Big Data: Public Sector Opportunities, Challenges, and Implications

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

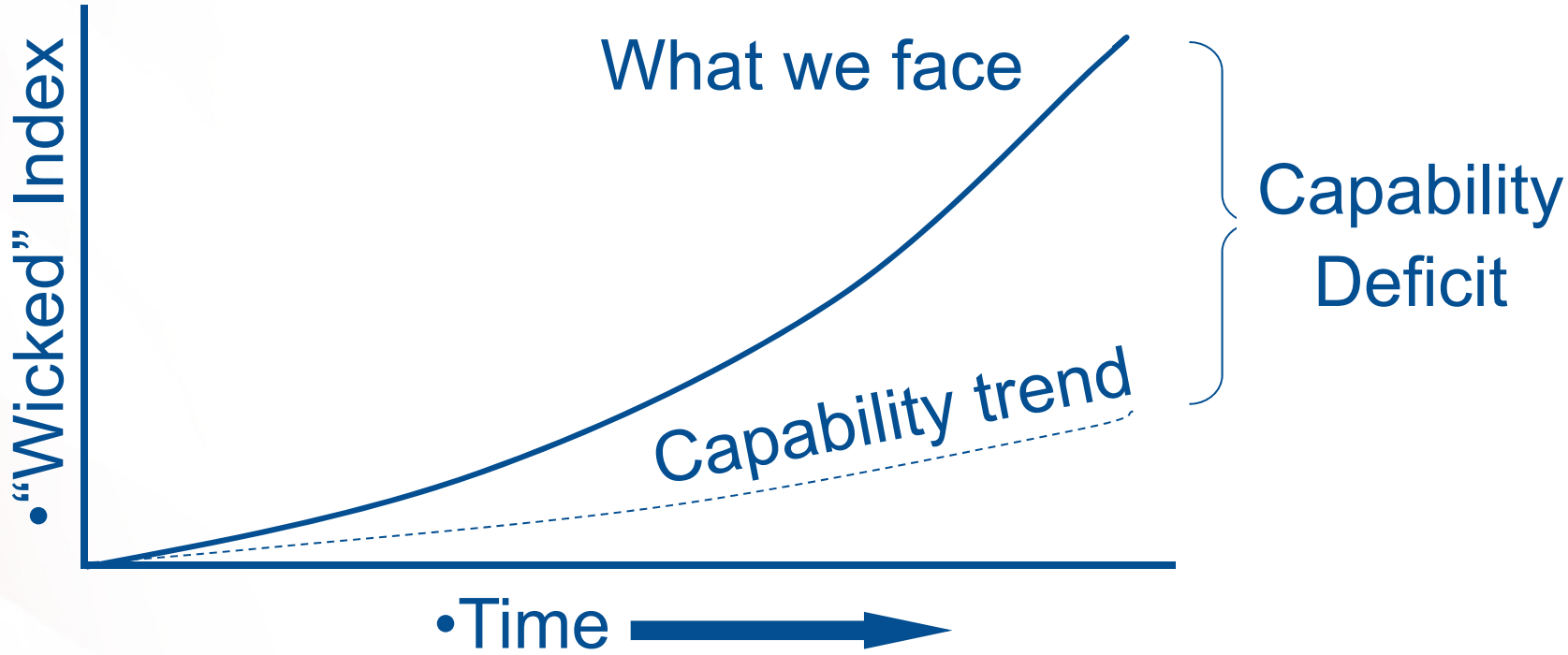
Presentation to the NEIAF Fall 2015 Conference

October 29, 2015

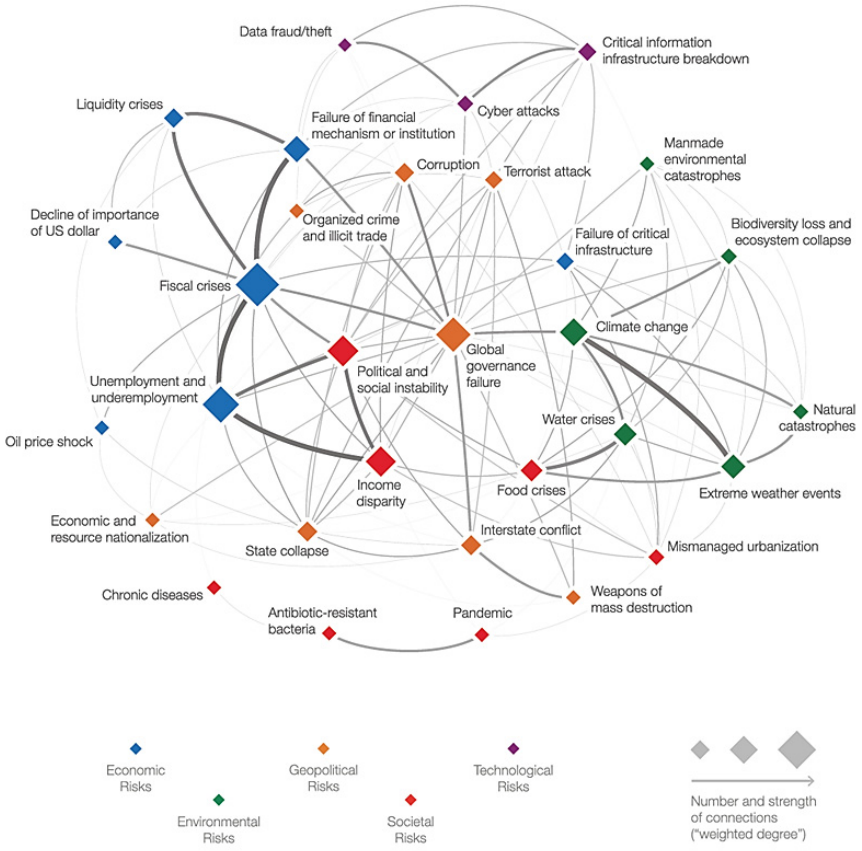
U.S. Government Accountability Office

- Is an independent, nonpartisan agency serving the Congress and helps improve the performance and ensure the accountability of the federal government.
- Core values are Accountability, Integrity, and Reliability
- Oversight, Insight, and Foresight
- To ensure independence, the Comptroller General (CG) is appointed to a 15-year term by the President. Other than the CG, there are no political appointees at GAO.

"Wicked" Problems



UNCERTAINTY & COMPLEXITY: Interconnected Risks



Source: Global Risks Perception Survey 2013-2014. Global Risks 2014 Report. World Economic Forum, Switzerland, 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.

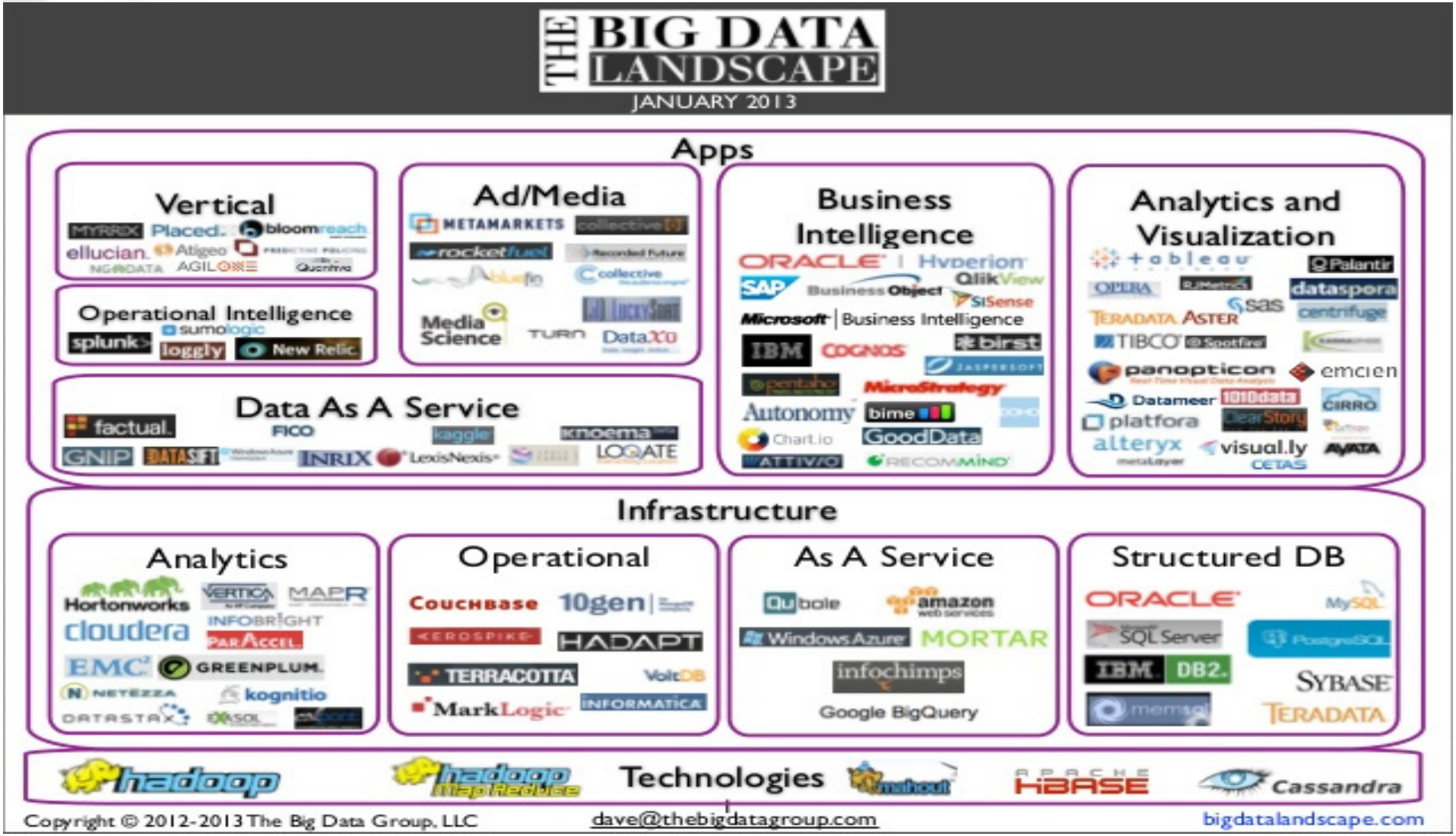
Technologies required to efficiently process large quantities of information, e.g., massively parallel processing frameworks such as shared nothing relational databases, MapReduce programming frameworks, and cloud infrastructure

***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

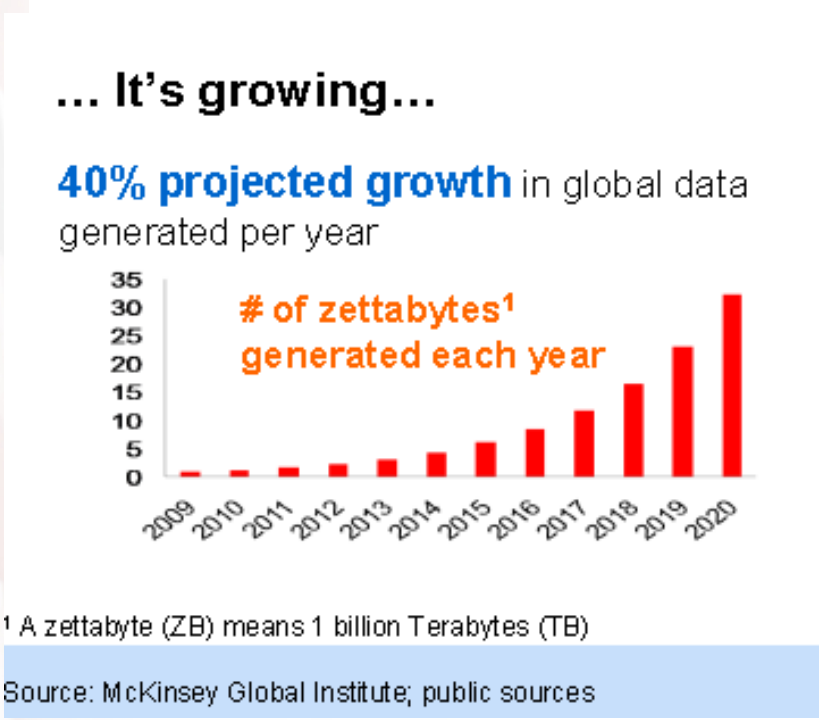


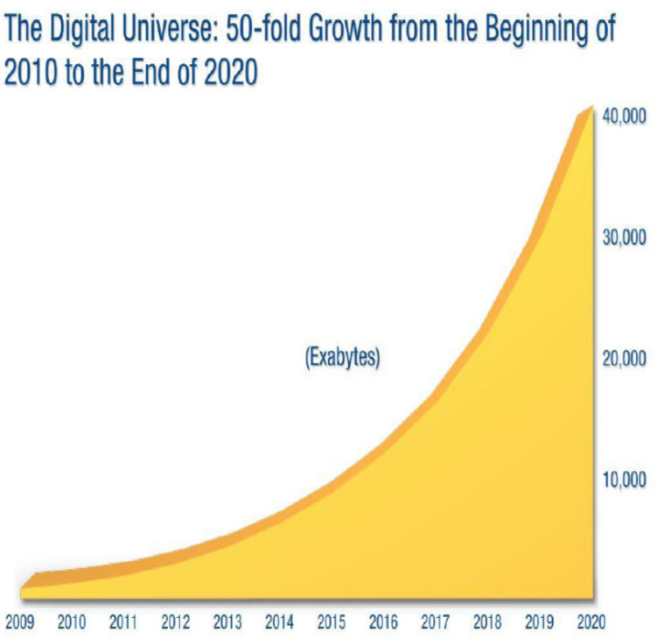
The Lexicon of Big Data

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}**

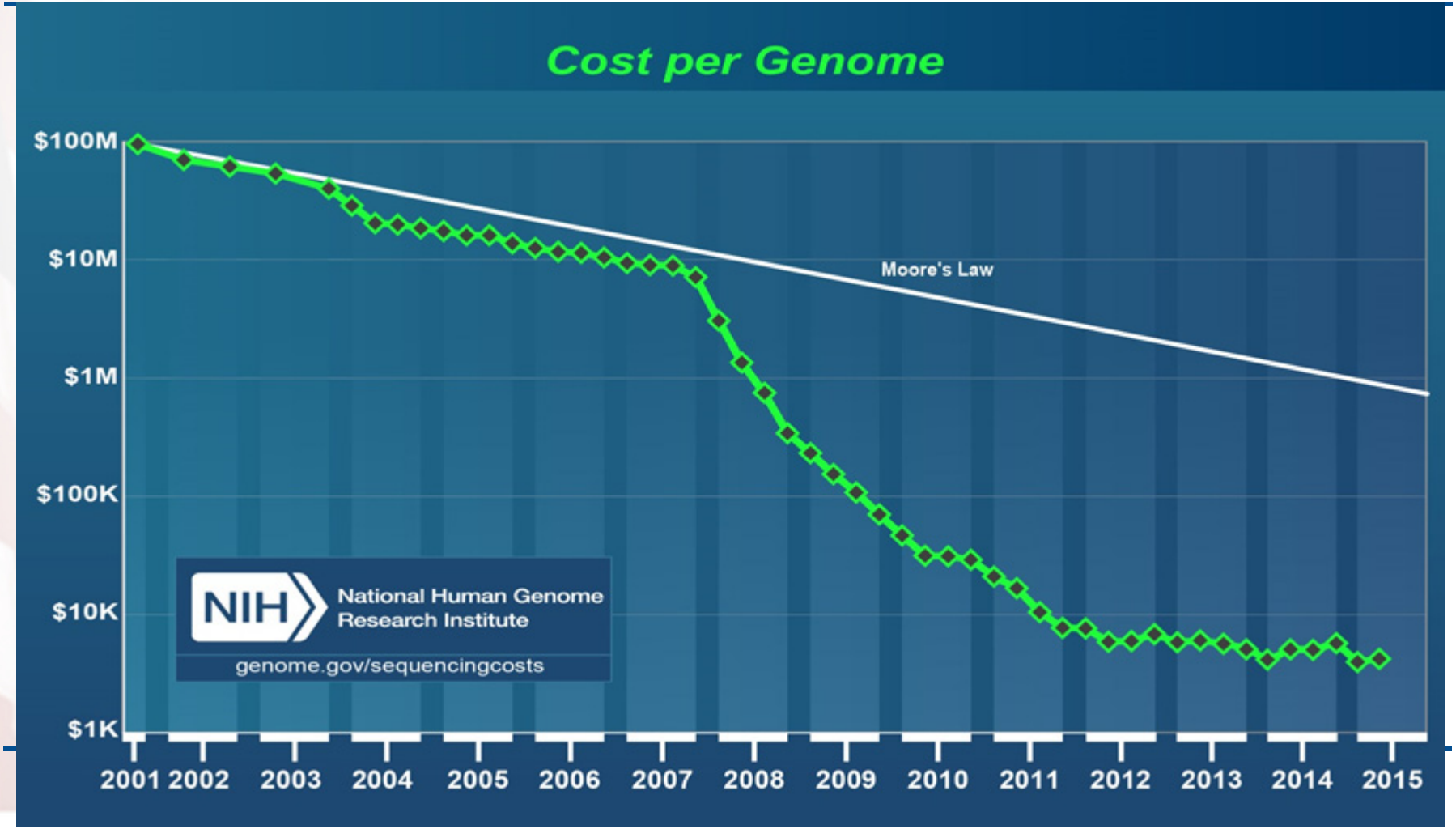
VOLUME:
EXPONENTIAL INCREASE IN GLOBAL DATA



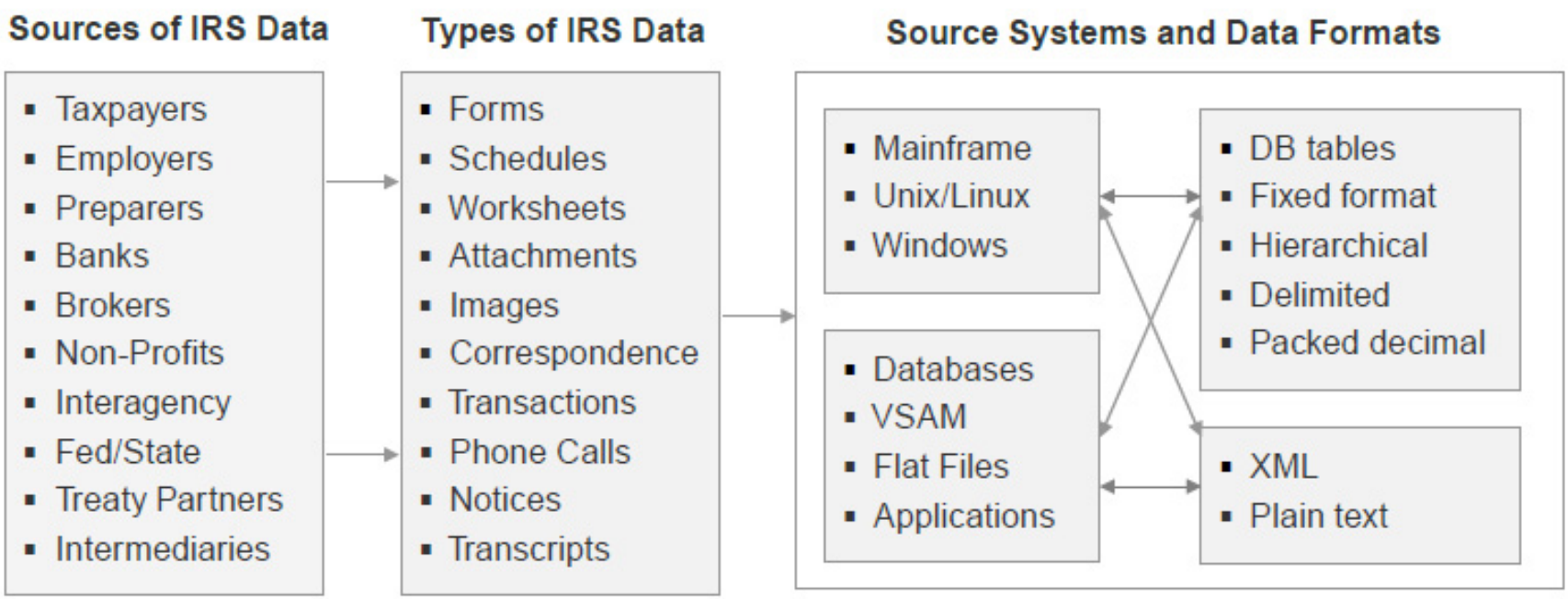


SOURCE: IDC Digital Universe Study, sponsored by EMC, Dec 2012

VELOCITY: (DATA TSUNAMI)
SEQUENCING COST PER GENOME, 2001-15

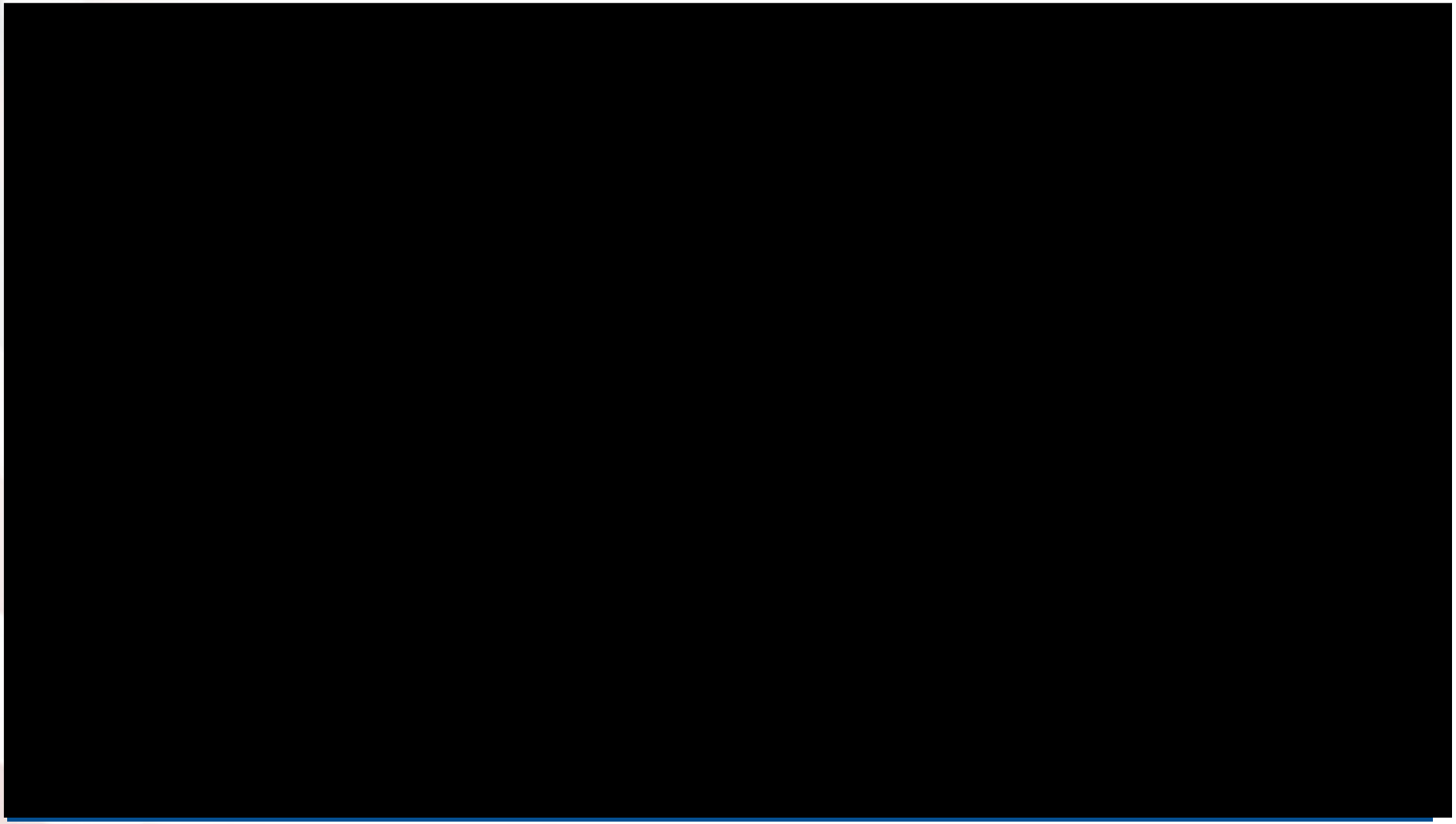


VARIETY:
IRS DATA CHARACTERIZED BY HETEROGENEITY



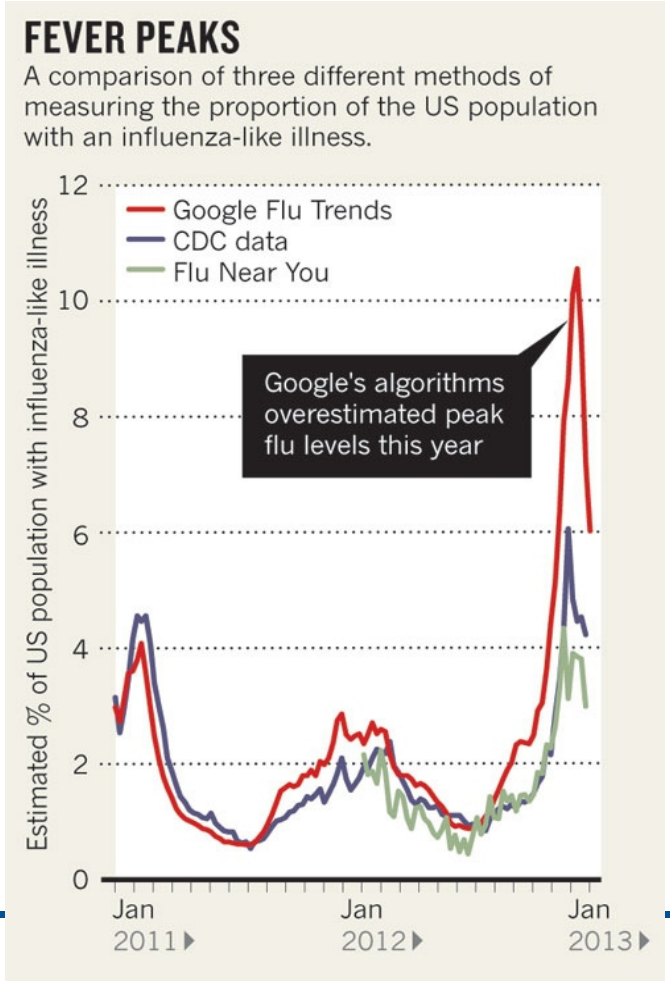
SOURCE: Jeff Butler, IRS Director of Research Databases, 2/13 Presentation on "Big Data and Analytics at IRS: Challenges and Opportunities"

VISUALIZATION: NASA Perpetual Ocean Data



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

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



SOURCE: NATURE, 2013

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.

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

USDA CROP INSURANCE PROGRAM

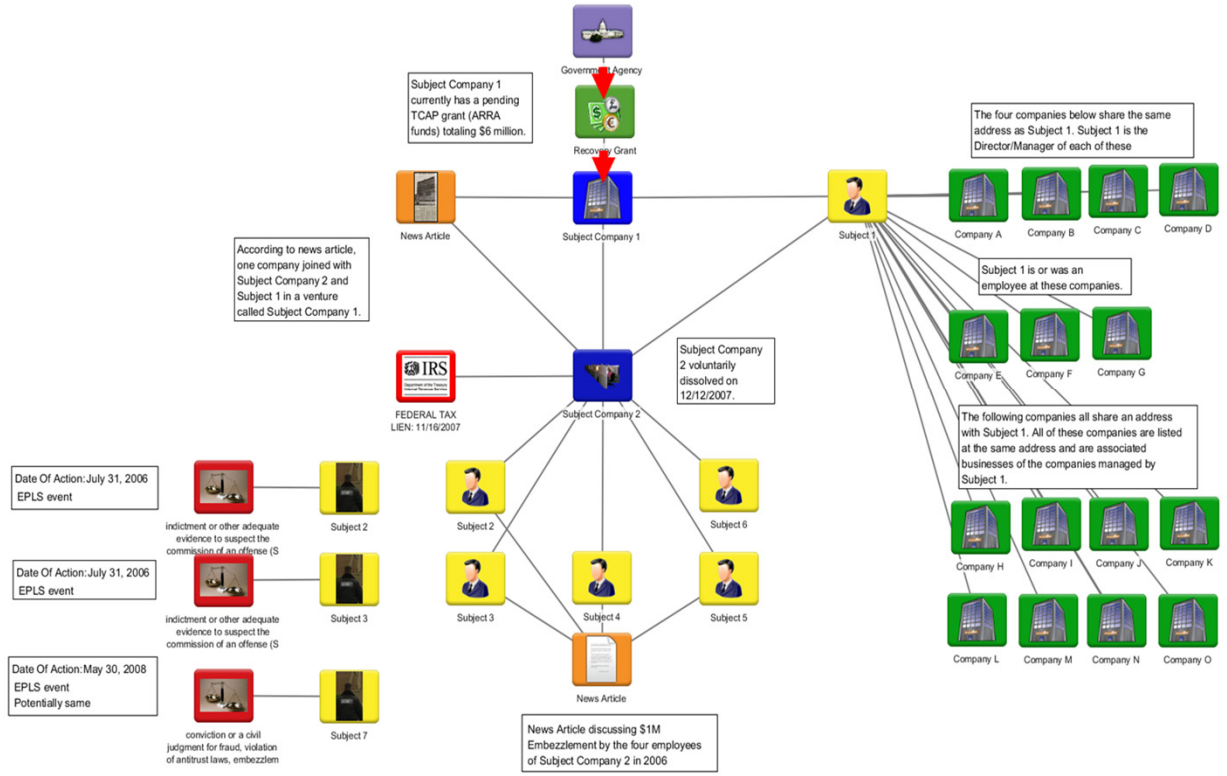
USDA Crop Insurance Program Compliance and Integrity Data Warehouse uses data to prevent fraudulent claim payments with estimate of more than $2.5 billion savings

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

RECOVERY BOARD & LINK ANALYSIS

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

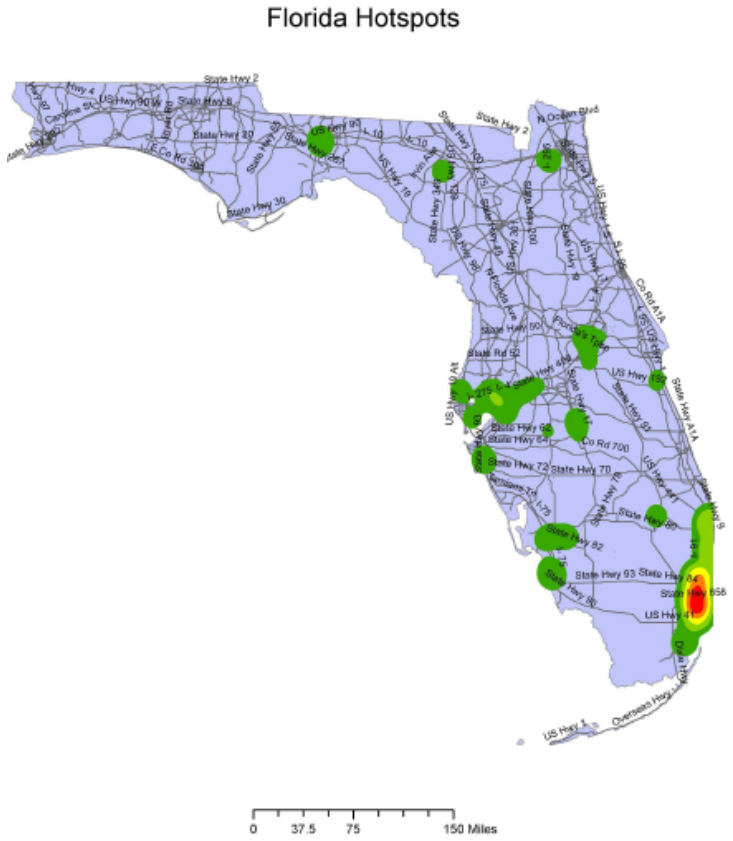


SOURCE: Recovery Board

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

LEGISLATION

Digital Accountability and Transparency Act of 2014 (DATA Act)

- Creates standards for reliable and transparent data to increase accountability

Government Performance and Results Act Modernization Act of 2010 (GPRA Modernization Act)

- Requires quarterly assessments to evaluate agency performance and improvement, which generates measurable data to help agencies achieve goals

Improper Payments Elimination and Recovery Improvement Act of 2012 (IPERIA)

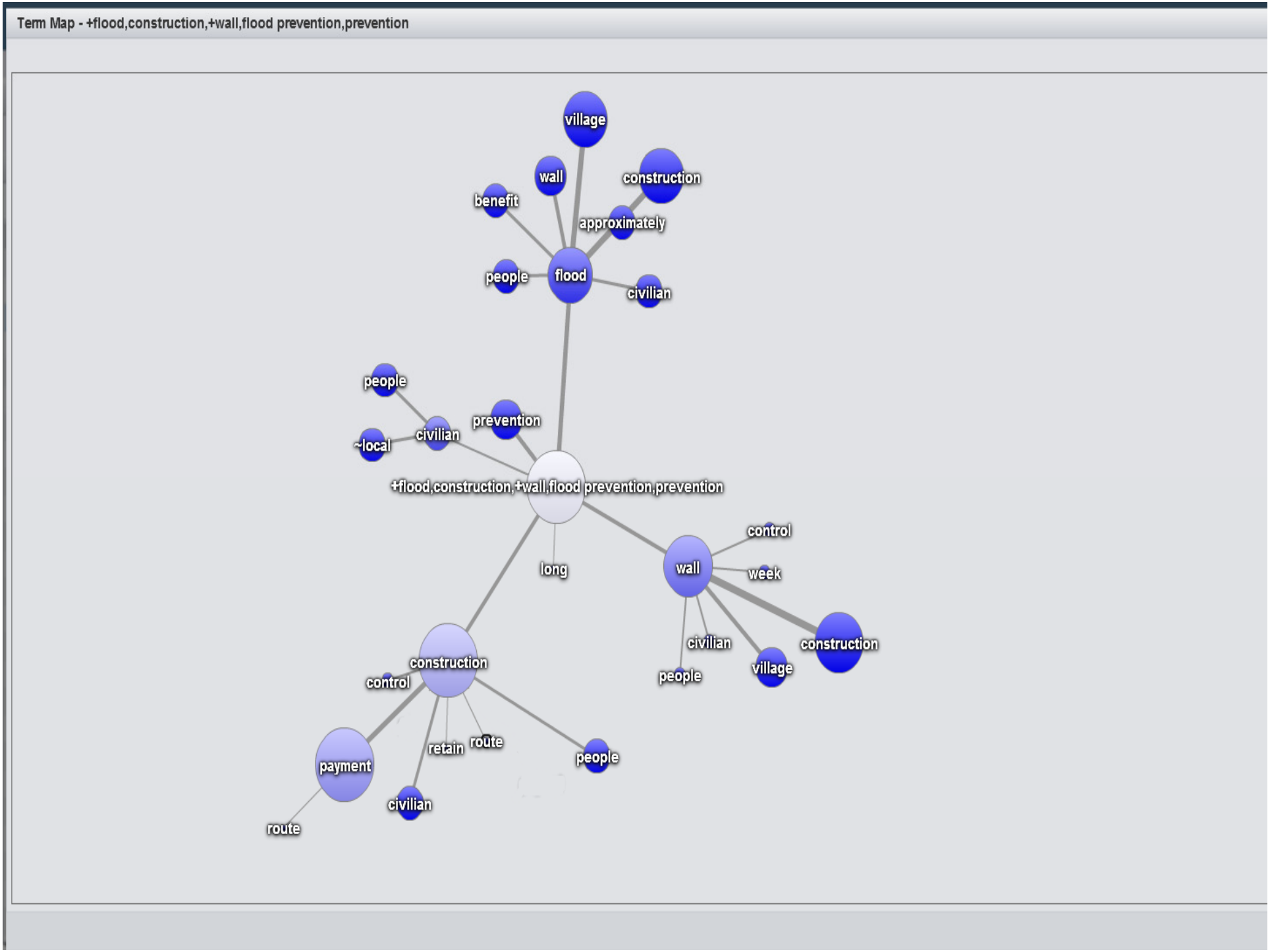
- Increases efforts to identify, prevent, and recover payment error, waste, fraud, and abuse by promoting data sharing techniques amongst federal agencies

How data analytics can be used: improper payments

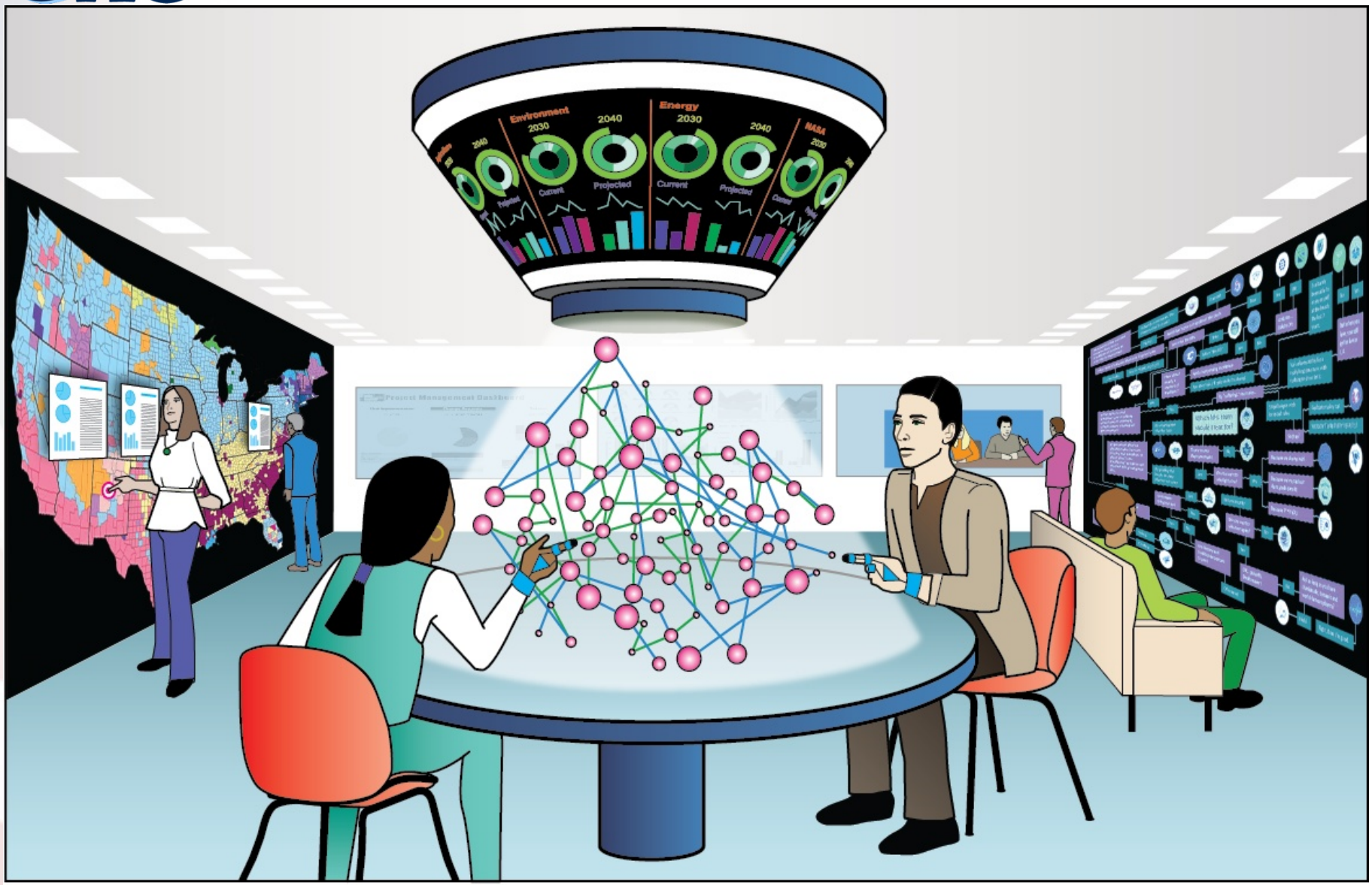
•Top 10 program improper payment estimates by dollar amount

Program	Agency	Reported Improper Payment Estimates	
		Dollars (in billions)	Error rate (percent of outlays)
Medicare Fee-for-Service	HHS	$36.0	10.1%
Earned Income Tax Credit	Treasury	$14.5	24.0%
Medicaid	HHS	$14.4	5.8%
Medicare Advantage (Part C)	HHS	$11.8	9.5%
Unemployment Insurance	Labor	$6.2	9.3%
Supplemental Security Income	SSA	$4.3	8.1%
Supplemental Nutrition Assistance Program	USDA	$2.6	3.4%
Old Age, Survivors, and Disability Insurance	SSA	$2.4	0.3%
Medicare Prescription Drug Benefit (Part D)	HHS	$2.1	3.7%
National School Lunch Program	USDA	$1.8	15.7%

•Source: Estimates reported by OMB for FY 2013.


Term Map - +flood,construction,+wall,flood prevention,prevention
village
wall
construction
benefit
approximately
people
flood
civilian
people
prevention
civilian
~local
+flood,construction,+wall,flood prevention,prevention
control
long
wall
week
civilian
construction
construction
people
village
control
payment
retain
route
people
civilian
route

WHAT WILL THE AUDIT OFFICE OF 2030 LOOK LIKE?



Source: GAO.

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.

DATA ANALYTICS @ 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

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

Fair Information Privacy Principles (FIPS)

- Transparency
- Individual Participation
- Purpose Specification
- Data Minimization
- Use Limitation
- Data Quality and Integrity
- Security
- Accountability and Auditing

Case Study – MIT Study Indicates Threat of Metadata

- Metadata shown to identify individuals without their names or account numbers (*i.e.*, even after data anonymization)
- MIT researchers analyzed anonymous credit card transactions of 1.1M people (3 months of purchase records across 10,000 stores)
- Using a new analytic formula, they needed only four bits of secondary information (metadata such as location or timing) to identify the unique individual purchasing patterns of 90% of people involved
- De-anonymyzation accomplished by linking unique purchasing pattern with publicly available info on LinkedIn and Facebook
- Researchers could tell women and men apart and could also pick out people in higher income brackets just by how long they lingered at different shops

Other Case Studies Indicating Big Data Challenges to Privacy

- Cambridge Univ. (2013) reported that the pattern of "likes" posted by people on Facebook unintentionally exposed their political and religious views, drug use, marital status, and sexuality
- UC-Riverside (2015) psychologists "likes" were a more accurate measure of someone's personality than the assessment of their close friends
- Uber (2014) showed how it could combine customer records of late-night trips in major cities with local crime reports to calculate the likelihood that any given rider was visiting prostitutes

Comptroller General Forum on 21st Century Data and Analytics: Cross-sectoral Impacts

Under the direction of the CG and at the request of several Committees and members, GAO, with the assistance of the U.S. National Academies, will conduct a forum on data and analytics in the 21st Century in order to support the strategic thinking of the U.S. *vis-à-vis* its transformative and disruptive nature as it continues to rapidly evolve and exhibit cross-sectoral impacts.

Objectives are:

1. Overview: Status and Trends (*e.g.*, the rise of the Internet of Things)
2. Opportunities (*i.e.*, innovation and competitiveness) and challenges (*i.e.*, privacy and civil liberties) from Big Data
3. Key Considerations for current and potential policies

Three profiles will illustrate benefits…and challenges including those to privacy and civil liberties. Economic sectors considered: Health care, Transportation systems, and Financial Markets

Comptroller General Forum on 21st Century Data and Analytics: Public Sector Impacts

Under the direction of the CG and at the request of several Committees and members, GAO, with the assistance of the U.S. National Academies, will conduct a forum on data and analytics in the 21st Century in order to support the strategic thinking of the U.S. *vis-à-vis* its transformative and disruptive nature as it continues to rapidly evolve and exhibit public sector impacts.

Objectives are:

1. Overview: Status and Trends
2. Opportunities (*i.e.*, fraud, waste, abuse, improper payments, program evaluation) and challenges (*i.e.*, privacy and civil liberties) from Big Data
3. Key Considerations for public sector function and current and potential policies

Three profiles will illustrate benefits…and challenges including those to privacy and civil liberties. Public sector issues considered: Fraud/waste/abuse mitigation, streamlining and efficiencies, improper payments, program evaluation/auditing

Thank you

personst@gao.gov

(202) 512-6412

@GAOChfScientist

http://www.linkedin.com/pub/timothy-persons/9/856/9ba/en

http://www.gao.gov/technology_assessment/key_reports