How Do We Facilitate the Use of Large Amounts of Heterogeneous Data

Steve Kempler
Steven.J.Kempler@nasa.gov
Codes 586/610.2
March 12, 2014

Information Science & Technology Colloquium Series

Is Gleaning Knowledge from Information in Reach?

Steve Kempler
Steven.J.Kempler@nasa.gov
Codes 586/610.2
March 12, 2014

Information Science & Technology Colloquium Series

How Do We Facilitate the Use of Large Amounts of Heterogeneous Data

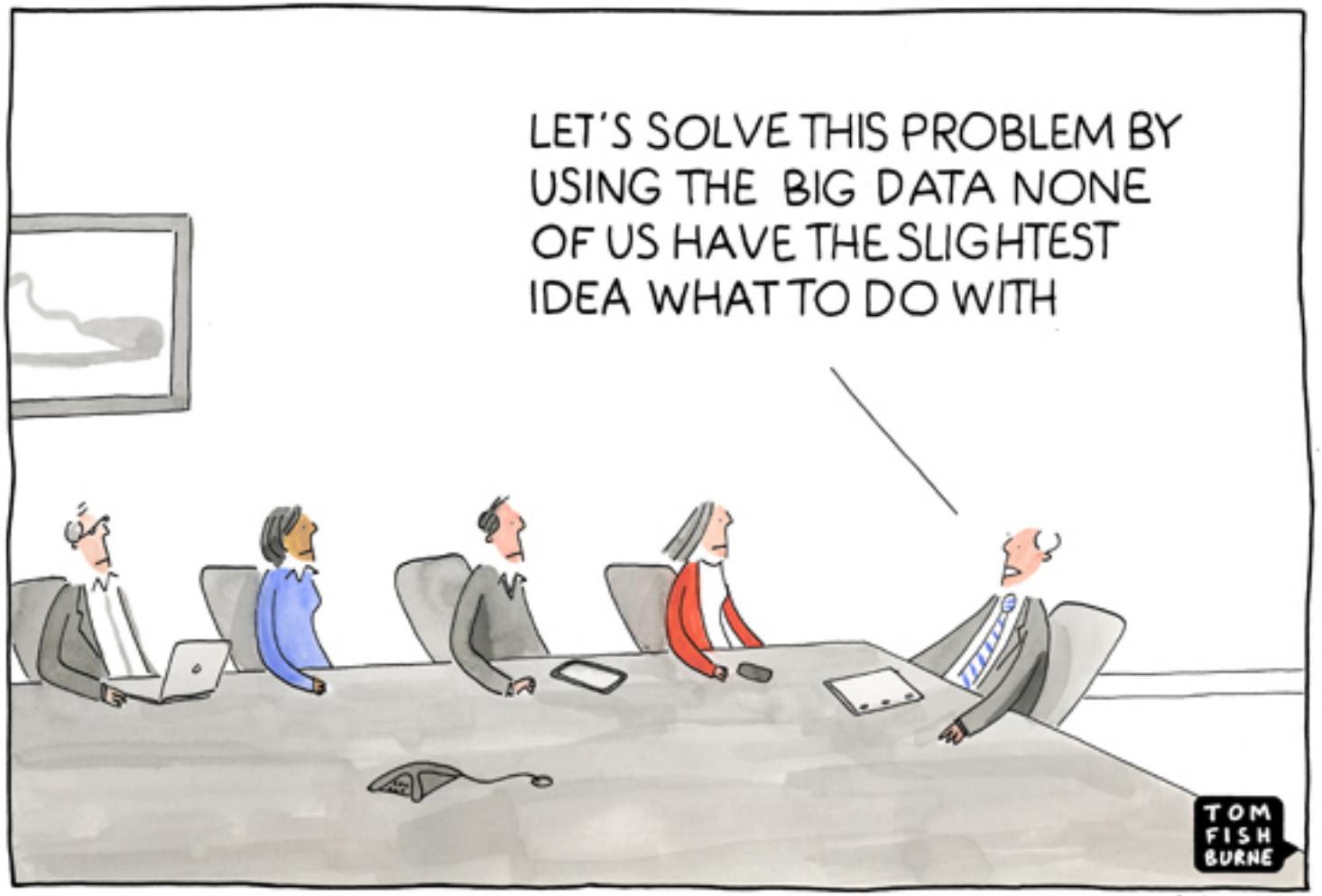
or…

Is Gleaning Knowledge from Information in Reach?

Steve Kempler
Steven.J.Kempler@nasa.gov
Codes 586/610.2

March 12, 2014

Information Science & Technology Colloquium Series



© marketoonist.com

http://dilbert.com/strips/comic/2007-05-16/

bit.ly/WXIWwe (Mar 21, 2013 cartoon)

But, what we really want to know is how we can maximize the use of (increasing amounts of heterogeneous) data?

From the Fourth Paradigm

(Editors: Hey, Tansley,Tolle)

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

"Based on data-intensive computing… we are at a stage of development that is analogous to when the printing press was invented.

Printing took a thousand years to develop and evolve into the many forms it takes today. Using computers to gain understanding from data created and stored in our electronic data stores will likely take decades—or less." (P. xi)

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)

Science Paradigms (from The Fourth Paradigm)

- 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

Today's Agenda

- Defining Terms
- Sorting Through All the Activities
- Evolving to Multiple Diverse Dataset Analysis
- Techniques and Technologies: At Least How it Looks Today
- Developing Skills for the Future

Let's Start the Top: What is Big Data

- An Adjective: Big Data methodology
- A Phenomenon: Rapid acceleration of increasing data volume….
- An Architecture: Solved by Big Data
- A Paradigm: The shift to Big Data analytics
- A Movement: Glean intelligence from data
- A Term: To describe the voluminous amount of unstructured and semi-structured data that would take too much time and cost too much money to load into a relational database for analysis.

  (http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data)

The next frontier is learning how to manage Big Data throughout its entire lifecycle.

(http://www.forbes.com/sites/ciocentral/2012/07/05/best-practices-for-managing-big-data/)

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

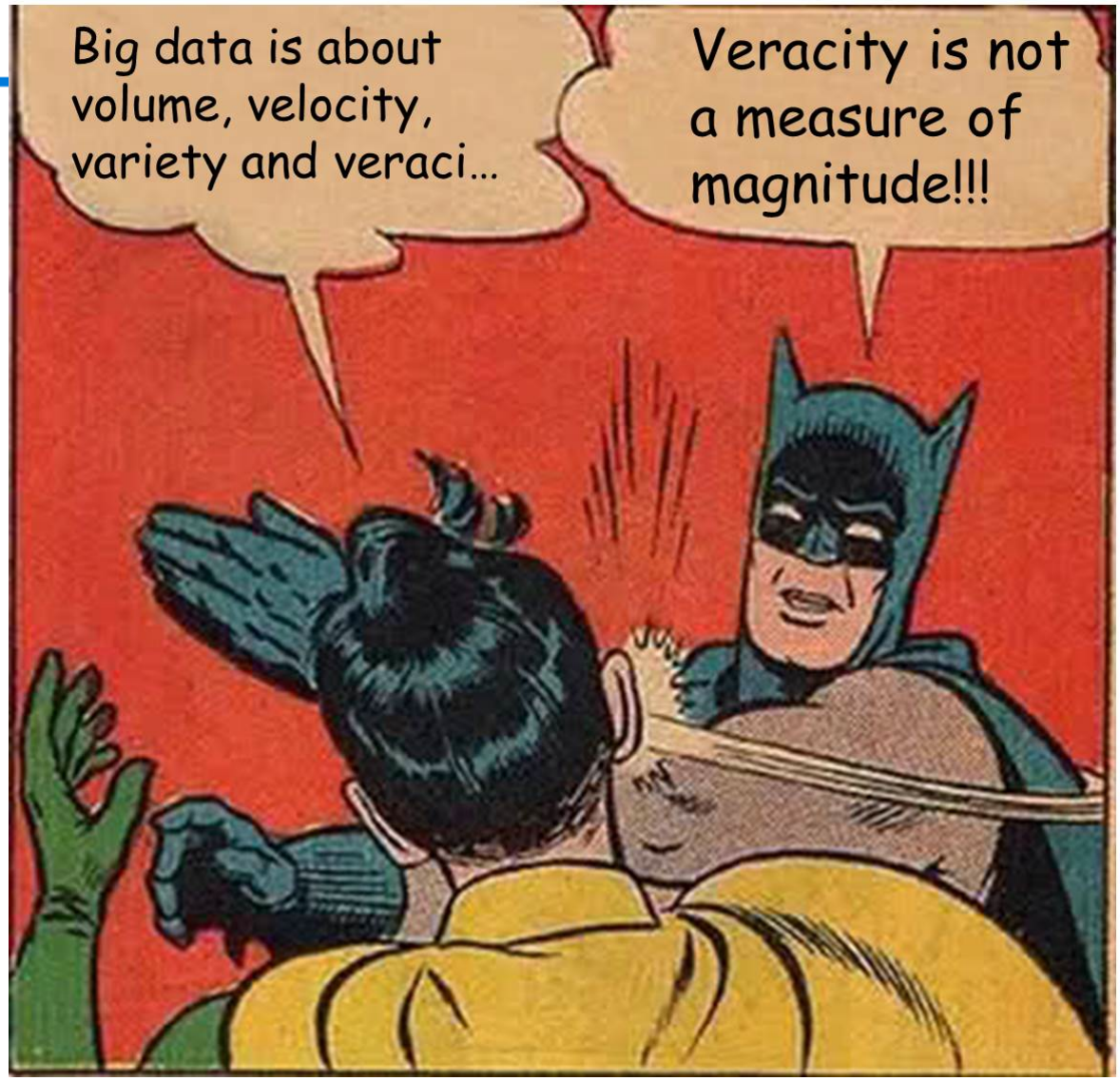
*"**Big Data**" is an umbrella term coined by Doug McLaney and IBM several years ago to denote data posing problems, summarized as the **four Vs**:*

- ***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*

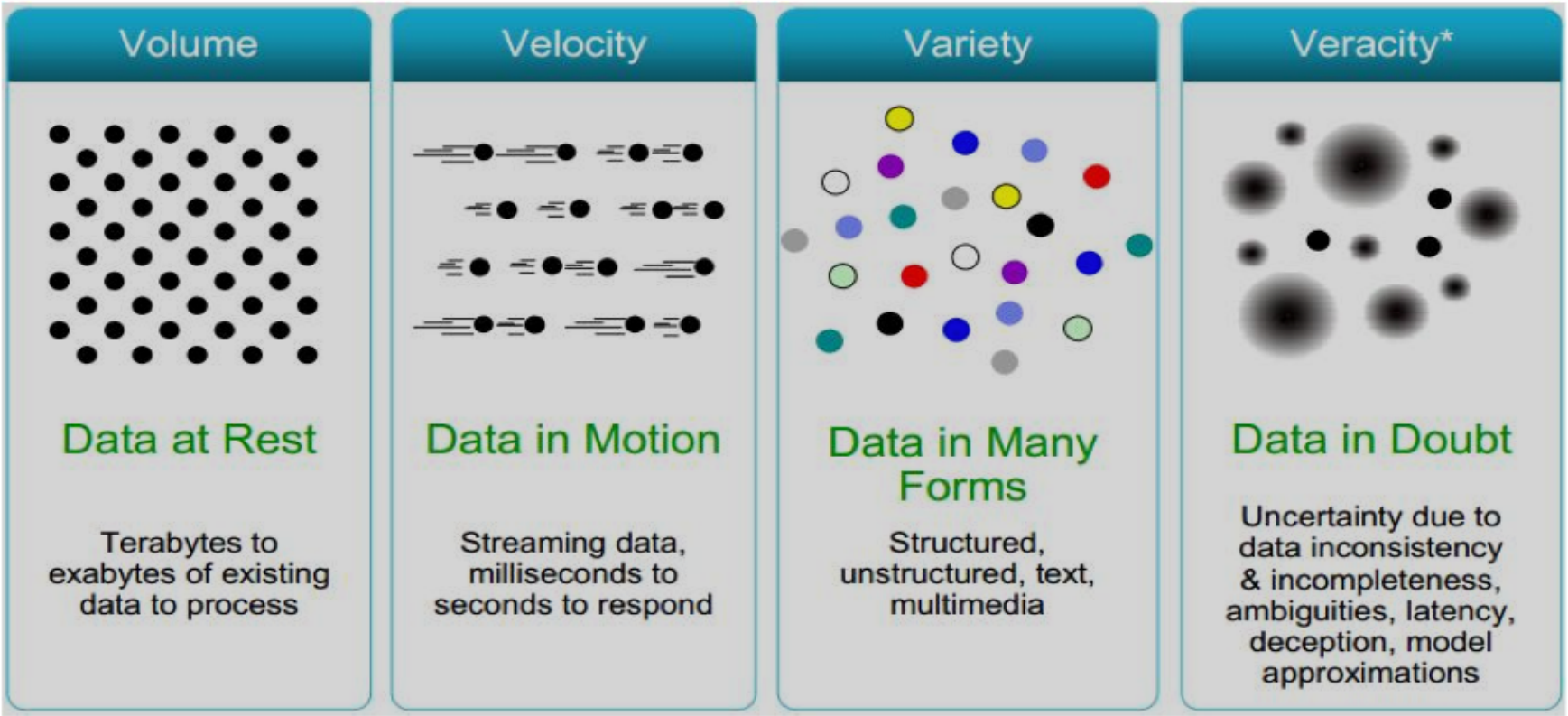


Thank you Doug Laney… whatever…
http://blogs.gartner.com/doug-laney/

From: Demystifying Data Science

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

4 V's of Big Data



IBM, 2012

Volume, Velocity, Variety, Veracity and NASA Earth Science Data

(EOSDIS FY2013 Annual Metrics Report)

- Volume
  - 9.8 PB
  - 9,800,000,000,000,000 bytes
- Velocity
  - 22 TB/day distributed, but…
  - … 2.3 TB/day Near Real Time, distributed within hours
- Variety
  - 6,861 Unique Datasets
  - Not to mention non-NASA datasets that can be co-analyzed
- Veracity
  - Metadata provides indicators to describe data quality
  - Ongoing data comparison and refinement analysis continuously improves data trustworthiness

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

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.

Big Data Activities - IEEE

Call for Papers: 2014 IEEE International Conference on Big Data (IEEE BigData 2014), http://cci.drexel.edu/bigdata/bigdata2014/index.htm,
October 27-30, 2014, Washington DC

Big Data Topics

1. Big Data Foundation
2. Big Data Infrastructure
3. Big Data Management
4. Big Data Search and Mining
5. Big Data Security & Privacy
6. Big Data Applications

Big Data Activities – OGC

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

BigData.DWG (Working Group) Goals (still evolving):

The OGC Big Data WG will specifically focus on spatio-temporal data, in line with OGC's mission. With the same inaccuracy as the term "Big Data" itself, we give them the working title "Big Earth Data" for now

Big Data Activities – NIST

(http://bigdatawg.nist.gov/home.php)

Deliverables:

- Develop Big Data Definitions
- Develop Big Data Taxonomies
- Develop Big Data Requirements
- Develop Big Data Security and Privacy Requirements
- Develop Big Data Security and Privacy Reference Architectures
- Develop Big Data Reference Architectures
- Develop Big Data Technology Roadmap

Big Data Activities – Research Data Sharing (RDA)

(https://rd-alliance.org/groups/big-data-analytics-ig/wiki/big-data-analytics-interest-group-charter.html)

Some objectives of the group are:

- BDA (Big Data Analytics) will aim to clarify some foundational terminologies …
- BDA will develop recommendation documents with a systematic classification of feasible combinations of analysis algorithms, analytical tools, data and resource characteristics and scientific queries.
- BDA will collaborate with external bodies and initiatives - such as NIST, OGC, ISO, EarthServer and others …

Big Data Activities – Earth Sciences Information Partners (ESIP) (http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics)

Goals are still evolving, but overall:

This cluster aims to explore the possibilities, interest, involvement, experience, and understanding that Partners are willing to bring to the table to enhance and foster the next evolutionary phase of enabling science research.

And many, many more…

Private and public organizations that sponsor special interest groups to promote and share techniques, technologies, and knowledge to address the growing amounts of heterogeneous datasets

It is Really About the V's

So, if you just look at 'Big Data' problem, it can indeed be overwhelming.

But, the provision and acquisition strategies of any particular data, that allows for their maximum use, are unique to that data and its user.

Thus, the problem really is the V's:

How to address the 'V'(s) that stands in the way of maximizing the use of particular data.

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			√	

Why Should NASA Earth Science be Concerned?

From a 'to advance science' point of view:

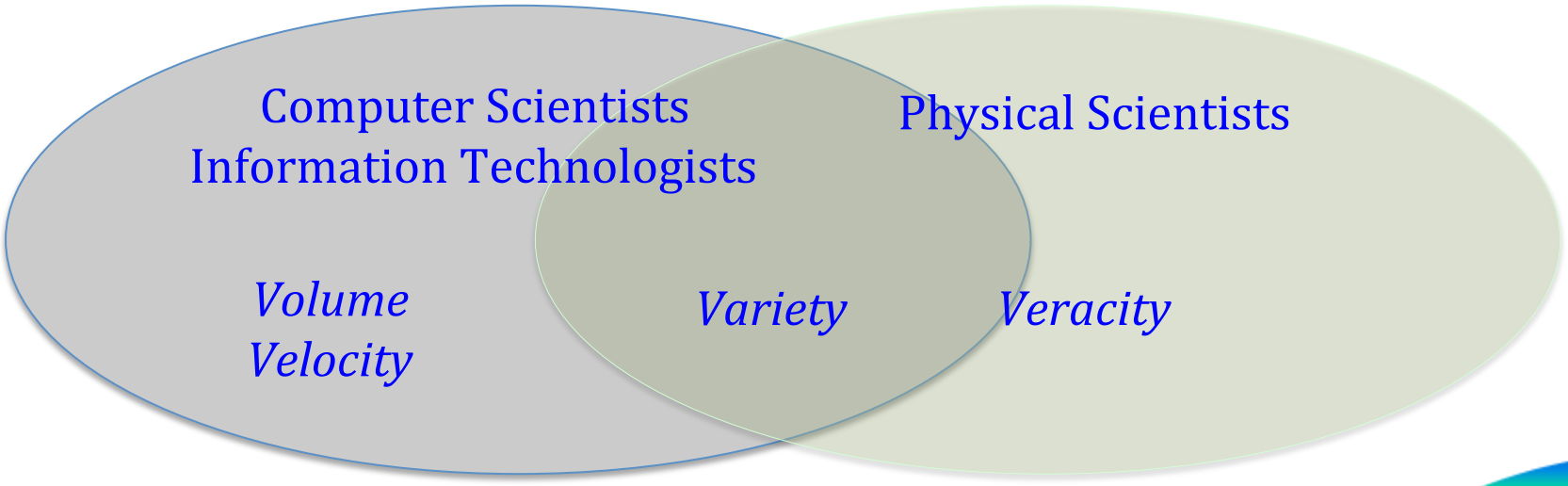
On the continuum of ever evolving data management systems, we need to understand and develop ways that allow for the **variety** of data relationships to be examined, and information to be manipulated, such that knowledge can be enhanced, to facilitate science.

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)

Data Analytics Example - Sabermetric Research

What is sabermetrics? *Originally defined by Bill James in 1980, sabermetrics is "the search for objective knowledge about baseball". (http://sabr.org/sabermetrics)*

In baseball, Runs Batted In (RBIs) is a tried and true baseball statistic showing a players team worth. However, additional information is revealed when applying data that yields the number of opportunities a player has to drive in runs.

Table 1 – NL RBI leaders, 2013

P Goldschmidt	125
F Freeman	109
J Bruce	109
B Phillips	103
P Alvarez	100
A Gonzalez	100
H Pence	99
A Craig	97
J Votto	94
M Holliday	94

Table 2 – Goldschmidt batting by situation, with effective outs

	AB	H	GIDP	SF	total outs	outs * {runners + batter}	AV
bases empty	329	87	0	0	242	242	.26
one man on	191	67	13	2	139	278	.35
two men on	68	23	11	2	58	174	.33
sacks full	14	5	1	1	11	44	.35
total	602	182	25	5	450	738	.30

Player	RBI	effective outs	ratio*
P Goldschmidt	125	738	.169
F Freeman	109	611	.176
J Bruce	109	793	.137

(Source: By the Numbers Volume 24, Number 1 The Newsletter of the SABR Statistical Analysis Committee February, 2014, abr.org/research/statistical-analysis-research-committee-newsletters)

Data Analytics Example – TRMM 3B42

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 the 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-degree by 0.25-degree spatial resolution*
- *Spatial coverage extends from 50 degrees south to 50 degrees north latitude.*

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

Data Analytics Example - 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: A Technology Challenge Case Study

- 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)

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

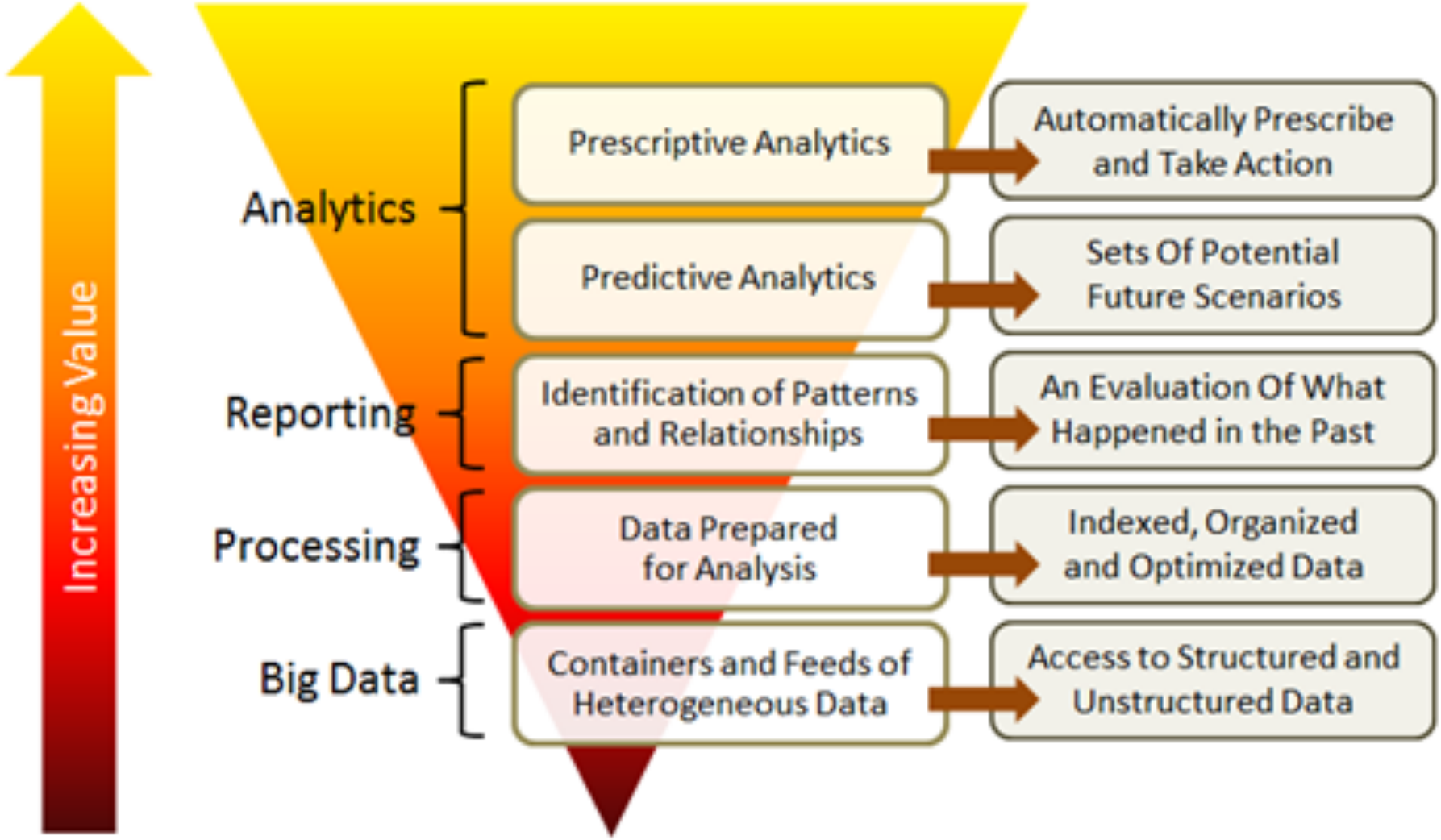
Regression

- 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

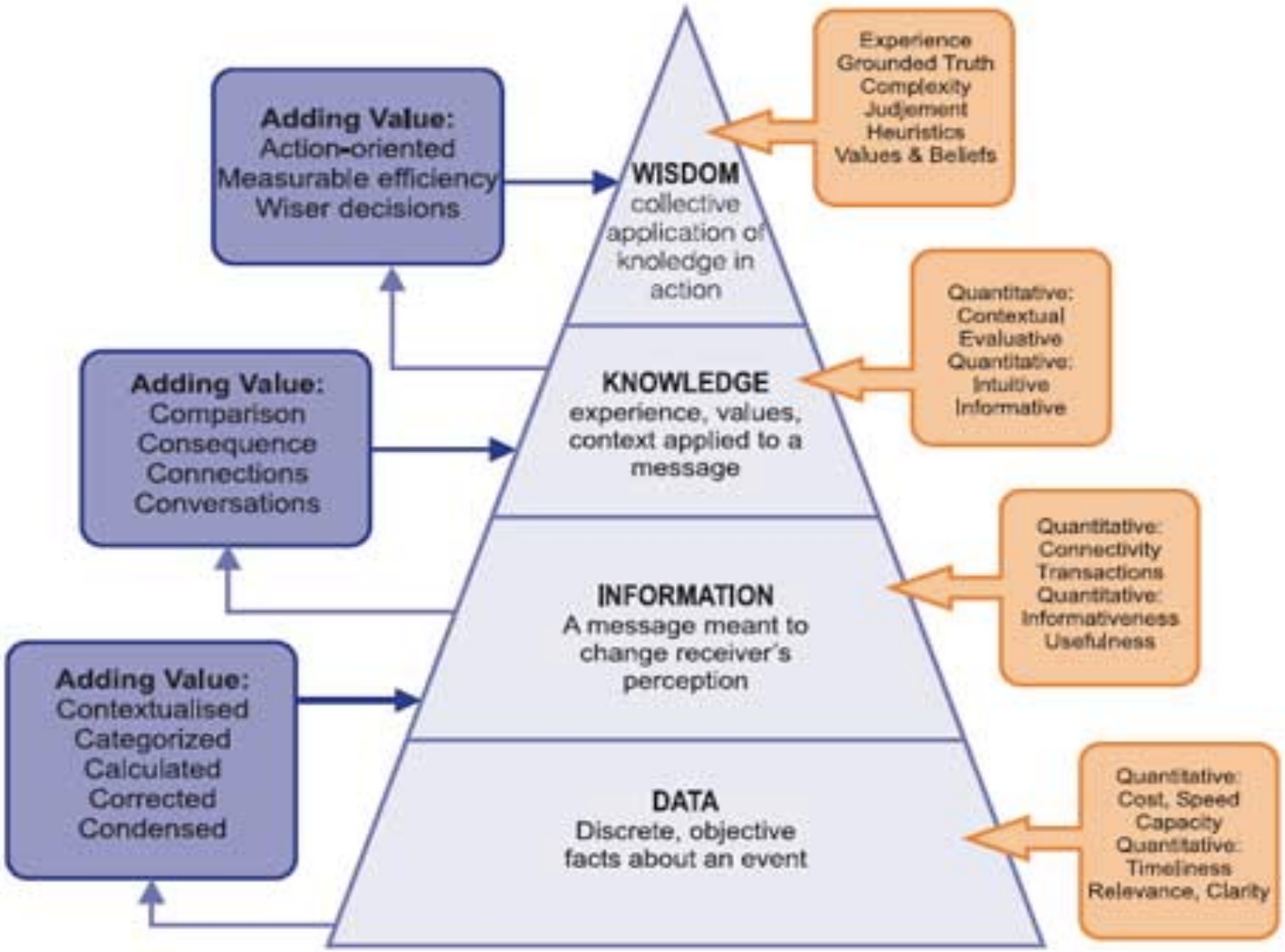
Analytics Value Chain

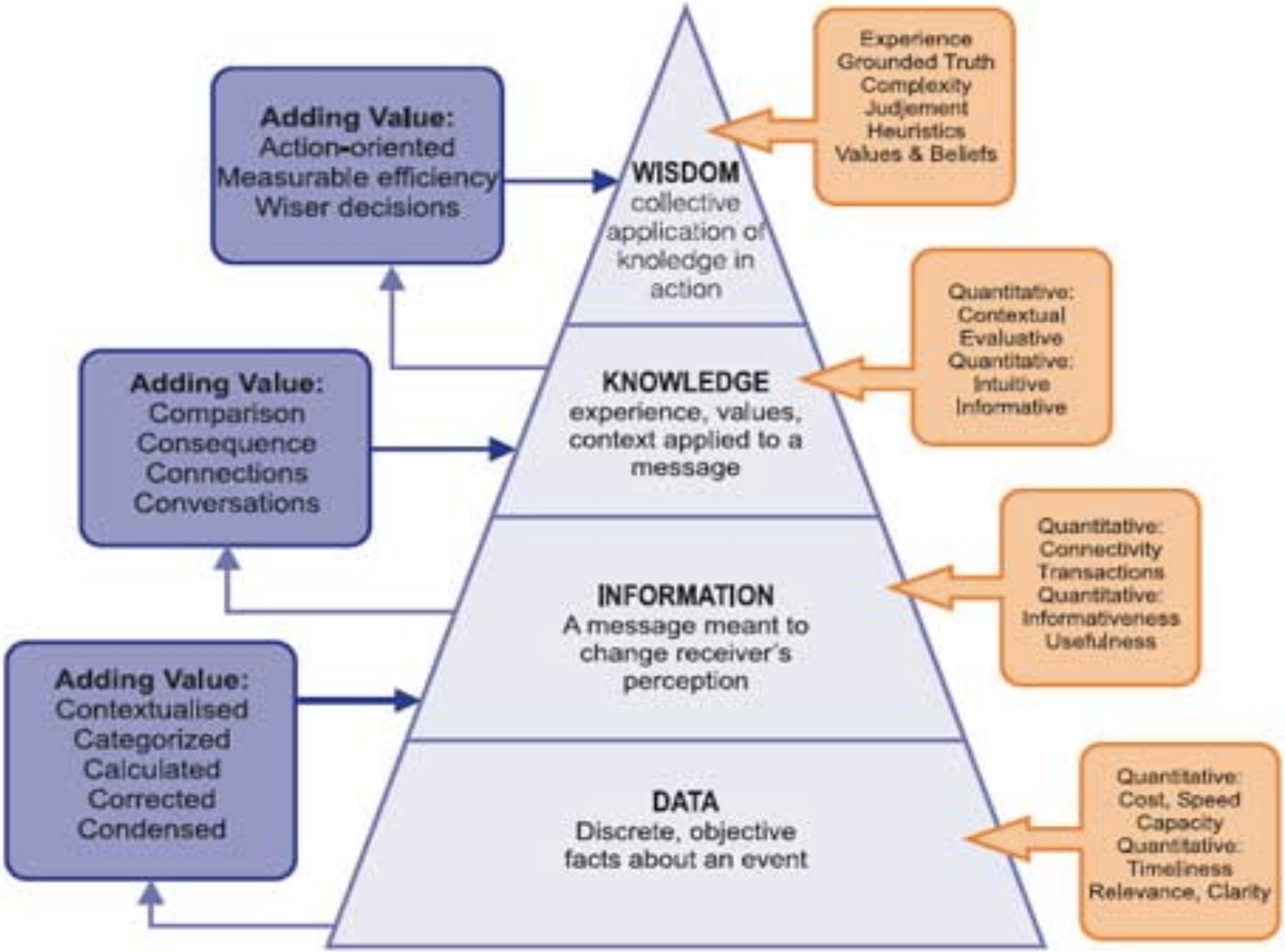
(http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/)

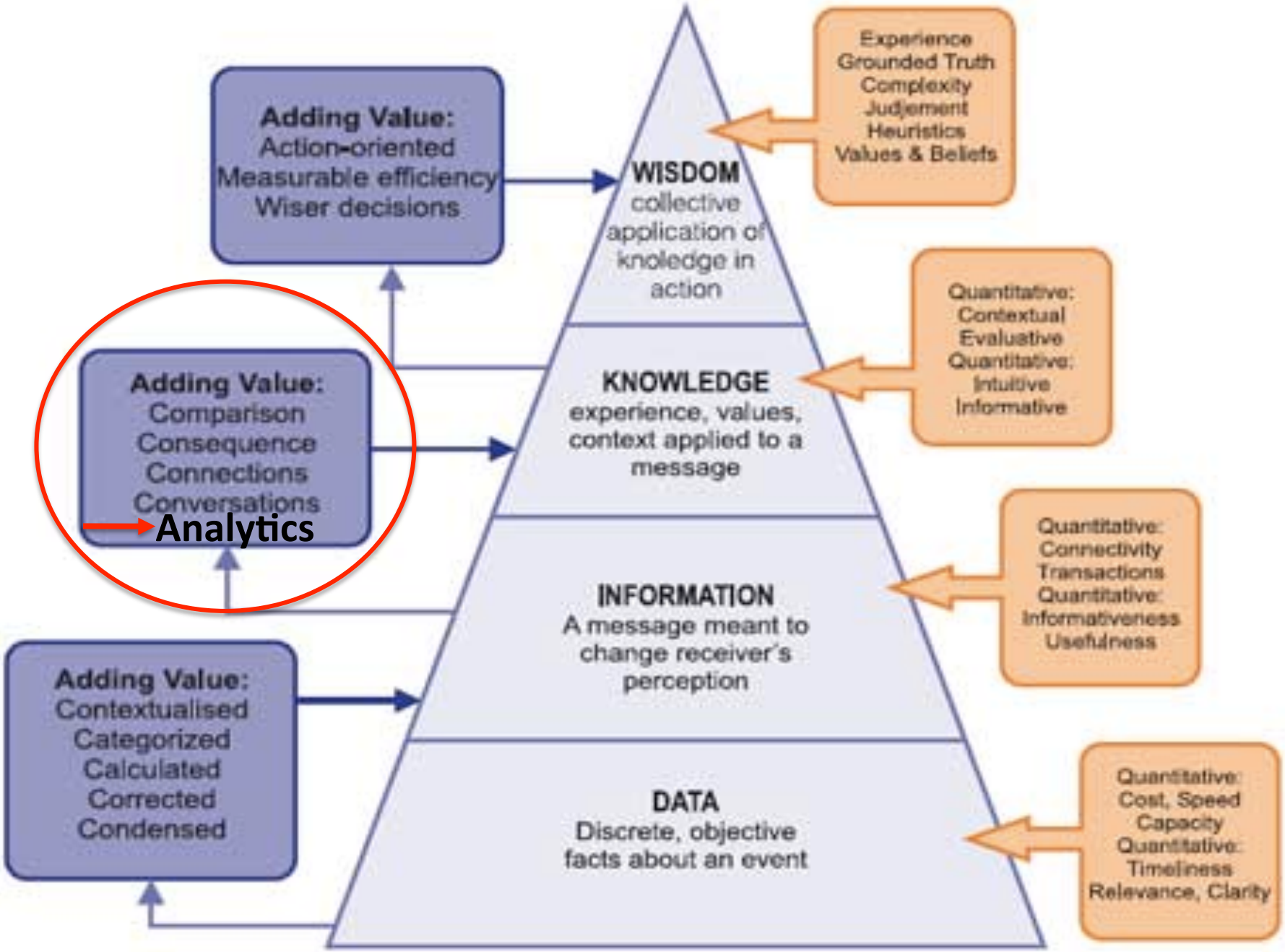


DIKW Pyramid

(http://km4meu.wordpress.com/2010/02/06/settling-the-eternal-semantic-debate-what-is-knowledge-what-is-information/)

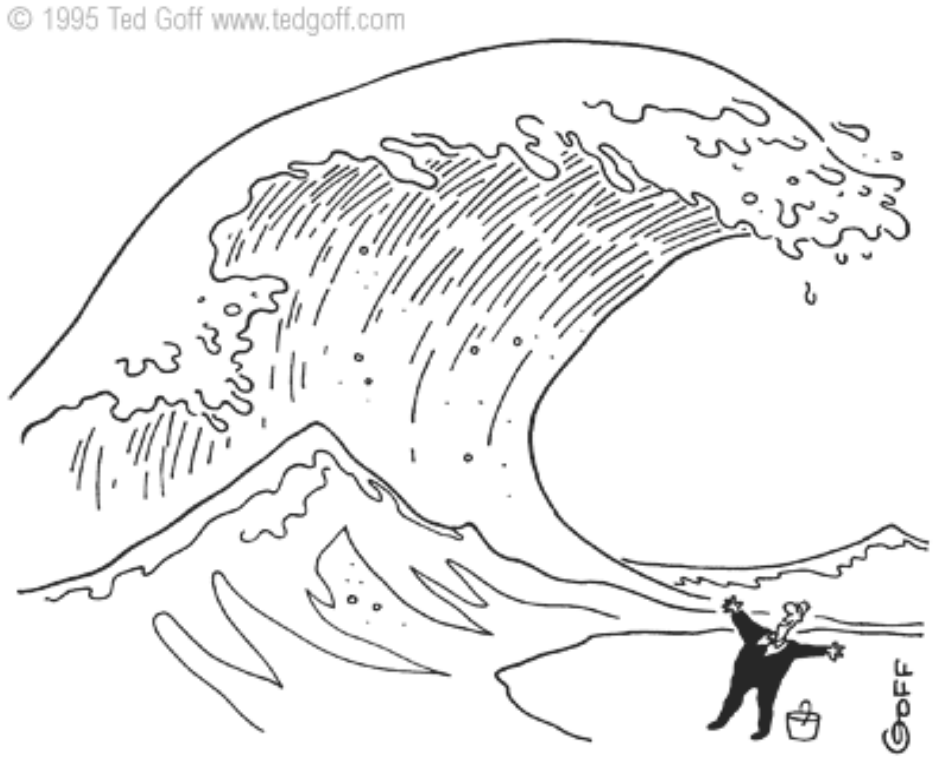



Adding Value:
Action-oriented
Measurable efficiency
Wiser decisions
Adding Value:
Comparison
Consequence
Connections
Conversations
Adding Value:
Contextualised
Categorized
Calculated
Corrected
Condensed
WISDOM
collective application of knoledge in action
KNOWLEDGE
experience, values, context applied 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


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

We are already seeing some attempts to infer knowledge based on the world's information.

(Fourth Paradigm, Parastatidis, P. 167)



"Eureka! More information!"

2011 McKinsey report suggests suitable analytics technologies to address…

(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.

What should NASA Earth science be thinking about to address Data Analytics?

- Data analytics technology. Today, these include NoSQL databases, DISC, Dryad, Hadoop and MapReduce
- Data and Information:
  - Making data/information transparent and usable
  - Collecting more accurate & detailed performance information
  - Addressing the narrowing segmentation of customers and therefore more precisely tailor products or services
  - Improving strategy decision-making
  - Improving & developing the next generation of products & services

What should NASA Earth science be thinking about to address Data Analytics?

- Learning the impacts of managing Heterogeneous Data, and being able to proactively react…
- … Determine specific areas that can benefit by having a Data Scientist, and specifying their qualifications

There is Much Work To Do

(Source: The Fourth Paradigm, Parastidis, P. 166)

- We are still a **long way from having computer systems that** can automatically discover, acquire, organize, analyze, correlate, interpret, infer, and reason over information that's on the Internet, that's hidden on researchers' hard drives, or that exists only in our brains.

- We do not yet have an infrastructure **capable of managing and processing knowledge on a global scale**, one that can act as the foundation for a generation of knowledge-driven services and applications.

- So, if the fourth paradigm is about data and information, it is **not unreasonable to foresee a future**, where we begin thinking about the challenges of managing knowledge and machine-based understanding on a very large scale.

We are Making Progress…

(Source: The Fourth Paradigm)

- Moving forward, the work of the **Semantic Web community** will continue to play a significant role in the interoperable exchange of information and knowledge. (Parastidis, P. 167)

- We are already seeing **attempts to infer knowledge based on the world's information.** Services such as OpenCyc, Freebase, Powerset, True Knowledge, and Wolfram|Alpha demonstrate how facts can be recorded … be combined and made available **…** (Parastidis, P. 167)

- The myGrid and myExperiment projects demonstrate benefits of **capturing and sharing, in a semantically rich way, the definitions of workflows** in science. (Parastidis, P. 168)

- The **workflows themselves**… can be generated and transformed dynamically to meet the requirements at hand**…** (Goble, P. 142)

But We Need to Develop Expertise

Attention is needed to develop Data Scientists who are:

- Experts in data analytics tools and techniques
- Knowledgeable in particular science domains where data analytics can advance our understanding of science

What Defines a Data Scientist?

One who has organizational Leadership and is learned in: Statistics, Data analysis, Probability, Algorithms for Data Science, Data Engineering, Machine Learning, Exploratory Data Analysis, Analytics

- Woodstock, AGU IN43A-1638: The role of the Data Scientist is a hybrid one... With the skills to support domain scientists with data and computational needs and communicate across domains, yet not quite able to do the domain science itself.

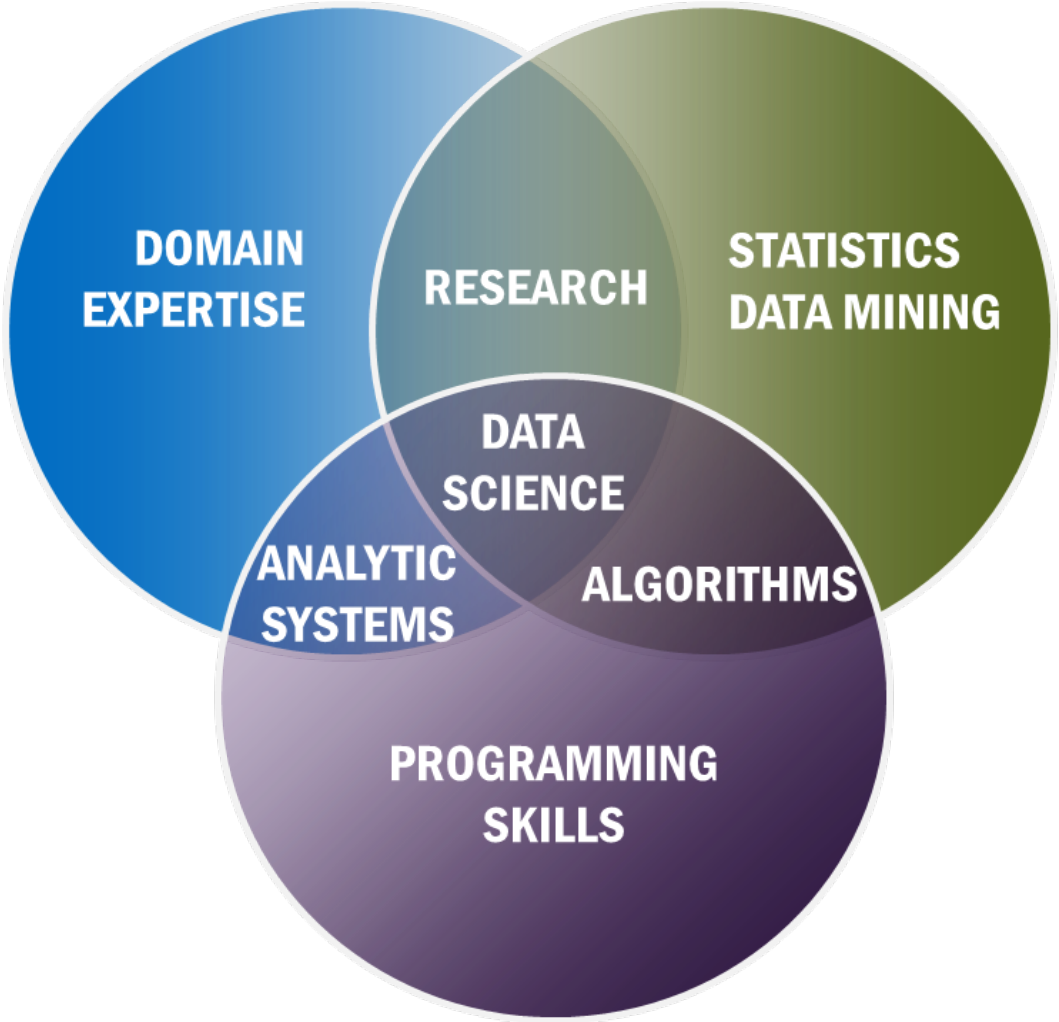
  Role: Provides access to unified technology (open standards); know seamless access, training scientists in IT, build community (with scientists)

What Defines a Data Scientist?

- Evans, 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.
- http://www.smartplanet.com/blog/bulletin/7-new-types-of-jobs-created-by-big-data/682: This emerging role is taking the lead in processing raw data and determining what types of analysis would deliver the best results.

"Science Now Takes a Community"

(Source: Big Data Effects on Weather and Climate, Informal Discussions on The New Economics, Nancy Grady, http://www.ametsoc.org/boardpges/cwce/docs/profiles/GradyNancyW/2013-04-AWF.pdf)



Big Data Analytics[1] and Data Scientist[2] Master's Degrees*

1. (http://www.informationweek.com/big-data/slideshows/big-data-analytics/big-data-analytics-masters-degrees-20/240145673)
2. (http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/#Arizona_State_University)

*** As it turns out, while data analytics and Data Science are still at their infancy, these two sources actually mostly point to the same universities, and the same programs.**

‘Top’ Universities with Big Data Analytics and Data Scientist Master's Degrees

1	Arizona State University
2	Bentley University
3	Carnegie Mellon University
4	Columbia University
5	DePaul University
6	Drexel University
7	Harvard University
8	Indiana Univeristy
9	Louisiana State University
10	Massachusetts Institute of Technology
11	Michigan State University
12	New York University
13	North Carolina State University

14	Northwestern University
15	Purdue University
16	Rutgers University
17	Stanford University
18	University of California, Berkeley
19	University of Cincinnati
20	University of Connecticut
21	University of Illinois
22	University of Maryland
23	University of Ottawa
24	University of San Francisco
25	University of Tennessee
26	York University

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Business analytics/ process analysis	√		√			√		√	√		√	√			√	√			√	√		√	√	√	√	√
Strategic marketing		√			√			√			√		√		√				√			√	√	√		
Data science							√											√				√				
Data analysis/ analytics							√				√		√	√							√	√				
Pedictive analytics/ modeling			√		√				√					√		√				√				√		
Information technology/ systems/ management		√	√											√			√	√	√	√	√		√	√		
Computer science				√			√						√			√	√									
Business management										√										√						
Data mining	√				√	√		√	√		√	√	√	√	√	√	√	√	√	√	√		√	√	√	
Statistics	√	√			√	√	√		√		√	√	√	√		√			√		√	√		√	√	√
Machine learning				√	√		√									√	√	√						√		
Data visualization			√	√			√						√	√				√								√
Decision support modeling								√																√		
Mathematical modeling						√	√						√								√					
Database processing					√								√			√	√						√			
Data management tools									√		√				√	√				√						
Healthcare						√			√					√												
Computational biology				√													√									

GES – DISC
Goddard Earth Sciences
Data Information Services Center

But What do These Programs Emphasize?

Business analytics/process analysis
Strategic marketing
Data science
Data analysis/analytics
Pedictive analytics/modeling
Information technology/systems/management
Computer science
Business management

19 - Business oriented
13 - Data Science/Analytics oriented
10 - Information systems oriented
5 - Computer science oriented
2 - Business management oriented

	B	D	I	C	M
Data mining	15	10	8	3	1
Statistics	14	10	5	4	0
Machine learning	3	5	3	3	0
Data visualization	3	6	3	2	0
Decision support modeling	2	1	1	0	0
Mathematical modeling	2	4	1	1	0
Database processing	4	4	2	2	0
Data management tools	5	4	1	1	1
Healthcare	2	2	1	0	0
Computational biology	0	0	1	2	0

Calls for Action

Highlights from Fox (The Fourth Paradigm, P. 151):

- Researchers in science must work with colleagues in computer science and informatics to develop field-specific requirements …
- Scientific and professional societies must provide the settings in which the needed rich interplay between science requirements and informatics capabilities can be realized…

Calls for Action

Highlights from Parastatidis (The Fourth Paradigm, P. 170-171):

- The process of combining information from existing scientific knowledge … including the specific methodologies that were followed to produce conclusions, should be automatic and implicitly supported…
- The interoperable exchange of information, whether representing facts or processes, is vital to successfully sharing knowledge.

From Kempler:

- Create academic curriculums that foster specific data science skills (per domain)

Summary

- The 'Big Data Problem' is nothing more than addressing, independently, or together, the need to:
  - Manage & access data of interest amongst large volumes of data
  - Acquire data of interest quickly
  - Understand and implement analytics to scientifically combine heterogeneous datasets
  - Understand and enhance the quality of data

- The **new and growing field of data analytics** will facilitate the usability of heterogeneous data and information

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, as applied to specific science domains; Provide feedback to academic curriculums so that data scientists can acquire the needed skills

- We need to understand and **develop promising data analytics** techniques and technologies to facilitate this.

Analytics Tools Mentioned

- DISC, www.pdl.cmu.edu/DISC
- MapReduce, http://labs.google.com/papers/mapreduce.html
- Hadoop, http://hadoop.apache.org
- Dryad, http://research.microsoft.com/en-us/projects/dryad
- OpenCyc, www.opencyc.org
- Freebase, www.freebase.com
- Powerset, www.powerset.com
- True Knowledge, www.trueknowledge.com
- Wolfram|Alpha, www.wolframalpha.com
- myGrid, www.mygrid.org.uk
- myExperiment, www.myexperiment.org

Other Links

- http://en.wikipedia.org/wiki/Analyt
- http://www.smartplanet.com/blog/bulletin/7-new-types-of-jobs-created-by-big-data/682
- http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
- http://searchbusinessanalytics.techtarget.com/definition/Data-scientist
- http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
- http://www-01.ibm.com/software/data/bigdata/
- http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
- http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/
- http://km4meu.wordpress.com/2010/02/06/settling-the-eternal-semantic-debate-what-is-knowledge-what-is-information/
- http://www.informationweek.com/big-data/slideshows/big-data-analytics/big-data-analytics-masters-degrees-20/240145673
- http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/#Arizona_State_University
- http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192

References

- The Fourth Paradigm: Data-Intensive Scientific Discovery, Edited by: Tony Hey, Stewart Tansley, Kristin Tolle, 2009, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf
- Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
- Frontiers in Massive Data Analysis, National Research Council, 2013, http://bigdatawg.nist.gov/FrontiersInMassiveDataAnalysisPrepub.pdf
- Doing Data Science: Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O'Reilly Media; ISBN-10: 1449358659, 2013
- National Institute of Standards and Technology(NIST) Big Data Working Group (NBD-WG), NIST Big Data Definitions and Taxonomies, February, 2014, http://bigdatawg.nist.gov/show_InputDoc.php, M0142
- National Institute of Standards and Technology(NIST) Big Data Working Group (NBD-WG), Big Data Analytics Reading List, October, 2013, http://bigdatawg.nist.gov/show_InputDoc.php, M0264
- Big Data Activities – Earth Sciences Information Partners (ESIP), http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics

References

- National Institute of Standards and Technology (NIST) Big Data Working Group (NBD-WG), Demystifying Data Science, Natasha Balac, Predictive Analytics Center of Excellence, San Diego Supercomputer Center, University of California, October, 2013, http://bigdatawg.nist.gov/show_InputDoc.php, M0169
- Big Data Effects on Weather and Climate, Informal Discussions on The New Economics, Nancy Grady, http://www.ametsoc.org/boardpges/cwce/docs/profiles/GradyNancyW/2013-04-AWF.pdf
- EOSDIS FY2013 Annual Metrics Report, NASA ESDIS: Hyo Duck Chang, Brian Krupp, Lalit Wanchoo, Young-In Won, February 2014
- IEEE BigData 2014, http://cci.drexel.edu/bigdata/bigdata2014/callforpaper.htm
- Big Data Activities – Research Data Sharing (RDA), https://rd-alliance.org/groups/big-data-analytics-ig/wiki/big-data-analytics-interest-group-charter.html
- Open Geospatial Consortium (OGC) Big Data Working Group, http://external.opengeospatial.org/twiki_public/BigDataDwg/WebHome
- Society of American Baseball Research, http://sabr.org/sabermetrics
- TRMM 3B42, http://disc.sci.gsfc.nasa.gov/precipitation/documentation/TRMM_README/TRMM_3B42_readme.shtml

Yes… Gleaning Knowledge from Information is in Reach.
The wheels are turning, and we need to fuel the fire.

Thank You

(Steven.J.Kempler@nasa.gov)

Learn more about Data Analytics… Come see:

John Schnase, Dan Duffy
NASA GSFC

Meeting the Big Data Challenges of Climate Science through Cloud-Enabled Climate Analytics-as-a-Service

WEDNESDAY, MARCH 26, 2014