Analytical Tools: What Auditors Need to Know About Big Data

Timothy M. Persons, Ph.D.
Chief Scientist
U.S. Government Accountability Office
personst@gao.gov / www.gao.gov / @GAOChfScientist

Presentation to the MAMIAF/SWIAF/SEIAF Joint Meeting

December 4, 2014

CHARACTERIZING BIG DATA

Massive data sets generated & stored with size beyond ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

Exponential growth in volume results from data creation in digital form, proliferation of sensors (ubiquitous, high bandwith), high resolution imagery and video, complex simulations. Use of algorithmic approaches to extract meaning from huge volumes of data.

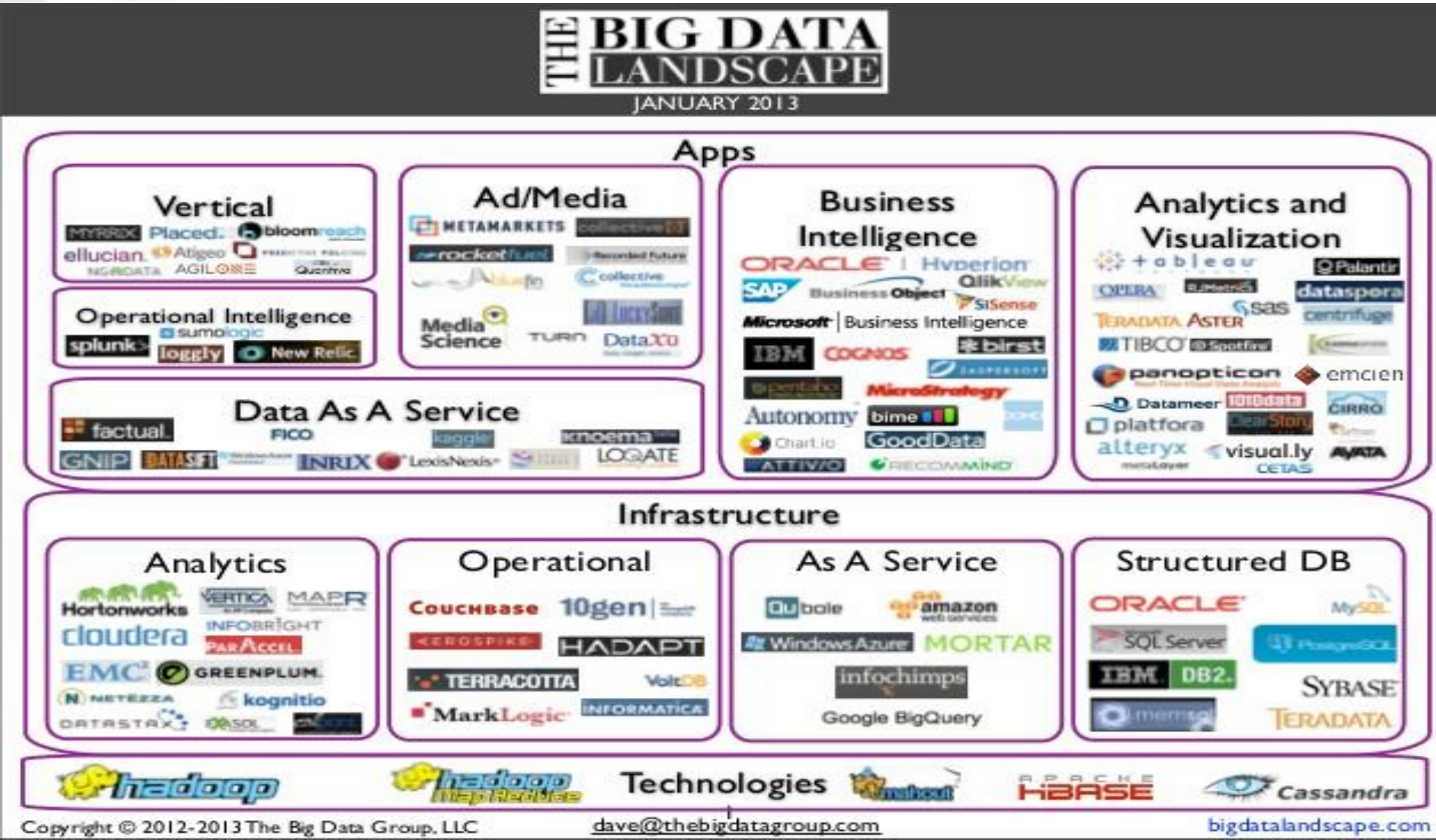
Technologies required to efficiently process large quantities of information, e.g., massively parallel processing frameworks such as shared nothing relational databases, MapReduce programming frameworks, and cloud infrastructure

***The V6 challenge of Big Data*: Volume, Velocity, Variety, Visualization, Verification, Value**

Source: David Feinleib, The Big Data Group, LLC

GAO

BIG DATA: INFRASTRUCTURE & APPS



dave@thebigdatagroup.com

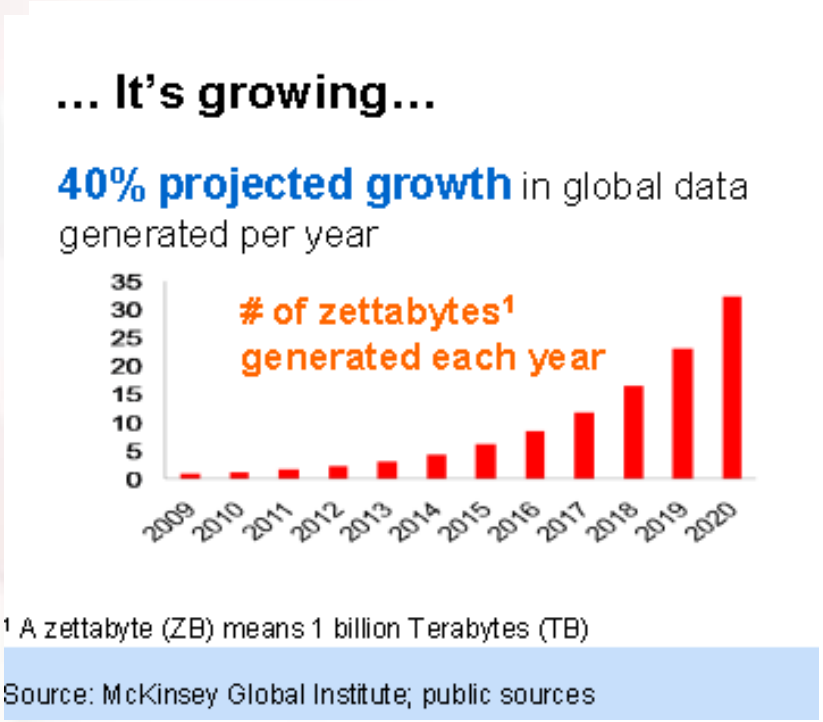
bigdatalandscape.com

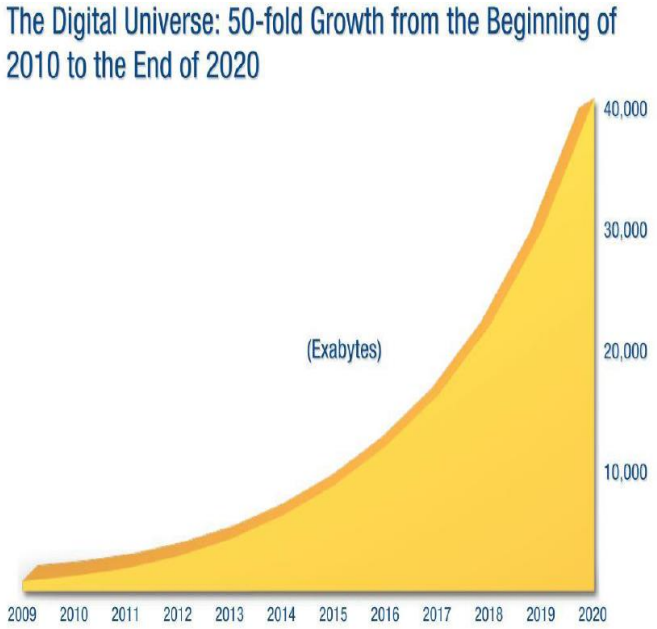
The Lexicon of Big Data

Unit	Size	Description of Scope
Bit (b)	single stored value of 0 or 1	Smallest computer memory element.
Byte (B)	8 bits	Basic unit of computing. Enough information to code a letter of the alphabet or a number.
Kilobyte (KB)	1,000 B ~ 2^{10} bytes	Derived from Greek, meaning thousand. 1 KB is approximately half a page of typed text.
Megabyte (MB)	1,000 KB ~ 2^{20} bytes	Derived from Greek, meaning great. 3.5-inch HD floppy disks held 1.44 MB of data. A CD can hold ~700 MB of data.
Gigabyte (GB)	1,000 MB ~ 2^{30} bytes	Derived from Greek, meaning giant. 1 DVD-R can hold ~4.7 GB of data.
Terabyte (TB)	1,000 GB ~ 2^{40} bytes	Derived from Greek, meaning monster. ~2,000 hours of CD quality audio. 20 years' of Hubble telescope observations has produced more than 45 TB of data.
Petabyte (PB)	1,000 TB ~ 2^{50} bytes	In 2009, Google could process ~1PB of data per hour.
Exabyte (EB)	1,000 PB ~ 2^{60} bytes	Cisco reported global IP traffic was approximately 31 EB per month in 2011.
Zettabyte (ZB)	1,000 EB ~ 2^{70} bytes	Cisco reported global IP traffic is expected to reach 1.3 ZB per year or 110 EB per month by 2016.
Yottabyte (YB)	1,000 ZB ~ 2^{80} bytes	If a 1TB hard drive costs about $100 today, it would cost $100 trillion to buy 1YB of storage.

Note: **Number of atoms in the known universe = 2^{272}**

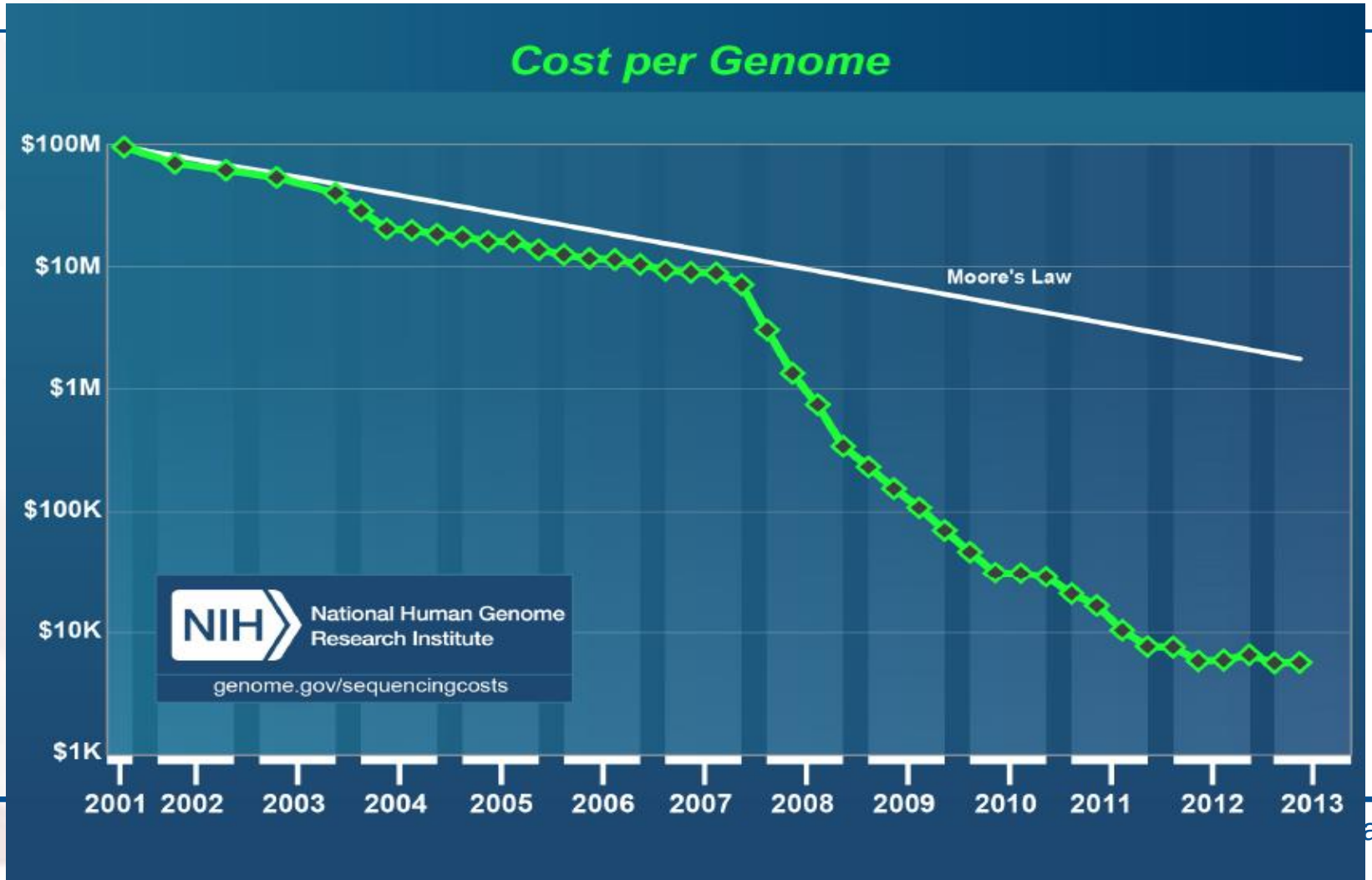
VOLUME:
EXPONENTIAL INCREASE IN GLOBAL DATA



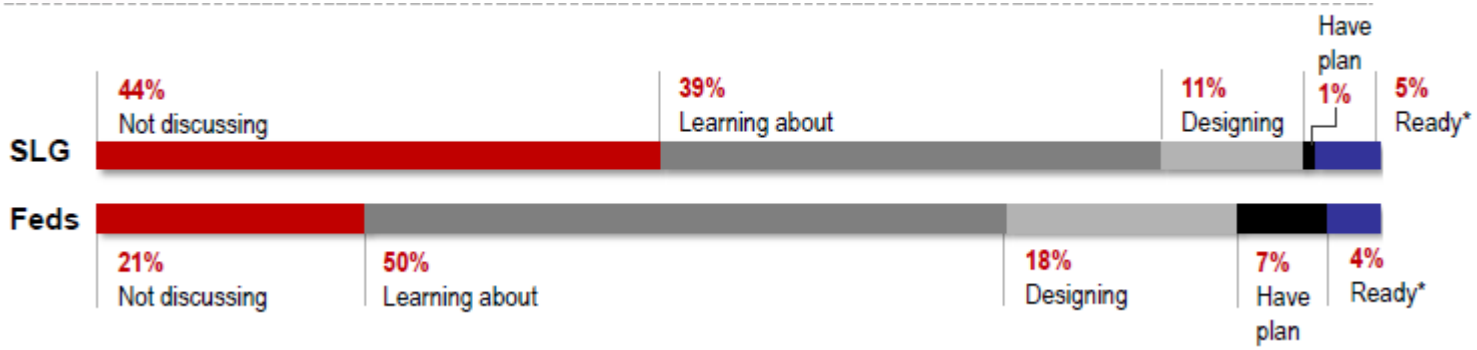


SOURCE: IDC Digital Universe Study, sponsored by EMC, Dec 2012

VELOCITY: (DATA TSUNAMI)
SEQUENCING COST PER GENOME, 2001-13



VELOCITY:
GROWTH IN STATE AND LOCAL DATA



Today, the average state and local agency stores **499 terabytes** of data

That's approximately **10 million** four-drawer filing cabinets full of text or **50 times** the printed collection of the Library of Congress*

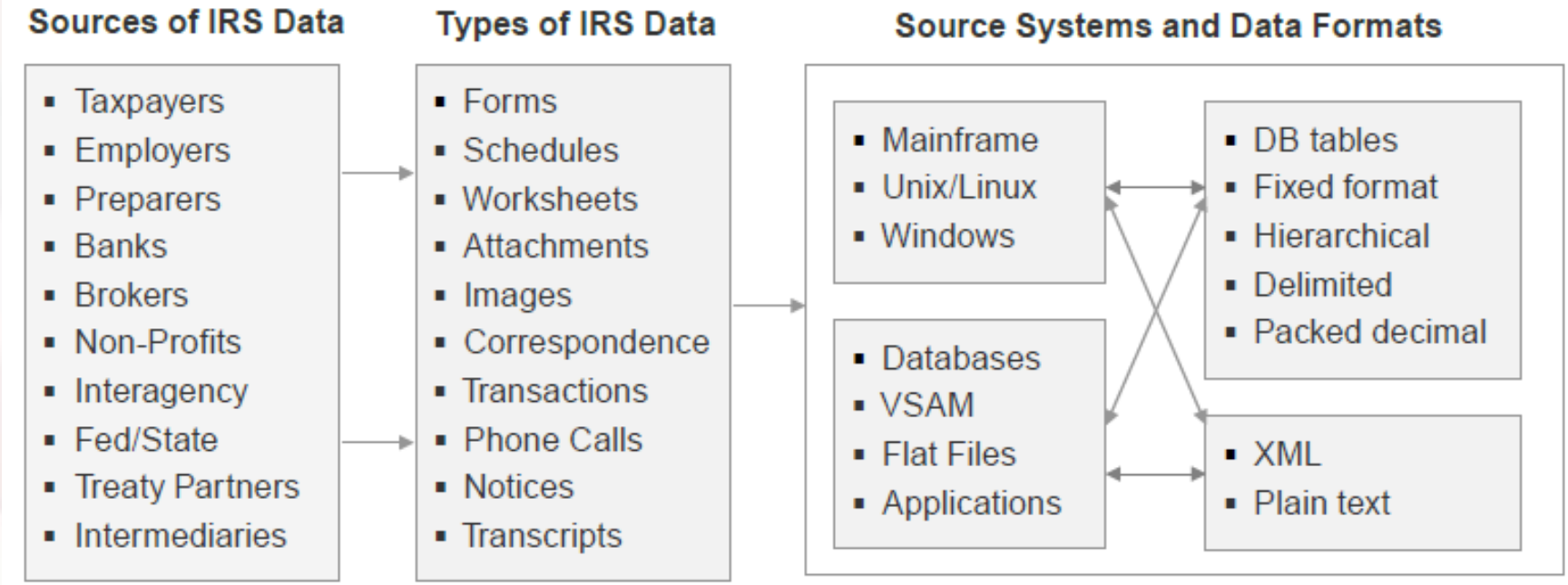
87% of state and local agencies say the size of their stored data has grown in the last two years

and...

97% expect data to grow in the next two years, by an ***average of 53%***

MeriTalk, The State and Local Big Data Gap, April 29, 2013

VARIETY:
IRS DATA CHARACTERIZED BY HETEROGENEITY

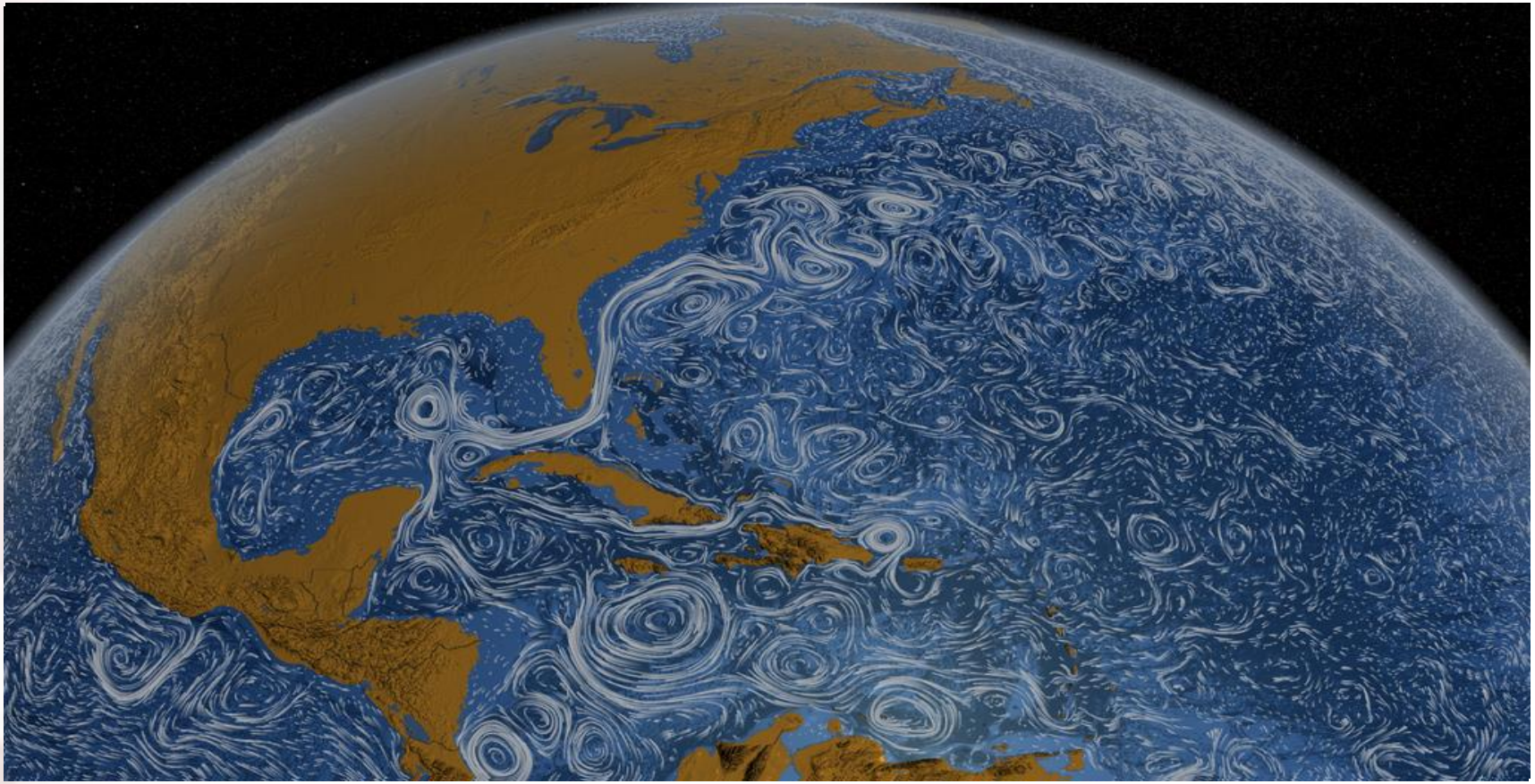


SOURCE: Jeff Butler, IRS Director of Research Databases, 2/13 Presentation on "Big Data and Analytics at IRS: Challenges and Opportunities"

VARIETY: FUSING DATA TO ANTICIPATE UNEXPECTED EVENTS (IARPA)

- Significant societal events preceded and/or followed by population-level changes in communication, consumption, and movement
- Currently analysis of news feeds, Twitter, blogs, and websearch queries for detecting disease outbreaks, forecasting product sales macroeconomic trends.
- Little research on/few methods for combining data from diverse sources
- New IARPA initiative focuses on developing new methods for aggregating multiple, noisy signals indicative of significant events

VISUALIZATION: NASA VISUALIZES THE OCEAN THROUGH MULTIPLE DATA STREAMS



SOURCE: http://www.nasa.gov/topics/earth/features/perpetual-ocean.html

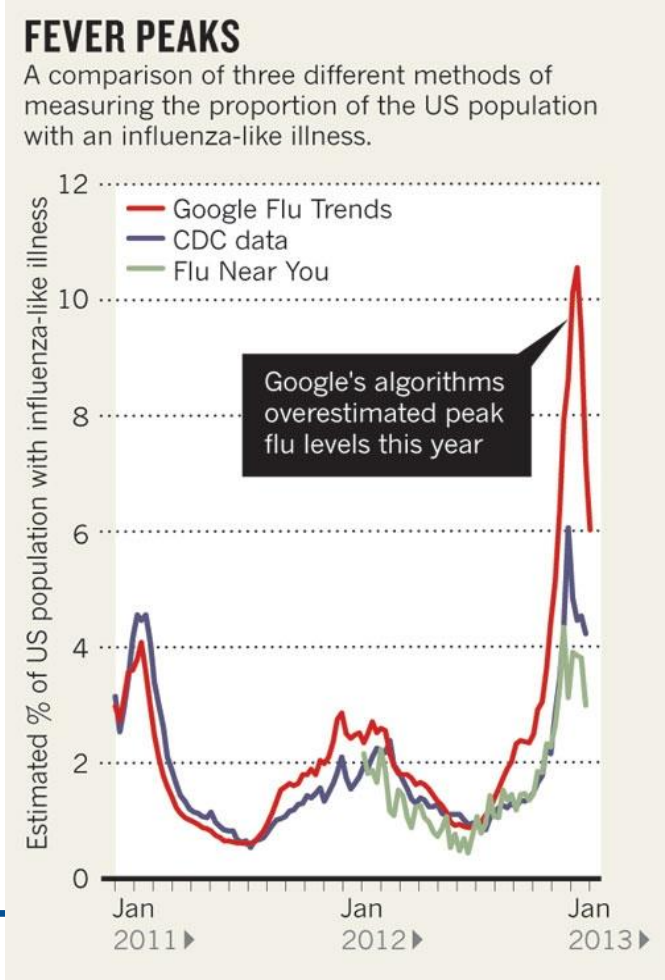
VERIFICATION: CHALLENGES OF BIG DATA

- Data sets too poorly organized to be usable; comprised of unstructured data; issues of heterogeneity; utility limited by ability to interpret & use it; more data being collected than can be stored; data sets too large to download or send over today's Internet; large & linked datasets may be exploited to identify individuals

- Boston bombing: Unchecked crowd-source investigations

- 2013 experience with Google Flu Trends indicates importance of verification

VERIFICATION: GOOGLE FLU TRENDS OVERESTIMATE PEAK FLU LEVELS (2013)



SOURCE: NATURE, 2013

VALUE: CONTRIBUTIONS OF BIG DATA

Improvements in understanding & predictive capabilities

- **Medical science: Aggregation of digital information from PET scans/other sources to illuminate progression of Alzheimer's disease in the human brain.**
- **Environmental science: Collection and retention of sensor data, along with the improvement of climate algorithms, have enabled prediction of size of the ozone hole with increasing accuracy.**
- **Weather predictions: NOAA predicts Hurricane Sandy**
- **Consumer profiling: Walmart, Facebook, Amazon**

VALUE: BIG DATA & CITIES

Data-Smart City Solutions (Stephen Goldsmith/Harvard) highlights best practices in government and data, top innovators, and promising case studies. Focus on combining integrated, cross-agency data with community data to better discover and preemptively address civic problems.

- **Predictive algorithms allow police departments to anticipate future crime hotspots.**
- **Analysis of accumulated data from subway smartcards could predict the effects of transit disruptions and give broad insight into transit-system operations.**
- **Integration of data from different human-services agencies could increase the effectiveness of social workers and others as they assist at-risk youth. Agencies and their workers could use digital tools both to collaborate and to gain new insight from their combined data resources.**

VALUE: NYC TARGETS ILLEGAL CONVERSIONS

- Bloomberg appoints NYC's first "director of analytics" (ca. 2010) to build a team of data scientists and harness city's "untapped troves of information to reap efficiencies" across multiple areas
- Datafied features of the city used to tackle "illegal conversions' (cutting up dwellings into smaller units) which can lead to fire hazards, crime, drugs, disease, pest infestation; only 200 inspectors to handle 25,000 illegal-conversion complaints a year
- Combined listing of 900,000 property lots in the city with datasets from 19 different agencies, e.g., delinquency by building owners in paying property taxes, foreclosure proceedings, anomalies in utilities usage, ambulance visits, crime rates, rodent complaints and then compared to 5 years of fire data ranked by severity; model developed refined by interviews in the field and allowed prioritization of complaints
- Improved efficiency of inspectors; spillover benefits for fire department since illegal conversions more likely to result in injury or death for firefighters

GAO

VALUE:
BENEFITS TO THE ACCOUNTABILITY COMMUNITY

- Central data analytics activities (such as the Recovery Board's Recovery Operations Center--ROC) enhances oversight capabilities of IGs by supporting fraud detection and prevention in Federal spending

- Rapidly aggregates & analyzes large, complex volumes of data to screen for potential risks or identify targets; through link analysis supports audits, investigations and prosecutions

- Potential for Congress to track funds and help reduce fraud, waste and abuse for emergency funding outlays such as the Hurricane Sandy relief effort

RECOVERY BOARD & LINK ANALYSIS

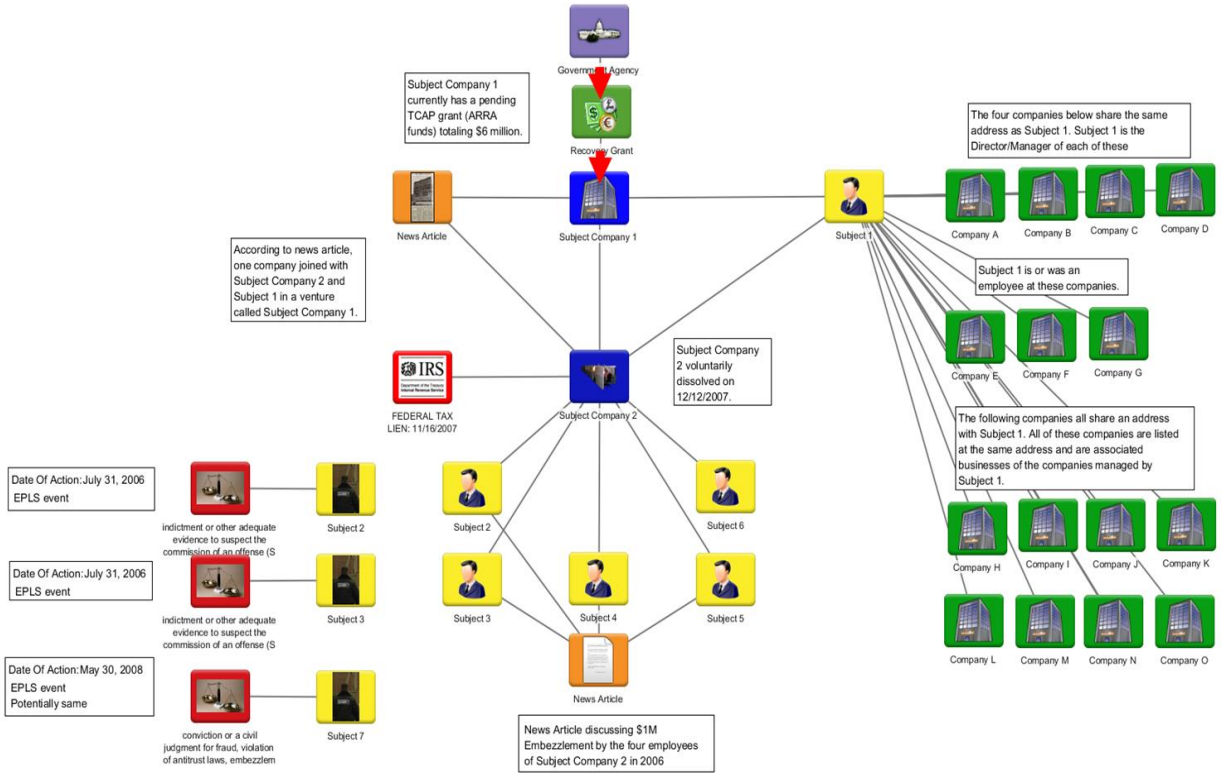
- **Analysis of 20,000 records from an OIG led to identification of 15 "bad actor" networks engaged in suspicious activities.**

- **ROC compared over 3 million people listed as employees of entities that were recipients of Recovery Act funds in CCR (Central Contractor Registration) to individuals with EPLS (Excluded Parties List System). Fuzzy logic employed to match and then the names compared to geographic information (thereby limiting instances of false positives). Numerous recipients, accounting for $2.6 million in ARRA funds, determined to have employees listed on EPLS.**

- **Both cases produced information that provided valuable intelligence into criminal activity & gave agency persons of interest to contact.**

RECOVERY BOARD & LINK ANALYSIS

How individuals and/or companies are connected, what awards an entity has received, other derogatory information (bankruptcies, arrests, and exclusions)

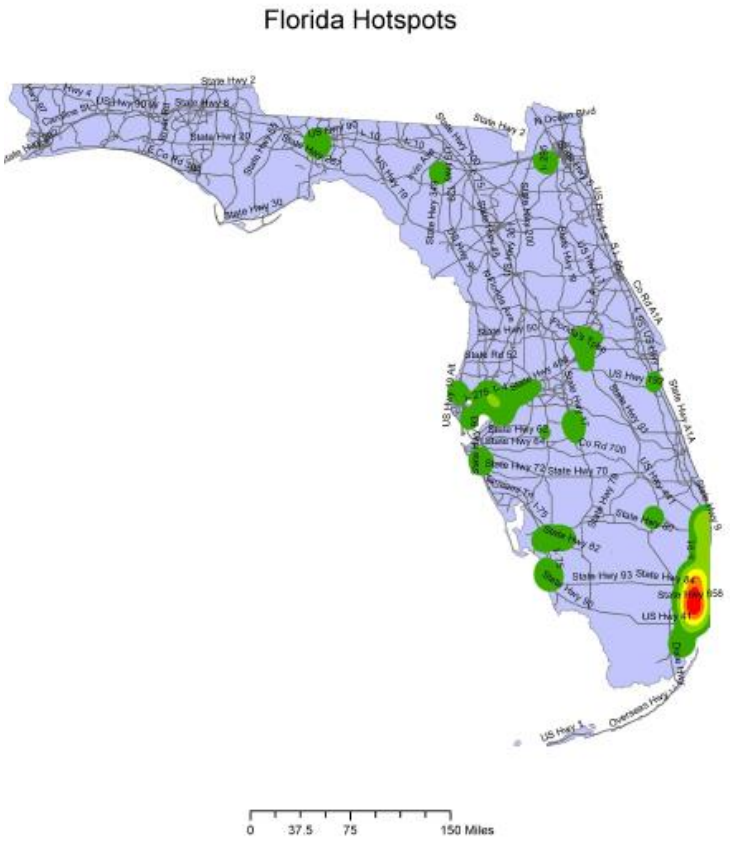


SOURCE: Recovery Board

RECOVERY BOARD & GEOSPATIAL MAPPING

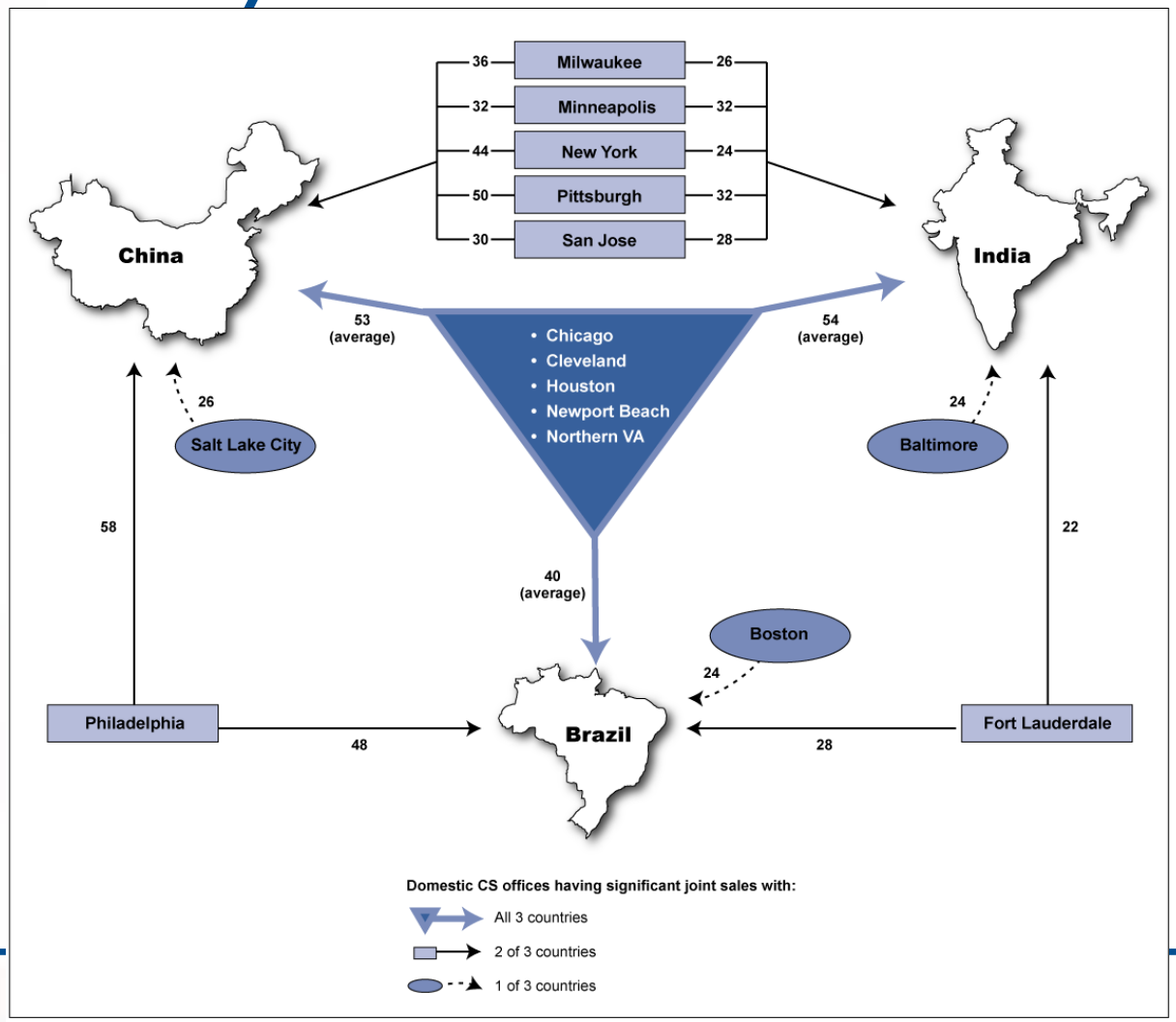
-Creation of heat maps helped investigators, auditors and evaluators focus on high-risk geographic areas

-Geospatial and mapping capabilities used to verify and validate questionable addresses found by mapping and comparing these addresses with legitimate facilities or businesses



SOURCE: Recovery Board

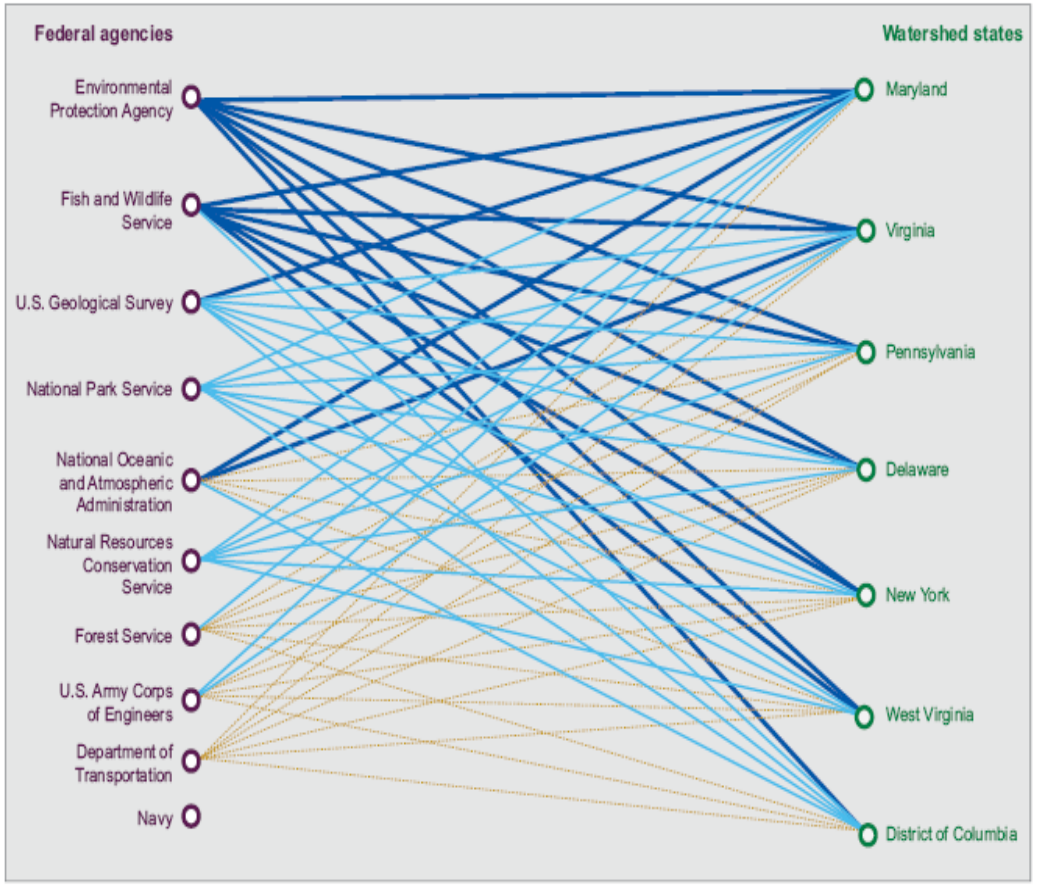
COMMERCIAL SERVICE EXPORT PROMOTION (GAO-11-909)



Sources: Commerce (data); GAO (analysis); Map Resources (maps).

CHESAPEAKE BAY CLEANUP (GAO-11-802)

Figure 3: Extent of Collaboration Needed between Federal Agencies and Watershed States to Accomplish Strategy Actions



Source: GAO analysis of survey responses.

AGENCY-SPECIFIC ACCOUNTABILITY ACTIVITIES

USDA Crop Insurance Program Compliance and Integrity Data Warehouse uses data to prevent fraudulent claim payments with estimate of more than $2.5 billion savings

170 data sources; 3 terabytes of RMA (USDA's Risk Management Agency) policy information; 120 terabytes of weather, satellite and other remotely sensed data; 1.3 million crop insurance policies; 3,200 counties

Looks for atypical patterns among insurance claims, cross-checking them with data from high-solution satellite images and weather records

RECENT GAO WORK/INITIATIVES HIGHLIGHT CHALLENGES

- GAO leads a community of practice for federal, state, and local government officials to discuss challenges and opportunities related to data sharing within and across government agencies (resulting from GAO January 2013 forum with the Council of the Inspectors General on Integrity and Efficiency and the Recovery Board; GAO-13-680SP)

- GAO report and testimony (*e.g.*, GAO-13-758; GAO-13-871T) concludes lack of data standards and a universal award identifier limit data sharing across the federal government and across federal, state, and local agencies.

- Data standardization and resolving privacy considerations are key (issues to be addressed in morning and afternoon forum panels)

HUMAN CAPITAL CONSIDERATION FOR THE ACCOUNTABILITY COMMUNITY: "ALGORITHMISTS"

- Importance of accountability, monitoring and transparency may require new types of expertise and institutions
- Algorithmists (i.e., "quants" - experts in computer science, mathematics, and statistics) would evaluate selection of data sources, choice of analytical and predictive tools, including algorithms and models, and interpretation of results
- Claim they could fill the need similar to the one accountants and auditors filled in early 20th century to handle deluge of financial information
- External algorithmists could review accuracy or validity of big-data predictions; internal algorithmists could vet big-data analysis for integrity and accuracy (analogous to external/internal audit function)

SOURCE: Mayer-Schonberger and Cukier, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK, 2013

SOURCE: McKinsey Global Institute, 2013

FUTURE HUMAN CAPITAL NEEDS?

The talent battle will become a critical strategic imperative

	Big data savvy	Deep analytical	Big data infrastructure
Definitions	Employees who can define key questions that data can answer and have basic knowledge of statistics	Specialists who have conduct data analysis and advanced training in statistics and/or machine learning and	IT personnel who serve as database administrators and programmers
Occupations[1]	▪ Business and functional managers ▪ Budget, credit and financial analysts ▪ Engineers ▪ Life scientists ▪ Market research analysts ▪ Survey researchers ▪ Industrial-organizational psychologists ▪ Sociologist	▪ Actuaries ▪ Mathematicians ▪ Operations research analysts ▪ Statisticians ▪ Mathematical technicians ▪ Mathematical scientists ▪ Industrial engineers ▪ Epidemiologist ▪ Economists	▪ Computer and information scientists ▪ Computer programmers ▪ Computer software engineers for applications ▪ Computer software engineers for system software ▪ Computer system analysts ▪ Database administrators
Potential gap by 2018	~1.5 Million	~150,000	~300,000

1 Occupations are defined by the Standard Occupational Code (SOC) of the US Bureau of Labor Statistics and used as the proxy for types of talent in labor force.

IMPLICATIONS OF BIG DATA TREND

- **Potential value of Big Data for government missions, *e.g.*, real-time information and decision-making, especially for data-dependent agencies (impact on regulatory process, conduct of surveys)**
- **Identification of principles for Big Data governance issues**
- **Identification of best practices, e.g., privacy, for use of Big Data**
- **Human capital considerations apply to both external evaluation/audit and ability to conduct internal analysis**
- **Big Data considerations for auditing**

Advanced Analytics at GAO

GAO is developing pilots around data analytic technologies.

Pilot concepts include:

- Data mining for improper payments analysis
- Link analysis for fraud identification
- Document clustering and text mining for overlap and duplication analysis
- Network analysis for program coordination assessment

•Preliminary indications include:

- A substantial decrease in labor and time inputs in analyzing documents and their content
- A possible increase in quality and number of findings
- Enhanced visualization for more efficient communication of key findings

Thank you

personst@gao.gov