SEMALYTIX

Dark Data Understanding

White Paper

Taming Dark Data: Data-Driven Insights For Smart Decision Making

I Dark Data

Data is growing at an impressive, exponential rate. With exabytes of new data being created every single day [1], it becomes increasingly challenging to make sense of this ever-growing stream of data. There are several indicators which clearly corroborate that the exponential growth of data will continue:

- **Volume:** The data streams already generated today are huge. For example, only one hour of customer transaction data at Wal-Mart, corresponding to 2.5 petabytes, provides 167 times the amount of data housed by the Library of Congress [3].

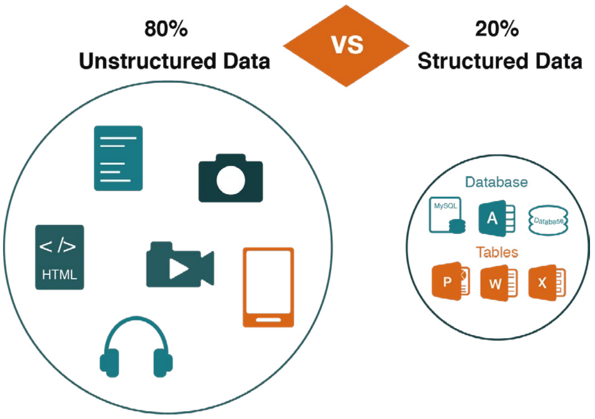
- **Growth Rate:** 90% of the data available today has been generated in the past two years [4]. The International Data Corporation (IDC) estimates that all digital data created, replicated or consumed will grow by a factor of 30 between 2005 and 2020, doubling every two years. By 2020, it is assumed that there will be over 40 trillion gigabytes of digital data, corresponding to 5,200 gigabytes per person on Earth [5].

- **Internet of Things:** Cisco estimates that currently less than 1% of physical objects are connected to IP networks. However, this is estimated to change radically to up to 50 billion devices being connected to the Internet by 2020, corresponding to between 6 and 7 devices per person on the planet [6]. These 50 billion devices will continuously generate data at an unprecedented scale. According to other analysts, this figure should be considered a lower bound at best, with current estimates, e.g. by Morgan Stanley, predicting 75 billion devices by 2020.[1]

Data has been referred to as the "*oil*" of the new economy [2]. The data available at an organisation represents an important intangible asset that can be exploited for decision making, both on operational and strategic levels. Data-derived insights have an important economic impact and value when leveraged correctly and in a timely manner. Using data-derived insights allows to move from decision making based on intuition to making informed and sound decisions. Leveraging data analytics for decision making can make the difference and set a company ahead of their competitors.

Dark data is the term used to describe information assets that organizations collect, process and store during regular business activities, but generally fail to reuse for other purposes. Dark data comprises mainly unstructured data (text, sensor data, videos, images, etc.) that can not be processed using standard operations of data including filtering, projecting, joining, aggregating, averaging, etc. Particularly challenging as dark data is textual content as it requires so called machine reading, that is the ability of a machine to understand natural language text.

Estimates suggest that only **half a percent of existing data is being analysed to generate insights** [5]. Furthermore, the **vast majority of existing data is unstructured and machine-generated** [7]. Structured data is typically found in databases. It is data that has been formatted into pre-defined schemas. Unstructured data includes sensor data, video, audio and text data, that has not been structured into a database according to a fixed schema. It is generally estimated that unstructured data amounts to about 80% of the data that is available at an organisation. In-depth analysis of unstructured data is typically hard to obtain in an automatic manner. Due the general lack of openly available analysis capabilities, unstructured data are often not analysed at all or only at shallow depth, which renders large amounts of unstructured data unavailable for decision making.



[1] http://www.businessinsider.com/75-billion-devices-will-be-connected-to-the-internet-by-2020-2013-10?IR=T

However, there is enormous value in unstructured data and failure in setting up processes to exploit insights from these assets might result in a loss of market to competitors. Take customer relationship management as an example. Most data about customers is unstructured, hidden in social media posts, blogs, recordings or transcripts of calls, etc. Analysing this rich data is key to infer 360° views of customers, e.g. in order to better understand customer journeys and customer value, prevent customer loss, develop targeted communication and marketing strategies. Such a holistic understanding is key to developing personalized offers supporting customer retention or up-selling.

Recently, the term **dark data** has emerged to describe data that is neither systematically collected nor used to derive any insights[2]. Yet, when appropriately analysed, dark data can have a huge impact and potential for organisations[3]. Even handling structured data may require the analysis of textual (hence, unstructured) data. Large datasets very rarely consist of purely numerical data; in fact, they always include natural language components such as column heads in databases, textual data in table content, metadata, documentation, summaries or links to documents.

The three classical challenges associated with processing unstructured data have become to be known as the „three V's": Volume, Variety and Velocity[4]. As 80% of the data relevant to organisations is unstructured, the *volume* of data to be analysed is potentially vast, ranging in the order of terabytes or exabytes depending on the size of the organisation. *Variety* refers to the challenge that data might come in very different formats, follow different standards and methods for data collection, contain implicit assumptions, even representing different ways of measurement, thus being hard to compare and aggregate. The *velocity* challenge refers to the problem that unstructured data might arrive in a rapid stream and require real-time processing, representing a computational and storage challenge. In addition to the three V's, one challenge is typically ignored: **Contextualization and interpretation**. Effective data analysis requires data interpretation in the context of the organisation as a whole to determine relevance, meaning and impact to the organisation. Our mission at Semalytix is to support our clients in generating actionable knowledge based on data-derived insights, focusing on dark data, in particular textual and semi-structured data. For this, Semalytix leverages the latest advances in artificial intelligence, machine learning, natural language processing and semantic technologies to develop tools which enable the generation of insights by analysing huge and heterogeneous data streams in real time.

The Three V's of Unstructured Data

Volume
The huge amount and growth of unstructured data can overtake traditional storage solutions

Variety
Traditional data management can't handle the changeable nature of big data

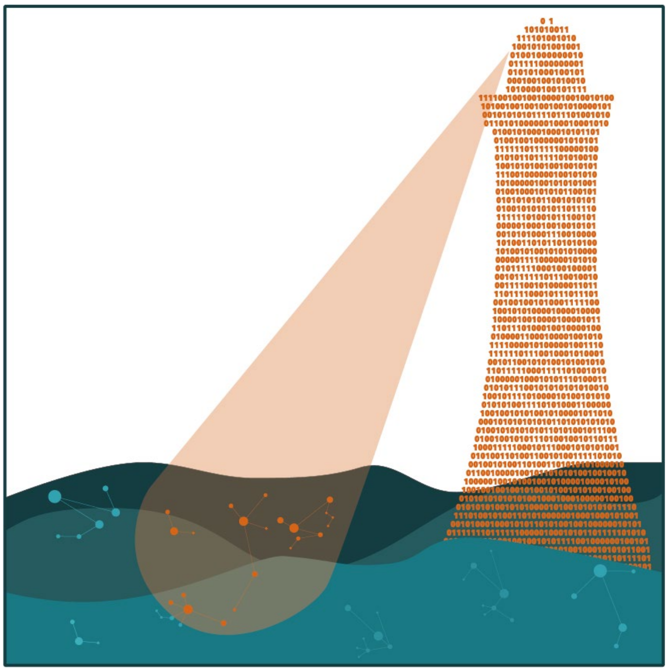
Velocity
Data is generated at an ongoing flow, making it harder to manage

The solutions of Semalytix are provided in an „Analytics-as-a-service" fashion, empowering our customers to get the analytics and insights they need by way of easy-to-understand and easy-to-reuse analytics, turning data into insights in real time to support decision making. Our solutions are deployed as a web application running on the cloud or on a virtual machine at the customers' organisation, depending on customer needs and preferences regarding data protection.

[2] https://www.gartner.com/it-glossary/dark-data

[3] https://www.forbes.com/sites/tomcoughlin/2017/07/24/analysis-of-dark-data-provides-market-advantages/#ab4dc10872b9

[4] http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/



II BEACON: Shedding light on your decision route

Our flagship product at Semalytix is **BEACON**. BEACON is delivered via a software-as-a-service model by a web application providing access to analytics in the form of intuitive and easy-to-use dashboards that support decision making in customer organisations. BEACON monitors a number of external (forums, social media, blogs) as well as company-internal data sources by a scalable data ingestion infrastructure. It applies state-of the-art machine learning and natural language processing technologies to read between the lines of unstructured data sources and extract the content and messages that are relevant to decision making. It relies on semantic technologies in order to disambiguate contents and assess their relevance, yielding accurate and precise analytics. BEACON comes in the form of a Google-like interface in order to provide a command console to extract insights from all relevant data sources. BEACON rests on five key principles:

- **Semantics:** The Semalytix technology stack extracts insights by computing meaning beyond individual words. Our semantic technologies normalize meaning by abstracting from particular word choices and disambiguating word meaning. Thus, we are able to capture meaning that is hidden below the textual surface, which results in higher accuracy than techniques that merely analyse text as collections of words.

- **Real-time:** Semalytix monitors a number of rapidly changing data sources in real time, ingesting and analysing the content. Thus, we are able to react to events and changes within seconds.

- **Relevance:** By applying state-of-the-art machine learning techniques, in particular classification techniques, Semalytix technology can effectively and efficiently narrow down the content to the information that is relevant for a particular need.

- **Actionability:** Each analysis / dashboard comes with a set of associated actions that are available as a direct option to support swift reaction or issue resolution.

- **Self-service:** The analytics-as-a-service platform developed by Semalytix empowers users to customize or create own dashboards and analyses easily.

BEACON is a web-based data management platform that relies on a Google-like interface to provide access to relevant data stored in a large knowledge-graph. New data is ingested into the knowledge-graph in real time from multiple sources. The data is analysed semantically using state-of-the-art natural language processing and machine learning to determine relevance to a given domain and task. Key extracted facts are stored in the knowledge graph to support generation of actionable insights delivered through dashboards that are updated automatically as new data arrives.

II.1 Machine reading between the lines with AI

Our solutions are powered by the latest advances in machine learning and natural language processing and knowledge representation. Our solutions rely on *machine learning*, in particular *deep learning*, as well as natural language processing technologies to extract insights from natural language text. This enables us to extract relevant knowledge such as facts mentioned about companies, wishes, needs and complaints by customers, the sentiment towards products or services as expressed in social media, etc. Our technologies are suited to find the needle in the haystack: the little piece of knowledge that can be crucial to change your business or strategy but that you could never find as the amount of data was too big; the little piece of evidence that allows you to up-sell a product to a customer or helps you to take an action in order to prevent a client from moving to a competitor.

Textual content cannot be merely regarded as a collection of words. There is a deeper message or content hidden behind the words that needs to be brought to the surface. This is where our technology, leveraging the state-of-the-art in information extraction, excels and makes a difference.

We deploy a stack of natural language processing components based on machine learned classifiers that are applied along a hierarchy to determine the relevance of content in an increasingly fine-grained fashion. This allows to shift the granularity of analyses from fine-grained over coarse-grained to a bird’s eye perspective depending on what view is needed for a given purpose.

Machine learning methods can be divided into supervised and unsupervised techniques. Unsupervised techniques are used to discover regularities or groups in data, e.g. in order to cluster customers into segments. Supervised learning techniques typically learn a classification function or model from so called training data consisting of data points with a correct classification that can then be applied to unseen examples. A classifier could for example be trained on a binary classification task to learn to distinguish spam content from non-spam content as a binary classification task. Deep learning methods are typically used for classification tasks and rely on so called neural networks that are hierarchically organized into multi-level classifier architectures. Deep learning methods have been shown to outperform standard machine learning classifiers on a number of tasks.

The machine learned classifiers are trained using annotated data to assess relevance of content for a given domain, thus performing implicit disambiguation. A post on Twitter, for example, might mention the abbreviation ILD, which can either stand for *Interstitial Lung Diseases* or for the *Institute for Liberty and Democracy*. Finding out which meaning is intended in a given context is key to ensure high analytical accuracy. The Semalytix technology stack performs disambiguation by relying on a complex classifier architecture that focuses only on the relevant meanings in a given application context. Sentiment classifiers are used to detect the sentiment towards a product that is expressed in a particular piece of text. Most importantly, the sentiment classifiers are trained to predict sentiment in a domain-specific manner. In the pharma domain, for instance, a statement such as: *“FDA approves new indication for Jardiance (empagliflozin)“* would be recognized as neutral for standard sentiment analysis systems. Instead, Semalytix trains sentiment analysers that are domain- and industry-specific, thus recognizing that the above tweet has a positive connotation in the pharma domain. However, information extraction methods can not only be used to extract sentiment, but also to extract comparative statements, such as the following: *„Pfizer‘s Torisel fails to improve* ***PFS*** *sig. more than Bayer,* ***Onyx***‘s *Nexavar in late-stage renal cancer study.“*

II.2 Powered by Semantics and Knowledge Graphs

At the core of our solutions is a *knowledge graph* that stores all the relevant knowledge that is extracted from a large number of heterogeneous sources. The knowledge graph functions as the backbone that is used to generate all analytics and dashboards.

The knowledge graph is dynamic and evolves as more and more information is added. It thus acts like a brain that contains all the knowledge relevant to an organization. As it evolves, new connections are added, others are strengthened, weakened or even eliminated if obsolete.

With increasing amounts of data being processed, the knowledge graph constantly accumulates more information. The resulting increase in size fosters the delivery of more — and more accurate — insights by means of linking and aggregating all individual pieces of knowledge available. The graph stores all the information about relevant entities in a given application domain, comprising competing organisations, customers, products, relations between these as well as all the relevant communication originating and affecting these entities.

The Semalytix knowledge graph comprises the following pieces of knowledge:

Knowledge graphs represent knowledge by way of triples or tuples of the form (subject, predicate, object) representing the fact that a given subject (e.g. Bayer Healthcare) stands in a certain relation represented by the predicate (e.g. markets) to a certain object (e.g. Aspirin). Knowledge graphs rely on open standards such as the Resource Description Framework (RDF) to represent such triples that altogether form a connected graph. The nodes in the graph represent so called resources, that is entities of relevance that we want to talk about (as Bayer Healthcare and Aspirin) and edges represent triples such as the above and are labeled with the corresponding predicate (markets). In this way, each resource can be connected to many other resources to represent different statements about the given resource. Instead of limiting two triples, the knowledge graphs at Semalytix can represent any relation between an arbitrary number of resources.

- **Entities** representing real things or events that exist in the world, such as a drug, e.g. Aspirin, or an event, e.g. the annual meeting of the American Society for Clinical Oncology (ASCO).

- **Facts** include objectively true knowledge, e.g. that the compound behind Aspirin is *acetylsalicylic acid*, that Bayer markets *acetylsalicylic acid* under the trademark „Aspirin“, etc.

- **Mentions** recording that some stakeholder mentioned a particular entity in a particular data source, such as the mention of the side-effect *stomach bleeding* in a Tweet about Aspirin.

- **Statements** or **arguments** raised by stakeholders towards an entity, e.g. the claim that *Ibuprofen* is superior to *Aspirin*, because the latter one has a higher likelihood of producing gastrointestinal bleeding.

- A **Lexicon** comprising relations between different words and terms, including the knowledge that ASA is an abbreviation for *“acetylsalicylic acid“*, that *„acetylsalicylic acid“* and *„acetylsalicylate“* are synonyms and that stomach bleeding and gastro-intestinal bleeding are strongly connected and semantically related side effects.

- **Provenance information:** For every mention and statement recorded in the knowledge graph we keep direct reference to the data source in which the mention or statement appeared to allow tracking of every analysis to a particular data source.

- **Confidence:** For every piece of knowledge in the graph, we record the confidence that a certain machine learning classifier had in its extraction in order to estimate the degree of reliablity for every piece of knowledge.

The knowledge graph relies on cutting-edge knowledge representation and data models such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Following the RDF data model, the knowledge is stored in triples of the form according to a given ontology:

Bayer Healthcare **markets** **Aspirin**

The knowledge graph is stored in an efficient and scalable distributed data architecture. Thus, it can scale to arbitrary data sizes and adapt dynamically to those, addressing the volume challenge of unstructured data.

II.3 Actionable insights in real-time

In order to address the variety and velocity challenge, BEACON comes with a set of dedicated crawlers and connectors for injecting potentially relevant data sources into the knowledge graph. The crawling and ingestion infrastructure can ingest tweets, blogs and forum posts, Word and PDF documents as well as spreadsheet data. The crawlers ingest data at the speed of up to 3000 documents per minute.

For any single data point that is ingested from the data stream, our infrastructure allows to determine its specific impact on corporate decision-making, and to trigger actions as previously specified by a user (see section below on self-service customizable analytics). As soon as new data arrives that is run through the natural language processing pipeline, dashboards are updated and trigger conditions are evaluated to decide whether the new data point has an impact on an existing analysis or dashboard. In case the dashboard or analysis is critical, immediate alerts can be triggered in order to support swift decision making.

We regard analytics as being created for the purpose of acting upon them. So each analysis /dashboard comes with a range of associated potential actions ranging from sending the dashboard via e-mail to a designated analyst in oder to inspect the underlying sources, drill down into potential causes of the development, to writing a targeted message to be sent to a group of relevant stakeholders or customers. By doing so, our infrastructure enables our customers to react early and timely to external and internal events as reflected in the data.

II.4 Self-service customizable analytics

There are two important obstacles hindering swift and timely decision making on the basis of data-derived analytics [8]. For one, line-of-business decision makers and data analysts are burdened with old-school legacy tools including Excel spreadsheets that do not support the generation of analytics from multiple sources and in real-time. Secondly, legacy tools do not support the easy creation of dashboards or analyses needed for a particular task. The creation of new analyses requires depositing a request to the IT or data analytics department who then implement the request as resources allow for it. This is very far from being able to produce the relevant analyses at the time they are really needed. In fact, a survey carried out by the Harvard Business Review shows that the majority of decision makers are dissatisfied with current tools for tasks such as data preparation, data integration, advanced analytics and data visualization [6].

In contrast, Semalytix BEACON provides a command control environment to decision makers in which they can easily create precise, customized analytics. This involves specifying the temporal regularity at which the analytics should be recomputed or indicating actions and trigger conditions that should be executed as soon as new data arrives that is relevant for the task at hand. Overall, this leads to faster analysis creation and quicker decision making at reduced costs.

III Markets and use cases

The solutions of Semalytix are applicable to a number of markets, including in particular the pharmaceutical industry, finance, manufacturing and service industries.

Our solution portfolio can be used to improve the following services:

- **Marketing:** Monitor and manage brand reputation, quantify the impact of your communication and marketing strategies; engage with actual or potential customers directly via micro-targeting.

- **Customer Relationship Management:** Get insights into customer journeys, identify customer dissatisfaction earlier and solve customer problems faster.

- **Personalized Recommendations / Products:** Offer personalized products to groups of customers or individuals; leverage data to up- or cross-sell.

- **Forecasting:** Use analytics technology to forecast sales by customer types and geographic segment.

- **Product Development:** exploit customer insights to develop new products / business models and to optimize pricing schemes.

- **Portfolio Management & Risk Assessment:** Predict the performance of assets and optimize investment decision making.

- **Due Diligence:** Identify relevant variables, risks and opportunities that support reaching an informed decision on acquisition or investment.

- **IT / Internal Operations:** Create rich, healthy, information life cycles within or on top of internal data platforms. Turn tedious documentation tasks into valid and revenue generating business intelligence proceses.

We discuss in more detail the impact that our solutions can make in four industries: 1) pharmaceutical industry, 2) finance and banking, 3) customer relationship management, and 4) due diligence.

III.1 Pharmaceutical Industry

The pharmaceutical industry is facing a number of challenges. On the one hand, while now stabilizing, costs for drug development have been increasing by a factor of 4 over the last 20 years. At the same time, there is increasing pressure by drug payers and patients to lower costs while sales volume is decreasing due to the appearance of generics. Further challenges include increased compliance costs, changing distribution channels and the move from sales-oriented to outcome-oriented payment models. In addition, the adoption of *biosimilars* has been slower than expected.

Nevertheless, there are huge opportunities associated to some of these challenges. On the one hand, increasing regulation and the need to demonstrate real-world evidence call for better approaches to collect and manage data assets. The changing distribution channels are creating pressure to leverage new communication channels for marketing and increasing stakeholder engagement in alternative channels (e.g. social media). To counterfeit increasing costs and reduced sales due to generics, R&D processes need to be streamlined. Personalized medicine as a cost-, data- and knowledge-intense model will require corresponding investments.

In the light of these opportunities, the need for data-analytic technologies is higher than ever. The market size for 3rd party analytics in the pharma industry is in fact expected to move from $120-200 Million in 2013 to between $680 Million and $1.1 Billion in 2020, corresponding to a Compound Annual Growth Rate (CAGR) of 23-27% [9].

Analytics are expected to have a positive impact on the following dimensions:

- **Regulatory Compliance:** Integrated approaches to data management can help to reduce the costs for regulatory compliance.

- **Marketing / Sales Support:** Learning from the information collected from competitors can help to refine own market strategy. Deeper insights into customers can be obtained by integrating / aggregating dark data coming from different sources, including but not limited to medical claims, health assessments, health screenings, wellness activities, pharmacy claims as well as general customer information from online / social media sources.

- **Product / Service Enhancement:** The integration of clinical trial data, hospital records, physical notes, research papers, patient demographic and social media information can contribute to deliver insights that aid early stages of drug development and testing. Health outcome analytics can help guide medical practitioners on the best medical process / approach available.

- **Clinical Pathway Delivery:** Following a patient's disease and treatments increases the probability of improved outcomes and increased productivity. Data-derived insights can contribute to reduce delivery risks by improving pathway design.

- **Lifetime Patient Data Management:** Lifetime patient datasets are useful to develop an integrated understanding of patients and help de-risk the process of discovery through a better focus on unmet needs as well as to improve selection of biomarker / subpopulations, and support faster identification of trial patients.

- **Digital Platforms:** Digital platforms can create a step change in the cost of engaging with patients and physicians, both delivering significant new value and helping to lower costs.

- **Real-World-Evidence:** Real-world evidence can complement controlled trials and clinical studies to better understand the values of drugs or (combined) therapies. Relevant data to be considered include non-interventional studies, observational studies, social media and forums, patient records, etc. Analysis of real-world evidence supports better analysis of the outcomes of therapies in daily healthcare, provides insights on which therapies are most effective, better understanding of expected outcomes, etc.

Bringing about the above added values requires better strategies to collect, integrate and analyze data to support decision making. It also requires consideration of non-standard sources to gather real-world evidence and customer/patient data as well as appropriate approaches to engage with different stakeholders via new communication channels (e.g. social media).

III.2 Finance & Banking

The finance industry is living in disruptive times, with higher regulatory pressure and strong competition from pure online services. New regulatory and compliance requirements are placing greater emphasis on governance and risk reporting, driving the need for deeper and more transparent analyses across global organisations. Financial services companies are seeking to leverage large amounts of consumer data across multiple service delivery channels (branch, web, mobile) to support new predictive analytic models supporting the discovery of consumer behavior patterns and thus increase conversion rates. Predictive credit risk models that tap into large amounts of data consisting of historical payment behavior are being adopted in consumer and commercial collections practices to help prioritize activities by determining the propensity for delinquency or payment. Mobile applications and internet-connected devices such as tablets and smartphone are creating greater pressure on the ability of technology infrastructures and networks to consume, index and integrate structured and unstructured data from a variety of sources. A tend from in-person to online banking resulting from the ease and affordability of executing financial transactions can be clearly observed. The availability of new data sources such as data from social media, blogs and other news feeds offer significant new opportunities. As with all online markets, banking is competitive and banks are interested in using every opportunity, identified through data, to cross-sell and up-sell customers.

In this situation, there is wide agreement that analytics is the key to develop new business models and services [10] [11] [12]. In fact, 76% of banks are convinced that the business driver for embracing big data is to enhance customer engagement, retention and loyalty [15]. 71% of banks say that in order to increase revenue, they need to better understand customers and big customer data will help them [15].

However, banks are currently not able to deliver effective personalized service as the level of customer intelligence is very low. The following business drivers are can be expected to play a key role in future banking models and services:

- Leverage big data to get a 360 degree view of each customer
- Drive revenues with one-to-one targeting and personalized offers in real-time
- Achieve greater customer loyalty with personalized retention offers

Other relevant areas for analytics are: Credit scoring/mortgage portfolio valuation, back testing, fraud detection, modelling market and consumer risk, supporting regulatory compliance. In general, our technology can be deployed to deliver predictive analytics, that is to predict how assets or shares will evolve or which events/effects are likely to happen in the future or to predict how likely a customer is to switch to another service or terminate a contract.

III.3 Due Diligence for mergers and acquisitions

Mergers and acquisitions require a substantial amount of due diligence from the side of the buying entity. Before committing to the transaction, the buyer needs to have complete information on the target company in terms of its obligations, liabilities, problematic contracts, litigation risks as well as intellectual property issues.

In particular, decision making within M&A requires careful scrutiny of the bought company along the following dimensions [13]:

- Financial Metrics and Matters
- Intellectual Property / Technology
- Customers / Sales
- Strategic Fit
- Material Contracts
- Employee / Management Issues
- Litigation
- Tax Matters
- Regulatory Issues
- Insurance Policies

A thorough analysis of a target company along all the aspects mentioned above represents a huge effort as it requires manually checking thousands and thousands of documents. A typical M&A data room consists of more than 30,000 documents. Relevant information might be contained in company agreements, employee contracts, rental, sales and lease contracts, credit agreements, proof-of-ownership documents, guarantees, property collaterals, maintenance documents, etc. Due to time constraints, only around 10% of the relevant documents can typically be examined.

Artificial intelligence techniques can support due diligence processes by providing intelligent / semantic search functionality, intelligent analysis of content, identification and extraction of relevant clauses (e.g. change-of-control clauses, red-flag clauses or deal breaker clauses, etc.), information classification, etc.

Machine reading techniques as developed by Semalytix can help here to support lawyers in effectively performing a more thorough / complete analysis in the same amount of time, to the benefit of both those performing the due diligence as well as their customers [14].

III.4 Customer Relationship Management (CRM)

In a recent report, the International Data Corporation (IDC) estimates that artificial intelligence (AI) technology applied to CRM might boost global business revenue in the orders of $1.1 trillion from 2017 to 2021 [15].

In particular, AI-driven CRM might lead to the creation of 800,000 direct jobs, and 2 Million of indirect jobs. The year 2018 is likely to turn out to be the mayor year for AI adoption.
On the basis of a survey, IDC has identified the top CRM categories where AI technology will benefit the most:

- Customer support (43%)
- Customer billing, inventory, logistics (41%)
- Partner management (41%)
- Customer influenced product or service design (40%)
- Digital commerce (40%)
- Product or service marketing (39% of respondents)
- Marketing operations (39%)
- Product or service pricing, finance (38%)
- Customer analytics (38%)
- Corporate marketing, branding, advertising (30% of respondents)

The most important use case for use of AI in CRM include:

- Email marketing (87%)
- Sales and marketing lead scoring (83% of responding AI adopters)
- Customer service case classification / routing (83%)
- Sales opportunity scoring (80%)
- Chatbots for customer service or product selection (75%)
- Cross-selling and upselling (68%)
- Fraud Detection (64%)
- Sales forecasting (61%)
- Credit risk scoring (61%)

The deployment of AI to support the above mentioned use cases will bring about huge economic benefits and surely represent a game changer with respect to traditional CRM methods. In the survey by IDC, 41% respondents answered that they would adopt AI within CRM solutions within two years from the data of the survey, that is by 2019. Semalytix technology can support to the analysis of heterogeneous customer-related data sources in real-time and thus to the development of holistic customer views that will enable automatization of many of the tasks and use cases mentioned above.

IV Summary

Dark and unstructured data comprises about 80% of the data that is available within organisations. It is data that is rarely exploited to generate insights as it is not available in a form readily accessible for data analytic tools. Yet, the potential value of dark data is huge. Semalytix develops methods that allow to analyse unstructured data, in particular textual data, at a large scale in order to generate actionable insights that support decision making. The main product of Semalytix is BEACON, a web-based platform and data management console that relies on a simple Google-like interface to provide relevant and actionable insights in real-time in the form of dashboards that update themselves as new data is ingested. BEACON monitors multiple and heterogeneous external or corporate data sources and channels in real time, analyses the data using cutting-edge machine reading technology and stores the relevant knowledge in a large knowledge graph that evolves as more and more connections are added, strengthened or weakened. The technology developed by Semalytix can act as an enabler for key business drivers and use cases in the pharmaceutical industry, in financial services and the banking industry, due diligence services as well as in customer relationship management. Taken together, the market for AI-enabled analytics in the above-mentioned domains is a multi-billion market with many opportunities for Semalytix and other players offering data analytics services. So far, the focus of Semalytix has been on developing solutions for the pharmaceutical industry. Semalytix is seeking partners to develop innovative solutions for the other industries identified in this white paper as being key beneficiaries from AI-enabled and semantics-based data analytic services.

About the authors

- **Prof. Dr. Philipp Cimiano**

Prof. Dr. Philipp Cimiano is co-founder of Semalytix. He is professor for computer science at Bielefeld University and leader of the semantic computing group. He has over 15 years of expertise in the development of cutting edge semantic technologies and artificial intelligence systems. Building on this expertise, at Semalytix he is responsible for strategy definition, outreach and academic relations.

- **Dr. Matthias Hartung**

Dr. Matthias Hartung is co-founder of Semalytix. With his background in computational linguistics, Matthias oversees the natural language processing pipeline underlying the Semalytix text understanding capabilities. In his role as Chief Research & Development Officer (CRDO), he is responsible for pulling state-of-the-art solutions into the Semalytix technology stack.

- **Janik Jaskolski**

Janik Jaskolski is a co-founder and CEO of Semalytix. His background lies in Cognitive Computer Science and Intelligent Systems. He has overseen the early development of prototypical solutions in Semalytix. His role has since evolved into managing the rapid growth of the company and he holds executive responsibilities for various product-related activities and designs.

References

[1]
Robert Pepper and John Garrity. The internet of everything: How the network unleashes the benefits of big data. In: The Global Information Technology Report 2014. World Economic Forum, 2014.
http://www3.weforum.org/docs/GITR/2014/GITR_Chapter1.2_2014.pdf

[2]
M. Palmer. Data is the new oil, November 2006.
http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

[3]
Beñat Bilbao-Osorio and Soumitra Dutta and Bruno Lanvin (eds.), The Global Information Technology Report, World Economic Forum, 2014.
http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf

[4]
SINTEF. Big data, for better or worse: 90% of world's data generated over last two years. ScienceDaily, May 2013.
https://www.sciencedaily.com/releases/2013/05/130522085217.htm

[5]
J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, December 2012.
https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[6], [6]
Dave Evans. The Internet of Things: How the Next Evolution of the Internet Is Changing Everything, April 2011.
https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[7]
Christine Taylor, Structured vs. Unstructured Data, Datamation,2017
https://www.datamation.com/big-data/structured-vs-unstructured-data.html

[8]
The untapped power of self-service data analytics, Harvard Business Review
http://pages.alteryx.com/rs/716-WAC-917/images/HBR_Alteryx_Report_6.pdf

[9]
Analytics in Pharma an Life Sciences, Everest Group, 2014
http://www.genpact.com/docs/default-source/resource-/analytics-in-pharma-and-life-sciences

[10]
Exploring Next Generation Financial Services: The Big Data Revolution, Accenture
https://www.accenture.com/t20170314T051509Z__w__/lu-en/_acnmedia/PDF-20/Accenture-Next-Generation-Financial.pdf

[11]
The Age of Analytics: Competing in a data-driven world, McKinsey Global Institute, December 2016
https://www.mckinsey.de/files/the-age-of-analytics-full-report.pdf

[12]
Daniel D. Gutierrez, Analytics: The real-world use of big data in financial services, IBM Institute for Business Value InsideBigData Guide To: Big Data for Finance,
https://insidebigdata.com/2014/09/22/insidebigdata-guide-big-data-finance/

[13]
see https://www.forbes.com/sites/allbusiness/2014/12/19/20-key-due-diligence-activities-in-a-merger-and-acquisition-transaction/#56c7b6ce4bfc

[14]
Jean Cumming, Artificial Intelligence: Due Diligence, Lexpert Magazine September 2017
http://www.lexpert.ca/article/due-diligence-30/?p=&sitecode=lex

[15], [15]
John F. Gantz, David Schubmehl, Mary Wardley, Gerry Murray, Dan Vesset, A Trillion-Dollar Boost. The Economic Impact of AI on Customer Relationship Management, IDC White Paper, June 2017
https://www.salesforce.com/content/dam/web/en_us/www/documents/white-papers/the-economic-impact-of-ai.pdf