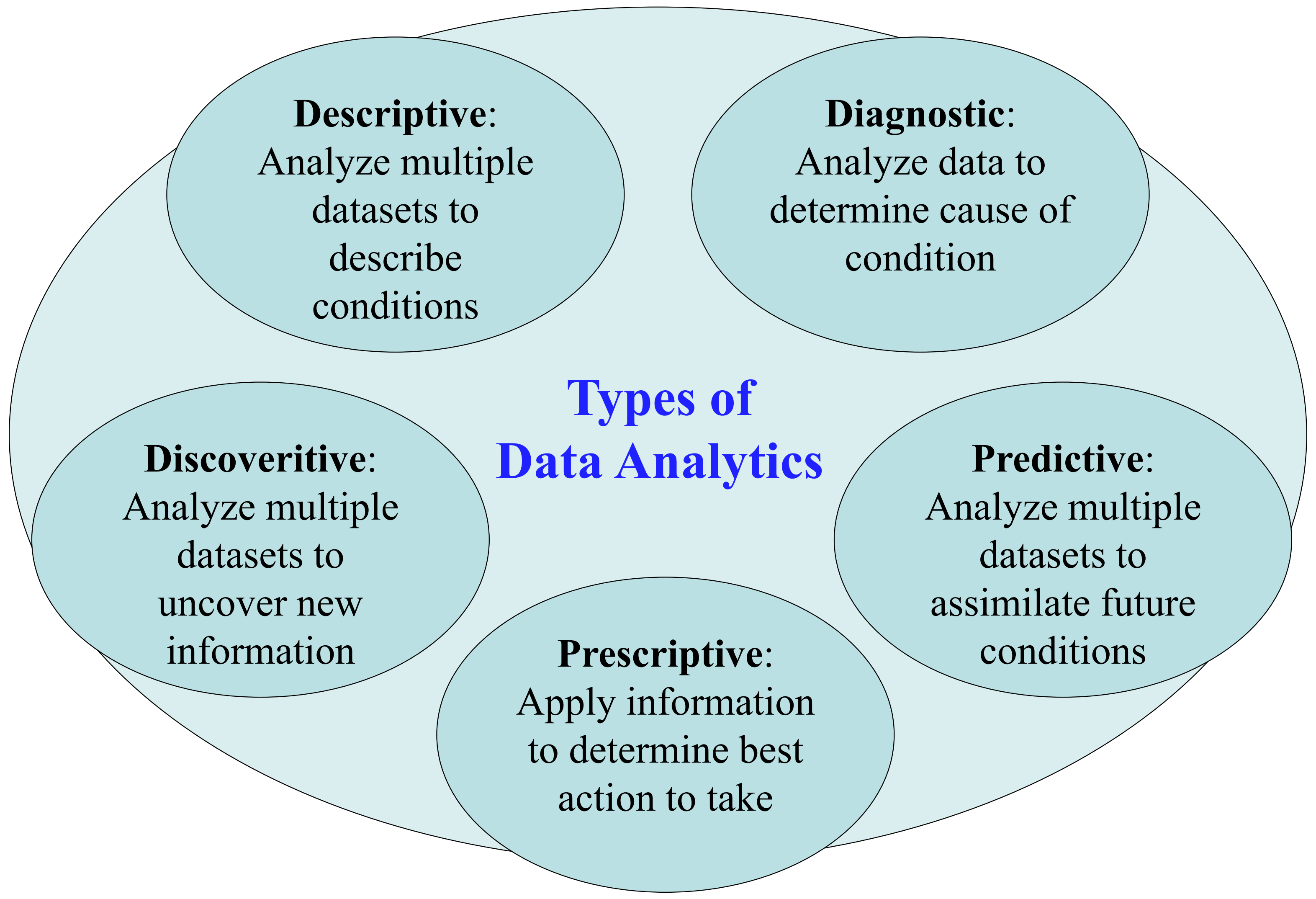
- ***Volume** – the sheer size of "data at rest" - **9.8 PB data** archived*
- ***Velocity** – the speed of new data ("data at move") - **22 TB/day** distributed*
- ***Variety** - the manifold different* - **6,861** Unique Datasets
- ***Veracity** – trustworthiness and issues of provenance* - Ongoing data comparison & refinement analysis **continuously improves** trustworthiness

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


Computer Scientists
Information Technologists
Physical Scientists
Volume
Velocity
Variety
Veracity

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)

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. (http://en.wikipedia.org/wiki/Analytics)


Types of Data Analytics

Descriptive: Analyze multiple datasets to describe conditions

Diagnostic: Analyze data to determine cause of condition

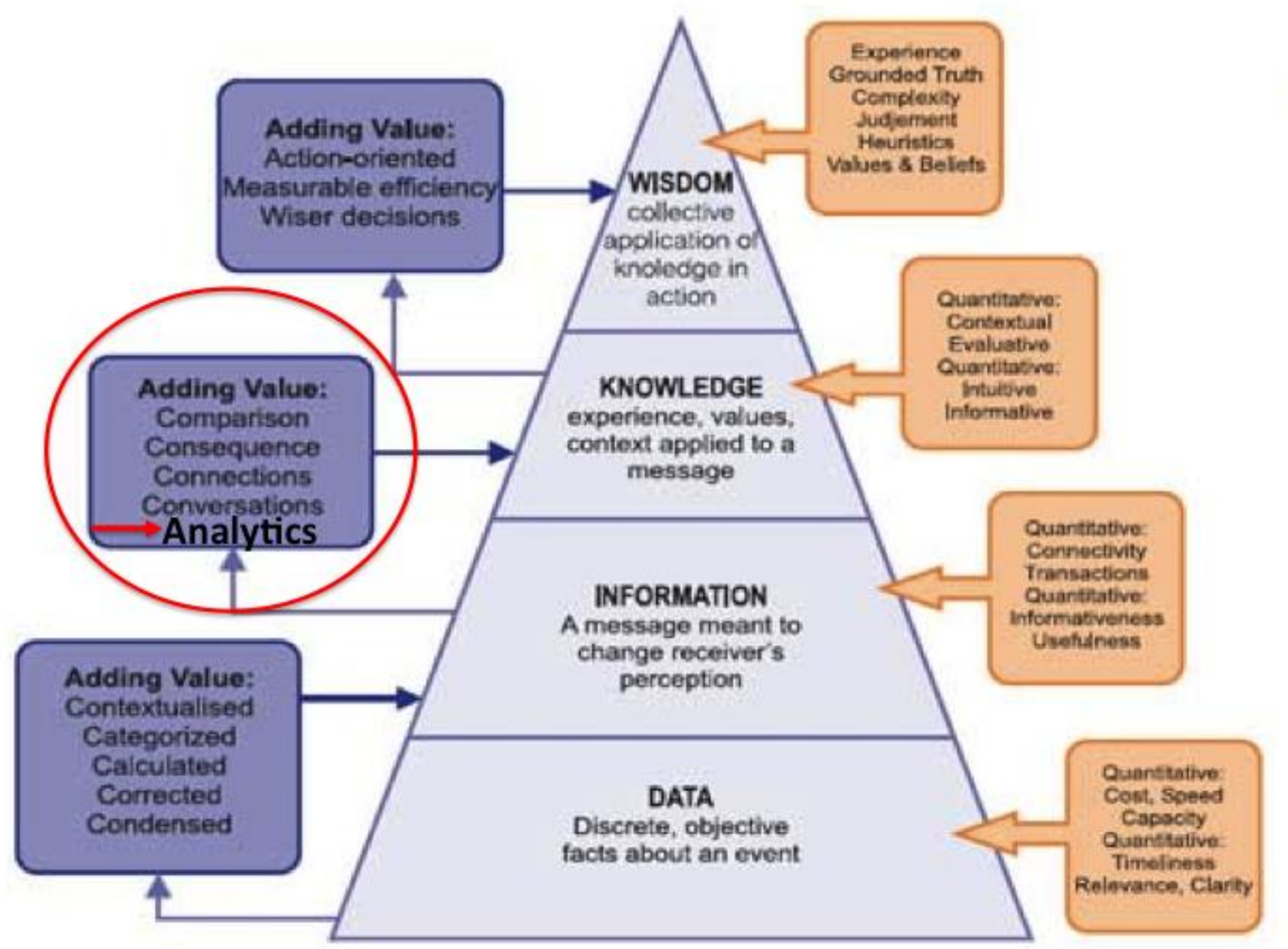
Discoveritive: Analyze multiple datasets to uncover new information

Predictive: Analyze multiple datasets to assimilate future conditions

Prescriptive: Apply information to determine best action to take

Summary

- We need to understand, **train, and inspire future generations** of Data (Information) Scientists to further maximize the value of large volumes of heterogeneous data
- We need to understand and **develop promising data analytics** techniques and technologies to facilitate science


Adding Value: Action-oriented Measurable efficiency Wiser decisions

Adding Value: Comparison Consequence Connections Conversations Analytics

Adding Value: Contextualised Categorized Calculated Corrected Condensed

WISDOM collective application of knoledge 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 Judjement Heuristics Values & Beliefs

Quantitative: Contextual Evaluative Quantitative: Intuitive Informative

Quantitative: Connectivity Transactions Quantitative: Informativeness Usefulness

Quantitative: Cost, Speed Capacity Quantitative: Timeliness Relevance, Clarity

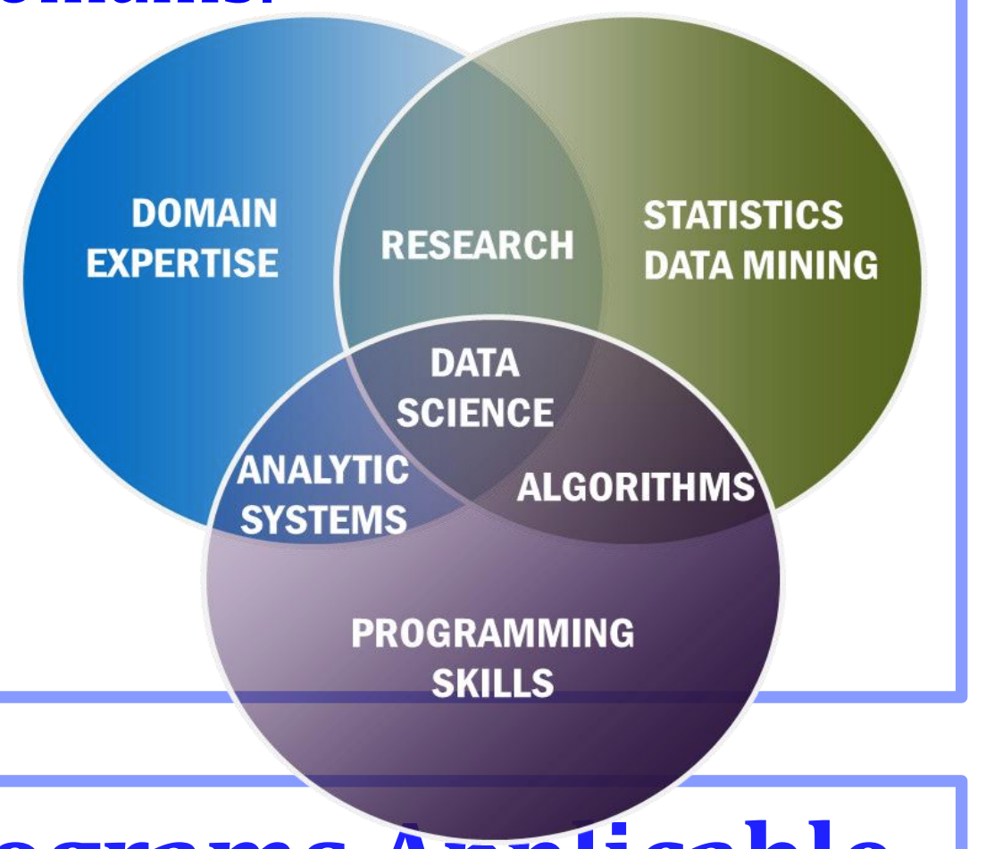
Skills of a Data Scientist

Expert in data analytics tools and techniques: e.g. rule learning, classification, cluster analysis, data fusion, ensemble learning machine learning, neural networks, anomaly detection, predictive modeling, time series analysis, visualization

Knowledge in particular science domains where data analytics can advance our understanding of science

Woodstock, 2013 AGU IN43A-1638: The role of the Data Scientist is a **hybrid one**… skills to support domain scientists with data and computational needs and communicate across domains.

Evans, 2013 AGU IN43A-1641: Data Scientist, who has a greater capacity in mathematical, numerical modeling, statistics, computational skills, software engineering and spatial skills and the ability to integrate data across multiple domains



Top Big Data & Analytics Masters Programs Applicable to Earth Science Data Scientist/Analytics

Caltech Virtual Summer School
DePaul University
George Mason University
George Washington University
Harrisburg University
Harvard University
Illinois Institute of Technology
Indiana University
Louisiana State University

NC State University
Northwestern University
Rensselaer Polytechnic Institute
Rutgers University
Texas A&M
University of Connecticut
University of Illinois
University of Maryland
University of San Francisco
University of Virginia

NO endorsements intended
(bisoftwareinsight.com/resource-center/big-data-analytics-programs/)
((http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/)

Acknowledgements: Thanks to the efforts of Tiffany Matthews, ASDC, and the ESIP Federation Earth Science Data Analytics Cluster