BUTLER
ANALYTICS

BUSINESS ANALYTICS YEARBOOK

2015

Version 2 – December 2014

Contents

* Version 2 additions are Business Process Management and Open Source BPMS

This Year Book is updated every month, and is freely available until November 2015.

Production of the sections dealing with Text Analytics and Prescriptive Analytics was supported by FICO.

Introduction

This yearbook is a summary of the research published on the Butler Analytics web site during 2014. The content is primarily of three types; technology lists and reviews, discussion of issues, and explanatory material.

The lists and reviews all have links to the supplier web sites and to the full reviews on the Butler Analytics web site.

Discussion of issues addresses controversial subjects such as the risks associated with predictive analytics, the problems associated with data visualization, and the feeding frenzy around big data.

Explanatory material provides a relatively easy introduction to issues such as text analytics, prescriptive analytics and other topics.

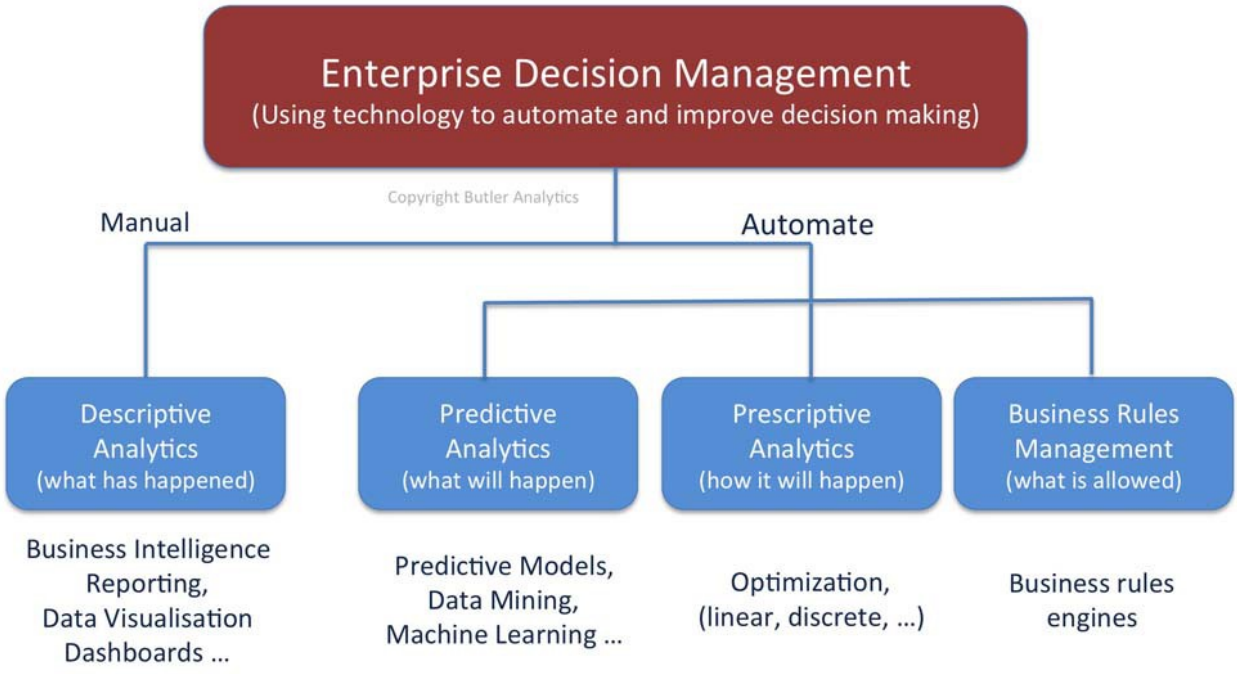
For 2015 the decision has been made to make this yearbook freely available – and it will be updated during 2015 regularly – so come back to download the latest version. Starting November 2015 a small fee will be introduced and this will give users of the yearbook access to monthly updates during 2015.

Finally it needs to be said that any ratings are our opinion and not recommendations.

Overview

Business analytics technologies can conveniently be placed under the umbrella of Enterprise Decision Management (EDM). These systems exploit methods and technologies to improve the efficacy and efficiency of decision making through the organization. This is in contrast to traditional systems which have almost exclusively been concerned with process automation and improved efficiency through labour displacement. The technologies employed in EDM include predictive analytics, business rule management, optimization, business intelligence, and in fact any technology which reduces the uncertainty involved in decision making, and increases decision making efficiency.

The traditional business use of information technology can be seen as an extension of the filing cabinet and desktop calculator. Computers have been used to store and process ever larger amounts of data and automate the calculation of quantities used in a business environment. These systems are also fairly inflexible, requiring the hard-coding of business rules and the use of complex programming languages to implement them. This is a deterministic world with no room for nuance or subtlety. Process automation peaked in the first decade of this century with ERP, CRM and other systems aimed at integrating and formalizing the transactional activity in an organization (with the unfortunate side effect of making the organization less flexible). This is now a domain of diminishing returns, and certainly much inferior returns compared with EDM.



In contrast EDM uses the computer as an uncertainty reduction machine, employing statistics, machine learning, data mining, optimization and business rules engines to fine tune decisions and massively increase

the speed at which they are made. In fact the current surge of interest in business intelligence (BI) tools and techniques is a testament to the urgent need to have technology help in the decision making process, although BI is labour intensive and prone to misinterpretation. As always the leaders in the use of EDM can be found in financial services with decision systems employed in loan approval, the detection of fraud, customer targeting and so on. The 'digitization' of business processes, an era that has persisted for fifty years, is now being complemented by the 'digitization' of decisions, and this new use for information technology will dwarf what has gone before it.

Any technology capable of reducing decision uncertainty, and reducing decision latency qualifies as an EDM enabler. Predictive analytics technologies scour historical data, looking for patterns that might be reliable enough to employ in future activities. Typical applications include reduction of customer churn, better sales targeting and other applications such as prediction of machine failure, or even the likelihood of hospital readmission. Technology startups are providing SaaS types services where business users, with little technical skill, can upload data and create their own predictive models. There are dangers associated with a 'black box' approach of this type, but it does at least indicate the way things will go. Larger organizations can afford to employ a team of data scientists and analysts to create bespoke predictive models and methods.

Optimization is another technology usually bundled in with EDM. This is primarily concerned with determining how resources should be deployed once the question of what will happen in the future is determined (the province of predictive analytics and statistics). Given a set of resources and constraints, optimization will work to maximize a given objective – usually profit and/or revenue. It answers the question 'how', given we know 'what'.

Finally the use of business rules engines complements both predictive analytics and optimization by saying what is, and is not permissible. A predictive model may suggest selling a given item at a certain price for example. However if the product has already been offered at a lower price to a subset of customers, it simply cannot be used in their cases. And optimization may suggest working practices that are unpopular or even illegal.

EDM is a sea-change in the way businesses use information technology, and the benefits that might be derived from it. Its effective use will distinguish the winners from the losers in a way we haven't seen before. Needless to say this all requires awareness at the very top of the organization, and there are profound organizational and cultural implications. We will after all be increasingly handing over the decision making process to machines – so we really do need to know what we are doing. Greater reward always implies greater risk, and EDM is no different. The risk mitigator is skill and knowledge – in a world of change some things never change.

Business Intelligence

Business Intelligence Evolves

This was the year of BI democratization. Business users are demanding direct access to their data and the tools to manipulate it. This is reflected in the rise of suppliers such as Qlik, Tableau, Yellowfin and Sisense. And the large established suppliers such as Microstrategy, Information Builders, SAS and Oracle are also making moves in this direction.

Cloud BI has also established itself in 2014, offering very capable, enterprise wide BI capability as a services. More notable suppliers include Birst, GoodData and Bime. These too focus on providing business users with direct acces to their data.

In reality BI is just a thin end of a very thick wedge. It provides businesses with information on what has happened or is happening, but is labor intensive and prone to error of interpretation. We await smarter BI tools and the automation of many BI tasks.

Dumb Technology

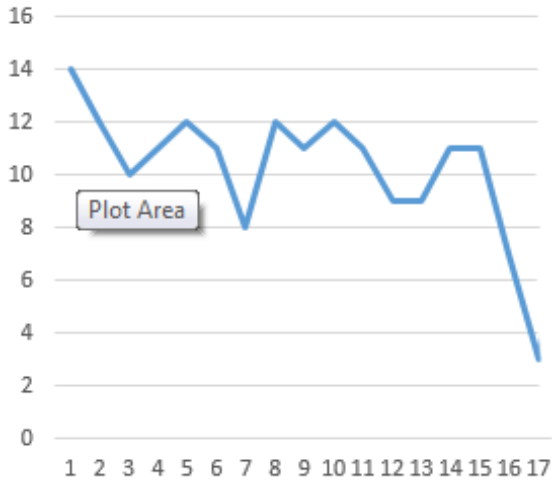
The data visualization tools being used by millions of business users around the world are essentially dumb. What this means is that the graphs, charts, gauges, dashboards and so on being produced in high volumes, within all types of businesses, come with no inbuilt intelligence to indicate what a visualization is actually saying. A simple example will clarify. Consider a chart with several points on it, all of which show a higher value than the previous one (apart from the first of course). Many people would identify this as a rising trend. But a random sequence of points will show such 'rising trends' quite regularly - and they mean absolutely nothing.

Dumb visualization technologies support the representation of any data a person may wish to analyse. Knowing when such visualizations mean something and when they do not is a tricky business. As far as I am aware all the data visualization tools available today are essentially dumb. Some may complain that users of this technology should know enough about statistics to work out the significance of the patterns that show up in visualizations. But this really misses the point. We don't expect users to be able to program in Java, or whatever language is used to build a visualization tool, and neither should we expect them to be statisticians.

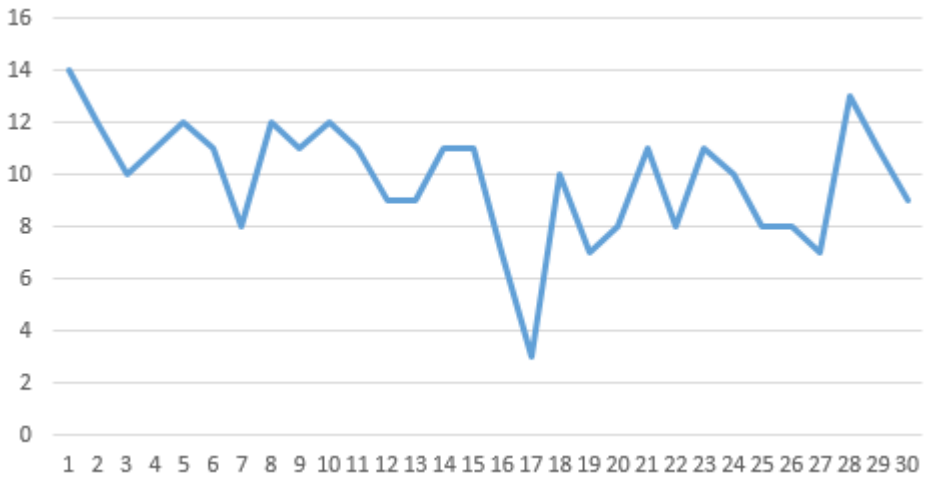
A smart data visualization tool will advise the user on the quality of the features a visualization is showing, and not simply display them without any guidance. This is desperately needed, since it will absolutely be the case that businesses around the globe will be basing decisions on spurious patterns (trends, comparisons etc) found by visualizing data.

Fooled By Data Visualization.

Please take a look at the first graph below. It could represent anything - sales by week number, a stock price, the temperature outside or a declining market share percentage. If a graph like this appears in your business then remedial action would almost certainly be taken. But here is the rub. This graph was produced in Excel by simulating 20 fair coin tosses and counting how many heads appeared in that 20 tosses. This process was then repeated 30 times and each point on the graph represents the number of heads from one set of 20 tosses. It is random.



Imagine you have 20 customers who on average each place an order once every two weeks. This graph could quite easily represent the number of orders taken each week for the first 17 weeks of the year. Alarm bells would be ringing, heads would roll and a hero would be born. Why a hero? Because things always tend to revert to the mean and it is fairly unlikely that three heads out of twenty coin tosses will repeat itself. And so our new Sales Director hero will be the unwitting recipient of a return to the mean - give her or him a pay rise.



The diagram above shows how this random sequence (which should average 10) progresses. Notice how the heroic efforts of the new head of sales immediately recovers the situation - not.

Blindness to the random is one of the greatest dangers when using data visualization tools. We will find trends when none really exist, and we will find other patterns (cyclical behaviour for example) when none really exists. We are hard wired to find patterns, and until recently it has served us quite well. But now we have to deal with probabilities, because in an uncertain world this is the only tool we really have. Our brains however have a hard time dealing with the uncertain. If you want more on this read Thinking Fast and Slow by Daniel Kahneman - a Nobel prize winning economist.

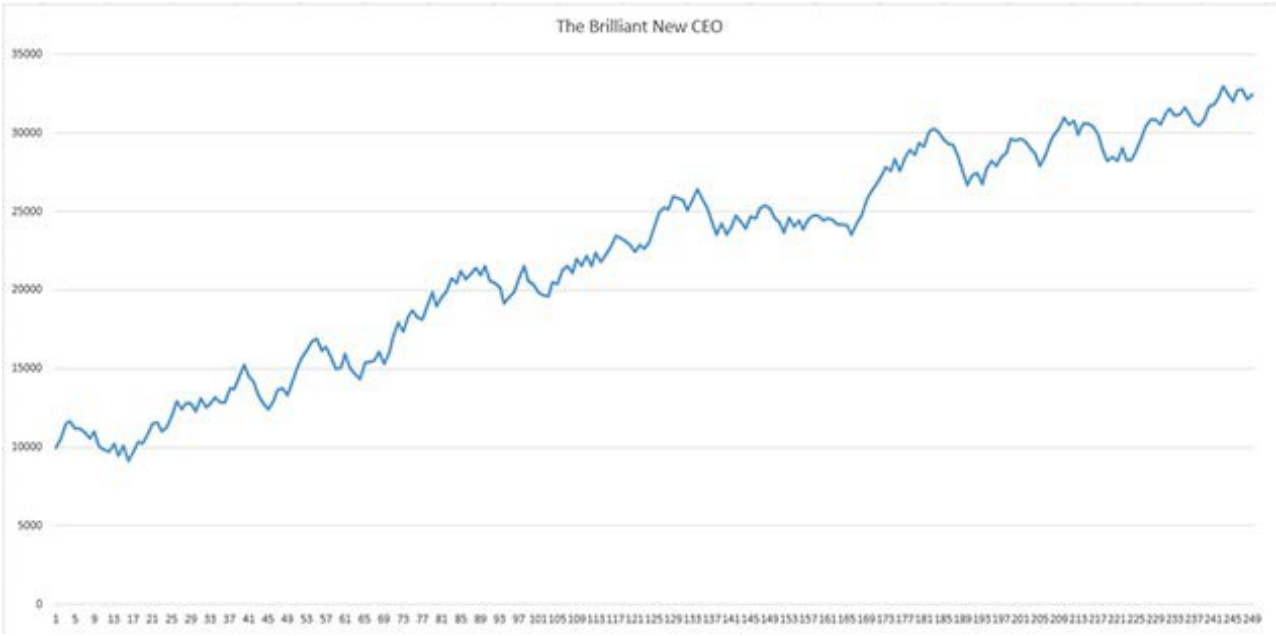
So let's be categorical. Many of the patterns you will find using the latest generation of eye-candy rich visualization tools will be nothing more than random accidents with no meaning whatsoever. It is highly likely that your business will suffer because of this willingness to see patterns where none exist. In fact the whole thing is becoming slightly silly with phrases such as 'the two second advantage'. Implying that data visualization will speed up recognition of patterns to such an extent that within two second you will be able to make a decision. I hope your résumé is up to date.

The solution to all of this is nothing to do with visualization, but depends on a modest understanding of randomness and probability and above all understanding your business - but of course no one is selling an expensive, career enhancing software package to do these unappealing things.

There is an excellent example of big decisions being made in business that are based on nothing more than random variation in the book The Drunkards Walk, by Leaonard Mlodinow. A Hollywood film boss was sacked because of the diminishing success of the movies the company produced. The thing with movies is that they take years to get funded and produce, and so the pipeline this sacked boss had in place continued for several years after he was gone. It was hugely successful and he was sacked for nothing other than random variation. And everyone has heard of the experiment with monkeys throwing darts at the financial pages to generate an investment portfolio. It outperformed most of the professionally managed funds.

There is a lot of ego at stake here, and so data visualization can certainly be used to glorify a rising trend and cast into hell those responsible for a falling trend - which of course is exactly how it will be used.

Just to reinforce the tricks randomness can play, here is another little story. An ailing corporation decided to set on a new CEO. Sales had dropped from 20,000 units a week to 10,000. The new CEO did what all new CEOs do and blamed everything on his predecessor. The success of this new CEO can be seen in the graph below. The first 15 week didn't look so good, but the old CEO could be blamed fairly and squarely for this.



Then things took off. Of course there were dips, but the trend was definitely upward. This CEO got his analytics team to produce any number of graphs - linear regressions, bubble charts, box plots and so on. After five years his 'strategy' was seen as an enormous success. He was seen on the front page of glossy business magazines, wrote a book called 'The Seven Secrets of Samurai Leadership' and saw the stock price of his business more than quadruple - which did wonders for his bonus. Lady Fortuna was very generous, because this too is a randomly generated chart. Start with 10,000 and cumulatively add a random number between -1000 and +1000 every week to simulate an increase or decrease in sales. And of course for every graph such as the one above there is another pointing the opposite way. If you wanted to see all the graphs together you would run a Monte Carlo simulation. Where is this graph going next? No one knows - but if I were that CEO I'd get out while the going was good and my book sales high.

Not all business success is down to luck, but inevitably much of it is, just as business failure is often down to bad luck. I'll be writing more articles on randomness in business, although I doubt they will be particularly popular.

Enterprise BI Platforms Compared

The enterprise BI platforms compared in this article represent the most common alternatives many organizations will consider. The analysis is high level, and not a feature by feature comparison - which is fairly pointless in our opinion. The five criteria used to compare the products are:

- Capability - the breadth of the offering.
- Performance - the resources needed to make a platform perform adequately. The greater the resources the lower the score.
- Extensibility - very important and a measure of how well a platform can be extended functionally and how well it scales.
- Productivity - the support a platform offers for productive work.
- Value - likely benefits versus costs.

This form of analysis creates some surprises, but you need to look at the full review to see why a particular offering does so well.

Important: If you want to see the full review click on the score. If you want to visit the vendor website click on the name.

4.3 InetSoft is one of the most understated BI suppliers. The capability of its BI offerings is quite unique (hence the high score) and the maturity of the products and company show in the technology they offer (InetSoft has been around since 1996). Licensing structures are flexible and will appeal to mid size businesses as well as large corporations.

4.3 - Information Builders is a long established supplier of BI, analytics and integration technologies. The integration of BI and data mining is quite unique and puts IB ahead of the crowd. The maturity, sophistication and value for money is very hard to beat.

4.3 - The QlikView BI platform has the ability to be all things to all people, and will satisfy business users, developers and enterprise needs. It sets the right balance between ease-of-use and sophistication in our opinion - something that very few other BI platforms achieve. We particularly like its extensibility and the ability to create new chart types and BI apps - there should be no running up against the buffers.

4.2 - BOARD provides a combined business intelligence and corporate performance management platform. The BI capability embraces reporting, multi-dimensional analysis, ad-hoc querying and dashboards. This is combined with the CPM functions of management and monitoring of performance planning, including budgeting, planning, forecasting, profitability analysis, scorecards and financial consolidation.

4.2 - Jedox is a BI workhorse and has a reputation for delivering good performance and ease-of-use. It's unpretentious profile belies the fact that many organizations have satisfied complex BI needs by using this technology.

4.2 - Yellowfin BI provides a powerful BI environment, which is also easy to use. This is a rare combination and probably accounts for the various accolades it has acquired in recent years. It provides dashboards, reporting, data discovery and visualization tools, storyboarding and collaborative BI as well as providing good mobile BI support.

4.1 JReport is a disarmingly straightforward BI environment capable of addressing the needs of large and medium size organizations. It is well regarded for the flexibility of the architecture and its support for the embedding of BI components into mainstream applications.

4.1 - Spotfire from TIBCO occupies a fairly unique position in the BI landscape, addressing both BI and predictive analytics needs. This makes the platform very powerful in our opinion since it becomes possible to display predictive models in a graphical environment. The recent acquisition of JasperSoft shows that TIBCO feels it needs to pitch at the less sophisticated end of the market too. Spotfire should be considered as a platform for a broad range of analytics actvities, and not simply a BI visualization platform.

4.0 - Birst offers a complete BI ecosystem, delivered as SaaS and if necessary as a software appliance within the organisation. Great emphasis is placed on various data storage mechanisms (data warehouse, ROLAP) and sophisticated ETL capability. The emphasis seems to be on getting stuff into their ecosystem rather than getting it out, and so there would appear at least to be some level of lock-in. Of course this is true of many suppliers (lock-in is a software supplier's dream), but it seems particularly emphasised with birst.

4.0 - IBM Cognos is a large, sprawling suite of BI tools and infrastructure best suited to the needs of large organizations, and most likely those with an existing IBM commitment. For senior executives with a 'Master of the Universe' complex Cognos will support every possible type of analysis. However for people who just want to know what is happening or has happened there are lighter and more nimble alternatives.

4.0 - Jaspersoft, now a TIBCO company, provides a BI platform based on open source software. This is a solid, widely used product set, that delivers a no frills BI environment and will meet the needs of most organizations with ease. All the usual functionality is included – reporting, dashboards, analysis and data integration.

4.0 - Microsoft, in typical style, has done BI its way. Much functionality is shoehorned into Excel and SQL Server and SharePoint is used as a distribution mechanism. But the ubiquity of Microsoft products means that many organizations will go this way, and with the necessary skills they will be able to deliver what they need, and at a modest cost (usually anyway).

4.0 - Pentaho often delivers everything an organization might need, and goes beyond the classical role of BI by also offering predictive analytics capability. This unglamorous technology stack with an open source heritage is designed to get the job done - and with the least possible fuss.

3.9 - Actuate provides a capable and extensible analytics platform capable of addressing most needs. It’s mainly based on open source technology and if you use the free open source BIRT project then it will cost you nothing, although this only covers data visualization and reporting. The full blown platform from Actuate addresses much more including dashboards, predictive analytics and self-service BI.

3.9 - arcplan is a 'no frills' BI platform, well regarded for just getting the job done, and providing planning, budgeting and forecasting functionality in addition to the more usual BI roles.

3.9 - MicroStrategy is in many ways a meeting of the old and new in business intelligence, and takes the positives from both. It is truly an enterprise solution meeting the less glamorous demands for regular reporting, complex dashboards and extensive admin, while offering up the sexier self-service BI users now expect from a BI solution. It is expensive, and for organizations with less demanding requirements other options will be more economical.

3.9 - SAS BI will attract larger organizations with complex needs. Even here however there may be resistance to the high price of SAS products and a need to find simpler solutions for some requirements.

3.8 - Oracle provides an extensive set of BI tools and specialized data management platforms. This will mostly be of interest to existing Oracle customers, being fairly expensive and highly proprietary. But for those wedded to the mantra that the Oracle way is the only way, the functionality and architectures offered by Oracle can do anything a business might require.

3.7 - SAP BI comes primarily in the form of BusinessObjects, a long established BI platform that is not particularly user friendly and can be fairly expensive. Users of SAP applications may choose to go this way, but there are no compelling reasons for anyone else to do so.

3.7 - SiSense provides easy-to-use business intelligence tools, and is targeted at the business user. The speed of query execution distinguishes the product and novel techniques have been used to deliver high levels of performance. Don't expect SiSense to satisfy all BI needs, it's ease-of-use is delivered at the price of reduced sophistication.

3.7 - Tableau is often the first supplier that comes to mind when businesses consider data visualization tools. While the product is easy to use, and produces very attractive visuals, it is not particularly sophisticated and may prove inadequate as needs mature. This has recently been addressed to some extent through an interface to R, but this is a classic choice between ease-of-use and sophistication - and Tableau has done very well by focusing on the former of these two.

Enterprise Reporting Platforms

BI Suites

IBM Cognos is a sprawling, heavy duty BI, analytics and reporting environment that is mainly of interest to larger organizations.

Jaspersoft supports data visualization, dashboards, ad-hoc reporting and analytics in a variety of editions to suit business size and needs.

Logi Analytics provides reporting, BI, mobile, dashboards, embedded applications, data discovery and visualization in a web browser environment.

MicroStrategy provides BI and reporting solutions for organizations of all sizes and even offers some completely free versions for smaller organizations.

Pentaho provides a broad suite of business intelligence and analytics tools including dashboards, reporting, data visualization, big data support, data integration, mobile BI and data mining.

QlikView provides a rich data visualization, reporting and discovery platform, with clever technology to make sure you get to the data that is relevant.

SpagoBI provides a broad suite of business intelligence tools based on open source components. These include reporting, OLAP support, charting, data mining, dashboards and consoles, mobile support and many other functions essential to broad BI implementations.

Tableau majors on data visualization with a large number of charts and graphs. Desktop and server (for web interface) versions available, and a free version for creating simple graphics.

Tibco Spotfire is a platform for enterprise analytics, reporting, data discovery and visualization, and BI. Distinguishes itself by the level of support for advanced analytics (data mining, statistics etc).

Yellowfin supports dashboards, data discovery, data visualization, mobile BI, collaborative BI and mapping.

Reporting

DBxtra provides ad-hoc reporting software with report designer, dashboard designer, free report viewer, report web services (secure, centralized repository for web based access to reports), schedule server for automating reporting tasks.

Eclipse BIRT is a very flexible and widely used reporting environment, but you will need to buy the commercial product ActuateOne to get bells and whistles (Dashboards, ad-hoc reporting etc).

i-net designer is a free graphical report designer which runs on Java. i-net Clear Reports server generates reports in a Java viewer applet. i-net JDBC drivers are available as a separate product.

JReport from Jinfonet supports dashboards, reporting, mobile reports and embedding reports within applications.

Navicat report builder contains three workspaces - data, design and preview. The data workspace supports the selection and manipulation of data and two wizards simplify the process. The design workspace is where the report layout is created and preview shows how the report will look when printed. Most databases are supported.

NextReports provides three, free open source utilities that support report creation using a variety of database platforms. NextReports Designer is an application to design in-grid reports, using connections to most popular databases. NextReports Engine is a lightweight (330 k) Java platform development library which can be used to run NextReports inside applications. NextReports Server supports the scheduling of NextReports and Jasper reports and to automatically deliver reports.

Open Source BI Platforms

All the open source BI suites detailed below offer ample reporting capabilities. Some, notably Pentaho Community and SpagoBI go well beyond this. Palo on the other hand provides an extremely effective Excel interface and good OLAP processing abilities.

Eclipse BIRT

Very flexible and widely used reporting environment, but you will need to buy the commercial product ActuateOne to get bells and whistles (Dashboards, ad-hoc reporting etc).

Actuate claim that Eclipse BIRT is the most widely used open source BI suite. As the name suggests it runs on the Eclipse IDE platform and provides various reporting and data visualization tools. This provides the foundation for report design and viewing.

The BIRT engine is at the heart of this offering, and this is a collection of Java classes and APIs which execute BIRT reports and generate them in appropriate formats. Reports include lists with grouping and calculations, charts (pie charts, line and bar charts etc) which can be rendered in SVG and support events to be interactive, crosstabs, documents (textual data) with embedded lists, charts etc, and compound reports (aggregated and embedded reports).

Actuate sell a version of BIRT on steroids called ActuateOne which includes BIRT Design (for report design), BIRT iHub and BIRT User Experience (Ad Hoc Reporting, viewers, Dahsboards, mobile support and Data Analyzer).

JasperSoft

JasperSoft provides several versions of its JasperSoft BI suite. The Community edition essentially provides a reporting and charting environment with supporting infrastructure. The report designer supports charts, images, crosstabs and sub-reports for sophisticated report layouts. Interactive report viewing is a browser based report viewer with sorting, filtering and formatting of report snapshot views. A centralised repository provides infrastructure for reporting and stores user profiles, reports, dashboards and analytic views.

Reports can be automatically scheduled and distributed using Report Scheduling and User Access and

Security provides report access by role or user. Mobile BI is also supported for iPhone and Android devices.

The commercial editions add much more including dashboards, a metadata layer, in-memory analysis, data integration capability and interactive visualizations. The list of JasperSoft customers is impressive, although most will undoubtedly be using the more capable commercial version.

Palo

Palo is an open source business intelligence suite focused around OLAP and Excel and web interfaces. The Palo OLAP Server is at the heart of the offering and provides multi-user, high performance access to data. It supports real-time aggregation and is an ideal platform for BI collaboration. Data is loaded into the OLAP server using the Palo ETL Server. This supports most data sources including relational databases, SAP and others.

Palo Web provides one of the user interfaces, both for designers and business users. Designers can administrate the system and create web-based reports, while users are able to view reports.

Palo for Excel is where business users will spend most of their time, and its central logic store avoids the bottlenecks that are often encountered using Excel for complex tasks. Because Palo is cell based (as opposed to record based) it is ideally suited to the Excel (or OpenOffice) interface.

The commercial version of Palo is supplied by Jedox in two versions. Jedox Base (effectively Palo) is free, while the premium edition offers considerably more functionality.

Pentaho Community

This is a very capable suite of BI, reporting, and data mining tools with sophisticated functionality, and will address the needs of many organisations.

Pentaho BI Suite Community Edition (CE) includes ETL, OLAP, metadata, data mining, reporting and dashboards. This is a very broad capability and forms the basis for the commercial offering provided by Pentaho. A variety of open source solutions are brought together to deliver the functionality including Weka for data mining, Kettle for data integration, Mondrian for OLAP and several others to address reporting, BI, dashboards, OLAP analytics and big data.

The Pentaho BI platform provides the environment for building BI solutions and includes authentication, a rules engine and web services. It includes a solution engine that facilitates the integration of reporting, analysis, dashboards and data mining. Pentaho BI server supports web based report management, application integration and workflow.

The Pentaho Report Designer, Reporting Engine and Reporting SDK support the creation of relational and analytical reports with many output formats and data sources.

If you want a version with support, training and consulting, as well as a few more bells and whistles then Pentaho provide such services and product.

ReportServer

This provides an extremely flexible open source reporting and dashboard environment. It supports Eclipse Birt, JasperReports and SAP Crystal Reports reporting engines in addition to its own ad-hoc oriented reporting tools. The user interface is web based and it supports a wide range of admin tools.

Central to ReportServer is the Dynamic List. This is the preferred method supporting a wide range of functions such as column selection, filtering, sorting, grouping, sub-totals, calculation and so on. JasperReports and Eclipse Birt tend to be used for 'pixel perfect' reporting with output to a pdf file. Finally Script Reports are used for particularly complex reports, and require programming skills to use. Interactive dashboards are supported and are generally constructed for items called dadgets (Dashboard Gadgets) - these can be anything from a report to an interactive HTML5 app.

Currently supported data sources include Oracle, Microsoft SQL Server, IBM Informix, IBM DB2, mySQL, PostgreSQL, h2 and of course csv files.

As with all open source BI bundles users can opt to get training, consulting and support from ReportServer if they wish.

SpagoBI

SpagoBI is essentially a very large collection of open source software brought together to create a broad business intelligence capability. In fact it goes beyond the traditional notion of BI to embrace domains such as data mining and BPM. This broad capability has encouraged large companies such as Fiat and Gamebay to adopt it as part of their strategic BI solution. SpagoBI is also unique among open source BI solutions in that the free version is the only version. There are no paid for versions with extra functionality. Users can elect to contract into support and training, but the product just comes in one version.

The breadth of the offering is impressive, and each area of functionality is often served by a number of engines to deliver all the functionality that might possibly be needed. The main areas of functionality include:

Reporting

- Multidimensional analysis
- Charts
- KPI
- Interactive Cockpits
- Ad-Hoc Reporting
- Location Intelligence
- Free Inquiry (Query by Example)

- Data Mining
- Real Time Dashboards and Console
- Collaboration
- Office Automation
- ETL
- Mobile
- Master Data Management

Users of SpagoBI will need support, training and consulting services, and obviously this is where the revenue model is pitched.

Free Dashboard Platforms

These free (not trials) dashboard platforms serve various functions. Microstrategy, Qlik, SpagoBI, SAP and InetSoft are fully blown analytics tools with sophisticated dashboard capability. Other such as Bittle, ClicData, DashZen and so on, are cloud based.

Bittle supports the creation of online dashboards specifically for the SME user. The free package restricts data storage and data sources, and comes with a standard graphics library (instead of a more advanced one). Bittle also supports report creation.

ClicData is a drag and drop dashboard tool where dashboard elements are dropped into the dashboard. The free version supports Excel, CSV, Dropbox, Google Drive, SkyDrive, FTP, and data refresh once a week. Collaboration features are supported, and schedules for automatic sharing. Refresh and sharing can both be made automatic. The dashboards & visualization widgets can be integrated within any html page and stay interactive. A formula editor supports alerts, trend lines and functions.

Dash supports real-time dashboards for websites, business and personal use. The free version supports a single dashboard.

Dashzen supports the creation of both private and public dashboards in the cloud. Private dashboards can be shared with nominated people. A variety of gadgets make up a dashboard, some of which are connected with various data sources (salesforce, Twitter, StackExchange etc).

InetSoft provide a free version of their excellent Style Scope platform. Style Scope Free Edition is a small-footprint server that delivers Web-based interactive Flash dashboards and visualizations that can be shared within an organization. The Java-based application can be installed on any Windows, Unix, or Mac desktop and can be connected to data in standard relational databases as well as spreadsheets.

SAP Lumira lets you understand your data by building visualizations using a drag and drop interface.

Combine and analyze data from Excel and other enterprise sources and quickly discover unique insight – no scripts, predefined queries or reports required.

Microstrategy Analytics Desktop is a sophisticated business analytics platform that is entirely free. Dashboards form a large part of the functionality and for users who want more analytics, an integration with R supports predictive analytics. Most data sources supported.

Netvibes is more geared to social media analytics, but can also be used outside this domain. The free service provides the dashboard and reader.

Qlik Sense is a next-generation, self-service data visualization and analysis application that empowers business users to easily create personalized visualizations, reports and dashboards with drag-and-drop simplicity.

SpagoBI provides dashboard capability as part of a much larger open source BI suite. SpagoBI offers a specific engine allowing the development of real-time monitoring consoles, to be used in Business, operational and BAM (Business Activity Monitoring) processes.

Zoho Reports is an online reporting and business intelligence service that helps you easily analyze your business data, and create insightful reports & dashboards for informed decision-making. It allows you to create and share powerful reports. The free service limits users (2) and data, and has cut down functionality.

Free MySQL Dashboard Platforms

All the free dashboard software listed below support MySQL - and most support may other data sources too.

Bittle supports the creation of online dashboards specifically for the SME user. The free package restricts data storage and data sources, and comes with a standard graphics library (instead of a more advanced one). Bittle also supports report creation.

ClicData is a drag and drop dashboard tool where dashboard elements are dropped into the dashboard. Collaboration features are supported, and schedules for automatic sharing. Refresh and sharing can both be made automatic. The dashboards & visualization widgets can be integrated within any html page and stay interactive. A formula editor supports alerts, trend lines and functions.

InetSoft provide a free version of their excellent Style Scope platform. Style Scope Free Edition is a small-footprint server that delivers Web-based interactive Flash dashboards and visualizations that can be shared within an organization. The Java-based application can be installed on any Windows, Unix, or Mac desktop and can be connected to data in standard relational databases as well as spreadsheets.

Microstrategy Analytics Desktop is a sophisticated business analytics platform that is entirely free. Dashboards form a large part of the functionality and for users who want more analytics, an integration with R supports predictive analytics. Most data sources supported.

Qlik Sense is a next-generation, self-service data visualization and analysis application that empowers business users to easily create personalized visualizations, reports and dashboards with drag-and-drop simplicity.

Free Dashboards for Excel Data

Bittle supports the creation of online dashboards specifically for the SME user. The free package restricts data storage and data sources, and comes with a standard graphics library (instead of a more advanced one). Bittle also supports report creation.

ClicData is a drag and drop dashboard tool where dashboard elements are dropped into the dashboard. The free version supports Excel, CSV, Dropbox, Google Drive, SkyDrive, FTP, and data refresh once a week. Collaboration features are supported, and schedules for automatic sharing. Refresh and sharing can both be made automatic. The dashboards & visualization widgets can be integrated within any html page and stay interactive. A formula editor supports alerts, trend lines and functions.

InetSoft provide a free version of their excellent Style Scope platform. Style Scope Free Edition is a small-footprint server that delivers Web-based interactive Flash dashboards and visualizations that can be shared within an organization. The Java-based application can be installed on any Windows, Unix, or Mac desktop and can be connected to data in standard relational databases as well as spreadsheets.

Microstrategy Analytics Desktop is a sophisticated business analytics platform that is entirely free. Dashboards form a large part of the functionality and for users who want more analytics, an integration with R supports predictive analytics. Most data sources supported.

Qlik Sense is a next-generation, self-service data visualization and analysis application that empowers business users to easily create personalized visualizations, reports and dashboards with drag-and-drop simplicity.

Open Source and Free OLAP Tools

Jedox Base is the free version of Jedox and comes with OLAP server and Excel add-in. Apart from multi-dimensional queries, data can be written back and consolidated in real-time. The Jedox server keeps all data in the cache for fast data access. APIs in Java, PHP, C/C++, or .NET can be used to integrate the Jedox OLAP database in other software environments. The engine controls permission management with stringent and secure access rights for users, groups, and roles. Data is encrypted with MD5 by default and supports the HTTPS using SSL to encrypt communication between clients with valid security certificates. The analytics supports SSO with Active Directory and LDAP on Linux and Windows.

The Excel add-in is used for communication between the Jedox OLAP database and the Excel front-end. The data entered in Excel is written back and aggregated to the OLAP cube structure through the add-in.

JsHypercube is a light-weight OLAP database written in JavaScript. It is useful for any application involving the aggregation of metrics for purposes of dynamic charting. Datasets can be "sliced and diced" in real-time, with low-latency.

Kylin is an open source Distributed Analytics Engine from eBay Inc. that provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets. It currently offers integration capability with BI Tools like Tableau. Integration with Microstrategy and Excel is coming soon. Features include:

- Job Management and Monitoring
- Compression and Encoding Support
- Incremental Refresh of Cubes
- Leverage HBase Coprocessor for query latency
- Approximate Query Capability for distinct Count (HyperLogLog)
- Easy Web interface to manage, build, monitor and query cubes
- Security capability to set ACL at Cube/Project Level
- Support LDAP Integration

Mondrian is an OLAP (online analytical processing) engine written in Java. It reads from JDBC data sources, aggregates data in a memory cache, and implements the MDX language and the olap4j and XML/A APIs.

olap4j is a common API for any OLAP server, so you can write an analytic application on one server and easily switch it to another. Built on that API, there is a growing collection of tools and components.

- olap4j - Core API, Query Model, Transformation and other auxiliary packages, along with the driver specification.
- olap4j-xmla - Driver implementation of olap4j for XML/A data sources. It is compatible with Mondrian, Palo, SAP BW and SQL Server 2005+.
- olap4j-tck - Technology compatibility kit. Consists mostly of JUnit tests.
- olap4j-jdk14 - Olap4j distribution compatible with Java 1.4. Includes the core API and the XML/A driver.

phpmyolap - OLAP application in PHP for MySQL databases. No java-based web service needed. No MDX language to use.

12 Cloud Business Intelligence Platforms Compared

The benefits of cloud based BI are obvious. Minimum set up time, no up front investment, reduced operational costs and no upgrade disruptions. The cloud BI solutions listed below vary enormously in capability, but depending on need it may be that simpler solutions might be preferable. In any case the products in this cloud BI platforms review are rated for capability, extensibility, productivity and value - the value shown is the average. Where a full review has been published the rating can be clicked.

4.5 Spotfire Cloud is an extraordinarily powerful cloud based deployment of the Spotfire analytics platform. For organizations wishing to think beyond pure BI, Spotfire also embraces other forms of analytics such as predictive and prescriptive analytics via its interface to other analytics tools (R for example). Comes in three flavours - Personal, Workgroup and Enterprise.

4.4 Jaspersoft for AWS (Amazon Web Services) can do pretty well anything you might imagine. It is more properly described as a BI development environment, and potentially a very cheap one at that (users can pay by the hour). For organizations that want to build an extensive, tailored BI environment, Jaspersoft will oblige - and the result will be wholly based on various open standards. So there is very little lock-in.

4.3 InfoCaptor is an extremely competent product, capable of addressing many BI, data visualization and analytics needs at a very modest price. Deployment can either be in-house or on the web, and in either case the interface is web based. This a pragmatic, 'get-the-job-done' solution without the surface gloss and high prices charged by other suppliers.

4.3 It is hard to fault Microstrategy's cloud BI solution. The needs of most organizations will be more than adequately met by its range of business user tools, data integration capabilities, mobile support and support for advanced analytics (predictive analytics particularly). The only fly in the ointment is pricing - but you get what you pay for.

4.1 BIME provides a good all-round cloud based business intelligence environment. It provides excellent support for diverse data sources, good collaboration and document distribution facilities, and easy to use chart and dashboard creation tools.

4.1 Birst offers an industrial strength cloud BI solution that will be mainly of interest to large organizations. Its data management and integration tools provide a bedrock for true enterprise wide BI, and business users are served with a rich set of capabilities through a simple user interface.

4.1 GoodData provides a particularly powerful analytics platform which, although it can serve general BI needs, is particularly targeted at sales and marketing functions. A number of dashboard solutions are offered for sales, marketing, Yammer and service analytics. More sophisticated than the average offering.

4.0 Chartio is a cloud based business intelligence environment that will mainly satisfy the needs of business users. It does come with moderately sophisticated querying and calculation tools, but isn't a full blown BI environment. Nonetheless it provides and easy-to-use environment that many users will find productive.

4.0 Domo has a good reputation for attractive visuals and a fairly easy route to data access (applications, cloud data, Excel spreadsheets etc). It isn't intended to be a sophisticated BI platform, and serves the straightforward needs of people who want to see charts and dashboards derived from a variety of datasources. A good complement to more mature, capable BI platforms, but not a full solution in itself. Astonishing web site - all bling, no content.

4.0 Microsoft Power BI provides collaboration and publishing enhancements to the BI facilities built in to the Office suite. The heavy work is done by Excel with extensions such as PowerPivot (for fast in-memory analysis across large data sets) and Power View (for visual data analysis).

3.9 Oracle Cloud BI is an enterprise class solution to BI requirements, and will be of particular interest to existing Oracle users. The staging environment will be particularly welcome, particularly as it is missing in so many other products of this nature.

3.9 Tableau Online is a hosted version of Tableau Server. So users get all the ease of use and attractive visuals that Tableau is well known for. It's primarily a platform for business users to create charts and dashboards, and distribute them as appropriate. Connectivity to data sources is good, and Tableau makes a good complement to enterprise BI platforms.

Data Integration Platforms

The platforms listed here are broken down into data integration platforms that are part of a larger suite of products, platforms that are independent and those that are open source. Some products are listed in their open source community editions and the enhanced, supported editions.

Data Integration as Part of Larger Product Suite

Actian DataConnect is part of the Actian big data analytics platform and delivers sophisticated, highly scalable data integration, profiling and matching capabilities that excel in big data environments. With Actian DataConnect you can integrate, migrate, sync, validate, standardize or enrich your data while maintaining data quality in every facet of your business. This includes:

- Streamline data integration using a convenient browser-based UI and visual design process and drag & drop link-style mapper – with no need to maintain custom-coding
- Monitoring integration server and execution status with message-driven integrations
- Deploying anywhere – with on-premise, cloud or hybrid options Scaling to any volume of data and connect to any endpoint Taking advantage of end-to-end lifecycle management

IBM's data integration solutions enable you to understand, cleanse, monitor, transform and deliver data, as well as to collaborate to bridge the gap between business and IT. IBM provides capabilities for delivering data in real time to business applications, whether through bulk (extract, transform, load (ETL)), virtual (federated) or incremental (change data capture) data delivery.

Information Builders iWay Integration Suite is part of an extensive BI and analytics platform. The iWay Integration Suite allows for direct access to all of your data, so you can design your architecture to address the unique information needs of all your users. iWay accelerates the deployment and reduces the risk of all types of data integration projects – including extract, transform, and load (ETL); enterprise information integration (EII) initiatives; and web services deployments.

- End-to-end integration of a wide variety of sources, including cloud-based information, social systems, and big data
- Support for real-time and batch integration
- Flexible extract, transform, and load (ETL) and message-based styles of integration

Oracle Data Integration is part of the broader Oracle range of products and delivers pervasive and continuous access to timely and trusted data across heterogeneous systems. Its comprehensive capabilities include real-time and bulk data movement, transformation, bi-directional replication, metadata management, data services, and data quality for customer and product domains.

Pentaho data integration is part of the Pentaho BI suite of products and prepares and blends data to create a complete picture of your business that drives actionable insights. The complete data integration platform

delivers accurate, “analytics ready” data to end users from any source. With visual tools to eliminate coding and complexity, Pentaho puts big data and all data sources at the fingertips of business and IT users alike.

QlikView Expressor is metadata management “the QlikView way” — a disruptive approach to data management. It is simple and descriptive, not complex and prescriptive. Consistently capture and manage metadata as you build analytic apps, rather than be locked into a semantic layer up front.

SAS Data Integration provides a powerful visual design tool for building, implementing and managing data integration processes regardless of data sources, applications, or platforms. An easy-to-manage, multiple-user environment enables collaboration on large enterprise projects with repeatable processes that are easily shared. The creation and management of data and metadata are improved with extensive impact analysis of potential changes made across all data integration processes. It enables users to quickly build and edit data integration, to automatically capture and manage standardized metadata from any source, and to easily display, visualize, and understand enterprise metadata and your data integration processes, and is a component in a number of SAS software offerings, including SAS Data Management Advanced

Independent Data Integration Products

Adeptia Suite covers data integration, application integration, B2B integration and BPM. Adeptia ETL Suite is a graphical, easy to use software that supports ANY TO ANY conversion. It consists of three distinct components. It has a web-based “Design Studio” that provides wizard-driven, graphical ability to document data rules as they relate to validations, mapping and edits. This tool includes a library of functions which can be pre-created and reused again and again. Data Mapper has a "preview" capability to see actual source and target data, while the rules are being specified, if the source data file is available. The second component is the “Central Repository” where all the rules and mapping objects are saved. The third component is the “Run-time Execution Engine” where the mapping rules and data flow transactions are executed on incoming data files and messages.

Apatar provides connectivity to many popular applications and data sources (Oracle, MS SQL, MySQL, Sybase, DB2, MS Access, PostgreSQL, XML, InstantDB, Paradox, BorlandJDataStore, Csv, MS Excel, Qed, HSQL, Compiere ERP, SalesForce.Com, SugarCRM, Goldmine, any JDBC data sources and more). Supports bi-directional integration, is platform independent and can be used without coding via the Visual Job Designer. An on-demand version supports Salesforce and QuickBooks.

Centerprise Data Integrator provides a powerful, scalable, high-performance, and affordable integration platform designed for ease and is robust enough to deal with complex data integration challenges. The complex data mapping capabilities make it a good platform for overcoming the challenges of complex hierarchical structures such as XML, electronic data interchange (EDI), web services, and more. The expanding library of Centerprise Connectors is preconfigured to provide a plethora of integration options, enabling high-speed integration and migration to quickly and easily integrate with, or migrate to, leading enterprise CRM and ERP applications, as well as connectors for SOAP and REST web services that can be used to connect to a wide range of web services, including search engines and social media platforms .

CloverETL product family comes in the free community edition with core functionality and three paid for versions that incrementally include more connectors, scheduling and automation, and parallel processing and big data support.

Elixir Data ETL is designed to provide on-demand, self-serviced data manipulation for business users as well as for enterprise level data processing needs. Its visual-modeling paradigm drastically reduces the time required to design, test and implement data extraction, aggregation and transformation - a critical process for any application processing, enterprise reporting and performance measurement, data mart or data warehousing initiatives. Ready for web-based deployment, Elixir Data ETL allows business users to quickly obtain the critical information for their business decisions and operational needs, freeing up the IT group to focus on enterprise level IT issues.

Informatica's family of enterprise data integration products access and integrate data from any business system, in any format, and deliver that data throughout the enterprise at scale and at any speed. Powered by Vibe™, these enterprise data integration products enable your IT organization to scale with your business needs, dramatically lower costs, boost productivity, and reduce risk. At the same time, they enable business-IT collaboration and co-development to deliver on business demands for timely, relevant, trustworthy data. Informatica PowerCenter caters for highly scalable, high-performance enterprise data integration software. By leveraging Vibe, it serves as the foundation for all data integration projects – from departmental and project-based work, to enterprise integration initiatives, and beyond, for Integration Competency Centers (ICC).

Talend's data integration products provide an extensible, highly-performant, open source set of tools to access, transform and integrate data from any business system in real time or batch to meet both operational and analytical data integration needs. With 800+ connectors, it integrates almost any data source. The broad range of use cases addressed include: massive scale integration (big data/ NoSQL), ETL for business intelligence and data warehousing, data synchronization, data migration, data sharing, and data services.

Syncsort products cover three main areas of functionality:

- DMX is full-featured data integration software that helps organizations extract, transform and load more data in less time
- DMX-h offers a unique approach to Hadoop Sort and Hadoop ETL, that lowers the barriers for wider adoption, helping organizations unleash the full potential of Hadoop. Eliminate the need for custom code, get smarter connectivity to all your data, and improve Hadoop's processing efficiency.
- Syncsort MFX delivers the fastest and most resource-efficient mainframe sort, copy, join technology available, and is the only mainframe sort solution that offloads CPU cycles to zIIP engines.

Open Source Platforms

Apatar provides connectivity to many popular applications and data sources (Oracle, MS SQL, MySQL, Sybase, DB2, MS Access, PostgreSQL, XML, InstantDB, Paradox, BorlandJDataStore, Csv, MS Excel, Qed, HSQL, Compiere ERP, SalesForce.Com, SugarCRM, Goldmine, any JDBC data sources and more). Supports bi-directional integration, is platform independent and can be used without coding via the Visual Job Designer. An on-demand version supports Salesforce and QuickBooks.

Clover Editions are built on an Open Source Engine. The engine is a Java library and does not come with any User Interface components such as a Graph Designer. The developer / embedding application is responsible for managing graphs rather than using the Clover Designer or Server UI. However, your application does have access to most of powerful data transformation and ETL features that Clover uses throughout its own product range. The CloverETL Open Source Engine can be embedded in any application, commercial ones as well.

Jaspersoft ETL is easy to deploy and out-performs many proprietary and open source ETL systems. It is used to extract data from your transactional system to create a consolidated data warehouse or data mart for reporting and analysis.

KETL™ is a premier, open source ETL tool. The data integration platform is built with portable, java-based architecture and open, XML-based configuration and job language. KETL™ features successfully compete with major commercial products available today. Highlights include:

- Support for integration of security and data management tools
- Proven scalability across multiple servers and CPU's and any volume of data
- No additional need for third party schedule, dependency, and notification tools

Pentaho's Data Integration, also known as Kettle, delivers powerful extraction, transformation, and loading (ETL) capabilities. You can use this stand-alone application to visually design transforms and jobs that extract your existing data and make it available for easy reporting and analysis.

Talend Open Studio is a powerful and versatile set of open source products for developing, testing, deploying and administrating data management and application integration projects. Talend delivers the only unified platform that makes data management and application integration easier by providing a unified environment for managing the entire lifecycle across enterprise boundaries. Developers achieve vast productivity gains through an easy-to-use, Eclipse-based graphical environment that combines data integration, data quality, MDM, application integration and big data.

Predictive Analytics

Predictive analytics is concerned with the analysis of historical data to discover patterns of behavior which might be useful in the future. The classic example is credit checking, and trying to establish who will, and who won't be a good credit risk in the future based on the analysis of historical data. Predictive analytics is quite different from business intelligence and reporting activities most organizations engage in. These tend to be called descriptive analytics, and as the name implies it is concerned simply with describing what has happened in the past and what might be happening now. Predictive analytics is concerned with the future.

To attempt predictive analytics an organization needs relevant data. If we are trying to establish which customers might be the best candidates for a promotion then we clearly need good quality customer based data. A great deal of effort is usually expended in making sure the data is as accurate as possible, since predictive analytics conforms to the garbage in – garbage out paradigm as much as anything else. But assuming we have decent data the next step is to use an analytics platform to prepare data and analyze it. There are many algorithms used in predictive analytics and knowing which ones to use requires a good deal of experience. In any case the algorithms will usually identify patterns of behavior and indicate which variables are important. The resulting models that are built need to be validated and checked by a domain expert, since analytics tools can be fooled by random coincidences.

Models deemed suitable can then be used to score new data. In the case of credit approval the predictive models will often produce a score where a threshold limit has to be reached to grant approval. In many industries it is absolutely essential that the inner workings of the model are understood – a person cannot be refused a loan simply because a piece of software says so – the reasons need to be understood by humans.

Predictive analytics is most widely used in customer related activities within many organizations, and can address issues such as churn rates, marketing, selling strategies and so on. But it is also used to identify when a machine may fail, or when a patient is likely to be readmitted into a hospital, or even when an employee might resign.

For 'standard applications' and particularly in sales and marketing activities there are many technology suppliers selling ready made solutions, and so the user does not need to understand how the technology works. In other applications however it is necessary to employ skilled data scientists and analysts to create a useable predictive model. This is especially true of big data (a meaningless term really – it's just data) where special consideration needs t be given to the nature of the data.

These are early days for predictive analytics and the underlying technology, mainly derived from statistics and machine learning, is advancing quickly. It is already used to determine how goods are laid out in a store, which movies are recommended to a particular viewer, and who might be at risk of various diseases. It will become ubiquitous and affect every aspect of business and our lives.

Predictive Analytics Economics

The desirable economics associated with any investment is easily enough stated – the benefits should exceed the costs. In an immature domain such as predictive analytics the economics are not so easily established, and this is made more difficult by the hype, over-expectation, inexperience and general confusion surrounding the topic. Predictive analytics and big data are everywhere, and we've become so enamored with the buzzwords and hype that the economics seem like a rather dull topic to address.

The first thing we need to establish is why a business would want to use predictive technologies. I shall ignore big data simply because it's a non-topic, despite the hype. It's plumbing for data – bigger pipes, more of them and different shapes – but just plumbing. So back to predictive analytics. The reason your organization should be interested in these new technologies is because they enable a second wave of business automation. The first wave lasted over four decades and was primarily concerned with the automation and efficiency of business processes. The swan song of this era was actually the Enterprise Resource Planning suite. It integrated transactional systems and made them much easier to manage – in theory at least.

So welcome to the second wave of automation – decision automation. This is the real reason why predictive technologies are so important – they allow us to automate (completely or partially) the decision making process in our businesses. As always the real pioneers are to be found in financial services. Loan approval, fraud detection, customer targeting and so on, almost always involve the use of predictive models. They not only automate the decision process, but they make it more effective. It's actually a much more powerful use of technology than the process automation that went before it.

This background is necessary to understand the benefits that might be derived from predictive analytics. Provided the models are accurate, that the data is available and of good quality, that the models can be deployed into a production environment, and that the effects of the models can be measured and managed, we have a formula for a very significant lift in business efficiency and efficacy. And what is more it is measurable – given the will and the tools to measure.

So somewhat surprisingly the difficult part of the equation – the benefits – is actually easier to measure that we might first have thought. Next we need to consider the costs, and thrown into this we need to consider risks. The costs are fairly easily listed – data, technology, skills, management overhead, training etc. etc. The risks however are harder to quantify, because unlike process automation decision automation has more profound risks associated with it. These can be roughly categorized as:

- Faulty models due to poor data quality and low skill levels.
- Poor integration with production systems.
- Unmanageable complexity and poor monitoring and management.

The immutable law of risk and return means that a technology that might deliver significant gains might also deliver significant risk. A well conceived and executed predictive model might easily revolutionise some part of a business. A sloppily developed and managed model might do a business considerable harm. A loan

application model that says 'no' to the majority of applicants is hardly going to help a business – but hopefully it would be quickly spotted. But when the number of predictive models used is counted in the hundreds, an errant model (or models) may be much harder to spot. The solution, as always, is model management and reporting; although such is the immaturity of predictive analytics that only a few suppliers provide any such capability.

To summarize. Predictive technologies are a key enabler in the automation of business decisions. The potential benefits are not only more efficient processes, but also more effective ones. The costs associated with decision automation (the reason we use predictive technologies) are easily listed. The risks however need more careful consideration, and frankly need to be taken very seriously – if you want greater gains you inevitably have to take greater risks.

This in a nutshell is the economics of predictive analytics – a key enabling technology in decision automation. Those who get it right will start to lap those who haven't got a clue – and it will show.

Predictive Analytics – The Idiot's Way

Here is a real idiot's guide to predictive analytics:

- Get the relevant data.
- Let the predictive analytics algorithms loose on the data.
- Find the best patterns.
- Apply them to new data to predict outcomes – who might be a bad loan risk, which customers will respond to an offer etc.

Suppliers may talk in these terms because it is in their interest to make it sound easy and without risk – the opposite is true. There are many reasons why your predictive models might end up being more of a liability than an asset, but I'll focus on just one – curve fitting, which is also known by several other names. An example will clarify. Imagine tossing a coin ten times, recording the outcome as H – head and T – tails. Lets say the outcome is H H T T H H T H H T. Now any pattern detection software worth its salt will proudly deliver the following pattern – that after 2 heads a tail follows. But wait a minute. If you are willing to wager that this will happen in the future then ruin will almost certainly be the outcome. Each flip of the coin is a random independent event. We all know this.

Now scale the whole thing up to billions of records in a database with possibly hundreds of attributes (Name, Phone, Age, Salary, Zip etc). Is it possible that random patterns appear in the data to mislead us – yes absolutely. Now the people who conduct the analysis generally know about these things, and so they will reserve part of the data set to test a pattern. In fact they may use something called k-fold cross validation where the reserved section of data is varied across multiple attempts to build a model. But look at our sequence of heads and tails. If we had reserved the last three flips to test our hypothesis then we would still have come to the conclusion that it is true. These random patterns, which data mining algorithms will happily throw out as candidate predictive models, are just ghosts in your data.

The whole issue of whether a pattern is valid or not is actually extremely complicated, and well beyond the understanding of many who practice in this field. The more frequently a data set is interrogated for patterns the more likely we are to find them, and the more likely they are to be erroneous – it's called data mining bias among other things. Big data with thousands of attributes attracts problems of its own, and despite the popular view that billions of records can't tell lies, in reality they tell a whole new genre of lies.

Fortunately there is a sanity check on all of this. Domain experts will know whether a pattern makes sense. If it doesn't then it may just be that it is newly discovered knowledge, but with a domain expert at hand we can be suspicious until the suspicions are negated. Just another example of how truly stupid computers are.

Why Your Predictive Models May Be Wrong

Some of the predictive models currently being used by organizations of all types will be causing damage. Suppliers and consultants don't talk about this because it isn't exactly going to generate business (quite the opposite). When a few disaster stories eventually hit the headlines no doubt everyone will become more interested in predictive model error rates.

First of all we need to clarify a point. The term 'error rates' is used in data mining to depict the rate of false predictions a model gives. In this article I'm using it at a higher level of granularity – the number of erroneous models as a fraction of all models. Yes, organizations are deploying defective models for a variety of reasons – some of which can be addressed, and some of which cannot. Here is a rundown of why some of the predictive models used in your organization might be erroneous (with statistically meaningless actual prediction rates):

1. The people developing models do not understand the nuances and procedures necessary to ensure a model is fit for purpose (the most common reason). Do not assume that a PhD Statistician will have the requisite knowledge – their thesis may well have been concerned with some obscure mathematical properties of a particular distribution function. Finding people who know about the application of important techniques such as the Bonferroni correction is not so easy.
2. Your data may be messed up and simply unable to deliver accurate models. Even mild amounts of mess can produce incorrect models (this is particularly true of techniques such as decision trees, which are inherently unstable).
3. Suppliers have convinced your management that it's all about dragging visually appealing icons around a graphical interface and pressing the 'go' button. The 'ease-of-use' promise is as old as the hills, and is the technology supplier's best friend when selling to confused managers. Trouble is that it works, but always leads to trouble.
4. The fundamental act of searching through a large space of variable, variable value, parameter and data set combinations means that a very high percentage of such combinations are meaningless – but of course your algorithms do not know this. Such a scenario is ideal for a simple application of Bayes rule, which invariably shows that error rates are going to be much higher than one might imagine.

5. Political pressure. 'We've got piles of this item in stock, produce a predictive model that shows if we drop its price by 50% we will also sell 200% more of this other item." – the Sales Director. "Oh and by-the-way, if it all goes belly-up I'll blame the model.' There is nothing to say here really is there, other than this was common practice in the banks prior to the 2008 collapse – and no doubt still is.
6. Things change. The data used to build a model is always historical (unless you have discovered time travel). What was fashionable one year (remember Facebook) might not be fashionable next. Predictive models assume everything remains the same – it doesn't.

I imagine there are other reasons, but 6 is already one too many for me. Reasons 1 and 2 are addressable – 3,4,5,6 probably not.

Predictive models are being used to great effect, but everyone will also be using models that are just plain wrong. The key to eliminating defective models is monitoring and management on an ongoing basis. Without such vigilance you may just end up with the dumbest smart applications in your industry.

Enterprise Predictive Analytics Platforms Compared

Predictive analytics is concerned with trawling through historical data to find useful patterns which might be used in the future. As such it employs data mining techniques to find the patterns, and once found and verified they are applied via some scoring mechanism, where each new event is scored in some way (e.g. new loan applicants are scored for suitability or not). The data mining platforms compared in this article represent the most common alternatives many organizations will consider. The analysis is high level, and not a feature by feature comparison – which is fairly pointless in our opinion. The five criteria used to compare the products are:

- Capability – the breadth of the offering.
- Integration – how well the analytics environment integrates with data, production applications and management controls.
- Extensibility – very important and a measure of how well a platform can be extended functionally and how well it scales.
- Productivity – the support a platform offers for productive work.
- Value – likely benefits versus costs.

This form of analysis creates some surprises, but you need to look at the full review to see why a particular offering does so well.

Important: If you want to see the full review click on the score. If you want to visit the vendor website click on the name.

4.4 – Revolution Analytics has taken R (the open source analytics platform) and sanitized it for enterprise use. Some technicians may feel it doesn't need sanitizing, but business and technology managers would probably disagree. In any case it is very hard to fault – which makes the review quite short.

4.3 – IBM Predictive Analytics. Large corporations looking for enterprise wide analytics capability would be foolish not to consider IBM. This behemoth of a supplier has got it all – at a price. You just have to decide whether you want to pay it.

4.2 – Actian provides a complete big data and analytics environment for enterprise analytics needs. What is more the technology is advanced, facilitating analytics tasks that simply are not possible with many other technologies. These are big claims, but Actian has been working quietly in the background to develop and acquire technology that is certainly way ahead of many big data analytics offerings.

4.2 – FICO provides quite unique technology and elegantly combines predictive analytics with prescriptive analytics and business rules management. It's a formidable combination of capabilities, and it is now available in the FICO Analytic Cloud, so the technology can be accessed by medium size businesses as well as the large corporations that have traditionally used it.

4.2 – KNIME has tended to be overshadowed by RapidMiner until recently. However, since RapidMiner recently assumed a more commercial posture (there is no longer a fully functional open source version), KNIME has assumed greater prominence as a leading analytics platform.

4.2 – SAS will almost certainly address every analytic need your organization could possibly face, and there is a large and skilled pool of SAS professionals around the world. The fly in the ointment is the cost of the technology, and the decision to go with SAS simply boils down to one of perceived value.

4.1 – Angoss provides business oriented data mining technology, avoiding technical complexity for its own sake, and oversimplified products that cannot deliver. This is a good compromise, and many businesses will derive benefits in their customer oriented operations from employing this technology.

4.1 – RapidMiner is an excellent data mining and statistics platform with a large following. It is in no way an end-user tool and requires a good deal of skill to use. With version 6 the product and company became much more commercial, and the recent acquisition of Radoop puts it in the big data league.

4.0 – Alpine Data Labs distinguishes itself through the excellent integration with the general business applications environment and its support for collaboration and management reporting.

4.0 – Salford Systems offers some very capable data mining technology indeed. It excels particularly in ensemble methods, and since these have proved to be some of the most powerful machine learning techniques Salford has won many competitions. Not for the novice, but something different to the usual algorithms that are used in predictive analytics.

4.0 – SAP InfiniteInsight (formerly known as KXEN prior to acquisition by SAP in 2013) addresses a particular set of predictive analytics problems in several well defined markets (typically, but not exclusively retail, financial services and telecoms). The two very significant features of InfiniteInsight are the speed with which predictive models can be built and the reliability of those models. It is not however a general purpose machine learning or data mining toolbox

3.9 – Blue Yonder is perhaps an appropriate name for this supplier. There are some pretty wild claims made concerning automation and its desirability when using predictive technologies, although the underlying technology is novel and sophisticated. Worth a look, but don't be taken in by the claims of the marketing people.

3.9 – Dotplot provides data mining, statistics, text mining and predictive analytics tools in an integrated, highly graphical cloud based environment. All that is needed to use dotplot is a browser. Resulting models can be integrated with other applications via web services using SOAP and REST protocols. Dig a little deeper and dotplot is actually a much needed graphical front end to R and Weka functions.

3.9 – Oracle has done a surprisingly good job with its predictive analytics platform. It will in the main only be of interest to existing Oracle users, but the in-database analytics does have a broader appeal.

3.9 – STATISTICA from Dell embraces most of the analytics tools many organizations will need, both large and small. One of the most powerful aspects of the product set is the level of integration, with seamless

connections between disparate modes. Statistics, machine learning, data mining and text mining are all at the disposal of the user without having to migrate from one environment to another.

3.8 – BigML aims to make predictive analytics easy, and certainly it provides an easy to use drag and drop, point and click interface. Whether predictive analytics will ever be easy is a different matter – there are many potholes even for experienced analysts. In the main BigML uses decision trees to create models, and some ensemble methods in a cloud based environment.

Open Source and Free Time Series Analytics Tools

GMDH Shell is a simple yet powerful forecasting software, developed by GMDH LLC. Based on neural networks, the software allows you to easily create predictive models, as well as preprocess data with a simple point-and-click interface. GMDH Shell is much faster than other tools based on neural networks thanks to optimization of core algorithms and excellent parallel processing capabilities.

MacAnova is a free, open source, interactive statistical analysis program for Windows, Macintosh, and Linux. MacAnova has many capabilities but its strengths are analysis of variance and related models, matrix algebra, time series analysis (time and frequency domain), and (to a lesser extent) uni- and multi-variate exploratory statistics. Core MacAnova has a functional/command oriented interface, but an increasing number of capabilities are available through a menu/dialog/mouse type interface.

R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are briefly summarized below. There is also a considerable overlap between the tools for time series and those in the Econometrics and Finance task views. Base R contains substantial infrastructure for representing and analyzing time series data. The fundamental class is "ts" that can represent regularly spaced time series (using numeric time stamps).

Weka now has a dedicated time series analysis environment that allows forecasting models to be developed, evaluated and visualized. This environment takes the form of a plugin tab in Weka's graphical "Explorer" user interface and can be installed via the package manager. Weka's time series framework takes a machine learning/data mining approach to modeling time series by transforming the data into a form that standard propositional learning algorithms can process. It does this by removing the temporal ordering of individual input examples by encoding the time dependency via additional input fields.

Zaitun Time Series is a free and open source software designed for statistical analysis of time series data. It provides easy way for time series modeling and forecasting. It provides several statistics and neural networks models, and graphical tools that will make your work on time series analysis easier, and provides several statistics and neural networks models, and graphical tools that will make your work on time series analysis easier. Zaitun Time Series has a capability to deal with the stock market data. It is facilitated with the stock data type which can help the visualization of the stock market data in a candle stick graph.

Customer Analytics Platforms

11Ants RAP is a cloud-based customer science platform which drives retail growth in medium to large retailers. RAP begins with the assumption that there is a huge amount of unrealized value in the transactional data you gather and store - telling detailed stories about your customers and their behaviour: what they buy, what they don't, when they shop, how often, how price sensitive they are, and much more.

The RAP Customer Science modules provide a strategic and tactical framework for retail growth. The modules include: Company Performance Module, Category Performance Module, Category Drill Down Module, Territory Drill Down Module, Promotion Impact Module, Cross-Sell Opportunity Module, Customer Retention Module, Loyalty Participation Module, Product Customer Profile Module, Product Substitution Module, Mailing List Builder Module and Look-Alike Customer Module.

Agilis - is an innovative leader of Customer and Operational Business Analytics. Its solutions use both internal and external data sets about your customer to win and keep customers, increase ARPU (average revenue per user), and reduce risk and expenses throughout the entire customer lifecycle.

- Acquisition and On-Boarding - Agilis Point of Sale (POS) Solution performs real-time evaluation of each prospect across all channels at the time of acquisition to present the right offers to new customers, identify customers who represent financial risk and enhance customer experience to drive future profitability and loyalty.
- Collections and Churn - Learning from subscriber's behavior from onboarding through the lifecycle; Agilis Customer Analytics scores a customer's propensity to churn, and to pay, providing proactive churn management and collections effectiveness. Utilizing the subscriber's scores and potential future value, Agilis Analytics helps target those customers for rescue along with providing proactive actions to minimize financial risk.
- Subscriber Management - Agilis subscriber management analytics maximize market share and lifetime customer value. Agilis Customer Analytics utilize subscriber and third party data, when applicable, to predict consumer behavior to allow you to optimize revenue opportunities and deliver enhanced customer experience.
- Financial Risk Management- - Agilis Risk Management solutions are out-of-the-box self learning adaptive analytics, purpose built to address current and new risks that are associated with next generation networks and services delivering end-to-end risk management solutions throughout the subscriber life cycle.

Alteryx - is specifically an analytics technology and solutions supplier. Its customer analytics solutions address multiple issues, including:

- Anticipate customer behavior so you can drive repeatable business.
- Target prospects with similar attributes as customers with hyper-local messaging.

- Drive significant improvements in marketing campaign effectiveness, customer retention, operational efficiency and brand loyalty.

Angoss - customer analytics can help build, host and maintain industry-based predictive models for analysts, in order to discover new patterns and trends in their data and accurately predict customer behavior.

There are 8 stages that make up the Customer Analytics Lifecycle:

- Customer Segmentation allows analysts to understand the landscape of the market in terms of customer characteristics and whether they naturally can be grouped into segments that have something in common.
- Customer Acquisition is used to acquire new customers and increase market share, and often involves offering products to a large number of prospects.
- Upsell/Cross Sell aim to provide existing customers with additional or more valued products.
- Next Product/Recommendation aims to promote additional products to existing customers when the time is right.
- Customer Retention/Loyalty/Churn aim at maintaining and rewarding customer loyalty and reduce customer defection.
- Customer Lifetime Value is used to design programs to appreciate and reward valuable customers.
- Product Segmentation allows for the optimization of using product affinity; in most cases using Market Basket Analysis.

Angoss customer analytics are offered through Angoss KnowledgeSEEKER®, KnowledgeSTUDIO®, KnowledgeREADER™, and KnowledgeSEEKER Marketing Edition and KnowledgeSCORE™ providing customer analytics solutions for data profiling and visualization, Decision Tree analysis, predictive modeling, and scoring and strategy building.

Banks, retailers, telcos, and healthcare providers use Angoss analytics to develop their customer analytics lifecycle.

FICO - is a leading provider of analytics solutions with a broad base of technology and solutions offerings, as will be apparent from the solutions listings below. It has a particularly dominating presence in financial services, insurance, pharmaceuticals and retail. The solutions domains it addresses include:

- Debt Management - Collections and Recovery, Customer Growth & Retention, Campaign Management, Customer Strategy Management, Multi-Channel Customer Communications
- Customer Originations - Originations
- Fraud & Security - Enterprise Fraud & Security, Compliance
- Scores - Scoring & Scoring Solutions

The products offered include:

Customer Growth & Retention Products

- FICO® TRIAD® Customer Manager
- FICO® Customer Dialogue Manager
- FICO® Customer Communications Services

Customer Originations Products

- FICO® Origination Manager
- FICO® Customer Communications Services
- FICO® Application Fraud Manager
- FICO® LiquidCredit® Service

Debt Management Products

- FICO® Debt Manager™ Solution
- FICO® Engagement Analyzer
- FICO® PlacementsPlus® Service
- FICO® Risk Intervention Manager
- FICO® Network

Fraud & Security Products

- FICO® Falcon® Fraud Manager
- FICO® Identity Resolution Engine
- FICO® Fraud Resolution Manager
- FICO® Application Fraud Manager
- FICO® Insurance Fraud Manager, Health Care Edition

Scores and Scoring Solutions

- FICO® Score
- FICO® Score Open Access
- FICO® Expansion Score
- myFICO® service

HP Vertica - addresses many analytics problems and a solution is offered specifically for customer analytics. It allows organizations to collect, manage, and analyze data from disparate sources, including web logs, third-party analytics tools, social media, and traditional CRM and customer records from enterprise systems.

Some solution examples include:

Online & Mobile

Online and mobile businesses have a need to improve the effectiveness of their website. Analyzing clickstream data provides rich insight into which pages are effective and which pages site visitors ignore. When combined with sales and conversion data, clickstream analysis can help you discover the most effective series of steps needed to encourage conversions, sales, and add-on purchases.

Retail

Increased competition and shrinking margins are compelling retailers to increase the amount of data they collect to gain a competitive advantage. Loyalty programs, customer tracking solutions, and market research, when combined with sales and inventory data, provides rich insights that drive decisions around products, promotions, price, and distribution management. Data driven decisions enable sales based on actual purchase patterns, instead of guess work.

Financial Services

Banks, insurance companies, and other financial institutions are using customer analytics to understand the lifetime value of customers, increase cross-sales, and manage customer attrition, among other programs. The ability to understand how customers use different banking programs – credit and debit cards, mortgages and loans, and online banking tools – allows financial services companies to develop targeted campaigns and value added offers that increase customer satisfaction and profits.

Communication Service Providers (CSPs)

Call Detailed Records (CDRs) contain a wealth of information such as the length of a call, number called, weather the call was dropped or not, etc. CDRs create massive amounts of valuable data for CSPs. The ability to analyze these massive volumes of data allows CSPs to develop customer focused promotions that attract and retain customers, reducing churn and increasing profitability.

IBM - offers a broad set of customer analytics capabilities. Its solutions aim to:

- Target customers with highly relevant offers across all channels, including digital, mobile and social.
- Understand customers in the context of their individual relationships with your brand.
- Engage with customers through the right channel, with the right message, at the right time.
- Predict which customers are at risk of churning and why, and take actions to retain them.
- Measure customer sentiment and spot emerging trends in social media and survey data.

- Maximize customer lifetime value through personalized up-sell and cross-sell offers.

The solution modules it provides to address these issues include:

- Customer acquisition
- Attract more valuable customers with smarter analytics.
- Lifetime Value
- Customer lifetime value
- Maximize the long-term profitability of every customer.
- Loyalty
- Customer loyalty
- Improve loyalty and satisfaction by understanding and engaging your customers.
- IBM Digital Analytics
- Digital marketing optimization
- Turn insights from web analytics and digital customer profiles into personalized campaigns.
- Profitability
- Understand the value your customers represent to your bottom line.
- IBM Social Media Analytics
- Social media analytics
- Unlock the value of customer sentiment in social media.
- Marketing Analytics
- Marketing performance analytics
- Deliver better results and bigger returns with customer analytics.

Optimove - is Web-based software that implements a systematic approach to running, measuring and optimizing customer marketing campaigns. With Optimove, marketers and customer retention experts maximize the lifetime value of every customer by consistently achieving the best match between every campaign and every customer.

In other words, Optimove's unique customer modeling technology understands every customer – and accurately predicts how each marketing campaign will impact their behavior. Armed with this insight, marketers unlock the magic of one-to-one marketing campaigns to convert more customers, increase spending and reduce churn.

Marketers use Optimove to build, track and optimize a comprehensive customer marketing plan. The software integrates with email service providers, message boards and multi-channel campaign execution systems to automatically deliver campaigns to customers at scheduled times.

The software then automatically analyzes the performance of every customer marketing campaign – in monetary terms – using test and control groups or A/B tests. Every campaign thus becomes a measurable marketing experiment which feeds the software's self-learning recommendation engine!

SAP - applications form the bedrock for many large organizations. It offers an innovative and powerful solution to customer analytics through InfiniteInsight. With traditional predictive analytics, you can expect to spend a great deal of time on activities that are manual, repetitive, and prone to human error. SAP InfiniteInsight (formerly KXEN) has changed all that – automating most of the effort so that users can gain unprecedented customer insight and make forward-looking decisions with ease.

SAS - provides a very broad suite of analytics technologies, and its Customer Intelligence offering specifically addresses the customer analytics space. Its solutions aim to:

- Increase response rates, customer loyalty and, ultimately, ROI by contacting the right customers with highly relevant offers and messages.
- Reduce campaign costs by targeting those customers most likely to respond.
- Decrease attrition by accurately predicting customers most likely to leave and developing the right proactive campaigns to retain them.
- Deliver the right message by segmenting customers more effectively and better understanding target populations.

The products which address these issues include:

- SAS® Enterprise Miner™ - Streamline the data mining process to create highly accurate predictive and descriptive models based on large volumes of data.
- SAS® Customer Link Analytics - Categorize customer relationships and identify the communities in which they interact.
- SAS® Rapid Predictive Modeler - Generate predictive models quickly and easily, and apply results to improve decision making.
- SAS/ETS® - Model, forecast and simulate business processes with econometric and time series analysis.
- SAS® Data Management - Ensure better, more reliable data integrated from any source.

Open Source and Free Data Mining Platforms

Free data mining software ranges from complete model development environments such as Knime and Orange, to a variety of libraries written in Java, C++ and most often in Python. The list below provides summaries of the most popular alternatives.

Apache Mahout supports mainly three use cases: Recommendation mining takes users' behavior and from that tries to find items users might like. Clustering takes e.g. text documents and groups them into groups of topically related documents. Classification learns from exisiting categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category.

Data.Mining.Fox (DMF) from easydatamining is a free data mining tool that hides much of the background complexity. The interface takes users through several well defined steps from data import through to predictions based on new data.

The Databionic ESOM Tools is a suite of programs to perform data mining tasks like clustering, visualization, and classification with Emergent Self-Organizing Maps (ESOM).

The gnome-datamine-tools is a growing collection of tools packaged to provide a freely available single collection of data mining tools.

Jubatus is the first open source platform for online distributed machine learning on the data streams of Big Data. Jubatus uses a loose model sharing architecture for efficient training and sharing of machine learning models, by defining three fundamental operations; Update, Mix, and Analyze, in a similar way with the Map and Reduce operations in Hadoop.

KEEL is an open source (GPLv3) Java software tool to assess evolutionary algorithms for Data Mining problems including regression, classification, clustering, pattern mining and so on. It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques (training set selection, feature selection, discretization, imputation methods for missing values, etc.), Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches (Pittsburgh, Michigan and IRL, ...), and hybrid models such as genetic fuzzy systems, evolutionary neural networks, etc.

Knime is a widely used open source data mining, visualization and reporting graphical workbench used by over 3000 organisations. Knime desktop is the entry open source version of Knime (other paid for versions are for organisations that need support and additional features). It is based on the well regarded and widely used Eclipse IDE platform, making it as much a development platform (for bespoke extensions) as a data mining platform.

MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.

ML-Flex uses machine-learning algorithms to derive models from independent variables, with the purpose

of predicting the values of a dependent (class) variable.

Mlpy is a Python module for Machine Learning built on top of NumPy/SciPy and the GNU Scientific Libraries. mlpy provides a wide range of state-of-the-art machine learning methods for supervised and unsupervisedproblems and it is aimed at finding a reasonable compromise among modularity, maintainability, reproducibility, usability and efficiency.

Modular toolkit for Data Processing (MDP) is a library of widely used data processing algorithms that can be combined according to a pipeline analogy to build more complex data processing software. Implemented algorithms include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Slow Feature Analysis (SFA), and many more

Orange is a very capable open source visualization and analysis tool with an easy to use interface. Most analysis can be achieved through its visual programming interface (drag and drop of widgets) and most visual tools are supported including scatterplots, bar charts, trees, dendograms and heatmaps. A large number (over 100) of widgets are supported.

R is a programming language, but there are literally thousands of libraries that can be incorporated into the R environment making it a powerful data mining environment. In reality R is probably the most flexible and powerful data mining environment available, but it does require high levels of skill.

Rattle (the R Analytical Tool To Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets.

RapidMiner is perhaps the most widely used open source data mining platform (with over 3 million downloads). It incorporates analytical ETL (Extract, Transform and Load), data mining and predictive reporting. The free version is now throttled and is called the trial version.

scikit learn provides many easy to use tools for data mining and analysis. It is built on python and specifically NumPy, SciPy and matplotlib.

Shogun machine learning toolbox's focus is on large scale kernel methods and especially on Support Vector Machines (SVM). It provides a generic SVM object interfacing to several different SVM implementations, among them the state of the art OCAS,Liblinear, LibSVM, SVMLight, SVMLin and GPDT. Each of the SVMs can be combined with a variety of kernels. The toolbox not only provides efficient implementations of the most common kernels, like the Linear, Polynomial, Gaussian and Sigmoid Kernel but also comes with a number of recent string kernels as e.g. the Locality Improved, Fischer, TOP, Spectrum, Weighted Degree Kernel (with shifts).

TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.

Vowpal Wabbit is the essence of speed in machine learning, able to learn from terafeature datasets with ease. Via parallel learning, it can exceed the throughput of any single machine network interface when

doing linear learning, a first amongst learning algorithms.

WEKA is set of data mining tools is incorporated into many other products (Knime and Rapid Miner for example), but it also a stand-alone platform for many data mining tasks including preprocessing, clustering, regression, classification and visualization. The support for data sources is extended through Java Database Connectivity, but the default format for data is the flat file.

Open Source and Free Social Network Analysis Tools

Cuttlefish runs on UNIX systems and employs many well known layout algorithms for visualization. It supports interactive manipulation of layout and process visualization.

Cytoscape is used for social network, biological, semantic web and general network analysis. It supports Apps which extend the functionality in areas such as new layouts, file format support and scripting. A very sophisticated offering.

Gephi is an open source visualization and exploration platform available on Windows, Linux and Mac OS X. It supports all types of networks – directed, undirected and mixed, and is capable of handling very large network graphs of up to one million nodes. Various metrics are supported including betweenness, closeness, diameter, clustering coefficient, average shortest path, pagerank and HITS. Dynamic filtering allows edges and/or nodes to be selected based on network structure or data. Ideal for social network analysis, link analysis and biological network analysis. Perhaps the most advanced of the open source tools.

GraphChi can run very large graph computations on just a single machine, by using a novel algorithm for processing the graph from disk (SSD or hard drive). Programs for GraphChi are written in the vertex-centric model, proposed by GraphLab and Google's Pregel. GraphChi runs vertex-centric programsasynchronously (i.e changes written to edges are immediately visible to subsequent computation), and in parallel. GraphChi also supports streaming graph updates and removal of edges from the graph. The promise of GraphChi is to bring web-scale graph computation, such as analysis of social networks, available to anyone with a modern laptop.

GraphInsight was once a commercial product, but is now open source. It is fast and highly scalable (5 million nodes and 4 million links can be accommodated). Visualisation supports 2 and 3 dimensional models, and interaction with visualizations can be accomplished using the embedded Python shell. A C++ graph drawing library is also available for bespoke projects.

JUNG — the Java Universal Network/Graph Framework--is a software library that provides a common and extendible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. It is written in Java, which allows JUNG-based applications to make use of the extensive built-in capabilities of the Java API, as well as those of other existing third-party Java libraries.

The JUNG architecture is designed to support a variety of representations of entities and their relations, such as directed and undirected graphs, multi-modal graphs, graphs with parallel edges, and hypergraphs. It provides a mechanism for annotating graphs, entities, and relations with metadata. This facilitates the creation of analytic tools for complex data sets that can examine the relations between entities as well as the metadata attached to each entity and relation.

The current distribution of JUNG includes implementations of a number of algorithms from graph theory, data mining, and social network analysis, such as routines for clustering, decomposition, optimization, random graph generation, statistical analysis, and calculation of network distances, flows, and importance

measures (centrality, PageRank, HITS, etc.).

JUNG also provides a visualization framework that makes it easy to construct tools for the interactive exploration of network data. Users can use one of the layout algorithms provided, or use the framework to create their own custom layouts. In addition, filtering mechanisms are provided which allow users to focus their attention, or their algorithms, on specific portions of the graph.

NetworkX is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

- Python language data structures for graphs, digraphs, and multigraphs.
- Nodes can be "anything" (e.g. text, images, XML records)
- Edges can hold arbitrary data (e.g. weights, time-series)
- Generators for classic graphs, random graphs, and synthetic networks
- Standard graph algorithms
- Network structure and analysis measures

MeerKat is suitable for many types of network analysis, including social networks. If provides filtering mechanisms, interactive editing, support for dynamic networks, various metrics and automatically detects communities.

NodeXL is a free, open-source template for Microsoft® Excel® 2007, 2010 and 2013 that makes it easy to explore network graphs. With NodeXL, you can enter a network edge list in a worksheet, click a button and see your graph, all in the familiar environment of the Excel window.

- Flexible Import and Export Import and export graphs in GraphML, Pajek, UCINet, and matrix formats.
- Direct Connections to Social Networks Import social networks directly from Twitter, YouTube, Flickr and email, or use one of several available plug-ins to get networks from Facebook, Exchange, Wikis and WWW hyperlinks.
- Zoom and Scale Zoom into areas of interest, and scale the graph's vertices to reduce clutter.
- Flexible Layout Use one of several "force-directed" algorithms to lay out the graph, or drag vertices around with the mouse. Have NodeXL move all of the graph's smaller connected components to the bottom of the graph to focus on what's important.
- Easily Adjusted Appearance Set the color, shape, size, label, and opacity of individual vertices by filling in worksheet cells, or let NodeXL do it for you based on vertex attributes such as degree, betweenness centrality or PageRank.
- Dynamic Filtering Instantly hide vertices and edges using a set of sliders—hide all vertices with

degree less than five, for example.

- Powerful Vertex Grouping Group the graph's vertices by common attributes, or have NodeXL analyze their connectedness and automatically group them into clusters. Make groups distinguishable using shapes and color, collapse them with a few clicks, or put each group in its own box within the graph. "Bundle" intergroup edges to make them more manageable.
- Graph Metric Calculations Easily calculate degree, betweenness centrality, closeness centrality, eigenvector centrality, PageRank, clustering coefficient, graph density and more.
- Task Automation Perform a set of repeated tasks with a single click.

sna is a R library with range of tools for social network analysis, including node and graph-level indices, structural distance and covariance methods, structural equivalence detection, network regression, random graph generation, and 2D/3D network visualization.

Social Networks Visualizer (SocNetV) is a cross-platform, user-friendly tool for the analysis and visualization of Social Networks. It lets you construct networks (mathematical graphs) with a few clicks on a virtual canvas, or load networks of various formats (GraphML, GraphViz, Adjacency,Pajek, UCINET, etc). Also, SocNetV enables you to modify the social networks, analyse their social and mathematical properties and apply visualization layouts for relevant presentation.

Furthermore, random networks (Erdos-Renyi, Watts-Strogatz, ring lattice, etc) and known social network datasets (i.e. Padgett's Florentine families) can be easily recreated. SocNetV also offers a built-in web crawler, allowing you to automatically create networks from links found in a given initial URL.

The application computes basic graph properties, such as density, diameter, geodesics and distances (geodesic lengths), connectedness, eccentricity, etc. It also calculates advanced structural measures for social network analysis such as centrality and prestige indices (i.e. closeness centrality, betweeness centrality, information centrality, power centrality, proximity and rank prestige), triad census, cliques, clustering coefficient, etc.

SocNetV offers various layout algorithms based either on prominence indices or dynamic models (i.e. Spring-embedder) for meaningful visualizations of social networks. There is also comprehensive documentation, both online and while running the application, which explains each feature and algorithm of SocNetV in detail.

Commercial products include:

Sentinel Visualizer provides advanced Link Analysis, Data Visualization, Geospatial Mapping, and Social Network Analysis (SNA). Features include:

- Integrated knowledgebase, link analysis, social network analysis, geospatial, and timelines
- Industry standard database format
- Supports network multi-user environments

- Data visualization capabilities including 2D and 3D link charts and stereoscopic graphs.
- Built-in Social Network Analysis and Advanced Network Analysis.
- Automated layout tools to organize complex network charts.
- Geospatial Integration with Google Earth.
- Detection of network cut-points.
- Temporal Analysis.
- De-clutter tools to filter out noise and non-credible information.
- Advanced tools such as shortest path, all paths, and gradient metrics.
- Automated cluster/cell detection.
- Supports an unlimited number of databases.
- Integrated scalable entity and relationship knowledgebase.
- Name disambiguation through aliases and fuzzy searching.
- Grading of reliability of source and credibility of information for relationship information.
- Built-in support for storing any type of document (binary or text) in the knowledgebase.
- Administrator-configurable data types.
- Extensible metadata model for complete customization.
- Data import wizard for easy integration with many data sources.
- Dataset Builder to build complex queries quickly.
- Rich reporting exports to Excel, Word, PDF, and HTML.
- Export data to industry-standard XML format.

InFlow 3.1 performs network analysis AND network visualization in one integrated product - there is no passing files back and forth between different programs like many other tools. What is mapped in one window is measured in the other window -- what you see, is what you measure. InFlow excels at what-if analysis -- change the network, get new metrics -- just 2 clicks of the mouse. InFlow is designed to work with Microsoft Office and the WWW. Users do not need to be an expert in statistics to use InFlow.

InFlow provides easy access to the most popular network metrics. With visualization and metrics in one interactive interface, almost unlimited what-if scenarios are possible.

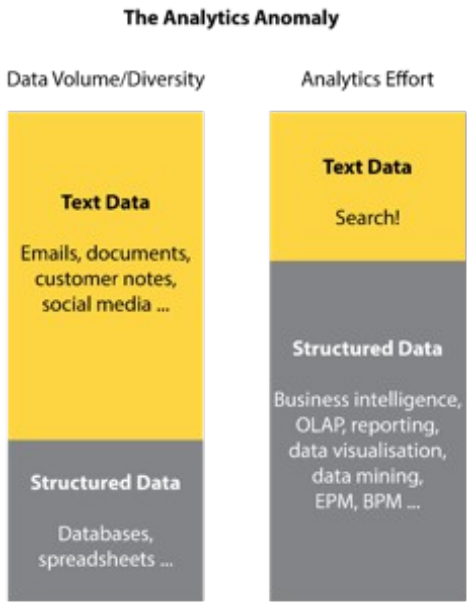
- Network Centrality / Centralization
- Small-World Networks

- Cluster Analysis
- Network Density
- Prestige / Influence
- Structural Equivalence
- Network Neighborhood
- External / Internal Ratio
- Weighted Average Path Length
- Shortest Paths & Path Distribution

NetMiner is a premium software tool for Exploratory Analysis and Visualization of Network Data. NetMiner allows you to explore your network data visually and interactively, and helps you to detect underlying patterns and structures of the network. It has the comprehensive data model expressing various types of nodes, links, node attributes and link attributes. Through its data model, NetMiner is able to represent most social, natural and physical phenomena as network data.

Text Analytics

The Business Value of Text Analytics



The business value of text analytics is fairly straightforward. The large amounts of text based data that most organizations possess, acquired and managed at considerable cost, can be analysed in such a way that insights can be gained to improve the efficacy and efficiency of business operations. Text based data are an untapped resource in many organizations. Structured data (customer details held in a database for example) on the other hand are very well exploited, primarily because they are specifically formatted for computer processing. While unstructured data, primarily text, is well suited for human communication, there have been significant hurdles to overcome to make it amenable to processing by computer systems. These barriers have been slowly eroded to the extent that significant value can now be extracted from text.

This is something of an irony since text based data typically accounts for eighty per cent of the data most organizations generate and process. Emails, documents, social data, comments and a wide variety of other text are usually archived, but typically not analysed in any meaningful way. The cost of creating, capturing and managing this information is considerable. In a service based business most employees can easily be categorized as information workers, and the cost of the information they generate is directly related to associated labour costs. Viewed in this way the cost of text data in many organizations is in excess of fifty per cent of all costs. Clearly any technology capable of automating the extraction of useful information from these data should be of interest.

The application of text analytics technologies has grown rapidly with increased regulation, the proliferation of social data, and efforts to record the thoughts and comments of customers. Embedded in the terabytes of unstructured data are patterns which can serve a diverse range of purposes, from flagging when a customer is likely to close their account, through to fraud detection. The value of text analytics is amplified when both structured and text data are combined, and to this end text mining technologies are witnessing significant uptake. In this scenario text data are converted into a form where they can be merged with structured data from transactional systems and are then scrutinized by data mining technologies, whose sole purpose is to uncover hidden structure in data and reveal exploitable patterns. It is then crucial that these patterns can be deployed in a production environment, with full monitoring of performance as scoring is performed on new incoming data. Managers will not be confident unless they can assess the benefits a predictive model is bringing on a real-time, ongoing basis.

To realize value from the very large sums invested in creating text data an organization needs to carefully plan and execute a business led initiative. This involves identification of business processes where text analytics might add value, the creation of text analytics capability, and a feedback loop in which information capture is informed by the outcome of analytics processes. This latter point is crucial, but somewhat surprisingly is often not mentioned by suppliers and consultants in this domain. If a certain type of information generates useful patterns then it becomes important to understand why, and attempt the capture of other information which might amplify the value of the analytics process.

Underlying all of this is some fairly simple economics - the cost of discovering and exploiting information derived from text analytics should be less than the value realized. Fortunately analytics often produces measurable outcomes captured by metrics such as lift. A two per cent increase in lift can mean a very considerable return on text analytics investments in many customer and marketing oriented activities.

Finally it should be noted that the scale and scope of text analytics will be accelerated by the current developments in big data technologies. The most heavily visited topic on the butleranalytics.com web site is text mining. We predict that this will become the largest growth area within the data analytics space, and a key differentiator in the benefits organizations reap from their analytics activities.

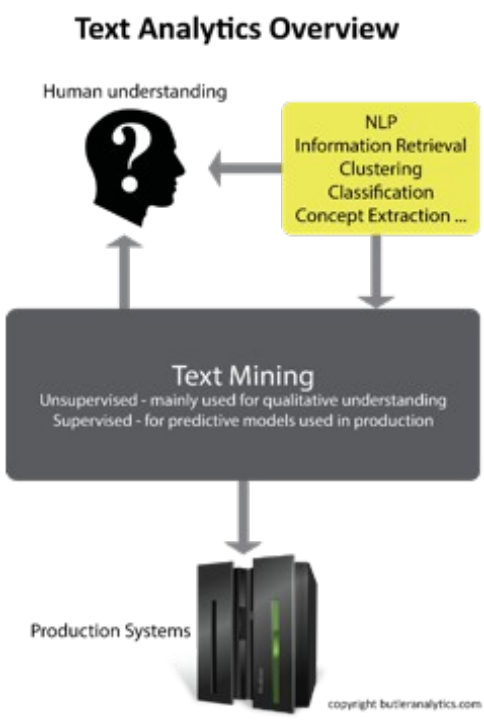
What is Text Analytics?

Text analytics convert unstructured text into useful information. This information can be in a format suitable for human consumption (categorized documents for example) or fed into computer systems to improve business processes (detecting customers who might defect). There are many techniques used in text analytics, but the target for the resulting information is always a computer system or people who need the information.

The information that text analytics can deliver to a person is very diverse. This ranges from language translation through to identifying important entities (people, places, products), categorizing documents, identifying important topics, establishing links between entities, establishing document similarities and so on. Much of this functionality comes under such headings as natural language processing (NLP), information retrieval, information extraction and several other domains which are still strongly associated with their academic roots. As far as the user is concerned this form of text analytics should simply reduce the overheads associated with finding and processing information, and many commercial products exist that perform exactly this function. Various surveys show that the average information worker spends up to a third of their time locating information and trying to make sense of it. If text analytics can reduce this overhead by just a few per cent, then simple math would show that the savings are considerable. In reality text analytics delivers much more than just a few per cent improvement, and tens of per cent improvement is common.

Processing unstructured text data so it can be processed by computer systems is a wholly different exercise. Powerful data mining algorithms, capable of identifying patterns within data, do not understand unstructured data. To this end many of the techniques mentioned above (NLP, concept extraction …) can be used to extract features from text (important entities for example) which can be used as input for the data

mining algorithms. These features are often combined with structured data to provide additional insights into behaviour. Text data in the form of customer notes might be processed to deliver features that show important terms used by customers, and when combined with customer records from a database will often improve the accuracy of patterns found. These might indicate suitable targets for a marketing campaign, or potential delinquency. The terms used for this type of activity are ambiguous, but for our purposes we can call this text mining and seen as an extension of data mining.



While text mining is often used to identify patterns which can be used in production systems, it too can provide output suitable for human consumption. This type of mining is called unsupervised learning - the data are simply fed into the algorithms and the output shows various clusters of documents, possibly revealing significant insights. A second type of text mining is more concerned with finding patterns that improve business processes through deployment in computer systems. This is called supervised learning where the text mining algorithms learn from a large sample of text data, and the resulting patterns are usually tested against new data the resulting pattern hasn't seen before. These patterns often classify new data (risk or no-risk for example), create probabilities of new data being in a particular class, or calculate a numerical value for new data (a credit limit for example).

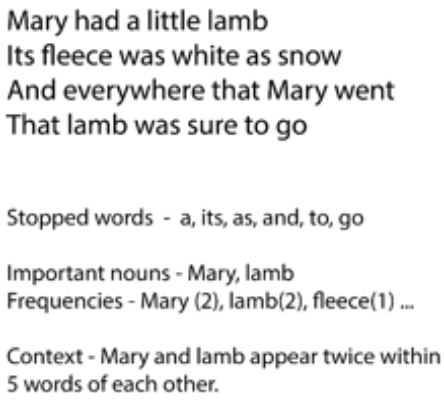
In summary text mining offers the potential to automate the analysis of text data and feed resulting patterns directly into production systems. Many other techniques exist to process language for human consumption, although some of these techniques can also provide input to business processes. Text mining employs many machine learning technologies, and since this is a domain of intense interest, it is here that many advances will be made. Coupled with the advances being made in the storage of text data (column databases for example), the use of text mining technologies will see accelerating uptake over the next few years. Of course the adoption of such technologies can happen through in-house initiatives or by employing ready-made solutions. As always the best route for many organizations will be the middle-way – technologies that address much of the problem at hand, but with a sufficiently powerful toolset that bespoke work is not problematical.

Text Analytics Methods

Natural language text is not a medium readily understood by computer systems, in contrast to the neatly arrange rows and columns in a database. This is the primary reason that text analytics has had such a long gestation before it could be usefully employed in a business arena. It also means that much of the effort involved in text analytics is preparatory work, to make sure the data are in a format that can be processed by text applications.

The first stage in dealing with text data is nearly always the process of identifying individual words and phrases (usually called tokens). Even this is not as simple as it sounds since abbreviations, acronyms, synonyms and ambiguity make the task quite involved (the word 'wave' has multiple meanings for example). It is also usually necessary to identify 'parts-of-speech', and specifically which words are nouns, verbs, adjectives and so on. Many words are meaningless as far as text analysis is concerned and can be 'stopped out'. Words such as 'a', 'it', 'and', 'to' and so on can usually be stopped and unsurprisingly are called stop words. A significant part of natural language processing is dedicated to these tasks, and it is a prerequisite before other work can be done. At the heart of this approach is an attempt to infer some level of meaning within documents (identify important entities, concepts and categories).

A wholly different approach can be adopted by using statistical methods. Here we might simply count the number of times various words appear in a corpus of documents and infer some level of importance from the resulting frequencies. One of the most useful metrics based on this approach is called the inverse document frequency. This increases in importance as a particular word appears frequently in a given document, but is not common in all documents. The word 'loan' may appear frequently in a corpus of documents and have no particular importance in a particular document. Whereas the word 'default' would appear less often (hopefully) and have more significance in a specific document. This approach can give useful results, but context and meaning is almost entirely lost. In an attempt to address this, short sequences of words called n-grams can be processed. This does at least offer the opportunity for frequent combinations of words to be identified. Significantly more sophisticated approaches are often used in commercial text analytics products, a good example being probabilistic latent sematic analysis where documents can be assigned to discovered topics.



The above methods, and many others, can be used to generate input to data mining activities. We might have a detailed transactional history of customer activity, but with little notion of how happy or otherwise the customers are. To this end we might use some of the above methods to identify customer sentiment and add additional variables (usually called features) to our customer records. This approach is proving to be successful in many applications.

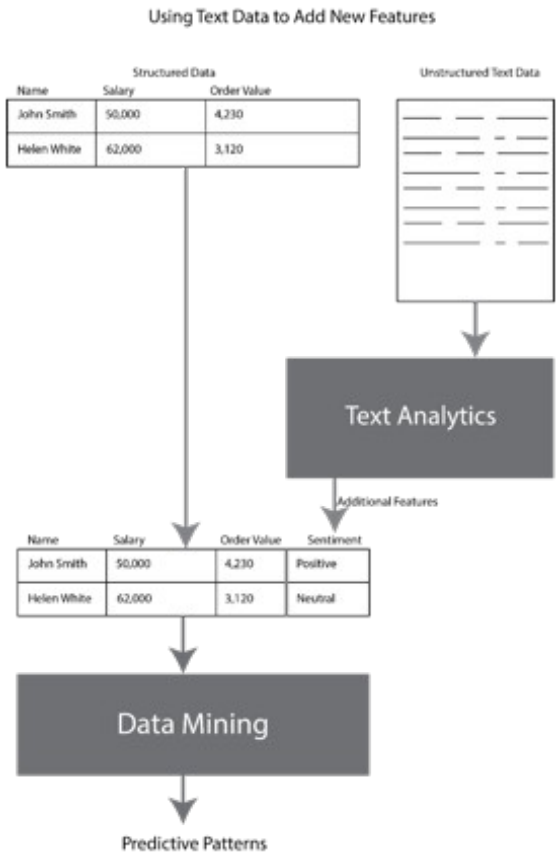
There are two ways to address the complexities associated with text analytics. The first is simply to buy a 'solution' for the task at hand. Various suppliers provide analytics solutions to a range of vertical and horizontal business needs. The benefits associated with this approach include fast time to implementation, reduced need for in-house staff and the ability to call upon a supplier that has experience in a particular domain. The downside is usually lack of tailoring to particular needs, less awareness of how an application actually works, and a potential dead-end if a solution cannot be modified sufficiently. The alternative is to build an in-house facility with associated costs, but with the opportunity to address specific needs accurately.

Text analytics can deliver simple to use functionality, often seen in information retrieval and characterised

by some form of search capability. But it is making its way into data mining activities and it is here that more capable organizations will realize significant advantage.

Unstructured Meets Structured Data

Text data are typically held as notes, documents and various forms of electronic correspondence (emails for example). Structured data on the other hand are usually contained in databases with fixed structures. Many data mining techniques have been developed to extract useful patterns from structured data and this process is often enhanced by the addition of variables (called features) which add new 'dimensions', providing information that is not implicitly contained in existing features. The appropriate processing of text data can allow such new features to be added, improving the effectiveness of predictive models or providing new insights.



Incorporating features derived from customer notes, email exchanges and comments can improve lead targeting, flag possible defection and even contribute to the identification of fraud. The methods used to extract useful features from text depend on the domain, the nature of the business application and the characteristics of the text data. However a statistical approach based on a frequency matrix (a count of words appearing in various text sources) often yields useful new features after the application of appropriate statistical techniques. Other techniques might employ named entity extraction (NEE) where probabilities can be assigned to the likelihood that a particular document refers to a given entity (people, places, products, dates etc.).

A prerequisite for combining text and structured data is some form of integrated data environment, since the sources of a data could be highly diverse and volatile. While building predictive models can be facilitated using various interfaces to unstructured data, implementing the resulting models requires tight data integration and a scalable computing environment. This can be achieved through big data infrastructure such as that offered by Hadoop and associated technologies, although this is definitely not a trivial undertaking. The alternative is to embrace integrated infrastructure and tools provided by some of the larger suppliers in this domain.

Such is the complexity of integrating text analytics with structured data that most organizations will opt to buy solutions for their needs. There is no point reinventing the wheel here, and some advanced solutions are already emerging for customer and marketing applications where text data are incorporated into data

mining activities. Probabilistic latent semantic analysis and specifically Latent Dirichlet Allocation is a highly sophisticated technique used to associate documents with a topic. Specialist knowledge is needed to employ such techniques and many businesses will simply opt to buy capability rather than explore such highly complex statistical methods. The techniques used are just a small part of the story with infrastructure, skill sets, presentation methods, management and performance monitoring representing the larger piece of the cake.

The integration of text data sources with structured data will see significant progress over the next few years. Organizations that are willing to integrate the missing 80% of their data that text represents (missing from current analytical activities) will gain insights and operational improvements that would otherwise not have been possible.

Business Application

Text analytics has wide application in a business environment. For users wishing to reap productivity gains, text analytics can automatically filter spam, categorize messages (emails for example), label documents and enable fast and relevant information search. This can be viewed as the 'traditional' role for text analytics, although more contemporary applications include fraud detection, warranty claim processing, competitor analysis and social media monitoring. Beyond this text data are being used to add new features to data mining activities with the aim of creating predictive models which might be used to detect possible customer defection, new selling opportunities and delinquency. It can also be used to provide new insights into customer behavior through identifying archetypes.

The potential benefits of text analytics are not 'automatic'. Domain experts are needed to provide input to the analytics process and sanitize results. This applies to almost every application of the technology regardless of whether a 'plug and play' type solution is being used or a bespoke application has been built. Even web based services which provide sentiment analysis of social media require considerable amounts of configuration if results are to be meaningful.

The focus for text analytics solutions is primarily in the customer and marketing domains. Such solutions are often cloud based, but for larger organizations a in-house deployment might be necessary because of latency and security issues. Either way text analytics provides insights into customer behavior that are not accessible through analysis of the structured data contained in databases. This extra dimension can be used to tailor a more relevant interaction with customers and predict future behavior. For example it may be possible to identify which customers are perfectly good credit risks, but sometimes make later payments because of lifestyle. It is much more likely that a customer 'sweet spot' can be identified through text analysis than by any other mechanism, since text contains significantly more potential information than the history, demographics and other data held in databases.

In a marketing environment, text data in the form of open surveys (where respondents can add free text comment), can be used to extract nuances which simply cannot be accommodated in a closed response form. This might enable sentiment scores to be created or the identification of terms and concepts that had not been anticipated. Obviously this is closely related to the sentiment analysis of social media, which at the

current time is over-hyped, but is quickly evolving to provide behavioral insights and trend analysis for marketing activities.

While customer and marketing applications might be the most obvious ones, text analytics applies to any domain where text data is acquired. In a manufacturing environment the notes made by maintenance staff might be used to improve the prediction of maintenance requirements and the avoidance of down time. In a medical environment text notes that are captured during the diagnostic process can provide valuable input to understanding patient concerns and the process of diagnosis itself.

Perhaps the most promising application of text analytics is the creation of new features for data mining processes. Combining structured and unstructured data in this way facilitates the meeting of two quite different information 'dimensions' and is already being used in sales, marketing and customer care applications.

As always business management will be tasked with the need to identify opportunities and decide how unique they want a solution to be. Packaged solutions potentially reduce risk, but also reduce opportunity. A bespoke solution introduces technical and business risk, but also provides the most opportunity. Fortunately a number of suppliers offer a middle way with many of the technical, architectural and business risks largely mitigated, but with an opportunity to deliver a tight fit with individual business needs.

Strategy

Strategy is always the meeting point of need and capability. The starting point in a text analytics strategy is the identification of business processes where text analytics might deliver benefit, and an awareness of what is possible. Business processes that are candidates obviously need access to relevant text data. At the current time this usually applies to customer and marketing applications, although as text analytics becomes more prevalent so businesses will collect more text data to enable the analytics process. As far as capability is concerned it is usually a given that provided good quality text data is available, so value will be created through the analytics process, although this is nearly always an iterative process.

Once candidate business processes have been identified it becomes a matter of fleshing out some form of cost/benefit analysis. With analytics technologies it becomes much more difficult to estimate benefits unless data are available from other organizations, or suppliers who have experience in the domain. Typically, an increase in lift (value created with new information divided by value created without information, multiplied by 100) of just a few percent will more than adequately reward many text and data mining activities. The cost of developing and deploying text analytics applications depends very much on the route taken. Text analytics may be part of a much larger 'big data' project, in which case it becomes more difficult to allocate costs. Costing discrete projects is usually much easier.

This does not mean however that determining costs will be necessarily straightforward. Unlike traditional process automation projects (e.g. ERP, CRM) where deployment is essentially linear, analytics projects are usually iterative in nature. Again it is very useful to have access to people who have knowledge of building and deploying analytics processes, although the nature of text data will be particular to each organization

and this inevitably introduces some variability. The cost components will include hardware and software (unless a cloud service is used), skill sets, domain expert involvement, performance management and monitoring and possibly the acquisition of new data. This latter point is more important than is immediately obvious. Sourcing external data (social data, market data etc.) is an obvious cost, but it is more than likely that the results of analysis will imply that greater data capture when dealing with customers for example might deliver more accurate insights and predictive models. There is a cost associated with this and it needs to be taken into account.

Finally there is a cost associated with the management and monitoring of processes involving analytics. The insights and models derived from analysis usually decay with time, simply because markets and customers change. Monitoring performance and feeding this back into the analytics process is not a trivial matter and will impose its own overhead.

While we have been conservative by suggesting just a few percent increase in lift, it does happen that the benefits can be considerably greater than this. A more useful model for modeling the return from an investment is Expected Return. This allows for multiple scenarios and will generate an expected return from an investment. While this is not widely used at the current time other than in some specific industries (petrochemicals for example), it does give a good feel for risk and is more appropriate for analytics projects where there are more unknowns.

Analytics projects do need a somewhat different approach to risk management than traditional IT systems. It really is not enough to develop a model and leave business users to get on with it, the whole process needs much finer integration between business, IT and data analysts.

Text Analytics Platforms

AlchemyAPI

AlchemyAPI provides cloud based text analytics services to support sentiment analysis, marketing, content discovery, business intelligence, and most tasks where natural language processing is needed. An on-site capability can also be provided if needed.

The capabilities offered by AlchemyAPI go beyond those most large organizations could build in-house, and not least because the training set used to model language is 250 times larger than Wikipedia. Innovative techniques using deep learning technologies (multi-layered neural networks) also go well beyond most of the competition, and AlchemyAPI distinguishes itself by using the technology for image recognition in addition to text analytics.

The functionality is broad and includes:

- Named entity recognition for identifying people, places, companies, products and other named items.
- Sentiment analysis with sophisticated capabilities such as negation recognition, modifiers, document level, entity, keyword and quotation level sentiment.
- Keyword extraction to identify content topics.
- Concept tagging, which is capable of identifying concepts not explicitly mentioned in a document.
- Relation extraction where sentences are parsed into subject, action and object.
- Text categorization to identify most likely categories.
- Other functionality such as author extraction, language detection and text extraction.

Alchemy API was founded in 2005 and is based in Denver Colorado. Pricing plans for the cloud based services are based on transaction per day and start with a free Starter subscription.

Angoss KnowledgeREADER

KnowledgeREADER from Angoss is part of a broad suite of analytics tools and specifically addresses text analytics in the context of customer oriented and marketing applications. It majors on visual representation including dashboards for sentiment and text analysis, and also provides a somewhat unique map of the results of association mining to display words that tend to occur together.

Many of the advanced features make use of the embedded Lexalytics text analytics engine - widely recognized as one of the best in class. Entity, theme and topic extraction are supported along with decision and strategy trees for profiling, segmentation and predictive modelling. Sentiment analysis supports the visual graphing of sentiment trends and individual documents can be marked up for sentiment.

Angoss provides its technology through the cloud or on-site implementation. High levels of end-user

functionality are claimed with much of the functionality available to business users. More advanced analysis can be achieved by combining text with structured data and text can be used to generate additional features for data mining activities.

Obviously this is a sophisticated product best suited to the needs of large organizations in the main, although the cloud based access will suit the needs of some mid-sized organizations too. Overall this is well suited to customer and marketing text analytics needs where text is used to gain insight into sentiment and customer behaviour.

Attensity

Attensity majors on social analytics, but also offers a general purpose text analytics engine. Four major components define the offering:

- Attensity Pipeline collects data from over one hundred million social sources as input for analysis.
- Attensity Respond provides a mechanism for responding to social comment.
- Attensity Analyze allows text in emails, call-center notes, surveys and other sources of text to be analyzed for sentiment and trend.
- Attensity Text Analytics provides an underlying engine that embraces several unique NLP technologies and a semantic annotation server for auto-classification, entity extraction and exhaustive extraction. It comes with good integration tools too so that the results of text analytics can be merged with structured data analytics.

Three horizontal solutions are offered for marketing, customer service and IT.

Basis Technology

Basis Technology delivers a variety of products and services based on multilingual text analytics and digital forensics. The Rosette platform provides morphological analysis, entity extraction, name matching and name translation in fields such as information retrieval, government intelligence, e-discovery and financial compliance.

The Rosette search and text analytics technology comes in five distinct functional units:

- RLI - Rosette Language Identifier - automatic language and character encoding identification.
- RBL - Rosette Base Linguistics - many search engines have used RBL to provide essential linguistic services such as tokenization, lemmatization, decompounding, part-of-speech tagging, sentence boundary detection, and noun phrase detection. Currently supports 40 languages.
- REX - Rosette Entity Extractor - finds entities such as names, places, organizations and dates.
- RNI - Rosette Name Indexer - matches the names of people, places and organizations written in different languages against a single, universal index.

- RNT - Rosette Name Translator - provides multilingual name translation through a combination of dictionaries, linguistic algorithms and statistical inference.

A Rosette plug-in is available for Lucene and Solr search technologies and Basis Technology provides solutions for government, social media monitoring, financial compliance, e-discovery and enterprise search.

Clarabridge

Clarabridge provides a text analytics solution with a customer experience focus. This embraces various sources of customer information including surveys, emails, social media and the call centre.

The technology addresses three essential steps in the analysis of textual information. It supports the aggregation of information from most sources imaginable, allows the information to be processed for linguistic content and the creation of categories, and finally provides a rich user interface so the results of analysis can be seen. There are three main areas of functionality:

- Clarabridge Analyze comes with the ability to tune classification models and the way sentiment is scored, and provides various reports and visualizations.
- Clarabridge Act provides a customer engagement environment for all customer facing employees by providing real-time dashboards and the mechanisms to address customer feedback.
- Clarabridge Intelligence Platform carries out analysis and is essentially a natural language processing (NLP) engine. Connections to other applications in the organization can be facilitated by Clarabridge Connect, and includes out-of-the-box connectors for salesforce, Radian 6, Lithium and other applications.

Mobile workers are well catered for by Clarabridge Go - a mobile application providing various reports and visuals. A variety of horizontal (product management, customer care, operations management, sales and marketing, human resources) and vertical solutions are also available.

Clustify

Clustify, used mainly by legal firms, groups related documents into clusters, providing an overview of the document set and aiding with categorization. This is done without preconceptions about keywords or taxonomies — the software analyzes the text and identifies the structure that arises naturally. Clustify can cluster millions of documents on a desktop computer in less than an hour, bringing organization to large projects.

Clustify identifies important keywords used for clustering and reports frequency information so that clusters can be browsed which contain a set of specified keywords. It also identifies a representative document for each cluster, allowing decisions to be made on other documents in the same cluster.

Uses of Clustify include taxonomy development, search engine enhancement, litigation and ad targeting. The technology is built on proprietary mathematical models which measure the similarity of documents.

Connexor

Connexor provides a suite of text analytics tools which embrace a wide variety of NLP methods. These include metadata discovery, name recognition, sentiment detection, language identification, automatic document summarization, document classification, text cleansing, language analysis (10 European languages) and machine translation.

Connexor's Machinese libraries transform text into linguistically analyzed structured data. This includes Machinese Phrase Tagger which splits text into word units, Machinese Syntax which shows the relationship between words and concepts and Machinese Metadata which will extract information in 10 languages.

Solutions are offered for organizations operating in defence and security, life sciences and media, and Connexor works with a wide variety of organizations (software houses, businesses, systems integrators etc.) to deliver NLP capability.

DatumBox

DatumBox provides a cloud based machine learning platform with 14 separate areas of functionality, much of which is relevant to text analytics. The various functions are called via a REST API and address the following types of application:

- Sentiment Analysis - classifies documents as positive, negative or neutral.
- Twitter Sentiment Analysis - specifically targeted at Twitter data.
- Subjectivity Analysis - classifies documents as subjective (personal opinions) or objective.
- Topic Classification - documents assigned to 12 thematic categories.
- Spam Detection - documents labeled as spam or nospam.
- Adult Content Detection.
- Readability Assessment - based on terms and idioms.
- Langauge Detection.
- Commercial Detection - commercial or non-commercial based on keywords and expressions.
- Educational Detection - based on context.
- Gender Detection - written by or targeting men/women based on words and idioms.
- Keyword Extraction.
- Text Extraction - extraction of important information from a web page.
- Document Similarity - to detect web page duplicates and plagiarism.

Eaagle

Eaagle provides text mining technology to marketing and research professionals. Data is loaded into Eaagle and a variety of reports and charts are returned showing relevant topics and words, word clouds, and other statistics. Both online and Windows based software is offered. The Windows offering is called Full Text Mapper with good visuals to explore topics and various word statistics.

ExpertSystem

ExpertSystem majors on semantic analysis, employing a semantic analysis engine and complete semantic network for a complete understanding of text, finding hidden relationships, trends and events, and transforming unstructured information into structured data. Its Cogito semantic technology offers a complete set of features including: semantic search and natural language search, text analytics, development and management of taxonomies and ontologies, automatic categorization, extraction of data and metadata, and natural language processing.

At the heart of Cogito is the Sensigrafo, a rich and comprehensive semantic network, which enables the disambiguation of terms, a major stumbling block in many text analytics technologies. Sensigrafo allows Cogito to understand the meaning of words and context (Jaguar: car or animal?; apple: the fruit or the company?) - a critical differentiator between semantic technology and traditional keyword and statistics based approaches.

Sensigrafo is available in different languages and contains more than 1 million concepts, more than 4 million relationships for the English language alone, and a rich set of attributes for each concept. The Cogito semantic network includes common words, which comprise 90% of all content, and rich vertical domain dictionaries including Corporate & Homeland Security, Finance, Media & Publishing, Oil & Gas, Life Sciences & Pharma, Government and Telecommunications, providing rich contextual understanding that improves precision and recall in the process information retrieval and management.

The technology has found uses in CRM applications, product development, competitive intelligence, marketing and many activities where knowledge sharing is critical.

FICO

FICO provides sophisticated text analytics capability in its analytics tools and in the form of specific business solutions. At the heart of the offering is its Data Management Integration Platform (DMIP) addressing the complex issues associated with accessing diverse data sources and supporting a variety of analytics tools. Linguistic analysis supports tagging, dependency analysis, named entity extraction and intention analysis. Model Builder supports most forms of text analysis with parsing, indexing, stop words, n-grams, stemming, 'bag of words' analysis and so on. Some particularly sophisticated text analytics solutions are offered for fraud detection, employing Latent Dirichlet Allocation as a method of categorizing customers. In its traditional banking, insurance and financial services markets FICO utilizes text analytics to provide additional features in its scorecard technology.

A cloud based solution text analytics solution will soon be available. While Model Builder is a large

sophisticated product, the cloud based offering will provide a much easier user interface when it is launched later in the year.

IBM

IBM provides text analytics support through two products. IBM Content Analytics is primarily an extension of enterprise search technologies that adds several useful visualizations to discover structure within text data. LangaugeWare on the other hand leverages natural language processing (NLP) to facilitate several types of text analysis.

A major component within IBM Content Analytics is IBM Content Analytics with Enterprise Search. This supports the visualization of trends, patterns within text and relationships. Facets feature highly in the analysis. These are categories which are derived from text analysis. For example documents on infectious diseases might be categorized by a 'hepatitis' facet. The facet-pair view shows how facets (categories) are related to each other, and a dashboard facility allows several analyses to be viewed simultaneously. A connections view displays relationships between various facet values and a sentiment view allows the sentiment behind facets to be displayed. Other components in IBM Content Analytics are targeted at specific applications including healthcare and fraud. Content Classification supports the organization of unstructured content.

LanguageWare uses NLP techniques at the document level. This includes entity and concept recognition, knowledge/information extraction and textual relationship discovery.

As always with IBM these capabilities are offered within the context of supporting infrastructure and services and will primarily be of interest to larger organizations. There is nothing particularly interesting here, and it is likely that less costly and more capable solutions will be available for many text analytics needs.

Intellexer

Intellexer provides a family of tools for natural language search, document management, document comparison and the summarization and analysis of documents and web content. Nine solutions are offered, all reasonably priced:

- Name recognition - extracts names (named entities) and defines relations between them.
- Summarizer - extracts main ideas in a document and creates a short summary.
- Categorizer - for automatic document categorization.
- Comparator - compares documents and determines the degree of proximity between them.
- Question-answering - looks for documents which answer a natural language query.
- Natural language interface - generates Boolean queries for any application.
- Related Facts - is an IE plugin for Google search and selects 5 main topics and supplements them

with related facts.

- Summarizer plug-in for IE - summarizes web pages and extracts concepts.
- PDF Converter - to incorporate PDF documents into text processing.

KBSPortal

KBSPortal provides an NLP capability which includes tagging and categorizing user submitted web site content, text summarization, document linking by entities, vulgarity detection, sentiment rating and association of sentiment with products and people. This functionality is available as a web service or through purchase of source code for in-house deployment.

Lexalytics

Lexalytics is one of the forerunners in text analytics and its Salience text analytics engine is used in market research, social media monitoring, survey analysis/voice of customer, enterprise search and public policy applications. The functionality offered by Salinece includes sentiment analysis, named entity extraction, theme extraction, entity-level sentiment analysis, summarization and facet and attribute extraction. The Salience engine can be integrated into other business applications via a flexible set of APIs, and can be tuned for very specific tasks and high levels of performance.

Another essential component in the Lexalytics approach is data directories. This effectively provides a parameter driven environment with files to set up relationship patterns, sentiment analysis, and the creation of themes. Non-English support is provided through this mechanism. Each directory can be configured to support a particular task delivering considerable flexibility and power.

Leximancer

Leximancer uses 'concepts' as a primary analytic structure, and these are automatically identified by the software without need for existing structures such as taxonomies or ontologies. Analysis is presented through a variety of useful visualizations, with drilling down to individual documents. It is used in survey analysis, market research, social media monitoring, customer loyalty and forensic analysis.

Leximancer Enterprise runs on a multi-user server providing users with a browser interface, and also provides a REST web services interface for application integration. A desktop version is available as a stand-alone environment, or users can access the LexiPortal via a web browser for a web based service (charging based on usage). Moderately priced academic versions are also available.

Linguamatics

Linguamatics provides a NLP capability with either in-house or cloud based implementation. A search engine approach to mining text comes with a good query interface and the ability to drill down to individual documents. A domain knowledge plug-in supports taxonomies, thesauri and ontologies.

The technology is widely used in life sciences and healthcare and the on-line service provides access to content in this domain. A web services API supports most programming languages.

Linguasys

Linguasys primarily satisfies the need to process text in multiple languages – and by multiple we mean English, Arabic, Chinese, German, French, Hebrew, Indonesian, Japanese, Korean, Malay, Spanish, Pashto, Persian, Portuguese, Russian, Thai, Vietnamese, Urdu and others under development. This may well be unique in the world of natural language processing, and is possible because all languages are transformed into a large collection of concepts, each with its own identifier. It is the concepts which link all the languages together. The concept 'mobile phone' for example has the same concept number in all languages and is given identifier 26300, along with all variants that mean the same thing – 'cellular phone' for example.

Luminoso

Luminoso is a cloud based text analytics service that calls upon a multi-lingual capability. Many of the current problems associated with text analytics (ambiguity for example) are at least partly solved by Luminoso. A variety of useful reports and visualizations provide users with a particularly good interface.

Megaputer

PolyAnalyst from Megaputer is a data and text mining platform which embraces the complete analytics lifecycle. Megaputer provides two separate software packages for text analysis. PolyAnalyst performs linguistic and semantic text analysis and coding, clustering and categorization of documents, entity extraction, visualization of patterns, automated or manual taxonomy creation, text OLAP, and generating interactive graphical reports on results. TextAnalyst provides a list of the most important keywords in a document, a set of related keywords for each word, and the ability to automatically summarize a document or perform natural language queries.

NetOwl

NetOwl provides both text and entity analytics in the cloud and in private deployments. Text analytics includes Extractor to perform entity extraction, DocMatcher which compares and categorizes documents according to user defined concepts, and TextMiner for mining large amounts of text. Entity analytics is used to accurately match and identify names - important in many areas, including CRM, anti-fraud and national security. This includes NameMatcher to identify name variants from large multicultural and multilingual name databases. EntityMatcher performs identity resolution on similar databases.

Provalis Research

Provalis provides a suite of text analytics tools, each of which facilitates a particular type of text analysis. QDA Miner (available in a free Lite version) supports qualitative analysis with coding, annotation, retrieval and analysis of document and image collections. WordStat on the other hand supports the extraction of themes and trends, taxonomy and ontology creation, clustering and proximity analysis, and machine learning tools for document classification. SimStat, as the name suggests provides statistical analysis tools for text analysis. These three components can be purchased separately or as ProSuite, and all components are integrated with each other.

Rocket AreoText

AeroText is a text extraction and text mining solution that derives meaning from content contained within unstructured text documents. AeroText is capable of discovering entities (people, products, dates, places, products) and the relationships between them, as well as event discovery (contract data, PO information etc.) and subject-matter determination. AeroText is also capable of resolving ambiguities, such as relative time references, 'one and the same' matches and semantic analysis, based on context at the document, paragraph or sentence-level.

SAS Text Analytics

SAS Text Analytics is part of the very broad analytics capability offered by SAS. Several modules are provided including:

- SAS Contextual Analysis - for the creation of document classification models.
- SAS Enterprise Content Categorization - for automated content categorization, and various add-on modules add extra capability as needed.
- SAS Ontology Management - to define semantic relationships.
- SAS Sentiment Analysis
- SAS Text Miner - use of various supervised and unsupervised techniques.

Statistica Text Miner

Statistica Text Miner is part of the extensive Statistica statistical analysis and data mining product set. Extensive pre-processing options are available with stemming and stub lists for most European languages. 'Bag of words' type analysis can be carried out with input to the data mining capabilities of Statistica.

Qualitative Data Analysis Tools

Qualitative data analysis (QDA) is a process of exploring, marking up, coding and annotating documents to extract concepts, links and other structures.

ATLAS.ti is one of the most powerful tools for qualitative research. Managed documents, multi-document view, high-performance multimedia engine, intuitive margin-area coding for all data types. ATLAS.ti offers state-of-the art multimedia processing. Frame-level and wave previews make coding audio and video material a joy; images can be fully coded in every detail and segments can even be moved and re-sized. The multi-document view makes it easy to link sections across various documents. Cloud views provide very quick, accurate, and yet intuitive analytical access to your data material. The query tool, co-occurence explorer and the codes-PD-table allow in-depth analysis.

Dedoose is a cross-platform web app for analyzing qualitative and mixed methods research with text, photos, audio, videos, and spreadsheet data.

f4analyse supports you in analyzing your textual data. You can develop codes, write memos, code (manually or automatically), and you can analyze cases and topics. The program is slim & easy to learn.

HyperRESEARCH is designed to assist you with any research project involving analysis of qualitative data. It's easy to use and works with both Mac and Windows computers. So if you're collaborating with multiple researchers, everyone gets to use their preferred computer. HyperRESEARCH is powerful, and flexible - which means that no matter how you want to approach your data, the software will allow you to "do it your way." HyperRESEARCH can help you analyze almost any kind of qualitative data, whether it's audio, video, graphical or textual. The intuitive interface and well-written documentation – and especially the step-by-step tutorials -- help get you up and running with your own data quickly and easily.

MAXQDA for Windows and Mac is a professional software for qualitative and mixed methods data analysis. MAXQDA is not limited to one kind of research approach or method. Organize, evaluate, code, annotate and interpret all kinds of data, create easy-to-read reports and visualizations, and connect and share with other researchers. Analyze interviews, reports, tables, online surveys, videos, audio files, images, and even bibliographical data sets with MAXQDA. Organize and categorize your data, retrieve your results and create impressive illustrations and reports. MAXQDA has transcription tools onboard and multimedia functionality to directly analyze all kinds of media files. Outstanding mixed methods features allow for the import of quantitative data alongside the qualitative information and results can be transformed into variables for further statistical analysis.

Government agencies use NVivo to deliver evidence-based findings and shape policy. Businesses use NVivo in pilot studies, program evaluation and to inform decision-making. Academics use NVivo to produce rigorous research. NVivo enables users to collect, organize and analyze content from interviews, focus group discussions, surveys, audio, social media, videos and webpages.

QDA Miner is an easy-to-use qualitative data analysis software package for coding, annotating, retrieving

and analyzing small and large collections of documents and images. QDA Miner qualitative data analysis tool may be used to analyze interview or focus group transcripts, legal documents, journal articles, speeches, entire books, as well as drawings, photographs, paintings, and other types of visual documents.

Quigga is a freemium QDA tool with multiple features, including - import PDFs into separate libraries. Automatic OCR and tag extraction. Populate missing metadata. Full-text search, duplicate paper detection, inbound and outbound links. Built-in PDF reader with annotating, highlighting, automated jump links, and so much more. Create printable summaries of notes, mindmaps of your thoughts, and directly cite your references within Microsoft Word™. Optionally sync to the private cloud with unlimited storage. Share library documents, metadata, and notes in private with selected friends or colleagues.

webQDA is software that supports the analysis of qualitative data in a collaborative and distributed environment. webQDA is focused on researchers who work in multiple contexts and need to analyse qualitative data as an individual or in group in a synchronous or asynchronous way. It offers online and real time collaborative work as well provides a service to support scientific research. This software is optimized to Internet Explorer, Firefox, Chrome, Opera and Safari browsers, and with Windows, Mac OS and Linux operating systems.

XSight is a feature rich QDA tool, the main features of which include: Customize the user friendly interface to suit your working style. Capture your ideas visually with 'maps', just like you would on a flip chart or whiteboard. Query your data with our powerful, state-of-the-art search engine. Use 'tags' - XSight's answer to highlighter pens - to capture and highlight information. Take a fresh look at your project - zoom in, zoom out or drill down. Build reports and presentations with ease. Work in virtually every language.

Free Qualitative Data Analysis Tools

AQUAD 7 is open-source freeware (according to the conditions of GNU GPL v.3). A full feature list of AQUAD 7 can be found on the informationen and features page. The recent version 7.3 allows to analyze all kinds of qualitative data: within the framework of the code-paradigm, along Boolean minimization to identify types and by means of textual sequence-analysis to reconstruct case structures based on strict hypothesis testing (following the approach of objective hermeneutics). An interface with the statistical software “R” (open source) allows to combine qualitative and quantitative analyses; the scripts were modified and more scripts were added to version 7.3. AQUAD supports the following data types:

- texts of any kind (e.g. transcripts of social interactions, letters, documents, …)
- audio-data (e.g. interview recordings)
- video-data (e.g. observations)
- pictures (e.g. photos, drawings)

The Coding Analysis Toolkit (or "CAT") consists of a ASP.NET based suite of tools to facilitate efficient and effective analysis of text datasets that have been coded using the CAT coding module or ATLAS.ti. It's functionality includes:

- Efficiently code raw text data sets
- Annotate coding with shared memos
- Manage team coding permissions via the Web
- Create unlimited collaborator sub-accounts
- Assign multiple coders to specific tasks
- Easily measure inter-rater reliability
- Adjudicate valid & invalid coder decisions
- Report validity by dataset, code or coder
- Export coding in RTF, CSV or XML format
- Archive or share completed projects
- Data can be Plain text, HTML, CAT XML, Merged ATLAS.ti coding

Cassandre is a free open source text analysis software tool. It uses semi-automatic coding, based on the identification of markers, grouped into registers, which represent analysis categories.

CATMA is a practical and intuitive tool for literary scholars, students and other parties with an interest in text analysis and literary research. Being implemented as a web application in the newest version, CATMA

also facilitates the exchange of analytical results via the internet. Features include:

- Freely producible Tagsets, suitable to apply analytical categories of individual choice to the text
- The possibility of advanced search in the text, using the natural language based Query Builder
- A set of predefined statistical and non-statistical analytical functions
- The visualization of the distribution of items of interest (e.g. words, word-groups or Tags) in the text
- The possibility to analyze whole corpora of texts in one work step
- Easy switching between the different modules
- The easy exchange of documents, Tagsets and Markup information, facilitating cooperative textual analysis
- A context sensitive help function and a user manual for better usability

FreeQDA is a software for professional qualitative research data analysis, such as interviews, manuscripts, journal articles, memos and field notes. FreeQDA requires Java >= 1.6

QCAmap is an open access web application for systematic text analysis in scientific projects based on the techniques of qualitative content analysis. It can be used within research projects in e.g. Psychology, Sociology, Education, Economics, Linguistic Sciences, to analyze small and large amounts of any text material coming from interviews, group discussions, observation protocols, documents, open-ended questionnaire items and others. Qualitative Content Analysis is a strictly rule-guided procedure containing qualitative steps (assignment of categories to text passages) and quantitative steps (analysis of category frequencies).

QDA Miner Lite is a free and easy-to-use version of the Provalis QDA Miner. It can be used for the analysis of textual data such as interview and news transcripts, open-ended responses, etc. as well as for the analysis of still images. It offers basic CAQDAS features such as:

- Importation of documents from plain text, RTF, HTML, PDF as well as data stored in Excel, MS Access, CSV, tab delimited text files,
- Importation from other qualitative coding software such as Altas.ti, HyperResearch, Etnograph, from transcription tools like Transana and Transcriber as well as from Reference Information System (.RIS) files.
- Intuitive coding using codes organized in a tree structure.
- Ability to add comments (or memos) to coded segments, cases or the whole project.
- Fast Boolean text search tool for retrieving and coding text segments.
- Code frequency analysis with bar chart, pie chart and tag clouds.

- Coding retrieval with Boolean (and, or , not) and proximity operators (includes, enclosed, near, before, after). Export tables to XLS, Tab Delimited, CSV formats, and Word format Export graphs to BMP, PNG, JPEG, WMF formats.
- Single-file (*.qdp) project format.
- Interface and help file in English, French and Spanish.

RDQA is a R package for Qualitative Data Analysis, a free (free as freedom) qualitative analysis software application (BSD license). It works on Windows, Linux/FreeBSD and (probably) the Mac OSX platforms. RQDA is an easy to use tool to assist in the analysis of textual data. At the moment it only supports plain text formatted data. All the information is stored in a SQLite database via the R package of RSQLite.

TAMS stands for Text Analysis Markup System. It is a convention for identifying themes in texts (web pages, interviews, field notes). It was designed for use in ethnographic and discourse research.

TAMS Analyzer is a program that works with TAMS to let you assign ethnographic codes to passages of a text just by selecting the relevant text and double clicking the name of the code on a list. It then allows you to extract, analyze, and save coded information. TAMS Analyzer is open source; it is released under GPL v2. The Macintosh version of the program also includes full support for transcription (back space, insert time code, jump to time code, etc.) when working with data on sound files.

Weft QDA is an easy-to-use, free and open-source tool for the analysis of textual data such as interview transcripts, fieldnotes and other documents. Features include:

- Import plain-text documents from text files or PDF
- Character-level coding using categories organised in a tree structure
- Retrieval of coded text and ‘coding-on’
- Simple coding statistics
- Fast free-text search
- Combine coding and searches using boolean queries AND, OR, AND NOT
- ‘Code Review’ to evaluate coding patterns across multiple documents
- Export to HTML and CSV formats

Open Source and Free Enterprise Search Platforms

Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Includes rank searching, powerful query types, fielded searching, multiple index search, flexible faceting etc

Constellio is the first open source comprehensive suite of enterprise search. It is the result of more than 2 years of research and development and is based on best practices and standards of market research information. Based on the popular search engine Apache Solr and using the architecture of connectors of Google Search Appliance, Constellio provides the solution to index all sources of information in your business. Constellio is compatible with all connectors from Google Search Appliance and can import any index from Solr and Lucene.

Elasticsearch is a flexible and powerful open source, distributed, real-time search and analytics engine. Architected from the ground up for use in distributed environments where reliability and scalability are must haves, Elasticsearch gives you the ability to move easily beyond simple full-text search. Through its robust set of APIs and query DSLs, plus clients for the most popular programming languages, Elasticsearch delivers on the near limitless promises of search technology.

Searchdaimon ES is based on ten years of research and development. ES is one of the most scalable, fastest and most accurate solutions for search today. Retrieval of information, analysis and storage constitutes the three most important parts of the system. With a highly advanced system for distributed information retrieval and analysis, combined with large capacity storage handling, we are ready to solve challenges for enterprises of all sizes.

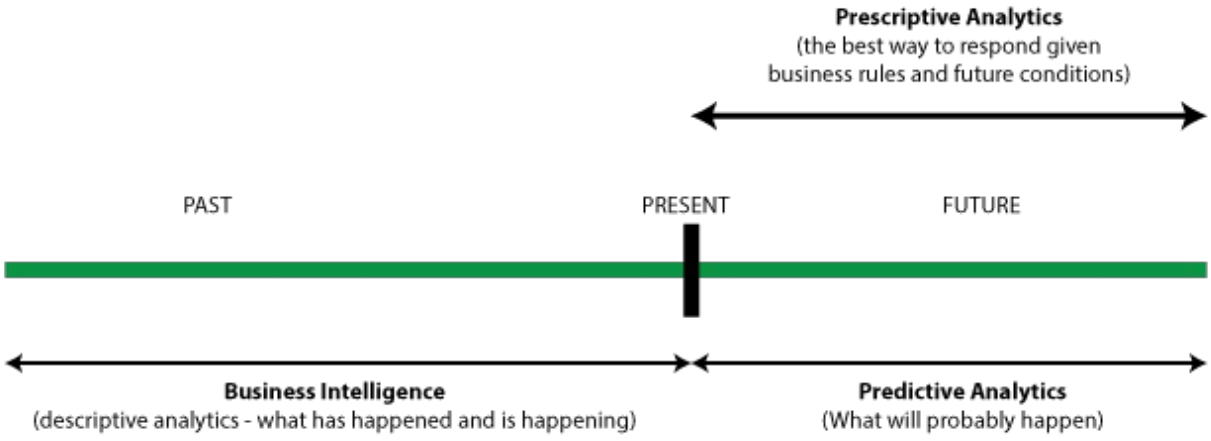
Solr is the popular, blazing fast open source enterprise search platform from the Apache LuceneTMproject. Its major features include powerful full-text search, hit highlighting, faceted search, near real-time indexing, dynamic clustering, database integration, rich document (e.g., Word, PDF) handling, and geospatial search. Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites.

Sphinx is an open source full text search server, designed from the ground up with performance, relevance (aka search quality), and integration simplicity in mind. It's written in C++ and works on Linux (RedHat, Ubuntu, etc), Windows, MacOS, Solaris, FreeBSD, and a few other systems. Sphinx lets you either batch index and search data stored in an SQL database, NoSQL storage, or just files quickly and easily — or index and search data on the fly, working with Sphinx pretty much as with a database server.

Prescriptive Analytics

The Business Value of Prescriptive Analytics

After fifty years of using information technology to increase the efficiency of business processes we are now firmly in the era where technology is also being used to provide us with information. Business intelligence allows us to establish what has happened and is happening in our business (often called descriptive analytics), and predictive analytics uncover patterns which can be useful in the prediction of future events. This doesn't complete the picture however. Descriptive and predictive analytics may tell us what has happened and what may happen, but they do not tell us the best way to deploy our resources to meet the demands of the future. An example will clarify. In a retail environment our descriptive analytics will tell us sales volumes, seasonal fluctuations and so on. Predictive analytics may give us insights into which products tend to be purchased together. Armed with this knowledge we then need to know how shelf space should best be allocated and more generally how resources should be utilized to maximize revenue and/or profitability. This is where prescriptive analytics fits in - think of it as a prescription for action.



The major part of prescriptive analytics is concerned with resource optimization given a set of business rules (constraints) and predictions relating to demand, customer behavior, the success of marketing campaigns and so on. In real business problems, optimization may involve thousands of variables and constraints, and finding the optimal use of resources, given an objective that is to be maximized or minimized, can only be achieved using powerful computerized optimization software. Examples abound. Airlines use prescriptive analytics to determine the allocation of seats to each particular class. Vehicle rental businesses optimize the positioning of vehicles to maximize revenue and profitability. Energy companies increasingly use prescriptive analytics and especially with the unpredictable nature of renewable energy sources.

Of course this all assumes that business managers buy into the resource utilization schedules created by prescriptive analytics techniques. As such the analytics initiative needs high level sponsorship and coordinated effort throughout the enterprise. Reporting mechanisms need to be put in place and procedures to deal with the inevitable changes of circumstances all businesses experience. To this end some businesses run some of their prescriptive analytics processes in near real-time to accommodate change, and such is the power of the optimization algorithms and computer hardware that this has become possible for complex analytics tasks.

Prescriptive analytics is clearly not a trivial undertaking. It needs close liaison between analytics teams and business management, and an integrated analytics environment capable of integrating business rules, predictive models and prescriptive analytics. The integration is important, and particularly in large complex businesses. Without such integration prescriptive analytics may be very difficult to achieve, if not impossible.

Expect to see prescriptive analytics technologies more widely used as the user interfaces become more user friendly, and business managers become empowered to address increasingly complex optimization problems without recourse to teams of analysts. However for large, complex prescriptive analytics tasks the analytics teams are here to stay.

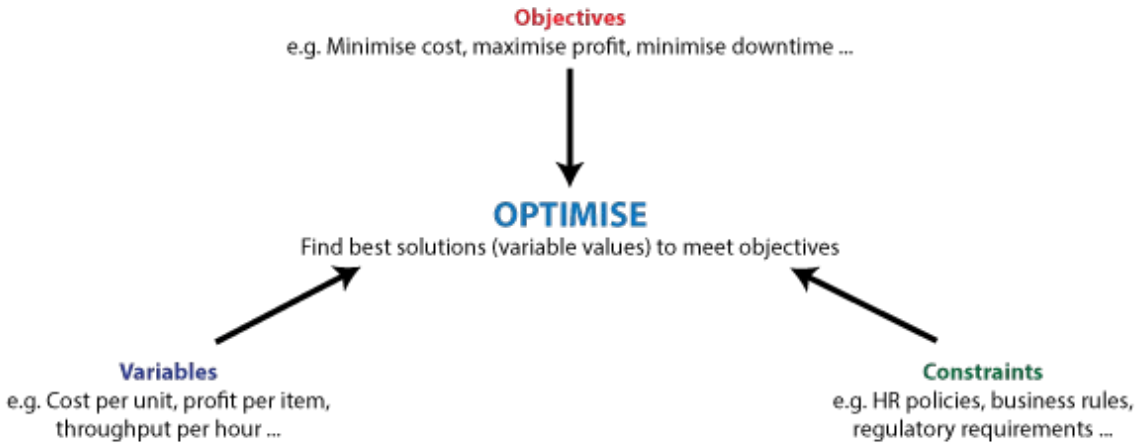
While most analytics technologies are concerned with **what** has happened or will happen, prescriptive analytics tells **how** to best deploy resources to optimize our operational activities - and the benefits are often substantial.

What is Prescriptive Analytics?

Optimization sits at the heart of prescriptive analytics technologies, and specifically the computation of best resource usage given a set of constraints and objectives. Work planning problems represent a classic application, where work is allocated to limited human resources in a manner that meets constraints and optimizes objectives.

While optimization has been used for decades in many large corporations, the compute intensive processing has traditionally been associated with very long compute times - typically days and weeks. This limited the application of the technology. However advances made in the mathematical algorithms and more powerful hardware mean that optimization can be applied to a much broader range of problems, and in some instance execute on a near real-time basis.

The three essential components in an optimization problem are variables, constraints and objectives. In a work planning problem the variables would typically represent the number of hours work allocated to various people from a given list of tasks. The constraints would limit the way the allocation of resources could take place - no more than 20% of the personnel from any department can be engaged on a project for example. Finally the objectives state what we are trying to achieve. Often this is simply to minimize costs, or maximize profits - or both. However in the work planning problem we might be most interested in minimizing the time a project takes. Each optimization problem has its own set of variables, constraints and objectives and much of the work goes into specifying what these are.



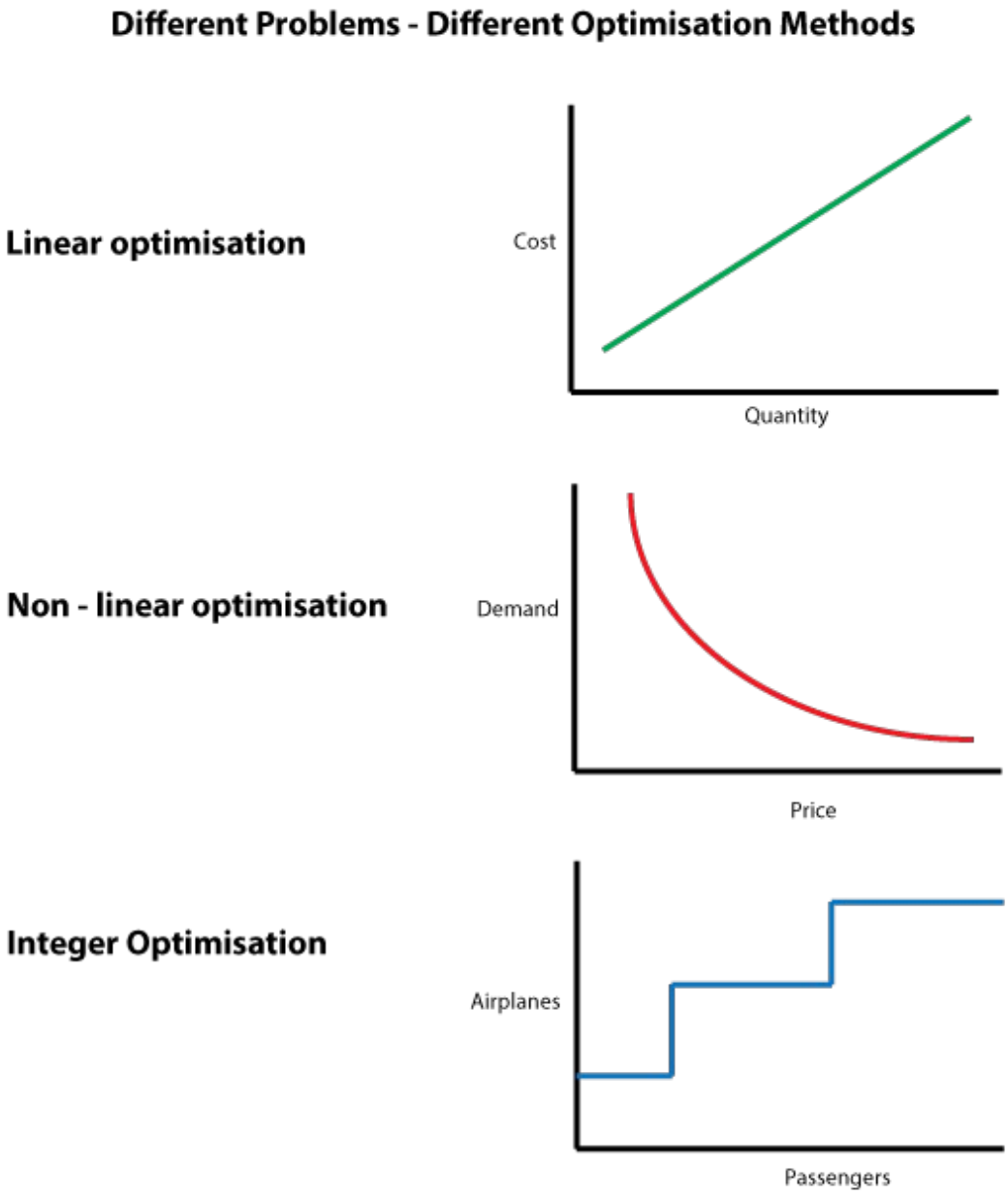
Prescriptive analytics can be divided into two primary activities. The first involves optimization when the input variables are known (a stock count, or balances in accounts for example). The problem here is simply to establish the best outcome given these variables along with associated constraints and given objectives. A second set of optimization problems comes under the heading of stochastic optimization, a suitably off-putting name which simply indicates there is uncertainty in the input data - next month's sales for example. This more complex category of problems will attempt to find the best solution to a business optimization problem for the most likely future situations. Obviously there is a strong link here with statistical modelling and other forms of predictive analytics, where probabilities are assigned to variables.

It is increasingly the case that prescriptive analytics is integrated with other systems. optimization has traditionally been an isolated activity, but today it can take inputs from business rules and predictive analytics processing, and benefits hugely from them. The business rules act as constraints (do not mail someone with an offer of a 5% discount when they have already been mailed a 10% discount - for example), and predictive analytics can provide inputs which predict variable values (the number of prospects likely to respond to a marketing campaign for example).

Prescriptive analytics is still relatively new (the term was first introduced about a decade ago) and only a handful of suppliers provide the integrated environment necessary to take advantage of outputs from other processes. However prescriptive analytics does complete the analytics picture - descriptive analytics (business intelligence) and predictive analytics say **what** has happened or will happen, while prescriptive analytics say **how** things should happen.

Prescriptive Analytics Methods

Optimizing complex business problems requires sophisticated technology. Recent years have witnessed major advances in the speed of optimization algorithms and in the complexity of problem that can be addressed. The net result is the proliferating use of optimization technologies to address everything from marketing campaign optimization to how many business class seats should be allocated on individual flights.



There are several well defined types of problem that optimization techniques can address - and some they

can't. The earliest and often easiest form of optimization assumed that variables and objectives were related to each other in a linear manner. If a resource usage is doubled, so is its cost. While there are some problems that are well served by this model (the use of material components in a mix for example), many are not. To cater for more complex optimization problems, non-linear relationships have been accommodated. A good example here is a price/demand curve where demand drops off rapidly as price exceeds a certain threshold, and increases exponentially as price drops below a critical level. The solution of non-linear optimization problems is much more complex than linear problems, but contemporary tools with good user interfaces help keep such problems manageable. Other problems require that variables can only take on integer values (we can't have 2.5 airplanes for example). Another class of problem makes use of network programming, where the aim is to minimize some function of the network. A good example here is minimizing the cost of transport as a given number of trucks ship goods to a network of stores.

Other techniques are also finding their way into prescriptive analytics, in addition to the optimization techniques mentioned above. Queuing problems are common in business and optimization techniques are used to address problems from traffic flow through to minimizing check-out queues in stores. Simulation is also used to model the performance of business systems and is a large domain in its own right. It is very often the case that the 'best' solution to various business problems simply cannot be found, and so looking for a good solution becomes necessary, and this is where both analyst and business managers need to really understand the problem they are attempting to solve.

Stochastic optimization takes prescriptive analytics into a realm where many uncertainties in business can be accommodated. Employee attendance, future sales, the response to marketing campaigns, wastage and hundreds of other variables are inherently uncertain in nature. The variables can be treated as random in many ways, with limits on how much they can vary. The stochastic optimization algorithms will find the best, or at least a good, solution for the most likely outcomes where uncertainty is present.

Such is the advanced nature of some prescriptive analytics tools and solutions that near real-time optimization can occur to accommodate changing business conditions. For the very largest optimization problems this still is not possible, but the frontier is being pushed forward rapidly and in volatile markets real-time optimization can, and often does deliver significant benefits.

Prescriptive analytics technologies will advance rapidly over the next four to five years with new entrants and new capabilities. As always integration is largely the determiner of how successfully prescriptive analytics can be used in a live production environment. Building prescriptive models is one thing, using them in a production environment requires extensive integration capabilities and good management and control tools.

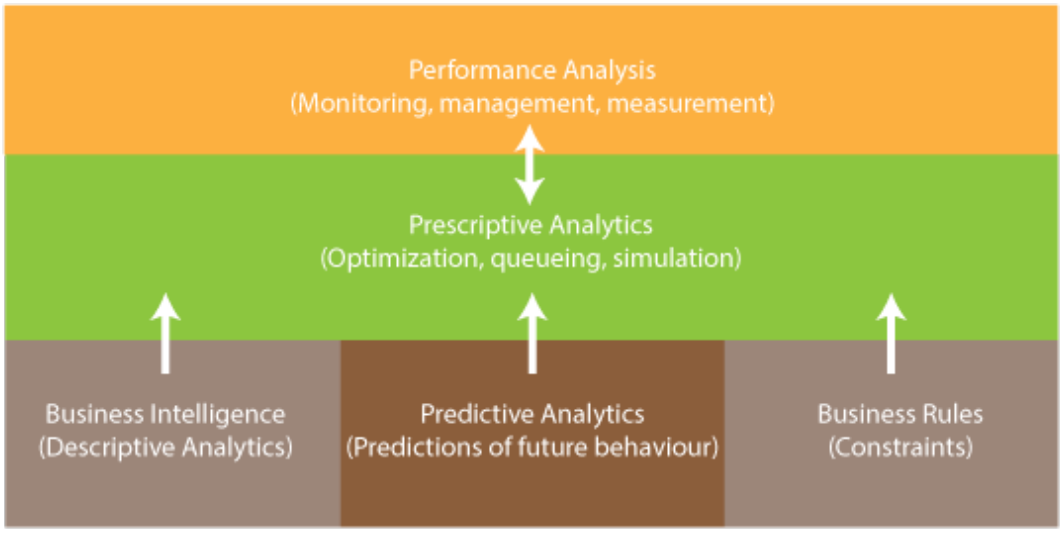
Integration

Prescriptive analytics, and specifically optimization, has traditionally been treated as a stand-alone domain. This has meant that the inputs for optimization have been manually created, and that the outputs have been produced in isolation from all other systems. In practical terms this equates to increased cost, delays, errors and a frustrating lack of flexibility. For these reasons it is important that the prescriptive analytics

environment is integrated into the overall systems environment as fully as possible.

One of the most significant overheads associated with performing prescriptive analytics is the creation and maintenance of business rules, or in the terminology of optimization, the constraints. Even modest optimization projects might involve hundreds or thousands of rules, and recreating them for every optimization problem is a heavy overhead prone to errors. Ideally the prescriptive analytics tools should have access to a business rules data base and management system and be able to convert them into a format that makes sense. In fact for large optimization projects such a facility is not really an option.

Business rules typically express how resources can be combined. It obviously does not make sense to offer a customer a five per cent discount after offering ten per cent on the same product. Neither is it likely to be acceptable that the whole of a workforce is laid off periodically. These and tens of thousands of other constraints typically influence the way large businesses operate, and since prescriptive analytics makes extensive use of them there should be a high level of integration.



However it is not just business rules which should be integrated. Many of the inputs involve the specification of forecasts and other types of prediction. Predictive analytics will generate many models with predictive power - the propensity of customers to respond to an offer for example. These can be used as input to a optimization project, and once again it is extremely useful if the predictive and prescriptive analytics environments are integrated. Optimization caters for probabilistic inputs (a sales forecast for example) through stochastic optimization techniques, and since these facilitate much greater sophistication, the integration with probabilistic predictive models is very important.

Business intelligence, or descriptive analytics, can also provide key inputs for prescriptive analytics, providing information on what has happened and is happening. Information such as the sales of items by region and period might constitute key information in production planning optimization, along with many other metrics. It is important that output from reporting and analytics are available to optimization projects.

Finally the results of optimization need to be deployed, and done so in an environment where performance can be monitored, measured and managed. The results of optimization may have a very short life, and it is essential that changes in circumstances can be compared with assumptions made when an optimized model was built and new models built in a timely manner.

A lack of integration is not so problematic for small, one-off projects, but it becomes a major headache as projects grow in size and frequency. Integration should be in the top three or four requirements when selecting a prescriptive analytics platform.

Business Application

Prescriptive analytics applications embrace most aspects of business operations. In fact whenever there is a resource allocation problem with constraining rules, many variables and well defined objectives, then it is likely that prescriptive analytics can be applied.

It's application in marketing is gaining momentum, particularly combined with predictive analytics where customer response to various initiatives can be predicted. Resource allocation can then be optimized based on these predictions and the relevant business rules which determine how customers are to be treated. Of course optimization has been applied for decades to workforce and other resource deployment problems, but it is only now with much faster execution that other problems are becoming feasible. Even sports bodies are using optimization to maximize television ratings while catering to the needs of players and fans.

Optimization is used widely in the airline industry. Since margins are so thin frequent optimization can help turn unprofitable scenarios into profitable ones. Bad weather will mean re-optimizing resource allocation, and as seats are sold on a flight it is desirable to re-optimize the allocation of various classes of seat. Unscheduled maintenance can also cause resources to be re-allocated and optimization is essential if inefficiencies are to be ironed out of operations. In this particular application it is very important that optimization can be performed in less than 24 hours, and often in much shorter times.

The energy industries have used optimization for decades, but increased optimization speed and capability have meant new applications. With new energy sources, and particularly renewable ones, it is desirable to combine prescriptive analytics with predictive analytics to create long term energy scenarios. The predictive analytics are used to create long term energy forecasts and optimization is then applied to explore optimization of energy production.

An application ideally suited to optimization is that presented by vehicle rental firms. With thousands of vehicles assigned to hundreds of locations there is an obvious need to make sure that each vehicle is located where it will generate most revenue.

The earliest application of optimization included problems such as 'least cost mix'. Here the aim is to produce a mix of component materials with a particular specification, given that each component has its own constituent properties. Animal food mix was the classic application where overall nutrient levels had to be met from mixing various ingredients, which each had their own nutrient levels. This and other similar problems are still solved using optimization technology, but look fairly primitive compared with the real-life

problems that are being solved today.

Optimization technology will grow in use as it becomes more user friendly, executes with greater speed, and is more tightly integrated into the over systems environment. For many businesses optimization will provide an edge that cannot be achieved any other way.

Strategy

Realizing the considerable benefits that prescriptive analytics can bring to an organization requires that several issues are adequately addressed. These include technical, organizational and operational factors and include:

1. Large, complex optimization problems require a team of professionals trained in operations research or some related discipline. Some smaller problems can be addressed by non-specialists using technology such as Excel Solver, but most real-life problems will need an experienced team and significantly more sophisticated technology.

2. High level sponsorship is needed, since prescriptive analytics usually span functional silos. The optimal solution to enterprise problems may seem sub-optimal at the department level for example, and so there will have to be mechanisms put in place to allow such issues to be resolved.

3. The technology platform must scale and offer high levels of performance. While initial projects may be comparatively modest the scale and scope will rapidly grow as benefits are realized. Performance and scaling bottlenecks will be experienced if the supporting technology is architecturally weak.

4. Integration with existing analytics tools and business applications means inefficiencies can be kept to a minimum and errors largely eradicated. Business intelligence, predictive analytics, rules based systems and some transactional applications will need to be integrated with the prescriptive analytics platform. Unless this can be achieved the speed and accuracy of optimization will be compromised.

5. Obviously there needs to be adequate monitoring and management of prescriptive analytics projects, with effective reporting mechanisms so that changes in the business environment can be responded to in a adequate manner, and changes in business strategy quickly implemented.

It should be clear that prescriptive analytics is deeply concerned with the operation efficiency of an organization and needs to be integrated into the information systems environment. Supporting information needs to be extracted from other systems and sent to operational systems to implement the resulting solutions. Parallel with this is the need for management and reporting structures so that associated issues can be resolved. Without this enterprise support the prescriptive analytics efforts tend to remain isolated and inefficient.

Finally it is necessary that the whole prescriptive analytics effort is business driven, with a good understanding of where the major payoffs are and how projects should be prioritized. For organizations inexperienced in the domain this may mean using external resources (such as consultants and experienced suppliers) to formulate a strategy. In some industries it may be possible to buy solutions to specific

problems, and inevitably the options here will grow rapidly over coming years. However it really is very important that organizations do not end up with multiple point solutions, and worse still with solutions that will not scale. And so the issues listed above are just as applicable to solutions as they are to deploying a prescriptive analytics platform.

Optimization Technologies

Suppliers of optimisation technology include:

Premium Solver Pro

Frontline Solvers provide a number of Excel Add-Ins, and it is claimed that Premium Solver Pro will solver larger optimization problems with much greater speed (between 2 and 50 times faster). Up to 2000 decision variables can be specified and users can specify their problems with an Excel Solver type utility or a Ribbon and Task Pane interface. Premium Solver Pro automatically chooses the best solver engine based on the model specification. A licence costs US$995.

SolverStudio

SolverStudio is a free Add-In for Excel that supports the creation of optimization models using a variety of modelling languages and solvers, including PuLP (included with SolverStudio), AMPL, GMPL, GAMS, Gurobi, COOPR/Pyomo, and SimPy. The models can be created and edited without leaving excel, and the model is stored within the workbook.

What'sBest

This is a an Add-In provided by Lindo Systems and is targeted at large industrial size optimization problems. It addresses linear, nonlinear (convex and nonconvex/Global), quadratic, quadratically constrained, second order cone, stochastic, and integer optimization. Some of the largest problems formulated in What'sBest use over 100,000 variables and it is claimed that execution speeds are still acceptable. A variety of options are available with a starting price of US$495, rising to several thousand dollars if all options are included.

FICO

Used in many of the world's largest corporations and addressing many complex optimisation problems, FICO leads the market in providing optimisation technology that integrates with analytics and business rule based systems. The technology is currently being made available to a broader audience via the FICO Analytic Cloud.

IBM ILOG

Since 2009 ILOG has been part of IBM and completes an impressive array of optimization capabilities. IBM ILOG CPLEX Optimization Studio provides an environment for model development via mathematical and constraint programming. IBM ILOG ODM Enterprise is a platform for building optimization based planning and scheduling applications. Supply Chain Optimization is also offered as a particular solution.

River Logic

Enterprise Optimizer has several components. Workstation supports rapid development of decision-support applications. EO Server supports deployment of planning and analytics solutions. EO IBP Framework facilitates integrated business planning. Strong integration with Microsoft systems architectures.

Business Process Management

People have always created and modified business processes, but until business process management (BPM) tools became available it was often an informal and fairly ad-hoc procedure. There is nothing particularly complex about BPM, despite the unnecessarily elaborate terminology. In essence it provides a language for people to design, analyze, build, modify and discuss business processes. And it also creates a bridge between the process and systems used to implement it – assuming technology is a factor, as it nearly always is.

The business process management cycle starts with analysis and design – as do all creative projects. A diagramming notation is used to show the activities, routing and messages involved in a process – usually based on the BPMN (Business Process Model and Notation) standard. Business Process Management Systems (BPMS) provide a framework for these activities and a variety of tools to help in the design, implementation and monitoring of business processes. During the design and analysis phase a BPMS usually allows a design to be simulated, so that any undesirable side effects can be detected.

If a business process is to be more than just a diagram it needs to be plugged in to the operational systems. To this end BPMS offer connectors to various applications and support for methods such as Service Oriented Architectures (SOA), where existing systems can be made to behave as a set of services which a business process uses. Obviously this can become fairly technical – but no more than any other system implementation.

Finally we are ready to put the business processes into production where process instances (individual transactions and workflows) are handled. An essential part of this is the creation of a log of all process activities – invaluable material for the next phase.

Evaluating the performance of processes is a major benefit of a BPMS. The logs created during process execution can be interrogated to identify bottlenecks, loopholes, inefficiencies and any other form of poor performance. To this end process mining is a relatively new addition to many BPMS, where logs are interrogated for patterns of behavior which are sub-optimal.

Of course this is just a brief over-view, and depending on the size of the organization the whole process can become quite involved with thousands of processes having to be orchestrated. Even so these basic principles still apply.

Open Source BPMS

Activiti is a light-weight workflow and Business Process Management (BPM) Platform targeted at business people, developers and system admins. Its core is a super-fast and rock-solid BPMN 2 process engine for Java. It's open-source and distributed under the Apache license. Activiti runs in any Java application, on a server, on a cluster or in the cloud. It integrates perfectly with Spring, it is extremely lightweight and based on simple concepts.

Activiti supports all aspects of Business Process Management (BPM) in the full context of software development. This includes non technical aspects like analysis, modeling and optimizing business processes as well as technical aspects of creating software support for business processes. Activiti recognizes that BPM as a management discipline is a completely different aspect then BPM as software engineering.

Activiti's primary purpose and focus is to implement the general purpose process language BPMN 2.0. And there is no single process language that can cover all the use cases well. In many cases a custom dedicated process language makes sense. So at the core, Activiti has the Process Virtual Machine architecture. That means that any custom process language can be build on top of it.

Bonita BPM improves business operations by connecting people, processes, and information systems into easily managed applications. Use Bonita Studio to map the organization, define the data structure, build the user interface, and create actionable reports. Bonita Portal creates a central location to perform tasks, monitor case completion, search for information, and collaborate with peers.

Camunda is an open source platform for workflow and business process automation. It executes BPMN 2.0, is very light-weight and scales very well. Camunda is written in Java and a perfect match for Java EE and Spring while providing a powerful REST API and script language support. You can use camunda BPM for system integration workflows as well as for human workflow and case management.

You can add camunda to your Java application as a library. You can also use it as a container service in Tomcat, JBoss etc., so it can be used by multiple applications which can be redeployed without shutting down the process engine. Some of the biggest companies in the world and most trusted public institutions rely on camunda.

Intalio bpms provides a comprehensive enterprise-class platform to design, deploy, and manage the most complex business processes; over 1000 organizations world-wide in all industries rely on the technology to manage their mission-critical business processes. Intalio bpms features an intuitive and powerful visual designer and a reliable high-performance process execution server. It also includes enterprise-level capabilities such as business activity and metrics monitoring, business rules and decision management, document management, mobility support, and system integration tools and portals.

jBPM is a flexible Business Process Management (BPM) Suite. It makes the bridge between business analysts and developers. Traditional BPM engines have a focus that is limited to non-technical people only. jBPM has a dual focus: it offers process management features in a way that both business users and

developers like it. The core of jBPM is a light-weight, extensible workflow engine written in pure Java that allows you to execute business processes using the latest BPMN 2.0 specification. It can run in any Java environment, embedded in your application or as a service.

jSonic BPM suite enables enterprise owners to align business processes with the dynamic market conditions, statutory compliances and, customer and partner requirements. It is a comprehensive solution that improves the bottom line of organization by increasing process efficiency, optimizing resource utilization and automating human workflow system.

jSonic BPM suite, the Open Source BPM Software offers an all-encompassing solution covering process designing, modeling, executing, automating and monitoring as per the business needs and wants. The major components of the suite include Process Management, Workflow Management and the Interface Designer.

Orchestra is a complete solution to handle long-running, service oriented processes. It provides out of the box orchestration functionalities to handle complex business processes. It is based on the OASIS standard BPEL (Business Process Execution Language). Orchestra is fully Open Source and is downloadable under the LGPL License.

ProcessMaker is a cost effective and easy to use open source business process management (BPM) or workflow software application. Workflow software such as ProcessMaker can assist organizations of any size with designing, automating and deploying business processes or workflows of various kinds.

ProcessMaker workflow software features an extensive toolbox which provides the ability to easily create digital forms and map out fully functioning workflows. The software is completely web based and accessed via any web browser, making it simple to manage and coordinate workflows throughout an entire organization – including user groups and departments. ProcessMaker workflow software can also interact with other applications and systems such as ERP, business intelligence, CRM and document management.

Red Hat JBoss BPM Suite is the JBoss platform for Business Process Management (BPM). It enables enterprise business and IT users to document, simulate, manage, automate and monitor business processes and policies. It is designed to empower business and IT users to collaborate more effectively, so business applications can be changed more easily and quickly. Create, test, deploy and monitor BPMN2-based business processes to optimize enterprise workflows and automate critical processes. Includes all the business rules and event processing capabilities of Red Hat JBoss BRMS. Easily create real-time dashboards to monitor key performance indicators for running processes and activities.

Talend's BPM products enable managers, business analysts, developers and end users to model current processes, collaborate on improvements, and rapidly create and optimize process-driven solutions in minutes. Talend combines three solutions in one: an innovative process modeler, a powerful BPM and workflow engine, and a breakthrough user interface for the creation of forms. You can create human interactive or process-based applications, and automate and optimize business processes in a single day.

About Butler Analytics

Butler Analytics was founded by Martin Butler, best known as founder of Butler Group – Europe's largest indigenous technology analyst firm until its acquisition by Datamonitor in 2005. Martin is widely recognized as one of the most eminent technology analysts in Europe.

The aim of Butler Analytics is evaluate and inform on all analytics technologies and suppliers. As the emphasis shifts away from the traditional use of technology as a means of process automation, to its use in aiding with, and automating business decisions, so business and technology managers need a source information to keep them abreast of developments. Butler Analytics aims to provide such a service.

Suppliers who feel they have been neglected or wrongly represented can contact us through:

info@butleranalytics.com

Feedback is always welcome and Martin can be contacted at:

martin@butleranalytics.com