CP7019-MANAGING BIG DATA

UNIT_I

Big data:

Big data refers to datasets whose size is beyond the ability of typical database software tool to capture, store, managed and analyze. Big data is data that goes beyond the traditional limits of data along three dimensions: i) Volume, ii) Variety, iii) Velocity

Data Volume:

Data Volume can be measured by quality of transactions, events and amount of history. Big Data isn't just a description of raw volume. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

Measuring data volume

Unit of Measure	Approximate Size	Mathematical Representation	Examples
KB = kilobyte	1,000 (10^3 or one thousand) bytes	2^{10} or 1024 bytes	A typical joke = 1KB
MB = megabyte	1,000,000 (10^6 or one million) bytes	2^{20} or 1,048,576 bytes	Complete work of Shakespeare = 5MB
GB = gigabyte	1,000,000,000 (10^9 or one billion) bytes	2^{30} or 1,073,741,824 bytes	Ten yards of books on a shelf = 1GB
TB = terabyte	1,000,000,000,000 (or 10^{12})	2^{40} or 1,099,511,627,776 bytes	All the X-rays for a large hospital = 1TB Tweets; created daily = 12+TB; U.S. Library of Congress = 235TB
PB = petabyte	1,000,000,000,000,000 (or 10^{15})	2^{50} or 1,125,899,906,842,624 bytes	All U.S. academic research libraries = 2PB Data processed in a day by Google = 24PB
EB = exabyte	1,000,000,000,000,000,000 (or 10^{18})	2^{60} or 1,152,921,504,606,846,976 bytes	Total data created in 2006 = 161EB
ZB = zettabyte	1,000,000,000,000,000,000,000 (or 10^{21})	2^{70} or 1,180,591,620,717,411,303,424 bytes	Total amount of global data expected to be 2.7 ZB by end of 2012
YB = yottabyte	1,000,000,000,000,000,000,000,000 (or 10^{24})	2^{80} or 1,208,925,819,614,629,174,706,176	Today, to save all those bytes you need a data center as big as the state of Delaware

Data Variety:

It is the assortment of data. Traditionally data, especially operational data, is "structured" as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.).

Wide variety of data:

Internet data(Social media ,Social Network-Twitter, Face book), Primary Research (Surveys, Experiences, Abservations), Secondary Research (Competitive and Market place data, Industry reports, Consumer data, Business data), Location data (Mobile device data,

Geospatial data), Image data (Video, Satellite image, Surveillance), Supply Chain data (vendor Catalogs, Pricing etc), Device data (Sensor data, RF device, Telemetry)

Structured Data

They have predefined data model and fit into relational database. Especially, operational data is "structured" as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.)

Semi-structured data

These are data that do not fit into a formal structure of data models. Semi-structured data is often a combination of different types of data that has some pattern or structure that is not as strictly defined as structured data. Semi-structured data contain tags that separate semantic elements which includes the capability to enforce hierarchies within the data.

Unstructured data

Do not have a predefined data model and /or do not fit into a relational database. Oftentimes, text, audio, video, image, geospatial, and Internet data (including click streams and log files) are considered unstructured data.

Data Velocity

Data velocity is about the speed at which data is created, accumulated, ingested, and processed. The increasing pace of the world has put demands on businesses to process information in real-time or with near real-time responses. This may mean that data is processed on the fly or while "streaming" by to make quick, real-time decisions or it may be that monthly batch processes are run inter-day to produce more timely decisions.

Why bother about Unstructured data?

- The amount of data (all data, everywhere) is doubling every two years.
- Our world is becoming more transparent. Everyone is accepting this and people don't mind parting with data that is considered sacred and private.
- Most new data is unstructured. Specifically, unstructured data represents almost 95 percent of new data, while structured data represents only 5 percent.
- Unstructured data tends to grow exponentially, unlike structured data, which tends to grow in a more linear fashion.
- Unstructured data is vastly underutilized.

Need to learn how to:

- Use Big data
- Capitalize new technology capabilities and leverage existing technology assets.
- Enable appropriate organizational change.
- Deliver fast and superior results.

Advantage of Big data Business Models:

Big data models

Improve Operational Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Reduce risks and costs	Sell to microtrends	Offer new services
Save time	Enable self service	Seize market share
Lower complexity	Improve customer experience	Incubate new ventures
Enable self service	Detect fraud	

Industry Examples of Big data:

I) Digital Marketing

i) Database Marketers Pioneers of Big data.
ii) Big data and New school of marketing.
iii) Cross channel life cycle marketing.
iv) Social and affiliate marketing.
v) Empowering marketing with social intelligence.

II) Financial Services

i) Fraud and Big data
Fraud detection framework.
ii) Risk and Big Data
Credit Risk management
Credit Risk Framework
iii) Big data and Algorithmic trading
Crunching through complex interrelated data
Intraday Risk Analytic a constant flow of big data
Calculating Risk in marketing
Benefits to other industries.

III) Big data and Advances in health care

IV) Pioneering New Frontiers in medicine

V) Advertising and Big data

i) Big data impacts on advertising market
ii) Reach, Resonance and Reaction
iii) Need to act quickly
iv) Real-time optimization
v) Complexity of measurement
vi) Content Delivery
vii) Marketing mixed modeling
viii) The Three Big data Vs in Advertising
ix) Using customer Products as a doorway

Industry Examples of Big Data

I) Digital Marketing

- Introduction
- Database Marketers, Pioneers of Big Data
- Big Data & New School of Marketing
- Cross Channel Life cycle Marketing
- Social and Affiliate Marketing
- Empowering marketing with Social Intelligence

Introduction

Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want.

Digital \marketing is easy when consumers interact with corporate` primary platform (ie. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (eg. Face book, Twitter, Google +).

One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (ie. There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions.

Big data on the web will completely transform a company's ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organizations will be able to make hundreds and thousands of smart decisions every day.

Database Marketers, Pioneers of Big Data

Database marketing is concerned with building databases containing info about individuals, using that information to better understand those individuals and communicating effectively with some of those individuals to drive business value.

Marketing databases are typically used for

i) Customer acquisition
ii) Retaining and cross-selling to existing customers which reactivates the cycle

As companies grew and systems proliferated, a situation where there was one system for one product and another for another product etc. was landed up (silos). Then companies began developing technologies to manage and duplicate data from multiple sources companies started developing software that could eliminate duplicate customer info (de-duping). This enable them to extract customer information from silos product systems, manage the info into single database, remove all the duplicates and then send direct mail to subsets of the customers in the database. Companies such as Reader's Digest and several other firms were early champions of this new kind of marketing and they used it very effectively. By the 1980's marketers developed the ability to run reports on the info in their databases which gave them better and deeper insights into buying habits and preferences of customers. Telemarketing became

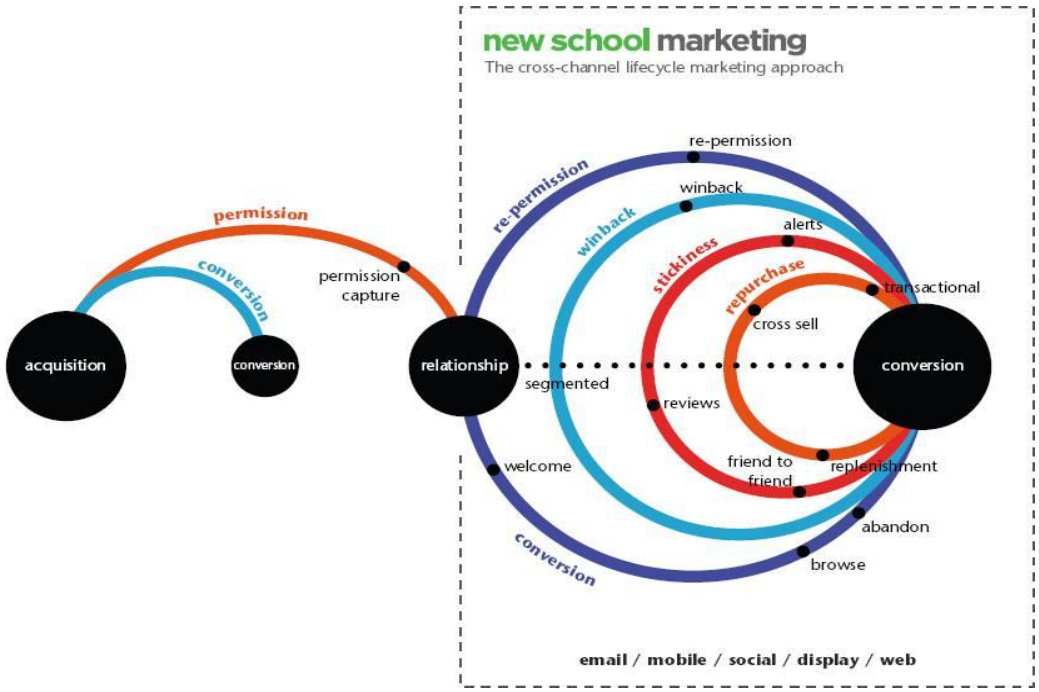
popular when marketers figured out how to feed information extracted from customer databases to call centers. In 1990's email entered the picture and marketers saw opportunities to reach customers via Internet and WWW. In the past five years there has been exponential growth in database marketing and the new scale is pushing up against the limits of technology.

Big Data & New School of Marketing

New school marketers deliver what today's consumers want ie. Relevant interactive communication across digital power channels
Digital power channels: email, mobile, social display and web.
Consumers have changed so must marketers

Right approach – Cross Channel Lifecycle Marketing

Cross-Channel Lifecycle Marketing really starts with the capture of customer permission, contact information, and preferences for multiple channels. It also requires marketers to have the right integrated marketing and customer information systems, so that (1) they can have complete understanding of customers through stated preferences and observed behavior at any given time; and (2) they can automate and optimize their programs and processes throughout the customer lifecycle. Once marketers have that, they need a practical framework for planning marketing activities. The various loops that guide marketing strategies and tactics in the Cross-Channel Lifecycle Marketing approach: conversion, repurchase, stickiness, win-back, and re-permission are shown in the following figure.



Social & Affiliate Marketing or Pay for Performance Marketing on the Internet

The concept of affiliate marketing, or pay for performance marketing on the Internet is often credited to William J. Tobin, the founder of PC Flowers & Gifts. Amazon.com launched its own affiliate program in 1996 and middleman affiliate

networks like Link-share and Commission Junction emerged preceding the 1990s Internet boom, providing the tools and technology to allow any brand to put affiliate marketing practices to use. Today, most of the major brands have a thriving affiliate program. Today, industry analysts estimate affiliate marketing to be a $3 billion industry. It's an industry that largely goes anonymous. Unlike email and banner advertising, affiliate marketing is a behind the scenes channel most consumers are unaware of.

In 2012, the emergence of the social web brings these concepts together. What only professional affiliate marketers could do prior to Facebook, Twitter, and Tumblr, now any consumer with a mouse can do. Couponmountain.com and other well known affiliate sites generate multimillion dollar yearly revenues for driving transactions for the merchants they promote. The expertise required to build, host, and run a business like Couponmountain.com is no longer needed when a consumer with zero technical or business background can now publish the same content simply by clicking "Update Status" or "Tweet." The barriers to enter the affiliate marketing industry as an affiliate no longer exist.

Empowering Marketing with Social intelligence

As a result of the growing popularity and use of social media around the world and across nearly every demographic, the amount of user-generated content—or "big data"—created is immense, and continues growing exponentially. Millions of status updates, blog posts, photographs, and videos are shared every second. Successful organizations will not only need to identify the information relevant to their company and products—but also be able to dissect it, make sense of it, and respond to it—in real time and on a continuous basis, drawing business intelligence—or insights—that help predict likely future customer behavior. Very intelligent software is required to parse all that social data to define things like the sentiment of a post.

Marketers now have the opportunity to mine social conversations for purchase intent and brand lift through Big Data. So, marketers can communicate with consumers regardless of the channel. Since this data is captured in real-time, Big Data is forcing marketing organizations to quickly optimize media and message. Since this data provides details on all aspects of consumer behavior, companies are eliminating silos within the organization to prescription across channels, across media, and across the path to purchase.

II) Financial Services

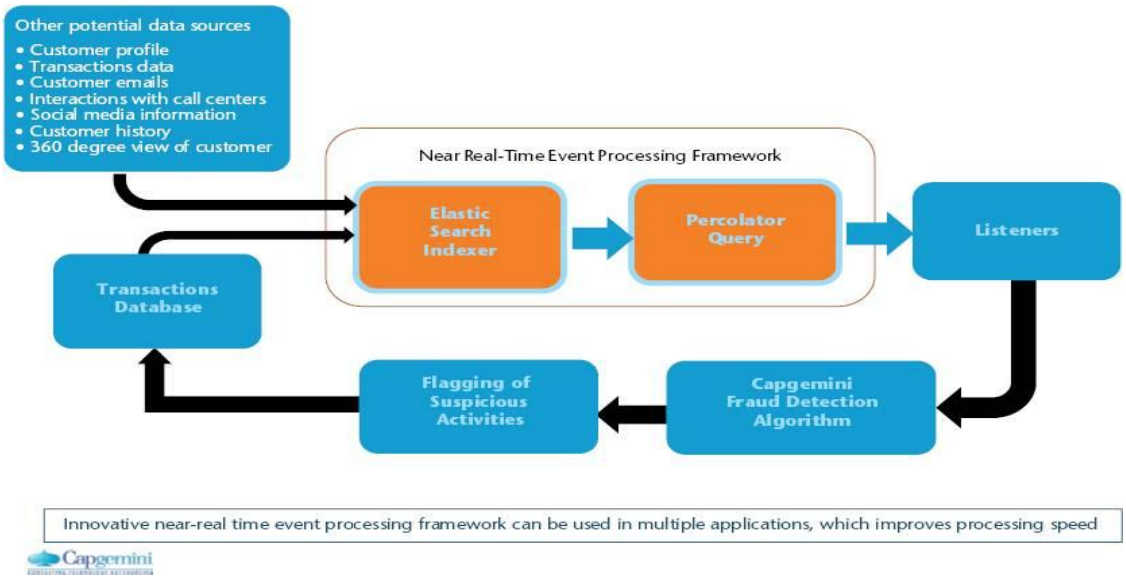
i) Fraud & Big Data

- Fraud is intentional deception made for personal gain or to damage another individual.
- One of the most common forms of fraudulent activity is credit card fraud.
- Social media and mobile phones are forming new frontiers fraud.
- Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the following three Vs :

1. **High volume:** Years of consumer records and transactions (150 billion +
2. records per year).
3. **High velocity:** Dynamic transactions and social media info.
4. **High variety:** Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

Fraud Detection Powered by Near Real-Time Event Processing Framework

- The near real-time event processing framework can be used in multiple applications which improves processing speed.



Innovative near-real time event processing framework can be used in multiple applications, which improves processing speed

Capgemini

- This fraud detection system uses an open source search server based on Apache Lucene. It can be used to search all kind of documents at near real-time. The tool is used to index new transactions which are sourced in real-time, which allows analytics to run in a distributed fashion utilizing the data specific to the index. Using this tool, large historical data sets can be used in conjunction with real-time data to identify deviation from typical payment patterns. The big data component allows overall historical patterns to be compared and contrasted and allows the number of attributes

and characteristics about consumer behavior to be very wide with little impact on overall performance.

- Percolator query performs the function of identifying new transactions that have raised profiles. Percolator query can handle both structured and unstructured data. This provides scalability to the event processing framework and allows specific suspicious transactions to be enriched with additional unstructured information (E.g. Phone location/geospatial records, customer travel schedules and so on). This ability to enrich the transaction further can reduce false positives and increase the experience of customer while redirecting fraud efforts to actual instances of suspicious activity.

- Capegemini's fraud Big Data initiative focuses on flagging the suspicious credit card transactions to prevent fraud in near real-time via multi-attribute monitoring. Real-time inputs involving transaction data and customers records are monitored via validity checks and detection rules. Pattern recognition is performed against the data to score and weight individual transactions across each of the rules and scoring dimensions. A cumulative score is then calculated for each transaction record and compared against thresholds to decide if the transaction is suspicious or not.

Social Network Analysis (SNA)

- This is another approach to solving fraud with Big data.
- SNA views social relationships and makes assumptions.
- SNA could reveal all individuals involved in fraudulent activity from perpetrators to their associates and understand their relationships and behavior to identify a bust out fraud case. Bust out is a hybrid credit and fraud problem and the scheme is typically defined by the following behavior.

    - The account in question is delinquent or charged off.
    - The balance is close to or over the limit.
    - One or more payments have been returned.
    - The customer cannot be located.
    - The above conditions exist with more than one account and/or financial institution.

- There are some Big Data solutions in the market like SAS's SNA solution, which helps institutions and goes beyond individual and account views to analyze all related activities and relationships at a network dimension. The network dimension allows visualization of social networks and helps to see hidden connections and relationships, which could be a group of fraudsters. There are huge amounts of data involved behind the scene, but the key to SNA solutions like SAS's is the visualization techniques for users to easily engage and take action.

ii) Risk and Big data

Types of Risk Management

- Credit risk management
- Market risk management
- Operational risk management (not common as credit & market)

-

The tactics of risk professionals include

- Avoiding risk
- Reducing the negative effect or probability of risk
- Accepting some or all of the potential consequences in exchange for potential gain.
- Credit risk analytics focus on past credit behaviors' to predict the likelihood that a borrower will default on any type of debt by failing to make payments which they accepted to do. E.g. Is this person likely to default on $300000 mortgage?
- Market risk analytics focus on understanding the likelihood that the value of a portfolio will decrease due to the change in stock prices, interest rates, foreign exchange rates and commodity prices. E.g. should we sell this share if the price drops another 10%?

Credit Risk Management

- Credit risk management focuses on reducing risks to acceptable levels that can boost profitability
- Risk data sources and advanced analytics are instrumental for credit risk management
- Social media and cell phone usage data are opening up new opportunities to whenever customer behavior that can be used for credit decisioning.
- There are four critical parts in the typical credit risk framework as illustrated in the following figure. They are planning, customer acquisition, account management and collections. All the four are handled through the use of Big data.

Credit Risk Framework

- As Figure illustrates, there are four critical parts of the typical credit risk framework: planning, customer acquisition, account management, and collections. All four parts are handled in unique ways through the use of Big Data.

iii) <u>Big Data and Algorithmic Trading</u>

- Financial institutions particularly investment banks have been at the forefront of applying analytics for risk management, proprietary trading and portfolio management.

- Many investment banks use algorithmic trading a highly sophisticated set of processes in which "insights" are made "actionable" via automated "decisions".
- Algorithmic trading relies on sophisticated mathematics to determine buy and sell orders for equities, commodities, interest rate and foreign exchange rates at blinding speed.
- A key component of algorithmic trading is determining return and risk of each trade and then making a decision to buy or sell.
- Algorithmic trading involves a huge number of transactions with complex interdependent data and every millisecond matters.

Crunching through complex interrelated data

- For market risk, the data explodes very quickly. Today, the portfolios being evaluated are quite large and include multiple financial instruments. Even for a single financial instrument it is possible to have hundreds of thousands of scenarios. Hence the problem of scenario analysis equity derivates different equities, different maturities and different strike prices becomes very complex**.**

Intraday Risk Analytics, a constant flow of Big Data

- Banks have moved from daily evaluation of risks to intraday risk evaluation and management.
- Intraday risk management involves pricing the entire portfolio and calculating the risk limits of each of the counter-parties within the bank's portfolio the problem gets complex and computationally intensive.
- For a bank to do their basic scenario analysis, it takes a million calculations for determining the value at risk for just one instrument. This must happen fast enough so that risk limits on the entire portfolio can be evaluated several times during the course of the day

Calculating Risk in Marketing

- While risk analytics is used for risk management, banks are using risk predictive analytics for marketing.
- When a bank scores its customers and prospects for credit card solicitations, it will use some risk management tools. This tool not only determines who has a high likelihood of responding to promotional offers but also considers the underlying risk for each of the prospects to whom the solicitations are being sent. Without taking risk profiles of individuals, bank promotion responses can result in customers with a higher risk profile
- One of the challenges for retail banks is to score large number of people for its marketing initiatives. The existing customers have to be scored to determine the borrowers whose probabilities of not paying on their credit card or on the mortgage is rising. Once these potential defaulters are identified steps could be taken to mitigate risk of default.
- The huge volume of population that have to be scored increases the problem .the scores have to be calculated quickly in order to take action promptly whether it is promotion or risk mitigation.

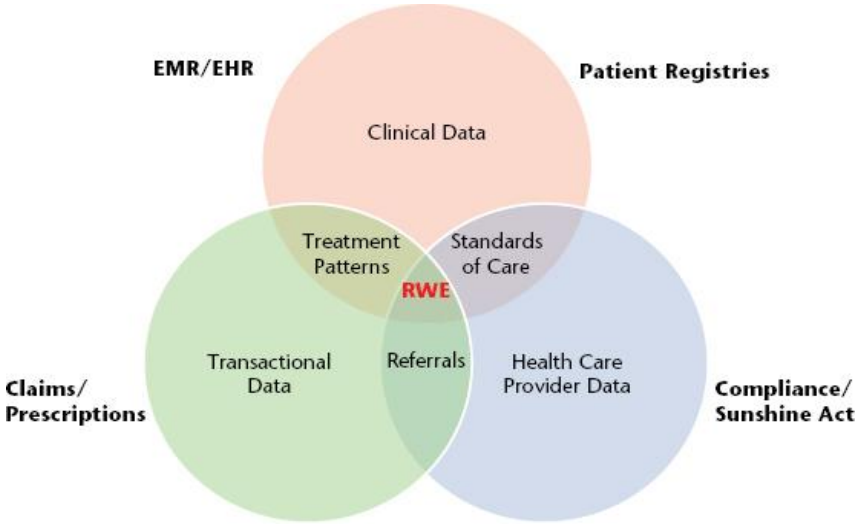
Benefits to other industries from financial services risk experience

- While adoption of analytics has been slower in industries such as retail, media and telecommunications momentum is starting to build around Bigdata Analytics.
- Marketing is an area that is mature in terms of adopting analytics. In marketing analytics can be used for marketing campaign management , targeted micromarketing(sending of different offers to different people depending on their likelihood to buy and market basket analysis which indicates what people buy together.
- In retail, forecasting is key area where analytics is being applied. Customer churn analysis has been used by banks to determine who is likely to cancel their credit card or account. This is the same technique used by telecommunication companies and retailers to determine customer defection. Churn is also a factor used in determining customer lifetime value. Customer lifetime value indicates how much money a firm can make, over the customer with the firm. Companies use their customer lifetime value to segment their customers and determine the customers to focus on.
- The insurance industry uses actuarial models for estimating losses. The emerging trend is to use Monte-Carlo simulations for estimating potential losses these computationally complex models require large hardware and hence adoption of analytics becomes difficult which has started changing with the advent of parallel and grid computing.
- Another use of Big Data analytics on banks is identifying manipulate behavior or fraudulent activities in real-time so that it can be mitigated or penalized immediately. For this we have to dig through voluminous transactions and find patterns quickly. A fair trading platform could be created by quickly catching and correcting market manipulating behavior.

III) Big data and Healthcare

- Big data promises enormous revolution in healthcare, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine.
- In addition to saving and improving lives, Big Data has the potential to transform the entire healthcare system by replacing guesswork and institution with objective data-driven science.
- The healthcare industry now has huge amount of data: from biological data such as gene expression, Special Needs Plans (SNPs), proteomics, metabolomics, and next-generation gene sequence data etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and Health Information Exchanges (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices.
- In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science (see Figure).

Figure: Data in the World of Healthcare



- The healthcare system is facing severe economic, effectiveness and quality challenges. These factors are forcing transformation in pharmaceutical business model. Hence the healthcare industry is moving from traditional model built on regulatory approval and settling of claims to medical evidence and proving economic effectiveness through improved analytics derived insights. The success of this model depends on the creation of robust analytics capability and harnessing integrated real-world patient level data.

Disruptive analytics

- Data science and disruptive analytics can have immediate beneficial impact on the healthcare systems.
- Data analytics makes it possible to create transparent approach to pharmaceutical decision making based on the aggregation and analysis of healthcare data such as electronic medical records and insurance claims data.
- Creating healthcare analytics framework has significant value for individual stakeholders.
- For providers (physicians), there is an opportunity to build analytics systems for evidence – based medicine(EBM) lifting through clinical and health outcomes data to determine the best clinical protocols that provide the best heath outcomes for patients and create defined standards of care.
- For producers(Pharmaceutical and medical device companies)there is an opportunity to build analytics systems to enable(transactional medicine) integrating externally generated post marketing safety, epidemiology and health outcomes data with internally generated clinical and discovery data (sequencing, expression, biomarkers) to enable improved strategic R&D decision making across the pharmaceutical value chain.
- For payers (ie, insurance companies) there is an opportunity to create analytics systems to enable comparative effectiveness research(CER) that will be used to drive reimbursement by mining large collections of claims, health care records(EMR/EHR), economic and geographic, demographic data sets to determine what treatment and therapies work best for which patients in which context and with what overall economic and outcomes benefit.

A Holistic Value Proposition

- The ability to collect, integrate, analyze and manage data can make health care data such as HER/EMR, valuable.
- Big data approach to analyze health care data creates methods and platform for analysis of large volumes of disparate kinds of data (Clinical, EMR, Claims, Labs etc.) to better answer questions of outcomes, epidemiology, safety, effectiveness and pharmaeconomic benefit.
- Big data technology and platforms such as Hadoop, R, Openhealthdata etc help clients create real-world evidence-based approaches to realize solutions for competitive effectiveness research, improve outcomes in complex populations and to improve decision making.

BI is not Data Science

- Traditional Business Intelligence and data warehousing skills do not help in predictive analytics. Like a lawyer who draws a conclusion and then looks for supporting evidence. Traditional BI is declarative and doesn't necessarily require any real domain understanding. Generating automated reports from aging data warehouses that are briefly scanned by senior management does not meet the definition of data science.
- Making data science useful to business is about identifying that question management really tries to answer question.

IV) Pioneering New Frontiers in Medicine

- In Medical Field, Big Data analytics are being used by researches to understand autoimmune disease such as Rheumatoid Arthritis, Diabetes and lupus and neurodegenerative disease such as multi sclerosis Parkinson's and Alzheimer's. In most these cases, the goal is to identify the genetic variations that causes the diseases. The data sets used for such identification contain thousands of genes. For example a research work on the role of environment factors and interactions between environmental factors in multiple sclerosis typically uses data sets that contain 100,000 to 500,000 genetic variations. The algorithms used to identify the interactions between environmental factors and diseases. They also have rapid search techniques built into them and should be able to do statistical analysis and permutation analysis which can be very, very time consuming if not properly done.

Challenges faced by pioneers of quantitative pharmacology

- The data set is very large 1000 by 2000 matrix.
- When an interactive analysis for first order and second order interactions are done each of the 500,000 genetic locations have to be compared to each of all the rest of the 500,000 genetics locations for the first order and this has to be done twice and then 500,000 may reduce to a third for the second order interaction and so on. Basically a second order interaction would be 500,000 squared, a third order would be 500,000 cubed and so on. Such huge computations are made possible in little time with the aid of big data technologies.

V) Advertising and Big Data

Big Data is changing the way advertisers address three related needs.

(i) How much to spend on advertisements.
(ii) How to allocate amount across all the marketing communication touch points.
(iii) How to optimize advertising effectiveness. Given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction.

Reach, Resonance, and Reaction

Reach: First part of reach is to identify the people who are most volumetrically responsive to their advertising and then answer questions such as what do those people watch? What do they do online? How to develop media plan against intended audience. The second part of reach is delivering advertisements to the right audience. That is, to understand if we are actually reaching our desired audience. If we think about the online world, it's a world where we can deliver 100 million impressions but we never really know for sure who our campaign was actually delivered to. If our intended audience is women aged 18 to 35, of our 100 million impressions, what percentage of impressions were actually delivered to the intended audience? What was the reach, what was the frequency, what was the delivery against the intended audience?

Resonance: If we know whom we want to reach and we're reaching them efficiently with your media spend, the next question is, are our ads breaking through? Do people know they're from our brand? Are they changing attitudes? Are they making consumers more likely to want to buy our brand? This is what is called "resonance."

Reaction: Advertising must drive a behavioral "reaction" or it isn't really working. We have to measure the actual behavioral impact. Whether we've identified the

highest potential audience, reached them efficiently with our media plan, delivered ads that broke through the clutter and increased their interest in buying our brand—did it actually result in a purchase? Did people actually buy our product or service based on exposure to our advertising?
The three guiding principles to measurement are
1. End to end measurement—reach, resonance and reaction
2. Across platforms (TV, digital, print, mobile, etc.)
3. Measured in real-time (when possible)

The Need to Act Quickly (Real-Time When Possible)

When we start executing a campaign, how do we know on a daily basis whether our advertising campaign is actually being delivered to our intended audience the way it's supposed to?
For example, in digital, ad performance will differ across websites. Certain websites are really good; certain websites are really bad. How do we optimize across sites? - By moving money out of weak performing sites and into better performing sites.

Real time optimization:
When new ad campaign is launched, it is good if the ad is break thru and is highly memorable and it is bad news if the consumers think the ad is for the key competitor.

If out of three ads aired, two have high breakthrough but one is weak, the weak performing ad could be quickly taken off air and the media spend can be rotated to the higher performing ads. This will make breakthrough scores go up.

Instead of 30 second ads, a mix of 15s and 30s ads, can be planned. Suppose real time data shows that 15s ads work as well as 30s ads. Instead of spending money on 30s ads, all money can be spent on 15-second ads and scores will continue to grow.

The measurement tools and capabilities are enabling real-time optimization this and so there's a catch-up happening both in terms of advertising systems and processes, but the industry infrastructure must be able to actually enable all of this real-time optimization.

Measurement Can Be Tricky

There are tools that allow us to tag digital advertising through a panel, we can read those people who were exposed to the advertising and those who were not and measure their actual offline purchase behavior.
A company doing this for a large beer client could see that this campaign generated (after the fact) a 20 percent sales increase among consumers exposed versus not exposed to the advertising. The average person would look at that and think that the advertising is working. But when the reach for this particular client was analyzed, their intended audience was males, aged 21–29. Of their 100 million delivered impressions, only about 40 million were actually delivered to males aged 21–29. Sixty million went to someone other than their intended audience; some went to kids (not good for a beer brand); some went to people 65+. It would have been good if instead of 40% of the impressions hitting the intended audience, 70 or 80% of the impressions had hit them. When the 40 percent of impressions that hit the intended audience were analyzed, the reach and frequency of those was 10 percent reach 65 frequency. In

other words, they only hit about 10 percent of their intended audience, but each of these people was bombarded with, on average, 65 ads. That's not quite the optimization one would hope for. There's a lot of science in advertising that shows that by maximizing reach and minimizing frequency, we can get best response. If the measuring of all of this were in real time, the plan could have been quickly adjusted to increase delivery to the intended audience, increase reach, and reduce frequency.

Content Delivery Matters

The ads were on twelve websites: four of the ads didn't perform well in those sites. The other ones were really good. If they had measured that in flight, they could have moved spending out of the bad performing sites, into good performing sites, and further improved results. End-to-end measurement is important. Reach times resonance equals reaction. Measuring the sales impact alone is great, but it's not enough. Sales impact could be great and still be completely non-optimized on the reach and resonance factors that cause the reaction.

Optimization and Marketing Mixed Modeling

Marketing Mixed Modeling (MMM) is a tool that helps advertisers understand the impact of their advertising and other marketing activities on sales results. MMM can generally provide a solid understanding of the relative performance of advertising by medium (e.g., TV, digital, print, etc.), and in some cases can even measure sales performance by creative unit, program genre, website, and so on.

Now, the impact on sales in social media can be measured through market mixed modeling. Market mixed modeling is a way that can take all the different variables in the marketing mix—including paid, owned, and earned media—and uses them as independent variables that regress against sales data and tries to understand the single variable impact of all these different things.

Since these methods are quite advanced, organizations use high-end internal analytic talent and advanced analytics platforms such as SAS or point solutions such as Unica and Omniture. Alternatively, there are several large analytics providers like Mu Sigma that supply it as a software-as-a-service (SaaS).

As the world becomes more digital, the quantity and quality of marketing data is improving, which is leading to more granular and insightful MMM analyses.

The Three Big Data Vs in Advertising

Impact of the three Vs (volume, velocity, and variety) in advertising:

Volume

The volume of information and data that is available to the advertiser has gone up exponentially versus what it was 20 years ago. In the old days, we would copy test our advertising. The agency would build a media plan demographically targeted and we'd execute it. Maybe 6 to 12 months later, we'd try to use whatever sales data we had to try to understand if there was any impact. In today's world, there is hugely more advertising effectiveness data. On TV advertising, we can measure every ad in every TV show every day, across about 70 percent of the viewing audience. We measure clients digital ad performance hourly—by ad, by site, by exposure, and by audience. On a daily or weekly basis, an advertiser can look at their advertising performance.

Velocity

There are already companies that will automate and optimize advertising on the web without any human intervention at all based on click-thru. It's now beginning to happen on metrics like breakthrough, branding, purchase intent etc. This is sometimes called programmatic buying. Literally, we'll have systems in place that will be measuring the impact of the advertising across websites or different placements within websites, figuring out where the advertising is performing best. It will be automated optimization and reallocation happening in real-time. The volume and the velocity of data, the pace at which we can get the data, make decisions and do things about it is dramatically increased.

Variety

Before, we really didn't have a lot of data about how our advertising was performing in market. We have a lot more data and it's a lot more granular. We can look at our brand's overall advertising performance in the market. But we can also decompose it to how much of a performance is because of the creative quality, the media weight, how much is because of the program that the ads sit in. How much is because of the placement: time of day, time of year, pod position, how much is because of cross-platform exposure, how much is because of competitive activity. Then we have the ability to optimize on most of those things—in real time. And now we can also measure earned (social) and owned media. Those are all things that weren't even measured before.

Using Consumer Products as a Doorway

Over the last year, RIM's Research in Motion (RIM), the maker of Blackberry share price has plunged 75 percent. The company once commanded more than half of the American smartphone market. Today it has 10 percent. Companies with deep pockets and resources such as Hewlett Packard (HP) failed to enter the tablet market, while Apple is selling iPads in its sleep.

Apple entered into the mobile and tablet market because of the iPod, which crushed giants like Sony in the MP3 market. For Apple, the market was not just about selling hardware or music on iTunes. It gave them a chance to get as close to a consumer as anyone can possibly get. This close interaction also generated a lot of data that help them expand and capture new customers. Again it's all about the data, analytics, and putting it into action.

Google gives away product that other companies, such as Microsoft, license for the same the reason. It also began playing in the mobile hardware space through the development of the Android platform and the acquisition of Motorola. It's all about gathering consumer data and monetizing the data. With Google Dashboard we can see every search we did, e-mails we sent, IM messages, web-based phone calls, documents we viewed, and so on. This is powerful for marketers.

The online retailer Amazon has created new hardware with Kindle and Barnes and Noble released the Nook. If both companies know every move we make, what we download, what we search for, they can study our behaviors to present new products that they believe will appeal to us. The connection with consumers and more importantly taking action on the derived data is important to win.

BIG DATA TECHNOLOGY

- Hadoop's Parallel World
  - Hadoop Distributed File System(HDFS)
  - Map Reduce
- Old Vs New Approaches
- Data Discovery
- Open Source Technology for Big Data Analytics
- The cloud and Big Data
- Predictive Analytics
- Software as a Service BI
- Mobile Business Intilligence
- Crowdsourcing Analytics
- Inter and Trans Firewall Analytics
- R & D Approach

i) Hadoop's parallel World(Creators-Doug cutting was at yahoo now at cloudera, Mike Cafarella-teaching at University of Micigan)

Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive value from all their data.

Advantages of Hadoop:

- The scalability and elasticity of free open-source Hadoop running on standard hardware allow organizations to hold onto more data and take advantage of all their data to increase operational efficiency and gain competitive edge. Hadoop supports complex analyses across large collections of data at one tenth the cost of traditional solutions.
- Hadoop handles a variety of workloads, including search, log processing, recommendations systems, data warehousing and video/image analysis.
- Apache Hadoop is an open-source project by the Apache Software foundations. The software was originally developed by the world's largest Internet companies to capture and analyze the data that they generate. Unlike traditional, structured platforms Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformation on that data. Hadoop stores terabytes and even peta bytes of data inexpensively. It is robust and reliable and handles hardware and system failures automatically without losing data analyses.
- Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

Components of Hadoop:

The two critical components of Hadoop are

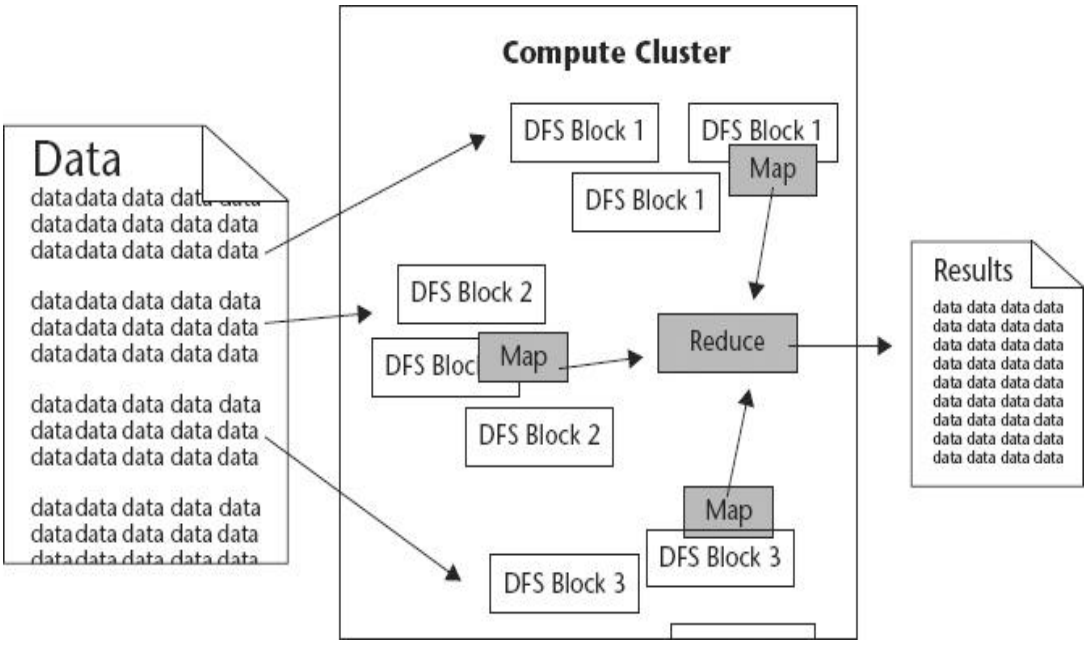
1) The Hadoop Distributed File System (HDFS)

- HDFS is the storage system for a cluster.
- When data lands in the cluster, HDFS breaks it into pieces and distribute those pieces among the different servers participating in the cluster.
- Each server stores just a small fragment of the complete data set and each piece of data is replicated on more than one server.

2) Map Reduce:

- Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed in parallel to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its result back for collation into a comprehensive answer. Map Reduce is the agent that distributes the work and collects the results.
- Both HDFS and Map Reduce are designed to continue to work even if there are failures.
- HDFS continuously monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails or data is damaged due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster.
- Map Reduce monitors the progress of each of the servers participating in the job, when an analysis job is running. If one of them is slow in returning an answer or fails before completing its work, Map Reduce automatically starts another instance of the task on another server that has a copy of the data.
- Because of the way that HDFS and Map Reduce work, Hadoop provides scalable, reliable and fault-tolerant services for data storage and analysis at very low cost.

Figure-DFS & Map Reduce



ii) Old Vs New approaches to Data Analytics

Old Approach (Database approach)	**New Approach** (Big data Analytics)
Follows data and analytics technology stack with different layers of cross-communicating data and working on "scale-up" expensive hardware.	Follows data and analytics platform that does all the data processing and analytics in one layer without moving data back and forth on cheap but scalable ("scale-out") commodity hardware.
Data is moved to places where they have to be processed.	Data must be processed and converted into usable business intelligence where it sits.
Massive parallel processing was not employed due to hardware and storage limitations.	Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing.
Due to technological limitations storing, managing and analyzing massive data sets were difficult.	New proprietary technologies and open source inventions enable different approaches that make it easier and more affordable to store, manage & analyze data.
Not able to handle unstructured data.	The variety of data and ability to handle unstructured data is on the rise. Big data approach provides solution to this.

iv) Data Discovery

- Data discovery is the term used to describe the new wave of business intelligence that enables users to explore data, make discoveries and uncover insights in a dynamic and intuitive way versus predefined queries and preconfigured drill-down dashboards. This approach is being followed by many business users due its freedom and flexibility to view Big Data. There are two software companies that stand out in the crowd by growing their businesses at unprecedented rates in this space: Tableau Software and QlikTech International.
- Both companies' approach to the market is much different than the traditional BI software vendor. They used a sales model referred to as "land and expand". In order to succeed at the BI game of the "land and expand model", we need a product that is easy to use with lots of attractive output. Analytics and reporting are produced by the people using the results. IT provides the infrastructure, but business people create their own reports and dashboards.
- The most important characteristic of rapid-fire BI is that business users, not specialized developers, drive the applications. The result is that everyone wins. The IT team can stop the backlog of change requests and instead spend time on strategic IT issues. Users can serve themselves data and reports when needed.

- There is a simple example of powerful visualization that the Tableau team is referring to. A company uses an interactive dashboard to track the critical metrics driving their business. Every day, the CEO and other executives are plugged in real-time to see how their markets are performing in terms of sales and profit, what the service quality scores look like against advertising investments, and how products are performing in terms of revenue and profit. Interactivity is key: a click on any filter lets the executive look into specific markets or products. She can click on any data point in any one view to show the related data in the other views. She can look into any unusual pattern or outlier by showing details on demand. Or she can click through the underlying information in a split-second.
- Business intelligence needs to work the way people's minds work. Users need to navigate and interact with data any way they want to—asking and answering questions on their own and in big groups or teams.
- Qliktech has designed a way for users to leverage direct—and indirect—search. With QlikView search, users type relevant words or phrases in any order and get instant, associative results. With a global search bar, users can search across the entire data set. With search boxes on individual list boxes, users can confine the search to just that field. Users can conduct both direct and indirect searches. For example, if a user wanted to identify a sales rep but couldn't remember the sales rep's name—just details about the person, such as that he sells fish to customers in the Nordic region—the user could search on the sales rep list box for "Nordic" and "fish" to narrow the search results to just the people who meet those criteria.

v) Open Source Technology for Big Data Analytics

- Open- source software is computer software that is available in source code form under an open- source license that permits users to study, change, improve and distribute the software.
- Hadoop is a open- source project.
- One of the key attributes of open- source projects is that it is not constrained by someone else's predetermined ideas or vision which makes it flexible, extensible and low cost.
- One disadvantage of open- source is that it has to coexist with the proprietary solution for a long time for many reasons. for example getting data from hadoop to a database required a hadoop expert in the middle to do the data cleansing and the data type translation .If the data was not 100% (clean which is the case with most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn't directly access and analyze data in hadoop clusters. SQL-H is software that is developed to solve this problem.

v) The Cloud and Big Data

- Market economics are demanding that capital- intensive infrastructure costs disappear and business challenges are forcing clients to consider newer models. The cloud-deployment model satisfies such needs. With a cloud model, payment is on subscription basis with no capital expense. Typical 30% maintenance fees are not incurred and all the updates on the platform are automatically available.

- The traditional cost of value chains is completely eliminated by massively scalable platforms (such as cloud) where marginal cost to deliver an incremental product/service is zero. Whether a private hosted model or a publicly shared one, the true value lies in delivering software, data and/or analytics in an "as a service' model.

vi)Predictive Analytics

- Enterprises will move from being in reactive positions (Business Intelligence) to forward learning positions (Predictive analysis). Using all the data available i.e traditional internal data sources combined with new rich external data sources will make the predictions more accurate and meaningful. Algorithm trading and supply chain optimizations are two examples where predictive analytics have greatly reduced the friction in business. Predictive analytics proliferate in every facet of our lives both personal and business. Some of the leading trends in business today are
- Recommendation engines similar to those used in Netflix, Amazon that use past purchases and buying behavior to recommend new purchases.
- Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk and portfolio risk.
- Innovation engines for new product innovation, drug discovery and consumer and fashion trends to predict new product formulations and new purchases.
- Consumer insight engines that integrate a wide variety of consumer-related information including sentiment, behavior and emotions.
- Optimization engines that optimize complex interrelated operations and decisions that are too complex to handle.

vii) Software as a Service Business Intelligence

- The basic principle is to make it easy for companies to gain access to solutions without building and maintaining their own onsite implementation.
- SaaS is less expensive.
- The solutions are typically sold by vendors on a subscription or pay-as-you-go basis instead of the more traditional software licensing model with annual maintenance fees.
- SaaS BI can be a good choice when there is little or no budget money available for buying BI software and related hardware. Because they don't involve upfront purchase costs or additional staffing requirements needed to manage the BI system, total cost of ownership may be lower than that of on-premise software.
- Another common buying factor for SaaS is the immediate access to talent especially in the world of information management, BI and predictive analytics.
- Omniture (now owned by adobe) is a successful SaaS BI. Omniture's success was due to its ability to handle big data in the form web log data.

Omniture's success was also due to the following reasons:

- **Scaling the Saas delivery model**
  Omniture was built from the ground up to be SaaS. It made use of concept called magic number that helps analyze and understand SaaS business.
- **Killer sales organizations**
  Omniture's well known customers like HP, eBay and Gannet were sale organizations.

- **A focus on customer success**
  Unlike traditional enterprise software, with SaaS business, it is easy for customers to leave if they are not satisfied. Today's BI is not designed for the end-user. It is not intuitive, not accessible, not real time and it thus not meet the expectations of today's customer technology who expect a mu8ch more connected experience.

viii) Mobile Business Intelligence

- Simplicity and ease of use had been the major barriers to BI adoption. But mobile device have made complicated actions to be performed very easily. For example, a young child can use an ipad or iphone easily but not a laptop. This ease of use will drive the wide adoption of mobile BI.
- Multi touch and software oriented devices have brought mobile analytics and intelligence to a much wider audience.
- Ease of mobile application development and development have also contributed to the wide adoption of mobile BI.

Three elements that have impacted the viability of mobile BI are

i) Location-GPS component enables finding location easy.
ii) Transaction can be done through smart phones.
iii) Multimedia functionality allows virtualization.

Three challenges with mobile BI include

i) Managing standards for these devices.
ii) Managing security (always a big challenge).
iii) Managing "bring your own device", where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

ix) Crowdsourcing Analytics

- Crowdsourcing is the recognition that organizations can't always have the best and brightest internal people to solve all their big problems. By creating an open, competitive environment with clear rules and goals problems can be solved.
- In October 2006, Netflix an online DVD rental business announced a contest to create a new predictive model for recommending movies based on past user ratings. The grand price was $1,000,000. Netflix already had an algorithm to solve the problem but thought there was an opportunity to improve the model which would turnout huge revenues.
- Kaggle is an Australian firm that provides an innovative solutions for statistical/analytics for outsourcing. Kaggle manages competitions among world's best data scientist corporation, governments and research laboratories that confront complex statistical challenges describe the problems to kaggle and provide data sets. Kaggle converts the problems and the data into contests that are posted on its website. The contest feature cash prizes ranging in values from $100 to $3 million. Kaggle's clients range in size from tiny start-ups to Multinational Corporations such as Ford Motor Company and government agencies such as NASA.

- The idea is that someone comes to Kaggle with a problem, they put it up on their website and then people from all over the world can compete to see who can produce the best solution. In essence Kaggle has developed an effective global platform for crowdsourcing complex analytic problems.
- There are various types of crowdsourcing such as crowd voting, crowd purchasing, wisdom of crowds, crowd funding and contests.
- Example:
  - ❖ 99designs.com/, does crowdsourcing of graphic design.
  - ❖ Agentanything.com/, posts missions where agents are invited to do various jobs.
  - ❖ 33needs.com/, allows people to contribute to charitable programs to make social impact.

x) Inter and Trans-Firewall Analytics

- Yesterday companies were doing functional silo-based analytics. Today they are doing intra-firewall analytics with data within the firewall. Tomorrow they will be collaborating on insights with other companies to do inter-firewall analytics as well as leveraging the public domain spaces to do trans-firewall analytics (Fig.1).

- As fig.2 depicts, setting up inter-firewall and trans-firewall analytics can add significant value. But this presents some challenges. When information is collected outside the firewall, the information to noise ratio increases, putting additional requirements on analytical methods and technology requirements.

- Further, organizations are limited by a fear of collaboration and overreliance on proprietary information. The fear of collaboration is driven by competitive fears, data privacy concerns and proprietary orientations that limit opportunities for cross-organizational learning and innovations. The transition to an inter-firewall and trans-firewall paradigm may not be easy but it continues to grow and will become a key weapon for decisions scientists to drive disruptive value and efficiencies.

Figure 1

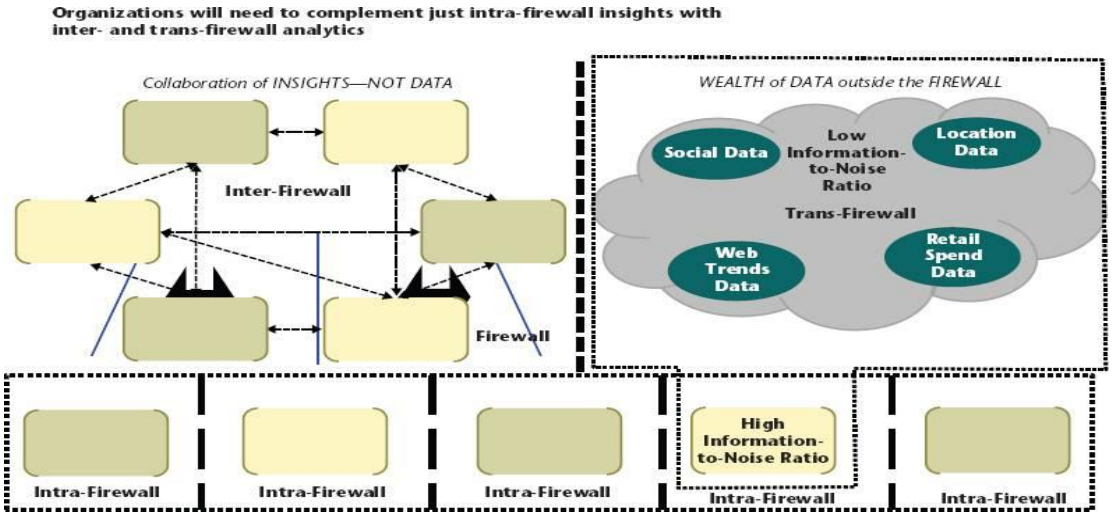
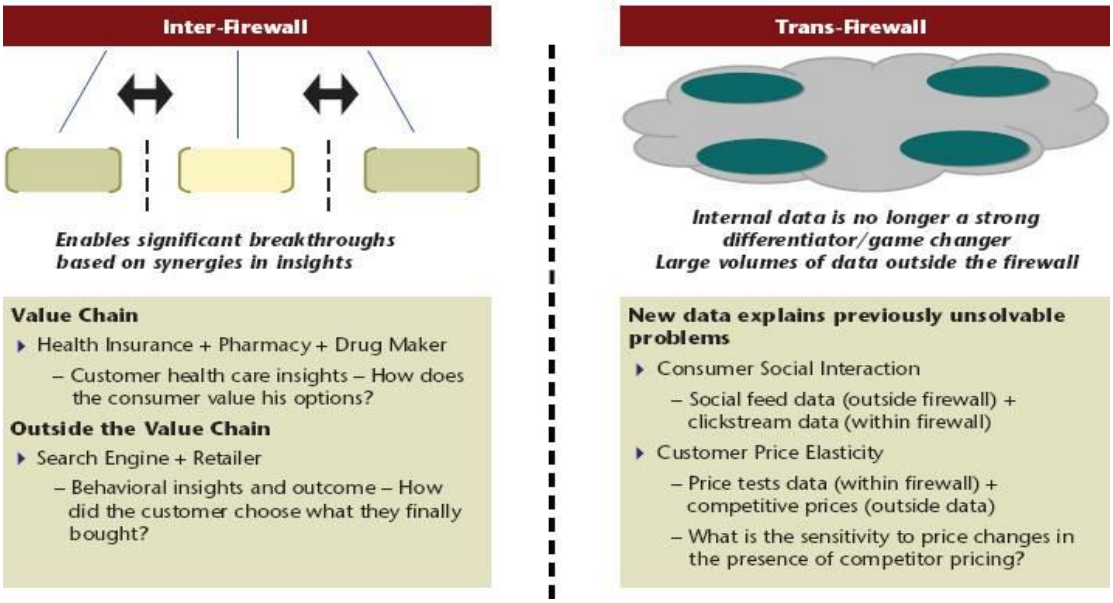


Figure 2

Disruptive value and efficiencies can be extracted by cooperating and exploring outside the boundaries of the firewall



xi) R & D Approved helps adapt new technology

- Business analytics can certainly help to a company embrace innovation and direction by leveraging critical data and acting on the results. For example, a market research executive analyzes customers and market data to anticipate new product features that will attract customers.

- For many reasons, organizations find it hard to make changes after spending many years implementing a data management, BI and analytic stack. So the organizations have to do lot of research and development on the new technologies before completely adopting the technologies to minimize the risk. The two core programs that have to focused but R & D teams are

Program	Goal	Core Elements
Innovation management	Tap into the latent creativity of all Visa employees, providing them with a platform to demonstrate mastery and engage collaboratively with their colleagues.	Employee personal growth
		Employee acquisition and retention
Research and open innovation	Look outside of the company and scan the environment for trends, new technology, and approaches	Innovation
		Competitive advantage

Adding Big Data Technology

The process of enterprises must follow to get started with the big data technology

1. Practical approach – Start with the problem and then find a solution.

2. Opportunistic Approach – Start with the technology and then find a home for it.

For the both approach the following activities have to be conducted,

(i) **Play** – R & D team members may request to install their lab to get more familiar with the technology.

(ii) **Initial Business review** – Talk with the business owner to validate the applicability and rank, the priorities to ensure that it is worth pursuing.

(iii) **Architecture Review** – Asses the validity of the underlying architecture and ensure that it maps to IT's standards.

(iv) **Pilot use cases** – find the use case to test the technology out.

(v) **Transfer Form R & D to PRODUCTION** – Negotiate internally regarding what it would take to more it from research to production using the following table.

Who needs to be involved in this process?

Operations	Nonfunctional integration	How does this integrate with our existing programs? (Monitoring, security, intrusion, etc.)
Engineering	Readiness	What do we need to do to prepare from a skill and infrastructure perspective to take on this project?
Application development	Functional requirements	How does this new technology map to the business functional requirements?
Business users	Derived value	How can new technology solve business problems I have or anticipate? How can I find new business opportunities?

Organizations may have a lot of smart people, but there are other smart people outside. Organizations need to be exposed to the value they are creating. A systematic program that formalizes relationships with a powerful ecosystem is shown in the following table.

Innovation ecosystem: Leveraging brain power from outside of your organization

Source	Example
Academic community	Tap into a major university who did a major study on social network analytics.
Vendors research arms	Leverage research a vendor completed in their labs demonstrating success leveraging unstructured data.
Research houses	Use research content to support a given hypothesis for a new endeavor.
Government agencies	Discuss fraud strategies with the intelligence community.
Venture capital orgs	Have a venture capital firm review some new trends they are tracking and investing in.
Start-ups	Invite BI and analytic technology start-ups in instead of just sticking with the usual suspects.