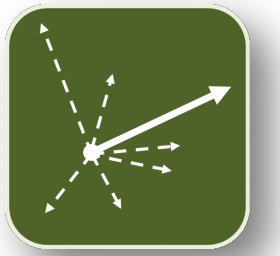
Waterford Institute *of* Technology

BIG DATA 2

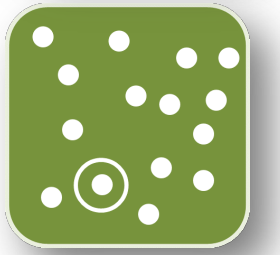
SAMITHA SOMATHILAKA

Department of Computing & Mathematics, WIT

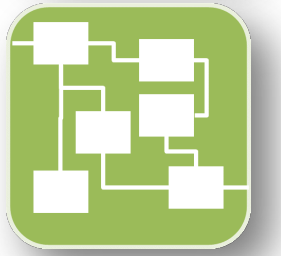
BIG DATA OPPORTUNITIES



Making better informed decisions
e.g. strategies, recommendations

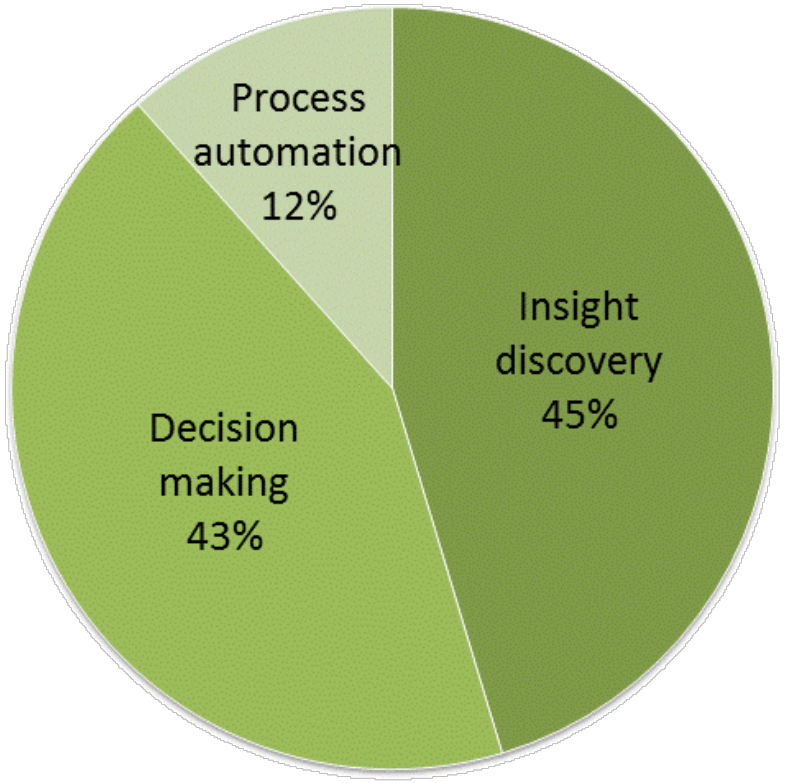


Discovering hidden insights
e.g. anomalies forensics, patterns, trends



Automating business processes
e.g. complex events, translation

WHICH IS THE BIGGEST OPPORTUNITY FOR BIG DATA IN YOUR ORGANIZATION?



- 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage.
- Business analytics needs will drive 70% of investments in the expansion and modernization of information infrastructure.

IDENTIFYING INSURANCE FRAUD

- Opportunity
  - Save and make money by reducing fraudulent auto insurance claims
- Data & Analytics
  - Predictive analytics against years of historical claims and coverage data
  - Text mining adjuster reports for hidden clues, e.g. missing facts, inconsistencies, changed stories
- Results
  - Improved success rate in pursuing fraudulent claims from 50% to 88%; reduced fraudulent claim investigation time by 95%
  - Marketing to individuals with low propensity for fraud

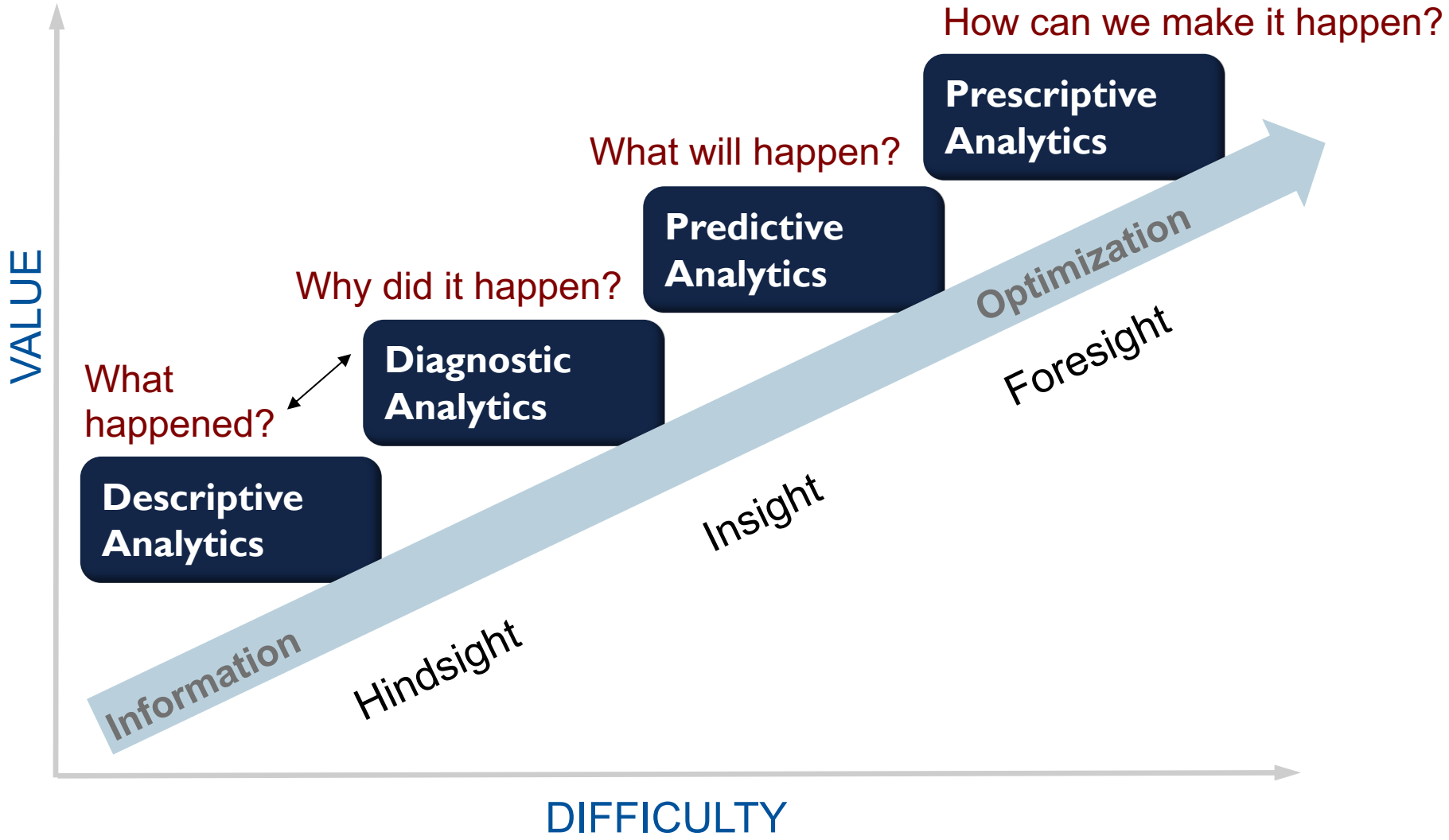
QUALITY IMPROVEMENT

- Opportunity
  - Move from manual to automated inspection of burger bun production to ensure and improve quality
- Data & Analytics
  - Photo-analyze over 1000 buns-per-minute for color, shape and seed distribution
  - Continually adjust ovens and process automatically
- Result
  - Eliminate 1000s of pounds of wasted product per year; speed production; save energy; Reduce manual labor costs

Is the company using all of its “senses” to observe, measure and optimize business processes?

BIG DATA ANALYTICS

ANALYTICS MODELS

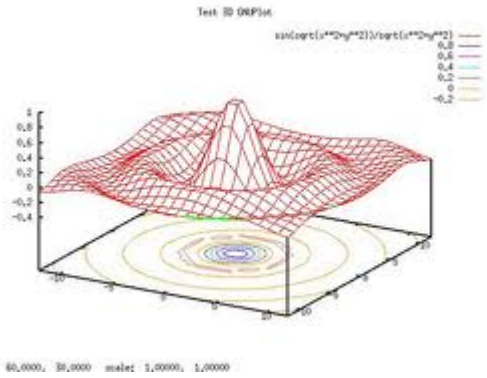


DESCRIPTIVE ANALYTICS

- Descriptive analytics, such as reporting/OLAP, dashboards, and data visualization, have been widely used for some time.
- They are the core of traditional BI.

Year 2000				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$6,851,006.49	$7,132,961.38	$4,322,514.74	$4,526,954.71
Marketing Expense	$750,179.20	$756,596.17	$455,048.05	$462,815.40
Research and Development Expense	$538,243.39	$538,014.73	$329,890.95	$336,808.13
Selling Expense	$1,632,921.64	$1,579,790.18	$986,887.49	$927,970.90
Taxes	$314,659.05	$319,390.19	$202,636.67	$200,205.01

Year 2001				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$2,554,556.31	$2,700,773.16	$1,726,031.16	$1,773,448.08
Marketing Expense	$294,766.22	$290,696.70	$187,757.29	$176,778.55
Research and Development Expense	$200,719.90	$193,236.83	$134,270.95	$125,725.88
Selling Expense	$620,427.30	$611,649.47	$405,092.93	$400,181.91
Taxes	$130,926.70	$122,526.31	$82,450.78	$80,671.87

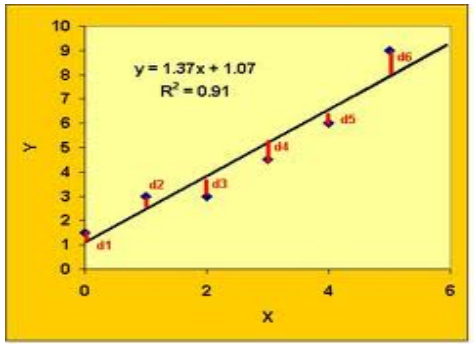


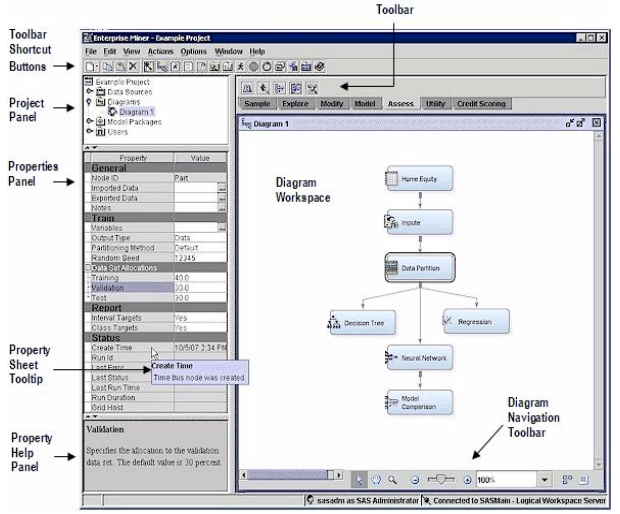
What has occurred?

- Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and predictive analytics.

PREDICTIVE ANALYTICS

- Algorithms for predictive analytics, such as regression analysis, machine learning, and neural networks, have also been around for some time.



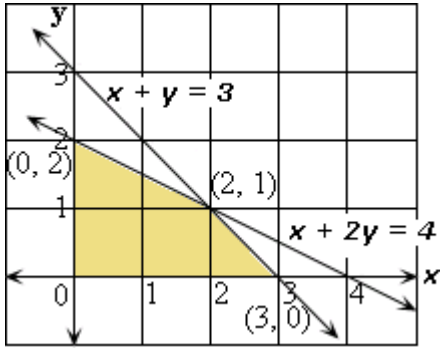


What will occur?

- Marketing is the target for many predictive analytics applications.
- Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and prescriptive analytics.

PRESCRIPTIVE ANALYTICS

- Prescriptive analytics are often referred to as advanced analytics.
- Prescriptive analytics provide an optimal solution, often for the allocation of scarce resources.
- They, too, have been in academia for a long time but are now finding wider use in practice.
- For example, the use of mathematical programming for revenue management is common for organizations that have “perishable” goods (e.g., rental cars, hotel rooms, airline seats).



TECHNOLOGIES FOR HANDLING BIG DATA

BIG DATA TECHNIQUES

- To analyse the datasets, there are many techniques available, some of which are as follows:
  - Massive Parallelism
  - Data Distribution
  - High-Performance Computing
  - Task and Thread Management
  - Data Mining and Analytics
  - Data Retrieval

DISTRIBUTED AND PARALLEL COMPUTING FOR BIG DATA

- Distributed computing is a method in which multiple computing resources are connected in a network and computing tasks are distributed across the resources, thereby increasing the computing power. Distributed computing is faster and more efficient than traditional computing, and, hence, of immense value when it comes to processing a huge amount of data in a limited time.
- To carry out complex computations, the processing power of a standalone personal computer can also be enhanced by adding multiple processing units, which can carry out the processing of a complex task by breaking it up into sub-tasks, and carrying out individual sub-tasks simultaneously. Such systems are often termed as parallel systems. The greater the processing power, the faster the computing.

BIG DATA TECHNIQUES: MASSIVE PARALLELISM

- According to the simplest definition available, a parallel system is a system where multiple processors are involved and associated to carry out the concurrent computations.
- Massive parallelism refers to a parallel system where multiple systems interconnected with each other pose as a single mighty conjoint processor and carry out tasks received from the data sets parallelly.
- In terms of Big Data dynamics, the systems can not only be processor, but also memory, hardware and even network conjoint to scale up the operational efficiency posing as a massive system that can eat humongous datasets parallelly without breaking a sweat.

DISTRIBUTED AND PARALLEL COMPUTING FOR BIG DATA

- Following Table differentiates between distributed and parallel computing systems:

Distributed Computing System	Parallel Computing System
An independent, autonomous system connected in a network for accomplishing specific tasks	A computer system with several processing units attached to it
Coordination is possible between connected computers that have their own memory and CPU	A common shared memory can be directly accessed by every processing unit in a network
Loose coupling of computers connected in a network that provides access to data and remotely located resources	Tight coupling of processing resources that are used for solving a single, complex problem

CLOUD COMPUTING AND BIG DATA

- One of the vital issues that organisations face with the storage and management of Big Data is the huge amount of investment to get the required hardware setup and software packages.
- Some of these resources may be overutilised or underutilised with varying requirements overtime. We can overcome these challenges by providing a set of computing resources that can be shared through cloud computing.
- The cloud computing environment saves costs related to infrastructure in an organisation by providing a framework that can be optimised and expanded horizontally.

BIG DATA TECHNIQUES: HIGH-PERFORMANCE COMPUTING

- High-performance computing is the simultaneous use of supercomputers and parallel processing techniques for solving intricate computation problems.
- It emphasises on making parallel processing systems and algorithms by joining both parallel and administrative computational methods.

BIG DATA TECHNIQUES: HIGH-PERFORMANCE COMPUTING

- The words supercomputing and high-performance computing are often used to resemble each other.
- High-performance computing is used for performing research activities and cracking advanced problems through computer simulation, modelling and analysis.