Federal Big Data Working Group

http://www.meetup.com/Federal-Big-Data-Working-Group/sponsors/

- Goodier
What Are We Doing?

Xcelerate Solutions sponsors the activities of the Federal Big Data Working Group for the advantage of our government clients.

- We meet both virtually and in our office in McLean VA at least once a month.
How Are we Doing it?

• Dr. Katherine Goodier, Data Sciences Practice Leader, Xcelerate Solutions
• Dr. Chuck Rehberg, CTO, Trigent and Semantic Insights
• Dr. Kirk Borne, Professor of Astrophysics and Computational Science, George Mason University
• Dr. Tom Rindflesch, Information Research Specialist at Cognitive Science Branch, National Institutes for Health (NIH)
• Ms. Mary Galvin, Managing Principal, AIC & Semantic Community
• Dr. Joan Aron, Independent Consultant Climate Data, Aron Environmental Consulting

BIG DATA is a TEAM SPORT
Federal Big Data and Cognitive Metadata

- Goodier
What is the Cognitive Metadata Solution

Cognitive metadata (i.e. metadata coming from enhancing machine learning with our human perception, reasoning, or intuition such as preference for a type of content), which is very useful for personalization purposes and conversely for limiting PII incidents.

Ultimately cognitive metadata improves automated reasoning.
The Internet was built without a way to know who or what you were connecting to

- Federal internet service providers workaround this with a patchwork of identity security controls and NIAP certifications
- No fair blaming the user – no framework, no cues, no control
2. Safeguarding and Sharing Information


“ For example, the United States Government Accountability Office (GAO) aggregates data from many agencies.

Recognizing the inherent risks, GAO sets up discrete network enclaves that are distinct from their agency-wide network, for Big Data. It assigns appropriate levels of security to each enclave driven by the sensitivity of the data therein.

• Other agencies note they ensure Big Data is stripped of personally identifiable information (PII) before it leaves the originating agency’s control.

• Data aggregation needs will expand as more elements of the critical infrastructure adopt increased cyber protection and detection capabilities that will drive enhanced data/information sharing.”

• www.meritalk.com

• Beacon Report

• Balancing the Cyber Big Data Equation
Over time, agencies, digital developers, and data users may also create, discover, or propose new and innovative ways to combine, share, or otherwise leverage the power of the digital data and content collected or disseminated by their digital services or programs. If data will be re-combined, used or shared in ways that individuals did not originally contemplate or expect, agencies must consider the need, under applicable law or policy, to provide such individuals with additional or updated notice of their privacy rights and choices.4

In determining precisely when, where, and how to give such notice, agencies, their digital developers, and partners will need to exercise creativity and ingenuity to ensure that required notices are clearly communicated to individuals at the right time and place, and in the right manner, without unduly interfering with the user experience. The timing and format of such notices may need to vary, depending on the digital or mobile platform involved.5
## What is PII?

**PII includes:** Name, email, home address, phone #

**Sensitive PII includes:**

<table>
<thead>
<tr>
<th>If Stand-Alone:</th>
<th>If Paired With Another Identifier:</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ Social Security number</td>
<td>➢ Citizenship or immigration status</td>
</tr>
<tr>
<td>➢ Driver’s license or state ID #</td>
<td>➢ Medical information</td>
</tr>
<tr>
<td>➢ Passport number</td>
<td>➢ Ethnic or religious affiliation</td>
</tr>
<tr>
<td>➢ Alien Registration Number</td>
<td>➢ Sexual orientation</td>
</tr>
<tr>
<td>➢ Financial account number</td>
<td>➢ Account passwords</td>
</tr>
<tr>
<td>➢ Biometric identifiers</td>
<td>➢ Last 4 digits of SSN</td>
</tr>
<tr>
<td></td>
<td>➢ Date of birth</td>
</tr>
<tr>
<td></td>
<td>➢ Criminal history</td>
</tr>
<tr>
<td></td>
<td>➢ Mother’s maiden name</td>
</tr>
</tbody>
</table>

https://it.ojp.gov/default.aspx?area=privacy&page=1295
3. Risk Exposure grows as our use of Federal Shared Services grows

- **Clinger-Cohen** 1996
- **Quicksilver** 2001
- **E-Government Act** 2002
- **E-Gov Initiatives** Initial 5 (HR, GM, FM, FHA,CM) 2004
  - **Lines of Business** Round 2 (Geo, BFE, ITI, ISS) 2006
- **Payroll Consolidation Completes** 2009
- **GAO Report: Opportunities to Reduce Potential Duplication** 2011
- **Cloud-First** 2010
- **E-Gov Initiatives** Round 2 (DAIP, ITDS, IAD-Loans/Grants) 2008
  - **Lines of Business** Round 2 (Geo, BFE, ITI, ISS) 2006
- **Shared Services** 2011
4. Ensuring adherence to Security and Privacy regulations across identities shared in the federal clouds

retain MEANING (aka, contextual semantics)
in loosely coupled, highly flexible multi-tenant environments
4. Solutions for the Federal Use Case from Research

Cognitive metadata: Advanced Streaming and Prediction for improved compliance to security and privacy regulations

Caching prediction algorithms will adjust according to risk exposure, and personal information protection capabilities improve over time

4. Cognitive Metadata solutions shared across government at the new scale of IT


1. Align budget and acquisitions with the technology cycle;
2. improve program management;
3. streamline governance and increase accountability;
4. increase engagement with the IT community; and
5. adopt lighter technologies and shared solutions—including the adoption of a "cloud-first" policy.

– www.cio.gov
4. Cognitive metadata employs predictive algorithms from Big Data Machine Learning combined with contextual Natural Language Processing

Cognitive metadata uses a three-step human-driven process that translates Policy documents into formal policy rule sets that computers can understand and evaluate.

1. Policy documents are translated into digital policies, using Natural Language Processing technologies.

2. Policy deconfliction ensures consistency and operational desirability. Automated deconfliction, using Turing methods and Theorem Proving Techniques that work with the constructs defined in XML, delivers active models of the resulting policy via a Policy Based Tool GUI. DPM delivers this new user interface to data stewards and Foreign Disclosure Officers (FDOs) giving them total control over both the design and the approval of the resulting model. Then the human-approved set of deconflicted digital policies are translated into standard QOS policy-labeled services.

3. Digital policies are defined in a computer interpretable language which is also friendly to humans.
4. Cognitive Metadata enhances data science

Convergence

Predictions that enhance machine learning fueled by encoded human knowledge at the Intersection of Our Digital Lives.

Substantive expertise

Hacking Skills

Math & Statistics Knowledge

Danger Zone!

Data Science

Traditional Research

Machine Learning
“A New Natural Language Understanding Technology for Research of Large Information Corpora.”

By Chuck Rehberg, CTO
Semantic Insights™
a Division of Trigent Software, Inc.
04 August 2014
“Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.”
- Robert Frost

**Road 1:** What can we do with what we have?

**Road 2:** What is required to solve it properly in the first place?
Large Corpora are everywhere and growing

- **The Internet**
  - Eric Schmidt, the CEO of Google, estimated the size of the internet at roughly 5 Million Terabytes of data. Schmidt further noted that in its [first] seven years of operations, Google has indexed roughly 200 terabytes of that, or .004% of the total size. -widely cited

- **The Library of Congress**

- **Medline (NIH)**

- **NSF**

- **Twitter**
  - “Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.” - [http://www.internetlivestats.com/twitter-statistics/](http://www.internetlivestats.com/twitter-statistics/)

How do we make this information immediately computable?
Road 1: What can we do with what we have?

- **Keyword search**
  - “As the Web gets increasingly large and complex, keyword search becomes less effective as a means for making sense of it. In fact, it will even decline in productivity in the future. Natural language search will be a bit better than keyword search, but ultimately won’t solve the problem either — because like keyword search it cannot really see or make use of the structure of information.” - [http://www.novaspivack.com/technology/diagram-beyond-keyword-and-natural-language-search](http://www.novaspivack.com/technology/diagram-beyond-keyword-and-natural-language-search)
  - Requires pre-indexing the corpus (index is fixed as long as document doesn’t change)

- **Semantic Tagging** (a method by which computers can categorize the content)
  - “Different individuals create substantially different conceptualisations of tagging data and tagging activities despite the fact that their purposes are similar.” - [http://dcpapers.dublincore.org/index.php/pubs/article/download/925/921](http://dcpapers.dublincore.org/index.php/pubs/article/download/925/921)
  - Requires pre-indexing the corpus (subject to change when the folksonomy/ontology changes)

- **Semantic Indexing** - e.g. Latent Semantic Indexing (LSA)
  - “...(LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text.” - [http://lsa.colorado.edu/whatis.html](http://lsa.colorado.edu/whatis.html)
  - “LSA cannot capture polysemy (i.e., multiple meanings of a word). Each occurrence of a word is treated as having the same meaning due to the word being represented as a single point in space. For example, the occurrence of “chair” in a document containing “The Chair of the Board” and in a separate document containing “the chair maker” are considered the same.” - [http://en.wikipedia.org/wiki/Latent_semantic_analysis#Limitations](http://en.wikipedia.org/wiki/Latent_semantic_analysis#Limitations)
  - Requires pre-indexing the corpus (index is fixed as long as document doesn’t change)
A closer look at the Problem

1. Since we often cannot know the true intent of the author, the available “meaning” of a text is ascribed by the reader and is not inherent in the text.
   - Different readers can legitimately hold that a given text means different things

2. The desired “Information of Interest” specified in any research investigation may vary based on the needs and POV of the Reader. For example:
   a) One reader can be reading for Risks
   b) Another reader can be reading for Opportunities

3. The meaning of a natural language text can be grammatically ambiguous.

4. The “correct” sense of the terms can be open to interpretation.

5. The information represented can be conflicting

6. As the reader learns more, interpretations can change even when the text does not

7. The rate of addition of documents to a corpus is generally increasing.
Road 2: What is needed to solve it properly?

Ability to:
1. “Speed read” Natural Language text.
2. Dynamically map Natural Language text to the appropriate/chosen Ontologies
3. Learn from teaching, training and experience

Support for:
1. Multiple evolving Ontologies (including POV)
2. Multiple simultaneous meanings
3. Evolving document corpora
Road 2: Real-time Document Research for the Masses

- The first release is here (in Beta Test).
  - Check out the products currently in Beta Test at Semantic Insights™ [http://www.semanticinsights.com/](http://www.semanticinsights.com/)

- Get Involved

- Learn What’s going on
  - Check out [http://semanticcommunity.info/](http://semanticcommunity.info/)
To learn more

- At *Semantic Insights* we focus on developing semantics-based information products that produce high-value results serving the needs of general users requiring little or no special training.

- Visit us at [www.semanticinsights.com](http://www.semanticinsights.com)
Chuck Rehberg

As CTO at Trigent Software and Chief Scientist at Semantic Insights, Chuck Rehberg has developed patented high performance rules engine technology and advanced natural language processing technologies that empower a new generation of semantic research solutions.

Chuck has more than thirty years in the high-tech industry, developing leading-edge solutions in the areas of Artificial Intelligence, Semantic Technologies, analysis and large-scale configuration software.
Effectively Exploiting Big Data with Semantics

Thomas C. Rindflesch, National Library of Medicine, National Institutes of Health
Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.
Background

- Basic biomedical research is crucial to medicine
  - Complexity of molecular pathophysiology
  - Challenges development of new therapies
- Big data can facilitate progress
  - MEDLINE: Biomedical research literature
- Need effective, automatic access to information in this data source
Background

- Basic biomedical research is crucial to medicine
  - Complexity of molecular pathophysiology
  - Challenges development of new therapies
- Big data can facilitate progress
  - MEDLINE: Biomedical research literature
- Need effective, automatic access to information in this data source
- Semantic processing
Semantic processing: **SemRep**

- Developed at National Library of Medicine
- Depends on domain knowledge
  - Unified Medical Language System (UMLS)
- Computable representation of meaning
  - Semantic predications
SemRep: semantic predication

Exemestane after non-steroidal aromatase inhibitors for post-menopausal women with advanced breast cancer

Aromatase Inhibitor TREATS Breast Carcinoma

Metathesaurus Concept

Semantic Network Relation

Metathesaurus Concept

Unified Medical Language System
Web application: Semantic MEDLINE

• Uses nearly 70 million semantic predications
  – From all of MEDLINE
• Guides the user through content
  – Makes connections “not visible to the naked eye”
• Exploits existing IR system
  – PubMed
• Displays results as an interactive graph
Semantic MEDLINE overview

- PubMed: Document retrieval
- MEDLINE citations
- SemRep: semantic processing
- Semantic predications
- Automatic summarization
- Graphical summary
- Biomedical information management
Use case: Inflammation and cancer

• With some exceptions, cancer therapy is not effective
• Scientific basis
  – Traditionally: kill cancer cells
  – More recently: manipulate non-cancer cells (immune system)
• Goal: look for trends in cancer immunotherapy
by promoting neo-angiogenesis. The aim of this study was to investigate whether inflammatory cytokines and vascular endothelial growth factor (VEGF) levels in exhaled breath condensate (EBC) and in serum were related to tumour size in patients with stage I-IIA non-small cell lung cancer (NSCLC). The age and sex-matched controls, while only serum IL-6 concentrations were measured, allowed for an assessment of the role of IL-6 in the carcinogenesis of NSCLC. The IL-6, IL-17, IL-1α, and IL-17α expression levels of the VEGF, IL-6, IL-1α, and IL-17α in EBC and serum of patients were measured by multiplex bead array. The results showed a significant correlation between EBC and serum levels of IL-6 and IL-17α (r = 0.78, p = 0.001). The tumor diameter was significantly correlated with EBC concentrations of VEGF (r = 0.58, p = 0.039), IL-6 (r = 0.67, p = 0.013) and IL-17α (r = 0.66, p = 0.017). Our results show a significant relationship between inflammatory cytokines and angiogenic markers, which indicates that IL23R may play an important role in the carcinogenesis of bladder cancer. While effective, the inflammation induced by these therapies is transient and not designed to induce long-lasting tumor-specific cytolytic T lymphocytes (CTLs) responses that have proven so adept at eradminating tumors. Therefore, in order to maintain the benefits of bacteria-induced acute inflammation but gain long-lasting anti-tumor immunity, nonpathogenic bacteria have been shown to prove part of the immune system in a manner wherein it is able to efficiently deliver tumor antigens to both the MHC Class I and II antigen presentation pathways for activation of tumor-targeting CTL-mediated immunity. Lm is a versatile bacterial vector as evidenced by its ability to induce therapeutic immunity against a wide-array of TAAs and specifically infect and kill tumor cells directly. It is for these reasons, among others, that Lm-based immunotherapies have delivered impressive therapeutic efficacious in preclinical models of cancer for two decades and are now showing promise clinically. In this review, we provide an overview of the history leading up to the development of current Lm-based immunotherapies, the advantages and mechanisms of Lm as a therapeutic vaccine vector, the preclinical experience with Lm-based immunotherapies targeting a number of malignancies, and the recent findings from clinical trials along with concluding remarks on the future of Lm-based tumor immunotherapies. Considerable evidence has suggested that chronic inflammation is a causative factor in the development of human colorectal cancer (CRC). Interleukin (IL)-17A produced mainly by Th17 cells is a novel proinflammatory cytokine and increased IL-17A is associated with colorectal neoplastic transformation. In this study, the expression of IL-17A mRNA was non-statistically increased (4-fold higher) in the CRC-adjacent tissues and it became significantly increased (3-fold higher) in the CRC-adjacent tissues as compared with the control. The expression level of IL-17A in the CRC-adjacent tissues was also examined. The results showed that the expression pattern of Th17 cytokines was not affected by the presence of CRC. The expression pattern of Th17 cytokines was not affected by the presence of CRC. These results are consistent with the previous reports that Th17 cytokines are not involved in the colorectal neoplastic transformation. The inhibition of cyclooxygenase (COX) activity by NSAIDs plays a role in their anti-tumorigenic properties. NSAIDs also have COX-independent activity which is not fully understood. In this study, we report a novel COX-independent mechanism of sulindac sulfide (SS), which facilitates a previously uncharacterized cleavage of epithelial cell adhesion molecule (EpCAM) protein. EpCAM is a type I transmembrane glycoprotein that has been implemented as an over-expressed oncogene in many cancers including colon, breast, pancreas, and prostate.
Semantic predicates as a graph
Cytokine CAUSES Tumorigenesis

**Interleukin-8 and interleukin-17 for cancer.**

Zarogoulidis P¹, Katsikogianni F, Tsiouda T, Sakkas A, Katsikogiannis N, Zarogoulidis K.

**Abstract**

Pro-inflammatory cytokines have been associated with chronic inflammation and inflammatory diseases. Increased levels of interleukins (ILs) have been associated with inflammatory disease exacerbation. ILs levels have been observed to be associated with advance stage cancer for several types of cancer and a poor prognostic maker for malignant disease. Moreover; increased levels of cytokines induce tumorigenesis. There are several paradigms such as the hepatocellular
Large-scale implementation

• Collaboration with Oak Ridge National Lab
  – YarcData Urika graph-appliance super computer
  – Knowledge discovery and graph mining library

• Exploit both semantics and graph theory

• Support interactive discovery browsing
  – For hypothesis generation in biomedical research

• Large graphs on a jumbo screen
Acknowledgments

• Michael J. Cairelli, D.O.
• Marcelo Fiszman, M.D., Ph.D.
• Halil Kilicoglu, Ph.D.
• Graciela Rosemblat, Ph.D.
• Dongwook Shin, Ph.D.
• Sreenivas R. Sukumar, Ph.D. (Oak Ridge National Lab)
Big Data:
Astronomy and beyond

Kirk Borne
@KirkDBorne
School of Physics, Astronomy, & Computational Sciences
College of Science, George Mason University, Fairfax, VA
Big Data: What is it good for?

The 3 D2D’s

✓ Knowledge Discovery
  – Data-to-Discovery

✓ Data-driven Decision Support
  – Data-to-Decisions

✓ Big ROI (Return On Innovation) !!!
  – Data-to-Dollars
Astronomy Big Data Example

The LSST (Large Synoptic Survey Telescope)
LSST = Large Synoptic Survey Telescope
http://www.lsst.org/

8.4-meter diameter primary mirror = 10 square degrees!
(mirror funded by private donors)

Hello!
LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

Construction begins July 2014

8.4-meter diameter primary mirror = 10 square degrees!

(mirror funded by private donors)

Hello!
LSST = Large Synoptic Survey Telescope
http://www.lsst.org/

8.4-meter diameter primary mirror = 10 square degrees!

- 100-200 Petabyte image archive
- 20-40 Petabyte database catalog

(mirror funded by private donors)
**LSST Key Science Drivers: Mapping the Dynamic Universe**

- Solar System Inventory (moving objects, NEOs, asteroids: census & tracking)
- Nature of Dark Energy (distant supernovae, weak lensing, cosmology)
- Optical transients (of all kinds, with alert notifications within 60 seconds)
- Digital Milky Way (proper motions, parallaxes, star streams, dark matter)

**LSST in time and space:**
- When? ~2022-2032
- Where? Cerro Pachon, Chile
LSST Summary
http://www.lsst.org/

- 3-Gigapixel camera
- One 6-Gigabyte image every 20 seconds
- 30 Terabytes every night for 10 years
- Repeat images of the entire night sky every 3 nights: *Celestial Cinematography*
- 100-Petabyte final image data archive anticipated – *all data are public!!!*
- 20-Petabyte final database catalog anticipated
- Real-Time Event Mining: ~10 million events per night, every night, for 10 years!
  - Follow-up observations required to classify these
  - Which ones should we follow up? …

… Decisions! Decisions! ( = D2D !)
Decision Analytics – based on massive amounts of information
(Big Data – What is it good for? Decision Support and Innovation!)

From Devices......

...... Intentions...

...... Demographics...

...... Location, weather, and other geographic attributes...

Slide provided by
http://www.syntasa.com/
• **Decision Analytics-as-a-Service™ (AaaS)**
  – Your business rules determine the goals, decision points, alerts, and responses.

• **Analytics-as-a-Service (AaaS)**
  – Moving beyond historical *hindsight* and *oversight* (**Descriptive Analytics**) to the new world of *insight* and *foresight* (**Predictive & Prescriptive AaaS**).

• **Multi-channel big data streams**
  – Mining and acting on internal motivators (endogenous factors) and external motivators (exogenous factors).

• **Personalization and Customization**

• **Decision Automation in a rich content (Big Data) environment**

---

*Based on Marketing Analytics-as-a-Service™ (MAaaS) from*

Big Data: What is it good for?

✓ Knowledge Discovery
- BD2K = Big Data-to-Knowledge  http://bd2k.nih.gov
- Class discovery: predictive power discovery
- Association discovery: recommender engines
- Correlation discovery
- Novelty discovery: *surprise!*

✓ Data-driven Decision Support
- Predictive & Prescriptive Analytics
- The Last Mile Challenge
- Data-to-Decisions (DTD)
- Decision Science-as-a-Service™ http://www.syntasa.com/

✓ Big ROI (Return On *Innovation*) !!!
Ms. Mary Galvin

- [http://www.aicnova.com/#!about1/c158v](http://www.aicnova.com/#!about1/c158v)

- AIC was formed in November 2013 to fulfill a market need for consultants with collective technology and business expertise. The company's founder, Mary Galvin, previously worked in technical and customer-facing roles at Northrop Grumman, Basis Technology, SRA International, and LexisNexis Special Services (LNSSI). Mary's domain expertise encompasses a variety of technologies, to include data intensive computing systems, entity/identity resolution, machine translation, foreign language search, digital forensics, and Unmanned Aerial Vehicles (UAVs). Mary has been involved in several initiatives pertaining to HPCC Systems since April 2011. The HPCC is an open source platform that excels at processing datasets exhibiting the '3 V's' of 'big data': volume, variety and velocity.

- Mary graduated from Villanova University in 2003 with a bachelor's degree in computer engineering and minors in computer science, Spanish and naval science. She has taken several extended learning courses in Arabic and Chinese and is expected to obtain a master's in information systems engineering from Johns Hopkins University in 2016.

- Outside of work Mary volunteers with Girls in Technology (GIT), a DC-based organization that encourages girls in the 6th through 12th grade to pursue careers in STEM (Science, Technology, Engineering, and Math). She also volunteers with Villanova's entrepreneurship program, helping undergraduate students from her alma mater understand the practical aspects to their entrepreneurial studies.
Big Data at LexisNexis

US Public Records
- 50 billion records
- 10k+ data sources
- 250 mil. unique identities
- 1.1 billion unique businesses

Patent Data
- 100+ patenting authorities
- Translations for all non-English content
- On average, sources go back roughly 30 years (some go back 100+ years)

Case Law
- 20 million + court records from federal, state and local governments
- Non-US countries include France, Australia, Hong Kong, Canada, and the UK

News Articles
- 20k+ sources, including traditional print (newspapers, magazines, trade journals, etc) and “new” media (ie, blogs, Twitter feeds, audio & video transcripts)
History of the HPCC

- **Late 90s/Early 2000s**
  - Designed and Developed from the Ground-Up to Meet LexisNexis’ Internal Big Data Needs.

- **2001**
  - United States Government Sought After Getting LexisNexis’ Data Capabilities In-House for their Internal Data Mining Needs.

- **2004**
  - Google’s MapReduce Paper is Published.

- **2007**
  - First Release of Hadoop Available (designed after Map Reduce Papers).

- **2009**
  - The Idea of Releasing the HPCC to the OSS Community was Presented to LexisNexis Corporate Management.

- **2011**
  - The HPCC is Officially Released to the Open Source Community!

- **2012**
  - The Spread of HPCC Users has Gone Global, and as a Result, Innovation Ignites.
HPCC Architectural Overview
ECL Overview

Task: Produce a set of records wherein a particular field contains a specific set of values

Start at top of MyFile
Loop through MyFile records
  If MyField = 1 or MyField = 3 or MyField = 4 or MyField = 7
    Include record in output set
  Else
    Throw out record and go back to top of loop
end if and loop

Typical approach for solving this in many programming languages
ECL Overview (cont’d)

Task: Produce a set of records wherein a particular field contains a specific set of values

SetValidValues := [1,3,4,7]; //Set Definition
IsValidRec := MyFile.MyField IN SetValidValues; //Boolean
ValRecsMyFile := MyFile(IsValidRec); //filtered Recordset
OUTPUT(ValRecsMyFile);
HPCC Modules & Plugins

• Exploratory Data Analysis (EDA) Toolkit

• Scalable Automated Linking Technology (SALT)
  • Data Ingest
  • Data Profiling
  • Data Hygiene
  • Clustering
  • Relationship Extraction

• Other
  • H2H Connector
  • Machine Learning Module
  • R Integration
  • Eclipse IDE
  • JDBC Driver
  • .........
HPCC Academic Program

• Audience: Colleges and Universities
• Benefits:
  • Internship opportunities
  • Invitation-only conferences
  • Free training for qualifying projects
  • Access to an external cluster, as available
HPCC Academic Community
Additional Learning Options

• Audience: Anyone!

• Learning Options:
  • Online:
    • Includes both prerequisites and tailored courses depending on role type (ie, developers, analysts, and administrators)
    • [http://hpccsystems.com/community/training-videos](http://hpccsystems.com/community/training-videos)
  • In-Person:
    • [http://hpccsystems.com/community/training-events/training](http://hpccsystems.com/community/training-events/training)
Getting Started: hpccsystems.com
Transforming Data-Driven Publications and Decision Support

Joan L. Aron, Ph.D.
Consultant
Federal Big Data Working Group

COM.BigData 2014
Mission
Federal Big Data Working Group

• **Federal** – supports Federal Big Data Initiative but not endorsed by federal govt or agencies

• **Big Data** – supports Federal Digital Government Strategy “treating all content as data”

• **Working Group** – data science teams (fed govt & non-fed govt experts) for big data products

• **Meetup** – world’s largest network of local groups to help people self-organize
Big Data Transformation of Science

• Vision of Federal Leadership
  – Networking and Information Technology Research and Development (NITRD) *interagency program*
  – National Science Foundation (NSF)

• Environment and Earth System Science Focus
  – Data Publishing and Public Access
  – Data Discovery and Decision Support
  – Workforce Development - Working With Data Science (Scientific Community and Decision-Makers)
NITRD Big Data Initiative

• Core Technologies
  – collection, storage, management, analytics
  – sharing, collaboration
• Domain Research Data
  – astronomy, data.gov, Earth observation systems
  – genomics, materials genome, nano S&T
  – NSF (DataOne, DataNet), particle physics, LHC
• Challenges/Competitions (engage broader public)
• Workforce Development (data science, Big Data)
NSF Strategic Plan 2014 – 2018

• NSF Agency Priority Goals
  – Increase Data Scientists and Data Infrastructure
  – Optimize Award Process to Level Workload
  – Ensure Public Access to Publications

• FBDWG/Semantic Community
  – Data Scientists Building Infrastructure to Improve Public Access to Publications
  – Participating in NSF Grants Funding Process
  – NSF Strategic Plan Knowledge Base (data publication with initial ontology)
    http://semanticommunity.info/Data_Science/NSF_Strategic_Plan#Story
Data Publishing and Public Access

• NSF Asst. Dir. Farnam Jahanian
  – Implementation plans for public access (to scientific research data) could vary by discipline, and new business models for universities, libraries, publishers and scholarly and professional societies could emerge.

• FBDWG/Semantic Community
  – Team can broker a win-win for scientific community and scientific publishers (e.g., Elsevier Data Research Services).
Data Publishing and Public Access

• NSF Geosciences (GEO) Aim - Data Citation (2012)
  – Scientific collaborators decide on datasets
  – Data centers, libraries, repositories, publishers develop data citation methods
  – Research institutions value data citations

• FBDWG/Semantic Community
  – Formed Data Science Publications Interest Group with NSF EarthCube (National Data Infrastructure for Earth System Science – sharing GEO data)
  – Scientific Data Publications in Data Browsers
  – Data Publications: Publications, Websites, Databases
Data Discovery and Decision Support

- FBDWG/Semantic Community – Environment and Earth System Science Focus
  - Can we better visualize the “invisible”?
  - Nonpoint source pollution is “invisible” to many.
    - Runoff (nitrogen, phosphorus, sediment, pathogens) is the major cause of U.S. water quality impairments.
    - Can Big Data improve speed and accuracy of monitoring to identify problems and target interventions faster?
    - Example: Lake Erie has history of harmful algal blooms (HABs). Municipal sewage treatment “solved” problem by 1980s but HABs came back in 2000s due to runoff.
Workforce Development
Working with Data Science

• FBDWG/Semantic Community – Environment and Earth System Science Focus
  – EarthCube Data Science Publications Interest Group
  – Implementation issues in scientific community
  – How much data should be published with study? Final set of data? Raw data? Audit trail?
  – Should it matter if data will be used for decisions affecting people’s lives and/or costing millions (medical treatment; environmental policy)?
  – Major decisions may end up in court.
  – Benefits of Big Data depend on workforce development for scientific community interacting with data science.
Workforce Development
Working with Data Science

• FBDWG/Semantic Community – Environment and Earth System Science Focus
  – Multiple decision-makers involved
  – Watersheds and water quality (agencies / organizations that manage water, land, coastal zones, public health; local landowners, land users)
  – Big Data can help to generate more options.
  – Big Data can support expertise in negotiation.
  – **BUT** Big Data cannot determine the decision outcome.
  – Benefits of Big Data depend on workforce development for decision-makers interacting with data science.
Return to Big Data Vision
George Strawn, Director NITRD

• Creating a science of big data?
• From mining knowledge to directing action?
• Bringing together diverse communities
• Enhancing big data education and training
Future Meetups

- September 8: OSTP FASTER CoP
- October 6: Wolfram Language
- November 3: Georgetown Massive Data Institute
- December 1: NSF GEO/EarthCube and ICER (Integrative and Collaborative Education & Research)