The Impact of Big Data on Auditing (and vice versa)

Doug Laney, VP Information Strategy & Innovation
@douglas_laney
Key Issues

• How can and should auditing practices leverage Big Data and Big Data technology?

• Has the eminence of Big Data made information itself a legitimate corporate asset?
Key Issues

• How can and should auditing practices leverage Big Data and Big Data technology?

• Has the eminence of Big Data made information itself a legitimate corporate asset?
Big Data is #1
What is Big Data Anyway?

Technical Challenges and Opportunities

New forms of PROCESSING

Business Challenges and Opportunities

VOLUME

VELOCITY

VARIETY

DECISIONS

INSIGHTS

AUTOMATION

© 2014 Gartner, Inc. and/or its affiliates. All rights reserved.
Beyond Basic Business Intelligence

Gartner Analytic Continuum

VALUE

Hindsight

Insight

Foresight

DIFFICULTY

Inertia

Descriptive Analytics

What happened?

 Drill Down/Across
 Olap
 Data Discovery
 Scheduled Reports
 Query & Ad hoc Reporting

Value

Diagnostic Analytics

What happened?

 Forecasing
 Regression Analysis
 Multivariate Stats
 Pattern Matching
 Visualization
 Network/Cluster Analysis
 Data/Text Mining

Predictive Analytics

What will happen?

Forecasting
Regression Analysis
Multivariate Stats
Pattern Matching
Visualization
Network/Cluster Analysis
Data/Text Mining

Prescriptive Analytics

How can we make it happen?

Graph Analytics
Heuristics/Rule Engines
Complex Event Processing (CEP)
Operations Research (Optimization)
Simulation (Monte Carlo)
Neural Networks

Imperative

Alerts
Semantic & Sentiment Analysis
DRILL DOWN/Across
OLAP
Data Discovery
Scheduled Reports
Query & Ad hoc Reporting

Gartner Analytic Continuum
Information Without Borders

Correlations and patterns from disparate, linked data sources yield the greatest insights and transformative opportunities.
Which is the biggest opportunity for Big Data?

Source: Gartner Big Data Strategy Essentials Webinar, March 2013 & 2014
By 2017, over 50% of analytics initiatives will make use of event data streams generated from instrumented machines, applications and/or individuals.
Risk and Fraud Examples
Making the grade by setting a new curve

• Opportunity
  - Predict future financial performance, health, issues, risks better than industry standard methods

• Data and Analytics
  - Continuously crunch over 5 dozen publically-available factors to generate a hundred-point FHR (financial health scale) vs. subjective quarterly financial analyst ABC grades

• Result
  - Most financial analyst firms excited/blinded by MF Global hiring Jon Corzine, but RapidRatings saw its FHR score slide from 49 to 26, advising investors to run not hide
Sniffing and snuffing insurance fraud

• Opportunity
  - Save and make money by reducing fraudulent auto insurance claims

• Data and Analytics
  - Predictive analytics against years of historical claims and coverage data
  - Text mining adjuster reports for hidden clues, e.g. missing facts, inconsistencies, changed stories

• Results
  - Improved success rate in pursuing fraudulent claims from 50% to 88% and claim investigation time by 95%
  - Additional $12 million in subrogation recoveries
  - Marketing to individuals with low propensity for claim fraud
Cutting through the corporate performance Fog

- **Opportunity**
  - Improve investor and portfolio performance by anticipating corporate earnings

- **Data and Analytics**
  - Text analyzed over 50,000 corporate annual reports using the Gunning Fog readability index (e.g. number of words per sentence and the number of syllables per word), and overall report length

- **Results**
  - More complicated annual reports are reflective of companies with lower year-over-year earnings persistence
  - Companies anticipating performance issues tend to use complex, obfuscatory language and writing
Projecting project success or failure

• Opportunity
  - 20 years of unused “dark data” (archived emails) incurring cost without benefit
  - Consulting project issues reported after the fact; no way to anticipate and address them

• Data and Analytics
  - Identify project communications (emails) from prior successful vs unsuccessful projects
  - Mine emails for communication and sentiment patterns (time-of-day, CC/BCC, length, content)

• Results
  - Early warnings of scope, budget, technical, personnel, compliance issues
  - Improved project success and reduced penalties
  - Productized project indicator benchmarks
Tomorrow’s news, today

• Opportunity
  - Identify systemic, related or mounting health, geopolitical or financial issues

• Data and Analytics
  - Continually capture, translate, classify and analyze 40,000 local news reports in 43 languages from around the world
  - Custom charting, alerts, animated map visualization of categories or search terms
  - Predict and highlight emerging hot topics and trends

• Results
  - Enable politicians, insurers, investors, disease control institutions and global support groups to respond quicker to local or global issues
Analytics Detects and Affects Defects

• Opportunity
  - Identify the root cause of production frailty to reduce total cost of production

• Data and Analytics
  - Big Data Warehouse (BDW) captured product line environmental metrics: humidity, pressure, temperature and lighting, from 1000s of sensors
  - EDW with production frailty data from production system.

• Results
  - Hybrid Analytics uncovered the responsiveness of the temperature control system to be the root cause
  - Lower production efficient ratio by 1%+, equivalent to $US millions per year.
Crediting Big Data and Analytics for Fraud Reduction

• Opportunity
  - Identify and reduce credit card transaction fraud

• Data and Analytics
  - Hadoop-based analytics can model fraud down to the actual merchant terminal
  - Has increased 10-fold to 500 the number of transaction aspects analyzed, and can add new ones to models in under one hour (vs. 2-3 days)
  - Runs 16 distinct fraud models for different geographic and market segments

• Results
  - Approaching ability to analyze all transactions, up from only 2%
  - Identifies $2B more fraud situations annually
Voice of L’Customer

• Opportunity
  - Build sustainable relationships with beauty product consumers

• Data and Analytics
  - Collects, cleanses, classifies and understands the sentiment of millions of online posts per day from dozens of listening posts via Clarabridge
  - Tracks interests of 17,000 Facebook fans and 1.5M “likes”, and 41,000 Twitter followers
  - Identifies actionable consumer comments and routing to the “Voice of Beauty” command center for appropriate handling

• Results
  - Transformed how brand awareness and loyalty are monitored and leveraged
  - Enables real-time, direct interaction with consumers
Finding Fraud Faster

• **Opportunity**
  - A rapidly growing Turkish credit card business targeting lower value segments caused an increase in fraud

• **Data and Analytics**
  - Replaced manual process of credit card application review with automated real-time scoring and flagging
  - Increase from 13% to 100% of applications reviewed
  - Implemented fraud modeling in 15 days using KXEN

• **Results**
  - Increased number of identified actual fraudulent applications by 3x; 92% of fraud cases identified
  - Reduced number of fraud alerts from 300,000 to 30,000 per quarter by tuning and discovering new patterns
  - Saving $25,000 per day; ROI achieved in *one week*
Using Data Dumps to Find Grease Dumpers

• Opportunity
  - Tracking grease dumpers in NYC that account for 60% of backups citywide in 7400 miles of sewer lines

• Data and Analytics
  - “Dark data” on local restaurants from the city’s Business Integrity Commission, license data from the Department of Health and Mental Hygiene were used to compare restaurants that did not have a grease carter with geo-spatial data on sewers

• Results
  - 95% success rate in tracking down dumpers
  - Eliminated 30 million pounds of debris from sewers providing 2 million gallons of extra sewer capacity
  - Reduced sewer backup-related costs for businesses and homeowners
Queasy Tweets

• **Opportunity**
  - Determine which restaurants are prone to serving up foodborne illnesses

• **Data and Analytics**
  - Researchers analyzed 3.8M tweets from 94,000 users tweeting about or while at 23,000 NYC area restaurants.
  - Their nEmesis machine learning algorithm then followed these users for 72 hours to look for mentions of terms related to nausea.

• **Results**
  - Over four months the system spotted 480 reports of possible food poisoning, which correlated with poor health department scores for these restaurants
  - Ability for NYC Dept. of Health to proactively investigate restaurants
Buildings Have Magnetic Personalities Too

- **Opportunity**
  - Understand activities within a building, without having to go inside

- **Data and Analytics**
  - Indoor Atlas maps the interior of a building using the unique magnetic “fingerprint” of the structure caused by distortions of the Earth’s magnetic field.
  - Determines location of internal structures and people to within six feet, including which floor

- **Results**
  - Emergency personnel can use a Google Map overlay to navigate a building
  - Competitors can see changes to layouts and foot traffic
  - Pay $99/mo to keep your building’s magnetic map private
Outperforming Traditional Outbreak Monitoring

• Opportunity
  - Early identification of disease outbreaks

• Data and Analytics
  - Automatically sifts through millions of posts on dozens of social media sites, local news reports, medical workers’ social networks and government websites to track instances of disease
  - Sophisticated filtering algorithm to reduce “noise”
  - Continually plots disease hotspots on a map

• Results
  - Identified a cluster of “mystery hemorrhagic fever” in Guinea over a week before the Ministry of Health of Guinea notified the World Health Organization (WHO), that a day later confirmed the Ebola outbreak
  - Assist the cash-strapped WHO
Wicking away content quality issues

• Opportunity
  - Improve the ultimate quality of Wikipedia entries

• Data and Analytics
  - The largest free user-generated content repository in the world: Wikipedia (10M+ articles)
  - Dynamic Bayesian network and information extraction techniques, multivariate Gaussian distributions to model data quality dimension distributions for reach quality class, and combined trained classifiers

• Results
  - Outperforms human reviewers by 23% in correctly classifying the quality rank of a given article
  - Potential to focus Wikipedia editors on articles with low quality rankings and automatically “lock” high quality articles yielding overall improvements to Wikipedia quality and editing productivity
Open Data is in da’ House

• Opportunity
  - Improve accuracy of home appraisals

• Data and Analytics
  - Newly open US government data sets on economy, population and housing in machine readable format (per president Obama’s executive order) plus historical and current home sale data
  - 250,000 proprietary analytic models in R run each night against 100M+ homes on a hybrid architecture (internal server farm plus Amazon Cloud)

• Results
  - Shrink home sale price prediction error rate from 13.6% to 6.9%
  - Forecasts the change in a home’s value over the next 12 months and now accommodates variability within a zip code
Homing in on the Predictive Power of Search Analytics

• Opportunity
  - Improve predictability of real estate market trends

• Data and Analytics
  - Housing related search data from Google Trends
  - Simple seasonal autoregressive (AR) and linear regression models

• Results
  - Beat National Association of Realtors predictions of future home sales by 23.6%, and out-predicted the Case-Shiller house price index
  - A 1% increase in search frequency about real estate agents ~ selling 3,520 homes per quarter per US state
  - Overcame systemic lag in government reporting of economic factors
  - Showed that consumers’ “digital traces” can be compiled to reveal economic intentions and activities
Key Issues

• How can and should auditing practices leverage Big Data and Big Data technology?

• Has the eminence of Big Data made information itself a legitimate corporate asset?
Worst Responders

Insurance Industry
Modified Commercial General Liability Policy Standard

Legal System
Conflicting Rulings Worldwide

Accounting Profession
Modified International Accounting Standard (IAS) 38
Hearing? In one ear and out the other.

US Senate Hearing on Adapting a 1930's Financial Reporting Model to the 21st Century

Our current measurement and reporting system, both internal and external to the organization, misclassifies or overlooks entirely significant sources of value.

Our current measurement and reporting systems in the United States do not capture emerging sources of value. The measurements we use don’t reflect all the ways that companies create or destroy value in the New Economy.

The current measurement and reporting system can exacerbate business risk and stock volatility.

http://banking.senate.gov/00_07hr0g/071900/samek.htm
80% of executives believe the value of their company’s information assets are reported on their balance sheet.
Infonomics is the economic theory of information as new asset class, and the discipline of accounting for, managing and deploying information just as any other enterprise asset.

Wikipedia: Infonomics
LinkedIn: Center for Infonomics discussion group
Twitter: #infonomics
Everyone Agrees; Nobody Agrees

Information is one of our greatest competitive assets.

Information is one of our greatest performance assets.

Information is one of our most risky assets.

Information is one of our greatest pains in the asset.

Throughout the enterprise, people talk about information as an asset, but do they really mean it?
So What Is an Asset Anyway?

- **Webster**: A single item of ownership having exchange value or convertible into cash. Total resources of a person or business such as cash, notes, and goodwill.
- **American Institute of CPAs**: Any economic resources (tangible/intangible) that can be owned or produce value. Assets have a positive economic value.
- **Financial Accounting Standards Board**: A probable future economic benefit obtained or controlled by a particular entity as a result of past transactions or events.
- **International Accounting Standards Board**: A resource controlled by the enterprise as a result of past events and from which future economic benefits are expected to flow to the enterprise.

*Information arguably meets all the criteria of what defines an asset.*
**So What Is an Asset Anyway?**

<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Webster</td>
<td>A single item of ownership having exchange value or convertible into cash. Total resources of a person or business such as cash, notes, and goodwill.</td>
</tr>
<tr>
<td>American Institute of CPAs</td>
<td>Any economic resources (tangible/intangible) that can be owned or produce value. Assets have a positive economic value.</td>
</tr>
<tr>
<td>Financial Accounting Standards Board</td>
<td>A probable future economic benefit obtained or controlled by a particular entity as a result of past transactions or events.</td>
</tr>
<tr>
<td>International Accounting Standards Board</td>
<td>A resource controlled by the enterprise as a result of past events and from which future economic benefits are expected to flow to the enterprise.</td>
</tr>
</tbody>
</table>

Information arguably meets all the criteria of what defines an asset.
Assets or Not?

Not just a trick question, but a tricky one that vexes businesses.
Investors Recognize Your Information’s Value, Why Don’t Your Accountants?

![Graph showing Tobin's q* ratio over time. The x-axis represents years from 1950 to 2010, and the y-axis represents the ratio of market value to tangible assets. The graph compares Infoprodct Companies (4.7) and Infocentric Companies (2.4) with the Market Average (1.1). The text states: "You are here?" and "Market caps eclipse tangible asset value." * Ratio of market value to tangible assets.
Three Degrees of Information Value

- **Vision Gap**
  - If you applied information to all relevant business processes

- **Performance Gap**
  - Your expected capabilities and plans

- **Realized**
  - Your current capabilities and execution

Periodically measuring and closing these gaps is more important than the measurement methods or values themselves.
Why Quantify Information's Value?

• Improve the management of information
• Create a common language for IT, business leaders and CFO to communicate about information
• Become a more infocentric business (optimizing and innovating)
• Justify and prove benefits of IT initiatives
• Better leverage or monetize one of the enterprise's most underutilized resources
• Drive improved corporate market valuations
ISO 27001
Guideline for Valuing Information Assets

You’ll never guess what it doesn’t include
Gartner’s Models for Quantifying Information Value?

- Intrinsic Value of Info (IVI)
- Business Value of Info (BVI)
- Performance Value of Info (PVI)

Noneconomic Value

- Cost Value of Info (CVI)
- Economic Value of Info (EVI)
- Market Value of Info (MVI)

Economic Value
Information Valuation in Practice

- Prioritizing analytics initiatives
- Justifying various EIM initiatives
- Changing employee behavior
- Planning for ways to monetize data
- Pricing syndicated data
- Validating infosec. investments
Recommendations

• Begin including exogenous and “dark” data sources in auditing and other risk-oriented analytics

• Brush aside basic BI capabilities. Apply data science techniques and technologies to discover and disclose anomalies, patterns, relationships and root causes.

• Anticipate and prepare for an auditing burden opportunity from an increasing reality (if not formal recognition) of information itself as a corporate asset

“Our present intuition about the nature of information will be remembered as narrow and shortsighted.

The dominant principle of the new economy, the information economy, has lately been to conceal the value of information, of all things.”

— Jaron Lanier
Who Owns the Future