www.pwc.com

Enhancing Quality In Transforming Public Sector Auditing

Big data analytics: how do you use it to find fraud?

pwc

Neil Meikle

- Associate Director, PricewaterhouseCoopers Consulting Services
- In charge of the forensic technology team in Malaysia
- Forensic investigations – reacting to fraud and corruption crisis situations
- Background in computer forensics and forensic data analytics in the UK
- Focus on large data analytics and e-Disclosure projects in London
- Advise clients on data as an opportunity and a risk

Overview

- Organisations are capturing ever larger amounts of data
- Through data analytics, the public and private sector are increasingly making data-driven decisions
- In recent years, we have been using data analytics tools and techniques to support our forensic investigations
- Electronic data is now one of the main sources of evidence in our forensic investigations
- Even though our investigations are usually reactive, versus the pro-active nature of auditing, is there an opportunity to use big data analytics in audits too?
- Can we use data analytics to uncover fraud pro-actively?

Increasing volumes of data

Five key drivers are changing the environment in which organisations operate today

Chart from The Washington Post @ http://www.washingtonpost.com/wp-dyn/content/article/2011/02/10/AR2011021004916.html

Big data analytics – consensus in the definition

Defined around the 3 "V's": Volumes, Velocity & Variety

Data Volumes

- Big data analytics deals with data volumes of 10's of Tb's to Pb's where traditional BI deals with Gb's to 10's of Tb's
- Volumes will be further driven by massive data growth in unstructured and externally, user generated data including text, speech and social media
- *"All data created from the beginning of time until 2003 is now being created every 2 days"*

Data feed Velocity

- High frequency of data generation & data delivery
- Data processing speeds to keep up with streaming data analysis and taking / recommending actions, in real time
- *"340 million tweets on Twitter and 250 million photos on Facebook added everyday"*

Data type Variety

- Structured (relational), semi structured & unstructured data
- In addition to internal structured data (transactions, operational), semi & unstructured data such as text, emails, logs, blogs, click streams, web / XML /RSS feeds, audio and video material, sensor data
- *"Only 5% of data is structured in a format suitable for traditional Business Intelligence"*

Increasingly data-driven decisions

Data analytics is the process of generating insight from raw structured data

- **Data** – low level information
  - Some inherent structure and context
  - Not easily understood in its raw state and in high volumes

- **Data Analytics** – applying a structured process to answer our questions
  - Inspecting, profiling, cleansing, matching, transforming, re-orienting, modelling…
  - Data mining and modelling are particular data analysis techniques that support knowledge discovery for predictive purposes
  - Identifying useful information
  - Drawing conclusions
  - Pro-actively supporting decision making

More data means more potential insight in all industry sectors

- ***Finance and retail*** (e.g. pricing and risk analytics)
- ***Utilities*** (e.g. smart usage analysis)
- ***Pharmaceuticals and health*** (e.g. smart patient monitoring and diagnosis)
- ***Supply chain and inventory*** (e.g. efficiency improvement through simulation modelling, stock management)
- ***Marketing and CRM*** (e.g. customer profiling and segmentation, customer acquisition and retention , customer value and profitability)
- ***Fraud investigation and prevention*** (e.g. insider fraud, bribery, corruption)


The Economist

Schumpeter

Building with big data

The data revolution is changing the landscape of business

May 26th 2011 | from the print edition

IN A short story called "On Exactitude in Science", Jorge Luis Borges described an empire in which cartographers became so obsessive that they produced a map as big as the empire itself. This was so cumbersome that future generations left it to disintegrate. ("[I]n the western deserts, tattered fragments of the map are still to be found, sheltering some occasional beast or beggar.")

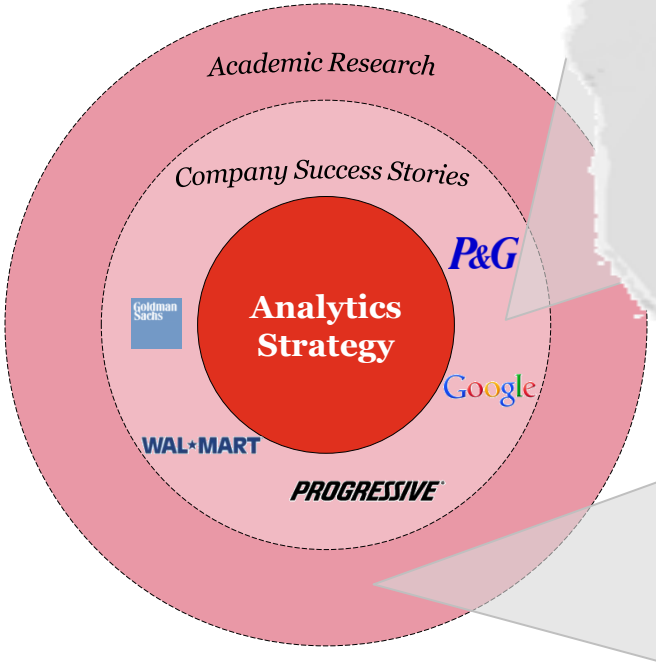
As usual, the reality of the digital age is outpacing fiction. Last year people stored enough data to fill 60,000 Libraries of Congress. The world's 4 billion mobile-phone users (12% of whom own smartphones) have turned themselves into data-streams. YouTube claims to receive 24 hours of video every minute. Manufacturers have embedded 30m sensors into their products, converting mute bits of metal into data-streaming nodes in the internet of things. The number of smartphones is increasing by [illegible] and the number of sensors by 30%.

Data analytics – why is it important?

1. Rapid increase in available data makes it a viable source of insight
   - In just four hours on "black Friday" 2012, Walmart in the USA handled 10 million cash register transactions – almost 5,000 items per second.
   - VISA processes more than 172,800,000 card transactions each day.

2. Data analytics techniques are being deployed across many different industries and for a range of purposes
   - "Companies are using [data analytics] tools to improve business efficiency, spot trends and opportunities, provide customers with more relevant products and services and, increasingly, to predict how people, or machines, will behave in the future."
     - Big data in the spotlight as never before, 26th June 2013, Financial Times

Companies that differentiate themselves through analytics have superior productivity and profitability; academic research confirms this



- There is an observed positive correlation between data-driven decision-making and firm performance
- In a study of 179 large public firms, those that emphasized decision-making based on data and business analytics experienced output and productivity that is **5-6% higher than what would be expected given their other investments and information technology usage**
- Data-driven decision making was defined not only by collecting data, but also by **how it is used — or not — in making crucial decisions**

Brynjolfsson & Hitt, MIT/Wharton 2011

- If the median Fortune 1000 business **increased the usability of its data by 10%,** it would translate to an increase in **$2.01 billion** in total revenue every year …

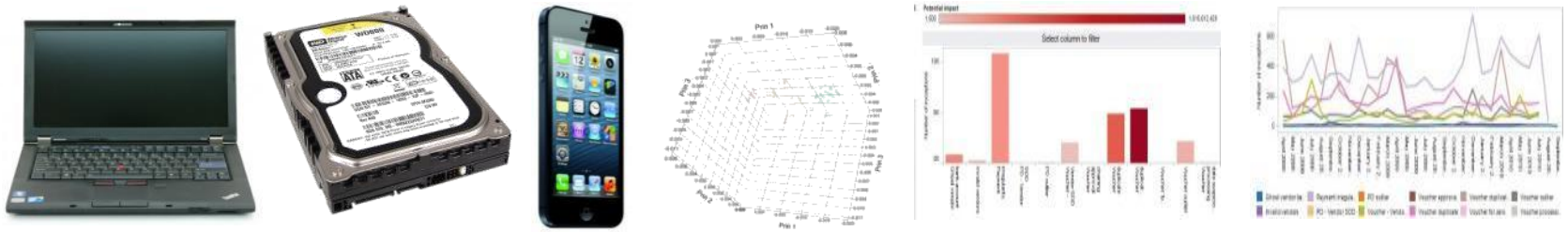
"Measuring the Business Impacts of Effective Data," University of Texas at Austin

How are we using data analytics on our forensic investigations?

Forensic investigations involve reacting to a crisis situation – such as…

- Bribery
- Corruption
- Asset misappropriation
- Fraud
- Accounting fraud
- Cyber-crime
- Competition / anti-trust issues
- Money laundering
- Market abuse
- Regulatory non-compliance

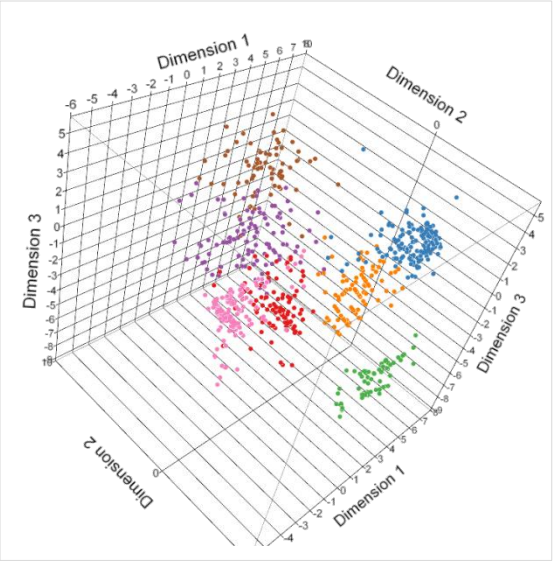
Forensic investigations now rely heavily on electronic evidence and forensic technology



- When fraud or corruption is identified or suspected (or even if we simply want to pro-actively prevent it), electronic data is a key source of evidence
- Forensic technology is used to uncover evidence and insight from data: either using forensic data analytics techniques or computer forensics techniques (e.g. deleted file recovery, handphone forensics, Internet evidence)
- We can use forensic technology tools and techniques to secure data in a forensically sound manner and search it for evidence
- We treat all potentially relevant evidence as if it will be presented in court

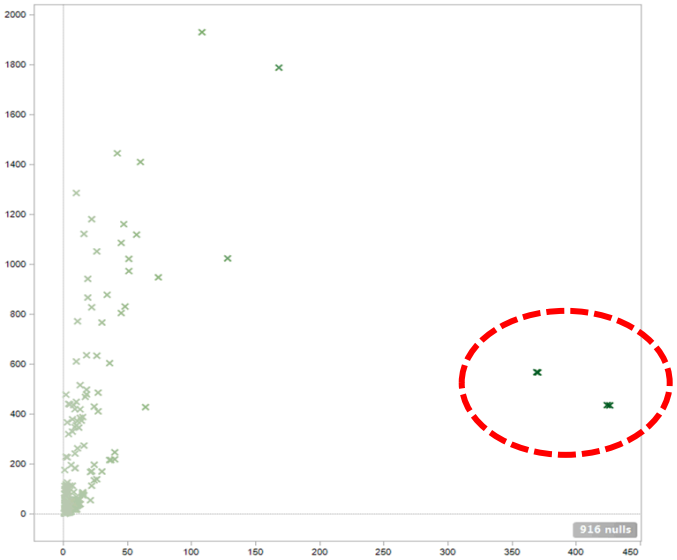
How we support our forensic investigations by transforming raw transactional data into insight

- Raw data = transactions
- Data recovered from financial or other operational database systems
- There are usually:
  - Large volumes of data
  - Many transaction / data types
  - Difficult questions to answer
  - Complex processing, querying and analysis
- And we need to get answers QUICKLY!

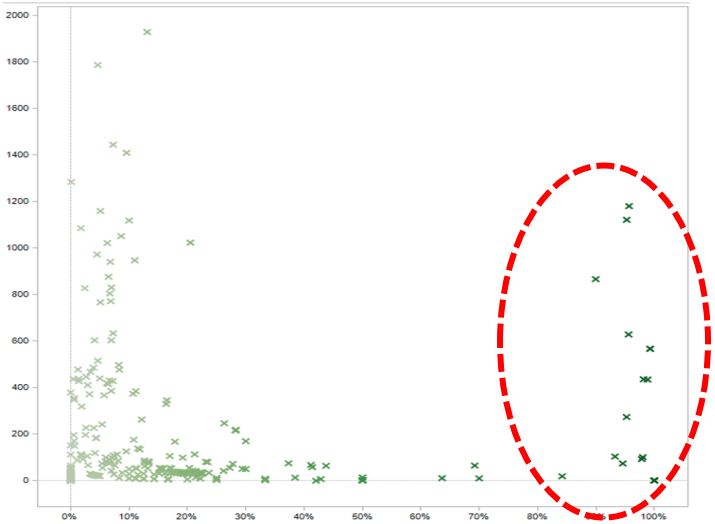


Forensic data analytics project: ORBIT
Data visualisation identifies competition cheats

Online competition run by a large company in Malaysia, using Facebook to vote



Number of votes vs number of repeated first four characters of an email address



Number of votes vs %age with 'later' Facebook profiles

Suspected cheating by participants… how to investigate voting data?

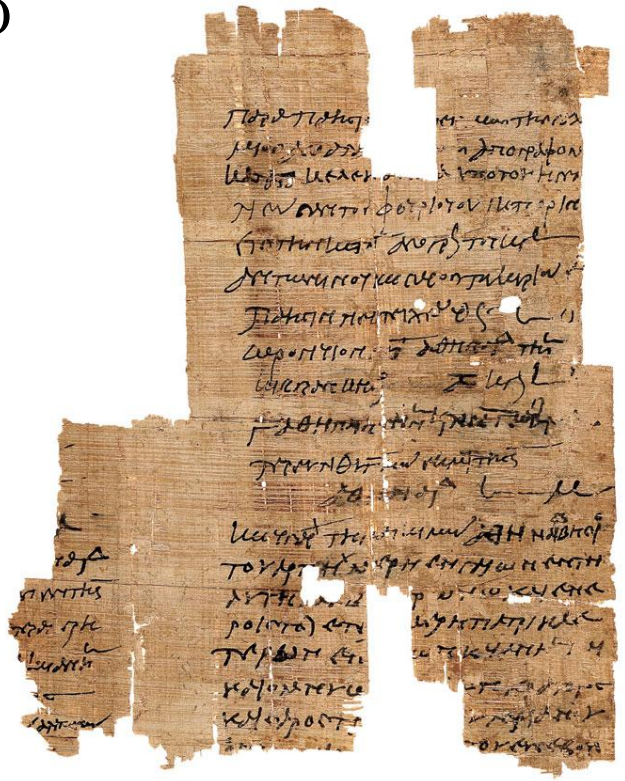
Forensic data analytics project: HYBRID
A fraud investigation on employee payroll

- Analysis of a Malaysian company's electronic payroll records
- > 5,000 employees
- Too many to review manually
- Ten data sources loaded into our database

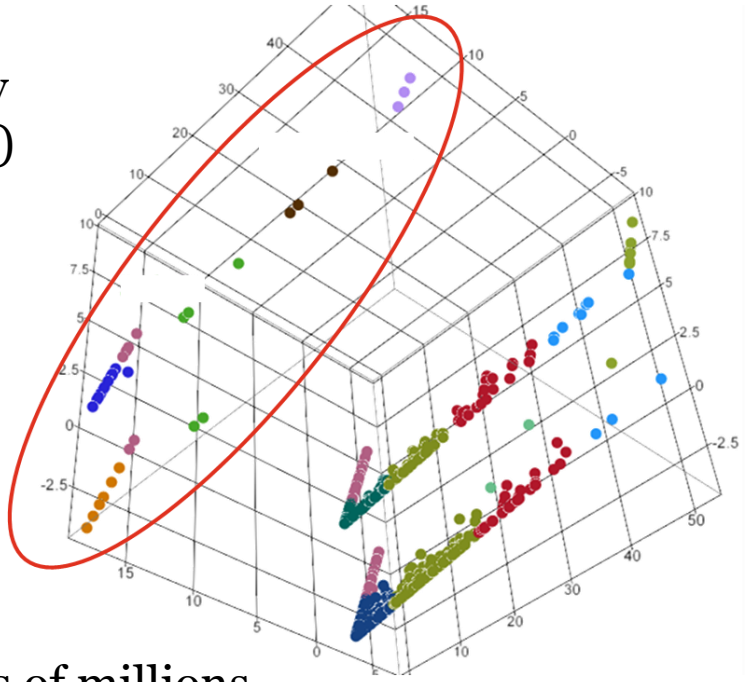
- Seventeen red-flag tests to identify clear failures against our expectations (e.g. payment > 1 month after employees' end date)
- Thirteen calculated metrics to allow us to identify employees that go outside defined thresholds (e.g. overtime pay as % of total remuneration)

Forensic data analytics project: CODEX
A fraud investigation on invoices and payments

- Analysis of a large Malaysian company's PO listing, journal invoices and vendor data
- A lot of data:
  - PO Listings Data: *1,147,713* *rows*
  - Journals Data: *57,190,537* *rows*
  - Vendor Master: *71,155* *rows*
- Not efficient to load into standard audit analytics software
- We used the data servers in our Forensic Technology Lab to load, query and visualise the procurement patterns over time and run red-flag tests



Forensic data analytics project: DIGITAL Using segmentation to detect procurement fraud



- A UK TV production and broadcast company uncovered a false invoicing fraud (by chance)
- They suspected other instances of false invoicing fraud over a period of two years
- For the time period in question, procurements totalled approx. 200,000 transactions and 9,500 vendors
- These transactions exhibited a huge range of PO values: from a few pounds to hundreds of millions
- We were not informed of which transactions had been identified as fraudulent
- Traditional red-flag fraud detection techniques have limitations in some cases (and wouldn't have worked for this project!)

Which key forensic technology capabilities could also be used to pro-actively detect fraud?

Which key forensic data analytics capabilities could be used to pro-actively detect fraud?

- **Data visualisation**
- **Suspicious transactions**
  - Red flag breaches
  - Outlier transactions
  - Behavioural anomalies based on data mining techniques
- **Suspicious trends**
  - Sudden increase in payment totals and volumes, especially to high risk entities / countries / new or one-time vendors / employees
  - A suspicious transaction distribution pattern
- **Large data loading and querying**
  - These tests are often only possible because we are able to efficiently handle huge volumes of data

An example of one of the tests… using transaction distributions to detect unusual payments

- Imagine that we have a list of all payments to suppliers over the previous two years
- Are any of these payments fictional (i.e. invented for the purposes of fraudulent payment)?
- One way of viewing the data is to take the first digit from each of the payment amounts, e.g. for RM23,000, we take the digit "2"

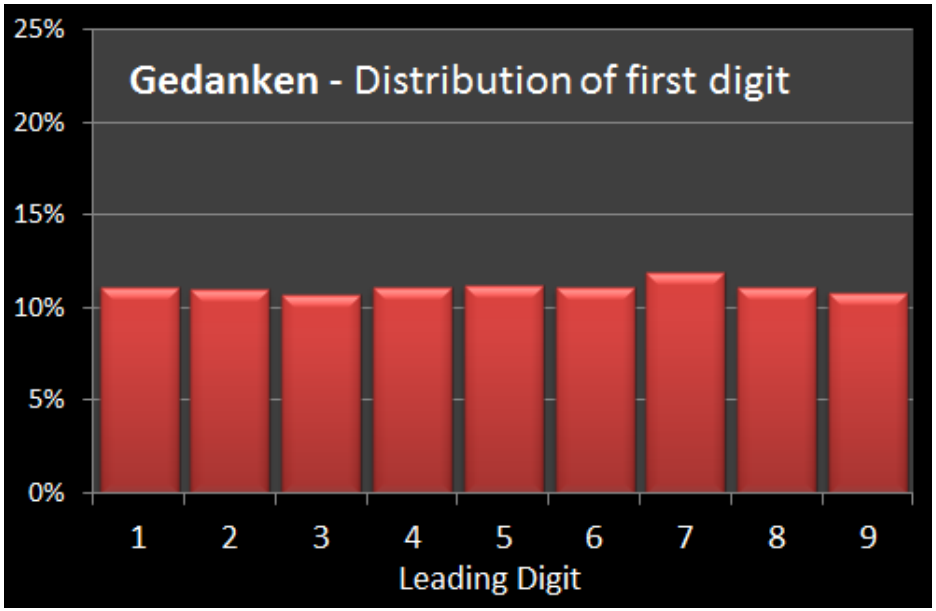


Chart from Data Genetics @ http://www.datagenetics.com/blog/march52012/

- How many of each of these "first digits" would we expect to be in our data?

But this is what really happens…

- The number 1 occurs significantly more often than the number 2 which, in turn, occurs more frequently than the number 3 … all the way down to the number 9
- This is called Benford's Law
- It is useful because it allows us to see where numerical amounts do not follow an expected pattern

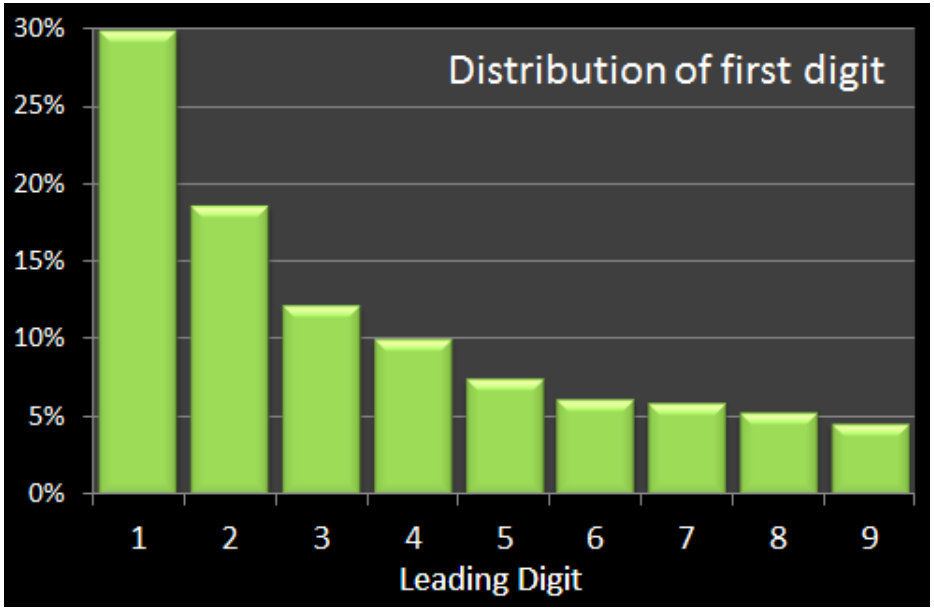


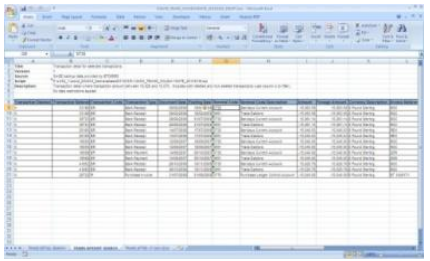
Chart from Data Genetics @ http://www.datagenetics.com/blog/march52012/

- Most people who make corrupt payments will not be aware of Benford's Law – so when they try to create seemingly random numbers, they try to match the distribution pattern most people assume (the chart on the previous slide!)

Wrap up

Recap

- Organisations are capturing and analysing ever larger amounts of data
- In recent years, we have been increasingly using data analytics tools and techniques to support our forensic investigations
- Forensic investigations are reactive, but there is also an opportunity to use big data analytics pro-actively in audits to uncover fraud
- Big data analytics provides us with the tools to:
  - Load large amounts of data
  - Visualise data rapidly
  - Identify suspicious transactions and trends
  - Detect and prevent fraud!



Thank you

Neil Meikle

Associate Director

Forensic Technology

Tel. +60 3 2173 0488

E: neil.meikle@my.pwc.com