

Data Analytics at GAO

Naba Barkakati
Chief Technologist
U.S. Government Accountability Office

DATA Act

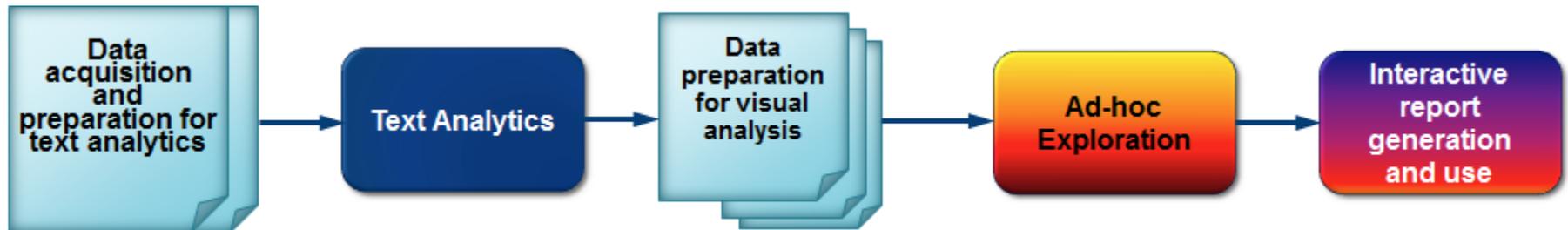


Some of its requirements:

- Treasury and OMB to develop government-wide financial data standards and issue related guidance
- Treasury to establish a data analysis center
- OMB and Treasury to consult with public and private stakeholders in establishing data standards --
<http://fedspendingtransparency.github.io/>
- Agency IGs to report on agencies' spending data quality and the use of data standards
- GAO to review IG reports and assess agencies' data quality and implementation of the data standards.

SAS Contextual Analytics applied to an old GAO review

Determining duplication of recovery efforts in Afghanistan; are two companies who are funding recovery efforts overlapping effort?



SAS Contextual Analysis for text analytics
SAS Enterprise Guide for data manipulation
SAS Visual Analytics for ad-hoc exploration and reporting

Results of SAS Analytics

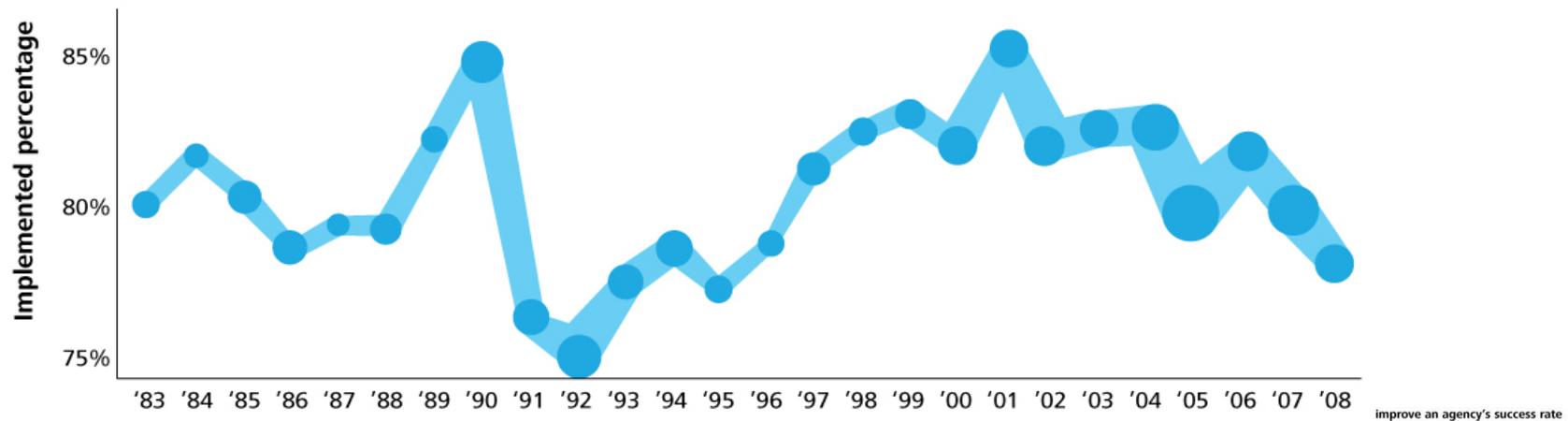
Identified 3 potential duplication areas and took 3 hours to complete the analysis



Deloitte's analysis of 26 years worth of GAO reports

Full details at: <http://dupress.com/articles/text-analytics-and-gao-reports/>

Figure 1. Completion rates for GAO recommendations over time



Note: Size of ticker indicates the total number of recommendations GAO made in each year.

Tools used by Deloitte:

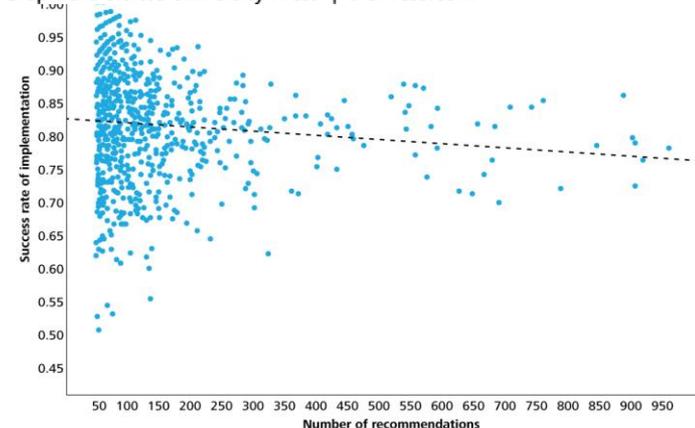
Python for website scraping

SPSS Text Analytics

Megaputer Polyanalyst

SAS for econometric analysis

Graphic: Deloitte University Press | DUPress.com



Source: Deloitte Research data analysis.

Graphic: Deloitte University Press | DUPress.com

Focus on Advanced Analytics at GAO

- 2013 GAO-CIGIE-RATB Forum on data analytics for oversight and law enforcement
- Community of practice focused on data-sharing challenges:

http://www.gao.gov/aac/gds_community_of_practice/overview Email: GovernmentDataShare@gao.gov

- DATA Act related activities
- Senior executives to lead analytics at GAO
- Acquiring tools and skills such as Python, R, SAS
Contextual analytics, etc.
- Pilot projects in planning stages



Potential Data Analytics pilots at GAO

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

Some Observations

- Not necessarily “Big Data” but messy data
- Text analytics helpful for analyzing documentary evidence -- the staple of GAO audits
- Link analysis for fraud investigations
- Knowledge of Python, R or SAS, text analytics + subject matter expertise -- difficult to get it all in one person
- XML-tagged text would help
- Plans for new content creation and distribution system at GAO, expectation is that the content would be XML

