Text and Data Mining: Technologies Under Construction

WHO'S INSIDE

Accenture
American Institute of Biological Sciences
Battelle
Bristol-Myers Squibb
Clinerion
Columbia Pipeline Group
Copyright Clearance Center, Inc.
CrossRef
Docear
Elsevier
Figshare
General Electric
IBM
Komatsu
Linguamatics Limited
MedAware
Mercedes-Benz
Meta
Novartis
OMICtools
Science Europe
SciTech Strategies
SPARC
Sparrho
Spotfire
Talix
UnitedHealth Group
Verisk
VisTrails
Wellcome Trust

Market Performance

Advancing the Business of Information

January 22, 2016

Table of Contents

Table of Contents for Figures & Tables

Why This Topic

Text and data mining (TDM), also referred to as content mining, is a major focus for academia, governments, healthcare, and industry as a way to unleash the potential for previously undiscovered connections among people, places, things, and, for the purpose of this report, scientific, technical, and healthcare information. Although there have been continuous advances in methodologies and technologies, the power of TDM has not yet been fully realized, and technology struggles to keep up with the ever-increasing flow of content.

Given that proliferation of information, the importance of TDM is undisputed: Connections between individual or cohort health data can lead to precise diagnoses and treatments for diseases or medical conditions, and previously undiscovered links in scientific and technical data can produce unimagined results in particle physics, biological sciences, social sciences, engineering, and more. Mining all study results, including negative findings, can prevent duplicative experimentation, as well as time and money spent on experiments already conducted.

Tools, users, uses, and solutions vary by industry, business function, the desired result, and the data set used. We focus on scientific, technical, and medical research, and healthcare information, outlining the components of TDM, deployment of TDM today, applications to come, and hurdles to overcome as TDM progresses.

Methodology

Primary research for this report comprised a series of approximately 20 in-depth interviews with content vendors, technology and tool providers, and TDM experts. Secondary research supplemented those interviews with information gathered from websites, industry blogs, webinars, published reports, and mainstream press. Access to Elsevier's ScienceDirect and Thomson Reuters' Web of Science was also invaluable.

Outsell's daily dialogue with the market and key stakeholders in STM and healthcare also added depth and breadth to primary and secondary research findings.

How It Works

The objective of TDM is to extract and create new knowledge, data, and insights from existing materials. The process flows as follows:

- Using TDM techniques, data scientists make previously undiscovered connections between apparently unrelated content (using "content" as an all-encompassing term including text, numbers, images, video, and so forth).

- Human curation ensures the trustworthiness of those connections through validating underlying information.
- Applying descriptive, predictive, and prescriptive analytics allows users to understand what those connections mean or can lead to.
- Visualizing the results supports processing and understanding new information.

Data mining typically refers to making connections or finding patterns within structured data. The structure consists of a database of discrete fields containing defined short textual or numeric data, with an overlying technology searching within those fields. The critical activity, however, is making connections, not searching. Common practice is to categorize data mining by the structure of the database, the type of information within the database, the algorithms used to do the mining, and utilization of the resulting data. Text mining refers to searching and connecting information in both structured and unstructured text such as doctors' notes, or collections of scientific documents.

Advances in data capture, transmission, storage, and processing, as well as machine learning and natural language processing (NLP), now allow organizations to combine and integrate databases of all sizes into data warehouses, a means of centralizing and managing data. Equally dramatic advances in analytical software and visualization tools allow users to analyze, understand, and apply new data generated from those activities.

These basic facts drive efforts to move text and data mining processes:

- Research articles and papers contain much of the world's scientific and medical knowledge;
- Approximately 80% of information, both scientific and medical, exists in unstructured text;
- Humans must curate a significant portion of mined data, as well as underlying content, to validate and realize its full value;
- In most cases, data must be significantly longitudinal (have critical mass) to generate meaningful results.

Scientific databases cover many components of biology – genes, protein sequences, metabolic pathways, just to name a few – as well as chemistry, botany, earth sciences, and social sciences such as anthropology and archaeology. Most of their content comes from scientific research, typically published in peer-reviewed journals and structured to enable efficient exploitation by means of data mining tools.

Some databases contain unstructured data, often including supporting images, data, or other information that may or may not be structured. The goal is to automate identification, tagging, and structuring that text and data to enable the most efficient methods of extracting new and important information, eliminating the need for human curation. That means addressing the current practice of manual transference of data from published articles into data repositories, leading to disconnects between literature and data – for example, DNA sequences appearing in genetic databases with no known function.

Providers in the TDM marketplace provide services around text and data tagging, warehousing, mining algorithms, visualization, and analytics. Table 1 presents a sampling of providers of TDM-related functions and the markets they serve. **Click on a company name in the table to visit the website.**

Table 1. Providers of TDM-Related Functions

	Types of Businesses Served				
Tool	Engineering	Research	Healthcare	Content Providers	Functionality
Advanced Miner		🟩			Data processing, analysis, and modeling
Alteryx Designer	🟩	🟩	🟩		Data blending, analytics, and reporting
Angoss Knowledge Suite	🟩	🟩	🟩	🟩	Predictive analytics and data mining
Ascribe			🟩		Text analytics
Averbis Text Analytics	🟩	🟩	🟩		Analysis of unstructured and structured data
Aylien		🟩	🟩		Natural language processing, information retrieval, machine learning
Azure HDInsight (Microsoft)	🟩		🟩		Analytics tools platform
CARP Language Technology		🟩	🟩		Processing unstructured text
Clarabridge			🟩		Text analytics
Cogito	🟩	🟩	🟩	🟩	Multilingual management of unstructured information
Data Applied	🟩		🟩		Analytics, data mining, and information visualization
Discovery 5 (Brainspace)	🟩			🟩	Unstructured data analysis
HealthLanguage Analytics			🟩		Natural language processing of clinical texts, clinical data analytics, language engineering infrastructure
HPE Haven (HP)			🟩		Big data management
IBM Analytics	🟩	🟩	🟩	🟩	Big data management and analysis
Lexalytics Text Analytics	🟩	🟩	🟩	🟩	Text mining technologies
LIONoso		🟩	🟩		Machine learning and optimization

	Types of Businesses Served				
Tool	Engineering	Research	Healthcare	Content Providers	Functionality
Meshlabs		🟩	🟩		Social listening, text analytics, and natural language processing
muText	🟩	🟩	🟩		Text mining of unstructured and semistructured data
NetOwl				🟩	Text and entity analytics
Neural Designer	🟩	🟩	🟩		Deep learning
OpenText Content Analytics			🟩		Text mining
Oracle Data Mining	🟩	🟩	🟩	🟩	Data mining and analytics
Pingar	🟩	🟩			Text analytics
Provalis Research	🟩	🟩	🟩	🟩	Text structuring
SAP InfiniteInsight	🟩	🟩	🟩		Text and data analytics
Saplo				🟩	Text analytics
SAS Enterprise Miner		🟩	🟩		Text and data analytics
Smartlogic	🟩	🟩	🟩	🟩	Text mining and analytics
STATISTICA Data Miner	🟩	🟩	🟩		Data mining and analytics
TEMIS		🟩		🟩	Structure and management of unstructured information
Teradata Warehouse Miner	🟩	🟩	🟩		Data management, mining, and analytics
Textual ETL		🟩	🟩	🟩	Structure and management of unstructured information
TIBCO Spotfire Miner		🟩	🟩		Data mining, visualization, and analytics
Veera		🟩	🟩		Data mining and analytics

Engineering = commercial data production, primarily utilities, seismology, etc.
Research = academic and corporate research
Healthcare = patient, payment, claims data
Content providers = publishers

Source: Company websites, company analysis

Applications

Data mining has long supported market research, customer analytics, and many other types of commercial activities, because it is easier to design systems around the structured data collected by commercial websites. Because text is primarily unstructured, however, the road is longer and more difficult to navigate. Academia and industry are on their way, but the push is on to find more and better ways of structuring and mining information. The following are examples of deployment of TDM in scientific research, healthcare, and engineering.

Scientific Research

Text mining has advanced from use cases in pharma, for example, along the pharma pipeline, to post-marketing business intelligence, competitive intelligence, key opinion leader analysis, discovery of adverse drug reactions (pharmacovigilance) in clinical trials, and more. Growing awareness of the importance of TDM is leading to more applications and a heightened focus on its benefits.

Examples of deployment of TDM in the research arena include:

- **Making previously undiscovered connections in research literature.** Extracting value from content isn't about search and discovery; it comprises making a correlation or connection that wouldn't otherwise be apparent. Text mining will produce value once there are multiple facets for making correlations or connections, at various confidence levels, in both structured and unstructured content. In the meantime, applications exist to combine scientific literature and relevant data. In 2013, for example, researchers at Bharathiar University in India developed HPIminer, a system incorporating information from curated databases (HPRD and KEGG), a text mining engine (a combination of three earlier text mining systems that the researchers used), and a web interface to upload biomedical literature. HPIminer visualizes all known protein-protein interactions (PPIs) and pathways of a given protein or interaction retrieved directly from the literature, providing valuable insight into biological systems. The number of PPIs in biomedical literature is huge and continually expanding, contributing to manually curated PPI databases and databases of related protein pathways.

  A more recent initiative, used primarily by pharma companies, is Copyright Clearance Center's RightFind XML for Mining. Launched in 2015, the tool allows users to search for articles, obtain permission to mine content, and download content for processing by a text mining tool provider such as Linguamatics, CCC's development partner, and mine it.

What we expect: Academic researchers, corporations with R&D capabilities, and tools providers such as Linguamatics and CCC are moving toward extracting maximum value from structured and unstructured content. It will come about with the mechanization of content tagging through unsupervised learning – having a computer understand what text says, using pattern matching to get to the cognitive value without human intervention – eliminating the need to curate data. An even bigger hurdle is getting access to multiple data and content sources by crossing silos and allowing access to nonpublic information.

- **Supporting medical research.** Much of pharmaceutical companies' research takes place through clinical trials. Clinicaltrials.gov lists 202,378 clinical trials in 190 countries, including all 50 states of the US. Of those, 59,415 are open studies, and all require a meaningful sample size of patients matching their criteria. Finding suitable participants typically entails combing through electronic medical records containing significant amounts of unstructured information in the form of clinical notes. These physician notes contain nuanced information not found in structured information, also critical to screening trial candidates. Companies such as Clinerion source patients for trials, but they use search methods rather than mining, requiring humans to review clinical notes. In 2015, researchers from Ohio State University and Universidad del Desarrollo in Chile developed baseline methods for identifying text in clinical notes pertinent to trial criteria using natural language processing to expedite the process. Although their discoveries indicate improved methodologies of recognizing systematic human reasoning employed in annotating medical records, the researchers also point out that it was only a start, and they plan to continue their research.

What we expect: More precise patient-clinical trial matching is in development at technology companies such as IBM Watson Health and pharma companies such as Novartis. It is critical to companies testing drugs to move quickly, both to avoid moves by competitors and to discover if a drug or treatment is effective with most clinical trial participants. Just as important is supporting post-trial pharmacovigilance through early detection of adverse drug events and reactions – currently a voluntary activity on the part of physicians and patients. Elsevier's *Journal of Biomedical Informatics* (in which the Ohio State and Chilean researchers' study appeared) published a special issue in 2015 dedicated to papers focused on technologies to help mine text-related pharmacovigilance sources, publishing 13 of 27 submissions – an indication of the importance of this activity, as well as highlighting the early stage of its development.

- **Measuring research impact.** In 2014, the Higher Education Funding Council for England (HEFCE), along with partners including other UK funding councils, Research Councils UK, and the Wellcome Trust, commissioned an analysis of 6,679 impact case studies submitted to the 2014 Research Excellence Framework (used primarily to determine government funding allocations) to prove impact of research by UK higher education institutions. The resulting report is part of a project led by Digital Science, Nature Publishing Group, and King's College London.

  The group used text mining to analyze the case studies across 149 fields of research, 60 impact topics, and 36 units of assessment. Overall, the group identified 3,709 paths to impact arising from multiple fields of research. Other findings include the global impact of UK research and unexpected impact from smaller institutions – as well as an absence of standardized presentation leading to difficulties in analysis and to the conclusion that development of robust impact metrics is unlikely, unless and until standards are imposed on case study submissions.

What we expect: Impact metrics continue to be a primary tool for funders determining deployment of funds, for institutions making hiring, tenure, and promotion decisions, and for authors determining optimal publication routes. With more attention to TDM and to standardizing and normalizing content, impact metrics will be more robust, more easily implemented, and refined to become the optimal tool for making important decisions around areas of research to focus on and to fund.

TDM technologies also support research workflows. Examples include:

- Monitoring emerging research trends through mining scientific literature has been a topic of discussion in the research arena for decades, but until recently there has there been little progress, primarily because of a dearth of electronic data and computing power. Researchers from SciTech Strategies published a study in late 2015 outlining a process of mining scientific literature using citation analysis to identify the taxonomy of scientific and technical knowledge, a critical tool in research planning and evaluation. Their findings include superior results when using documents rather than journals, and improved metrics over those used today. The authors envision utilizing their methodology not only to inform funding and research decisions, but also to identify resources, both human and physical, to better advance innovation in scientific discovery.
- Recommendation engines utilize TDM to alert researchers to literature similar to what they are reading or wish to know about, based on reading patterns or stated preferences. This is a valued way to cope with the massive amount of information researchers deal with today. A number of recommendation tools, such as Meta, PubChase, Sparrho, and Docear, are already available for researchers. Databases such as Elsevier's ScienceDirect utilize recommendation engines also: When a user searches on a term, recommended books appear to the right of the results list, and when a user downloads the PDF of an article, a box pops up with three articles directly related to the downloaded article.
- Systematic reviews play an important role in fields of research such as evidence-based medicine and software engineering. The rapidly increasing number of published studies makes the task of identifying relevant studies in an unbiased way for inclusion in systematic reviews both complex and time-consuming, often lasting up to three years and requiring two reviewers to assess each article to minimize errors. A 2014 study found that automating the process through text mining could reduce the workload by 30% to 70%, with only a 5% loss of relevant studies. Another study in 2015, by researchers at the University of Bristol, also found that supervised machine learning saves time as well as assisting with risk-of-bias assessments by reducing human errors and subjectivity.

- Sentiment analysis can validate citation counts: An article may have hundreds of citations, but knowing if the references are negative ("Joan Smith was proved wrong in her previous research," "the referenced data set is worthless," and the like) is important to recognizing the value of the referenced study. This type of application is not yet in use, but metrics companies will look to deploy it in the future. Recent sentiment analysis studies analyze Twitter mentions of research, but findings to date indicate that researchers primarily use social media to disseminate information about a study rather than express an opinion.
- Literature and patent analysis can be time-intensive: Researchers look for benchmarks, previous findings, and other information buried in mountains of text. TDM finds and makes connections between patent filings or journal articles, creating new information for researchers to incorporate in studies or to support grant proposals. It can also point to anomalies or errors, or buttress ongoing research. It has become more and more imperative to analyze patent information to identify new trends to improve current workflows or to guide and project future business strategies as patent information reflects the innovation capabilities. In 2014, Bristol-Myers Squibb (BMS) partnered with Linguamatics to mine patents to identify a specific technology as well as supporting information. They concluded that mining patent information provides valuable knowledge driving business decisions. Extracting key concepts from patent documents and correlating them with metadata from those documents using Linguamatics I2E text analytics tool, combined with Spotfire data visualization software, for the first time provided a feasible approach at BMS to automatically extract meaningful information from unstructured text in patents and identify trending information in a productively efficient manner.

What we expect: Each research workflow application represents a focal point for platform developers today; all are undergoing development or further enhancement. The future is enrichments of what already exists, including eliminating human intervention and curation, as well as new tools supporting robust outcomes. Literature such as biomedical patents are complex and growing rapidly in volume, and so effective text mining and visualization tools will continue to come into play; scientists and businesses require accurate and timely information to drive research and business decisions.

Healthcare

The healthcare industry is an advanced user of TDM and analytics, including mining patient data, claims data, hospital data such as readmissions, and more. Applications are critical to the health of patients as well as to the financial health of providers – recognized by the US government in 2015 through the appointment of the country's first US Chief Data Scientist. Developments to date include:

- **Advancing patient health and effective claims management.** Tools are critical for analyzing and making the right clinical decisions by streamlining formerly manual processes of chart reviews and reimbursement coding. Knowledge about a patient, often only implied in a clinical record, is why humans must review records – 38% of diabetics in the US do not have “diabetic” coded in their medical record, for example. Chart coding may relate only to a complaint, not diabetes – even if there is medication in the chart for diabetes. TDM tools can indicate a patient is likely a diabetic and therefore worth more under the US’s value-based reimbursement program, because under Medicare Advantage, for example, a risk-sharing program with payers and providers, that patient now generates higher reimbursement for the physician or hospital.

  Using a proprietary taxonomy and NLP, Coding InSight from Talix (spun off from Healthline Networks) identifies high-risk patients to ensure correct coding and timely intervention. The tool’s goal is to match resources to requirements, not to fill providers’ coffers. Verisk Health is another provider using NLP and TDM to provide fraud prevention to its customers, as well as shifting insurers’ mind-sets from reactive to proactive preventive measures.

What we expect: With enough data and advanced mining algorithms, fraud prevention and payment optimization will be possible without human intervention, saving providers and payers money and bringing to light issues such as gaps in coding that lead to potential diagnostic errors.

- **Preventing diagnostic and medication errors.** Companies such as MedAware address issues such as diagnostic errors, medication errors, patient medication compliance, recommended treatment paths, and more. The company offers a platform integrating with a hospital’s electronic medical records (EMR) system to detect prescription errors before they happen. The technology detects patterns in patient records to flag prescription anomalies and blocks the order until the doctor confirms it or cancels and places a reorder.

  Although EMR systems have fail-safes around drug interactions, dosage anomalies, and duplicate prescriptions, they do not use data to determine in real time if a prescribed drug may be the wrong one. MedAware’s algorithm uses millions of patient records to determine if dosage or medication is out of the norm. Retrospective studies in Israel (the company’s base) and in Boston indicate that its use will benefit both patients and providers once installed and operational in hospital systems.

What we expect: Platforms such as MedAware will prevent the thousands of prescription errors currently made annually with computerized order entry by providers. This will save lives, lower readmission rates, and cut costs for providers and patients alike.

- **Accelerating healthcare research.** OptumLabs, an open research and innovation collaborative owned by UnitedHealth Group’s management and analytics subsidiary, Optum, and founded by Optum, AARP, and the Mayo Clinic, continues to grow its members, studies, and goals. The group competes with the likes of Watson Health, Geisinger Health

System's xG Health Solutions, and Aetna, looking to capitalize on big data and its potential. OptumLabs provides partners with robust de-identified data, analytical tools, an open forum for various perspectives on healthcare, and support in the form of experts. The goal is a program of discovery through "constellations" of multiyear, multiproject, and multipartner research programs around high-profile diseases such as Alzheimer's, heart failure, and cancer, as well as a focus on complex comorbidity, performance measure development, big data methodologies, and more.

The purpose of the group initially was to accelerate healthcare research, as well as its application, through comparing treatment options, but now it is to become a greater vision of developing more accurate predictors of illness, advancing cognitive computing, leading to better tools for TDM. Partners include leading universities, healthcare systems, corporations, not-for-profit organizations, and the US Department of Health and Human Services.

What we expect: Bringing together academic programs, mission-driven organizations, profit-driven corporations, and government entities will lead to including diverse perspectives and knowledge in pursuit of breakthrough healthcare. Developments driving those advances include understanding what data to collect and how to combine it, as well as what to do with it, hurdles that today stymie many healthcare systems just starting down the path of managing big data.

- **Using sensors to monitor health.** Battelle, the world's largest nonprofit research and development organization, conducts research and development, and designs and manufactures products in segments such as consumer and industrial, energy and environment, health and pharmaceutical, and national security. The company's NeuroLife Neural Bypass technology, using medical sensors to aid in the recovery of nervous system injuries, incorporates its EluciData technology, an analytics program using advanced machine learning and pattern recognition to translate complex neurosignals into clear information for diagnostics and therapies. This allows paralyzed patients to regain conscious control of fingers, hands, wrists, and arms. The EluciData technology combines data from sources such as EMRs, medical sensors worn by patients, clinical devices, diagnostics, imaging, and billing records, as well as information sources such as online diet and exercise journals and weather reports. Using this data, EluciData can predict outcomes based on patient profiles, develop personalized medical recommendations, and send automated alerts to healthcare providers. It is also used to monitor patient compliance for clinical trials and analyze variables to understand variances in outcomes.

What we expect: Advances in materials science, device engineering, and data analytics are converging to make medical sensors smaller, more sensitive, and smarter than ever before. This means expansion in the potential uses of medical sensors and improved biocompatibility for long-term monitoring. Smaller sensors translate into the potential for injection into bloodstreams to monitor organ function, or implantation to monitor individual nerves. More sensitive sensors produce more data for diagnostics and treatments, and smarter sensors means more advances like NeuroLife.

When providers begin to connect data silos – and share information with each other – creating longitudinal databases, deeper discovery can begin. This will lead to cost reduction, objective and complete analysis, and timely discovery of issues. Analytic tools are early-stage, as is NLP technology. The rapidly evolving field of health economics is only a few years old, and advancements are moving at a rate faster than the tools and technologies used to study them. This means healthcare is currently a treatment-naïve environment where technologies, assessment, and analytic tools have not kept pace with the rate of data generation (bound to accelerate even more with proliferation of wearable devices), treatment options, and drug discovery. The balance will shift as advances in TDM continue, and as healthcare systems learn what type of data to collect, how to collect it, and how to analyze it effectively. At that point, machine learning, TDM, and analytics will help predict the "flight path" of a patient, opening the door to full implementation of prescriptive analytics.

Engineering

With the explosion of technology around sensors, engineers of all types have the ability to innovate and address issues facing businesses, individuals, and populations. Examples include:

- **Advancing systems maintenance.** A significant portion of lost revenue in the oil and gas pipeline industry is a result of poor equipment reliability and unexpected production losses. Pipeline integrity is also critical to reducing maintenance costs and safety. Users such as trending and plant historians can apply predictive analytics to data from multiple control-level sources. Such users typically rely on traditional trending and historian software (trending software captures plant management information about production status, performance monitoring, quality assurance, tracking and genealogy, and product delivery). With the amount of data available from meters, inspections, asset management systems, leak detection sensors, and other instruments, it is now possible to make better decisions for pipeline maintenance. Columbia Pipeline Group (CPG) was the first to implement predictive analytics tools from GE and Accenture, beginning in 2015, analyzing pipeline integrity across 15,000 miles of interstate natural gas pipelines.

  Although these analytics are quickly developing to serve industry, they still have the same human-intervention needs as all big data efforts: Engineers must advise analytics experts on the underlying data, and the amount and source of the data is critical, depending on up-to-date installed instrumentation and real-time data.

What we expect: Both the amount of data and the algorithms will continue to develop, eventually providing enough information to prevent maintenance issues, let alone environmental disasters. Mechanizing the process will lead to further cost savings and safety measures.

- **Boosting efficiency in areas ranging from excavation to transport and power generation.** Komatsu and General Electric announced in April 2015 a partnership to provide services for mining projects. GE, providing data analysis services to companies in aviation, healthcare, energy production and distribution, transportation, and manufacturing since its investment in Pivotal in 2013, will use the same type of technology to partner with Komatsu, a Japanese equipment company. The partnership will support mining clients by sending operational data from large dump trucks – some of them driverless – to a GE data center in the US in order to calculate optimal routes and positioning, as well as determine speed and braking requirements based on the terrain. In addition, Komatsu will install equipment to increase fuel efficiency. Komatsu's technology alone can boost fuel efficiency by 5%. Combined with GE's technology, that number rises to 13%. That is meaningful when a construction site with 300 trucks can save over $4 million with a 1% improvement.

What we expect: Companies with access to detailed operating information through the use of sensors will be able to cut costs and increase productivity through mining and analysis of data, supported by companies such as GE that have the requisite computing power and technology. Not only will these types of services provide operational support to industrial corporations, they will also support sustainability by incorporating environmental data into the analytical processes.

- **Supporting real-time business decisions.** Demands on automobile manufacturers have increased as they manage a greater number of models and customization options while also managing shorter product life cycles. At the same time, networked sensors and complex machinery combine to produce data to improve production processes. Combining diverse data with production demands led Mercedes-AMG, the performance unit of Mercedes-Benz, to implement data mining technologies to drive real-time business performance. Starting with accounting and finance, then including development and manufacturing, the company in 2014 piloted a quality assurance platform using predictive analytics to optimize engine-testing processes. The results are efficient use of expensive equipment, as well as the ability to correlate historical test data with sensor data from the engines undergoing testing, all in real time. No longer do the company's engineers have to finish testing, analyze the data, and address issues. The time freed by stopping testing when it identifies an issue allows for more testing, as well as more time for engineers to focus on refining engine quality. It also enables the company to launch more models into more market segments, and to allow for more customization of vehicles. The company is now investigating and testing application of real-time analytics to project management and inventory management.

What we expect: Real-time insights through real-time data are coming to fruition through industrial applications. Business processes will continue to evolve and change with applications of both internal and external data to production systems, providing immediate insights to those designing and creating new products. Those insights range from financial considerations to production efficiencies, inventory management, resource utilization, and more.

Advances in smart sensors become advances in technology and methods for analyzing and deploying strategies for manufacturing operations, service logistics, maintenance management, and more. Predictive analytics tools are being developed to extract and analyze information from multistream sensor signals to predict future performance of complex engineering systems, and they support decisions around the actions impacting those systems.

The same is true in healthcare, with sensors in mobile devices sending data to healthcare organizations at unprecedented volume and velocity. Looking to a future of data-driven decisions, being able to manage, analyze, and deploy data, is critical to optimizing performance in healthcare, engineering, and research. There are, however, hurdles to reaching that goal.

Challenges

TDM applications providers and their markets face challenges to the utopian vision of open content, fully structured or accessible to mining algorithms, to extract optimal value from digital information. These include:

- **High barriers of data integration because of domain silos.** Many fundamental scientific questions require data from two or more domains; barriers to accomplishing that primarily rest on data ownership and lack of common standards. A report by the American Institute of Biological Sciences points out that community agreement around standards – not more standards – will advance data integration. Linking between data systems is occurring, but technological, commercial, and legal issues hamper progress.

  Integrating information applies also to social issues, such as preventing pandemics, raising questions around laws, privacy, and data ownership. An example is the potential to take location data from sources such as cell phones and security cameras and combine it with online medical data to construct a "heat map" of the individuals who may have been in proximity to someone carrying an infectious disease. In this way, exposed individuals might be contacted and alerted to symptoms and precautious to take.

- **Lack of trust.** The purpose of peer review is to ensure accuracy and reliability in published research. It is a flawed system with instances of peer review fraud, lack of reproducibility, and lack of knowledge about unreliable research inputs or methods. Even with stringent peer review, use of devices or reagents known to be defective or of inconsistent quality, or experiments produced in a low-quality lab, make a study undependable. The untrustworthiness of that underlying information can cause a company to discount TDM results. Allowing the signal to rise over the noise by increasing the amount of underlying information may help, but that can happen only by federating data and allowing for wholesale mining – something that is still over the horizon. Today, researchers spend time vetting underlying information, diminishing the benefit of time and effort saved by TDM methodologies.

  The issue of trustworthiness can combine with that of cross-integration: Researchers recently compared papers that gauged the effect of chronic morphine exposure on genes, based on the same original microarray data deposited in a gene database. What they found was inconsistent gene identifiers in the studies, as well as a lack of documentation for approaches each study used. Even after rectifying these inconsistencies, they found that the results between the studies differed widely.

- **Copyright restrictions on access to content.** In 2014, British copyright law changed to include TDM as an exception – with the restriction of applying only to noncommercial text and data mining initiatives. This excludes corporations and their institutional partners from benefitting from the exemption. The move is a topic of heated discussion because publishers can restrict mining of content if large-scale TDM activities "impact system performance." Considerable lobbying has resulted, to eliminate the restrictions to provide wider access for TDM activities. Publisher objections include erosion of copyright protection, the potential for creating competing content offerings through incorporating copyright-protected content, and potential impact on the reputation and value of branded offerings through lack of quality control of those competing products.

  Efforts to offset concerns include:

  - A statement by STM Association publishers committing to facilitate TDM for noncommercial scientific research in the EU, including granting licenses permitting mining of copyright-protected content on reasonable terms;
  - Crossref's Text and Data Mining Services, a collection of voluntarily submitted publisher TDM license URLs to help researchers see how they can use TDM legally with relevant content;
  - A Scholarly Publishing and Academic Resources Coalition (SPARC) and Science Europe report addressing why copyright is a barrier to TDM, evidence for the demand of TDM, and its benefit to academic research and innovation.

- **Lack of harmonization of metadata.** Particularly important to effective and efficient TDM are standards around metadata, ontologies, and technologies that can cross-mine content types and formats. Harmonizing metadata formats across publishers, repositories, and so forth is key to creating databases large enough to be effectively mined.

- **Technology not keeping up with discovery.** Storage is an issue: Data sets can require – and should be – multiple terabytes to generate reliable data. Adequate storage is a challenge for many enterprises looking to accelerate their work through TDM.

  Debates around TDM methodologies uncover another issue. Starting with machine learning, for example, the following key questions arise:

  - How accurate is a machine-learning algorithm when used against a particular type and quantity of training data?

- Can the algorithm deal appropriately with errors in modeling assumptions or in the training data itself?
- Can there even be a single algorithm for every body of training data?

Going beyond technical issues are education and an understanding of what text and data mining can do, as well as having the right tools and the people to use those tools. Linguamatics notes a change over the past five years in customers looking to hire informatics experts used to dealing with structured databases rather than traditional information scientists, deriving primarily from the focus on text mining.

Finally, privacy and security restrictions are issues that will always exist in the digital information marketplace. There is a fine line between personal privacy and the greater good, a subjective delineation potentially crossed through governmental legislation, by inappropriate interference by hackers, or simply by accident.

Implications

It isn't difficult to imagine how TDM will benefit research and healthcare, building on today's technologies. Those technologies already provide much-needed insights and information, but they are only the beginning. As systems get smarter, the goal will be to conduct an assessment on a patient to determine if that patient is likely to get a disease before they even get it, much like BRCA1 (breast cancer) genetic testing. Understanding risk characteristics around specific patients or cohorts will lower the cost of care and improve patient outcomes, as systems get smarter. Population health studies will add to the data, providing doctors with tools that "learn" from continued data input, leading to better diagnostics and treatments. Today, developments revolve in large part around getting the right money to the right place; the future will focus more on actual clinical decision support at the point of care: Coding reimbursement is well on its way, gaps in care still exist.

It is these gaps that companies like IBM's Watson Health are addressing. Watson for Oncology, for example, is being "trained" by Memorial Sloan Kettering to create expertise around oncology. Cancers are very different, as are treatments. No oncologist can understand the intricacies to personalize care because there are so many different types of treatments based on a trove of information around patient reaction to those treatments. If Watson can distill those using TDM and machine learning, the value is significant because the result is evidence-based, personalized treatment options as well as money saved by eliminating unnecessary tests.

What tomorrow looks like is technology driven: better modeling, methodologies, data, and computing power. These new technologies will provide new types of information:

- Genomics, such as complete genomic maps that can indicate the appropriate treatment for an individual, will support personalized medicine.
- Game theory in wellness and services in the medical arena – understanding how a person feels pointing to triggers for patient understanding of, and compliance with, a course of treatment.
- New types of data from sensors and wearables building on new technologies, expanding on today's functionality. Just a few of the innovations already on the market include:
  - Luna mattress covers, integrating WiFi, a microphone, and an array of sensors to monitor temperature, breathing, and heart rate, as well as ambient light and humidity. A connected smartphone app provides insights into how well a user is sleeping at night.
  - Owlet baby booties sensing vital signs, sending parent alerts, and saving lives.
  - Abilify tablets embedded with the Proteus ingestible sensor digitally recording ingestion, transmitting information to healthcare professionals.

We are experiencing the evolution of a revolution; even with enormous growth in models and analytics over the past decade, it is still early stages. In all fields, the goal is to get beyond tagging content – a human-driven (supervised) activity – to a level where unsupervised technologies create subvertical ontologies, knowledge graphs, or health graphs to help create experts in every arena. This means experts in not just medicine but pediatric oncology, for example. With the Internet of Things, the explosion of content and data will continue to be significant, and it will have to be ingested without tagging. The goal isn't to make a machine think better than a human, it is to enable humans to make better decisions – and at this point, it isn't a matter of if, it is a matter of when.

10 to Watch

Outsell's 10 to Watch are helping to take the market to where it needs to go with TDM-related initiatives, enabling and supporting text and data mining in exemplary ways.

Accenture

Accenture provides a broad range of services and solutions in strategy, consulting, digital, technology, and operations, including developing and implementing technology solutions to improve clients' productivity and efficiency. The company services 13 focused industry groups including health and energy. With nearly 2 million miles of transmission pipeline networks globally, Accenture's Intelligent Pipeline Solution utilizes data mining functionality to provide predictive analytics. In partnership with GE's Predictivity, users can improve operations and respond even more quickly to potential events. Given Accenture's resources, positive developments in commercial applications of TDM are sure to continue.

Copyright Clearance Center (CCC)

Copyright Clearance Center's RightFind XML for Mining is the first commercial tool of its kind directed at the information industry, coupling content licensing with text mining-ready content. The tool addresses research pain points with the aggregation of articles in XML, a structured format for mining, normalization of the content, and licensing that content for legal use. CCC's offering reaches the research market at a critical point in the TDM timeline, when researchers are looking for ways to plow through mountains of information. Although only at the beginning of its evolution, it is leading the way to further innovations, with early participation in the market and filling a critical position between content and technology.

Crossref

Crossref maintains a database of DOIs (digital object identifiers) for its approximately 4,000 publisher members, each with associated bibliographic metadata. The organization launched its TDM services in May 2014, providing a Crossref Metadata API for researchers to use to access the full text of content identified by DOIs across publisher sites. The API, free to users, provides automated TDM tools with direct links to full text on publisher sites, indication of whether the underlying text is TDM-enabled, and rate-limiting window identification (how many times full text can be requested over a certain time period so as not to overburden a publisher's system). It is a strong start to providing free TDM resources for researchers – an important step in furthering scientific research.

Figshare

Figshare's next-gen platform, customized for academic institutions, is an open online digital repository for researchers, with institutional functionality allowing licensees to highlight institutional or departmental research, group content in collections to create a single citable unit, and support content discovery through enhanced search, categorization, and metadata. The platform also provides data curation capabilities and administrative workflows – including compliance with funder mandates – along with flexible storage options. In this way, the platform supports institutions with a data repository containing advanced tools, functionality, and scalability, not to mention user-friendly aspects such as institutional single sign-on systems. Figshare is a leader in the field of research data management, and more developments supporting open science, including supporting TDM, are sure to come.

General Electric (GE)

GE harnesses what it calls the "Industrial Internet" to leverage connectivity and analytics, allowing customers to connect machines with large industrial data sets. Primarily used in the area of early anomaly detection in asset performance, GE's software lets users not only identify deviations from normal or desired performance well before they are visible to standard operational systems, but also draw on a collective knowledge base to answer "what if" as well as "what should I do" questions. Using GE's SmartSignal predictive analytics software, users can also identify what is going to fail, the apparent cause of the failure, and the priority of the impending failure. SmartSignal provides predictive analytics solutions to generate early warnings, allowing customers to shift to efficiently planned maintenance, thereby reducing costs and time, as well as protecting assets and, in many cases, the environment.

IBM Watson

Using machine learning and NLP, IBM Watson scans massive quantities of data – structured and unstructured – and produces answers to questions. Watson has healthcare-related offerings to accelerate discovery, including Watson Discovery Advisor for Life Sciences, creating links between data sets; Watson for Oncology, providing clinicians with evidence-based treatment options based on Memorial Sloan Kettering training; and Watson for Clinical Trial Matching, allowing doctors to generate a list of available cancer trials and as potential matches for patients.

Linguamatics

Linguamatics provides NLP-based text mining for the pharma-biotech and healthcare industries with the goal of speeding up the drug-discovery cycle and improving patient outcomes. The company partners and collaborates with companies as well as academic and governmental organizations such as ChemAxon, a cheminformatics software platform provider; Biovia (formerly Accelrys), an R&D software and services company; Knime, an open platform for data-driven innovation; and Cambridge Semantics, provider of an open, standards-based software suite called Anzo driven by semantic web technology. For content, the company partners with Thomson Reuters, CCC, and IFI Claims Patent Services. These partnerships support the functionality underlying I2E, Linguamatics' leading text mining platform, with interfaces for both occasional and expert users.

OMICtools

OMICtools is a manually curated metadatabase providing an overview of more than 4,400 web-accessible tools related to genomics, transcriptomics, proteomics, and metabolomics. Information about each tool comes from developers, the scientific literature, or voluntary submissions. The goal is to help experimental researchers and clinicians find appropriate data mining tools for their needs, developers to stay up to date and avoid redundancy, and funding agencies to ensure submitted projects are high value.

Talix

Talix's Coding InSight, built on the Talix HealthData Engine, automates analysis of vast amounts of structured and unstructured patient data to uncover missed coding opportunities prospectively as well as retrospectively, enabling healthcare providers and payers to close risk adjustment gaps by improving coding efficiency, receive reimbursements based on correct coding, and improve patient outcomes. The Talix HealthData Engine is a patient-data analytics platform using NLP, a medical taxonomy, and a coding and clinical rules database to enable mining of both structured and unstructured patient data buried in electronic medical records.

VisTrails

VisTrails is a National Science Foundation (NSF)-supported suite of visualization tools developed by Juliana Freire and Claudio Silva, associate professors of computer science at the University of Utah. The tool enables researchers to create complex visualizations, mine data, and apply other large-scale data analysis applications by automating much of the difficult and time-consuming back-end work that goes into producing them. Applications in medicine include visualizing a patient's heart or *brain* so a surgeon can see what will come under the scalpel, supporting improved patient outcomes.

Essential Actions

To take advantage of the benefits of TDM today and tomorrow, Outsell recommends the following for content providers as well as TDM tool vendors.

✓ **Build scalable solutions or look to partner.**

Information providers must ensure that TDM initiatives can scale as technologies develop. Continuously populate, link, or combine databases to provide monetizable longitudinal data, and stay on top of technology advancements that will support ever-growing data sets for diverse industries. Understanding how millennials or other demographic groups or markets access information is also key: Think about designing interfaces that support user preferences, such as different types of visualization or tables. Partnering with TDM tool vendors – having them shoulder the risks of scalability and provide consistency of results across different bodies of content – may be a better strategy than internal development.

✓ **Think big picture – the time is now.**

Copyright restrictions protect, but they also limit. Be sure protections aren't creating barriers to research advancements based on "what if" scenarios. Don't wait until technology catches up with content proliferation; now is the time to participate in those advancements, including increased access. Along those lines, as TDM exposes underlying content to scrutiny, ensure reliability through reproducibility, equating to high value for users.

✓ Support data standards.

It is a never-ending song – the tune may change but the words stay the same: Support standards. Understanding the importance of standards and keeping abreast of changes, processes, and methodologies behind TDM is critical to developing and maintaining value in content assets. Standards will allow optimal delivery of that value, which means requiring submission of data according to community-defined standards, as well as more vendor collaboration around establishing those standards.

✓ Prioritize data integrity.

Integrity of data is essential to making decisions based on that data. That means understanding what is in the data collected and ensuring as much as possible that those using it can be confident that it is reliable and trustworthy. Although peer review is under intense scrutiny in today's market, it is imperative that the scrutiny not slow progress in establishing processes for reviewing data supporting a publication. Work with organizations like Dryad and the Research Data Alliance to accomplish this, to ensure the value of scientific content to researchers, and other information consumers, by raising confidence levels in that content.

✓ Ensure data security.

Ensure data security by establishing data governance, thereby ensuring adherence to privacy and security regulations. Customer data, for example, is sensitive not only to customers, but it is also sensitive to an organization's marketing and sales strategies. Carefully monitor research data, such as clinical trials information, to ensure patient privacy and prevent data breaches. Data governance committees can help ensure adherence through familiarity with regulations, laws, and administrative policies.

Imperatives for Information Managers

Outsell recommends the following actions for information managers as they support investigators in their research activities.

✓ Seek out data experts.

Hire or collaborate with data scientists to better support research efforts in academic and corporate environments. Data experts not only identify potential issues with data sets or cross-integration of data, they can also lead efforts to confirm application of standards to data. Quality control is essential; experts can verify if a data set can and should be used in a study.

✓ Collaborate externally and internally.

To advance understanding around TDM, its standards, and its uses, it is imperative that managers maintain a continued dialogue, internally and externally. Doing so will advance the cause of standards as well as comprehension of how and why TDM benefits and advances knowledge. It also means giving input to information providers about how best to serve their researcher customers, both through licensing discussions as well as suggesting additional services to support TDM activities.

✓ Keep abreast of developments.

To support researchers fully requires a basic awareness of what is happening in the TDM space – what technologies are available or in development, who are leaders in the space. An understanding of these dynamics is critical to providing insight into the future of research. By deepening that understanding, information managers can better support researchers who may not have the same level of understanding.

✓ Negotiate TDM licensing rights.

Because of the effort involved in negotiating permissions with individual information vendors to mine content, information managers must remember: This is a rapidly growing technology that will serve future researchers. To avoid going back to vendors for TDM rights, routinely include requests for text mining rights of full-text articles in XML format when negotiating library licenses. This will serve institutions and researchers, and it will also reinforce with vendors the importance of TDM.

✓ Employ data mining internally.

Develop new data sets from collections that can be gathered and analyzed. Institutions now have opportunities to measure something new, something previously unachievable due to software and hardware constraints. External vendors can digitize special collections and employ linked data technologies to support TDM activities, expanding offerings and creating added value for library customers.

Related Research

Reports

Data Business Fundamentals	November 24, 2015
IBM Watson – A Hammer in Search of Nails	May 27, 2015
Health Information Markets in the US: The Changing Landscape	June 25, 2013
Five Technologies to Watch 2012–2013	January 25, 2012

Insights

Online Data Security Threats Drive New Legal Solutions	October 6, 2015
Chemical Abstracts Service Presents PatentPak Interactive Chemistry Viewer	September 2, 2015
Public Sector Open Data: Trying But Must Try Harder	June 17, 2015
Copyright Clearance Center’s Newest Service: RightFind XML for Mining	June 11, 2015
Text and Data Mining Are Set to Become a Reality in 2014	January 28, 2014

Deni Auclair
VP & Lead Analyst
T. +1-508-785-8384
dauclair@outsellinc.com

See additional reports published in *all coverage areas.*

Does this report meet your needs? Provide feedback **HERE**

About Outsell

The rapid convergence of information, media and technology is reshaping businesses every day. Enter Outsell, Inc., the only research and advisory firm focusing on these three sectors. As the trusted advisor to executives, our analysts turn complexity into clarity, and provide the facts and insights necessary to make the right decisions. Our proven blend of big data, research, proprietary intelligence, and exclusive leadership communities produces tangible results and a strong ROI. We promise to deliver "wow" and ensure clients stay more focused, save time, and grow revenue in a fast-changing digital world.

www.outsellinc.com
info@outsellinc.com

Burlingame, CA USA
+1 650-342-6060

London, United Kingdom
+44 (0)20 8090 6590

The information, analysis, and opinions (the "Content") contained herein are based on the qualitative and quantitative research methods of Outsell, Inc. and its staff's extensive professional expertise in the industry. Outsell has used its best efforts and judgment in the compilation and presentation of the Content and to ensure to the best of its ability that the Content is accurate as of the date published. However, the industry information covered by this report is subject to rapid change. Outsell makes no representations or warranties, express or implied, concerning or relating to the accuracy of the Content in this report and Outsell assumes no liability related to claims concerning the Content of this report.

OUTSELL®

Advancing the Business of Information