This Conference brought to you by

www.ttcus.com

Linkedin/Group:
Technology Training Corporation

@Techtrain

Technology Training Corporation

www.ttcus.com
GEOINT Big Data:
Evolving How We Understand the World
Through Research and Development

Dr. Ann Carbonell
Director, Office of Information Integration, NGA InnoVision

Big Data Analytics and Applications for Defense, Intelligence
and Homeland Security Symposium - 24 April 2013

This briefing may contain domestic imagery and/or U.S.
persons (USPER) information, which has been included
consistent with applicable laws, directives, and policies.
Agenda

- What is GEOINT?
- Sea of Change
- The Data Explosion
- Our Challenges
- Big Data Solutions
- “Yes” - Big Data 7 Vs
- R&D Focus Areas
Know The Earth... Show The Way... Understand the World

ANTICIPATE

Where something is... Why it’s there... What may happen...

Target-Based ➔ Activity-Based

Approved for Public Release 13-160
NGA is the lead federal agency responsible for GEOINT

**GEOINT Answers…**

- Where am I?
- Where are the natural obstacles?
- Where are the man-made structures?
- How do I navigate them?
- What activities are taking place there?
- Where are my adversaries/enemies?
- What did the area look like before?
- What does the area look like now?
- What might it look like after an event?

"The geospatial intelligence field… is legendary for its standard of customer support. It is the base foundation for integration."

— James R. Clapper, Director of National Intelligence

**GEOINT Anticipates…**

- Where something will be
- What may happen next

**GEOINT Provides…**

- A foundation for decision-making
- An ability to plan and execute
- Information dominance
- Decision superiority

Approved for Public Release 13-160
NGA Has a **Bold** Vision

**Putting the Power of GEOINT in Your Hands**

**NGA Strategy**

Provide online, on-demand access to our GEOINT knowledge

Broaden and deepen our analytic expertise to produce new value

“We will continue to deliver to our varied customer set what they need, when they need it, how they need it. But we have to be thinking about the future. We have to be continually pushing ourselves so that we do remain at the forefront.”

– Letitia A. Long, Director, NGA

Approved for Public Release 13-160
NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY
Know the Earth... Show the Way... Understand the World

Sea of Change

THE INTERNET in 2015
IS THE DAWN OF THE ZETTABYTE ERA

But how much data are we talking about?

1000 MEGABYTES = 1 GB
1000 GIGABYTES = 1 TB
1000 TERABYTES = 1 PB
1000 PETABYTES = 1 EB
1000 EXABYTES = 1 ZB

How much is a zettabyte really?
1 zettabyte equals to 36,000 years of HD-TV video, or the equivalent of streaming the entire Netflix catalog 3.177 times.

WHAT IS A PETABYTE?
TO UNDERSTAND A PETABYTE WE MUST FIRST UNDERSTAND A GIGABYTE.

1 GIGABYTE
7 MINUTES OF HD-TElevision VIDEO
2 GIGABYTES
20 YARDS OF BOOKS ON SHELF
4.7 GIGABYTES
SIZE OF A STANDARD DVD-R

A PETABYTE IS A LOT OF DATA

1 EXABYTE
20 MILLION FOUR-HOUR MOVIES FILLED WITH TEXT
1.5 EXABYTES
SIZE OF THE 10 BILLION PHOTOS ON FACEBOOK
15+ EXABYTES
INTERNET UPSIP DATA BACKED UP ON MOZY.COM
20 EXABYTES
THE AMOUNT OF DATA PER PERSON PROCESSED BY GOOGLE DAILY
50 EXABYTES
TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995

By 2015, nearly 3 billion people will be online, pushing the data created and shared to nearly 8 zettabytes.

IT WOULD TAKE OVER 5 YEARS TO WATCH THE AMOUNT OF VIDEO THAT WILL CROSS GLOBAL NETWORKS EVERY SECOND IN 2015

How will internet traffic grow?

What type of video?

How will people be getting their videos?

What does internet traffic look like right now?

©Cisco
The Data Explosion

Every second, one hour of video is uploaded to YouTube

Experiments | Simulations | Archives | Social Media | Sensors

Information Technology

The Challenge
Enable Discovery

Deliver capability to mine, search, and analyze this data in near real-time

Petabytes
Exabytes
Zettabytes

The Response
Science itself is evolving

Approved for Public Release 13-160

Slide Courtesy of Jeff Nichols
Associate Laboratory Director
Oak Ridge National Laboratory
GEOINT Big Data - Challenges

- Metadata poor and inconsistent across data sources
- New catalog for every data type
- Schema is not easily modified
- Costly maintenance of many databases
- Analysts’ biggest problem is finding the data and wasted time searching several repositories
- Need to find the right balance between richness of metadata, analysts time, and value to multi-INT analysis
- Costly to scale in size or computational capacity to accommodate NGA’s needs
Big Data Solutions: One Size DOES NOT Fit All
Big Data - 7Vs

Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, creation, storage, search, sharing, analysis, and visualization.

- **Visualization**: data patterns
- **Variety**: data in many forms
- **Velocity**: data in motion
- **Volume**: data at rest
- **Veracity**: data in doubt
- **Vulnerability**: data at risk
- **Value**: data that is meaningful

Harness and utilize massive data in new ways and bring together sensing, perception, and decision support.
Visualization

- Immersive
- Collaborative
- 3D/4D
- Connecting Disconnected Users

Source: NGA Pathfinder

Midwest US Tornado Hot Spots
Image Source: NOAA

Approved for Public Release 13-160
Crowdsourcing provides a mechanism for capturing image/video “information intelligence” content to support forensic research.
Variety

GEOINT Types of Data are Vast

Phenomenologies
- Geophysics
- Hyperspectral
- Infrared
- LiDAR
- Motion/Video
- Panchromatic
- Polarimetric
- Radar

Platforms
- Airborne
- Spaceborne
- Handheld
- Surface
- Subsurface

Providers
- National
- DoD
- Local/State
- Commercial
- Foreign/International
- Open Source

Foundation
- Elevation
- Features
- Gravity
- Controlled Imagery
- GPS
- Sails
- Bathymetry

Providers
- NGA
- NCOMS
- Commercial
- Open Source

Services
- Request
- Task
- Collection
- Discovery
- Retrieval
- Dissemination
- Processing
- Analysis
- Archive

User / Problem Solver
- Detect
- Classify
- Characterize
- Identify
- Understand

Structured Data
- Transactional
- Time-phased data

Semi &/or Unstructured Data
- Text Report
- Emails
- Presentations
- Videos

Patterns
- Points of interest over time

Activity

Networks

Associations

Approved for Public Release 13-160
Automation and Conflation

The development of automation algorithms for extracting, attributing, and updating foundation data holds the promise of order-of-magnitude improvements in processing time, information extraction, geolocation accuracy, and analyst efficiency.

The “Spaghetti Bowl” test conflates (combines) Urban Feature Data and Open Street Maps and compares the result against two human conflated data sets.

Manual conflation took 29 hours; with automation the process took 20 seconds.
Velocity

Extracting Change: Movers, Tracks, Tips, Patterns

Only keep vital information

Goals: discover, find and characterize activity, patterns of life, networks and anomalies
Paving the Way for IC-ITE

The Volume Variety and Velocity of geospatial data results in new hardware and software architecture challenges which need to be addressed prior to major community acquisitions.

**Architecture 1**
- Amazon
- IC-ITE (Utility)
- IC-ITE (Storage)
- NSG
- Virtual Machines
- GPUs?
- We can use this

**Architecture 2**
- Google
- IC-ITE (Data)
- Bare Metal
- GPUs?
- We are just beginning here!

- How do we store and divide geospatial data (Extending Apache Accumulo) so that the world will be able to perform Geospatial operations in this new type of environment?

- Which geospatial algorithms are best run in HPC architecture, Utility architecture or on specialty hardware like GPU platforms, shared memory machines, or Tilera based machines that may be geographically distributed or disconnected?

**Move the computation to the data rather than the data to the computation**
Volume

Change Management – Heat Maps – Tipping and Cueing

What is contributing to data explosion?

Increasing Data Variety, Complexity and Producers

- Sensor and Processing Data
- More Spectral Bands
  - LIDAR
  - Giga-pixel Sensors
  - 30 FPS WAMI
  - Video SAR
- Higher-quality Optics
- Source Metadata
- Crowdsourcing
- Data from Humans
- Tables, Documents, Chat
- Records

New Sensor Capabilities Driving Need For New PED Solutions

Much of DoD data is high-definition imagery from drones, satellites and battlefield sensors. Post 9/11 this data has increased 1600%.

Source: 1105 Gov Info Group, 2012
Data Reduction

Source for Video: Air Force Research Lab, COMPASE Center, Test at Creech AFB by PNNL

Approved for Public Release 13-160
Veracity

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Ground Photo Geolocation

Determining the location where a ground photo was taken without any metadata is a hard problem requiring cloud techniques to enable sufficient scaling

Ground Photo of Interest
Source: DARPA Photo of Mt Whitney, CA Nat'l Park

LIDAR terrain grid view showing identical perspective

- Standards
- GEOINT Accuracy
- Metadata Tagging
- Ontologies
- Discoverability
- Governance
- Precision
- Data Quality
- Data Retention
- Analytic Confidence/Rating
- Calibration

Source: DARPA Photo of Mt Whitney, CA Nat'l Park

Approved for Public Release 13-160
Vulnerability

Security of the Data

- Data Governance
- Security Data Warehouse
- Compliance Requirements
- Threat and Risks

- Purpose-based Access
- Protocols
- Privacy
Understanding the “Story” the Data is Telling

A broad mission, as part of the U.S. Intelligence Community, to provide timely, relevant, and accurate geospatial intelligence in support of national security – in times of peace, in times of crisis and in times of war.
R&D Focus Areas

• **Produce Metadata from the BLOB**
  - Trackers (Velocity, Volume)
  - Tipping and Cueing (Volume)
  - Feature Extraction (Visualization)
  - Automatic Target Recognition ATR (Volume, Velocity and Visualization)
  - 3D from Motion Imagery (Visualization)
  - Compression and Data Reduction (Volume)
  - Crowdsourcing (Volume)
  - Prototype implementation for improved video quality off platforms and through networks (Veracity, Value)
  - Structured Observation Management/Associations (Veracity, Velocity)
  - Trustworthiness/Confidence (Vulnerability, Veracity)
  - Prioritization (Volume)

• **Advanced Analytics**
  - Relationships and Multi-INT Fusion (Variety)
  - Community and Open Metadata Standards (Veracity, Value)
  - Interface Standards (Veracity, Volume)
  - IC-ITE (Community Cloud Computing Standards, Veracity)
  - MapReduce Analytics (Velocity, Volume)
  - Automation Tools (Velocity, Visualization)
  - Gaming Engines (Visualization, Velocity)
Questions?
Backup for Notes
GEOINT Functional Manager: D/NGA’s Community Leadership Role

- Heads the GEOINT Community (by EO 12333, ICD 113, and DoDD 5105.60)
- Advises the Director of National Intelligence (DNI) and Secretary of Defense on:
  - Resource management
  - Policies and procedures
  - Collection capabilities and gaps
  - Processing and dissemination
  - Technical architectures
  - Other issues or activities as determined by the DNI
- Sets training and tradecraft standards
- Develops and implements strategic guidance, policies, and procedures
- Ensures coordination within and across the Intelligence Community and Department of Defense
GEOINT Research and Development (GRAND) Subcommittee

- Includes representation from R&D/S&T organizations across the IC, DoD, and NSG
- Focuses on GEOINT R&D/S&T from a cross-community functional management perspective
- Recent achievements include:
  - Expanding Partnerships
  - R&D Strategic Framework
  - NSG R&D Emerging Technologies List
- Continue to expand the GRAND into the Civil Community, Cross-INT and in GEOINT Communities of Interest
Future Cloud Opportunities

Store and Divide Geospatial Data
Find the most effective tools to:
• Support feature extraction
• Improve indexing strategies
• Combine, fused and conflated data
• 4D probabilistic or predictive results
• Extend Apache Accumulo

Geospatial Data Flow Based on Use
Determine the best indexing and flow for:
• Geospatial data in MapReduce
• Streaming detection environment
• Standard databases in the Utility cloud
• Optimal use of the PRESSUREWAVE and Machineshop environments

Extracting Features from Imagery
• Develop automated content indexing with still imagery
• Correlate / conflate extracted features from imagery with geospatial data
• Developing structured observation management via image-semantics

Research

Hardware Architecture Challenges
Determine the best architectures for critical algorithms:
• Data cloud or utility cloud
• GPU platforms
• Field programmable gate arrays
• Shared memory machines
• Geographically distributed data/proc.

Show Stopper
Critical

Improve MapReduce Queries
Effectively broker interactive queries:
• In a map-reduce environment
• To ensure expensive queries are only undertaken when necessary
• Support Copilot-ABR or Distillery-Ghostmachine

Adapt Multimedia for the Cloud
Develop multimedia intelligence analysis, reporting and dissemination tools such as:
• Contemporaneous video merging in a distributed environment
• Optimizing image processing/display in the IC-ITE environment

Important

Approved for Public Release 13-160